Week 13:
Multiple hypothesis testing

Phase IV

Key ideas

- The number of statistical tests performed affects the chances of getting a false positive.
- We need to adjust our p-values to have our errors under control.
- We control the family-wise error rate (FWER) if we are worried about observing any false positives
- We control the false discovery rate (FDR) in exploratory scenarios where we can live with false positives.

Refresh 1: Null hypothesis testing

Null Hypothesis:

An assumed statement (e.g., there are no differences in means).

Statistic:

It quantifies the observed data assuming that our null hypothesis is true.

P-value

Probability of obtaining an statistic at least as extreme as the observed one.

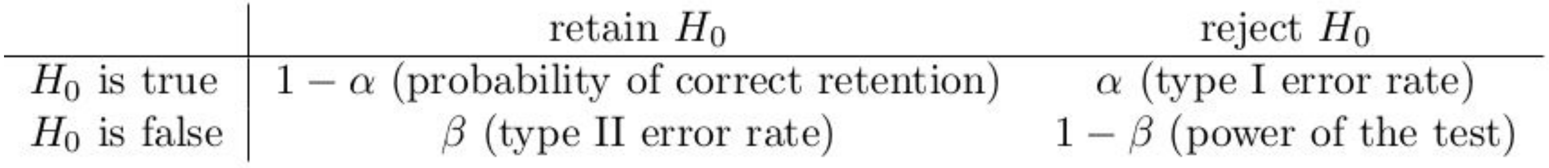
Significance level:

Threshold that controls the proportion of false positives that we tolerate.

Refresh 2: mistakes

Any time that we make a decision about whether to trust the null hypothesis or not, we are subject to committing errors!

- ➢ <u>Type I error</u>: If we reject H$_0$ (accept H$_a$) when in fact H $_0$ is true.
- ➢ <u>Type II error</u>: if we accept H$_0$ (reject H$_a$) when in fact H$_a$ is true.



	retain H_0	reject H_0
H_0 is true	$1-\alpha$ (probability of correct retention)	α (type I error rate)
H_0 is false	β (type II error rate)	$1-\beta$ (power of the test)

Refresh 2: mistakes

Any time that we make a decision about whether to trust the null hypothesis or not, we are subject to committing errors!

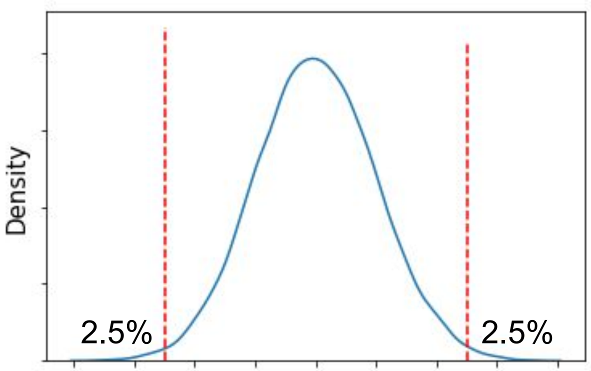
- ➢ Type I error: If we reject H_0 (accept H_a) when in fact H_0 is true.
- ➢ Type II error: if we accept H_0 (reject H_a) when in fact H_a is true.

unbiasedresearch.blogspot.com

Refresh 3: two-sided tests and error adjustment

- Many parametric tests have **directionality** (e.g. t-test).

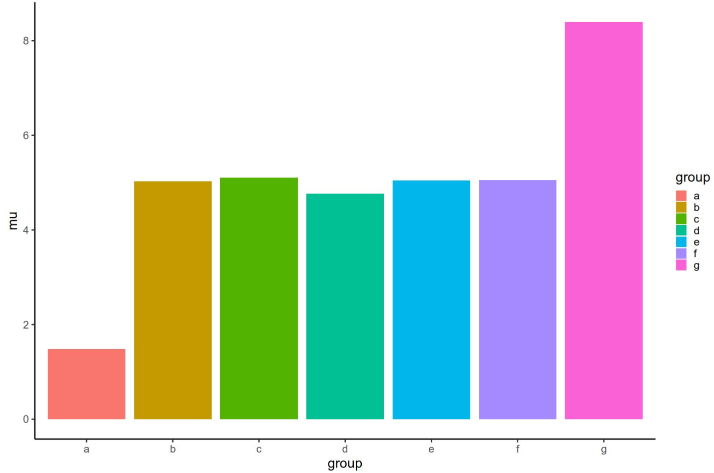
- We need to **adjust** the amount of critical region depending on whether we include one or both tails.



This keeps our total Type I error rate the same.

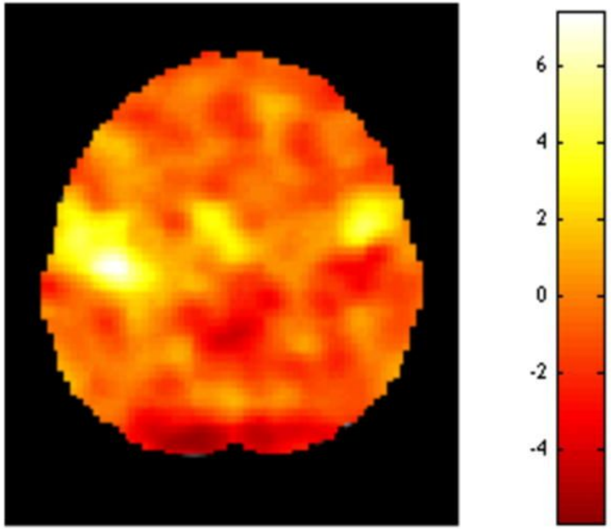
The multiple testing problem

- In modern research, we often need to perform multiple hypothesis tests at the same time
    - Example 1: ANOVA/Kruskal-Wallis post-hoc analysis.
    - Example 2: Brain voxels that are significantly activated.
    - Example 3: Association between many variables.



The multiple testing problem

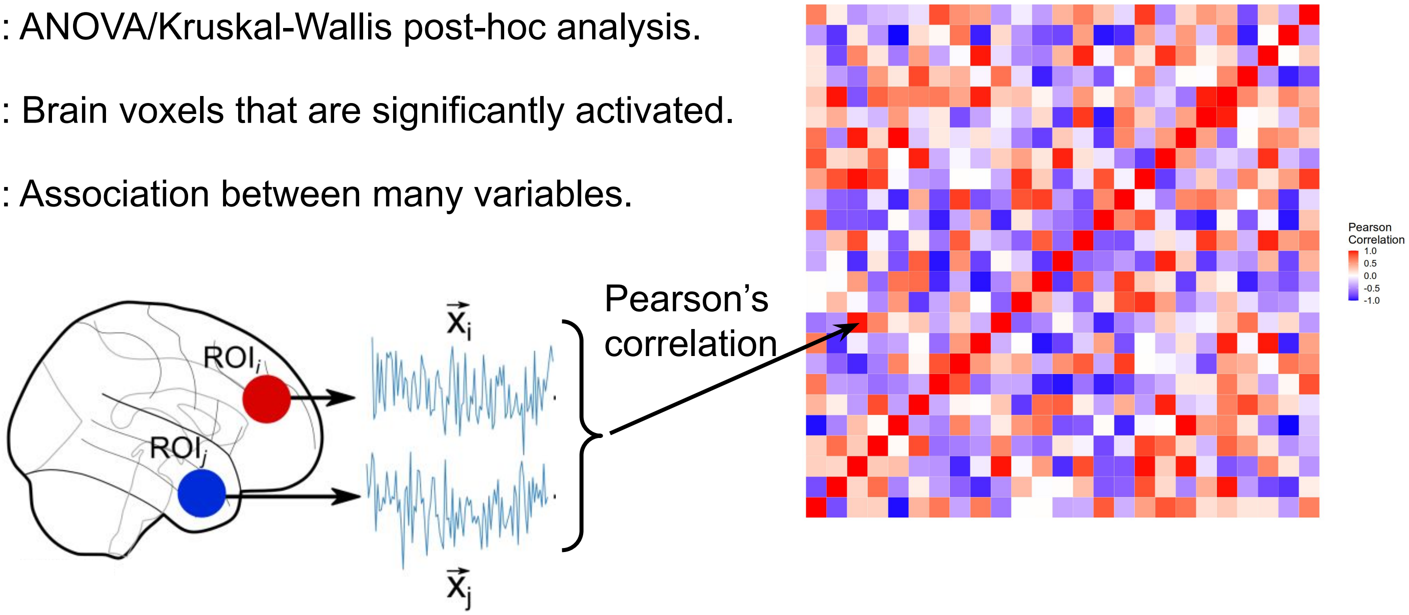
- In modern research, we often need to perform multiple hypothesis tests at the same time
  - Example 1: ANOVA/Kruskal-Wallis post-hoc analysis.
  - Example 2: Brain voxels that are significantly activated.
  - Example 3: Association between many variables.



Lindquist 2015

The multiple testing problem

- In modern research, we often need to perform multiple hypothesis tests at the same time
  - Example 1: ANOVA/Kruskal-Wallis post-hoc analysis.
  - Example 2: Brain voxels that are significantly activated.
  - Example 3: Association between many variables.



Pearson's correlation

The multiple testing problem

- In modern research, we often need to perform multiple hypothesis tests at the same time.

- In this case, we deal with a ***family of tests***: a set of tests that are related to the same research goal and for which significance statements are reached.

The multiple testing problem

- In modern research, we often need to perform multiple hypothesis tests at the same time.

- In this case, we deal with a ***family of tests***: a set of tests that are related to the same research goal and for which significance statements are reached.

- Multiple testing **inflates** the number of **false positives**: the bigger the family is, the greater the type I error rate (see tutorial).

The multiple testing problem

- In modern research, we often need to perform multiple hypothesis tests at the same time.

- In this case, we deal with a ***family of tests***: a set of tests that are related to the same research goal and for which significance statements are reached.

- Multiple testing **inflates** the number of **false positives**: the bigger the family is, the greater the type I error rate (see tutorial).

- We need to choose an **appropriate significance threshold** to account for this increase of false positives (see tutorial).

The multiple testing problem

Two main techniques to adjust the Type I error rate when performing *m* hypotheses depending on what to control for:

1. The **family-wise error rate**, FWER = P(V>1).
2. The **false discovery rate**, FDR = E(V/R).

	Null was true	**Null was not true**	**Total**
Null rejected	V	S	R
Null not rejected	U	T	m - R
Total	m_0	$m - m_0$	m

Family-wise error rate (FWER)

- The family-wise error rate (FWER): under the null hypothesis, the probability of getting one or more Type I errors in a family of tests:

$$FWER = P(V>1)$$

- FWER correction was initially important in the context of **post-hoc** comparisons in **ANOVA**: Tukey's procedure, Scheffé’s procedure, Dunnett's correction.

- The following are methods that can be used in more broadly situations:
  - Bonferroni correction.
  - Sequential Holm’s method.
  - Hochberg.

Family-wise error rate (FWER)

- The family-wise error rate (FWER): under the null hypothesis, the probability of getting one or more Type I errors in a family of tests:

$$\text{FWER} = P(V>1)$$

- FWER correction was initially important in the context of **post-hoc** comparisons in **ANOVA**: Tukey's procedure, Scheffé’s procedure, Dunnett's correction.

- The following are methods that can be used in more broadly situations:
  - **Bonferroni correction.**
  - Sequential Holm’s method.
  - Hochberg.

Bonferroni correction

- If we have *m* independent tests, the probability of obtaining at least one false positive is the following:

$$\mathrm{FWER} = 1 - (1 - \alpha)^m$$

- The bonferroni correction controls for this by using *α’* instead of *α*, where *α’* is the nominal type I error divided by the *m* number of tests.

$$\alpha' = \frac{\alpha}{m}$$

- This correction keeps the FWER under a desired threshold (e.g. 0.05, see tutorial).

Family-wise error correction: caveats

- Controlling for the FWER is appropriate when you are concerned about getting **any** false positives.
- This kind of correction might be **too restrictive** and lead to a **decrease** in **statistical power** (i.e. Type I error vs Type II error trade-off, see tutorial).
- Sometimes we can live with having a certain number of false positives.
- In these cases, controlling instead for the false discovery rate (FDR) might be more appropriate.

False discovery rate (FDR) correction

- In contrast to FWER that controls the probability of getting any false positive, the False Discovery Rate (FDR) controls the **fraction of false positive** results among all the rejected ones.

$$\mathrm{FDR} = E(V/R|R > 0) \cdot P(R > 0)$$

- Controlling for FDR ensures that on average the FDR is smaller than or equal to a specific threshold q which lies between 0 and 1

Benjamini-Hochberg (BH) procedure

1. Select desired limit *q* on FDR (e.g. 0.05).
2. Rank the p-values in ascending order from smallest to largest.
3. Adjust each p-values as follows:

$$\text{adj. } p_i = \frac{p_i \cdot m}{rank_i}$$

4. Determine the **largest** rank *i* for which the adj. p_i is less than or equal to the FDR threshold *q*.
5. Reject all the tests below this rank.

FDR correction: remarks

- If **all** null hypotheses are **true**, the FDR is **equivalent** to the FWER.
- When we control the FWER we are also controlling the FDR.
- Controlling the FDR only is then **less stringent** and lead to an **increase in statistical power** (See tutorial).
- As a result, It can be appropriate, for example, when one is more interested in discovering new findings.

Multiple testing correction in R

- For the most broadly situations (see tutorial):

```
p.adjust(our vector of p-values, correction method)
```

- In the context of ANOVA/Kruskal-wallis post-hoc analysis:

```
pairwise.t.test( … )
pairwise.wilcox.test( … )
TukeyHSD( one anova object )
```

Key ideas

- The number of statistical tests performed affects the chances of getting a false positive.
- We need to adjust our p-values to have our errors under control.
- We control the family-wise error rate (FWER) if we are worried about observing any false positives
- We control the false discovery rate (FDR) in exploratory scenarios where we can live with false positives.