Hypothesis tests

Patrick Breheny

January 24

Recap

- 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:
  - Hypothesis testing
  - Confidence intervals
- Both methods address the question of generalization, but do so in different ways and provide different, and complimentary, information

Mere coincidence?

- Today we will cover hypothesis testing, the simpler of the two (in some sense)
- The simplest question we could ask about the polio vaccine study is whether it is possible that what we witnessed was just a coincidence
- In other words, is it possible that the vaccine makes no difference at all, but it just happened to have a lower polio rate than the control group due to random chance?

Hypotheses

- This is a hypothetical situation ("what if the vaccine had no effect") that corresponds to a certain *hypothesis* about the world
- This particular kind of hypothesis, that the observed differences are entirely due to random chance, is called the *null hypothesis*, "null" referring to the notion that nothing is different between the two groups
- The goal of hypothesis testing is to weigh the evidence and quantify the extent to which 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

The scientific method: Proof and disproof

- 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 scientific method: Summing up

- 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 is evidence that directly measures the plausibility of the skeptic's claim

Polio study: what does hypothesis testing tell us?

- 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, $p = .0000000008$, 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
- To wrap up, then, we have now ruled out confounding factors, perception and diagnostic bias, *and chance* as possible explanations for why the vaccinated group had lower rates of polio than the control group; the only other explanation is that the vaccine itself actually does reduce the risk of polio

p-values do not assess the design of the study

- 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 as they were randomized was 0.51)

Fisher's scale of evidence

- As I remarked earlier, the smaller the p-value is, the stronger the evidence against the null
- There is a generally agreed-upon scale for interpreting p-values with regard to the strength of evidence that they represent:

p	Evidence against null
0.1	Borderline
0.05	Moderate
0.025	Substantial
0.01	Strong
0.001	Overwhelming

"Significance"

- The term "statistically significant" is often used to describe p-values below .05, possibly with a modifier:
  - "Borderline significant" $(p < .1)$
  - "Highly significant" $(p < .01)$
- 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"

p-value cutoffs

- Nevertheless, it is worth considering the long-run implications of making decisions based on certain p-value cutoffs
- Suppose we establish a decision-making of cutoff of 0.05; i.e., we will conclude that the null hypothesis is false if $p < 0.05$
- If $p < 0.05$ and the null hypothesis is indeed false, then we arrive at the correct conclusion
- If $p > 0.05$ 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, if we simplify our decision-making down to a simple "reject null" or "don't reject null", there are four possible outcomes of a hypothesis test:

	Null hypothesis	
	True	False
$p > \alpha$ (don't reject)	Correct	Type II error
$p < \alpha$ (reject)	Type I error	Correct

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

Error rates

- Suppose, then, that a large number of hypotheses are tested, with the following results:

	Null hypothesis	
	True	False
$p > \alpha$ (don't reject)	a	b
$p < \alpha$ (reject)	c	d

- Let us define three quantities:
  - *Type I error rate* $= c/(a+c)$: The fraction of null hypotheses that are falsely rejected
  - *Type II error rate* $= b/(b+d)$: The fraction of non-null hypotheses that fail to be rejected
  - *False discovery rate* $= c/(c+d)$: The fraction of null hypothesis rejections that were incorrect

Type I error rate guarantees

- A fundamental property of p-values is that if we use $p < \alpha$ as a cutoff, the Type I error rate is guaranteed (in the long run) to be no more than α
- However, $p < \alpha$ cutoff guarantees us nothing about the Type II error rate, nor about the false discovery rate
- Indeed, a p-value *can't* directly tell us about anything in the right-hand column of the table on the previous slide, since it is calculated based on assuming that the null hypothesis is true

Example

Suppose an investigator sets out to test 200 null hypotheses, of which half are true and half are not. Suppose further that the investigator's hypothesis tests have a Type I error rate of 5% and a Type II error rate of 20%.

(a) Out of the 200 hypothesis tests that the investigator carries out, how many are type I errors?

(b) How many are type II errors?

(c) How many null hypotheses are correctly rejected?

(d) How many times did the investigator correctly fail to reject the null hypothesis?

(e) Out of all the times in which a null hypothesis was rejected, in what percent was the null hypothesis actually true?

p-value misconceptions

- Certainly, p-values are widely used, have a purpose, and can be informative
- However, p-values also have a number of limitations, and people often try to use them to address questions that they just can't answer
- Because p-values are often misused and misunderstood, we will now take some time to cover several 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, highly publicized 2009 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 (i.e., moderately convincing grounds for rejecting the null hypothesis)

Interpretation

- 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

Hypothetical example

- 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"

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
- In particular, just because we can rule out the null hypothesis that two groups are *exactly* the same, this tells us nothing about *how different* the two groups are

Nexium

- 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

Nexium (cont'd)

- The FDA approved Nexium (which, some would argue, was basically the same thing as the now-generic Prilosec, only much more expensive)
- 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 limitations 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, they also have a number of clear limitations:
  - They tell us nothing about how different two groups are
  - They tell us nothing about whether the null hypothesis is true
- Confidence intervals, which we will discuss next time, provide much more information than p-values, and tell us a great deal about both how different the two groups are and what we can and cannot conclude in the absence of significance

Summary

- Hypothesis tests are used to discredit the null hypothesis that nothing is going on besides random chance
- Hypothesis tests are based on p-values, probability of obtaining results as extreme or more extreme than the one observed in the sample, given that the null hypothesis is true
- The lower the p-value, the more convincing the evidence against the null hypothesis
- Know the terms:
  - Type I error (rate)
  - Type II error (rate)
  - False discovery rate
- If we use a threshold of $p < \alpha$ for rejecting hypotheses, then the type I error rate is α