Testing Errors, Power, and Multiple Comparisons

Ryan Miller

Introduction

- Hypothesis testing is an important statistical tool, but it needs to be applied appropriately within the broader investigative process that underlies statistical thinking (and scientific inquiry more generally)
- In this lecture, we will cover a few important considerations, and the vocabulary associated with them, when using hypothesis testing as a decision making tool

Clofibrate

- In 1980, the *New England Journal of Medicine* published results from a randomized, placebo-controlled, double-blind experiment involving the cholesterol-lowering drug *clofibrate*
- Of the subjects randomly assigned to take clofibrate, adherers were defined as those who took more than 80% of their prescribed pills:

	Number	Deaths
Took at least 80%	708	15%
Took less than 80%	357	25%
Total	1103	20%

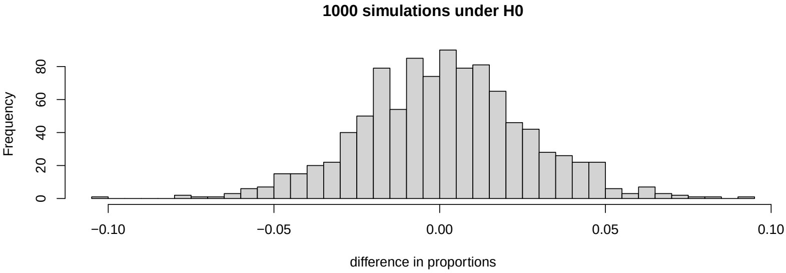
Clofibrate

Is clofibrate effective? We might evaluate the hypothesis:

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

The study observed:

$$\hat{p}_{\text{death}|\text{adherer}} - \hat{p}_{\text{death}|\text{nonadherer}} = 106/708 - 89/357 = -0.10$$



Clofibrate

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

Clofibrate

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

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

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

Clofibrate

- This experiment should have been analyzed using the **intent-to-treat** principle:

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

- The corresponding hypothesis test yields an unconvincing *p*-value of 0.51
  - Using a **significance level** (evidence threshold) of $\alpha = 0.05$, we don't have evidence to refute 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 a high p-value doesn't *prove* the Null Hypothesis)
    - This would mean that our experiment/analysis resulted in a *decision error*
    - Put differently, we failed to reject H_0 because $p \geq \alpha$, but that was a mistake because H_0 was false and should've been rejected

Clofibrate

- Yes, clofibrate *could* be better (remember a high p-value doesn't *prove* the Null Hypothesis)
  - This would mean that our experiment/analysis resulted in a *decision error*
  - Put differently, we failed to reject H_0 because $p \geq \alpha$, but that was a mistake because H_0 was false and should've been rejected
- A second type of error would be incorrectly rejecting a null hypothesis that is actually true
  - Any guesses on the *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

Which type of error might have been made in the clofibrate study?

Practice

Describe (in words) what a Type I and Type II error would be for the following scnarios:

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 have 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 strength of hypothesis testing is **type I error control**

- Setting a significance threshold of α limits the *probability of making a type I error* to α

Type I Error Control

A major strength of hypothesis testing is **type I error control**

- Setting a significance threshold of α limits the *probability of making a type I error* to α
- Imagine 100 different hypothesis tests where the null hypotheses are all true
  - Using $\alpha = 0.05$, you'd expect 5 type I errors (on average)
  - Trivially, how could we guarantee we make zero type I errors?

Type I Error Control

A major strength of hypothesis testing is **type I error control**

- Setting a significance threshold of α limits the *probability of making a type I error* to α
- Imagine 100 different hypothesis tests where the null hypotheses are all true
  - Using $\alpha = 0.05$, you'd expect 5 type I errors (on average)
  - Trivially, how could we guarantee we make zero type I errors?
- Type I error rates are *controllable*, as they depend entirely on the null distribution (namely the tail-areas defined by α)
  - Type II errors are not easily controlled, as they require you to know the *true* effect size (something you're usually trying to *estimate*!)

Power

Rather than fixating on controlling Type II errors, statisticians instead focus on a quantity known as **statistical power**:

- Let β denote the probability of making a Type II error
  - **Power** is defined as $1 - \beta$, meaning it is the probability of *correctly rejecting* a *false null hypothesis*

Power

Rather than fixating on controlling Type II errors, statisticians instead focus on a quantity known as **statistical power**:

- Let β denote the probability of making a Type II error
  - **Power** is defined as $1 - \beta$, meaning it is the probability of *correctly rejecting* a *false null hypothesis*
- Calculating the power of an experiment requires us to specify an *effect size* (usually based upon *clinical significance*)
  - Power also depends upon sample size and α
  - Trivially, how could we guarantee 100% power?

Power Calculations

- Statisticians are often asked to perform *power calculations* to help plan future studies, addressing the question:
  - *What sample size(s) are needed to achieve a certain probability of rejecting a false null hypothesis?*

Power Calculations

- Statisticians are often asked to perform *power calculations* to help plan future studies, addressing the question:
  - *What sample size(s) are needed to achieve a certain probability of rejecting a false null hypothesis?*
- This link is an example of a power calculator for difference in proportions tests
- If the death rates observed in the clofibrate study (0.20 and 0.21) are true at the population level, sample sizes of ~ 25000 in each group are needed to have an 80% chance of rejecting the null hypothesis and detecting this difference!
  - If we're willing to accept a 10% type I error rate, the requirement drops to ~ 20000

Takeaways

- Statisticians use significance thresholds (ie: α) to limit the probability of making a *type I error*
  - These thresholds control the long-run rate of “false positives” in scientific experiments
- *Type II errors* are more complicated, and statisticians usually focus on *power* instead
  - Power depends upon n, α, and the effect size
  - Planning an experiment usually involves calculating the necessary sample size(s) to achieve reasonable power to detect a clinically significant effect without compromising type I error control

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)

Drug Use and Tailgating

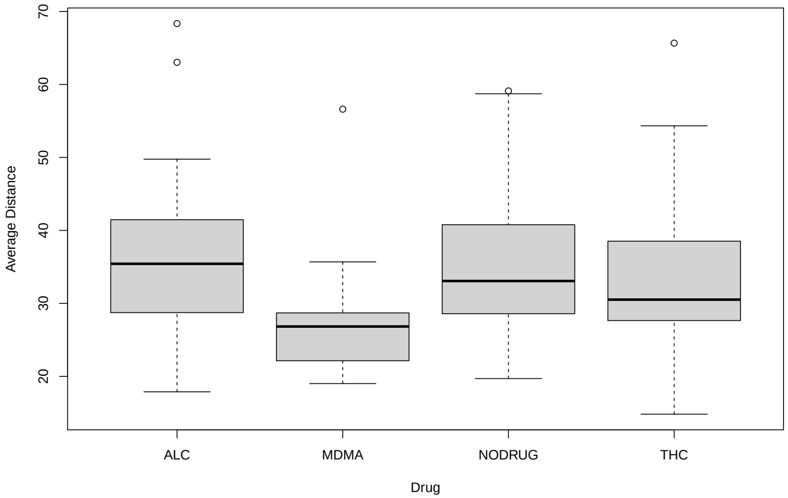
- This example comes from a study done at the National Advanced Driving Simulator (NADS), which attempted to link drug use with risky behavior in other areas (driving)
- In a driving simulator, subjects were told to follow a lead vehicle that was programmed to vary its speed unpredictably
  - As the lead vehicle erratically changed speed, more cautious drivers follow at a larger distance, while riskier drivers tailgate the vehicle

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.

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 all 6 tests (evaluated vs. $\alpha = 0.05$), does this experiment still have a 5% chance of making a Type I error?

The Bonferroni Adjustment

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

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

The Bonferroni Adjustment

The Type I error rate for this *family of tests* is inflated, if the null hypothesis is true for all 6 tests in the tailgating study (and if 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. Now:

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

The Bonferroni Adjustment

Setting $\alpha^* = \alpha/h$ is known as the **Bonferroni Adjustment** (or *Bonferroni Correction*). Now, how many of the six hypotheses can be rejected while still achieving 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

The Bonferroni Adjustment

Setting $\alpha^* = \alpha/h$ is known as the **Bonferroni Adjustment** (or *Bonferroni Correction*). Now, how many of the six hypotheses can be rejected while still achieving 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 likelihood of our *entire analysis* making a Type I error 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

- Hypothesis testing can be considered a decision making tool, but this can lead to errors
  - There is an inherent trade-off between Type I and Type II errors that must be managed by the data analyst

Conclusion

- Hypothesis testing can be considered a decision making tool, but this can lead to errors
    - There is an inherent trade-off between Type I and Type II errors that must be managed by the data analyst
- Performing multiple hypothesis tests within the same experiment can be problematic
    - Taken to the extreme (like genetic association example), it's possible that "significant findings" are more likely to be Type I errors than real discoveries
    - Approaches like the Bonferroni correction can be used to control the *family-wise* Type I error rate of an experiment