Multiple Hypothesis Testing

- This session focuses on *multiple hypothesis testing*.

Multiple Hypothesis Testing

- This session focuses on *multiple hypothesis 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*.

Multiple Hypothesis Testing

- This session focuses on *multiple hypothesis 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*.

Multiple Hypothesis Testing

- This session focuses on *multiple hypothesis 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.

Multiple Testing

- Now suppose that we wish to test m null hypotheses, $H_{01}, \ldots, H_{0m}$.

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?

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*.

A Thought Experiment

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

A Thought Experiment

- Suppose that we flip a fair coin ten times, and we wish to test *H_0: the coin is fair*.
  - We'll probably get approximately the same number of heads and tails.
  - The p-value probably won't be small. We do not reject H_0.
- But what if we flip 1,024 fair coins ten times each?

A Thought Experiment

- Suppose that we flip a fair coin ten times, and we wish to test H_0*: the coin is fair*.
  - We'll probably get approximately the same number of heads and tails.
  - The p-value probably won't be small. We do not reject H_0.
- But what if we flip 1,024 fair coins ten times each?
  - We'd expect one coin (on average) to come up all tails.

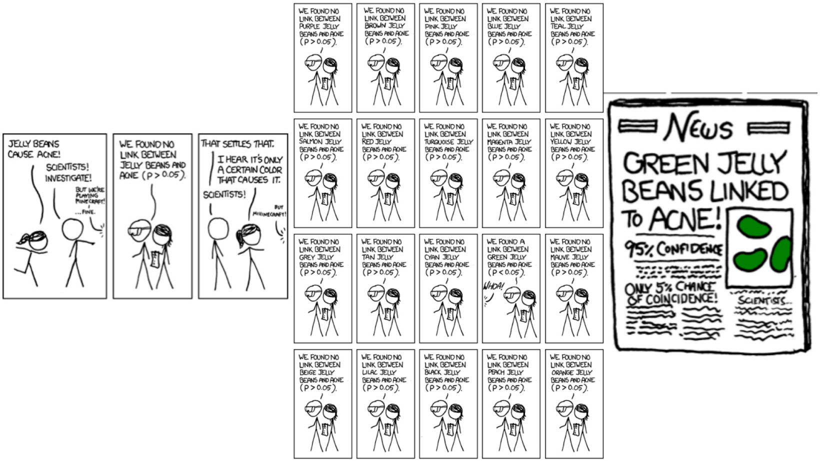
A Thought Experiment

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

A Thought Experiment

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



`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.

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.

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!

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.

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.
- $\text{FWER} = \Pr(V \geq 1)$

	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

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

Challenges in Controlling the Family-Wise Error Rate

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

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

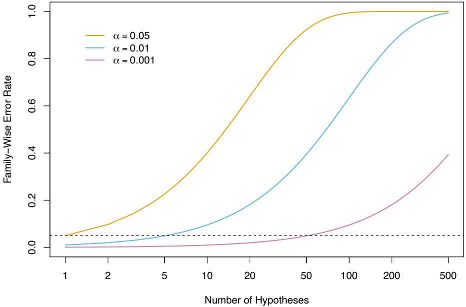
$$\text{FWER} = 1 - \prod_{j=1}^{m} (1-\alpha) = 1 - (1-\alpha)^m.$$

Challenges in Controlling the Family-Wise Error Rate

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

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

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



The Bonferroni Correction

$$\begin{aligned}
\text{FWER} &= \Pr(\text{falsely reject at least one null hypothesis}) \\
&= \Pr(\cup_{j=1}^{m} A_j) \\
&\leq \sum_{j=1}^{m} \Pr(A_j)
\end{aligned}$$

where A_j is the event that we falsely reject the jth null hypothesis.

The Bonferroni Correction

$$\begin{aligned}
\text{FWER} &= \Pr(\text{falsely reject at least one null hypothesis}) \\
&= \Pr(\cup_{j=1}^{m} A_j) \\
&\leq \sum_{j=1}^{m} \Pr(A_j)
\end{aligned}$$

where A_j is the event that we falsely reject the jth null hypothesis.

- If we only reject hypotheses when the p-value is less than α/m, then

$$\text{FWER} \leq \sum_{j=1}^{m} \Pr(A_j) \leq \sum_{j=1}^{m} \frac{\alpha}{m} = m \times \frac{\alpha}{m} = \alpha,$$

  because $\Pr(A_j) \leq \alpha/m$.

- This is the *Bonferroni Correction*: to control FWER at level α, reject any null hypothesis with p-value below α/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

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 jth manager's expected excess return equals zero*.
- If we reject H_{0j} if the p-value is less than $\alpha = 0.05$, then we will conclude that the *first* and *third* managers have significantly non-zero excess returns.

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 jth manager's expected excess return equals zero*.
- If we reject H_{0j} if the p-value is less than $\alpha = 0.05$, then we will conclude that the *first* and *third* managers have significantly non-zero excess returns.
- However, we have tested multiple hypotheses, so the FWER is *greater* than 0.05.

Fund Manager Data with 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

- Using a Bonferroni correction, we reject for p-values less than $\alpha/m = 0.05/5 = 0.01$.

Fund Manager Data with 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

- Using a Bonferroni correction, we reject for p-values less than $\alpha/m = 0.05/5 = 0.01$.
- Consequently, we will reject the null hypothesis only for the *first* manager.

Fund Manager Data with 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

- Using a Bonferroni correction, we reject for p-values less than $\alpha/m = 0.05/5 = 0.01$.
- Consequently, we will reject the null hypothesis only for the *first* manager.
- Now the FWER is at most 0.05.

Holm's Method for Controlling the FWER

Holm's Method for Controlling the FWER

1. Compute p-values, $p_1, \ldots, p_m$, for the m null hypotheses $H_{01}, \ldots, H_{0m}$.

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 $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$.

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

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

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 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 α.

Holm's Method on the 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

- The ordered p-values are $p_{(1)} = 0.006, p_{(2)} = 0.012$, $p_{(3)} = 0.601$, $p_{(4)} = 0.756$ and $p_{(5)} = 0.918$.

Holm's Method on the 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

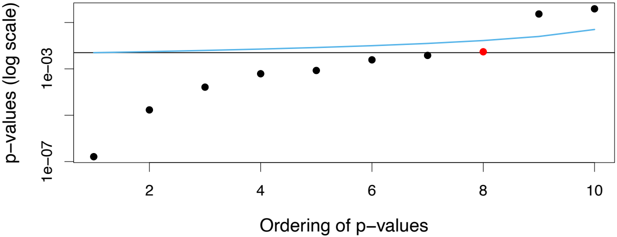
- The ordered p-values are $p_{(1)} = 0.006, p_{(2)} = 0.012$, $p_{(3)} = 0.601$, $p_{(4)} = 0.756$ and $p_{(5)} = 0.918$.
- The Holm procedure rejects the first two null hypotheses, because
  - $p_{(1)} = 0.006 < 0.05/(5+1-1) = 0.0100$
  - $p_{(2)} = 0.012 < 0.05/(5+1-2) = 0.0125$,
  - $p_{(3)} = 0.601 > 0.05/(5+1-3) = 0.0167$.

Holm's Method on the 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

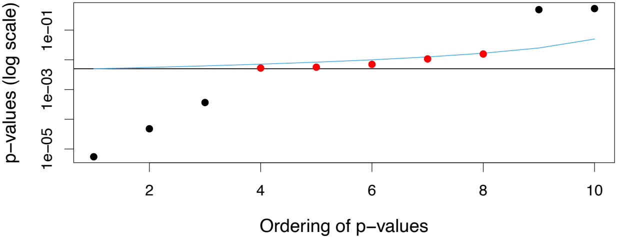
- The ordered p-values are $p_{(1)} = 0.006, p_{(2)} = 0.012$, $p_{(3)} = 0.601$, $p_{(4)} = 0.756$ and $p_{(5)} = 0.918$.
- The Holm procedure rejects the first two null hypotheses, because
  - $p_{(1)} = 0.006 < 0.05/(5 + 1 - 1) = 0.0100$
  - $p_{(2)} = 0.012 < 0.05/(5 + 1 - 2) = 0.0125$,
  - $p_{(3)} = 0.601 > 0.05/(5 + 1 - 3) = 0.0167$.
- Holm rejects H_0 for the *first* and *third* managers, but Bonferroni only rejects H_0 for the *first* manager.

A Comparison with $m = 10$ p-values



- Aim to control FWER at 0.05.
- p-values below the black horizontal line are rejected by Bonferroni.
- p-values below the blue line are rejected by Holm.
- Holm and Bonferroni make the same conclusion on the black points, but only Holm rejects for the red point.

A More Extreme Example



- Now five hypotheses are rejected by Holm but not by Bonferroni ....
- .... even though both control FWER at 0.05.

Holm or Bonferroni?

- Bonferroni is simple ... reject any null hypothesis with a p-value below α/m.
- Holm is slightly more complicated, but it will lead to more rejections while controlling FWER!!
- So, *Holm is a better choice*!