Lecture 16: Confidence intervals

Criminology 250

Prof Maria Cuellar

University of Pennslyvania

ggplot

- How does ggplot work?

- Esquisse: A nice add-in that helps create plots using ggplot more easily.

https://dreamrs.github.io/esquisse/

Exam 2 grading notes

- Great answers! I like how you said your intuition is that probably there is a relationship, but this linear model didn't have proper fit so something else might be going on. It requires further analysis.
- I like how you separated between different types of problems:
  - Model assumptions not satisfied
  - Relationship between variables is not significant
  - Can't make causal conclusions because we don't know if there are unobserved confounders
- Some of you interpreted p-values incorrectly. We'll go over this today in class.

Exam 2 grading notes

- Independence: Cannot test this directly, although clumping might be evidence of it. Usually we need to think about the context (e.g. could the participants have been affecting each other for some reason?)
- Transformations: Yes, wages are usually logged, but do you understand why?
- Language about assumptions: You don't "satisfy" or "prove" an assumption. You can say, "To test assumption X, I have drawn this plot. Based on the plot, I have evidence to say that the assumption is not satisfied."
- Assumptions can be "weak" or "strong". Usually weak is better, but harder to satisfy. You can also "weaken" assumptions.
- What if some assumptions are met and others aren't? It's a gamble, but try to have them all be satisfied. Normality might be better with more data.

Exam 2 grading notes

- Can stick with correlation: that's fine. But must be careful about inference (size of estimated coefficient and p-value).
- If you want a title to be split up into two lines, use \n in the title.
- Make sure your plots *and* labels are large enough to be seen and read.
- In your EDA, sometimes plots or tables are not very informative. Always try to think, what should the reader take from this graph? In your EDA it's fine just to get a sense for what's going on in the data. In a research report, it should usually be something surprising, something the reader didn't already know.
- Although I like your use of color, try to keep it professional. Usually, pastel tones are used if you want to make graphs look good. Otherwise, tones like red and yellow are used when you're trying to point something specific out to the reader.
- Use professional language.

What is a linear model?

- Usually it's where the variables are linearly related to each other, as in $y = mx + b$, but you can make this much more flexible, so actually many of the models used in statistics and machine learning *are* linear models.
- You can play the game, IJALM (it's just a linear model): https://threadreaderapp.com/thread/1286420597505892352.html

How to interpret p-values correctly

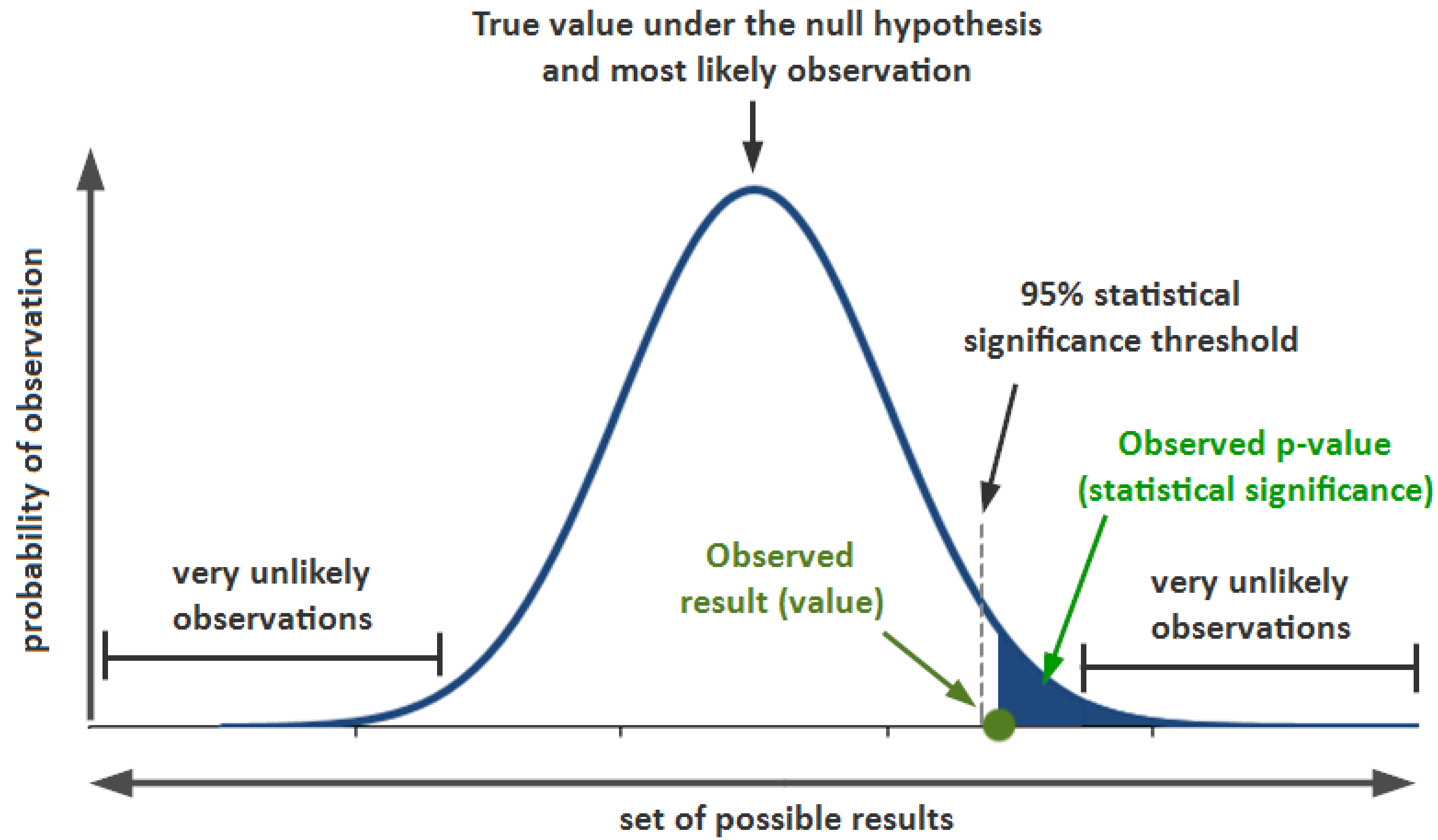
Correct:

- "If the model assumptions are satisfied, there is evidence that we can reject the null hypothesis in favor of the alternative at the 5% level."
- "If the model assumptions are satisfied, if we assume that there is no effect, you'd obtain the observed coefficient (or effect) or larger in 3% of studies because of random sample error.
- How probable are your sample data if the null hypothesis is correct? That’s the only question that p-values answer.

Incorrect:

- $p<0.05$: "We can state that the observed result was not due to random chance with 95% confidence."
- $p=0.03$: "There's a 3% chance of making a mistake by rejecting the null hypothesis."
- The idea that p-values are the probability of making a mistake is WRONG! They are also not error rates.

p-value


True value under the null hypothesis
and most likely observation
95% statistical
significance threshold
Observed p-value
(statistical significance)
probability of observation
very unlikely
observations
Observed
result (value)
very unlikely
observations
set of possible results

Confidence intervals definition

Besides p-values, another way to express what the evidence of an experiment is telling us is to compute confidence intervals (CI):

$$CI = \text{Estimate} \pm \text{Margin of error} = (\text{Estimate} - ME,\ \text{Estimate} + ME).$$

Confidence interval definiton for mean

- First: difference between a *parameter* (e.g. μ) and a *sample estimate* (e.g. $\overline{X}$.

Definition (for a mean):

A (1- α)100% (e.g. 95%) confidence interval for the population parameter of the mean, μ, is given by

$$\overline{X} \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$

A 95% CI has $z_{\alpha/2} = 1.96 \approx 2$:

$$\overline{X} \pm 2 \times \frac{\sigma}{\sqrt{n}}$$

Confidence interval definiton for proportion

Definition (for a proportion):

A (1- α) 100% (e.g. 95%) confidence interval for the population parameter of a proportion, p, is given by

$$\hat{p} \pm 2\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

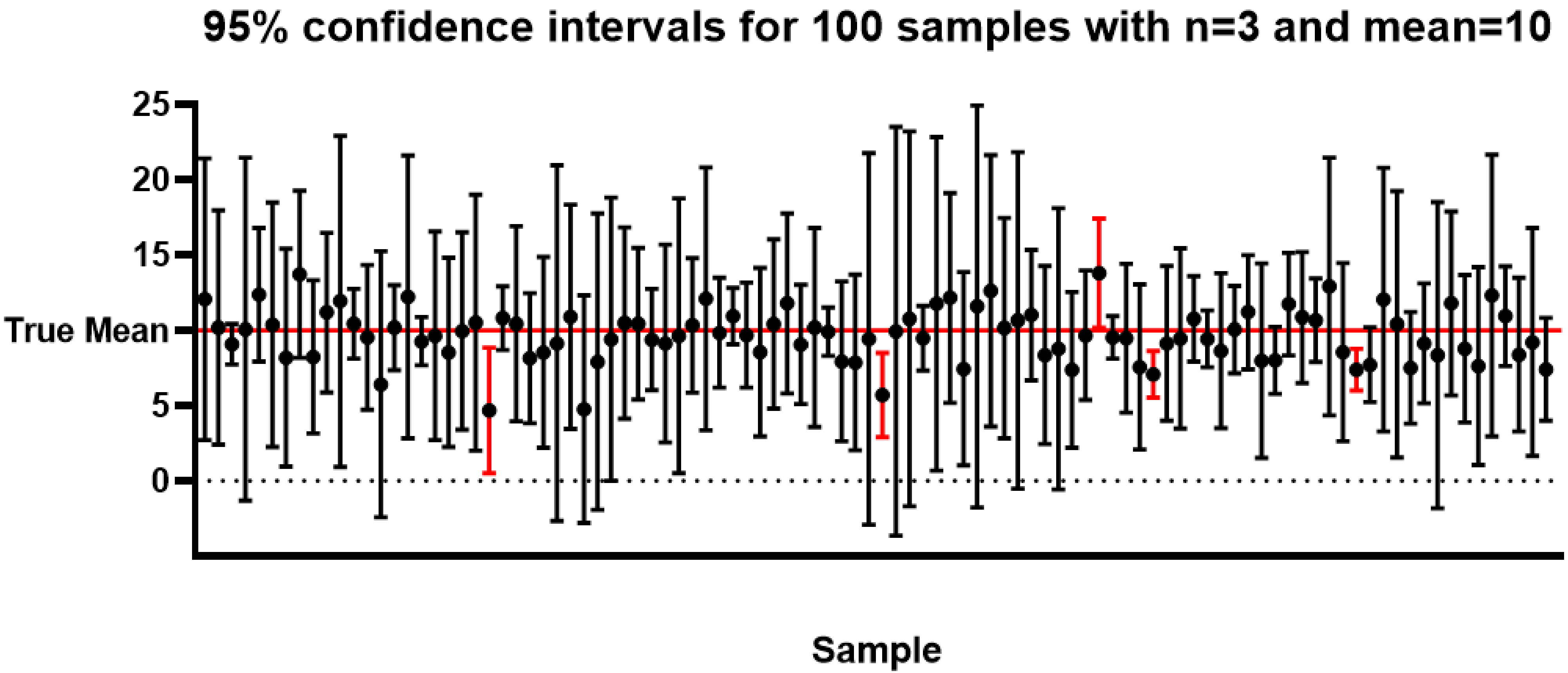
Example of confidence interval

- **Example**: Every day, newspapers report opinion polls. For example, they might say that **"83 percent of the population favor arming pilots with guns."** Usually, you will see a statement like **"this poll is accurate to within 4 points 95 percent of the time."** They are saying that 83 ± 4 is a 95 percent confidence interval for the true but unknown proportion p of people who favor arming pilots with guns. If you form a confidence interval this way every day for the rest of your life, 95 percent of your intervals will contain the true parameter. This is true even though you are estimating a different quantity (a different poll question) every day. (Wasserman, All of Statistics.)

How to understand the meaning of 95% CI?

- Interpretation: If all the assumptions are met, and if we repeat the experiment many times, the *random* interval that we compute each time will contain the single, fixed, true parameter value 95% of the time.
- But we probably won't repeat the same experiment over and over, so this interpretation is pretty useless.
- Instead, it's also correct to say that if you use the same method of computing CIs, and you calculate CIs for many different samples of different datasets, then 95% of your intervals will trap the true parameter value. There is no need to introduce the idea of repeating the same experiment over and over. (Wasserman, All of Statistics.)

Visualizing a confidence interval



Confidence intervals from lm model

You can use R to translate the p-values from a linear model into confidence intervals:

```
dat <- read.csv(file = 'sim.data.csv')
reg.output <- lm(po.brut ~ funds, dat)
confint(reg.output)
```

```
##                  2.5 %     97.5 %
## (Intercept) 39.9859678 41.1001708
## funds       -0.3759656 -0.3582317
```

How to interpret confidence intervals correctly

Correct:

- "95% of samples of this size will produce confidence intervals that capture the true proportions."
- "We are 95% confident that the true population parameter lies in our interval."

Incorrect:

- 95% CI (A to B), there is a 95% probability that the true population mean lies between A and B. (Actually, the mean is either in the interval or it's not. What is random is the interval itself.)
- Everything we have seen in this class and these slides so far is frequentist (or classical) statistics. There is also another branch of statistics that is Bayesian.
- Bayesian credible intervals are what we wish frequentist intervals did: "The probability that μ is in the confidence interval, given the data, is 95%." Note: Bayesian intervals will not, in general, trap the parameter 95% of the time.

Relationship between p-values and confidence intervals

- CIs and p-values are closely related although they provide different information. While p-values are the outcome of hypothesis tests and indicate whether or not the sample data provide sufficient evidence to reject the null hypothesis, CIs indicate whether the estimate is a precise one or only a very “rough” estimate.
- Usually CIs are more informative than p-values.
- How are they calculated? If the underlying data are normally distributed (or have a large sample size for a sum or mean, so we can call the CLT), then a confidence interval for a statistic is the statistic plus or minus the appropriate "multiplier" times the esimated standard error of the quantity.
- The multiplier depends on the desired confidence level (e.g., 95, 90), and the degrees of freedom for the standard error.
- The extent of the interval on either side of the estimate is called the margin of error (ME).