Errors in hypothesis testing

- Types of errors.
- Multiple hypothesis testing.

Review of Hypothesis testing

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

- Is smoking independent from cardiovascular disease?
- Does the average blood pressure among mice in the treatment group equal the average blood pressure among mice in the control group?

Review of Hypothesis testing

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

- Is smoking independent from cardiovascular disease?
- Does the average blood pressure among mice in the treatment group equal the average 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 hypothesis

- 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. Smoking is independent of cardiovascular disease.
  2. There is no difference in the average blood pressures.
- The alternative hypothesis, H_1, represents the complement of the null. For instance:
  1. Smoking and cardiovascular disease are not independent.
  2. There is a difference in the average blood pressures.

2. Construct a test statistic

- The test statistic summarizes the extent to which our data are (in) 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 $H_0 : \hat{\mu}_t = \hat{\mu}_c$, we use a two-sample statistic:

$$T = \frac{(n_t + n_c - 2)^{1/2}(\hat{\mu}_t - \hat{\mu}_c)}{(\frac{1}{n_t} + \frac{1}{n_c})^{1/2}(S_x^2 + S_y^2)^{1/2}}$$

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 $H_0 : \hat{\mu}_t = \hat{\mu}_c$
- Under H_0, $T \sim t_{n_t+n_c-2}$ 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 the null hypothesis

- 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").
- How small is small enough?

Types of errors

	Retain Null	Reject Null
H_0 true	✓	type I error
H_1 true	type II error	✓

- A Type I Error is rejecting the null hypothesis when it is true.
- A Type II Error is failing to reject the null hypothesis when the alternative is true.
- Ideally we want both types of errors to happen with a low probability, but there is a trade-off.

Type I error

- As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a significance level of 0.05, $\alpha = 0.05$.
- Type I error rate:

$$P(\text{Type I error}) = P(\text{Reject } H_0 | H_0 \text{ is true}) \leq \alpha$$

- Increasing α increases the Type I error rate.
- When we select α we control for the tolerance we have for type I errors.

Type II error

- If the alternative hypothesis is actually true, what is the chance that we make a Type II Error, i.e. we fail to reject the null hypothesis even when we should reject it?
- The answer is not obvious, but
  - If the true population average is very close to the null hypothesis value, it will be difficult to detect a difference (and reject H_0).
  - If the true population average is very different from the null hypothesis value, it will be easier to detect a difference.
- The probability of correctly rejecting the null is the **power** of the test.

	Retain Null	Reject Null
H_0 true	$1-\alpha$	α(significance)
H_1 true	β	$1-\beta$(power)

Multiple testing

- Now assume we want to test multiple hypotheses $H_{01}, H_{02}, \ldots H_{0m}$.
- If we reject all null hypotheses for which the p-value falls below 0.05, 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'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.
- 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 the coin is fair.
- If we test a lot of hypotheses, we are almost certain to get one very small p-value by chance!

Multiple hypotheses testing

- Suppose we test $H_{01}, \ldots, H_{0m}$, all of which are true, and reject any null hypothesis with a p-value below 0.05.
- Then we expect to falsely reject approximately $0.05 \times m$ null hypotheses.
- If m = 10,000, then we expect to falsely reject 500 null hypotheses by chance!
- That's a lot of Type I errors, i.e. false discoveries/false positives!
- Example: Genome-wide association studies.

Family-wise error rate

	Retain Null	Reject Null	
H_0 true	U	V	m_0
H_1 true	W	S	$m - m_0$
	$m - R$	R	m

$$\text{FWER} = 1 - P(\text{do not falsely reject any null hypothesis}) =$$
$$1 - \cap_{j=1}^{m} P(\text{do not falsely reject } H_{0j})$$

Family-wise error rate

	Retain Null	Reject Null	
H_0 true	U	V	m_0
H_1 true	W	S	$m - m_0$
	$m - R$	R	m

$$\text{FWER} = 1 - P(\text{do not falsely reject any null hypothesis}) =$$
$$1 - \cap_{j=1}^{m} P(\text{do not falsely reject } H_{0j})$$

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

$$\text{FWER} = 1 - \prod_{j=1}^{m} P(\text{do not falsely reject } H_{0j}) = 1 - (1 - \alpha)^m$$

Multiple hypotheses testing

$$\text{FWER} = P(\text{falsely reject at least one null hypothesis}) =$$

$$P(\cup_{j=1}^{m} A_j) \leq \sum_{j=1}^{m} P(A_j)$$

where A_j is the event that we falsely reject the j-th null hypothesis. If we only reject hypotheses when the p-value is less than α/m, then

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

because $P(A_j) \leq \alpha/m$

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

Example: Video Games and ADHD

vs	Internet	TV	VG-C	VG-I
Young's Addiction Scale	0.804	**0.040**	**< 0.001**	**<0.001**
Conner's Scale: Oppositional	0.096	0.397	0.917	0.826
Conner's Scale: Inattention	0.289	0.311	**0.001**	**<0.001**
Conner's Scale: Hyperactivity	0.901	0.397	0.800	0.142
Conner's Scale: ADHD	0.115	0.343	**0.018**	**0.020**

- If we reject H_{0j} if the p-value is less than $\alpha = 0.05$, we will conclude that TV, VG-C, VG-I significantly affect YAS, VG-C and VG-I significantly affect Inattention and ADHD.
- However, we have tested multiple hypotheses, so the FWER is greater than 0.05.

Example: Video Games and ADHD

vs	Internet	TV	VG-C	VG-I
Young's Addiction Scale	0.804	0.040	$<$ **0.001**	$<$**0.001**
Conner's Scale: Oppositional	0.096	0.397	0.917	0.826
Conner's Scale: Inattention	0.289	0.311	**0.001**	$<$**0.001**
Conner's Scale: Hyperactivity	0.901	0.397	0.800	0.142
Conner's Scale: ADHD	0.115	0.343	0.018	0.020

- Using the Bonferroni correction we will reject p-values less than $\alpha/20 = 0.0025$.
- If we reject H_{0j} if the p-value is less than 0.0025, we will conclude that VG-C, VG-I significantly affect YAS, VG-C and VG-I significantly affect Inattention.

Holm's method for controling FWER

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

$$L = \min_j : p_{(j)} > \frac{\alpha}{m + 1 - j}$$

- Reject all null hypotheses H_{0j} for which $p_j < p_{(L)}$
- Holm's method controls the FWER at level α.

Bonferroni vs Holm

- 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

$$p_{(1)} = 0.006, p_{(2)} = 0.012, p_{(3)} = 0.601, p_{(4)} = 0.756, p_{(5)} = 0.918.$$

  - Bonferroni?
  - Bonferroni-Holm?

Bonferroni vs Holm

- 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!!
- Holm is a better choice

The False Discovery Rate

	Retain Null	Reject Null	
H_0 true	U	V	m_0
H_1 true	W	S	$m - m_0$
	$m - R$	R	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:

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

The False Discovery Rate

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

- 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 P(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 for controlling 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 $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$.
4. Define $L = max_j : p_{(j)} < qj/m$.
5. Reject all null hypotheses H_{0j} for which $p_{(j)} \leq p_{(L)}$.

Then, FDR $\leq q$.

FWER vs FDR

- 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

$$p_{(1)} = 0.006, p_{(2)} = 0.012, p_{(3)} = 0.601, p_{(4)} = 0.756, p_{(5)} = 0.918.$$

  - Bonferroni?
  - Bonferroni-Holm?
  - Benjamini- Hochberg?