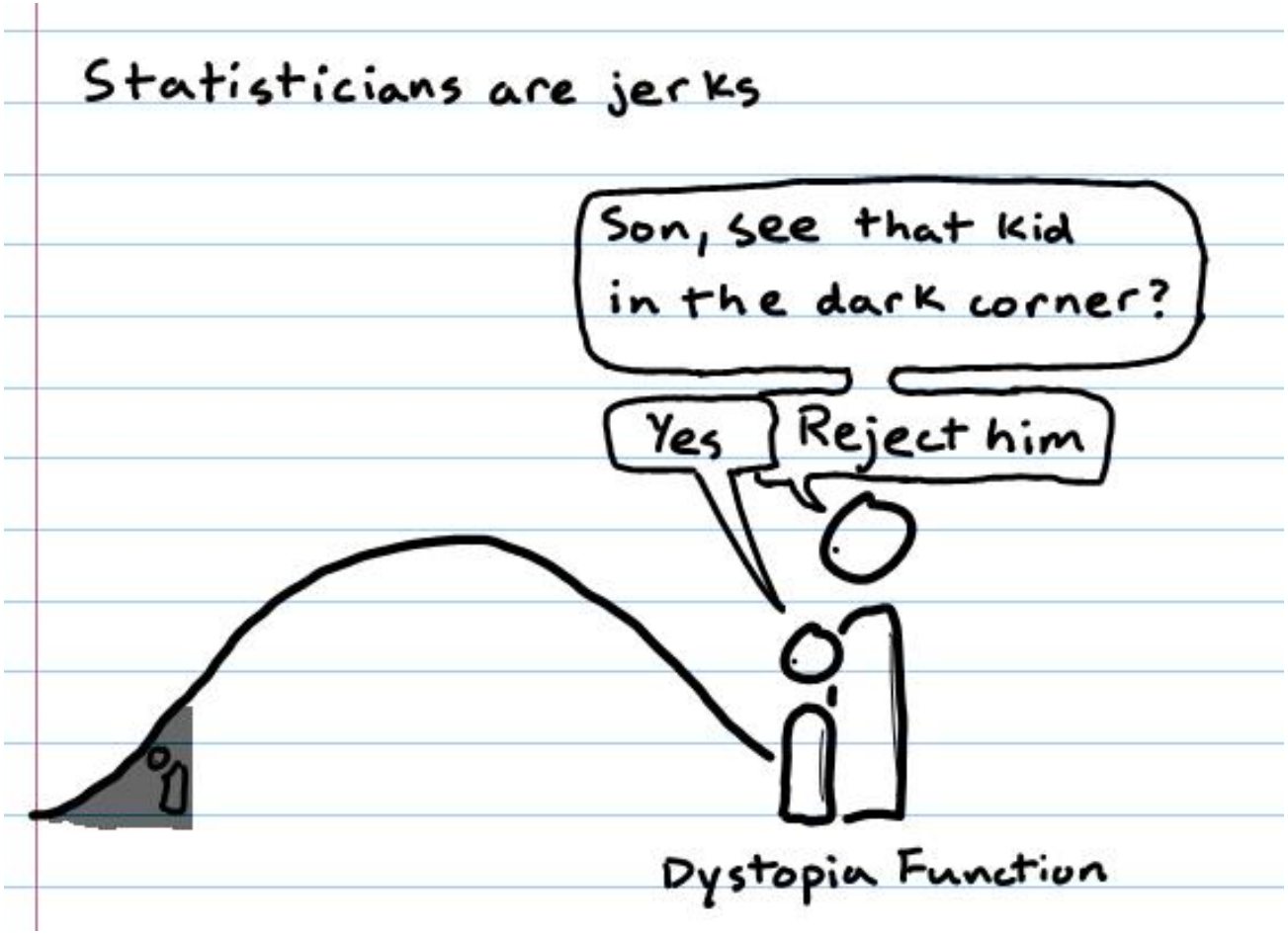
Statistical Hypothesis Testing

Javier E. Flores

February 20, 2019

Obligatory Joke



(The joke will make sense eventually...)

Statistical Testing

So far, we've discussed several statistical concepts that have allowed us to work with and understand data.

We've learned about study design, data visualization, interval estimation, and a myriad of other important statistical ideas.

While each of these tools/concepts offers a great degree of utility, none provide us with a way to formally test a belief or *hypothesis* that we might have.

With this in mind, we'll use the historic polio epidemic of the 1950s to introduce a statistical testing framework.

Polio

Polio, or poliomyelitis, is a viral disease which primarily affects children and, in the most severe of cases, may lead to paralysis, difficulty breathing, and death.

In the early 1950's this disease was spreading like wildfire among US children, with about 58000 new cases in 1952.

In response to this outbreak, the US Public Health Service organized a large study in 1954 involving nearly one million children.

Before the start of this study, Jonas Salk had developed his (now) famous anti-polio vaccine but only had preliminary laboratory data to support its efficacy.

The primary goal of the larger scale 1954 study was to obtain definitive proof of the efficacy of Salk's vaccine in preventing polio.

Question: If you were running this trial in 1954, would you perform a randomized experiment?

1954 Polio Study

In this situation, performing a randomized experiment would be controversial.

Considering the potentially deadly nature of this disease, leaving some children unvaccinated would be ethically compromising.

To address the randomized experiment issue, the trialists thought to offer the vaccine to all children whose parents provided consent and use those children whose parents refused the vaccine as a control.

Question: Is this a true workaround to the problem?

Nope!

Any well-trained statistician should suspect that those parents who provide consent are most likely characteristically different from those that don't.

In this case, this suspicion would be correct: parents who provided consent typically had higher incomes and, consequently, had children who were more likely to develop polio.

What's the connection between wealth and polio?

> Children from wealthier families were thought to have been raised in cleaner, more sterile environments.
>
> As a result, these children were not exposed (during early childhood) to milder cases of polio from which they could build immunity.

Catch 22

In performing a randomized experiment an issue of ethics is raised, whereas using non-consenting children as controls introduces confounding bias to the study.

So what was decided?

> A randomized trial among children of consenting parents.

While certainly ethically questionable, physicians were able to sleep (somewhat) because the study would potentially save thousands of lives (and was double-blinded so physicians wouldn't know if/when they were leaving children untreated).

> Children (randomly) assigned to the control group received a placebo saline injection, and the treatment group received Salk's vaccine.
>
> Neither the child, their parent, or the administering physician were aware of which vaccine - placebo or not - was being given.

Study Results

1954 Polio Trial Results

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

What do these data tell us?

Should we worry about confounding? Any forms of bias?

Can we assume that the population at large will see these same effects?

Study Results

Within the study sample, the vaccine was clearly effective. Polio incidence in the vaccine group was nearly a third of the control.

This study met the "gold standard" in trial design as it was a double-blinded, placebo controlled, randomized experiment. Therefore, there should be little to no concern of confounding or any other biases influencing these results.

The question of generalizability is what statistical hypothesis testing (SHT) helps us answer.

Statistical Hypothesis Testing (SHT)

Knowing what we know about statistics, it is highly unlikely that the population will see exactly the same effects observed in this study sample. (Recall that statistics are our best guesses for population parameters but are subject to some error.)

It's even possible that the observations in our data are entirely due to chance! This would imply that the population would see *no* benefit from the introduction of the Salk vaccine.

With SHT, we assume this "worst case scenario" of no effect and ask ourselves how likely our sample results are under this assumption.

As an example, for this experiment we would ask:

"If the vaccine truly made no difference, how likely would it be to observe an incidence rate nearly a third lower in the vaccinated group than the non-vaccinated group?"

Null Hypothesis

We call this assumption of a null effect (i.e. "the vaccine truly makes no difference") our **null hypothesis**.

Under the null hypothesis, both population parameters are exactly the same (e.g. incidence rates among vaccinated and unvaccinated children) and any differences observed in a given sample are due to random chance.

The following notation is typically used to describe the null hypothesis:

$$H_0 : \mu_A = \mu_B,$$

where A and B represent two populations of interest (e.g. vaccinated and unvaccinated children).

p-values

In order to assess the plausibility of our null hypothesis, we use our sample data to determine how likely our results (or those that are more extreme) are under the assumption of a null effect.

The **p-value** refers to the *probability* of obtaining results as or more extreme than those observed in our sample, provided the null hypothesis is true.

Smaller p-values are indicative of greater evidence against the null hypothesis.

Alternative Hypothesis

It is generally the case that our null hypothesis is paired with some **alternative hypothesis**.

In the event that sufficient evidence is brought against our null hypothesis, the alternative hypothesis provides recourse in making a conclusion.

Using statistical notation,

$$H_A : \mu_A \neq \mu_B$$

Quick Practice

Scenario: Two Youtube stars, Jake and Logan Paul (brothers), are in a seemingly constant battle to make news headlines for pulling a controversial stunt. "Jake-Paulers", fans of Jake Paul, claim that the head of the "Logang", Logan Paul, is the more controversial Youtuber. In a mission to prove their claim, "Jake-Paulers" collected data comparing the difference in proportions of headlines that were controversial between these brothers and obtained a p-value of 0.01.

- In this scenario, what is the null hypothesis?
- What is the alternative hypothesis?
- With a p-value of 0.01, is it correct to say that there is a 1% chance that the null hypothesis is true? Why or why not?

Burden of Proof

In the context of the polio study in particular, you might see some parallels between the framework of statistical testing and the scientific method.

Statistical hypothesis testing may be thought of as a formal way of executing the scientific method.

We use p-values in order to quantify evidence against some postulated theory (i.e. the null hypothesis) and arrive at some new conclusion (i.e. alternative hypothesis) should sufficient evidence be available to disprove the original idea.

One point of emphasis here is that we can only disprove some null hypothesis. As Albert Einstein once said, "No amount of experimentation can ever prove me right, but a single experiment can prove me wrong."

Statistical Significance

Ronald Fisher, a legend among statisticians, was also the mind behind this measure of evidence that we call the p-value.

Fisher suggested the following guidelines in using the p-value to assess evidence against the null hypothesis:

p-value	**Evidence against H_0**
0.100	Borderline
0.050	Moderate
0.025	Substantial
0.010	Strong
0.001	Overwhelming

It is by these recommendations that modern science has decided to use 0.05 as a threshold for *rejecting* the null hypothesis and for claiming *statistical significance*.

Statistical Significance

Despite being based on recommendations by the man who some have credited as being "a genius who almost single-handedly created the foundations of modern statistical science", these are still somewhat arbitrary cutoffs.

Even more important than these cutoffs are a correct understanding of what a p-value truly is.

> Is there truly a difference between a p-value of 0.0499 and 0.0501? By these cutoffs, one is statistically significant and the other is not. (Pretty silly!)
>
> On the other hand even though both 0.0001 and 0.04 are statistically significant, one is substantially more compelling than the other!

When reporting results, you should always include the p-value itself and not just whether it was "statistically significant".

Randomization Distributions

How does one obtain a p-value, exactly?

There are several ways, but we'll first learn how to compute p-values using **randomization distributions**.

Randomization distributions are very much like sampling distributions but with one crucial difference: randomization distributions are formed assuming some null hypothesis.

To demonstrate what I mean by this, we'll take a look at data obtained from a randomized experiment performed on mice.

Randomization Distribution

This experiment randomized young mice to live in either complete darkness or with a light on at night in order to determine whether weight gain was associated with having a light on at night.

The data from this experiment are provided in the table below.

Mouse ID	Group	BMGain
1	Light	1.71
2	Light	4.67
3	Light	4.99
4	Light	5.33
5	Light	5.43
6	Light	6.94
7	Light	7.15
8	Light	9.17
9	Light	10.26
10	Light	11.67
11	Dark	2.27
12	Dark	2.53
13	Dark	2.83
14	Dark	4.00
15	Dark	4.21
16	Dark	4.60
17	Dark	5.95
18	Dark	6.52

Randomization Distribution

Prior to forming our randomization distribution, we first need to address a couple of questions:

- What is our null hypothesis?
- What statistic can we use to assess the claim specified by our null?

For this randomized experiment, our null hypothesis would be that the weight gain is the same in both groups. In other words, being assigned to live in light or darkness does not matter.

One statistic we could use to assess this claim is the difference in means between each group, $\bar{x}_{light} - \bar{x}_{dark}$.

Randomization Distribution

For this particular example, we've determined the null hypothesis as well as the statistic needed to test it (i.e. $\bar{x}_{light} - \bar{x}_{dark}$).

Question: If we wanted to form the sampling distribution for the mean difference, what would we do?

Question: If we wanted to form the bootstrap distribution for the mean difference, what would we do?

With both the bootstrap and sampling distributions, the fundamental idea is to draw repeated samples and compute the statistic of interest from each to form a distribution.

Randomization Distribution

This fundamental idea holds also for the randomization distribution, but each new sample is formed by repeatedly **permuting**, or shuffling, the group labels of the cases in our sample.

Question: Why do we resample in this way?

Original Sample

Mouse ID	Group	BMGain
1	Light	1.71
2	Light	4.67
3	Light	4.99
4	Light	5.33
5	Light	5.43
6	Light	6.94
7	Light	7.15
8	Light	9.17
9	Light	10.26
10	Light	11.67
11	Dark	2.27
12	Dark	2.53
13	Dark	2.83
14	Dark	4.00
15	Dark	4.21
16	Dark	4.60
17	Dark	5.95
18	Dark	6.52

"New Sample"

Mouse ID	Group	BMGain
1	Light	1.71
2	Dark	4.67
3	Dark	4.99
4	Light	5.33
5	Light	5.43
6	Dark	6.94
7	Light	7.15
8	Light	9.17
9	Dark	10.26
10	Light	11.67
11	Dark	2.27
12	Dark	2.53
13	Light	2.83
14	Dark	4.00
15	Light	4.21
16	Light	4.60
17	Light	5.95
18	Dark	6.52

Randomization Distribution

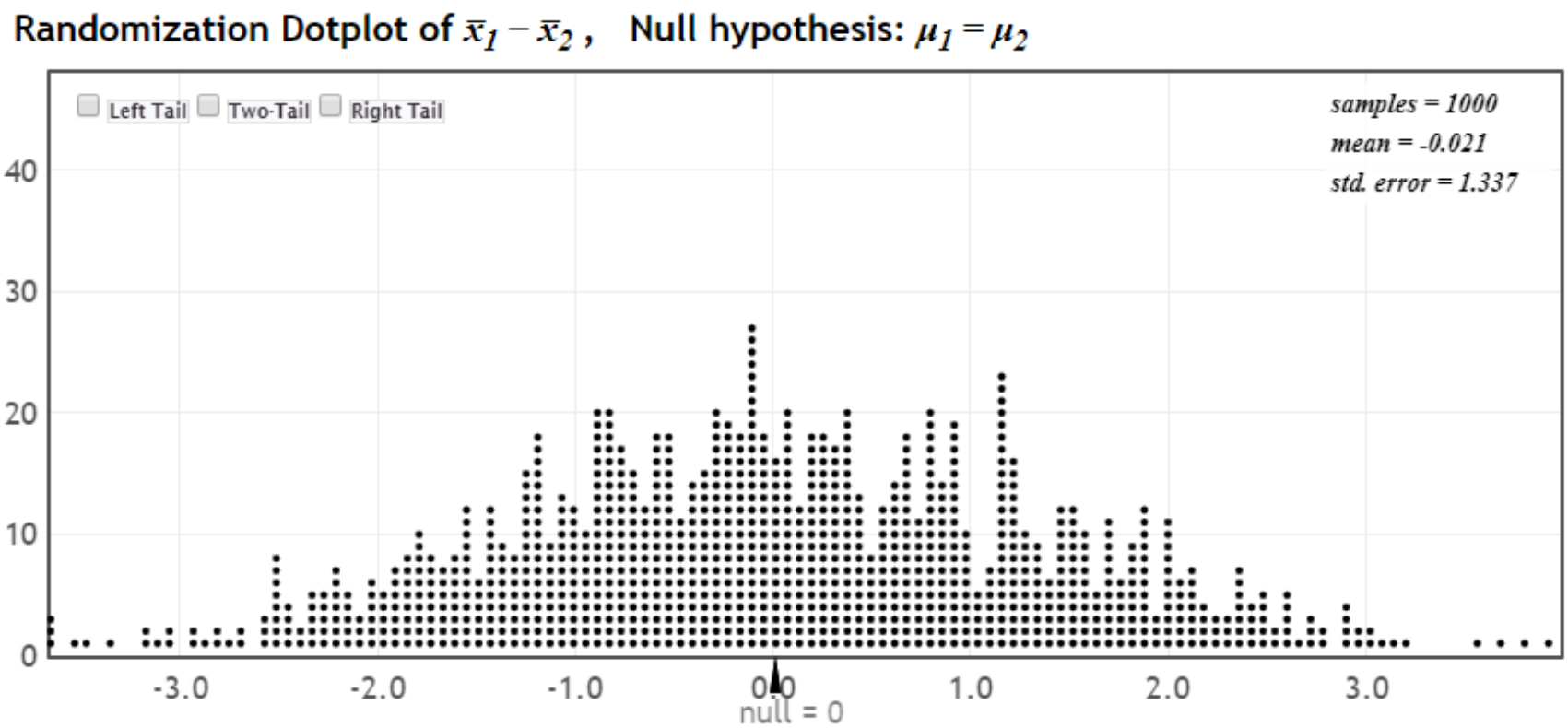
Remember, the null hypothesis states that the treatment group does not matter.

Therefore, when assuming the null is true, it wouldn't matter which cases are labeled as "Light" and which are "Dark".

When we then compute the statistic $(\bar{x}_{light} - \bar{x}_{dark})$ for each permuted sample, we expect there to generally be no difference between groups.

Our randomization distribution should then be centered at 0.

Randomization Distribution



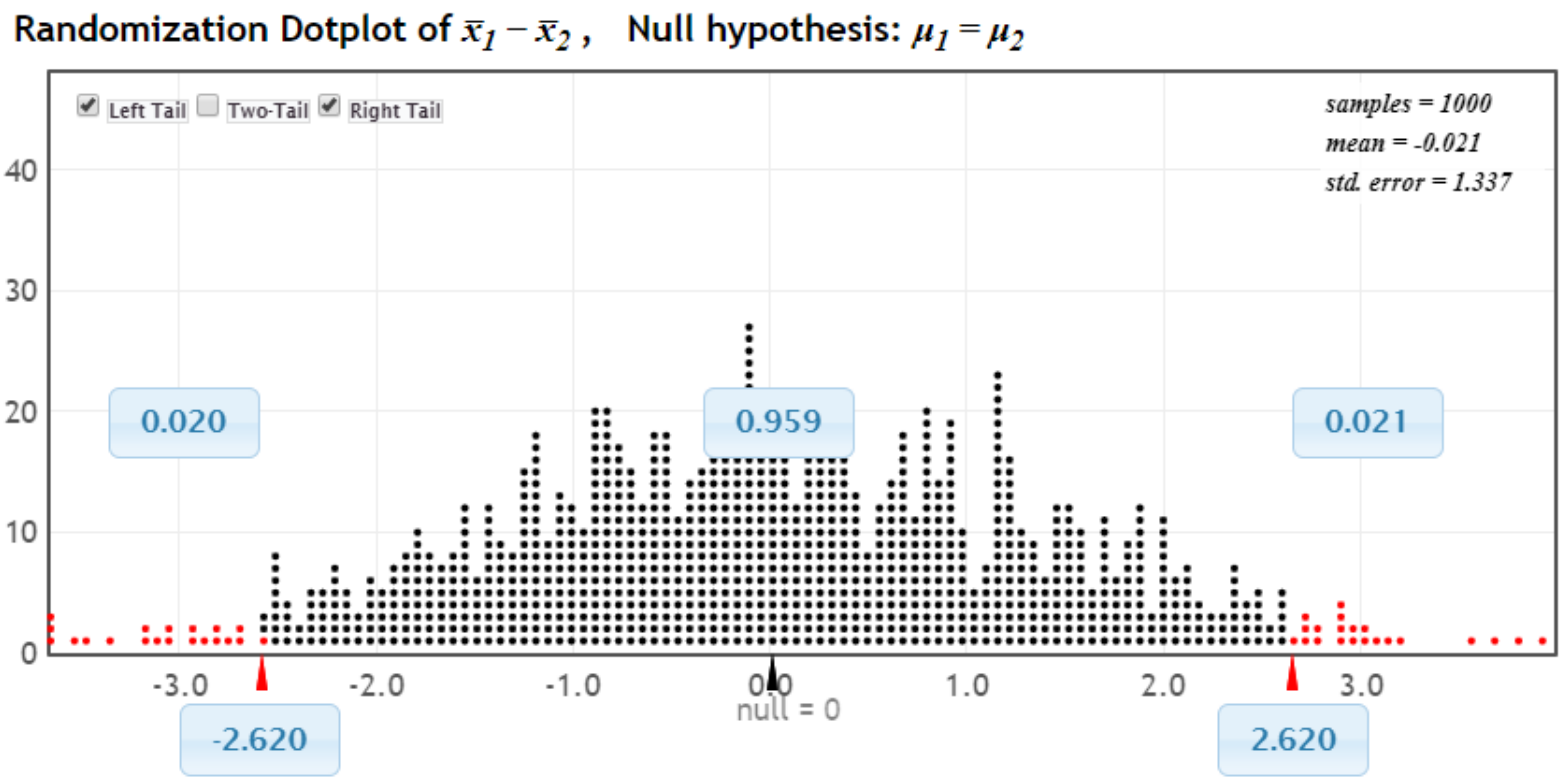
Finding the p-value

Remember that the p-value is the probability of obtaining results as or more extreme than those observed in our sample, provided the null hypothesis is true.

Therefore, in order to get our p-value, we count all the statistics whose values were as or more extreme than the statistic in our original sample and divide by the total number of permuted samples generated.

Finding the p-value

The mean difference in our original sample was 2.62. In forming the randomization distribution, there were 41 samples whose statistics were as or more extreme than 2.62. Our p-value is then .041.



Statistical Hypothesis Testing

In summary, statistical hypothesis testing consists of the following steps:

1) State the null and alternative hypotheses.
2) Determine an appropriate test statistic for evaluating these hypotheses and specify a plan for collecting data necessary to compute this statistic.
3) Calculate and catalog the test statistic using collected data.
4) Compare the observed data test statistic to a **reference distribution**, such as the randomization distribution, to obtain a p-value
5) Use the p-value to determine the validity of the null hypothesis. Any conclusions should be expressed in terms of the original research question.

Practice

On a hot summer weekend, a couple of scientists wanted to find out whether drinking beer or water had an effect on the number of mosquito bites received. These scientists threw a summer party inviting several of their friends over and randomly assigned them to drink either beer or water.

- State the null and alternative hypotheses.
- Go to www.lock5stat.com/StatKey/ and choose the appropriate test statistic from one of the options under the column "Randomization Hypothesis Tests".
- Change the dataset to "Mosquitos", and determine the test statistic for the original sample.
- Generate 1 randomized sample and compute the test statistic.
- Generate 2000 randomized samples and find the p-value.

Solution

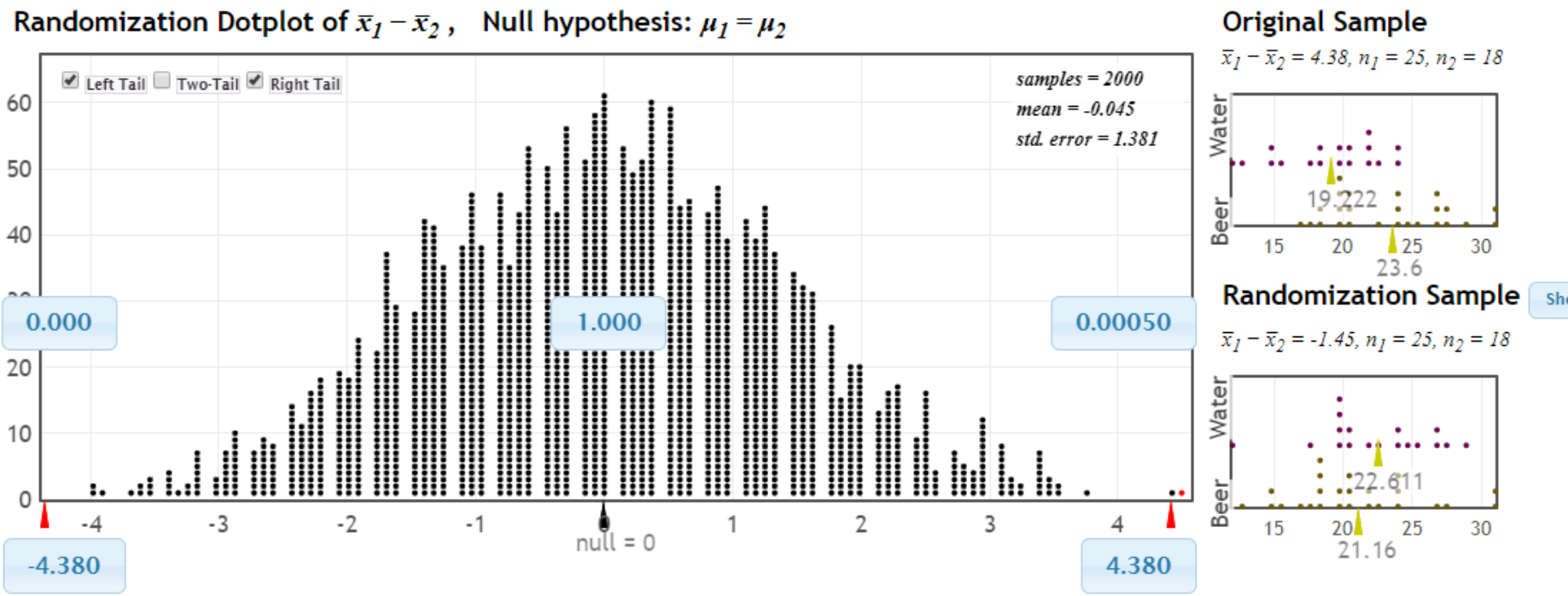
The null hypothesis is that there is no difference in mosquito bite frequency between beer and water drinkers. The alternative hypothesis is that there is.

The appropriate test statistic would be the difference in means. If the null is true, we would expect the average difference in number of bites in beer drinkers and water drinkers to be 0.

For the original sample, the mean difference $(\bar{x}_{beer} - \bar{x}_{water})$ is 4.38.

Solution

The p-value is 0.0005.



Drawing Conclusions ($p < 0.05$)

In the previous example, we found a p-value of 0.0005. Given that this is below the threshold of 0.05, we would deem this **statistically significant**.

Despite finding statistical significance, there are still two distinct possibilities to consider:

1) The null hypothesis is indeed false and our results lead us to the correct conclusion.
2) The null hypothesis is actually true and our results lead us to an incorrect conclusion.

We call the second possibility a **type I error**.

Drawing Conclusions ($p > 0.05$)

If, on the other hand, we had found a p-value above 0.05, we would no longer have a statistically significant result.

In this scenario, there are also two distinct possibilities to consider:

1) The null hypothesis is indeed true and our results lead us to the correct conclusion.
2) The null hypothesis is actually false and our results lead us to an incorrect conclusion.

We call the second possibility a **type II error**.

Statistical Significance

		The Truth (Based on Entire Population)	
		Nothing Is There (H_0 Is True)	Something Is There (H_0 Is False)
Your Conclusion (Based on Your Sample)	I Don't See Anything (Nonsignificant)	Right!	Wrong (Type II Error)
	I See Something (Significant)	Wrong (Type I Error)	Right!

Practice

Suppose you are the sitting judge on a murder trial.

- If you were to judge the defendant as guilty when they were truly innocent, which type of error would you be making?
- If you were to judge the defendant as innocent when they were truly guilty, which type of error would you be making?

Suppose that in this world there are only two types of people: Idiots and Non-idiots. Suppose also that there exists a test (that is not 100% reliable) capable of differentiating between these two types of people.

- Assuming people are considered idiots unless proven otherwise, which type error describes the situation in which an idiot passes the test?
- What if a non-idiot fails the test?

Solution

Provided that the defendant is "Innocent until proven guilty", we have the following:

	Defendant Innocent	**Defendant Guilty**
Innocent Verdict	Correct	Type II Error
Guilty Verdict	Type I Error	Correct

Giving a guilty verdict when the defendant was truly innocent would be committing a type I error.

A type II error would be judging the defendant as innocent when they were truly guilty.

Solution

We are told that people are considered idiots unless proven otherwise. Therefore:

	Actually an Idiot	**Actually a Non-Idiot**
Fail Test	Correct	Type II Error
Pass Test	Type I Error	Correct

If an idiot were to pass the test, a type I error would have been committed.

On the other hand, a type II error occurs when a non-idiot fails the test.

Consequences

As you may have gleamed from the previous examples, the consequences in making a type I or type II error differ.

In the murder trial example, type I errors lead to the conviction of innocent men whereas type II errors allow guilty men to walk free!

In a scientific context, type I errors introduce erroneous conclusions that may be used as the basis for false beliefs within a field of study.

On the other hand, type II errors may close off avenues of research and stymie scientific progress.

Error Rates

Oftentimes our interests are in answering multiple questions. Suppose that for each question, we conduct a hypothesis test at a **significance level** of $\alpha = 0.05$. If we had access to the underlying truth, we could construct the following table:

	H_0 **true**	H_0 **false**
Fail to Reject H_0	a	b
Reject H_0	c	d

From this table, we define a few key quantities:

- The **type I error rate** $= c/(a+c)$, which is the rate at which H_0 is falsely rejected.
- The **type II error rate** $= b/(b+d)$, which is the rate at which H_0 is falsely not rejected.
- The **false discovery rate** $= c/(c+d)$, which is the fractions of null hypothesis rejections that were incorrect.

Significance Level

On the previous slide, I used the phrase significance level when describing the performed hypothesis tests.

The significance level is a user-defined quantity that guarantees our testing procedure type I error rate is less than the specified quantity, α.

Traditionally, the significance level is set to 0.05 which results in, on average, 1/20 situations in which a type I error is committed.

The significance level does not control the type II error rate or false discovery rate.

Controversies

Since their inception, p-values have been increasingly misused and misinterpreted.

This has become so much of a problem that the largest professional organization of statisticians, the American Statistical Association (ASA), felt compelled to release a public statement on p-values.

In addition, there are some journals (e.g. *Basic and Applied Psychology*) that have banned the use of p-values entirely!

Misconceptions

Of the myriad of mistakes made with p-values, one of the most common is to believe the p-value is a probability that the null hypothesis is true.

Having a high p-value does not translate to a high probability of the null being true. If you aren't convinced of this, consider the following hypothetical scenario:

- 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 shooting, the p-value of a result as or more extreme than this is 0.17
- Since this p-value is "high" (above 0.05), does this mean I'm as good at shooting as Steph Curry?

Misconceptions

As ridiculous as the implied conclusion of that example was, this kind of thing manifests in some way or another more often than you'd think:

- In 2006, the Woman's Health Initiative found that low-fat diets were not 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"

Clinical and Practical Significance

Aside from conceptual misconceptions about the p-value, people often conflate the ideas of statistical and practical significance.

> Results that are statistically significant are those that are unlikely under an assumed null distribution.
>
> Results that are clinically (or practically) significant are those that have substantial meaning in the context of the research question.

Statistical significance does not imply practical significance.

In order to demonstrate this, consider the following example...

Nexium vs. Prilosec

Prilosec is a popular heartburn medication that was developed by AstraZeneca in the 1980's.

In 2001, the FDA patent for this medication expired. This compelled AstraZeneca to replace *Prilosec* with a new drug *Nexium* to avoid losing profit to competing post-patent variants of their already successful drug.

Omeprazole and Esomeprazole are the active ingredients for *Prilosec* and *Nexium*, respectively.

In the development of *Nexium*, it was found that using Esomeprazole over Omeprazole provided twice the effective dose at the same amount of drug.

Nexium vs. Prilosec

However, in comparing *Nexium* to its predecessor, AstraZeneca showed that the difference in healing rate for erosive esophagitis between the two was only 3%! (*Nexium*: 90%, *Prilosec*: 87%)

Largely due to the size of sample used to test this comparison (nearly 6000 subjects), this difference was statistically significant with a p-value well below 0.05.

Because AstraZeneca was able to show a statistically significant difference, the FDA approved Nexium and AstraZeneca spent millions marketing *Nexium* under the slogan, "Better is better".

The marketing campaign worked and AstraZeneca has since made over 47 BILLION DOLLARS from *Nexium*.

Nexium vs. Prilosec

With only a 3% difference in healing rate, the efficacy of either drug was practically the same.

The 95% confidence interval for the improvement factor was (1.02, 1.06).

And yet, despite this, doctors prescribed this marginally better, but substantially more expensive brand name drug over the equally effective, much more affordable off-brand variants of *Prilosec*.

So just to drive the point home: **statistical testing does not measure practical importance!**

Practice

The statements below may be rated as either "Terrible", "Bad", "OK", "Good", or "Excellent" depending on how well they communicate a meaningful conclusion. With your group, classify each statement. Be sure to discuss the reasoning behind each classification.

1) Our results are statistically significant so we reject the null hypothesis.
2) The p-value is 0.01, indicating strong evidence that Nexium is more effective than Prilosec in treating heartburn.
3) The p-value is 0.17, indicating that there is a 17% chance that *Nexium* and *Prilosec* are equally effective in treating heartburn.
4) The study provided borderline evidence ($p = 0.07$) that low-fat diets reduce breast cancer risk. While this fails to meet the threshold for statistical significance, the observed results are still relatively unlikely under the assumption of no effect. Therefore, it is plausible that low-fat diets have a small protective effect.
5) The study failed to reject the hypothesis that diet isn't associated with breast cancer risk.

Statistical Significance

1) Our results are statistically significant so we reject the null hypothesis. **BAD**
2) The p-value is 0.01, indicating strong evidence that Nexium is more effective than Prilosec in treating heartburn. **GOOD**
3) The p-value is 0.17, indicating that there is a 17% chance that *Nexium* and *Prilosec* are equally effective in treating heartburn. **TERRIBLE**
4) The study provided borderline evidence ($p = 0.07$) that low-fat diets reduce breast cancer risk. While this fails to meet the threshold for statistical significance, the observed results are still relatively unlikely under the assumption of no effect. Therefore, it is plausible that low-fat diets have a small protective effect. **EXCELLENT**
5) The study failed to reject the hypothesis that diet isn't associated with breast cancer risk. **OK**

Statistical Hypothesis Testing

One benefit to statistical testing a p-values are that p-values have the same interpretation regardless of the application and testing procedure.

Provided that you have a solid understanding of what a p-value actually is, you also are able to understand its implications without having to know the sometimes complicated mathematical details needed to obtain it.

On the other hand, p-values are limited in that they provide no information about clinical significance or effect size (i.e. how different two groups are).

p-values also do not tell us whether the null hypothesis is true.

Connection to Confidence Intervals

In describing some of the controversies surrounding p-values, I mentioned that certain journals have banned their use entirely.

For these journals, using p-values have been replaced by the (arguably) more informative confidence interval:

- When the parameter value specified in H_0 is outside of the 95% confidence interval, a hypothesis would *reject* H_0 at the $\alpha = 0.05$ level.
- Otherwise, if the interval contains the parameter value specified in H_0, we would *fail to reject* H_0 at the $\alpha = 0.05$ level.

More generally, a $(1 - \alpha) * 100\%$ confidence interval corresponds to a test with significance level α.

Wrap-Up

Right now, you should...

- Understand null hypotheses and their relation to p-values.
- Understand how to construct and utilize randomization distributions to test hypotheses.
- Recognize the limitations and misuses of p-values.
- Be aware of the relationship between hypothesis testing and confidence intervals.

These notes cover sections 4.1 - 4.5 of the textbook. Please read through the section and its examples along with any links provided in this lecture.