Statistical Testing

Ryan Miller

Polio Epidemic Case Study

- In the early 1950s the US experienced an outbreak of polio that reached 58,000 new cases in 1952
- At this point several vaccines had been developed, with one developed by Jonas Salk appearing to be particularly promising based upon laboratory studies
- 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 each child recruited into the study 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 somewhat 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 this 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 due to children from poorer backgrounds being 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 was placebo controlled and double-blinded
  - Children in the control group were given an injection of a saline solution
  - The child, their parents, or their doctors didn't know 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 is lower in the treatment group, but what is causing this?

- Confounding? No, proper randomization was used
- Perception bias among the treated? No, a placebo was used
- Diagnostic bias? No, the doctors/participants were blinded
- 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, but how can we determine whether the results observed in our sample are convincing evidence that the population will benefit from the vaccine?

One question we could ask of the experiment is:

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

Statistical Tests

The hypothetical situation: *"What if the vaccine made no difference"* represents a **null hypothesis**, or the notion 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:

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

- The goal of statistical testing is to quantify how plausible the null hypothesis is in light of the data we've 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 formally called the *p*-value
- The smaller the *p*-value, the stronger the evidence is against the null
  - A *p*-value of 0.01 means: if the null hypothesis were true, only 1/100 samples would produce results as or more extreme than what we observed in our sample

Null and Alternative Hypotheses

Generally we pair the null hypothesis with an **alternative hypothesis**:

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

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

- The alternative hypothesis offers a sensible conclusion if our sample indicates the null hypothesis is unlikely
- Honestly, I don't suggest paying too much attention to it, most of our focus will be on the null hypothesis

The Burden of Proof

- Statistical tests are a formal way of executing the scientific method
- *p*-values provide evidence as to whether or not a theory should be rejected if the observed data are too different from what the theory predicts
- A subtle but crucially important piece of the scientific method is that it can never prove a hypothesis, it can only disprove a competitor
  - Albert Einstein: "No amount of experimentation can ever prove me right, but a single experiment can prove me wrong"
  - Put into our terms: "We can never prove the null hypothesis, we can only disprove it""

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" suggested 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

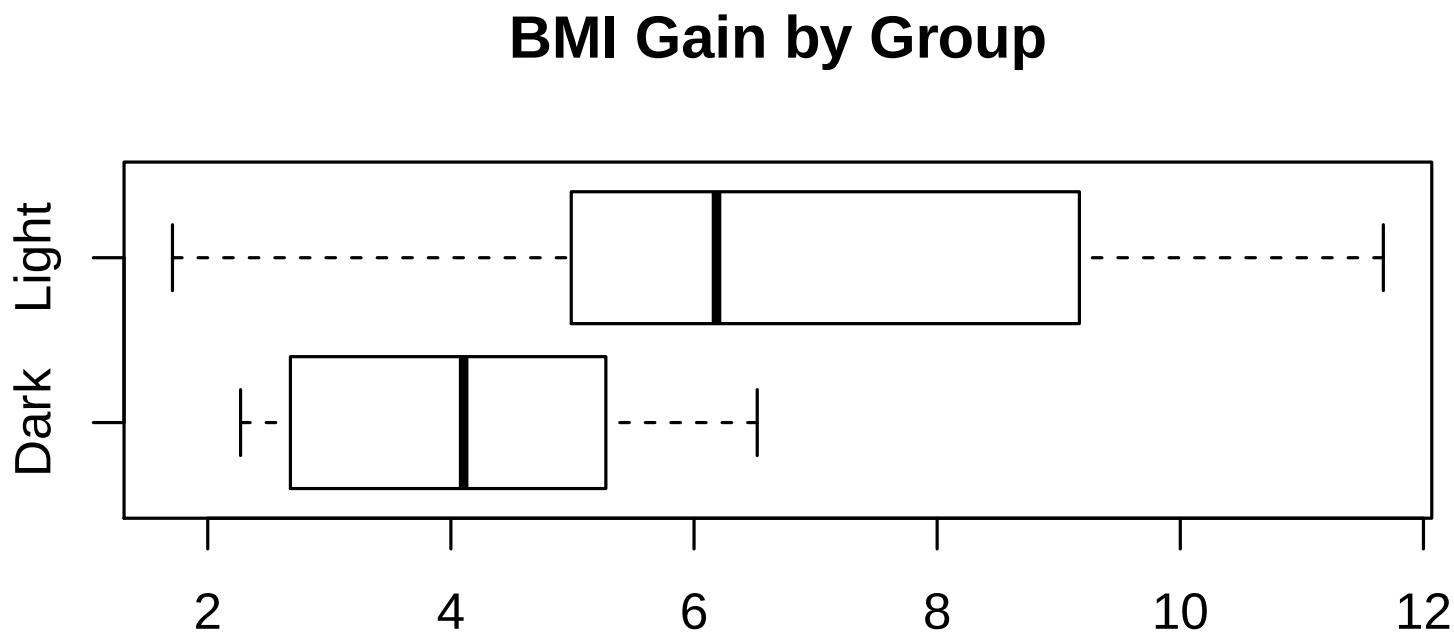
- 0.05 is an arbitrary cutoff that shouldn't distract you from the main idea behind p-values
- A p-value of 0.0001 doesn't tell you the same thing as a p-value of 0.04, even though both are "statistically significant"
- When reporting results you should include the p-value itself, not just whether or 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

Randomization Distributions

The "LightatNight" experiment randomized young mice to live in complete darkness, or a light on at night, to evaluate the research question "does light at night lead to weight gain?"

	1	2	3	4	5	6	7	8	9	10
Group	Light	Light	Light	Light	Light	Light	Light	Light	Light	Light
BMGain	1.71	4.67	4.99	5.33	5.43	6.94	7.15	9.17	10.26	11.67

	11	12	13	14	15	16	17	18
Group	Dark	Dark	Dark	Dark	Dark	Dark	Dark	Dark
BMGain	2.27	2.53	2.83	4.00	4.21	4.60	5.95	6.52



Randomization Distributions

To answer this research question using statistical testing, we need to do 3 things:

1. Determine an appropriate statistic to address the research question
2. Find the distribution of that statistic, *given the null hypothesis is true*, this is called the **randomization distribution**
3. Locate the value of the test statistic that we observed in our sample in the randomization distribution

- The randomization distribution approximates the sampling distribution of our chosen statistic, *when the null is true*
- The boostrap distribution approximates the sampling distribution, *in reality*
- This distinction is very important in understanding how the two distributions relate

Randomization Distributions

For the "LightatNight" experiment, we can construct the randomization distribution by repeatedly *shuffling* (permuting) the group labels of the cases in our sample:

Table 1: Original Data

Group	Light	Light	Light	Light	Light	Light	Light	Light	Light	Light
BMGain	1.71	4.67	4.99	5.33	5.43	6.94	7.15	9.17	10.26	11.67

Group	Dark	Dark	Dark	Dark	Dark	Dark	Dark	Dark
BMGain	2.27	2.53	2.83	4.00	4.21	4.60	5.95	6.52

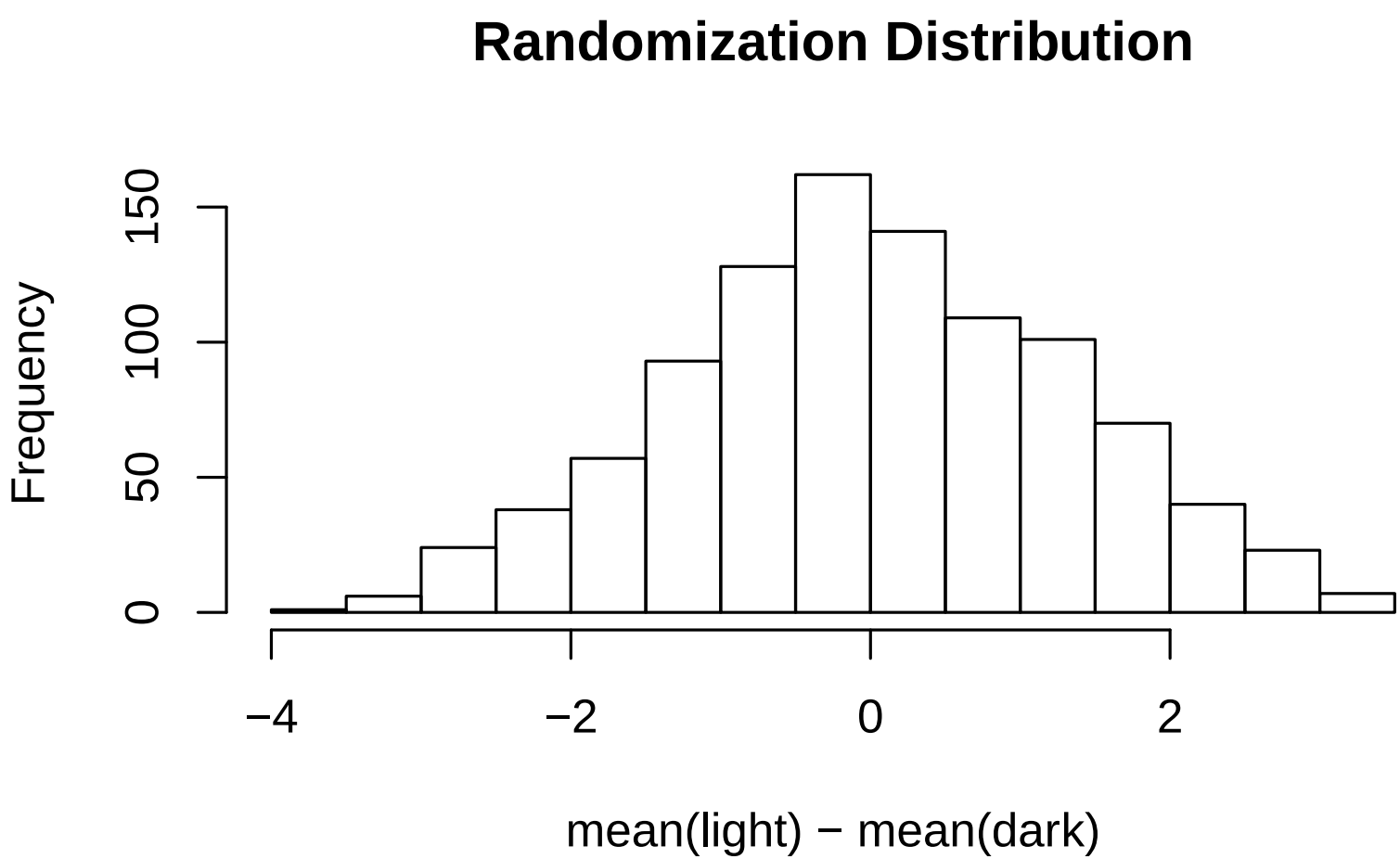
Table 2: Shuffled Data

Group	Light	Dark	Light	Dark	Dark	Light	Light	Dark	Light	Light
BMGain	1.71	4.67	4.99	5.33	5.43	6.94	7.15	9.17	10.26	11.67

Group	Light	Light	Light	Dark	Light	Dark	Dark	Dark
BMGain	2.27	2.53	2.83	4.00	4.21	4.60	5.95	6.52

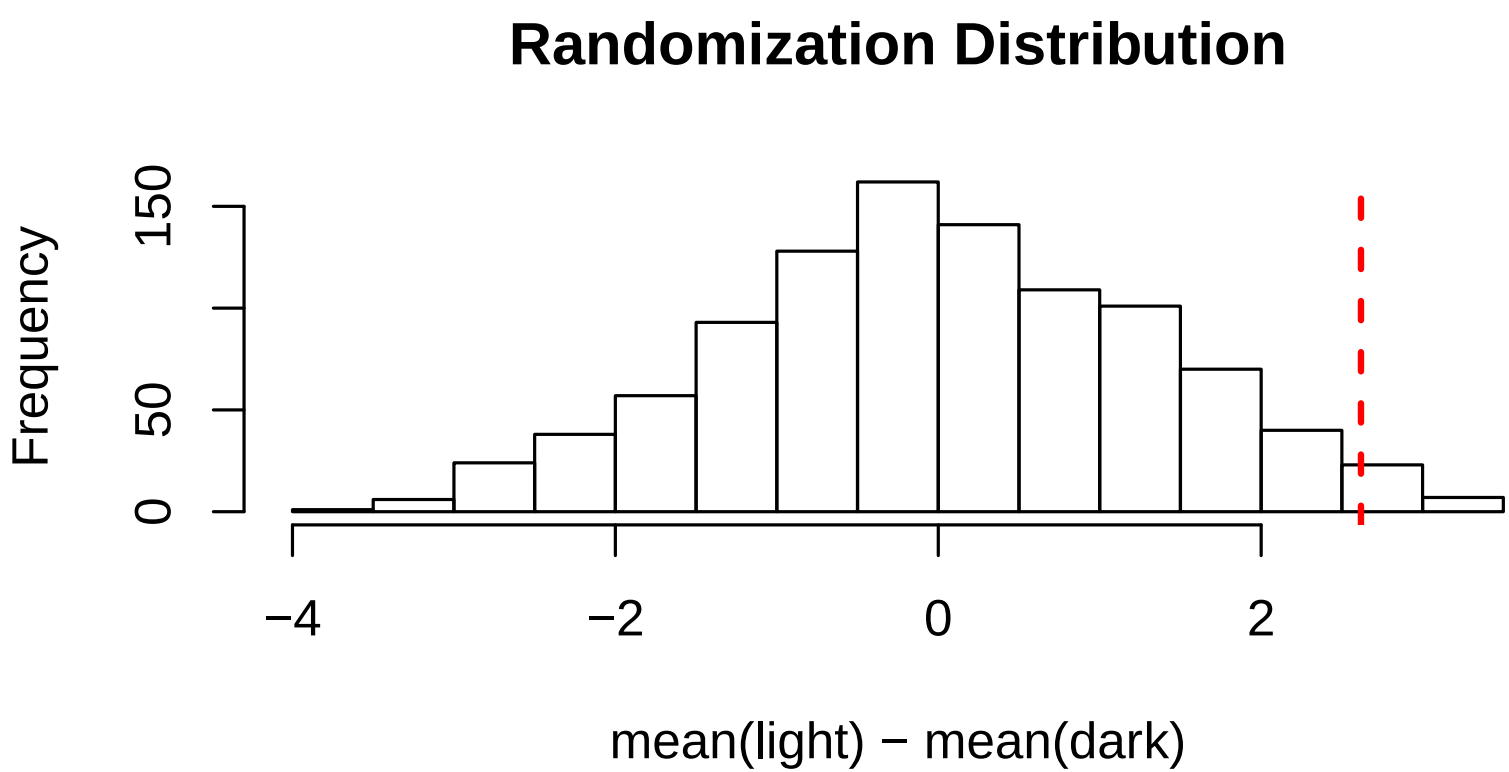
Randomization Distributions

Similar to bootstrapping, we calculate a statistic in each of our shuffled samples. For the "LightatNight" data, a suitable statistic is the difference in means: $\bar{x}_{\text{light}} - \bar{x}_{\text{dark}}$



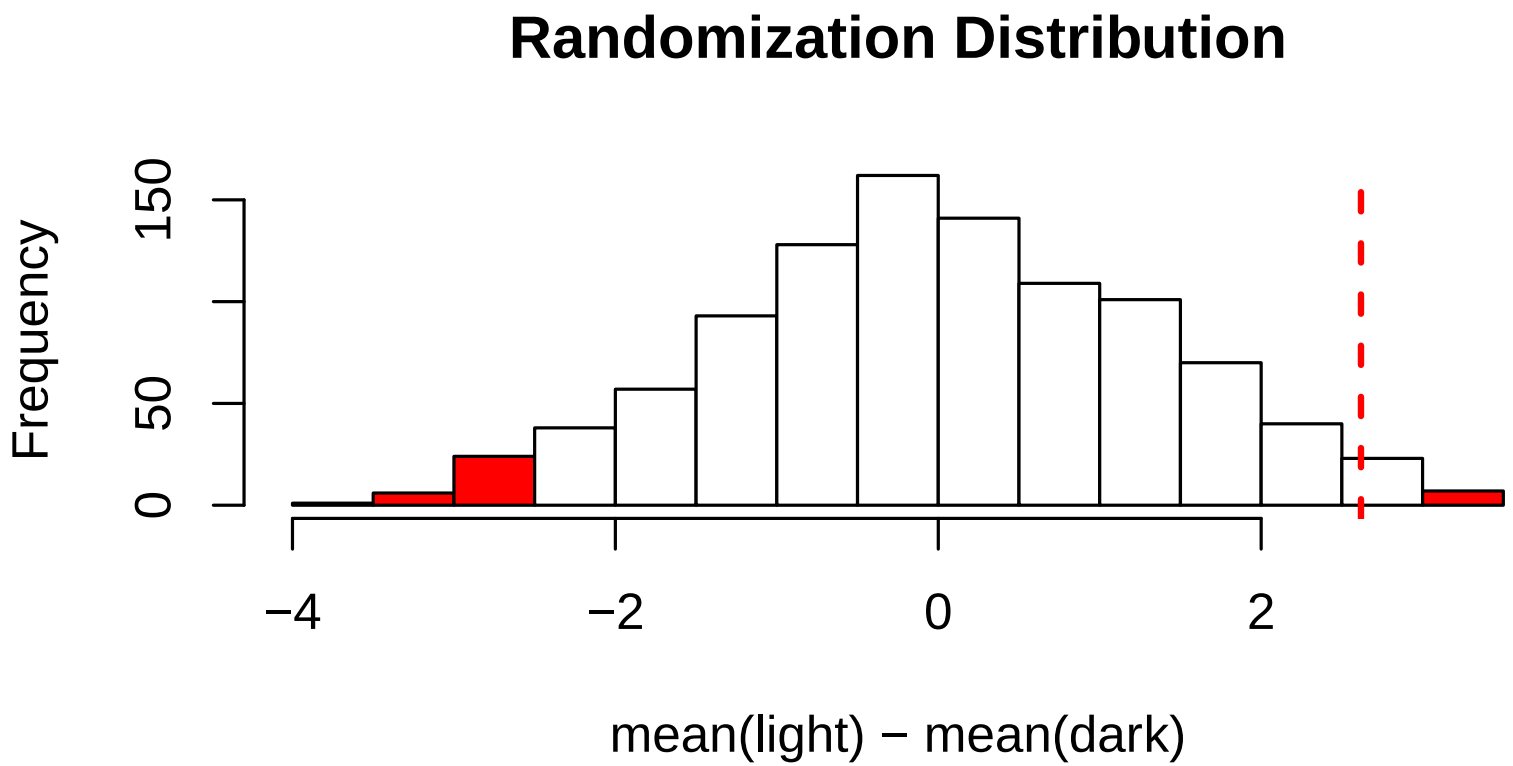
Randomization Distributions

The randomization distribution of $\bar{x}_{\text{light}} - \bar{x}_{\text{dark}}$ is centered at 0, which is exactly what we'd expect to be the most likely result under the null hypothesis. We can use this distribution to determine the p-value by locating the statistic actually observed in our sample:



Randomization Distributions

Only 16 shuffled samples had larger statistics, another 22 had more extreme negative statistics. Thus, the p-value for this experiment is approximately $38/1000 = 0.038$



Randomization Practice in StatKey

1. Go to "Test for Difference in Means" under "Randomization Tests" in StatKey
2. Change the dataset to "Mosquitos". In this dataset, individuals were assigned to drink water or beer, researchers then tracked the number of mosquito bites they received
3. Closely examine the original sample, pay close attention to the total number of individuals who received each number of bites (ie: 5 total people had 24 bites)
4. Generate 1 randomized sample, pay close attention to the number of individuals who received each number of bites. How is this randomized sample incorporated into the randomization distribution?
5. Now generate 2000 randomized samples, use this to find the p-value of the observed difference in means of the actual experiment

Putting it all together

Statistical testing is a formal procedure consisting of the following steps:

1. Clearly stating null and alternative hypotheses
2. Specifying an appropriate test statistic and a plan for collecting the data to construct that statistic
3. Calculating a test statistic for the observed data
4. Comparing the test statistic to a **reference distribution**, such as the randomization distribution, to obtain the p-value
5. Using the p-value to make a conclusion about the plausibility of the null hypothesis, this should be expressed in the context of the research question

Making Mistakes

Suppose we use a p-value cutoff of 0.05 to decide whether or not our findings are **statistically significant**. This threshold leads to the following possibilities:

- If $p < 0.05$ and the null hypothesis is false, we reach the correct conclusion
- If $p > 0.05$ and the null hypothesis is true, we reach the correct conclusion
- If $p < 0.05$ and the null hypothesis is true, our conclusion is wrong
- If $p > 0.05$ and the null hypothesis is false, our conclusion is wrong

So there are two types of mistakes we can make, statisticians demonstrated their creativity by giving them the exciting names: **type I** and **type II errors**

Making Mistakes

	H_0 True	H_0 False
Reject H_0	Type I Error	Correct Rejection
Fail to Reject H_0	Correct Decision	Type II Error

Making Mistakes

- Type I and II errors are each different types of mistakes and they have different consequences
- Type I errors inject a false conclusion into the scientific conversation
  - It can take tremendous amounts of time and resources to invalidate the original finding
- Type II errors generally go unnoticed, but they are a missed opportunity for scientific progress and can potentially deter future research

Making Mistakes

Suppose we conduct a lot of hypothesis tests using a significance level of $\alpha = 0.05$, we can summarize the results in the following two-way frequency table:

	H_0 is true	H_0 is false
Don't reject H_0	a	b
Reject H_0	c	d

From this table, we define a few key quantities:

- The **type I error rate** $= c/(a + c)$: The rate at which null hypotheses are falsely rejected
- The **type II error rate** $= b/(b + d)$: The rate at which non-null hypotheses fail to be rejected
- The **false discovery rate** $= c/(c + d)$: The fraction of null hypothesis rejections that were incorrect

Making Mistakes

- A fundamental property of the *p*-value is that using a significance level of α to determine statistical significance guarantees that the type I error rate of the testing protocol is less than α
- This means that we can control the long run type I error rate with our choice of significance level
- The traditional cutoff of 0.05 results in, on average, 1/20 situations with a true null hypothesis being type I errors
- The significance level tells us nothing about the type II error rate, or the false discovery rate!

Controversies

- 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

- One 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 us absolutely nothing 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
  - Does this justify the conclusion that Steph Curry and I are equally good shooters?

p-value Misconceptions

While that hypothetical example illustrates the problem, you're likely thinking that no one actually 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 frequent mistake is mistaking a statistically significant result for a *clinically significant* result
  - Statistical significant means 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

Nexium vs. Prilosec

- 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**

Summarizing the Pros and Cons of Hypothesis Testing

- Pros:
  - The most attractive feature of statistical testing is that the p-value has the same meaning regardless of the application, the testing procedure
  - You can understand the implications of a particular p-value without understanding the potentially complicated mathematical details of hypothesis test used to obtain it
- Cons:
  - p-values are limited by the fact that they don't actually tell us how different the groups are (effect size)
  - They also can't tell us if the null hypothesis is true

Connecting Confidence Intervals and Hypothesis Tests

1. The **sampling distribution** shows the distribution of sample statistics from a population
2. The **bootstrap distribution** simulates the distribution of sample statistics from a population using a single sample
3. The **randomization distribution** simulates the distribution of sample statistics, *under the null hypothesis*, from a population using a single sample

Bootstrapping and randomization are important and complimentary tools in understanding our sample and making inferences about the population

Connecting Confidence Intervals and Hypothesis Tests

- When the parameter value specified in H_0 is *outside* of the 95% confidence interval, a hypothesis test would *reject* H_0 at the $\alpha = 0.05$ level
- When the parameter value specified in H_0 is *inside* of the 95% confidence interval, a hypothesis test would *not reject* H_0 at the $\alpha = 0.05$ level

This relationship is flexible, a 99% confidence interval coresponds with a test at the $\alpha = 0.01$ level and a 90% confidence interval correpsonds with a test at the $\alpha = 0.1$ level

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 - 4.5 of the textbook, I encourage you to read through those sections and their examples