



BROWN
Computer Science

CS1951A: Data Science

Lecture 13: Multiple Hypothesis Testing

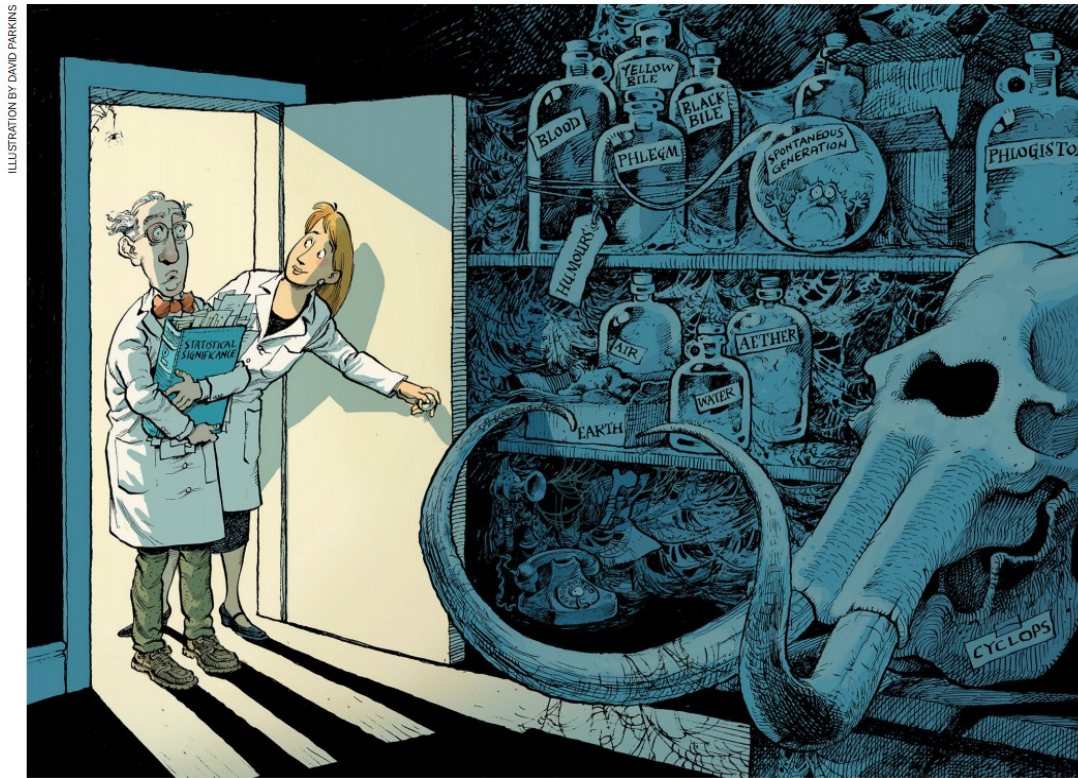
Lorenzo De Stefani

Spring 2021

Overview

- Problems with hypothesis testing
- p-hacking
- Publication Bias
- Multiple hypothesis testing
- Family Wise Error Rate
- False Discovery Rate

Interpreting p-values



Retire statistical significance

Valentin Amrhein, Sander Greenland, Blake McShane and more than 800 signatories call for an end to hyped claims and the dismissal of possibly crucial effects.

<https://www.nature.com/articles/d41586-019-00857-9>

Interpreting p-values

“In my experience teaching many academic physicians, when physicians are presented with a single-sentence summary of a study that produced a surprising result with $P = 0.05$, the overwhelming majority will confidently state that there is a 95% or greater chance that the null hypothesis is incorrect.

This is an understandable but categorically wrong interpretation because the P value is calculated on the assumption that the null hypothesis is true. It cannot, therefore, be a direct measure of the probability that the null hypothesis is false. This logical error reinforces the mistaken notion that the data alone can tell us the probability that a hypothesis is true.”

Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. *Ann Intern Med.* 1999;130:995-1004.

Interpreting p-values

“In my experience teaching many academic physicians, when physicians are presented with a single-sentence summary of a study that produced a surprising result with $P = 0.05$, the overwhelming majority will confidently state that there is a 95% or greater chance that the null hypothesis is incorrect.

This is an understandable but categorically wrong interpretation because the P value is calculated on the assumption that the null hypothesis is true. It cannot, therefore, be a direct measure of the probability that the null hypothesis is false. This logical error reinforces the mistaken notion that the data alone can tell us the probability that a hypothesis is true.”

Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. *Ann Intern Med.* 1999;130:995-1004.

Interpreting p-values

“In my experience teaching many academic physicians, when physicians are presented with a single-sentence summary of a study that produced a surprising result with $P = 0.05$, the overwhelming majority will confidently state that there is a 95% or greater chance that the null hypothesis is incorrect.

This is an understandable but **categorically wrong interpretation because the P value is calculated on the assumption that the null hypothesis is true**. It cannot, therefore, be a direct measure of the probability that the null hypothesis is false. This logical error reinforces the mistaken notion that the data alone can tell us the probability that a hypothesis is true.”

Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. *Ann Intern Med.* 1999;130:995-1004.

Interpreting p-values

“In my experience teaching many academic physicians, when physicians are presented with a single-sentence summary of a study that produced a surprising result with $P = 0.05$, the overwhelming majority will confidently state that there is a 95% or greater chance that the null hypothesis is incorrect.

This is an understandable but categorically wrong interpretation because the P value is calculated on the assumption that the null hypothesis is true. **It cannot, therefore, be a direct measure of the probability that the null hypothesis is false.** This logical error reinforces the mistaken notion that the data alone can tell us the probability that a hypothesis is true.”

Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. *Ann Intern Med.* 1999;130:995-1004.

Interpreting p-values

“In my experience teaching many academic physicians, when physicians are presented with a single-sentence summary of a study that produced a surprising result with $P = 0.05$, the overwhelming majority will confidently state that there is a 95% or greater chance that the null hypothesis is incorrect.

This is an understandable but categorically wrong interpretation because the P value is calculated on the assumption that the null hypothesis is true. **It cannot, therefore, be a direct measure of the probability that the null hypothesis is false.** This logical error reinforces the mistaken notion that the data alone can tell us the probability that a hypothesis is true.”

Goodman SN. Toward evidence-based medical statistics. 1: The P value fallacy. *Ann Intern Med.* 1999;130:995-1004.

Interpreting P-Values

It is **NOT** the probability that the null or the alternative hypothesis are correct or incorrect

Probability of observing an effect equal to or more extreme than the one observed, **assuming the null hypothesis is true**

Question time 1

We run a test and we obtain a p-value $p=0.05$. This means that the null hypothesis has a 5% chance of being true

- a) Agree
- b) Disagree
- c) Cannot say

Question time 2

We run a test and we obtain a p-value $p=0.05$. If we reject the null, the probability of that being a wrong decision (thus assuming the null is true) is at most 5%

- a) Agree
- b) Disagree
- c) Cannot say

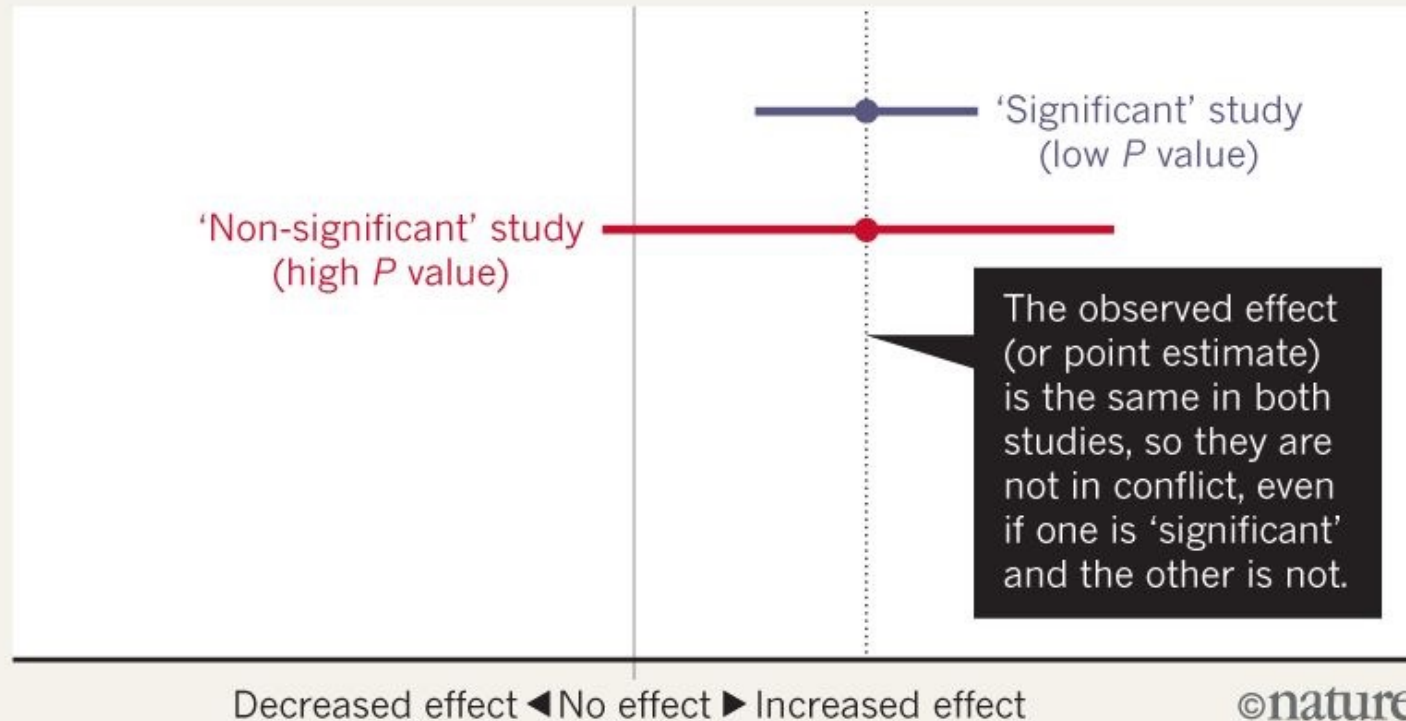
Question time 3

You read a study showing that a new drug leads to a significant decrease in cholesterol. You later read a **newer study** that shows that **there is a decrease in cholesterol but it is *not* statistically significant**. These studies are contradictory, **one of them must be wrong**.

- a) Agree
- b) Disagree
- c) Cannot say

BEWARE FALSE CONCLUSIONS

Studies currently dubbed 'statistically significant' and 'statistically non-significant' need not be contradictory, and such designations might cause genuine effects to be dismissed.



Question time 4

I test a new cancer treatment and find **a statistically significant decrease in tumor size for patients receiving the treatment** compared to a control group. I should prescribe this treatment to all of my patients now.

- a) Agree
- b) Disagree
- c) Cannot say

Question time 5

A p-value $p=0.05$ means that the probability of observing a statistical phenomena at least as extreme as the one seen in the data is at most 0.05 assuming the null-hypothesis is correct

- a) Agree
- b) Disagree
- c) Cannot say

p-values have problems

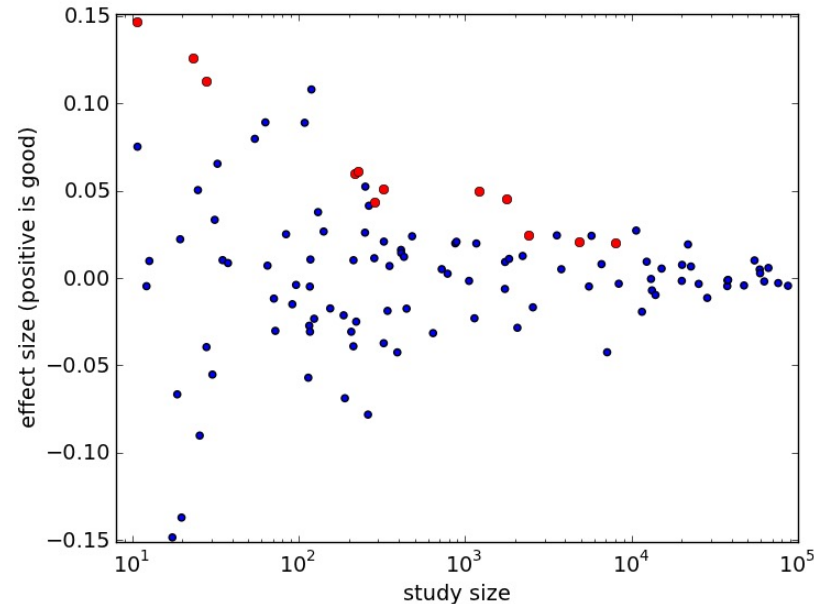
- Never ever use it **as the only measure or validation of statistical accuracy**
- Still done in many disciplines
- Several important resulting issues
 - Publication Bias
 - Overfitting due to abuse of adaptive data analysis
 - Multiple Hypothesis Fallacy

Background: Effect Size

$$\text{Effect size} = \frac{|[\text{Mean of experimental group}] - [\text{Mean of control group}]|}{\text{standard deviation}}$$

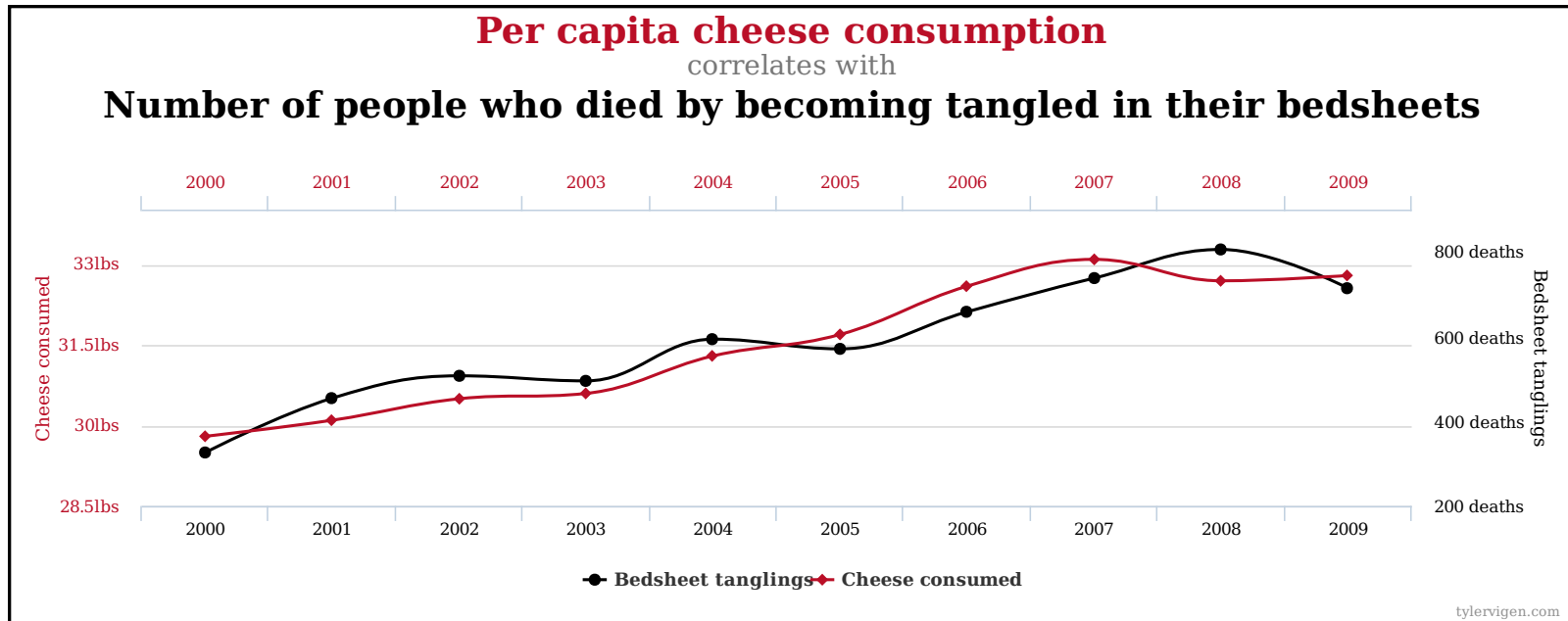
- The higher the effect size, **the stronger the apparent impact of the recommended method/procedure**
 - Something that measures **how “good” the result is**
 - Not **just “significant” – how significant?**
- Used prolifically in meta-analysis to combine results from multiple studies
 - Careful! Averaging results from different experiments can produce nonsense
- Caveat: Other definitions of effect size exist: odds-ratio, correlation coefficient

Publication Bias: the decline effect



- As the **study size increases**, the **effect size diminishes** (Rhine 1934)
- The largest the study size the lower the effect size
- Lower effect size → weaker **apparent statistical evidence** supporting the result

Spurious correlations



Spurious correlations

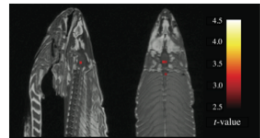
Subject. One mature Atlantic Salmon (*Salmo salar*) participated in the fMRI study. The salmon was approximately 18 inches long, weighed 3.8 lbs, and was not alive at the time of scanning.

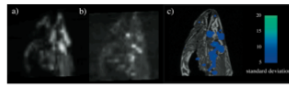
Task. The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence. The salmon was asked to determine what emotion the individual in the photo must have been experiencing.

Design. Stimuli were presented in a block design with each photo presented for 10 seconds followed by 12 seconds of rest. A total of 15 photos were displayed. Total scan time was 5.5 minutes.

Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction
Craig M. Bennett¹, Abigail A. Baird², Michael B. Miller¹, and George L. Wolford³
¹ Psychology Department, University of California Santa Barbara, Santa Barbara, CA; ² Department of Psychology, Vassar College, Poughkeepsie, NY; ³ Department of Psychological & Brain Sciences, Dartmouth College, Hanover, NH

INTRODUCTION
With the extreme dimensionality of functional neuroimaging data comes extreme risk for false positives. Across the 130,000 voxels in a typical fMRI volume the probability of a false positive is almost certain. Correction for

GLM RESULTS

A *t*-contrast was used to test for regions with significant BOLD signal change during the photo condition compared to rest. The parameters for this comparison were $t(131) > 3.15$, $p(\text{uncorrected}) < 0.001$, 3 voxel extent threshold.
Several active voxels were discovered in a cluster located within the salmon's brain cavity (Figure 1, see above). The size of this cluster was 81 mm³ with a cluster-level significance of $p = 0.001$. Due to the coarse resolution of the echo-planar image acquisition and the relatively small size of the salmon brain further discrimination between brain regions could not be completed. Out of a search volume of 8064 voxels a total of 16 voxels were significant.
Identical *t*-contrasts controlling the false discovery rate (FDR) and familywise error rate (FWER) were completed. These contrasts indicated no active voxels, even at relaxed statistical thresholds ($p = 0.25$).

VOXELWISE VARIABILITY

To examine the spatial configuration of false positives we completed a variability analysis of the fMRI timeseries. On a voxel-by-voxel basis we calculated the standard deviation of signal values across all 140 volumes.
We observed clustering of highly variable voxels into groups near areas of high voxel signal intensity. Figure 2a shows the mean EPI image for all 140 image volumes. Figure 2b shows the standard deviation values of each voxel. Figure 2c shows thresholded standard deviation values overlaid onto a high-resolution T₂-weighted image.
To investigate this effect in greater detail we conducted a Pearson correlation to examine the relationship between the signal in a voxel and its variability. There was a significant positive correlation between the mean voxel value and its variability over time ($r = 0.54$, $p < 0.001$). A scatterplot of mean voxel signal intensity against voxel standard deviation is presented to the right.

DISCUSSION
Can we conclude from this data that the salmon is engaging in the perspective-taking task? Certainly not. What we can determine is that random noise in the fMRI timeseries may yield spurious results if multiple comparisons are not controlled for. Adaptive methods for controlling the FDR and FWER are excellent options and are widely available in all major fMRI analysis packages. We argue that relying on standard statistical thresholds ($p < 0.001$) and low minimum cluster sizes ($k > 8$) is an ineffective control for multiple comparisons. We further argue that the vast majority of fMRI studies should be utilizing multiple comparisons correction as standard practice in the computation of their statistics.

REFERENCES
Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57:289-300.
Friston KJ, Worsley KJ, Frackowiak RSJ, Mazziotta JC, and Evans AC. (1994). Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, 1:214-220.

If you search long enough, you are almost guaranteed to find something that looks interesting and statistically significant

Testing multiple hypotheses

- Suppose we want to evaluate multiple null-hypotheses H_1, H_2, \dots, H_m at the same time

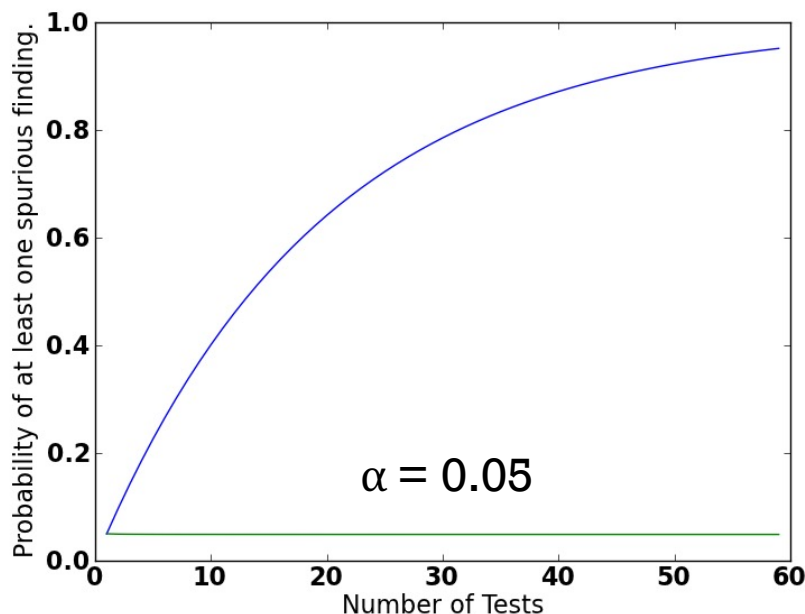
- V, U, S, T are **random variables**
- Only the number of rejections R and the number of null hypotheses non rejected $m - R$ are observable

Testing multiple hypotheses

- Our efforts so far have been towards avoiding **false positives**
 - For a given value α , our procedures guarantee that **the probability of a false positive is asymptotically controlled at level α**
 - We would rather not deem something as significant if it is, than missing some possible discovery
- Statistical procedures have **a bias towards conservativeness**
- Obtaining guarantees on **no False Negatives is very challenging**
 - Requires major assumptions on the data and the randomness of the observed distribution

Challenges in MHT

- Can we use the methods we are already familiar with?
- Suppose that we test each hypothesis
 - $P(\text{detecting an effect when there is none}) = \alpha = 0.05$
 - $P(\text{detecting an effect when it exists}) = 1 - \alpha$
 - $P(\text{detecting an effect when there is none on at least one out of } k \text{ experiments}) = 1 - (1 - \alpha)^k$



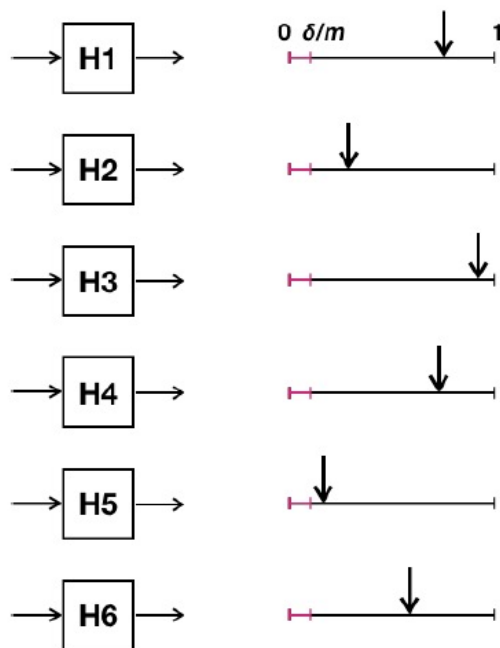
The probability of having at least one false positive increases **exponentially** with the number of tests

Family Wise Error Rate

- Given a family of hypotheses and a testing procedure the **Family Wise Error Rate (FWER)** is the probability of having **at least one false positive**
 - We want to guarantee **control of the FWER at a given level α**
 - Emphasis on controlling **false positives**
 - We need to opportunistically **correct the value α** to enforce control for each hypothesis

Bonferroni correction

- Controls FWER at level α
- Simply **divide the threshold α by the number of hypotheses**
- Decide whether to reject each null using the **corrected threshold $\alpha_c = \alpha/m$**



- Based on **union bound**
 - For any pair of events E_1, E_2
 - $P(E_1 \vee E_2) \leq P(E_1) + P(E_2)$
 - We are controlling the probability of **a wrong decision for each hypothesis at α/m**
- Hypotheses being tested may be **dependent** on each other

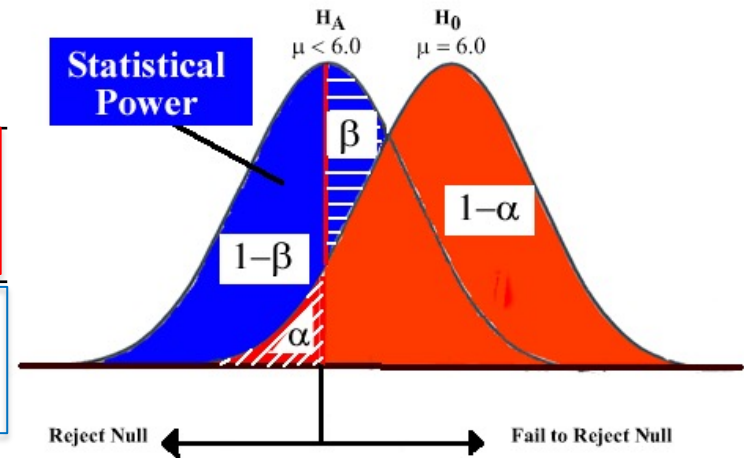
Bonferroni control procedure

1. Given null hypotheses H_1, H_2, \dots, H_m compute their p-values $p_1 \leq p_2 \leq \dots \leq p_m$
2. Sort the p-values $p_1 \leq p_2 \leq \dots \leq p_m$
3. Find the largest i such that $p_i \leq \alpha/m$
4. Reject all null hypotheses corresponding to the i p-values $p_1 \leq p_2 \leq \dots \leq p_i$, fail to reject all others
5. Guarantees **FWER control at level α**

$$\Pr(V > 0) \leq \alpha$$

Statistical power

	Null is true	Null is false
Reject	Type I error False Positive Probability $\leq \alpha$	Correct Confidence Probability $\geq 1 - \alpha$
Fail to reject	Correct True Negative Probability $\geq 1 - \beta$	Type II error False Negative Probability $\leq \beta$



- Statistical power $1 - \beta$ is the probability of not having false negatives
 - The higher, the less likely we are to fail to reject a null which is not true
 - The parameter β is generally unknown and it depends on the relation between the true distribution and the null H_0
 - How “different” they are
- There is a natural trade off between the Statistical power and Confidence level
 - The lower α , the more conservative the test (higher precision)
 - The higher α , the more likely to reject false null (higher recall)
- Can you think of a procedure which 100% guarantees no false discoveries?

Precision and recall

- **Precision of a test** $= \frac{\text{Correct rejections}}{\text{Total Rejections}} = \frac{S}{R}$
 - In FWER we want to control the probability of Precision < 1
- **Recall of a test** $= \frac{\text{Correct rejections}}{\text{Total False Null Hypotheses}} = \frac{S}{S+T}$
- Generally, **the higher the precision the lower the recall and vice versa**
- **F1 score** combines these two attributes

$$F_1 = \frac{2}{\text{recall}^{-1} + \text{precision}^{-1}} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

- Optimizing F_1 is hard
 - We need **assumptions** on the data generating model!

Trading off performance and assumptions

- The Bonferroni procedure is **correct and very general**
 - No assumptions on the hypotheses or their relations
 - VERY conservative procedure! (**Low recall**)
 - Can we do better?
- Improving recall will generally be achieved by **trading away some generality**
 - We will **add some assumptions to our model**
 - Such improvements in the recall are **mostly seen experimentally rather than analytically**

The Simes Procedure

- Used to **test the global null hypothesis**

$$H_0 = \cap H_{0,i}, \text{ for } i = 1, 2, \dots, m$$

You can think of the global null as saying **all of the null hypotheses are correct**

- Requires each null has a p-value $p_i \sim U(0,1)$ under $H_{0,i}$
- Requires p_i 's , and hence $H_{i,0}$, to be **independent!**
 - Actually, it sufficient for the **not to be negatively dependent.**

Examples

- We flip 1000 coins 20 times each. For each coin we are testing the nulls

$H_{0,i}$: the i – th coin is fair

- Are these independent? Yes!
- We have gathered the records of purchases at Walmart in 2019. We are interest in testing the correlation between items being purchased
 - *$H_{0,(i,j)}$: there is no correlation between purchasing item i and item j*
 - Are these independent? Generally, they are not!

Simes control procedure

1. Given hypotheses H_1, H_2, \dots, H_m compute their p-values
2. Sort the p-values $p_1 \leq p_2 \leq \dots \leq p_m$
3. Find $T_n = \min_i \left\{ \frac{p_i m}{i} \right\}$
Under H_0 and independence of the p_i 's, $T_n \sim U(0,1)$
4. If $T_n \leq \alpha$ reject the global null hypothesis
 $H_0 = \cap H_i$

Assuming **independence** of the p_i 's, Simes rejects H_0 with confidence α

Comparison

- Whenever Bonferroni rejects **at least one null hypothesis**, Simes always rejects the global null
- When the global null is rejected using Simes when can have different behaviors using Bonferroni:
 - Consider a scenario for which all of the $p_i = \alpha$
 - Bonferroni would not reject any of the m hypotheses
 - Instead, Simes would reject the global null

Comparison

- Is Simes procedure better than Bonferroni?
 - a) Yes
 - b) No
 - c) Bad Question
 - d) Apples to oranges

Weak control of FWER

- Simes guarantees **Weak Control of the FWER** under the global null hypothesis
- Vice versa, procedures that do not require this constraint (e.g., Bonferroni) are said to **Control FWER in the Strong Sense**
 - We omit the “Strong” denomination as it is the default desired control

Holm procedure

Holm's Step-Down procedure (1979)

1. Given hypotheses H_1, H_2, \dots, H_m compute their p-values
2. Sort the p-values $p_1 \leq p_2 \leq \dots \leq p_m$
3. Find the **minimum index i** such that $p_i > \frac{\alpha}{m-i+1}$
4. Reject all null hypotheses corresponding to the p-values $p_1 \leq p_2 \leq \dots \leq p_{i-1}$, fail to reject the remaining

The Holm procedure (strongly) controls FWER with confidence α

Holm vs Bonferroni

Holm's procedure is **strictly more powerful** than Bonferroni!

1. If a null hypothesis with p-value p_i is rejected by Bonferroni (thus, $p_i \leq \alpha/m$), then that hypothesis will also be rejected by Holm!
 2. Vice versa is generally **not true!**
 - Consider a scenario such that $p_i = \frac{\alpha}{m-i+1}$
 - Bonferroni would only reject the null corresponding to p_1
 - Holm would reject all m nulls
- **Same precision under the same set of assumptions**
 - **Holm has, generally, much stronger recall**

Hochberg's step-up procedure (1988)

1. Given hypotheses H_1, H_2, \dots, H_m compute their p-values
2. Sort the p-values $p_1 \leq p_2 \leq \dots \leq p_m$
3. Find the **largest index i such that $p_i \leq \frac{\alpha}{m-i+1}$**
4. Reject all null hypotheses corresponding to the p-values $p_1 \leq p_2 \leq \dots \leq p_i$, fail to reject the remaining

The Hochberg procedure controls FWER with confidence α **assuming non-negative dependence of the hypotheses**

Holm vs Hochberg

- Holm **does not require non-negative dependence assumption but Hochberg does**
 - Holm's procedure is based on the Bonferroni correction, while Hochberg is based on Simes method
 - Mismatch in the scope must be accounted for in the comparison
- Assuming non-negative dependence of the hypotheses we have that Hochberg's procedure is **statistically more powerful** than Holm
 1. If a null hypothesis with p-value p_i is rejected by Holm (thus, $p_i \leq \alpha / (m - k - i)$), the that hypothesis will also be rejected by Hochberg!
 2. Vice versa is generally not true!
 - Consider a scenario such that all $p_i = \frac{\alpha}{m-k+1}$, where $1 < k \leq m$
 - Holm would not reject any null hypothesis
 - Holm would reject the first the hypotheses corresponding to the first k p-values in the sorted list
- Same precision under non –negative dependence assumption
- Hochberg has –generally- stronger recall

Is FWER the right kind of control?

- FWER is a **very** conservative criterion
 - We want to control the probability of making even a single "mistake" **regardless of how many hypotheses would be rejected**
- In some cases, we would be inclined **to accept some mistakes if that would allow to improve the recall**

False Discovery Rate

- We want to control the **expected ratio of the number of wrong rejections to total**
- $Q \triangleq \frac{V}{R}$ if $R > 0$, and $Q \triangleq 0$ if $R = 0$.
- $FDR = E[Q] = E\left[\frac{V}{R} | R > 0\right] P(R > 0)$
- As done for FWER, we select a value $\alpha \in (0,1]$ to the desired critical level control for FDR
 - Typical values 0.01, 0.05,...
- A procedure that guarantees $E[Q] \leq \alpha$ is said to **control FDR at level α** .

The Benjamini-Hochberg (BH) procedure

1. Given hypotheses H_1, H_2, \dots, H_m compute their p-values
2. Sort the p-values $p_1 \leq p_2 \leq \dots \leq p_m$, we assume the tests used to obtain them are independent
3. Find the **largest index i such that $p_i \leq \frac{i}{m} \alpha$**
4. Reject all null hypotheses corresponding to the p-values $p_1 \leq p_2 \leq \dots \leq p_{i-1}$, fail to reject the remaining

Assuming the hypotheses being tested are independent the BH procedure controls FDR at level α

The Benjamini-Yakutieli (BY) procedure

The BY procedure generalizes BH control of FDR for dependence

1. Given hypotheses H_1, H_2, \dots, H_m compute their p-values
2. Sort the p-values $p_1 \leq p_2 \leq \dots \leq p_m$,
3. Find the **largest index i such that $p_i \leq \frac{i}{m c(m)} \alpha$** where
 - $c(m) = 1$ if the hypotheses are independent or positively dependent – This is just the BH procedure
 - $c(m) = \sum_{i=1}^m 1/i$ - the Harmonic number – for arbitrary dependence
4. Reject all null hypotheses corresponding to the p-values $p_1 \leq p_2 \leq \dots \leq p_{i-1}$, fail to reject the remaining

FDR and recall

- The main difference between FWER and FDR is that in the latter **we are willing to tolerate some imprecision (i.e., false positives) in order to improve recall**
- Whether FDR or FWER are the “right” type of control is highly context-dependent!

There is a lot more!

- Literature is rich in many variations of FWER and FDR
 - FWER1, FWER2,
 - pFDR, mFDR, MFDR,...
 - Or altogether different notions! (NFDR, sFDR,...)
- Many other control procedures
 - Sidak's, Dunnett's, Resampling, Bootstrapping
- “Historic reasons” and application dependent
- Assert the level of control that “**you can claim**”