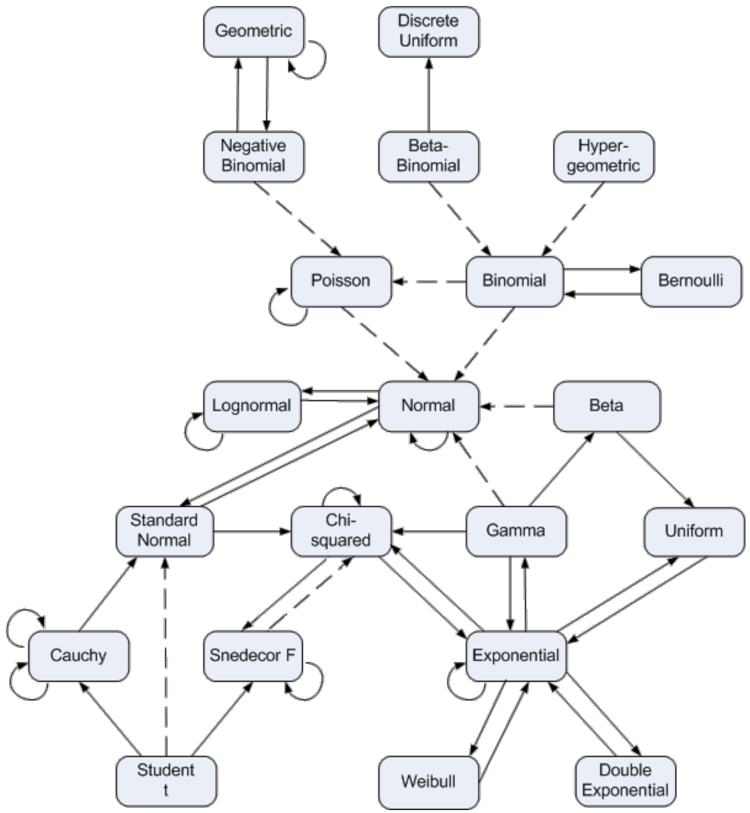
Probability and Statistic for ML (Part 2)

Module 1.2

carlos.chavez@immune.institute


Kahoot!

Main Distribution Review



Continuous Distributions: A probability distribution in which the random variable X can take on any value (is continuous).

Example: The height of students in a class represented by a normal distribution

Discrete Distributions: A discrete distribution means that X can assume one of a countable (usually finite) number of values.

Example: roll a fair dice.

Discrete Distributions

Bernoulli distribution: is a discrete distribution consisting of only one trial with 2 outcomes: success and failure. It’s considered to be the basis for defining more complex discrete distributions

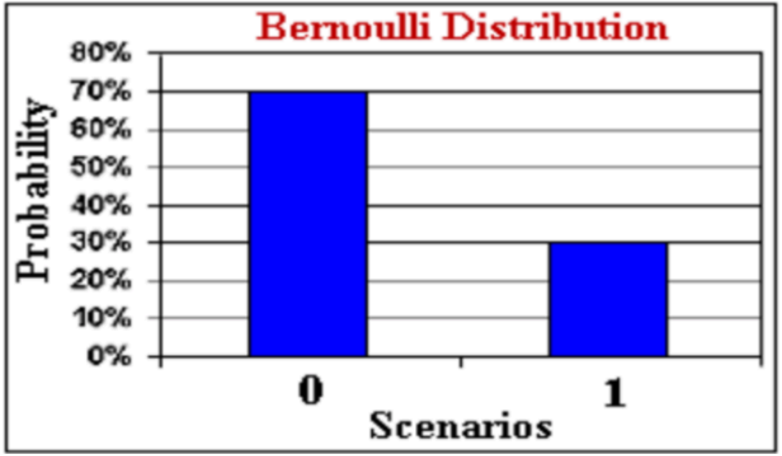
Examples:

- Probability of either pass or fail an exam
- A player either wins or loss a match

$$P(x) = \begin{cases} 1-p & \text{for } x=0 \\ p & \text{for } x=1 \end{cases}$$

$$Mean = p$$

$$Standard\ Deviation = \sqrt{p(1-p}$$



Discrete Distributions

Binomial distribution: computes the probability of *k* successes within *n* trials. It is Based on the Bernoulli distribution.

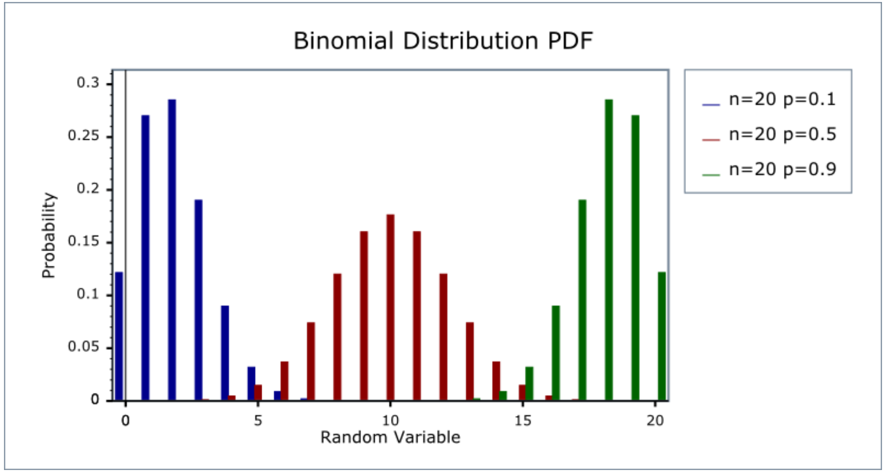
Example:

Toss a fair coin **three times** … what is the chance of getting **two Heads**?

$$P(x) = \frac{n!}{x!(n-x)!}p^x(1-p)^{n-x} \text{ for } n > 0$$

$$Mean = np$$

$$Standard\ Deviation = \sqrt{np(1-p)}$$



Continuos Distributions

Exponential distribution: describe the amount of time between events occurring at random moments. It is considered that time has no effect on future outcomes (memoryless).

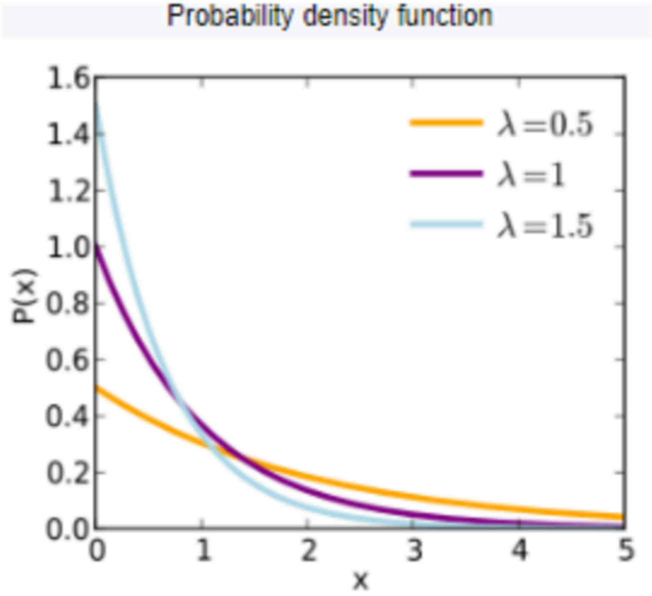
Examples:

- How long it will take until someone receives the next phone call?
- How long will a product function before breaking down?

$$f(x) = \lambda e^{\lambda x} \text{ for } \lambda > 0$$

$$Mean = \frac{1}{\lambda}$$

$$Standard\ Deviation = \frac{1}{\lambda}$$



Continuos Distributions

Normal distribution: is used to describe natural phenomena, such as the distribution of shoes sizes or population age.

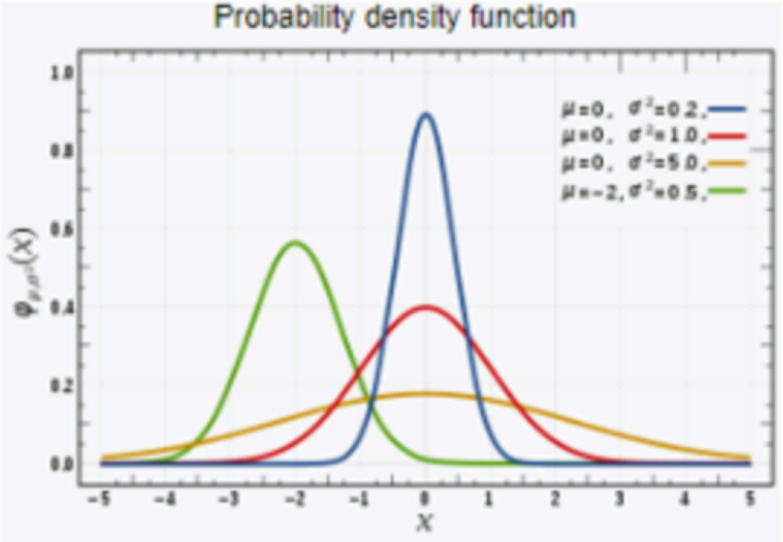
Notes:

- Standard Normal: $\mu = 0 \quad \sigma = 1$
- Every normal distribution can be transformed to a standard normal.

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

$$Mean = \mu$$

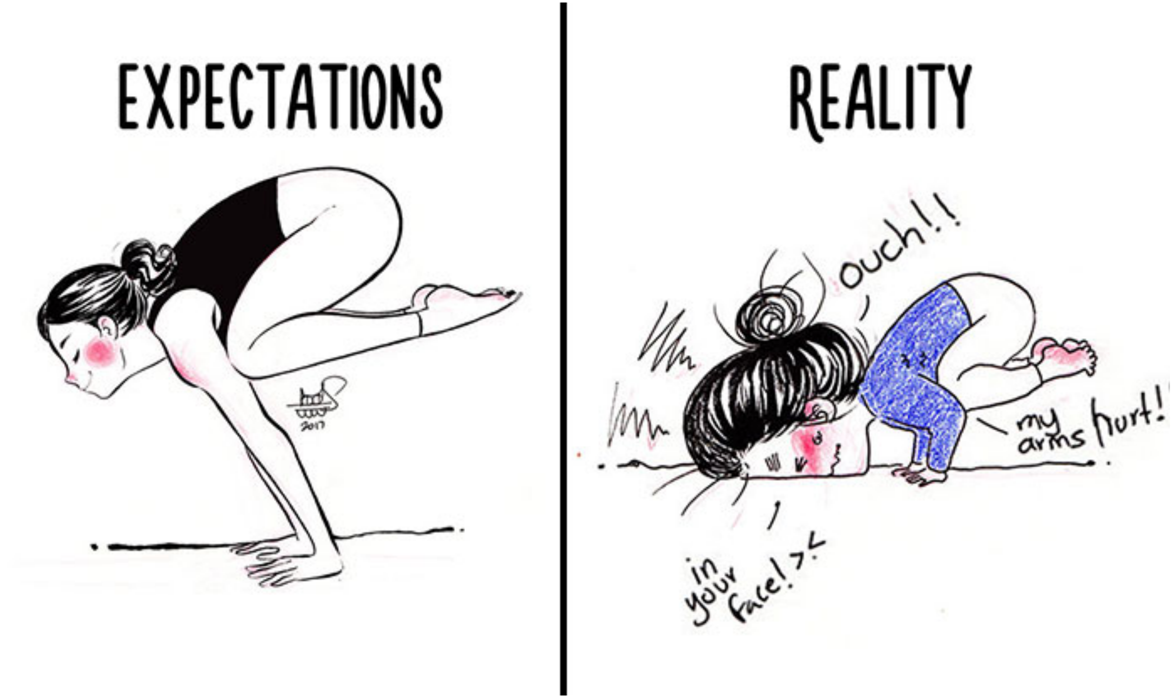
$$Standard\ Deviation = \sigma$$



The red curve is the *standard normal distribution*

Probability vs Statistic

Probability vs Statistic



Probability vs Statistic

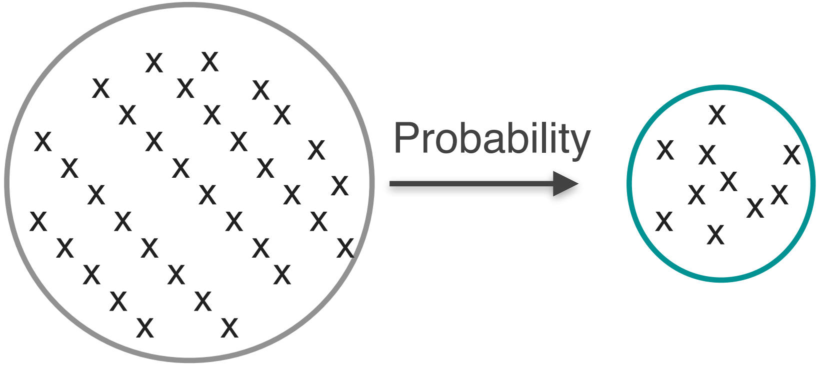
Population Distributions	Samples
probabilities	estimations
parameters	statistics

IMMUNE CODING INSTITUTE

Probability vs Statistic

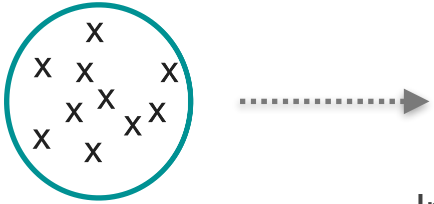
Probability

Population



Statistic

Sample



Descriptive Statistics

Estimate the Distribution

Inferential Statistics

IMMUNE
CODING INSTITUTE

Distribution Parameters

All the distribution we have studied are determined by some parameters.

Normal

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

Binomial

$$P(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \text{ for } n > 0$$

In general, any deterministic function of the distribution is a parameter or property

mean

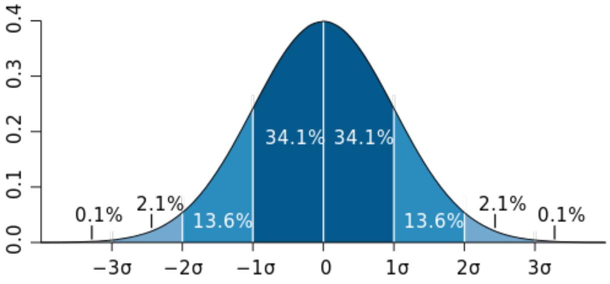
mode

min, max

variance

standard deviation

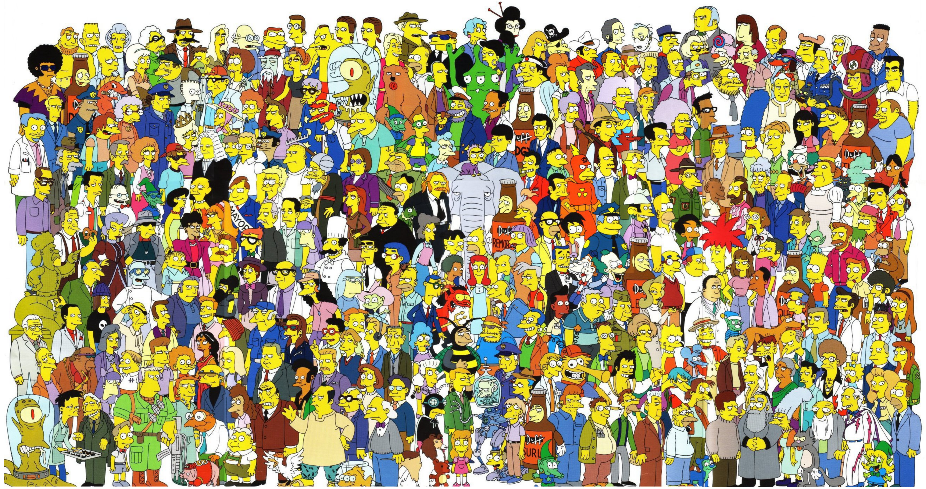
Difference between variance and standard deviation



VARIANCE
VERSUS
STANDARD DEVIATION

Variance	Standard Deviation
It is the statistical measure of how far the numbers are spread in a data set from their average.	It is the measure of dispersion of values in a given data set relative to their mean.
It helps determine the size of the data spread.	It measures the absolute variability of the dispersion.
It is calculated by taking the average of the squared deviation of the each value in the data set from the mean.	It is calculated by taking the square root of the variance.
$S^2 = \sum (x - M)^2 / n$, where S^2 = variance, x = a value in the data set, M = mean, and n = number of values.	$S = \sqrt{\sum (x - M)^2 / n}$, where S = standard deviation.
Variance is one of the key aspects of asset allocation in investing portfolios.	Standard deviation can be used as a measure of market and security volatility in finance.

Population



Sample n objects much smaller than population size

Deduce some population parameter from sample

We can suppose that this collection of ages correspond to a distribution

Samples

The importance of the sample

- Abraham Wald

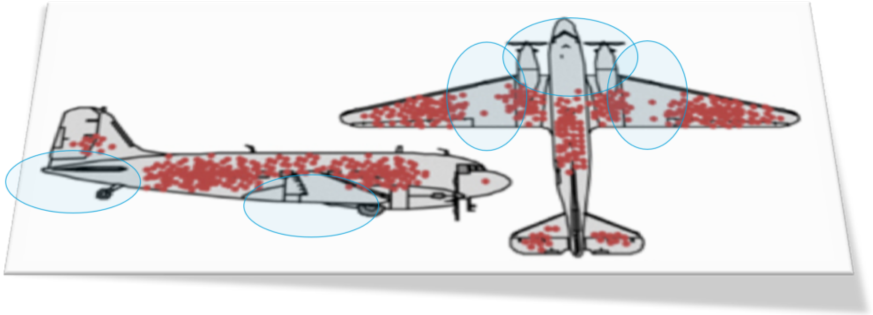
Wald was the one who convinced the Navy that they were about to armor the completely wrong parts of their planes, saving hundreds of flight crews in the process.

Samples

The importance of the sample

- Abraham Wald

Wald was the one who convinced the Navy that they were about to armor the completely wrong parts of their planes, saving hundreds of flight crews in the process.



“Don’t arm(our) the places that sustained the most damage on planes that came back – simply because, they came back and these areas can sustain damage!”

Statistic

Definition: Statistical learning refers to a set of tools for modeling and understanding a population (or datasets)

Statistics: is any quantity computed from values in a sample, often the average or the variation.

Use statistic to infer properties of the distribution or population:

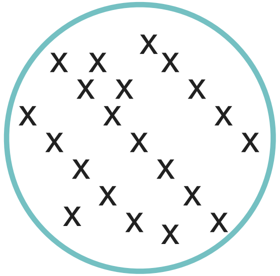
$\boxed{\text{type of distribution}}$ $\boxed{\text{parameter}}$

Some statistics are

average max or min observed variation

IMMUNE
CODING INSTITUTE

Parameter Estimation

sample  $sample = \{X_1, X_2, \cdots, X_{n-1}, X_n\}$

$\min_i \{X_i\}$

$Mode\{X_i\}$

$Variance = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

$\max_i \{X_i\}$

$Median\{X_i\}$

$mean\{X\} = \frac{1}{n}\sum_{i=1}^{n} X_i$

sample = [1.1, 2.0, 8.5, 3.0, 4.75]

Mean

Median

Mode

sample = [1.1, 2.0, 8.5, 3.0, 4.75]

Mean	Median	Mode
4,83		

IMMUNE
CODING INSTITUTE

sample = [1.1, 2.0, 8.5, 3.0, 4.75]

Mean **Median** **Mode**

~~4,83~~ 5

sample = [1.1, 2.0, 8.5, 3.0, 4.75]

Mean	**Median**	**Mode**
4,83	[1.1, 2.0, 8.5, 3.0, 4.75]	

sample = [1.1, 2.0, 8.5, 3.0, 4.75]

Mean	Median	Mode
4,83	[1.1, 2.0, 8.5, 3.0, 4.75] [1.1, 2.0, 3.0, 4.75, 8.5,]	

IMMUNE
CODING INSTITUTE

sample = [1.1, 2.0, 8.5, 3.0, 4.75]

Mean

4,83

Median

[1.1, 2.0, 8.5, 3.0, 4.75]

[1.1, 2.0, 3.0, 4.75, 8.5,]

Mode

IMMUNE
CODING INSTITUTE

sample = [1.1, 2.0, 8.5, 3.0, 4.75]

Mean

4,83

Median

[1.1, 2.0, 8.5, 3.0, 4.75]

[1.1, 2.0, 3.0, 4.75, 8.5,]

Mode

[1, 1, 2, 3, 4, 5, 7,
7, 7,7, 4, 2, 8, 8, 8,
9, 200]

IMMUNE
CODING INSTITUTE

sample = [1.1, 2.0, 8.5, 3.0, 4.75]

Mean	Median	Mode
4,83	[1.1, 2.0, 8.5, 3.0, 4.75] [1.1, 2.0, $\boxed{3.0}$, 4.75, 8.5,]	[1, 1, 2, 3, 4, 5, 7, 7, 7,7, 4, 2, 8, 8, 8, 9, 200] 7

IMMUNE
CODING INSTITUTE

sample = [1.1, 2.0, 8.5, 3.0, 4.75]

Mean

4,83

Median

[1.1, 2.0, 8.5, 3.0, 4.75]

[1.1, 2.0, 3.0, 4.75, 8.5,]

Mode

[1, 1, 2, 3, 4, 5, 7, 7, 7,7, 4, 2, 8, 8, 8, 9, 200]

7

sample = [1.1, 200.0, 8.5, 3.0, 4.75] ?

IMMUNE
CODING INSTITUTE

Data Validation and Clean up

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S

```
In [192]: df.isna().sum()

Out[192]: PassengerId      0
          Survived         0
          Pclass           0
          Name             0
          Sex              0
          Age            177
          SibSp            0
          Parch            0
          Ticket           0
          Fare             0
          Cabin          687
          Embarked         2
          dtype: int64
```

19,86% Numeric

Categorical
77,10%

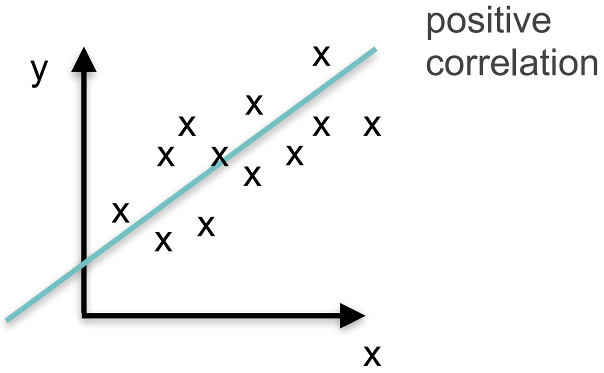
0,22%
Categórico

Correlation Analysis

Correlation coefficient r(X,Y) is a statistic that measures the degree that Y is a function of X and vice versa. Correlation values range from -1 to 1, where 1 means fully correlated, -1 means negatively-correlated, and 0 means no correlation.

Some examples of data that have a **high correlation:**

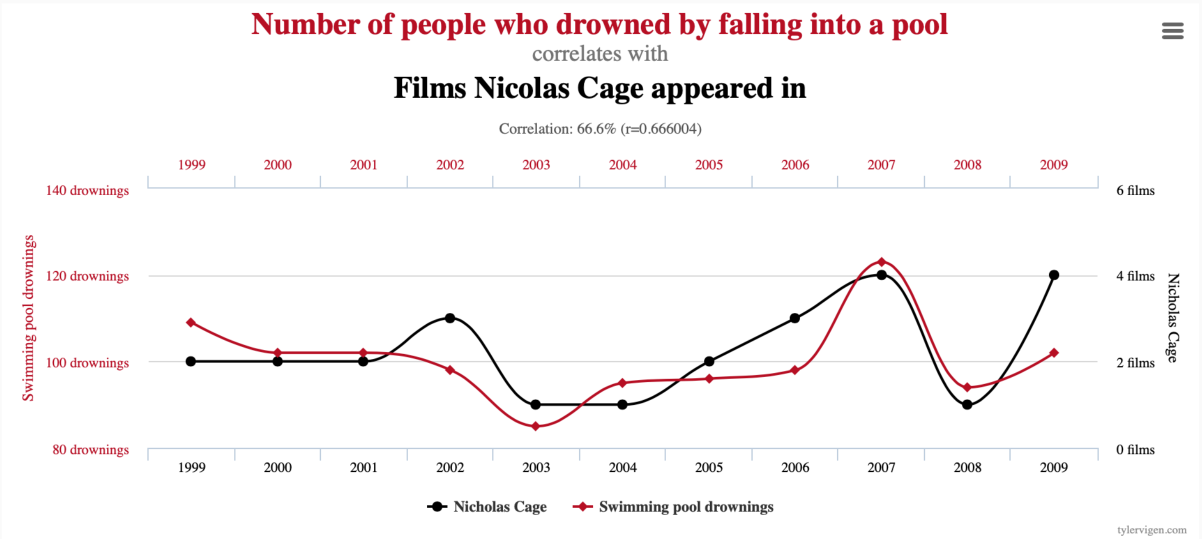
- Your caloric intake and your weight.
- Your eye color and your relatives' eye colors.
- The amount of time your study and your GPA.



IMMUNE
CODING INSTITUTE

Correlation Analysis

Correlation coefficient r(X,Y) is a statistic that measures the degree that Y is a function of X and vice versa. Correlation values range from -1 to 1, where 1 means fully correlated, -1 means negatively-correlated, and 0 means no correlation.



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

Note: Correlation does not imply causation!

Other [examples]

IMMUNE
CODING INSTITUTE

p-value

Definition: a *p-value* helps you determine the significance of your results.

Hypothesis test are used to test the validity of a claim that is made about the population

- A small *p-value* (typically $\leq$ 0.05) indicates strong evidence agains the null hypothesis, so you reject the null hypothesis.

- A large *p-value* (typically > 0.05) indicates weak evidence agains the null hypothesis, so you fail to reject the null hypothesis.

p-value

Example:

p-value

Example:

Randomly sample some delivery times and run the data through the hypothesis test

$$H0 : \bar{X} \neq \mu = 45'$$

p-value

Example:

Randomly sample some delivery times and run the data through the hypothesis test

$$H0 : \bar{X} \neq \mu = 45'$$

- A small *p-value* (typically $\leq$ 0.05) indicates strong evidence agains the null hypothesis, so you reject the null hypothesis.

IMMUNE

CODING INSTITUTE