Chapter 4. SAMPLING DISTRIBUTIONS

In agricultural research, we commonly take a number of plots or animals for experimental use. In effect we are working with a number of individuals drawn from a large population. Usually we don't know the exact characteristics of the parent population from which the plots or animals are drawn. Hopefully the samples we draw and the statistics we compute from them are close approximations of the parameters of the parent populations. To ensure a representative sample we use the principle of randomization. A random sample is one drawn so that each individual in the population has the same chance of being included.

The parameters of a population are based on all of its variates and are therefore fixed. The statistics vary from sample to sample. Therefore the possible values of a statistic constitute a new population, a distribution of the sample statistic.

4.1 Distribution of Sample Means

Consider a population of N variates with mean μ and standard deviation σ, and draw all possible samples of r variates. Assume that the samples have been replaced before each drawing, so that the total number of different samples which can be drawn is the combination of N things taken r at a time, that is $M = \binom{N}{r}$. The mean of all these sample means $(\overline{Y}_1, \overline{Y}_2, ..., \overline{Y}_M)$ is denoted by $\mu_{\bar{y}}$ and their standard deviation by $\sigma_{\bar{y}}$, also known as the standard error of a mean. The mean of the sample means is the same as the mean of the parent population, μ, e.g.

$$\mu_{\bar{y}} = \Sigma\overline{Y}_i / M = \mu = \Sigma Y_i / N$$

The variance of the sample means $(\sigma^2_{\bar{y}})$ equals the variance of the parent (σ^2) population divided by the sample size (r) and multiplied by a factor f.

$$\sigma^2_{\bar{y}} = \Sigma(\overline{Y}_i - \mu_{\bar{y}})^2 / M = (\sigma^2 / r) \bullet f$$

where $f = (N - r) / (N - 1)$

Note that the standard error of a mean approaches the standard deviation of the parent population divided by the square root of the sample size, $\sigma_{\bar{y}} = \sigma / \sqrt{r}$ for a large population (i.e., f approaches unity). The larger the size of a sample, the smaller the variance of the sample mean.

Consider samples taken from a normal population. Figure 4-1 illustrates the relationship of the parent population (r = 1) with the sampling distributions of the means of samples of size r = 8 and r = 16.

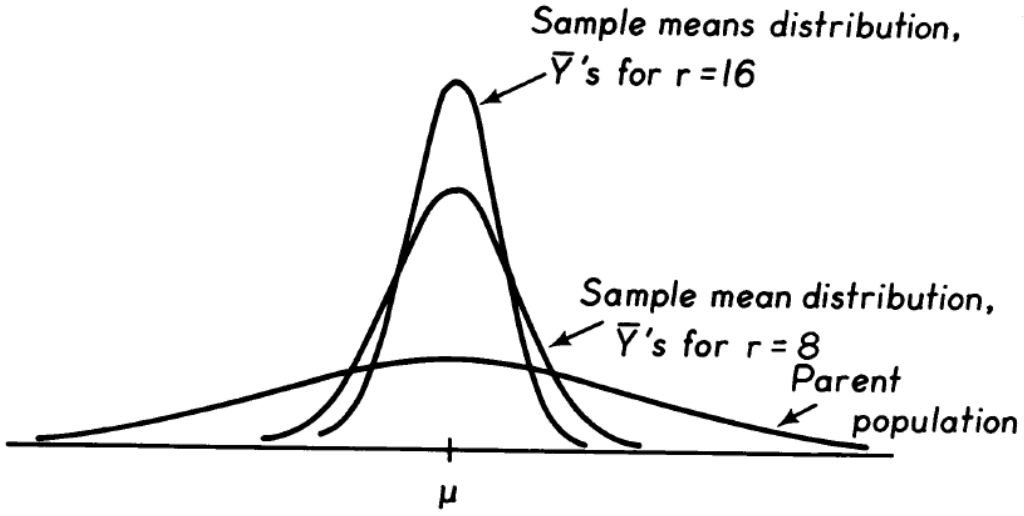

Figure 4-1

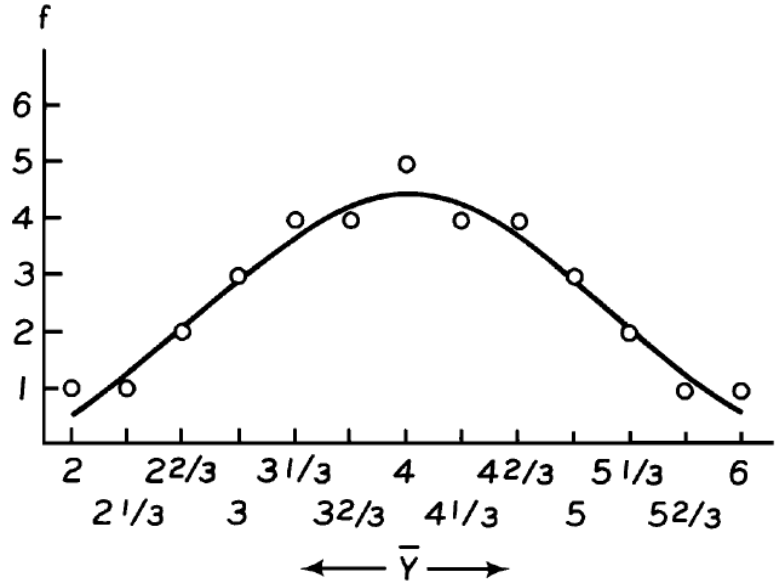

Figure 4-2. The relation of the frequencies of means for r = 3 from the population 1,2,3,4,5,6,7 and the normal distribution.

Even when the variates of the parent population are not normally distributed, the means generated by samples tend to be normally distributed. This can be illustrated by considering samples of size 3 from a simple non-normal population with variates 1,2,3,4,5,6, and 7. Table 4-1 presents all possible sample means, and Figure 4-2 shows the frequency distribution of the means which approaches the normal frequency curve.

Table 4-1. All possible different samples of size 3 from the population 1, 2, 3, 4, 5, 6, 7 with $\mu = 4$ and $\sigma = 2$.

Sample No.	Sample			$\bar{y}$	Sample No.	Sample			
1	1	2	3	2	19	2	3	7	4
2	1	2	4	2 1/3	20	2	4	5	3 2/3
3	1	2	5	2 2/3	21	2	4	6	4
4	1	2	6	3	22	2	4	7	4 1/3
5	1	2	7	3 1/3	23	2	5	6	4 1/3
6	1	3	4	2 2/3	24	2	5	7	4 2/3
7	1	3	5	3	25	2	6	7	5
8	1	3	6	3 1/3	26	3	4	5	4
9	1	3	7	3 2/3	27	3	4	6	4 1/3
10	1	4	5	3 1/3	28	3	4	7	4 2/3
11	1	4	6	3 2/3	29	3	5	6	4 2/3
12	1	4	7	4	30	3	5	7	5
13	1	5	6	4	31	3	6	7	5 1/3
14	1	5	7	4 1/3	32	4	5	6	5
15	1	6	7	4 2/3	33	4	5	7	5 1/3
16	2	3	4	3	34	4	6	7	5 2/3
17	2	3	5	3 1/3	35	5	6	7	6
18	2	3	6	3 2/3					

The mean and standard deviation of the distribution of the sample means are:

$$\mu_{\bar{y}} = \frac{1}{35}(2 + 21/3 + 22/3 + ... + 52/3 + 6) = 4 = \mu$$

$$\sigma\frac{2}{\bar{y}} = \frac{1}{35}\{(2-4)^2 + (21/3-4)^2 + ... + (52/3-4)^2 + (6-4)^2\} = \frac{\sigma^2}{r} \bullet (\frac{N-r}{N-1}) = \frac{4}{3} \bullet (\frac{4}{6}) = \frac{8}{9}$$

$$\sigma_{\bar{y}} = \sqrt{8/9}$$

Note that in this particular case, we have used a simple population with only seven elements. Sample means from samples with increasing size, from a large population will more closely approach the normal curve. This tendency of sample means to approach a normal distribution with increasing sample size is called the central limit theorem.

4.2 The Distribution of Sample Mean Differences

In section 4.1 we mentioned that the means of all possible samples of a given size (r_1) drawn from a large population of Y's are approximately normally distributed with $\mu_{\bar{y}} = \mu_y$ and $\sigma^2_{\bar{y}} = \sigma^2_y / r_1$. Now consider drawing samples of size r_2 from another large population, X's. The parameters of these sample means are also approximately normally distributed with $\mu_{\bar{x}} = \mu_x$ and $\sigma^2_{\bar{x}} = \sigma^2_x / r_2$. An additional approximately normal population is generated by taking differences between all possible means, $\bar{Y} - \bar{X} = \bar{d}$, with the parameters $\mu_{\bar{d}}$ and $\sigma^2_{\bar{d}}$

$$\mu_{\bar{d}} = \mu_{\bar{y}} - \mu_{\bar{x}} = \mu_y - \mu_x$$

and

$$\sigma^2_{\bar{d}} = \sigma^2_{\bar{y}} + \sigma^2_{\bar{x}} = \sigma^2_y / r_1 + \sigma^2_x / r_2$$

When the variances of the parent populations are equal,

$\sigma^2_y = \sigma^2_x \ (= \sigma^2)$ and sample sizes are the same, $r = r_1 = r_2$ then $\sigma^2_{\bar{d}} = 2\,\sigma^2 / r$.

The square root of the variance of mean differences, $\sigma_{\bar{d}}$, is usually called the standard error of the difference between sample means. Figure 4-3 diagrams the generation of a population of mean differences by repeated sampling from two populations of individual variates and indicates relationships among the parameters.

The relationships among the population parameters developed in Sections 4-1 and 4-2 are important in statistical evaluation. With information about the parent population one can estimate parameters associated with a sample mean or the difference between two sample means. This will be discussed further in later chapters.

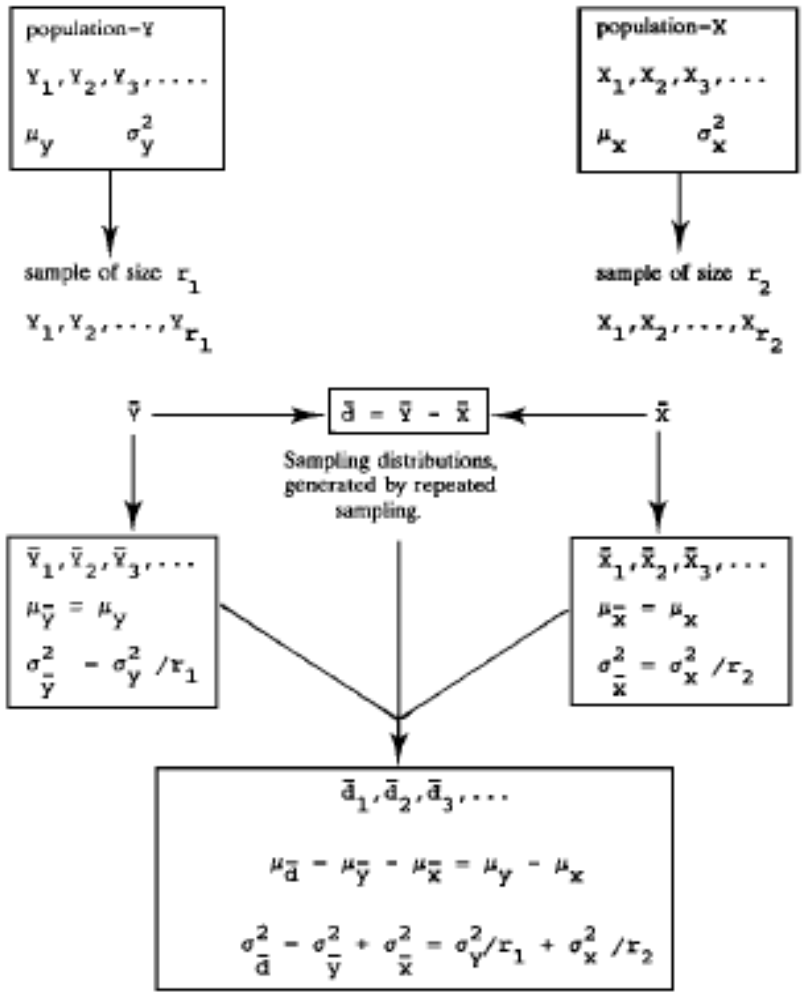

Figure 4-3. Relationships between parameters of a population of sample mean differences and parent populations.

4.3 Normal Approximation to Binomial

Although a formula is given in Chapter 3 to calculate probabilities for binomial events, if the number of trials (n) is large the calculations become tedious. Since many practical problems involve large samples of repeated trials, it is important to have a more rapid method of finding binomial probabilities.

It was also pointed out in Chapter 3 that the normal distribution is useful as a close approximation to many discrete distributions when the sample size is large. When $n \geq 30$, the sample is usually considered large. In this section we will show how the normal distribution is used to approximate a binomial distribution for ease in the calculation of probabilities.

Since the normal frequency curve is always symmetric, whereas the binomial histogram is symmetric only when $p = q = 1/2$, it is clear that the normal curve is a better approximation of the binomial histogram if both p and q are equal to or nearly equal to 1/2. The more p and q differ from 1/2, the greater the number of trials are required for a close approximation. Figure 4-4 shows how closely a

normal curve can approximate a binomial distribution with n = 10 and p = q = 1/2. Figure 4-5 illustrates a case where the normal distribution closely approximates the binomial when p is small but the sample size is large.

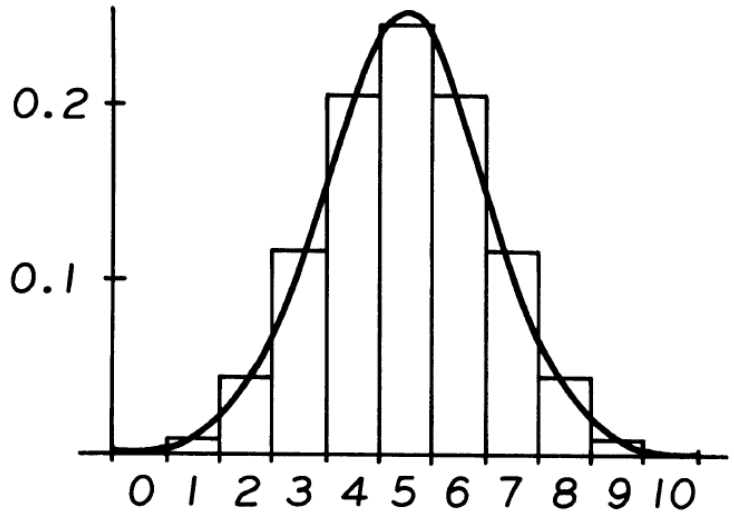

Figure 4-4. Binomial distribution for p = 0.5 and n = 10.

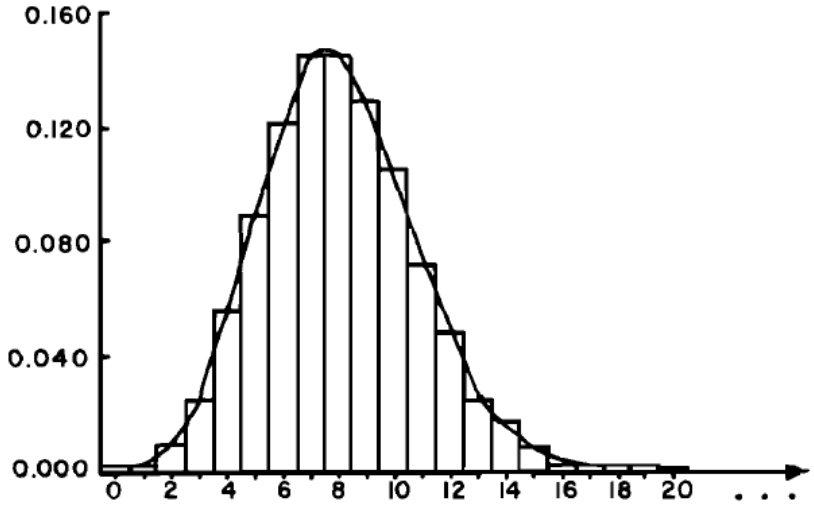

Figure 4-5. Binomial distribution for p = 0.08 and n = 100.

To use the normal curve to approximate discrete binomial probabilities, the area under the curve must include the area of the block of the histogram at any value of r, the number of occurrences under consideration. To include the block centered at r, the value of Y to be used in the normal curve equation for the normal deviate must be adjusted by adding 1/2 to, or subtracting 1/2 from the value of r. The calculation can be described by the following steps:

Step 1. Compute the mean and the standard deviation

$\mu = np, \quad \sigma = \sqrt{npq}$

Step 2. In order to find the corresponding normal deviate (Y) for a given r, 1/2 must be either added to or subtracted from r to include the block centered at r.

$Y = r - 1/2$ or $Y = r + 1/2$

Step 3. Standardize the normal deviate Y, by computing Z.

$Z = (Y - \mu) / \sigma = (Y - np) / \sqrt{npq}$

Step 4. From Appendix Table A-4, find the probability of the occurrence of a random standard normal deviate that is equal to or greater than, or equal to or smaller than Z.

Step 5. Compute the required probability. This depends on the nature of the problem and is illustrated by the four cases below.

Example 4-1. If 8% of a particular canned product is known to be underweight, what is the probability that a random sample of 100 cans will contain (a) 14 or more underweight cans (b) 4 or fewer underweight cans, (c) 5 or more underweight cans, (d) more than 4 but less than 15 underweight cans?

Step 1. $\mu = np = 100\,(0.08) = 8.0$
$\sigma = \sqrt{npq} = \sqrt{100(0.08)(0.92)} = 2.71$

(a) To find the probability of 14 or more underweight cans see Figure 4-6.

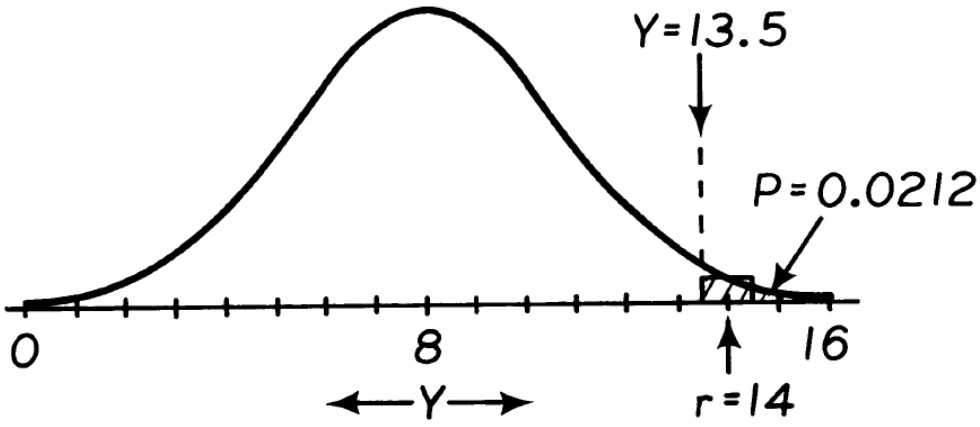

Figure 4-6.

Step 2. $Y = 5 - 1/2 = 14 - 1/2 = 13.5$

Step 3. $Z = (13.5 - 8.0) / 2.71 = 2.03$

Step 4. P ($Z \geq 2.03$) = 0.0212 from Appendix Table A-4.

Step 5. The required probability in this case is the one obtained from Step 4, 0.0212.

(b) To find the probability of 4 or fewer underweight cans, see Figure 4-7.

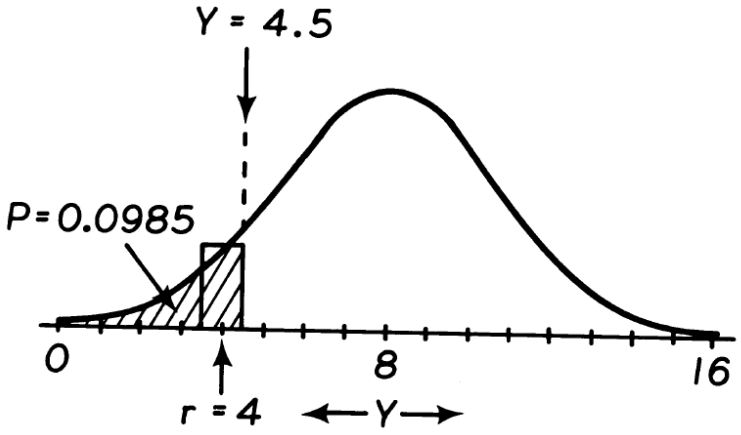

Figure 4-7.

Step 2. $Y = r + 1/2 = 4 + 0.5 = 4.5$

Step 3. $Z = (4.5 - 8.0) / 2.71 = 1.29$

Step 4. Appendix Table A-4 gives only positive values for Z, i.e., for $Z \geq 0$. since the distribution is symmetrical about Z = 0, probabilities for negative values of Z are determined by ignoring the sign. Therefore, P ($Z \leq -1.29$) = P($Z \geq 1.29$) = 0.0985.

Step 5. The required probability in this case is the one obtained from Step 4, 0.0985, or about 10%.

(c) To find the probability of 5 or more underweight cans, see Figure 4-8.

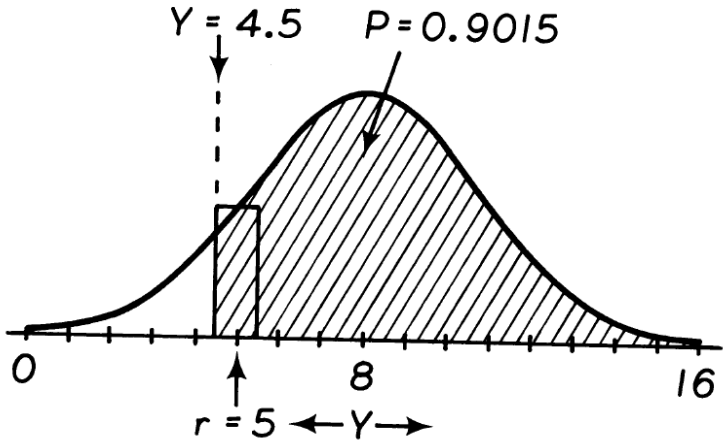

Figure 4-8.

Step 2. $Y = r - 1/2 = 5 - 1/2 = 4.5$

Step 3. $Z = (4.5 - 8.0) / 2.71 = -1.29$

Step 4. From Appendix Table A-4, $P(Z \leq -1.29) = P(Z \geq 1.29) = 0.0985$.

Step 5. The problem is to find the probability that $P(Z \leq -1.29) = 1 - 0.0985 = 0.9015$.

(d) To find the probability that more than 4 but less than 15 cans are underweight, we must find the probability of 5 more and 14 or less underweight cans as in Figure 4-9.

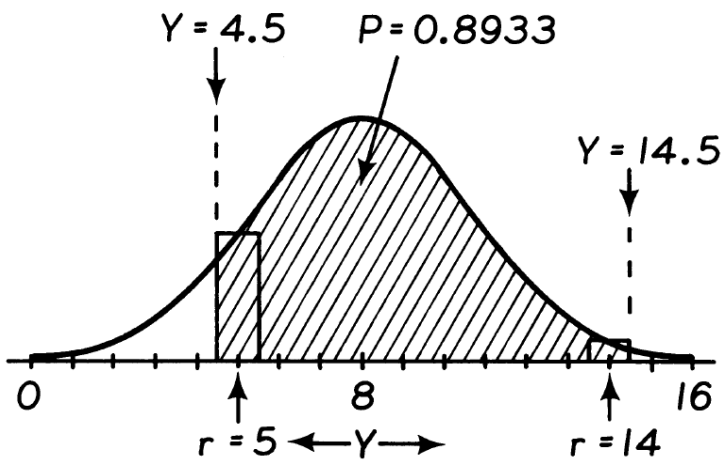

Figure 4-9.

Step 2. Note in this case that we need to find the probability between two r values, r_1 and r_2.

$Y_1 = r_1 - 1/2 = 5 - 1/2 = 4.5$

$Y_2 = r_2 + 1/2 = 14 + 1/2 = 14.5$

Step 3. Now we have to calculate two standardized Z values, Z_1 and Z_2.

$$Z_1 = \frac{4.5 - 8.0}{2.71} = -1.29$$

$$Z_2 = \frac{14.5 - 8.0}{2.71} = 2.40$$

Step 4. From Appendix Table A-4, we find the two probabilities:

$P(Z \leq -1.29) = (P(Z \geq 1.29) = 0.0985$

$P(Z \geq 2.40) = 0.0082$

Step 5. The required probability is the area between the two Z values, which is equal to:

$1 - P(Z \leq -1.29) - P(Z \geq 2.40)$
$= 1 - 0.0985 - 0.0082$
$= 1 - 0.1067 = 0.8933$

4.4 Chi-Square Distribution

We now introduce a distribution called chi-square. This distribution is related to S^2, the variance of a sample. We have shown that $Z = (Y - \mu)/\sigma$, is a standard normal deviate. The square of a single standard normal deviate is called chi-square with 1 degree of freedom. The sum of squares of r independent standard normal deviates is called the chi-square with r degrees of freedom, i.e.,

$$\chi^2 = \sum_{\Sigma}^{r} Z^2 = \sum_{\Sigma}^{r} \left(\frac{Y - \mu}{\sigma}\right)^2$$

In sampling from a normal distribution, the sample variance is:

$$S^2 = \frac{\Sigma(Y - \overline{Y})^2}{r - 1}$$

It can be shown mathematically that S^2 relates to χ^2 in the following way:

$$\chi^2 = [(r - 1)\, S^2] / \sigma^2, \text{ with } r - 1 \text{ degrees of freedom.}$$

Obviously, χ^2 cannot have negative values.

The disturbance of χ^2 depends on the degrees of freedom of the sample variance. For each degree of freedom, there is a χ^2 distribution. Figure 4-10 gives ^{2} distributions for several selected degrees of freedom.

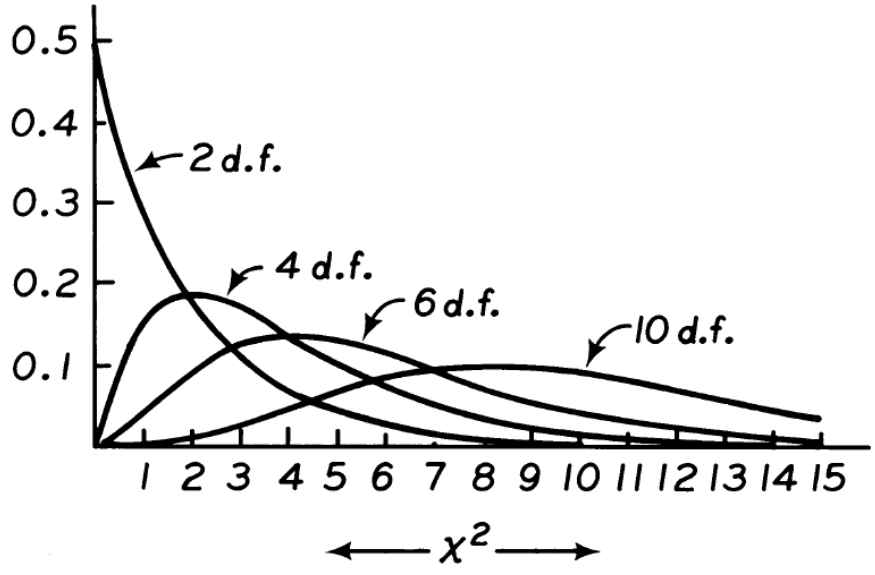

Figure 4-10. Some X^2 distribution curves.

Note that the χ^2 curve becomes more symmetrical as the degrees of freedom increase. Appendix Table A-5 gives the χ^2 values which have certain specific cumulative probabilities for various degrees of freedom. Since the χ^2 distribution is related to a measure of dispersion from a sample, it has many applications to problems of the significance of an estimated dispersion, an association between two classification and the closeness of fitting observations to a theoretical model. These applications are discussed in Chapter 11.

4.5. Student t-Distributions

In section 4.1, it was pointed out that a population of sample means of size r is approximately normally distributed with mean $\mu_{\bar{y}} = \mu$ and $\sigma^2_{\bar{y}} = \sigma^2 / r$, where μ and σ^2 are the mean and variance of the parent population. This population can also be standardized (section 3.4) to become the standard normal distribution,

$$Z = \frac{\overline{Y} - \mu}{\sigma_{\bar{y}}}$$

Seldom, if ever, do we know the population parameters, μ and $\sigma_{\bar{y}}$. When the standard error, $\sigma_{\bar{y}}$, is unknown, it is estimated from the sample by $S_{\bar{y}}$, the sample standard error.

$$S_{\bar{y}} = S / \sqrt{r} \text{ and } S^2 = \Sigma(Y - \overline{Y})^2 / (r - 1)$$
$$= [\Sigma Y^2 - (\Sigma Y)^2 / r] / (r - 1)$$

If $\sigma_{\bar{y}}$ is replaced in the above Z formula by its estimator, $S_{\bar{y}}$, the distribution is no longer normal but has a similar distribution called Student's t.

$$t = (\overline{Y} - \mu) / S_{\bar{y}} \qquad \text{with r - 1 degrees of freedom.}$$

There is a t-distribution for every sample size, since the standard error depends on the number of variates in the sample. This distribution is named after its discover, W. S. Gosset (1876-1937), a statistician at Guinness Brewery in Dublin, Ireland, who wrote under the pseudonym "Student." Student t-distributions are tabulated in Appendix Table A-6. These distributions are symmetrical, bell-shaped curves that look normal but are flatter. The mean of a t-distribution is zero and its variance is r/(r - 2) for r > 2. This is always greater than 1, the variance of Z. However, as sample size increases, the variance approaches 1. Therefore, the t-distribution approaches the standard normal. This can be seen from the last line of Appendix Table A-6 where the d.f. becomes infinite, since the values entered are the same as the Z-scores (appendix Table A-4) for any specific probability listed at the top of Appendix Table A-6. Figure 4-11 illustrates how the t-distribution relates to the standard normal distribution as sample size increases.

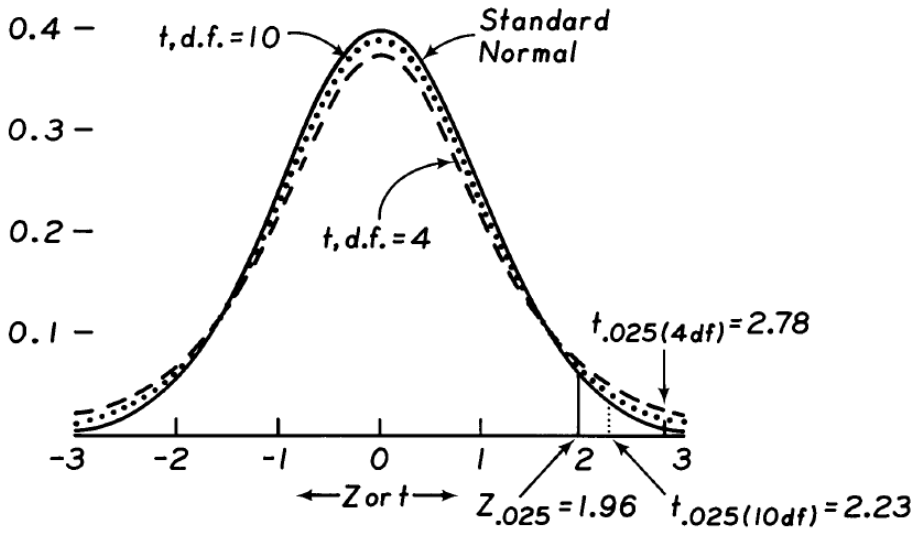

Figure 4-11. t-distribution curves.

With sample sizes greater than 30, the t-distribution is indistinguishable from the normal for all practical purposes.

The numerator of the t-formula, $\overline{Y} - \mu$, gives the magnitude of the difference between a sample mean $\overline{Y}$ and the population mean μ. The likelihood that $\overline{Y}$ is not different from μ, taking into account chance variation, is evaluated by the t-distribution. Since μ is not known, it is usually replaced by a hypothetical value for calculations. This hypothetical value can either be a guessed population constant or an important practical threshold value. One of the uses of the t-distribution is to compare the observed $\overline{Y}$ to a hypothetical value. Another use is to calculate two values, based on the t-formula, within which will fall with a specific probability. These uses relate to the concept of testing hypotheses and confidence intervals to be discussed in Chapter 5.

4.6 Fi Distributions

Another important sampling distribution is the F distribution, named in honor of R. A. Fisher (1890-1962), who first developed and described it. This distribution relates to a random variable which is defined as the ratio of the unbiased estimates of two population variances, that is

$$F = S_1^2 / S_2^2$$

This ratio has several applications, two of which are widely used: the test of equality between two variances and the test of equality of two or more means. For comparing two means (numerator df = 1), the F value is the square of the Student's t, i.e., $F = t^2$ for any given probability. Mathematically, the F distribution is derived from the ratio of two positive quantities, the range of the F values is from zero to infinity. The shape of the F curve depends on two parameters, the degrees of freedom associated with the sample of variance of the numerator and the degrees of freedom associated with the sample variance of the denominator. Generally, it is skewed to the right. However, when one or both degrees of freedom increase, it tends to be more symmetrical. Some specific F curves are shown in Figure 4-12.

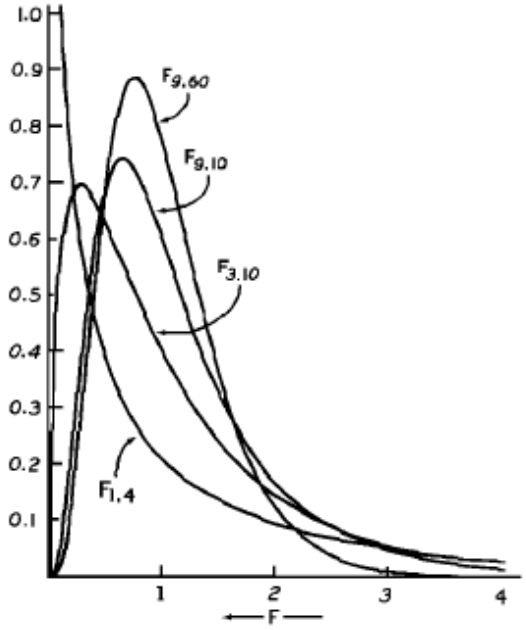

Figure 4-12

The F distribution is tabulated in Appendix Table A-7. Note that the distribution is identified by degrees of freedom for a numerator and denominator and thus a single table only permits the listing of a few possibilities. That is, the table entry F value at the joint point of the values of df_1 , the degrees of freedom associated with the numerator in the upper row of the table, and df_2, the degrees of freedom associated with the denominator in the first column, will cut off the right hand 10%, 5%, or 1% of the distribution.

The mean and variance of an F distribution are

$$\mu_F = df_2 / (df_2 - 2)$$

and

$$\sigma_F^2 = 2(df_2)^2 \bullet (df_1 + df_2 - 2) / df_1 \bullet (df_2 - 2)^2 \bullet (df_2 - 4)$$

For $df_2 \leq 2$, the mean cannot be estimated, and for $df_2 \leq 4$ the variance cannot be estimated from the above formulas.

SUMMARY

1. For any population with mean μ, and variance σ^2, the mean of all sample means of size r also equals μ and the variance equals σ^2/r. The standard error of a sample is $\sigma / \sqrt{r}$.

2. The distribution of sample means tends to the normal as the sample size increases, regardless of the shape of the parent population from which the samples are drawn. This tendency is known as the Central Limit theorem.

3. The difference between the means of two samples $(\bar{d} = \overline{Y} - \overline{X})$ of sizes r_1 and r_2 that are independently drawn from two populations with parameters μ_y, σ^2_y, and μ_x, σ^2_x, respectively, has a mean equal to $\mu_{\bar{d}}$ and variance $\sigma^2_{\bar{d}}$ where,

$$\mu_{\bar{d}} = \mu_y - \mu_x$$

and

$$\sigma^2_{\bar{d}} = \sigma^2_y / r_1 + \sigma^2_x / r_2$$

4. the standard normal curve can be used to approximate probabilities of many types of discrete events. For example, the probability of r occurrences from n binomial trials ($n \geq 30$), can be approximated by applying the following formula:

$$Z = (y - np) - \sqrt{npq}$$

5. Chi-square distribution is widely used for inference about a measure of dispersion, i.e., it is used to determine whether the variance of a sample is equal to its population variance. Chi-square can also be used to test hypotheses and to test fitness of observations to a theoretical model. Appendix Table A-5 lists some critical values with selected degrees of freedom and probabilities. The mean and variance of a ^{2} are (r - 1) and 2 (r - 1) respectively, where (r - 1) is the degrees of freedom.

6. A Student t-distribution is used for inferences about a population mean, where the sample size is small ($r \leq 30$), and the population variance is unknown. Student-4 related to a sample mean is expressed as,

$$t = (\overline{Y} - \mu) / S_{\bar{y}}$$

The t distribution depends on the degrees of freedom of the sample (r - 1). Student t-distributions are tabulated in Appendix Table A-6. The mean of a 1 distribution is 0 and the variance is r/(r - 2) for r > 2.

7. The F distribution is widely used to test the equality of two variances and the equality among several means. The mathematical expression of the F distribution is very complicated and depends on two degrees of freedom: df_1, the degrees of freedom associated with the numerator and df_2, the degrees of freedom associated with the denominator. The mean and variance of an F variable are,

$$\mu_F = df_2 / (df_2 - 2) \qquad \text{for } df_2 > 2$$

$$\sigma^2_F = \frac{2(df_2)^2 \ (df_1 + df_2 - 2)}{df_1 (df_2 - 2)^2 \ (df_2 - 4)} \qquad \text{for } df_2 > 4$$

Some tabular F values are shown in Appendix Table A-7.

EXERCISES

1. Suppose samples are taken from a population of heifers with known mean and variance of weight gains of 188 lbs. and 2200 square lbs., respectively.

   a) What are the mean and the variance of an average weight gain estimated from a sample of 20 heifers? (188,110)

   b) What are the mean and variance from a sample of 30 heifers? (188, 73.3)

   c) How many heifers should be in a sample in order for the variance of sample means to equal 50? (44)

2. Suppose the yield (lb/acre) of corn from 2 types of soil is studied for years and the means and variances are known to be

	mean	variance
Type A	4500	455,625
Type B	3900	152,100

   a) What is the mean and variance of the differences between the sample means from the two types of soil with 10 and 15 observations, respectively? What is the standard error of the differences? (600; 55, 702.5; 236.01)

   b) Calculate the mean and standard error of the difference assuming the above sample sizes were 40 and 60, respectively. (600, 118.01)

3. Assume that the mean weight of fried chicken dinners from a food store is 12 ounces and the standard deviation is 0.6 ounces.

   a) What is the probability that in a random sample of 5 of these dinners, the average weight will be less than 11.5? More than 11.9? Between 11.5 and 11.9? (0.0314, 0.6443, 0.3243)

   b) Answer the above question for a sample size of 10. (0.0043, 0.7019, 0.2938)

4. Assume that the heights of men are normally distributed with a standard deviation of 2.5 inches. How large a sample should be taken in order to be 90% certain that the sample mean will not differ from the population mean by more than 0.5 inches? (68)

5. In a manufacturing process it is known that 5% of the items are defective. What is the probability that a lot of 100 items will contain fewer than 3 defective items? (0.125)

6. It is known that only 60% of the seeds of a rare plant germinate. What is the probability that out of 100 seeds planted:

   a) 70 or more germinate: (0.0262)
   b) fewer than 45 germinate; (0.0008)
   c) more than 50 and fewer than 80 will germinate? (0.9838)

d) between 65 and 70 will germinate? (0.1052)

7. Suppose that in the past, 20% of all graduating students were in the College of Agricultural and Environmental Sciences at UCD. Among the 2,500 freshmen in 1979, what is your estimate of the probability that,

   a) There are less than 480 students who will graduate from the College of AES in 1983? (0.1515)

   b) Between 480 and 530 students will graduate from the College of AES in 1983? (0.6730)

   c) Suppose that the University administrators must be 95% certain that at least 750 students will graduate from the College of AES in 1983. How many freshmen should be admitted in 1979? (3748)

8. From Appendix Table A-5, plot the approximate probability curves of chi-square distributions with 3 and 9 degrees of freedom. What are the means and standard deviations of these two distributions? (3, 6), (9, 18)

9. Find the chi-square values that will cut off 50% of the right-hand tail probabilities for 1, 8, 18, and 100 degrees of freedom. Subtract these values from the mean and divide them by the standard deviation of their corresponding distributions. Explain how these values change.

10. A sample of 7 observations, 14, 12, 11, 10, 13, 15, 13, was taken from a population with a mean of 12. What is the probability of obtaining a mean from another sample of size 7 which is greater than the observed mean? $(0.2 < p < 0.25)$