This is “Estimation”, chapter 7 from the book Beginning Statistics (index.html) (v. 1.0).

Chapter 7

Estimation

If we wish to estimate the mean μ of a population for which a census is impractical, say the average height of all 18-year-old men in the country, a reasonable strategy is to take a sample, compute its mean $\overline{x}$, and estimate the unknown number μ by the known number $\overline{x}$. For example, if the average height of 100 randomly selected men aged 18 is 70.6 inches, then we would say that the average height of all 18-year-old men is (at least approximately) 70.6 inches.

Estimating a population parameter by a single number like this is called **point estimation**; in the case at hand the statistic $\overline{x}$ is a **point estimate** of the parameter μ. The terminology arises because a single number corresponds to a single point on the number line.

A problem with a point estimate is that it gives no indication of how reliable the estimate is. In contrast, in this chapter we learn about **interval estimation**. In brief, in the case of estimating a population mean μ we use a formula to compute from the data a number E, called the **margin of error**[1] of the estimate, and form the interval $[\overline{x} - E, \overline{x} + E]$ .We do this in such a way that a certain proportion, say 95%, of all the intervals constructed from sample data by means of this formula contain the unknown parameter μ. Such an interval is called a **95% confidence interval**[2] **for** μ.

Continuing with the example of the average height of 18-year-old men, suppose that the sample of 100 men mentioned above for which $\overline{x} = 70.6$ inches also had sample standard deviation s = 1.7 inches. It then turns out that E = 0.33 and we would state that we are 95% confident that the average height of all 18-year-old men is in the interval formed by 70.6 ± 0.33 inches, that is, the average is between 70.27 and 70.93 inches. If the sample statistics had come from a smaller sample, say a sample of 50 men, the lower reliability would show up in the 95% confidence interval being longer, hence less precise in its estimate. In this example the 95% confidence interval for the same sample statistics but with n = 50 is 70.6 ± 0.47 inches, or from 70.13 to 71.07 inches.

1. E, the number added to and subtracted from the point estimate to produce the interval estimate.

2. An interval with endpoints $\overline{x} \pm E$, computed from the sample data in such a way that a specified proportion of all intervals constructed by this process will contain the parameter of interest.

7.1 Large Sample Estimation of a Population Mean

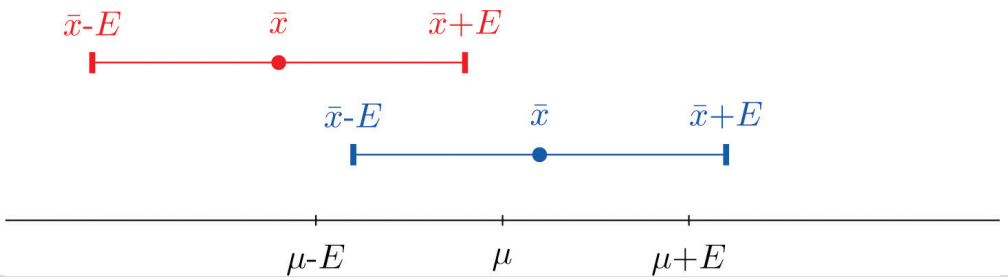
LEARNING OBJECTIVES

1. To become familiar with the concept of an interval estimate of the population mean.
2. To understand how to apply formulas for a confidence interval for a population mean.

The Central Limit Theorem says that, for large samples (samples of size $n \geq 30$), when viewed as a random variable the sample mean $\overline{X}$ is normally distributed with mean $\mu_{\overline{X}} = \mu$ and standard deviation $\sigma_{\overline{X}} = \sigma \Big/ \sqrt{n}$. The Empirical Rule says that we must go about two standard deviations from the mean to capture 95% of the values of $\overline{X}$ generated by sample after sample. A more precise distance based on the normality of $\overline{X}$ is 1.960 standard deviations, which is $E = 1.960\sigma \Big/ \sqrt{n}$.

The key idea in the construction of the 95% confidence interval is this, as illustrated in Figure 7.1 "When Winged Dots Capture the Population Mean": because in sample after sample 95% of the values of $\overline{X}$ lie in the interval $\left[\mu - E, \mu + E\right]$, if we adjoin to each side of the point estimate $\overline{x}$ a "wing" of length E, 95% of the intervals formed by the winged dots contain μ. The 95% confidence interval is thus $\overline{x} \pm 1.960\sigma \Big/ \sqrt{n}$. For a different **level of confidence**[3], say 90% or 99%, the number 1.960 will change, but the idea is the same.

Figure 7.1 *When Winged Dots Capture the Population Mean*

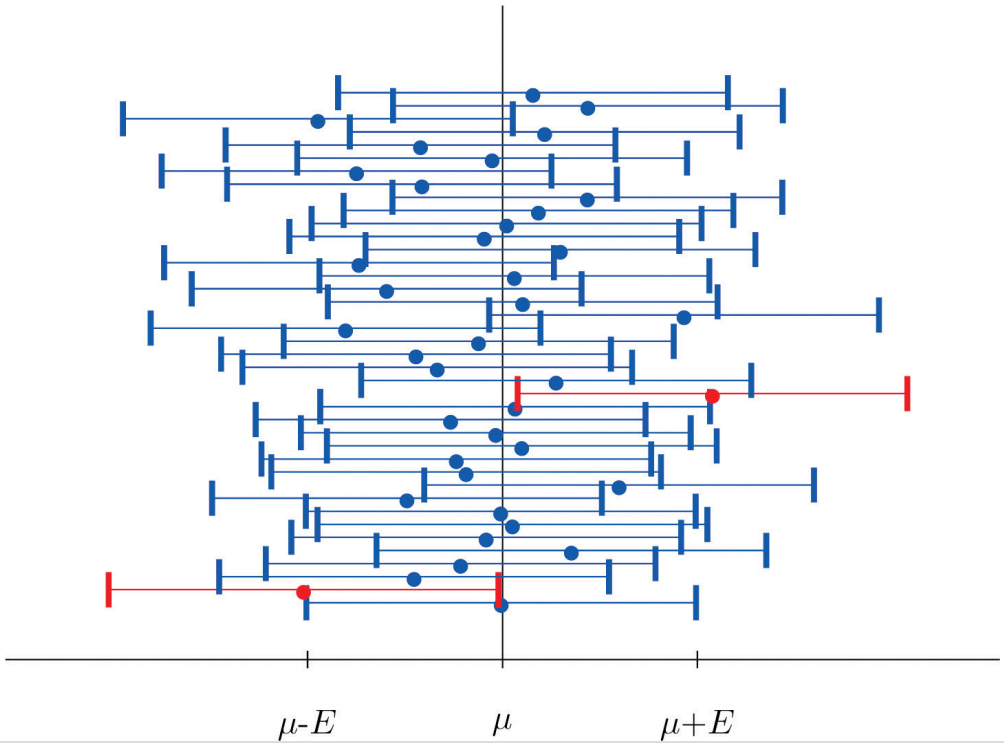


3. The proportion of confidence intervals which, if under repeated random sampling were always constructed according to the formula of the text, would contain the parameter of interest.

Figure 7.2 "Computer Simulation of 40 95% Confidence Intervals for a Mean" shows the intervals generated by a computer simulation of drawing 40 samples from a normally distributed population and constructing the 95% confidence interval for

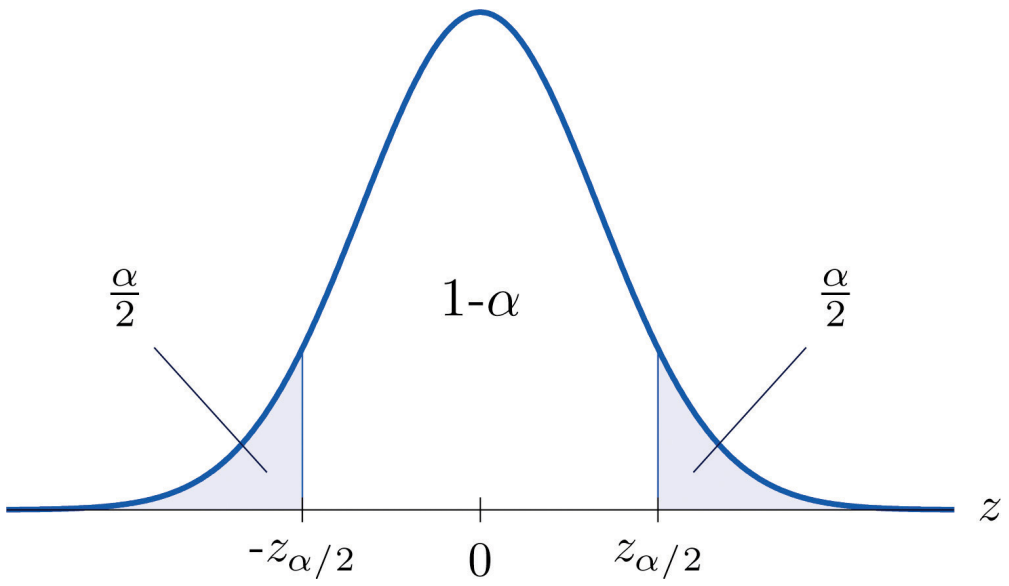
each one. We expect that about $(0.05)\ (40) = 2$of the intervals so constructed would fail to contain the population mean μ, and in this simulation two of the intervals, shown in red, do.

Figure 7.2 *Computer Simulation of 40 95% Confidence Intervals for a Mean*



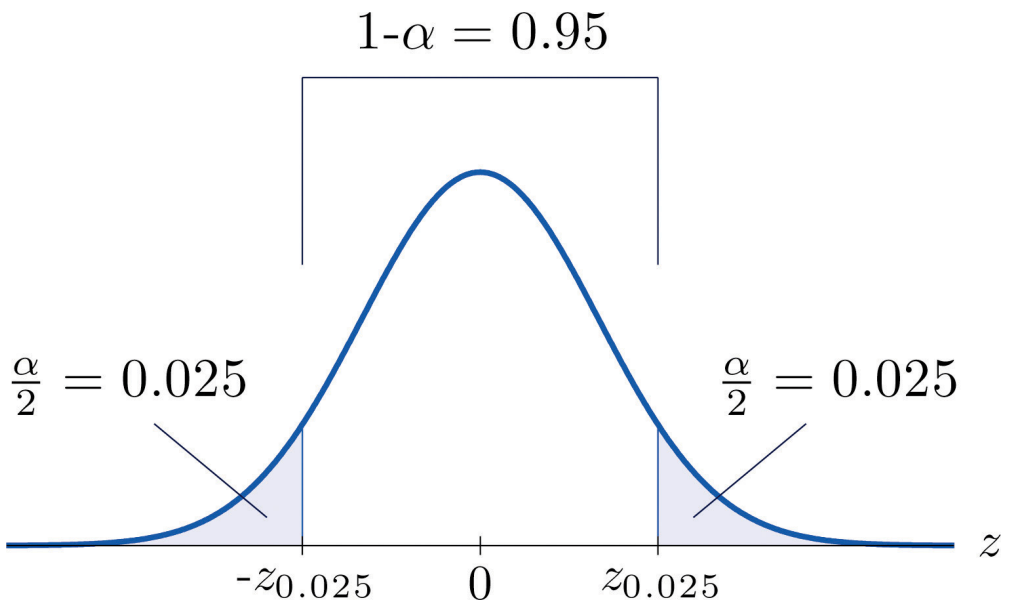
It is standard practice to identify the level of confidence in terms of the area α in the two tails of the distribution of $\overline{X}$ when the middle part specified by the level of confidence is taken out. This is shown in Figure 7.3, drawn for the general situation, and in Figure 7.4, drawn for 95% confidence. Remember from Section 5.4.1 "Tails of the Standard Normal Distribution" in Chapter 5 "Continuous Random Variables" that the z-value that cuts off a right tail of area c is denoted z_c. Thus the number 1.960 in the example is $z_{.025}$, which is $z_{\alpha/2}$ for $\alpha = 1 - 0.95 = 0.05$.

Figure 7.3



For $100\,(1-\alpha)$*% confidence the area in each tail is* $\alpha\,/\,2$.

Figure 7.4



For 95% confidence the area in each tail is $\alpha\,/\,2 = 0.025$.

The level of confidence can be any number between 0 and 100%, but the most common values are probably 90% ($\alpha = 0.10$), 95% ($\alpha = 0.05$), and 99% ($\alpha = 0.01$).

Thus in general for a $100\,(1-\alpha)$% confidence interval, $E = z_{\alpha/2}\left(\sigma \,/\, \sqrt{n}\right)$, so the formula for the confidence interval is $\overline{x} \pm z_{\alpha/2}\left(\sigma \,/\, \sqrt{n}\right)$. While sometimes the population standard deviation σ is known, typically it is not. If not, for $n \geq 30$ it is generally safe to approximate σ by the sample standard deviation s.

Large Sample $100\,(1-\alpha)\,\%$ Confidence Interval for a Population Mean

If σ is known: $\overline{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$

If σ is unknown: $\overline{x} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$

A sample is considered large when $n \geq 30$.

As mentioned earlier, the number $E = z_{\alpha/2}\sigma \,/\, \sqrt{n}$ or $E = z_{\alpha/2}s \,/\, \sqrt{n}$ is called the *margin of error* of the estimate.

EXAMPLE 1

Find the number $z_{\alpha/2}$ needed in construction of a confidence interval:

a. when the level of confidence is 90%;
b. when the level of confidence is 99%.

Solution:

a. For confidence level 90%, $\alpha = 1 - 0.90 = 0.10$, so $z_{\alpha/2} = z_{0.05}$. The procedure for finding this number was given in Section 5.4.1 "Tails of the Standard Normal Distribution". Since the area under the standard normal curve to the right of $z_{.05}$ is 0.05, the area to the left of $z_{0.05}$ is 0.95. We search for the area 0.9500 in Figure 12.2 "Cumulative Normal Probability". The closest entries in the table are 0.9495 and 0.9505, corresponding to z-values 1.64 and 1.65. Since 0.95 is exactly halfway between 0.9495 and 0.9505 we use the average 1.645 of the z-values for $z_{0.05}$.
b. For confidence level 99%, $\alpha = 1 - 0.99 = 0.01$, so $z_{\alpha/2} = z_{0.005}$. Since the area under the standard normal curve to the right of $z_{0.005}$ is 0.005, the area to the left of $z_{0.005}$ is 0.9950. We search for the area 0.9950 in Figure 12.2 "Cumulative Normal Probability". The closest entries in the table are 0.9949 and 0.9951, corresponding to z-values 2.57 and 2.58. Since 0.995 is halfway between 0.9949 and 0.9951 we use the average 2.575 of the z-values for $z_{0.005}$.

EXAMPLE 2

Use Figure 12.3 "Critical Values of " to find the number $z_{\alpha/2}$ needed in construction of a confidence interval:

a. when the level of confidence is 90%;
b. when the level of confidence is 99%.

Solution:

a. In the next section we will learn about a continuous random variable that has a probability distribution called the Student t-distribution. Figure 12.3 "Critical Values of " gives the value t_c that cuts off a right tail of area c for different values of c. The last line of that table, the one whose heading is the symbol ∞ for infinity and $[z]$, gives the corresponding z-value z_c that cuts off a right tail of the same area c. In particular, $z_{0.05}$ is the number in that row and in the column with the heading $t_{0.05}$. We read off directly that $z_{0.05} = 1.645$.
b. In Figure 12.3 "Critical Values of " $z_{0.005}$ is the number in the last row and in the column headed $t_{0.005}$, namely 2.576.

Figure 12.3 "Critical Values of " can be used to find z_c only for those values of c for which there is a column with the heading t_c appearing in the table; otherwise we must use Figure 12.2 "Cumulative Normal Probability" in reverse. But when it can be done it is both faster and more accurate to use the last line of Figure 12.3 "Critical Values of " to find z_c than it is to do so using Figure 12.2 "Cumulative Normal Probability" in reverse.

EXAMPLE 3

A sample of size 49 has sample mean 35 and sample standard deviation 14. Construct a 98% confidence interval for the population mean using this information. Interpret its meaning.

Solution:

For confidence level 98%, $\alpha = 1 - 0.98 = 0.02$, so $z_{\alpha/2} = z_{0.01}$. From Figure 12.3 "Critical Values of " we read directly that $z_{0.01} = 2.326$. Thus

$$\overline{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} = 35 \pm 2.326 \left(\frac{14}{\sqrt{49}} \right) = 35 \pm 4.652 \approx 35 \pm 4.7$$

We are 98% confident that the population mean μ lies in the interval $[30.3, 39.7]$, in the sense that in repeated sampling 98% of all intervals constructed from the sample data in this manner will contain μ.

EXAMPLE 4

A random sample of 120 students from a large university yields mean GPA 2.71 with sample standard deviation 0.51. Construct a 90% confidence interval for the mean GPA of all students at the university.

Solution:

For confidence level 90%, $\alpha = 1 - 0.90 = 0.10$, so $z_{\alpha/2} = z_{0.05}$. From Figure 12.3 "Critical Values of " we read directly that $z_{0.05} = 1.645$. Since $n = 120$, $\overline{x} = 2.71$, and $s = 0.51$,

$$\overline{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} = 2.71 \pm 1.645 \left(\frac{0.51}{\sqrt{120}} \right) = 2.71 \pm 0.0766$$

One may be 90% confident that the true average GPA of all students at the university is contained in the interval
$(2.71 - 0.08, 2.71 + 0.08) = (2.63, 2.79)$.

KEY TAKEAWAYS

- A confidence interval for a population mean is an estimate of the population mean together with an indication of reliability.
- There are different formulas for a confidence interval based on the sample size and whether or not the population standard deviation is known.
- The confidence intervals are constructed entirely from the sample data (or sample data and the population standard deviation, when it is known).

EXERCISES

BASIC

1. A random sample is drawn from a population of known standard deviation 11.3. Construct a 90% confidence interval for the population mean based on the information given (not all of the information given need be used).

   a. $n = 36, \overline{x} = 105.2, s = 11.2$
   b. $n = 100, \overline{x} = 105.2, s = 11.2$

2. A random sample is drawn from a population of known standard deviation 22.1. Construct a 95% confidence interval for the population mean based on the information given (not all of the information given need be used).

   a. $n = 121, \overline{x} = 82.4, s = 21.9$
   b. $n = 81, \overline{x} = 82.4, s = 21.9$

3. A random sample is drawn from a population of unknown standard deviation. Construct a 99% confidence interval for the population mean based on the information given.

   a. $n = 49, \overline{x} = 17.1, s = 2.1$
   b. $n = 169, \overline{x} = 17.1, s = 2.1$

4. A random sample is drawn from a population of unknown standard deviation. Construct a 98% confidence interval for the population mean based on the information given.

   a. $n = 225, \overline{x} = 92.0, s = 8.4$
   b. $n = 64, \overline{x} = 92.0, s = 8.4$

5. A random sample of size 144 is drawn from a population whose distribution, mean, and standard deviation are all unknown. The summary statistics are $\overline{x} = 58.2$ and $s = 2.6$.

   a. Construct an 80% confidence interval for the population mean μ.
   b. Construct a 90% confidence interval for the population mean μ.
   c. Comment on why one interval is longer than the other.

6. A random sample of size 256 is drawn from a population whose distribution, mean, and standard deviation are all unknown. The summary statistics are $\overline{x} = 1011$ and $s = 34$.

   a. Construct a 90% confidence interval for the population mean μ.

b. Construct a 99% confidence interval for the population mean μ.
c. Comment on why one interval is longer than the other.

APPLICATIONS

7. A government agency was charged by the legislature with estimating the length of time it takes citizens to fill out various forms. Two hundred randomly selected adults were timed as they filled out a particular form. The times required had mean 12.8 minutes with standard deviation 1.7 minutes. Construct a 90% confidence interval for the mean time taken for all adults to fill out this form.

8. Four hundred randomly selected working adults in a certain state, including those who worked at home, were asked the distance from their home to their workplace. The average distance was 8.84 miles with standard deviation 2.70 miles. Construct a 99% confidence interval for the mean distance from home to work for all residents of this state.

9. On every passenger vehicle that it tests an automotive magazine measures, at true speed 55 mph, the difference between the true speed of the vehicle and the speed indicated by the speedometer. For 36 vehicles tested the mean difference was –1.2 mph with standard deviation 0.2 mph. Construct a 90% confidence interval for the mean difference between true speed and indicated speed for all vehicles.

10. A corporation monitors time spent by office workers browsing the web on their computers instead of working. In a sample of computer records of 50 workers, the average amount of time spent browsing in an eight-hour work day was 27.8 minutes with standard deviation 8.2 minutes. Construct a 99.5% confidence interval for the mean time spent by all office workers in browsing the web in an eight-hour day.

11. A sample of 250 workers aged 16 and older produced an average length of time with the current employer ("job tenure") of 4.4 years with standard deviation 3.8 years. Construct a 99.9% confidence interval for the mean job tenure of all workers aged 16 or older.

12. The amount of a particular biochemical substance related to bone breakdown was measured in 30 healthy women. The sample mean and standard deviation were 3.3 nanograms per milliliter (ng/mL) and 1.4 ng/mL. Construct an 80% confidence interval for the mean level of this substance in all healthy women.

13. A corporation that owns apartment complexes wishes to estimate the average length of time residents remain in the same apartment before moving out. A

sample of 150 rental contracts gave a mean length of occupancy of 3.7 years with standard deviation 1.2 years. Construct a 95% confidence interval for the mean length of occupancy of apartments owned by this corporation.

14. The designer of a garbage truck that lifts roll-out containers must estimate the mean weight the truck will lift at each collection point. A random sample of 325 containers of garbage on current collection routes yielded $\overline{x} = 75.3$ lb, s = 12.8 lb. Construct a 99.8% confidence interval for the mean weight the trucks must lift each time.

15. In order to estimate the mean amount of damage sustained by vehicles when a deer is struck, an insurance company examined the records of 50 such occurrences, and obtained a sample mean of $2,785 with sample standard deviation $221. Construct a 95% confidence interval for the mean amount of damage in all such accidents.

16. In order to estimate the mean FICO credit score of its members, a credit union samples the scores of 95 members, and obtains a sample mean of 738.2 with sample standard deviation 64.2. Construct a 99% confidence interval for the mean FICO score of all of its members.

ADDITIONAL EXERCISES

17. For all settings a packing machine delivers a precise amount of liquid; the amount dispensed always has standard deviation 0.07 ounce. To calibrate the machine its setting is fixed and it is operated 50 times. The mean amount delivered is 6.02 ounces with sample standard deviation 0.04 ounce. Construct a 99.5% confidence interval for the mean amount delivered at this setting. Hint: Not all the information provided is needed.

18. A power wrench used on an assembly line applies a precise, preset amount of torque; the torque applied has standard deviation 0.73 foot-pound at every torque setting. To check that the wrench is operating within specifications it is used to tighten 100 fasteners. The mean torque applied is 36.95 foot-pounds with sample standard deviation 0.62 foot-pound. Construct a 99.9% confidence interval for the mean amount of torque applied by the wrench at this setting. Hint: Not all the information provided is needed.

19. The number of trips to a grocery store per week was recorded for a randomly selected collection of households, with the results shown in the table.

2	2	2	1	4	2	3	2	5	4
2	3	5	0	3	2	3	1	4	3
3	2	1	6	2	3	3	2	4	4

Construct a 95% confidence interval for the average number of trips to a grocery store per week of all households.

20. For each of 40 high school students in one county the number of days absent from school in the previous year were counted, with the results shown in the frequency table.

x	0	1	2	3	4	5
f	24	7	5	2	1	1

Construct a 90% confidence interval for the average number of days absent from school of all students in the county.

21. A town council commissioned a random sample of 85 households to estimate the number of four-wheel vehicles per household in the town. The results are shown in the following frequency table.

x	0	1	2	3	4	5
f	1	16	28	22	12	6

Construct a 98% confidence interval for the average number of four-wheel vehicles per household in the town.

22. The number of hours per day that a television set was operating was recorded for a randomly selected collection of households, with the results shown in the table.

3.7	4.2	1.5	3.6	5.9
4.7	8.2	3.9	2.5	4.4
2.1	3.6	1.1	7.3	4.2
3.0	3.8	2.2	4.2	3.8
4.3	2.1	2.4	6.0	3.7
2.5	1.3	2.8	3.0	5.6

Construct a 99.8% confidence interval for the mean number of hours that a television set is in operation in all households.

LARGE DATA SET EXERCISES

23. Large Data Set 1 records the SAT scores of 1,000 students. Regarding it as a random sample of all high school students, use it to construct a 99% confidence interval for the mean SAT score of all students.

http://www.gone.2012books.lardbucket.org/sites/all/files/data1.xls

24. Large Data Set 1 records the GPAs of 1,000 college students. Regarding it as a random sample of all college students, use it to construct a 95% confidence interval for the mean GPA of all students.

    http://www.gone.2012books.lardbucket.org/sites/all/files/data1.xls

25. Large Data Set 1 lists the SAT scores of 1,000 students.

    http://www.gone.2012books.lardbucket.org/sites/all/files/data1.xls

    a. Regard the data as arising from a census of all students at a high school, in which the SAT score of every student was measured. Compute the population mean μ.
    b. Regard the first 36 students as a random sample and use it to construct a 99% confidence for the mean μ of all 1,000 SAT scores. Does it actually capture the mean μ?

26. Large Data Set 1 lists the GPAs of 1,000 students.

    http://www.gone.2012books.lardbucket.org/sites/all/files/data1.xls

    a. Regard the data as arising from a census of all freshman at a small college at the end of their first academic year of college study, in which the GPA of every such person was measured. Compute the population mean μ.
    b. Regard the first 36 students as a random sample and use it to construct a 95% confidence for the mean μ of all 1,000 GPAs. Does it actually capture the mean μ?

ANSWERS

1. a. 105.2 ± 3.10
   b. 105.2 ± 1.86

3. a. 17.1 ± 0.77
   b. 17.1 ± 0.42

5. a. 58.2 ± 0.28
   b. 58.2 ± 0.36
   c. Asking for greater confidence requires a longer interval.

7. 12.8 ± 0.20

9. -1.2 ± 0.05

11. 4.4 ± 0.79

13. 3.7 ± 0.19

15. 2785 ± 61

17. 6.02 ± 0.03

19. 2.8 ± 0.48

21. 2.54 ± 0.30

23. $(1511.43, 1546.05)$

25. a. $\mu = 1528.74$
    b. $(1428.22, 1602.89)$

7.2 Small Sample Estimation of a Population Mean

LEARNING OBJECTIVES

1. To become familiar with Student's *t*-distribution.
2. To understand how to apply additional formulas for a confidence interval for a population mean.

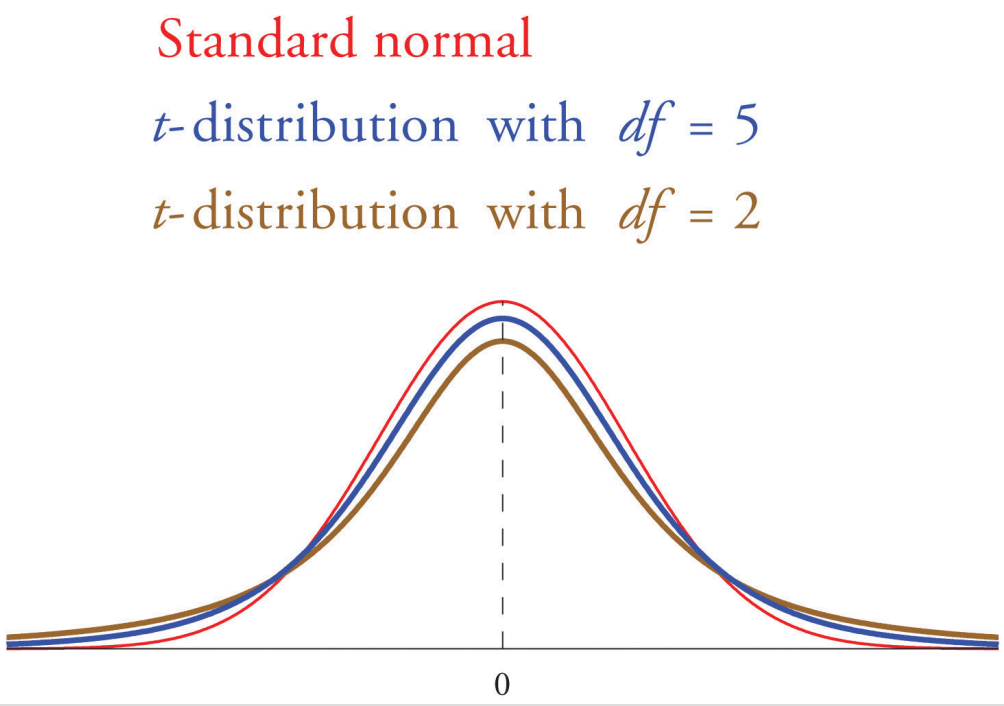
The confidence interval formulas in the previous section are based on the Central Limit Theorem, the statement that for large samples $\overline{X}$ is normally distributed with mean μ and standard deviation $\sigma \,/\, \sqrt{n}$. When the population mean μ is estimated with a small sample ($n < 30$), the Central Limit Theorem does not apply. In order to proceed we assume that the numerical population from which the sample is taken has a normal distribution to begin with. If this condition is satisfied then when the population standard deviation σ is known the old formula $\overline{x} \pm z_{\alpha/2}\left(\sigma \,/\, \sqrt{n}\right)$ can still be used to construct a $100\,(1-\alpha)\%$ confidence interval for μ.

If the population standard deviation is unknown and the sample size n is small then when we substitute the sample standard deviation s for σ the normal approximation is no longer valid. The solution is to use a different distribution, called **Student's *t*-distribution**[4] **with $n-1$ degrees of freedom**[5]. Student's *t*-distribution is very much like the standard normal distribution in that it is centered at 0 and has the same qualitative bell shape, but it has heavier tails than the standard normal distribution does, as indicated by Figure 7.5 "Student's ", in which the curve (in brown) that meets the dashed vertical line at the lowest point is the *t*-distribution with two degrees of freedom, the next curve (in blue) is the *t*-distribution with five degrees of freedom, and the thin curve (in red) is the standard normal distribution. As also indicated by the figure, as the sample size n increases, Student's *t*-distribution ever more closely resembles the standard normal distribution. Although there is a different *t*-distribution for every value of n, once the sample size is 30 or more it is typically acceptable to use the standard normal distribution instead, as we will always do in this text.

4. A distribution of a continuous random variable that resembles that standard normal distribution but has heavier tails.

5. A number that specifies a particular *t*-distribution and that is computed based on the sample size.

Figure 7.5 *Student's t-Distribution*



Just as the symbol z_c stands for the value that cuts off a right tail of area c in the standard normal distribution, so the symbol t_c stands for the value that cuts off a right tail of area c in the standard normal distribution. This gives us the following confidence interval formulas.

Small Sample $100\,(1-\alpha)\,\%$ Confidence Interval for a Population Mean

If σ is known: $\overline{x} \pm z_{\alpha/2}\left(\frac{\sigma}{\sqrt{n}}\right)$

If σ is unknown: $\overline{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$ (degrees of freedom $df = n-1$)

The population must be normally distributed.

A sample is considered small when $n < 30$.

To use the new formula we use the line in Figure 12.3 "Critical Values of " that corresponds to the relevant sample size.

EXAMPLE 5

A sample of size 15 drawn from a normally distributed population has sample mean 35 and sample standard deviation 14. Construct a 95% confidence interval for the population mean, and interpret its meaning.

Solution:

Since the population is normally distributed, the sample is small, and the population standard deviation is unknown, the formula that applies is

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

Confidence level 95% means that $\alpha = 1 - 0.95 = 0.05$ so $\alpha\ /\ 2 = 0.025.$ Since the sample size is $n = 15$, there are $n-1 = 14$ degrees of freedom. By Figure 12.3 "Critical Values of " $t_{0.025} = 2.145.$ Thus

$$\overline{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) = 35 \pm 2.145\left(\frac{14}{\sqrt{15}}\right) = 35 \pm 7.8$$

One may be 95% confident that the true value of μ is contained in the interval $\left(35 - 7.\,8,35 + 7.8\right) = \ (27.\,2,42.\,8)\,.$

EXAMPLE 6

A random sample of 12 students from a large university yields mean GPA 2.71 with sample standard deviation 0.51. Construct a 90% confidence interval for the mean GPA of all students at the university. Assume that the numerical population of GPAs from which the sample is taken has a normal distribution.

Solution:

Since the population is normally distributed, the sample is small, and the population standard deviation is unknown, the formula that applies is

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

Confidence level 90% means that $\alpha = 1 - 0.90 = 0.10$ so $\alpha / 2 = 0.05$. Since the sample size is $n = 12$, there are $n-1 = 11$ degrees of freedom. By Figure 12.3 "Critical Values of " $t_{0.05} = 1.796$. Thus

$$\overline{x} \pm t_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right) = 2.71 \pm 1.796\left(\frac{0.51}{\sqrt{12}}\right) = 2.71 \pm 0.26$$

One may be 90% confident that the true average GPA of all students at the university is contained in the interval $(2.71 - 0.26, 2.71 + 0.26) = (2.45, 2.97)$.

Compare Note 7.9 "Example 4" in Section 7.1 "Large Sample Estimation of a Population Mean" and Note 7.16 "Example 6". The summary statistics in the two samples are the same, but the 90% confidence interval for the average GPA of all students at the university in Note 7.9 "Example 4" in Section 7.1 "Large Sample Estimation of a Population Mean", $(2.63, 2.79)$, is shorter than the 90% confidence interval $(2.45, 2.97)$, in Note 7.16 "Example 6". This is partly because in Note 7.9 "Example 4" the sample size is larger; there is more information pertaining to the true value of μ in the large data set than in the small one.

KEY TAKEAWAYS

- In selecting the correct formula for construction of a confidence interval for a population mean ask two questions: is the population standard deviation σ known or unknown, and is the sample large or small?
- We can construct confidence intervals with small samples only if the population is normal.

EXERCISES

BASIC

1. A random sample is drawn from a normally distributed population of known standard deviation 5. Construct a 99.8% confidence interval for the population mean based on the information given (not all of the information given need be used).

   a. $n = 16, \overline{x} = 98, s = 5.6$
   b. $n = 9, \overline{x} = 98, s = 5.6$

2. A random sample is drawn from a normally distributed population of known standard deviation 10.7. Construct a 95% confidence interval for the population mean based on the information given (not all of the information given need be used).

   a. $n = 25, \overline{x} = 103.3, s = 11.0$
   b. $n = 4, \overline{x} = 103.3, s = 11.0$

3. A random sample is drawn from a normally distributed population of unknown standard deviation. Construct a 99% confidence interval for the population mean based on the information given.

   a. $n = 18, \overline{x} = 386, s = 24$
   b. $n = 7, \overline{x} = 386, s = 24$

4. A random sample is drawn from a normally distributed population of unknown standard deviation. Construct a 98% confidence interval for the population mean based on the information given.

   a. $n = 8, \overline{x} = 58.3, s = 4.1$
   b. $n = 27, \overline{x} = 58.3, s = 4.1$

5. A random sample of size 14 is drawn from a normal population. The summary statistics are $\overline{x} = 933$ and $s = 18$.

   a. Construct an 80% confidence interval for the population mean μ.
   b. Construct a 90% confidence interval for the population mean μ.
   c. Comment on why one interval is longer than the other.

6. A random sample of size 28 is drawn from a normal population. The summary statistics are $\overline{x} = 68.6$ and $s = 1.28$.

   a. Construct a 95% confidence interval for the population mean μ.

b. Construct a 99.5% confidence interval for the population mean μ.
c. Comment on why one interval is longer than the other.

APPLICATIONS

7. City planners wish to estimate the mean lifetime of the most commonly planted trees in urban settings. A sample of 16 recently felled trees yielded mean age 32.7 years with standard deviation 3.1 years. Assuming the lifetimes of all such trees are normally distributed, construct a 99.8% confidence interval for the mean lifetime of all such trees.

8. To estimate the number of calories in a cup of diced chicken breast meat, the number of calories in a sample of four separate cups of meat is measured. The sample mean is 211.8 calories with sample standard deviation 0.9 calorie. Assuming the caloric content of all such chicken meat is normally distributed, construct a 95% confidence interval for the mean number of calories in one cup of meat.

9. A college athletic program wishes to estimate the average increase in the total weight an athlete can lift in three different lifts after following a particular training program for six weeks. Twenty-five randomly selected athletes when placed on the program exhibited a mean gain of 47.3 lb with standard deviation 6.4 lb. Construct a 90% confidence interval for the mean increase in lifting capacity all athletes would experience if placed on the training program. Assume increases among all athletes are normally distributed.

10. To test a new tread design with respect to stopping distance, a tire manufacturer manufactures a set of prototype tires and measures the stopping distance from 70 mph on a standard test car. A sample of 25 stopping distances yielded a sample mean 173 feet with sample standard deviation 8 feet. Construct a 98% confidence interval for the mean stopping distance for these tires. Assume a normal distribution of stopping distances.

11. A manufacturer of chokes for shotguns tests a choke by shooting 15 patterns at targets 40 yards away with a specified load of shot. The mean number of shot in a 30-inch circle is 53.5 with standard deviation 1.6. Construct an 80% confidence interval for the mean number of shot in a 30-inch circle at 40 yards for this choke with the specified load. Assume a normal distribution of the number of shot in a 30-inch circle at 40 yards for this choke.

12. In order to estimate the speaking vocabulary of three-year-old children in a particular socioeconomic class, a sociologist studies the speech of four children. The mean and standard deviation of the sample are $\overline{x} = 1120$ and $s = 215$ words. Assuming that speaking vocabularies are normally distributed,

construct an 80% confidence interval for the mean speaking vocabulary of all three-year-old children in this socioeconomic group.

13. A thread manufacturer tests a sample of eight lengths of a certain type of thread made of blended materials and obtains a mean tensile strength of 8.2 lb with standard deviation 0.06 lb. Assuming tensile strengths are normally distributed, construct a 90% confidence interval for the mean tensile strength of this thread.

14. An airline wishes to estimate the weight of the paint on a fully painted aircraft of the type it flies. In a sample of four repaintings the average weight of the paint applied was 239 pounds, with sample standard deviation 8 pounds. Assuming that weights of paint on aircraft are normally distributed, construct a 99.8% confidence interval for the mean weight of paint on all such aircraft.

15. In a study of dummy foal syndrome, the average time between birth and onset of noticeable symptoms in a sample of six foals was 18.6 hours, with standard deviation 1.7 hours. Assuming that the time to onset of symptoms in all foals is normally distributed, construct a 90% confidence interval for the mean time between birth and onset of noticeable symptoms.

16. A sample of 26 women's size 6 dresses had mean waist measurement 25.25 inches with sample standard deviation 0.375 inch. Construct a 95% confidence interval for the mean waist measurement of all size 6 women's dresses. Assume waist measurements are normally distributed.

ADDITIONAL EXERCISES

17. Botanists studying attrition among saplings in new growth areas of forests diligently counted stems in six plots in five-year-old new growth areas, obtaining the following counts of stems per acre:

    9,432 11,026 10,539
    8,773 9,868 10,247

    Construct an 80% confidence interval for the mean number of stems per acre in all five-year-old new growth areas of forests. Assume that the number of stems per acre is normally distributed.

18. Nutritionists are investigating the efficacy of a diet plan designed to increase the caloric intake of elderly people. The increase in daily caloric intake in 12 individuals who are put on the plan is (a minus sign signifies that calories consumed went down):

$$\begin{array}{cccccc} 121 & 284 & -94 & 295 & 183 & 312 \\ 188 & -102 & 259 & 226 & 152 & 167 \end{array}$$

Construct a 99.8% confidence interval for the mean increase in caloric intake for all people who are put on this diet. Assume that population of differences in intake is normally distributed.

19. A machine for making precision cuts in dimension lumber produces studs with lengths that vary with standard deviation 0.003 inch. Five trial cuts are made to check the machine's calibration. The mean length of the studs produced is 104.998 inches with sample standard deviation 0.004 inch. Construct a 99.5% confidence interval for the mean lengths of all studs cut by this machine. Assume lengths are normally distributed. Hint: Not all the numbers given in the problem are used.

20. The variation in time for a baked good to go through a conveyor oven at a large scale bakery has standard deviation 0.017 minute at every time setting. To check the bake time of the oven periodically four batches of goods are carefully timed. The recent check gave a mean of 27.2 minutes with sample standard deviation 0.012 minute. Construct a 99.8% confidence interval for the mean bake time of all batches baked in this oven. Assume bake times are normally distributed. Hint: Not all the numbers given in the problem are used.

21. Wildlife researchers tranquilized and weighed three adult male polar bears. The data (in pounds) are: 926, 742, 1,109. Assume the weights of all bears are normally distributed.

    a. Construct an 80% confidence interval for the mean weight of all adult male polar bears using these data.
    b. Convert the three weights in pounds to weights in kilograms using the conversion 1 lb = 0.453 kg (so the first datum changes to $(926)\ (0.453) = 419$). Use the converted data to construct an 80% confidence interval for the mean weight of all adult male polar bears expressed in kilograms.
    c. Convert your answer in part (a) into kilograms directly and compare it to your answer in (b). This illustrates that if you construct a confidence interval in one system of units you can convert it directly into another system of units without having to convert all the data to the new units.

22. Wildlife researchers trapped and measured six adult male collared lemmings. The data (in millimeters) are: 104, 99, 112, 115, 96, 109. Assume the lengths of all lemmings are normally distributed.

    a. Construct a 90% confidence interval for the mean length of all adult male collared lemmings using these data.

b. Convert the six lengths in millimeters to lengths in inches using the conversion 1 mm = 0.039 in (so the first datum changes to (104)(0.039) = 4.06). Use the converted data to construct a 90% confidence interval for the mean length of all adult male collared lemmings expressed in inches.
c. Convert your answer in part (a) into inches directly and compare it to your answer in (b). This illustrates that if you construct a confidence interval in one system of units you can convert it directly into another system of units without having to convert all the data to the new units.

ANSWERS

1. a. 98 ± 3.9
   b. 98 ± 5.2
3. a. 386 ± 16.4
   b. 386 ± 33.6
5. a. 933 ± 6.5
   b. 933 ± 8.5
   c. Asking for greater confidence requires a longer interval.
7. 32.7 ± 2.9
9. 47.3 ± 2.19
11. 53.5 ± 0.56
13. 8.2 ± 0.04
15. 18.6 ± 1.4
17. 9981 ± 486
19. 104.998 ± 0.004
21. a. 926 ± 200
    b. 419 ± 90
    c. 419 ± 91

7.3 Large Sample Estimation of a Population Proportion

LEARNING OBJECTIVE

1. To understand how to apply the formula for a confidence interval for a population proportion.

Since from Section 6.3 "The Sample Proportion" in Chapter 6 "Sampling Distributions" we know the mean, standard deviation, and sampling distribution of the sample proportion $\hat{p}$, the ideas of the previous two sections can be applied to produce a confidence interval for a population proportion. Here is the formula.

Large Sample $100\,(1-\alpha)\,\%$ Confidence Interval for a Population Proportion

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}$$

A sample is large if the interval $\left[p-3\,\sigma_{\hat{P}},\ p+3\,\sigma_{\hat{P}}\right]$ lies wholly within the interval $[0,1]$.

In actual practice the value of p is not known, hence neither is $\sigma_{\hat{P}}$. In that case we substitute the known quantity $\hat{p}$ for p in making the check; this means checking that the interval

$$\left[\hat{p}-3\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}},\ \hat{p}+3\sqrt{\frac{\hat{p}\,(1-\hat{p})}{n}}\right]$$

lies wholly within the interval $[0,1]$.

EXAMPLE 7

To estimate the proportion of students at a large college who are female, a random sample of 120 students is selected. There are 69 female students in the sample. Construct a 90% confidence interval for the proportion of all students at the college who are female.

Solution:

The proportion of students in the sample who are female is
$\hat{p} = 69 / 120 = 0.575.$

Confidence level 90% means that $\alpha = 1 - 0.90 = 0.10$ so $\alpha / 2 = 0.05.$ From the last line of Figure 12.3 "Critical Values of " we obtain $z_{0.05} = 1.645.$

Thus

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}\left(1-\hat{p}\right)}{n}} = 0.575 \pm 1.645\sqrt{\frac{\left(0.575\right)\ \left(0.425\right)}{120}} = 0.575 \pm 0.$$

One may be 90% confident that the true proportion of all students at the college who are female is contained in the interval
$\left(0.575 - 0.074,0.575 + 0.074\right) = \left(0.501,0.649\right).$

KEY TAKEAWAYS

- We have a single formula for a confidence interval for a population proportion, which is valid when the sample is large.
- The condition that a sample be large is not that its size n be at least 30, but that the density function fit inside the interval $\left[0,1\right]$.

EXERCISES

BASIC

1. Information about a random sample is given. Verify that the sample is large enough to use it to construct a confidence interval for the population proportion. Then construct a 90% confidence interval for the population proportion.

   a. $n = 25, \hat{p} = 0.7$
   b. $n = 50, \hat{p} = 0.7$

2. Information about a random sample is given. Verify that the sample is large enough to use it to construct a confidence interval for the population proportion. Then construct a 95% confidence interval for the population proportion.

   a. $n = 2500, \hat{p} = 0.22$
   b. $n = 1200, \hat{p} = 0.22$

3. Information about a random sample is given. Verify that the sample is large enough to use it to construct a confidence interval for the population proportion. Then construct a 98% confidence interval for the population proportion.

   a. $n = 80, \hat{p} = 0.4$
   b. $n = 325, \hat{p} = 0.4$

4. Information about a random sample is given. Verify that the sample is large enough to use it to construct a confidence interval for the population proportion. Then construct a 99.5% confidence interval for the population proportion.

   a. $n = 200, \hat{p} = 0.85$
   b. $n = 75, \hat{p} = 0.85$

5. In a random sample of size 1,100, 338 have the characteristic of interest.

   a. Compute the sample proportion $\hat{p}$ with the characteristic of interest.
   b. Verify that the sample is large enough to use it to construct a confidence interval for the population proportion.
   c. Construct an 80% confidence interval for the population proportion p.
   d. Construct a 90% confidence interval for the population proportion p.
   e. Comment on why one interval is longer than the other.

6. In a random sample of size 2,400, 420 have the characteristic of interest.

   a. Compute the sample proportion $\widehat{p}$ with the characteristic of interest.
   b. Verify that the sample is large enough to use it to construct a confidence interval for the population proportion.
   c. Construct a 90% confidence interval for the population proportion p.
   d. Construct a 99% confidence interval for the population proportion p.
   e. Comment on why one interval is longer than the other.

APPLICATIONS

7. A security feature on some web pages is graphic representations of words that are readable by human beings but not machines. When a certain design format was tested on 450 subjects, by having them attempt to read ten disguised words, 448 subjects could read all the words.

   a. Give a point estimate of the proportion p of all people who could read words disguised in this way.
   b. Show that the sample is not sufficiently large to construct a confidence interval for the proportion of all people who could read words disguised in this way.

8. In a random sample of 900 adults, 42 defined themselves as vegetarians.

   a. Give a point estimate of the proportion of all adults who would define themselves as vegetarians.
   b. Verify that the sample is sufficiently large to use it to construct a confidence interval for that proportion.
   c. Construct an 80% confidence interval for the proportion of all adults who would define themselves as vegetarians.

9. In a random sample of 250 employed people, 61 said that they bring work home with them at least occasionally.

   a. Give a point estimate of the proportion of all employed people who bring work home with them at least occasionally.
   b. Construct a 99% confidence interval for that proportion.

10. In a random sample of 1,250 household moves, 822 were moves to a location within the same county as the original residence.

    a. Give a point estimate of the proportion of all household moves that are to a location within the same county as the original residence.
    b. Construct a 98% confidence interval for that proportion.

11. In a random sample of 12,447 hip replacement or revision surgery procedures nationwide, 162 patients developed a surgical site infection.

    a. Give a point estimate of the proportion of all patients undergoing a hip surgery procedure who develop a surgical site infection.
    b. Verify that the sample is sufficiently large to use it to construct a confidence interval for that proportion.
    c. Construct a 95% confidence interval for the proportion of all patients undergoing a hip surgery procedure who develop a surgical site infection.

12. In a certain region prepackaged products labeled 500 g must contain on average at least 500 grams of the product, and at least 90% of all packages must weigh at least 490 grams. In a random sample of 300 packages, 288 weighed at least 490 grams.

    a. Give a point estimate of the proportion of all packages that weigh at least 490 grams.
    b. Verify that the sample is sufficiently large to use it to construct a confidence interval for that proportion.
    c. Construct a 99.8% confidence interval for the proportion of all packages that weigh at least 490 grams.

13. A survey of 50 randomly selected adults in a small town asked them if their opinion on a proposed "no cruising" restriction late at night. Responses were coded 1 for in favor, 0 for indifferent, and 2 for opposed, with the results shown in the table.

1	0	2	0	1	0	0	1	1	2
0	2	0	0	0	1	0	2	0	0
0	2	1	2	0	0	0	2	0	1
0	2	0	2	0	1	0	0	2	0
1	0	0	1	2	0	0	2	1	2

    a. Give a point estimate of the proportion of all adults in the community who are indifferent concerning the proposed restriction.
    b. Assuming that the sample is sufficiently large, construct a 90% confidence interval for the proportion of all adults in the community who are indifferent concerning the proposed restriction.

14. To try to understand the reason for returned goods, the manager of a store examines the records on 40 products that were returned in the last year. Reasons were coded by 1 for "defective," 2 for "unsatisfactory," and 0 for all other reasons, with the results shown in the table.

0 2 0 0 0 0 0 2 0 0
0 0 0 0 0 0 0 0 0 2
0 0 2 0 0 0 0 2 0 0
0 0 0 0 0 1 0 0 0 0

a. Give a point estimate of the proportion of all returns that are because of something wrong with the product, that is, either defective or performed unsatisfactorily.
b. Assuming that the sample is sufficiently large, construct an 80% confidence interval for the proportion of all returns that are because of something wrong with the product.

15. In order to estimate the proportion of entering students who graduate within six years, the administration at a state university examined the records of 600 randomly selected students who entered the university six years ago, and found that 312 had graduated.

    a. Give a point estimate of the six-year graduation rate, the proportion of entering students who graduate within six years.
    b. Assuming that the sample is sufficiently large, construct a 98% confidence interval for the six-year graduation rate.

16. In a random sample of 2,300 mortgages taken out in a certain region last year, 187 were adjustable-rate mortgages.

    a. Give a point estimate of the proportion of all mortgages taken out in this region last year that were adjustable-rate mortgages.
    b. Assuming that the sample is sufficiently large, construct a 99.9% confidence interval for the proportion of all mortgages taken out in this region last year that were adjustable-rate mortgages.

17. In a research study in cattle breeding, 159 of 273 cows in several herds that were in estrus were detected by means of an intensive once a day, one-hour observation of the herds in early morning.

    a. Give a point estimate of the proportion of all cattle in estrus who are detected by this method.
    b. Assuming that the sample is sufficiently large, construct a 90% confidence interval for the proportion of all cattle in estrus who are detected by this method.

18. A survey of 21,250 households concerning telephone service gave the results shown in the table.

	Landline	No Landline
Cell phone	12,474	5,844
No cell phone	2,529	403

a. Give a point estimate for the proportion of all households in which there is a cell phone but no landline.
b. Assuming the sample is sufficiently large, construct a 99.9% confidence interval for the proportion of all households in which there is a cell phone but no landline.
c. Give a point estimate for the proportion of all households in which there is no telephone service of either kind.
d. Assuming the sample is sufficiently large, construct a 99.9% confidence interval for the proportion of all all households in which there is no telephone service of either kind.

ADDITIONAL EXERCISES

19. In a random sample of 900 adults, 42 defined themselves as vegetarians. Of these 42, 29 were women.

    a. Give a point estimate of the proportion of all self-described vegetarians who are women.
    b. Verify that the sample is sufficiently large to use it to construct a confidence interval for that proportion.
    c. Construct a 90% confidence interval for the proportion of all all self-described vegetarians who are women.

20. A random sample of 185 college soccer players who had suffered injuries that resulted in loss of playing time was made with the results shown in the table. Injuries are classified according to severity of the injury and the condition under which it was sustained.

	Minor	Moderate	Serious
Practice	48	20	6
Game	62	32	17

    a. Give a point estimate for the proportion p of all injuries to college soccer players that are sustained in practice.
    b. Construct a 95% confidence interval for the proportion p of all injuries to college soccer players that are sustained in practice.
    c. Give a point estimate for the proportion p of all injuries to college soccer players that are either moderate or serious.

d. Construct a 95% confidence interval for the proportion p of all injuries to college soccer players that are either moderate or serious.

21. The body mass index (BMI) was measured in 1,200 randomly selected adults, with the results shown in the table.

	BMI		
	Under 18.5	**18.5–25**	**Over 25**
Men	36	165	315
Women	75	274	335

a. Give a point estimate for the proportion of all men whose BMI is over 25.
b. Assuming the sample is sufficiently large, construct a 99% confidence interval for the proportion of all men whose BMI is over 25.
c. Give a point estimate for the proportion of all adults, regardless of gender, whose BMI is over 25.
d. Assuming the sample is sufficiently large, construct a 99% confidence interval for the proportion of all adults, regardless of gender, whose BMI is over 25.

22. Confidence intervals constructed using the formula in this section often do not do as well as expected unless n is quite large, especially when the true population proportion is close to either 0 or 1. In such cases a better result is obtained by adding two successes and two failures to the actual data and then computing the confidence interval. This is the same as using the formula

$$\tilde{p} \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}\left(1-\tilde{p}\right)}{\tilde{n}}}$$

where

$$\tilde{p} = \frac{x+2}{n+4} \text{ and } \tilde{n} = n+4$$

Suppose that in a random sample of 600 households, 12 had no telephone service of any kind. Use the adjusted confidence interval procedure just described to form a 99.9% confidence interval for the proportion of all households that have no telephone service of any kind.

LARGE DATA SET EXERCISES

23. Large Data Sets 4 and 4A list the results of 500 tosses of a die. Let p denote the proportion of all tosses of this die that would result in a four. Use the sample data to construct a 90% confidence interval for p.

    http://www.gone.2012books.lardbucket.org/sites/all/files/data4.xls

    http://www.gone.2012books.lardbucket.org/sites/all/files/data4A.xls

24. Large Data Set 6 records results of a random survey of 200 voters in each of two regions, in which they were asked to express whether they prefer Candidate A for a U.S. Senate seat or prefer some other candidate. Use the full data set (400 observations) to construct a 98% confidence interval for the proportion p of all voters who prefer Candidate A.

    http://www.gone.2012books.lardbucket.org/sites/all/files/data6.xls

25. Lines 2 through 536 in Large Data Set 11 is a sample of 535 real estate sales in a certain region in 2008. Those that were foreclosure sales are identified with a 1 in the second column.

    http://www.gone.2012books.lardbucket.org/sites/all/files/data11.xls

    a. Use these data to construct a point estimate $\hat{p}$ of the proportion p of all real estate sales in this region in 2008 that were foreclosure sales.
    b. Use these data to construct a 90% confidence for p.

26. Lines 537 through 1106 in Large Data Set 11 is a sample of 570 real estate sales in a certain region in 2010. Those that were foreclosure sales are identified with a 1 in the second column.

    http://www.gone.2012books.lardbucket.org/sites/all/files/data11.xls

    a. Use these data to construct a point estimate $\hat{p}$ of the proportion p of all real estate sales in this region in 2010 that were foreclosure sales.
    b. Use these data to construct a 90% confidence for p.

ANSWERS

1. a. (0.5492, 0.8508)
   b. (0.5934, 0.8066)

3. a. (0.2726, 0.5274)
   b. (0.3368, 0.4632)

5. a. 0.3073
   b. $\hat{p} \pm 3\sqrt{\frac{\hat{\hat{p}}\hat{\hat{q}}}{n}} = 0.31 \pm 0.04$

      and

      $[0.27, 0.35] \subset [0,1]$

   c. (0.2895, 0.3251)
   d. (0.2844, 0.3302)
   e. Asking for greater confidence requires a longer interval.

7. a. 0.9956
   b. (0.9862, 1.005)

9. a. 0.244
   b. (0.1740, 0.3140)

11. a. 0.013
    b. (0.01, 0.016)
    c. (0.011, 0.015)

13. a. 0.52
    b. (0.4038, 0.6362)

15. a. 0.52
    b. (0.4726, 0.5674)

17. a. 0.5824
    b. (0.5333, 0.6315)

19. a. 0.69
    b. $\hat{p} \pm 3\sqrt{\frac{\hat{\hat{p}}\hat{\hat{q}}}{n}} = 0.69 \pm 0.21$

       and

       $[0.48, 0.90] \subset [0,1]$

    c. 0.69 ± 0.12

21. a. 0.6105
    b. (0.5552, 0.6658)
    c. 0.5583
    d. (0.5214, 0.5952)

23. $(0.1368, 0.1912)$

25. a. $\hat{p} = 0.2280$
    b. $(0.1982, 0.2579)$

7.4 Sample Size Considerations

LEARNING OBJECTIVE

1. To learn how to apply formulas for estimating the size sample that will be needed in order to construct a confidence interval for a population mean or proportion that meets given criteria.

Sampling is typically done with a set of clear objectives in mind. For example, an economist might wish to estimate the mean yearly income of workers in a particular industry at 90% confidence and to within $500. Since sampling costs time, effort, and money, it would be useful to be able to estimate the smallest size sample that is likely to meet these criteria.

Estimating μ

The confidence interval formulas for estimating a population mean μ have the form $\overline{x} \pm E$. When the population standard deviation σ is known,

$$E = \frac{z_{\alpha/2}\sigma}{\sqrt{n}}$$

The number $z_{\alpha/2}$ is determined by the desired level of confidence. To say that we wish to estimate the mean to within a certain number of units means that we want the margin of error E to be no larger than that number. Thus we obtain the minimum sample size needed by solving the displayed equation for n.

Minimum Sample Size for Estimating a Population Mean

The estimated minimum sample size n needed to estimate a population mean μ to within E units at $100\,(1-\alpha)$% confidence is

$$n = \frac{\left(z_{\alpha/2}\right)^2 \sigma^2}{E^2} \quad \text{(rounded } up\text{)}$$

To apply the formula we must have prior knowledge of the population in order to have an estimate of its standard deviation σ. In all the examples and exercises the population standard deviation will be given.

EXAMPLE 8

Find the minimum sample size necessary to construct a 99% confidence interval for μ with a margin of error $E = 0.2$. Assume that the population standard deviation is $\sigma = 1.3$.

Solution:

Confidence level 99% means that $\alpha = 1 - 0.99 = 0.01$ so $\alpha / 2 = 0.005$. From the last line of Figure 12.3 "Critical Values of " we obtain $z_{0.005} = 2.576$. Thus

$$n = \frac{\left(z_{\alpha/2}\right)^2 \sigma^2}{E^2} = \frac{\left(2.576\right)^2 (1.3)^2}{(0.2)^2} = 280.361536$$

which we round up to 281, since it is impossible to take a fractional observation.

EXAMPLE 9

An economist wishes to estimate, with a 95% confidence interval, the yearly income of welders with at least five years experience to within $1,000. He estimates that the range of incomes is no more than $24,000, so using the Empirical Rule he estimates the population standard deviation to be about one-sixth as much, or about $4,000. Find the estimated minimum sample size required.

Solution:

Confidence level 95% means that $\alpha = 1 - 0.95 = 0.05$ so $\alpha / 2 = 0.025$. From the last line of Figure 12.3 "Critical Values of " we obtain $z_{0.025} = 1.960$.

To say that the estimate is to be "to within $1,000" means that $E = 1000$. Thus

$$n = \frac{\left(z_{\alpha/2}\right)^2 \sigma^2}{E^2} = \frac{\left(1.960\right)^2 (4000)^2}{(1000)^2} = 61.4656$$

which we round up to 62.

Estimating *p*

The confidence interval formula for estimating a population proportion p is $\hat{p} \pm E$, where

$$E = z_{\alpha/2} \sqrt{\frac{\hat{p}\left(1 - \hat{p}\right)}{n}}$$

The number $z_{\alpha/2}$ is determined by the desired level of confidence. To say that we wish to estimate the population proportion to within a certain number of percentage points means that we want the margin of error E to be no larger than that number (expressed as a proportion). Thus we obtain the minimum sample size needed by solving the displayed equation for n.

Minimum Sample Size for Estimating a Population Proportion

The estimated minimum sample size n needed to estimate a population proportion p to within E at $100\,(1-\alpha)$% confidence is

$$n = \frac{\left(z_{\alpha/2}\right)^2 \hat{p}\left(1-\hat{p}\right)}{E^2} \quad \text{(rounded } up\text{)}$$

There is a dilemma here: the formula for estimating how large a sample to take contains the number $\hat{p}$, which we know only after we have taken the sample. There are two ways out of this dilemma. Typically the researcher will have some idea as to the value of the population proportion p, hence of what the sample proportion $\hat{p}$ is likely to be. For example, if last month 37% of all voters thought that state taxes are too high, then it is likely that the proportion with that opinion this month will not be dramatically different, and we would use the value 0.37 for $\hat{p}$ in the formula.

The second approach to resolving the dilemma is simply to replace $\hat{p}$ in the formula by 0.5. This is because if $\hat{p}$ is large then $1-\hat{p}$ is small, and vice versa, which limits their product to a maximum value of 0.25, which occurs when $\hat{p} = 0.5$. This is called the **most conservative estimate**[6], since it gives the largest possible estimate of n.

6. The estimate obtained using $\hat{p} = 0.5$, which gives the largest estimate of n.

EXAMPLE 10

Find the necessary minimum sample size to construct a 98% confidence interval for p with a margin of error $E = 0.05$,

a. assuming that no prior knowledge about p is available; and
b. assuming that prior studies suggest that p is about 0.1.

Solution:

Confidence level 98% means that $\alpha = 1 - 0.98 = 0.02$ so $\alpha / 2 = 0.01$. From the last line of Figure 12.3 "Critical Values of " we obtain $z_{0.01} = 2.326$.

a. Since there is no prior knowledge of p we make the most conservative estimate that $\hat{p} = 0.5$. Then

$$n = \frac{\left(z_{\alpha/2}\right)^2 \hat{p}\left(1-\hat{p}\right)}{E^2} = \frac{\left(2.326\right)^2 \left(0.5\right)\ \left(1-0.5\right)}{0.05^2} = 541.0276$$

which we round up to 542.

b. Since $p \approx 0.1$ we estimate $\hat{p}$ by 0.1, and obtain

$$n = \frac{\left(z_{\alpha/2}\right)^2 \hat{p}\left(1-\hat{p}\right)}{E^2} = \frac{\left(2.326\right)^2 (0.1)\ (1-0.1)}{0.05^2} = 194.769936$$

which we round up to 195.

EXAMPLE 11

A dermatologist wishes to estimate the proportion of young adults who apply sunscreen regularly before going out in the sun in the summer. Find the minimum sample size required to estimate the proportion to within three percentage points, at 90% confidence.

Solution:

Confidence level 90% means that $\alpha = 1 - 0.90 = 0.10$ so $\alpha / 2 = 0.05$. From the last line of Figure 12.3 "Critical Values of " we obtain $z_{0.05} = 1.645$.

Since there is no prior knowledge of p we make the most conservative estimate that $\hat{p} = 0.5$. To estimate "to within three percentage points" means that $E = 0.03$. Then

$$n = \frac{\left(z_{\alpha/2}\right)^2 \hat{p}\left(1-\hat{p}\right)}{E^2} = \frac{\left(1.645\right)^2 \left(0.5\right)\ \left(1-0.5\right)}{0.03^2} = 751.6736111$$

which we round up to 752.

KEY TAKEAWAYS

- If the population standard deviation σ is known or can be estimated, then the minimum sample size needed to obtain a confidence interval for the population mean with a given maximum error of the estimate and a given level of confidence can be estimated.
- The minimum sample size needed to obtain a confidence interval for a population proportion with a given maximum error of the estimate and a given level of confidence can always be estimated. If there is prior knowledge of the population proportion p then the estimate can be sharpened.

EXERCISES

BASIC

1. Estimate the minimum sample size needed to form a confidence interval for the mean of a population having the standard deviation shown, meeting the criteria given.

   a. $\sigma = 30$, 95% confidence, $E = 10$
   b. $\sigma = 30$, 99% confidence, $E = 10$
   c. $\sigma = 30$, 95% confidence, $E = 5$

2. Estimate the minimum sample size needed to form a confidence interval for the mean of a population having the standard deviation shown, meeting the criteria given.

   a. $\sigma = 4$, 95% confidence, $E = 1$
   b. $\sigma = 4$, 99% confidence, $E = 1$
   c. $\sigma = 4$, 95% confidence, $E = 0.5$

3. Estimate the minimum sample size needed to form a confidence interval for the proportion of a population that has a particular characteristic, meeting the criteria given.

   a. $p \approx 0.37$, 80% confidence, $E = 0.05$
   b. $p \approx 0.37$, 90% confidence, $E = 0.05$
   c. $p \approx 0.37$, 80% confidence, $E = 0.01$

4. Estimate the minimum sample size needed to form a confidence interval for the proportion of a population that has a particular characteristic, meeting the criteria given.

   a. $p \approx 0.81$, 95% confidence, $E = 0.02$
   b. $p \approx 0.81$, 99% confidence, $E = 0.02$
   c. $p \approx 0.81$, 95% confidence, $E = 0.01$

5. Estimate the minimum sample size needed to form a confidence interval for the proportion of a population that has a particular characteristic, meeting the criteria given.

   a. 80% confidence, $E = 0.05$
   b. 90% confidence, $E = 0.05$
   c. 80% confidence, $E = 0.01$

6. Estimate the minimum sample size needed to form a confidence interval for the proportion of a population that has a particular characteristic, meeting the criteria given.

   a. 95% confidence, $E = 0.02$
   b. 99% confidence, $E = 0.02$
   c. 95% confidence, $E = 0.01$

APPLICATIONS

7. A software engineer wishes to estimate, to within 5 seconds, the mean time that a new application takes to start up, with 95% confidence. Estimate the minimum size sample required if the standard deviation of start up times for similar software is 12 seconds.

8. A real estate agent wishes to estimate, to within \$2.50, the mean retail cost per square foot of newly built homes, with 80% confidence. He estimates the standard deviation of such costs at \$5.00. Estimate the minimum size sample required.

9. An economist wishes to estimate, to within 2 minutes, the mean time that employed persons spend commuting each day, with 95% confidence. On the assumption that the standard deviation of commuting times is 8 minutes, estimate the minimum size sample required.

10. A motor club wishes to estimate, to within 1 cent, the mean price of 1 gallon of regular gasoline in a certain region, with 98% confidence. Historically the variability of prices is measured by $\sigma = \$0.03$. Estimate the minimum size sample required.

11. A bank wishes to estimate, to within \$25, the mean average monthly balance in its checking accounts, with 99.8% confidence. Assuming $\sigma = \$250$, estimate the minimum size sample required.

12. A retailer wishes to estimate, to within 15 seconds, the mean duration of telephone orders taken at its call center, with 99.5% confidence. In the past the standard deviation of call length has been about 1.25 minutes. Estimate the minimum size sample required. (Be careful to express all the information in the same units.)

13. The administration at a college wishes to estimate, to within two percentage points, the proportion of all its entering freshmen who graduate within four years, with 90% confidence. Estimate the minimum size sample required.

14. A chain of automotive repair stores wishes to estimate, to within five percentage points, the proportion of all passenger vehicles in operation that are at least five years old, with 98% confidence. Estimate the minimum size sample required.

15. An internet service provider wishes to estimate, to within one percentage point, the current proportion of all email that is spam, with 99.9% confidence. Last year the proportion that was spam was 71%. Estimate the minimum size sample required.

16. An agronomist wishes to estimate, to within one percentage point, the proportion of a new variety of seed that will germinate when planted, with 95% confidence. A typical germination rate is 97%. Estimate the minimum size sample required.

17. A charitable organization wishes to estimate, to within half a percentage point, the proportion of all telephone solicitations to its donors that result in a gift, with 90% confidence. Estimate the minimum sample size required, using the information that in the past the response rate has been about 30%.

18. A government agency wishes to estimate the proportion of drivers aged 16–24 who have been involved in a traffic accident in the last year. It wishes to make the estimate to within one percentage point and at 90% confidence. Find the minimum sample size required, using the information that several years ago the proportion was 0.12.

ADDITIONAL EXERCISES

19. An economist wishes to estimate, to within six months, the mean time between sales of existing homes, with 95% confidence. Estimate the minimum size sample required. In his experience virtually all houses are re-sold within 40 months, so using the Empirical Rule he will estimate σ by one-sixth the range, or $40 / 6 = 6.7$.

20. A wildlife manager wishes to estimate the mean length of fish in a large lake, to within one inch, with 80% confidence. Estimate the minimum size sample required. In his experience virtually no fish caught in the lake is over 23 inches long, so using the Empirical Rule he will estimate σ by one-sixth the range, or $23 / 6 = 3.8$.

21. You wish to estimate the current mean birth weight of all newborns in a certain region, to within 1 ounce (1/16 pound) and with 95% confidence. A sample will cost \$400 plus \$1.50 for every newborn weighed. You believe the

standard deviations of weight to be no more than 1.25 pounds. You have $2,500 to spend on the study.

a. Can you afford the sample required?
b. If not, what are your options?

22. You wish to estimate a population proportion to within three percentage points, at 95% confidence. A sample will cost $500 plus 50 cents for every sample element measured. You have $1,000 to spend on the study.

a. Can you afford the sample required?
b. If not, what are your options?

ANSWERS

1. a. 35
   b. 60
   c. 139

3. a. 154
   b. 253
   c. 3832

5. a. 165
   b. 271
   c. 4109

7. 23

9. 62

11. 955

13. 1692

15. 22,301

17. 22,731

19. 5

21. a. no
    b. decrease the confidence level