Confidence Intervals (CI)

A **confidence interval** is a range of plausible values for a population parameter. In this class, we will be constructing confidence intervals for two population parameters:

- population mean, μ
- population proportion, p

While there are many kinds of confidence intervals in statistics (which you can learn about in more advanced classes), we are focusing on confidence intervals with the following form:

$$\text{point estimate} \pm z^* \times SE$$

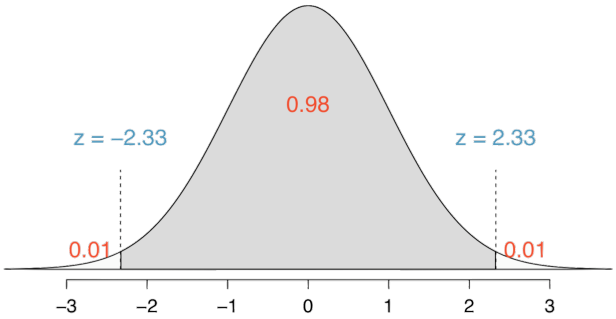
This kind of interval is a Wald confidence interval. **It relies on an assumption of approximate normality. Before we construct such confidence intervals, we must assess whether approximate normality is an appropriate assumption (i.e. is the parent distribution normal? is the sample size "sufficiently large"?).**

The **point estimate** is a

- sample mean, $\bar{x}$ (if we wish to estimate the population mean), or
- sample proportion, $\hat{p}$ (if we wish to estimate the population proportion).

The z^* is a number which is sometimes referred to as the critical value. This comes from a standard normal distribution and changes based on the confidence level we choose for our interval. For example, for a 95% confidence interval, z^*=1.96; this corresponds to cutoff points between which 95% of the probability lies.

The figure below gives an illustration of this for a 98% confidence interval; z^*=2.33.



The standard error (SE) is

- $\frac{\sigma}{\sqrt{n}}$ (if we wish to estimate the population mean), or
- $\sqrt{\frac{p(1-p)}{n}}$ (if we wish to estimate the population proportion).

Example R Code

To find z^* for a 98% confidence interval:

- `qnorm(p=0.01, mean=0, sd=1, lower.tail=FALSE)`

Interpretation

Confidence intervals are:

- always about the population
- are not probability statements
- only about population parameters, not individual observations
- only reliable if the sample statistic they're based on is an unbiased estimator of the population parameter

The general interpretation for a 95% confidence interval for the population mean is as follows:

We are 95% confident that the population mean is between $\bar{x} - \underbrace{1.96}_{z^* \text{ for a 95\% CI}} \times \sigma/\sqrt{n}$ *(the lower bound of the interval) and* $\bar{x} + 1.96 \times \sigma/\sqrt{n}$ *(the upper bound of the interval).*

The general interpretation for a 95% confidence interval for the population proportion is as follows:

We are 95% confident that the population proportion is between $\bar{x} - 1.96 \times \sqrt{p(1-p/)n}$ *(the lower bound of the interval) and* $\bar{x} + 1.96 \times \sqrt{p(1-p)/n}$ *(the upper bound of the interval).*

What does it mean to be 95% confident?

Imagine going out and collecting 100 different samples, and then constructing a separate confidence interval for each one. We expect that about 95 of these hundred confidence intervals will contain the true population parameter (either the population mean or the population proportion, depending on which one you were interested in, given the data).

Practice Problems

1. According to official census figures, 8% of couples living together are not married (in the United States). A researcher took a random sample of 400 couples and found that 9.5% of them are not married. Estimate the proportion of couples in the United States that are not married. Show all your work. (Note, estimate here indicates that you should create a confidence interval.)

   (a) Is this problem about a mean or a proportion?

   This is about a proportion; the observations are categorical - either a couple lives together or they don't.

   (b) What is the value of (i) n; (ii) $\hat{p}$?

   - $n = 400$ (the sample size);
   - $\hat{p} = 0.095$ (the proportion of successes from the sample)

   (c) Check that the conditions of the relevant Central Limit Theorem are satisfied for this problem.

   - Independence: this is a random sample, so we take random to indicate that independence is satisfied. To check this condition, you should be looking for a statement that the observations were collected randomly, or that they are independent.
   - "Sufficiently large sample size": since this is a problem about a proportion, we have to check that the number of successes AND the number of failures are at least 10. Since we are creating a confidence interval (with no hypothesis test), we are only focued on $\hat{p}$, so we are going to plug in $\hat{p}$ to the conditions to check that they are satisfied.

     $np \approx n\hat{p} = 400 * 0.095 = 38 \geq 10$ - number of successes is greater than 10

     $n(1-p) \approx n(1-\hat{p}) = 400 * (1-0.095) = 362 \geq 10$ - number of failures is greater than 10

   Both conditions are satisfied for the Central Limit Theorem (for proportions). We can trust what we get for our confidence interval in (d).

(d) Estimate the proportion of couples in the United States who live together but are not married. Show all your work. (Note, estimate here indicates that you should create a confidence interval.)

Since it wasn't specified, construct a 95% confidence interval.

$$\begin{aligned}
\hat{p} \pm z^* \times \sqrt{\frac{p(1-p)}{n}} &\approx \hat{p} \pm z^* \times \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\
&= 0.095 \pm \texttt{qnorm(p=0.025, lower.tail=FALSE)} \times \sqrt{\frac{0.095(1-0.095)}{400}} \\
&= 0.095 \pm 1.96 \times \sqrt{\frac{0.095(1-0.095)}{400}} \\
&= (0.066, 0.124)
\end{aligned}$$

(e) Interpret your result from (d).

We are 95% confident that the true proportion of couples in the United States who live together but are not married is between 0.066 and 0.124 (between 6.6 and 12.4 percent).

2. The board of a major credit card company requires that the mean wait time for customers for service calls is at most 3.00 minutes. To make sure that the mean wait time is not exceeding the requirement, an assignment manager tracks the wait times of 45 randomly selected calls. The mean wait time was calculated to be 3.4 minutes. Assume the population standard deviation is 1.45 minutes.

   (a) Is this problem about a mean or a proportion?

   This is about a mean - wait time is quantitative.

   (b) Check that the conditions of the relevant Central Limit Theorem are satisfied for this problem.

   - Independence: "...the assistant manager tracks the wait time of 45 randomly selected calls." The calls are randomly selected, so independence is satisfied.
   - "Sufficiently large" sample size: since this is a problem about the mean, we check if the sample size is greater than or equal to 30. It is 45, so we are all set.

   (c) Estimate the mean wait time for customers for service calls. Show all your work.

$$\begin{aligned} \bar{x} \pm z^* \times \sigma/\sqrt{n} &= 3.4 \pm 1.96 \times 1.45/\sqrt{45} \\ &= (2.976, 3.824) \end{aligned}$$

   (d) Interpret your result from (c).

   We are 95% confident that the true mean wait time for customers for service calls is between 2.976 and 3.824 minutes.

3. The population standard deviation for waiting times to be seated at a restaurant is know to be 10 minutes. An expensive restaurant claims that the average waiting time for dinner is approximately 1 hour, but we suspect that this claim is inflated to make the restaurant appear more exclusive and successful. A random sample of 30 customers yielded a sample average waiting time of 50 minutes.

   (a) Is this problem about a mean or a proportion?

   This is about a mean - wait time is quantitative

   (b) Check that the conditions of the relevant Central Limit Theorem are satisfied for this problem.

   - Independence: "A random sample of 30 customers yielded..." The customers are randomly selected, so independence is satisfied.
   - "Sufficiently large" sample size: since this is a problem about the mean, we check if the sample size is greater than or equal to 30. It is 30, so we are all set.

   (c) Estimate the mean wait time for dinner. Show all your work.

   It's easiest if you convert everything to minutes.

$$\begin{aligned} \bar{x} \pm z^* \times \sigma/\sqrt{n} &= 50 \pm 1.96 \times 10/\sqrt{30} \\ &= (46.422, 53.578) \end{aligned}$$

   (d) Interpret your result from (c).

   We are 95% confident that the true mean wait time to be seated at this restaurant is between 46.42 and 53.58 minutes.