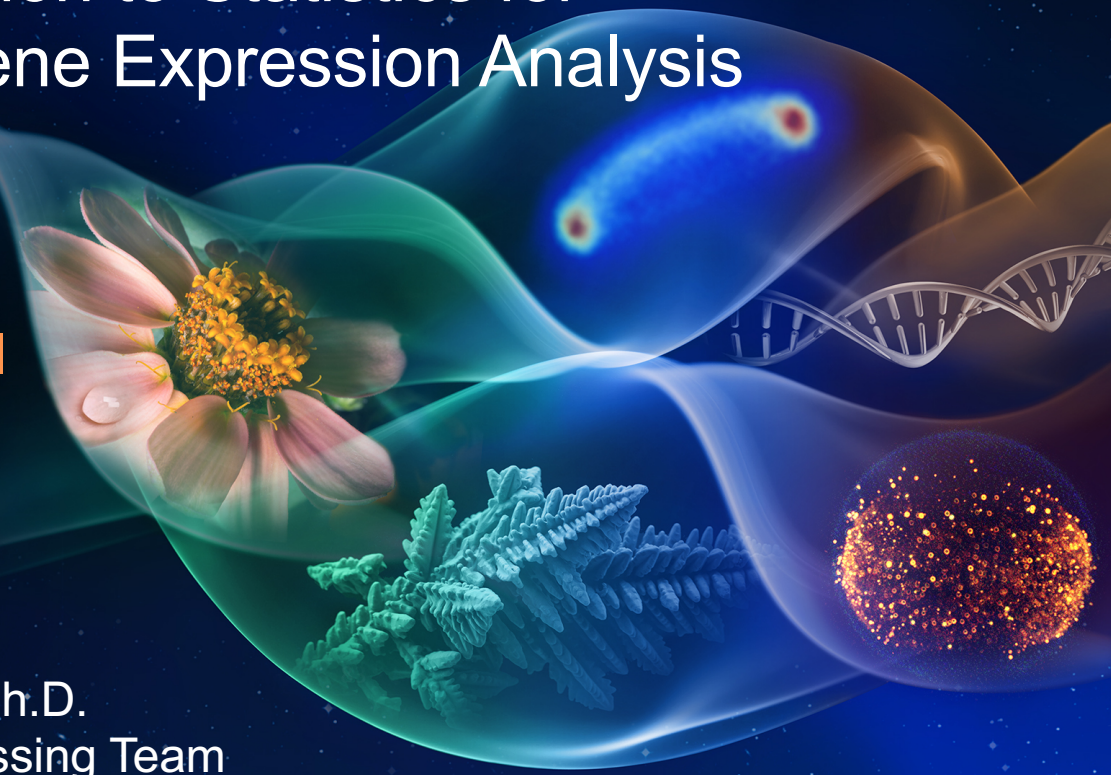
National Aeronautics and Space Administration

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

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)

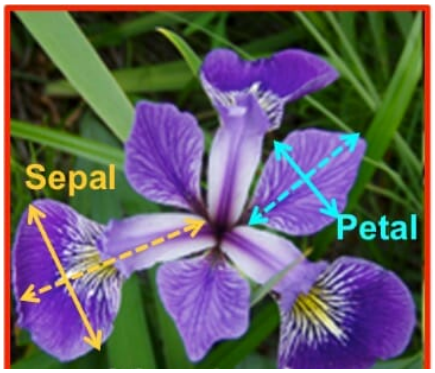
Mel, an advanced carbon-based life form, sets out from a planet in the Tadpole Galaxy, 420 million light years from Earth…

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…

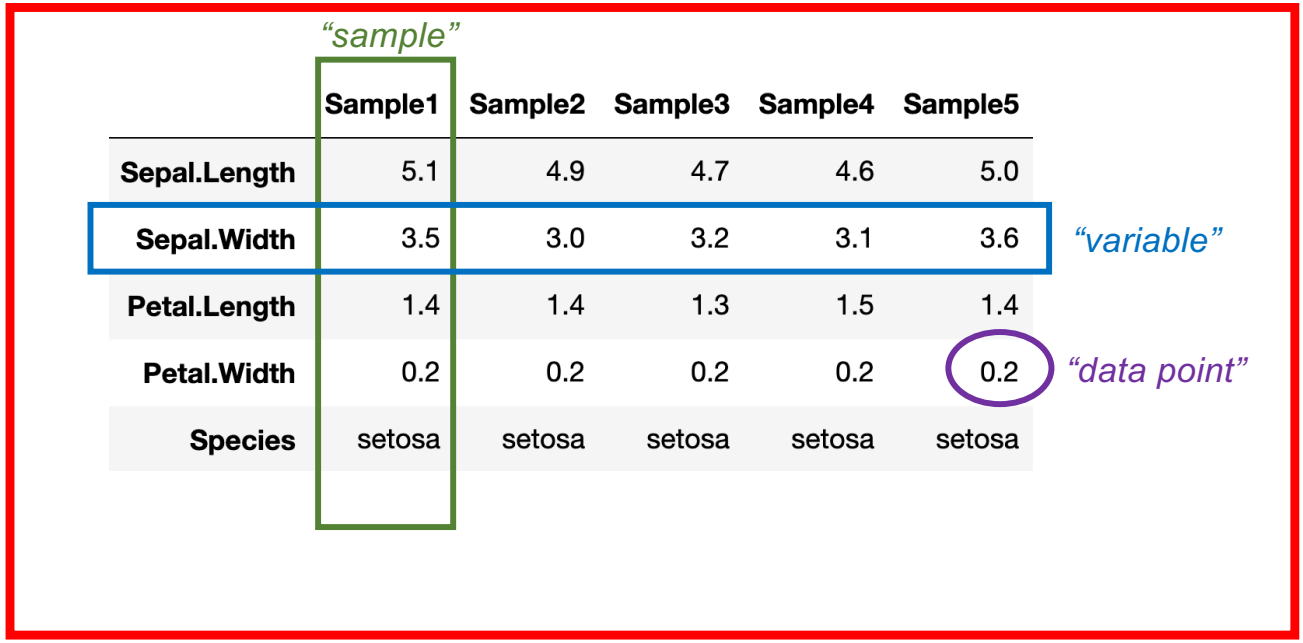


"Iris" dataset from R (first 5 samples):

	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
```

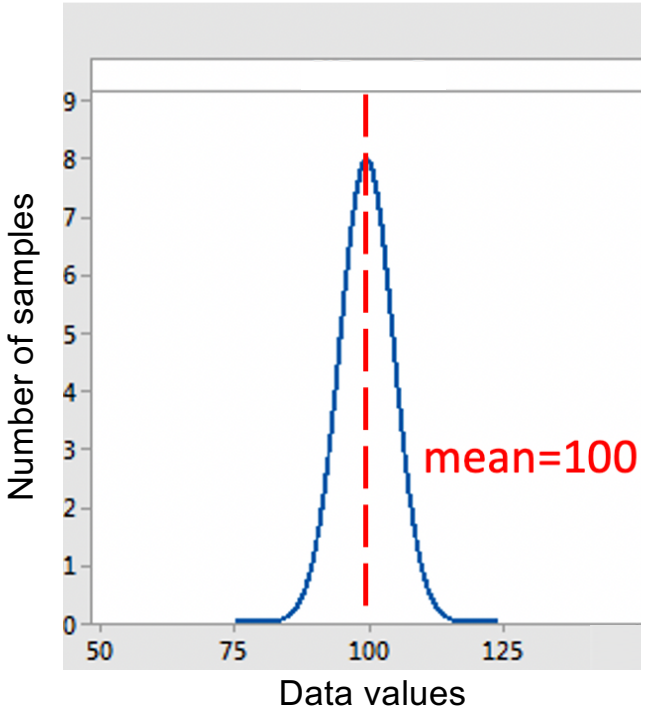
```
In [ ]:
```

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

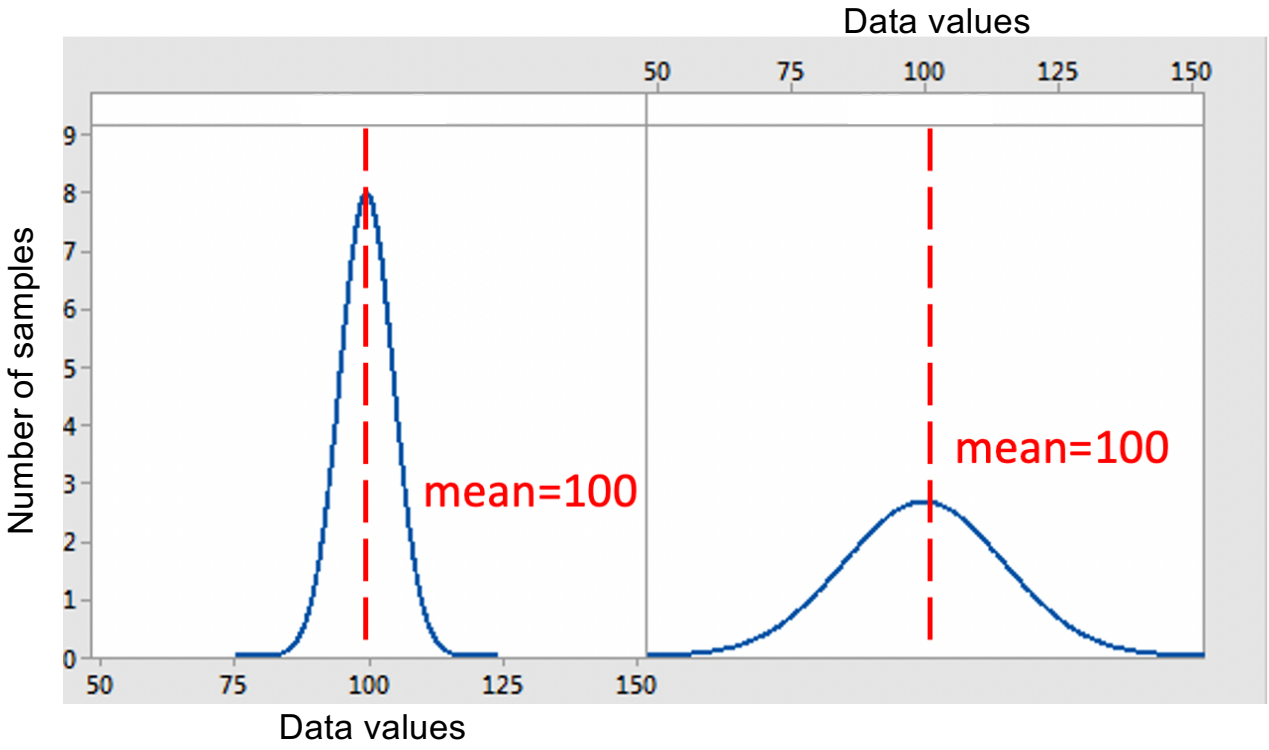
Mel invents another quantification method, the variance or dispersion of a set of numbers:

- 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!



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.

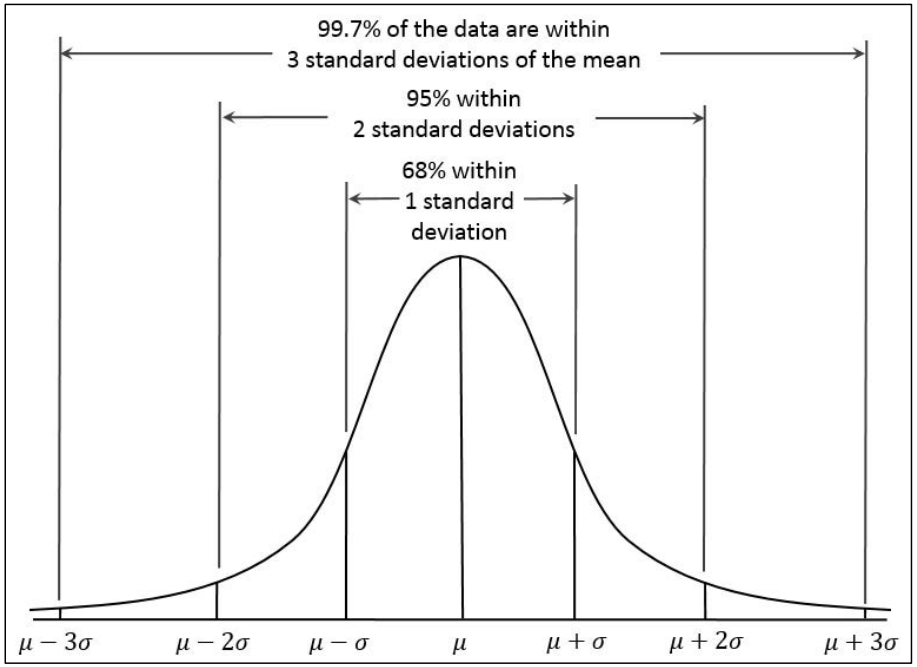
Mel calls this new metric the standard deviation (SD)

- **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: $\sqrt{0.046} = 0.21$

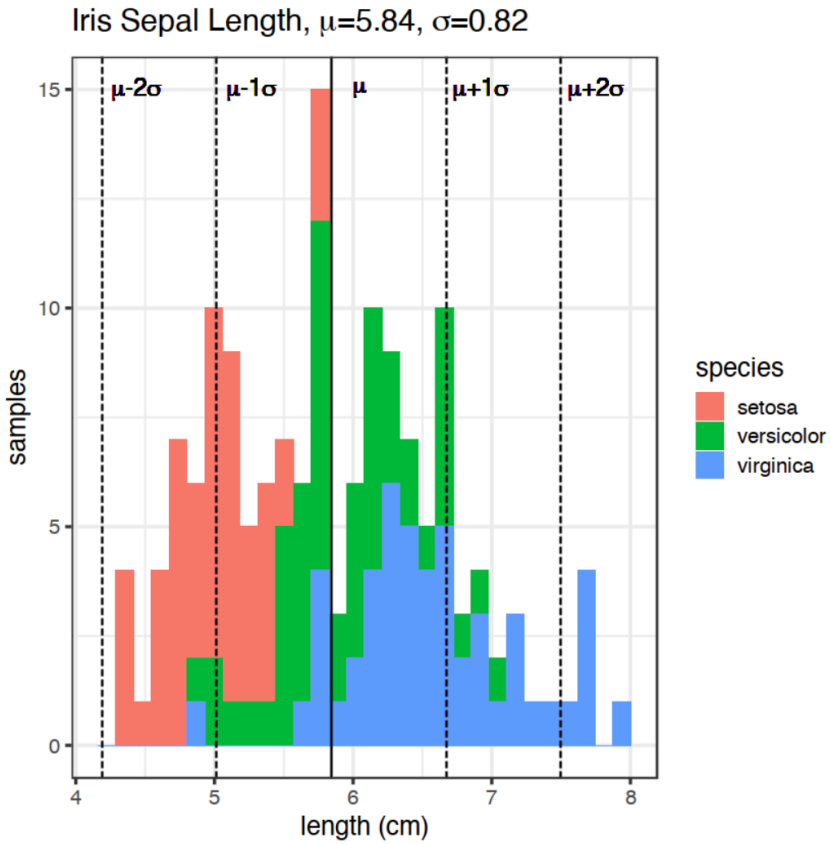
Programmatically (R):

```
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

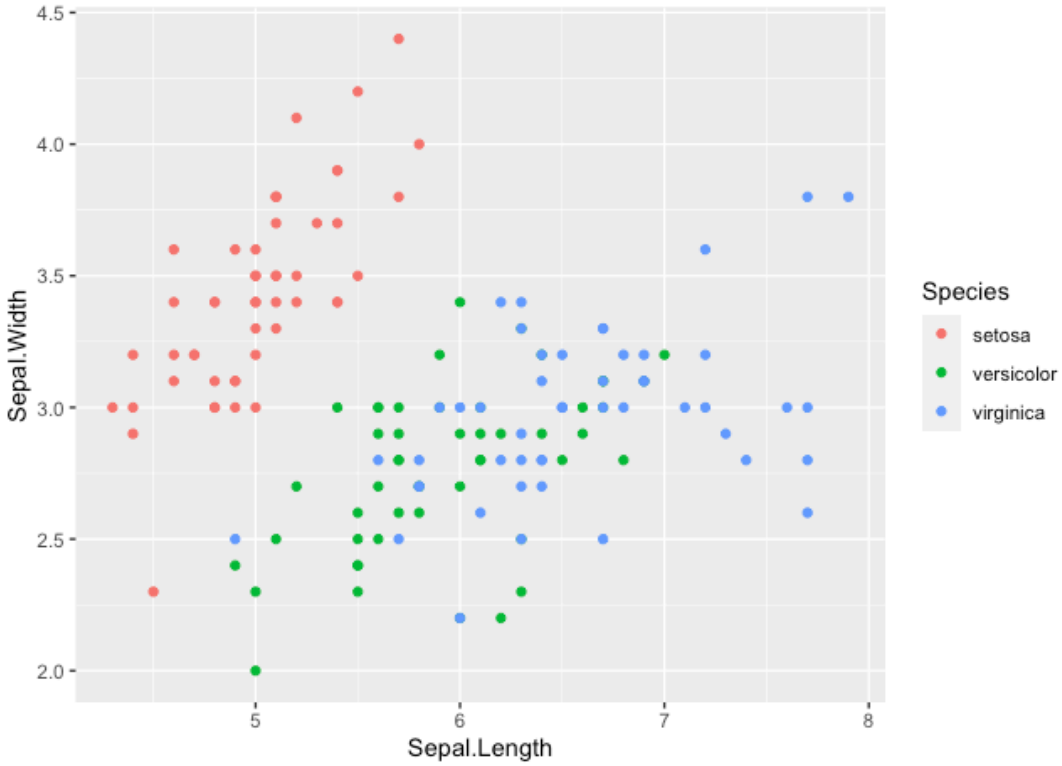
- A negative LFC indicates lower overall values in the group being compared (the numerator, *setosa* here).
- A positive LFC indicates higher overall values in the group being compared (the numerator, *versicolor* here).

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

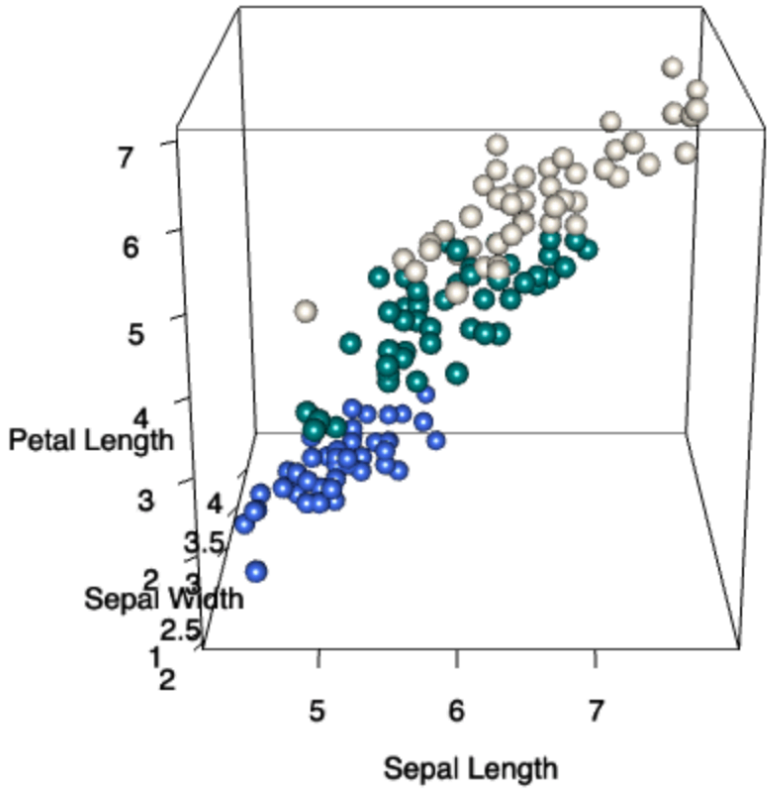
	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6	Sample7	Sample8	Sample9	Sample10	⋯	Sample141	Sample142	Sample143
Sepal.Length	5.1	4.9	4.7	4.6	5.0	5.4	4.6	5.0	4.4	4.9	⋯	6.7	6.9	5.8
Sepal.Width	3.5	3.0	3.2	3.1	3.6	3.9	3.4	3.4	2.9	3.1	⋯	3.1	3.1	2.7
Petal.Length	1.4	1.4	1.3	1.5	1.4	1.7	1.4	1.5	1.4	1.5	⋯	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)

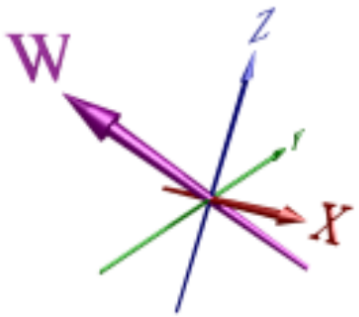


3D scatterplot (relationship between 3 variables)



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

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



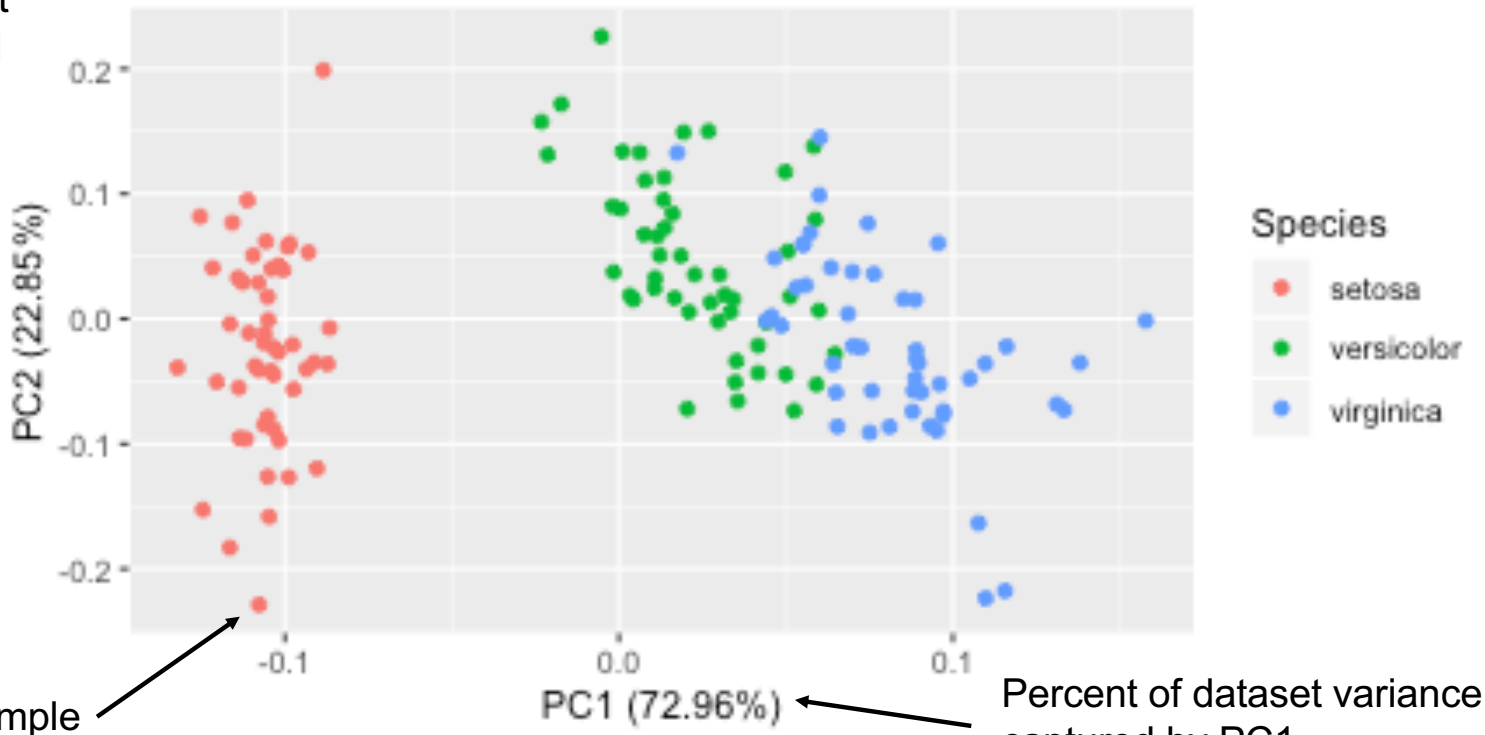
So Mel gets really ambitious and invents Principal Component Analysis (PCA)

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**

PCA plot of all 4 iris variables condensed to 2 PCs

Percent of dataset variance captured by PC2

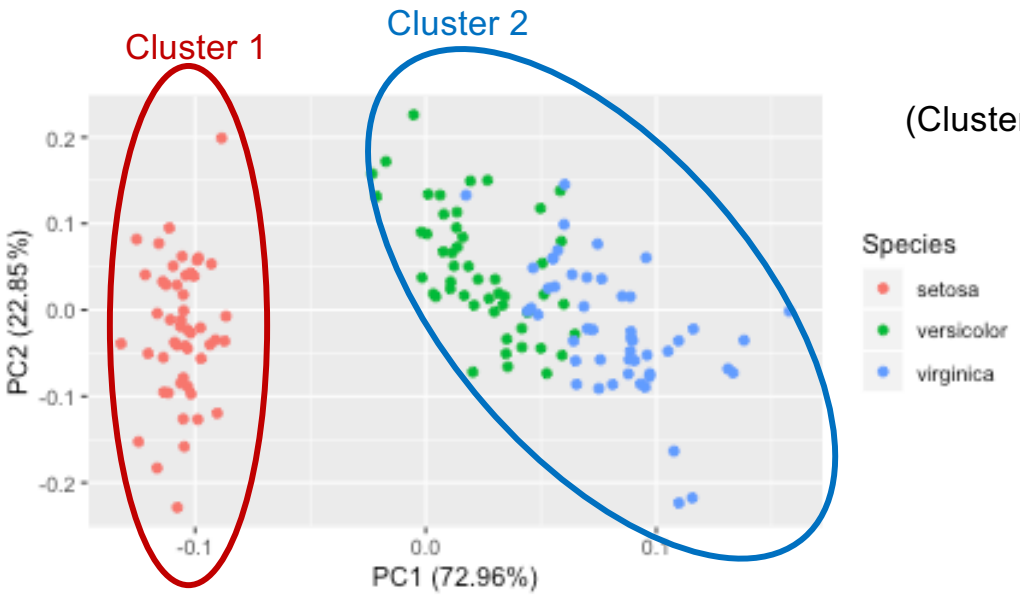


One sample

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.