Lecture 19: Confidence Intervals for a Mean with a Known Standard Deviation

Chapter 14 in book

Corinne Riddell (Instructors Alan Hubbard and Tomer Altman)

October 14, 2024

Learning objectives for today

- What is a confidence interval?
- How to make a confidence interval for your sample's mean when the population standard deviation is known

Readings

- Chapter 14 of Baldi and Moore
- Online resource: Estimating a confidence interval see 10.5 and 10.5.2 (we get to the material in 10.5.1 on a later lecture); page 320

Statistical Inference

Statistical Inference provides methods for drawing conclusions about a population from sample data. This includes:

- Confidence intervals for point estimates (this lecture)
- Hypothesis tests (covered in the next lecture)

Conditions for inference about a mean

For the methods we discuss today, the following conditions (also called assumptions) need to be present to use your sample mean $\bar{x}$ to make inference about an underlying population mean μ:

1. The sample is a simple random sample from the population of interest. There is no non-response or other systematic bias (i.e., no confounding, no measurement error, no selection bias). Note: We don't talk much about systematic error in this class, but it is super important if you have observational data. Take epidemiology to learn about about systematic error!
2. Either the population distribution of the variable follows a Normal distribution $N(\mu, \sigma)$ or the sample size is big enough to invoke the Central Limit Theorem (CLT).
3. Either the standard deviation in the population σ is known **OR** the sample size is "big enough" so that the sample standard deviation can be used.

Mean height example

A recent National Health and Nutrition Examination Survey (NHANES) reports that the mean height of a sample of 217 eight-year old boys was $\bar{x} = 132.5$ cm. On the basis of this sample, we want to estimate the mean μ in the population of >1 million American eight-year-old boys.

First, we need to check if the problem description meets the conditions / assumptions required:

- Assumption 1: SRS

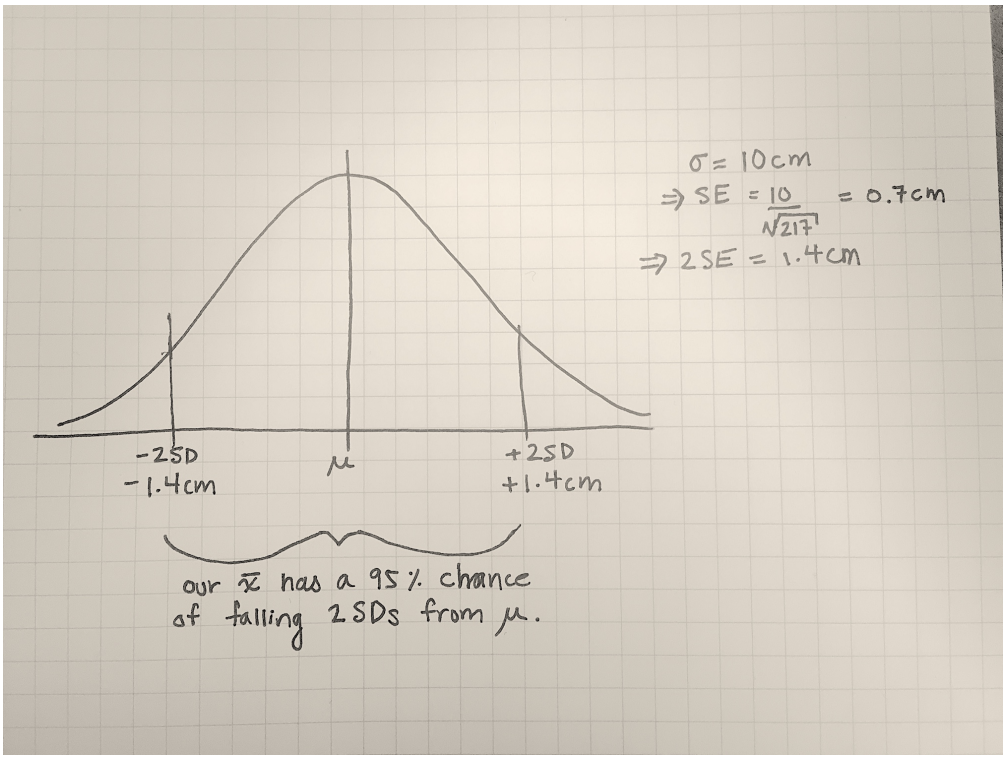
- Assumption 2: Normality or n big enough
- Assumption 3: Known SD or n big enough

Mean height example

- Assumption 1: Is it a simple random sample (SRS)? Look here for how NHANES participants are chosen. It is not an SRS, though random sampling was done within multiple stages. For this question, we will pretend the sample is an SRS.
- Assumption 2: Assume that the distribution of heights in the total population is Normally distributed. This is an okay assumption to make about measurements like height. If you had access to the sample, we could make a histogram and Q-Q plot of the data and see if the sample appears to be roughly Normally distributed and use that as evidence in support of this assumption. In addition, the sample size is probably quite large and thus one can invoke the CLT.
- Assumption 3: We are not provided the population standard deviation σ, but perhaps we could do some research and find that $\sigma = 10$ cm can be assumed as the population standard deviation. In this class, if you are asked to assume a standard deviation, it will be provided to you by the question.

Calculating a 95% confidence interval

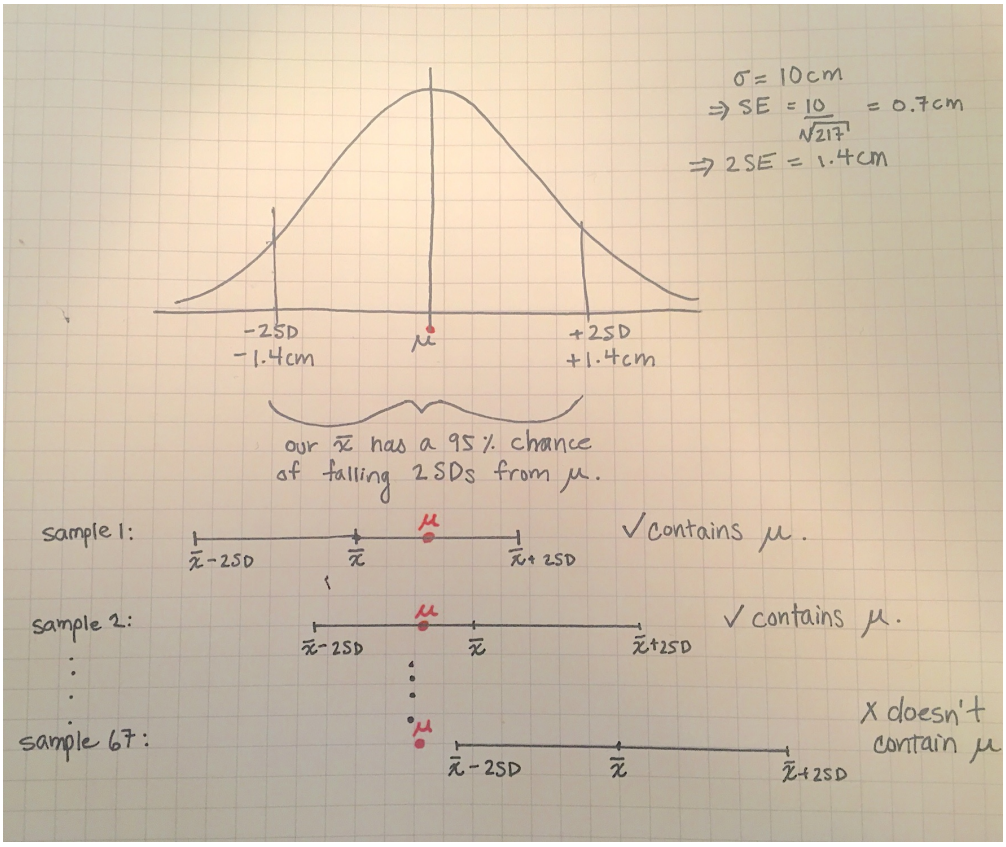
- Recall from last class that $\bar{x}$ is an unbiased estimator of μ. This means that if you took multiple samples, the average of $\bar{x}$ from each sample equals μ.
- Under repeated sampling, the sampling distribution of $\bar{x}$ is Normally distributed with a mean of μ and standard deviation $\frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{217}} = 0.7$ cm. This follows from the previous lecture.
- On scratch paper, we can draw the Normal distribution for the sampling distribution, and shade in the middle 95% of the area within 1.96 (≈ 2) standard deviations of the mean.



Calculating a 95% confidence interval (continued)

- An $\bar{x}$ from any random sample has a 95% chance of being within 1.96 (≈ 2) SD of the population mean μ

- This implies that for 95% of samples, 1.4 cm is the maximum distance separating $\bar{x}$ and μ
- If we estimate that the value μ is somewhere in the interval from $\bar{x} - 1.4$ to $\bar{x} + 1.4$, we'll be right 95% of the time
- That is, 95% of the intervals we make will contain the true parameter value



Calculating a 95% confidence interval (continued)

Using the sample estimate $\bar{x} = 132.5$ and the 1.96 (≈ 2) times the standard error of the sampling distribution of 1.4, our interval has a lower bound of:

$$\bar{x} - 1.4 = 132.5 - 1.4 = 131.1$$

And an upper bound of:

$$\bar{x} + 1.4 = 132.5 + 1.4 = 133.9$$

IMPORTANT: Interpretation of a confidence interval

- Our best guess for μ is 132.5
- Given we only took one sample of size n=217, this best estimate is imprecise
- Our 95% confidence interval for μ is 131.1 to 133.9
- If our model assumptions are correct and there is only random error affecting the estimate, then 95% of the intervals we make will contain the true value μ. That is, 19 times out of 20, the intervals we make will contain the true value.

- This means that the interval $\bar{x} \pm 1.4$ has a 95% success rate in capturing within that interval the mean height μ of all eight-year-old American boys
- I emphasize this as **important** because many people get the interpretation wrong, and it is often misinterpreted on the Internet and in other sources!
- Thus, the technical definition of a 95% CI is a random interval that has the property that in **repeated experiments** the true parameter (true mean, μ) will fall within the CI 95% of the time
- We can not use the data to test this, so must use theory to develop a CI procedure that has this property
- **Do not use the textbook's shorthand that "we are 95% confident that μ is contained in the CI"**. This description is ambiguous and imprecise.

What would make the CI smaller (and more precise)?

- If we increase the sample size, the confidence interval becomes narrower and more precise
- If the underlying variability in the data was smaller (i.e., σ was smaller), than the CI would be more precise

Definitions: Margin of error and confidence level

The 95% confidence interval we made took this format:

$$estimate \pm 2 \times SE$$

Here SE is the standard error, which is $\frac{\sigma}{\sqrt{n}}$ here. Let $2 \times SE$ be called the "margin of error". Then:

$$estimate \pm \text{margin of error}$$

For a 95% confidence interval:

- 95% is called the **confidence level**
- The **margin of error** is $2 \times SE$ for a 95% confidence level **if** the sampling distribution of $\bar{x}$ is Normal!
- The margin of error is different for different confidence levels. For example, if we wanted to make a 99% confidence interval, would the margin of error increase or decrease?

Confidence intervals for the mean μ

This table summarizes the number (which we refer to as the **critical value** $z*$ to multiply by the SE for different confidence levels:

Confidence level C	90%	95%	99%
Critical value $z*$	1.645	1.960 (≈ 2)	2.576

- These numbers correspond to the value on the x-axis corresponding to having 90%, 95%, or 99% of the area under the Normal density curve between $-z$ and z. For example, the middle 90% of the area under a Normal density curve lies between -1.645 and +1.645.
- Thus, a 90% confidence interval is of the form:

$$\bar{x} \pm 1.645 \frac{\sigma}{\sqrt{n}}$$

Confidence interval for the mean of a Normal population

Draw an SRS of size n from a Normally-distributed population having unknown mean μ and known standard deviation σ. A level C confidence interval for μ is:

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

We can rewrite this as:

unbiased estimate $\pm$ (critical value) $\times$ (sd of the distribution of the estimate)

unbiased estimate $\pm$ (critical value) $\times$ (standard error)

unbiased estimate $\pm$ margin of error

PPDAC Steps in finding confidence intervals

1. Problem: Statement of the problem in terms of the parameter you would like to estimate
2. Plan: How will you estimate this parameter? What type of data will you collect?
3. Data: After you plan the study, collect the data you need to answer the problem
4. Analysis: Evaluate whether the assumptions required to compute a confidence interval are satisfied. Calculate the estimate of the mean and its confidence interval.
5. Conclusion: Return to the practical question to describe your results in this setting

Example on IQ scores (Example 14.3)

We are interested in the mean IQ scores of 7th grade girls in a Midwest school district. Here are the scores for 31 randomly selected seventh-grade girls. We also know that the standard deviation of IQ scores is 15 points:

```
scores <- c(114, 100, 104, 89, 102, 91, 114, 114, 103, 105,
            108, 130, 120, 132, 111, 128, 118, 119, 86, 72,
            111, 103, 74, 112, 107, 103, 98, 96, 112, 112, 93)

iq_data <- data.frame(scores)

known_sigma <- 15
```

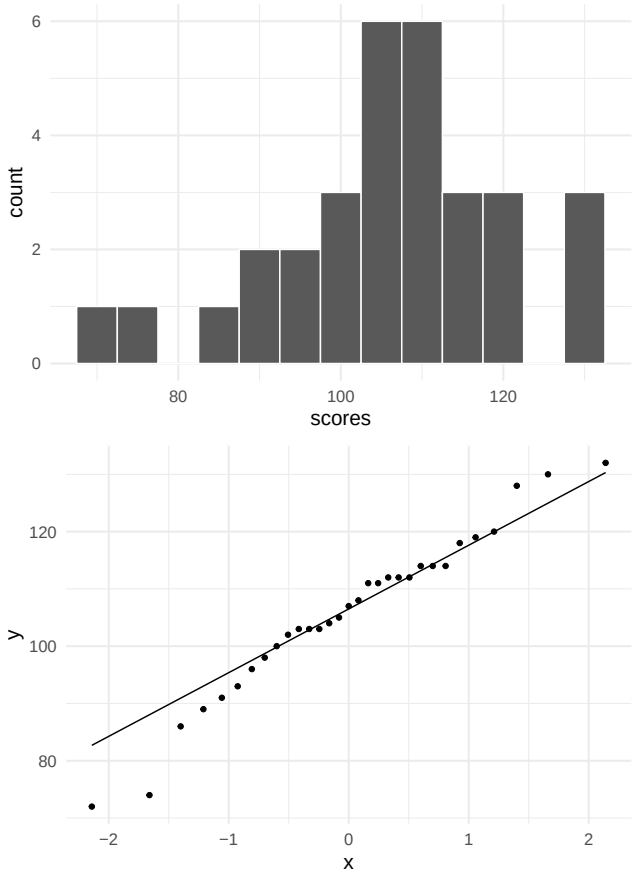
Estimate the mean IQ score μ for all seventh grade girls in this Midwest school district by giving a 95% confidence interval.

Example on IQ scores (Example 14.3)

First check the three conditions (also called assumptions):

1. Normality: We can evaluate this using a histogram and Q-Q plot
2. SRS: Does the information provided say this is an SRS? We cannot evaluate this assumption by looking at a plot.
3. Known σ: Is σ known?

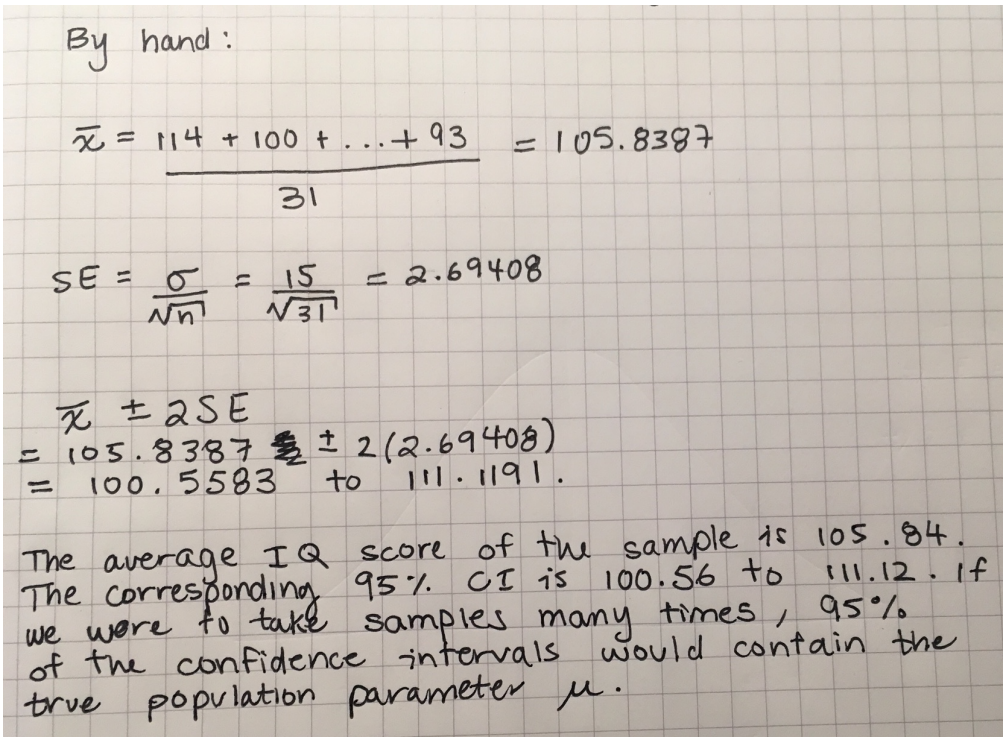
Assess Normality



We can't examine the Normality of the population (because we don't have data on the entire population) but we can make a plot for the sample. These data appear slightly left-skewed, but since there is not much data, it may actually follow a Normal distribution.

Calculating the estimated mean and its confidence interval

Option 1: Perform calculations by hand (approximating the critical value $z* \approx 2$)



Calculating the estimated mean and its confidence interval

Option 2: Perform calculations using R

```
sample_mean <- mean(scores)

standard_error <- known_sigma/sqrt(length(scores))
critical_value <- 1.96

lower_bound <- sample_mean - critical_value*standard_error
upper_bound <- sample_mean + critical_value*standard_error

sample_mean
```

```
## [1] 105.8387
```

```
standard_error
```

```
## [1] 2.69408
```

```
lower_bound
```

```
## [1] 100.5583
```

```
upper_bound
```

```
## [1] 111.1191
```

The sample estimate of the mean is 105.84. Its 95% confidence interval is from 100.56 to 111.12. If our model assumptions are correct and there is only random error affecting the estimate, this method for calculating confidence intervals will contain the true value μ within the margin of error 95% of the time (19 times out of 20).

Repeat using a user-written function in R

- There are many packages in R that can take as input a vector of observations and return the estimate and CI
- Following is an example of one user-written function. It takes as input a known SD.

```
CI_z <- function (x, standard_deviation, ci = 0.95)
{
`%>%` <- magrittr::`%>%`
sample_size <- length(x)
Margin_Error <- abs(qnorm((1-ci)/2))* standard_deviation/sqrt(sample_size)
df_out <- data.frame( sample_size=length(x), Mean=mean(x), sd=sd(x),
Margin_Error=Margin_Error,
'CI lower limit'=(mean(x) - Margin_Error),
'CI Upper limit'=(mean(x) + Margin_Error)) %>%
tidyr::pivot_longer(names_to = "Measurements", values_to ="values", 1:6 )
return(df_out)
}
CI_z(x=scores,standard_deviation = known_sigma,ci=0.95)
```

```
## # A tibble: 6 x 2
##   Measurements   values
##   <chr>           <dbl>
## 1 sample_size     31
## 2 Mean           106.
## 3 sd              14.3
## 4 Margin_Error     5.28
## 5 CI.lower.limit 101.
## 6 CI.Upper.limit 111.
```

User written function when SD is unknown

- Function `CI_t` calculates the sample SD and uses the t-distribution as opposed to the Normal distribution to derive the CI
- If the data are Normally distributed, and the population SD is unknown, then the t-distribution accounts for the additional error for having to calculate the sample SD (results in wider CI's than using the Normal distribution)
- As n gets bigger, this difference becomes negligible

```
CI_t <- function (x, ci = 0.95)
{
`%>%` <- magrittr::`%>%`
Margin_Error <- qt(ci + (1 - ci)/2, df = length(x) - 1) * sd(x)/sqrt(length(x))
df_out <- data.frame( sample_size=length(x), Mean=mean(x), sd=sd(x),
Margin_Error=Margin_Error,
'CI lower limit'=(mean(x) - Margin_Error),
'CI Upper limit'=(mean(x) + Margin_Error)) %>%
tidyr::pivot_longer(names_to = "Measurements", values_to ="values", 1:6 )
return(df_out)
}

CI_t(x=scores,ci=0.95)
```

```
## # A tibble: 6 x 2
##   Measurements   values
##   <chr>           <dbl>
```

```
## 1 sample_size     31
## 2 Mean           106.
## 3 sd              14.3
## 4 Margin_Error     5.23
## 5 CI.lower.limit 101.
## 6 CI.Upper.limit 111.
```

Recap

- We learned how to create a confidence interval for the mean when the standard deviation for the population is known
- We learned about the three required assumptions and how to check the Normality assumption using a histogram and Q-Q plot
- We learned how to interpret the confidence interval and the definitions for the confidence level and the margin of error
- We introduced the situation where the SD of the data is unknown