

Week 6 Discussion (Group)

Group 4

2025-10-07

General section

1. What does a confidence interval tell us about a population parameter?

A confidence interval is an “interval estimate” that gives a range of plausible values for the population and reflects ‘sampling variability’. For example you could say “We are 95% confident the population mean is between 80 and 93.”

2. If we have a 95% confidence interval of [2.5, 3.5] for a population mean, how should we interpret this interval?

We are 95% confident that the population mean lies between 2.5 and 3.5. Another way to also interpret this is if you were to repeat this sampling process several times, 95% of the confidence intervals calculated in the same way would contain the population true mean. True mean = the population mean and reflects the actual average of the whole population, which is usually unknown and we try to estimate it through sampling.

3. How does changing the confidence level (e.g., from 90% to 99%) affect the width of the confidence interval? Why does this happen?

Changing the confidence level from 90% to 99% would affect the width of the CI by it becoming wider, not shorter. This happens due to it having a larger “critical value” which is used in the following formula. (Margin of error = critical value x standard error). A larger critical value (90% -> 99%) would then make the margin of error wider and therefore the width.

4. How is the margin of error related to confidence intervals?

Margin of error is the half-width of a confidence interval. again, $ME = (\text{Critical Value}) \times (SE)$. Increasing the confidence level raises the critical value, so the margin of error and therefore the interval would get wider.

Real-life paper section

1. Find a research paper where confidence intervals are used to present findings. Use APA citations.

Lininger, M. R., & Riemann, B. L. (2016). Statistical primer for athletic trainers: Using confidence intervals and effect sizes to evaluate clinical meaningfulness. *Journal of Athletic Training*, 51(12), 1045-1048. <https://doi.org/10.4085/1062-6050-51.12.14>.

2. Analyze how well the confidence interval is explained in the report.

In the paper the concept of the confidence interval (CI) is mentioned several times and is explained very well, especially for readers who can relate or work in clinical and health fields. The authors begin by defining a CI as a “range of values within which the population value is believed to lie at a given level of statistical confidence”. The Author then goes beyond the mathematical or technical definition by emphasizing the practical importance helping clinicians understand not only if a treatment effect exists but also how precise that estimate is. This is continued to be strengthened by a practical example involving ankle dorsiflexion range of motion (ROM), which shows how the CI can help determine whether an observed difference is likely to be meaningful in a clinical setting. The author explains that a narrow CI (e.g. 5°-8°) indicates a more precise and useful estimate, while a wide CI (e.g. 1°-15°) leads to uncertainty and limited decision-making value. They also discuss factors influencing CI width, such as sample size, variability, and confidence level, and caution that even a 95% CI does not guarantee every patient will experience the same result. Overall, the report explains confidence intervals in a thorough, logical, and clinically relevant manner, making a complex statistical concept accessible and meaningful for practitioners.

3. How does the variability of the data impact the confidence interval?

The variability of the data has a direct influence on the width of the confidence interval (CI). When there is greater variability in the sample data, which means that participants respond differently to the treatment then the CI becomes wider, indicating less precision in estimating the true population effect. This happens because inconsistent responses among individuals would make it harder to determine the impact of the intervention. On the other hand, a lower variability results in a narrower CI, providing a more precise estimate of the true effect. The authors state that “variability within the sample indicates that the treatment effect was not consistent among study participants, which can occur when inclusion or exclusion criteria are broad”. Therefore, reducing variability in a study helps produce narrower confidence intervals and allows clinicians to interpret results with greater confidence in their accuracy.

4. What might happen if the confidence interval is too wide to provide meaningful insights?

The variability of the data directly affects the width and precision of the confidence interval (CI). According to the paper, when there is greater variability among study participants, this means that individuals respond differently to the intervention and this is reflected by the confidence interval becoming wider. This happens because inconsistent responses make it harder to estimate the

true population effect with precision. In contrast, when data variability is low, which means the participants respond more repetitively, the CI becomes narrower and provides a more precise estimate of the treatment's effect. The authors explain that variability within the sample indicates that the treatment effect was not consistent among study participants, which can occur when inclusion or exclusion criteria are broad. They note that wide intervals caused by high variability make it difficult for clinicians to be confident about the actual benefit of an intervention. This means that minimizing variability through careful study design and participant selection would lead to narrower, more useful confidence intervals for clinical decision-making.

Real-life Case Study Section

1. Explore a real-life case study where biased sampling led to incorrect conclusions

A famous historical example of biased sampling occurred during the 1936 U.S. presidential election, when *The Literary Digest* magazine conducted one of the largest opinion polls in history. The magazine sent out more than ten million questionnaires to predict whether Franklin D. Roosevelt or Alf Landon would win the presidency. Approximately 2.4 million responses were received, and the poll predicted a decisive victory for Landon. However, the actual election results were the opposite - Roosevelt won in a landslide.

The failure of this prediction stemmed from selection bias and nonresponse bias. The sample was drawn from automobile registration lists, telephone directories, and magazine subscription records. During the Great Depression, people who owned cars or telephones tended to be wealthier and thus more likely to vote Republican. On top of that, individuals who responded voluntarily were not representative of the overall population, adding to the distortion (*Squire, 1988*). This case demonstrates that a massive sample size cannot overcome a fundamentally biased sampling.

2. What were the consequences, and how could the issue have been avoided?

The poll's inaccuracy had serious consequences for both the magazine's credibility and the public's trust in polling. *The Literary Digest's* reputation was completely damaged, leading to a decline in readership and eventual closure. In contrast, George Gallup, who used a smaller but scientifically randomized sample, correctly predicted Roosevelt's victory. This event marked a turning point in modern statistics and public opinion research, proving that representation matters more than just sample size (*Bryson, 1976*).

The issue could have been avoided through probability sampling instead of convenience sampling. Specifically:

- Drawing a random sample that represented various socioeconomic groups.
- Using stratified sampling to ensure proportional representation from urban and rural populations.
- Conducting follow-ups to reduce nonresponse bias.

3. Why is it important to use proper sampling methods in statistics?

Proper sampling methods are essential because statistical inferences depend on how well a sample represents the population. Without representation, even precise calculations like means, confidence intervals, and margins of error can be misleading. Sound sampling allows researchers to generalize findings with confidence and minimize systematic error. The *Literary Digest* poll demonstrated that biased data collection can lead to false conclusions, even when millions of data points are available. Good sampling design is thus the foundation of reliable data analysis.

4. Discuss the risks of bias in sampling

- **Selection Bias:** Occurs when some groups are more likely to be included than others. For example, in 1936, only wealthier Americans were sampled because they owned telephones or cars.
- **Nonresponse Bias:** Individuals who fail to respond may differ systematically from those who do; in the Digest poll, low-income voters were underrepresented.
- **Measurement Bias:** Poorly designed questions or data collection methods can influence responses.
- **Convenience Bias:** Relying on easily accessible participants rather than a random cross-section of the population.

Each of these biases distorts statistical estimates and weakens the validity of conclusions. Recognizing and minimizing bias ensures that results are accurate, generalizable, and trustworthy.

References (APA 7th Edition)

Bryson, L. (1976). *The Literary Digest poll: Making of a statistical myth*. *Public Opinion Quarterly*, 40, 431-436. <https://doi.org/10.1086/268319>

Squire, P. (1988). *Why the 1936 Literary Digest poll failed*. *Public Opinion Quarterly*, 52, 125-133. <https://doi.org/10.1086/269082>

Group dataset section:

- Describe one of your datasets and research question(s). Use APA citations (where necessary).
- Select a key variable from this dataset. Describe the variable.
- Calculate the 95% confidence interval for the mean of the variable.
- Discuss how the intervals vary based on changes in sample size, variability, and confidence levels.
- Discuss how sample size affects the confidence interval width.
- Is it possible to have a narrow confidence interval with a small sample size?
- Reflect on the accuracy of using small versus large samples and what role statistical sampling plays in ensuring reliable results.

Group Dataset Section

1. Describe one of your datasets and research question(s)

Our group used the Coronavirus Dataset - France from Kaggle (*Mclikmb4, 2021*), which compiles daily COVID-19 indicators reported by French public health officials. The dataset includes variables like *date*, *new_cases*, *new_deaths*, *total_cases*, *total_deaths*, *confirmed_cases*, and *hospitalizations*. It provides a detailed overview of how COVID-19 evolved across France from 2020 onward, including both absolute and population-adjusted values.

Research Question:

What is the average number of confirmed COVID-19 cases in France, and how confidently can we estimate this population mean using a 95% confidence interval?

This question allows us to explore how confidence intervals and sampling theory help interpret real-world data reliability.

2. Select a key variable from this dataset and describe it

We selected the variable *confirmed_cases*, which measures the number of newly confirmed COVID-19 cases each day.

- **Type:** Quantitative, continuous variable
- **Units:** Confirmed cases per day
- **Purpose:** This variable reflects the spread of the virus and helps determine trends in transmission and infection over time.

By focusing on *confirmed_cases*, we can better understand the daily progression of the pandemic across France.

3. Calculate the 95% confidence interval for the mean of the variable

95% Confidence Interval:

- Lower Bound: 40978.48
- Upper Bound: 375073.8

Interpretation:

We are 95% confident that the true mean number of confirmed COVID-19 cases in France lies between the lower and upper bounds of the calculated interval. This means that if several random samples were drawn, approximately 95% of those intervals would contain the actual population mean.

4. Discuss how the intervals vary based on changes in sample size, variability, and confidence levels

- **Sample Size:** Increasing the sample size narrows the confidence interval because the standard error decreases, resulting in greater precision.
- **Variability:** Higher variability widens the interval and makes data points more spread out around the mean.
- **Confidence Level:** Raising the confidence level from something like 95% to 99% widens the interval because more certainty needs a larger margin of error.

Confidence Interval Comparison

90% Confidence Interval:

- Lower Bound: 68661.41
- Upper Bound: 347390.9

99% Confidence Interval:

- Lower Bound: -14747.17
- Upper Bound: 430799.5

As you can see from the test above, the 90% CI has a range of (68,661.41 | 347,390.9). Whereas the 99% CI has a range of (-14,747.17 | 430,799.5). As you can see from the results, as the confidence interval scales, the interval range also scales to be wider.

5. Discuss how sample size affects the confidence interval width

The width of a confidence interval decreases as the sample size increases, following the formula: Margin of Error = $t * (s / \sqrt{n})$.

Because the sample size (n) appears in the denominator, increasing n makes the overall error smaller. For instance, if you take a small sample of around 30 observations, the confidence interval might be pretty wide. If you increase the sample size to about 100 observations, the interval becomes much narrower. With a much larger sample of 300 observations, the interval becomes even tighter. This helps to show that larger samples have tighter and more reliable confidence intervals, even when variability remains constant.

6. Is it possible to have a narrow confidence interval with a small sample size? Reflect on accuracy and sampling reliability

No, it's not really possible to have that. Narrow confidence intervals with small sample sizes are unusual and can be misleading. A small sample may produce a tight interval only when variability is extremely low, but it would be very rare in real-world data like COVID-19 case counts.

Larger samples reduce random error and provide more reliable estimates of the true population mean. This is important in public health data because small or biased samples might exaggerate or underestimate infection trends, leading to inaccurate conclusions and poor health decisions. Proper statistical sampling helps to make sure representation, stability, and meaningful confidence levels so the results we get are actually correct and actionable.

7. References (APA 7th Edition)

mcikmb4, (2021, April 4), Coronavirus-dataset France, Kaggle, <https://www.kaggle.com/datasets/mcikmb4/coronavirusdataset-france?select=chiffres-cles.csv>