Confidence intervals and hypothesis tests

Patrick Breheny

January 19
In our last lecture, we discussed at some length the Public Health Service study of the polio vaccine

We discussed the careful design of the study to ensure that human perception and confounding factors could not bias the results in favor of or against the vaccine

However, there was one factor we could not yet rule out: the role of random chance in our findings
Are our results generalizable?

- Recall that in the study, the incidence of polio was cut by $71/28 \approx 2.5$ times.
- This is what we saw in our sample, but remember – this is not what we really want to know.
- What we want to know is whether or not we can generalize these results to the rest of the world’s population.
- The two most common ways of addressing that question are:
  - Confidence intervals
  - Hypothesis testing
- Both methods address the question of generalization, but do so in different ways and provide different, and complimentary, information.
Why we would like an interval

- Not to sound like a broken record, but
  - 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).
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
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
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 intervals 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.
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: after all, hundreds of thousands of children were involved. 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 about 200 children in the study actually contracted polio, and these are the children who tell us how effective the vaccine is (note that 200 is a lot smaller than 400,000!)
The final factor that determines the width of a confidence interval is the precision with which things are measured.

I mentioned that the diagnosis of polio is not black and white – misdiagnoses are possible.

Every misdiagnosis increases our uncertainty about the effect of the vaccine.

As another 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.
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

This is difficult to do: as Sherlock Holmes puts it in *A Study in Scarlet*, “In solving a problem of this sort, the grand thing is to be able to reason backward.”
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. Is this an important distinction, or are we splitting hairs here? Depends on who you ask.
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.

But – and this is an important but – it is unlikely that the vaccine increases the risk, or has no effect, and that what we saw was due to chance.

Our conclusions may be very different if our confidence interval looked like (0.5, 7), in which case our study would be inconclusive.
Not all values in an interval are equally likely

- It is important to note, however, that not all values in a confidence interval are equally likely.
- The ones in the middle of the interval are more likely than the values toward the edges.
- One way to visualize this is with a multilevel confidence bar.
Specific values of interest

- Although confidence intervals are excellent and invaluable ways to express a range of likely values for the parameter an investigator is studying, we are often interested in a particular value of a parameter.
- In the polio study, it is of particular interest to know whether or not the vaccine makes any difference at all.
- In other words, is the ratio between the risk of contracting polio for a person taking the vaccine and the risk of contracting polio for a person who got the placebo equal to 1?
- Because we are particularly interested in that one value, we often want to know how likely/plausible it is.
Hypotheses

- The specific value corresponds to a certain *hypothesis* about the world.
- For example, in our polio example, a ratio of 1 corresponded to the hypothesis that the vaccine provides no benefit or harm compared to placebo.
- This specific value of interest is called the *null hypothesis* ("null" referring to the notion that nothing is different between the two groups – the observed differences are entirely due to random chance).
- The goal of hypothesis testing is to weigh the evidence and deliver a number that quantifies whether or not the null hypothesis is plausible in light of the data.
**p values**

- All hypothesis tests are based on calculating the probability of obtaining results as extreme or more extreme than the one observed in the sample, given that the null hypothesis is true.
- This probability is denoted $p$ and called the $p$-value of the test.
- The smaller the $p$-value is, the stronger the evidence against the null:
  - A $p$-value of 0.5 says that if the null hypothesis was true, then we would obtain a sample that looks like the observed sample 50% of the time; the null hypothesis looks quite reasonable.
  - A $p$-value of 0.001 says that if the null hypothesis was true, then only 1 out of every 1,000 samples would resemble the observed sample; the null hypothesis looks doubtful.
The scientific method

- Hypothesis tests are a formal way of carrying out the scientific method, which is usually summarized as:
  - Form a hypothesis
  - Predict something observable about the world on the basis of your hypothesis
  - Test that prediction by performing an experiment and gathering data

- The idea behind hypothesis testing and $p$-values is that a theory should be rejected if the data are too far away from what the theory predicts
There is a subtle but very fundamental truth to the scientific method, which is that one can never really prove a hypothesis with it – only disprove hypotheses.

In the words of Albert Einstein, “No amount of experimentation can ever prove me right; a single experiment can prove me wrong.”

Hence all the fuss with the null hypothesis:
The healthy application of the scientific method rests on the ability to rebut the arguments of skeptics, who propose other explanations for the results you observed in your experiment.

One important skeptical argument is that your results may simply be due to chance.

The $p$-value – which directly measures the plausibility of the skeptic’s claim – is the evidence that will settle the argument.
In the polio study, for the null hypothesis that contracting polio is just as probable in the vaccine group as it is in the placebo group, \( p = 0.000000008 \), or about 1 in a billion

So, if the vaccine really had no effect, the results of the polio vaccine study would be a one-in-a-billion finding

Is it possible that the vaccine has no effect? Yes, but very, very unlikely
As another example from class last week, let’s calculate a $p$-value for the clofibrate study, where 15% of adherers died, compared with 25% on nonadherers.

The $p$-value turns out to be 0.0001.

So the drop in survival is unlikely to be due to chance, but it isn’t due to clofibrate either: recall, the drop was due to confounding.

It is important to consider the entire study and how well it was designed and run, not just look at $p$-values (FYI: the $p$-value comparing Clofibrate to placebo was 0.51).
Confidence intervals tell us about $p$-values

- It may not be obvious, but there is a close connection between confidence intervals and hypothesis tests.
- For example, suppose we construct a 95\% confidence interval.
- Whether or not this interval contains the null hypothesis tells us something about the $p$-value we would get if we were to perform a hypothesis test.
- If it does, then $p > .05$.
- If it doesn’t, then $p < .05$. 

Patrick Breheny  
STA 580: Biostatistics I
\textit{p}-values tell us about confidence intervals

- We can reason the other way around, also:
- If we get a \textit{p}-value above .05, then the 95\% confidence interval will contain the null hypothesis (and vice versa)
- In general, a $100(1 - \alpha)\%$ confidence interval tells us whether a \textit{p}-value is above $\alpha$ or not
Conclusion

- In general, then, 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 is an attempt to provide likely values for a parameter of interest, while the hypothesis test is an attempt to measure the evidence against the hypothesis that the parameter is equal to a certain, specific number.
To be sure, there are shades of gray when it comes to interpreting $p$-values; how low does a $p$-value have to be before one would say that we’ve collected sufficient evidence to refute the null hypothesis?

Suppose we used a cutoff of .05

If $p < .05$ and the null hypothesis is indeed false, then we arrive at the correct conclusion

If $p > .05$ and the null hypothesis is indeed true, then we once again fail to make a mistake
Types of error

- However, there are two types of errors we can commit; statisticians have given these the incredibly unimaginative names *type I error* and *type II error*.
  - A type I error consists of rejecting the null hypothesis in a situation where it was true.
  - A type II error consists of failing to reject the null hypothesis in a situation where it was false.
Possible outcomes of comparing $p$ to a cutoff

Thus, there are four possible outcomes of a hypothesis test:

<table>
<thead>
<tr>
<th>Null hypothesis</th>
<th>$p &gt; \alpha$ (accept)</th>
<th>$p &lt; \alpha$ (reject)</th>
</tr>
</thead>
<tbody>
<tr>
<td>True</td>
<td>Correct</td>
<td>Type I error</td>
</tr>
<tr>
<td>False</td>
<td>Type II error</td>
<td>Correct</td>
</tr>
</tbody>
</table>
Consequences of type I and II errors

• Type I and type II errors are different sorts of mistakes and have different consequences.

• A type I error introduces a false conclusion into the scientific community and can lead to a tremendous waste of resources before further research invalidates the original finding.

• Type II errors can be costly as well, but generally go unnoticed.

• A type II error – failing to recognize a scientific breakthrough – represents a missed opportunity for scientific progress.
The proper balance of these two sorts of errors certainly depends on the situation and the type of research being conducted.

That being said, the scientific community generally starts to be convinced at around the $p = .01$ to $p = .10$ level.

The term “statistically significant” is often used to describe $p$-values below .05; the modifiers “borderline significant” ($p < .1$) and “highly significant” ($p < .01$) are also used.

However, don’t let these clearly arbitrary cutoffs distract you from the main idea that $p$-values measure how far off the data are from what the theory predicts – a $p$-value of .04 and a $p$-value of 0.000001 are not at all the same thing, even though both are “significant”.
Certainly, \( p \)-values are widely used, and when used and interpreted correctly, very informative.

However, \( p \)-values are also widely misunderstood and misused – by everyone from students in introductory stat courses to leading scientific researchers.

For this reason, we will now take some time to cover several of the most common \( p \)-value misconceptions.
Reporting \( p \)-values

- One common mistake is taking the 5% cutoff too seriously.
- Indeed, some researchers fail to report their \( p \)-values, and only tell you whether it was “significant” or not.
- This is like reporting the temperature as “cold” or “warm”.
- Much better to tell someone the temperature and let them decide for themselves whether they think it’s cold enough to wear a coat.
Example: HIV Vaccine Trial

- For example, a recent study involving a vaccine that may protect against HIV infection found that, if they analyzed the data one way, they obtained a \( p \)-value of .08.

- If they analyzed the data a different way, they obtained a \( p \)-value of .04.

- Much debate and controversy ensued, partially because the two ways of analyzing the data produce \( p \)-values on either side of .05.

- Much of this debate and controversy is fairly pointless; both \( p \)-values tell you essentially the same thing – that the vaccine holds promise, but that the results are not yet conclusive.
Another big mistake is misinterpreting the $p$-value.

A $p$-value is the probability of getting data that looks a certain way, given that the null hypothesis is true.

Many people misinterpret a $p$-value to mean the probability that the null hypothesis is true, given the data.

These are completely different things.
Conditional probability

- The probability of $A$ given $B$ is not the same as the probability of $B$ given $A$
- For example, in the polio study, the probability that a child got the vaccine, given that he/she contracted polio, was 28%
- The probability that the child contracted polio, given that they got the vaccine, was 0.03%
Absence of evidence is not evidence of absence

- Another mistake (which is, in some sense, a combination of the first two mistakes) is to conclude from a high $p$-value that the null hypothesis is probably true.
- We have said that if our $p$-value is low, then this is evidence that the null hypothesis is incorrect.
- If our $p$-value is high, what can we conclude?
  - Absolutely nothing.
- Failing to disprove the null hypothesis is not the same as proving the null hypothesis.
As a hypothetical example, suppose you and Michael Jordan shoot some free throws
You make 2 and miss 3, while he makes all five
If two people equally good at shooting free throws were to have this competition, the probability of seeing a difference this big is 17% (i.e., $p = .17$)
Does this experiment constitute proof that you and Michael Jordan are equally good at shooting free throws?
Real example

- You may be thinking, “that’s clearly ridiculous; no one would reach such a conclusion in real life”
- Unfortunately, you would be mistaken: this happens all the time
- As an example, the Women’s Health Initiative found that low-fat diets reduce the risk of breast cancer with a $p$-value of .07
- The *New York Times* headline: “Study finds low-fat diets won’t stop cancer”
- The lead editorial claimed that the trial represented “strong evidence that the war against fats was mostly in vain”, and sounded “the death knell for the belief that reducing the percentage of total fat in the diet is important for health”
What should people do when confronted with a high $p$-value?

- Turn to the confidence interval

In this case, 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
- Maybe a low-fat diet won’t affect your risk of breast cancer
- On the other hand, it could reduce it to 83% of what it would otherwise be
A closer look at “significance”

- A final mistake is reading too much into the term “statistically significant”:
  - Saying that results are statistically significant informs the reader that the findings are unlikely to be due to chance alone
  - However, it says nothing about the clinical or scientific significance of the study

- A study can be important without being statistically significant, and can be statistically significant but of no medical/clinical relevance
As an example of statistical vs. clinical significance, consider the story of Nexium, a heartburn medication developed by AstraZeneca.

AstraZeneca originally developed the phenomenally successful drug Prilosec.

However, with the patent on the drug set to expire, the company modified Prilosec slightly and showed that for a condition called erosive esophagitis, the new drug’s healing rate was 90%, compared to Prilosec’s 87%.

Because the sample size was so large (over 5,000), this finding was statistically significant, and AstraZeneca called the new drug Nexium.
The FDA approved Nexium (which, some would argue, was basically the same thing as the now-generic Prilosec, only for 20 times the price)

AstraZeneca went on to spend half a billion dollars in marketing to convince patients and doctors that Nexium was a state of the art improvement over Prilosec

It worked – Nexium became one of the top selling drugs in the world and AstraZeneca made billions of dollars

The ad slogan for Nexium: “Better is better.”
Benefits and drawbacks of hypothesis tests

- The attractive feature of hypothesis tests is that $p$ always has the same interpretation.
- No matter how complicated or mathematically intricate a hypothesis test is, you can understand its result if you understand $p$-values.
- Unfortunately, the popularity of $p$-values has led to overuse and abuse: $p$-values are used in cases where they are meaningless or unnecessary, and $p < .05$ cutoffs used when they make no sense.
- This overuse has also led people to confuse low $p$-values with clinical and practical significance.
- Confidence intervals, which make a statement about both uncertainty and effect size, are very important, less vulnerable to abuse, and should be included alongside $p$-values whenever possible.