## Key Concepts

<table>
<thead>
<tr>
<th>Estimate, estimator</th>
<th>Maximum likelihood estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard error (of the mean)</td>
<td>Normal range, confidence limits or confidence interval</td>
</tr>
<tr>
<td>Unbiased, biased</td>
<td>(Student's) t-distribution, degrees of freedom</td>
</tr>
<tr>
<td>Minimum variance unbiased estimator, efficient estimator, robust estimator</td>
<td>Pooled variance</td>
</tr>
</tbody>
</table>
Estimates and Confidence Limits

SYMBOLS AND ABBREVIATIONS

d.f. degrees of freedom
s sample standard deviation (estimate)
S sample standard deviation (estimator)
s^2 sample variance (estimate)
S^2 sample variance (estimator)
s.e.m. standard error of the mean
t percentile of Student’s t-distribution or corresponding test statistic
x, y particular values of (random) variables
X, Y random variables
\bar{y} sample mean of y (estimate)
\bar{Y} sample mean of Y (estimator)
z particular value of a standardized normal random variable
Z standardized normal random variable
\pi population proportion
\mu population mean
\sigma population standard deviation
\sigma^2 population variance

ESTIMATES AND ESTIMATORS

It will be helpful at this point to distinguish between an estimate and an estimator. An estimator is a rule that tells us how to determine from any sample a numerical value to estimate a certain population parameter, whereas an estimate is the actual numerical value obtained from a particular sample. Suppose we select a random sample of 10 students from a class of 200, measure their heights, and find the sample
mean. If we consider the ‘sample mean’ to be the rule that tells us to add up the 10 heights and divide by 10, it is an estimator, and we shall denote it $\bar{Y}$. If, on the other hand, we consider the ‘sample mean’ to be the result we obtain from a particular sample – say, 70 inches – it is an estimate and we shall denote it $\bar{y}$. The estimator $\bar{Y}$ is a random variable that takes on different values from sample to sample. The estimate $\bar{y}$, on the other hand, is the numerical value obtained from one particular sample. We say that the sample mean $\bar{Y}$ is an estimator, and $\bar{y}$ an estimate, of the population mean $\mu$. Thus, particular values of an estimator are estimates. Different samples might yield $\bar{y} = 69$ inches or $\bar{y} = 72$ inches as estimates of the mean of the entire class of 200 students, but the estimator $\bar{Y}$ is always the same: add up the 10 heights and divide by 10.

Note that we have retained our convention of using capital, or uppercase, letters to denote random variables (estimators) and lowercase letters to denote specific values of random variables (estimates). The word ‘estimate’ is often used in both senses, and there is no harm in this provided you understand the difference. To help you appreciate the difference, however, we shall be careful to use the two different words and symbols, as appropriate.

Because our set of study units is not expected to be exactly the same from sample to sample, an estimator, which gives us the rule for calculating an estimate from a sample, is itself a random variable – the estimates vary from sample to sample. As a result, an estimator has a distribution and it is of interest to determine its variance and standard deviation (and also the shape and other general characteristics of its distribution). By selecting many different samples and obtaining the corresponding estimate for each sample, we obtain a set of data (estimates) from which we can compute a variance and standard deviation in the usual way. The problem with this approach is that it requires that many samples be studied and is therefore not practical. Fortunately, there is a way of estimating the variance and standard deviation of an estimator from the information available in a single sample. If the population variance of a random variable is $\sigma^2$, it can be proved mathematically that the sample mean $\bar{Y}$ of $n$ independent observations has variance equal to $\sigma^2/n$ and standard deviation equal to $\sigma/\sqrt{n}$. We can estimate these quantities by substituting $s$ for $\sigma$, where $s$ is computed from a single sample in the usual way. We see immediately from the formula $\sigma^2/n$ that the larger the sample size ($n$), the smaller are the variance and standard deviation of the sample mean. In other words, the sample mean calculated from large samples varies less, from sample to sample, than the sample mean calculated from small samples.

It is usual practice to call the standard deviation of an estimator the standard error (often abbreviated s.e.) of that estimator. The standard deviation of the sample mean ($\sigma/\sqrt{n}$) is, therefore, often called the standard error of the mean. When applied to the sample mean, there is no real difference between the terms ‘standard
deviation’ and ‘standard error’, but the latter is often used when we are referring to variability due to error, as opposed to natural variability. The fact that men have variable heights is in no way due to ‘error’, but the fact that our estimate of mean height differs from sample to sample and almost all the specific estimates differ from the true mean can be considered error – sampling error. Furthermore, from now on in this book, as in so much of the literature, we shall use the term ‘standard error’ to indicate the estimated standard deviation of the estimator. Thus, the standard error of the mean (often abbreviated s.e.m.) is $s/\sqrt{n}$.

**NOTATION FOR POPULATION PARAMETERS, SAMPLE ESTIMATES, AND SAMPLE ESTIMATORS**

We have already indicated that we denote population parameters by Greek letters and sample statistics by Latin letters. The former are always fixed constants, whereas the latter, in the context of repeated sampling, can be considered random variables. The particular estimates calculated from one sample (such as 69 inches or 72 inches in the above example) are fixed constant quantities, but they can be viewed as particular outcomes of random variables in the context of examining many, many samples. Depending on how they are viewed, we use lowercase or uppercase Latin letters. Assuming a sample of size $n$, our fundamental notation is as follows:

<table>
<thead>
<tr>
<th>Name</th>
<th>Parameter</th>
<th>Estimate</th>
<th>Estimator</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>$\mu$, $\sigma^2$, $\sigma_\bar{Y}$</td>
<td>$\bar{y}$, $s^2$, $s_\bar{Y}$</td>
<td>$\bar{Y}$, $S^2$, $S_\bar{Y}$</td>
</tr>
<tr>
<td>Variance</td>
<td>$\sigma$, $\sigma_Y$</td>
<td>$s$, $s_Y$</td>
<td>$S$, $S_Y$</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>$\sigma = \sigma_Y/\sqrt{n}$</td>
<td>$s_\bar{Y} = s_Y/\sqrt{n}$</td>
<td>$S_\bar{Y} = S_Y/\sqrt{n}$</td>
</tr>
<tr>
<td>Standard deviation of the mean</td>
<td>$\sigma_\bar{Y} = \sigma_Y/\sqrt{n}$</td>
<td>$s_\bar{Y} = s_Y/\sqrt{n}$</td>
<td>$S_\bar{Y} = S_Y/\sqrt{n}$</td>
</tr>
</tbody>
</table>

Thus $\mu$ is the population mean (parameter), $\bar{y}$ the sample mean (estimate). Similarly, $\sigma$ is the population standard deviation, $s$ the sample standard deviation; in other words, $\sigma$ is the standard deviation of a random variable $Y$, and $s$ is the estimate of $\sigma$ we calculate from a sample. When we want to stress the fact that the relevant random variable is $Y$, we write $\sigma_Y$ or $s_Y$ – the standard deviation of $Y$. Analogously, $\sigma_\bar{Y}$ is the standard deviation of $\bar{Y}$, the standard deviation of the sample mean of $n$ observations, while $s_\bar{Y}$ is the sample estimate of this quantity, which, as noted earlier, is often called the standard error of the mean. Thus $\sigma_\bar{Y} = \sigma_Y/\sqrt{n}$ and $s_\bar{Y} = s_Y/\sqrt{n}$. 
PROPERTIES OF ESTIMATORS

We usually estimate the population mean by the sample mean; and we often estimate the population variance by the sample variance. Why do we do this? There are obviously many possible estimators to choose from so why not, for example, estimate the population mean by the average of the smallest and largest values in our sample? The answer is that we choose estimators, and hence estimates, that have certain ‘good’ properties. As discussed earlier, an estimator is a random variable and has a distribution. The characteristics of this distribution determine the goodness of the estimator. In the previous example, suppose we take all possible samples of 10 students from the class of 200 and compute the mean height from each sample. These sample means vary in the sense that they have different values from sample to sample, and if we average them, we obtain a ‘mean of means’, which is equal to the population mean – the mean height of all 200 students. In a case such as this, in which the average of the sample estimates for all possible samples equals the value of the parameter being estimated, we say the estimator is unbiased: the mean of the estimator’s distribution is equal to the parameter being estimated. When this does not occur, the estimator is biased. In general, when we have a random sample, the sample mean is an unbiased estimator of the population mean and the sample variance is an unbiased estimator of the population variance; but the sample standard deviation is not an unbiased estimator of the population standard deviation.

Recall that we use $n - 1$ rather than $n$ as a divisor when we average the sum of squared deviations from the sample mean in computing the sample variance. If we used $n$ as a divisor, we would find that the average of all possible estimates is equal to $(n - 1)/n$ times the population variance and it follows that the estimator that uses $n$ as a divisor is biased; such an estimator leads to an estimate that is, on an average, smaller than the population value. With $n - 1$ as a divisor, we have an unbiased estimator of the variance.

We should like our estimate to be close to the parameter being estimated as often as possible. It is not very helpful for an estimate to be correct ‘on average’ if it fluctuates widely from sample to sample. Thus, for an estimator to be good, its distribution should be concentrated fairly closely about the true value of the parameter of interest; in other words, if an estimator is unbiased the variance of the estimator’s distribution should be small. If we have a choice among several estimators that are competing, so to speak, for the job of estimating a parameter, we might proceed by eliminating any that are biased and then, from among those that are unbiased, choose the one with the smallest variance. Such an estimator is called a minimum variance unbiased estimator, or an efficient estimator. It can be shown mathematically that if the underlying population is normally distributed, the sample mean and the sample variance are minimum variance unbiased estimators of
the population mean and the population variance, respectively. In other situations, however, the sample descriptive statistic may not have this property. The sample mean is a minimum variance unbiased estimator of the population mean when the underlying population is binomially or Poisson-distributed, but in these two cases an analogous statement cannot be made about the sample variance. In fact, the efficient estimator of the variance of a Poisson distribution is provided by the sample mean!

We have discussed only two important properties of good estimators. Other properties are also of interest, but the important thing is to realize that criteria exist for evaluating estimators. One estimator may be preferred in one situation and a second, competing estimator may be preferred in another situation. There is a tendency to use estimators whose good properties depend on the data following a normal distribution, even though the data are not normally distributed. This is unfortunate because more appropriate methods of analysis are available. You cannot hope to learn all of the considerations that must be made in choosing estimators that are appropriate for each situation that might arise, but you should be aware of the necessity of examining such issues.

An estimator that usually has good properties, whatever the situation, is called robust. It is to our advantage if we can find estimators that are robust. The sample mean is a robust estimator of the population mean.

**MAXIMUM LIKELIHOOD**

Having accepted the fact that estimators should have certain desirable properties, how do we find an estimator with such properties in the first place? Again there are many ways of doing this, and no single approach is preferable in every situation. A full discussion of these methods requires mathematical details that are beyond the scope of this book, but we must describe two very important approaches to deriving estimators because they are frequently mentioned in the literature. One is based on a method known as maximum likelihood estimation, and the other on a method called least-squares estimation. The method of least squares will be discussed in Chapter 10. Maximum likelihood estimates of parameters such as the population mean and variance are those values of the parameters that make the probability of, or likelihood for, our sample as large as possible – a maximum. Intuitively, they are those parameter values that make the data we observe ‘most likely’ to occur in the sample.

A simple example will illustrate the principle. Suppose we wish to estimate the proportion of a population that is male. We take a random sample of $n$ persons from the population and observe in the sample $y$ males and $n - y$ females. Suppose now that the true population proportion is $\pi$, so that each person in the sample has
a probability $\pi$ of being male. What then is the probability of, or likelihood for, our sample which contains $y$ males and $n - y$ females? We learned in Chapter 5 that the probability of this happening is given by the binomial distribution with parameters $n$ and $\pi$, that is,

$$P(y \text{ males}) = \frac{n!}{y!(n-y)!} \pi^y (1 - \pi)^{n-y}.$$

Fixing $y$, which is known once the sample has been taken, and considering $\pi$ to be a variable, we now ask: What value of $\pi$ makes this probability a maximum? We shall call that value the maximum likelihood estimate of $\pi$. We shall not prove it here, but it can be shown that in this case the likelihood is largest when we set $\pi$ equal to $y/n$, the sample proportion of male persons. (It is instructive to take the time to verify this numerically. Suppose, for example, that $n = 5$ and $y = 2$. Calculate $P(y \text{ male})$ for various values of $\pi$ (e.g. $\pi = 0.2, 0.3, 0.4, 0.5$). You will find that it is largest when $\pi = 0.4$.) Thus, the maximum likelihood estimate of $\pi$ in this example is $y/n$. Analogously, the maximum likelihood estimator of $\pi$ is $Y/n$, where $Y$ is the random variable denoting the number of male persons we find in samples of size $n$ from the population.

Except as noted in the Appendix, maximum likelihood estimators have the following important properties for samples comprising a very large number of study units; that is, as the number of study units tends to infinity, maximum likelihood estimators have the following (so-called asymptotic) properties:

1. They are unbiased: the mean of the estimator (i.e. the mean of many, many estimates) will equal (or at least be arbitrarily close to – very, very close to) the true value of the parameter being estimated.
2. They are efficient: the variance of the estimator (i.e. the variance of many, many estimates) is the smallest possible for any asymptotically unbiased estimator.
3. The estimators are normally distributed. The utility of this last property will become apparent later in this chapter.

In certain instances, maximum likelihood estimators have some of these properties for all sample sizes $n$; in general, however, they have these properties only for very large sample sizes. How large $n$ must be for these properties to be enjoyed, at least approximately, depends on the particular situation, and is often unknown. Other problems can occur with maximum likelihood estimators: we must know the mathematical formula for the distribution, they may be difficult to compute, they may not exist and, if they do exist, they may not be unique. Nevertheless, the principle of maximum likelihood estimation has a great deal of intuitive appeal and it is widely used.
ESTIMATING INTERVALS

In the literature, estimates are often given in the form of a number plus or minus another number. For example, the mean serum cholesterol level for a group of 100 persons might be reported as 186 ± 32 mg/dl. Unfortunately, these numbers are often reported without any explanation of what the investigator is attempting to estimate. It is a standard convention for the first number (186) to be the parameter estimate, but there is no standard convention regarding what the number after the ± sign represents. In some instances, the second number is a simple multiple (usually 1 or 2) of the estimated standard deviation of the random variable of interest. In other instances, it may be some multiple of the standard error of the estimator. In our example, 186 ± 32 mg/dl represents the mean serum cholesterol level plus or minus one standard deviation, estimated from a group of 100 persons. If we divide 32 mg/dl by the square root of the sample size (i.e. by 10), we obtain the standard error of the mean, equal to 3.2 mg/dl. Thus the interval 186 ± 3.2 mg/dl represents the mean serum cholesterol level plus or minus one standard error of the mean. Clearly it is important to know which is being quoted, and to know whether the number after the ± sign is once, twice or some other multiple of the estimated standard deviation or standard error of the mean. We now turn to a discussion of how the different intervals should be interpreted.

We shall see later in this chapter that the standard error of the mean is used to define an interval that we believe contains the true value of the mean. For now, suppose that we wish to estimate an interval that includes most of the population values. Suppose that our population of cholesterol values is normally distributed, with mean \( \mu \) and standard deviation \( \sigma \). Then we know that approximately two thirds of the population lies between \( \mu - \sigma \) and \( \mu + \sigma \) – that is, in the interval \( \mu \pm \sigma \). We can estimate this interval by \( \bar{y} \pm s \) (i.e., 186 ± 32 mg/dl). Thus, provided cholesterol levels are normally distributed and our sample size is large enough for the estimates to be close to the true parameters, we can expect about two thirds of the population to have cholesterol levels between 154 and 218 mg/dl. Similarly, we know that about 95% of the population lies in the interval \( \mu \pm 2\sigma \), which for our example is estimated by 186 ± 64. Thus, we might estimate that 95% of the population cholesterol levels are between 122 and 250 mg/dl.

Such intervals, or limits, are often quoted in the literature in an attempt to define a normal range for some quantity. They must, however, be very cautiously interpreted on two counts. First, the assumption of normality is not a trivial one. Suppose it was quoted, on the basis of the data in Table 3.1, that \( \bar{y} \pm 2s \) for triglyceride levels in male students is 111 ± 118 mg/dl (\( \bar{y} = 111, s = 59 \)). Should we expect 95% of the population to be in this interval? Note that this interval is from −7 to 229 mg/dl. Triglyceride levels are not at all normally distributed, but follow a very positively skewed distribution in the population (see Figures 3.2 and 3.4);
this explains why $\bar{y} - 2s$ is negative and hence an impossible quantity. We must not assume that $\mu \pm 2\sigma$ includes about 95% of the population for a non-normally distributed random variable. Second, there is an implicit assumption (in addition to normality) that the sample estimates are close to the true parameter values. It is rarely realized that it takes very large sample sizes for this assumption to be adequately met. If the sample size is less than 50, then $\bar{y} \pm 2s$ includes, on an average, less than 95% of a normally distributed population. When the sample size is 10, it includes on an average only 92%, and when the sample size is down to 8, it includes on an average less than 90%. Although these consequences might be considered small enough to neglect, it must be realized that these percentages are only averages. The results from any one sample could well include less than 75% of the population, as can be seen from Table 6.1. We see from this table, for example, that when the sample size is 10, there is a 0.12 probability that $\bar{y} \pm 2s$ includes less than 75% of the population. Even when the sample size is 100, there is still a 0.06 probability that $\bar{y} \pm 2s$ includes less than 75% of the population. Thus, if cholesterol levels are normally distributed, the interval 122 to 250 mg/dl has a small but non-negligible probability (0.06) of including less than 75% of the population.

Table 6.1  Approximate proportion of samples from a normal distribution in which the estimated mean $\pm 2s.d.$ will fail to include the indicated percentage of the population

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>75</th>
<th>90</th>
<th>95</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.12</td>
<td>0.18</td>
<td>0.24</td>
</tr>
<tr>
<td>20</td>
<td>0.09</td>
<td>0.12</td>
<td>0.15</td>
</tr>
<tr>
<td>30</td>
<td>0.08</td>
<td>0.10</td>
<td>0.12</td>
</tr>
<tr>
<td>40</td>
<td>0.07</td>
<td>0.09</td>
<td>0.11</td>
</tr>
<tr>
<td>50</td>
<td>0.07</td>
<td>0.09</td>
<td>0.11</td>
</tr>
<tr>
<td>100</td>
<td>0.06</td>
<td>0.07</td>
<td>0.08</td>
</tr>
</tbody>
</table>

DISTRIBUTION OF THE SAMPLE MEAN

We have seen that if $Y$ is normally distributed, we can write, approximately,

$$P(\mu - 2\sigma \leq Y \leq \mu + 2\sigma) = 0.95$$

and

$$P\left(-2 \leq \frac{Y - \mu}{\sigma} \leq 2\right) = 0.95.$$
The 97.5th percentile of the standard normal distribution is about 2, and the 2.5th percentile is about \(-2\). Now consider \(Y\), the mean of \(n\) such normally distributed random variables. We have already learned that \(Y\) has mean \(\mu\) and standard deviation \(\sigma/\sqrt{n}\). Furthermore, it is normally distributed. As an example, suppose we have a population in which the mean height is 67 inches and the standard deviation is 3 inches. We take samples of four persons each and average their heights. Then the distribution of these averages is normal with mean 67 inches and standard deviation \(3/\sqrt{4}\) inches = 1.5 inches. This is pictured in Figure 6.1. Notice that whereas about 95% of the population lies in the interval 61 to 73 inches, about 95% of the means lie in the interval 64 to 70 inches.

![Figure 6.1](image)

**Figure 6.1** Normal density function of a person’s height, \(Y\), with mean 67 inches and standard deviation 3 inches, and of the average of a sample of four persons’ heights, \(\bar{Y}\), with the same mean 67 inches but standard deviation 1.5 inches.

Now let us subtract the mean from \(\bar{Y}\) and divide the difference by its standard deviation, that is,

\[
Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}.
\]

We denote the result \(Z\), because that is the letter usually used for a standardized normal random variable; with about 95% probability it lies between \(-2\) and \(+2\). Consequently, we can write

\[
P(-2 \leq Z \leq 2) = 0.95
\]
or

\[
P \left( -2 \leq \frac{\bar{Y} - \mu}{\sigma_T} \leq 2 \right) = 0.95,
\]

which is equivalent to

\[
P \left( \bar{Y} - 2\sigma_T \leq \mu \leq \bar{Y} + 2\sigma_T \right) = 0.95.
\]

(The derivation of this equivalence is given in the Appendix.) Remember, \( \sigma_T = \sigma_Y / \sqrt{n} \) for a sample of size \( n \).

Let us now summarize the various facts we have learned about the sample mean \( \bar{Y} \):

1. The mean of \( \bar{Y} \) is \( \mu \).
2. The variance of \( \bar{Y} \) is \( \sigma^2 / n \) (i.e. \( \sigma_{\bar{Y}}^2 = \sigma_Y^2 / n \)).
3. The standard deviation of \( \bar{Y} \) is \( \sigma / \sqrt{n} \) (i.e. \( \sigma_{\bar{Y}} = \sigma_Y / \sqrt{n} \)).
4. \( \bar{Y} \) is normally distributed. This is strictly true only if \( Y \) is normally distributed. It is a remarkable fact, however, that it also tends to be true, for moderate to very large sample sizes, almost irrespective of how \( Y \) is distributed (the word ‘almost’ is added to allow for some special situations that, although mathematically possible, do not usually occur in practice). Usually, a mean of five or more observations is for all intents and purposes normally distributed.
5. It therefore follows, provided \( \bar{Y} \) is based on five or more observations, that

\[
P \left( \bar{Y} - \frac{2\sigma_T}{\sqrt{n}} \leq \mu \leq \bar{Y} + \frac{2\sigma_T}{\sqrt{n}} \right) = 0.95.
\]

**CONFIDENCE LIMITS**

Consider once more the triglyceride levels in Table 3.1, for which \( \bar{Y} = 111 \text{ mg/dl}, s = 59 \text{ mg/dl}, \) and \( n = 30 \). The standard error of the mean is \( \frac{59}{\sqrt{30}} = 11 \text{ mg/dl} \).

Let us assume for the moment that the true population values are \( \sigma_Y = 59 \) and \( \sigma_T = 11 \text{ mg/dl} \). (We shall see how to avoid this assumption later.) The interval \( \bar{Y} \pm 2\sigma_T \) is \( 111 \pm 22 \), or 89 to 133 mg/dl. Can we therefore say that there is about a 95% probability that \( \mu \) lies in this interval? In other words, can we write

\[
P \left( 89 \leq \mu \leq 133 \right) = 0.95?
\]

Regardless of how good the approximation might be, with the definitions we have given such an expression is impossible. The true mean \( \mu \) is a fixed quantity, not a
random variable, and the fact that we do not know it does not alter this. Similarly, 89 and 133 are fixed quantities. Either \( \mu \) lies between 89 and 133 mg/dl, or it does not. We cannot talk about the ‘probability’ of this being the case, since probability is a property associated with random variables. Thus, it is meaningful to write

\[
P(\overline{Y} - 2\sigma_Y \leq \mu \leq \overline{Y} + 2\sigma_Y) = 0.95
\]

but not to write

\[
P(\overline{y} - 2\sigma_Y \leq \mu \leq \overline{y} + 2\sigma_Y) = 0.95.
\]

Now you can see why we have taken pains to distinguish between \( \overline{Y} \) and \( \overline{y} \). The former is a random variable, about which we can make a probability statement; the latter is a fixed quantity, about which we cannot make a probability statement.

Despite the fact that we cannot talk about the probability of \( \mu \) lying between 89 and 133 mg/dl, this interval is clearly somehow related to the unknown mean \( \mu \), and we should be reasonably certain (perhaps ‘95% certain’) that \( \mu \) is in fact between 89 and 133 mg/dl. We call these numbers the 95% confidence limits, or confidence interval, for \( \mu \), and we say that \( \mu \) lies between 89 and 133 mg/dl with 95% confidence.

Thus \( \overline{y} \pm 2\sigma_Y \) (i.e. \( \overline{y} \pm 2\sigma_Y / \sqrt{n} \)) gives an approximate 95% confidence interval for the mean. This is strictly true only if \( Y \) is normally distributed, but tends to be true, for samples that are at least moderately large, whatever the distribution of \( Y \). A confidence interval is to be interpreted as follows: if we were to find many such intervals, each from a different sample but in exactly the same fashion, then, although the intervals may be different and bounce around randomly, in the long run about 95% of our intervals in fact would include the true mean and 5% would not. We cannot say that there is a 95% probability that the true mean lies between the two values we obtain from a particular sample, but we can say that we have 95% confidence that it does so. We know that if we estimate \( \mu \) by the mean \( \overline{y} \) of a single sample, we cannot expect to be so lucky as to have \( \overline{y} = \mu \). With a single number as our estimate, we have no feel for how far off we might be with our estimate. By using a confidence interval to estimate \( \mu \), we have a range of values that we think, with some degree of confidence, contains the true value \( \mu \). We shall now briefly indicate how confidence limits are calculated in several specific situations.

### CONFIDENCE LIMITS FOR A PROPORTION

Suppose we wish to estimate the probability a newborn is male. We take a random sample of 1000 births and find 526 are male. We therefore estimate the proportion of births that are male as being 0.526. What would the 95% confidence limits be? Since 0.526 is a maximum likelihood estimate based on a large sample, it can be
considered as the outcome of a normally distributed random variable. Now if $Y$ follows a binomial distribution with parameters $n$ and $\pi$, then it can be shown that the variance of $Y/n$ ($y/n$ is 526/1000 in our particular sample) is $\pi(1 - \pi)/n$. Thus $Y/n$ is approximately normally distributed with mean $\pi$ and standard deviation $\sqrt{\pi(1 - \pi)/n}$, and so an approximate 95% confidence interval for $\pi$ is

$$\frac{y}{n} \pm 2\sqrt{\frac{\pi(1 - \pi)}{n}}$$

which we estimate by

$$0.526 \pm 2\sqrt{\frac{0.526(1 - 0.526)}{1000}} = 0.526 \pm 0.032.$$  

Thus, we have about 95% confidence that the true proportion lies between 0.494 and 0.558.

Notice that to calculate this interval we substituted our estimate, 0.526, for $\pi$. For a large sample this is adequate. Had the sample been small, we should not assume $Y/n$ is normally distributed, and it would have been necessary to use special tables that have been calculated from the binomial distribution. As a rule of thumb, the approximation is adequate provided both $n\pi$ and $n(1 - \pi)$ are greater than 5. Note also that because the estimator is approximately normally distributed, in large samples a 95% confidence interval can always be obtained from a maximum likelihood estimate by adding and subtracting twice the standard error of the estimator.

### CONFIDENCE LIMITS FOR A MEAN

Consider our example in which we determined 95% confidence limits for mean triglyceride level as being 89 to 133 mg/dl. Recall that we assumed we knew $\sigma_Y = 59$ and $\sigma_T = 11$ mg/dl, whereas in fact these were really the sample estimates $s_Y$ and $s_T$, respectively. Usually we do not know the true standard deviation, $\sigma_Y$, so to be of any practical use we must be able to calculate confidence limits without it. Does it make any difference if we simply substitute $s_T$ for $\sigma_T$? The answer is that it does not, for all practical purposes, if the sample size is more than 30. When the sample size is smaller than this, we must allow for the fact that we do not know $\sigma_T$, as follows.

We have seen that if $Y$ is normally distributed, then $(\overline{Y} - \mu)/\sigma_T$ follows a standard normal distribution, for which the approximate 2.5th and 97.5th percentiles are, respectively, $-2$ and $+2$. Analogously, substituting $S_T$ for $\sigma_T$, $(\overline{Y} - \mu)/S_T$ follows a distribution called *Student's t-distribution* with $n - 1$ degrees of freedom. We mentioned earlier that we may be interested in knowing characteristics of specific
distributions other than just the mean and variance of the relevant random variable. Knowing the shape of Student’s t-distribution is one such example. Like the standard normal distribution, the t-distribution is symmetric about zero, so that the 2.5th percentile is simply the negative of the 97.5th percentile. Denote the 97.5th percentile \( t_{97.5} \). Then the 95% confidence limits are \( \bar{y} \pm t_{97.5} s_Y \).

The distribution of \( (\bar{Y} - \mu)/s_Y \) was derived by a mathematician who worked for the Guinness Brewing Company in Ireland and who published his statistical papers under the pseudonym ‘Student’. This quantity is denoted \( t \), hence the name ‘(Student’s) t-distribution.’ Just as the normal distribution is a family of distributions, a particular one being determined by the parameters \( \mu \) and \( \sigma \), so is the t-distribution a family of distributions — but in this case the particular distribution is determined by a parameter known as the ‘number of degrees of freedom’, a concept we shall explain shortly. Each t-distribution is similar to the standard normal distribution but has thicker tails, as illustrated in Figure 6.2. The fewer the degrees of freedom, the thicker the tails are. As the number of degrees of freedom becomes indefinitely large (in which case \( \sigma = s \)), the t-distribution becomes the same as the standard normal distribution. This can be seen in the following 97.5th percentiles, abstracted from a table of the t-distribution:

<table>
<thead>
<tr>
<th>Degrees of Freedom</th>
<th>12.706</th>
<th>2.228</th>
<th>2.042</th>
<th>1.960</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(standard normal)</td>
<td></td>
<td></td>
<td>1.960</td>
<td></td>
</tr>
</tbody>
</table>

Now you can see the basis for saying that for all practical purposes it makes no difference if we substitute \( s_Y \) for \( \sigma_Y \) when the sample size is more than 30: 2 is almost as close an approximation to 2.042 (\( t_{97.5} \) when there are 30 degrees of freedom, corresponding to \( n = 31 \)) as it is to 1.960 (the 97.5th percentile of the standard normal distribution).

![Figure 6.2](image)

**Figure 6.2** Example of the density function of Student’s t-distribution compared to that of the standard normal distribution.
Now let us go back once more to our example of triglyceride levels (from the data in Table 3.1), for which $\bar{y} = 111$ mg/dl, $s_Y = 59$ mg/dl, and $n = 30$. Calculating with more accuracy than previously, we find $s_Y = 59/\sqrt{30} = 10.77$ mg/dl, and from a table of the $t$-distribution we find that, for $n - 1 = 29$ degrees of freedom, $t_{97.5} = 2.045$. From this we find that $\bar{y} \pm t_{97.5}s_Y$ is $111 \pm 2.045 \times 10.77 = 111 + 22.0$ mg/dl. Clearly in this case our earlier approximation was adequate. Had the sample size been much smaller (e.g. 10 or less), however, the approximation $\bar{y} \pm 2s_Y$ would have led to an interval that is much too short.

**THE CONCEPT OF DEGREES OF FREEDOM**

The term *degrees of freedom*, abbreviated d.f., will occur again and again, not only in connection with the $t$-distribution, but also in many other problems. Basically, the degrees of freedom refer to the number of ‘independent’ observations in a quantity. We give the following simple examples to illustrate the degrees of freedom concept. You know that the sum of the angles in a triangle is equal to $180^\circ$. Suppose you were asked to choose the three angles of a triangle. You have only two degrees of freedom in the sense that you may choose two of the angles, but then the other is automatically determined because of the restriction that the sum of the three is $180^\circ$. Suppose you are asked to choose three numbers with no restrictions on them. You have complete freedom of choice in specifying all three numbers and hence, in that case, you have three degrees of freedom.

Now suppose you are asked to choose six numbers (which we shall call $y_1, y_2, y_3, y_4, y_5, \text{ and } y_6$) such that the sum of the first two is 16 and also such that the sum of all of them is 40. There are six numbers to be specified, but you do not have freedom of choice for all six. You have to take into account the restrictions

$$y_1 + y_2 = 16$$

and

$$y_1 + y_2 + y_3 + y_4 + y_5 + y_6 = 40.$$

As soon as you select $y_1$, then $y_2 = 16 - y_1$, and so $y_2$ is completely determined. Of the remaining numbers, $y_3 + y_4 + y_5 + y_6 = 40 - 16 = 24$. Thus, only three of the numbers $y_3, y_4, y_5, y_6$ can be freely chosen. If we choose $y_3, y_4, \text{ and } y_5$, for example, $y_6$ is predetermined as follows:

$$y_6 = 24 - (y_3 + y_4 + y_5)$$

Hence, the total number of degrees of freedom in this example is $1 + 3 = 4$. 
In computing a variance, we use as the divisor a number that makes the variance an unbiased estimator of the population variance. This divisor is the number of degrees of freedom associated with the estimator once an estimate of the mean has been made. Recall that our divisor for the variance is the size of the sample minus one (i.e. \( n - 1 \)). Once the mean is fixed at its sample value, there are only \( n - 1 \) degrees of freedom associated with permissible values of the numbers used to compute the variance. This same number is also the number of degrees of freedom associated with the estimated standard deviation, and with the \( t \)-distribution used to obtain confidence limits for the mean from the formula \( \bar{y} \pm ts_\bar{y} \).

Other types of problems with a variety of restrictions and degrees of freedom are considered in subsequent chapters of this book. In every instance the number of degrees of freedom is associated with a particular statistic (such as \( s_\bar{y} \)). It is also the appropriate value to use as the parameter of a distribution (such as the \( t \)-distribution) when using that statistic for a particular purpose (such as calculating confidence limits).

**CONFIDENCE LIMITS FOR THE DIFFERENCE BETWEEN TWO MEANS**

Suppose we compare two drugs, A and B, each aimed at lowering serum cholesterol levels. Drug A is administered to one group of patients (sample 1) and drug B to a second group (sample 2), with the patients randomly assigned to the two groups so that the samples are independent. If we use the estimators \( \bar{Y}_1 \) and \( \bar{Y}_2 \) to find estimates \( \bar{y}_1 \) and \( \bar{y}_2 \) of the post-treatment serum cholesterol means for drug A and drug B, respectively, we might want to construct a confidence interval for the true difference \( \mu_1 - \mu_2 \). In this situation a 95% confidence interval would be given by

\[
\bar{y}_1 - \bar{y}_2 \pm t_{97.5} s_{\bar{Y}_1 - \bar{Y}_2},
\]

where \( t_{97.5} \) is the 97.5th percentile of the \( t \)-distribution, with degrees of freedom equal to that associated with \( s_{\bar{Y}_1 - \bar{Y}_2} \), the standard error of \( \bar{Y}_1 - \bar{Y}_2 \). If the two samples have the same true variance \( \sigma^2 \), then the respective sample variances \( s_1^2 \) and \( s_2^2 \) are both estimates of the same quantity \( \sigma^2 \). In such instances we can average or ‘pool’ the sample variances to obtain a pooled estimate \( s_p^2 \) of \( \sigma^2 \):

\[
s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.
\]

Note that when we pool \( s_1^2 \) and \( s_2^2 \), we weight each by the number of degrees of freedom associated with it. Note also that we obtain a (weighted) average of sample variances, not of standard deviations. We take the square root of this pooled variance to obtain the (sample) pooled standard deviation \( s_p \). Next, you need to know that
when we have two independent random variables, the variance of their difference is equal to the sum of their variances. Thus the variance of $\bar{Y}_1 - \bar{Y}_2$ is the variance of $\bar{Y}_1$ plus the variance of $\bar{Y}_2$, that is,

$$\sigma^2_{\bar{Y}_1 - \bar{Y}_2} = \frac{\sigma^2}{n_1} + \frac{\sigma^2}{n_2} = \sigma^2 \left( \frac{1}{n_1} + \frac{1}{n_2} \right),$$

and it follows that the standard deviation of $\bar{Y}_1 - \bar{Y}_2$ is the square root of this, or $\sigma \sqrt{1/n_1 + 1/n_2}$. The standard error of $\bar{Y}_1 - \bar{Y}_2$ is obtained by substituting the estimate $s_p$ for $\sigma$ in this expression, so that the confidence interval is

$$\bar{y}_1 - \bar{y}_2 \pm t_{97.5} s_{\bar{Y}_1 - \bar{Y}_2} = \bar{y}_1 - \bar{y}_2 \pm t_{97.5} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$$

in which $t_{97.5}$ is the 97.5th percentile of the $t$-distribution with $n_1 + n_2 - 2$ degrees of freedom. The number of degrees of freedom associated with $s_p$ is the sum of the number of degrees of freedom associated with $s_1$ and the number associated with $s_2$. Once $\bar{y}_1$ and $\bar{y}_2$ are known, there are $n_1 + n_2 - 2$ independent observations used in computing $s_p$.

If the two samples do not have the same true variance, other methods, which we shall not detail here, must be used. The same method, however, leads to a good approximation of a 95% confidence interval even if the true variances are unequal, provided $n_1 = n_2$ (i.e. provided the two samples have the same size). When the two samples have different sizes, we must first determine whether it is reasonable to suppose that the two true variances are equal. This and other similar topics are the subjects of the next chapter.

**SUMMARY**

1. An estimator is a rule for calculating an estimate from a set of sample values. It is a random variable that takes on different values (estimates) from sample to sample. The mean of $n$ independent random variables, each with variance $\sigma^2$, has variance $\sigma^2/n$ and standard deviation $\sigma/\sqrt{n}$. The latter, or its estimate $s/\sqrt{n}$, is called the standard error of the mean.

2. An unbiased estimator is one whose mean is equal to the parameter being estimated. The mean and variance of a random sample are unbiased estimators of the
population mean and variance. An efficient estimator is an unbiased estimator that has minimum variance. In the case of normally, binomially, or Poisson-distributed random variables, the sample mean is a minimum variance unbiased estimator. The sample mean is a robust estimator (i.e. a good estimator of the population mean in a wide variety of situations).

3. Maximum likelihood estimates are parameter values that make the likelihood (probability) for the sample a maximum. In large samples, maximum likelihood estimators are usually unbiased, efficient, and normally distributed. Sometimes they also have these properties in small samples.

4. The estimated mean $\pm 2$ standard deviations is often calculated as a ‘normal range’ that contains about 95% of the population values. There is no guarantee that this approximation is good unless the population is normally distributed and the interval is calculated from a large sample. Sample means, on the other hand, tend to be normally distributed regardless of the form of the distribution being sampled.

5. A parameter can be said to lie within a specified interval with a certain degree of confidence, not with any degree of probability. The estimator, but not the estimate, of a 95% confidence interval can be said to have 95% probability of including the parameter. A particular 95% confidence interval for the mean should be interpreted as follows: if many such intervals were to be calculated in the same fashion, each from a different sample, then in the long run 95% of such intervals would include the true mean.

6. For large samples, the maximum likelihood estimate of a parameter $\pm$ twice the standard error is an approximate 95% confidence interval for that parameter. For sample sizes larger than 30, the sample mean $\pm$ two standard errors of the mean provides approximate 95% confidence limits for the mean. For smaller samples and a normally distributed random variable, 95% confidence limits are given by $\bar{y} \pm t_{97.5} s / \sqrt{n}$, where $t_{97.5}$ is the 97.5th percentile of Student’s $t$-distribution with $n - 1$ degrees of freedom.

7. If two independent samples come from populations with the same common variance, a pooled estimate, $s_p^2$, of the variance can be obtained by taking a weighted average of the two sample variances, $s_1^2$ with $n_1 - 1$ degrees of freedom and $s_2^2$ with $n_2 - 1$ degrees of freedom, weighting by the number of degrees of freedom. The pooled estimate then has $n_1 + n_2 - 2$ degrees of freedom and can be used to determine a 95% confidence interval for the difference between the two means: $\bar{y}_1 - \bar{y}_2 \pm t_{97.5} s_p \sqrt{1/n_1 + 1/n_2}$. This same interval is about correct even if the two variances are different, provided the two sample sizes, $n_1$ and $n_2$, are equal.
PROBLEMS

1. An unbiased estimator
   A. is equal to the true parameter
   B. has the smallest variance of all possible estimators
   C. is never an efficient estimate
   D. has mean equal to the true parameter
   E. is always a maximum likelihood estimate

2. If the standard error of the mean obtained from a sample of nine observations is quoted as being three units, then nine units is
   A. the true variance of the population
   B. the estimated variance of the population
   C. the true standard deviation of the population
   D. the estimated standard deviation of the population
   E. none of the above

3. We often choose different estimators for different statistical problems. An estimator that has good properties, even when the assumptions made in choosing it over its competitors are false, is said to be
   A. unbiased
   B. efficient
   C. maximum likelihood
   D. robust
   E. minimum variance

4. For samples comprising a very large number of study units, all the following are true of maximum likelihood estimators except
   A. they are unbiased
   B. they are efficient
   C. they are normally distributed
   D. they are unique in all applications
   E. they are suitable for constructing confidence intervals

5. Parameter values that make the data we observe ‘most likely’ to occur in a sample we have obtained are called
   A. asymptotic estimates
   B. confidence limits
   C. robust estimates
   D. maximum likelihood estimates
   E. interval estimates
6. An experimenter reports that on the basis of a sample of size 10, he calculates the 95% confidence limits for mean height to be 66 and 74 inches. Assuming his calculations are correct, this result is to be interpreted as meaning

A. there is a 95% probability that the population mean height lies between 66 and 74 inches
B. we have 95% confidence that a person's height lies between 66 and 74 inches
C. we have 95% confidence that the population mean height lies between 66 and 74 inches
D. 95% of the population has a height between 66 and 74 inches
E. none of the above

7. A 99% confidence interval for a mean

A. is wider than a 95% confidence interval
B. is narrower than a 95% confidence interval
C. includes the mean with 99% probability
D. excludes the mean with 99% probability
E. is obtained as the sample average plus two standard deviations

8. In a series of journal articles, investigator A reported her data, which are approximately normally distributed, in terms of a mean plus or minus two standard deviations, while investigator B reported his data in terms of a mean plus or minus two standard errors of the mean. The difference between the two methods is

A. investigator A is estimating the extreme percentiles, whereas investigator B is estimating the most usual percentiles
B. investigator A is estimating the range that she thinks contains 95% of the means, whereas investigator B is estimating the range that he thinks contains 95% of the medians
C. investigator A is estimating the range that she thinks contains about 95% of her data values, whereas investigator B is estimating the range that he thinks (with 95% confidence) contains the true mean being estimated
D. investigators A and B are really estimating the same range, but are just using different systems of reporting
E. none of the above
9. A 95% confidence interval implies that
   A. the $t$-test gives correct intervals 95% of the time
   B. if we repeatedly select random samples and construct such interval estimates, 95 out of 100 of the intervals would be expected to bracket the true parameter
   C. the hypothesis will be false in 95 out of 100 such intervals
   D. the probability that the interval is false is 95%
   E. there is a 95% probability that the underlying distribution is normal

10. For Student’s $t$-distribution with one degree of freedom, all the following are true except
   A. it has variance 1
   B. it has fatter tails than a normal distribution
   C. it can be used to obtain confidence limits for the mean of a normal distribution from a sample of two observations
   D. it has mean 0
   E. it is symmetric

11. In a sample of 100 normal women between the ages of 25 and 29 years, systolic blood pressure was found to follow a normal distribution. If the sample mean pressure was 120 mmHg and the standard deviation was 10 mmHg, what interval of blood pressure would represent an approximate 95% confidence interval for the true mean?
   A. 118 to 122 mmHg
   B. 100 to 140 mmHg
   C. 119 to 121 mmHg
   D. 110 to 130 mmHg
   E. 90 to 150 mmHg

12. An investigator is interested in the mean cholesterol level of patients with myocardial infarction. On the basis of a random sample of 50 such patients, a 95% confidence interval for the mean has a width of 10 mg/dl. How large a sample would be expected to have given an interval with a width of about 5 mg/dl?
   A. 100
   B. 200
   C. 300
   D. 400
   E. 800

13. A researcher is interested in the population variability of a normally distributed trait and finds two estimates of its standard deviation in the literature.
These two estimates are similar, and the researcher wishes to average them to obtain one overall estimate. The best procedure is to
A. take the simple average of the estimated standard deviations
B. take a weighted average of the estimated standard deviations, weighting them by their degrees of freedom.
C. take a simple average of the squares of the estimated standard deviations, and then take the square root of the result
D. take a weighted average of the squares of the estimated standard deviations, weighting them by their degrees of freedom, and then take the square root of the result
E. none of the above

14. A sample of five numbers is selected, and it is found that their mean is $\bar{y} = 24$. Given this information, the number of degrees of freedom available for computing the sample standard deviation is
A. 1
B. 2
C. 3
D. 4
E. 5

15. An investigator wishes to estimate the mean cholesterol level in a pediatric population. He decides, on the basis of a small sample, to calculate 95% confidence limits for the population mean. Since the data appear to be normally distributed, the appropriate statistical distribution to use in calculating the confidence interval is the
A. normal distribution
B. $t$-distribution
C. uniform distribution
D. binomial distribution
E. Poisson distribution