Multiple Testing

Adjusted *p*-values, False Discovery Rate (FDR), overfitting, mitigation strategies

“Torture the data long enough, and it will confess.“

Nagiza F. Samatova, samatova@csc.ncsu.edu
Professor, Department of Computer Science
North Carolina State University

The Key Message

If you look at the data through enough different perspectives, and ask enough questions, you can almost invariably find a statistically significant effect.

Multiple Testing: Examples

Un-supervised Statistical Learning Tasks

- **Comparing multiple treatment groups, you end up asking multiple questions:**
  - Is A different from B?
  - Is B different from C?
  - Is A different from C?
- **Studies with a treatment evaluated at multiple stages (e.g., clinical trials)**
  - You end up asking multiple questions as you follow each stage of the treatment

By asking multiple questions, with each question, you are increasing the chance of being fooled by chance.

Illustrative Example

- **Data**
  - 20 predictor variables
  - 1 outcome (response) variable
  - All are *randomly* generated

Question:

What are the odds that at least one predictor will (falsely) turn out to be statistically significant (Type I Error)?

- **if you do a series of 20 tests**
- **at the α=0.05 significance level**

Answer:

1. **Calculate the probability that all will correctly test non-significant at the 0.05 level**

$$P(\text{all are non-significant}) = 0.95 \times 0.95 \times \cdots \times 0.95 = (0.95)^{20} = 0.36$$

2. **Calculate the probability that at least one predictor will (falsely) test significant**

$$P(\text{at least one is significant}) = 1 - P(\text{all are non-significant}) = 0.64$$

Overfitting and How to Mitigate it?

- **The issue is related to the problem of overfitting**
  - fitting the model to the noise
- **Punchline**
  - The more variables you add, or the more models you run, the greater the probability that something will emerge as significant just by chance
- **How to mitigate or deal with this problem?**
  - Supervised learning tasks:
    - **Cross-validation (hold-out set): models are assessed on data that the model has not seen before**
  - Un-supervised statistical learning tasks
    - **Adjustment procedures:**
      - Dividing up the α according to the number of tests:
        - This results in smaller α (i.e., more stringent bar for statistical significance) for each test
        - For a 20-variable assessment (if original $\alpha = 0.05$)): $\alpha_{new} = \frac{0.05}{20} = 0.0025$
      - Bonferroni adjustment
        - Dividing up the α according to the number of observations, n

Illustrative Example: Adjusted α

- **Data**
  - 20 predictor variables
  - 1 outcome (response) variable
  - All are *randomly* generated

Question:

What are the odds that at least one predictor will (falsely) turn out to be statistically significant (Type I Error)?

- if you do a series of 20 tests
- at the $\alpha_{new} = 0.0025$ significance level

Answer:

1. Calculate the probability that all will correctly test non-significant at the 0.05 level

$$P(\text{all are non-significant}) = (1 - 0.0025)^{20} = 0.95$$

2. Calculate the probability that at least one predictor will (falsely) test significant

$$P(\text{at least one is significant}) = 1 - P(\text{all are non-significant}) = 0.05$$

Multiplicity Issues:

Multiple comparisons, many variables, many models, etc.

- **Checking for multiple pairwise differences across groups**
- **Looking at multiple subgroup results:**
  - we found no significant treatment effect overall, but we find an effect for unmarried woman younger than 20
- **Trying lots of statistical models**
- **Including lots of variables in models**
- **Asking a number of different questions (i.e., different possible outcomes)**

Summary: Key Ideas and Concepts

- **Sources of multiplicity issues:**
  - Multiple comparisons in multiple tests of significance
  - Many variables
  - Many models
- **Multiplicity increases the risk of concluding that something is significant just by chance (Type I error)**
- **Mitigation strategies for multiple statistical comparisons**
  - Adjustment procedure: dividing alpha by the number of tests
  - Bonferroni adjustment: dividing alpha by the number of observations, *n*
- **Mitigation strategy for supervised modeling**
  - Cross-validation: holdout sample with labeled outcome variables

Multiple Testing

Term	Definition	Examples/Comments
Type I Error	Mistakenly concluding that an effect is statistically significant	
False Discovery Rate (FDR)	Across multiple tests, the rate of making Type I error	
Adjustment of p-value	Accounting for doing multiple tests on the same data	
Overfitting	Fitting the noise	