Hypothesis Testing

Part 2 - Decision errors and multiple tests

Ryan Miller

Introduction

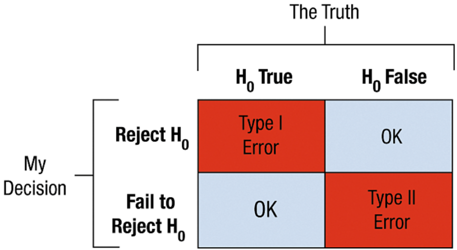
Statistical tests adopt the following framework:

1. Researchers form a null hypothesis that they seek evidence against
2. The compatibility of the observed data and null hypothesis is assessed using the p-value
3. If the p-value provides sufficient evidence against the null hypothesis the researcher will reject it in favor of an alternative

The culmination of this framework is the decision to either reject the null hypothesis or conclude that data provides insufficient evidence against it.

Decision Errors

Because we don't know the true status of the null hypothesis, any decision from a hypothesis test might be correct or incorrect:



		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

- 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*

Trade-offs

- Recall that α is a threshold used to determine "statistical significance", with $\alpha = 0.05$ being a widely used threshold
    - Using $\alpha = 0.05$ we can expect a Type I error in 5% of instances where H_0 is true
    - What could we do to reduce the rate of Type I errors? How would this impact the chances of making a Type II error?

Trade-offs

- Recall that α is a threshold used to determine "statistical significance", with $\alpha = 0.05$ being a widely used threshold
    - Using $\alpha = 0.05$ we can expect a Type I error in 5% of instances where H_0 is true
    - What could we do to reduce the rate of Type I errors? How would this impact the chances of making a Type II error?
- Setting a stricter criteria for statistical significance by reducing α decreases the chances of making a Type I error, but it increases the chances of making a Type II error

Error Rates

Suppose a large number of hypotheses are tested and the results are recorded in the table below:

	True Null Hypothesis	False Null Hypothesis
Fail to Reject Null	a	b
Reject Null	c	d

- The **Type I error rate** is defined as $\frac{c}{a+c}$, or the fraction of true null hypotheses that are incorrectly rejected
- The **Type II error rate** is defined as $\frac{b}{b+d}$, or the fraction of false null hypotheses that were not rejected
  - The complement of the Type II error rate, or $\frac{d}{b+d}$, is known as the testing procedure's **statistical power**
- The **false discovery rate** is defined as $\frac{c}{c+d}$, or the fraction of rejected hypotheses that were incorrectly rejected

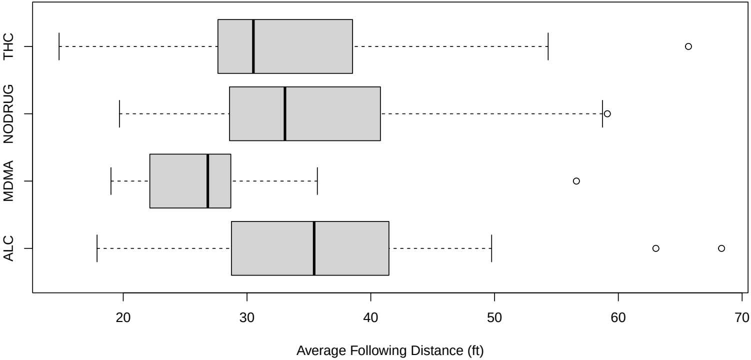
Multiple Hypothesis Tests

Many scientific studies involve multiple hypotheses, an example is presented below:

- The NADS organization looked at the relationship between drug use and tailgating behavior while driving
- They classified participants into 4 groups according to the "hardest" substance they regularly used (No Drug, Alcohol, THC, or MDMA)
- These participants then drove a simulated route in an advanced driving simulator, and the researchers recorded their average following distance behind a lead vehicle as one of the study's outcomes

Multiple Hypothesis Tests

If the researchers wanted to compare the mean following distances in each drug use group, how many different tests would they need to perform?



Grinnell College
Statistics

Multiple Hypothesis Tests

Since there are 4 different groups we'd like to compare, 6 different hypothesis tests are possible:

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 compare each test's p-value against $\alpha = 0.05$, will the *entire set of conclusions* from this experiment (as a whole) still have a 5% Type I error rate?

The Bonferroni Adjustment

If the null hypothesis is true for all 6 pairwise tests, and the tests are independent, 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

If the null hypothesis is true for all 6 pairwise tests, and the tests are independent, 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**. If we apply this correction, how many of the 6 hypotheses can be rejected with 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 $= 0.00216$

Using $\alpha^* = 0.05/6 = 0.0083$ only 2 of 6 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 a study report **adjusted p-values**
- For the Bonferroni adjustment, these are found by multiplying the original *p*-values by h (the number of tests)
- "Bonferroni Adjusted *p*-values" can then be compared directly to the target family-wise Type I error rate
    - For example, comparing the adjusted *p*-values against 0.05 will achieve a 5% family-wise Type I error rate

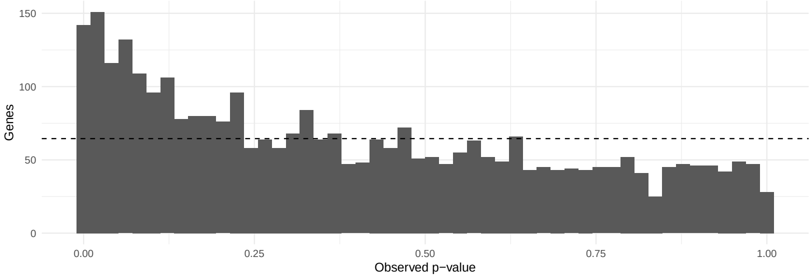
Practice

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

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" genes would be expected?
2) Suppose 783 genes had p-values less than 0.01, do you believe there is association between some 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*?

False Discovery Rates vs. Type I Error Control

A genomics study measured the expression levels of 17,322 genes to identify genes that are co-expressed with BRCA1, a gene that is well-known to be associated with breast cancer. For each gene a hypothesis test was performed, and the *p*-values of these tests are displayed using a histogram:



False Discovery Rates vs. Type I Error Control

- Suppose we apply the Bonferroni adjustment to control the family-wise type I error rate at 10%
  - $\alpha^* = 0.1/3226 = 0.00003$
  - The study yields 2 statistically significant genes (with *p*-values less than 0.00003)
- Suppose we seek to control the false discovery rate at 10%
  - This isn't as easy to do ourselves, but there are 24 genes that can be selected
  - Among these 24 genes we'd expect 2 or 3 to be false positives
- This example illustrates the overly stringent nature of the Bonferroni adjustment

Conclusion

- Hypothesis tests provide a tool for making a decision about some aspect of a population, such as deciding whether two variables are associated
  - The truth about the population is unknown, so we'll never know for certain if a hypothesis test leads to the correct conclusion
  - Fortunately, setting a significance threshold (such as $\alpha = 0.05$) will limit the chances of making a Type 1 Error (to 5%, for example)
- When multiple hypothesis tests are conducted on the same data we should be mindful that each individual test has its own chance of producing an incorrect conclusion
  - The Bonferroni Adjustment and false discovery rate control methods are useful in these scenarios, particularly when the number of tests is very large (ie: hundreds or thousands)