STAT 22000 Lecture Slides
Overview of Confidence Intervals

Yibi Huang
Department of Statistics
University of Chicago

Outline

This set of slides covers section 4.2 in the text

- Overview of Confidence Intervals

Confidence intervals

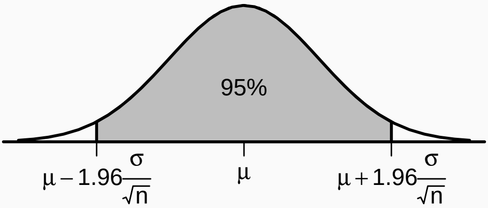
- A plausible range of values for the population parameter is called a *confidence interval*.
- Using only a sample statistic to estimate a parameter is like fishing in a murky lake with a spear, and using a confidence interval is like fishing with a net.

We can throw a spear where we saw a fish but we will probably miss. If we toss a net in that area, we have a good chance of catching the fish.

- If we report a point estimate, we probably won't hit the exact population parameter. If we report a range of plausible values we have a good shot at capturing the parameter.

Photos by Mark Fischer (http://www.flickr.com/photos/fischerfotos/7439791462) and Chris Penny

Recall that CLT says, for large n, $\overline{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$. For a normal curve, 95% of its area is within 1.96 SDs from the center. That means, **for 95% of the time,** $\overline{X}$ **will be within** $1.96\frac{\sigma}{\sqrt{n}}$ **from** μ.



Alternatively, we can also say, **for 95% of the time,** μ **will be within** $1.96\frac{\sigma}{\sqrt{n}}$ **from** $\overline{X}$.

Hence, we call the interval

$$\overline{X} \pm 1.96\frac{\sigma}{\sqrt{n}} = \left(\overline{X} - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

a **95% confidence interval for** μ.

Procedures to Construct a 95% Confidence Interval for μ

1. Take a simple random sample (or i.i.d. sample) of size n and find the sample mean $\overline{X}$.
2. If n is large, the 95% confidence interval for μ is given by

$$\overline{X} \pm 1.96 \frac{\sigma}{\sqrt{n}}$$

But σ is usually unknown ...

But σ is usually unknown ...

When the population SD σ is unknown, we replace it with our best guess — the sample SD s. So an approximate 95% confidence interval for μ is

$$\overline{X} \pm 1.96\, s/\sqrt{n}$$

- However, this replacement is hazardous because
  - s is a poor estimate of σ if the sample size n is small and
  - s is very sensitive to outliers
- So we require $n \geq 30$ and sample shouldn't have any outlier nor be too skewed $\Rightarrow$ Need to check histogram of the data
- We will discuss working with samples where $n < 30$ in the next chapter

Other Conditions Required to Use a Confidence Interval

Independence: Observations in the sample must be independent

- If the observations are from a simple random sample and consist of $< 10\%$ of the population, then they are nearly independent.
- Subjects in an experiment are considered independent if they undergo random assignment to the treatment groups.
- If a sample is from a seemingly random process, e.g. the lifetimes of wrenches used in a particular manufacturing process, checking independence is more difficult. In this case, use your best judgement.

Example: Average number of exclusive relationships

A random sample of 50 college students were asked how many exclusive relationships they have been in so far. This sample yielded a mean of 3.2 and a standard deviation of 1.74. Estimate the true average number of exclusive relationships using this sample.

$$\bar{x} = 3.2 \qquad s = 1.74$$

The approximate 95% confidence interval is about

$$\begin{aligned} \bar{x} \pm 1.96 \times SE &= \bar{x} \pm 1.96 \times \frac{s}{\sqrt{n}} \\ &= 3.2 \pm 1.96 \times \frac{1.74}{\sqrt{50}} \\ &\approx 3.2 \pm 0.5 = (2.7, 3.7) \end{aligned}$$

True or False

True or False and explain: We are 95% confident that the average number of exclusive relationships college students in this sample have been in is between 2.7 and 3.7.

False. The confidence interval $\overline{X} \pm 1.96SE$ definitely (100%) contains the sample mean $\overline{X}$, not just with probability 95%.

True or False and explain: 95% of college students have been in 2.7 to 3.7 exclusive relationships.

False. The confidence interval is for covering the population mean μ, not for covering 95% of the entire population. If 95% of college students have been in 2.7 to 3.7 exclusive relationships, the SD won't be as large as 1.74.

True or False

True or False and explain: There is 0.95 probability that the true mean number of exclusive relationships of college students falls in the interval $(2.7, 3.7)$

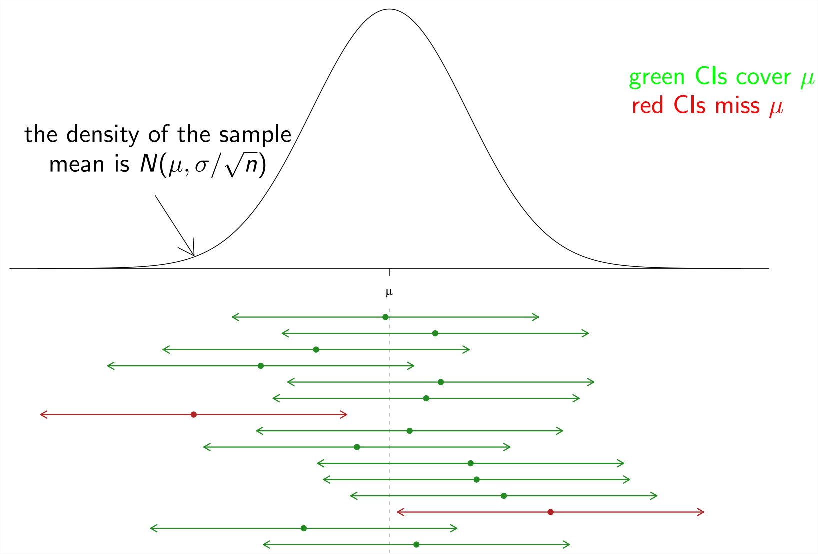
True or False and explain: The interval $(2.7, 3.7)$ has probability of 0.95 of enclosing the true mean number of exclusive relationships of college students.

Both are False. The population mean μ is a fixed number, not random. It is either in the interval (2.7, 3.7), or not in the interval. There is no uncertainty involved.

What does “95% confidence” mean?

What is the thing that has a 95% chance to happen?

- It is the **procedure to construct the 95% interval**.
- About 95% of the intervals constructed following the procedure (taking a SRS and then calculating $\overline{X} \pm 1.96\, s/\sqrt{n}$) will cover the true population mean μ.
- After taking the sample and an interval is constructed, the constructed interval either covers μ or it doesn’t. We don’t know. Only God knows.
- Just like lottery, before you pick the numbers and buy a lottery ticket, you have some chance to win the prise. After you get the ticket, you either win or lose.


green CIs cover μ
red CIs miss μ
the density of the sample mean is $N(\mu, \sigma/\sqrt{n})$
μ

True or False

True or False and explain: If a new random sample of size 50 is taken, we are 95% confident that the new sample mean will be between 2.7 and 3.7.

False. The confidence interval is for covering the population mean μ*, not for covering the mean of another sample. The SE* $\sigma/\sqrt{n}$ *or* $s/\sqrt{n}$ *is a typical distance between the sample mean and population mean, not a typical distance between two sample means.*

True or False

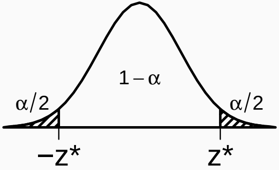
True or False and explain: This confidence interval $\overline{X} \pm 1.96\, s/\sqrt{n}$ is not valid since the number of exclusive relationships is integer-valued. Neither the population nor sample is normally distributed.

False. The construction of the CI $\overline{X} \pm 1.96\, s/\sqrt{n}$ only uses the normality of the sampling distribution of the sample mean. Neither the population nor the sample is required to be normally distributed. By the central limit theorem, with a large enough sample size we can assume that the sampling distribution is nearly normal and calculate a confidence interval.

Confidence Intervals at Other Confidence Levels

For a given confidence level $(1 - \alpha)$, we want to find a z^* such that

$$P(-z^* < Z < z^*) = 1 - \alpha \quad \text{or}$$



In general, a confidence intervals at confidence level $(1 - \alpha)$ is

$$\text{sample mean} \pm z^* SE$$

- $z^\star \times SE$ is called the *margin of error*

Commonly used confidence levels:

- 90% C.I.: $\alpha = 0.1$, $z^* = 1.645$
- 95% C.I.: $\alpha = 0.05$, $z^* = 1.96$
- 99% C.I.: $\alpha = 0.01$, $z^* = 2.58$

Margin of Error

THE STAR VANCOUVER

MP-commissioned poll finds 12 per cent of British Columbians would engage in civil disobedience

There's 'a deep, deep frustration with the Trudeau government' over pipeline, Kennedy Stewart says.

By Ainslie CruickshankStarMetro Vancouver
Sat., April 28, 2018

VANCOUVER — Twelve per cent of British Columbians are willing to engage in civil disobedience to oppose the Trans Mountain expansion project, a new poll has found, underscoring what a Burnaby MP says is a "deep frustration" with the federal government.

The online poll, conducted this month by Insights West and commissioned by NDP MP Kennedy Stewart, asked 1,021 people in B.C. between April 16-18 if they would consider civil disobedience to stop or disrupt the pipeline's construction. It found men and women were equally likely to consider civil disobedience, a release said.

. . . . . . (several lines omitted). . . . . .

The poll had a margin of error of plus or minus three percentage points 19 times out of 20.

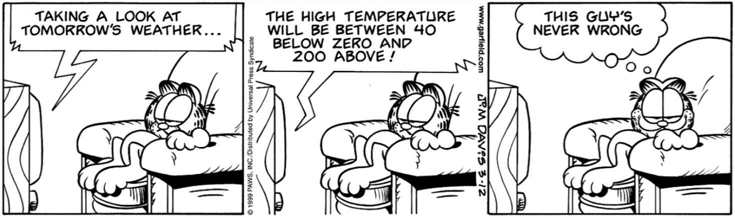
95% CI for the percentage of British Columbians that are willing to engage in civil disobedience to oppose the pipeline's construction is $12\% \pm 3\%$.

Width of an Interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

Can you see any drawbacks to using a wider interval?



If the interval is too wide it may not be very informative.

Image source: http://web.as.uky.edu/statistics/users/earo227/misc/garfield_weather.gif