L19: Statistical Inference with confidence intervals

Statistics is everywhere“‘

From The New York Times, March 6, 2020

From Bloomberg news on, Feb 4, 2022

We see articles like these all the time.

Statistics is everywhere

From the articles:

"The economy's remarkably steady job-creation machine sputtered in February and produced a mere 20,000 jobs. It was the smallest gain in well over a year and came on top of other signs that the economy was off to a sluggish start in 2019."

and

"The U.S. labor market showed unexpected strength last month despite record Covid-19 infections, extending momentum into the new year as surging wages added more pressure on the Federal Reserve to raise interest rates."

But also later in the 2022 article (not mentioned in 2020):

"A broad-based 467,000 gain... followed a 709,00 total upward revision to the prior two months"

Numbers are actually revised twice, once in the month following the first report, and again the month after that.

Statistics is everywhere

What we rarely see included in the articles talking about the jobs numbers is the margin of error.
For the 2022 estimate, the revision was more than 1.5 times the current number !

You can read more about this from the article on fivethirtyeight

Statistical Inference

Theory to practice

So far in part II we have been talking about probability and underlying probability distributions.

Now we are going to think about how to use these theoretical distributions to put some boundaries around estimates we calculate from samples.

Statistical Inference

Statistical Inference provides methods for drawing conclusions about a population from sample data. We are using data from a sample to infer something about the underlying population.

Today we will talk about

- Confidence intervals for point estimates
- Margins of error

Simple conditions for inference about a mean

1. We have 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).
2. The quantitative variable we measure has a perfectly Normal distribution $N(\mu, \sigma)$
3. We don't know the population mean μ, and want to estimate it. But we do know the population standard deviation.

Note that these conditions are idealized and not often realistic, however we will use this idealized version as a base which we will adapt as we move forward and discuss more realistic scenarios.

Example 14.1 Baldi and Moore

A recent NHANES reports that the mean height of a sample of 217 eight-year old boys was $\bar{x} = 132.5$ cm.

We want to use this sample 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 simple conditions required:

Condition 1: sampling

- Is it a SRS?

Example 14.1 Baldi and Moore

Condition 2: distribution in the population

- Assume that the distribution of heights in the total population is Normally distributed

Condition 3: Unknown population mean but known population σ

- We also need to assume a standard deviation. We will assume $\sigma = 10$ cm. (Note that if you are asked to assume a standard deviation, it will be provided to you by the question.)

Calculating a confidence interval

- Recall that $\bar{x}$ is an unbiased estimator of μ. Under repeated sampling, the sampling distribution of $\bar{x}$ is Normally distributed with a mean of μ and standard deviation $\sigma/\sqrt{n} = 10/\sqrt{217} = 0.7$ cm.

Calculating a confidence interval

- We can draw the Normal distribution for the sampling distribution, and shade in the middle 95% of the area within 2 standard deviations of the mean. Thus, an $\bar{x}$ from any random sample has a 95% chance of being within 2 SD of the population mean μ. This means that for 95% of samples, 1.4 cm is the maximum distance separating $\bar{x}$ and μ. Therefore, 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 times we take a sample.

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

to

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

Interpretation of a confidence interval

- Our best estimate of μ is 132.5
- But, given we only took one sample of size n=217, this best estimate is imprecise
- The 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, this method for calculating confidence intervals will contain the true value μ 95% of the time (19 times out of 20).
- 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.

Interpretation of a confidence interval

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)?

Remember that we are building the CI based on + and - 2 X standard deviation

$$standarddeviation = \frac{\sigma}{\sqrt{n}}$$

What would make the CI smaller?

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

$$standarddeviation = \frac{\sigma}{\sqrt{n}}$$

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

Margin of error and confidence level

Form of a confidence interval:

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

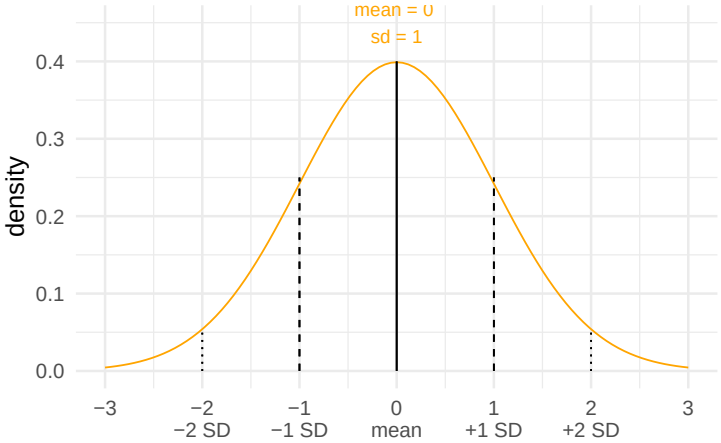
The margin of error will differ based on the confidence level (often 90%, 95%, or 99%) that is chosen.

Will a 99% confidence interval be wider or narrower than a 95% confidence interval?

The standard Normal distribution

Recall the Standard Normal

- The standard Normal distribution N(0,1) has $\mu = 0$ and $\sigma = 1$.
- $X \sim N(0, 1)$, implies that the random variable X is Normally distributed.



Standardizing Normally distributed data

- Any random variable that follows a Normal distribution can be standardized
- If x is an observation from a distribution that has a mean μ and a standard deviation σ,

$$z = \frac{x - \mu}{\sigma}$$

What's the Z

By converting our variable of interest X to Z we can use the probabilities of the standard normal probability distribution to estimate the probabilities associated with X.

- A standardized value is often called a z-score
- Interpretation: z is the number of standard deviations that x is above or below the mean of the data.

Confidence intervals for the mean μ

Confidence intervals for the mean μ

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 between -z and z.

The generic format of a confidence interval is then:

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

Confidence intervals for the mean μ

- For example, the middle 90% of the area under the Normal density 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 a SRS of size n from a Normal population having unknown mean μ and known standard deviation σ. A level C confidence interval for μ is:

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

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

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

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? What theoretical distribution (if any) is appropriate as a model for the data generating distribution?
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 using the method you have chosen 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 from literature: Daylight savings

7 Things to Know About Daylight Saving Time, Published March 09, 2023 By Morgan Coulson Article published on Hopkins website

"Adolescents who get less sleep often have behavioral, learning, and attention issues, as well as an increased risk of accidents, injuries, high blood pressure, obesity, diabetes, and mental health problems. A 2015 study published in the Journal of Clinical Sleep Medicine found that during school days after the time change, students were sleepier, had slower reaction times, and were less attentive."

Pre-DST sleep

	Week 1 Baseline					
	Monday Mean ± SD	Tuesday Mean ± SD	Wednesday Mean ± SD	Thursday Mean ± SD	Friday Mean ± SD	Week 1 Average
Actigraphy						
TST (min)	444.76 ± 81.24	457.65 ± 81.94	445.03 ± 91.64	452.43 ± 70.17	527.71 ± 91.20	471.04 ± 57.57
SL (min)	13.31 ± 17.23	15.39 ± 16.17	13.23 ± 26.34	14.73 ± 21.16	17.03 ± 29.05	15.44 ± 15.60
SE (%)	95.96 ± 3.07	97.37 ± 2.30	96.37 ± 4.39	96.45 ± 3.74	96.57 ± 3.34	96.40 ± 2.37
WASO (min)	18.17 ± 15.66	11.94 ± 11.52	17.19 ± 26.68	16.10 ± 17.87	19.13 ± 19.78	17.31 ± 12.35
Sleep Diary						
TST (min)	401.46 ± 96.99	428.07 ± 57.44	418.48 ± 68.65	429.52 ± 75.86	475.96 ± 76.61	431.97 ± 50.03
SL (min)	19.53 ± 15.54	20.78 ± 20.59	17.26 ± 15.47	15.89 ± 14.23	15.89 ± 16.87	17.81 ± 11.07
Awakenings (count)	0.94 ± 1.25	0.50 ± 0.76	0.76 ± 1.09	0.39 ± 0.79	1.03 ± 1.47	0.78 ± 0.82
Sleep Quality Score	2.97 ± 0.85	3.32 ± 0.75	3.12 ± 0.94	3.43 ± 0.73	3.81 ± 0.75	3.36 ± 0.50
KSS Score	4.55 ± 1.61	4.47 ± 1.8	4.74 ± 1.95	4.24 ± 1.52	3.82 ± 1.33	4.30 ± 1.09

DST, daylight saving time; TST, total sleep time; SL, sleep latency; SE, sleep efficiency; WASO, wake after sleep onset; KSS, Karolinska Sleepiness Scale; SD, standard deviation.

Post-DST sleep

	Week 2 Post-DST						
	Monday Mean ± SD	Tuesday Mean ± SD	Wednesday Mean ± SD	Thursday Mean ± SD	Friday Mean ± SD	Week 2 Average	p value for weekly difference
Actigraphy							
TST (min)	440.41 ± 80.45	431.29 ± 64.39*	423.61 ± 68.50	433.50 ± 69.83	442.29 ± 93.04**	438.61 ± 53.53	0.001
SL (min)	11.79 ± 9.48	12.03 ± 11.86	10.42 ± 10.12	13.50 ± 22.08	12.23 ± 11.84	12.41 ± 8.84	0.249
SE (%)	94.82 ± 6.62	97.01 ± 2.58	96.44 ± 3.74	96.16 ± 3.57	95.50 ± 6.44	95.79 ± 2.92	0.245
WASO (min)	23.03 ± 36.15	11.74 ± 10.69	14.97 ± 17.57	16.27 ± 16.18	19.16 ± 26.68	17.90 ± 13.43	0.825
Sleep Diary							
TST (min)	441.96 ± 107.98	409.78 ± 58.72	404.04 ± 80.57	425.85 ± 77.79	365.08 ± 89.99**	404.78 ± 47.45	0.004
SL (min)	15.30 ± 16.84	16.28 ± 16.32	19.38 ± 22.95	18.66 ± 24.88	15.94 ± 18.10	15.61 ± 14.29	0.308
Awakenings (count)	0.44 ± 1.08*	0.59 ± 1.21	0.72 ± 1.30	0.47 ± 0.84	0.61 ± 0.96	0.58 ± 0.83	0.150
Sleep Quality Score	3.17 ± 0.85	3.07 ± 0.81	3.11 ± 0.64	3.21 ± 0.74	3.00 ± 0.93**	3.10 ± 0.48	0.021
KSS Score	4.94 ± 1.83	5.09 ± 1.44*	4.89 ± 1.53	4.82 ± 1.59	4.7 ± 1.57**	5.08 ± 0.89	< 0.001

*$p < 0.05$, **$p < 0.01$ (significance level for weekday difference). DST, daylight saving time; TST, total sleep time; SL, sleep latency; SE, sleep efficiency; WASO, wake after sleep onset; KSS, Karolinska Sleepiness Scale; SD, standard deviation.

Example on IQ scores (pg. 354)

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
```

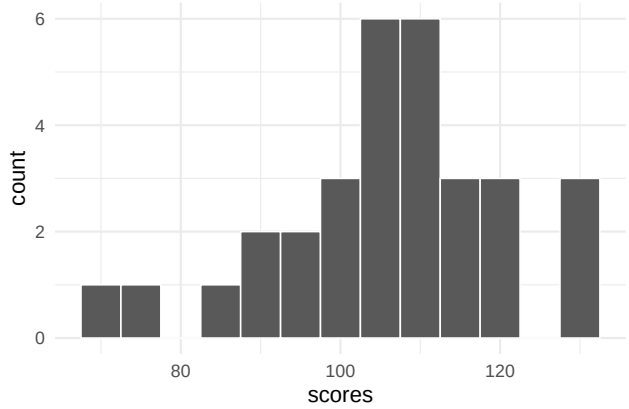
Estimate the mean IQ score μ for all seventh grade girls in this Midwest school district by giving a 95% confidence interval.

Example on IQ scores (pg. 354)

First check the three assumptions:

1. Normality: Can evaluate this using a histogram
2. SRS: Can only use information provided in the problem to assess with an SRS was taken.
3. Known σ: Can use information provided in the problem to determine if σ is known.

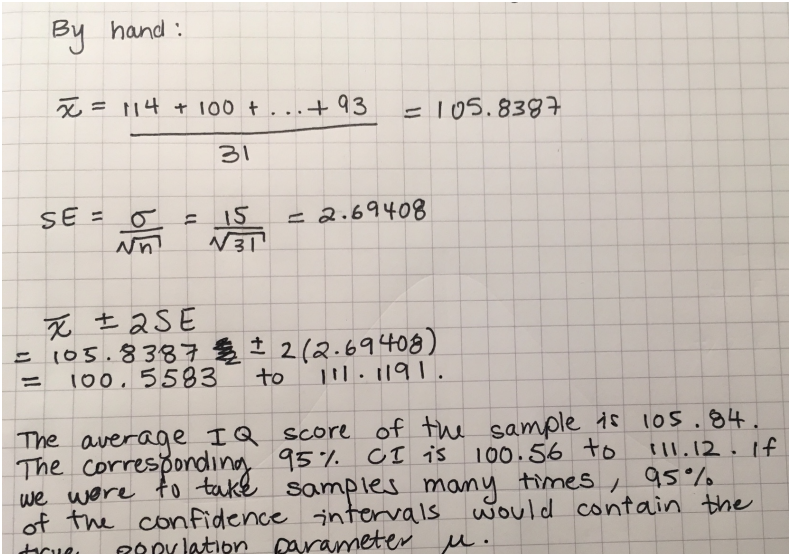
Checking Normality



We can't examine the Normality of the population (because we don't have data on everyone) 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



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
```

Calculating the estimated mean and its confidence interval

```
sample_mean
```

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

```
standard_error
```

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

```
lower_bound
```

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

```
upper_bound
```

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

Calculating the estimated mean and its confidence interval

Thus, our best estimate of the population mean is 105.84.

Its 95% confidence interval is 100.56 to 111.12.

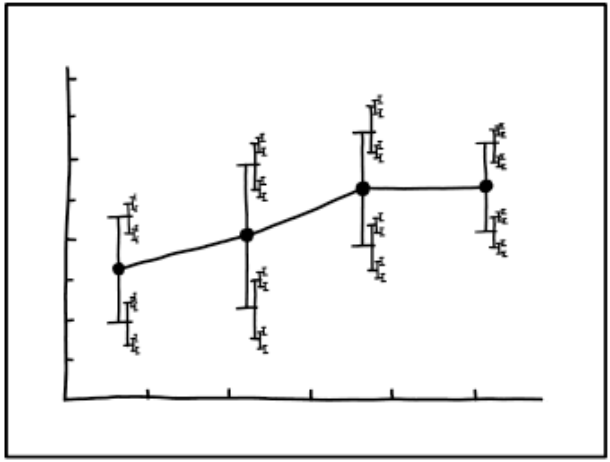
If our model assumptions are correct and there is only random error affecting the estimate, this method for calculating confidence intervals will contain the true value μ 95% of the time (19 times out of 20).

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.
- We learned how to interpret the confidence interval and the definitions for the confidence level and the margin of error
- We introduced our "recipe" for a margin of error

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

Central limit theorem



I DON'T KNOW HOW TO PROPAGATE ERROR CORRECTLY, SO I JUST PUT ERROR BARS ON ALL MY ERROR BARS.