Hypothesis Testing

Ryan Miller

Statistical Inference

So far, we've introduced two different areas where statisticians apply probability:

1) **Estimation** - using sample data to learn something about a broader population
2) **Hypothesis Testing** - using sample data to evaluate the plausibility of a particular null model for a population

This presentation will provide a detailed look at *hypothesis testing*

Polio Epidemic - Introduction

- In the early 1950s the US experienced an outbreak of polio that reached 58,000 new cases in 1952
- Several vaccines had been developed, with one created by Jonas Salk seeming particularly promising. How might the effectiveness of Salk's vaccine be established?

Polio Epidemic - Introduction

- In the early 1950s the US experienced an outbreak of polio that reached 58,000 new cases in 1952
- Several vaccines had been developed, with one created by Jonas Salk seeming particularly promising. How might the effectiveness of Salk's vaccine be established?
- In 1954, the US Public Health Service organized a large study involving nearly 1 million children in grades 1, 2, and 3, the most vulnerable age groups for polio
  - Do you have any concerns with performing a randomized experiment in this setting?

Polio Epidemic - Ethics

- Parents must provide consent for their children to receive the vaccination
  - But is it ethical to deliberately leave some of these consenting children unvaccinated?
- A more ethical design would offer the vaccine to all consenting children and use those whose parents refused the vaccine as the control group
  - Do you have any statistical concerns with the ethical design?

Polio Epidemic - Confounding

- Higher-income parents tended to be more likely to consent, and their children tended to be more likely to contract polio
  - This is thought to be because children from poorer backgrounds are more likely to come into contact with mild cases of polio during early childhood when they are protected by antibodies from their mothers
- Thus, family background would be a major source of confounding in the ethical design
  - Any observed differences could be attributable to this confounding variable and not the efficacy of the vaccine

Polio Epidemic - Randomization and Blinding

- To avoid confounding variables, the treatment and control groups needed to be *randomly assigned* from the same population: *children whose parents consented to treatment*
- This meant that some children whose parents consented would be randomly chosen to not receive the vaccine

Polio Epidemic - Randomization and Blinding

- To avoid confounding variables, the treatment and control groups needed to be *randomly assigned* from the same population: *children whose parents consented to treatment*
- This meant that some children whose parents consented would be randomly chosen to not receive the vaccine
- Additionally, the Salk vaccine trial included a placebo and was double-blinded
  - Children in the control group received an injection of a saline solution
  - Neither the child, their parents, nor their doctors knew who had received vaccine and who had received placebo

Polio Epidemic - Salk Vaccine Trial Results

The incidence of polio was lower in the treatment group. But to attribute this decrease to the vaccine all other explanations must be ruled out...

Group	n	Polio Cases	Rate per 100,000
Treatment	200000	56	28
Control	200000	142	71
Refused Consent	350000	161	46

- Confounding?

Polio Epidemic - Salk Vaccine Trial Results

The incidence of polio was lower in the treatment group. But to attribute this decrease to the vaccine all other explanations must be ruled out. . .

Group	n	Polio Cases	Rate per 100,000
Treatment	200000	56	28
Control	200000	142	71
Refused Consent	350000	161	46

- Confounding? No, random assignment balanced the vaccinated and unvaccinated groups
- Sampling bias?

Polio Epidemic - Salk Vaccine Trial Results

The incidence of polio was lower in the treatment group. But to attribute this decrease to the vaccine all other explanations must be ruled out...

Group	n	Polio Cases	Rate per 100,000
Treatment	200000	56	28
Control	200000	142	71
Refused Consent	350000	161	46

- Confounding? No, random assignment balanced the vaccinated and unvaccinated groups
- Sampling bias? No, both groups were randomly chosen from the same population
- Other biases?

Polio Epidemic - Salk Vaccine Trial Results

The incidence of polio was lower in the treatment group. But to attribute this decrease to the vaccine all other explanations must be ruled out...

Group	n	Polio Cases	Rate per 100,000
Treatment	200000	56	28
Control	200000	142	71
Refused Consent	350000	161	46

- Confounding? No, random assignment balanced the vaccinated and unvaccinated groups
- Sampling bias? No, both groups were randomly chosen from the same population
- Other biases? No, a placebo was used and the doctors/participants were blinded
- Random chance? ...

The Role of Random Chance

In a well-designed study, researchers are able to reduce the set of plausible explanations to just two:

1) A real relationship between the explanatory and response variables (ie: The vaccine effectively reduces the rate of polio)
2) Random chance (ie: The vaccine makes no difference and any observed differences can be explained by randomness. After all, it's extremely unlikely for two groups to have polio rates that are exactly identical.)

- **Hypothesis testing** is used to rule out random chance as a plausible explanation
- In the context of this study, hypothesis testing answers the question "how likely would it be for the vaccinated group to have a polio rate that is 43 cases per 100k lower than the unvaccinated group *if the vaccine had made no difference*?"

Hypothesis Testing

- This hypothetical scenario, *"what if the vaccine made no difference"*, is a **null hypothesis** (also called a **null model**)
  - Statistically speaking, it implies the *population parameters* (the polio rates for vaccinated and unvaccinated children) *are identical*, and the differences we observed in the *sample data* are due to *random chance*

Hypothesis Testing

- This hypothetical scenario, *"what if the vaccine made no difference"*, is a **null hypothesis** (also called a **null model**)
  - Statistically speaking, it implies the *population parameters* (the polio rates for vaccinated and unvaccinated children) *are identical*, and the differences we observed in the *sample data* are due to *random chance*
- In statistical notation:

$$\text{Null Hypothesis } (H_0) : p_{\text{trt}} = p_{\text{ctrl}} \text{ or } p_{\text{trt}} - p_{\text{ctrl}} = 0 \text{ or } \frac{p_{\text{trt}}}{p_{\text{ctrl}}} = 1$$

- **Hypothesis testing** evaluates how compatible the sample data are with a null model
  - If it's extremely unlikely for the sample data to arise from the null model, we conclude the null model is implausible (thus ruling out random chance as a plausible explanation for what was observed in the sample)

p-values

- Probability allows us to quantify how compatible/incompatible the sample data are with a null model
  - The **p-value** is defined as *the probability of seeing an outcome at least as extreme as what was observed in our sample if the null model were true*

p-values

- Probability allows us to quantify how compatible/incompatible the sample data are with a null model
  - The **p-value** is defined as *the probability of seeing an outcome at least as extreme as what was observed in our sample if the null model were true*
- The smaller the p-value, the more incompatible the sample data are with the null model, and thus the stronger the evidence is against random chance as a viable explanation
  - For example, a p-value of 0.01 indicates a 1/100 chance of seeing results as extreme as the sample data if the null model were true

Null Distributions

- In order to calculate a p-value, we need to know what could have happened if the null model were true
  - So, hypothesis testing is really just estimation with an added constraint (the data arose from the null model)

Null Distributions

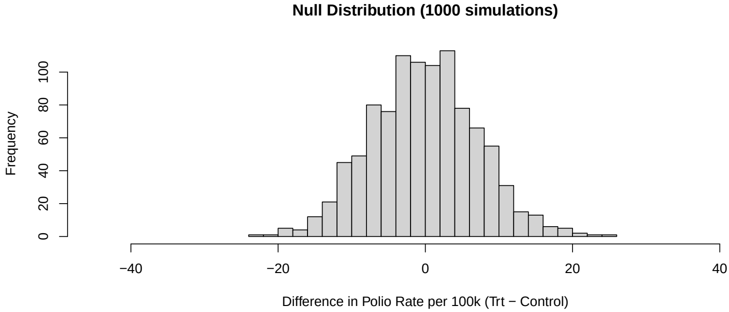
- In order to calculate a *p*-value, we need to know what could have happened if the null model were true
  - So, hypothesis testing is really just estimation with an added constraint (the data arose from the null model)
- We've previously approached estimation by finding the distribution of possible estimates that could have been observed if a study were repeated (ie: repeatedly taking different samples)
  - We called the distribution of these possible estimates the *sampling distribution* (though I prefer "the distribution of sample averages")

Null Distributions

- In order to calculate a p-value, we need to know what could have happened if the null model were true
  - So, hypothesis testing is really just estimation with an added constraint (the data arose from the null model)
- We've previously approached estimation by finding the distribution of possible estimates that could have been observed if a study were repeated (ie: repeatedly taking different samples)
  - We called the distribution of these possible estimates the *sampling distribution* (though I prefer "the distribution of sample averages")
- Hypothesis testing focuses on finding the *null sampling distribution*, or "null distribution" for short, which is the distribution of possible sample estimates that could occur if a particular null model were true

Null Distributions

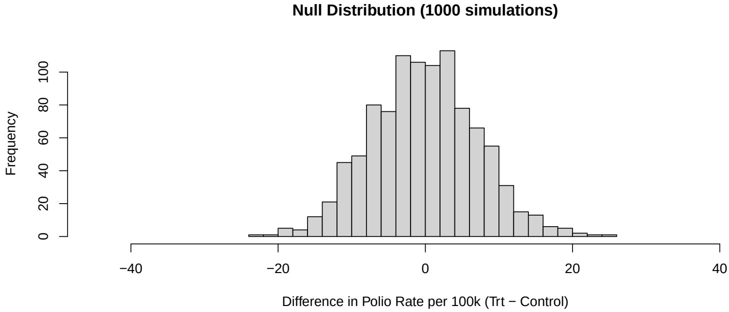
In our polio example, here is a simulated null distribution of the expected difference in polio rates had the vaccine made no difference



The actual experiment showed a difference of 43. what is the *p*-value?

Null Distributions

In our polio example, here is a simulated null distribution of the expected difference in polio rates had the vaccine made no difference



The actual experiment showed a difference of 43. what is the p-value? Very small, less than 1/1000!

The *p*-value as Evidence Against the Null

Ronald Fisher, creator of the *p*-value, and described by his peers as "a genius who almost single-handedly created the foundations of modern statistical science", suggests the following guidelines:

p-value	Evidence against the null
0.100	Borderline
0.050	Moderate
0.025	Substantial
0.010	Strong
0.001	Overwhelming

The p-value as Evidence Against the Null

Ronald Fisher, creator of the p-value, and described by his peers as "a genius who almost single-handedly created the foundations of modern statistical science", suggests the following guidelines:

p-value	Evidence against the null
0.100	Borderline
0.050	Moderate
0.025	Substantial
0.010	Strong
0.001	Overwhelming

- Many scientific fields use $\alpha = 0.05$ as a "significance threshold" for *rejecting* a null hypothesis
- Given this threshold, p-values < 0.05 are described as "statistically significant"

Statistical Significance

- $p < 0.05$ is an arbitrary cutoff that shouldn't distract you from the main idea behind p-values
- That is, a p-value of 0.0001 doesn't tell you the same thing as a p-value of 0.04, even though both are "statistically significant"

Statistical Significance

- $p < 0.05$ is an arbitrary cutoff that shouldn't distract you from the main idea behind p-values
- That is, a p-value of 0.0001 doesn't tell you the same thing as a p-value of 0.04, even though both are "statistically significant"
- When reporting results you should always include the p-value itself, not just whether it met some arbitrary significance threshold
  - Imagine your weather app only telling you: "it's cold" or "it's not cold"
  - This is bad because "cold" is subjective, it's better to provide the temperature and let you decide for yourself

Alternatives to the Null Model

Null hypotheses are intended serve as a "straw man" for a *complementary* **alternative hypothesis** that we want to establish:

$$\text{Null Hypothesis } (H_0) : p_{\text{trt}} = p_{\text{ctrl}}$$

$$\text{Alternative Hypothesis } (H_a) : p_{\text{trt}} < p_{\text{ctrl}}$$

The idea that the *p*-value will provide enough reason to doubt the null hypothesis that the alternative hypothesis is the only sensible thing to believe

One-sided vs. Two-sided Tests

- You may have noticed the null and alternative hypotheses on the prior slide aren't technically complementary (they don't account for $p_{\text{trt}} > p_{\text{ctrl}}$)
- This is called a *one-sided* hypothesis test, an approach that *is not* widely used in published research
  - Technically, the proper hypotheses in this test should be:

$$\text{Null Hypothesis } (H_0) : p_{\text{trt}} \geq p_{\text{ctrl}}$$

$$\text{Alternative Hypothesis } (H_a) : p_{\text{trt}} < p_{\text{ctrl}}$$

- This null hypothesis might seem a bit confusing, as it implies there are actually many different null models to consider - However, it suffices to consider only the "strongest" null model where $p_{\text{trt}} = p_{\text{ctrl}}$

Undesirability of One-sided Tests

- In order to stay true to the scientific method, the null and alternative hypotheses should be set up before researchers ever see the data
  - Using the hypotheses on the prior slide, what could you conclude if the treatment turns out to be very harmful in the sample?

Undesirability of One-sided Tests

- In order to stay true to the scientific method, the null and alternative hypotheses should be set up before researchers ever see the data
  - Using the hypotheses on the prior slide, what could you conclude if the treatment turns out to be very harmful in the sample?
- The answer is nothing! The one-sided p-value will be extremely large if the *direction* of the one-sided alternative is incorrect
  - This is undesirable because we'd want to be able to conclude that the treatment is harmful

Undesirability of One-sided Tests

- In order to stay true to the scientific method, the null and alternative hypotheses should be set up before researchers ever see the data
  - Using the hypotheses on the prior slide, what could you conclude if the treatment turns out to be very harmful in the sample?
- The answer is nothing! The one-sided p-value will be extremely large if the *direction* of the one-sided alternative is incorrect
  - This is undesirable because we'd want to be able to conclude that the treatment is harmful
- One-sided tests are also ripe for fraud, as there's no way of knowing that a researcher didn't change their hypotheses after seeing the data

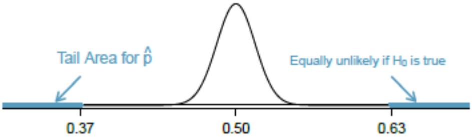
Two-sided p-values

- Statisticians avoid the ambiguity of one-sided hypotheses by favoring two-sided alternatives:

$$\text{Null Hypothesis } (H_0) : p_{\text{trt}} = p_{\text{ctrl}}$$

$$\text{Alternative Hypothesis } (H_a) : p_{\text{trt}} \neq p_{\text{ctrl}}$$

- In most scenarios, this doubles the one-sided p-value and is an effective way to prevent "cheating"



Misinterpretations of the p-value

The logic underlying the p-value is as follows:

1) Determine a null model you'd like to disprove
2) Measure how compatible the sample data is with the null model (that is, what is the probability such data were observed if the null model were true)
3) If there is enough incompatibility (a small p-value), *reject* the null model in favor of an alternative

Misinterpretations of the *p*-value

The logic underlying the *p*-value is as follows:

1) Determine a null model you'd like to disprove
2) Measure how compatible the sample data is with the null model (that is, what is the probability such data were observed if the null model were true)
3) If there is enough incompatibility (a small *p*-value), *reject* the null model in favor of an alternative

But does a large *p*-value mean we should *accept* the null model that is being evaluated?

Hypothetical Example

- Suppose Steph Curry and Professor Miller each shoot 5 three-point shots
  - I make 2 of 5 and Steph makes 5 of 5
- We can use a hypothesis test to evaluate the null model that we're both equally good three-point shooters ($H_0 : p_{\text{Miller}} = p_{\text{Curry}}$
  - The p-value for this scenario is 0.17
  - Are we equally good 3-pt shooters?

A Non-hypothetical Example

- It might seem like no one would make the mistake illustrated in that silly Steph Curry example, but unfortunately it happens quite often

A Non-hypothetical Example

- It might seem like no one would make the mistake illustrated in that silly Steph Curry example, but unfortunately it happens quite often
- In 2006, the Woman's Health Initiative evaluated the relationship between low-fat diets and reduced risk of breast cancer risk and found a p-value of 0.07
- The NY Times ran the headline: "Study Finds Lowfat Diets Won't Stop Cancer or Heart Disease"
- The article described the study's results as: "The death knell for the belief that reducing the percentage of fat in the diet is important for health"

“Proving” the Null Hypothesis

- Hypothesis testing is not designed to “prove” a null hypothesis, so you cannot use it to do so
- The closest thing to “proving” a null hypothesis is finding a very narrow confidence interval around the null value
  - Such an interval would suggest the only plausible values for the parameter are extremely close to what the null hypothesis suggests

Hypthosis Testing vs. Confidence Intervals

- Hypothesis tests and confidence intervals can both be used to evaluate whether random chance is a likely explanation for a phenomenon observed in sample data
    - Because both methods are based upon *sampling variability*, you can use one of these methods to infer things about the other (the null distribution is a sampling distribution with an added constraint)

Hypthosis Testing vs. Confidence Intervals

- Hypothesis tests and confidence intervals can both be used to evaluate whether random chance is a likely explanation for a phenomenon observed in sample data
  - Because both methods are based upon *sampling variability*, you can use one of these methods to infer things about the other (the null distribution is a sampling distribution with an added constraint)
- Consider a null model that two group means are equal, or $H_0 : \mu_1 - \mu_2 = 0$
  - If a sample produces a 95% confidence interval estimate of (3.2, 10.1), do you think this null model is plausible? What do you think the *p*-value might be?

Hypthosis Testing vs. Confidence Intervals

- Hypothesis tests and confidence intervals can both be used to evaluate whether random chance is a likely explanation for a phenomenon observed in sample data
  - Because both methods are based upon *sampling variability*, you can use one of these methods to infer things about the other (the null distribution is a sampling distribution with an added constraint)
- Consider a null model that two group means are equal, or $H_0 : \mu_1 - \mu_2 = 0$
  - If a sample produces a 95% confidence interval estimate of (3.2, 10.1), do you think this null model is plausible? What do you think the *p*-value might be?
  - This null model is *not plausible* because 0 isn't in the 95% CI, thus the two-sided *p*-value must be < 0.05

Hypthosis Testing vs. Confidence Intervals

- Suppose the p-value for the hypothesis $H_0 : \mu_1 - \mu_2 = 0$ were 0.13, what does this result tell you about the 95% confidence interval estimate for $\mu_1 - \mu_2$? What about the 80% confidence interval estimate?

Hypthosis Testing vs. Confidence Intervals

- Suppose the *p*-value for the hypothesis $H_0 : \mu_1 - \mu_2 = 0$ were 0.13, what does this result tell you about the 95% confidence interval estimate for $\mu_1 - \mu_2$? What about the 80% confidence interval estimate?
  - The 95% CI wouldn't contain 0, but the 80% CI would!
- Now, what if the *p*-value for this hypothesis were 0.001, what can you infer about the plausible differences between μ_1 and μ_2?

Hypthosis Testing vs. Confidence Intervals

- Suppose the *p*-value for the hypothesis $H_0 : \mu_1 - \mu_2 = 0$ were 0.13, what does this result tell you about the 95% confidence interval estimate for $\mu_1 - \mu_2$? What about the 80% confidence interval estimate?
  - The 95% CI wouldn't contain 0, but the 80% CI would!
- Now, what if the *p*-value for this hypothesis were 0.001, what can you infer about the plausible differences between μ_1 and μ_2?
  - A difference of zero is not plausible, but the test cannot tell us what differences are plausible...

Prilosec vs. Nexium

- Confidence intervals and hypothesis tests lead to similar conclusions, but that doesn't mean they provide the same information
- In the 1980s, *AstraZeneca* developed *Prilosec*, a very successful medication for healing erosive esophagitis
  - In the 2001, just before the company's patent on *Prilosec* was about to expire, *AstraZeneca* developed a new drug, *Nexium*

Prilosec vs. Nexium

- Confidence intervals and hypothesis tests lead to similar conclusions, but that doesn't mean they provide the same information
- In the 1980s, *AstraZeneca* developed *Prilosec*, a very successful medication for healing erosive esophagitis
  - In the 2001, just before the company's patent on *Prilosec* was about to expire, *AstraZeneca* developed a new drug, *Nexium*
- To get *Nexium* approved by the FDA, *AstraZeneca* conducted a large randomized experiment comparing it to *Prilosec*
  - The experiment resulted in a p-value < 0.001, well below significance threshold of $\alpha = 0.05$ used by the FDA
- After its approval, *AstraZeneva* spent millions of dollars marketing *Nexium* and it soon became one of the top selling drugs in the world, leading to billions in profits

Clincial Significance vs. Statistical Significance

- Despite a p-value < 0.001, the observed healing rates were 87% for Prilosec and 90% for Nexium
  - The factor by which Nexium improved healing had a 95% CI of (1.02, 1.06)

Clincial Significance vs. Statistical Significance

- Despite a p-value < 0.001, the observed healing rates were 87% for Prilosec and 90% for Nexium
  - The factor by which Nexium improved healing had a 95% CI of (1.02, 1.06)
- Further, the active ingredients of these drugs are:
  - Omeprazole (Prilosec)
  - Esomeprazole (Nexium)
- Without getting too far into the chemistry (not my area of expertise), Omeprazole is a 50-50 mix of active and inactive isomers, while Esomeprazole only contains active "S" isomers

Clincial Significance vs. Statistical Significance

- Despite a p-value < 0.001, the observed healing rates were 87% for Prilosec and 90% for Nexium
  - The factor by which Nexium improved healing had a 95% CI of (1.02, 1.06)
- Further, the active ingredients of these drugs are:
  - Omeprazole (Prilosec)
  - Esomeprazole (Nexium)
- Without getting too far into the chemistry (not my area of expertise), Omeprazole is a 50-50 mix of active and inactive isomers, while Esomeprazole only contains active "S" isomers
- Critics argue the results of the Nexium study are not **clinically significant**, meaning the differences in the two drugs aren't substantial enough to be influencing clinical practices

Reporting the Results of a Hypothesis Test

Below are several example statements ranging from "Really Really Bad", "Really Bad", "Bad", "Okay", "Good", and "Really Good". Take a moment to try and classify each statement:

1. $p < 0.05$ so we reject the null hypothesis
2. $p = 0.01$, indicating strong evidence that Nexium is more effective than Prilosec at treating heartburn
3. The study failed to reject the hypothesis that diet isn't associated with breast cancer risk
4. The study provided borderline evidence ($p = 0.07$) that low-fat diets reduce breast cancer risk, it is possible that diet has no effect, but it is also possible that low-fat diets have a small protective effect
5. The study rejected the hypothesis that Nexium and Prilosec are equally good
6. $p > 0.05$, so the null hypothesis is likely true

Putting it all together

1. $p < 0.05$ so we reject the null hypothesis **Really Bad**
2. $p = 0.01$, indicating strong evidence that Nexium is more effective than Prilosec at treating heartburn **Good**
3. The study failed to reject the hypothesis that diet isn't associated with breast cancer risk **Okay**
4. The study provided borderline evidence ($p = 0.07$) that low-fat diets reduce breast cancer risk, it is possible that diet has no effect but it is also possible that low-fat diets have a small protective effect **Really Good**
5. The study rejected the hypothesis that Nexium and Prilosec are equally good **Bad**
6. $p > 0.05$, so the null hypothesis is probably true **Really Really Bad**

Conclusion

- In the coming weeks we'll learn how to use conduct hypothesis tests for a variety of different summary measures, using a variety of different statistical methods:
  - Normal models based upon the Central Limit Theorem
  - Exact tests using the binomial distribution
  - Simulation-based approaches
- Before diving into these methods, we have a few more conceptual loose ends related to hypothesis testing that we need to address