Multiple Hypothesis Testing

- In a study we are currently planning, we will be measuring various biomarkers in pregnant women at 6 weeks of pregnancy
  - Compare the values of the biomarkers between those who develop preeclampsia and those who don’t
  - Idea is to be able to develop a test for early prediction of risk of preeclampsia
- Study design seems fairly straightforward
  - Measure the biomarker values in each of our ~40 biomarkers
  - Compare the values between those who develop preeclampsia and those who don’t, and compute the p-value for each
  - Biomarkers that have a p=0.05 would be considered predictive
- What is the problem with this approach?

Controlling the family-wise error rate

- Suppose you are performing n independent tests
- Choose a significance level, $\alpha$ (usually 0.05)
  - We will aim for the probability of any false positives to be $\alpha$, assuming all null hypotheses are true
- Divide $\alpha$ by the number of comparisons, n
- For each individual test, compute the p-value
- The p-value is considered significant if $p<\alpha/n$
- This is called a Bonferroni correction

Approaches to dealing with multiple comparisons

- Here are some approaches to dealing with multiple hypothesis testing:
  - This may be ok if
    - You report all your p-values and let the reader interpret them, or
    - You have one or two “main” tests and the others are related to them
  - Can be done if the number of hypotheses is very large
- Set a significance level so that the chance of any false positives is 0.05, assuming all null hypotheses are true
  - Called a family-wise error rate
  - Set a target false discovery rate instead of a false positive rate
  - Can be done if the number of hypotheses is very large

p-values and multiple hypotheses

- Remember, the p-value is the probability of a false positive, assuming the null hypothesis is true
- Suppose we do 40 such tests, and the null hypotheses are true
  - i.e. none of our biomarkers are predictive
  - We would hope to have no statistically significant results (i.e. all true negatives)
- For each test, the chance of a true negative is 0.95 (with $\alpha=0.05$)
- If we assume all biomarkers are independent, the chance of all true negatives is
  $0.95 \times 0.95 \times \ldots \times 0.95 = 0.129$
  - In other words, the chances of at least one false positive would be $0.871$, or 87.1%.

Preeclampsia Example with Bonferroni Correction

- In the preeclampsia example, we have 40 tests, so to establish a family-wise error rate of $\alpha=0.05$, we would calculate $\alpha/40=0.00125$
- So each individual biomarker would be considered statistically significant if the p-value for that biomarker is less than 0.00125.
  - Note this will increase the type 2 error rate (number of false negatives)
High throughput gene expression experiments
- (microarray or RNA-Seq), comparisons between the expression of all genes are made between two (or more) different conditions
  - expression changes after treatment of a drug, etc.
- Effectively performing tens of thousands of different tests simultaneously
  - One for each gene in the genome

Bonferroni correction for high-throughput experiments
- Bonferroni correction is not appropriate for high-throughput experiments
  - individual gene would be ~0.05/20000=0.0000025
  - The expression of each gene is not independent, so the Bonferroni correction is overly conservative
  - In many of these experiments, some false positives can be tolerated

Controlling the False Discovery Rate in high-throughput experiments
- An alternative approach is to control the False Discovery Rate (FDR)
- A method devised by Benjamini and Hochberg in 1995 allows us to do this
- Idea is to choose a target FDR
  - Usually called Q
- If Q is 0.1, then 10% of the genes we consider "significant" would be false positives, and 90% would be true positives
  - Of course, you don’t know which!
  - This is an estimate

The Benjamini-Hochberg procedure
- The idea behind the Benjamini-Hochberg procedure is that if all null hypotheses were true, the p-values would be evenly distributed between 0 and 1
- Suppose we are performing n tests, and we choose a target FDR of Q.
- Compute the p-values for each of the tests, and order them with the smallest first. For each p-value, call it’s position in the order (it’s rank) i.
- We reject the null hypothesis for the largest p-value with ip/n < Q, and all smaller p-values.
  - sometimes termed “discoveries”

Ways to use the FDR
- When computing the False Discovery Rate like this, there are three different ways to use it
  - Choose a target FDR (say 5% or 10%) as above, and then see which tests are discoveries within this FDR
  - Choose some number of tests which you consider important (e.g. the 200 genes with the biggest fold change). Compute the FDR which would make all these tests discoveries.
  - For each test, compute the FDR which would just include that test as a discovery. This assigns a FDR for each individual test – this is usually called a q-value for that test. Report the q-values for all tests.

Multiple Testing Example
- Example
- Tested associations between 25 dietary variables and mammographic density (risk factor in breast cancer)

<table>
<thead>
<tr>
<th>Dietary Variable</th>
<th>p-value</th>
<th>q-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Carbohydrates</td>
<td>0.094</td>
<td>0.094</td>
</tr>
<tr>
<td>Total meat</td>
<td>0.042</td>
<td>0.042</td>
</tr>
<tr>
<td>Blue fish</td>
<td>0.039</td>
<td>0.039</td>
</tr>
<tr>
<td>Legumes</td>
<td>0.004</td>
<td>0.004</td>
</tr>
<tr>
<td>Eggs</td>
<td>0.009</td>
<td>0.009</td>
</tr>
<tr>
<td>Diet</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td>Carbohydrates</td>
<td>0.005</td>
<td>0.005</td>
</tr>
<tr>
<td>Total meat</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>Blue fish</td>
<td>0.012</td>
<td>0.012</td>
</tr>
<tr>
<td>Legumes</td>
<td>0.003</td>
<td>0.003</td>
</tr>
<tr>
<td>Eggs</td>
<td>0.008</td>
<td>0.008</td>
</tr>
<tr>
<td>Diet</td>
<td>0.001</td>
<td>0.001</td>
</tr>
</tbody>
</table>
Bonferroni Correction for García-Arenzana data

- To use a Bonferroni Correction for these data, and setting a family-wise error rate (FWER) of 0.05, we would look for p-values less than 0.05/25=0.002
- Only one variable, total calories, would be considered significant for a FWER of 0.05

Benjamini-Hochberg Procedure for García-Arenzana data

- If we choose a FDR of 25%, we find five of the variables are "discoveries".
  - Recall, the interpretation of the FDR is that we estimate 25% of these are false positives
- At a FDR of 20%, only two are considered discoveries
- Demo in Excel.

p-values and high-throughput experiments

- When dealing with very large numbers of tests, individual p-values become almost meaningless
- In a genome-wide experiment, testing 20,000 genes, even with a significance level of 0.001, we would expect 20 false positives on average
  - But we have little control over the distribution of the number of false positives
  - How likely is it we have 5 false positives? 50? 200?
  - Given this, how do you interpret a p-value for an individual test (gene)?
- In experiments such as these the p-value tells us very little
  - The FDR is much more useful

Broader multiple hypothesis testing concerns

- Previous discussion looked at multiple hypothesis testing in the context of a single study
  - Multiple tests were planned and analysis was planned to account for them
- Concerns about multiple hypothesis testing also arise in an "unplanned" manner

Analyzing data without a plan

- The framework for hypothesis testing assumes all aspects of the experimental design are defined before the experiment and the analysis are performed
  - Not doing this can invalidate the interpretation of a p-value
  - Easy trap to fall into

Examples of multiple comparisons

- three groups instead and try an ANOVA
- same test
- In multiple regression (we will discuss this later...), choosing to include or exclude different independent variables
  - For complex data sets, trying enough approaches will almost always result in a "statistically significant" result
### Sequential Analyses
- Another common approach is to try an experiment, and if the result is not statistically significant, to then repeat it with additional samples or experimental replicates
  - Another form of multiple comparisons
  - Problem with this approach is that it is biased towards a statistically significant result
    - Stop experimenting if a result is statistically significant
    - Continue experimenting otherwise
    - In theory you can get a statistically significant result with this approach

### Publication Bias
- Remember, the interpretation of a p-value is the probability of observing data at least as extreme as the data observed, assuming the null hypothesis is true
  - This is not the same as the probability the null hypothesis is true
- Most p-values we see are in journal articles
- There is a strong preference to publish results which are "statistically significant," i.e. which have $p < 0.05$
- Some of these results are "real" and some are false positives
- Because the publications are selected based on the p-value, the interpretation of a p-value in published results is skewed
  - If we assume the null hypothesis is true, the probability of a false positive can be much higher than 5%

---

Marshall University School of Medicine