



Week 5 Video 1

Relationship Mining

Correlation Mining

Relationship Mining

- Discover relationships between variables in a data set with many variables
- Many types of relationship mining

Correlation Mining

- Perhaps the simplest form of relationship mining
- Finding substantial linear correlations between variables
 - ▣ Remember this from earlier in the class?
- In a large set of variables

Use Cases

- You have 100 variables, and you want to know how each one correlates to a variable of interest
 - ▣ Not quite the same as building a prediction model
- You have 100 variables, and you want to know how they correlate to each other

Many Uses...

- Studying relationships between questionnaires on traditional motivational constructs (goal orientation, grit, interest) and student reasons for taking a MOOC
- Correlating features of the design of mathematics problems to a range of outcome measures
- Correlating features of schools to a range of outcome measures

The Problem

- You run 100 correlations (or 10,000 correlations)
- 9 of them come up statistically significant
- Which ones can you “trust”?

If you...

- Set $p=0.05$
- Then, assuming just random noise
- 5% of your correlations will still turn up statistically significant

The Problem

- Comes from the paradigm of conducting a single statistical significance test

The Solution

- Adjust for the probability that your results are due to chance, using a *post-hoc control*

Two paradigms

- **FWER** – Familywise Error Rate
 - ▣ Control for the probability that any of your tests are falsely claimed to be significant (Type I Error)
- **FDR** – False Discovery Rate
 - ▣ Control for the overall rate of false discoveries

Bonferroni Correction

- The classic approach to FWER correction is the Bonferroni Correction



Bonferroni Correction

- Ironically, derived by Miller rather than Bonferroni

Bonferroni Correction

- Ironically, derived by Miller rather than Bonferroni
- Also ironically, there appear to be no pictures of Miller on the internet

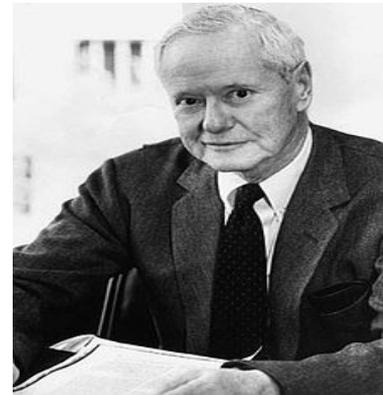
Bonferroni Correction

- A classic example of Stigler's Law of Eponymy
 - “No scientific discovery is named after its original discoverer”



Bonferroni Correction

- A classic example of Stigler's Law of Eponymy
 - ▣ “No scientific discovery is named after its original discoverer”
 - ▣ Stigler's Law of Eponymy was proposed by Robert Merton



Bonferroni Correction

- If you are conducting n different statistical tests on the same data set
- Adjust your significance criterion α to be
 - ▣ α / n
- E.g. For 4 statistical tests, use statistical significance criterion of 0.0125 rather than 0.05

Bonferroni Correction: Example

- Five tests
 - ▣ $p=0.04$, $p=0.12$, $p=0.18$, $p=0.33$, $p=0.55$
- Five corrections
 - ▣ All p compared to $\alpha = 0.01$
 - ▣ None significant anymore
 - ▣ $p=0.04$ seen as being due to chance

Bonferroni Correction: Example

- Five tests
 - ▣ $p=0.04$, $p=0.12$, $p=0.18$, $p=0.33$, $p=0.55$
- Five corrections
 - ▣ All p compared to $\alpha=0.01$
 - ▣ None significant anymore
 - ▣ $p=0.04$ seen as being due to chance
 - ▣ Does this seem right?

Bonferroni Correction: Example

- Five tests
 - ▣ $p=0.001$, $p=0.011$, $p=0.02$, $p=0.03$, $p=0.04$
- Five corrections
 - ▣ All p compared to $\alpha = 0.01$
 - ▣ Only $p=0.001$ still significant

Bonferroni Correction: Example

- Five tests
 - ▣ $p=0.001$, $p=0.011$, $p=0.02$, $p=0.03$, $p=0.04$
- Five corrections
 - ▣ All p compared to $\alpha = 0.01$
 - ▣ Only $p=0.001$ still significant
 - ▣ Does this seem right?

Quiz

- If you run 100 tests, which of the following are statistically significant?
 - A) 0.05
 - B) 0.01
 - C) 0.005
 - D) 0.001
 - E) All of the Above
 - F) None of the Above

Bonferroni Correction

- Advantages
 - You can be “certain” that an effect is real if it makes it through this correction
 - Does not assume tests are independent
 - In our “100 correlations with the same variable” case, they aren’t!
- Disadvantages
 - Massively over-conservative
 - Throws out everything if you run a lot of correlations

Often attacked these days

- Arguments for rejecting the sequential **Bonferroni** in ecological studies. MD Moran - Oikos, 2003 - JSTOR
- Beyond **Bonferroni**: less conservative analyses for conservation genetics. SR Narum - Conservation Genetics, 2006 – Springer
- What's wrong with **Bonferroni** adjustments. TV Perneger - Bmj, 1998 - bmj.com
- p Value fetishism and use of the **Bonferroni** adjustment. JF Morgan - Evidence Based Mental Health, 2007

There are FWER corrections that are a little less conservative...

- Holm Correction/Holm's Step-Down (Toothaker, 1991)
- Tukey's HSD (Honestly Significant Difference)
- Sidak Correction

- Still generally very conservative
- Lead to discarding results that probably should not be discarded

FDR Correction

- (Benjamini & Hochberg, 1995)



FDR Correction

- Different paradigm, arguably a better match to the original conception of statistical significance

Statistical significance

- $p < 0.05$
- A test is treated as rejecting the null hypothesis if there is a probability of under 5% that the results could have occurred if there were only random events going on
- This paradigm accepts from the beginning that we will accept junk (e.g. Type I error) 5% of the time

FWER Correction

- $p < 0.05$
- Each test is treated as rejecting the null hypothesis if there is a probability of under 5% divided by N that the results could have occurred if there were only random events going on
- This paradigm accepts junk far less than 5% of the time

FDR Correction

- $p < 0.05$
- Across tests, we will attempt to accept junk exactly 5% of the time
 - ▣ Same degree of conservatism as the original conception of statistical significance

FDR Procedure (Benjamini & Hochberg, 1995)

- Order your n tests from most significant (lowest p) to least significant (highest p)
 - Test your first test according to significance criterion $\alpha * 1 / n$
 - Test your second test according to significance criterion $\alpha * 2 / n$
 - Test your third test according to significance criterion $\alpha * 3 / n$
 - Quit as soon as a test is not significant

FDR Correction: Example

- Five tests
 - $p=0.001$, $p=0.011$, $p=0.02$, $p=0.03$, $p=0.04$

FDR Correction: Example

- Five tests
 - $p=0.001$, $p=0.011$, $p=0.02$, $p=0.03$, $p=0.04$
- First correction
 - $p = 0.001$ compared to $\alpha = 0.01$
 - Still significant!

FDR Correction: Example

- Five tests
 - $p=0.001$, $p=0.011$, $p=0.02$, $p=0.03$, $p=0.04$
- Second correction
 - $p = 0.011$ compared to $\alpha = 0.02$
 - Still significant!

FDR Correction: Example

- Five tests
 - $p=0.001$, $p=0.011$, $p=0.02$, $p=0.03$, $p=0.04$
- Third correction
 - $p = 0.02$ compared to $\alpha = 0.03$
 - Still significant!

FDR Correction: Example

- Five tests
 - $p=0.001, p=0.011, p=0.02, p=0.03, p=0.04$

- Fourth correction
 - $p = 0.03$ compared to $\alpha = 0.04$
 - Still significant!

FDR Correction: Example

- Five tests
 - $p=0.001, p=0.011, p=0.02, p=0.03, p=0.04$
- Fifth correction
 - $p = 0.04$ compared to $\alpha = 0.05$
 - Still significant!

FDR Correction: Example

- Five tests

- $p=0.04$, $p=0.12$, $p=0.18$, $p=0.33$, $p=0.55$

FDR Correction: Example

- Five tests
 - $p=0.04$, $p=0.12$, $p=0.18$, $p=0.33$, $p=0.55$
- First correction
 - $p = 0.04$ compared to $\alpha = 0.01$
 - Not significant; stop

Conservatism

- Much less conservative than Bonferroni Correction
- Much more conservative than just accepting $p < 0.05$, no matter how many tests are run

q value extension in FDR (Storey, 2002)

q value extension in FDR (Storey, 2002)

- p = probability that the results could have occurred if there were only random events going on
- q = probability that the current test is a false discovery, given the post-hoc adjustment

q value extension in FDR (Storey, 2002)

- q can actually be lower than p
- In the relatively unusual case where there are many statistically significant results

Closing thought

- Correlation mining can be a powerful way to see what factors are mathematically associated with each other
- Important to get the right level of conservatism

Next lecture

- Causal mining