Testing Errors and Power

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

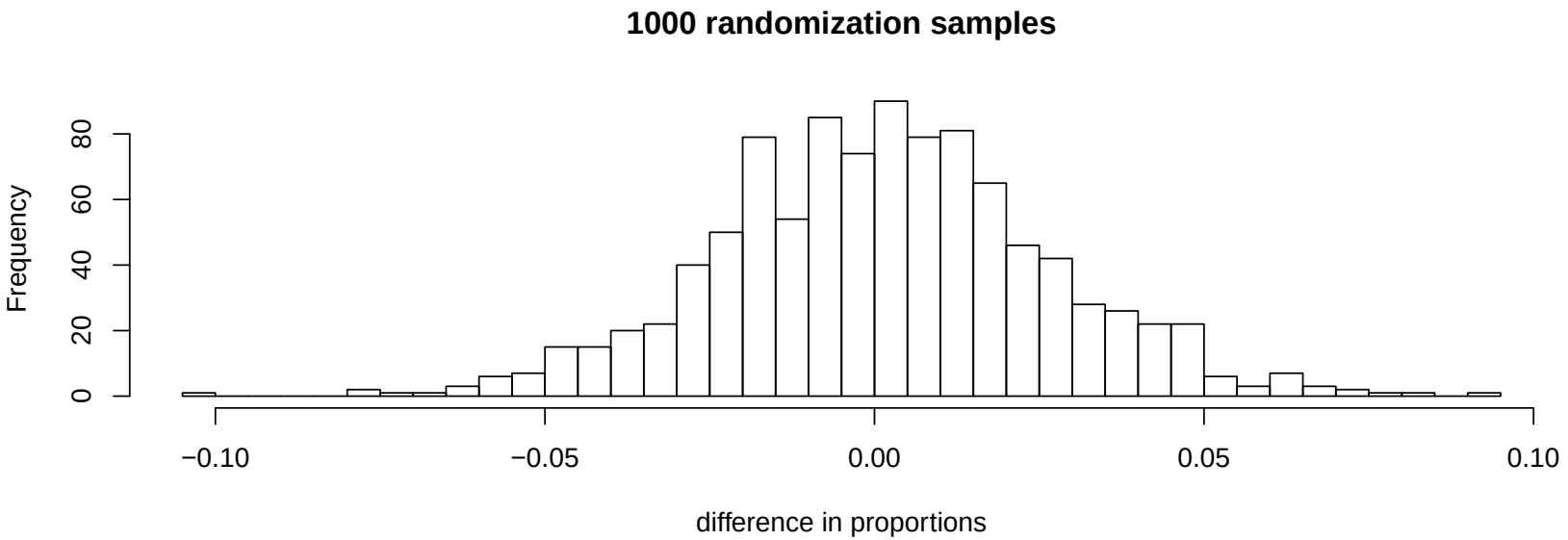
Is clofibrate effective? Let's use hypothesis testing:

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

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 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 $\alpha = 0.05$, we'd fail to reject the null hypothesis of clofibrate and placebo being equally effective
- But is it possible that clofibrate really is better than placebo?

Clofibrate

- Yes, clofibrate *could* be better
  - This would imply that our hypothesis test resulted in an error
  - In other words, we failed to reject H_0 with $p \geq \alpha$, but this was a mistake because H_0 was false and should be rejected
- Another type of error we could make is rejecting a null hypothesis that is actually true
- Any ideas 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 both 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

How could we decrease the chances of making a Type I error?

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**

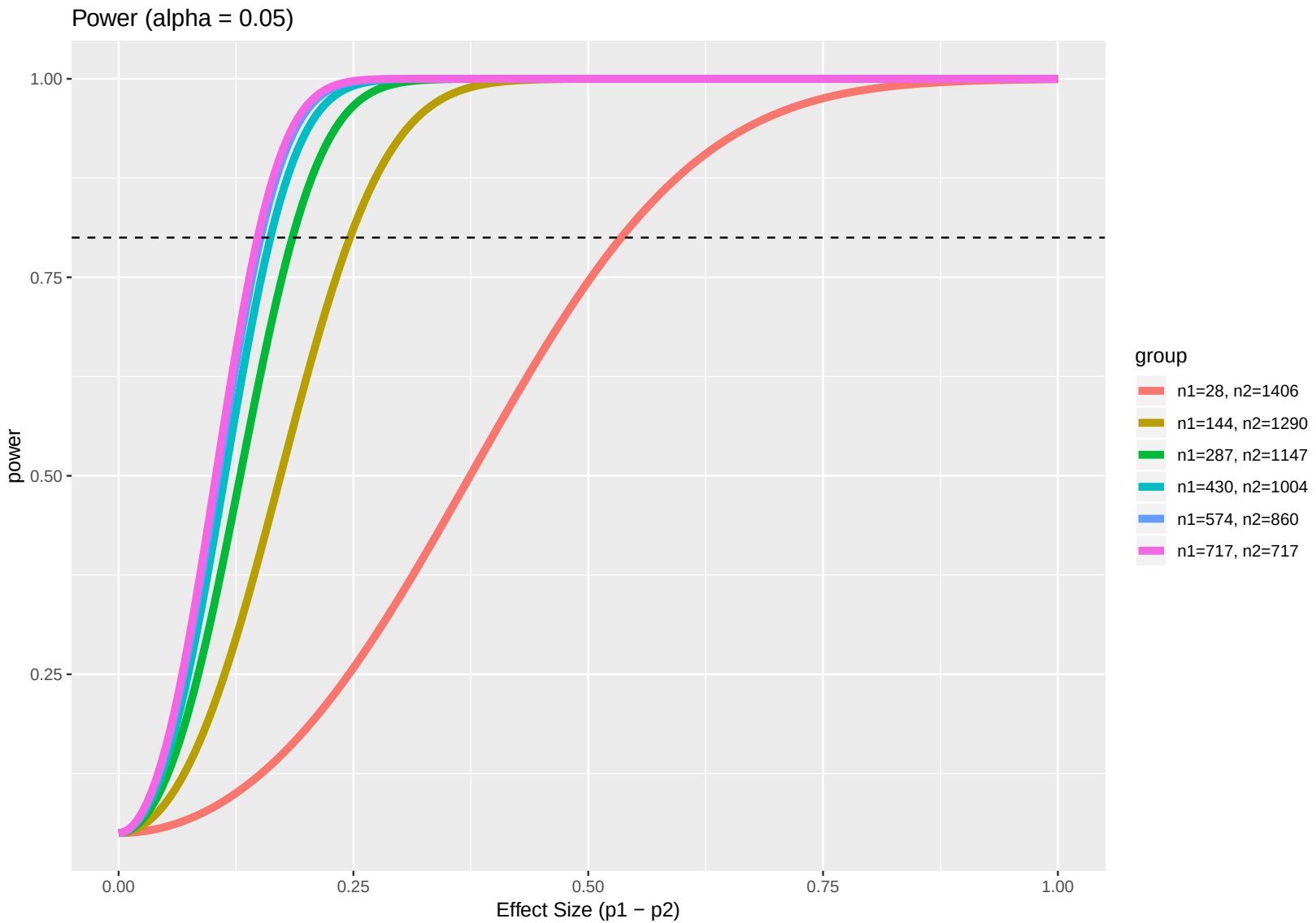
- Using a significance threshold of α limits the *probability of making a type I error* to α
- Imagine 100 hypothesis tests where the null hypothesis is true
  - Setting $\alpha = 0.05$ would lead to 5 type I errors (on average)
  - Trivially, how could we guarantee a 0 type I errors?
- Type I error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

Power

Rather than fixating on type II error control, statisticians use a term called **power**:

- Let the probability of making a type II error be denoted by β
- **Power** is defined as $1 - \beta$, it is 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



Take-aways

- We use significance thresholds to limit the probability of making a *type I error*
  - This controls the long-run proportion of "false positives" in scientific experiments
- *Type II errors* are harder to quantify and we usually talk about *power* instead
  - 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, this is 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, but *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 applied fields, many journal publications include statements like:

- "Free prostate specific antigen levels were significantly higher in controls ($p = 0.001$)"
- "The expected cancer incidence was 1.4 per 100,000 (95% CI: 0.7, 2.1)"

How should whether to report a *p*-value or a confidence interval? Should we report both?

Hypothesis Test or Confidence Interval?

- Use hypothesis testing if you are interested in a particular value (ie: is 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 the population?)
- There is no harm in reporting both and letting the reader choose which is relavant
  - Confidence intervals are particularly valuable for non-significant p-values because the reader can decide if the results are due a lacking sample size or due to the lack of an effect

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

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