Biostatistics: Data and Models

John D. Reeve
Southern Illinois University Carbondale
Carbondale, IL 62901

December 2022

Copyright Notice

Acknowledgments

I would like to thank my parents, Ann M. and John H. Reeve, and my partner Kimberly A. Cole, for their enduring support. I would also like to thank James T. Cronin, Luis Miguel Flores-Campaña, Jamie M. Kneitel, and Fernando T. Maestre for providing data sets used in this book.

Contents

Chapter 1

Introduction

1.1 Why this textbook?

Welcome to **Biostatistics: Data and Models**! This textbook provides a survey of statistical methods commonly used in the life sciences, an introduction to statistical theory, and significant exposure to the statistical software package SAS 9.4 (©2020 SAS Institute Inc.). The textbook is designed for graduate students and upper division undergraduates in the life sciences, and assumes some familiarity with mathematical notation, functions, and algebra. A review of these topics is also presented early in the textbook. Knowledge of calculus is helpful but not essential. No previous courses in statistics are needed.

There are many useful introductory statistical textbooks (e.g., Sokal & Rohlf 1995, Steel et al. 1997, Schork & Remington 2000), so what is different about this one? One is the close integration of the text with SAS programs and output. Many texts do not discuss a particular software package, provide only abbreviated examples, or present them under separate cover. However, these packages play an essential role in modern statistical analyses, and fluency in a statistical language is a basic tool for the practicing scientist. I selected SAS as the statistical package for this textbook because of its popularity, extensive documentation, and strong support of mixed models, a common statistical procedure. An alternative is the free software package called R (R Core Team 2021). For those interested in learning this software, R programs similar in function to the SAS code can be downloaded at the website for this text.

Another difference in this textbook is the integration of statistical procedures and theory. Most introductory textbooks present the statistical procedures and a mechanistic explanation of how they work, without discussing the underlying theory. The theory is typically presented in advanced courses to a more mathematically inclined audience. However, I feel that some knowledge of the theory is essential for students in the life sciences, and so some theoretical concepts are included in this text. For example, likelihood is used throughout the text to explain how parameters are estimated and statistical tests derived. Besides many basic statistical procedures, likelihood theory also plays a role in model building and selection using information criteria, as well as Bayesian statistics, an expanding field of statistical analysis.

As part of this integration of theory, statistical models are presented throughout the text. What is a statistical model? Suppose we are interested in fitting a line through some data points, which are in the form of (Y, X) pairs. A standard statistical model for fitting a line through such data is the linear regression model:

$$Y = \alpha + \beta X + \epsilon, \tag{1.1}$$

where α is the intercept of the line, β is the slope, and ϵ represents random variation of the data around the line. **There is always some random variation around the line, especially with biological data.** If there were no random variation, a statistical approach would not be needed – one could simply draw a line through the data.

Fig. 1.1 shows an example of this model, fitted to data on the number of reptile species on islands of varying size in the West Indies (Wright 1981). We will examine how the parameters of such models (α and β) can be estimated using likelihood theory, and how to test whether there is indeed a relationship between Y and X (as it appears in Fig. 1.1). It is also possible to make predictions from statistical models. For example, we could use this model to potentially predict the number of reptile species expected on other islands, ones not included in this data set.

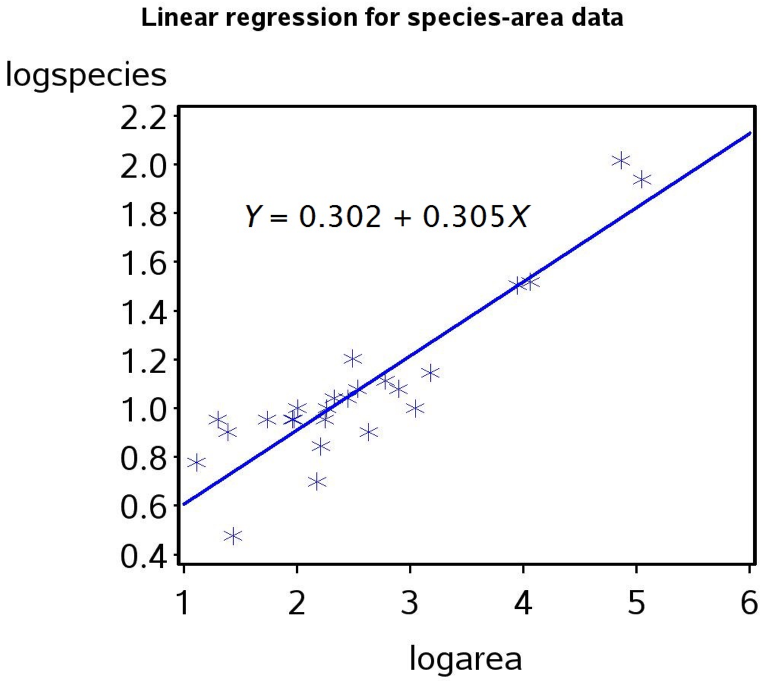


Figure 1.1: Number of reptile species vs. island area in the West Indies (Wright 1981). The number of species and island area were log-transformed before analysis. The fitted line and model equation are also shown.

1.2 Types of data

The first step faced by the statistical analyst is determining the form of the data. There are four types of data frequently encountered by scientists and statisticians: continuous, discrete, rank, and categorical data. **Continuous data** are quantities like the length and weight of an organism, concentrations of chemicals in the environment, or the growth rate of a population. The distinguishing feature of continuous data is that the observations can be described using real numbers. For example, the length of an organism might be 4.53 cm, while its weight 1.23 g. In contrast, **discrete data** always take integer values. They can be counts of organisms in a location, the number of vertebra in the spine, or quantities like the number of disease cases in a month. Typically, discrete data are non-negative integers, i.e., $0, 1, 2, 3, 4$ and so forth. The number of species in Fig. 1.1 could be treated as either continuous or discrete - although they are integer values, they are large enough to take many potential values and approximated as continuous data.

Rank or ordinal data are observations that indicate the relative ordering of the data. For example, suppose an entomologist wants to rapidly assess the level of damage caused by caterpillars to their host plants. It may be easy to quickly assess whether the plants have no damage (a rank of 1), or light (2), medium (3), and heavy damage (4), but finer gradations would be difficult. Rank data also play an important role in a set of procedures called nonparametric statistics, because these procedures often convert continuous or discrete data to rank data. **Categorical data** are observations that fall into separate categories. For example, we might classify specimens of an animal as male, female, or juvenile. No numbers are associated with these categories, although we would likely be interested in how many animals occur in each category, i.e., their frequencies.

1.3 Data and models

Once the data are classified into one of the above types, this determines to a large extent the statistical analysis. For example, suppose the data are (Y, X) pairs as in Fig. 1.1. A linear regression model like Eq. 1.1 would seem appropriate, because the data lie near a straight line. How could we model the random variation around the line? One common choice is to assume that ϵ has a normal or bell-shaped distribution, which we later examine in detail. Once a statistical model is chosen, this largely determines the analysis including how model parameters (α, β, and parameters for ϵ) are estimated and statistical tests conducted, often using likelihood theory. Another important task in statistics is model building, in which a number of different models are fitted to the data and the best-fitting one selected (there are various criteria for determining which is best). Fig. 1.2 shows this general process.

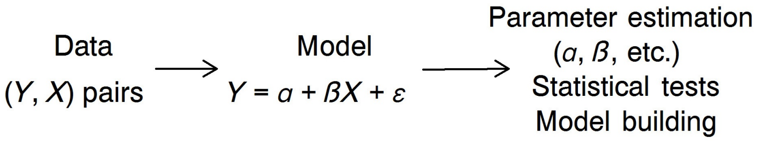


Figure 1.2: Sequence of analysis for many statistical problems.

1.4 Sequence of topics

The next chapter in this text is a brief review of the mathematics useful in statistics, and an introduction to SAS programming (Chapter 2). We then introduce descriptive statistics, which are quantities like the mean or average, designed to summarize the properties of a data set (Chapter 3). The next topic is probability theory, which provides an explanation for many natural processes that apparently have random components, and provides a foundation for statistics (Chapter 4). We then turn to probability distributions for both discrete and continuous data, which are essentially models for random processes (Chapter 5 and 6), and how means and other quantities are defined for these distribution (Chapter 7). We then examine how parameters for these distributions are estimated using likelihood, along with a measure of the reliability of these estimates (Chapter 8, 9), and how hypotheses concerning the parameters are tested (Chapter 10).

Several chapters are devoted to analysis of variance, or ANOVA, used to compare the means of different groups (Chapter 11-15). These groups are often generated by different experimental treatments, and ANOVA and related techniques provide a way of examining whether the treatments produces differences among these groups. Nonparametric alternatives to ANOVA are also considered (Chapter 16). We then examine linear regression and correlation, which are alternate methods of examining the relationship between two variables (Chapter 17, Chapter 18). These methods are designed for continuous variables, but can be adapted to discrete ones. Chapter 19 presents more complicated designs including three-way and nested ANOVA, and analysis of covariance (ANCOVA). In Chapter 20, we examine several techniques useful for analyzing categorical data. Chapter 21 provides an introduction to multiple regression, which examines how one continuous variable is affected by several other variables. Several large data sets used as examples are listed in Chapter 22, while Chapter 23 contains statistical tables used throughout the text.

1.5 References

R Core Team (2021). *R: A Language and Environment for Statistical Computing.* The R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.

Schork, M. A. & Remington, R. D. (2000) *Statistics with Applications to the Biological and Health Sciences, Third Edition.* Prentice Hall, Upper Saddle River, New Jersey, NJ.

Sokal, R. R. & Rohlf, F. J. (1995) *Biometry, Third Edition.* W. H. Freeman and Company, New York, NY.

Steel, R. G. D., Torrie, J. H. & Dickey, D. A. (1997) *Principles and Procedures of Statistics: A Biometrical Approach, Third Edition.* McGraw-Hill, Boston, MA.

Wright, S. J. (1981) Intra-archipelago vertebrate distributions: the slope of the species-area relation. *American Naturalist* 118: 726-748.

Chapter 2

Review of Mathematics

In this chapter, we will briefly review some of the mathematical concepts used in this textbook. Knowing these concepts will make it much easier to understand the mathematical underpinnings of statistics, especially the formulas used in statistics as well as their derivations. A particularly important concept is that of a function. We will commonly encounter several types of functions in statistics, including probability densities or distributions, likelihood functions, the functions used in statistical models, and ones used to transform the observations before statistical analysis.

2.1 Exponents

This section provides a brief summary of useful rules concerning exponents that often appear in statistical functions. Let a and b be two real numbers (numbers of any kind between $-\infty$ and ∞) that form the base of the exponent. This includes the special numbers $e \approx 2.71828$ and $\pi \approx 3.14159$ that often occur in statistics. As exponents or powers, let m and n be any positive integers $(1, 2, 3, ...)$. We then have

$$a^m a^n = a^{m+n} \tag{2.1}$$
$$(a^m)^n = a^{mn} \tag{2.2}$$
$$\frac{a^m}{a^n} = a^{m-n} \tag{2.3}$$
$$(a \times b)^n = a^n b^n \tag{2.4}$$
$$\left(\frac{a}{b}\right)^n = \frac{a^n}{b^n} \tag{2.5}$$

(Schmidt & Ayres 2003). For example, suppose that $a = 2$, $b = 3$, $m = 5$, and $n = 4$. We have

$$a^m a^n = a^{m+n} \tag{2.6}$$
$$2^5 2^4 = 2^{5+4} = 2^9 = 512 \tag{2.7}$$

$$(a^m)^n = a^{mn} \tag{2.8}$$
$$(2^5)^4 = 2^{5\times 4} = 2^{20} = 1048576 \tag{2.9}$$

$$\frac{a^m}{a^n} = a^{m-n} \tag{2.10}$$
$$\frac{2^5}{2^4} = 2^{5-4} = 2^1 = 2 \tag{2.11}$$

$$(a \times b)^n = a^n b^n \tag{2.12}$$
$$(2 \times 3)^5 = 2^5 3^5 = 32 \times 243 = 7776 \tag{2.13}$$

$$\left(\frac{a}{b}\right)^n = \frac{a^n}{b^n} \tag{2.14}$$
$$\left(\frac{2}{3}\right)^5 = \frac{2^5}{3^5} = \frac{32}{243} = 0.132 \tag{2.15}$$

These rules also hold for m and n any real number provided a and b are positive. Some special cases of the above rules are also commonly encountered in statistics. We have

$$a^0 = 1, a \neq 0 \tag{2.16}$$

$$0^0 = 0 \tag{2.17}$$

$$a^{1/2} = \sqrt{a} \tag{2.18}$$

$$a^{-m} = \frac{1}{a^m}, a \neq 0 \tag{2.19}$$

Now suppose that $a = 4$ and $m = 2$. We have

$$4^0 = 1 \tag{2.20}$$

$$4^{1/2} = \sqrt{4} = 2 \tag{2.21}$$

$$4^{-2} = \frac{1}{4^2} = \frac{1}{16} = 0.0625 \tag{2.22}$$

2.2 Inequalities

Statistical statements often involve the use of inequalities. For example, suppose that you are interested in the size distribution of a fish population. You might be interested in estimating the probability or proportion of fish that equal or exceed the legal catch size, say 12 inches. If y stands for fish size, then you would be interested in estimating the probability of fish for which $y \geq 12$ inches. You might also be interested in fish which lie within a certain range of size, say 6 to 12 inches. This could be written as $6 < y < 12$ inches using inequalities. The results of statistical tests are often reported using inequalities as well. You will commonly encounter statements of the form '$P < 0.05$' in scientific papers, which says that the probability P of a certain event occurring is less than 5%, or 1 chance in 20.

Inequalities can be manipulated much like equalities in algebra, with some exceptions. Let x and y stand for any two numbers, or more complex mathematical quantities. If $x < y$, then

$$x + b < y + b \tag{2.23}$$

where b is another number or quantity, and

$$ax < ay \tag{2.24}$$

where a is a **positive** number or other quantity. If a is **negative**, then

$$ax > ay. \tag{2.25}$$

Thus, multiplying an inequality by a negative number flips the direction of the inequality. For example, let $x = 5$, $y = 6$, and $a = -2$. We have $x < y$, but clearly $-2(5) = -10$ is greater than $-2(6) = -12$.

Another exception involves the inverse or reciprocal of an inequality. If $x < y$ and both are positive (or both negative), then

$$\frac{1}{x} > \frac{1}{y}. \tag{2.26}$$

Note the changed direction of the inequality. For example, if $x = 5$ and $y = 6$ so that $x < y$, the inequality is reversed because we have $1/5 > 1/6$. However, if $x < y$ and x is negative, then

$$\frac{1}{x} < \frac{1}{y}. \tag{2.27}$$

For example, if $x = -5$ and $y = 6$ then we have $1/-5 < 1/6$, or $-1/5 < 1/6$. These results can also be obtained through direct application of Eq. 2.24 and 2.25.

2.3 Functions

A variable is a symbol such as x or y chosen to represent a set of numbers, typically real numbers. A function is a relationship between x and y such that each value of x generates a single value of y (Schmidt & Ayres 2003). When such a relationship holds, it is customary to say that y is a function of x. An example of a function is the equation

$$y = 2x + 1 \tag{2.28}$$

This happens to be the equation of a line with a slope of 2 and an intercept of 1. In general, we can write a function using the notation

$$y = f(x) \tag{2.29}$$

where $f(x)$ stands for any possible function of x. In this context, x is often called the independent variable and y the dependent variable.

2.3.1 Functions in Statistics

One commonly used function in statistics is the equation for a line, namely

$$y = ax + b \tag{2.30}$$

where a is the slope and b is the intercept of the line. This function plays an important role in linear regression, a statistical procedure that fits a line to a series of points of the form (x, y) (see Chapter 17). Also common are quadratic functions of the form

$$y = ax^2 + bx + c \tag{2.31}$$

where a, b, and c are constants. Rather than a straight line, quadratic functions are shaped like a parabola.

Exponential and log functions are also commonly used in statistics. Examples of exponential functions are

$$y = 10^x \tag{2.32}$$

and

$$y = e^x, \tag{2.33}$$

where $e = 2.71828\ldots$, also written as

$$y = \exp(x). \tag{2.34}$$

Examples of log functions are the natural log and base 10 log, written as

$$y = \ln(x) \tag{2.35}$$

and

$$y = \log(x). \tag{2.36}$$

Confusingly, the natural log is sometimes written as $\log(x)$, while base 10 log is written as $\log_{10}(x)$. SAS uses this notation for log functions. The log functions are only defined for $x > 0$.

The exponential and log functions are inverses, meaning they reverse the action of each other. For example, we have

$$\exp(\ln(x)) = x \tag{2.37}$$

and

$$\ln(\exp(x)) = x. \tag{2.38}$$

For example, if you find $\ln(x)$ for some value of x, then apply the exp function to $\ln(x)$, you get the original value of x as the answer. Suppose that $x = 2$. We have $\ln(x) = \ln(2) = 0.693$, and then $\exp(\ln(2)) = \exp(0.693) = 2$. The same thing happens for the functions 10^x and $\log(x)$.

Another common function in statistics is the absolute value function, written as

$$y = |x|. \tag{2.39}$$

It is defined as follows. If x is positive or zero then $|x|$ is simply equal to x, while if x is negative then $|x| = -x$. For example, if $x = -2$ then $y = |-2| = -(-2) = 2$. A common use of the absolute value in statistics is to define a symmetric interval around zero. For example, the inequality $-3 < x < 3$ can also be written as $|x| < 3$.

The most commonly used distribution in statistics is the normal distribution, which can be written as a combination of several simpler functions:

$$y = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2.40}$$

Here μ and σ^2 are two parameters that govern the shape of the normal distribution, in particular its mean and variance (see Chapter 6).

2.3.2 Plotting functions using SAS - SAS demo

It can be difficult to discern the shape of a function without a graph. For example, the function describing the normal distribution gives you the famous bell-shaped curve, but this is not obvious from the equation. We will develop a SAS program that will plot any function, given its mathematical form, the values of any constants, and the range of x values for which a plot is needed. We will examine this plotting program in some detail, because it illustrates the structure of the programs used throughout this textbook.

SAS programs consist of a series of steps or instructions that enable you to input and manipulate data and then generate statistical results and graphs. Data are entered and manipulated using SAS `data` steps, while statistical results and graphs are generated using SAS procedures or `proc` steps. Note that SAS is not case-sensitive, so programs can be in either upper or lower case.

The first line of the program is a comment, used here to give the file name of the program. Any line of a SAS program beginning with an asterisk (*) is a comment, which are used to describe the program and its actions but are not executed by SAS.

```
* fplot.sas;
```

The next three lines consist of the instructions

```
title "Plot a function y = f(x)";
title2 "Linear function";
```

The two `title` lines add a main title and subtitle to the output. Note that each of the lines ends with a semicolon (;). This is absolutely critical in SAS programming, because it tells SAS where a particular statement or command ends. A misplaced or absent semicolon will typically cause errors when running the program.

The next part of the SAS code is a `data` step (SAS Institute Inc. 2016a). The idea here is to generate a data set with a sequence of x and $y = f(x)$ values that will later be plotted. The minimum and maximum values of x are set by specifying values for `xmin` and `xmax`, while the number of divisions is set by `xdiv` (the more divisions the finer the x scale and the smoother the graph). The program then calculates the step length between x values (`xlength`) using these quantities. The values of x and $y = f(x)$ are calculated in a programming loop using a `do` statement . Each pass through the loop calculates a new value of x, then finds $y = f(x)$ for that value of x. The results are then sent to a SAS data file using an `output` statement. You can set the name of the data file in the first line of the `data` step, which in this case is `fplot`. Note that six different functions are listed in this `data` step, but only one would be active (the line function) because the remainder are comments. This is a useful programming trick to deactivate sections of code.

```
data fplot;
    * Minimum and maximum values of x;
    xmin = -5;
    * Use for ln function, must have x > 0;
    *xmin = 0.001;
    xmax = 5;
    * Divisions between xmin and xmax (more = smoother graph);
    xdiv = 100;
    * Calculate step length;
    xlength = (xmax-xmin)/xdiv;
    * Find x and y = f(x) values for the plot;
    do i=0 to xdiv;
        x = xmin + i*xlength;
        * Insert f(x) formula here;
        * line function;
        y = 2*x + 1;
        * quadratic function;
        *y = -x**2 + 2*x + 5;
        * exponential function;
        *y = exp(x);
        * ln function;
        *y = log(x);
        * absolute value function;
        *y = abs(x);
        * normal distribution;
        *mu = 1;
        *sig2 = 1;
        *y = (1/sqrt(2*3.14159*sig2))*exp(-((x-mu)**2)/(2*sig2));
        * Output x and y to SAS data file;
        output;
    end;
run;
```

The resulting data are then printed using the SAS `print` procedure (SAS Institute Inc. 2016b), using the syntax below. The option `data=fplot` tells the `print` procedure to use this particular data file. If this option were omitted, the last data file created would automatically be used. The `run` statement tells SAS that the `proc print` command is complete and that it should get busy printing the data file.

```
* Print data;
proc print data=fplot;
run;
```

The gplot procedure is used to plot the function using the new data set (see below) (SAS Institute Inc. 2016c). The plot statement tells SAS which of your SAS variables are the x and y variables - the variable before the asterisk (*) is the y variable, after it the x variable (it is the position that is important, not the name of the variable). The href = 0 and vref = 0 options make SAS draw vertical and horizontal lines through the origin $(0, 0)$. The symbol1 statement tells SAS to join the points with a line (i=join), draw no symbol for each data point (v=none), and make the line connecting the points red (c=red). The remainder of the options listed in the program are intended to make the graph more legible by increasing the thickness of the lines and size of the axes labels. If you are curious how they work, try experimenting with the numbers given in the options. The quit statement returns control to SAS after running the program.

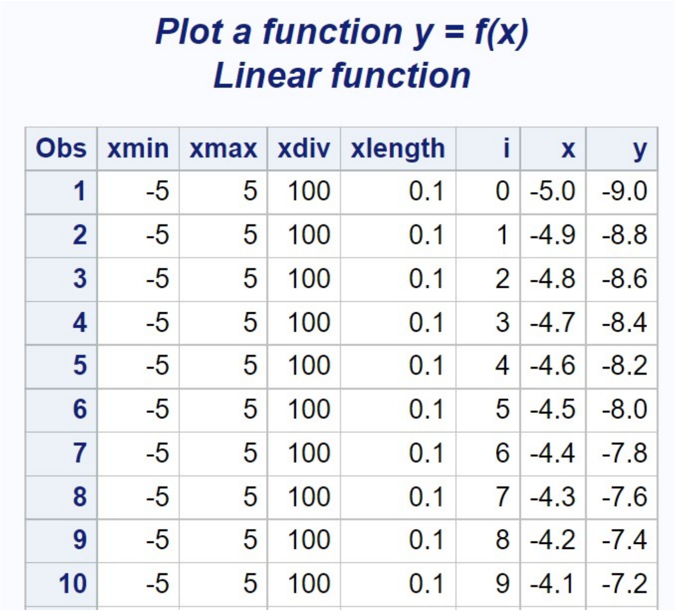
```
* Plot y = f(x);
proc gplot data=fplot;
    plot y*x=1 / href=0 vref=0 whref=3 wvref=3 vaxis=axis1 haxis=axis1;
    symbol1 i=join v=none c=red width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
quit;
```

See the full program listing below, a portion of the printed output, and graphs for the various functions included in the program. Both the program and complete output can be found on the website for this textbook.

SAS program

```
* fplot.sas;
title "Plot a function y = f(x)";
title2 "Linear function";
data fplot;
    * Minimum and maximum values of x;
    xmin = -5;
    * Use for ln function, must have x > 0;
    *xmin = 0.001;
    xmax = 5;
    * Divisions between xmin and xmax (more = smoother graph);
    xdiv = 100;
    * Calculate step length;
    xlength = (xmax-xmin)/xdiv;
    * Find x and y = f(x) values for the plot;
    do i=0 to xdiv;
        x = xmin + i*xlength;
        * Insert f(x) formula here;
        * line function;
        y = 2*x + 1;
        * quadratic function;
        *y = -x**2 + 2*x + 5;
        * exponential function;
        *y = exp(x);
        * ln function;
        *y = log(x);
        * absolute value function;
        *y = abs(x);
        * normal distribution;
        *mu = 1;
        *sig2 = 1;
        *y = (1/sqrt(2*3.14159*sig2))*exp(-((x-mu)**2)/(2*sig2));
        * Output x and y to SAS data file;
        output;
    end;
run;
* Print data;
proc print data=fplot;
run;
* Plot y = f(x);
proc gplot data=fplot;
    plot y*x=1 / href=0 vref=0 whref=3 wvref=3 vaxis=axis1 haxis=axis1;
    symbol1 i=join v=none c=red width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
```

```
run;
quit;
```



Plot a function y = f(x)
Linear function

Obs	xmin	xmax	xdiv	xlength	i	x	y
1	-5	5	100	0.1	0	-5.0	-9.0
2	-5	5	100	0.1	1	-4.9	-8.8
3	-5	5	100	0.1	2	-4.8	-8.6
4	-5	5	100	0.1	3	-4.7	-8.4
5	-5	5	100	0.1	4	-4.6	-8.2
6	-5	5	100	0.1	5	-4.5	-8.0
7	-5	5	100	0.1	6	-4.4	-7.8
8	-5	5	100	0.1	7	-4.3	-7.6
9	-5	5	100	0.1	8	-4.2	-7.4
10	-5	5	100	0.1	9	-4.1	-7.2

etc.

Figure 2.1: `fplot.sas - proc print`

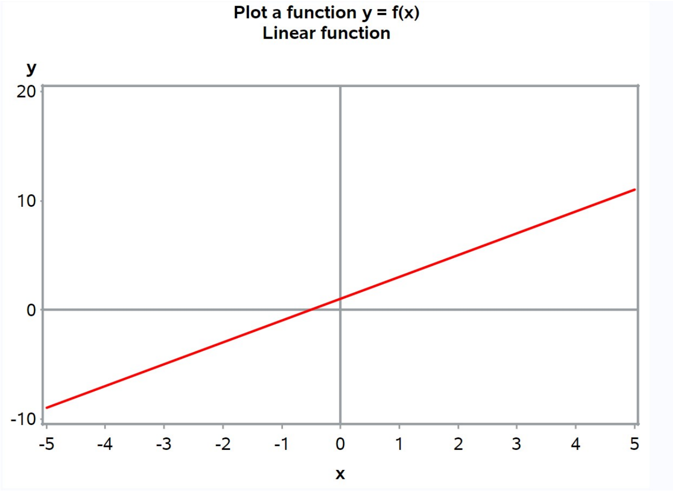


Figure 2.2: `fplot.sas - proc gplot`

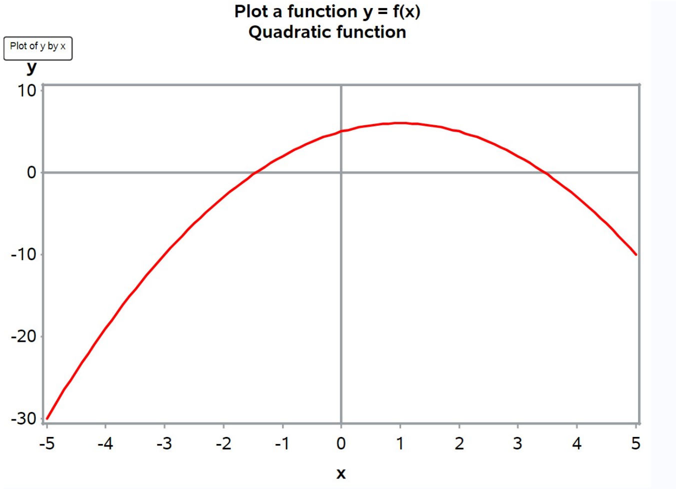


Figure 2.3: `fplot.sas - proc gplot`

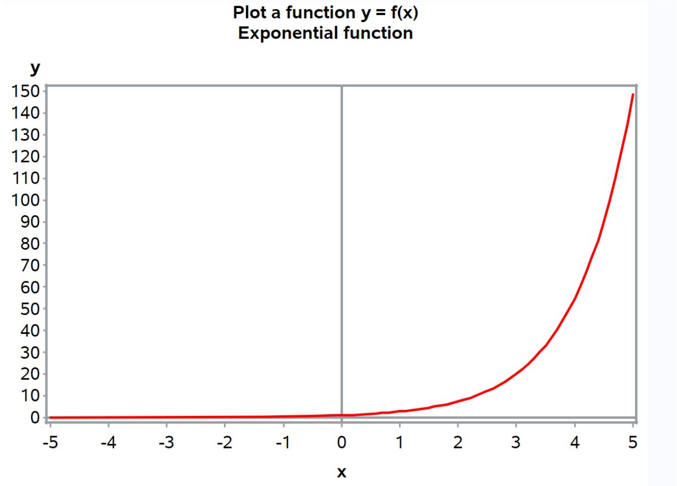


Figure 2.4: `fplot.sas - proc gplot`

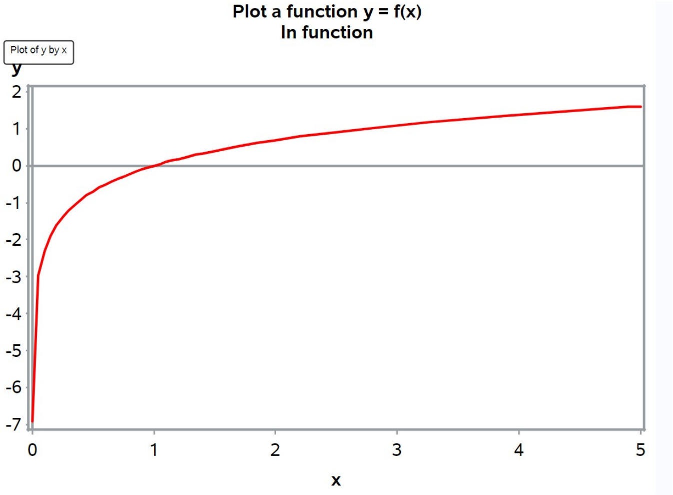


Figure 2.5: `fplot.sas - proc gplot`

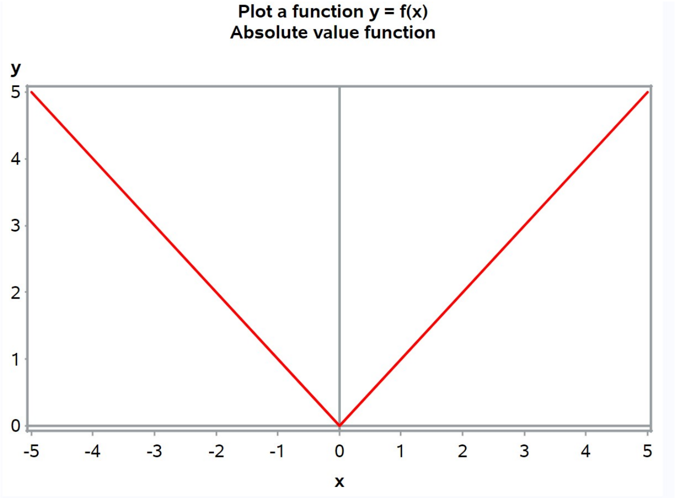


Figure 2.6: `fplot.sas - proc gplot`

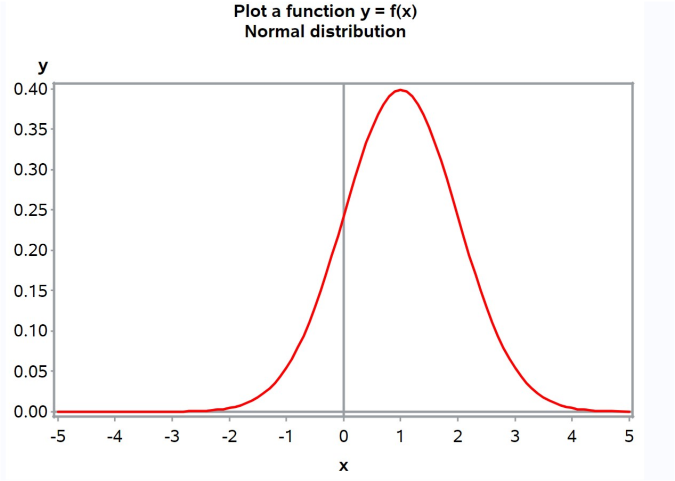


Figure 2.7: `fplot.sas - proc gplot`

2.4 Solving linear equations

We next review how to solve a linear equation for x, a procedure that will be useful in later developments. A linear equation has the general form

$$ax + b = cx + d \tag{2.41}$$

where a, b, c, and d are constants that are possibly zero, while x is a variable. We want to find a value of x that makes this equation true, meaning the two sides of the equation are equal. To solve this problem, you perform the same operations on both sides of the equation until you have x alone on one side of the equation. The other side is then the answer to this problem (Schmidt & Ayres 2003). More generally, a-d and x could also be more complicated expressions that one manipulates to obtain an expression for x.

To illustrate this procedure, suppose we have the equation

$$5x - 4 = 3x - 3. \tag{2.42}$$

Subtracting $3x$ from both sides of the equation, we get

$$2x - 4 = -3. \tag{2.43}$$

We next add 4 to both sides to obtain

$$2x = 1. \tag{2.44}$$

Dividing both sides by 2 we obtain the solution

$$x = 1/2. \tag{2.45}$$

If you want to check if the solution is correct, you can always substitute it back into the original equation. We have

$$5(1/2) - 4 = 3(1/2) - 3 \tag{2.46}$$
$$2.5 - 4 = 1.5 - 3 \tag{2.47}$$
$$-1.5 = -1.5. \tag{2.48}$$

So $x = 1/2$ is in fact the correct solution.

2.5 Roots of equations

For a particular function $y = f(x)$, it is often useful to find the values of x for which $y = f(x) = 0$. Values of x for which this is true are called the roots of the equation $f(x) = 0$ (Schmidt & Ayres 2003). Graphically, the roots are the values of x where the function crosses the x-axis, i.e., the function is equal to zero. It is possible to find the roots for many functions algebraically, but not every function has roots, and for some functions they can only be found numerically using software and a computer.

Roots are easy to find for linear functions. Recall that a linear function takes the general form

$$y = a + bx \tag{2.49}$$

where a and b are constants. We want to find values of x for which

$$a + bx = 0 \tag{2.50}$$

We then use the rules for solving linear equations to find x. Subtracting a from both sides and dividing by b, we obtain

$$x = \frac{-a}{b} \tag{2.51}$$

Suppose that $a = 1$ and $b = 2$, so that our function is

$$y = 1 + 2x. \tag{2.52}$$

It follows that the root of this function is $x = -a/b = -1/2$. If we examine the graph generated earlier for this function, we see that the function indeed crosses the x-axis at $x = -1/2$.

We can also find the roots for quadratic functions using, logically enough, the quadratic formula. Recall that a quadratic function takes the general form

$$y = ax^2 + bx + c \tag{2.53}$$

We want to find values of x for which

$$ax^2 + bx + c = 0. \tag{2.54}$$

The quadratic formula says that the roots are given by the equation

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a} \tag{2.55}$$

We previously plotted a quadratic function of the form

$$y = -x^2 + 2x + 5 \tag{2.56}$$

To find the roots, we need to solve the equation

$$-x^2 + 2x + 5 = 0 \tag{2.57}$$

Inspecting this equation, we see that $a = -1$, $b = 2$, and $c = 5$. Inserting these values in the quadratic formula, we obtain

$$x = \frac{-2 \pm \sqrt{2^2 - 4(-1)5}}{2(-1)} = \frac{-2 \pm \sqrt{24}}{-2} \tag{2.58}$$

$$= \frac{-2 \pm 4.90}{-2} = \frac{-6.90}{-2}, \frac{2.90}{-2} = 3.45, -1.45 \tag{2.59}$$

The roots of this quadratic equation are therefore equal to 3.45, 1.45. This result agrees with the graph drawn earlier.

2.6 Calculus

We will make only limited use of calculus in this course, but it is useful to review the concepts of derivatives and integrals. Derivatives are often used in estimating the parameters of statistical models through a method called maximum likelihood (Chapter 8). Integrals are used to generate the probabilities associated with confidence intervals, statistical tests, and other procedures. For example, the statistical tables given in Chapter 23 were all generated using integrals.

2.6.1 Derivatives

A derivative of a function $y = f(x)$ is defined to be the slope of the function at a particular value of x. Recall that the slope is defined as the change in y divided by the change in x. The mathematical definition of a derivative is given by the equation

$$\lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(x + \Delta x) - f(x)}{\Delta x} \tag{2.60}$$

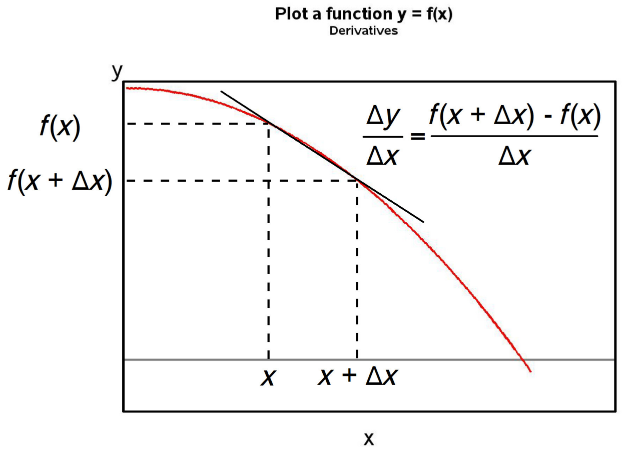


Figure 2.8: Definition of a derivative

where Δx is the change in x, while Δy is the change in y, defined as $f(x + \Delta x) - f(x)$ (Schmidt & Ayres 2003). This equation says that the derivative is given by the limit, as Δx goes to zero, of the slope $\Delta y/\Delta x$. See also Fig. 2.8. The derivative of a function may be written as $\frac{dy}{dx}$ or $f'(x)$.

Now suppose we have a linear function like

$$y = ax + b. \tag{2.61}$$

The derivative of this function is simply a, the slope of the line. It is equal to a regardless of the value of x, because a line has the same slope everywhere. We would write this as $\frac{dy}{dx} = a$ or $f'(x) = a$.

Assume now that we have a quadratic function. There is a formula for the derivative of a power of x that is often useful. If $y = f(x) = kx^n$, where k and n are any constants, then

$$\frac{dy}{dx} = knx^{n-1}. \tag{2.62}$$

We can use this formula to find the derivative of a quadratic function of the form

$$y = ax^2 + bx + c. \tag{2.63}$$

We have

$$\frac{dy}{dx} = a(2)x^{2-1} + b(1)x^{1-1} + 0 = 2ax + b. \tag{2.64}$$

To obtain this result, we also made use of the fact that the derivative of a constant (c in this case) is always zero (because it is unchanging), and that the derivative of a sum of functions is the sum of the derivatives.

One important application of the derivative in statistics is to find the maximum or minimum of a function. In particular, the derivative of a function is equal to zero at the maximum or minimum. This follows because a function that has a maximum must eventually stop rising and begin to fall, and at that point the slope is equal to zero. The same reasoning applies to a minimum.

To find the maximum or minimum for our general quadratic function, we set $dy/dx = 0$ and solve for x. We have

$$\frac{dy}{dx} = 2ax + b = 0. \tag{2.65}$$

Solving this linear equation for x, we find that the maximum or minimum will occur at $x = \frac{-b}{2a}$.

2.6.2 Function plot with derivative - SAS demo

We will plot a quadratic function and its derivative to observe the relationship between the two. Suppose that we have the following quadratic function:

$$y = -x^2 + 2x + 5. \tag{2.66}$$

The derivative of this function is

$$\frac{dy}{dx} = -2x^{2-1} + 2(1)x^{1-1} + 0 = -2x + 2. \tag{2.67}$$

We can find the minimum or maximum of this function by setting the derivative equal to zero and solving for x. We have

$$-2x + 2 = 0 \tag{2.68}$$

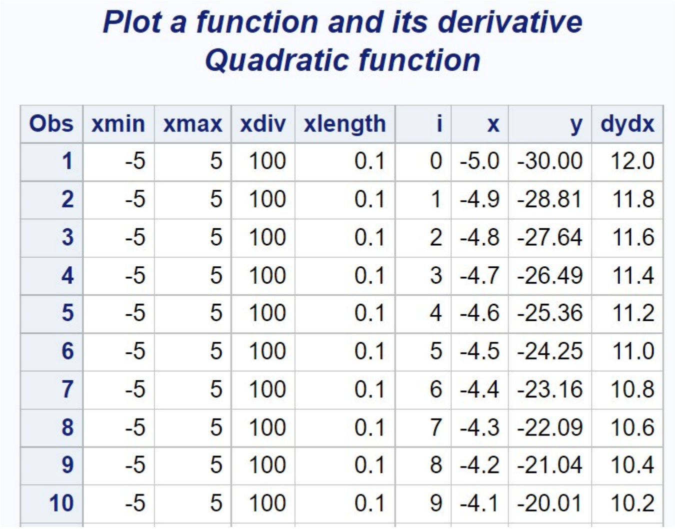
for which the solution is $x = 1$.

We will now plot both y and dy/dx using a revised version of our plotting program. This program calculates both y and dy/dx within the `do` loop, then plots both sets of points on the same graph using the overlay option in `proc gplot`. See SAS program and output below.

Note that the derivative of this quadratic function is a straight line with a slope of -2 and an intercept of 2. It equals zero at the point where it intercepts the x-axis, which also corresponds to the maximum of the quadratic function. Our calculation above shows this occurs at $x = 1$.

SAS program

```
* fplot_deriv.sas;
title "Plot a function and its derivative";
title2 "Quadratic function";
data fplot2;
    * Minimum and maximum values of x;
    xmin = -5;
    xmax = 5;
    * Divisions between xmin and xmax (more = smoother graph);
    xdiv = 100;
    * Calculate step length;
    xlength = (xmax-xmin)/xdiv;
    * Find x, y = f(x), and dy/dx values for the plot;
    do i=0 to xdiv;
        x = xmin + i*xlength;
        * quadratic function;
        y = -x**2 + 2*x + 5;
        * derivative of this function;
        dydx = -2*x + 2;
        * Output x, y, and dydx to SAS data file;
        output;
    end;
run;
* Print data;
proc print data=fplot2;
run;
* Plot y = f(x) and dydx;
proc gplot data=fplot2;
    plot y*x=1 dydx*x=2 / href=0 vref=0 overlay whref=3 wvref=3 vaxis=axis1
    haxis=axis1;
    symbol1 i=join v=none c=red width=3;
    symbol2 i=join v=none c=blue width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
quit;
```



Plot a function and its derivative
Quadratic function

Obs	xmin	xmax	xdiv	xlength	i	x	y	dydx
1	-5	5	100	0.1	0	-5.0	-30.00	12.0
2	-5	5	100	0.1	1	-4.9	-28.81	11.8
3	-5	5	100	0.1	2	-4.8	-27.64	11.6
4	-5	5	100	0.1	3	-4.7	-26.49	11.4
5	-5	5	100	0.1	4	-4.6	-25.36	11.2
6	-5	5	100	0.1	5	-4.5	-24.25	11.0
7	-5	5	100	0.1	6	-4.4	-23.16	10.8
8	-5	5	100	0.1	7	-4.3	-22.09	10.6
9	-5	5	100	0.1	8	-4.2	-21.04	10.4
10	-5	5	100	0.1	9	-4.1	-20.01	10.2

etc.

Figure 2.9: `fplot_deriv.sas - proc print`

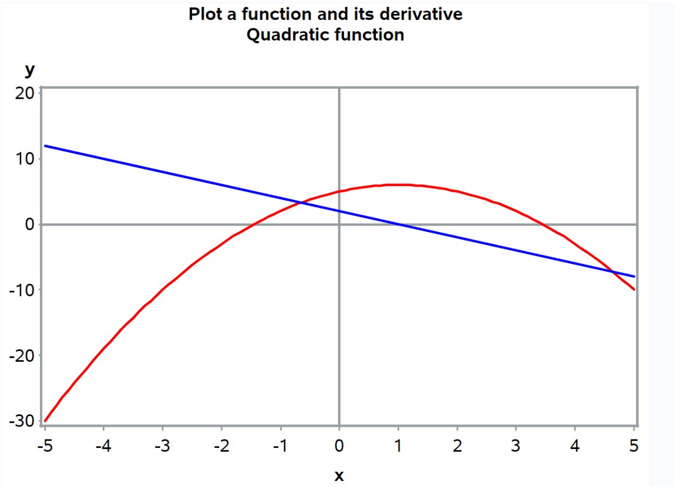


Figure 2.10: `fplot_deriv.sas - proc gplot`

2.6.3 Integrals

Statistics makes heavy use of integrals in working with the normal and other statistical distributions, although statistical tables or software typically do the work for the end user. For example, tables of the normal distribution provide probabilities for certain intervals - these probabilities are actually areas under the bell-shaped curve and are calculated by integration.

One kind of integral often encountered in statistics is a called a definite integral. It is basically the area A under a function $f(x)$ over some range of x values, say $a < x < b$. It is written mathematically as the equation

$$A = \int_a^b f(x)dx. \tag{2.69}$$

Here the symbol $\int$ is the integral sign, with the range of x values $(a < x < b)$ shown as sub- and superscripts of the integral sign.

To make things more concrete, we will illustrate definite integrals using the normal distribution function. Consider this function for $\mu = 5$ and $\sigma^2 = 1$, and the area A under it from $x = 5$ to $x = 6$ (Fig. 2.11). If we were modeling the behavior of some biological variable (say body mass of a small animal) using this distribution, the area A would be the probability that an animal falls within this range of x values. It would be expressed in mathematical terms as the integral

$$A = \int_a^b \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \int_5^6 \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-5)^2}{2}} dx. \tag{2.70}$$

How is the area A actually calculated through integration? We can approximate this area by dividing it into strips of width $\Delta x = 0.25$ and height $f(x)$ given by the normal distribution function (Fig. 2.12). Adding the areas of these strip, we obtain $A \approx 0.099 + 0.093 + 0.080 + 0.068 = 0.340$. If we increased the number of strips while simultaneously decreasing the width of the strips Δx, we would get an even more accurate approximation to A. The integral is defined as the limit of this process, as the number of strips approaches infinity and their width $\Delta x \to 0$ (Schmidt & Ayres 2003). The exact value of the area obtained through this process is $A = 0.341$.

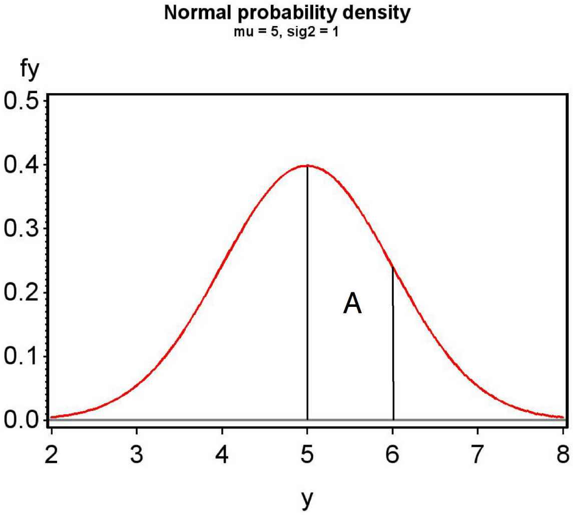


Figure 2.11: Plot of the normal distribution for $\mu = 5$ and $\sigma^2 = 1$.

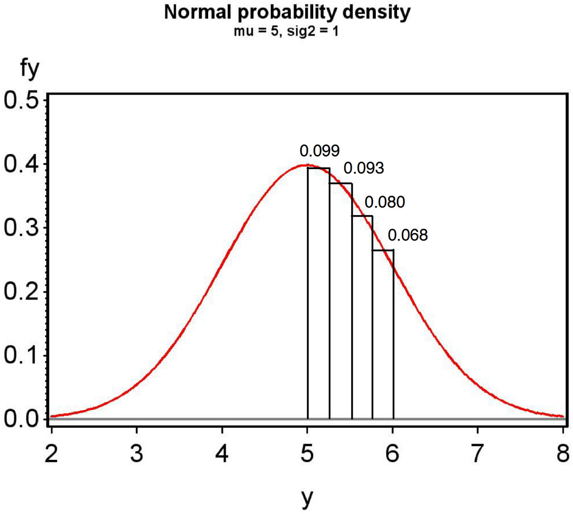


Figure 2.12: Plot of the normal distribution for $\mu = 5$ and $\sigma^2 = 1$.

2.7 References

SAS Institute Inc. (2016a) *SAS 9.4 Language Reference: Concepts, Sixth Edition.* SAS Institute Inc., Cary, NC.

SAS Institute Inc. (2016b) *Base SAS 9.4 Procedures Guide, Sixth Edition.* SAS Institute Inc., Cary, NC.

SAS Institute Inc. (2016c) *SAS/GRAPH 9.4: Reference, Fifth Edition.* SAS Institute Inc., Cary, NC.

Schmidt, P. A. & Ayres, F. Jr. (2003) *Schaums Outline of Theory and Problems of College Mathematics, 3rd Edition.* The McGraw-Hill Companies, Inc., New York, NY.

2.8 Problems

1. Suppose that you have the quadratic function $y = x^2 - 2x - 8$. Find the roots of this function, then determine the value of x that minimizes it. Plot the function using SAS, attaching your program, output, and graph.

2. Consider the quadratic function $y = -2x^2 + 5x + 5$. Find the roots of this function, then determine the value of x that maximizes it. Plot the function and its derivative dy/dx using SAS, attaching your program, output, and graph.

3. Plot the function $y = 0.5\lambda^3 x^2 \exp(-\lambda x)$ for $\lambda = 2$ and $0 \leq x \leq 5$ using SAS. Attach your program, output, and graph. This function is a special case of the gamma distribution, a probability distribution often used to model continuous data (see Chapter 6).

Chapter 3

Populations and Statistics

This chapter covers two topics that are fundamental in statistics. The first is the concept of a statistical population, which is the basic unit on which statistics are conducted and inferences made. We then examine descriptive statistics and frequency distributions, which are used quantify the properties of samples from a statistical population.

3.1 Statistical populations

Suppose we want to estimate the body length of an insect species in a particular location, say a forest stand. We sample the insects in some way (traps, sweep nets, locate them visually, etc.), and average their lengths to obtain an estimate of insect length. We can therefore make some inference about insect lengths in this particular forest stand, which we can call a **statistical population**. A statistical population is defined by both the question of interest (insect length) as well as the sampling method. If we sample insects in only a single forest stand, then the statistical population is length in that stand, not other stands. This is commonly called the **scope of inference** of the study. If we sampled within multiple stands in a forest, then we could potentially examine length for the forest as whole, which would be a different statistical population and the scope of inference would be broader. The sampling technique itself can also affect the statistical population. For example, only a subset of insects might be caught with sweep nets (maybe slower, smaller ones) and this would be a different set than those found visually. The two sampling techniques might therefore define different statistical populations.

Biologists are continually searching for better methods of sampling organisms, ones that better represent their true properties. In many cases the idea is to approximate what is known as **random sample** of the statistical population (see Chapter 8).

In the insect length example above, the statistical population coincides with individual insects in a location. However, the observations comprising a statistical population can be other quantities. For example, suppose we want to estimate the abundance of these insects using traps. We could deploy several traps in the stand, and then average the number of insects caught to estimate their abundance. The statistical population in this case would consist of number of insects caught in traps deployed at that location, rather than individual insects. Or one might be interested in soil nitrogen levels in the stand, estimated using core samples. In this case, the statistical population would be the nitrogen levels in core samples at this location.

Another type of statistical population involves experiments. Suppose we are interested in trapping the same insects in the forest stand, but now have traps baited with different attractants, say A, B, and C. Several traps are baited with each attractant, and the number of insects caught observed for each trap. We are interested in whether the number of insects caught varies with the attractant used. In this case, the statistical population would be trap catches for the different attractants. Similarly, suppose we were interested in the effect of different commercial diets on the growth rate of fish. Different fish would be fed the various diets and their growth rate observed. Here the statistical population would be the growth rate of individual fish for the different diets. Experiments also have a scope of inference. If we use four particular diets to grow fish, our conclusions are restricted to these four diets and not other diets. If the experiment used a particular strain of fish, our inferences would also be restricted to this strain.

3.2 Descriptive statistics and frequency

Given a sample from a statistical population, the first step in understanding its properties is to calculate a number of descriptive statistics. Some statistics give you an idea of the overall magnitude or location of the data, and are traditionally called **statistics of location**. We will examine two such statistics, the sample mean and the median. Other statistics give an indication of the scatter or spread of the data, and are called **statistics of**

dispersion. These include the sample variance, standard deviation, the coefficient of variation, and range of the data. Another important tool is the **frequency distribution** of the sample, often plotted as a histogram indicating the frequency of different values in the sample. Three other statistics, the mode, skewness, and kurtosis, provide information on the shape of this frequency distribution.

To illustrate how the various descriptive statistics are calculated, we will use a small subset of a larger data set on the elytra length for a predatory beetle, *Thanasimus dubius* (Coleoptera: Cleridae). This predator attacks insects known as bark beetles, some species of which are serious pests of coniferous forests (Berryman 1988). Beetles have two pairs of wings. The first pair, the elytra, act as covers for a membraneous second pair that are used in flight. The data are drawn from a rearing study of *T. dubius*, in which elytra length (mm) was used as a overall index of body size (Reeve et al. 2003). The subset data are for eight female *T. dubius* and are listed below:

```
5.2 4.2 5.7 5.4 4.0 4.5 5.2 4.2
```

We will later examine the full data set consisting of 130 individuals using SAS programs.

3.2.1 Sample mean

The sample mean is the average of the values in the sample, and is symbolized as $\bar{Y}$. It is commonly used as a measure of the location or center of the observations. If $Y_1, Y_2, \ldots, Y_n$ represent the observations in a sample from a statistical population, where n is the sample size, the sample mean is calculated using the formula

$$\bar{Y} = \frac{Y_1 + Y_2 + \ldots + Y_n}{n} = \frac{\sum_{i=1}^{n} Y_i}{n}. \tag{3.1}$$

The symbol $\sum_{i=1}^{n}$ stands for summing the observations, beginning with $i = 1$ and ending with $i = n$. The units of $\bar{Y}$ are the same as those for the Y_i values.

For our sample data set involving $n = 8$ elytra from female *T. dubius* beetles, we have

$$\bar{Y} = \frac{5.2 + 4.2 + 5.7 + 5.4 + 4.0 + 4.5 + 5.2 + 4.2}{8} = \frac{38.4}{8} = 4.8 \text{ mm}. \tag{3.2}$$

3.2.2 Median

The median is defined as the middle value of the sample, after ordering the sample from the smallest to the largest value. Suppose that $Y_{[j]}$ is the jth value in the ordered data set, with $Y_{[1]}$ the smallest value and $Y_{[n]}$ the largest. If n is odd, the median is equal to the middle value in the ordered data set, or $Y_{[n/2+1/2]}$. If n is even then the median is the average of the two middle values, or $(Y_{[n/2]} + Y_{[n/2+1]})/2$.

To find the median for the elytra data set, we first order the observations from smallest to largest. We have

j (order):	1	2	3	4	5	6	7	8
$Y_{[j]}$:	4.0	4.2	4.2	4.5	5.2	5.2	5.4	5.7

Because $n = 8$ is even, the median is the average of the middle two observations, or $(Y_{[n/2]} + Y_{[n/2+1]})/2 = (Y_{[8/2]} + Y_{[8/2+1]})/2 = (Y_{[4]} + Y_{[5]})/2 = (4.5 + 5.2)/2 = 4.85$.

Suppose now we had only $n = 7$ observations, with the ordered data set equal to

j (order):	1	2	3	4	5	6	7
$Y_{[j]}$:	4.0	4.2	4.2	4.5	5.2	5.2	5.4

Because $n = 7$ is odd, the median is the middle observation, or $Y_{[n/2+1/2]} = Y_{[7/2+1/2]} = Y_{[4]} = 4.5$ mm.

The median is also a measure of the location of the data, like the sample mean $\bar{Y}$, but is less sensitive to very large or small values in the sample. For example, suppose that the largest observation in the elytra data set was 100.0. The median would be unchanged because the ordering of the observations is unchanged, but now $\bar{Y} = 16.8$ mm, much larger than before.

The median represents a value that essentially divides the data in half, with 50% of the observations lying above or below it. This is an example of a statistic generically called **quantiles** or **percentiles**, with the median a 50% quantile. Other commonly used quantiles are the 25% and 75% quantiles. They and the median are sometime called **quartiles** because they divide the data into four quarters.

3.2.3 Sample variance

The sample variance, written as s^2, is a measure of the dispersion or scatter in the data around the sample mean. It is calculated using the formula

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1} \tag{3.3}$$

The sample variance s^2 will be small if the observations cluster tightly around $\bar{Y}$, because this makes $(Y_i - \bar{Y})^2$ small. Conversely, if the observations are widely scattered these terms will be large, making s^2 large. The units of s^2 are those of Y_i, but squared.

To find s^2 for the elytra data set, we first need to calculate the sample mean. We previously found that $\bar{Y} = 4.8$ mm. We then calculate s^2 using the above formula. We have

$$\begin{aligned} s^2 &= \frac{(5.2-4.8)^2 + (4.2-4.8)^2 + \ldots + (4.2-4.8)^2}{8-1} \qquad (3.4)\\ &= \frac{0.16 + 0.36 + 0.81 + 0.36 + 0.64 + 0.09 + 0.16 + 0.36}{7} \qquad (3.5)\\ &= \frac{2.94}{7} = 0.42 \text{ mm}^2. \qquad (3.6) \end{aligned}$$

3.2.4 Standard deviation

The sample standard deviation, written as s, is simply the square root of s^2. We have

$$s = \sqrt{s^2} \tag{3.7}$$

For the elytra example, we have $s = \sqrt{s^2} = \sqrt{0.42} = 0.645$ mm. The units of s are the same as those of Y_i, which makes it more comparable to statistics of location like $\bar{Y}$.

3.2.5 Coefficient of variation

The coefficient of variation, or CV, provides a measure of the variability of the observations expressed as a percentage of the sample mean. It is calculated using the formula

$$CV = 100\% \times \frac{s}{\bar{Y}}. \tag{3.8}$$

Using the elytra data where $s = 0.645$ mm and $\bar{Y} = 4.8$ mm, we have

$$CV = 100\% \times \frac{0.645}{4.8} = 13.4\% \tag{3.9}$$

The CV allows one to compare the variability of observations on variables that have different means. For example, suppose that we want to compare variability in *T. dubius* elytra length with variability in another predator that has a longer overall length. For biological variables like length, the standard deviation s often seems proportional to the sample mean $\bar{Y}$. If we divide s by $\bar{Y}$, as in the CV, we can control to some extent the influence of $\bar{Y}$ on variability. This allows us to compare variability in length across the two predators on a more even basis.

3.2.6 Range

The range is defined as the difference between the largest and smallest observations, i.e.,

$$\text{range} = Y_{\text{max}} - Y_{\text{min}}, \tag{3.10}$$

where Y_{max} is the largest observation and Y_{min} is the smallest. For the elytra data, we have $Y_{\text{max}} = 5.7$ and $Y_{\text{min}} = 4.0$, so

$$\text{range} = 5.7 - 4.0 = 1.7 \text{ mm}. \tag{3.11}$$

The range is another statistic of dispersion, but has some problems. The range tends to increase in size as the sample size n increases, because larger samples are more likely to yield very small or large observations. This is not the case for s^2 or s.

3.2.7 Frequency distributions - SAS demo

Frequency distributions are another way of summarizing and describing a sample from a statistical population. They typically take the form of a histogram showing the frequency of different observations in the sample. We will use SAS to construct frequency distributions as well as calculate descriptive statistics like $\bar{Y}$, s^2, and so forth. We will use the full elytra data set for *T. dubius* (Reeve et al. 2003) to illustrate these calculations (see Chapter 22). This data set contains both male and female beetles, and we will conduct separate analyses for each sex.

The program first uses a `data` step to read in the observations and make a data file (SAS Institute Inc. 2016a). The line

```
data elytra;
```

tells SAS to set up a data file named `elytra`. If you omit a name from this statement, SAS will automatically generate one for you. The line

```
    input sex $ length;
```

tells SAS to read in two variables and give them the names `sex` and `length`. It also tells SAS to expect the data in the form of two columns. The `$` symbol after `sex` tells SAS that it is a character variable, consisting of a word or letters rather than a number. The default is for a numeric variable. The line

```
    datalines;
```

tells SAS that the following lines in the program are the actual data. The program then lists the data, followed by another semicolon and then a `run` statement (see below). The `etc.` in the data is not SAS code, but shorthand for a longer data set. The `run` statement tells SAS the `data` step is over, and also that it should process the data and generate a SAS data file.

```
M   4.9
F   5.2
M   4.9
F   4.2
F   5.7

etc.

M   5.1
F   4.4
M   4.8
M   4.6
F   3.7
;
run;
```

We are now ready to do something with our newly minted SAS data file, named `elytra`. It is usually a good idea just to print the data file to make sure SAS correctly read the data. This is accomplished using the `proc print` code listed below.

```
* Print data set;
proc print data=elytra;
run;
```

The final lines of the SAS program invoke `proc univariate` to generate the histogram and calculate a number of descriptive statistics (SAS Institute Inc. 2016b). The first and third lines are comments. The second line tells SAS to call `proc univariate` using the `elytra` data set. The `class` statement tells the procedure to conduct a separate analysis for each sex in the data set, while the `var` statements tells it which variable to analyze, in this case the variable `length`. The `histogram` statement asks for a histogram of `length`. The option `vscale=count` tells SAS to make the vertical axis using counts of the observations (the default uses percentages).

```
* Descriptive statistics and histograms;
proc univariate data=elytra;
    * Separate analyses for each sex;
    class sex;
    var length;
    histogram length / vscale=count;
run;
quit;
```

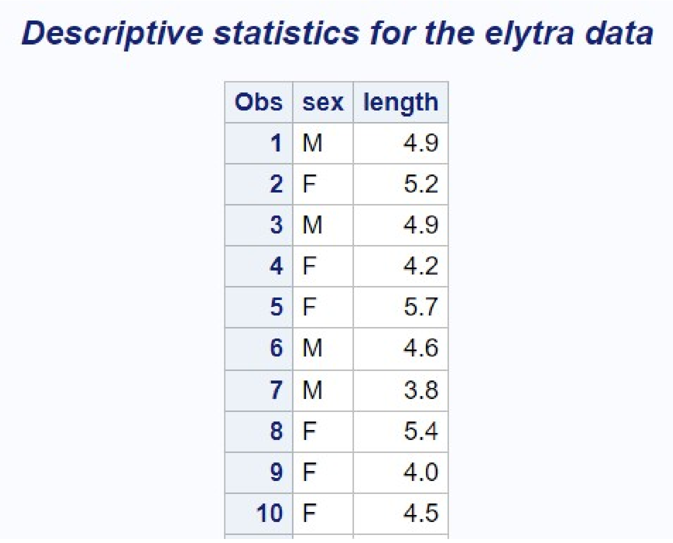
After running the program, we obtain output with various statistics of location and dispersion, including the sample mean, median range, variance, and standard deviation, as well as a graph showing the frequency distribution. A separate analysis is generated for each sex (`M` or `F`) of the beetles. We see that females have somewhat longer elytra than males ($\bar{Y} = 4.940$ mm vs. 4.713 mm), and there are small differences in other statistics. See a complete program listing below and selected portions of the SAS output. The complete output can be found on the website for this textbook.

SAS Program

```
* descriptive.sas;
title ’Descriptive statistics for the elytra data’;
data elytra;
    input sex $ length;
    datalines;
M   4.9
F   5.2
M   4.9
F   4.2
F   5.7

etc.

M   5.1
F   4.4
M   4.8
M   4.6
F   3.7
;
run;
* Print data set;
proc print data=elytra;
run;
* Descriptive statistics and histograms;
proc univariate data=elytra;
    * Separate analyses for each sex;
    class sex;
    var length;
    histogram length / vscale=count;
run;
quit;
```

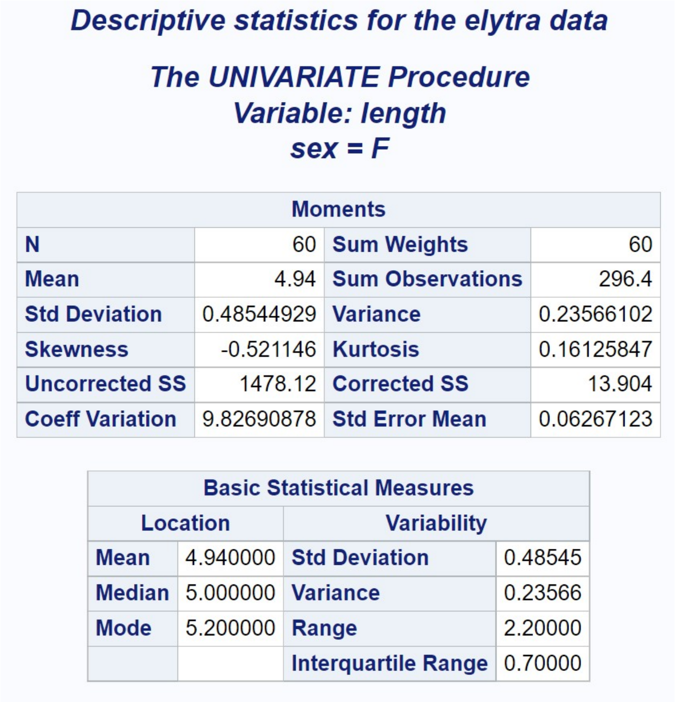


Descriptive statistics for the elytra data

Obs	sex	length
1	M	4.9
2	F	5.2
3	M	4.9
4	F	4.2
5	F	5.7
6	M	4.6
7	M	3.8
8	F	5.4
9	F	4.0
10	F	4.5

etc.

Figure 3.1: `descriptive.sas - proc print`



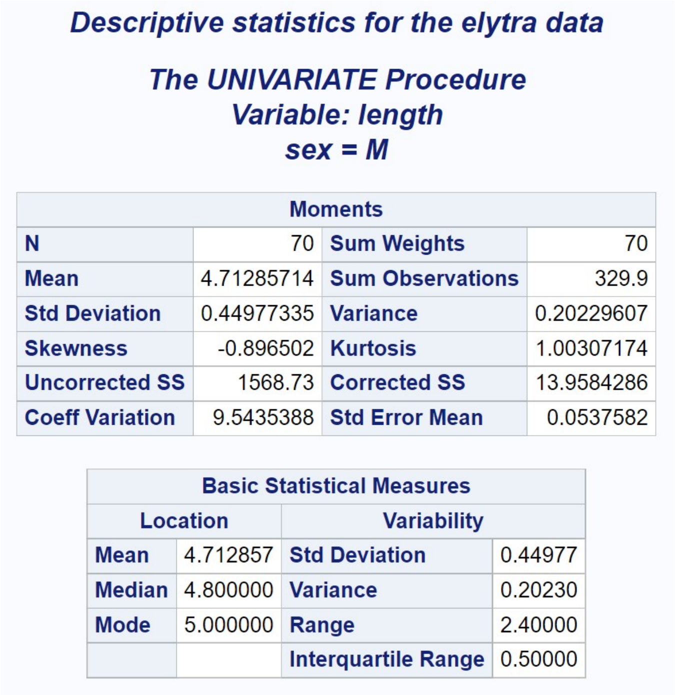
Descriptive statistics for the elytra data

The UNIVARIATE Procedure
Variable: length
sex = F

Moments			
N	60	Sum Weights	60
Mean	4.94	Sum Observations	296.4
Std Deviation	0.48544929	Variance	0.23566102
Skewness	-0.521146	Kurtosis	0.16125847
Uncorrected SS	1478.12	Corrected SS	13.904
Coeff Variation	9.82690878	Std Error Mean	0.06267123

Basic Statistical Measures			
Location		Variability	
Mean	4.940000	Std Deviation	0.48545
Median	5.000000	Variance	0.23566
Mode	5.200000	Range	2.20000
		Interquartile Range	0.70000

Figure 3.2: `descriptive.sas - proc univariate`



Descriptive statistics for the elytra data

The UNIVARIATE Procedure
Variable: length
sex = M

Moments			
N	70	Sum Weights	70
Mean	4.71285714	Sum Observations	329.9
Std Deviation	0.44977335	Variance	0.20229607
Skewness	-0.896502	Kurtosis	1.00307174
Uncorrected SS	1568.73	Corrected SS	13.9584286
Coeff Variation	9.5435388	Std Error Mean	0.0537582

Basic Statistical Measures			
Location		Variability	
Mean	4.712857	Std Deviation	0.44977
Median	4.800000	Variance	0.20230
Mode	5.000000	Range	2.40000
		Interquartile Range	0.50000

Figure 3.3: `descriptive.sas - proc univariate`

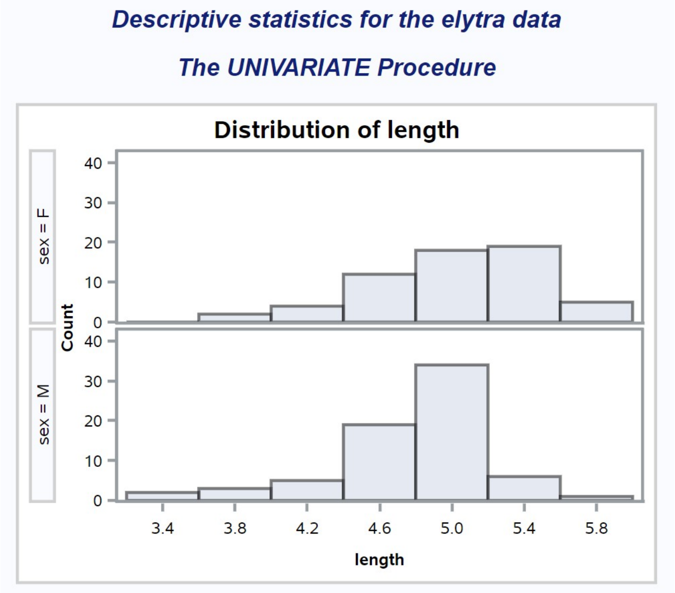


Figure 3.4: `descriptive.sas - proc univariate`

3.2.8 Mode

The mode is defined to be the most frequent value in the data set, and is another statistic of location. The mode in itself does not have many applications in biology, but is commonly used to describe the shape of a frequency distribution for the sample (see above). For example, we describe a frequency distribution as being unimodal if it has a single peak, and bimodal if there are two peaks. Examining the SAS output listed above, we see that female *T. dubius* beetles have a mode of 5.2 mm, while the mode for males is 5.0 mm. Both distributions appear to be unimodal.

3.2.9 Skewness

Skewness is a measure of the symmetry of the frequency distribution. Distributions that show an extended left tail to the frequency distribution, as well as the pattern mode > median > mean, are said to be skewed to the left. Fig. 3.5 shows an example of a left-skewed frequency distribution for some variable `y`. Conversely, distributions with an extended right tail and the pattern mean > median > mode are skewed to the right (Fig. 3.6). Skewness can be quantified by calculating the statistic g_1, given by the formula

$$g_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{Y_i - \bar{Y}}{s} \right)^3 . \tag{3.12}$$

The cubic terms here measure the asymmetry of the distribution. If the distribution is skewed to the left, with more values farther to the left than the right of $\bar{Y}$, there will tend to be large negative cubic terms, making $g_1 < 0$. Conversely, distributions skewed to the right will have large positive cubic terms and $g_1 > 0$. For distributions that are symmetrical we have $g_1 \approx 0$. For example, a frequency distribution for normally-distributed data would be symmetrical with $g_1 \approx 0$ (Fig. 3.7). For the elytra example, both male and female *T. dubius* have frequency distributions that appear skewed to the left, and also have negative g_1 values. Skewness is most often used as a description of the general shape of a distribution.

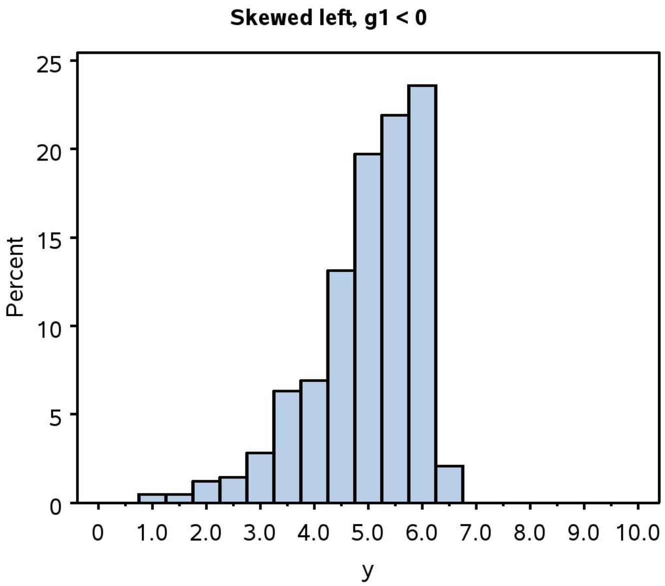


Figure 3.5: Frequency distribution that is skewed left ($g_1 < 0$).

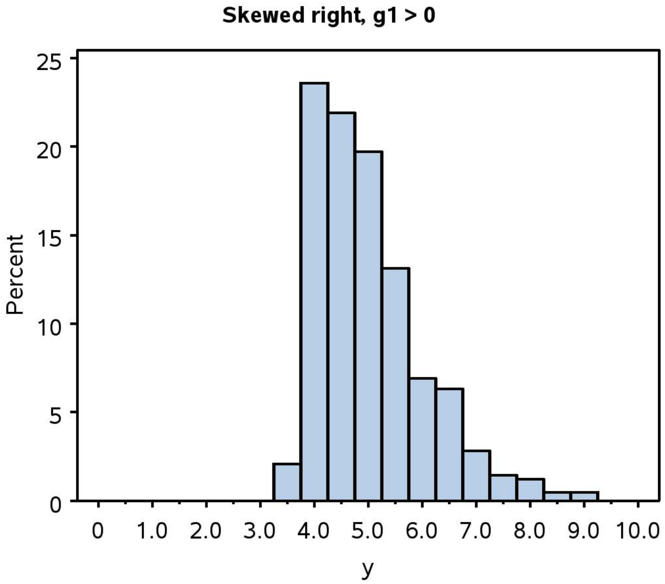


Figure 3.6: Frequency distribution that is skewed right ($g_1 > 0$).

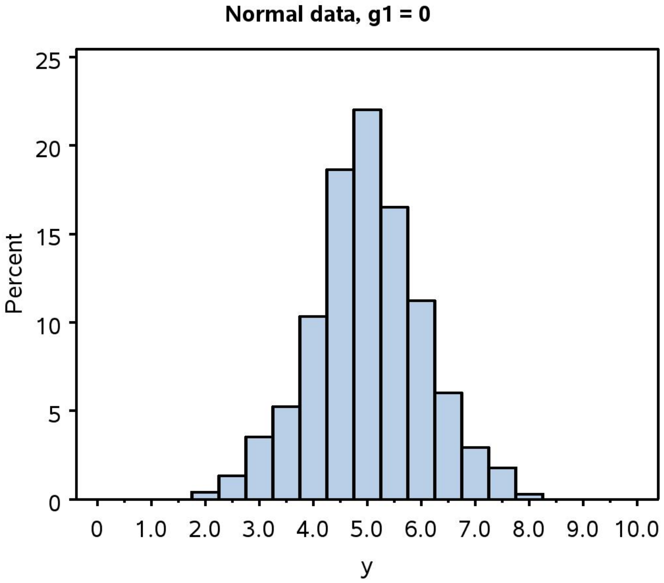


Figure 3.7: Frequency distribution for normal data ($g_1 \approx 0$).

3.2.10 Kurtosis

Kurtosis is a measure of how peaked or flat is a frequency distribution relative to the normal distribution. Distributions with a stronger central peak than the normal, and heavier left and right tails, are called leptokurtic (compare Fig. 3.8 and 3.10). Conversely, distributions with a weak peak and tails are called platykurtic (see Fig. 3.9 vs. 3.10). Kurtosis is quantified by calculating the statistic g_2:

$$g_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{Y_i - \bar{Y}}{s} \right)^4 - \frac{3(n-1)^2}{(n-2)(n-3)}. \tag{3.13}$$

The behavior of the terms in g_2 is less intuitive than those in the skewness statistic g_1. In any event, distributions that are leptokurtic have values of $g_2 > 0$, while platykurtic ones have $g_2 < 0$, with $g_2 \approx 0$ for distributions resembling the normal. For the elytra example, male *T. dubius* have a leptokurtic distribution with $g_2 = 1.003$, and the frequency distribution shows a strong central peak with heavy tails. The value of $g_2 = 0.161$ is smaller for female *T. dubius*, suggesting a shape more similar to the normal distribution. Like skewness, kurtosis is used to describe the general shape of the distribution.

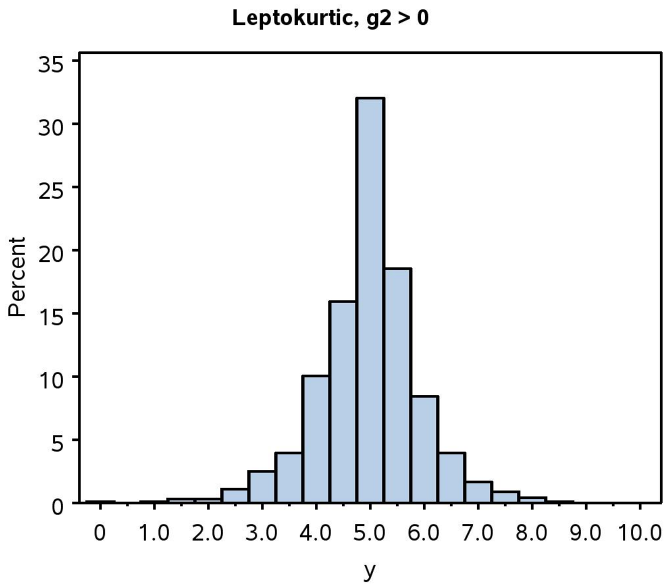


Figure 3.8: Frequency distribution that is leptokurtic ($g_2 > 0$).

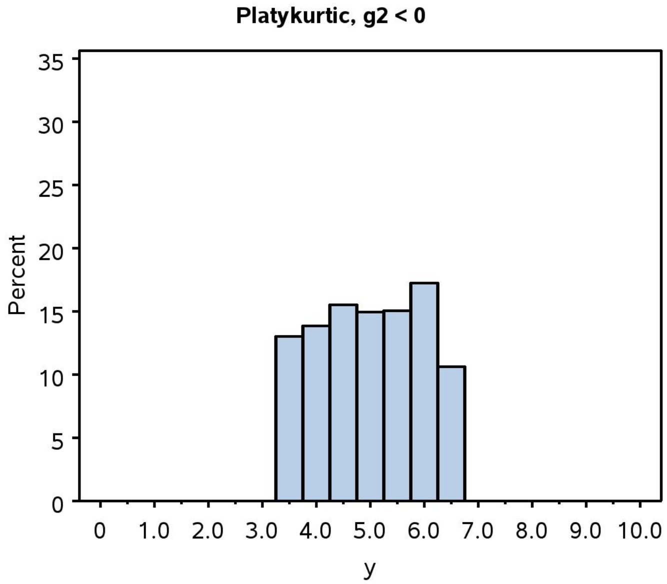


Figure 3.9: Frequency distribution that is platykurtic ($g_2 < 0$).

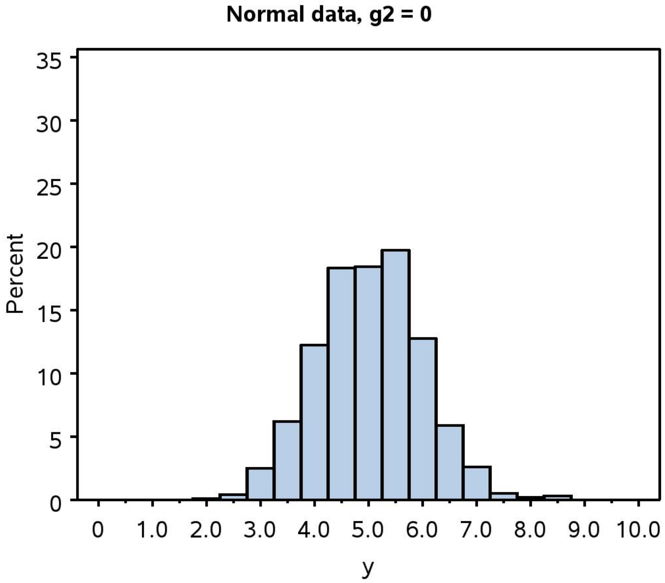


Figure 3.10: Frequency distribution for normal data ($g_2 \approx 0$).

3.2.11 Development time - SAS demo

We now examine another data set involving the development time of *T. dubius* reared under laboratory conditions (Reeve et al. 2003). Two different development times were measured, the time from the first larval stage until the prepupal stage, and the prepupal to adult stage. The program used to analyze these data is listed below. The `input` line is different than our previous program, because there are two variables (`time_pp` and `time_adult`) to analyze for each insect listed, which occur in two columns. The `var` and `histogram` statements in `proc univariate` are similar, listing the two variables so that descriptive statistics and frequency distributions are generated for both.

Note the periods (. values) given in the data set - these indicate missing values to SAS. In this study, observations were missing usually because the insect died before reaching the adult stage, but missing values can also be used to indicate lost data. The full data set for this example is listed in Chapter 22.

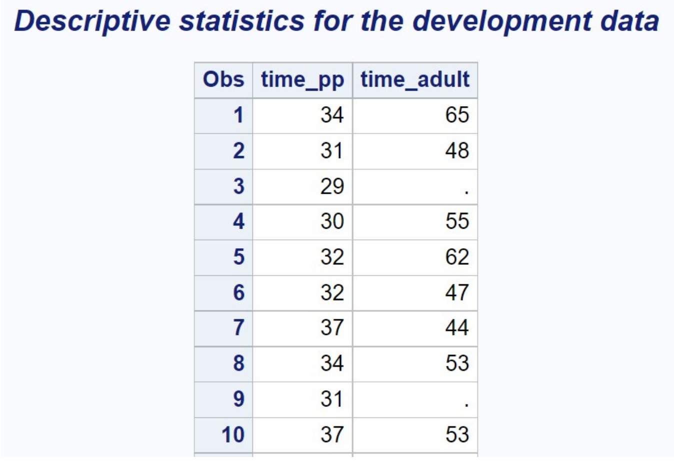
After running the program, we obtain output with statistics of location and dispersion as well as a frequency distribution, with a separate analysis for each variable. Clearly the larval-prepupal development time (`time_pp`) is shorter than the prepupal adult (`time_adult`) one ($\bar{Y} = 31.354$ vs. 75.353 days), and also shows less variability as indicated by the sample standard deviation ($s = 3.328$ vs. 26.347 days). Both variables appear to be skewed to the right, as indicated by positive values of g_1 as well as the result that mean > median > mode. Larval-prepupal development time shows little kurtosis ($g_2 = 0.047$), while prepupal-adult time apparently has a platykurtic distribution ($g_2 = -0.624$). This can also be observed in the frequency distribution for this variable, which is relatively flat compared to previous examples. Note that the distribution also appears to be somewhat bimodal, with two peaks of development time.

SAS Program

```
* descriptive_2.sas;
title 'Descriptive statistics for the development data';
data devel_time;
    input time_pp time_adult;
    datalines;
34  65
31  48
29   .
30  55
32  62

etc.

29   .
29 108
31 103
33   .
29  92
;
run;
* Print data set;
proc print data=devel_time;
run;
* Descriptive statistics, histograms, and normal quantile plots;
proc univariate data=devel_time;
    var time_pp time_adult;
    histogram time_pp time_adult / vscale=count;
run;
quit;
```

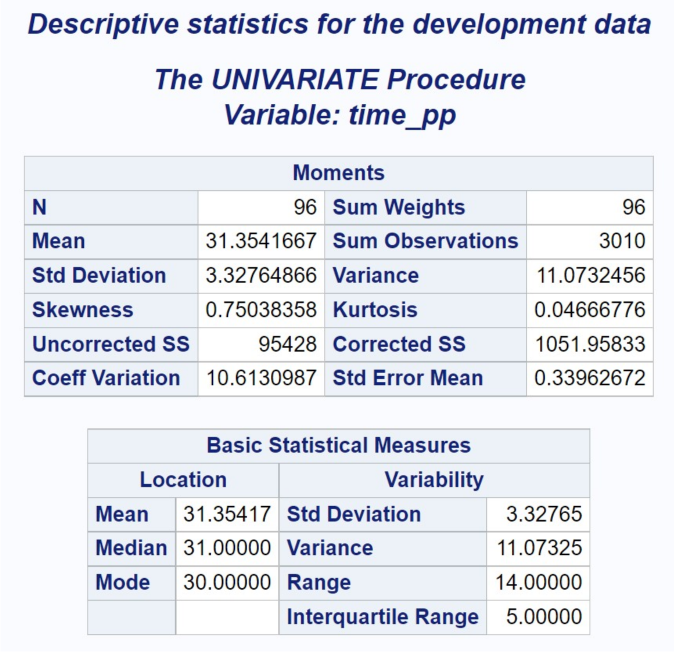


Descriptive statistics for the development data

Obs	time_pp	time_adult
1	34	65
2	31	48
3	29	.
4	30	55
5	32	62
6	32	47
7	37	44
8	34	53
9	31	.
10	37	53

etc.

Figure 3.11: `descriptive2.sas - proc print`



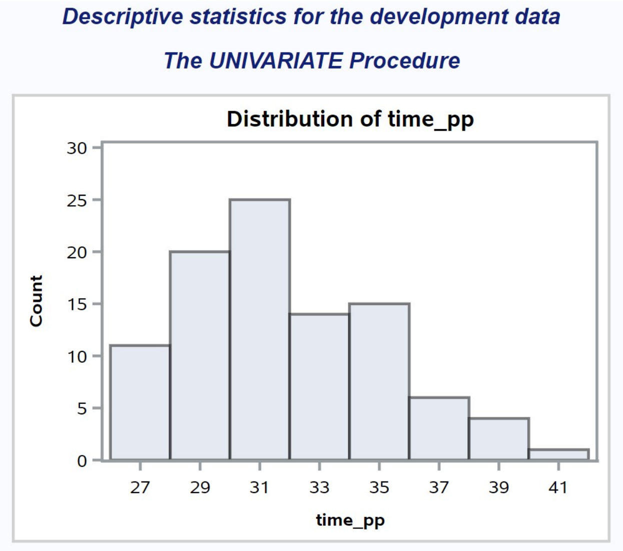
Descriptive statistics for the development data

The UNIVARIATE Procedure
Variable: time_pp

Moments			
N	96	Sum Weights	96
Mean	31.3541667	Sum Observations	3010
Std Deviation	3.32764866	Variance	11.0732456
Skewness	0.75038358	Kurtosis	0.04666776
Uncorrected SS	95428	Corrected SS	1051.95833
Coeff Variation	10.6130987	Std Error Mean	0.33962672

Basic Statistical Measures			
Location		Variability	
Mean	31.35417	Std Deviation	3.32765
Median	31.00000	Variance	11.07325
Mode	30.00000	Range	14.00000
		Interquartile Range	5.00000

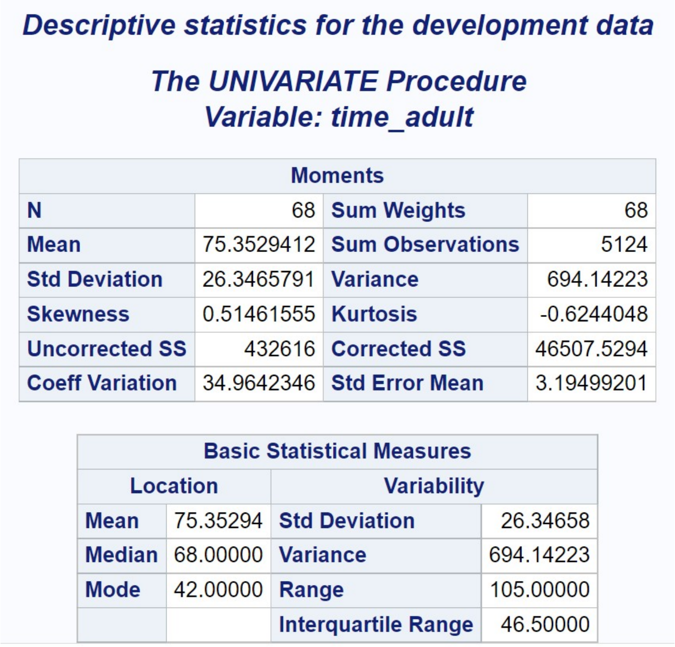
Figure 3.12: `descriptive2.sas - proc univariate`



Descriptive statistics for the development data

The UNIVARIATE Procedure

Figure 3.13: `descriptive2.sas - proc univariate`



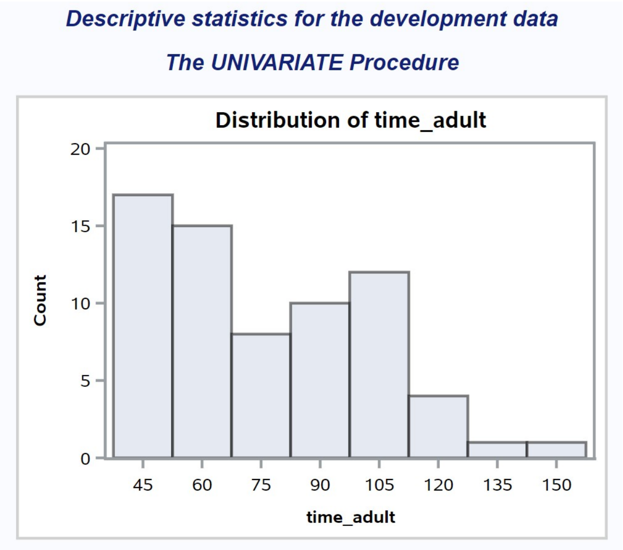
Descriptive statistics for the development data

The UNIVARIATE Procedure
Variable: time_adult

Moments			
N	68	Sum Weights	68
Mean	75.3529412	Sum Observations	5124
Std Deviation	26.3465791	Variance	694.14223
Skewness	0.51461555	Kurtosis	-0.6244048
Uncorrected SS	432616	Corrected SS	46507.5294
Coeff Variation	34.9642346	Std Error Mean	3.19499201

Basic Statistical Measures			
Location		Variability	
Mean	75.35294	Std Deviation	26.34658
Median	68.00000	Variance	694.14223
Mode	42.00000	Range	105.00000
		Interquartile Range	46.50000

Figure 3.14: `descriptive2.sas - proc univariate`



Descriptive statistics for the development data

The UNIVARIATE Procedure

Figure 3.15: `descriptive2.sas - proc univariate`

3.2.12 Frequency distributions for categorical data - SAS demo

The descriptive statistics we have developed so far are appropriate for continuous or discrete data. What about categorical data? One common way of summarizing categorical data is a frequency distribution, showing the number of occurrences in each category and possibly also their percentages. We can illustrate this process using the elytra data. There is one categorical variable in this data set, the sex of the beetle, and we might be interested in whether there were equal numbers of males and females. It also possible to derive categorical variables from the observations themselves. Suppose we classify a beetle as being ‘small’ if `length` is less than 5.0 mm, and ‘large’ otherwise. We can define this new variable within the SAS `data` set using an `if-then-else` statement. The code necessary to generate this new variable for the elytra data is shown below. It generates a new variable called `size` that takes the value `small` or `large` depending on the value of `length`.

```
* descriptive_freq.sas;
title ’Frequency distribution for the elytra data’;
data elytra;
    input sex $ length;
    * Classify insects into two groups by size;
    if length < 5.0 then size="small"; else size="large";
    datalines;
M   4.9
F   5.2
M   4.9
F   4.2
F   5.7

etc.

M   5.1
F   4.4
M   4.8
M   4.6
F   3.7
;
run;
```

We can then generate a frequency distribution for both `sex` and `size` using `proc freq` (SAS Institute Inc. 2016b). The `tables sex*size` statement will generate a two-way table of frequencies, classifying each observation into one of four categories (female-large, female-small, male-large, male-small). See below.

```
* Frequency distribution;
proc freq data=elytra;
    table sex*size;
run;
```

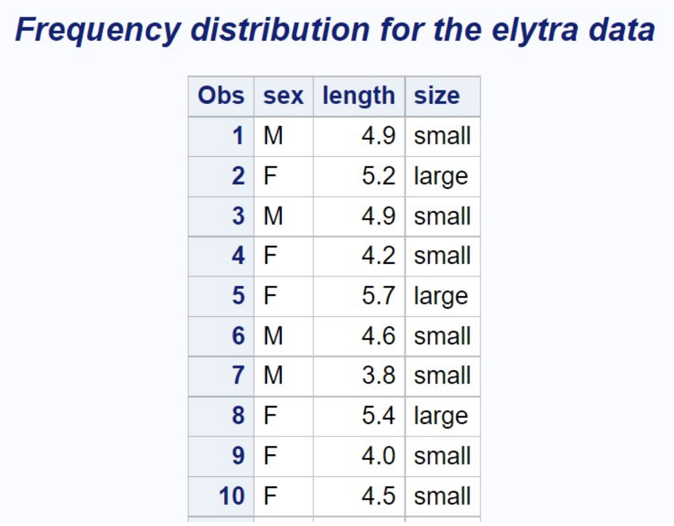
The complete program and output are listed below. From the frequency table generated by `proc freq`, we see that there are more males than females in the data set, and more small vs. large insects. Female beetles have a greater proportion of large insects than males.

SAS Program

```
* descriptive_freq.sas;
title 'Frequency distribution for the elytra data';
data elytra;
    input sex $ length;
    * Classify insects into two groups by size;
    if length < 5.0 then size="small"; else size="large";
    datalines;
M   4.9
F   5.2
M   4.9
F   4.2
F   5.7

etc.

M   5.1
F   4.4
M   4.8
M   4.6
F   3.7
;
run;
* Print data set;
proc print data=elytra;
run;
* Frequency distribution;
proc freq data=elytra;
    table sex*size;
run;
quit;
```

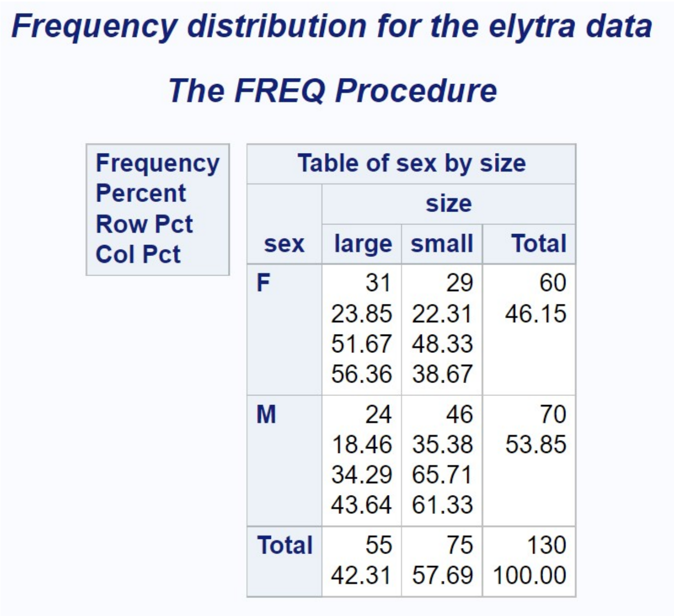


Frequency distribution for the elytra data

Obs	sex	length	size
1	M	4.9	small
2	F	5.2	large
3	M	4.9	small
4	F	4.2	small
5	F	5.7	large
6	M	4.6	small
7	M	3.8	small
8	F	5.4	large
9	F	4.0	small
10	F	4.5	small

etc.

Figure 3.16: `descriptive_freq.sas - proc print`



Frequency distribution for the elytra data

The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of sex by size			
	size		
sex	large	small	Total
F	31 23.85 51.67 56.36	29 22.31 48.33 38.67	60 46.15
M	24 18.46 34.29 43.64	46 35.38 65.71 61.33	70 53.85
Total	55 42.31	75 57.69	130 100.00

Figure 3.17: `descriptive_freq.sas - proc freq`

3.3 References

Berryman, A. A. (1988) *Dynamics of Forest Insect Populations: Patterns, Causes, Implications*. Plenum Press, New York, NY.

Lei, C.-H. & Armitage, K. B. (1980) Growth, development and body size of field and laboratory population of *Daphnia ambigua*. *Oikos* 35: 31-48.

Reeve, J. D., Rojas, M. G. & Morales-Ramos, J. A. (2003) Artificial diet and rearing methods for *Thanasimus dubius* (Coleoptera: Cleridae), a predator of bark beetles (Coleoptera: Scolytidae). *Biological Control* 27: 315-322.

SAS Institute Inc. (2016a) *SAS 9.4 Language Reference: Concepts, Sixth Edition.* SAS Institute Inc., Cary, NC.

SAS Institute Inc. (2016b) *Base SAS 9.4 Procedures Guide, Sixth Edition.* SAS Institute Inc., Cary, NC.

3.4 Problems

1. For the data below, find the mean, median, variance, standard deviation and CV using the formulas for these quantities and a calculator. Show the steps in your calculations. Feel free to check your answers using SAS.

   ```
   88.6  88.0  89.8  92.0  108.1  113.6  103.4  109.9  94.5  96.7  101.7
   ```

2. Ten adult females of the zooplankton species *Daphnia ambigua* were selected and their carapace length measured (μm) (Lei & Armitage 1980). The following data were obtained:

   ```
   487 429 428 378 410 401 358 392 414 480
   ```

   Calculate the mean, median, variance, standard deviation, and *CV* for these data by hand. Show all your calculations. Check your answers using SAS.

3. A laboratory study was conducted on the development time of another bark beetle predator, *Temnochila virescens* (Coleoptera: Trogositidae). The numbers listed below are the larval development time (days) of 35 insects.

   ```
    73  65  58  54  78  57  90
   103  59  52  73  67  67  53
    59  55  58  78  64  60  52
    96  68  81  76  77  57  79
    71  74  65  65  64  56  62
   ```

   (a) Use SAS to find the mean, median, mode, variance, standard deviation, and *CV* of these data, then plot a frequency distribution. Attach your program, output, and graph.

   (b) Examine the frequency distribution and skewness value (g_1) for these data. Do the data appear to be skewed, and if so in what direction? Explain your answer.

Chapter 4

Probability Theory

Probability theory is a branch of mathematics that is an essential component of statistics. It originally evolved from efforts to understand the odds and probabilities involved in games of chance, called classical probability theory (Weatherford 1982). The modern theory is developed from a small number of *a priori* axioms (like other mathematical theories) from which the rest the theory is deduced, including the behavior of probabilities and various rules for calculating them (Kolmogorov 1951, Weatherford 1982). While theoretical in origin, probability theory has proven to be spectacularly useful because it provides explanations for many natural processes, as well as the mathematical underpinnings for an enormous range of statistical procedures in the sciences.

4.1 Probability theory

4.1.1 Events

We can develop many elements of probability theory using a simple example, a single throw of a dice cube. If we throw the cube once, there are six possible outcomes corresponding to $1, 2, \ldots,$ or 6 spots appearing on the cube. We call the possible outcomes of this single throw of a dice cube a **sample space** S, commonly written using set notation as

$$S = \{1, 2, 3, 4, 5, 6\} \tag{4.1}$$

We now define as **events** various subsets of the elements in S. **Simple events** contain exactly one element of S. For S defined above, the simple events are $\{1\}, \{2\}, \{3\}, \{4\}, \{5\}$, and $\{6\}$. More specifically, the event $\{2\}$ signifies that a single throw of a dice cube showed two spots. More complex events contain more than one element of S. For example, consider the event A that the number of spots is odd, meaning that either one, three, or five spots showed on the dice cube after a single throw. We would write this event as the set $A = \{1,\ 3,\ 5\}$. Another possible event B is that the number of spots is less than or equal to three, or $B = \{1, 2, 3\}$. An event C such that number of spots is even would be written as $C = \{2, 4, 6\}$. Technically, both S itself and the empty set $\phi = \{\}$ are also possible events. S would always happen no matter the outcome of the throw, because some number of spots always occurs. The event corresponding to the empty set $\phi = \{\}$ would never happen because some number of spots always occurs after a throw. Sometimes certain events are subsets of other ones. For example, the event A defined above is a subset of S because every element of A is contained in S. This is written as $A \subset S$ using set notation.

4.1.2 Union, intersection, and complement of events

We now consider various combinations of events, again using our dice example. The **union** of two events A and B is defined to be the set containing all the simple events in A and B. The union is written using the notation $A \cup B$. For example, consider two of the events defined above for the dice example, $A = \{1, 3, 5\}$ (the number of spots is odd) and $B = \{1, 2, 3\}$ (the number of spots is less than or equal to three). We have

$$A \cup B = \{1, 3, 5\} \cup \{1, 2, 3\} = \{1, 2, 3, 5\}. \tag{4.2}$$

The union of two events can also be visualized using Venn diagrams, with the events A and B represented by circles and the shaded area their union (Fig. 4.1). The rectangle labeled S represents the entire sample space.

The **intersection** of two events A and B is defined to be the set containing simple events present in both A and B. The intersection is written using the notation $A \cap B$ or just AB. For example, consider the events A and B from the dice example. We have

$$A \cap B = \{1, 3, 5\} \cap \{1, 2, 3\} = \{1, 3\}. \tag{4.3}$$

The intersection of these two events is shown by the shaded area in Fig. 4.2. It is possible to have the intersection of two events be the empty set ϕ.

Consider the events A (spots is odd) and C (spots even) for the dice example. We have

$$A \cap C = \{1,3,5\} \cap \{2,4,6\} = \{\} = \phi. \tag{4.4}$$

Fig. 4.3 shows this outcome, with no shaded area because the intersection is empty. When the intersection of two events is the empty set, we say the two events are **mutually exclusive**. This means either one or the other event has occurred – it is impossible for them to happen at the same time.

The **complement** of an event A is the set of simple events in S remaining after we subtract those in A, typically written as A^c. For the event $A = \{1,3,5\}$ from the dice example, we have $A^c = \{2,4,6\}$, the simple events remaining in S after we substract those in A. Using set notation, $A^c = S - A = \{1,2,3,4,5,6\} - \{1,3,5\} = \{2,4,6\}$. We can also represent A^c in a diagram with the shaded area representing all outcomes outside of A (Fig. 4.4). Complements of events frequently arise in the use of statistical tables.

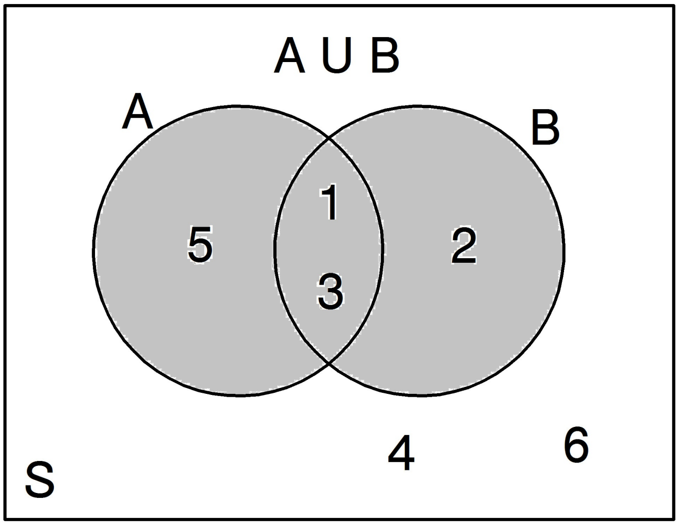


Figure 4.1: $A \cup B = \{1,2,3,5\}$.

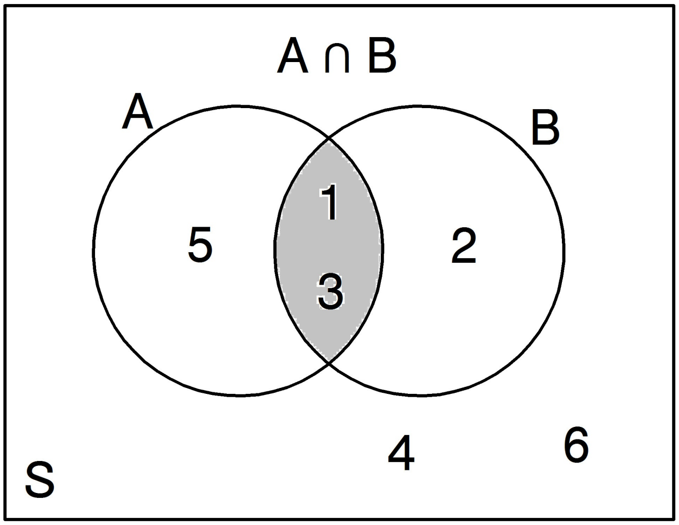


Figure 4.2: $A \cap B = \{1, 3\}$.

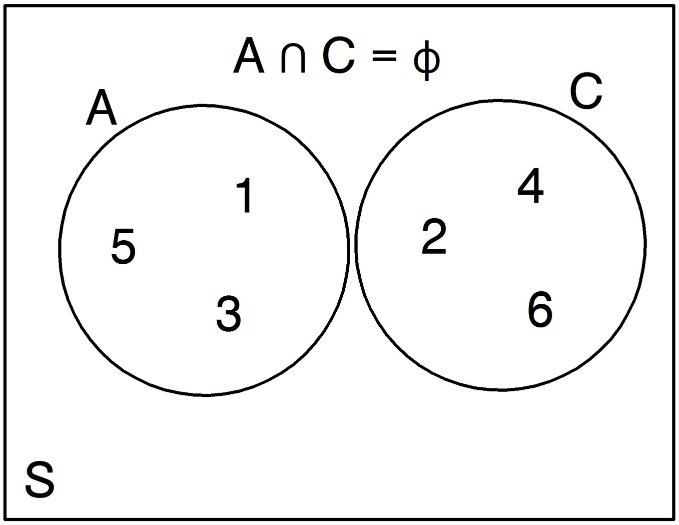


Figure 4.3: $A \cap C = \phi$.

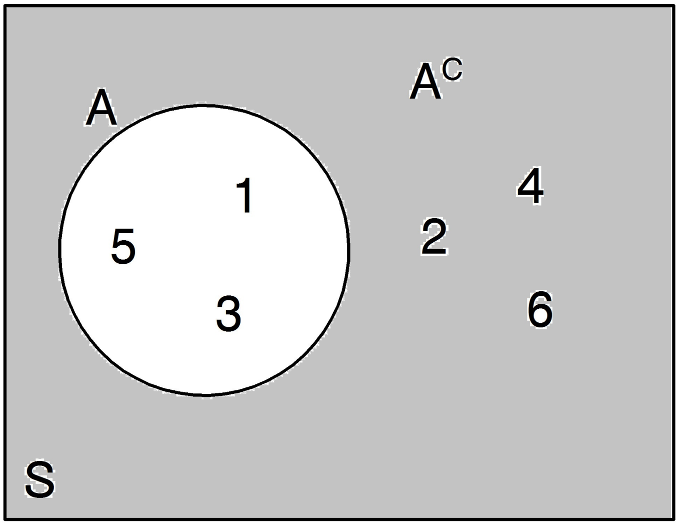


Figure 4.4: $A^C = \{2, 4, 6\}$.

4.1.3 Probability distributions

We now describe how probabilities are attached to events using a probability distribution, which can be mathematically defined based on certain axioms (Mood et al. 1974). Here, we simply list a number of properties of probability distributions that are useful to the practicing statistician. Let the symbol $P[A]$ stand for the probability of some event A. For a sample space S and any two events A and B, we have

1. $P[A] \geq 0$.
2. $P[S] = 1$.
3. $P[\phi] = 0$, where ϕ is the empty set.
4. $P[A \cup B] = P[A] + P[B] - P[A \cap B]$.
5. $P[A^c] = 1 - P[A]$.

While we have listed some of the properties of a probability distribution, we have not actually defined one yet. Recall the dice example, in which a single dice cube is thrown and the number of spots observed. If the cube is fair, then it is reasonable to assume that each number is equally likely to occur, and there are six possible numbers, so we assign a probability of 1/6 to each number. In particular, we have

$$P[\{1\}] = P[\{2\}] = P[\{3\}] = P[\{4\}] = P[\{5\}] = P[\{6\}] = 1/6 \tag{4.5}$$

How should we assign probabilities to events like $A = \{1, 3, 5\}$? We define these events to have a probability equal to the sum of the probabilities for each simple event within them. For example, we have

$$\begin{aligned} P[A] = P[\{1,3,5\}] &= P[\{1\}] + P[\{3\}] + P[\{5\}] && (4.6) \\ &= 1/6 + 1/6 + 1/6 = 3/6 = 1/2. && (4.7) \end{aligned}$$

This result also makes intuitive sense for the event A, because we would expect the dice cube to produce an odd number of spots half of the time. We can view this probability distribution as a model of the dice cube's behavior, which would be accurate if the dice cube is fair. This is a common task faced by a statistician in analyzing a problem – determine an appropriate probability distribution to describe a particular type of data.

We will now calculate the probabilities for certain events to illustrate how this probability distribution can be used. Recall that the sample space for this distribution is $S = \{1, 2, 3, 4, 5, 6\}$. Suppose we have three events, namely $A = \{1, 3, 5\}$ (an odd number of spots), $B = \{1, 2, 3\}$ (less than or equal to three spots), and $C = \{2, 4, 6\}$ (an even number of spots).

We have already illustrated how to find the probability for A. For B, we have

$$\begin{aligned} P[B] &= P[\{1,2,3\}] = P[\{1\}] + P[\{2\}] + P[\{3\}] \qquad (4.8)\\ &= 1/6 + 1/6 + 1/6 = 3/6 = 1/2. \qquad (4.9)\end{aligned}$$

For C the probability is

$$\begin{aligned} P[C] &= P[\{2,4,6\}] = P[\{2\}] + P[\{4\}] + P[\{6\}] \qquad (4.10)\\ &= 1/6 + 1/6 + 1/6 = 3/6 = 1/2. \qquad (4.11)\end{aligned}$$

For the sample space S, which is also an event, we have

$$\begin{aligned} P[S] &= P[\{1,2,3,4,5,6\}] \qquad (4.12)\\ &= P[\{1\}] + P[\{2\}] + P[\{3\}] + P[\{4\}] + P[\{5\}] + P[\{6\}] \qquad (4.13)\\ &= 1/6 + 1/6 + 1/6 + 1/6 + 1/6 + 1/6 = 6/6 = 1. \qquad (4.14)\end{aligned}$$

We also have $P[\{\}] = P[\phi] = 0$ because it is impossible to have no spots showing on the dice cube.

What is the probability for $A \cap B$? We have

$$\begin{aligned} P[A \cap B] &= P[\{1,3,5\} \cap \{1,2,3\}] = P[\{1,3\}] \qquad (4.15)\\ &= P[\{1\}] + P[\{3\}] = 1/6 + 1/6 = 1/3. \qquad (4.16)\end{aligned}$$

For $A \cup B$ we can calculate the probability in two ways. We can directly find it as follows. We have

$$\begin{aligned} P[A \cup B] &= P[\{1,3,5\} \cup \{1,2,3\}] = P[\{1,2,3,5\}] \qquad (4.17)\\ &= P[\{1\}] + P[\{2\}] + P[\{3\}] + P[\{5\}] \qquad (4.18)\\ &= 1/6 + 1/6 + 1/6 + 1/6 = 2/3. \qquad (4.19)\end{aligned}$$

We can also use the formula listed in Property 4 to find this probability. We have

$$\begin{aligned} P[A \cup B] &= P[A] + P[B] - P[A \cap B] \qquad (4.20)\\ &= 1/2 + 1/2 - 1/3 = 2/3. \qquad (4.21)\end{aligned}$$

We obtain the same answer as by direct calculation.

We can understand how the Property 4 formula works by considering the diagram for $A \cup B$ (Fig. 4.1). Suppose that the shaded area for event A represents the probability for A, and similarly for event B. If we add $P[A]$ and $P[B]$ together, this would actually be greater than $P[A \cup B]$ because it counts the area of the intersection ($A \cap B$) twice. This explains why we need to subtract $P[A \cap B]$ in Property 4 to obtain $P[A \cup B]$.

We now find the probability for A^c. We can directly calculate it by finding the probability for $A^c = S - A = \{1, 2, 3, 4, 5, 6\} - \{1, 3, 5\} = \{2, 4, 6\}$, so $P[A^c] = P[\{2, 4, 6\}] = 1/2$. Alternately, by Property 5 above,

$$P[A^c] = 1 - P[A] = 1 - 1/2 = 1/2. \tag{4.22}$$

Property 5 can also be explained by a diagram. The rectangle in Fig. 4.4 represents the sample space S, and by Property 2 we have $P[S] = 1$. If the circle for event A represents $P[A]$, then clearly $P[A^c] = 1 - P[A]$.

4.1.4 Probability spaces

The combination of a sample space S, a collection of all possible events on the sample space (A, B, S, ϕ, etc.), and a probability distribution is called a **probability space**.

4.1.5 Independence of events

Independence of events is an important concept in statistics, and basically implies that an event A has no effect on whether B occurs, and vice versa. In terms of probabilities, two events A and B are defined to be independent if

$$P[A \cap B] = P[A]P[B]. \tag{4.23}$$

Are the events $A = \{1, 3, 5\}$ and $B = \{1, 2, 3\}$ defined for the dice cube example independent? We have

$$P[A \cap B] = P[\{1, 3, 5\} \cap \{1, 2, 3\}] = P[\{1, 3\}] = 1/3. \tag{4.24}$$

However,

$$P[A]P[B] = 1/2 \times 1/2 = 1/4. \tag{4.25}$$

This implies that A and B are not independent because $P[A \cap B] \neq P[A]P[B]$. To see why this happens, observe that when the number of spots is less than or equal to three (B occurs), the number of spots is more likely to be odd (A occurs) because two of the three outcomes in B are odd.

We now work an example where the two events are independent. Suppose that $D = \{1, 2, 3, 4\}$, the event that the number of spots is less than or equal to four. Are A and D independent? We have

$$\begin{aligned} P[A \cap D] &= P[\{1, 3, 5\} \cap \{1, 2, 3, 4\}] && (4.26) \\ &= P[\{1, 3\}] = 1/6 + 1/6 = 1/3, && (4.27) \end{aligned}$$

and

$$\begin{aligned} P[A]P[D] &= 1/2 \times P[\{1, 2, 3, 4\}] && (4.28) \\ &= 1/2 \times (1/6 + 1/6 + 1/6 + 1/6) && (4.29) \\ &= 1/2 \times 2/3 = 1/3. && (4.30) \end{aligned}$$

This implies that A and D are independent because $P[A \cap D] = P[A]P[D]$. This outcome seems reasonable – when the number of spots is less than or equal to four (D occurs), the probability of the number of spots being odd is still equal to 1/2 because half of the outcomes in D are odd.

4.1.6 Conditional probability

Suppose that an event B has already happened, so that we have some information on a particular system or situation. Could this affect the probability that some other event A would occur? This is the idea behind **conditional probability**, an important concept in statistics that is related to independence. The conditional probability of an event A, given that B has occurred, is given by the formula

$$P[A|B] = \frac{P[A \cap B]}{P[B]}. \tag{4.31}$$

The notation '$A|B$' is read as A given B. For the dice cube example, what is the conditional probability of $A = \{1, 3, 5\}$ given that $B = \{1, 2, 3\}$ has occurred? We have

$$P[A|B] = \frac{P[A \cap B]}{P[B]} = \frac{1/3}{1/2} = 2/3. \tag{4.32}$$

Note that the $P[A|B] > P[A]$ because $2/3 > 1/2$. Thus, if B has occurred it is more likely that A occurs, because two of three outcomes in B are odd.

If two events are independent, implying that $P[A \cap B] = P[A]P[B]$, then we have

$$P[A|B] = \frac{P[A \cap B]}{P[B]} = \frac{P[A]P[B]}{P[B]} = P[A]. \tag{4.33}$$

Thus, if two events are independent then the fact that B has occurred does not alter the probability for A. We can illustrate this for the dice cube example using the events $A = \{1, 3, 5\}$ and $D = \{1, 2, 3, 4\}$, which we earlier showed to be independent. We have

$$P[A|D] = \frac{P[A \cap D]}{P[D]} = \frac{P[A]P[D]}{P[D]} \tag{4.34}$$

$$= \frac{1/3}{2/3} = 1/2 = P[A]. \tag{4.35}$$

Thus, if D has occurred it has no effect on the probability of A occurring. This follows because half the events in D are odd, and so the probability of obtaining an odd number (1/2) is exactly the same as the original probability.

4.1.7 A biological probability distribution

We now examine a more biological example involving the infection of amphibians by the chytrid fungus *Batrachochytrium dendrobatidis*, which appears responsible for the decline of amphibians in some regions (Lips et al. 2006). Certain amphibian species appear less susceptible than others by virtue of natural immunity or their ecological traits (Lips et al. 2003), and we would expect infection rates to therefore vary among species. Suppose we know that at a particular location the amphibians can be classified into three common species (A, B, and C) that can also be divided into infected and uninfected individuals, with the frequency of individuals in each category having the distribution given in Table 4.1. In practice, we would need to estimate these proportions, but we will assume they are already known.

Table 4.1: Proportions of individuals from three amphibian species (A, B, and C), classified as infected (Yes) or free of chytrid fungus (No).

		Species		
		A	B	C
Infected	No	0.25	0.2	0.15
	Yes	0.25	0.1	0.05

Suppose we now sample a single individual from this location. The sample space would be $S = \{$A-Yes, A-No, B-Yes, B-No, C-Yes, C-No$\}$. Here 'A-No' stands for an amphibian of species A that is free of fungus, and is one of six simple events. The probability of sampling an A-No individual would be $P[\text{A-No}] = 0.25$, with the probabilities for other simple events given by the entries in Table 4.1. Note that $P[S] = P[\{$A-No, A-Yes, B-No, B-Yes, C-No, C-Yes$\} = 0.25 + 0.25 + 0.2 + 0.1 + 0.15 + 0.05 = 1$ as is necessary for a probability distribution.

We now calculate the probabilities for certain events. Suppose that A is the event that species A is sampled, implying that $A = \{$A-No, A-Yes$\}$. We have

$$\begin{aligned} P[A] &= P[\{\text{A-No}, \text{A-Yes}\}] && (4.36) \\ &= P[\{\text{A-No}\}] + \mathrm{P}[\{\text{A-Yes}\}] && (4.37) \\ &= 0.25 + 0.25 = 0.5 && (4.38) \end{aligned}$$

Thus, we would expect half the amphibians sampled to be species A. Suppose we also want to find the probability for A^c. By Property 5 above, we have

$$P[A^c] = 1 - P[A] = 1 - 0.5 = 0.5 \qquad (4.39)$$

Now let B be the event that species B is sampled, so that $B = \{$B-No, B-Yes$\}$ and $P[B] = P[\{\text{B-No}, \text{B-Yes}\}] = \mathrm{P}[\{\text{B-No}\}] + \mathrm{P}[\{\text{B-Yes}\}] = 0.2 + 0.1 = 0.3$. What is the probability for $A \cap B$? We see that A and B share no simple events, so $P[A \cap B] = P[\{\}] = P[\phi] = 0$. The two events are therefore mutually exclusive, which is not surprising because the sampled amphibian can only be species A or B, not both.

What happens for $A \cup B$? We can directly calculate this probability by finding the simple events in $A \cup B$. We have

$$\begin{aligned} P[A \cup B] &= P[\{\text{A-No}, \text{A-Yes}\} \cup \{\text{B-No}, \text{B-Yes}\}] && (4.40)\\ &= P[\{\text{A-No}, \text{A-Yes}, \text{B-No}, \text{B-Yes}\}] && (4.41)\\ &= P[\{\text{A-No}\}] + \text{P}[\{\text{A-Yes}\}] + \text{P}[\{\text{B-No}\}] + \text{P}[\{\text{B-Yes}\}] && (4.42)\\ &= 0.25 + 0.25 + 0.20 + 0.10 = 0.80. && (4.43) \end{aligned}$$

An alternate way to calculate this probability uses Property 4 listed above. In particular,

$$\begin{aligned} P[A \cup B] &= P[A] + P[B] - P[A \cap B] && (4.44)\\ &= 0.5 + 0.3 - 0 = 0.8, && (4.45) \end{aligned}$$

the same answer as before.

We now define an event I which stands for infected amphibians, meaning $I = \{$A-Yes, B-Yes, C-Yes$\}$. We have

$$\begin{aligned} P[I] &= P[\{\text{A-Yes}, \text{B-Yes}, \text{C-Yes}\}] && (4.46)\\ &= P[\{\text{A-Yes}\}] + \text{P}[\{\text{B-Yes}\}] + \text{P}[\{\text{C-Yes}\}] && (4.47)\\ &= 0.25 + 0.1 + 0.05 = 0.4 && (4.48) \end{aligned}$$

This means that the overall probability of sampling an infected animal is 0.4. Suppose that we already know the sampled amphibian is species C. What is the probability that it is infected given it is species C, or $P[I|C]$? We have

$$\begin{aligned} P[I|C] &= \frac{P[I \cap C]}{P[C]} = \frac{P[\{\text{A-Yes}, \text{B-Yes}, \text{C-Yes}\} \cap \{\text{C-No}, \text{C-Yes}\}]}{P[\{\text{C-No}, \text{C-Yes}\}]} && (4.49)\\ &= \frac{P[\{\text{C-Yes}\}]}{P[\{\text{C-No}, \text{C-Yes}\}]} && (4.50)\\ &= \frac{0.05}{0.2} = 0.25. && (4.51) \end{aligned}$$

Thus, if an individual of species C has been sampled the probability of it being infected is 0.25. We can also see this by examining the column for species C in Table 4.1, where the proportion of infected animals is $0.05/(0.15 + 0.05) = 0.25$.

4.1.8 Bayes theorem

Another use of conditional probability involves Bayes Theorem, named for the Reverend Thomas Bayes, an eighteenth century clergyman who first derived the theorem. The theorem is often used in the interpretation of medical tests as well as the field of Bayesian statistics (Ellison 1996).

Recall the example above involving amphibians and their infection by chytrid fungus. Let D be the event an amphibian actually has the disease while D^C implies they are disease-free. Now suppose a particular test is used to determine if a sampled amphibian has the disease. Let T be the event the amphibian tests positive for the disease, while T^c means the amphibian tests negative. The test is less than perfect, however, and sometimes gives a positive result when the amphibian is disease-free (a false positive) and a negative one when it is diseased (a false negative). What we would like to calculate is the probability that an amphibian actually has the disease given that it tests positive, or $P[D|T]$. This is called the **positive predictive value** of the test.

What is known for the test is the probability of testing positive for amphibians with the disease, $P[T|D]$, called the **sensitivity** of the test. This would be determined by testing a large number of amphibians that are known to have the disease by other means, and finding the proportion that test positive. Also known is the probability of testing negative for amphibians that are disease-free, $P[T^c|D^c]$, called the **specificity** of the test. We will also need an estimate of the probability that an amphibian has the disease in the population as a whole, $P[D]$, called the **prevalence** of the disease.

To find $P[D|T]$, we begin by using the definition of conditional probability:

$$P[D|T] = \frac{P[D \cap T]}{P[T]} = \frac{P[T \cap D]}{P[T]} \tag{4.52}$$

We can also write

$$P[T|D] = \frac{P[T \cap D]}{P[D]}, \tag{4.53}$$

which implies that $P[T \cap D] = P[T|D]P[D]$. Inserting this result into Eq. 4.52, we obtain

$$P[D|T] = \frac{P[T|D]P[D]}{P[T]}. \tag{4.54}$$

We are nearly there, except that we need to express $P[T]$ in terms of known quantities. The event T is made up of two mutually exclusive groups, amphibians that test positive and have the disease ($T \cap D$), and ones that test positive that are disease-free ($T \cap D^c$). From above, we have $P[T \cap D] = P[T|D]P[D]$ and can similarly show that $P[T \cap D^c] = P[T|D^c]P[D^c]$. Because the two groups are mutually exclusive, we can write $P[T]$ as the sum of the probabilities for each group:

$$P[T] = P[T \cap D] + P[T \cap D^c] = P[T|D]P[D] + P[T|D^c]P[D^c]. \tag{4.55}$$

Substituting this quantity into Eq. 4.52, we obtain Bayes' theorem:

$$P[D|T] = \frac{P[T|D]P[D]}{P[T|D]P[D] + P[T|D^c]P[D^c]}. \tag{4.56}$$

Because $P[T|D^c] = 1 - P[T^c|D^c]$ and $P[D^c] = 1 - P[D]$, we can also write Bayes' theorem as

$$P[D|T] = \frac{P[T|D]P[D]}{P[T|D]P[D] + (1 - P[T^c|D^c])(1 - P[D])}. \tag{4.57}$$

or

$$P[D|T] = \frac{\text{sensitivity} \times \text{prevalence}}{\text{sensitivity} \times \text{prevalence} + (1 - \text{specificity}) \times (1 - \text{prevalence})}. \tag{4.58}$$

We can thus express the theorem in terms of the sensitivity and specificity of the test, and the overall prevalence of the disease, which are known quantities.

Bayes theorem – sample calculation

Suppose that the test for amphibian disease has a high sensitivity ($P[T|D] = 0.95$) as well as a high specificity ($P[T^c|D^c] = 0.90$). A particular amphibian population has a fairly high prevalence of the disease ($P[D] = 0.25$, implying 25% are infected). What is the probability that an animal that tests positive from this population has the disease, $P[D|T]$? Inserting these quantities in Bayes theorem (Eq. 4.57 or 4.58), we obtain

$$\begin{aligned}
P[D|T] &= \frac{P[T|D]P[D]}{P[T|D]P[D] + (1 - P[T^c|D^c])(1 - P[D])} && (4.59)\\
&= \frac{0.95 \times 0.25}{0.95 \times 0.25 + (1 - 0.9) \times (1 - 0.25)} && (4.60)\\
&= \frac{0.2375}{0.2375 + 0.075} && (4.61)\\
&= 0.76 && (4.62)
\end{aligned}$$

So, the probability that an animal that tests positive actually has the disease is 0.76. We now examine what happens if the prevalence of the disease is lower, say $P[D = 0.05]$, implying only 5% are infected. We have

$$\begin{aligned} P[D|T] &= \frac{P[T|D]P[D]}{P[T|D]P[D] + (1 - P[T^c|D^c])(1 - P[D])} \quad (4.63) \\ &= \frac{0.95 \times 0.05}{0.95 \times 0.05 + (1 - 0.9) \times (1 - 0.05)} \quad (4.64) \\ &= \frac{0.0475}{0.0475 + 0.095} \quad (4.65) \\ &= 0.3333 \quad (4.66) \end{aligned}$$

Now the probability that the animal has the disease is only 0.3333, despite using exactly the same sensitivity and specificity values for the test. What has happened here?

The explanation is that when prevalence is low, the majority of positive test results are actually false positives, in which disease-free animals test positive. This is reflected in the denominator of Eq. 4.63, where the term 0.095 (the probability of testing positive and being disease-free) is actually larger than the term .0475 (the probability of testing positive and having the disease). To fix this problem it would be helpful to have a test with higher specificity to reduce the incidence of false positives.

Bayesian statistics

Another type of probability theory, called subjective or Bayesian probability theory, equates probability with a degree of belief on the part of the analyst (Weatherford 1982). This theory makes use of Bayes theorem but with a different interpretation of the probabilities. Suppose that $P[D]$ is the belief by an investigator that a particular animal has the disease before the test, rather than the prevalence (frequency of the disease) in the amphibian population. The value of $P[D|T]$ calculated using Bayes' theorem now represents the investigator's belief that the animal has the disease after observing a positive test result. These two probabilities are simple examples of the prior and posterior distributions used in Bayesian statistics. See Ellison et al. (1996) and Ellison (2004) for a summary of arguments for Bayesian statistics, which is based on this interpretation of probability as belief. Dennis (1996) and Lele and Dennis (2009) provide arguments against Bayesian statistics and in favor of ‘frequentist’ statistics, the kind of statistics based on the form of probability developed in this chapter. While frequentist statistics has its problems, it remains the most commonly used method in many fields and has a long successful record in science.

4.2 References

Dennis, B. (1996) Discussion: should ecologists become Bayesians? *Ecological Applications* 6: 1095-1103.

Ellison, A. M. (1996) An introduction to Bayesian inference for ecological research and environmental decision-making. *Ecological Applications* 6: 1036-1046.

Ellison, A. M. (2004) Bayesian inference in ecology. *Ecology Letters* 7: 509-520.

Lele, S. R., and B. Dennis (2009) Bayesian methods for hierarchical models: Are ecologists making a Faustian bargain? *Ecological Applications* 19: 581-584.

Lips, K. R., F. Brem, R. Brenes, J. D. Reeve, R. A. Alford, J. Voyles, C. Carey, L. Livo, A. P. Pessier, and J. P. Collins (2006) Emerging infectious disease and the loss of biodiversity in a Neotropical amphibian community. *Proceedings of the National Academy of Sciences* 103: 3165-3170.

Lips, K. R., J. D. Reeve, and L. R. Witters (2003) Ecological traits predicting amphibian population declines in Central America. *Conservation Biology* 17: 1078-1088.

Kolmogorov, A. (1951) *Foundations of the Theory of Probability.* Chelsea, New York, NY.

Mood, A. M., Graybill, F. A. & Boes, D. C. (1974) *Introduction to the Theory of Statistics.* McGraw-Hill, Inc., New York, NY.

Mettlin, C., Littrup, P. J., Kane, R. A., Murphy, G. P, Lee, F., Chesley, A., Badalament, R. & Mostofi, F. K. (1994) Relative sensitivity and specificity of serum prostate specific antigen (PSA) level compared with age-referenced PSA, PSA density, and PSA change. *Cancer* 74: 1615-1620.

Weatherford, R. (1982) *Philosophical Foundations of Probability Theory.* Routledge & Kegan Paul Ltd., Boston, MA.

4.3 Problems

1. Suppose you have a loaded dice cube, such that $P[\{1\}] = 0.1, P[\{2\}] = 0.1, P[\{3\}] = 0.1, P[\{4\}] = 0.2, P[\{5\}] = 0.2$, and $P[\{6\}] = 0.3$. The cube is tossed a single time and the number of spots observed. Answer the following questions. Note that 'and' denotes an intersection of events, 'or' the union of events, and 'given' a conditional probability.

   (a) What is the probability that the number is even?

   (b) What is the probability that the number is odd and ≥ 3?

   (c) Are the events odd and ≥ 3 independent?

   (d) What is the probability that the number is odd or ≥ 3?

   (e) What is the probability that the number is odd, given that it is ≥ 3?

2. The PSA (prostate specific antigen) test is used to screen older men for prostate cancer. This test has a sensitivity of 0.90 and specificity of 0.719 (Mettlin et al. 1994). Assuming a prevalence of 0.1, find the probability that an individual with a positive test has cancer. Show your calculations.

3. Suppose you know that a particular animal population consists of 40% juveniles and 60% adults, and have a sample of two animals selected at random from the population.

   (a) What is the sample space for this scenario?

   (b) In a sample of two animals, what is the probability of obtaining two juveniles in a row?

   (c) What is the probability of obtaining one adult and one juvenile, in that order?

   (d) What is the probability of obtaining one juvenile and one adult, in that order?

   (e) What is the probability of obtaining two adults in a row?

Chapter 5

Discrete Random Variables

Random variables and their associated probability distributions are a basic component of statistical analyses. A statistician will examine the experiment or study and determine the type of observations or data it produces (continuous, discrete, or categorical) and then select a random variable and its distribution to model these data. We examine here three discrete random variables, the binomial, Poisson, and negative binomial, and their probability distributions. There are other discrete random variables but these three are the most commonly encountered in practice. These variables only take integer values and are typically used to model discrete or count data. We will also see how to calculate the mean and variance for a discrete random variable, using its probability distribution and a quantity called the **expected value**.

The basic concept of a **random variable** is to map the outcome of some random event into a number. For example, consider the dice cube example from Chapter 4. Define a number Y that is the number of spots showing on the dice – Y is a random variable. The sample space for Y would be $S = \{1, 2, 3, 4, 5, 6\}$ and the events any combination of these values. One requirement for Y to be a random variable is that events of the form $Y \leq y$ for any real number y are events in the probability space (Mood et al. 1974). For example, suppose that $y = 3.5$ for the dice cube example. The set defined by $Y \leq 3.5$ corresponds to the event $A = \{1, 2, 3\}$ and so is a member of the probability space for this example. This requirement is necessary in order to calculate probabilities for the random variable, and there is always a probability distribution associated with a particular random variable.

5.1 Binomial distribution

Binomial random variables are commonly used to model categorical observations or data that have two outcomes or states. For example, suppose we are sampling animals and classifying them into two age classes, say either adult (an event A) or juvenile (J). If we sample a single individual and classify it, the sample space would be $S = \{A, J\}$. We could then define a probability distribution such that $P[\{A\}] = p$ and $P[\{J\}] = 1 - p$, where p is the probability of observing an adult. Then, a random variable Y equal to the **number** of adults would be a binomial random variable. The random variable Y would have a sample space $S = \{0, 1\}$ corresponding to the number of adults. We could write the probability distribution for these two events as

$$P[Y = y] = p^y(1-p)^{1-y}, \tag{5.1}$$

where $y = 0$ or 1. To see how this formula works, suppose we want the probability for $Y = 1$, so that $y = 1$. Inserting $y = 1$ in the above formula, we obtain

$$P[Y = 1] = p^1(1-p)^{1-1} = p^1(1-p)^0 = p. \tag{5.2}$$

To find the probability for $Y = 0$, we insert $y = 0$ in the formula to find

$$P[Y = 0] = p^0(1-p)^{1-0} = p^0(1-p)^1 = 1 - p. \tag{5.3}$$

Suppose that we now sample two animals and let Y again be the number of adults. The sample space for Y would now be $S = \{0, 1, 2\}$. What would be the probability distribution for this random variable? Assuming the two animals sampled are independent events, the probability of seeing two adults ($Y = 2$) in a row would be $p \times p = p^2$, while two juveniles ($Y = 0$) would be $(1-p) \times (1-p) = (1-p)^2$. There are two ways of having one adult and one juvenile, a adult first and a juvenile second, or vice versa. The probability for each is $p \times (1-p)$, so the probability of seeing one adult would be twice that, or $2p(1-p)$. A general formula describing the probability distribution for this variable would be

$$P[Y = y] = \binom{2}{y} p^y(1-p)^{2-y}. \tag{5.4}$$

where

$$\binom{2}{y} = \frac{2!}{y!(2-y)!}. \tag{5.5}$$

The quantity $\binom{2}{y}$, known as a binomial coefficient, provides a way of calculating the number of ways y adults can occur among 2 sampled animals. It is often read as '2 choose y'. It makes use of factorials, which are defined for an integer j as the product $j \times (j-1) \times (j-2)... \times 1$. For example, $4! = 4 \times 3 \times 2 \times 1$. By convention, $0! = 1$.

To see how this distribution works, we will calculate the probability for different values of y. We have

$$\begin{aligned} P[Y=0] &= \binom{2}{0} p^0 (1-p)^{2-0} = \frac{2!}{0!(2-0)!}(1-p)^2 && (5.6) \\ &= \frac{2 \times 1}{1(2 \times 1)}(1-p)^2 && (5.7) \\ &= \frac{2}{2}(1-p)^2 = (1-p)^2 && (5.8) \end{aligned}$$

and

$$\begin{aligned} P[Y=1] &= \binom{2}{1} p^1 (1-p)^{2-1} = \frac{2!}{1!(2-1)!} p(1-p) && (5.9) \\ &= \frac{2 \times 1}{1(1)} p(1-p) && (5.10) \\ &= \frac{2}{1} p(1-p) = 2p(1-p). && (5.11) \end{aligned}$$

Finally, we have

$$\begin{aligned} P[Y=2] &= \binom{2}{2} p^2 (1-p)^{2-2} = \frac{2!}{2!(2-2)!} p^2 && (5.12) \\ &= \frac{2 \times 1}{(2 \times 1)1} p^2 && (5.13) \\ &= \frac{2}{2} p^2 = p^2. && (5.14) \end{aligned}$$

Do these probabilities sum to 1, satisfying this requirement for a probability distribution? We have $(1-p)^2 + 2p(1-p) + p^2 = (1-p)(1-p) + 2p - 2p^2 + p^2 = 1 - 2p + p^2 + 2p - 2p^2 + p^2 = 1$.

Suppose that we continue to sample l different animals, and let Y be the number of adults. The sample space for this binomial random variable would be $S = \{0, 1, 2, ..., l\}$. The probability distribution for this random variable is called the **binomial distribution**, and can be written using the formula

$$P[Y = y] = f(y) = \binom{l}{y} p^y (1-p)^{l-y} \tag{5.15}$$

where $y = 0, 1, 2, ..., l$ (Mood et al. 1974). The notation $f(y)$ is often used to denote a probability distribution, which is a function of y given the parameter values.

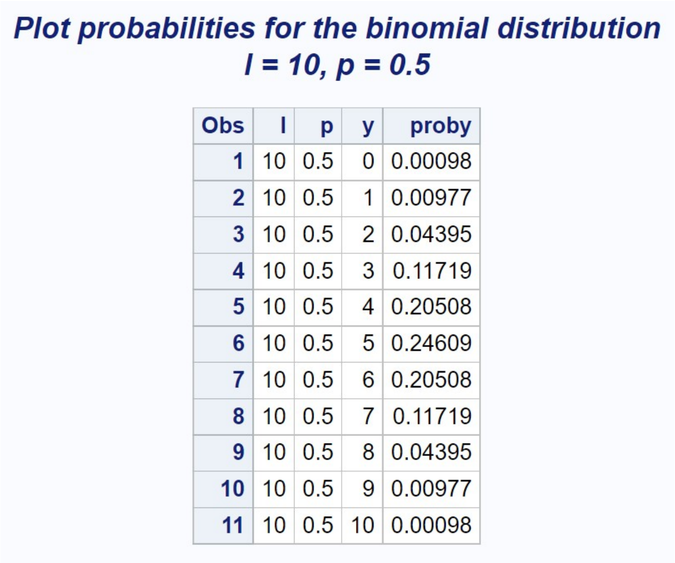
5.1.1 Binomial distribution - SAS demo

The SAS program below calculates and plots the binomial probabilities for different values of y using the SAS function `pdf`, given the values of the binomial parameters l and p. The probabilities are plotted for three different values of p, with $l = 10$. We see that for $p = 0.5$ the probability distribution has a peak at $y = 5$ (Fig. 5.2), indicating that five adults is the most likely outcome in 10 sampled animals. For $p = 0.25$ an adult occurs only 25% of the time, and so the probability distribution shifts to the left, with $y = 2$ having the highest probability (Fig. 5.3). For an adult almost certain, $p = 0.9$, then the probability distribution is shifted to the right with the peak at $y = 9$ (Fig. 5.4).

SAS Program

```
* binom_plot.sas;
title "Plot probabilities for the binomial distribution";
title2 "l = 10, p = 0.5";
data binom_plot;
    * Binomial parameters here;
    l = 10;
    p = 0.5;
    do y=0 to l;
        * Binomial distribution function;
        proby = pdf('binomial',y,p,l);
        * Output y and proby to SAS data file;
        output;
    end;
run;
* Print data;
proc print data=binom_plot;
run;
* Plot probabilities;
```

```
proc gplot data=binom_plot;
    plot proby*y=1 / vref=0 wvref=3 vaxis=axis1 haxis=axis1;
    symbol1 i=needle v=dot c=red width=3 height=2;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
quit;
```



Plot probabilities for the binomial distribution
l = 10, p = 0.5

Obs	l	p	y	proby
1	10	0.5	0	0.00098
2	10	0.5	1	0.00977
3	10	0.5	2	0.04395
4	10	0.5	3	0.11719
5	10	0.5	4	0.20508
6	10	0.5	5	0.24609
7	10	0.5	6	0.20508
8	10	0.5	7	0.11719
9	10	0.5	8	0.04395
10	10	0.5	9	0.00977
11	10	0.5	10	0.00098

Figure 5.1: binom_plot.sas - proc print

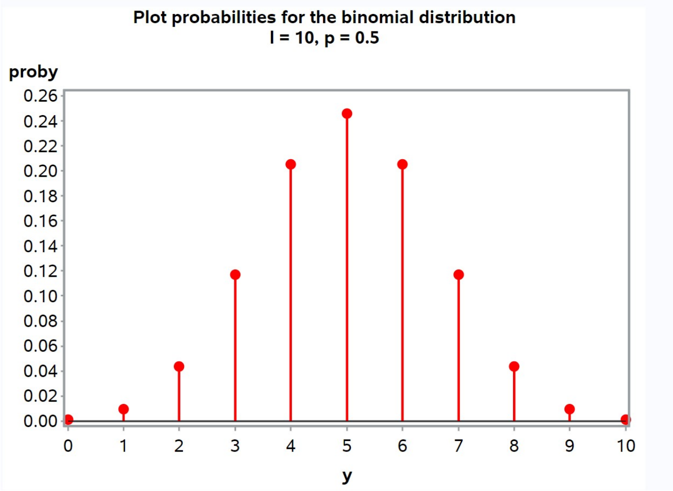


Figure 5.2: `binom_plot.sas - proc gplot`

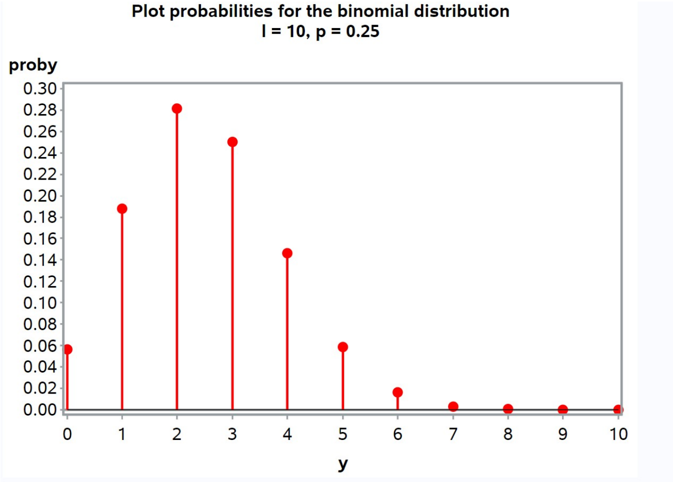


Figure 5.3: `binom_plot.sas - proc gplot`

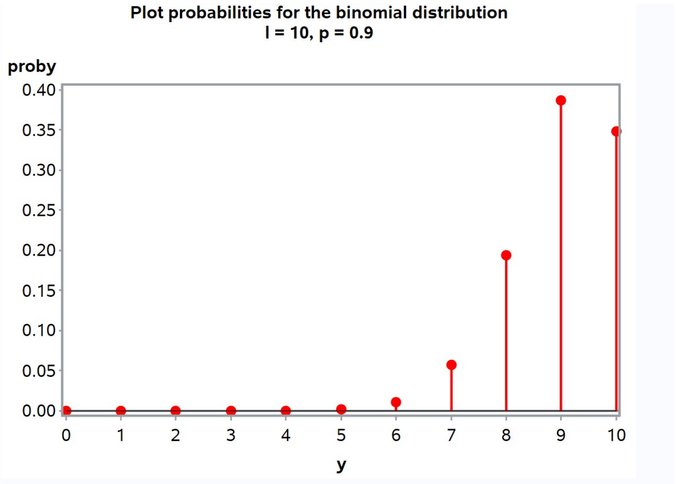


Figure 5.4: `binom_plot.sas - proc gplot`

5.2 Poisson distribution

Poisson random variables are commonly used to model counts of organisms or events in either space or time. For example, a Poisson random variable could be used to model the number of organisms in a sampling quadrat, or the number of flu infections per week in a city. The sample space for a Poisson random variable Y is $S = \{0, 1, 2, ..., \infty\}$, implying there is no upper limit on the counts. The Poisson distribution is given by the formula

$$P[Y = y] = f(y) = \frac{e^{-\lambda}\lambda^y}{y!} \tag{5.16}$$

where $y = 0, 1, 2, ..., \infty$. The parameter λ controls the shape of the distribution and is equal to the mean value of Y. For example, suppose the $\lambda = 2$. We have

$$P[Y = 0] = f(0) = \frac{e^{-2}2^0}{0!} = \frac{0.13534(1)}{1} = 0.13534, \tag{5.17}$$

$$P[Y = 1] = f(1) = \frac{e^{-2}2^1}{1!} = \frac{0.13534(2)}{1} = 0.27068, \tag{5.18}$$

$$P[Y = 2] = f(2) = \frac{e^{-2}2^2}{2!} = \frac{0.13534(4)}{2} = 0.27068, \tag{5.19}$$

$$P[Y = 3] = f(3) = \frac{e^{-2}2^3}{3!} = \frac{0.13534(8)}{6} = 0.18045, \tag{5.20}$$

$$P[Y = 4] = f(4) = \frac{e^{-2}2^4}{4!} = \frac{0.13534(16)}{24} = 0.09023 \tag{5.21}$$

and so forth.

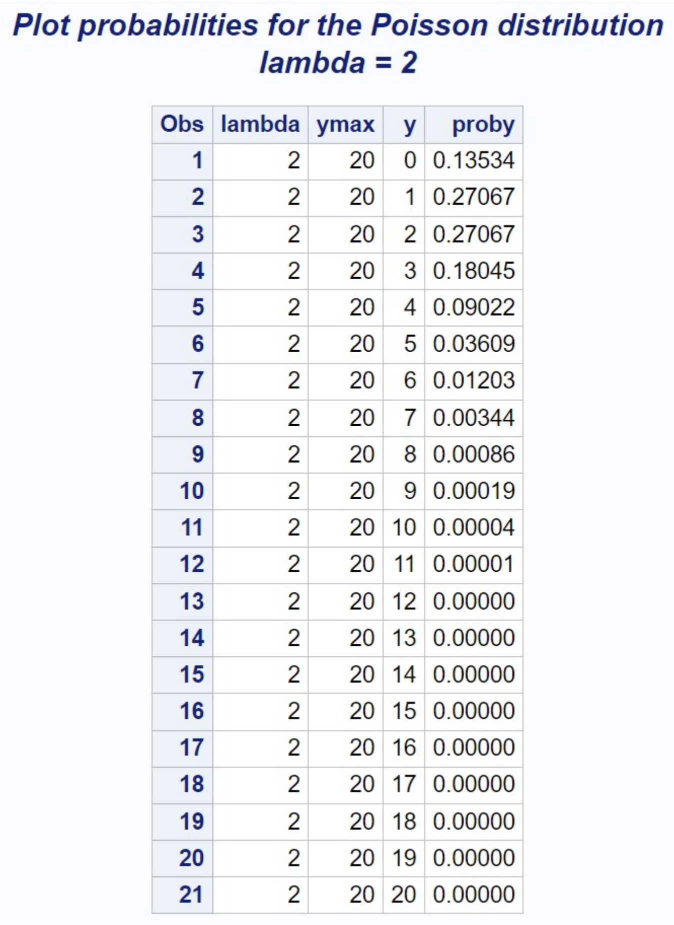
The Poisson distribution can arise in nature if certain assumptions hold true about the underlying process generating the data or observations (Mood et al. 1974, Snyder & Miller 1991). Suppose that we define an occurrence as a plant being present in a quadrat, or a case of disease occurring in a particular interval of time. **For the distribution of occurrences to be Poisson, we first need the probability of more than one occurrence to be small relative to the probability of exactly one occurrence, for a sufficiently small area of space (or short period of time).** In other words, two events are unlikely to occur in a small area or period of time. **Second, the number of occurrences in different areas of space (or time intervals) should be independent.** Another way of obtaining the Poisson distribution is as a limiting case of the binomial distribution. It can be shown that if lp is held constant (by making p small) while $l \rightarrow \infty$, the binomial distribution approaches a Poisson with $\lambda = lp$.

5.2.1 Poisson distribution - SAS demo

The following SAS program illustrates how the Poisson distribution varies for different values of λ. It is similar to the binomial distribution program, using the SAS function `pdf` to again find the probabilities (see below). We see that as λ increases, the Poisson distribution shifts to the right (Fig. 5.6, 5.7).

SAS Program

```
* Poisson_plot.sas;
title "Plot probabilities for the Poisson distribution";
title2 "lambda = 2";
data poisson_plot;
    * Poisson parameter here;
    lambda = 2;
    * Maximum value of y for plot;
    ymax = 20;
    do y=0 to ymax;
        * Poisson distribution function;
        proby = pdf('poisson',y,lambda);
        * Output y and proby to SAS data file;
        output;
    end;
run;
* Print data;
proc print data=poisson_plot;
run;
* Plot probabilities;
proc gplot data=poisson_plot;
    plot proby*y=1 / vref=0 wvref=3 vaxis=axis1 haxis=axis1;
    symbol1 i=needle v=dot c=red width=3 height=2;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
quit;
```



Plot probabilities for the Poisson distribution
lambda = 2

Obs	lambda	ymax	y	proby
1	2	20	0	0.13534
2	2	20	1	0.27067
3	2	20	2	0.27067
4	2	20	3	0.18045
5	2	20	4	0.09022
6	2	20	5	0.03609
7	2	20	6	0.01203
8	2	20	7	0.00344
9	2	20	8	0.00086
10	2	20	9	0.00019
11	2	20	10	0.00004
12	2	20	11	0.00001
13	2	20	12	0.00000
14	2	20	13	0.00000
15	2	20	14	0.00000
16	2	20	15	0.00000
17	2	20	16	0.00000
18	2	20	17	0.00000
19	2	20	18	0.00000
20	2	20	19	0.00000
21	2	20	20	0.00000

etc.

Figure 5.5: `Poisson_plot.sas - proc print`

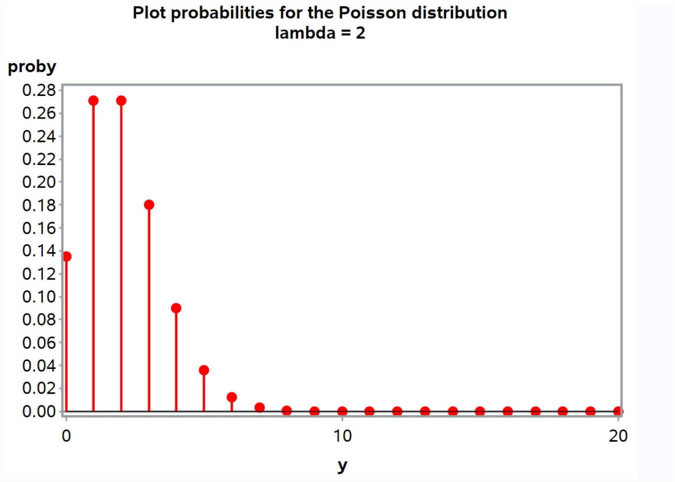


Figure 5.6: `Poisson_plot.sas - proc gplot`

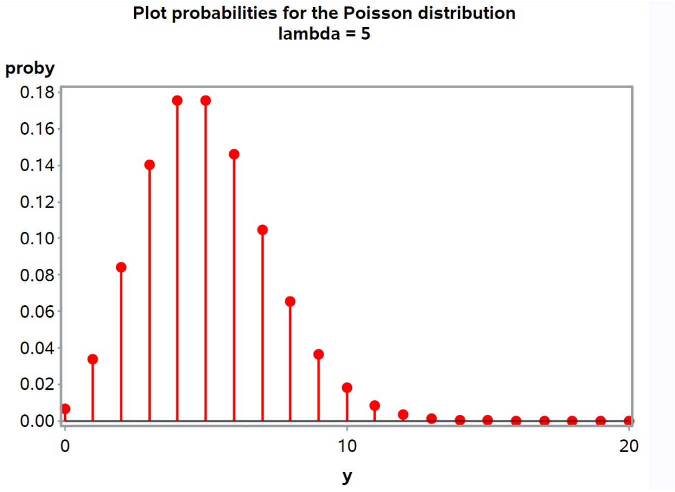


Figure 5.7: `Poisson_plot.sas - proc gplot`

5.3 Negative binomial distribution

Another useful tool for modeling count data is the negative binomial distribution. **It can be thought of as a mixture of Poisson distributions, each with a different value of λ.** For example, suppose that we are sampling insects in a forest across a number of locations. At the *ith* location the distribution of insects might be Poisson with parameter λ_i, but λ_i also differs among locations. Then the distribution of insects, considered across all locations, may have a negative binomial distribution. Because the density of most organisms typically varies in space, the negative binomial distribution often provides a better description of count data than the Poisson. The sample space for a negative binomial random variable Y is $S = \{0, 1, 2, ..., \infty\}$, the same as the Poisson. The probability distribution for the negative binomial is given by the formula

$$P[Y = y] = f(y) = \frac{\Gamma(k + y)}{\Gamma(y + 1)\Gamma(k)} \frac{(m/(k + m))^y}{(1 + m/k)^k} \tag{5.22}$$

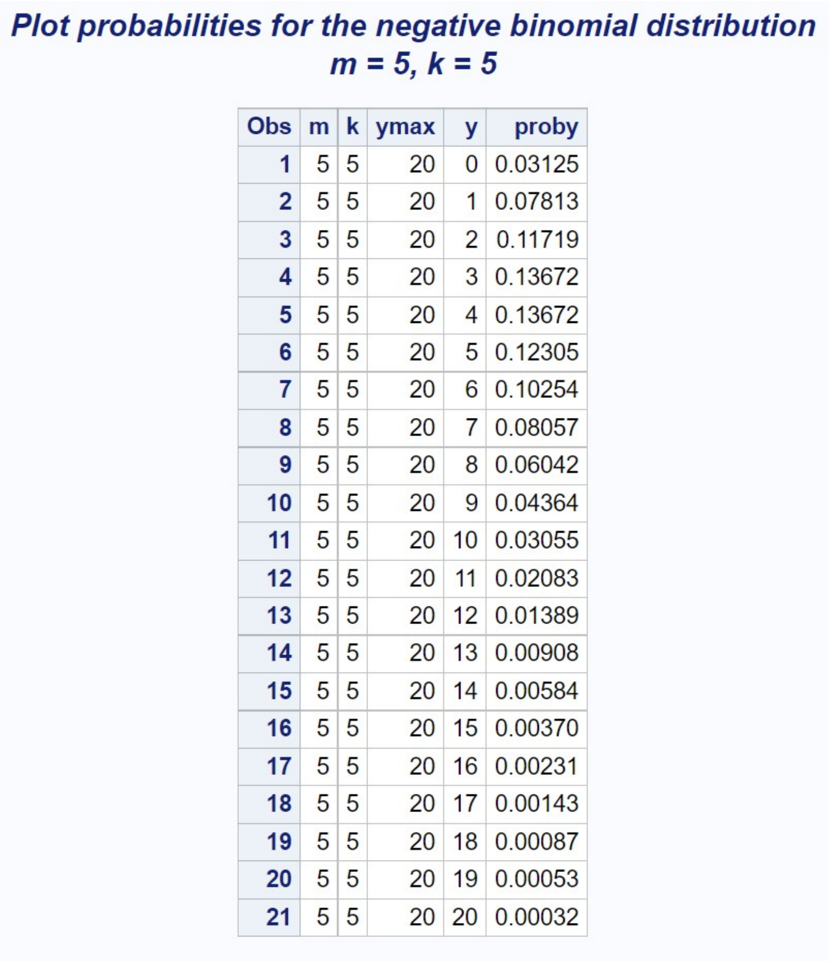
where $y = 0, 1, 2, ..., \infty$. The Γ symbol stands for the gamma function, which behaves like the factorial function but can be applied to non-integer quantities. The negative binomial distribution has two parameters, m and k, with m the mean of the distribution and k controlling its shape. For large values of k the negative binomial distribution approaches the Poisson distribution, while for small k the distribution becomes increasingly skewed to the right. See Bliss and Fisher (1953) for further information on this distribution.

5.3.1 Negative binomial distribution - SAS demo

The SAS program below shows how the shape of the negative binomial distribution varies with the parameter k. The program directly calculates the probabilities using the formula above, rather than the SAS `pdf` function, because we are using a different parameterization of the distribution. We see that distribution becomes more skewed to the right as k decreases (Fig. 5.9, 5.10).

SAS Program

```
* negbin_plot.sas;
title "Plot probabilities for the negative binomial distribution";
title2 "m = 5, k = 5";
data negbin_plot;
    * negative binomial parameters here;
    m = 5; k = 5;
    * Maximum value of y for plot;
    ymax = 20;
    do y=0 to ymax;
        * Negative binomial distribution function;
        proby = (gamma(k+y)/(gamma(y+1)*gamma(k)))*((m/(k+m))**y/(1+m/k)**k);
        * Output y and proby to SAS data file;
        output;
    end;
run;
* Print data;
proc print data=negbin_plot;
run;
* Plot probabilities;
proc gplot data=negbin_plot;
    plot proby*y=1 / vref=0 wvref=3 vaxis=axis1 haxis=axis1;
    symbol1 i=needle v=dot c=red width=3 height=2;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
quit;
```



Plot probabilities for the negative binomial distribution
***m* = 5, *k* = 5**

Obs	m	k	ymax	y	proby
1	5	5	20	0	0.03125
2	5	5	20	1	0.07813
3	5	5	20	2	0.11719
4	5	5	20	3	0.13672
5	5	5	20	4	0.13672
6	5	5	20	5	0.12305
7	5	5	20	6	0.10254
8	5	5	20	7	0.08057
9	5	5	20	8	0.06042
10	5	5	20	9	0.04364
11	5	5	20	10	0.03055
12	5	5	20	11	0.02083
13	5	5	20	12	0.01389
14	5	5	20	13	0.00908
15	5	5	20	14	0.00584
16	5	5	20	15	0.00370
17	5	5	20	16	0.00231
18	5	5	20	17	0.00143
19	5	5	20	18	0.00087
20	5	5	20	19	0.00053
21	5	5	20	20	0.00032

etc.

Figure 5.8: `negbin_plot.sas - proc print`

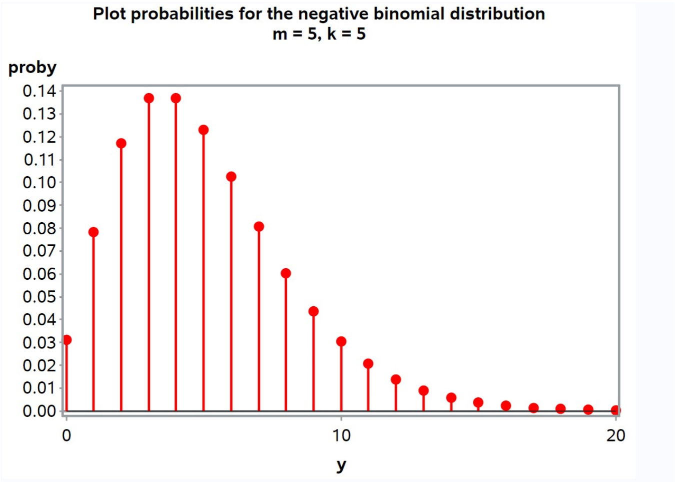


Figure 5.9: `negbin_plot.sas - proc gplot`

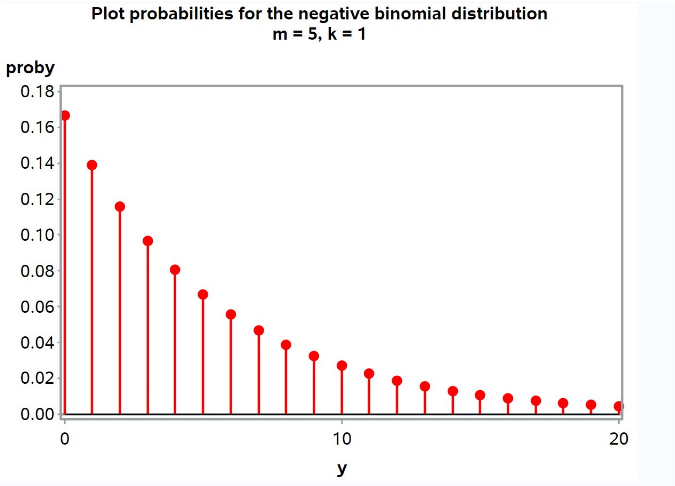


Figure 5.10: `negbin_plot.sas - proc gplot`

5.4 Expected values for discrete distributions

We have already seen how to calculate the mean, variance, and standard deviation for a set of observations (see Chapter 3). It is possible to calculate analogous quantities for probability distributions, such as the binomial, using the concept of an **expected value**.

Let Y be a random variable with some discrete probability distribution, such as the binomial, Poisson, or other distribution. The expected value or theoretical mean of Y, denoted by the expression $E[Y]$, is defined by the equation

$$E[Y] = \sum_y yP[Y = y] = \sum_y yf(y). \tag{5.23}$$

Here the summation is taken over all possible values of y for the probability distribution. **The expected value is a weighted average of each possible value of y, with the weights being the probability associated with each y.** It is a measure of the central location of the distribution of Y, in analogy to the sample mean $\bar{Y}$ for a data set. The expected value of Y can also be thought of as the sample mean $\bar{Y}$ of an infinitely large number of observations of Y.

For example, let Y have a binomial distribution with $l = 5$ and $p = 0.2$. We will first calculate some probabilities for the binomial distribution, then use them to calculate the expected value of Y, or $E[Y]$. We have

$$\begin{aligned}
P[Y = 0] = f(0) &= \binom{5}{0} 0.2^0 (1 - 0.2)^{5-0} && (5.24)\\
&= \frac{5!}{0!(5-0)!} 1(0.8^5) && (5.25)\\
&= \frac{120}{1(120)} 0.32768 && (5.26)\\
&= 0.32768. && (5.27)
\end{aligned}$$

$$P[Y=1]=f(1)=\binom{5}{1}0.2^1(1-0.2)^{5-1} \tag{5.28}$$

$$=\frac{5!}{1!(5-1)!}0.2(0.8^4) \tag{5.29}$$

$$=\frac{120}{1(24)}0.08192 \tag{5.30}$$

$$=0.40960. \tag{5.31}$$

$$P[Y=2]=f(2)=\binom{5}{2}0.2^2(1-0.2)^{5-2} \tag{5.32}$$

$$=\frac{5!}{2!(5-2)!}0.04(0.8^3) \tag{5.33}$$

$$=\frac{120}{2(6)}0.02048 \tag{5.34}$$

$$=0.20480. \tag{5.35}$$

$$P[Y=3]=f(3)=\binom{5}{3}0.2^3(1-0.2)^{5-3} \tag{5.36}$$

$$=\frac{5!}{2!(5-2)!}0.008(0.8^2) \tag{5.37}$$

$$=\frac{120}{2(6)}0.00512 \tag{5.38}$$

$$=0.05120. \tag{5.39}$$

$$P[Y=4]=f(4)=\binom{5}{4}0.2^4(1-0.2)^{5-4} \tag{5.40}$$

$$=\frac{5!}{4!(5-4)!}0.0016(0.8^1) \tag{5.41}$$

$$=\frac{120}{24(1)}0.00128 \tag{5.42}$$

$$=0.00640. \tag{5.43}$$

$$
\begin{aligned}
P[Y=5] = f(5) &= \binom{5}{5} 0.2^5 (1-0.2)^{5-5} && (5.44)\\
&= \frac{5!}{5!(5-5)!} 0.00032(0.8^0) && (5.45)\\
&= \frac{120}{120(1)} 0.00032 && (5.46)\\
&= 0.00032. && (5.47)
\end{aligned}
$$

These probabilities sum to 1, indicating our calculations are correct. Alternately, we could use the SAS program `binom_plot.sas` to find these probabilities.

We will now calculate $E[Y]$ using these probabilities and the formula for $E[Y]$ given above. We have

$$
\begin{aligned}
E[Y] = \sum_y y f(y) &= 0(0.32768) + 1(0.40960) + 2(0.20480) && (5.48)\\
&+ 3(0.05120) + 4(0.00640) + 5(0.00032) && (5.49)\\
&= 0 + 0.40960 + 0.40960 && (5.50)\\
&+ 0.15360 + 0.02560 + 0.00160 && (5.51)\\
&= 1.00000 && (5.52)
\end{aligned}
$$

So, $E[Y] = 1$ for the binomial distribution with $l = 5$ and $p = 0.2$.

For the binomial distribution in general, it can be shown that

$$E[Y] = lp \tag{5.53}$$

for any value of l and p. Thus, the expected value or theoretical mean for the binomial distribution can be easily calculated given the parameters of this distribution. Plugging $l = 5$ and $p = 0.2$ into this equation, we obtain $E[Y] = 5 \times 0.2 = 1.0$, the same value as obtained using the expected value formula.

Other probability distributions would have a different formula for the expected value or theoretical mean, but the formula always involves the parameters of the distribution. For the Poisson distribution it can be shown that $E[Y] = \lambda$, while for the negative binomial distribution $E[Y] = m$.

5.4.1 Variance for discrete distributions

We can also define the theoretical variance for a random variable Y using expected values. This variance measures the dispersion of Y, and can also be thought of as the sample variance s^2 of an infinite number of observations. The variance of a discrete random variable Y, denoted by $Var[Y]$, is defined as

$$\begin{aligned} Var[Y] = E[(Y - E[Y])^2] &= \sum_y (y - E[Y])^2 P[Y = y] && (5.54) \\ &= \sum_y (y - E[Y])^2 f(y). && (5.55) \end{aligned}$$

Note that this formula makes use of $E[Y]$, so it must be calculated first. As an example, let Y have the same binomial distribution as before, with $l = 5$ and $p = 0.2$, for which $E[Y] = 1$. Using the probabilities calculated above, we have

$$\begin{aligned} Var[Y] &= \sum_y (y - E[Y])^2 f(y) && (5.56) \\ &= (0-1)^2(0.32768) + (1-1)^2(0.40960) + (2-1)^2(0.20480) && (5.57) \\ &+ (3-1)^2(0.05120) + (4-1)^2(0.00640) + (5-1)^2(0.00032) && (5.58) \\ &= 1(0.32768) + 0(0.40960) + (1)0.20480 && (5.59) \\ &+ 4(0.05120) + 9(0.00640) + (16)0.00032 && (5.60) \\ &= 0.32768 + 0 + 0.20480 + 0.20480 + 0.05760 + 0.00512 && (5.61) \\ &= 0.8. && (5.62) \end{aligned}$$

For the binomial distribution, it can be mathematically shown that for any value of l and p, we have

$$Var[Y] = lp(1 - p). \tag{5.63}$$

Thus, the theoretical variance for the binomial distribution can also be calculated using the parameters of this distribution. Plugging $l = 5$ and $p = 0.2$ into this equation, we obtain $Var[Y] = 5(0.2)(1 - 0.2) = 0.8$, the same value as obtained using the variance formula.

Other probability distributions would have a different formula for the theoretical variance. For the Poisson distribution it can be shown that $Var[Y] = \lambda$. Because $E[Y] = \lambda$ for the Poisson, this implies the mean and variance of a Poisson random variable are equal. For the negative binomial distribution, $Var[Y] = m + m^2/k$, while $E[Y] = m$. This implies the variance of the negative binomial is always greater than its mean. The theoretical standard deviation is simply $\sqrt{Var[Y]}$.

5.5 Discrete random variables and samples

Discrete random variables like the binomial and Poisson are used to model real observations that are counts. But how well do these mathematical quantities match the behavior of the observations? We will now develop a graphical method of comparing the observed data with the pattern expected for discrete random variables, in particular the Poisson and negative binomial distributions. There are also statistical procedures called goodness-of-fit tests that are used for this purpose, but we defer this to Chapter 20.

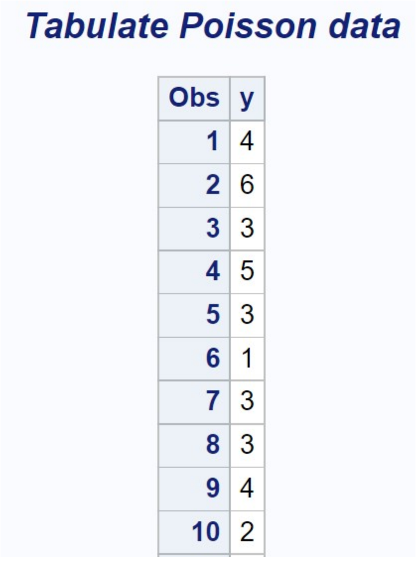
5.5.1 Parasitic wasps - SAS demo

Small insects are often sampled using sticky-traps, which are small cards covered with a substance called Tanglefoot®(The Tanglefoot Company, Grand Rapids, MI). For example, Reeve & Cronin (2010) used this method to sample populations of the parasitic wasp *Anagrus columbi*, which attacks eggs of the planthopper *Prokelisia crocea*. Suppose $n = 100$ traps are deployed for some period of time, then the traps collected and the wasps counted. If individual wasps are randomly and independently distributed across the field, we would expect the number of wasps per trap to have a Poisson distribution. We can then compare the observed distribution with the expected one for the Poisson distribution, to see if they resemble one another. If so, we can say the Poisson distribution provides an adequate description of these observations.

The first step in this procedure is simply to tabulate the number of traps with $0, 1, 2, 3, ...$ wasps, which is the observed frequency distribution. We can use `proc freq` in SAS to accomplish this task as in the following program. The numbers listed as data here are the number of wasps for each of the $n = 100$ sticky-traps. The statement `tables y` tells `proc freq` to count the number of observations for each value of `y` in the data set. The output generated is a table of these frequencies.

SAS Program

```
* poisson_freq.sas;
title 'Tabulate Poisson data';
data poisson;
    input y @@;
    datalines;
4 6 3 5 3 1 3 3 4 2
4 0 2 3 1 3 4 6 5 1
3 3 4 3 2 3 7 4 3 3
4 3 4 3 4 0 3 0 3 3
4 8 2 2 4 2 5 3 3 2
1 4 1 1 5 2 4 1 2 6
3 3 3 1 1 2 1 5 3 5
3 2 4 3 4 1 2 3 1 3
4 4 4 6 6 2 0 1 4 2
2 2 3 4 3 0 1 1 0 2
;
run;
* Print observations;
proc print data=poisson;
run;
* Tabulate data into frequencies;
proc freq data=poisson;
    tables y;
run;
quit;
```

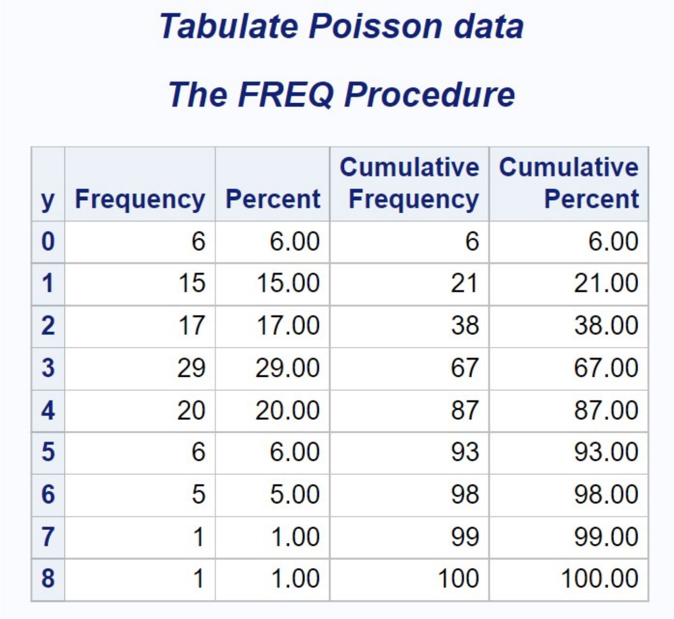


Tabulate Poisson data

Obs	y
1	4
2	6
3	3
4	5
5	3
6	1
7	3
8	3
9	4
10	2

etc.

Figure 5.11: `Poisson_freq.sas - proc print`



Tabulate Poisson data

The FREQ Procedure

y	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	6	6.00	6	6.00
1	15	15.00	21	21.00
2	17	17.00	38	38.00
3	29	29.00	67	67.00
4	20	20.00	87	87.00
5	6	6.00	93	93.00
6	5	5.00	98	98.00
7	1	1.00	99	99.00
8	1	1.00	100	100.00

Figure 5.12: `Poisson_freq.sas - proc freq`

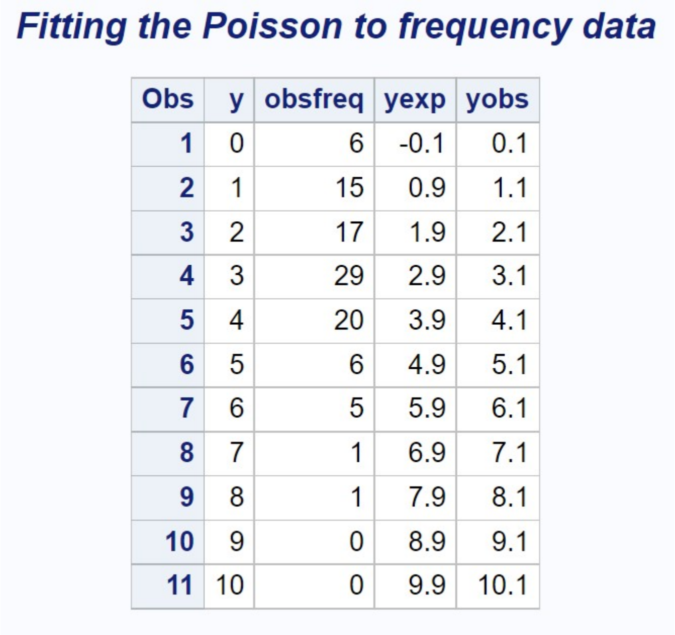
We now want to compare these observed frequencies with those expected for the Poisson distribution. We first need to estimate the Poisson parameter λ from the observed data using $\bar{Y}$ (see Chapter 8 for a justification). We then calculate the Poisson probabilities for $\lambda = \bar{Y}$, obtaining $P[Y = y]$ for values of y that spans or better exceeds the range of y values in the data set. Because $P[Y = y]$ is the probability or proportion of observations that take the value y, the expected frequency with n observations is therefore equal to $n \times P[Y = y]$. We can then compare the observed frequencies with the expected ones generated using the Poisson distribution. These calculations can be automated using the SAS program listed below. The program first uses `proc univariate` to find n, $\bar{Y}$, and the sample variance s^2 for the observed frequencies. We let `proc univariate` know that the data are in the form of frequencies (the variable `obsfreq`), rather than individual observations, by adding the command `freq obsfreq`.

The program then passes these results to a `data` step where the Poisson probabilities and expected frequencies are calculated, which are then plotted across a range of y values using `proc gplot`. See SAS output and graph below. We first see that sample mean and variance are similar in magnitude ($\bar{Y} = 2.910$ vs. $s^2 = 2.628$), suggesting these data are close to Poisson (recall that $E[Y] = Var[Y] = \lambda$ for this distribution). In addition, the observed and expected frequencies are quite similar, again implying an adequate fit by the Poisson distribution. There are some small differences in the observed and expected frequencies, which is to be expected because the observed ones are random quantities.

SAS Program

```
* Poisson_fit.sas;
title 'Fitting the Poisson to frequency data';
data poisson;
    input y obsfreq;
    * Generate offset y values for plot;
    yexp = y - 0.1; yobs = y + 0.1;
    datalines;
0   6
1  15
2  17
3  29
4  20
5   6
6   5
```

```
7   1
8   1
9   0
10  0
;
run;
* Print data set;
proc print data=poisson;
run;
* Descriptive statistics, save ybar, n, and var to data file;
proc univariate data=poisson;
    var y;
    histogram y / vscale=count;
    freq obsfreq;
    output out=stats mean=ybar n=n var=var;
run;
* Print output data file;
proc print data=stats;
run;
* Calculate expected frequencies using ybar;
data poisfit;
    if _n_ = 1 then set stats;
    set poisson;
    poisprob = pdf('poisson',y,ybar);
    expfreq = n*poisprob;
run;
* Print observed and expected frequencies;
proc print data=poisfit;
run;
* Plot observed and expected frequencies;
proc gplot data=poisfit;
    plot expfreq*yexp=1 obsfreq*yobs=2 / overlay legend=legend1 vref=0 wvref=3
    vaxis=axis1 haxis=axis1;
    symbol1 i=needle v=circle c=red width=3 height=2;
    symbol2 i=needle v=square c=blue width=3 height=2;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
    legend1 label=(height=2) value=(height=2);
run;
quit;
```

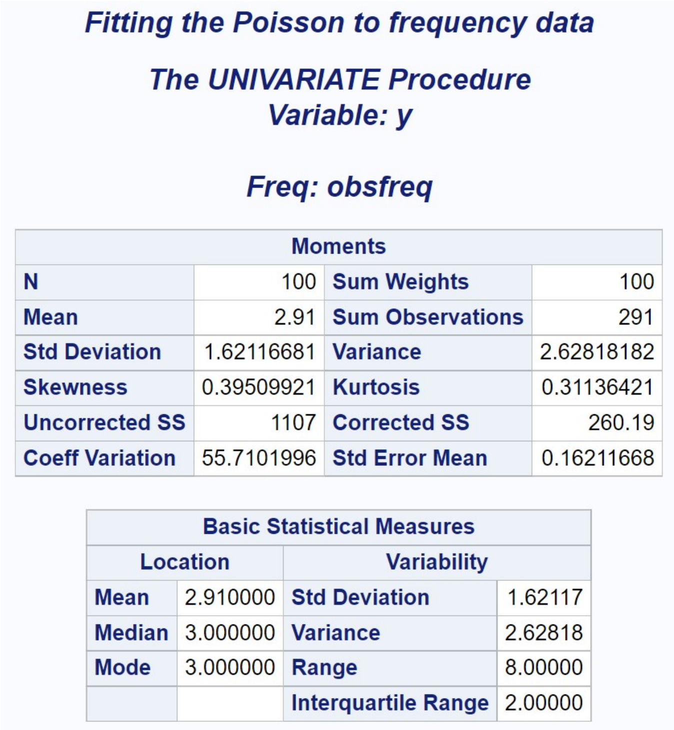


Fitting the Poisson to frequency data

Obs	y	obsfreq	yexp	yobs
1	0	6	-0.1	0.1
2	1	15	0.9	1.1
3	2	17	1.9	2.1
4	3	29	2.9	3.1
5	4	20	3.9	4.1
6	5	6	4.9	5.1
7	6	5	5.9	6.1
8	7	1	6.9	7.1
9	8	1	7.9	8.1
10	9	0	8.9	9.1
11	10	0	9.9	10.1

etc.

Figure 5.13: `Poisson_fit.sas - proc print`



Fitting the Poisson to frequency data

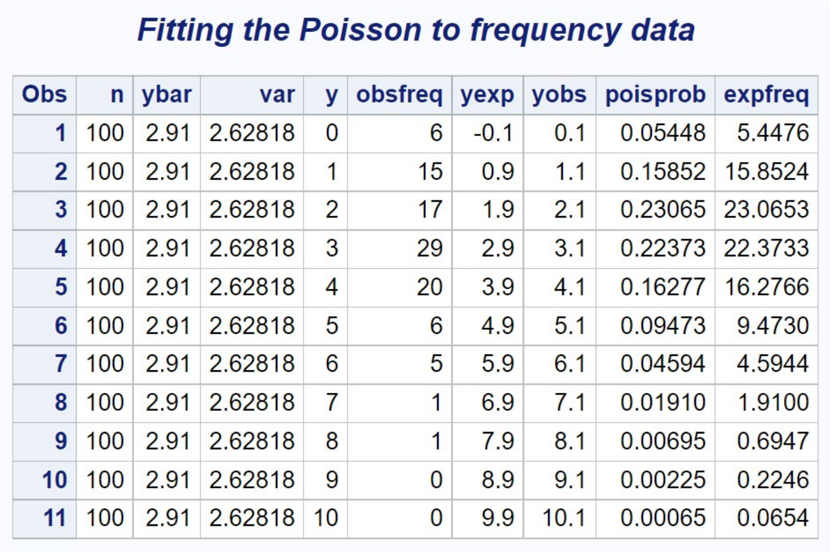
The UNIVARIATE Procedure
Variable: y

Freq: obsfreq

Moments			
N	100	Sum Weights	100
Mean	2.91	Sum Observations	291
Std Deviation	1.62116681	Variance	2.62818182
Skewness	0.39509921	Kurtosis	0.31136421
Uncorrected SS	1107	Corrected SS	260.19
Coeff Variation	55.7101996	Std Error Mean	0.16211668

Basic Statistical Measures			
Location		Variability	
Mean	2.910000	Std Deviation	1.62117
Median	3.000000	Variance	2.62818
Mode	3.000000	Range	8.00000
		Interquartile Range	2.00000

Figure 5.14: `Poisson_fit.sas - proc univariate`



Fitting the Poisson to frequency data

Obs	n	ybar	var	y	obsfreq	yexp	yobs	poisprob	expfreq
1	100	2.91	2.62818	0	6	-0.1	0.1	0.05448	5.4476
2	100	2.91	2.62818	1	15	0.9	1.1	0.15852	15.8524
3	100	2.91	2.62818	2	17	1.9	2.1	0.23065	23.0653
4	100	2.91	2.62818	3	29	2.9	3.1	0.22373	22.3733
5	100	2.91	2.62818	4	20	3.9	4.1	0.16277	16.2766
6	100	2.91	2.62818	5	6	4.9	5.1	0.09473	9.4730
7	100	2.91	2.62818	6	5	5.9	6.1	0.04594	4.5944
8	100	2.91	2.62818	7	1	6.9	7.1	0.01910	1.9100
9	100	2.91	2.62818	8	1	7.9	8.1	0.00695	0.6947
10	100	2.91	2.62818	9	0	8.9	9.1	0.00225	0.2246
11	100	2.91	2.62818	10	0	9.9	10.1	0.00065	0.0654

Figure 5.15: `Poisson_fit.sas - proc print`

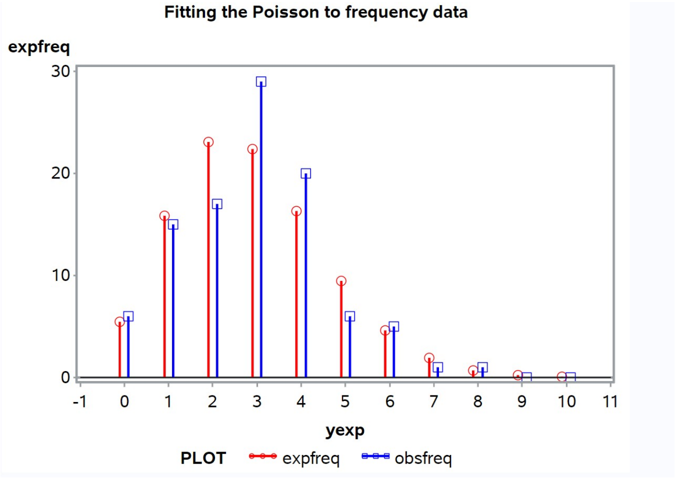


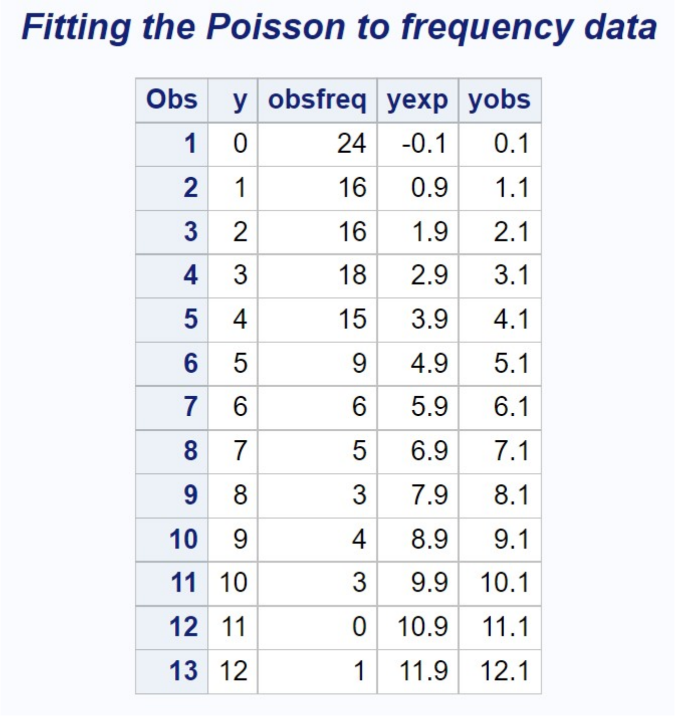
Figure 5.16: `Poissonfit.sas - proc gplot`

5.5.2 Corn borers - SAS demo

We now examine the spatial distribution of an insect pest, the European corn borer *Ostrinia nubilalis*, as reported by Bliss and Fisher (1953). The number of borers was recorded for 120 hills in which corn was planted. These data are already tabulated and can be directly inserted in the SAS program `poisson_fit2.sas` (see below). For this example, we see that the Poisson distribution provides a relatively poor fit (see Fig. 5.20) - there are more zeroes ($y = 0$) and large values ($y \geq 7$) in the observed frequencies than predicted by the Poisson. We also note that the sample variance $s^2 = 7.770$ is considerably larger than the mean $\bar{Y} = 3.167$, while for the Poisson these two quantities should be equal. This finding also suggests that these data are not Poisson in distribution.

SAS Program

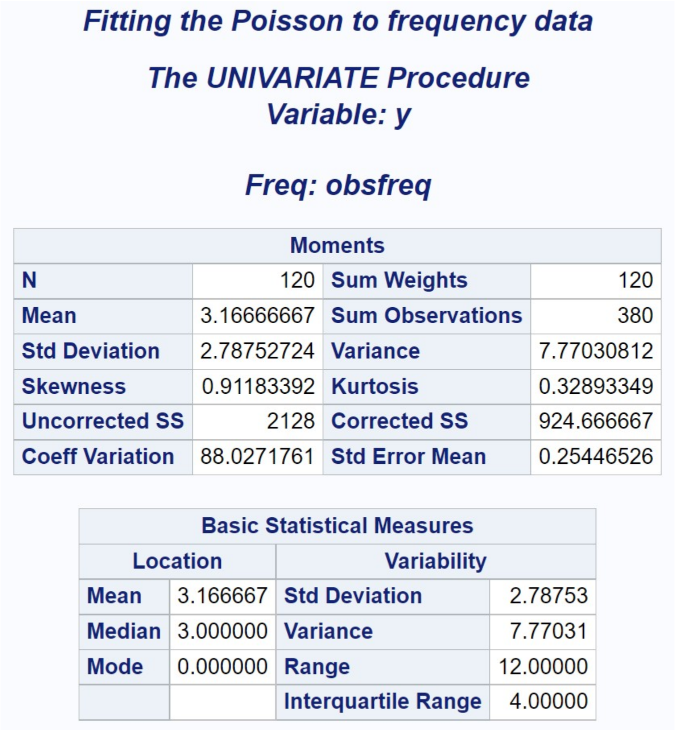
```
* Poisson_fit2.sas;
title 'Fitting the Poisson to frequency data';
data poisson;
    input y obsfreq;
    * Generate offset y values for plot;
    yexp = y - 0.1; yobs = y + 0.1;
    datalines;
0  24
1  16
2  16
3  18
4  15
5   9
6   6
7   5
8   3
9   4
10  3
11  0
12  1
;
run;
* Print data set;
proc print data=poisson;
run;
* Descriptive statistics, save ybar, n, and var to data file;
proc univariate data=poisson;
    var y;
    histogram y / vscale=count;
    freq obsfreq;
    output out=stats mean=ybar n=n var=var;
run;
* Print output data file;
proc print data=stats;
run;
* Calculate expected frequencies using ybar;
data poisfit;
    if _n_ = 1 then set stats;
    set poisson;
    poisprob = pdf('poisson',y,ybar);
    expfreq = n*poisprob;
run;
* Print observed and expected frequencies;
proc print data=poisfit;
run;
* Plot observed and expected frequencies;
proc gplot data=poisfit;
    plot expfreq*yexp=1 obsfreq*yobs=2 / overlay legend=legend1 vref=0 wvref=3
    vaxis=axis1 haxis=axis1;
    symbol1 i=needle v=circle c=red width=3 height=2;
    symbol2 i=needle v=square c=blue width=3 height=2;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
    legend1 label=(height=2) value=(height=2);
run;
quit;
```



Fitting the Poisson to frequency data

Obs	y	obsfreq	yexp	yobs
1	0	24	-0.1	0.1
2	1	16	0.9	1.1
3	2	16	1.9	2.1
4	3	18	2.9	3.1
5	4	15	3.9	4.1
6	5	9	4.9	5.1
7	6	6	5.9	6.1
8	7	5	6.9	7.1
9	8	3	7.9	8.1
10	9	4	8.9	9.1
11	10	3	9.9	10.1
12	11	0	10.9	11.1
13	12	1	11.9	12.1

Figure 5.17: `Poisson_fit2.sas - proc print`



Fitting the Poisson to frequency data

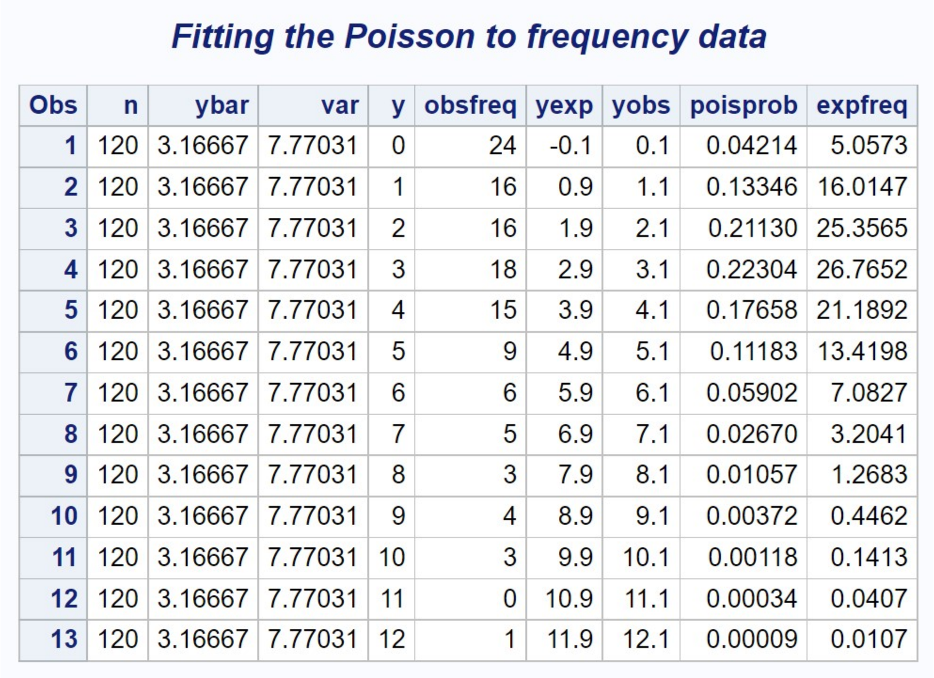
The UNIVARIATE Procedure
Variable: y

Freq: obsfreq

Moments			
N	120	Sum Weights	120
Mean	3.16666667	Sum Observations	380
Std Deviation	2.78752724	Variance	7.77030812
Skewness	0.91183392	Kurtosis	0.32893349
Uncorrected SS	2128	Corrected SS	924.666667
Coeff Variation	88.0271761	Std Error Mean	0.25446526

Basic Statistical Measures			
Location		Variability	
Mean	3.166667	Std Deviation	2.78753
Median	3.000000	Variance	7.77031
Mode	0.000000	Range	12.00000
		Interquartile Range	4.00000

Figure 5.18: `Poisson_fit2.sas - proc univariate`



Fitting the Poisson to frequency data

Obs	n	ybar	var	y	obsfreq	yexp	yobs	poisprob	expfreq
1	120	3.16667	7.77031	0	24	-0.1	0.1	0.04214	5.0573
2	120	3.16667	7.77031	1	16	0.9	1.1	0.13346	16.0147
3	120	3.16667	7.77031	2	16	1.9	2.1	0.21130	25.3565
4	120	3.16667	7.77031	3	18	2.9	3.1	0.22304	26.7652
5	120	3.16667	7.77031	4	15	3.9	4.1	0.17658	21.1892
6	120	3.16667	7.77031	5	9	4.9	5.1	0.11183	13.4198
7	120	3.16667	7.77031	6	6	5.9	6.1	0.05902	7.0827
8	120	3.16667	7.77031	7	5	6.9	7.1	0.02670	3.2041
9	120	3.16667	7.77031	8	3	7.9	8.1	0.01057	1.2683
10	120	3.16667	7.77031	9	4	8.9	9.1	0.00372	0.4462
11	120	3.16667	7.77031	10	3	9.9	10.1	0.00118	0.1413
12	120	3.16667	7.77031	11	0	10.9	11.1	0.00034	0.0407
13	120	3.16667	7.77031	12	1	11.9	12.1	0.00009	0.0107

Figure 5.19: `Poisson_fit2.sas - proc print`

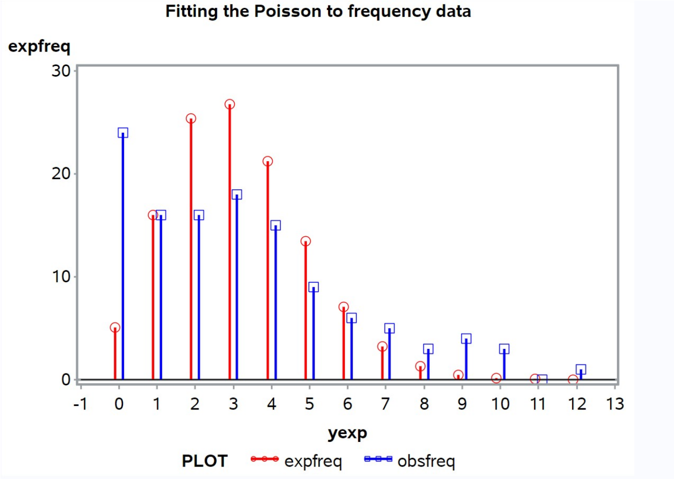


Figure 5.20: `Poissonfit2.sas - proc gplot`

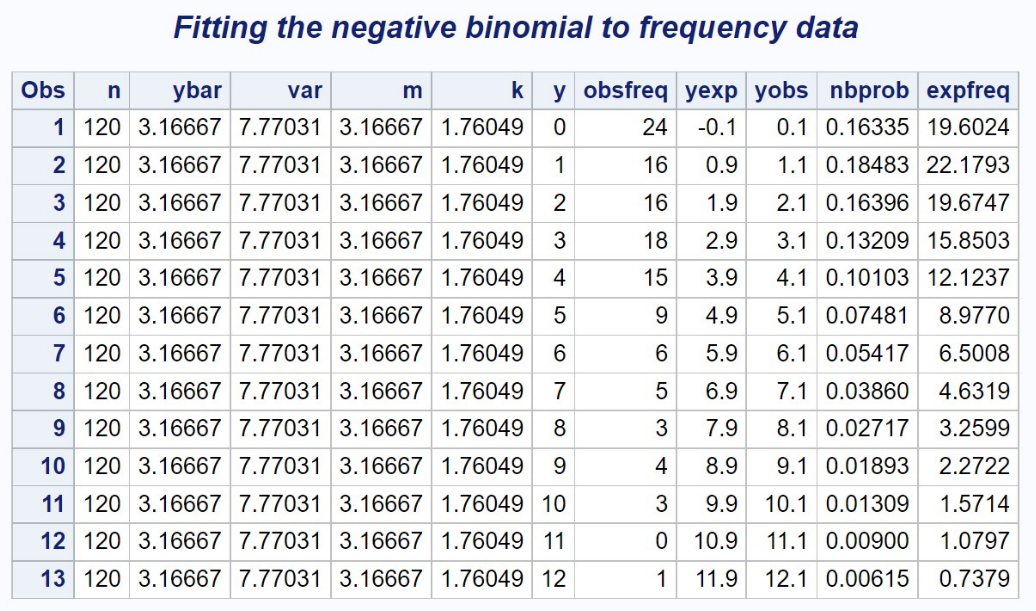
As an alternative to the Poisson, we can try fitting the negative binomial distribution using a similar SAS program. This distribution has two parameters, m and k, that must also be estimated before we can fit the distribution. The parameter m can be estimated using $\bar{Y}$ as with the Poisson, but k is best estimated using a technique called maximum likelihood (see Chapter 8). We will use a SAS procedure that can model count data using the negative binomial distribution, `proc genmod`, in order to estimate k (SAS Institute Inc. 2018). The output of `proc genmod` is manipulated in several `data` steps to combine these estimates with the observed frequency data, and then the negative binomial probabilities and expected frequencies calculated and plotted. See SAS program and output below.

We see that the expected frequencies for the negative binomial distribution provide a better match to the observed ones for this data set (Fig. 5.22). We also note that the variance predicted for the negative binomial distribution is close to the observed variance. From the negative binomial fit, we have $m = 3.167$ and $k = 1.760$, and so the estimated variance is $m + m^2/k = 3.167 + 3.167^2/1.760 = 7.459$, while the observed variance is $s^2 = 7.770$. This further implies the negative binomial provides a better fit to these data than the Poisson distribution.

SAS Program

```
* negbin_fit2.sas;
title 'Fitting the negative binomial to frequency data';
data negbin;
    input y obsfreq;
    * Generate offset y values for plot;
    yexp = y - 0.1; yobs = y + 0.1;
    datalines;
0  24
1  16
2  16
3  18
4  15
5   9
6   6
7   5
8   3
9   4
10  3
11  0
12  1
;
run;
* Print data set;
proc print data=negbin;
run;
* Descriptive statistics, save ybar, n, and var to data file;
proc univariate data=negbin;
    var y;
    histogram y / vscale=count;
    freq obsfreq;
    output out=stats mean=ybar n=n var=var;
run;
* Print output data file;
proc print data=stats;
run;
* Estimate m and k for the negative binomial distribution;
proc genmod data=negbin;
    model y = / dist=negbin;
    freq obsfreq;
    ods output ParameterEstimates=params;
run;
* Pick out value of m from genmod output;
data m;
```

```
    set params;
    if _n_ = 1;
    m = exp(Estimate);
    keep m;
run;
* Pick out value of k from genmod output;
data k;
    set params;
    if _n_ = 2;
    k = 1/Estimate;
    keep k;
run;
* Put m and k in one data file;
data params;
    merge m k;
run;
* Calculate expected frequencies using m and k;
data nbfit;
    if _n_ = 1 then set stats;
    if _n_ = 1 then set params;
    set negbin;
    nbprob = (gamma(k+y)/(gamma(y+1)*gamma(k)))*((m/(k+m))**y/(1+m/k)**k);
    expfreq = n*nbprob;
run;
* Print observed and expected frequencies;
proc print data=nbfit;
run;
* Plot observed and expected frequencies;
proc gplot data=nbfit;
    plot expfreq*yexp=1 obsfreq*yobs=2 / overlay legend=legend1 vref=0 wvref=3
    vaxis=axis1 haxis=axis1;
    symbol1 i=needle v=circle c=red width=3 height=2;
    symbol2 i=needle v=square c=blue width=3 height=2;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
    legend1 label=(height=2) value=(height=2);
run;
quit;
```



Fitting the negative binomial to frequency data

Obs	n	ybar	var	m	k	y	obsfreq	yexp	yobs	nbprob	expfreq
1	120	3.16667	7.77031	3.16667	1.76049	0	24	-0.1	0.1	0.16335	19.6024
2	120	3.16667	7.77031	3.16667	1.76049	1	16	0.9	1.1	0.18483	22.1793
3	120	3.16667	7.77031	3.16667	1.76049	2	16	1.9	2.1	0.16396	19.6747
4	120	3.16667	7.77031	3.16667	1.76049	3	18	2.9	3.1	0.13209	15.8503
5	120	3.16667	7.77031	3.16667	1.76049	4	15	3.9	4.1	0.10103	12.1237
6	120	3.16667	7.77031	3.16667	1.76049	5	9	4.9	5.1	0.07481	8.9770
7	120	3.16667	7.77031	3.16667	1.76049	6	6	5.9	6.1	0.05417	6.5008
8	120	3.16667	7.77031	3.16667	1.76049	7	5	6.9	7.1	0.03860	4.6319
9	120	3.16667	7.77031	3.16667	1.76049	8	3	7.9	8.1	0.02717	3.2599
10	120	3.16667	7.77031	3.16667	1.76049	9	4	8.9	9.1	0.01893	2.2722
11	120	3.16667	7.77031	3.16667	1.76049	10	3	9.9	10.1	0.01309	1.5714
12	120	3.16667	7.77031	3.16667	1.76049	11	0	10.9	11.1	0.00900	1.0797
13	120	3.16667	7.77031	3.16667	1.76049	12	1	11.9	12.1	0.00615	0.7379

Figure 5.21: `negbin_fit2.sas - proc print`

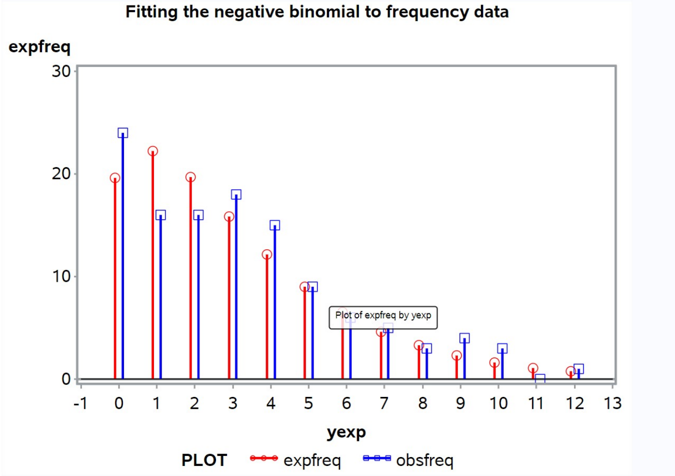


Figure 5.22: `negbin_fit2.sas - proc gplot`

5.6 Classifying spatial or temporal patterns

The spatial distribution of organisms, or the temporal occurrence of events like cases of disease, is often compared with the Poisson distribution. This distribution essentially assumes a random, independent distribution of organisms or events, and if the observed distribution differs from the Poisson then this could indicate some interesting biology. For example, if the observed frequencies have a distribution with more extreme values (low or high) than the Poisson, with $s^2 > \bar{Y}$, this implies organisms are unevenly distributed in space, or events in time. A pattern like this is often called an **overdispersed** distribution, or alternatively a clumped, aggregated, or contagious distribution (Pielou 1977, Begon et al. 2006). One method of quantifying the level of overdispersion is to fit the negative binomial distribution to the data and use the value of k as an index. Small values of k (say $k < 5$) imply an overdispersed distribution, while larger ones indicate a distribution close to Poisson. More rarely, an observed distribution may have fewer extreme values than the Poisson, with $s^2 < \bar{Y}$, implying the organisms are evenly distributed in space (or events in time). This is called an **underdispersed distribution**, also known as a regular, even, or repulsed distribution.

The figures below provide examples of spatial distributions that are overdispersed, Poisson, or underdispersed. Note the obvious clusters of organisms in the overdispersed example (Fig. 5.23). This might occur because the clusters are offspring from a single parent, the organisms are responding to resources that are clumped in space, or because the organisms are attracted to one another. The Poisson data also show a few clusters (Fig. 5.24), but these are chance occurrences. If we were to divide this graph into quadrats and count the number of organisms per quadrat, we would find the frequency distribution is close to Poisson. In contrast to the other examples, the organisms are spaced apart to some extent in the underdispersed example (Fig. 5.25). This could occur because they are territorial, compete for resources, or otherwise regulate their numbers in some fashion (Ridout & Besbeas 2004).

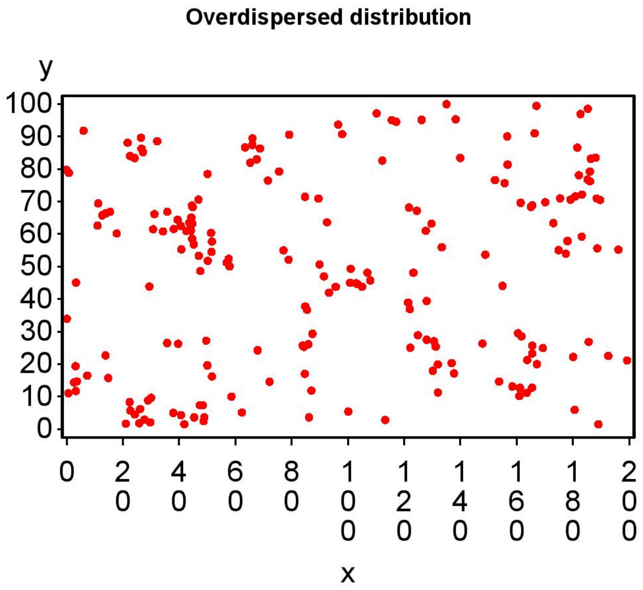


Figure 5.23: Overdispersed distribution of organisms in space

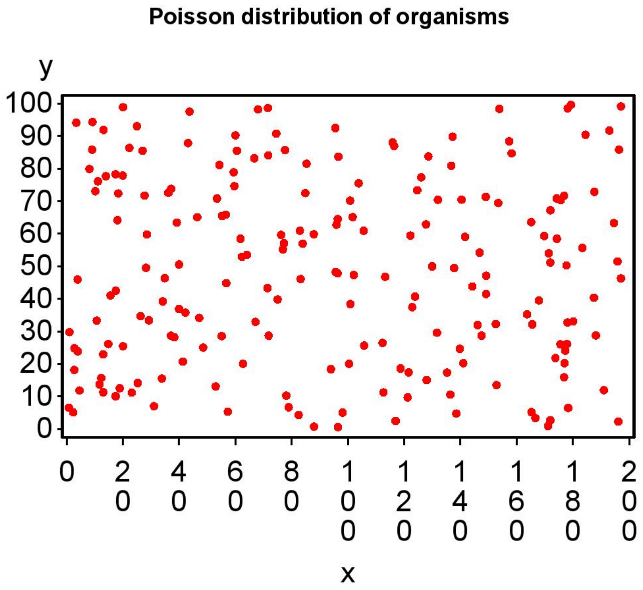


Figure 5.24: Poisson distribution of organisms in space

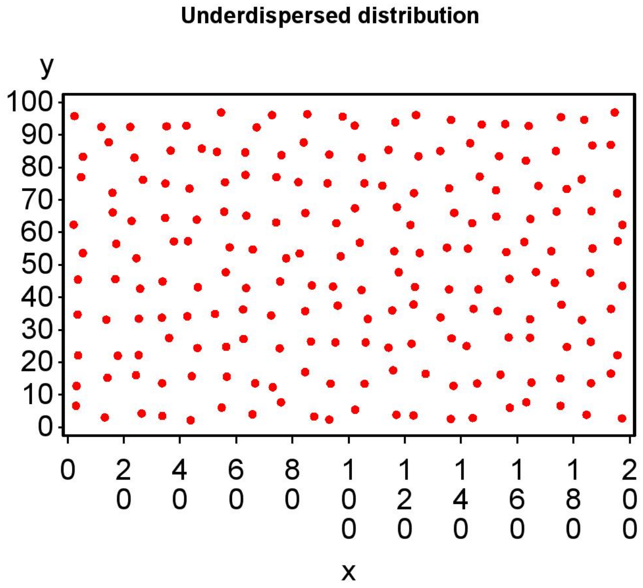


Figure 5.25: Underdispersed distribution of organisms in space

5.7 References

Begon, M., Townsend, C. R. & Harper, J. L. (2006) *Ecology: From Individuals to Ecosystems.* Blackwood Publishing, Malden, MA.

Bliss, C. I. & Fisher, R. A. (1953). Fitting the negative binomial distribution to biological data. *Biometrics* 9: 176-200.

Mood, A. M., Graybill, F. A. & Boes, D. C. (1974) *Introduction to the Theory of Statistics.* McGraw-Hill, Inc., New York, NY.

Pielou, E. (1977) *Mathematical Ecology.* John Wiley & Sons, Inc., New York, NY.

Reeve, J. D., and J. T. Cronin (2010) Edge behaviour in a minute parasitic wasp. *Journal of Animal Ecology*, 79: 483-490.

Ridout, M. S. & Besbeas, P. (2004) An empirical model for underdispersed data. *Statistical Modelling* 4: 77-89.

SAS Institute Inc. (2018) *SAS/STAT 15.1 Users Guide* SAS Institute Inc., Cary, NC

Snyder, D. L. & Miller, M. I. (1991) *Random Point Processes in Time and Space*, 2nd edition. Springer-Verlag New York Inc., New York, NY.

5.8 Problems

1. Consider the dice cube example from Chapter 4, and define a random variable Y that is the number of spots showing on the dice cube. Find $E[Y]$ and $Var[Y]$ for this random variable. Show your work.

2. Suppose that a random variable Y has a discrete distribution with the following probabilities:

y	$P[Y = y]$
0	0.5000
1	0.2500
2	0.1250
3	0.0625
4	0.0625

   (a) What is the expected value of Y, or $E[Y]$?

   (b) What is the variance of Y, or $Var[Y]$?

3. An entomologist studies the spatial distribution of aphids in a field. They randomly select 100 locations within the field and count the number of aphids on the plants at each location. The following observed frequency distribution was obtained:

Aphids (y)	Frequency
0	19
1	22
2	16
3	10
4	11
5	11
6	6
7	2
8	1
9	1
10	1
11	0

(a) Use the SAS program `Poisson_fit.sas` to calculate $\bar{Y}$ and s^2, and generate a plot of the observed frequencies vs. those expected for the Poisson distribution. Attach your SAS program and output.

(b) Based on the above results, do the data have a Poisson distribution? Explain your answer using the pattern of observed and expected frequencies, and the values of $\bar{Y}$ and s^2. Is the pattern random (Poisson), overdispersed, or underdispersed?

(c) What are some possible biological explanations for this pattern?

4. A field is surveyed for golden mice (*Ochrotomys nuttalli*) using a grid of baited traps. A total of 100 traps were deployed and the number of mice counted in each trap. The following frequency distribution was obtained:

Mice (y)	Frequency
0	55
1	21
2	10
3	7
4	4
5	2
6	1
7	0
8	0

(a) Use the program `Poisson_fit.sas` to calculate to calculate $\bar{Y}$ and s^2, and generate a plot of the observed frequencies vs. those expected for the Poisson distribution. Attach your program and output.

(b) Based on the above results, do the data have a Poisson distribution? Explain your answer using the pattern of observed and expected frequencies, and the values of $\bar{Y}$ and s^2. Is the pattern random (Poisson), overdispersed, or underdispersed?

Chapter 6

Continuous Random Variables

We previously examined several different probability distributions for discrete random variables, in particular the binomial, Poisson, and negative binomial distributions. These distributions are suitable for modeling observations that are counts of some type, such as the number of plants in a quadrat or the number of females vs. males in a sample. Many variables in biology are continuous, however, such as the length and weight of organisms, quantities associated with populations such as birth, mortality, and growth rates, and chemical concentrations. We will now examine continuous random variables and their associated distributions that are used to model these quantities, in particular the **uniform and normal distributions**. The uniform distribution is often used to generate random sampling points in one- and two-dimensional areas. For example, we could use the uniform distribution to select a random point along a transect to sample, or a random x, y coordinate within a field to place a sampling quadrat. It also a useful starting point for understanding continuous distributions because of its simplicity. We then turn to the normal distribution, which forms the basis of many statistical procedures. Many biological variables have a distribution close to normal, or if initially non-normal can often be transformed to more closely resemble the normal distribution.

Discrete random variables have a function $f(y)$ that directly provides the probabilities for events that are integers, such as $Y = 0$, $Y = 3$, and so forth (see Chapter 5). However, events for continuous random variables are in the form of intervals. For example, we will be interested in finding the probability for events like $1 < Y < 3$ or $Y > 5$. Continuous random variables use a different kind of function, called a **probability density function**, to find the probabilities for events. For an event like $1 < Y < 3$, probabilities are found by integrating the probability density function (finding the area under the function) over this interval. This process will be explained in more detail below. For many continuous random variables, such as the normal distribution, there exist tables of these integrals and probabilities for certain useful intervals. Note that events like $Y = 3$ have zero probability for continuous random variables, because this implies an interval of zero width and so the integral is zero. This makes some intuitive sense, because it is unlikely that a continuous quantity Y would take a value exactly equal to 3 to many decimal places.

6.1 Uniform distribution

Suppose that we have two constants, a and b, with $a < b$. A random variable Y has a uniform distribution if an observation is equally likely to occur anywhere between a and b, but never occurs outside this interval. The probability density for the uniform distribution is defined by the equation

$$f(y) = \frac{1}{b-a} \tag{6.1}$$

for $a \leq y \leq b$ (Mood et al. 1974). Outside of this interval, we have $f(y) = 0$. The quantities a and b are the parameters of the uniform distribution. The uniform distribution for $a = 0$, $b = 1$ is shown below (Fig. 6.1). The uniform distribution gets its name from the fact that its density is uniform over the interval a to b.

Note that the density simply describes a square with a length and width of one, implying an area equal to one. This is an important property of probability density functions in general – the area under $f(y)$ is always equal to one. Also shown is the uniform density for $a = 0$ and $b = 2$ (Fig. 6.2). It is lower but wider than the previous example, and also has an area of one.

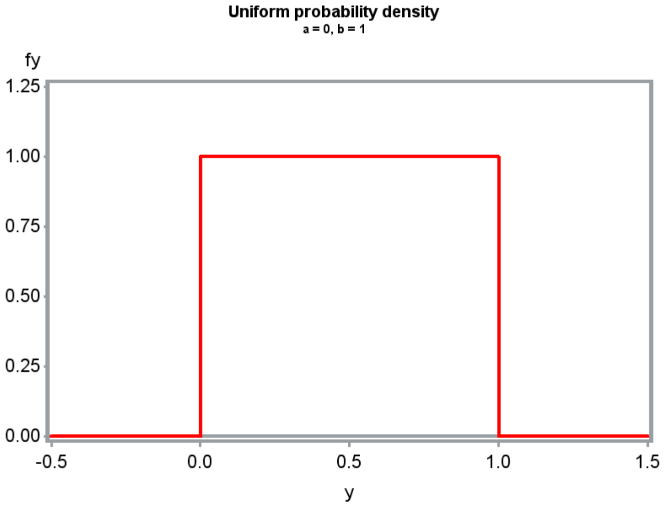


Figure 6.1: Uniform probability density for $a = 0, b = 1$

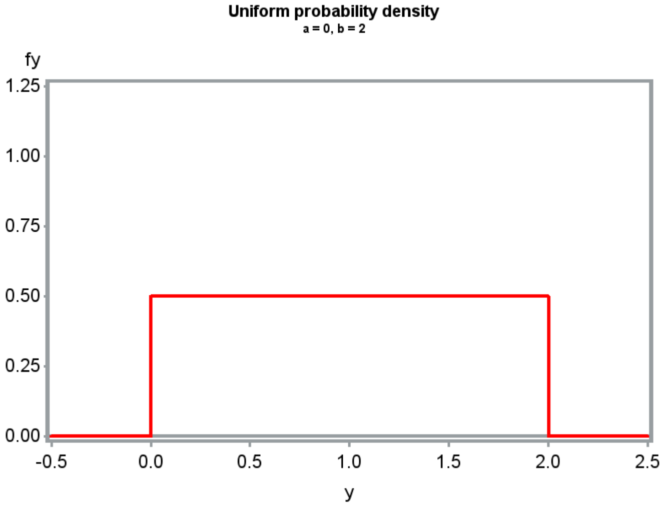


Figure 6.2: Uniform probability density for $a = 0, b = 2$

Probabilities for the uniform distribution are calculated by finding the area under the probability density function, using integration (see Chapter 2). This is relatively easy to do because of the simple form of the probability density. Suppose Y is a uniform random variable, and $a = 0$ and $b = 1$. What is the probability that an observed Y lies within the interval 0.5 to 0.75? We have

$$\begin{aligned} P[0.5 < Y < 0.75] &= \int_{0.5}^{0.75} \frac{1}{b-a} dy \qquad (6.2) \\ &= \int_{0.5}^{0.75} \frac{1}{1-0} dy = y|_{0.5}^{0.75} \qquad (6.3) \\ &= 0.75 - 0.5 = 0.25. \qquad (6.4) \end{aligned}$$

We could also have found this probability without any calculus. It is just the area under $f(y)$ between 0.5 and 0.75, calculated as length × height $= (0.75 - 0.5) \times 1 = 0.25$.

Here are two more examples. Suppose that for $a = 0$ and $b = 2$, we want to find the probability that $0.2 < Y < 0.4$. The height of the density function in this case is $1/(b - a) = 1/(2 - 0) = 0.5$. We therefore have $P[0.2 < Y < 0.4] = (0.4 - 0.2) \times 0.5 = 0.1$. Now suppose we want the probability that $0 < Y < 2$. We have $P[0 < Y < 2] = (2-0) \times 0.5 = 1$. This also follows from the fact that $f(y)$ is a probability density function which has an area of one, and the interval $0 < Y < 2$ encompasses the entire range of $f(y)$.

The **cumulative distribution function** for a continuous random variable is defined as the quantity

$$F(y) = P[Y < y] = \int_{-\infty}^{y} f(z) dz. \qquad (6.5)$$

This function is just the probability to the left of y. The function $F(y)$ increases from 0 to 1 as y increases. If we carry out this integral for the uniform distribution, we get the function

$$F(y) = \frac{y-a}{b-a} \qquad (6.6)$$

for $a \leq y \leq b$. In addition, $F(y) = 0$ for $y < a$, and $F(y) = 1$ for $y > b$. Figure 6.3 shows the cumulative distribution function for the uniform distribution corresponding to Fig. 6.2. Note that it increases linearly between a and b, as the probability to the left of y accumulates. The cumulative distribution function has many uses in statistics, especially for continuous random variables.

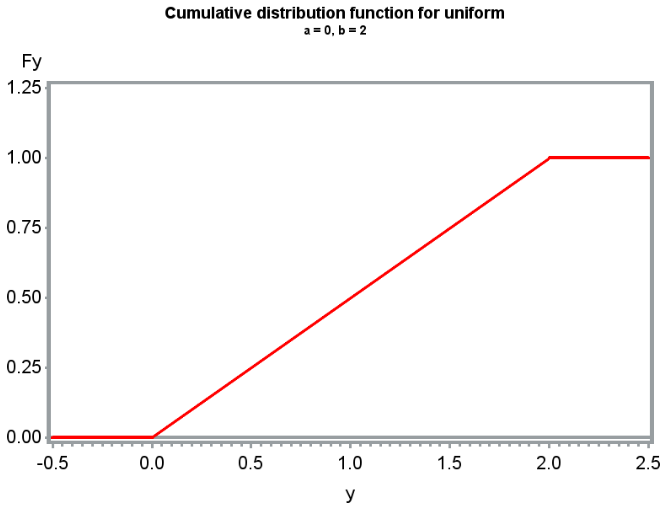


Figure 6.3: Cumulative distribution function for the uniform distribution, with $a = 0, b = 2$

The uniform distribution has a number of common applications. It is possible to generate a stream of random numbers that have a uniform distribution using software, and from these values produce random observations for other distributions, including discrete distributions as well as the normal distribution. The uniform distribution can also be used to generate random sampling points along a transect for ecological studies, or random x, y coordinates for placing quadrats within an area (see below). It can also be used to randomly sample from a population, or to randomize the order of treatments in an experiment.

6.1.1 Random sampling coordinates - SAS demo

A common application of the uniform distribution is to generate random sampling coordinates. SAS can produce random observations with a uniform distribution using the function `ranuni`. For this function, the parameter values of the uniform distribution are set at $a = 0$ and $b = 1$.

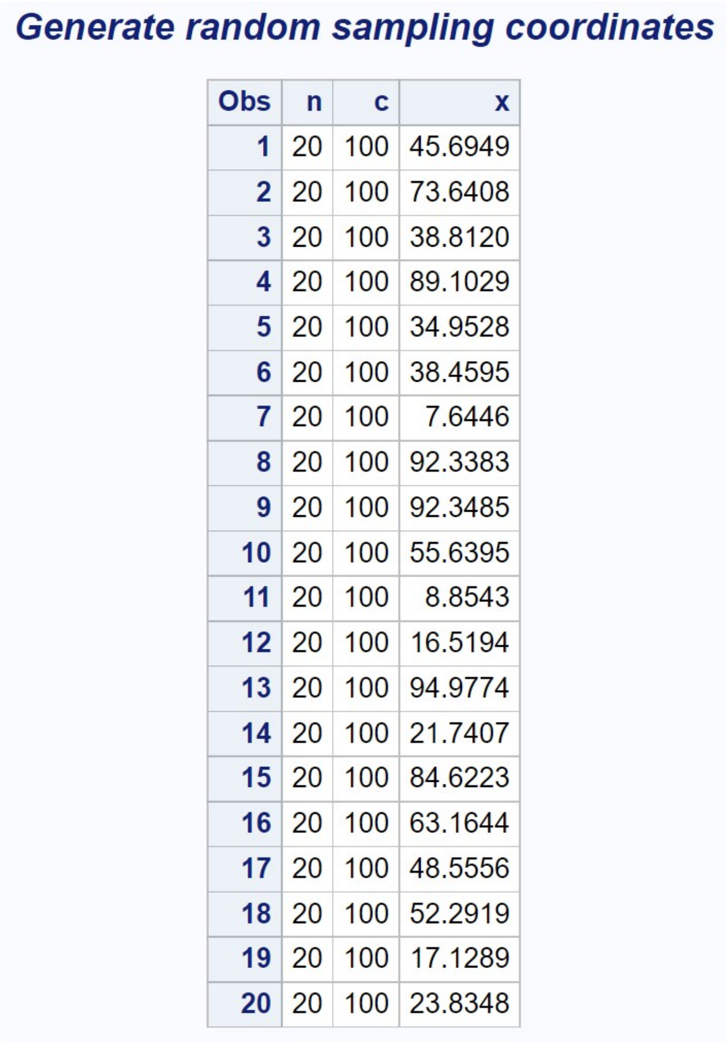
However, we will often want observations for other parameter values, especially other values of b. It can be shown that if Y has a uniform distribution with $a = 0$ and $b = 1$, then the variable $Y' = cY$ has a uniform distribution with $a = 0$ and $b = c$, where c is any positive number. This fact enables us to generate uniform random variables with any value of b.

For example, suppose we want to produce random sampling coordinates along a 100 m transect using the uniform distribution. If Y has a uniform distribution with $a = 0$ and $b = 1$, then $Y' = 100Y$ has a uniform distribution with $a = 0$ and $b = 100$. Values of Y generated in this fashion will give us sampling coordinates uniformly distributed between 0 and 100 m.

We will illustrate this process using a SAS program to generate random sampling coordinates for a 100 m transect and also a 200 × 100 m rectangular area. A call to `gplot` is used to plot the random coordinates. See SAS program and output below.

SAS Program

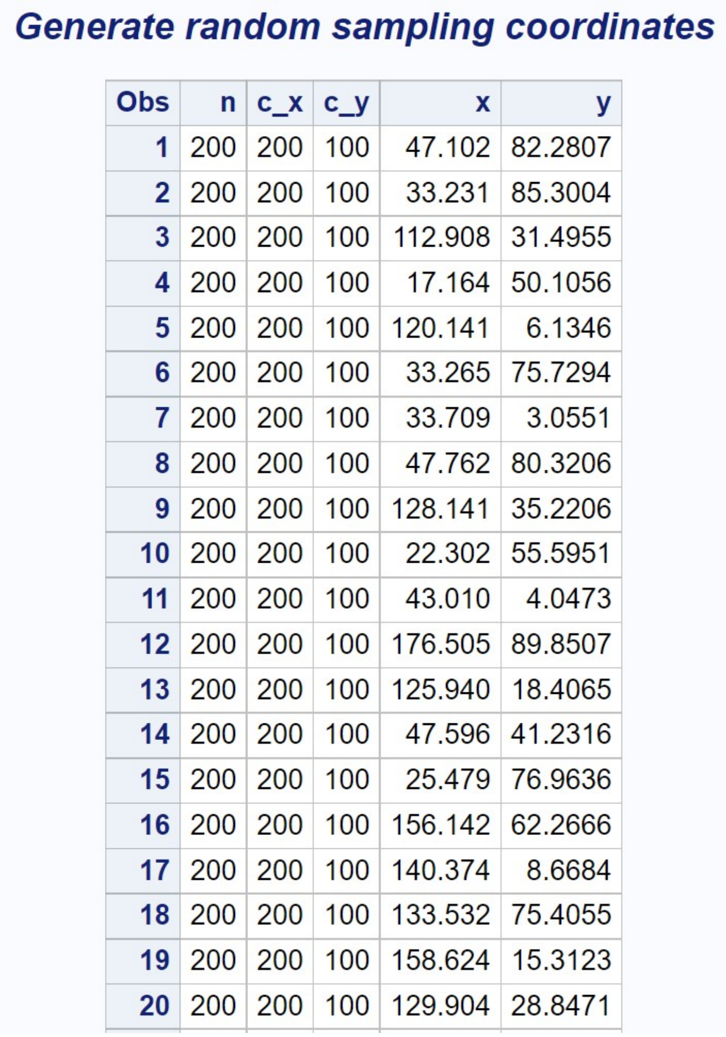
```
* randcoords.sas;
title "Generate random sampling coordinates";
* Generate n random coordinates along a c m transect;
data transect;
    * Sample size n;
    n = 20;
    * Multiplying by c gives a uniform random variable with a=0, b=c;
    c = 100;
    do i = 1 to n;
        x = c*ranuni(0);
        output;
    end;
    drop i;
run;
* Print coordinates;
proc print data=transect;
run;
* Generate n random coordinates within a 200 x 100 m area;
data coords;
    * Sample size n;
    n = 200;
    * Multiplying by c_x gives a uniform random variable with a=0, b=c_x;
    c_x = 200;
    * Multiplying by c_y gives a uniform random variable with a=0, b=c_y;
    c_y = 100;
    do i = 1 to n;
        x = c_x*ranuni(0);
        y = c_y*ranuni(0);
        output;
    end;
    drop i;
run;
* Print first 25 coordinates;
proc print data=coords(obs=25);
run;
* Show coordinates as a scatterplot;
proc gplot data=coords;
    plot y*x / vaxis=axis1 haxis=axis2;
    symbol1 v=dot c=red;
    axis1 order=(0 to 100 by 10) label=(height=2) value=(height=2)
    width=3 major=(width=2) minor=none;
    axis2 order=(0 to 200 by 20) label=(height=2) value=(height=2)
    width=3 major=(width=2) minor=none;
run;
quit;
```



Generate random sampling coordinates

Obs	n	c	x
1	20	100	45.6949
2	20	100	73.6408
3	20	100	38.8120
4	20	100	89.1029
5	20	100	34.9528
6	20	100	38.4595
7	20	100	7.6446
8	20	100	92.3383
9	20	100	92.3485
10	20	100	55.6395
11	20	100	8.8543
12	20	100	16.5194
13	20	100	94.9774
14	20	100	21.7407
15	20	100	84.6223
16	20	100	63.1644
17	20	100	48.5556
18	20	100	52.2919
19	20	100	17.1289
20	20	100	23.8348

Figure 6.4: `randcoords.sas - proc print`



Generate random sampling coordinates

Obs	n	c_x	c_y	x	y
1	200	200	100	47.102	82.2807
2	200	200	100	33.231	85.3004
3	200	200	100	112.908	31.4955
4	200	200	100	17.164	50.1056
5	200	200	100	120.141	6.1346
6	200	200	100	33.265	75.7294
7	200	200	100	33.709	3.0551
8	200	200	100	47.762	80.3206
9	200	200	100	128.141	35.2206
10	200	200	100	22.302	55.5951
11	200	200	100	43.010	4.0473
12	200	200	100	176.505	89.8507
13	200	200	100	125.940	18.4065
14	200	200	100	47.596	41.2316
15	200	200	100	25.479	76.9636
16	200	200	100	156.142	62.2666
17	200	200	100	140.374	8.6684
18	200	200	100	133.532	75.4055
19	200	200	100	158.624	15.3123
20	200	200	100	129.904	28.8471

etc.

Figure 6.5: `randcoords.sas - proc print`

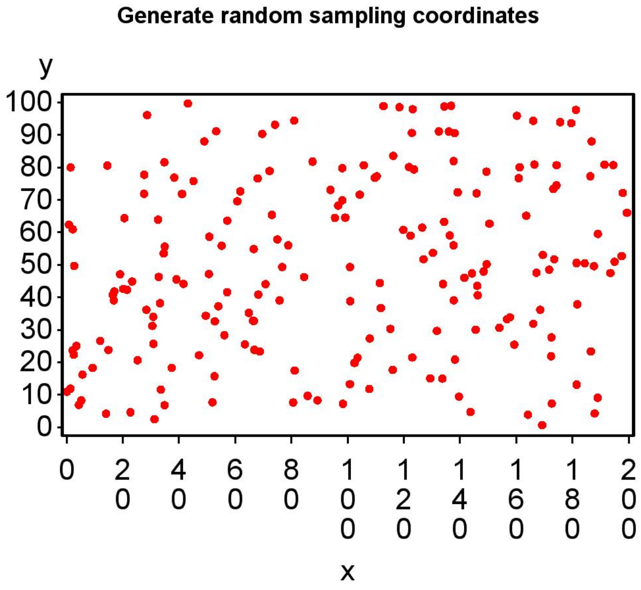


Figure 6.6: `randcoords.sas - proc gplot`

6.2 Normal distribution

The normal distribution plays an important role in statistics, with good reason. Biological variables often have a distribution that can be approximated by the normal or can be transformed to be normal. The normal distribution is thus a valid choice for modeling many variables encountered in practice. Many statistical quantities will also have a distribution approaching the normal for large sample sizes. For example, the distribution of the sample mean $\bar{Y}$ will approach the normal distribution as the sample size n increases, thanks to the central limit theorem (see Chapter 7). So, even if the underlying data are non-normal, statistics like $\bar{Y}$ will be normally-distributed for sufficiently large n.

The probability density for the normal distribution is defined by the function

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}} \tag{6.7}$$

for $\infty < \mu < \infty$ and $\sigma^2 > 0$ (Mood et al. 1974). The normal distribution has two parameters, μ and σ^2. The parameter μ is the mean of the distribution and basically controls its location, while σ^2 is its variance and determines its dispersion or spread. A random variable Y with a normal distribution is often written as $Y \sim N(\mu, \sigma^2)$, where the symbol '$\sim$' stands for 'is distributed as' while 'N' signifies the normal. A random variable with a **standard normal distribution** assumes that $\mu = 0$ and $\sigma^2 = 1$, or $Y \sim N(0, 1)$. The symbol Z is often used to denote a standard normal random variable.

Figure 6.7 shows the bell-shaped normal distribution for three different sets of μ and σ^2 values, and illustrates how these parameters affect its location and shape. As μ is increased the distribution shifts to the right, while an increase in σ^2 causes the distribution to spread out.

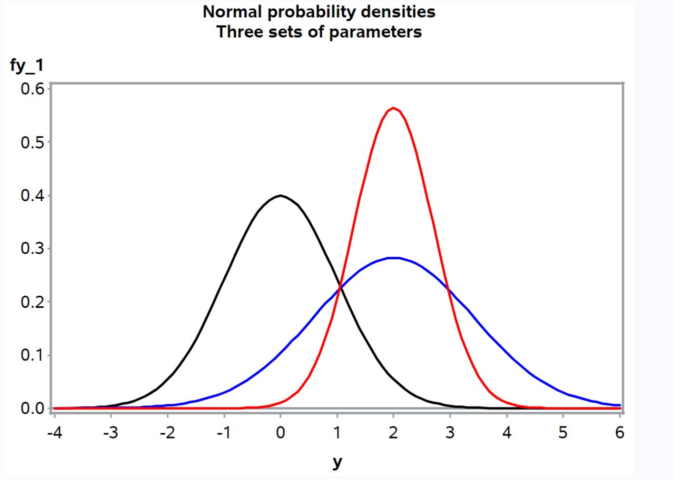


Figure 6.7: `normal_plot3.sas - proc gplot`

6.2.1 Normal distribution - SAS demo

The SAS program used to generate Fig. 6.7 is listed below. Three different sets of μ and σ^2 values are given in the `data` step of the program (feel free to experiment with other values). The different curves are specified in the `plot` statement for `proc gplot`. The `overlay` option is used to generate a single graph with all three curves, each with different colors specified by the `symbol` statements.

SAS Program

```
* normal_plot3.sas;
options pageno=1 linesize=80;
goptions reset=all;
title "Normal probability densities";
title2 "Three sets of parameters";
data normal_plot;
    * Three sets of normal parameters here;
    mu_1 = 0; sig2_1 = 1;
    mu_2 = 2; sig2_2 = 2;
    mu_3 = 2; sig2_3 = 0.5;
    * Minimum and maximum values of y;
    ymin = -4;
    ymax = 6;
    * Divisions between ymin and ymax (more = smoother graph);
    ydiv = 100;
    * Calculate step length;
    ylength = (ymax-ymin)/ydiv;
    * Find y and f(y) values for the plot;
    do i=0 to ydiv;
        y = ymin + i*ylength;
        * normal probability density function;
        fy_1 = (1/sqrt(2*3.14159*sig2_1))*exp(-((y-mu_1)**2)/(2*sig2_1));
        fy_2 = (1/sqrt(2*3.14159*sig2_2))*exp(-((y-mu_2)**2)/(2*sig2_2));
        fy_3 = (1/sqrt(2*3.14159*sig2_3))*exp(-((y-mu_3)**2)/(2*sig2_3));
        * Output y and fy1, fy2, fy3 to SAS data file;
        output;
    end;
run;
* Print data;
proc print data=normal_plot;
run;
* Plot probability density function;
proc gplot data=normal_plot;
    plot fy_1*y=1 fy_2*y=2 fy_3*y=3 / vref=0 wvref=3 vaxis=axis1 haxis=axis1 overlay;
    symbol1 i=join v=none c=black width=3;
    symbol2 i=join v=none c=blue width=3;
    symbol3 i=join v=none c=red width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
quit;
```

The cumulative distribution function for the normal distribution is defined as the quantity

$$F(y) = P[Y < y] = \int_{-\infty}^{y} f(z)dz = \int_{-\infty}^{y} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(z-\mu)^2}{2\sigma^2}} dz. \qquad (6.8)$$

The values of this integral have to be numerically calculated. Fig. 6.8 shows the cumulative distribution functions for the three normal distributions shown in Fig. 6.7. Note that the mean and variance for the different normal distributions affect the overall location and shape of $F(y)$.

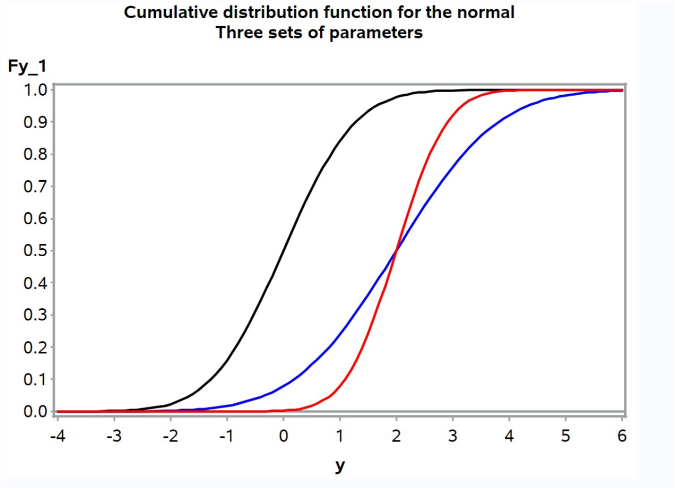


Figure 6.8: Cumulative distribution function for three normal distributions

Like other continuous random variables, events for the normal distribution are in the form of intervals. We can calculate the probabilities for events by finding the area under the normal density function corresponding to the interval. This process is more difficult than for the uniform distribution because $f(y)$ has a more complex shape. However, there exist tables of the area under $f(y)$ for certain intervals that can be used for this purpose, as well as the SAS function `probnorm`. Table Z gives the probabilities for intervals of the form $Z < z$, where Z has a standard normal distribution and $z \geq 0$ (see Chapter 23). The first two digits of z are specified in the left-most column of Table Z, while the third digit is the top row. The values within the table correspond to the probability that $Z < z$, or $P[Z < z]$, i.e., the cumulative distribution function for the standard normal.

6.2.2 Sample calculations - standard normal distribution

We illustrate how Table Z is used to calculate the probabilities for various events listed below. The general strategy is to sketch the interval on the standard normal bell curve, and deduce from this picture how to obtain the probability using Table Z.

1. Find the probability that $Z < 0.55$, or $P[Z < 0.55]$. From Table Z, we see that $P[Z < 0.55] = 0.7088$. See Fig. 6.9 for an illustration of this probability.

2. Find the probability that $0.40 < Z < 1.96$. In this case, the interval is not the same as shown in Table Z, and additional calculations are required. We first find the probabilities for the intervals $Z < 1.96$ and $Z < 0.4$ using Table Z. The probability for $0.40 < Z < 1.96$ should then be the difference between these two probabilities (see Fig. 6.10). We have $P[Z < 1.96] = 0.9750$ and $P[Z < 0.40] = 0.6554$ from Table Z, so $P[0.40 < Z < 1.96] = P[Z < 1.96] - P[Z < 0.40] = 0.9750 - 0.6554 = 0.3196$.

3. Find the probability that $Z > 0.55$. We will use the complement rule to obtain this probability (see Chapter 4). For any event A, we have $P[A^c] = 1 - P[A]$. If A is the event $Z < 0.55$, then A^C corresponds to $Z > 0.55$. Therefore, $P[Z > 0.55] = 1 - P[Z < 0.55] = 1 - 0.7088 = 0.2912$. See also Fig. 6.11.

4. Find the probability that $Z < -1.23$. This problem makes use of the symmetry of the standard normal distribution around zero, as well as the complement rule. By symmetry, we have $P[Z < -1.23] = P[Z > 1.23]$. The complement of $Z < 1.23$ is $Z > 1.23$, and so

$P[Z > 1.23] = 1 - P[Z < 1.23] = 1 - 0.8907 = 0.1093$. See Fig. 6.12.

5. Find the probability that $-0.44 < Z < 2.15$. This problem can also be handled using symmetry and the complement rule. We first have $P[Z < 2.15] = 0.9842$ using Table Z (Fig. 6.13). We then have $P[Z < -0.44] = P[Z > 0.44] = 1 - P[Z < 0.44] = 1 - 0.6700 = 0.3300$ by symmetry (Fig. 6.14). Therefore, $P[-0.44 < Z < 2.15] = P[Z < 2.15] - P[Z < -0.44] = 0.9842 - 0.3300 = 0.6542$.

6. Find a number z_0 such that $P[Z < z_0] = 0.95$. This problem is the inverse of the previous ones. Here, we want to find a value z_0 that gives a certain probability, rather than z_0 being a given quantity and determining the probability. To find z_0, we scan Table Z until we find a value that gives a probability close 0.95. We see that $z_0 = 1.64$ or 1.65 give approximately the right probability.

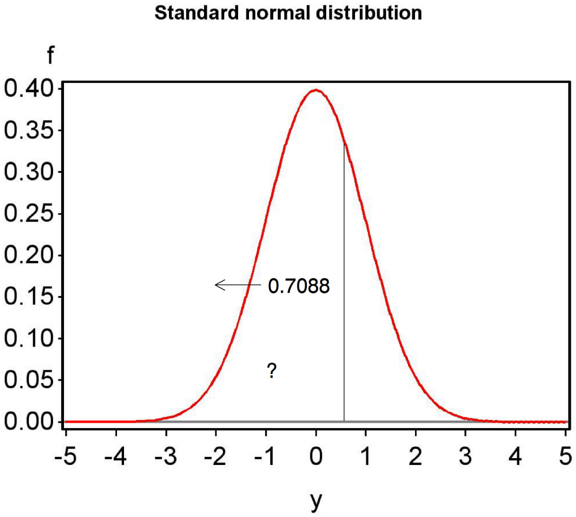


Figure 6.9: Sample calculation 1

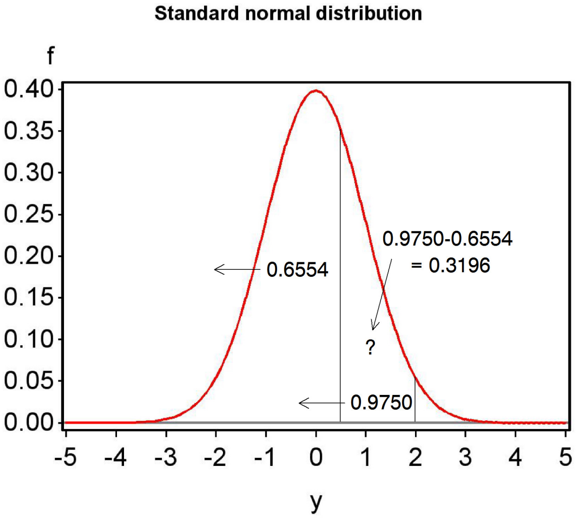


Figure 6.10: Sample calculation 2

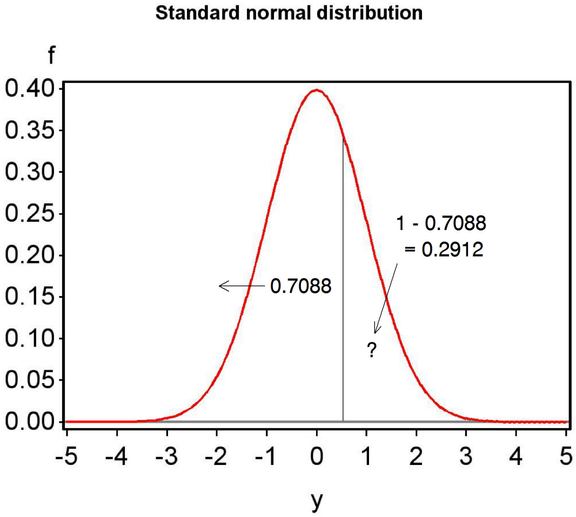


Figure 6.11: Sample calculation 3

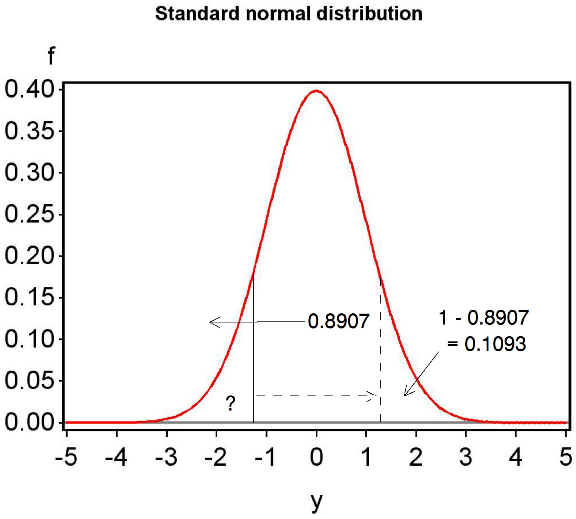


Figure 6.12: Sample calculation 4

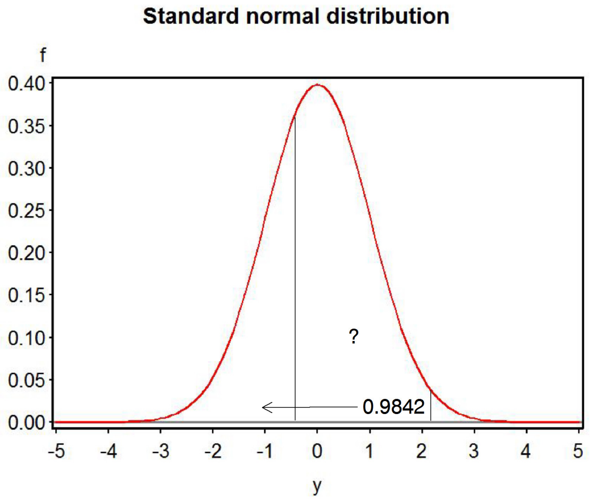


Figure 6.13: Sample calculation 5 - part 1

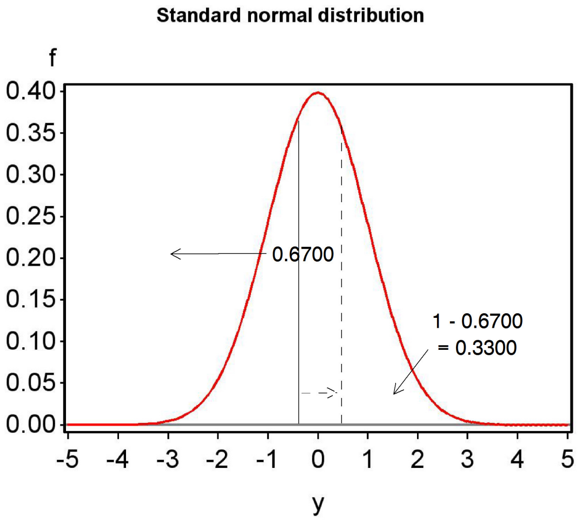


Figure 6.14: Sample calculation 5 - part 2

6.2.3 Sample calculations - other normal distributions

We now examine how probabilities can be calculated for normal distributions that are not standard normal. If $Y \sim N(\mu, \sigma^2)$, it can be shown that the quantity

$$Z = \frac{Y - \mu}{\sigma} \sim N(0, 1) \tag{6.9}$$

Thus, a random variable Y with a normal distribution having any μ or σ^2 can be transformed to a standard normal Z. The transformation works by first centering the random variable Y around zero by subtracting μ, and then dividing by σ so that it has a standard deviation and variance of one. Once Y is transformed to a standard normal Z, we can find probabilities for any event involving Y using Table Z. This process is illustrated below in several sample calculations.

1. Suppose that $Y \sim N(50, 16)$. Find the probability that $Y < 55$. First, we find $\sigma = \sqrt{\sigma^2} = \sqrt{16} = 4$. Using the above equation, we then have

$$\begin{aligned} P[Y < 55] &= P\left[Y - \mu < 55 - \mu\right] && (6.10) \\ &= P\left[\frac{Y - \mu}{\sigma} < \frac{55 - \mu}{\sigma}\right] && (6.11) \\ &= P\left[Z < \frac{55 - 50}{4}\right] && (6.12) \\ &= P[Z < 1.25]. && (6.13) \end{aligned}$$

   We then use Table Z to find that $P[Z < 1.25] = 0.8944$, and so $P[Y < 55] = 0.8944$.

2. Find the probability that $52 < Y < 56$, assuming $Y \sim N(50, 16)$. To find this probability, we first convert the problem to one involving Z. We have

$$\begin{aligned} P[52 < Y < 56] &= P\left[52 - \mu < Y - \mu < 56 - \mu\right] && (6.14) \\ &= P\left[\frac{52 - \mu}{\sigma} < \frac{Y - \mu}{\sigma} < \frac{56 - \mu}{\sigma}\right] && (6.15) \\ &= P\left[\frac{52 - 50}{4} < Z < \frac{56 - 50}{4}\right] && (6.16) \\ &= P[0.50 < Z < 1.50]. && (6.17) \end{aligned}$$

We next find the probabilities for the intervals $Z < 1.50$ and $Z < 0.50$ using Table Z, and then substract them to obtain $P[0.50 < Z < 1.50]$. We have $P[Z < 1.50] = 0.9332$ and $P[Z < 0.50] = 0.6915$, so $P[0.50 < Z < 1.50] = 0.9332-0.6915 = 0.2417$. Thus, $P[52 < Y < 56] = 0.2417$.

3. Find the probability that $Y > 54$. We have

$$\begin{aligned} P[Y > 54] &= P\left[Y - \mu > 54 - \mu\right] && (6.18)\\ &= P\left[\frac{Y-\mu}{\sigma} > \frac{54-\mu}{\sigma}\right] && (6.19)\\ &= P\left[Z > \frac{54-50}{4}\right] && (6.20)\\ &= P[Z > 1.00]. && (6.21) \end{aligned}$$

We next use the complement rule to obtain this probability. We have $P[Z > 1.00] = 1 - P[Z < 1.00] = 1 - 0.8413 = 0.1587$, so $P[Y > 54] = 0.1587$.

4. Find the probability that $Y < 46.5$. We have

$$\begin{aligned} P[Y < 46.5] &= P\left[Y - \mu < 46.5 - \mu\right] && (6.22)\\ &= P\left[\frac{Y-\mu}{\sigma} < \frac{46.5-\mu}{\sigma}\right] && (6.23)\\ &= P\left[Z < \frac{46.5-50}{4}\right] && (6.24)\\ &= P[Z < -0.88]. && (6.25) \end{aligned}$$

By symmetry, we have $P[Z < -0.88] = P[Z > 0.88]$. The complement of $Z < 0.88$ is $Z > 0.88$, and so $P[Z > 0.88] = 1 - P[Z < 0.88] = 1 - 0.8106 = 0.1093$. So, $P[Y < 46.5] = 0.1093$.

5. Find the probability that $46 < Z < 52$. We have

$$\begin{aligned} P[46 < Y < 52] &= P\left[46 - \mu < Y - \mu < 52 - \mu\right] && (6.26)\\ &= P\left[\frac{46-\mu}{\sigma} < \frac{Y-\mu}{\sigma} < \frac{52-\mu}{\sigma}\right] && (6.27)\\ &= P\left[\frac{46-50}{4} < Z < \frac{52-50}{4}\right] && (6.28)\\ &= P[-1.00 < Z < 0.50]. && (6.29) \end{aligned}$$

We then use symmetry and the complement rule to find this probability involving Z. We first have $P[Z < 0.50] = 0.6915$ using Table Z. We then have $P[Z < -1.00] = P[Z > 1.00] = 1 - P[Z < 1.00] = 1 - 0.8413 = 0.1587$ by symmetry. Therefore, $P[-1.00 < Z < 0.50] = P[Z < 0.50] - P[Z < -1.00] = 0.6915 - 0.1587 = 0.5328$, and so $P[46 < Y < 52] = 0.5328$.

6. Find a number y_0 such that $P[Y < y_0] = 0.70$. This problem can also be handled by converting it to one involving Z. We have

$$\begin{aligned} P[Y < y_0] &= P[Y - \mu < y_0 - \mu] && (6.30)\\ &= P\left[\frac{Y-\mu}{\sigma} < \frac{y_0-\mu}{\sigma}\right] && (6.31)\\ &= P\left[Z < \frac{y_0 - 50}{4}\right] && (6.32)\\ &= P[Z < z_0] && (6.33)\end{aligned}$$

where $z_0 = \frac{y_0-50}{4}$. We then search for a value of z_0 such that $P[Z < z_0] = 0.70$, and obtain $z_0 = 0.52$ from Table Z. We then solve for y_0 as follows:

$$\begin{aligned} z_0 &= \frac{y_0 - 50}{4} && (6.34)\\ 0.52 &= \frac{y_0 - 50}{4} && (6.35)\\ 4(0.52) &= y_0 - 50 && (6.36)\\ 2.08 &= y_0 - 50 && (6.37)\\ 2.08 + 50 &= y_0 && (6.38)\\ 52.08 &= y_0. && (6.39)\end{aligned}$$

So, $y_0 = 52.08$ is the answer. In general, one would have $z_0 = \frac{y_0-\mu}{\sigma}$, so $y_0 = \sigma z_0 + \mu$ for any σ and μ.

6.3 Expected values and variance for continuous distributions

We saw earlier how a theoretical mean, variance, and standard deviation could be calculated for a discrete random variable, using the concept of expectation and its probability distribution. The same concepts can be extended to continuous random variables and probability densities.

Let Y be a continuous random variable with some probability density. The expected value of Y, or its theoretical mean, is defined by the equation

$$E[Y] = \int_{-\infty}^{\infty} y f(y) dy \tag{6.40}$$

where $f(y)$ is the probability density of Y, and the integral is carried out over the interval $-\infty$ to ∞ (Mood et al. 1974). This equation is analogous to the definition of expected value for a discrete random variable, except that we use integration rather than summation to make the calculation.

Similar to discrete random variables, we can also define the theoretical variance of a continuous random variable using expectation. The variance of a continuous random variable Y is defined as

$$Var[Y] = E[(Y - E[Y])^2] = \int_{-\infty}^{\infty} (y - E[Y])^2 f(y) dy. \tag{6.41}$$

We can directly calculate these quantities for the uniform distribution. Recall from calculus that $\int u du = u^2/2$. We therefore have

$$\begin{aligned} E[Y] &= \int_{-\infty}^{\infty} y f(y) dy = \int_a^b \frac{y}{b-a} dy & (6.42) \\ &= \frac{1}{b-a} \frac{y^2}{2} |_a^b = \frac{1}{b-a} \frac{b^2 - a^2}{2} & (6.43) \\ &= \frac{(b-a)(b+a)}{2(b-a)} = \frac{b+a}{2} & (6.44) \end{aligned}$$

Thus, the expected value (or theoretical mean) of a uniform random variable is located at the center of the interval, midway between a and b. It can also be shown using the above formula that the variance of the uniform distribution is

$$Var[Y] = \frac{(b-a)^2}{12} \tag{6.45}$$

The theoretical standard deviation is just the square root of this quantity.

What are these quantities for the normal distribution? Recall that the normal distribution is specified by the two parameters μ and σ^2. If $Y \sim N(\mu, \sigma^2)$, it can be shown (by evaluating the above integrals using the normal density) that

$$E[Y] = \mu \tag{6.46}$$

and

$$Var[Y] = \sigma^2. \tag{6.47}$$

Thus, the parameters μ and σ^2 for this distribution are the theoretical mean and variance $E[Y]$ and $Var[Y]$.

6.4 Continuous random variables and samples

Suppose we have a set of observations and want to determine if they can be modeled using the normal distribution. We now develop a graphical method of comparing these observed data with the pattern expected for the normal distribution, called a **normal quantile plot**. These plots exist for other continuous distributions as well, and are generally called quantile-quantile plots. The idea is to plot the quantiles for the observed data vs. the quantiles for the normal distribution, with the quantiles for the normal on the x-axis and the data quantiles on the y-axis. If the data are normally distributed, then this plot will resemble a straight diagonal line. This is because we are essentially plotting the quantiles for one normal distribution (the data) vs. the quantiles for the normal distribution itself (Wilk & Gnanadesikan 1968). This is like plotting the function $y = ax$, which is the equation of a line with slope a. See Chapter 3 for a review of quantiles such as the median, the 25% and 75% quartiles, and so forth.

We will illustrate the calculations for a normal quantile plot using a small data set. Suppose we have $n = 9$ data points that take the values 5.33, 4.98, 5.80, 4.37, 3.83, 2.76, 3.82, 4.02, and 3.09. We first order or rank the data points from smallest to largest, similar to finding the median (Table 6.1). We then find the proportion p of observations less than each data point, using the formula $p = (j - 3/8)/(n + 1/4)$, where j is the order of the data point and n is the sample size. Note that the median of these data (the value 4.02) corresponds to $p = 0.5$. The values 3/8 and 1/4 in the formula are there to prevent p from taking the value 0 or 1 for the largest and smallest observations.

Table 6.1: Calculations for a normal quantile plot

j (order)	$Y_{[j]}$	p	z
1	2.76	0.068	-1.49
2	3.09	0.176	-0.93
3	3.82	0.284	-0.57
4	3.83	0.392	-0.27
5	4.02	0.500	0.00
6	4.37	0.608	0.27
7	4.98	0.716	0.57
8	5.33	0.824	0.93
9	5.80	0.932	1.49

We then determine the quantiles of the standard normal distribution that correspond to the values of p for these data. For example, suppose we want to find a value z such that $P[Z < z] = 0.5$, the median of the standard normal distribution. We see from Table Z that $z = 0$ give the correct probability. For $p = 0.932$, we find that $z = 1.49$ gives close to the correct probability. We can similarly find the values of z for the other values of p, giving the last column in Table 6.1. The final step is then to plot the ordered data vs. the normal quantiles (Fig. 6.15). If the data are normally distributed, there should be a linear relationship between the observed data and the normal quantiles, and the normal quantile plot will be a diagonal line. This appears to be the case for these data. If the data are non-normal, however, all manner of curved relationships are possible.

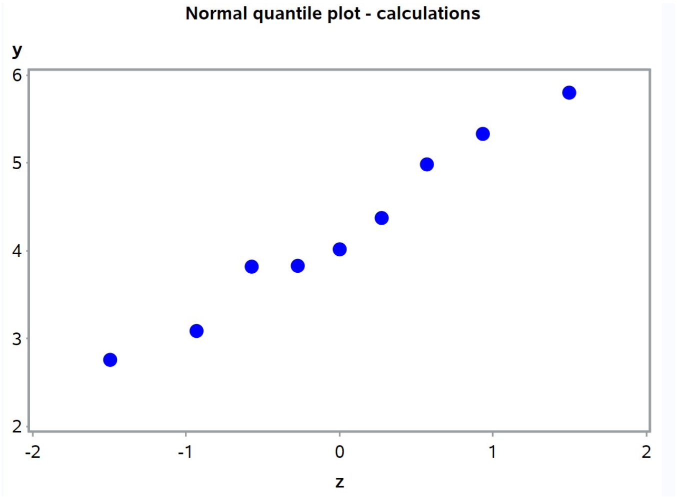


Figure 6.15: Normal quantile plot using Table 6.1

6.4.1 Elytra lengths - SAS demo

We previously examined a data set involving the elytra lengths of male and female *T. dubius* beetles and calculated various descriptive statistics using `proc univariate` (see Chapter 3). We now examine whether these data are normally-distributed using normal quantile plots. A normal quantile plot is requested by adding the command `qqplot` with the `normal` option to the program (see below). A histogram and fitted normal curve can also be generated using the `histogram` command with the `normal` option. Separate analyses are requested for male and female beetles using a `class` statement, because the two sexes differ in size and could also have potentially different distributions. We observe that the normal quantile plots for female beetles is close to linear, suggesting a normal distribution, while the males show some curvature.

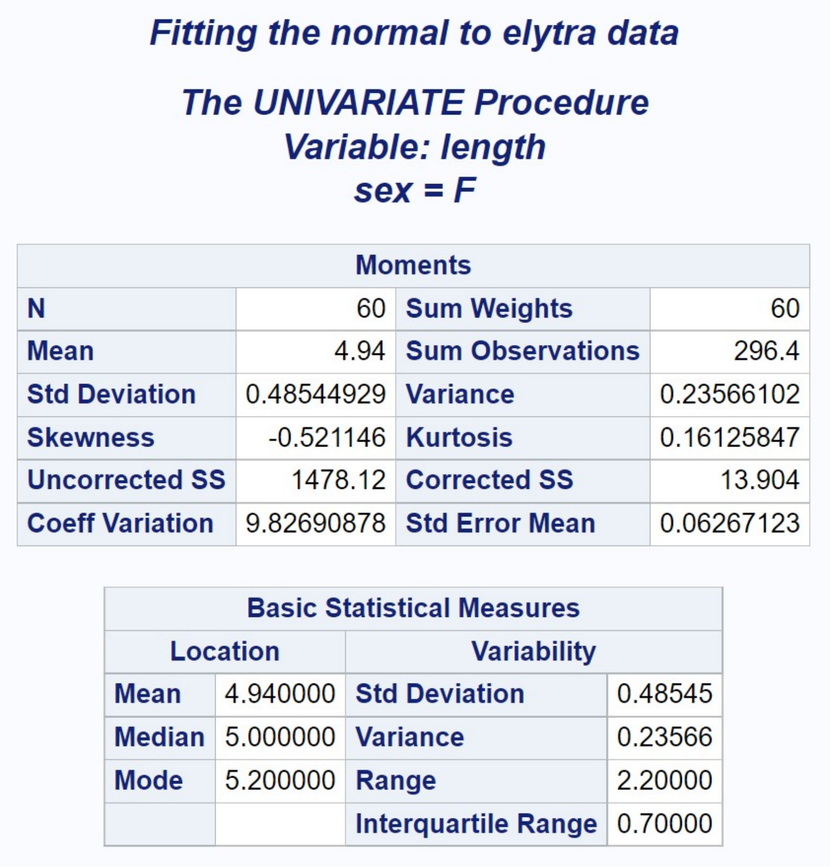
SAS Program

```
* normal_quantile_plot.sas;
title 'Fitting the normal to elytra data';
data elytra;
    input sex $ length;
    datalines;
M   4.9
F   5.2
M   4.9
F   4.2
F   5.7

etc.

M   5.1
F   4.4
M   4.8
M   4.6
F   3.7
;
run;
* Descriptive statistics, histograms, and normal quantile plots;
proc univariate plots data=elytra;
    * Separate analyses for each sex;
    class sex;
    var length;
    histogram length/ vscale=count normal;
    qqplot length / normal;
run;
```

```
quit;
```



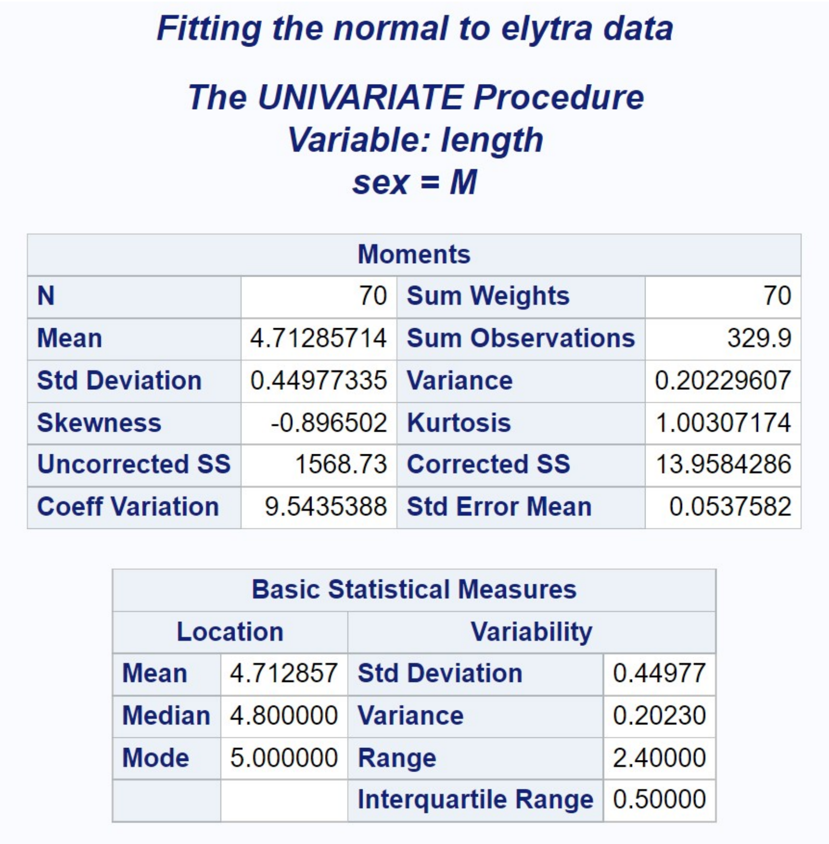
Fitting the normal to elytra data

The UNIVARIATE Procedure
Variable: length
sex = F

Moments			
N	60	Sum Weights	60
Mean	4.94	Sum Observations	296.4
Std Deviation	0.48544929	Variance	0.23566102
Skewness	-0.521146	Kurtosis	0.16125847
Uncorrected SS	1478.12	Corrected SS	13.904
Coeff Variation	9.82690878	Std Error Mean	0.06267123

Basic Statistical Measures			
Location		Variability	
Mean	4.940000	Std Deviation	0.48545
Median	5.000000	Variance	0.23566
Mode	5.200000	Range	2.20000
		Interquartile Range	0.70000

Figure 6.16: `normal_quantile_plot.sas - proc univariate`



Fitting the normal to elytra data

The UNIVARIATE Procedure
Variable: length
sex = M

Moments			
N	70	Sum Weights	70
Mean	4.71285714	Sum Observations	329.9
Std Deviation	0.44977335	Variance	0.20229607
Skewness	-0.896502	Kurtosis	1.00307174
Uncorrected SS	1568.73	Corrected SS	13.9584286
Coeff Variation	9.5435388	Std Error Mean	0.0537582

Basic Statistical Measures			
Location		Variability	
Mean	4.712857	Std Deviation	0.44977
Median	4.800000	Variance	0.20230
Mode	5.000000	Range	2.40000
		Interquartile Range	0.50000

Figure 6.17: `normal_quantile_plot.sas - proc univariate`

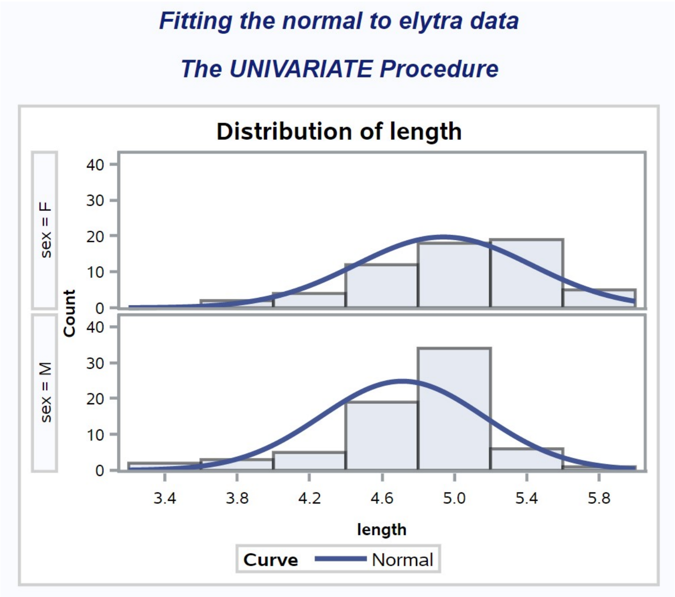


Figure 6.18: `normal_quantile_plot.sas - proc univariate`

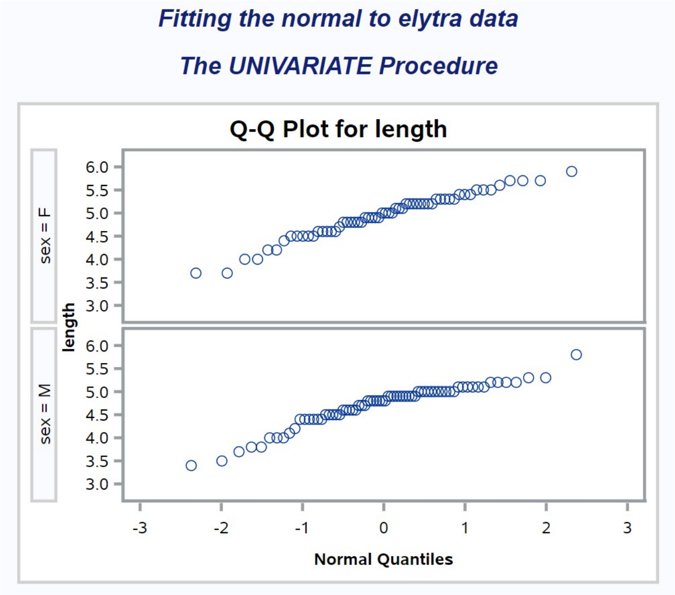


Figure 6.19: `normal_quantile_plot.sas - proc univariate`

6.4.2 Development time - SAS demo

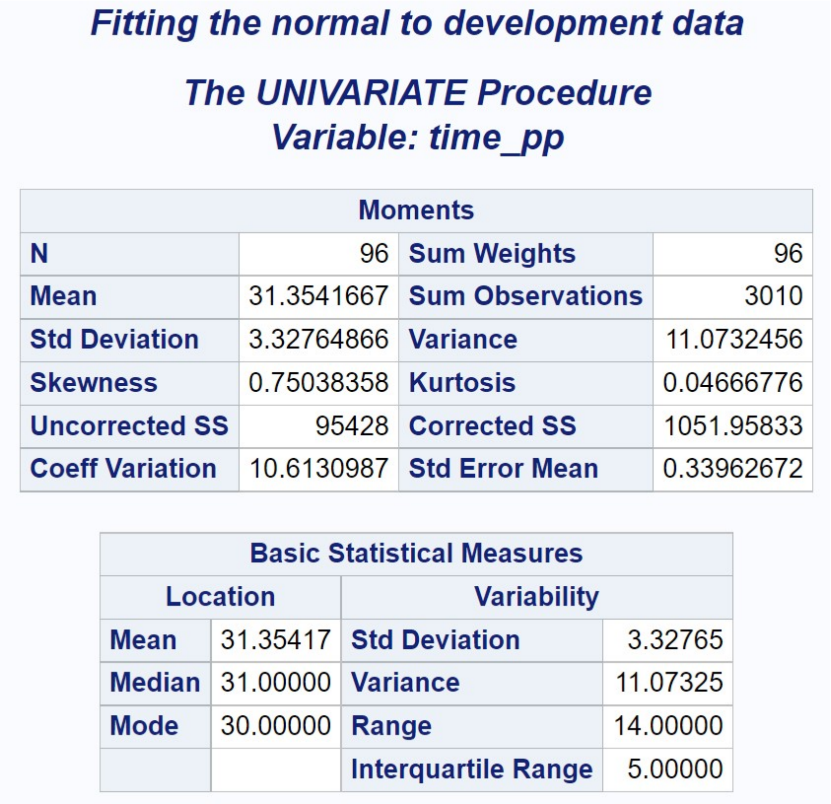
We now examine a data set involving the development time of *T. dubius* beetles in various stages, in particular the time from the larval to prepupal stage, and then from the prepupal to adult stage (Reeve et al. 2003). See program below for details of this analysis. We see that the normal quantile plots for both stages are quite nonlinear, suggesting a distribution different from normal. This is a reflection of the skewed distributions of development time we saw earlier for these data (Chapter 3). Skewed and nonnormal distributions are a common feature of insect development data (Wagner et al. 1984).

SAS Program

```
* normal_quantile_plot_2.sas;
title 'Fitting the normal to development data';
data devel_time;
    input time_pp time_adult;
    datalines;
34  65
31  48
29   .
30  55
32  62

etc.

29   .
29 108
31 103
33   .
29  92
;
run;
* Descriptive statistics, histograms, and normal quantile plots;
proc univariate plots data=devel_time;
    var time_pp time_adult;
    histogram time_pp time_adult / vscale=count normal;
    qqplot time_pp time_adult / normal;
run;
quit;
```



Fitting the normal to development data

The UNIVARIATE Procedure
Variable: time_pp

Moments			
N	96	Sum Weights	96
Mean	31.3541667	Sum Observations	3010
Std Deviation	3.32764866	Variance	11.0732456
Skewness	0.75038358	Kurtosis	0.04666776
Uncorrected SS	95428	Corrected SS	1051.95833
Coeff Variation	10.6130987	Std Error Mean	0.33962672

Basic Statistical Measures			
Location		Variability	
Mean	31.35417	Std Deviation	3.32765
Median	31.00000	Variance	11.07325
Mode	30.00000	Range	14.00000
		Interquartile Range	5.00000

Figure 6.20: `normal_quantile_plot_2.sas - proc univariate`

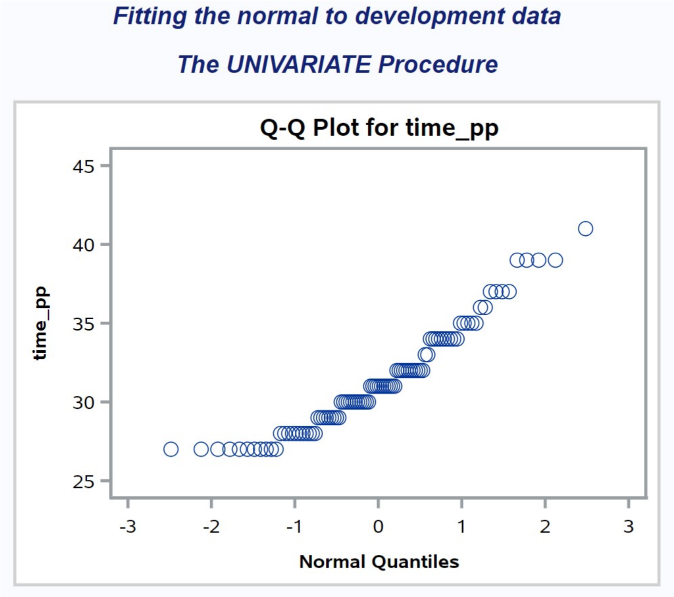
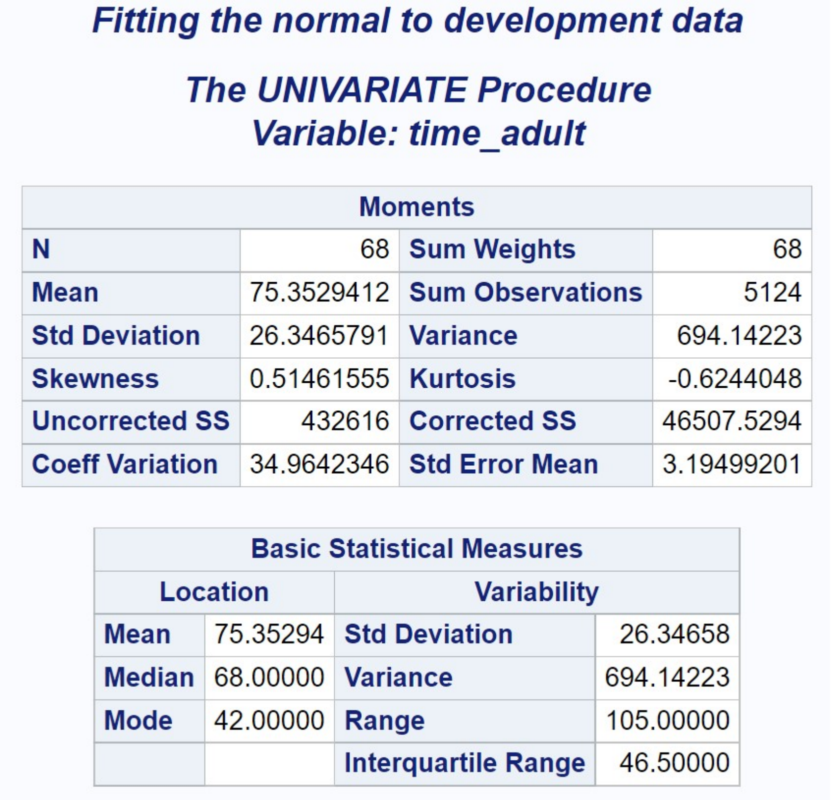


Figure 6.21: `normal_quantile_plot_2.sas - proc univariate`



Fitting the normal to development data

The UNIVARIATE Procedure
Variable: time_adult

Moments			
N	68	Sum Weights	68
Mean	75.3529412	Sum Observations	5124
Std Deviation	26.3465791	Variance	694.14223
Skewness	0.51461555	Kurtosis	-0.6244048
Uncorrected SS	432616	Corrected SS	46507.5294
Coeff Variation	34.9642346	Std Error Mean	3.19499201

Basic Statistical Measures			
Location		Variability	
Mean	75.35294	Std Deviation	26.34658
Median	68.00000	Variance	694.14223
Mode	42.00000	Range	105.00000
		Interquartile Range	46.50000

Figure 6.22: `normal_quantile_plot_2.sas - proc univariate`

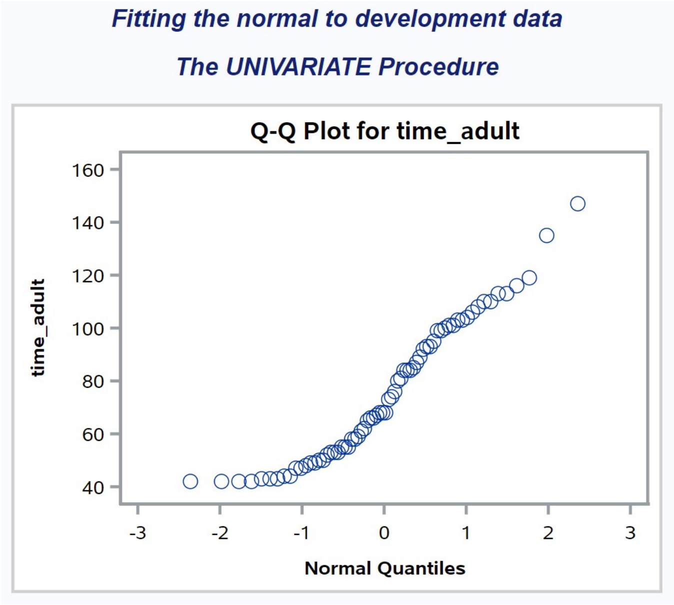


Figure 6.23: `normal_quantile_plot_2.sas - proc univariate`

6.5 References

Mood, A. M., Graybill, F. A. & Boes, D. C. (1974) *Introduction to the Theory of Statistics*. McGraw-Hill, Inc., New York, NY.

Reeve, J. D., Rojas, M. G. & Morales-Ramos, J. A. (2003) Artificial diet and rearing methods for *Thanasimus dubius* (Coleoptera: Cleridae), a predator of bark beetles (Coleoptera: Scolytidae). *Biological Control* 27: 315-322.

SAS Institute Inc. (2016) *Base SAS 9.4 Procedures Guide, Sixth Edition.* SAS Institute Inc., Cary, NC.

Wagner, T. L., Wu, H., Sharpe, P. J. H. & Coulson, R. N. (1984) Modeling distributions of insect development time: A literature review and application of the Weibull function. *Annals of the Entomological Society of America* 77: 475-487.

Wilk, M. B. & Gnanadesikan, R. (1968) Probability plotting methods for the analysis of data. *Biometrika* 55: 1-17.

6.6 Problems

1. A random variable Y has a uniform probability density with $a = 0$ and $b = 2$.

   (a) What is the expected value of Y, or $E[Y]$? What is the variance of Y, or $Var[Y]$?

   (b) What are the 25%, 50%, and 80% quantiles or percentiles of Y?

   (c) Find the probability that $Y < 0.05$.

   (d) Find a symmetric interval centered around $y = 1$ that has a probability of 0.95.

2. Suppose that Y has a normal distribution with $\mu = 1$ and $\sigma^2 = 3$, or $Y \sim N(1, 3)$. Find the following quantities using Table Z.

   (a) The probability that $Y > 2$.

   (b) The probability that $1 < Y < 3$.

   (c) The probability that $Y < 0.5$.

   (d) The probability that Y is not inside the interval given in b.

   (e) A value of y_0 such that the probability that $Y < y_0$ is 0.9.

3. Suppose that Y has a normal distribution with $\mu = 2$ and $\sigma^2 = 4$, or $Y \sim N(2, 4)$. Find the following quantities using Table Z:

   (a) The probability that $Y < 2.5$.

   (b) The probability that $0.5 < Y < 2.5$.

   (c) The probability that $Y < 1$.

   (d) The probability that Y is not inside the interval given in b.

   (e) A value of y_0 such that the probability that $Y < y_0$ is 0.4.

Chapter 7

Expected Value, Variance, and Samples

7.1 Expected value and variance

Previously, we determined the expected value and variance for a random variable Y, which we can think of as a single observation from a distribution. We will now extend these concepts to a linear function of Y and also the sum of n random variables. We will use these results to derive the expected value and variance of the sample mean $\bar{Y}$ and variance s^2, and so describe their basic statistical properties. The idea of an unbiased estimator is also expressed in terms of expected values, and we will show that $\bar{Y}$ and s^2 are unbiased estimators of the theoretical mean and variance of Y, i.e., $E[Y]$ and $Var[Y]$. This is true regardless of the distribution of Y.

We begin by reviewing the definition of expected value and variance. Recall that if Y has a discrete distribution, the expected value (theoretical mean) of Y, or $E[Y]$, is given by the equation

$$E[Y] = \sum_y yP[Y = y] = \sum_y yf(y). \tag{7.1}$$

Here $f(y)$ is the probability distribution of Y, with the summation is taken over all possible values of y. If Y has a continuous distribution, the expected value is defined as the integral

$$E[Y] = \int_{-\infty}^{\infty} yf(y)dy, \tag{7.2}$$

where $f(y)$ is the probability density of Y. For both discrete and continuous random variables, the expected value is essentially a weighted average of all possible values of Y, with the weights being probabilities or densities.

We also defined the theoretical variance of a random variable using expectation. The variance of a random variable Y, denoted by $Var[Y]$, is defined as

$$Var[Y] = E[(Y - E[Y])^2] = \sum_y (y - E[Y])^2 P[Y = y] \tag{7.3}$$

$$= \sum_y (y - E[Y])^2 f(y). \tag{7.4}$$

The variance is a measure of the dispersion of the distribution of Y. The variance of a continuous random variable Y is similarly defined as

$$Var[Y] = E[(Y - E[Y])^2] = \int_{-\infty}^{\infty} (y - E[Y])^2 f(y) dy. \tag{7.5}$$

Table 7.1 summarizes the expected value and variance for the different distributions we have examined so far. These quantities are a function of the parameters in the distribution. Note that for the binomial, Poisson, negative binomial and uniform distributions, there is some relationship between $E[Y]$ and $Var[Y]$, because the formulas share the same parameters. For example, in the Poisson distribution the theoretical mean and variance are both equal to λ. This is not the case for the normal distribution, where the mean and variance are two separate parameters.

Table 7.1: Expected value and variance for five common probability distributions

Distribution	Parameters	$E[Y]$	$Var[Y]$
Binomial	l, p	lp	$lp(1-p)$
Poisson	λ	λ	λ
Negative binomial	m, k	m	$m + m^2/k$
Uniform	a, b	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
Normal	μ, σ^2	μ	σ^2

The significance of this result is that many statistical procedures assume the mean and variance are unrelated, because they are based on the normal distribution. If we wish to apply these procedures to other distributions, we will need to transform the observations to reduce the relationship between the mean and variance. This type of transformation is known as a **variance-stabilizing transformation** (see Chapter 15).

7.2 Linear functions and sums - expected value and variance

Before we turn to samples, we first need to determine the expected value of a linear function of Y. Let Y be a random variable with any distribution, and define a new variable $Y' = aY + b$, where a and b are constants. This is called a linear function of Y because there is a straight-line relationship between Y' and Y. What is the expected value of Y', or $E[Y']$? It can be shown that

$$E[Y'] = E[aY + b] = aE[Y] + b. \tag{7.6}$$

Thus, multiplying a random variable by a constant and then adding another constant just shifts the theoretical mean in the same way (Mood et al. 1974). This result holds for random variables with either a discrete or continuous distribution.

Now suppose we have n random variables of any type, $Y_1, Y_2, \ldots, Y_n$, which may or may not be independent. The random variables could also have unequal means and variances, and even different distributions. What is the expected value of the sum of these variables? One can show that

$$E[Y_1 + Y_2 + \ldots + Y_n] = E[Y_1] + E[Y_2] + \ldots + E[Y_n] = \sum E[Y_i]. \tag{7.7}$$

So, **the expected value of a sum is equal to the sum of the expected values** (Mood et al. 1974).

We will now examine how the theoretical variance is affected by a linear function. Let Y be a variable with any distribution with an associated variance of $Var[Y]$. Define a new random variable $Y' = aY + b$, where a and b are constants. What is the variance of Y', or $Var[Y']$? It can be shown that

$$Var[Y'] = Var[aY + b] = a^2 Var[Y]. \tag{7.8}$$

This implies that a linear function of a random variable increases its variance by a factor of a^2, with b playing no role in the variance. This makes intuitive sense, because multiplying a random variable by a constant (a) should affect its breadth or dispersion, while adding a constant (b) only shifts its location and not its dispersion.

Now suppose we have n random variables of any type, $Y_1, Y_2, \ldots, Y_n$. The random variables can have unequal means and variances, but we will assume they are independent. What is the variance of the sum of these observations? It can be shown that

$$Var[Y_1 + Y_2 + \ldots + Y_n] = Var[Y_1] + Var[Y_2] + \ldots + Var[Y_n] = \sum Var[Y_i]. \tag{7.9}$$

Thus, **the variance of a sum is equal to the sum of the variances** (Mood et al. 1974). As you add more and more random variables together, the variance of the sum also increases. This result only holds when the random variables are independent of each other – if they were dependent a much more complicated formula would be required. This is one advantage of working with a random sample in which the observations are independent, because it simplifies parameter estimation and other statistical procedures (see Chapter 8).

7.3 Sample mean - expected value and variance

We will now use the preceding results to find the expected value and variance of the sample mean. Suppose we have a set of observations $Y_1, Y_2, \ldots, Y_n$ drawn from some statistical population, say the body lengths of n randomly selected individuals. The random variables Y_i are independent, and because they are drawn from the same population, they also have the same expected value $E[Y_i]$ and variance $Var[Y_i]$.

The sample mean is defined using the familiar formula:

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}. \tag{7.10}$$

What is the expected value of the sample mean or $\bar{Y}$? Using our results for sums of variables and linear transformations, we have

$$E[\bar{Y}] = E\left[\frac{\sum Y_i}{n}\right] = \frac{E[\sum Y_i]}{n} = \frac{\sum E[Y_i]}{n} = \frac{nE[Y_i]}{n} = E[Y_i]. \tag{7.11}$$

The expected value of the mean is thus equal to the expected value of the individual variables (Mood et al. 1974).

The fact that $E[\bar{Y}] = E[Y_i]$ means that $\bar{Y}$ is an **unbiased estimator** of the theoretical mean of the distribution of Y_i. In less technical terms, it implies that on average $\bar{Y}$ will be equal to the underlying mean of the random variable Y_i. This is often a desirable property in an estimator, although there are useful biased estimators as well.

We also need to calculate the theoretical variance of the sample mean, written as $Var[\bar{Y}]$. Using the properties of the expected value and variance, we have

$$Var[\bar{Y}] = Var\left[\frac{\sum Y_i}{n}\right] = \frac{Var[\sum Y_i]}{n^2} = \frac{\sum Var[Y_i]}{n^2} = \frac{nVar[Y_i]}{n^2} = \frac{Var[Y_i]}{n}. \tag{7.12}$$

Thus, the variance of the sample mean is the variance of Y_i divided by n (Mood et al. 1974).

What does this result imply? **As you collect larger and larger samples, the variance of the sample mean $\bar{Y}$ becomes smaller.** In other words, $\bar{Y}$ becomes less variable when it includes more data. This result underlies many of the desirable effects of larger sample sizes in statistics, including better estimates of parameters (Chapter 8), smaller confidence intervals (Chapter 9), and statistical tests with more power (Chapter 10).

The standard deviation of the sample mean $\bar{Y}$ is defined to be the square root of the above quantity:

$$\sqrt{Var[\bar{Y}]} = \sqrt{\frac{Var[Y_i]}{n}} = \frac{\sqrt{Var[Y_i]}}{\sqrt{n}}. \tag{7.13}$$

This formula makes it clear that the standard deviation of the mean is a function of the standard deviation of the individual observations and the sample size used in the mean. The common name for this quantity is the **standard error**. In general, a standard error is the standard deviation of a particular statistic, in this case the sample mean $\bar{Y}$.

7.4 Sample variance - expected value

Recall that the sample variance is defined using the formula

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}. \tag{7.14}$$

It can be shown that $E[s^2] = Var[Y_i]$, implying that the sample variance is an unbiased estimator of the underlying variance of Y_i.

It is important to note that all our results for the sample mean $\bar{Y}$ and variance s^2 hold true for any distribution, not just the normal distribution. The basic requirement is that the observations Y_1, $Y_2, \ldots$, Y_n are randomly drawn from some statistical population, implying they are independent and have the same expected value $E[Y_i]$ and variance $Var[Y_i]$.

7.5 Sample calculations and simulation - SAS demo

As an example of these rules of expectation and variance, suppose that Y has a normal distribution with mean $\mu = 1$ and variance $\sigma^2 = 1$, namely $Y \sim N(1, 1)$. Suppose we want to find the expected value and variance of $Y' = 2Y + 1$. Note that Y' is a linear function of Y with $a = 2$ and $b = 1$. Using the formulas for the expected value and variance of a linear function, we have $E[Y'] = aE[Y] + b = 2E[Y] + 1 = 2(1) + 1 = 3$, and also $Var[Y'] = a^2Var[Y] = 2^2Var[Y] = 4(1) = 4$.

Now suppose we have three variables Y_1, Y_2, and Y_3 with the same distribution as above, and assumed to be independent. What is the expected value and variance of the sum of these two variables, $Y_1 + Y_2 + Y_3$? Using the formulas for sums of random variables, we have $E[Y_1 + Y_2 + Y_3] = E[Y_1] + E[Y_2] + E[Y_3] = 1 + 1 + 1 = 3$, and $Var[Y_1 + Y_2 + Y_3] = Var[Y_1] + Var[Y_2] + Var[Y_3] = 1 + 1 + 1 = 3$.

We can also calculate the expected value and variance of the sample mean $\bar{Y}$ for Y_1, Y_2, and Y_3. Using the preceding results, we have $E[\bar{Y}] = E[Y_i] = 1$, and $Var[\bar{Y}] = Var[Y_i]/n = 1/3$.

We can verify that these theoretical rules for the expected value and variance have some basis in reality by conducting an experiment. Recall that the expected value for a random variable can also be thought of as the sample mean $\bar{Y}$ for an infinite number of observations of that random variable. Similarly, its theoretical variance is the sample variance s^2 of an infinite number of observations. It is easy to generate a very large number of observations using SAS, and then compare the result predicted by these theoretical rules with the sample mean and variance of the observations. The SAS program listed below first generates 1,000 observations having the

Table 7.2: Expected value and variance

	Theory		Simulation	
Variable	$E[\cdot]$	$Var[\cdot]$	$\bar{Y}$	s^2
Y	1	1	1.032	0.980
Y'	3	4	3.063	3.919
$Y_1 + Y_2 + Y_3$	3	3	3.052	3.069
$\bar{Y}$	1	1/3	1.017	0.341
s^2	1	-	1.001	-

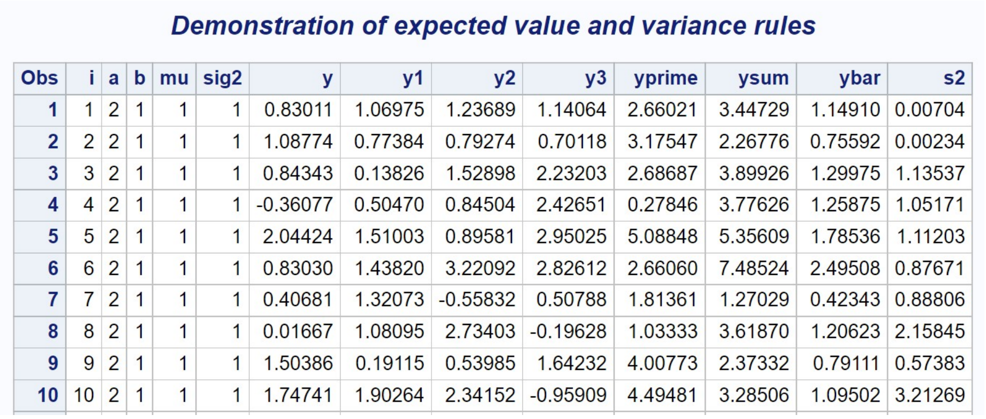
specified distribution $[Y, Y_i \sim N(1,1)]$ in a `data` step. Formulas are then used to calculate Y', $Y_1 + Y_2 + Y_3$, $\bar{Y}$, and s^2. The SAS procedure `proc univariate` is then used to calculate the sample mean and variance of these quantities. See SAS output below.

If the theory involving expected values and variances is correct, it should predict the behavior of the mean and variance in this large sample. A comparison between the results predicted using our expected value formulas and the observed simulation results is given in Table 7.2. The theoretical predictions and sample mean and variance are in close agreement.

Notice also from the SAS output that the distributions of Y', $Y_1 + Y_2 + Y_3$, and $\bar{Y}$ appear to be normally distributed (see Fig. 7.8 - 7.10). In fact, linear functions and sums of normal random variables are always normally distributed, as is the sample mean. This may not be the case for variables with other distributions. We also see that the variance of $\bar{Y}$ is lower than Y (1/3 vs. 1), an important property of this statistic (see Fig. 7.8 vs. 7.10).

SAS Program

```
* Linear.sas;
title ’Demonstration of expected value and variance rules’;
data linear;
    * Loop to generate 1000 random observations;
    do i = 1 to 1000;
        a = 2;
        b = 1;
        * Generate y, y1, y2, y3 with N(1,1) distribution;
        mu = 1; sig2 = 1;
        y = sqrt(sig2)*rannor(0) + mu;
        y1 = sqrt(sig2)*rannor(0) + mu;
        y2 = sqrt(sig2)*rannor(0) + mu;
        y3 = sqrt(sig2)*rannor(0) + mu;
        * Calculate a linear function of y, then sum, mean, and s2;
        yprime = a*y + b;
        ysum = y1 + y2 + y3;
        ybar = ysum/3;
        s2 = ((y1-ybar)**2+(y2-ybar)**2+(y3-ybar)**2)/(3-1);
        output;
    end;
run;
* Print simulated data, first 25 observations;
proc print data=linear(obs=25);
run;
* Descriptive statistics, histograms, and normal quantile plots;
proc univariate data=linear;
    var y yprime ysum ybar s2;
    histogram y yprime ysum ybar s2 / vscale=count normal midpoints=-6 to 12 by 0.5;
    qqplot y yprime ysum ybar s2 / normal;
run;
quit;
```

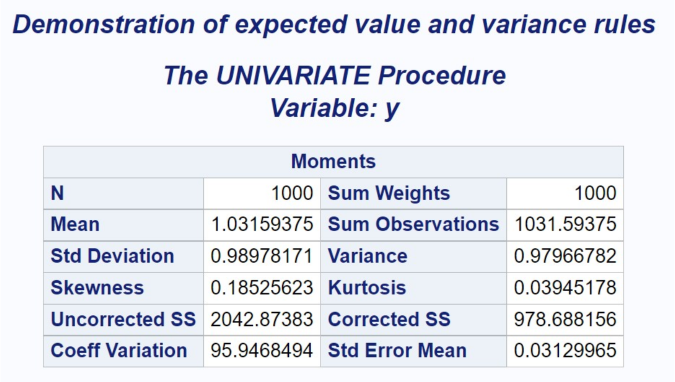


Demonstration of expected value and variance rules

Obs	i	a	b	mu	sig2	y	y1	y2	y3	yprime	ysum	ybar	s2
1	1	2	1	1	1	0.83011	1.06975	1.23689	1.14064	2.66021	3.44729	1.14910	0.00704
2	2	2	1	1	1	1.08774	0.77384	0.79274	0.70118	3.17547	2.26776	0.75592	0.00234
3	3	2	1	1	1	0.84343	0.13826	1.52898	2.23203	2.68687	3.89926	1.29975	1.13537
4	4	2	1	1	1	-0.36077	0.50470	0.84504	2.42651	0.27846	3.77626	1.25875	1.05171
5	5	2	1	1	1	2.04424	1.51003	0.89581	2.95025	5.08848	5.35609	1.78536	1.11203
6	6	2	1	1	1	0.83030	1.43820	3.22092	2.82612	2.66060	7.48524	2.49508	0.87671
7	7	2	1	1	1	0.40681	1.32073	-0.55832	0.50788	1.81361	1.27029	0.42343	0.88806
8	8	2	1	1	1	0.01667	1.08095	2.73403	-0.19628	1.03333	3.61870	1.20623	2.15845
9	9	2	1	1	1	1.50386	0.19115	0.53985	1.64232	4.00773	2.37332	0.79111	0.57383
10	10	2	1	1	1	1.74741	1.90264	2.34152	-0.95909	4.49481	3.28506	1.09502	3.21269

etc.

Figure 7.1: `linear.sas - proc print`

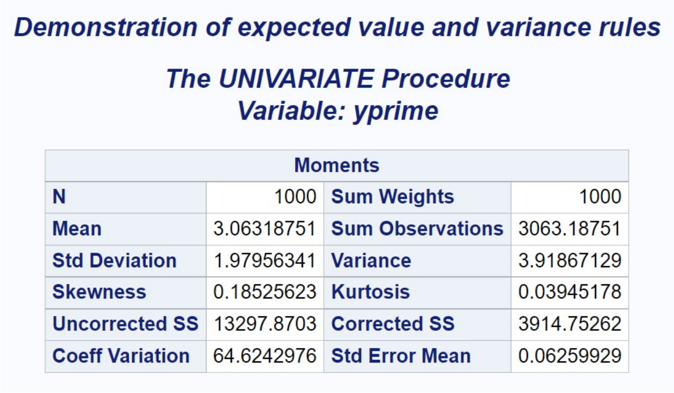


Demonstration of expected value and variance rules

The UNIVARIATE Procedure
Variable: y

Moments			
N	1000	Sum Weights	1000
Mean	1.03159375	Sum Observations	1031.59375
Std Deviation	0.98978171	Variance	0.97966782
Skewness	0.18525623	Kurtosis	0.03945178
Uncorrected SS	2042.87383	Corrected SS	978.688156
Coeff Variation	95.9468494	Std Error Mean	0.03129965

Figure 7.2: `linear.sas - proc univariate`

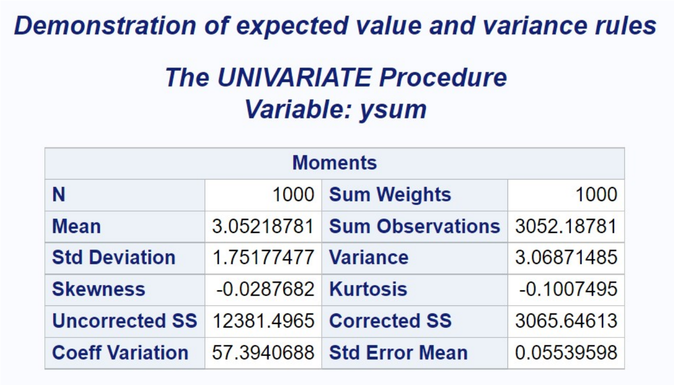


Demonstration of expected value and variance rules

The UNIVARIATE Procedure
Variable: yprime

Moments			
N	1000	Sum Weights	1000
Mean	3.06318751	Sum Observations	3063.18751
Std Deviation	1.97956341	Variance	3.91867129
Skewness	0.18525623	Kurtosis	0.03945178
Uncorrected SS	13297.8703	Corrected SS	3914.75262
Coeff Variation	64.6242976	Std Error Mean	0.06259929

Figure 7.3: `linear.sas - proc univariate`

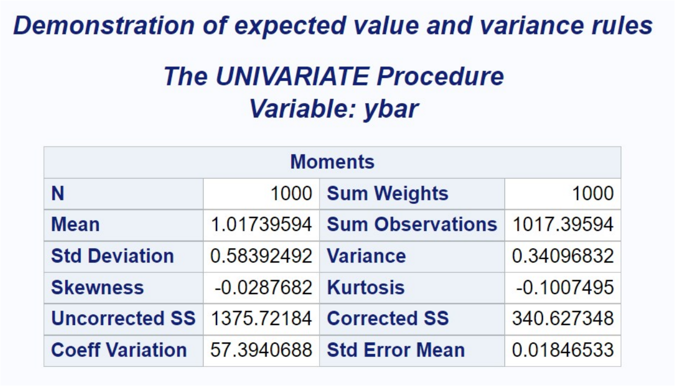


Demonstration of expected value and variance rules

The UNIVARIATE Procedure
Variable: ysum

Moments			
N	1000	Sum Weights	1000
Mean	3.05218781	Sum Observations	3052.18781
Std Deviation	1.75177477	Variance	3.06871485
Skewness	-0.0287682	Kurtosis	-0.1007495
Uncorrected SS	12381.4965	Corrected SS	3065.64613
Coeff Variation	57.3940688	Std Error Mean	0.05539598

Figure 7.4: `linear.sas - proc univariate`

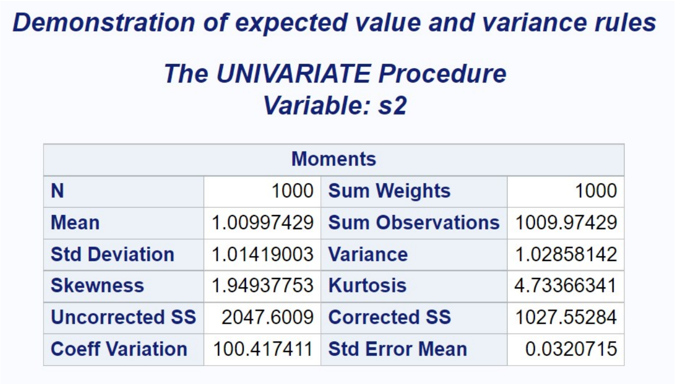


Demonstration of expected value and variance rules

The UNIVARIATE Procedure
Variable: ybar

Moments			
N	1000	Sum Weights	1000
Mean	1.01739594	Sum Observations	1017.39594
Std Deviation	0.58392492	Variance	0.34096832
Skewness	-0.0287682	Kurtosis	-0.1007495
Uncorrected SS	1375.72184	Corrected SS	340.627348
Coeff Variation	57.3940688	Std Error Mean	0.01846533

Figure 7.5: `linear.sas - proc univariate`



Demonstration of expected value and variance rules

The UNIVARIATE Procedure
Variable: s2

Moments			
N	1000	Sum Weights	1000
Mean	1.00997429	Sum Observations	1009.97429
Std Deviation	1.01419003	Variance	1.02858142
Skewness	1.94937753	Kurtosis	4.73366341
Uncorrected SS	2047.6009	Corrected SS	1027.55284
Coeff Variation	100.417411	Std Error Mean	0.0320715

Figure 7.6: `linear.sas - proc univariate`

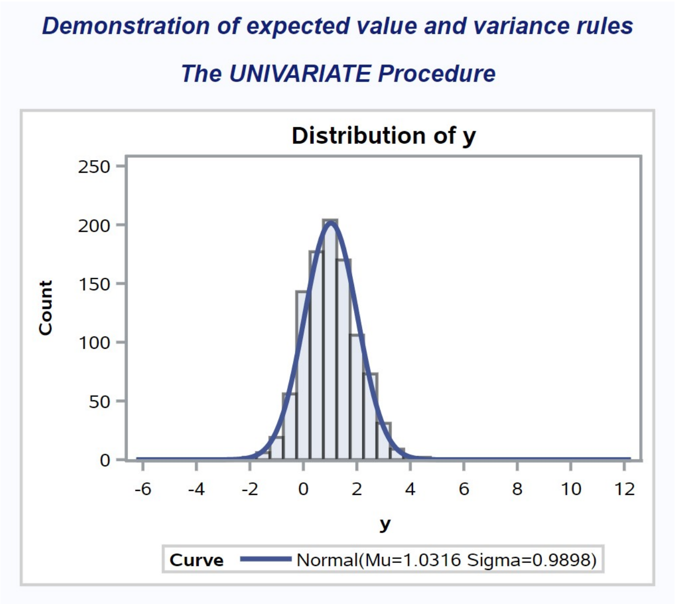


Figure 7.7: `linear.sas - proc univariate`

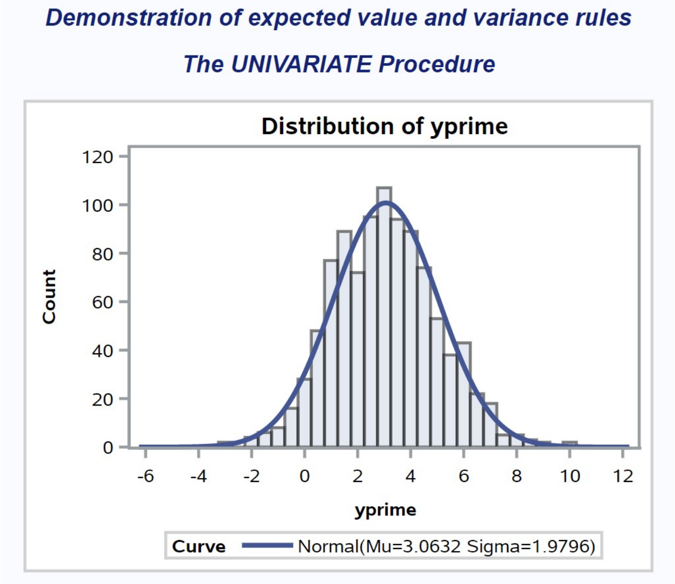


Figure 7.8: `linear.sas - proc univariate`

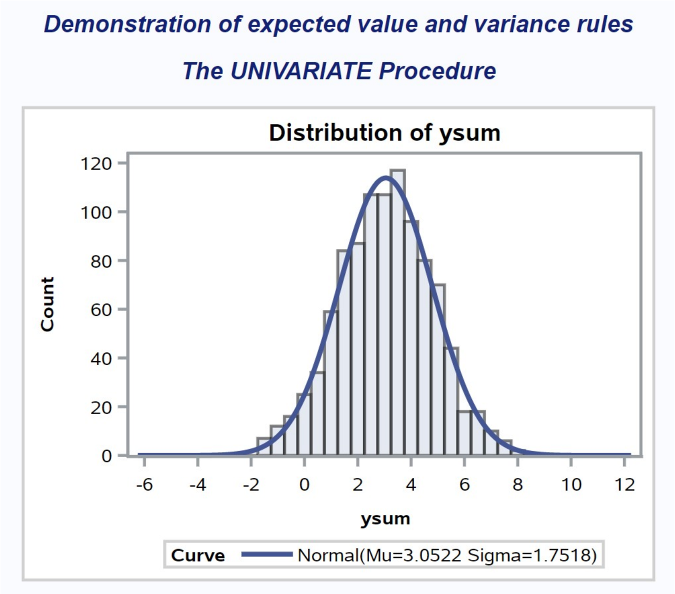


Figure 7.9: `linear.sas - proc univariate`

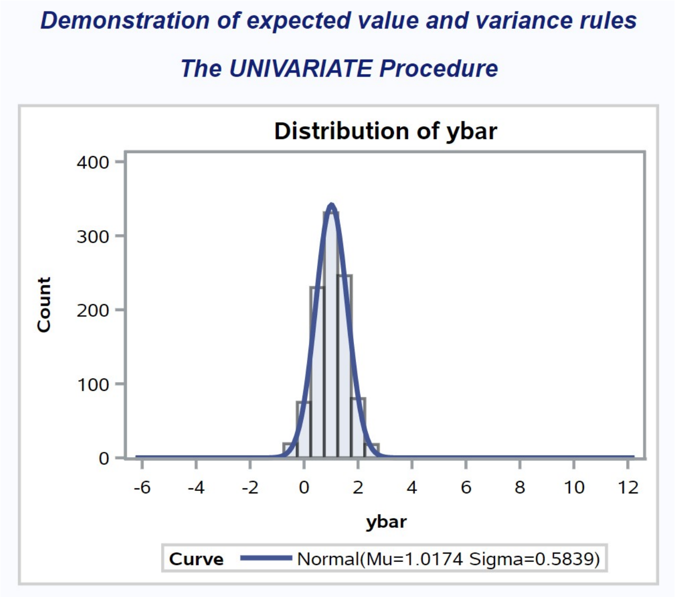


Figure 7.10: `linear.sas - proc univariate`

7.6 Central limit theorem

Suppose we randomly draw a sample $Y_1, Y_2, \ldots, Y_n$ of size n from some statistical population. In this situation, the observations are independent and have a common expected value $E[Y_i]$ and variance $Var[Y_i]$. **They may have any probability distribution, known or unknown.**

The **central limit theorem** states that the distribution of the sample mean of these random variables, namely $\bar{Y}$, approaches a normal distribution with mean $E[Y_i]$ and variance $Var[Y_i]/n$ as the sample size n becomes large (Mood et al. 1974). In particular, we have $\bar{Y} \sim N(E[Y_i], Var[Y_i]/n)$ for large n. The central limit theorem also holds for sums of random variables, and in this case we have $\sum Y_i \sim N(nE[Y_i], nVar[Y_i])$ for large n. **These results are true for any probability distribution - $\bar{Y}$ and $\sum Y_i$ will have a normal distribution for large sample sizes.** Note also that the variance of $\bar{Y}$ decreases as the sample size n increases. We would also expect this from our earlier results concerning the variance of $\bar{Y}$.

7.6.1 Central limit theorem - SAS demo

The operation of the central limit theorem can be demonstrated in a simple experiment using a SAS program (see below). The program models Y as a Poisson random variable with $\lambda = 1$, implying $E[Y_i] = 1$ and $Var[Y_i] = 1$. Sample means are then generated for different sample sizes, ranging from $n = 1$ to $n = 50$, in a SAS `data` step. A total of 1,000 sample means are generated for each value of n in the simulation. The program then used `proc univariate` to calculate summary statistics for these data, as well as histograms and normal quantile plots (not shown). See SAS output below.

Examining the histograms, we see that as n increases the distribution of $\bar{Y}$ approaches the normal distribution. A sample size of $n = 50$ appears sufficient to produce a distribution almost indistinguishable from normal. What is especially interesting here is that fact that the Poisson is a discrete random variable, yet the distribution of $\bar{Y}$ approaches the normal distribution, a continuous random variable.

We also observe that the variance of $\bar{Y}$ decreases as the sample size n increases, as predicted by the central limit theorem and our earlier results on the variance of $\bar{Y}$. See Table 7.3.

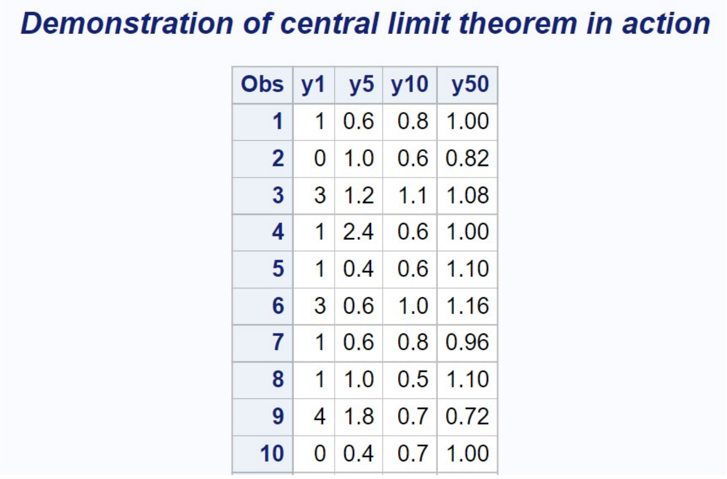
Table 7.3: Mean and variance of $\bar{Y}$

	Theory		Simulation	
n	$E[Y_i]$	$Var[Y_i]/n$	Mean of $\bar{Y}$	Variance of $\bar{Y}$
1	1.000	1.000	0.993	1.036
5	1.000	0.200	0.994	0.213
10	1.000	0.100	1.001	0.111
50	1.000	0.020	0.995	0.019

SAS Program

```
* central_limit_theorem.sas;
title 'Demonstration of central limit theorem in action';
data cntrlmt;
    * Loop to generate 1000 random observations;
    do i = 1 to 1000;
        * A single Poisson observations with lambda = 1;
        y1 = ranpoi(0,1);
        * Mean of 5 Poisson observations;
        y5 = 0;
        do j = 1 to 5;
            y5 = y5 + ranpoi(0,1);
        end;
        y5 = y5/5;
        * Mean of 10 Poisson observations;
        y10 = 0;
        do j = 1 to 10;
            y10 = y10 + ranpoi(0,1);
        end;
        y10 = y10/10;
        * Mean of 50 Poisson observations;
        y50 = 0;
        do j = 1 to 50;
            y50 = y50 + ranpoi(0,1);
        end;
        y50 = y50/50;
        * Mean of 100 Poisson observations;
        output;
    end;
    drop i j;
run;
* Print simulated data (first 25 observations);
proc print data=cntrlmt(obs=25);
```

```
run;
* Descriptive statistics, histograms, and normal quantile plots;
proc univariate data=cntrlmt;
    var y1 y5 y10 y50;
    histogram y1 y5 y10 y50 / vscale=count normal
    qqplot y1 y5 y10 y50 / normal;
    symbol1 h=3;
run;
quit;
```

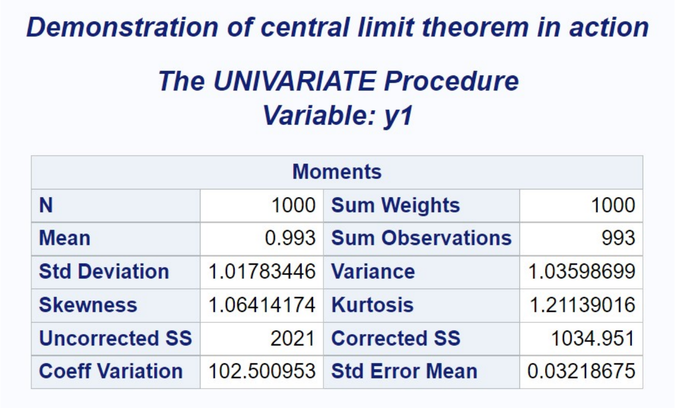


Demonstration of central limit theorem in action

Obs	y1	y5	y10	y50
1	1	0.6	0.8	1.00
2	0	1.0	0.6	0.82
3	3	1.2	1.1	1.08
4	1	2.4	0.6	1.00
5	1	0.4	0.6	1.10
6	3	0.6	1.0	1.16
7	1	0.6	0.8	0.96
8	1	1.0	0.5	1.10
9	4	1.8	0.7	0.72
10	0	0.4	0.7	1.00

etc.

Figure 7.11: `central_limit_theorem.sas - proc print`

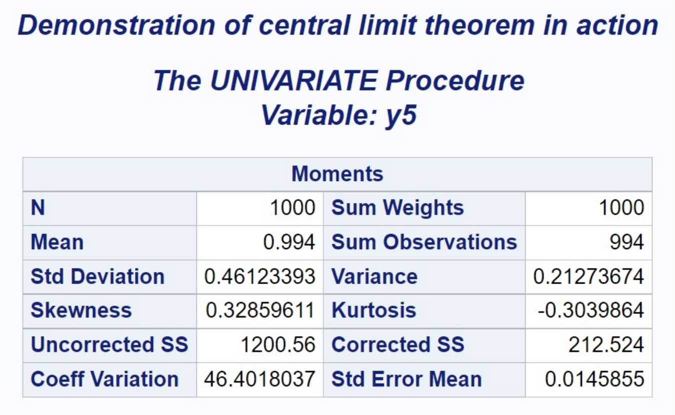


Demonstration of central limit theorem in action

The UNIVARIATE Procedure
Variable: y1

Moments			
N	1000	Sum Weights	1000
Mean	0.993	Sum Observations	993
Std Deviation	1.01783446	Variance	1.03598699
Skewness	1.06414174	Kurtosis	1.21139016
Uncorrected SS	2021	Corrected SS	1034.951
Coeff Variation	102.500953	Std Error Mean	0.03218675

Figure 7.12: `central_limit_theorem.sas - proc univariate`

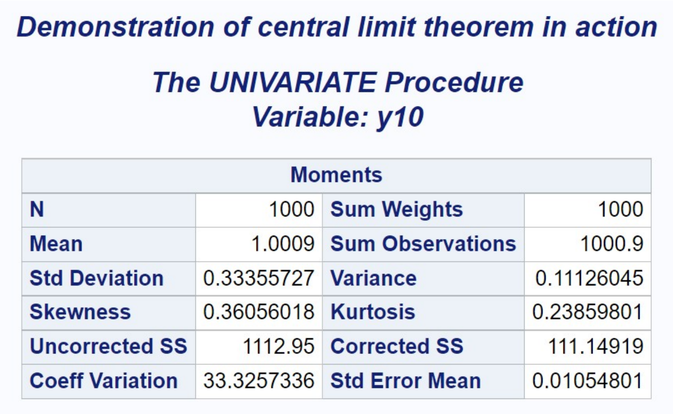


Demonstration of central limit theorem in action

The UNIVARIATE Procedure
Variable: y5

Moments			
N	1000	Sum Weights	1000
Mean	0.994	Sum Observations	994
Std Deviation	0.46123393	Variance	0.21273674
Skewness	0.32859611	Kurtosis	-0.3039864
Uncorrected SS	1200.56	Corrected SS	212.524
Coeff Variation	46.4018037	Std Error Mean	0.0145855

Figure 7.13: `central_limit_theorem.sas - proc univariate`

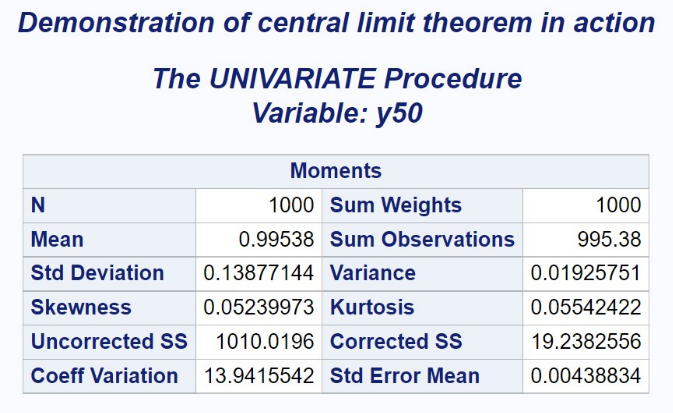


Demonstration of central limit theorem in action

The UNIVARIATE Procedure
Variable: y10

Moments			
N	1000	Sum Weights	1000
Mean	1.0009	Sum Observations	1000.9
Std Deviation	0.33355727	Variance	0.11126045
Skewness	0.36056018	Kurtosis	0.23859801
Uncorrected SS	1112.95	Corrected SS	111.14919
Coeff Variation	33.3257336	Std Error Mean	0.01054801

Figure 7.14: `central_limit_theorem.sas - proc univariate`



Demonstration of central limit theorem in action

The UNIVARIATE Procedure
Variable: y50

Moments			
N	1000	Sum Weights	1000
Mean	0.99538	Sum Observations	995.38
Std Deviation	0.13877144	Variance	0.01925751
Skewness	0.05239973	Kurtosis	0.05542422
Uncorrected SS	1010.0196	Corrected SS	19.2382556
Coeff Variation	13.9415542	Std Error Mean	0.00438834

Figure 7.15: `central_limit_theorem.sas - proc univariate`

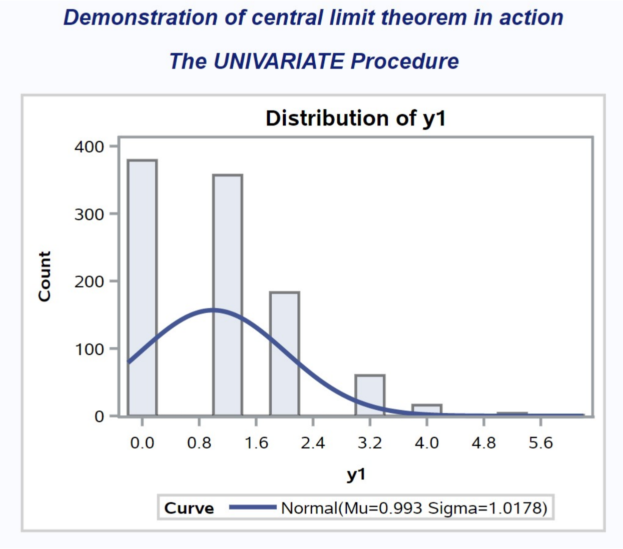


Figure 7.16: `central_limit_theorem.sas - proc univariate`

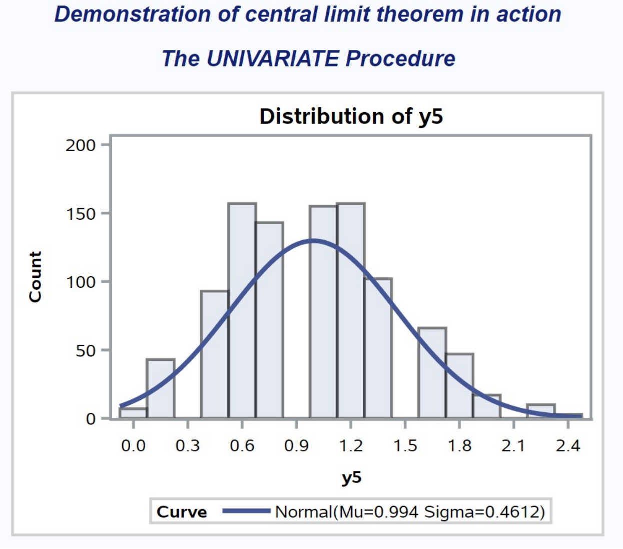


Figure 7.17: `central_limit_theorem.sas - proc univariate`

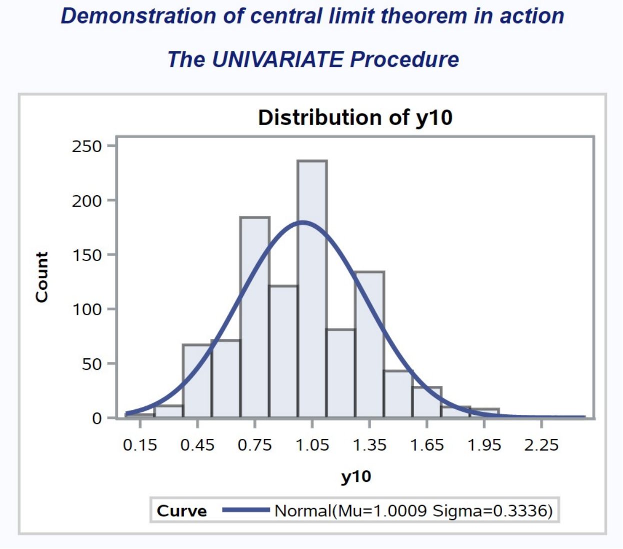


Figure 7.18: `central_limit_theorem.sas - proc univariate`

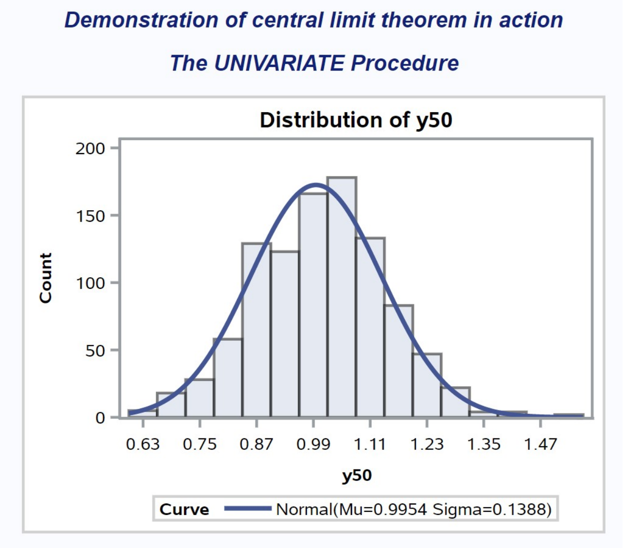


Figure 7.19: `central_limit_theorem.sas - proc univariate`

7.7 Applications of the central limit theorem

The central limit theorem provides a potential explanation why so many biological variables like the length of an organism and other continuous variables are apparently normal in distribution. These variables are often under the control of multiple genes and environmental factors that can behave like sums and means of random variables, and so their combined effect should generate a normal distribution of outcomes by the central limit theorem (Hartl & Clark 1989).

The theorem also applies to measurements of ecological variables like population density. To estimate population density, we often average the results of several quadrats (or whatever sampling units) to yield a single number for a given location. By the central limit theorem, these average densities will have a normal distribution for sufficiently large n.

Most of the statistical methods we will study are based on the assumption that the observations in a study or experiment have a normal distribution. This would seem a risky assumption, since many natural processes yield random variables that are not strictly normal, some examples being count data that are better modeled using the binomial and Poisson distributions. However, the tests themselves are often based on means that are assumed to have a normal distribution. The central limit theorem guarantees these means are normal provided sample sizes are sufficiently large. Thus, statistical tests based on normality should be valid for non-normal data given large enough sample sizes (see Stewart-Oaten 1995 for further discussion).

The central limit may not be sufficient to guarantee normality for smaller sample sizes, and so other approaches may be needed. One possibility would be a transformation of the observations to make their distribution closer to normal (Chapter 15). If that fails, there are nonparametric statistical procedures (Chapter 16) that are valid for any distribution, as well as ones that allow the use of other probability distributions.

7.8 References

Hartl, D. L. & Clark, A. G. *Principles of Population Genetics, Second Edition.* Sinauer Associates, Inc., Sunderland, MA.

Mood, A. M., Graybill, F. A. & Boes, D. C. (1974) *Introduction to the Theory of Statistics.* McGraw-Hill, Inc., New York, NY.

Stewart-Oaten, A. (1995) Rules and judgments in statistics: three examples. *Ecology* 76: 2001-2009.

7.9 Problems

1. Let Y_1, Y_2, and Y_3 be three independent random variables with $E[Y_i] = 2$ and $Var[Y_i] = 1$. Using the rules for expected value and variance, calculate the expected value and variance of the following quantities:

   (a) $3Y_1 + 1$.

   (b) $Y_1 + Y_2 + Y_3$.

   (c) $(Y_1 + Y_2 + Y_3)/3$.

2. Suppose that Y_1, Y_2, and Y_3 are three independent random variables, with $E[Y_i] = 3$ and $Var[Y_i] = 2$. Using the rules for expected value and variance, calculate the expected value and variance of the following quantities:

   (a) $0.5Y_2 + 2$.

   (b) $(Y_1 + Y_2 + Y_3)/3$.

   (c) $2(Y_1 + Y_2) + 3$.

3. The exponential distribution is often used to model the time until an event happens, such as the radioactive decay of an atom or mortality processes in population models. The probability density for the exponential distribution is defined as

   $$f(y) = \frac{e^{-y/\lambda}}{\lambda} \tag{7.15}$$

   for $y \geq 0$. The distribution has one parameter, λ, which is the mean decay time ($E[Y] = \lambda$). A single random observation with an exponential distribution can be generated in SAS using the expression `ranexp(0)*lambda`. Modify the program `central_limit.sas` so that is generates exponential observations instead of Poisson ones, using $\lambda = 2$. Discuss how the distribution of $\bar{Y}$ changes as the sample size increases.

Chapter 8

Sampling and Estimation

We discuss in this chapter two topics that are critical to most statistical analyses. The first is **random sampling**, which is a method for obtaining observations from a statistical population that has many advantages. After obtaining a random sample, the next step of the analysis is the selection of a probability distribution to model the observations, such as the Poisson or normal distributions. One then seeks to **estimate the parameters** of these distributions (λ, μ, σ^2, etc.) using the information contained in the random sample, the second topic of this chapter. We will examine one common method of parameter estimation called maximum likelihood.

8.1 Random samples

A basic assumption of many statistical procedures is that the observations are a **random sample** from a statistical population (see Chapter 3). A sample from a statistical population is a random sample if (1) each element of the population has an equal probability of being sampled, and (2) the observations in the sample are independent (Thompson 2002). This definition has a number of implications. It implies that a random sample will resemble the statistical population from which it is drawn, especially as the sample size n increases, because each element of the population has an equal chance of being in the sample. Random sampling also implies there is no connection or relationship between the observations in the sample, because they are independent of one another.

What are some ways of obtaining a random sample? Suppose we are interested in the distribution of body length for insects of a given species, say in a particular forest. This defines the statistical population of interest. One way to obtain a random sample would be to number all the insects, and then write the numbers on pieces of paper and place them in a hat. After mixing the pieces, one would draw n numbers from the hat (without peeking) and collect only those insects corresponding to these numbers. This method of sampling would yield a random sample, because each individual would have an equal probability of being selected, and the observations would also be independent. For many insect species this method would be impractical, however, because they can be difficult to find and number. It would be more useful for statistical populations where the number of elements is known and they can be uniquely identified, as in surveys of human populations (Thompson 2002).

A more feasible way of sampling insects would be to place traps in the forest and in this way sample the population. If we want to successfully approximate a random sample with our trapping scheme, however, some knowledge of the biology of the organism is essential. For example, suppose that insect size varies in space because of differences in food plants or microclimate. A single trap deployed at only one location could therefore yield insects different in length than those in the overall population. A better sampling scheme would deploy multiple traps at several locations within the forest. The location of the traps could be randomly chosen to avoid conscious or unconscious biases by the trapper, such as deploying the traps close to a road for convenience. There is also the problem that insects susceptible to trapping could differ in length from the general population. This implies that the population actually sampled could differ from the target statistical population, and a careful analyst would consider this possibility. Thus, the biology of the organism plays an integral role in designing an appropriate sampling scheme.

8.2 Parameter estimation

Suppose we have obtained a random sample from some statistical population, say the lengths of insects trapped in a forest, or the counts of the insects in each trap. The first step faced by the analyst is to chose a probability distribution to model the data in the sample. For insect lengths, a normal distribution could be a plausible model, while counts of the insects per trap might have a Poisson distribution. Once a distribution has been selected, the next task is to estimate the parameters of the distribution using the sample data. The dominant method of parameter estimation in modern statistics is **maximum likelihood**. This method has a number of desirable statistical properties although it can also be computationally intensive.

Maximum likelihood obtains estimates of the parameters using a mathematical function (see Chapter 2) known as the likelihood function. The likelihood function gives the probability or density of the observed data as a function of the parameters in the probability distribution. For example, the likelihood function for Poisson data would be a function of the Poisson parameter λ. We then seek the maximum value of the likelihood function (hence the name maximum likelihood) across the potential range of parameter values. The parameter values that maximize the likelihood are the maximum likelihood estimates. In other words, **the maximum likelihood estimates are the parameter values that give the largest probability (or probability density) for the observed data.**

8.2.1 Maximum likelihood for Poisson data

We will first illustrate estimation using maximum likelihood with a random sample drawn from a statistical population where the observations are Poisson. For simplicity, let $n = 3$ and suppose the observed values are $Y_1 = 8$, $Y_2 = 5$, and $Y_3 = 6$. We begin by calculating the probability of observing this sample, which in fact is its likelihood function. Because we have a random sample, the Y_i values are independent of each other, and so this probability is the product of the probability for each Y_i. We have

$$\begin{aligned} L(\lambda) &= P[Y_1 = 8] \times P[Y_2 = 5] \times P[Y_3 = 6] && (8.1) \\ &= \frac{e^{-\lambda}\lambda^8}{8!} \times \frac{e^{-\lambda}\lambda^5}{5!} \times \frac{e^{-\lambda}\lambda^6}{6!} && (8.2) \end{aligned}$$

The notation $L(\lambda)$ is used for likelihood functions and indicates the likelihood is a function of the parameter λ of the Poisson distribution. The method of maximum likelihood estimates λ by finding the value of λ that maximizes this function (Mood *et al.* 1974). Note that the location of the maximum will vary with the data in the sample.

We can find the maximum likelihood estimate graphically by plotting $L(\lambda)$ as function of λ (Fig. 8.1). For these particular data values, the maximum occurs at $\lambda = 6.3$, and so the maximum likelihood estimate (often abbreviated MLE) of λ is this value. This is also the value of $\bar{Y}$ for these data, which suggests that $\bar{Y}$ might be the maximum likelihood estimator of λ in general.

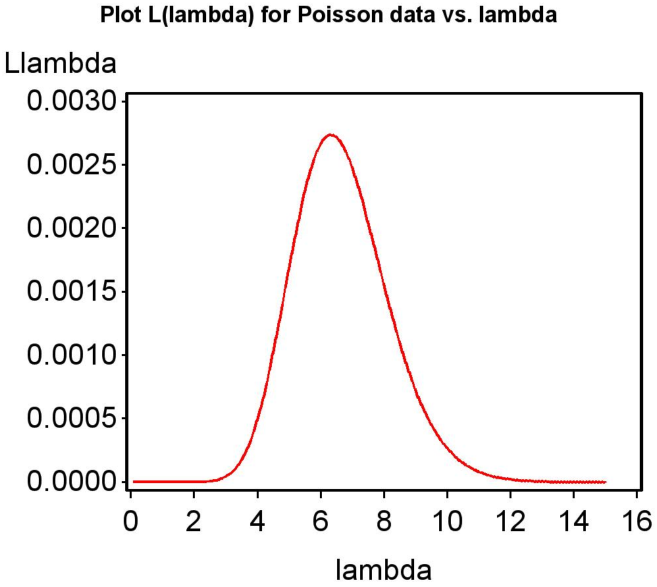


Figure 8.1: Plot of $L(\lambda)$ vs. λ

For readers interested in the mathematics, this also can be shown using derivatives. Let y_1, y_2, and y_3 be the observed values of Y_1, Y_2, and Y_3. The likelihood function can then be written as

$$L(\lambda) = \frac{e^{-\lambda}\lambda^{y_1}}{y_1!} \times \frac{e^{-\lambda}\lambda^{y_2}}{y_2!} \times \frac{e^{-\lambda}\lambda^{y_3}}{y_3!} = \frac{e^{-3\lambda}\lambda^{y_1+y_2+y_3}}{y_1!y_2!y_3!} \tag{8.3}$$

We want to find the maximum of $L(\lambda)$ (Eq. 8.3), which should occur when the derivative of this function with respect to λ equals zero. This follows because the derivative is the slope of a function, and at the maximum the slope is equal to zero. Differentiating $L(\lambda)$ with respect to λ and simplifying, we obtain

$$\frac{dL(\lambda)}{d\lambda} = \frac{e^{-3\lambda}}{y_1!y_2!y_3!}\left[(y_1+y_2+y_3)\lambda^{y_1+y_2+y_3-1} - 3\lambda^{y_1+y_2+y_3}\right]. \tag{8.4}$$

This derivative can only equal zero if the term in square brackets is zero:

$$\left[(y_1 + y_2 + y_3)\lambda^{y_1+y_2+y_3-1} - 3\lambda^{y_1+y_2+y_3}\right] = 0 \tag{8.5}$$

or

$$(y_1 + y_2 + y_3)\lambda^{y_1+y_2+y_3-1} = 3\lambda^{y_1+y_2+y_3}. \tag{8.6}$$

Canceling the quantity $\lambda^{y_1+y_2+y_3}$ from both sides of this equation, we find that

$$(y_1 + y_2 + y_3)\lambda^{-1} = 3, \tag{8.7}$$

or

$$\hat{\lambda} = \frac{y_1 + y_2 + y_3}{3}. \tag{8.8}$$

Note that this is the sample mean $\bar{Y}$ for $n = 3$, and it is can be shown that $\bar{Y}$ is the maximum likelihood estimator of λ for any n. Statisticians often write the estimator of a parameter like λ using the notation $\hat{\lambda}$, pronounced 'λ-hat.' An **estimator** can be thought of as the formula or recipe for obtaining an estimate of a parameter, with the **estimate** itself obtained by plugging actual data values into the estimator.

8.2.2 Poisson likelihood function - SAS demo

We can use a SAS program to further illustrate the behavior of the likelihood function for Poisson data (see program listing below). In particular, we will show how $L(\lambda)$ changes as the observed data and the sample size n changes. The program first generates n random Poisson observations for a specified Poisson parameter value of $\lambda = 6$ (`mu_parameter = 6`). It then plots $L(\lambda)$ across a range of λ values. In this scenario we actually know the underlying value of λ and can see how well maximum likelihood estimates its value. See SAS program below.

The program makes extensive use of loops in the data step, to generate the Poisson data and also values of the likelihood function for different values of λ. One new feature of this program is the use of a SAS macro variable(SAS Institute Inc. 2016). In this case, a macro variable labeled `n` is defined and assigned a value of 3 using the command

```
%let n = 3;
```

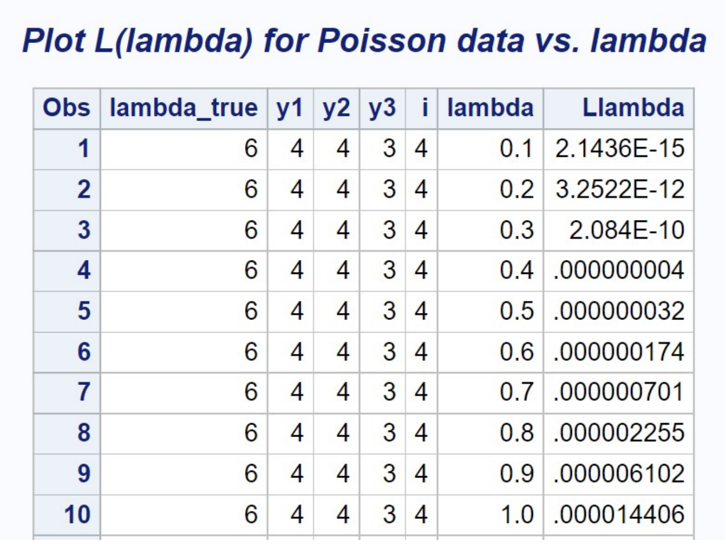
We can then refer to this value throughout the program using the notation `&n`. Otherwise, if we wanted to change the sample size n in the program we would have to type in a new value everywhere sample size is used in the calculations.

SAS program

```
* likepois_random.sas;
title "Plot L(lambda) for Poisson data vs. lambda";
data likepois;
    * Generate n random Poisson observations with parameter lambda;
    %let n = 3;
    lambda_parameter = 6;
    array ydata (&n) y1-y&n;
    do i=1 to &n;
        ydata(i) = ranpoi(0,lambda_parameter);
    end;
    * Find likelihood as function of lambda;
    do lambda=0.1 to 15 by 0.1;
        Llambda = 1;
        do i=1 to &n;
            Llambda = Llambda*pdf(’poisson’,ydata(i),lambda);
        end;
        output;
    end;
run;
* Print data;
proc print data=likepois;
run;
* Plot likelihood as a function of lambda;
proc gplot data=likepois;
    plot Llambda*lambda=1 / vaxis=axis1 haxis=axis1;
    symbol1 i=join v=none c=red width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
quit;
```

Examining the SAS output and graphs from the first two runs of the program (Fig. 8.3, 8.4), we see that the likelihood function is different. This is because the observed data are different for each run. The peak in the likelihood function always occurs at the value of $\bar{Y}$ for each data set, and this is the maximum likelihood estimate of λ.

The last run shows the effect of increasing the sample size in the program, from $n = 3$ to $n = 10$. Note that the peak of the likelihood function lies quite close to the specified value $\lambda = 6$ (Fig. 8.5). This illustrates an important property of maximum likelihood estimators - they converge on the true value as $n \to \infty$. This property is known as consistency in mathematical statistics.



Plot L(lambda) for Poisson data vs. lambda

Obs	lambda_true	y1	y2	y3	i	lambda	Llambda
1	6	4	4	3	4	0.1	2.1436E-15
2	6	4	4	3	4	0.2	3.2522E-12
3	6	4	4	3	4	0.3	2.084E-10
4	6	4	4	3	4	0.4	.000000004
5	6	4	4	3	4	0.5	.000000032
6	6	4	4	3	4	0.6	.000000174
7	6	4	4	3	4	0.7	.000000701
8	6	4	4	3	4	0.8	.000002255
9	6	4	4	3	4	0.9	.000006102
10	6	4	4	3	4	1.0	.000014406

etc.

Figure 8.2: `likepois_random.sas - proc print`

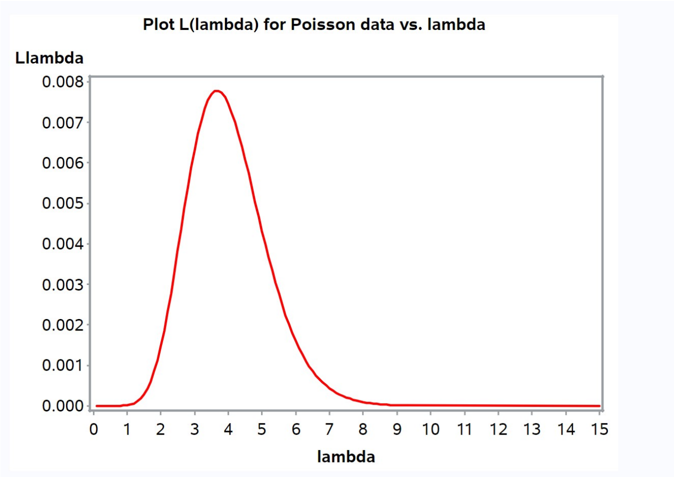


Figure 8.3: `likepois_random.sas - proc gplot` $(n = 3)$

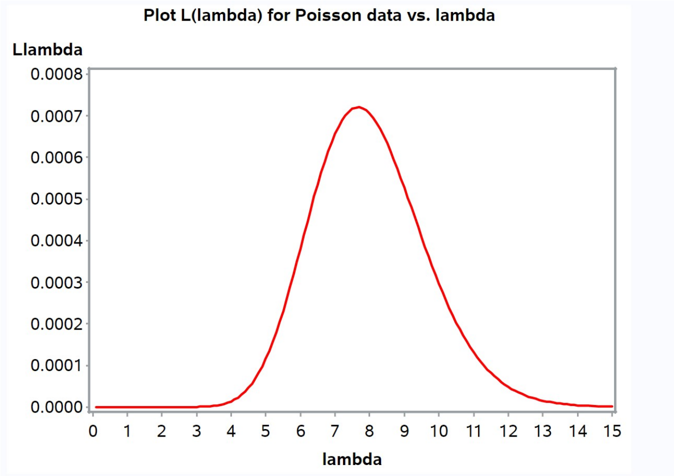


Figure 8.4: `likepois_random.sas - proc gplot` $(n = 3)$

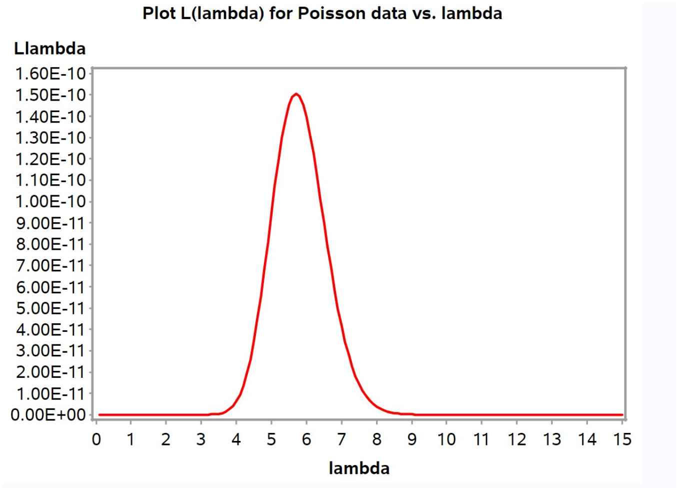


Figure 8.5: `likepois_random.sas - proc gplot` $(n = 10)$

8.2.3 Maximum likelihood for normal data

Now suppose we draw a random sample from a population with a normal distribution, such as body lengths, etc. For simplicity, let $n = 3$ again and the observed values be $Y_1 = 4.5$, $Y_2 = 5.4$, and $Y_3 = 5.3$. The likelihood function in this case is the probability density values for the observed data:

$$L(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(4.5-\mu)^2}{\sigma^2}} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(5.4-\mu)^2}{\sigma^2}} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(5.3-\mu)^2}{\sigma^2}}. \tag{8.9}$$

Note that the terms in the likelihood for normal data are probability densities, instead of probabilities as with Poisson data.

We can find the maximum likelihood estimate graphically by plotting $L(\mu, \sigma^2)$ as function of μ and σ^2. The likelihood function in this case describes a dome-shaped surface (Fig. 8.6). With these particular data, the maximum occurs at about $\mu = 5.07$ and $\sigma^2 = 0.16$, and so these are the maximum likelihood estimates of μ and σ^2.

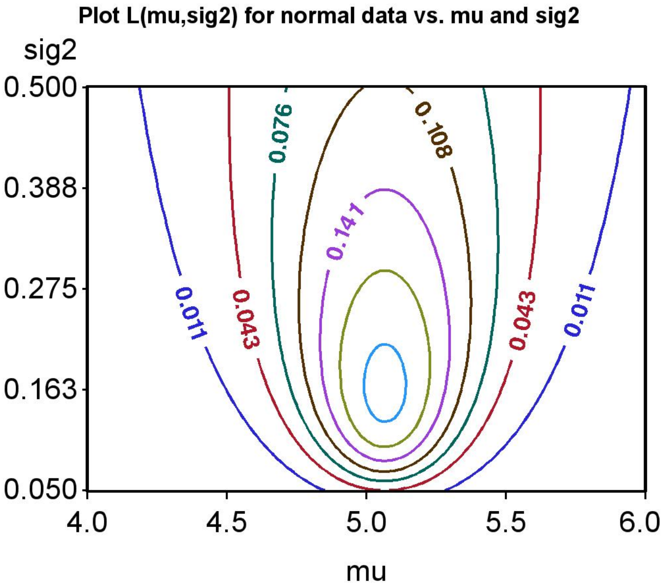


Figure 8.6: Plot of $L(\mu, \sigma^2)$ vs. μ and σ^2

Using a bit of calculus, it can be shown that the maximum likelihood estimators of these parameters are, for any sample size n:

$$\hat{\mu} = \bar{Y} \tag{8.10}$$

and

$$\hat{\sigma}^2 = \frac{\Sigma_{i=1}^n (Y_i - \bar{Y})^2}{n}. \tag{8.11}$$

Note that does not quite equal the sample variance s^2, which uses $n-1$ (rather than n) in the denominator:

$$s^2 = \frac{\Sigma_{i=1}^n (Y_i - \bar{Y})^2}{n-1}. \tag{8.12}$$

Recall that s^2 is an unbiased estimator of σ^2, and so $\hat{\sigma}^2$ derived using maximum likelihood is actually a biased estimator of σ^2. It would consistently generate values that underestimate σ^2 because n is greater than $n-1$. For cases like this one where bias is known, it is common to use a bias-corrected version of the maximum likelihood estimator (i.e., $n-1$ rather than n in the denominator).

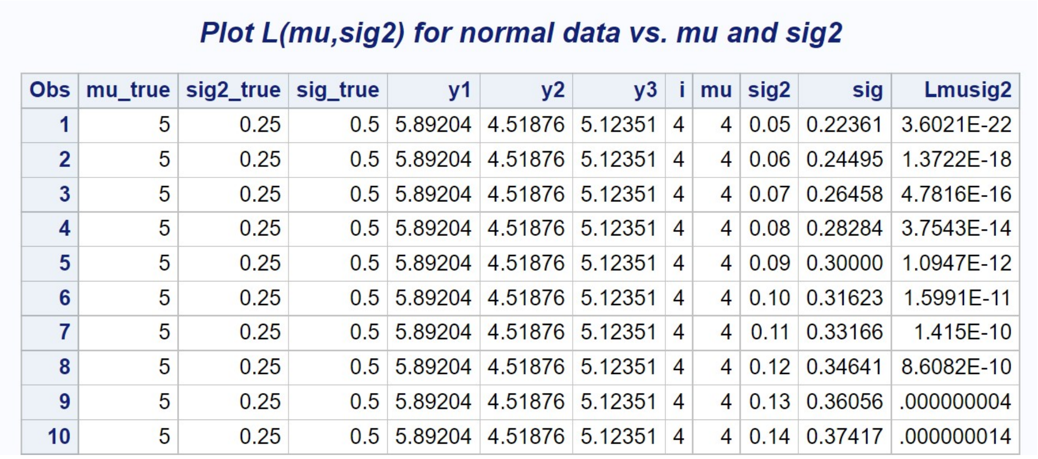
8.2.4 Normal likelihood function - SAS demo

We will use another SAS program to illustrate the behavior of the likelihood function for normal data. The program first generates n random normal observations for a specified, known value of $\mu = 5$ and $\sigma^2 = 0.25$. It then plots the likelihood function across a range of possible μ and σ^2 values. See SAS program below.

Examining the SAS output and graphs from the first two runs of the program (Fig. 8.8, 8.9), we see that the likelihood function changes with the observed data. The peak always occurs at $\hat{\mu}$ and $\hat{\sigma}^2$ for each data set. The last run shows the effect of increasing the sample size from $n = 3$ to $n = 10$. Note that the peak of the likelihood function lies quite close to the specified values of $\mu = 5$ and $\sigma^2 = 0.25$ (Fig. 8.10). This again illustrates the consistency of maximum likelihood estimates.

SAS program

```
* likenorm_random.sas;
title "Plot L(mu,sig2) for normal data vs. mu and sig2";
data likenorm;
    * Generate n random normal observations with parameters mu and sig2;
    %let n = 3;
    mu_parameter = 5; sig2_parameter = 0.25; sig_parameter = sqrt(sig2_parameter);
    array ydata (&n) y1-y&n;
    do i=1 to &n;
        ydata(i) = mu_parameter + sig_parameter*rannor(0);
    end;
    * Find likelihood as a function of mu and sig2;
    do mu=4 to 6 by 0.01;
        do sig2=0.05 to 0.5 by 0.01;
            sig = sqrt(sig2);
            Lmusig2 = 1;
            do i=1 to &n;
                Lmusig2 = Lmusig2*pdf(’normal’,ydata(i),mu,sig);
            end;
            output;
        end;
    end;
run;
* Print data, first 25 observations;
proc print data=likenorm(obs=25);
run;
* Plot likelihood as a function of mu and sig2;
* Contour plot version;
proc gcontour data=likenorm;
    plot sig2*mu=Lmusig2 / autolabel nolegend vaxis=axis1 haxis=axis1;
    symbol1 height=1.5 font=swissb width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
quit;
```



Plot L(mu,sig2) for normal data vs. mu and sig2

Obs	mu_true	sig2_true	sig_true	y1	y2	y3	i	mu	sig2	sig	Lmusig2
1	5	0.25	0.5	5.89204	4.51876	5.12351	4	4	0.05	0.22361	3.6021E-22
2	5	0.25	0.5	5.89204	4.51876	5.12351	4	4	0.06	0.24495	1.3722E-18
3	5	0.25	0.5	5.89204	4.51876	5.12351	4	4	0.07	0.26458	4.7816E-16
4	5	0.25	0.5	5.89204	4.51876	5.12351	4	4	0.08	0.28284	3.7543E-14
5	5	0.25	0.5	5.89204	4.51876	5.12351	4	4	0.09	0.30000	1.0947E-12
6	5	0.25	0.5	5.89204	4.51876	5.12351	4	4	0.10	0.31623	1.5991E-11
7	5	0.25	0.5	5.89204	4.51876	5.12351	4	4	0.11	0.33166	1.415E-10
8	5	0.25	0.5	5.89204	4.51876	5.12351	4	4	0.12	0.34641	8.6082E-10
9	5	0.25	0.5	5.89204	4.51876	5.12351	4	4	0.13	0.36056	.000000004
10	5	0.25	0.5	5.89204	4.51876	5.12351	4	4	0.14	0.37417	.000000014

etc.

Figure 8.7: `likenorm_random.sas - proc print`

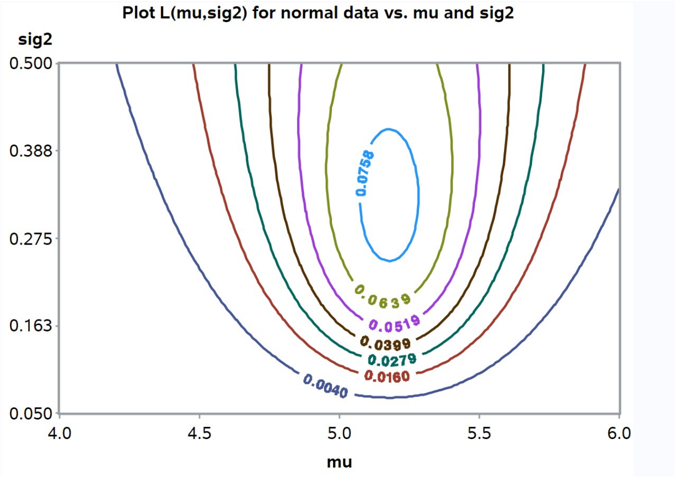


Figure 8.8: `likenorm_random.sas - proc gcontour` $(n = 3)$

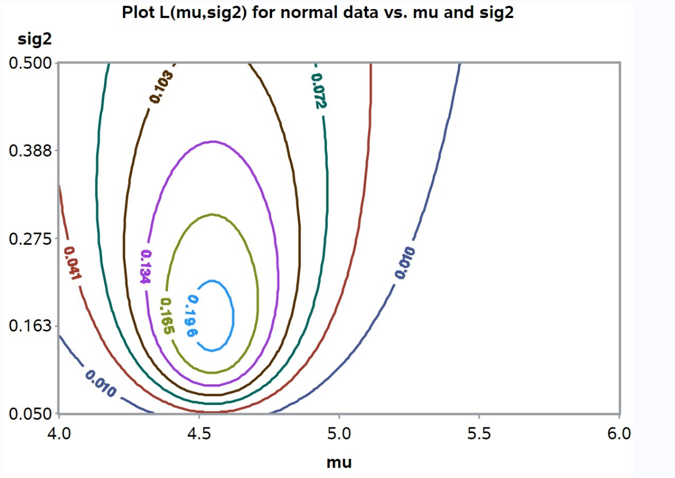


Figure 8.9: `likenorm_random.sas - proc gcontour` $(n = 3)$

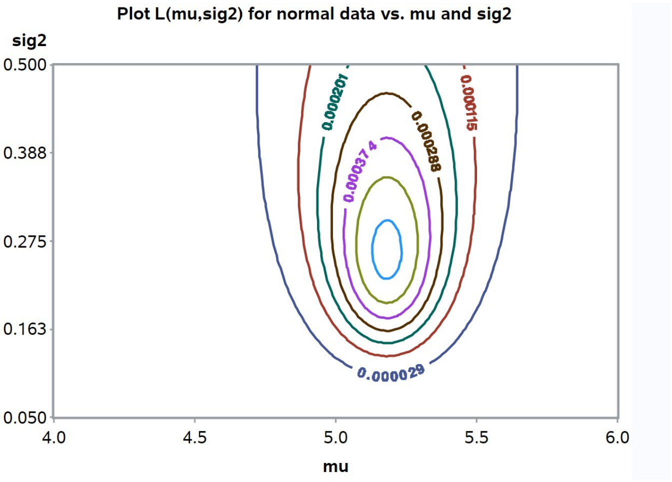


Figure 8.10: `likenorm_random.sas - proc gcontour` $(n = 10)$

8.3 Optimality of maximum likelihood estimates

Why should we use maximum likelihood estimates? There are other methods of parameter estimation, but maximum likelihood estimates are optimal in a number of ways (Mood *et al.* 1974). We have already seen that they are **consistent**, approaching the true parameter values as sample size increases. Increasing the sample size also reduces the variance of these estimators. We can observe this behavior for $\hat{\mu} = \bar{Y}$, the estimator of μ for the normal distribution. Recall that the variance of $\bar{Y}$ is σ^2/n, which decreases for large n. Maximum likelihood estimates are also **asymptotically unbiased**, meaning their expected value approaches the true value of the parameter as the sample size n increases. We can see this in operation for $\hat{\sigma}^2$ (Eq. 8.11), the maximum likelihood estimator of σ^2, vs. s^2 (Eq. 8.12), an unbiased estimator of σ^2. Note that the difference between n vs. $n - 1$ in the denominator becomes very small as n increases. Finally, maximum likelihood estimates are **asymptotically normal**, meaning their distribution approaches the normal distribution for large n.

There are other uses for the likelihood function besides parameter estimation. We will later see how the likelihood function can be used to develop statistical tests called likelihood ratio tests. Many of the statistical tests we will study are actually likelihood ratio tests. Likelihood methods provide an essential tool for developing new statistical procedures, provided that we can specify a probability distribution for the data.

8.4 References

Mood, A. M., Graybill, F. A. & Boes, D. C. (1974) *Introduction to the Theory of Statistics.* McGraw-Hill, Inc., New York, NY.

Thompson, S. K. (2002) *Sampling.* John Wiley & Sons, Inc., New York, NY.

SAS Institute Inc. (2016) *SAS 9.4 Macro Language: Reference, Fifth Edition.* SAS Institute Inc., Cary, NC.

8.5 Problems

1. The exponential distribution is a continuous distribution that is used to model the time until a particular event occurs. For example, the time when a radioactive particle decays is often modeled using an exponential distribution. If a variable Y has a exponential distribution, then its probability density is given by the formula

   $$f(y) = \frac{e^{-y/\lambda}}{\lambda} \tag{8.13}$$

   for $y \geq 0$. The distribution has one parameter, λ, which is the mean decay time ($E[Y] = \lambda$).

   (a) Use SAS and the program `fplot.sas` to plot the exponential probability density with $\lambda = 2$, for $0 \leq y \leq 5$. Attach your SAS program and output.

   (b) Suppose you have a sample of four observations y_1, y_2, y_3 and y_4 from the exponential distribution. What would be the likelihood function for these observations?

   (c) Plot the likelihood function for $y_1 = 1$, $y_2 = 2$, $y_3 = 2$ and $y_4 = 3$ over a range of λ values. Show that the maximum occurs at $\hat{\lambda} = \bar{Y}$, the maximum likelihood estimator of λ. Attach your SAS program and output.

2. The geometric distribution is a discrete distribution that is used to model the time until a particular event occurs. Consider tossing a coin – the number of tosses before a head appears would have a geometric distribution. If a variable Y has a geometric distribution, then the probability that Y takes a particular value y is given by the formula

   $$P[Y = y] = f(y) = p(1-p)^y \tag{8.14}$$

   where p is the probability of observing the event on a particular trial, and $y = 0, 1, 2, \ldots, \infty$. The distribution has only one parameter, p.

   (a) Use SAS and the program `fplot.sas` to plot this probability distribution for $p = 0.5$, for $y = 0, 1, \ldots, 10$. Attach your SAS program and output.

(b) Suppose you have a sample of three observations y_1, y_2, and y_3 from the geometric distribution. What would be the likelihood function for these observations?

(c) Plot the likelihood function for $y_1 = 1$, $y_2 = 2$, and $y_3 = 3$ over a range of p values. Show that the maximum occurs at $\hat{p} = 1/(\bar{Y} + 1)$, the maximum likelihood estimator of p. Attach your SAS program and output.

Chapter 9

Confidence Intervals

In the preceding chapter, we examined the maximum likelihood method for estimating the parameters of a statistical population, using a random sample from that population. For example, if we have a sample from a population with a normal distribution, we can estimate the parameter μ of this population using the sample mean $\bar{Y}$. We will now examine a common method for characterizing the precision of these estimates, known as **confidence intervals**. Given an estimate $\bar{Y}$ of μ, say, we will learn how to calculate an interval that will contain the true population μ with a certain probability. A narrow interval indicates the parameter μ is reliably estimated, while a broad one indicates substantial uncertainty as to its value.

9.1 Preliminaries to confidence intervals

We now discuss some material that is essential for the construction of confidence intervals and later in hypothesis testing. We first review some results from Chapter 8 on parameter estimation for the normal distribution, then derive some new results. We then examine some distributions associated with sampling from the normal distributions, not surprisingly called **sampling distributions**.

9.1.1 Parameters and estimates

Confidence intervals are based on estimates of population parameters, such as μ and σ^2 for populations with a normal distribution. Our previous results on parameter estimation suggest that $\bar{Y}$ and s^2 are reasonable estimators of μ and σ^2. The sample standard deviation $s = \sqrt{s^2}$ is typically used to estimate the population standard deviation σ.

We also want to estimate the variance and standard deviation of the sample mean $\bar{Y}$. Recall that for a random sample Y_1, Y_2, ... Y_n with any distribution,

$$Var[\bar{Y}] = \frac{Var[Y_i]}{n} \tag{9.1}$$

where $Var[Y_i]$ is the variance of Y_i (Chapter 7). For a random sample where the observations are normal, this translates to

$$Var[\bar{Y}] = \frac{\sigma^2}{n} \tag{9.2}$$

because $Var[Y_i] = \sigma^2$ for the normal. If we use s^2 to estimate σ^2, we can therefore estimate $Var[\bar{Y}]$ using s^2/n and $\sigma/\sqrt{n}$ using $s/\sqrt{n}$.

The table below summarizes the different parameters, their estimators, and common terminology for these quantities:

Table 9.1: Parameters and their estimators

Parameter	Estimator	Terminology
μ	$\bar{Y}$	Sample mean
σ^2	s^2	Sample variance
σ	s	Sample standard deviation
$\frac{\sigma^2}{n}$	$\frac{s^2}{n}$	Sample variance of the mean
$\frac{\sigma}{\sqrt{n}}$	$\frac{s}{\sqrt{n}}$	Standard error of the mean

Recall that the term standard error always refers to the standard deviation of a statistic, such as $\bar{Y}$. The term standard deviation used without qualification usually means the standard deviation s of items in a random sample from a population.

9.1.2 Sampling distributions

In this section, we will first examine the probability distribution of the estimator $\bar{Y}$. We then examine the distributions of some quantities involving $\bar{Y}$ and the sample variance s^2, known as sampling distributions. These sampling distributions will be used to construct confidence intervals and also play an important role in hypothesis testing (Chapter 10).

Distribution of $\bar{Y}$

Suppose we have a random sample $Y_1, Y_2, ..., Y_n$ from a statistical population with a normal distribution, in particular that $Y_i \sim N(\mu, \sigma^2)$ and are independent of each other. It can be shown that

$$\bar{Y} \sim N\left(\mu, \frac{\sigma^2}{n}\right). \tag{9.3}$$

Thus, the sample mean of normal observations also has a normal distribution with the same mean μ, but with variance equal to σ^2/n, not σ^2 (Mood *et al.* 1974).

Note that the distribution of $\bar{Y}$ will be approximately normal for any distribution provided n is large, thanks to the central limit theorem (see Chapter 7). Thus, for large sample sizes we have $\bar{Y} \sim N(E[Y], Var[Y]/n)$ for any probability distribution. This result has important statistical implications. **Confidence intervals and hypothesis testing procedures often assume that $\bar{Y}$ is normally distributed, and this will be approximately true if n is sufficiently large.** These statistical procedures are therefore robust to departures from normality in the data for large n.

We also learned earlier that if $Y \sim N(\mu, \sigma^2)$, then the transformed variable $(Y - \mu)/\sigma$ has a standard normal distribution, or $(Y - \mu)/\sigma = Z \sim N(0, 1)$. Combining these two results, we find that

$$\frac{\bar{Y} - \mu}{\sqrt{\sigma^2/n}} = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1) \tag{9.4}$$

Thus, the quantity $\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$ has a standard normal distribution. We will use this sampling distribution to obtain a confidence interval for μ, for the case where σ^2 is known from other information.

We will also need to find certain intervals with a specified probability using the standard normal distribution, in order to construct confidence intervals. In general, we will need to find a positive value c such that

$$P[-c_\alpha < Z < c_\alpha] = 1 - \alpha \tag{9.5}$$

for this purpose, where typically $\alpha = 0.05$ or 0.01. The values of c_α that satisfy this probability are often called **critical values**, a term that also applies to other probability distributions. We use the notation c_α because this quantity depends on the value of α. To find c_α, we first express this probability in terms of Table Z. We have

$$\begin{aligned} P[-c_\alpha < Z < c_\alpha] &= P[Z < c_\alpha] - P[Z < -c_\alpha] && (9.6)\\ &= P[Z < c_\alpha] - (1 - P[Z < c_\alpha]) && (9.7)\\ &= 2P[Z < c_\alpha] - 1. && (9.8) \end{aligned}$$

If we set $2P[Z < c_\alpha] - 1 = 1 - \alpha$ and rearrange, we get

$$P[Z < c_\alpha] = (2 - \alpha)/2 = 1 - \alpha/2. \qquad (9.9)$$

Therefore, we examine Table Z for a value of c_α such that $P[Z < c_\alpha] = 1 - \alpha/2$. For $\alpha = 0.05$, we would look for $c_{0.05}$ such that $P[Z < c_{0.05}] = 1 - 0.05/2 = 0.975$ and find that $c_{0.05} = 1.96$ is the answer. Similarly, for $\alpha = 0.01$ we seek $c_{0.01}$ such that $P[Z < c_{0.01}] = 1 - 0.01/2 = 0.995$. There is no value in Table Z that gives quite this probability, although we can see 2.57 and 2.58 are close. The exact answer is $c_{0.01} = 2.576$.

t distribution

Another important sampling distribution is the t distribution. This distribution has a single parameter, called the degrees of freedom, that governs the shape of the distribution. It can be shown that the quantity

$$\frac{\bar{Y} - \mu}{s/\sqrt{n}} \sim t_{n-1} \qquad (9.10)$$

(Mood *et al.* 1974). Here the symbol 't_{n-1}' stands for the t distribution with $n-1$ degrees of freedom, where n is the sample size in $\bar{Y}$. Degrees of freedom is often abbreviated as '*df*'.

The t distribution resembles the standard normal distribution in being bell-shaped, except that it has more probability in the tails and less in the center of the distribution (Fig. 9.1). Roughly speaking, the t distribution has heavier tails than the normal because $\bar{Y}$ and s are both random quantities in Eq. 9.10, making their ratio more variable than for Eq. 9.4 where only $\bar{Y}$ is random. However, as $n \to \infty$ the t distribution does approach the standard normal distribution. We will use this sampling distribution to obtain a confidence interval for μ, when σ^2 is estimated using the sample variance s^2.

What is the origin of the term degrees of freedom? Recall that the sample standard deviation s is obtained from the sample variance, calculated using the formula

$$s^2 = \frac{\Sigma_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}. \tag{9.11}$$

Notice that the sample variance s^2 is composed of terms of the form $Y_i - \bar{Y}$. Although there are n of these terms, they also sum to zero ($\Sigma_i^n(Y_i - \bar{Y}) = 0$). This implies that if $n-1$ terms are known, we can always determine the remaining term because of this relationship, implying there are really only $n-1$ free, independent terms in s^2 (Mood et al. 1974). Hence the name degrees of freedom.

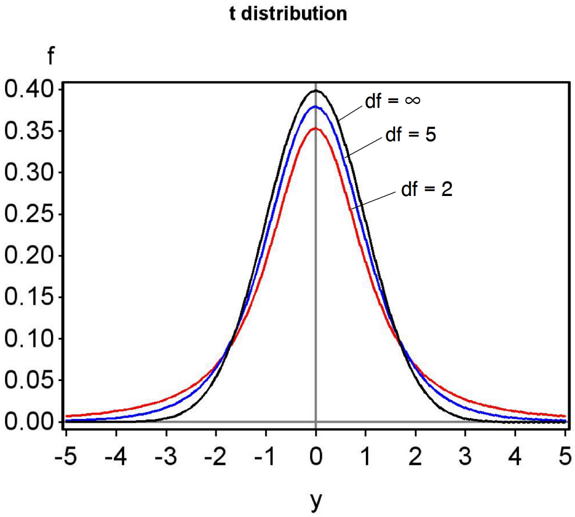


Figure 9.1: Plot of the t distribution for different degrees of freedom

Table T gives the quantiles of the t distribution for different values of the degrees of freedom and the cumulative probability p. We will also need to find intervals of the form

$$P[-c_{\alpha,df} < T < c_{\alpha,df}] = 1 - \alpha, \tag{9.12}$$

where $c_{\alpha,df}$ is a positive number, T has a t distribution, for $\alpha = 0.05$ or 0.01. We use the notation $c_{\alpha,df}$ because this quantity will depend on both α and the degrees of freedom. We proceed as before by expressing this probability in terms of Table T. We have

$$\begin{aligned} P[-c_{\alpha,df} < T < c_{\alpha,df}] &= P[T < c_{\alpha,df}] - P[T < -c_{\alpha,df}] && (9.13) \\ &= P[T < c_{\alpha,df}] - (1 - P[T < c_{\alpha,df}]) && (9.14) \\ &= 2P[T < c_{\alpha,df}] - 1. && (9.15) \end{aligned}$$

If we set $2P[T < c_{\alpha,df}] - 1 = 1 - \alpha$ and rearrange, we get

$$2(1 - P[T < c_{\alpha,df}]) = \alpha. \tag{9.16}$$

Because $P[T < c_{\alpha,df}]$ is essentially p for this table, we simply look across the row corresponding to $2(1 - p)$ at the top and find the column corresponding to α. For $\alpha = 0.05$, we see that for $df = 10$ the answer is $c_{0.05,10} = 2.228$. For $\alpha = 0.01$ and $df = 10$, the answer is $c_{0.01,10} = 3.169$.

χ^2 distribution

One other common sampling distribution is the χ^2 (chi-square) distribution, which also has a parameter called the degrees of freedom. It can be shown that the quantity

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1} \tag{9.17}$$

(Mood *et al.* 1974). Here the symbol 'χ^2_{n-1}' stands for a χ^2 distribution with $n - 1$ degrees of freedom. The degrees of freedom parameter controls the shape of the χ^2 distribution (Fig. 9.2). The χ^2 distribution is only defined for positive values, because s^2 is always positive, and its distribution shifts to the right (large values become more likely) as n and the degrees of freedom increases. We will use this sampling distribution to obtain a confidence interval for σ^2 and σ.

Table C gives the quantiles of the χ^2 distribution for different values of the degrees of freedom and the cumulative probability p. We will need to find the probabilities for certain intervals, but this is more complicated with the χ^2 distribution because it is asymmetrical, unlike the normal or t distributions. In this case, we want to find two positive numbers $c_{\alpha/2,df}$ and $c_{1-\alpha/2,df}$ such that

$$P[c_{\alpha/2,df} < X < c_{1-\alpha/2,df}] = 1 - \alpha, \tag{9.18}$$

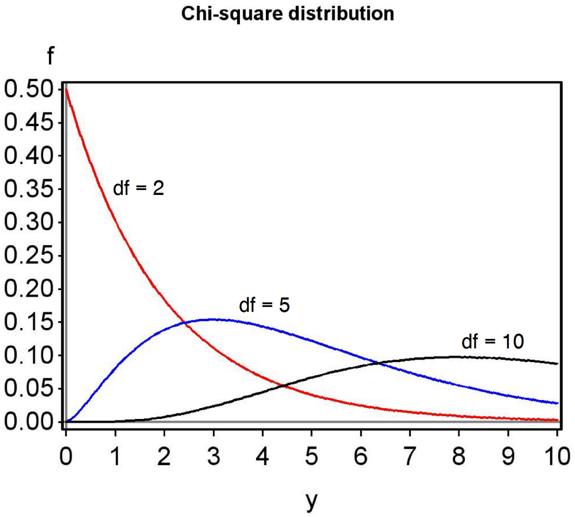


Figure 9.2: Plot of the χ^2 distribution for different degrees of freedom

where X has a χ^2 distribution and $\alpha = 0.05$ or $\alpha = 0.01$. The subscripts $\alpha/2$ and $1-\alpha/2$ for c essentially correspond to values of p in Table C. This gives the correct probability because

$$P[c_{\alpha/2,df} < X < c_{1-\alpha/2,df}] = P[X < c_{1-\alpha/2,df}] - P[X < c_{\alpha/2,df}] \tag{9.19}$$
$$= 1 - \alpha/2 - \alpha/2 = 1 - \alpha. \tag{9.20}$$

To see how these values are obtained from Table C, suppose that $\alpha = 0.05$ and $df = 10$. To find $c_{\alpha/2,df} = c_{0.05/2,10} = c_{0.025,10}$, we look in the column for $p = 0.025$ and row for $df = 10$, and obtain $c_{0.025,10} = 3.247$. To find $c_{1-\alpha/2,df} = c_{1-0.05/2,10} = c_{0.975,10}$, we look in the column for $p = 0.975$ and row for $df = 10$, and obtain $c_{0.975,10} = 20.483$.

Now suppose that $\alpha = 0.01$. Using the same technique, we find that $c_{\alpha/2,df} = c_{0.01/2,10} = c_{0.005,10} = 2.156$, and $c_{1-\alpha/2,df} = c_{1-0.01/2,10} = c_{0.995,10} = 25.188$.

9.2 Confidence intervals

We now have the information needed to calculate confidence intervals. We will begin with a simple but unrealistic case, finding a confidence interval for

μ when σ^2 is known through other means. This case is unrealistic because σ^2 is almost always estimated from the data, but the calculations are simple and illustrate a general method for finding confidence intervals. We then turn to finding a confidence intervals for μ, and then σ^2, where all parameters are estimated from the data.

9.2.1 Confidence intervals for μ when σ^2 is known

We will use the fact that the quantity $\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$ has a standard normal distribution to find a confidence interval for μ. Suppose that α is given and we have found c_α such that

$$P\left[-c_\alpha < Z < c_\alpha\right] = 1 - \alpha. \tag{9.21}$$

(see previous section). Substituting $\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$ for Z we obtain

$$P\left[-c_\alpha < \frac{\bar{Y}-\mu}{\sigma/\sqrt{n}} < c_\alpha\right] = 1 - \alpha. \tag{9.22}$$

Multiplying both sides by $\sigma/\sqrt{n}$ gives you

$$P\left[-c_\alpha \frac{\sigma}{\sqrt{n}} < \bar{Y} - \mu < c_\alpha \frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha. \tag{9.23}$$

Multiplying all parts inside the brackets by -1 reverses the signs and inequalities to give

$$P\left[c_\alpha \frac{\sigma}{\sqrt{n}} > \mu - \bar{Y} > -c_\alpha \frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha. \tag{9.24}$$

We now add to $\bar{Y}$ to all parts inside the brackets to give

$$P\left[\bar{Y} + c_\alpha \frac{\sigma}{\sqrt{n}} > \mu > \bar{Y} - c_\alpha \frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha, \tag{9.25}$$

or equivalently

$$P\left[\bar{Y} - c_\alpha \frac{\sigma}{\sqrt{n}} < \mu < \bar{Y} + c_\alpha \frac{\sigma}{\sqrt{n}}\right] = 1 - \alpha. \tag{9.26}$$

We call the terms $\bar{Y} - c_\alpha \frac{\sigma}{\sqrt{n}}$ and $\bar{Y} + c_\alpha \frac{\sigma}{\sqrt{n}}$ the lower and upper $100(1-\alpha)\%$ confidence limits for μ (Mood et al. 1974). Confidence intervals are often reported in the form $(\bar{Y} - c_\alpha \frac{\sigma}{\sqrt{n}}, \bar{Y} + c_\alpha \frac{\sigma}{\sqrt{n}})$. Note that the center of the confidence interval is at $\bar{Y}$, our estimate of μ. This interval would be expected to include the true value of μ with a probability of $1 - \alpha$, because this was the probability set in Eq. 9.21.

It is common practice to set $\alpha = 0.05$, which corresponds to a $100(1 - 0.05)\% = 95\%$ confidence interval. For this case, we would have $c_\alpha = c_{0.05} = 1.96$ (see previous section). Therefore, the 95% confidence interval would be

$$(\bar{Y} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{Y} + 1.96\frac{\sigma}{\sqrt{n}}). \tag{9.27}$$

We would expect this interval to include the true μ with a probability of 0.95, or 95% of the time. However, it follows that the interval will miss μ with a probability of 0.05, or 5% of the time. **This is an important feature of confidence intervals - they will often but not always enclose the true parameter value for the population, with the probability set by α.**

If we wanted to be more certain of including μ, we could choose a smaller α, say $\alpha = 0.01$, which corresponds to a $100(1 - 0.01)\% = 99\%$ confidence interval. Here we have $c_\alpha = c_{0.01} = 2.576$, and so the 99% confidence interval would be

$$(\bar{Y} - 2.576\frac{\sigma}{\sqrt{n}}, \bar{Y} + 2.576\frac{\sigma}{\sqrt{n}}). \tag{9.28}$$

A 99% confidence interval will necessarily be broader than a 95% one, because it is constructed to have a higher probability of including μ.

Confidence intervals - sample calculation

Suppose we have a sample of $n = 10$ elytra from female *T. dubius* beetles (see Chapter 3 for a description of these data), yielding the values listed below:

```
5.0 5.1 5.2 5.9 4.8 5.5 4.8 5.1 5.0 5.1
```

For this sample, we calculate that $\bar{Y} = 5.150$. Suppose we have *a priori* knowledge that $\sigma = 0.3$, although that would be rare in practice. We will calculate a 95% and 99% confidence interval for μ.

The formula for a 95% confidence interval is

$$(\bar{Y} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{Y} + 1.96\frac{\sigma}{\sqrt{n}}). \tag{9.29}$$

Substituting $n = 10, \bar{Y} = 5.150$, and $\sigma = 0.3$ in the above formula, we obtain

$$(5.150 - 1.96\frac{0.3}{\sqrt{10}}, 5.150 + 1.96\frac{0.3}{\sqrt{10}}), \tag{9.30}$$

or

$$(5.150 - 0.186, 5.150 + 0.186), \tag{9.31}$$

or

$$(4.964, 5.336). \tag{9.32}$$

So, the 95% confidence interval for μ is $(4.964, 5.336)$.

For a 99% confidence interval, we use the formula

$$(\bar{Y} - 2.576\frac{\sigma}{\sqrt{n}}, \bar{Y} + 2.576\frac{\sigma}{\sqrt{n}}). \tag{9.33}$$

Substituting as before, we obtain

$$(5.150 - 2.576\frac{0.3}{\sqrt{10}}, 5.150 + 2.576\frac{0.3}{\sqrt{10}}), \tag{9.34}$$

or

$$(5.150 - 0.244, 5.150 + 0.244), \tag{9.35}$$

or

$$(4.906, 5.394). \tag{9.36}$$

The 99% confidence interval is therefore $(4.906, 5.394)$. Note that the 99% confidence interval is broader than the 95% one, because its lower limit is lower and upper limit higher.

9.2.2 Confidence intervals for μ when σ^2 is estimated

Confidence intervals for μ can also be derived when σ^2 is estimated using the sample variance s^2, as will usually be the case in practice. We will make use of the fact that

$$\frac{\bar{Y} - \mu}{s/\sqrt{n}} \sim t_{n-1}. \tag{9.37}$$

Using Table T, we can find a value of $c_{\alpha,n-1}$ for $n-1$ degrees of freedom such that the following equation is true:

$$P\left[-c_{\alpha,n-1} < \frac{\bar{Y} - \mu}{s/\sqrt{n}} < c_{\alpha,n-1}\right] = 1 - \alpha. \tag{9.38}$$

Rearranging this equation using the same procedures as before, we obtain

$$P\left[\bar{Y} - c_{\alpha,n-1}\frac{s}{\sqrt{n}} < \mu < \bar{Y} + c_{\alpha,n-1}\frac{s}{\sqrt{n}}\right] = 1 - \alpha. \tag{9.39}$$

The terms $\bar{Y} - c_{\alpha,n-1}\frac{s}{\sqrt{n}}$ and $\bar{Y} + c_{\alpha,n-1}\frac{s}{\sqrt{n}}$ are the lower and upper $100(1-\alpha)\%$ confidence limits for μ (Mood et al. 1974). The interval would be reported in the form $(\bar{Y} - c_{\alpha,n-1}\frac{s}{\sqrt{n}}, \bar{Y} + c_{\alpha,n-1}\frac{s}{\sqrt{n}})$. The center of the confidence interval is located at $\bar{Y}$, the estimate of μ.

For example, if we let $\alpha = 0.05$ this corresponds to a 95% confidence interval of the form

$$(\bar{Y} - c_{0.05,n-1}\frac{s}{\sqrt{n}}, \bar{Y} + c_{0.05,n-1}\frac{s}{\sqrt{n}}). \tag{9.40}$$

The value of $c_{0.05,n-1}$ would need to be determined from Table T, using the column for $2(1-p) = \alpha = 0.05$ and the row for $n-1$ degrees freedom.

For $\alpha = 0.01$, we obtain a 99% confidence interval of the form

$$(\bar{Y} - c_{0.01,n-1}\frac{s}{\sqrt{n}}, \bar{Y} + c_{0.01,n-1}\frac{s}{\sqrt{n}}). \tag{9.41}$$

In this case, we would use the column for $2(1-p) = \alpha = 0.01$ to find the value of $c_{0.01,n-1}$, using $n-1$ degrees freedom.

Confidence interval for μ - sample calculation

We return to the elytra data set, for which we previously calculated that $\bar{Y} = 5.150$, $s^2 = 0.109$, and $s = 0.331$ for $n = 10$. We will calculate 95% and 99% confidence intervals for μ.

The formula for a 95% confidence interval is

$$(\bar{Y} - c_{0.05,n-1}\frac{s}{\sqrt{n}}, \bar{Y} + c_{0.05,n-1}\frac{s}{\sqrt{n}}). \tag{9.42}$$

For $n = 10$, we have $df = n - 1 = 10 - 1 = 9$. For a 95% confidence interval, we therefore look up $c_{0.05,n-1} = c_{0.05,9}$ using the column for $2(1-p) = 0.05$ in Table T, choosing the value for 9 degrees of freedom. We obtain $c_{0.05,9} = 2.262$. Substituting $n = 10, \bar{Y} = 5.150, s = 0.331$, and $c_{0.05,9} = 2.262$ in the above formula, we obtain

$$(5.150 - 2.262\frac{0.331}{\sqrt{10}}, 5.150 + 2.262\frac{0.331}{\sqrt{10}}), \tag{9.43}$$

or

$$(5.150 - 0.237, 5.150 + 0.237), \tag{9.44}$$

or

$$(4.913, 5.387). \tag{9.45}$$

So, the 95% confidence interval for μ is $(4.913, 5.387)$. For a 99% confidence interval, we find $c_{0.01,n-1} = c_{0.01,9}$ for $2(1-p) = 0.01$ and 9 degrees of freedom in Table T, obtaining $c_{0.01,9} = 3.250$. Substituting this value in the above formula, we obtain

$$(5.150 - 3.250\frac{0.331}{\sqrt{10}}, 5.150 + 3.250\frac{0.331}{\sqrt{10}}), \tag{9.46}$$

or

$$(5.150 - 0.340, 5.150 + 0.340), \tag{9.47}$$

or

$$(4.810, 5.490). \tag{9.48}$$

The 99% confidence interval is therefore $(4.810, 5.490)$, and as expected is broader than the 95% one.

9.2.3 Confidence intervals for σ^2 and σ

Confidence intervals for σ^2 and σ can also be derived, using the fact that

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1} \tag{9.49}$$

Using Table C for the χ^2 distribution, we can find values $c_{\alpha/2,n-1}$ and $c_{1-\alpha/2,n-1}$ for $n-1$ degrees of freedom such that the following equation is true:

$$P\left[c_{\alpha/2,n-1} < \frac{(n-1)s^2}{\sigma^2} < c_{1-\alpha/2,n-1}\right] = 1 - \alpha. \tag{9.50}$$

We now rearrange this equation to obtain a confidential interval for σ^2. If we take the inverse of all the inside terms, we obtain

$$P\left[\frac{1}{c_{\alpha/2,n-1}} > \frac{\sigma^2}{(n-1)s^2} > \frac{1}{c_{1-\alpha/2,n-1}}\right] = 1 - \alpha. \tag{9.51}$$

Note that taking the inverse changes the direction of the inequality signs. Multiplying each term by $(n-1)s^2$ we obtain

$$P\left[\frac{(n-1)s^2}{c_{\alpha/2,n-1}} > \sigma^2 > \frac{(n-1)s^2}{c_{1-\alpha/2,n-1}}\right] = 1-\alpha, \tag{9.52}$$

or equivalently

$$P\left[\frac{(n-1)s^2}{c_{1-\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)s^2}{c_{\alpha/2,n-1}}\right] = 1-\alpha. \tag{9.53}$$

The terms $\frac{(n-1)s^2}{c_{1-\alpha/2,n-1}}$ and $\frac{(n-1)s^2}{c_{\alpha/2,n-1}}$ are the lower and upper $100(1-\alpha)\%$ confidence limits for σ^2, and the interval $(\frac{(n-1)s^2}{c_{1-\alpha/2,n-1}}, \frac{(n-1)s^2}{c_{\alpha/2,n-1}})$ is a $100(1-\alpha)\%$ confidence interval for σ^2 (Mood et al. 1974). The confidence interval for σ^2 is not symmetrical around the value s^2, our estimate of σ^2.

For a 95% confidence interval with $\alpha = 0.05$, the confidence interval formula is

$$\left(\frac{(n-1)s^2}{c_{0.975,n-1}}, \frac{(n-1)s^2}{c_{0.025,n-1}}\right) \tag{9.54}$$

To find $c_{0.025,n-1}$, we look across the top row of Table C and find the column corresponding to $p = 0.025$, then look for the row corresponding to $n-1$ degrees of fredom. To find $c_{0.975,n-1}$, we use the column corresponding to $p = 0.975$, again looking for the row with $n-1$ degrees of freedom.

For a 99% confidence interval with $\alpha = 0.01$, the confidence interval formula is

$$\left(\frac{(n-1)s^2}{c_{0.995,n-1}}, \frac{(n-1)s^2}{c_{0.005,n-1}}\right) \tag{9.55}$$

To find $c_{0.005,n-1}$, we use the column corresponding to $p = 0.005$, while the column for $c_{0.995,n-1}$ corresponds to $p = 0.995$. We again use the entries corresponding to $n-1$ degrees of freedom.

We can also obtain a confidence interval for $\sigma = \sqrt{\sigma^2}$ by taking the square root of the above confidence limits. In particular, a confidence interval for σ would be $(\sqrt{\frac{(n-1)s^2}{c_{1-\alpha/2,n-1}}}, \sqrt{\frac{(n-1)s^2}{c_{\alpha/2,n-1}}})$.

Confidence interval for σ^2 and σ - sample calculation

Recall the elytra data set, for which $\bar{Y} = 5.150$ and $s^2 = 0.109$ for $n = 10$. Calculate a 95% and 99% confidence interval for σ^2 and then σ.

The formula for a 95% confidence interval is

$$\left(\frac{(n-1)s^2}{c_{0.975,n-1}}, \frac{(n-1)s^2}{c_{0.025,n-1}}\right) \tag{9.56}$$

For $n = 10$, we have $df = n - 1 = 10 - 1 = 9$.

For a 95% confidence interval, with $\alpha = 0.05$, we find from Table C that $c_{0.025,n-1} = c_{0.025,9} = 2.700$, and $c_{0.975,n-1} = c_{0.975,9} = 19.023$. Substituting $n = 10, s^2 = 0.110, c_{0.025,9} = 2.700$ and $c_{0.975,9} = 19.023$ in the above formula, we obtain

$$\left(\frac{(10-1)0.109}{19.023}, \frac{(10-1)0.109}{2.700}\right) \tag{9.57}$$

or

$$(0.052, 0.363). \tag{9.58}$$

So, the 95% confidence interval for σ^2 is $(0.052, 0.363)$. To obtain a 95% confidence interval for σ we simply take the square root of these values, or $(\sqrt{0.052}, \sqrt{0.363}$, to obtain $(0.228, 0.603)$.

For a 99% confidence interval, the formula is

$$\left(\frac{(n-1)s^2}{c_{0.995,n-1}}, \frac{(n-1)s^2}{c_{0.005,n-1}}\right) \tag{9.59}$$

We use Table C to find $c_{0.005,n-1} = c_{0.005,9} = 1.735$, and $c_{0.995,n-1} = c_{0.995,9} = 23.589$. Substituting these values in the above formula, we obtain

$$\left(\frac{(10-1)0.109}{23.589}, \frac{(10-1)0.109}{1.735}\right) \tag{9.60}$$

or

$$(0.042, 0.565). \tag{9.61}$$

The 99% confidence interval of σ^2 is therefore $(0.042, 0.565)$. To obtain a 99% confidence interval for σ, we take the square root and obtain $(0.205, 0.752)$. Note that the 99% intervals are wider than the corresponding 95% ones.

9.2.4 Confidence intervals - SAS demo

These same calculations can be readily accomplished using `proc univariate` in SAS (SAS Institute Inc. 2016). We obtain 95% confidence intervals by including the option `cibasic` in the `proc univariate` line of the program.

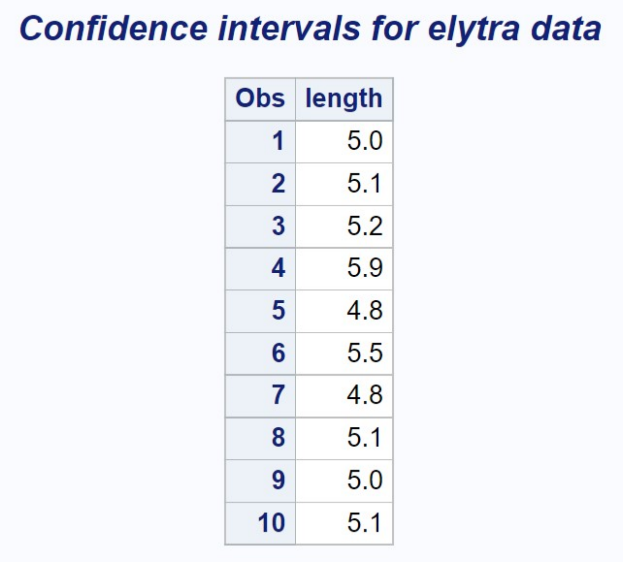
99% confidence intervals may be obtained by specifying `alpha=0.01` in the `proc univariate` line. See SAS program and Fig. 9.3 - 9.6 below. Similar to our earlier calculations, the 95% confidence interval was $(4.913, 5.387)$ for μ, $(0.052, 0.365)$ for σ^2, and $(0.228, 0.604)$ for σ. The 99% confidence intervals can be found further in the output.

9.2.5 Confidence interval size

Confidence intervals are a method of characterizing the precision of parameter estimates, with narrower intervals generally indicating a population parameter like μ is better estimated. How then can an investigator reduce the size of these confidence intervals? The simplest way is to increase the sample size n on which the estimate is based. This reduces the size of confidence intervals for μ because it reduces the magnitude of the quantity $c_{\alpha,n-1}s/\sqrt{n}$, which determines the width of the interval (see Eq. 9.26). Most of this effect is through the $\sqrt{n}$ term here, but $c_{\alpha,n-1}$ also becomes smaller for larger n. Increasing the sample size n also reduces the size of the confidence intervals for σ^2 and σ, although the mechanism is more complex in this case.

SAS Program

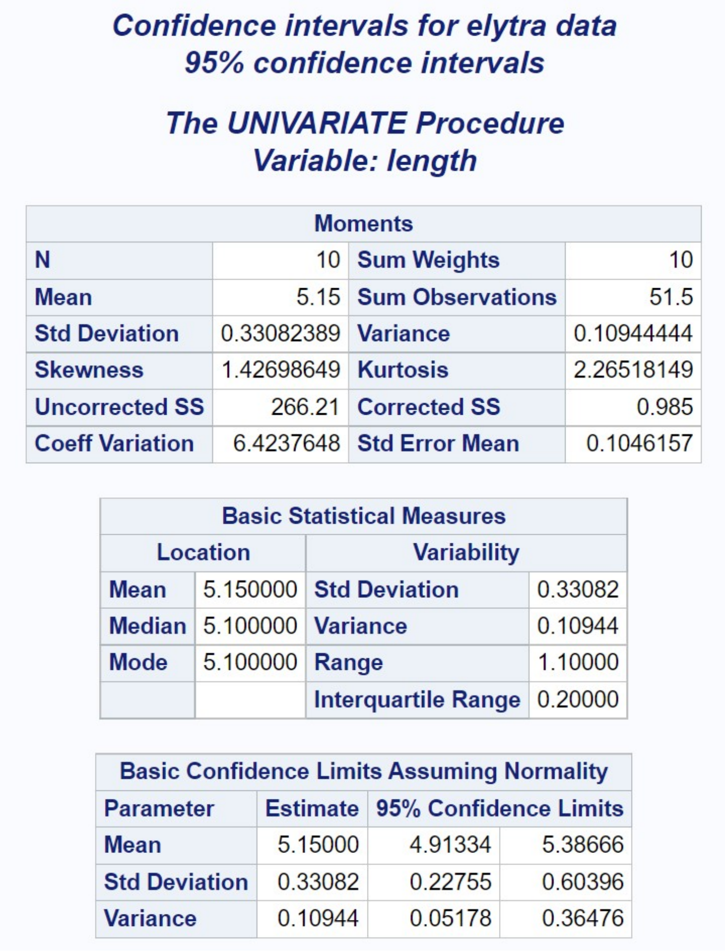
```
* Confidence_intervals.sas;
title ’Confidence intervals for elytra data’;
data elytra;
    input length;
    datalines;
5.0
5.1
5.2
5.9
4.8
5.5
4.8
5.1
5.0
5.1
;
run;
* Print data set;
proc print data=elytra;
run;
* Generate 95% confidence intervals and plots;
title2 "95% confidence intervals";
proc univariate cibasic data=elytra;
    var length;
    histogram length / vscale=count normal;
    qqplot length / normal;
run;
* Generate 99% confidence intervals;
title2 "99% confidence intervals";
proc univariate cibasic alpha = 0.01 data=elytra;
    var length;
run;
quit;
```



Confidence intervals for elytra data

Obs	length
1	5.0
2	5.1
3	5.2
4	5.9
5	4.8
6	5.5
7	4.8
8	5.1
9	5.0
10	5.1

Figure 9.3: `confidence_intervals.sas - proc print`



Confidence intervals for elytra data
95% confidence intervals

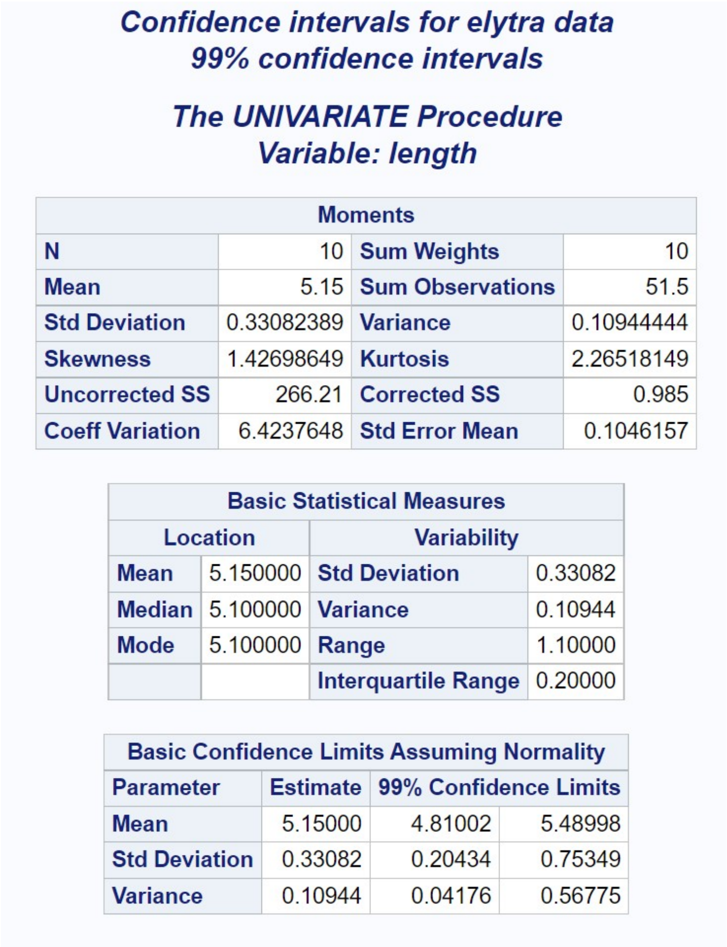
The UNIVARIATE Procedure
Variable: length

Moments			
N	10	Sum Weights	10
Mean	5.15	Sum Observations	51.5
Std Deviation	0.33082389	Variance	0.10944444
Skewness	1.42698649	Kurtosis	2.26518149
Uncorrected SS	266.21	Corrected SS	0.985
Coeff Variation	6.4237648	Std Error Mean	0.1046157

Basic Statistical Measures			
Location		Variability	
Mean	5.150000	Std Deviation	0.33082
Median	5.100000	Variance	0.10944
Mode	5.100000	Range	1.10000
		Interquartile Range	0.20000

Basic Confidence Limits Assuming Normality			
Parameter	Estimate	95% Confidence Limits	
Mean	5.15000	4.91334	5.38666
Std Deviation	0.33082	0.22755	0.60396
Variance	0.10944	0.05178	0.36476

Figure 9.4: `confidence_intervals.sas - proc univariate`



Confidence intervals for elytra data
99% confidence intervals

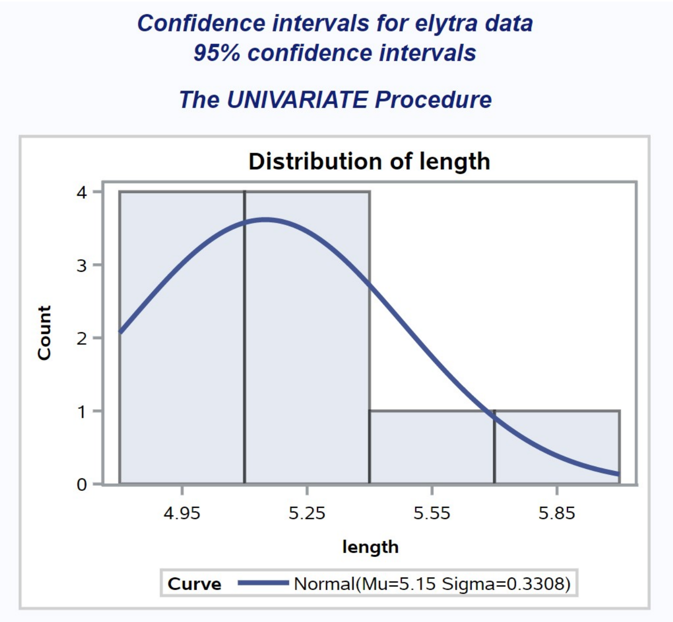
The UNIVARIATE Procedure
Variable: length

Moments			
N	10	Sum Weights	10
Mean	5.15	Sum Observations	51.5
Std Deviation	0.33082389	Variance	0.10944444
Skewness	1.42698649	Kurtosis	2.26518149
Uncorrected SS	266.21	Corrected SS	0.985
Coeff Variation	6.4237648	Std Error Mean	0.1046157

Basic Statistical Measures			
Location		Variability	
Mean	5.150000	Std Deviation	0.33082
Median	5.100000	Variance	0.10944
Mode	5.100000	Range	1.10000
		Interquartile Range	0.20000

Basic Confidence Limits Assuming Normality			
Parameter	Estimate	99% Confidence Limits	
Mean	5.15000	4.81002	5.48998
Std Deviation	0.33082	0.20434	0.75349
Variance	0.10944	0.04176	0.56775

Figure 9.5: `confidence_intervals.sas - proc univariate`



Confidence intervals for elytra data
95% confidence intervals

The UNIVARIATE Procedure

Figure 9.6: `confidence_intervals.sas - proc univariate`

9.3 References

Mood, A. M., Graybill, F. A. & Boes, D. C. (1974) *Introduction to the Theory of Statistics.* McGraw-Hill, Inc., New York, NY.

SAS Institute Inc. (2016) *Base SAS 9.4 Procedures Guide, Sixth Edition.* SAS Institute Inc., Cary, NC.

9.4 Problems

1. Ten adult female *Daphnia ambigua* (Lei and Armitage 1980) were cultured under laboratory conditions, and their longevity (days) determined. The following data were obtained.

   ```
   28 4 22 21 17 21 22 26 15 19
   ```

   (a) Find $\bar{Y}$, s^2, and s for these data, then calculate a 95% confidence interval for μ, σ^2 and then σ. Show all your calculations.

   (b) Find a 99% confidence interval for μ, σ^2 and then σ. Show your calculations.

   (c) Use SAS to find the same confidence intervals as in parts a and b. List the confidence intervals and test results below. Attach your SAS program(s) and output.

2. A study was conducted to measure the population growth rate of a laboratory culture of nematodes. A hundred nematodes were each added to 8 petri dishes of a new growth media, and the number of offspring counted one generation later. The number of offspring divided by the initial number of organisms (100) provides an estimate of λ, the finite growth rate of the population. It is customary to log-transform the values of λ in such studies, yielding $r = \ln(\lambda)$. The following values of r were obtained:

   ```
   2.1 0.8 1.8 1.9 0.8 1.7 0.5 1.6
   ```

   (a) Find $\bar{Y}$, s^2, and s for these data, then calculate a 95% confidence interval for μ, σ^2 and then σ. Show all your calculations.

   (b) Find a 99% confidence interval for μ, σ^2 and then σ. Show your calculations.

   (c) Use SAS to find the same confidence intervals as in parts a and b. List the confidence intervals and test results below. Attach your SAS program(s) and output.

Chapter 10

Hypothesis Testing

We previously examined how the parameters for a probability distribution can be estimated using a random sample and maximum likelihood (Chapter 8), as then showed how confidence intervals provide a measure of the reliability of these estimates (Chapter 9). In hypothesis testing, the subject of this chapter, we examine the consistency of observed data sets with a null hypothesis, commonly a statement about the parameter values within a statistical model. We conduct a statistical test of this null hypothesis, with the result being a decision to accept or reject the null hypothesis based on the magnitude of a quantity called a P value. Small values of P indicate a test result inconsistent with the null hypothesis, suggesting it might be false and some alternative hypothesis more valid. In the following, we discuss the different components and steps of hypothesis testing.

10.1 The null and alternative hypotheses

As an example of hypothesis testing, suppose that we rear n tilapia on a commercial diet, and want to compare their body size with ones reared using a natural diet. Fish reared on natural food are already known to have a weight of 500 g at a certain age, and weight is normally distributed. We could test whether the fish reared on the commercial diet have the same mean weight as ones reared on natural food (500 g) using the **null hypothesis** that $\mu = 500$ g, where μ is the mean parameter for the normal distribution. This can be written as $H_0 : \mu = 500$ g, where H_0 stands for null hypothesis. Null hypotheses of this type can be written more generally as $H_0 : \mu = \mu_0$, where

μ_0 is the hypothesized mean of the distribution. For the tilapia problem, we would have $\mu_0 = 500$ g.

An **alternative hypothesis** for this example is that the mean weight of tilapia on commercial diet is different from 500 g. This can be written as $H_1 : \mu \neq 500$ g, where H_1 stands for the alternative hypothesis. Alternative hypotheses of this type are written generally as $H_1 : \mu \neq \mu_0$. We may also be interested in particular values of the alternative mean, such as $H_1 : \mu = 490$ g or $H_1 : \mu = 530$ g, or more generally $H_1 : \mu = \mu_1$.

10.2 Test statistics

A test statistic is a quantity that measures the consistency of the observed data with the null hypothesis. Test statistics are usually chosen so that large values occur when the data are inconsistent with H_0. What would be a suitable test statistic for the tilapia problem, using $H_0 : \mu = \mu_0$ as the null hypothesis? Suppose we rear n fish on the commercial diet, and then calculate the sample mean $\bar{Y}$ of their weights. The statistic $\bar{Y}$ is an estimator of the true mean μ for this statistical population, which may or may not be equal to the μ_0 under the null hypothesis. A value of $\bar{Y}$ substantially greater than μ_0, or smaller than μ_0, would be inconsistent with H_0. This suggests using the quantity $\bar{Y} - \mu_0$ as the test statistic for the problem. What about the other parameter for the normal distribution, σ^2 or σ? For simplicity, we will assume that it is a known quantity, although this is rare in practice. We could then use the test statistic

$$Z_s = \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} \tag{10.1}$$

to test $H_0 : \mu = \mu_0$ (Bickel & Doksum 1977). We use this statistic because it has a standard normal distribution under H_0 ($Z_s \sim N(0, 1)$, see Chapter 9) which makes it straightforward to employ the test. Note that Z_s becomes large (positive or negative) if the sample mean $\bar{Y}$ differs greatly from μ_0. In general, tests based on the standard normal distribution are called Z tests.

10.3 Acceptance and rejection regions – Type I error

Given a suitable test statistic, how large must it be before we decide the data are inconsistent with H_0? This is determined by finding an interval that defines an **acceptance region** for the test, and its complement, called the **rejection** or **critical region** (Bickel & Doksum 1977). We then accept H_0 if the test statistic falls within the acceptance region, and reject H_0 if it falls outside or lies on its boundary. The boundaries of the acceptance and rejection regions are determined by setting the probability of a Type I error. **A Type I error is defined as the test rejecting H_0 when H_0 is true. The probability of committing a Type I error is called the Type I error rate, usually denoted with the symbol α.** It is common practice to set $\alpha = 0.05$, meaning there is a 1 in 20 chance that the test will reject H_0 even when it is true. It follows that the probability of the test accepting H_0 if it is true is $1 - \alpha$. For $\alpha = 0.05$, we have $1 - \alpha = 1 - 0.05 = 0.95$.

The acceptance region is determined as follows. Suppose that $H_0 : \mu = \mu_0$ is true. Because the test statistic $Z_s \sim N(0, 1)$ under H_0, the following is a true statement:

$$P[-c_\alpha < Z_s < c_\alpha] = P[-c_\alpha < \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} < c_\alpha] = 1 - \alpha. \tag{10.2}$$

The quantity c_α would be chosen using Table Z to satisfy this equation (for details see Chapter 9). The interval $(-c_\alpha, c_\alpha)$ is the acceptance region of a test with a Type I error rate of α. Under H_0, the test statistic Z_s would lie within this interval with probability $1 - \alpha$ and outside this region with probability α, which is the required Type I error rate. The rejection region would be the complement of the acceptance region, i.e., all values on the boundary or outside of $(-c_\alpha, c_\alpha)$.

For example, with $\alpha = 0.05$ we find that $c_{0.05} = 1.96$, and so we would accept H_0 if Z_s lies within $(-1.96, 1.96)$ and reject H_0 if it lies outside this interval or exactly on the boundary (see Fig. 10.1). The acceptance region for this test can also be expressed using absolute values - we would accept H_0 if $|Z_s| < 1.96$ and reject it if $|Z_s| \geq 1.96$.

The acceptance region becomes larger (and the rejection region smaller) for smaller α values. For $\alpha = 0.01$, we find that $c_{0.01} = 2.576$ and so the acceptance region is $(-2.576, 2.576)$ (Fig. 10.2). Using absolute values, we would accept H_0 if $|Z_s| < 2.576$ and reject it otherwise. Using a smaller value of α indicates we are more concerned about making a Type I error. For $\alpha = 0.01$ there is only a 1 in 100 chance we would reject H_0 if H_0 were true, but this also reduces the power of the test (see below) to detect whether H_0 is false.

The acceptance and rejection regions we just developed are for a **two-tailed test**, which tests the null hypothesis $H_0 : \mu = \mu_0$ with $H_1 : \mu \neq \mu_0$ the alternative hypothesis. This test statistic will reject H_0 for either large and small values of the test statistic Z_s, which occurs when $\bar{Y}$ is greater than μ_0 or less than μ_0. We will later examine the behavior of **one-tailed tests**, where the null is $H_0 : \mu = \mu_0$ while the alternative is of the form $H_1 : \mu > \mu_0$, or $H_1 : \mu < \mu_0$. Note that the two alternative hypotheses here specify that μ is either greater or less than μ_0. One-tailed tests are designed to reject H_0 in only one direction.

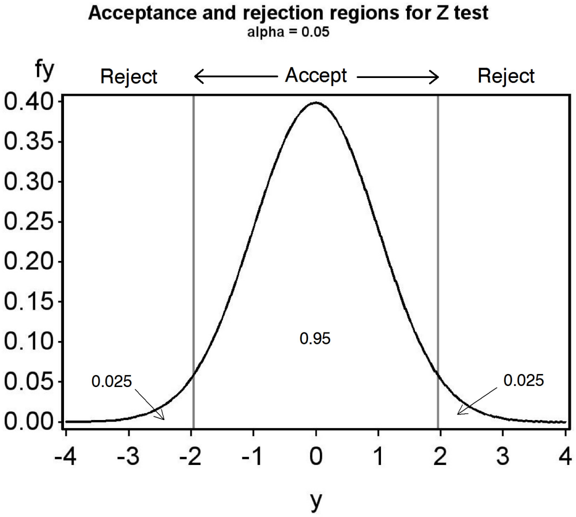


Figure 10.1: Acceptance and rejection regions for a one-sample Z test, $\alpha = 0.05$. Also shown is the distribution of Z_s under H_0.

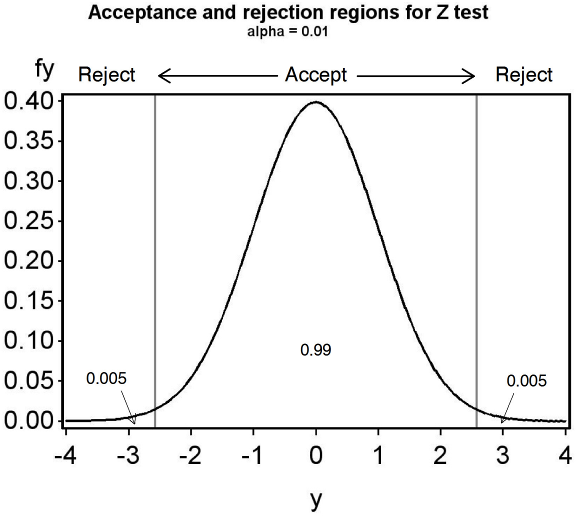


Figure 10.2: Acceptance and rejection regions for a one-sample Z test, $\alpha = 0.01$. Also shown is the distribution of Z_s under H_0.

10.3.1 One-sample Z test - sample calculation

We will now do an example of this test, known as a one-sample Z test. Recall the tilapia diet example, where it is known that fish reared on natural food have a mean weight of 500 g. We rear $n = 10$ fish on a commercial diet, and want to compare the weight of fish on the commercial diet with ones reared on natural food. In particular, we want to test $H_0 : \mu = 500$ g. We find that $\bar{Y} = 495$ g for the fish reared on the commercial diet, and already know that $\sigma^2 = 49$ g^2, so $\sigma = 7$ g. Because $\bar{Y} = 495$ g is less than 500 g, it already appears that the commercial diet produces smaller fish than natural food, but a statistical test is still needed to provide convincing evidence against H_0. For the test statistic, we have

$$Z_s = \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}} = \frac{495 - 500}{7/\sqrt{10}} = \frac{-5}{2.214} = -2.258 \tag{10.3}$$

For a Type I error rate of $\alpha = 0.05$, the acceptance region for Z_s is $(-1.96, 1.96)$. $Z = -2.258$ lies outside this interval, so we would reject H_0 at the $\alpha = 0.05$ level. For $\alpha = 0.01$ the acceptance region is $(-2.576, 2.576)$. Because Z_s lies within this interval, we would accept H_0 at this α level. Thus, the decision to accept or reject H_0 depends on both the test statistic value and the value of α.

10.4 P values

As noted above, the value of α can affect whether we accept or reject H_0. Rather than force a particular α on the analyst, the test results can also be presented in the form of a P value. **A P value is defined as the smallest value of α for which one can just reject** H_0 (Bickel & Doksum 1977). It is calculated by finding an α such that the test statistic Z_s is equal to c_α.

Recall from Chapter 9 that c_α is defined so that the following equation is true:

$$P[Z < c_\alpha] = 1 - \alpha/2. \tag{10.4}$$

To find the P value for the tilapia example, we substitute the test statistic value Z_s for c_α in the above equation, ignoring the fact that Z_s is negative. We have

$$P[Z < Z_s] = P[Z < 2.258] = 1 - \alpha/2. \tag{10.5}$$

From Table Z, we see that $P[Z < 2.258] \approx 0.9881$. We then solve the equation

$$0.9881 = 1 - \alpha/2 \tag{10.6}$$

for α to obtain the P value. We have $\alpha = 2(1 - 0.9881) = 0.0238$. This is the P value for the test, reported as $P = 0.0238$. Given the P value, the analyst or other interested parties can decide for themselves whether to reject or accept H_0.

A P value can also be thought of as the probability of obtaining a test statistic equal to or more extreme than the observed one, under the null hypothesis. We can see this from a graph of the acceptance and rejection regions for the tilapia example, where $Z_s = -2.258$ and $P = 0.0238$ (Fig. 10.3). The probabilities outside the acceptance region correspond to $P[Z_s \leq -2.258]$ and $P[Z_s \geq 2.258]$, which are the probabilities of observing values of Z_s equal to or more extreme than the observed value of $Z_s = -2.258$. The two definitions of a P value are equivalent.

A P value is also a measure of the consistency of the observed data with the null hypothesis. If the P value is large, say $P > 0.05$, then the observed data generated a test statistic value that is fairly likely under the null hypothesis. On the other hand, if P is small then the observed data has generated a test statistic that is unlikely under the null hypothesis. This suggests the observed data are inconsistent with the null hypothesis, and the null may be false.

There are specific phrases generally used to describe the significance of a statistical test result. If a test yields $P \leq 0.05$, it is described as being **significant**, while if $P \leq 0.01$ it is **highly significantly**. If $P > 0.05$ the test is described as **nonsignificant**. The tilapia example with $P = 0.0238$ would be described as significant because $0.0238 < 0.05$, but not highly significant.

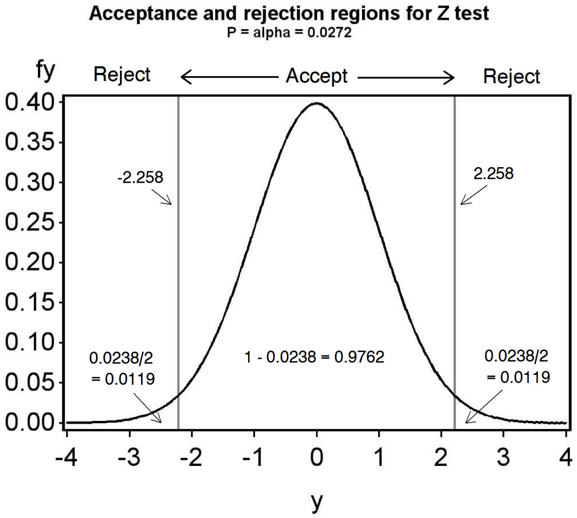


Figure 10.3: Acceptance-rejection region for a one-sample Z test, exact $P = 0.0238$

10.5 Type II error and power

Suppose now that H_0 is actually false and some alternative hypothesis H_1 is true. **A Type II error is defined as failing to reject H_0 when H_0 is false**. The probability of committing a Type II error is called the Type II error rate, usually denoted by the symbol β. It follows that the probability of the test rejecting H_0 if it is false is $1 - \beta$, and this quantity is called the **power** of the test (Bickel & Doksum 1977). High power values indicate the test is capable of detecting departures from the null hypothesis.

The power and Type II error rate of a statistical test depends on the sample size n of the test, the standard deviation of the observations σ, the Type I error rate α, and the particular alternative hypothesis chosen. An analyst interested in determining the power of a test will fix some of these values, often α and σ, and then examine how changes in n and the alternative hypothesis affect power. This procedure is called a **power analysis**. A power value of 0.8 is believed to be adequate in most situations (Cohen 1988). This implies that a statistical test will reject H_0 when it is false 80% of the time.

It is relatively easy to calculate the power for a one-sample Z test, using the distribution of Z_s under H_1. Suppose that we choose $\alpha = 0.05$, so that the acceptance region is the interval $(-1.96, 1.96)$, and that the alternative hypothesis is $H_1 : \mu = \mu_1$ for some μ_1. Under $H_0 : \mu = \mu_0$ the test statistic has a standard normal distribution, implying $Z_s \sim N(0, 1)$, but what is its distribution under H_1? Using the expected value and variance rules in Chapter 7, one can show that

$$E[Z_s] = \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} = \phi \tag{10.7}$$

and also that $Var[Z_s] = 1$. So, Z_s has the same variance under both H_1 and H_0, but the mean under H_1 is equal to ϕ, not zero as under H_0. It follows that under H_1 the test statistic $Z_s \sim N(\phi, 1)$. The probability of rejecting H_0 when H_1 is true, the power of the test, is the probability that Z_s lies outside the interval $(-1.96, 1.96)$, or

$$\text{power} = P[Z_s \leq -1.96] + P[Z_s \geq 1.96]. \tag{10.8}$$

The Type II error rate β can be calculated as $1-$power, or directly by finding

$$\beta = P[-1.96 < Z_s < 1.96] \tag{10.9}$$

when H_1 is true.

Fig. 10.4 shows the power and Type II error for the tilapia example with $H_0 : \mu = 500$ vs. a particular alternative hypothesis, $H_1 : \mu = 495$. We assume $\sigma = 7$ as before, with $n = 10$ and $\alpha = 0.05$. For this alternative hypothesis, we have

$$\phi = \frac{\mu_1 - \mu_0}{\sigma/\sqrt{10}} = \frac{(495 - 500)}{7/\sqrt{10}} = -2.26. \tag{10.10}$$

Thus, under H_1 we have $Z_s \sim N(-2.26, 1)$, and this distribution is shown as well as the distribution of Z_s under H_0 and the acceptance and rejection regions for the test. The power is the area Z_s under H_1 outside the acceptance region, while β is the area in the region.

What happens to the power as we vary μ_1? Suppose now that $H_1 : \mu_1 = 490$ is the alternative hypothesis. As we can see from Fig. 10.5, in this case the power is substantially higher and β is lower. Fig. 10.6 shows how power changes as we vary μ_1 across a range of values. Power is quite high (nearly 1) for μ_1 far from μ_0, but approaches a minimum value of α for μ_1 near μ_0. The minimum power is α, not zero, because the test will reject H_0 even if it is true ($\mu_1 = \mu_0$) at this rate.

Power is also affected by sample size. If we use $H_1 : \mu = 495$ and increase the sample size from $n = 10$ to $n = 20$, this also increases the power (Fig. 10.7). However, an increase in the standard deviation from $\sigma = 7$ to $\sigma = 10$ lowers the power (Fig. 10.8).

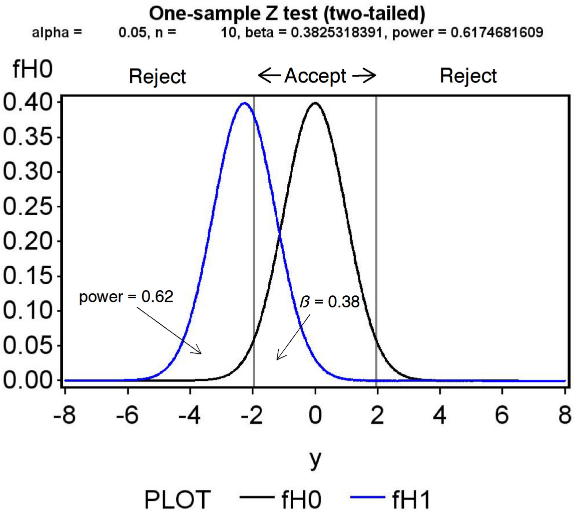


Figure 10.4: Distribution of Z_s under $H_1 : \mu = 495$, with $\sigma = 7, n = 10$ ($\phi = -2.26$). Almost all of the power occurs to the left of the acceptance region, but there is also a small amount to the right. Also shown is the distribution of Z_s under H_0.

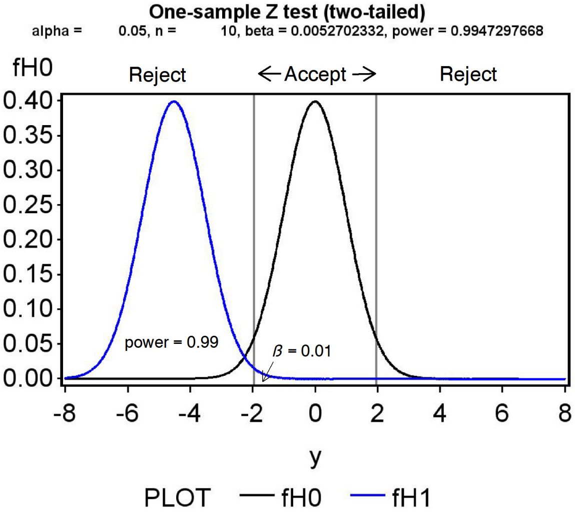


Figure 10.5: Distribution of Z_s under $H_1 : \mu = 490$, with $\sigma = 7, n = 10$ ($\phi = -4.52$).

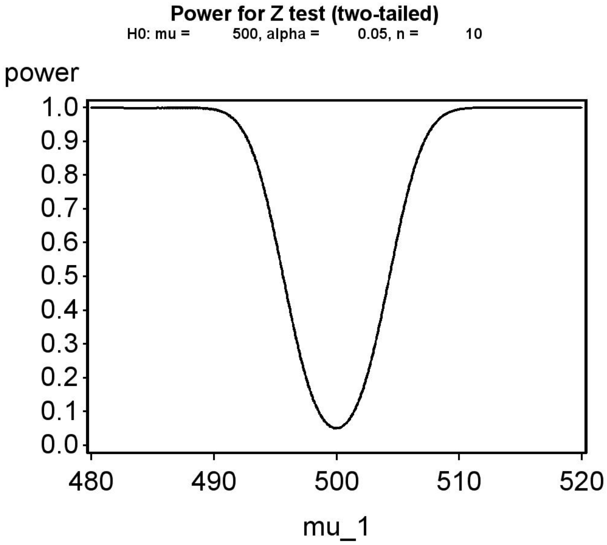


Figure 10.6: Power across a range of μ_1 values, for $H_0 : \mu = 500$, $\sigma = 7$, and $n = 10$

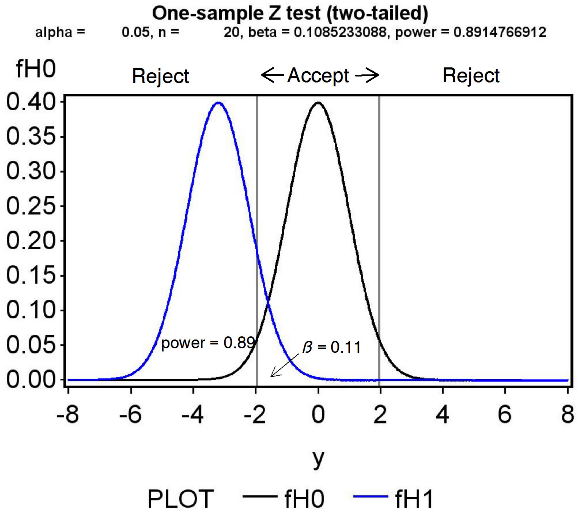


Figure 10.7: Distribution of Z_s under $H_1 : \mu = 495$, with $\sigma = 7, n = 20$ ($\phi = -3.19$).

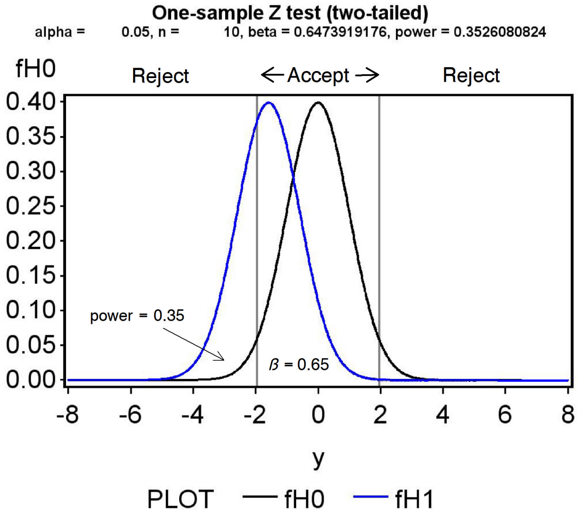


Figure 10.8: Distribution of Z_s under $H_1 : \mu = 495$, with $\sigma = 10, n = 10$ ($\phi = -1.58$).

Table 10.1: Effects on power and the Type II error rate β of changes in various parameters. The arrows indicate if a particular quantity increases or decreases.

Parameter	Direction	ϕ	power	β
$\lvert\mu_1 - \mu_0\rvert$	$\uparrow$	$\uparrow$	$\uparrow$	$\downarrow$
n	$\uparrow$	$\uparrow$	$\uparrow$	$\downarrow$
σ	$\uparrow$	$\downarrow$	$\downarrow$	$\uparrow$
α	$\uparrow$	no change	$\uparrow$	$\downarrow$

All of these effects on power can be understood through their influence on ϕ. Any change in a parameter value that makes ϕ larger increases power and reduces β, because it shifts the distribution of Z_s under H_1 away from the acceptance and into the rejection region. Thus, large differences between μ_1 and μ_0, large n, and small σ will all increase power because they increase ϕ. Conversely, close values of μ_1 and μ_0, small n, and large σ would all reduce power. Table 10.1 summarizes how the different parameter values influence ϕ, power, and the Type II error rate β. Also shown is the effect of the Type I error rate α on power. If an investigator can accept a larger value of α, so that Type I errors are more common, this reduces the acceptance and increases the rejection region size, and thus increases power.

Note that a sufficiently large value of n can generate a large value of ϕ, even when μ_1 and μ_0 are close or σ is large. Thus, large sample sizes can yield adequate power even when the data are noisy, or the two means are close in value. This basically arises from the inverse relationship between the variance of $\bar{Y}$ and n, i.e., $Var[\bar{Y}] = \sigma^2/n$, which is incorporated in the test statistic Z_s (see Eqn. 10.1).

10.6 Summary table

A common way of summarizing the different outcomes in hypothesis testing is the table below. The null hypothesis H_0 can be either true or false. If H_0 is true, then the test may accept H_0 and make a correct decision, or reject it and make a Type I error, with a Type I error rate of α. If H_0 is false, then the test may accept H_0 and make a Type II error with an error rate of β, or reject it and make a correct decision.

Table 10.2: Table summarizing the different outcomes in hypothesis testing, with the corresponding Type I (α) and Type II (β) error rates.

	Accept H_0	Reject H_0
H_0 true	Correct 1-α	Type I error α
H_0 false	Type II error β	Correct 1-β = power

10.7 One-sample t test

In the preceding, we used the test statistic Z_s to test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$, for the case where σ^2 or σ was known. Although this simplifies the statistics, in most cases we will need to estimate σ^2 and σ from the data using the sample variance s^2 and standard deviation s. We then use the test statistic

$$T_s = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} \tag{10.11}$$

to conduct the test (Bickel & Doksum 1977). T_s has a t distribution with $n-1$ degrees of freedom under H_0 (see Chapter 9). The following is therefore a true statement:

$$P[-c_{\alpha,n-1} < T_s < c_{\alpha,n-1}] = P[-c_{\alpha,n-1} < \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} < c_{\alpha,n-1}] = 1 - \alpha. \tag{10.12}$$

The quantity $c_{\alpha,n-1}$ would be chosen using Table T, using the entry for $2(1-p)$ corresponding to α and the appropriate degrees of freedom (see Chapter 9). The interval $(-c_{\alpha,n-1}, c_{\alpha,n-1})$ is the acceptance region of a test with a Type I error rate of α, while the rejection region is its complement.

For example, with $\alpha = 0.05$ and $n = 10$, we have $c_{0.05,9} = 2.262$. We would therefore accept H_0 if T_s lies within $(-2.262, 2.262)$, and reject it if T_s lies outside this interval (see Fig. 10.9). Using absolute values, we would accept H_0 if $|T_s| < 2.262$ and reject it otherwise. For $\alpha = 0.01$ and $n = 10$, we have $c_{0.01,9} = 3.250$, and would accept H_0 if T_s lies within $(-3.250, 3.250)$ and reject it otherwise.

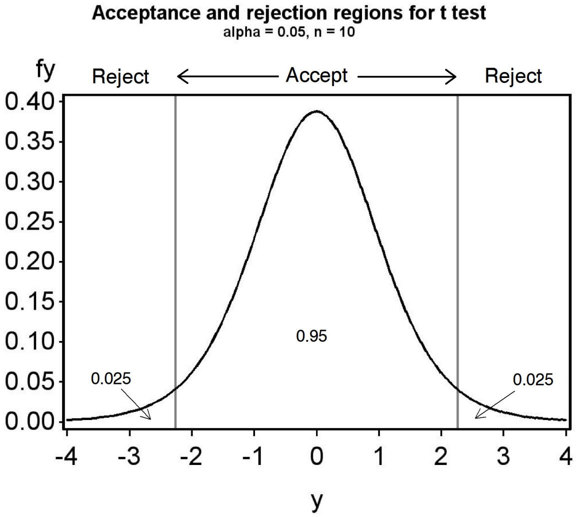


Figure 10.9: Acceptance and rejection regions for a one-sample t test, $\alpha = 0.05, n = 10$. The distribution shown is for the t distribution with $n - 1 = 9$ degrees of freedom.

10.7.1 One-sample t test - sample calculation

Recall the tilapia example, and suppose that $\bar{Y} = 493$ g and $s^2 = 48.2$ g^2, so that $s = 6.94$ g, with $n = 10$. We wish to test $H_0 : \mu = 500$ g vs. $H_1 : \mu \neq 500$ g. For the test statistic, we have

$$T_s = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} = \frac{493 - 500}{6.94/\sqrt{10}} = \frac{-7}{2.19} = -3.196 \tag{10.13}$$

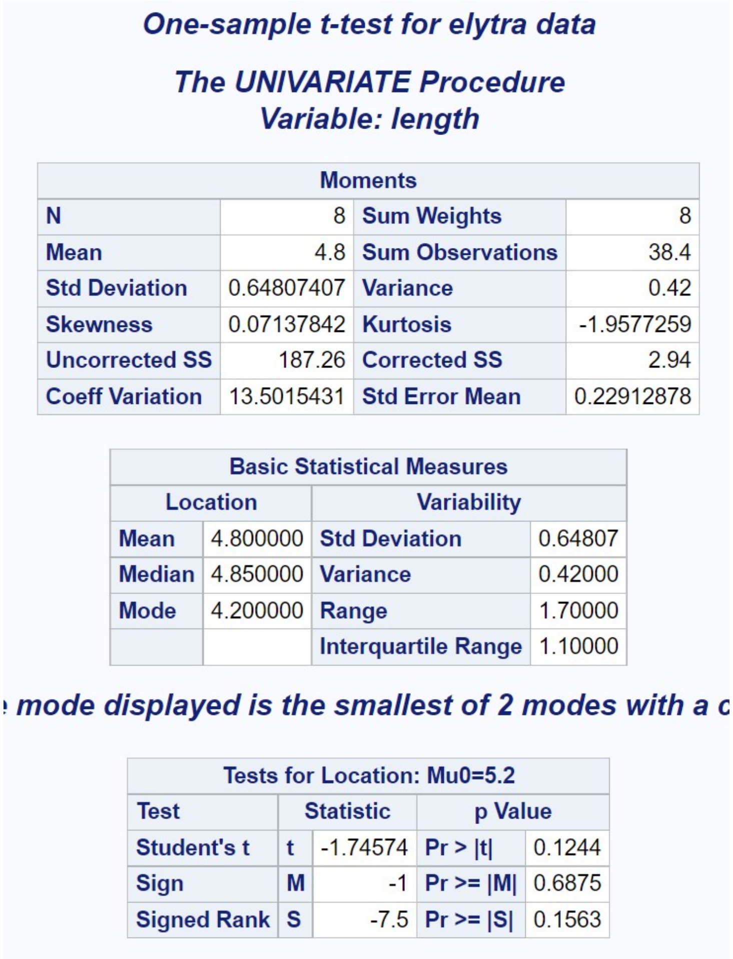
For $\alpha = 0.05$, the acceptance region for T_s is $(-2.262, 2.262)$ with $n - 1 = 10 - 1 = 9$ degrees of freedom (Fig. 10.9). $T_s = -3.196$ lies outside this interval, so we would reject H_0 at the $\alpha = 0.05$ level. For $\alpha = 0.01$ the acceptance region is $(-3.250, 3.250)$. Because T_s lies within this interval, we would accept H_0 at this α level. We can also determine a P value for this test using Table T. The P value is found by scanning along the row in the table corresponding to 9 degrees of freedom, looking for two values that bracket T_s while ignoring its sign. We see that the values 2.821 and 3.250 bracket $T_s = -3.196$. Looking at the values for $2(1 - p)$, which correspond to α, this implies that $0.010 < P < 0.020$. This is the best accuracy that can be accomplished using Table T, and to obtain an exact P value would require the use of SAS.

10.7.2 Hypothesis testing - SAS demo

To illustrate hypothesis testing using SAS, we will use a subset ($n = 8$) of the elytra data for the insect predator *Thanasimus dubius* (see Chapter 3). These observations are from a study that used an artificial diet to rear the insects, and we would like to compare their size to wild individuals. Suppose that wild predators have an elytral length of 5.2 mm. This suggests testing $H_0 : \mu = 5.2$ mm vs. $H_1 : \mu \neq 5.2$ mm. We can conduct a one-sample t test for this null hypothesis using `proc univariate`, by adding the option `mu0=5.2` as an option. See SAS program and Fig. 10.10 below. The test statistic T_s and its P value are listed on one line at the bottom of the output. We see that $T_s \approx -1.75$ for this test. What is its P value? The notation Pr > |t| in the output is shorthand for the $P[T_s < -1.75] + P[T_s > 1.75]$, the P value for this two-tailed test. We thus have $P = 0.1244$, a non-significant test result because $P > 0.05$. The degrees of freedom for the test are not reported by SAS, but are equal to $n - 1 = 8 - 1 = 7$. A sentence reporting this test result in a scientific journal would be something like 'A one-sample t test comparing the elytra length of individuals reared on artificial diet vs. wild individuals was non-significant ($t_7 = -1.746, P = 0.1244$).' Note that the degrees of freedom are reported as a subscript on the test statistic.

SAS Program

```
* one-sample_t_test.sas;
title 'One-sample t-test for elytra data';
data elytra;
    input sex $ length;
    datalines;
F   5.2
F   4.2
F   5.7
F   5.4
F   4.0
F   4.5
F   5.2
F   4.2
;
run;
* Generate t test and plots;
proc univariate mu0=5.2 data=elytra;
    var length;
    histogram length / vscale=count normal;
    qqplot length / normal;
run;
quit;
```



One-sample t-test for elytra data

The UNIVARIATE Procedure
Variable: length

Moments			
N	8	Sum Weights	8
Mean	4.8	Sum Observations	38.4
Std Deviation	0.64807407	Variance	0.42
Skewness	0.07137842	Kurtosis	-1.9577259
Uncorrected SS	187.26	Corrected SS	2.94
Coeff Variation	13.5015431	Std Error Mean	0.22912878

Basic Statistical Measures			
Location		Variability	
Mean	4.800000	Std Deviation	0.64807
Median	4.850000	Variance	0.42000
Mode	4.200000	Range	1.70000
		Interquartile Range	1.10000

mode displayed is the smallest of 2 modes with a c

Tests for Location: Mu0=5.2				
Test	Statistic		p Value	
Student's t	t	-1.74574	Pr > |t|	0.1244
Sign	M	-1	Pr >= |M|	0.6875
Signed Rank	S	-7.5	Pr >= |S|	0.1563

Figure 10.10: `one-sample_t_test.sas - proc univariate`

10.7.3 Power analysis for one-sample t tests - SAS demo

A power analysis can be used to determine an adequate sample size n for a one-sample t test, as well as many other statistical tests. To conduct a power analysis, you need to specify a null and alternative hypothesis, a Type I error rate α, and have some estimate of the standard deviation σ of the population in question. The analysis then calculates the power for a range of n values. **The idea is to choose a value of n that gives power close to 0.8, often regarded as an adequate level of power (Cohen 1988).** The power analysis for a one-sample t test involves the same quantity

$$\phi = \frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}} \tag{10.14}$$

as for the one-sample Z test, and its power is influenced by the same factors (see Table 10.1). The power calculation involves the **non-central t distribution** with a non-centrality parameter of ϕ. One subtle difference is that acceptance and rejection regions for the t test depends on n through the degrees of freedom, unlike the Z test. Larger values of n lead to smaller values of $c_{\alpha,n-1}$, shrinking the acceptance region and affecting the power calculation in this way.

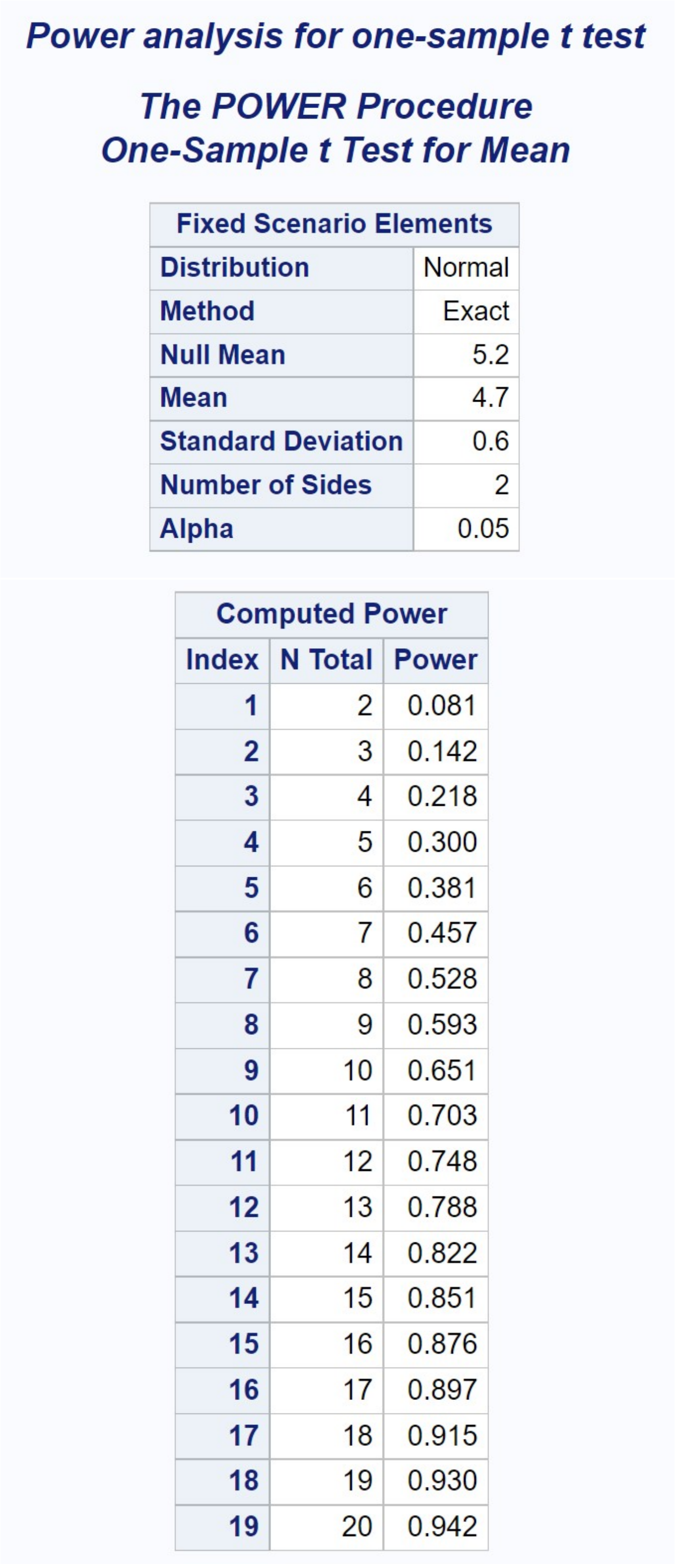
Returning to the elytra example, suppose we want to test if the length of predators reared on an artificial diet differs from wild individuals, which have a length of 5.2 mm. This implies $H_0 : \mu = 5.2$ mm. For biological reasons, we are interested in detecting an decrease or increase in length of approximately 10% on the artificial diet, about 0.5 mm. This suggests an alternative hypothesis of the form $H_1 : \mu = 5.2 - 0.5 = 4.7$ mm (or $H_1 : \mu = 5.2 + 0.5 = 5.7$ mm). How many predators need to be reared on artificial diet to give a power of at least 0.8? Assume we already have an estimate of σ from another study, say $s = 0.6$ mm, and let $\alpha = 0.05$.

We can use `proc power` to find the sample size n that gives this power (SAS Institute Inc. 2018). See program plus Fig. 10.11 and 10.12 below. We first specify a one-sample t test using the `onesamplemeans` option, followed by values for μ under H_0 (`nullmean = 5.2`), σ (`stddev = 0.6`), and μ under H_1 (`mean = 4.7`). The default value of α is 0.05. We then specify a range of sample sizes (n) for which we want the power to be calculated, using the option `ntotal = 2 to 20 by 1`. This finds the power for $n = 2, 3, \ldots, 20$. The `power = .` option tells SAS solve for power (there are other possibilities, like finding n for a given power value). The option `plot x=n` generates a plot of power vs. n. We see that a sample size of $n = 14$ gives power > 0.8 for this scenario.

While power increases rapidly for small sample sizes, there are diminishing returns once the power exceeds about 0.8. In other words, obtaining higher power values requires many more observations.

SAS Program

```
* One-sample_t_test_power2.sas;
title 'Power analysis for one-sample t test';
proc power;
    onesamplemeans
        nullmean = 5.2
        stddev = 0.6
        mean = 4.7
        ntotal = 2 to 20 by 1
        power = . ;
    plot x=n;
run;
quit;
```



Power analysis for one-sample t test

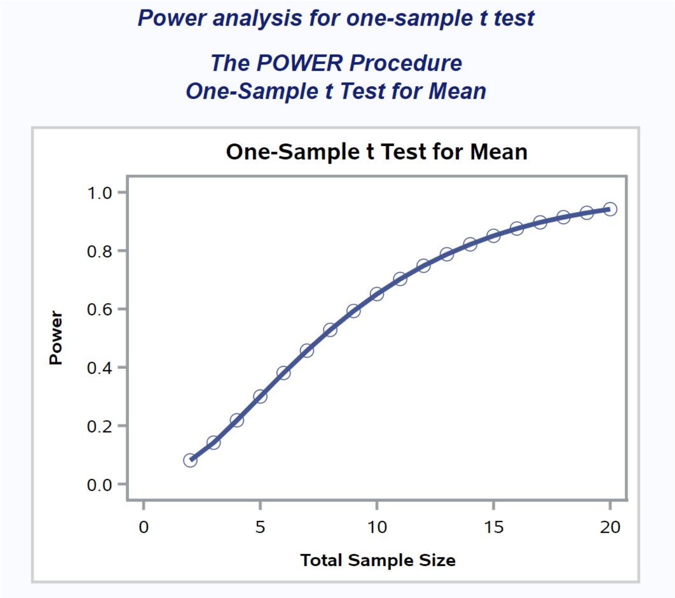
The POWER Procedure
One-Sample t Test for Mean

Fixed Scenario Elements	
Distribution	Normal
Method	Exact
Null Mean	5.2
Mean	4.7
Standard Deviation	0.6
Number of Sides	2
Alpha	0.05

Computed Power

Index	N Total	Power
1	2	0.081
2	3	0.142
3	4	0.218
4	5	0.300
5	6	0.381
6	7	0.457
7	8	0.528
8	9	0.593
9	10	0.651
10	11	0.703
11	12	0.748
12	13	0.788
13	14	0.822
14	15	0.851
15	16	0.876
16	17	0.897
17	18	0.915
18	19	0.930
19	20	0.942

Figure 10.11: `one-sample_t_test_power2.sas - proc power`



Power analysis for one-sample t test

The POWER Procedure
One-Sample t Test for Mean

Figure 10.12: `one-sample_t_test_power2.sas - proc power`

10.8 One-tailed t test

The tests we have examined so far are known as two-tailed tests. They are called this because the test statistic Z_s or T_s can detect departures from $H_0 : \mu = \mu_0$ in both directions, for $H_1 : \mu > \mu_0$ and $H_1 : \mu < \mu_0$, although the alternative for these tests is usually written more compactly as $H_1 : \mu \neq \mu_0$. We will now examine one-tailed tests, which have the same null hypothesis but the alternative is one direction or the other.

Suppose we are interested in testing $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$. We can use the same test statistic as before, namely

$$T_s = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}}. \tag{10.15}$$

If H_1 is true, we would expect to see $\bar{Y}$ values larger than μ_0, and so T_s would be positive. We would reject H_0 if T_s was sufficiently positive, with the acceptance and rejection regions determined as before by controlling the Type I error rate. Therefore, if the Type I error rate is α we want to determine a constant $c'_{\alpha,n-1}$ such that the following statement is true:

$$P[T_s < c'_{\alpha,n-1}] = 1 - \alpha \tag{10.16}$$

The quantity $c'_{\alpha,n-1}$ would be chosen using Table T, using the entry for p corresponding to $1 - \alpha$. We would accept H_0 if $T_s < c'_{\alpha,n-1}$ and reject it if $T_s \geq c'_{\alpha,n-1}$.

For example, with $\alpha = 0.05$ so that $p = 0.95$, and $n = 10$ (degrees of freedom $= n - 1 = 10 - 1 = 9$), we have $c'_{0.05,9} = 1.833$. We would therefore accept H_0 if $T_s < 1.833$ and reject it if $T_s \geq 1.833$ (see Fig. 10.13). For $\alpha = 0.01$ and $n = 10$, we have $c'_{0.01,9} = 2.822$, and would accept H_0 if $T_s < 2.822$ and reject it otherwise.

If we now wish to test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu < \mu_0$, we would use the same test statistic as above. However, if H_1 is true we would expect $\bar{Y}$ to be smaller than μ_0, and so T_s would be negative. To determine the acceptance and rejection regions we would find $c'_{\alpha,n-1}$ in the same way as above, except we would use its negative. We would accept H_0 if $T_s > -c'_{\alpha,n-1}$ and reject it if $T_s \leq -c'_{\alpha,n-1}$. For example, if $\alpha = 0.05$ and $n = 10$, we would accept H_0 if $T_s > -1.833$ and reject it if $T_s \leq -1.833$ (Fig. 10.14). For $\alpha = 0.01$, we would accept H_0 if $T_s > -2.822$ and reject it otherwise.

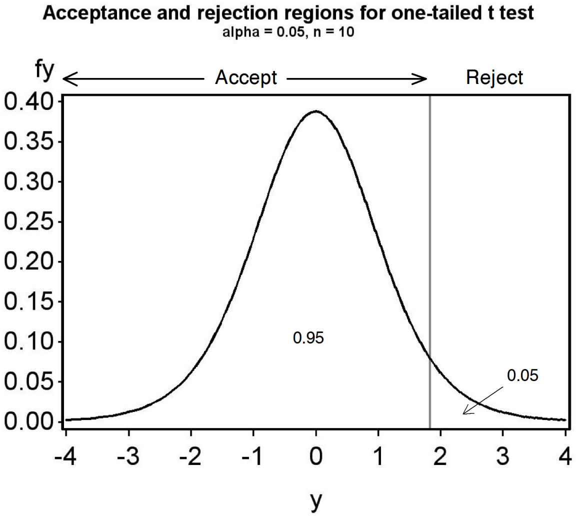


Figure 10.13: Acceptance and rejection regions for one-tailed t test, $H_0 : \mu = \mu_0$ vs. $H_1 : \mu > \mu_0$, for $\alpha = 0.05$ and $n = 10$.

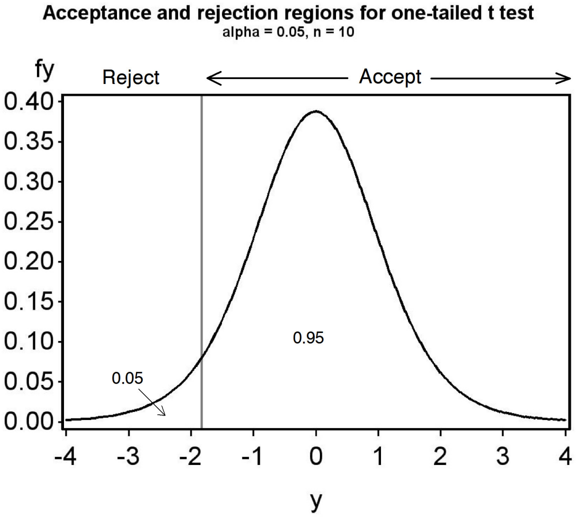


Figure 10.14: Acceptance and rejection regions for a one-tailed t test, $H_0 : \mu = \mu_0$ vs. $H_1 : \mu < \mu_0$, for $\alpha = 0.05$ and $n = 10$.

10.8.1 One-tailed t test - sample calculation

Recall the tilapia example, with $\bar{Y} = 493$ g, $s^2 = 48.2$ g^2, $s = 6.94$ g, and $n = 10$. Suppose we are only interested in detecting diets that produce fish of lower weight than natural food, implying we wish to test $H_0 : \mu = 500$ g vs. $H_1 : \mu < 500$ g. The test statistic value is again

$$T_s = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} = \frac{493 - 500}{6.94/\sqrt{10}} = \frac{-7}{2.19} = -3.196 \qquad (10.17)$$

For $\alpha = 0.05$ and $n - 1 = 10 - 1 = 9$ degrees of freedom, we have $-c'_{0.05,9} = -1.833$. Because $T_s = -3.916 < -1.833$, we would reject H_0 at the $\alpha = 0.05$ level. For $\alpha = 0.01$, we have $-c'_{0.01,9} = -2.821$, and again $T_s = -3.196 < -2.821$. Thus, we can also reject H_0 at the $\alpha = 0.01$ level. We could continue this process with successively smaller values of α by scanning the row corresponding to 9 degrees of freedom in Table T, but cannot reject H_0 for smaller ones. Therefore, we have $P < 0.01$ for this test.

Suppose we had wanted to test $H_0 : \mu = 500$ g vs. $H_1 : \mu > 500$ g using the same data and test statistic value, namely $T_s = -3.196$. The scenario here could be that we want a commercial diet that actually increases the weight of tilapia over natural food, and are not interested in ones that yield lower weights. In this case, for $\alpha = 0.05$ we would not reject H_0, because $T_s = -3.196 < 1.833$. The test was non-significant, with $P > 0.05$.

10.8.2 One-tailed t test - SAS demo

Recall the elytra length example, where we tested $H_0 : \mu = 5.2$ mm vs. $H_1 : \mu \neq 5.2$ mm using SAS (Fig. 10.10). While there is no option for one-tailed tests in `proc univariate`, we can reinterpret the output and so derive a P value for a one-tailed test.

Suppose we want to test $H_0 : \mu = 5.2$ mm vs. $H_1 : \mu < 5.2$ mm. This implies we want to test whether predators reared on artificial diet are smaller than those reared on natural food, which have a length of 5.2 mm. This would be reasonable if we want to detect diets that are deficient in some manner. If H_1 were true we would expect to see a negative value of T_s, because $\bar{Y}$ would likely be smaller than μ_0. This is what occurred in the SAS output, because $\bar{Y} = 4.8 < 5.2$ mm and $T_s = -1.75$. The one-tailed P value in this case is simply half the two-tailed P value, or P(one-tailed) = P(two-tailed)/2= 0.1244/2 = 0.0622. This is because the two-tailed test gives the P value for both tails (see Fig. 10.9), but for this one-tailed test we only need the probability for the left tail of the t distribution (Fig. 10.14).

Now suppose we want to test $H_0 : \mu = 5.2$ mm vs. $H_1 : \mu > 5.2$ mm. This implies we want to test whether predators reared on artificial diet are larger than those reared on natural food. If H_1 were true we would expect to see a positive value of T_s, because $\bar{Y}$ would likely be greater than μ_0. This is not what occurred in the SAS output, because $\bar{Y} = 4.8 < 5.2$ mm and $T_s = -1.75$. The P value should therefore be large in this case, and in fact the one-tailed P value is $1 - P(\text{two-tailed})/2 = 1 - 0.1244/2 = 0.9378$. This is the probability for the right tail of the t distribution, which is large because T_s is negative.

We can distill the above procedures to a simple rule that will convert the SAS two-tailed P value to the appropriate one-tailed one. Assume $H_0 : \mu = \mu_0$ is the null hypothesis. **If the test statistic favors the alternative hypothesis, then the one-tailed P value is $P(\textbf{two-tailed})/2$, otherwise it is $1 - P(\textbf{two-tailed})/2$.** For example, if we have $H_1 : \mu > \mu_0$ and $T_s > 0$, the test statistic favors H_1 and the P value is $P(\text{two-tailed})/2$. This procedure also works for tests calculated by hand. You first find the P value for the two-tailed test, then convert it to a one-tailed P value using the same rule.

10.8.3 One-tailed tests - a warning

As discussed above, the P value for a one-tailed test may sometimes be half the two-tailed P value. This makes it tempting to employ a one-tailed test after a two-tailed test yields a nonsignificant result. However, the proper procedure is to determine whether a one-tailed alternative hypothesis and test is appropriate for the situation **before** conducting the test. For example, artificial diets for insects are unlikely to yield larger insects than natural diets, and so it seems reasonable to use an alternative hypothesis of the form $H_1 : \mu < \mu_0$, where μ_0 is the size of insects reared on natural foods. This choice of an alternative hypothesis can be justified based on prior knowledge of the system.

10.9 Confidence intervals and tests

Confidence intervals are typically used as measures of the accuracy or reliability of parameter estimates, but can also be used for hypothesis testing. Why might you do this? There are cases where the statistical software only provides confidence intervals for a parameter, but a test can still be developed using these intervals. Also, a publication may only provide confidence intervals for a parameter, but the reader can still conduct a test if required using these intervals. Some statisticians argue that this makes confidence intervals more useful than hypothesis testing, because they also provide information on the magnitude of a population parameter, and how reliably it is estimated (see Yaccoz 1991).

We will now demonstrate how a confidence interval for μ is equivalent to a one-sample t test. Recall that a $100(1-\alpha)\%$ confidence interval for μ has the form

$$\left(\bar{Y} - c_{\alpha,n-1}\frac{s}{\sqrt{n}}, \bar{Y} + c_{\alpha,n-1}\frac{s}{\sqrt{n}}\right) \tag{10.18}$$

(see Chapter 9). Suppose that we want to test $H_0 : \mu = \mu_0$. If we accept H_0 when this confidence interval includes μ_0, and reject it if the interval does not include μ_0, this is an α level test of H_0, equivalent to running a one-sample t test.

To see this connection, note that we would accept H_0 if μ_0 was inside this interval, or

$$\bar{Y} - c_{\alpha,n-1}\frac{s}{\sqrt{n}} < \mu_0 < \bar{Y} + c_{\alpha,n-1}\frac{s}{\sqrt{n}}. \tag{10.19}$$

Rearranging these inequalities, we see it is equivalent to saying

$$-c_{\alpha,n-1} < \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} < c_{\alpha,n-1}, \tag{10.20}$$

or

$$-c_{\alpha,n-1} < T_s < c_{\alpha,n-1}, \tag{10.21}$$

where T_s is the test statistic for a one-sample t test. We would reject H_0 if T_s falls outside this interval. Note that this acceptance region is exactly the same as for the t test with Type I error rate of α, which is of the form $(-c_{\alpha,n-1}, c_{\alpha,n-1})$. Thus, the test based on a $100(1-\alpha)\%$ confidence interval is equivalent to an α level test. In particular, a 95% confidence interval is equivalent to an $\alpha = 0.05$ test.

Conversely, it is often possible to reverse this process and obtain a confidence interval from a statistical test. The procedure is called 'inverting the test' (Bickel & Doksum 1977).

10.10 Likelihood ratio tests

We saw earlier how statisticians use the concept of maximum likelihood to estimate population parameters (Chapter 8). The maximum likelihood method begins by constructing a likelihood function based on the distribution of the data (Poisson, normal, etc.) and the observed data. We then maximize the likelihood as a function of the parameters of the distribution (μ, σ^2, etc). The values of the parameters that maximize the likelihood are the maximum likelihood estimates of the parameters. The likelihood function is not a fixed quantity but instead varies with the observed data, so that different data sets yield different estimates of the population parameters. Maximum likelihood estimators have desirable statistical properties and in many cases yield estimators that seem reasonable (like using $\bar{Y}$ to estimate μ).

Likelihood methods can also be used to develop statistical tests called **likelihood ratio tests**. These tests also have desirable statistical properties and in many cases are identical to classical statistical tests. Likelihood methods thus provide a theoretical framework for many statistical problems, including parameter estimation, confidence intervals, and hypothesis testing. The main drawback of these methods is that one must be willing to specify the distribution of the data, be it Poisson, binomial, normal, or more exotic distributions.

10.10.1 Example of a likelihood ratio test

We will now develop a likelihood ratio test that leads to the familiar one-sample t test (Mood et al. 1974) . We suppose that the data are normally distributed and we wish to test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0$. A random sample with n observations has been obtained.

We can think of H_0 and H_1 as two different statistical models for the data. Under H_0, the data are assumed to be normally distributed with $\mu = \mu_0$, but can have any value of σ^2 because this parameter is left unspecified. Under H_1, the data are permitted to have any value of μ and σ^2.

The first step in constructing a likelihood ratio test is to find the maximum likelihood estimates of the parameters for each of these two statistical models. We have already dealt with this problem for the model specified by H_1 – this is just maximum likelihood estimation of μ and σ^2 for the normal distribution. The same methods can be used to estimate σ^2 under H_0, but we will not go into the details.

This process can be illustrated by plotting the likelihood function as a function of μ and σ^2. To make things more concrete, we show the likelihood function for a data set with three data points ($Y_1 = 4.5$, $Y_2 = 5.3$, and $Y_3 = 5.4$). Also shown is a possible null hypothesis for these data, such as $H_0 : \mu = 4.7$. See Fig. 10.15 below.

The maximum likelihood estimates of μ and σ^2 under H_1 are the values of μ and σ^2 found at the peak of the likelihood function. However, the maximum likelihood estimate of σ^2 under H_0 occurs at a different location. Because μ is fixed at 4.7 under H_0, σ^2 is only free to vary along the vertical line shown in the figure. The maximum likelihood estimate of σ^2 under H_0 is the value of σ^2 that maximizes the likelihood along this line.

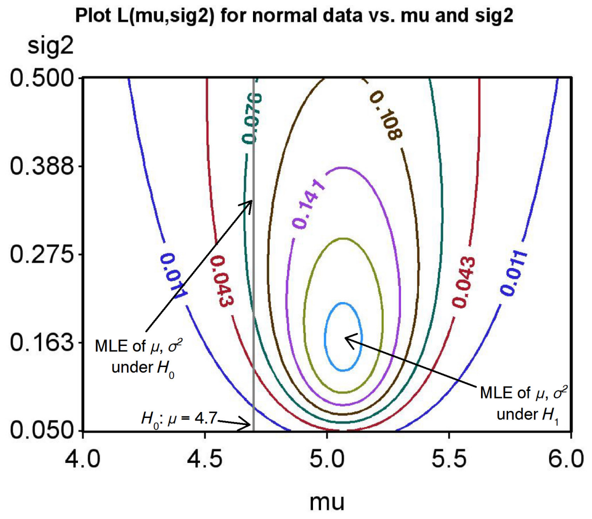


Figure 10.15: Likelihood ratio test for $H_0 : \mu = 4.7$

We are now ready to construct the likelihood ratio test statistic. Let L_{H_0} be the maximum height of the likelihood surface under H_0, which occurs at the maximum likelihood estimate of σ^2 under H_0. Similarly, let let L_{H_1} be the maximum height under H_1, which occurs at the estimates of μ and σ^2 under H_1. The test statistic λ is just the ratio of these two quantities:

$$\lambda = \frac{L_{H_0}}{L_{H_1}}. \tag{10.22}$$

How does this statistic behave? If H_0 is true, the peak of the likelihood function will often be near the vertical line, and the height of the likelihood function will be similar at the two locations. This implies a value of $\lambda \approx 1$ because $L_{H_0} \approx L_{H_1}$. If H_0 is false and H_1 true, however, we would expect to see $L_{H_0} < L_{H_1}$ and so $\lambda < 1$. We would therefore reject H_0 for sufficiently small values of λ.

More formally, we reject H_0 if $\lambda < c$ and accept H_0 otherwise. The value of c is determined using the Type I error rate α and the distribution of λ under H_0.

An alternate form of the test uses $-2\ln(\lambda)$ rather than λ itself, and rejects H_0 for values of $-2\ln(\lambda) > d$, where d is a constant that controls the Type I error rate. This form of the test rejects for large values of the test statistic, similar to other tests we have developed. Note that

$$-2\ln(\lambda) = 2\ln(L_{H_1}) - 2\ln(L_{H_0}) \tag{10.23}$$

by the properties of logarithms, and is a positive quantity. SAS provides values of the likelihood function in this format for some statistical procedures, and these can be used to construct likelihood ratio tests.

How is the likelihood ratio test related to a t test? It can be shown mathematically that the value of the test statistic

$$T_s = \frac{\bar{Y} - \mu_0}{s/\sqrt{n}} \tag{10.24}$$

is directly proportional to $-2\ln(\lambda)$, the likelihood ratio test statistic (Mood et al. 1974). Figure 10.16 plots the value of $-2\ln(\lambda)$ vs. T_s for a scenario matching our example data set. We observe there is a one-to-one correspondence between the two test statistics. When such a correspondence occurs between two test statistics, the tests are considered to be statistically equivalent. We will later see that many statistical tests are in fact likelihood ratio tests. These include tests in analysis of variance, regression, and methods for categorical data such as χ^2 tests.

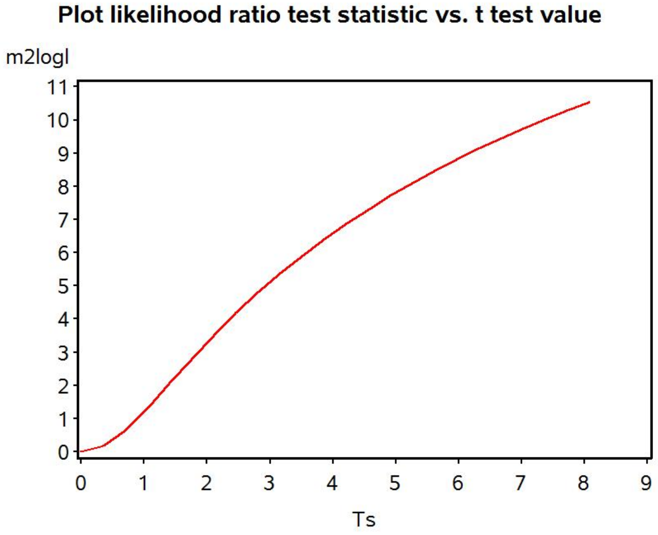


Figure 10.16: Likelihood ratio vs. t test statistics.

10.11 References

Bickel, P. J. & Doksum, K. A. (1977) *Mathematical Statistics: Basic Ideas and Selected Topics.* Holden-Day, Inc., San Francisco, CA.

Mood, A. M., Graybill, F. A. & Boes, D. C. (1974) *Introduction to the Theory of Statistics.* McGraw-Hill, Inc., New York, NY.

SAS Institute Inc. (2018) *SAS/STAT 15.1 Users Guide* SAS Institute Inc., Cary, NC

Yaccoz, N. G. (1991) Use, overuse, and misuse of significance tests in evolutionary biology and ecology. *Bulletin of the Ecological Society of America* 72: 106-111.

10.12 Problems

1. A company that rears beneficial insects produces lacewings (Chrysopidae: Neuroptera) whose mean length is 10 mm. A new method of rearing is being tested and the company wants to determine if the new method changes lacewing length. A sample of 10 insects is collected for the new method, yielding the following lengths:

    ```
    10.3  14.1  11.5  9.9  12.6  9.7  11.0  9.5  12.4  13.5
    ```

    (a) Test whether the lacewings produced using the new method have the same length as before ($H_0 : \mu = 10$ vs. $H_1 : \mu \neq 10$), using a two-tailed test and Table T. Provide a P value and discuss the significance of the test. Show your calculations.

    (b) Suppose the company is only interested in rearing methods that yield larger lacewing lengths, because bigger is better with beneficial insects. Test $H_0 : \mu = 10$ vs. $H_1 : \mu > 10$. Provide a P value and discuss the significance of the test.

    (c) Use SAS and `proc univariate` to carry out the same two tests. What are the exact P values for these tests? Attach your SAS program and printout.

2. A study is done to measure the concentration of a particular chemical (ppm) in drinking water, with samples taken at eight locations. The samples were analyzed and the following results obtained:

    ```
    23 20 24 20 23 24 21 22
    ```

    (a) Test whether the concentration of the chemical is significantly different from 20 ppm, the level set by the EPA, using a two-tailed test and Table T. Provide a P value and discuss the significance of the test. Show your calculations.

    (b) The EPA actually requires that the concentration of the chemical be equal to or below 20 ppm. Test whether the chemical concentration exceeds this level using a one-tailed test and Table T. In particular, test $H_0 : \mu = 20$ vs. $H_1 : \mu > 20$. Provide a P value and discuss the significance of the test.

(c) Use SAS and `proc univariate` to carry out the same two tests. What are the exact P values for these tests? Attach your SAS program and printout.

Chapter 11

Analysis of Variance (One-Way)

We now develop a statistical procedure for comparing the means of two or more groups, known as analysis of variance or ANOVA. These groups might be the result of an experiment in which organisms are exposed to different treatments. Alternately, the groups might be different species or different age classes of the same species, populations in different locations, or different genetic families. The test works by comparing the variance among the group means to the variance of the observations within each group – if the variance among group means is large (implying differences in their means) relative to the variance within groups, the test is significant. This chapter will examine tests for one-way ANOVA, in which a single factor like a treatment affects the observations. More complex designs are possible in which several factors may influence the observations and may also interact (see Chapter 14 and 19).

What do the data look like for a one-way ANOVA design? Suppose we are interested in trapping bark beetles (Coleoptera: Curculionidae: Scolytinae) using different chemical baits, which could involve the beetle's sex pheromones or odors of the trees they colonize. Suppose there are three different baits (A, B, and C), with $a = 3$ denoting the number of treatments. The baits are deployed on traps in the forest, with $n = 5$ replicate traps for each bait type. A typical experimental design would establish 15 traps in the forest, and then randomly assign a bait to each trap. After a period of time, the traps would be checked and the number of insects caught in each trap recorded (Table 11.1). Because the data are counts, it would not be normally distributed but more likely have a Poisson or negative binomial distribution (see Chapter 5). However, it is often possible to **transform** count data to have a distribution closer to the normal by taking the square root or log of the counts (see Chapter 15). The third column in Table 11.1 shows the count data after applying a log transformation. The notation Y_{ij} is often used to refer to the observations in ANOVA designs. The i subscript refers to the group or treatment, while j is the observation within the treatment. For example, Y_{13} refers to the third observation in the first treatment, which is 2.41.

Another one-way ANOVA design for bark beetles might simply look at variability in their density across sites. Suppose there is a large collection of study sites, and we randomly select five sites for trapping. Five traps are deployed at each of the five sites and the number of beetles caught per trap is recorded. Data for a study of this type are listed below, also with a log transformation (Table 11.2). There appears to be substantial variability in beetle abundance across sites, with Site 4 having very high beetle catches, while Site 5 has low ones.

The data sets presented in this section represent **balanced designs**, because there are the same number of replicates for each treatment or group. An **unbalanced design** would have an unequal number of replicates, possibly very unequal. We will present tests and theory for balanced designs in this chapter, because this greatly simplifies the formulas and equations. However, these results can be readily extended to unbalanced designs, and unbalanced designs require no changes in the SAS programs presented.

Table 11.1: Example 1 - Bark beetles captured in a trapping experiment comparing the attraction to different baits. There were three baits (A, B, and C) and five replicate traps per bait treatment. Also shown are the log-transformed counts (Y_{ij}) and subscript values (i, j), and some preliminary one-way ANOVA calculations.

Treatment	Count	$Y_{ij} =$ $\log_{10}(\text{Count})$	i	j	$\bar{Y}_{i\cdot}$	$(Y_{ij} - \bar{Y}_{i\cdot})^2$	$\sum(Y_{ij} - \bar{Y}_{i\cdot})^2$
A	373	2.57	1	1	2.3600	0.0441	0.2110
A	126	2.10	1	2		0.0676	
A	255	2.41	1	3		0.0025	
A	138	2.14	1	4		0.0484	
A	379	2.58	1	5		0.0484	
B	25	1.40	2	1	1.7160	0.0999	0.1325
B	64	1.81	2	2		0.0088	
B	62	1.79	2	3		0.0055	
B	71	1.85	2	4		0.0180	
B	54	1.73	2	5		0.0002	
C	449	2.65	3	1	2.2220	0.1832	0.4581
C	249	2.40	3	2		0.0317	
C	69	1.84	3	3		0.1459	
C	199	2.30	3	4		0.0061	
C	84	1.92	3	5		0.0912	

Table 11.2: Example 2 - Bark beetles captured in a trapping study comparing their abundance at five randomly chosen study sites. There were five replicate traps per site. Also shown are the log-transformed counts (Y_{ij}) and subscript values (i, j), and some preliminary one-way ANOVA calculations.

Site	Count	$Y_{ij} =$ $\log_{10}$(Count)	i	j	$\bar{Y}_{i\cdot}$	$(Y_{ij} - \bar{Y}_{i\cdot})^2$	$\sum(Y_{ij} - \bar{Y}_{i\cdot})^2$
1	137	2.14	1	1	2.0120	0.0164	0.1598
1	101	2.00	1	2		0.0001	
1	113	2.05	1	3		0.0014	
1	48	1.68	1	4		0.1102	
1	155	2.19	1	5		0.0317	
2	156	2.19	2	1	2.4700	0.0784	0.4730
2	165	2.22	2	2		0.0625	
2	652	2.81	2	3		0.1156	
2	179	2.25	2	4		0.0484	
2	757	2.88	2	5		0.1681	
3	278	2.44	3	1	2.2460	0.0376	0.3419
3	197	2.29	3	2		0.0019	
3	95	1.98	3	3		0.0708	
3	395	2.60	3	4		0.1253	
3	83	1.92	3	5		0.1063	
4	2540	3.40	4	1	2.6960	0.4956	0.7600
4	613	2.79	4	2		0.0088	
4	200	2.30	4	3		0.1568	
4	251	2.40	4	4		0.0876	
4	390	2.59	4	5		0.0112	
5	18	1.26	5	1	1.1940	0.0044	0.0459
5	16	1.20	5	2		0.0000	
5	11	1.04	5	3		0.0237	
5	21	1.32	5	4		0.0159	
5	14	1.15	5	5		0.0019	

11.1 ANOVA models

We now examine the statistical models that are used in one-way ANOVA. There are two models for one-way ANOVA, known as fixed or random effects models, but sometimes called Model I and II. This classification is based on how the groups in the design are defined or generated. We begin by defining fixed and random effects, then present the statistical models and hypotheses for each type.

11.1.1 Fixed and random effects

For groups generated by different treatments in an experiment, or purposely chosen groups of organisms such as different species, sexes, or ages, the groups are classified as **fixed effects**. They are called fixed effects because these groups are the only ones of interest to the investigator, and the only ones on which a statistical inference can be made (Littell et al. 1996, McCulloch and Searle 2001). They are also incorporated in statistical models as fixed parameters. Groups that are generated by a process of random sampling are classified as a **random effects** (Littell et al. 1996, McCulloch and Searle 2001). For example, suppose we want to examine the fish populations in a large number of lakes, and are interested in how body length varies across lakes. If we randomly sample the lakes to be examined, from a large collection of lakes, then lake would be classified as a random effect. In many genetic experiments, families are chosen at random from a larger collection of families, making family a random effect. Random effects are incorporated in statistical models as random variables, typically with a normal distribution.

These definitions suggest a simple test for fixed vs. random effects – if the groups are a random sample from a large collection you have a random effect, otherwise a fixed effect. Although it is usually possible to declare an effect as either fixed or random, in practice it is sometimes difficult to decide. For example, suppose that a particular organism occurs at only a small number of locations. If we randomly select a subset of these locations to sample, seemingly implying a random effect, the overall number of locations is still finite. In this scenario, location may be better classified as a fixed effect.

11.1.2 Fixed effects model

Suppose that we want to model the observations in the bark beetle trapping experiment, Example 1, where different baits are used. Recall that the symbol Y_{ij} stand for the jth observation in the ith treatment group, where $i = 1, 2, 3$ and $j = 1, 2, 3, 4, 5$. For example, $Y_{11} = 2.57$ and $Y_{12} = 2.10$, while $Y_{32} = 2.40$ (see Table 11.1). One commonly used model for such a design is

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \tag{11.1}$$

where μ is a parameter setting the grand mean (the overall mean) of the observations, α_i is the deviation from the grand mean caused by the ith treatment (McCulloch and Searle 2001), and $\epsilon_{ij} \sim N(0, \sigma^2)$. It is usually assumed that $\sum \alpha_i = 0$, i.e., the treatment effect terms sum to zero. The ϵ_{ij} term represents random departures from the mean value for the ith treatment, due to natural variability among the observations. The ϵ_{ij} values are also assumed to be independent (Chapter 4). In practice, these parameters would be unknown but could be estimated from the data. The same model can be used to describe the observations for experiments with any number of treatments (any a value) as well as replicates per treatments (any n), as well as experiments where the number of observations is unequal among treatments.

It follows that for the ith treatment, $E[Y_{ij}] = \mu + \alpha_i$ and $Var[Y_{ij}] = \sigma^2$, using the rules for expected values and variances. Thus, for the ith treatment we have $Y_{ij} \sim N(\mu + \alpha_i, \sigma^2)$. We can illustrate how the different parameters work in this model with a diagram that plots the distribution for each group. Suppose that we want to model an experiment similar to the bark beetle trapping one, with $a = 3$ treatments. Suppose that $\mu = 2.1$, $\alpha_1 = 0.25$, $\alpha_2 = -0.40$, and $\alpha_3 = 0.15$, with $\sigma^2 = 0.1$. Fig. 11.1 shows the distribution of the observations in each treatment group. Note that the means for treatment 1 and 3 are shifted upward from the grand mean by their positive values of α_i, while the mean for treatment 2 is shifted downward by its negative value. The distribution for each treatment has the same variance, namely $\sigma^2 = 0.1$.

The usual objective in ANOVA is to test whether the means of the different groups are significantly different, implying there is treatment or group effect. In terms of the fixed effects model, this amounts to testing whether the α_i values are significantly different from zero, because it is these parameters that produce shifts in the group means from the grand mean. More formally, we are interested in testing the null hypothesis H_0 : all $\alpha_i = 0$.

Under H_0, all the groups have the same mean μ because the α_i terms are zero (Fig. 11.2). The alternative hypothesis would be H_1 : some $\alpha_i \neq 0$, i.e., there is some treatment effect on some (perhaps all) groups (Fig. 11.1). We will discuss how this null hypothesis is actually tested later in the chapter.

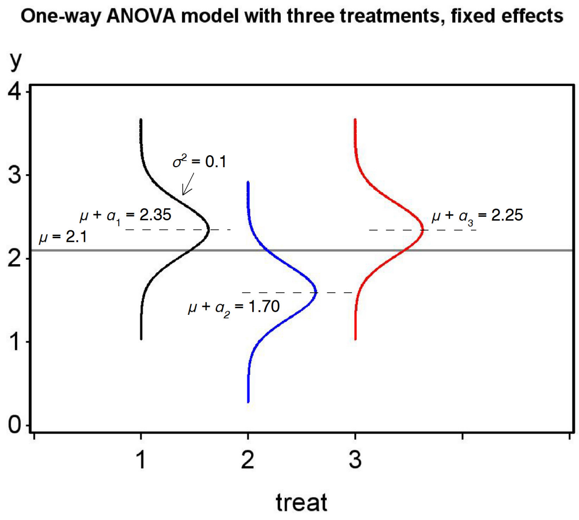


Figure 11.1: Fixed effects model for one-way ANOVA, under H_1 : some $\alpha_i \neq 0$.

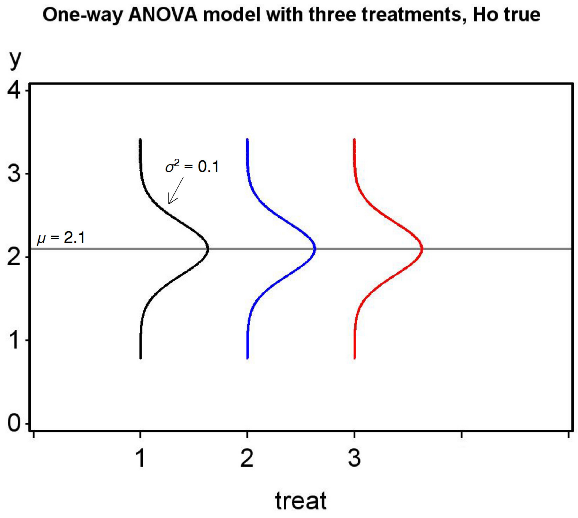


Figure 11.2: Fixed effects model for one-way ANOVA, under H_0 : all $\alpha_i = 0$.

11.1.3 Random effects model

Suppose that we now want to model the variability in bark beetle abundance across different sites, such as in the Example 2 study. Let Y_{ij} stand for the jth observation at the ith sampled site, with $i = 1, 2, 3, 4, 5$ and $j = 1, 2, 3, 4, 5$. We have $Y_{11} = 4.92$, $Y_{12} = 4.62$, and so forth (see Table 11.2). A common statistical model for this design is

$$Y_{ij} = \mu + A_i + \epsilon_{ij} \tag{11.2}$$

where μ is again a parameter setting the grand mean of the observations, with A_i a random deviation from the grand mean due to the ith site (McCulloch and Searle 2001), and $\epsilon_{ij} \sim N(0, \sigma^2)$. It is assumed that A_i is normally distributed with mean zero and variance σ_A^2, or $A_i \sim N(0, \sigma_A^2)$. Note that in the random effects model the group effect is indeed a random variable, one whose variance is unknown but can be estimated from the data. The variances σ_A^2 and σ^2 are collectively called the **variance components** of the model.

For the ith group sampled, it can be shown that $E[Y_{ij}] = \mu + A_i$ and $Var[Y_{ij}] = \sigma^2$, using the rules for expected values and variances. Thus, for the ith treatment we have $Y_{ij} \sim N(\mu + A_i, \sigma^2)$. Because the A_i values are themselves random quantities, however, the expected value is itself a random quantity and would differ for each group sampled. We again illustrate how the model works using a diagram showing the distribution for each group. Suppose that we want to model a study similar to the second bark beetle one (Table 11.2), with $a = 5$ sites randomly selected from a larger collection of sites. Suppose that $\mu = 2.1$ and $\sigma^2 = 0.1$. The first time we did this study, we might see a pattern like Fig. 11.3. If we redid the study and randomly selected another five sites, we would get a different pattern (Fig. 11.4). This illustrates that this model is not static like the fixed effects one, but instead would vary with the sites actually sampled. In the random effects model, we are usually interested in testing whether the variance of A_i is zero vs. greater than zero, or $H_0 : \sigma_A^2 = 0$ vs. $H_1 : \sigma_A^2 > 0$. Under $H_0 : \sigma_A^2 = 0$, all the A_i values must be zero (to give $\sigma_A^2 = 0$), and so all the groups have the same mean μ. A plot of the model under H_0 would therefore be similar to Fig. 11.2. This null hypothesis is tested in the same way as the one for the fixed effects model (see below).

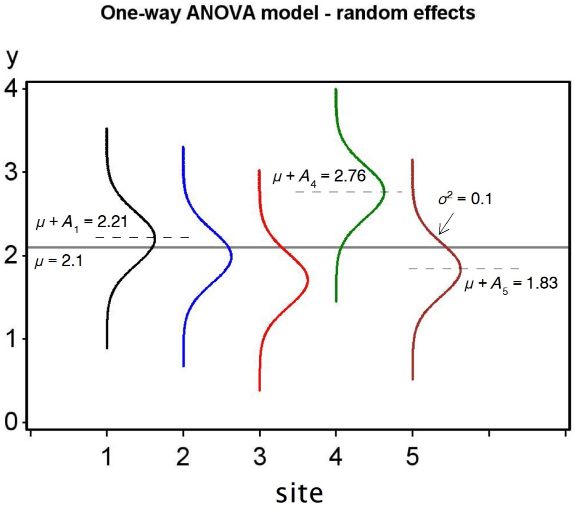


Figure 11.3: Random effects model for one-way ANOVA, for the first time sites are sampled.

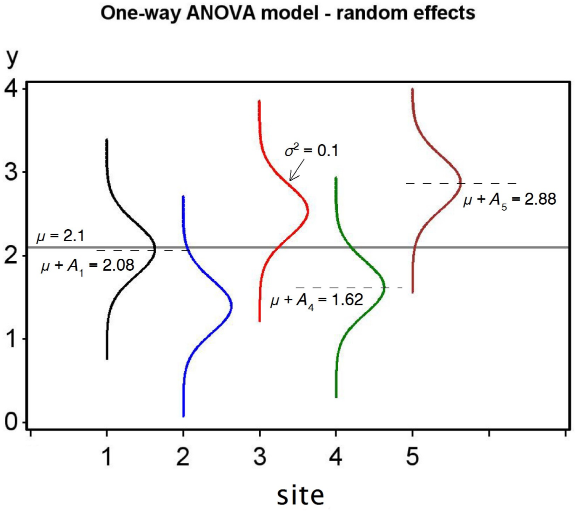


Figure 11.4: Random effects model for one-way ANOVA, for the second time sites are sampled.

11.2 Hypothesis testing for ANOVA

We now develop a statistical test for the null hypotheses in both fixed and random effects models, either H_0 : all $\alpha_i = 0$ or $H_0 : \sigma_A^2 = 0$. We will first present the test and explain how it works in terms of different estimates of the variance, then later show it is another example of a likelihood ratio test.

11.2.1 Sums of squares and mean squares

Suppose the data are described by a fixed effects model, for which the hypotheses are H_0 : all $\alpha_i = 0$ vs. H_1 : some $\alpha_i \neq 0$. It is clear that if H_1 is true, then the observations for the different groups will be shifted from the grand mean, as shown in Fig. 11.1, and in particular $Y_{ij} \sim N(\mu + \alpha_i, \sigma^2)$ for each group. For a random effects model, we have $H_0 : \sigma_A^2 = 0$ vs. $H_1 : \sigma_A^2 > 0$. If H_1 is true, we would also expected the observations for the different groups to be shifted away from the grand mean (Fig. 11.3), and in particular $Y_{ij} \sim N(\mu + A_i, \sigma^2)$. How can we estimate this shift in actual data? How large must this shift be to be judged statistically significant?

We begin by calculating the means for each group using the data. These are labeled as $\bar{Y}_{i\cdot}$ and are called group means. The '$\cdot$' subscript implies the mean was calculated using all the observations in that group ($j = 1, 2, \ldots, n$). We then calculate the mean of the group means, called the grand mean and labeled as $\bar{\bar{Y}}$. If the *ith* group is shifted from the grand mean, we can measure this shift using the quantity $\bar{Y}_{i\cdot} - \bar{\bar{Y}}$. In fact, this quantity estimates α_i for the *ith* group, and so is a direct measure of any group effect (see Section 11.3 on maximum likelihood estimation). If these quantities are small then this suggests H_0 might be true, whereas if they are large this provides evidence for H_1. We can obtain a single measure of these shifts by squaring and summing them across all groups, to obtain a quantity called the sum of squares among groups or SS_{among}, because it measures variation in the observations among groups:

$$SS_{among} = n \sum_{i=1}^{a} (\bar{Y}_{i\cdot} - \bar{\bar{Y}})^2. \tag{11.3}$$

Note the sample size n in this expression, which we will justify below. To make this quantity more concrete, we will calculate SS_{among} for Example 1, the bark beetle trapping experiment. We first calculate the sample mean for each group for the log-transformed data, as shown in Table 11.1. Then, the grand mean is estimated using the mean of these means, or

$$\bar{\bar{Y}} = \frac{\sum_{i=1}^{a} \bar{Y}_{i\cdot}}{a} = \frac{2.3600 + 1.7160 + 2.2220}{3} = \frac{6.2980}{3} = 2.0993. \tag{11.4}$$

We then have

$$\begin{aligned} SS_{among} &= n \sum_{j=1}^{a} (\bar{Y}_{i\cdot} - \bar{\bar{Y}})^2 && (11.5) \\ &= 5 \left[(2.3600 - 2.0993)^2 + (1.7160 - 2.0993)^2 + (2.2220 - 2.0993)^2 \right] && (11.6) \\ &= 5 \left[0.0680 + 0.1469 + 0.0151 \right] && (11.7) \\ &= 1.1500 && (11.8) \end{aligned}$$

SS_{among} has $a-1$ degrees of freedom, where a is the number of groups. There are $a-1$ degrees of freedom because there are a terms of the form $\bar{Y}_{i\cdot} - \bar{\bar{Y}}$ in the sum of squares, but these sum to zero so there are really only $a-1$ independent terms (similar to the $n-1$ degrees of freedom for the sample variance s^2). The next step is to convert SS_{among} to a sample variance, dividing it by $a-1$. This quantity is called the mean square among groups:

$$MS_{among} = \frac{SS_{among}}{a-1} = \frac{n \sum_{j=1}^{a} (\bar{Y}_{i\cdot} - \bar{\bar{Y}})^2}{a-1}. \tag{11.9}$$

For the bark beetle experiment, we find that

$$MS_{among} = \frac{SS_{among}}{a-1} = \frac{1.1500}{3-1} = 0.5750. \tag{11.10}$$

So, what variance does MS_{among} estimate? If H_0 is true and there are no group effects, we would expect $\bar{Y}_{i\cdot}$ to have a variance of σ^2/n, because it is a sample mean composed of n observations in the ith group (which have a variance of σ^2). MS_{among} estimates this variance multiplied by n, because of the n term in numerator, and so actually estimates $n\sigma^2/n = \sigma^2$. On the other hand, if H_1 is true then there are group effects, and we would expect the group means to be shifted away from the grand mean. This should increase the size of MS_{among}. **Thus, MS_{among} estimates σ^2 if H_0 is true but becomes larger if H_1 is true.**

We next develop an estimate of the variance σ^2 that is free of any effects, fixed or random. This variance estimate is based on a quantity called the sum of squares within groups or SS_{within}, because it measures variation of the observations within each group. It is defined by the formula

$$\begin{aligned} SS_{within} &= \sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i\cdot})^2 && (11.11)\\ &= \sum_{j=1}^{n}(Y_{1j} - \bar{Y}_{1\cdot})^2 + \ldots + \sum_{j=1}^{n}(Y_{aj} - \bar{Y}_{a\cdot})^2. && (11.12)\end{aligned}$$

It has $a(n-1)$ degrees of freedom, because there are a sum of squares terms each with $n-1$ degrees of freedom. We can obtain an estimate of σ^2 by dividing this sum of squares by its degrees of freedom, to obtain the mean square within groups:

$$MS_{within} = \frac{SS_{within}}{a(n-1)} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i\cdot})^2}{a(n-1)}. \qquad (11.13)$$

This quantity estimates σ^2 because it simply averages estimates of σ^2 for each group. With some rearrangement, we can write MS_{within} as

$$\begin{aligned} MS_{within} &= \frac{\sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i\cdot})^2}{a(n-1)} && (11.14)\\ &= \frac{\sum(Y_{1j} - \bar{Y}_{1\cdot})^2 + \ldots + \sum(Y_{aj} - \bar{Y}_{a\cdot})^2}{a(n-1)} && (11.15)\\ &= \frac{\frac{\sum(Y_{1j} - \bar{Y}_{1\cdot})^2}{n-1} + \ldots + \frac{\sum(Y_{aj} - \bar{Y}_{a\cdot})^2}{n-1}}{a} && (11.16)\\ &= \frac{s_1^2 + \ldots + s_a^2}{a}. && (11.17)\end{aligned}$$

Each term in the numerator of this expression is the sample variance s^2 for each group, which is then averaged across all groups to yield an overall or **pooled** estimate of σ^2. The word 'pooled' in statistics often indicates a combined estimate of a variance. It can also be shown that $E[MS_{within}] = \sigma^2$, regardless of any group effects.

We now calculate MS_{within} for the bark beetle experiment. We first need to calculate the quantity $(Y_{ij} - \bar{Y}_{i\cdot})^2$ for the observations in each group and then sum these for each group (see Table 11.1). Summing these quantities in turn across all groups, we obtain

$$SS_{within} = 0.2110 + 0.1325 + 0.4581 = 0.8016. \tag{11.18}$$

$$\tag{11.19}$$

We then have

$$MS_{within} = \frac{SS_{within}}{a(n-1)} = \frac{0.8016}{3(5-1)} = 0.0668. \tag{11.20}$$

$$\tag{11.21}$$

There is one more sum of squares that can be calculated in one-way ANOVA, the total sum of squares. It is defined as

$$SS_{total} = \sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij} - \bar{\bar{Y}})^2. \tag{11.22}$$

It measures the variability of the observations around the grand mean of the data ($\bar{\bar{Y}}$) and has $an - 1$ degrees of freedom. Applying this formula to the Example 1 data set, we obtain $SS_{total} = 1.9516$ after much calculation.

An interesting feature of the sum of squares is that they add to the total sum of squares, as do the degrees of freedom. In particular, we have

$$SS_{among} + SS_{within} = SS_{total} \tag{11.23}$$

and

$$(a - 1) + a(n - 1) = an - 1. \tag{11.24}$$

Thus, the sum of squares and degrees of freedom can be neatly partitioned into components corresponding to among group and within group variation. We will illustrate this relationship further in the section below on ANOVA tables.

11.2.2 F statistic and distribution

We next describe the statistic used to test H_0 : all $\alpha_i = 0$ for the fixed effect model, and $H_0 : \sigma_A^2 = 0$ for the random effects one. It is simply the ratio of MS_{among} and MS_{within}, or

$$F_s = \frac{MS_{among}}{MS_{within}}. \tag{11.25}$$

If H_0 is true for either model, both MS_{among} and MS_{within} estimate σ^2 and we would expect their ratio, F_s, to be small and on the order of one. However, if H_0 is false and H_1 is true, we would expect MS_{among} to become larger and F_s to increase. We would therefore reject H_0 for large values of F_s.

To complete our testing procedure and find P values, we need to know the distribution of F_s under H_0. It turns out this statistic has an F distribution under H_0, whose shape and location is governed by two parameters, the degrees of freedom for MS_{among} and MS_{within}. These are called the numerator and denominator degrees of freedom, which we abbreviate as df_1 and df_2. In particular, for one-way ANOVA we have $df_1 = a - 1$ and $df_2 = a(n - 1)$. Figure 11.5 shows the F distribution for three different sets of parameter values. Note that distribution can have a maximum at $y = 0$ for small values of df_1, while larger values of df_2 decrease the probability in the right tail of the distribution.

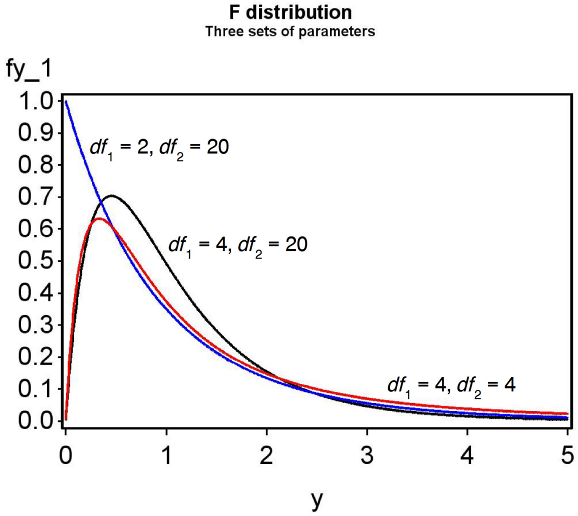


Figure 11.5: The F distribution for three different sets of parameter values

Table F gives the quantiles of the F distribution for different values of the degrees of freedom and the cumulative probability p. Statistical tests that make use of the F distribution are typically called F tests.

Calculating the test statistic F_s for the bark beetle experiment, we have

$$F_s = \frac{MS_{among}}{MS_{within}} = \frac{0.5750}{0.0668} = 8.6078, \tag{11.26}$$

with $df_1 = a - 1 = 3 - 1 = 2$ and $df_2 = a(n-1) = 3(5-1) = 12$.

As with previous tests, we seek acceptance and rejection regions for a particular value of α, the Type I error rate. In particular, we seek a quantity c_{α,df_1,df_2} such that

$$P[0 < F_s < c_{\alpha,df_1,df_2}] = 1 - \alpha. \qquad (11.27)$$

The region is of this form because the test is designed to reject H_0 for large values of F_s, and accept it for small ones. To find c_{α,df_1,df_2}, we look in Table F for the column corresponding to $1 - p = \alpha$, for the appropriate degrees of freedom. The acceptance region would therefore be $(0, c_{\alpha,df_1,df_2})$, and we would reject H_0 if F_s lies outside this region.

For $\alpha = 0.05$, $df_1 = 2$, and $df_2 = 12$, we see from Table F that $c_{0.05,2,12} = 3.885$. Our acceptance region is therefore $(0, 3.885)$, and we reject H_0 at the $\alpha = 0.05$ level if $F_s \geq 3.885$ (Fig. 11.6). We see this is the case because $F_s = 8.6078 > 3.885$. We can continue this process for increasingly smaller α and eventually find that for $\alpha = 0.005$ we can still reject H_0, but not for $\alpha = 0.001$. We therefore have $P < 0.005$ for this test, because $\alpha = 0.005$ is the smallest value of α for which we can reject H_0 (see Chapter 10). An F test in ANOVA would often be reported as follows: 'There was a highly significant difference among the different baits in the number of bark beetles trapped ($F_{2,12} = 8.6078, P < 0.005$).' Note that the degrees of freedom are given as subscripts.

11.2.3 ANOVA tables

We can organize the different sum of squares and mean squares into an ANOVA table. It lists the different sources of variation in the data (among, within, and total), their degrees of freedom, sums of squares and mean squares, and then the F statistic and its P value. Table 11.3 shows the general layout of such a table for one-way ANOVA designs, while Table 11.4 gives the results for the Example 1 analysis. Note the additive relationship for the degrees of freedom and sum of squares.

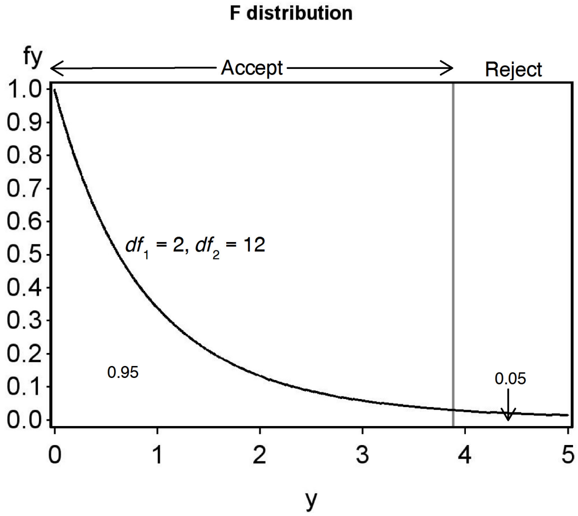


Figure 11.6: Acceptance and rejection regions for $\alpha = 0.05$

Table 11.3: General ANOVA table for one-way designs with a groups and n observations per group, showing formulas for different mean squares and the F test.

Source	*df*	Sum of squares	Mean square	F_s
Among	$a-1$	$SS_{among} = n\sum_{i=1}^{a}(\bar{Y}_{i\cdot} - \bar{Y})^2$	$MS_{among} = SS_{among}/(a-1)$	MS_{among}/MS_{within}
Within	$a(n-1)$	$SS_{within} = \sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij} - \bar{Y}_{i\cdot})^2$	$MS_{within} = SS_{within}/a(n-1)$	
Total	$an-1$	$SS_{total} = \sum_{i=1}^{a}\sum_{j=1}^{n}(Y_{ij} - \bar{Y})^2$		

Table 11.4: ANOVA table for the Example 1 data set, including a P value for the test.

Source	*df*	Sum of squares	Mean square	F_s	P
Among	2	1.1500	0.5750	8.6078	< 0.005
Within	12	0.8016	0.0668		
Total	14	1.9516			

11.2.4 One-way ANOVA for Example 1 - SAS demo

The same calculations for the bark beetle experiment can be carried out in SAS using `proc glm` (SAS Institute Inc. 2018). This procedure is primarily intended for fixed effects ANOVA models, with `proc mixed` the best choice for random effects models. However, the F test would be the same in either procedure.

We will also use SAS and `proc gplot` (SAS Institute Inc. 2016) to visualize the data. The basic idea is to plot, for each treatment group, the individual data points along with their mean ($\bar{Y}$) ± one standard error ($s/\sqrt{n}$). These plots are useful for comparing the relative effects of the treatments, a concept called **effect size**, as well as the variability of the observations. Effect size is used to judge the **biological significance** of the treatments – are the differences among the treatments biologically meaningful? This is distinct from the statistical significance of the ANOVA. For example, you could observe large differences among the treatment means that could be biological significant, but the F test could be non-significant because the data were highly variable. Conversely, the differences among the means could be small and not biologically meaningful, but the F test could be significant because n is large, and so the test can detect even small differences.

The SAS program for one-way ANOVA is a bit more complicated than previous programs, so we will examine it a section at a time. The first step is to read in the observations using a `data` step, with one variable denoting the treatment (`treat`) and a second the number of beetles captured (`count`). As discussed earlier, it is common to log-transform count data, and so we generate a variable `y` that is the `log10` (log base 10) of `count`. The `data` step is followed by a `print` statement to print the data set. See section below.

```
* bark_beetle_experiment.sas;
title "One-way ANOVA for bark beetle trapping experiment";
data bark_beetle;
    input treat $ count;
    * Apply transformations here;
    y = log10(count);
    datalines;
A   373
A   126
A   255

etc.
```

```
C   199
C    84
;
run;
* Print data set;
proc print data=bark_beetle;
run;
```

We next plot the data using `proc gplot` (SAS Institute Inc. 2016). The `plot` statement tells `gplot` to plot the variable `y` on the y-axis and `treat` on the x-axis of the plot. The appearance of the points is controlled by the `symbol1` statement, which among other things specifies that the points be plotted along with their means ± one standard error, with the means joined by a line, using the option `i=std1mjt`. Other options in the `symbol` statement control the type and size of the points, and line width. The `vaxis=axis1` and `haxis=axis1` options control the visual appearance of the x- and y-axes. See below.

```
* Plot means, standard errors, and observations;
proc gplot data=bark_beetle;
    plot y*treat=1 / vaxis=axis1 haxis=axis1;
    symbol1 i=std1mjt v=star height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
```

The next section of the program conducts the one-way ANOVA and F test using `proc glm`. The `plots=diagnostics` option generates graphs that are used to examine some of the assumptions of ANOVA – we will defer their discussion to Chapter 15. The `class` statement tells SAS that the variable `treat` is the one that defines different groups in the ANOVA (see listing below). The `model` statement basically tells SAS the form of the ANOVA model. Recall that the model for fixed effects one-way ANOVA is given by the equation

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}. \tag{11.28}$$

If we equate Y_{ij} with `y`, and α_i with `treat`, we see there are similarities between the fixed effects model and the SAS `model` statement. In fact, SAS assumes you want a grand mean μ unless otherwise specified, as well as the error term ϵ_{ij}. As we examine more complex ANOVA models in later chapters, we will see there is nearly a one-to-one correspondence between these models and the corresponding SAS `model` statement.

```
* One-way ANOVA with all fixed effects;
proc glm plots=diagnostics data=bark_beetle;
    class treat;
    model y = treat;
    * Calculate means for each group;
    means treat;
run;
```

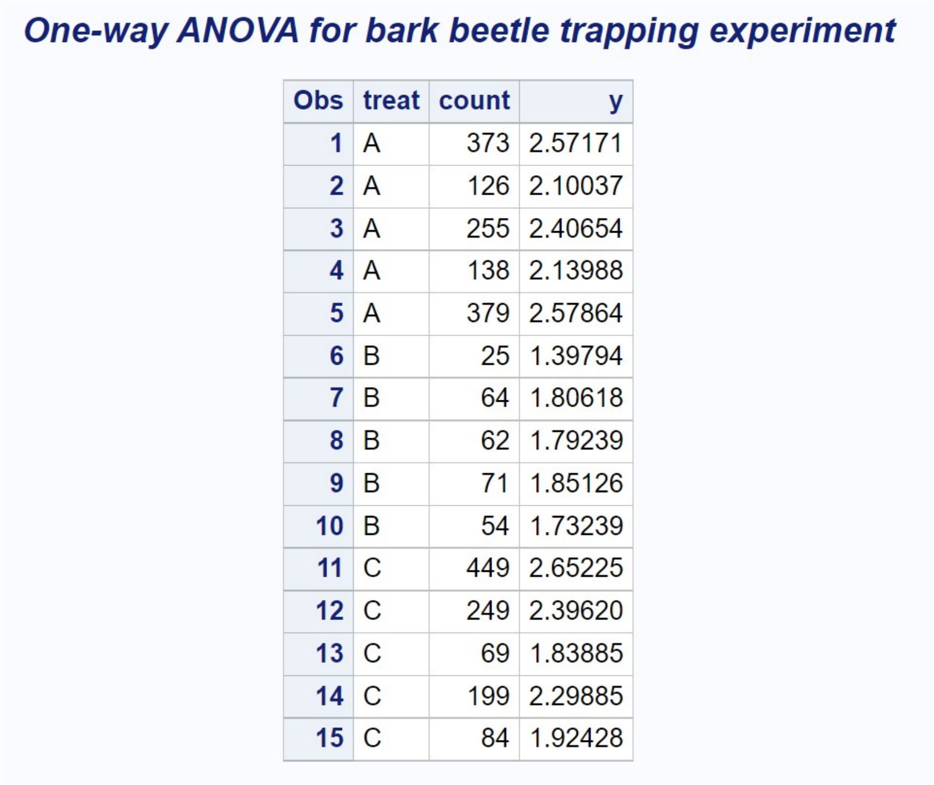
The `means` statement causes `glm` to calculate means for each `treat` group.

The complete SAS program and output are listed below. The output shows the same F test in three different locations within the `proc glm` output (Fig. 11.9). The first is in a format resembling an ANOVA table, and then two other times corresponding to Type I and III sums of squares. These are different ways of calculating the sums of squares and tests, with Type III sums of squares more generally useful for ANOVA designs. For one-way ANOVA the results are the same, and we see that there was a highly significant difference among groups ($F_{2,12} = 8.60, P = 0.0048$). Inspection of the graph (Fig. 11.8) and means suggests that treatment A caught the most beetles, followed by C and then B.

SAS Program

```
* bark_beetle_experiment.sas;
title "One-way ANOVA for bark beetle trapping experiment";
data bark_beetle;
    input treat $ count;
    * Apply transformations here;
    y = log10(count);
    datalines;
A   373
A   126
A   255
A   138
A   379
B    25
B    64
B    62
B    71
B    54
C   449
C   249
C    69
C   199
C    84
```

```
;
run;
* Print data set;
proc print data=bark_beetle;
run;
* Plot means, standard errors, and observations;
proc gplot data=bark_beetle;
    plot y*treat=1 / vaxis=axis1 haxis=axis1;
    symbol1 i=std1mjt v=star height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* One-way ANOVA with all fixed effects;
proc glm plots=diagnostics data=bark_beetle;
    class treat;
    model y = treat;
    * Calculate means for each group;
    means treat;
run;
quit;
```



One-way ANOVA for bark beetle trapping experiment

Obs	treat	count	y
1	A	373	2.57171
2	A	126	2.10037
3	A	255	2.40654
4	A	138	2.13988
5	A	379	2.57864
6	B	25	1.39794
7	B	64	1.80618
8	B	62	1.79239
9	B	71	1.85126
10	B	54	1.73239
11	C	449	2.65225
12	C	249	2.39620
13	C	69	1.83885
14	C	199	2.29885
15	C	84	1.92428

Figure 11.7: `bark_beetle_experiment.sas - proc print`

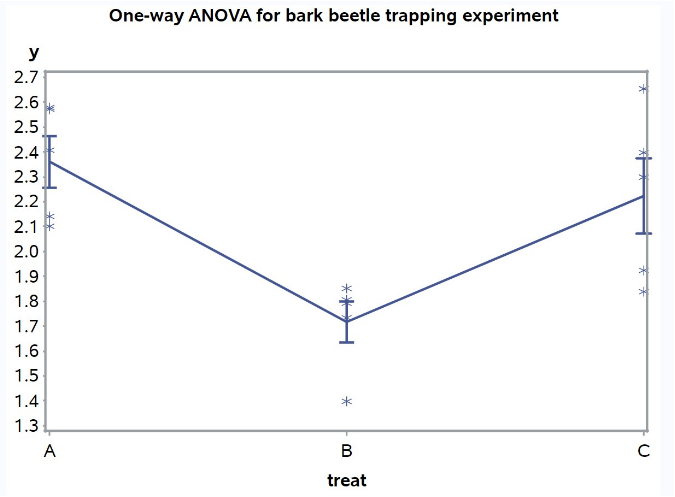
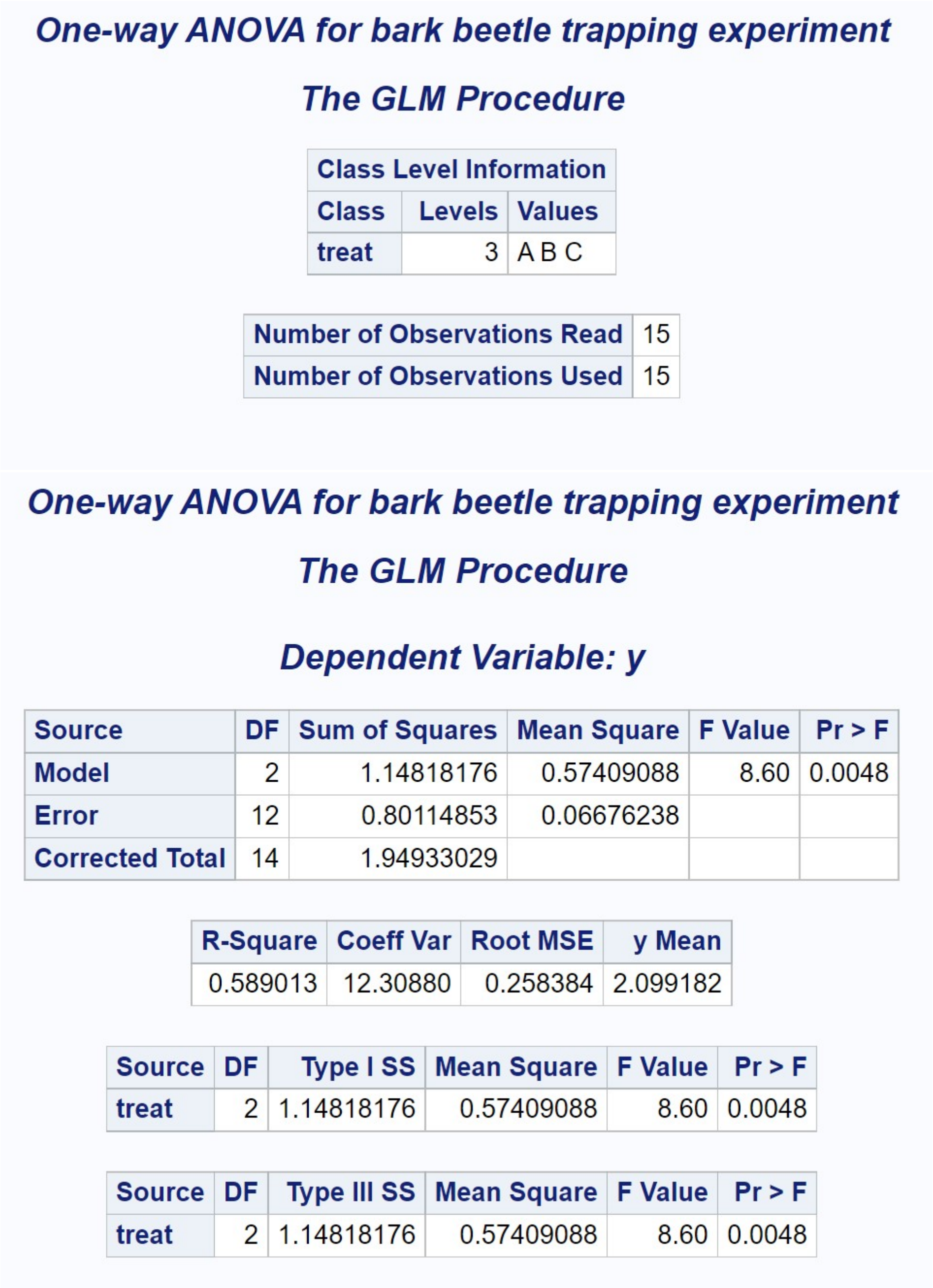


Figure 11.8: `bark_beetle_experiment.sas - proc gplot`



One-way ANOVA for bark beetle trapping experiment

The GLM Procedure

Class Level Information		
Class	Levels	Values
treat	3	A B C

Number of Observations Read	15
Number of Observations Used	15

One-way ANOVA for bark beetle trapping experiment

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	1.14818176	0.57409088	8.60	0.0048
Error	12	0.80114853	0.06676238		
Corrected Total	14	1.94933029			

R-Square	Coeff Var	Root MSE	y Mean
0.589013	12.30880	0.258384	2.099182

Source	DF	Type I SS	Mean Square	F Value	Pr > F
treat	2	1.14818176	0.57409088	8.60	0.0048

Source	DF	Type III SS	Mean Square	F Value	Pr > F
treat	2	1.14818176	0.57409088	8.60	0.0048

Figure 11.9: `bark_beetle_experiment.sas - proc glm`

11.2.5 One-way ANOVA for Example 2 - sample calculation

We will conduct an F test for our second data set, involving a study of bark beetles trapped at five different sites ($a = 5$) selected at random from a collection of sites, with five traps per site ($n = 5$). This implies a random effects model, and we are therefore interested in testing $H_0 : \sigma_A^2 = 0$ vs. $H_1 : \sigma_A^2 > 0$. Some preliminary calculations for the F test are shown in Table 11.2. We first find the mean $\bar{Y}_{i\cdot}$ for each site, then calculate the grand mean as the average of the site means:

$$\begin{aligned}\bar{\bar{Y}} &= \frac{\sum_{i=1}^{a} \bar{Y}_{i\cdot}}{a} &(11.29)\\ &= \frac{2.0120 + 2.4700 + 2.2460 + 2.6960 + 1.1940}{5} &(11.30)\\ &= \frac{10.6180}{5} = 2.1236. &(11.31)\end{aligned}$$

We then have

$$\begin{aligned}SS_{among} &= n \sum_{j=1}^{a} (\bar{Y}_{i\cdot} - \bar{\bar{Y}})^2 &(11.32)\\ &= 5\left[(2.0120 - 2.1236)^2 + \ldots + (1.1940 - 2.1236)^2\right] &(11.33)\\ &= 5\left[0.0125 + 0.1200 + 0.0150 + 0.3276 + 0.8642\right] &(11.34)\\ &= 6.6965 &(11.35)\end{aligned}$$

We next calculate MS_{among}:

$$MS_{among} = \frac{SS_{among}}{a-1} = \frac{6.6965}{5-1} = 1.6741. \qquad (11.36)$$

$$(11.37)$$

Now we find SS_{within}, first calculating $(Y_{ij} - \bar{Y}_{i\cdot})^2$ for the observations in each group and then summing these for each group (see Table 11.2). Summing these quantities in turn across all groups, we obtain

$$SS_{within} = 0.1598 + 0.4730 + 0.3419 + 0.7600 + 0.0459 = 1.7806. \qquad (11.38)$$

$$(11.39)$$

We then have

$$MS_{within} = \frac{SS_{within}}{a(n-1)} = \frac{1.7806}{5(5-1)} = 0.0890. \tag{11.40}$$

$$\tag{11.41}$$

Calculating the test statistic F_s, we obtain

$$F_s = \frac{MS_{among}}{MS_{within}} = \frac{1.6741}{0.0890} = 18.8101, \tag{11.42}$$

$$\tag{11.43}$$

with $df_1 = a - 1 = 4 - 1 = 4$ and $df_2 = a(n-1) = 5(5-1) = 20$. From Table F, we find $P < 0.001$. The variance among sites was highly significant $(F_{4,12} = 18.8101, P < 0.001$.

11.2.6 One-way ANOVA for Example 2 - SAS demo

We can carry out the F test as well as estimate the variance components (σ_A^2 and σ^2) for the random effects model using SAS. The first section of the program involving the `data` step and `gplot` graph is similar to the fixed effects program. The next section of the program fits the random effects model to the data and conducts the F test, using `proc mixed` (see listing below). As before, the `class` statement tells SAS that the variable `site` is the one that defines different groups in the ANOVA. Now recall that the model for random effects one-way ANOVA is given by the equation

$$Y_{ij} = \mu + A_i + \epsilon_{ij}. \tag{11.44}$$

Note that A_i corresponds to `site` in the bark beetle study. In `proc mixed`, fixed effects in the model are placed in a `model` statement, while any random effects are listed in a `random` statement (SAS Institute Inc. 2018). Because our random effects model only has one random effect, `site`, this is listed in the `random` statement. There are no fixed effects in this model, so the `model` statement lists nothing after the equals sign. The option `ddfm=kr` specifies a general method of calculating the degrees of freedom that works well under many circumstances, including more complicated models.

```
* One-way ANOVA with random effects - F test;
proc mixed method=type3 data=bark_beetle;
    class site;
    model y = / ddfm=kr;
    random site;
run;
* One-way ANOVA with random effects - variance components;
proc mixed cl plots=residualpanel data=bark_beetle;
    class site;
    model y = / ddfm=kr;
    random site;
run;
```

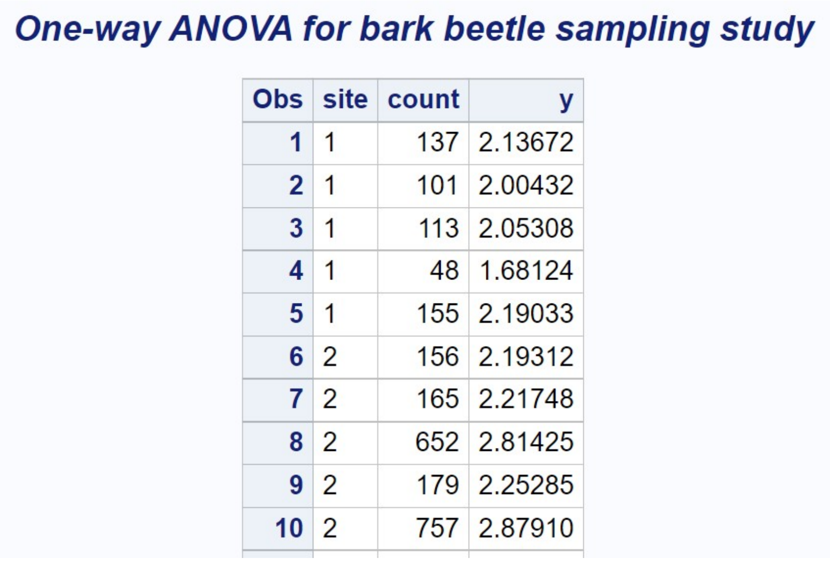
Why is `proc mixed` invoked twice in this program? The first one generates the F statistic for testing $H_0 : \sigma_A^2 = 0$ vs. $H_1 : \sigma_A^2 > 0$, using the option `method=type3`. This is not the default in `proc mixed`, which appears more designed to estimate the variance components in random effects (Littell et al. 1996). If we drop this option, as in the second `proc mixed` statement, we get only these estimates and no F test. Confidence intervals for the variance components are requested using the `cl` option. The variance components estimated in the second `proc mixed` using a version of maximum likelihood, the preferred method of estimating these quantities.

The complete SAS program and output are listed below. The F test found in the first `proc mixed` output (Fig. 11.12) was highly significant ($F_{4,12} = 18.77, P < 0.0001$), suggesting $\sigma^2_A > 0$. The second `proc mixed` output provides estimates and confidence intervals for the two variance components (Fig. 11.13). We have $\hat{\sigma}^2_A = 0.3174$ for which the 95% confidence interval is $(0.1093, 3.1458)$, and $\hat{\sigma}^2 = 0.0893$ with confidence interval $(0.0523, 0.1863)$. From these results, we see that the variance among sites was greater than the variance within sites ($0.3174 > 0.0893$). This can also be seen in Fig. 11.11 – beetle numbers vary considerably among sites relative to within them.

SAS Program

```
* bark_beetle_random.sas;
title "One-way ANOVA for bark beetle sampling study";
data bark_beetle;
    input site $ count;
    * Apply transformations here;
    y = log10(count);
    datalines;
1   137
1   101
1   113
1    48
1   155
2   156
2   165
2   652
2   179
2   757
3   278
3   197
3    95
3   395
3    83
4  2540
4   613
4   200
4   251
```

```
4   390
5    18
5    16
5    11
5    21
5    14
;
run;
* Print data set;
proc print data=bark_beetle;
run;
* Plot means, standard errors, and observations;
proc gplot data=bark_beetle;
    plot y*site=1 / vaxis=axis1 haxis=axis1;
    symbol1 i=std1mjt v=star height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* One-way ANOVA with random effects - F test;
proc mixed method=type3 data=bark_beetle;
    class site;
    model y = / ddfm=kr;
    random site;
run;
* One-way ANOVA with random effects - variance components;
proc mixed cl plots=residualpanel data=bark_beetle;
    class site;
    model y = / ddfm=kr;
    random site;
run;
quit;
```



One-way ANOVA for bark beetle sampling study

Obs	site	count	y
1	1	137	2.13672
2	1	101	2.00432
3	1	113	2.05308
4	1	48	1.68124
5	1	155	2.19033
6	2	156	2.19312
7	2	165	2.21748
8	2	652	2.81425
9	2	179	2.25285
10	2	757	2.87910

etc.

Figure 11.10: `bark_beetle_random.sas - proc print`

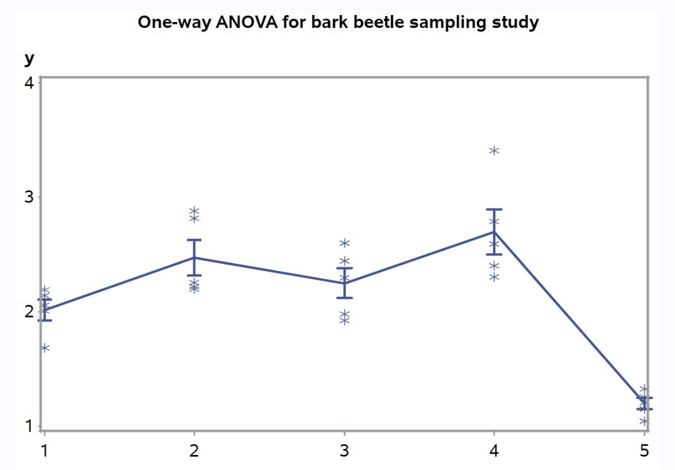
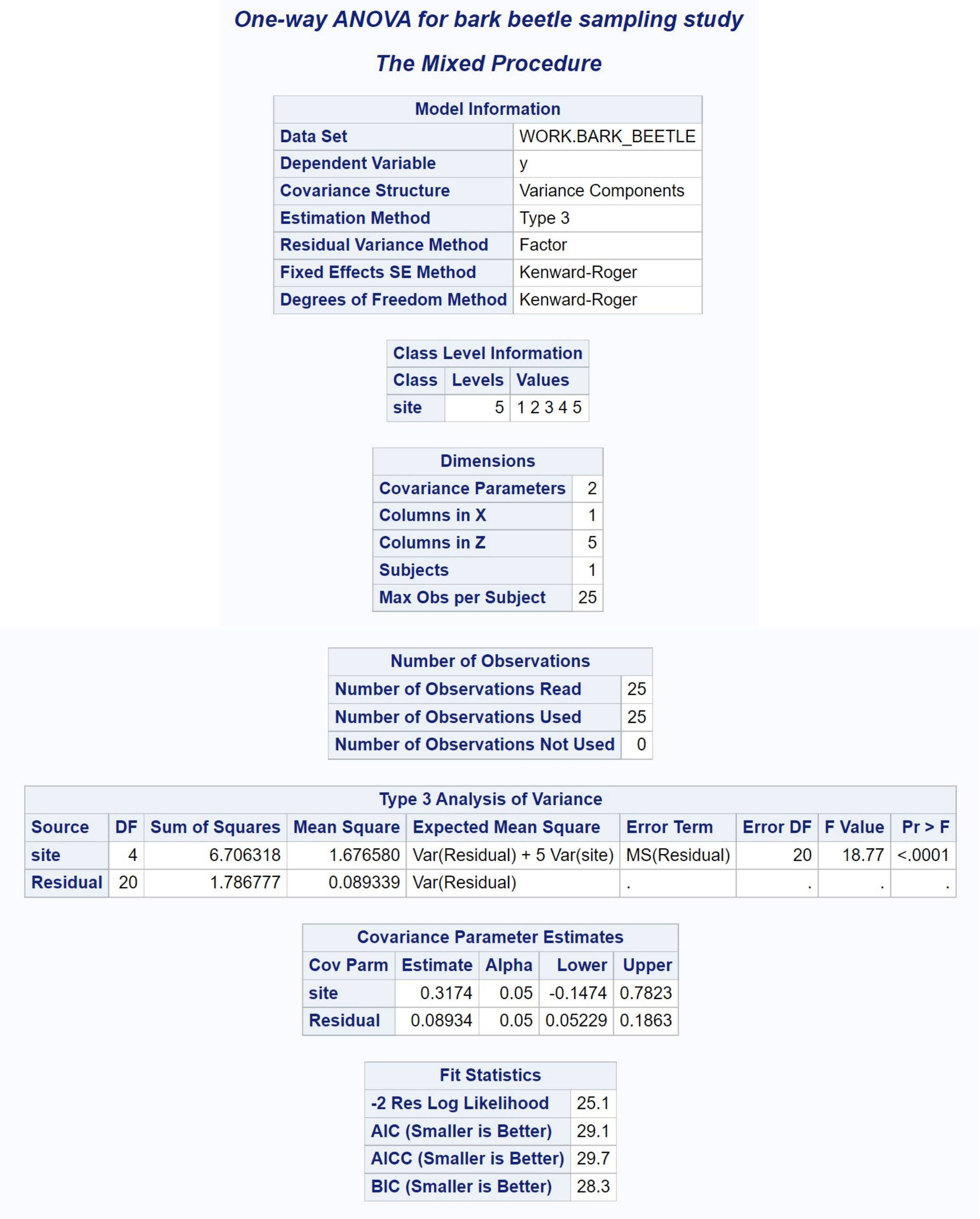


Figure 11.11: `bark_beetle_random.sas - proc gplot`



One-way ANOVA for bark beetle sampling study

The Mixed Procedure

Model Information	
Data Set	WORK.BARK_BEETLE
Dependent Variable	y
Covariance Structure	Variance Components
Estimation Method	Type 3
Residual Variance Method	Factor
Fixed Effects SE Method	Kenward-Roger
Degrees of Freedom Method	Kenward-Roger

Class Level Information		
Class	Levels	Values
site	5	1 2 3 4 5

Dimensions	
Covariance Parameters	2
Columns in X	1
Columns in Z	5
Subjects	1
Max Obs per Subject	25

Number of Observations	
Number of Observations Read	25
Number of Observations Used	25
Number of Observations Not Used	0

Type 3 Analysis of Variance

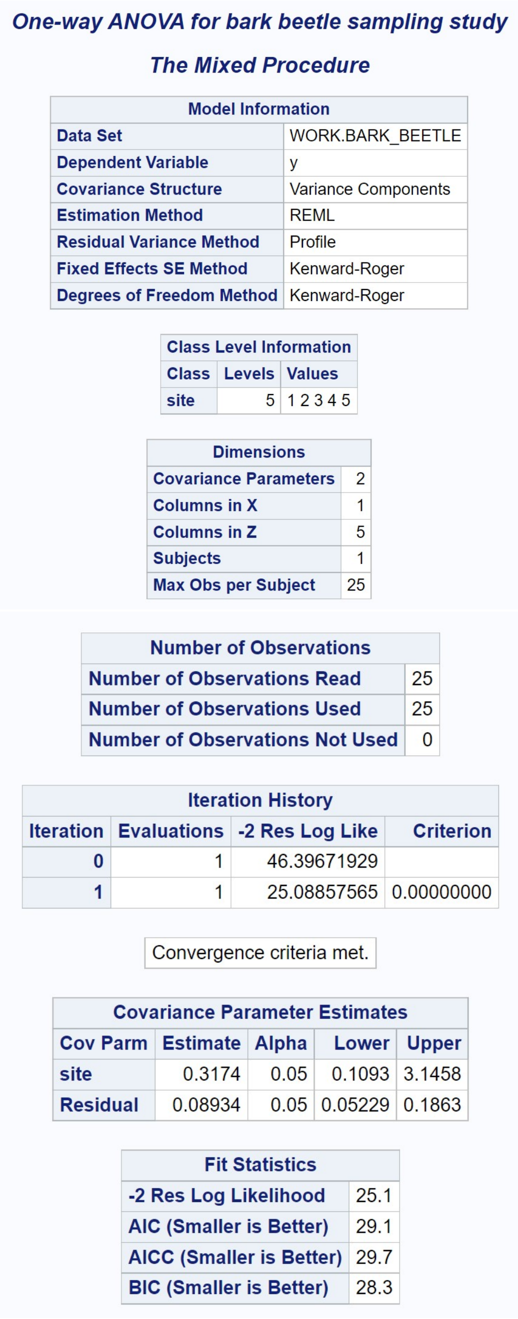
Source	DF	Sum of Squares	Mean Square	Expected Mean Square	Error Term	Error DF	F Value	Pr > F
site	4	6.706318	1.676580	Var(Residual) + 5 Var(site)	MS(Residual)	20	18.77	<.0001
Residual	20	1.786777	0.089339	Var(Residual)	.	.	.	.

Covariance Parameter Estimates

Cov Parm	Estimate	Alpha	Lower	Upper
site	0.3174	0.05	-0.1474	0.7823
Residual	0.08934	0.05	0.05229	0.1863

Fit Statistics	
-2 Res Log Likelihood	25.1
AIC (Smaller is Better)	29.1
AICC (Smaller is Better)	29.7
BIC (Smaller is Better)	28.3

Figure 11.12: `bark_beetle_random.sas - proc mixed` (1)



One-way ANOVA for bark beetle sampling study

The Mixed Procedure

Model Information	
Data Set	WORK.BARK_BEETLE
Dependent Variable	y
Covariance Structure	Variance Components
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Kenward-Roger
Degrees of Freedom Method	Kenward-Roger

Class Level Information		
Class	Levels	Values
site	5	1 2 3 4 5

Dimensions	
Covariance Parameters	2
Columns in X	1
Columns in Z	5
Subjects	1
Max Obs per Subject	25

Number of Observations	
Number of Observations Read	25
Number of Observations Used	25
Number of Observations Not Used	0

Iteration History			
Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	46.39671929	
1	1	25.08857565	0.00000000

Convergence criteria met.

Covariance Parameter Estimates				
Cov Parm	Estimate	Alpha	Lower	Upper
site	0.3174	0.05	0.1093	3.1458
Residual	0.08934	0.05	0.05229	0.1863

Fit Statistics	
-2 Res Log Likelihood	25.1
AIC (Smaller is Better)	29.1
AICC (Smaller is Better)	29.7
BIC (Smaller is Better)	28.3

Figure 11.13: `bark_beetle_random.sas - proc mixed` (2)

11.3 Maximum likelihood estimates

This section sketches how the parameters in one-way ANOVA can be estimated using maximum likelihood. Recall that the likelihood for a random sample of three observations ($Y_1 = 4.5, Y_2 = 5.4, Y_2 = 5.3$) from a normal distribution (see Chapter 8) was of the form

$$L(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(4.5-\mu)^2}{\sigma^2}} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(5.4-\mu)^2}{\sigma^2}} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(5.3-\mu)^2}{\sigma^2}}. \tag{11.45}$$

We found maximum likelihood estimates of the normal distribution parameters by maximizing this quantity with respect to μ and σ^2.

Suppose now we have a data set that can be modeled using the fixed effects one-way ANOVA model, in particular

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}. \tag{11.46}$$

This model has a number of parameters to estimate, such as μ, α_i for $i = 1, 2, \ldots, a$, and σ^2. What would the likelihood function look like for these data? Consider the first group for the bark beetle experiment (Example 1), for which we have $Y_{11} = 2.576$, $Y_{12} = 2.10$, $Y_{13} = 2.41$, $Y_{14} = 2.14$, and $Y_{15} = 2.58$. For the first group the model assumes that $Y_{1j} \sim N(\mu + \alpha_1, \sigma^2)$, and so the likelihood would be

$$L_1 = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(2.57-(\mu+\alpha_1))^2}{\sigma^2}} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(2.10-(\mu+\alpha_1))^2}{\sigma^2}} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(2.41-(\mu+\alpha_1))^2}{\sigma^2}} \tag{11.47}$$

$$\times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(2.14-(\mu+\alpha_1))^2}{\sigma^2}} \times \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(2.58-(\mu+\alpha_1))^2}{\sigma^2}}. \tag{11.48}$$

The likelihood L_2 for the second group would be similar, except that $Y_{2j} \sim N(\mu + \alpha_2, \sigma^2)$, and L_3 similarly defined. The overall likelihood would then be defined as

$$L(\mu, \alpha_1, \alpha_2, \alpha_3, \sigma^2) = L_1 \times L_2 \times L_3. \tag{11.49}$$

Finding the maximum likelihood estimates involves maximizing this quantity with respect to the parameters $\mu, \alpha_1, \alpha_2, \alpha_3$, and σ^2. The likelihood for designs with any number of treatment groups and replicates would be similar. Using a bit of calculus to find the maximum, it can be shown that the maximum likelihood estimates of these parameters, in general, are

$$\hat{\mu} = \bar{\bar{Y}}, \tag{11.50}$$

$$\hat{\alpha}_i = \bar{Y}_{i\cdot} - \bar{\bar{Y}}, \tag{11.51}$$

and

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} (Y_{ij} - \bar{Y}_{i\cdot})^2}{a(n-1)} = MS_{within}. \tag{11.52}$$

(McCulloch & Searle 2001). These estimators seem quite reasonable. They use the grand mean of the data, $\bar{\bar{Y}}$, to estimate the grand mean μ of the model, and the difference between the *ith* group mean and the grand mean, $\bar{Y}_{i\cdot} - \bar{\bar{Y}}$, to estimate the deviation from the group mean α_i. Note that $\hat{\sigma}^2$ is equal to MS_{within}, which we have already encountered in our ANOVA calculations.

Suppose now we have a data set suited to the random effects model, in particular

$$Y_{ij} = \mu + A_i + \epsilon_{ij}. \tag{11.53}$$

This model has three parameters to be estimated: μ, σ_A^2, and σ^2. The likelihood for this model is more complex because of the random effect A_i, but one can show that the maximum likelihood estimators of these parameters are

$$\hat{\mu} = \bar{\bar{Y}}, \tag{11.54}$$

$$\hat{\sigma}_A^2 = \frac{MS_{among} - MS_{within}}{n}, \tag{11.55}$$

and

$$\hat{\sigma}^2 = MS_{within}. \tag{11.56}$$

An intuitive explanation of the formula for $\hat{\sigma}_A^2$ is that MS_{among} incorporates variance from both A_i and ϵ_{ij}, while MS_{within} only has ϵ_{ij}. Subtracting MS_{within} from MS_{among} leaves only the variance due to A_i, so that the numerator of this expression estimates $n\sigma_A^2$. We then divide by n to obtain an estimate of σ_A^2.

Suppose that for an unusual data set we obtain $MS_{among} < MS_{within}$, implying a negative estimate of $\hat{\sigma}_A^2 = 0$ according to the above equation. An inherent feature of maximum likelihood is that is restricts variance components to plausible values (McCulloch & Searle 2001), so in this case it would simply say that $\hat{\sigma}_A^2 = 0$, the smallest possible nonnegative value. This would be reflected in the SAS output for `proc mixed`, which would report that the variance component in question was zero. The estimates presented here are actually obtained using a variant of maximum likelihood called restricted maximum likelihood or REML. This method is the default in SAS, and has some theoretical advantages over straight maximum likelihood (McCulloch and Searle 2001).

11.4 F test as a likelihood ratio test

The F test in one-way ANOVA can be derived as a likelihood ratio test, similar to the development of the t test in Chapter 10. We first find the maximum likelihood estimates of various parameters under H_1 vs. H_0, where the parameters under consideration are the ANOVA model parameters. Recall that the observations in the fixed effects model are described as

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \tag{11.57}$$

where μ is the grand mean, α_i is the effect of the ith treatment, and $\epsilon_{ij} \sim N(0, \sigma^2)$. This is the statistical model under the alternative hypothesis, where $\alpha_i \neq 0$ for some i. Under H_0 : all $\alpha_i = 0$, the model reduces to just

$$Y_{ij} = \mu + \epsilon_{ij}. \tag{11.58}$$

We would need to find the maximum likelihood estimates under both H_1 (see previous section) and H_0, as well as L_{H_0} and L_{H_1}, the maximum height of the likelihood function under H_0 and H_1. We would then use the likelihood ratio test statistic

$$\lambda = \frac{L_{H_0}}{L_{H_1}}. \tag{11.59}$$

It can be shown that there is a one-to-one correspondence between $-2\ln(\lambda)$ and F_s in one-way ANOVA, and so the F test is actually a likelihood ratio test (McCulloch & Searle 2001). A similar argument can be made to justify the F test for the random effects model. Like all likelihood ratio tests, large values of the test statistic $-2\ln(\lambda)$ or F_s indicate a lower value of the likelihood under H_0 relative to H_1, and thus a poorer fit of the H_0 model.

11.5 One-way ANOVA and two-sample t tests

There is an alternative to one-way ANOVA when there are only two groups to be compared, the two-sample t test. Let μ_1 be the mean of the first group and μ_2 the second one, and suppose that the two groups have the same variance σ^2 and sample size n. We are interested in testing $H_0 : \mu_1 = \mu_2$ vs. $H_1 : \mu_1 \neq \mu_2$, to determine if there are differences in the means of the two groups. Under H_0, the test statistic

$$T_s = \frac{(\bar{Y}_{1\cdot} - \bar{Y}_{2\cdot})}{\sqrt{\frac{s_1^2+s_2^2}{2}}} \sim t_{2(n-1)}. \tag{11.60}$$

Here $\bar{Y}_{1\cdot}$ and $\bar{Y}_{2\cdot}$ are the sample means for each group, and s_1^2 and s_2^2 the sample variances. For a Type I error rate of α, the acceptance region of the test would be the interval $(-c_{\alpha,2(n-1)}, c_{\alpha,2(n-1)})$, where $c_{\alpha,2(n-1)}$ is determined using Table T (see Chapter 10). We would reject H_0 if it falls on the edge or outside this interval. There are also versions of this test statistic for unequal sample sizes.

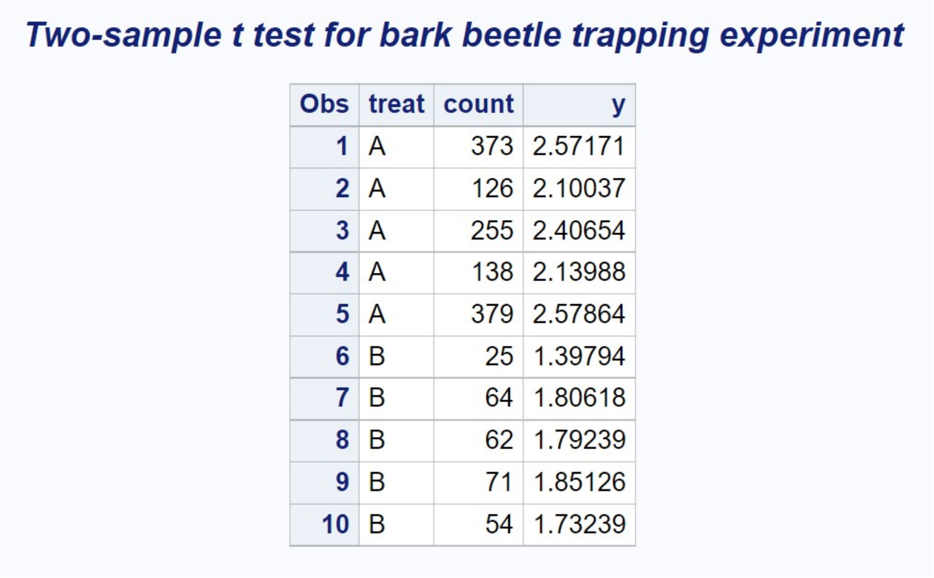
Although a two-sample t test is often used for comparing two groups, in the form above it is equivalent to the F test in one-way ANOVA. To see this, note that $T_s^2 = F_s$ for one-way ANOVA with two groups. It can also be shown that the acceptance and rejection regions are the same for the two tests. Unlike ANOVA, though, a two-sample t test can also be used for one-tailed alternative hypotheses, such as $H_1 : \mu_1 > \mu_2$ or $H_1 : \mu_1 < \mu_2$. The procedure is similar to one-sample t tests for one-tailed alternatives (see Chapter 10).

11.5.1 Two-sample t test for Example 1 - SAS demo

We can illustrate this test by comparing treatment A and B in the Example 1 study, deleting the data for the third treatment. See SAS program and output below. The `data` and `proc gplot` portions of the program are similar to our previous one-way ANOVA code. The two-sample t test is carried out using `proc ttest` (SAS Institute Inc. 2018), with the `class` statement indicating the variable that codes for different groups (`treat`), while the `var` statement designates the dependent variable (`y`). From Fig. 11.16, we see there was a highly significant difference between treatment A and B ($t_8 = 4.90, P = 0.0012$), with treatment A catching more beetles than B (Fig. 11.15).

SAS Program

```
* bark_beetle_experiment_ttest.sas;
title "Two-sample t test for bark beetle trapping experiment";
data bark_beetle;
    input treat $ count;
    * Apply transformations here;
    y = log10(count);
    datalines;
A   373
A   126
A   255
A   138
A   379
B    25
B    64
B    62
B    71
B    54
;
run;
* Print data set;
proc print data=bark_beetle;
run;
* Plot means, standard errors, and observations;
proc gplot data=bark_beetle;
    plot y*treat=1 / vaxis=axis1 haxis=axis1;
    symbol1 i=std1mjt v=star height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* Two-sample t test;
proc ttest data=bark_beetle;
    class treat;
    var y;
run;
quit;
```



Two-sample t test for bark beetle trapping experiment

Obs	treat	count	y
1	A	373	2.57171
2	A	126	2.10037
3	A	255	2.40654
4	A	138	2.13988
5	A	379	2.57864
6	B	25	1.39794
7	B	64	1.80618
8	B	62	1.79239
9	B	71	1.85126
10	B	54	1.73239

Figure 11.14: `bark_beetle_experiment_ttest.sas - proc print`

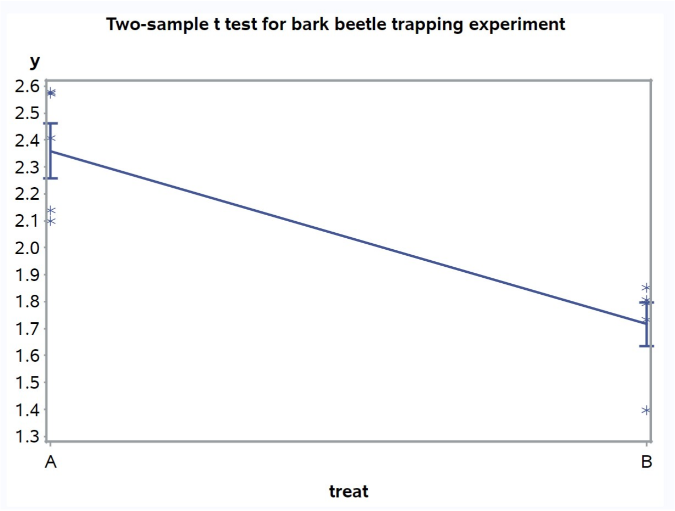
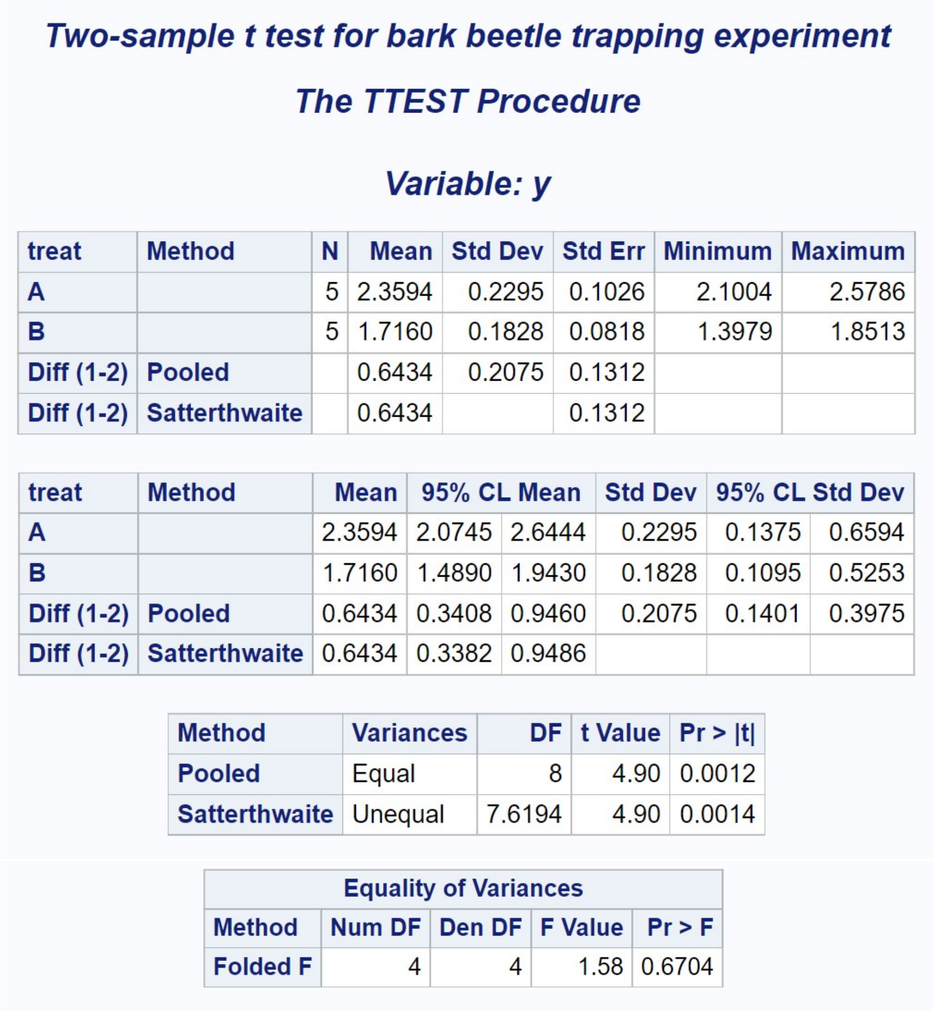


Figure 11.15: `bark_beetle_experiment_ttest.sas - proc gplot`



Two-sample t test for bark beetle trapping experiment

The TTEST Procedure

Variable: y

treat	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
A		5	2.3594	0.2295	0.1026	2.1004	2.5786
B		5	1.7160	0.1828	0.0818	1.3979	1.8513
Diff (1-2)	Pooled		0.6434	0.2075	0.1312		
Diff (1-2)	Satterthwaite		0.6434		0.1312		

treat	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
A		2.3594	2.0745	2.6444	0.2295	0.1375	0.6594
B		1.7160	1.4890	1.9430	0.1828	0.1095	0.5253
Diff (1-2)	Pooled	0.6434	0.3408	0.9460	0.2075	0.1401	0.3975
Diff (1-2)	Satterthwaite	0.6434	0.3382	0.9486			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	8	4.90	0.0012
Satterthwaite	Unequal	7.6194	4.90	0.0014

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	4	4	1.58	0.6704

Figure 11.16: `bark_beetle_experiment_ttest.sas - proc ttest`

11.6 References

McCulloch, C. E. & Searle, S. R. (2001) *Generalized, Linear, and Mixed Models.* John Wiley & Sons, Inc., New York, NY.

Littell, R. C., Milliken, G. A., Stroup, W. W. & Wolfinger, R. D. (1996) *The SAS System for Mixed Models.* SAS Institute Inc., Cary, NC.

SAS Institute Inc. (2016) *SAS/GRAPH 9.4: Reference, Fifth Edition.* SAS Institute Inc., Cary, NC.

SAS Institute Inc. (2018) *SAS/STAT 15.1 Users Guide* SAS Institute Inc., Cary, NC

Winer, B. J., Brown, D. R. & Michels, K. M. (1991) *Statistical Principles in Experimental Design*, 3rd edition. McGraw-Hill, Inc., Boston, MA.

11.7 Problems

1. A doctor conducts an experiment in which men are placed on four different diets, consisting of a standard weight loss regimen (a control treatment) and three new diets (Diets 1, 2, 3). The weight losses (lbs) after six months are given in the following table.

Control	Diet 1	Diet 2	Diet 3
19.5	20.0	20.8	25.9
20.5	16.4	17.4	25.9
16.6	11.9	16.7	25.8
19.3	22.1	16.8	22.5

   (a) Test whether there is a significant difference among the four treatments using one-way ANOVA, using manual calculations. Report the P value and discuss the significance of the test, and then interpret the results of the experiment. Show all your calculations.

   (b) Repeat the analysis using SAS and `proc glm`. Attach your program and output.

2. An experiment was conducted on the fecundity of a predatory insect reared on an artificial diet using four different concentrations of the preservative sorbic acid: (1) no sorbic acid, (2) 0.1% sorbic acid, (3) 0.2% sorbic acid, and (4) 0.5% sorbic acid. Twenty insects were reared at each concentration and the fecundity of the resulting adults measured. See table below.

Treatment	Observations
No sorbic acid	87, 124, 105, 87, 100, 89, 95, 79, 102, 112 92, 87, 115, 96, 111, 90, 86, 92, 109, 76
0.1% sorbic acid	105, 94, 97, 94, 83, 97, 107, 99, 104, 83 101, 71, 100, 75, 87, 106, 88, 99, 90, 74
0.2% sorbic acid	73, 94, 81, 83, 100, 98, 76, 91, 68, 82 92, 105, 76, 82, 95, 96, 101, 89, 92, 67
0.5% sorbic acide	83, 54, 86, 76, 74, 81, 79, 72, 80, 78 70, 83, 83, 85, 90, 70, 85, 94, 82, 75

   Test whether there is a difference among the four treatments using one-way ANOVA and SAS. Interpret the results of this analysis, providing a P value and discussing the significance of the test. Using a graph, explain what happens to fecundity as the concentration of sorbic acid changes.

Chapter 12

Power Analysis for One-Way ANOVA

Recall that the power of a statistical test is the probability of rejecting H_0 when H_0 is false, and some alternative hypothesis H_1 is true. We saw earlier (Chapter 10) that power for one-sample Z and t tests is a function of the quantity

$$\phi = \frac{(\mu_1 - \mu_0)}{\sigma/\sqrt{n}}, \tag{12.1}$$

where μ_1 and μ_0 are the means under H_1 and H_0, σ is the standard deviation of the observations, and n is the sample size. Anything that increases ϕ increases the power of the test, including greater differences between μ_1 and μ_0, decreasing σ, or increasing the sample size n. Larger values of the Type I error rate α also increase the power of the test, because they make it more likely the test will reject H_0 under any circumstances. Although one-way ANOVA is a more complicated design, we will see that exactly the same factors influence the power of its associated F test.

A power analysis for a one-way ANOVA design is usually conducted before running the experiment or study. This is known as a **prospective power analysis**. We then use the information from this analysis to refine our experimental design, most often the sample sizes needed for each treatment group to yield adequate power. Conversely, a **retrospective power analysis** is one conducted after an experiment or study, using the results from the study in the power calculation. This is a controversial procedure that some statisticians find questionable (Steidl et al. 1997).

Cohen (1988) recommends using a default power value of 0.8 when designing an experiment, if there is no other basis for setting the power. One reason is that achieving higher power values usually requires disproportionately larger sample sizes. He also recommends a power value of 0.8 on the basis of the ratio of Type II (β) to Type I error (α). He suggests that an optimal ratio of β/α is about four, implying that Type I errors are four times more serious than Type II errors. If you use $\alpha = 0.05$ as the Type I error rate, and choose power $= 0.8$, then $\beta = 1 -$ power $= 0.2$, and so $\beta/\alpha = 4$.

12.1 Power analysis for one-way ANOVA

Suppose we want to design an experiment involving several treatments that has adequate power. Assuming we know the treatments we will apply, the first step in a power analysis is to specify the actual values of the treatment means under H_1, the alternative hypothesis. If the experiment has five treatments, we might speculate that the treatment means take the following values under H_1:

$$H_1 : \mu_1 = 20, \mu_2 = 22, \mu_3 = 22, \mu_4 = 25, \mu_5 = 18. \tag{12.2}$$

For example, these values could be the final weights of fish reared on five different diets. This is the form of H_1 needed by `proc power` (SAS Institute Inc. 2018). We can also express H_1 in terms of the usual model for this design, the fixed effects model of the form

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}. \tag{12.3}$$

In terms of the parameters of this model, H_1 is equivalent to saying

$$H_1 : \alpha_1 = -1.4, \alpha_2 = 0.6, \alpha_3 = 0.6, \alpha_4 = 3.6, \alpha_5 = -3.4, \tag{12.4}$$

where $\alpha_i = \mu_i - \mu$, and μ is the grand mean ($\mu = \sum \mu_i/5 = 107/5 = 21.4$) (Winer et al. 1991; Montgomery 1997).

The null hypothesis in terms of group means would have the form

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = \mu, \tag{12.5}$$

where μ is the grand mean. This is equivalent to the usual null hypothesis for one-way ANOVA, which is $H_0 : \alpha_i = 0$ for all i.

We also need to specify a standard deviation σ for the power analysis. We could potentially estimate σ from similar studies in the literature or through a pilot study. If the paper provides an ANOVA table, we can estimate σ using $\sqrt{MS_{within}} = \sqrt{MS_{error}}$. SAS actually calculates this quantity and labels it `Root MSE` – see previous printouts for `proc glm`. In other situations, you may not know σ precisely but can specify a plausible range of values. Continuing our example, we suppose that previous experiments suggest $\sigma = 3$.

To calculate the power for this example, we also need to specify a sample size n for the treatments. Usually we are interested in determining the power for a range of n values, so we can determine the minimal sample size to needed to reject H_0 with adequate power. Most power analyses assume an equal sample size for each treatment, because this usually yields a higher power than unbalanced designs. We also need to specify the Type I error rate for the overall ANOVA, and $\alpha = 0.05$ is customary.

The power is then calculated using the distribution of the statistic F_s under H_1, called the non-central F distribution (the distribution under H_0 is the F distribution). The non-central F distribution has three parameters, the usual two degrees of freedom plus an additional parameter λ, defined by the formula

$$\lambda = \frac{n \sum_{i=1}^{a} \alpha_i^2}{\sigma^2}, \tag{12.6}$$

where $\alpha_i = \mu_i - \mu$, and $\mu = \sum \mu_i / a$ (Winer et al. 1991, Montgomery 1997). Note that λ is a function of the α_i values, σ, and the sample size n. The non-central F distribution is equal to the F distribution when $\lambda = 0$, which can only happen if there are no treatment effects and $\alpha_i = 0$ for all i. As the value of λ increases, however, the noncentral F distribution will shift to the right, away from the position held by the F distribution. Note the similarity of this quantity with ϕ, which determines the power for one-sample Z and t tests.

Figure 12.1 shows the F and noncentral F distributions for the power analysis example described above, with $a = 5$, the α_i values as specified, and $\sigma = 3$. We also assume for the moment that $n = 5$, and set $\alpha = 0.05$. For this design, we have $df_1 = a - 1 = 5 - 1 = 4$, $df_2 = a(n-1) = 5(5-1) = 20$. For $\alpha = 0.05$, we would reject H_0 if F_s, the test statistic for one-way ANOVA (Chapter 11), exceeded 2.866 (see Table F). We also need to calculate a value of λ for the noncentral F distribution. We have

$$\lambda = \frac{5\left[(-1.4)^2 + 0.6^2 + 0.6^2 + 3.6^2 + (-3.4)^2\right]}{3^2} = 15.111. \tag{12.7}$$

We see that the noncentral F lies to the right of the F distribution, because λ is fairly large in this example. What is the power of the test? It is the area of the noncentral F distribution lying to the right of 2.866, because this is the probability that F_s will exceed 2.866 under H_1, i.e., the probability of rejecting H_0 if it is false and H_1 is true.

What would happen to the power for other values of n or σ, or for that matter smaller or larger differences among groups under H_1 (implying smaller or larger α_i values)? Any change that increases the value of λ will increase the power of the test, because it reduces the amount of overlap between the two distributions. Examining λ, we see that larger n, larger differences among groups, and smaller σ values would all increase λ and so increase the power of the test. Larger α (Type I error rate) values also increase the power of the test, because they reduce the acceptance and increase the rejection region size. Sample size n also has an effect on power through the acceptance region – larger n reduces its upper boundary through its effect on $df_2 = a(n-1)$. Fig. 12.2 shows the F and noncentral F distributions for the power example, now using $n = 8$. Note how the overlap between the two distributions is reduced for larger n, increasing the power of the test. See Table 12.1 for a summary of how these factors affect power and β.

Table 12.1: Effects on power and the Type II error rate β of changes in various parameters. The arrows indicate if a particular quantity increases or decreases.

Parameter	Direction	λ	power	β
α_i values	↑	↑	↑	↓
n	↑	↑	↑	↓
σ	↑	↓	↓	↑
α	↑	no change	↑	↓

The effect of n on λ implies that a sufficient large sample size can generate adequate power, even when the α_i values are small or σ is large. Thus, large sample sizes should make it possible to detect small treatment effects, and can also compensate for noisy data.

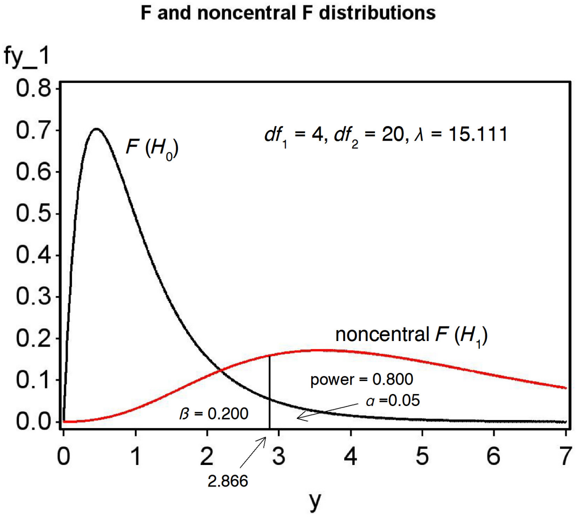


Figure 12.1: The F and noncentral F distributions for the power example, using $n = 5$.

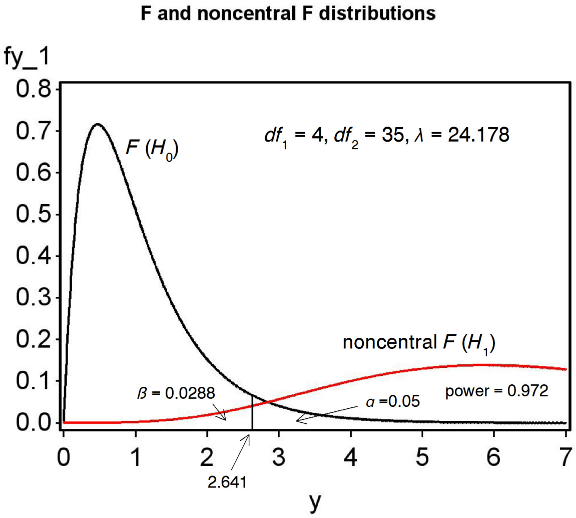


Figure 12.2: The F and noncentral F distributions for the power example, for $n = 8$.

12.2 Power analysis - SAS Demo

SAS makes power analysis relatively easy and provides specific methods for one-way ANOVA and many other designs. Consider our previous example involving five different treatments. We are interested in determining the power of a one-way ANOVA, when the following alternative hypothesis is true:

$$H_1 : \mu_1 = 20, \mu_2 = 22, \mu_3 = 22, \mu_4 = 25, \mu_5 = 18. \tag{12.8}$$

We need another piece of information for the power analysis, the value of σ. From preliminary studies or a previously published paper, we estimate that $\sigma = 3$. We also specify the Type I error rate, setting $\alpha = 0.05$.

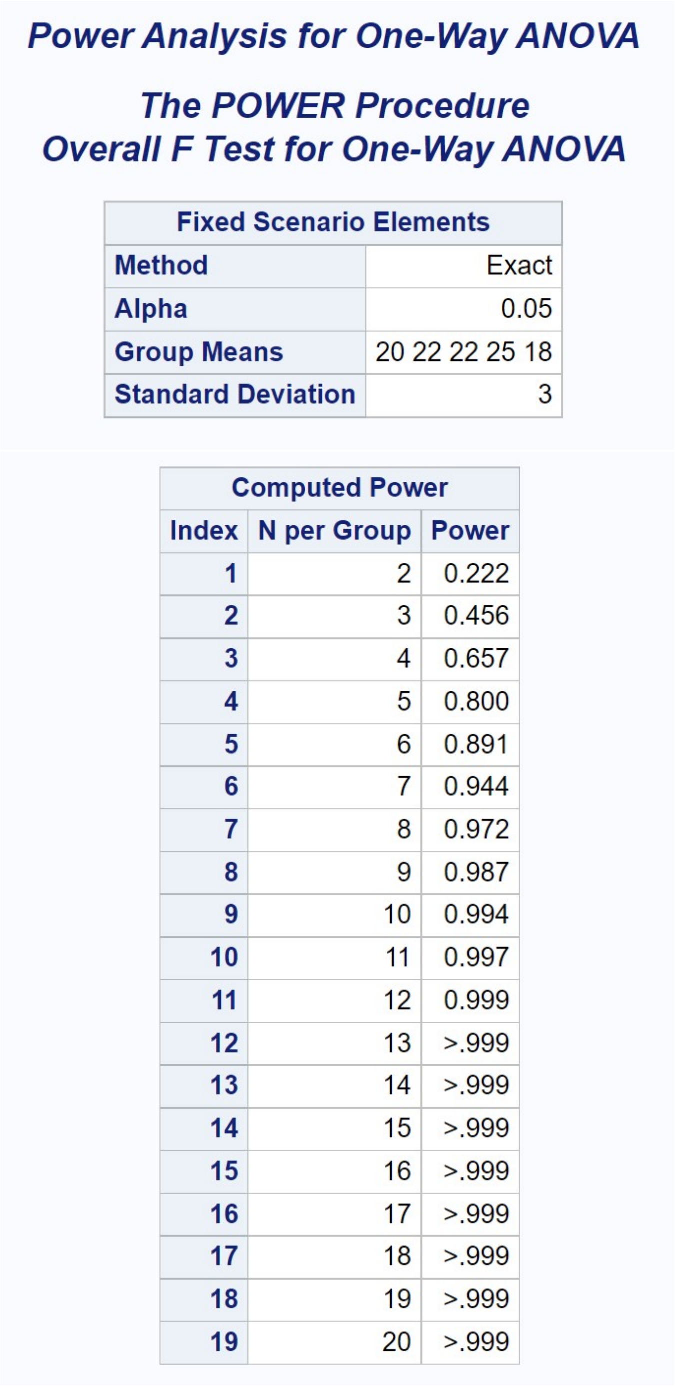
This is everything required to carry out a power analysis using `proc power` (SAS Institute Inc. 2018). We first specify that we want a power analysis for one-way ANOVA using the option `onewayanova`. The means for each treatment group are specified using the `groupmeans` option, with the means listed in parentheses. See program listing below.

The values of σ and α are similarly specified using the `stddev` and `alpha` options. We are interested in determining the power for a range of n values, the sample size per group. This is specified using the `npergroup` option. You can either give a list of n values or use the syntax `x to y by z` to specify a sequence of values. The `power` option is specified as a missing value (a period), because we want SAS to solve for power as a function of sample size per group. The `plot` command generates a plot of power vs. sample size.

We see that power increases rapidly with sample size per group (n), from both the SAS output and graph (Fig. 12.3, 12.4). A power value of 0.8 is achieved for $n = 5$ in this example.

SAS Program

```
* oneway_power.sas;
title ’Power Analysis for One-Way Anova’;
proc power;
    onewayanova
        groupmeans = (20 22 22 25 18)
        stddev = 3
        alpha = 0.05
        npergroup = 2 to 20 by 1
        power = . ;
    plot x=n;
run;
quit;
```



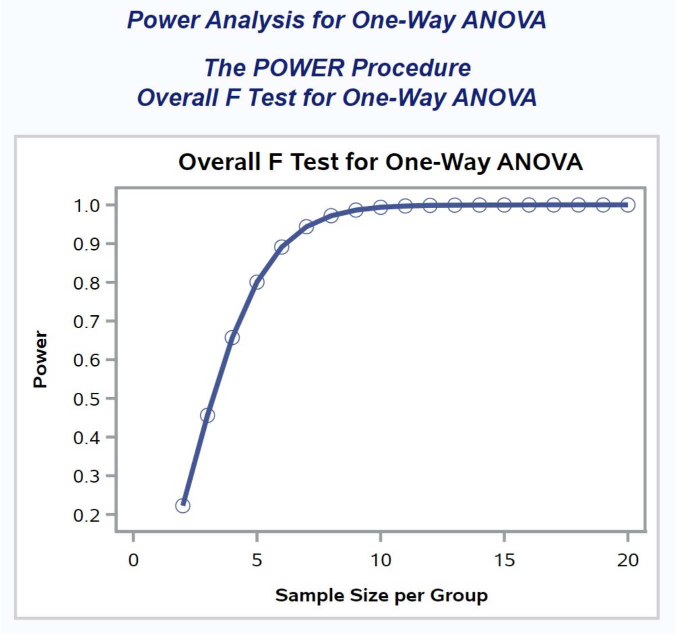
Power Analysis for One-Way ANOVA

The POWER Procedure
Overall F Test for One-Way ANOVA

Fixed Scenario Elements	
Method	Exact
Alpha	0.05
Group Means	20 22 22 25 18
Standard Deviation	3

Computed Power		
Index	N per Group	Power
1	2	0.222
2	3	0.456
3	4	0.657
4	5	0.800
5	6	0.891
6	7	0.944
7	8	0.972
8	9	0.987
9	10	0.994
10	11	0.997
11	12	0.999
12	13	>.999
13	14	>.999
14	15	>.999
15	16	>.999
16	17	>.999
17	18	>.999
18	19	>.999
19	20	>.999

Figure 12.3: `one-way_power.sas - proc power`



Power Analysis for One-Way ANOVA

The POWER Procedure
Overall F Test for One-Way ANOVA

Figure 12.4: `one-way_power.sas - proc power`

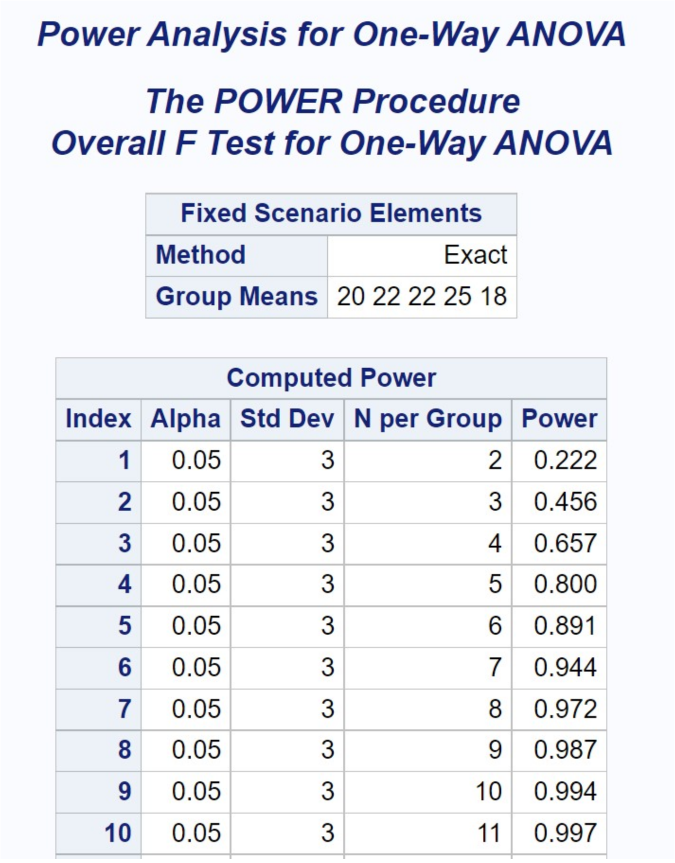
12.3 Power analysis continued - SAS demo

It is often worthwhile to compare power curves for different values of σ and α, to see how these influence power. We can obtain this from `proc power` by specifying several different values of these parameters. We will examine the results for $\alpha = 0.05$ vs. 0.01 and $\sigma = 3$ vs. 6. These are requested by listing both values under the `alpha` and `stddev` statements. See program and output below.

The effects of α and σ on power can be readily seen in Fig. 12.6. Lower α reduces the power of the test across all sample sizes, because it makes it harder to reject H_0 under any circumstances. Larger values of σ also decrease the power at all sample sizes. The larger the value of σ, the more variable the data, and the harder it is for the statistical test to distinguish between the null and alternative hypotheses. Adequate power is only obtained for a larger sample size.

SAS Program

```
* oneway_power2.sas;
title 'Power Analysis for One-Way Anova';
proc power;
    onewayanova
        groupmeans = (20 22 22 25 18)
        stddev = 3 6
        alpha = 0.05 0.01
        npergroup = 2 to 20 by 1
        power = . ;
    plot x=n;
run;
quit;
```



Power Analysis for One-Way ANOVA

The POWER Procedure
Overall F Test for One-Way ANOVA

Fixed Scenario Elements	
Method	Exact
Group Means	20 22 22 25 18

Computed Power				
Index	Alpha	Std Dev	N per Group	Power
1	0.05	3	2	0.222
2	0.05	3	3	0.456
3	0.05	3	4	0.657
4	0.05	3	5	0.800
5	0.05	3	6	0.891
6	0.05	3	7	0.944
7	0.05	3	8	0.972
8	0.05	3	9	0.987
9	0.05	3	10	0.994
10	0.05	3	11	0.997

etc.

Figure 12.5: `one-way_power2.sas - proc power`

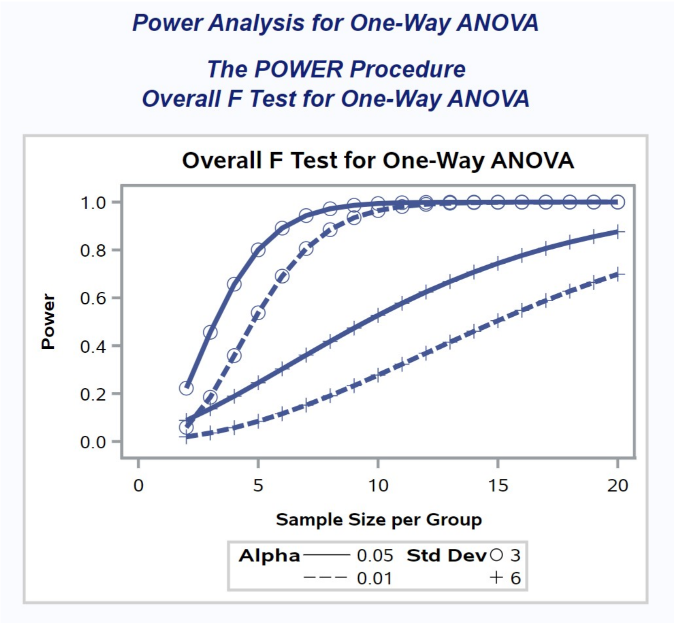


Figure 12.6: `one-way_power2.sas - proc power`

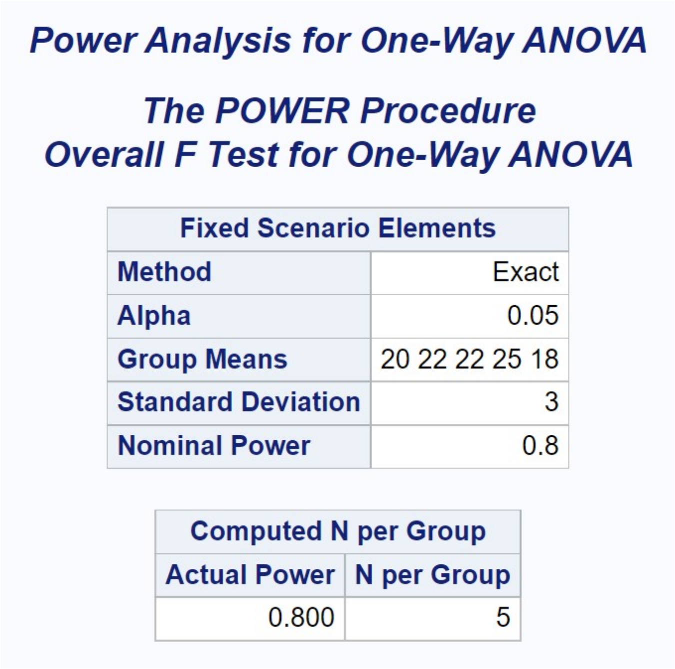
12.4 Power analysis continued - SAS demo

The SAS procedure `power` can be used to directly find the sample size n for a power of 0.8. One way is to simply read the value of n from a power vs. sample size graph, choosing the smallest n that gives power greater than or equal to 0.8. Returning to the first output we generated using `power` with $\alpha = 0.05$ and $\sigma = 3$, we see that for $n = 5$ the power exactly equals 0.8, so this is our sample size.

Alternately, you can set a power value of 0.8 and have `proc power` find the sample size. We first set the power option equal to 0.8 in the program, then change the `npergroup` option to a missing value, which tells `power` to solve for it. See program below and attached SAS output. SAS indicates that a sample size of $n = 5$ would give power = 0.8. This is the same result as obtained earlier by inspecting the power curve. For this particular example, there was a value of n that gave exactly the required power. More often, `power` will provide an n that guarantees power ≥ 0.8, not exactly 0.8.

SAS Program

```
* oneway_power3.sas;
title 'Power Analysis for One-Way ANOVA';
proc power;
    onewayanova
        groupmeans = (20 22 22 25 18)
        stddev = 3
        alpha = 0.05
        npergroup = .
        power = 0.8;
run;
quit;
```



Power Analysis for One-Way ANOVA

The POWER Procedure
Overall F Test for One-Way ANOVA

Fixed Scenario Elements	
Method	Exact
Alpha	0.05
Group Means	20 22 22 25 18
Standard Deviation	3
Nominal Power	0.8

Computed N per Group	
Actual Power	N per Group
0.800	5

Figure 12.7: `one-way_power3.sas - proc power`

12.5 References

Cohen, J. (1988) *Statistical Power Analysis for the Behavioral Sciences, Second Edition.* Lawrence Erlbaum Associates, Inc., Hillsdale, NJ.

Montgomery, D. C. (1997) *Design and Analysis of Experiments.* John Wiley & Sons, New York, NY.

SAS Institute Inc. (2018) *SAS/STAT 15.1 Users Guide* SAS Institute Inc., Cary, NC

Steidl, R. J., Hayes, J. P. & Schauber, E. (1997) Statistical power analysis in wildlife research. *Journal of Wildlife Management* 61: 270-279.

Winer, B. J., Brown, D. R. & Michels, K. M. (1991) *Statistical Principles in Experimental Design.* McGraw-Hill, Inc., Boston, MA.

12.6 Problems

1. Suppose you want to compare the effect of four different diets on the weight of prawns reared in aquaculture ponds. There is a standard diet (S) and three other diets (A, B and C) that will be fed to prawns in replicate ponds. Relative to diet S, you would like to see a 20% increase in weight on diet A, a 20% increase on diet B, and a 30% increase on diet C. If the mean weight on diet S is 100 g, this translates into the following alternative hypothesis:

   $$H_1 : \mu_S = 100, \mu_A = 120, \mu_B = 120, \mu_C = 130. \tag{12.9}$$

   From previous studies the researchers estimate that $\sigma = 22$. Assume a Type I error rate of $\alpha = 0.05$.

   (a) Use SAS and `proc power` to determine the sample size per treatment (number of ponds) necessary to give power ≥ 0.8. Attach your SAS program and output.

   (b) Repeat the same analysis for $\alpha = 0.01$. How does this change in the Type I error rate affect the sample size? Why?

2. Suppose you want to compare the effect of five different diets on the weight of fish reared in aquaculture. There is a control diet (C) and four other diets (D1, D2, D3, and D4). Relative to diet S, you would like to see a 10% increase in weight on diet D1, a 15% increase on diet D2, and 20% increases on diets D3 and D4. If the weight on the control diet C is 100 g, this translates into the following alternative hypothesis:

   $$H_1 : \mu_C = 100, \mu_{D1} = 110, \mu_{D2} = 115, \mu_{D3} = 120, \mu_{D4} = 120. \tag{12.10}$$

   Previous studies suggest that $\sigma = 10$. Assume a Type I error rate of $\alpha = 0.05$.

   (a) Use SAS to determine the sample size per treatment necessary to give power ≥ 0.8. Attach your program and output.

   (b) Repeat the same analysis for the following alternative hypothesis:

   $$H_1 : \mu_C = 100, \mu_{D1} = 105, \mu_{D2} = 108, \mu_{D3} = 110, \mu_{D4} = 110. \tag{12.11}$$

   How does this change affect the sample size? Why?

Chapter 13

Multiple Comparisons

One-way ANOVA, as well as more complex variants, provides a test of an overall null hypothesis of the form $H_0 : \alpha_i = 0$ for all i vs. H_1 : some $\alpha_i \neq 0$. If we obtain a small P value for this test, it provides evidence against H_0 and in favor of H_1. However, this overall test provides little information on whether particular groups are different. We now turn to statistical methods designed to compare pairs of groups for one-way ANOVA designs. These procedures allow comparisons to be made among all possible pairs of groups, or sometimes one group vs. all others, and are collectively called **multiple comparisons**. Although multiple comparisons are often conducted in association with ANOVA, they are in fact stand-alone procedures (Hsu 1996). There is no need to conduct an ANOVA before using these procedures, although SAS will generate an overall F test regardless. Moreover, significant differences between groups in multiple comparisons may not coincide with a significant overall F test, or vice versa.

13.1 Models for multiple comparisons

The statistical model for multiple comparisons is basically the one-way ANOVA model expressed in a different form. The one-way ANOVA model is

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \tag{13.1}$$

where μ is the grand mean, α_i is the deviation from the grand mean caused by the ith group, and $\epsilon_{ij} \sim N(0, \sigma^2)$. For multiple comparison procedures it is common to define $\mu_i = \mu + \alpha_i$, and so the one-way model becomes

$$Y_{ij} = \mu_i + \epsilon_{ij}. \tag{13.2}$$

We can think of μ_i as the mean of the ith group, where there are a total groups.

Now consider two groups i and j in a study which have means μ_i and μ_j, where $i \neq j$. We will be interested in estimating the difference in the means of these two groups, $\mu_i - \mu_j$, and finding a confidence interval to accompany this estimate for all possible pairs of groups. We will also be interested in testing whether the means of the two groups are equal, namely $H_0 : \mu_i = \mu_j$ or equivalently $H_0 : \mu_i - \mu_j = 0$, again for all possible pairs of groups. For a study with a groups, this amounts to $a(a-1)/2$ pairs of groups. For example, if there are $a = 3$ groups there are $3(3-1)/2 = 3$ possible pairwise comparisons (groups 1-2, 2-3, and 1-3). There are multiple comparison methods that provide estimates, confidence intervals, and tests, while others provide only tests but have more statistical power. The basic purpose of these procedures is to statistically test which pairs of treatments are different, and provide some idea of the magnitude of the difference. We will examine three procedures in this category, known as **all possible pairwise comparisons**. The procedures are called Fisher's least significant difference, the Tukey procedure, and the Ryan-Einot-Gabriel-Welsch (REGW) procedure (Hsu 1996).

For experiments that have a clearly identifiable control group, it may be appropriate to compare each group with only the control. For example, suppose the control is a standard drug treatment for a disease. We may only be interested in treatments that give a significantly better (or maybe worse) result compared to the control, and are not interested in other comparisons among the treatments. For a study with a groups including the control, this amounts to $a - 1$ pairs of groups with the control. For example, if there are $a = 3$ groups with the first group ($i = 1$) the control, there are $3 - 1 = 2$ possible comparisons (groups 1-2 and 1-3). We will examine Dunnett's procedure in this category, known as **multiple comparisons with a control** (Hsu 1996).

13.2 Error rates in multiple comparisons

There are two error rates commonly used to describe multiple comparison procedures. One is the **per comparison** error rate, which is the Type I error rate for a single test comparing a single pair of groups. This rate is like that used in other statistical tests we have encountered, where only a single test is considered. The second is the **experimentwise error rate**, or **EER**. **The EER is defined as the probability of one or more Type I errors (rejecting H_0 when it is true) in a set of comparisons.**

Why do we need two error rates? Multiple comparison procedures such as the ones mentioned above can involve a substantial number of statistical tests, one test for each pair of groups. For example, with $a = 5$ groups there would be $5(5-1)/2 = 10$ possible pairwise comparisons, while for $a = 10$ groups we would have $10(10-1)/2 = 45$ comparisons! Given this many comparisons and tests, it is quite possible that some pairs would yield a significant test result even if the null hypothesis were true, i.e., we would reject $H_0 : \mu_i = \mu_j$ for one or more pairs of groups, even though there is no difference between the groups. For example, suppose that the per comparison error rate is set at the typical $\alpha = 0.05$ value, which amounts to a 1 in 20 chance of rejecting H_0 when it is true. Given $a = 10$ and 45 total tests, we would expect to see a few significant test results just by chance. This difficulty has been called the **multiplicity problem** (Westfall et al. 1999).

To see the magnitude of the multiplicity problem, we can plot the EER for the least significant difference procedure, which controls the per comparison error rate but not the EER. Fig. 13.1 shows a plot of the EER vs. the number of groups or treatments (a). The least significant difference procedure is a t test that compares the means for each pair of groups, with each test conducted at the same α level, in this case $\alpha = 0.05$. We see that the EER, and the number of pairwise comparisons, increases rapidly with the number of groups. Thus, it becomes more likely that any significant differences reported among groups are in fact Type I errors. In contrast, methods designed to control the EER, such as the Tukey procedure, would maintain an EER of 0.05 regardless of the number of groups. These tests manage the EER by essentially reducing the per comparison error rate for each test. **The penalty of controlling the EER is a loss of power to detect differences among groups where they do exist.**

Multiple comparison procedures have been the subject of considerable controversy in the ecological and statistical literature. A number of methods were popular because they gave significant results more often than competing ones. These include the least significant difference procedure, Fisher's protected least significant difference, Duncan's multiple range test, and the Student-Newman-Keuls test. Unfortunately, these particular tests do not control the experimentwise error rate (Day & Quinn 1989, Hsu 1996).

Another error rate in common use is the **false discovery rate** or **FDR** (Benjamini & Hochberg 1995). **This rate is defined as the expected proportion of significant tests that are Type I errors.** Procedures that use the FDR have more power than those controlling the EER, but with more Type I errors. We will examine the rationale for FDR procedures later in the chapter.

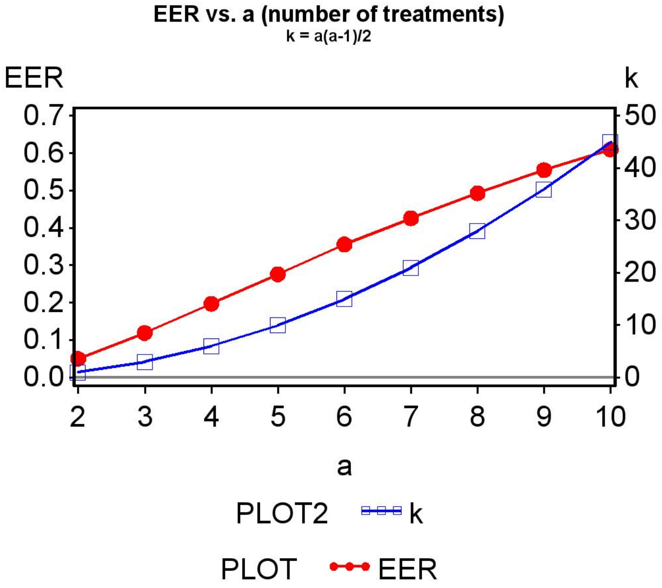


Figure 13.1: Plot of the experimentwise error rate vs. a, the number of treatments or groups, using $\alpha = 0.05$ for each comparison. Also shown is the number of pairwise comparisons ($k = a(a-1)/2$) vs. a.

13.3 All pairwise comparisons

This section examines three different methods for all pairwise comparisons among groups, the least significant difference, Tukey, and REGW methods. The least significant difference method does not control the EER, but is simple in form and a useful starting point. It provides estimates and confidence intervals for $\mu_i - \mu_j$, the difference between the group means for any pair of groups, as well as a statistical test for $H_0 : \mu_i - \mu_j$. The Tukey procedure is similar to the least significant difference except that it controls the EER. We also examine the REGW method, an example of a **multiple range test**. Multiple range procedures only provide tests, not confidence intervals, but are more powerful procedures.

13.3.1 Least significant difference

We first develop confidence intervals and construct statistical tests for the least significant difference procedure, using methods similar to those in Chapter 9 and 10. For multiple comparisons, we are interested in estimating $\mu_i - \mu_j$ and finding a confidence interval for this quantity. It seems reasonable to use $\bar{Y}_i - \bar{Y}_j$ to estimate $\mu_i - \mu_j$, but what is the variance of this estimate? Using the rules for calculating the variance of a sum of random variables (Chapter 7), we have

$$Var[\bar{Y}_i - \bar{Y}_j] = Var[\bar{Y}_i] + (-1)^2 Var[\bar{Y}_j] = \sigma^2/n + \sigma^2/n = 2\sigma^2/n. \quad (13.3)$$

ANOVA provides an estimate of σ^2, namely MS_{within}, and so we can estimate the variance of $\bar{Y}_i - \bar{Y}_j$ using the quantity $2MS_{within}/n$, which has $a(n-1)$ degrees of freedom. Using these results, it can be shown that the quantity

$$\frac{(\bar{Y}_i - \bar{Y}_j) - (\mu_i - \mu_j)}{\sqrt{\frac{2MS_{within}}{n}}} \sim t_{a(n-1)}. \quad (13.4)$$

We use this quantity to first derive a confidence interval for $\mu_i - \mu_j$. Using Table T, we can find a value of $c_{\alpha,a(n-1)}$ for $a(n-1)$ degrees of freedom such that the following equation is true:

$$P\left[-c_{\alpha,a(n-1)} < \frac{(\bar{Y}_i - \bar{Y}_j) - (\mu_i - \mu_j)}{\sqrt{\frac{2MS_{within}}{n}}} < c_{\alpha,a(n-1)}\right] = 1 - \alpha. \quad (13.5)$$

Rearranging this equation, we obtain

$$P\left[\bar{Y}_i - \bar{Y}_j - c_{\alpha,a(n-1)}\sqrt{\frac{2MS_{within}}{n}} < \mu_i - \mu_j < \bar{Y}_i - \bar{Y}_j + c_{\alpha,a(n-1)}\sqrt{\frac{2MS_{within}}{n}}\right] = 1 - \alpha. \quad (13.6)$$

The confidence interval would therefore be the interval

$$\left(\bar{Y}_i - \bar{Y}_j - c_{\alpha,a(n-1)}\sqrt{\frac{2MS_{within}}{n}}, \bar{Y}_i - \bar{Y}_j + c_{\alpha,a(n-1)}\sqrt{\frac{2MS_{within}}{n}}\right). \quad (13.7)$$

The center of the confidence interval is located at $\bar{Y}_i - \bar{Y}_j$, the estimate of $\mu_i - \mu_j$. We will later illustrate how this interval is calculated in a SAS demo of the least significant difference procedure.

Now suppose we want to test $H_0 : \mu_i = \mu_j$ or equivalently $H_0 : \mu_i - \mu_j = 0$. Under H_0, the test statistic

$$T_s = \frac{(\bar{Y}_i - \bar{Y}_j) - 0}{\sqrt{\frac{2MS_{within}}{n}}} = \frac{(\bar{Y}_i - \bar{Y}_j)}{\sqrt{\frac{2MS_{within}}{n}}} \sim t_{a(n-1)}. \quad (13.8)$$

Using a Type I error rate of α, the acceptance region of the test would be the interval $(-c_{\alpha,a(n-1)}, c_{\alpha,a(n-1)})$, where $c_{\alpha,a(n-1)}$ is determined using Table T (see Chapter 10). We would reject H_0 if it falls on the edge or outside this interval.

We can rearrange the test given above into a different form, one that is commonly used for multiple comparisons. Recall that one would accept H_0 if T_s falls inside the acceptance region $(-c_{\alpha,a(n-1)}, c_{\alpha,a(n-1)})$, which implies

$$-c_{\alpha,a(n-1)} < \frac{(\bar{Y}_i - \bar{Y}_j)}{\sqrt{\frac{2MS_{within}}{n}}} < c_{\alpha,a(n-1)}. \quad (13.9)$$

We can rearrange this into the form

$$-c_{\alpha,a(n-1)}\sqrt{\frac{2MS_{within}}{n}} < \bar{Y}_i - \bar{Y}_j < c_{\alpha,a(n-1)}\sqrt{\frac{2MS_{within}}{n}}, \quad (13.10)$$

or

$$-LSD < \bar{Y}_i - \bar{Y}_j < LSD, \quad (13.11)$$

where

$$LSD = c_{\alpha,a(n-1)}\sqrt{\frac{2MS_{within}}{n}}. \quad (13.12)$$

The quantity LSD is called the least significant difference. We would accept H_0 if $\bar{Y}_i - \bar{Y}_j$ falls inside the interval $(-LSD, LSD)$, or equivalently if $|\bar{Y}_i - \bar{Y}_j| < LSD$. Conversely, we would reject H_0 if $|\bar{Y}_i - \bar{Y}_j| \geq LSD$. This same rule applies to any pair of groups, because LSD would take the same value. Any pair of means that equals or exceeds this value is declared to be significantly different.

The confidence intervals we derived for $\mu_i - \mu_j$ can also be expressed in this format. In particular, the confidence interval would have the form

$$\left(\bar{Y}_i - \bar{Y}_j - LSD, \bar{Y}_i - \bar{Y}_j + LSD\right). \tag{13.13}$$

13.3.2 Least significant difference - SAS demo

Kneitel & Lessin (2010) studied the effect of eutrophication on vernal pools in California. They were interested in the effect of eutrophication (nutrient addition) on algae cover during the period the pools were filled with water, as well as vascular plant cover later in the season. Experimental pools were subjected to five different treatments: low, medium, high, and very high nutrient addition levels, and a control to which no nutrients were added. We will use a simplified data set from this study to illustrate the least significant difference procedure in SAS. We first examine the data involving algae cover. Algae cover was expressed as a percentage of the pool covered, and for data of this type it is common to transform the data. The data were first converted to a proportion by dividing the percentage by 100, then the arcsine-square root transformation applied (see Chapter 15). See the `data` step in the SAS program below.

The program is similar to our previous one-way ANOVA programs, with the addition of a `means` statement within `proc glm`:

```
means treat / t cldiff lines;
```

This statement requests a mean for each level of `treat`, the treatment variable (SAS Institute Inc. 2018). The `t` option requests the least significant difference procedure, because it is essentially a t test. The option `cldiff` requests 95% confidence intervals for $\mu_i - \mu_j$ for all pairs of groups, while `lines` generates a diagram that indicates which pairs of groups are significantly different at the $\alpha = 0.05$ level. See the full program listing and SAS output below.

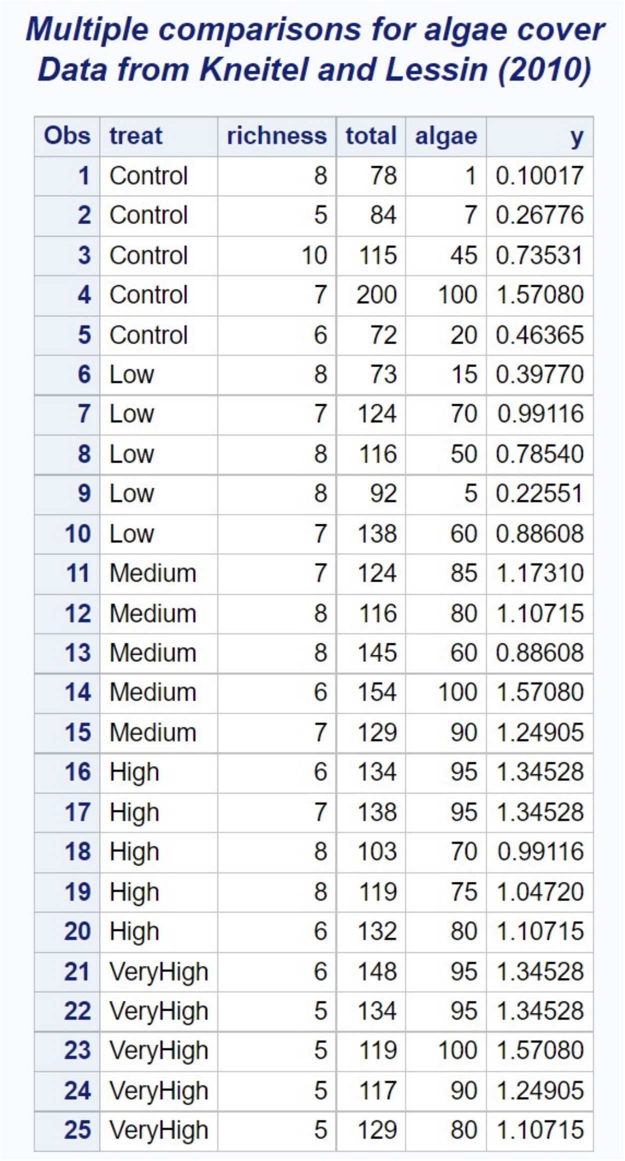
According to the one-way ANOVA output (Fig. 13.4), there was a highly significant difference among the nutrient treatments ($F_{4,20} = 4.76, P < 0.0073$). Confidence intervals for $\mu_i - \mu_j$ and $\mu_j - \mu_i$ are given for every pair of groups (Fig. 13.5). For example, SAS gives a confidence interval for $\mu_{\text{medium}} - \mu_{\text{control}}$ as well as $\mu_{\text{control}} - \mu_{\text{medium}}$. Also shown in the output is the diagram (Fig.

13.6) generated by the `lines` command, interpreted as follows. **Treatments covered by the same line are not significantly different, while if they share no lines they are significantly different.** There were six significant differences among pairs of treatments (`VeryHigh-Low`, `VeryHigh-Control`, `Medium-Low`, `Medium-Control`, `High-Low`, and `High-Control`). Note that the per comparison error rate used in the tests and confidence intervals is labeled `Alpha` in the SAS output.

The results from Fig. 13.6 can also be used to indicate significant differences on a graph, using letters instead of lines (Fig. 13.3). Treatments with the same letter are not significantly different. Note that SAS does not provide these letters – they were added to the graph using photo-editing software.

SAS Program

```
* Kneitel_2010_algae_lsd2.sas;
title 'Multiple comparisons for algae cover';
title2 'Data from Kneitel and Lessin (2010)';
data kneitel;
    input treat $ richness total algae;
    * Apply transformations here;
    y = arsin(sqrt(algae/100));
    datalines;
Control   8   78    1
Control   5   84    7
Control  10  115   45
Control   7  200  100
Control   6   72   20
Low       8   73   15
Low       7  124   70
Low       8  116   50
Low       8   92    5
Low       7  138   60
Medium    7  124   85
Medium    8  116   80
Medium    8  145   60
Medium    6  154  100
Medium    7  129   90
High      6  134   95
High      7  138   95
High      8  103   70
High      8  119   75
High      6  132   80
VeryHigh  6  148   95
VeryHigh  5  134   95
VeryHigh  5  119  100
VeryHigh  5  117   90
VeryHigh  5  129   80
;
run;
* Print data set;
proc print data=kneitel;
run;
* Plot means, standard errors, and observations;
proc gplot data=kneitel;
    plot y*treat=1 / vaxis=axis1 haxis=axis1;
    symbol1 i=std1mjt v=star height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* One-way anova with comparisons;
proc glm plots=diagnostics data=kneitel;
    class treat;
    model y = treat;
    * LSD or Students t - only controls the per comparison error rate;
    means treat / t cldiff lines;
run;
quit;
```



Multiple comparisons for algae cover
Data from Kneitel and Lessin (2010)

Obs	treat	richness	total	algae	y
1	Control	8	78	1	0.10017
2	Control	5	84	7	0.26776
3	Control	10	115	45	0.73531
4	Control	7	200	100	1.57080
5	Control	6	72	20	0.46365
6	Low	8	73	15	0.39770
7	Low	7	124	70	0.99116
8	Low	8	116	50	0.78540
9	Low	8	92	5	0.22551
10	Low	7	138	60	0.88608
11	Medium	7	124	85	1.17310
12	Medium	8	116	80	1.10715
13	Medium	8	145	60	0.88608
14	Medium	6	154	100	1.57080
15	Medium	7	129	90	1.24905
16	High	6	134	95	1.34528
17	High	7	138	95	1.34528
18	High	8	103	70	0.99116
19	High	8	119	75	1.04720
20	High	6	132	80	1.10715
21	VeryHigh	6	148	95	1.34528
22	VeryHigh	5	134	95	1.34528
23	VeryHigh	5	119	100	1.57080
24	VeryHigh	5	117	90	1.24905
25	VeryHigh	5	129	80	1.10715

Figure 13.2: `Kneitel_2010_algae_lsd2.sas - proc print`

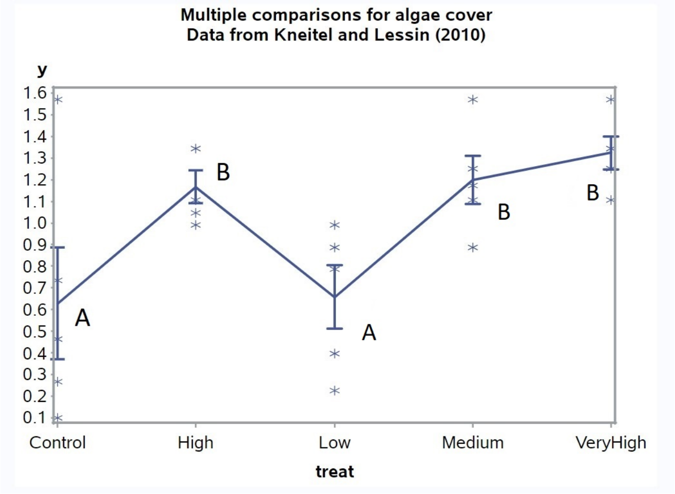
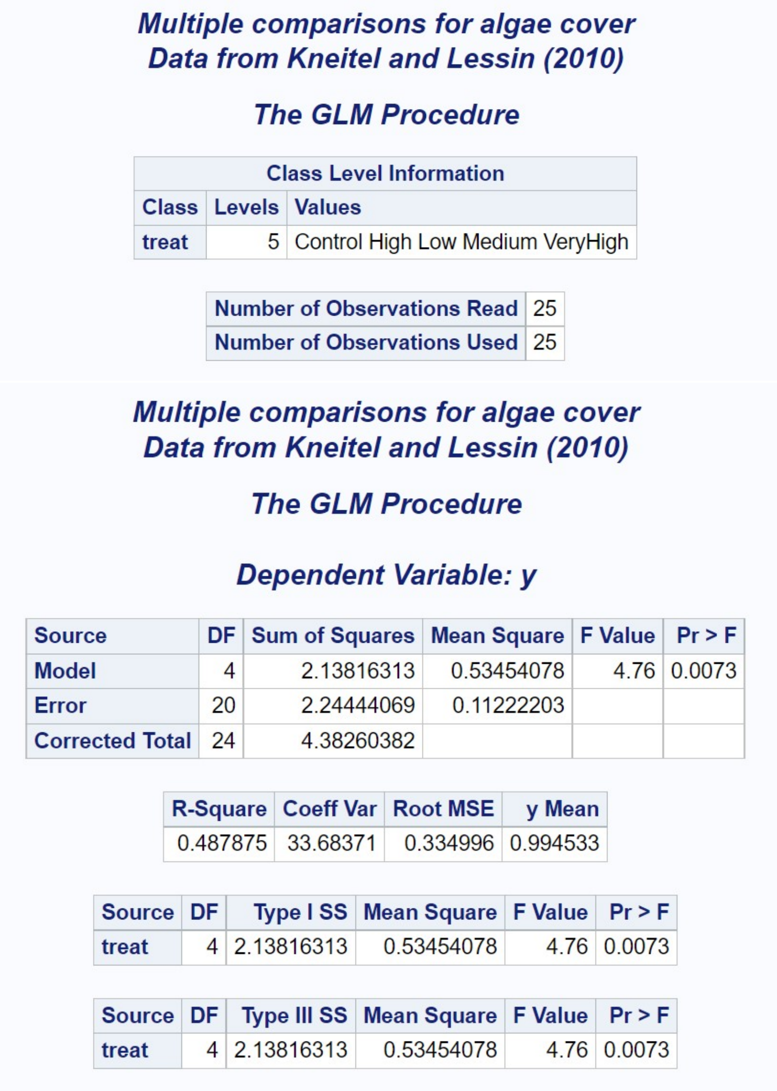


Figure 13.3: `Kneitel_2010_algae_lsd2.sas - proc gplot`



Multiple comparisons for algae cover
Data from Kneitel and Lessin (2010)

The GLM Procedure

Class Level Information		
Class	**Levels**	**Values**
treat	5	Control High Low Medium VeryHigh

Number of Observations Read	25
Number of Observations Used	25

Multiple comparisons for algae cover
Data from Kneitel and Lessin (2010)

The GLM Procedure

Dependent Variable: y

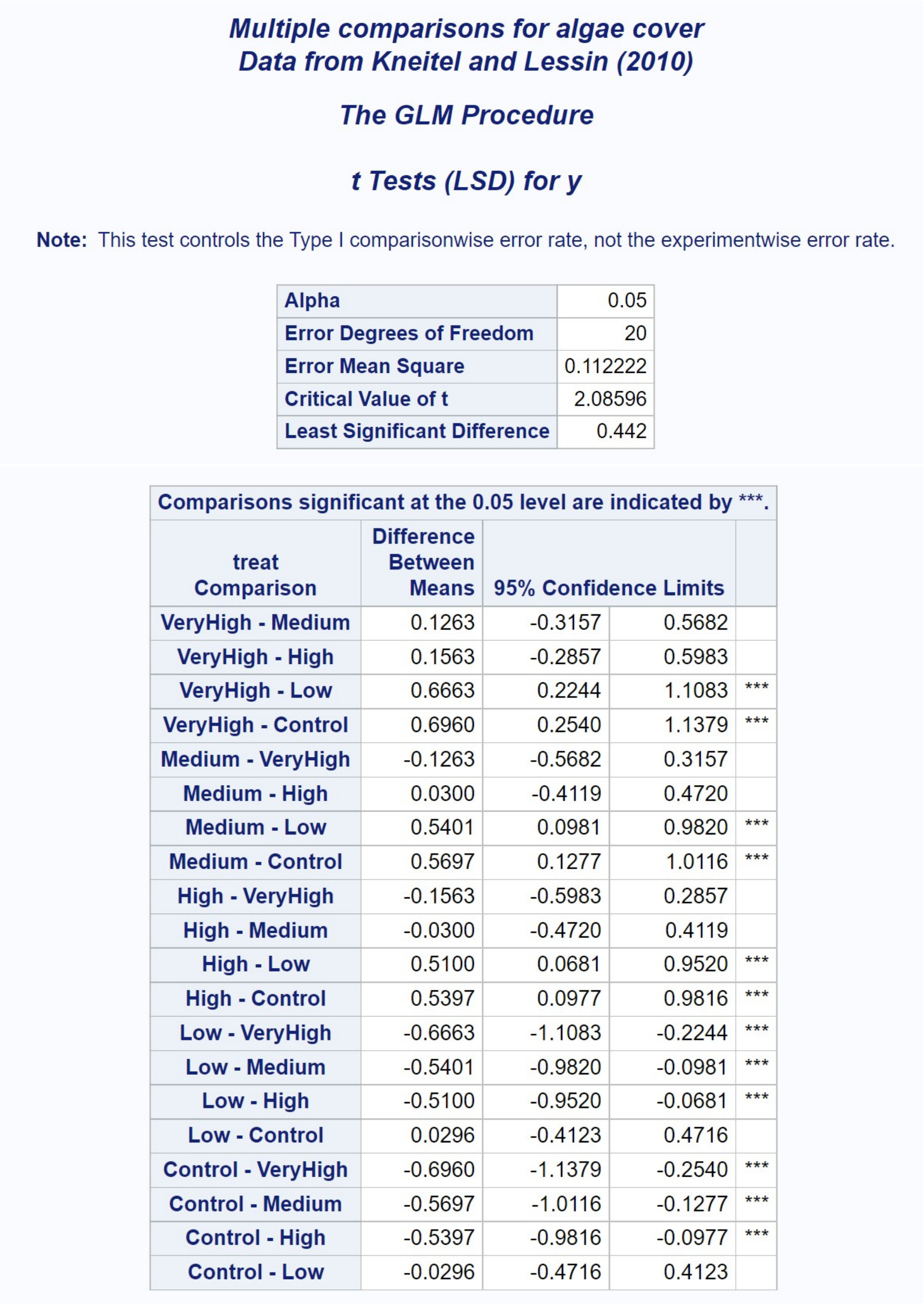
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	2.13816313	0.53454078	4.76	0.0073
Error	20	2.24444069	0.11222203		
Corrected Total	24	4.38260382			

R-Square	Coeff Var	Root MSE	y Mean
0.487875	33.68371	0.334996	0.994533

Source	DF	Type I SS	Mean Square	F Value	Pr > F
treat	4	2.13816313	0.53454078	4.76	0.0073

Source	DF	Type III SS	Mean Square	F Value	Pr > F
treat	4	2.13816313	0.53454078	4.76	0.0073

Figure 13.4: `Kneitel_2010_algae_lsd2.sas - proc glm`



Multiple comparisons for algae cover
Data from Kneitel and Lessin (2010)

The GLM Procedure

t Tests (LSD) for y

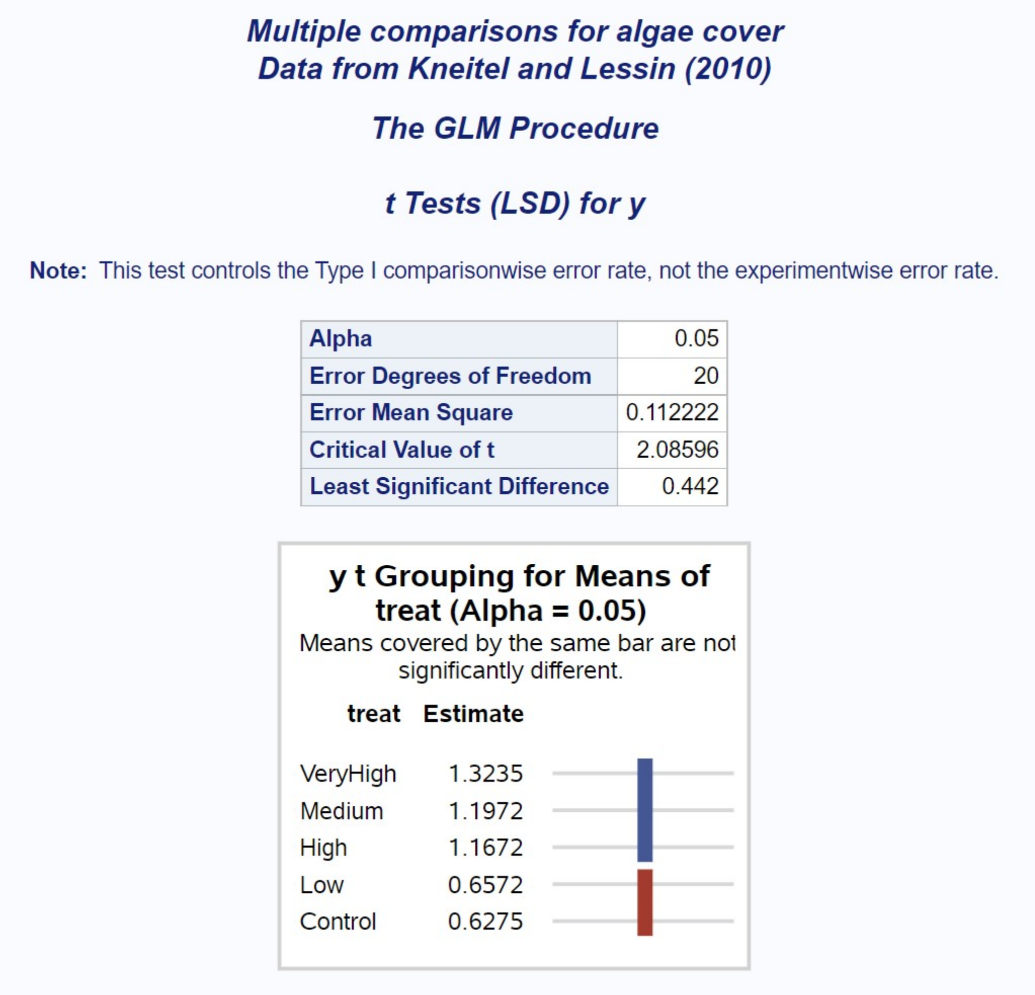
Note: This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	20
Error Mean Square	0.112222
Critical Value of t	2.08596
Least Significant Difference	0.442

Comparisons significant at the 0.05 level are indicated by ***.

treat Comparison	Difference Between Means	95% Confidence Limits		
VeryHigh - Medium	0.1263	-0.3157	0.5682	
VeryHigh - High	0.1563	-0.2857	0.5983	
VeryHigh - Low	0.6663	0.2244	1.1083	***
VeryHigh - Control	0.6960	0.2540	1.1379	***
Medium - VeryHigh	-0.1263	-0.5682	0.3157	
Medium - High	0.0300	-0.4119	0.4720	
Medium - Low	0.5401	0.0981	0.9820	***
Medium - Control	0.5697	0.1277	1.0116	***
High - VeryHigh	-0.1563	-0.5983	0.2857	
High - Medium	-0.0300	-0.4720	0.4119	
High - Low	0.5100	0.0681	0.9520	***
High - Control	0.5397	0.0977	0.9816	***
Low - VeryHigh	-0.6663	-1.1083	-0.2244	***
Low - Medium	-0.5401	-0.9820	-0.0981	***
Low - High	-0.5100	-0.9520	-0.0681	***
Low - Control	0.0296	-0.4123	0.4716	
Control - VeryHigh	-0.6960	-1.1379	-0.2540	***
Control - Medium	-0.5697	-1.0116	-0.1277	***
Control - High	-0.5397	-0.9816	-0.0977	***
Control - Low	-0.0296	-0.4716	0.4123	

Figure 13.5: `Kneitel_2010_algae_lsd2.sas - proc glm`



Multiple comparisons for algae cover
Data from Kneitel and Lessin (2010)

The GLM Procedure

t Tests (LSD) for y

Note: This test controls the Type I comparisonwise error rate, not the experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	20
Error Mean Square	0.112222
Critical Value of t	2.08596
Least Significant Difference	0.442

y t Grouping for Means of treat (Alpha = 0.05)

Means covered by the same bar are not significantly different.

treat	Estimate
VeryHigh	1.3235
Medium	1.1972
High	1.1672
Low	0.6572
Control	0.6275

Figure 13.6: `Kneitel_2010_algae_lsd2.sas - proc glm`

We will now calculate the value of LSD for this example to show how it is used to construct confidence intervals and tests. From the ANOVA output for `proc glm`, we see that $MS_{within} = 0.1122$ with 20 degrees of freedom. From Table T (Chapter 23), using $\alpha = 0.05$ we see that $c_{0.05,20} = 2.086$. There are also $n = 5$ replicates per treatment. We then have

$$LSD = c_{\alpha,a(n-1)}\sqrt{\frac{2MS_{within}}{n}} = 2.086\sqrt{\frac{2(0.1122)}{5}} = 0.4419. \qquad (13.14)$$

Note that SAS also displays the value of LSD in the output (Fig. 13.5, 13.6). We next calculate a 95% confidence interval for $\mu_{\text{medium}} - \mu_{\text{control}}$. Recall that the formula for the interval is

$$\left(\bar{Y}_i - \bar{Y}_j - LSD, \bar{Y}_i - \bar{Y}_j + LSD\right). \qquad (13.15)$$

Inserting the estimated means for these two treatments (see SAS output) in this formula, and the LSD value, we obtain

$$(1.1972 - 0.6275 - 0.4419, 1.1972 - 0.6275 + 0.4419) \qquad (13.16)$$

or $(0.1278, 1.0116)$. This confidence interval and the LSD value are quite close to the values obtained by SAS.

We next show how the LSD value is used to test $H_0 : \mu_{\text{medium}} - \mu_{\text{control}} = 0$ or equivalently $H_0 : \mu_{\text{medium}} = \mu_{\text{control}}$. We would reject H_0 if $|\bar{Y}_i - \bar{Y}_j| \geq LSD$. Inserting the estimated means for these two treatments, we see that $|1.1972 - 0.6275| = 0.5687 \geq 0.4419$, and so this pair of means was significantly different.

13.3.3 The Tukey procedure

The Tukey method for multiple comparisons is similar to the least significant difference procedure, except that it uses the **studentized range distribution** in place of the t distribution. The studentized range distribution is designed to control the EER for all pairwise comparisons among group means (Hsu 1996). Another advantage is that the confidence intervals constructed using this distribution are **simultaneous confidence intervals**. This means that the overall probability the confidence intervals include the true value of $\mu_i - \mu_j$, for all pairs of groups, is equal to $1 - \alpha$ for some specified α. The overall probability α is also the EER for the family of all pairwise tests.

The Tukey procedure makes use of a quantity called the honestly significant difference (HSD), defined as

$$HSD = q_{\alpha,a,a(n-1)} \sqrt{\frac{MS_{within}}{n}}. \tag{13.17}$$

The quantity $q_{\alpha,a,a(n-1)}$ is obtained from the studentized range distribution, and depends on α (the desired EER), the number of groups a, as well as the degrees of freedom for MS_{within}.

To test $H_0 : \mu_i = \mu_j$ or $H_0 : \mu_i - \mu_j = 0$, we accept H_0 if $|\bar{Y}_i - \bar{Y}_j| < HSD$, and reject it $|\bar{Y}_i - \bar{Y}_j| \geq HSD$. This same rule applies to any pair of groups, because HSD would take the same value. Any pair of means that equals or exceeds this value is declared to be significantly different. The Tukey confidence intervals are of the form

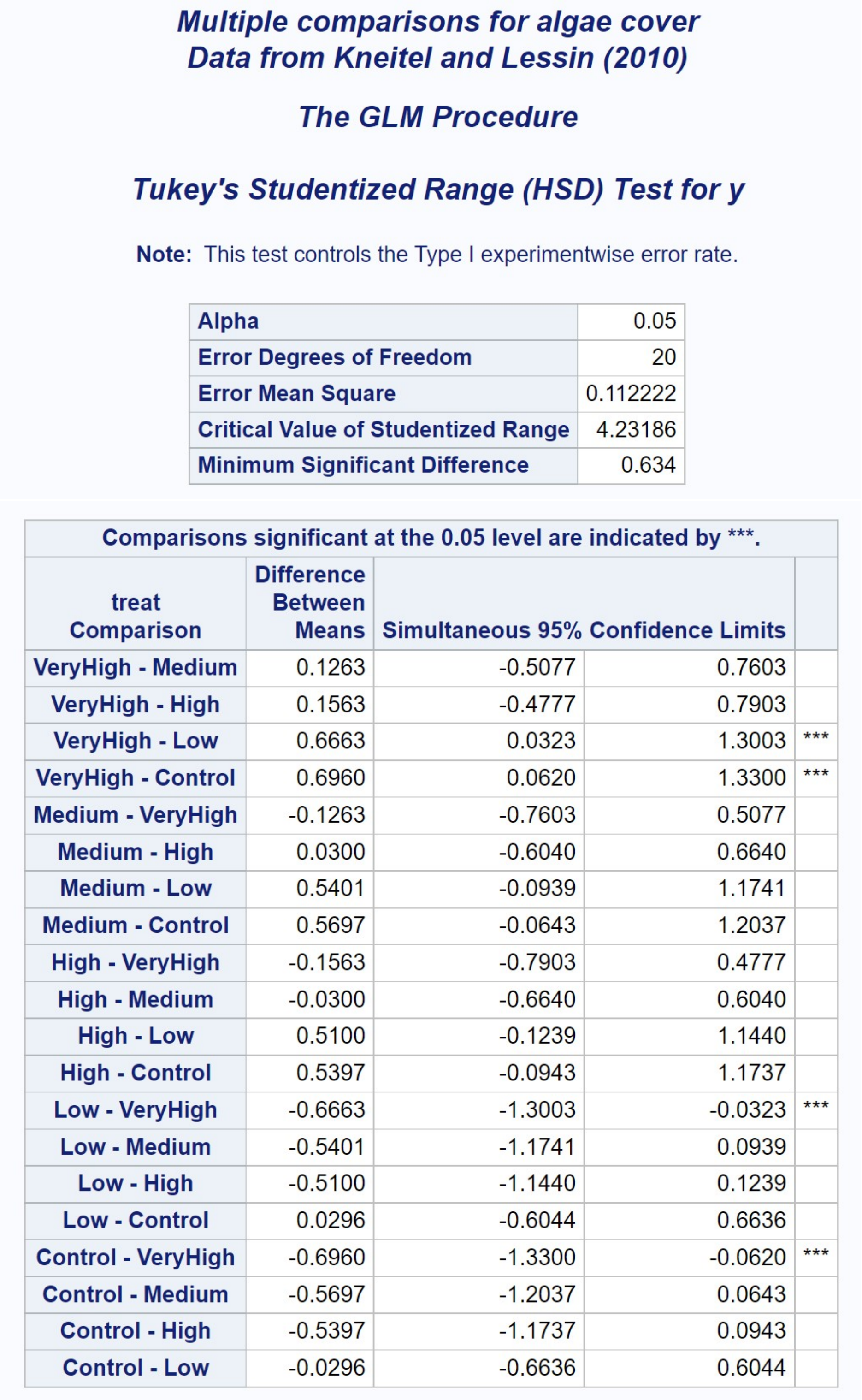
$$\left(\bar{Y}_i - \bar{Y}_j - HSD, \bar{Y}_i - \bar{Y}_j + HSD\right). \tag{13.18}$$

13.3.4 Tukey procedure - SAS demo

Implementing the Tukey procedure requires only a small change in our previous SAS program. The `means` statement within `proc glm` becomes

```
means treat / tukey cldiff lines;
```

Confidence intervals for $\mu_i - \mu_j$ and $\mu_j - \mu_i$ are given for every pair of groups, as well as a diagram indicating which treatments are significantly different (Fig. 13.7, 13.8). The value of `Alpha` listed in the output is the EER. Note that the Tukey method finds fewer significant comparisons than the least significant difference procedure. We see there are only two significant ones, `VeryHigh-Low` and `VeryHigh-Control`. This is a common pattern observed with multiple comparison tests, a few significant differences but also substantial overlap among treatments or groups.



Multiple comparisons for algae cover
Data from Kneitel and Lessin (2010)

The GLM Procedure

Tukey's Studentized Range (HSD) Test for y

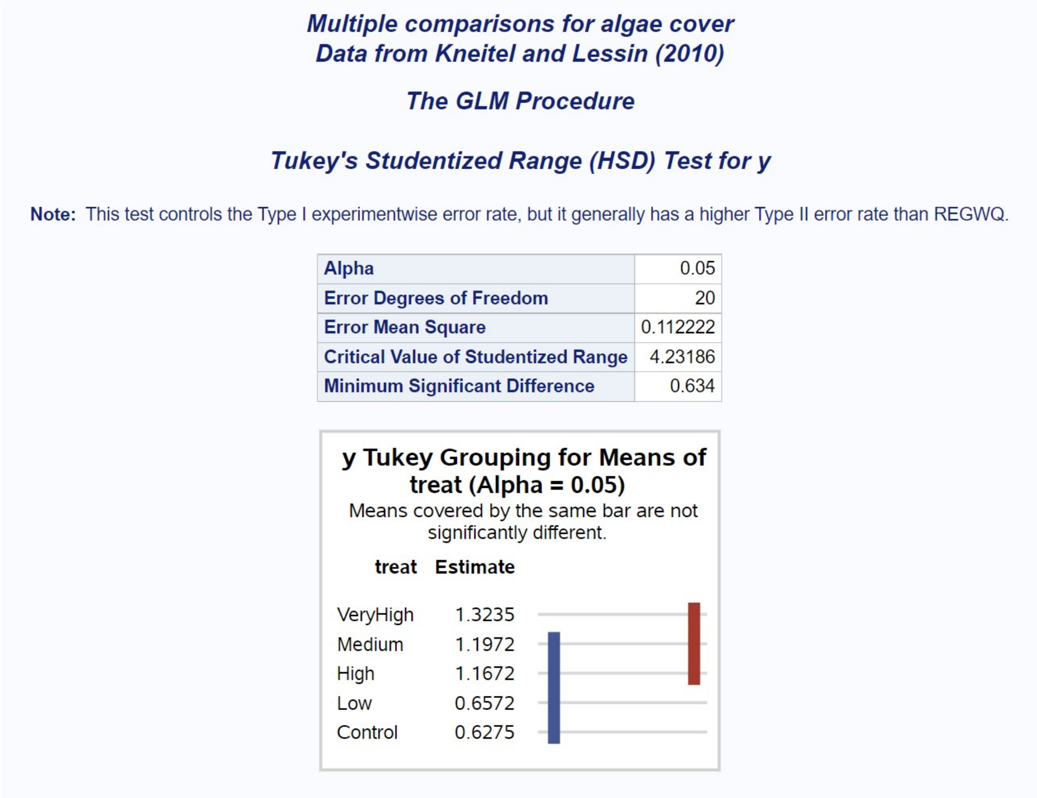
Note: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	20
Error Mean Square	0.112222
Critical Value of Studentized Range	4.23186
Minimum Significant Difference	0.634

Comparisons significant at the 0.05 level are indicated by ***.

treat Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
VeryHigh - Medium	0.1263	-0.5077	0.7603	
VeryHigh - High	0.1563	-0.4777	0.7903	
VeryHigh - Low	0.6663	0.0323	1.3003	***
VeryHigh - Control	0.6960	0.0620	1.3300	***
Medium - VeryHigh	-0.1263	-0.7603	0.5077	
Medium - High	0.0300	-0.6040	0.6640	
Medium - Low	0.5401	-0.0939	1.1741	
Medium - Control	0.5697	-0.0643	1.2037	
High - VeryHigh	-0.1563	-0.7903	0.4777	
High - Medium	-0.0300	-0.6640	0.6040	
High - Low	0.5100	-0.1239	1.1440	
High - Control	0.5397	-0.0943	1.1737	
Low - VeryHigh	-0.6663	-1.3003	-0.0323	***
Low - Medium	-0.5401	-1.1741	0.0939	
Low - High	-0.5100	-1.1440	0.1239	
Low - Control	0.0296	-0.6044	0.6636	
Control - VeryHigh	-0.6960	-1.3300	-0.0620	***
Control - Medium	-0.5697	-1.2037	0.0643	
Control - High	-0.5397	-1.1737	0.0943	
Control - Low	-0.0296	-0.6636	0.6044	

Figure 13.7: `Kneitel_2010_algae_tukey2.sas - proc glm`



Multiple comparisons for algae cover
Data from Kneitel and Lessin (2010)

The GLM Procedure

Tukey's Studentized Range (HSD) Test for y

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

Alpha	0.05
Error Degrees of Freedom	20
Error Mean Square	0.112222
Critical Value of Studentized Range	4.23186
Minimum Significant Difference	0.634

y Tukey Grouping for Means of treat (Alpha = 0.05)

Means covered by the same bar are not significantly different.

treat	Estimate
VeryHigh	1.3235
Medium	1.1972
High	1.1672
Low	0.6572
Control	0.6275

Figure 13.8: `Kneitel_2010_algae_tukey2.sas - proc glm`

We will now calculate the value of HSD for this example, to show how it is used to construct confidence intervals and tests. As before, we have $MS_{within} = 0.1122$ with 20 degrees of freedom. The SAS output gives the value of $q_{0.05,5,20} = 4.2319$, and there are $n = 5$ replicates per treatment. We then have

$$HSD = q_{\alpha,a,a(n-1)}\sqrt{\frac{MS_{within}}{n}} = 4.2319\sqrt{\frac{(0.1122)}{5}} = 0.6339. \tag{13.19}$$

This value agrees with the SAS output labeled `Minimum Significant Difference`. We now calculate a 95% confidence interval for $\mu_{\text{medium}} - \mu_{\text{control}}$. The formula for the confidence interval is

$$\left(\bar{Y}_i - \bar{Y}_j - HSD, \bar{Y}_i - \bar{Y}_j + HSD\right). \tag{13.20}$$

Inserting the estimated means for these two treatments (see SAS output) in this formula, and the HSD value, we obtain

$$(1.1972 - 0.6275 - 0.6339, 1.1972 - 0.6275 + 0.6339). \tag{13.21}$$

or $(-0.0642, 1.2036)$. This confidence interval is close to the value provided by SAS. Now suppose we want to test $H_0 : \mu_{\text{medium}} - \mu_{\text{control}} = 0$ using our HSD value. We would reject H_0 if $|\bar{Y}_i - \bar{Y}_j| \geq HSD$. Inserting the means for these treatments, we find that $|1.1972 - 0.6275| = 0.5687 < 0.6339$, and so this pair of means was not significantly different.

How does this procedure control the EER as well as provide simultaneous confidence intervals? **The Tukey procedure basically controls the EER by making each pairwise test more conservative, through the use of the studentized range distribution.** Notice that $HSD > LSD$ for the same data set (0.6339 vs. 0.4419). This means that the Tukey procedure requires a larger difference between groups before declaring they are significantly different, and the confidence intervals are also broader. As a consequence, there is lower power to detect differences among groups when they do exist. This is the price paid for controlling the EER.

13.3.5 Multiple range tests - REGW

The multiple comparison procedures we have examined so far yield both tests and confidence intervals. Another type of multiple comparison procedure are multiple range tests. These procedures provide only tests, but are also more powerful procedures because they essentially conduct fewer overall tests than the methods we studied earlier. There are a number of different multiple range tests, but we will only examine the REGW (Ryan-Einot-Gabriel-Welsch) procedure because it controls the EER (Hsu 1996).

The test works as follows (Hsu 1996). Suppose we order the sample means of the a different groups from smallest to largest:

$$\bar{Y}_{[1]} \leq \bar{Y}_{[2]} \leq \ldots \bar{Y}_{[a-1]}, \leq \bar{Y}_{[a]} \tag{13.22}$$

where $\bar{Y}_{[1]}$ is the smallest and $\bar{Y}_{[a]}$ the largest sample mean.

We then examine the range (difference) between the largest and smallest sample mean, namely $\bar{Y}_{[a]} - \bar{Y}_{[1]}$. If

$$\bar{Y}_{[a]} - \bar{Y}_{[1]} < q_a \sqrt{\frac{MS_{within}}{n}} \tag{13.23}$$

then we stop and declare there are no significant differences among groups. Otherwise, we assert that these two groups are significantly different and continue the process. We next examine the next innermost ranges $\bar{Y}_{[a-1]} - \bar{Y}_{[1]}$ and $\bar{Y}_{[a]} - \bar{Y}_{[2]}$. If

$$\bar{Y}_{[a-1]} - \bar{Y}_{[1]]} < q_{a-1} \sqrt{\frac{MS_{within}}{n}} \tag{13.24}$$

and

$$\bar{Y}_{[a1]} - \bar{Y}_{[2]]} < q_{a-1} \sqrt{\frac{MS_{within}}{n}} \tag{13.25}$$

then we stop the testing process. Otherwise, we assert that one or both groups are significantly different. This process is continued until no more significant differences are found.

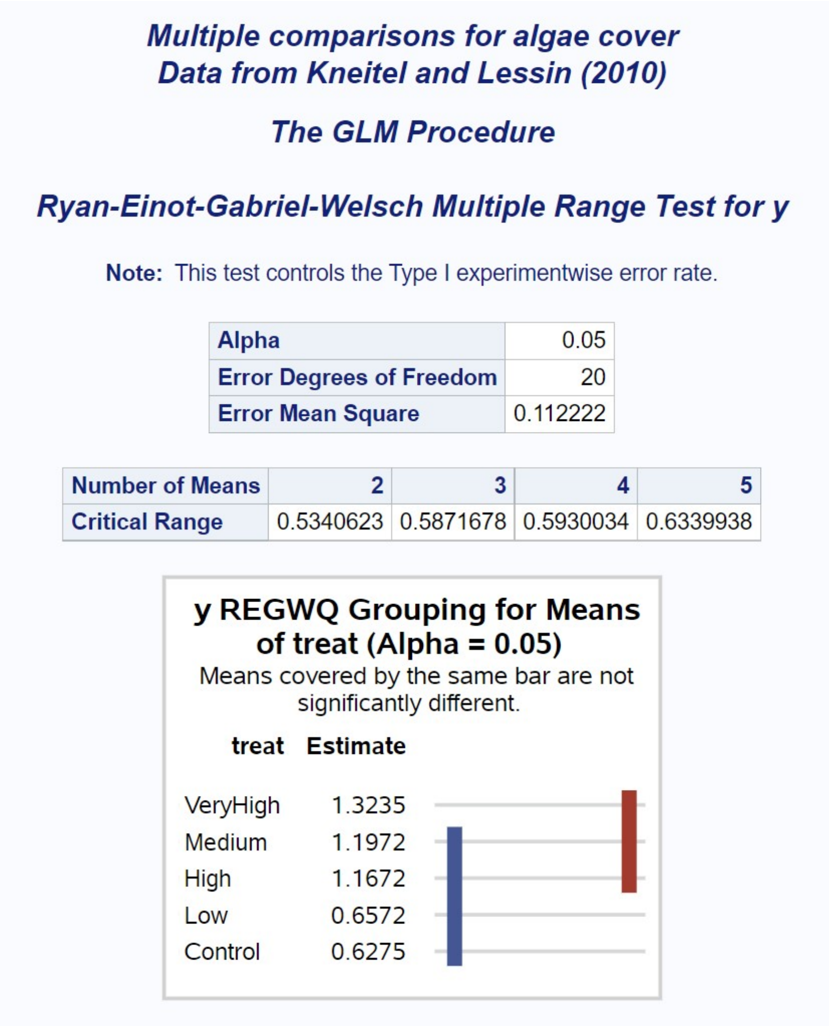
Note that the values of q are not the same for every step of the test. They are constructed so that $q_a > q_{a-1} > \ldots > q_2$, meaning that the largest range is tested using q_a, the next two ranges with q_{a-1} (a smaller value), and so forth. This implies that the largest range must have the largest difference in means to be judged significant, while later tests allow for smaller differences. The values of q are chosen so that the experimentwise error rate has a specified value, usually $\alpha = 0.05$ (Hsu 1996). The studentized range distribution is involved in this process. The value of q_a used in the first step of the procedure is the same as that used by the Tukey procedure, as well as the difference in the means judged to be significant. The two procedures diverge after this point.

13.3.6 REGW procedure - SAS demo

We can use the REGW procedure by adding the `regwq` option to the `means` statement, as follows

```
means treat / regwq;
```

SAS then generates a diagram indicating which groups are significantly different (Fig. 13.9). For this example, the REGW procedure gives the same pattern of significant differences among groups as the Tukey method. The REGW procedure may become liberal (not fully control the EER) when the data are unbalanced, and SAS prints a warning note in this situation.



Multiple comparisons for algae cover
Data from Kneitel and Lessin (2010)

The GLM Procedure

Ryan-Einot-Gabriel-Welsch Multiple Range Test for y

Note: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	20
Error Mean Square	0.112222

Number of Means	**2**	**3**	**4**	**5**
Critical Range	0.5340623	0.5871678	0.5930034	0.6339938

y REGWQ Grouping for Means of treat (Alpha = 0.05)

Means covered by the same bar are not significantly different.

treat	**Estimate**
VeryHigh	1.3235
Medium	1.1972
High	1.1672
Low	0.6572
Control	0.6275

Figure 13.9: Kneitel_2010_algae_regw2.sas - proc glm

13.4 Comparisons with a control - Dunnett procedure

Many studies include some sort of control group or treatment, and the experimenter may only be interested in comparing the control group with each of the other $a-1$ groups. For example, the control could represent a standard medical treatment for a disease while the other treatments represent alternative forms of therapy. The physician only wants to know if the alternative forms are better or worse than the standard method.

In this situation, there are only $a-1$ comparisons to be made rather than the full $a(a-1)/2$ comparisons of all pairs of means. The Dunnett procedure is designed to control the EER for just these $a-1$ comparisons, and hence has more power than other pairwise methods (Hsu 1996). The calculations are similar to the Tukey method, but use the quantity

$$DSD = d_{\alpha,a,a(n-1)}\sqrt{\frac{2MS_{within}}{n}}, \tag{13.26}$$

where DSD stands for Dunnett's significant difference. The values of $d_{\alpha,a,a(n-1)}$ are obtained from a distribution analogous to the studentized range distribution, except that it controls the EER for $a-1$ comparisons. The value of d depends on α (the desired EER), the number of groups a, and the degrees of freedom for MS_{within}.

Let μ_c be the mean of the control group, while μ_i is any other group. Dunnett's procedure can be used to test for $H_0 : \mu_i = \mu_c$ or equivalently $H_0 : \mu_i - \mu_c = 0$. We would accept H_0 if $|\bar{Y}_i - \bar{Y}_c| < DSD$. Conversely, we would reject H_0 if $|\bar{Y}_i - \bar{Y}_c| \geq DSD$. This same rule applies to all comparisons with the control group.

Confidence intervals for $\mu_i - \mu_c$ have the form

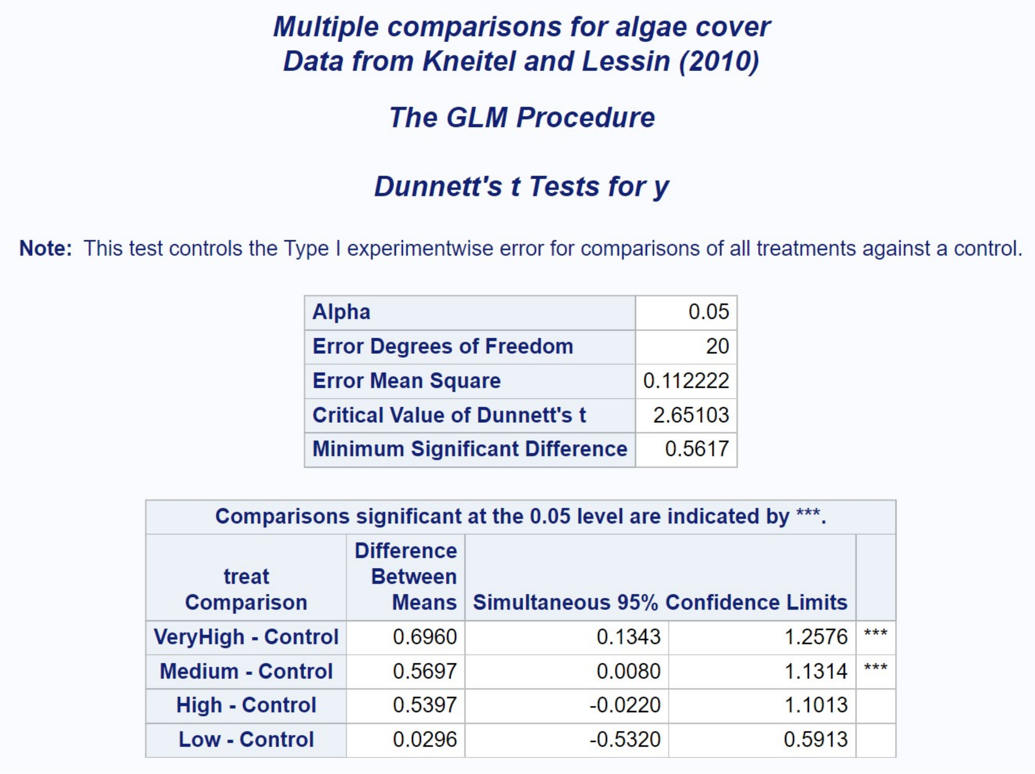
$$\left(\bar{Y}_i - \bar{Y}_c - DSD, \bar{Y}_i - \bar{Y}_c + DSD\right). \tag{13.27}$$

13.4.1 Dunnett's procedure - SAS demo

Dunnett's procedure is invoked using the `dunnett` option in the `means` statement, with the control group specified in parentheses. For this data set, the control group is coded as `Control`, and so we have

```
means treat / dunnett('Control');
```

Confidence intervals for $\mu_i - \mu_c$ are given in the SAS output, with the symbol `***` indicating which comparisons of the control are significantly different (Fig. 13.10). We see that the `VeryHigh` and `Medium` treatments are significantly different from `Control`.



Multiple comparisons for algae cover
Data from Kneitel and Lessin (2010)

The GLM Procedure

Dunnett's t Tests for y

Note: This test controls the Type I experimentwise error for comparisons of all treatments against a control.

Alpha	0.05
Error Degrees of Freedom	20
Error Mean Square	0.112222
Critical Value of Dunnett's t	2.65103
Minimum Significant Difference	0.5617

Comparisons significant at the 0.05 level are indicated by *.**

treat Comparison	**Difference Between Means**	**Simultaneous 95% Confidence Limits**		
VeryHigh - Control	0.6960	0.1343	1.2576	***
Medium - Control	0.5697	0.0080	1.1314	***
High - Control	0.5397	-0.0220	1.1013	
Low - Control	0.0296	-0.5320	0.5913	

Figure 13.10: `Kneitel_2010_algae_dunnett2.sas - proc glm`

13.5 Bonferroni and Sidak corrections

One way of controlling the EER in a set of comparisons is to use a distribution designed to control it, such as the studentized range distribution. These procedures control the EER by essentially making the per comparison rate for each test more conservative. This adjustment of the per comparison error rate is built into the studentized range distribution.

The Bonferroni correction provides another way of controlling the EER, by explicitly reducing the per comparison error rate and then using a simple t test (like the least significant difference procedure) to compare group means. Suppose that we are interested in k possible comparisons, either all $a(a-1)/2$ pairwise comparisons or $a-1$ comparisons with a control, where a is the number of groups. The Bonferroni correction adjusts the per comparison error rate as follows. Let α be the per comparison error rate, while α' is the desired EER. If we conduct each comparison at the per comparison rate of

$$\alpha = \frac{\alpha'}{k}, \tag{13.28}$$

then it can be shown the EER will not exceed α' (Hsu 1996). For example, suppose we are interested in all $k = a(a-1)/2$ pairwise comparison among groups. We would then conduct each test at the

$$\alpha = \frac{\alpha'}{k} = \frac{\alpha'}{a(a-1)/2} \tag{13.29}$$

level. We would use the same t test as in the least significant difference procedure, but adjust the value α according to this formula. We then have

$$BSD = c_{\frac{\alpha'}{a(a-1)/2}, a(n-1)} \sqrt{\frac{2MS_{within}}{n}}, \tag{13.30}$$

where BSD is the difference judged to be significant given the Bonferroni correction. We would accept $H_0 : \mu_i = \mu_j$ (or $H_0 : \mu_i - \mu_j = 0$) if $\bar{Y}_i - \bar{Y}_j$ falls inside the interval $(-BSD, BSD)$, or equivalently if $|\bar{Y}_i - \bar{Y}_j| < BSD$. Conversely, we would reject H_0 if $|\bar{Y}_i - \bar{Y}_j| \geq BSD$. A confidence interval for $\mu_i - \mu_j$ based on the Bonferroni correction would have the form

$$\left(\bar{Y}_i - \bar{Y}_j - BSD, \bar{Y}_i - \bar{Y}_j + BSD\right). \tag{13.31}$$

To make things more concrete, we can calculate the value of BSD for the algae cover example (Kneitel & Lessin 2010). From our previous output, we have $a = 5$ groups, $n = 5$ replicates per group, and $MS_{within} = 0.1122$. If we set the EER to be $\alpha' = 0.05$, by the above formula we have

$$\alpha = \frac{\alpha'}{a(a-1)/2} = \frac{0.05}{5(5-1)/2} = \frac{0.05}{10} = 0.005. \tag{13.32}$$

For $\alpha = 0.005$, we have $c_{0.005,20} = 3.1534$, and so

$$BSD = c_{\frac{\alpha'}{a(a-1)/2},a(n-1)}\sqrt{\frac{2MS_{within}}{n}} = 3.1534\sqrt{\frac{2(0.1122)}{5}} = 0.6681. \tag{13.33}$$

Note that the value of $BSD = 0.6681$ is larger than $HSD = 0.6339$ value for the Tukey procedure. Thus, the Bonferroni method requires a greater difference among means before declaring they are significantly different, implying it has lower power than the Tukey procedure. It would also generate larger confidence intervals and so provides less precision in estimation.

Given these drawbacks, why would the Bonferroni correction be used? The Bonferroni procedure is quite general and can be used to control the EER for other testing procedures, not just comparisons among means in ANOVA. For example, it is common to have a collection of statistical tests that address a particular question. We might have a single experiment in which a number of different Y variables are measured, with a separate ANOVA conducted on each variable. If enough variables are examined it is possible that some could be significant by chance, and we could control the EER for all these tests using the Bonferroni correction, with k being the number of Y variables. There is also a version of this procedure similar in spirit to REGW, called the **sequential Bonferroni method** (Rice 1989). The sequential Bonferroni alleviates to some extent the lack of power in the standard Bonferroni correction. This procedure is implemented in `proc multtest` in SAS.

The Sidak correction is another procedure used to control the EER, which provides slightly more power than the Bonferroni method. Let α be the per comparison error rate, while α' is the desired EER. If we conduct each comparison at the per comparison rate of

$$\alpha = 1 - (1 - \alpha')^{1/k}, \tag{13.34}$$

then the actual EER will not exceed α'. For example, suppose we are interested in all $k = a(a-1)/2$ pairwise comparison among groups. We would then conduct each test at the

$$\alpha = 1 - (1 - \alpha')^{1/k} = 1 - (1 - \alpha')^{1/[a(a-1)/2]} \tag{13.35}$$

level. For $\alpha' = 0.05$ and $a = 5$ groups, we obtain

$$\alpha = 1 - (1 - \alpha')^{1/[a(a-1)/2]} = 1 - (1 - 0.05)^{1/10} = 0.0051. \tag{13.36}$$

We would then compare pairs of means using the same test as for the Bonferroni correction, except that we would use $\alpha = 0.0051$ rather than $\alpha = 0.005$. This value of α is a bit larger than the corresponding Bonferroni one, making the Sidak correction slightly more powerful.

SAS implements both the Bonferroni and Sidak corrections in the `means` statement with the options `bon` or `sidak`, similar to using the `tukey` option.

13.6 Vascular plant cover - SAS demo

Kneitel & Lessin (2010) also examined vascular plant cover in their study of the effect of eutrophication on vernal pools in California. Vascular plant cover (`cover`) was derived by subtracting algal cover (`algae`) from total cover (`total`), then arcsine-square root transformed before analysis (see Chapter 15). See `data` step in the SAS program below.

The `proc glm` code compares all possible pairs of group means using the Tukey procedure, and also compares the `Control` treatment with the other treaments using Dunnett's procedure. This was done to provide more examples of these procedures. **In practice, you should choose one procedure for comparing the means.**

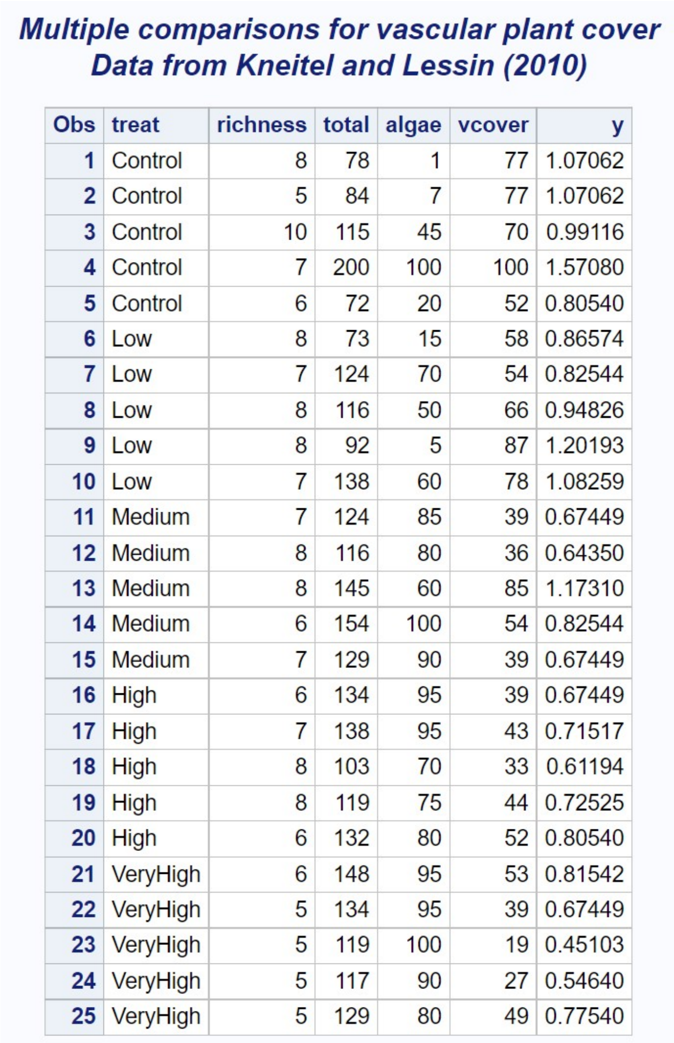
The diagram generated by the Tukey procedure indicates two significant differences among treatments (Fig. 13.15). Reading the diagram, we see the `Control-High` and `Control-VeryHigh` comparisons were significant, because they have different lines. No other pairs of groups were significantly different. Figure 13.12 indicates how these results could be graphically displayed using letters instead of lines. We see that vascular plant cover actually decreased with increased nutrient levels, likely due to inhibition from the algal mats that form (Kneitel and Lessin 2010).

If the lines diagram is confusing, we can also determine which groups are significantly different by examining the confidence intervals generated by the Tukey procedure (Fig. 13.14). Confidence intervals that do not include zero indicate a significant difference among groups, because of the duality between confidence intervals and tests (see Chapter 10). The significant tests are indicated by `***` in the SAS output.

The output for Dunnett’s procedure shows that the `High` and `VeryHigh` treatments were significantly different from the `Control` group (Fig. 13.16).

SAS Program

```
* Kneitel_2010_cover2.sas;
title ’Multiple comparisons for vascular plant cover’;
title2 ’Data from Kneitel and Lessin (2010)’;
data kneitel;
    input treat $ richness total algae;
    * Apply transformations here;
    vcover = total-algae;
    y = arsin(sqrt(vcover/100));
    datalines;
Control   8   78    1
Control   5   84    7
Control  10  115   45
Control   7  200  100
Control   6   72   20
Low       8   73   15
Low       7  124   70
Low       8  116   50
Low       8   92    5
Low       7  138   60
Medium    7  124   85
Medium    8  116   80
Medium    8  145   60
Medium    6  154  100
Medium    7  129   90
High      6  134   95
High      7  138   95
High      8  103   70
High      8  119   75
High      6  132   80
VeryHigh  6  148   95
VeryHigh  5  134   95
VeryHigh  5  119  100
VeryHigh  5  117   90
VeryHigh  5  129   80
;
run;
* Print data set;
proc print data=kneitel;
* Plot means, standard errors, and observations;
proc gplot data=kneitel;
    plot y*treat=1 / vaxis=axis1 haxis=axis1;
    symbol1 i=std1mjt v=star height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* One-way anova with comparisons;
proc glm order=data plots=diagnostics data=kneitel;
    class treat;
    model y = treat;
    * Tukey procedure - controls the EER;
    means treat / tukey cldiff lines;
    * Dunnett’s procedure - controls EER for comparisons with a control;
    means treat / dunnett(’Control’);
run;
quit;
```



Multiple comparisons for vascular plant cover
Data from Kneitel and Lessin (2010)

Obs	treat	richness	total	algae	vcover	y
1	Control	8	78	1	77	1.07062
2	Control	5	84	7	77	1.07062
3	Control	10	115	45	70	0.99116
4	Control	7	200	100	100	1.57080
5	Control	6	72	20	52	0.80540
6	Low	8	73	15	58	0.86574
7	Low	7	124	70	54	0.82544
8	Low	8	116	50	66	0.94826
9	Low	8	92	5	87	1.20193
10	Low	7	138	60	78	1.08259
11	Medium	7	124	85	39	0.67449
12	Medium	8	116	80	36	0.64350
13	Medium	8	145	60	85	1.17310
14	Medium	6	154	100	54	0.82544
15	Medium	7	129	90	39	0.67449
16	High	6	134	95	39	0.67449
17	High	7	138	95	43	0.71517
18	High	8	103	70	33	0.61194
19	High	8	119	75	44	0.72525
20	High	6	132	80	52	0.80540
21	VeryHigh	6	148	95	53	0.81542
22	VeryHigh	5	134	95	39	0.67449
23	VeryHigh	5	119	100	19	0.45103
24	VeryHigh	5	117	90	27	0.54640
25	VeryHigh	5	129	80	49	0.77540

Figure 13.11: `Kneitel_2010_cover2.sas - proc print`

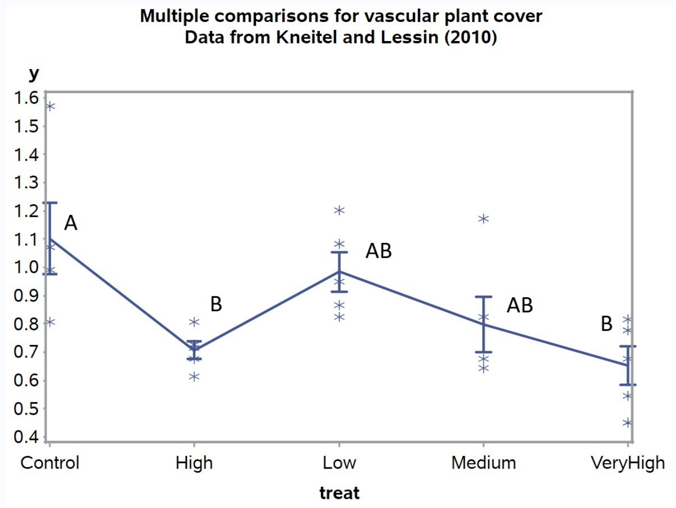
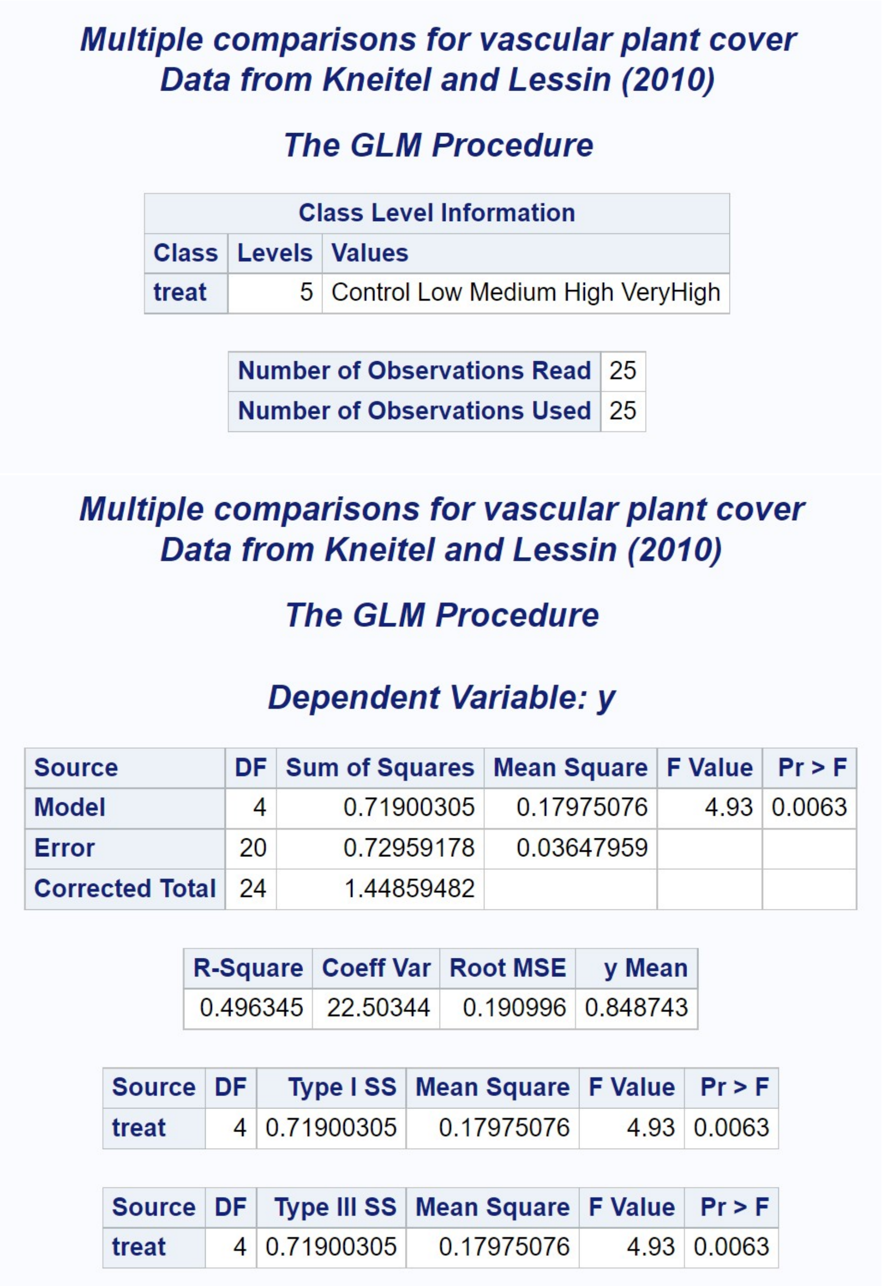


Figure 13.12: `Kneitel_2010_cover2.sas - proc gplot`



Multiple comparisons for vascular plant cover
Data from Kneitel and Lessin (2010)

The GLM Procedure

Class Level Information		
Class	**Levels**	**Values**
treat	5	Control Low Medium High VeryHigh

Number of Observations Read	25
Number of Observations Used	25

Multiple comparisons for vascular plant cover
Data from Kneitel and Lessin (2010)

The GLM Procedure

Dependent Variable: y

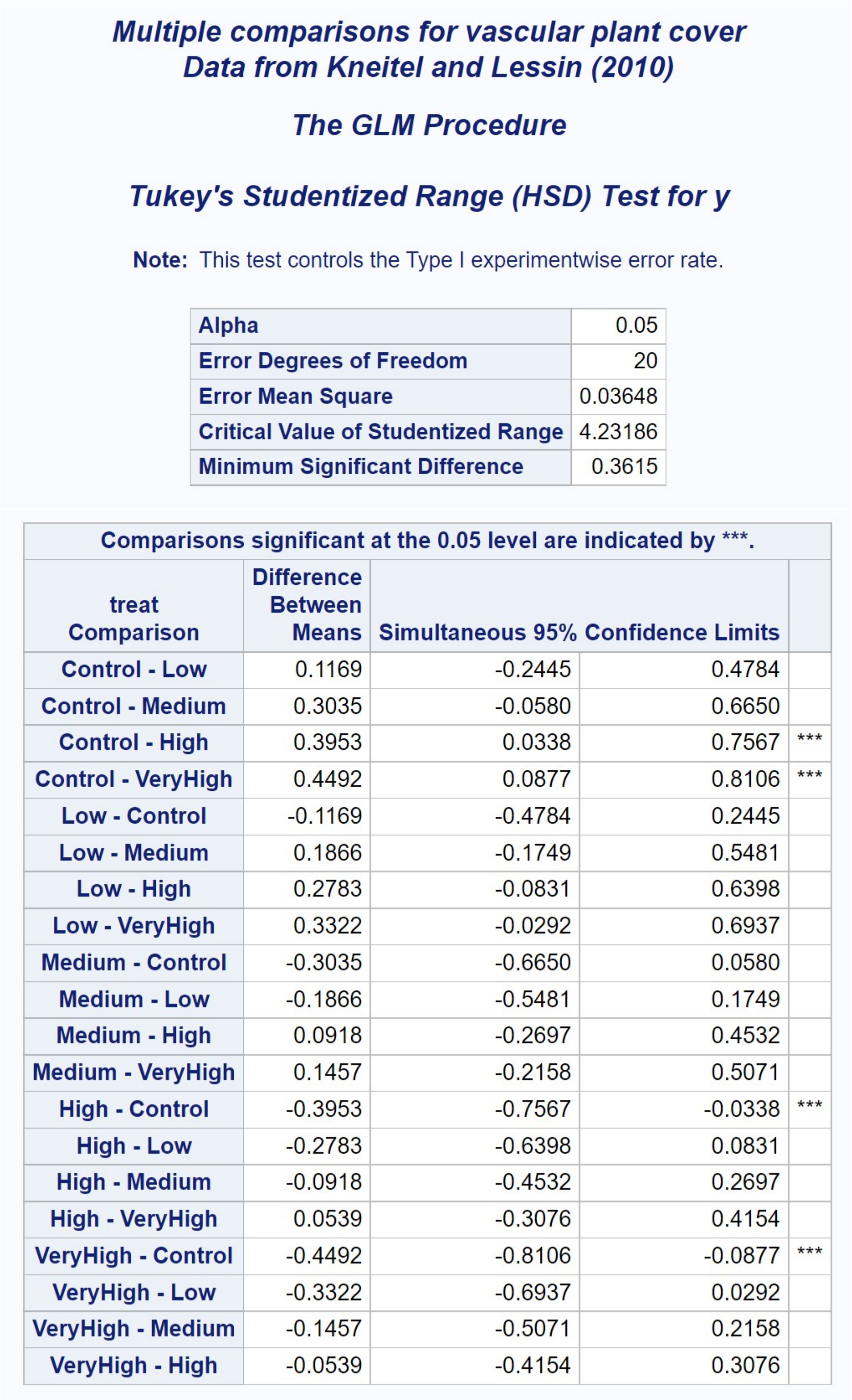
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	0.71900305	0.17975076	4.93	0.0063
Error	20	0.72959178	0.03647959		
Corrected Total	24	1.44859482			

R-Square	Coeff Var	Root MSE	y Mean
0.496345	22.50344	0.190996	0.848743

Source	DF	Type I SS	Mean Square	F Value	Pr > F
treat	4	0.71900305	0.17975076	4.93	0.0063

Source	DF	Type III SS	Mean Square	F Value	Pr > F
treat	4	0.71900305	0.17975076	4.93	0.0063

Figure 13.13: `Kneitel_2010_cover2.sas - proc glm`



Multiple comparisons for vascular plant cover
Data from Kneitel and Lessin (2010)

The GLM Procedure

Tukey's Studentized Range (HSD) Test for y

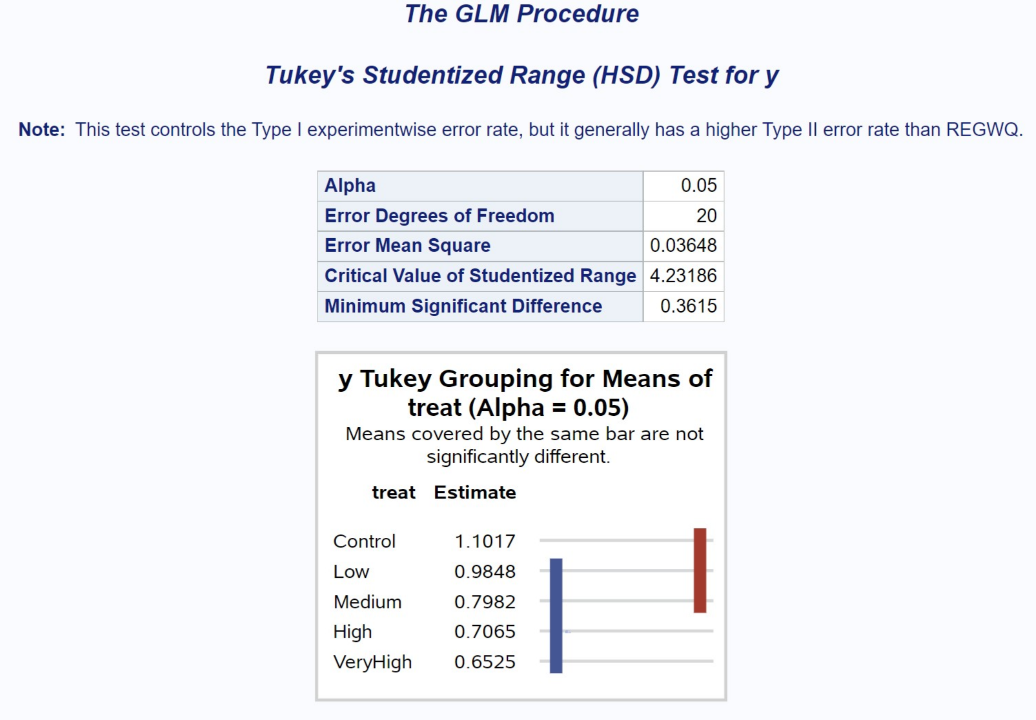
Note: This test controls the Type I experimentwise error rate.

Alpha	0.05
Error Degrees of Freedom	20
Error Mean Square	0.03648
Critical Value of Studentized Range	4.23186
Minimum Significant Difference	0.3615

Comparisons significant at the 0.05 level are indicated by ***.

treat Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
Control - Low	0.1169	-0.2445	0.4784	
Control - Medium	0.3035	-0.0580	0.6650	
Control - High	0.3953	0.0338	0.7567	***
Control - VeryHigh	0.4492	0.0877	0.8106	***
Low - Control	-0.1169	-0.4784	0.2445	
Low - Medium	0.1866	-0.1749	0.5481	
Low - High	0.2783	-0.0831	0.6398	
Low - VeryHigh	0.3322	-0.0292	0.6937	
Medium - Control	-0.3035	-0.6650	0.0580	
Medium - Low	-0.1866	-0.5481	0.1749	
Medium - High	0.0918	-0.2697	0.4532	
Medium - VeryHigh	0.1457	-0.2158	0.5071	
High - Control	-0.3953	-0.7567	-0.0338	***
High - Low	-0.2783	-0.6398	0.0831	
High - Medium	-0.0918	-0.4532	0.2697	
High - VeryHigh	0.0539	-0.3076	0.4154	
VeryHigh - Control	-0.4492	-0.8106	-0.0877	***
VeryHigh - Low	-0.3322	-0.6937	0.0292	
VeryHigh - Medium	-0.1457	-0.5071	0.2158	
VeryHigh - High	-0.0539	-0.4154	0.3076	

Figure 13.14: `Kneitel_2010_cover2.sas - proc glm`



The GLM Procedure

Tukey's Studentized Range (HSD) Test for y

Note: This test controls the Type I experimentwise error rate, but it generally has a higher Type II error rate than REGWQ.

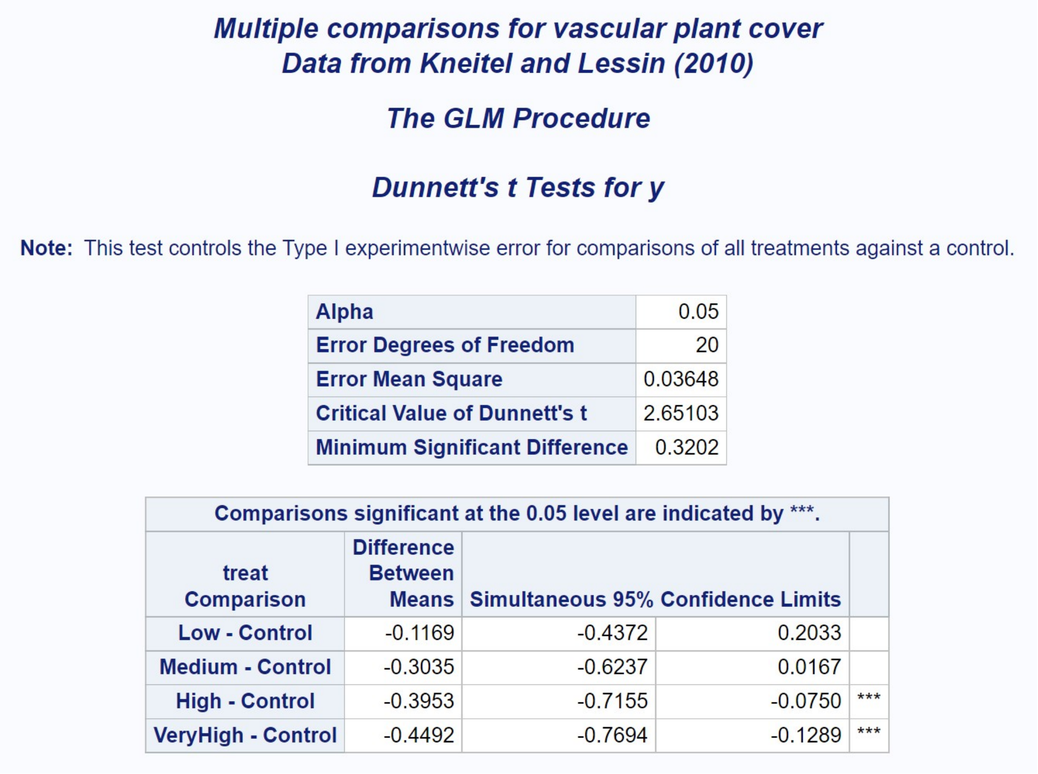
Alpha	0.05
Error Degrees of Freedom	20
Error Mean Square	0.03648
Critical Value of Studentized Range	4.23186
Minimum Significant Difference	0.3615

y Tukey Grouping for Means of treat (Alpha = 0.05)

Means covered by the same bar are not significantly different.

treat	Estimate
Control	1.1017
Low	0.9848
Medium	0.7982
High	0.7065
VeryHigh	0.6525

Figure 13.15: `Kneitel_2010_cover2.sas - proc glm`



Multiple comparisons for vascular plant cover
Data from Kneitel and Lessin (2010)

The GLM Procedure

Dunnett's t Tests for y

Note: This test controls the Type I experimentwise error for comparisons of all treatments against a control.

Alpha	0.05
Error Degrees of Freedom	20
Error Mean Square	0.03648
Critical Value of Dunnett's t	2.65103
Minimum Significant Difference	0.3202

Comparisons significant at the 0.05 level are indicated by *.**

treat Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
Low - Control	-0.1169	-0.4372	0.2033	
Medium - Control	-0.3035	-0.6237	0.0167	
High - Control	-0.3953	-0.7155	-0.0750	***
VeryHigh - Control	-0.4492	-0.7694	-0.1289	***

Figure 13.16: `Kneitel_2010_cover2.sas - proc glm`

13.7 False discovery rate procedure

The multiple comparison procedures we have examined control the EER, but at the cost of power. This is especially true for studies with many treatments or groups. For example, suppose we have $a = 5$ treatments and want to conduct all pairwise comparisons using the Bonferroni method, with an EER of $\alpha' = 0.05$. There are $k = a(a-1)/2 = 5(4)/2 = 10$ pairwise comparisons, and so we would conduct each comparison at the $\alpha = \alpha'/k = 0.05/10 = 0.005$ level. For $a = 10$ treatments, a similar calculation suggests that each comparison should be conducted at the $\alpha = 0.0011$ level, yielding a much more conservative test. As the number of treatments increases, this makes it less likely significant differences will be found, and so the power to detect differences among treatments decreases. The number of treatments has similar effects on other multiple comparison procedures that control the EER.

The **false discovery rate** or **FDR** method provides an alternative approach to multiple comparisons and tests. For example, suppose that the LSD method finds a number of significant comparisons among treatments, which are termed discoveries. The LSD method only controls the per comparison error rate, and if there are many comparisons some of the significant ones could be Type I errors. The FDR method controls the **expected proportion** of these significant comparisons that are Type I errors, and therefore are false discoveries (Benjamini & Hochberg 1995). This differs substantially from methods that control the EER, which are concerned with keeping the **number** of Type I errors low. One will have more Type I errors using the FDR, but the proportion of them is controlled, and the power to detect differences among treatments will be higher than EER methods. This approach seems particularly useful for studies that screen many treatments or groups, possibly for future work, and it is more important to identify possible effects than controlling the number of Type I errors (Verhoeven et al. 2005).

The FDR method for multiple comparisons works as follows (Benjamini & Hochberg 1995). Suppose you have k pairwise comparisons, and obtain a P value for each one using the LSD procedure. Let $P_{[1]} \leq P_{[2]} \leq \ldots \leq P_{[k]}$ be the P values for these tests, ordered from smallest to largest, with $P_{[i]}$ the *ith* one. Let α^* be the specified false discovery rate. We then examine the ordered P values from largest to smallest (from $i = k$ to 1), examining at each step whether

$$P_{[i]} \leq \frac{i}{k}\alpha^*. \tag{13.37}$$

We can see that the right side of this equation decreases from α^* to α^*/k as i decreases. The first time this inequality is true, we declare that this pairwise comparison and all further ones are significantly different. Benjamini & Hochberg (1995) show that this procedure controls the false discovery rate. The same method can also be used in other multiple testing scenarios, not just multiple comparisons among means.

As an example of this procedure, consider the algae cover example we examined earlier (Kneitel and Lessin 2010). There are ten pairwise comparisons among the different nutrient treatments. We first obtain the P values for each comparison using the LSD method (see SAS demo below), and order these from largest to smallest (Table 13.1). We then compare the P values with the right side of Eq. 13.37, beginning at the top of the table. We see that first comparison that satisfies Eq. 13.37 is high vs. low, and so we declare this comparison and all further ones to be significant. Thus, the FDR procedure found six of ten pairwise comparisons to be significant, similar to the LSD procedure. The Tukey and REGW procedures, which control the EER, found only two significant comparisons.

Table 13.1: Ordered P values for LSD comparisons of algae cover in different nutrient treatments (Kneitel and Lessin 2010). The last column calculates the right side of Eq. 13.37 for $\alpha^* = 0.05$ and $k = 10$ pairwise comparisons.

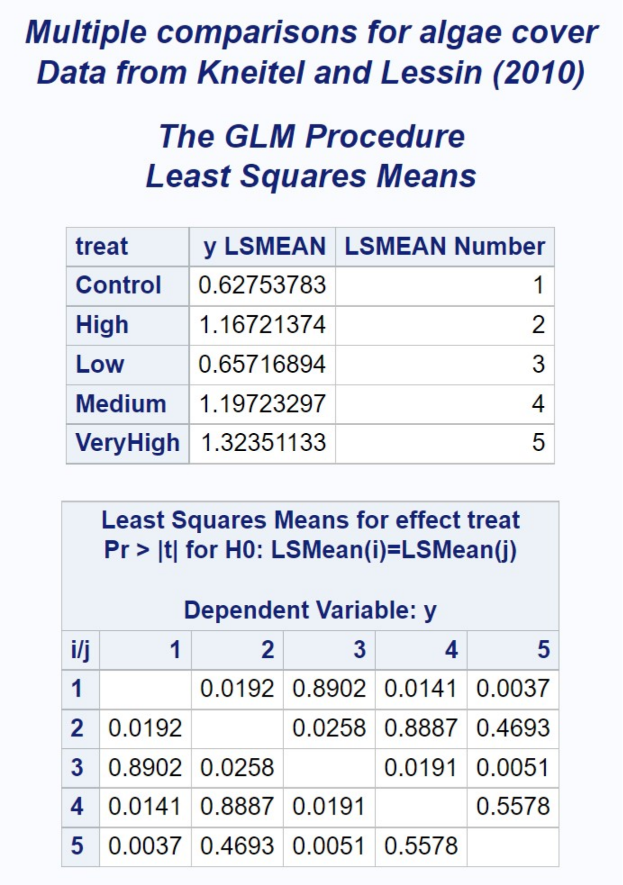
Comparison	i	$P_{[i]}$	$\frac{i}{k}\alpha^*$
control–low	10	0.8902	0.0500
medium–high	9	0.8887	0.0450
medium–very high	8	0.5578	0.0400
high–very high	7	0.4693	0.0350
high–low	6	0.0258	0.0300
control–high	5	0.0192	0.0250
low–medium	4	0.0191	0.0200
control–medium	3	0.0141	0.0150
low–very high	2	0.0051	0.0100
control–very high	1	0.0037	0.0010

13.7.1 False discovery rate - SAS demo

The FDR procedure can be implemented in two steps using SAS. We first need to obtain the P values for the LSD procedure. This can be accomplished by adding an `lsmeans` statement to our previous program, with a `pdiff` option:

```
lsmeans treat / adjust=t pdiff;
```

The result is a table of P values for each comparison (Fig. 13.17).



Multiple comparisons for algae cover
Data from Kneitel and Lessin (2010)

The GLM Procedure
Least Squares Means

treat	y LSMEAN	LSMEAN Number
Control	0.62753783	1
High	1.16721374	2
Low	0.65716894	3
Medium	1.19723297	4
VeryHigh	1.32351133	5

Least Squares Means for effect treat
Pr > \|t\| for H0: LSMean(i)=LSMean(j)

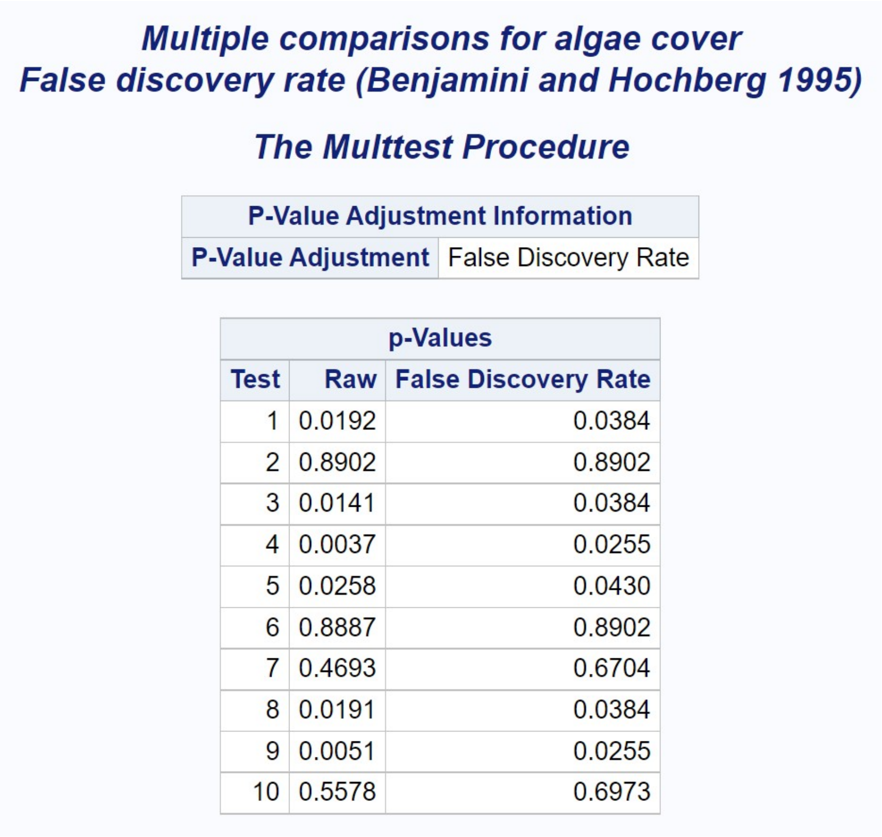
Dependent Variable: y

i/j	1	2	3	4	5
1		0.0192	0.8902	0.0141	0.0037
2	0.0192		0.0258	0.8887	0.4693
3	0.8902	0.0258		0.0191	0.0051
4	0.0141	0.8887	0.0191		0.5578
5	0.0037	0.4693	0.0051	0.5578	

Figure 13.17: `Kneitel_2010_algae_fdr1.sas - proc glm`

We then use `proc multtest` to carry out the FDR procedure. The P values for each comparison are supplied in a SAS data set, labeled as `raw_p`. The data set is specified using the `inpvalues` option, while the FDR procedure is requested using the `fdr` option. The output consists of the original and adjusted P values, with the adjustment made according to the FDR procedure. For an FDR of 0.05, adjusted P values less than 0.05 are judged to be significant. See complete program below. We observe that six of ten pairwise comparisons have an adjusted P value less than 0.05, and so were significant by the FDR procedure (Fig. 13.18).

```
* Kneitel_2010_algae_fdr2.sas;
title ’Multiple comparisons for algae cover’;
title2 ’False discovery rate (Benjamini and Hochberg 1995)’;
data pvalues;
    input comparison :$18. raw_p;
    datalines;
Control-High     0.0192
Control-Low      0.8902
Control-Medium   0.0141
Control-VeryHigh 0.0037
High-Low         0.0258
High-Medium      0.8887
High-VeryHigh    0.4693
Low-Medium       0.0191
Low-VeryHigh     0.0051
Medium-VeryHigh  0.5578
;
* Multiple comparisons using fdr;
proc multtest inpvalues=pvalues fdr;
run;
quit;
```



Multiple comparisons for algae cover
False discovery rate (Benjamini and Hochberg 1995)

The Multtest Procedure

P-Value Adjustment Information	
P-Value Adjustment	False Discovery Rate

p-Values		
Test	**Raw**	**False Discovery Rate**
1	0.0192	0.0384
2	0.8902	0.8902
3	0.0141	0.0384
4	0.0037	0.0255
5	0.0258	0.0430
6	0.8887	0.8902
7	0.4693	0.6704
8	0.0191	0.0384
9	0.0051	0.0255
10	0.5578	0.6973

Figure 13.18: `Kneitel_2010_algae_fdr2.sas - proc multcomp`

13.8 References

Benjamini, Y. & Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57: 289-300.

Day, R. W. & Quinn, G. P. (1989) Comparisons of treatments after an analysis of variance in ecology. *Ecological Monographs* 59: 433-463.

Hsu, J. C. (1996) *Multiple Comparisons: Theory and Methods.* Chapman & Hall/CRC Press, Boca Raton, FL.

Kneitel, J. M. & Lessin, C. L. (2010) Ecosystem-phase interactions: aquatic eutrophication decreases terrestrial plant diversity in California vernal pools. *Oecologia* 163: 461-469.

Kohler, C. K, Heidinger, R. C. & Call, T. (1990) Levels of PCBs and trace metal in Crab Orchard Lake sediment, benthos, zooplankton and fish. Waste Management and Research Center Report RR-E43, Illinois Department of Natural Resources.

Rice, W. R. (1989) Analyzing tables of statistical tests. *Evolution* 43: 223-225.

SAS Institute Inc. (2018) *SAS/STAT 15.1 Users Guide* SAS Institute Inc., Cary, NC

Verhoeven, K. J. F., Simonsen, K. L. & McIntyre, L. M. (2005) Implementing false discovery rate control: increasing your power. *Oikos* 108: 643-647.

Westfall, P. H., Tobias, R. D., Rom, D., Wolfinger, R. D. & Hochberg, Y. (1999) *Multiple Comparisons and Multiple Tests Using the SAS System.* SAS Institute Inc., Cary, NC.

13.9 Problems

1. White-tailed deer are voracious consumers of landscaping plants. A frustrated homeowner/professor is interested in testing whether different repellents actually reduce deer herbivory. Replicate plots of houseplants are established and four different treatments applied to the plots: (1) a control with no treatment, (2) hot pepper oil repellent, (3) rotten egg repellent, and (4) livestock blood repellent. There were 4 replicate plots per treatment. The amount of herbivory (percentage of plants eaten) after one month are given in the following table.

Control	Hot pepper	Rotten eggs	Blood
61.1	54.4	32.0	36.2
64.9	67.9	28.5	38.3
61.6	54.6	21.6	31.1
67.8	58.1	38.8	44.1

   (a) Test whether there is an overall effect of treatment on the percentage of plants eaten, using one-way anova and SAS. Report your results using P values and discuss the significance of the test.

   (b) Use the Tukey procedure to compare the different treatments, and interpret your results. Which pairs of treatments are significantly different? Do the treatments fall into particular groups?

   (c) Suppose the homeowner is only interested in treatments that are different from the control. Use the Dunnett method to compare the three treatments with the control one. Which treatments are significantly different from the control?

2. PCB concentrations were measured in the sediment of Crab Orchard Lake, at 11 different sites (Kohler et al. 1990). Three samples were taken at each site, yielding the data shown in the table below. Site 10 is near an abandoned dump site for a manufacturer of electrical transformers.

Site	PCB (mg/kg), sample 1-3
1	0.0453, 0.0626, 0.527
2	0.0395, 0.0494, 0.0416
3	0.0234, 0.0451, 0.0541
4	0.033, 0.0643, 0.0517
5	0.0394, 0.0810, 0.0266
6	0.0294, 0.0425, 0.0538
7	0.0255, 0.0440, 0.0427
8	0.0323, 0.0382, 0.0360
9	0.0533, 0.0407, 0.0626
10	0.160, 0.437, 0.343
11	0.135, 0.142, 0.0592

(a) Test whether there is an overall effect of site on PCB concentration, using one-way ANOVA and SAS. Treat site as a fixed effect. Report your results using P values and discuss the significance of the test. A log transformation should be applied before analysis.

(b) Use the REGW procedure to compare the different sites, and interpret your results. Which pairs of sites are significantly different? Do the sites fall into particular groups?

3. An entomologist wants to compare the attractiveness of nine different baits (A-I) for bark beetles. There were three replicate traps for each bait treatment. The table below lists the number of beetles captured in each trap.

Bait	Beetles, trap 1-3
A	27, 36, 26
B	25, 19, 37
C	8, 16, 12
D	15, 8, 12
E	68, 42, 57
F	43, 32, 47
G	10, 12, 19
H	71, 62, 53
I	19, 11, 21

(a) Test whether there is an overall effect of bait on beetle captures, using one-way ANOVA and SAS. Report your results using P

values and discuss the significance of the test. Apply a log transformation before analysis.

(b) Use the FDR procedure to compare the different baits, and interpret your results. Which baits are significantly different?

Chapter 14

Analysis of Variance (Two-Way)

Two-way ANOVA examines how two different factors, such as different experimental treatments, affect the means of the different groups. For example, we might be interested in how different baits, as well as trap color, affect the number of insects caught in the traps. If we conducted an experiment where traps were deployed with different combinations of bait and trap color, this would be a **two-way factorial design**, where the term factorial implies all possible combinations of the two factors. If there were three different baits (A, B, and C) and two trap colors (black, white), a factorial design implies there would be six different treatment combinations in the experiment (A-black, A-white, B-black, B-white, C-black, C-white). There would be one or more traps deployed with each treatment combination. It is customary to call one of the factors in a two-way design 'Factor A', while the other is 'Factor B'.

Similar to one-way ANOVA designs, the factors in two-way ANOVA can be either fixed or random. In the insect trapping experiment discussed above, both bait and trap color would be fixed effects because they were selected by the investigator. There are then F tests for each factor in the design, and potentially a test for the **interaction** of the two factors. **An interaction between two factors implies there is a joint effect of the two factors beyond that predicted by each factor operating additively.** For example, insects might be strongly attracted to A-black traps, more than would be predicted by the bait and trap color effects observed in the rest of the treatments. We will focus some effort on the analysis of this design because it is one of the more common ones.

There are other possible two-way designs, including one fixed and one random effect, or more rarely both effects are random. We will examine a popular design where one factor is fixed and the other random, called a **randomized block design**. There is an F test for the fixed effect in this design, and this test is often the primary goal of the analysis. With respect to the random effects, it is common to simply estimate the variance components associated with these effects and not conduct any tests, although these are still available. This design is ubiquitous in field studies because it helps control for certain forms of spatial or temporal heterogeneity in the observations, permitting a more powerful test of any treatment or group effects.

What do the data look like for a two-way ANOVA design? We will first examine a simplified data set from a trapping study of the bark beetle predator *T. dubius* (Reeve et al. 2009). These predators feed on bark beetles which attack and kill pine trees, and are attracted to the pheromones of the bark beetles as well as odors emitted by damaged pines. Visual cues may also play a role in their behavior, in particular the dark vertical silhouette provided by the bole of the tree. Three different baits were used: frontalin + turpentine (FRT), ipsdienol + turpentine (IDT), and ipsenol + turpentine (IST). Frontalin, ipsdienol, and ipsenol are bark beetle pheromones, while turpentine contains volatiles similar to those in pine resin. The traps were also painted two different colors, black vs. white, to manipulate their appearance to the predators. Thus, there were a total of six treatments (three baits, two colors) in the design. The different treatments were randomly assigned to trapping locations along transects in a pine forest, with four replicates per treatment. The number of predators caught in each trap were counted after several weeks of trapping (Table 14.1). The fourth column in the table shows the values after applying a log transformation, which is commonly used with count data (see Chapter 15).

We will use the notation Y_{ijk} to reference the observations in two-way ANOVA designs. The i subscript refers to the group or treatment within Factor A (bait), j the group or treatment within Factor B (trap color), while k refers to the observation within the treatment. For example, Y_{123} refers to the third observation in the FRT bait - W color treatment, which is 0.903.

We will also examine data from an experiment that examined how nutrient and water availability, as well as resource heterogeneity in space or time, affect biomass production in grassland plants (Maestre & Reynolds

2007). Plants from a grassland community were seeded in small containers in the greenhouse, with the treatments consisting of different levels of nitrogen and watering. There were three nitrogen and three watering levels in the experiment, for a total of nine treatments, with four replicate containers per treatment. The experiment also included treatments were the nitrogen was heterogeneously distributed in the container and watering was pulsed in time, but we will defer analysis of these other factors to Chapter 19. The total biomass of the plants was then determined after 100 d of growth (Table 14.2).

The data sets presented in this chapter are balanced designs with the same number of replicates per group, because this simplifies the formulas. They can be extended to unbalanced designs, but we will let SAS handle the details of the calculations in this case. We will later see how unbalanced data sets can influence the tests in two-way ANOVA.

Table 14.1: Example 1 - Effect of bait and trap color on catches of *T. dubius*, a bark beetle predator (Reeve et al. 2009). The baits used were frontalin + turpentine (FRT), ipsdienol + turpentine (IDT), and ipsenol + turpentine (IST), and the traps were painted either black (B) or white (W). Also shown are the means for each treatment group ($\bar{Y}_{ij\cdot}$) and preliminary calculations to find SS_{within}

Bait	Color	*T. dubius*	$Y_{ijk} = \log_{10}(T.dubius + 1)$	i	j	k	$\bar{Y}_{ij\cdot}$	$(Y_{ijk} - \bar{Y}_{ij\cdot})^2$
FRT	B	18	1.279	1	1	1	1.150	1.664×10^{-2}
FRT	B	12	1.114	1	1	2		1.296×10^{-3}
FRT	B	22	1.362	1	1	3		4.494×10^{-2}
FRT	B	6	0.845	1	1	4		9.303×10^{-2}
FRT	W	12	1.114	1	2	1	0.980	1.796×10^{-2}
FRT	W	15	1.204	1	2	2		5.018×10^{-2}
FRT	W	7	0.903	1	2	3		5.929×10^{-3}
FRT	W	4	0.699	1	2	4		7.896×10^{-2}
IDT	B	0	0.000	2	1	1	0.369	1.363×10^{-1}
IDT	B	2	0.477	2	1	2		1.161×10^{-2}
IDT	B	1	0.301	2	1	3		4.658×10^{-3}
IDT	B	4	0.699	2	1	4		1.087×10^{-1}
IDT	W	2	0.477	2	2	1	0.314	2.665×10^{-2}
IDT	W	1	0.301	2	2	2		1.626×10^{-4}
IDT	W	2	0.477	2	2	3		2.665×10^{-2}
IDT	W	0	0.000	2	2	4		9.844×10^{-2}

Bait	Color	*T. dubius*	$Y_{ijk} = \log_{10}(T.dubius + 1)$	i	j	k	$\bar{Y}_{ij.}$	$(Y_{ijk} - \bar{Y}_{ij.})^2$
IST	B	2	0.477	3	1	1	0.725	6.126×10^{-2}
IST	B	2	0.477	3	1	2		6.126×10^{-2}
IST	B	10	1.041	3	1	3		1.002×10^{-1}
IST	B	7	0.903	3	1	4		3.186×10^{-2}
IST	W	1	0.301	3	2	1	0.719	1.745×10^{-1}
IST	W	4	0.699	3	2	2		3.901×10^{-4}
IST	W	14	1.176	3	2	3		2.091×10^{-1}
IST	W	4	0.699	3	2	4		3.901×10^{-4}

Table 14.2: Example 2 - Effect of nutrient and water availability on the total biomass of grassland plants grown in microcosms (Maestre & Reynolds 2007).

N (mg)	Water (ml/week)	Y_{ijk} = Biomass	i	j	k
40	125	4.372	1	1	1
40	125	4.482	1	1	2
40	125	4.221	1	1	3
40	125	3.977	1	1	4
40	250	7.400	1	2	1
40	250	8.027	1	2	2
40	250	7.883	1	2	3
40	250	7.769	1	2	4
40	375	7.226	1	3	1
40	375	8.126	1	3	2
40	375	6.840	1	3	3
40	375	7.901	1	3	4
80	125	5.140	2	1	1
80	125	3.913	2	1	2
80	125	4.669	2	1	3
80	125	4.306	2	1	4
80	250	9.099	2	2	1
80	250	9.711	2	2	2
80	250	9.123	2	2	3
80	250	9.709	2	2	4
80	375	10.701	2	3	1
80	375	11.552	2	3	2
80	375	11.356	2	3	3
80	375	9.759	2	3	4

N (mg)	Water (ml)	Y_{ijk} = Biomass	i	j	k
120	125	5.021	3	1	1
120	125	4.970	3	1	2
120	125	5.055	3	1	3
120	125	4.862	3	1	4
120	250	9.029	3	2	1
120	250	10.791	3	2	2
120	250	9.115	3	2	3
120	250	10.319	3	2	4
120	375	12.189	3	3	1
120	375	14.381	3	3	2
120	375	13.153	3	3	3
120	375	14.066	3	3	4

14.1 Random assignment of treatments

A essential step in executing ANOVA designs is the random assignment of treatments to experimental units. For instance, in the Example 2 experiment we would want to randomly assign nitrogen and watering levels to the microcosms. This avoids any bias on the part of the experimenter in assigning the treatments to the containers, and also ensures that the replicates for each treatment are spread and intermingled throughout the greenhouse. What could happen if the treatments are not randomly assigned? Suppose that all the replicates for a given treatment in Example 2 are placed next to each other in the greenhouse, perhaps because this is convenient when applying the treatments. If a particular location happens to be warmer or receive more sunlight than another location, then the plants may be larger in that location and so bias the results of the experiment. We may falsely conclude a particular treatment has an effect on biomass because of this location effect. The random assignment of treatments avoids biases of this sort and also ensures independence of the observations, a basic assumption of most statistical models (Hurlbert 1984; Potvin 1993). Experiments with this feature are also known as **completely randomized designs**. We will illustrate the random assignment of treatments using a SAS program below.

14.1.1 Random assignment of treatments - SAS Demo

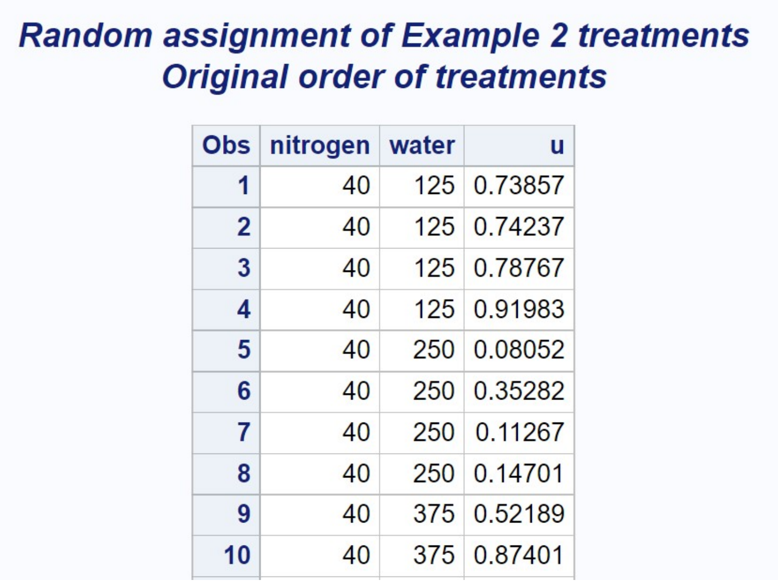
The program below shows one way of randomly assigning treatments to containers for the Example 2 experiment. We first input the different treatment combinations using a `data` step, with one line in the data set for each replicate. The `data` step also assigns a random number to each observation. The program uses a uniform random variable generated by the `ranuni` function, but any continuous random variable would work. We then use `proc sort` to sort the observations in ascending order by this random variable, thereby randomly shuffling the treatments (see Fig. 14.2). We would then assign to the first container the first treatment combination in the shuffled observations, the second container the second treatment combination, and so forth.

```
* Rand_treatments.sas;
title "Random assignment of Example 2 treatments";
data treat;
    input nitrogen water;
    * Generate a uniform random variable;
    u = ranuni(0);
    datalines;
 40  125
 40  125
 40  125
 40  125
 40  250
 40  250
 40  250
 40  250

etc.

;
run;
title2 "Original order of treatments";
proc print data=treat;
run;
* Sort treatments by value of u;
proc sort out=shuffled data=treat;
    by u;
run;
title2 "Randomly shuffled treatments";
proc print data=shuffled;
run;
```

```
quit;
```

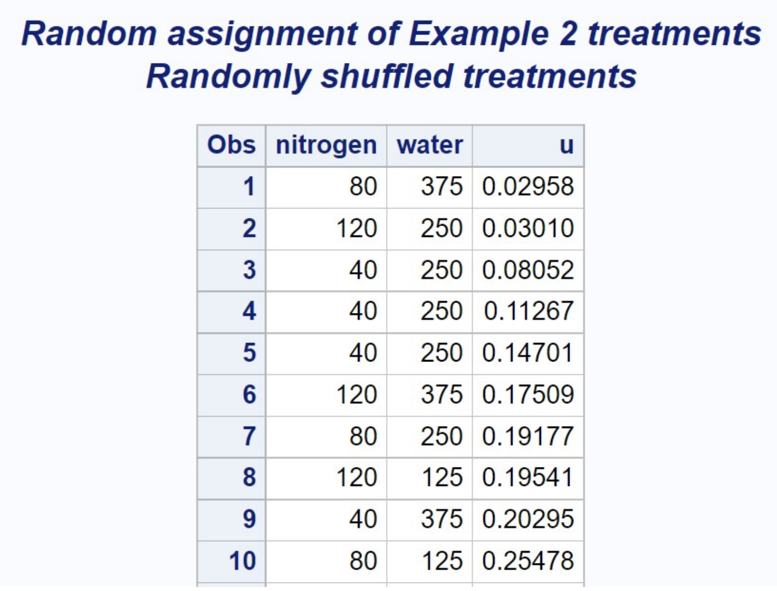


Random assignment of Example 2 treatments
Original order of treatments

Obs	nitrogen	water	u
1	40	125	0.73857
2	40	125	0.74237
3	40	125	0.78767
4	40	125	0.91983
5	40	250	0.08052
6	40	250	0.35282
7	40	250	0.11267
8	40	250	0.14701
9	40	375	0.52189
10	40	375	0.87401

etc.

Figure 14.1: `rand_treatments.sas - proc print`



Random assignment of Example 2 treatments
Randomly shuffled treatments

Obs	nitrogen	water	u
1	80	375	0.02958
2	120	250	0.03010
3	40	250	0.08052
4	40	250	0.11267
5	40	250	0.14701
6	120	375	0.17509
7	80	250	0.19177
8	120	125	0.19541
9	40	375	0.20295
10	80	125	0.25478

etc.

Figure 14.2: `rand_treatments.sas - proc print`

14.2 Two-way fixed effects model

Suppose that we want to model the observations in studies like Example 1 or 2, where there are two factors that are manipulated and are fixed effects. Let Factor A be one treatment (such as bait type) while Factor B is the other treatment (trap color). Let the symbol Y_{ijk} stand for the kth observation ($k = 1, 2, \ldots, n$) in the ith Factor A treatment and jth Factor B treatment. For example, with the Example 1 data set we have $Y_{111} = 1.279$ while $Y_{222} = 0.301$ (see Table 14.1). One commonly used model for such a design (Searle 1971) is

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}. \tag{14.1}$$

Here μ is the grand mean of the observations, while α_i is the deviation from μ caused by the ith treatment in Factor A, while β_j is the deviation caused by the jth treatment in Factor B. These terms are called the **main effects** in the model. The term $(\alpha\beta)_{ij}$ represents an **interaction** between Factors A and B, implying a shift in the mean for a particular treatment combination beyond the effects of Factor A and B. An interaction between two factors A and B is often symbolized as 'A × B.' It is also considered a fixed effect when both A and B are fixed effects. The ϵ_{ijk} term represents random departures from the mean value predicted by the main effects and interaction due to natural variability among the observations, and are also assumed to be independent. The model also assumes that $\sum \alpha_i = 0, \sum \beta_j = 0$, and $\sum (\alpha\beta)_{ij} = 0$, but this does not affect its generality. The same model can also be used to describe the observations for studies where there are a groups or levels for Factor A, and b for Factor B, with any number of replicates (n) per treatment combination, as well as unbalanced designs with different numbers of replicates.

It follows for the ith level of Factor A and jth of Factor B that $E[Y_{ijk}] = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$ and $Var[Y_{ijk}] = \sigma^2$, using the rules for expected values and variances. Thus, for the ith and jth level we have $Y_{ijk} \sim N(\mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}, \sigma^2)$. We can illustrate how the different parameters work in this model by plotting the distribution of the data for different parameter values. The behavior of the model is described for four different scenarios below. We will model an experiment similar to Example 1, where there are three levels for Factor A ($a = 3$) and two for Factor B ($b = 2$).

14.2.1 Factor A effect

Suppose that Factor A has a strong effect on Y_{ijk}, but there is only a minimal effect of Factor B and no interaction between the two factors. To make things concrete, let $\mu = 1.5$, $\alpha_1 = 0.5, \alpha_2 = 0, \alpha_3 = -0.5, \beta_1 = 0.1, \beta_2 = -0.1$, $(\alpha\beta)_{ij} = 0$ for all i and j, and $\sigma^2 = 0.05$. Figure 14.3 shows the distribution of the observations in each treatment group. Note that the mean for treatment 1 under Factor A is shifted upward from μ while treatment 3 is shifted downward, for both levels of Factor B. The distribution for each treatment combination has the same variance, namely $\sigma^2 = 0.05$.

14.2.2 Factor B effect

Suppose the reverse situation is now true, with Factor B having a strong effect on Y_{ijk} while Factor A has a minimal effect, again with no interaction. This could be modeled using $\alpha_1 = 0.1, \alpha_2 = 0, \alpha_3 = -0.1, \beta_1 = 0.5, \beta_2 = -0.5$, and $(\alpha\beta)_{ij} = 0$ for all i and j. Figure 14.4 shows the pattern that results. Note that the mean for treatment 1 under Factor B is shifted upward from μ, while treatment 2 is shifted downward, for all three levels of Factor A.

14.2.3 Factor A and B effect

If both factors have an effect on Y_{ijk}, we would expect to see a combination of the previous patterns, with the treatment groups shifted away from each other (Fig. 14.5). This figure uses $\alpha_1 = 0.5, \alpha_2 = 0, \alpha_3 = -0.5, \beta_1 = 0.3, \beta_2 = -0.3$, and $(\alpha\beta)_{ij} = 0$ for all i and j.

14.2.4 Interaction effect

We now examine how an A × B interaction influences the model. Suppose that $\alpha_1 = 0.5, \alpha_2 = 0, \alpha_3 = -0.5, \beta_1 = 0.3$ and $\beta_2 = -0.3$ as in the previous figure, but now $(\alpha\beta)_{11} = 0.2, (\alpha\beta)_{12} = -0.2, (\alpha\beta)_{21} = 0, (\alpha\beta)_{22} = 0, (\alpha\beta)_{31} = -0.2$, and $(\alpha\beta)_{32} = 0.2$. We see that Factor B has a substantial effect under treatment 1 for Factor A, and smaller effect under treatment 2, and almost no effect under treatment 3 (Fig. 14.6). Note that the distributions under the different treatment combinations no longer move in parallel as in Fig. 14.5. This pattern is diagnostic of an interaction in the analysis of real data. We will later examine a data set where there is strong interaction between the two factors.

The objective in two-way ANOVA is to test whether Factor A, B, or both have an effect on the group means, and whether there is interaction between the two factors. For Factor A this amounts to testing H_0 : all $\alpha_i = 0$, while for Factor B we would test H_0 : all $\beta_j = 0$. For interaction between the two factors, we would test H_0 : all $(\alpha\beta)_{ij} = 0$. The corresponding alternative hypotheses are H_1 : some $\alpha_i \neq 0$, H_1 : some $\beta_j \neq 0$, and H_1 : some $(\alpha\beta)_{ij} \neq 0$. We will discuss how these null hypotheses are tested in the next section.

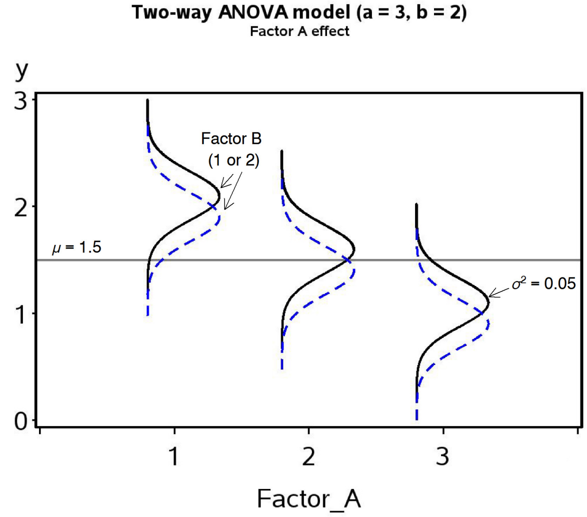


Figure 14.3: Fixed effects model for two-way ANOVA showing a Factor A effect.

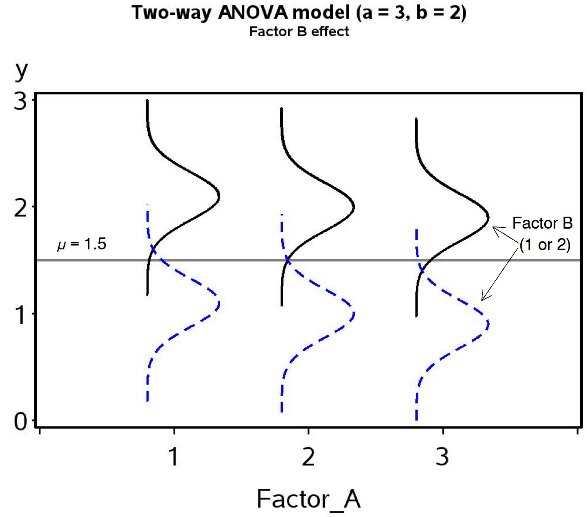


Figure 14.4: Fixed effects model for two-way ANOVA showing a Factor B effect.

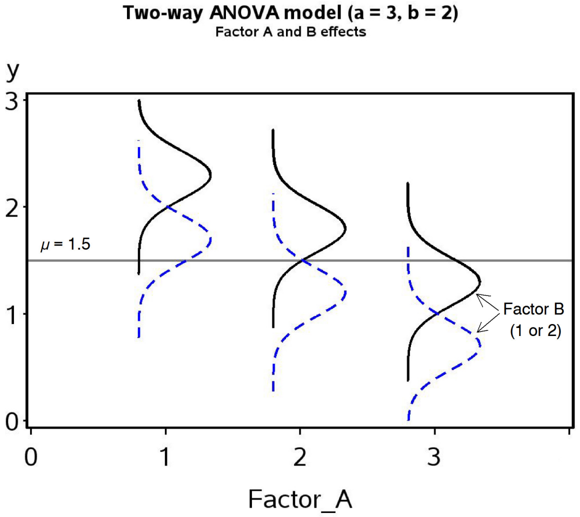


Figure 14.5: Fixed effects model for two-way ANOVA showing both Factor A and B effects.

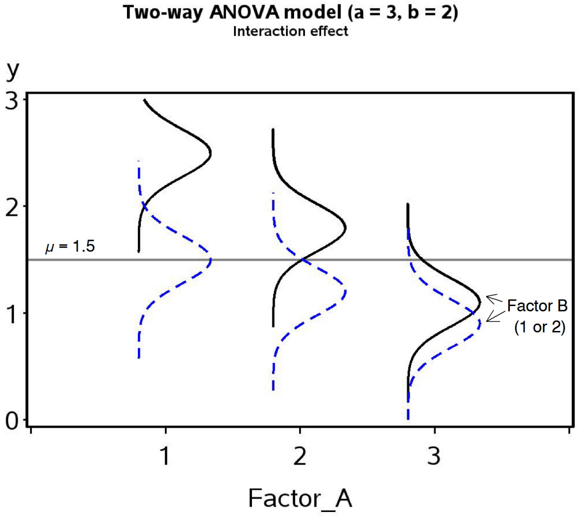


Figure 14.6: Fixed effects model for two-way ANOVA showing an A × B interaction between the two factors.

14.3 Hypothesis testing for two-way ANOVA

We now develop statistical tests for each of the null hypotheses listed above. All work in a similar fashion to the F test for one-way ANOVA. For Factor A and B in the model, as well as the interaction term, there is a corresponding sum of squares and mean square term. There is also an overall sum of squares and mean square within groups. These quantities are used to construct three different F tests, one for Factor A, Factor B, and the A × B interaction. These three tests are also examples of likelihood ratio tests, in which the fit is compared between the null and alternative models (Searle 1971). We will illustrate the calculations for these tests using the Example 1 data set, with Factor A being bait while Factor B is trap color.

14.3.1 Sum of squares and mean squares

We begin by calculating the group means for each treatment combination. For the Example 1 data, this amounts to calculating a group mean for each combination of bait and trap color. These group means are shown in Table 14.1 and labeled as $\bar{Y}_{ij\cdot}$. Here the '·' notation implies the mean was calculated using all the observations in that group ($k = 1, 2, \ldots, n$). A grand mean can then be calculated as the mean of these group means, or equivalently by summing all the observations and dividing by their total number. We label this grand mean as $\bar{\bar{\bar{Y}}}$. It can be generally calculated using the formula

$$\bar{\bar{\bar{Y}}} = \frac{\sum_{i=1}^{a}\sum_{j=1}^{b}\bar{Y}_{ij\cdot}}{ab}. \tag{14.2}$$

For the Example 1 data set, we have

$$\bar{\bar{\bar{Y}}} = \frac{1.150 + 0.980 + 0.369 + 0.314 + 0.725 + 0.719}{6} = 0.709. \tag{14.3}$$

We next calculate a mean corresponding to each level of Factor A by averaging across the levels of Factor B, which we denote as $\bar{\bar{Y}}_{i\cdot\cdot}$. It can be calculated using the formula

$$\bar{\bar{Y}}_{i\cdot\cdot} = \frac{\sum_{j=1}^{b}\bar{Y}_{ij\cdot}}{b}. \tag{14.4}$$

For the Example 1 data set, we have

$$\bar{\bar{Y}}_{1\cdot\cdot} = \frac{1.150 + 0.980}{2} = 1.065, \tag{14.5}$$

$$\bar{\bar{Y}}_{2\cdot\cdot} = \frac{0.369 + 0.314}{2} = 0.342, \tag{14.6}$$

and

$$\bar{\bar{Y}}_{3\cdot\cdot} = \frac{0.725 + 0.719}{2} = 0.722. \tag{14.7}$$

The difference $\bar{\bar{Y}}_{i\cdot\cdot} - \bar{\bar{\bar{Y}}}$ is a measure of the shift generated by Factor A in the observations, as well as an estimate of α_i for each level of Factor A. We can obtain a single measure of these shifts by squaring and summing them across all groups to obtain a sum of squares for Factor A, or SS_A. It can be calculated using the general formula

$$SS_A = nb \sum_{i=1}^{a} (\bar{\bar{Y}}_{i\cdot\cdot} - \bar{\bar{\bar{Y}}})^2. \tag{14.8}$$

SS_A has $a-1$ degrees of freedom. We can calculate a mean square for Factor A using the formula

$$MS_A = \frac{SS_A}{a-1}. \tag{14.9}$$

Note the factor nb in the expression for SS_A, which scales MS_A so that it estimates σ^2 if H_0 : all $\alpha_i = 0$ is true (no Factor A effect). If H_1 is true, implying some $\alpha_i \neq 0$, then MS_A will become larger. For the Example 1 data, we have

$$SS_A = 4(2)\left[(1.065 - 0.709)^2 + (0.342 - 0.709)^2 + (0.722 - 0.709)^2\right] \tag{14.10}$$

$$= 8\left[1.265 \times 10^{-1} + 1.353 \times 10^{-1} + 1.501 \times 10^{-4}\right] = 2.096. \tag{14.11}$$

and

$$MS_A = \frac{2.096}{3-1} = 1.048. \tag{14.12}$$

We can similarly calculate a mean corresponding to each level of Factor B, averaging across levels of Factor A. The general formula for these means is

$$\bar{\bar{Y}}_{\cdot j\cdot} = \frac{\sum_{i=1}^{a} \bar{Y}_{ij\cdot}}{a}. \tag{14.13}$$

For the Example 1 data set, we have

$$\bar{\bar{Y}}_{\cdot 1\cdot} = \frac{1.150 + 0.369 + 0.725}{3} = 0.748 \tag{14.14}$$

and

$$\bar{\bar{Y}}_{\cdot 2\cdot} = \frac{0.980 + 0.314 + 0.719}{3} = 0.671. \tag{14.15}$$

The difference $\bar{\bar{Y}}_{\cdot j\cdot} - \bar{\bar{\bar{Y}}}$ is a measure of the shift generated by Factor B in the observations, as well as an estimate of β_j for each level of Factor B. Squaring and summing them across all groups, we obtain a sum of squares for Factor B, or SS_B. It can be calculated using the general formula

$$SS_B = na \sum_{j=1}^{b} (\bar{\bar{Y}}_{\cdot j\cdot} - \bar{\bar{\bar{Y}}})^2. \tag{14.16}$$

SS_B has $b - 1$ degrees of freedom. We can then calculate a mean square for Factor B using the formula

$$MS_B = \frac{SS_B}{b-1}. \tag{14.17}$$

For the Example 1 data, we have

$$SS_B = 4(3)\left[(0.748 - 0.709)^2 + (0.671 - 0.709)^2\right] \tag{14.18}$$

$$= 12\left[1.485 \times 10^{-3} + 1.485 \times 10^{-3}\right] = 3.565 \times 10^{-2} \tag{14.19}$$

and

$$MS_B = \frac{3.565 \times 10^{-2}}{2-1} = 3.565 \times 10^{-2}. \tag{14.20}$$

We can also calculate a sum of squares and mean square to test for the A × B interaction. The sum of squares for interaction, SS_{AB}, is calculated in general using the formula

$$SS_{AB} = n \sum_{i=1}^{a} \sum_{j=1}^{b} (\bar{Y}_{ij\cdot} - \bar{\bar{Y}}_{i\cdot\cdot} - \bar{\bar{Y}}_{\cdot j\cdot} + \bar{\bar{\bar{Y}}})^2. \tag{14.21}$$

The terms within this expression estimate $(\alpha\beta)_{ij}$, and are measures of the difference between the means for each treatment combination and the values predicted by the model without any interaction. SS_{AB} has $(a-1)(b-1)$ degrees of freedom. Its associated mean square is defined by the formula

$$MS_{AB} = \frac{SS_{AB}}{(a-1)(b-1)}. \tag{14.22}$$

For the Example 1 data, we have

$$\begin{aligned} SS_{AB} = 4[&(1.150 - 1.065 - 0.748 + 0.709)^2 && (14.23)\\ &+(0.980 - 1.065 - 0.671 + 0.709)^2 && (14.24)\\ &+(0.369 - 0.342 - 0.748 + 0.709)^2 && (14.25)\\ &+(0.314 - 0.342 - 0.671 + 0.709)^2 && (14.26)\\ &+(0.725 - 0.722 - 0.748 + 0.709)^2 && (14.27)\\ &+(0.719 - 0.722 - 0.671 + 0.709)^2 && (14.28)\\ = 5[2.111 \times 10^{-3} + \cdots + 2.836 \times 10^{-2}] = 2.836 \times 10^{-2}. && && (14.29) \end{aligned}$$

and

$$MS_{AB} = \frac{2.836 \times 10^{-2}}{(3-1)(2-1)} = 1.418 \times 10^{-2}. \quad (14.30)$$

These sum of squares and mean squares measure how Factor A, B, and the A × B interaction influence the means of each treatment combination. What about variability within each group? We can calculate SS_{within} using the general formula

$$SS_{within} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk} - \bar{Y}_{ij\cdot})^2 \quad (14.31)$$

which has $ab(n-1)$ degrees of freedom. The associated mean square is calculated as

$$MS_{within} = \frac{SS_{within}}{ab(n-1)}. \quad (14.32)$$

The last column of Table 14.1 shows the preliminary calculations for SS_{within}. Adding this column across all the treatment groups yields

$$SS_{within} = 1.644 \times 10^{-2} + \cdots + 3.901 \times 10^{-4} = 1.361 \quad (14.33)$$

and

$$MS_{within} = \frac{1.361}{(3)(2)(4-1)} = 7.561 \times 10^{-2}. \quad (14.34)$$

There is one more sum of squares that is often calculated in two-way ANOVA, the total sum of squares. It is defined as

$$SS_{total} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n}(Y_{ijk} - \bar{\bar{\bar{Y}}})^2. \quad (14.35)$$

It measures the variability of the observations around the grand mean of the data ($\bar{\bar{\bar{Y}}}$) and has $abn - 1$ degrees of freedom. An interesting feature of the sum of squares is that they add to the total sum of squares when the design is balanced, as do the degrees of freedom. In particular, we have

$$SS_A + SS_B + SS_{AB} + SS_{within} = SS_{total} \tag{14.36}$$

and

$$(a-1) + (b-1) + (a-1)(b-1) + ab(n-1) = abn - 1. \tag{14.37}$$

Thus, the sum of squares and degrees of freedom can be partitioned into components corresponding to every source of variation in the study. For Example 1, we have $SS_{total} = 3.521$ with $3(2)(4) - 1 = 23$ degrees of freedom.

14.3.2 ANOVA tables and tests

We can organize the different sum of squares and mean squares into an ANOVA table. It lists the different sources of variation in the data (Factor A, B, A × B interaction, within groups, and total) and their degrees of freedom. Table 14.3 shows the general layout of such a table for two-way ANOVA designs.

Also shown in the table are F statistics used test to whether Factor A, Factor B, and their interaction have an effect on the observations. The numerator of the test statistic is the mean square for each factor (MS_A, MS_B, or MS_{AB}), while the denominator is always MS_{within}. Thus, we use $F_s = MS_A/MS_{within}$ to test for the effect of Factor A. Under H_0 : all $\alpha_i = 0$ this statistic has an F distribution with $df_1 = a-1$ and $df_2 = ab(n-1)$. Similarly, we use $F_s = MS_B/MS_{within}$ to test for an effect of Factor B. Under H_0 : all $\beta_j = 0$ it has an F distribution with $df_1 = b-1$ and $df_2 = ab(n-1)$. Finally, we use $F_s = MS_{AB}/MS_{within}$ to test for an interaction between A and B. Under H_0 : all $(\alpha\beta)_{ij} = 0$ it has an F distribution with $df_1 = (a-1)(b-1)$ and $df_2 = ab(n-1)$.

All these tests are examples of likelihood ratio tests. For example, consider the test for the A × B interaction. To construct the likelihood ratio test for the interaction, we first find the maximum likelihood estimates of various parameters under H_1 vs. H_0. Recall that the observations in the two-way ANOVA model are described as

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}. \tag{14.38}$$

where μ is the grand mean, α_i is the effect of the ith level of Factor A, β_j is the effect of the jth level of Factor B, $(\alpha\beta)_{ij}$ is effect of the interaction, and $\epsilon_{ijk} \sim N(0, \sigma^2)$. This is the statistical model under the alternative hypothesis H_1 : some $(\alpha\beta)_{ij} \neq 0$, implying an interaction effect. Under H_0 : all $(\alpha\beta)_{ij} = 0$, the model reduces to

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}. \tag{14.39}$$

We would need to find the maximum likelihood estimates under both H_1 and H_0, as well as L_{H_0} and L_{H_1}, the maximum height of the likelihood function under H_0 and H_1. We would then use the likelihood ratio test statistic

$$\lambda = \frac{L_{H_0}}{L_{H_1}}. \tag{14.40}$$

It can be shown that there is a one-to-one correspondence between $-2\ln(\lambda)$ and F_s for the interaction effect, and so the F test is actually a likelihood ratio test (Searle 1971), as are the tests for the other effects. Large values of the test statistic $-2\ln(\lambda)$ or F_s indicate a lower value of the likelihood under H_0 relative to H_1, and thus a poorer fit of the H_0 model.

Table 14.4 shows the results for the Example 1 data set, including the F statistics and P values obtained using Table F. **In examining the test results, it is customary to examine the test for the interaction first, followed by the main effects.** If the interaction is nonsignificant this suggests the two main effects have a simple additive effect on the observations, provided they are significant. If the interaction is significant the interpretation requires more attention. If one or more of the main effects are significant, it suggests the observations are driven by both interaction and main effects. Fig. 14.6 shows a theoretical example where an interaction, Factor A, and Factor B all influence the observations.

For the bait × trap color interaction, we see that $F_s = 0.19$ with $df_1 = 2$ and $df_2 = 18$, and from Table F find that $P > 0.100$. Thus, the interaction was nonsignificant for these data ($F_{2,18} = 0.19, P > 0.100$). The color effect was also nonsignificant ($F_{1,18} = 0.47$, $P > 0.100$), but the bait effect was highly significant ($F_{2,18} = 13.86, P < 0.001$). Each bait represents a different bark beetle pheromone, and apparently some baits are more attractive than others for *T. dubius*.

Table 14.3: General ANOVA table for two-way designs with replication, showing formulas for different mean squares and F tests.

Source	*df*	Sum of squares	Mean square	F_s
Factor A	$a-1$	SS_A	$MS_A = SS_A/(a-1)$	MS_A/MS_{within}
Factor B	$b-1$	SS_B	$MS_B = SS_B/(b-1)$	MS_B/MS_{within}
AB interaction	$(a-1)(b-1)$	SS_{AB}	$MS_{AB} = SS_{AB}/(a-1)(b-1)$	MS_{AB}/MS_{within}
Within	$ab(n-1)$	SS_{within}	$MS_{within} = SS_{within}/ab(n-1)$	
Total	$abn-1$	SS_{total}		

Table 14.4: ANOVA table for the Example 1 data set, including P values for the tests.

Source	*df*	Sum of squares	Mean square	F_s	P
Bait	2	2.096	1.048	13.86	< 0.001
Color	1	3.565×10^{-2}	3.565×10^{-2}	0.47	> 0.100
Bait × Color	2	2.836×10^{-2}	1.418×10^{-2}	0.19	> 0.100
Within	18	1.361	7.561×10^{-2}		
Total	23	3.521			

14.3.3 Two-way ANOVA for Example 1 - SAS demo

The same calculations for the Example 1 study can be carried out using `proc glm` (SAS Institute Inc. 2018). This procedure is primarily intended for fixed effects ANOVA models, and this study has two fixed effects, bait type and trap color, plus the interaction is also considered a fixed effect.

The first step in the program (see below) is to read in the observations using a `data` step, with one variable denoting the bait treatment (`bait`), another the trap color (`color`), and the third the number of *T. dubius* captured per trap (`Tdubius`). These numbers are then log-transformed using a SAS function to yield the variable `y = log10(Tdubius+1)`. We add one to the observations before taking the log to avoid problems with zeroes.

The data are then plotted using `proc gplot`, with the `bait` treatment on the x-axis and separate lines drawn for each color (SAS Institute Inc. 2016). This is accomplished with the command `plot y*bait=color`. The rest of the `gplot` statements control the appearance of the symbols and axes.

The next section of the program conducts the two-way ANOVA using `proc glm`. The `class` statement tells SAS that both `bait` and `color` are used to classify the observations into the six treatment groups. The `model` statement tells SAS the form of the ANOVA model. Recall that the model for fixed effects two-way ANOVA is given by the equation

$$Y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk} \tag{14.41}$$

The α_i, β_j, and $(\alpha\beta)_{ij}$ terms in this model equate directly with the `bait`, `color`, and `bait*color` entries in the `model` statement. The `lsmeans` statement causes `glm` to calculate quantities called least squares means for each level of `bait` and `color`. When the data are balanced these are equivalent to the means for each treatment group, but least squares means have some advantages for unbalanced data and other statistical models. The option `adjust=tukey` requests multiple comparisons among treatments using the Tukey method. This is useful for comparing the different `bait` treatments, but for color there is only one comparison (black vs. white) and in this case would be equivalent to the F test for `color`.

The `proc glm` output provides information similar to that summarized in an ANOVA table (see Fig. 14.9). The degrees of freedom, sum of squares, mean squares, F statistics and P values for the bait, color, and bait × color interaction are listed near the bottom of the output under `Type III SS`. The degrees of freedom, sum of squares, and mean square for the variation within groups are labeled as `Error` above this section (this terminology will be explained in Chapter 15). The output labeled `Type I SS` is produced by sequentially fitting the different terms in the model, in the order listed in the `model` statement. Type III sums of squares are more generally useful than Type I for ANOVA designs, although the results are the same when the design is balanced. The output labeled `Model` refers to the combined variation due to bait, color, plus their interaction. The associated F statistic tests whether any or all of these effects influence the observations vs. the null hypothesis that they have no effect. This particular test is not used much with ANOVA designs.

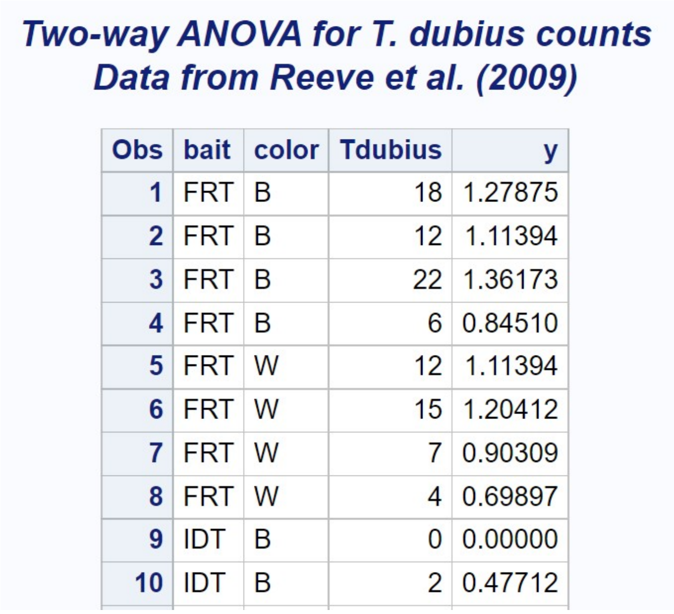
We now examine the results of these tests, beginning with the interaction (Fig. 14.9). We see that the bait × color interaction was nonsignificant ($F_{2,18} = 0.19, P = 0.8311$). The color effect was also nonsignificant ($F_{1,18} = 0.47, P = 0.5011$), while bait was highly significant ($F_{2,18} = 13.85, P = 0.0002$). Examining the graph and Tukey results (Fig. 14.8, 14.10), we see that predator densities for the FRT and IST treatments were significantly higher than for IDT. Note that the lines connecting the different treatments are roughly parallel, further indicating an absence of interaction. The effect of trap color appears minimal in this study, although trap catches were somewhat higher for black traps.

SAS Program

```
* Tdubius_bait_color.sas;
title "Two-way ANOVA for T. dubius trapping";
title2 "Data from Reeve et al. (2009)";
data Tdubius;
    input bait $ color $ Tdubius;
    * Apply transformations here;
    y = log10(Tdubius+1);
    datalines;
FRT B   18
FRT B   12
FRT B   22
FRT B   6
FRT W   12
FRT W   15
FRT W   7
FRT W   4

etc.
```

```
;
run;
* Print data set;
proc print data=Tdubius;
run;
* Plot means, standard errors, and observations;
proc gplot data=Tdubius;
    plot y*bait=color / vaxis=axis1 haxis=axis1 legend=legend1;
    symbol1 i=std1mjt v=star height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
    legend1 label=(height=2) value=(height=2);
run;
* Two-way ANOVA with all fixed effects;
proc glm plots=diagnostics data=Tdubius;
    class bait color;
    model y = bait color bait*color;
    lsmeans bait color / adjust=tukey cl lines;
run;
quit;
```



Two-way ANOVA for T. dubius counts
Data from Reeve et al. (2009)

Obs	bait	color	Tdubius	y
1	FRT	B	18	1.27875
2	FRT	B	12	1.11394
3	FRT	B	22	1.36173
4	FRT	B	6	0.84510
5	FRT	W	12	1.11394
6	FRT	W	15	1.20412
7	FRT	W	7	0.90309
8	FRT	W	4	0.69897
9	IDT	B	0	0.00000
10	IDT	B	2	0.47712

etc.

Figure 14.7: `Tdubius_bait_color.sas - proc print`

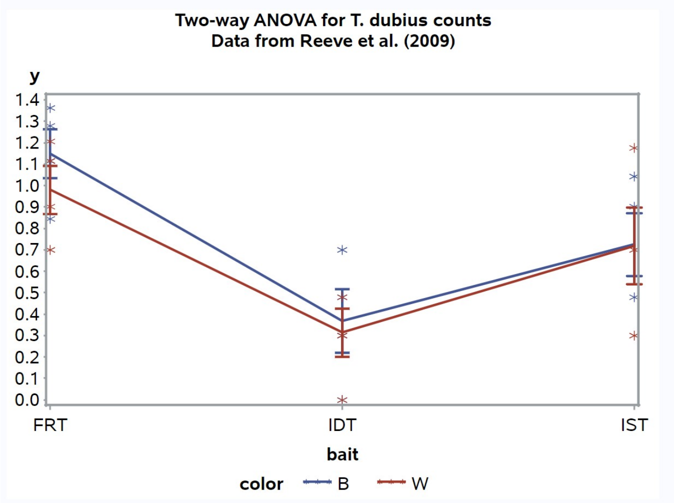
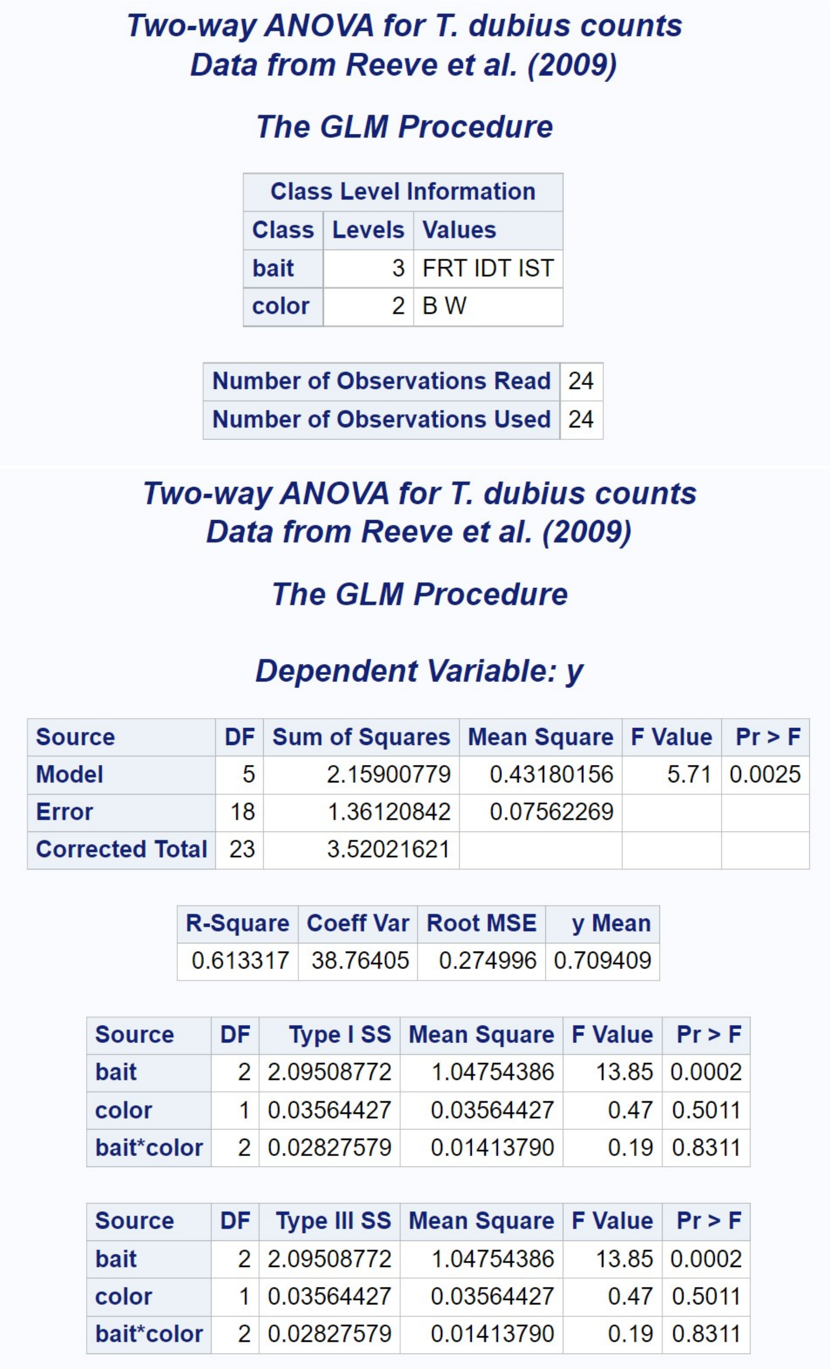


Figure 14.8: `Tdubius_bait_color.sas - proc gplot`



Two-way ANOVA for T. dubius counts
Data from Reeve et al. (2009)

The GLM Procedure

Class Level Information		
Class	**Levels**	**Values**
bait	3	FRT IDT IST
color	2	B W

Number of Observations Read	24
Number of Observations Used	24

Two-way ANOVA for T. dubius counts
Data from Reeve et al. (2009)

The GLM Procedure

Dependent Variable: y

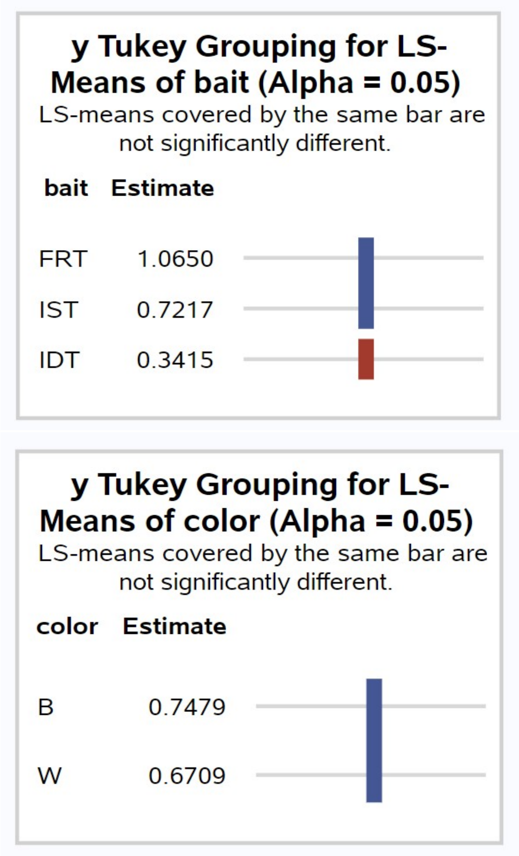
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	2.15900779	0.43180156	5.71	0.0025
Error	18	1.36120842	0.07562269		
Corrected Total	23	3.52021621			

R-Square	Coeff Var	Root MSE	y Mean
0.613317	38.76405	0.274996	0.709409

Source	DF	Type I SS	Mean Square	F Value	Pr > F
bait	2	2.09508772	1.04754386	13.85	0.0002
color	1	0.03564427	0.03564427	0.47	0.5011
bait*color	2	0.02827579	0.01413790	0.19	0.8311

Source	DF	Type III SS	Mean Square	F Value	Pr > F
bait	2	2.09508772	1.04754386	13.85	0.0002
color	1	0.03564427	0.03564427	0.47	0.5011
bait*color	2	0.02827579	0.01413790	0.19	0.8311

Figure 14.9: `Tdubius_bait_color.sas - proc glm`



y Tukey Grouping for LS-Means of bait (Alpha = 0.05)

LS-means covered by the same bar are not significantly different.

bait	Estimate
FRT	1.0650
IST	0.7217
IDT	0.3415

y Tukey Grouping for LS-Means of color (Alpha = 0.05)

LS-means covered by the same bar are not significantly different.

color	Estimate
B	0.7479
W	0.6709

Figure 14.10: `Tdubius_bait_color.sas - proc glm`

14.3.4 Two-way ANOVA for Example 2 - SAS demo

We next analyze the Example 2 data set using SAS. These data involve the total biomass of grass plants grown in small containers, where the treatments are nitrogen or water availability. The SAS program is similar to the previous example but with different variable names. Examining the output in Fig. 14.13, we see that the nitrogen × water interaction was highly significant ($F_{4,27} = 11.31, P < 0.0001$). The interaction can be observed in Fig. 14.12, which shows that the lines connecting the treatments are not parallel. Note that the greatest response of biomass to nitrogen occurred at the highest water level, while the response was minimal at the lowest level (Maestre & Reynolds (2007). Thus, low water levels apparently prevent growth even when nitrogen is abundant.

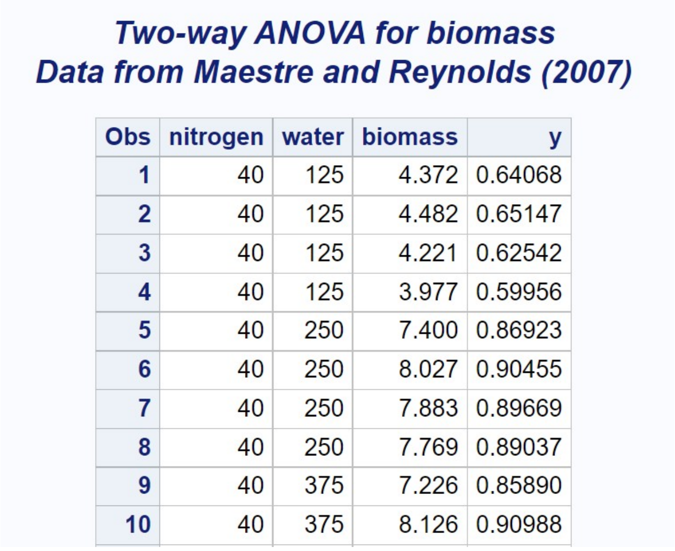
The analysis also found highly significant main effects of nitrogen ($F_{2,27} = 64.28, P < 0.0001$) and water ($F_{2,27} = 456.46, P < 0.0001$) on biomass (Fig. 14.13). There were also significant differences between every nitrogen or water treatment (Fig. 14.14). We can judge the relative strength of these effects by examining Fig. 14.12 as well as their sum of squares values, which are a measure of the amount of variation explained by each effect. They suggest that watering had the most effect on biomass, followed by nitrogen and the nitrogen × water interaction.

SAS Program

```
* Maestre_biomass.sas;
title "Two-way ANOVA for total biomass";
title2 "Data from Maestre and Reynolds (2007)";
data maestre;
    input nitrogen water biomass;
    * Apply transformations here;
    y = log10(biomass);
    datalines;
 40  125   4.372
 40  125   4.482
 40  125   4.221
 40  125   3.977
 40  250   7.400
 40  250   8.027
 40  250   7.883
 40  250   7.769

etc.

;
run;
* Print data set;
proc print data=maestre;
run;
* Plot means, standard errors, and observations;
proc gplot data=maestre;
    plot y*nitrogen=water / vaxis=axis1 haxis=axis1 legend=legend1;
    symbol1 i=std1mjt v=star height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
    legend1 label=(height=2) value=(height=2);
run;
* Two-way ANOVA with all fixed effects;
proc glm plots=diagnostics data=maestre;
    class nitrogen water;
    model y = nitrogen water nitrogen*water;
    lsmeans nitrogen water / adjust=tukey cl lines;
run;
quit;
```



Two-way ANOVA for biomass
Data from Maestre and Reynolds (2007)

Obs	nitrogen	water	biomass	y
1	40	125	4.372	0.64068
2	40	125	4.482	0.65147
3	40	125	4.221	0.62542
4	40	125	3.977	0.59956
5	40	250	7.400	0.86923
6	40	250	8.027	0.90455
7	40	250	7.883	0.89669
8	40	250	7.769	0.89037
9	40	375	7.226	0.85890
10	40	375	8.126	0.90988

etc.

Figure 14.11: `Maestre_biomass.sas.sas - proc print`

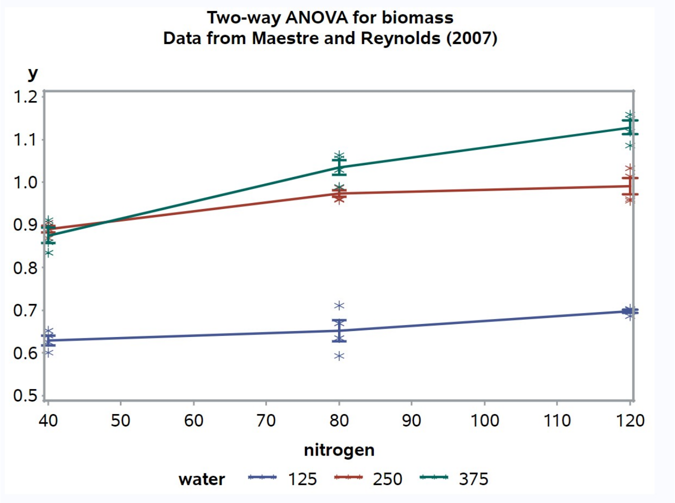
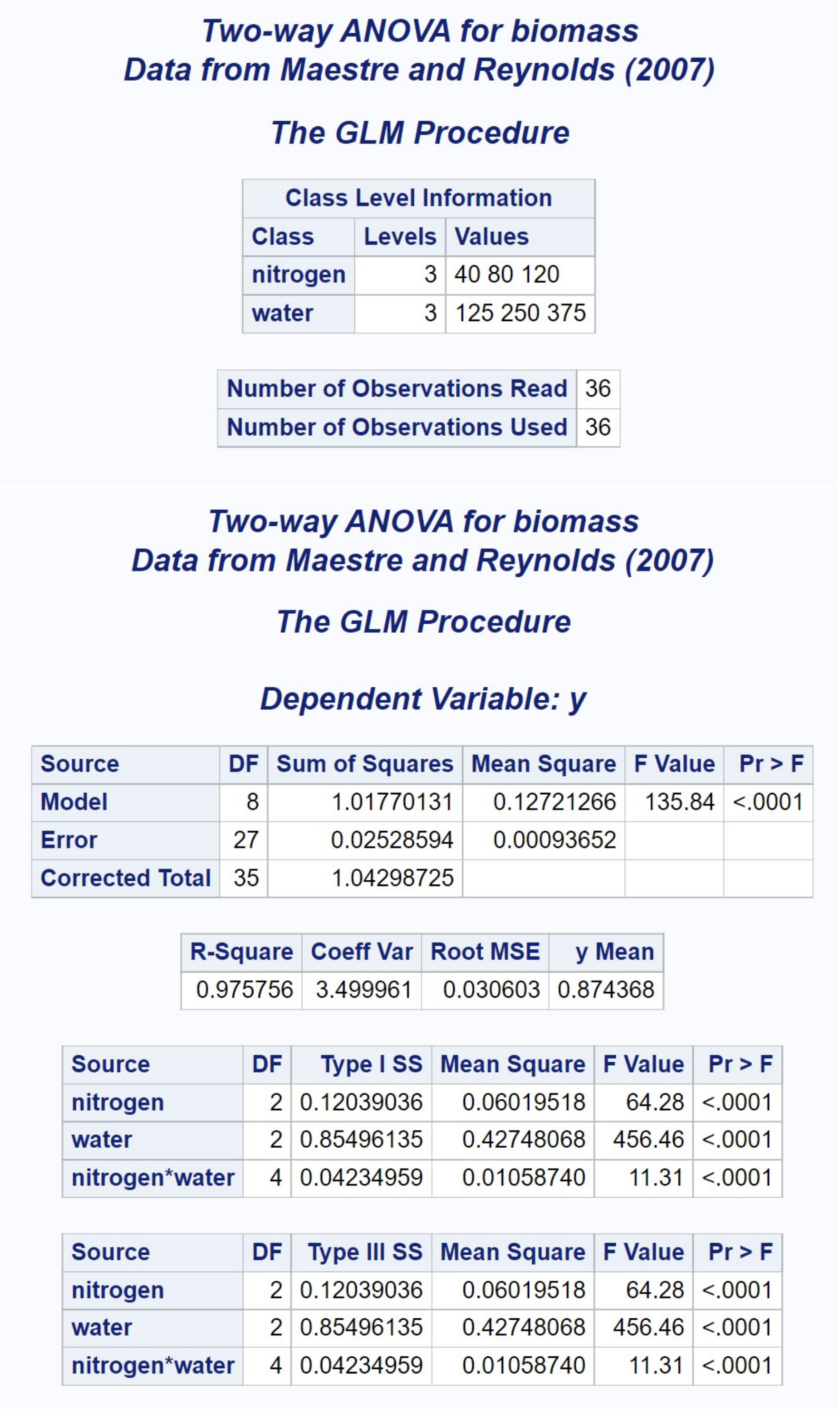


Figure 14.12: `Maestre_biomass.sas.sas - proc gplot`



Two-way ANOVA for biomass
Data from Maestre and Reynolds (2007)

The GLM Procedure

Class Level Information		
Class	Levels	Values
nitrogen	3	40 80 120
water	3	125 250 375

Number of Observations Read	36
Number of Observations Used	36

Two-way ANOVA for biomass
Data from Maestre and Reynolds (2007)

The GLM Procedure

Dependent Variable: y

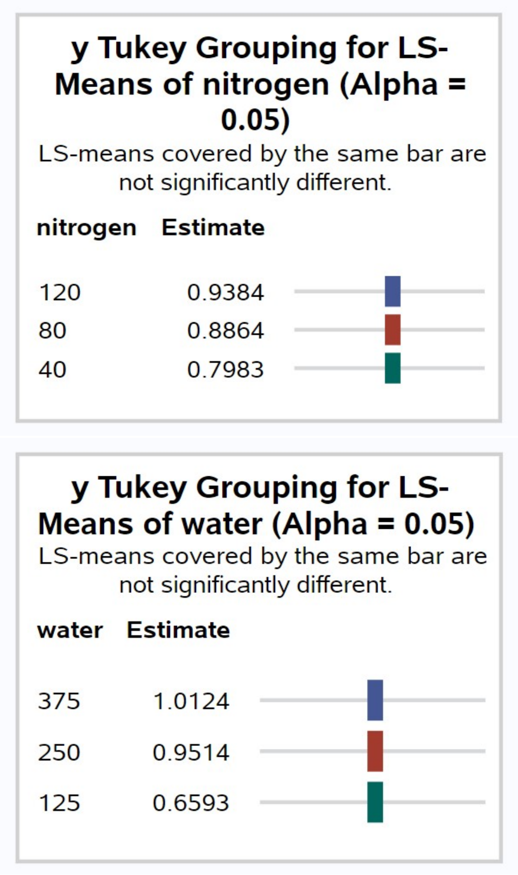
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	1.01770131	0.12721266	135.84	<.0001
Error	27	0.02528594	0.00093652		
Corrected Total	35	1.04298725			

R-Square	Coeff Var	Root MSE	y Mean
0.975756	3.499961	0.030603	0.874368

Source	DF	Type I SS	Mean Square	F Value	Pr > F
nitrogen	2	0.12039036	0.06019518	64.28	<.0001
water	2	0.85496135	0.42748068	456.46	<.0001
nitrogen*water	4	0.04234959	0.01058740	11.31	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
nitrogen	2	0.12039036	0.06019518	64.28	<.0001
water	2	0.85496135	0.42748068	456.46	<.0001
nitrogen*water	4	0.04234959	0.01058740	11.31	<.0001

Figure 14.13: `Maestre_biomass.sas - proc glm`



y Tukey Grouping for LS-Means of nitrogen (Alpha = 0.05)

LS-means covered by the same bar are not significantly different.

nitrogen	Estimate	
120	0.9384	🟦
80	0.8864	🟥
40	0.7983	🟩

y Tukey Grouping for LS-Means of water (Alpha = 0.05)

LS-means covered by the same bar are not significantly different.

water	Estimate	
375	1.0124	🟦
250	0.9514	🟥
125	0.6593	🟩

Figure 14.14: `Maestre_biomass.sas - proc glm`

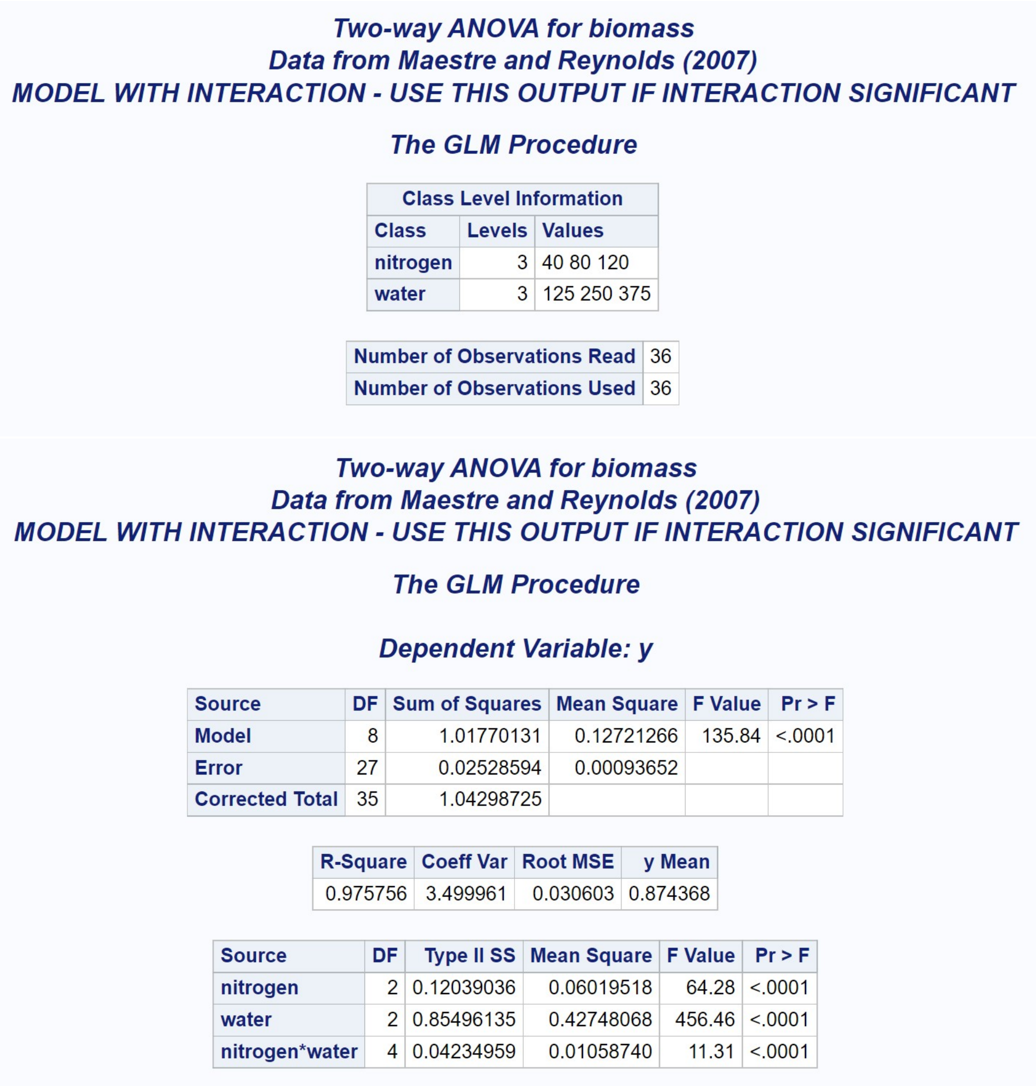
14.3.5 Tests for main effects with interaction

There is disagreement among statisticians on whether tests of the main effects are appropriate when there is significant interaction. Two different procedures have been developed. The SAS one involves fitting models with and without a given main effect, but always including interaction terms, yielding what SAS calls Type III sums of squares and tests (Speed et al. 1978, Shaw & Mitchell-Olds 1993, SAS Institute Inc. 2018). This has the benefit of generating tests for the interaction and main effects in a single pass (see preceding SAS demo). However, there are authors that believe tests of the main effects are questionable in the presence of interaction (e.g., Cox 1984, Winer et al. 1991, Stewart-Oaten 1995). One issue is whether a model with interaction but lacking a main effect is even plausible (Stewart-Oaten 1995). These considerations motivate a different procedure. The first step is to examine the test for interaction using the full two-way ANOVA model. If interaction appears weak or absent, there are two alternate ways of testing the main effects. One is to drop the interaction and rerun the model, examining the main effects in the usual fashion. Another method is to use what SAS calls Type II sums of squares, obtained using the option `\ss2` in the `model` statement. The tests based on these sums of squares assume there is no interaction. If the interaction is significant the main effects tests are ignored, although one can still test for Factor A effects at each level of Factor B, or vice versa (Winer et al. 1991). These are called tests of **simple effects**, and can be conducted using the SAS `slice` option for `lsmeans`.

The modified SAS code to implement these procedures is listed below, along with the corresponding output for the Example 2 data set. We see that the nitrogen × water interaction was highly significant ($F_{4,27} = 11.31, P < 0.0001$), and so we skip the tests of the main effects (Fig. 14.15). Note that the main effects sum of squares are identical to our previous ones using SAS Type III tests, but this would only be true for the special case of balanced designs with equal n for each treatment (see next section for unbalanced designs). The `slice` option is used to test for a nitrogen effect at every level of water, and vice versa (Fig. 14.16). We see that the effect of nitrogen was significant at the lowest water level, while highly significant at the other two levels. It appears the nitrogen effect was smaller at low water levels (see Fig. 14.12). The water effect was highly significant at every level of nitrogen.

SAS Program

```
* Two-way ANOVA with interaction;
title3 "MODEL WITH INTERACTION - USE THIS OUTPUT IF INTERACTION SIGNIFICANT";
proc glm plots=diagnostics data=maestre;
    class nitrogen water;
    model y = nitrogen water nitrogen*water / ss2;
    lsmeans nitrogen*water / slice=water slice=nitrogen;
run;
* Two-way ANOVA without interaction;
title3 "MODEL WITHOUT INTERACTION - USE THIS OUTPUT IF INTERACTION NS";
proc glm data=maestre;
    class nitrogen water;
    model y = nitrogen water / ss2;
    lsmeans nitrogen water / adjust=tukey cl lines;
run;
```



Two-way ANOVA for biomass
Data from Maestre and Reynolds (2007)
MODEL WITH INTERACTION - USE THIS OUTPUT IF INTERACTION SIGNIFICANT

The GLM Procedure

Class Level Information		
Class	Levels	Values
nitrogen	3	40 80 120
water	3	125 250 375

Number of Observations Read	36
Number of Observations Used	36

Two-way ANOVA for biomass
Data from Maestre and Reynolds (2007)
MODEL WITH INTERACTION - USE THIS OUTPUT IF INTERACTION SIGNIFICANT

The GLM Procedure

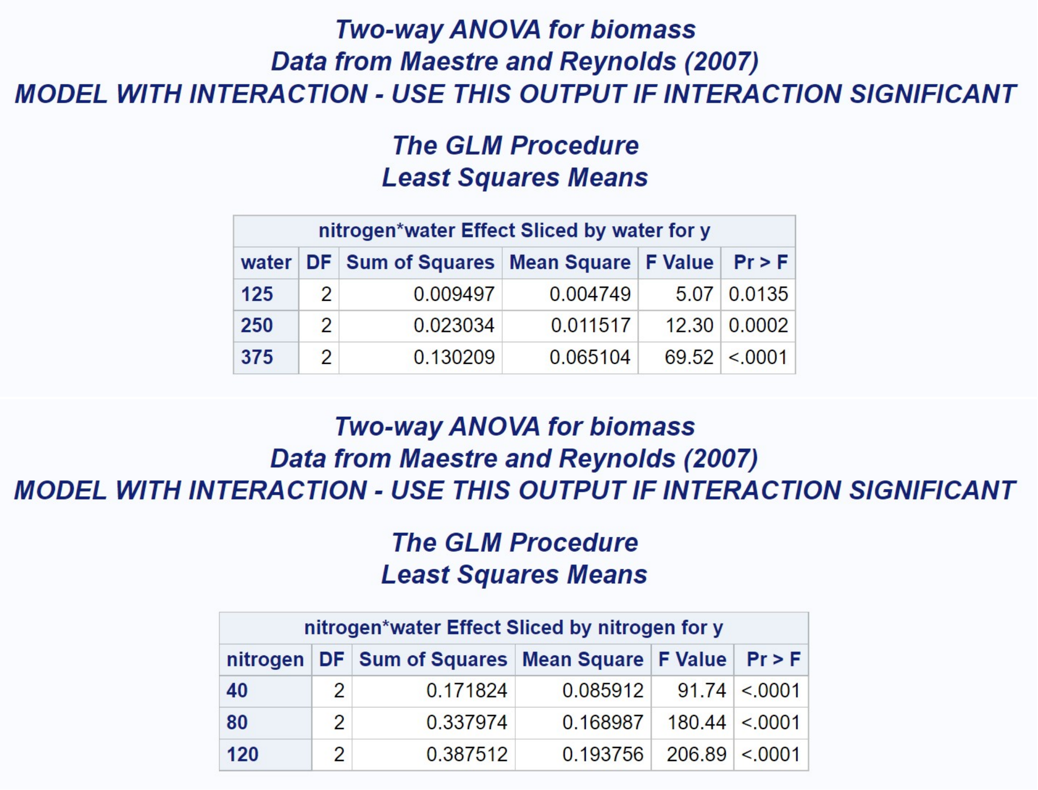
Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	8	1.01770131	0.12721266	135.84	<.0001
Error	27	0.02528594	0.00093652		
Corrected Total	35	1.04298725			

R-Square	Coeff Var	Root MSE	y Mean
0.975756	3.499961	0.030603	0.874368

Source	DF	Type II SS	Mean Square	F Value	Pr > F
nitrogen	2	0.12039036	0.06019518	64.28	<.0001
water	2	0.85496135	0.42748068	456.46	<.0001
nitrogen*water	4	0.04234959	0.01058740	11.31	<.0001

Figure 14.15: `Maestre_biomass2_new.sas - proc glm`



Two-way ANOVA for biomass
Data from Maestre and Reynolds (2007)
MODEL WITH INTERACTION - USE THIS OUTPUT IF INTERACTION SIGNIFICANT

The GLM Procedure
Least Squares Means

nitrogen*water Effect Sliced by water for y					
water	DF	Sum of Squares	Mean Square	F Value	Pr > F
125	2	0.009497	0.004749	5.07	0.0135
250	2	0.023034	0.011517	12.30	0.0002
375	2	0.130209	0.065104	69.52	<.0001

Two-way ANOVA for biomass
Data from Maestre and Reynolds (2007)
MODEL WITH INTERACTION - USE THIS OUTPUT IF INTERACTION SIGNIFICANT

The GLM Procedure
Least Squares Means

nitrogen*water Effect Sliced by nitrogen for y					
nitrogen	DF	Sum of Squares	Mean Square	F Value	Pr > F
40	2	0.171824	0.085912	91.74	<.0001
80	2	0.337974	0.168987	180.44	<.0001
120	2	0.387512	0.193756	206.89	<.0001

Figure 14.16: `Maestre_biomass2_new.sas - proc glm`

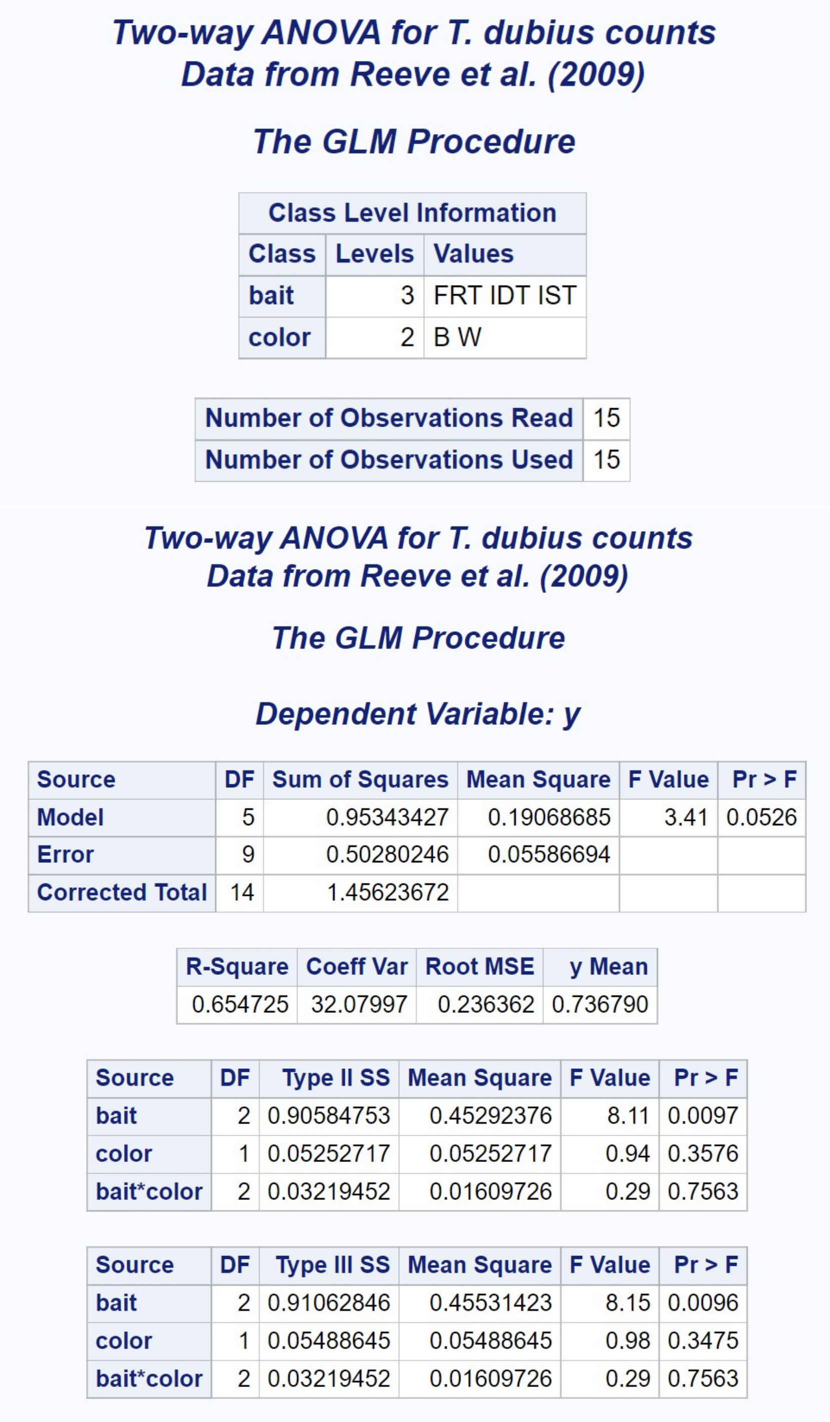
14.4 Unbalanced designs and two-way ANOVA

The examples we have examined so far are balanced designs, with equal numbers of observations in each treatment combination. For these designs, the various sums of squares are independent and additive ($SS_A + SS_B + SS_{AB} + SS_{within} = SS_{total}$), the different methods of calculating the sum of squares (Type I, II, and III) yield the same results, and the resulting tests are the same. This is not the case for unbalanced two-way (or higher) designs, which occur frequently in practice. These are designs where there are fewer observations in some treatments than others, possible only a single observation. These designs can be analyzed using the same SAS procedures and programs as before, but the various sums of squares are no longer additive, and the tests are not independent (Shaw & Mitchell-Olds 1993). For this reason, if the lack of balance is severe the analysis should be interpreted with some caution.

We will use the Example 2 data set, with nine observations removed, to illustrate the analysis of unbalanced designs (see Table 14.5). The number of observations varies from $n = 1$ to 4 across treatments. These data can be analyzed using the same program as before. To show the results for both Type II and III sums of squares, the option `\ ss2 ss3` was added to the `model` statement. Examining the output (Fig. 14.17), we see that the bait × trap color interaction was nonsignificant ($F_{2,9} = 0.29, P = 0.7563$). The color effect was also nonsignificant (Type II: $F_{1,9} = 0.94, P = 0.3576$, Type III: $F_{1,9} = 0.98, P = 0.3475$), but the bait effect was highly significant (Type II: $F_{2,9} = 8.11, P < 0.0097$, Type III: $F_{2,9} = 8.15, P = 0.0096$). This is basically the same result as we obtained earlier for this study, despite the lack of balance. We can also see that the sums of squares are no longer additive. For example, with Type III sums of squares we have $SS_A + SS_B + SS_{AB} + SS_{within} = 0.9106 + 0.0549 + 0.0322 + 0.5028 = 1.5005$. This does not equal $SS_{total} = 1.4562$.

Table 14.5: Example 2 - Unbalanced design.

Bait	Color	*T. dubius*
FRT	B	18
FRT	W	12
FRT	W	15
FRT	W	7
FRT	W	4
IDT	B	2
IDT	B	1
IDT	B	4
IDT	W	2
IDT	W	1
IST	B	2
IST	B	2
IST	B	10
IST	B	7
IST	W	4



Two-way ANOVA for T. dubius counts
Data from Reeve et al. (2009)

The GLM Procedure

Class Level Information		
Class	Levels	Values
bait	3	FRT IDT IST
color	2	B W

Number of Observations Read	15
Number of Observations Used	15

Two-way ANOVA for T. dubius counts
Data from Reeve et al. (2009)

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	0.95343427	0.19068685	3.41	0.0526
Error	9	0.50280246	0.05586694		
Corrected Total	14	1.45623672			

R-Square	Coeff Var	Root MSE	y Mean
0.654725	32.07997	0.236362	0.736790

Source	DF	Type II SS	Mean Square	F Value	Pr > F
bait	2	0.90584753	0.45292376	8.11	0.0097
color	1	0.05252717	0.05252717	0.94	0.3576
bait*color	2	0.03219452	0.01609726	0.29	0.7563

Source	DF	Type III SS	Mean Square	F Value	Pr > F
bait	2	0.91062846	0.45531423	8.15	0.0096
color	1	0.05488645	0.05488645	0.98	0.3475
bait*color	2	0.03219452	0.01609726	0.29	0.7563

Figure 14.17: `Tdubius_bait_color_unbalanced.sas - proc glm`

14.5 Two-way ANOVA without replication

The designs we have examined so far assume there are multiple observations for each treatment combination, implying $n > 1$ for each group. However, it is possible to analyze studies where there is only replicate per group ($n = 1$) although this requires a change in the model. With so little data, it is not possible to estimate the interaction terms nor easily conduct a test for the interaction. However, we can fit a simplified model of the form

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}. \tag{14.42}$$

Note that the interaction term is absent. In addition, we no longer need the third subscript k for the observations because there is only one observation per treatment group. One can visualize the behavior of this model using the same figures as for the two-way model with replication (see Fig. 14.3-14.5), except that the model does not incorporate interaction.

It is important to realize that interaction could still be present in the data, even though we cannot test for it using this model. If interaction is present it will reduce the power to detect main effects, because it adds variability to the observations in a way not accounted for by the model. Even if interaction is absent, this design will obviously have less power than a design with replication.

For these designs, we will be interested in testing whether Factor A or B have an effect on the groups means. For Factor A, this amounts to testing H_0 : all $\alpha_i = 0$, while for Factor B we would test H_0 : all $\beta_j = 0$. No test of this type is possible for the interaction.

14.5.1 Hypothesis testing

The sums of squares, mean squares, and other quantities for two-way ANOVA without replication are similar to those for designs with replication. We will illustrate the calculations using another data set for the insect predator *T. dubius* (Example 3, Table 14.6). This predator is most abundant during cool periods of the year in the southern USA, possibly because it cannot tolerate high temperatures. A study was conducted to see how temperature (which we call Factor A) and relative humidity (Factor B) affect the mortality rate of its eggs in the laboratory (Reeve 2000). Eggs and environmental chambers were in short supply, however, so only a single replicate was conducted at each temperature and humidity combination. Six temperatures

$(15°, 20°, 25°, 30°, 35°$, and 37.5°C) and three relative humidity treatments (55%, 75%, and 100%) were used. This corresponds to $a = 6$ and $b = 3$ in the formulas below. An arcsine-square root transformation was applied to the mortality rate observations, a common practice for data in the form of proportions.

We begin by calculating a mean corresponding to each level of Factor A by averaging across the levels of Factor B, which we denote as $\bar{Y}_{i\cdot}$. It can be calculated using the formula

$$\bar{Y}_{i\cdot} = \frac{\sum_{j=1}^{b} Y_{ij}}{b}. \tag{14.43}$$

For example, we have

$$\bar{Y}_{1\cdot} = \frac{0.379 + 0.325 + 0.615}{3} = 0.440 \tag{14.44}$$

for the first temperature treatment (15°C) in Example 3. The means for other temperature values are given in Table 14.6. We similarly can find means corresponding to each level of Factor B by averaging across the levels of Factor A. The general formula is

$$\bar{Y}_{\cdot j} = \frac{\sum_{i=1}^{a} Y_{ij}}{a}. \tag{14.45}$$

For the first humidity treatment in Example 3, we have

$$\bar{Y}_{\cdot 1} = \frac{0.379 + 0.439 + 0.358 + 0.466 + 0.970 + 1.571}{6} = 0.697. \tag{14.46}$$

The means for the other humidity treatments are $\bar{Y}_{\cdot 2} = 0.731$ and $\bar{Y}_{\cdot 3} = 0.719$. A grand mean $\bar{\bar{Y}}$ can then be calculated by averaging across the values of $\bar{Y}_{i\cdot}$ or equivalently by summing all the observations and dividing by their total number. It can be generally calculated using the formula

$$\bar{\bar{Y}} = \frac{\sum_{i=1}^{a} \bar{Y}_{i\cdot}}{a}. \tag{14.47}$$

For the Example 3 data set, we have

$$\bar{\bar{Y}} = \frac{0.440 + 0.502 + 0.454 + 0.521 + 0.806 + 1.571}{6} = 0.716. \tag{14.48}$$

We next develop sums of squares and means squares for this design. The difference $\bar{Y}_{i\cdot} - \bar{\bar{Y}}$ is a measure of the shift generated by Factor A in the observations, and also estimates α_i. Squaring and summing them across all the levels of Factor A, we obtain SS_A. It is calculated using the general formula

$$SS_A = b\sum_{i=1}^{a}(\bar{Y}_{i\cdot} - \bar{\bar{Y}})^2. \tag{14.49}$$

SS_A has $a-1$ degrees of freedom. Its mean square is calculated using the formula

$$MS_A = \frac{SS_A}{a-1}. \tag{14.50}$$

Note the factor b in the expression for SS_A, which as usual scales MS_A so that it estimates σ^2 under H_0. For the Example 3 data, we have

$$\begin{aligned} SS_A &= 3\left[(0.440 - 0.716)^2 + (0.502 - 0.716)^2 + \cdots + (1.571 - 0.716)^2\right] && (14.51)\\ &= 3\left[0.076176 + 0.045796 + 0.068644 + 0.038025 + 0.008100 + 0.731025\right] && (14.52)\\ &= 2.903298 && (14.53)\\ & && (14.54) \end{aligned}$$

and

$$MS_A = \frac{2.903298}{6-1} = 0.580660. \tag{14.55}$$

We similarly define SS_B using the general formula

$$SS_B = a\sum_{j=1}^{b}(\bar{Y}_{\cdot j} - \bar{\bar{Y}})^2. \tag{14.56}$$

SS_B has $b-1$ degrees of freedom. We can then calculate a mean square for Factor B using the formula

$$MS_B = \frac{SS_B}{b-1}. \tag{14.57}$$

For the Example 3 data, we have

$$\begin{aligned} SS_B &= 6\left[(0.697 - 0.716)^2 + (0.731 - 0.716)^2 + (0.719 - 0.716)^2\right] && (14.58)\\ &= 6\left[0.000361 + 0.000225 + 0.000009\right] && (14.59)\\ &= 0.003570. && (14.60) \end{aligned}$$

and

$$MS_B = \frac{0.003570}{3-1} = 0.001785. \tag{14.61}$$

We now need a measure of the variability of the observations. We previously used SS_{within} for this purpose, which measured the variability of the observations within each treatment group. However, in two-way designs without replication there is only a single observation in these groups ($n = 1$). If we assume there is no interaction, however, we can use an interaction-like sum of squares as a measure of variability. In particular, we have

$$SS_{within} = \sum_{i=1}^{a} \sum_{j=1}^{b} (Y_{ij} - \bar{Y}_{i\cdot} - \bar{Y}_{\cdot j} + \bar{\bar{Y}})^2. \tag{14.62}$$

The squared terms within this expression measure the difference between the one observation for each treatment combination and the values predicted by the model without any interaction. Note the similarity to SS_{AB} for designs with replication. SS_{within} has $(a-1)(b-1)$ degrees of freedom, and the associated mean square is defined by the formula

$$MS_{within} = \frac{SS_{within}}{(a-1)(b-1)}. \tag{14.63}$$

The last column of Table 14.6 shows the preliminary calculations for SS_{within}. Adding this column across all the treatment groups yields

$$SS_{within} = 0.131334 \tag{14.64}$$

and

$$MS_{within} = \frac{0.131334}{(6-1)(3-1)} = 0.013133. \tag{14.65}$$

The total sum of squares is given by the formula

$$SS_{total} = \sum_{i=1}^{a} \sum_{j=1}^{b} (Y_{ij} - \bar{\bar{Y}})^2 \tag{14.66}$$

and has $ab-1$ degrees of freedom. For Example 3, we calculate that $SS_{total} = 3.038202$ with 17 degrees of freedom.

As before, we can organize the different sum of squares and mean squares into an ANOVA table. Table 14.7 shows the general layout of such a table for two-way designs without replication. We use $F_s = MS_A/MS_{within}$ to test for the effect of Factor A. Under H_0 : all $\alpha_i = 0$ this statistic has an F distribution with $df_1 = a - 1$ and $df_2 = (a-1)(b-1)$. Similarly, we use $F_s = MS_B/MS_{within}$ to test for an effect of Factor B. Under H_0 : all $\beta_j = 0$ it has an F distribution with $df_1 = b - 1$ and $df_2 = (a-1)(b-1)$.

Table 14.8 shows the results for the Example 3 data set, including the F statistics and P values obtained using Table F. The temperature effect was highly significant ($F_{5,10} = 44.214, P < 0.001$) while humidity was nonsignificant ($F_{2,10} = 0.136, P > 0.100$). Examining the data in Table 14.6, we see that mortality rates sharply increased as temperature increased.

Table 14.6: Example 3 - Effect of temperature and relative humidity on the mortality rate of *T. dubius* eggs. Also shown are the means for each temperature level ($\bar{Y}_{i.}$) and preliminary calculations to find SS_{within}

Temp. (°C)	Humidity (%)	Mortality	$Y_{ij} = \sin^{-1}(\sqrt{\text{Mortality}})$	i	j	$\bar{Y}_{i.}$	$(Y_{ij} - \bar{Y}_{i.} - \bar{Y}_{.j} + \bar{\bar{Y}})^2$
15	55	0.137	0.379	1	1		0.001764
15	75	0.102	0.325	1	2	0.440	0.016900
15	100	0.333	0.615	1	3		0.029584
20	55	0.181	0.439	2	1		0.001936
20	75	0.337	0.619	2	2	0.502	0.010404
20	100	0.188	0.448	2	3		0.003249
25	55	0.123	0.358	3	1		0.005929
25	75	0.259	0.534	3	2	0.454	0.004225
25	100	0.205	0.470	3	3		0.000169
30	55	0.202	0.466	4	1		0.001296
30	75	0.321	0.602	4	2	0.521	0.004356
30	100	0.226	0.495	4	3		0.000841
35	55	0.680	0.970	5	1		0.033489
35	75	0.447	0.732	5	2	0.806	0.007921
35	100	0.431	0.716	5	3		0.008649
37.5	55	1.000	1.571	6	1		0.000361
37.5	75	1.000	1.571	6	2	1.571	0.000225
37.5	100	1.000	1.571	6	3		0.000036

Table 14.7: General ANOVA table for two-way designs without replication, showing formulas for different mean squares and F tests.

Source	df	Sum of squares	Mean square	F_s
Factor A	$a-1$	SS_A	$MS_A = SS_A/(a-1)$	MS_A/MS_{within}
Factor B	$b-1$	SS_B	$MS_B = SS_B/(b-1)$	MS_B/MS_{within}
Within	$(a-1)(b-1)$	SS_{within}	$MS_{within} = SS_{within}/(a-1)(b-1)$	
Total	$ab-1$	SS_{total}		

Table 14.8: ANOVA table for the Example 3 data set, including P values for the tests.

Source	df	Sum of squares	Mean square	F_s	P
Temperature	5	2.903298	0.580660	44.214	< 0.001
Humidity	2	0.003570	0.001785	0.136	> 0.100
Within	10	0.131334	0.013133		
Total	17	3.038198			

14.5.2 Two-way ANOVA no replication - SAS demo

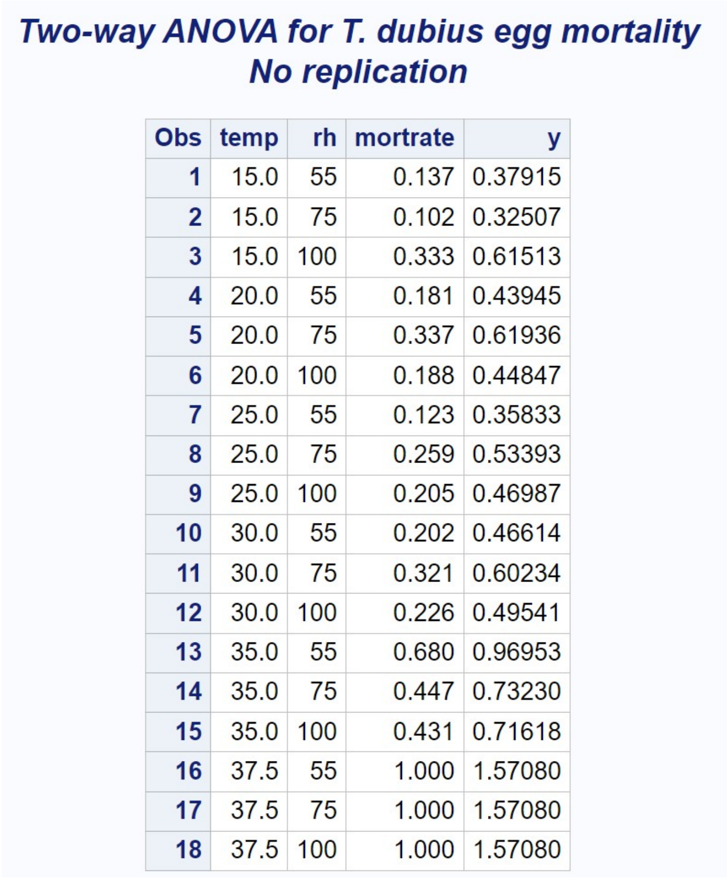
We now analyze these same data using SAS. The program is similar to previous ones for two-way designs with replication, except that the interaction term needs to be deleted from the `model` statement. Because there are several levels of temperature (`temp`) and relative humidity (`rh`) in the experimental design, it seems reasonable to use multiple comparisons to compare the different groups using an `lsmeans` statement. See SAS program and output below.

Examining Fig. 14.20, we see a highly significant effect of temperature on egg mortality ($F_{5,10} = 44.31, P < 0.0001$), while the effect of humidity was nonsignificant ($F_{2,10} = 0.13, P = 0.8777$). The results are similar to the manual calculations in Table 14.6. The Tukey procedure (Fig. 14.21) finds that 37.5°C was significantly different from all the other temperatures, while 35°C was significantly different from 15°C and 25°C. No other differences were significant. There were also no significant differences among the humidity treatments. Figure 14.19 suggests that mortality was constant up to 30°C, then rapidly increased.

SAS Program

```
* Clerid_eggs_th.sas;
title "Two-way ANOVA for T. dubius egg mortality";
title2 "No replication";
data mortality;
    input temp rh mortrate;
    * Apply transformations here;
    y = arsin(sqrt(mortrate));
    datalines;
15  55  0.137
15  75  0.102
15 100  0.333
20  55  0.181
20  75  0.337
20 100  0.188
25  55  0.123
25  75  0.259
25 100  0.205
30  55  0.202
30  75  0.321
30 100  0.226
35  55  0.680
35  75  0.447
```

```
35 100  0.431
37.5 55 1.000
37.5 75 1.000
37.5 100 1.000
;
run;
* Print data set;
proc print data=mortality;
run;
* Plot means, standard errors, and observations;
proc gplot data=mortality;
    plot y*temp=rh / vaxis=axis1 haxis=axis1 legend=legend1;
    symbol1 i=j v=star height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
    legend1 label=(height=2) value=(height=2);
run;
* Two-way ANOVA with all fixed effects;
proc glm plots=diagnostics data=mortality;
    class temp rh;
    model y = temp rh;
    lsmeans temp rh / adjust=tukey cl lines;
run;
quit;
```



Two-way ANOVA for T. dubius egg mortality
No replication

Obs	temp	rh	mortrate	y
1	15.0	55	0.137	0.37915
2	15.0	75	0.102	0.32507
3	15.0	100	0.333	0.61513
4	20.0	55	0.181	0.43945
5	20.0	75	0.337	0.61936
6	20.0	100	0.188	0.44847
7	25.0	55	0.123	0.35833
8	25.0	75	0.259	0.53393
9	25.0	100	0.205	0.46987
10	30.0	55	0.202	0.46614
11	30.0	75	0.321	0.60234
12	30.0	100	0.226	0.49541
13	35.0	55	0.680	0.96953
14	35.0	75	0.447	0.73230
15	35.0	100	0.431	0.71618
16	37.5	55	1.000	1.57080
17	37.5	75	1.000	1.57080
18	37.5	100	1.000	1.57080

etc.

Figure 14.18: `Clerid_eggs_th.sas - proc print`

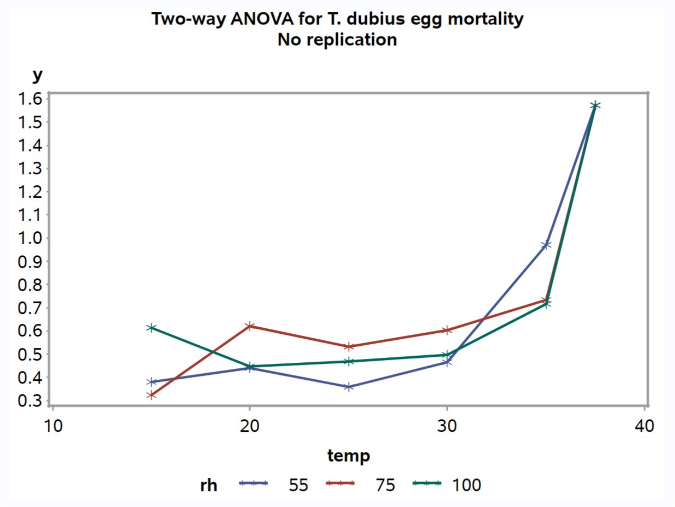
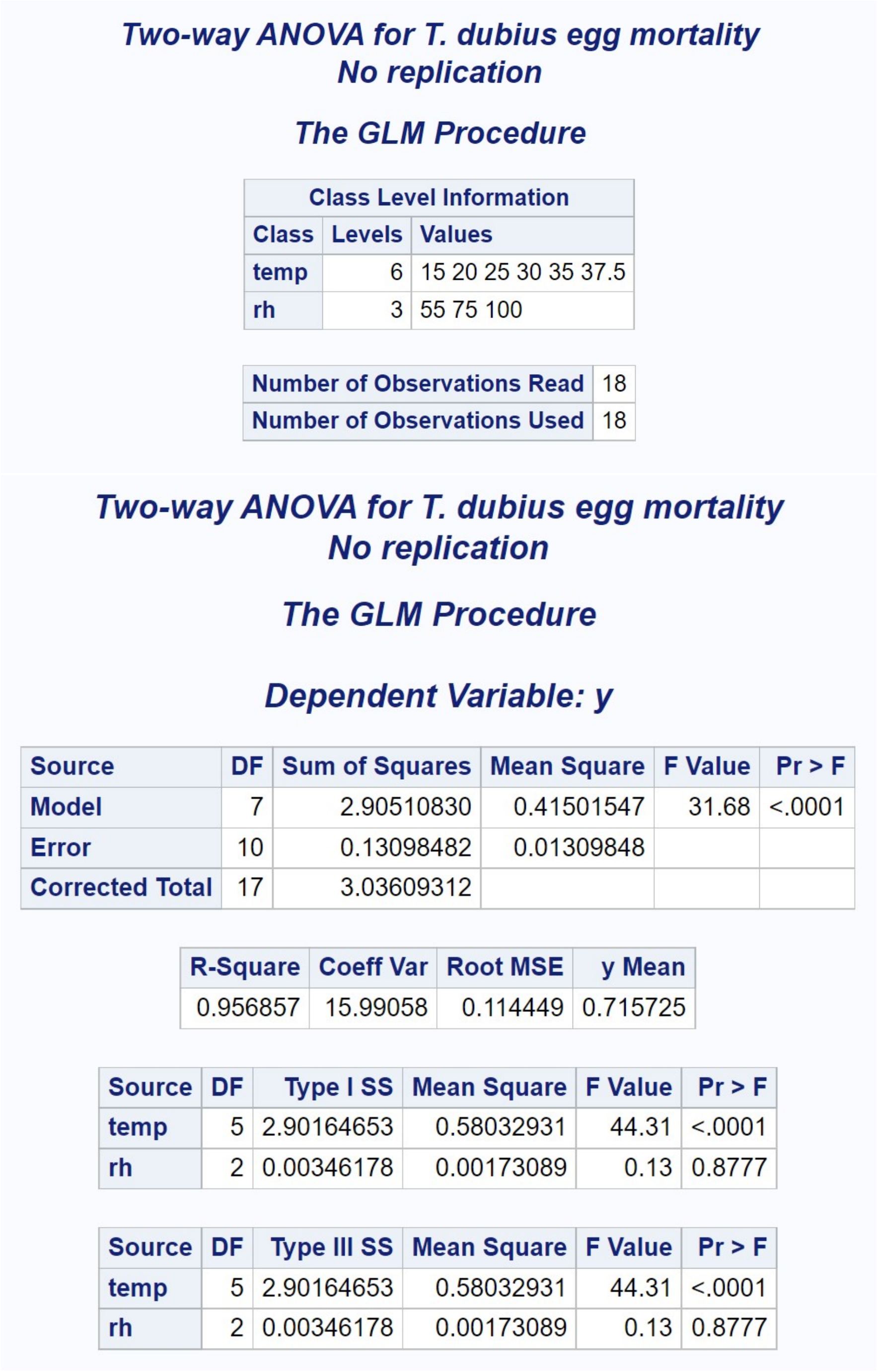


Figure 14.19: `Clerid_eggs_th.sas - proc gplot`



Two-way ANOVA for T. dubius egg mortality
No replication

The GLM Procedure

Class Level Information		
Class	**Levels**	**Values**
temp	6	15 20 25 30 35 37.5
rh	3	55 75 100

Number of Observations Read	18
Number of Observations Used	18

Two-way ANOVA for T. dubius egg mortality
No replication

The GLM Procedure

Dependent Variable: y

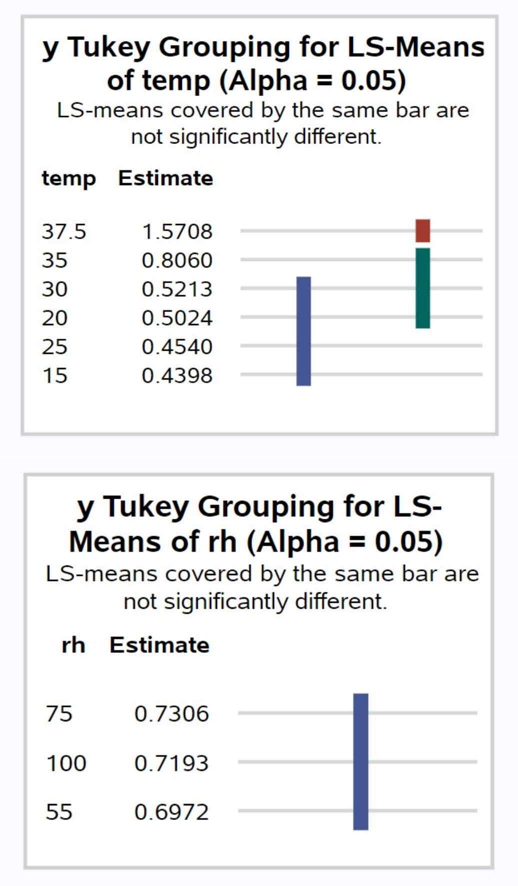
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	7	2.90510830	0.41501547	31.68	<.0001
Error	10	0.13098482	0.01309848		
Corrected Total	17	3.03609312			

R-Square	Coeff Var	Root MSE	y Mean
0.956857	15.99058	0.114449	0.715725

Source	DF	Type I SS	Mean Square	F Value	Pr > F
temp	5	2.90164653	0.58032931	44.31	<.0001
rh	2	0.00346178	0.00173089	0.13	0.8777

Source	DF	Type III SS	Mean Square	F Value	Pr > F
temp	5	2.90164653	0.58032931	44.31	<.0001
rh	2	0.00346178	0.00173089	0.13	0.8777

Figure 14.20: `Clerid_eggs_th.sas - proc glm`



y Tukey Grouping for LS-Means of temp (Alpha = 0.05)

LS-means covered by the same bar are not significantly different.

temp	Estimate
37.5	1.5708
35	0.8060
30	0.5213
20	0.5024
25	0.4540
15	0.4398

y Tukey Grouping for LS-Means of rh (Alpha = 0.05)

LS-means covered by the same bar are not significantly different.

rh	Estimate
75	0.7306
100	0.7193
55	0.6972

Figure 14.21: `Clerid_eggs_th.sas - proc glm`

14.6 Randomized block designs

Suppose that we are interested in the yield of five different strains (A, B, C, D, and E) of corn, with five replicates per strain. One possible design would be to randomly assign the strain treatments to 30 small plots scattered throughout a large field, in a completely randomized design (Fig. 14.22). The resulting data from this design could be analyzed using one-way ANOVA (Chapter 11), with strain as the treatment. One problem with this design is soil fertility, moisture, and other factors could vary across this large field. This spatial heterogeneity would make it more difficult to see any treatment effects because it would increase the variance among replicate plots.

A common two-way design, the **randomized block design**, provides a possible solution to this spatial heterogeneity problem. Suppose that soil fertility and moisture are more homogeneous on smaller spatial scales, as often seems to be true. We could then select six plots within this field, called **blocks**, and within sections of each block plant the five corn strains (see Fig. 14.23). The order of the different treatments within each block would be randomized, hence the name randomized blocks. This ensures that the sequence of treatments varies across blocks, and that each treatment has different strains for neighbors in each block. The resulting data would then be analyzed using a two-way model with a fixed treatment effect and a random block effect, which helps account and control for spatial heterogeneity in the system. The block is considered a random effect because the blocks are usually selected from a potentially large collection of possible blocks. **A statistical model with both fixed and random effects is called a mixed model.**

Another example of a randomized block design could be insect traps baited with different attractants, say A, B, C, D, and E. Different stands in the forest would be the blocks. Five traps would be deployed in each stand along a transect, with baits randomly assigned to the traps within the transect. In another type of randomized block design, the blocks are different times rather than locations in space. For example, suppose that we want to test six different diets for rearing fish in ponds, but only have six ponds available. We could randomly assign the diets to the ponds and conduct the experiment, obtaining one replicate of each treatment. We would then repeat the study several more times using the same ponds, with the treatments randomly assigned each time. Each time would be treated as a separate block in the analysis.

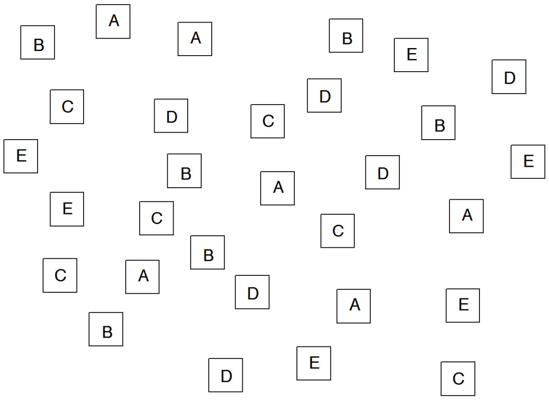


Figure 14.22: Completely randomized design with five treatments (A, B, C, D, and E) and six replicates per treatment.

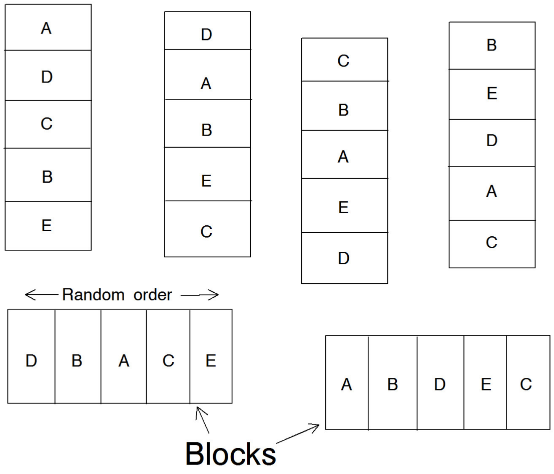


Figure 14.23: Randomized block design with five treatments (A, B, C, D, and E) and six blocks.

14.6.1 Randomized block models

There are two effects in a randomized block design, a fixed treatment and a random block effect, usually denoted as Factor A and B. The model commonly used to analyze these designs has the form

$$Y_{ij} = \mu + \alpha_i + B_j + \epsilon_{ij}. \tag{14.67}$$

Here μ, α_i, and ϵ_{ij} are defined as in previous models, while $B_j \sim N(0, \sigma^2_B)$. The model thus has two variance components, the variance among blocks (σ^2_B) and the variance of ϵ_{ij} (σ^2).

Note that there is no interaction term in this model, although there could be interaction in the data. A randomized block design has just one observation per combination of treatment and block, and so there are insufficient data to estimate an A × B interaction. However, there are variants of the randomized block design that have two or more replicates of each Factor A treatment per block. In this, case, we could fit a model with interaction of the form

$$Y_{ij} = \mu + \alpha_i + B_j + (\alpha B)_{ij} + \epsilon_{ij}. \tag{14.68}$$

Here $(\alpha B)_{ij} \sim N(0, \sigma^2_{AB})$. The interaction term in these designs is considered to be a random effect because it involves the random block effect. This model has three variance components, the interaction variance (σ^2_{AB}), the block variance (σ^2_B), and the variance of ϵ_{ij} (σ^2).

14.6.2 Hypothesis testing and variance components

We will use `proc mixed` in SAS to analyze the data for randomized block designs (SAS Institute Inc. 2018). The default method in SAS estimates the variance components in the model using a method called restricted maximum likelihood, or REML. This process involves separating the fixed effects from the likelihood function, then estimating the variance components of the random effects by maximizing this restricted likelihood (hence the name). Once these are determined, the fixed effects parameters are estimated and F tests generated for those effects (Littell et al. 1996, McCulloch & Searle 2001). For a randomized block design, the null hypothesis tested for Factor A would be H_0 : all $\alpha_i = 0$. However, there is no ANOVA table nor related quantities like sum of squares and mean squares. The emphasis in `proc mixed` is on the estimation of variance components rather than tests on them, although tests can be constructed if necessary (see below).

14.6.3 Randomized block design - SAS demo

We will illustrate a `proc mixed` analysis for the randomized block design using a different trapping study of *T. dubius* (Reeve et al. 2009). Six different stands were located in the forest and considered to be blocks. Five traps were placed in a line at 30 m intervals within each stand, and then a bait treatment randomly assigned to each trap. There were five such treatments: blank trap (`BLANK`), α-pinene (`AP`), frontalin + α-pinene (`FRAP`), ipsdienol + α-pinene (`IDAP`), and ipsenol + α-pinene (`ISAP`). As mentioned earlier, frontalin, ipsdienol, and ipsenol are bark beetle pheromones while α-pinene is a major component of pine resin. The number of predators caught in each trap was then counted. See SAS program with data below.

The count data were manipulated in two ways before analysis. A log transformation was applied to predator counts to ensure the observations meet the assumptions of ANOVA (see Chapter 15). All observations for the `BLANK` treatment were also removed using the statement

```
if treat="BLANK" then delete;
```

because this treatment caught no insects. The `proc mixed` portion of the program basically implements the model for randomized block designs. We first need to tell SAS the variables categorizing the groups in the data, using a `class` statement. For the trapping study, the variables `treat` and `block` identify the treatment and block variables, so we use the statement

```
class treat block;
```

Next, recall that the randomized block model has the form

$$Y_{ij} = \mu + \alpha_i + B_j + \epsilon_{ij}. \tag{14.69}$$

Here, the SAS variable `treat` corresponds to α_i, the fixed effect in the model, while `block` corresponds to B_j, the random effect. One feature of `proc mixed` is the separation of fixed and random effects in the model – all fixed effects are placed in the `model` statement while random effects are included in the `random` statement. Thus, the `model` statement for the trapping data would be

```
model y = treat / ddfm=kr;
```

while the `random` statement is

```
random block;
```

The `ddfm=kr` option specifies the Kenward-Rogers method of calculating the degrees of freedom (SAS Institute Inc. 2018), a general method for calculating the degrees of freedom that works in a variety of circumstances. An `lsmeans` statement of the form

```
lsmeans treat / pdiff=all adjust=tukey adjdfe=row;
```

is also used to compare the different bait treatments using the Tukey method. See complete program listing and output below.

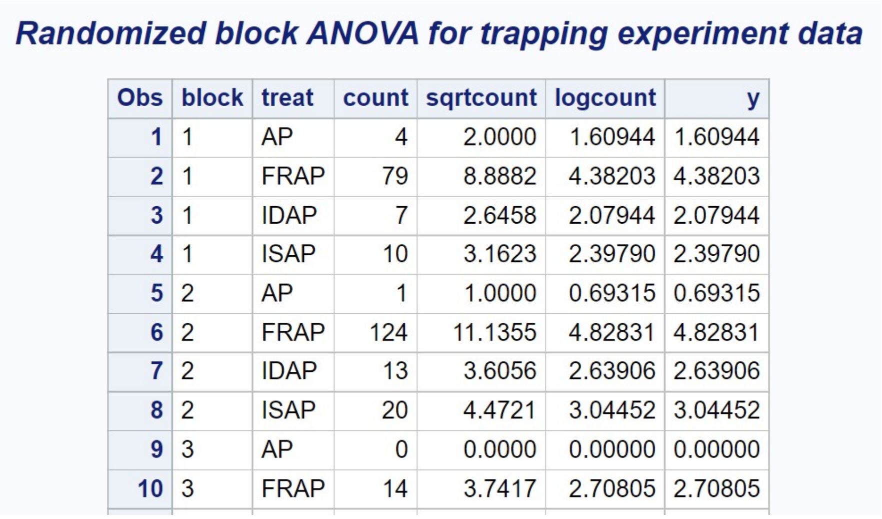
From Fig. 14.27, we see there was a highly significant effect of bait treatment on the number of predators trapped ($F_{3,13.9} = 54.68, P < 0.0001$). Note the non-integer degrees of freedom for this F statistic. This has occurred because the data are unbalanced (one observation is a missing value) and `proc mixed` is adjusting the test. The Tukey output (Fig. 14.28) shows a column of adjusted P values, with the adjustment made according to the Tukey procedure. Adjusted P values less than 0.05 are judged to be significant. We see that every pair of bait treatments was significantly different except for IDAP vs. ISAP. The graph (Fig. 14.25) and least squares means show that `FRAP` caught the most insects, IDAP and ISAP were intermediate, while AP caught the fewest.

The `proc mixed` output also provides estimates of the two variance components in the model, the block variance (σ^2_B) and the variance of ϵ_{ij} (σ^2). They are listed under the `Covariance Parameter Estimates` in the SAS output (Fig. 14.27), labeled as `block` and `Residual`, along with confidence intervals for these estimates. We see that the block variance $\sigma^2_B = 0.3332$ was large relative to $\sigma^2 = 0.1831$. The block variance can be directly observed in Fig. 14.25 as the vertical spread between different blocks. In most cases, we are primarily interested in testing the fixed effects in the model, with the random effects and their associated variance components of less importance. They are included in the model and analysis to account and control for spatial heterogeneity in the observations. We will examine a likelihood ratio test for the block variance in the next section.

SAS Program

```
* TrapRCBD_clerids.sas;
title "Randomized block ANOVA for trapping experiment data";
data trapexp;
    input block $ treat $ count;
    * Apply transformations here;
    sqrtcount = sqrt(count);
    logcount = log(count+1);
    * Choose which variable is used for plots and anova;
    y = logcount;
    * Delete blank traps;
    if treat="BLANK" then delete;
    datalines;
1   AP      4
1   BLANK   0
1   FRAP    79
1   IDAP    7
1   ISAP    10
2   AP      1
2   BLANK   0
2   FRAP    124
2   IDAP    13
2   ISAP    20
3   AP      0
3   BLANK   0
3   FRAP    14
3   IDAP    .
3   ISAP    2
4   AP      0
4   BLANK   0
4   FRAP    15
4   IDAP    11
4   ISAP    7
5   AP      0
5   BLANK   0
5   FRAP    29
5   IDAP    7
5   ISAP    7
6   AP      2
6   BLANK   0
6   FRAP    70
6   IDAP    14
6   ISAP    20
;
```

```
run;
* Print data set;
proc print data=trapexp;
run;
* Plot means, standard errors, and observations;
proc gplot data=trapexp;
    plot y*treat=block / vaxis=axis1 haxis=axis1;
    symbol1 i=j v=star height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
proc mixed cl plots=residualpanel data=trapexp;
    class treat block;
    model y = treat / ddfm=kr;
    random block;
    lsmeans treat / pdiff=all adjust=tukey adjdfe=row;
run;
quit;
```



Randomized block ANOVA for trapping experiment data

Obs	block	treat	count	sqrtcount	logcount	y
1	1	AP	4	2.0000	1.60944	1.60944
2	1	FRAP	79	8.8882	4.38203	4.38203
3	1	IDAP	7	2.6458	2.07944	2.07944
4	1	ISAP	10	3.1623	2.39790	2.39790
5	2	AP	1	1.0000	0.69315	0.69315
6	2	FRAP	124	11.1355	4.82831	4.82831
7	2	IDAP	13	3.6056	2.63906	2.63906
8	2	ISAP	20	4.4721	3.04452	3.04452
9	3	AP	0	0.0000	0.00000	0.00000
10	3	FRAP	14	3.7417	2.70805	2.70805

etc.

Figure 14.24: `TrapRCBD_clerids.sas - proc print`

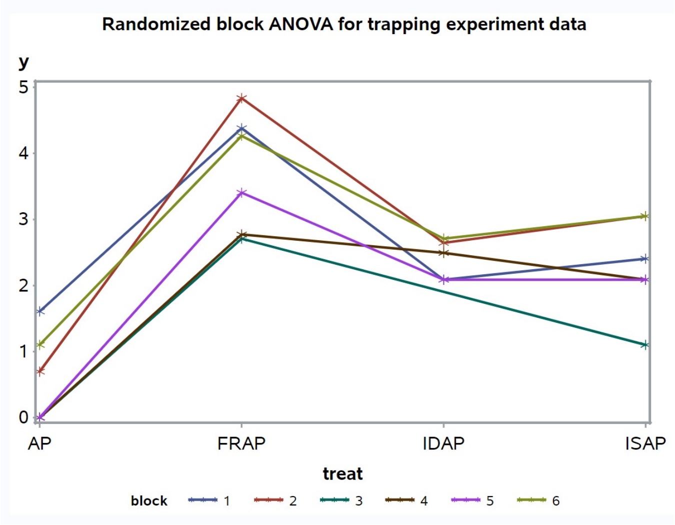
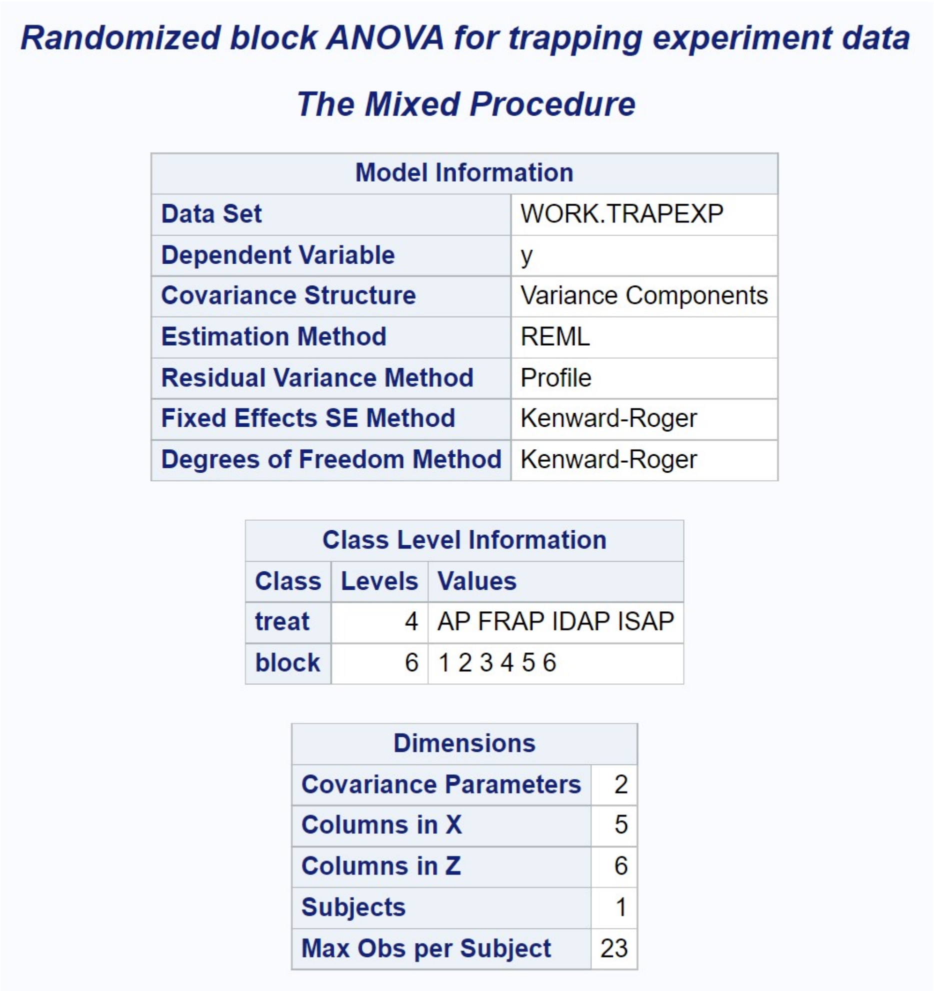


Figure 14.25: `TrapRCBD_clerids.sas - proc gplot`



Randomized block ANOVA for trapping experiment data

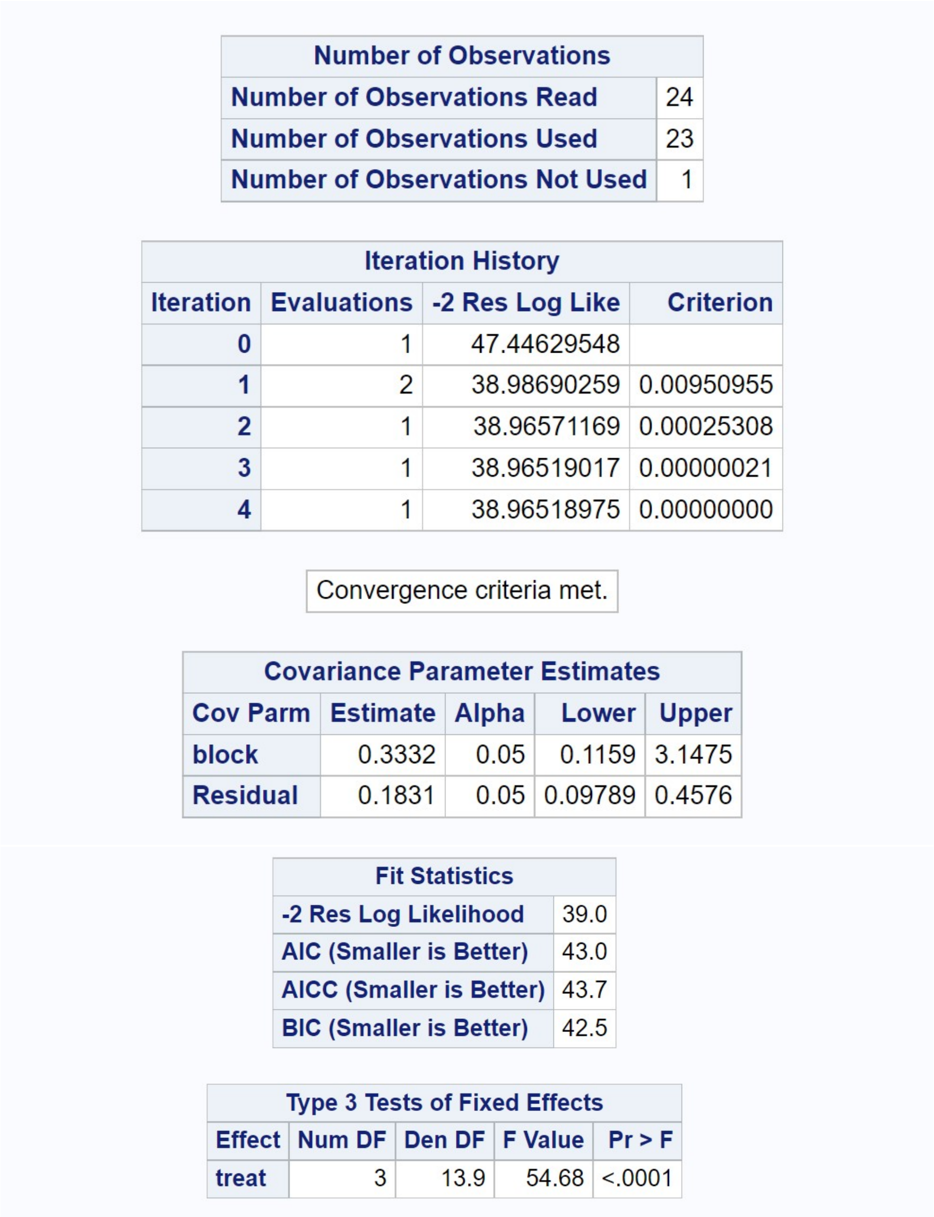
The Mixed Procedure

Model Information	
Data Set	WORK.TRAPEXP
Dependent Variable	y
Covariance Structure	Variance Components
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Kenward-Roger
Degrees of Freedom Method	Kenward-Roger

Class Level Information		
Class	Levels	Values
treat	4	AP FRAP IDAP ISAP
block	6	1 2 3 4 5 6

Dimensions	
Covariance Parameters	2
Columns in X	5
Columns in Z	6
Subjects	1
Max Obs per Subject	23

Figure 14.26: `TrapRCBD_clerids.sas - proc mixed`



Number of Observations	
Number of Observations Read	24
Number of Observations Used	23
Number of Observations Not Used	1

Iteration History

Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	47.44629548	
1	2	38.98690259	0.00950955
2	1	38.96571169	0.00025308
3	1	38.96519017	0.00000021
4	1	38.96518975	0.00000000

Convergence criteria met.

Covariance Parameter Estimates

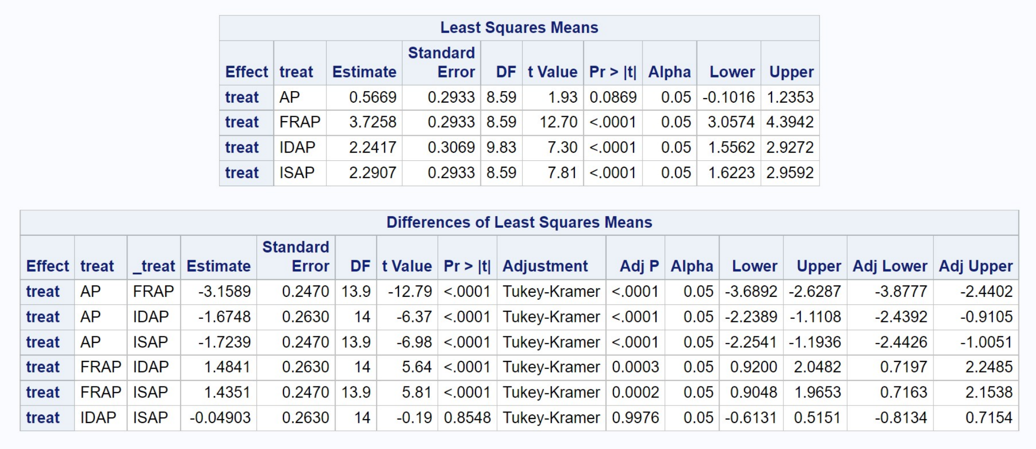
Cov Parm	Estimate	Alpha	Lower	Upper
block	0.3332	0.05	0.1159	3.1475
Residual	0.1831	0.05	0.09789	0.4576

Fit Statistics	
-2 Res Log Likelihood	39.0
AIC (Smaller is Better)	43.0
AICC (Smaller is Better)	43.7
BIC (Smaller is Better)	42.5

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
treat	3	13.9	54.68	<.0001

Figure 14.27: `TrapRCBD_clerids.sas - proc mixed`



Least Squares Means

Effect	treat	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
treat	AP	0.5669	0.2933	8.59	1.93	0.0869	0.05	-0.1016	1.2353
treat	FRAP	3.7258	0.2933	8.59	12.70	<.0001	0.05	3.0574	4.3942
treat	IDAP	2.2417	0.3069	9.83	7.30	<.0001	0.05	1.5562	2.9272
treat	ISAP	2.2907	0.2933	8.59	7.81	<.0001	0.05	1.6223	2.9592

Differences of Least Squares Means

Effect	treat	_treat	Estimate	Standard Error	DF	t Value	Pr > |t|	Adjustment	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
treat	AP	FRAP	-3.1589	0.2470	13.9	-12.79	<.0001	Tukey-Kramer	<.0001	0.05	-3.6892	-2.6287	-3.8777	-2.4402
treat	AP	IDAP	-1.6748	0.2630	14	-6.37	<.0001	Tukey-Kramer	<.0001	0.05	-2.2389	-1.1108	-2.4392	-0.9105
treat	AP	ISAP	-1.7239	0.2470	13.9	-6.98	<.0001	Tukey-Kramer	<.0001	0.05	-2.2541	-1.1936	-2.4426	-1.0051
treat	FRAP	IDAP	1.4841	0.2630	14	5.64	<.0001	Tukey-Kramer	0.0003	0.05	0.9200	2.0482	0.7197	2.2485
treat	FRAP	ISAP	1.4351	0.2470	13.9	5.81	<.0001	Tukey-Kramer	0.0002	0.05	0.9048	1.9653	0.7163	2.1538
treat	IDAP	ISAP	-0.04903	0.2630	14	-0.19	0.8548	Tukey-Kramer	0.9976	0.05	-0.6131	0.5151	-0.8134	0.7154

Figure 14.28: `TrapRCBD_clerids.sas - proc mixed`

14.6.4 Likelihood ratio test for the block effect

In the preceding example, the block variance $\sigma_B^2 = 0.3332$ appeared large relative to $\sigma^2 = 0.1831$, the variance due to ϵ_{ij}. The block effect was also clearly visible in Fig. 14.25. A further step would be a test of $H_0 : \sigma_B^2 = 0$ vs. $H_1 : \sigma_B^2 > 0$. If the test is significant it provides further evidence for variability among blocks in the density of insects. Littell et al. (1996) recommend a likelihood ratio test for this purpose.

We can construct this test by fitting two different models to the data, corresponding to H_0 vs. H_1. Under $H_0 : \sigma_B^2 = 0$ the statistical model for a randomized block design reduces to

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \tag{14.70}$$

because $B_j = 0$ for all j under H_0. The statistical model under $H_1 : \sigma_B^2 > 0$ is just the full model for randomized block designs:

$$Y_{ij} = \mu + \alpha_i + B_j + \epsilon_{ij} \tag{14.71}$$

We now need to find maximum likelihood estimates of the model parameters under both H_1 and H_0, as well as L_{H_0} and L_{H_1}, the maximum height of the likelihood function under H_0 and H_1. We would then use the likelihood ratio test statistic

$$-2\ln(\lambda) = 2\ln(L_{H_1}) - 2\ln(L_{H_0}). \tag{14.72}$$

The SAS program below finds the likelihoods for both models using `proc mixed`. Two separate calls to `proc mixed` are required, one for each model. The likelihoods are labeled `-2 Res Log Likelihood` in the output, which is almost the form required above except for the sign (see Fig. 14.29, 14.30). Examining the output, we see that $-2\ln(L_{H_0}) = 47.4$ and $-2\ln(L_{H_1}) = 39.0$. We then have

$$-2\ln(\lambda) = -39.0 - (-47.4) = -39.0 + 47.4 = 8.4 \tag{14.73}$$

How do we obtain a P value for this test statistic? For any likelihood ratio test, the quantity $-2\ln(\lambda)$ has approximately a χ^2 distribution under H_0. The degrees of freedom for the test are equal to the difference in the number of parameters for the two models (H_1 vs. H_0). There is a difference in one parameter between the two models here, because H_1 has the block variance σ_B^2 while under H_0 this is assumed to be zero. We therefore have $df = 1$, and from Table C find that $P < 0.005$. We are actually conducting a one-tailed test, however, because H_1 is a one-tailed alternative. Thus, the P value is half this quantity, or $P < 0.0025$. It appears the variance due to blocks was highly significant.

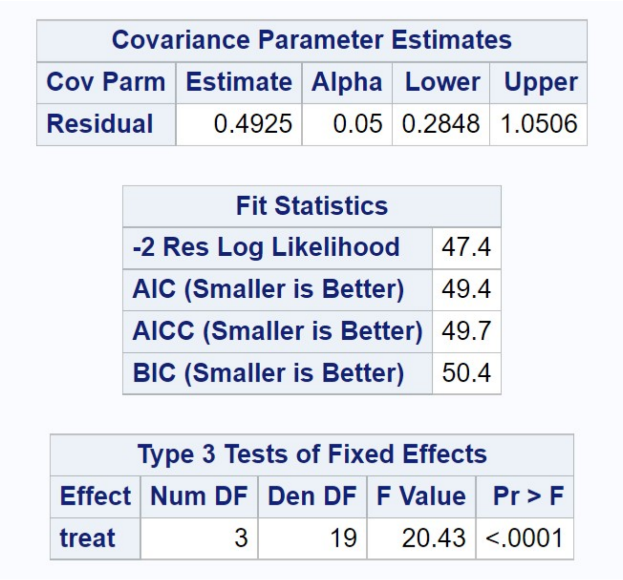
We can calculate the P value more exactly using a simple SAS program (see below). In the `data` step, the program reads in the values of $-2\ln(L_{H_0})$, $-2\ln(L_{H_1})$, and df, then calculates the P value using the SAS function `probchi`. We find that $P = 0.0019$ (Fig. 14.31).

SAS Program

```
* TrapRCBD_clerids_block_test.sas;
title "Randomized block ANOVA for trapping experiment data";
data trapexp;
    input block $ treat $ count;
    * Apply transformations here;
    sqrtcount = sqrt(count);
    logcount = log(count+1);
    * Choose which variable is used for plots and anova;
    y = logcount;
    * Delete blank traps;
    if treat="BLANK" then delete;
    datalines;
1   AP      4
1   BLANK   0
1   FRAP    79
1   IDAP    7
1   ISAP    10

etc.

6   AP      2
6   BLANK   0
6   FRAP    70
6   IDAP    14
6   ISAP    20
;
run;
title2 "H0 true - no block effect";
proc mixed cl data=trapexp;
    class treat;
    model y = treat / ddfm=kr;
run;
title2 "H1 true - there is a block effect";
proc mixed cl data=trapexp;
    class treat block;
    model y = treat / ddfm=kr;
    random block;
run;
quit;
```

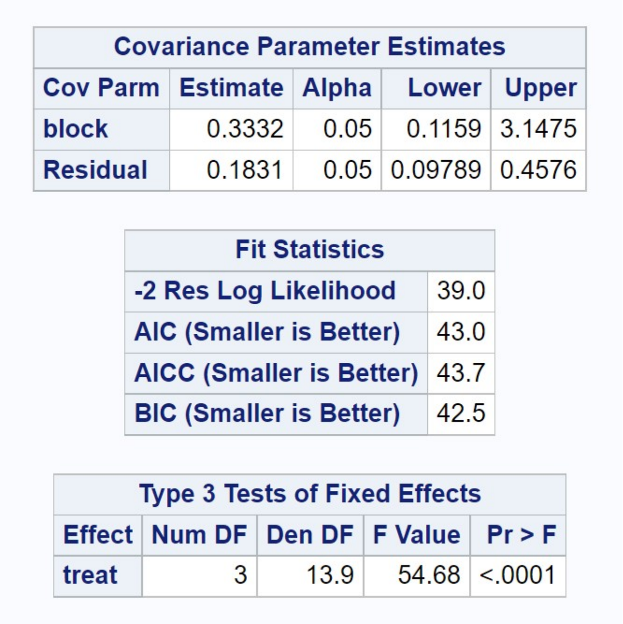


Covariance Parameter Estimates				
Cov Parm	Estimate	Alpha	Lower	Upper
Residual	0.4925	0.05	0.2848	1.0506

Fit Statistics	
-2 Res Log Likelihood	47.4
AIC (Smaller is Better)	49.4
AICC (Smaller is Better)	49.7
BIC (Smaller is Better)	50.4

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
treat	3	19	20.43	<.0001

Figure 14.29: `TrapRCBD_clerids_block_test.sas - proc mixed` (1)



Covariance Parameter Estimates

Cov Parm	Estimate	Alpha	Lower	Upper
block	0.3332	0.05	0.1159	3.1475
Residual	0.1831	0.05	0.09789	0.4576

Fit Statistics

-2 Res Log Likelihood	39.0
AIC (Smaller is Better)	43.0
AICC (Smaller is Better)	43.7
BIC (Smaller is Better)	42.5

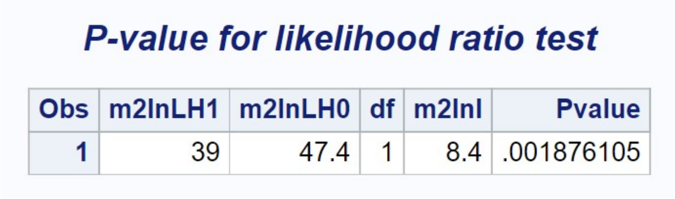
Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
treat	3	13.9	54.68	<.0001

Figure 14.30: `TrapRCBD_clerids_block_test.sas - proc mixed` (2)

SAS Program

```
* lrtpvalue.sas;
title "P-value for likelihood ratio test";
data values;
    *Data are -2lnL values under H0 and H1, plus degrees of freedom;
    input m2lnLH1 m2lnLH0 df;
    m2lnl = -m2lnLH1 - (-m2lnLH0);
    * Find P-value;
    Pvalue = (1 - probchi(m2lnl,df))/2;
    datalines;
39.0 47.4 1
;
run;
proc print data=values;
run;
```



P-value for likelihood ratio test

Obs	m2lnLH1	m2lnLH0	df	m2lnl	Pvalue
1	39	47.4	1	8.4	.001876105

Figure 14.31: `lrtpvalue.sas - proc print`

14.7 References

Cox, D. R. (1984) Interaction. *International Statistical Review* 52: 1-24.

Hurlbert, S. H. (1984) Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* 54: 187-211.

Littell, R. C., Milliken, G. A., Stroup, W. W. & Wolfinger, R. D. (1996) *The SAS System for Mixed Models.* SAS Institute Inc., Cary, NC.

Maestre, F. T. & Reynolds, J. F. (2007) Amount or pattern? Grassland responses to the heterogeneity and availability of two key resources. *Ecology* 88: 501-511.

McCulloch, C. E. & Searle, S. R. (2001) *Generalized, Linear, and Mixed Models.* John Wiley & Sons, Inc., New York, NY.

Potvin, C. (1993) ANOVA: experiments in controlled environments. Pages 46-68 in *Design and Analysis of Ecological Experiments*, S. M. Scheiner and J. Gurevitch eds. Chapman & Hall, New York, NY.

Reeve, J. D., Rojas, M. G., & Morales-Ramos, J. A. (2003) Artificial diet and rearing methods for *Thanasimus dubius* (Coleoptera: Cleridae), a predator of bark beetles (Coleoptera: Scolytidae). *Biological Control* 27: 315-322.

Reeve, J. D., Strom, B. L., Rieske-Kinney, L. K., Ayres, B. D. & Costa, A. (2009) Geographic variation in prey preference in bark beetle predators. *Ecological Entomology* 34: 183-192.

SAS Institute Inc. (2016) *SAS/GRAPH 9.4: Reference, Fifth Edition.* SAS Institute Inc., Cary, NC.

SAS Institute Inc. (2018) *SAS/STAT 15.1 Users Guide* SAS Institute Inc., Cary, NC

Searle, S. R. (1971) *Linear Models.* John Wiley & Sons, Inc., New York, NY.

Shaw, R. G. & Mitchell-Olds, T. (1993) ANOVA for unbalanced data: an overview. *Ecology* 74: 1638-1645.

Speed, F. M., Hocking, R. R. & Hackney, O. P. (1978) Methods of analysis of linear models with unbalanced data. *Journal of the American Statistical Association* 73: 105-112.

Stewart-Oaten, A. (1995) Rules and judgments in statistics: three examples. *Ecology* 76: 2001-2009.

Winer, B. J., Brown, D. R. & Michels, K. M. (1991) *Statistical Principles in Experimental Design*, 3rd edition. McGraw-Hill, Inc., Boston, MA.

14.8 Problems

1. An entomologist is interested in how bark beetles respond to traps baited with two treatments, their own pheromone (P) vs. the pheromone plus a repellent chemical (PR). They also want to see if trap color (black vs. white) affects the response of the beetles. They conduct an experiment in which these two factors are randomly assigned to traps in one section of the forest, with five replicate traps for each treatment. The counts of bark beetles responding to each trap are listed below.

Bait	Trap color	Counts for five replicate traps
P	Black	138, 569, 196, 139, 726
PR	Black	96, 168, 25, 36, 152
P	White	174, 99, 293, 67, 122
PR	White	52, 27, 11, 57, 93

   (a) Write an appropriate ANOVA model for this design, and state which effects are fixed or random. Is it possible to include an interaction term in the model?

   (b) Use SAS to analyze these data using your ANOVA model, log transforming the observations. Interpret the results of all the tests. Attach your SAS program and output.

2. A research group is interested in the effects of diet and temperature on the growth rate of fish in aquaculture. They conduct an experiment with three different diet treatments (A, B and C) crossed with three rearing temperatures (15, 20 and 25°C). Two fish tanks are assigned to each treatment combination and the growth rate (g/week) determined for each tank. The following data were obtained:

Diet	Temp	Growth rate (two tanks)
A	15	24.7, 22.3
A	20	31.9, 28.9
A	25	32.6, 31.3
B	15	19.6, 14.2
B	20	30.5, 26.5
B	25	25.5, 32.8
C	15	21.1, 21.3
C	20	23.4, 23.4
C	25	28.2, 25.8

(a) Write an appropriate ANOVA model for this design, and state which effects are fixed or random. Is it possible to include an interaction term in the model?

(b) Use SAS to analyze these data using your ANOVA model. You may use any method for dealing with interactions. Interpret the results of all the tests.

(c) Use the Tukey method to compare the different diet treatments, and then the temperature treatments. Interpret the results.

Chapter 15

Assumptions and Transformations

Analysis of variance as well as regression analysis (see Chapter 17) make a number of assumptions about the nature of the observations. These assumptions are embodied in the statistical model used in the analysis. For example, recall the model for fixed effects one-way ANOVA:

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}. \tag{15.1}$$

Here μ is the grand mean while α_i is the deviation from μ caused by the *ith* level of Factor A. The ϵ_{ij} term represents random departures from the mean value predicted by Factor A due to natural variability. It is assumed that $\epsilon_{ij} \sim N(0, \sigma^2)$ and that these random variables are also independent of one another. We examine these assumptions in more detail below and discuss how their violation can affect the validity of the statistical analyses. We then describe how **variance-stabilizing transformations** are used to fix certain violations of these assumptions. We also present a common method for identifying these violations known as **residual analysis**.

15.1 ANOVA assumptions

15.1.1 Independence of observations

One key assumption embodied in the above model is that the error terms ϵ_{ij} are independent, implying that the observations Y_{ij} are also independent.

How would a lack of independence influence the results of ANOVA? The consensus is that a lack of independence can greatly influence the validity of ANOVA, including the Type I error rate and power of the F test, as well as the estimation of group effects (Glass et al. 1972).

As an example of an experimental design where the observations are not independent, suppose that we conduct an insect trapping experiment with two bait treatments, A and B. We place all of the bait A traps in one location and bait B ones in a second location. If location influences the abundance of insects, we would expect the trap catches at a particular site to be high or low for this reason, separate of any treatment effect. As a consequence, the observations at a particular location are related to one another and so not independent. We would also be more likely to find a bait effect if these data were analyzed using one-way ANOVA, simply because of the location effect. Thus, the Type I error rate of the F test would be higher. This combination of poor experimental design and an inappropriate statistical analysis has been called **pseudoreplication** (Hurlbert 1984). While there are multiple traps within each location, they are not true replicates because the observations are not independent, and treatment and location effects cannot be separated. This design basically has only one replicate per treatment, one for each location.

Fortunately, the assumption of independence will usually be satisfied by good experimental design and execution (Hurlbert 1984). In the insect bait experiment, a better experimental design would randomly allocate bait types to traps at both locations, and the analysis would also include a location (block) effect in the statistical model. Randomization also helps ensure that estimates of the treatment effects are unbiased. For example, bait type A might be messier to use than B, and the experimenter might be tempted to do those replicates last or place them in a different location. This potential source of bias by the experimenter is avoided by randomization of the treatments.

15.1.2 Homogeneity of variances

Another key assumption of ANOVA is that the variance is similar among treatment groups, also known as the **homogeneity of variances** assumption or **homoscedasticity**. This follows from the assumption that ϵ_{ij} has a variance of σ^2 regardless of the treatment group. We can also see this from a graphical presentation of the one-way ANOVA model, where each treatment group has the same distribution with the same variance except for shifts due to Factor A (see Fig. 11.1 in Chapter 11). The condition of unequal variances is also called **heteroscedasticity**.

If the homogeneity of variances assumption is not satisfied this can strongly affect the validity of the F test in ANOVA, especially when the design is unbalanced (Glass et al. 1972). If the treatments with higher variances have smaller sample sizes, then the actual Type I error rate will be higher than its nominal value (say $\alpha = 0.05$). Conversely, if the treatments with higher variances have larger sample sizes, the actual Type I error rate will be smaller than its nominal value. We will see later in this chapter how **variance-stabilizing transformations** can be used to equalize the variance among groups, making the observations better conform to this assumption.

15.1.3 Normality

A further assumption of ANOVA is that the error term ϵ_{ij} is normally distributed, and as a consequence so are the observations (Y_{ij} values). The assumption of normality appears to be less important for the validity of ANOVA than homogeneity of variances. Many studies indicate that the ANOVA F test has the nominal Type I error rate ($\alpha = 0.05$) even when the observations have distributions quite different from the normal, although power may be increased or decreased relative to the normal (see Table 16, Glass et al. 1972). For large values of n per group, ANOVA is likely to be a valid procedure regardless of the distribution of the observations due to the central limit theorem (Chapter 7). In practice, a transformation that equalizes the variance among groups also seems to normalize the observations, solving both problems.

15.1.4 Absence of outliers

An assumption of ANOVA related to normality is the absence of outliers. **Outliers are observations that lie far from the other observations in a particular study.** The source of the outlier could be a rare biological event, or simply a data entry error or bad measurement with an instrument. Because it lies far from the other observations, an outlier will increase the size of MS_{within} and alter the estimated effect of its treatment group. If the outlier is a data error then there is justification for deleting it from the observations. If the source is unclear or the outlier is a valid observation, then one common approach is to conduct the statistical analysis with and without the outlier and present both results. Outliers can be often be identified using residual analysis (see below).

15.1.5 Additivity

ANOVA models are known as additive models because the observations are modeled as the sum of several factors. For example, the model for two-way fixed effects ANOVA without replication is

$$Y_{ij} = \mu + \alpha_i + \beta_j + \epsilon_{ij}. \tag{15.2}$$

Thus, the Y_{ij} values are modeled as the sum of the grand mean, the effects of Factor A and B, and a random term representing variability among the observations. Additivity of effects is a basic assumption of ANOVA.

However, some biological processes like survival and reproduction are inherently multiplicative processes. For example, suppose our observations are the number of offspring surviving to maturity from a single female. This number will be the product of the fecundity of the female and the survival rate of the offspring. We now apply a number of treatments that could potentially influence both these factors. The resulting observations could be described using the model

$$Y_{ij} = \lambda s_i f_j \gamma_{ij}, \tag{15.3}$$

where λ is the average number of offspring surviving to maturity, while s_i and f_j are the differential effects of the survival and fecundity treatments. The term γ_{ij} is a multiplicative error term with a distribution that takes only positive values, and it is typically required that $E[\gamma_{ij}] = 1$. Note that these must all be positive quantities in order for the number of offspring (Y_{ij}) to be positive.

Can data of this type be analyzed using ANOVA? The answer is yes, because we can use a log transformation to make the data additive. Taking the log of both sides of this model, we obtain

$$\log Y_{ij} = \log \lambda + \log s_i + \log f_j + \log \gamma_{ij}. \tag{15.4}$$

The result is an additive model the same as for unreplicated two-way ANOVA, and the data can be analyzed using standard ANOVA methods. This is one reason why studies of reproduction and survival as well as population dynamics routinely use the log transformation.

15.2 Variance-stabilizing transformations

Variance-stabilizing transformations are often used by statisticians to equalize the variance of observations across different treatment groups, so that the homogeneity of variances assumption is better satisfied. We have already employed these transformations in some of our analyses, including the log and arcsine-square root transformations.

The different transformations are derived as follows. Suppose we have a random variable Y that describes the data, and there is a functional relationship between its variance $Var[Y] = v$ and its mean $E[Y] = m$. More specifically, suppose that we have

$$v = f(m) \tag{15.5}$$

where f is some function. For example, with the Poisson distribution for parameter λ we have $Var[Y] = E[Y] = \lambda$ (Chapter 7), and so $v = m$ is the functional relationship. It can then be shown that a function g that satisfies the equation

$$g(m) = \int \frac{\theta dm}{\sqrt{f(m)}}, \tag{15.6}$$

where θ is a constant, will be a variance-stabilizing transformation (Bartlett 1947). To see how this process works, suppose that a random variable Y has a Poisson distribution. We find that

$$g(m) = \int \frac{\theta dm}{\sqrt{m}} = \theta \frac{m^{1/2}}{1/2} + C = 2\theta\sqrt{m} + C \propto \sqrt{m}. \tag{15.7}$$

Thus, the variance-stabilizing transformation for Poisson data is $\sqrt{Y}$.

As another example, suppose that $v = m^2$ so that the variance increases with the square of the mean. Negative binomial data will have this form for large m, because $v = m + m^2/k$ for this distribution (Chapter 7). For this relationship between v and m, we have

$$g(m) = \int \frac{\theta dm}{\sqrt{m^2}} = \int \frac{\theta dm}{m} = \theta \log m + C \propto \log m, \tag{15.8}$$

implying that $\log Y$ is the variance-stabilizing transformation. Either natural or base 10 log transformations can be used and will yield identical results for the statistical tests in ANOVA. The $\log Y$ transformation is a ‘stronger’ transformation than the $\sqrt{Y}$ because it corrects for a stronger relationship between v and m.

A variance-stabilizing transformation is also needed for proportions, because the variance of a proportion depends on its mean. To see this, suppose that we observe l different individuals from some population and record their sex. Let Y be the number of individuals in the sample that are female. The variable Y would be a binomial random variable with parameters l and p, where p is the proportion of females in the population, and so $E[Y] = lp$ and $Var[Y] = lp(1-p)$ (see Chapter 5). Then, a **binomial proportion** would be Y/l, the proportion of females in the sample. For this proportion, we have $E[Y/l] = lp/l = p$ while $Var[Y/l] = lp(1-p)/l^2 = p(1-p)/l$. If we set $m = p$, then $v = Var[Y/l] = m(1-m)/l$ and so v is a function of m. Using the same method as above, we find that the variance-stabilizing transformation for binomial proportions is $\sin^{-1}(\sqrt{Y})$ or $\arcsin(\sqrt{Y})$. This transformations maps proportions from 0 to 1 to the interval 0 to $\pi/2$. The largest effect of the transformation is on proportions close to 0 or 1.

Table 15.1 lists the commonly used variance-stabilizing transformations. Also listed are variants of the transformations that are useful when the data include zeroes, as often occurs in count data. In the next section, we will illustrate the use of these transformations, and how the appropriate transformation can be determined through residual analysis.

Table 15.1: Variance-stabilizing transformations for various $v = f(m)$ and the data for which they are useful.

$v = f(m)$	Transformation	Comments
$v = m$	$\sqrt{Y}, \sqrt{Y+1/2}$ (zeroes)	Poisson data
$v = m^2$	$\log Y, \log(Y+1)$ (zeroes)	Overdispersed count data, many other types
$v = m(1-m)/l$	$\arcsin(\sqrt{Y})$	Proportions

15.3 Residual analysis

The details of residual analysis are presented in this section. We begin by defining predicted and residual values using one-way ANOVA as an example, for both fixed and random effects (similar results hold for more complex designs). We then illustrate residual analysis and the use of variance-stabilizing transformations with some examples.

15.3.1 Models, estimates, and predictors

ANOVA is based on statistical models that contain a number of parameters. For example, the statistical model for fixed effects one-way ANOVA has the form

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}, \tag{15.9}$$

where μ is the grand mean, α_i is the deviation from the μ caused by the *ith* treatment, and $\epsilon_{ij} \sim N(0, \sigma^2)$. We saw earlier how likelihood methods could be used to estimate the parameters μ, α_i, and σ^2 for this model. For the random effects version, the model contained a random variable $A_i \sim N(0, \sigma^2_A)$, and is written as

$$Y_{ij} = \mu + A_i + \epsilon_{ij}. \tag{15.10}$$

The parameters in this model are μ, σ^2_A, and σ^2, and these quantities can also be estimated using likelihood methods. It is also possible to estimate the random variable A_i itself, more specifically the value realized in a particular group and study. Estimators of A_i are often called **predictors** in this context, because they concern random variables rather than model parameters (Searle et al. 1992).

15.3.2 Predicted and residual values

We can use these estimates to generate a **predicted value** for each observation Y_{ij} in the data set. For the fixed effects model listed above, the predicted value of Y_{ij} is $\hat{Y}_{ij} = \hat{\mu} + \hat{\alpha}_i$, where $\hat{\mu}$ and $\hat{\alpha}_i$ are the estimated values of μ and α_i. Note that all observations in the *ith* group would have the same predicted value.

What actually are the predicted values here? Recall that for the fixed effects model, the maximum likelihood estimates of these parameters are

$$\hat{\mu} = \bar{\bar{Y}} \tag{15.11}$$

and

$$\hat{\alpha}_i = \bar{Y}_{i\cdot} - \bar{\bar{Y}}. \tag{15.12}$$

Thus,

$$\hat{Y}_{ij} = \hat{\mu} + \hat{\alpha}_i = \bar{\bar{Y}} + \bar{Y}_{i\cdot} - \bar{\bar{Y}} = \bar{Y}_{i\cdot}. \tag{15.13}$$

So, the predicted value for the *ith* group is just the mean of that group.

Similarly, for the random effects model the predicted value of Y_{ij} is $\hat{Y}_{ij} = \hat{\mu} + \hat{A}_i$, where $\hat{\mu} = \bar{\bar{Y}}$ and $\hat{A}_i$ is the predictor of A_i. It turns out that the best predictor for the realized value of A_i is 'shrunk' relative to α_i and has the form

$$\hat{A}_i = \frac{\sigma_A^2}{\sigma_A^2 + \sigma^2/n}(\bar{Y}_{i\cdot} - \bar{\bar{Y}}) \tag{15.14}$$

(Searle et al. 1992). It depends on σ_A^2 and σ^2 as well as $\bar{Y}_{i\cdot}$ and $\bar{\bar{Y}}$. It follows that

$$\hat{Y}_{ij} = \hat{\mu} + \hat{A}_i = \bar{\bar{Y}} + \frac{\sigma_A^2}{\sigma_A^2 + \sigma^2/n}(\bar{Y}_{i\cdot} - \bar{\bar{Y}}) \tag{15.15}$$

for the random effects model. Thus, $\hat{Y}_{ij}$ is not equal to $\bar{Y}_{i\cdot}$ in this situation but lies closer to the grand mean $\bar{\bar{Y}}$, unless n is large. In practice, estimates of the two variance components are used to generate the predicted value.

In assessing the validity of our statistical models, we will also be interested in the **residuals** of the observations, which are defined as the difference $Y_{ij} - \hat{Y}_{ij}$. The residuals essentially provide an estimate of the error term ϵ_{ij} for each observation, which we can call $\hat{\epsilon}_{ij}$. Why is this so? The model for one-way ANOVA can be expressed as

$$Y_{ij} - (\mu + \alpha_i) = \epsilon_{ij}. \tag{15.16}$$

If we insert estimates for μ and α_i in this equation, we obtain an estimate of ϵ_{ij}:

$$Y_{ij} - (\hat{\mu} + \hat{\alpha}_i) = Y_{ij} - \hat{Y}_i = \hat{\epsilon}_{ij}. \tag{15.17}$$

There is an interesting relationship between these residual values and MS_{within}. Suppose that we use the sample variance of the $\hat{\epsilon}_{ij}$ values to estimate the variance of ϵ_{ij}, namely σ^2. The sum of squares associated with this sample variance is

$$SS = \sum_{i=1}^{a}\sum_{j=1}^{n}(\hat{\epsilon}_{ij})^2 = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(Y_{ij} - (\hat{\mu} + \hat{\alpha}_i)\right)^2, \tag{15.18}$$

and the degrees of freedom are $a(n-1)$. Dividing SS by its degrees of freedom, we obtain an estimator of σ^2 based on the residuals:

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n}\left(Y_{ij} - (\hat{\mu} + \hat{\alpha}_i)\right)^2}{a(n-1)}. \tag{15.19}$$

How is this quantity related to MS_{within}, our other estimate of σ^2? If we plug $\hat{\mu} = \bar{\bar{Y}}$ and $\hat{\alpha}_i = \bar{Y}_{i\cdot} - \bar{\bar{Y}}$ into this equation, we obtain

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{a}\sum_{j=1}^{n}\left(Y_{ij} - (\bar{\bar{Y}} + \bar{Y}_{i\cdot} - \bar{\bar{Y}})\right)^2}{a(n-1)} \tag{15.20}$$

$$= \frac{\sum_{i=1}^{a}\sum_{j=1}^{n}\left(Y_{ij} - \bar{Y}_{i\cdot})\right)^2}{a(n-1)} \tag{15.21}$$

$$= MS_{within}. \tag{15.22}$$

Thus, MS_{within} can be expressed in terms of the residuals from the ANOVA estimation process. This relationship is true for all ANOVA models (and regression as well). Because MS_{within} can be expressed using the residual or error terms, MS_{within} is also called $MS_{residual}$ or MS_{error}, and SS_{within} similarly named $SS_{residual}$ or SS_{error}. This terminology is used in SAS output as well.

It is also possible to express MS_{among} in terms of the maximum likelihood estimates of the parameters. Because $\hat{\alpha}_i = \bar{Y}_{i\cdot} - \bar{\bar{Y}}$, we have

$$MS_{among} = \frac{n\sum_{i=1}^{a}(\bar{Y}_{i\cdot} - \bar{\bar{Y}})^2}{a-1} = \frac{n\sum_{i=1}^{a}\hat{\alpha}_i^2}{a-1}. \tag{15.23}$$

From this result, it is clear that MS_{among} is an increasing function of the values of $\hat{\alpha}_i$, the estimated treatment effects (Winer et al. 1991).

15.3.3 Evaluating ANOVA assumptions

Residuals play a key role in determining if a particular data set satisfies the assumptions of ANOVA. They can be used to evaluate three of the assumptions: (1) homogeneity of variances among groups, (2) absence of outliers, and (3) normality of the error terms.

We can evaluate the homogeneity of variances assumption through a plot of the residuals vs. predicted values. **If the variances are homogeneous among groups, the points should be equally scattered for**

each group. This is because the residuals are estimates of the ϵ_{ij} values and are supposed to have the same variance across groups. If the residual vs. predicted plot shows a definite pattern, such as a increase or decrease in the scatter as the predicted values increase, this suggests a variance-stabilizing transformation may be needed. This type of plot is also useful for detecting any outliers in the data. **If an outlier is present it will have a very large residual value.** The normality assumption can be evaluated using a normal quantile plot of the residuals. **If the residuals are normal, then this plot will be a straight diagonal line.**

15.3.4 Residual analysis and transformations - SAS demo

We will illustrate residual analysis and the use of transformations with data from a trapping study of the predatory insect *Thanasiumus dubius* (Reeve et al. 2009). This study used a randomized block design with five bait treatments and six blocks, previously analyzed in Chapter 14. Note that the model for this design contains both fixed and random effects, but predicted values and residuals can still be generated through a more complex process (Searle et al. 1992)

The complete program for this example is listed below for reference. We can generate a residual vs. predicted plot, and a normal quantile plot, by adding the option `plots=residualpanel` to the `proc mixed` statement. We first analyze the data using no transformation by setting `y = count` in the `data` step. Examining the residual vs. predicted plot, we see an increase in the scatter of the residuals as the predicted values increase (Fig. 15.1, top left), especially for the largest predicted values. This implies that the variance of the observations increases with their mean (v is some function of m). In addition, the normal quantile plot (bottom left) does not appear to be a straight diagonal line. Neither assumption appears to be satisfied in this analysis.

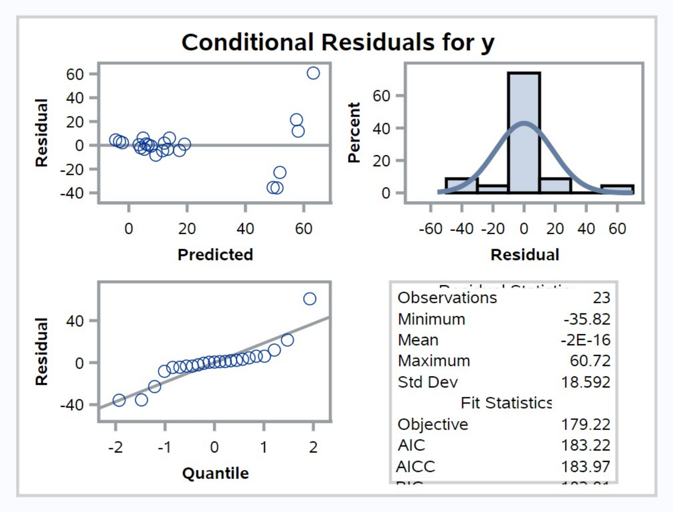
We next analyze the data using a square root transformation by setting `y = sqrtcount` in the `data` step. The residual vs. predicted plot shows less scatter of the residuals for larger predicted values, although there is still some spread (Fig. 15.2). The normal quantile plot is now a straight diagonal line.

We then try a log transformation of the data, setting `y = logcount` in the `data` step. The residual vs. predicted plot shows the same scatter across the range of predicted values (Fig. 15.3), and the normal quantile plot is a straight diagonal line. This is the desired outcome with the data now satisfying the homogeneity of variances and normality assumptions. There also appear to be no outliers (extreme residual values) in these observations. **We can then proceed to interpret the rest of the analysis, such as the F test and multiple comparisons. They should be valid at this point because the ANOVA assumptions are satisfied.** See Chapter 14 for the interpretation of this analysis.

SAS Program

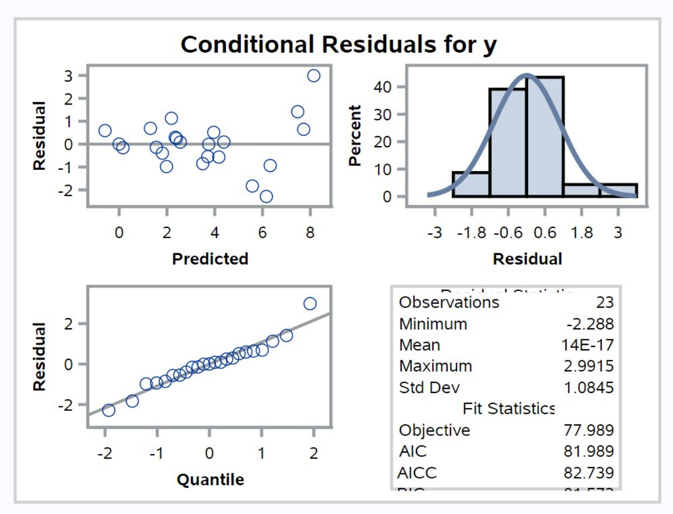
```
* TrapRCBD_clerids.sas;
title "Randomized block anova for trapping experiment data";
data trapexp;
    input block $ treat $ count;
    * Apply transformations here;
    sqrtcount = sqrt(count);
    logcount = log(count+1);
    * Choose which variable is used for plots and anova;
    y = logcount;
    * Delete blank traps;
    if treat="BLANK" then delete;
    datalines;
1   AP      4
1   BLANK   0
1   FRAP    79
1   IDAP    7
1   ISAP    10
2   AP      1
2   BLANK   0
2   FRAP    124
2   IDAP    13
2   ISAP    20
3   AP      0
3   BLANK   0
3   FRAP    14
3   IDAP    .
3   ISAP    2
4   AP      0
4   BLANK   0
4   FRAP    15
4   IDAP    11
4   ISAP    7
5   AP      0
5   BLANK   0
5   FRAP    29
5   IDAP    7
5   ISAP    7
6   AP      2
6   BLANK   0
6   FRAP    70
6   IDAP    14
6   ISAP    20
;
```

```
run;
* Print data set;
proc print data=trapexp;
run;
* Plot means, standard errors, and observations;
proc gplot data=trapexp;
    plot y*treat=block / vaxis=axis1 haxis=axis1;
    symbol1 i=j v=star height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* Mixed model analysis;
proc mixed cl plots=residualpanel data=trapexp;
    class treat block;
    model y = treat / ddfm=kr;
    random block;
    lsmeans treat / pdiff=all adjust=tukey;
run;
quit;
```



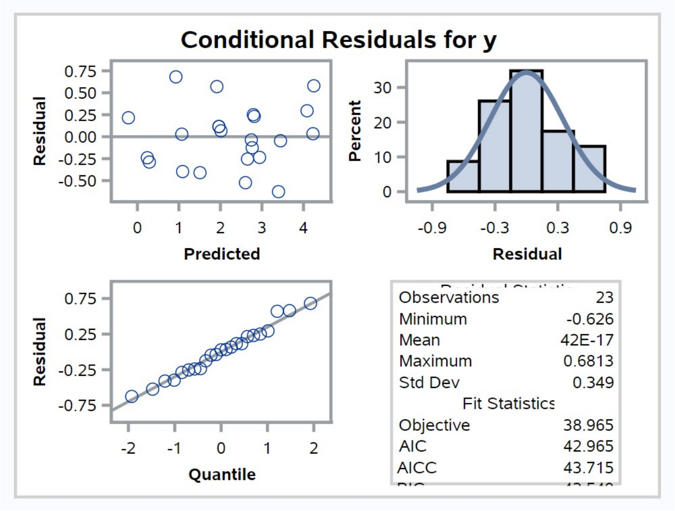
```
Observations        23
Minimum         -35.82
Mean            -2E-16
Maximum          60.72
Std Dev         18.592
      Fit Statistics
Objective       179.22
AIC             183.22
AICC            183.97
```

Figure 15.1: `TrapRCBD.sas - proc mixed` (no transform)



```
Observations        23
Minimum         -2.288
Mean            14E-17
Maximum         2.9915
Std Dev         1.0845
      Fit Statistics
Objective       77.989
AIC             81.989
AICC            82.739
```

Figure 15.2: `TrapRCBD.sas - proc mixed` ($\sqrt{Y}$ transform)



```
Observations        23
Minimum         -0.626
Mean            42E-17
Maximum         0.6813
Std Dev          0.349
      Fit Statistics
Objective       38.965
AIC             42.965
AICC            43.715
```

Figure 15.3: `TrapRCBD.sas - proc mixed` ($\ln(Y+1)$ transform)

15.3.5 $\arcsin(\sqrt{Y})$ transformation - SAS demo

As another example of residual analysis and transformation, we will analyze the observations from an experiment involving an insect predator and the survival of a pest insect on which it feeds. Plots are established each containing 20 pest insects, and a predator treatment (0, 1, or 2 predators) randomly assigned to each plot. There were $n = 10$ plots per predator treatment. The proportion of pest insects surviving was determined for each plot. We will analyze this experiment using one-way ANOVA and `proc glm`, with the `predator` treatment a fixed effect. Residual plots can be requested using the option `plots=diagnostics`. See complete program below.

We first analyze these data using untransformed proportions, using `y = prop` in the `data` step, where `prop` is the proportion of surviving pest insects. Examining the residual vs. predicted plot (Fig. 15.4, top left), we see that the variability of the observations for one treatment is smaller. This is the 0 predator treatment and has a very high survival rate. The normal quantile plot (second row, left) is a straight diagonal line, so this assumption is apparently satisfied.

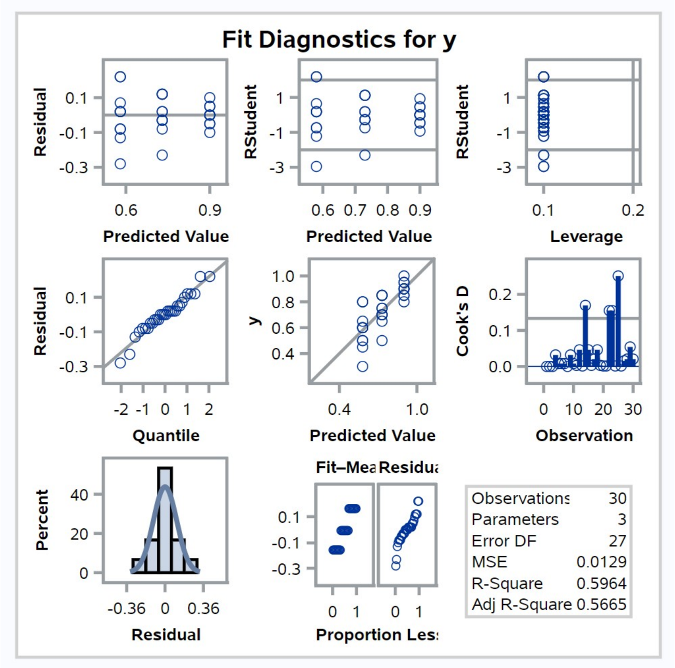
We then analyze the experiment using the transformation $\arcsin(\sqrt{Y})$ where Y is the proportion, using `y = arsin(sqrt(prop))` in the `data` step. The residual vs. predicted plot shows an equal scatter of the residuals across the predicted values, suggesting the homogeneity of variances assumption is satisfied (Fig. 15.5). The normal quantile plot is a straight diagonal line once more. What has happened here? The transformation has spread out the survival rates for the 0 predator treatment, thus equalizing the variances among the treatment groups.

Examining the ANOVA output (Fig. 15.8), we see there was a highly significant effect of the predator treatment on the survival rate of the pest insect ($F_{2,27} = 21.26, P < 0.0001$). Pest survival decreased as the number of predators increased (Fig. 15.7).

SAS Program

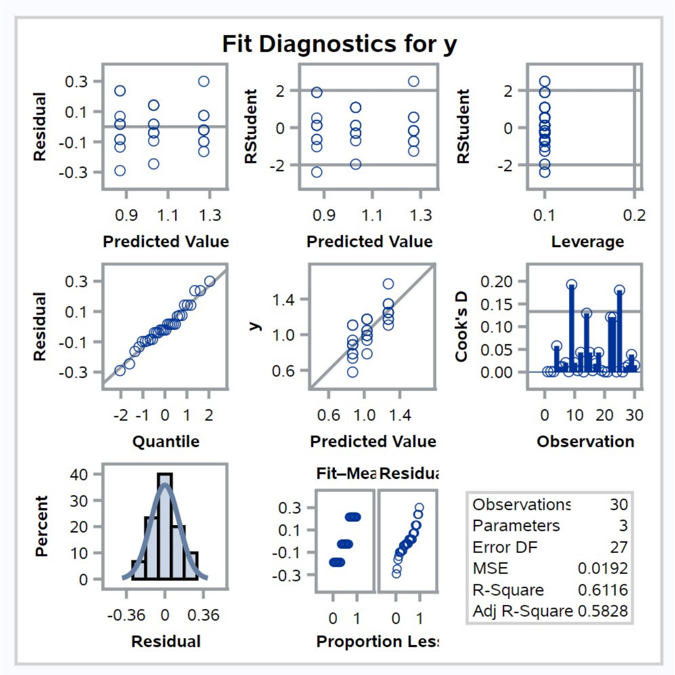
```
* arcsine.sas;
title 'One-way ANOVA for proportions';
data arcsine;
    input predators survivors;
    prop = survivors/20;
    * Apply transformations here;
    y = arsin(sqrt(prop));
    datalines;
0 18
0 18
0 18
0 16
0 19
0 19
0 17
0 18
0 20
0 17
1 14
1 17
1 15
1 10
1 17
1 14
1 13
1 17
1 14
1 15
2 12
2 16
2 16
2 12
2 6
2 12
2 13
2 10
2 9
2 10
;
run;
* Print data set;
proc print data=arcsine;
run;
```

```
* Plot means, standard errors, and observations;
proc gplot data=arcsine;
    plot y*predators=1 / vaxis=axis1 haxis=axis1;
    symbol1 i=std1mjt v=star height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* One-way anova with all fixed effects;
proc glm plots=diagnostics data=arcsine;
    class predators;
    model y = predators;
run;
quit;
```



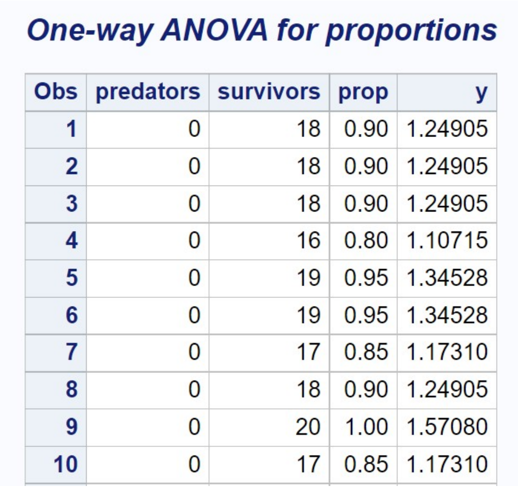
Observations	30
Parameters	3
Error DF	27
MSE	0.0129
R-Square	0.5964
Adj R-Square	0.5665

Figure 15.4: `arcsine.sas - proc glm` (no transform)



Observations	30
Parameters	3
Error DF	27
MSE	0.0192
R-Square	0.6116
Adj R-Square	0.5828

Figure 15.5: `arcsine.sas - proc glm` (arcsin($\sqrt{Y}$) transform)



One-way ANOVA for proportions

Obs	predators	survivors	prop	y
1	0	18	0.90	1.24905
2	0	18	0.90	1.24905
3	0	18	0.90	1.24905
4	0	16	0.80	1.10715
5	0	19	0.95	1.34528
6	0	19	0.95	1.34528
7	0	17	0.85	1.17310
8	0	18	0.90	1.24905
9	0	20	1.00	1.57080
10	0	17	0.85	1.17310

etc.

Figure 15.6: `arcsine.sas - proc print`

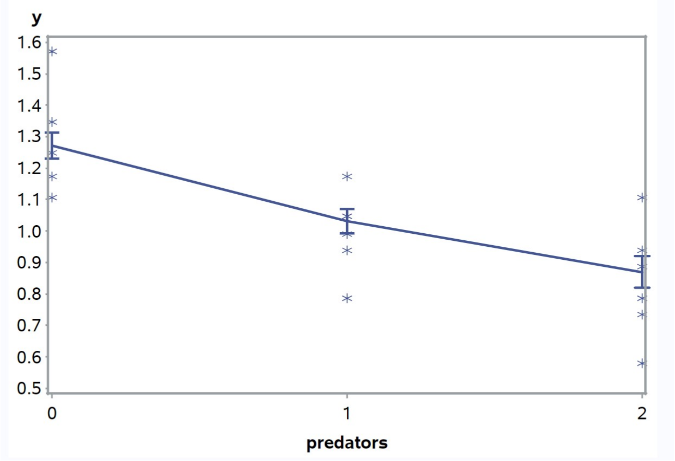
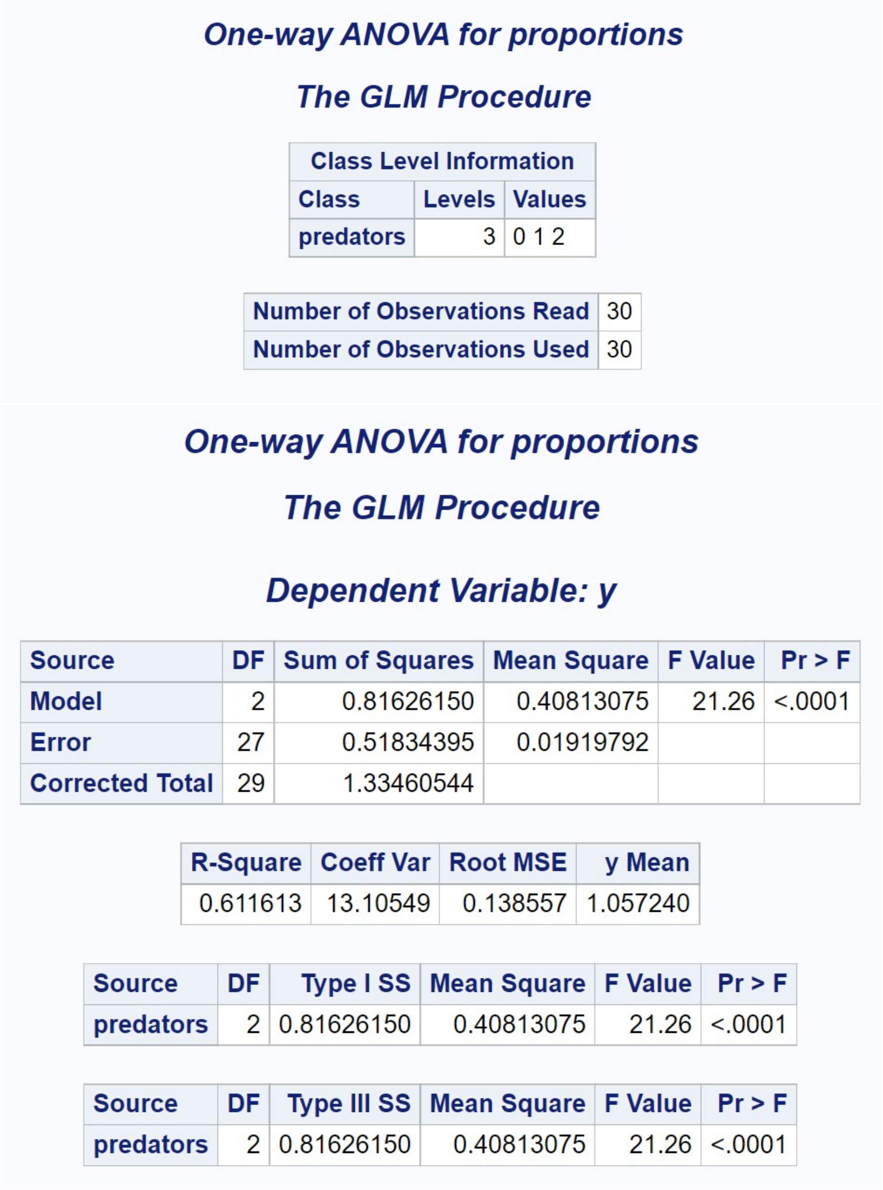


Figure 15.7: `arcsine.sas - proc gplot`



One-way ANOVA for proportions

The GLM Procedure

Class Level Information		
Class	Levels	Values
predators	3	0 1 2

Number of Observations Read	30
Number of Observations Used	30

One-way ANOVA for proportions

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.81626150	0.40813075	21.26	<.0001
Error	27	0.51834395	0.01919792		
Corrected Total	29	1.33460544			

R-Square	Coeff Var	Root MSE	y Mean
0.611613	13.10549	0.138557	1.057240

Source	DF	Type I SS	Mean Square	F Value	Pr > F
predators	2	0.81626150	0.40813075	21.26	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
predators	2	0.81626150	0.40813075	21.26	<.0001

Figure 15.8: `arcsine.sas - proc glm`

15.3.6 Transformations when data are limited

In many real studies, we will have insufficent data to determine the appropriate variance-stabilizing transformation using residual analysis. For example, we may not have enough points to determine if the variance is related to the mean, or whether the normality assumption is satisfied. In this situation you may have to guess the appropriate transformation. For count data you would use the $\sqrt{Y}$ or $\log Y$ transformation. Most count data are more overdispersed or clumped than the Poisson distribution, however, and so the $\log Y$ transformation will usually be a better choice than $\sqrt{Y}$. You would use the $\arcsin(\sqrt{Y})$ transformation for proportion data, especially if there are some proportions near 0 or 1.

15.4 References

Bartlett, M. S. (1947). The use of transformations. *Biometrics* 3: 39-52.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972) Consequences of failure to meet assumptions underlying fixed effects analysis of variance and covariance. *Review of Educational Research* 42: 237-288.

Hurlbert, S. H. (1984) Pseudoreplication and the design of ecological field experiments. *Ecological Monographs* 54: 187-211.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992) *Variance Components.* John Wiley & Sons, Inc., New York, NY.

Reeve, J. D., Strom, B. L., Rieske-Kinney, L. K., Ayres, B. D. Ayres, & Costa, A. (2009) Geographic variation in prey preference in bark beetle predators. *Ecological Entomology* 34: 183-192.

Winer, B. J., Brown, D. R. & Michels, K. M. (1991) *Statistical Principles in Experimental Design*, 3rd edition. McGraw-Hill, Inc., Boston, MA.

Chapter 16

Nonparametric Tests

The statistical tests we have examined so far are called **parametric tests**, because they assume the data have a known distribution, such as the normal, and test hypotheses about the parameters of this distribution. Examples of such tests are the F test in ANOVA, and one- or two-sample t tests. Parametric tests can also be constructed for other distributions, such as the Poisson and binomial.

While ANOVA and other procedures are derived assuming the data are normal, they can also be validly applied to non-normal data provided sample sizes are large, due to the central limit theorem (Glass et al. 1972). For example, the means used in the ANOVA F tests are assumed to have a normal distribution, which will be true for normal data. This will also hold for non-normal data, provided the sample sizes are sufficiently large for the central limit theorem to operate (Chapter 7). Thus, the tests used in ANOVA will still be valid for large sample sizes, regardless of the distribution of the data. Valid in this context means the tests have the correct Type I error rate (such as $\alpha = 0.05$) and power levels.

There are conditions where parametric procedures are less than ideal, such as non-normal data and relatively small sample sizes. We cannot rely on the central limit theorem here, and so parametric tests based on the normal distribution might be invalid. **Nonparametric tests** are often useful in this situation, because they do not assume a particular probability distribution for the data. For this reason they are also known as **distribution-free** methods. Nonparametric tests can be more powerful than parametric tests for non-normal data (Conover 1999; Hollander et al. 2014). The increase in power can be substantial for distributions with heavy tails compared to the normal distribution, which implies that extreme observations are more common. While nonparametric tests are less powerful than parametric ones for normal data, the loss of power is often quite minimal.

We will examine three types of nonparametric tests for one-way designs. The first are tests based on ranks. These replace the data values with their rank values, obtained by ordering the data from smallest to largest. They then utilize test statistics that are functions of these ranks rather than the original data values. We will cover rank tests for two or more groups, in particular the Wilcoxon and Kruskal-Wallis tests (Conover 1999; Hollander et al. 2014). They are used to test whether the distributions for each group differ in location, and serve a function similar to parametric tests like one-way ANOVA. We will also examine the two-sample Kolmogorov-Smirnov test, which can detect differences in both the shape and location of two distributions (Conover 1999; Hollander et al. 2014). It makes use of the empirical distribution function for each group, the empirical counterpart of the cumulative distribution function for continuous random variables (Chapter 6). The last type of nonparametric test we will consider are randomization tests. These tests examine whether the data are consistent with a null hypothesis of randomness (Hinkelmann & Kempthorne 1994; Manly 1997). The behavior of a test statistic (often a parametric one like an F statistic) is examined under this null hypothesis, in a process that involves randomly permuting or rearranging observations across the groups many times.

We will use data from a study of chitons (a kind of mollusk) in the intertidal zone (Flores-Campaña et al. 2012) to illustrate the use of nonparametric tests. Populations of *Chiton albolineatus* were sampled from three islands in Mazatlan Bay, Mexico. For each island, samples were taken from sites that were exposed or sheltered from wave action, and the body length of the chitons measured. The authors found that the distribution of chiton length was non-normal, and so used the nonparametric Kruskal-Wallis test to compare the lengths of chitons across islands and sites. They found significant differences in length among various combinations of island and site, and a tendency for chiton to be larger in exposed sites. We will use a small subset of these data in our calculations, shown in Tables 16.1 and 16.2.

Table 16.1: Example 1 - Body lengths of *Chiton albolineatus* in the intertidal zone of the island of Venados (Flores-Campaña et al. 2012). Chitons were sampled from sites sheltered or exposed to wave action. Also shown are the rank values (R_{ij}) for each observation, and the sum of the ranks for each groups ($\sum_{j=1}^{n_i} R_{ij}$, where n_i is the sample size for each group.)

Site	Y_{ij} = Length (mm)	R_{ij}	i	j	$\sum_{j=1}^{n_i} R_{ij}$
Sheltered	44.39	20	1	1	
Sheltered	22.30	3	1	2	
Sheltered	21.31	2	1	3	
Sheltered	23.80	5	1	4	
Sheltered	26.23	8	1	5	70
Sheltered	27.98	10	1	6	
Sheltered	28.10	11	1	7	
Sheltered	24.39	6	1	8	
Sheltered	22.32	4	1	9	
Sheltered	15.16	1	1	10	
Exposed	30.20	16	2	1	
Exposed	29.36	14	2	2	
Exposed	28.88	12	2	3	
Exposed	32.23	19	2	4	
Exposed	26.54	9	2	5	140
Exposed	24.85	7	2	6	
Exposed	30.54	17	2	7	
Exposed	31.36	18	2	8	
Exposed	28.98	13	2	9	
Exposed	29.49	15	2	10	

Table 16.2: Example 2 - Body length of *C. albolineatus* on the sheltered side of three islands, located in Mazatlan Bay, Mexico (Flores-Campaña et al. 2012). Also shown are the rank values (R_{ij}) for each observation, and the sum of the ranks for each group ($\sum_{j=1}^{n_i} R_{ij}$)

Site	Y_{ij} = Length (mm)	R_{ij}	i	j	$\sum_{j=1}^{n_i} R_{ij}$
Lobos	23.86	16	1	1	
Lobos	20.20	6	1	2	
Lobos	29.32	27	1	3	
Lobos	23.56	13	1	4	
Lobos	24.32	17	1	5	157
Lobos	22.33	12	1	6	
Lobos	23.69	14	1	7	
Lobos	26.78	21	1	8	
Lobos	27.32	23	1	9	
Lobos	21.22	8	1	10	
Pajaros	32.90	29	2	1	
Pajaros	32.73	28	2	2	
Pajaros	26.94	22	2	3	
Pajaros	29.09	26	2	4	
Pajaros	12.32	1	2	5	142
Pajaros	15.25	5	2	6	
Pajaros	25.87	19	2	7	
Pajaros	20.21	7	2	8	
Pajaros	13.96	3	2	9	
Pajaros	12.48	2	2	10	
Venados	44.39	30	3	1	
Venados	22.30	10	3	2	
Venados	21.31	9	3	3	
Venados	23.80	15	3	4	
Venados	26.23	20	3	5	166
Venados	27.98	24	3	6	
Venados	28.10	25	3	7	
Venados	24.39	18	3	8	
Venados	22.32	11	3	9	
Venados	15.16	4	3	10	

16.1 Wilcoxon two-sample test

The Wilcoxon test provides a nonparametric alternative to a two-sample t test or a one-way ANOVA for two groups (see Chapter 11). It does not assume any particular distribution of the data, except that it is a continuous one (see Chapter 6). The null and alternative hypotheses for the Wilcoxon test are expressed in terms of the cumulative distribution for the two groups, say $F_1(y)$ and $F_2(y)$. Under the null hypothesis the two distribution are supposed to be identical, which can be expressed as $H_0 : F_2(y) = F_1(y)$ for all y (Fig. 16.1). Under the alternative, one distribution is shifted from the other by a distance Δ, but they otherwise have the same shape (Conover 1999; Hollander et al. 2014). This can be expressed as $H_1 : F_2(y) = F_1(y-\Delta)$ (Fig. 16.2).

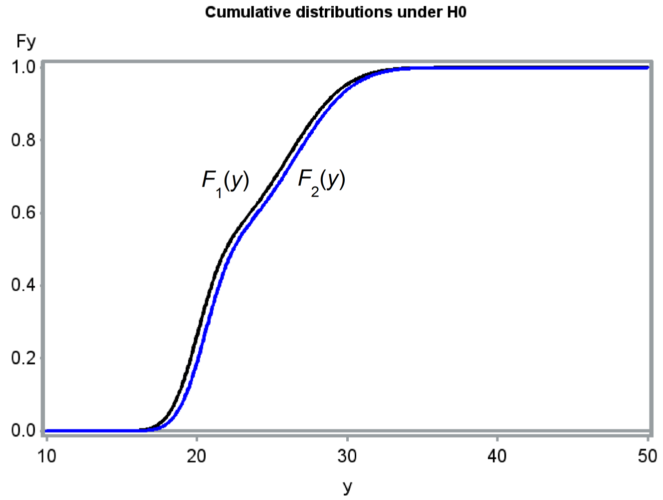


Figure 16.1: Cumulative distributions for two groups under $H_0 : \Delta = 0$.

The Wilcoxon test statistic W is based on the ranks of the observations. The observations are first assigned ranks from the smallest to the largest across the two groups. The test statistic is then the sum of the ranks for one of the groups. Typically the one with the smallest sample size is chosen, or if the sample sizes are equal, one is arbitrarily selected (SAS uses group order). For the Example 1 data the sample sizes are equal, so we could use

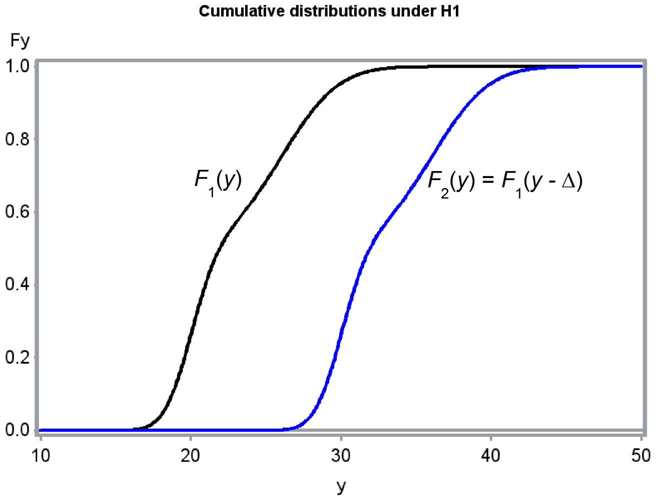


Figure 16.2: Cumulative distributions for two groups under $H_1 : \Delta = 10$.

the summed ranks for the Sheltered chiton group, namely

$$W = \sum_{j=1}^{n_1} R_{1j} = 70 \tag{16.1}$$

(Conover 1999; Hollander et al. 2014). We would expect small values of this statistic when F_1 is located to the left of F_2 ($\Delta > 0$), because this implies that values of Y_{1j} are more likely to be small relative to Y_{2j} ones. Conversely, large values of the statistic would occur when F_1 is to the right of F_2 ($\Delta < 0$). W is also sensitive to differences in the expected values (means) of the two distributions, because of the relationship between expected values and distributions. For a two-tailed test, we would reject H_0 if W is sufficiently large, or sufficiently small. An exact P value for both one- and two-tailed tests can be calculated using the distribution of W. We will let SAS handle the calculations for exact tests.

For large sample sizes, the distribution of W under H_0 approaches the normal distribution with mean and variance given by

$$E_{H_0}[W] = \frac{n_1(n_1 + n_2 + 1)}{2} \tag{16.2}$$

and

$$Var_{H_0}[W] = \frac{n_1 n_2(n_1 + n_2 + 1)}{12}. \tag{16.3}$$

The expected value formula assumes W is calculated using the first group. We then have

$$Z = \frac{W - E_{H_0}[W]}{\sqrt{Var_{H_0}[W]}} \sim N(0,1) \tag{16.4}$$

for large sample sizes. We can use this approximation to find P values for both one- and two-tailed tests (Hollander et al. 2014).

The Wilcoxon statistic W can be derived starting with a two-sample t test (see Chapter 11), and simply replacing the observations with their rank values (Bickel & Doksum 1977). It is also equivalent to the Mann-Whitney U test, another common nonparametric test. Modifications of the Wilcoxon test are also available to deal with the problem of tied observations. The tied observations are assigned the average of the tied ranks, and the variance equation is modified to account for the number of ties (Hollander et al. 2014).

Wilcoxon test - sample calculation

For the Example 1 data, we see that $W = 70$ for the Sheltered chitons (see Table 16.1). We will use the normal approximation for this statistic to obtain a two-tailed P value for the test. We have $E_{H_0}[W] = 10(10+10+1)/2 = 105$ and $Var_{H_0}[W] = 10 \cdot 10(10+10+1)/12 = 175$, and so

$$Z = \frac{70 - 105}{\sqrt{175}} = -2.646. \tag{16.5}$$

From Table Z, we find that $P[Z < -2.646] = 1 - P[Z < 2.646] \approx 1 - 0.9960 = 0.0040$. The two-tailed P value is then twice this value, or $P = 2(0.0040) = 0.0080$.

16.1.1 Wilcoxon test for Example 1 - SAS demo

We now conduct the Wilcoxon test using the Example 1 data and the SAS procedure `npar1way`, which implements a number of nonparametric procedures for one-way (single factor) designs (SAS Institute Inc. 2018). See program listing below. The Wilcoxon test is invoked by adding the `wilcoxon` option in the `proc npar1way` statement. The `class` statement identifies the group variable, while `var` selects the dependent variable. The `exact wilcoxon` line generates exact P values for the test. The program also includes `proc gplot` code to plot the group means (SAS Institute Inc. 2016a). For purposes of comparison, a one-way ANOVA is also conducted using `proc glm`. See program and output below (Fig. 16.3-16.6).

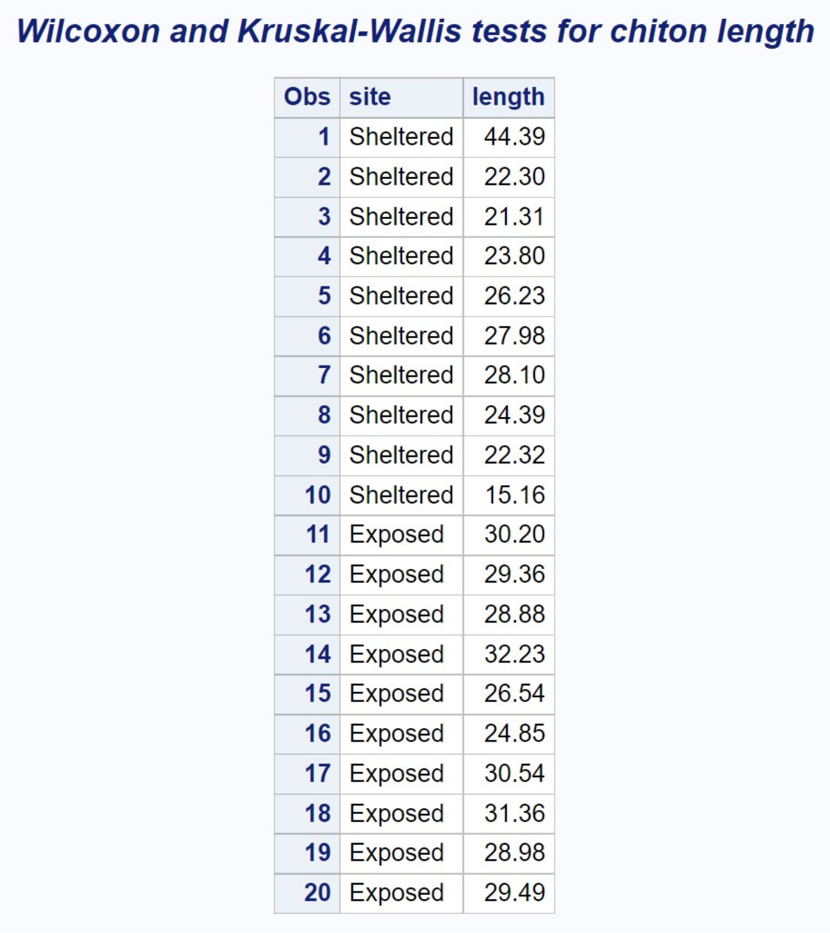
We see that the Wilcoxon two-tailed test was highly significant, for both the exact test ($W = 70, P = 0.0068$) and the normal approximation ($Z = -2.6080, P = 0.0091$). The value of Z calculated by SAS differs slightly from our earlier result, because it includes a correction that improves the normal approximation. From the summed ranks for each group, as well as the graph, it appears that the Sheltered chitons were smaller than the Exposed ones. Note that the parametric one-way ANOVA for these data was non-significant ($F_{1,18} = 2.13, P = 0.1619$). This likely occurred because of one very large and one small chiton at the Sheltered site, which would be outliers in the ANOVA. In the analysis using ranks, these are simply the largest and smallest rank values, only one step away from the next ones.

SAS Program

```
* WKWtest_chitons_Venados.sas;
title 'Wilcoxon and Kruskal-Wallis tests for chiton length';
data chitons;
    input site :$10. length;
    datalines;
Sheltered   44.39
Sheltered   22.30
Sheltered   21.31
Sheltered   23.80
Sheltered   26.23

etc.

;
run;
* Print data set;
proc print data=chitons;
run;
* Plot means, standard error, and observations;
proc gplot data=chitons;
    plot length*site / vaxis=axis1 haxis=axis1;
    symbol1 i=std1mjt v=star height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* Kruskal-Wallis/Wilcoxon tests;
proc npar1way wilcoxon data=chitons;
    class site;
    var length;
    exact wilcoxon;
run;
* One-way ANOVA for comparison;
proc glm data=chitons;
    class site;
    model length = site;
run;
quit;
```



Wilcoxon and Kruskal-Wallis tests for chiton length

Obs	site	length
1	Sheltered	44.39
2	Sheltered	22.30
3	Sheltered	21.31
4	Sheltered	23.80
5	Sheltered	26.23
6	Sheltered	27.98
7	Sheltered	28.10
8	Sheltered	24.39
9	Sheltered	22.32
10	Sheltered	15.16
11	Exposed	30.20
12	Exposed	29.36
13	Exposed	28.88
14	Exposed	32.23
15	Exposed	26.54
16	Exposed	24.85
17	Exposed	30.54
18	Exposed	31.36
19	Exposed	28.98
20	Exposed	29.49

Figure 16.3: `WKWtest_chitons_Venados.sas - proc print`

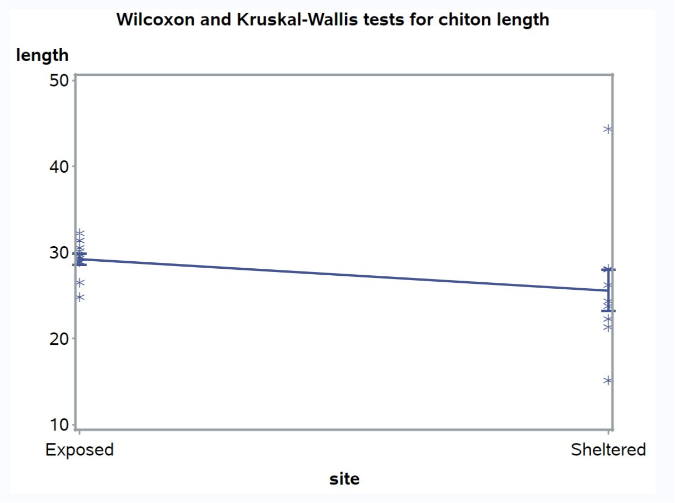
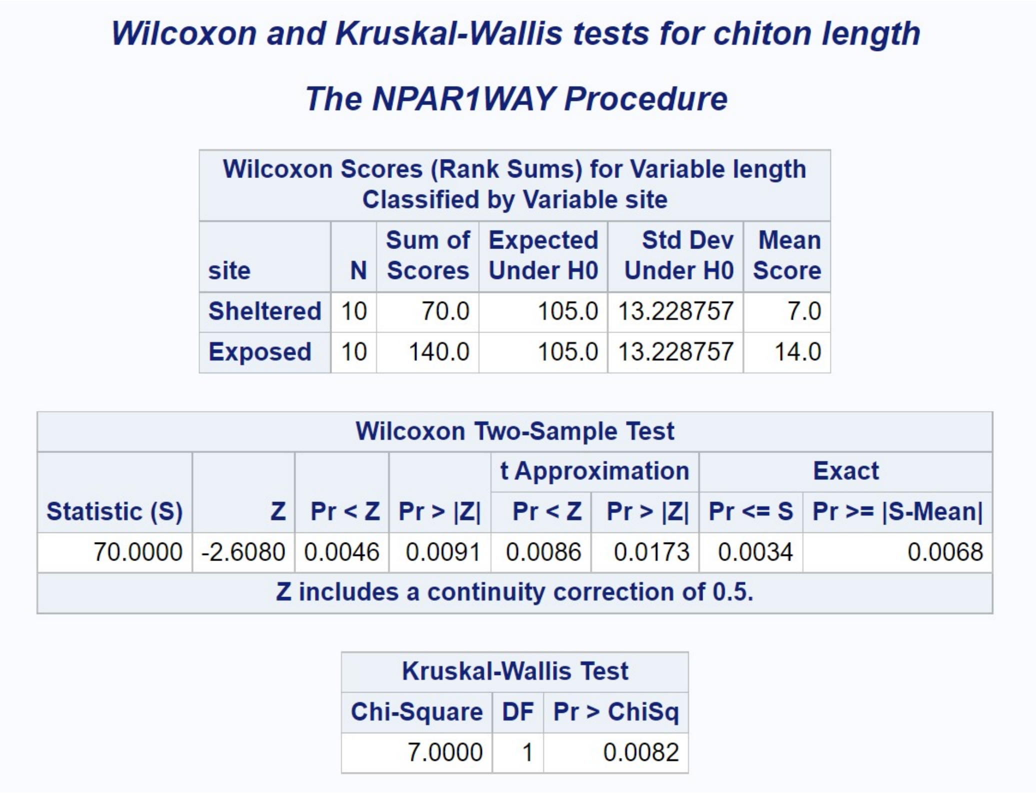


Figure 16.4: `WKWtest_chitons_Venados.sas - proc gplot`



Wilcoxon and Kruskal-Wallis tests for chiton length

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable length Classified by Variable site

site	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Sheltered	10	70.0	105.0	13.228757	7.0
Exposed	10	140.0	105.0	13.228757	14.0

Wilcoxon Two-Sample Test

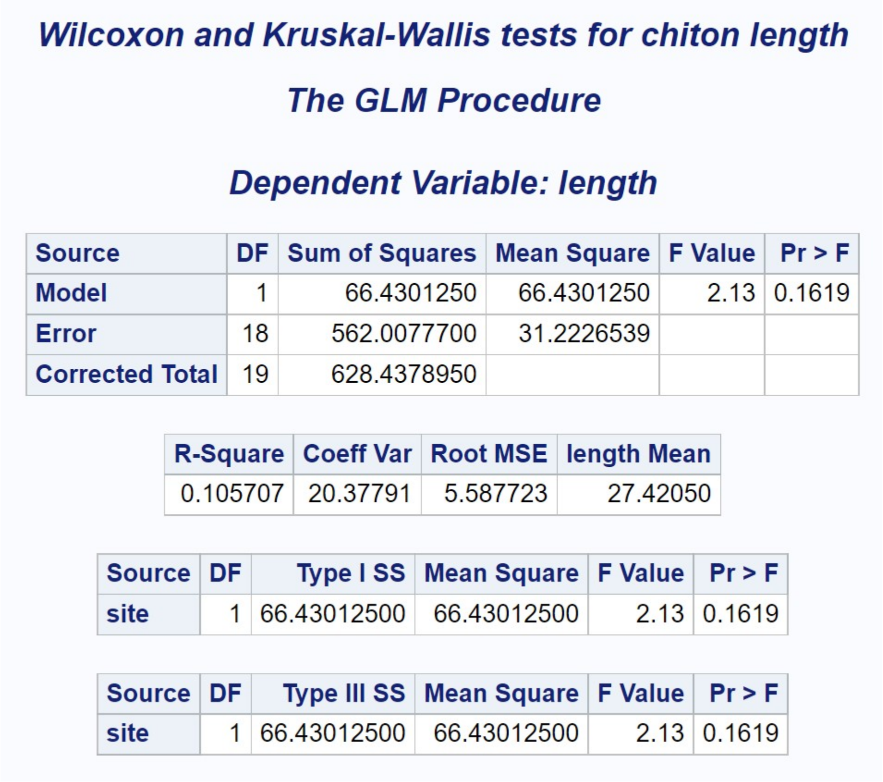
				t Approximation		Exact	
Statistic (S)	Z	Pr < Z	Pr > |Z|	Pr < Z	Pr > |Z|	Pr <= S	Pr >= |S-Mean|
70.0000	-2.6080	0.0046	0.0091	0.0086	0.0173	0.0034	0.0068

Z includes a continuity correction of 0.5.

Kruskal-Wallis Test

Chi-Square	DF	Pr > ChiSq
7.0000	1	0.0082

Figure 16.5: `WKWtest_chitons_Venados.sas - proc npar1way`



Wilcoxon and Kruskal-Wallis tests for chiton length

The GLM Procedure

Dependent Variable: length

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	66.4301250	66.4301250	2.13	0.1619
Error	18	562.0077700	31.2226539		
Corrected Total	19	628.4378950			

R-Square	Coeff Var	Root MSE	length Mean
0.105707	20.37791	5.587723	27.42050

Source	DF	Type I SS	Mean Square	F Value	Pr > F
site	1	66.43012500	66.43012500	2.13	0.1619

Source	DF	Type III SS	Mean Square	F Value	Pr > F
site	1	66.43012500	66.43012500	2.13	0.1619

Figure 16.6: `WKWtest_chitons_Venados.sas - proc glm`

16.2 Kruskal-Wallis test

The Kruskal-Wallis test is an extension of rank methods to one-way designs with three or more groups. The null and alternative hypotheses are similar to the Wilcoxon test, with the cumulative distributions for the different groups the same under H_0, and differing by shift parameters under H_1. The Kruskal-Wallis test is sensitive to these shifts as well as differences among the means of the groups.

The Kruskal-Wallis test statistic H is calculated using the ranks of the observations across all groups. Suppose we have a different groups, and for simplicity assume the same sample size n for each group. The Kruskal-Wallis test statistic is

$$H = \frac{12n}{an(an+1)} \sum_{i=1}^{a} \left(\frac{\sum_{j=1}^{n} R_{ij}}{n} - \frac{an+1}{2} \right)^2 \tag{16.6}$$

(Conover 1999; Hollander et al. 2014). Note that the left term in parentheses is the mean rank for each group, while the right one is the mean rank across all the groups. This implies that H will become large when the mean rank differs among groups, similar to the way differences in the group means affect the F statistic for one-way ANOVA. In fact, the Kruskal-Wallis statistic can be derived from the F test by substituting ranks for the observations (Bickel & Doksum 1977). A more complex form of H is used when sample sizes are unequal, or when there are ties in the data. Under H_0, H has approximately a χ^2 distribution with $a - 1$ degrees of freedom.

Kruskal-Wallis test - sample calculation

We will illustrate the Kruskal-Wallis test using both the Example 1 and 2 data sets. For Example 1, we have two groups with ten observations each, so $a = 2$ and $n = 10$. The summed ranks for the two groups are 70 (Sheltered) and 140 (Exposed). It follows that

$$
\begin{aligned}
H &= \frac{12 \cdot 10}{2 \cdot 10(2 \cdot 10 + 1)} \left[\left(\frac{70}{10} - \frac{2 \cdot 10 + 1}{2} \right)^2 + \left(\frac{140}{10} - \frac{2 \cdot 10 + 1}{2} \right)^2 \right] \\
&= \frac{120}{420} \left[(7 - 10.5)^2 + (14 - 10.5)^2 \right] \\
&= 0.2857 \left[12.25 + 12.25 \right] \\
&= 7.00.
\end{aligned}
$$

The degrees of freedom are $a - 1 = 2 - 1 = 1$. From Table C, we find that $P < 0.01$, and so the Exposed and Sheltered chitons were significantly different in length $(H = 7.00, df = 1, P < 0.01)$.

The Example 2 data involves chitons collected from three different islands $(a = 3)$, with ten chitons sampled per island $(n = 10)$. The summed ranks for the three islands are 157, 142, and 166. From this information, we calculate that

$$
\begin{aligned}
H = {} & \frac{12 \cdot 10}{3 \cdot 10(3 \cdot 10 + 1)} \\
& \cdot \left[\left(\frac{157}{10} - \frac{3 \cdot 10 + 1}{2} \right)^2 + \left(\frac{142}{10} - \frac{3 \cdot 10 + 1}{2} \right)^2 + \left(\frac{166}{10} - \frac{3 \cdot 10 + 1}{2} \right)^2 \right] \\
= {} & \frac{120}{930} \left[(15.7 - 15.5)^2 + (14.2 - 15.5)^2 + (16.6 - 15.5)^2 \right] \\
= {} & 0.129 \left[0.04 + 1.69 + 1.21 \right] \\
= {} & 0.38.
\end{aligned}
$$

The degrees of freedom are $a - 1 = 3 - 1 = 2$. From Table C, we find that $P < 0.9$. There was no significant difference in length among the three islands $(H = 0.38, df = 2, P < 0.9)$.

16.2.1 Kruskal-Wallis test for Example 1 - SAS demo

The Kruskal-Wallis test is automatically calculated when the `wilcoxon` option for `proc npar1way` is used (see previous output). We see there was a highly significant difference in length betwee the Sheltered and Exposed sites $(H = 7.00, df = 1, P = 0.0082)$.

16.2.2 Kruskal-Wallis test for Example 2 - SAS demo

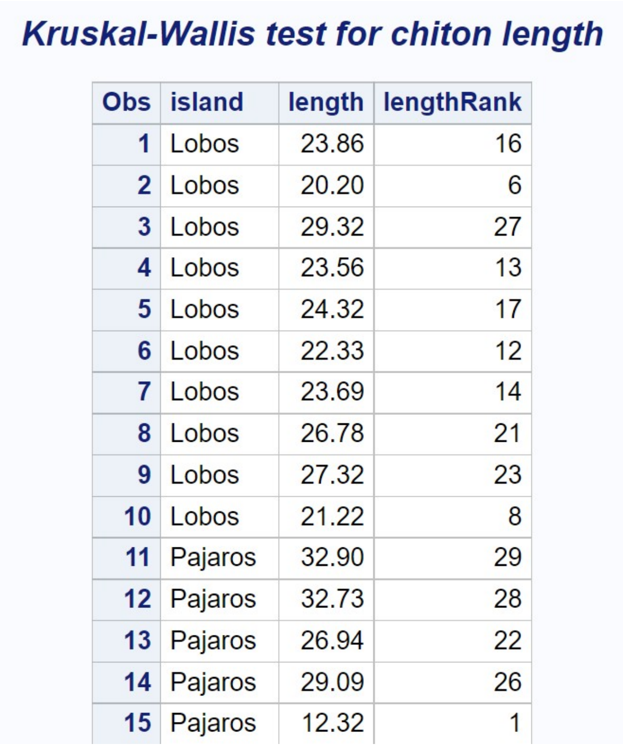
The Kruskal-Wallis test for the Example 2 data is shown below (Fig. 16.9). There was no significant difference in length among the three islands ($H = 0.38, df = 2, P = 0.8272$). Note that an exact version of this test is also provided ($P = 0.8386$).

SAS Program

```
* KWtest_chitons_3islands.sas;
title 'Kruskal-Wallis test for chiton length';
data chitons;
    input island $ length;
    datalines;
Lobos       23.86
Lobos       20.20
Lobos       29.32
Lobos       23.56
Lobos       24.32

etc.

;
run;
* Print data set;
proc print data=chitons;
run;
* Plot means, standard error, and observations;
proc gplot data=chitons;
    plot length*island / vaxis=axis1 haxis=axis1;
    symbol1 i=std1mjt v=star height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* Kruskal-Wallis/Wilcoxon tests;
proc npar1way wilcoxon data=chitons;
    class island;
    var length;
    exact wilcoxon;
run;
quit;
```



Kruskal-Wallis test for chiton length

Obs	island	length	lengthRank
1	Lobos	23.86	16
2	Lobos	20.20	6
3	Lobos	29.32	27
4	Lobos	23.56	13
5	Lobos	24.32	17
6	Lobos	22.33	12
7	Lobos	23.69	14
8	Lobos	26.78	21
9	Lobos	27.32	23
10	Lobos	21.22	8
11	Pajaros	32.90	29
12	Pajaros	32.73	28
13	Pajaros	26.94	22
14	Pajaros	29.09	26
15	Pajaros	12.32	1

etc.

Figure 16.7: `KWtest_chitons_3islands.sas - proc print`

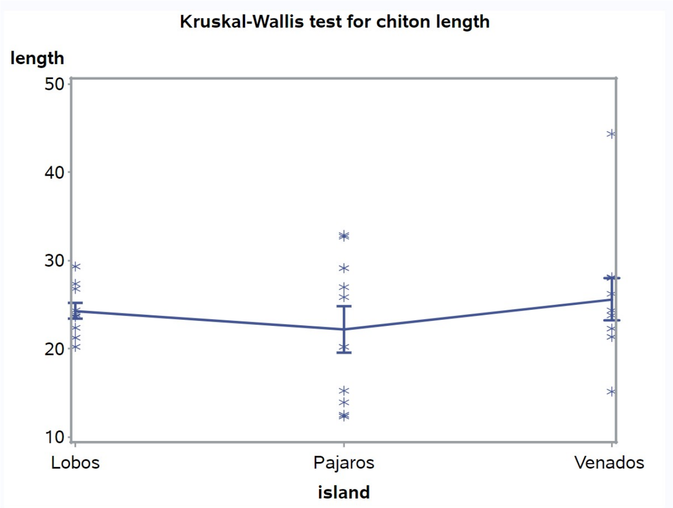
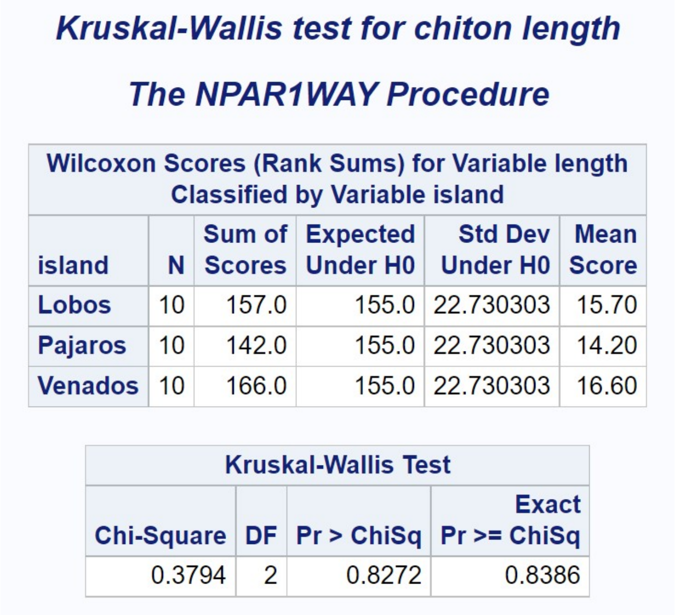


Figure 16.8: `KWtest_chitons_3islands.sas - proc gplot`



Kruskal-Wallis test for chiton length

The NPAR1WAY Procedure

Wilcoxon Scores (Rank Sums) for Variable length Classified by Variable island					
island	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
Lobos	10	157.0	155.0	22.730303	15.70
Pajaros	10	142.0	155.0	22.730303	14.20
Venados	10	166.0	155.0	22.730303	16.60

Kruskal-Wallis Test			
Chi-Square	DF	Pr > ChiSq	Exact Pr >= ChiSq
0.3794	2	0.8272	0.8386

Figure 16.9: `KWtest_chitons_3islands.sas - proc npar1way`

16.3 Kolmogorov-Smirnov test

The Kolmogorov-Smirnov test is a nonparametric procedure used to compare the probability densities of two groups or samples, using their cumulative distributions (see Chapter 6). Let $F_1(y)$ be the cumulative distribution function for the first group, while $F_2(y)$ is the second. The null hypothesis for the Kolmogorov-Smirnov test is $H_0 : F_2(y) = F_1(y)$, which means that the two groups have the same distribution. The alternative hypothesis is $H_1 : F_2(y) \neq F_1(y)$ for some y, implying there is some difference in the distributions, which could involve their location, general shape, variance, and so forth. This is a broader alternative hypothesis than the rank tests we examined earlier, where the distributions had the same shape but differed by location.

The Kolmogorov-Smirnov test statistic is calculated using the empirical distribution functions of the two groups, which estimates the underlying cumulative distribution function. For a sample with n_i observations, the empirical distribution function is defined as

$$G_i(y) = \frac{\text{Number of } Y_{ij} \text{ values} \leq y}{n_i}. \tag{16.7}$$

$G_i(y)$ increases in a step-like fashion as y increases, with a jump occurring at every value of Y_{ij} (Conover 1999; Hollander et al. 2014). Fig. 16.10 shows these functions for the two sites in Example 1. The Kolmogorov-Smirnov test uses the maximum vertical distance between the two functions as the test statistic. The distance is defined using the formula

$$D = \max_{y} |G_1(y) - G_2(y)| \tag{16.8}$$

(Conover 1999; Hollander et al. 2014). D is the largest distance between $G_1(y)$ and $G_2(y)$ over all values of y, with the absolute value making it a positive quantity. We would then reject H_0 for sufficiently large values of D. The P value for the test can be calculated exactly for small sample sizes, and there is also a large sample approximation for the test. We will let SAS handle the details. This test can also be used when there are ties in the observations, in which case it is conservative, meaning it is less likely to reject H_0 (Hollander et al. 2014).

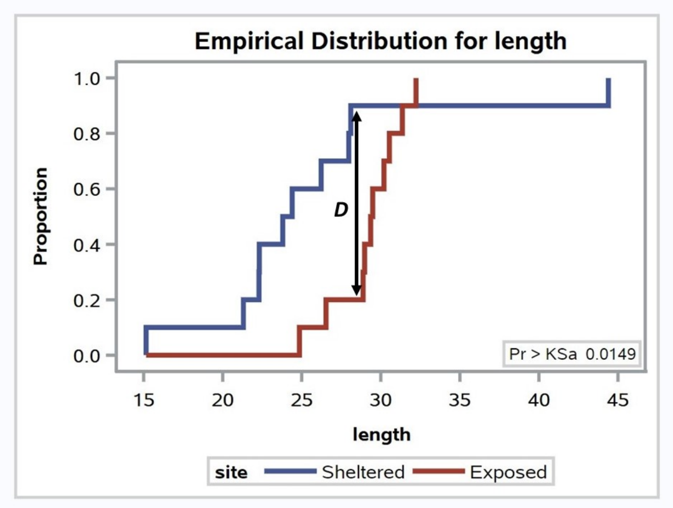


Figure 16.10: Empirical distribution functions for the Example 1 data. Also shown is the maximum value of D for the two samples.

16.3.1 Kolmogorov-Smirnov test for Example 1 - SAS demo

The SAS procedure `npar1way` can also be used for the Kolmogorov-Smirnov test (SAS Institute Inc. 2018). It is invoked by adding the `edf` option in the `proc npar1way` statement (see program below). This option also generates a graph of the empirical distribution function for the two groups (Fig. 16.10). An exact version of test can be calculated using the line `exact ks`. The program also includes `proc gchart` code to generate histograms of the two groups (SAS Institute Inc. 2016a). This seems more appropriate for the Kolmogorov-Smirnov test than plotting the means, because this test can detect differences in both shape and location. Examining the SAS output, we see that $D = 0.7$ (Fig. 16.12). The P value for the exact version of the test was significant ($P = 0.0123$), implying there was some difference in the distributions of the two sites. The graph generated by `proc gchart` suggests they differed in both location and variance (Fig. 16.11).

SAS Program

```
* KStest_chitons_Venados.sas;
title 'Kolmogorov-Smirnov test for chiton length';
data chitons;
    input site :$10. length;
    datalines;
Sheltered   44.39
Sheltered   22.30
Sheltered   21.31
Sheltered   23.80
Sheltered   26.23

etc.

;
run;
* Print data set;
proc print data=chitons;
run;
* Histograms for the two groups;
proc gchart data=chitons;
    vbar length / group=site axis=axis1 gaxis=axis1 maxis=axis2;
    axis1 label=(height=2) value=(height=2) width=3 minor=none;
    axis2 label=(height=1.5) value=(height=1.5) width=1.5;
run;
* Kolmogorov-Smirnov test;
proc npar1way edf data=chitons;
    class site;
    var length;
    exact ks;
run;
quit;
```

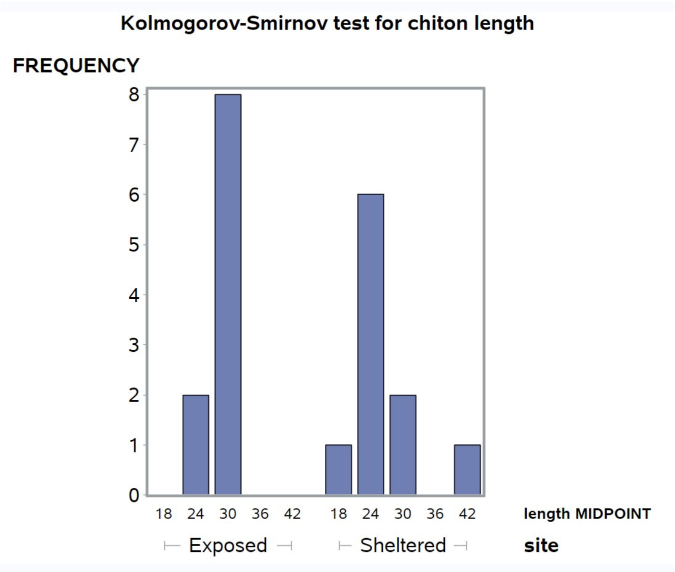
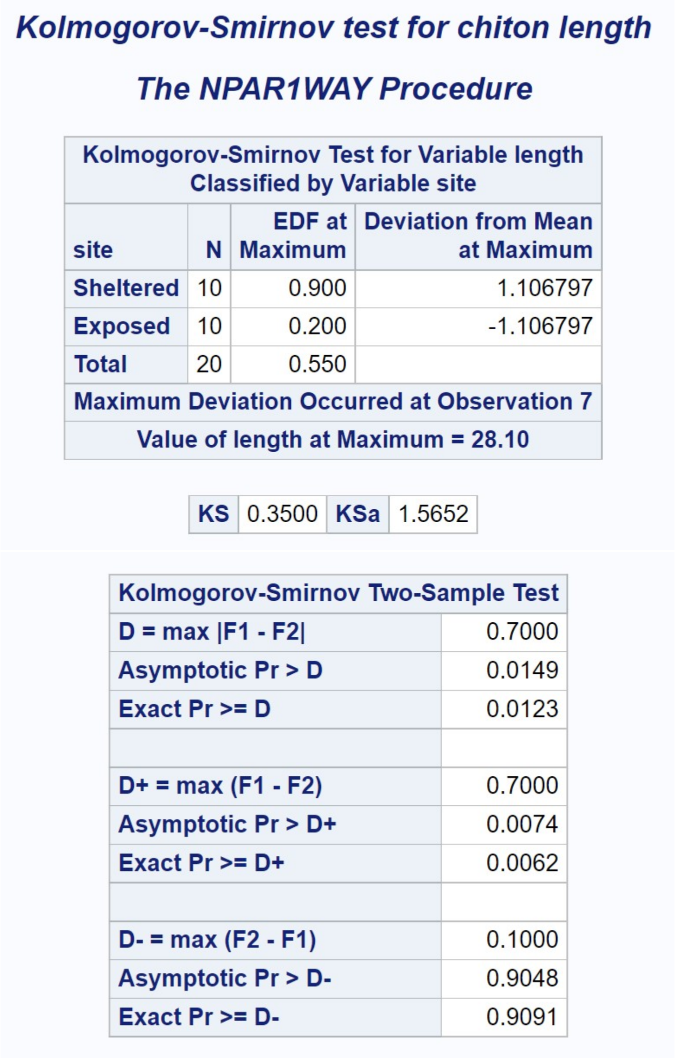


Figure 16.11: `KStest_chitons_Venados.sas - proc gchart`



Kolmogorov-Smirnov test for chiton length

The NPAR1WAY Procedure

Kolmogorov-Smirnov Test for Variable length
Classified by Variable site

site	N	EDF at Maximum	Deviation from Mean at Maximum
Sheltered	10	0.900	1.106797
Exposed	10	0.200	-1.106797
Total	20	0.550	

Maximum Deviation Occurred at Observation 7

Value of length at Maximum = 28.10

KS	0.3500	KSa	1.5652

Kolmogorov-Smirnov Two-Sample Test	
D = max |F1 - F2|	0.7000
Asymptotic Pr > D	0.0149
Exact Pr >= D	0.0123
D+ = max (F1 - F2)	0.7000
Asymptotic Pr > D+	0.0074
Exact Pr >= D+	0.0062
D- = max (F2 - F1)	0.1000
Asymptotic Pr > D-	0.9048
Exact Pr >= D-	0.9091

Figure 16.12: `KStest_chitons_Venados.sas - proc npar1way`

16.4 Randomization tests

Randomization tests are another common kind of nonparametric test used for one-way designs, as well as more complex ones (Hinkelmann & Kempthorne 1994; Manly 1997). The null hypothesis for these tests is different from other tests we have considered, which involved statements about probability distributions and their parameters. For randomization tests, the null hypothesis is that all possible permutations (rearrangements) of the data among groups are equally likely, given no treatment or group effects, with the observed data being one such arrangement (Hinkelmann & Kempthorne 1994; Manly 1997). These tests commonly employ a parametric test statistic to examine the null hypothesis, one that is sensitive to potential differences among groups. For one-way designs, the F_s statistic from one-way ANOVA (Chapter 11) is often used to detect differences in the group means. To conduct a randomization test using this statistic, we first calculate the value of $F_s(obs)$ for the observed data. Similar to one-way ANOVA, we then need to determine if $F_s(obs)$ is sufficiently large to consider rejecting H_0. This is accomplished by permuting or rearranging the observations many times across groups, and calculating the value of F_s for each permutation. The justification for this procedure follows directly from the definition of H_0. The P value for the test is defined as the proportion of the F_s values greater than or equal to $F_s(obs)$, including $F_s(obs)$ as one of the values.

For small data sets it may be possible to carry out all possible permutations, but for larger data sets this may be impractical. Instead, the observations are randomly rearranged across groups a large number of times, in effect drawing a random sample from all possible permutations. The collection of F_s values obtained by this process is called the **randomization distribution**. How many of these randomizations are needed to generate an accurate P value for the test? Some guidance is provided by Manly (1997), who suggests that 1000 randomizations should be sufficient for $P \approx 0.05$, and 5000 for $P \approx 0.01$.

An interesting feature of randomization tests is that the randomization distribution of F_s under H_0 can be approximated by the parametric F distribution (Hinkelmann & Kempthorne 1974) under some conditions. This provides another justification for the use of F tests when the normality assumption of these tests is violated.

We will use data on nematode intensities for male vs. female bobcats (*Lynx rufus*) to illustrate randomization tests. The sampled bobcats were recent roadkill collected from the Southern Illinois region (Francisco A. Jimenez-Ruiz and Eliot A. Zieman, unpublished data). The guts were examined for nematodes as well as other parasites, and the total number counted (Table 16.3). These data have many zeroes as well as large values, as is common for parasite intensity data. The data are clearly non-normal and so a nonparametric test seems warranted.

Table 16.3: Example 3 - Number of nematode parasites found in the gut of male and female bobcats collected from Southern Illinois .

Sex	Nematodes	Sex	Nematodes	Sex	Nematodes	Sex	Nematodes
F	0	F	0	M	6	M	8
F	8	F	5	M	10	M	0
F	0	F	0	M	1	M	60
F	0	F	0	M	0	M	25
F	0	F	0	M	5	M	1
F	0	F	11	M	59	M	0
F	0	F	0	M	2	M	74
F	1	F	5	M	3	M	3
F	2	F	11	M	0	M	1
F	1	F	0	M	44	M	15
F	1	F	24	M	1	M	0
F	6	F	13	M	1	M	7
F	1	F	2	M	0	M	0
F	6			M	2	M	0
F	2			M	17		
F	1			M	5		
F	13			M	3		
F	0			M	26		
F	0			M	20		
F	7			M	3		

16.4.1 Randomization test for Example 3 - SAS demo

We will analyze the bobcat data using both one-way ANOVA and the analogous randomization test, comparing the parasite intensities for male vs. female cats. The SAS program below first generates a graph showing the mean intensities for both sexes, then conducts a standard one-way ANOVA (Fig. 16.14, 16.15). We see that the mean intensity for male bobcats was higher than females, and the ANOVA showed this difference was significant ($F_{1,65} = 5.50, P = 0.0221$).

The program then uses two SAS macro programs to conduct the randomization test (Cassell 2002). SAS macros are chunks of code that are used to carry out custom calculations, ones not available in standard SAS procedures (SAS Institute Inc. 2016b). They are inserted into a main program through the use of `%include` statements, which point to the file locations of the macros on the user's computer. Note that the percent sign (`%`) tells SAS that a particular line contains macro code. The first macro, `%rand_gen.sas`, is used to generate the desired number of random permutations of the data. Once the macro is included in the program, it can be called using the following arguments. The input data set is specified using the `indata=parasites` statement, while the output data set specified by `outdata=outrand` contains all the randomizations. The statement `numreps=5000` sets the number of randomizations, with the dependent variable specified by `depvar=nematodes`.

The next step in the randomization test is to conduct a one-way ANOVA for each one of the randomizations, as well as the original data set. This is accomplished using `proc glm` with a `by replicate` statement. The variable `replicate` is generated by the `rand_gen` macro to number the different randomizations. In addition, a data file containing the statistical output of the ANOVA is specified using the statement `outstat=outstat1`. The ANOVA for the original data corresponds to a `replicate = 0` in this output file. The `noprint` option is used to suppress the printing of each ANOVA.

The last step in the randomization test uses the second macro, `%rand_anl.sas`, to determine the P value for the test. The data file containing the statistical output from `proc glm` is specified using a `randdata=outstat1` argument. The `where=_source_='sex' and _type_='SS3'` argument tells the macro which part of the statistical output to use, in particular the test associated with the sex effect and Type III sum of squares. The `testprob=prob` statement tells the macro to use the P value for this F test in calculating the P value for the randomization test. The macro uses the P rather than F_s value to provide some additional flexibility for other kinds of tests (Cassell 2002). As the F_s and P value for the ANOVA are related, it yields the same result. *The P value for the randomization test is provided in the SAS log.* The `testlabel=Model F test` argument provides some labeling for this output. Examining the SAS log, we find that the randomization test was significant ($P = 0.0182$). The P value for this test was smaller than the one found using one-way ANOVA, and makes no assumptions about the distribution of the data.

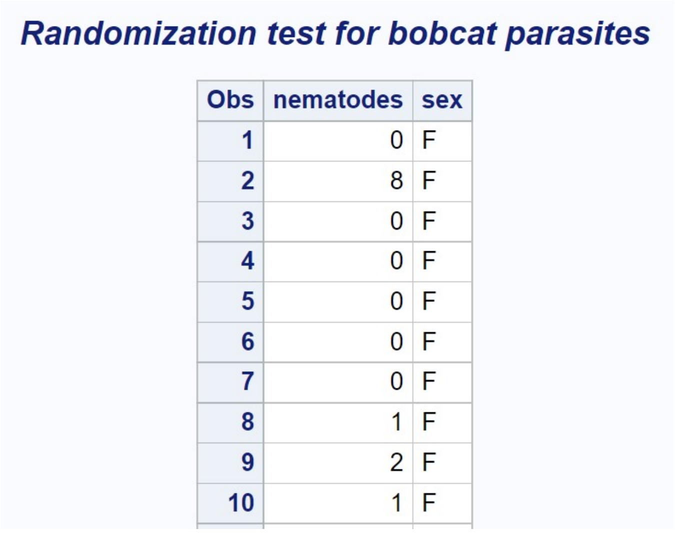
The remaining portion of the program generates a graph of the randomization distribution of F_s, and displays the value of this statistic for the original distribution (Fig. 16.16). We see that the original value of F_s lies far above most of the randomizations. This illustrates the pattern for a significant randomization test. For a non-significant test, we would see an F_s value that is more central within the randomization distribution.

SAS Program

```
* Randtest_bobcat_parasites.sas;
title 'Randomization test for bobcat parasites';
data parasites;
    input nematodes sex $;
    datalines;
0  F
8  F
0  F
0  F
0  F

etc.

;
run;
* Print data set;
proc print data=parasites;
run;
* Plot means, standard error, and observations;
proc gplot data=parasites;
    plot nematodes*sex / vaxis=axis1 haxis=axis1;
    symbol1 i=std1mjt v=star height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* One-way ANOVA;
proc glm data=parasites;
    class sex;
    model nematodes = sex;
run;
* Include two macros for randomization test;
%include "/home/u49852288/sasuser.v94/Statistics Book 2/Chapter 16/rand_gen.sas";
%include "/home/u49852288/sasuser.v94/Statistics Book 2/Chapter 16/rand_anl.sas";
* Sampled randomization test;
%rand_gen(indata=parasites,outdata=outrand,depvar=nematodes,numreps=5000)
proc glm data=outrand noprint outstat=outstat1;
    by replicate;
    class sex;
    model nematodes = sex;
run;
%rand_anl(randdata=outstat1,where=_source_='sex' and _type_='SS3',testprob=prob,testlabel=Mo
* Extract F values from outstat1 for null distribution graph;
data nulldist;
    set outstat1;
    if _type_="SS3";
    * Assign original F value to macro variable;
    if replicate=0 then call symput(’F’,F);
run;
* Null distribution;
title2 "Null distribution";
proc univariate data=nulldist noprint;
    var F;
    histogram F / vscale=count href=&F hreflabel="F";
run;
quit;
```



Randomization test for bobcat parasites

Obs	nematodes	sex
1	0	F
2	8	F
3	0	F
4	0	F
5	0	F
6	0	F
7	0	F
8	1	F
9	2	F
10	1	F

etc.

Figure 16.13: `Randtest_bobcat_parasites.sas - proc print`

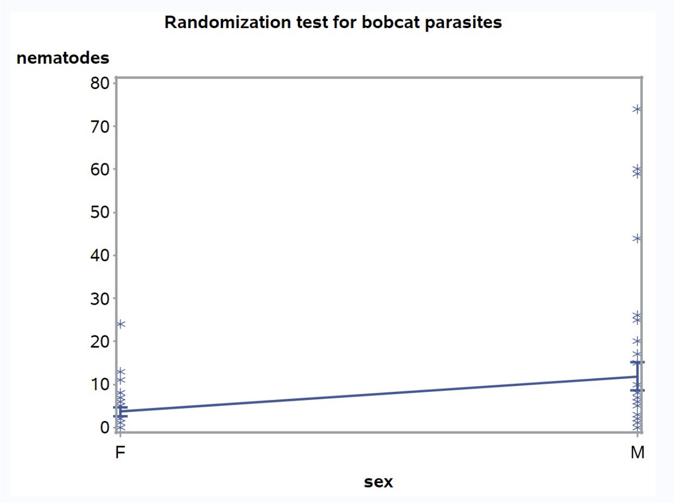
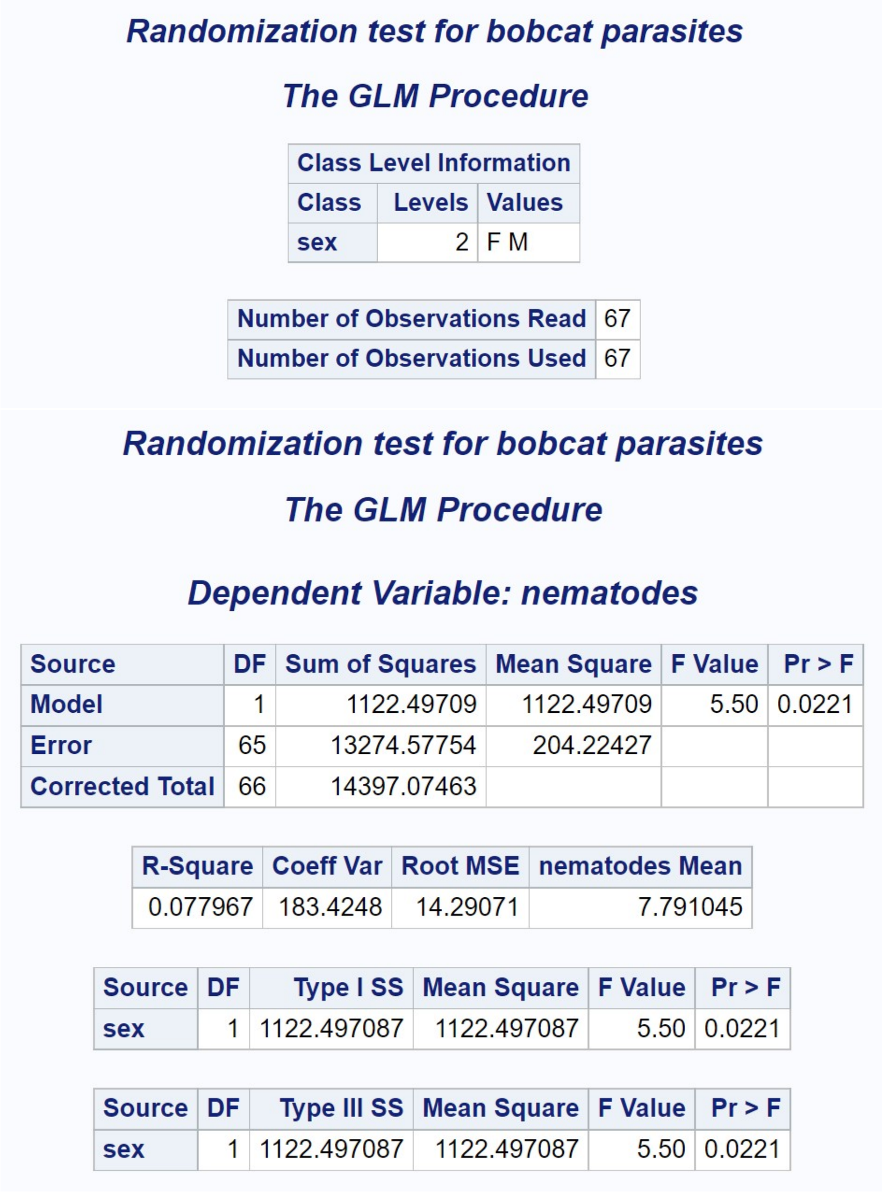


Figure 16.14: `Randtest_bobcat_parasites.sas - proc gplot`



Randomization test for bobcat parasites

The GLM Procedure

Class Level Information		
Class	Levels	Values
sex	2	F M

Number of Observations Read	67
Number of Observations Used	67

Randomization test for bobcat parasites

The GLM Procedure

Dependent Variable: nematodes

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1122.49709	1122.49709	5.50	0.0221
Error	65	13274.57754	204.22427		
Corrected Total	66	14397.07463			

R-Square	Coeff Var	Root MSE	nematodes Mean
0.077967	183.4248	14.29071	7.791045

Source	DF	Type I SS	Mean Square	F Value	Pr > F
sex	1	1122.497087	1122.497087	5.50	0.0221

Source	DF	Type III SS	Mean Square	F Value	Pr > F
sex	1	1122.497087	1122.497087	5.50	0.0221

Figure 16.15: `Randtest_bobcat_parasites.sas - proc glm`

SAS Log

```
Randomization test for Model F test where _source_='sex' and _type_='SS3'
has significance level of 0.0182
```

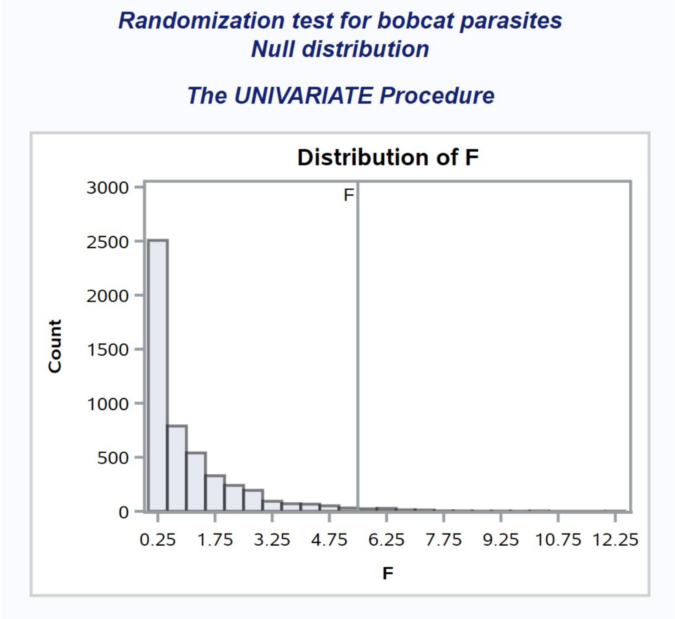


Figure 16.16: `Randtest_bobcat_parasites.sas - proc univariate`

16.5 Limitations of nonparametric tests

While nonparametric tests can be useful for non-normal data, they do have some drawbacks. One is that the number of designs that have nonparametric tests are fairly limited. We have seen nonparametric tests analogous to one-way ANOVA and two-sample t tests. There is also a rank test for randomized block designs called Friedman's test, as well as procedures for multiple comparisons (Hollander et al. 2014). Unfortunately, for more complex designs there are few available procedures.

Although nonparametric tests are not based on a particular distribution, they do make some assumptions. Consider the null and alternative hypotheses for the Wilcoxon test. The two groups are assumed to have the same cumulative distribution function, differing only by a shift parameter Δ. This implies the two groups have the same variance under both hypotheses, similar to parametric tests. When the variances are unequal as well as the sample sizes, both parametric and nonparametric tests may not be valid (Stewart-Oaten 1995). In particular, they may not have the correct Type I error rate.

Table 16.4 illustrates how unequal variances and sample sizes can affect the Type I error rate. It summarizes a simulation study comparing the validity of several different methods of comparing samples from two groups, including parametric and nonparametric methods. The first six columns give the theoretical mean, variance, and the sample sizes for the two groups. The simulated data were normally distributed with these parameters. Each data set was then analyzed using a two-sample t test, a Welch t test that implements a correction for unequal variances, the Wilcoxon test, and a randomization test. Any significant differences detected by these tests are Type I errors, because the two groups have the same mean. A total of 5000 simulated data sets were generated and analyzed. The proportion of simulated data sets showing significant results is an estimate of the Type I error rate (α) for each test. If the test is conducted using $\alpha = 0.05$, for example, we would expect this proportion of the simulations to be significant.

Regardless of differences in the variance between the two groups, when the sample sizes are equal all methods yielded a Type I error rate near the nominal $\alpha = 0.05$ level. When sample sizes are unequal, the t test, Wilcoxon test, and the randomization test all yielded Type I error rates higher or lower than $\alpha = 0.05$. Note that the pattern depends on which group (high or low variance) has the smaller sample size. Thus, the validity of these procedures depends on equal variances, especially when sample sizes are unequal across groups. This assumption needs to be carefully examined within applying both parametric and nonparametric tests.

The only valid test in this scenario was the Welch t test, which employs a correction for unequal variances. The correction alters the degrees of freedom for the test, based on the sample sizes and variances of the two groups (Stuart et al. 1999). It is conducted automatically by `proc ttest` in SAS, with the output labeled `Satterthwaite` (see Chapter 11). There is also a similar procedure for one-way designs called Welch ANOVA. It can conducted under `proc glm` using the `welch` option for the `means` statement.

Table 16.4: Effect of unequal variances and sample sizes on the estimated Type I error rate for common parametric and nonparametric tests, using $\alpha = 0.05$ for all tests. See text for further details.

μ_1	σ_1^2	n_1	μ_2	σ_2^2	n_2	t	Welch	Wilcoxon	Randomization
10	1	10	10	1	10	0.0474	0.0454	0.0422	0.0484
10	1	10	10	2	10	0.0516	0.0504	0.0514	0.0524
10	1	5	10	2	15	0.0208	0.0510	0.0236	0.0214
10	1	15	10	2	5	0.0956	0.0578	0.0662	0.0954
10	1	10	10	4	10	0.0510	0.0452	0.0464	0.0510
10	1	5	10	4	15	0.0104	0.0494	0.0170	0.0108
10	1	15	10	4	5	0.1588	0.0574	0.0836	0.1598

16.6 References

Cassell, D. L. (2002) A randomization-test wrapper for SAS PROCs. SUGI 27: Paper 251-27.

Conover, W. J. (1999) *Practical Nonparametric Statistics.* John Wiley & Sons, Inc., New York, NY.

Flores-Campaña, L. M., Arzola-González, J. F., & León-Herrera, R. (2012) Body size structure, biometric relationships and density of *Chiton albolineatus* (Mollusca: Polyplacophora) on the intertidal rocky zone of three islands of Mazatlan Bay, SE of the Gulf of California. *Revista de Biologia Marina y Oceanographía* 47: 203-211.

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972) Consequences of failure to meet assumptions underlying fixed effects analysis of variance and covariance. *Review of Educational Research* 42: 237-288.

Hinkelmann, K., & Kempthorne, O. (1994) *Design and Analysis of Experiments, Volume I: Introduction to Experimental Design.* John Wiley & Sons, Inc., New York, NY.

Hollander, M., Wolfe, D. A., & Chicken, E. (2014) *Nonparametric Statistical Methods, Third Edition.* John Wiley & Sons, Inc., Hoboken, NJ.

SAS Institute Inc. (2018) *SAS/STAT 15.1 Users Guide* SAS Institute Inc., Cary, NC

SAS Institute Inc. (2016a) *SAS/GRAPH 9.4: Reference, Fifth Edition.* SAS Institute Inc., Cary, NC.

SAS Institute Inc. (2016b) *SAS 9.4 Macro Language: Reference, Fifth Edition.* SAS Institute Inc., Cary, NC.

Stuart, A., Ord, J. K., & Arnold, S. (1999) *Kendall's Advanced Theory of Statistics, Volume 2A, Classical Inference and the Linear Model.* Oxford University Press Inc., New York, NY.

Manly, B. F. J. (1997) *Randomization, Bootstrap, and Monte Carlo Methods in Biology.* Chapman & Hall, New York, NY.

16.7 Problems

1. Using the Example 3 data, conduct a Wilcoxon test comparing parasite intensity in male vs. female bobcats. How do the results compare to the randomization test for these data in the text?

2. Data were also collected on the number of cestode parasites found in the bobcats from Example 3 (see below). Cestodes are another common type of gut parasite. Conduct a randomization test comparing the cestode intensity for male vs. female bobcats.

Sex	Cestodes	Sex	Cestodes	Sex	Cestodes	Sex	Cestodes
F	1	F	0	M	9	M	3
F	7	F	7	M	31	M	2
F	9	F	6	M	5	M	2
F	0	F	33	M	0	M	0
F	1	F	2	M	10	M	3
F	1	F	1	M	6	M	7
F	8	F	18	M	0	M	2
F	0	F	6	M	0	M	5
F	0	F	1	M	6	M	1
F	32	F	14	M	9	M	1
F	11	F	12	M	6	M	4
F	4	F	6	M	18	M	0
F	3	F	0	M	4	M	3
F	13			M	9	M	1
F	2			M	6		
F	2			M	5		
F	12			M	17		
F	4			M	4		
F	1			M	8		
F	3			M	11		

Chapter 17

Linear Regression

Linear regression is a statistical method for examining the relationship between two continuous variables, typically called Y and X. It assumes a linear relationship between the two variables, with a slope and intercept. One common purpose of linear regression is to establish whether changes in X cause changes in Y, by testing whether the slope of this line is significantly different from zero. Another purpose is prediction. Given a value of X, linear regression can be used to predict the value of Y and generate a confidence interval for this prediction. The variable X is sometimes under the control of the investigator, similar to a fixed effect in ANOVA, but can also be a random variable.

A basic assumption of linear regression is that X could be causing changes in Y, but not the reverse. For this reason, Y is often called the **dependent variable** while X is the **independent variable** in the analysis. The term **regressor** is also used for the independent variable in this context. For example, we might be interested in the effect of temperature on the growth rate of fish. Temperature might cause an increased growth rate, but clearly growth rate cannot influence temperature. This causal relationship is a distinguishing feature of regression as opposed to **correlation** analysis. Correlation is used to examine the **association** between two continuous variables and no causal direction is assumed (see Chapter 18). For example, we might be interested in the relationship between fish length and weight but there is no obvious causal relationship between the two variables.

Although linear regression assumes a different statistical model than ANOVA, there are a number of similarities in the estimation process and statistical tests for the two types. For example, both ANOVA and linear regression models use likelihood methods for parameter estimation and test construction, and employ F statistics to test various hypotheses. Both are examples of **general linear models**, in which the model parameters and error terms enter the model in an additive (linear) fashion.

What do the data look like for linear regression? As an example, we will use data from study on the southern pine beetle, *Dendroctonus frontalis* (Reeve et al. 1998). The study used cages to experimentally manipulate the density of *D. frontalis* attacking pine trees. The independent or X variable in the study was the number of beetles added to the cages, while the dependent or Y variable was the number of attacks the beetles made through the bark into the tree (Table 17.1). The notation Y_i and X_i refers to the values for the *ith* pair of numbers. For example, $Y_2 = 2.660$ and $X_2 = 1.000$. We will later see there is a positive relationship between the two variables, with attack density increasing as more beetles are added to the cages. Besides establishing the relationship between the two variables, there was also some interest in predicting the attack density as a function of the number of beetles added to the cage, for use in future studies. We will use the linear regression model to predict attack density for $X = 1.75$, a value not occurring in the data set.

Table 17.1: Example 1 - Observations from an experiment in which different numbers of the bark beetle *D. frontalis* were introduced into cages and the resulting attack density recorded (Reeve et al. 1998). Here Y is the attack density (attacks per 100 cm^2 of bark) while X is the number of beetles added ($\times 10^3$). Also shown are some preliminary calculations for the regression analysis.

i	Y_i	X_i	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})(X_i - \bar{X})$	$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$	$Y_i - \hat{Y}_i$	$(Y_i - \hat{Y}_i)^2$	$(\hat{Y}_i - \bar{Y})^2$	$(Y_i - \bar{Y})^2$
1	1.250	0.100	0.740	2.779	2.206	-0.956	0.914	5.176	10.440
2	2.660	1.000	0.002	-0.073	4.586	-1.926	3.711	0.011	3.316
3	7.330	2.000	1.081	2.962	7.231	0.099	0.010	7.563	8.116
4	1.600	1.250	0.084	-0.835	5.248	-3.648	13.305	0.588	8.301
5	2.620	0.500	0.212	0.856	3.264	-0.644	0.415	1.481	3.464
6	1.000	0.200	0.578	2.646	2.471	-1.471	2.162	4.042	12.118
7	4.340	1.500	0.291	-0.076	5.909	-1.569	2.461	2.038	0.020
8	5.230	0.750	0.044	-0.157	3.925	1.305	1.702	0.309	0.561
9	2.500	0.250	0.504	1.407	2.603	-0.103	0.011	3.528	3.925
10	3.250	0.500	0.212	0.567	3.264	-0.014	0.000	1.481	1.516
11	6.000	2.000	1.081	1.579	7.231	-1.231	1.516	7.563	2.307
12	4.750	1.500	0.291	0.145	5.909	-1.159	1.343	2.038	0.072
13	2.500	0.250	0.504	1.407	2.603	-0.103	0.011	3.528	3.925
14	8.750	2.000	1.081	4.439	7.231	1.519	2.307	7.563	18.223
15	6.000	1.000	0.002	0.060	4.586	1.414	1.998	0.011	2.307
16	5.000	0.500	0.212	-0.239	3.264	1.736	3.014	1.481	0.269
17	7.150	1.000	0.002	0.106	4.586	2.564	6.572	0.011	7.123
18	6.750	1.500	0.291	1.225	5.909	0.841	0.708	2.038	5.158
19	7.500	1.500	0.291	1.630	5.909	1.591	2.532	2.038	9.114
20	2.500	0.500	0.212	0.912	3.264	-0.764	0.584	1.481	3.925
21	5.000	2.000	1.081	0.540	7.231	-2.231	4.979	7.563	0.269
22	2.250	0.250	0.504	1.585	2.603	-0.353	0.124	3.528	4.978

i	Y_i	X_i	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})(X_i - \bar{X})$	$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$	$Y_i - \hat{Y}_i$	$(Y_i - \hat{Y}_i)^2$	$(\hat{Y}_i - \bar{Y})^2$	$(Y_i - \bar{Y})^2$
23	1.250	0.125	0.698	2.699	2.272	-1.022	1.045	4.879	10.440
24	4.750	1.000	0.002	0.011	4.586	0.164	0.027	0.011	0.072
25	4.500	0.250	0.504	-0.013	2.603	1.897	3.599	3.528	0.000
26	9.560	2.000	1.081	5.281	7.231	2.329	5.423	7.563	25.795
27	5.000	0.500	0.212	-0.239	3.264	1.736	3.014	1.481	0.269
$\sum$			11.798	31.203			63.486	82.528	146.014

17.1 Linear regression model

Suppose that we want to model the observations in studies like Example 1, where Y is observed for a number of X values. Let Y_i and X_i stand for the ith pair of values. The linear regression model takes the form

$$Y_i = \alpha + \beta X_i + \epsilon_i, \tag{17.1}$$

where α is the intercept and β the slope of a line, while $\epsilon_i \sim N(0, \sigma^2)$ (Searle 1971). Thus, the linear regression model represents the relationship between Y_i and X_i as a line on which random deviations due to natural variability (ϵ_i) are imposed. The slope β is also called the **regression coefficient**.

For the ith pair of values, we have $E[Y_i] = \alpha + \beta X_i$ and $Var[Y_i] = \sigma^2$ using the rules for expected values and variances. Thus, $Y_i \sim N(\alpha + \beta X_i, \sigma^2)$ for any X_i value. The behavior of the linear regression model can be illustrated by plotting this distribution across a range of X_i values. When β is positive, the mean of Y_i will increase as X_i increases (Fig. 17.1), while if β is negative the mean would decrease (not shown). The variance remains the same for all X_i. Note that the linear regression model has assumptions similar to the ANOVA models – the observations are assumed be normal and have the same variance.

The first objective in linear regression is to estimate the model parameters, especially the slope β, and then test whether it is different from zero. In particular, we will be interested in testing $H_0 : \beta = 0$. If a test of this hypothesis is significant this suggests a causal relationship (positive or negative) between Y and X. The alternative hypothesis can be written as $H_1 : \beta \neq 0$. It is also possible to test whether the intercept differs from zero although this is less common. We will discuss how these parameters are estimated and hypotheses tested in the next section.

17.2 Linear regression and likelihood

The maximum likelihood method can be used to estimate the parameters for regression models, similar to ANOVA models. Suppose we have n observations conforming to the linear regression model

$$Y_i = \alpha + \beta X_i + \epsilon_i. \tag{17.2}$$

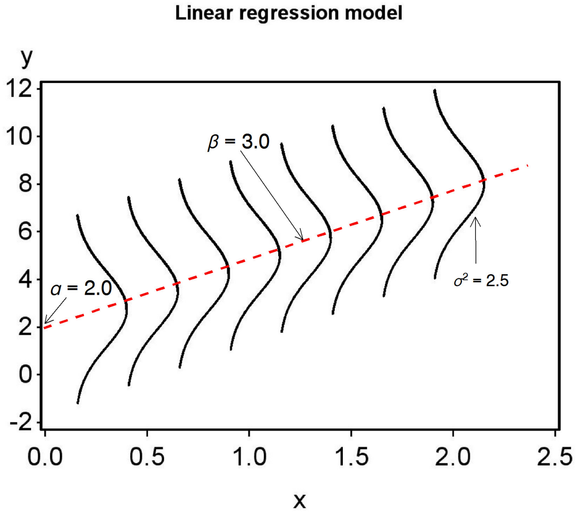


Figure 17.1: The linear regression model plotted across a range of X values, with $\alpha = 2.0$, $\beta = 3.0$, and $\sigma^2 = 2.5$.

This model has three parameters to estimate, namely α, β, and σ^2 (the variance of ϵ_i). What would the likelihood function be for these data? Consider the first observation in the *D. frontalis* cage experiment, for which $Y_1 = 1.250$ and $X_1 = 0.100$. For this observation, the model states that $Y_1 \sim N(\alpha + \beta X_1, \sigma^2)$, and so the likelihood would be

$$L_1 = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(Y_1-(\alpha+\beta X_1))^2}{\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(1.250-(\alpha+\beta 0.100))^2}{\sigma^2}} \tag{17.3}$$

The likelihood L_i for the ith observation would be similar, and the overall likelihood is defined as their product:

$$L(\alpha, \beta, \sigma^2) = L_1 \times L_2 \times \ldots \times L_n. \tag{17.4}$$

Finding the maximum likelihood estimates involves maximizing this quantity with respect to the parameters α, β, and σ^2. Using some calculus to find the maximum, it can be shown that estimators of these parameters are

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \tag{17.5}$$

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} \tag{17.6}$$

and

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(Y_i - (\hat{\alpha} + \hat{\beta}X_i))^2}{n-2} = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{n-2}. \tag{17.7}$$

Here $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$, the value of Y_i predicted by the model at X_i.

We can gain some insight into the estimation process by rearranging the likelihood function. It can be written in the form

$$L(\alpha, \beta, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2}\frac{\sum_{i=1}^{n}(Y_i - (\alpha+\beta X_i))^2}{\sigma^2}}. \tag{17.8}$$

Now examine the terms in the sum, which are of the form $(Y_i - (\alpha + \beta X_i))^2$. Values of α and β that minimize these terms will make the overall likelihood larger, because of the negative sign in the exponent. The likelihood will reach its maximum when this sum is smallest. Thus, values of α and β that minimize

$$\sum_{i=1}^{n}(Y_i - (\alpha + \beta X_i))^2 \tag{17.9}$$

are the maximum likelihood estimates. These estimates are also called **least squares** estimates because they minimize the sum of these squared terms. In fact, we could directly estimate α and β using this method without recourse to likelihood (Searle 1971). The two methods yield the same results when the data have a normal distribution.

A likelihood ratio test for linear regression can be constructed as follows. Suppose we want to test $H_0 : \beta = 0$ vs. $H_1 : \beta \neq 0$, the latter implying a linear relationship between Y and X. The statistical model under H_0 would be

$$Y_i = \alpha + \beta X_i + \epsilon_i \tag{17.10}$$

$$= \alpha + \epsilon_i \tag{17.11}$$

because $\beta = 0$ under H_0. The statistical model under H_1 would be the full model including a slope term, namely

$$Y_i = \alpha + \beta X_i + \epsilon_i. \tag{17.12}$$

We would need to find the maximum likelihood estimates under both H_1 (see previous section) and H_0, as well as L_{H_0} and L_{H_1}, the maximum height of the likelihood function under H_0 and H_1. We would then use the likelihood ratio test statistic

$$\lambda = \frac{L_{H_0}}{L_{H_1}}. \tag{17.13}$$

There is a one-to-one correspondence between $-2\ln(\lambda)$ and the statistic F_s used to test this null hypothesis (McCulloch & Searle 2001).

We can gain further insight into this test by defining various sum of squares and mean squares used to calculate F_s. In particular, we will define SS_{error}, $SS_{regression}$, and SS_{total} and their associated mean squares, which have functions similar to those in ANOVA. We will also summarize the calculations in an ANOVA table.

SS_{error} describes variation in the data around the regression line, or variation not explained by the model. It is defined as

$$SS_{error} = \sum_{i=1}^{n} \Big(Y_i - (\hat{\alpha} + \hat{\beta}X_i)\Big)^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2. \tag{17.14}$$

SS_{error} has $n - 2$ degrees of freedom. We can therefore define

$$MS_{error} = \frac{SS_{error}}{n-2} = \hat{\sigma}^2. \tag{17.15}$$

Thus, MS_{error} is equivalent to $\hat{\sigma}^2$, the maximum likelihood estimate of σ^2, the same relationship as found in ANOVA. SS_{error} and MS_{error} will be small if the data lie on a straight line and large if the data are scattered around the line.

$SS_{regression}$ describes variation in the data explained by the regression model. It is defined as

$$SS_{regression} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \tag{17.16}$$

and has one degree of freedom. We therefore have

$$MS_{regression} = \frac{SS_{regression}}{1} = SS_{regression}. \tag{17.17}$$

$SS_{regression}$ and $MS_{regression}$ will be large if the data have a strong positive or negative slope. To see this, recall that $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$. If the estimated slope

$\hat{\beta}$ is large, the values of $\hat{Y}_i$ will vary strongly as X_i changes and so generate a large sum of squares.

The total sum of squares is defined (as in ANOVA) to be

$$SS_{total} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \tag{17.18}$$

and has $n-1$ degrees of freedom. There is also a familiar relationship among the different sums of squares, namely

$$SS_{regression} + SS_{error} = SS_{total}. \tag{17.19}$$

The likelihood ratio statistic used to test $H_0 : \beta = 0$ is defined as

$$F_s = \frac{MS_{regression}}{MS_{error}}. \tag{17.20}$$

Under H_0, F_s has an F distribution with $df_1 = 1$ and $df_2 = n-2$ the degrees of freedom. Given the definitions of $MS_{regression}$ and MS_{error}, we can see that F_s tends to be large when the data have a strong slope (the numerator of this expression) relative to the amount of scatter in the data (the denominator).

We can organize the different sum of squares and mean squares into an ANOVA table for linear regression. It lists the different sources of variation in the data (regression, error, and total), their degrees of freedom, as well as the F test. Table 17.2 shows the general layout for linear regression.

Table 17.2: General ANOVA table for linear regression, showing formulas for different mean squares and the F test.

Source	*df*	Sum of squares	Mean square	F_s
Regression	1	$SS_{regression}$	$MS_{regression} = SS_{regression}/1$	$MS_{regression}/MS_{error}$
Error	$n-2$	SS_{error}	$MS_{error} = SS_{error}/(n-2)$	
Total	$n-1$	SS_{total}		

Table 17.3: ANOVA table for the Example 1 data set, including a P value for the test.

Source	*df*	Sum of squares	Mean square	F_s	P
Regression	1	82.528	82.528	32.504	< 0.001
Error	25	63.486	2.539		
Total	26	146.014			

17.2.1 Sample calculation - $\hat{\beta}$, $\hat{\alpha}$, and F test

We will illustrate the above calculations using the Example 1 data set, where Y is *D. frontalis* attack density and X is the number of beetles added to the cage. We are interested in estimating the slope and intercept (β and α) of the relationship between the two variables, and then testing whether the slope is significantly different from zero ($H_0 : \beta = 0$).

The first step is to calculate the sample mean for both Y and X, and we obtain $\bar{Y} = 4.481$ and $\bar{X} = 0.960$. We then calculate $(X_i - \bar{X})^2$ for each value of X_i (see Table 17.1) and sum these values to obtain

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 = 11.798. \tag{17.21}$$

We then calculate $(Y_i - \bar{Y})(X_i - \bar{X})$ for each pair of numbers and sum these to obtain

$$\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X}) = 31.203. \tag{17.22}$$

The estimate of β can then be calculated, and we find

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{31.203}{11.798} = 2.645. \tag{17.23}$$

We can then estimate α using the formula

$$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X} = 4.481 - 2.645(0.960) = 1.942. \tag{17.24}$$

The next step is to calculate the predicted values of Y_i using the formula $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$, for each value of X_i (see Table 17.1). We then calculate $Y_i - \hat{Y}_i$ in another column - these are the residuals for each observation. Squaring and summing the residuals, we find

$$SS_{error} = \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = 63.486, \tag{17.25}$$

and

$$MS_{error} = \frac{SS_{error}}{n-2} = \frac{63.486}{27-2} = 2.539. \tag{17.26}$$

We next calculate a column consisting of $(\hat{Y}_i - \bar{Y})^2$ for each observation, then sum these values to obtain

$$SS_{regression} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 = 82.528, \tag{17.27}$$

and so

$$MS_{regression} = SS_{regression}/1 = 82.528. \tag{17.28}$$

We are now in a position to calculate F_s, the statistic used to test $H_0 : \beta = 0$. We have

$$F_s = \frac{MS_{regression}}{MS_{error}} = \frac{82.528}{2.539} = 32.504. \tag{17.29}$$

Under H_0, F_s has an F distribution with $df_1 = 1$ and $df_2 = 27 - 2 = 25$ degrees of freedom. Using Table F, we find the $P < 0.001$. There was a highly significant effect of beetles numbers on the attack density of *D. frontalis* ($F_{1,25} = 32.504, P < 0.001$).

The last column in Table 17.1 calculates $(Y_i - \bar{Y})^2$, the components of SS_{total}. Summing these components we obtain $SS_{total} = 146.014$. It can also be calculated using the formula $SS_{regression} + SS_{error} = SS_{total}$. Table 17.3 shows the completed ANOVA table.

The observations for Example 1 and the fitted linear regression model are shown in Fig. 17.2. The estimation procedure (maximum likelihood or least squares) finds values of α and β that minimize the sum of the squared differences between the data points and the line. In particular, it minimizes the sum of the squared residuals, where the residuals are $Y_i - \hat{Y}_i = Y_i - (\hat{\alpha} + \hat{\beta} X_i)$.

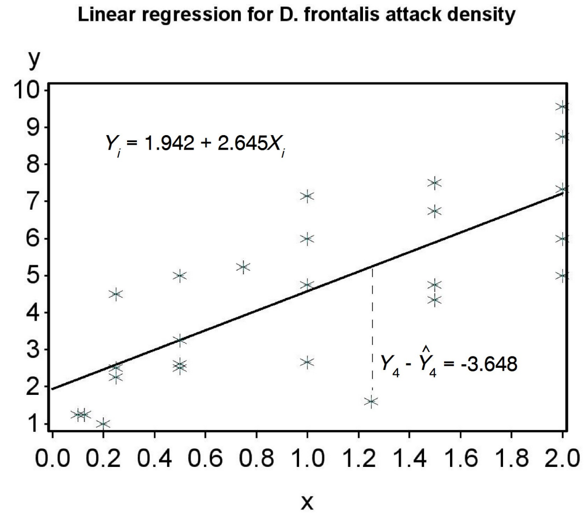


Figure 17.2: Linear regression model fitted to the Example 1 data, where Y is attack density and X is beetles added to the cages. The vertical dashed line shows the residual $Y_4 - \hat{Y}_4 = -3.648$ for the $i = 4$ observation.

17.3 Confidence and prediction intervals

In this section, we will derive confidence intervals for the parameters of the regression model (α and β) that provide a measure of their precision (see Chapter 9). We will also find confidence intervals for the mean value of Y_i at a given value of X_i. Another type of interval for linear regression are **prediction intervals**. These are used to set limits for future Y_i values given some value of X_i. Both of these intervals are used in prediction, another common purpose for linear regression. See Draper & Smith (1981) for further details.

The confidence interval for the slope β is based on $\hat{\beta}$, the maximum likelihood estimate of β, and the standard error of this estimate $s_{\hat{\beta}}$, given by the formula

$$s_{\hat{\beta}} = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}}, \tag{17.30}$$

where $\hat{\sigma}^2 = MS_{error}$. Note that $s_{\hat{\beta}}$ depends on the scatter of the data around the line ($\hat{\sigma}^2$) as well as the amount of variability in X_i. **A study using a larger range of X_i values will thus provide a more precise estimate of β, because it reduces $s_{\hat{\beta}}$. Increasing the sample size n would also increase the precision, by increasing the sum of squares in the denominator for $s_{\hat{\beta}}$.**

It can be shown that the quantity

$$\frac{\hat{\beta} - \beta}{s_{\hat{\beta}}} \tag{17.31}$$

has a t distribution with $n-2$ degrees of freedom, the same as for MS_{error}. This fact can be used to derive a confidence interval for β. Using Table T, we first find a value of $c_{\alpha,n-2}$ for $n-2$ degrees of freedom such that the following equation is true:

$$P\left[-c_{\alpha,n-2} < \frac{\hat{\beta} - \beta}{s_{\hat{\beta}}} < c_{\alpha,n-2}\right] = 1 - \alpha. \tag{17.32}$$

Rearranging this equation we obtain

$$P\left[\hat{\beta} - c_{\alpha,n-2}s_{\hat{\beta}} < \beta < \hat{\beta} + c_{\alpha,n-2}s_{\hat{\beta}}\right] = 1 - \alpha. \tag{17.33}$$

It follows that the interval

$$(\hat{\beta} - c_{\alpha,n-2}s_{\hat{\beta}}, \hat{\beta} + c_{\alpha,n-2}s_{\hat{\beta}}) \tag{17.34}$$

is a $100(1-\alpha)\%$ confidence interval for β. The center of the confidence interval would be $\hat{\beta}$.

We may also want to test various null hypotheses concerning β. For example, we may want to test $H_0 : \beta = \beta_0$ vs. $H_1 : \beta \neq \beta_0$, where β_0 takes some value of interest. Similar to the approach in Chapter 10, we would use the test statistic

$$T_s = \frac{\hat{\beta} - \beta_0}{s_{\hat{\beta}}}. \tag{17.35}$$

Under H_0, T_s has a t distribution with $n-2$ degrees of freedom, and we would reject H_0 for sufficiently large values of this statistic. For $\beta_0 = 0$, this test is equivalent to the F test we developed earlier for $H_0 : \beta = 0$, and in fact $T_s^2 = F_s$. The t test is more general, however, because we can also test $H_0 : \beta = \beta_0$ for any value of β_0.

It is possible to derive similar t tests and confidence intervals for the intercept parameter α. The t test is most commonly used to test $H_0 : \alpha = 0$. If the test is significant this implies an intercept different from zero. We will let SAS handle the calculations here.

We can also derive a confidence interval for the theoretical mean of Y_i at a given X_i value. Recall that according to the linear regression model, $E[Y_i] = \alpha + \beta X_i$. Thus, Y_i has a mean of $\mu_i = \alpha + \beta X_i$ for any X_i value. The confidence interval is based on $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$, the predicted value of Y_i at X_i. It also depends on the standard error $s_{\hat{Y}}$ of $\hat{Y}$, which is given by the formula

$$s_{\hat{Y}} = \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right]}. \tag{17.36}$$

Note that the standard error $s_{\hat{Y}}$ depends on the value of $(X_i - \bar{X})^2$, which is the squared distance of X_i from $\bar{X}$. The farther X_i is from $\bar{X}$, the larger the value of $s_{\hat{Y}}$.

Using methods similar to the confidence interval for β, it can be shown that a $100(1-\alpha)$ confidence interval for $\mu_i = \alpha + \beta X_i$ has the form

$$(\hat{Y}_i - c_{\alpha,n-2}s_{\hat{Y}}, \hat{Y}_i + c_{\alpha,n-2}s_{\hat{Y}}). \tag{17.37}$$

The interval will be broader for values of X_i far from $\bar{X}$ because $s_{\hat{Y}}$ will be larger. In other words, the precision of the confidence interval decreases with the distance from $\bar{X}$.

We next examine **prediction intervals**. Here, we are trying to find an interval that contains a defined percentage of future Y_i values for a given value of X_i. These are similar in form to the intervals for the theoretical mean $\mu_i = \alpha + \beta X_i$, but are always wider because you are trying to enclose a single future observation rather than a mean value.

The prediction interval is based on $\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$, the predicted value of Y_i at X_i, and the standard error $s_{\hat{Y}(1)}$ of $\hat{Y}_i$, which is given by the formula

$$s_{\hat{Y}(1)} = \sqrt{\hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2} \right]}. \tag{17.38}$$

Note the additional term (1+) within the square brackets, which makes this standard error larger than $s_{\hat{Y}}$. It also depends on the value of $(X_i - \bar{X})^2$, and so the farther X_i is from $\bar{X}$, the larger the value of $s_{\hat{Y}(1)}$. It can be shown that a $100(1-\alpha)$ prediction interval for a single future Y_i has the form

$$(\hat{Y}_i - c_{\alpha,n-2} s_{\hat{Y}(1)}, \hat{Y}_i + c_{\alpha,n-2} s_{\hat{Y}(1)}). \tag{17.39}$$

17.3.1 Sample calculation - confidence and prediction intervals

We now illustrate the calculations for confidence intervals using the Example 1 data. We earlier found that $\hat{\beta} = 2.645$ and $\hat{\alpha} = 1.942$. To find a confidence interval for β, we first need to calculate $s_{\hat{\beta}}$. From Table 17.1, we see that $\sum_{i=1}^{n} (X_i - \bar{X})^2 = 11.798$, and we earlier calculated that $\hat{\sigma}^2 = MS_{error} = 2.539$. Inserting these quantities into the formula for $s_{\hat{\beta}}$, we find

$$s_{\hat{\beta}} = \sqrt{\frac{\hat{\sigma}^2}{\sum_{i=1}^{n} (X_i - \bar{X})^2}} = \sqrt{\frac{2.539}{11.798}} = 0.464. \tag{17.40}$$

A 95% confidence interval for β has the form

$$(\hat{\beta} - c_{0.05,n-2} s_{\hat{\beta}}, \hat{\beta} + c_{0.05,n-2} s_{\hat{\beta}}) \tag{17.41}$$

From Table T with $df = n - 2 = 27 - 2 = 25$, we find that $c_{0.05,25} = 2.060$. Inserting this value, $\hat{\beta} = 2.645$, and $s_{\hat{\beta}} = 0.464$ in the above formula, we obtain

$$(2.645 - 2.060(0.464), 2.645 + 2.060(0.464)) \tag{17.42}$$

or

$$(1.689, 3.601). \tag{17.43}$$

We next find a confidence interval for the theoretical mean $\mu_i = \alpha + \beta X_i$ at $X_i = 1.75$. We first need to find the predicted value $\hat{Y}_i$ for this value of X_i, using the estimated intercept and slope. We have

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i = 1.942 + 2.645(1.75) = 6.571. \tag{17.44}$$

The standard error $s_{\hat{Y}}$ for this interval also uses $\sum_{i=1}^{n}(X_i - \bar{X})^2 = 11.798$ and $\hat{\sigma}^2 = 2.539$, and we earlier found that $\bar{X} = 0.960$. Inserting these quantities into the formula for $s_{\hat{Y}}$, we find that

$$s_{\hat{Y}} = \sqrt{\hat{\sigma}^2 \left[\frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]} \tag{17.45}$$

$$= \sqrt{2.539 \left[\frac{1}{27} + \frac{(1.75 - 0.960)^2}{11.798}\right]} \tag{17.46}$$

$$= \sqrt{2.539 \left[0.037 + \frac{0.624}{11.798}\right]} \tag{17.47}$$

$$= 0.478. \tag{17.48}$$

A 95% confidence interval for the theoretical mean $\mu_i = \alpha + \beta X_i$ has the form

$$(\hat{Y} - c_{0.05,n-2} s_{\hat{Y}}, \hat{Y} + c_{0.05,n-2} s_{\hat{Y}}) \tag{17.49}$$

Inserting $\hat{Y} = 6.571$, $s_{\hat{Y}} = 0.478$, and $c_{0.05,25} = 2.060$ in the above formula, we find

$$(6.571 - 2.060(0.478), 6.571 + 2.060(0.478)) \tag{17.50}$$

or

$$(5.586, 7.556). \tag{17.51}$$

Lastly, we calculate a prediction interval for a single future observation Y_i at $X_i = 1.75$. We earlier calculated that $\hat{Y}_i = 6.571$ for this value of X_i, and will again make use of $\sum_{i=1}^{n}(X_i - \bar{X})^2 = 11.798$, $\bar{X} = 0.960$ and $\hat{\sigma}^2 = 2.539$. Inserting these quantities into the formula for $s_{\hat{Y}(1)}$, we obtain

$$s_{\hat{Y}(1)} = \sqrt{\hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}\right]} \tag{17.52}$$

$$= \sqrt{2.539 \left[1 + \frac{1}{27} + \frac{(1.75 - 0.960)^2}{11.798}\right]} \tag{17.53}$$

$$= \sqrt{2.539 \left[1 + 0.037 + \frac{0.624}{11.798}\right]} \tag{17.54}$$

$$= 1.663. \tag{17.55}$$

A 95% prediction interval for a single Y_i has the form

$$(\hat{Y} - c_{0.05,n-2}s_{\hat{Y}(1)}, \hat{Y} + c_{0.05,n-2}s_{\hat{Y}(1)}) \tag{17.56}$$

Inserting $\hat{Y} = 6.571$, $s_{\hat{Y}(1)} = 1.663$, and $c_{0.05,25} = 2.060$ in this formula, we obtain

$$(6.571 - 2.060(1.663), 6.571 + 2.060(1.663)) \tag{17.57}$$

or

$$(3.145, 9.997). \tag{17.58}$$

Note this interval is much wider than the interval for the theoretical mean $\mu_i = \alpha + \beta X_i$, which was $(5.586, 7.556)$. This is because you are trying to enclose a single future observation, a random variable Y_i, rather than a theoretical mean.

17.4 R^2 values

R^2 values are a measure of how well a statistical model explains the data. Recall that the following relationship holds among the sum of squares in linear regression:

$$SS_{regression} + SS_{error} = SS_{total}. \tag{17.59}$$

We can think of the different sum of squares as partitioning the variability in the data into different sources. $SS_{regression}$ represents variability explained by the regression line, SS_{error} represents variability of the observations around the regression line, while SS_{total} is the total amount of variability in the data. The R^2 value for a linear regression model is the proportion of total variability explained by the model, or

$$R^2 = \frac{SS_{regression}}{SS_{total}} = \frac{SS_{regression}}{SS_{regression} + SS_{error}}. \tag{17.60}$$

It is clear from this formula that R^2 must range between 0 and 1 ($0 \leq R^2 \leq 1$). For the Example 1 data, we have

$$R^2 = 82.528/146.014 = 0.565. \tag{17.61}$$

Thus, 56.5% of the variation is explained by the regression model for these data. Small R^2 values indicate there is substantial variability in the data not explained by the model, while large ones indicate the model explains most of the variation.

More generally, we can define an R^2 value for both ANOVA and regression models as

$$R^2 = \frac{SS_{model}}{SS_{total}} = \frac{SS_{model}}{SS_{model} + SS_{error}}. \tag{17.62}$$

For example, we have $SS_{model} = SS_{among}$ for one-way ANOVA while $SS_{error} = SS_{within}$. The R^2 value here is the proportion of the variation explained by the one-way ANOVA model, in particular the variation among the group means. The SAS output for `proc glm` provides an R^2 for ANOVA models of this form.

17.5 Linear regression for Example 1 - SAS demo

The linear regression analysis can be conducted using `proc glm` and a program similar in structure to ANOVA ones (see SAS program below). We first input the observations using a `data` step, with the first variable standing for attack density (`attacks`) while the second is the number of beetles added (`beetles`). The next two lines in the `data` step define which of these two variables are the dependent and independent ones. The line `y = attacks` sets attack density as the dependent variable, while `x = beetles` is the independent variable. The remainder of the program then uses `y` and `x` rather than the original variables and so does not need to be changed for other data sets. Transformations of the observations could also be applied at this point.

Note the additional observation at end of the data set, for which beetles is 1.75 but attacks is a missing value. The purpose is to make proc glm calculate confidence and prediction intervals for attack density for that particular number of beetles.

The data are then plotted along with the fitted line plus confidence and prediction intervals. This accomplished using the following proc gplot code (SAS Institute Inc. 2016). The three y*x statements in the plot command plot the same data in three different ways, which are then combined into one graph using the overlay option. The first plot, using the symbol1 command, draws the data points. The second plot, using the symbol2 command, draws a regression line through the points and also plots 95% confidence intervals for the mean of Y_i at X_i, or $\mu_i = \alpha + \beta X_i$, across the range of X_i values. The third plot, using the symbol3 command, plots 95% prediction intervals for a single future observation, again across the range of X_i values. A similar plot is also generated by proc glm for linear regression models (see Fig. 17.9).

The regression analysis is conducted using proc glm as shown below (SAS Institute Inc. 2018). There is no class statement because the independent variable x is a continuous variable and does not fall into discrete groups as with ANOVA. Note the similarity of the model statement to the linear regression model. The option clparm is used to generate 95% confidence intervals for α and β, while clm generates a 95% confidence interval for the mean of Y_i at each value of X_i. If we want prediction intervals it is necessary to run proc glm a second time using the cli option in the model statement (see below). This is necessary because proc glm cannot generate both types of intervals at the same time.

The data points, regression line, and confidence or prediction intervals are shown in Fig. 17.4. The prediction intervals are much wider than the confidence intervals, because the prediction intervals are for single future Y_i while the confidence intervals enclose a mean. Note that both types of interval increase in width as you move away from the center of the X values. This follows from the fact that the standard errors involved in these calculations are a function of $(X_i - \bar{X})^2$, which increases as X_i moves away from $\bar{X}$.

Examining the output for proc glm, first note that the slope β is labeled as x while the intercept α is Intercept (Fig. 17.5). We see that attack density y increases with beetle numbers x, because $\hat{\beta} = 2.645$ and is positive. The effect of beetle numbers on attack density was highly significant ($F_{1,25} =$

$32.5, P < 0.0001$). There are several F tests to chose from in the output, but all give the same result for simple linear regression. Alternately, we could report the t test for β ($t_{25} = 5.70, P < 0.0001$), which also tests $H_0 : \beta = 0$. We see that $R^2 = 0.565$, indicating that 56.5% of the variation is explained by the regression model.

The `proc glm` output also provides 95% confidence intervals for α and β. A 95% confidence interval for the mean of Y_i at each X_i value is also given (Fig. 17.6), as well as 95% prediction intervals for a single future Y_i (Fig. 17.7). These intervals were also calculated for $X_i = 1.75$ and match our earlier results.

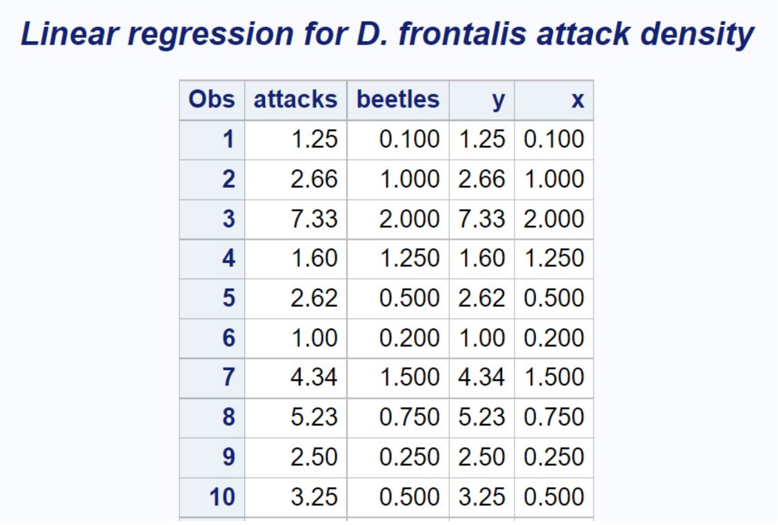
Note that the estimated intercept is some distance from zero ($\hat{\alpha} = 1.942$), and in fact the t test of $H_0 : \alpha = 0$ reported by SAS was highly significant ($t_{25} = 3.59, P = 0.0014$). This cannot really be true because the addition of zero beetles should give you an attack density of zero. A more resonable (and possibly non-linear) model would require that the intercept be zero. This is a potential pitfall when using linear regression. Many biological phenomenon are approximately linear over some range of the data but the approximation breaks down for more extreme values. A linear regression does not take this possibility into account and so cannot provide a general explanation of some phenomena.

SAS Program

```
* SPBattack2.sas;
title 'Linear regression for D. frontalis attack density';
data frontalis;
    input attacks beetles;
    * Apply transformations here;
    y = attacks;
    x = beetles;
    datalines;
1.25 0.100
2.66 1.000
7.33 2.000
1.60 1.250
2.62 0.500

etc.

5.00 0.500
.    1.750
;
run;
* Print data set;
proc print data=frontalis;
run;
* Plot data and regression line;
proc gplot data=frontalis;
    plot y*x y*x y*x / overlay vaxis=axis1 haxis=axis1;
    symbol1 i=none v=star c=black height=2 width=3;
    symbol2 i=rlclm v=none c=red height=2 width=3;
    symbol3 i=rlcli v=none c=blue height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* Regression analysis with confidence intervals;
proc glm plots=diagnostics data=frontalis;
    model y = x / clparm clm;
run;
* Regression analysis with prediction intervals;
proc glm data=frontalis;
    model y = x / clparm cli;
run;
quit;
```



Linear regression for D. frontalis attack density

Obs	attacks	beetles	y	x
1	1.25	0.100	1.25	0.100
2	2.66	1.000	2.66	1.000
3	7.33	2.000	7.33	2.000
4	1.60	1.250	1.60	1.250
5	2.62	0.500	2.62	0.500
6	1.00	0.200	1.00	0.200
7	4.34	1.500	4.34	1.500
8	5.23	0.750	5.23	0.750
9	2.50	0.250	2.50	0.250
10	3.25	0.500	3.25	0.500

etc.

Figure 17.3: `SPBattack.sas - proc print`

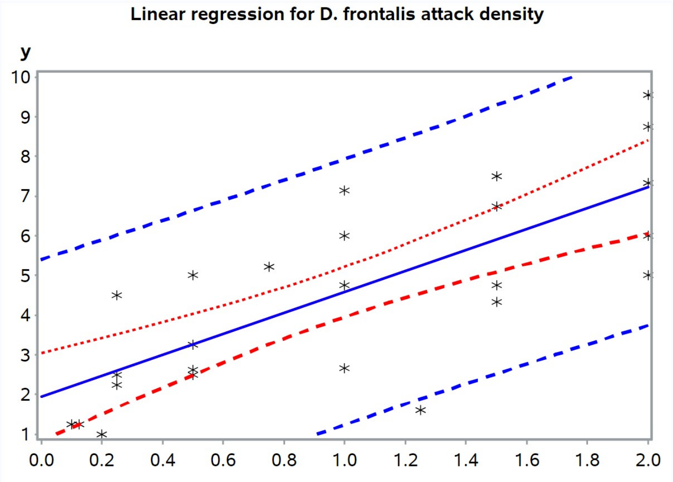
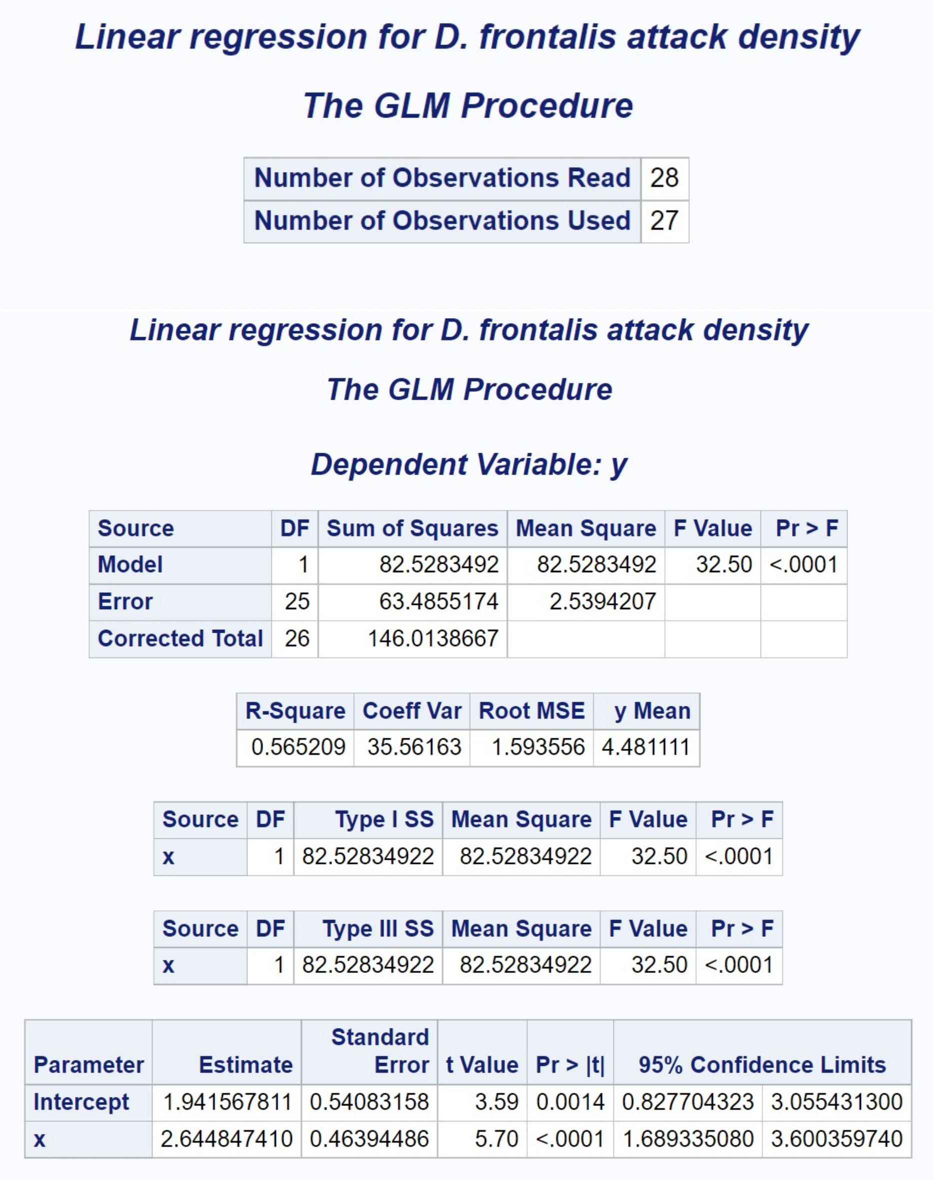


Figure 17.4: `SPBattack.sas - proc gplot`



Linear regression for D. frontalis attack density

The GLM Procedure

Number of Observations Read	28
Number of Observations Used	27

Linear regression for D. frontalis attack density

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	82.5283492	82.5283492	32.50	<.0001
Error	25	63.4855174	2.5394207		
Corrected Total	26	146.0138667			

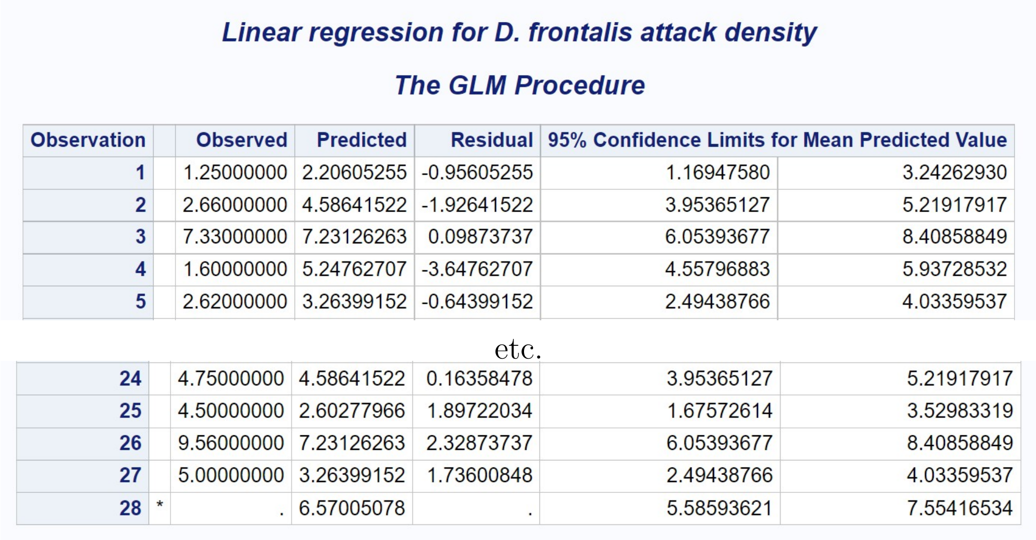
R-Square	Coeff Var	Root MSE	y Mean
0.565209	35.56163	1.593556	4.481111

Source	DF	Type I SS	Mean Square	F Value	Pr > F
x	1	82.52834922	82.52834922	32.50	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
x	1	82.52834922	82.52834922	32.50	<.0001

		Standard				
Parameter	Estimate	Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1.941567811	0.54083158	3.59	0.0014	0.827704323	3.055431300
x	2.644847410	0.46394486	5.70	<.0001	1.689335080	3.600359740

Figure 17.5: `SPBattack.sas - proc glm`



Linear regression for D. frontalis attack density

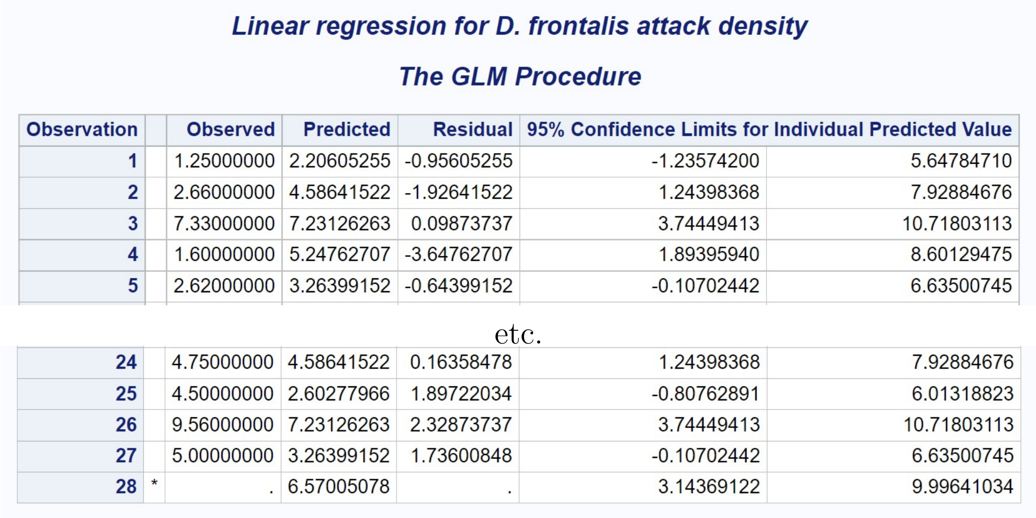
The GLM Procedure

Observation		Observed	Predicted	Residual	95% Confidence Limits for Mean Predicted Value	
1		1.25000000	2.20605255	-0.95605255	1.16947580	3.24262930
2		2.66000000	4.58641522	-1.92641522	3.95365127	5.21917917
3		7.33000000	7.23126263	0.09873737	6.05393677	8.40858849
4		1.60000000	5.24762707	-3.64762707	4.55796883	5.93728532
5		2.62000000	3.26399152	-0.64399152	2.49438766	4.03359537

etc.

24		4.75000000	4.58641522	0.16358478	3.95365127	5.21917917
25		4.50000000	2.60277966	1.89722034	1.67572614	3.52983319
26		9.56000000	7.23126263	2.32873737	6.05393677	8.40858849
27		5.00000000	3.26399152	1.73600848	2.49438766	4.03359537
28	*	.	6.57005078	.	5.58593621	7.55416534

Figure 17.6: `SPBattack.sas - proc glm`



Linear regression for D. frontalis attack density

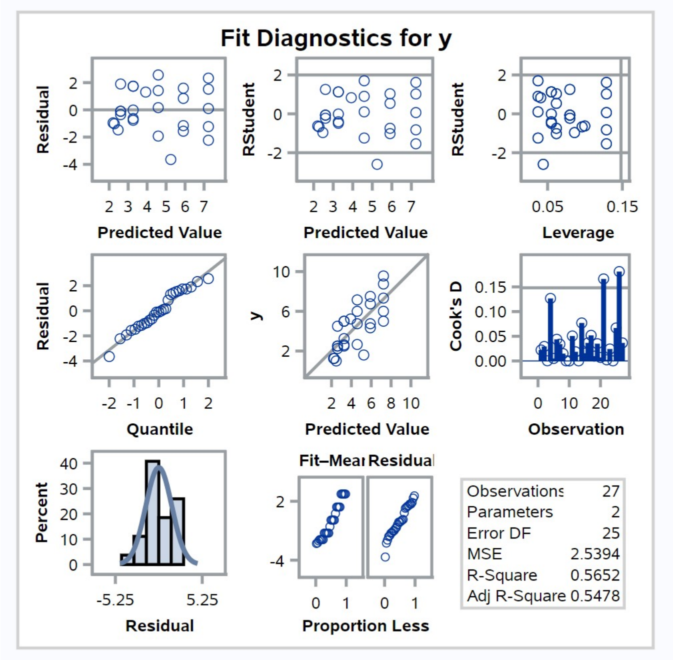
The GLM Procedure

Observation		Observed	Predicted	Residual	95% Confidence Limits for Individual Predicted Value	
1		1.25000000	2.20605255	-0.95605255	-1.23574200	5.64784710
2		2.66000000	4.58641522	-1.92641522	1.24398368	7.92884676
3		7.33000000	7.23126263	0.09873737	3.74449413	10.71803113
4		1.60000000	5.24762707	-3.64762707	1.89395940	8.60129475
5		2.62000000	3.26399152	-0.64399152	-0.10702442	6.63500745

etc.

24		4.75000000	4.58641522	0.16358478	1.24398368	7.92884676
25		4.50000000	2.60277966	1.89722034	-0.80762891	6.01318823
26		9.56000000	7.23126263	2.32873737	3.74449413	10.71803113
27		5.00000000	3.26399152	1.73600848	-0.10702442	6.63500745
28	*	.	6.57005078	.	3.14369122	9.99641034

Figure 17.7: `SPBattack.sas - proc glm`



```
Observations      27
Parameters         2
Error DF          25
MSE           2.5394
R-Square      0.5652
Adj R-Square  0.5478
```

Figure 17.8: `SPBattack.sas - proc glm`

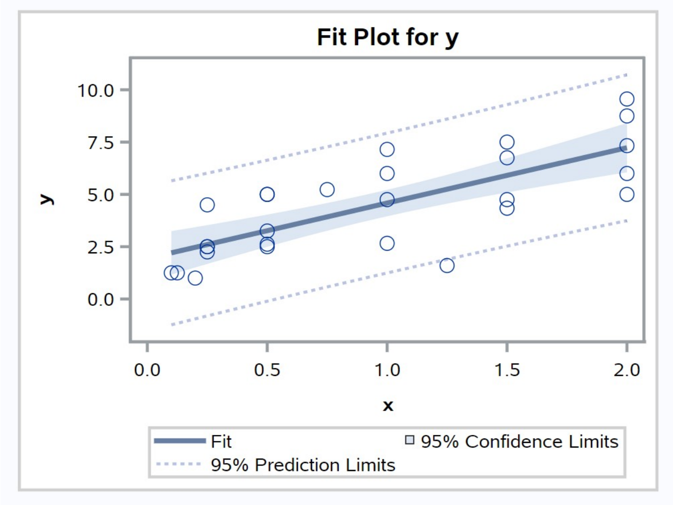


Figure 17.9: `SPBattack.sas - proc glm`

17.6 Assumptions and transformations

Linear regression makes the same assumptions as ANOVA, including homogeneity of variances and normality, and the same types of plots can be used to assess them. If the homogeneity of variances assumption is satisfied, the points in a residual vs. predicted plot should be equally scattered across the range of predicted values. Outliers can also be identified using this plot. The normality assumption can be evaluated using a normal quantile plot of the residuals, with a straight diagonal line indicating this assumption is satisfied.

Examining the residuals from the Example 1 analysis, we see no obvious pattern in the residual vs. predicted plot, suggesting the homogeneity of variances assumption is satisfied (Fig. 17.8). No outliers were present. The normal quantile plot suggests the normality assumption is satisfied.

Linear regression makes another key assumption, namely that the relationship between Y and X is linear. This assumption can be checked by examining a plot of Y vs. X. What can be done if the relationship seems nonlinear? We can sometimes fix this problem by applying a transformation to Y, X, or both Y and X, so that linear regression can be applied to the transformed data. **This use of transformations greatly extends the utility of linear regression.** Some commonly used transformations are $\log Y$ vs. X, $\log Y$ vs. $\log X$, Y vs. $\log X$, and $1/Y$ vs. X. A transformation that linearizes the data sometimes corrects for problems with the homogeneity of variances and normality assumptions.

A transformation may be selected based on prior information about the data and system. For example, a conservation biologist may be interested in the relationship between island area A and the number of species S on the island, and previous studies suggest that this relationship will be linear on a log scale (MacArthur & Wilson 1967). Another approach is to try a number of transformations and chose the one that makes the data most linear. We will illustrate each approach with an example below.

In cases where no transformation can linearize the data, another possibility would be **nonlinear regression** (Juliano 1993). This type of analysis requires that the user specify a model $Y = f(X, \theta_1, \theta_2, \ldots) + \epsilon$ for the data, where f is a function with parameters $\theta_1, \theta_2, \ldots$ to be estimated. SAS implements this type of nonlinear regression in `proc nlin`, while `proc nlmixed` allows for nonlinear functions as well as random effects and nonnormal distributions.

17.6.1 Species-area data - SAS demo

For many organisms there is a relationship between a defined area of habitat, such as an island, and the number of species found there. If S is the number of species, and A the area of habitat, then the model $S = cA^z$ seems to describe many data sets (MacArthur & Wilson 1967). Applying the $\log_{10}$ function to both sides of this equation, we obtain

$$\log_{10} S = \log_{10} c + z \log_{10} A. \tag{17.63}$$

This form of the model is linear and suggests linear regression could be used to analyze species-area data. The SAS program listed below shows how these transformations can be applied to the bird fauna on archipelagos and islands of varying areas. The data are the number of species vs. island area (square miles) for 23 islands. The data were simulated to resemble Fig. 9 in MacArthur & Wilson (1967). An extra observation is included with a missing value for the number of species, but an island area of 5000 square miles, to make `proc glm` calculate a confidence interval for the mean of this island.

We first conduct the analysis without any transformation, with the line `y = species` defining species as the dependent variable while `x = area` is the independent one. Examining the `proc gplot` graph, note the nonlinear nature of the relationship between the number of species and island area (Fig. 17.10). The picture improves after a $\log_{10}$ transformation is applied to both variables (Fig. 17.11).

Now that the linearity assumption is satisfied, we can interpret the rest of the SAS output (Fig. 17.13). We see that the number of species increased with island area ($\hat{\beta} = 0.241$) and the effect was highly significant ($F_{1,21} = 148.16, P < 0.0001$). In terms of the original model, where $S = cA^z$, we see that $\hat{\beta} = 0.241$ is also an estimate of z. The R^2 value is 0.876, indicating that 87.6% of the variation is explained by the regression model. Confidence intervals are also provided for the intercept and slope.

The `proc glm` output also includes a predicted value $\hat{Y}_i = 1.800$ at $X_i = 3.699$, which corresponds to an island area of 5000 (see Fig. 17.14). We need to convert this predicted value to the original scale of measurement using antilogs. We have $\hat{S}_i = 10^{\hat{Y}_i} = 10^{1.800} = 63.10$ species. So, we predict there would be 63 species on an island of 5000 square miles. The confidence interval for the mean is $(1.746, 1.855)$, which we can similarly convert to $(10^{1.745}, 10^{1.855})$ or $(55.72, 71.61)$.

Examining the residual plots from this analysis, it appears the homogeneity of variances and normality assumptions were satisfied (Fig. 17.15).

SAS Program

```
* SAprob2.sas;
title 'Linear regression for species-area data';
data sa;
    input species area;
    * Apply transformations here;
    y = log10(species);
    x = log10(area);
    datalines;
 15      28
104  113480
165  380358
116   33252
 35    1010
 33     305
 78   37620
 93    4762
 50     213
 76    2976
 18      23
 28     186
 20     423
121  108512
 53     364
 22     269
102   11163
 28     487
158  445409
 19      70
111   38309
152  100873
 55    1354
  .    5000
;
run;
* Print data set;
proc print data=sa;
run;
* Plot data and regression line;
proc gplot data=sa;
    plot y*x=1 y*x=2 y*x=3 / overlay vaxis=axis1 haxis=axis1;
    symbol1 i=none v=star c=black height=2 width=3;
    symbol2 i=rlclm v=none c=red height=2 width=3;
    symbol3 i=rlcli v=none c=blue height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* Regression analysis with confidence intervals;
proc glm plots=diagnostics data=sa;
    model y = x / clparm clm;
run;
* Regression analysis with prediction intervals;
proc glm data=sa;
    model y = x / clparm cli;
run;
quit;
```

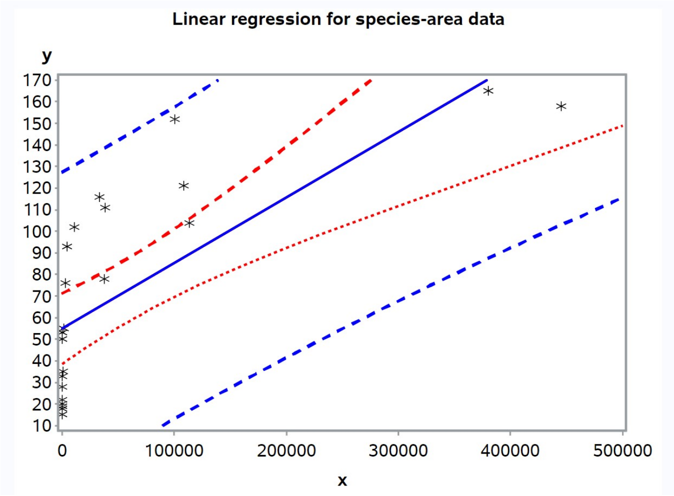


Figure 17.10: `SAprob2.sas - proc gplot`

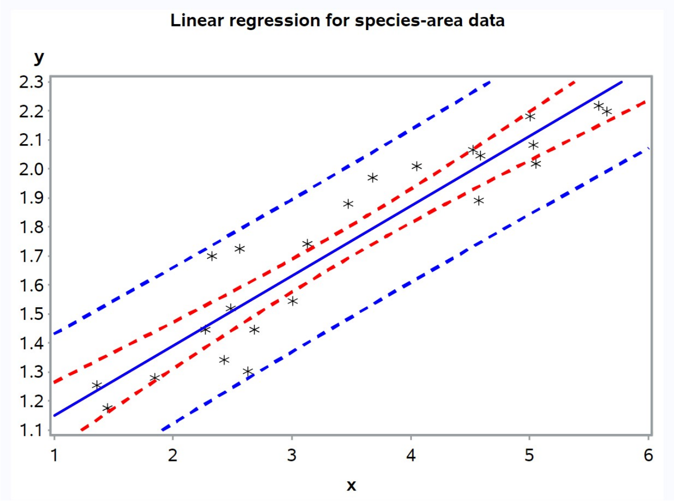
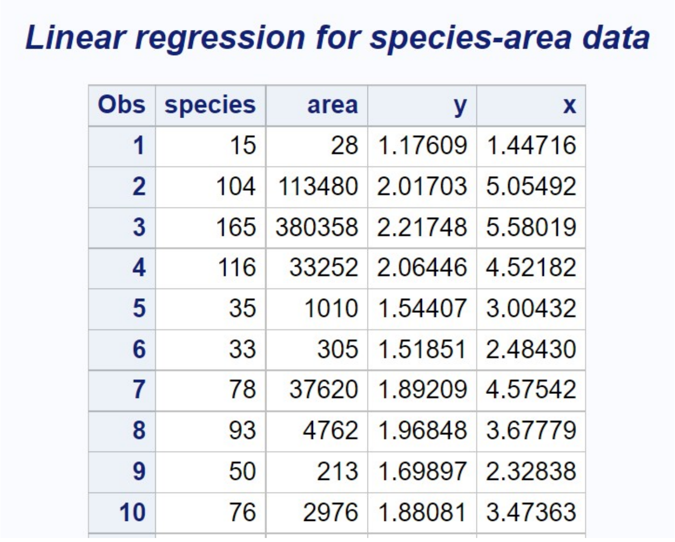


Figure 17.11: `SAprob2.sas - proc gplot`

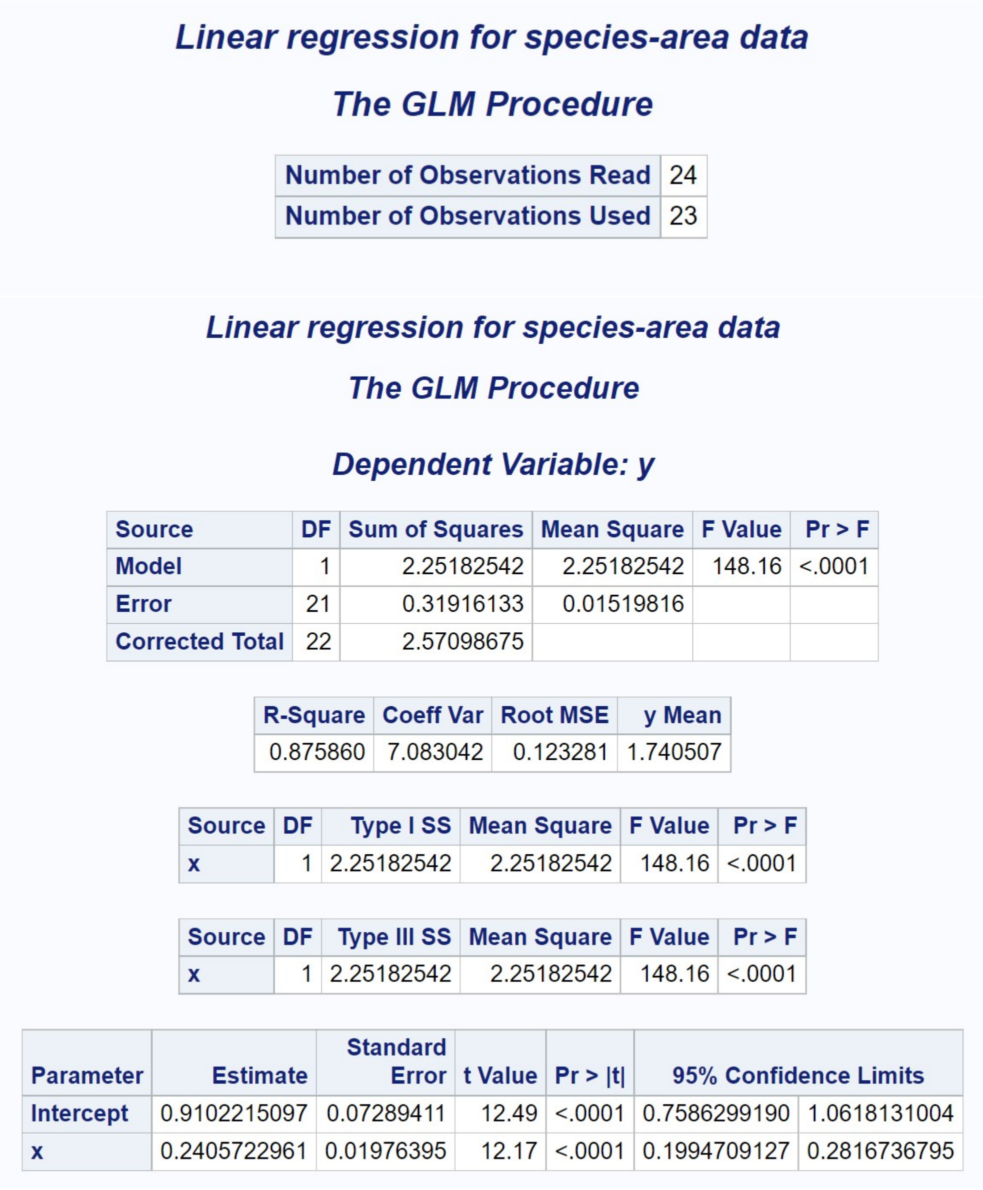


Linear regression for species-area data

Obs	species	area	y	x
1	15	28	1.17609	1.44716
2	104	113480	2.01703	5.05492
3	165	380358	2.21748	5.58019
4	116	33252	2.06446	4.52182
5	35	1010	1.54407	3.00432
6	33	305	1.51851	2.48430
7	78	37620	1.89209	4.57542
8	93	4762	1.96848	3.67779
9	50	213	1.69897	2.32838
10	76	2976	1.88081	3.47363

etc.

Figure 17.12: `SAprob2.sas - proc print`



Linear regression for species-area data

The GLM Procedure

Number of Observations Read	24
Number of Observations Used	23

Linear regression for species-area data

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	2.25182542	2.25182542	148.16	<.0001
Error	21	0.31916133	0.01519816		
Corrected Total	22	2.57098675			

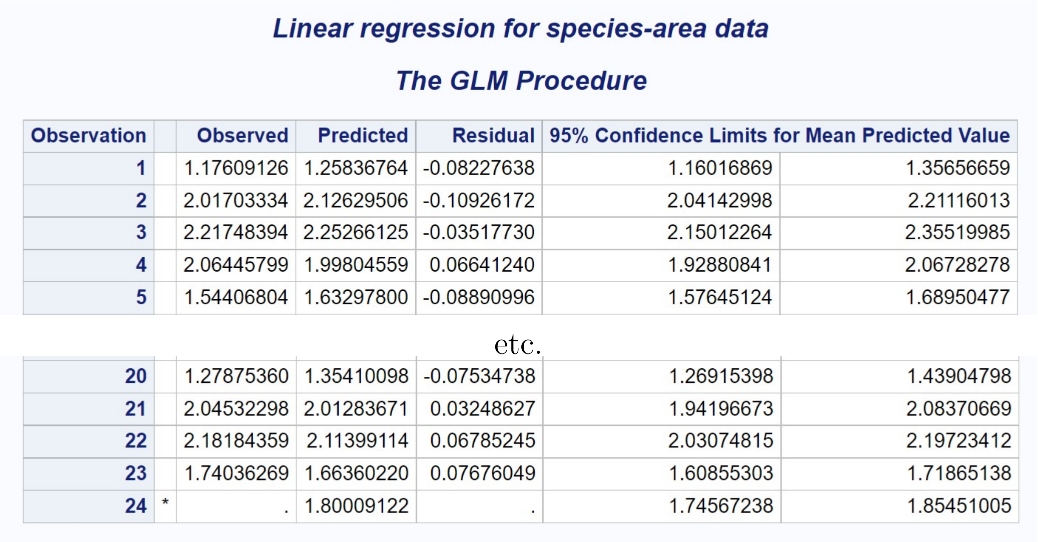
R-Square	Coeff Var	Root MSE	y Mean
0.875860	7.083042	0.123281	1.740507

Source	DF	Type I SS	Mean Square	F Value	Pr > F
x	1	2.25182542	2.25182542	148.16	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
x	1	2.25182542	2.25182542	148.16	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	0.9102215097	0.07289411	12.49	<.0001	0.7586299190	1.0618131004
x	0.2405722961	0.01976395	12.17	<.0001	0.1994709127	0.2816736795

Figure 17.13: `SAprob2.sas - proc glm`



Linear regression for species-area data

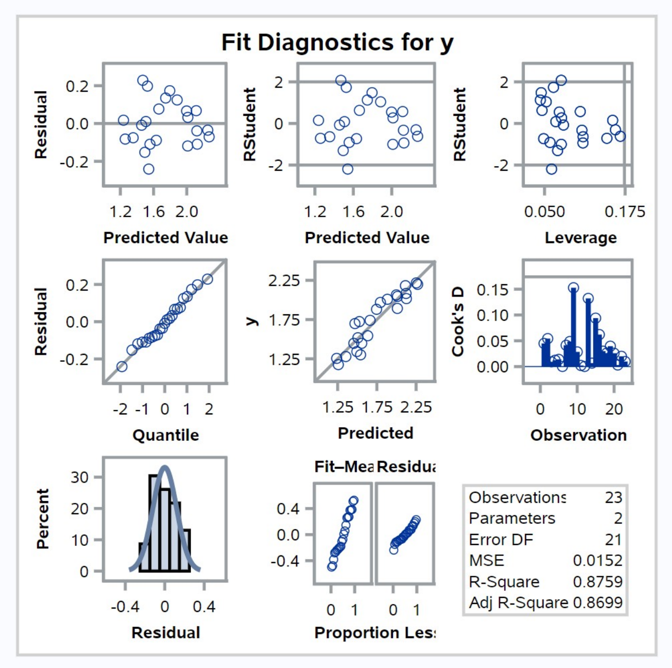
The GLM Procedure

Observation		Observed	Predicted	Residual	95% Confidence Limits for Mean Predicted Value	
1		1.17609126	1.25836764	-0.08227638	1.16016869	1.35656659
2		2.01703334	2.12629506	-0.10926172	2.04142998	2.21116013
3		2.21748394	2.25266125	-0.03517730	2.15012264	2.35519985
4		2.06445799	1.99804559	0.06641240	1.92880841	2.06728278
5		1.54406804	1.63297800	-0.08890996	1.57645124	1.68950477

etc.

Observation		Observed	Predicted	Residual	95% Confidence Limits for Mean Predicted Value	
20		1.27875360	1.35410098	-0.07534738	1.26915398	1.43904798
21		2.04532298	2.01283671	0.03248627	1.94196673	2.08370669
22		2.18184359	2.11399114	0.06785245	2.03074815	2.19723412
23		1.74036269	1.66360220	0.07676049	1.60855303	1.71865138
24	*	.	1.80009122	.	1.74567238	1.85451005

Figure 17.14: `SAprob2.sas - proc glm`



Observations	23
Parameters	2
Error DF	21
MSE	0.0152
R-Square	0.8759
Adj R-Square	0.8699

Figure 17.15: `SAprob2.sas - proc glm`

17.6.2 Population growth rates - SAS demo

As another example of transformations, consider a study of the population growth of phytophagous mites on leaf sections. An experiment was conducted in which leaf sections are inoculated with a range of mite densities and the number of offspring recorded one generation later. The number of offspring per initial mite is the finite growth of the population, usually symbolized as λ. This is the dependent variable in the analysis while mite density is the independent one. The SAS program listed below gives the mite densities and the λ values for this experiment.

We first conduct the analysis without any transformation. Looking at the plot of Y (λ) vs. X (density), we see a curvilinear relationship (Fig. 17.16). A transformation is clearly needed, but which one? A natural log transformation usually a good starting point for population data, both for growth rates and numbers. We begin by log-tranforming the dependent variable λ and find that the plot is now linear (Fig. 17.17).

Interpreting the `proc glm` output below (Fig. 17.19), we see that λ decreased with mite density ($\hat{\beta} = -0.020$) and the effect was highly significant ($F_{1,15} = 1695.22, P < 0.0001$). The R^2 value was 0.991, indicating that almost all the variation in the data was explained by the regression line. It appears that the growth rate of the mites was adversely affected by their density, probably through competition for resources or other intraspecific interactions. The residual plots suggest the homogeneity of variances and normality assumptions were satisfied (Fig. 17.20).

SAS Program

```
* logistic.sas;
title ’Linear regression for growth rate-density data’;
data grd;
    input lambda density;
    * Apply transformations here;
    y = log(lambda);
    x = density;
    datalines;
  7.32   5
  4.82  15
  4.69  25
  3.90  35
  2.65  45
  2.52  55
  1.70  65
  1.68  75
  1.43  85
  1.07  95
  0.74 105
  0.72 115
  0.64 125
  0.47 135
  0.40 145
  0.38 155
  0.25 165
;
run;
* Print data set;
proc print data=grd;
run;
* Plot data and regression line;
proc gplot data=grd;
    plot y*x=1 y*x=2 y*x=3 / overlay vaxis=axis1 haxis=axis1;
    symbol1 i=none v=star c=black height=2 width=3;
    symbol2 i=rlclm v=none c=red height=2 width=3;
    symbol3 i=rlcli v=none c=blue height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* Regression analysis with confidence intervals;
proc glm plots=diagnostics data=grd;
    model y = x / clparm clm;
run;
* Regression analysis with prediction intervals;
proc glm data=grd;
    model y = x / clparm cli;
run;
quit;
```

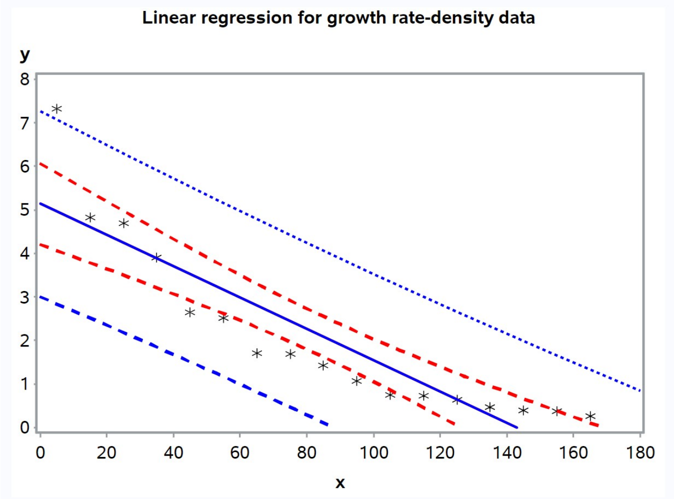


Figure 17.16: `logistic.sas - proc gplot`

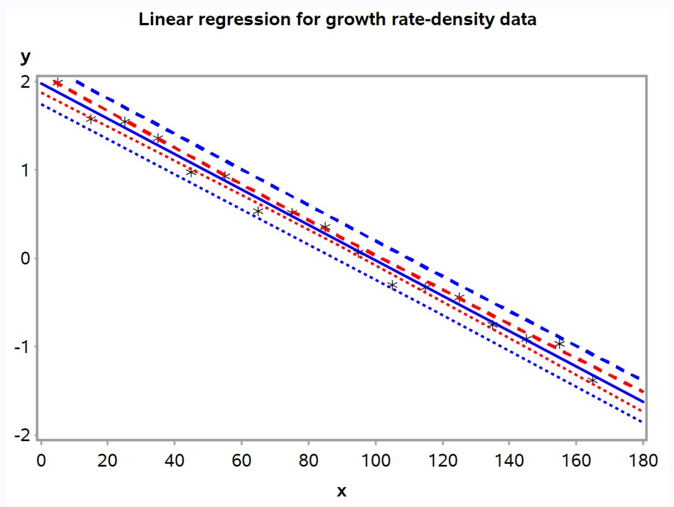
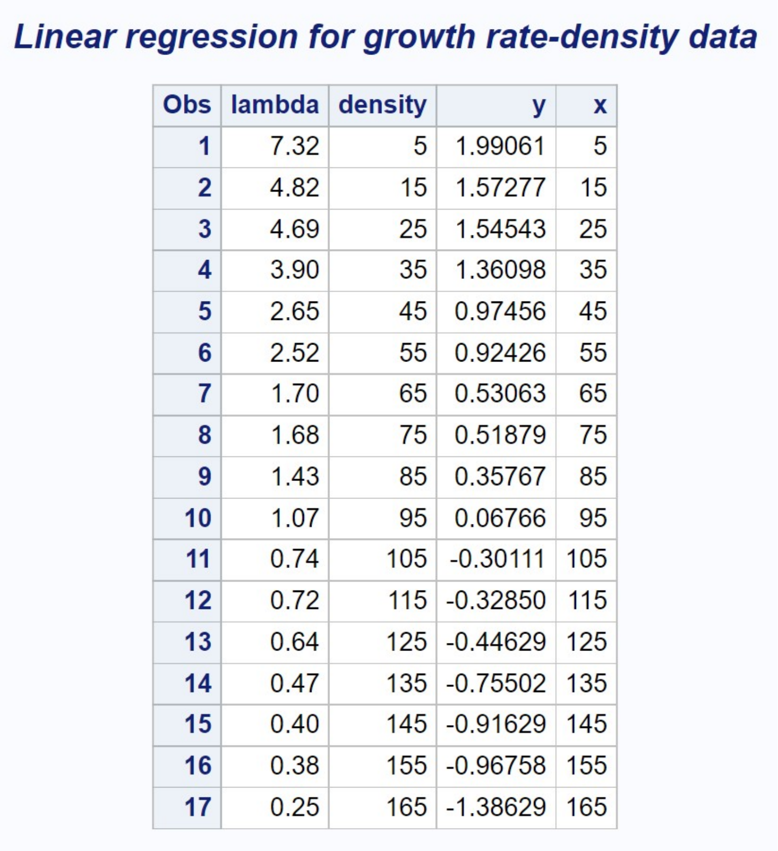


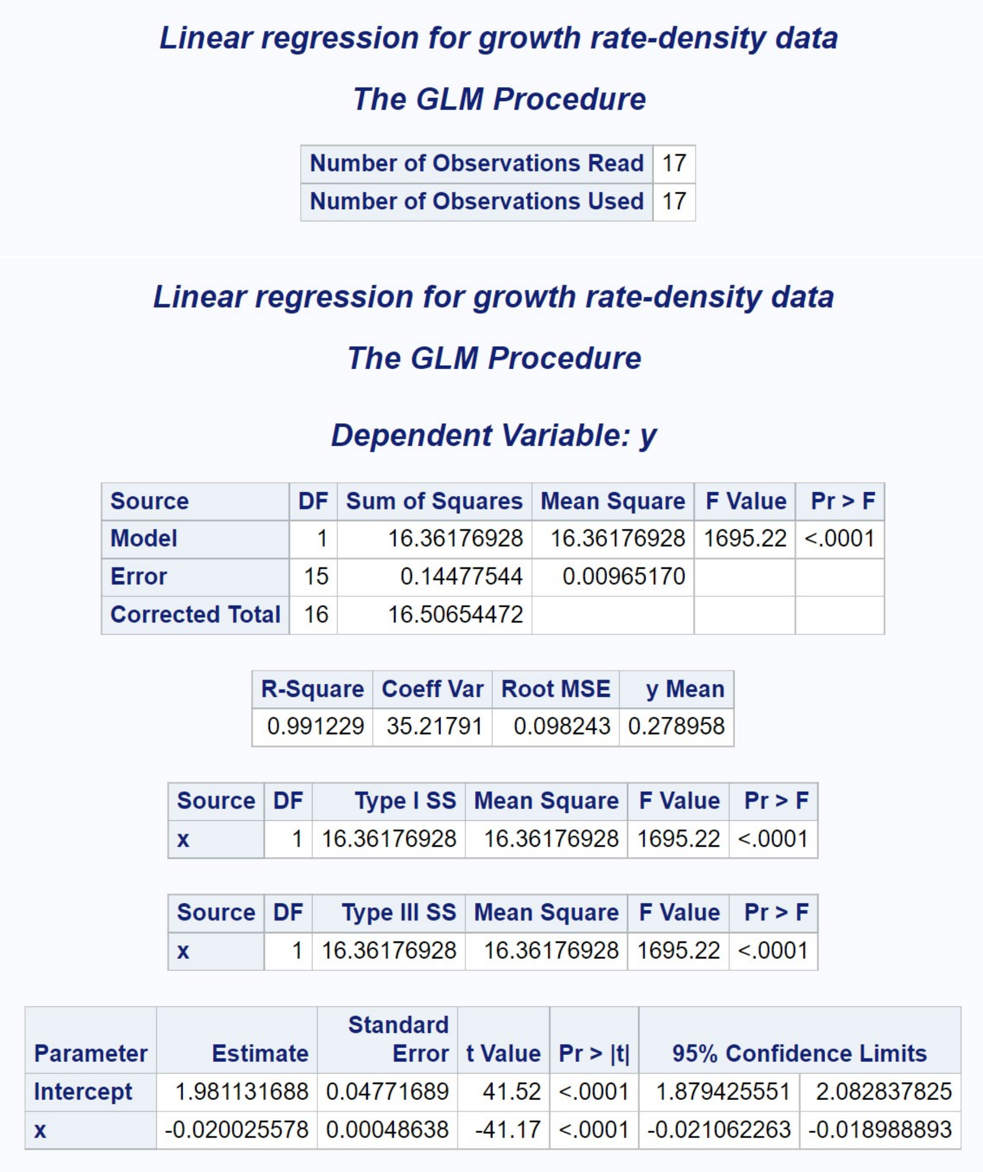
Figure 17.17: `logistic.sas - proc gplot`



Linear regression for growth rate-density data

Obs	lambda	density	y	x
1	7.32	5	1.99061	5
2	4.82	15	1.57277	15
3	4.69	25	1.54543	25
4	3.90	35	1.36098	35
5	2.65	45	0.97456	45
6	2.52	55	0.92426	55
7	1.70	65	0.53063	65
8	1.68	75	0.51879	75
9	1.43	85	0.35767	85
10	1.07	95	0.06766	95
11	0.74	105	-0.30111	105
12	0.72	115	-0.32850	115
13	0.64	125	-0.44629	125
14	0.47	135	-0.75502	135
15	0.40	145	-0.91629	145
16	0.38	155	-0.96758	155
17	0.25	165	-1.38629	165

Figure 17.18: `logistic.sas - proc print`



Linear regression for growth rate-density data

The GLM Procedure

Number of Observations Read	17
Number of Observations Used	17

Linear regression for growth rate-density data

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	16.36176928	16.36176928	1695.22	<.0001
Error	15	0.14477544	0.00965170		
Corrected Total	16	16.50654472			

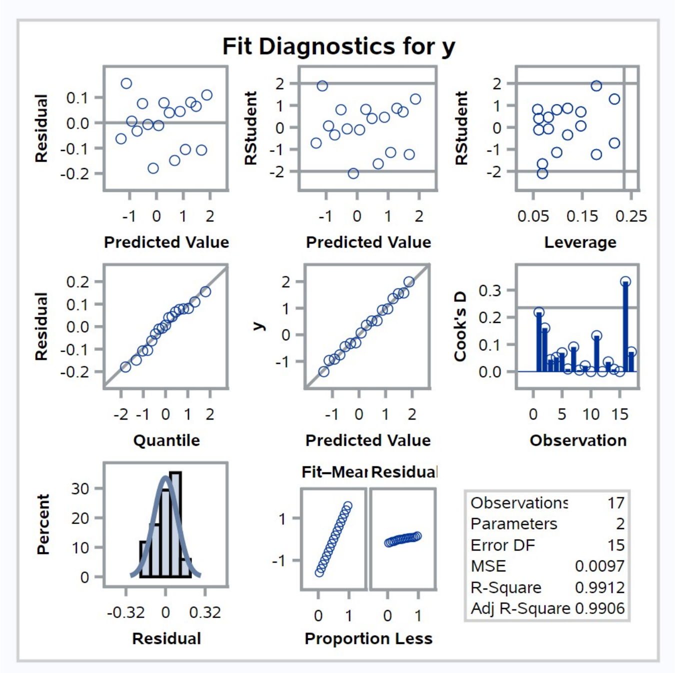
R-Square	Coeff Var	Root MSE	y Mean
0.991229	35.21791	0.098243	0.278958

Source	DF	Type I SS	Mean Square	F Value	Pr > F
x	1	16.36176928	16.36176928	1695.22	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
x	1	16.36176928	16.36176928	1695.22	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1.981131688	0.04771689	41.52	<.0001	1.879425551	2.082837825
x	-0.020025578	0.00048638	-41.17	<.0001	-0.021062263	-0.018988893

Figure 17.19: `logistic.sas - proc glm`



```
Observations      17
Parameters         2
Error DF          15
MSE           0.0097
R-Square      0.9912
Adj R-Square  0.9906
```

Figure 17.20: `logistic.sas - proc glm`

17.7 References

MacArthur, R. H. & Wilson, E. O. (1967) *The Theory of Island Biogeography.* Princeton University Press, Princeton, NJ.

McCulloch, C. E. & Searle, S. R. (2001) *Generalized, Linear, and Mixed Models.* John Wiley & Sons, Inc., New York, NY.

Reeve, J. D., Rhodes, D. J. & Turchin, P. (1998) Scramble competition in southern pine beetle (Coleoptera: Scolytidae). *Ecological Entomology* 23: 433-443.

SAS Institute Inc. (2016) *SAS/GRAPH 9.4: Reference, Fifth Edition.* SAS Institute Inc., Cary, NC.

SAS Institute Inc. (2018) *SAS/STAT 15.1 Users Guide* SAS Institute Inc., Cary, NC

Searle, S. R. (1971) *Linear Models.* John Wiley & Sons, Inc., New York, NY.

17.8 Problems

1. An experiment was conducted to measure the effect of density on the rate of egg laying in cowpea weevils. Ten different densities were used in the experiment, and the rate of egg laying determined for each density. The following data were obtained:

Density	Eggs per day
100	7.629
200	4.530
500	3.820
700	2.718
1200	2.403
1500	1.756
1700	1.772
2000	1.508
2200	1.518
2500	1.359

   (a) Plot the rate of egg laying (Y) vs. density (X), and observe the nonlinear relationship between Y and X. Find a transformation of Y and/or X that linearizes this relationship using SAS.

   (b) For the transformed data, use SAS to plot a 95% confidence interval for the mean of Y_i and a 95% prediction interval for a single value of Y_i. Label the intervals (confidence or prediction) on the `gplot` graph.

   (c) Analyze the transformed data set using linear regression and SAS. In your SAS output, label the 95% confidence intervals for the intercept (α) and slope (β) in your SAS printout.

   (d) Interpret the results of the regression analysis. Is there a significant effect of density on the rate of egg production? What direction is the effect?

2. A zoologist wants to establish the relationship between the length of an animal and its weight. He wants to use length to predict weight in future studies, because length is easier to measure. The lengths and weights of a random sample of 20 animals were determined, yielding the following data:

Length (mm)	Weight (g)
14.7	1.65
19.9	4.86
15.8	2.04
19.0	3.53
8.4	0.32
10.2	0.46
13.5	1.68
22.1	6.24
16.2	1.85
8.2	0.28
10.1	0.48
19.8	4.18
20.6	4.77
22.0	6.10
18.1	2.78
22.4	5.26
10.5	0.55
14.5	1.56
11.9	1.07
14.7	1.74

(a) Plot the weight (Y) vs. length (X) using SAS, and observe the nonlinear relationship between Y and X. Attach your graph of this relationship. Then, find a transformation of Y and/or X that linearizes this relationship. What transformation did you use? Attach your graph showing the transformed relationship.

(b) Analyze the transformed data using linear regression and SAS. Briefly interpret your results using P values. Is there a significant effect of length on weight? What direction is the effect? Attach your program and output.

(c) For animals that are 21 mm long, find a 95% confidence interval for the mean weight.

Chapter 18

Correlation

Correlation is a statistical technique used to examine the **association** between two continuous variables. Unlike regression, correlation does not assume a particular direction to the relationship among the variables, and there is no dependent or independent variable. Instead, there are two random variables Y_1 and Y_2 that could be related in some way. Correlation may be used to examine the relationship between just two variables, or as a screening tool to examine the pairwise relationships among many variables.

We will use a classic data set to illustrate correlation, the iris flowers examined by Fisher (1936). The data set contains measurements of iris flowers for three different *Iris* species, but we will only examine *I. setosa*. The variables measured were sepal length and width, and petal length and width, for a total of 50 observations. We will use sepal length and width for the first ten flowers to illustrate the calculations in a correlation analysis (Table 18.1). The notation Y_{1i} and Y_{2i} refer to the values for the *ith* pair of numbers. For example, $Y_{11} = 5.1$ and $Y_{21} = 3.5$. Figure 18.1 shows there is a positive association between the two variables, with sepal length (Y_{1i}) and width (Y_{2i}) appearing to increase together. We will later examine the correlations among all four variables in the full data set.

Table 18.1: Example 1 - Sepal length and width measurements for ten flowers of *I. setosa* (Fisher 1936), showing some preliminary calculations for the correlation analysis. For these data, $\bar{Y}_1 = 4.86$ and $\bar{Y}_2 = 3.31$. See Chapter 22 for the full data set.

i	Y_{1i} = Sepal length	Y_{2i} = Sepal width	$(Y_{1i} - \bar{Y}_1)(Y_{2i} - \bar{Y}_2)$	$(Y_{1i} - \bar{Y}_1)^2$	$(Y_{2i} - \bar{Y}_2)^2$
1	5.1	3.5	4.56×10^{-2}	5.76×10^{-2}	3.61×10^{-2}
2	4.9	3.0	-1.24×10^{-2}	1.60×10^{-3}	9.61×10^{-2}
3	4.7	3.2	1.76×10^{-2}	2.56×10^{-2}	1.21×10^{-2}
4	4.6	3.1	5.46×10^{-2}	6.76×10^{-2}	4.41×10^{-2}
5	5.0	3.6	4.06×10^{-2}	1.96×10^{-2}	8.41×10^{-2}
6	5.4	3.9	3.19×10^{-1}	2.92×10^{-1}	3.48×10^{-1}
7	4.6	3.4	-2.34×10^{-2}	6.76×10^{-2}	8.10×10^{-3}
8	5.0	3.4	1.26×10^{-2}	1.96×10^{-2}	8.10×10^{-3}
9	4.4	2.9	1.89×10^{-1}	2.12×10^{-1}	1.68×10^{-1}
10	4.9	3.1	-8.40×10^{-3}	1.60×10^{-3}	4.41×10^{-2}
$\sum$	48.6	33.1	6.34×10^{-1}	7.64×10^{-1}	8.49×10^{-1}

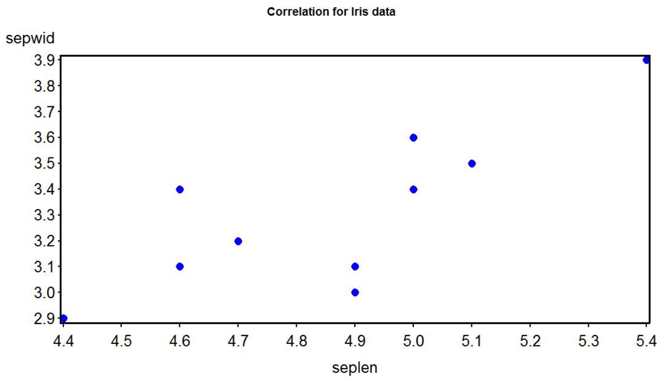


Figure 18.1: Scatterplot of *I. setosa* sepal length and width

18.1 Correlation model

The statistical model for correlation is the **bivariate normal distribution**. This is an extension of the normal distribution to a pair of random variables, Y_1 and Y_2, that have a joint probability distribution. This differs from the continuous (and discrete) random variables we previously studied, that model the behavior of a single observation Y and so are classified as **univariate distributions**.

The bivariate normal distribution has five parameters, the mean and standard deviation for Y_1 and Y_2 $(\mu_1, \sigma_1, \mu_2, \sigma_2)$ and the parameter ρ, which ranges between -1 and 1 (Stuart et al. 1999). This parameter describes the association between Y_1 and Y_2: if $\rho > 0$ then the two variables are positively related, as in Fig. 18.1, while if the $\rho < 0$ they are inversely related. If $\rho = 0$ the two variables are independent of one another. The probability density for the bivariate normal distribution is given by the function

$$f(y_1, y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \times \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{y_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{y_1-\mu_1}{\sigma_1}\frac{y_2-\mu_2}{\sigma_2} + \left(\frac{y_2-\mu_2}{\sigma_2}\right)^2\right\}\right]. \tag{18.1}$$

A interesting property of this distribution is that each Y variable, when considered alone, also has a normal distribution. In particular, $Y_1 \sim N(\mu_1, \sigma_1^2)$ and $Y_2 \sim N(\mu_2, \sigma_2^2)$. These are known as the **marginal distributions** of Y_1 and Y_2.

Figure 18.2 and Fig. 18.3 shows this distribution as a surface or contour plot, for $\rho = 0.7$. This value of ρ implies a strong positive relationship between the two variables, and so the probability density has a ridge-like shape because Y_1 and Y_2 are likely to increase or decrease together. Fig. 18.4 shows a sample data set generated for the same parameter values of this distribution. Note the relationship between Y_1 and Y_2 and the elliptical cloud of points.

Figure 18.5 shows the distribution for a strong negative relationship between the variables ($\rho = -0.7$). A sample data set for the same parameter values is shown in Fig. 18.6. Figure 18.7 and Fig. 18.8 show the patterns when the two variables are unassociated or independent ($\rho = 0$).

The usual goal in correlation is to estimate the value of ρ and then test $H_0 : \rho = 0$. This null hypothesis means the two variables are independent, and if we can reject this suggests the two variables are associated or dependent. It is also possible to test null hypotheses of the form $H_0 : \rho = \rho_0$, where ρ_0 is any value.

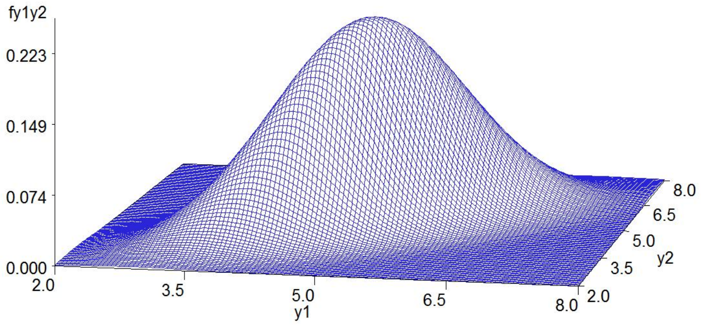


Figure 18.2: Surface plot of the bivariate normal distribution for $\mu_1 = \mu_2 = 5, \sigma_1^2 = \sigma_2^2 = 1$, and $\rho = 0.7$.

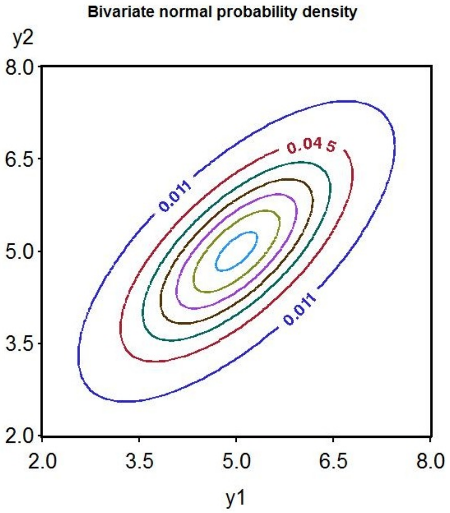


Figure 18.3: Contour plot of the bivariate normal for the same parameter values as Fig. 18.3

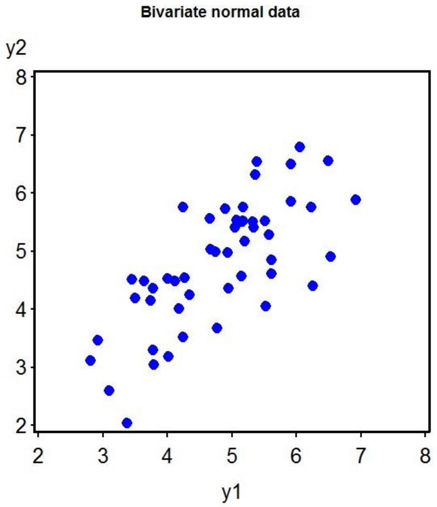


Figure 18.4: Simulated data for the bivariate normal distribution with $\mu_1 = \mu_2 = 5, \sigma_1^2 = \sigma_2^2 = 1$, and $\rho = 0.7$.

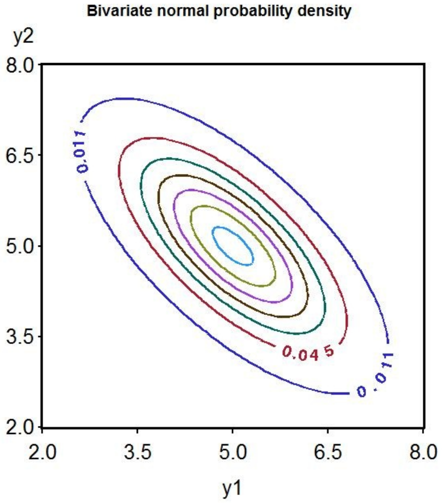


Figure 18.5: Contour plot of the bivariate normal for $\mu_1 = \mu_2 = 5, \sigma_1^2 = \sigma_2^2 = 1$, and $\rho = -0.7$.

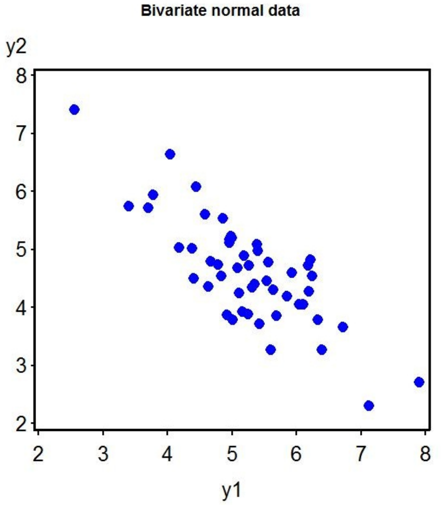


Figure 18.6: Simulated data for the bivariate normal distribution with $\mu_1 = \mu_2 = 5, \sigma_1^2 = \sigma_2^2 = 1$, and $\rho = -0.7$.

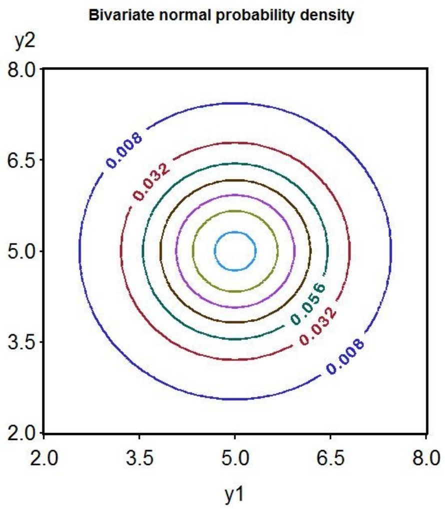


Figure 18.7: Contour plot of the bivariate normal for $\mu_1 = \mu_2 = 5, \sigma_1^2 = \sigma_2^2 = 1$, and $\rho = 0$.

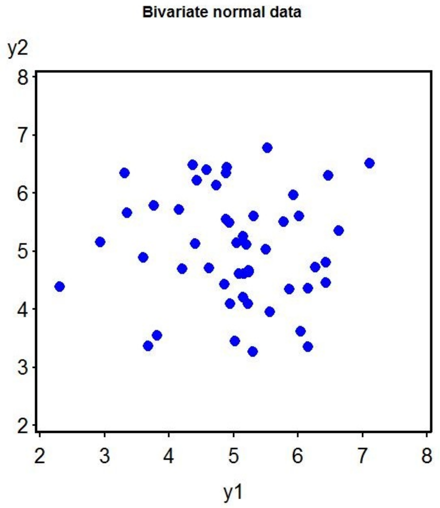


Figure 18.8: Simulated data for the bivariate normal distribution with $\mu_1 = \mu_2 = 5, \sigma_1^2 = \sigma_2^2 = 1$, and $\rho = 0$.

18.2 Correlation and maximum likelihood

Maximum likelihood can be used to estimate the parameters for the bivariate normal distribution, using methods like those for simpler ones. It turns out that the sample mean $\bar{Y}$ and standard deviation s can be used to estimate μ_1, σ_1, μ_2, and σ_2 for this distribution. For the Example 1 data set, we have $\bar{Y}_1 = 4.86, s_1 = 0.29136, \bar{Y}_2 = 3.31$, and $s_2 = 0.30714$. The maximum likelihood estimator of ρ is the sample **correlation coefficient**, r, given by the formula

$$r = \frac{\sum_{i=1}^{n}(Y_{1i} - \bar{Y}_1)(Y_{2i} - \bar{Y}_2)}{\sqrt{\sum_{i=1}^{n}(Y_{1i} - \bar{Y}_1)^2 \sum_{i=1}^{n}(Y_{2i} - \bar{Y}_2)^2}} \tag{18.2}$$

(Stuart et al. 1999). Note that the sign of r depends on the numerator of this expression. If Y_1 and Y_2 are positively or negatively associated, the numerator will be positive or negative. For the Example 1 data, we have

$$\sum_{i=1}^{n}(Y_{1i} - \bar{Y}_1)(Y_{2i} - \bar{Y}_2) = 0.634, \tag{18.3}$$

$$\sum_{i=1}^{n}(Y_{1i} - \bar{Y}_1)^2 = 0.764, \tag{18.4}$$

$$\text{and} \sum_{i=1}^{n}(Y_{2i} - \bar{Y}_2)^2 = 0.849. \tag{18.5}$$

Using these values, the correlation coefficient can then be calculated:

$$r = \frac{6.34 \times 10^{-1}}{\sqrt{7.64 \times 10^{-1} \times 8.49 \times 10^{-1}}} = 0.787. \tag{18.6}$$

The equation for r can also be expressed using the standard deviations of the two variables, and a quantity called the **sample covariance**. The sample covariance is given by the formula

$$s_{12} = \frac{\sum_{i=1}^{n}(Y_{1i} - \bar{Y}_1)(Y_{2i} - \bar{Y}_2)}{n-1}. \tag{18.7}$$

Dividing the top and bottom of the equation for r by $n - 1$, we have

$$r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(Y_{1i} - \bar{Y}_1)(Y_{2i} - \bar{Y}_2)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_{1i} - \bar{Y}_1)^2 \frac{1}{n-1}\sum_{i=1}^{n}(Y_{2i} - \bar{Y}_2)^2}} \tag{18.8}$$

$$= \frac{s_{12}}{\sqrt{s_1^2 s_2^2}} = \frac{s_{12}}{s_1 s_2} \tag{18.9}$$

Thus, r can be expressed as the sample covariance s_{12} scaled by the standard deviation s_1 and s_2 for each variable. This quantity is also known as the **Pearson correlation coefficient**.

The square of the correlation coefficient is called the **coefficient of determination**, and provides an indication of the amount of variability in Y_1 explained by Y_2, or vice versa. It is typically written as R^2 like in linear regression or ANOVA. The value of R^2 ranges from zero to one, with values near one implying a strong relationship (positive or negative) between Y_1 and Y_2, while values near zero imply a weak one. For the Example 1 data, we have $R^2 = 0.787^2 = 0.619$. About 62% of the variability in Y_1 is explained by Y_2, or vice versa.

There is also a likelihood ratio test for $H_0 : \rho = 0$ vs. $H_1 : \rho \neq 0$, equivalent to testing whether Y_1 is independent of Y_2. Under H_0, the test statistic

$$T_s = r\sqrt{\frac{n-2}{1-r^2}} \tag{18.10}$$

has a t distribution with $n-2$ degrees of freedom, and we would reject H_0 for sufficiently large values (Stuart et al. 1999). For the Example 1 data, we have

$$T_s = 0.787\sqrt{\frac{10-2}{1-0.787^2}} = 3.608. \tag{18.11}$$

Using Table T with $10 - 2 = 8$ degrees of freedom, we see that $P < 0.01$. The correlation between sepal length and width was highly significant ($t_8 = 3.608, P < 0.01$), and so the two variables appear dependent, not independent.

There is an approximate test for $H_0 : \rho = \rho_0$ vs. $H_0 : \rho \neq \rho_0$, for values ρ_0 different from zero. It uses a special transformation for r, the inverse hyperbolic tangent function:

$$\operatorname{arctanh}(r) = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right), \tag{18.12}$$

defined for $-1 < r < 1$. The effect of this transformation is to spread out the distribution of r and make it more normal. Under H_0, we have $E[\operatorname{arctanh}(r)] \approx \operatorname{arctanh}(\rho_0)$ and $Var[\operatorname{arctanh}(r)] \approx 1/(n-3)$, and so

$$\begin{aligned} Z_s &= \frac{\operatorname{arctanh}(r) - \operatorname{arctanh}(\rho_0)}{\sqrt{1/(n-3)}} && (18.13) \\ &= \sqrt{n-3}\,[\operatorname{arctanh}(r) - \operatorname{arctanh}(\rho_0)] \sim N(0,1) && (18.14) \end{aligned}$$

for large n (Stuart et al. 1999). As an example of this test, suppose we want to test $H_0 : \rho = 0.5$ for the Example 1 data set. We have

$$Z_s = \sqrt{10-3}\,[\text{arctanh}(0.787) - \text{arctanh}(0.5)] \tag{18.15}$$

$$= 2.646(1.064 - 0.549) = 1.363. \tag{18.16}$$

Using Table Z and the method in Chapter 10, we find that $P = 0.1738$. The correlation coefficient was not significantly different from 0.5 ($Z_s = 1.363, P = 0.1738$).

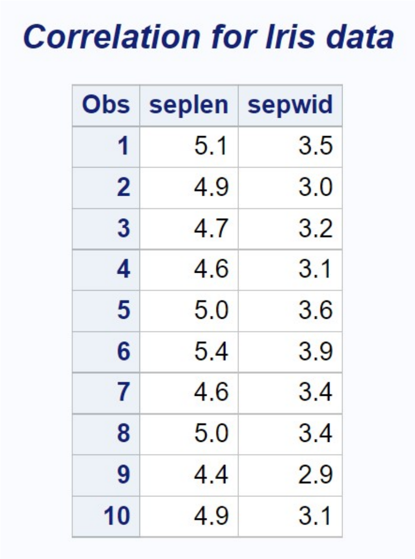
18.2.1 Correlation for Example 1 - SAS demo

We can conduct a correlation analysis using `proc corr` in SAS (see program below). We first input the observations using a `data` step. Within the `proc corr` section of the program, we specify the variables to be analyzed using a `var` statement (SAS Institute Inc. 2016). The option `plots=(scatter matrix)` generates pairwise scatterplots of all the variables, and then a scatterplot matrix that plots all possible pairs of variables in one graph.

From the `proc corr` output (Fig. 18.10), we see that the correlation between sepal length and width was highly significant ($r = 0.787, P = 0.0069$). The pairwise scatterplot and scatterplot matrix are also shown (Fig. 18.11).

SAS Program

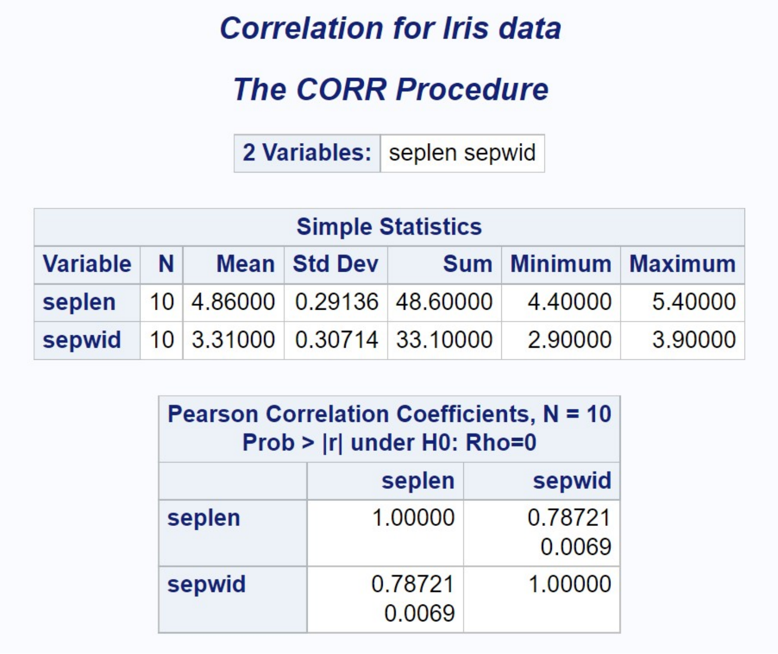
```
* Iris.sas;
title "Correlation for Iris data";
data iris;
    input seplen sepwid;
    datalines;
5.1 3.5
4.9 3.0
4.7 3.2
4.6 3.1
5.0 3.6
5.4 3.9
4.6 3.4
5.0 3.4
4.4 2.9
4.9 3.1
;
run;
* Print data set;
proc print data=iris;
run;
* Correlation analysis and scatterplots;
proc corr data=iris plots=(scatter matrix);
    var seplen sepwid;
run;
quit;
```



Correlation for Iris data

Obs	seplen	sepwid
1	5.1	3.5
2	4.9	3.0
3	4.7	3.2
4	4.6	3.1
5	5.0	3.6
6	5.4	3.9
7	4.6	3.4
8	5.0	3.4
9	4.4	2.9
10	4.9	3.1

Figure 18.9: `Iris.sas - proc print`



Correlation for Iris data

The CORR Procedure

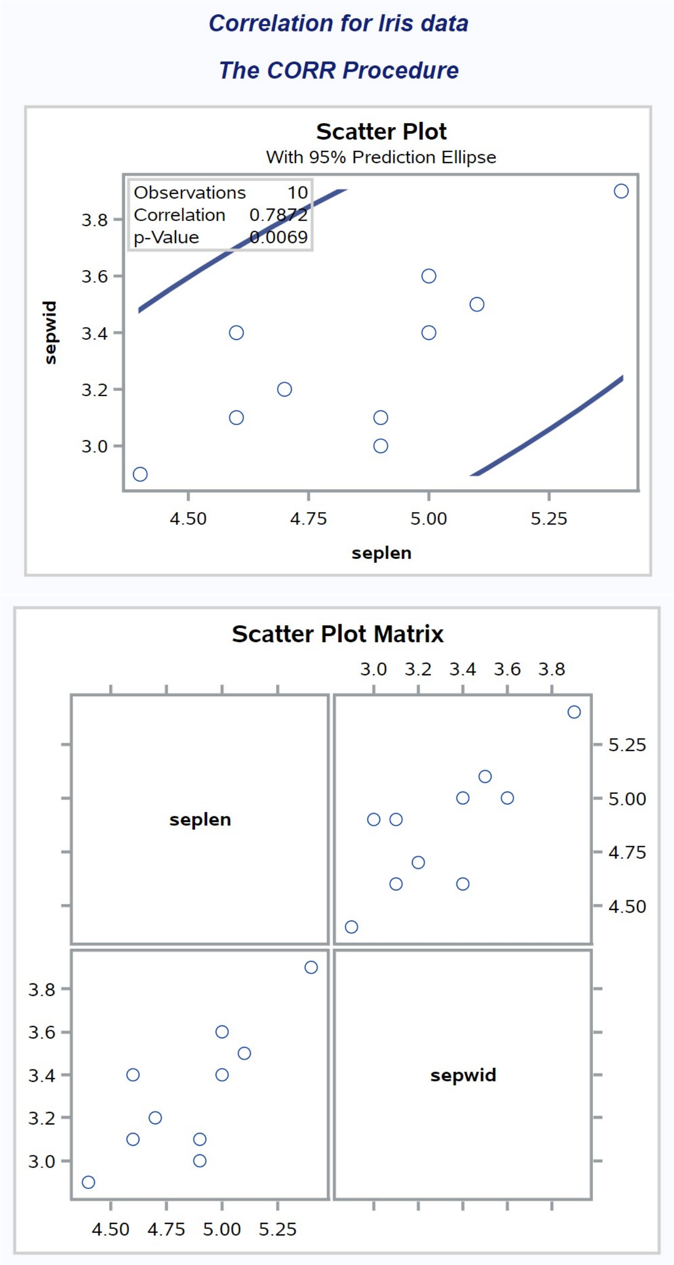
2 Variables:	seplen sepwid

Simple Statistics						
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
seplen	10	4.86000	0.29136	48.60000	4.40000	5.40000
sepwid	10	3.31000	0.30714	33.10000	2.90000	3.90000

Pearson Correlation Coefficients, N = 10
Prob > |r| under H0: Rho=0

	seplen	sepwid
seplen	1.00000	0.78721 0.0069
sepwid	0.78721 0.0069	1.00000

Figure 18.10: `Iris.sas - proc corr`



Correlation for Iris data

The CORR Procedure

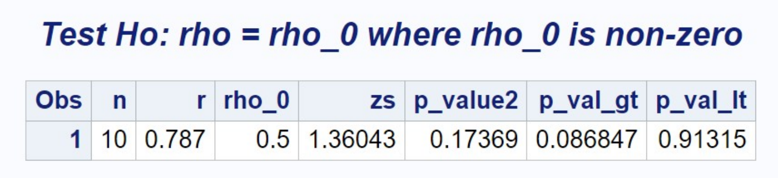
Figure 18.11: `Iris.sas - proc corr`

18.2.2 Testing $H_0 : \rho = \rho_0$ - SAS demo

We can use a short SAS program to test $H_0 : \rho = 0.5$ vs. $H_1 : \rho \neq 0.5$ for the Example 1 data (see program and output below). The program calculates the P value for this two-tailed alternative (`pvalue2`) as well as both one-tailed ones (`p_val_gt`,`p_val_lt`). We see that the correlation between sepal length and width is not significantly different from 0.5 ($Z_s = 1.360, P = 0.1737$).

SAS Program

```
* rhocalc.sas;
title ’Test Ho: rho = rho_0 where rho_0 is non-zero’;
data rhocalc;
    * Input sample size, rho, and rho_0;
    n = 10;
    r = 0.787;
    rho_0 = 0.5;
    zs = sqrt(n-3)*(artanh(r)-artanh(rho_0));
    * P-value for two-tailed test;
    p_value2 = 2*(1 - probnorm(abs(zs)));
    * P-values for one-tailed tests;
    * Ho: rho = rho_0 vs. H1: rho > rho_0;
    p_val_gt = 1 - probnorm(zs);
    * Ho: rho = rho_0 vs. H1: rho < rho_0;
    p_val_lt = probnorm(zs);
run;
* Print test results;
proc print data=rhocalc;
run;
```



Test Ho: rho = rho_0 where rho_0 is non-zero

Obs	n	r	rho_0	zs	p_value2	p_val_gt	p_val_lt
1	10	0.787	0.5	1.36043	0.17369	0.086847	0.91315

Figure 18.12: `rhocalc.sas - proc print`

18.2.3 Correlation for *I. setosa*, all data - SAS demo

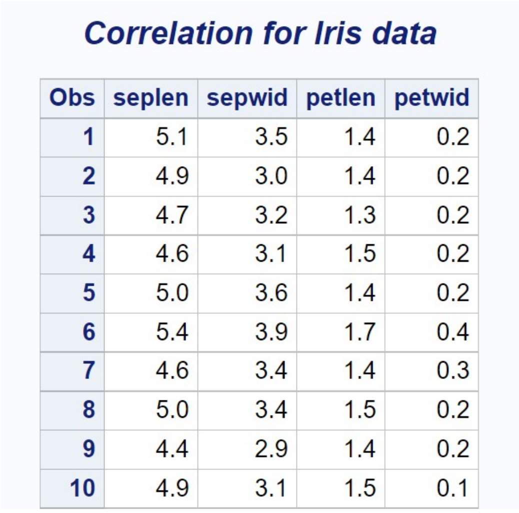
We now analyze the full data set for *I. setosa*, as listed in Chapter 22. We will examine the correlation between sepal length, sepal width, petal length, and petal width for all 50 flowers. The SAS program is similar to the Example 1 analysis, except that all four variables are listed in the `data` and `proc corr` steps. From Fig. 18.14, we see there was a highly significant correlation between sepal length and width ($r = 0.743, P < 0.0001$), and petal length and width were also significantly correlated ($r = 0.332, P = 0.0186$). All the remaining correlations were nonsignificant. It appears that measurements of the same structure (petal or sepal) are correlated, but the correlation is weaker between structures. The scatterplot matrix (Fig. 18.15) reflects these patterns, with sepal length and width showing a strong positive association, with a weaker one for petal length and width. The remaining pairs show no obvious relationships.

SAS Program

```
* Iris_all.sas;
title "Correlation for Iris data";
data iris;
    input seplen sepwid petlen petwid;
    datalines;
5.1 3.5 1.4 0.2
4.9 3.0 1.4 0.2
4.7 3.2 1.3 0.2
4.6 3.1 1.5 0.2
5.0 3.6 1.4 0.2

etc.

4.8 3.0 1.4 0.3
5.1 3.8 1.6 0.2
4.6 3.2 1.4 0.2
5.3 3.7 1.5 0.2
5.0 3.3 1.4 0.2
;
run;
* Print data set;
proc print data=iris;
run;
* Correlation analysis and scatterplots;
proc corr data=iris plots=(scatter matrix);
    var seplen sepwid petlen petwid;
run;
quit;
```

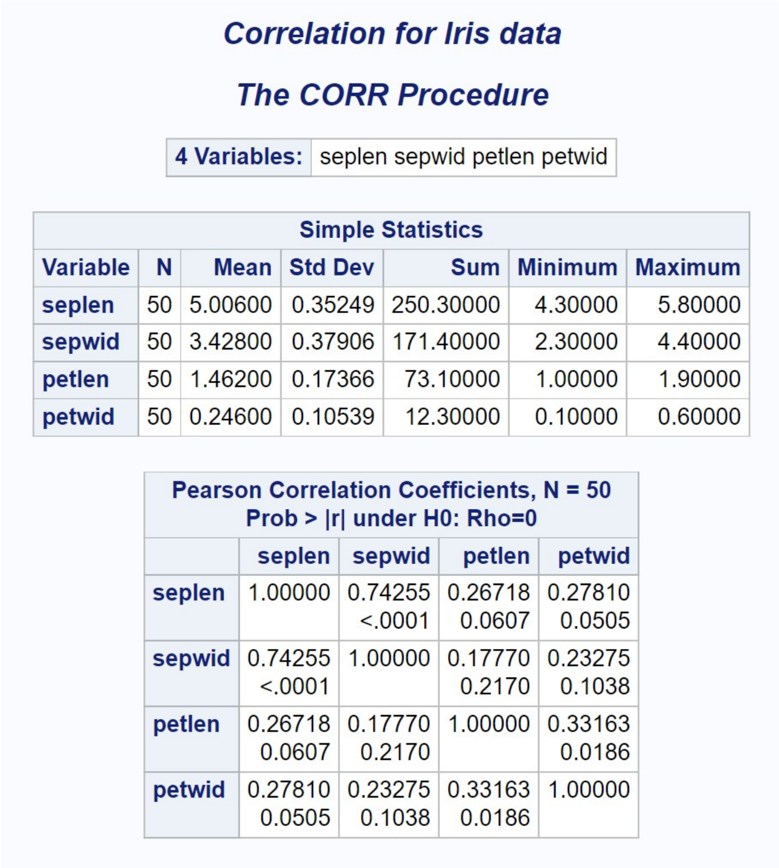


Correlation for Iris data

Obs	seplen	sepwid	petlen	petwid
1	5.1	3.5	1.4	0.2
2	4.9	3.0	1.4	0.2
3	4.7	3.2	1.3	0.2
4	4.6	3.1	1.5	0.2
5	5.0	3.6	1.4	0.2
6	5.4	3.9	1.7	0.4
7	4.6	3.4	1.4	0.3
8	5.0	3.4	1.5	0.2
9	4.4	2.9	1.4	0.2
10	4.9	3.1	1.5	0.1

etc.

Figure 18.13: `Iris_all.sas - proc print`



Correlation for Iris data

The CORR Procedure

4 Variables:	seplen sepwid petlen petwid

Simple Statistics						
Variable	**N**	**Mean**	**Std Dev**	**Sum**	**Minimum**	**Maximum**
seplen	50	5.00600	0.35249	250.30000	4.30000	5.80000
sepwid	50	3.42800	0.37906	171.40000	2.30000	4.40000
petlen	50	1.46200	0.17366	73.10000	1.00000	1.90000
petwid	50	0.24600	0.10539	12.30000	0.10000	0.60000

Pearson Correlation Coefficients, N = 50 Prob > |r| under H0: Rho=0				
	seplen	**sepwid**	**petlen**	**petwid**
seplen	1.00000	0.74255 <.0001	0.26718 0.0607	0.27810 0.0505
sepwid	0.74255 <.0001	1.00000	0.17770 0.2170	0.23275 0.1038
petlen	0.26718 0.0607	0.17770 0.2170	1.00000	0.33163 0.0186
petwid	0.27810 0.0505	0.23275 0.1038	0.33163 0.0186	1.00000

Figure 18.14: `Iris_all.sas - proc corr`

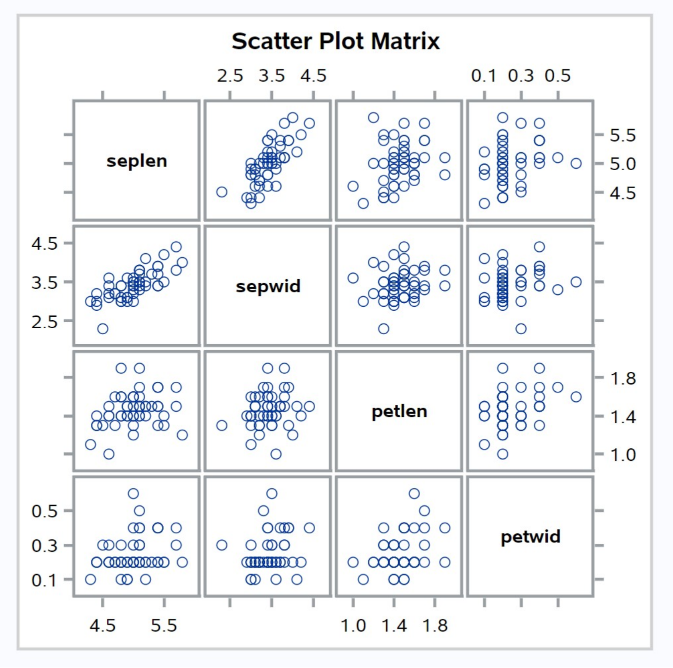


Figure 18.15: `Iris_all.sas - proc corr`

18.3 Correlation assumptions

The main assumption of correlation is that the data have a bivariate normal distribution. If the data do not appear to be bivariate normal, it may be useful to transform one or both variables. The same transformations used in linear regression may be helpful (see Chapter 17). For example, suppose that the relationship between Y_1 and Y_2 appears to be curved (Fig. 18.16). A log transformation of Y_2 makes the overall distribution more similar to the bivariate normal (Fig. 18.17). Once the distribution appears correct, we would calculate the correlation coefficient r and conduct our tests.

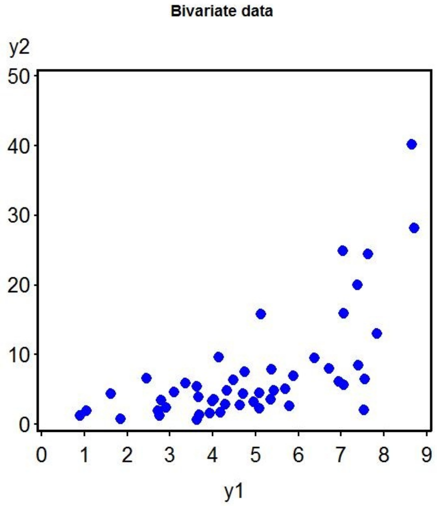


Figure 18.16: Simulated data showing a curved relationship between Y_1 and Y_2.

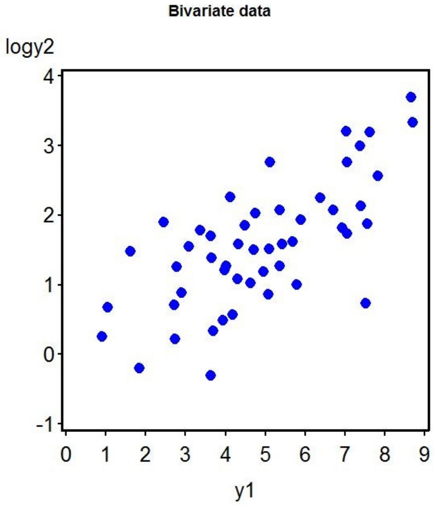


Figure 18.17: Simulated data showing a bivariate normal distribution for Y_1 and $\ln(Y_2)$.

18.4 Nonparametric correlation

There are also nonparametric correlation methods useful when the observations are not bivariate normal. One common method is the Spearman rank correlation test (Hollander et al. 2014). This procedure simply substitutes the rank values of Y_1 and Y_2 in the formula for r, then proceeds as before. We are still interested in testing whether Y_1 and Y_2 are independent, but no distribution is specified.

We will illustrate the Spearman rank correlation procedure using the Example 1 data set. The initial calculations are shown in Table 18.2. We next calculate the Spearman rank correlation r_s using the results from this table. We have

$$r_s = \frac{62.00}{\sqrt{81.00 \times 81.50}} = 0.763. \tag{18.17}$$

If we want to test whether Y_1 and Y_2 are independent, we can use the same test procedure as before, but substituting r_s for r. For the Table 18.2 data, we have

$$T_s = r_s\sqrt{\frac{n-2}{1-r_s^2}} = 0.763\sqrt{\frac{10-2}{1-0.763^2}} = 3.339. \tag{18.18}$$

Using Table T with $10-2=8$ degrees of freedom, we see that $P < 0.02$. The test was significant ($r_s = 0.763, P < 0.02$), which suggests the two variables are not independent.

Table 18.2: Preliminary calculations for Spearman rank correlation using the Example 1 data. Here R_{1i} and R_{2i} are the rank values of sepal length and width, with $\bar{R}_1 = \bar{R}_2 = 5.5$. Note that tied ranks are assigned their average value.

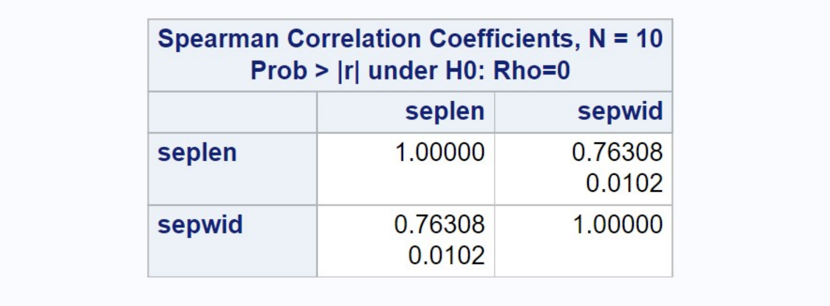
i	Y_{1i} = Sepal length	Y_{2i} = Sepal width	R_{1i}	R_{2i}	$(R_{1i} - \bar{R}_1)(R_{2i} - \bar{R}_2)$	$(R_{1i} - \bar{R}_1)^2$	$(R_{2i} - \bar{R}_2)^2$
1	5.1	3.5	9	8	8.75	12.25	6.25
2	4.9	3.0	5.5	2	0.00	0.00	12.25
3	4.7	3.2	4	5	0.75	2.25	0.25
4	4.6	3.1	2.5	3.5	6.00	9.00	4.00
5	5.0	3.6	7.5	9	7.00	4.00	12.25
6	5.4	3.9	10	10	20.25	20.25	20.25
7	4.6	3.4	2.5	6.5	-3.00	9.00	1.00
8	5.0	3.4	7.5	6.5	2.00	4.00	1.00
9	4.4	2.9	1	1	20.25	20.25	20.25
10	4.9	3.1	5.5	3.5	0.00	0.00	4.00
$\sum$	-	-	55	55	62.00	81.00	81.50

18.4.1 Spearman rank correlation for Example 1 - SAS demo

The Spearman rank correlation and tests can be conducted in SAS by adding the `spearman` option to the `proc corr` statement. For the Table 18.2 data, we obtain $r_s = 0.763, P = 0.0102$. See SAS code and output below.

SAS Program

```
proc corr data=iris plots=(scatter matrix) spearman;
```



Spearman Correlation Coefficients, N = 10
Prob > |r| under H0: Rho=0

	seplen	sepwid
seplen	1.00000	0.76308 0.0102
sepwid	0.76308 0.0102	1.00000

Figure 18.18: `Iris.sas - proc corr`

18.5 References

Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7: 179188.

Hollander, M., Wolfe, D. A., & Chicken, E. (2014) *Nonparametric Statistical Methods, Third Edition.* John Wiley & Sons, Inc., Hoboken, NJ.

Moser, J. C., Reeve, J. D., Bento, J. M. S., Della Lucia, T. M. C., Cameron, R. S. & Heck, N. M. (2004) Eye size and behaviour of day- and night-flying leafcutting ant alates. *Journal of Zoology* 264: 69-75.

SAS Institute Inc. (2016) *Base SAS 9.4 Procedures Guide, Sixth Edition.* SAS Institute Inc., Cary, NC.

Stuart, A., Ord, J. K. & Arnold, S. (1999) *Kendall's Advanced Theory of Statistics.* Oxford University Press Inc., New York, NY.

18.6 Problems

1. An entomologist was interested in variation in eye and head size for leaf-cutting ants (Moser et al. 2004). A microscope was used to measure the width of the head (mm) and the surface area of the eyes and ocelli (mm^2). The surface areas were then square-root transformed. The following data were obtained for the females of one species (*Atta sexdens*).

   (a) Calculate all pairwise correlations among these variables using SAS. Interpret the results of this analysis, providing a P value and discussing the significance of the test. Provide a biological explanation for the positive correlations among these variables.

   (b) Test whether each of the pairwise correlations is significantly different from 0.2.

   (c) Calculate all pairwise Spearman rank correlations using SAS. Interpret the results of this analysis.

Head	Eye	Ocelli	Head	Eye	Ocelli
4.1	0.660	0.311	3.8	0.633	0.290
4.1	0.651	0.301	3.9	0.659	0.293
4.1	0.614	0.287	4.0	0.633	0.287
4.1	0.668	0.301	4.1	0.614	0.295
4.0	0.659	0.298	4.2	0.678	0.295
4.1	0.659	0.306	4.2	0.668	0.292
4.1	0.678	0.311	4.1	0.668	0.304
4.0	0.668	0.311	4.2	0.678	0.298
4.0	0.601	0.285	4.2	0.678	0.286
3.9	0.651	0.288	3.9	0.646	0.295
4.1	0.678	0.303	4.0	0.633	0.295
4.1	0.665	0.298	4.1	0.659	0.295
4.2	0.668	0.306	4.0	0.646	0.296
4.0	0.668	0.306	4.1	0.655	0.298
4.1	0.678	0.306	4.0	0.659	0.290
4.0	0.659	0.301	4.1	0.678	0.298
3.9	0.659	0.298	4.1	0.678	0.301
4.1	0.678	0.304	4.1	0.668	0.298
4.2	0.668	0.299	4.1	0.659	0.295
4.1	0.659	0.304	4.2	0.678	0.301
4.1	0.665	0.301	3.9	0.687	0.296
4.2	0.665	0.307	4.0	0.614	0.293
4.1	0.651	0.306	4.1	0.668	0.298
4.2	0.659	0.293	4.3	0.678	0.304
4.1	0.659	0.301	4.1	0.646	0.297
4.0	0.659	0.301	4.2	0.655	0.301

Chapter 19

More Complex ANOVA Designs

This chapter examines three designs that incorporate more factors and introduce some new elements of experimental design. They are three-way ANOVA, one-way nested ANOVA, and analysis of covariance (ANCOVA). These are common designs whose elements can be combined to generate more elaborate ones. A useful guide to complex ANOVA designs is Winer et al. (1991), who provide a description and statistical model for each design. Once a particular design is identified, the statistical model can be used to program the analysis in SAS or other software.

19.1 Three-way ANOVA

We will first discuss three-way ANOVA, an analysis which examines how three different factors influence the means of the different groups. The three factors may be any combination of fixed or random effects and are typically referred to a Factors A, B, and C. In this design, there are one or more replicate observations for each combination of the three factors. The statistical analysis for three-way ANOVA designs may include F tests for the main effects of the factors as well as the interactions among them. For example, if the design has replication and all three factors are fixed, there are F tests for the main effects (Factor A, B, C), each pairwise interaction (A × B, A × C, B × C), and even a three-way interaction (A × B × C). The additional complexity of this design with its many interactions can make interpretation of the results quite challenging.

As an example of three-way ANOVA, we will analyze data from an experiment by Maestre & Reynolds (2007). This study examined how overall nutrient and water availability, and nutrient heterogeneity, affected grassland biomass production (Table 19.1). Nutrient heterogeneity was manipulated by placing the nitrogen at a particular location within the container vs. an even distribution. See Chapter 14 for further description of this experiment. We will use the notation Y_{ijkl} to reference the observations in three-way ANOVA designs. The i subscript refers to the group or treatment within Factor A (in this case nitrogen heterogeneity), j the treatment within Factor B (nitrogen levels), k the treatment within Factor C (water levels), while l refers to the observation within the treatment. For example, Y_{1134} refers to the fourth observation in the no nutrient heterogeneity, 40 mg N, 375 ml water treatment, which is 7.901.

Table 19.1: Example 1 - Effect of nitrogen heterogeneity, nitrogen availability, and water availability on the total biomass of grassland plants grown in microcosms (Maestre & Reynolds 2007). The table illustrates how the subscripts for Y_{ijkl} vary across treatments for a portion of the data set (see Chapter 22 for the full version).

N het. (Y/N)	N (mg)	Water (ml/week)	Y_{ijkl} = Biomass	i	j	k	l
N	40	125	4.372	1	1	1	1
N	40	125	4.482	1	1	1	2
N	40	125	4.221	1	1	1	3
N	40	125	3.977	1	1	1	4
N	40	250	7.400	1	1	2	1
N	40	250	8.027	1	1	2	2
N	40	250	7.883	1	1	2	3
N	40	250	7.769	1	1	2	4
N	40	375	7.226	1	1	3	1
N	40	375	8.126	1	1	3	2
N	40	375	6.840	1	1	3	3
N	40	375	7.901	1	1	3	4
etc.							
Y	120	250	10.731	2	3	2	1
Y	120	250	12.640	2	3	2	2
Y	120	250	10.350	2	3	2	3
Y	120	250	11.550	2	3	2	4
Y	120	375	14.697	2	3	3	1
Y	120	375	17.826	2	3	3	2
Y	120	375	14.711	2	3	3	3
Y	120	375	13.614	2	3	3	4

19.1.1 Three-way fixed effects model

Suppose that we want to model the observations in a study like Example 1, where there are Factors A, B, and C. Assume the design is factorial with every possible combination of the three factors, with $n > 1$ observations of each one. This design is often called three-way ANOVA with replication. A common model for the observations Y_{ijkl} in such designs (Winer et al. 1991) is

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\beta\gamma)_{jk} + (\alpha\gamma)_{ik} + (\alpha\beta\gamma)_{ijk} + \epsilon_{ijkl}. \quad (19.1)$$

Here μ is the grand mean of the observations, while α_i is the deviation from μ caused by the ith level or treatment of Factor A, β_j the deviation caused by the jth level of Factor B, and γ_k is the deviation caused by the kth level of Factor C. These terms are the **main effects** in the model. The terms $(\alpha\beta)_{ij}$, $(\beta\gamma)_{jk}$, and $(\alpha\gamma)_{ik}$ are **pairwise or first-order interactions** among Factors A and B, B and C, and A and C (A × B, B × C, and A × C). They are similar to the interaction term in two-way ANOVA, but with three factors in the design there are more possibilities for interaction among them. The term $(\alpha\beta\gamma)_{ijk}$ models a **three-way or second-order interaction** (A × B × C) among all three factors. It can be thought of as an interaction of interactions, i.e., the interaction between Factors A and B could change across levels of C. The ϵ_{ijkl} term represents the usual random departures from the mean value predicted by the main effects and interactions due to natural variability.

The objective in three-way ANOVA is to test whether Factor A, B, and C have an effect on the group means, and whether there are interactions among these factors. For Factor A this amounts to testing H_0 : all $\alpha_i = 0$, and similarly H_0 : all $\beta_j = 0$ for Factor B and H_0 : all $\gamma_k = 0$ for Factor C. For the A × B interaction, we would test $H_0 : (\alpha\beta)_{ij} = 0$, and similarly $H_0 : (\alpha\gamma)_{ik} = 0$ for the A × C and $H_0 : (\beta\gamma)_{jk} = 0$ for the B × C interactions. For the three-way interaction, A × B × C, we are interested in testing H_0 : all $(\alpha\beta\gamma)_{ijk} = 0$. The F tests for these hypotheses can be constructed using various sums of squares and mean squares, similar to two-way ANOVA, and are also examples of likelihood ratio tests. We will not consider this process in detail but instead proceed directly to the analysis of the Example 1 data set using SAS.

19.1.2 Three-way ANOVA for Example 1 - SAS demo

The first step in the program (see below) is to read in the observations using a `data` step, with the first variable (`nitrohet`) denoting the nitrogen heterogeneity treatment, while `nitrogen` and `water` represent the nitrogen and water levels. The variable `biomass` is then log-transformed before analysis, yielding the dependent variable `y = log10(biomass}`. Three separate plots then requested using `proc gplot` (SAS Institute Inc. 2016), one for every pairwise combination of `nitrohet`, `nitrogen`, and `water`. These plots will allow us to examine the main effects and all pairwise interactions among the treatments. The choice as to whether a particular treatment is plotted on the x-axis or appears as separate groups (lines) on the graph is arbitrary. Like two-way ANOVA, if the lines are not parallel in a plot this suggests there is an interaction between the factors. The second set of `proc gplot` graphs is intended to illustrate the three-way interaction among the factors. Each plot illustrates the interaction between `nitrogen` and `water` at one level of `nitrohet`. These plots will appear different if there is substantial interaction among the three factors.

The next section of the program conducts the three-way ANOVA using `proc glm` (SAS Institute Inc. 2018). The `class` statement tells SAS that `nitrohet`, `nitrogen`, and `water` are used to classify the observations into the 18 different treatment groups. The `model` statement tells SAS the form of the ANOVA model. Recall that the model for fixed effects three-way ANOVA (Eq. 19.1). The statement `nitrohet|nitrogen|water` is SAS shorthand for this model, and will automatically generate all the possible main effects and interactions of the three factors.

The `lsmeans` statement causes `proc glm` to calculate quantities called least squares means for each level of `nitrohet`, `nitrogen`, and `water`. When the data are balanced these are equivalent to the means for each treatment group, but least squares means have some advantages for unbalanced data and other statistical models. The option `adjust=tukey` requests multiple comparisons among treatments using the Tukey method. This is useful for comparing the different levels of the main effects. However, tests for the main effects as well as multiple comparisons should be treated with caution in the presence of strong interaction (see Chapter 14 for discussion of this issue).

We now examine the results of the tests generated by SAS, examining the interactions first (Fig. 19.6). We are primarily interested in the results for Type III sums of squares. We see that the three-way nitrogen heterogeneity

× nitrogen × water interaction was nonsignificant ($F_{4,54} = 1.39, P = 0.2492$). The two graphs that illustrate this interaction appear similar, further indicating this interaction is weak or absent (Fig. 19.5). Turning to the pairwise interactions, we see that the nitrogen heterogeneity × nitrogen interaction was nonsignificant ($F_{2,54} = 0.93, P = 0.4017$). In agreement with this result, the corresponding graph for this interaction (Fig. 19.2) suggests these two treatments are additive. The nitrogen × water interaction ($F_{4,54} = 12.90, P < 0.0001$) was highly significant. Examining Fig. 19.3, we see that the source of this interaction was a reduced effect of watering at lower nitrogen levels. The nitrogen heterogeneity × water interaction was also highly significant ($F_{2,54} = 13.10, P < 0.0001$). This interaction was apparently generated by a stronger effect of nitrogen heterogeneity at the lowest water level (Fig. 19.4). Overall, the significant interactions suggest that effects of these factors on biomass are not additive (Maestre & Reynolds 2007).

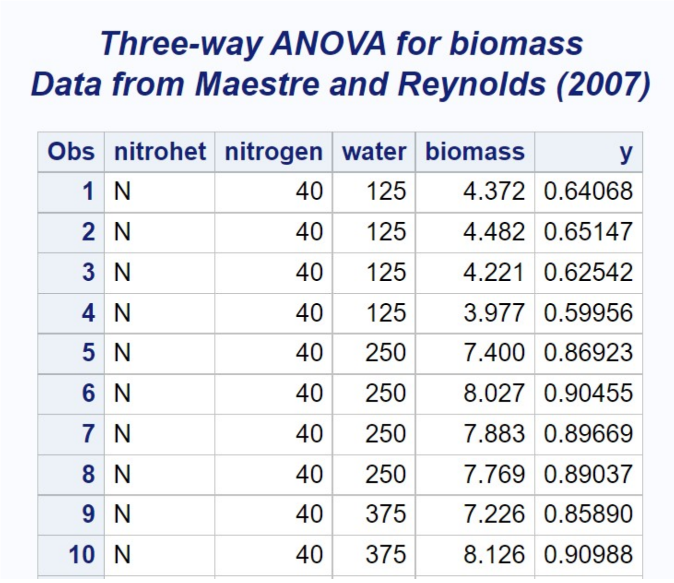
The SAS analysis also found highly significant main effects of nitrogen heterogeneity ($F_{1,54} = 144.14, P < 0.0001$), nitrogen ($F_{2,27} = 129.71, P < 0.0001$) and water ($F_{2,27} = 657.00, P < 0.0001$) on biomass, as well as significant differences among all levels of these treatments (Fig. 19.7). We can judge the strength of these effects through the interaction plots as well as the sum of squares values. Watering appears to have the largest effect on biomass, followed by nitrogen and nitrogen heterogeneity. The heterogeneity result is particularly intriguing, because more biomass was generated when this nutrient was heterogeneously distributed in space. Maestre & Reynolds (2007) suggest this occurred because nutrient patches encourage root proliferation, leading to increased nutrient uptake and overall growth. Even though there were significant interactions in this analysis, the main effects were larger and explained most of the variation in these data.

SAS program

```
* Maestre_biomass_3way.sas;
title "Three-way ANOVA for biomass";
title2 "Data from Maestre and Reynolds (2007)";
data maestre;
    input nitrohet $ nitrogen water biomass;
    * Apply transformations here;
    y = log10(biomass);
    datalines;
N   40  125   4.372
N   40  125   4.482
N   40  125   4.221
N   40  125   3.977
N   40  250   7.400
N   40  250   8.027
N   40  250   7.883
N   40  250   7.769

etc.

Y  120  375  14.697
Y  120  375  17.826
Y  120  375  14.711
Y  120  375  13.614
;
run;
* Print data set;
proc print data=maestre;
run;
proc gplot data=maestre;
    plot y*nitrohet=nitrogen y*nitrogen=water y*nitrohet=water / vaxis=axis1
    haxis=axis1 legend=legend1;
    symbol1 i=std1mjt v=star height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
    legend1 label=(height=2) value=(height=2);
run;
* Sort data by nitrohet levels;
proc sort data=maestre;
    by nitrohet;
run;
* Plots to show three-way interaction;
proc gplot data=maestre;
    by nitrohet;
    plot y*nitrogen=water / vaxis=axis1 haxis=axis1 legend=legend1;
    symbol1 i=std1mjt v=star height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
    legend1 label=(height=2) value=(height=2);
run;
* Three-way ANOVA with all fixed effects;
proc glm plots=diagnostics data=maestre;
    class nitrohet nitrogen water;
    model y = nitrohet|nitrogen|water;
    lsmeans nitrohet nitrogen water / adjust=tukey cl lines;
run;
quit;
```



Three-way ANOVA for biomass
Data from Maestre and Reynolds (2007)

Obs	nitrohet	nitrogen	water	biomass	y
1	N	40	125	4.372	0.64068
2	N	40	125	4.482	0.65147
3	N	40	125	4.221	0.62542
4	N	40	125	3.977	0.59956
5	N	40	250	7.400	0.86923
6	N	40	250	8.027	0.90455
7	N	40	250	7.883	0.89669
8	N	40	250	7.769	0.89037
9	N	40	375	7.226	0.85890
10	N	40	375	8.126	0.90988

etc.

Figure 19.1: `Maestre_biomass_3way.sas - proc print`

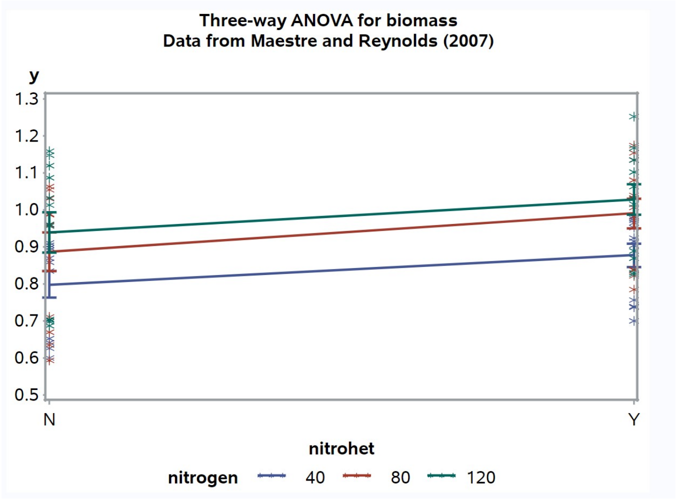


Figure 19.2: `Maestre_biomass_3way.sas - proc gplot`

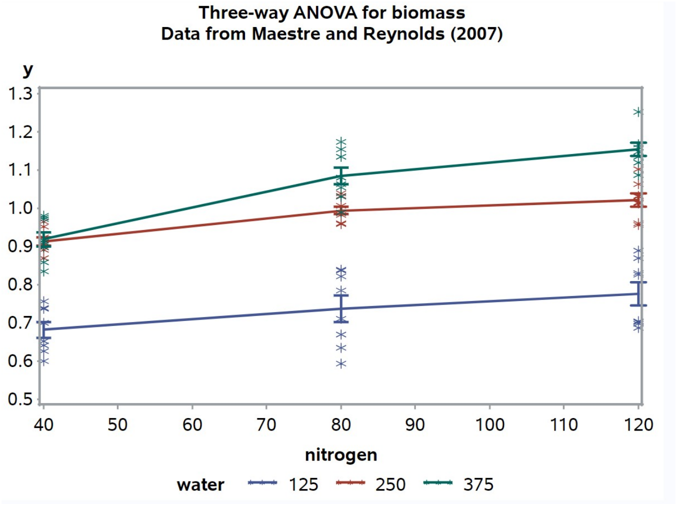


Figure 19.3: `Maestre_biomass_3way.sas - proc gplot`

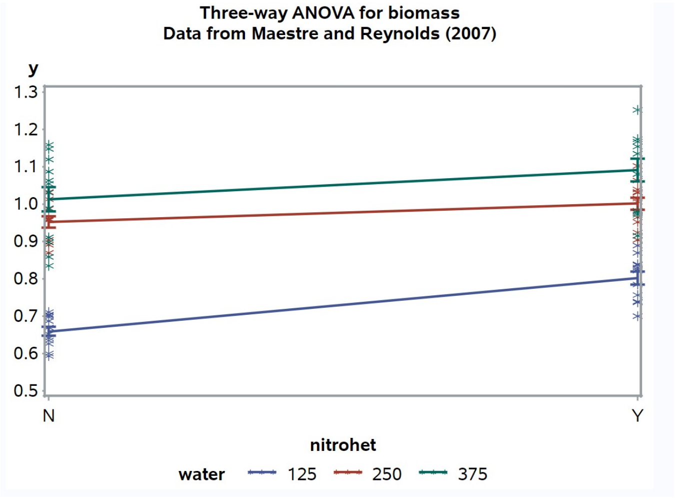
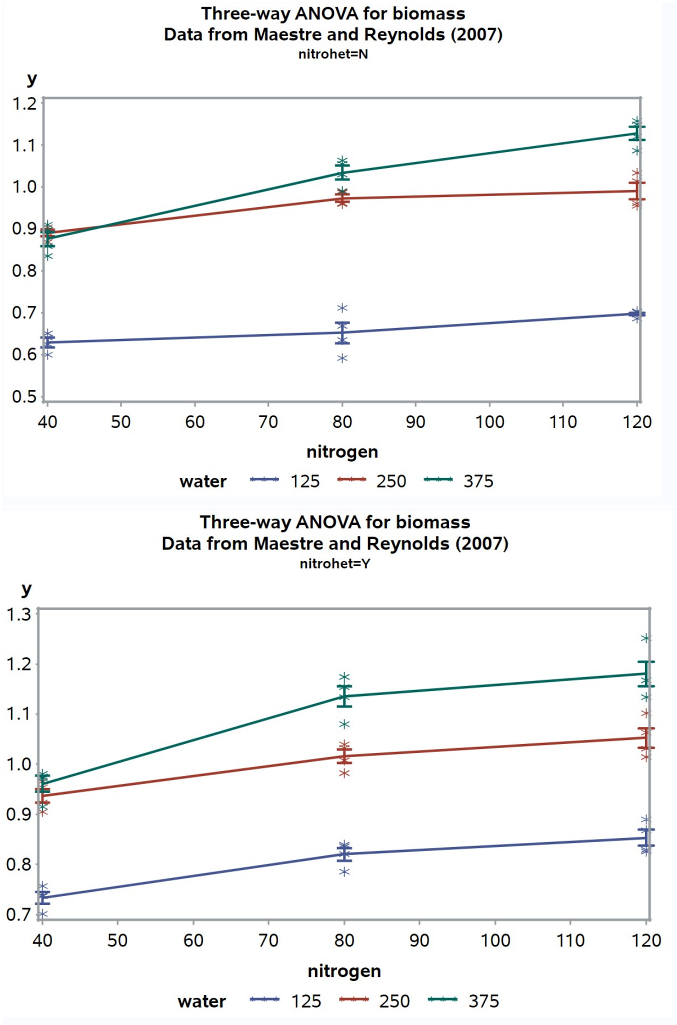


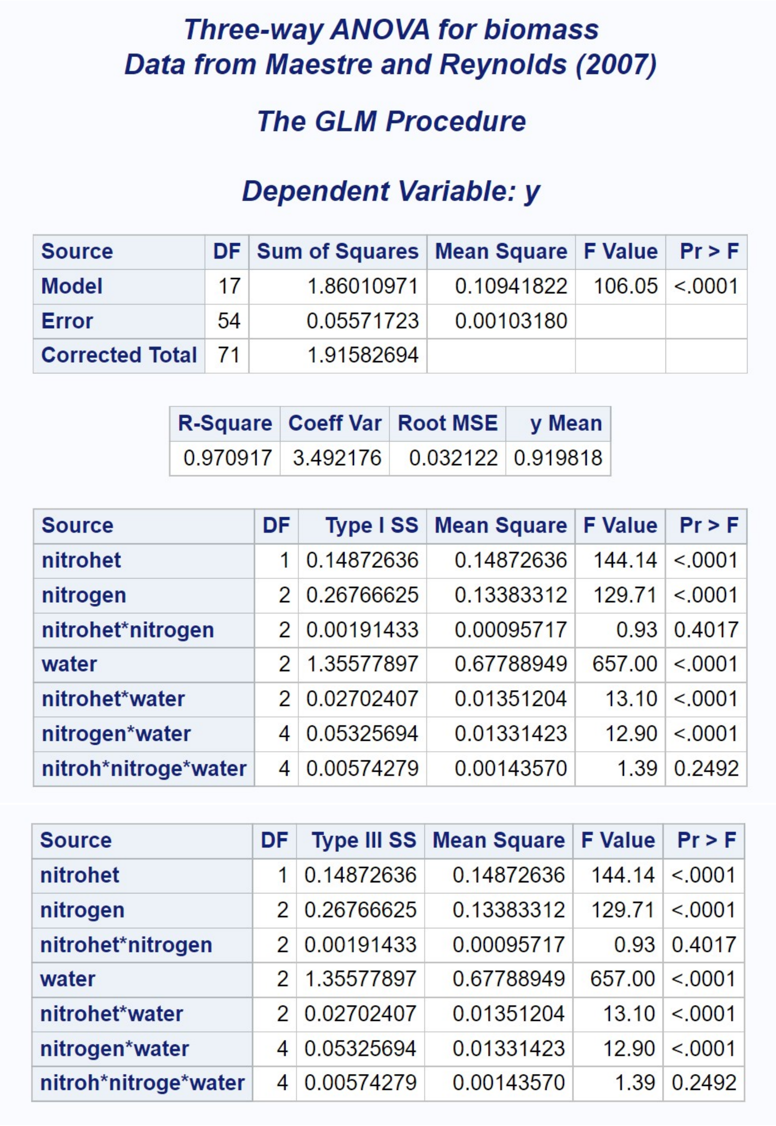
Figure 19.4: `Maestre_biomass_3way.sas - proc gplot`



Three-way ANOVA for biomass
Data from Maestre and Reynolds (2007)
nitrohet=N

Three-way ANOVA for biomass
Data from Maestre and Reynolds (2007)
nitrohet=Y

Figure 19.5: `Maestre_biomass_3way.sas - proc gplot`



Three-way ANOVA for biomass
Data from Maestre and Reynolds (2007)

The GLM Procedure

Dependent Variable: y

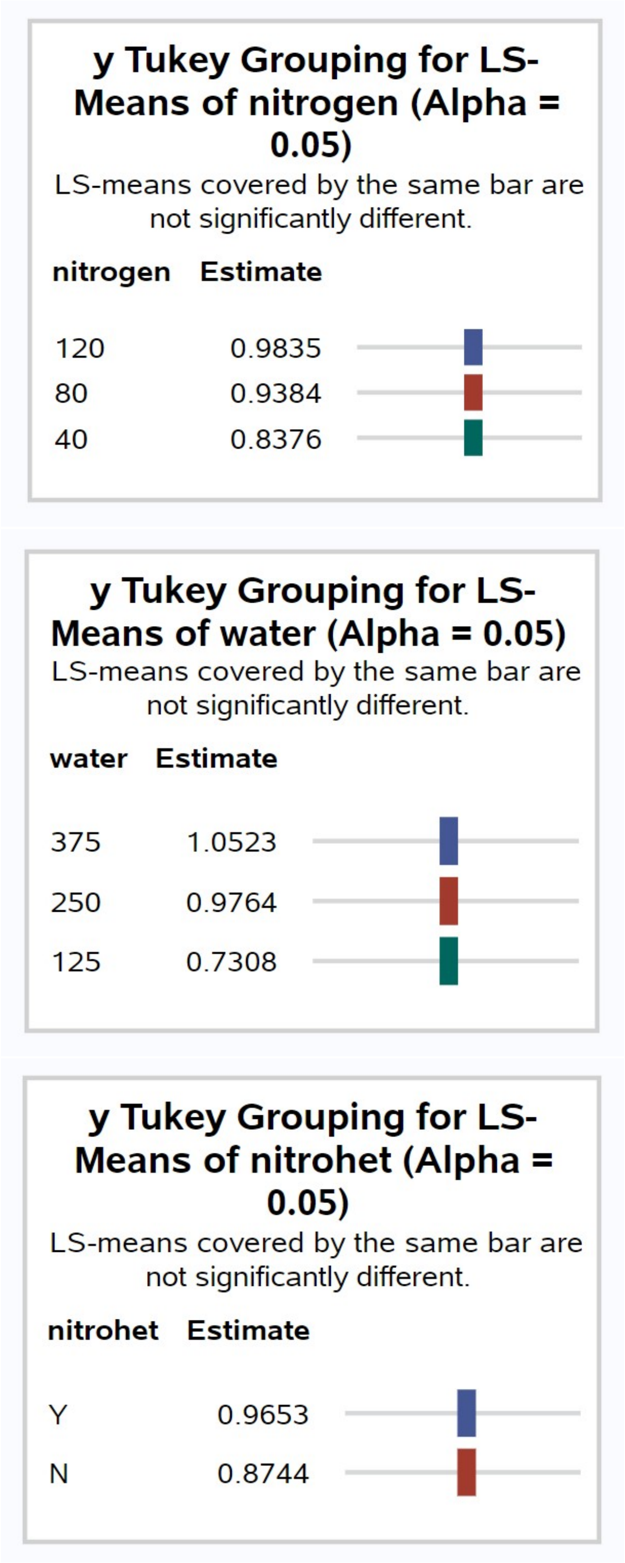
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	17	1.86010971	0.10941822	106.05	<.0001
Error	54	0.05571723	0.00103180		
Corrected Total	71	1.91582694			

R-Square	Coeff Var	Root MSE	y Mean
0.970917	3.492176	0.032122	0.919818

Source	DF	Type I SS	Mean Square	F Value	Pr > F
nitrohet	1	0.14872636	0.14872636	144.14	<.0001
nitrogen	2	0.26766625	0.13383312	129.71	<.0001
nitrohet*nitrogen	2	0.00191433	0.00095717	0.93	0.4017
water	2	1.35577897	0.67788949	657.00	<.0001
nitrohet*water	2	0.02702407	0.01351204	13.10	<.0001
nitrogen*water	4	0.05325694	0.01331423	12.90	<.0001
nitroh*nitroge*water	4	0.00574279	0.00143570	1.39	0.2492

Source	DF	Type III SS	Mean Square	F Value	Pr > F
nitrohet	1	0.14872636	0.14872636	144.14	<.0001
nitrogen	2	0.26766625	0.13383312	129.71	<.0001
nitrohet*nitrogen	2	0.00191433	0.00095717	0.93	0.4017
water	2	1.35577897	0.67788949	657.00	<.0001
nitrohet*water	2	0.02702407	0.01351204	13.10	<.0001
nitrogen*water	4	0.05325694	0.01331423	12.90	<.0001
nitroh*nitroge*water	4	0.00574279	0.00143570	1.39	0.2492

Figure 19.6: `Maestre_biomass_3way.sas - proc glm`



y Tukey Grouping for LS-Means of nitrogen (Alpha = 0.05)

LS-means covered by the same bar are not significantly different.

nitrogen	Estimate
120	0.9835
80	0.9384
40	0.8376

y Tukey Grouping for LS-Means of water (Alpha = 0.05)

LS-means covered by the same bar are not significantly different.

water	Estimate
375	1.0523
250	0.9764
125	0.7308

y Tukey Grouping for LS-Means of nitrohet (Alpha = 0.05)

LS-means covered by the same bar are not significantly different.

nitrohet	Estimate
Y	0.9653
N	0.8744

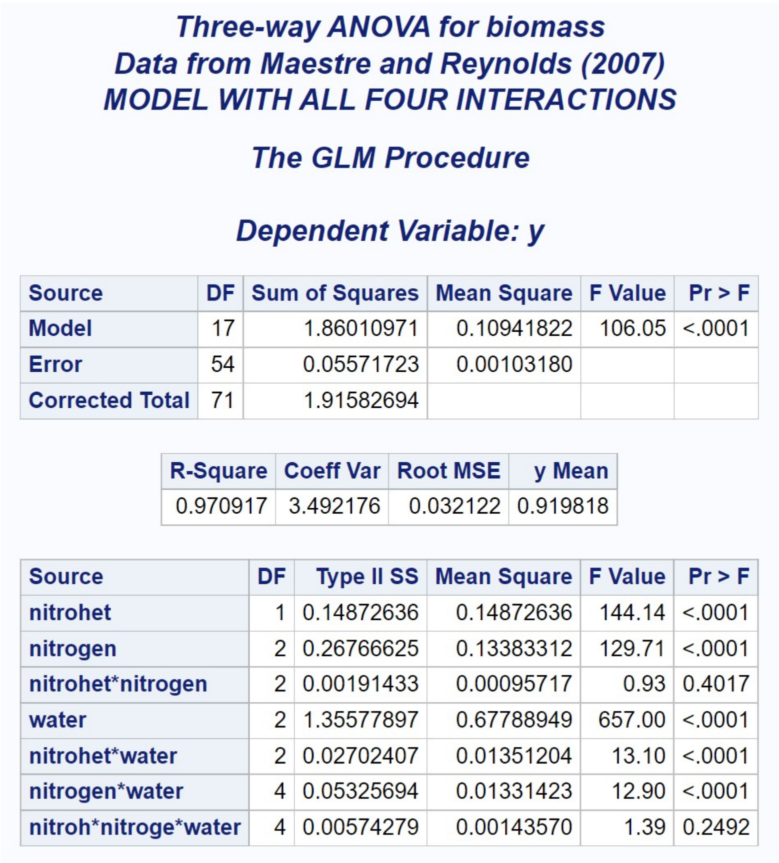
Figure 19.7: `Maestre_biomass_3way.sas - proc glm`

19.1.3 Tests for main effects with interaction

As discussed in Chapter 14, there are questions as to whether tests of main effects are appropriate when interaction is significant, and these extend to three-way designs. As an alternative, we can use the `slice` option for `lsmeans` to avoid tests of the main effects. The modified SAS code is listed below along with the output. We first fit the full model including all the interactions (see Fig. 19.8), and observe that the nitrogen heterogeneity × nitrogen × water interaction was nonsignificant ($F_{4,54} = 1.39, P = 0.2492$), as was the nitrogen heterogeneity × nitrogen interaction ($F_{2,54} = 0.93, P = 0.4017$). We then drop these interactions and refit the model (Fig. 19.9). The remaining two interactions were both highly significant in this reduced model (nitrogen heterogeneity × water, $F_{2,60} = 12.79, P < 0.0001$; nitrogen × water, $F_{4,60} = 12.61, P < 0.0001$). We skip the tests of the main effects because of these highly significant interactions, and instead use the `slice` option to test for a nitrogen heterogeneity effect at each water level, and vice versa. These tests were all highly significant, suggesting that nitrogen heterogeneity affected biomass at every water level, and water affected biomass at every nitrogen heterogeneity level (Fig. 19.10). Similar tests could be conducted to examine the effects of nitrogen and water.

SAS Program

```
* Three-way ANOVA with interaction;
title3 "MODEL WITH ALL FOUR INTERACTIONS";
proc glm data=maestre;
    class nitrohet nitrogen water;
    model y = nitrohet|nitrogen|water / ss2;
    output out=resids p=pred r=resid;
run;
* Three-way ANOVA dropping ns interactions;
title3 "MODEL WITH ONLY SIGNIFICANT INTERACTIONS";
proc glm data=maestre;
    class nitrohet nitrogen water;
    model y = nitrohet nitrogen water nitrohet*water nitrogen*water / ss2;
    lsmeans nitrohet*water / slice=water slice=nitrohet;
run;
```



Three-way ANOVA for biomass
Data from Maestre and Reynolds (2007)
MODEL WITH ALL FOUR INTERACTIONS

The GLM Procedure

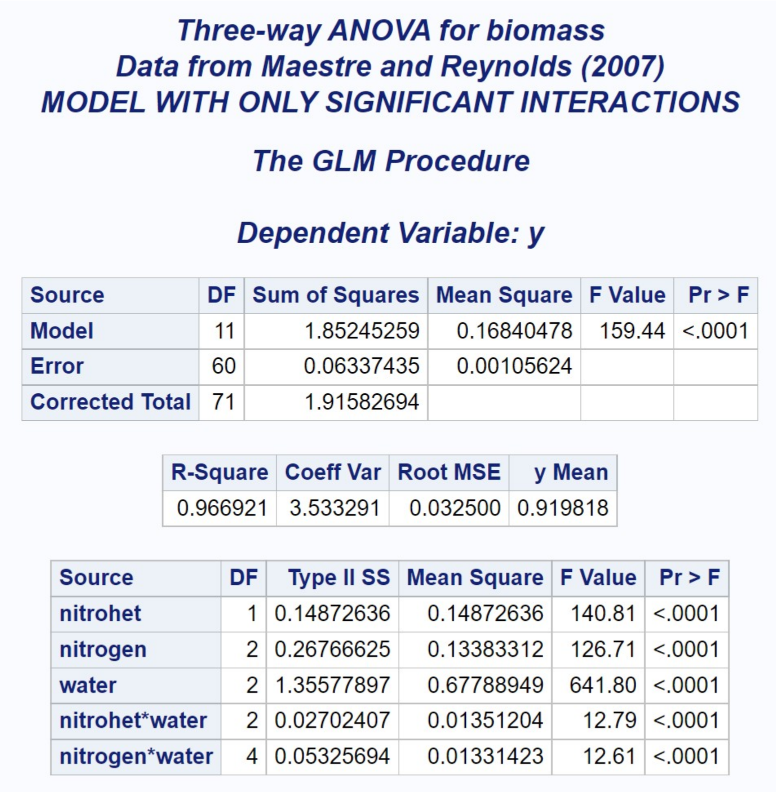
Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	17	1.86010971	0.10941822	106.05	<.0001
Error	54	0.05571723	0.00103180		
Corrected Total	71	1.91582694			

R-Square	Coeff Var	Root MSE	y Mean
0.970917	3.492176	0.032122	0.919818

Source	DF	Type II SS	Mean Square	F Value	Pr > F
nitrohet	1	0.14872636	0.14872636	144.14	<.0001
nitrogen	2	0.26766625	0.13383312	129.71	<.0001
nitrohet*nitrogen	2	0.00191433	0.00095717	0.93	0.4017
water	2	1.35577897	0.67788949	657.00	<.0001
nitrohet*water	2	0.02702407	0.01351204	13.10	<.0001
nitrogen*water	4	0.05325694	0.01331423	12.90	<.0001
nitroh*nitroge*water	4	0.00574279	0.00143570	1.39	0.2492

Figure 19.8: `Maestre_biomass_3way_new.sas - proc glm` (1)



Three-way ANOVA for biomass
Data from Maestre and Reynolds (2007)
MODEL WITH ONLY SIGNIFICANT INTERACTIONS

The GLM Procedure

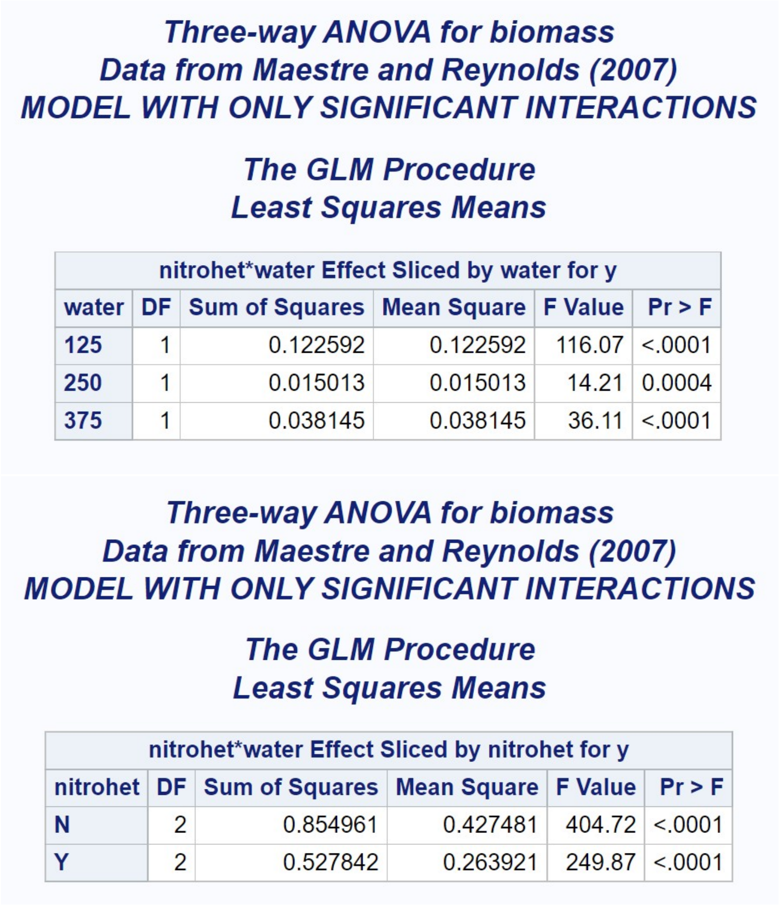
Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	1.85245259	0.16840478	159.44	<.0001
Error	60	0.06337435	0.00105624		
Corrected Total	71	1.91582694			

R-Square	Coeff Var	Root MSE	y Mean
0.966921	3.533291	0.032500	0.919818

Source	DF	Type II SS	Mean Square	F Value	Pr > F
nitrohet	1	0.14872636	0.14872636	140.81	<.0001
nitrogen	2	0.26766625	0.13383312	126.71	<.0001
water	2	1.35577897	0.67788949	641.80	<.0001
nitrohet*water	2	0.02702407	0.01351204	12.79	<.0001
nitrogen*water	4	0.05325694	0.01331423	12.61	<.0001

Figure 19.9: `Maestre_biomass_3way_new.sas - proc glm` (2)



Three-way ANOVA for biomass
Data from Maestre and Reynolds (2007)
MODEL WITH ONLY SIGNIFICANT INTERACTIONS

The GLM Procedure
Least Squares Means

nitrohet*water Effect Sliced by water for y					
water	DF	Sum of Squares	Mean Square	F Value	Pr > F
125	1	0.122592	0.122592	116.07	<.0001
250	1	0.015013	0.015013	14.21	0.0004
375	1	0.038145	0.038145	36.11	<.0001

Three-way ANOVA for biomass
Data from Maestre and Reynolds (2007)
MODEL WITH ONLY SIGNIFICANT INTERACTIONS

The GLM Procedure
Least Squares Means

nitrohet*water Effect Sliced by nitrohet for y					
nitrohet	DF	Sum of Squares	Mean Square	F Value	Pr > F
N	2	0.854961	0.427481	404.72	<.0001
Y	2	0.527842	0.263921	249.87	<.0001

Figure 19.10: `Maestre_biomass_3way_new.sas - proc glm` (2)

19.1.4 Other three-way designs

The Maestre & Reynolds (2007) experiment had four replicate containers for each treatment combination ($n = 4$), and so it was possible to fit a model with a three-way interaction, namely nitrogen heterogeneity × nitrogen × water. Suppose now there was only observation for each treatment combination ($n = 1$). It is still possible to analyze these data using three-way ANOVA, but the data are not sufficient to fit a model with a three-way interaction. We would therefore use the model

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\beta\gamma)_{jk} + (\alpha\gamma)_{ik} + \epsilon_{ijk}. \tag{19.2}$$

The equivalent model statement for `proc glm` would be

```
model y = nitrohet nitrogen water nitrohet*nitrogen nitrohet*water
nitrogen*water;
```

There is no shorthand method of specifying this model. The SAS output would be interpreted in the same way as the model with replication, except there would be no test for a three-way interaction.

Another common three-way design could have one or more factors that are random effects. For example, suppose that one manipulated nitrogen and water levels similar to Maestre & Reynolds (2007) but conducted the experiment in three different blocks, either different locations in the greenhouse or points in time. Block could be a random effect in this design, and the corresponding model would be

$$Y_{ijkl} = \mu + \alpha_i + \beta_j + C_k + (\alpha\beta)_{ij} + (\beta C)_{jk} + (\alpha C)_{ik} + (\alpha\beta C)_{ijk} + \epsilon_{ijkl}. \tag{19.3}$$

Here C stands for a random block effect, with $C \sim N(0, \sigma^2_C)$. Note that every interaction term involving C is also considered a random effect. This model could be analyzed with `proc mixed` (SAS Institute Inc. 2018) using the following SAS statements:

```
proc mixed cl;
    class nitrogen water block;
    model y = nitrogen water nitrogen*water / ddfm=kr;
    random block block*nitrogen block*water block*nitrogen*water;
run;
```

19.2 One-way nested ANOVA

The second design we will examine are called one-way nested designs. There are two factors in this design, a Factor A that may be a fixed or random effect, and a random nested Factor B. Nested means that for each level of Factor A, there are several levels of Factor B that are unique to that level of A. There are several replicate observations for each combination of Factor A and B.

As an example of this design, we will examine a genetic study of a minute parasitic wasp, *Anagrus delicatus* (Hymenoptera: Mymaridae). This wasp attacks eggs of the planthopper *Prokelisia marginata* (Homoptera: Delphacidae), a salt marsh insect that feeds on *Spartina* plants. Cronin & Strong (1996) were interested in the genetics of various wasp traits, including the number of eggs carried by the wasps themselves, ovipositor length, and various behavioral traits. They collected female wasps from three separate sites in San Francisco Bay and established genetically identical isolines from individual wasps collected from each site. They then measured the traits for a number of individuals from each isoline. Isolines are the nested factor in this design, because each isoline was established from a single site. Sites were classified as a fixed effect because there were essentially only three sites available for sampling, and so the sites were not randomly selected from a population of sites. Example 2 below shows a simulated data set based on this study, with three sites, 14 isolines per site, and eight individuals per isoline.

Table 19.2: Example 2 - Fecundity for *Anagrus delicatus* collected from three different sites, with 14 isolines per site and eight wasps per isoline. The data were simulated from results presented in Cronin and Strong (1996). Note that the values in the site, isoline, and wasp columns also correspond to the subscripts for Y_{ijk}. See Chapter 22 for the full version of this data set.

Site	Isoline	Wasp	Y_{ijk} = eggs
1	1	1	37
1	1	2	41
1	1	3	46
1	1	4	44
1	1	5	43
1	1	6	41
1	1	7	38
1	1	8	37
1	2	1	37
1	2	2	28
1	2	3	34
1	2	4	37
1	2	5	35
1	2	6	39
1	2	7	36
etc.			
3	13	1	36
3	13	2	39
3	13	3	36
3	13	4	30
3	13	5	37
3	13	6	32
3	13	7	38
3	13	8	39
3	14	1	32
3	14	2	34
3	14	3	41
3	14	4	33
3	14	5	35
3	14	6	35
3	14	7	34
3	14	8	31

19.2.1 Nested ANOVA models

Suppose that we want to model the observations in a study like Example 2, where there is a fixed Factor A and a nested Factor B. A common model for the observations Y_{ijk} in such designs (Winer et al. 1991) is

$$Y_{ijk} = \mu + \alpha_i + B_{j(i)} + \epsilon_{ijk}. \tag{19.4}$$

Here μ is the grand mean of the observations, α_i the deviation from μ caused by the *ith* level or treatment of Factor A, and $B_{j(i)}$ the random deviation caused by the *jth* level of Factor B nested within the *ith* level of Factor A. $B_{j(i)}$ is assumed to be normally distributed with mean zero and variance $\sigma^2_{B(A)}$, or $B_{j(i)} \sim N(0, \sigma^2_{B(A)})$, while $\epsilon_{ijk} \sim N(0, \sigma^2)$ as usual. $B_{j(i)}$ and ϵ_{ijk} are assumed to be independent. This model has two variance components, namely $\sigma^2_{B(A)}$ and σ^2.

The behavior of this model is illustrated in Fig. 19.11, for $a = 3$ levels of Factor A and $b = 4$ levels of Factor B nested within each A. The figure illustrates how the value of α_i shifts the mean of the observations away from μ, similar to other ANOVA models. The $B_{j(i)}$ values, which are random variables, shift the observations for each nested level away from the values set by $\mu + \alpha_i$. The values of $B_{j(i)}$ are different for each level of Factor A because they are random quantities.

The usual objectives for this nested ANOVA design are to test for Factor A effects, and estimate the variance components $\sigma^2_{B(A)}$ and σ^2. For Factor A, this amounts to testing H_0 : all $\alpha_i = 0$. We will not consider this process in detail but proceed to the analysis and interpretation of the Example 2 data set. We will use `proc mixed` for the analysis because this design involves a mixed model.

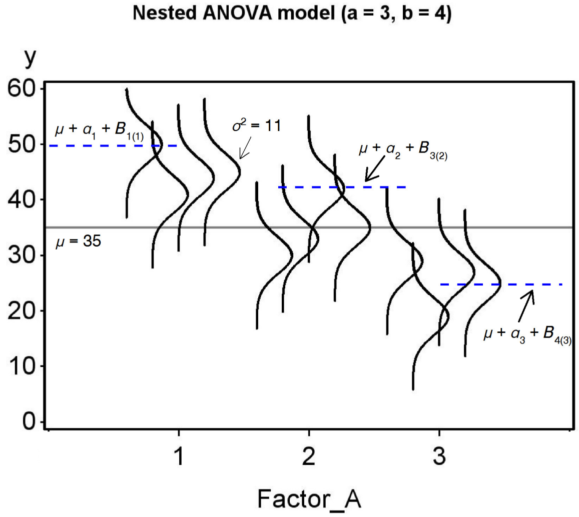


Figure 19.11: Mixed model for nested ANOVA showing the Factor A and B effects.

19.2.2 Nested ANOVA for Example 2 - SAS demo

The first step in analyzing the Example 2 data is to read the observations using a `data` step, with the variables `site` and `isoline` denoting the collection site and *Anagrus* isoline (see program below), while the dependent variable is `eggs`. Although the isolines are numbered similarly across the three sites, note they are actually unique to each site and so are nested within sites. The variable `wasp` refers to a particular wasp within each isoline, but is not used in the analyses. Two plots are then requested using `proc gplot` (SAS Institute Inc. 2016), one showing the mean for each site and so illustrating the site effect. The second plot shows the individual wasps color-coded by isoline, allowing for a visual comparison of variation among and within isolines. The x-axis position of each wasp is jittered to keep the points from overlapping. This involves adding a small random quantity to the `site` value, generating a new variable called `site_jit` that differs for each wasp.

The next section of the program conducts the nested ANOVA using `proc mixed` (SAS Institute Inc. 2018). The `class` statement tells SAS that `site` and `isoline` are used to classify the observations. Next, the fixed effect site is listed in the `model` statement, while the random, nested effect of isoline is incorporated in the `random` statement. SAS uses the syntax `isoline(site)` to indicate that isoline is nested within site. An `lsmeans` statement is used to compared the different sites using the Tukey method.

There appears to be little difference among the sites in the mean number of eggs per wasp (Fig. 19.13), and the test of the site effect was non-significant ($F_{2,39} = 2.3, P = 0.1323$) (Fig. 19.16). We next look at the estimates of the variance components. The variance among isolines within sites ($\hat{\sigma}^2_{B(A)} = \hat{\sigma}^2_{\text{isoline(site)}} = 10.17$) was substantial relative to the variance among wasps within isolines ($\hat{\sigma}^2 = 11.02$). This pattern can be observed in Fig. 19.14, with the observations for each isoline falling into discernable groups.

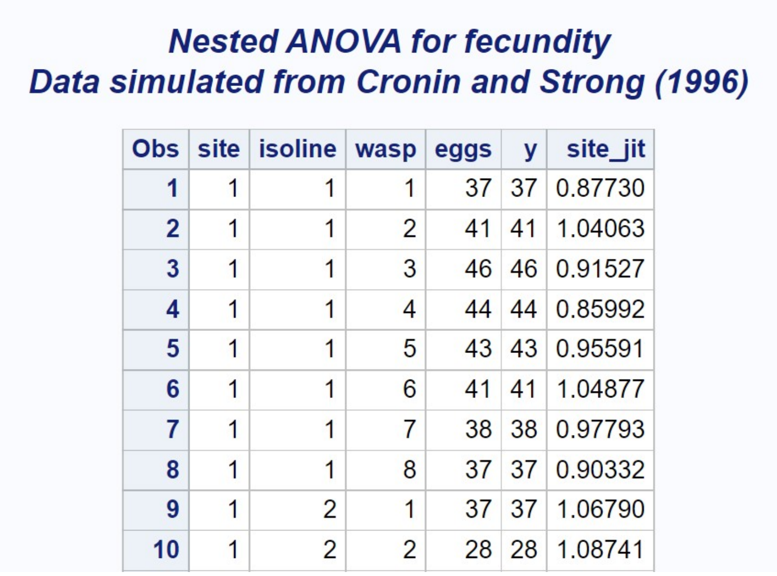
We can use the two variance components to estimate the heritability of egg number, which is the proportion of the variance due to genotypic vs. phenotypic differences among individuals (Falconer & Mackay 1996). The genotypic variance, V_G, is estimated by the variance among isolines within sites, because each isoline represents a different genetic group. For the wasp example, we have $V_G = \hat{\sigma}^2_{\text{isoline(site)}} = 10.17$. The environmental variance, V_E, is estimated by the variance among individuals within isolines, and represents variation among individuals not due to genotype. It is estimated by the variance among wasps within isolines, or $V_E = \hat{\sigma}^2 = 11.02$. The phenotypic variance is defined as the sum of the genotypic and environmental variance, or $V_P = V_G + V_E$. Heritability is then defined $h^2 = V_G/V_P = V_G/(V_G + V_E)$. It follows that $h^2 = 10.17/(10.17 + 11.02) = 0.48$ for the number of eggs in the wasps. This is relatively large value, suggesting that egg number could readily evolve in response to selection pressure.

SAS program

```
* Nested_ANOVA_Anagrus.sas;
title "Nested ANOVA for fecundity";
title2 "Data simulated from Cronin and Strong (1996)";
data anagrus;
    input site isoline wasp eggs;
    * Apply transformations here;
    y = eggs;
    * Make jittered data for plots;
    site_jit = site + 0.1*rannor(0);
    datalines;
1   1  1  37
1   1  2  41
1   1  3  46
1   1  4  44
1   1  5  43
1   1  6  41
1   1  7  38
1   1  8  37
1   2  1  37
1   2  2  28

etc.

;
run;
* Print data set;
proc print data=anagrus;
run;
* Plot means and standard errors for each site;
proc gplot data=anagrus;
    plot y*site=1 / vaxis=axis1 haxis=axis1;
    symbol1 i=std1jmt v=none height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* Plot observations for each site and isoline;
proc gplot data=anagrus;
    plot y*site_jit=isoline / vaxis=axis1 haxis=axis1;
    symbol1 i=none v=dot height=0.5;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* Nested ANOVA mixed model;
proc mixed cl plots=residualpanel data=anagrus;
    class site isoline;
    model y = site / ddfm=kr;
    random isoline(site);
    * Compare levels of fixed effect using Tukey's HSD;
    lsmeans site / diff=all adjust=tukey cl adjdfe=row;
run;
quit;
```



Nested ANOVA for fecundity
Data simulated from Cronin and Strong (1996)

Obs	site	isoline	wasp	eggs	y	site_jit
1	1	1	1	37	37	0.87730
2	1	1	2	41	41	1.04063
3	1	1	3	46	46	0.91527
4	1	1	4	44	44	0.85992
5	1	1	5	43	43	0.95591
6	1	1	6	41	41	1.04877
7	1	1	7	38	38	0.97793
8	1	1	8	37	37	0.90332
9	1	2	1	37	37	1.06790
10	1	2	2	28	28	1.08741

etc.

Figure 19.12: `nested_ANOVA_Anagrus.sas - proc print`

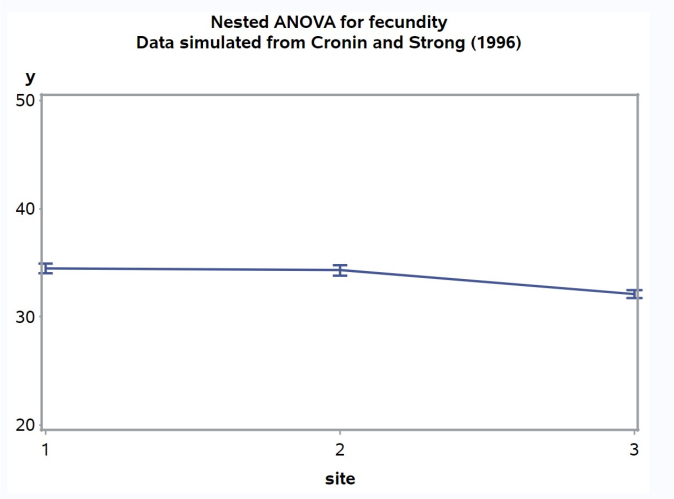


Figure 19.13: `nested_ANOVA_Anagrus.sas - proc gplot` (1)

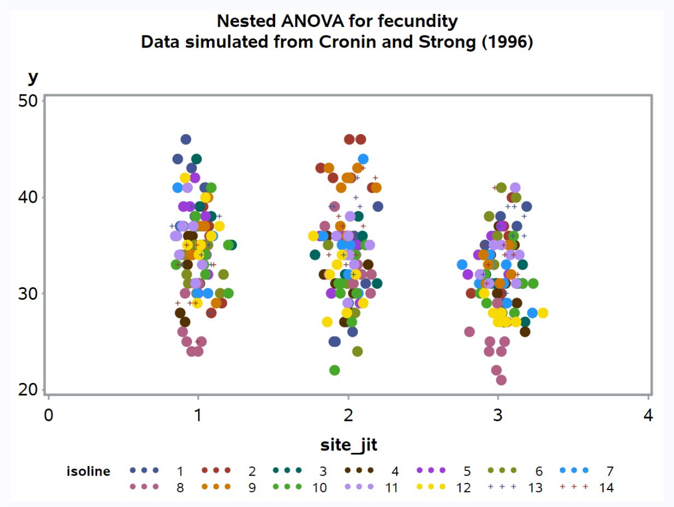
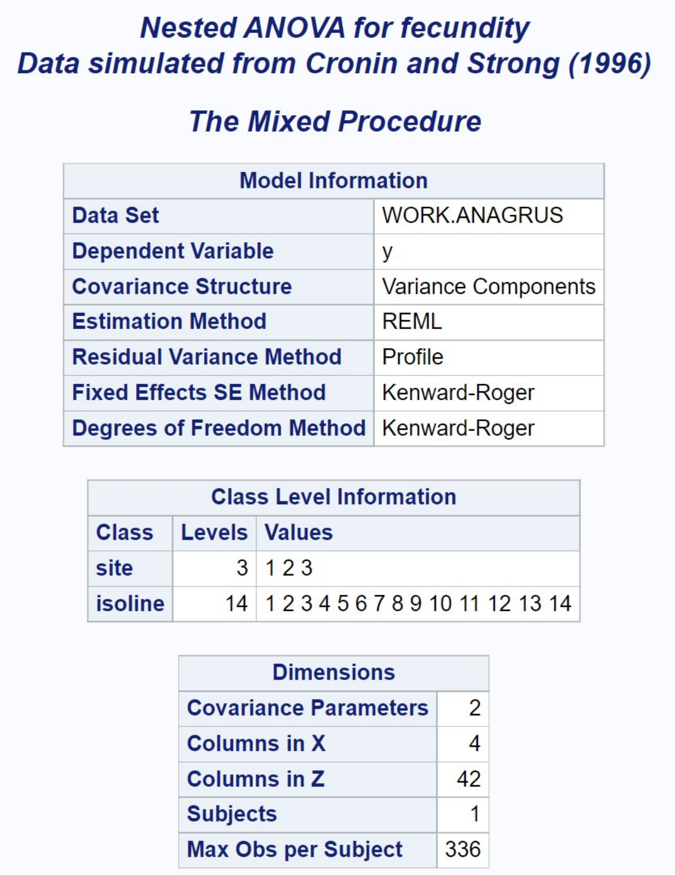


Figure 19.14: `nested_ANOVA_Anagrus.sas - proc gplot` (2)



Nested ANOVA for fecundity
Data simulated from Cronin and Strong (1996)

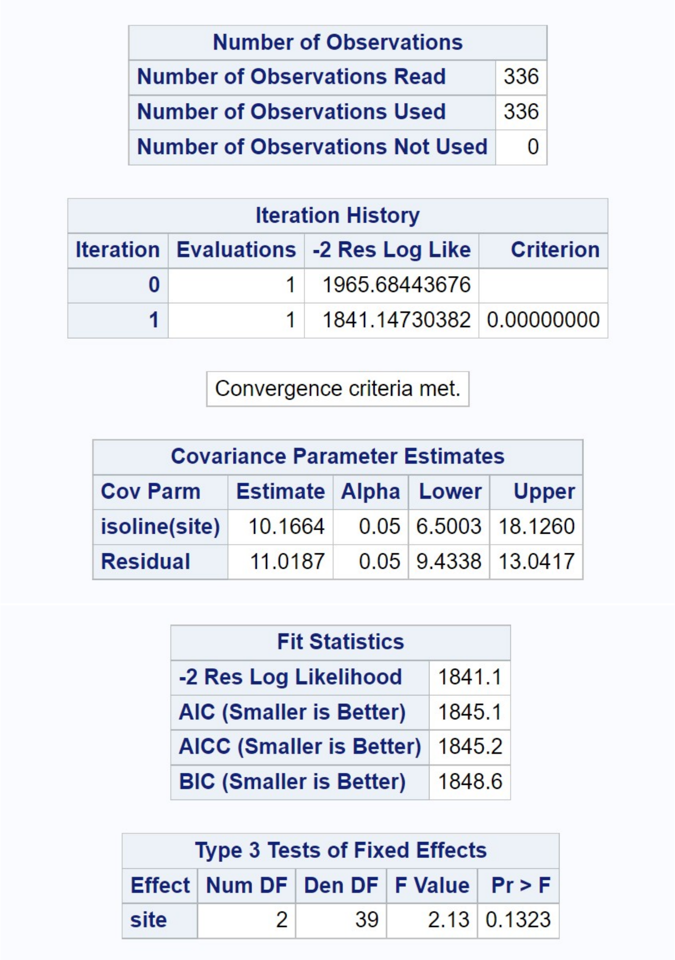
The Mixed Procedure

Model Information	
Data Set	WORK.ANAGRUS
Dependent Variable	y
Covariance Structure	Variance Components
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Kenward-Roger
Degrees of Freedom Method	Kenward-Roger

Class Level Information		
Class	Levels	Values
site	3	1 2 3
isoline	14	1 2 3 4 5 6 7 8 9 10 11 12 13 14

Dimensions	
Covariance Parameters	2
Columns in X	4
Columns in Z	42
Subjects	1
Max Obs per Subject	336

Figure 19.15: `nested_ANOVA_Anagrus.sas - proc mixed`



Number of Observations	
Number of Observations Read	336
Number of Observations Used	336
Number of Observations Not Used	0

Iteration History			
Iteration	Evaluations	-2 Res Log Like	Criterion
0	1	1965.68443676	
1	1	1841.14730382	0.00000000

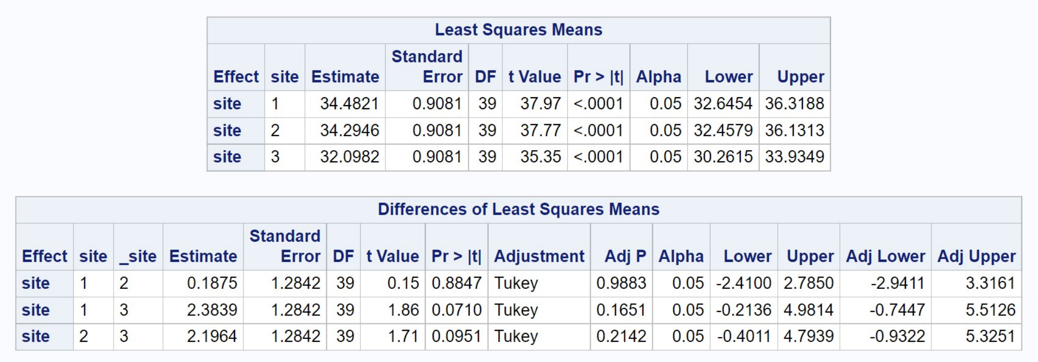
Convergence criteria met.

Covariance Parameter Estimates				
Cov Parm	Estimate	Alpha	Lower	Upper
isoline(site)	10.1664	0.05	6.5003	18.1260
Residual	11.0187	0.05	9.4338	13.0417

Fit Statistics	
-2 Res Log Likelihood	1841.1
AIC (Smaller is Better)	1845.1
AICC (Smaller is Better)	1845.2
BIC (Smaller is Better)	1848.6

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
site	2	39	2.13	0.1323

Figure 19.16: `nested_ANOVA_Anagrus.sas - proc mixed`



Least Squares Means

Effect	site	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
site	1	34.4821	0.9081	39	37.97	<.0001	0.05	32.6454	36.3188
site	2	34.2946	0.9081	39	37.77	<.0001	0.05	32.4579	36.1313
site	3	32.0982	0.9081	39	35.35	<.0001	0.05	30.2615	33.9349

Differences of Least Squares Means

Effect	site	_site	Estimate	Standard Error	DF	t Value	Pr > |t|	Adjustment	Adj P	Alpha	Lower	Upper	Adj Lower	Adj Upper
site	1	2	0.1875	1.2842	39	0.15	0.8847	Tukey	0.9883	0.05	-2.4100	2.7850	-2.9411	3.3161
site	1	3	2.3839	1.2842	39	1.86	0.0710	Tukey	0.1651	0.05	-0.2136	4.9814	-0.7447	5.5126
site	2	3	2.1964	1.2842	39	1.71	0.0951	Tukey	0.2142	0.05	-0.4011	4.7939	-0.9322	5.3251

Figure 19.17: `nested_ANOVA_Anagrus.sas - proc mixed`

19.3 Analysis of covariance

Analysis of covariance, or ANCOVA, is a design that combines elements of ANOVA and regression. The simplest ANCOVA design is a combination of one-way ANOVA and linear regression. The Factor A in the design is typically a fixed effect, such as a treatment. For each observation Y in a given treatment, a **covariate** X is also measured. This covariate is thought to explain some level of variation in Y, and including it in the analysis could increase the power to detect treatment effects. Y is often assumed to be linearly related to X, although nonlinear relationships can be accomodated. More generally, a study might involve a mixture of factors and covariates, and the covariate effects may be of equal or greater interest than the factors.

As an example of ANCOVA, we will analyze a study of the fitness of adult *Thanasimus dubius*, a bark beetle predator, reared on an artificial diet vs. individuals collected from the wild (Reeve et al. 2003). The fitness variables measured were the total number of eggs laid (fecundity) and elytral length (Table 19.3). Body size and fecundity are often related in insects, so elytral length was used as a covariate in the analysis. This helps control for natural variation in body size to better see the treatment effect. The three treatments in the study were (1) artificial diet as larvae and *Ips grandicollis* (a bark beetle) as adults (`DietIG`), (2) artificial diet as larvae and cowpea weevils (a substitute prey) as adults (`DietCPW`), and (3) wild adults fed cowpea weevils (`WildCPW`). The wild adults were collected from the field and so reared on natural prey as larvae. We will use the notation Y_{ij} to reference the observations in ANCOVA designs, with the i subscript refering to the Factor A or treatment group, while j is the observation within the treatment.

Table 19.3: Example 3 - Fitness of the predator *T. dubius*, reared on an artificial diet as larvae vs. wild individuals collected from the field (Reeve et al. 2003). See Chapter 22 for the full data set.

Y_{ij} = Eggs	X_{ij} = Length (mm)	Treatment	i	j
290	5.7	DietIG	1	1
99	5.2	DietIG	1	2
340	5.5	DietIG	1	3
271	4.8	DietIG	1	4
200	5.2	DietIG	1	5
etc.				
66	4.6	DietCPW	2	1
93	5.0	DietCPW	2	2
9	5.4	DietCPW	2	3
404	5.4	DietCPW	2	4
244	5.1	DietCPW	2	5
etc.				
62	4.7	WildCPW	3	1
290	5.0	WildCPW	3	2
488	5.8	WildCPW	3	3
336	5.2	WildCPW	3	4
337	5.8	WildCPW	3	5
etc.				

19.3.1 ANCOVA model

The following model is commonly used for simple ANCOVA designs (Winer et al. 1991). We have

$$Y_{ij} = \mu + \alpha_i + \beta(X_{ij} - \bar{X}) + \epsilon_{ij}, \tag{19.5}$$

where μ is the grand mean and α_i is the deviation from μ caused by the ith level of Factor A. The term X_{ij} is the value of the covariate for observation Y_{ij}, while $\bar{X}$ is the average of all the covariate values. The parameter β is the slope of the relationship between Y_{ij} and X_{ij}. This slope is assumed to be the same across all levels of Factor A. We will later see how to test this assumption. As usual, the model assumes $\epsilon_{ij} \sim N(0, \sigma^2)$.

The model can also be written in the form

$$Y'_{ij} = Y_{ij} - \beta(X_{ij} - \bar{X}) = \mu + \alpha_i + \epsilon_{ij}. \tag{19.6}$$

Displayed this way, we can see that ANCOVA is equivalent to carrying out a one-way ANOVA on values of Y_{ij} that have been adjusted for the covariate X, namely the values of Y'_{ij}.

Another adjustment of the model is needed by SAS and other statistical software. Combining some elements, the model can be written as

$$Y_{ij} = \mu' + \alpha_i + \beta X_{ij} + \epsilon_{ij}, \tag{19.7}$$

where $\mu' = \mu - \beta\bar{X}$. The quantity μ' represents a grand mean adjusted for the effect of the covariate. The objective in ANCOVA is to test whether Factor A and the covariate have an effect, and so test H_0 : all $\alpha_i = 0$ and $H_0 : \beta = 0$ with separate F tests. However, we will first need to test the assumption that the slopes across Factor A levels are the same. This is accomplished by adding a treatment × covariate interaction to the SAS model, which allows each group to have a different slope. If the test for this effect is significant, we would have a scenario similar to two-way ANOVA when interaction is present (see Chapter 14). In particular, if the interaction is significant tests of the main effects in ANCOVA (Factor A and the covariate X) may not make sense.

19.3.2 ANCOVA for Example 3 - SAS demo

The first step in the analysis (see program below) is to plot the number of eggs (`y`) for each treatment (`treat`) against elytral length, the covariate (`x`), using `proc gplot` (SAS Institute Inc. 2016). This gives some idea whether each treatment group has the same slope, a key assumption of ANCOVA. The slopes do appear to be similar (Fig. 19.19). We then fit the ANCOVA model using `proc glm`, because all the effects in the model are fixed effects (SAS Institute Inc. 2018). The first step is to fit a model with an interaction between the treatment and covariate, and examine the test for the interaction (Fig. 19.20). We see that it was non-significant ($F_{2,35} = 0.02, P = 0.9781$), and so can assume the slopes are the same across treatments. We then rerun the program using the model without interaction (Fig. 19.21). The covariate effect was highly significant ($F_{1,37} = 9.99, P = 0.0031$), suggesting there is a relationship between fecundity and body size. The treatment effect was nonsigificant ($F_{2,37} = 0.52, P = 0.5976$), implying the treatments themselves had no effect on egg numbers. Predators reared on the artificial diet were apparently similar to wild predators on this measure of fitness, controlling for elytral length and so body size. The `proc glm` output also includes a plot of the fitted model and points (Fig. 19.22).

The program also includes an `lsmeans` statement to calculate the least squares means for each treatment group, and test for differences among them using the Tukey method. Least squares means are means adjusted for the effect of other variables in the model, and in the case of ANCOVA are the treatment means adjusted for the covariate. In particular, they have the form

$$\bar{Y}_i(adj) = \bar{Y}_i - \hat{\beta}(\bar{X}_i - \bar{\bar{X}}). \tag{19.8}$$

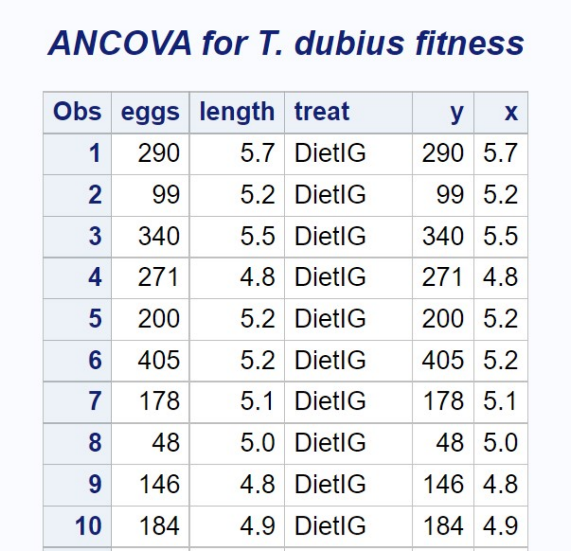
We can see they are composed of two terms, the treatment means and the adjustment for the covariate. Treatment groups that have covariate means ($\bar{X}_i$ values) far from the overall covariate mean ($\bar{\bar{X}}$) receive a larger adjustment. No significant differences were found among the treatment groups (Fig. 19.23), which is not surprising given the overall treatment effect was nonsignificant.

SAS Program

```
* ANCOVA_fitness.sas;
title 'ANCOVA for T. dubius fitness';
data fitness;
    input eggs length treat $;
    * Choose y and x variables;
    y = eggs;
    x = length;
    datalines;
290    5.7   DietIG
 99    5.2   DietIG
340    5.5   DietIG
271    4.8   DietIG
200    5.2   DietIG

etc.

;
run;
* Print data set;
proc print data=fitness;
run;
* Plot data and regression line;
proc gplot data=fitness;
    plot y*x=treat / vaxis=axis1 haxis=axis1 legend=legend1;
    symbol1 i=rl v=star height=2 width=3;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
    legend1 label=(height=2) value=(height=2);
run;
* ANCOVA;
proc glm plots=diagnostics data=fitness;
    class treat;
    * Model with interaction;
    *model y = treat x treat*x;
    * Model without interaction;
    model y = treat x;
    lsmeans treat / pdiff=all adjust=tukey cl lines;
run;
quit;
```



ANCOVA for T. dubius fitness

Obs	eggs	length	treat	y	x
1	290	5.7	DietIG	290	5.7
2	99	5.2	DietIG	99	5.2
3	340	5.5	DietIG	340	5.5
4	271	4.8	DietIG	271	4.8
5	200	5.2	DietIG	200	5.2
6	405	5.2	DietIG	405	5.2
7	178	5.1	DietIG	178	5.1
8	48	5.0	DietIG	48	5.0
9	146	4.8	DietIG	146	4.8
10	184	4.9	DietIG	184	4.9

etc.

Figure 19.18: `ANCOVA_fitness.sas - proc print`

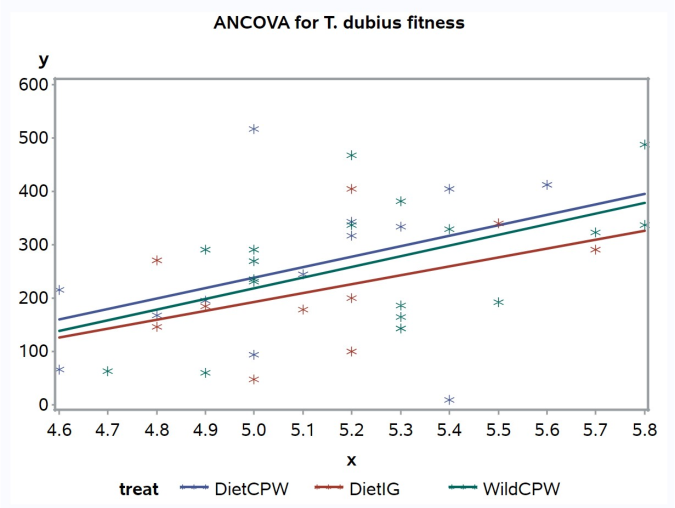
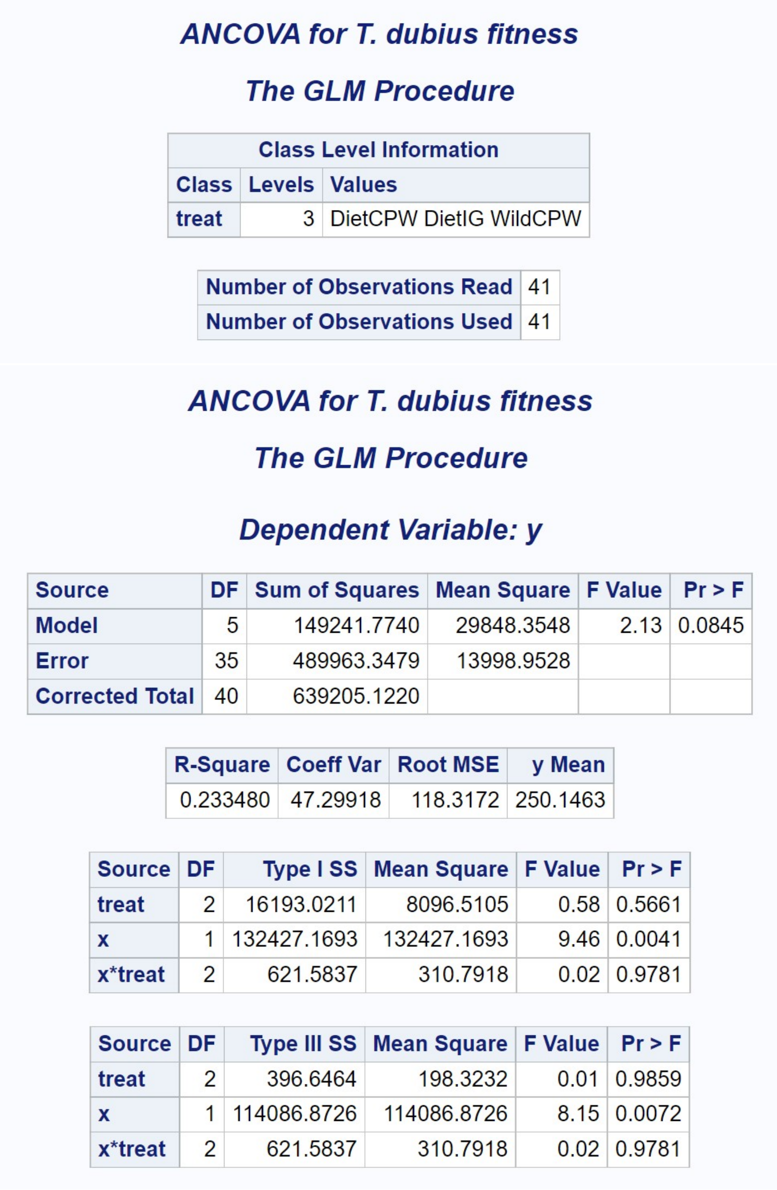


Figure 19.19: `ANCOVA_fitness.sas - proc gplot`



ANCOVA for T. dubius fitness

The GLM Procedure

Class Level Information		
Class	**Levels**	**Values**
treat	3	DietCPW DietIG WildCPW

Number of Observations Read	41
Number of Observations Used	41

ANCOVA for T. dubius fitness

The GLM Procedure

Dependent Variable: y

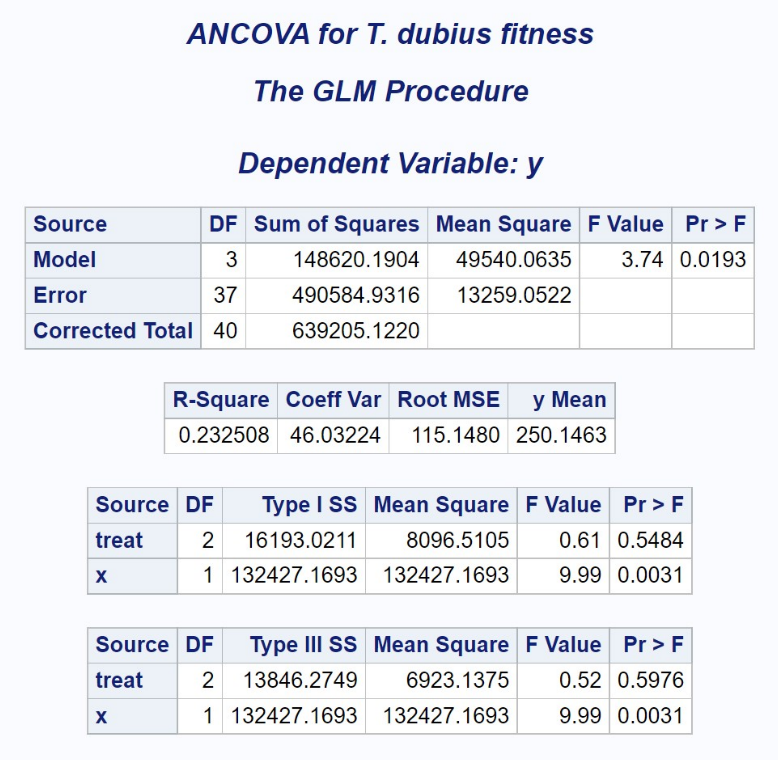
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	149241.7740	29848.3548	2.13	0.0845
Error	35	489963.3479	13998.9528		
Corrected Total	40	639205.1220			

R-Square	Coeff Var	Root MSE	y Mean
0.233480	47.29918	118.3172	250.1463

Source	DF	Type I SS	Mean Square	F Value	Pr > F
treat	2	16193.0211	8096.5105	0.58	0.5661
x	1	132427.1693	132427.1693	9.46	0.0041
x*treat	2	621.5837	310.7918	0.02	0.9781

Source	DF	Type III SS	Mean Square	F Value	Pr > F
treat	2	396.6464	198.3232	0.01	0.9859
x	1	114086.8726	114086.8726	8.15	0.0072
x*treat	2	621.5837	310.7918	0.02	0.9781

Figure 19.20: `ANCOVA_fitness.sas - proc glm` (with interaction)



ANCOVA for T. dubius fitness

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	148620.1904	49540.0635	3.74	0.0193
Error	37	490584.9316	13259.0522		
Corrected Total	40	639205.1220			

R-Square	Coeff Var	Root MSE	y Mean
0.232508	46.03224	115.1480	250.1463

Source	DF	Type I SS	Mean Square	F Value	Pr > F
treat	2	16193.0211	8096.5105	0.61	0.5484
x	1	132427.1693	132427.1693	9.99	0.0031

Source	DF	Type III SS	Mean Square	F Value	Pr > F
treat	2	13846.2749	6923.1375	0.52	0.5976
x	1	132427.1693	132427.1693	9.99	0.0031

Figure 19.21: `ANCOVA_fitness.sas - proc glm` (without interaction)

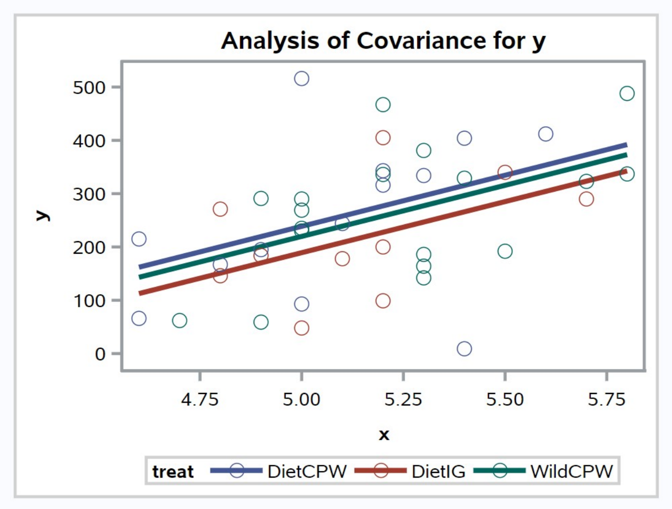
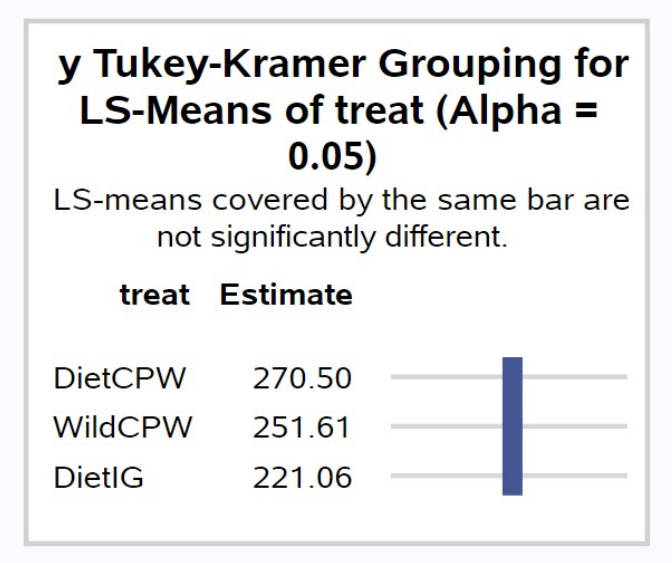


Figure 19.22: ANCOVA_fitness.sas - proc glm (without interaction)



y Tukey-Kramer Grouping for LS-Means of treat (Alpha = 0.05)

LS-means covered by the same bar are not significantly different.

treat	Estimate
DietCPW	270.50
WildCPW	251.61
DietIG	221.06

Figure 19.23: ANCOVA_fitness.sas - proc glm (without interaction)

19.4 References

Cronin, J. T. & Strong, D. R. (1996) Genetics of oviposition success of a thelytokous fairyfly parasitoid, *Anagrus delicatus*. *Heredity* 76: 43-54.

Falconer, D. S. & MacKay, T. F. C. (1996) *Introduction to Quantitative Genetics*, 4th edition. Longman Group Ltd., Essex, England.

Maestre, F. T. & Reynolds, J. F. (2007) Amount or pattern? Grassland responses to the heterogeneity and availability of two key resources. *Ecology* 88: 501-511.

Reeve, J. D., Rojas, M. G. & Morales-Ramos, J. A. (2003) Artificial diet and rearing methods for *Thanasimus dubius* (Coleoptera: Cleridae), a predator of bark beetles (Coleoptera: Scolytidae). *Biological Control* 27: 315-322.

SAS Institute Inc. (2016) *SAS/GRAPH 9.4: Reference, Fifth Edition.* SAS Institute Inc., Cary, NC.

SAS Institute Inc. (2018) *SAS/STAT 15.1 Users Guide* SAS Institute Inc., Cary, NC

Winer, B. J., Brown, D. R. & Michels, K. M. (1991) *Statistical Principles in Experimental Design*, 3rd edition. McGraw-Hill, Inc., Boston, MA.

19.5 Problems

1. A limnologist wants to examine the length of a zooplankton species reared using four different algal growth media (1, 2, 3, and 4). She is also interested in whether there is variation among the containers used to rear the organisms. An experiment is conducted where three containers are used for each rearing medium, for a total of 12 different containers. The containers were randomly selected from a box of containers. The length of four animals was determined for each container, yielding the following data:

Medium	Container	Lengths 1-4 (mm)
1	1	3.1, 3.0, 3.2, 3.0
1	2	3.3, 3.6, 2.8, 2.5
1	3	3.7, 3.4, 3.4, 3.6
2	1	2.7, 2.9, 3.2, 3.0
2	2	2.9, 3.4, 3.5, 2.9
2	3	3.5, 3.5, 3.7, 4.0
3	1	2.8, 2.7, 1.8, 2.5
3	2	2.6, 2.5, 3.2, 2.4
3	3	2.6, 2.9, 1.8, 2.4
4	1	4.1, 4.6, 3.3, 4.5
4	2	3.7, 3.9, 4.0, 3.9
4	3	4.4, 4.4, 3.9, 4.6

   (a) Write an appropriate ANOVA model for this design, stating which factors are fixed, random, and possibly nested.

   (b) Use SAS to analyze these data using your ANOVA model, transforming the observations only if necessary. Is there a significant difference among the four media in zooplankton length?

   (c) Use the Tukey method to compare the media treatments. Interpret your results.

   (d) Compare the magnitude of your variance components. Does there appear to be much variation among containers?

2. An ecologist is interested in the effect of three management treatments (labeled 1, 2, and 3) on the abundance of an endangered snail. Treatment 2 is a control treatment. Twenty-four plots are established and

the three treatments assigned at random to the plots. The density of snails is then measured at a later time, as well as a covariate in the form of a habitat index. Larger values of the habitat index are thought to indicate better snail habitat. See data set below.

Treatment	Index	Snails
1	9.3	23.0
1	9.8	24.9
1	9.9	24.7
1	10.1	24.6
1	8.9	23.4
1	10.8	27.1
1	9.6	25.4
1	10.7	25.4
2	11.9	21.8
2	9.6	18.8
2	10.3	21.0
2	10.8	21.5
2	9.9	20.9
2	10.9	22.6
2	8.9	19.8
2	10.2	22.4
3	11.2	23.4
3	10.3	18.5
3	11.1	22.3
3	9.8	20.5
3	11.2	20.5
3	8.7	18.4
3	8.4	18.7
3	10.5	19.2

(a) Test for equality of slopes among the different treatment groups using SAS. Is this key assumption of ANCOVA satisfied?

(b) Use ANCOVA and SAS to test for overall treatment and covariate effects in this experiment, and the Tukey method to compare the different treatments. Interpret and discuss your results. Is there a significant treatment and covariate effect? How do the different treatments compare?

3. A scientist interested in aquaculture raises fish using three kinds of treatments in a factorial design. There were two fish diets (A and B), two strains of fish (1 and 2), and three temperatures (22^o, 24^o, and 26^oC). Two fish were reared for each combination of the treatments. The following data were obtained:

Diet	Strain	Temp	Weight (lb)
A	1	22	5.5
A	1	22	5.8
A	1	24	5.9
A	1	24	5.7
A	1	26	6.2
A	1	26	5.9
A	2	22	5.2
A	2	22	5.0
A	2	24	5.4
A	2	24	5.6
A	2	26	5.0
A	2	26	4.9
B	1	22	5.4
B	1	22	4.8
B	1	24	5.4
B	1	24	5.4
B	1	26	5.7
B	1	26	5.5
B	2	22	5.2
B	2	22	4.8
B	2	24	5.1
B	2	24	5.1
B	2	26	4.8
B	2	26	4.5

(a) Write an appropriate ANOVA model for this design, stating which factors are fixed or random.

(b) Use SAS to analyze these data using your ANOVA model, transforming the observations only if necessary. Interpret the results of your analysis.

Chapter 20

Methods for Categorical Data

Categorical data are observations that fall into two or more discrete categories, such as female vs. male organisms, age or size classes, or different phenotypes in genetic studies (Chapter 1). This requires a different type of statistical model than in previous chapters, where the observations were assumed to have a normal distribution. We will instead use the binomial and multinomial distributions to model categorical data, and derive likelihood ratio and chi-square tests of various hypotheses. Recall that the binomial distribution can be used to model data with two categories (see Chapter 5). **The multinomial distribution is a generalization of the binomial to data with more than two categories.**

One class of test we will examine are called **goodness-of-fit tests**. These tests compare the observed frequencies of different categories of observations with those expected under some null hypothesis. For example, recall the laboratory rearing study of *Thanasimus dubius* described in Chapter 3. We might be interested in whether the sex ratio for these predatory beetles is close to 1:1 (50% females, 50% males), as occurs in many diploid sexual organisms. This is our null hypothesis and it implies that the probability p a sampled individual is female is 0.5, or $H_0 : p = 0.5$. Suppose we have a sample of $n = 130$ beetles as in this data set. What are the expected frequencies of females and males in this sample? Recall that $E[Y] = np$ for the binomial distribution, where n is the sample size (Chapter 5). Under H_0, we would therefore expect $E_1 = np = 130(0.5) = 65$ females and $E_2 = n(1 - p) = 130(0.5) = 65$ males. The observed frequencies are $O_1 = 60$ females and $O_2 = 70$ males for this data set. It is common to organize these results into following form (Table 20.1):

Table 20.1: Observed and expected frequencies of female and male *T. dubius* from a laboratory rearing study (Reeve et al. 2003).

	Females	Males	$\sum$
i	1	2	
O_i	60	70	130
E_i	65	65	130

A goodness-of-fit test for $H_0 : p = 0.5$ provides a way of comparing these observed and expected frequencies, generating a test statistic and P value for the test. Based on these results we may accept or reject this null hypothesis, and in this case the result was non-significant ($P = 0.3805$). We will later see how goodness-of-fit tests may be applied to data with more categories and cases where certain model parameters are estimated from the data.

Tests of independence are a second class of tests for categorical data. Suppose that the observations in a data set can be classified in two different ways. For example, a sample of amphibians could be classified into different species and whether individuals of a given species are infected with a pathogen. Using a test of independence, we can test whether species and infection status are independent events (see Chapter 4). Equivalently, we can test whether the probability of being infected is the same across species. To make things more concrete, suppose that four amphibian species (A, B, C, and D) are randomly sampled and scored for infection, yielding Table 20.2. The null hypothesis of independence, or an equal probability of being infected across all species, can be expressed as follows. Let p_A be the overall probability an individual of species A is sampled (infected or not), while p_I is the probability it is infected (across all four species). If species and infection status are independent, we would expect by definition that the probability of sampling an infected individual of species A would be $p_A p_I$ (see Chapter 4). A similar relationship would hold for the other possible outcomes, and the null hypothesis of independence can be expressed in this form.

Tests of independence also make use of observed and expected frequencies, with the expected frequencies calculated under the null hypothesis of independence (see Table 20.2). Subscripts are commonly used to indicate the observed and expected frequencies in particular cells of the table, with the first subscript indicating the row and the second the column in the table. For example, in Table 20.2 we have $O_{11} = 7, O_{21} = 18, O_{12} = 12, O_{22} = 38$, and so forth. We will later see how to calculate the expected frequencies under the null hypothesis of independence. There appear to be substantial differences between the observed and expected frequencies in this table, and in fact the test of independence was highly significant ($P = 0.0002$), suggesting that amphibian species and infection status are **not** independent. We will focus on two-way tables like the one below, but it is also possible to conduct tests of independence for three-way or higher tables. However, these problems are more commonly addressed using **loglinear models**, which have an ANOVA-like structure and feel but focus on testing the interactions between factors, which are equivalent to tests of independence (Agresti 1990).

Table 20.2: Observed frequencies of infected and non-infected individuals in four amphibian species. Below each observed frequency is the expected frequency under the null hypothesis of independence.

			Species		
Infected	A	B	C	D	$\sum$
Yes	7	12	15	27	61
	10.167	20.333	14.233	16.267	
No	18	38	20	13	89
	14.833	29.667	20.767	23.733	
$\sum$	25	50	35	40	150

20.1 Goodness-of-fit tests

As a simple example of a goodness-of-fit test, consider the data set involving male and female *T. dubius*. Suppose we want to test the hypothesis that the sex ratio is 1:1 (50% female, 50% male) in this species. The population falls into two categories, female or male, which suggests using the binomial distribution to model the observations. Suppose that we have a sample of size n from this population and let Y be the number of females in the sample, a binomial random variable. If p is the probability that a *T. dubius* adult is female, then the probability the sample will have y females is given by the formula

$$P[Y = y] = \binom{n}{y} p^y (1-p)^{n-y}. \tag{20.1}$$

The null hypothesis that the sex ratio is 1:1 implies that $p = 0.5$, which can be written as $H_0 : p = 0.5$. The alternative is that the sex ratio differs from 1:1, or $H_1 : p \neq 0.5$. More generally, we will be interested in testing $H_0 : p = p_0$ vs. $H_1 : p \neq p_0$ where p_0 is some probability.

We now develop a likelihood ratio test for $H_0 : p = p_0$ vs. $H_1 : p \neq p_0$, assuming the observations have a binomial distribution. It is a goodness-of-fit test because we will be comparing the observed frequencies of females and males with that expected under H_0, and if observed and expected frequencies are substantially different we will reject H_0. The likelihood ratio test uses the ratio of the likelihoods under H_0 and H_1 as the test statistic (see Chapter 10).

Recall that the likelihood function for discrete distributions is just the probability of the observed data (see Chapter 8). The data are fixed quantities in this function, while the parameters of the distribution are free to vary. In this case, the value of y (the number of females in the sample) is the data while p is the parameter that is free to vary, and so the likelihood function for binomial data would be

$$L(p) = \binom{n}{y} p^y (1-p)^{n-y}. \tag{20.2}$$

We first need to find the maximum value of the likelihood under H_0. Under the null hypothesis the parameter p is set equal to p_0, and so we have

$$L_{H_0} = \binom{n}{y} p_0^y (1-p_0)^{n-y}. \tag{20.3}$$

This is the only value that can be taken by L_{H_0}, because all the other quantities are fixed, and so this is also its maximum. Under H_1, the parameter p is free to vary in $L(p)$. The maximum value of the likelihood function occurs at $\hat{p} = y/n$, the maximum likelihood estimate of p. This is simply the proportion of females in the sample. Thus,

$$L_{H_1} = \binom{n}{y} \hat{p}^y (1-\hat{p})^{n-y} = \binom{n}{y} (y/n)^y (1-y/n)^{n-y}. \tag{20.4}$$

The test statistic is the ratio of these two likelihoods:

$$\begin{aligned}
\lambda &= \frac{L_{H_0}}{L_{H_1}} && (20.5)\\
&= \frac{\binom{n}{y} p_0^y (1-p_0)^{n-y}}{\binom{n}{y} (y/n)^y (1-y/n)^{n-y}} && (20.6)\\
&= \frac{p_0^y (1-p_0)^{n-y}}{(y/n)^y (1-y/n)^{n-y}} && (20.7)\\
&= \left(\frac{p_0}{y/n}\right)^y \left(\frac{1-p_0}{1-y/n}\right)^{n-y} && (20.8)\\
&= \left(\frac{np_0}{y}\right)^y \left(\frac{n(1-p_0)}{n-y}\right)^{n-y} && (20.9)\\
&= \left(\frac{E_1}{O_1}\right)^{O_1} \left(\frac{E_2}{O_2}\right)^{O_2}. && (20.10)
\end{aligned}$$

Here O_1 and O_2 would be the observed frequencies of females and males, while $E_1 = np_0$ and $E_2 = n(1-p_0)$ are the corresponding expected frequencies (see Table 20.1). Under H_0, the quantity

$$G^2 = -2 \ln \lambda \qquad (20.11)$$

has approximately a χ^2 distribution with one degree of freedom, with the approximation improving as n increases (Agresti 1990). In terms of the observed and expected frequencies, we have

$$\begin{aligned}
G^2 &= -2 \ln \lambda && (20.12)\\
&= -2 \ln \left[\left(\frac{E_1}{O_1}\right)^{O_1} \left(\frac{E_2}{O_2}\right)^{O_2} \right] && (20.13)\\
&= -2[O_1 \ln(E_1/O_1) + O_2 \ln(E_2/O_2)] && (20.14)\\
&= 2[O_1 \ln(O_1/E_1) + O_2 \ln(O_2/E_2)]. && (20.15)
\end{aligned}$$

Similar to other likelihood ratio tests that utilize the χ^2 distribution, the degrees of freedom are equal to the difference in the number of parameters free between the H_1 and H_0 models (see Chapter 14). There is one free parameter under H_1, namely p, but under H_0 we have $p = p_0$, a fixed quantity. Thus, there is a difference of one parameter between the two models, implying one degree of freedom. G^2 values will become large if the observed and expected frequencies are different.

Another commonly used statistic for this goodness-of-fit test is the quantity

$$X^2 = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} \tag{20.16}$$

(Agresti 1990). Under H_0, X^2 has approximately a χ^2 distribution with one degree of freedom. Although the two test statistics G^2 and X^2 are different in form, they usually yield similar values and test results. X^2 values also become large as the observed and expected frequencies diverge. This test is often called a 'chi-square' or 'χ^2' test, although the likelihood ratio test also uses the χ^2 distribution.

Goodness-of-fit test - sample calculation

We now conduct a goodness-of-fit test for the Table 20.1 data, testing $H_0 : p = 0.5$. We have

$$\begin{aligned} G^2 &= 2[O_1 \ln(O_1/E_1) + O_2 \ln(O_2/E_2)] && (20.17) \\ &= 2[60 \ln(60/65) + 70 \ln(70/65)] && (20.18) \\ &= 2[-4.803 + 5.188] && (20.19) \\ &= 0.770. && (20.20) \end{aligned}$$

We next find the P value from Table C and obtain a non-significant result ($G^2 = 0.770, df = 1, P < 0.5$). Thus, there was no evidence against a 1:1 sex ratio in this study.

We next calculate the equivalent X^2 statistic for these data. We have

$$\begin{aligned} X^2 &= \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} && (20.21) \\ &= \frac{(60 - 65)^2}{65} + \frac{(70 - 65)^2}{65} && (20.22) \\ &= 0.385 + 0.385 && (20.23) \\ &= 0.770. && (20.24) \end{aligned}$$

The result is identical to G^2 and so the P value is the same ($X^2 = 0.770, df = 1, P < 0.5$). The test results are often similar for these two statistics, although seldom identical as in this case.

Goodness-of-fit test - SAS demo

We can use proc freq in SAS to conduct a goodness-of-fit test for the Table 20.1 data using the X^2 statistic (SAS Institute Inc. 2016). This procedure does not provide the likelihood ratio test involving G^2, but there is another option that is actually better than both. SAS can conduct an exact chi-square (X^2) test where the distribution of the test statistic under H_0 is determined exactly, instead of approximating it with a χ^2 distribution. This approach is computationally intensive and may be impractical for large sample sizes, but in this case the chi-square (X^2) test would be valid and the exact test unnecessary.

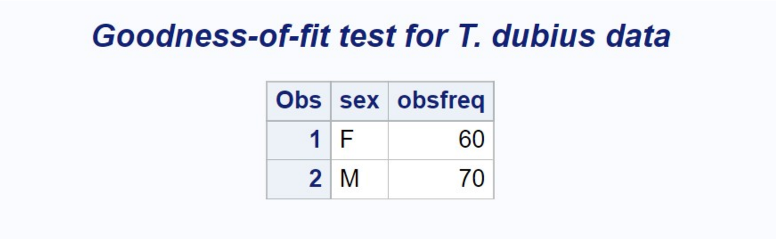
The first step in the analysis is to make a SAS data set using the observed frequencies in Table 20.1. The variable obsfreq contains this information for each value of sex (see SAS program below). The data could also have been entered as individual observations with a single data line for each observation, as in the original data set (see Chapter 3). We would then use proc freq to tabulate the data.

Now examine the proc freq portion of the program. The order=data option asks SAS to use the order of the categories (values of sex) given by the data, rather than alphabetically. The tables line requests a frequency table for sex. The next step is to tell SAS the probabilities under H_0 for each sex, which are $p = 0.5$ for females and $1 - p = 0.5$ for males. This is accomplished using the option testp = (0.5 0.5). The order of the probabilities in the testp statement should match the order of the categories in the data. The weight command tells proc freq that the data are in the form of frequencies, and the name of the variable containing these frequencies (obsfreq). An exact chi-square (X^2) test is requested by the command exact chisq.

Examining the SAS output (Fig. 20.2), we find that the exact chi-square (X^2) test was non-significant ($X^2 = 0.769, df = 1, P = 0.4300$). There is no evidence that the sex ratio differs from 1:1 in this organism.

SAS Program

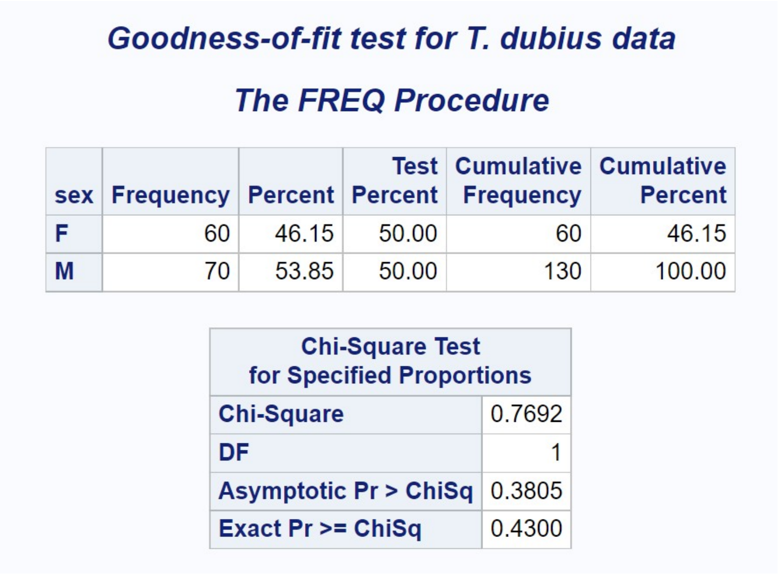
```
* gof_clerids.sas;
title 'Goodness-of-fit test for T. dubius data';
data elytra;
    input sex \$ obsfreq;
    datalines;
F  60
M  70
;
run;
* Print data set;
proc print data=elytra;
run;
* Goodness-of-fit test (Chi-square only);
proc freq data=elytra order=data;
    tables sex / testp=(0.5 0.5) chisq cellchi2 expected;
    weight obsfreq;
    * Compute exact test if frequencies low, takes too long for large data sets;
    exact chisq;
run;
quit;
```



Goodness-of-fit test for T. dubius data

Obs	sex	obsfreq
1	F	60
2	M	70

Figure 20.1: `gof_clerids.sas - proc print`



Goodness-of-fit test for T. dubius data

The FREQ Procedure

sex	Frequency	Percent	Test Percent	Cumulative Frequency	Cumulative Percent
F	60	46.15	50.00	60	46.15
M	70	53.85	50.00	130	100.00

Chi-Square Test for Specified Proportions	
Chi-Square	0.7692
DF	1
Asymptotic Pr > ChiSq	0.3805
Exact Pr >= ChiSq	0.4300

Figure 20.2: `gof_clerids.sas - proc freq`

20.1.1 Goodness-of-fit tests for a categories

We now examine goodness-of-fit tests for data with a different categories. A common type occurs in genetic studies where different genotypes are crossed, such as Mendel's classic experiments involving pea plants (Mendel 1865). One of his experiments created hybrids for two genes governing the shape (round or wrinkled) and color (yellow or green) of the peas, which were then crossed and the phenotypes of the offspring scored. A total of $n = 556$ peas were observed (Table 20.3).

Table 20.3: Observed and expected frequencies for a dihybrid cross (Mendel 1865).

	Round yellow	Round green	Wrinkled yellow	Wrinkled green	$\sum$
i	1	2	3	4	
O_i	315	101	108	32	556
E_i	312.75	104.25	104.25	34.75	556

This table has $a = 4$ categories. If we assume Mendelian genetics, with the round allele dominant over the wrinkled one and yellow color dominant over green, we would expect to see these four phenotypes in a 9:3:3:1 ratio. This forms the null hypothesis for this problem. We can express it in the form $H_0 : p_1 = 9/16 = 0.5625, p_2 = 3/16 = 0.1875, p_3 = 3/16 = 0.1875$, and $p_4 = 1/16 = 0.0625$. The alternative H_1 is that the probabilities differ from these values. More generally, we will be interested in testing $H_0 : p_1 = p_{10}, p_2 = p_{20}, p_3 = p_{30}$, and $p_4 = p_{40}$ vs. some alternative hypothesis H_1 where the probabilities differ from these values.

Also shown in Table 20.3 are the expected frequencies under H_0, calculated using the formula $E_i = np_i$. We have $E_1 = 556(0.5625) = 312.75$, $E_2 = 556(0.1875) = 104.25 = E_3$, and $E_4 = 556(0.0625) = 34.75$. These are the expected numbers of peas for each phenotype assuming that H_0 is true.

We need a different distribution to model these observations, a generalization of the binomial called the **multinomial distribution**. Suppose that n total peas are sampled, and let Y_1, Y_2, Y_3 and Y_4 be random variables corresponding to the four phenotypes, with y_1 the observed number of round and yellow peas, y_2 the number of round and green, y_3 the number of wrinkled and yellow, while y_4 is wrinkled and green. Because $n = Y_1 + Y_2 + Y_3 + Y_4$ there is some dependence among the four variables (if we know three, the fourth is determined by this relationship). Let p_1 be the probability that a pea is round and yellow, with p_2, p_3, and p_4 similarly defined. The four probabilities sum to one ($p_1 + p_2 + p_3 + p_4 = 1$), which implies the distribution really has only three parameters. Then, the probability of observing y_1, y_2, y_3, and y_4 peas of each type is given by the multinomial distribution, which has the form

$$P[Y_1 = y_1, Y_2 = y_2, Y_3 = y_3, Y_4 = y_4] = \frac{n!}{y_1!y_2!y_3!y_4!} p_1^{y_1} p_2^{y_2} p_3^{y_3} p_4^{y_4}. \qquad (20.25)$$

This distribution can be readily extended to any number of categories.

Using the multinomial distribution as a model for the observations, we can extend the G^2 goodness-of-fit statistic to a categories by adding more terms of the form $O_i \ln(O_i/E_i)$. For a table with a categories, we have

$$G^2 = 2 \sum_{i=1}^{a} O_i \ln(O_i/E_i). \qquad (20.26)$$

Under H_0, G^2 has a χ^2 distribution with $a - 1$ degrees of freedom. They are equal to $a - 1$ because there are $a - 1$ free parameters (p_1, p_2, etc.) under H_1 but none free under H_0. Similarly, the X^2 statistic can be generalized as

$$X^2 = \sum_{i=1}^{a} \frac{(O_i - E_i)^2}{E_i}. \qquad (20.27)$$

This statistic also has $a - 1$ degrees of freedom under H_0.

Goodness-of-fit test - sample calculation

We illustrate a goodness-of-fit test for $a = 4$ categories using the pea data, testing $H_0 : p_1 = 0.5625, p_2 = 0.1875, p_3 = 0.1875$, and $p_4 = 0.0625$. Table 20.3 presents the observed and expected frequencies, from which we can calculate G^2. We have

$$\begin{aligned}
G^2 &= 2\sum_{i=1}^{a} O_i \ln(O_i/E_i) && (20.28)\\
&= 2[315\ln(315/312.75) + 101\ln(101/104.25) && (20.29)\\
&+ 108\ln(108/104.25) + 32\ln(32/34.75)] && (20.30)\\
&= 2[2.258 - 3.199 + 3.817 - 2.638] && (20.31)\\
&= 0.476. && (20.32)
\end{aligned}$$

The degrees of freedom for the test are $a - 1 = 4 - 1 = 3$. We next find the P value from Table C and obtain a non-significant result ($G^2 = 0.476, df = 3, P < 0.95$). The observed frequencies apparently agree with the Mendelian ratios of 9:3:3:1.

We next conduct a chi-square (X^2) test for these data. We have

$$\begin{aligned}
X^2 &= \sum_{i=1}^{a} \frac{(O_i - E_i)^2}{E_i} && (20.33)\\
&= \frac{(315 - 312.75)^2}{312.75} + \frac{(101 - 104.25)^2}{104.25} && (20.34)\\
&+ \frac{(108 - 104.25)^2}{104.25} + \frac{(32 - 34.75)^2}{34.75} && (20.35)\\
&= 0.016 + 0.101 + 0.135 + 0.218 && (20.36)\\
&= 0.470 && (20.37)
\end{aligned}$$

We also obtain a non-significant result with this test ($X^2 = 0.470, df = 3, P < 0.95$).

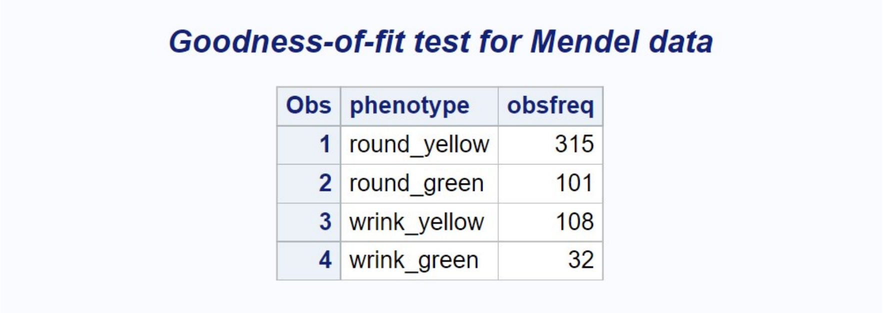
Goodness-of-fit test - SAS demo 2

The chi-square (X^2) test for the Table 20.3 data can also be conducted in SAS. A data set is first made using the observed frequencies, with `proc freq` then used to carry out the test. The `testp` statement lists the probabilities under $H_0 : p_1 = 0.5625, p_2 = 0.1875, p_3 = 0.1875$, and $p_4 = 0.0625$. The order of the probabilities matches the order of the phenotypes in the data set. See SAS program and output below. An exact chi-square test is also requested which may take SAS some period of time to calculate.

We see from the SAS output (Fig. 20.4) that the exact chi-square (X^2) test was non-significant ($X^2 = 0.470, df = 3, P = 0.9272$). There is no evidence that the ratios of the phenotypes differ from the Mendelian 9:3:3:1 ratio.

SAS Program

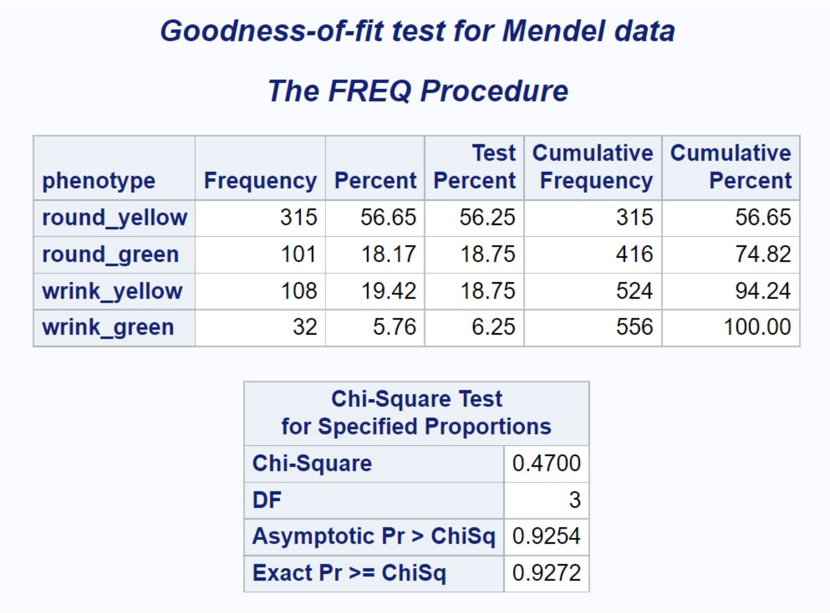
```
* gof_peas.sas;
title 'Goodness-of-fit test for Mendel data';
data peas;
    input phenotype :\$12. obsfreq;
    datalines;
round_yellow  315
round_green   101
wrink_yellow  108
wrink_green    32
;
run;
* Print data set;
proc print data=peas;
run;
* Goodness-of-fit test (Chi-square only);
proc freq data=peas order=data;
    tables phenotype / testp=(0.5625 0.1875 0.1875 0.0625) chisq cellchi2 expected;
    weight obsfreq;
    * Compute exact test if frequencies low, takes too long for large data sets;
    exact chisq;
run;
quit;
```



Goodness-of-fit test for Mendel data

Obs	phenotype	obsfreq
1	round_yellow	315
2	round_green	101
3	wrink_yellow	108
4	wrink_green	32

Figure 20.3: gof_peas.sas - proc print



Goodness-of-fit test for Mendel data

The FREQ Procedure

phenotype	Frequency	Percent	Test Percent	Cumulative Frequency	Cumulative Percent
round_yellow	315	56.65	56.25	315	56.65
round_green	101	18.17	18.75	416	74.82
wrink_yellow	108	19.42	18.75	524	94.24
wrink_green	32	5.76	6.25	556	100.00

Chi-Square Test for Specified Proportions	
Chi-Square	0.4700
DF	3
Asymptotic Pr > ChiSq	0.9254
Exact Pr >= ChiSq	0.9272

Figure 20.4: `gof_peas.sas - proc freq`

20.1.2 Goodness-of-fit tests with estimated parameters

Another common type of goodness-of-fit test compares the observed frequencies with that expected for some theoretical distribution, such as the Poisson. We previously fitted a Poisson distribution to count data and compared graphically the observed and expected frequencies (Chapter 5). We now compare these frequencies using a goodness-of-fit test similar to previous examples. The null hypothesis in this case is that the observations are Poisson in distribution, while the alternative is that some other distribution describes them.

There are two additional considerations with these goodness-of-fit tests. One is that the Poisson parameter λ must be estimated from the observations, using the estimator $\hat{\lambda} = \bar{Y}$. This requires an adjustment to the degrees of freedom for the test (Agresti 1990). **In particular, one degree of freedom is subtracted from the total for every parameter estimated.** For the Poisson distribution we have to estimate λ, and so the degrees of freedom are $a - 1 - 1 = a - 2$. A second consideration involves the expected frequencies in the tests. The distributions of both G^2 and X^2 are approximately χ^2 under H_0, but this approximation works better if the expected frequencies are not too small, although there is no universal rule on what constitutes small (Agresti 1990). **One commonly used but overly conservative rule is $E_i \geq 5$ - the expected frequencies must equal or exceed five for all cells.** We have not encountered this problem in previous examples but it does occur with goodness-of-fit tests for the Poisson and other discrete distributions. **The solution is to combine adjacent cells in the table until the expected frequencies equal or exceed five. The observed frequencies are also combined to match the expected ones.**

20.1.3 Corn borers - SAS demo

We will use a SAS program to automate most of the calculations for this goodness-of-fit test. The test cannot be totally automated, however, because the expected frequencies need to be manually combined at some point. Recall the corn borers data and SAS program from Chapter 5. The program listed below is similar, except that some additional quantities needed for the tests are calculated in the second `data` step. In particular, the program calculates the individual terms for the X^2 and G^2 tests, defined as the SAS variables `cellchi2` and `olnoe`, and keeps a running total of these values in the variables `sumchi2` and `sumlike`. See Fig. 20.6 for the results of these calculations.

As before, define E_1 to be the expected frequency for the first cell ($y = 0$), E_2 the expected frequency for the second cell ($y = 1$), and so forth. We see that the expected frequency $E_8 = 3.2041 < 5$, as are the remaining values. We therefore add them together so that the combined expected frequency is greater than five. We have

$$\begin{aligned} E_{\text{combined}} &= 3.204 + 1.268 + 0.446 && (20.38)\\ &+ 0.141 + 0.041 + 0.011 && (20.39)\\ &= 5.111. && (20.40)\end{aligned}$$

We also need to combine the observed frequencies for these cells, to obtain

$$\begin{aligned} O_{\text{combined}} &= 5 + 3 + 4 + 3 + 0 + 1 && (20.41)\\ &= 16. && (20.42)\end{aligned}$$

We then calculate an overall G^2 statistic as follows. First, we calculate the component of this test statistic for the combined cells, obtaining

$$O_{\text{combined}} \ln(O_{\text{combined}}/E_{\text{combined}}) = 16 \ln(16/5.111) = 18.259. \qquad (20.43)$$

We then find the running total of these components (`sumlike`) prior to the combined cells from the SAS output, which is 13.078. The overall test statistic is therefore equal to

$$G^2 = 2[13.078 + 18.259] = 62.674. \qquad (20.44)$$

There are $a = 8$ categories in the test, so the degrees of freedom are $a - 2 = 8 - 2 = 6$. Using Table C, we find that the test was highly significant ($G^2 = 62.674, df = 6, P < 0.001$). This result strongly suggests the observations do not have a Poisson distribution. Instead, they appear to have an overdispersed pattern with an excess of zeros and large values relative to the Poisson (Fig. 20.7).

We now calculate a chi-square (X^2) goodness-of-fit test for these observations. We first calculate the component of this statistic for the combined cells, obtaining

$$\frac{(O_{\text{combined}} - E_{\text{combined}})^2}{E_{\text{combined}}} = \frac{(16 - 5.111)^2}{5.111} = 23.199. \qquad (20.45)$$

The running total of these components (sumchi2) prior to the combined cells is 80.705, and so the overall test statistic is

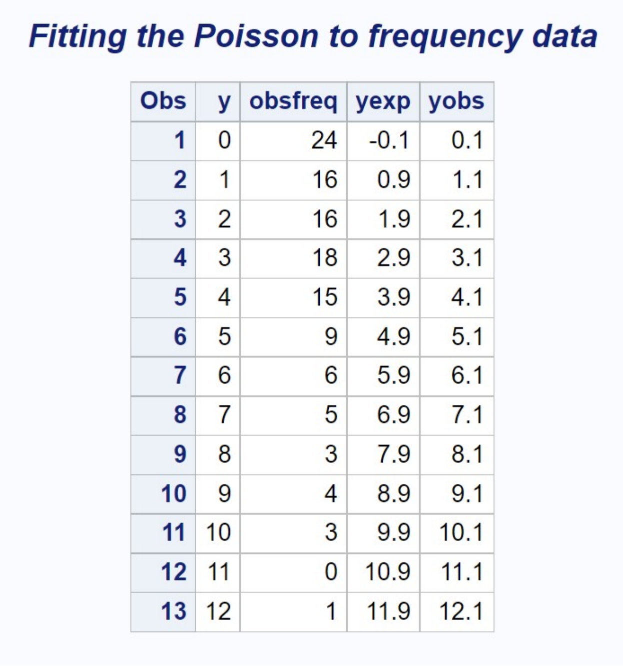
$$X^2 = 80.705 + 23.199 = 103.904. \tag{20.46}$$

The degrees of freedom are $a - 2 = 7 - 2 = 6$, the same as above. The test was again highly significant ($X^2 = 103.904, df = 6, P < 0.001$).

SAS Program

```
* Poisson_fit2_gof.sas;
title ’Fitting the Poisson to frequency data’;
data poisson;
    input y obsfreq;
    * Generate offset y values for plot;
    yexp = y - 0.1; yobs = y + 0.1;
    datalines;
0  24
1  16
2  16
3  18
4  15
5   9
6   6
7   5
8   3
9   4
10  3
11  0
12  1
;
run;
* Print data set;
proc print data=poisson;
run;
* Descriptive statistics, save ybar, n, and var to data file;
proc univariate data=poisson;
    var y;
    histogram y / vscale=count;
    freq obsfreq;
    output out=stats mean=ybar n=n var=var;
run;
* Print output data file;
proc print data=stats;
run;
```

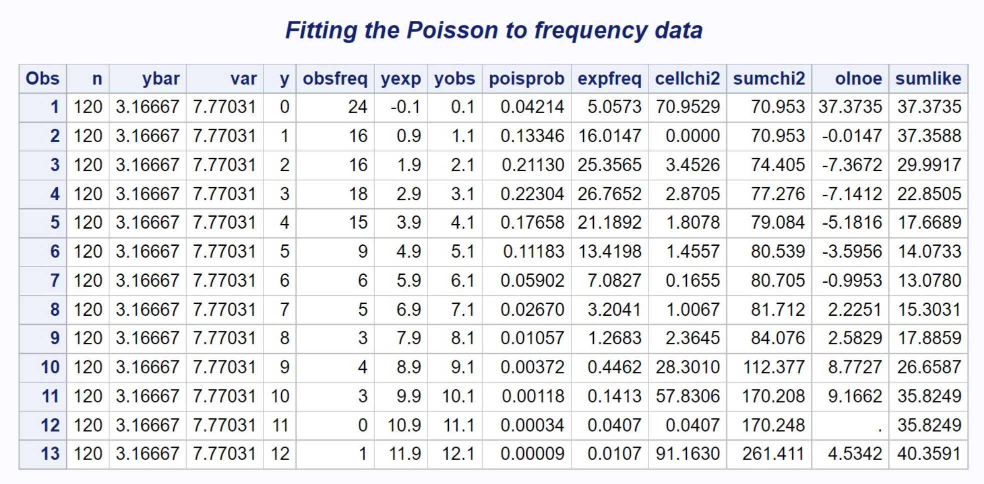
```
* Calculate expected frequencies using ybar;
data poisfit;
    if _n_ = 1 then set stats;
    set poisson;
    poisprob = pdf(’poisson’,y,ybar);
    expfreq = n*poisprob;
    * Calculate test values for each cell;
    cellchi2 = ((obsfreq - expfreq)**2)/expfreq;
    sumchi2 + cellchi2;
    olnoe = obsfreq*log(obsfreq/expfreq);
    sumlike + olnoe;
run;
* Print observed and expected frequencies;
proc print data=poisfit;
run;
* Plot observed and expected frequencies;
proc gplot data=poisfit;
    plot expfreq*yexp=1 obsfreq*yobs=2 / overlay legend=legend1 vref=0 wvref=3
    vaxis=axis1 haxis=axis1;
    symbol1 i=needle v=circle c=red width=3 height=2;
    symbol2 i=needle v=square c=blue width=3 height=2;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
    legend1 label=(height=2) value=(height=2);
run;
quit;
```



Fitting the Poisson to frequency data

Obs	y	obsfreq	yexp	yobs
1	0	24	-0.1	0.1
2	1	16	0.9	1.1
3	2	16	1.9	2.1
4	3	18	2.9	3.1
5	4	15	3.9	4.1
6	5	9	4.9	5.1
7	6	6	5.9	6.1
8	7	5	6.9	7.1
9	8	3	7.9	8.1
10	9	4	8.9	9.1
11	10	3	9.9	10.1
12	11	0	10.9	11.1
13	12	1	11.9	12.1

Figure 20.5: `Poisson_fit2_gof.sas - proc print`



Fitting the Poisson to frequency data

Obs	n	ybar	var	y	obsfreq	yexp	yobs	poisprob	expfreq	cellchi2	sumchi2	olnoe	sumlike
1	120	3.16667	7.77031	0	24	-0.1	0.1	0.04214	5.0573	70.9529	70.953	37.3735	37.3735
2	120	3.16667	7.77031	1	16	0.9	1.1	0.13346	16.0147	0.0000	70.953	-0.0147	37.3588
3	120	3.16667	7.77031	2	16	1.9	2.1	0.21130	25.3565	3.4526	74.405	-7.3672	29.9917
4	120	3.16667	7.77031	3	18	2.9	3.1	0.22304	26.7652	2.8705	77.276	-7.1412	22.8505
5	120	3.16667	7.77031	4	15	3.9	4.1	0.17658	21.1892	1.8078	79.084	-5.1816	17.6689
6	120	3.16667	7.77031	5	9	4.9	5.1	0.11183	13.4198	1.4557	80.539	-3.5956	14.0733
7	120	3.16667	7.77031	6	6	5.9	6.1	0.05902	7.0827	0.1655	80.705	-0.9953	13.0780
8	120	3.16667	7.77031	7	5	6.9	7.1	0.02670	3.2041	1.0067	81.712	2.2251	15.3031
9	120	3.16667	7.77031	8	3	7.9	8.1	0.01057	1.2683	2.3645	84.076	2.5829	17.8859
10	120	3.16667	7.77031	9	4	8.9	9.1	0.00372	0.4462	28.3010	112.377	8.7727	26.6587
11	120	3.16667	7.77031	10	3	9.9	10.1	0.00118	0.1413	57.8306	170.208	9.1662	35.8249
12	120	3.16667	7.77031	11	0	10.9	11.1	0.00034	0.0407	0.0407	170.248	.	35.8249
13	120	3.16667	7.77031	12	1	11.9	12.1	0.00009	0.0107	91.1630	261.411	4.5342	40.3591

Figure 20.6: `Poisson_fit2_gof.sas - proc print`

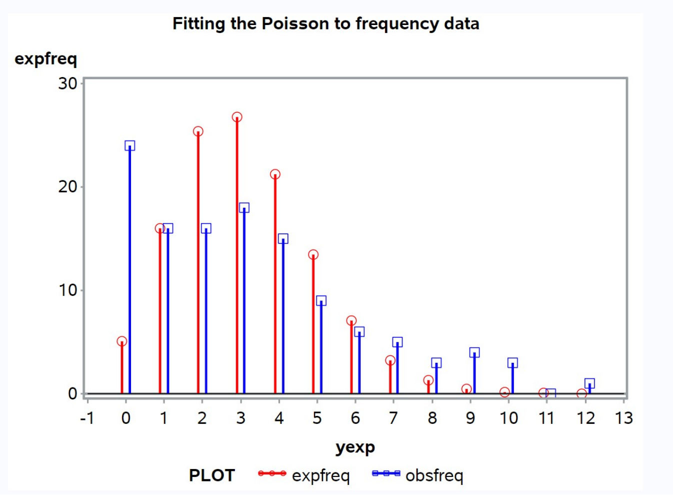


Figure 20.7: `Poisson_fit2_gof.sas - proc gplot`

20.2 Tests of independence

We now develop tests of independence for tables in which the observations are classified in two different ways, known as two-way tables. The test statistics are similar to previous likelihood ratio (G^2) and chi-square (X^2) goodness-of-fit tests, and use the multinomial distribution to model the observations. Because the null hypothesis is different for tests of independence, however, the expected frequencies are calculated differently as are the degrees of freedom. Further details are provided in Agresti (1990).

We first examine how the expected frequencies are constructed for tests of independence, but these calculations will require estimates of the probabilities for certain events. Recall the Table 20.2 example where amphibians were sampled and classified by species and infection status. What is the overall probability of sampling species A, regardless of infection status? Let the quantity p_{+1} stand for this probability, where the + symbol indicates the overall probability combining infected and uninfected individuals while '1' stands for the first column in Table 20.2, which is species A. We can estimate this probability by summing the number of infected and uninfected individuals for species A and dividing by the sample size n. If we let O_{+1} stand for this sum, we have

$$\hat{p}_{+1} = \frac{O_{+1}}{n} = \frac{25}{150} = 0.167. \tag{20.47}$$

This is just the column total for species A divided by the sample size n. We can similarly calculate the probability of sampling species B, obtaining

$$\hat{p}_{+2} = \frac{O_{+2}}{n} = \frac{50}{150} = 0.333. \tag{20.48}$$

For species C, we obtain $\hat{p}_{+3} = 0.233$, while for species D we have $\hat{p}_{+4} = 0.267$.

What about the overall probability of being infected, across all species? Let the quantity p_{1+} stand for this probability, where '1' stands for the first row in Table 20.2, while + indicates the overall probability combining species A through D. We can estimate this probability by summing the infected individuals across all four species and dividing by the sample size n. If we let O_{1+} stand for this sum, we obtain

$$\hat{p}_{1+} = \frac{O_{1+}}{n} = \frac{61}{150} = 0.407. \tag{20.49}$$

This is just the row total of the infected amphibians divided by n. The overall probability of not being infected, p_{2+}, is estimated using the formula

$$\hat{p}_{2+} = \frac{O_{2+}}{n} = \frac{89}{150} = 0.593. \tag{20.50}$$

We are now in a position to calculate the expected frequencies under the null hypothesis of independence. If p_{11} is the probability of sampling an individual of species A that is infected, then if species and infection status are independent we can estimate this probability using

$$\hat{p}_{11} = \hat{p}_{1+}\hat{p}_{+1}. \tag{20.51}$$

The expected frequency for this cell, E_{11}, would be n times this probability, or

$$\begin{align} E_{11} &= n\hat{p}_{11} \tag{20.52}\\ &= n\hat{p}_{1+}\hat{p}_{+1} \tag{20.53}\\ &= n\frac{O_{1+}}{n}\frac{O_{+1}}{n} \tag{20.54}\\ &= \frac{O_{1+}O_{+1}}{n}. \tag{20.55} \end{align}$$

Thus, the expected frequency for this cell is the product of its column and row totals divided by the sample size. Using the Table 20.2 data, we find that

$$E_{11} = \frac{61(25)}{150} = 10.167. \tag{20.56}$$

All other cells are calculated in a similar manner. For example, we have

$$E_{13} = \frac{O_{1+}O_{+3}}{n} = \frac{61(35)}{150} = 14.233. \tag{20.57}$$

The remaining expected values are given in Table 20.2. The general formula for any cell would be

$$E_{ij} = \frac{O_{i+}O_{+j}}{n}. \tag{20.58}$$

This formula says that the expected value for any cell is the product of the row and column totals for that cell, divided by the sample size n.

Now suppose a particular two-way table has r rows and c columns. The likelihood ratio test statistic (G^2) for a test of independence is given by the general formula

$$G^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij}\ln(O_{ij}/E_{ij}). \tag{20.59}$$

G^2 has a χ^2 distribution under H_0 with $(r-1)(c-1)$ degrees of freedom. The explanation for the degrees of freedom is as follows (Agresti 1990). Under H_1, where the observations are not independent, the probability of an observation falling into a particular cell could be anything. Thus, there are rc values of p_{ij} that are free to vary except that they must sum to one, so there are $rc-1$ free parameters under H_1. Under H_0 there are r values of p_{i+} but only $r-1$ free to vary because these probabilities also sum to one. Similarly, there are $c-1$ values of p_{+j} free to vary. The difference in the number of free parameters under H_1 vs. H_0 is the degrees of freedom for the test, similar to goodness-of-fit tests. We therefore have

$$df = rc - 1 - (r-1) - (c-1) = rc - r - c + 1 = (r-1)(c-1). \tag{20.60}$$

The chi-square (X^2) statistic for a test of independence is given by the general formula

$$X^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}. \tag{20.61}$$

Under H_0, X^2 also has a χ^2 distribution with $(r-1)(c-1)$ degrees of freedom.

20.2.1 Test of independence - sample calculation

We illustrate these tests of independence using the Table 20.2 data, for which the expected frequencies have already been calculated. For the likelihood ratio test, we have

$$
\begin{aligned}
G^2 &= 2\sum_{i=1}^{r}\sum_{j=1}^{c} O_{ij}\ln(O_{ij}/E_{ij}) && (20.62)\\
&= 2[7\ln(7/10.167) + 12\ln(12/20.333) + 15\ln(15/14.233) && (20.63)\\
&+ 27\ln(27/16.267) + 18\ln(18/14.833) + 38\ln(38/29.667) && (20.64)\\
&+ 20\ln(20/20.767) + 13\ln(13/23.733)] && (20.65)\\
&= 2[-2.613 - 6.328 + 0.787 + 13.681 && (20.66)\\
&+ 3.483 + 9.407 - 0.753 - 7.825] && (20.67)\\
&= 2[9.839] && (20.68)\\
&= 19.678. && (20.69)
\end{aligned}
$$

There are $r = 2$ rows and $c = 4$ columns in the table, so the degrees of freedom are $(r-1)(c-1) = (2-1)(4-1) = 3$. From Table C, we see that the test was highly significant ($G^2 = 19.678, df = 3, P < 0.001$). This provides some evidence that species and infection status are not independent.

For the chi-square (X^2) version of this test, we have

$$
\begin{aligned}
X^2 &= \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij}-E_{ij})^2}{E_{ij}} && (20.70)\\
&= \frac{(7-10.167)^2}{10.167} + \frac{(12-20.333)^2}{20.333} + \frac{(15-14.233)^2}{14.233} && (20.71)\\
&+ \frac{(27-16.267)^2}{16.267} + \frac{(18-14.833)^2}{14.833} + \frac{(38-29.667)^2}{29.667} && (20.72)\\
&+ \frac{(20-20.767)^2}{20.767} + \frac{(13-23.733)^2}{23.733} && (20.73)\\
&= 0.987 + 3.415 + 0.041 + 7.082 + 0.676 + 2.341 && (20.74)\\
&+ 0.028 + 4.854 && (20.75)\\
&= 19.424. && (20.76)
\end{aligned}
$$

The test was also highly significant ($X^2 = 19.424, df = 3, P < 0.001$), similar to the likelihood ratio test.

20.2.2 Test of independence - SAS demo

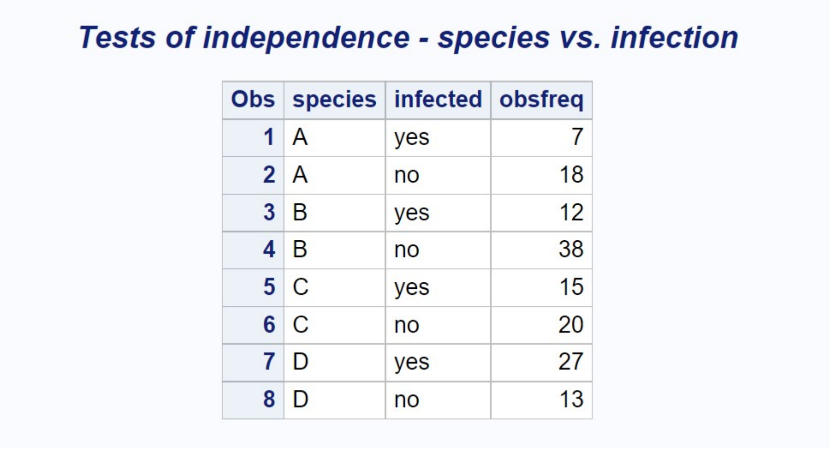
We can carry out the same calculations using SAS and `proc freq` (SAS Institute Inc. 2016). See program below. A two-way table of infection status and species is requested using the command `tables infected*species`. Likelihood ratio (G^2) and chi-square (X^2) tests are then requested using the `chisq` option. Because sample sizes are relatively small in this example, we can also request an exact version of both tests using the `exact chisq` option.

The option `out=percents outpct` requests an output data file called `percents` that contains various percentages, including the column percents from the two-way table. This file is used by `proc gchart` to generate a vertical bar chart with `species` on the x-axis (SAS Institute Inc. 2018). The percentage of infected and uninfected amphibians shown within each bar are generated using the option `subgroup=infected`.

Examining the SAS output in Fig. 20.9, we see that both tests were highly significant ($G^2 = 19.618, df = 3, P = 0.0002; X^2 = 19.425, df = 3, P = 0.0002$). The exact tests gave similar results in this case. The graph generated by `proc gchart` suggests that the infection rate is low for species A and B, intermediate for species C, and highest for species D (Fig. 20.10).

SAS Program

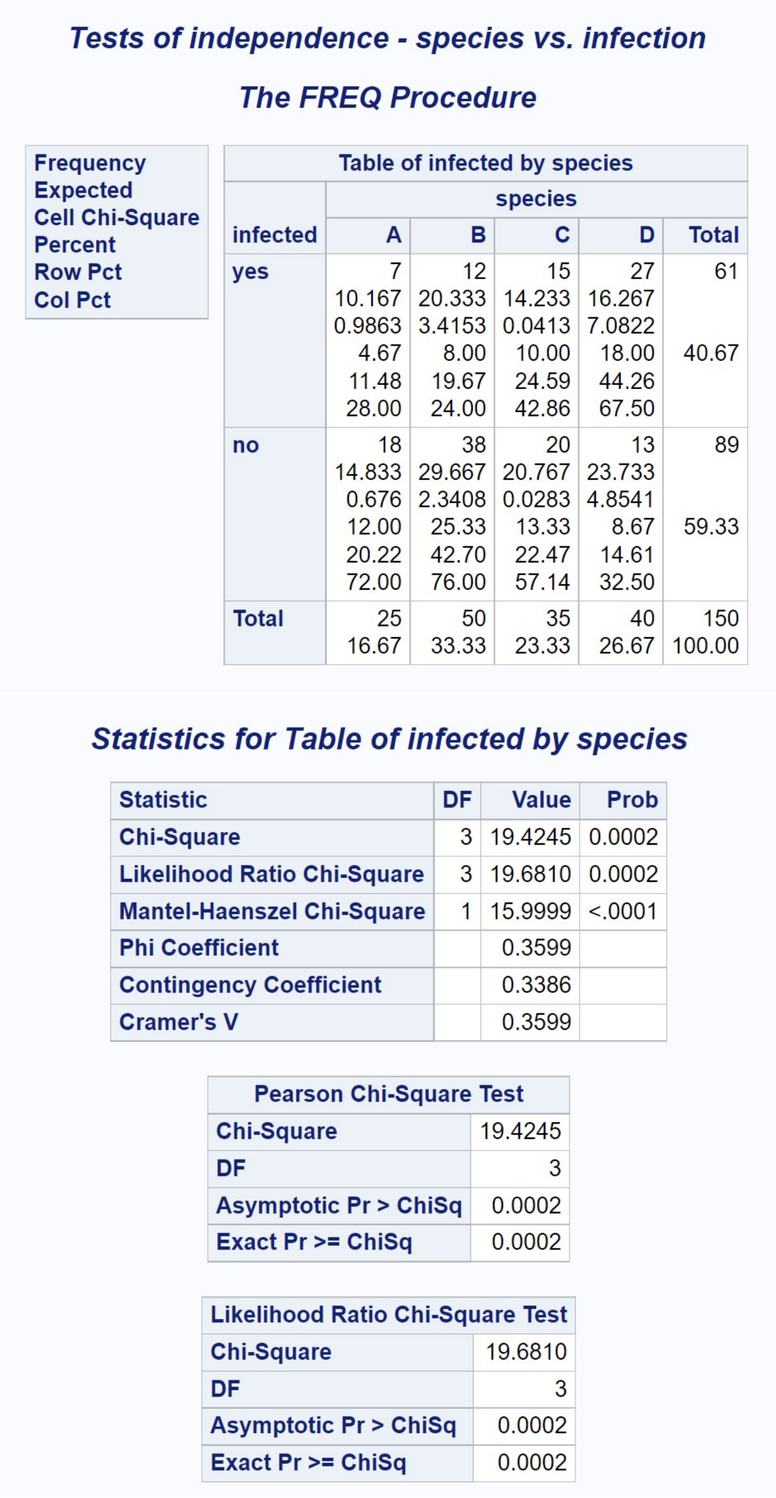
```
* chytrid.sas;
title "Tests of independence - species vs. infection";
data chytrid;
    input species $ infected $ obsfreq;
    datalines;
A  yes   7
A  no   18
B  yes  12
B  no   38
C  yes  15
C  no   20
D  yes  27
D  no   13
;
run;
* Print data set;
proc print data=chytrid;
run;
* Tests of independence;
proc freq data=chytrid order=data;
    tables infected*species / chisq cellchi2 expected out=percents outpct;
    weight obsfreq;
    * Can compute an exact test if frequencies are low;
    * Not recommended for large data sets;
    exact chisq;
run;
* Print output data file containing percents;
proc print data=percents;
run;
* Generate bar chart showing percentages;
proc gchart data=percents;
    vbar species / sumvar=pct_col subgroup=infected width=10 woutline=3
    raxis=axis1 maxis=axis2 legend=legend1;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
    axis2 label=(height=2) value=(height=2) width=3;
    legend1 label=(height=2) value=(height=2);
run;
quit;
```



Tests of independence - species vs. infection

Obs	species	infected	obsfreq
1	A	yes	7
2	A	no	18
3	B	yes	12
4	B	no	38
5	C	yes	15
6	C	no	20
7	D	yes	27
8	D	no	13

Figure 20.8: `chytrid.sas - proc print`



Tests of independence - species vs. infection

The FREQ Procedure

Frequency Expected Cell Chi-Square Percent Row Pct Col Pct

Table of infected by species					
	species				
infected	A	B	C	D	Total
yes	7 10.167 0.9863 4.67 11.48 28.00	12 20.333 3.4153 8.00 19.67 24.00	15 14.233 0.0413 10.00 24.59 42.86	27 16.267 7.0822 18.00 44.26 67.50	61 40.67
no	18 14.833 0.676 12.00 20.22 72.00	38 29.667 2.3408 25.33 42.70 76.00	20 20.767 0.0283 13.33 22.47 57.14	13 23.733 4.8541 8.67 14.61 32.50	89 59.33
Total	25 16.67	50 33.33	35 23.33	40 26.67	150 100.00

Statistics for Table of infected by species

Statistic	DF	Value	Prob
Chi-Square	3	19.4245	0.0002
Likelihood Ratio Chi-Square	3	19.6810	0.0002
Mantel-Haenszel Chi-Square	1	15.9999	<.0001
Phi Coefficient		0.3599	
Contingency Coefficient		0.3386	
Cramer's V		0.3599	

Pearson Chi-Square Test	
Chi-Square	19.4245
DF	3
Asymptotic Pr > ChiSq	0.0002
Exact Pr >= ChiSq	0.0002

Likelihood Ratio Chi-Square Test	
Chi-Square	19.6810
DF	3
Asymptotic Pr > ChiSq	0.0002
Exact Pr >= ChiSq	0.0002

Figure 20.9: `chytrid.sas - proc freq`

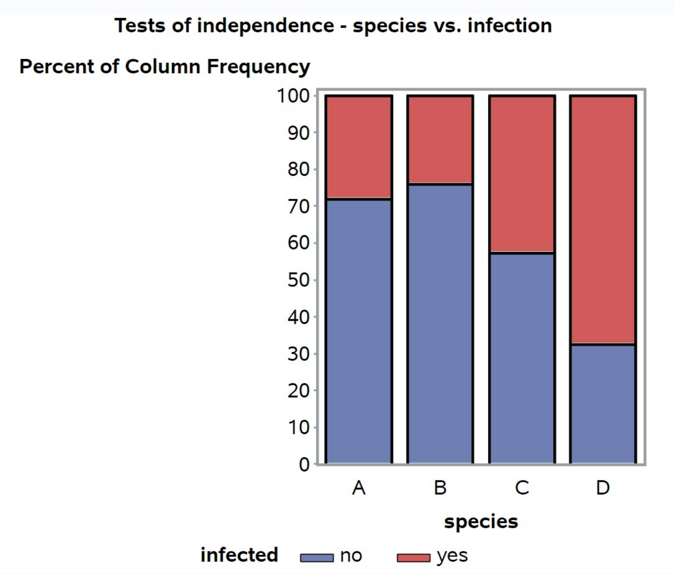


Figure 20.10: `chytrid.sas - gchart`

20.2.3 Test of independence - SAS demo 2

Ecologists often study the age structure of plant or animal populations, because this can provide clues about their birth and death rates. For example, a population with a higher proportion of young individuals could indicate the population is increasing through higher birth rates. Suppose that an ecologist wants to compare the age structure of three different populations of a bird species. One hundred individuals from each population are sampled and classified by age. There are five age classes, beginning with the nestlings (age 0) and individuals 1, 2, 3, or 4+ years old. See Table 20.4 for the results.

Table 20.4: Observed frequencies of age 0, 1, 2, 3, and 4 year old individuals for three different populations.

		Population		
Age class	1	2	3	$\sum$
0	36	48	60	144
1	22	24	21	67
2	18	14	12	44
3	13	10	12	28
4	11	4	2	17
$\sum$	100	100	100	300

These data were obtained using a sampling scheme that selected 100 individuals for each population, so that the column totals are fixed at 100 while the row totals are free to vary. This differs from the previous example (Table 20.2), where amphibians in general were sampled and the number of each species was a random quantity. It turns out the multinomial distribution can be used to describe both sampling methods, and the tests for independence are the same (Agresti 1990).

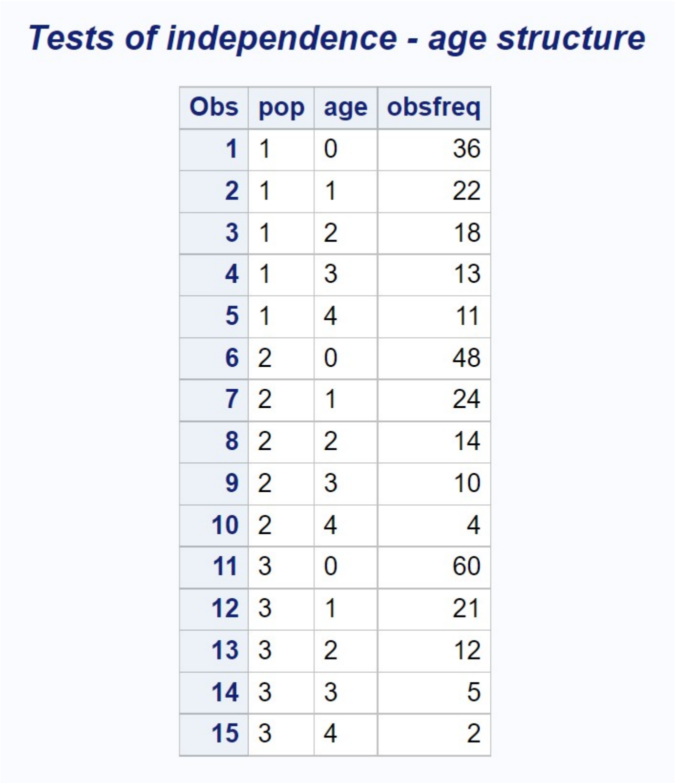
We will conduct tests of independence for these data using SAS and `proc freq` (see program below). As before, we will conduct both the likelihood ratio (G^2) and chi-square (X^2) tests. One difference in this program is that the option for exact tests is turned off, because they are quite time consuming (and unnecessary) for large data sets. An output file is used by `proc gchart` to generate a vertical bar chart with `pop` on the x-axis, with the divisions within each bar the percentages of each age group. These were generated using the option `subgroup=age`.

The likelihood ratio test of independence was significant ($G^2 = 18.920, df = 8, P = 0.0153$) as was the chi-square test ($X^2 = 18.864, df = 8, P = 0.0156$) (see Fig. 20.13). Examining the bar chart, we see that the percentage of younger individuals was lowest for population 1 and highest for population 3 (Fig. 20.14). One possible explanation is that population 3 has the highest birth rate while population 1 has the lowest.

SAS Program

```
* age_structure.sas;
title "Tests of independence - age structure";
data age;
    input pop $ age $ obsfreq;
    datalines;
1  0  36
1  1  22
1  2  18
1  3  13
1  4  11
2  0  48
2  1  24
2  2  14
2  3  10
2  4   4
3  0  60
3  1  21
3  2  12
3  3   5
3  4   2
;
run;
* Print data set;
proc print data=age;
run;
* Tests of independence;
proc freq data=age order=data;
    tables age*pop / chisq cellchi2 expected out=percents outpct;
    weight obsfreq;
    * Can compute an exact test if frequencies are low;
    * Not recommended for large data sets;
    *exact chisq;
run;
* Print output data file containing percents;
proc print data=percents;
run;
* Generate bar chart showing percentages;
proc gchart data=percents;
    vbar pop / sumvar=pct_col subgroup=age width=10 woutline=3
    raxis=axis1 maxis=axis2 legend=legend1;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
    axis2 label=(height=2) value=(height=2) width=3;
    legend1 label=(height=2) value=(height=2);
```

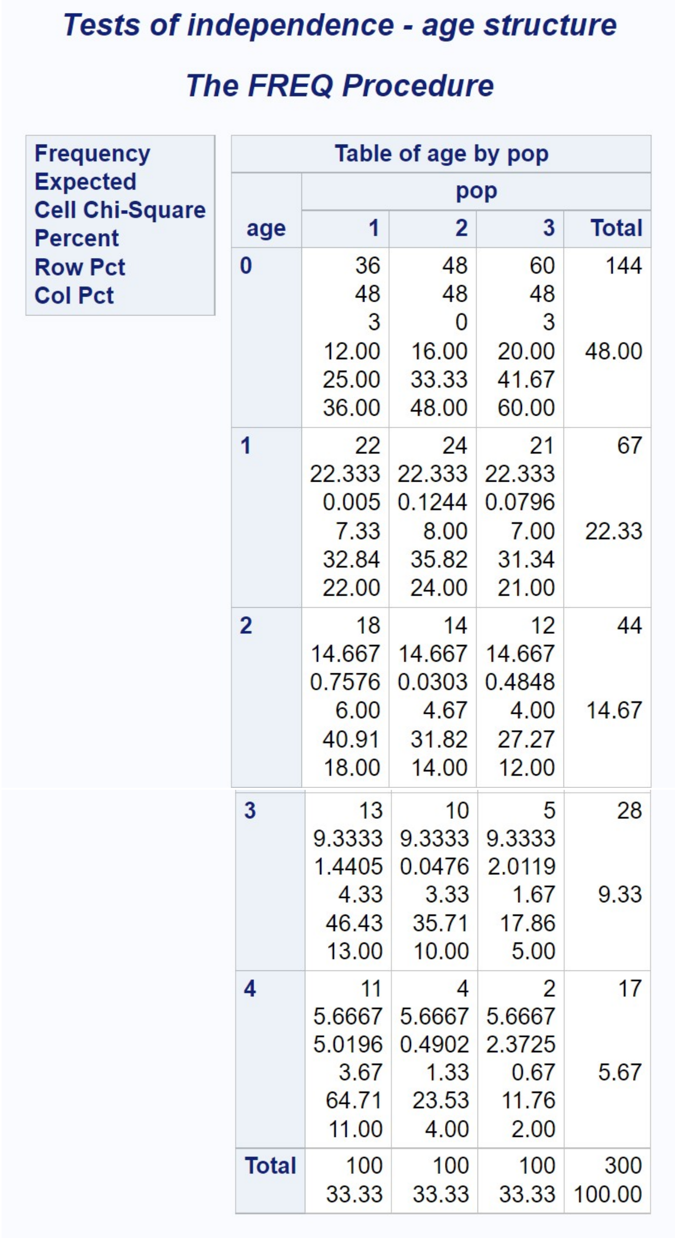
```
run;
quit;
```



Tests of independence - age structure

Obs	pop	age	obsfreq
1	1	0	36
2	1	1	22
3	1	2	18
4	1	3	13
5	1	4	11
6	2	0	48
7	2	1	24
8	2	2	14
9	2	3	10
10	2	4	4
11	3	0	60
12	3	1	21
13	3	2	12
14	3	3	5
15	3	4	2

Figure 20.11: `age_structure.sas - proc print`



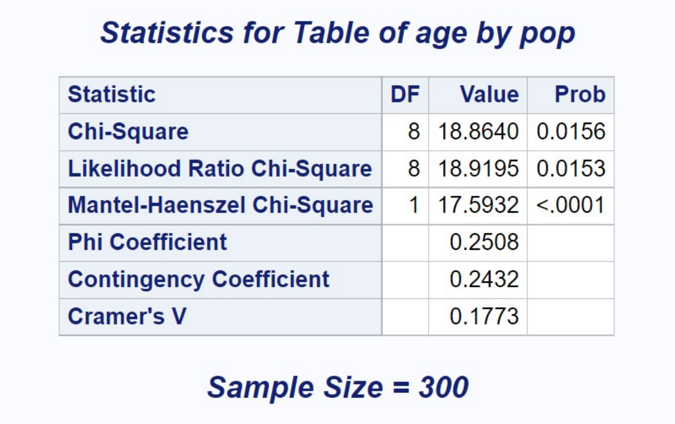
Tests of independence - age structure

The FREQ Procedure

Frequency Expected Cell Chi-Square Percent Row Pct Col Pct

Table of age by pop				
	pop			
age	1	2	3	Total
0	36 48 3 12.00 25.00 36.00	48 48 0 16.00 33.33 48.00	60 48 3 20.00 41.67 60.00	144 48.00
1	22 22.333 0.005 7.33 32.84 22.00	24 22.333 0.1244 8.00 35.82 24.00	21 22.333 0.0796 7.00 31.34 21.00	67 22.33
2	18 14.667 0.7576 6.00 40.91 18.00	14 14.667 0.0303 4.67 31.82 14.00	12 14.667 0.4848 4.00 27.27 12.00	44 14.67
3	13 9.3333 1.4405 4.33 46.43 13.00	10 9.3333 0.0476 3.33 35.71 10.00	5 9.3333 2.0119 1.67 17.86 5.00	28 9.33
4	11 5.6667 5.0196 3.67 64.71 11.00	4 5.6667 0.4902 1.33 23.53 4.00	2 5.6667 2.3725 0.67 11.76 2.00	17 5.67
Total	100 33.33	100 33.33	100 33.33	300 100.00

Figure 20.12: `age_structure.sas - proc freq`



Statistics for Table of age by pop

Statistic	DF	Value	Prob
Chi-Square	8	18.8640	0.0156
Likelihood Ratio Chi-Square	8	18.9195	0.0153
Mantel-Haenszel Chi-Square	1	17.5932	<.0001
Phi Coefficient		0.2508	
Contingency Coefficient		0.2432	
Cramer's V		0.1773	

Sample Size = 300

Figure 20.13: `age_structure.sas - proc freq`

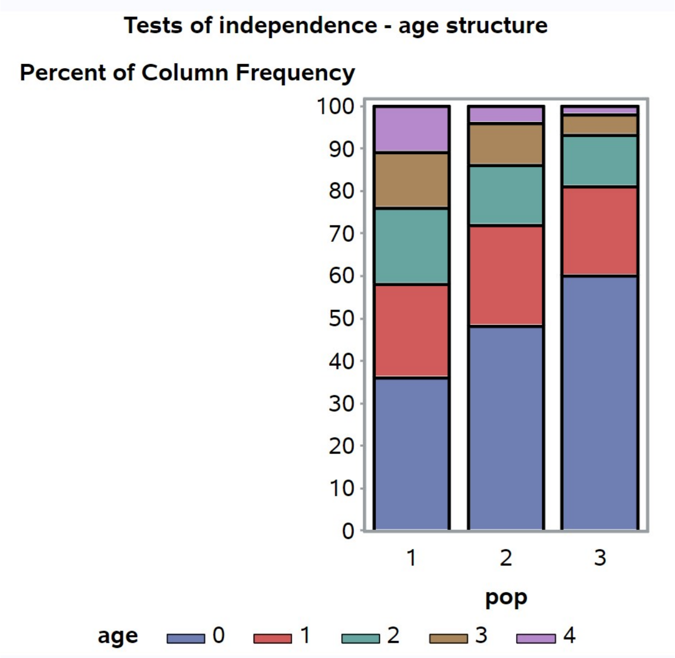


Figure 20.14: `age_structure.sas - proc gchart`

20.3 References

Agresti, A. (1990). *Categorical Data Analysis.* John Wiley & Sons, New York, NY.

Mendel, G. (1865) Experiments in plant hybridization. http://www.mendelweb.org/Mendel.html

Reeve, J. D., Rojas, M. G. & Morales-Ramos, J. A. (2003) Artificial diet and rearing methods for *Thanasimus dubius* (Coleoptera: Cleridae), a predator of bark beetles (Coleoptera: Scolytidae). *Biological Control* 27: 315-322.

SAS Institute Inc. (2016) *Base SAS 9.4 Procedures Guide, Sixth Edition.* SAS Institute Inc., Cary, NC.

SAS Institute Inc. (2016) *SAS/GRAPH 9.4: Reference, Fifth Edition.* SAS Institute Inc., Cary, NC.

20.4 Problems

1. An ecologist wants to characterize the spatial distribution of an uncommon plant species in the forest. One hundred quadrats are established and the number of plants counted in each quadrat. The following data were obtained:

Plants	Frequency
0	42
1	23
2	12
3	8
4	4
5	3
6	3
7	2
8	1
9	1
10	0
11	0
12	0

Test whether these data have a Poisson distribution, using both likelihood ratio (G^2) and X^2 (χ^2) tests, using the program `Poisson_fit2_gof.sas` to help with the calculations. Discuss your results. Do the data appear to be Poisson, overdispersed, or underdispersed?

2. Some species of snakes can imitate a rattlesnake and thereby avoid being eaten by predators, a phenomenon known as Batesian mimicry. Individuals of one such species were randomly selected from locations where rattlesnakes were absent, at moderate density, and at high density. Each snake was then scored for whether or not it imitated a rattlesnake when disturbed. The following results were obtained.

		Rattlesnake density	
Imitated a rattlesnake?	Absent	Moderate	High
Yes	65	76	82
No	35	24	18

   (a) Test if imitation of a rattlesnake is independent of rattlesnake density using a manual likelihood ratio (G^2) test. Show your calculations.

   (b) Test if imitation of a rattlesnake is independent of rattlesnake density using a manual $X^2(\chi^2)$ test. Show your calculations.

   (c) Check your above answers by having SAS carry out the same two tests.

   (d) Interpret the results of your tests. Does the frequency of rattlesnake imitation vary significantly with the density of rattlesnakes, and if so what is the pattern?

Chapter 21

Multiple Regression

Multiple regression is a statistical technique for examining the relationship between a dependent variable Y and multiple independent variables or regressors $X_1, X_2, \ldots, X_k$. Like with linear regression, the independent variables or regressors may be fixed values under experimental control, or random variables. One purpose of multiple regression is to determine whether changes in any of the independent variables cause changes in Y. This involves testing whether the slope β_j for a given independent variable X_j is significantly different from zero, for each of the independent variables. There is also an overall test that examines whether any of independent variables (alone or in combination) affect Y. Another purpose of multiple regression is prediction, using a set of values for the independent variable to predict the value of Y along with a confidence interval. A third use is model selection. The objective here is to find a model that approximates the data with the fewest variables, involving a trade-off between model fit and model complexity. We will examine a popular method of model selection that uses Akaike's Information Criterion or *AIC* (Akaike 1974; Anderson et al. 2000, Burnham & Anderson 2002).

We will first illustrate multiple regression using a relatively simple data set from a study of southern pine beetle, *Dendroctonus frontalis* (Reeve et al. 1998). We previously used this study to examine the relationship between the number of beetles added to caged trees and how this affected their attack density. We now examine how attack density and the density of a competitor, bluestain fungus, affects the survival rate of beetle offspring (from egg to emerging adult). High attack densities imply a high density of adult beetles within the tree, and this crowding could reduce survival of their offspring (see also Coulson et al. 1976). High levels of bluestain fungus are also known to reduce survival, by interfering with the beetle's own symbiotic fungus (Hofstetter et al. 2006).

Table 21.1: Example 1 - Effects of attack density and bluestain fungus on the survival of *D. frontalis* brood from egg to emergence (Reeve et al. 1998). The dependent variable was the log-transformed survival rate of the beetle offspring, while attack density (attacks per 100 cm^2 of bark) and the proportion of bluestained phloem were the independent variables.

X_{1i} = Attack density	X_{2i} = Bluestain	Survival	Y_i = ln(Survival)	i
1.250	0.000	0.107	-2.235	1
2.656	0.481	0.715	-0.335	2
7.334	0.171	0.036	-3.324	3
1.603	0.352	0.188	-1.671	4
2.622	0.016	0.438	-0.826	5
1.000	0.000	0.585	-0.536	6
4.342	0.185	0.115	-2.163	7
5.233	0.018	0.257	-1.359	8
2.500	0.410	0.032	-3.442	9
3.250	0.015	0.350	-1.050	10
6.000	0.007	0.161	-1.826	11
4.750	0.000	0.073	-2.617	12
2.500	0.095	0.219	-1.519	13
8.750	0.033	0.028	-3.576	14
6.000	0.015	0.294	-1.224	15
5.000	0.105	0.207	-1.575	16
7.149	0.025	0.227	-1.483	17
6.750	0.015	0.040	-3.219	18
7.500	0.043	0.089	-2.419	19
2.500	0.073	0.176	-1.737	20
5.000	0.055	0.084	-2.477	21
2.250	0.023	0.203	-1.595	22
1.250	0.123	0.074	-2.604	23
4.750	0.035	0.126	-2.071	24
4.500	0.212	0.290	-1.238	25
9.557	0.166	0.010	-4.605	26
5.000	0.338	0.207	-1.575	27

We will use another data set to illustrate prediction in multiple regression. Soul et al. (2013) were interested in predicting endocranial volume (brain size) in extinct mammals, where only the skull length, height, and width are available. For this purpose, they developed a multiple regression model using existing species as the observations, with endocranial volume the dependent variable, and skull length, width and height the independent ones. A portion of these observations are listed below (see https://datadryad.org for the full data set). We will fit a multiple regression model to these observations, then use them to predict endocranial volume for two hypothetical fossils, a mouse and a bear.

Table 21.2: Example 2 - Skull length, width, height, endocranial volume, and species name (Soul et al. 2013). The dependent variable was endocranial volume, estimated using the mass of glass beads filling the skull.

Length (mm)	Width (mm)	Height (mm)	Volume (g)	i	Common name
15.04	11.29	6.61	0.38	1	Pygmy glider
52.40	30.94	25.68	12.36	2	Rufous kangaroo rat
75.87	52.79	39.45	56.70	3	Howler monkey
41.73	25.70	16.79	5.68	4	Scaley-tailed squirrel
39.71	26.87	17.13	5.92	5	Lord derby's flying squirrel
18.90	12.62	7.61	0.51	6	Yellow-footed antechinus
15.10	11.69	7.06	0.46	7	Brown antechinus
123.70	73.89	63.93	150.53	8	Pronghorn
46.75	28.70	18.45	6.51	9	Mountain beaver
154.32	103.77	71.95	284.03	10	Antarctic fur seal
133.39	59.75	72.60	128.49	11	Babiroussa
etc.					
32.90	19.83	14.73	3.19	185	Tree shrew
32.15	20.33	13.95	3.17	186	Painted tree shrew
200.23	98.99	84.53	358.82	187	Brown bear
179.70	95.48	75.51	302.72	188	Sloth bear
67.48	42.35	29.66	24.79	189	Ruffled lemur
70.78	30.98	28.08	17.91	190	Rasse
67.05	54.15	44.99	56.71	191	Wombat
70.36	45.09	37.72	38.43	192	Arctic fox
80.73	47.96	39.45	48.55	193	Fox
13.54	9.24	7.13	0.36	194	Meadow jumping mouse
13.15	9.05	7.00	-	195	Fossil mouse
190.17	97.32	80.31	-	196	Fossil bear

21.1 Multiple regression model

Suppose we want to model the observations for a data set like Example 1, where a dependent variable Y is observed along with two independent variables X_1 and X_2. Let Y_i, X_{1i}, and X_{2i} be the *ith* set of values. The multiple regression model takes the form

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i, \tag{21.1}$$

where β_0 is the intercept, β_1 and β_2 are the slopes or regression coefficients for X_1 and X_2, and $\epsilon_i \sim N(0, \sigma^2)$ (Draper & Smith 1981; Kutner et al. 2005, Sheather 2009). This equation defines a plane in three dimensions, which we will later visualize for the Example 1 data.

More generally, the model for k different independent variables $X_1, X_2, \ldots, X_k$ takes the form

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \epsilon_i. \tag{21.2}$$

Here, the parameters $\beta_1, \beta_2, \ldots, \beta_k$ are the slopes for each independent variable. While this model appears complicated, there is a simple interpretation of the regression coefficients. **The slope β_j can be thought of as the change in Y per unit change in X_j, while holding all the other variables constant.** There are also specific plots designed to visualize this model for any number of independent variables.

21.2 Multiple regression in matrix form

We will now show how the multiple regression model can be expressed in matrix form (Draper & Smith 1981). This will greatly simplify later developments, and in any event the matrix form of the model is commonly used in the statistical literature as well as software documentation. If you are unfamiliar with matrices, there are many online resources that provide an introduction to matrices and linear algebra. The textbook by Tabachnik and Fidell (2001) also provides a useful summary of essential concepts (see their Appendix A). In the following, we will briefly review various matrix operations and then apply them to multiple regression. Chapter 24 of this text lists a SAS program that carries out these operations using `proc iml` (SAS Institute Inc. 2018a).

A matrix is a rectangular collection of numbers (or other quantities) arranged in rows and columns, enclosed in a set of parentheses or brackets. A vector is a simple type of matrix consisting of a single column or row of numbers. Matrices and vectors can be added, multiplied, transposed, and even inverted in their own unique way, and these operations allow one to express the multiple regression model in a compact way as well as estimate the parameters of this model.

We will first make use of **matrix addition** and **multiplication** to write the multiple regression model. Suppose we have two vectors $\boldsymbol{A}$ and $\boldsymbol{B}$ of the following form:

$$\boldsymbol{A} = \begin{pmatrix} a \\ b \\ c \end{pmatrix} \text{ and } \boldsymbol{B} = \begin{pmatrix} d \\ e \\ f \end{pmatrix}. \tag{21.3}$$

To add these two vectors, we simply add the elements of each one to obtain

$$\boldsymbol{A} + \boldsymbol{B} = \begin{pmatrix} a+d \\ b+e \\ c+f \end{pmatrix}. \tag{21.4}$$

For example, suppose

$$\boldsymbol{A} = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \text{ and } \boldsymbol{B} = \begin{pmatrix} 4 \\ 5 \\ 6 \end{pmatrix}. \tag{21.5}$$

Then

$$\boldsymbol{A} + \boldsymbol{B} = \begin{pmatrix} 1+4 \\ 2+5 \\ 3+6 \end{pmatrix} = \begin{pmatrix} 5 \\ 7 \\ 9 \end{pmatrix}. \tag{21.6}$$

Note that the two vectors (or matrices) must have the same dimensions or shape for addition to work.

For the multiple regression model, we will also need to multiply a matrix by a vector. Suppose that we have two matrices $\boldsymbol{C}$ and $\boldsymbol{D}$ of the following form:

$$\boldsymbol{C} = \begin{pmatrix} a & d \\ b & e \\ c & f \end{pmatrix} \text{ and } \boldsymbol{D} = \begin{pmatrix} g \\ h \end{pmatrix}. \tag{21.7}$$

Then

$$\boldsymbol{CD} = \begin{pmatrix} a & d \\ b & e \\ c & f \end{pmatrix} \times \begin{pmatrix} g \\ h \end{pmatrix} = \begin{pmatrix} ag + dh \\ bg + eh \\ cg + fh \end{pmatrix}. \tag{21.8}$$

Note the pattern in the multiplication process. You take the elements in each row of $\boldsymbol{C}$ and multiply them by the column elements of $\boldsymbol{D}$, then add the result to obtain $\boldsymbol{CD}$. In this case, the multiplication process takes a 3 × 2 matrix (3 rows and 2 columns) and a 2 × 1 matrix, and produces a 3 × 1 matrix. Thus, the numbers of rows and columns in the product depends on the number of rows in first matrix and columns in the second – this is true of matrix multiplication in general. The number of columns in the first matrix and rows in the second matrix must also match for matrix multiplication to be possible.

As an example of matrix multiplication, suppose that

$$\boldsymbol{C} = \begin{pmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{pmatrix} \text{ and } \boldsymbol{D} = \begin{pmatrix} 7 \\ 8 \end{pmatrix}. \tag{21.9}$$

Then

$$\boldsymbol{CD} = \begin{pmatrix} 1 \cdot 7 + 4 \cdot 8 \\ 2 \cdot 7 + 5 \cdot 8 \\ 3 \cdot 7 + 6 \cdot 8 \end{pmatrix} = \begin{pmatrix} 39 \\ 54 \\ 69 \end{pmatrix}. \tag{21.10}$$

Another matrix operation we will use later is the **transpose** of a matrix. This operation takes the columns of a matrix and turns them into the rows of a new matrix. For example, suppose we have a matrix

$$\boldsymbol{F} = \begin{pmatrix} a & e \\ b & f \\ c & g \\ d & h \end{pmatrix}. \tag{21.11}$$

The transpose of $\boldsymbol{F}$ (written as $\boldsymbol{F'}$) is defined to be

$$\boldsymbol{F'} = \begin{pmatrix} a & b & c & d \\ e & f & g & h \end{pmatrix}. \tag{21.12}$$

For example, suppose

$$\boldsymbol{F} = \begin{pmatrix} 1 & 5 \\ 2 & 6 \\ 3 & 7 \\ 4 & 8 \end{pmatrix}. \tag{21.13}$$

Then

$$\boldsymbol{F'} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{pmatrix}. \tag{21.14}$$

Now suppose we have a multiple regression problem with $k = 2$ independent variables and n observations, similar to the Example 1 data set. The standard model equation for this problem would be

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i, \tag{21.15}$$

for $i = 1$ to n. If we write out the full system of equations for each observation or value of i, we would obtain n different equations:

$$\begin{pmatrix} Y_1 = \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} + \epsilon_1 \\ Y_2 = \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} + \epsilon_2 \\ Y_3 = \beta_0 + \beta_1 X_{13} + \beta_2 X_{23} + \epsilon_3 \\ \vdots \\ Y_n = \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} + \epsilon_n \end{pmatrix}. \tag{21.16}$$

Using the definition of matrix addition, these equations can be rewritten in matrix form as

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \beta_0 + \beta_1 X_{11} + \beta_2 X_{21} \\ \beta_0 + \beta_1 X_{12} + \beta_2 X_{22} \\ \beta_0 + \beta_1 X_{13} + \beta_2 X_{23} \\ \vdots \\ \beta_0 + \beta_1 X_{1n} + \beta_2 X_{2n} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}. \tag{21.17}$$

Using the definition of matrix multiplication, a further simplification is possible:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{21} \\ 1 & X_{12} & X_{22} \\ 1 & X_{13} & X_{23} \\ \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}. \tag{21.18}$$

As a final step, this equation can be written in the form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{21.19}$$

where

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}, \text{and} \tag{21.20}$$

$$\boldsymbol{X} = \begin{pmatrix} 1 & X_{11} & X_{21} \\ 1 & X_{12} & X_{22} \\ 1 & X_{13} & X_{23} \\ \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} \end{pmatrix} \tag{21.21}$$

For an actual data set, these matrices and vectors would contain the values of Y, X_{1i}, and X_{2i}. The matrix $\boldsymbol{X}$ is often called the **design matrix**, because it basically describes the design of the study, including the values of the independent variables, their number, and the overall sample size.

In general, the multiple regression model for k independent variables or regressors can be expressed in the same simple form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{21.22}$$

where

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}, \text{and} \tag{21.23}$$

$$\boldsymbol{X} = \begin{pmatrix} 1 & X_{11} & X_{21} & \dots & X_{k1} \\ 1 & X_{12} & X_{22} & \dots & X_{k2} \\ 1 & X_{13} & X_{23} & \dots & X_{k3} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{1n} & X_{2n} & \dots & X_{kn} \end{pmatrix}. \tag{21.24}$$

21.3 Multiple regression and likelihood

We will use maximum likelihood to estimate the parameters in the multiple regression model, making use of the matrix form of the model. Suppose we have $k = 2$ independent variables similar to the Example 1 data. The multiple regression model in this case would be

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i. \tag{21.25}$$

This model has four parameters to estimate, in particular β_0, β_1, β_2, and σ^2. Consider the first observation in the Example 1 data, for which $Y_1 = -2.235, X_{11} = 1.250$, and $X_{21} = 0.000$. For this observation, the model states that $Y_1 \sim N(\beta_0 + \beta_1 X_{11} + \beta_2 X_{21}, \sigma^2)$, and so the likelihood would be

$$L_1 = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(Y_1-(\beta_0+\beta_1 X_{11}+\beta_2 X_{21}))^2}{\sigma^2}} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\frac{(-2.235-(\beta_0+\beta_1 1.25+\beta_2 0.000))^2}{\sigma^2}} \tag{21.26}$$

The overall likelihood is then defined as the product of the likelihoods for each observation, in particular

$$L(\beta_0, \beta_1, \beta_2, \sigma^2) = L_1 \times L_2 \times \ldots \times L_n. \tag{21.27}$$

Finding the maximum likelihood estimates involves maximizing this quantity with respect to the parameters β_0, β_1, β_2, and σ^2. Similar to linear regression, we can gain some insight into the estimation process by rearranging the likelihood function. It can be written in the form

$$L(\beta_0, \beta_1, \beta_2, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2}\frac{\sum_{i=1}^n (Y_i-(\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}))^2}{\sigma^2}}. \tag{21.28}$$

Focusing on the sum in this expression, we see that values of β_0, β_1, and β_2 that minimize the sum of the squared terms will maximize the overall likelihood. Similar to linear regression, these are also the **least squares** estimates because they minimize the sum of these squared terms (Draper & Smith 1981). We will later see that they minimize the sum of the squared residuals from the plane defined by β_0, β_1, and β_2.

Now consider the case where there are k independent variables, so that the model has $k + 2$ parameters $(\beta_0, \beta_1, \beta_2, \ldots, \beta_k, \sigma^2)$. The likelihood L would have the same structure as above, but with more parameters and independent variables. The maximum likelihood estimates can be found by taking the derivative of L (actually $\log L$) with respect to every parameter, setting these derivatives equal to zero, then solving for the parameter values that satisfy these equations. The result is a complex system of equations involving the data set and parameters. Using matrix algebra, however, the equations for $\beta_0, \beta_1, \ldots, \beta_k$ can expressed in a very compact form:

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{Y} \tag{21.29}$$

Here $\boldsymbol{X}, \boldsymbol{\beta}$, and $\boldsymbol{Y}$ are from the matrix version of the multiple regression model. The idea then is solve this equation for $\boldsymbol{\beta}$ using matrix operations. This set of equations are called the **normal equations** (Draper and Smith 1981; Kutner et al 2005; Sheather 2009). They look a bit like the simple equation $xb = y$, where x and y are known values. You would solve this equation for b by multiplying both sides by x^{-1}, to obtain $x^{-1}xb = x^{-1}y$, or $b = x^{-1}y = y/x$. What we need is the matrix equivalent of x^{-1}.

Note that the inverse of x has the property $x^{-1}x = 1$. The inverse of a matrix has the same property, but the equivalent of the number 1 is called the **identity matrix**, written as $\boldsymbol{I}$. It is defined as a square matrix with ones on the diagonal and zeroes everywhere else. For example, the 3×3 identity matrix is

$$\boldsymbol{I} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}. \tag{21.30}$$

Similar to the number 1, if you multiply a matrix $\boldsymbol{A}$ by $\boldsymbol{I}$ the result is equal to $\boldsymbol{A}$. For example, suppose that $\boldsymbol{A}$ is defined by the matrix

$$\boldsymbol{A} = \begin{pmatrix} 1 & 6 & 4 \\ 3 & 7 & 6 \\ 4 & 1 & 9 \end{pmatrix}. \tag{21.31}$$

Then we have

$$\boldsymbol{A}\boldsymbol{I} = \begin{pmatrix} 1 & 6 & 4 \\ 3 & 7 & 6 \\ 4 & 1 & 9 \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{21.32}$$

$$= \begin{pmatrix} 1\cdot 1 + 6\cdot 0 + 4\cdot 0 & 1\cdot 0 + 6\cdot 1 + 4\cdot 0 & 1\cdot 0 + 6\cdot 0 + 4\cdot 1 \\ 3\cdot 1 + 7\cdot 0 + 6\cdot 0 & 3\cdot 0 + 7\cdot 1 + 6\cdot 0 & 3\cdot 0 + 7\cdot 0 + 6\cdot 1 \\ 4\cdot 1 + 1\cdot 0 + 9\cdot 0 & 4\cdot 0 + 1\cdot 1 + 9\cdot 0 & 4\cdot 0 + 1\cdot 0 + 9\cdot 1 \end{pmatrix}$$

$$= \begin{pmatrix} 1 & 6 & 4 \\ 3 & 7 & 6 \\ 4 & 1 & 9 \end{pmatrix} = \boldsymbol{A}$$

Now we can define the inverse of a matrix for a square matrix like $\boldsymbol{A}$. The inverse of $\boldsymbol{A}$, written as $\boldsymbol{A^{-1}}$, is a matrix for which $\boldsymbol{A^{-1}A} = I$ and also $\boldsymbol{AA^{-1}} = I$. Note that the order of multiplication does not matter in this case, although it would for other types of matrices. There are a number of numerical techniques for finding the inverse of a matrix, but we will not be concerned with these details. The inverse of $\boldsymbol{A}$ is the matrix

$$\boldsymbol{A^{-1}} = \begin{pmatrix} -0.934 & 0.820 & -0.131 \\ 0.049 & 0.115 & -0.098 \\ 0.410 & -0.377 & 0.180 \end{pmatrix}. \tag{21.33}$$

Multiplying $\boldsymbol{A^{-1}}$ and $\boldsymbol{A}$, we obtain

$$\boldsymbol{A^{-1}A} = \begin{pmatrix} -0.934 \cdot 1 + 0.820 \cdot 3 - 0.131 \cdot 4 & \dots & \dots \\ 0.049 \cdot 1 + 0.115 \cdot 3 - 0.098 \cdot 4 & \dots & \dots \\ 0.041 \cdot 1 - 0.377 \cdot 3 + 0.180 \cdot 4 & \dots & \dots \end{pmatrix} \tag{21.34}$$

$$= \begin{pmatrix} 1.002 & 0.005 & 0.005 \\ 0.002 & 1.001 & 0.004 \\ -0.001 & 0.001 & 0.998 \end{pmatrix} \approx \boldsymbol{I}. \tag{21.35}$$

The result is not exact because of rounding in the values of $\boldsymbol{A^{-1}}$.

We are now ready to solve the normal equations for $\boldsymbol{\beta}$ using matrix operations. Recall that these equations are of the form

$$\boldsymbol{X'X\beta = X'Y} \tag{21.36}$$

Multiplying both sides of this equation by the inverse of $\boldsymbol{X'X}$, denoted by $\boldsymbol{(X'X)^{-1}}$, we obtain

$$\boldsymbol{(X'X)^{-1}X'X\beta = (X'X)^{-1}X'Y} \tag{21.37}$$

or

$$\boldsymbol{I\beta = (X'X)^{-1}X'Y} \tag{21.38}$$

from which it follows that

$$\boldsymbol{\hat{\beta} = (X'X)^{-1}X'Y} \tag{21.39}$$

Here $\boldsymbol{\hat{\beta}}$ is a vector containing the maximum likelihood (or least squares) estimates $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_k$ of the model parameters, except for σ^2. This is the method used by SAS and other statistical packages to estimate the model parameters. We will later see how the elements of $\boldsymbol{(X'X)^{-1}}$ are also used to calculate standard errors and confidence intervals. Similar methods are used to estimate the parameters for ANOVA models. In this case, the design matrix encodes the various treatment combinations and interactions.

The estimates of the model parameters can be used to generated a predicted value for each observation in the data set, of the form

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \ldots + \hat{\beta}_k X_{ki}. \tag{21.40}$$

The residual of each observation is the difference between the observed and predicted values, namely $Y_i - \hat{Y}_i$.

Maximum likelihood also provides an estimator of σ^2 similar to linear regression. Define an error sum of squares by the equation

$$SS_{error} = \sum_{i=1}^{n} \Big(Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \ldots + \hat{\beta}_k X_{ki})\Big)^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2. \tag{21.41}$$

The multiple regression form of MS_{error}, and an estimator of σ^2, is obtained by dividing SS_{error} by $n - k - 1$ degrees of freedom:

$$MS_{error} = \frac{SS_{error}}{n - k - 1} = \hat{\sigma}^2. \tag{21.42}$$

$SS_{regression}$ describes variation in the data explained by the regression model, similar to linear regression. It is defined as

$$SS_{regression} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \tag{21.43}$$

and has k degrees of freedom. We therefore have

$$MS_{regression} = \frac{SS_{regression}}{k}. \tag{21.44}$$

$SS_{regression}$ and $MS_{regression}$ will be large if $\hat{Y}_i$ varies strongly with respect to one or more of the independent variables $(X_{1i}, X_{2i}, \ldots, X_{ki})$.

The total sum of squares for multiple regression is defined as

$$SS_{total} = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 \tag{21.45}$$

and has $n - 1$ degrees of freedom. Similar to linear regression, there is an additive relationship among the different sums of squares:

$$SS_{regression} + SS_{error} = SS_{total}. \tag{21.46}$$

We can use the two mean squares to construct an overall F test for the multiple regression, which tests $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$. This null hypothesis basically says none of the independent variables $(X_{1i}, \ldots, X_{ki})$ affect the dependent one (Y_i). The alternative hypothesis is that one or more slopes are different from zero ($H_1 : \beta_j \neq 0$ for some j). If this test is significant, it suggests one or more of the independent variables are affecting the dependent variable, but not which ones. The test statistic is

$$F_s = \frac{MS_{regression}}{MS_{error}}. \tag{21.47}$$

Under H_0, F_s has an F distribution with $df_1 = k$ and $df_2 = n - k - 1$ the degrees of freedom. Note that we encountered a similar test in the SAS output for ANOVA designs, but in ANOVA we were more concerned with tests of each treatment effect, not in testing the overall model. It is also a likelihood ratio test using the H_0 and H_1 models for the data (McCulloch & Searle 2001).

We can organize the different sum of squares and mean squares into an ANOVA table for multiple regression (Table 21.3). It lists the different sources of variation in the data (regression, error, and total), their degrees of freedom, as well as the overall F test.

Table 21.3: General ANOVA table for multiple regression, showing formulas for different mean squares and the overall F test.

Source	*df*	Sum of squares	Mean square	F_s
Regression	k	$SS_{regression}$	$MS_{regression} = SS_{regression}/k$	$MS_{regression}/MS_{error}$
Error	$n-k-1$	SS_{error}	$MS_{error} = SS_{error}/(n-k-1)$	
Total	$n-1$	SS_{total}		

21.4 Tests and confidence intervals for $\boldsymbol{\beta}$

We next develop tests and confidence intervals for the parameters of the multiple regression model, in particular the slope parameters $\beta_1, \beta_2, \ldots, \beta_k$ and also the intercept β_0. These will help us evaluate which (if any) of the independent variables affect the dependent one. These tests and confidence intervals are based on the maximum likelihood estimates of each β_j and its standard error s_{β_j}, given by the formula

$$s_{\hat{\beta}_j} = \sqrt{\hat{\sigma}^2 d_{j+1,j+1}}, \tag{21.48}$$

for $j = 0, 1, \ldots, k$. Here $\hat{\sigma}^2 = MS_{error}$ and $d_{j+1,j+1}$ is the entry in the $(j+1)th$ row and column of the matrix $\boldsymbol{(X'X)^{-1}}$, i.e., the diagonal entries of this matrix (Draper & Smith 1981). For example, for β_0 and $j = 0$ we would use $d_{0+1,0+1} = d_{11}$, the entry in the first row and column. It can then be shown that the quantity

$$\frac{\hat{\beta}_j - \beta_j}{s_{\hat{\beta}_j}} \tag{21.49}$$

has a t distribution with $n-k-1$ degrees of freedom, the same as for MS_{error}. This fact can be used to derive tests and confidence intervals for each β_j.

Suppose we want to test $H_0 : \beta_j = \beta_{j0}$ vs. $H_1 : \beta_j \neq \beta_{j0}$, where β_{j0} takes some value of interest. We would use the test statistic

$$T_s = \frac{\hat{\beta}_j - \beta_{j0}}{s_{\hat{\beta}_j}}. \tag{21.50}$$

Under H_0, T_s has a t distribution with $n-k-1$ degrees of freedom, and we would reject H_0 for sufficiently large values of this statistic. The most commonly used null hypothesis tested is $H_0 : \beta_j = 0$ – if this test is significant it suggests the slope for X_j differs from zero, and so X_j is causing a change in Y. Note that this test examines the unique effect of X_j on Y with all the other independent variables in the model, in effect pitting X_j against all the other independent variables.

Confidence intervals can also be derived using the t distribution with $n-k-1$ degrees of freedom. The interval

$$(\hat{\beta}_j - c_{\alpha,n-k-1} s_{\hat{\beta}_j}, \hat{\beta}_j + c_{\alpha,n-k-1} s_{\hat{\beta}_j}) \tag{21.51}$$

is a $100(1-\alpha)\%$ confidence interval for β_j, where $c_{\alpha,n-k-1}$ could be obtained from Table T (see Chapter 9 for details). We will let SAS handle the details for these confidence intervals as well as tests.

Chapter 24 of this text lists a SAS program that carries out a multiple regression analysis for the Example 1 data using `proc iml` and matrix operations. This includes constructing the design matrix $\boldsymbol{X}$ and vector $\boldsymbol{Y}$ from the observations, estimating $\boldsymbol{\beta}$, then calculating MS_{error}, $MS_{regression}$, and $s_{\hat{\beta}_j}$. It also conducts the overall F test of the model and t tests for the regression coefficients.

21.5 Standardized regression coefficients

The regression coefficient β_j is the change in Y per unit of X_j (the slope) given the other independent variables in the model. The magnitude of β_j is affected by the strength of this relationship as well as the units of measurement for the variables. This can make it difficult to compare the relative effects of the different independent variables on Y, because their units could be quite different. Standardized regression coefficients solve this problem by expressing the slope in units of the standard deviation of Y and X_j (Kutner et al. 2005). They are calculated using the formula

$$\hat{\beta}'_j = \hat{\beta}_j \frac{s_{X_j}}{s_Y}, \tag{21.52}$$

where s_{X_j} is the sample standard deviation of X_j and s_Y is the sample standard deviation of Y. As a result of this scaling the standardized coefficients are dimensionless, similar to a correlation coefficient (Chapter 18).

21.6 R^2 values

We can define an R^2 value for multiple regression similar to one for linear regression. It is the proportion of the total sum of squares explained by the regression model, or

$$R^2 = \frac{SS_{regression}}{SS_{total}} = \frac{SS_{regression}}{SS_{regression} + SS_{error}}. \tag{21.53}$$

Large R^2 values suggest the regression model explains most of the variation (sum of squares) in the data, and vice versa for small R^2 values.

21.7 Multiple regression for Example 1 - SAS demo

We next conduct a multiple regression analysis of the Example 1 data using `proc reg` (SAS Institute Inc. 2018b). See program below. It is similar in structure to previous linear regression and ANOVA programs, but we will use `proc reg` rather than `proc glm` because it has several useful features for multiple regression. We first input the observations using a `data` step, applying transformations if necessary. Theoretical models of competition suggest a linear relationship between the log of the survival rate and measures of density like attack density and bluestain levels, so we define `y = log(survival)` in the `data` step. The two independent variables are attack density (defined as `satkden`) and bluestain levels (`blueden`).

As a first step in the analysis, it is often useful to plot the values of the dependent variables vs. the independent ones, to see their individual effects. We will use `proc gplot` (SAS Institute Inc. 2016) for this purpose, using commands similar to the ones for linear regression (Chapter 17). The program fits a regression line through the points in each graph, but these are the lines for linear, not multiple, regression. Special techniques are needed visualize the fitted model for multiple regression, which we will later examine.

The next section of the program conducts the multiple regression using `proc reg`. The `plots=diagnostics` option generates graphs that are used to examine the assumptions of multiple regression, similar to ANOVA and linear regression. The `model` statement tells SAS the multiple regression model, including the dependent variable (`y`) and the two independent variables (`satkden` and `blueden`). Note the similarity of the `model` statement to the multiple regression model with two independent variables. The option `clb` requests confidence intervals for the model parameters while `stb` displays the standardized regression coefficients. We will examine the remaining options later.

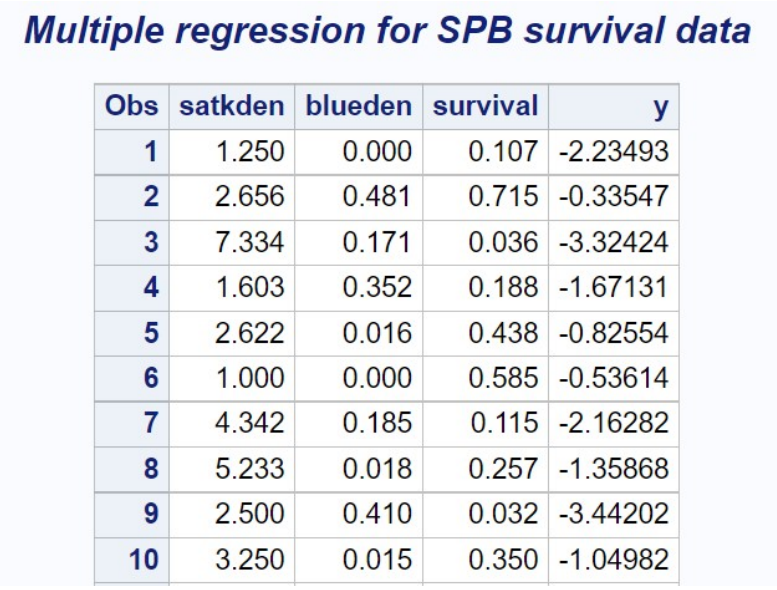
Examining the two `proc gplot` graphs, we see that log survival rate appeared to decrease with attack density, while bluestain had no obvious effect (Fig. 21.2, 21.3). The `proc reg` output contains the overall F for the multiple regression as well as separate t tests for the independent variables (Fig. 21.4). We see that the overall test was significant ($F_{2,24} = 5.45, P = 0.0112$), suggesting one or more of the independent variables affected survival. The t test for attack density was highly significant ($t_{24} = -3.30, P = 0.0030$) while bluestain was nonsignificant ($t_{24} = -0.65, P = 0.5243$). The slope or regression coefficient for attack density was negative ($\beta = -0.2391$), indicating survival decreases with attack density, as was the coefficient for bluestain ($\beta = -0.8096$). This suggests that bluestain actually had a greater effect than attack density, but this is because their units are quite different. If we examine the standardized regression coefficients, we see that attack density had a larger coefficient ($\beta' = -0.5682$) than bluestain ($\beta' = -0.1113$), and so had a larger effect on survival. The multiple regression model explained about 31% of the variation in the data ($R^2 = 0.3122$). The usual homogeneity of variances and normality assumptions also appear satisfied (Fig. 21.5).

SAS Program

```
* SPBsurvival.sas;
title "Multiple regression for SPB survival data";
data SPB;
    input satkden blueden survival;
    * Apply transformations here;
    y = log(survival);
    datalines;
1.250 0.000 0.107
2.656 0.481 0.715
7.334 0.171 0.036
1.603 0.352 0.188
2.622 0.016 0.438

etc.

5.000 0.338 0.207
;
run;
* Print data set;
proc print data=SPB;
run;
* Plot y vs. x variables;
proc gplot data=SPB;
    plot y*(satkden blueden) / vaxis=axis1 haxis=axis1;
    symbol1 i=rl v=star c=black height=2 width=2;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* Multiple regression analysis;
proc reg plots=diagnostics data=SPB;
    * Specify regression model and request residual-residual plots;
    model y = satkden blueden / clb stb tol vif partial;
run;
quit;
```



Multiple regression for SPB survival data

Obs	satkden	blueden	survival	y
1	1.250	0.000	0.107	-2.23493
2	2.656	0.481	0.715	-0.33547
3	7.334	0.171	0.036	-3.32424
4	1.603	0.352	0.188	-1.67131
5	2.622	0.016	0.438	-0.82554
6	1.000	0.000	0.585	-0.53614
7	4.342	0.185	0.115	-2.16282
8	5.233	0.018	0.257	-1.35868
9	2.500	0.410	0.032	-3.44202
10	3.250	0.015	0.350	-1.04982

etc.

Figure 21.1: `SPBsurvival.sas - proc print`

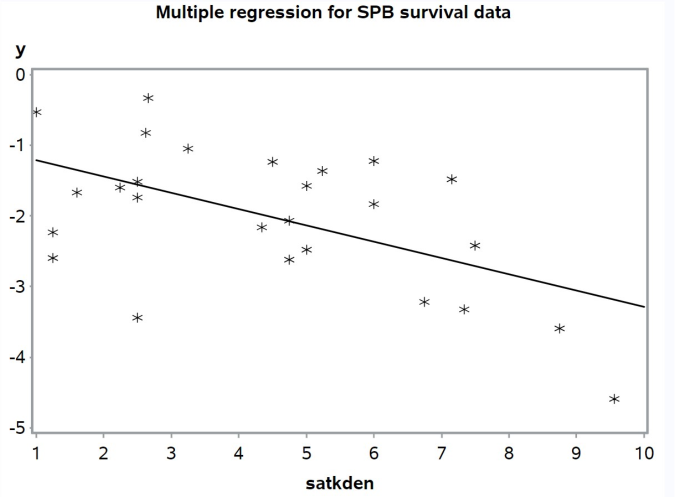


Figure 21.2: `SPBsurvival.sas - proc gplot`

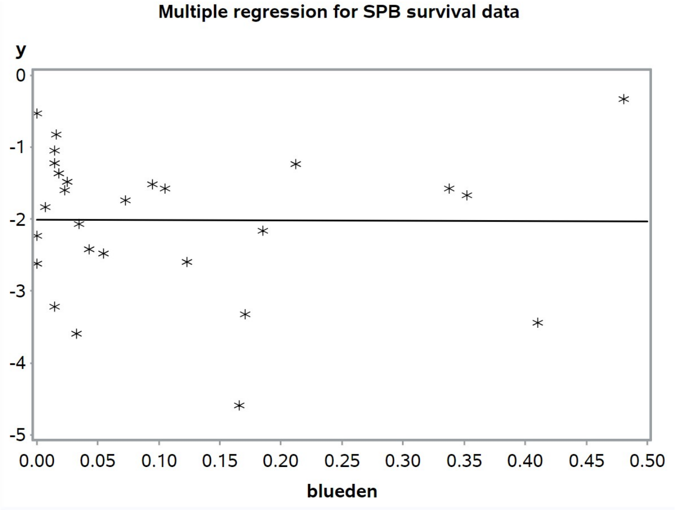
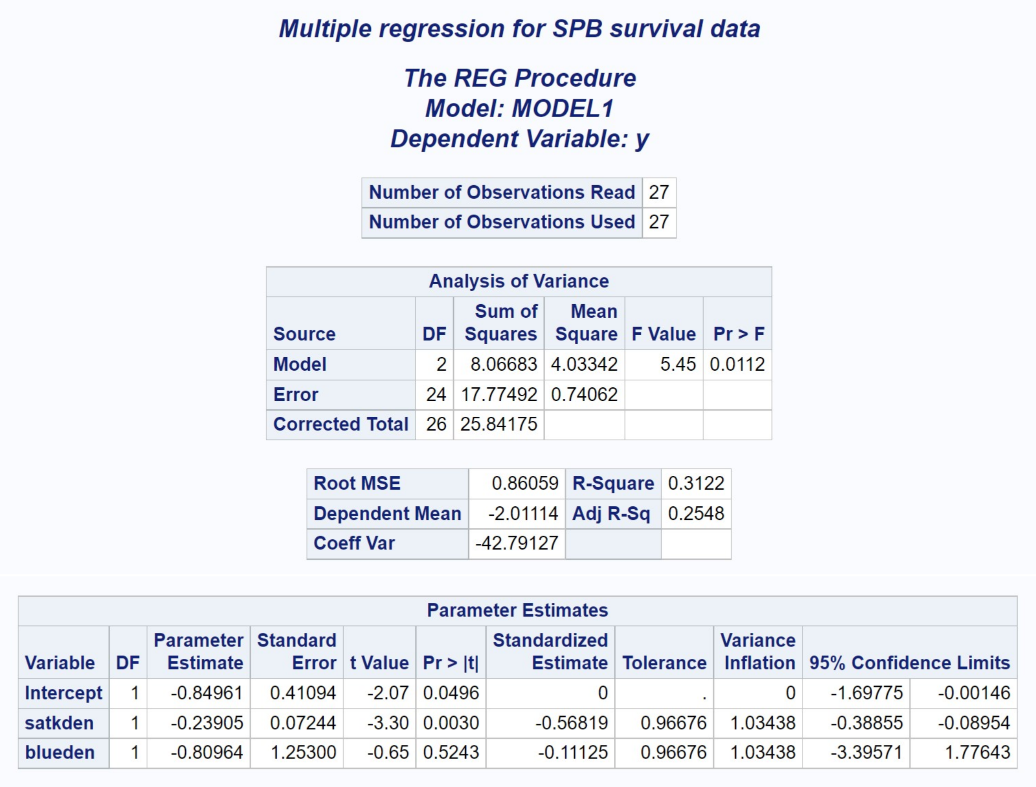


Figure 21.3: `SPBsurvival.sas - proc gplot`



Multiple regression for SPB survival data

The REG Procedure
Model: MODEL1
Dependent Variable: y

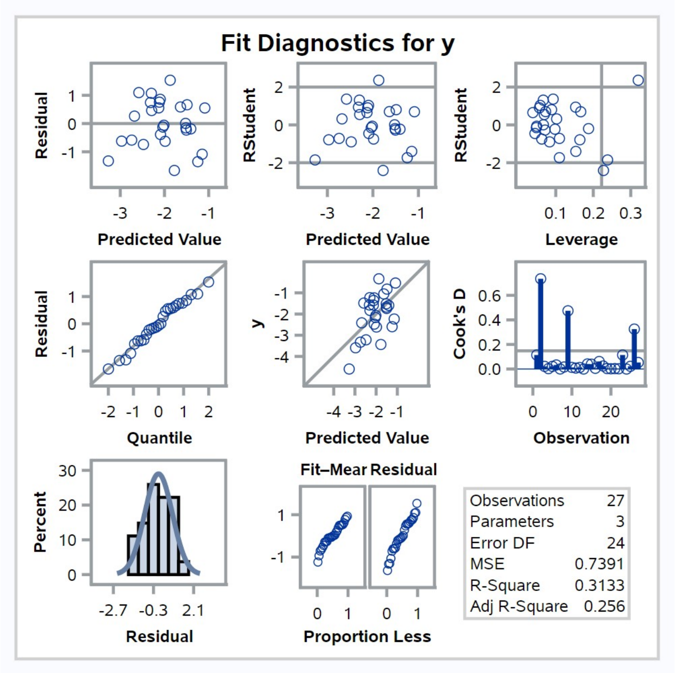
Number of Observations Read	27
Number of Observations Used	27

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	8.06683	4.03342	5.45	0.0112
Error	24	17.77492	0.74062		
Corrected Total	26	25.84175			

Root MSE	0.86059	R-Square	0.3122
Dependent Mean	-2.01114	Adj R-Sq	0.2548
Coeff Var	-42.79127		

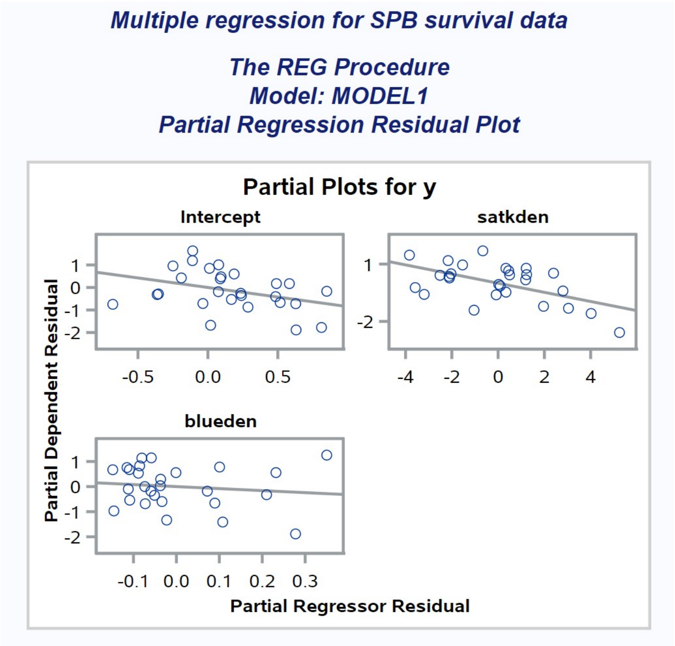
Parameter Estimates										
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	1	-0.84961	0.41094	-2.07	0.0496	0	.	0	-1.69775	-0.00146
satkden	1	-0.23905	0.07244	-3.30	0.0030	-0.56819	0.96676	1.03438	-0.38855	-0.08954
blueden	1	-0.80964	1.25300	-0.65	0.5243	-0.11125	0.96676	1.03438	-3.39571	1.77643

Figure 21.4: `SPBsurvival.sas - proc reg`



```
Observations      27
Parameters         3
Error DF          24
MSE           0.7391
R-Square      0.3133
Adj R-Square   0.256
```

Figure 21.5: `SPBsurvival.sas - proc reg`



Multiple regression for SPB survival data

The REG Procedure
Model: MODEL1
Partial Regression Residual Plot

Figure 21.6: `SPBsurvival.sas - proc reg`

21.8 Visualizing the multiple regression model

We can visualize the model fitted to the Example 1 data using a three-dimensional scatter plot (Fig. 21.7). The maximum likelihood (and least squares) process minimizes the squared residuals between the observations and the plane defined by the multiple regression model. From this graph, we can see that survival decreased with increasing attack density while bluestain had a minimal effect. The slope of the plane with respect to attack density is the same as the estimated slope in Fig. 21.4, and similarly for bluestain. This kind of graph would not work for more than two independent variables, because it would have more than three dimensions.

Another type of graph that works for any number of independent variables are **residual-residual plots**, or added-variable plots (Kutner et al. 2005). As the name suggests, they are constructed using two sets of residuals. Suppose we are interested in visualizing the effect of X_1 on Y. The first set of residuals is obtained from a multiple regression of X_1 on $X_2, X_3, \ldots, X_k$, with X_1 the dependent variable. The second set of residuals is from a multiple regression of Y on $X_2, X_3, \ldots, X_k$, excluding X_1. This procedure essentially subtracts the effect of $X_2, X_3, \ldots, X_k$ on both Y and X_1. If we plot the two sets of residuals against each other, this would show the unique effect of X_1 on Y. If we were to fit a line through these residuals using linear regression, the slope of the line would be equal to $\hat{\beta}_1$ from the full multiple regression (Y vs. $X_1, X_2, \ldots, X_k$).

Residual-residual plots are requested in SAS using the `partial` option in the `model` statement for `proc reg`, generating the output in Fig. 21.6. Besides visualizing the relationships between the dependent and independent variables, these plots can be used to identify outliers and observations that strongly influence the regression lines, known as high **leverage** points (Sheather 2009). They can also be used to determine whether the relationship between Y and a given X variable is in fact linear, as assumed by the multiple regression model. Examining these plots for the Example 1 data, the relationship between survival rates and attack density (or bluestain) appeared linear and there were no large outliers.

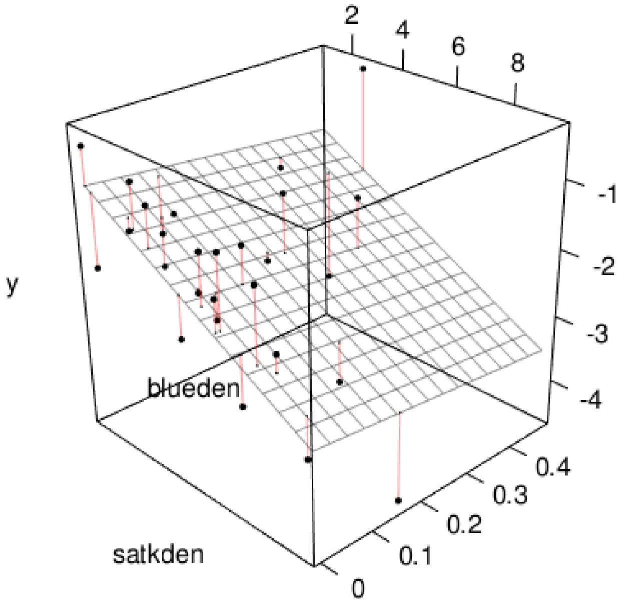


Figure 21.7: Multiple regression model fitted to the Example 1 data (see SAS program for variable definitions). The vertical red lines are the residuals for each observation ($Y_i - \hat{Y}_i$). This plot used R code from Chang (2023).

21.9 Collinearity in multiple regression

In a multiple regression analysis, there may sometimes be strong linear relationships or correlations among two or more independent variables, a problem called **collinearity**. This can cause issues in estimating the regression coefficients, including large standard errors and confidence intervals, and potentially large values for the estimates themselves. Another symptom of collinearity are independent variables that are nonsignificant even though the overall F test is significant. See Kutner et al. (2005) and Sheather (2009) for further details.

One diagnostic tool for detecting collinearity are **tolerance values**. They are calculated as follows. Suppose we want the tolerance value for the independent variable X_1. We would run a multiple regression of X_1 on $X_2, \ldots, X_k$ and find the $R^2(X_1)$ value for this regression. The tolerance value for X_1 is defined as $1 - R^2(X_1)$. If X_1 is strongly collinear with one or more independent variables, it will have a small tolerance value because $R^2(X_1)$ will be large. Another common measure is the **variance inflation factor**, defined as $1/(1 - R^2(X_1))$. This is just the inverse of the tolerance value, and will be large if there is strong collinearity among the independent variables. A common rule of thumb is that collinearity is a problem when a variance inflation factor is sufficient large, say 5 or 10 (Kutner et al. 2005; Sheather 2009)

The tolerance and variance inflation factors are requested using the options `tol` and `vif` the `model` statement for `proc reg`. Examining these quantities for the Example 1 data set, we see the variance inflation factors were small for both independent variables (Fig. 21.4). The variance inflation factors were the same here because there were only two independent variables in the model.

21.10 Multiple regression for Example 2 - SAS demo

We now analyze the Example 2 data set using SAS and `proc reg` (see program below). Here the objective is to predict endocranial volume for fossil skulls using a multiple regression model fitted to existing species. We first log-transform all the variables in a `data` step. This makes intuitive sense, because we would expect endocranial volume to be the product of length, height, and width. After log-transform this would yield an additive model that can be fitted using multiple regression. The dependent variable in the analysis is then `logV`, while `logL`, `logH`, and `logW` are the independent ones. Note the last two observations have missing values for endocranial volume – we will use multiple regression to predict it for these fossil skulls where endocranial volume is unavailable.

Plots generated using `proc gplot` show a strong linear relationship between `logV` and all three independent variables (Fig. 21.9-21.11). We then conduct the multiple regression using `proc reg` and the same syntax as in Example 1. Two new options for the `model` statement are `clm` and `cli`. The `clm` option generates a 95% confidence interval for the mean of Y_i for each observation, while `cli` generates a 95% prediction interval for a single Y_i (see Chapter 17). These intervals are calculated for all the observations, including the two fossil skulls. Examining the output (Fig. 21.12), we see that the overall F test was highly significant ($F_{3,190} = 8498.88, P < 0.0001$), as were the individual t tests for length ($t_{190} = 3.77, P = 0.0002$), height ($t_{190} = 9.55, P < 0.0001$), and width ($t_{190} = 14.11, P < 0.0001$). The standardized regression coefficients suggest that width had the greatest effect on endocranial volume ($\beta' = 0.5097$), followed by height ($\beta' = 0.3873$) and then width ($\beta' = 0.1052$). Combined, these three variables explained 99.3% of the variation in volume ($R^2 = 0.9926$), suggesting the model would be useful for prediction. The confidence and prediction intervals for the two fossil skulls are shown at the bottom of Fig. 21.13.

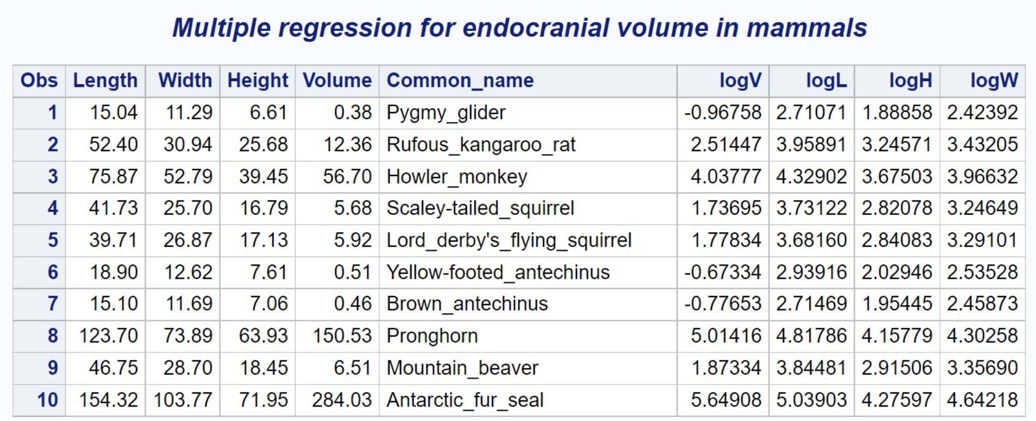
A possible concern with this analysis were large variance inflation factors for all three independent variables (Fig. 21.12). Despite these large values, the individual t tests for these variables were all highly significant, suggesting they each contribute something unique to the model. Kutner et al. (2005) also argue that collinearity is less important when prediction is primary goal of the analysis, as in the Example 2 regression.

SAS Program

```
* Endocranial4.sas;
title "Multiple regression for endocranial volume in mammals";
data ECVdat;
    input Length Width Height Volume Common_name :$30.;
    * Apply transformations here;
    logV = log(Volume);
    logL = log(Length);
    logH = log(Height);
    logW = log(Width);
    datalines;
15.04   11.29   6.61    0.38    Pygmy_glider
52.40   30.94   25.68   12.36   Rufous_kangaroo_rat
75.87   52.79   39.45   56.70   Howler_monkey
41.73   25.70   16.79   5.68    Scaley-tailed_squirrel
39.71   26.87   17.13   5.92    Lord_derby's_flying_squirrel

etc.

70.36   45.09   37.72   38.43   Arctic_fox
80.73   47.96   39.45   48.55   Fox
13.54   9.24    7.13    0.36    Meadow_jumping_mouse
13.15   9.05    7.00     .      Fossil_mouse
190.17  97.32   80.31    .      Fossil_bear
;
run;
* Print data set;
proc print data=ECVdat;
run;
* Plot y vs. x variables;
proc gplot data=ECVdat;
    plot logV*(logL logH logW) / vaxis=axis1 haxis=axis1;
    symbol1 i=rl v=star c=black height=2 width=2;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* Multiple regression;
proc reg plots=diagnostics data=ECVdat;
    * Specify variables in regression model;
    model logV = logL logH logW / clb stb tol vif partial clm cli;
run;
quit;
```



Multiple regression for endocranial volume in mammals

Obs	Length	Width	Height	Volume	Common_name	logV	logL	logH	logW
1	15.04	11.29	6.61	0.38	Pygmy_glider	-0.96758	2.71071	1.88858	2.42392
2	52.40	30.94	25.68	12.36	Rufous_kangaroo_rat	2.51447	3.95891	3.24571	3.43205
3	75.87	52.79	39.45	56.70	Howler_monkey	4.03777	4.32902	3.67503	3.96632
4	41.73	25.70	16.79	5.68	Scaley-tailed_squirrel	1.73695	3.73122	2.82078	3.24649
5	39.71	26.87	17.13	5.92	Lord_derby's_flying_squirrel	1.77834	3.68160	2.84083	3.29101
6	18.90	12.62	7.61	0.51	Yellow-footed_antechinus	-0.67334	2.93916	2.02946	2.53528
7	15.10	11.69	7.06	0.46	Brown_antechinus	-0.77653	2.71469	1.95445	2.45873
8	123.70	73.89	63.93	150.53	Pronghorn	5.01416	4.81786	4.15779	4.30258
9	46.75	28.70	18.45	6.51	Mountain_beaver	1.87334	3.84481	2.91506	3.35690
10	154.32	103.77	71.95	284.03	Antarctic_fur_seal	5.64908	5.03903	4.27597	4.64218

etc.

Figure 21.8: `Endocranial4.sas - proc print`

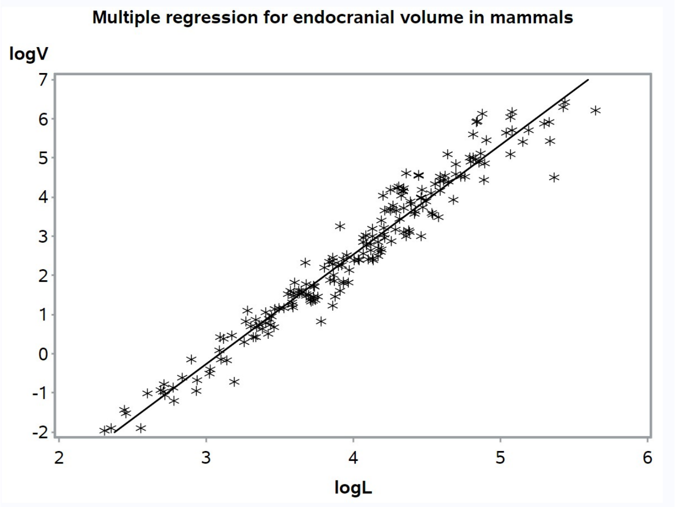


Figure 21.9: `Endocranial4.sas - proc gplot`

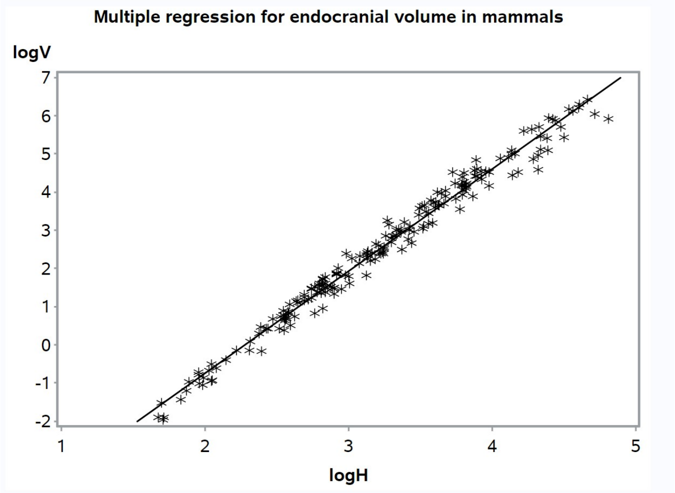


Figure 21.10: `Endocranial4.sas - proc gplot`

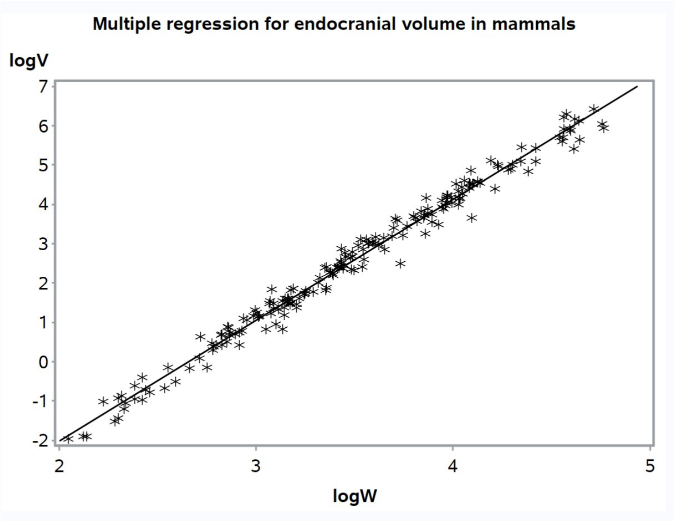
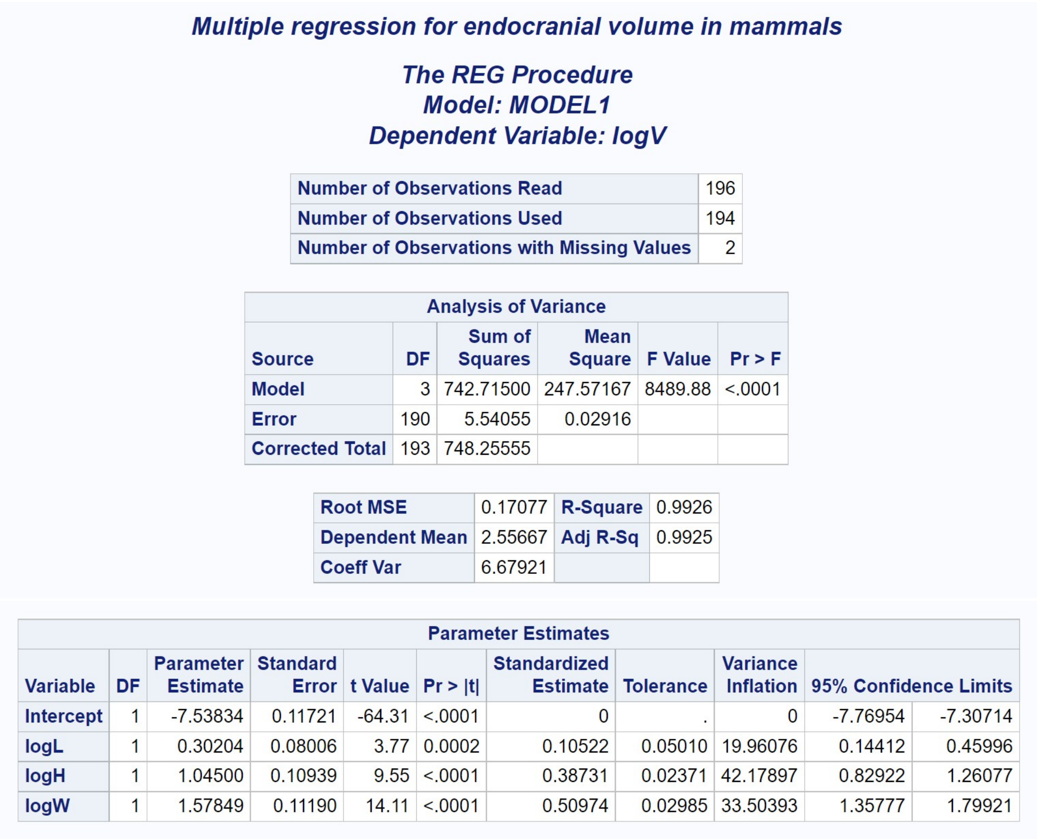


Figure 21.11: `Endocranial4.sas - proc gplot`



Multiple regression for endocranial volume in mammals

The REG Procedure
Model: MODEL1
Dependent Variable: logV

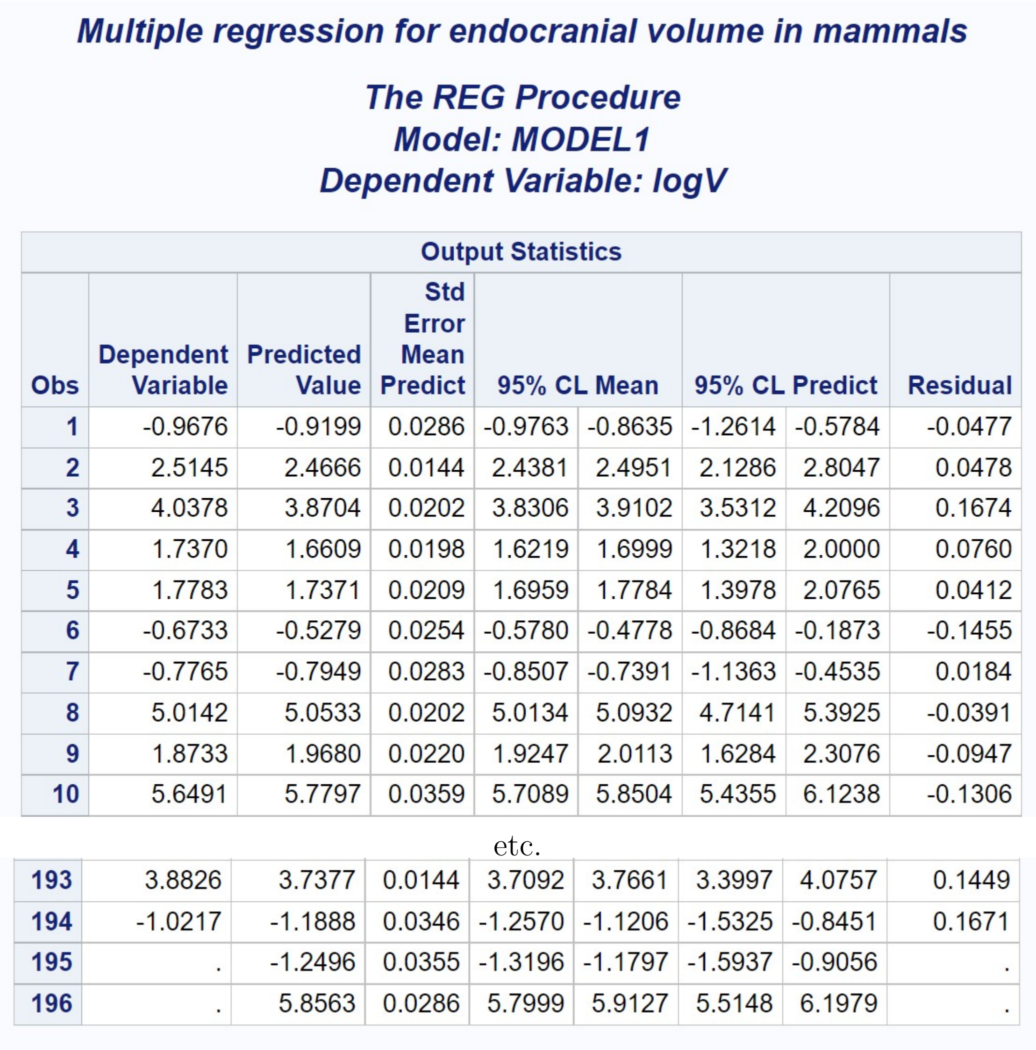
Number of Observations Read	196
Number of Observations Used	194
Number of Observations with Missing Values	2

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	742.71500	247.57167	8489.88	<.0001
Error	190	5.54055	0.02916		
Corrected Total	193	748.25555			

Root MSE	0.17077	R-Square	0.9926
Dependent Mean	2.55667	Adj R-Sq	0.9925
Coeff Var	6.67921		

Parameter Estimates										
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	1	-7.53834	0.11721	-64.31	<.0001	0	.	0	-7.76954	-7.30714
logL	1	0.30204	0.08006	3.77	0.0002	0.10522	0.05010	19.96076	0.14412	0.45996
logH	1	1.04500	0.10939	9.55	<.0001	0.38731	0.02371	42.17897	0.82922	1.26077
logW	1	1.57849	0.11190	14.11	<.0001	0.50974	0.02985	33.50393	1.35777	1.79921

Figure 21.12: `Endocranial4.sas - proc reg`



Multiple regression for endocranial volume in mammals

The REG Procedure
Model: MODEL1
Dependent Variable: logV

Output Statistics

Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
1	-0.9676	-0.9199	0.0286	-0.9763	-0.8635	-1.2614	-0.5784	-0.0477
2	2.5145	2.4666	0.0144	2.4381	2.4951	2.1286	2.8047	0.0478
3	4.0378	3.8704	0.0202	3.8306	3.9102	3.5312	4.2096	0.1674
4	1.7370	1.6609	0.0198	1.6219	1.6999	1.3218	2.0000	0.0760
5	1.7783	1.7371	0.0209	1.6959	1.7784	1.3978	2.0765	0.0412
6	-0.6733	-0.5279	0.0254	-0.5780	-0.4778	-0.8684	-0.1873	-0.1455
7	-0.7765	-0.7949	0.0283	-0.8507	-0.7391	-1.1363	-0.4535	0.0184
8	5.0142	5.0533	0.0202	5.0134	5.0932	4.7141	5.3925	-0.0391
9	1.8733	1.9680	0.0220	1.9247	2.0113	1.6284	2.3076	-0.0947
10	5.6491	5.7797	0.0359	5.7089	5.8504	5.4355	6.1238	-0.1306

etc.

193	3.8826	3.7377	0.0144	3.7092	3.7661	3.3997	4.0757	0.1449
194	-1.0217	-1.1888	0.0346	-1.2570	-1.1206	-1.5325	-0.8451	0.1671
195	.	-1.2496	0.0355	-1.3196	-1.1797	-1.5937	-0.9056	.
196	.	5.8563	0.0286	5.7999	5.9127	5.5148	6.1979	.

Figure 21.13: `Endocranial4.sas - proc reg`

21.11 Power analysis for multiple regression

The appropriate sample size for a multiple regression study can be determined through a power analysis. Similar to power analysis in ANOVA, we must specify the Type I error rate α, the desired power level, and the size of the effect we wish to detect. The effect size in power analyses for multiple regression is often expressed in terms of an R^2 value, which combines the effects of the independent variables (through $SS_{regression}$) and the variability of the observations (SS_{error}).

The SAS procedure `power` can do a power analysis for multiple regression using the `multreg` option. We first specify the Type I error rate and desired power using the `alpha` and `power` options (see SAS program below). We will be interested in the sample sizes needed for the overall F test of $H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$, which is equivalent to testing $H_0 : R^2 = 0$. This value of R^2 is specified using the `rquaredreduced` option. The values of R^2 under the alternative hypothesis ($H_1 : \beta_j \neq 0$ for some j) are then specified using the `rsquarefull` option. Some plausible values for ecological or behavioral data are 0.1, 0.3, and 0.6, but any value can be used. We must also specify the number of independent variables (k) under H_0 and H_1, using the `nreducedpredictors` and `nfullpredictors` options. We set the `ntotal` option to a missing value, which tells `power` to solve for the sample size n that gives the desired power.

SAS Program

```
* multreg_power.sas;
title 'Power Analysis for Multiple Regression';
proc power;
    multreg
    model = fixed
    alpha = 0.05
    power = 0.8
    rsquarereduced = 0
    rsquarefull = 0.1 0.3 0.6
    nreducedpredictors = 0
    nfullpredictors = 1 2 3 4 5 6 7 8 9 10 20 30 40 50
    ntotal = . ;
run;
quit;
```

Table 21.4 summarizes the result of this analysis, with the entries the sample size n to obtain the desired power. Note that the effect size (R^2 under H_1) strongly influences sample size, and that more observations are necessary to maintain power as the number of independent variables (k) is increased. For one predictor, the sample size specified is for a simple linear regression.

The `power` procedure can be used to find the sample size for other scenarios, including tests of the individual regression coefficients ($H_0 : \beta_j = 0$). The basic idea is to specify an R^2 value with and without X_j in the model, with the number of predictors in the full and reduced model differing by 1.

Table 21.4: Power for Multiple Regression - Effect of R^2 and the number of independent variables (k) on the sample size n for the overall F test of the model ($H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0$). See text for further details.

		R^2	
k	0.1	0.3	0.6
1	73	21	8
2	90	26	11
3	103	30	12
4	113	33	14
5	122	36	16
6	130	39	17
7	137	42	18
8	144	44	20
9	150	46	21
10	156	48	22
20	205	67	34
30	244	82	45
40	278	97	55
50	308	110	66

21.12 Polynomial regression

In a linear regression, we sometimes saw a nonlinear relationship between Y and X for some data sets (Chapter 17). This problem could often be fixed by applying a transformation to Y or X , but this approach sometimes fails. An alternative solution is to fit a flexible polynomial in X to the data. The observations would be modeled using the equation

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \ldots + \epsilon_i. \tag{21.54}$$

This is a **polynomial regression** model. It is similar in structure to multiple regression, except the independent variables $X_1, X_2, \ldots, X_k$ are replaced with increasing powers of X.

As we add more powers of X, the polynomial regression model becomes increasingly flexible. A model using only X and X^2 would fit a quadratic polynomial (a parabola) to the data, while one with X, X^2, and X^3 would fit a cubic one, which is S-shaped. While higher powers of X would allow even more flexibility, they are seldom needed to obtain an adequate fit. Another issue is extrapolation beyond the range of X values, where higher order polynomials can generate unrealistic estimates (Kutner et al. 2005). For these reasons, it is desirable to find the lowest order polynomial that adequately describes the data.

One issue with using the powers of X in a regression is that they are collinear with one another. For example, we would expect X, X^2, and X^3 to be strongly correlated. A common strategy is to use **centered polynomials** to reduce this collinearity. This is accomplished by centering the independent variable around its mean before finding the power. In particular, we first define $x = X - \bar{X}$ and then raise x to the desired power, using these centered variables in the polynomial regression.

21.13 Population growth experiment - SAS demo

As an example of polynomial regression, we will analyze data from a hypothetical experiment on a stored grain insect, where varying numbers of adult insects (N) are added to a container with grain, and then the number of offspring per adult estimated (R). We would expect that R would decrease as N was increased because of intraspecific competition among the insects. The Ricker model is often used as a simple description of intraspecific competition and could be suitable for these data (Ricker 1954). The model has two parameters, the intrinsic growth rate r of the organism and its carrying capacity K. For this model, we would expect the following relationship between $\log R$ and N:

$$\log R = r(1 - (N/K)) = r - (r/K)N = \alpha - \beta N, \tag{21.55}$$

where $\alpha = r$ and $\beta = r/K$. This is essentially a linear regression model for $\log R$ vs. N. What we would like to determine is whether this model is adequate, or whether a more complex nonlinear one is needed. We can answer this question using a polynomial regression model with different powers of N. If the tests for these terms are significant, it suggests a more complex model is needed for these observations.

The SAS program below lists the observations from this hypothetical experiment in a `data` step. Also listed is the mean of value of N (`nbar = 50.455`). This is used in the centering process, which first calculates a centered density `x` and then the powers of `x` (`x2`, `x3`). The data are then plotted along with a smooth line using `proc gplot` and the `symbol1 i=sm70` option. The smooth line is constructed using cubic splines, which are themselves a kind of polynomial. This graph helps visualize the relationship between `logR` and `n`.

We then use `proc glm` to conduct the polynomial regression (SAS Institute Inc. 2018b). We use this procedure rather than `proc reg` because it can generate Type I sums of squares and tests. These are produced by sequentially fitting the different terms in the `model` statement, and can be used to determine the lowest order polynomial needed to describe the data. For example, the Type I test for `x3` tests whether this power is needed with `x` and `x2` already in the model.

The results from `proc glm` output suggested a quadratic polynomial provides an adequate description of these data (see discussion below). The remainder of the program plots the observations with a quadratic polynomial line plus a confidence interval (`proc gplot` with the `symbol` option `i=rqclm`). It then uses `proc reg` to finish the analysis, using syntax similar to previous multiple regression analyses.

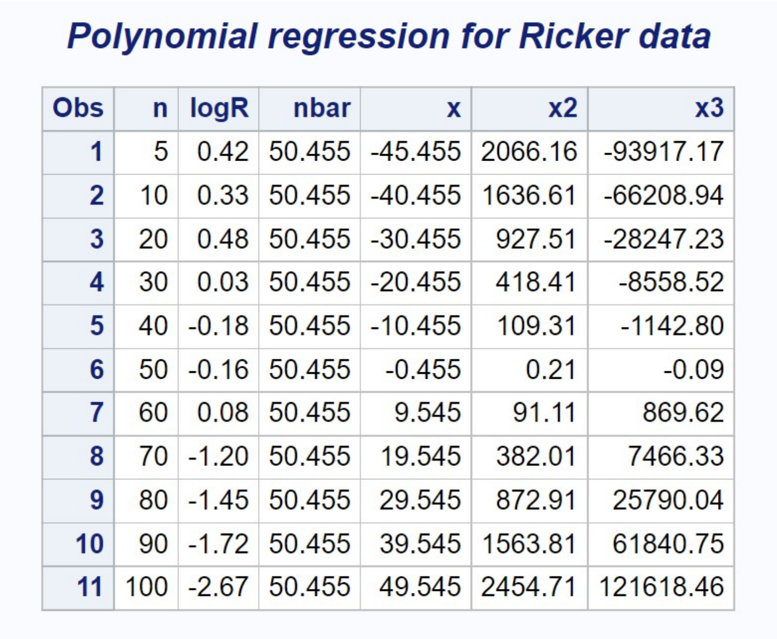
SAS Program

```
* Ricker_polynomial.sas;
title "Polynomial regression for Ricker data";
data ricker;
    input n logR;
    * For centered polynomials, you'll need the mean X value;
    nbar = 50.455;
    x = n-nbar;
    x2 = x**2;
    x3 = x**3;
    datalines;
 5  0.42
10  0.33
20  0.48
30  0.03
40 -0.18
50 -0.16
60  0.08
70 -1.20
80 -1.45
90 -1.72
100 -2.67
;
run;
* Print data set;
proc print data=ricker;
run;
* Plot data and fit smooth line;
proc gplot data=ricker;
    plot logR*n / vaxis=axis1 haxis=axis1;
    symbol1 i=sm70 v=star c=black height=2 width=2;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
run;
* Polynomial regression;
proc glm data=ricker;
    * Look at Type I tests to determine order of polynomial;
    model logR = x x2 x3;
run;
* Preceding analysis suggests second-order polynomial adequate;
* Plot the data and second-order polynomial;
proc gplot data=ricker;
    plot logR*n / vaxis=axis1 haxis=axis1;
    symbol1 i=rqclm v=star c=black height=2 width=2;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
```

```
run;
* Polynomial regression with second-order polynomial;
proc reg data=ricker;
    model logR = x x2 / clb stb tol vif partial;
run;
quit;
```

Examining the first `proc gplot` graph, we observe `logR` decreased with density `n`, suggesting that reproduction was affected by intraspecific competition. The relationship appears curved, however, and so the Ricker model may not be adequate (Fig. 21.15). The Type I tests from `proc glm` yielded highly significant results for `x` ($F_{1,7} = 109.81, P < 0.0001$) and `x2` ($F_{1,7} = 12.45, P = 0.0096$), but a nonsignificant one for `x3` ($F_{1,7} = 0.43, P = 0.5328$) (Fig. 21.16). This pattern suggests a quadratic polynomial would be sufficient to describe these data. In addition, the highly significant test for `x2` means we can definitively reject the linear Ricker model.

The second `proc gplot` graph shows that a quadratic provides a reasonable approximation to the observations (Fig. 21.17). Examining the `proc reg` output (Fig. 21.18), we see that overall F test was highly significant ($F_{3,8} = 40.89, P < 0.0001$), as were the individual tests for `x` ($t_8 = -10.59, P < 0.0001$) and `x2` ($t_8 = -3.66, P = 0.0064$). The polynomial regression model explained about 94% of the variation in the data ($R^2 = 0.943$). Due to centering, the variance inflation factors show no collinearity issues with `x` and `x2`.



Polynomial regression for Ricker data

Obs	n	logR	nbar	x	x2	x3
1	5	0.42	50.455	-45.455	2066.16	-93917.17
2	10	0.33	50.455	-40.455	1636.61	-66208.94
3	20	0.48	50.455	-30.455	927.51	-28247.23
4	30	0.03	50.455	-20.455	418.41	-8558.52
5	40	-0.18	50.455	-10.455	109.31	-1142.80
6	50	-0.16	50.455	-0.455	0.21	-0.09
7	60	0.08	50.455	9.545	91.11	869.62
8	70	-1.20	50.455	19.545	382.01	7466.33
9	80	-1.45	50.455	29.545	872.91	25790.04
10	90	-1.72	50.455	39.545	1563.81	61840.75
11	100	-2.67	50.455	49.545	2454.71	121618.46

Figure 21.14: `Ricker_polynomial.sas - proc print`

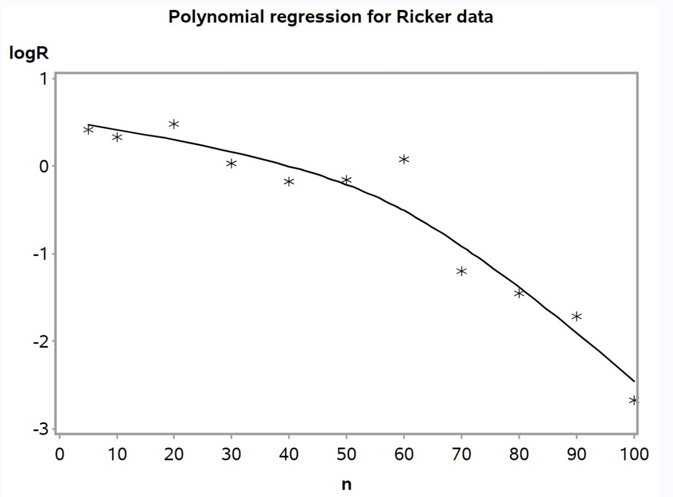
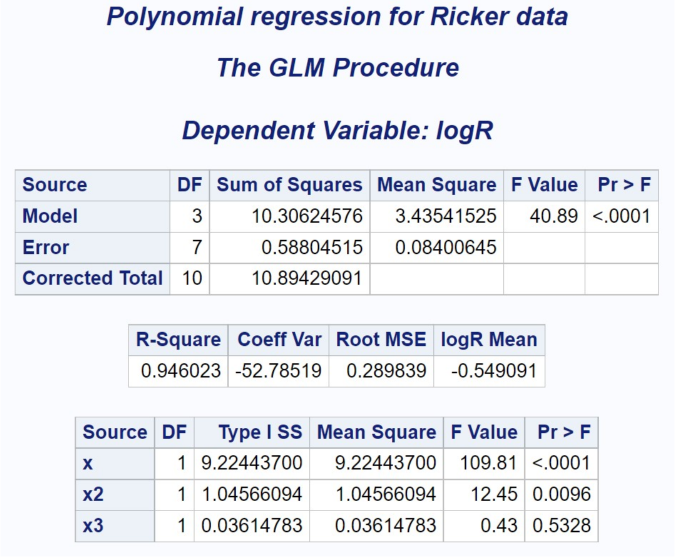


Figure 21.15: `Ricker_polynomial.sas - proc gplot (1)`



Polynomial regression for Ricker data

The GLM Procedure

Dependent Variable: logR

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	10.30624576	3.43541525	40.89	<.0001
Error	7	0.58804515	0.08400645		
Corrected Total	10	10.89429091			

R-Square	Coeff Var	Root MSE	logR Mean
0.946023	-52.78519	0.289839	-0.549091

Source	DF	Type I SS	Mean Square	F Value	Pr > F
x	1	9.22443700	9.22443700	109.81	<.0001
x2	1	1.04566094	1.04566094	12.45	0.0096
x3	1	0.03614783	0.03614783	0.43	0.5328

Figure 21.16: `Ricker_polynomial.sas - proc glm`

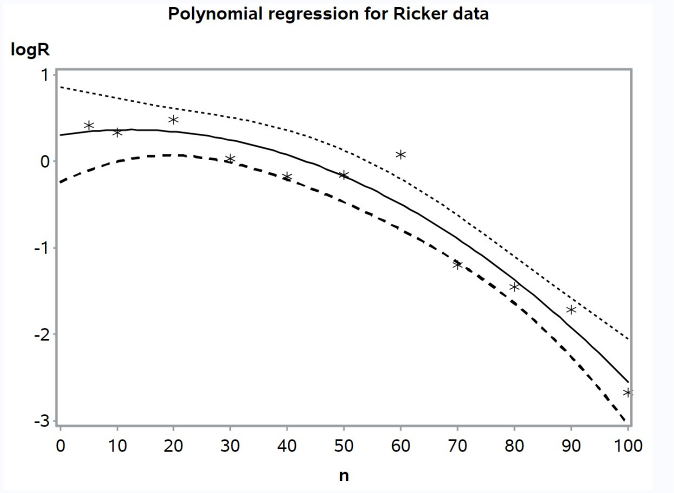
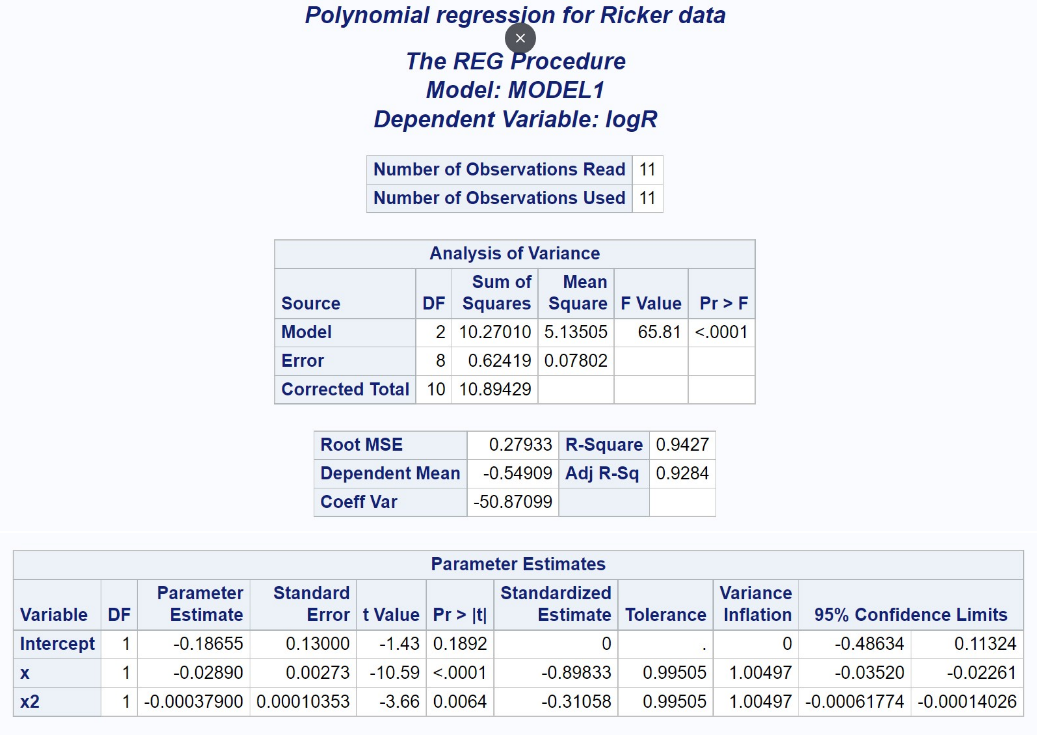


Figure 21.17: `Ricker_polynomial.sas - proc gplot (2)`



Polynomial regression for Ricker data

The REG Procedure
Model: MODEL1
Dependent Variable: logR

Number of Observations Read	11
Number of Observations Used	11

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	10.27010	5.13505	65.81	<.0001
Error	8	0.62419	0.07802		
Corrected Total	10	10.89429			

Root MSE	0.27933	R-Square	0.9427
Dependent Mean	-0.54909	Adj R-Sq	0.9284
Coeff Var	-50.87099		

Parameter Estimates										
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate	Tolerance	Variance Inflation	95% Confidence Limits	
Intercept	1	-0.18655	0.13000	-1.43	0.1892	0	.	0	-0.48634	0.11324
x	1	-0.02890	0.00273	-10.59	<.0001	-0.89833	0.99505	1.00497	-0.03520	-0.02261
x2	1	-0.00037900	0.00010353	-3.66	0.0064	-0.31058	0.99505	1.00497	-0.00061774	-0.00014026

Figure 21.18: `Ricker_polynomial.sas - proc reg`

21.14 Model selection using information criteria

There is a substantial literature on problems with hypothesis testing and P values in scientific research, as well as defenses of this approach. Recent papers that summarize these issues include Aho et al. (2014), Burnham and Anderson (2014), Murtaugh (2014), and de Valpine (2014), in an ecological context. The most common alternative to hypothesis testing is model selection using Akaike's Information Criterion, or AIC (Akaike 1974; Anderson et al. 2000; Burnham and Anderson 2002; Burnham and Anderson 2014). The basic idea is to formulate a collection of models to describe the data, and then choose the best one based on AIC values, defined as the model with the smallest AIC. For example, in a multiple regression setting we might be interested in determining the best model among different subsets of the independent variables. There is no explicit hypothesis testing in this approach, but confidence intervals can be calculated to describe the magnitude of an effect.

While the hypothesis testing and AIC approaches seem different, they often use similar statistical models with the same sets of assumptions. There is also a common scenario, nested models, where the two approaches would produce similar results. Models are nested when a simpler model is a special case of a more complex one, with fewer parameters or variables. Procedures like ANOVA and multiple regression utilize nested models, with the tests constructed using a simpler H_0 model nested within a more complex H_1 model. Murtaugh (2014) showed there is a direct relationship between P values and changes in AIC under these conditions. Suppose that a test comparing H_0 and H_1 was highly significant, favoring the H_1 model. The AIC value for the H_1 model would also be substantially smaller than H_0, and so this approach would also select the H_1 model. However, it is important to note there are scenarios where the models are not nested, which precludes hypothesis testing and P values but where AIC is useful. For example, Burnham & Anderson (2002) used AIC to compare nine different nonlinear models of the relationship between the number of bird species and sample size, with the models of such different forms they could not be nested.

So what is AIC? The AIC uses the concept of **Kullback-Leibler information**. We suppose that the data have some probability distribution f, and we would like to approximate it with another distribution g. These two distributions can be thought of as different models for the data, with f the true one. Kullback-Leibler information is a measure of the distance between f and g, denoted as $I(f, g)$. In mathematical terms, it is defined as the expected value of $\ln(f/g)$, where the expected value is calculated using the f distribution:

$$I(f,g) = E_f \left[\ln \left(\frac{f(x)}{g(x|\boldsymbol{\theta})} \right) \right]. \tag{21.56}$$

(Akaike 1974; Anderson et al. 2000; Burnham and Anderson 2002). The notation $g(x|\boldsymbol{\theta})$ is used to emphasize that g has a number of parameters (say θ_1, θ_2, etc.) that could affect $I(f, g)$. $I(f, g)$ is always positive unless $f = g$, for which $I(f, g) = 0$. Because the true distribution f and the parameters of g are typically unknown, $I(f, g)$ is not useful in this form because it cannot be calculated.

To see how $I(f, g)$ behaves, suppose that f and g are simple continuous distributions like the normal. Equation 21.56 can then be expressed as an integral of the form

$$I(f,g) = \int f(x) \ln \left(\frac{f(x)}{g(x|\boldsymbol{\theta})} \right) dx. \tag{21.57}$$

If f and g are quite distinct from each other $I(f, g)$ will be large, because positive values of $\ln(f/g)$ will mostly coincide with f, and so receive more weight in the integral (Fig. 21.19). This effect is diminished when f and g are closely overlapping. One can think of $I(f, g)$ as measuring the mismatch between the two distributions, or more formally as the loss of information when approximating the true distribution f using $g(x|\boldsymbol{\theta})$.

We can break the expected value in Equation 21.56 into two pieces, using the fact that $\ln(a/b) = \ln a - \ln b$ and formulas for the expected value of a sum (see Chapter 7). We have

$$I(f,g) = E_f[\ln f(x)] - E_f[\ln g(x|\boldsymbol{\theta})]. \tag{21.58}$$

The first term in this equation does not involve g, and in any event would be a constant because f is fixed. This suggests that to minimize $I(f, g)$, we should compare the relative values of the second term. It can be shown that smaller values of $-E_f[\ln g(x|\boldsymbol{\theta})]$ would make $I(f, g)$ smaller, minimizing the loss of information (Burnham and Anderson 2002).

The contribution of Akaike (1974) was to find an estimator of $-E_f[\ln g(x|\boldsymbol{\theta})]$ using maximum likelihood, which also provides estimates of the parameters of g. Suppose we have a data set that could be used to estimate $\boldsymbol{\theta}$ using maximum likelihood (see Chapter 8). He showed that $-E_f[\ln g(x|\boldsymbol{\theta})]$ could be estimated using

$$AIC = -2 \ln L(\hat{\boldsymbol{\theta}}) + 2K, \tag{21.59}$$

where $L(\hat{\boldsymbol{\theta}})$ is the likelihood function for g at the maximum likelihood estimate of $\boldsymbol{\theta}$, by definition the largest value of L (Akaike 1974; Anderson et al. 2000; Burnham and Anderson 2002). Here K is the number of parameters in $\boldsymbol{\theta}$. An interesting feature of the AIC is that K is actually a bias correction for this estimate.

Now suppose we have a number of different g distributions that are models for our data, with different numbers of parameters. Models with the smallest value of AIC would also have the smallest $I(f, g)$, and so the smallest loss of information in approximating f by g. We can gain further insight into this process by examining the two terms in the AIC formula. Models with more parameters could potentially fit the data better, generating a larger L and so smaller $-2 \ln L$, but they would also have larger values of $2K$. Thus, the AIC imposes a tradeoff between the fit of the model and its complexity.

In multiple regression, ANOVA, and other general linear models, $-2 \ln L$ and so AIC are a function of SS_{error} and the number of parameters in the model. In particular, for models of this type we have

$$AIC = n \ln(SS_{error}/n) + 2K \tag{21.60}$$

where n is sample size. We can see from this expression that better models will tend to have smaller values of SS_{error} and also fewer parameters, for a given sample size. Note that different software packages may count K and calculate AIC in different ways, so that the values of reported are different. These differences, while confusing, have no effect on the relative ranking of models by AIC.

A quantity related to AIC is the Bayesian Information Criterion or BIC (Schwarz 1978). The BIC was derived using the Bayesian interpretation of probability as a belief, but is valid outside this framework. BIC is calculated using the formula

$$BIC = -2 \ln L(\hat{\theta}) + \ln(n)K \tag{21.61}$$

where as before n is sample size and K the number of parameters. The only difference between the formulas for AIC and BIC is the multiplier for K – it is a constant (2) for AIC but $\ln(n)$ for BIC. In terms of regression and

ANOVA models, BIC can be calculated using the formula

$$BIC = n\ln(SS_{error}/n) + \ln(n)K \tag{21.62}$$

The BIC is used in the same fashion as AIC, with smaller values indicating a better model. It is clear from this formula that BIC penalizes complex models more heavily as sample size increases, because of the $\ln(n)$ multiplier.

Which criterion, AIC vs. BIC, performs best in model selection? Brewer et al. (2016) compared the two methods using simulated data intended to mimic the hidden heterogeneity likely present in real data sets, where the data could be mixture of observations with different parameter values. Performance was measured by how well the selected models predicted the observations of similar data sets, separate from the ones used in model selection. This tests how well the predictions of the model generalize to new observations. When heterogeneity was low AIC generally performed best, but BIC was better when heterogenity was large, so there was no clear winner.

We will use a more complex data set to illustrate model selection using AIC. Kaul and Wilsey (2020) wanted to determine which factors affect the success of tallgrass prairie restorations located in Iowa, USA. These prairies were restored using seed mixes, and as one measure of success they compared the species diversity of the seed mix with the diversity at the restored site, using the Bray-Curtis dissimilarity index as the dependent variable. This index ranges from 0 (all species shared) to 1 (none in common), so larger values suggest the restoration has failed. The independent variables were the age of the site and its linearity (shape), soil pH and organic matter, temperature and precipitation at establishment as well as annual averages, and exotic species abundance. A subset of these observations is shown in Table 21.5 (see https://datadryad.org for the full data set).

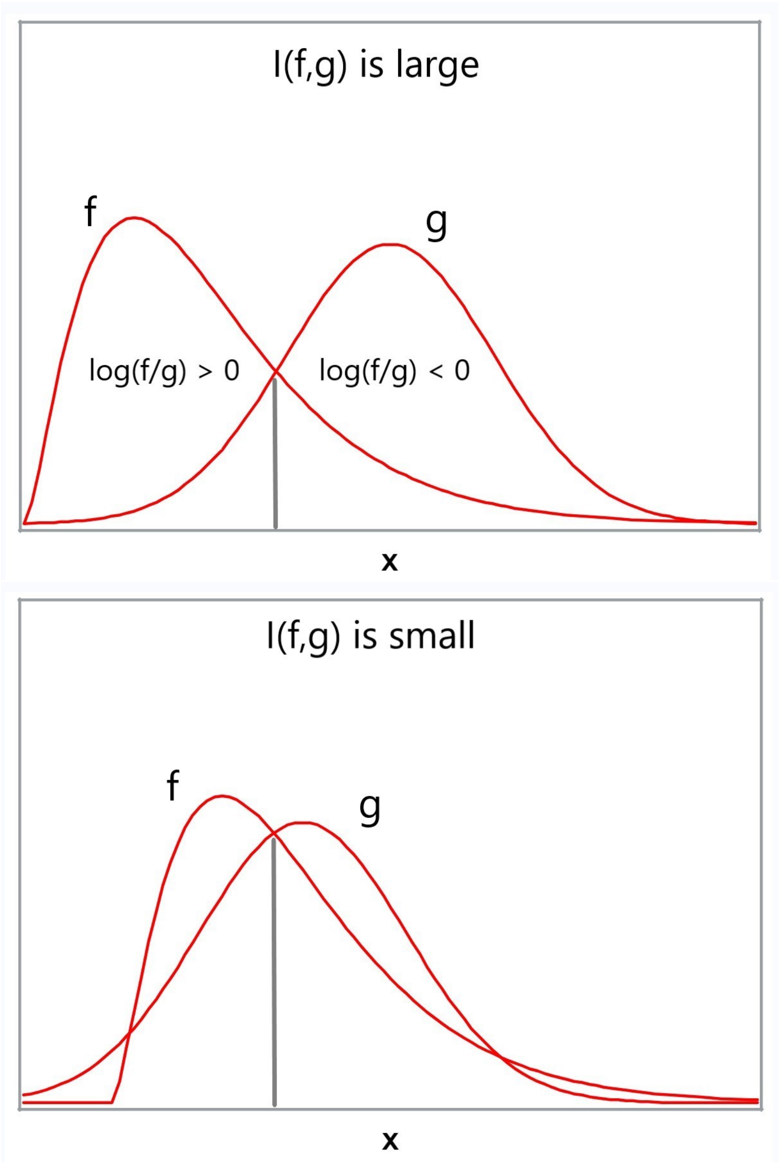


Figure 21.19: Graphical illustration of $I(f, g)$ under two scenarios.

Table 21.5: Example 3 - Site variables for restored prairies and Bray-Curtis dissimilarity (Kaul and Wilsey 2020). Here TE = temperature at establishment, PE = precipitation at establishment, TA = average annual temperature, and PA = average annual precipitation. See text for further details.

Site	Age	Linearity	pH	Organic	TE	PE	TA	PA	Exotic	Bray-Curtis	i
3	5	1.29	7.33	12.28	12.25	31.77	9.94	35.61	24.05	0.982	1
4	6	1.39	7.93	8.54	8.69	45.50	8.89	36.80	19.87	0.898	2
5	14	1.21	7.90	7.11	9.31	31.83	9.11	35.52	3.68	0.791	3
7	5	1.24	8.03	6.28	9.42	28.26	8.00	36.48	10.25	1.000	4
8	3	1.28	7.67	5.69	6.61	43.13	8.00	36.48	18.00	0.998	5
etc.											
100	17	1.30	7.87	10.42	9.17	33.58	10.72	37.59	25.03	0.970	40
101	3	1.26	7.93	8.64	8.56	40.66	10.72	37.59	15.37	0.772	41
102	11	1.25	8.10	3.61	11.69	40.36	10.00	36.28	9.57	0.972	42
105	13	1.05	7.93	14.54	8.53	35.57	8.00	36.48	13.77	0.718	43
106	10	1.02	6.97	8.19	7.64	47.79	8.00	36.48	14.25	0.624	44

21.15 Model selection for Example 3 - SAS demo

We will use *AIC* to select the best model for the Example 3 data, with multiple regression the underlying model (see program below). We first input the observations using a `data` step, selecting the Bray-Curtis index (`bc`) as the dependent variable `y`. We then plot `y` vs. all the independent variables (`age`, `linear`, ..., `exotic`) using `proc gplot`.

The next section of the program conducts a standard multiple regression using `proc reg`. We will later compare the results of this analysis with that generated by `proc glmselect`, a SAS procedure that implements various types of model selection (SAS Institute Inc. 2018b). The `model` statement for `proc glmselect` is similar to `proc reg`, but with a `class` statement it can also accomodate ANOVA-like factors. Model selection using *AIC* is implemented using the `selection=stepwise(select=AICC)` option. Stepwise refers to the search method, with the procedure adding or dropping individual variables until it finds the best model. The option `AICC` requests a version of *AIC* corrected for small sample sizes. Model selection using *BIC* could be requested using the `select=SBC` option (Schwarz's Bayesian Criterion or *BIC*).

Examining a subset of `proc gplot` graphs, we see that the Bray-Curtis dissimilarity index (`y`) increased with the linearity of the site (`linear`) and exotic species abundance (`exotic`), and decreased with precipitation during establishment (`PE`) (Fig. 21.21-21.23). From the `proc reg` output (Fig. 21.24), we see that the overall model was highly significant ($F_{9,34} = 11.11, P < 0.0001$) as were the individual tests for linearity ($t_{34} = 3.43, P = 0.0016$), exotic abundance ($t_{34} = 4.19, P = 0.0002$), and precipitation during establishment ($t_{34} = -3.07, P = 0.0042$). These variables also had the largest standardized regression coefficients. No other variables approached significance.

Model selection using *AIC* and `proc glmselect` chose linearity, exotic abundance, and precipitation during establishment for the best model (Fig. 21.25). These were the same variables that were significant in the multiple regression. Kaul and Wilsey (2020) found these same three variables in their model search using stepwise regression, a method of model selection where variables are added or removed based on repeated tests at some α level ($\alpha = 0.15$ in this case). The different model selection methods all yielded the same result, suggesting it is a robust one. These authors conclude that high exotic species abundance interfered with the restoration process, so that the restored site shared fewer species with the seed mix used to restore it. Linearity also affected restoration, likely because highly linear sites had more edges for exotic species to invade. Presumably precipitation during establishment aided the initial success of the seed mix, and so had the opposite effect.

Note that `proc glmselect` does not provide P values for the t tests of the independent variables, because they would not be valid in this context. The Type I error rates for these tests assume a single multiple regression analysis, not a selection process where many different models were considered.

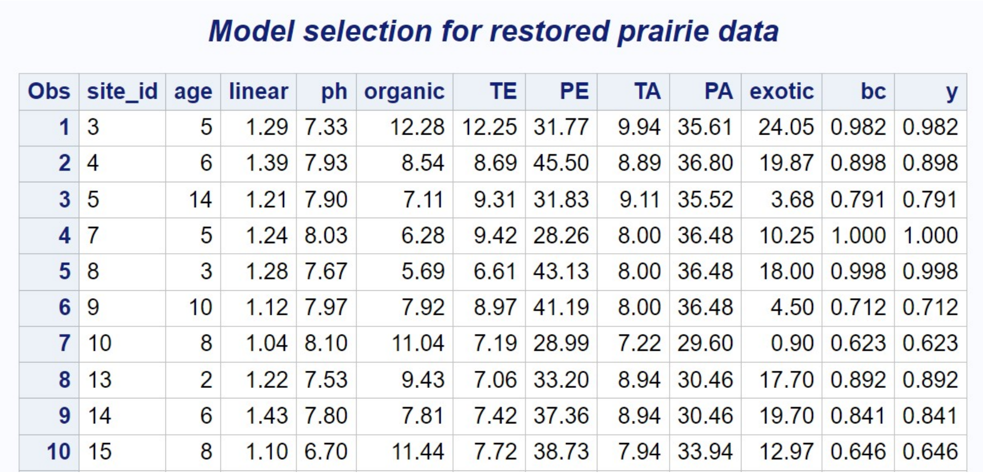
SAS Program

```
* Restored6.sas;
title "Model selection for restored prairie data";
data RPdat;
    input site_id $ age linear ph organic TE PE TA PA exotic bc;
    * Bray-Curtis (bc) measures dissimilarity of the site vs.
    restoration seed mix;
    * 0 = all species in common, 1 = none in common;
    * Kaul and Wilsey (2020) say similarity in paper;
    * Apply transformations here;
    y = bc;
    datalines;
3       5       1.29   7.33 12.28 12.25   31.77  9.94    35.61   24.05 0.982
4       6       1.39   7.93 8.54  8.69    45.50  8.89    36.80   19.87 0.898
5       14      1.21   7.90 7.11  9.31    31.83  9.11    35.52   3.68  0.791
7       5       1.24   8.03 6.28  9.42    28.26  8.00    36.48   10.25 1.000
8       3       1.28   7.67 5.69  6.61    43.13  8.00    36.48   18.00 0.998

etc.

100     17      1.30   7.87 10.42 9.17    33.58  10.72   37.59   25.03 0.970
101     3       1.26   7.93 8.64  8.56    40.66  10.72   37.59   15.37 0.772
102     11      1.25   8.10 3.61  11.69   40.36  10.00   36.28   9.57  0.972
105     13      1.05   7.93 14.54 8.53    35.57  8.00    36.48   13.77 0.718
106     10      1.02   6.97 8.19  7.64    47.79  8.00    36.48   14.25 0.624
;
run;
* Print data set;
proc print data=RPdat;
run;
* Plot y vs. x variables;
proc gplot data=RPdat;
```

```
    plot y*(age linear ph organic TE PE TA PA exotic) / vaxis=axis1
haxis=axis1;
    symbol1 i=rl v=star c=black height=2 width=2;
    axis1 label=(height=2) value=(height=2) width=3 major=(width=2) minor=none;
    symbol1 i=rl v=star c=black;
run;
* Multiple regression;
proc reg data=RPdat;
    * Specify variables in regression model;
    model y = age linear ph organic TE PE TA PA exotic / clb stb tol vif partial;
run;
* Model selection using AICc (stepwise);
proc glmselect data=RPdat;
    * Specify variables in regression model and method of selection;
    model y = age linear ph organic TE PE TA PA exotic /
    selection=stepwise(select=AICC);
run;
quit;
```



Model selection for restored prairie data

Obs	site_id	age	linear	ph	organic	TE	PE	TA	PA	exotic	bc	y
1	3	5	1.29	7.33	12.28	12.25	31.77	9.94	35.61	24.05	0.982	0.982
2	4	6	1.39	7.93	8.54	8.69	45.50	8.89	36.80	19.87	0.898	0.898
3	5	14	1.21	7.90	7.11	9.31	31.83	9.11	35.52	3.68	0.791	0.791
4	7	5	1.24	8.03	6.28	9.42	28.26	8.00	36.48	10.25	1.000	1.000
5	8	3	1.28	7.67	5.69	6.61	43.13	8.00	36.48	18.00	0.998	0.998
6	9	10	1.12	7.97	7.92	8.97	41.19	8.00	36.48	4.50	0.712	0.712
7	10	8	1.04	8.10	11.04	7.19	28.99	7.22	29.60	0.90	0.623	0.623
8	13	2	1.22	7.53	9.43	7.06	33.20	8.94	30.46	17.70	0.892	0.892
9	14	6	1.43	7.80	7.81	7.42	37.36	8.94	30.46	19.70	0.841	0.841
10	15	8	1.10	6.70	11.44	7.72	38.73	7.94	33.94	12.97	0.646	0.646

etc.

Figure 21.20: `Restored6.sas - proc print`

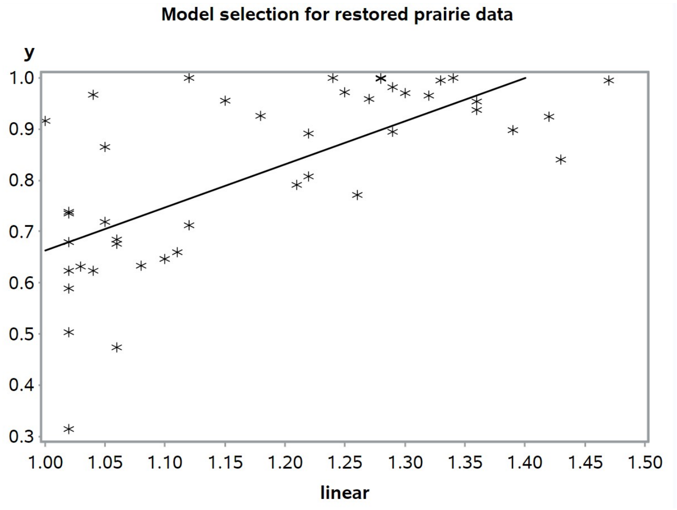


Figure 21.21: `Restored6.sas - proc gplot`

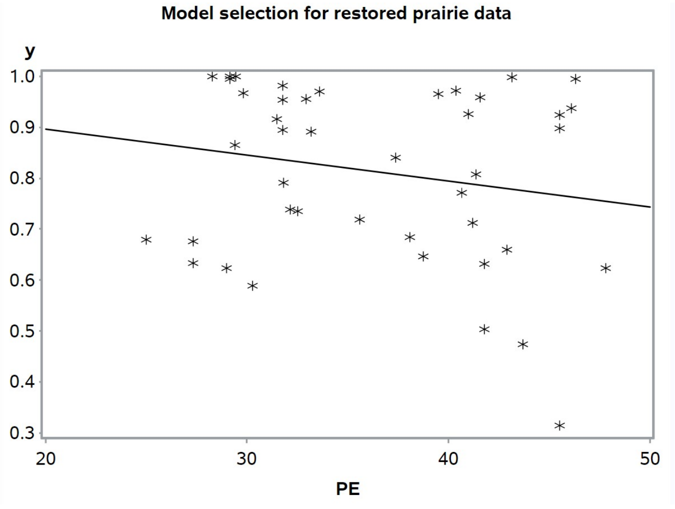


Figure 21.22: `Restored6.sas - proc gplot`

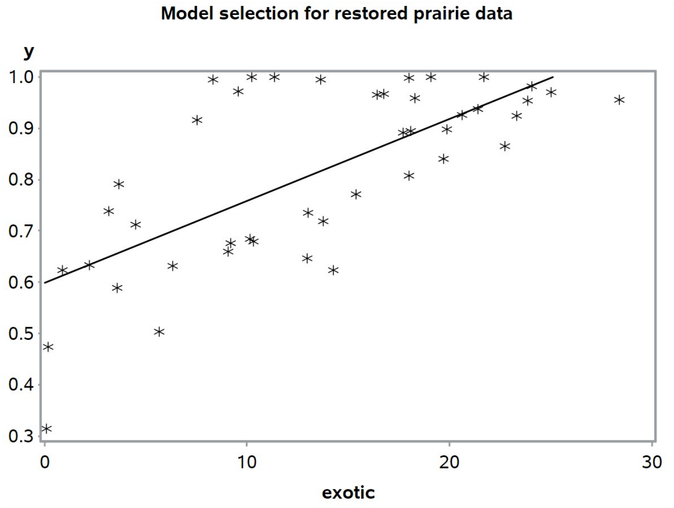
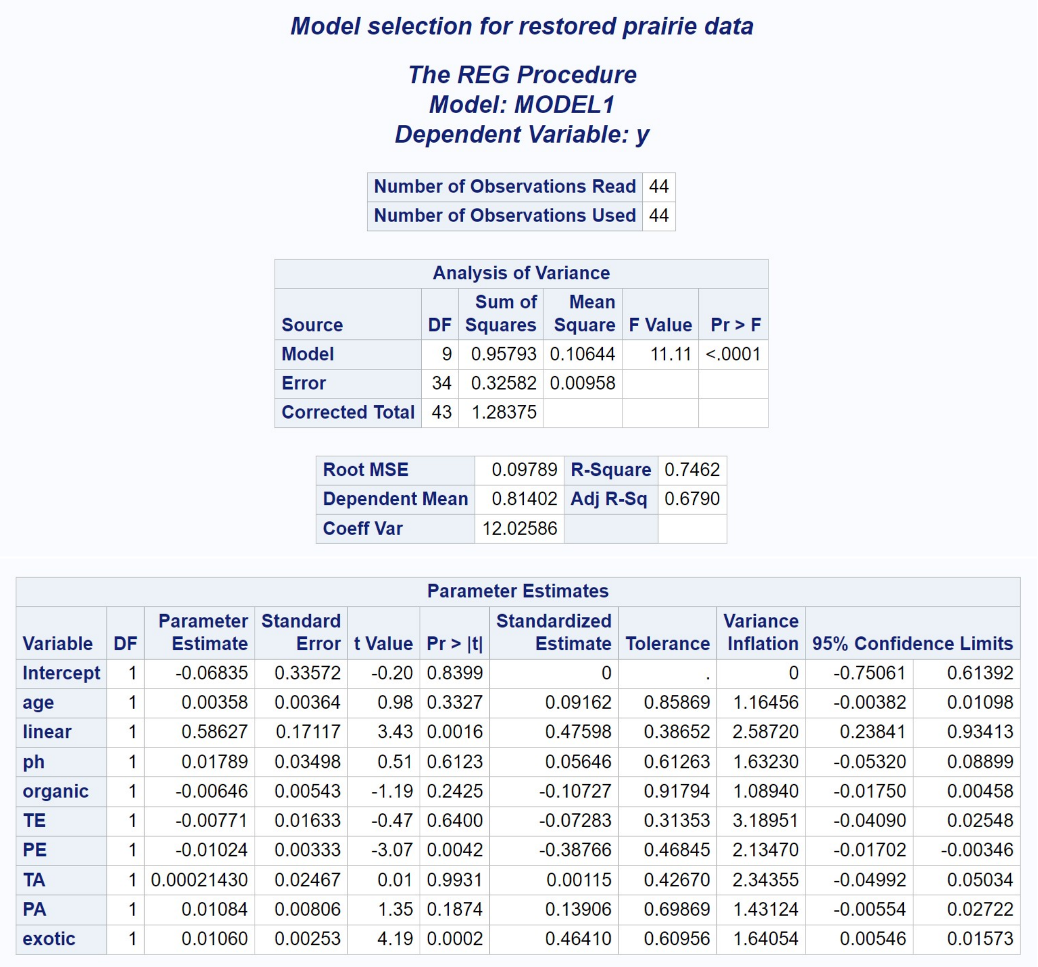


Figure 21.23: `Restored6.sas - proc gplot`



Model selection for restored prairie data

The REG Procedure
Model: MODEL1
Dependent Variable: y

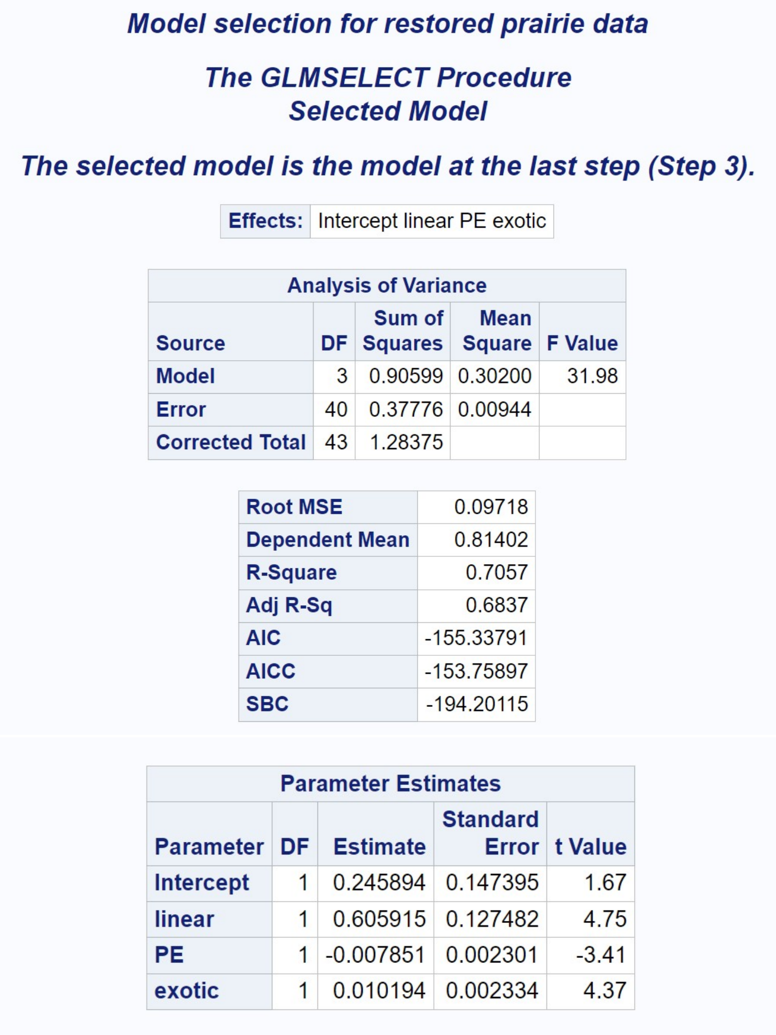
Number of Observations Read	44
Number of Observations Used	44

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	9	0.95793	0.10644	11.11	<.0001
Error	34	0.32582	0.00958		
Corrected Total	43	1.28375			

Root MSE	0.09789	**R-Square**	0.7462
Dependent Mean	0.81402	**Adj R-Sq**	0.6790
Coeff Var	12.02586		

Parameter Estimates										
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Standardized Estimate**	**Tolerance**	**Variance Inflation**	**95% Confidence Limits**	
Intercept	1	-0.06835	0.33572	-0.20	0.8399	0	.	0	-0.75061	0.61392
age	1	0.00358	0.00364	0.98	0.3327	0.09162	0.85869	1.16456	-0.00382	0.01098
linear	1	0.58627	0.17117	3.43	0.0016	0.47598	0.38652	2.58720	0.23841	0.93413
ph	1	0.01789	0.03498	0.51	0.6123	0.05646	0.61263	1.63230	-0.05320	0.08899
organic	1	-0.00646	0.00543	-1.19	0.2425	-0.10727	0.91794	1.08940	-0.01750	0.00458
TE	1	-0.00771	0.01633	-0.47	0.6400	-0.07283	0.31353	3.18951	-0.04090	0.02548
PE	1	-0.01024	0.00333	-3.07	0.0042	-0.38766	0.46845	2.13470	-0.01702	-0.00346
TA	1	0.00021430	0.02467	0.01	0.9931	0.00115	0.42670	2.34355	-0.04992	0.05034
PA	1	0.01084	0.00806	1.35	0.1874	0.13906	0.69869	1.43124	-0.00554	0.02722
exotic	1	0.01060	0.00253	4.19	0.0002	0.46410	0.60956	1.64054	0.00546	0.01573

Figure 21.24: `Restored6.sas - proc reg`



Model selection for restored prairie data

The GLMSELECT Procedure
Selected Model

The selected model is the model at the last step (Step 3).

Effects:	Intercept linear PE exotic

Analysis of Variance				
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**
Model	3	0.90599	0.30200	31.98
Error	40	0.37776	0.00944	
Corrected Total	43	1.28375		

Root MSE	0.09718
Dependent Mean	0.81402
R-Square	0.7057
Adj R-Sq	0.6837
AIC	-155.33791
AICC	-153.75897
SBC	-194.20115

Parameter Estimates				
Parameter	**DF**	**Estimate**	**Standard Error**	**t Value**
Intercept	1	0.245894	0.147395	1.67
linear	1	0.605915	0.127482	4.75
PE	1	-0.007851	0.002301	-3.41
exotic	1	0.010194	0.002334	4.37

Figure 21.25: `Restored6.sas - proc glmselect`

21.16 References

Aho, K., Derryberry, D., & Peterson, T. (2014) Model selection for ecologists: the worldviews of AIC and BIC. *Ecology* 95: 631-636.

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control* AC-19: 716-723.

Anderson, D. R., Burnham, K. P., & Thompson, W. L. (2000) Null hypothesis testing: problems, prevalence, and an alternative approach. *The Journal of Wildlife Management* 64: 912-923.

Brewer, M. J., Butler, A., & Cooksley, S. L. (2016) The relative performance of AIC, AICc and BIC in the presence of unobserved heterogeneity. *Methods in Ecology and Evolution* 7: 679-692.

Burnham, K. P. & Anderson, D. R. (2002) *Model Selection and Inference: A Practical Information-Theoretic Approach, Second Edition.* Springer-Verlag, New York, NY.

Burnham, K. P. & Anderson, D. R. (2014) *P* values are only an index to evidence: 20th- vs. 21st-century statistical science. *Ecology* 95: 627-630.

Chang, W. (2023) *R Graphics Cookbook, Second Edition.* https://r-graphics.org/

Coulson, R. N., Mayyasi, A. M., Foltz, J. L., Hain, F. P. & Martin, W. C. (1976) Resource utilization by the southern pine beetle, *Dendroctonus frontalis* (Coleoptera: Scolytidae). *Canadian Entomologist* 108: 353-362.

de Valpine, P. (2014) The common sense of *P* values. *Ecology* 95: 617-621.

Draper, N. R. & Smith, H. (1981) *Applied Regression Analysis, Second Edition.* John Wiley & Sons, New York, NY.

Gotelli, N. J. (2008) *A Primer of Ecology, Fourth Edition.* Sinauer Associates, Inc., Sunderland, MA.

Hofstetter, R. W., Klepzig, K. D., Moser, J. C. & Ayres, M. P. (2006) Seasonal dynamics of mites and fungi and their interaction with southern pine beetle. *Enviromental Entomology* 35: 22-30.

Kaul, A. D., & Wilsey, B. J. (2020) Exotic species drive patterns of plant species diversity in 93 restored tallgrass prairies. *Ecological Applications* (2020): e2252. doi: 10.1002/eap.2252.

Kutner, M. H., Nachtsheim, C. J., Neter, J. & Li, W. (2005) *Applied Linear Statistical Models, Fifth Edition.*

McCulloch, C. E. & Searle, S. R. (2001) *Generalized, Linear, and Mixed Models.* John Wiley & Sons, Inc., New York, NY.

Murtaugh, P. A. (2014) In defense of *P* values. *Ecology* 95: 611-617.

Reeve, J. D., Rhodes, D. J. & Turchin, P. (1998) Scramble competition in southern pine beetle (Coleoptera: Scolytidae). *Ecological Entomology* 23: 433-443.

Ricker, W. E. (1954) Stock and recruitment. *Journal of the Fisheries Research Board of Canada* 11: 559623.

SAS Institute Inc. (2016) *SAS/GRAPH 9.4: Reference, Fifth Edition.* SAS Institute Inc., Cary, NC.

SAS Institute Inc. (2018a) *SAS/IML 15.1 User's Guide.* SAS Institute Inc., Cary, NC.

SAS Institute Inc. (2018b) *SAS/STAT 15.1 Users Guide.* SAS Institute Inc., Cary, NC.

Schwarz, G. (1978) Estimating the dimension of a model. *Annals of Statistics* 6: 461-464.

Sheather, S. J. (2009) *A Modern Approach to Regression with R.* Springer Science+Business Media LLC, New York, NY.

Soul, L. D., Benson, R. B. J., & Weisbecker, V. (2013) Multiple regression modeling for estimating the endocranial volume in extinct Mammalia. *Paleobiology* 39: 149-162.

Tabachnick, B. G., & Fidell, L. S. (2001) *Using Multivariate Statistics, Fourth Edition.* Allyn and Bacon, Boston, MA.

21.17 Problems

1. This problem involves the matrix calculations for linear regression, a special case of multiple regression. Suppose you have a data set with four observations:

Y_i	X_i
4	1
6	2
9	3
10	4

   (a) What is the design matrix $\boldsymbol{X}$ and the vector $\boldsymbol{Y}$ for this data set?

   (b) What is the transpose of $\boldsymbol{X}$, or $\boldsymbol{X'}$? The answer should be a 2×4 matrix.

   (c) Calculate $\boldsymbol{X'X}$ using matrix multiplication. The answer should be a 2×2 matrix.

   (d) Show that the matrix below is the inverse of $\boldsymbol{X'X}$, by multiplying them together to obtain $\boldsymbol{I}$ (the identity matrix).

$$(\boldsymbol{X'X})^{-\mathbf{1}} = \begin{pmatrix} 1.5 & -0.5 \\ -0.5 & 0.2 \end{pmatrix} \tag{21.63}$$

   (e) Calculate $(\boldsymbol{X'X})^{-\mathbf{1}}\boldsymbol{X'}$ using matrix multiplication. The answer should be a 2×4 matrix.

   (f) Finally, calculate $\boldsymbol{\beta} = (\boldsymbol{X'X})^{-\mathbf{1}}\boldsymbol{X'Y}$ using matrix multiplication. The answer should be a 2×1 matrix, with elements equal to the regression intercept and slope. You can check your answer by running a linear regression (see Chapter 17).

2. Ecologists who study predator-prey interactions are often interested in the mortality inflicted by the predator as a function of prey abundance. Data were collected on the proportion of prey eaten by a single predator as the number of prey were increased, in a laboratory experiment (see table below). The proportion eaten was an average over multiple replicates.

   (a) Fit a flexible model to these observations using polynomial regression and SAS. What order polynomial was needed to describe these observations? Attach your program and output.

   (b) Use the polynomial model to predict the proportion eaten for 35 and 45 prey, including confidence intervals for the predictions.

   (c) The proportion eaten vs. prey curve can take different shapes depending on the **functional response** of the predator. For example, the curve would be flat for a Type I response, strictly decreasing for a Type II response, and hump-shaped for a Type III response (Gotelli 2008). How would you classify the response in this experiment?

Number of Prey	Proportion Eaten
1	0.00
2	0.05
3	0.10
4	0.13
5	0.14
7	0.21
10	0.24
15	0.31
20	0.39
25	0.39
30	0.42
40	0.40
50	0.30

3. Data were collected on the abundance of an insect species (Y) as a function of five environmental variables (X_1, X_2, X_3, X_4, and X_5). See table below.

   (a) Conduct a multiple regression analysis of these data using SAS, with Y the dependent variable and X_1, X_2, X_3, X_4, and X_5 the regressors. Discuss the significance of the overall test of the model and the tests for each independent variable. Attach your program and output.

   (b) Use standardized regression coefficients to compare the size and direction of the different effects, especially the significant ones. Which independent variables have the most effect on insect abundance? Which ones increase or decrease it?

   (c) Select the best model for these data using AIC, and write the answer below. Attach your SAS program and output.

X_1	X_2	X_3	X_4	X_5	Y
13.4	10.4	12.7	126	15.4	18.6
11.4	11.9	8.7	115	12.8	21.2
9.2	13.3	10.3	143	11.2	24.1
15.5	11.0	12.6	88	14.9	18.9
13.1	9.9	10.3	156	11.9	26.4
16.3	13.2	9.2	146	14.0	19.8
10.1	15.5	8.9	135	11.3	21.8
7.6	10.0	16.7	128	7.0	26.7
12.9	11.7	16.4	89	13.2	18.6
11.0	11.3	14.8	171	12.3	20.8
11.1	13.3	6.4	128	14.1	21.6
14.3	10.1	9.3	92	15.1	18.4
10.1	13.7	11.3	129	13.0	18.5
12.2	8.9	11.9	143	12.9	24.8
10.9	10.3	12.8	154	12.3	27.9
12.6	13.4	13.7	177	16.1	22.6
12.6	12.4	11.1	165	16.4	26.5
12.4	10.2	7.1	118	13.9	23.9
15.2	13.4	8.1	111	14.2	19.5
13.7	14.0	11.3	123	12.3	16.3
14.1	16.5	7.0	58	7.8	7.1
20.5	8.5	8.1	143	11.9	27.3
5.9	12.0	8.8	149	8.2	27.2
12.7	13.8	19.3	122	12.7	19.7
15.2	9.5	11.9	126	9.2	24.0
17.5	14.0	16.0	130	15.0	19.1
7.7	10.0	10.0	81	9.9	16.0
16.7	13.8	9.3	132	11.2	18.4
14.9	16.9	11.1	124	11.4	17.2
10.6	15.3	13.1	145	15.4	19.4

Chapter 22

Data Sets

22.1 Elytra Length

Elytra length of male and female clerid beetles (*Thanasimus dubius*) including a sample SAS `data` step. Data drawn from Reeve et al. (2003).

```
data elytra;
    input sex $ length;
    datalines;
M   4.9
F   5.2
M   4.9
F   4.2
F   5.7
M   4.6
M   3.8
F   5.4
F   4.0
F   4.5
M   4.9
F   5.2
M   4.9
F   4.2
F   5.7
M   4.6
M   3.8
F   5.4
F   4.0
F   4.5
F   5.2
F   4.9
M   5.0
M   4.4
M   5.0
M   5.0
M   4.9
F   4.5
F   4.5
M   5.1
F   5.5
M   4.8
F   4.9
M   4.8
M   4.5
```

```
M   4.5
M   4.4
M   5.2
M   4.1
F   5.0
M   4.4
F   4.9
M   4.7
M   4.4
F   4.8
F   4.5
M   4.0
M   3.4
F   5.5
M   4.7
M   4.8
F   4.8
F   3.7
M   5.3
M   4.6
F   4.8
M   4.5
M   5.0
M   4.4
F   4.6
M   4.4
M   4.9
F   5.3
F   5.0
F   4.7
F   5.2
M   5.0
M   5.0
M   4.8
M   5.8
F   5.7
F   5.2
M   4.9
F   5.1
F   5.3
F   5.3
F   5.9
F   5.3
M   4.5
F   5.2
```

```
M   5.1
F   4.6
M   4.8
M   3.5
F   4.6
F   5.3
M   5.2
F   4.8
M   5.1
M   5.2
M   4.9
M   5.3
M   5.2
F   4.9
F   5.6
M   5.0
M   5.0
F   5.1
M   5.1
F   5.5
M   5.1
F   4.8
F   4.9
F   5.0
M   4.9
M   5.0
F   5.0
M   4.9
M   4.8
F   5.2
F   4.8
M   4.7
F   5.1
M   4.5
M   5.0
F   5.4
F   4.6
M   4.0
M   4.2
F   5.2
F   4.6
M   5.0
M   3.7
M   4.6
M   4.0
```

```
M   5.1
F   4.4
M   4.8
M   4.6
F   3.7
;
run;
```

22.2 Development Time

Development times for the clerid beetle *Thanasimus dubius* The variables `time_pp` and `time_adult` are the development time from the larval to the prepupal stage, and the prepupal to the adult stage, respectively (Reeve et al. 2003).

```
data devel_time;
    input time_pp time_adult;
    datalines;
34  65
31  48
29   .
30  55
32  62
32  47
37  44
34  53
31   .
37  53
32   .
31  42
29   .
35   .
39   .
34  43
32   .
34   .
34 113
32  47
32 100
41   .
32  49
29   .
32  53
39   .
39  84
35   .
32   .
35  74
36  43
31  50
34   .
```

```
35  44
35 116
34   .
34   .
37  58
36 101
32  67
34  68
34  61
28  66
31  84
30  68
28 106
28  42
31  58
31  42
28  68
32  55
32   .
30 101
30  99
39  43
30  80
28  52
27  50
28 110
28  42
30   .
28  66
28 147
27   .
37 135
30 119
29 113
30 103
30  95
27  87
29  89
33   .
27  76
27   .
30   .
30  49
30  81
29  85
```

```
27   .
31 104
27  73
27 110
27   .
31  99
31  55
31  59
27   .
30  93
27   .
28  84
28  93
29   .
29 108
31 103
33   .
29  92
;
run;
```

22.3 Plant Biomass

Effect of nitrogen heterogeneity, nitrogen availability, and water availability on the total biomass of grassland plants grown in microcosms (Maestre & Reynolds 2007).

```
data maestre;
    input nitrohet $ nitrogen water biomass;
    datalines;
N   40  125   4.372
N   40  125   4.482
N   40  125   4.221
N   40  125   3.977
N   40  250   7.400
N   40  250   8.027
N   40  250   7.883
N   40  250   7.769
N   40  375   7.226
N   40  375   8.126
N   40  375   6.840
N   40  375   7.901
N   80  125   5.140
N   80  125   3.913
N   80  125   4.669
N   80  125   4.306
N   80  250   9.099
N   80  250   9.711
N   80  250   9.123
N   80  250   9.709
N   80  375  10.701
N   80  375  11.552
N   80  375  11.356
N   80  375   9.759
N  120  125   5.021
N  120  125   4.970
N  120  125   5.055
N  120  125   4.862
N  120  250   9.029
N  120  250  10.791
N  120  250   9.115
N  120  250  10.319
N  120  375  12.189
N  120  375  14.381
```

```
N  120  375  13.153
N  120  375  14.066
Y   40  125   5.458
Y   40  125   5.017
Y   40  125   5.479
Y   40  125   5.714
Y   40  250   8.972
Y   40  250   9.234
Y   40  250   8.032
Y   40  250   8.372
Y   40  375   9.464
Y   40  375   9.563
Y   40  375   9.385
Y   40  375   8.226
Y   80  125   6.616
Y   80  125   6.909
Y   80  125   6.851
Y   80  125   6.098
Y   80  250  10.792
Y   80  250  10.164
Y   80  250  10.947
Y   80  250   9.582
Y   80  375  14.936
Y   80  375  13.607
Y   80  375  14.231
Y   80  375  12.038
Y  120  125   7.389
Y  120  125   6.683
Y  120  125   7.759
Y  120  125   6.752
Y  120  250  10.731
Y  120  250  12.640
Y  120  250  10.350
Y  120  250  11.550
Y  120  375  14.697
Y  120  375  17.826
Y  120  375  14.711
Y  120  375  13.614
;
run;
```

22.4 *Anagrus* fecundity

Fecundity for the parasitoid *Anagrus delicatus* collected from different sites, with 14 isolines per site and eight individual wasps per isoline. The data were simulated from the results presented in Cronin and Strong (1996).

```
data anagrus;
    input site isoline wasp eggs;
    datalines;
1   1  1  37
1   1  2  41
1   1  3  46
1   1  4  44
1   1  5  43
1   1  6  41
1   1  7  38
1   1  8  37
1   2  1  37
1   2  2  28
1   2  3  34
1   2  4  37
1   2  5  35
1   2  6  39
1   2  7  36
1   2  8  29
1   3  1  35
1   3  2  37
1   3  3  40
1   3  4  39
1   3  5  37
1   3  6  44
1   3  7  35
1   3  8  38
1   4  1  28
1   4  2  36
1   4  3  31
1   4  4  27
1   4  5  36
1   4  6  33
1   4  7  31
1   4  8  35
1   5  1  34
1   5  2  35
```

```
1   5  3  30
1   5  4  39
1   5  5  42
1   5  6  39
1   5  7  38
1   5  8  32
1   6  1  30
1   6  2  32
1   6  3  35
1   6  4  35
1   6  5  32
1   6  6  31
1   6  7  34
1   6  8  30
1   7  1  30
1   7  2  36
1   7  3  37
1   7  4  30
1   7  5  41
1   7  6  35
1   7  7  34
1   7  8  37
1   8  1  25
1   8  2  31
1   8  3  24
1   8  4  26
1   8  5  30
1   8  6  31
1   8  7  25
1   8  8  24
1   9  1  34
1   9  2  35
1   9  3  29
1   9  4  34
1   9  5  34
1   9  6  40
1   9  7  37
1   9  8  37
1  10  1  38
1  10  2  30
1  10  3  33
1  10  4  32
1  10  5  33
1  10  6  34
1  10  7  35
```

```
1  10  8  41
1  11  1  36
1  11  2  33
1  11  3  36
1  11  4  34
1  11  5  37
1  11  6  41
1  11  7  37
1  11  8  31
1  12  1  35
1  12  2  36
1  12  3  35
1  12  4  37
1  12  5  40
1  12  6  34
1  12  7  29
1  12  8  42
1  13  1  33
1  13  2  39
1  13  3  33
1  13  4  37
1  13  5  28
1  13  6  35
1  13  7  34
1  13  8  38
1  14  1  35
1  14  2  33
1  14  3  25
1  14  4  29
1  14  5  29
1  14  6  35
1  14  7  33
1  14  8  29
2   1  1  26
2   1  2  39
2   1  3  36
2   1  4  27
2   1  5  25
2   1  6  31
2   1  7  30
2   1  8  25
2   2  1  42
2   2  2  46
2   2  3  46
2   2  4  42
```

2	2	5	43
2	2	6	36
2	2	7	36
2	2	8	41
2	3	1	38
2	3	2	36
2	3	3	35
2	3	4	31
2	3	5	36
2	3	6	32
2	3	7	29
2	3	8	34
2	4	1	28
2	4	2	36
2	4	3	33
2	4	4	32
2	4	5	27
2	4	6	31
2	4	7	30
2	4	8	32
2	5	1	30
2	5	2	35
2	5	3	32
2	5	4	31
2	5	5	36
2	5	6	34
2	5	7	29
2	5	8	36
2	6	1	28
2	6	2	34
2	6	3	34
2	6	4	35
2	6	5	32
2	6	6	31
2	6	7	24
2	6	8	31
2	7	1	35
2	7	2	34
2	7	3	44
2	7	4	34
2	7	5	35
2	7	6	36
2	7	7	32
2	7	8	30
2	8	1	37

```
2   8  2  32
2   8  3  33
2   8  4  39
2   8  5  30
2   8  6  31
2   8  7  32
2   8  8  34
2   9  1  41
2   9  2  41
2   9  3  43
2   9  4  36
2   9  5  43
2   9  6  42
2   9  7  42
2   9  8  37
2  10  1  34
2  10  2  30
2  10  3  35
2  10  4  27
2  10  5  30
2  10  6  22
2  10  7  31
2  10  8  31
2  11  1  34
2  11  2  36
2  11  3  38
2  11  4  36
2  11  5  34
2  11  6  33
2  11  7  35
2  11  8  29
2  12  1  28
2  12  2  29
2  12  3  27
2  12  4  36
2  12  5  33
2  12  6  32
2  12  7  34
2  12  8  32
2  13  1  40
2  13  2  39
2  13  3  39
2  13  4  34
2  13  5  32
2  13  6  42
```

```
2  13  7  36
2  13  8  39
2  14  1  38
2  14  2  42
2  14  3  37
2  14  4  37
2  14  5  34
2  14  6  33
2  14  7  43
2  14  8  34
3   1  1  30
3   1  2  35
3   1  3  36
3   1  4  37
3   1  5  29
3   1  6  27
3   1  7  39
3   1  8  38
3   2  1  30
3   2  2  37
3   2  3  30
3   2  4  31
3   2  5  27
3   2  6  31
3   2  7  36
3   2  8  40
3   3  1  27
3   3  2  33
3   3  3  31
3   3  4  32
3   3  5  34
3   3  6  31
3   3  7  31
3   3  8  31
3   4  1  26
3   4  2  27
3   4  3  37
3   4  4  30
3   4  5  29
3   4  6  35
3   4  7  34
3   4  8  31
3   5  1  36
3   5  2  32
3   5  3  34
```

```
3   5  4  37
3   5  5  32
3   5  6  34
3   5  7  33
3   5  8  32
3   6  1  33
3   6  2  40
3   6  3  34
3   6  4  38
3   6  5  36
3   6  6  35
3   6  7  41
3   6  8  34
3   7  1  31
3   7  2  33
3   7  3  31
3   7  4  34
3   7  5  29
3   7  6  33
3   7  7  28
3   7  8  33
3   8  1  22
3   8  2  25
3   8  3  29
3   8  4  24
3   8  5  24
3   8  6  26
3   8  7  25
3   8  8  21
3   9  1  32
3   9  2  31
3   9  3  28
3   9  4  28
3   9  5  35
3   9  6  34
3   9  7  33
3   9  8  31
3  10  1  31
3  10  2  32
3  10  3  29
3  10  4  30
3  10  5  28
3  10  6  31
3  10  7  28
3  10  8  36
```

```
3  11  1  32
3  11  2  31
3  11  3  34
3  11  4  35
3  11  5  35
3  11  6  31
3  11  7  41
3  11  8  34
3  12  1  28
3  12  2  27
3  12  3  27
3  12  4  27
3  12  5  27
3  12  6  30
3  12  7  28
3  12  8  28
3  13  1  36
3  13  2  39
3  13  3  36
3  13  4  30
3  13  5  37
3  13  6  32
3  13  7  38
3  13  8  39
3  14  1  32
3  14  2  34
3  14  3  41
3  14  4  33
3  14  5  35
3  14  6  35
3  14  7  34
3  14  8  31
;
run;
```

22.5 Fitness of *T. dubius*

Fitness of adult *T. dubius*, a bark beetle predator, reared on an artificial diet as larvae vs. wild individuals collected from the field (Reeve et al. 2003). The adults were fed either *Ips grandicollis* or cowpea weevils.

```
data fitness;
    input eggs longevity length treat $;
    datalines;
290    78   5.7   DietIG
 99    40   5.2   DietIG
340    70   5.5   DietIG
271    67   4.8   DietIG
200    84   5.2   DietIG
405    80   5.2   DietIG
178    80   5.1   DietIG
 48    23   5.0   DietIG
146    62   4.8   DietIG
184    82   4.9   DietIG
 66    67   4.6   DietCPW
 93    45   5.0   DietCPW
  9    49   5.4   DietCPW
404   121   5.4   DietCPW
244   114   5.1   DietCPW
195    72   4.9   DietCPW
343   126   5.2   DietCPW
516   138   5.0   DietCPW
215   108   4.6   DietCPW
412   156   5.6   DietCPW
167    79   4.8   DietCPW
316   117   5.2   DietCPW
334   127   5.3   DietCPW
 62   221   4.7   WildCPW
290   180   5.0   WildCPW
488   175   5.8   WildCPW
336   177   5.2   WildCPW
337   164   5.8   WildCPW
230    93   5.0   WildCPW
381   155   5.3   WildCPW
192   152   5.5   WildCPW
186   143   5.3   WildCPW
467   140   5.2   WildCPW
 59    42   4.9   WildCPW
```

```
323   138   5.7   WildCPW
291   117   4.9   WildCPW
164   112   5.3   WildCPW
142   112   5.3   WildCPW
269   110   5.0   WildCPW
329    91   5.4   WildCPW
235    84   5.0   WildCPW
;
run;
```

22.6 *Iris* flower measurements

Sepal and petal measurements for *I. setosa* (Fisher 1936).

```
data iris;
    input seplen sepwid petlen petwid;
    datalines;
5.1 3.5 1.4 0.2
4.9 3.0 1.4 0.2
4.7 3.2 1.3 0.2
4.6 3.1 1.5 0.2
5.0 3.6 1.4 0.2
5.4 3.9 1.7 0.4
4.6 3.4 1.4 0.3
5.0 3.4 1.5 0.2
4.4 2.9 1.4 0.2
4.9 3.1 1.5 0.1
5.4 3.7 1.5 0.2
4.8 3.4 1.6 0.2
4.8 3.0 1.4 0.1
4.3 3.0 1.1 0.1
5.8 4.0 1.2 0.2
5.7 4.4 1.5 0.4
5.4 3.9 1.3 0.4
5.1 3.5 1.4 0.3
5.7 3.8 1.7 0.3
5.1 3.8 1.5 0.3
5.4 3.4 1.7 0.2
5.1 3.7 1.5 0.4
4.6 3.6 1.0 0.2
5.1 3.3 1.7 0.5
4.8 3.4 1.9 0.2
5.0 3.0 1.6 0.2
5.0 3.4 1.6 0.4
5.2 3.5 1.5 0.2
5.2 3.4 1.4 0.2
4.7 3.2 1.6 0.2
4.8 3.1 1.6 0.2
5.4 3.4 1.5 0.4
5.2 4.1 1.5 0.1
5.5 4.2 1.4 0.2
4.9 3.1 1.5 0.2
5.0 3.2 1.2 0.2
5.5 3.5 1.3 0.2
```

```
4.9 3.6 1.4 0.1
4.4 3.0 1.3 0.2
5.1 3.4 1.5 0.2
5.0 3.5 1.3 0.3
4.5 2.3 1.3 0.3
4.4 3.2 1.3 0.2
5.0 3.5 1.6 0.6
5.1 3.8 1.9 0.4
4.8 3.0 1.4 0.3
5.1 3.8 1.6 0.2
4.6 3.2 1.4 0.2
5.3 3.7 1.5 0.2
5.0 3.3 1.4 0.2
;
run;
```

22.7 References

Cronin, J. T. & Strong, D. R. (1996) Genetics of oviposition success of a thelytokous fairyfly parasitoid, *Anagrus delicatus*. *Heredity* 76: 43-54.

Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 7: 179188.

Maestre, F. T. & Reynolds, J. F. (2007) Amount or pattern? Grassland responses to the heterogeneity and availability of two key resources. *Ecology* 88: 501-511.

Reeve, J. D., Rojas, M. G. & Morales-Ramos, J. A. (2003) Artificial diet and rearing methods for *Thanasimus dubius* (Coleoptera: Cleridae), a predator of bark beetles (Coleoptera: Scolytidae). *Biological Control* 27: 315-322.

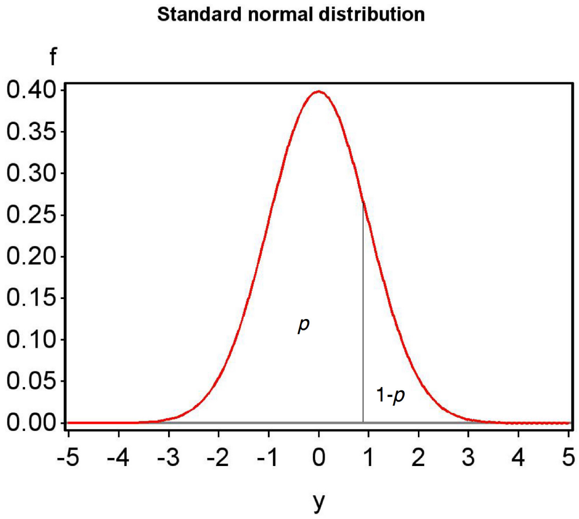
Chapter 23

Statistical Tables

23.1 Table Z: Probabilities for the standard normal distribution.

Suppose a random variable Z has a standard normal distribution ($Z \sim N(0,1)$). This table gives $P[Z < z] = p$ where the first two digits of z are given on the left, while the last digit is given in the top row. The values in the table were generated using the SAS function `probnorm` (SAS Institute Inc. 2016).

Figure 23.1: Plot of the standard normal distribution illustrating the probability shown in the table below.



References

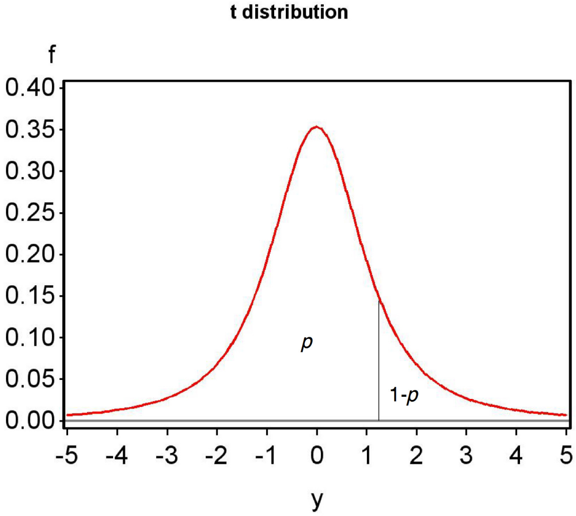
SAS Institute Inc. (2016) *SAS 9.4 Functions and CALL Routines: Reference, Fifth Edition.* SAS Institute Inc., Cary, NC.

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999

23.2 Table T: Quantiles of the t distribution

Suppose a random variable T has a t distribution. This table gives values of the quantile q such that $P[T < q] = p$, where $p = 0.75, 0.9, ..., 0.9995$. Degrees of freedom are given on the left. The values in the table were generated using the SAS function `tinv` (SAS Institute Inc. 2016).

Figure 23.2: Plot of the t distribution illustrating p and $1 - p$ in the table below.



References

SAS Institute Inc. (2016) *SAS 9.4 Functions and CALL Routines: Reference, Fifth Edition.* SAS Institute Inc., Cary, NC.

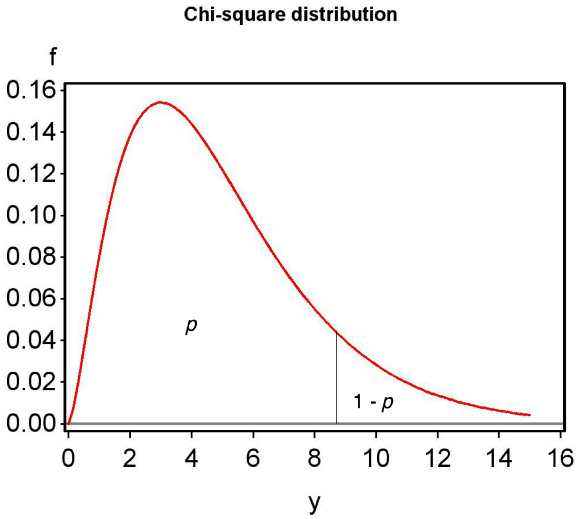
	p	0.75	0.90	0.95	0.975	0.990	0.995	0.9995
	$1-p$	0.25	0.10	0.05	0.025	0.010	0.005	0.0005
	$2(1-p)$	0.50	0.20	0.10	0.050	0.020	0.010	0.0010
df	1	1.000	3.078	6.314	12.706	31.821	63.657	636.619
	2	0.816	1.886	2.920	4.303	6.965	9.925	31.599
	3	0.765	1.638	2.353	3.182	4.541	5.841	12.924
	4	0.741	1.533	2.132	2.776	3.747	4.604	8.610
	5	0.727	1.476	2.015	2.571	3.365	4.032	6.869
	6	0.718	1.440	1.943	2.447	3.143	3.707	5.959
	7	0.711	1.415	1.895	2.365	2.998	3.499	5.408
	8	0.706	1.397	1.860	2.306	2.896	3.355	5.041
	9	0.703	1.383	1.833	2.262	2.821	3.250	4.781
	10	0.700	1.372	1.812	2.228	2.764	3.169	4.587
	11	0.697	1.363	1.796	2.201	2.718	3.106	4.437
	12	0.695	1.356	1.782	2.179	2.681	3.055	4.318
	13	0.694	1.350	1.771	2.160	2.650	3.012	4.221
	14	0.692	1.345	1.761	2.145	2.624	2.977	4.140
	15	0.691	1.341	1.753	2.131	2.602	2.947	4.073
	16	0.690	1.337	1.746	2.120	2.583	2.921	4.015
	17	0.689	1.333	1.740	2.110	2.567	2.898	3.965
	18	0.688	1.330	1.734	2.101	2.552	2.878	3.922
	19	0.688	1.328	1.729	2.093	2.539	2.861	3.883
	20	0.687	1.325	1.725	2.086	2.528	2.845	3.850
	21	0.686	1.323	1.721	2.080	2.518	2.831	3.819
	22	0.686	1.321	1.717	2.074	2.508	2.819	3.792
	23	0.685	1.319	1.714	2.069	2.500	2.807	3.768
	24	0.685	1.318	1.711	2.064	2.492	2.797	3.745
	25	0.684	1.316	1.708	2.060	2.485	2.787	3.725
	26	0.684	1.315	1.706	2.056	2.479	2.779	3.707
	27	0.684	1.314	1.703	2.052	2.473	2.771	3.690
	28	0.683	1.313	1.701	2.048	2.467	2.763	3.674
	29	0.683	1.311	1.699	2.045	2.462	2.756	3.659
	30	0.683	1.310	1.697	2.042	2.457	2.750	3.646
	31	0.682	1.309	1.696	2.040	2.453	2.744	3.633
	32	0.682	1.309	1.694	2.037	2.449	2.738	3.622
	33	0.682	1.308	1.692	2.035	2.445	2.733	3.611
	34	0.682	1.307	1.691	2.032	2.441	2.728	3.601
	35	0.682	1.306	1.690	2.030	2.438	2.724	3.591

	p	0.75	0.90	0.95	0.975	0.990	0.995	0.9995
	$1-p$	0.25	0.10	0.05	0.025	0.010	0.005	0.0005
	$2(1-p)$	0.50	0.20	0.10	0.050	0.020	0.010	0.0010
df	36	0.681	1.306	1.688	2.028	2.434	2.719	3.582
	37	0.681	1.305	1.687	2.026	2.431	2.715	3.574
	38	0.681	1.304	1.686	2.024	2.429	2.712	3.566
	39	0.681	1.304	1.685	2.023	2.426	2.708	3.558
	40	0.681	1.303	1.684	2.021	2.423	2.704	3.551
	50	0.679	1.299	1.676	2.009	2.403	2.678	3.496
	60	0.679	1.296	1.671	2.000	2.390	2.660	3.460
	70	0.678	1.294	1.667	1.994	2.381	2.648	3.435
	80	0.678	1.292	1.664	1.990	2.374	2.639	3.416
	90	0.677	1.291	1.662	1.987	2.368	2.632	3.402
	100	0.677	1.290	1.660	1.984	2.364	2.626	3.390
	∞	0.674	1.282	1.645	1.960	2.326	2.576	3.291

23.3 Table C: Quantiles of the χ^2 distribution

Suppose a random variable X has a χ^2 distribution with *df* degrees of freedom. This table gives values of the quantile q such that $P[X < q] = p$, where $p = 0.005, ..., 0.999$. The values in the table were generated using the SAS function `cinv` (SAS Institute Inc. 2016).

Figure 23.3: Plot of the χ^2 distribution ($df = 5$) illustrating p and $1 - p$ in the table below.



References

SAS Institute Inc. (2016) *SAS 9.4 Functions and CALL Routines: Reference, Fifth Edition*. SAS Institute Inc., Cary, NC.

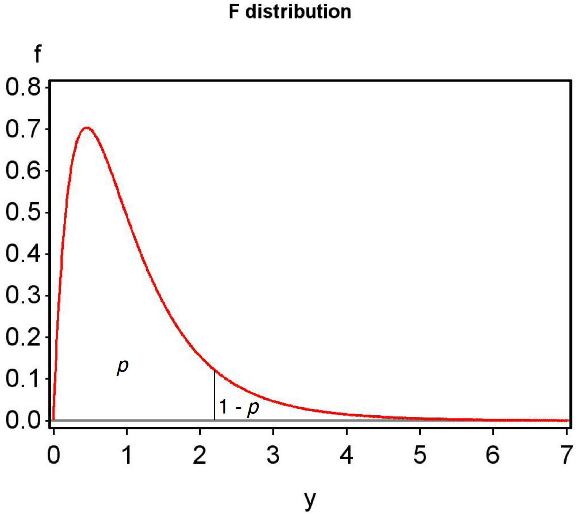
	p	0.005	0.010	0.025	0.050	0.100	0.250	0.500	0.750	0.900	0.950	0.975	0.990	0.995	0.999
	$1-p$	0.995	0.990	0.975	0.950	0.900	0.750	0.500	0.250	0.100	0.050	0.025	0.010	0.005	0.001
df	1	$3.93e^{-5}$	$1.57e^{-4}$	$9.82e^{-4}$	$3.93e^{-3}$	0.016	0.102	0.455	1.323	2.706	3.841	5.024	6.635	7.879	10.828
	2	0.010	0.020	0.051	0.103	0.211	0.575	1.386	2.773	4.605	5.991	7.378	9.210	10.597	13.816
	3	0.072	0.115	0.216	0.352	0.584	1.213	2.366	4.108	6.251	7.815	9.348	11.345	12.838	16.266
	4	0.207	0.297	0.484	0.711	1.064	1.923	3.357	5.385	7.779	9.488	11.143	13.277	14.860	18.467
	5	0.412	0.554	0.831	1.145	1.610	2.675	4.351	6.626	9.236	11.07	12.833	15.086	16.750	20.515
	6	0.676	0.872	1.237	1.635	2.204	3.455	5.348	7.841	10.645	12.592	14.449	16.812	18.548	22.458
	7	0.989	1.239	1.690	2.167	2.833	4.255	6.346	9.037	12.017	14.067	16.013	18.475	20.278	24.322
	8	1.344	1.646	2.180	2.733	3.490	5.071	7.344	10.219	13.362	15.507	17.535	20.09	21.955	26.124
	9	1.735	2.088	2.700	3.325	4.168	5.899	8.343	11.389	14.684	16.919	19.023	21.666	23.589	27.877
	10	2.156	2.558	3.247	3.940	4.865	6.737	9.342	12.549	15.987	18.307	20.483	23.209	25.188	29.588
	11	2.603	3.053	3.816	4.575	5.578	7.584	10.341	13.701	17.275	19.675	21.920	24.725	26.757	31.264
	12	3.074	3.571	4.404	5.226	6.304	8.438	11.34	14.845	18.549	21.026	23.337	26.217	28.300	32.909
	13	3.565	4.107	5.009	5.892	7.042	9.299	12.34	15.984	19.812	22.362	24.736	27.688	29.819	34.528
	14	4.075	4.660	5.629	6.571	7.790	10.165	13.339	17.117	21.064	23.685	26.119	29.141	31.319	36.123
	15	4.601	5.229	6.262	7.261	8.547	11.037	14.339	18.245	22.307	24.996	27.488	30.578	32.801	37.697
	16	5.142	5.812	6.908	7.962	9.312	11.912	15.338	19.369	23.542	26.296	28.845	32.000	34.267	39.252
	17	5.697	6.408	7.564	8.672	10.085	12.792	16.338	20.489	24.769	27.587	30.191	33.409	35.718	40.790
	18	6.265	7.015	8.231	9.390	10.865	13.675	17.338	21.605	25.989	28.869	31.526	34.805	37.156	42.312
	19	6.844	7.633	8.907	10.117	11.651	14.562	18.338	22.718	27.204	30.144	32.852	36.191	38.582	43.820
	20	7.434	8.260	9.591	10.851	12.443	15.452	19.337	23.828	28.412	31.410	34.170	37.566	39.997	45.315
	21	8.034	8.897	10.283	11.591	13.240	16.344	20.337	24.935	29.615	32.671	35.479	38.932	41.401	46.797
	22	8.643	9.542	10.982	12.338	14.041	17.24	21.337	26.039	30.813	33.924	36.781	40.289	42.796	48.268
	23	9.260	10.196	11.689	13.091	14.848	18.137	22.337	27.141	32.007	35.172	38.076	41.638	44.181	49.728
	24	9.886	10.856	12.401	13.848	15.659	19.037	23.337	28.241	33.196	36.415	39.364	42.980	45.559	51.179
	25	10.520	11.524	13.120	14.611	16.473	19.939	24.337	29.339	34.382	37.652	40.646	44.314	46.928	52.620
	26	11.160	12.198	13.844	15.379	17.292	20.843	25.336	30.435	35.563	38.885	41.923	45.642	48.29	54.052
	27	11.808	12.879	14.573	16.151	18.114	21.749	26.336	31.528	36.741	40.113	43.195	46.963	49.645	55.476
	28	12.461	13.565	15.308	16.928	18.939	22.657	27.336	32.620	37.916	41.337	44.461	48.278	50.993	56.892
	29	13.121	14.256	16.047	17.708	19.768	23.567	28.336	33.711	39.087	42.557	45.722	49.588	52.336	58.301
	30	13.787	14.953	16.791	18.493	20.599	24.478	29.336	34.800	40.256	43.773	46.979	50.892	53.672	59.703
	31	14.458	15.655	17.539	19.281	21.434	25.390	30.336	35.887	41.422	44.985	48.232	52.191	55.003	61.098
	32	15.134	16.362	18.291	20.072	22.271	26.304	31.336	36.973	42.585	46.194	49.480	53.486	56.328	62.487
	33	15.815	17.074	19.047	20.867	23.110	27.219	32.336	38.058	43.745	47.400	50.725	54.776	57.648	63.870
	34	16.501	17.789	19.806	21.664	23.952	28.136	33.336	39.141	44.903	48.602	51.966	56.061	58.964	65.247
	35	17.192	18.509	20.569	22.465	24.797	29.054	34.336	40.223	46.059	49.802	53.203	57.342	60.275	66.619
	36	17.887	19.233	21.336	23.269	25.643	29.973	35.336	41.304	47.212	50.998	54.437	58.619	61.581	67.985
	37	18.586	19.960	22.106	24.075	26.492	30.893	36.336	42.383	48.363	52.192	55.668	59.893	62.883	69.346
	38	19.289	20.691	22.878	24.884	27.343	31.815	37.335	43.462	49.513	53.384	56.896	61.162	64.181	70.703
	39	19.996	21.426	23.654	25.695	28.196	32.737	38.335	44.539	50.660	54.572	58.120	62.428	65.476	72.055
	40	20.707	22.164	24.433	26.509	29.051	33.660	39.335	45.616	51.805	55.758	59.342	63.691	66.766	73.402
	41	21.421	22.906	25.215	27.326	29.907	34.585	40.335	46.692	52.949	56.942	60.561	64.950	68.053	74.745
	42	22.138	23.650	25.999	28.144	30.765	35.510	41.335	47.766	54.090	58.124	61.777	66.206	69.336	76.084
	43	22.859	24.398	26.785	28.965	31.625	36.436	42.335	48.840	55.230	59.304	62.990	67.459	70.616	77.419
	44	23.584	25.148	27.575	29.787	32.487	37.363	43.335	49.913	56.369	60.481	64.201	68.710	71.893	78.750
	45	24.311	25.901	28.366	30.612	33.350	38.291	44.335	50.985	57.505	61.656	65.410	69.957	73.166	80.077
	46	25.041	26.657	29.160	31.439	34.215	39.220	45.335	52.056	58.641	62.830	66.617	71.201	74.437	81.400
	47	25.775	27.416	29.956	32.268	35.081	40.149	46.335	53.127	59.774	64.001	67.821	72.443	75.704	82.720
	48	26.511	28.177	30.755	33.098	35.949	41.079	47.335	54.196	60.907	65.171	69.023	73.683	76.969	84.037
	49	27.249	28.941	31.555	33.930	36.818	42.010	48.335	55.265	62.038	66.339	70.222	74.919	78.231	85.351
	50	27.991	29.707	32.357	34.764	37.689	42.942	49.335	56.334	63.167	67.505	71.420	76.154	79.490	86.661

	p	0.005	0.010	0.025	0.050	0.100	0.250	0.500	0.750	0.900	0.950	0.975	0.990	0.995	0.999
	$1-p$	0.995	0.990	0.975	0.950	0.900	0.750	0.500	0.250	0.100	0.050	0.025	0.010	0.005	0.001
df	51	28.735	30.475	33.162	35.600	38.56	43.874	50.335	57.401	64.295	68.669	72.616	77.386	80.747	87.968
	52	29.481	31.246	33.968	36.437	39.433	44.808	51.335	58.468	65.422	69.832	73.810	78.616	82.001	89.272
	53	30.230	32.018	34.776	37.276	40.308	45.741	52.335	59.534	66.548	70.993	75.002	79.843	83.253	90.573
	54	30.981	32.793	35.586	38.116	41.183	46.676	53.335	60.600	67.673	72.153	76.192	81.069	84.502	91.872
	55	31.735	33.570	36.398	38.958	42.060	47.610	54.335	61.665	68.796	73.311	77.380	82.292	85.749	93.168
	56	32.490	34.350	37.212	39.801	42.937	48.546	55.335	62.729	69.919	74.468	78.567	83.513	86.994	94.461
	57	33.248	35.131	38.027	40.646	43.816	49.482	56.335	63.793	71.040	75.624	79.752	84.733	88.236	95.751
	58	34.008	35.913	38.844	41.492	44.696	50.419	57.335	64.857	72.160	76.778	80.936	85.950	89.477	97.039
	59	34.770	36.698	39.662	42.339	45.577	51.356	58.335	65.919	73.279	77.931	82.117	87.166	90.715	98.324
	60	35.534	37.485	40.482	43.188	46.459	52.294	59.335	66.981	74.397	79.082	83.298	88.379	91.952	99.607
	61	36.301	38.273	41.303	44.038	47.342	53.232	60.335	68.043	75.514	80.232	84.476	89.591	93.186	100.888
	62	37.068	39.063	42.126	44.889	48.226	54.171	61.335	69.104	76.630	81.381	85.654	90.802	94.419	102.166
	63	37.838	39.855	42.950	45.741	49.111	55.110	62.335	70.165	77.745	82.529	86.830	92.010	95.649	103.442
	64	38.610	40.649	43.776	46.595	49.996	56.050	63.335	71.225	78.860	83.675	88.004	93.217	96.878	104.716
	65	39.383	41.444	44.603	47.450	50.883	56.990	64.335	72.285	79.973	84.821	89.177	94.422	98.105	105.988
	66	40.158	42.240	45.431	48.305	51.770	57.931	65.335	73.344	81.085	85.965	90.349	95.626	99.330	107.258
	67	40.935	43.038	46.261	49.162	52.659	58.872	66.335	74.403	82.197	87.108	91.519	96.828	100.554	108.526
	68	41.713	43.838	47.092	50.020	53.548	59.814	67.335	75.461	83.308	88.250	92.689	98.028	101.776	109.791
	69	42.494	44.639	47.924	50.879	54.438	60.756	68.334	76.519	84.418	89.391	93.856	99.228	102.996	111.055
	70	43.275	45.442	48.758	51.739	55.329	61.698	69.334	77.577	85.527	90.531	95.023	100.425	104.215	112.317
	71	44.058	46.246	49.592	52.600	56.221	62.641	70.334	78.634	86.635	91.670	96.189	101.621	105.432	113.577
	72	44.843	47.051	50.428	53.462	57.113	63.585	71.334	79.690	87.743	92.808	97.353	102.816	106.648	114.835
	73	45.629	47.858	51.265	54.325	58.006	64.528	72.334	80.747	88.850	93.945	98.516	104.010	107.862	116.092
	74	46.417	48.666	52.103	55.189	58.900	65.472	73.334	81.803	89.956	95.081	99.678	105.202	109.074	117.346
	75	47.206	49.475	52.942	56.054	59.795	66.417	74.334	82.858	91.061	96.217	100.839	106.393	110.286	118.599
	76	47.997	50.286	53.782	56.920	60.690	67.362	75.334	83.913	92.166	97.351	101.999	107.583	111.495	119.850
	77	48.788	51.097	54.623	57.786	61.586	68.307	76.334	84.968	93.270	98.484	103.158	108.771	112.704	121.100
	78	49.582	51.910	55.466	58.654	62.483	69.252	77.334	86.022	94.374	99.617	104.316	109.958	113.911	122.348
	79	50.376	52.725	56.309	59.522	63.380	70.198	78.334	87.077	95.476	100.749	105.473	111.144	115.117	123.594
	80	51.172	53.540	57.153	60.391	64.278	71.145	79.334	88.130	96.578	101.879	106.629	112.329	116.321	124.839
	81	51.969	54.357	57.998	61.261	65.176	72.091	80.334	89.184	97.680	103.010	107.783	113.512	117.524	126.083
	82	52.767	55.174	58.845	62.132	66.076	73.038	81.334	90.237	98.780	104.139	108.937	114.695	118.726	127.324
	83	53.567	55.993	59.692	63.004	66.976	73.985	82.334	91.289	99.880	105.267	110.090	115.876	119.927	128.565
	84	54.368	56.813	60.540	63.876	67.876	74.933	83.334	92.342	100.980	106.395	111.242	117.057	121.126	129.804
	85	55.170	57.634	61.389	64.749	68.777	75.881	84.334	93.394	102.079	107.522	112.393	118.236	122.325	131.041
	86	55.973	58.456	62.239	65.623	69.679	76.829	85.334	94.446	103.177	108.648	113.544	119.414	123.522	132.277
	87	56.777	59.279	63.089	66.498	70.581	77.777	86.334	95.497	104.275	109.773	114.693	120.591	124.718	133.512
	88	57.582	60.103	63.941	67.373	71.484	78.726	87.334	96.548	105.372	110.898	115.841	121.767	125.913	134.745
	89	58.389	60.928	64.793	68.249	72.387	79.675	88.334	97.599	106.469	112.022	116.989	122.942	127.106	135.978
	90	59.196	61.754	65.647	69.126	73.291	80.625	89.334	98.650	107.565	113.145	118.136	124.116	128.299	137.208
	91	60.005	62.581	66.501	70.003	74.196	81.574	90.334	99.700	108.661	114.268	119.282	125.289	129.491	138.438
	92	60.815	63.409	67.356	70.882	75.100	82.524	91.334	100.750	109.756	115.390	120.427	126.462	130.681	139.666
	93	61.625	64.238	68.211	71.760	76.006	83.474	92.334	101.800	110.850	116.511	121.571	127.633	131.871	140.893
	94	62.437	65.068	69.068	72.640	76.912	84.425	93.334	102.850	111.944	117.632	122.715	128.803	133.059	142.119
	95	63.250	65.898	69.925	73.520	77.818	85.376	94.334	103.899	113.038	118.752	123.858	129.973	134.247	143.344
	96	64.063	66.730	70.783	74.401	78.725	86.327	95.334	104.948	114.131	119.871	125.000	131.141	135.433	144.567
	97	64.878	67.562	71.642	75.282	79.633	87.278	96.334	105.997	115.223	120.990	126.141	132.309	136.619	145.789
	98	65.694	68.396	72.501	76.164	80.541	88.229	97.334	107.045	116.315	122.108	127.282	133.476	137.803	147.010
	99	66.510	69.230	73.361	77.046	81.449	89.181	98.334	108.093	117.407	123.225	128.422	134.642	138.987	148.230
	100	67.328	70.065	74.222	77.929	82.358	90.133	99.334	109.141	118.498	124.342	129.561	135.807	140.169	149.449

23.4 Table F: Quantiles of the F distribution

Suppose a random variable Y has an F distribution, with df_1 and df_2 the numerator and denominator degrees of freedom. This table gives values of the quantile q such that $P[Y < q] = p$, where $p = 0.005, ..., 0.999$. The values in the table were generated using the SAS function `finv` (SAS Institute Inc. 2016).

Figure 23.4: Plot of the F distribution ($df_1 = 4$, $df_2 = 20$) illustrating p and $1 - p$ in the table below.



References

SAS Institute Inc. (2016) *SAS 9.4 Functions and CALL Routines: Reference, Fifth Edition.* SAS Institute Inc., Cary, NC.

		0.900	0.950	0.975	0.990	0.995	0.999	p
		0.100	0.050	0.025	0.010	0.005	0.001	$1-p$
df_1	df_2							
1	4	4.545	7.709	12.218	21.198	31.333	74.137	
2	4	4.325	6.944	10.649	18.000	26.284	61.246	
3	4	4.191	6.591	9.979	16.694	24.259	56.177	
4	4	4.107	6.388	9.605	15.977	23.155	53.436	
5	4	4.051	6.256	9.364	15.522	22.456	51.712	
6	4	4.010	6.163	9.197	15.207	21.975	50.525	
1	5	4.060	6.608	10.007	16.258	22.785	47.181	
2	5	3.780	5.786	8.434	13.274	18.314	37.122	
3	5	3.619	5.409	7.764	12.060	16.530	33.202	
4	5	3.520	5.192	7.388	11.392	15.556	31.085	
5	5	3.453	5.050	7.146	10.967	14.940	29.752	
6	5	3.405	4.950	6.978	10.672	14.513	28.834	
1	6	3.776	5.987	8.813	13.745	18.635	35.507	
2	6	3.463	5.143	7.260	10.925	14.544	27.000	
3	6	3.289	4.757	6.599	9.780	12.917	23.703	
4	6	3.181	4.534	6.227	9.148	12.028	21.924	
5	6	3.108	4.387	5.988	8.746	11.464	20.803	
6	6	3.055	4.284	5.820	8.466	11.073	20.030	
1	7	3.589	5.591	8.073	12.246	16.236	29.245	
2	7	3.257	4.737	6.542	9.547	12.404	21.689	
3	7	3.074	4.347	5.890	8.451	10.882	18.772	
4	7	2.961	4.120	5.523	7.847	10.050	17.198	
5	7	2.883	3.972	5.285	7.460	9.522	16.206	
6	7	2.827	3.866	5.119	7.191	9.155	15.521	
1	8	3.458	5.318	7.571	11.259	14.688	25.415	
2	8	3.113	4.459	6.059	8.649	11.042	18.494	
3	8	2.924	4.066	5.416	7.591	9.596	15.829	
4	8	2.806	3.838	5.053	7.006	8.805	14.392	
5	8	2.726	3.687	4.817	6.632	8.302	13.485	
6	8	2.668	3.581	4.652	6.371	7.952	12.858	
1	9	3.360	5.117	7.209	10.561	13.614	22.857	
2	9	3.006	4.256	5.715	8.022	10.107	16.387	
3	9	2.813	3.863	5.078	6.992	8.717	13.902	
4	9	2.693	3.633	4.718	6.422	7.956	12.560	
5	9	2.611	3.482	4.484	6.057	7.471	11.714	
6	9	2.551	3.374	4.320	5.802	7.134	11.128	
1	10	3.285	4.965	6.937	10.044	12.826	21.040	
2	10	2.924	4.103	5.456	7.559	9.427	14.905	
3	10	2.728	3.708	4.826	6.552	8.081	12.553	
4	10	2.605	3.478	4.468	5.994	7.343	11.283	
5	10	2.522	3.326	4.236	5.636	6.872	10.481	
6	10	2.461	3.217	4.072	5.386	6.545	9.926	

		0.900	0.950	0.975	0.990	0.995	0.999	p
		0.100	0.050	0.025	0.010	0.005	0.001	$1-p$
df_1	df_2							
1	11	3.225	4.844	6.724	9.646	12.226	19.687	
2	11	2.860	3.982	5.256	7.206	8.912	13.812	
3	11	2.660	3.587	4.630	6.217	7.600	11.561	
4	11	2.536	3.357	4.275	5.668	6.881	10.346	
5	11	2.451	3.204	4.044	5.316	6.422	9.578	
6	11	2.389	3.095	3.881	5.069	6.102	9.047	
1	12	3.177	4.747	6.554	9.330	11.754	18.643	
2	12	2.807	3.885	5.096	6.927	8.510	12.974	
3	12	2.606	3.490	4.474	5.953	7.226	10.804	
4	12	2.480	3.259	4.121	5.412	6.521	9.633	
5	12	2.394	3.106	3.891	5.064	6.071	8.892	
6	12	2.331	2.996	3.728	4.821	5.757	8.379	
1	13	3.136	4.667	6.414	9.074	11.374	17.815	
2	13	2.763	3.806	4.965	6.701	8.186	12.313	
3	13	2.560	3.411	4.347	5.739	6.926	10.209	
4	13	2.434	3.179	3.996	5.205	6.233	9.073	
5	13	2.347	3.025	3.767	4.862	5.791	8.354	
6	13	2.283	2.915	3.604	4.620	5.482	7.856	
1	14	3.102	4.600	6.298	8.862	11.060	17.143	
2	14	2.726	3.739	4.857	6.515	7.922	11.779	
3	14	2.522	3.344	4.242	5.564	6.680	9.729	
4	14	2.395	3.112	3.892	5.035	5.998	8.622	
5	14	2.307	2.958	3.663	4.695	5.562	7.922	
6	14	2.243	2.848	3.501	4.456	5.257	7.436	
1	15	3.073	4.543	6.200	8.683	10.798	16.587	
2	15	2.695	3.682	4.765	6.359	7.701	11.339	
3	15	2.490	3.287	4.153	5.417	6.476	9.335	
4	15	2.361	3.056	3.804	4.893	5.803	8.253	
5	15	2.273	2.901	3.576	4.556	5.372	7.567	
6	15	2.208	2.790	3.415	4.318	5.071	7.092	
1	16	3.048	4.494	6.115	8.531	10.575	16.120	
2	16	2.668	3.634	4.687	6.226	7.514	10.971	
3	16	2.462	3.239	4.077	5.292	6.303	9.006	
4	16	2.333	3.007	3.729	4.773	5.638	7.944	
5	16	2.244	2.852	3.502	4.437	5.212	7.272	
6	16	2.178	2.741	3.341	4.202	4.913	6.805	
1	17	3.026	4.451	6.042	8.400	10.384	15.722	
2	17	2.645	3.592	4.619	6.112	7.354	10.658	
3	17	2.437	3.197	4.011	5.185	6.156	8.727	
4	17	2.308	2.965	3.665	4.669	5.497	7.683	
5	17	2.218	2.810	3.438	4.336	5.075	7.022	
6	17	2.152	2.699	3.277	4.102	4.779	6.562	

		0.900	0.950	0.975	0.990	0.995	0.999	p
		0.100	0.050	0.025	0.010	0.005	0.001	$1-p$
df_1	df_2							
1	18	3.007	4.414	5.978	8.285	10.218	15.379	
2	18	2.624	3.555	4.560	6.013	7.215	10.390	
3	18	2.416	3.160	3.954	5.092	6.028	8.487	
4	18	2.286	2.928	3.608	4.579	5.375	7.459	
5	18	2.196	2.773	3.382	4.248	4.956	6.808	
6	18	2.130	2.661	3.221	4.015	4.663	6.355	
1	19	2.990	4.381	5.922	8.185	10.073	15.081	
2	19	2.606	3.522	4.508	5.926	7.093	10.157	
3	19	2.397	3.127	3.903	5.010	5.916	8.280	
4	19	2.266	2.895	3.559	4.500	5.268	7.265	
5	19	2.176	2.740	3.333	4.171	4.853	6.622	
6	19	2.109	2.628	3.172	3.939	4.561	6.175	
1	20	2.975	4.351	5.871	8.096	9.944	14.819	
2	20	2.589	3.493	4.461	5.849	6.986	9.953	
3	20	2.380	3.098	3.859	4.938	5.818	8.098	
4	20	2.249	2.866	3.515	4.431	5.174	7.096	
5	20	2.158	2.711	3.289	4.103	4.762	6.461	
6	20	2.091	2.599	3.128	3.871	4.472	6.019	
1	21	2.961	4.325	5.827	8.017	9.830	14.587	
2	21	2.575	3.467	4.420	5.780	6.891	9.772	
3	21	2.365	3.072	3.819	4.874	5.730	7.938	
4	21	2.233	2.840	3.475	4.369	5.091	6.947	
5	21	2.142	2.685	3.250	4.042	4.681	6.318	
6	21	2.075	2.573	3.090	3.812	4.393	5.881	
1	22	2.949	4.301	5.786	7.945	9.727	14.380	
2	22	2.561	3.443	4.383	5.719	6.806	9.612	
3	22	2.351	3.049	3.783	4.817	5.652	7.796	
4	22	2.219	2.817	3.440	4.313	5.017	6.814	
5	22	2.128	2.661	3.215	3.988	4.609	6.191	
6	22	2.060	2.549	3.055	3.758	4.322	5.758	
1	23	2.937	4.279	5.750	7.881	9.635	14.195	
2	23	2.549	3.422	4.349	5.664	6.730	9.469	
3	23	2.339	3.028	3.750	4.765	5.582	7.669	
4	23	2.207	2.796	3.408	4.264	4.950	6.696	
5	23	2.115	2.640	3.183	3.939	4.544	6.078	
6	23	2.047	2.528	3.023	3.710	4.259	5.649	
1	24	2.927	4.260	5.717	7.823	9.551	14.028	
2	24	2.538	3.403	4.319	5.614	6.661	9.339	
3	24	2.327	3.009	3.721	4.718	5.519	7.554	
4	24	2.195	2.776	3.379	4.218	4.890	6.589	
5	24	2.103	2.621	3.155	3.895	4.486	5.977	
6	24	2.035	2.508	2.995	3.667	4.202	5.550	

		0.900	0.950	0.975	0.990	0.995	0.999	p
		0.100	0.050	0.025	0.010	0.005	0.001	$1-p$
df_1	df_2							
1	25	2.918	4.242	5.686	7.770	9.475	13.877	
2	25	2.528	3.385	4.291	5.568	6.598	9.223	
3	25	2.317	2.991	3.694	4.675	5.462	7.451	
4	25	2.184	2.759	3.353	4.177	4.835	6.493	
5	25	2.092	2.603	3.129	3.855	4.433	5.885	
6	25	2.024	2.490	2.969	3.627	4.150	5.462	
1	26	2.909	4.225	5.659	7.721	9.406	13.739	
2	26	2.519	3.369	4.265	5.526	6.541	9.116	
3	26	2.307	2.975	3.670	4.637	5.409	7.357	
4	26	2.174	2.743	3.329	4.140	4.785	6.406	
5	26	2.082	2.587	3.105	3.818	4.384	5.802	
6	26	2.014	2.474	2.945	3.591	4.103	5.381	
1	27	2.901	4.210	5.633	7.677	9.342	13.613	
2	27	2.511	3.354	4.242	5.488	6.489	9.019	
3	27	2.299	2.960	3.647	4.601	5.361	7.272	
4	27	2.165	2.728	3.307	4.106	4.740	6.326	
5	27	2.073	2.572	3.083	3.785	4.340	5.726	
6	27	2.005	2.459	2.923	3.558	4.059	5.308	
1	28	2.894	4.196	5.610	7.636	9.284	13.498	
2	28	2.503	3.340	4.221	5.453	6.440	8.931	
3	28	2.291	2.947	3.626	4.568	5.317	7.193	
4	28	2.157	2.714	3.286	4.074	4.698	6.253	
5	28	2.064	2.558	3.063	3.754	4.300	5.656	
6	28	1.996	2.445	2.903	3.528	4.020	5.241	
1	29	2.887	4.183	5.588	7.598	9.230	13.391	
2	29	2.495	3.328	4.201	5.420	6.396	8.849	
3	29	2.283	2.934	3.607	4.538	5.276	7.121	
4	29	2.149	2.701	3.267	4.045	4.659	6.186	
5	29	2.057	2.545	3.044	3.725	4.262	5.593	
6	29	1.988	2.432	2.884	3.499	3.983	5.179	
1	30	2.881	4.171	5.568	7.562	9.180	13.293	
2	30	2.489	3.316	4.182	5.390	6.355	8.773	
3	30	2.276	2.922	3.589	4.510	5.239	7.054	
4	30	2.142	2.690	3.250	4.018	4.623	6.125	
5	30	2.049	2.534	3.026	3.699	4.228	5.534	
6	30	1.980	2.421	2.867	3.473	3.949	5.122	
1	31	2.875	4.160	5.549	7.530	9.133	13.202	
2	31	2.482	3.305	4.165	5.362	6.317	8.704	
3	31	2.270	2.911	3.573	4.484	5.204	6.993	
4	31	2.136	2.679	3.234	3.993	4.590	6.067	
5	31	2.042	2.523	3.010	3.675	4.196	5.480	
6	31	1.973	2.409	2.851	3.449	3.918	5.070	

		0.900	0.950	0.975	0.990	0.995	0.999	p
		0.100	0.050	0.025	0.010	0.005	0.001	$1-p$
df_1	df_2							
1	32	2.869	4.149	5.531	7.499	9.090	13.117	
2	32	2.477	3.295	4.149	5.336	6.281	8.639	
3	32	2.263	2.901	3.557	4.459	5.171	6.936	
4	32	2.129	2.668	3.218	3.969	4.559	6.014	
5	32	2.036	2.512	2.995	3.652	4.166	5.429	
6	32	1.967	2.399	2.836	3.427	3.889	5.021	
1	33	2.864	4.139	5.515	7.471	9.050	13.039	
2	33	2.471	3.285	4.134	5.312	6.248	8.579	
3	33	2.258	2.892	3.543	4.437	5.141	6.883	
4	33	2.123	2.659	3.204	3.948	4.531	5.965	
5	33	2.030	2.503	2.981	3.630	4.138	5.382	
6	33	1.961	2.389	2.822	3.406	3.861	4.976	
1	34	2.859	4.130	5.499	7.444	9.012	12.965	
2	34	2.466	3.276	4.120	5.289	6.217	8.522	
3	34	2.252	2.883	3.529	4.416	5.113	6.833	
4	34	2.118	2.650	3.191	3.927	4.504	5.919	
5	34	2.024	2.494	2.968	3.611	4.112	5.339	
6	34	1.955	2.380	2.808	3.386	3.836	4.934	
1	35	2.855	4.121	5.485	7.419	8.976	12.896	
2	35	2.461	3.267	4.106	5.268	6.188	8.470	
3	35	2.247	2.874	3.517	4.396	5.086	6.787	
4	35	2.113	2.641	3.179	3.908	4.479	5.876	
5	35	2.019	2.485	2.956	3.592	4.088	5.298	
6	35	1.950	2.372	2.796	3.368	3.812	4.894	
1	36	2.850	4.113	5.471	7.396	8.943	12.832	
2	36	2.456	3.259	4.094	5.248	6.161	8.420	
3	36	2.243	2.866	3.505	4.377	5.062	6.744	
4	36	2.108	2.634	3.167	3.890	4.455	5.836	
5	36	2.014	2.477	2.944	3.574	4.065	5.260	
6	36	1.945	2.364	2.785	3.351	3.790	4.857	
1	37	2.846	4.105	5.458	7.373	8.912	12.771	
2	37	2.452	3.252	4.082	5.229	6.135	8.374	
3	37	2.238	2.859	3.493	4.360	5.038	6.703	
4	37	2.103	2.626	3.156	3.873	4.433	5.799	
5	37	2.009	2.470	2.933	3.558	4.043	5.224	
6	37	1.940	2.356	2.774	3.334	3.769	4.823	
1	38	2.842	4.098	5.446	7.353	8.882	12.714	
2	38	2.448	3.245	4.071	5.211	6.111	8.331	
3	38	2.234	2.852	3.483	4.343	5.016	6.665	
4	38	2.099	2.619	3.145	3.858	4.412	5.763	
5	38	2.005	2.463	2.923	3.542	4.023	5.190	
6	38	1.935	2.349	2.763	3.319	3.749	4.790	

		0.900	0.950	0.975	0.990	0.995	0.999	p
		0.100	0.050	0.025	0.010	0.005	0.001	$1-p$
df_1	df_2							
1	39	2.839	4.091	5.435	7.333	8.854	12.660	
2	39	2.444	3.238	4.061	5.194	6.088	8.290	
3	39	2.230	2.845	3.473	4.327	4.995	6.629	
4	39	2.095	2.612	3.135	3.843	4.392	5.730	
5	39	2.001	2.456	2.913	3.528	4.004	5.158	
6	39	1.931	2.342	2.754	3.305	3.731	4.759	
1	40	2.835	4.085	5.424	7.314	8.828	12.609	
2	40	2.440	3.232	4.051	5.179	6.066	8.251	
3	40	2.226	2.839	3.463	4.313	4.976	6.595	
4	40	2.091	2.606	3.126	3.828	4.374	5.698	
5	40	1.997	2.449	2.904	3.514	3.986	5.128	
6	40	1.927	2.336	2.744	3.291	3.713	4.731	
1	41	2.832	4.079	5.414	7.296	8.803	12.561	
2	41	2.437	3.226	4.042	5.163	6.046	8.214	
3	41	2.222	2.833	3.454	4.299	4.957	6.562	
4	41	2.087	2.600	3.117	3.815	4.356	5.668	
5	41	1.993	2.443	2.895	3.501	3.969	5.100	
6	41	1.923	2.330	2.736	3.278	3.696	4.703	
1	42	2.829	4.073	5.404	7.280	8.779	12.516	
2	42	2.434	3.220	4.033	5.149	6.027	8.179	
3	42	2.219	2.827	3.446	4.285	4.940	6.532	
4	42	2.084	2.594	3.109	3.802	4.339	5.640	
5	42	1.989	2.438	2.887	3.488	3.953	5.073	
6	42	1.919	2.324	2.727	3.266	3.680	4.677	
1	43	2.826	4.067	5.395	7.264	8.757	12.472	
2	43	2.430	3.214	4.024	5.136	6.008	8.146	
3	43	2.216	2.822	3.438	4.273	4.923	6.503	
4	43	2.080	2.589	3.101	3.790	4.324	5.613	
5	43	1.986	2.432	2.879	3.476	3.937	5.048	
6	43	1.916	2.318	2.719	3.254	3.665	4.653	
1	44	2.823	4.062	5.386	7.248	8.735	12.431	
2	44	2.427	3.209	4.016	5.123	5.991	8.115	
3	44	2.213	2.816	3.430	4.261	4.907	6.476	
4	44	2.077	2.584	3.093	3.778	4.308	5.588	
5	44	1.983	2.427	2.871	3.465	3.923	5.024	
6	44	1.913	2.313	2.712	3.243	3.651	4.630	
1	45	2.820	4.057	5.377	7.234	8.715	12.392	
2	45	2.425	3.204	4.009	5.110	5.974	8.086	
3	45	2.210	2.812	3.422	4.249	4.892	6.450	
4	45	2.074	2.579	3.086	3.767	4.294	5.564	
5	45	1.980	2.422	2.864	3.454	3.909	5.001	
6	45	1.909	2.308	2.705	3.232	3.638	4.608	

		0.900	0.950	0.975	0.990	0.995	0.999	p
		0.100	0.050	0.025	0.010	0.005	0.001	$1-p$
df_1	df_2							
1	46	2.818	4.052	5.369	7.220	8.695	12.355	
2	46	2.422	3.200	4.001	5.099	5.958	8.057	
3	46	2.207	2.807	3.415	4.238	4.877	6.425	
4	46	2.071	2.574	3.079	3.757	4.280	5.541	
5	46	1.977	2.417	2.857	3.444	3.896	4.979	
6	46	1.906	2.304	2.698	3.222	3.625	4.587	
1	47	2.815	4.047	5.361	7.207	8.677	12.319	
2	47	2.419	3.195	3.994	5.087	5.943	8.030	
3	47	2.204	2.802	3.409	4.228	4.864	6.401	
4	47	2.068	2.570	3.073	3.747	4.267	5.519	
5	47	1.974	2.413	2.851	3.434	3.883	4.958	
6	47	1.903	2.299	2.691	3.213	3.612	4.566	
1	48	2.813	4.043	5.354	7.194	8.659	12.286	
2	48	2.417	3.191	3.987	5.077	5.929	8.005	
3	48	2.202	2.798	3.402	4.218	4.850	6.379	
4	48	2.066	2.565	3.066	3.737	4.255	5.498	
5	48	1.971	2.409	2.844	3.425	3.871	4.938	
6	48	1.901	2.295	2.685	3.204	3.601	4.547	
1	49	2.811	4.038	5.347	7.182	8.642	12.253	
2	49	2.414	3.187	3.981	5.066	5.915	7.980	
3	49	2.199	2.794	3.396	4.208	4.838	6.357	
4	49	2.063	2.561	3.060	3.728	4.243	5.478	
5	49	1.968	2.404	2.838	3.416	3.860	4.919	
6	49	1.898	2.290	2.679	3.195	3.589	4.529	
1	50	2.809	4.034	5.340	7.171	8.626	12.222	
2	50	2.412	3.183	3.975	5.057	5.902	7.956	
3	50	2.197	2.790	3.390	4.199	4.826	6.336	
4	50	2.061	2.557	3.054	3.720	4.232	5.459	
5	50	1.966	2.400	2.833	3.408	3.849	4.901	
6	50	1.895	2.286	2.674	3.186	3.579	4.512	
1	51	2.807	4.030	5.334	7.159	8.610	12.192	
2	51	2.410	3.179	3.969	5.047	5.889	7.934	
3	51	2.194	2.786	3.385	4.191	4.814	6.317	
4	51	2.058	2.553	3.049	3.711	4.221	5.441	
5	51	1.964	2.397	2.827	3.400	3.838	4.884	
6	51	1.893	2.283	2.668	3.178	3.568	4.495	
1	52	2.805	4.027	5.328	7.149	8.595	12.164	
2	52	2.408	3.175	3.963	5.038	5.877	7.912	
3	52	2.192	2.783	3.379	4.182	4.803	6.298	
4	52	2.056	2.550	3.044	3.703	4.210	5.424	
5	52	1.961	2.393	2.822	3.392	3.828	4.867	
6	52	1.891	2.279	2.663	3.171	3.558	4.479	

		0.900	0.950	0.975	0.990	0.995	0.999	p
		0.100	0.050	0.025	0.010	0.005	0.001	$1-p$
df_1	df_2							
1	53	2.803	4.023	5.322	7.139	8.581	12.137	
2	53	2.406	3.172	3.958	5.030	5.865	7.892	
3	53	2.190	2.779	3.374	4.174	4.793	6.280	
4	53	2.054	2.546	3.038	3.695	4.200	5.407	
5	53	1.959	2.389	2.817	3.384	3.818	4.852	
6	53	1.888	2.275	2.658	3.163	3.549	4.464	
1	54	2.801	4.020	5.316	7.129	8.567	12.111	
2	54	2.404	3.168	3.953	5.021	5.854	7.872	
3	54	2.188	2.776	3.369	4.167	4.783	6.262	
4	54	2.052	2.543	3.034	3.688	4.191	5.391	
5	54	1.957	2.386	2.812	3.377	3.809	4.836	
6	54	1.886	2.272	2.653	3.156	3.540	4.449	
1	55	2.799	4.016	5.310	7.119	8.554	12.085	
2	55	2.402	3.165	3.948	5.013	5.843	7.853	
3	55	2.186	2.773	3.364	4.159	4.773	6.246	
4	55	2.050	2.540	3.029	3.681	4.181	5.375	
5	55	1.955	2.383	2.807	3.370	3.800	4.822	
6	55	1.884	2.269	2.648	3.149	3.531	4.435	
1	56	2.797	4.013	5.305	7.110	8.541	12.061	
2	56	2.400	3.162	3.943	5.006	5.833	7.834	
3	56	2.184	2.769	3.359	4.152	4.763	6.230	
4	56	2.048	2.537	3.024	3.674	4.172	5.361	
5	56	1.953	2.380	2.803	3.363	3.791	4.808	
6	56	1.882	2.266	2.644	3.143	3.523	4.421	
1	57	2.796	4.010	5.300	7.102	8.529	12.038	
2	57	2.398	3.159	3.938	4.998	5.823	7.817	
3	57	2.182	2.766	3.355	4.145	4.754	6.214	
4	57	2.046	2.534	3.020	3.667	4.164	5.346	
5	57	1.951	2.377	2.798	3.357	3.783	4.794	
6	57	1.880	2.263	2.639	3.136	3.514	4.408	
1	58	2.794	4.007	5.295	7.093	8.517	12.015	
2	58	2.396	3.156	3.934	4.991	5.813	7.800	
3	58	2.181	2.764	3.351	4.138	4.746	6.199	
4	58	2.044	2.531	3.016	3.661	4.156	5.333	
5	58	1.949	2.374	2.794	3.351	3.775	4.781	
6	58	1.878	2.260	2.635	3.130	3.507	4.396	
1	59	2.793	4.004	5.290	7.085	8.506	11.994	
2	59	2.395	3.153	3.929	4.984	5.804	7.784	
3	59	2.179	2.761	3.347	4.132	4.737	6.185	
4	59	2.043	2.528	3.012	3.655	4.148	5.319	
5	59	1.947	2.371	2.790	3.345	3.767	4.769	
6	59	1.876	2.257	2.631	3.124	3.499	4.384	

		0.900	0.950	0.975	0.990	0.995	0.999	p
		0.100	0.050	0.025	0.010	0.005	0.001	$1-p$
df_1	df_2							
1	60	2.791	4.001	5.286	7.077	8.495	11.973	
2	60	2.393	3.150	3.925	4.977	5.795	7.768	
3	60	2.177	2.758	3.343	4.126	4.729	6.171	
4	60	2.041	2.525	3.008	3.649	4.140	5.307	
5	60	1.946	2.368	2.786	3.339	3.760	4.757	
6	60	1.875	2.254	2.627	3.119	3.492	4.372	

Chapter 24

Matrix Programs

24.1 Matrix calculations

Examples of various matrix operations using SAS `proc iml` (SAS Institute Inc. 2018).

```
* matrix2.sas;
title 'Matrix calculations';
proc iml;
reset print;
* Define matrix A and B;
A = {1, 2, 3};
B = {4, 5, 6};
* Add A and B;
AplusB = A + B;
* Define matrix C and D;
C = {1 4, 2 5, 3 6};
D = {7, 8};
* Multiply C and D;
CD = C*D;
* Transpose of F;
F = {1 5, 2 6, 3 7, 4 8};
transposeF = t(F);
* Define another matrix A;
A = {1 6 4, 3 7 6, 4 1 9};
* Inverse of A;
Ainv = inv(A);
* Check that Ainv*A = I;
AinvA = Ainv*A;
quit;
```

24.2 Multiple regression in matrix form

A multiple regression analysis using matrix operations and `proc iml` (SAS Institute Inc. 2018). The data are from Reeve et al. (1998).

```
* multreg.sas;
title 'Multiple regression in matrix form';
data multdata;
        input X1 X2 survival;
        * Apply transformations here;
        Y = log(survival);
        datalines;
1.250 0.000 0.107
2.656 0.481 0.715
7.334 0.171 0.036
1.603 0.352 0.188
2.622 0.016 0.438
1.000 0.000 0.585
4.342 0.185 0.115
5.233 0.018 0.257
2.500 0.410 0.032
3.250 0.015 0.350
6.000 0.007 0.161
4.750 0.000 0.073
2.500 0.095 0.219
8.750 0.033 0.028
6.000 0.015 0.294
5.000 0.105 0.207
7.149 0.025 0.227
6.750 0.015 0.040
7.500 0.043 0.089
2.500 0.073 0.176
5.000 0.055 0.084
2.250 0.023 0.203
1.250 0.123 0.074
4.750 0.035 0.126
4.500 0.212 0.290
9.557 0.166 0.010
5.000 0.338 0.207
;
run;
* Print the data;
proc print data=multdata;
run;
```

```
* Matrix calculations;
proc iml;
reset print;
* Read in data set;
use multdata var {Y X1 X2};
read all;
close multdata;
* Design matrix X;
n = nrow(Y); * Find sample size;
ones = shape(1,n,1);
X = ones||X1||X2;
* Y values;
print Y;
* X’ or X transpose;
Xt = t(X);
* X’X;
XtX = Xt*X;
* X’X inverse;
XtXinv = inv(XtX);
* Show this is the inverse;
test = XtXinv*XtX;
* (X’X inverse)X’;
XtXinvXt = XtXinv*Xt;
* beta = (X’X inverse)X’Y;
beta = XtXinvXt*Y;
* Yhat;
Yhat = X*beta;
* SSerror and MSerror;
SSerror = sum((Y-Yhat)##2);
dfnum = nrow(beta)-1;
dfdenom = n - dfnum - 1;
MSerror = SSerror/dfdenom;
* SSreg and MSreg;
Ymean = mean(Y);
SSreg = sum((Yhat-Ymean)##2);
MSreg = SSreg/dfnum;
* F statistic and P value for overall test;
F = MSreg/MSerror;
P = 1 - probf(F,dfnum,dfdenom);
* Standard errors for beta;
sebeta = sqrt(MSerror*vecdiag(XtXinv));
* t tests for beta_i = 0;
Tvec = beta/sebeta;
Pvec = 2*(1-probt(abs(Tvec),dfdenom));
quit;
```

24.3 References

Reeve, J. D., Rhodes, D. J. & Turchin, P. (1998) Scramble competition in southern pine beetle (Coleoptera: Scolytidae). *Ecological Entomology* 23: 433-443.

SAS Institute Inc. (2018) *SAS/IML 15.1 User's Guide*. SAS Institute Inc., Cary, NC.

Index