8

CONFIDENCE INTERVALS

Figure 8.1 Have you ever wondered what the average number of M&Ms in a bag at the grocery store is? You can use confidence intervals to answer this question. (credit: modification of work "sweet, orange, food, green, red, color, brown, blue, colorful, yellow, chocolate, snack, dessert, toy, plain, candy, sweetness, treat, confectionery, coated, m ms, hard shell, snack food, jelly bean"/ Pxhere, Public Domain)

Introduction

Suppose you were trying to determine the rent of a two-bedroom apartment in your town. You might look in the classified section of the newspaper, write down several rents listed, and average them together. You would have obtained a point estimate of the true mean. If you are trying to determine the percentage of times you make a basket when shooting a basketball, you might count the number of shots you make and divide that by the number of shots you attempted. In this case, you would have obtained a point estimate for the true proportion the parameter p in the binomial probability density function.

We use sample data to make generalizations about an unknown population. This part of statistics is called **inferential statistics**. **The sample data help us to make an estimate of a population parameter**. We realize that the point estimate is most likely not the exact value of the population parameter, but close to it. After calculating point estimates, we construct interval estimates, called confidence intervals. What statistics provides us beyond a simple average, or point estimate, is an estimate to which we can attach a probability of accuracy, what we will call a confidence level. We make inferences with a known level of probability.

In this chapter, you will learn to construct and interpret confidence intervals. You will also learn a new distribution, the Student's-t, and how it is used with these intervals. Throughout the chapter, it is important to keep in mind that the confidence interval is a random variable. It is the population parameter that is fixed.

If you worked in the marketing department of an entertainment company, you might be interested in the mean number of songs a consumer downloads a month from Apple Music. If so, you could conduct a survey and calculate the sample mean, $\overline{x}$, and the sample standard deviation, s. You would use $\overline{x}$ to estimate the population mean and s to estimate the population standard deviation. The sample mean, $\overline{x}$, is the **point estimate** for the population mean, μ. The sample

standard deviation, s, is the point estimate for the population standard deviation, σ.

$\overline{x}$ and s are each called a statistic.

A **confidence interval** is another type of estimate but, instead of being just one number, it is an interval of numbers. The interval of numbers is a range of values calculated from a given set of sample data. The confidence interval is likely to include the unknown population parameter.

Suppose, for the Apple Music example, we do not know the population mean μ, but we do know the sample mean and that the population standard deviation is $\sigma = 1$ and our sample size is 100. Then, by the central limit theorem, the standard deviation of the sampling distribution of the sample means is

$$\frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{100}} = 0.1$$

The **Empirical Rule**, which applies to the normal distribution, says that in approximately 95% of the samples, the sample mean, $\overline{x}$, will be within two standard deviations of the population mean μ. For our Apple Music example, two standard deviations is (2)(0.1) = 0.2. The sample mean $\overline{x}$ is likely to be within 0.2 units of μ.

Because $\overline{x}$ is within 0.2 units of μ, which is unknown, then μ is likely to be within 0.2 units of $\overline{x}$ with 95% probability. The population mean μ is contained in an interval whose lower number is calculated by taking the sample mean and subtracting two standard deviations (2)(0.1) and whose upper number is calculated by taking the sample mean and adding two standard deviations. In other words, μ is between $\overline{x} - 0.2$ and $\overline{x} + 0.2$ in 95% of all the samples.

For the Apple Music example, suppose that a sample produced a sample mean $\overline{x} = 2$. Then with 95% probability the unknown population mean μ is between

$\overline{x} - 0.2 = 2 - 0.2 = 1.8$ and $\overline{x} + 0.2 = 2 + 0.2 = 2.2$

We say that we are **95% confident** that the unknown population mean number of songs downloaded from Apple Music per month is between 1.8 and 2.2. **The 95% confidence interval is (1.8, 2.2).** Please note that we talked in terms of 95% confidence using the empirical rule. The empirical rule for two standard deviations is only approximately 95% of the probability under the normal distribution. To be precise, two standard deviations under a normal distribution is actually 95.44% of the probability. To calculate the exact 95% confidence level we would use 1.96 standard deviations.

The 95% confidence interval implies two possibilities. Either the interval (1.8, 2.2) contains the true mean μ, or our sample produced an $\overline{x}$ that is not within 0.2 units of the true mean μ. The first possibility happens for 95% of well-chosen samples. It is important to remember that the second possibility happens for 5% of samples, even though correct procedures are followed.

Remember that a confidence interval is created for an unknown population parameter like the population mean, μ.

For the confidence interval for a mean the formula would be:

$$\mu = \overline{X} \pm Z_{\alpha} {}^{\sigma}\Big/\sqrt{n}$$

Or written another way as:

$$\overline{X} - Z_{\alpha} {}^{\sigma}\Big/\sqrt{n} \leq \mu \leq \overline{X} + Z_{\alpha} {}^{\sigma}\Big/\sqrt{n}$$

Where $\overline{X}$ is the sample mean. Z_{α} is determined by the level of confidence desired by the analyst, and ${}^{\sigma}\Big/\sqrt{n}$ is the standard deviation of the sampling distribution for means given to us by the Central Limit Theorem.

8.1 A Confidence Interval When the Population Standard Deviation Is Known or Large Sample Size

A confidence interval for a population mean, when the population standard deviation is known based on the conclusion of the Central Limit Theorem that the sampling distribution of the sample means follow an approximately normal distribution.

Calculating the Confidence Interval

Consider the standardizing formula for the sampling distribution developed in the discussion of the Central Limit Theorem:

$$Z_1 = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

Notice that μ is substituted for $\mu_{\bar{x}}$ because we know that the expected value of $\mu_{\bar{x}}$ is μ from the Central Limit theorem and $\sigma_{\bar{x}}$ is replaced with $\sigma/\sqrt{n}$, also from the Central Limit Theorem.

In this formula we know $\bar{X}$, $\sigma_{\bar{x}}$ and n, the sample size. (In actuality we do not know the population standard deviation, but we do have a point estimate for it, s, from the sample we took. More on this later.) What we do not know is μ or Z_1. We can solve for either one of these in terms of the other. Solving for μ in terms of Z_1 gives:

$$\mu = \bar{X} \pm Z_1 \sigma/\sqrt{n}$$

Remembering that the Central Limit Theorem tells us that the distribution of the $\overline{X}$'s, the sampling distribution for means, is normal, and that the normal distribution is symmetrical, we can rearrange terms thus:

$$\overline{X} - Z_\alpha\left(\sigma/\sqrt{n}\right) \leq \mu \leq \overline{X} + Z_\alpha\left(\sigma/\sqrt{n}\right)$$

This is the formula for a confidence interval for the mean of a population.

Notice that Z_α has been substituted for Z_1 in this equation. This is where a choice must be made by the statistician. The analyst must decide the level of confidence they wish to impose on the confidence interval. α is the probability that the interval will not contain the true population mean. The confidence level is defined as (1-α). Z_α is the number of standard deviations $\overline{X}$ lies from the mean with a certain probability. If we chose Z_α = 1.96 we are asking for the 95% confidence interval because we are setting the probability that the true mean lies within the range at 0.95. If we set Z_α at 1.64 we are asking for the 90% confidence interval because we have set the probability at 0.90. These numbers can be verified by consulting the Standard Normal table. Divide either 0.95 or 0.90 in half and find that probability inside the body of the table. Then read on the top and left margins the number of standard deviations it takes to get this level of probability.

In reality, we can set whatever level of confidence we desire simply by changing the Z_α value in the formula. It is the analyst's choice. Common convention in Economics and most social sciences sets confidence intervals at either 90, 95, or 99 percent levels. Levels less than 90% are considered of little value. The level of confidence of a particular interval estimate is called by (1-α).

A good way to see the development of a confidence interval is to graphically depict the solution to a problem requesting a confidence interval. This is presented in Figure 8.2 for the example in the introduction concerning the number of downloads from Apple Music. That case was for a 95% confidence interval, but other levels of confidence could have just as easily been chosen depending on the need of the analyst. However, the level of confidence MUST be pre-set and not subject to revision as a result of the calculations.

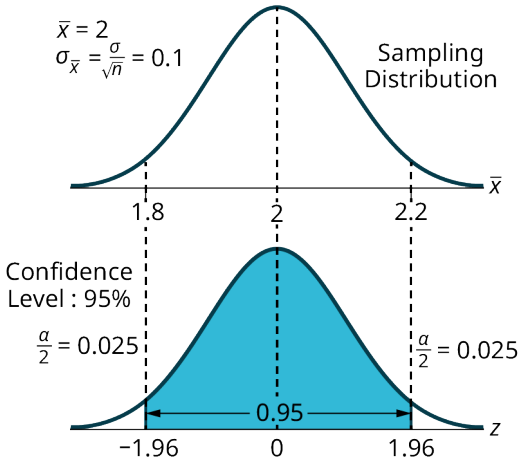

Figure 8.2

$$\mu = \overline{X} \pm Z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right)$$
$$= 2 \pm 1.96(0.1)$$
$$= 2 \pm 0.196$$
$$1.804 \leq \mu \leq 2.196$$

For this example, let's say we know that the actual population mean number of Apple Music downloads is 2.1. The true population mean falls within the range of the 95% confidence interval. There is absolutely nothing to guarantee that this will happen. **Further, if the true mean falls outside of the interval we will never know it. We must always remember that we will never ever know the true mean.** Statistics simply allows us, with a given level of probability (confidence), to say that the true mean is within the range calculated. This is what was called in the introduction, the "level of ignorance admitted".

Changing the Confidence Level or Sample Size

Here again is the formula for a confidence interval for an unknown population mean assuming we know the population standard deviation:

$$\overline{X} - Z_\alpha\left(\sigma\Big/\sqrt{n}\right) \leq \mu \leq \overline{X} + Z_\alpha\left(\sigma\Big/\sqrt{n}\right)$$

It is clear that the confidence interval is driven by two things, the chosen level of confidence, Z_α, and the standard deviation of the sampling distribution. The Standard deviation of the sampling distribution is further affected by two things, the standard deviation of the population and the sample size we chose for our data. Here we wish to examine the effects of each of the choices we have made on the calculated confidence interval, the confidence level and the sample size.

For a moment we should ask just what we desire in a confidence interval. Our goal was to estimate the population mean from a sample. We have forsaken the hope that we will ever find the true population mean, and population standard deviation for that matter, for any case except where we have an extremely small population and the cost of gathering the data of interest is very small. In all other cases we must rely on samples. With the Central Limit Theorem we have the tools to provide a meaningful confidence interval with a given level of confidence, meaning a known probability of being wrong. By meaningful confidence interval we mean one that is useful. Imagine that you are asked for a confidence interval for the ages of your classmates. You have taken a sample and find a mean of 19.8 years. You wish to be very confident so you report an interval between 9.8 years and 29.8 years. This interval would certainly contain the true population mean and have a very high confidence level. However, it hardly qualifies as meaningful. The very best confidence interval is narrow while having high confidence. There is a natural tension between these two goals. The higher the level of confidence the wider the confidence interval as the case of the students' ages above. We can see this tension in the equation for the confidence interval.

$$\mu = \overline{x} \pm Z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right)$$

The confidence interval will increase in width as $Z\alpha$ increases, $Z\alpha$ increases as the level of confidence increases. There is a tradeoff between the level of confidence and the width of the interval. Now let's look at the formula again and we see that the sample size also plays an important role in the width of the confidence interval. The sample size, n, shows up in the denominator of the standard deviation of the sampling distribution. As the sample size increases, the standard deviation of the sampling distribution decreases and thus the width of the confidence interval, while holding constant the level of confidence. Again we see the importance of having large samples for our analysis although we then face a second constraint, the cost of gathering data.

EXAMPLE 8.1

Suppose we are interested in the mean scores on an exam. A random sample of 36 scores is taken and gives a sample mean (sample mean score) of 68 ($\overline{X}$ = 68). In this example we have the unusual knowledge that the population standard deviation is 3 points. Do not count on knowing the population parameters outside of textbook examples. Find a confidence interval estimate for the population mean exam score (the mean score on all exams).

Problem

Find a 90% confidence interval for the true (population) mean of statistics exam scores.

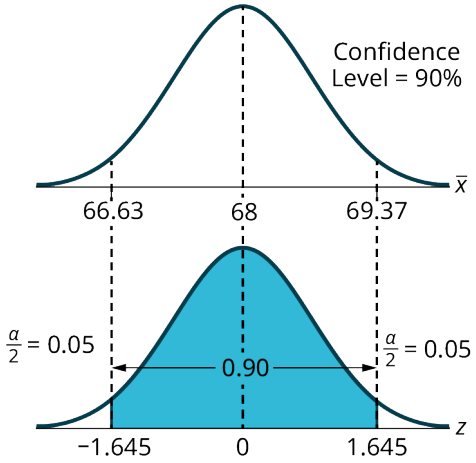

Figure 8.3

Solution

The solution is shown step by step:

The formula for a confidence interval for an unknown population mean assuming we know the population standard deviation is:

$$\overline{X} - Z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \overline{X} + Z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right)$$

For a 90% confidence interval, visualize an area of 0.90 centered under the normal curve (See Figure 8.3). The remaining area for the two tails of the normal distribution is then 0.10, which indicates that the area in the left tail is one-half of 0.10, which is 0.05. The corresponding z-score that cuts off an area of 0.05 in the left tail is 1.645.

In this example we are given that the population standard deviation $\sigma = 3$.

We are also given that the sample size $n = 36$ and the sample mean $\overline{X} = 68$.

Substituting these values in the confidence interval formula results in the following:

$$\overline{X} - Z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right) \leq \mu \leq \overline{X} + Z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$68 - 1.645\left(\frac{3}{\sqrt{36}}\right) \leq \mu \leq 68 + 1.645\left(\frac{3}{\sqrt{36}}\right)$$

$$68 - 0.8225 \leq \mu \leq 68 + 0.8225$$

$$67.1775 \leq \mu \leq 68.8225$$

We estimate with 90% confidence that the true population mean exam score for all statistics students is between 67.18 and 68.82.

TRY IT 8.1

Suppose average pizza delivery times are normally distributed with an unknown population mean and a population standard deviation of six minutes. A random sample of 28 pizza delivery restaurants is taken and has a sample mean delivery time of 36 minutes.

Find a 90% confidence interval estimate for the population mean delivery time.

EXAMPLE 8.2

Problem

Suppose we change the original problem in Example 8.1 by using a 95% confidence level. Find a 95% confidence interval for the true (population) mean statistics exam score.

Solution

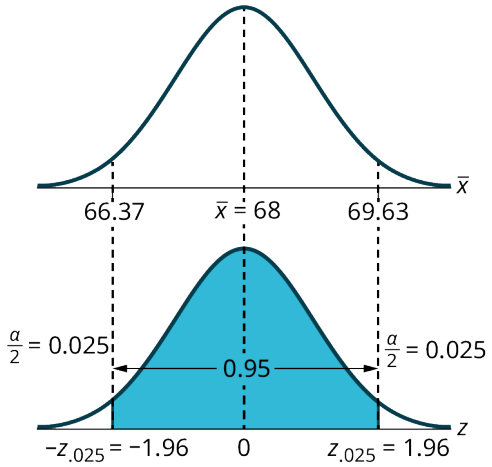

Figure 8.4

$$\mu = \bar{x} \pm Z_\alpha \left(\frac{\sigma}{\sqrt{n}} \right)$$

$$\mu = 68 \pm 1.96 \left(\frac{3}{\sqrt{36}} \right)$$

$$67.02 \leq \mu \leq 68.98$$

$\sigma = 3$; $n = 36$; The confidence level is 95% ($CL = 0.95$).

$CL = 0.95$ so $\alpha = 1 - CL = 1 - 0.95 = 0.05$

$$Z_{\frac{\alpha}{2}} = Z_{0.025} = 1.96$$

Notice that the plus/minus term in the equation is larger for a 95% confidence level in the original problem.

Comparing the results: The 90% confidence interval is (67.18, 68.82). The 95% confidence interval is (67.02, 68.98). The 95% confidence interval is wider. If you look at the graphs, because the area 0.95 is larger than the area 0.90, it makes sense that the 95% confidence interval is wider. To be more confident that the confidence interval actually does contain the true value of the population mean for all statistics exam scores, the confidence interval necessarily needs to be wider. This demonstrates a very important principle of confidence intervals. There is a trade off between the level of confidence and the width of the interval. Our desire is to have a narrow confidence interval, huge wide intervals provide little information that is useful. But we would also like to have a high level of confidence in our interval. This demonstrates that we cannot have both.

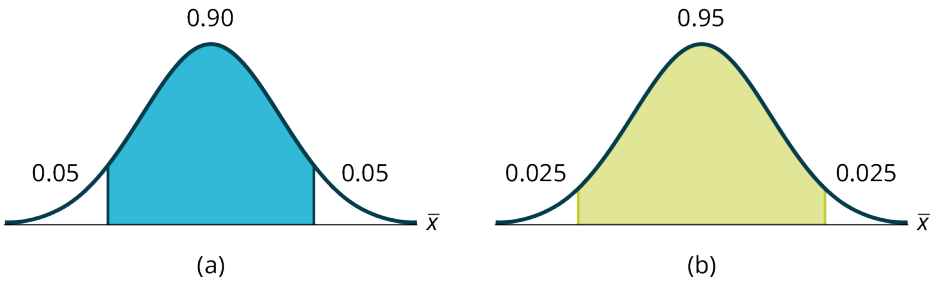

Figure 8.5

Summary: Effect of Changing the Confidence Level

- Increasing the confidence level makes the confidence interval wider.
- Decreasing the confidence level makes the confidence interval narrower.

And again here is the formula for a confidence interval for an unknown mean assuming we have the population standard deviation:

$$\overline{X} - Z_\alpha\left(\sigma\Big/\sqrt{n}\right) \leq \mu \leq \overline{X} + Z_\alpha\left(\sigma\Big/\sqrt{n}\right)$$

The standard deviation of the sampling distribution was provided by the Central Limit Theorem as $\sigma/\sqrt{n}$. While we infrequently get to choose the sample size it plays an important role in the confidence interval. Because the sample size is in the denominator of the equation, as n increases it causes the standard deviation of the sampling distribution to decrease and thus the width of the confidence interval to decrease. We have met this before as we reviewed the effects of sample size on the Central Limit Theorem. There we saw that as n increases the sampling distribution narrows until in the limit it collapses on the true population mean.

TRY IT 8.2

Refer back to the pizza-delivery Try It 8.1 exercise. The population standard deviation is six minutes and the sample mean deliver time is 36 minutes. Use a sample size of 20. Find a 95% confidence interval estimate for the true mean pizza delivery time.

Changing the Sample Size

EXAMPLE 8.3

Suppose we change the original problem in Example 8.1 to see what happens to the confidence interval if the sample size is changed.

Problem

Leave everything the same except the sample size. Use the original 90% confidence level. What happens to the confidence interval if we increase the sample size and use $n = 100$ instead of $n = 36$? What happens if we decrease the sample size to $n = 25$ instead of $n = 36$?

Solution

$$\mu = \overline{x} \pm Z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$\mu = 68 \pm 1.645\left(\frac{3}{\sqrt{100}}\right)$$

$$67.5065 \leq \mu \leq 68.4935$$

If we **increase** the sample size n to 100, we **decrease** the width of the confidence interval relative to the original sample size of 36 observations.

Solution

$$\mu = \overline{x} \pm Z_\alpha\left(\frac{\sigma}{\sqrt{n}}\right)$$

$$\mu = 68 \pm 1.645\left(\frac{3}{\sqrt{25}}\right)$$

$$67.013 \leq \mu \leq 68.987$$

If we **decrease** the sample size n to 25, we **increase** the width of the confidence interval by comparison to the original sample size of 36 observations.

Summary: Effect of Changing the Sample Size

- Increasing the sample size makes the confidence interval narrower.

- Decreasing the sample size makes the confidence interval wider.

TRY IT 8.3

Refer back to the pizza-delivery Try It 8.1 exercise. The mean delivery time is 36 minutes and the population standard deviation is six minutes. Assume the sample size is changed to 50 restaurants with the same sample mean. Find a 90% confidence interval estimate for the population mean delivery time.

We have already seen this effect when we reviewed the effects of changing the size of the sample, *n*, on the Central Limit Theorem. See Figure 7.8 to see this effect. Before we saw that as the sample size increased the standard deviation of the sampling distribution decreases. This was why we choose the sample mean from a large sample as compared to a small sample, all other things held constant.

Thus far we assumed that we knew the population standard deviation. This will virtually never be the case. We will have the sample standard deviation, *s*, however. This is a point estimate for the population standard deviation and can be substituted into the formula for confidence intervals for a mean under certain circumstances. We just saw the effect the sample size has on the width of confidence interval and the impact on the sampling distribution for our discussion of the Central Limit Theorem. We can invoke this to substitute the point estimate for the standard deviation if the sample size is large "enough". Simulation studies indicate that 30 observations or more will be sufficient to eliminate any meaningful bias in the estimated confidence interval.

EXAMPLE 8.4

Spring break can be a very expensive holiday. A sample of 80 students is surveyed, and the average amount spent by students on travel and beverages is $593.84. The sample standard deviation is approximately $369.34.

Problem

Construct a 92% confidence interval for the population mean amount of money spent by spring breakers.

Solution

We begin with the confidence interval for a mean. We use the formula for a mean because the random variable is dollars spent and this is a continuous random variable. The point estimate for the population standard deviation, *s*, has been substituted for the true population standard deviation because with 80 observations there is no concern for bias in the estimate of the confidence interval.

$$\mu = \overline{x} \pm \left[Z_{(\mathrm{a}\ /2)} \frac{s}{\sqrt{n}} \right]$$

Substituting the values into the formula, we have:

$$\mu = 593.84 \pm \left[1.75 \frac{369.34}{\sqrt{80}} \right]$$

$Z_{(a/2)}$ is found on the standard normal table by looking up 0.46 in the body of the table and finding the number of standard deviations on the side and top of the table; 1.75. The solution for the interval is thus:

$$\mu = 593.84 \pm 72.2636 = (521.57, 666.10)$$

$$\$521.58 \leq \mu \leq \$666.10$$

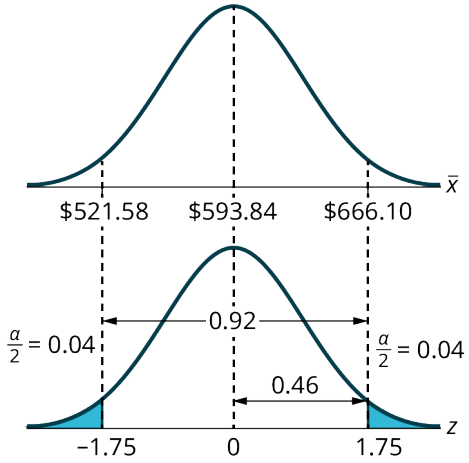

Figure 8.6

TRY IT 8.4

The price of a chair is a large range of cost. The average cost of 25 chairs in a store is $100. The sample standard deviation is $50. Construct a 92% confidence interval for the population mean of the cost of chairs.

Formula Review

The general form for a confidence interval for a single population mean, known standard deviation, normal distribution is given by $\overline{X} - Z_\alpha\left(\sigma/\sqrt{n}\right) \leq \mu \leq \overline{X} + Z_\alpha\left(\sigma/\sqrt{n}\right)$ This formula is used when the population standard deviation is known.

CL = confidence level, or the proportion of confidence intervals created that are expected to contain the true population parameter

$\alpha = 1 - CL$ = the proportion of confidence intervals that will not contain the population parameter

$z_{\frac{\alpha}{2}}$ = the *z*-score with the property that the area to the right of the z-score is $\frac{\alpha}{2}$ this is the *z*-score used in the calculation where α = 1 – *CL*.

8.2 A Confidence Interval When the Population Standard Deviation Is Unknown and Small Sample Case

In practice, we rarely know the population **standard deviation**. In the past, when the sample size was large, this did not present a problem to statisticians. They used the sample standard deviation *s* as an estimate for σ and proceeded as before to calculate a **confidence interval** with close enough results. This is what we did in Example 8.4 above. The point estimate for the standard deviation, s, was substituted in the formula for the confidence interval for the population standard deviation. In this case the 80 observations are well above the suggested 30 observations to eliminate any bias from a small sample. However, statisticians ran into problems when the sample size was small. A small sample size caused inaccuracies in the confidence interval.

William S. Goset (1876–1937) of the Guinness brewery in Dublin, Ireland ran into this problem. His experiments with hops and barley produced very few samples. Just replacing σ with *s* did not produce accurate results when he tried to calculate a confidence interval. He realized that he could not use a normal distribution for the calculation; he found that the actual distribution depends on the sample size. This problem led him to "discover" what is called the **Student's t-distribution**. The name comes from the fact that Gosset wrote under the pen name "A Student."

Up until the mid-1970s, some statisticians used the **normal distribution** approximation for large sample sizes and used the Student's t-distribution only for sample sizes of at most 30 observations.

If you draw a simple random sample of size *n* from a population with mean μ and unknown population standard

deviation σ and calculate the t-score $t = \frac{\overline{x}-\mu}{\left(\frac{s}{\sqrt{n}}\right)}$, then the t-scores follow a **Student's t-distribution with *n* – 1 degrees of freedom**. The t-score has the same interpretation as the ***z*-score**. It measures how far in standard deviation units $\overline{x}$ is from its mean μ. For each sample size n, there is a different Student's t-distribution.

The **degrees of freedom**, ***n* – 1**, come from the calculation of the sample standard deviation ***s***. Remember when we first calculated a sample standard deviation we divided the sum of the squared deviations by n - 1, but we used n deviations $(x - \overline{x}\text{values})$ to calculate ***s***. Because the sum of the deviations is zero, we can find the last deviation once we know the other ***n* – 1** deviations. The other ***n* – 1** deviations can change or vary freely. **We call the number *n* – 1 the degrees of freedom (df)** in recognition that one is lost in the calculations. The effect of losing a degree of freedom is that the t-value increases and the confidence interval increases in width.

Properties of the Student's t-Distribution

- The graph for the Student's t-distribution is similar to the standard normal curve and at infinite degrees of freedom it is the normal distribution. You can confirm this by reading the bottom line at infinite degrees of freedom for a familiar level of confidence, e.g. at column 0.05, 95% level of confidence, we find the t-value of 1.96 at infinite degrees of freedom.
- The mean for the Student's t-distribution is zero and the distribution is symmetric about zero, again like the standard normal distribution.
- The Student's t-distribution has more probability in its tails than the standard normal distribution because the spread of the t-distribution is greater than the spread of the standard normal. So the graph of the Student's t-distribution will be thicker in the tails and shorter in the center than the graph of the standard normal distribution.
- The exact shape of the Student's t-distribution depends on the degrees of freedom. As the degrees of freedom increases, the graph of Student's t-distribution becomes more like the graph of the standard normal distribution.
- The underlying population of individual observations is assumed to be normally distributed with unknown population mean μ and unknown population standard deviation σ. This assumption comes from the Central Limit theorem because the individual observations in this case are the $\overline{x}$s of the sampling distribution. The size of the underlying population is generally not relevant unless it is very small. If it is normal then the assumption is met and doesn't need discussion.

A probability table for the Student's t-distribution is used to calculate t-values at various commonly-used levels of confidence. The table gives t-scores that correspond to the confidence level (column) and degrees of freedom (row). When using a t-table, note that some tables are formatted to show the confidence level in the column headings, while the column headings in some tables may show only corresponding area in one or both tails. Notice that at the bottom the table will show the t-value for infinite degrees of freedom. Mathematically, as the degrees of freedom increase, the t-distribution approaches the standard normal distribution. You can find familiar Z-values by looking in the relevant alpha column and reading value in the last row.

A Student's t table (See Appendix A Statistical Tables) gives t-scores given the degrees of freedom and the right-tailed probability.

The Student's t-distribution has one of the most desirable properties of the normal distribution: it is symmetrical. What the Student's t-distribution does is spread out the horizontal axis so it takes a larger number of standard deviations to capture the same amount of probability. In reality there are an infinite number of Student's t-distributions, one for each adjustment to the sample size. As the sample size increases, the Student's t-distribution become more and more like the normal distribution. When the sample size reaches 30 the normal distribution is usually substituted for the Student's t because they are so much alike. This relationship between the Student's t-distribution and the normal distribution is shown in Figure 8.7.

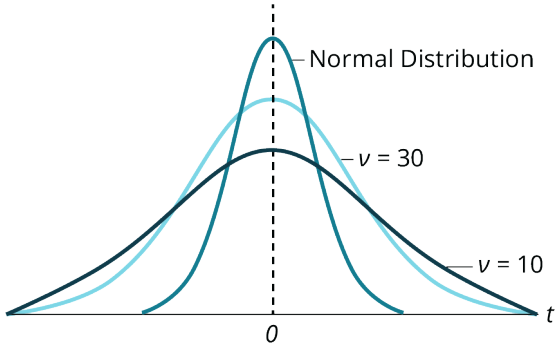

Figure 8.7

This is another example of one distribution limiting another one, in this case the normal distribution is the limiting distribution of the Student's t when the degrees of freedom in the Student's t approaches infinity. This conclusion comes directly from the derivation of the Student's *t*-distribution by Mr. Gosset. He recognized the problem as having few observations and no estimate of the population standard deviation. He was substituting the sample standard deviation and getting volatile results. He therefore created the Student's *t*-distribution as a ratio of the normal distribution and Chi squared distribution. The Chi squared distribution is itself a ratio of two variances, in this case the sample variance and the unknown population variance. The Student's *t*-distribution thus is tied to the normal distribution, but has degrees of freedom that come from those of the Chi squared distribution. The algebraic solution demonstrates this result.

Development of Student's t-distribution:

$$t = \frac{z}{\sqrt{\frac{\chi^2}{v}}}$$

where z is the standard normal variable and χ^2 is the chi-squared distribution with v degrees of freedom. Substitute values and simplify:

$$t = \frac{\frac{(\overline{x}-\mu)}{\sigma}}{\sqrt{\frac{\frac{s^2}{(n-1)}}{\frac{\sigma^2}{(n-1)}}}} = \frac{\frac{(\overline{x}-\mu)}{\sigma}}{\sqrt{\frac{s^2}{\sigma^2}}} = \frac{\frac{(\overline{x}-\mu)}{\sigma}}{\frac{s}{\sigma}} = \frac{\overline{x}-\mu}{\frac{s}{\sqrt{n}}}$$

$$t = \frac{s}{\sqrt{n}}$$

Restating the formula for a confidence interval for the mean for cases when the sample size is smaller than 30 and we do not know the population standard deviation, σ:

$$\overline{x} - t_{\mathrm{v},\alpha}\left(\frac{s}{\sqrt{n}}\right) \leq \mu \leq \overline{x} + t_{\mathrm{v},\alpha}\left(\frac{s}{\sqrt{n}}\right)$$

Here the point estimate of the population standard deviation, s has been substituted for the population standard deviation, σ, and $t_{\nu,\alpha}$ has been substituted for Z_α. The Greek letter ν (pronounced nu) is placed in the general formula in recognition that there are many Student t_ν distributions, one for each sample size. ν is the symbol for the degrees of freedom of the distribution and depends on the size of the sample. Often df is used to abbreviate degrees of freedom. **For this type of problem,** the degrees of freedom is ν = n-1, where n is the sample size. To look up a probability in the Student's t table we have to know the degrees of freedom in the problem.

EXAMPLE 8.5

Problem

The average earnings per share (EPS) for 10 industrial stocks randomly selected from those listed on the Dow-Jones Industrial Average was found to be $\overline{X}$ = 1.85 with a standard deviation of s=0.395. Calculate a 99% confidence interval for the average EPS of all the industrials listed on the DJIA.

$$\overline{x} - t_{\mathrm{v},\alpha}\left(\frac{s}{\sqrt{n}}\right) \leq \mu \leq \overline{x} + t_{\mathrm{v},\alpha}\left(\frac{s}{\sqrt{n}}\right)$$

Solution

To help visualize the process of calculating a confident interval we draw the appropriate distribution for the problem. In this case this is the Student's t because we do not know the population standard deviation and the sample is small, less than 30.

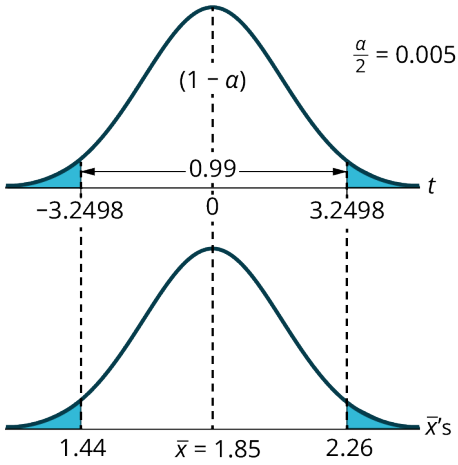

Figure 8.8

To find the appropriate t-value requires two pieces of information, the level of confidence desired and the degrees of freedom. The question asked for a 99% confidence level. On the graph this is shown where (1-α) , the level of confidence , is in the unshaded area. The tails, thus, have .005 probability each, α/2. The degrees of freedom for this type of problem is n-1= 9. From the Student's t table, at the row marked 9 and column marked .005, is the number of standard deviations to capture 99% of the probability, 3.2498. These are then placed on the graph remembering that the Student's t is symmetrical and so the t-value is both plus or minus on each side of the mean.

Inserting these values into the formula gives the result. These values can be placed on the graph to see the relationship between the distribution of the sample means, $\overline{X}$'s and the Student's *t*-distribution.

$$\mu = \overline{X} \pm t_{\alpha/2,\text{df=n-1}} \frac{s}{\sqrt{n}} = 1.851 \pm 3.2498 \frac{0.395}{\sqrt{10}} = 1.8551 \pm 0.406$$

$$1.445 \leq \mu \leq 2.257$$

We state the formal conclusion as :

With 99% confidence level, the average EPS of all the industries listed at DJIA is from $1.44 to $2.26.

TRY IT 8.5

You do a study of hypnotherapy to determine how effective it is in increasing the number of hours of sleep subjects get each night. You measure hours of sleep for 12 subjects with the following results. Construct a 95% confidence interval for the mean number of hours slept for the population (assumed normal) from which you took the data.

8.2; 9.1; 7.7; 8.6; 6.9; 11.2; 10.1; 9.9; 8.9; 9.2; 7.5; 10.5

8.3 A Confidence Interval for A Population Proportion

During an election year, we see articles in the newspaper that state **confidence intervals** in terms of proportions or percentages. For example, a poll for a particular candidate running for president might show that the candidate has 40% of the vote within three percentage points (if the sample is large enough). Often, election polls are calculated with 95% confidence, so, the pollsters would be 95% confident that the true proportion of voters who favored the candidate would be between 0.37 and 0.43.

Investors in the stock market are interested in the true proportion of stocks that go up and down each week. Businesses that sell personal computers are interested in the proportion of households in the United States that own personal

computers. Confidence intervals can be calculated for the true proportion of stocks that go up or down each week and for the true proportion of households in the United States that own personal computers.

The procedure to find the confidence interval for a population proportion is similar to that for the population mean, but the formulas are a bit different although conceptually identical. While the formulas are different, they are based upon the same mathematical foundation given to us by the Central Limit Theorem. Because of this we will see the same basic format using the same three pieces of information: the sample value of the parameter in question, the standard deviation of the relevant sampling distribution, and the number of standard deviations we need to have the confidence in our estimate that we desire.

How do you know you are dealing with a proportion problem? First, the underlying **distribution has a binary random variable and therefore is a binomial distribution**. (There is no mention of a mean or average.) If X is a binomial random variable, then $X \sim B(n, p)$ where n is the number of trials and p is the probability of a success. To form a sample proportion, take X, the random variable for the number of successes and divide it by n, the number of trials (or the sample size). The random variable P' (read "P prime") is the sample proportion,

$$P' = \frac{X}{n}$$

(Sometimes the random variable is denoted as $\widehat{P}$, read "P hat".)

p' = the **estimated proportion** of successes or sample proportion of successes (p' is a **point estimate** for p, the true population proportion, and thus q is the probability of a failure in any one trial.)

x = the **number** of successes in the sample

n = the size of the sample

The formula for the confidence interval for a population proportion follows the same format as that for an estimate of a population mean. Remembering the sampling distribution for the proportion from The Central Limit Theorem, the standard deviation was found to be:

$$\sigma_{p'} = \sqrt{\frac{p(1-p)}{n}}$$

The confidence interval for a population proportion, therefore, becomes:

$$p = p' \pm \left[Z_{\left(\frac{a}{2}\right)} \sqrt{\frac{p'(1-p')}{n}} \right]$$

$Z_{\left(\frac{a}{2}\right)}$ is set according to our desired degree of confidence and $\sqrt{\frac{p'(1-p')}{n}}$ is the standard deviation of the sampling distribution.

The **sample proportions p' and q' are estimates of the unknown population proportions p and q**. The estimated proportions p' and q' are used because p and q are not known.

Remember that as p moves further from 0.5 the binomial distribution becomes less symmetrical. Because we are estimating the binomial with the symmetrical normal distribution the further away from symmetrical the binomial becomes the less confidence we have in the estimate.

This conclusion can be demonstrated through the following analysis. Proportions are based upon the binomial probability distribution. The possible outcomes are binary, either “success” or “failure”. This gives rise to a proportion, meaning the percentage of the outcomes that are “successes”. It was shown that the binomial distribution could be fully understood if we knew only the probability of a success in any one trial, called p. The mean and the standard deviation of the binomial were found to be:

$$\mu = \text{np}$$
$$\sigma = \sqrt{\text{np}\,q}$$

It was also shown that the binomial could be estimated by the normal distribution if BOTH np AND nq were greater than 5. From the discussion above, it was found that the standardizing formula for the binomial distribution is:

$$Z = \frac{\text{p}' - p}{\sqrt{\left(\frac{pq}{n}\right)}}$$

which is nothing more than a restatement of the general standardizing formula with appropriate substitutions for μ and σ from the binomial. We can use the standard normal distribution, the reason Z is in the equation, because the normal distribution is the limiting distribution of the binomial. This is another example of the Central Limit Theorem. We have already seen that the sampling distribution of means is normally distributed. Recall the extended discussion in The Central Limit Theorem concerning the sampling distribution of proportions and the conclusions of the Central Limit Theorem.

We can now manipulate this formula in just the same way we did for finding the confidence intervals for a mean, but to find the confidence interval for the binomial population parameter, p.

$$p' - Z_{\frac{\alpha}{2}}\sqrt{\frac{p'q'}{n}} \leq p \leq p' + Z_{\frac{\alpha}{2}}\sqrt{\frac{p'q'}{n}}$$

Where p′ = x/n, the point estimate of p taken from the sample. Notice that p′ has replaced p in the formula. This is because we do not know p, indeed, this is just what we are trying to estimate.

Unfortunately, there is no correction factor for cases where the sample size is small so np′ and nq' must always be greater than 5 to develop an interval estimate for p.

EXAMPLE 8.6

Problem

Suppose that a market research firm is hired to estimate the percent of adults living in a large city who have smartphones. Five hundred randomly selected adult residents in this city are surveyed to determine whether they have smartphones. Of the 500 people sampled, 421 responded yes - they own smartphones. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of adult residents of this city who have smartphones.

Solution

- The solution step-by-step.

Let *X* = the number of people in the sample who have smartphones. *X* is binomial: the random variable is binary, people either have a smartphone or they do not.

To calculate the confidence interval, we must find *p′*, *q′*.

n = 500

x = the number of successes in the sample = 421

$p' = \frac{x}{n} = \frac{421}{500} = 0.842$

p′ = 0.842 is the sample proportion; this is the point estimate of the population proportion.

q′ = 1 – *p′* = 1 – 0.842 = 0.158

Since the requested confidence level is *CL* = 0.95, then $\alpha = 1 - CL = 1 - 0.95 = 0.05$ $\left(\frac{\alpha}{2}\right) = 0.025$.

Then $z_{\frac{\alpha}{2}} = z_{0.025} = 1.96$

This can be found using the Standard Normal probability table in Appendix A Statistical Tables. This can also be found in the students t table at the 0.025 column and infinity degrees of freedom because at infinite degrees of freedom the students *t*-distribution becomes the standard normal distribution, Z.

The confidence interval for the true binomial population proportion is

$$\text{p}' - Z_{\alpha}\sqrt{\frac{\text{p}'\text{q}'}{n}} \leq p \leq \text{p}' + Z_{\alpha}\sqrt{\frac{\text{p}'\text{q}'}{n}}$$

$$\text{Substituting in the values from above we find the confidence interval is :} 0.810 \leq p \leq 0.874$$

Interpretation: We estimate with 95% confidence that between 81% and 87.4% of all adult residents of this city have smartphones.

Explanation of 95% Confidence Level: Ninety-five percent of the confidence intervals constructed in this way would contain the true value for the population proportion of all adult residents of this city who have smartphones.

TRY IT 8.6

Suppose 250 randomly selected people are surveyed to determine if they own a tablet. Of the 250 surveyed, 98 reported owning a tablet. Using a 95% confidence level, compute a confidence interval estimate for the true proportion of people who own tablets.

EXAMPLE 8.7

Problem

The Dundee Dog Training School has a larger than average proportion of clients who compete in competitive professional events. A confidence interval for the population proportion of dogs that compete in professional events from 150 different training schools is constructed. The lower limit is determined to be 0.08 and the upper limit is determined to be 0.16. Determine the level of confidence used to construct the interval of the population proportion of dogs that compete in professional events.

Solution

We begin with the formula for a confidence interval for a proportion because the random variable is binary; either the client competes in professional competitive dog events or they don't.

$$p = p' \pm \left[Z_{\left(\frac{a}{2}\right)} \sqrt{\frac{p'(1-p')}{n}} \right]$$

Next we find the sample proportion:

$$p' = \frac{0.08 + 0.16}{2} = 0.12$$

The ± that makes up the confidence interval is thus 0.04; 0.12 + 0.04 = 0.16 and 0.12 - 0.04 = 0.08, the boundaries of the confidence interval. Finally, we solve for *Z*.

$\left[Z \cdot \sqrt{\frac{0.12(1-0.12)}{150}} \right] = 0.04$, ***therefore Z* = 1.51**

And then look up the probability for 1.51 standard deviations on the standard normal table.

$p(Z = 1.51) = 0.4345$, $p(Z) \cdot 2 = 0.8690$ ***or*** 86.90%.

TRY IT 8.7

A student polls their school to see if students in the school district are for or against the new legislation regarding school uniforms. They survey 600 students and finds that 480 are against the new legislation.

a. Compute a 90% confidence interval for the true percent of students who are against the new legislation, and interpret the confidence interval.

b. In a sample of 300 students, 68% said they own an iPod and a smart phone. Compute a 97% confidence interval for the true percent of students who own an iPod and a smartphone.

EXAMPLE 8.8

Problem

A financial officer for a company wants to estimate the percent of accounts receivable that are more than 30 days overdue. They survey 500 accounts and find that 300 are more than 30 days overdue. Compute a 90% confidence interval for the true percent of accounts receivable that are more than 30 days overdue, and interpret the confidence interval.

Solution

- The solution is step-by-step:

$x = 300$ and $n = 500$

$p' = \frac{x}{n} = \frac{300}{500} = 0.600$

$q' = 1 - p' = 1 - 0.600 = 0.400$

Since confidence level = 0.90, then α = 1 – confidence level = (1 – 0.90) = $0.10\left(\frac{\alpha}{2}\right) = 0.05$

$Z_{\frac{\alpha}{2}} = Z_{0.05} = 1.645$

This Z-value can be found using a standard normal probability table. The student's t-table can also be used by entering the table at the 0.05 column and reading at the line for infinite degrees of freedom. The t-distribution is the normal distribution at infinite degrees of freedom. This is a handy trick to remember in finding Z-values for commonly used levels of confidence. We use this formula for a confidence interval for a proportion:

$$\text{p'} - Z_\alpha \sqrt{\frac{\text{p'q'}}{n}} \leq p \leq \text{p'} + Z_\alpha \sqrt{\frac{\text{p'q'}}{n}}$$

Substituting in the values from above we find the confidence interval for the true binomial population proportion is $0.564 \leq \text{p} \leq 0.636$

Interpretation:

- We estimate with 90% confidence that the true percent of all accounts receivable overdue 30 days is between 56.4% and 63.6%.
- Alternate Wording: We estimate with 90% confidence that between 56.4% and 63.6% of ALL accounts are overdue 30 days.

Explanation of 90% Confidence Level: Ninety percent of all confidence intervals constructed in this way contain the true value for the population percent of accounts receivable that are overdue 30 days.

TRY IT 8.8

A student polls his school to see if students in the school district are for or against the new legislation regarding school uniforms. She surveys 600 students and finds that 480 are against the new legislation.

a. Compute a 90% confidence interval for the true percent of students who are against the new legislation, and interpret the confidence interval.

b. In a sample of 300 students, 68% said they own an iPod and a smart phone. Compute a 97% confidence interval for the true percent of students who own an iPod and a smartphone.

8.4 Calculating the Sample Size n: Continuous and Binary Random Variables

Continuous Random Variables

Usually we have no control over the sample size of a data set. However, if we are able to set the sample size, as in cases where we are taking a survey, it is very helpful to know just how large it should be to provide the most information. Sampling can be very costly in both time and product. Simple telephone surveys will cost approximately $30.00 each, for example, and some sampling requires the destruction of the product.

If we go back to our standardizing formula for the sampling distribution for means, we can see that it is possible to solve it for n. If we do this we have $\left(\overline{X} - \mu\right)$ in the denominator.

$$n = \frac{Z_{\frac{\alpha}{2}}{}^2 \sigma^2}{(\overline{X} - \mu)^2} = \frac{Z_{\frac{\alpha}{2}}{}^2 \sigma^2}{e^2}$$

Because we have not taken a sample yet we do not know any of the variables in the formula except that we can set Z_α to the level of confidence we desire just as we did when determining confidence intervals. If we set a predetermined acceptable error, or tolerance, for the difference between $\overline{X}$ and μ, called e in the formula, we are much further in solving for the sample size n. We still do not know the population standard deviation, σ. In practice, a pre-survey is usually done which allows for fine tuning the questionnaire and will give a sample standard deviation that can be used. In other cases, previous information from other surveys may be used for σ in the formula. While crude, this method of determining the sample size may help in reducing cost significantly. It will be the actual data gathered that determines the inferences about the population, so caution in the sample size is appropriate calling for high levels of confidence and small sampling errors.

Binary Random Variables

What was done in cases when looking for the mean of a distribution can also be done when sampling to determine the population parameter p for proportions. Manipulation of the standardizing formula for proportions gives:

$$n = \frac{Z_{\frac{\alpha}{2}}{}^2 pq}{e^2}$$

where e = (p′-p), and is the acceptable sampling error, or tolerance, for this application. This will be measured in percentage points.

In this case the very object of our search is in the formula, p, and of course q because q =1-p. This result occurs because the binomial distribution is a one parameter distribution. If we know p then we know the mean and the standard deviation. Therefore, p shows up in the standard deviation of the sampling distribution which is where we got this formula. If, in an abundance of caution, we substitute 0.5 for p we will draw the largest required sample size that will provide the level of confidence specified by Zα and the tolerance we have selected. This is true because of all combinations of two fractions that add to one, the largest multiple is when each is 0.5. Without any other information concerning the population parameter p, this is the common practice. This may result in oversampling, but certainly not under sampling, thus, this is a cautious approach.

There is an interesting trade-off between the level of confidence and the sample size that shows up here when considering the cost of sampling. Table 8.1 shows the appropriate sample size at different levels of confidence and different level of the acceptable error, or tolerance.

Required sample size (90%)	Required sample size (95%)	Tolerance level
1691	2401	2%
752	1067	3%
271	384	5%
68	96	10%

Table 8.1

This table is designed to show the maximum sample size required at different levels of confidence given an assumed p= 0.5 and q=0.5 as discussed above.

The acceptable error, called tolerance in the table, is measured in plus or minus values from the actual proportion. For example, an acceptable error of 5% means that if the sample proportion was found to be 26 percent, the conclusion would be that the actual population proportion is between 21 and 31 percent with a 90 percent level of confidence if a sample of 271 had been taken. Likewise, if the acceptable error was set at 2%, then the population proportion would be between 24 and 28 percent with a 90 percent level of confidence, but would require that the sample size be increased from 271 to 1,691. If we wished a higher level of confidence, we would require a larger sample size. Moving from a 90 percent level of confidence to a 95 percent level at a plus or minus 5% tolerance requires changing the sample size from 271 to 384. A very common sample size often seen reported in political surveys is 384. With the survey results it is

frequently stated that the results are good to a plus or minus 5% level of "accuracy".

EXAMPLE 8.9

Problem

Suppose a mobile phone company wants to determine the percentage of customers who would upgrade to the latest-version smartphone from their current smartphone. How many customers should the company survey in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of customers who would upgrade to the latest-version smartphone?

Solution

From the problem, we know that the acceptable error, e, is **0.03** (3%=0.03) and $z_{\frac{\alpha}{2}}$ $z_{0.05}$ = 1.645 because the confidence level is 90%. The acceptable error, e, is the difference between the actual population proportion p, and the sample proportion we expect to get from the sample.

However, in order to find n, we need to know the estimated (sample) proportion p'. Remember that q' = 1 – p'. But, we do not know p' yet. Since we multiply p' and q' together, we make them both equal to 0.5 because $p'q'$ = (0.5)(0.5) = 0.25 results in the largest possible product. (Try other products: (0.6)(0.4) = 0.24; (0.3)(0.7) = 0.21; (0.2)(0.8) = 0.16 and so on). The largest possible product gives us the largest n. This gives us a large enough sample so that we can be 90% confident that we are within three percentage points of the true population proportion. To calculate the sample size n, use the formula and make the substitutions.

$n = \frac{z^2 p' q'}{e^2}$ gives $n = \frac{1.645^2 (0.5)(0.5)}{0.03^2} = 751.7$

Round the answer to the next higher value. The sample size should be 752 smartphone customers in order to be 90% confident that the estimated (sample) proportion is within three percentage points of the true population proportion of all customers who would upgrade their smartphone.

TRY IT 8.9

Suppose an internet marketing company wants to determine the current percentage of customers who click on ads on their smartphones. How many customers should the company survey in order to be 90% confident that the estimated proportion is within five percentage points of the true population proportion of customers who click on ads on their smartphones?