

Hypothesis Testing (part 4, decision-making errors)

Ryan Miller

- ▶ The previous presentations have introduced the general framework of hypothesis testing, which includes the p -value as a measure of evidence against a null hypothesis
- ▶ At the heart of hypothesis testing is *decision making*
 - ▶ The goal is to decide whether a particular null model is compatible with the sample data
 - ▶ The smaller the p -value, the higher the degree of incompatibility (suggesting an alternative is more believable)

Decision Thresholds

- ▶ Many scientific fields use $\alpha = 0.05$ as a “significance threshold” for *rejecting* a null hypothesis
- ▶ More generally, we could let α denote a *decision threshold*
 - ▶ If $p\text{-value} \leq \alpha$ we'd reject H_0 in favor of the alternative
 - ▶ If $p\text{-value} > \alpha$ we'd decide there isn't enough evidence to reject H_0

Decision Errors

		The Truth	
		H_0 True	H_0 False
My Decision	Reject H_0	Type I Error	OK
	Fail to Reject H_0	OK	Type II Error

Example #1

- ▶ Consider a jury trial for Person A
 - ▶ H_0 : Person A is not guilty vs. H_A : Person A is guilty
- ▶ In words, what would a Type I and Type II error represent?

Example #1 (solution)

- ▶ A Type I error would mean that Person A is not guilty (H_0 is true), but the jury decides they are guilty (reject H_0)
- ▶ A Type II error would mean that Person A is guilty (H_0 is false), but the jury decides they are not guilty (not enough evidence to reject H_0)

Example #2

- ▶ Consider a clinical trial evaluating a new medication for disease B
 - ▶ H_0 : The medication doesn't cure disease B vs. H_A : The medication cures disease B
- ▶ In words, what would a Type I and Type II error represent?

Example #2 (solution)

- ▶ A Type I error would mean the new medication is not effective (H_0 is true), but the study concludes it cures disease B (reject H_0)
- ▶ A Type II error would mean the new medication cures disease B (H_0 is false), but the study concludes it is ineffective (not enough evidence to reject H_0)

- ▶ By design, using a *decision threshold* of α means the probability of making a Type I error (when H_0 is true) is α
 - ▶ If $\alpha = 0.05$, we'd expect a Type I to occur in 5% of tests where the null hypothesis is true

- ▶ By design, using a *decision threshold* of α means the probability of making a Type I error (when H_0 is true) is α
 - ▶ If $\alpha = 0.05$, we'd expect a Type I to occur in 5% of tests where the null hypothesis is true
- ▶ If we wanted to reduce the rate of Type I errors, we might consider a more stringent threshold of $\alpha = 0.01$
 - ▶ This comes at the expense of making more Type II errors (we've made it harder to reject H_0 , which includes scenarios when H_0 is false)

- ▶ By design, using a *decision threshold* of α means the probability of making a Type I error (when H_0 is true) is α
 - ▶ If $\alpha = 0.05$, we'd expect a Type I to occur in 5% of tests where the null hypothesis is true
- ▶ If we wanted to reduce the rate of Type I errors, we might consider a more stringent threshold of $\alpha = 0.01$
 - ▶ This comes at the expense of making more Type II errors (we've made it harder to reject H_0 , which includes scenarios when H_0 is false)

Error Rates and Study Replication

- ▶ The *decision threshold* of $\alpha = 0.05$ is very widely used because it is thought to balance the rates of Type I and Type II errors
- ▶ While we'd expect a Type I error in 5% of studies, if others are repeating the same research the chance of two independent studies both resulting in a Type I error is very small
 - ▶ $0.05 * 0.05 = 0.0025$ (or 1/400)

Error Rates and Multiple Tests

- ▶ It is important to draw a distinction between testing the same hypothesis in multiple different studies and testing multiple hypotheses in the same study
 - ▶ In the later scenario, a *decision threshold* of $\alpha = 0.05$ can be problematic

Error Rates and Multiple Tests

- ▶ It is important to draw a distinction between testing the same hypothesis in multiple different studies and testing multiple hypotheses in the same study
 - ▶ In the later scenario, a *decision threshold* of $\alpha = 0.05$ can be problematic
- ▶ As an example, consider a genetic association study testing differences in the expression levels of 7129 genes across two patients with two different types of leukemia
 - ▶ This single study involves 7129 different hypothesis tests
 - ▶ If all of the tests used $\alpha = 0.05$, and none of the genes were related to the type of leukemia, we'd expect to see 356 “statistically significant” genes

Error Rates and Multiple Tests

- ▶ It is important to draw a distinction between testing the same hypothesis in multiple different studies and testing multiple hypotheses in the same study
 - ▶ In the later scenario, a *decision threshold* of $\alpha = 0.05$ can be problematic
- ▶ As an example, consider a genetic association study testing differences in the expression levels of 7129 genes across two patients with two different types of leukemia
 - ▶ This single study involves 7129 different hypothesis tests
 - ▶ If all of the tests used $\alpha = 0.05$, and none of the genes were related to the type of leukemia, we'd expect to see 356 "statistically significant" genes
- ▶ As you'd expect, it is wise to use a more stringent significance threshold in this type of study (one involving many different related hypotheses)

The Bonferroni Adjustment

- ▶ A simple fix is to divide the desired Type I error rate (α) by the number of hypothesis tests (h) to get a new significance threshold ($\alpha^* = \alpha/h$)
- ▶ This procedure is known as the “Bonferroni Adjustment” and it will limit the *entire study's* family-wise Type I error rate to $\alpha\%$
 - ▶ In our leukemia example that tested 7129, we might use an adjusted significance threshold of $\alpha^* = 0.05/7139 = 0.00007$ if we wanted to limit the probability of making at least one Type I error to 5%

Conclusion

- ▶ Hypothesis testing is a decision making tool, but it isn't perfect
 - ▶ Type I errors occur when the null hypothesis is *true*, but the data say to *reject it*
 - ▶ Type II errors occur when the null hypothesis is *false*, but the data *do not provide enough evidence to reject it*

Conclusion

- ▶ Hypothesis testing is a decision making tool, but it isn't perfect
 - ▶ Type I errors occur when the null hypothesis is *true*, but the data say to *reject it*
 - ▶ Type II errors occur when the null hypothesis is *false*, but the data *do not provide enough evidence to reject it*
- ▶ The Type I error rate is controlled by the *significance threshold*, α
 - ▶ There is a trade-off between using more/less stringent values of α (lowering α will reduce the chances of making a Type I error but increase the likelihood of making a Type II error)

Conclusion

- ▶ Hypothesis testing is a decision making tool, but it isn't perfect
 - ▶ Type I errors occur when the null hypothesis is *true*, but the data say to *reject it*
 - ▶ Type II errors occur when the null hypothesis is *false*, but the data *do not provide enough evidence to reject it*
- ▶ The Type I error rate is controlled by the *significance threshold*, α
 - ▶ There is a trade-off between using more/less stringent values of α (lowering α will reduce the chances of making a Type I error but increase the likelihood of making a Type II error)
- ▶ Performing a large number of hypothesis tests within a single study can be problematic