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.

A Quick Review of Hypothesis Testing

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

- Is the true coefficient β_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?

A Quick Review of Hypothesis Testing

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

- Is the true coefficient β_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

A Quick Review of Hypothesis Testing

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

- Is the true coefficient β_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

A Quick Review of Hypothesis Testing

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

- Is the true coefficient β_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

A Quick Review of Hypothesis Testing

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

- Is the true coefficient β_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

1. Define the Null and Alternative Hypotheses

- We divide the world into *null* and *alternative* hypotheses.
- The null hypothesis, H_0, is the default state of belief about the world. For instance:
  1. The true coefficient β_j equals zero.
  2. There is no difference in the expected blood pressures.

1. Define the Null and Alternative Hypotheses

- We divide the world into *null* and *alternative* hypotheses.
- The null hypothesis, H_0, is the default state of belief about the world. For instance:
  1. The true coefficient β_j equals zero.
  2. There is no difference in the expected blood pressures.
- The alternative hypothesis, H_a, represents something different and unexpected. For instance:
  1. The true coefficient β_j is non-zero.
  2. There is a difference in the expected blood pressures.

2. Construct the Test Statistic

- The test statistic summarizes the extent to which our data are consistent with H_0.

2. Construct the Test Statistic

- The test statistic summarizes the extent to which our data are consistent with H_0.
- Let $\hat{\mu}_t$ / $\hat{\mu}_c$ respectively denote the average blood pressure for the n_t / n_c mice in the treatment and control groups.

2. Construct the Test Statistic

- The test statistic summarizes the extent to which our data are consistent with H_0.
- Let $\hat{\mu}_t$ / $\hat{\mu}_c$ respectively denote the average blood pressure for the n_t / n_c mice in the treatment and control groups.
- To test $H_0 : \mu_t = \mu_c$, we use a two-sample t-statistic

$$T = \frac{\hat{\mu}_t - \hat{\mu}_c}{s\sqrt{\frac{1}{n_t} + \frac{1}{n_c}}}$$

3. Compute the p-Value

- The p-value is the probability of observing a test statistic at least as extreme as the observed statistic, *under the assumption that H_0 is true*.

3. Compute the p-Value

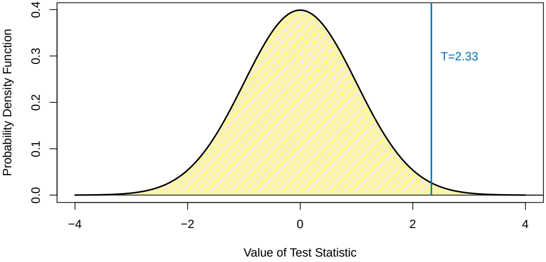
- The p-value is the probability of observing a test statistic at least as extreme as the observed statistic, *under the assumption that H_0 is true*.
- A small p-value provides evidence *against* H_0.

3. Compute the p-Value

- The p-value is the probability of observing a test statistic at least as extreme as the observed statistic, *under the assumption that H_0 is true*.
- A small p-value provides evidence *against* H_0.
- Suppose we compute $T = 2.33$ for our test of $H_0 : \mu_t = \mu_c$.

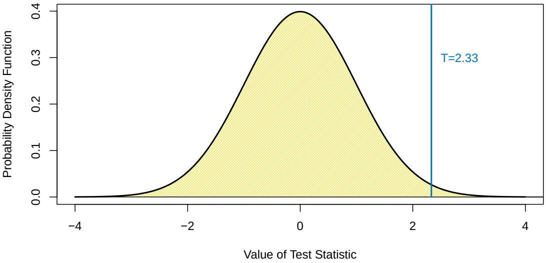
3. Compute the p-Value

- The p-value is the probability of observing a test statistic at least as extreme as the observed statistic, *under the assumption that H_0 is true*.
- A small p-value provides evidence *against* H_0.
- Suppose we compute $T = 2.33$ for our test of $H_0 : \mu_t = \mu_c$.
- Under H_0, $T \sim N(0, 1)$ for a two-sample t-statistic.



3. Compute the p-Value

- The p-value is the probability of observing a test statistic at least as extreme as the observed statistic, *under the assumption that H_0 is true*.
- A small p-value provides evidence *against* H_0.
- Suppose we compute $T = 2.33$ for our test of $H_0 : \mu_t = \mu_c$.
- Under H_0, $T \sim N(0, 1)$ for a two-sample t-statistic.



- The p-value is 0.02 because, if H_0 is true, we would only see $|T|$ this large 2% of the time.

4. Decide Whether to Reject H_0, Part 1

- A small p-value indicates that such a large value of the test statistic is unlikely to occur under H_0.

4. Decide Whether to Reject H_0, Part 1

- A small p-value indicates that such a large value of the test statistic is unlikely to occur under H_0.
- So, a small p-value provides evidence against H_0.

4. Decide Whether to Reject H_0, Part 1

- A small p-value indicates that such a large value of the test statistic is unlikely to occur under H_0.
- So, a small p-value provides evidence against H_0.
- If the p-value is sufficiently small, then we will want to *reject* H_0 (and, therefore, make a potential "discovery").

4. Decide Whether to Reject H_0, Part 1

- A small p-value indicates that such a large value of the test statistic is unlikely to occur under H_0.
- So, a small p-value provides evidence against H_0.
- If the p-value is sufficiently small, then we will want to *reject* H_0 (and, therefore, make a potential “discovery”).
- *But how small is small enough*? To answer this, we need to understand the *Type I error*.

4. Decide Whether to Reject H_0, Part 2

		Truth	
		H_0	H_a
Decision	Reject H_0	Type I Error	Correct
	Do Not Reject H_0	Correct	Type II Error

4. Decide Whether to Reject H_0, Part 2

The null hypothesis holds, and we didn't reject it!

		Truth	
		H_0	H_a
Decision	Reject H_0	Type I Error	Correct
	Do Not Reject H_0	$\boxed{\text{Correct}}$	Type II Error

4. Decide Whether to Reject H_0, Part 2

The null hypothesis doesn't hold, and we rejected it!

		Truth	
		H_0	H_a
Decision	Reject H_0	Type I Error	$\boxed{\text{Correct}}$
	Do Not Reject H_0	Correct	Type II Error

4. Decide Whether to Reject H_0, Part 2

The null hypothesis doesn't hold, and we didn't reject it!

		Truth	
		H_0	H_a
Decision	Reject H_0	Type I Error	Correct
	Do Not Reject H_0	Correct	Type II Error

4. Decide Whether to Reject H_0, Part 2

The null hypothesis holds, and we rejected it!

		Truth	
		H_0	H_a
Decision	Reject H_0	Type I Error	Correct
	Do Not Reject H_0	Correct	Type II Error

4. Decide Whether to Reject H_0, Part 3

- The *Type I error rate* is the probability of making a Type I error.
- We want to ensure a small Type I error rate.

4. Decide Whether to Reject H_0, Part 3

- 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 α, then the Type I error rate will be at most α.

4. Decide Whether to Reject H_0, Part 3

- 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 α, then the Type I error rate will be at most α.
- So, *we reject H_0 when the p-value falls below some α*: often we choose α to equal 0.05 or 0.01 or 0.001.

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.

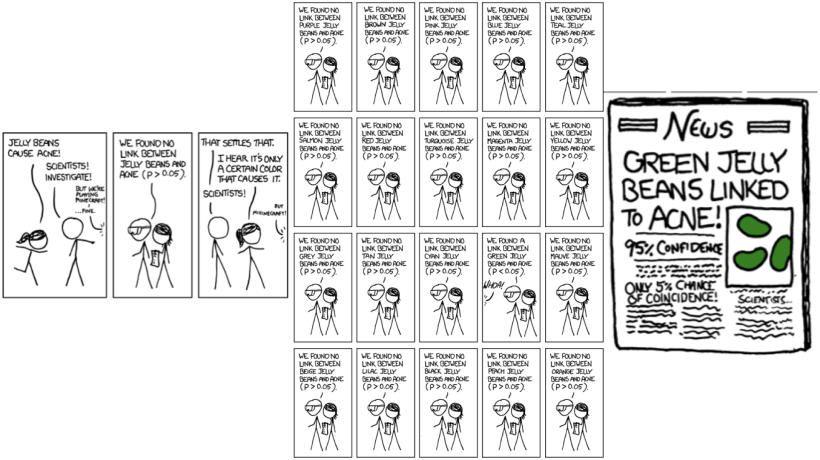
A Thought Experiment

- 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

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

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

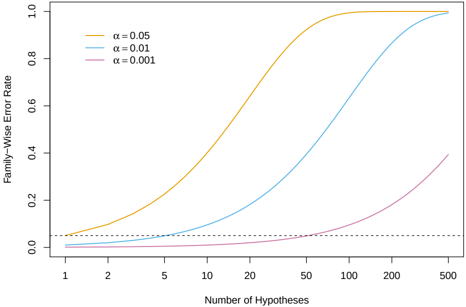
$$\text{FWER} = 1 - \prod_{j=1}^{m} (1 - \alpha) = 1 - (1 - \alpha)^m.$$

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* j*th 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* j*th 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

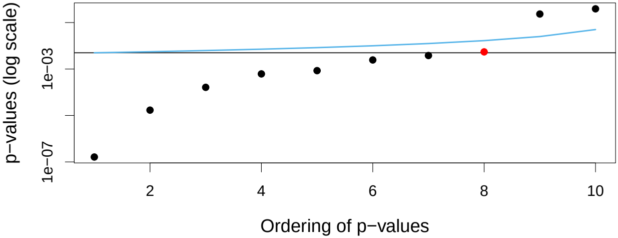
- The ordered p-values are $p_{(1)} = 0.006, p_{(2)} = 0.012$, $p_{(3)} = 0.601$, $p_{(4)} = 0.756$ and $p_{(5)} = 0.918$.
- The Holm procedure rejects the first two null hypotheses, because
  - $p_{(1)} = 0.006 < 0.05/(5 + 1 - 1) = 0.0100$
  - $p_{(2)} = 0.012 < 0.05/(5 + 1 - 2) = 0.0125$,
  - $p_{(3)}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

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

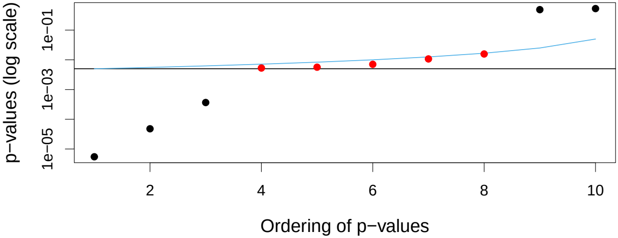
- The ordered p-values are $p_{(1)} = 0.006, p_{(2)} = 0.012$, $p_{(3)} = 0.601$, $p_{(4)} = 0.756$ and $p_{(5)} = 0.918$.
- The Holm procedure rejects the first two null hypotheses, because
    - $p_{(1)} = 0.006 < 0.05/(5+1-1) = 0.0100$
    - $p_{(2)} = 0.012 < 0.05/(5+1-2) = 0.0125$,
    - $p_{(3)}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

Other Methods

- There are lots of specialized approaches to control FWER.

Other Methods

- There are lots of specialized approaches to control FWER.
- For example:
  - *Tukey's Method*: for pairwise comparisons of the difference in expected means among a number of groups.

Other Methods

- There are lots of specialized approaches to control FWER.
- For example:
    - *Tukey's Method*: for pairwise comparisons of the difference in expected means among a number of groups.
    - *Scheffé's Method*: for testing arbitrary linear combinations of a set of expected means, e.g.

$$H_0 : \frac{1}{2}\left(\mu_1 + \mu_3\right) = \frac{1}{3}\left(\mu_2 + \mu_4 + \mu_5\right).$$

Other Methods

- There are lots of specialized approaches to control FWER.
- For example:
  - *Tukey's Method*: for pairwise comparisons of the difference in expected means among a number of groups.
  - *Scheffé's Method*: for testing arbitrary linear combinations of a set of expected means, e.g.

$$H_0 : \frac{1}{2}\left(\mu_1 + \mu_3\right) = \frac{1}{3}\left(\mu_2 + \mu_4 + \mu_5\right).$$

- 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*.

The False Discovery Rate

The False Discovery Rate

- Back to this table:

	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 False Discovery Rate

- Back to this table:

	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 $\Pr(V > 1)$, i.e., the probability of falsely rejecting *any* null hypothesis.

The False Discovery Rate

- Back to this table:

	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 $\Pr(V > 1)$, 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).

The False Discovery Rate

- Back to this table:

	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 $\Pr(V > 1)$, 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*:

$$\text{FDR} = \text{E}(V/R).$$

Intuition Behind the False Discovery Rate

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

Intuition Behind the False Discovery Rate

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

- A scientist conducts a hypothesis test on each of $m = 20,000$ drug candidates.

Intuition Behind the False Discovery Rate

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

- 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.

Intuition Behind the False Discovery Rate

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

- 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.

Intuition Behind the False Discovery Rate

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

- 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 Pr(at least one false rejection).

Intuition Behind the False Discovery Rate

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

- 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 Pr(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

Benjamini-Hochberg Procedure to Control FDR

1. Specify q, the level at which to control the FDR.

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 null hypotheses $H_{01}, \ldots, H_{0m}$.

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 null hypotheses $H_{01}, \ldots, H_{0m}$.
3. Order the p-values so that $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$.

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 null hypotheses $H_{01}, \ldots, H_{0m}$.
3. Order the p-values so that $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$.
4. Define $L = \max\left\{j : p_{(j)} < qj/m\right\}$.

Benjamini-Hochberg Procedure to Control FDR

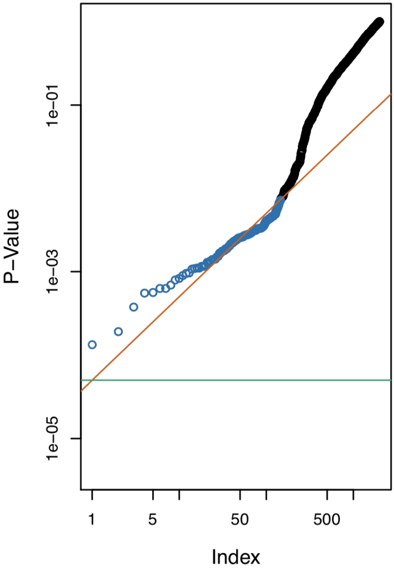
1. Specify q, the level at which to control the FDR.
2. Compute p-values $p_1, \ldots, p_m$ for the null hypotheses $H_{01}, \ldots, H_{0m}$.
3. Order the p-values so that $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$.
4. Define $L = \max\left\{j : p_{(j)} < qj/m\right\}$.
5. Reject all null hypotheses H_{0j} for which $p_j \leq p_{(L)}$.

Benjamini-Hochberg Procedure to Control FDR

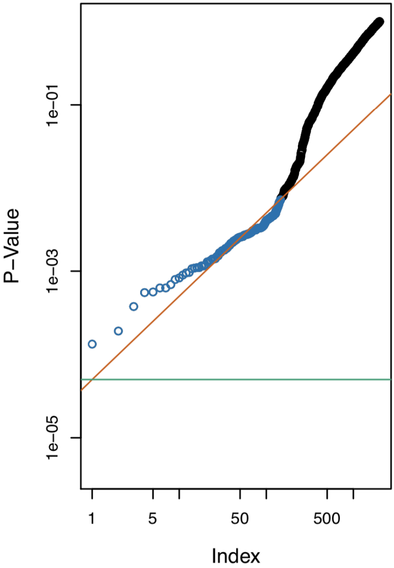
1. Specify q, the level at which to control the FDR.
2. Compute p-values $p_1, \ldots, p_m$ for the null hypotheses $H_{01}, \ldots, H_{0m}$.
3. Order the p-values so that $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(m)}$.
4. Define $L = \max\left\{j : p_{(j)} < qj/m\right\}$.
5. Reject all null hypotheses H_{0j} for which $p_j \leq p_{(L)}$.

Then, FDR $\leq q$.

A Comparison of FDR Versus FWER, Part 1

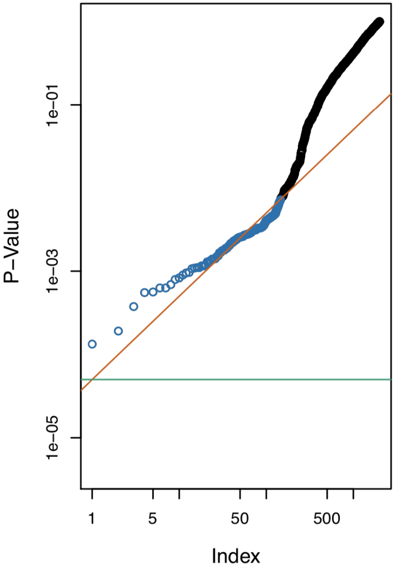


A Comparison of FDR Versus FWER, Part 1



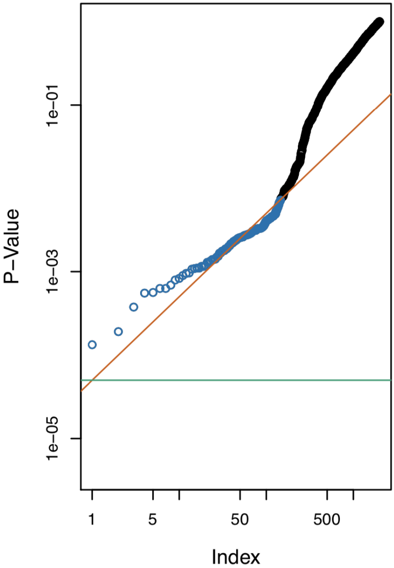
- Here, p-values for $m = 2,000$ null hypotheses are displayed.

A Comparison of FDR Versus FWER, Part 1



- Here, p-values for $m = 2,000$ null hypotheses are displayed.
- To control FWER at level $\alpha = 0.1$ with Bonferroni: reject hypotheses below green line. *(No rejections!)*

A Comparison of FDR Versus FWER, Part 1



- Here, p-values for $m = 2,000$ null hypotheses are displayed.
- To control FWER at level $\alpha = 0.1$ with Bonferroni: reject hypotheses below green line. *(No rejections!)*
- To control FDR at level $q = 0.1$ with Benjamini-Hochberg: reject hypotheses shown in blue.

A Comparison of FDR Versus FWER, Part 2

A Comparison of FDR Versus FWER, Part 2

- Consider $m = 5$ p-values from the *Fund* data:
  $p_1 = 0.006, p_2 = 0.918, p_3 = 0.012, p_4 = 0.601, p_5 = 0.756.$

A Comparison of FDR Versus FWER, Part 2

- Consider $m = 5$ p-values from the *Fund* data:
  $p_1 = 0.006, p_2 = 0.918, p_3 = 0.012, p_4 = 0.601, p_5 = 0.756.$
- Then $p_{(1)} = 0.006$, $p_{(2)} = 0.012$, $p_{(3)} = 0.601$, $p_{(4)} = 0.756$, and $p_{(5)} = 0.918$.

A Comparison of FDR Versus FWER, Part 2

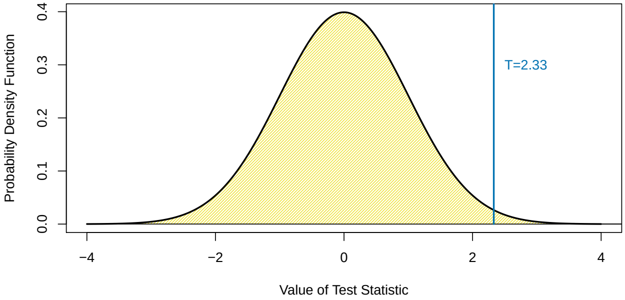
- Consider $m = 5$ p-values from the *Fund* data:
  $p_1 = 0.006, p_2 = 0.918, p_3 = 0.012, p_4 = 0.601, p_5 = 0.756.$
- Then $p_{(1)} = 0.006$, $p_{(2)} = 0.012$, $p_{(3)} = 0.601$, $p_{(4)} = 0.756$, and $p_{(5)} = 0.918$.
- To control FDR at level $q = 0.05$ using Benjamini-Hochberg:
  - Notice that $p_{(1)} < 0.05/5$, $p_{(2)} < 2 \times 0.05/5$, $p_{(3)} > 3 \times 0.05/5$, $p_{(4)} > 4 \times 0.05/5$, and $p_{(5)} > 5 \times 0.05/5$.
  - So, we reject H_{01} and H_{03}.

A Comparison of FDR Versus FWER, Part 2

- Consider $m = 5$ p-values from the *Fund* data:
  $p_1 = 0.006, p_2 = 0.918, p_3 = 0.012, p_4 = 0.601, p_5 = 0.756.$
- Then $p_{(1)} = 0.006$, $p_{(2)} = 0.012$, $p_{(3)} = 0.601$, $p_{(4)} = 0.756$, and $p_{(5)} = 0.918$.
- To control FDR at level $q = 0.05$ using Benjamini-Hochberg:
  - Notice that $p_{(1)} < 0.05/5$, $p_{(2)} < 2 \times 0.05/5$, $p_{(3)} > 3 \times 0.05/5$, $p_{(4)} > 4 \times 0.05/5$, and $p_{(5)} > 5 \times 0.05/5$.
  - So, we reject H_{01} and H_{03}.
- To control FWER at level $\alpha = 0.05$ using Bonferroni:
  - We reject any null hypothesis for which the p-value is less than $0.05/5$.
  - So, we reject only H_{01}.

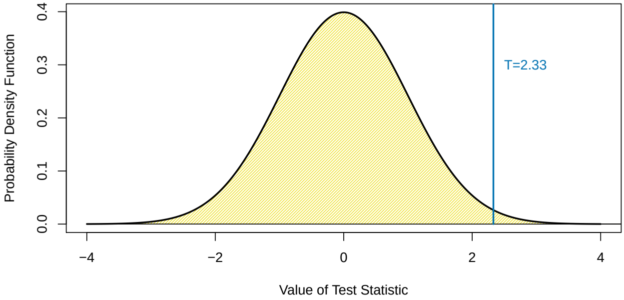
Re-Sampling Approaches

- So far, we have assumed that we want to test some null hypothesis H_0 with some test statistic T, and that we know (or can assume) the distribution of T under H_0.
- This allows us to compute the p-value.



Re-Sampling Approaches

- So far, we have assumed that we want to test some null hypothesis H_0 with some test statistic T, and that we know (or can assume) the distribution of T under H_0.
- This allows us to compute the p-value.



- What if this *theoretical null distribution* is unknown?

A Re-Sampling Approach for a Two-Sample t-Test, Part 1

- Suppose we want to test $H_0 : E(X) = E(Y)$ versus $H_a : E(X) \neq E(Y)$, using n_X independent observations from X and n_Y independent observations from Y.
- The two-sample t-statistic takes the form

$$T = \frac{\hat{\mu}_X - \hat{\mu}_Y}{s\sqrt{1/n_X + 1/n_Y}}.$$

A Re-Sampling Approach for a Two-Sample t-Test, Part 1

- Suppose we want to test $H_0 : E(X) = E(Y)$ versus $H_a : E(X) \neq E(Y)$, using n_X independent observations from X and n_Y independent observations from Y.
- The two-sample t-statistic takes the form

$$T = \frac{\hat{\mu}_X - \hat{\mu}_Y}{s\sqrt{1/n_X + 1/n_Y}}.$$

- If n_X and n_Y are large, then T approximately follows a $N(0, 1)$ distribution under H_0.

A Re-Sampling Approach for a Two-Sample t-Test, Part 1

- Suppose we want to test $H_0 : E(X) = E(Y)$ versus $H_a : E(X) \neq E(Y)$, using n_X independent observations from X and n_Y independent observations from Y.
- The two-sample t-statistic takes the form

$$T = \frac{\hat{\mu}_X - \hat{\mu}_Y}{s\sqrt{1/n_X + 1/n_Y}}.$$

- If n_X and n_Y are large, then T approximately follows a $N(0,1)$ distribution under H_0.
- If n_X and n_Y are small, then we don't know the theoretical null distribution of T.

A Re-Sampling Approach for a Two-Sample t-Test, Part 1

- Suppose we want to test $H_0 : E(X) = E(Y)$ versus $H_a : E(X) \neq E(Y)$, using n_X independent observations from X and n_Y independent observations from Y.
- The two-sample t-statistic takes the form

$$T = \frac{\hat{\mu}_X - \hat{\mu}_Y}{s\sqrt{1/n_X + 1/n_Y}}.$$

- If n_X and n_Y are large, then T approximately follows a $N(0, 1)$ distribution under H_0.
- If n_X and n_Y are small, then we don't know the theoretical null distribution of T.
- Let's take a *permutation* or *re-sampling* approach....

A Re-Sampling Approach for a Two-Sample t-Test, Part 2

A Re-Sampling Approach for a Two-Sample t-Test, Part 2

1. Compute the two-sample t-statistic T on the original data $x_1, \ldots, x_{n_X}$ and $y_1, \ldots, y_{n_Y}$.

A Re-Sampling Approach for a Two-Sample t-Test, Part 2

1. Compute the two-sample t-statistic T on the original data $x_1, \ldots, x_{n_X}$ and $y_1, \ldots, y_{n_Y}$.
2. For $b = 1, \ldots, B$ (where B is a large number, like $1,000$):

A Re-Sampling Approach for a Two-Sample t-Test, Part 2

1. Compute the two-sample t-statistic T on the original data $x_1, \ldots, x_{n_X}$ and $y_1, \ldots, y_{n_Y}$.
2. For $b = 1, \ldots, B$ (where B is a large number, like $1,000$):
   2.1 Randomly shuffle the $n_x + n_Y$ observations.

A Re-Sampling Approach for a Two-Sample t-Test, Part 2

1. Compute the two-sample t-statistic T on the original data $x_1, \ldots, x_{n_X}$ and $y_1, \ldots, y_{n_Y}$.
2. For $b = 1, \ldots, B$ (where B is a large number, like $1,000$):
   2.1 Randomly shuffle the $n_x + n_Y$ observations.
   2.2 Call the first n_X shuffled observations $x_1^*, \ldots, x_{n_X}^*$ and call the remaining observations $y_1^*, \ldots, y_{n_Y}^*$.

A Re-Sampling Approach for a Two-Sample t-Test, Part 2

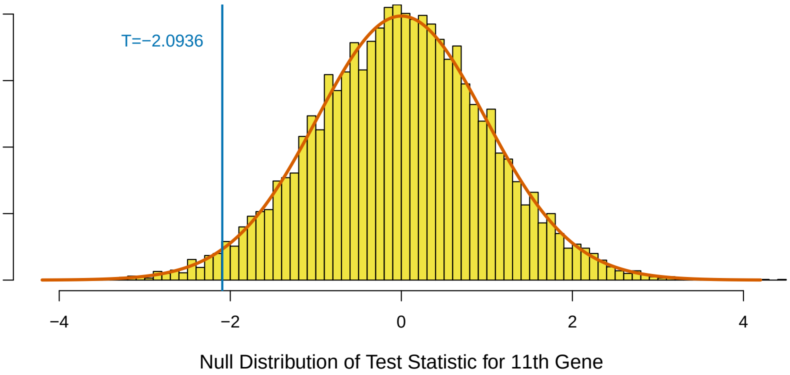
1. Compute the two-sample t-statistic T on the original data $x_1, \dots, x_{n_X}$ and $y_1, \dots, y_{n_Y}$.
2. For $b = 1, \dots, B$ (where B is a large number, like $1,000$):
   2.1 Randomly shuffle the $n_x + n_Y$ observations.
   2.2 Call the first n_X shuffled observations $x_1^*, \dots, x_{n_X}^*$ and call the remaining observations $y_1^*, \dots, y_{n_Y}^*$.
   2.3 Compute a two-sample t-statistic on the shuffled data, and call it T^{*b}.

A Re-Sampling Approach for a Two-Sample t-Test, Part 2

1. Compute the two-sample t-statistic T on the original data $x_1, \ldots, x_{n_X}$ and $y_1, \ldots, y_{n_Y}$.
2. For $b = 1, \ldots, B$ (where B is a large number, like $1,000$):
    2.1 Randomly shuffle the $n_x + n_Y$ observations.
    2.2 Call the first n_X shuffled observations $x_1^*, \ldots, x_{n_X}^*$ and call the remaining observations $y_1^*, \ldots, y_{n_Y}^*$.
    2.3 Compute a two-sample t-statistic on the shuffled data, and call it T^{*b}.
3. The p-value is given by

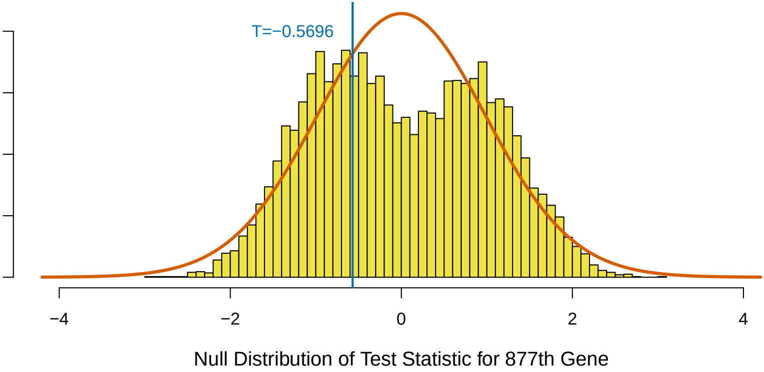
$$\frac{\sum_{b=1}^{B} 1_{(|T^{*b}| \geq |T|)}}{B}.$$

Application to Gene Expression Data, Part 1



Theoretical p-value is 0.041. Re-sampling p-value is 0.042.

Application to Gene Expression Data, Part 2



Theoretical p-value is 0.571. Re-sampling p-value is 0.673.

More on Re-Sampling Approaches

More on Re-Sampling Approaches

- Re-sampling approaches are useful if the theoretical null distribution is unavailable, or requires stringent assumptions. *(So, they're always useful!)*

More on Re-Sampling Approaches

- Re-sampling approaches are useful if the theoretical null distribution is unavailable, or requires stringent assumptions. *(So, they're always useful!)*
- An extension of the re-sampling approach to compute a p-value can be used to control FDR.

More on Re-Sampling Approaches

- Re-sampling approaches are useful if the theoretical null distribution is unavailable, or requires stringent assumptions. *(So, they're always useful!)*
- An extension of the re-sampling approach to compute a p-value can be used to control FDR.
- This example involved a two-sample t-test, but similar approaches can be developed for other test statistics.