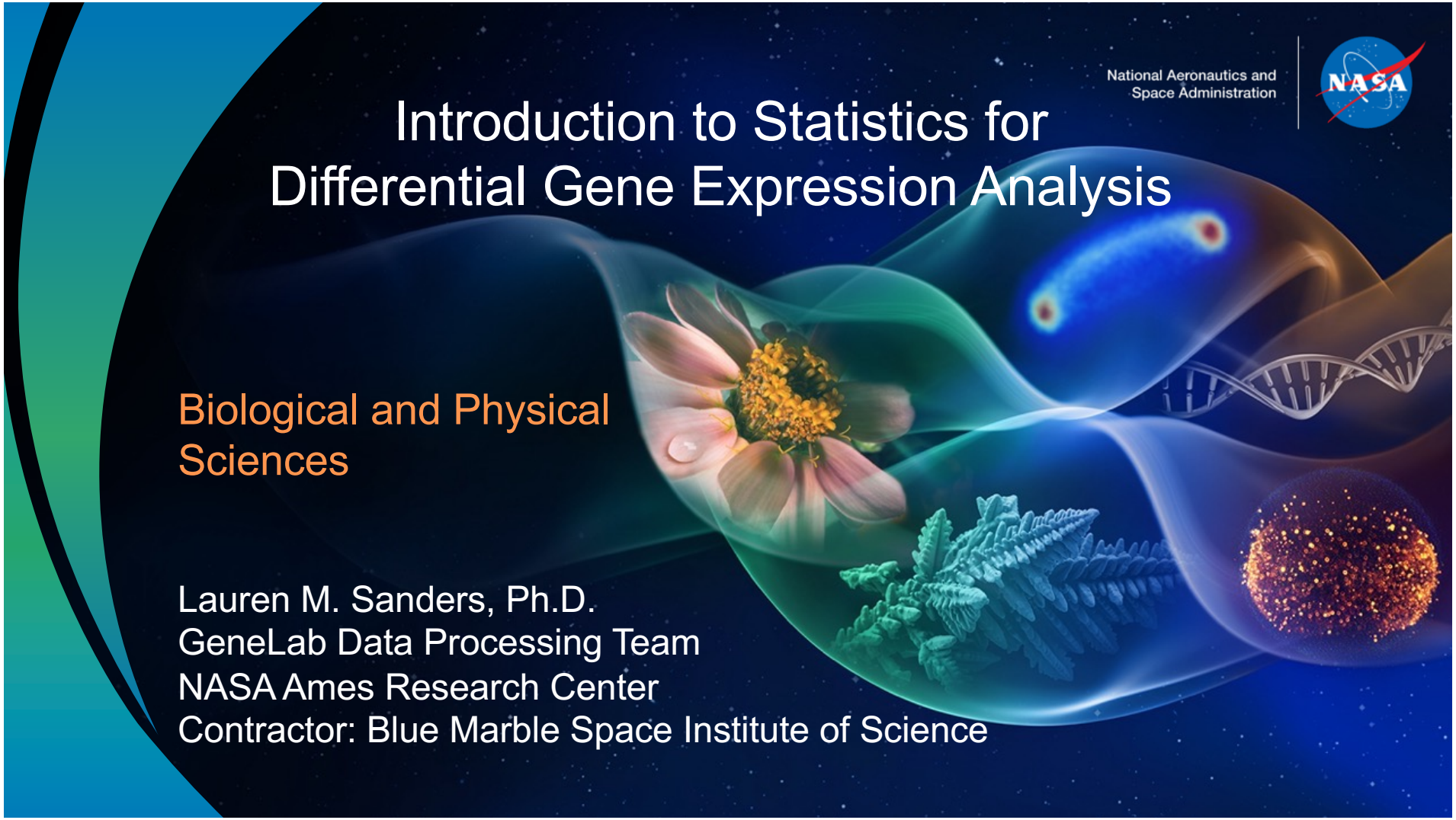


Introduction to Statistics for Differential Gene Expression Analysis

Biological and Physical Sciences

Lauren M. Sanders, Ph.D.
GeneLab Data Processing Team
NASA Ames Research Center
Contractor: Blue Marble Space Institute of Science

Outline

1. Mean
2. Variance / Dispersion
3. Standard Deviation
4. Fold Change
5. Hypothesis Testing
6. Null Hypothesis
7. P-value
8. P-value adjusted for multiple comparisons
9. T-tests and Wald Tests
10. Principal Component Analysis
11. Median of Ratios method (used in DESeq2 step 1)
12. Maximum Likelihood Estimation (used in DESeq2 step 2)
13. General Linear Model / Negative Binomial Model (used in DESeq2 step 3)

- Metrics to quantify (describe) data
- Form the basis for more complex calculations
- In bioinformatics, these are usually calculated by computational programs

- Commonly used concepts and calculations in probability and significance

- Key methods used in the DESeq2 R package for differential gene expression analysis
- Advanced topics in probability and statistics

Important Terminology

- *sample data, dataset* – a set of data points from an experiment
- *a sample* – a self-contained measurement of one or more variables
- *data point* – a single value in a dataset
- *variable* – a quantifiable aspect of each sample (can be numeric or non-numeric)

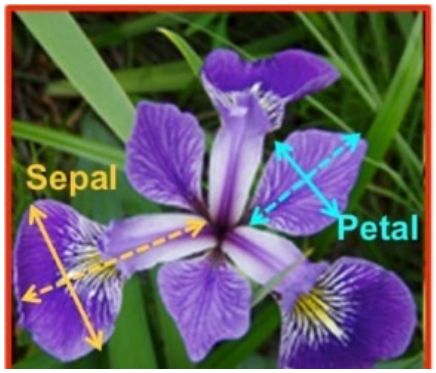
Mel, an advanced carbon-based life form, sets out from a planet in the Tadpole Galaxy, 420 million light years from Earth…

Upon arrival to Earth, Mel accidentally lands in a large field at the “iris capital of the world”, Portland, Oregon…

Because all there is to see are irises, Mel thinks they are the dominant life form on planet Earth, and spends a year collecting detailed data on petal length and width, and sepal length and width.

Mel decides there are three main iris species, and names them *setosa*, *versicolor*, and *virginica*…

Mel sends the iris dataset back to the home planet, and the Tadpolish scientists back home have a lot of questions…

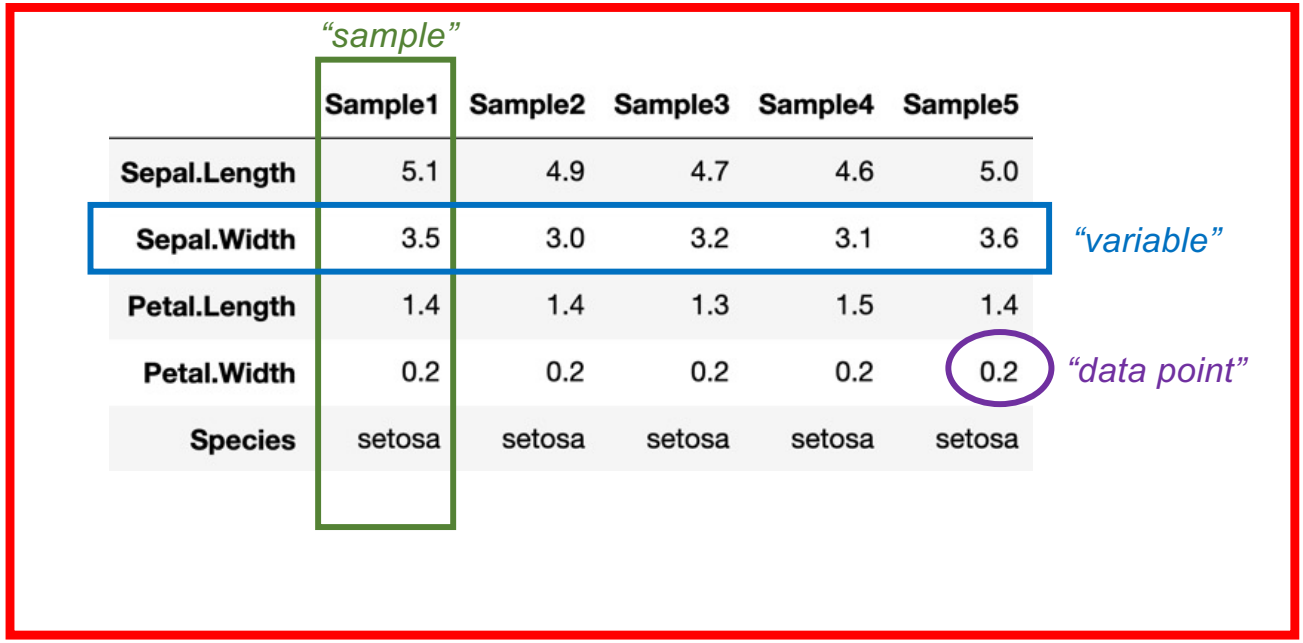


"Iris" dataset from R (first 5 samples):

	Sample1	Sample2	Sample3	Sample4	Sample5
Sepal.Length	5.1	4.9	4.7	4.6	5.0
Sepal.Width	3.5	3.0	3.2	3.1	3.6
Petal.Length	1.4	1.4	1.3	1.5	1.4
Petal.Width	0.2	0.2	0.2	0.2	0.2
Species	setosa	setosa	setosa	setosa	setosa

https://www.snaplogic.com/machine-learning-showcase/iris-flower-classificatio

Important Terminology



The Tadpolish scientists want some basic descriptors of the different species on the new planet Earth…

- **So, Mel comes up with a new quantification, the average or mean of a set of numbers**

- **To calculate the mean of a set of data points:**
  1. Add all the data points together
  2. Divide the sum by the total number of data points

- Equation: $m = \frac{\Sigma x_i}{n}$
  - m = mean
  - Σ = sum
  - x = a data point
  - n = total number of points

Let's calculate the mean sepal length for Samples 1 through 5:

	Sample1	Sample2	Sample3	Sample4	Sample5
Sepal.Length	5.1	4.9	4.7	4.6	5.0
Sepal.Width	3.5	3.0	3.2	3.1	3.6
Petal.Length	1.4	1.4	1.3	1.5	1.4
Petal.Width	0.2	0.2	0.2	0.2	0.2
Species	setosa	setosa	setosa	setosa	setosa

By hand: $$\frac{5.1 + 4.9 + 4.7 + 4.6 + 5.0}{5} = 4.86$$

Programmatically (R):

```
In [26]: mean(c(5.1, 4.9, 4.7, 4.6, 5.0))
```

```
4.86
```

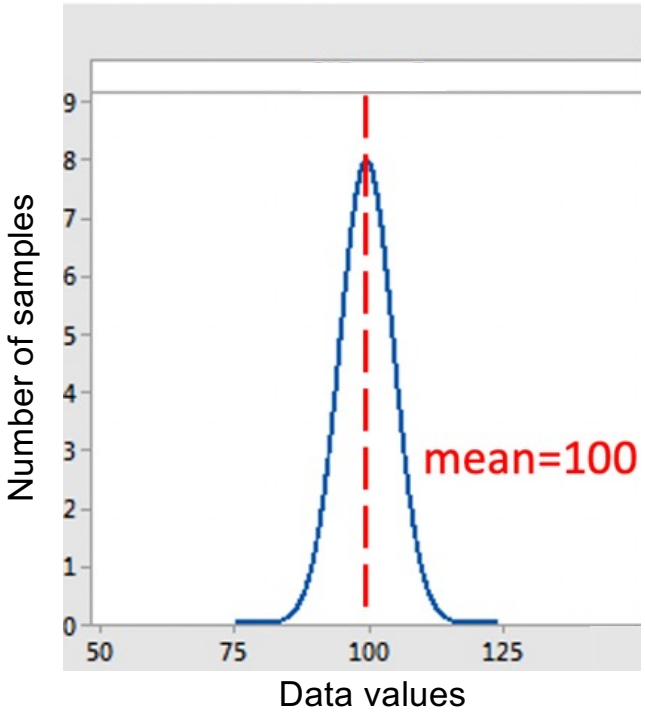
```
In [ ]:
```

Realizing that the mean only describes the **center** of the dataset, not the outer edges, and the edge cases are often the most interesting…

Mel invents another quantification method, the variance or dispersion of a set of numbers:

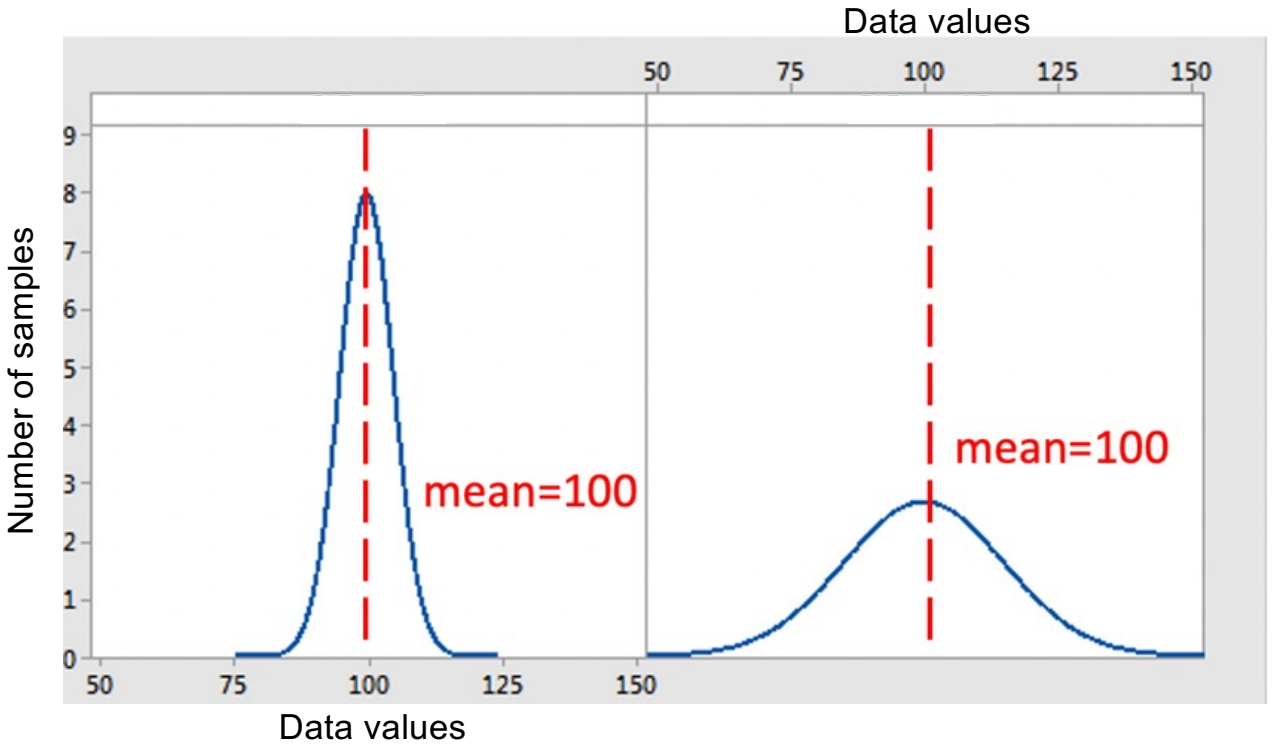
- The variance tells you the consistency of the points in a dataset
- A dataset with low variance has data points closer together (more consistent) than a dataset with high variance
- Knowing the variance helps you predict the likelihood of unusual events

Datasets can have the same mean, but different variance!



statisticsbyjim.com/basics/variability-range-interquartile-variance-standard-deviation/

Datasets can have the same mean, but different variance!



Calculating variance

- **How we calculate variance depends on whether we are calculating variance for:**
  - the whole population (e.g. all irises on Earth)
  - a sample from the population (e.g. Mel's irises data set)

- **The Tadpolish scientists really want to know the variance for the entire iris population, but there is only 1 Mel and time is short**

- **So, Mel samples a reasonable amount**
  - We hope that the sample is big enough to be representative of the population

- **NOTE: There is a part of statistics that deals with how BIG a sample needs to be, but we won't be going into that today**

Calculating variance

- **To calculate the variance of a dataset that was sampled from a larger population (the "sample variance"):**
  1. Calculate the mean of the dataset
  2. Subtract the mean from each data point
  3. Square each result
  4. Sum up the squared results
  5. Divide by one fewer than the total number of datapoints

- **Equation:** $$V = \frac{\Sigma(x_i - m)^2}{n - 1}$$

V = variance
Σ = sum
x = a data point
m = mean
n = total number of data points

Variance calculation example

- **Let's find the variance of the sepal length for samples 1 through 5:**

	Sample1	Sample2	Sample3	Sample4	Sample5
Sepal.Length	5.1	4.9	4.7	4.6	5.0
Sepal.Width	3.5	3.0	3.2	3.1	3.6
Petal.Length	1.4	1.4	1.3	1.5	1.4
Petal.Width	0.2	0.2	0.2	0.2	0.2
Species	setosa	setosa	setosa	setosa	setosa

sample variance = the variance of these 5 irises
population variance = the variance of all the irises

Let's find the variance of the sepal length for samples 1 through 5:

By hand:

- **Calculate the mean:**

$$\frac{5.1 + 4.9 + 4.7 + 4.6 + 5.0}{5} = 4.86$$

- **Subtract the mean from each data point:**

$$[(5.1 - 4.86), (4.9 - 4.86), (4.7 - 4.86), (4.6 - 4.86), (5.0 - 4.86)] = [0.24, 0.04, -0.16, -0.26, 0.14]$$

- **Square each result:**

$$[(0.24)^2, (0.04)^2, (-0.16)^2, (-0.26)^2, (0.14)^2] = [0.06, 0.002, 0.03, 0.07, 0.02]$$

- **Sum results and divide by one fewer than total number of data points:**

$$\frac{0.06 + 0.002 + 0.03 + 0.07 + 0.02}{4} = 0.046$$

Programmatically (R):

```
In [28]: var(c(5.1, 4.9, 4.7, 4.6, 5.0))
         0.043

In [ ]:
```

Mel then realizes that the variance, as a stand-alone number, is a bit difficult to interpret and visualize, even for the scientists from the Tadpole Galaxy …

But if you take the square root of the variance, the resulting value describes the average distance that each data point lies away from the mean.

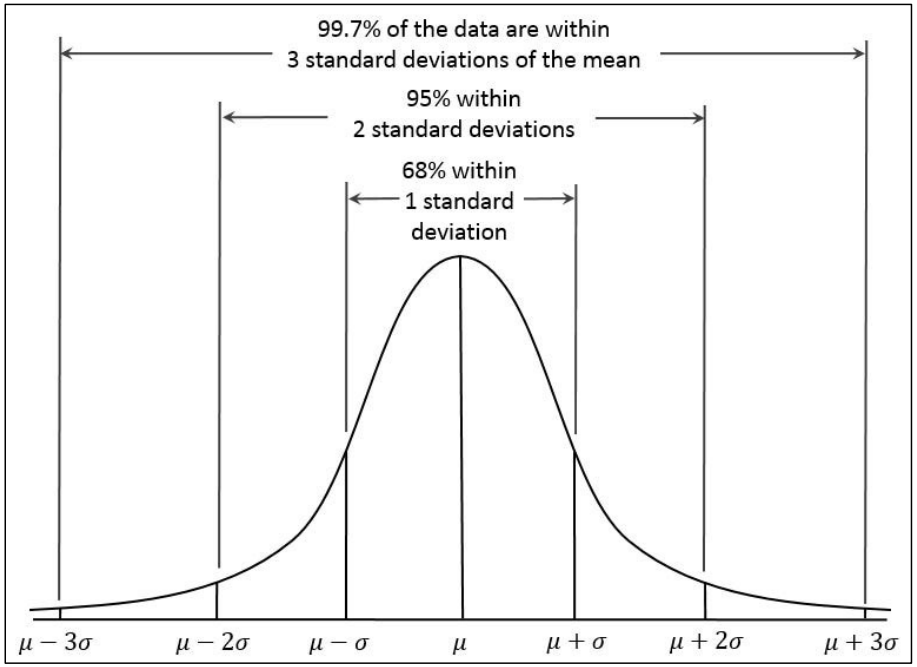
Mel calls this new metric the standard deviation (SD)

- **Equation:** $\sqrt{V}$

  V = variance

The standard deviation ("SD" or σ) is a measure of how "spread out" your data points are

Beautiful fake data distribution:



μ = mean

σ = standard deviation

- The less "typical" a data point is, the more standard deviations it is away from the mean

Let's find the standard deviation of the sepal length for iris samples 1 through 5:

- **Take the square root of the variance we just calculated:**

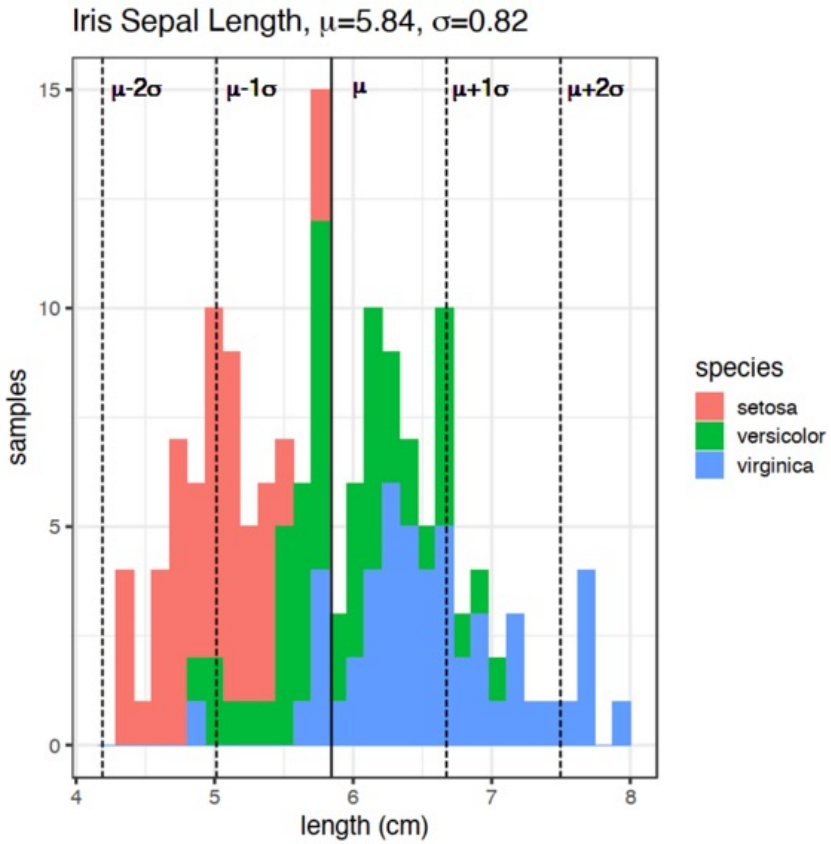
By hand: $\sqrt{0.046} = 0.21$

Programmatically (R):

```
In [40]: sd(c(5.1, 4.9, 4.7, 4.6, 5.0))
         0.207364413533277

In [ ]:
```

Let's look at the distribution of ALL the sepal length data



1. Is the SD of all the data greater or smaller than the SD of just the first 5 samples? Why do you think this is?
   Greater; because the full dataset has more spread out data points.
2. Take a look at the data between μ and $\mu-1\sigma$. Is there a species-specific pattern? What about the data between μ and $\mu-2\sigma$?
   The *setosa* species makes up the majority of the samples between μ and $\mu-1\sigma$, and almost all of the samples between μ and $\mu-2\sigma$.
3. If we removed all the *setosa* samples, would the mean of the dataset increase or decrease?
   Increase.
4. Based on this graph, does one of the iris species appear more different from the other two?
   Setosa has less overlap with the other two than they do with each other.

Mel then discovers that you can use the **fold change** metric to quantitate a difference between two groups of samples

To calculate fold change between two data points (x,y), you simply take the ratio of x over y:

$$\frac{x}{y}$$

Let's calculate the fold change between data points (6,10):

$$\frac{6}{10} = 0.6$$

Fold change between two groups of samples

- **In bioinformatics, we usually want to know the fold change of some measured variable between two groups or datasets**

- **In that case, we calculate the mean value of the variable of interest for each group, then take the fold change of the mean values.**

Let's calculate **fold change** of sepal length between all the iris species:

	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8	Sample9	Sample10	⋯	Sample141	Sample142	Sample143
Sepal.Length	5.1	4.9	4.7	4.6	5.0	5.4	4.6	5.0	4.4	4.9	⋯	6.7	6.9	5.8
Sepal.Width	3.5	3.0	3.2	3.1	3.6	3.9	3.4	3.4	2.9	3.1	⋯	3.1	3.1	2.7
Petal.Length	1.4	1.4	1.3	1.5	1.4	1.7	1.4	1.5	1.4	1.5	⋯	5.6	5.1	5.1
Petal.Width	0.2	0.2	0.2	0.2	0.2	0.4	0.3	0.2	0.2	0.1	⋯	2.4	2.3	1.9
Species	setosa	setosa	setosa	setosa	setosa	setosa	setosa	setosa	setosa	setosa	⋯	virginica	virginica	virginica

First, calculate the means:

R magic to get a LIST from a ROW in a dataframe

```
In [70]: mean(unname(unlist(subset(iris, Species == 'setosa')['Sepal.Length'])))
```

5.006

```
In [71]: mean(unname(unlist(subset(iris, Species == 'versicolor')['Sepal.Length'])))
```

5.936

```
In [72]: mean(unname(unlist(subset(iris, Species == 'virginica')['Sepal.Length'])))
```

6.588

Let's calculate **fold change** of sepal length between all the iris species:

Then, calculate each fold change:

setosa / versicolor

```
In [84]: round(5.006 / 5.936, digits=2)
```

0.84

setosa / virginica

```
In [85]: round(5.006 / 6.588, digits=2)
```

0.76

versicolor / virginica

```
In [86]: round(5.936 / 6.588, digits=2)
```

0.9

Based on the fold change values:

1. Which iris species are the most similar?
2. Which are the most different?

Log Fold Change

- **In practice, if we are looking at many variables (for example, thousands of genes) the values for fold change can be greatly spaced out and difficult to compare and visualize directly.**
- **Thus, we often take the log of fold change values for data science purposes = log fold change (LFC).**

setosa / versicolor

```
In [9]: round(log(5.006 / 5.936), digits=2)
-0.17
```

versicolor / setosa

```
In [10]: round(log(5.936 / 5.006), digits=2)
0.17
```

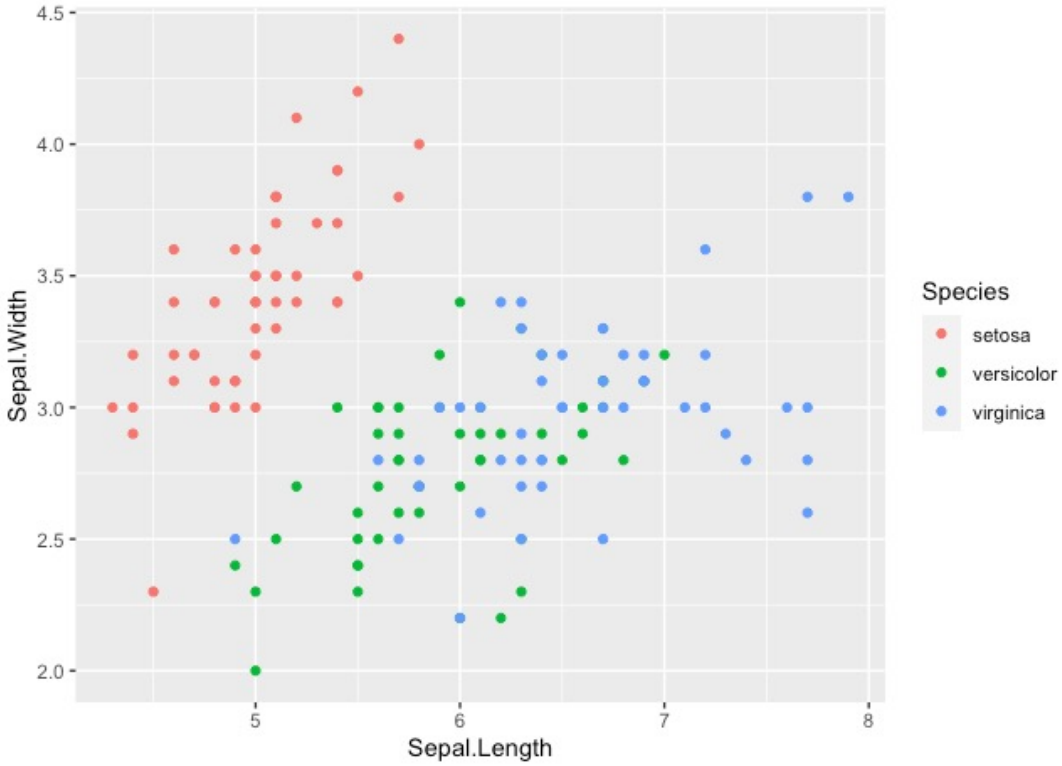
- A negative LFC indicates lower overall values in the group being compared (the numerator, *setosa* here).
- A positive LFC indicates higher overall values in the group being compared (the numerator, *versicolor* here).

Taking a look at the iris dataset again, what has Mel been ignoring so far?

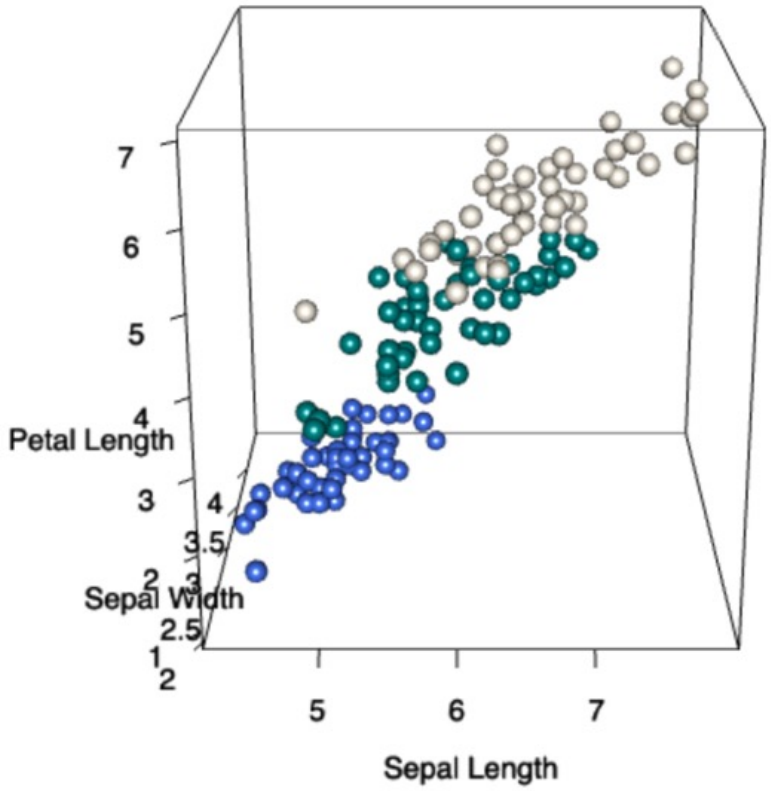
	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8	Sample9	Sample10	⋯	Sample141	Sample142	Sample143
Sepal.Length	5.1	4.9	4.7	4.6	5.0	5.4	4.6	5.0	4.4	4.9	⋯	6.7	6.9	5.8
Sepal.Width	3.5	3.0	3.2	3.1	3.6	3.9	3.4	3.4	2.9	3.1	⋯	3.1	3.1	2.7
Petal.Length	1.4	1.4	1.3	1.5	1.4	1.7	1.4	1.5	1.4	1.5	⋯	5.6	5.1	5.1
Petal.Width	0.2	0.2	0.2	0.2	0.2	0.4	0.3	0.2	0.2	0.1	⋯	2.4	2.3	1.9
Species	setosa	setosa	setosa	setosa	setosa	setosa	setosa	setosa	setosa	setosa	⋯	virginica	virginica	virginica

- **So far Mel has been using only 1 variable (Sepal Length)**
- **But this is multi-dimensional data with more than one variable**
  (1 dimension = 1 variable)
- **How can Mel work with all the variables to more completely analyze the differences between the 3 iris species?**

2D scatterplot (relationship between 2 variables)

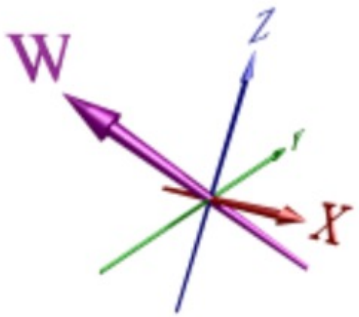


3D scatterplot (relationship between 3 variables)



https://www.r-graph-gallery.com/3d_scatter_plot.html

Even with Tadpole Galaxy technology, it is very difficult to visualize 4 dimensions…



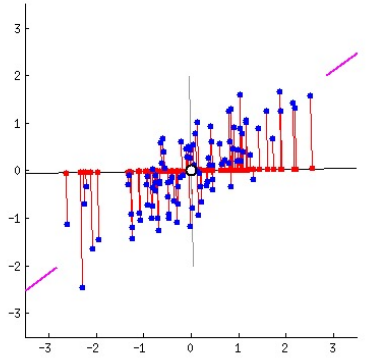
So Mel gets really ambitious and invents Principal Component Analysis (PCA)

Principal Component Analysis (PCA)

- **PCA reduces the number of dimensions (variables) by condensing all variables into a small number of vectors**
  - Vector = summary of key information which was previously spread out across many variables
- **The vectors represent the main sources of information, or variance, in the dataset**
  - The vectors are called principal components or PCs
- **High-dimensionality data are usually visualized with a scatterplot of the first 2 PCs, which represent the majority of the variance in the dataset**

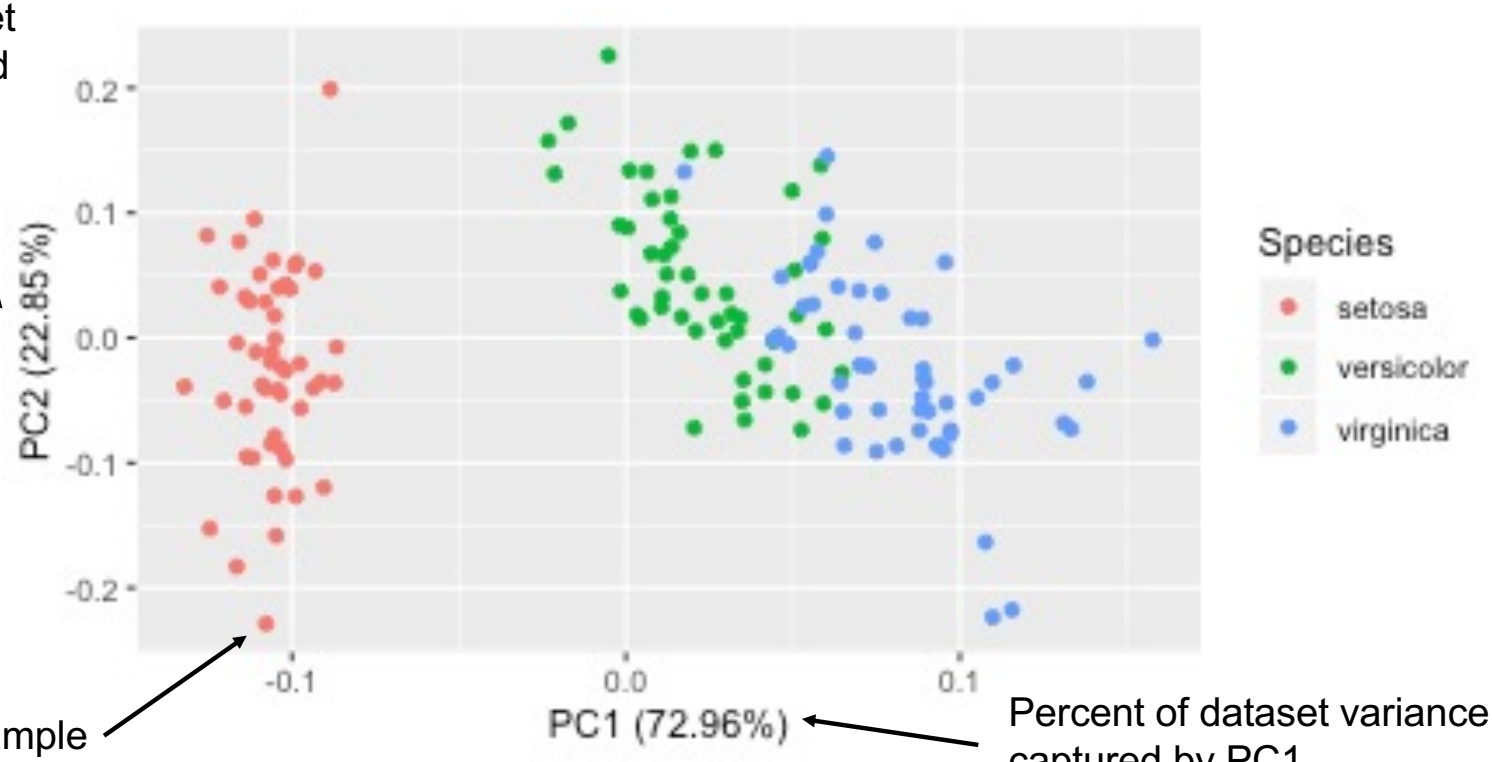
How are Principal Components calculated?

- **Calculate the "covariance matrix" between all data points**
  - Variance = how a data point varies from the mean
  - Covariance = how similarly 2 data points vary from the mean
  - E.g. is there any relationship between these 2 points (correlation)
- **Calculate the "eigenvectors" of the covariance matrix**
  - Eigenvectors = directions of the axes where there is the most variance
  - These are the PCs



PCA plot of all 4 iris variables condensed to 2 PCs

Percent of dataset variance captured by PC2

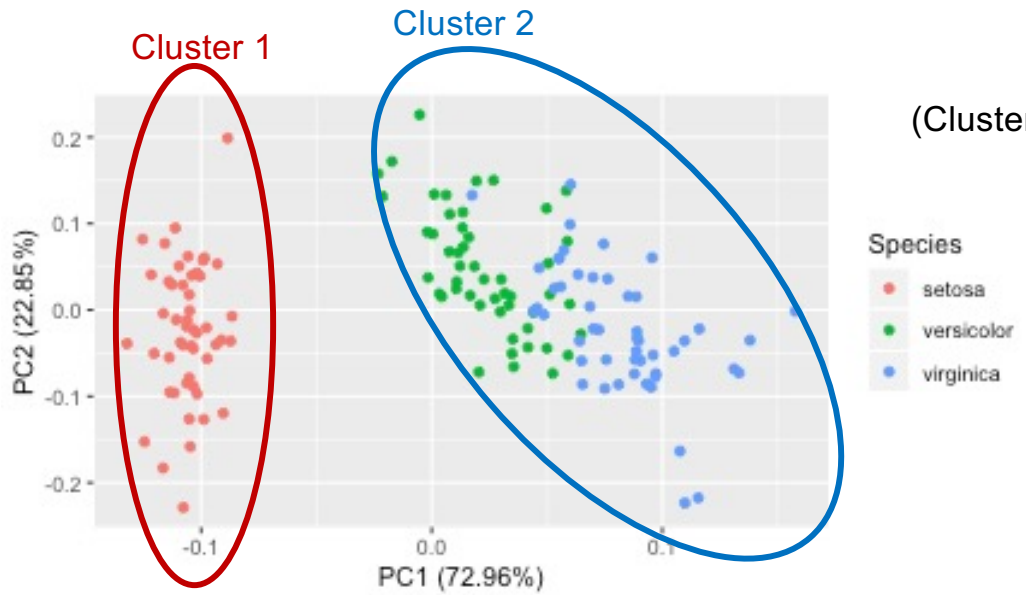


One sample

Percent of dataset variance captured by PC1

cran.r-project.org/web/packages/ggfortify/vignettes/plot_pca.html

PCA finds **clusters** of similar samples



(Clusters manually assigned)

- Mel, who knows about DNA, now thinks that the differences observed between the iris species must be due to a gene that is turned ON in *setosa* flowers but OFF in *versicolor* and *virginica*. **What is the name for this type of statement?**

A hypothesis

Hypothesis testing: What is a **hypothesis?**

- **An "educated guess" based on some data that you have**
- **Must be testable in some way**
  - –Remember the sheep hypothesis from Dr. Heller's statistics lecture?
  - –Hypothesis: all sheep are white
  - –Lots of white sheep support this hypothesis, but one purple sheep is enough to disprove it

Null Hypothesis (H_0)

- **Remember from Dr. Heller's lecture that the null hypothesis is the current, or existing explanation**
  - In science, the null hypothesis can be "the pattern we observe is just coincidence"
- **For example, Daenerys decisively disproved H_0 "Dragons are extinct"**
- **BUT in science it is almost never decisive.**

In science (not in GoT), how do we test hypotheses?

- **To accept our hypothesis, we must reject the null hypothesis**
- **If we usually can't decisively disprove the null hypothesis, how do we ever reject it?**
    1. Design and perform experiments to get data that supports our hypothesis
    2. Perform statistical analysis on the data to see whether our results occurred by chance
    3. If we can demonstrate that our results likely did not occur by chance, we can reject the null hypothesis

Null Hypothesis (H_0)

What is the null hypothesis for Mel's hypothesis: "*I hypothesize that the differences I observe between setosa and versicolor and virginica are due to a gene with that is turned <u>on</u> in setosa but <u>off</u> in the other species.*"

- "The differences observed between *setosa* and *versicolor* and *virginica* are not attributable to a gene turned on in *setosa* but off in the other species."

What is a good experiment Mel could perform to test this hypothesis?

- Extract RNA from several replicates of each iris species, perform RNA sequencing

P-value (“probability value”)

- **The p-value is the probability that your null hypothesis is true, given the data you get from your experiment.**
  - The *lower* the p-value, the *less likely* the null hypothesis is to be true.

- **A p-value of less than 0.05 is generally accepted as low enough to reject the null hypothesis (because there is only 5% chance it is true) and accept an alternate hypothesis.**
  - A p-value<0.05 indicates *statistical significance*, i.e. that the outcome of our experiment was likely not due to random chance.

“Proving” a hypothesis?

Given what we have just learned, is it ever possible to definitively prove a hypothesis using hypothesis testing?

NO, the best we can do is to assign a probability to the null hypothesis being true.

Adjusted p-value

- **If Mel performed RNA sequencing on the irises…**
  - –Some iris genomes have over 100,000 genes!
  - –Each gene is a variable
  - –RNA sequencing would give ~100,000 variables (genes) to compare between the iris species

- **For data with thousands of variables, it is important to mathematically adjust p-values for multiple comparison testing to avoid false positives.**
  - – Multiple comparison testing = comparing multiple variables between conditions

Why is it important to adjust p-values for multiple comparisons?

- **If Mel looked at the difference in expression of a random set of 20 genes:**
  - –*By chance* we would expect that on average 1 out of 20 genes would appear significant using the 0.05 (5%) p-value cutoff.
  - –This is a false positive - a result that appears significant but in reality is due to random chance.
  - –Mathematical p-value adjustment takes into account the probability of false positives given the number of variables

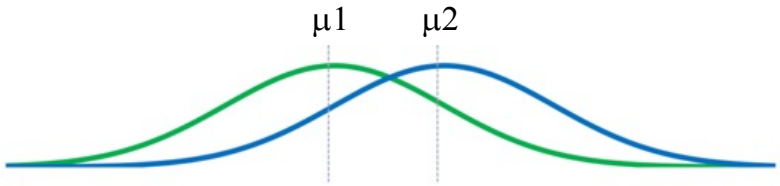
- **Many bioinformatics statistical tools will generate, for each variable:**
  - –a p-value (does not take false positives into account)
  - –an adjusted p-value (does take false positives into account)

Adjusted p-value questions

1. **What should be used when determining the significance of an experiment where ~100,000 genes (aka variables) are considered, p-value or adjusted p-value, and why?**
   – Adjusted p-value, because it takes into account the likelihood of false positives.
2. **When is it appropriate to use a p-value rather than an adjusted p-value to determine significance, given a p-value cutoff of 0.05 (1/20)?**
   – In an experiment where there are few enough variables that false positives are not reasonably expected. If you are using a p-value cutoff of 0.05 (i.e., 1 out of 20), you would want far fewer variables than 20.
3. **Which will generate more significantly differentially expressed genes in Mel's experiment, the p-value or the adjusted p-value?**
   – The p-value, since it is a less stringent metric.

How do we actually calculate a p-value?

- **Use a statistical hypothesis testing method such as a T-Test**
  - –Compares the means of all measured variables between two groups



Expression of Gene X in *setosa*

Expression of Gene X in *virginica*

$$tScore = \frac{signal}{noise} = \frac{difference\ in\ group\ means}{variability\ of\ groups} = \frac{\mu_1 - \mu_2}{\sigma_1 - \sigma_2}$$

μ_1 = mean expression of gene X in *setosa*
μ_2 = mean expression of gene X in *virginica*

σ_1 = standard deviation of gene X in *setosa*
σ_2 = standard deviation of gene X in *virginica*

Standard deviation comes in handy again!

What are the results returned from a T-test?

- **A *t-score* / *t-value* for each variable**
  - –The greater the t-value, the more difference there is within that variable between the two groups

- **A p-value associated with each t-value**
  - –Describes the likelihood of the null hypothesis being true for that variable

Another hypothesis testing method: **Wald test**

- **Creates a mathematical model for the null hypothesis and the testing hypothesis, and identifies which model best maximizes the likelihood of the data**
  - Mathematical model = mathematical description of a situation, specifying rules about how data should behave within that situation.

- **Put another way: the Wald test identifies a mathematical model for which the data we have is *the most likely result***

- **Returns:**
  - a Wald statistic for each variable
  - a p-value associated with each Wald statistic

Wald statistic

- **A Wald statistic is like a *t-score***
  - –A value that describes whether we can reject the null hypothesis for that variable

- **The Wald statistic is useful for ranking genes for downstream applications**
  - –Gene Set Enrichment Analysis (GSEA) identifies gene sets that involve a ranked gene list
  - –Ingenuity Pathway Analysis (IPA) finds biological pathways enriched in a gene data set

Questions about T-tests and Wald Tests

1. **Suppose Mel's final gene expression data has measurements from 100,000 genes. If Mel performs a T-test to compare the two groups, how many t-values (and p-values) are expected?**
   – We would expect 100,000 t-values and 100,000 p-values: one for each gene, because in gene expression data, each gene is a variable.

2. **If Mel performs a Wald test to compare these two groups, how many Wald statistics and p-values are expected?**
   – We would expect 100,000 of each.

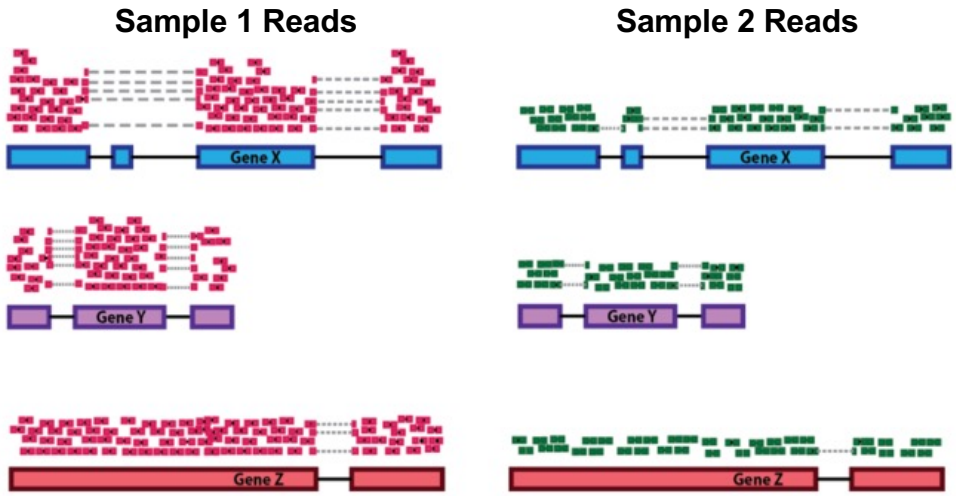
Normalizing gene expression data for differences in **read depth**

- **Gene expression data often have different *read depth* levels per sample**
  - –Read depth = the total number of reads that come off the sequencer for each sample

- **The number of reads in a sequencing run is split up between the number of samples**
  - –Thus, the read depth for each sample depends on the total number of samples

Sample 1 has twice as many reads as Sample 2

- **If Mel directly compares Sample 1 and Sample 2 without normalization, it looks like all genes are more highly expressed in Sample 1**



- **Thus, Mel first must normalize the samples by their read depth, to allow for direct comparison of gene expression values between samples.**

Normalization with Median of Ratios Method (DESeq2 Step 1)

➢**DESeq2 normalizes gene expression data *by read depth* using the Median of Ratios method.**

1. For each gene, a pseudo-reference sample is created that is equal to the geometric mean across all samples.

$$GM = \sqrt[n]{x_1 * x_2 * x_3 \ldots * x_n}$$

gene	Sample1	Sample2	pseudo-reference sample
EF2A	1489	906	sqrt(1489 * 906) = **1161.5**
ABCD1	22	13	sqrt(22 * 13) = **17.7**
...	...	...	...

https://hbctraining.github.io/DGE_workshop/lessons/02_DGE_count_normalization.html

Normalization with Median of Ratios Method

2. For every gene in a sample, the ratios (sample/pseudo-reference) are calculated. This is performed for each sample in the dataset. Since the majority of genes are not differentially expressed, the majority of genes in each sample should have similar ratios within the sample.

gene	Sample1	Sample2	pseudo-reference sample	ratio of Sample1/ref	ratio of Sample2/ref
EF2A	1489	906	1161.5	1489/1161.5 = **1.28**	906/1161.5 = **0.78**
ABCD1	22	13	16.9	22/16.9 = **1.30**	13/16.9 = **0.77**
MEFV	793	410	570.2	793/570.2 = **1.39**	410/570.2 = **0.72**
BAG1	76	42	56.5	76/56.5 = **1.35**	42/56.5 = **0.74**
MOV10	521	1196	883.7	521/883.7 = **0.590**	1196/883.7 = **1.35**
...	...	...	...		

Normalization with Median of Ratios Method

3. The median (middle) value of all ratios for a given sample is taken as the normalization factor (size factor) for that sample.

gene	Sample1	Sample2	pseudo-reference sample	ratio of Sample1/ref	ratio of Sample2/ref
EF2A	1489	906	1161.5	1489/1161.5 = **1.28**	906/1161.5 = **0.78**
ABCD1	22	13	16.9	22/16.9 = **1.30**	13/16.9 = **0.77**
MEFV	793	410	570.2	793/570.2 = **1.39**	410/570.2 = **0.72**
BAG1	76	42	56.5	76/56.5 = **1.35**	42/56.5 = **0.74**
MOV10	521	1196	883.7	521/883.7 = **0.590**	1196/883.7 = **1.35**
...	...	...	...		

```
normalization_factor_Sample1 <- median(c(1.28, 1.3, 1.29, 1.35, 0.59))
normalization_factor_Sample2 <- median(c(0.78, 0.77, 0.72, 0.74, 1.35))
```

Normalization with Median of Ratios Method

4. Each raw count value in each sample is divided by that sample's normalization factor to generate normalized count values.

Raw Counts

gene	Sample1	Sample2
EF2A	1489	906
ABCD1	22	13
...	...	...

Normalized Counts

gene	Sample1	Sample2
EF2A	1489 / 1.3 = **1145.39**	906 / 0.77 = **1176.62**
ABCD1	22 / 1.3 = **16.92**	13 / 0.77 = **16.88**
...	...	...

Normalization with Median of Ratios Method

1. **If we did not perform read depth normalization, what result would you expect if we compared gene expression values between Sample 1 (read depth = 5 million reads) and Sample 2 (read depth = 20 million reads)?**

   Most, if not all of the differentially expressed genes would display overexpression in Sample 2 because more raw counts are assigned to each gene in Sample 2 due to deeper sequencing depth.

2. **Why would this result be problematic?**

   The differentially expressed genes may or may not be representative of the true biology of the two groups, and it would be impossible to tell which genes' overexpression were due to true biology and which were due to the sequencing depth difference.

Differential expression analysis in DESeq2

- **After normalization, Mel is ready to perform differential expression analysis**

- **~100,000 hypothesis tests run at once (one for each gene)**

- **DESeq2 uses a Wald test for differential expression analysis hypothesis testing**

- **What is the null hypothesis we are testing for each gene?**

  "This gene does not change expression significantly between the tested groups"

Estimating gene-wise variance/dispersion before Wald test (DESeq2 Step 2)

- **For genes with low mean counts, variance/dispersion estimates are often unreliable**

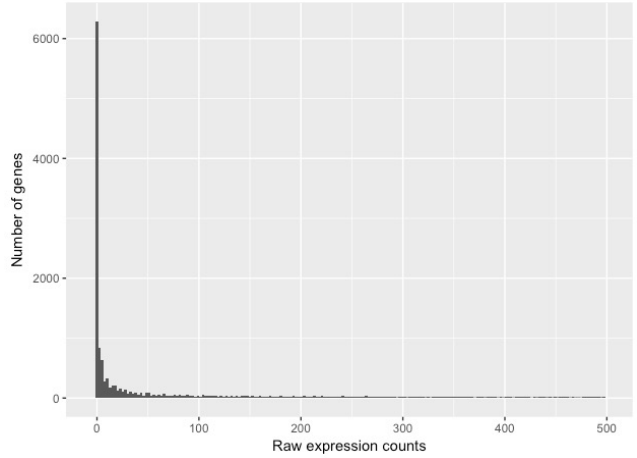
- **To solve this problem, DESeq2 estimates variance/dispersion by sharing information across genes**

  - With this approach, DESeq2 assumes that genes with similar expression levels have similar dispersion/variance

- **The dispersion for each gene is estimated using maximum likelihood estimation. This means that the most likely estimate of dispersion is calculated for each gene, given the gene expression values (aka gene counts) from all the replicates in each group.**

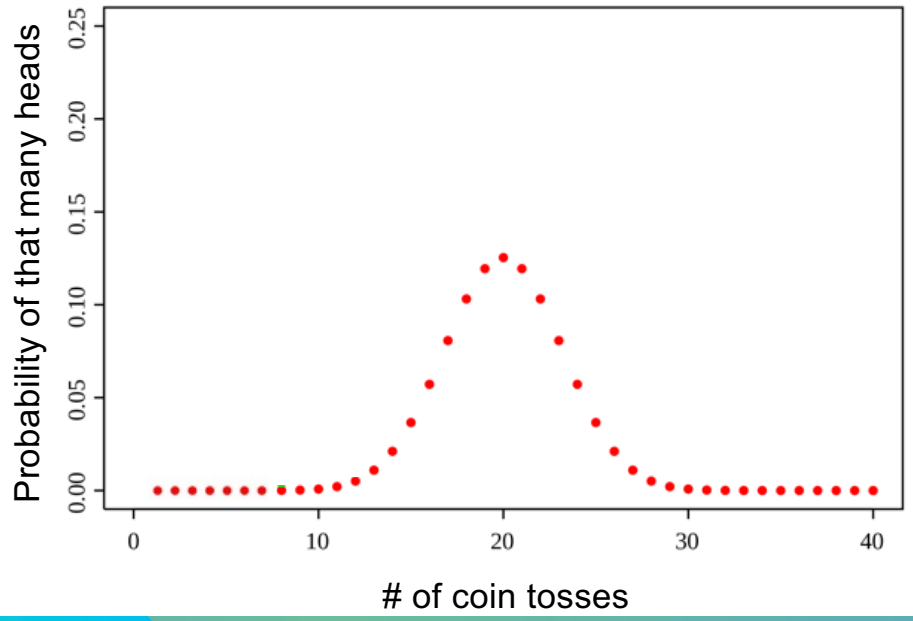
Picking a mathematical model for a Wald test on gene counts data (DESeq2 Step 3)

- **Recall that the Wald test creates a mathematical model for the null hypothesis and the testing hypothesis, and identifies which model best maximizes the likelihood of the data**
- **The type of model is dependent on the shape ("distribution") of the data**
  - A model = a distribution shape, and the numbers that specify that shape (e.g. mean and variance)
  - The Wald test has to be told which general type of distribution to use
- **RNAseq data has a specific shape: the majority of the genes have an expression value close to 0, and there is a long right tail because there is no upper limit for maximum expression value:**



Count data is sometimes modeled with the **binomial distribution**

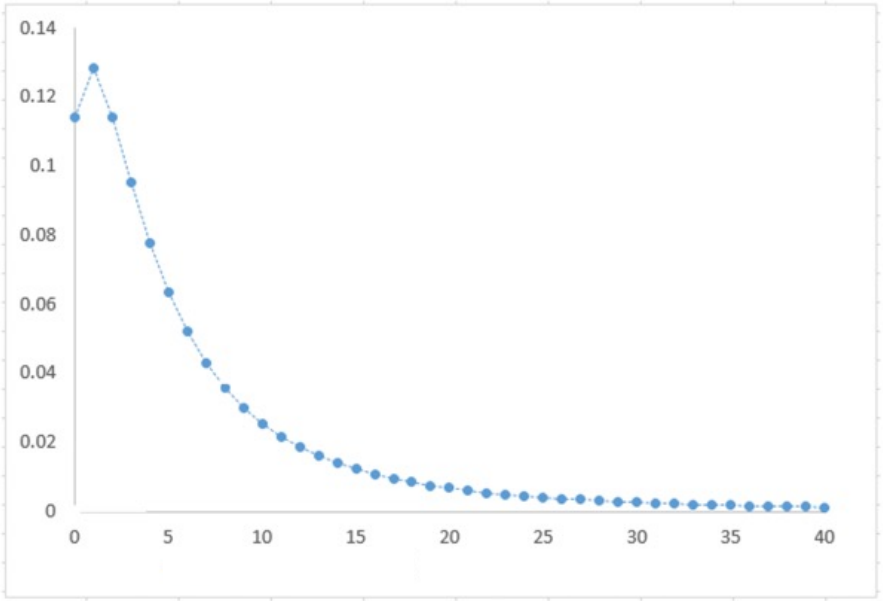
- **The binomial distribution is the “coin toss” distribution**
- **Only appropriate for relatively small number of data points**

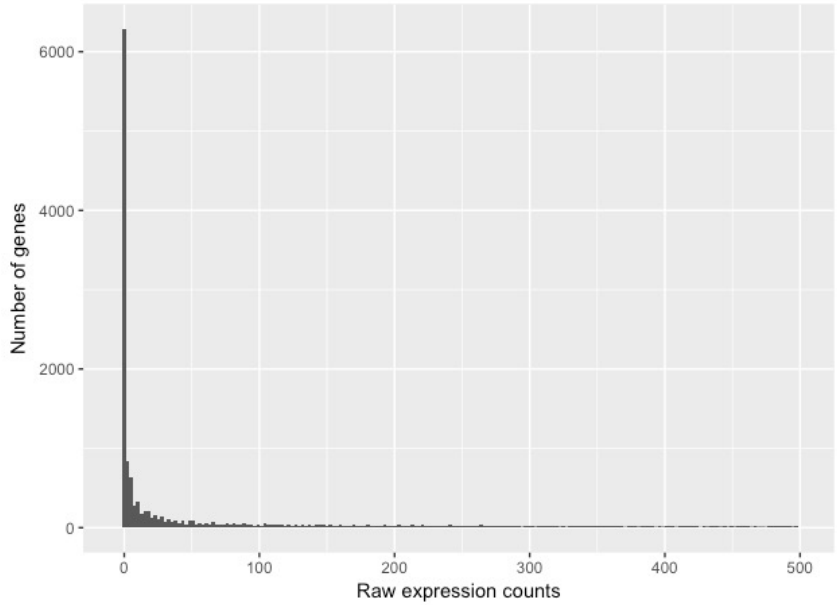


https://en.wikipedia.org/wiki/Binomial_distribution#/media/File:Binomial_distribution_pmf.svg

The negative binomial distribution is appropriate for RNA sequencing data

- **Therefore, DESeq2 uses the negative binomial distribution to create models in the Wald test for differential gene expression analysis**





https://www.vosesoftware.com/riskwiki/BetaNegBindistribution.php

Visualizing RNA sequencing data

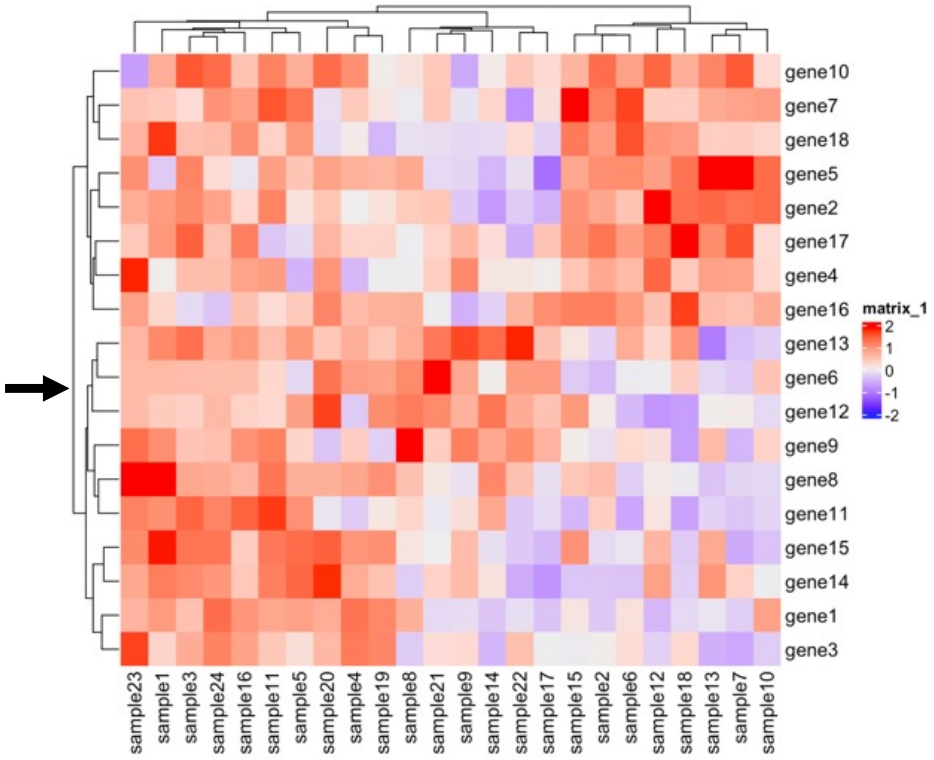
- **After differential expression analysis, we may have hundreds or thousands of differentially expressed genes**

- **For the human eye, it is easier to interpret a visualization or figure**

- **Two common visualizations for RNA sequencing are**
  - Heatmap
  - Volcano plot

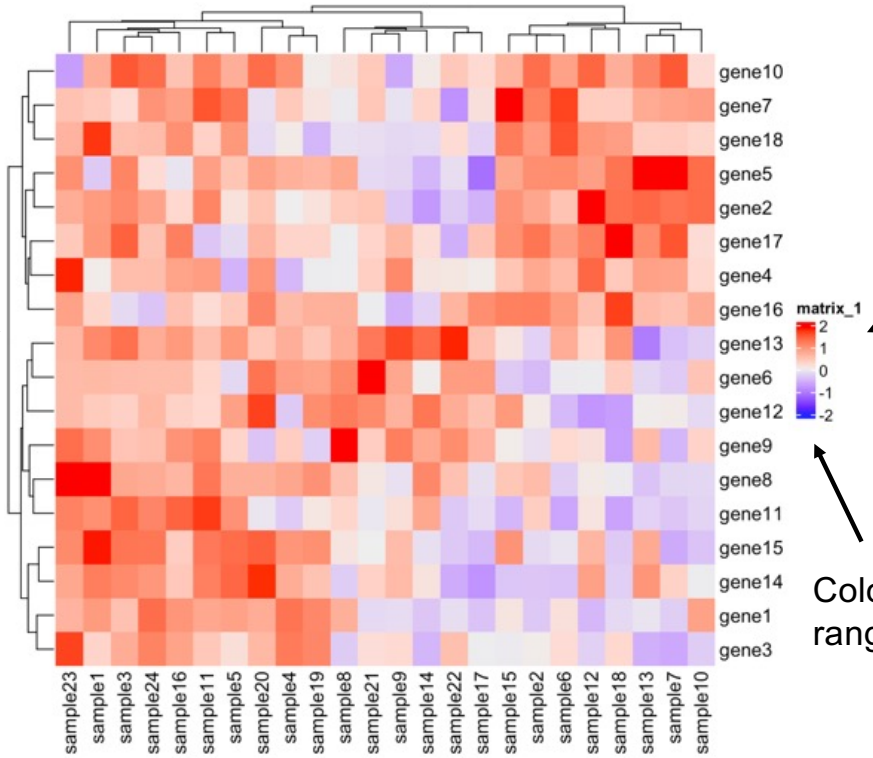
Heatmap

	sample 1	sample 2	sample 3	sample 4	sample 5	sample 6	sample 7	sample 8	sample 9	…
gene1	0.9	-0.35	0.5	1.27	0.83	0.16	-0.29	0.68	-0.16	…
gene2	0.91	0.79	1.07	0.01	0.14	0.47	1.28	0.39	-0.35	…
gene3	0.28	0.03	0.71	1.22	0.17	0.21	-0.64	-0.31	0.22	…
gene4	0.03	0.76	0.53	-0.5	-0.53	0.57	0.81	-0.01	1.08	…
gene5	-0.33	1.03	1.12	0.67	0.45	1.04	2.62	0.76	-0.23	…
gene6	0.58	-0.47	0.55	0.87	-0.2	-0.04	-0.32	1.06	0.78	…
gene7	0.4	1.15	0.21	0.41	1.26	1.63	0.81	-0.04	-0.1	…
gene8	2.2	0.55	0.77	0.79	0.68	-0.29	-0.22	0.5	-0.11	…
gene9	1.03	-0.13	0.46	0.37	0.28	0.23	-0.51	1.89	1.18	…
gene10	0.69	1.32	1.5	1.01	0.69	0.83	1.49	0.13	-0.63	…
gene11	1.01	0.35	1.41	-0.32	0.99	-0.66	-0.36	0.27	0.17	…
gene12	0.39	0.06	0.33	-0.33	0.84	-0.48	0.05	1.2	0.66	…
gene13	1.08	-0.25	1.3	0.69	0.92	0.72	-0.42	0.72	1.61	…
gene14	1.19	-0.36	1.07	0.72	1.38	-0.39	0.31	-0.3	0.57	…
gene15	1.82	-0.18	1.26	0.95	1.34	-0.09	-0.62	0.12	0.55	…
gene16	0.27	1.18	-0.18	0.56	0.4	0.9	0.48	0.68	-0.56	…
gene17	0.93	1.26	1.43	0.29	-0.18	0.88	1.53	-0.02	0.6	…
gene18	1.71	0.89	0.51	0.05	0.93	1.55	0.37	-0.11	-0.2	…



https://jokergoo.github.io/ComplexHeatmap-reference/book/a-single-heatmap.html

Heatmap



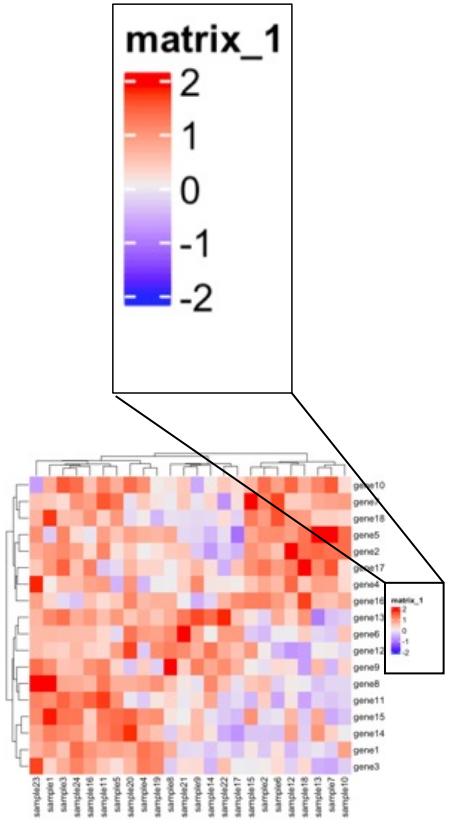
Columns clustered based on similarity of values

Legend/color bar

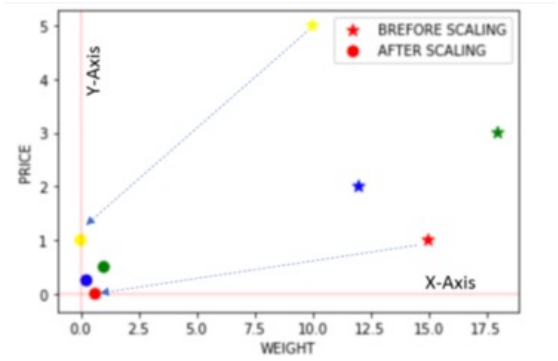
Rows clustered based on similarity of values

Colors mapped to the range of values in the matrix

Scaling data



- **Do you think that these data values naturally fall between -2 and 2?**



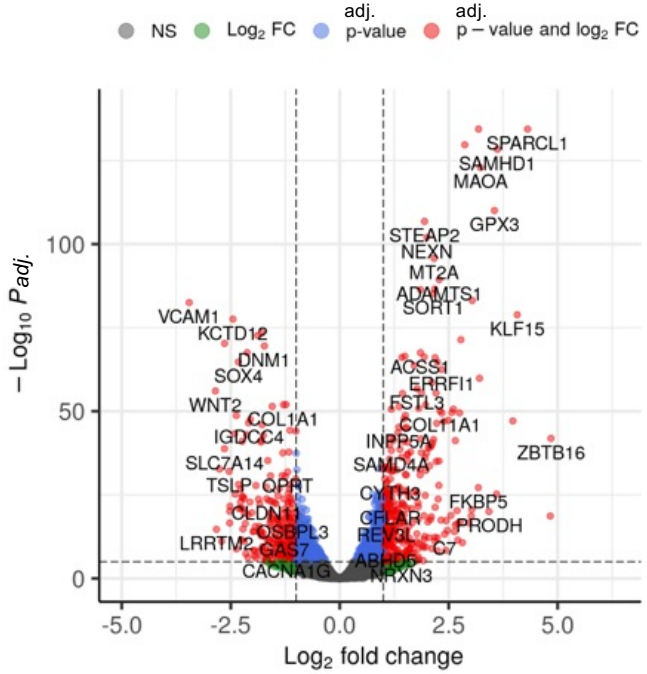
- **These data have been SCALED**
- **Pre-scaling, gene count values can range from ~30,000 to 1. The larger values can skew the data non-representatively**
- **Scaling brings all the values into similar proportions so they can be reasonably compared**

Volcano Plot

Results of a differential expression analysis:

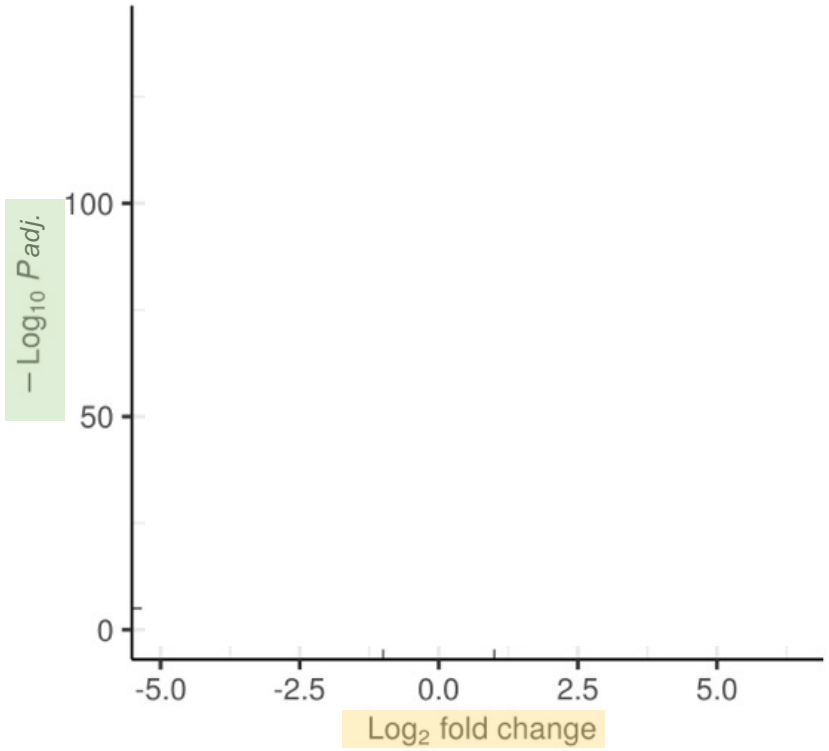
	log2FoldChange	pvalue	padj
TSPAN6	-0.38	0	0
TNMD	NA	NA	NA
DPM1	0.2	0.07	0.19
SCYL3	0.03	0.81	0.91
C1orf112	-0.09	0.75	0.88
FGR	-0.08	0.69	NA
CFH	0.42	0	0
FUCA2	-0.24	0.01	0.03
GCLC	-0.05	0.76	0.88
NFYA	-0.49	0	0
STPG1	-0.12	0.48	0.69
...	...	...	...

genes



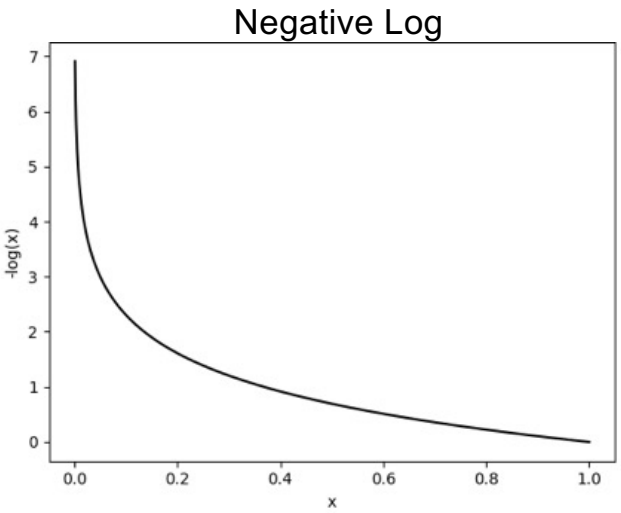
Volcano Plot

	log2FoldChange	pvalue	padj
TSPAN6	-0.38	0	0
TNMD	NA	NA	NA
DPM1	0.2	0.07	0.19
SCYL3	0.03	0.81	0.91
C1orf112	-0.09	0.75	0.88
FGR	-0.08	0.69	NA
CFH	0.42	0	0
FUCA2	-0.24	0.01	0.03
GCLC	-0.05	0.76	0.88
NFYA	-0.49	0	0
STPG1	-0.12	0.48	0.69
…	…	…	…

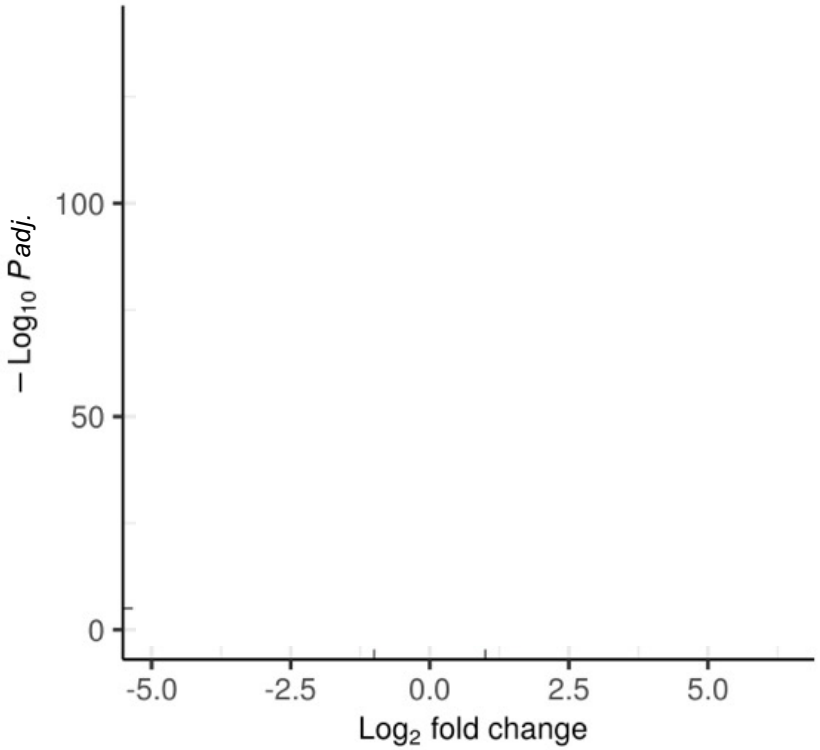


https://livmiranda921.github.io/notebook/2017/08/13/softmax-and-the-negative-loglikelihood/

Volcano Plot



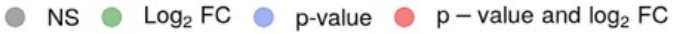
- As the adjusted pvalue gets smaller, the negative log adjusted pvalue gets bigger!



https://livmiranda921.github.io/notebook/2017/08/13/softmax-and-the-negative-log-likelihood/

Volcano Plot

<u>Significance cutoffs:</u>

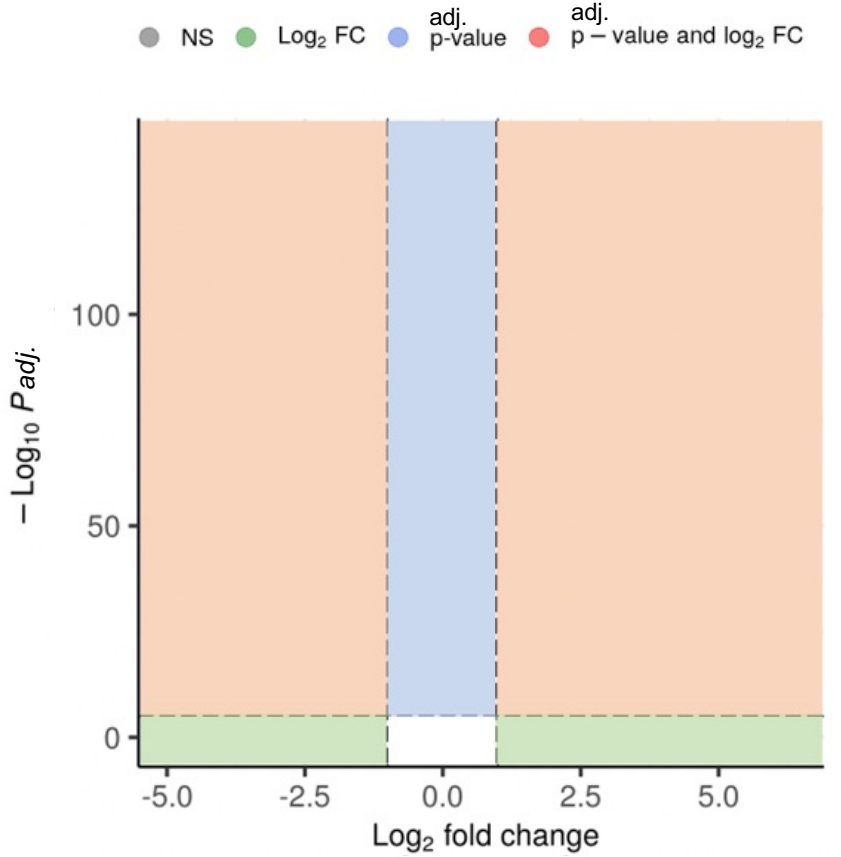


On the plot, these cutoffs are shown with dotted lines

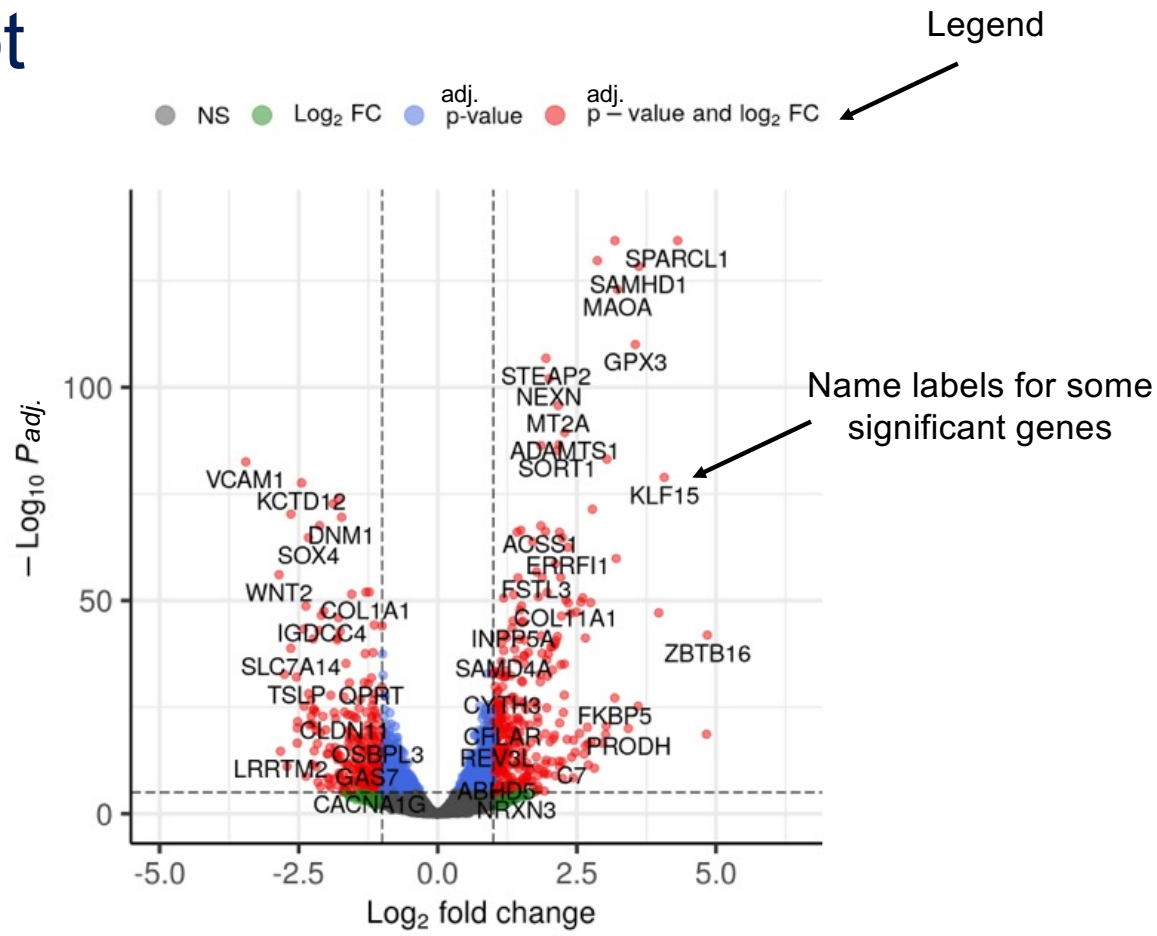
A gene is colored green if:
- Its LFC between the compared conditions is >|2|

A gene is colored blue if:
- Its adj pvalue is <10e-6

A gene is colored red if:
- Its LFC is >|2| AND its adj p value is <10e-6



Volcano Plot



Now you are ready to run DESeq2 on your gene counts data to find differentially expressed genes between spaceflight and ground control conditions!