Statistical Testing

Ryan Miller

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 developed by Jonas Salk appearing to be particularly promising. How might the effectiveness of Salk's vaccine be confirmed?
- 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
  - The parents of recruited children needed to consent to receive the vaccination
  - What are some concerns involved with a performing a randomized experiment in this setting?

Polio Epidemic - Ethics

- It is controversial whether it is ethical to deliberately leave some of the consenting children unvaccinated
- A more ethical design might be to offer the vaccine to all consenting children and use those whose parents refused the vaccine as the control group
- What are some problems with the aforementioned 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 due to this factor and not the efficacy of the vaccine

Polio Epidemic - Randomization and Blinding

- To avoid confounding, the treatment and control groups needed to be randomly selected 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 were given an injection of a saline solution
  - Neither the child, their parents, nor their doctors knew who received vaccine and who received placebo

Polio Epidemic - Salk Vaccine Trial Results

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

The incidence of polio was lower in the treatment group, what explanations are plausible?

- Confounding? No, proper randomization was used
- Diagnostic bias? No, the doctors/participants were blinded
- Sampling bias? No, the same population is represented by each group
- Random chance? ...

Statistical Tests

- In the Salk Vaccine Trial, the incidence of polio was reduced by a factor of roughly 2.5 (71/28)
- But this is just what we saw in the sample, we really want to generalize these findings to a broader population
- It is unlikely that the broader population will see a reduction of *exactly* 2.5, so how can we determine whether the results observed in our sample are convincing evidence that the population will benefit from the vaccine?

We could try using confidence intervals, but instead we might ask:

> *Had the vaccine made no difference, how likely would be for the vaccinated group to have a 2.5 times lower incidence rate?*

Statistical Tests

- This hypothetical situation: *"What if the vaccine made no difference"* represents a **null hypothesis**
- It implies that both population parameters (the polio incidence rates for vaccinated and unvaccinated children) *are the same* and any differences we observed in the sample are *due to random chance*. Using statistical notation:

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

- The goal of statistical testing is to quantify how plausible the null hypothesis is given the data we observed in our sample

p-values

Statistical tests are based upon determining: *The probability of obtaining results as extreme or more extreme than those observed in our sample, provided the null hypothesis is true*

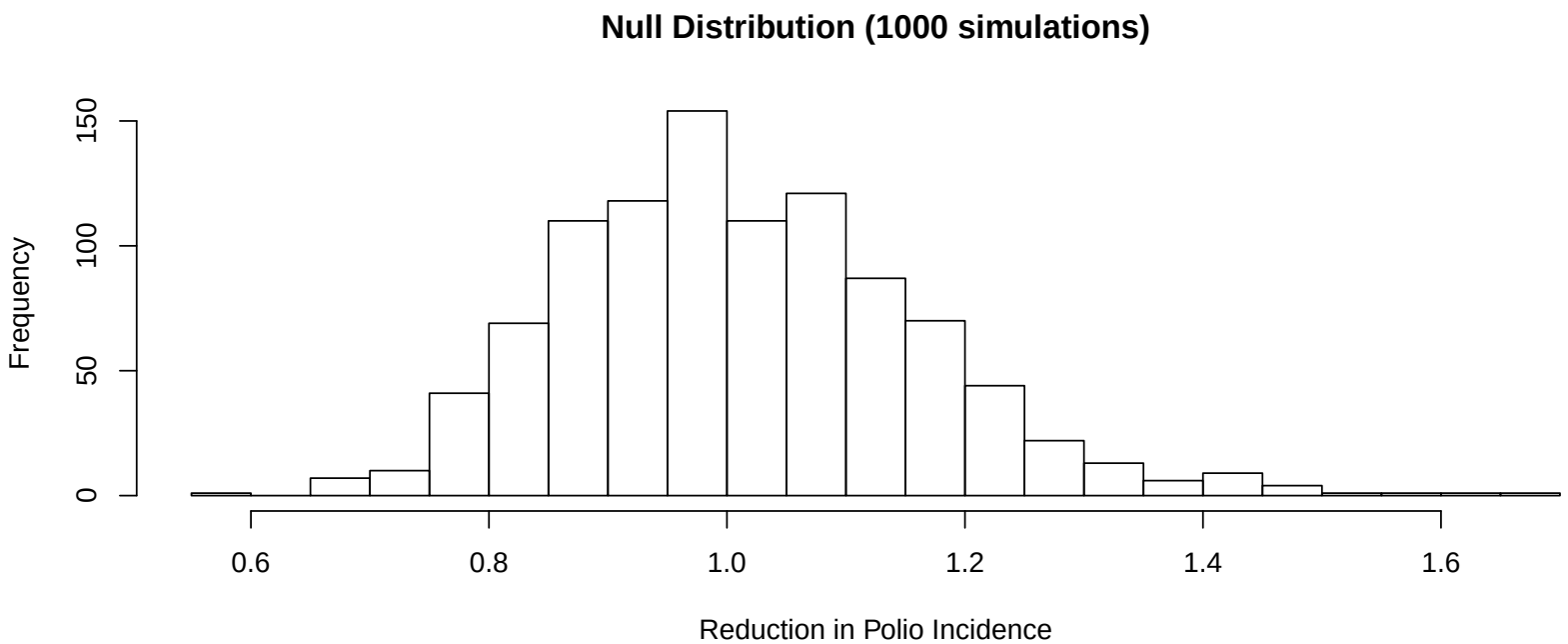
- This probability is called the *p*-value
- The smaller the *p*-value, the stronger the evidence is against the null hypothesis
  - A *p*-value of 0.01 means that if the null hypothesis were true, only 1/100 samples would be expected to produce an outcome as or more extreme as the one we observed in our sample

The Null Distribution

- The way we calculate *p*-values is similar to the logic underlying confidence intervals
- *Interval estimation* was based around finding plausible values of a statistic that could occur when repeatedly sampling
- Statistical testing seeks to find plausible values of a statistic that could occur when repeatedly sampling *if the null hypothesis were true*
- Thus, the *sampling distribution* in the hypothetical world where the null hypothesis is true is called the **null distribution**
- The null distribution is centered at the value specified in the null hypothesis, and it reflects the possible sample estimates we could expect to see if the null hypothesis were true

The Null Distribution

In the polio example, here is the null distribution for the factor by which polio was reduced.



The actual experiment showed a reduction of ~2.5, what do you think the p-value is?

Null and Alternative Hypotheses

Generally, we pair the null hypothesis with an **alternative hypothesis** that we'd like to establish:

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

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

- The alternative hypothesis offers a sensible conclusion if our data suggests the null hypothesis is unlikely
- We'll first look at *one-sided* hypothesis tests because they are easy to understand. But in reality most tests are *two-sided*

Hypothesis Testing - Example

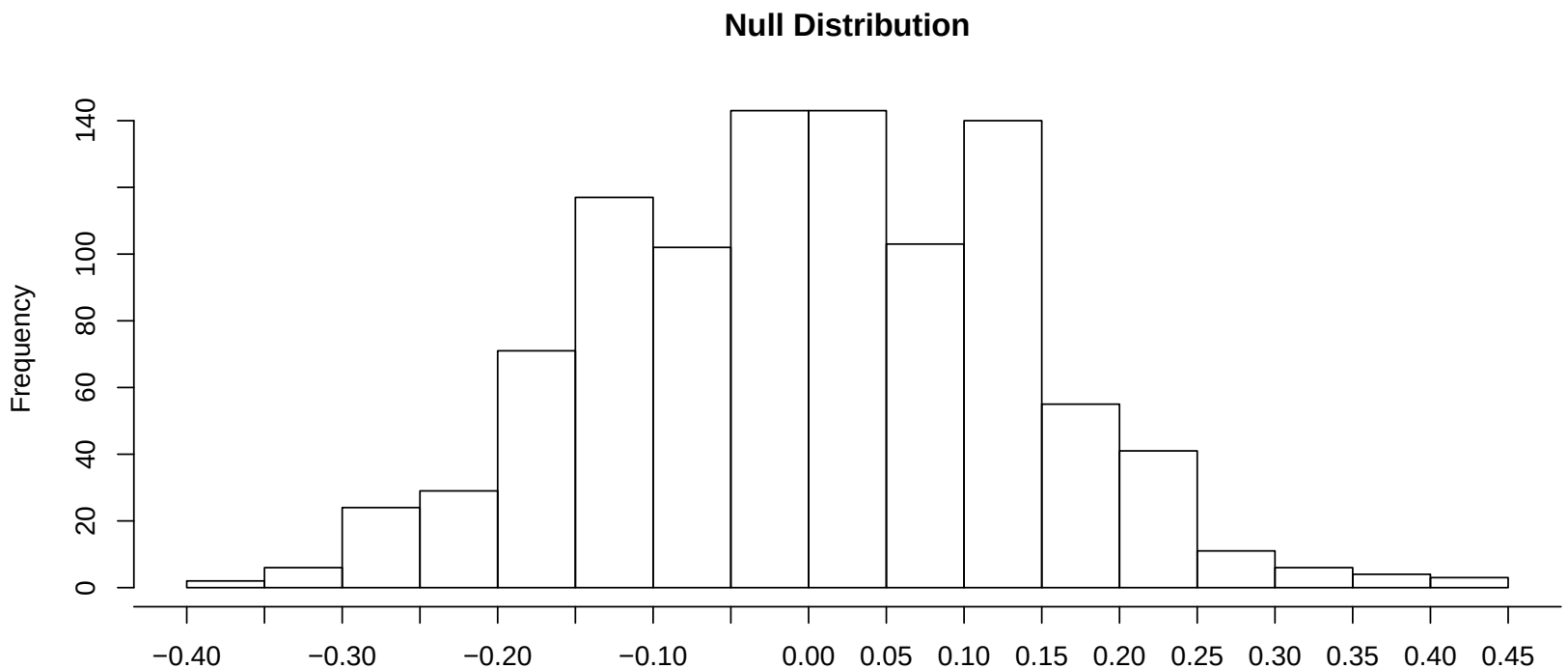
The TV show Mythbusters uses experiments to evaluate popular beliefs that might not be true. One myth the show investigates is whether yawning is contagious.

- 50 people were recruited with the premise that they were looking for people to appear on the show
- The recruiter met with each person in a small room and either intentionally yawned or did not yawn during the interview
- After the recruiter left, each subject was alone in the room for a period of time while being recorded on video
- Whether or not each subject yawned at any point during or after the interview was recorded
  - When the recruiter didn't yawn, 4 of 16 subjects also yawned
  - When the recruiter yawned, 10 of 34 subjects also yawned

Hypothesis Testing - Example

With your group:

1. Using the information given regarding this experiment, come up with suitable null and alternative hypotheses
2. Report your estimate of the statistic your test will use
3. Estimate the p-value using the null distribution below



Hypothesis Testing - Example (Solution)

1. $H_0 : p_{y|\text{yawn}} = p_{y|\text{no yawn}}$ and $H_A : p_{y|\text{yawn}} > p_{y|\text{no yawn}}$
2. The observed difference in proportions is $\hat{p}_{y|\text{yawn}} - \hat{p}_{y|\text{no yawn}} = 0.044$
3. A really large portion of the histogram is more extreme in favor of the alternative than our estimate of 0.044, so the p-value is likely around 0.4

We conclude that this experiment does not provide any conclusive evidence that yawning is contagious

Statistical Significance

Ronald Fisher, the developer of the *p*-value who has been described 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

- Generally, modern science uses 0.05 as a threshold for *rejecting* the 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
- 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 include the p-value itself, not just whether or not it was below the 0.05 threshold for significance
  - Think about someone asking about the weather and you answering "it's cold" or "it's not cold"
  - It is better to provide the exact temperature and let them decide

p-value Misconceptions

- *p*-values have been much maligned over the last several years, so much so that the largest professional organization of statisticians, the American Statistical Association (ASA), recently issued a statement on *p*-values
- The statement addresses several different *p*-value misconceptions, the proliferation of these mistakes has led some to abandon *p*-values entirely (They've been banned from the journal: *Basic and Applied Psychology*)

p-value Misconceptions

- A common mistake is to conclude that a high *p*-value means the null hypothesis *is likely to be true*
- In reality, a high *p*-value tells you very little about how likely the null hypothesis is to be true!
- We'll illustrate this with a hypothetical example:
  - Suppose Steph Curry and I each shoot 5 three-point shots
  - I make 2/5 and he makes 5/5
  - Under the null hypothesis that we are equally good at three-point shooting, the probability (*p*-value) of a result this extreme is 0.17
  - Do these results justify the conclusion that Steph Curry and I are equally good shooters?

p-value Misconceptions

While that hypothetical example illustrates the problem, but maybe you're thinking that no makes conclusions like that in real life...

Unfortunately, it happens all the time:

- In 2006, the Woman's Health Initiative found that low-fat diets are associated with reduced breast cancer risk with 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"

p-value Misconceptions

- Another common mistake is mistaking a *statistically significant result* for a *clinically significant* result
  - Statistical significance simply suggests that the observed differences are unlikely to be due to random chance
  - It doesn't mean that the observed differences are of any practical importance

Statistical vs. Clinical Significance

- In the 1980s pharmaceutical company AstraZeneca developed an incredibly successful heartburn medication *Prilosec*
- The FDA patent for Prilosec ran out in 2001, prompting AstraZeneca to try to replace Prilosec with a new drug *Nexium*
- The active ingredients of these drugs are:
  - Omeprazole (Prilosec)
  - Esomeprazole (Nexium)
- Without getting in to the chemistry, Omeprazole is a 50-50 mix of active and inactive isomers, while Esomeprazole only contains active "S" isomers
- Thus, taking the same amount of Nexium provides twice the effective dose of the active isomer

Nexium vs. Prilosec

- With this "modification", AstraZeneca showed that Nexium had a healing rate of 90% for erosive esophagitis, while Prilosec only had a 87% success rate
- Because the sample size of the trial was large (nearly 6,000), the difference was statistically significant with a p-value well below 0.05
- This led the FDA to approve Nexium, while AstraZeneca spent hundreds of millions of dollars marketing the drug to patients and doctors as a state-of-the-art improvement over Prilosec under the slogan: "better is better"
- The marketing campaign worked, AstraZeneca has since made *over 47 billion dollars* from Nexium

Nexium vs. Prilosec

- Practically speaking, the success rate of the two drugs was roughly the same, it was the large sample size that led to a statistically significant difference
- The 95% confidence interval for the factor by Nexium improved the healing rate was (1.02, 1.06)
- Furthermore, the small observed difference is almost surely due to Nexium containing more of the active isomer, not a groundbreaking development
- This is an example of when statistical hypothesis testing can go wrong
  - Statistical testing doesn't measure practical importance
  - Statistical testing needs to be informed by other sources of scientific knowledge

Putting it all together

An important part of this class is translating the results of statistical test to a meaningful conclusion. Below are several examples ranging from "Really Really Bad", "Really Bad", "Bad", "Okay", "Good", and "Really Good". With your group try to 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**

The Next Steps

- So far we've seen how to determine the p-value when given the null distribution
- In theory, the null distribution not only requires repeated sampling but also for the null hypothesis to be true... so how do we estimate it?
- In our next lab we will learn about *randomization approaches* aimed at *simulating* the null distribution

Conclusion

Right now you should...

1. Understand null hypotheses and how p-values measure the evidence against the null
2. Understand how randomization allows us to replicate the study/experiment under the null hypothesis
3. Know how to perform a randomization test using StatKey
4. Be aware of p-value misconceptions

These notes cover Sections 4.1 - 4.3 of the textbook, I encourage you to read through those sections and their examples