Lecture 6: Confidence Intervals
STAT 630, Fall 2021

- **Point estimate**: our best guess for the value of a population parameter (e.g., $\bar{x}$ is a point estimate of μ).
- **Interval estimate**: a plausible range of values for the population parameter. Similar to point estimates, interval estimates are random and vary from sample to sample.

Confidence interval for the population mean μ when the sample size n is large:

Let $X_1, X_2, \cdots, X_n$ be a random sample of size n. For large n, the central limit theorem states that $\bar{X} \sim N(\mu, \sigma/\sqrt{n})$. Using this we can write the following probability statement:

$$P\left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95$$

This follows since the $(1 - 0.95)/2 = 0.025$ quantile of the standard normal distribution is `qnorm(0.025)` $= -1.96$, which implies $P(-1.96 < Z < 1.96) = 0.95$ for $Z \sim N(0, 1)$.
Rearranging terms in the above probability statement gives:

$$P(\bar{X} - 1.96\sigma/\sqrt{n} < \mu < \bar{X} + 1.96\sigma/\sqrt{n}) = 0.95$$

We call $\bar{X} \pm 1.96\sigma/\sqrt{n}$ a 95% confidence interval (CI) for the population mean μ. Specifically, $(\bar{X} - 1.96\sigma/\sqrt{n}, \bar{X} + 1.96\sigma/\sqrt{n})$ is a random interval that contains the population mean μ with probability 0.95. The interval is random since it is computed from a sample (each sample gives a different interval).

What does 95% confidence mean? Suppose we repeatedly take random samples of size n from the population, and construct a 95% confidence interval using each sample. Then approximately 95% of those intervals (19 out of every 20) should contain the population mean μ.

Remarks:

- The population standard deviation σ is usually unknown. When σ is unknown we can estimate it with the sample standard deviation s. This introduces another source of random error in the intervals we construct since s also varies from sample to sample. However, $\bar{x} \pm 1.96s/\sqrt{n}$ is a good approximation of $\bar{x} \pm 1.96\sigma/\sqrt{n}$ when the sample size is large ($n \geq 30$), and when the population distribution is not too heavily skewed or non-normal.
- For a **given realization** of a confidence interval (i.e., the interval is calculated using values from a single, observed sample) it is incorrect to say that there is a 0.95 probability that the population mean μ is contained in that interval. Since μ is a fixed number, it is either contained or not contained in the interval calculated using a given sample. Instead, we say that we are 95% confident that μ is contained in the interval.

Ex1: A random sample of 100 US high school students were asked, "How many days were you physically active for over an hour in the last 7 days?" The sample mean was $\bar{x} = 3.75$ days with a standard deviation of $s = 2.6$ days.[1] Calculate and interpret a 95% confidence interval for the population mean number of physically active days per week for high school students in the US.

Given $n = 100$, $\bar{x} = 3.75$, $s = 2.6$
Sine n is large, a 95% confidence interval can be calculated as

$$\bar{x} \pm 1.96\frac{s}{\sqrt{n}} \implies 3.75 \pm 1.96 \cdot \frac{2.6}{\sqrt{100}} \implies (3.24, 4.26)$$

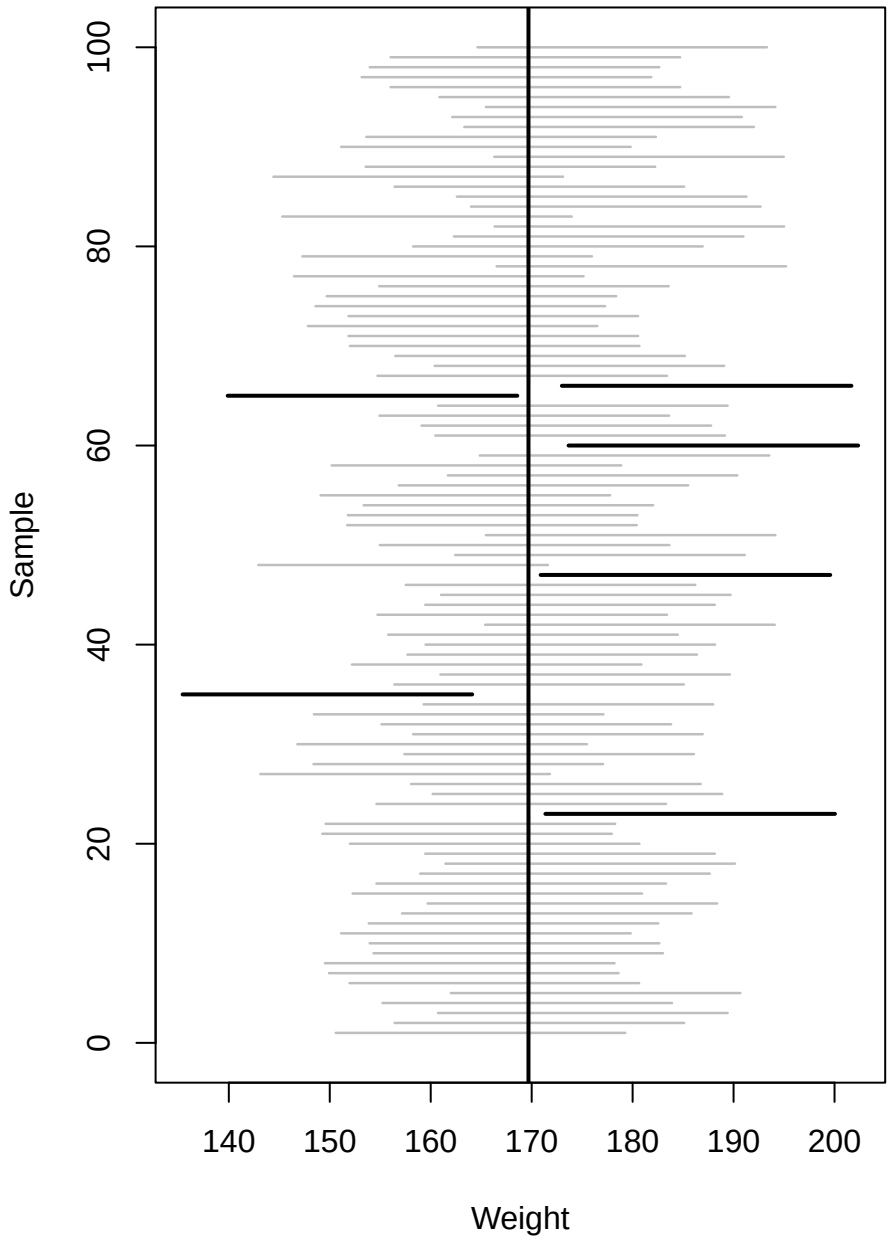
We are 95% confident that the population mean, μ, is between 3.24 and 4.26 days.

Ex2: Critique the following statement: There is a 0.95 probability that the population mean, μ, is between 3.24 and 4.26 days.

μ is a fixed number, so $P(\mu \in (3.24, 4.26)) = 0$ or 1

[1] Data from the Youth Risk Behavior Surveillance System (YRBSS): `https://www.cdc.gov/healthyyouth/data/yrbs/data.htm`

Simulation Example: The figure below shows 100 confidence intervals constructed by repeatedly taking random samples of size $n = 30$ from the weights of individuals in the CDC data set. Each sample was used to construct a different 95% confidence interval. The population mean $\mu = 169.7$ is illustrated with the vertical line (it was calculated as `mean(cdc$weight)`, the mean weight of all 20000 individuals in the data set). 94 out of the 100 intervals contain the population mean μ.

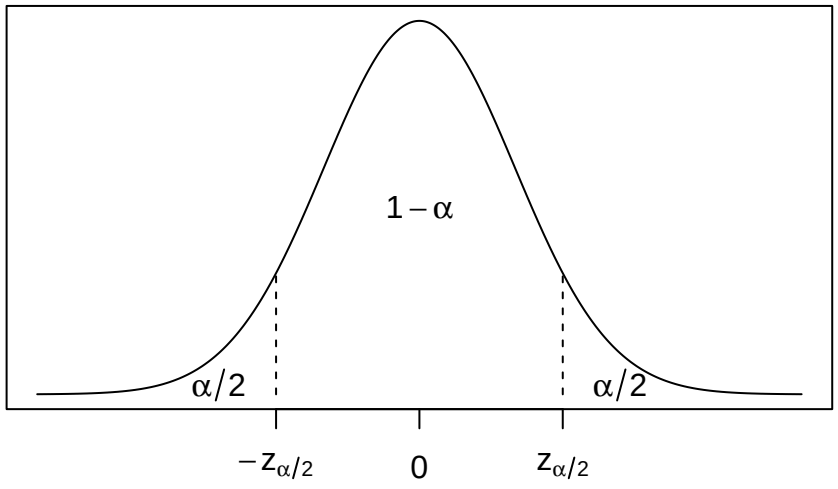


Changing the confidence level:
A $100(1-\alpha)\%$ confidence interval for the population mean μ is given by

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Here, we assume that the population size n is sufficiently large so the CLT applies. Also, since n is large, when σ is unknown, we can replace it with s to obtain an approximate interval.

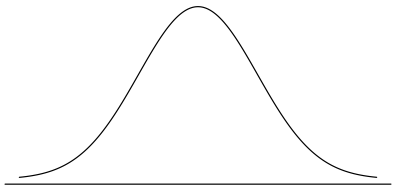
$z_{\alpha/2}$ is the value such that $P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1-\alpha$, where $Z \sim N(0,1)$.



Confidence level $1-\alpha$	Value of $\alpha/2$	$z_{\alpha/2}$
0.9	0.05	1.645
0.95	0.025	1.96
0.98	0.01	2.326
0.99	0.005	2.576

The values in this table can be computed using the R command:
$z_{\alpha/2} =$ `qnorm(1-alpha/2)`

Ex3: Calculate a 99% confidence interval for the population mean number of physically active days per week for US high school students. Recall, the sample size $n = 100$, sample mean $\bar{x} = 3.75$ days, and the sample standard deviation $s = 2.6$ days.



Some Terminology
A $100(1-\alpha)\%$ confidence interval for the population mean μ:

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

- Standard Error = SE = $\sigma/\sqrt{n}$
- Margin of Error = $z_{\alpha/2}$SE = $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
- Confidence Level = $1-\alpha$ (or $100(1-\alpha)\%$ when expressed as a percentage)
- $z_{\alpha/2}$ is called the critical value

Sample Size Determination
Suppose we want to estimate a confidence interval for μ with a specified margin of error. What sample size n is required so that the margin of error of the interval is $\pm$ E with confidence level $1-\alpha$?

Ex4: What sample size is needed to estimate the mean number of physically active days per week for US high school students using a 95% confidence interval with a ± 0.25 margin of error? Use $s = 2.6$ as the estimate for σ.

$$n = \left(\frac{z_{\alpha/2}s}{E}\right)^2 = \left(\frac{1.96 \cdot 2.6}{0.25}\right)^2 = 415.5$$

Need to sample $n = 416$ high school students.

Practice Problems

Practice Problem 1: The 2010 General Social Survey asked the question: "For how many days during the past 30 days was your mental health, which includes stress, depression, and problems with emotions, not good?" Based on responses from 1,151 US residents, the survey reported a 95% confidence interval of 3.40 to 4.24 days in 2010.

(a) Interpret this interval in context of the data.

(b) Suppose the researcher thinks a 99% confidence level would be more appropriate. Will this new interval be smaller or larger than the 95% confidence interval?

(c) If a new survey were to be done with 500 Americans, would the standard error of the estimate be larger, smaller, or about the same. Assume the standard deviation has remained constant since 2010.

Practice Problem 2: Suppose 20 years ago, the mean cholesterol level of adult men in a certain town was 185 mg/dl with a standard deviation of 50 mg/dl.

(a) Suppose you obtain a random sample of size 100 and find the mean cholesterol to be $\bar{x} = 210$. Assuming that σ has not changed, find a 90% confidence interval for the mean cholesterol level of the population (of adult men in this town).

(b) Suppose you decide to conduct a new study to determine the mean cholesterol level of adult men in this town. Assuming that the standard deviation has not changed, how many people should you include in your sample if you want the margin of error to be ± 10 mg/dl, using 95% confidence.

(c) If you want to be 99% confident, then how large should your sample size be?

Practice Problem 3: In any given situation, if the level of confidence and the standard deviation are kept constant, how much would you need to increase the sample size to decrease the width of the interval to half its original size?

Solutions to Practice Problems

Practice Problem 1

(a) We are 95% confident that the population mean μ for the number of days per month US residents report being in poor mental health is between 3.4 and 4.24.

(b) A 99% confidence interval would be larger.

(c) The standard error $SE = \sigma/\sqrt{n}$ would be larger since when n decreases the SE will increase, assuming σ is fixed.

Practice Problem 2

(a) Since n is large (≥ 30), and the sample is random, the confidence interval is valid by the CLT. Thus a 90% confidence interval for μ is given by

$$\bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \implies 210 \pm 1.645\frac{50}{\sqrt{100}} \implies (201.775, 218.225)$$

(b)

$$n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{1.96 \cdot 50}{10}\right)^2 = 96.04$$

A sample size of 96 is needed.

(c)

$$n = \left(\frac{z_{\alpha/2}\sigma}{E}\right)^2 = \left(\frac{2.576 \cdot 50}{10}\right)^2 = 165.9$$

A sample size of 166 is needed.

Practice Problem 3

You would need to increase the sample size n by a factor of 4. The margin of error $E = z_{\alpha/2}\sigma/\sqrt{n}$. Substituting $4n$ into the margin of error formula gives

$$z_{\alpha/2}\frac{\sigma}{\sqrt{4n}} = z_{\alpha/2}\frac{\sigma}{2\sqrt{n}} = \frac{E}{2}$$

Code Appendix

Here is the code I wrote to visualize confidence intervals (for mean weight) constructed by repeatedly taking samples from the population. You are not expected to know how to make a plot like this; but I decided to include the code in case you are interested.

```
cdc <- readRDS(url("https://ericwfox.github.io/data/cdc.rds"))

set.seed(200)
mu <- mean(cdc$weight)
sigma <- sd(cdc$weight)
ci_lower <- rep(0, 100)
ci_upper <- rep(0, 100)
contains <- rep(0, 100)
for(i in 1:100) {
  samp <- sample(cdc$weight, 30)
  ci_lower[i] <- mean(samp) - 1.96 * sigma / sqrt(30)
  ci_upper[i] <- mean(samp) + 1.96 * sigma / sqrt(30)
  if(mu >= ci_lower[i] & mu <= ci_upper[i]) {
    contains[i] <- 1
  }
}

par(mar=c(5, 4, 2, 2)) # format margins
xmin <- min(ci_lower)
xmax <- max(ci_upper)
plot(c(xmin, xmax), c(0, 100), type="n", xlab = "Weight", ylab = "Sample")
for(i in 1:100) {
  lines(c(ci_lower[i], ci_upper[i]), c(i, i), col="grey", lwd=1.25)
  if(contains[i] == 0) {
    lines(c(ci_lower[i], ci_upper[i]), c(i, i), col="black", lwd=2)
  }
}
abline(v=mu, col="black", lwd=2)
```