# Statistical Learning

## Lecture 11a - Multiple Hypothesis Testing

ANU - RSFAS

Last Updated: Thu May 19 17:31:35 2022

# Muliple Hypthesis Testing

- A single null hypothesis might look like $H_0$: the expected blood pressures of mice in the control and treatment groups are the same.
- We will now consider testing $m$ null hypotheses, $H_{01}, \ldots, H_{0m}$, where e.g. $H_{0j}$ : the expected values of the j$^{th}$ biomarker among mice in the control and treatment groups are equal.
- In this setting, we need to be careful to avoid incorrectly rejecting too many null hypotheses, i.e. having too many false positives.

# Review of Hypothesis Testing

Hypothesis tests allow us to answer simple "yes-or-no" questions, such as:

- Is the true coefficient $\beta_j$ in a linear regression equal to zero?
- Does the expected blood pressure among mice in the treatment group equal the expected blood pressure among mice in the control group?

Hypothesis testing proceeds as follows:

1. Define the null and alternative hypotheses
2. Construct the test statistic
3. Compute the $p$-value
4. Decide whether to reject the null hypothesis

# Decision Outcomes

| | | Truth | |
|---|---|---|---|
| | | $H_0$ | $H_a$ |
| **Decision** | Reject $H_0$ | Type I Error | Correct |
| | Do Not Reject $H_0$ | Correct | Type II Error |

# Decision Outcomes

- The **Type I error rate** is the probability of making a **Type I error**.
- We want to ensure a small **Type I error rate**.
- If we only reject $H_0$ when the $p$-value is less than $\alpha$, then the Type I error rate will be at most $\alpha$.
- So, we reject $H_0$ when the $p$-value falls below some $\alpha$ - often we choose (i.e. we control) $\alpha$ to be equal 0.05 or 0.01 or 0.001.
- $\alpha = 0.05$ was due to R.A. Fisher stating that in a particular problem it seemed reasonable.

# Multiple Testing

- Now suppose that we wish to test $m$ null hypotheses, $H_{01}, \ldots, H_{0m}$
- Can we simply reject all null hypotheses for which the corresponding $p$-value falls below (say) 0.01?
- If we reject all null hypotheses for which the $p$-value falls below 0.01, then how many **Type I errors** will we make?

# A Thought Experiment

- Suppose that we flip a fair coin ten times, and we wish to test $H_0$: the coin is fair.
  - We have a binomial set-up here
  - We'll probably get approximately the same number of heads and tails.
  - The $p$-value probably won't be small. We do not reject H0.

# A Thought Experiment

- But what if we flip 1,024 fair coins ten times each?
  - We'd expect one coin (on average) to come up all tails.
  - The $p$-value for the null hypothesis that this particular coin is fair is less than 0.002!
  - So we would conclude it is not fair, i.e. we reject $H_0$, even though it's a fair coin.
- If we test a lot of hypotheses, we are almost certain to get one very small $p$-value by chance!

# Multiple Testing: **Even** XKCD Weighs In

- **Even** posted on an office door at CSIRO!


https://xkcd.com/882/

# The Challenge of Multiple Testing

- Suppose we test $H_{01}, \ldots, H_{0m}$, all of which are true, and reject any null hypothesis with a $p$-value below 0.01.
- Then we expect to falsely reject approximately $0.01 \times m$ null hypotheses.
- If $m = 10,000$, then we expect to falsely reject 100 null hypotheses by chance!
- **That's a lot of Type I errors, i.e. false positives!**

# The Family-Wise Error Rate

- The family-wise error rate (FWER) is the probability of making at least one Type I error when conducting m hypothesis tests.
- $\text{FWER} = P(V \geq 1)$

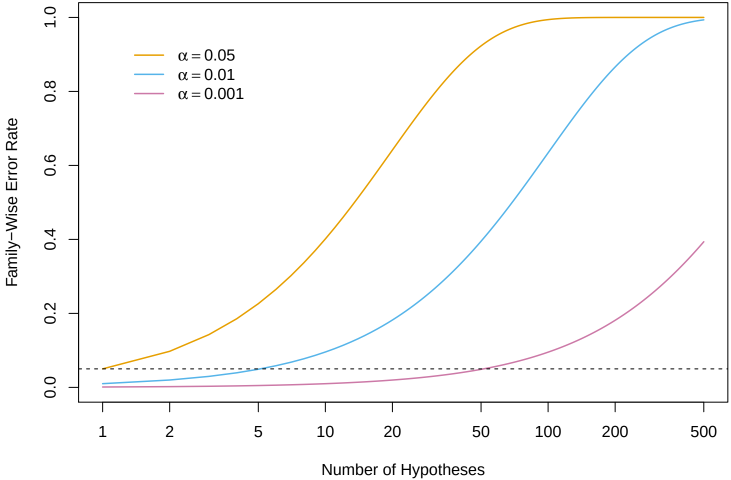
| | $H_0$ is True | $H_0$ is False | Total |
|---|---|---|---|
| Reject $H_0$ | $V$ | $S$ | $R$ |
| Do Not Reject $H_0$ | $U$ | $W$ | $m - R$ |
| Total | $m_0$ | $m - m_0$ | $m$ |

# Challenges in Controlling the Family-Wise Error Rate

$$\begin{aligned} \text{FWER} &= 1 - P(\text{do not falsely reject any null hypotheses}) \\ &= 1 - P\left(\cap_{j=1}^{m} \{\text{do not falsely reject } H_{0j}\}\right) \end{aligned}$$

- If the tests are independent and all $H_{0j}$ are true then:

$$\text{FWER} = 1 - \prod_{j=1}^{m}(1-\alpha) = 1 - (1-\alpha)^m$$


α = 0.05
α = 0.01
α = 0.001
Family−Wise Error Rate
0.0
0.2
0.4
0.6
0.8
1.0
1
2
5
10
20
50
100
200
500
Number of Hypotheses

# The Bonferroni Correction

$$\begin{aligned} \text{FWER} &= P(\text{falsely reject at least one null hypotheses}) \\ &= P\left(\cup_{j=1}^{m} A_j\right) \leq \sum_{i=1}^{n} P(A_j) \end{aligned}$$

- $A_j$ is the event we falsely reject the $\text{j}^{th}$ null hypothesis
- Note: the inequality is due to Boole's inequality https://en.wikipedia.org/wiki/Boole's_inequality
- If we only reject hypotheses when the $p$-value is less than $\alpha/m$, then

$$\text{FWER} \leq \sum_{i=1}^{n} P(A_j) \leq \sum_{i=1}^{n} \alpha/m = m \times \alpha/m = \alpha$$

- This is the **Bonferroni Correction**: to control FWER at level $\alpha$, reject any null hypothesis with $p$-value below $\alpha/m$.

# Fund Manager Data

| Manager | Mean, $\bar{x}$ | $s$ | $t$-statistic | $p$-value |
|---|---|---|---|---|
| One | 3.0 | 7.4 | 2.86 | 0.006 |
| Two | -0.1 | 6.9 | -0.10 | 0.918 |
| Three | 2.8 | 7.5 | 2.62 | 0.012 |
| Four | 0.5 | 6.7 | 0.53 | 0.601 |
| Five | 0.3 | 6.8 | 0.31 | 0.756 |

- $H_{0j}$ : the j$^{th}$ manager's expected excess return equals zero.
- Set $\alpha = 0.05$, which do we reject?
- However, we have tested multiple hypotheses, so the **FWER is greater than 0.05**.

# Fund Manager Data - Bonferroni Correction

| Manager | Mean, $\bar{x}$ | $s$ | $t$-statistic | $p$-value |
|---|---|---|---|---|
| One | 3.0 | 7.4 | 2.86 | 0.006 |
| Two | -0.1 | 6.9 | -0.10 | 0.918 |
| Three | 2.8 | 7.5 | 2.62 | 0.012 |
| Four | 0.5 | 6.7 | 0.53 | 0.601 |
| Five | 0.3 | 6.8 | 0.31 | 0.756 |

- Set $\alpha^* = \alpha/m = 0.05/5 = 0.001$
- Now we only reject the first manager.
- The FWER is 0.05.

# Holm's Method for Controlling the FWER

1. Compute $p$-values, $p_1, \ldots p_m$, for the $m$ null hypotheses $H_{01}, \ldots, H_{0m}$.
2. Order the $m$ $p$-values so that $p(1) \leq p(2) \leq \cdots \leq p(m)$.
3. Define

$$L = \min \left\{ j : p(j) > \frac{\alpha}{m + 1 - j} \right\}.$$

4. Reject all null hypotheses $H_{0j}$ for which $p_j < p(L)$.

- **Holm's method** controls the FWER at level $\alpha$.

# Holm's Method

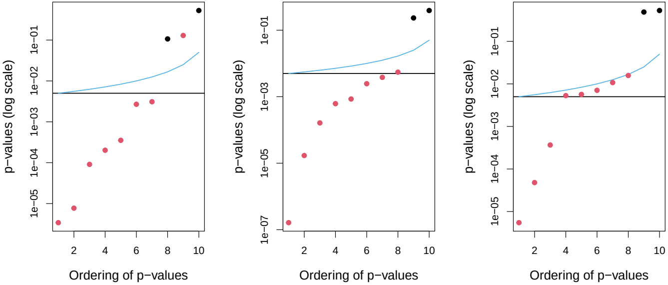
| Manager | Mean, $\bar{x}$ | $s$ | $t$-statistic | $p$-value |
|---|---|---|---|---|
| One | 3.0 | 7.4 | 2.86 | 0.006 |
| Two | -0.1 | 6.9 | -0.10 | 0.918 |
| Three | 2.8 | 7.5 | 2.62 | 0.012 |
| Four | 0.5 | 6.7 | 0.53 | 0.601 |
| Five | 0.3 | 6.8 | 0.31 | 0.756 |

- The Holm procedure rejects the first two null hypotheses:

$$
\begin{aligned}
p(1) &= 0.006 < 0.05/(5+1-1) = 0.0100 \\
p(2) &= 0.012 < 0.05/(5+1-2) = 0.0125 \\
p(3) &= 0.601 > 0.05/(5+1-3) = 0.0167
\end{aligned}
$$

- Note: $L = 3$

# A Comparison with m = 10 p-values




- 3 different simulations
- m = 10 (black dots $m_0 = 2$ true null hypotheses)
- Bonferroni correction $\Rightarrow$ reject all below the black line
- Holm procedure $\Rightarrow$ reject all below the blue line
- The FWER is 0.05

## Other Methods

- **Tukey's Method**: All pairwise differences among means:

$$H_0 : \mu_i - \mu_j = 0 \quad \forall i, j.$$

- **Scheffé's Method** for testing arbitrary linear combinations of a set of expected means:

$$H_0 : \frac{1}{2}(\mu_1 + \mu_2) = \frac{1}{3}(\mu_2 + \mu_4 + \mu_5)$$

- Bonferroni and Holm are general procedures that will work in most settings.
- However, in certain special cases, methods such as Tukey and Scheffé can give better results: i.e. more rejections while maintaining FWER control.

# False Discovery Rate - A Different Idea

| | $H_0$ is True | $H_0$ is False | Total |
|---|---|---|---|
| Reject $H_0$ | $V$ | $S$ | $R$ |
| Do Not Reject $H_0$ | $U$ | $W$ | $m - R$ |
| Total | $m_0$ | $m - m_0$ | $m$ |

- The **FWER** rate focuses on controlling $P(V > 1)$, i.e., the probability of falsely rejecting any null hypothesis.
- This is a tough ask when $m$ is large! It will cause us to be super conservative (i.e. to very rarely reject).
- Instead, we can control the **false discovery rate**:

$$\text{FDR} = E(V/R) = E\left(\frac{\text{number of false rejections}}{\text{total number of rejections}}\right)$$

# False Discovery Rate

- A scientist conducts a hypothesis test on each of $m = 20,000$ drug candidates.
- She wants to identify a smaller set of promising candidates to investigate further.
- She wants reassurance that this smaller set is really "promising", i.e. not too many falsely rejected $H_0$'s.
- FWER controls $P(\text{at least one false rejection})$.
- **FDR controls the fraction of candidates in the smaller set that are really false rejections. This is what she needs!**

# Benjamini-Hochberg Procedure to Control FDR

1. Specify $q$, the level at which to control the FDR.
2. Compute $p$-values, $p_1, \ldots p_m$, for the $m$ null hypotheses $H_{01}, \ldots, H_{0m}$.
3. Order the $m$ $p$-values so that $p(1) \leq p(2) \leq \cdots \leq p(m)$.
4. Define

$$L = \max \left\{ j : p(j) < \frac{qj}{m} \right\}.$$

5. Reject all null hypotheses $H_{0j}$ for which $p_j \leq p(L)$.

- Then the $FDR \leq q$.

# FDR - Fund Managers

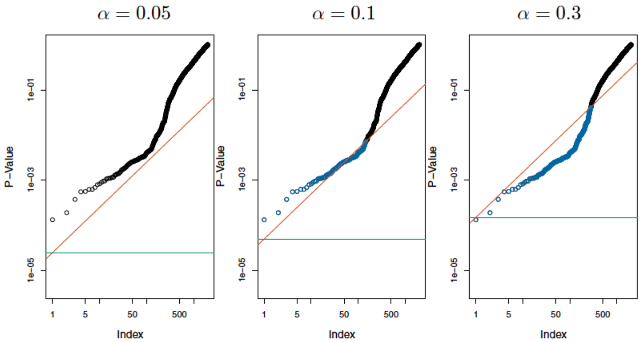
| Manager | Mean, $\bar{x}$ | $s$ | $t$-statistic | $p$-value |
|---|---|---|---|---|
| One | 3.0 | 7.4 | 2.86 | 0.006 |
| Two | -0.1 | 6.9 | -0.10 | 0.918 |
| Three | 2.8 | 7.5 | 2.62 | 0.012 |
| Four | 0.5 | 6.7 | 0.53 | 0.601 |
| Five | 0.3 | 6.8 | 0.31 | 0.756 |

- To control FDR at level $q = 0.05$ using Benjamini-Hochberg:

$$
\begin{aligned}
p(1) &= 0.006 < 0.05/(5) = 0.010 \\
p(2) &= 0.012 < 0.05(2)/(5) = 0.020 \\
p(3) &= 0.601 > 0.05(3)/(5) = 0.030 \\
p(4) &= 0.756 > 0.05(4)/(5) = 0.040 \\
p(5) &= 0.918 > 0.05(5)/(5) = 0.050
\end{aligned}
$$

- So, we reject $H_{01}$ and $H_{03}$ and $L = 2$.

# A Comparison of FDR Versus FWER




- $p$-values for $m = 2,000$ null hypotheses
- To control FWER at various levels with the Bonferroni method: reject hypotheses below green line. (Only one rejection! [graph on the right])
- The orange lines indicate the p-value thresholds corresponding to FDR control, via Benjamini-Hochberg, at levels $q = 0.05, q = 0.1, q = 0.3$ - rejected hypotheses shown in blue.