Hypothesis tests

Patrick Breheny

September 8

Specific values of interest

- Confidence intervals express a range of likely values for the parameter θ an investigator is studying; however, we are often particularly interested in one specific 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* and denoted θ_0 ("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

- 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
- Mathematically, we must choose some way of quantifying "extreme-ness"; this is known as a *test statistic*, which is just a function $T : \text{Data} \to \mathbb{R}$
- Let X denote random data that could be obtained from a repeatable experiment and x denote the actual data we obtained the single realization of this experiment. The *p-value* of a test statistic T is defined as

$$p(x) = P\{T(X) \geq T(x) | \theta = \theta_0\}$$

Interpretation

- 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 "larger" (with respect to T) than 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 be as large as the observed sample; the null hypothesis looks doubtful

Fisher's scale of evidence

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 "evidence against the null" from the previous slide is often referred to as the "statistical significance" of a study
- For example, if a study results in a p-value of 0.08, it might be described in words as "borderline significant"; a p-value of 0.04 might be described as "moderately significant"; a p-value of 0.0006 would be described as "overwhelmingly significant", and so on

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

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)

p-value cutoffs

- In the last lecture, we remarked that if science goes about constructing 95% confidence intervals, then 95% of those intervals will contain the truth
- It is worth considering the same long-run frequency implications for hypothesis testing
- Suppose we establish a decision-making of cutoff of α; i.e., we will conclude that the null hypothesis is false if $p < \alpha$
- If $p < \alpha$ and the null hypothesis is indeed false, then we arrive at the correct conclusion
- If $p > \alpha$ 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 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:

	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

- **Theorem:** For any value θ_0 and for any $\alpha \in [0, 1]$,

$$P\{p(X) \leq \alpha \,|\theta = \theta_0\} \leq \alpha$$

- In other words, using $p < \alpha$ as a cutoff, the Type I error rate is guaranteed (in the long run) to be no more than α
- Note that using α as a cutoff guarantees us nothing about the Type II error rate or 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?

Confidence intervals tell us about p-values

- There is a close connection between the long-run frequency properties of confidence intervals and hypothesis tests
- For example, consider testing a null hypothesis not by calculating a p-value, but by constructing a 95% confidence interval and seeing whether the interval contains the null hypothesis
- The 95% guarantee of the confidence interval means that we have a 5% guarantee on the Type I error rate of this test
- So, for example, if the 95% confidence interval contains the null hypothesis, we don't know the exact p-value, but we would know that $p > 0.05$

p-values tell us about confidence intervals

- The same logic works the other way around, also
- Consider constructing a confidence interval by testing all possible values of θ; if θ is rejected, we place it outside the confidence interval, and if it isn't rejected, we place it inside the interval
- The Type I error rate guarantee of the hypothesis test means that we have a coverage guarantee for the resulting confidence interval
- So, for example, if $p < 0.05$, we don't know the exact limits of the 95% confidence interval, but we know that it doesn't contain the null hypothesis

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 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

p-value misconceptions

- p-values are widely used and have a purpose in answering one very specific question
- However, p-values are also widely misunderstood and misused by many people
- For this reason, I'd like to take a little time to cover some 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
- However, a p-value of 0.04 and p-value of 0.00001 obviously represent very different levels of evidence against the null, even though both are "significant"

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 (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

- $P(A|B)$ is not the same as $P(B|A)$
- 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"

Women's Health Initiative: Confidence interval

- 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 a woman's risk of breast cancer by 17%

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

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 offered a small advantage over the now-generic Prilosec, but was 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 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
  - $p < 0.05$ cutoffs used even when they make little sense
  - p values encourage confusion between clinical and practical significance
- Confidence intervals avoid these problems, although take more work to interpret

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
- There is a close connection between hypothesis tests and confidence intervals; however, confidence intervals are far less prone to misinterpretation
- Know the terms:
  - Type I error (rate)
  - Type II error (rate)
  - False discovery rate