Confidence intervals

Patrick Breheny

January 25

Limits of hypothesis testing

- In our last lecture, we saw that p-values provide a simple way of testing the hypothesis that an observed difference is due entirely to chance
- This is useful, but as we saw, there are clear limitations:
  - Suppose we reject the null hypothesis that a treatment is completely ineffective; we would of course want to know *how effective* the treatment is
  - Suppose we don't reject the null hypothesis; what *can* we conclude?
- Hypothesis testing provides no answers to these questions; to address them we need confidence intervals

Why we would like an interval

- In our polio vaccine study, we saw 28 cases per 100,000 in the vaccine group and 71 cases per 100,000 in the control group
  - What we know: People in our sample were 2.5 times less likely to contract polio if vaccinated
  - What we want to know: How much less likely would the rest of the population be to contract polio if they were vaccinated?
- This second number is almost certainly different from 2.5 – maybe by a little, maybe by a lot
- Since it is highly unlikely that we got the exactly correct answer in our sample, it would be nice to instead have an interval that we could be reasonably confident contained the true number (the parameter)

What is a confidence interval?

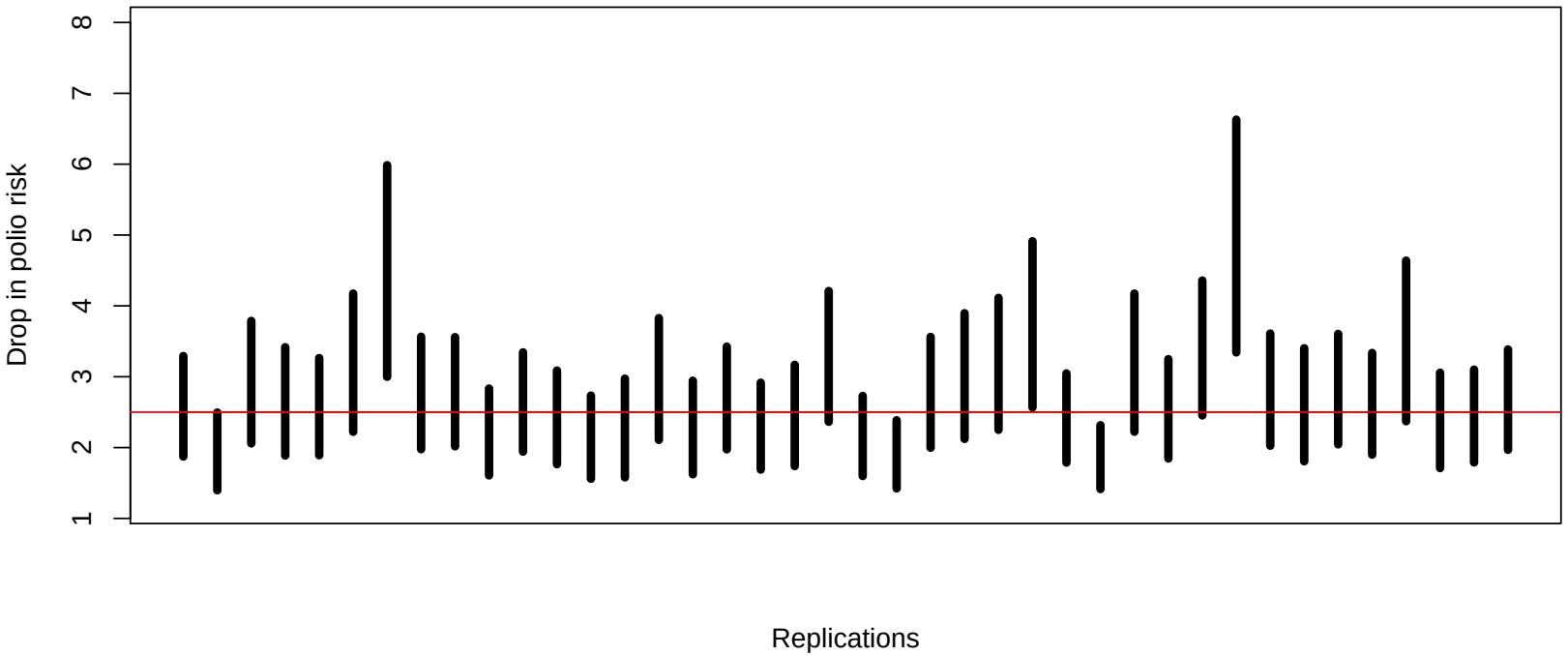
- It turns out that the interval (1.9,3.5) does this job, with a confidence level of 95%
- We will discuss the nuts and bolts of constructing confidence intervals often during the rest of the course
- First, we need to understand what a confidence interval is
- Why (1.9,3.5)? Why not (1.6,3.3)?
- And what the heck does "a confidence level of 95%" mean?

What a 95% confidence level means

- There's nothing special about the interval (1.9,3.5), but there is something special about the procedure that was used to create it
- The interval (1.9,3.5) was created by a procedure that, when used repeatedly, contains the true population parameter 95% of the time
- Does (1.9,3.5) contain the true population parameter? Who knows?
- However, in the long run, our method for creating confidence intervals will successfully do its job 95% of the time (it has to, otherwise it wouldn't be a 95% confidence interval), so this is how much confidence we can place in the interval

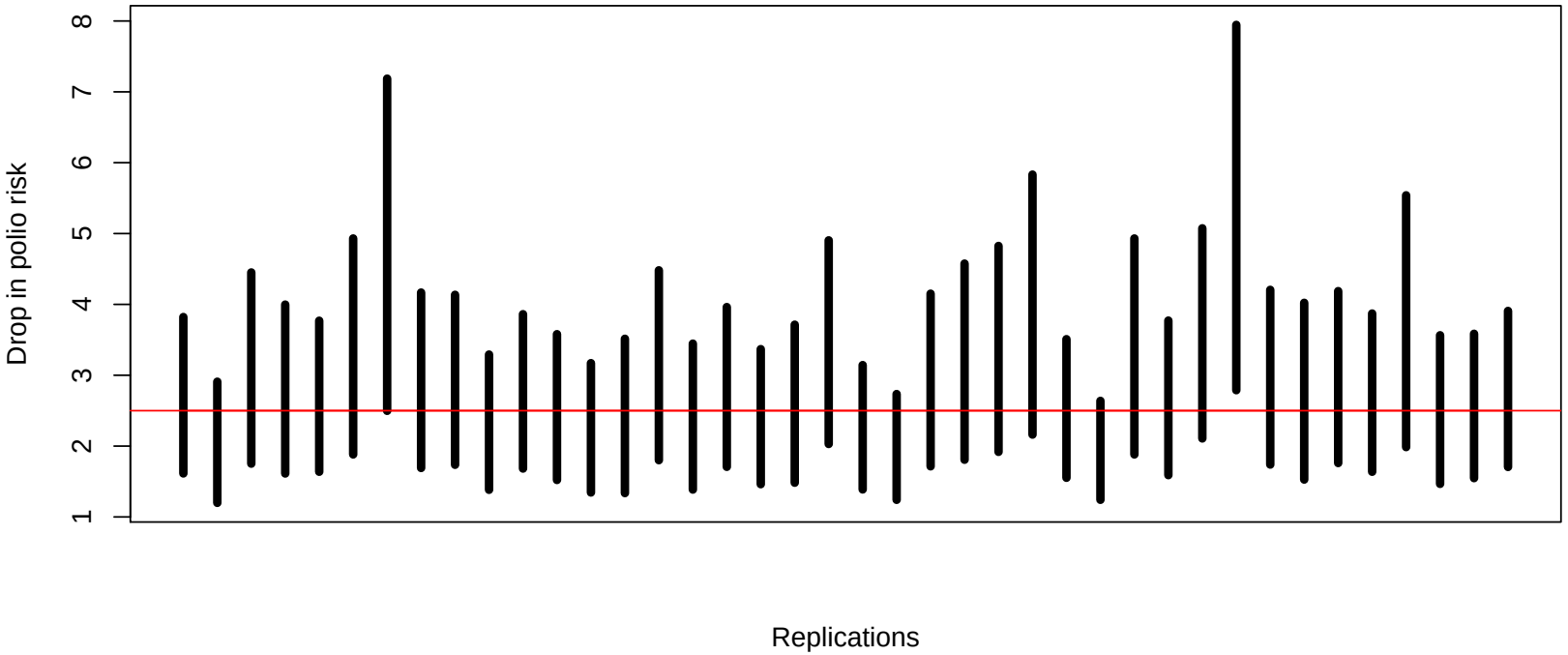
Simulated 80% confidence intervals

Imagine replicating the polio study 40 times (red line = truth):



Simulated 95% confidence intervals

Same studies, same data, difference confidence level:



What's special about 95%?

- The vast majority of confidence intervals in the world are constructed at a confidence level of 95%
- What's so special about 95%?
- Nothing
- However, it does make things easier to interpret when everyone sticks to the same confidence level, and the convention that has stuck in the scientific literature is 95%, so we will largely stick to 95% intervals in this class as well

Consequences

- Thus, if science as a whole goes about constructing these intervals, we can trust that its conclusions will be correct 95% of the time
- This is the sort of long-run guarantee that makes these intervals so appealing to the scientific community
- In reality, however, that percentage is somewhat lower than 95% due to factors such as incorrect assumptions and bias resulting from the experimental design
- For example, a 95% confidence interval for the results of the *Literary Digest* poll would be wrong nearly 100% of the time due to the fundamentally biased nature of the study

The subtle task of inference

Inference is a complicated business, as it requires us to think in a manner opposite than we are used to:

- Usually, we think about what will happen, taking for granted that the laws of the universe work in a certain way
- When we infer, we see what happens, then try to conclude something about the way that the laws of the universe must work

Confidence interval subtleties

- This subtlety leads to some confusion with regard to confidence intervals – for example, is it okay to say, "There is a 95% probability that the true reduction in polio risk is between 1.9 and 3.5"?
- Well, not exactly – the true reduction is some fixed value, and once we have calculated the interval (1.9,3.5), it's fixed too
- Thus, there's really nothing random anymore – the interval either contains it or it doesn't
- Now, is this an important distinction, or are we splitting hairs here? Depends on who you ask, and we'll talk about this more later in the course

What do confidence intervals tell us?

- So, in the polio study, what does the confidence interval of (1.9,3.5) tell us?
- It gives us a range of likely values by which the polio vaccine cuts the risk of contracting polio: it could cut the risk by as much as 3.5 times less risk, or as little as 1.9 times less risk
- However, it is unlikely that the vaccine increases the risk or has no effect
- Our conclusions would be very different if our confidence interval looked like (0.5,7), in which case our study would be inconclusive

The width of a confidence interval

- The width of a confidence interval reflects the degree of our uncertainty about the truth
- Three basic factors determine the extent of this uncertainty, and the width of any confidence interval:
  - The confidence level
  - The amount of information we collect
  - The precision with which the outcome is measured

Confidence levels

- As we saw, the width of a confidence interval is affected by whether it was, say, an 80% confidence interval or a 95% confidence interval
- This percentage is called the *confidence level*
- Confidence levels closer to 100% always produce larger confidence intervals than confidence levels closer to 0%
- If I need to contain the right answer 95% of the time, I need to give myself a lot of room for error
- On the other hand, if I only need my interval to contain the truth 10% of the time, I can afford to make it quite small

Amount of information

- It is hopefully obvious that the more information you collect, the less uncertainty you should have about the truth
- Doing this experiment on thousands of children should allow you to pin down the answer to a tighter interval than if only hundreds of children were involved
- It may be surprising that the interval is as wide as it is for the polio study since the sample size was $n = 400,000$
- However, keep in mind that a very small percentage of those children actually contracted polio – the 99.9% of children in both groups who never got polio tell us very little about whether the vaccine worked or not
- Only 198 children in the study actually contracted polio; this is the more meaningful measure here in terms of the amount of information we collected

Precision of measurement

- The final factor that determines the width of a confidence interval is the precision with which things are measured
- For example, consider a study of whether an intervention reduces blood pressure
- Blood pressure is quite variable, so researchers in such studies will often measure subjects' blood pressure several times at different points in the day, then take the average
- The average will be more precise than any individual measurement, and they will reduce their uncertainty about the effect of the treatment

p-values tell us about confidence intervals

- It may not be obvious, but there is a close connection between confidence intervals and hypothesis tests
- For example, suppose that we construct a confidence interval by systematically testing all possible values of the quantity we are interested in, and we include in our interval any value that is not rejected by a $p < 0.05$ rule (this actually is how a lot of confidence intervals are constructed)
- Thus, if our p-value was above 0.05, we know that a 95% confidence interval will contain the null hypothesis
- Alternatively, if $p < 0.05$, we know that a 95% confidence interval will not contain the null hypothesis
- This is true in general even if we don't literally construct the interval via hypothesis testing

Confidence intervals tell us about p-values

- Likewise, if we know the 95% confidence interval, we can say whether $p < 0.05$ or not:
  - If the confidence interval contains the null hypothesis, then $p > .05$
  - If it doesn't, then $p < .05$
- In general, a $100(1-\alpha)\%$ confidence interval tells us whether a p-value is above α or not

Summary

- Thus, confidence levels and hypothesis tests lead to similar conclusions
- For example, in our polio example, both methods indicated that the study provided strong evidence that the vaccine reduced the probability of contracting polio well beyond what you would expect by chance alone
- This is a good thing – it would be confusing otherwise
- However, the information provided by each technique is different: the confidence interval provides a range of values for a parameter of interest that are consistent with the data, while the hypothesis test measures whether a single specific value is consistent with the data

Confidence intervals tell us about effect size

- In the previous lecture, we said that hypothesis tests tell us nothing about the effect size
- For example, the p-value for Nexium vs. Prilosec was $p < 0.0001$, even though the difference in healing rates was only 90% vs. 87%
- Confidence intervals, on the other hand, tell us a great deal about possible effect sizes
- In the Nexium example, the confidence interval for the factor by which Nexium increases the healing rate above that of Prilosec is $(1.02, 1.06)$
- This interval tells us that although Nexium certainly provides a benefit, that benefit is rather small

CIs are useful even in the absence of significance

- We also said that high p-values do not allow us to draw any conclusions
- Confidence intervals, however, are still useful
- In this Women's Health Initiative breast cancer study, the confidence interval for the drop in risk was $(0.83, 1.01)$
  - The study suggests that a woman could likely reduce her risk of breast cancer by about 10% by switching to a low-fat diet
  - So, maybe a low-fat diet won't affect your risk of breast cancer (recall that $p = 0.07$)
  - On the other hand, it is highly unlikely that it increases risk, and could reduce a woman's risk of breast cancer by up to 17%

CIs and "proving" the null hypothesis

- In our previous lecture, we said that it was impossible to prove a hypothesis, only to disprove one
- With confidence intervals, however, we can explore this issue in a little more depth
- Specifically, let's consider a huge study of over 1.2 million children, which found that if we divide the rate of autism among vaccinated children by the rate of autism among unvaccinated children, we get 0.99 (i.e., the rates are almost exactly the same)

CI for vaccines and autism risk

- Furthermore, because this is such a large study, the 95% confidence interval is very narrow: (0.92, 1.06)
- Effectively, this is about as close as you can come to "proving" the null hypothesis: there is no evidence that vaccines increase the risk of autism, and we can even rule out the idea that they have a large effect on autism risk
- However, it is still true that we can never truly prove the null hypothesis; here, we cannot rule out the possibility that vaccines confer a very small increase in risk (on the order of shifting the probability of a child developing autism from 1.4% to 1.45%)
- And, of course, it's also possible that vaccines lower the risk of autism by a small amount

Comments

- As this example hopefully illustrates, we can *never* rule out the possibility that two groups might be just very slightly different from each other (i.e., conclude that the null hypothesis is true)
- In many scenarios, the null hypothesis is almost certainly not true – surely, receiving surgery and receiving a drug to treat a condition will not produce *exactly* the same success rate
- What a non-significant ($p > 0.05$) finding means, though, is that the 95% confidence interval will contain the null, and therefore, surgery might be better than drug, or drug might be better than surgery and it will be difficult to make that decision in the presence of uncertainty

Summaries: Introduction

- Often in this class, I will provide you with a description of a study and the data they collected, then ask you to carry out a hypothesis test, construct a confidence interval, and write a sentence summarizing the main findings of the study
- We haven't learned how to calculate p-values and confidence intervals, but I want to take a moment here to discuss writing a summary sentence

Summaries: Expectations

For the purposes of this class, your summary must include the following components:

- Describe the conclusion in terms of the scientific content of the study (i.e., do not use the words "null hypothesis")
- Indicate the strength of evidence/significance
- If an association is found, indicate the direction of association

You may also describe the effect size, or how different the two groups are – I would certainly not penalize you for this – but I will usually ask about effect size separately

Summaries: Examples

- REALLY BAD: $p < 0.05$, so we reject the null hypothesis.
- BAD: The study rejected the hypothesis that Nexium and Prilosec are equally good.
- GOOD: The study provides strong evidence that Nexium is more effective than Prilosec at treating heartburn.
- REALLY REALLY BAD: $p > 0.05$, so the null hypothesis is true.
- OK: The study failed to reject the hypothesis that diet isn't associated with cancer.
- GOOD: The study provided only borderline evidence that low-fat diets reduce the incidence of breast cancer. It is possible that diet has no effect, although it is also possible that low-fat diets have a small protective benefit.

Summary

- There is always a range (i.e., an interval) of values of a parameter that are consistent with the data
- A 95% confidence interval means that the procedure used to construct the interval will contain the true value 95% of the time
- The higher the desired confidence level, the wider we need to make the interval
- The width of a confidence interval is also affected by the amount of information we collect and the accuracy with which we collect it