Chapter 24: Multiple Testing

TXCL7565/PHSC7565

What This Lecture Covers

- What multiple testing is and why it is a problem
- Where does multiple testing arise
- Methods to avoid false positives
- False discovery rate

WHAT IS MULTIPLE TESTING AND WHY IS IT A PROBLEM

What is multiple testing?

A single statistical test carries a 5% risk of producing a false positive conclusion - generally considered an acceptable level of risk.

Repeated tests rack up a much greater (and ultimately unacceptable) risk.

Multiple testing example

- 10 index cards
  - 8 cards have simple words of encouragement
  - 2 card says STARBUCKS and can be turned in for a $5 Starbucks gift card.

- What is the probability that you get a Starbucks gift card?
- If everyone in the class draws one card (with replacement), how many Starbucks cards will I have to give away?

Why is multiple testing a problem?

- If we test 20 hypotheses where there is truly no difference, we expect that at one test will have ‘significant’ p-value ($p<0.05$), e.g., a false positive.

- Likewise, the probability of at least one false positive if we perform even 5 tests is:

$$\begin{aligned}
\text{Probability of at least one mistake} &= 1 - \text{Probability of no mistakes} \\
&= 1 - Pr(\text{no mistake})^5 \\
&= 1 - (0.95)^5 \\
&= 1 - 0.77 \\
&= 0.22
\end{aligned}$$

Why is multiple testing a problem?

- Taken to extremes, multiple testing is virtually guaranteed to find ‘statistical significance’ even in the absence of any real effects.

- Probability of at least one mistake out of 100 tests:

$$\begin{aligned}
\text{Probability of at least one mistake} &= 1 - \text{Probability of no mistakes} \\
&= 1 - Pr(\text{no mistake})^{100} \\
&= 1 - (0.95)^{100} \\
&= 1 - 0.0059 \\
&= 0.9941
\end{aligned}$$

WHERE DOES MULTIPLE TESTING ARISE

Where does multiple testing arise?

- Comparing several treatment groups
- Recording numerous endpoints and testing each one for changes
- Measuring the same end point on several occasions and testing at each time point
- Breaking the data into numerous sub-sets and testing within each of these

Comparing several treatment groups

- If we assess several possible treatments, it is tempting to make all pairwise comparisons.
- As the number of groups increases, the number of comparisons increases dramatically, e.g.,
  - 3 groups = 3 comparisons
  - 4 groups = 6 comparisons
  - 10 groups = 45 comparisons

Comparing several endpoints

- It is tempting to measure lots of variables in a study and testing each outcome compared to our intervention.

- This is especially true in omics studies where the number of endpoints can range from a few hundred to over a hundred thousand.

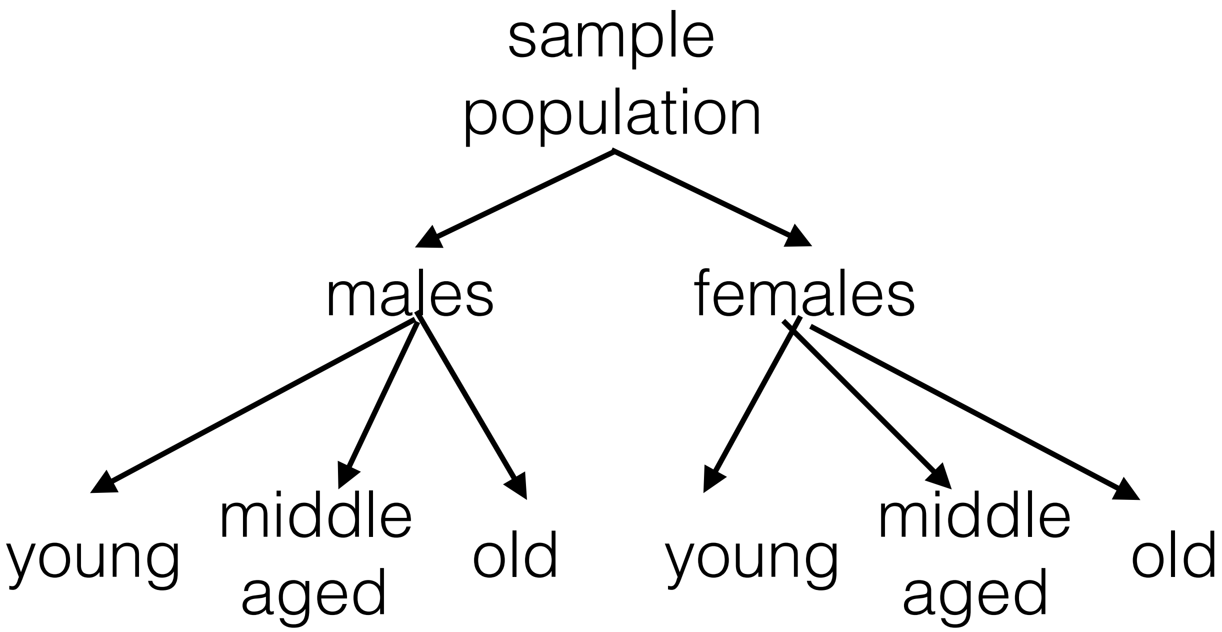
- This also applies to multiple predictors of the same outcome, e.g., genome-wide association studies

Testing at several time points

- The relevant endpoint is often measured at several time points and it is tempting to test for treatment/ intervention differences as each time point.

Testing within numerous groups

- It is tempting when results are not statistically significant in the total sample population to begin to break the sample populations into subgroups and check for statistical significance.
- But is becomes a slippery slope…



METHODS TO AVOID FALSE POSITIVES

Use a single (‘omnibus’) test to avoid a series of pairwise comparisons

- *Multiple treatments* - e.g., ANOVA, chi-square with more than two groups, multiple regression, …

- *Multiple endpoints* - multivariate statistics can handle multiple endpoints and allow omnibus testing

- *Multiple time points* - mixed linear models/repeated measures model multiple time points and allow omnibus testing

Bonferroni correction

- Bonferroni correction is used to control the type 1 errors at 5% (i.e., family-wise error rate)

- Bonferroni raises the standard of proof for all the individual tests.

- Critical P = 0.05/number of test

- Or, the each p-value can be multiplied by the number of tests.

- BUT, this type of correction reduces the power and is only appropriate if each test is independent of the other tests

Distinction between primary and secondary (exploratory) analyses

Another way to avoid the hazards of multiple testing is to highlight one particular route through the experimentation and data analysis.

- What is the primary question being asked?
- What will be the primary endpoint that answers that question?
- What will be the primary statistical analysis of that endpoint?

The identification of a primary analysis must be finalized before the experimental data has been seen.

Secondary (exploratory) analyses

Hypothesis generating vs. hypothesis testing

- Not really answers to questions, rather they provide guidance as to what might be useful further research.

“Enjoy the result you have found by exploratory data analysis, for you will not find it again.” — Stephen Senn

Look for patterns of significant results

In situations with multiple tests with no multiple testing correction:

- 2 out of 30 tests are ‘marginally’ significant = likely no true differences

- 25 out of 30 tests are significant = likely some of the tests are correct

- Do the detected differences agree with other known biology and research results?

FALSE DISCOVERY RATE

False Discovery Rate

- Many times for genetic studies, we use a false discovery rate (FDR) rather than a traditional p-value to help account for multiple comparisons.

- FDR is the estimated proportion of “significant” tests that are false positives.

- An FDR value is calculated for each test (e.g., gene), but it is dependent on the distribution of the other test results (e.g., other genes).

- When we use a 5% FDR threshold for significance, we are estimating that 5% of the significant genes are false positives.

What did we learn?

- Unless special precautions are used, multiple testing will increase the risk of generating false positive findings beyond the level of 5% that is normally tolerated.

- Where there are several treatments compared against each other, the first step in analysis should be an **omnibus test** and if this proves significant, more detailed analyses can be undertaken.

- Data should not be broken down into multiple subgroups unless either a Bonferroni correction or distinction between primary and secondary analyses is used to provide additional protection.

- **Bonferroni correction** maintains the overall risk of a false positive in any test at 5% at the cost of power.

- It is legitimate to declare a primary endpoint/analysis and then use secondary (exploratory) analyses to generate additional hypotheses.