# Tutorial #3

Lyssa Buissereth
Liam Shamir
Michael Russo

# Format of this Presentation

We will introduce a statistical test, show a sample hypothesis that you may want to answer, and finally sample calculations.

***Always pick your tests when forming your hypothesis and before data collection.*** It helps prevent the temptation of switching tests after collecting data to force a conclusion.

# Dataset Used - QIIME's Moving Pictures

[1] Moving Pictures Tutorial From QIIME:
https://docs.qiime2.org/2020.2/tutorials/moving-pictures/

[2] Tutorial's Github Repo (provided by Zhengqiao):
https://github.com/EESI/qiime2_tutorial

This dataset consists of "**human microbiome samples from two individuals at four body sites at five timepoints, the first of which immediately followed antibiotic usage**" [1]

After getting Moving Pictures dataset ⇒ Convert it into a **.tsv (tab-separated values)** file so you can view it in Excel/MATLAB/viewer of choice

# t-test

# t-test: generally speaking

- A generalized t-test is used to compare the estimated value of a parameter to the hypothesized value of that parameter

t = [(parameter estimate) - (parameter  value (hypothesized))] / (standard error of parameter estimate)

- But it is most commonly used to compare means

t = [(mean est.) - (mean (hypothesized))] / (standard error of mean est.)

- 2 assumptions:
  - Population is *normally distributed*
  - Samples are taken *randomly*

# t-test: compared to Z-tests

- It functions much the same way as a Z-test, but is used when the variance of the population from which the sample came is unknown (i.e. you are analyzing a sample)

Z = [(parameter est.) - (parameter  value (hypothesized))] / (standard error of parameter)

t = [(parameter  est.) - (parameter value (hypothesized))] / (standard error of parameter est.)

- When the sample size increases substantially enough, the population's variance being unknown becomes less of an issue, and one can simply perform a standard Z-test
    - This is typically agreed upon as taking place when the sample size, n, becomes ≥ 30 [pg. 136, *Biostatistical Analysis 5th Ed.*, Zar]

# t-test: 2 sample

- Comparing 2 sample means is a bit different from comparing one sample mean to the population's mean
- Need to adjust the standard deviation calculation
- If unequal variances are assumed between the samples, std becomes

std = mean1 - mean2 / sqrt ((sample_variance1 / sample_size1) + (sample_variance2 / sample_size2))

- If equal variances are assumed, the test becomes:

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{s_p\sqrt{1/N_1 + 1/N_2}}$$

where

$$s_p^2 = \frac{(N_1-1)s_1^2 + (N_2-1)s_2^2}{N_1+N_2-2}$$

https://www.itl.nist.gov/div898/handbook/eda/section3/eda353.htm

# t-test: hypotheses

- Null hypothesis - **$H_0$: parameter est. ≥ OR = OR ≤ parameter value (hypothesized)**
- Alternative hypothesis - **$H_A$: parameter est. > OR ≠ OR < parameter value (hypothesized)**
  - *Note: by convention, the null hypothesis always uses the equality version of a comparator (≥, =, ≤)*

- Can compare a sample to the population or a sample to another sample
  - $H_0$: parameter est. = parameter value hypothesized
  - $H_0$: parameter est. 1 = parameter est. 2

- Can set expectations of a where you expect the parameter estimate to be distributed
  - To either side of the parameter value hypothesized ⇒ $H_A$: parameter est. ≠ parameter value hypothesized
  - Larger than the parameter value hypothesized ⇒ $H_A$: parameter est. > parameter value hypothesized
  - Smaller than the parameter value hypothesized ⇒ $H_A$: parameter est. < parameter value hypothesized

# t-test: calculations

- To determine if the null hypothesis is rejected or not, need to calculate a threshold ("critical") value

- $t_{CRIT}$ is taken from (or interpolated using) a "t table" whose parameters are
  - Alpha (α) value (prob. of finding parameter est. this far away from the parameter value hypoth.)
  - Number of tails ($H_A$ includes a ≠ ⇒ 2 tails, $H_A$ includes > *OR* < ⇒ 1 tail)
  - Degrees of freedom (*df* = sample size - 1)

If $|t_{CALC}| \geq |t_{CRIT}| = t_{\alpha\ (\text{tail \#}),\ df}$ then $H_0$ is *rejected*

# Hypothesis #1 (for t test)

“The average count of bacteria on the left hand of subject 1 is statistically significantly different from the average count of bacteria on the right hand of subject 1”

- No mention of which hand should have a higher or lower count ⇒ **2 tail**
- Comparing **two sample** means
- **Assuming equal variances** between samples (because the samples came from the same individual)

$H_0$: mean(Left Palm Count(day 1)) = mean(Right Palm Count(day 1))
$H_A$: mean(Left Palm Count(day 1)) ≠ mean(Right Palm Count(day 1))

To make sure I can test my hypothesis, I will need to ensure:

- Random sampling
- Normal distribution ⇒ Or gather more than 30 samples

# Data Collection

Col's C-K are the palm samples from subject 1 on every checkup day.

Col. A has each bacteria's taxonomy, col. B has IDs called the OTUs for each bacteria, and the cell values are counts of that bacteria in the sample

| | A | B | C | D | E | F | G | H | I | J | K |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | # Constructed from biom file | lp 1 | lp 1 | lp 1 | lp 1 | rp 1 | rp 1 | rp 1 | rp 1 | rp 1 |
| 2 | | #OTU ID | L2S155 | L2S175 | L2S204 | L2S222 | L3S242 | L3S294 | L3S313 | L3S341 | L3S360 |
| 3 | k__Bacteria; p__Pr | 01173073677a287321be3484fbed0007 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | k__Bacteria; p__Ve | 0160e14a78b18b903618f11bc732746e | 0 | 0 | 0 | 0 | 23 | 0 | 0 | 0 | 0 |
| 5 | k__Bacteria; p__Pr | 01b99cb344ed2530f7d80897ffe257a9 | 30 | 0 | 0 | 0 | 0 | 11 | 0 | 13 | 0 |
| 6 | k__Bacteria; p__Fir | 01ce91fd8dbecf637eb5e67cdab5c5aa | 2 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 |
| 7 | k__Bacteria; p__Fir | 01e0b7ac306895be84179f2715af269b | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 8 | k__Bacteria; p__Ba | 025dd300b0ccbd9d5898969c2a1ab138 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 9 | k__Bacteria; p__Fir | 02878fe3ccc81d4c884ca5574178d6a0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 10 | k__Bacteria; p__Ba | 029cc71dc93341de90188b686798aa0d | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 11 | k__Bacteria; p__Fir | 02ef9a59d6da8b642271166d3ffd1b52 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 12 | k__Bacteria; p__Ba | 0305a4993ecf2d8ef4149fdfc7592603 | 0 | 0 | 0 | 0 | 28 | 0 | 0 | 0 | 0 |
| 13 | k__Bacteria; p__Ba | 0310f41e594c49368dde5c5993a7a5d0 | 0 | 0 | 14 | 0 | 0 | 0 | 0 | 0 | 0 |
| 14 | k__Bacteria; p__Ba | 0316af109bb877b5c7a85ddecd6dfd1e | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 15 | k__Bacteria; p__Ac | 03178a409a41853ea09a94055e138e19 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 16 | k__Archaea; p__Cr | 033511c7ff4fe93866075cfb9129aa3b | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 17 | k__Bacteria; p__Fir | 0335b1664150f1e151340b1450eae898 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 18 | k__Bacteria; p__Fir | 047b7fb62a5e9d2711e639ae1cb1519a | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 19 | k__Bacteria; p__Ba | 04c7e0ea3038f942f5a28778a74cd1c0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 20 | k__Bacteria; p__Fir | 04fd81a94c775a5906f2d92cf0548e4d | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 21 | k__Bacteria; p__Fir | 052128d7d424728578efe7852b0afd0d | 0 | 0 | 0 | 0 | 8 | 0 | 0 | 0 | 0 |
| 22 | k__Bacteria; p__Pr | 0524cb6dcc8f60dafc09c46911c9d0ac | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 9 |
| 23 | k__Bacteria; p__Fir | 059408ca99f1001f64e6df1e06fb951d | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 24 | k__Bacteria; p__Pr | 063fa4f5a6fa894f7f8243468324ac5e | 0 | 15 | 0 | 21 | 0 | 0 | 0 | 0 | 0 |
| 25 | k__Bacteria; p__Pr | 06845c67bc4203081a981200f33e87eb | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

*Dear reader,*

*Pay no attention to the row colors. I simply forgot to unhighlight them*

# Calculations

```
[h,p]=ttest2(L(:,1),R(:,1)) % comparing left & right palms on sampling day 1
```

```
>> [h,p]

ans =

    1.0000    0.0047
```

MATLAB's ttest2 automatically computes a 2 sample t-test and compares it to $t_{CRIT}$ assuming $\alpha = 0.05$

h = null **h**ypothesis rejection.

- h=1 ⇒ null hypothesis is rejected
- h=0 ⇒ null hypothesis is not rejected

p = **p**-value

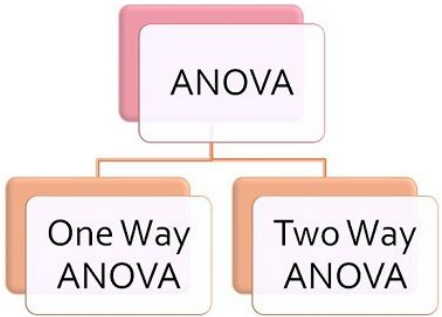
# ANOVA (one/two way)

# ANOVA

- **An**alysis **o**f **Var**iance
- **“One” way or “two” way =** refers to the number of independent variables in the test
- Statistical test used to analyze the difference between the means of more than two groups.

Assumptions:

1. Data is normally distributed
2. Homogeneity  of variances






Marshall, E., Owen, A., & Gillard, J. (n.d.). *The Statistics Tutor’s Quick Guide to Commonly Used Statistical Tests.* Retrieved from www.statstutor.ac.uk

# One-way ANOVA

**Dependent variable**: Continuous
**Independent variable**: Categorical (at least 3 categories)

**Application**: Detect the difference in means of 3 or more independent groups.

- Can be thought of as an extension of the t-test.

- ANOVA uses the ratio of the between group variance to the within group variance to decide whether there are statistically significant differences between the groups or not.



Marshall, E., Owen, A., & Gillard, J. (n.d.). *The Statistics Tutor's Quick Guide to Commonly Used Statistical Tests.* Retrieved from www.statstutor.ac.uk

# One-way ANOVA (cont'd)

Interpretation:

- ANOVA tests null hypothesis **"H0: all group means are equal"** using an F-test.
- The p-value concludes whether or not there is at least one pairwise difference.
    - Null hypothesis (H0) of ANOVA is that there is no difference among group means.
    - If the p-value < 0.05, we reject H0 and conclude that there is a significant difference between at least one pair of means.



Bevans, R. (2020, March 6 ). *An introduction to the one-way ANOVA*, Retrieved from Scribbr: https://www.scribbr.com/statistics/one-way-anova/
Marshall, E., Owen, A., & Gillard, J. (n.d.). *The Statistics Tutor's Quick Guide to Commonly Used Statistical Tests.* Retrieved from www.statstutor.ac.uk

# Two-way ANOVA

**Dependent variable**: Continuous
**Independent variables**: Two categorical (2+ levels within each)

**Application**: : Comparing means for combinations of two independent categorical variables (factors)

- “Double-testing” that same group
- E.g. Testing one set of individuals before and after taking a medication to see if the medication worked or not



Marshall, E., Owen, A., & Gillard, J. (n.d.). *The Statistics Tutor’s Quick Guide to Commonly Used Statistical Tests.* Retrieved from www.statstutor.ac.uk

# Two-way ANOVA (cont'd)

Interpretation:

There are three sets of hypothesis with a two-way ANOVA.

**H0 for each set** is as follows:

- The population means of the first factor are equal – equivalent to a one-way ANOVA for the **row factor**.

- The population means of the second factor are equal – equivalent to a one-way ANOVA for the **column factor**.

- There is no interaction between the two factors – no interaction between columns (data sets) and rows.
    - any systematic differences between columns are the same for each row and that any systematic differences between rows are the same for each column



Marshall, E., Owen, A., & Gillard, J. (n.d.). *The Statistics Tutor's Quick Guide to Commonly Used Statistical Tests.* Retrieved from www.statstutor.ac.uk

# ANOVA Overview

| Number of Independent Variables | ANOVA Used | Example |
|---|---|---|
| *One* | One-way ANOVA | Independent:<br>*-coffee brand*<br>Dependent:<br>-heart rate |
| *Two or more* | Two-way ANOVA | Independent:<br>*-coffee brand*<br>*-anxiety level*<br>Dependent:<br>-heart rate |

*Image cred: socratic.org, 2017*

# TEST 1

Question: “Is there a difference in microbiome populations between two body sites - **LEFT** and **RIGHT** palms?”

- Relative abundance of microbe counts (using OTU Table) obtained from two subjects.
- Presence (counts) of 30 microbes considered.
- Because hands are constantly interacting, expecting the microbes to be interchangeable between these two locations.
- One-way ANOVA

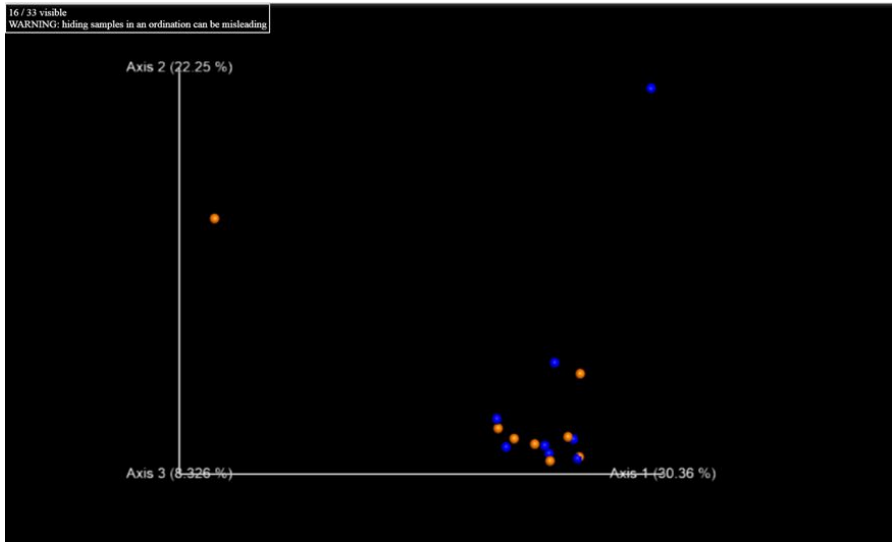
# ANOVA1 results supports PCOA visual

P value (<.05) allows us to reject the hypothesis that there is variation between microbes in palms.

File Edit View Insert Tools Desktop Window Help

**ANOVA Table**

| Source | SS | df | MS | F | Prob>F |
|---|---|---|---|---|---|
| Groups | 0.04049 | 29 | 0.0014 | 3.5 | 0.0005 |
| Error | 0.01198 | 30 | 0.0004 | | |
| Total | 0.05247 | 59 | | | |




PCOA shows microbes to be clustered, therefore samples are similar.

LEFT PALM
RIGHT PALM

# TEST 2

"What factor impacts the microbe variation? **Body site** or **time**?"

- **Body sites**: tongue and gut
- **Time**: immediate start of antibiotic and 140 days of taking antibiotics
- Presence (counts) of 20 microbes considered.
- Each item to be classified in two ways, as opposed to only one way.
  - Column, row (interaction)
- Two-way ANOVA
  - Partitions the overall variance of the outcome variable into three components to test the three null hypotheses.

| | COLUMN | COLUMN |
|---|---|---|
| ROW | Tongue - Day 1 | Tongue - 140 days |
| ROW | Gut - Day 1 | Gut - 140 days |

# ANOVA2 Results

| | COLUMN | COLUMN |
|---|---|---|
| ROW | Tongue - Day 1 | Tongue - 140 days |
| ROW | Gut - Day 1 | Gut - 140 days |

*null hypothesis = no interaction/mean is the same

File Edit View Insert Tools Desktop Window Help

**ANOVA Table**

| Source | SS | df | MS | F | Prob>F |
|---|---|---|---|---|---|
| Columns | 0.00014 | 1 | 0.00014 | 0.04 | 0.8372 |
| Rows | 0.00451 | 1 | 0.00451 | 1.39 | 0.2425 |
| Interaction | 0.00041 | 1 | 0.00041 | 0.13 | 0.7225 |
| Error | 0.24705 | 76 | 0.00325 | | |
| Total | 0.25212 | 79 | | | |

1. Variability among columns.
2. Variability among rows.
3. Interactions between row and column.
   - These are differences between rows that are not the same at each column, equivalent to variation between columns that is not the same at each row.

- These results indicate that the differences that are seen are due to chance - 83%, 24% and 72%
- On average, the treatment effect is indistinguishable from random variation.



David J. Sheskin. Handbook of Parametric and Nonparametric Statistical Procedures: Third Edition IBSN:1584884401.

# Dataset Used -

[1] **Multivariate ANOVA (MANOVA) -- Notes and R Code:**
https://gaopinghuang0.github.io/2017/11/20/MANOVA-notes-and-R-code#35-contrasts

The effects of cognitive behavior therapy (CBT) on obsessive compulsive disorder (OCD). Two dependent variables (DV1 and DV2) are considered: the occurrence of obsession-related behaviors (**Actions**) and the occurrence of obsession-related cognitions (**Thoughts**). OCD sufferers are grouped into three conditions: with CBT, with behavior therapy (BT), and with no-treatment (NT). [1]

# MANOVA

# Multivariate **ANOVA**

- Only ***difference*** being that MANOVA has **TWO** (or more) dependent variables, not **one**.
- MANOVA uses the covariance between means

# MANOVA

# WAIT

- Previously mentioned data contains only **one** dependent variable: bacterial occurrence value, and in order to perform MANOVA, **TWO** or more dependent variables are needed

# MANOVA

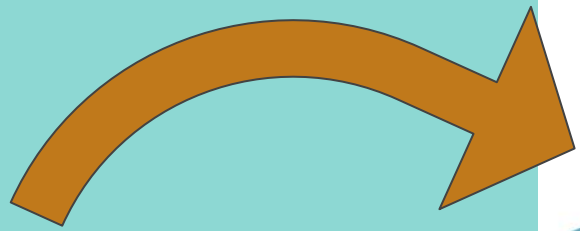
## Example of Suitable Data

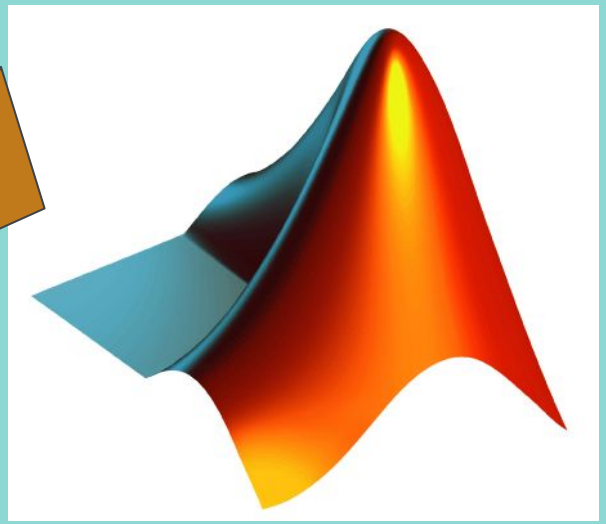
Suppose we want to know whether higher education (college, high school) attendance affects men and women in their professional development. Also assume we have each person's salary, duration of employment, number of promotions, highest schooling level, and job satisfaction.

| Dependent | Independent |
|---|---|
| Salary | Gender |
| Duration of Employment | Highest Schooling Level |
| Number of Promotions | |
| Job Satisfaction | |

# MANOVA - Example

```
Group,Actions,Thoughts
CBT,5,14
CBT,5,11
CBT,4,16
CBT,4,13
CBT,5,12
CBT,3,14
CBT,7,12
CBT,6,15
CBT,6,16
CBT,4,11
BT,4,14
BT,4,15
BT,1,13
BT,1,14
BT,4,15
BT,6,19
BT,5,13
BT,5,18
BT,2,14
BT,5,17
No Treatment Control,4,13
No Treatment Control,5,15
No Treatment Control,5,14
No Treatment Control,4,14
No Treatment Control,6,13
No Treatment Control,4,20
No Treatment Control,7,13
No Treatment Control,4,16
No Treatment Control,6,14
No Treatment Control,5,18
```





# MANOVA - Example

| Dependent | Independent |
| --- | --- |
| Actions | Group |
| Thoughts | |

```
r = fitrm(OCD1, 'Actions-Thoughts ~ Group ') ;
[manovatbl,A,C,D] = manova(r)
```

The **fitrm** function uses a table (OCD1) and then a grouping notation to delineate (based on table headers) the dependent and independent variables to produce a MatLab RepeatedMeasuresModel class object that can be used by the manova function.

# MANOVA - Example

| Within | Between | Statistic | Value | F | RSquare | df1 | df2 | pValue |
|---|---|---|---|---|---|---|---|---|
| Constant | (Intercept) | Pillai | 0.95329 | 551.02 | 0.95329 | 1 | 27 | 1.6946e-19 |
| Constant | (Intercept) | Wilks | 0.046711 | 551.02 | 0.95329 | 1 | 27 | 1.6946e-19 |
| Constant | (Intercept) | Hotelling | 20.408 | 551.02 | 0.95329 | 1 | 27 | 1.6946e-19 |
| Constant | (Intercept) | Roy | 20.408 | 551.02 | 0.95329 | 1 | 27 | 1.6946e-19 |
| Constant | Group | Pillai | 0.23438 | 4.1327 | 0.23438 | 2 | 27 | 0.027178 |
| Constant | Group | Wilks | 0.76562 | 4.1327 | 0.23438 | 2 | 27 | 0.027178 |
| Constant | Group | Hotelling | 0.30612 | 4.1327 | 0.23438 | 2 | 27 | 0.027178 |
| Constant | Group | Roy | 0.30612 | 4.1327 | 0.23438 | 2 | 27 | 0.027178 |

# Multiple Comparison Corrections

# Multiple Comparison/Testing

- Testing many features at the same time - high dimensional data.
    - Gencomics = Lots of Data = Lots of Hypotheses
- Therefore, P-value is no longer useful
    - Because testing many hypotheses simultaneously, even the slightest p-value cut off can propagate many false positives with high probability
- To solve this problem,
    1. Define a procedure
    2. Estimate an informative error rate for the procedure (controlling Type 1 Error rate)
    3. Control this error rate by adapting the procedure to it to guaranteed an error rate below a predefined value

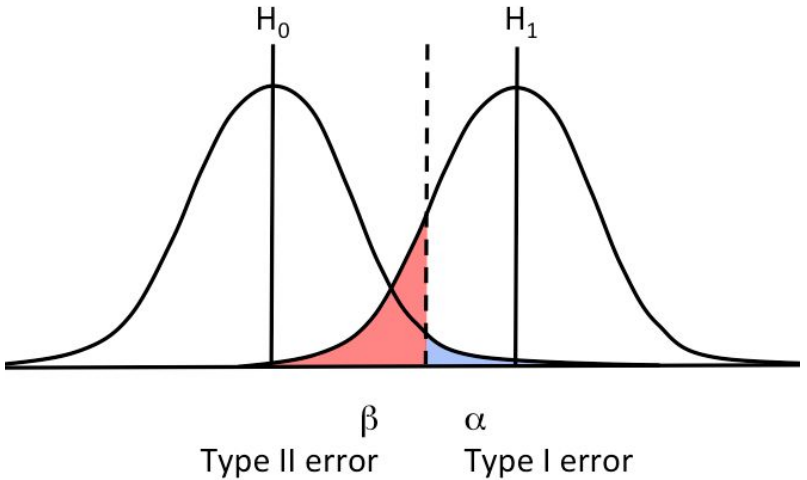


image cred: educational research techniques



Irizarry, R., & Love, M. (n.d.). *Multiple Testing*. Retrieved from Github: https://genomicsclass.github.io/book/pages/multiple_testing.html

# Family-wise error rate

- The probability of at least one type I error
- Want to guard against ANY false positives
    - Control false discovery rates
- Two general types of FWER corrections:
    - **Single step:** equivalent adjustments made to each p-value
    - **Sequential:** adaptive adjustment made to each p-value

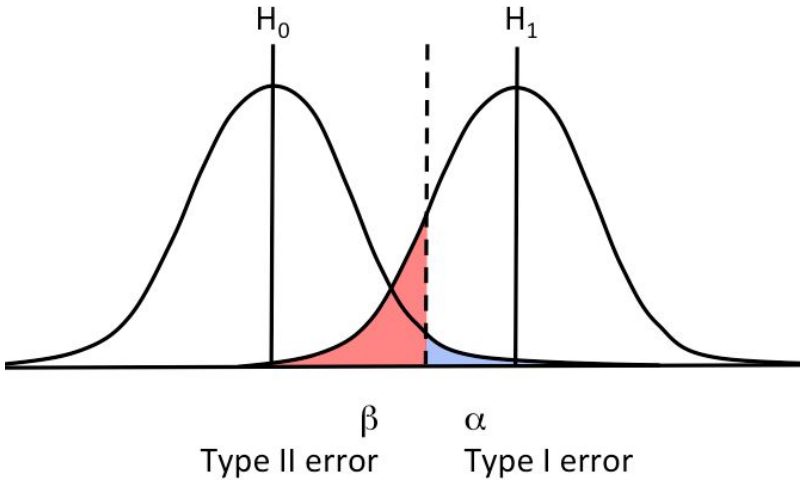


image cred: educational research techniques



https://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture10.p

# Benjamini-Hochberg Procedure

*Decreases the false discovery rate

| | O | P | Q | R | S |
|---|---|---|---|---|---|
| 1 | **Benjamini-Hochberg Approach** | | | | |
| 2 | | | | | |
| 3 | FDR | 0.05 | | | |
| 4 | k | 6 | | | |
| 5 | | | | | |
| 6 | | p-value | rank | adj α | BH sig |
| 7 | B | 0.00356 | 1 | 0.008333 | yes |
| 8 | F | 0.01042 | 2 | 0.016667 | yes |
| 9 | A | 0.01208 | 3 | 0.025 | yes |
| 10 | D | 0.02155 | 4 | 0.033333 | yes |
| 11 | E | 0.03329 | 5 | 0.041667 | yes |
| 12 | C | 0.11542 | 6 | 0.05 | no |

Image cred: real-statistics

1. Arrange p-values in ascending order
2. Assign ranks to p-values
3. Calculate each p-value's B-H critical value using:

$$(i/m)Q$$

i = the individual p-value's rank
m = total # tests
Q = false discovery rate (a percentage, chosen by you)

4. Compare original p-values to the critical B-H from Step 3 - find the largest p-value that is smaller than the critical value (to test for significance)



https://www.statisticshowto.com/benjamini-hochberg-procedure/

# Git Link