YOU HAVE DATA...

NOW WHAT?

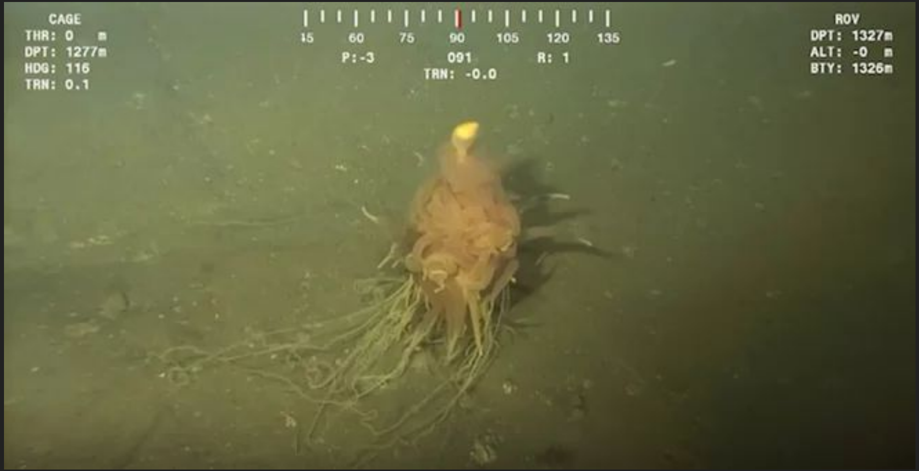
BIMM 143: Intro to Statistical Analysis of Large Datasets

Alex Sharp
Dr. Barry Grant

How do we know whether our data is real?

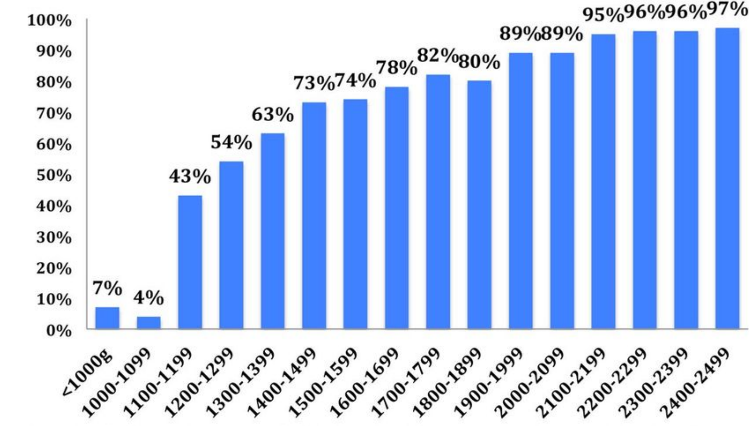
How might we structure a hypothesis regard to whether the sea creature below is the flying spaghetti monster?

=



What can we learn from our data?
Birth weight and chances of survival

We want to know the association between birth weight and survival because it determines necessary postnatal care.



Statistics are important!

How can we make sure our association is true?

1. What is the difference between a population and a sample?

2. Why do measurements of a phenomenon vary?

(Give one example for each)

Variance and Standard Deviation

= A measure of how much data varies

How to calculate it: Square the difference between measurements and the sample mean and take the average:

Population (σ_p^2) or Sample Variance (σ_s^2):

$$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$$

Adjusted Sample Variance (s^2):

$$s^2 = \frac{\sum (x - \bar{x})^2}{n-1}$$

The **standard deviation** is the square-root of the variance. The standard deviation is a standard measure of **dispersion**

Sample Standard Deviation

$$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$$

Basically, taking the square root of the variation puts it on the same **scale** as the mean

More on sample variance and standard deviation:

If you were ever curious, we divide by (n-1) instead of (n) because someone figured out that it leads to a more accurate estimation of the population variance.

Thus the *adjusted* sample variance is an “unbiased estimation of the population variance”

$$\sigma_p^2 = \sigma_s^2 + \sigma_p^2/n$$

$$\sigma_p^2 = \sigma_s^2\,(n/n-1) \rightarrow s^2$$

If you are curious to see what this mean, you can try out the algebra on your own

How to find the variance and standard deviation in R

```
#A minute to try it out!

#Find the functions to calculate the variance and standard deviation for the following data sets:

X=c(5,6,4,6.5)

Y=c(2,10,7,3,5)

x10=10*c(5,6,4,6.5)
```

Clarifying questions: Variance

What do I mean when I say that the variance of a measurement is high?

What is the relationship between standard deviation and variance?

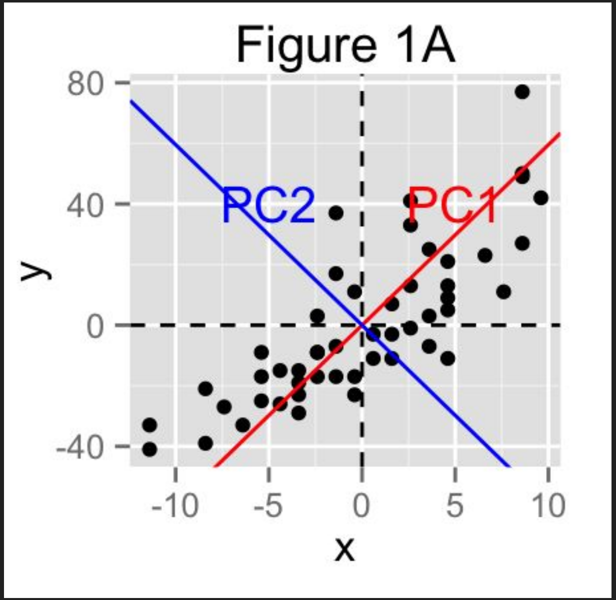
How does the adjusted variance/standard deviation relate to the population and sample variances/standard deviations?

Example: Variance in Principal Component Analysis

Which PC is best, 1 or 2?

What happens to the variance when I scale a measurement from 0→ 100 (mean = 50) to 0→ 1 (mean = 0.5)?

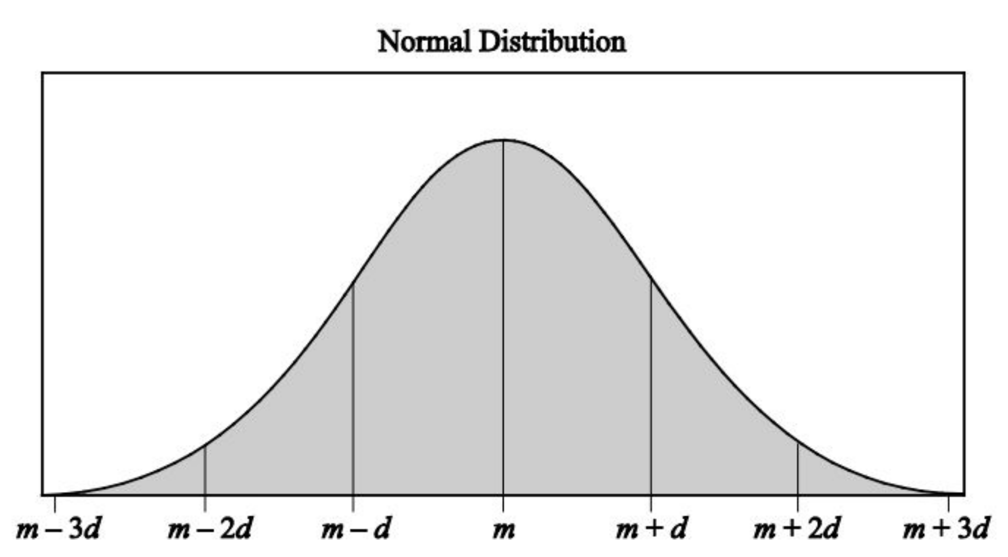
How does scaling measurements allow for a better comparison of the variances between principal components?



Next Up: Probability Distributions

What is the total of the probabilities for all outcomes of an event?

What do you think is the sum of the area under the curve of a probability distribution?



Probability Distributions

Use dnorm() to plot the normal distribution (use ?functionName to figure out how it works!)

Set the mean to 5 and standard deviation to 2. Try plotting 100 points and connecting them with lines()

Use hist() and rnorm() to make a histogram of a set of random variables

Set the mean to 5 and standard deviation to 2. Generate 5,000 points

Define the number of bars in the histogram with the breaks argument

What happens when you increase the sample size and breaks?

Use abline() to draw two red lines at 1 standard deviation from the mean

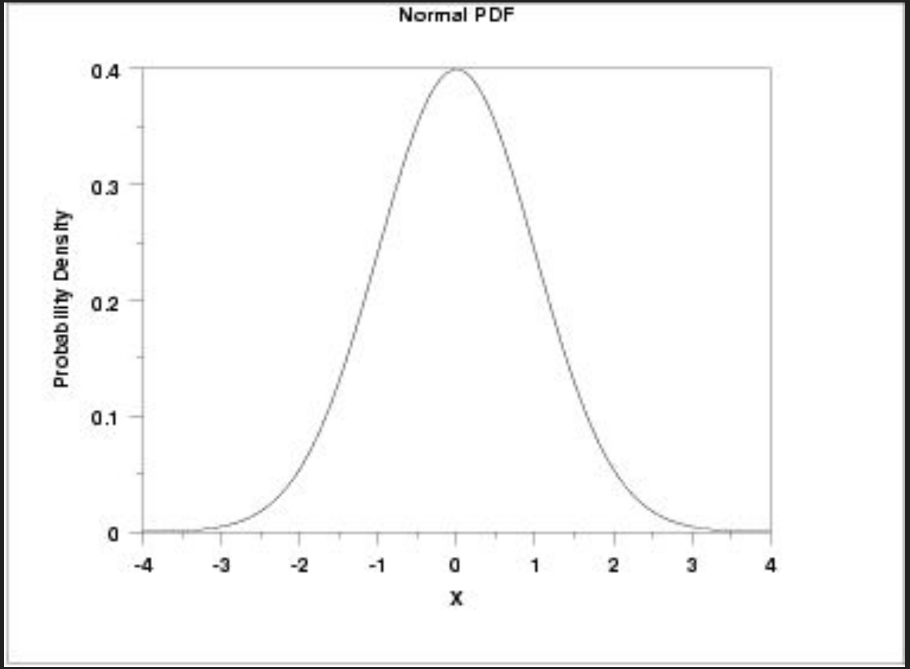
Use pnorm() to calculate the probability that x is between 2 and 6 if mean = 5 and sd = 2

Hint0: pnorm() integrates a normal distribution between a lower bound and +∞

Hint2: Use pnorm() twice

Probability Distributions

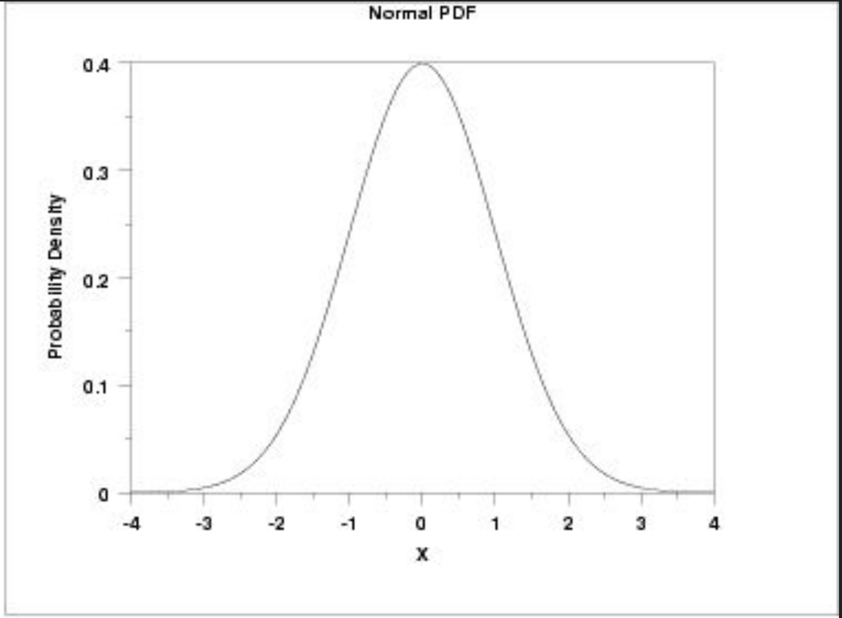
What is the relationship between the histogram and the probability density curve we made in the previous exercise?



The Normal Distribution

$$y = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

$$\mu = \text{ Mean}$$
$$\sigma = \text{ Standard Deviation}$$
$$\pi \approx 3.14159\cdots$$
$$e \approx 2.71828\cdots$$



This looks bloody complicated, but let’s see if we can write a function that will calculate this in R.

Building a function to plot the normal distribution

Let’s give it a try!

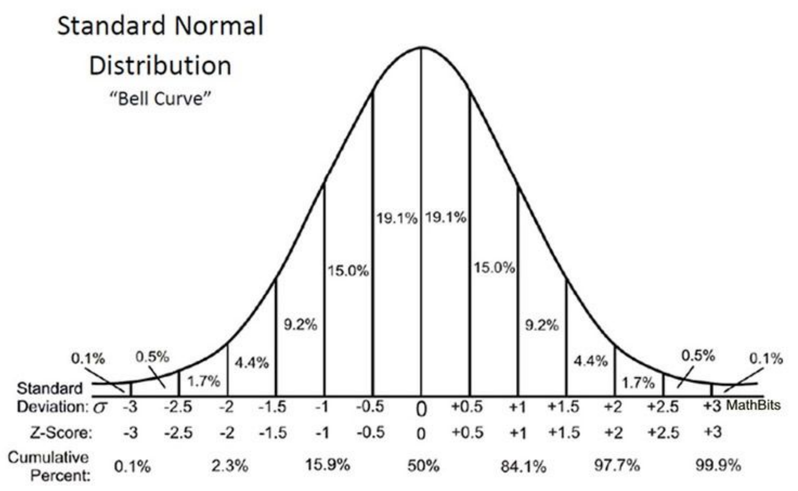
```
x = c(1:100)

mean = 50

stdev = 10

coolnorm <- function(x,mean,stdev){

      #Write your body here (Use the equation from the previous slide)

      return(y)

}

plot(coolnorm(x,mean,stdev))
```

Confidence Intervals on Normal Distributions

= range of values, equidistant from the mean, for which the probability of measurement of those values is defined

In the normal distribution the 95% confidence Intervals correspond to 1.96 standard deviations above and below the mean!

This 1.96 is called the Z-score for 95% confidence (we will learn more about z-scores later)



Confidence Intervals on Normal Distributions

Use dnorm() to plot the normal distribution

Set the mean to 5 and standard deviation to 2.

Eg. x = 5; sd = 2

Use **abline()** to draw red lines at 1 standard deviation from the mean

Use **abline()** to draw blue lines at 1.96 standard deviations from the mean

Q: What do you notice about the positions of 95% confidence interval when you change the standard deviation?

Why might it be useful to model our data using the normal distribution?

(1 reason)

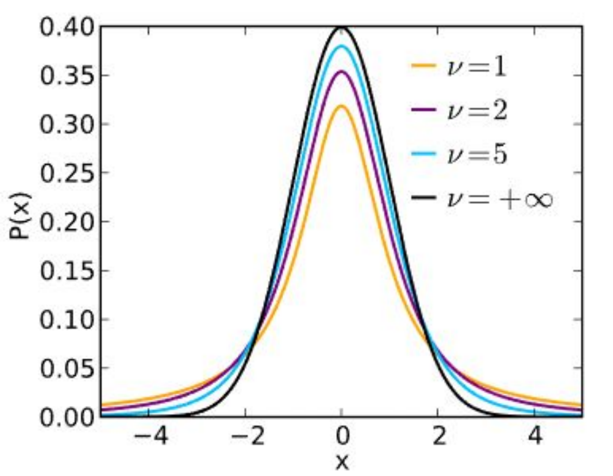
What are some obstacles to building a normal distribution from our data?

(1 obstacles)

Use the t-distribution when we do not have enough measurements to build a normal distribution

At ***low sample sizes***, the t-distribution more accurately estimates the shape of the probability distribution. The t-distribution tends toward the normal distribution when the sample size is large.

v =
degrees of freedom, which can be n-1 but it may also be calculated with alternative methods



$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}},$$

$$\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\,dx$$

As a rule of thumb, at less than 30 samples the sample mean and standard deviation does not accurately reflect the population mean.

Plotting probability from a t-distribution

Use **dnorm()** and **dt()** to plot the normal and t-distributions on the same graph

```
x <- seq(-4, 4, length=100)
hx <- dnorm(x)

degf <- c(1, 3, 8, 30)
colors <- c("red", "blue", "darkgreen", "gold", "black")
labels <- c("df=1", "df=3", "df=8", "df=30", "normal")

plot(x, hx, type="l", lty=2, xlab="x value",
  ylab="Density", main="Comparison of t Distributions")

for (i in 1:4){
  lines(x, dt(x,degf[i]), lwd=2, col=colors[i])
}

legend("topright", inset=.05, title="Distributions",
  labels, lwd=2, lty=c(1, 1, 1, 1, 2), col=colors)
```

Clarifying Questions: The normal and t-distributions

When is it more appropriate to model our data with the t-distribution than the normal distribution?

What happens to the curve of the t-distribution at low sample sizes and high sample sizes?

Next Section:
Let’s think about how scientists compare data

What are some of the differences between these two data sets?

City/Year	2014	2015	2016	2017	Average
Tijuana	490	670	910	1740	950
San Diego	80	90	120	100	100

Do Tijuana and San Diego serve as good representations of murder rates in Mexico and the US?

In other words, are these cities good **samples** of the **populations** of Mexico and the US?

What additional information do we need to test this hypothesis?

How do we compare sets of measurements?

Example:

The mouse behavioral response to cat urine.

Dr. Grant wants to know:

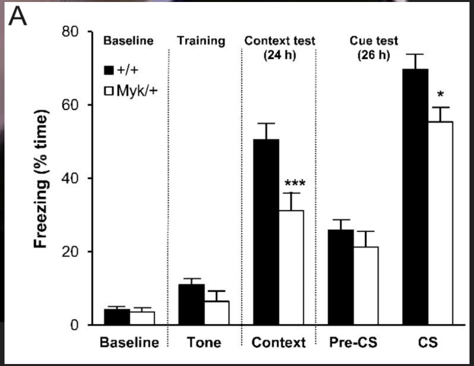
Did you see more or less freezing behavior?

How would you set up this experiment? Hint: What is your control?

How do we compare two sets of measurements?

Dr. Grant asks whether the results were significant.

In science we calculate the p-value to determine significance.



What is the p-value? What do you think a “good” p-value is?

Inferential statistics

- Hypothesis testing is a common form of inferential statistics. A null hypothesis is first stated, such as:
    - “There is no difference in freezing behavior in normal and diseased samples.” The alternative hypothesis is that there is a difference.
    - “The spaghetti monster does not exist”
- We use a test statistic to decide whether to accept or reject the null hypothesis. The test statistic gives us the probability that we can accept our null hypothesis.

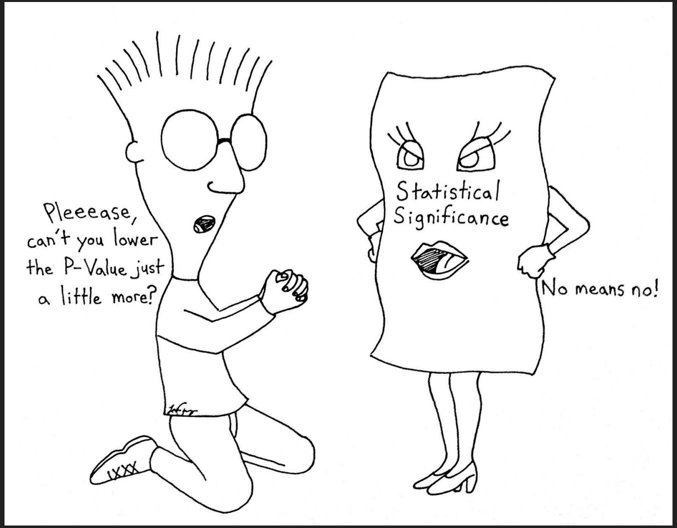
Inferential Statistics: P-values

The p-value is the probability that two or more sets of measurements are from the same phenomenon.

Hypothesis 1: Two sets of measurements are different

Hypothesis 0: (null hypothesis) Two sets of measurements are the same

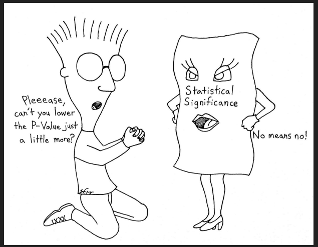
The p-value is the probability that the null hypothesis is true, which means that low p-values signify that the two data sets are more likely to be different phenomenon.



P-values: A rule of thumb

In general, a p-value of 0.05 or less is sufficient to declare a significant difference between datasets (such that we can reject the null hypothesis with 95% confidence or more)

How do we get the p-value?
It depends on what you are comparing



Z-test: Used to compare a sample mean (n>30) to a normally distributed population

One-sample t-test: Used to compare a population mean to a t-distributed set of sample measurements

Two-sample t-test: The t-statistic and the t-distribution is used to compare two sets of sample measurements (n<30)

ANOVA: Used to compare two or more normally distributed sets of measurements with similar standard deviations

Z-test

Z-test: Used to a set of sample measurements (n > 30) to a normally distributed population

```
x1 = rnorm(31, mean = 4.5, sd = 0.2)
```

Use the z.test() function to test whether x1 lies within a normally distributed population of mean 3 and standard deviation of 0.5

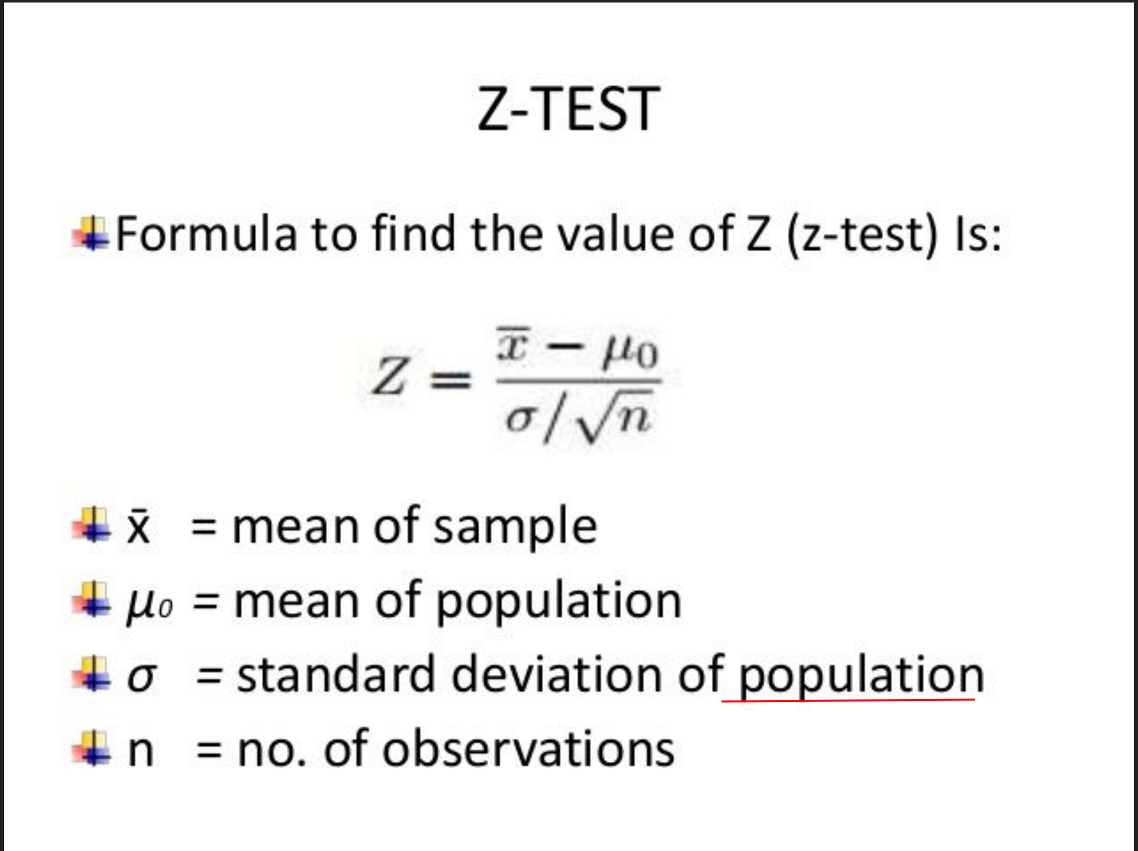
Hint: z.test(x1, mu = , sigma.x =)

Hint2: Use summary() to read the z.test object

What is the z-score? Is it above or below 1.96, and is the sample mean within the 95% confidence interval of the population mean?

Use $ and [] to extract the p-value

Try to calculate the z-statistic without the z-test and find the p-value from a table



Z-TEST

- Formula to find the value of Z (z-test) Is:

$$Z = \frac{\overline{x} - \mu_0}{\sigma/\sqrt{n}}$$

- $\bar{x}$ = mean of sample
- μ_0 = mean of population
- σ = standard deviation of population
- n = no. of observations

Note: We use the standard deviation of the population, and not of the sample

Can we use the Z-test to determine whether Tijuana and San Diego representative of their respective countries in 2017?

City/Year	Population	2014	2015	2016	2017
Tijuana	1.5 million	490	670	910	1740
San Diego	1.5 million	80	90	120	100
Mexico	127 million				29,000
USA	320 million				17,000

Hint: What additional data do we need to perform the Z-test?

One-Sample t-test

Used to compare a population mean to a t-distributed set of measurements

```
x1 = c(2,3,2.4)
```

Use t.test() to determine whether x1 is part of a population with a mean of 4. Find the t-statistic.

Q1: In this case, are we using the standard deviation from the population or the sample? (Hint: Go back 4 slides)

Q2: Does the t-distribution model the population or the sample (Hint: Think about the previous question)?

One Sample T-test:
Are 4-year homicide rates in Tijuana and San Diego representative of their respective countries in 2017?

City/Year	Population	2014	2015	2016	2017
Tijuana	1.5 million	490	670	910	1740
San Diego	1.5 million	80	90	120	100
Mexico	127 million				29,000
USA	320 million				17,000

How could we improve our sampling methodology if we wanted to estimate homicide rates for 2018?

Two-Sample t-test

Most commonly used statistical test in biology

Used to compare two t-distributed sets of measurements

```
x2 = c(4,5,8)
```

Use t.test() to determine whether x1 and x2 are likely from the same population

Go on to the next slide.

Two-Sample t-test

Two-sample t-test: Used to compare two t-distributed sets of measurements

```
Next try these two:

Test1 = t.test(x = c(8, 12, 9, 11), y = c(18, 19, 22, 21))

Test2 = t.test(x = c(8, 12, 9, 11), z = c(12, 12, 12))

In which test can we reject the null hypothesis, and why?

(Hint: how does the p-value relate to the probability that
the null hypothesis is true?)

Which tests show a significant difference?
```

Does playing Super Mario 64 improve cognitive abilities?

*Based on actual data, although not the first study Mario and the brain.

Participants were approximately 60 years old. For 6 months, participants were either left alone, trained to play piano with Synthesia software, or trained to play Super Mario 64.

Two types of tests administered: MoCa is a general cognitive assessment and short-term memory was assessed by recall of speech sounds

Treatment, Test	Pretreatment	Posttreatment
Control, MoCa	24.75924, 26.98667, 24.94788, 27.73277, 26.32180, 24.33580, 23.93787	25.53444, 29.59317, 27.64380, 24.93487, 28.18750, 28.66826, 25.12095
Control, STM	30.20258, 24.95905, 22.94211, 27.28486, 25.69725, 25.14251, 15.00047	20.71227, 25.36827, 27.51740, 24.15624, 24.58701, 25.95468, 21.62241
Music Control, MoCa	30.34595, 26.44569, 24.20254, 27.93985, 27.55459, 28.06074	26.18530, 25.45878, 27.88757, 27.77321, 24.92148, 26.81723
Music Control, STM	25.28486, 24.88745, 25.00814, 27.31849, 24.26051, 27.03386	24.23587, 23.61379, 24.64021, 25.79427, 21.92580, 28.31796
Video Game, MoCa	27.04761, 27.04171, 28.94454, 26.86962, 25.55325, 27.40797, 26.10106	30.00503, 29.29816, 27.00058, 28.77350, 29.11423, 30.94159, 28.30345
Video Game, STM	25.56566, 22.69519, 23.73103, 22.05665, 20.80492, 25.96523, 17.07626	28.35187, 29.37445, 25.75735, 21.65600, 27.92469, 24.79261, 23.83318

Steps to perform the two-sample t-test

- Consider two groups for which you obtain measurements.

- Make the null hypothesis (H_0) that there is no difference in the means of these two groups.

- Make the alternate hypothesis (H_1) that there is a difference.

- Calculate the t-statistic first: In the numerator take the absolute value of the difference of the two group means.

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- In the denominator calculate the “noise”.

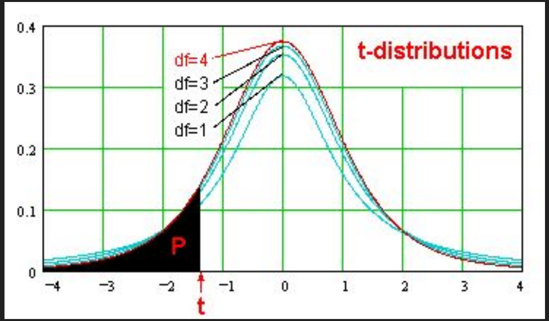
- From the t-statistic obtain a probability value.

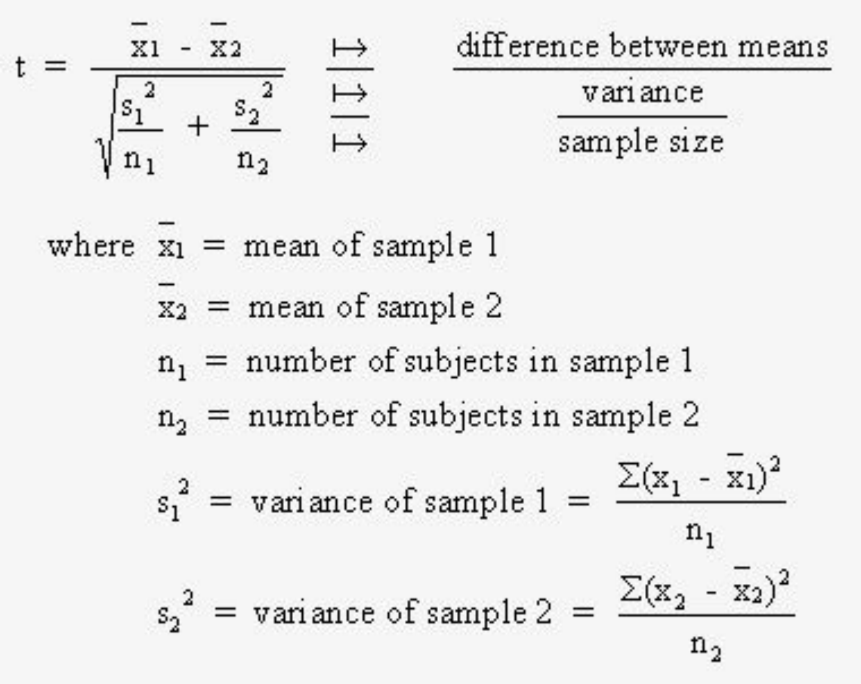
What does the t-test do?

The executive summary:

The t-test calculates the t-statistic

The t-statistic is then used to calculate the p-value by **integrating** over the t-distribution from -∞ to the t-statistic





$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \quad \begin{matrix} \mapsto \\ \mapsto \\ \mapsto \end{matrix} \quad \frac{\text{difference between means}}{\frac{\text{variance}}{\text{sample size}}}$$

where $\bar{x}_1$ = mean of sample 1

$\bar{x}_2$ = mean of sample 2

n_1 = number of subjects in sample 1

n_2 = number of subjects in sample 2

s_1^2 = variance of sample 1 = $\frac{\Sigma(x_1 - \bar{x}_1)^2}{n_1}$

s_2^2 = variance of sample 2 = $\frac{\Sigma(x_2 - \bar{x}_2)^2}{n_2}$

What does the t-test do?

If you are interested, take a look:

https://docs.google.com/document/d/1AR74nQho8TgznDo_i4UMYGmZTzEwk6tGvwoJIztybFs/edit?usp=sharing

What do we mean when we say that a p-value leads us to the wrong inference?

What are some reasons why a p-value may be inaccurate? Come up with 2 reasons and how they might contribute to an alteration in the p-value.

See here for some inspiration…

https://xkcd.com/882/

Two ways to be wrong means two types of error

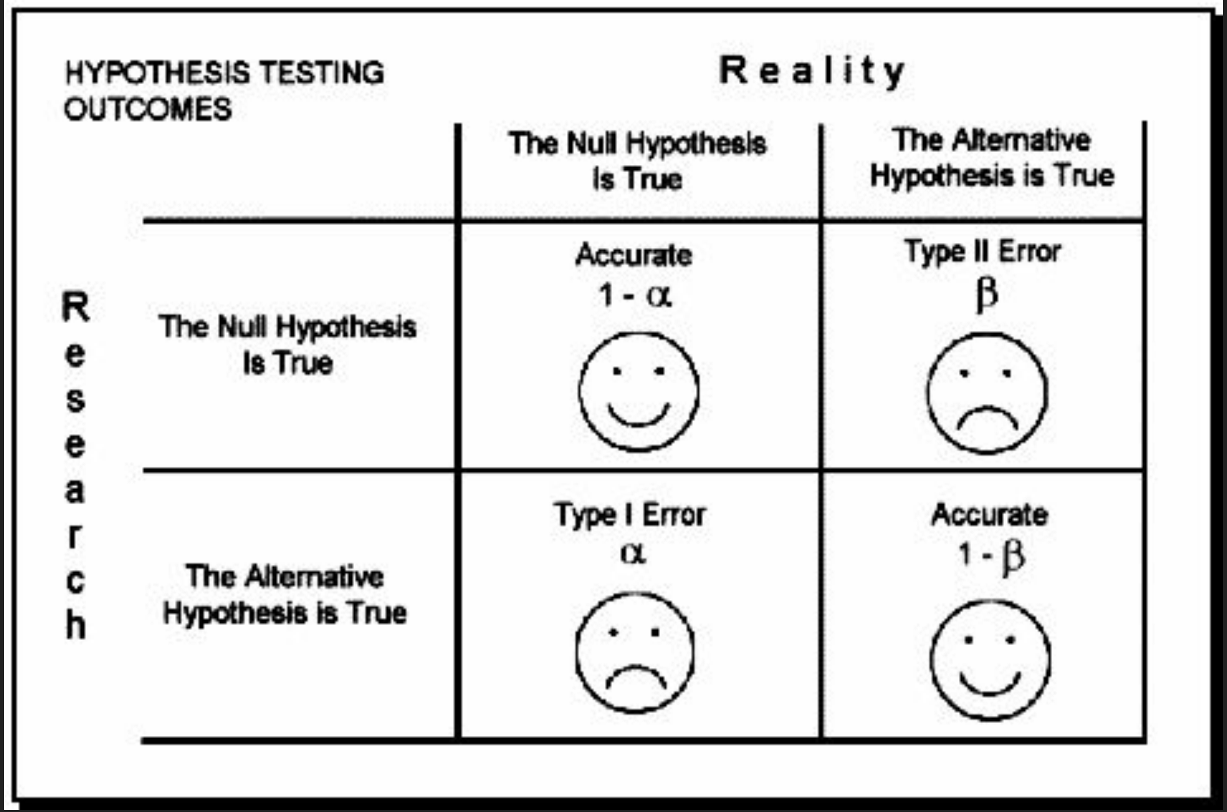
Way to be wrong #1

We think null hypothesis is true because **the p-value we measure is more than 0.05,** but the null hypothesis is actually false. This means that there actually is a significant difference between datasets, and our inference is therefore wrong.

Way to be wrong #2

We think the null hypothesis is false because the p-value we *measure is less than 0.05*, but our null hypothesis is actually true.

Four outcomes of hypothesis testing (2 bad ones)



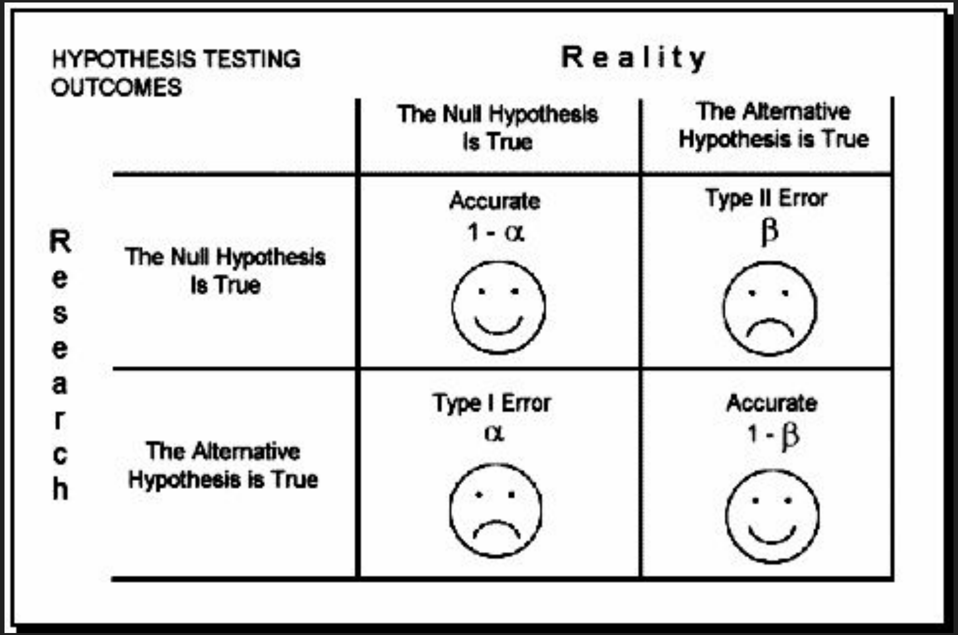
HYPOTHESIS TESTING OUTCOMES		Reality	
		The Null Hypothesis Is True	The Alternative Hypothesis is True
Research	The Null Hypothesis Is True	Accurate $1 - \alpha$ ☺	Type II Error β ☹
	The Alternative Hypothesis is True	Type I Error α ☹	Accurate $1 - \beta$ ☺

Type 1 error (also known as alpha, or significance level)

= The probability that you reject the null hypothesis when it is true

In other words, you think there is a difference (low p-value), when there actually is not one

The ***False Discovery Rate*** is the probability of committing Type 1 error, which is also the rate of false positives



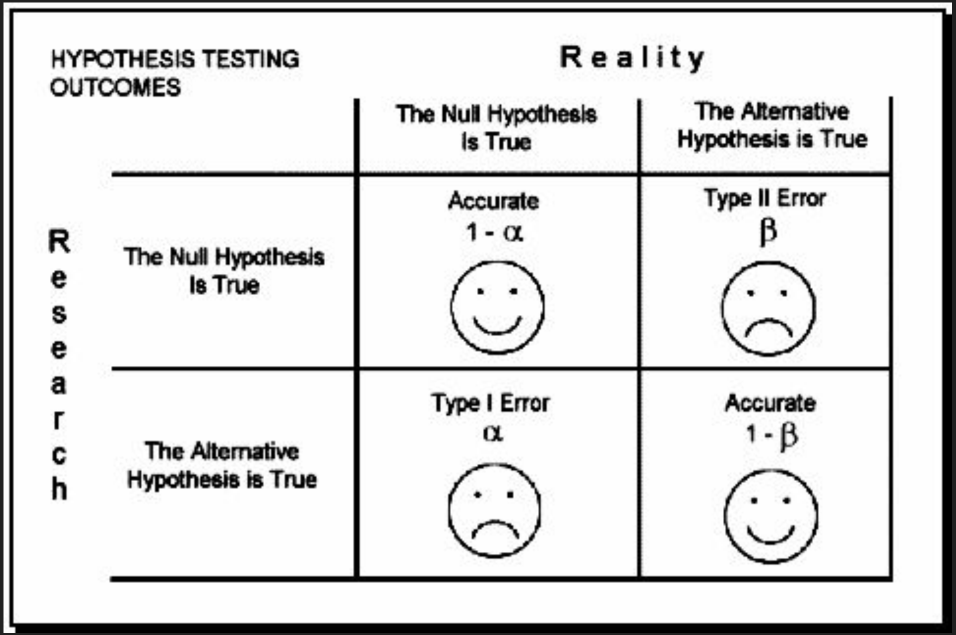
HYPOTHESIS TESTING OUTCOMES		Reality	
		The Null Hypothesis Is True	The Alternative Hypothesis is True
Research	The Null Hypothesis Is True	Accurate 1 - α ☺	Type II Error β ☹
	The Alternative Hypothesis is True	Type I Error α ☹	Accurate 1 - β ☺

Type 2 error (or beta)

= The probability that you will accept the null hypothesis when it is false

In other words, you think there is no difference between data sets ($p > 0.05$), when there actually is a difference.

The ***Power*** is 1-beta, or the probability that you will reject a null hypothesis when it is false.



HYPOTHESIS TESTING OUTCOMES		Reality	
		The Null Hypothesis Is True	The Alternative Hypothesis is True
Research	The Null Hypothesis Is True	Accurate $1 - \alpha$ ☺	Type II Error β ☹
	The Alternative Hypothesis is True	Type I Error α ☹	Accurate $1 - \beta$ ☺

false discovery rate

- The false discovery rate (FDR) is a popular multiple corrections correction. A false positive (also called a type I error) is sometimes called a false discovery.

FDR = number of false positives / number called significant

- The FDR equals the p value of the t-test times the number of genes measured (e.g. for 10,000 genes and a p value of 0.01, there are 100 expected false positives).

false discovery rate

- You can adjust the false discovery rate. For example:

RNA-seq: Comparing two conditions for differential expression

Would you report 100 differentially regulated transcripts of which 10 are likely to be false positives, or 20 transcripts of which one is likely to be a false positive?

FDR	# regulated transcripts	# false discoveries
0.1	100	10
0.05	45	3
0.01	20	1

t-test: power calculation

- Power is the fraction of true positives that will be detected. It is a value between 0 and 1. The larger the sample size, the larger the power.

- You can use the R packages pwr or power.t.test.

- Power is the probability that you can reject the null hypothesis when it is false

Calculating Power

```
#Use the pwr.t.test() function to calculate the power for a given FDR
(significance level).

#Let’s set the sample size (number in each group) to n=11, the threshold for
significance to 0.05, and see that the power is 0.6. (set d=1) Do it
yourself:

pwr.t.test(n,sig.level,d,type,alternative)

#if we do not give the sample size, the function will calculate one for us
if we give it power instead of sig.level. Try it out

pwr.t.test(sig.level,d,power,type,alternative)

#What sample size do we need to get a power of 0.9?
```

Power: A rule of thumb

In general, a power of 0.8 for a significance level of 0.05 is ok.

To calculate the power, we need to first calculate the effect size for our data, or Cohen’s d (aka delta)

Effect size = the size of the significant or actual difference between two groups of data.

```
Use the following function:

     effect_size <- function(m1,m2,s1,s2,n1,n2)
     {
       #Cohen's method (1988), effect size is called Cohen's d

       pooled_sd = sqrt(((n1-1)*s1^2+(n2-1)*s2^2)/(n1+n2-2))

       d = abs(m1-m2)/pooled_sd

       return(d)
     }
```

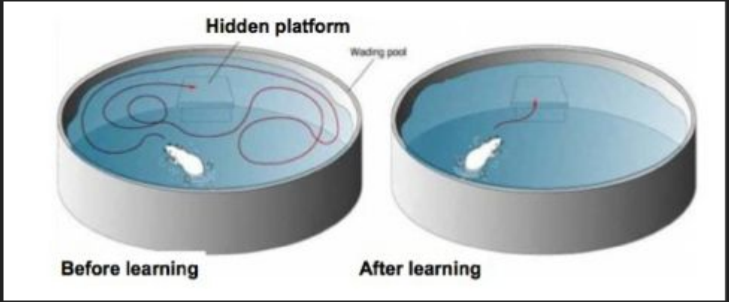
Hint: Use this function to calculate the effect size for the next activity

Does living in an enriched environment improve memory in mice?

*Again, based on actual data

Learning assessed by the length of swim path in the Morris Water Maze

CTL mice were raised in a cage with 4 other mice, RUN mice were given a wheel on which to run (also raised with 4 other mice), and ENR were given larger cages, new toys each week, and were raised with an additional 5 mice.



Condition	Pre-training	Post-training (3 days)
CTL	296.0262, 195.1290, 216.6341, 279.1301, 262.5093, 185.4587, 299.3648	107.8990, 107.7202, 105.2005, 106.5831, 109.8243, 116.8620, 115.4566
RUN	242.0761, 195.5219, 249.3518, 239.2630, 235.7053, 233.4579, 259.3812	75.93327, 36.42067, 112.04678, 70.77384, 98.65486, 76.85637, 74.01117
ENR	203.0949, 231.6910, 204.7919, 239.5369, 236.8510, 229.3702, 216.9117	37.45427, 37.92696, 58.13280, 28.55817, 64.23409, 48.59342, 31.46771

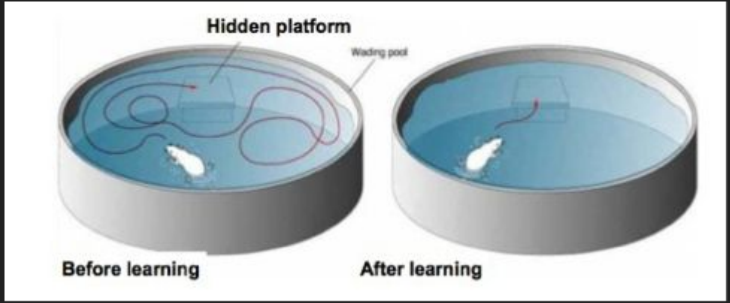
Does living in an enriched environment improve memory in mice?

Calculate the log-2 fold decrease from pre-training to post-training (use c())

Test for statistical significance.

Calculate power for a significance level of 0.05.

Do you accept or reject the null hypothesis?



Condition	Pre-training	Post-training (3 days)
CTL	296.0262, 195.1290, 216.6341, 279.1301, 262.5093, 185.4587, 299.3648	107.8990, 107.7202, 105.2005, 106.5831, 109.8243, 116.8620, 115.4566
RUN	242.0761, 195.5219, 249.3518, 239.2630, 235.7053, 233.4579, 259.3812	75.93327, 36.42067, 112.04678, 70.77384, 98.65486, 76.85637, 74.01117
ENR	203.0949, 231.6910, 204.7919, 239.5369, 236.8510, 229.3702, 216.9117	37.45427, 37.92696, 58.13280, 28.55817, 64.23409, 48.59342, 31.46771

See Super Mario 64 and Mouse papers below for more information:

http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0187779

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5049654/

Clarifying Questions: Hypothesis Testing

What do I mean when I say that the effect size is small?

Why would we want to adjust the false discovery rate when comparing two samples for differential gene expression?

What is the relationship between the power and the sample size?

Power gives you the power to do what?

Next up: Multiple Testing Correction

When we have a lot of p-values, some of them are bound to be wrong

We can adjust the p-values to new values which better account for the false discovery rate within a large number of p-values

Simply input a vector of p-values into p.adjust() and select the adjustment you would like to perform

Methods “BH” and “BY” stand for Benjamani & Hotchberg (1996), and Benjamani & Yekutieli (2001). These methods control for the FDR while the others (see documentation) control for the family-wise error rate, which is related to the rate of false rejection.

One last note on statistical measures:
Fold change

Fold Change = (observation in condition 1) / (observation in condition 2)

Fold change: Is it worth talking about?

- To a statistician fold change is sometimes considered meaningless. Fold change can be large (e.g. >> two-fold up- or down-regulation) without being statistically significant.
- Why might fold change be important to a biologist?

Fold change: Is it worth talking about?

To a biologist fold change is almost always considered important for two reasons:

- First, a very small but statistically significant fold change might not be relevant to a cell’s function.
- Second, it is of interest to know which genes are most dramatically regulated, as these are often thought to reflect changes in biologically meaningful transcripts and/or pathways.

Calculate the p-value and power for an RNA-seq data set

1. Download the file from your email and read it into separate variables R Studio
2. Use na.omit to remove rows without values if necessary
3. Use cbind() and %in% (if necessary) to append the variables into a dataframe
4. Select rows without zero expression with “[ ]” (eg. df[df[, 1:4] != 0,])
5. Perform the t-test on the two conditions and calculate the $\log_2$ fold-change
6. For the differentially expressed genes, calculate the power and decide which genes are well powered and have a significant fold change
7. Plot the p-values, and then plot the adjusted p-values (use “BH” or “BY”)
8. Read: http://varianceexplained.org/statistics/interpreting-pvalue-histogram/
   Is our p-value histogram ideal? Why or why not? Does correcting it improve the distribution?

Muddy Points Assessment and R Quiz

Please fill out the Muddy Points Assessment below

https://goo.gl/forms/YBPM7MoMTqKNnLoV2

If there is time:
Analysis of dependence between variables

Linear Regression: Can be used to find a relationship between two or more continuous variables (A continuous variable has an infinite number of numerical values, like length measured in feet) lm()

Logistic Regression: Can be used to find a relationship between two or more continuous and categorical variables (A categorical variable has a limited number of values, like yes/no, true/false) glm(family = binomial())

Covariance: The joint variance between two continuous variables is used as a measure of the dependence between those variables. Calculated as the difference from a proposed linear model of the data. cov()

Chi-Square Test: Used to compare two or more continuous and categorical variables. You get a p-value! chisq.test()