Testing Errors, Power, and Multiple Comparisons

Ryan Miller

Clofibrate

- In 1980, a study was published in the New England Journal of Medicine describing a randomized, placebo-controlled, double-blind experiment involving the drug clofibrate, which reduces blood cholesterol levels.
- Of the subjects randomly assigned to take clofibrate, adherers were those who took more than 80% of their prescribed pills:

	Number	Deaths
Adherers	708	15%
Nonadherers	357	25%
Total	1103	20%

Clofibrate

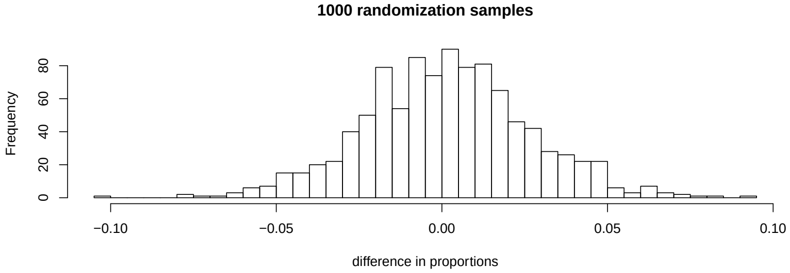
Is clofibrate effective? Let's use hypothesis testing:

$$H_0 : p_{\text{death}|\text{adherer}} - p_{\text{death}|\text{nonadherer}} = 0$$

We observed:

$$\hat{p}_{\text{death}|\text{adherer}} - \hat{p}_{\text{death}|\text{nonadherer}} = -0.10$$

The *randomization distribution* looks like this:



Clofibrate

With a *p*-value of approximately 0.001 (1/1000), we should be convinced that the observed difference in survival was not due to random chance. But does that mean that the difference was due to clofibrate?

	Clofibrate		Placebo	
	Number	Deaths	Number	Deaths
Adherers	708	15%	1813	15%
Nonadherers	357	25%	882	28%
Total	1103	20%	2789	21%

If we consider the experiment's placebo group, clofibrate no longer appears to be effective

Clofibrate

- This experiment should be analyzed using the **intent-to-treat** principle, applying ITT we see:

$$\hat{p}_{\text{death}|\text{clofibrate}} - \hat{p}_{\text{death}|\text{placebo}} = -0.01$$

- The corresponding hypothesis test yields an unconvincing *p*-value of 0.51
  - Using a **significance level** of $\alpha = 0.05$, we'd fail to reject the null hypothesis that clofibrate and placebo are equally effective
  - But is it *possible* that prescribing clofibrate really is better than prescribing placebo?

Clofibrate

- Yes, clofibrate *could* be better (remember that a high *p*-value doesn't *prove* the Null Hypothesis)
  - This would imply that our experiment and hypothesis test resulted in an error
  - In other words, we failed to reject H_0 with $p \geq \alpha$, but that was a mistake because H_0 is false and should be rejected
- Another type of error we could make is rejecting a null hypothesis that is actually true
- Any guesses on what *exciting names* statisticians have given these *two types of errors*?

Type I and Type II Errors

- A **type I error** occurs when the null hypothesis is *rejected*, but in reality it is *true*
- A **type II error** occurs when the null hypothesis *cannot be rejected*, but in reality it is *false*

	H0 is true	H0 is false
Don't Reject H0	Correct	Type II Error
Reject H0	Type I Error	Correct

Practice

For each scenario, describe (in words) what a Type I and Type II error would mean:

1. H_0 : Person A is not guilty of the crime vs. H_A : Person A is guilty of the crime
2. H_0 : Drug A doesn't cure disease B vs. H_A : Drug A cures disease B

Additionally, how do you think a data analyst could decrease the chances of making a Type I error? (assuming the data has already been collected)

Practice (Solution)

1. A type I error would be deciding an innocent person is guilty, a type II error would be deciding a guilty person is innocent
2. A type I error would be deciding that an ineffective drug is beneficial, a type II error would be deciding a beneficial drug is not effective

We could reduce our chances of making a type I error by lowering our significance threshold.

Type I Error Control

A major reason for the popularity of hypothesis testing is **type I error control**

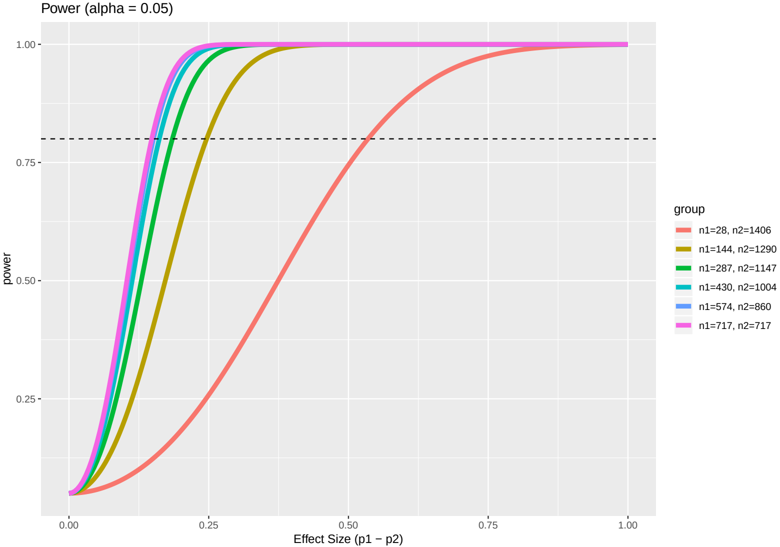
- Using a significance threshold of α limits the *probability of making a type I error* to α
- Imagine 100 hypothesis tests where the null hypotheses are all true
    - Setting $\alpha = 0.05$ would lead to 5 type I errors (on average)
    - Trivially, how could we guarantee a 0 type I errors?
- Type I error rates are controllable because they depend entirely on the null distribution (namely the tail-areas defined by α)
    - Type II error control is difficult, as failing to reject an incorrect null hypothesis depends on what is true in reality

Power

Rather than fixating on type II error control, statisticians instead focus on something called **power**:

- Let the probability of making a type II error be denoted by β
- **Power** is defined as $1 - \beta$, or the probability that we correctly reject a false null hypothesis
- To calculate power, we need to specify an *effect size*, or what we think is true in reality
  - Power also depends upon sample size and α
  - Trivially, How could we guarantee 100% power?

Power Curves



Takeaways

- We use significance thresholds (ie: α) to limit the probability of making a *type I error*
  - This controls the long-run rate of "false positives" in scientific experiments
- *Type II errors* are harder to quantify and we usually talk about *power* instead
  - Power depends upon n, α, and the effect size
  - When designing experiments, we try to achieve a reasonable power without compromising type I error control

Relating Confidence Intervals and Hypothesis Tests

- Suppose the 95% confidence interval for a difference in means is (3.2, 10.1), what do you think the two-sided *p*-value looks like when testing $H_0 : \mu_1 - \mu_2 = 0$?
  - The *p*-value will be *less than* 0.05 because the 95% confidence interval doesn't contain zero (the value specified in the null hypothesis)
- Hypothesis testing is based upon plausible values *when the null hypothesis is true*, while confidence intervals are based upon *plausible values in reality*
  - The variation in these plausible values *depends* on the data itself, it *doesn't depend* on the null hypothesis being true

Relating Confidence Intervals and Hypothesis Tests

- Suppose the hypothesis test for a difference in proportions $H_0 : p_1 - p_2 = 0$ yields a p-value of 0.16, what do you think the 99% confidence interval looks like?
  - The 99% confidence interval *will contain* 0, this is because the p-value is *larger than* 0.01
- Suppose the hypothesis test for a difference in means $H_0 : \mu_1 - \mu_2 = 0$ yields a p-value of 0.004, what do you think the 99% confidence interval looks like?
  - The 99% confidence interval *won't contain* 0, this is because the p-value is *less than* 0.01

Hypothesis Test or Confidence Interval?

In many fields journal publications tend to include statements like:

- “Free prostate specific antigen levels were significantly higher in controls ($p = 0.003$)”
- “The expected cancer incidence was 1.4 per 100,000 (95% CI: 0.7, 2.1)”
- “The rate reduction in the intervention group was 0.9 (CI: 0.86-0.94, $p < 0.001$)”

So, should you report the *p*-value or a confidence interval? Should you report both?

Hypothesis Test or Confidence Interval?

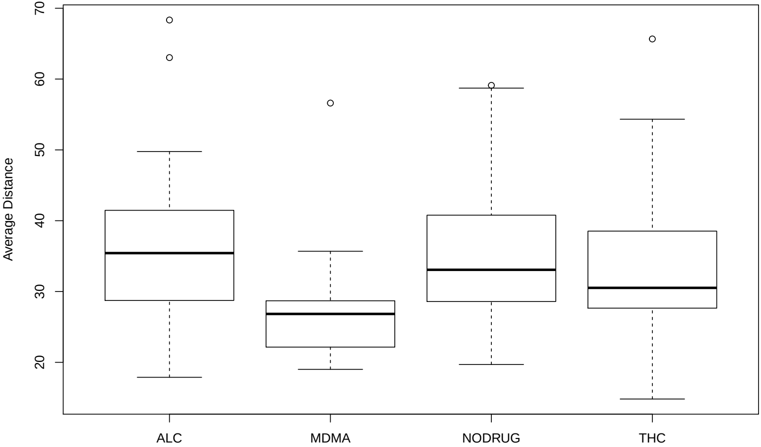
- Use hypothesis testing if you are interested in a specific null value (ie: could $p_1 - p_2 = 0$?)
- Use confidence intervals when you are concerned with estimating an effect (ie: what is the cancer incidence rate for this population?)
- There is no harm in reporting both and letting the reader decide which is more informative
  - Confidence intervals are particularly valuable for non-significant p-values because the reader can themselves decide if the results appear to be due a lacking sample size or due to the lack of an effect

Drug Use and Tailgating

- This example comes from a study done at the National Advanced Driving Simulator (NADS), which attempted to link drug use with risky behavior in other areas (driving)
- In a driving simulator, subjects were told to follow a lead vehicle that was programmed to vary its speed unpredictably
    - As the lead vehicle erratically changed speed, more cautious drivers follow at a larger distance, while riskier drivers tailgate the vehicle
- The study's outcome variable was the average following distance of each participant
- The study's explanatory variable was the participant's drug use group: Alcohol, MDMA, THC, or no drugs used
    - Participants who used multiple drugs were classified according to the "hardest" drug they used (MDMA $>$ THC $>$ Alcohol)

Drug Use and Tailgating

After removing a couple of outliers, here's what the data look like:



Multiple Testing

In these data there are four different groups we'd like to compare, requiring six different hypothesis tests.

1. ALC vs NODRUG, p-value $= 0.5102$
2. ALC vs MDMA, p-value $= 0.00417$
3. ALC vs THC, p-value $= 0.8959$
4. THC vs NODRUG, p-value $= 0.4782$
5. THC vs MDMA, p-value $= 0.01383$
6. MDMA vs NODRUG, p-value $= 0.00216$

If we use the results of these 6 tests (comparing vs. $\alpha = 0.05$), does our experiment still have a 5% Type I error rate?

The Bonferroni Adjustment

The Type I error rate for this *family of tests* is inflated, suppose the null hypothesis is true for all 6 pairwise tests in the tailgating study (and the tests are independent); Then, using $\alpha = 0.05$:

$$\begin{aligned} Pr(\text{At least one type I error}) &= 1 - Pr(\text{No type I errors}) \\ &= 1 - (1 - 0.05)^6 = 26.5\% \end{aligned}$$

This suggests a simple correction to significance threshold: $\alpha^* = \alpha/h$, where h is the number of hypothesis tests being performed. Then:

$$\begin{aligned} Pr(\text{At least one type I error}) &= 1 - Pr(\text{No type I errors}) \\ &= 1 - (1 - 0.05/6)^6 \approx 5\% \end{aligned}$$

The Bonferroni Adjustment

Setting $\alpha^* = \alpha/h$ is known as the **Bonferroni Adjustment**. How many of the six hypotheses can be rejected while still acheiving a family-wise Type I error rate of 5%?

1. ALC vs NODRUG, p-value $= 0.5102$
2. ALC vs MDMA, p-value $= 0.00417$
3. ALC vs THC, p-value $= 0.8959$
4. THC vs NODRUG, p-value $= 0.4782$
5. THC vs MDMA, p-value $= 0.01383$
6. MDMA vs NODRUG, p-value $= 0.00216$

Since $\alpha^* = 0.05/6 = 0.0083$, only two of six tests are now considered “statistically significant”; but we’ve controlled the *family-wise* Type I error rate at 5%.

Bonferroni Adjusted *p*-values

- Occasionally you'll see **adjusted p-values** get reported (rather than an explanation of how to compare the original *p*-values to an adjusted significance threshold)
- For the Bonferroni adjustment this simply entails multiplying each of the original *p*-values by h (the number of tests)
- "Bonferroni Adjusted *p*-values" can then be compared directly with a significance threshold describing the desired Type I error rate
  - For example, you could compare the adjusted *p*-values to 0.05 to achieve a 5% family-wise Type I error rate

Practice

A genetic association study tested for differences in gene expression between two types of leukemia. The study tested 7129 genes.

1. If all 7129 tests were done using $\alpha = 0.01$, and there are no genetic differences between these two types of leukemia, how many "statistically significant" results would you expect?
2. Suppose 783 genes had *p*-values less than 0.01, do you believe there is some association between genes and type of leukemia
3. Suppose you wanted to use the Bonferroni adjustment to ensure a Type I error rate no larger than 5%. What would your adjusted significance threshold be?
4. Suppose the "most significant" gene had a *p*-value of 0.000001, what is its *Bonferroni Adjusted p-value*?

Practice - Solution

1. You'd expect $7129 * 0.01 = 71$ Type I errors
2. Yes, there were over 10 times (712) more significant results than expected
3. $\alpha^* = 0.05/7129 = 0.000007$
4. The adjusted p-value is $0.000001 * 7129$, or $p^* = 0.007$

Conclusion

Right now, you should...

1. Understand the errors that can occur when hypothesis testing
2. Understand statistical power and how it relates to hypothesis testing errors
3. Know the relationship between hypothesis tests and confidence intervals
4. Understand the problems with multiple testing, and how to use the Bonferonni adjustment

These notes cover Sections 4.4 - 4.5 of the textbook, I encourage you to read through those sections and examples