Our goal is to estimate the value of an unknown population parameter, such as a population mean or a proportion from a binomial population. We want to use the sample information to estimate the population parameter of interest (called the target parameter) and to assess the reliability of the estimate. Different techniques will be used for estimating a mean or proportion, depending on whether a sample contains a large or small number of measurements.

**Definition 5.1** The unknown population parameter (e.g., mean or proportion) that we are interested in estimating is called the **target parameter**.

### Determining the Target Parameter

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Key Words or Phrases</th>
<th>Type of Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>Mean; Average</td>
<td>Quantitative</td>
</tr>
<tr>
<td>$p$</td>
<td>Proportion; percentage; fraction; rate</td>
<td>Qualitative (success or failure)</td>
</tr>
</tbody>
</table>

For examples, the words *mean* in “mean gas mileage” and *average* in “average life expectancy” imply that the target parameter is the population mean $\mu$. The word *proportion* in “proportion of Iraq War veterans with post-traumatic stress syndrome” indicates that the target parameter is the binomial proportion $p$.

### 5.2 Large-Sample Confidence Interval for a Population Mean

**Motivative example:**
Suppose a large hospital wants to estimate the average length of time patients remain in the hospital. Hence, the hospital’s target parameter is the population mean $\mu$. To accomplish this objective, the hospital administrators plan to randomly sample 100 of all previous patients’ records and to use the sample mean $\bar{X}$ of the lengths of stay to estimate $\mu$, the mean of all patients’ visits. The sample mean $\bar{X}$ represents a point estimator of the population mean $\mu$. How can we assess the accuracy of this large-sample point estimator?

By the central limit theorem, we know that the sampling distribution of the sample mean is approximately normal for large samples. What are the chances that the interval $\bar{X} \pm 2\sigma_{\bar{X}} = \bar{X} \pm \frac{2\sigma}{\sqrt{n}}$ will enclose $\mu$, the population mean?

---

1<sup>1</sup> Last update: July 9, 2012
Definition 5.2 An interval estimator (or confidence interval) is a formula that tells us how to use sample data to calculate an interval that estimates a population parameter.

Definition 5.3 The confidence coefficient is the probability that an interval estimator encloses the population parameter – that is, the relative frequency with which the interval estimator encloses the population parameter when the estimator is used repeatedly a very large number of times. The confidence level is the confidence coefficient expressed as a percentage.

For example, if our confidence level is 95%, then in the long run, 95% of our sample confidence intervals will contain \( \mu \).

Definition 5.4 The value \( z_\alpha \) is defined as the value of the standard normal random variable \( Z \) such that the area \( \alpha \) will lie to its right. In other words, \( P(Z > z_\alpha) = \alpha \).

We can construct a confidence interval with any desired confidence coefficient by increasing or decreasing the area (call it \( \alpha \)) assigned to the tails of the sampling distribution. For example, if we place the area \( \alpha/2 \) in each tail and if \( z_{\alpha/2} \) is the \( z \) value such that \( \alpha/2 \) will lie to its right, then the confidence interval with confidence coefficient is \((1 - \alpha)\) is

\[
\bar{x} \pm z_{\alpha/2} \sigma_{\bar{x}}
\]

Definition 5.5 The value \( z_\alpha \) is defined as the value of the standard normal random variable \( Z \) such that the area \( \alpha \) will lie to its right. In other words, \( P(z > z_\alpha) = \alpha \).

\[
\begin{array}{cccc}
\text{Confidence Level} & 100(1-\alpha) & \alpha & \alpha/2 & z_{\alpha/2} \\
90\% & 0.10 & 0.05 & 1.645 \\
95\% & 0.05 & 0.025 & 1.96 \\
99\% & 0.01 & 0.005 & 2.575 \\
\end{array}
\]

Large-Sample 100(1-\( \alpha \))% Confidence Interval for \( \mu \)

The large-sample 100(1-\( \alpha \))% confidence interval for \( \mu \) is

\[
\bar{x} \pm z_{\alpha/2} \sigma_{\bar{x}} = \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}
\]

where \( z_{\alpha/2} \) is the \( z \) value with an area \( \alpha/2 \) to its right and \( \sigma_{\bar{x}} = \sigma/\sqrt{n} \). The parameter \( \sigma \) is the standard deviation of the sampled population and \( n \) is the sample size.
Note: When $\sigma$ is unknown (as is almost always the case) and $n$ is large (say, $n \geq 30$), the confidence interval is approximately equal to

$$\bar{x} \pm z_{\alpha/2} \left( \frac{s}{\sqrt{n}} \right)$$

where $s$ is the sample standard deviation.

**Conditions Required for a Valid Large-Sample Confidence Interval for $\mu$**

1. A random sample is selected from the target population.

2. The sample size $n$ is large (i.e., $n \geq 30$). (Due to the central limit theorem, this condition guarantees that the sampling distribution of $\bar{x}$ is approximately normal.)

**Interpretation for a Confidence Interval for a Population Mean**

When we form a $100(1-\alpha)\%$ confidence interval for $\mu$, we usually express our confidence in the interval with a statement such as “We can be $100(1-\alpha)\%$ confident that $\mu$ lies between the lower and upper bounds of the confidence interval,” where, for a particular application we substitute the appropriate numerical values for the level of confidence and for the lower and upper bounds. The statement reflects our confidence in the estimation process, rather than in the particular interval that is calculated from the sample data. We know that repeated application of the same procedure will result in different lower and upper bounds on the interval. Furthermore, we know that $100(1-\alpha)\%$ of the resulting intervals will contain $\mu$. There is (usually) no way to determine whether any particular interval is one of those which contain $\mu$ or one of those which do not. However, unlike point estimators, confidence intervals have some measure of reliability – the confidence coefficient – associated with them. For that reason, they are generally preferred to point estimators.
5.3 Small-Sample Confidence Interval for a Population Mean

The use of a small sample in making an inference about \( \mu \) presents two immediate problems.

**Problem 1** The shape of the sampling distribution of the sample mean \( \bar{X} \) now depends on the shape of the population that is sampled. We can no longer assume that the sampling distribution of \( \bar{X} \) is approximately normal, because the central limit theorem ensures normality only for samples that are sufficiently large.

**Solution** The sampling distribution of \( \bar{X} \) is exactly normal even for relatively small samples if the population is normal. It is approximately normal if the sampled population is approximately normal.

**Problem 2** The population standard deviation \( \sigma \) is almost always unknown. Although it is still true that \( \sigma_{\bar{X}} = \sigma/\sqrt{n} \), the sample standard deviation \( s \) may provide a poor approximation for \( \sigma \) when the sample size is small.

**Solution** Instead of using the standard normal statistic

\[
Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}
\]

which requires knowledge of, or a good approximation to, \( \sigma \), we define and use the statistic

\[
t = \frac{\bar{X} - \mu}{s/\sqrt{n}}
\]

in which the sample standard deviation \( s \) replaces the population standard deviation \( \sigma \).

If we are sampling from a normal distribution, the \textbf{t-statistic} has a sampling distribution very much like that of the \( z \)-statistic: mound shaped, symmetric, and with mean 0. The primary difference between the sampling distribution of \( t \) and \( Z \) is that the \( t \)-statistic is more variable than the \( Z \), a property that follows intuitively when you realize that \( t \) contains two random quantities (\( \bar{X} \) and \( s \)), whereas \( z \) contains only one (\( \bar{x} \)).

The actual amount of variability in the sampling distribution of \( t \) depends on the sample size \( n \). A convenient way of expressing this dependence is to say that the \( t \) statistic has \((n-1)\) \textbf{degrees of freedom (df)}. Recall that the quantity \((n-1)\) is the divisor that appears in the formula for \( s^2 \). This number plays a key role in the sampling distribution of \( s^2 \) and appears in discussions of other statistics in later lectures. In particular, the smaller the number of degrees of freedom associated with the \( t \)-statistic, the more variable will be its sampling distribution.
Small-Sample Confidence Interval for $\mu$

The small-sample confidence interval for $\mu$ is

$$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$$

where $t_{\alpha/2}$ is based on $(n - 1)$ degrees of freedom.

Conditions Required for a Valid Small-Sample Confidence Interval for $\mu$

1. A random sample is selected from the target population.
2. The population has a relative frequency distribution that is approximately normal.

5.4 Large-Sample Confidence Interval for a Population Proportion

Problem: Public-opinion polls are conducted regularly to estimate the fraction of U.S. citizens who trust the president. Suppose 1,000 people are randomly chosen and 637 answer that they trust the president. How would you estimate the true fraction of all U.S. citizens who trust the president?

Solution: What we have really asked is how you would estimate the probability $p$ of success in a binomial experiment in which $p$ is the probability that a person chosen trusts the president. One logical method of estimating $p$ for the population is to use the proportion of successes in the sample. That is, we can estimate $p$ by calculating

$$\hat{p} = \frac{\text{Number of people sampled who trust the president}}{\text{Number of people sampled}}$$

where $\hat{p}$ is read "p hat." Thus, in this case,

$$\hat{p} = \frac{637}{1,000} = 0.637$$

Sampling Distribution of $\hat{p}$

1. The mean of the sampling distribution of $\hat{p}$ is $p$; that is, $\hat{p}$ is an unbiased estimator of $p$.
2. The standard deviation of the sampling distribution of $\hat{p}$ is $\sqrt{pq/n}$; that is, $\sigma_{\hat{p}} = \sqrt{pq/n}$, where $q = 1 - p$.
3. For large samples, the sampling distribution of $\hat{p}$ is approximately normal. A sample size is considered large if the interval $\hat{p} \pm 3\sigma_{\hat{p}}$ does not include 0 or 1.
Large-Sample Confidence Interval for $p$

The large-sample confidence interval for $p$ is

\[ \hat{p} \pm z_{\alpha/2} \sigma_{\hat{p}} = \hat{p} \pm z_{\alpha/2} \sqrt{\frac{pq}{n}} = \hat{p} \pm z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} \]

where $\hat{p} = \frac{x}{n}$ and $\hat{q} = 1 - \hat{p}$.

*Note:* When $n$ is large, $\hat{p}$ can approximate the value of $p$ in the formula for $\sigma_{\hat{p}}$.

**Conditions Required for a Valid Large-Sample Confidence Interval for $p$**

1. A random sample is selected from the target population.

2. The sample size $n$ is large. (this condition will be satisfied if both $n\hat{p} \geq 15$ and $n\hat{q} \geq 15$. Note that $n\hat{p}$ and $n\hat{q}$ are simply the number of successes and number of failures, respectively, in the sample.

*Unless* $n$ *is extremely large, the large-sample procedure presented in this section performs poorly when $p$ is near 0 or 1.* To overcome this potential problem, an *extremely* large sample size is required. Since the value of $n$ required to satisfy “extremely large” is difficult to determine, statisticians have proposed an alternative method, based on the Willson (1927) point estimator of $p$. Researchers have shown that the confidence interval with *Wilson’s adjustment for estimating $p$* works well for any $p$, even when the sample size $n$ is very small.

**Adjusted (1 − $\alpha$)100% Confidence Interval for a Population Proportion $p$**

An adjusted confidence interval for $p$ is

\[ \tilde{p} \pm z_{\alpha/2} \sqrt{\frac{\tilde{p}(1 - \tilde{p})}{n + 4}} \]

where $\tilde{p} = \frac{x + 2}{n + 4}$ is the adjusted sample proportion of observations with the characteristic of interest, $x$ is the number of successes in the sample, and $n$ is the sample size.
5.5 Determining the Sample Size

In this section, we show the appropriate sample size for making an inference about a population mean or proportion depends on the desired reliability.

**Determination of Sample Size for 100(1 – α)% Confidence Intervals for μ**

In order to estimate μ with a sampling error SE, half-width of the confidence interval, and with 100(1 – α)% confidence, the required sample size is found as follows:

\[
z_{\alpha/2} \left( \frac{\sigma}{\sqrt{n}} \right) = SE
\]

The solution for \( n \) is given by the equation

\[
n = \left( \frac{z_{\alpha/2}}{SE} \right)^2 \frac{\sigma^2}{σ^2}
\]

The value of \( σ \) is usually unknown. It can be estimated by the standard deviation \( s \) from a previous sample. Alternatively, we may approximate the range \( R \) of observations in the population and (conservatively) estimate \( σ \approx \frac{R}{4} \). In any case, you should round the value of \( n \) obtained upward to ensure that the sample size will be sufficient to achieve the specified reliability.

**Determination of Sample Size for 100(1 – α)% Confidence Interval for \( p \)**

In order to estimate a binomial probability \( p \) with sampling error SE and with 100(1 – α)% confidence, the required sample size is found by solving the following equation for \( n \):

\[
z_{\alpha/2} \sqrt{\frac{pq}{n}} = SE
\]

The solution for \( n \) can be written as follows:

\[
n = \left( \frac{z_{\alpha/2}}{SE} \right)^2 \frac{(pq)}{(σ^2)}
\]

Since the value of the product \( pq \) is unknown, it can be estimated by the sample fraction of successes, \( \hat{p} \), from a previous sample. We can show that the value of \( pq \) is at its maximum when \( p \) equals 0.5, so you can obtain conservatively large values of \( n \) by approximating \( p \) by 0.5 or values close to 0.5. In any case, you should round the value of \( n \) obtained upward to ensure that the sample size will be sufficient to achieve the specified reliability.