Learning Path 2 - Module 1

Reviewing Hypothesis Testing (Learning Path 1)

Applied Epistemology:
A Framework for how we know things scientifically.

A Refresher on Learning Path I: Hypothesis Testing

Agenda:

1. H_O/H_A: Our model of the test universe (the distribution of the variable)

2. Test & assumptions: are the assumptions met? Is the test valid?

3. Quantitative evidence: **p-value**, or critical value.
   - False positive =Type I (α), False Negative = Type II (β), Type III errors
   - Sensitivity, Specificity, Power → confusion matrix, ROC/AUC curve
   - Confusion Matrix

4. Conclusion & uncertainty/estimation

5. **<mark>Z-scores</mark>, χ² Goodness-of-fit test, and χ² Contingency test**

Perspectivism – why assumptions are important: https://hdsr.mitpress.mit.edu/pub/qasl4fza/release/3

What is "Statistical Thinking"?

Understanding complexity via:

- Think in probabilities;
- Understand Distributions;
- Apply models and list their assumptions;
- Quantify uncertainty;
- Utilize systematic criteria –>that is: automate decision making.

Retraining our brains to not rely on heuristics/shortcuts and bias.

- *We see patterns (i.e., faces) when there are none. Funny examples: https://www.boredpanda.com/objects-with-faces/*

Your pipeline for hypothesis testing in statistics

Step 1

Formulate your null **hypothesis**
- Null hypothesis is *only hypothesis that is tested*
- Falsification: *want* to reject your null

Step 2

Identify appropriate **test statistic**
- What number summarizes the evidence (mean, count etc.)
- Assumptions of your test

Step 3

Quantify the results of your test
- **P value** or comparison to **critical values**
- How *unusual* is your data (if Ho is true)?

Step 4

Conclude: reject or fail to reject
- based on alpha value
- if appropriate, confidence interval of the parameter

You are a biostatistician. Choose the correct test and justify it.

- Most of the work involved in statistics is clearly stating your null hypothesis
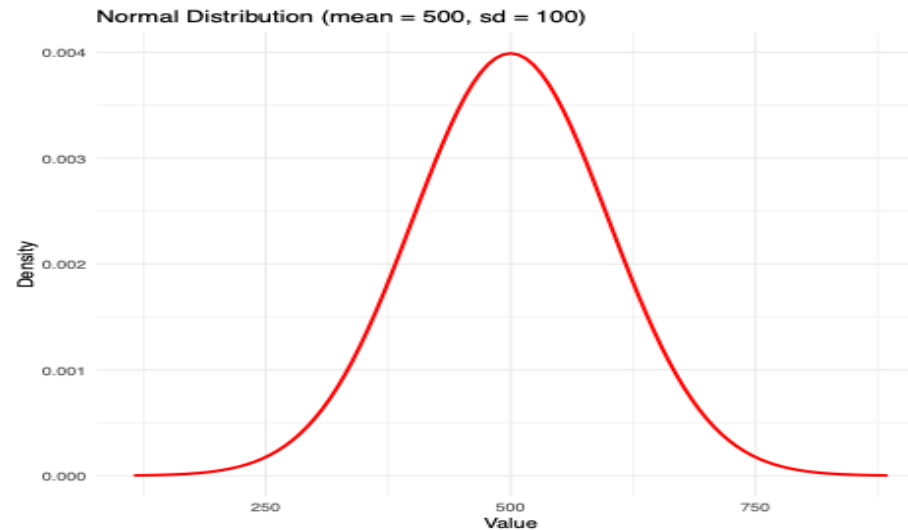
    What is your data type? What is your expectation? Can you quantify it? What is the null sampling distribution?

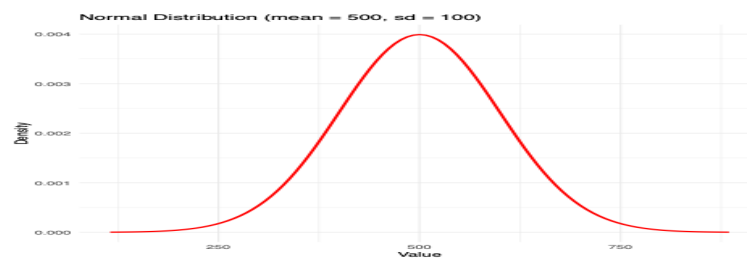- Hypothesis testing allows you to ask if a parameter **significantly** differs from the **null** expectation

    It quantifies how unusual the data are *if you assume that the null hypothesis is true.*

- Hypotheses are about populations but are tested with data from samples

    Assumes that the sampling is random & the sampling distribution follows a normal distribution (Thanks, Central Limit Theorem)

Example: The Verbal Ability and Skills Test (VAST) is a standardized test with a range of **0** to **1000**. VAST scores are normally distributed within the population of test takers with a mean score, μ, of **500** with a standard deviation, σ, of **100**. The red curve shows the distribution of this population of scores.

Normal Distribution (mean = 500, sd = 100)

The ACE training program boasts that their graduates score higher on average than the population of individuals who do not participate in a training course (the null population). That is, ACE claims test takers who first complete the ACE training program on average score higher than 500 on the VAST.

Normal Distribution (mean = 500, sd = 100)

$\mu = 500$
$\sigma = 100$

That is, ACE claims test takers who first complete the ACE training program on average score higher than 500 on the VAST.
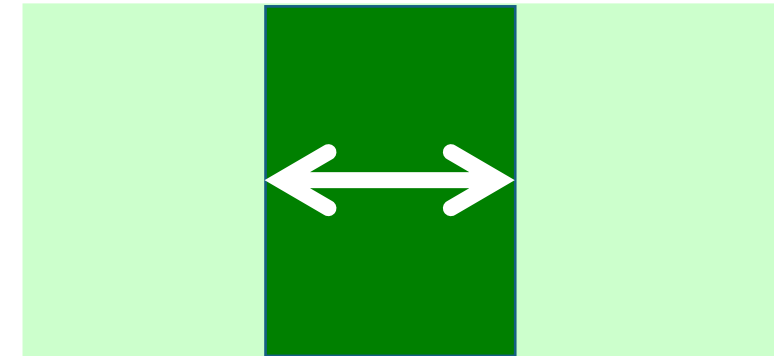
Step 1: Constructing the null hypothesis

Formulate Hypothesis
- o Quantifies how unusual data is *if you assume that the null hypothesis is true*
- o H_0 and H_A - mutually exclusive

What is the best Null Hypothesis to test this claim?

A. Ho: ACE $(\mu_{ACE}) \geq 500$
B. Ho: ACE $(\mu_{ACE}) = 500$
C. Ho: ACE $(\mu_{ACE}) \leq 500$
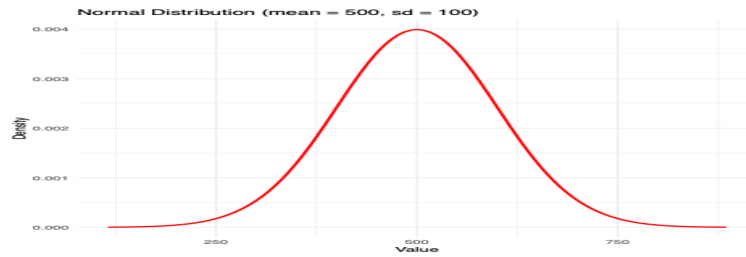D. Ho: ACE $(\mu_{ACE}) \neq 500$

Step 1: Making and using hypotheses:

The Null Hypothesis (H_0):

A specific statement about a population parameter made for the purpose of the argument. Usually carefully worded so that it can be rejected (falsified).

The Alternate Hypothesis (H_A):

Represents all other possible parameter values except that stated in H_0. It is often what the researcher hopes is true and remains after the Ho has been rejected.

Falsification of the null hypothesis (Karl Popper, 1959), discussion here: https://pubmed.ncbi.nlm.nih.gov/23249368/
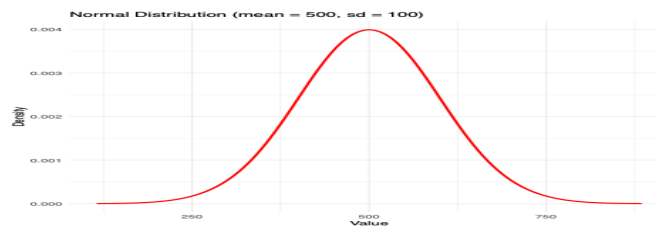
Normal Distribution (mean = 500, sd = 100)

$\mu = 500$
$\sigma = 100$

That is, ACE claims test takers who first complete the ACE training program on average score higher than 500 on the VAST.

Other concepts to review:

A. Would you be convinced that the average score for ACE graduates is greater than 500 if you were told that **one** randomly selected ACE graduate had a score of 550? (What is the probability that a randomly sampled score from the normally distributed population is 550 or greater, given that the population mean is 500 and the standard deviation is 100?)

B. Would it be more convincing if **25** randomly selected ACE graduates had an average score of 550? Why?

Normal Distribution (mean = 500, sd = 100)

$\mu = 500$
$\sigma = 100$

A random sample of 30 ACE graduates yields a sample mean of 510 (i.e., $X = 510$) and sd of 90.

Step 2: Identify the test statistic

- We can use a Z-score. (5B, Biostats I)

- <u>Assumptions:</u> Random sample, Normal Distribution

- Mechanics:

- Z = $\dfrac{510 - 500}{16.43}$ = 0.609

$$Z = \frac{\overline{Y} - \mu}{\sigma_{\overline{Y}}}$$

$$\sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}}$$
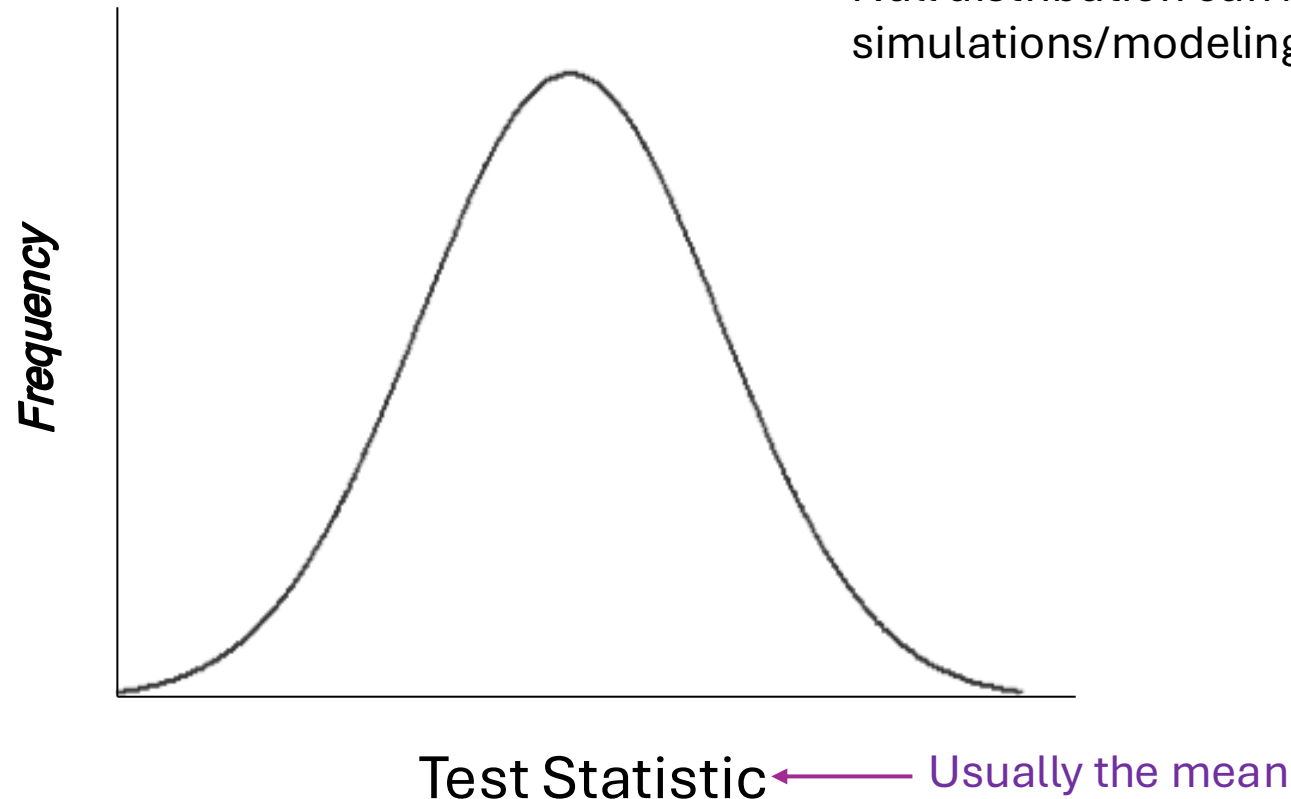
Step 2: Identify a Test Statistic:

Quantity calculated from the data that is used to evaluate how compatible the results are with those expected the null hypothesis.
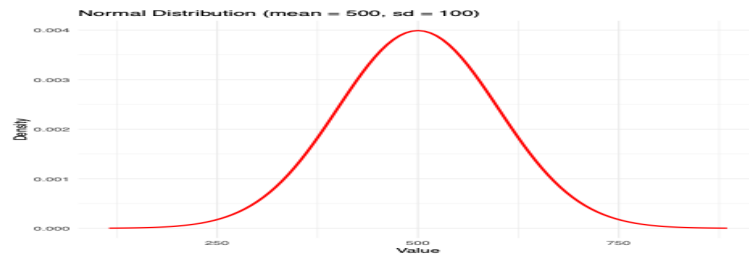
- How 'weird' are your results?
- Do your data support the assumptions of your test statistic?

Null Sampling Distribution:

Probability of the test statistic assuming the null hypothesis

- Usually assume Normal Distribution (for means, we can usually rely on CTL!)
- Null distribution can be acquired via computer simulations/modeling

Test Statistic ← Usually the mean

Normal Distribution (mean = 500, sd = 100)

μ = 500
σ = 100

A random sample of 30 ACE graduates yields a sample mean of 510 (i.e., X = 510) and sd of 90.

Z = 0.609

Step 3: Quantify results

Hypothesis testing automates **binary** decision making:

1. If p-value < α
   - Reject null hypothesis

2. If p-value > α
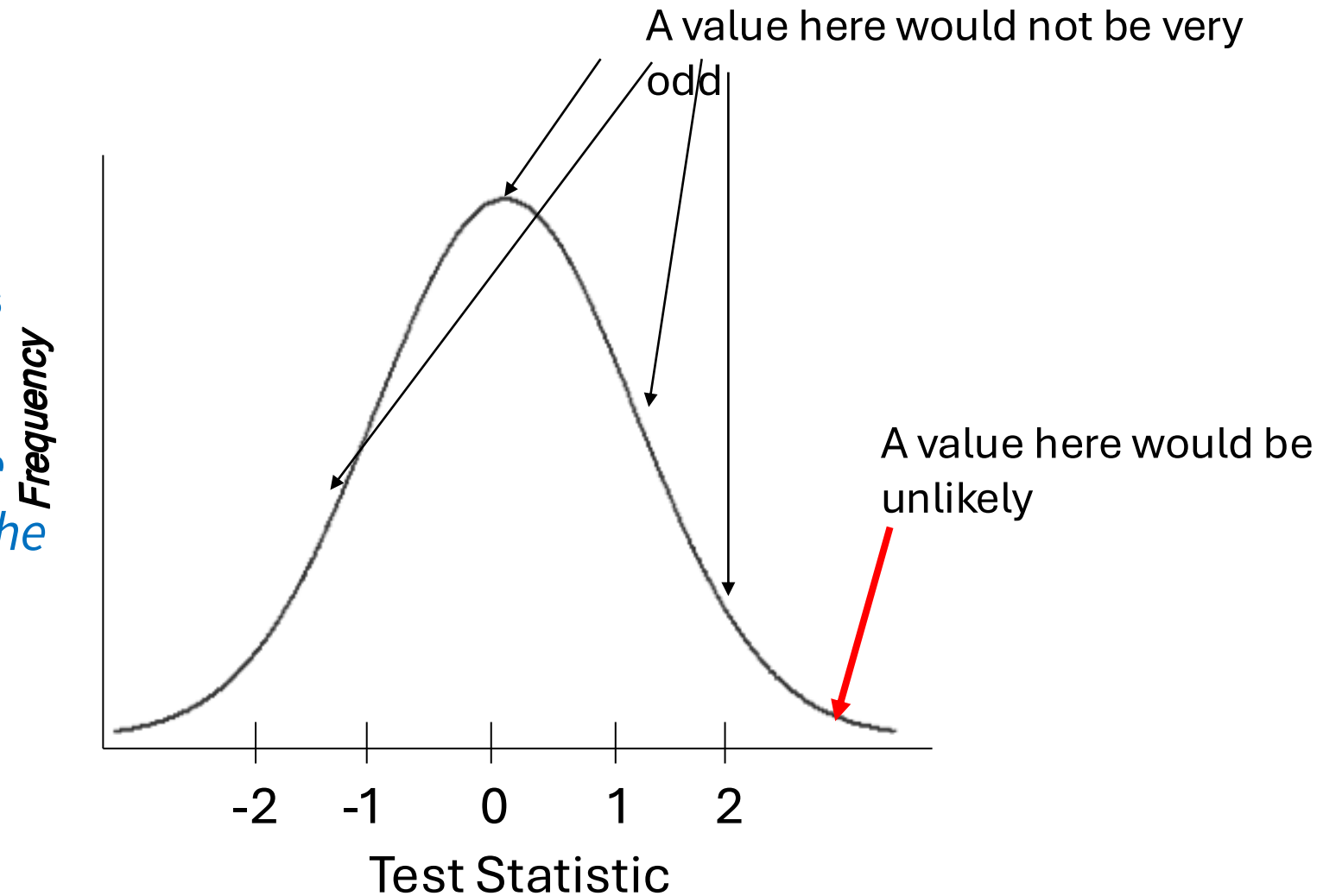   - Fail to reject null hypothesis

P-Value:

Probability of obtaining data that are equal to or even more extreme than the value assuming the null hypothesis is true

α (significance level):

Decided *before* experiment.
Conventional values: α = 0.05 or
α = 0.01.

A value here would not be very odd

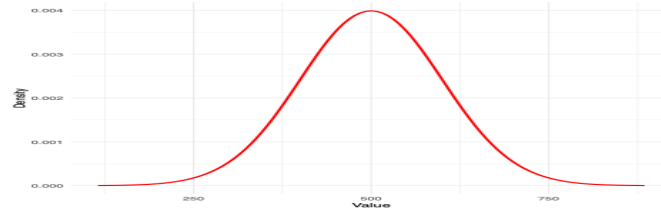A value here would be unlikely

P-Value:

Probability of obtaining data that are <u>equal to or even more extreme</u> than the value assuming the null hypothesis is true

"How weird is your result, if your Ho is true?"

A P-value is not:

- *A probability that Ho is true (it is already assumed that the Ho is true)*

- *A probability that the results are due to chance*
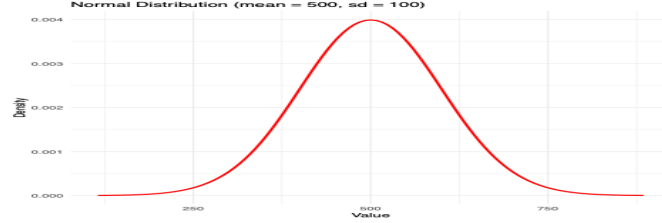
- *an effect size*

μ = 500
σ = 100

A random sample of 30 ACE graduates yields a sample mean of 510 (i.e., X = 510) and sd of 90. **Z = 0.609**

Step 3: Quantify results

- https://z-table.com/

- We won't calculate a p-value directly, but we can use cut-off values (critical values) that correspond to levels of α. The critical value is defined on the sampling distribution for means of samples of size N drawn from the null distribution, where the probability of obtaining a sample mean as large as or larger than the critical value is equal to a pre-defined alpha.

- **P(X > Z) = 0.2451**

- We don't double this, because this is a one-sided test (we would double it for a two-sided test, most will be)

- **0.2451 > 0.05**

We can compute what the necessary critical value cut-off would be:

1. Identify the p-value for the critical value. Here the p-value is 0.05 on the upper tail.

2. Convert the p-value to a z-score. Here the p–z converter (see table) tells us that a z-score of 1.645 cuts off .05 on the upper tail.

3. Z=0.609 is << 1.645 (and the probability of 0.2451 is >> 0.05)

Normal Distribution (mean = 500, sd = 100)

μ = 500
σ = 100

A random sample of 30 ACE graduates yields a sample mean of 510 (i.e., X = 510) and sd of 90. **Z = 0.609**

Step 4: Quantify results

- **Probability > α** so we <u>Fail-to-reject</u> the null hypothesis

- You will also want to put a confidence interval on your results (we will see this in Module 2 when we introduce the t test).

RNA-seq uses Z-scores to create heatmaps:

- In RNA-seq analysis, Z-scores are used to compare expression levels between samples. The Z-score of a gene is calculated by comparing its expression level in a given sample to the expression level of that gene across all samples.
    - A Z-score of **zero** indicates that the gene's expression level <u>is the same as the mean expression level across all samples</u>, while a positive Z-score indicates that the gene is expressed at a higher level than the mean, and a negative Z-score indicates that the gene is expressed at a lower level than the mean.
    - Z-scores are computed on a gene-by-gene (row-by-row) basis by subtracting the mean and then dividing by the standard deviation. It is used as scaling.

- Once Z-scores are calculated, they can be used to identify differentially expressed genes.
    - genes with a Z-score greater than a certain threshold (such as 2 or 3) can be considered as differentially expressed.
    - Z-scores can be used to create a heatmap or volcano plot, which can be a valuable way to visualize the data and identify patterns of expression

Simple example of using z-scores for RNA-seq data:

Gene	Control 1	Control 2	Treat 1	Treat 2
G1	5	6	9	10
G2	200	210	220	230
G3	1.2	1.1	1.3	1.4

This data set has two controls, two treatments for three genes. Each number is log-scaled or normalized expression value.

For each gene, ask:

"How far above or below that gene's own average is each sample?"

$$z = \frac{x - \bar{x}}{s}$$

x = value for one sample
$\bar{x}$ = *mean for that gene*
s = *Standard deviation for that gene*

Simple example of using z-scores for RNA-seq data:

Gene	Control 1	Control 2	Treat 1	Treat 2
G1	5	6	9	10
G2	200	210	220	230
G3	1.2	1.1	1.3	1.4

$$z = \frac{x - \bar{x}}{s}$$

For Gene 1:

Sample	Value	Mean ($\bar{x}$)	Std Dev (s)	z-score
Control 1	5	7.5	≈ 2.38	(5 – 7.5)/2.38 = -1.05
Control 2	6	7.5	≈ 2.38	-0.63
Treat 1	9	7.5	≈ 2.38	0.63
Treat 2	10	7.5	≈ 2.38	1.05

-0.84

+0.84

For Gene 2:

Sample	Value	Mean ($\bar{x}$)	Std Dev (s)	z = (x – x̄)/s
Control 1	200	215	≈ 12.9	(200 – 215)/12.9 = **-1.16**
Control 2	210	215	≈ 12.9	–0.39
Treat 1	220	215	≈ 12.9	+0.39
Treat 2	230	215	≈ 12.9	+1.16

-0.78

+0.78

← Way higher s, and mean than gene 1
← Similar up-regulation

For Gene 3:

Sample	Value	Mean ($\bar{x}$)	Std Dev (s)	z = (x – x̄)/s
Control 1	1.2	1.25	≈ 0.13	(1.2 – 1.25)/0.13 = **-0.38**
Control 2	1.1	1.25	≈ 0.13	–1.15
Treat 1	1.3	1.25	≈ 0.13	+0.38
Treat 2	1.4	1.25	≈ 0.13	+1.15

-0.76

+0.76

← Lower s, and mean than gene 1
← Similar up-regulation

Heatmap

In addition to plotting subsets, we could also extract the normalized values of *all* the significant genes and plot a heatmap of their expression using `pheatmap()` .

In this heatmap Z-scores are calculated for each row (each gene) and these are plotted instead of the normalized expression values; this ensures that the expression patterns/trends that we want to visualize are not overwhelmed by the expression values.

> *Z-scores are computed on a gene-by-gene basis by subtracting the mean and then dividing by the standard deviation. The Z-scores are computed **after the clustering**, so that it only affects the graphical aesthetics and the color visualization is improved.*

https://hbctraining.github.io/GCC-BOSC-2018/lessons/data_visualization.html

Which statement(s) is true about p-values?
----
a. p-value is the probability that the null hypothesis is true or false

b. p-value reflects the weight of evidence against the null hypothesis

c. p-value measures the size of the effect

d. if p value is less than or equal to the significance level, then the null hypothesis is not rejected.