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

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

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

$$Var[Y] = E[(Y - \mu)^2] = \int_{-\infty}^{\infty} (y - \mu)^2 f(y) dy. \tag{7.4}$$

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

$$E[Y'] = E[aY + b] = aE[Y] + b. \tag{7.5}$$

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

$$E[Y_1 + Y_2 + \ldots + Y_n] = E[Y_1] + E[Y_2] + \ldots + E[Y_n] = \sum E[Y_i]. \tag{7.6}$$

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

$$Var[Y'] = Var[aY + b] = a^2 Var[Y]. \tag{7.7}$$

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

$$Var[Y_1 + Y_2 + \ldots + Y_n] = Var[Y_1] + Var[Y_2] + \ldots + Var[Y_n] = \sum Var[Y_i]. \tag{7.8}$$

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:

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}. \tag{7.9}$$

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

$$E[\bar{Y}] = E\left[\frac{\sum Y_i}{n}\right] = \frac{E[\sum Y_i]}{n} = \frac{\sum E[Y_i]}{n} = \frac{nE[Y_i]}{n} = E[Y_i]. \tag{7.10}$$

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

$$Var[\bar{Y}] = Var\left[\frac{\sum Y_i}{n}\right] = \frac{Var[\sum Y_i]}{n^2} = \frac{\sum Var[Y_i]}{n^2} = \frac{nVar[Y_i]}{n^2} = \frac{Var[Y_i]}{n}. \tag{7.11}$$

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:

$$\sqrt{Var[\bar{Y}]} = \sqrt{\frac{Var[Y_i]}{n}} = \frac{\sqrt{Var[Y_i]}}{\sqrt{n}}. \tag{7.12}$$

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

$$s^2 = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}. \tag{7.13}$$

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 $Var[Y'] = a^2 Var[Y] = 2^2 Var[Y] = 4(1) = 4$.

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 100,000 observations having the 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`

Table 7.2: Expected value and variance

	Theory		Simulation	
Variable	$E[\cdot]$	$Var[\cdot]$	$\bar{Y}$	s^2
Y	1	1	1.002	0.999
Y'	3	4	3.003	3.997
$Y_1 + Y_2 + Y_3$	3	3	3.009	3.018
$\bar{Y}$	1	1/3	1.003	0.335
s^2	1	-	1.001	-

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. 7.2 - 7.4). 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. 7.2 vs. 7.4).

SAS Program

```
* Linear.sas;
options pageno=1 linesize=80;
goptions reset=all;
title 'Demonstration of expected value and variance rules';
data linear;
        * Loop to generate 100000 random observations;
        do i = 1 to 100000;
                a = 2;
                b = 1;
                * Generate y, y1, y2, y3 with N(1,2) 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(w=3)
        haxis=axis1 vaxis=axis2 wbarline=3 waxis=3 height=4;
        qqplot y yprime ysum ybar s2 / normal waxis=3 height=4;
        symbol1 h=3;
        axis1 order=(-6 to 12 by 0.2);
        axis2 order=(0 to 10000 by 2000);
run;
quit;
```

SAS Output

Demonstration of expected value and variance rules 1
11:24 Tuesday, August 30, 2011

Obs	i	a	b	mu	sig2	y	y1	y2	y3
1	1	2	1	1	1	2.40343	1.16387	1.40458	-0.42999
2	2	2	1	1	1	0.88987	1.40552	0.96997	-0.03135
3	3	2	1	1	1	3.33301	0.32146	-0.29182	-1.54156
4	4	2	1	1	1	1.07186	2.12205	1.59675	0.77104
5	5	2	1	1	1	2.00620	1.49543	1.58517	0.17379
6	6	2	1	1	1	0.19762	2.14458	0.60593	4.30395
7	7	2	1	1	1	-0.81464	1.75308	1.30842	0.60105
8	8	2	1	1	1	2.80298	1.06916	0.78637	0.65529
9	9	2	1	1	1	-0.56801	1.22604	0.15375	0.29170
10	10	2	1	1	1	-0.36265	1.34079	-0.36673	-2.41139
11	11	2	1	1	1	0.18388	-1.54080	1.11047	1.30994
12	12	2	1	1	1	1.62454	1.05907	0.65800	2.33563
13	13	2	1	1	1	1.83046	1.43715	-0.37539	0.83023
14	14	2	1	1	1	0.63400	0.38407	2.14804	1.30881
15	15	2	1	1	1	-0.87082	1.83641	0.60312	-0.56471
16	16	2	1	1	1	1.30249	2.13018	1.30823	0.54144
17	17	2	1	1	1	1.11287	-0.95954	-0.10480	-0.40775
18	18	2	1	1	1	1.58593	0.50497	1.22156	-0.10476
19	19	2	1	1	1	0.94855	1.95200	-0.19290	0.73783
20	20	2	1	1	1	2.76269	-1.04592	0.28742	2.47228
21	21	2	1	1	1	1.35196	1.55843	-0.65659	0.11237
22	22	2	1	1	1	0.90882	1.64321	1.11038	0.58658
23	23	2	1	1	1	-0.11425	1.50310	0.71810	-0.02761

Obs	yprime	ysum	ybar	s2
1	5.80686	2.13846	0.71282	0.99400
2	2.77973	2.34414	0.78138	0.54282
3	7.66603	-1.51192	-0.50397	0.90147
4	3.14372	4.48984	1.49661	0.46383
5	5.01240	3.25439	1.08480	0.62446
6	1.39523	7.05446	2.35149	3.45095
7	-0.62927	3.66255	1.22085	0.33754
8	6.60595	2.51082	0.83694	0.04474
9	-0.13603	1.67149	0.55716	0.34031
10	0.27470	-1.43732	-0.47911	3.52919
11	1.36775	0.87961	0.29320	2.53262
12	4.24908	4.05270	1.35090	0.76748

13	4.66092	1.89199	0.63066	0.85120
14	2.26799	3.84092	1.28031	0.77850
15	-0.74163	1.87482	0.62494	1.44171
16	3.60497	3.97985	1.32662	0.63128
17	3.22575	-1.47209	-0.49070	0.18780
18	4.17186	1.62176	0.54059	0.44073
19	2.89710	2.49693	0.83231	1.15685
20	6.52538	1.71377	0.57126	3.15485
21	3.70392	1.01421	0.33807	1.26478
22	2.81764	3.34017	1.11339	0.27912
23	0.77151	2.19358	0.73119	0.58590

Obs	i	a	b	mu	sig2	y	y1	y2	y3
24	24	2	1	1	1	2.08307	0.88622	1.07412	-0.47401
25	25	2	1	1	1	0.78354	3.10312	0.25982	1.98184

Obs	yprime	ysum	ybar	s2
24	5.16614	1.48633	0.49544	0.71371
25	2.56709	5.34477	1.78159	2.05116

The UNIVARIATE Procedure
Variable: y

Moments

N	100000	Sum Weights	100000
Mean	1.00156793	Sum Observations	100156.793
Std Deviation	0.99964147	Variance	0.99928307
Skewness	0.00181563	Kurtosis	-0.0101015
Uncorrected SS	200241.14	Corrected SS	99927.3079
Coeff Variation	99.8076557	Std Error Mean	0.00316114

Demonstration of expected value and variance rules 6
11:24 Tuesday, August 30, 2011

The UNIVARIATE Procedure
Variable: yprime

Moments

N	100000	Sum Weights	100000
Mean	3.00313586	Sum Observations	300313.586
Std Deviation	1.99928294	Variance	3.99713229
Skewness	0.00181563	Kurtosis	-0.0101015
Uncorrected SS	1301591.73	Corrected SS	399709.232
Coeff Variation	66.5731767	Std Error Mean	0.00632229

Demonstration of expected value and variance rules 9
11:24 Tuesday, August 30, 2011

The UNIVARIATE Procedure
Variable: ysum

Moments

N	100000	Sum Weights	100000
Mean	3.00918311	Sum Observations	300918.311
Std Deviation	1.73737373	Variance	3.01846747
Skewness	0.01349896	Kurtosis	0.01654247
Uncorrected SS	1207362.03	Corrected SS	301843.729
Coeff Variation	57.7357264	Std Error Mean	0.00549406

Demonstration of expected value and variance rules 12
11:24 Tuesday, August 30, 2011

The UNIVARIATE Procedure
Variable: ybar

Moments

N	100000	Sum Weights	100000
Mean	1.00306104	Sum Observations	100306.104
Std Deviation	0.57912458	Variance	0.33538527
Skewness	0.01349896	Kurtosis	0.01654247
Uncorrected SS	134151.337	Corrected SS	33538.1921
Coeff Variation	57.7357264	Std Error Mean	0.00183135

```
Demonstration of expected value and variance rules            15
                                        11:24 Tuesday, August 30, 2011

                    The UNIVARIATE Procedure
                         Variable:  s2

                              Moments

N                         100000    Sum Weights             100000
Mean                   1.00110176    Sum Observations    100110.176
Std Deviation          1.00051177    Variance            1.00102381
Skewness               2.05787409    Kurtosis            6.75960569
Uncorrected SS         200321.854    Corrected SS         100101.38
Coeff Variation        99.9410659    Std Error Mean       0.0031639
```

Figure 7.1: Frequency distribution for $Y \sim N(1, 1)$

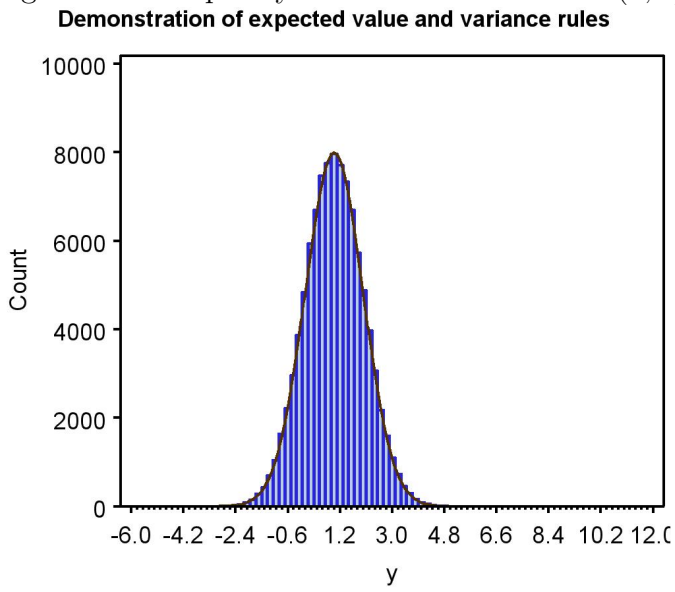


Figure 7.2: Frequency distribution for $Y' = 2Y + 1$

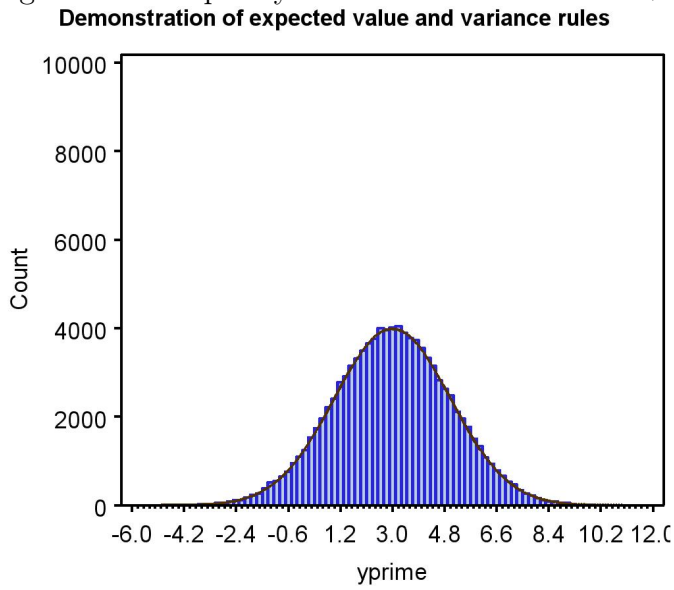


Figure 7.3: Frequency distribution for $Y_1 + Y_2 + Y_3$

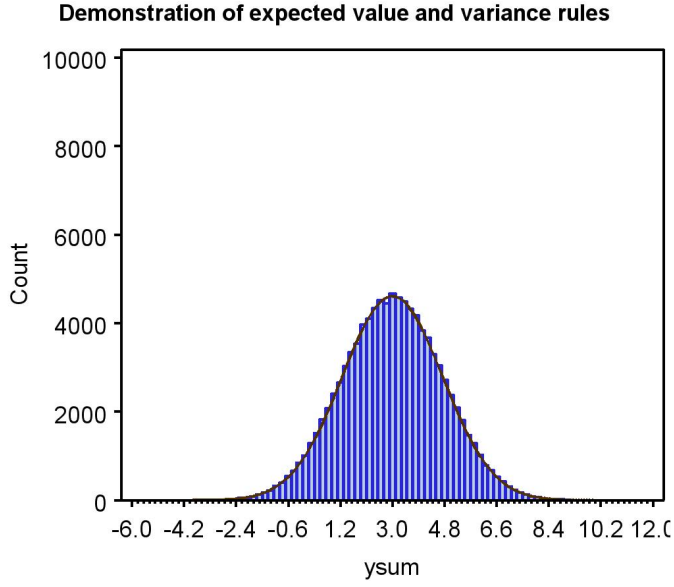
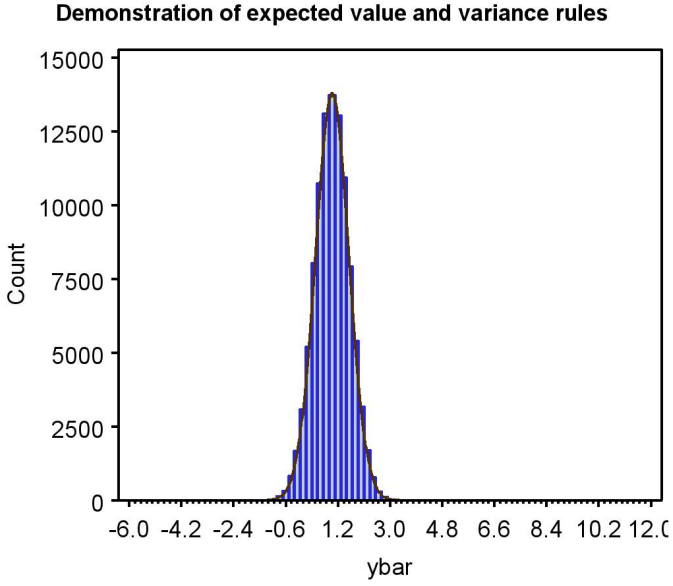


Figure 7.4: Frequency distribution for $\bar{Y}$



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 $Var[Y_i]$ 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 $n = 100$, in a SAS `data` step. A total of 100,000 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 $\bar{Y}$

n	Mean of $\bar{Y}$	Variance of $\bar{Y}$	$E[Y_i]$	$Var[Y_i]/n$
1	0.997	0.998	1.000	1.000
5	1.001	0.201	1.000	0.200
10	1.000	0.100	1.000	0.100
50	1.000	0.020	1.000	0.020

SAS Program

```
* central_limit_theorem.sas;
options pageno=1 linesize=80;
goptions reset=all;
title 'Demonstration of central limit theorem in action';
data cntrlmt;
        * Loop to generate 100000 random observations;
        do i = 1 to 100000;
                * 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(w=3) wbarline=3 waxis=3 height=4;
        qqplot y1 y5 y10 y50 / normal waxis=3 height=4;
        symbol1 h=3;
run;
quit;
```

SAS Output

Demonstration of central limit theorem in action 1
09:25 Thursday, May 6, 2010

Obs	y1	y5	y10	y50
1	3	1.4	0.7	1.08
2	0	1.0	1.3	1.22
3	0	0.4	0.6	0.98
4	1	1.6	0.9	0.96
5	1	0.4	1.0	1.18
6	1	1.4	1.1	0.76
7	0	1.0	0.8	0.88
8	1	1.0	1.1	0.96
9	2	1.8	1.0	1.10
10	4	1.0	1.0	1.00
11	0	0.4	1.3	0.94
12	1	0.2	1.1	0.84
13	0	1.8	0.9	0.80
14	1	1.2	0.8	1.10
15	0	0.8	0.8	0.86
16	2	0.6	0.7	1.00
17	0	1.4	0.8	1.00
18	1	2.2	0.6	0.98
19	0	1.6	0.4	1.26
20	2	0.6	0.9	0.80
21	1	0.8	0.7	0.86
22	0	0.8	1.3	0.90
23	4	1.0	0.9	0.84
24	2	1.2	0.7	1.04
25	0	0.8	0.8	0.82

Demonstration of central limit theorem in action 2
09:25 Thursday, May 6, 2010

The UNIVARIATE Procedure
Variable: y1

Moments

N	100000	Sum Weights	100000
Mean	0.99708	Sum Observations	99708
Std Deviation	0.99900023	Variance	0.99800145
Skewness	0.99072058	Kurtosis	0.937192
Uncorrected SS	199216	Corrected SS	99799.1474
Coeff Variation	100.192585	Std Error Mean	0.00315912

Demonstration of central limit theorem in action 5
09:25 Thursday, May 6, 2010

The UNIVARIATE Procedure
Variable: y5

Moments

N	100000	Sum Weights	100000
Mean	1.00095	Sum Observations	100095
Std Deviation	0.44812041	Variance	0.20081191
Skewness	0.45210017	Kurtosis	0.19930735
Uncorrected SS	120271.08	Corrected SS	20080.9897
Coeff Variation	44.7695104	Std Error Mean	0.00141708

Demonstration of central limit theorem in action 8
09:25 Thursday, May 6, 2010

The UNIVARIATE Procedure
Variable: y10

Moments

N	100000	Sum Weights	100000
Mean	1.000173	Sum Observations	100017.3
Std Deviation	0.31593491	Variance	0.09981487
Skewness	0.31057365	Kurtosis	0.11464285
Uncorrected SS	110015.99	Corrected SS	9981.38701

Coeff Variation	31.5880264	Std Error Mean	0.00099907

Demonstration of central limit theorem in action 11
09:25 Thursday, May 6, 2010

The UNIVARIATE Procedure
Variable: y50

Moments

N	100000	Sum Weights	100000
Mean	1.0000688	Sum Observations	100006.88
Std Deviation	0.14104417	Variance	0.01989346
Skewness	0.12418096	Kurtosis	0.00838865
Uncorrected SS	102003.086	Corrected SS	1989.32593
Coeff Variation	14.1034468	Std Error Mean	0.00044602

Figure 7.5: Frequency distribution for Y

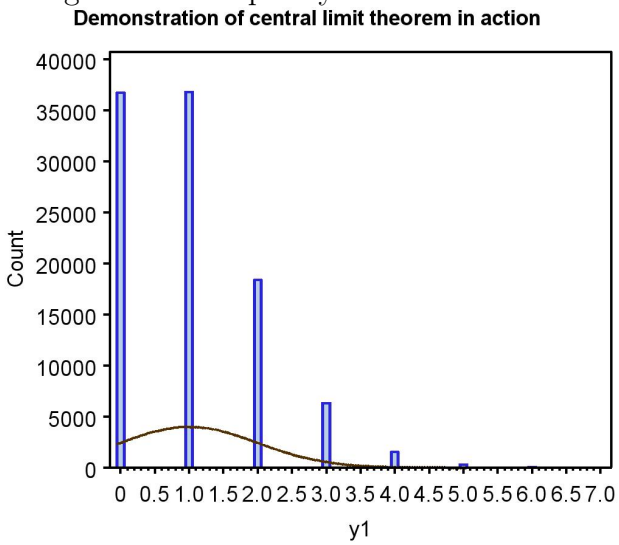


Figure 7.6: Frequency distribution for $\bar{Y}$ with $n = 5$

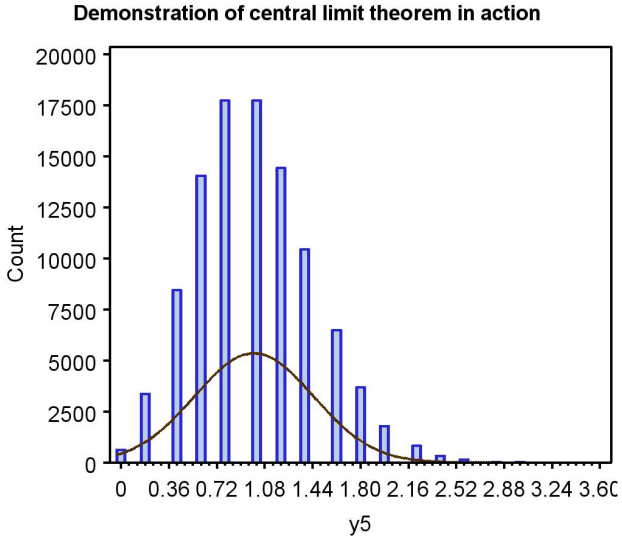


Figure 7.7: Frequency distribution for $\bar{Y}$ with $n = 10$

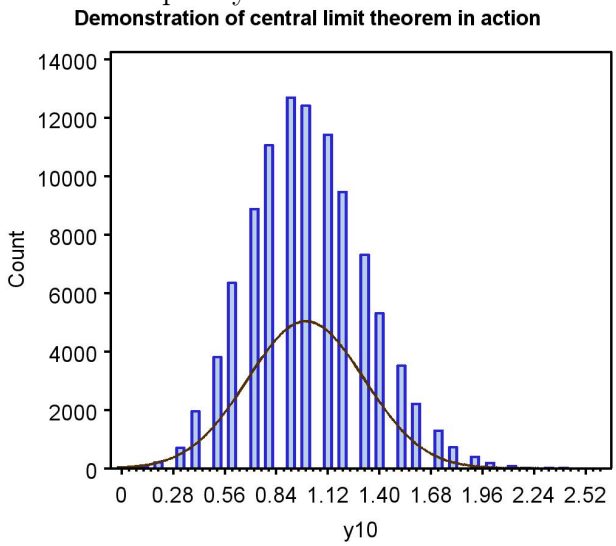
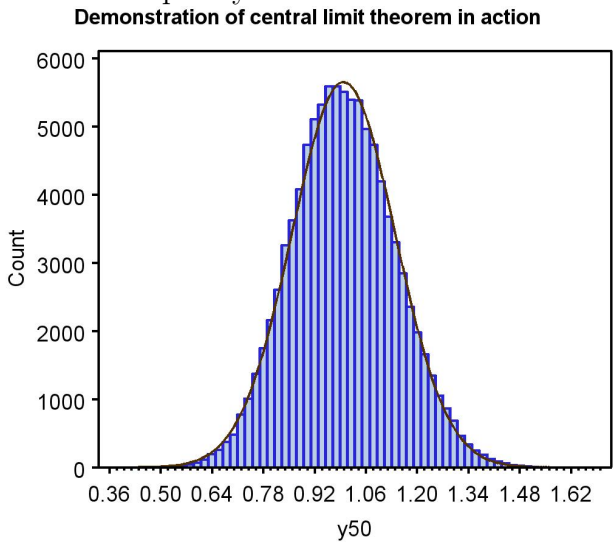


Figure 7.8: Frequency distribution for $\bar{Y}$ with $n = 50$



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

$$f(y) = \frac{e^{-y/\lambda}}{\lambda} \tag{7.14}$$

   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.