Data Science Unit 2

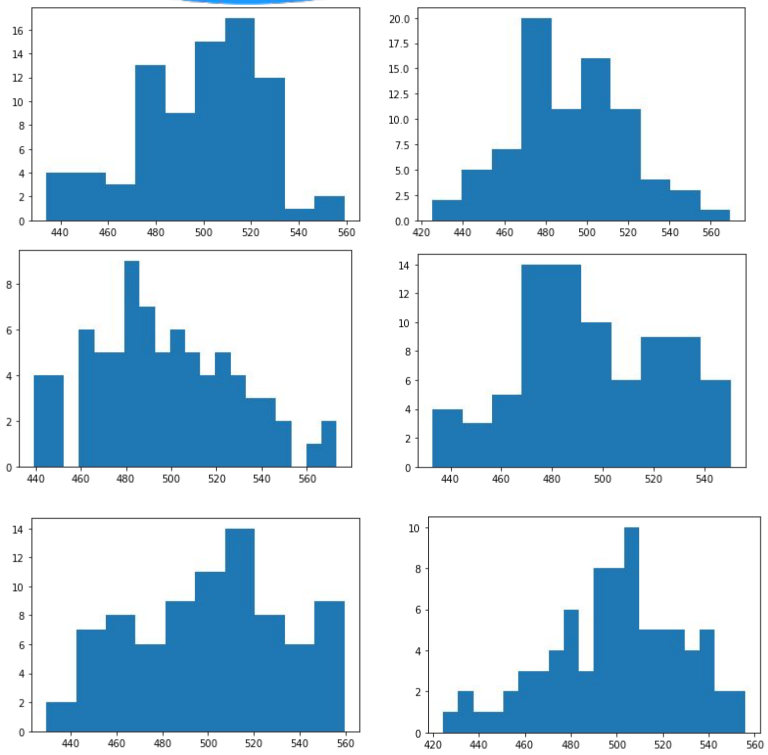
Statistics in Python

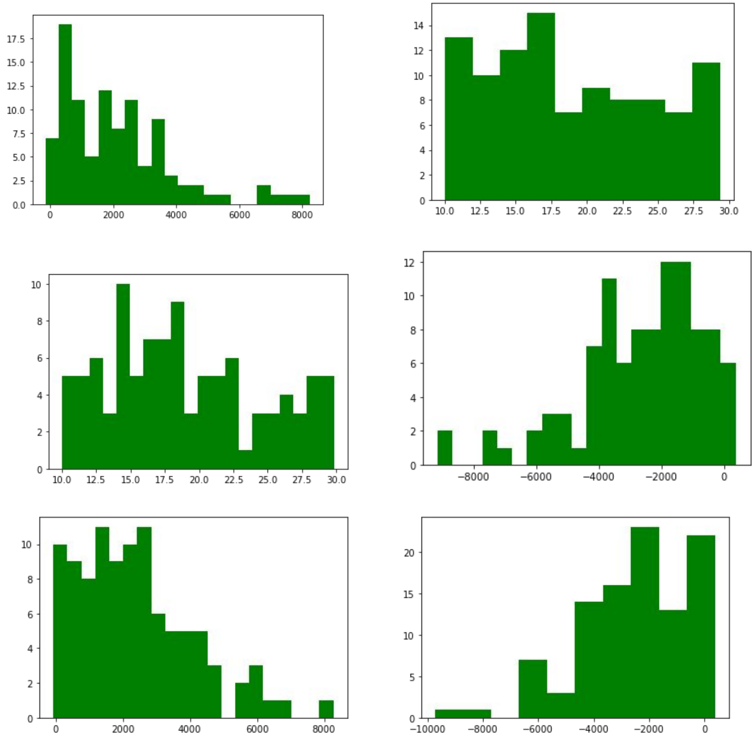
In this session we will...

1. Learn to calculate descriptive statistics in Python
2. Learn to conduct a hypothesis test in Python

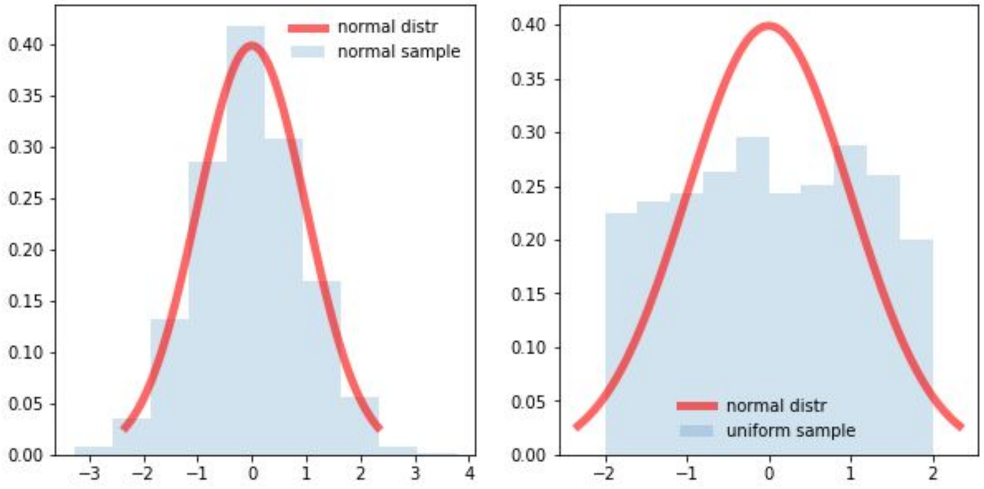
Normal Distribution

What's the difference?





The Normal Distribution



The Normal Distribution

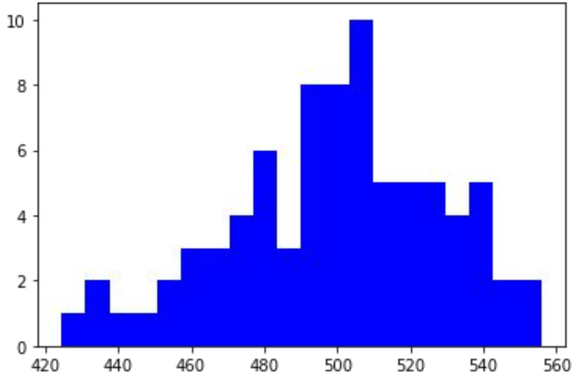
Examples of typically normally distributed variables: the height of adult females, IQ, the speed cars travel on the motorway

Examples of non-normally distributed variables: the income of people in the UK, the amount of time it takes your friend to reply to a text, the age people die

The Normal Distribution

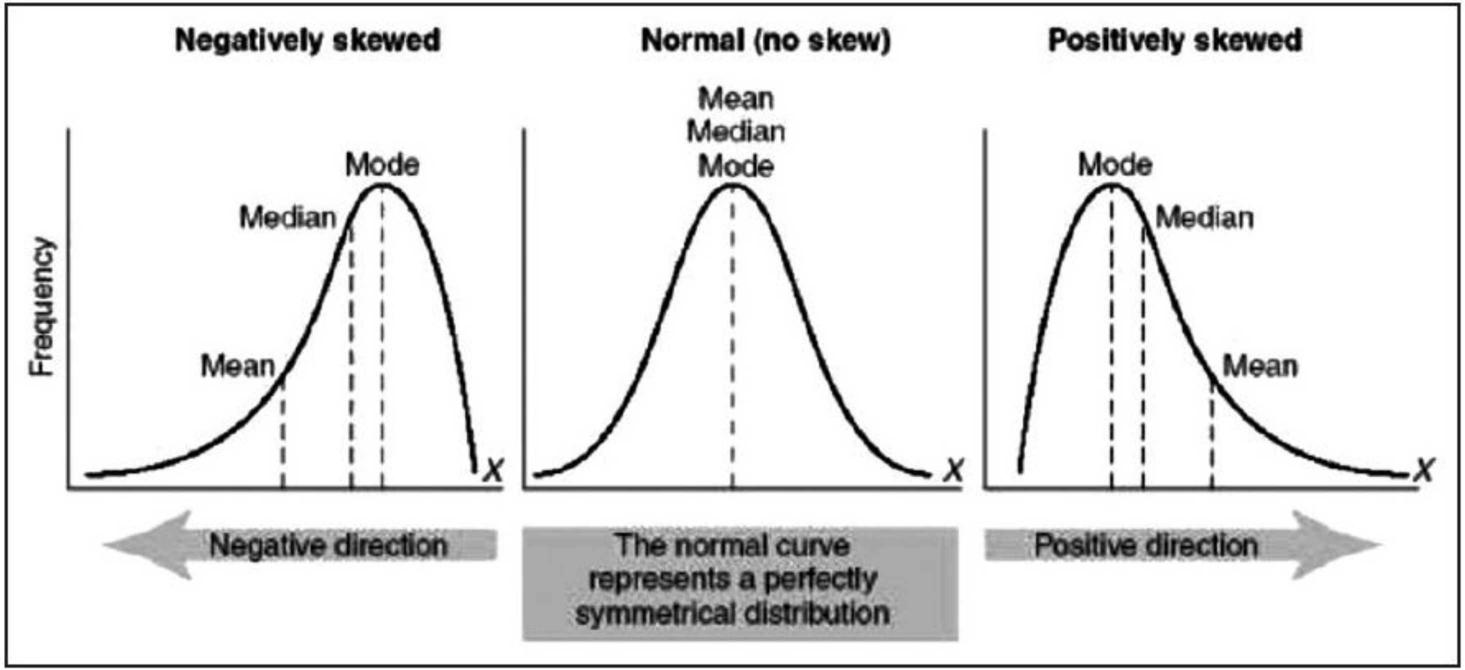
We can use Python to look at the distribution of a variable and assess whether we think it's normal or not.

```
from matplotlib import pyplot as plt

plt.hist(data, bins=20, facecolor='blue')
plt.show()
```

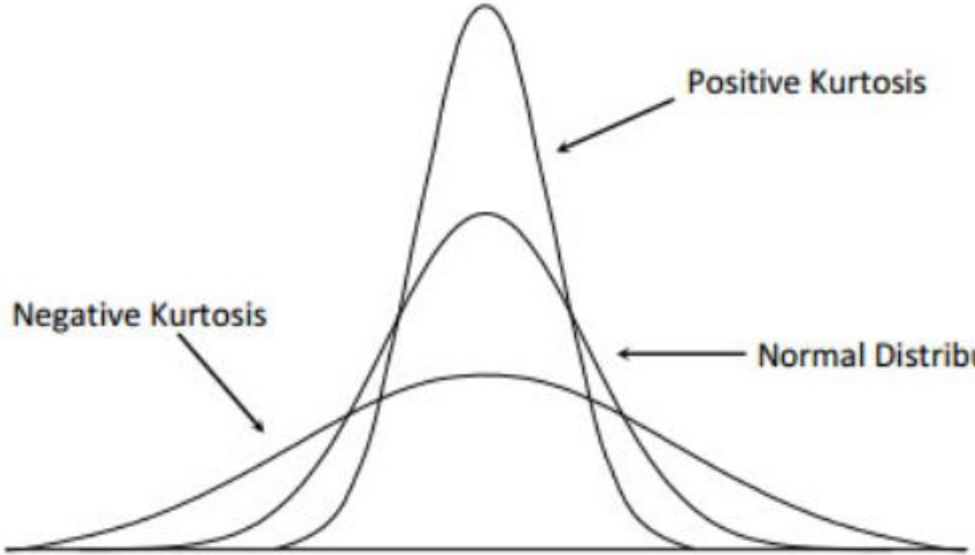


“Looks normal”

Skew


Negatively skewed
Normal (no skew)
Positively skewed
Frequency
Mode
Median
Mean
X
Mean
Median
Mode
X
Mode
Median
Mean
X
Negative direction
The normal curve represents a perfectly symmetrical distribution
Positive direction

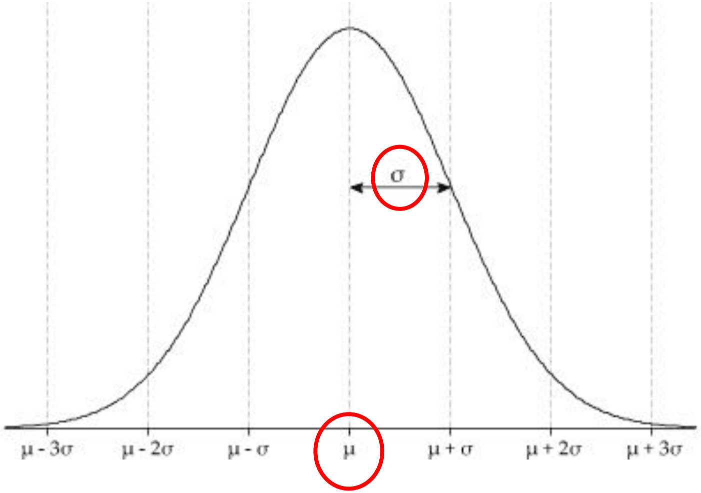
Kurtosis


Positive Kurtosis
Negative Kurtosis
Normal Distribution

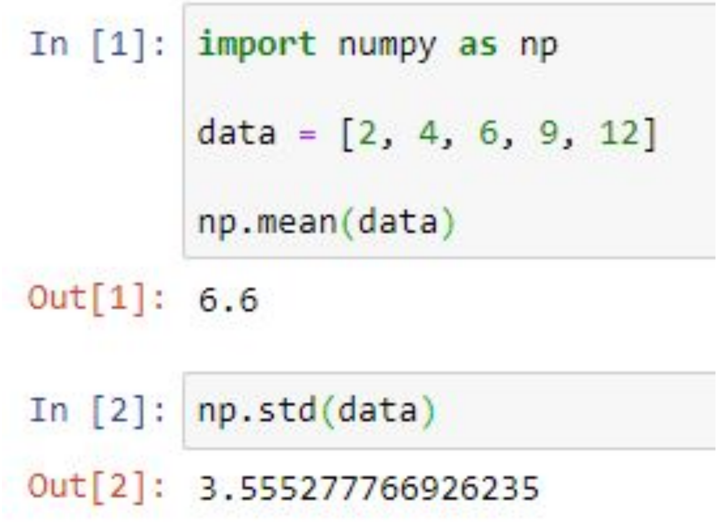
Describing Normal Variables

The key information that describes a normal variable is…

1. **mean:** add up all the values and divide by how many values there are
2. **standard deviation:** how far the values tend to be from the mean



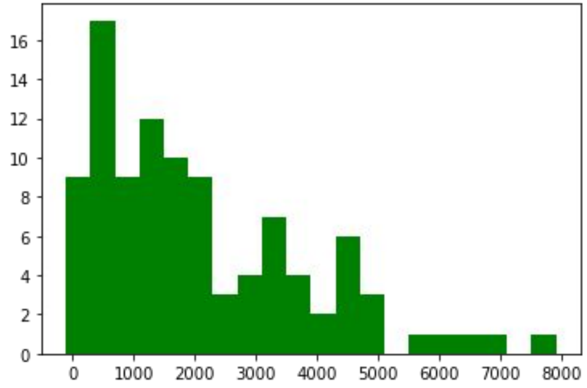
Mean & SD in python



```
In [1]: import numpy as np

        data = [2, 4, 6, 9, 12]

        np.mean(data)

Out[1]: 6.6

In [2]: np.std(data)

Out[2]: 3.55277766926235
```

If our data looks non-normal, particularly if it looks heavily skewed, we'll want the median and IQR to describe our data.

```
from matplotlib import pyplot as plt

plt.hist(data, bins=20, facecolor='blue')
plt.show()
```

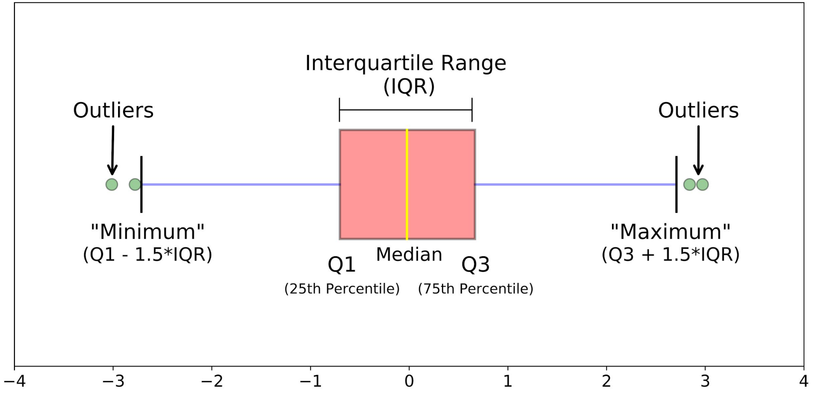


"Looks skewed - not normal"

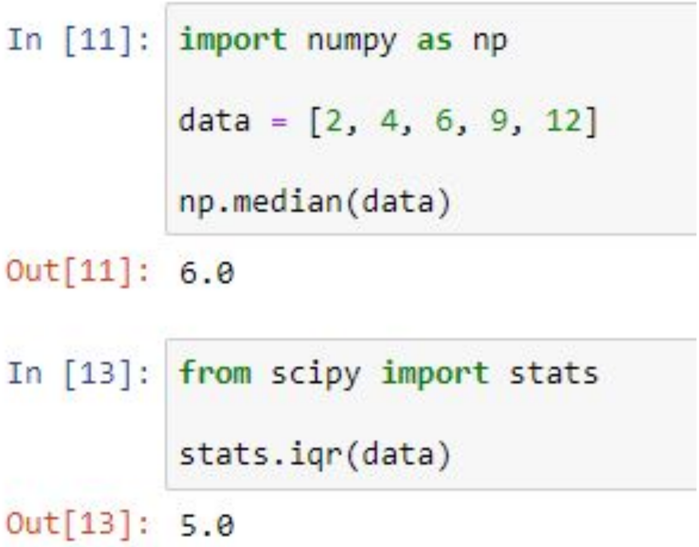
Non-Normal: Median and InterQuartile Range

Put all value in order, the median is in the middle

The span of the central 50% of the values is the IQR



Median and IQR in python



```
In [11]: import numpy as np

         data = [2, 4, 6, 9, 12]

         np.median(data)

Out[11]: 6.0

In [13]: from scipy import stats

         stats.iqr(data)

Out[13]: 5.0
```

Measures of Central Tendency

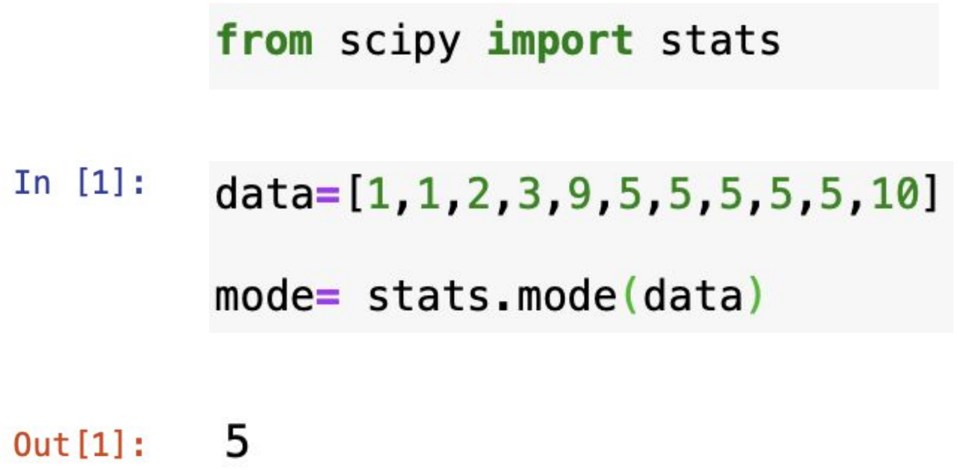
Mode

The value that appears most often in our data

It is the value that contributes to the largest part of the mean

Measures of Central Tendency

Mode in python



```
         from scipy import stats

In [1]:  data=[1,1,2,3,9,5,5,5,5,5,10]

         mode= stats.mode(data)

Out[1]:  5
```

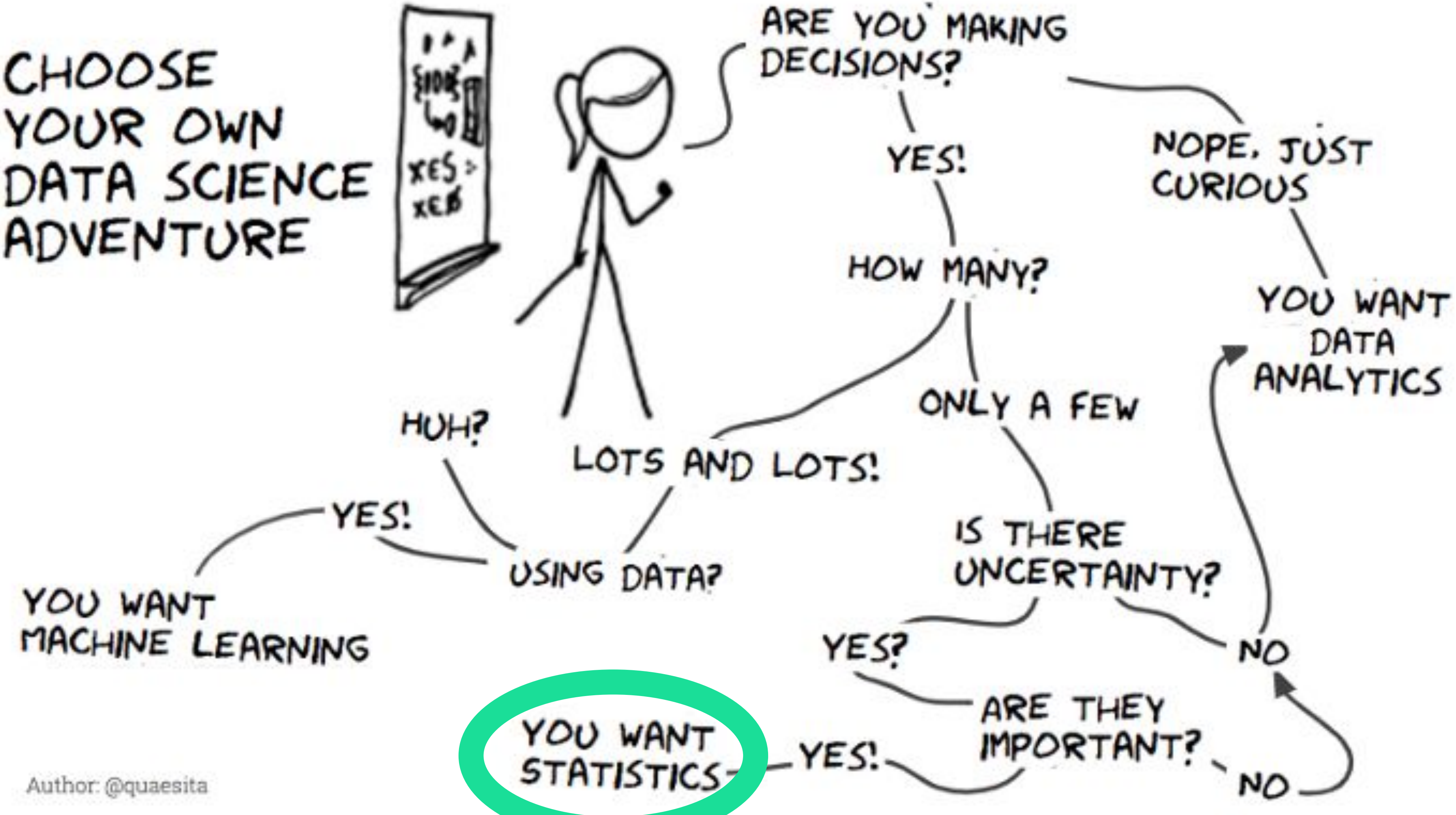
Practice

Measures of Central Tendency

Outliers

- Mean- Heavily affected by outliers
- Median- Barely affected by outliers
- Mode- Unaffected by outliers

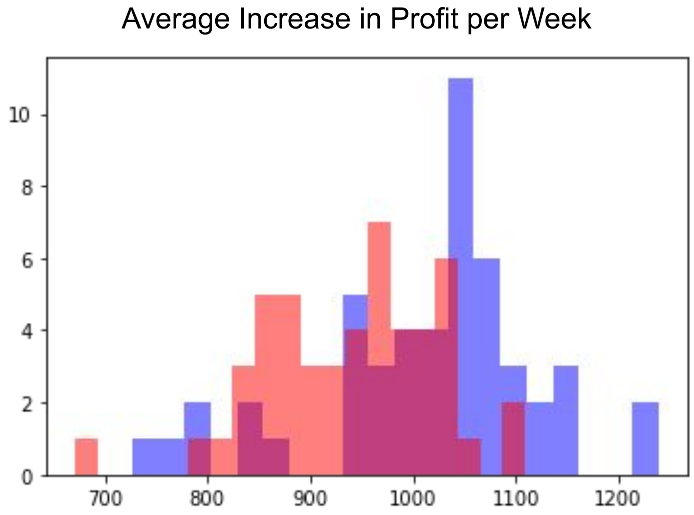
Hypothesis Testing



Author: @quaesita

Motivation

Let’s say you own a chain of convenience stores and you employ two area managers, Jeff and Jane. You are looking to award one ‘Manager of the Year’. On average, Jeff’s stores have increased their profits more than Jane’s over the past year, but was this just down to chance… or has something different being going on in his stores?



Hypothesis Testing steps:

1. Create a ‘null hypothesis’

Null hypothesis: *There’s no significant difference between the increased profit between Jeff’s stores and Jane’s stores*

2. Calculate a p-value (the probability that the difference you’ve observed was due to random chance)

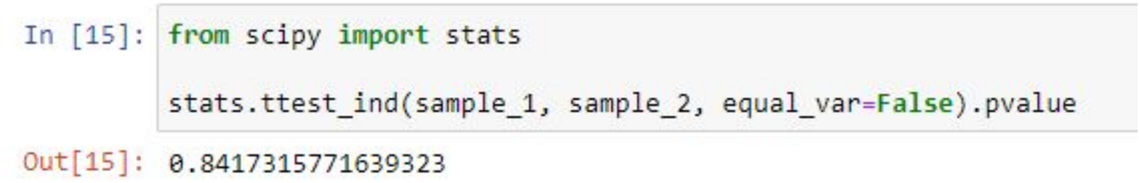
The process for this depends on the type of data you have, we’ll look at a couple of common ways today.

3. Reject your null hypothesis if your p-value is less than 5% = 0.05 (as it’s then pretty unlikely that the difference was due to random chance)

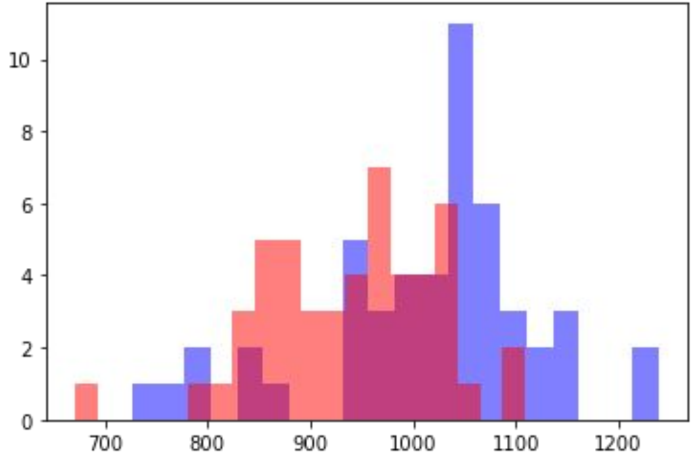
You might also want to mention the opposite of your null hypothesis, called an alternate or experimental hypothesis, in this case: *There is a significant difference between the increased profit between Jeff’s stores and Jane’s stores*

Welch’s t-test

If you have two normally distributed samples then you can calculate a p-value using Welch’s t-test in Python.



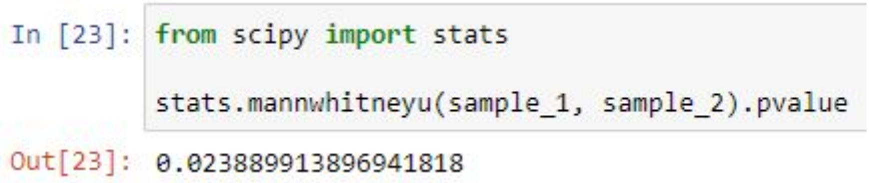
```
In [15]: from scipy import stats

         stats.ttest_ind(sample_1, sample_2, equal_var=False).pvalue

Out[15]: 0.8417315771639323
```



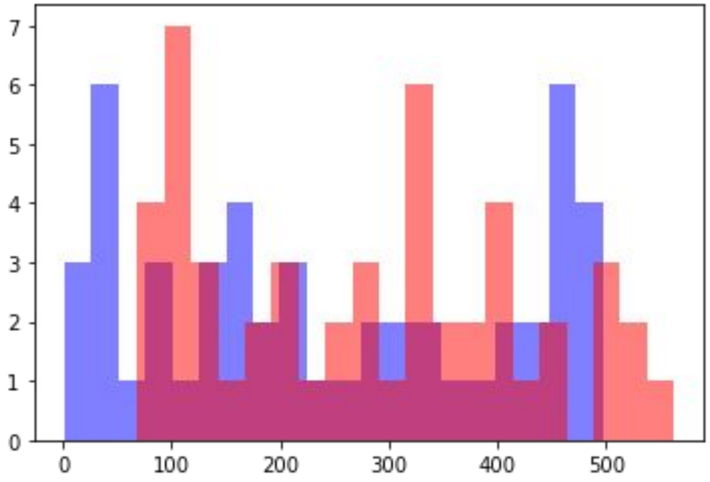
Here, the p-value was greater than 0.05 so we have *failed to reject our null hypothesis*

Mann Whitney U Test

If you're in a similar position, but both of your samples aren't normal, than you can calculate a p-value using the Mann Whitney U Test in Python.



```
In [23]: from scipy import stats

         stats.mannwhitneyu(sample_1, sample_2).pvalue

Out[23]: 0.023889913896941818
```



Here, the p-value was less than 0.05 so we have enough evidence to *reject our null hypothesis*

Practice

There are some other tests that we need if our data is a little different. Here's two common scenarios...

More than 2 samples = ANOVA Test

Both variables are categorical = Chi-squared Test (or Fisher's exact Test)

Errors

Remember Jeff and Jane?

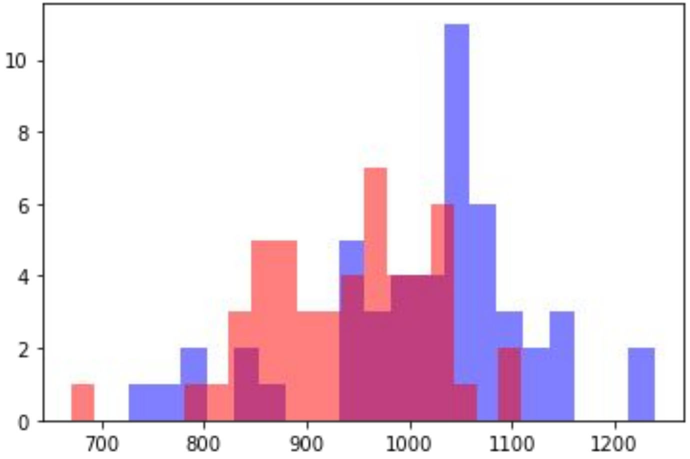
There were two errors we could make here...

1. We used sales data to make Jeff Manager of the Year, when in reality the difference in profit was due to random chance.

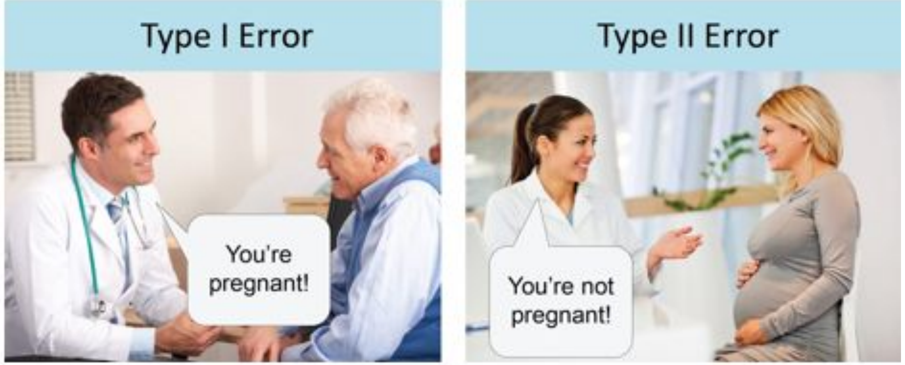
1. We didn't use this data to help us make our decision, despite it being informative.

These are called Type I and Type II errors.

Average Increase in Profit per Week

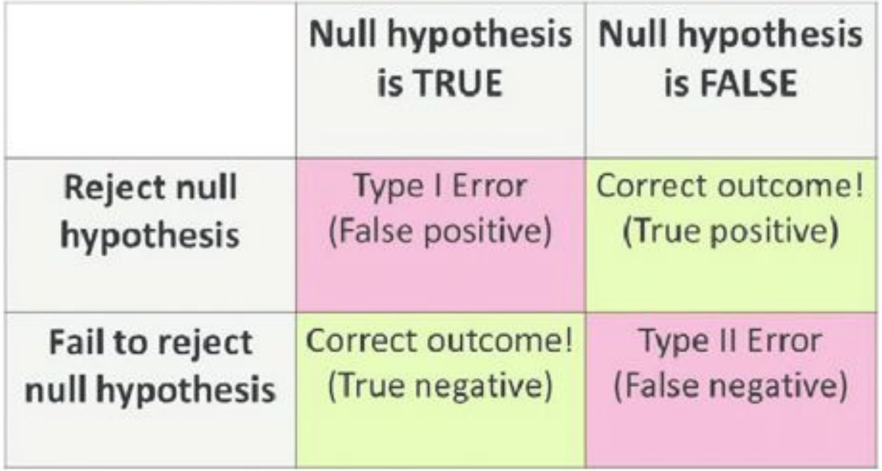


Errors



Type I Error

Type II Error



	Null hypothesis is TRUE	Null hypothesis is FALSE
Reject null hypothesis	Type I Error (False positive)	Correct outcome! (True positive)
Fail to reject null hypothesis	Correct outcome! (True negative)	Type II Error (False negative)

Errors

Which type of error is worst in the following scenarios...

1. A lab technician performing a HIV Test.
2. An adult deciding whether or not to propose marriage to their partner.
3. A jury deciding whether or not to convict a suspected murderer.

Missing Data

About missing data:

Sometimes we are unable to collect every attribute for a particular observation. Unfortunately, this makes the observation unusable until we decide how to deal with it.

We have to decide whether to:

- Drop the observation.
- Drop the attribute (i.e. remove the column)
- Impute a value for that specific attribute and observation.

So, how do we decide?

Types of missing data:

- **Missing completely at random (MCAR)**
    - The reason that the data are missing is completely random and introduces no sampling bias.
    - In this case, it's safe to drop or impute.
    - We can test for this by looking at other attributes for missing and non-missing groups to see if they match.

Types of missing data:

- **Missing at random (MAR)**
    - The data are missing in a way that is related to another factor.
    - This is a form of sampling bias.
    - Like other instances of sampling bias, we can fix this by modeling the selection process.
        - This is done by building a model to impute the missing value based on other variables.

Types of missing data:

- **Missing not at random (MNAR)**
    - The response is missing in a way that relates to its own value.
    - We can't test for this.
    - We also can't fix this in a reasonable way.

Sampling Bias

Sampling Bias:

Sampling bias occurs when a sample is collected in such a way that some members of the intended population are more or less likely to be included than others.

This can happen when a sample is taken non-randomly — either implicitly or explicitly.

When we have non-random sampling that results in sampling bias, it can affect the inferences or results of our analyses. We must be sure not to attribute our results to the process we observe when they could actually be because of non-random sampling.

Conceptually, this is straightforward: When we have sampling bias, we aren't measuring what we think we are measuring.

Examples of Sampling Bias:

- **Pre-screening:** Purposely restricting the sample to a specific group or region.
    - This typically happens when people try to study priority areas to save costs and assume priority areas are the same as random areas.
- **Self-selection:** When someone has the ability to non-randomly decide what is included in a sample.
    - This typically happens in surveys and polls but can also be an issue with other kinds of reporting.
- **Survivorship bias:** When we select only surviving subjects in a sample over time.
    - This might happen when we only look at existing customers and assume they have the same characteristics as new customers.

Recovering from Sampling Bias:

- Working out causal DAGs can help you identify when to watch out for sampling bias.
- Generally, it's best to prevent sampling bias whenever possible.
- We can't really do anything if we ENTIRELY exclude an important group of data.
- However, if portions of our data are overrepresented or underrepresented, there are ways to correct that effect.
    - Typically, we explicitly model the selection process, which means we need data on factors that determine whether or not someone participates.

Stratified Random Sampling:

An elegant way to reduce sample bias is to employ stratified sampling. If we know the proportions of a population we can ensure the sample is the same. For example, in a field of 300 cows, 100 are striped and the rest are spotted.

- If we wanted to take a random sample of 30 cows we would stratify by ensuring we took 10 cows randomly from the stripes, and 20 randomly from the spotted

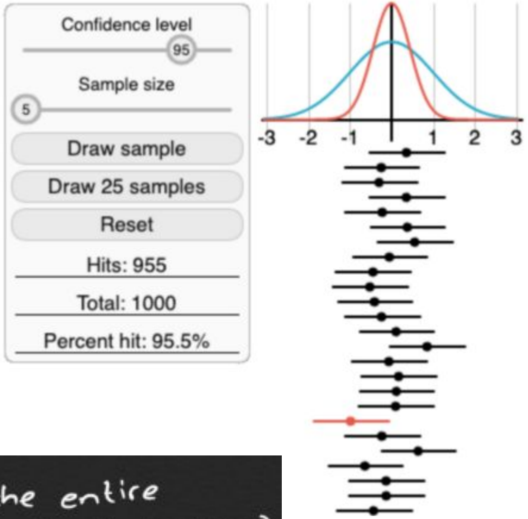
Confidence Intervals

Confidence Intervals

A closely related concept is a **confidence interval** for a statistic, such as the mean. A confidence interval is a range of values around the observed value of the statistic, (eg, the *sample* mean), which we expect to contain the true value. A 95% confidence interval can be interpreted like so:

- Under infinite sampling of the population, we would expect that the *true* value of the parameter we are estimating (eg, the *population* mean) would fall within that range 95% of the time.
- If the null hypothesis says that the statistic has a particular value (eg, that the population mean is 7) and our confidence interval does not include that value (eg, the sample mean is 9, and the confidence interval runs from 7.5 to 10.5), then we reject the null hypothesis, because the null hypothesis value of 7 lies outside our confidence interval.

Confidence Intervals



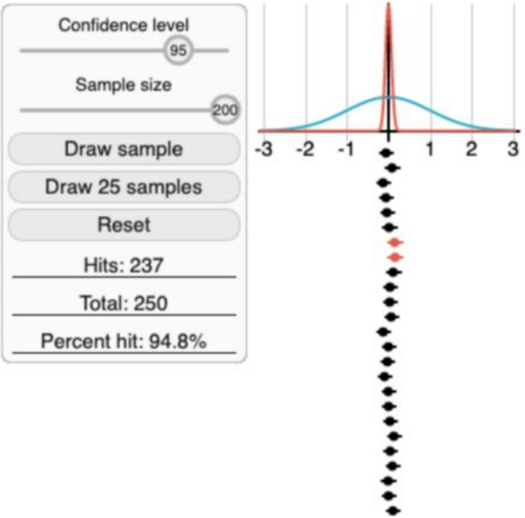
Confidence level: 95

Sample size: 5

Hits: 955

Total: 1000

Percent hit: 95.5%



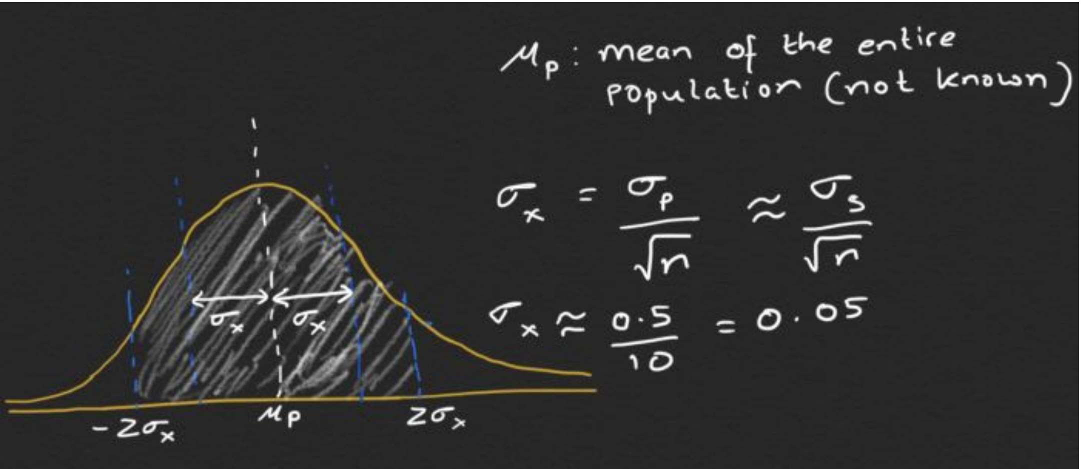
Confidence level: 95

Sample size: 200

Hits: 237

Total: 250

Percent hit: 94.8%



μ_p : mean of the entire population (not known)

$$\sigma_x = \frac{\sigma_p}{\sqrt{n}} \approx \frac{\sigma_s}{\sqrt{n}}$$

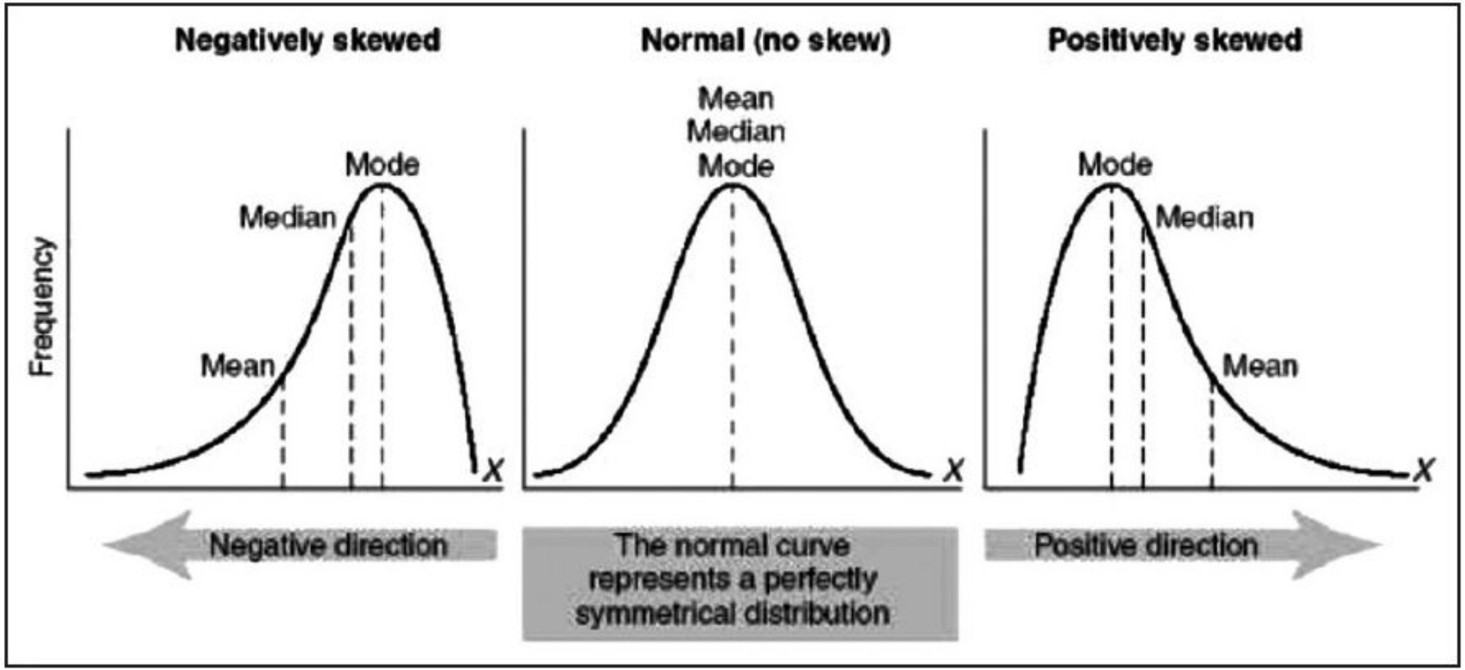
$$\sigma_x \approx \frac{0.5}{10} = 0.05$$

Skew

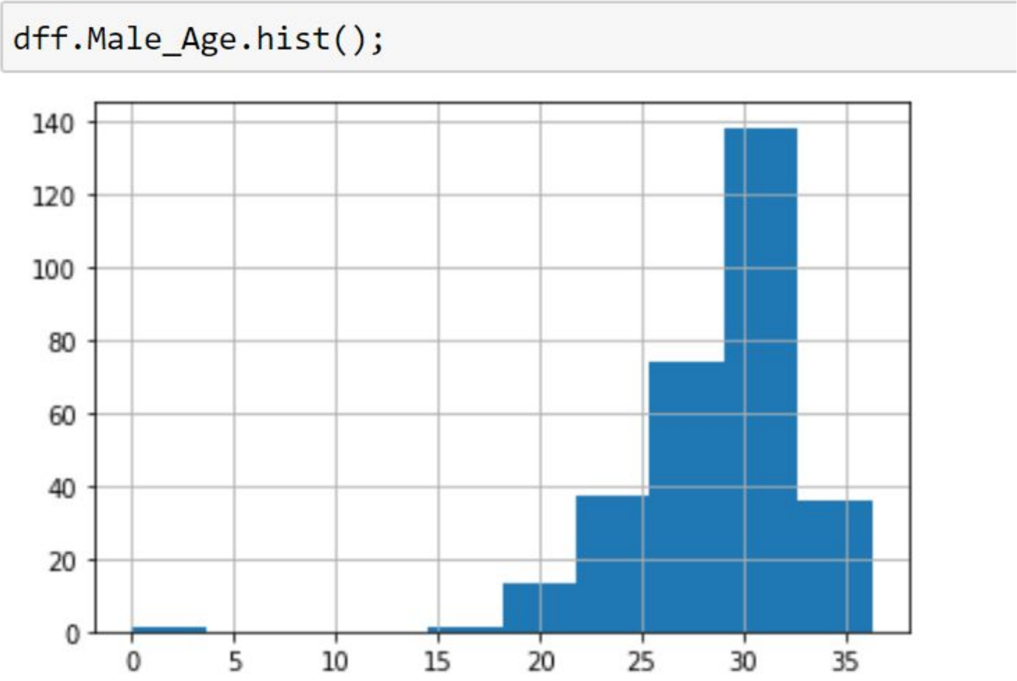
Skew

- Skewness is surprisingly important.
- Most algorithms implicitly use the mean by default when making approximations.
- If you know your data is heavily skewed, you may have to either transform your data or set your algorithms to work with the median.

Skew

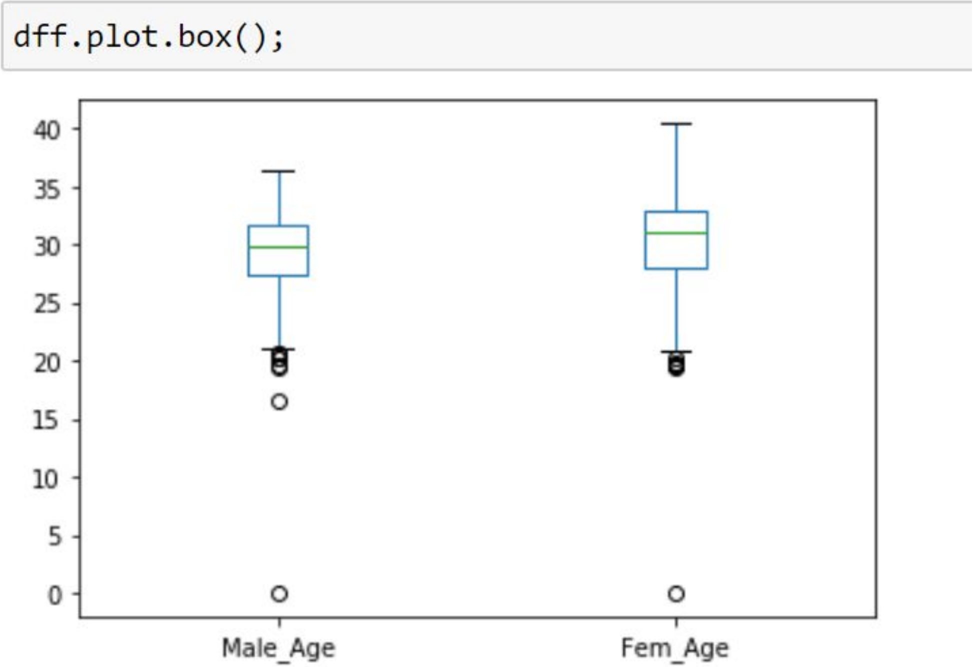

Negatively skewed
Normal (no skew)
Positively skewed
Frequency
Mode
Median
Mean
X
Mean
Median
Mode
X
Mode
Median
Mean
X
Negative direction
The normal curve represents a perfectly symmetrical distribution
Positive direction

Skew



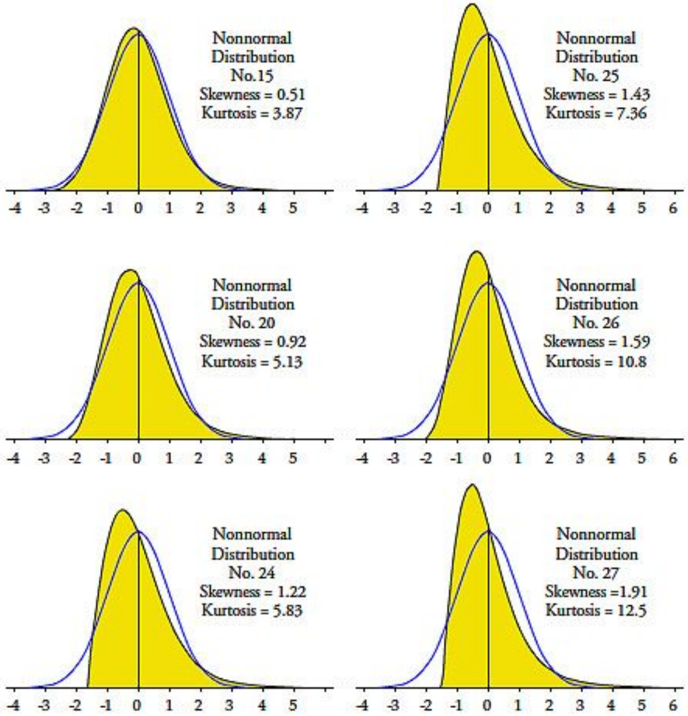
```
dff.Male_Age.hist();
```

Skew



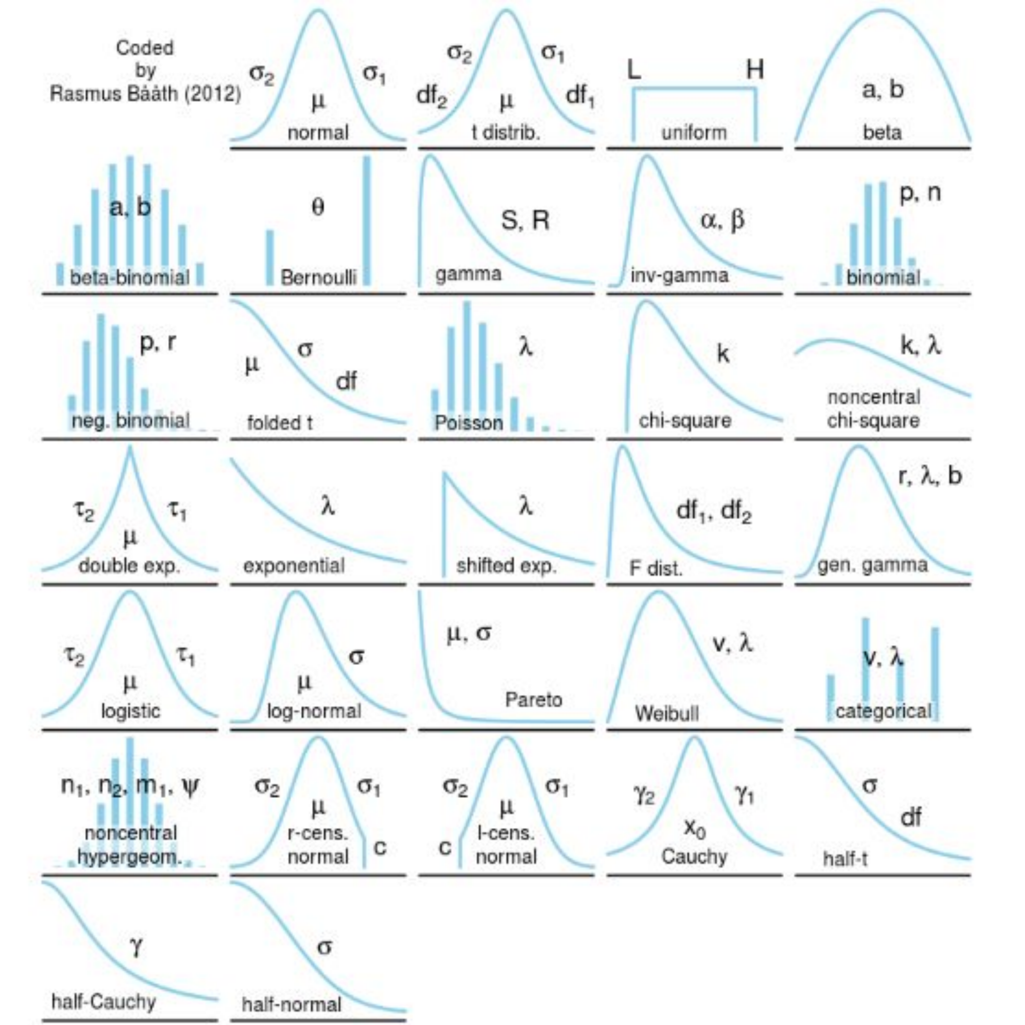
```
dff.plot.box();
```

Skew


Nonnormal Distribution No.15
Skewness = 0.51
Kurtosis = 3.87
-4 -3 -2 -1 0 1 2 3 4 5
Nonnormal Distribution No. 25
Skewness = 1.43
Kurtosis = 7.36
-4 -3 -2 -1 0 1 2 3 4 5 6
Nonnormal Distribution No. 20
Skewness = 0.92
Kurtosis = 5.13
-4 -3 -2 -1 0 1 2 3 4 5
Nonnormal Distribution No. 26
Skewness = 1.59
Kurtosis = 10.8
-4 -3 -2 -1 0 1 2 3 4 5 6
Nonnormal Distribution No. 24
Skewness = 1.22
Kurtosis = 5.83
-4 -3 -2 -1 0 1 2 3 4 5
Nonnormal Distribution No. 27
Skewness =1.91
Kurtosis = 12.5
-4 -3 -2 -1 0 1 2 3 4 5 6

Skew

1. Look at the Titanic data variables.
2. Are any of them normal?
3. Are any skewed?
4. How might this affect our modeling?



Which one?

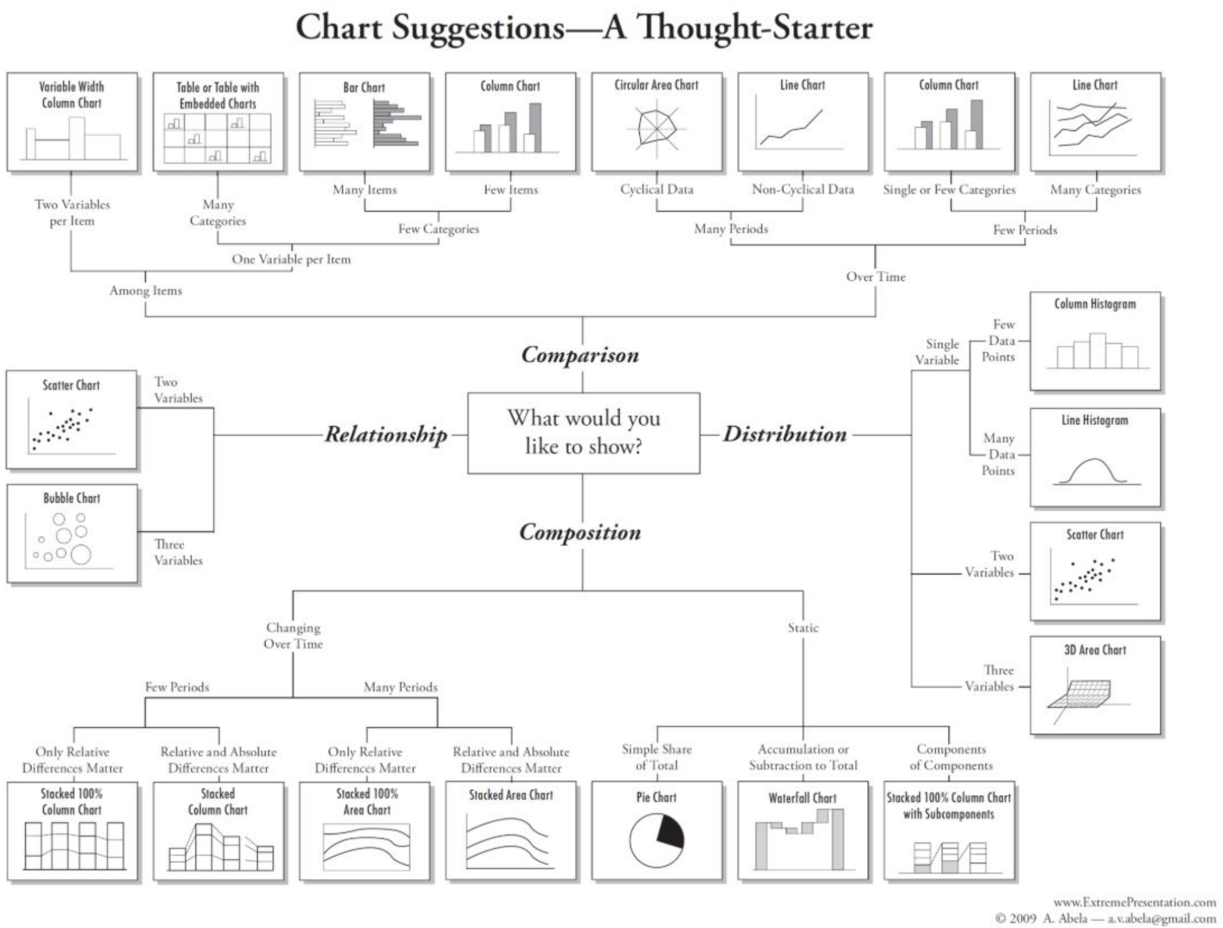
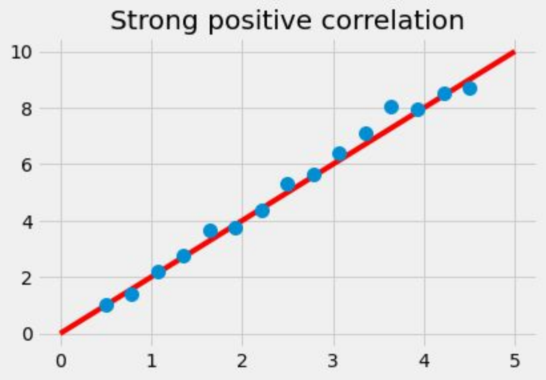


Chart Suggestions—A Thought-Starter

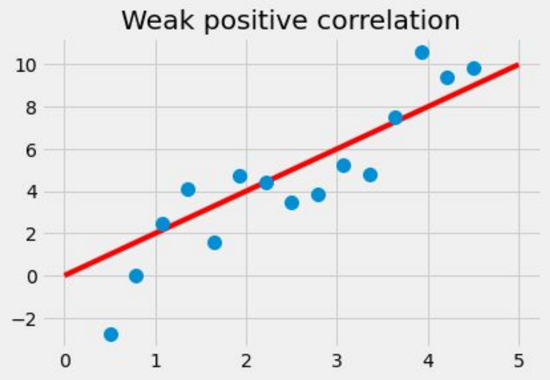
www.ExtremePresentation.com
© 2009 A. Abela — a.v.abela@gmail.com

Correlation and Causation

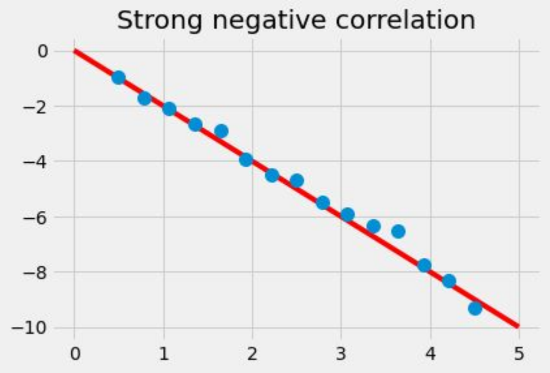
Correlation and Association



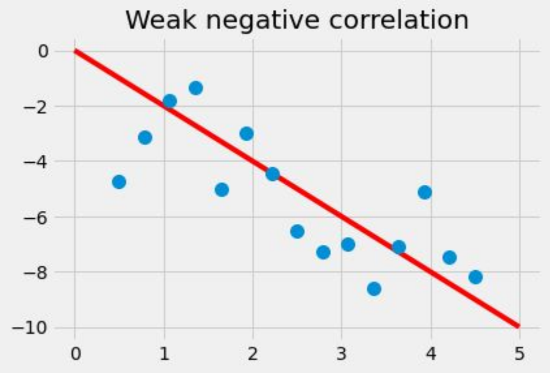
Strong positive correlation



Weak positive correlation

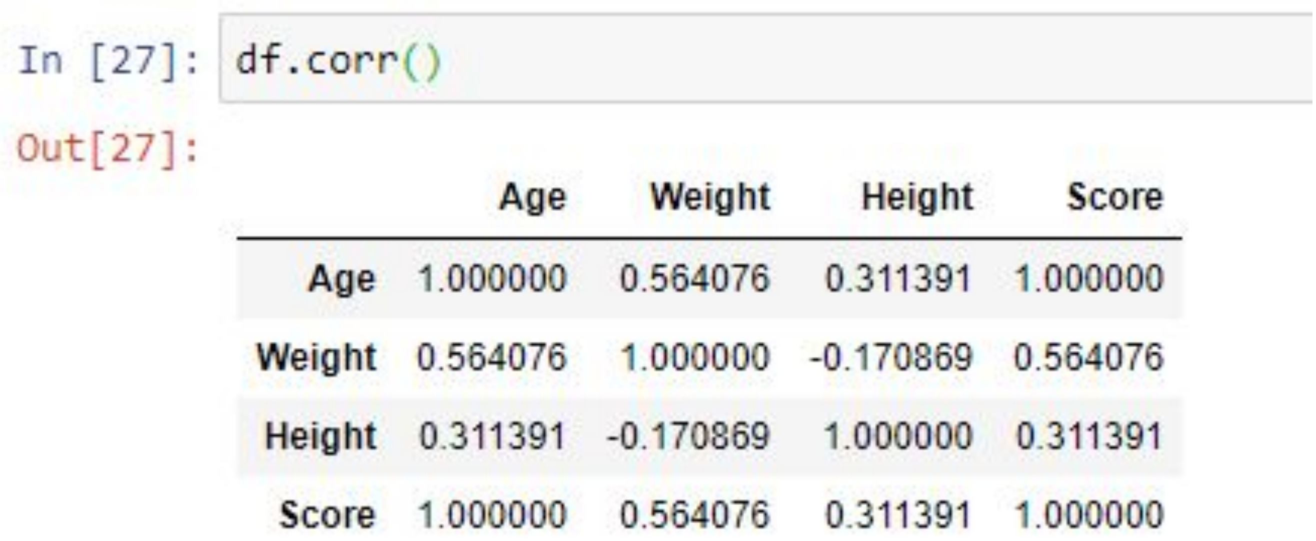


Strong negative correlation



Weak negative correlation

Correlation and Association

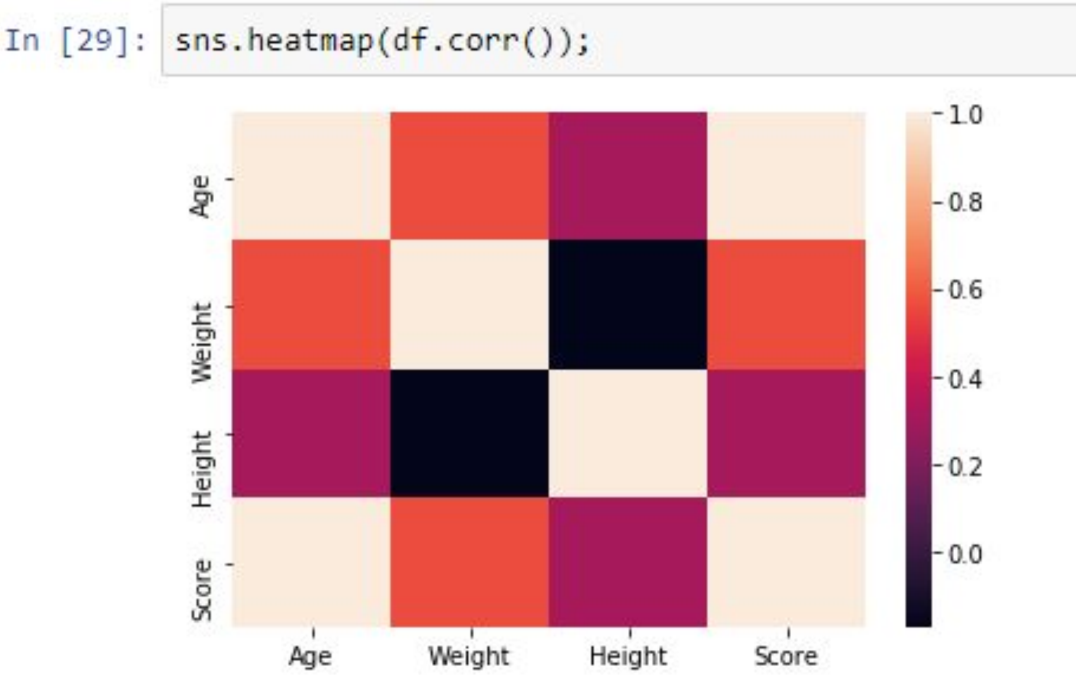


```
In [27]: df.corr()
```

Out[27]:

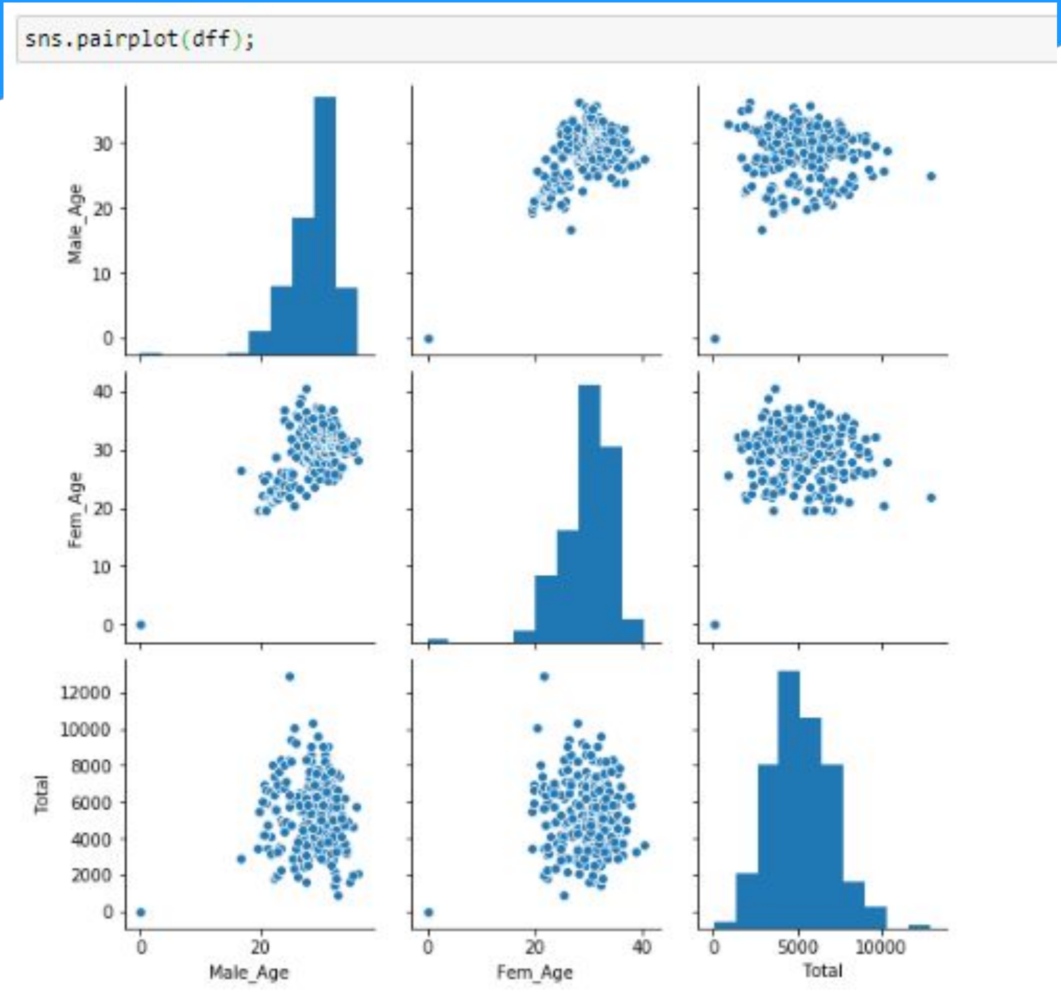
	Age	Weight	Height	Score
Age	1.000000	0.564076	0.311391	1.000000
Weight	0.564076	1.000000	-0.170869	0.564076
Height	0.311391	-0.170869	1.000000	0.311391
Score	1.000000	0.564076	0.311391	1.000000

Correlation and Association



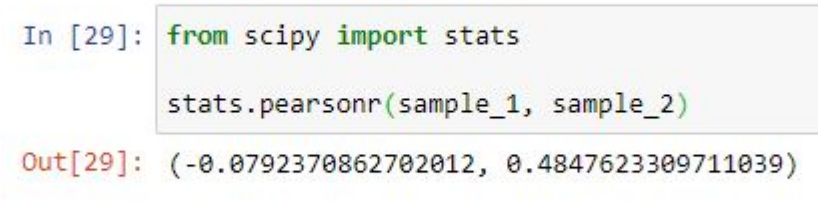
```
In [29]: sns.heatmap(df.corr());
```

Correlation and Association



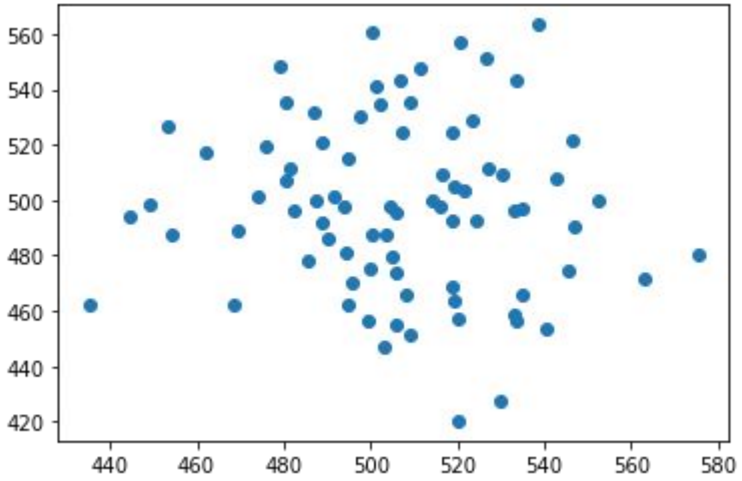
```
sns.pairplot(dff);
```

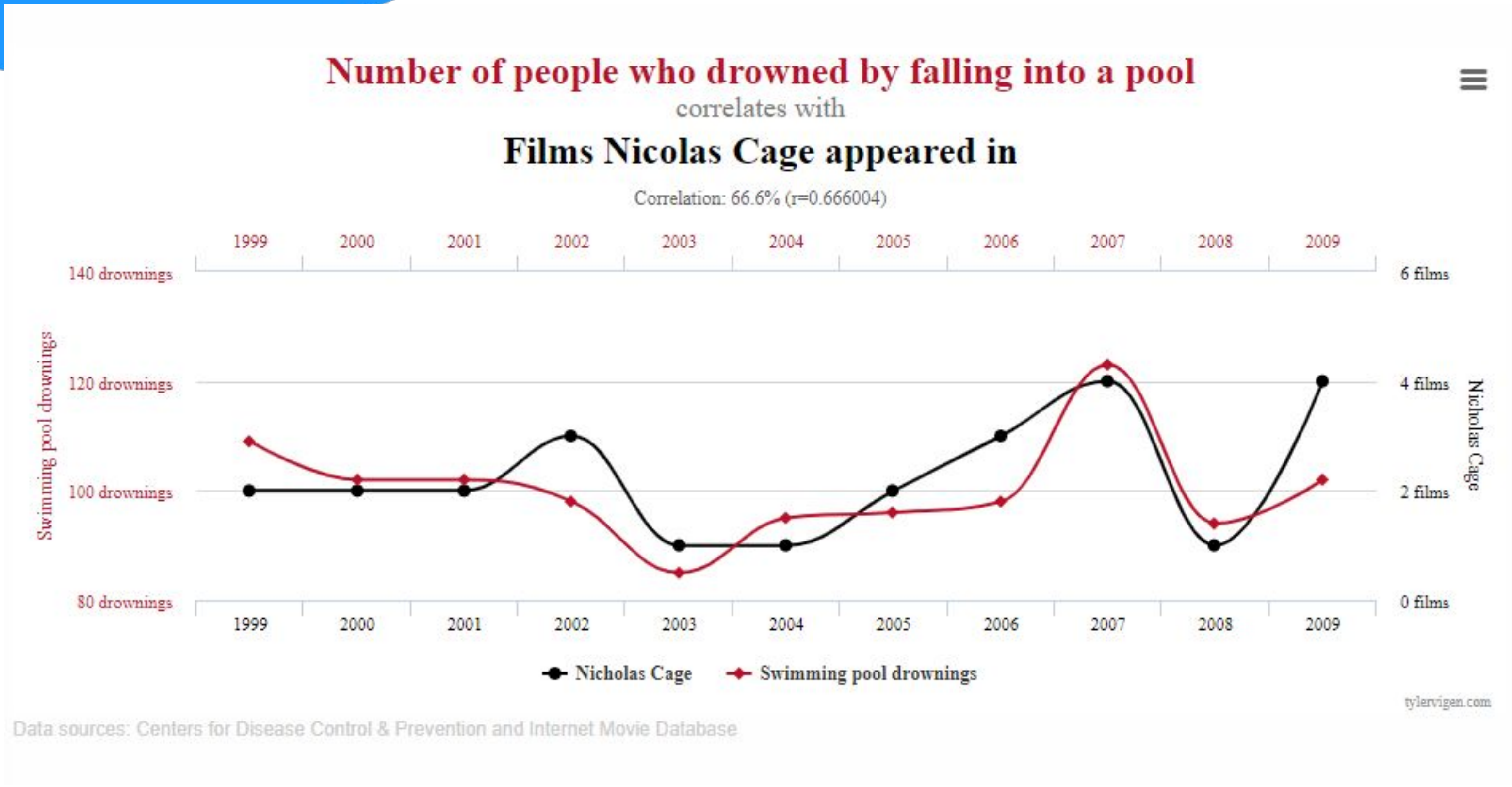
P-values for linear correlation



```
In [29]: from scipy import stats

         stats.pearsonr(sample_1, sample_2)

Out[29]: (-0.0792370862702012, 0.4847623309711039)
```

We can use Python to calculate how correlated two variables are. Doing this, we might have a null hypothesis that *they are not correlated at all*. The second number in the pair above is a p-value. As this is greater than 0.05, we are unable to reject this null hypothesis.





Number of people who drowned by falling into a pool

correlates with

Films Nicolas Cage appeared in

Correlation: 66.6% (r=0.666004)

Data sources: Centers for Disease Control & Prevention and Internet Movie Database

Structure of Causal claims

- If X happens, Y must happen.
- If Y happens, X must have happened.
    - (You need X and something else for Y to happen.)
- If X happens, Y will probably happen.
- If Y happens, X probably happened.

Why do we care?

- Understanding this difference is critical for executing the data science workflow, especially when identifying and acquiring data.
- We need to fully articulate our question and use the right data to answer it while also considering any confounders.
- We don’t want to overstate what our model measures.
- Be careful not to say “caused” when you really mean “measured” or “associated.”

Investigating Causal Relationships

Controlled experiments

- The most foolproof way to measure an effect is to control all of the confounders and directly intervene and control our variable of interest.
- This way we know that any correlation we find is not because of the confounders but instead because of the variable we control.
- This also means that all the effects we see are due to the variable we control.
- However, experiments are not always possible and take longer than using observational data.

When is it OK to rely on Association?

- **When any intervention that arises from your model affects only the outcome variable.**
    - In other words, you only need to predict Y.
    - This works because we only need to observe explanatory variables and implicitly know the confounders' effect.
    - Decision-making and intervention based on your model are hidden dangers that can shift confounders.
    - You can always retrain your model to work with a new set of confounders if they shift.

When is it OK to rely on Association?

- **When correlation is causal.**
    - If you are sure there are no confounding factors or selection bias, then that association might be a causation (risky).
    - It's OK to exclude confounders that have very unlikely or small effects.
    - This is a saving grace. To create a good model, you only need variables that correlate with your outcome.
        - Those variables merely need to meaningfully correlate with your outcome.

How does Association relate to Causation?

- Most commonly, we find an association between two variables if:
    - There is an observed correlation between the variables.
    - There is an observed correlation in a subset of data.
    - We find that the descriptive statistics significantly differ in two subsets of data.

How does Association relate to Causation?

- We may not still fully understand the causal direction (e.g., does smoking cause cancer or does cancer cause smoking?).
  - A causes B, B causes A, or a third factor causes both.
    - A and B never cause each other! [FALSE]
- We also might not understand other factors influencing the association.

Section Summary:

1) **It's important to have deep subject area knowledge.** You'll develop this over time and it will help you move through your analysis in a logical manner. However, keep in mind that you can show a strong association and still be wrong.

2) **A DAG (directed acyclic graph) can be a handy tool for thinking through the logic of your models.**

3) **There is a distinction between causation and correlation.** In our smoking example, it's relatively obvious that there's a flaw in our logic; however, this won't always be so readily apparent — especially in cutting-edge fields where there are many other unknown variables.

4) **Good data are essential.** Throughout this course we will be developing your data intuition so you can spot gaps and bias more readily. You'll also be introduced to tools that can help. However, your analysis is only as good as your understanding of the problem and the data.

Statistical tests

Each test makes various assumptions:

- ANOVA assumes the residuals are normally distributed and data have equal variances.
- The Welch t-test assumes normal distributions but not necessarily equal variances and more effectively accounts for small sample sizes.
- The Mann-Whitney test assumes nothing about the distributions but requires at least 20 data points in each set, producing a weaker p value.

Additional Resources

For more information on this topic, check out the following resources:

- An Introduction to Statistical Learning
- The more advanced book: Elements of Statistical Learning
- Spurious Correlations
- Wikipedia pages on ANOVA, Welch's t-test, Mann-Whitney test

Any Questions