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.

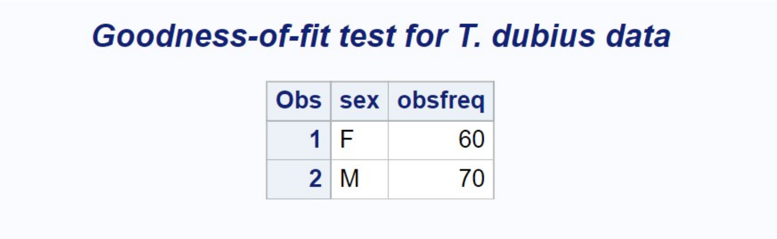
The first step in the analysis is to make a SAS data set using the observed frequencies in Table 20.1. The variable `obsfreq` contains this information for each value of `sex` (see SAS program below). The data could also have been entered as individual observations with a single data line for each observation, 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

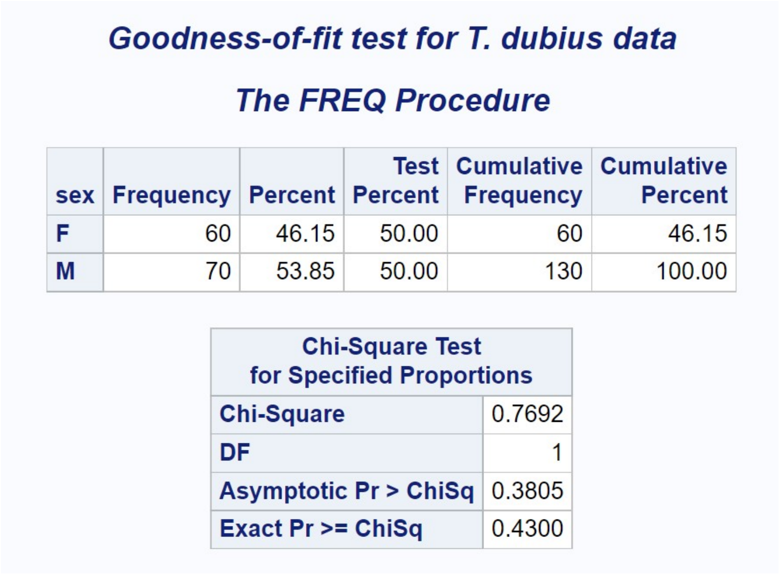
```
* gof_clerids.sas;
title 'Goodness-of-fit test for T. dubius data';
data elytra;
    input sex \$ obsfreq;
    datalines;
F  60
M  70
;
run;
* Print data set;
proc print data=elytra;
run;
* Goodness-of-fit test (Chi-square only);
proc freq data=elytra order=data;
    tables sex / testp=(0.5 0.5) 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

$$P[Y_1 = y_1, Y_2 = y_2, Y_3 = y_3, Y_4 = y_4] = \frac{n!}{y_1!y_2!y_3!y_4!} p_1^{y_1} p_2^{y_2} p_3^{y_3} p_4^{y_4}. \tag{20.25}$$

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

$$G^2 = 2\sum_{i=1}^{a} O_i \ln(O_i/E_i). \tag{20.26}$$

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

$$X^2 = \sum_{i=1}^{a} \frac{(O_i - E_i)^2}{E_i}. \tag{20.27}$$

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 ($X^2 = 0.470, df = 3, P < 0.95$).

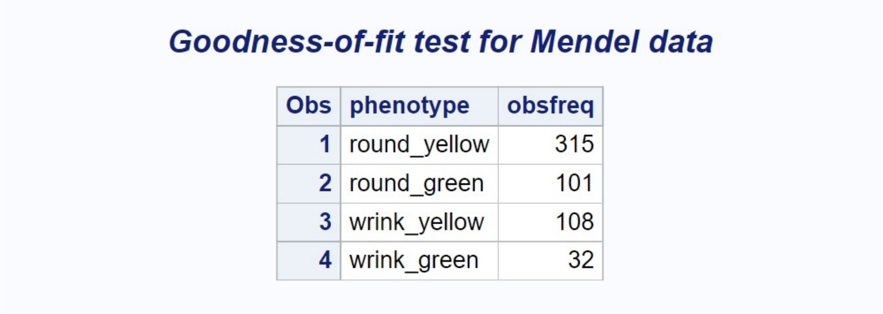
Goodness-of-fit test - SAS demo 2

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

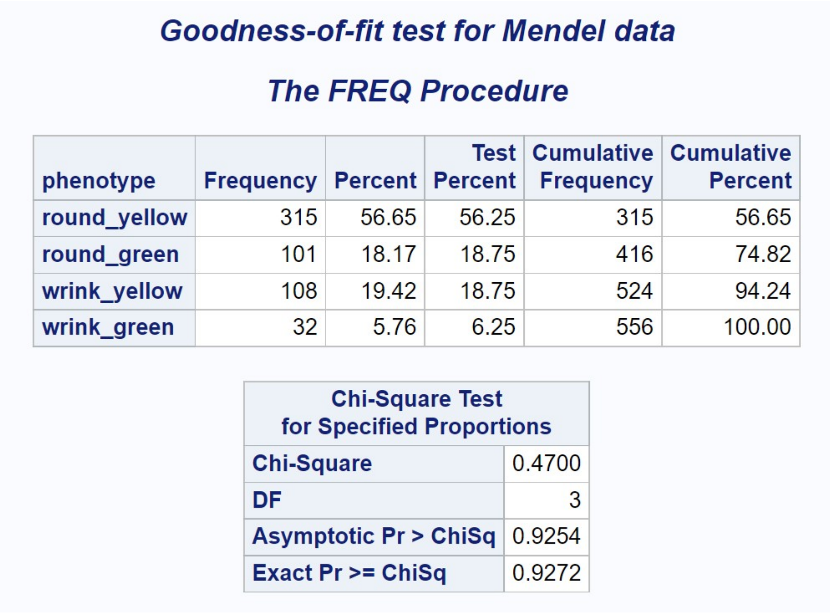
```
* gof_peas.sas;
title 'Goodness-of-fit test for Mendel data';
data peas;
    input phenotype :\$12. obsfreq;
    datalines;
round_yellow  315
round_green   101
wrink_yellow  108
wrink_green    32
;
run;
* Print data set;
proc print data=peas;
run;
* Goodness-of-fit test (Chi-square only);
proc freq 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

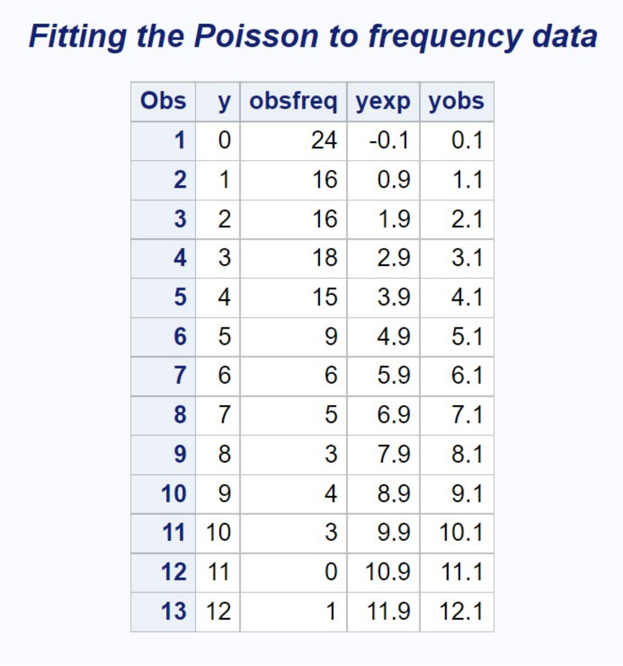
$$X^2 = 80.705 + 23.199 = 103.904. \tag{20.46}$$

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;
```

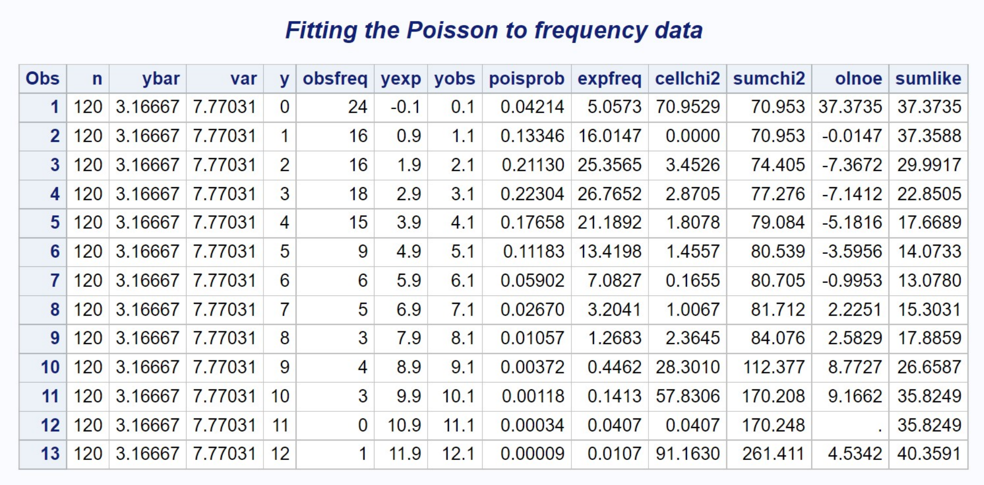
```
* Calculate expected frequencies using ybar;
data poisfit;
    if _n_ = 1 then set stats;
    set poisson;
    poisprob = pdf('poisson',y,ybar);
    expfreq = n*poisprob;
    * Calculate test values for each cell;
    cellchi2 = ((obsfreq - expfreq)**2)/expfreq;
    sumchi2 + cellchi2;
    olnoe = 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`

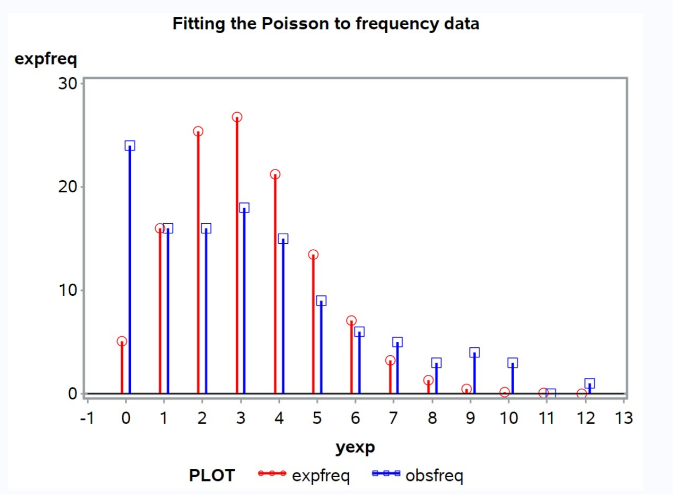


Figure 20.7: `Poisson_fit2_gof.sas - proc gplot`

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

$$\begin{aligned} E_{11} &= n\hat{p}_{11} && (20.52)\\ &= n\hat{p}_{1+}\hat{p}_{+1} && (20.53)\\ &= n\frac{O_{1+}}{n}\frac{O_{+1}}{n} && (20.54)\\ &= \frac{O_{1+}O_{+1}}{n}. && (20.55) \end{aligned}$$

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

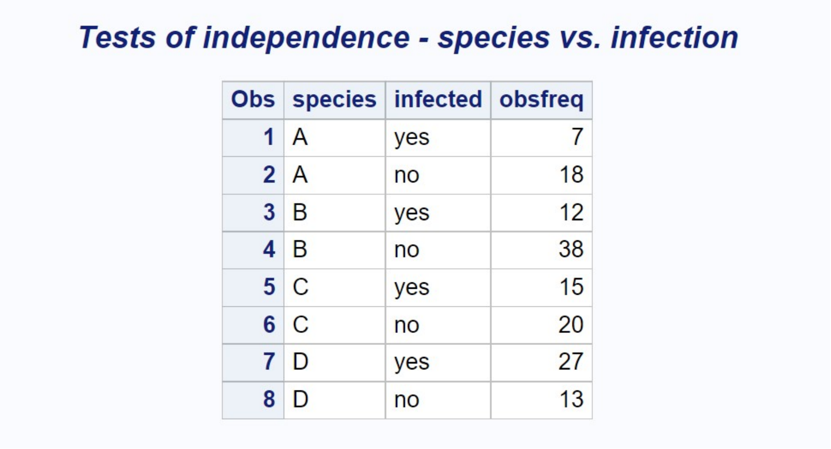
We can carry out the same calculations using SAS and `proc freq` (SAS Institute Inc. 2016). See program below. A two-way table of infection status and species is requested using the command `tables infected*species`. Likelihood ratio (G^2) and chi-square (X^2) tests are then requested using the `chisq` 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

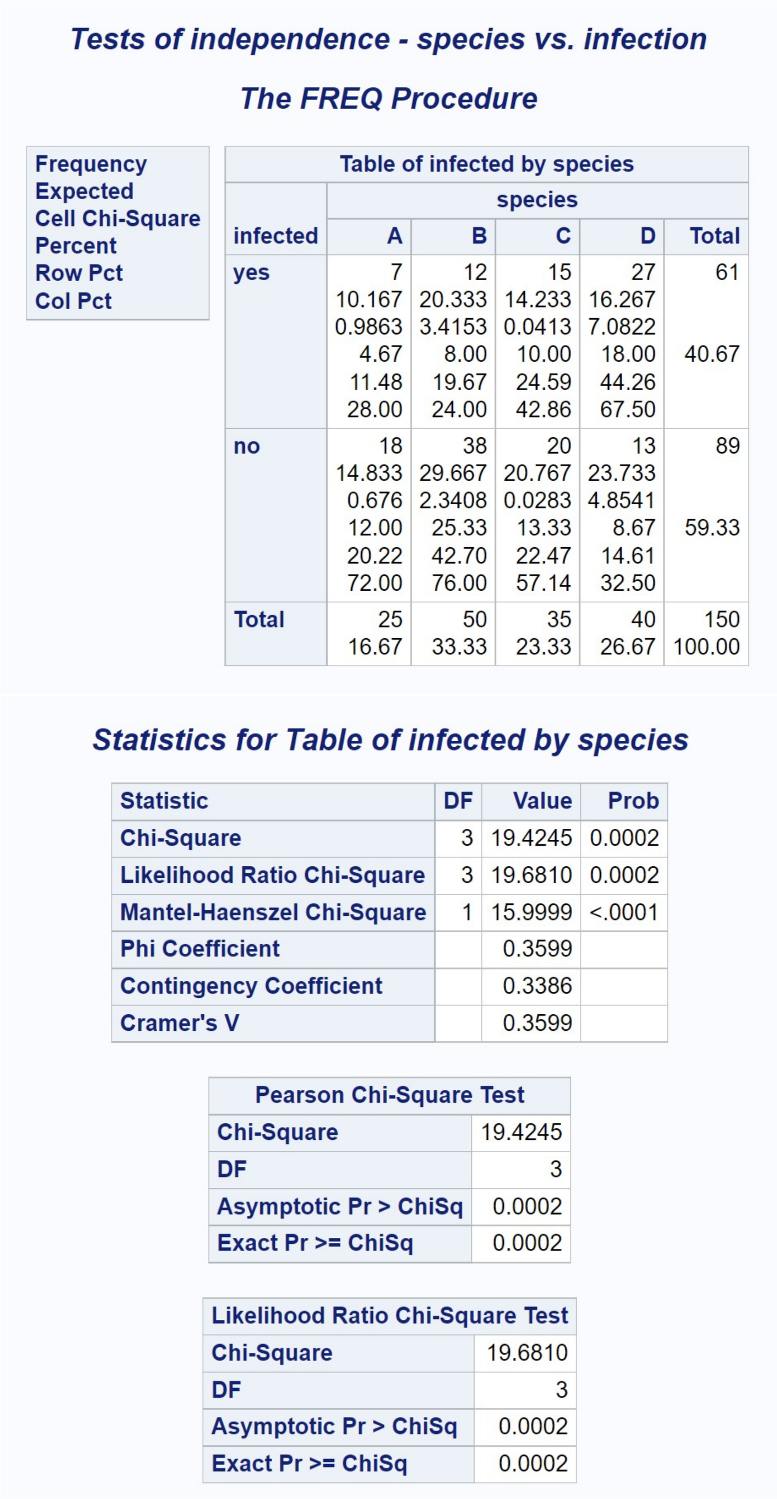
```
* chytrid.sas;
title "Tests of independence - species vs. infection";
data chytrid;
    input species $ infected $ obsfreq;
    datalines;
A  yes   7
A  no   18
B  yes  12
B  no   38
C  yes  15
C  no   20
D  yes  27
D  no   13
;
run;
* Print data set;
proc print data=chytrid;
run;
* Tests of independence;
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`

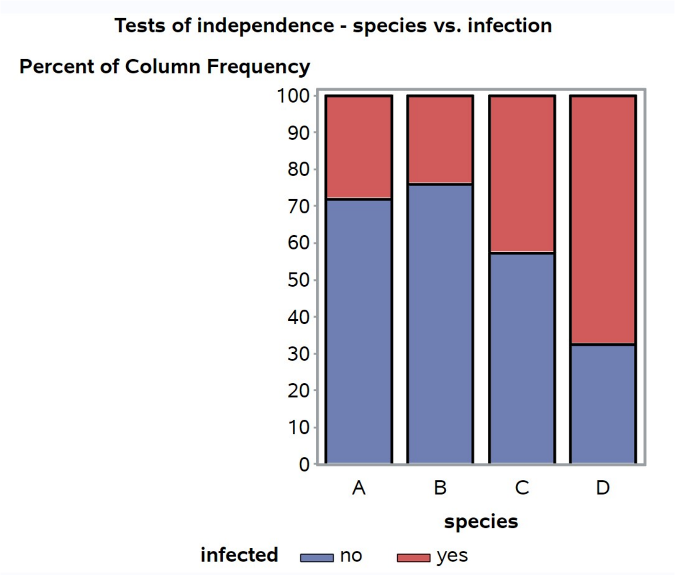


Figure 20.10: `chytrid.sas - gchart`

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).

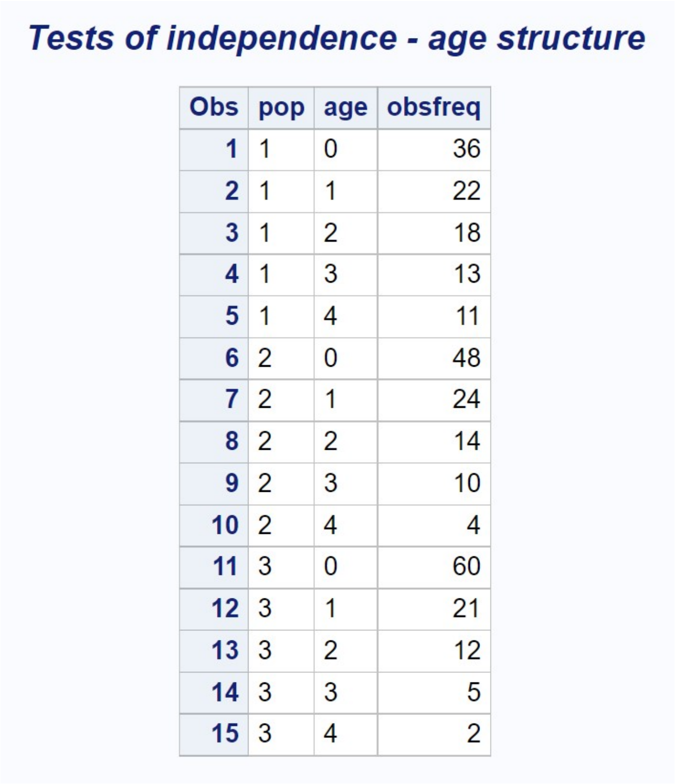
We will conduct tests of independence for these data using SAS and `proc freq` (see program below). As before, we will conduct both the likelihood ratio (G^2) and chi-square (X^2) tests. One difference in this program is that the option for exact tests is turned off, because they are quite time consuming (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);
```

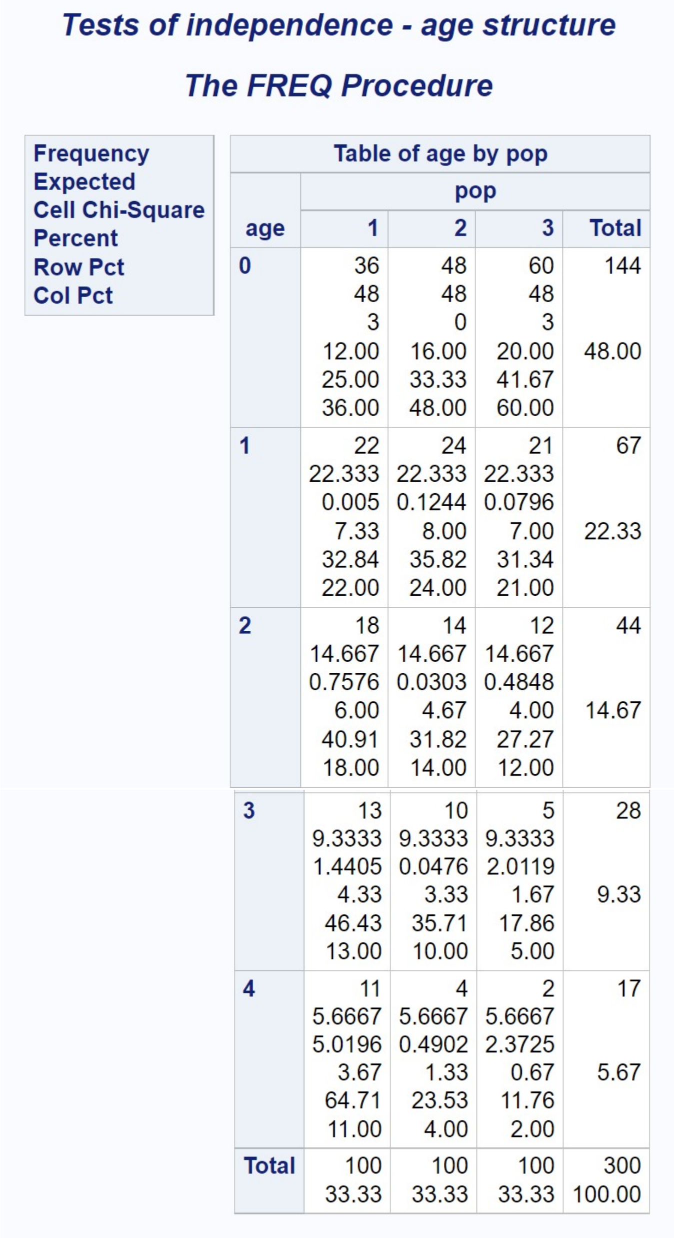
```
run;
quit;
```



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`



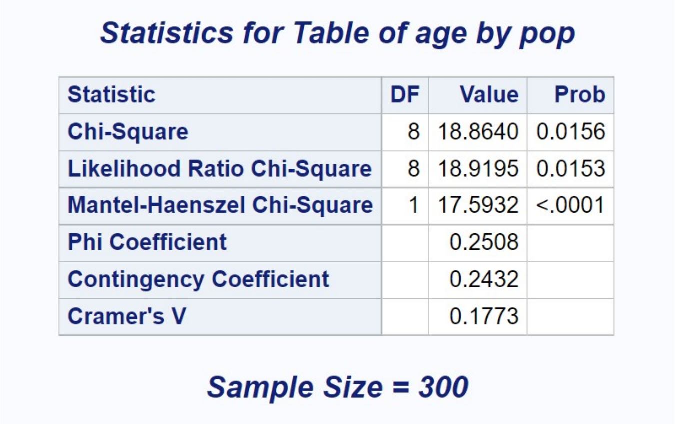
Tests of independence - age structure

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`

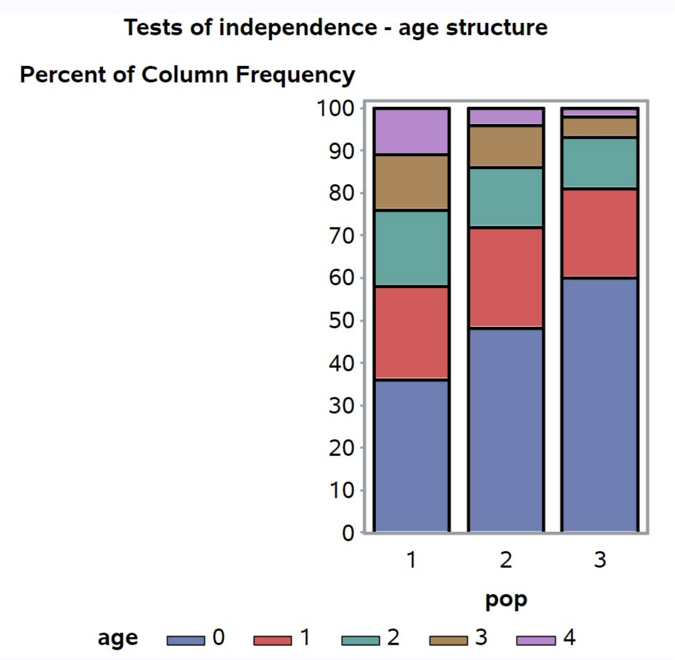


Figure 20.14: `age_structure.sas - proc gchart`

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?