11 Special Sampling Distributions

If a random sample of 25 persons was drawn from the population of college students and each person's athletic ability and intelligence were measured, a correlation between the two variables could be computed. Whatever the value of the correlation, a different value would have been obtained if a different sample of 25 persons had been drawn. Hence the correlation coefficient with a sample size of 25 varies from sample to sample. It therefore has a distribution of its own, called a sampling distribution.

When a statistic, such as a correlation coefficient, is computed, it is necessary to know its sampling distribution to be able to interpret it. If all statistics had radically different sampling distributions, data analysis would be an impossible task. Fortunately most statistics used in data analysis have one of four distributions. These four sampling distributions serve as important reference points in data analysis. Moreover, the sampling distributions of other statistics are closely approximated by these four distributions. That is, the statistics are not exactly distributed as one of the four, but one of the four can be used as a reasonable approximation. This chapter considers these four special sampling distributions.

The material presented here is quite abstract. It may be more useful to some students merely to skim the chapter now. Then, when the distributions are presented later, this chapter can be used as a reference.

The Standard Normal Distribution

The first sampling distribution is one that was presented in Chapter 10: the standard normal distribution. The Z or standard normal distribution is defined as follows: If X is a normally distributed variable with mean μ and variance σ^2, then

$$\frac{X - \mu}{\sigma}$$

is also normally distributed, with a mean of zero and variance of one. The standard normal distribution is a normal distribution with a mean of zero and a variance of one.

Many statistics, especially those involving means, have a standard normal distribution when the distribution of scores from which the means were computed is normal.

If X is normally distributed with mean μ and variance σ^2, the sample mean $\bar{X}$ is also normally distributed with mean μ and variance of σ^2/n, where n is sample size. The quantity $\sigma/\sqrt{n}$ is called the standard error of the mean. This is a very important fact:

$$\begin{array}{ccc} \text{standard error} & & \text{standard deviation divided by the} \\ \text{of the mean} & = & \text{square root of the sample size} \end{array}$$

This fact is only guaranteed if the observations are sampled randomly and independently. However, the formula for the standard error of the mean is true of any distribution, not just the normal.

In words, the variability of the mean equals the variability of the observations used to form the mean divided by the square root of the sample size. How far the sample mean is away from the population mean depends on the inherent variability in the population and the sample size. The larger the sample size, the closer on the average is the sample mean to the population mean. Consider separately the two special cases of n equal to one and n equal to infinity. If n is one, the mean is a single observation and its variability is σ^2 divided by one, which remains σ^2. If n is very large or infinite, σ^2/n equals zero. This implies that the sample mean is identical to the parameter μ when the sample size is quite large.

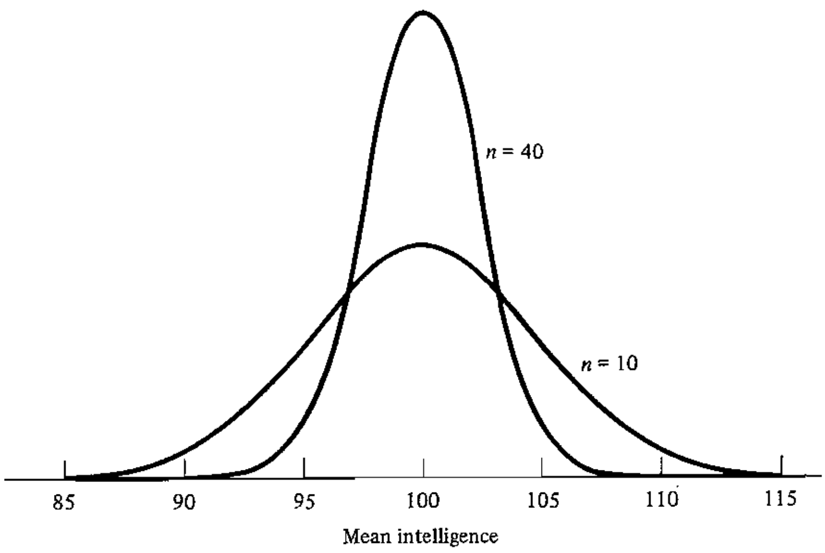
The relationship between sample size and the standard error of the mean can be examined graphically. Consider the variable IQ, which is assumed to be normally distributed, with a mean of 100 and a standard deviation of 15. The standard error of the mean is $15/\sqrt{n}$. In Figure 11.1 are two sampling distributions of $\bar{X}$ for sample sizes of 10 and 40. Note how much more variable the sample mean is when n is 10 and than when n is 40.

If the population mean is subtracted from the sample mean and if this difference is divided by its standard error, the following quantity is obtained.

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

If X is normally distributed, the above expression has a standard normal or Z distribution. Even if X does not have a normal distribution, $\bar{X}$ has approximately a normal distribution given the central limit theorem, discussed in the

FIGURE 11.1 **Sampling distribution of intelligence mean ($\mu = 100$, $\sigma^2 = 225$) for sample sizes of 10 and 40.**



previous chapter. And, the larger n is, the closer to normal is the distribution of $\bar{X}$.

Other quantities have standard normal distributions. Imagine that two samples are drawn from the same normally distributed variable with variance σ^2 and two sample means are computed. The two sample means are denoted as $\bar{X}_1$ and $\bar{X}_2$, and their sample sizes are n_1 and n_2, respectively. The quantity $\bar{X}_1 - \bar{X}_2$ is normally distributed with a mean of zero and a variance of

$$\sigma^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$$

given that the observations are normally distributed and independently and randomly sampled. It then follows that

$$\frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\dfrac{1}{n_1} + \dfrac{1}{n_2}}}$$

has a Z distribution.

Table 11.1 summarizes these facts about normally distributed variables. The quantities in the table have a normal distribution given that X has a normal distribution.

TABLE 11.1 **Sampling Distributions That Are Normal**

Variable	Mean	Standard Deviation
X	μ	σ
$\dfrac{X-\mu}{\sigma}$	0	1
$\bar{X}$	μ	$\dfrac{\sigma}{\sqrt{n}}$
$\bar{X}-\mu$	0	$\dfrac{\sigma}{\sqrt{n}}$
$\dfrac{\bar{X}-\mu}{\dfrac{\sigma}{\sqrt{n}}}$	0	1
$\bar{X}_1-\bar{X}_2$	0	$\sigma\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}$
$\dfrac{\bar{X}_1-\bar{X}_2}{\sigma\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$	0	1

t Distribution

The second sampling distribution that is commonly used in statistical analysis is the *t* distribution. Consider a normally distributed variable X. If a score is sampled from this population, and if the population mean is subtracted from this score, and then if this difference is divided by the population standard deviation, the resulting quantity is

$$\frac{X-\mu}{\sigma}$$

As was previously discussed, this quantity has a standard normal or Z distribution. But what would happen if the population standard deviation or σ is replaced with s, the sample estimate of the standard deviation? The quantity

$$\frac{X - \mu}{s}$$

does not have a Z distribution but rather has a *t* distribution.

The *t* distribution looks very much like the Z distribution. Like Z it has a mean of zero, is symmetric, and has bounds of plus and minus infinity. However, the *t* distribution is not as peaked as Z at zero, and so its tails are somewhat fatter than Z. These fatter tails make the variance of *t* greater than one. A *t* distribution looks like a bloated Z distribution.

Actually how closely the *t* distribution approaches the Z distribution depends on how closely *s* approaches σ. Recall that *t* differs from Z in that *s* is substituted for σ. Because *s* is a statistic and so it has sampling error, the quantity *s* does not equal σ. How close *s* is to σ depends on what are called the *degrees of freedom* used to estimate *s*. The degrees of freedom for *s* are usually $n - 1$. As the degrees of freedom get larger, the *t* distribution approaches Z. For very large degrees of freedom, *t* and Z are virtually indistinguishable. Although there is only one Z distribution, there are many *t* distributions. These different *t* distributions are denoted by *t*(*df*), where *df* stands for degrees of freedom.

Many quantities have a *t* distribution. In fact, all of the statistics in Table 11.1 that have a standard normal or Z distribution have a *t* distribution when the sample standard deviation is substituted for the population standard deviation.

So the quantities

$$\frac{X - \mu}{\frac{s}{\sqrt{n}}}$$

and

$$\frac{\bar{X}_1 - \bar{X}_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

have a *t* distribution. In each case the *t* distribution involves a statistic that has a normal distribution minus its population mean divided by its estimated standard error. The facts concerning the *t* distribution are summarized in Table 11.2.

The *t* distribution is useful for testing hypotheses about means, and these facts will be used in Chapters 12 and 13. Also, it will be seen in Chapter 16 that the *t* distribution is used to test hypotheses concerning correlation and regression coefficients.

TABLE 11.2 **Various Statistics That Have t, χ^2, and F Distributions[a]**

	t Distribution	
$\dfrac{X-\mu}{s}$	$\dfrac{X-\mu}{\dfrac{s}{\sqrt{n}}}$	$\dfrac{\bar{X}_1-\bar{X}_2}{s\sqrt{\dfrac{1}{n_1}+\dfrac{1}{n_2}}}$
	χ^2 Distribution	
Z^2	$\dfrac{(n-1)s^2}{\sigma^2}$	
	F Distribution	
$\dfrac{s_1^2}{s_2^2}$	$\dfrac{\chi_1^2/df_1}{\chi_2^2/df_2}$	

[a] The variable X has a normal distribution with a mean of μ and a variance of σ^2. The samples from which $\bar{X}$ and s^2 are computed are randomly and independently drawn.

Chi Square Distribution

Like t, the chi square distribution is closely related to Z. Consider the variable X, which has a normal distribution. To measure relative position a Z score can be computed. To measure how deviant or unusual a Z score is, it could be squared: Z_i^2. So if $\mu = 20$, $\sigma^2 = 100$, and $X_i = 22$, then $Z_i = (22 - 20)/10 = .2$ and $Z_i^2 = .04$. The variable Z^2 has a *chi square distribution*. The chi square distribution is symbolized by χ^2.

All chi square distributions have a positive skew and a lower limit of zero. The lower limit must be zero since Z^2 must be positive.

It is possible to take repeated random and independent samples from X. Many Z scores and Z^2 could be computed. The sum of k independent Z^2 values has a chi square distribution with k degrees of freedom. Thus, χ^2 is not one distribution, but rather a family of distributions that differ by their degrees of freedom which equal the number of Z's that are squared. The term k is called the degrees of freedom of the χ^2 distribution. The different chi square distributions are symbolized by $\chi^2(k)$.

A χ^2 distribution with k degrees of freedom has a mean of k and a variance of $2k$. So a χ^2 with 10 degrees of freedom has a mean of 10 and a variance of 20. The shape of the distribution of χ^2 is positively skewed with a lower bound of zero. As the degrees of freedom get larger, the skew becomes less pronounced and χ^2 with k degrees of freedom approaches a normal distribution with a mean of k and a variance of $2k$. So $(\chi^2 - k)/\sqrt{2k}$ has approximately a Z distribution, if k is appreciable, say greater than 20.

It can also be shown that

$$\frac{(n-1)s^2}{\sigma^2}$$

has a χ^2 distribution with $n - 1$ degrees of freedom. In words, the sample variance times $n - 1$ divided by the population variance has a chi square distribution. Facts about χ^2 are presented in Table 11.2.

The main use of the χ^2 distribution is testing models concerning frequency data. The chi square distribution is used for this purpose in Chapter 17. Chi square is also used to test hypotheses of differences between medians which is described in Chapter 18, and differences between correlations which is described in Chapter 16. Often when χ^2 is used, it is used to approximate a sampling distribution.

F *Distribution*

Again let X be a normally distributed variable. Two random, independent samples of X's of sizes n_1 and n_2 are chosen from the population. The variances, ${s_1}^2$ and ${s_2}^2$, are computed from each sample. If the ratio of ${s_1}^2/{s_2}^2$ is computed, the quantity would have an F distribution. Like χ^2, F is positively skewed with a lower bound of zero. (Because variances are always nonnegative, their ratio must be nonnegative.) Its peak comes near the value of one.

Like χ^2, F is actually a family of distributions. To determine which F distribution is being referred to, one needs to know the degrees of freedom of the numerator, ${s_1}^2$, and the degrees of freedom of the denominator, ${s_2}^2$. The number of degrees of freedom equals the denominator of the formula for the variance for each sample. So for an F, the degrees of freedom on the numerator and the degrees of freedom on the denominator must be determined. A given F distribution is denoted as $F(df_n, df_d)$ where df_n are the degrees of freedom on the numerator and df_d the degrees of freedom on the denominator.

The F distribution is closely related to χ^2. It can be shown that F is a ratio of two independent χ^2 variables, each divided by its degrees of freedom.

The main use of the F distribution is to test hypotheses about means. A procedure for doing so, called analysis of variance, is described in Chapters 14 and 15.

Relation Between Sampling Distributions

The four major sampling distributions, though distinct, are closely related. One aspect that ties the four together is the normal distribution. The Z distribution is itself normal. The χ^2 distribution is based on the sum of squared scores that are normally distributed. The t distribution also presumes that the numerator is normally distributed. Finally, the F distribution can be viewed as the ratio of two independent variances whose scores are normally distributed. Hence, the normal distribution is the starting point for all four of these distributions.

But the four distributions are more closely linked. In some cases their distributions are identical. There are four major equivalences between pairs of the major sampling distributions. First a χ^2 with one degree of freedom is identical to a Z^2 value. So the probabilities in Appendix C can be used to determine the probability of various χ^2 events. For instance, what is the probability of obtaining a χ^2 value with one degree of freedom larger than 1.0? The answer to this question lies in finding the probability of obtaining a value of Z greater than 1.0 or less than –1.0. The answer, using Appendix C, is $1 - (2)(.3413) = .3174$.

The second fact linking the distributions is that a t with an infinite number of degrees of freedom equals Z. This holds because if t has an infinite number of degrees of freedom its denominator becomes σ, and so t becomes Z.

The third fact is that a t with q degrees of freedom when squared equals an F with one degree of freedom on the numerator and q on the denominator. This fact is not so obvious. If a t^2

$$\frac{(\bar{X} - \mu)^2}{s^2/n}$$

is examined, both the numerator and the denominator estimate σ^2/n, and so both are estimates of the same population variance.

The final fact is that an F with q and infinite degrees of freedom is identical to a χ^2 distribution with q degrees of freedom which is divided by q. This is due the fact that F equals the ratio of two χ^2's divided by their degrees of freedom, and so

$$F(df_n, df_d) = \frac{\chi^2(df_n)/df_n}{\chi^2(df_d)/df_d}$$

A χ^2/df with an infinite degrees of freedom equals one. Substituting this fact into the denominator of the above equation, the result is

$$F(df_n, \infty) = \frac{\chi^2(df_n)}{df_n}$$

These facts are summarized in Table 11.3.

TABLE 11.3 **Equivalences Between the Four Major Sampling Distributions**

$$\begin{aligned} \chi^2(1) &= Z^2 \\ Z &= t(\infty) \\ t(q)^2 &= F(1, q) \\ F(q, \infty) &= \chi^2(q)/q \end{aligned}$$

Summary

Every statistic has a sampling distribution. There are four major sampling distributions. Many statistics' sampling distributions exactly or nearly exactly correspond to one of the four distributions. They are Z, t, χ^2, and F.

The Z or standard normal distribution is a normal distribution with a mean of zero and a variance of one. If the observations have a normal distribution, the sampling distribution of the sample mean also has a normal distribution. The standard deviation of the sampling distribution of the mean is the standard deviation of the observations divided by the square root of the sample size.

The t distribution is identical to Z, but the denominator is the sample standard deviation and not the population standard deviation. The t distribution looks like Z but it is less peaked and has fatter tails. Like Z, t has a mean of zero and is symmetrically distributed.

The χ^2 distribution with k degrees of freedom is a positively skewed distribution with a mean of k and a variance of $2k$. A χ^2 statistic can be viewed as the sum of k independent Z^2 values. The value k is called the degrees of freedom.

The F distribution is the ratio of two independently computed variances drawn from the same normally distributed population. Like χ^2, F is positively skewed with a lower limit of zero. The peak in the F distribution is near one.

These sampling distributions of Z, t, χ^2, and F are routinely used in testing statistical models. It is this topic of testing models that is presented in the next chapter.

Problems

1. Let X be a normally distributed variable with a mean of 40 and a standard deviation of 9. What is the distribution of the following statistics?

   a. $\bar{X}$ b. s^2
   c. $(\bar{X} - \mu)/s$ d. $(X - \mu)^2/\sigma^2$

2. If X is a variable with a mean of 20 and a variance of 49, determine the standard error of the mean for sample sizes of

   a. 100 b. 10 c. 1000 d. 50

3. If Y has a mean of 80 and a variance of 64, what would n have to be for the standard error of the mean to be 1.00?

4. Describe the distribution of the sample mean with a sample size of 49 if the numbers are drawn from a normal distribution with a mean of 10.0 and a variance of 64.

5. If X is normally distributed with a mean of 20 and a variance of 100, determine the probability that $\bar{X}$ is greater than 22 if

   a. $n = 25$ b. $n = 50$ c. $n = 100$ d. $n = 200$

6. Using the Z distribution determine the following probabilities.

   a. $\chi^2(1) > 1.44$ b. $\chi^2(1) > 4.00$ c. $t(\infty) > 1.00$

7. Compute the standard error of the mean for the following cases.

   a. $\sigma^2 = 100, \mu = 50, n = 25$
   b. $\sigma^2 = 9, \mu = 0, n = 16$
   c. $\sigma^2 = 25, \mu = -5, n = 64$
   d. $\sigma^2 = 81, \mu = 4, n = 100$

8. Using the facts in Table 11.3, show that $t(\infty)^2 = \chi^2(1)$.

9. Describe how χ^2 and F are similar and different in their shape. Consider the following.

   a. skew b. central tendency
   c. upper and lower limit d. variability