

Lesson

- Welcome to STAT 100!
- ▶

Lesson 1: Statistics: Benefits, Risks, and Measurements
- ▶

Lesson 2: Characteristics of Good Sample Surveys and Comparative Studies
- ▶

Lesson 3: Getting the Big Picture and Summaries
- ▶

Lesson 4: Bell-Shaped Curves and Statistical Pictures
- ▶

Lesson 5: Relationships Between Measurement Variables
- ▶

Lesson 6: Relationships Between Categorical Variables
- ▶

Lesson 7: Understanding Uncertainty
- ▶

Lesson 8: The Diversity of Samples
- ▼

Lesson 9: Confidence Intervals

○

9.1 - Confidence Intervals for a Population Proportion

○

9.2 - Confidence Intervals for a Population Mean

○

9.3 - Confidence Intervals for the Difference Between Two Population Proportions or Means

○

9.4 - Test Yourself!

○

9.5 - Have Fun With It!

▶

Lesson 10: Hypothesis Testing

▶

Lesson 11: Significance Testing Caveats & Ethics of Experiments

Resources

- References
- ▶

Reviews

9.1 - Confidence Intervals for a Population Proportion



A random sample is gathered to estimate the percentage of American adults who believe that parents should be required to vaccinate their children for diseases like measles, mumps, and rubella. We know that estimates arising from surveys like that are random quantities that vary from sample-to-sample. In [Lesson 8](#) we learned what probability has to say about how close a sample proportion will be to the true population proportion.

In an unbiased random survey

$$sample\ proportion = population\ proportion + random\ error.$$

The **Normal Approximation** tells us that the distribution of these random errors over all possible samples follows the normal curve with a standard deviation of

$$\sqrt{\frac{population\ proportion(1 - population\ proportion)}{n}} = \sqrt{\frac{p(1 - p)}{n}}$$

The random error is just how much the sample estimate differs from the true population value. The fact that random errors follow the normal curve also holds for many other summaries like sample averages or differences between two sample proportions or averages - you just need a different formula for the standard deviation in each case (see sections 9.3 and 9.4 below).

Notice how the formula for the standard deviation of the sample proportion depends on the true population proportion p . When we do probability calculations we know the value of p so we can just plug that in to get the standard deviation. But when the population value is unknown, we won't know the standard deviation exactly. However, we can get a very good approximation by plugging in the sample proportion. We call this estimate the standard error of the sample proportion

Standard Error of Sample Proportion = estimated standard deviation of the sample proportion =

$$\sqrt{\frac{sample\ proportion(1 - sample\ proportion)}{n}}$$

Example 9.1

The EPA considers indoor radon levels above 4 picocuries per liter (pCi/L) of air to be high enough to warrant amelioration efforts. Tests in a sample of 200 Centre County Pennsylvania homes found 127 (63.5%) of these sampled households to have indoor radon levels above 4 pCi/L. What is the population value being estimated by this sample percentage? What is the standard error of the corresponding sample proportion?



Solution: The population value is the percentage of all Centre County homes with indoor radon levels above 4 pCi/L.

The standard error of the sample proportion = $\sqrt{\frac{0.635(1 - 0.635)}{200}} = 0.034$

Recap: the estimated percent of Centre Country households that don't meet the EPA guidelines is 63.5% with a standard error of 3.4%. The Normal approximation tells us that

- for 68% of all possible samples, the sample proportion will be within one standard error of the true population proportion and
- for 95% of all possible samples, the sample proportion will be within two standard errors of the true population proportion.

Thus, a 68% confidence interval for thepercent of all Centre Country households that don't meet the EPA guidelines is given by

$$63.5\% \pm 3.4\%.$$

A 95% confidence interval for the percent of all Centre Country households that don't meet the EPA guidelines is given by

$$63.5\% \pm 6.8\%.$$

Note: when you see a margin of error in a news report, it almost always referring to a 95% confidence interval. But other levels of confidence are possible.

Confidence Intervals for a proportion:

For large random samples a confidence interval for a population proportion is given by

$$sample\ proportion \pm z * \sqrt{\frac{sample\ proportion(1 - sample\ proportion)}{n}}$$

where z^* is a multiplier number that comes form the normal curve and determines the *level of confidence* (see Table 9.1 for some common multiplier numbers).

Table 9.1. Commonly Used Multipliers

Multiplier Number (z^*)	Level of Confidence
3.0	99.7%
2.58 (2.576)	99%
2.0 (more precisely 1.96)	95%
1.645	90%
1.282	80%
1.15	75%
1.0	68%

Interpreting Confidence Intervals

To interpret a confidence interval remember that the sample information is random - but there is a pattern to its behavior if we look at all possible samples. Each possible sample gives us a different sample proportion and a different interval. But, even though the results vary from sample-to-sample, we are "confident" because the margin-of-error would be satisfied for 95% of all samples (with $z^*=2$).

The margin-of-error being satisfied means that the interval includes the true population value.

Properties of Confidence Intervals

- There is a trade-off between the level of confidence and the precision of the interval. If you want more confidence, you will have to settle for a wider interval (bigger z^*).
- Our formula for the confidence interval depends on the normal approximation, so you must check that you have independent trials and a large enough sample to be sure that the normal approximation is appropriate.
- The standard error calculation involves estimating the true standard deviation by substituting the sample proportion for the population proportion in the formula. Luckily, this works well in situations where the normal curve is appropriate [i.e. when np and $n(1-p)$ are both bigger than 5].
- A confidence Interval is only related to sampling variability. The probability that your interval captures the true population value could be much lower if your survey is biased (e.g. bad question wording, low response rate, etc...).

Example 9.2

We take a random sample of 50 households in order to estimate the percentage of all homes in the United States that have a refrigerator. It turns out that 49 of the 50 homes in our sample have a refrigerator. Can we use the formulas above to make a confidence interval in this situation?

Solution: No, in such a skewed situation- with only 1 home that does not have a refrigerator - the normal curve would be a very poor approximation to the distribution of sample proportions.

