

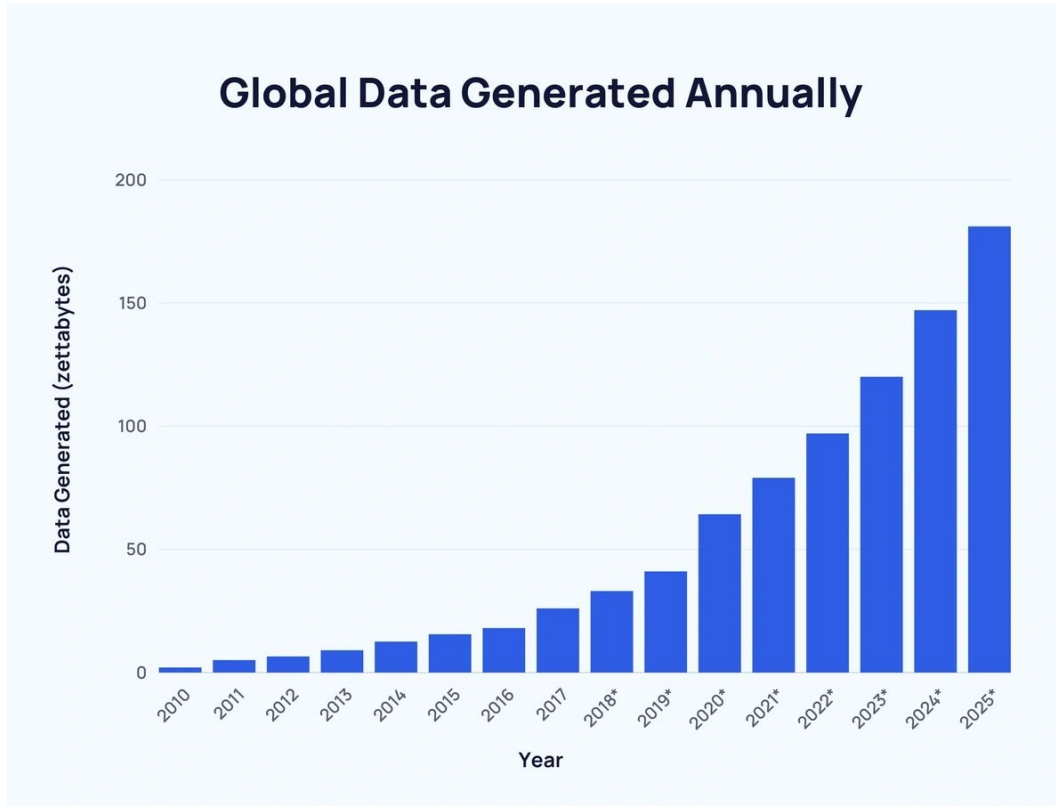
Applied Data science

Unit-I

Data units

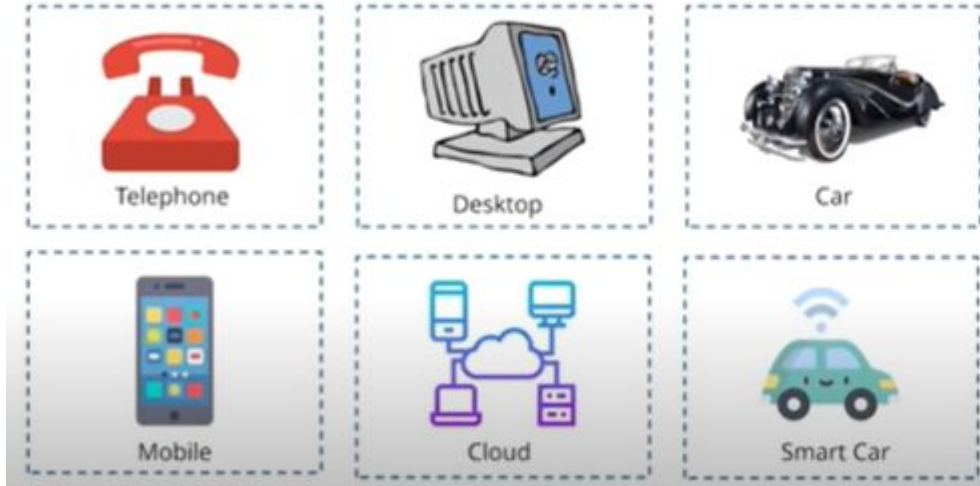
Name	Equal to
Bit (Bit)	1 Bit
Byte (Byte)	8 Bits
Kilobyte (KB)	1,024 Bytes
Megabyte (MB)	1,024 Kilobytes
Gigabyte (GB)	1,024 Megabytes
Terrabyte (TB)	1,024 Gigabytes
Petabyte (PB)	1,024 Terrabytes
Exabyte (EB)	1,024 Petabytes
Zettabyte (ZB)	1,024 Exabytes
Yottabyte (YB)	1,024 Zettabytes

Data growth



Data Sources

Evolution of Technology



Data Sources

Social Media



1,736,111 pictures



347,222 tweets



204,000,000
emails



4,166,667 likes &
200,000 photos



300 hours of video
uploaded

Data Sources

Other Factors



Data Analysis at Walmart

Halloween and cookie sales



Data scientist at Walmart found a connection between Halloween and the sales of cookies.

Data Analysis at Walmart

Hurricane and strawberry pop tarts



Data Analysis at Walmart

Social media and cake pops



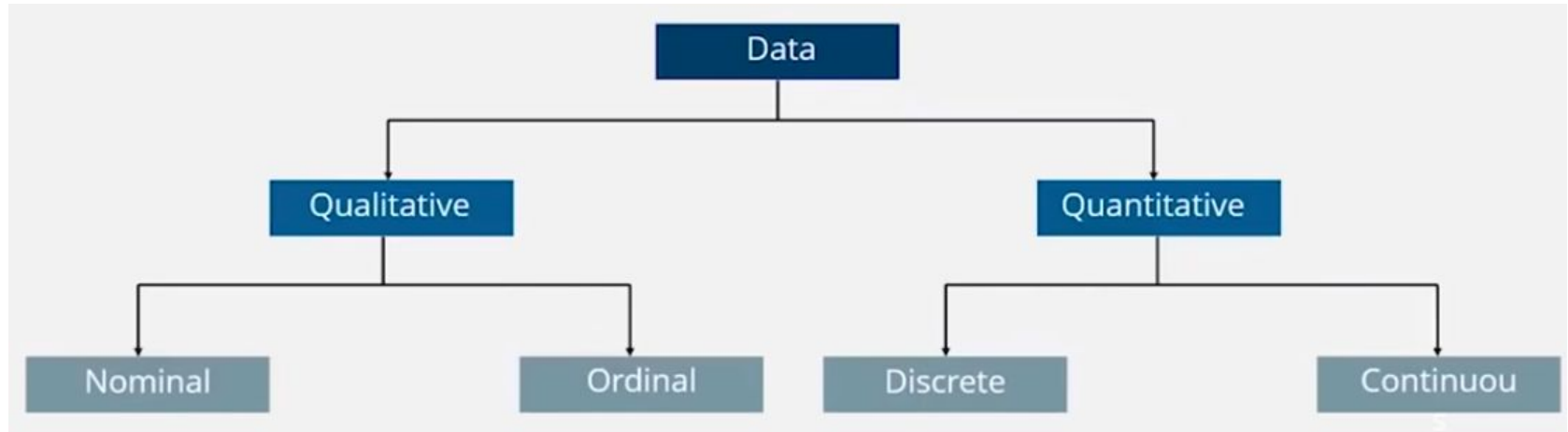
Walmart is leveraging social media data to find about the trending products so that they can be introduced to the Walmart stores across the world

What is Data?

- Data refers to facts and statistics collected together for reference or analysis.



Categories of data

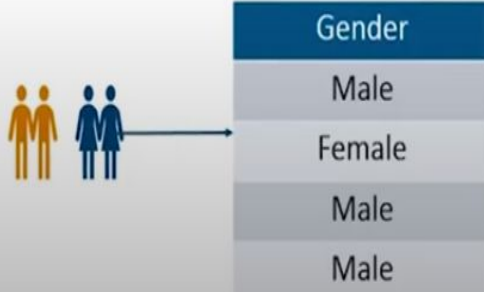


Qualitative data

Qualitative data deals with characteristics and descriptors that can't be easily measured, but can be observed subjectively.

Nominal Data

Data with no inherent order or ranking such as gender or race, such kind of data is called Nominal data



The diagram illustrates nominal data using gender. On the left, there are four stylized human figures: two orange (representing males) and two blue (representing females). An arrow points from these figures to a table. The table has a header 'Gender' and four rows of data: 'Male', 'Female', 'Male', and 'Male'.

Gender
Male
Female
Male
Male

Ordinal Data

Data with an ordered series, such as shown in the table such kind of data is called Ordinal data



The table illustrates ordinal data using customer ratings. It has two columns: 'Customer ID' and 'Rating'. The data is ordered from highest to lowest rating.

Customer ID	Rating
001	Good
002	Average
003	Average
004	Bad

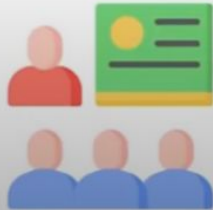
Quantitative data

Quantitative data deals with numbers and things you can measure objectively.

Discrete Data

Also known as categorical data, it can hold finite number of possible values.

Example: Number of students in a class



Continuous Data

Data that can hold infinite number of possible values.

Example: Weight of a person



Statistics

Statistics is an area of applied mathematics concerned with the data collection, analysis, interpretation and presentation.



Statistics: Example

Your company has created a new drug that may cure cancer. How would you conduct a test to confirm the drug's effectiveness?



Statistics: Example

You and a friend are at a baseball game, and out of the blue he offers you a bet that neither team will hit a home run in that game. Should you take the bet?

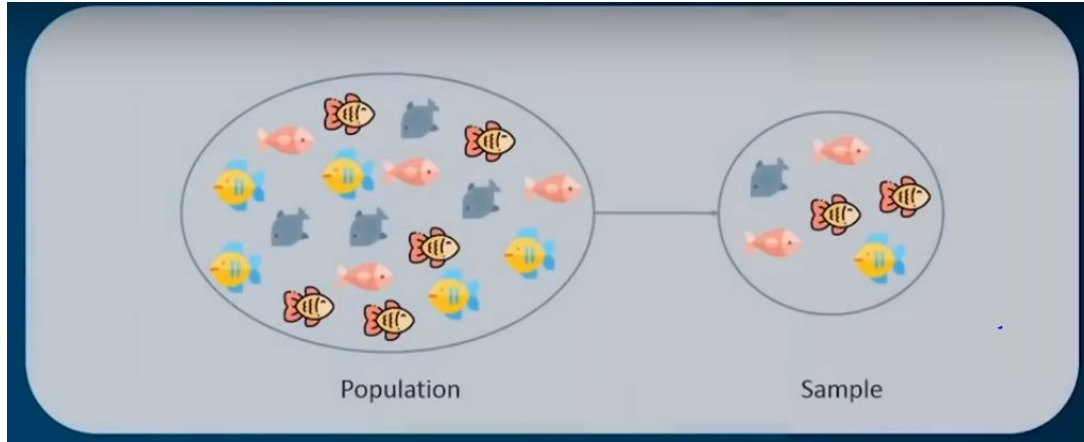


Statistics: Example

The latest sales data have just come in, and your boss wants you to prepare a report for management on places where the company could improve its business. What should you look for? What should you not look for?



Statistics: Terminology



Population is a collection or set of individuals or objects or events whose properties are to be analyzed.
Sample is a subset of population

Populations and Samples

Population can be any set of objects that we are interested in.

Example: All the customers of Netflix

Every car manufacturer

All set of tweets

Sample is a subset of population

Example: 500 Netflix customers

Toyoto, BMW, Ford

Tweets of politicians

When we take a sample, we take a subset of the units of size n in order to examine the observations to draw conclusions and make inferences about the population.

Sampling methods: Random Sampling

Random Sampling

Systematic Sampling

Stratified Sampling



Each member of the population has equal chance of being selected in the sample.

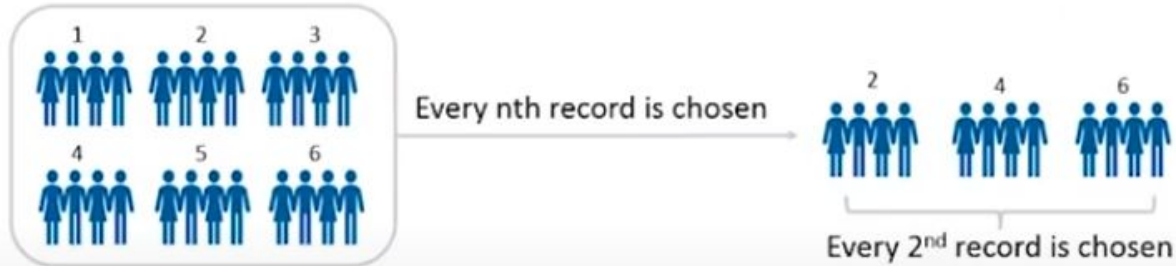
Sampling

Suppose your population was all tweets posted last one month.

Make a list of all the tweets and select 1/10th of those people at random and take all the tweets they posted, and that would be your sample. — Simple random sampling

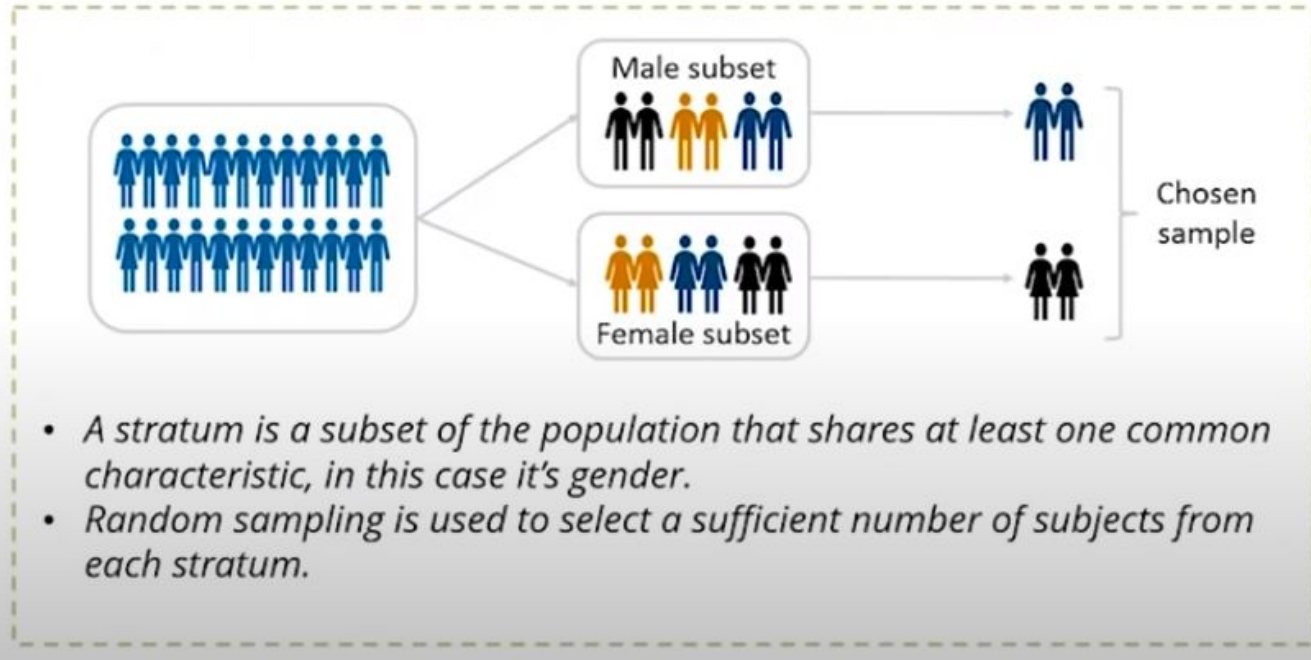
Sample 1/10th of all tweets sent each day at random, and that would be your sample. — Simple random sampling

Sampling methods: Systematic sampling



In Systematic sampling every nth record is chosen from the population to be a part of the sample.

Sampling methods: Stratified Sampling

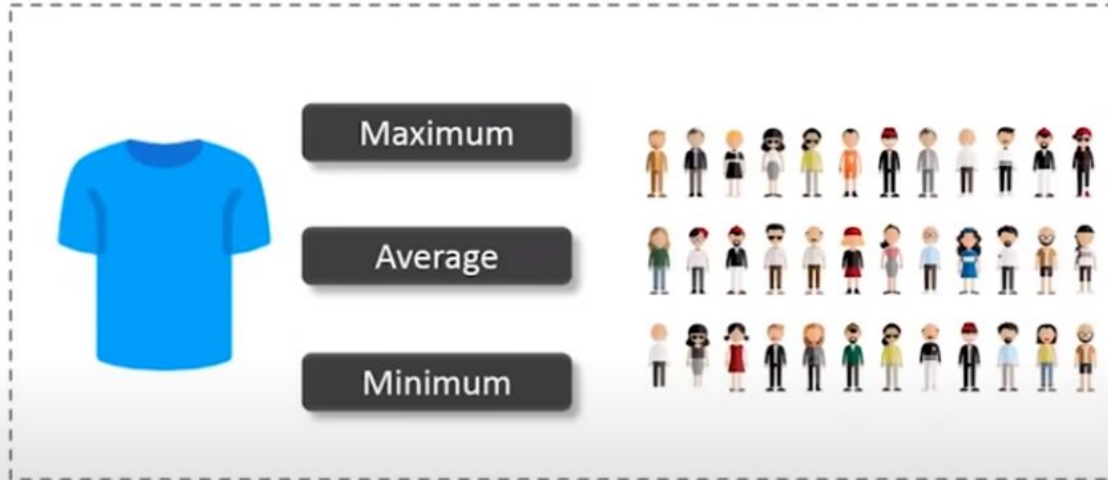


Types of statistics

1. Descriptive statistics
2. Inferential statistics

Descriptive Statistics

***Descriptive statistics** uses the data to provide descriptions of the population, either through numerical calculations or graphs or tables.*

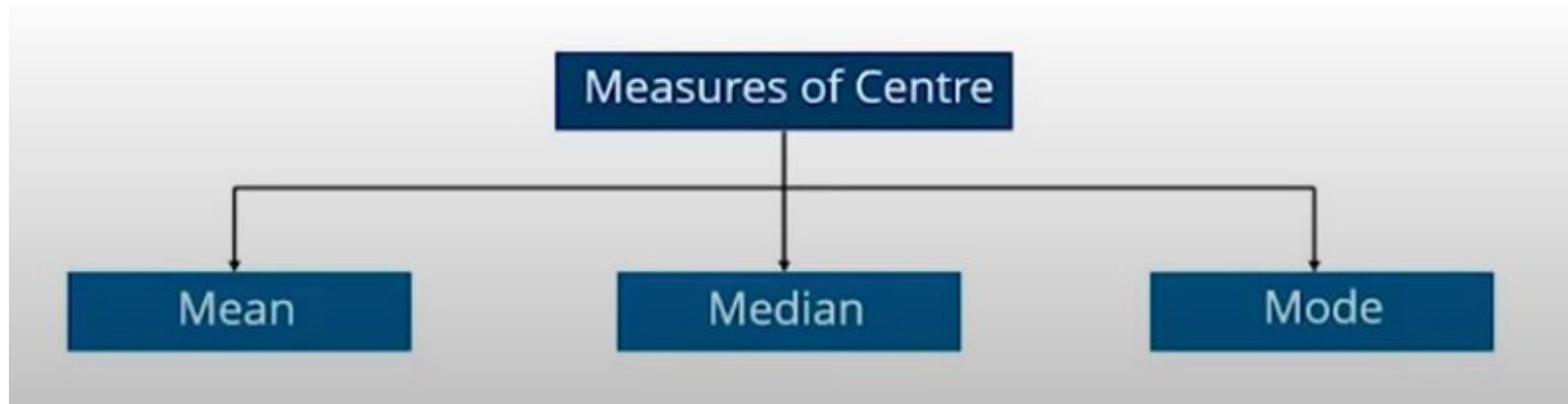


Descriptive Statistics is mainly focused upon the main characteristics of data. It provides graphical summary of the data.

Descriptive Statistics

Descriptive statistics is broken down into two categories:

- Measures of central tendency
- Measures of variability (spread)



Descriptive Statistics



How do we measure statistics?

Measures of center

Measures of center are how we define the middle, or center, of a dataset. There are different ways of defining the center of data.

1. Arithmetic mean

The arithmetic mean of a dataset is found by adding up all of the values and then dividing it by the number of data values.

This is likely the most common way to define the center of data.

Arithmetic mean is sensitive to outliers. Example: Mean of 11, 15, 17, and 14 is 14.25

What if a new value 31 is added to the data? It greatly affects the mean.

2. Median

The median is the number found in the middle of the dataset when it is sorted in order.

Example: Median of 11,14,15, and 17 is 14.5, Median of 11, 14,15, 17, and 31 is 15.

Introduction of 31 did not affect the median of the dataset greatly. This is because the median is less sensitive to outliers.

Measures of center

Mean

Measure of average of all the values in a sample is called Mean.

Here is a sample dataset of cars containing the variables:

- Cars,
- Mileage per Gallon(mpg)
- Cylinder Type (cyl)
- Displacement (disp)
- Horse Power(hp)
- Real Axle Ratio(drat)

Cars	mpg	cyl	disp	hp	drat
MazdaRX4	21	6	160	110	3.9
MazdaRX4_W					
AG	21	6	160	110	3.9
Datsun_710	22.8	4	108	93	3.85
Alto	21.3	6	108	96	3
WagonR	23	4	150	90	4
Toyata_11	23	6	108	110	3.9
Honda_12	23	4	160	110	3.9
Ford_11	23	6	160	110	3.9

To find out the average horsepower of the cars among the population of cars, we will check and calculate the average of all values:

$$\frac{110 + 110 + 93 + 96 + 90 + 110 + 110 + 110}{8} = 103.625$$

Measures of Center

Median

Measure of the central value of the sample set is called **Median**.

Here is a sample dataset of cars containing the variables:

- Cars,
- Mileage per Gallon(mpg)
- Cylinder Type (cyl)
- Displacement (disp)
- Horse Power(hp)
- Real Axle Ratio(drat)

Cars	mpg	cyl	disp	hp	drat
MazdaRX4	21	6	160	110	3.9
MazdaRX4_W					
AG	21	6	160	110	3.9
Datsun_710	22.8	4	108	93	3.85
Alto	21.3	6	108	96	3
WagonR	23	4	150	90	4
Toyota_11	23	6	108	110	3.9
Honda_12	23	4	160	110	3.9
Ford_11	23	6	160	110	3.9

To find out the center value of mpg among the population of cars, arrange records in *Ascending order*, i.e., **21, 21, 21.3, 22.8, 23, 23, 23, 23**

In case of even entries, take average of the two middle values, i.e. $(22.8+23)/2 = 22.9$

Measures of Center

Mode

The value most recurrent in the sample set is known as Mode.

Here is a sample dataset of cars containing the variables:

- Cars,
- Mileage per Gallon(mpg)
- Cylinder Type (cyl)
- Displacement (disp)
- Horse Power(hp)
- Real Axle Ratio(drat)

Cars	mpg	cyl	disp	hp	drat
MazdaRX4	21	6	160	110	3.9
MazdaRX4_W					
AG	21	6	160	110	3.9
Datsun_710	22.8	4	108	93	3.85
Alto	21.3	6	108	96	3
WagonR	23	4	150	90	4
Toyata_11	23	6	108	110	3.9
Honda_12	23	4	160	110	3.9
Ford_11	23	6	160	110	3.9

To find the most common type of cylinder among the population of cars, check the value which is repeated most number of times, i.e., *cylinder type 6*

Measures of spread

A measure of spread, sometimes also called a measure of dispersion, is used to describe the variability in a sample or population.

Range

Inter Quartile Range

Variance

Standard Deviation

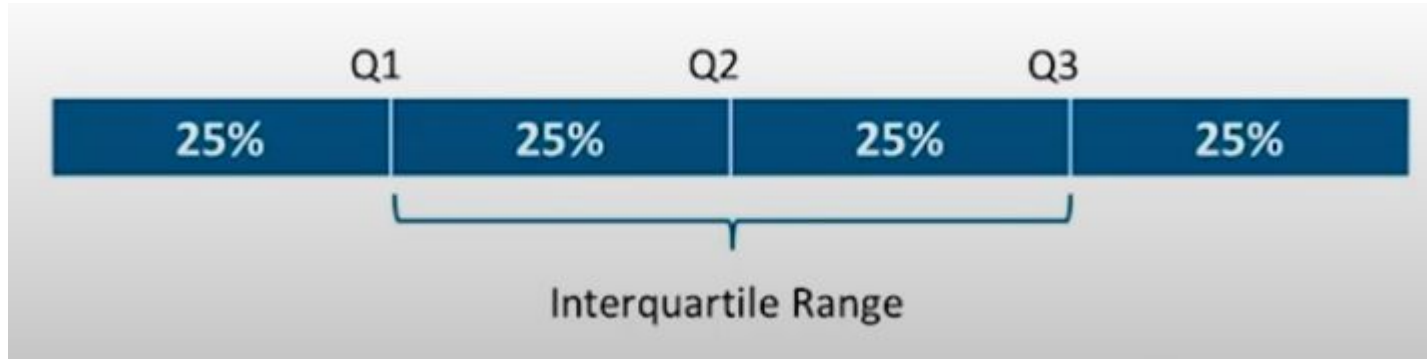
Measures of spread: Range

Range is the given measure of how spread apart the values in a dataset are.

$$\text{Range} = \text{Max}(x_i) - \text{Min}(x_i)$$

Measures of spread: Interquartile range

Quartiles tell us about the spread of a data set by breaking the data set into quarters, just like the median breaks it in half.



Measures of spread: Interquartile range

Consider the marks of the 100 students below, ordered from the lowest to the highest scores

The first quartile (Q1) lies between the 25th and 26th.
 $Q1 = (45 + 45) \div 2 = 45$

Order	Score	Order	Score	Order	Score	Order	Score	Order	Score
1st	35	21st	42	41st	53	61st	64	81st	74
2nd	37	22nd	42	42nd	53	62nd	64	82nd	74
3rd	37	23rd	44	43rd	54	63rd	65	83rd	74
4th	38	24th	44	44th	55	64th	66	84th	75
5th	39	25th	45	45th	55	65th	67	85th	75
6th	39	26th	45	46th	56	66th	67	86th	76
7th	39	27th	45	47th	57	67th	67	87th	77
8th	39	28th	45	48th	57	68th	67	88th	77
9th	39	29th	47	49th	58	69th	68	89th	79
10th	40	30th	48	50th	58	70th	69	90th	80
11th	40	31st	49	51st	59	71st	69	91st	81
12th	40	32nd	49	52nd	60	72nd	69	92nd	81
13th	40	33rd	49	53rd	61	73rd	70	93rd	81
14th	40	34th	49	54th	62	74th	70	94th	81
15th	40	35th	51	55th	62	75th	71	95th	81
16th	41	36th	51	56th	62	76th	71	96th	81
17th	41	37th	51	57th	63	77th	71	97th	83
18th	42	38th	51	58th	63	78th	72	98th	84
19th	42	39th	52	59th	64	79th	74	99th	84
20th	42	40th	52	60th	64	80th	74	100th	85

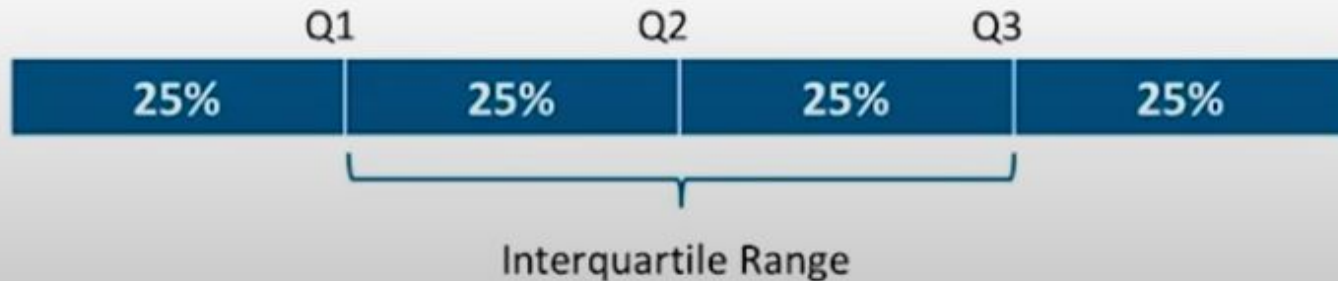
The second quartile (Q2) lies between the 50th and 51st.
 $Q2 = (58 + 59) \div 2 = 58.5$

The third quartile (Q3) lies between the 75th and 76th.
 $Q3 = (71 + 71) \div 2 = 71$

Measures of spread: Interquartile range

Inter Quartile Range(IQR) is the measure of variability, based on dividing a dataset into quartiles.

- *Quartiles divide a rank-ordered data set into four equal parts, denoted by Q1, Q2, and Q3, respectively*
- *The interquartile range is equal to Q3 minus Q1, i.e.. $IQR = Q3 - Q1$*



Measures of spread: Variance

*Variance describes how much a random variable differs from its expected value.
It entails computing squares of deviations.*

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

x : Individual data points

n : Total number of data points

\bar{x} : Mean of data points

Deviation is the difference between each element from the mean.

$$\text{Deviation} = (x_i - \mu)$$

Measures of spread: Variance

Population Variance is the average of squared deviations.

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2$$

Sample Variance is the average of squared differences from the mean.

$$s^2 = \frac{1}{(n-1)} \sum_{i=1}^N (x_i - \bar{x})^2$$

Measures of spread: Standard Deviation

Standard Deviation is the measure of the dispersion of a set of data from its mean.

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$

Standard Deviation use case

Standard Deviation Use Case: Daenerys has 20 Dragons. They have the numbers 9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4. Work out the Standard Deviation.

STEP 1

Find out the mean for your sample set.

The Mean is:

$$\frac{9+2+5+4+12+7+8+11+9+3+7+4+12+5+4+10+9+6+9+4}{20}$$

$$\mu=7$$

Standard Deviation use case

STEP 2

Then for each number, subtract the Mean and square the result.

$$(x_i - \mu)^2$$

$$(9-7)^2 = 2^2 = 4$$

$$(2-7)^2 = (-5)^2 = 25$$

$$(5-7)^2 = (-2)^2 = 4$$

And so on...

□ We get the following results:

4, 25, 4, 9, 25, 0, 1, 16, 4, 16, 0, 9, 25, 4, 9, 9, 4, 1, 4, 9

Standard Deviation use case

STEP 3

Then work out the mean of those squared differences.

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$

$$\frac{4+25+4+9+25+0+1+16+4+16+0+9+25+4+9+9+4+1+4+9}{20}$$

$$\sigma^2 = 8.9$$

Standard Deviation use case

STEP 4

Take square root of σ^2 .

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2}$$

$$\sigma = 2.983$$

Descriptive Statistics using R

- [..\..\Descriptive Statistics\mean_median_mode.R](#)
- Click here for code
https://docs.google.com/document/d/1TvD9n60EPaeWF8_sff2juNFyDgi8-BtKNC_DnHiTMGCA/edit?usp=sharing

Understand mode calculation:

https://rpubs.com/ashishgopal1414/Function_Mode

Measures of variation (Spread): Revisit

Measuring how "spread out" the data we collect is.

Consider that we take a random sample of 24 of our friends on Facebook and wrote down how many friends that they had on Facebook. Here's the list:

```
friends = [109, 1017, 1127, 418, 625, 957, 89, 950, 946, 797, 981, 125,  
455, 731, 1640, 485, 1309, 472, 1132, 1773, 906, 531, 742, 621]
```

The most basic measure of variation: the range. The range is simply the maximum value minus the minimum value, i.e. $1773 - 89 = 1684$.

The most commonly used measure of variation is standard deviation - it measures how much data values deviate from the arithmetic mean.

The general formula to calculate the standard deviation is

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

s is our sample standard deviation

x is each individual data point.

\bar{x} is the mean of the data

n is the number of data points

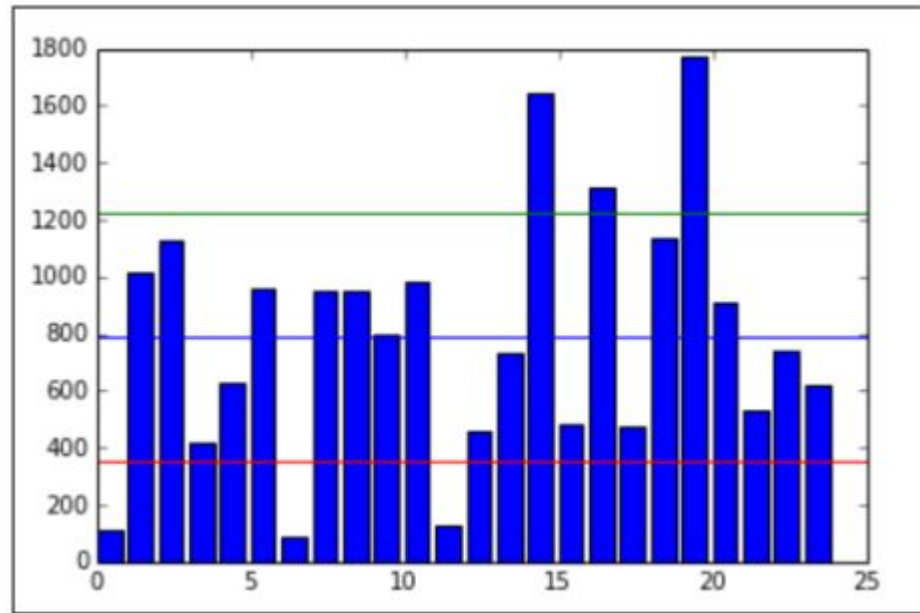
$$s = \sqrt{\frac{(109 - 789)^2 + (1017 - 789)^2 + \dots + (621 - 789)^2}{24}}$$

$s=425.2$, it is the average distance the data values are from the mean, the data is pretty spread out.

In the following plot, every person will be represented by a single bar in the bar chart, and the height of the bars represent the number of friends that the individuals have.

The blue line in the center is drawn at the mean (789), the red line on the bottom is drawn at the mean minus the standard deviation ($789 - 425 = 364$), and, finally, the green line towards the top is drawn at the mean plus the standard deviation ($789 + 425 = 1,214$).

Most of the data lies between the green and the red lines while the outliers lie outside the lines. There are three people who have friend counts below the red line and three people who have a friend count above the green line.



Coefficient of variation

The coefficient of variation is defined as the ratio of the data's standard deviation to its mean.

This is a way to standardize the standard deviation, which makes it easier to compare across datasets.

The last column in the above table is coefficient of variation.

Everyone in the mailroom, while not making as much money, are making just about the same as everyone else in the mailroom, which is why their coefficient of variation is only 8%.

The people in the executive department may be getting paid more but employees in the executive department are getting wildly different salaries.

Salaries of Company XYZ			
Department	Mean Salary	SD	CoV
Mailroom	\$25,000	\$2,000	8.0%
Human Resources	\$52,000	\$7,000	13.5%
Executive	\$124,000	\$42,000	33.9%

Measures of relative standing

Combine both the measures of centers and variations to create measures of relative standings.

The z-score is a way of telling us how far away a single data value is from the mean.

The z-score of a x data value is as follows:

$$z = \frac{x - \bar{x}}{s}$$

x is the data point

\bar{x} is the mean

s is the standard deviation.

```
friends = [109, 1017, 1127, 418, 625, 957, 89, 950, 946, 797, 981,  
125, 455, 731, 1640, 485, 1309, 472, 1132, 1773, 906, 531, 742,  
621]
```

From the above friends data, find the Z-score of each person and plot a histogram.

Measures of relative standing

A z-score measures exactly how many standard deviations above or below the mean a data point is.

A positive z-score says the data point is above average.

A negative z-score says the data point is below average.

A z-score close to 0 says the data point is close to average.

Statistical Inference

Statistical inference is the discipline that concerns itself with the development of procedures, methods, and theorems that allow us to extract meaning and information from data that has been generated by stochastic (random) processes.

-Cathy O'Neil and Rachel Schutt *Doing Data Science*, Straight Talk from The Frontline. O'Reilly. 2014

A statistic is a descriptive measure computed from data of a sample. For example, the sample mean (average), median (middle value), or sample standard deviation (a measure of typical deviation) are all statistics.

Statistical inference is the process of drawing conclusions about an underlying population based on a sample or subset of the data.

-Thomas D. Gauthier, Mark E. Hawley, in [Introduction to Environmental Forensics \(Third Edition\)](#), 2015

Probability

- <https://ell.brainpop.com/level3/unit4/lesson3/movie/>

Probability

Probability is a mathematical method used for statistical analysis.

Probability and statistics are interconnected branches of mathematics that deal with the analyzing the relative frequency of events.

Probability is the measure of how likely an event will occur.

- Probability is the ratio of desired outcomes to total outcomes:
(desired outcomes) / (total outcomes)

- Probabilities of all outcomes always sums to 1

Example:

- On rolling a dice, you get 6 possible outcomes
- Each possibility only has one outcome, so each has a probability of $1/6$
- For example, the probability of getting a number '2' on the dice is $1/6$



Terminologies in Probability

Random Experiment

An experiment or a process for which the outcome cannot be predicted with certainty

Sample Space

The entire possible set of outcomes of a random experiment is the sample space (S) of that experiment

Event

One or more outcomes of an experiment. It is a subset of sample space(S)

Terminologies in Probability

Random Experiments:

Before rolling a die you do not know the result. This is an example of a random experiment.

An outcome is a result of a random experiment.

The set of all possible outcomes is called the sample space.

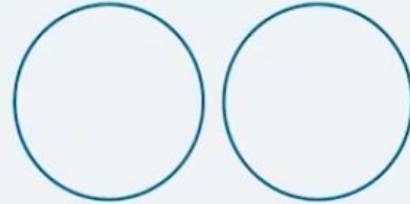
Random experiment: toss a coin; sample space: $S=\{H,T\}$

Random experiment: roll a die; sample space: $S=\{1,2,3,4,5,6\}$

Types of events

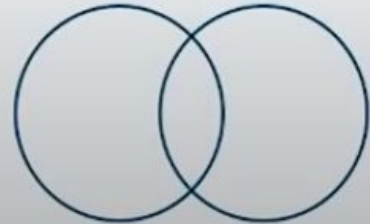
Disjoint Events do not have any common outcomes.

- The outcome of a ball delivered cannot be a sixer and a wicket
- A single card drawn from a deck cannot be a king and a queen
- A man cannot be dead and alive



Non-Disjoint Events can have common outcomes

- A student can get 100 marks in statistics and 100 marks in probability
- The outcome of a ball delivered can be a no ball and a six



Types of events

Dependent events

Not paying your power bill on time and having your power cut off.

Being the first person to enter a movie theater and finding a good seat.

Independent events

Taking a cab home and finding your favorite movie on cable.

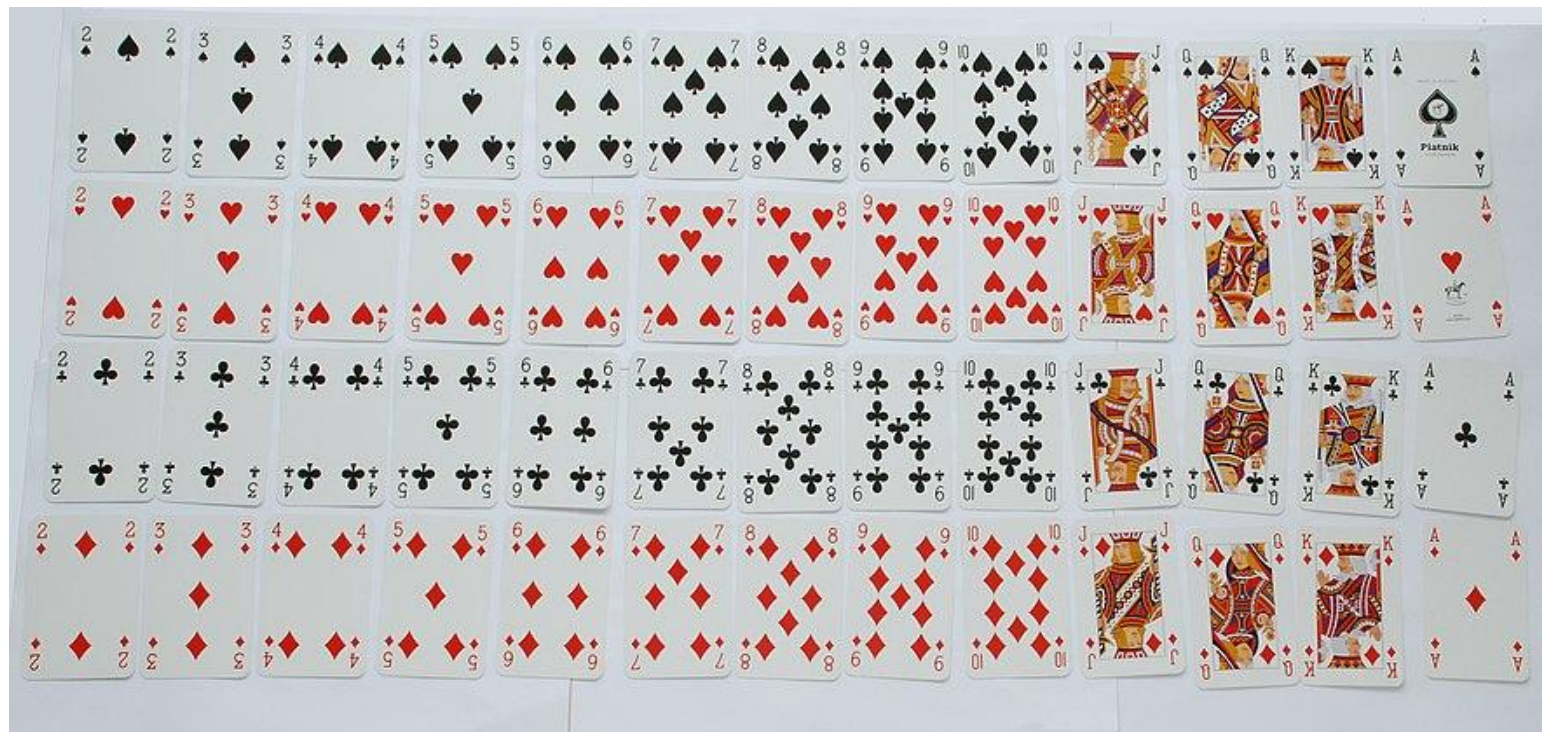
Owning a dog and having a car.

If we roll a die twice, the outcome of the first roll and second roll have no effect on each other.

Types of probability

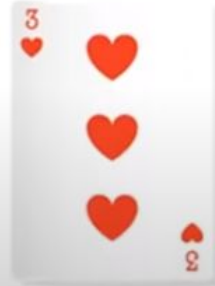
1. Marginal Probability
2. Joint Probability
3. Conditional Probability

Deck of cards



Marginal Probability

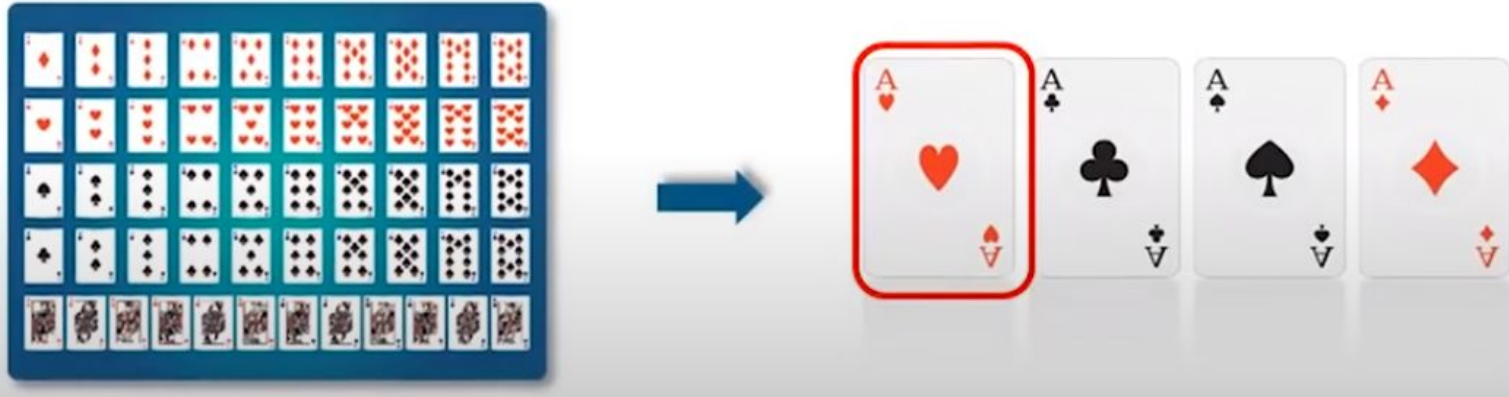
Marginal Probability is the probability of occurrence of a single event.



$$\text{Marginal Probability} = \frac{13}{52}$$

Joint Probability

Joint Probability is a measure of two events happening at the same time



Probability of a card drawn is 4 and red $P(4 \cap \text{red}) = P(4) \times P(\text{red})$
 $= 4/52 \times 26/52 = 2/52 = 1/26$

Conditional Probability

- *Probability of an event or outcome based on the occurrence of a previous event or outcome*
- *Conditional Probability of an event B is the probability that the event will occur given that an event A has already occurred*

If A and B are dependent events then the expression for conditional probability is given by:

$$P(B|A) = P(A \text{ and } B) / P(A)$$

If A and B are independent events then the expression for conditional probability is given by:

$$P(B|A) = P(B)$$

Conditional Probability

- Consider that a fair die has been rolled and you are asked to give the probability that it was a five. You get extra information that the number rolled was odd.
- A – {The number is five}
- B – {The number is odd}

$$P(A|B)=P(A\cap B)/P(B) = (1/6)/(3/6) = 1/3$$

Probability distributions

- A probability distribution is the mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment.
- For instance, if X is used to denote the outcome of a coin toss ("the experiment"), then the probability distribution of X would take the value 0.5 (1 in 2 or $1/2$) for $X = \text{heads}$, and 0.5 for $X = \text{tails}$

We can describe the probability distribution of one coin flip using a probability table:

Outcome	Probability
Heads	Tails
.5	.5

Probability distributions

- There are two types of probability distributions:
 - Discrete probability distributions
 - Continuous probability distributions
- A discrete probability distribution is a probability distribution of a categorical or discrete variable.
- **A probability mass function (PMF)** is a mathematical function that describes a discrete probability distribution. It gives the probability of every possible value of a variable.
- Probability mass function assigns a particular probability to every possible value of a discrete random variable.

Discrete probability distributions

- Let X be the discrete random variable. Then the formula for the probability mass function, $f(x)$, evaluated at x , is given as follows:
- $f(x) = P(X = x)$
- The probability mass function associated with a random variable can be represented with the help of a table or by using a graph.
- Example: X , represents the number of heads in the coin tosses.
- The sample space created is $[HH, TH, HT, TT]$. This shows that X can take the values 0 (no heads), 1 (1 head), and 2 (2 heads).

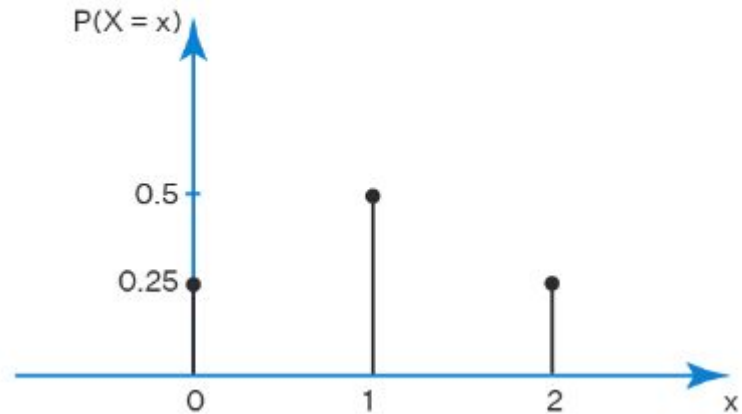
A probability mass function table displays the various values that can be taken up by the discrete random variable

x	$P(X = x)$
0	0.25
1	0.5
2	0.25

Discrete probability distributions

The probability mass function graph is used to display the probabilities associated with the possible values of the random variable.

Probability Mass Function Graph



Properties of Probability mass function

- The cumulative distribution function, $P(X \leq x)$, can be determined by summing up the probabilities of x values.
- $P(X = x) = f(x) > 0$. This implies that for every element x associated with a sample space, all probabilities must be positive.
- The sum of all probabilities associated with x values of a discrete random variable will be equal to 1.
- The probability associated with an event T can be determined by adding all the probabilities of the x values in T .

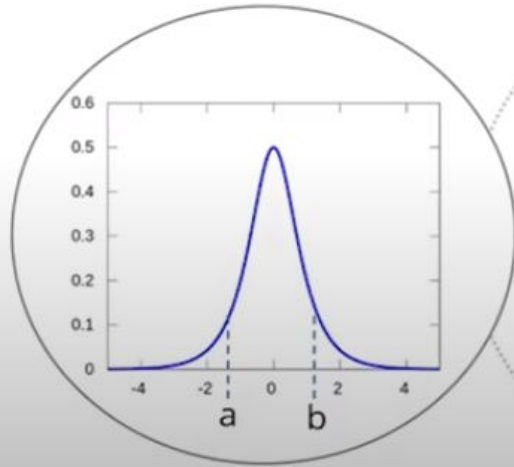
Continuous probability distribution

- A continuous probability distribution is the probability distribution of a continuous variable.
- A continuous random variable can be defined as a variable that can take on infinitely many values.
- A continuous variable can have any value between its lowest and highest values.
- Therefore, continuous probability distributions include every number in the variable's range.
- If X is continuous, the probability that X takes on any specific value x is 0.
- A **probability density function** (PDF) is a mathematical function that describes a continuous probability distribution.

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

Probability density function

The equation describing a continuous probability distribution is called a Probability Density Function



Property 01



Graph of a PDF will be continuous over a range



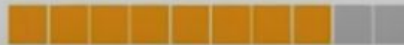
Property 02



Area bounded by the curve of density function and the x-axis is equal to 1



Property 03



Probability that a random variable assumes a value between a & b is equal to the area under the PDF bounded by a & b

Probability density function

Let X be a continuous random variable with the PDF given by:

$$f(x) = \begin{cases} x; & 0 < x < 1 \\ 2 - x; & 1 < x < 2 \\ 0; & x > 2 \end{cases}$$

Find $P(0.5 < x < 1.5)$.

Solution:

Given PDF is:

$$f(x) = \begin{cases} x; & 0 < x < 1 \\ 2 - x; & 1 < x < 2 \\ 0; & x > 2 \end{cases}$$

$$P(0.5 < X < 1.5) = \int_{0.5}^{1.5} f(x)dx$$

Let us split the integral by taking the intervals as given below:

$$= \int_{0.5}^1 f(x)dx + \int_1^{1.5} f(x)dx$$

Substituting the corresponding values of $f(x)$ based on the intervals, we get;

$$= \int_{0.5}^1 xdx + \int_1^{1.5} (2 - x)dx$$

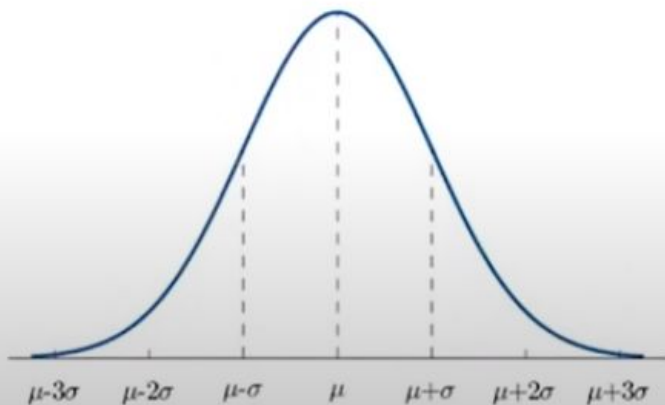
Probability density function

Integrating the functions, we get;

$$\begin{aligned} &= \left(\frac{x^2}{2}\right)_{0.5}^1 + \left(2x - \frac{x^2}{2}\right)_1^{1.5} \\ &= [(1)^2/2 - (0.5)^2/2] + \{[2(1.5) - (1.5)^2/2] - [2(1) - (1)^2/2]\} \\ &= [(1/2) - (1/8)] + \{[3 - (9/8)] - [2 - (1/2)]\} \\ &= (3/8) + [(15/8) - (3/2)] \\ &= (3 + 15 - 12)/8 \\ &= 6/8 \\ &= 3/4 \end{aligned}$$

Normal Distribution

The Normal Distribution is a probability distribution that associates the normal random variable X with a cumulative probability



$$Y = [1/\sigma * \text{sqrt}(2\pi)] * e^{-(x - \mu)^2/2\sigma^2}$$

Where,

- X is a normal random variable
- μ is the mean and
- σ is the standard deviation

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

$f(x)$ = probability density function

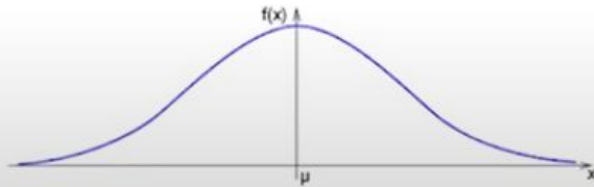
σ = standard deviation

μ = mean

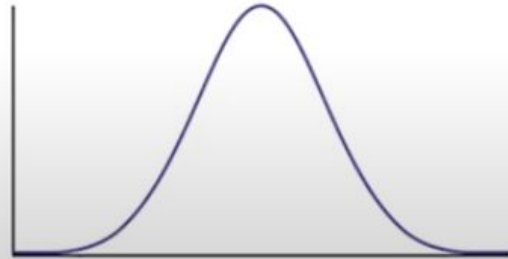
Normal Distribution

The graph of the Normal Distribution depends on two factors: the *Mean* and the *Standard Deviation*

- **Mean:** *Determines the location of center of the graph*
- **Standard Deviation:** *Determines the height of the graph*



If the standard deviation is large,
the curve is short and wide.



If the standard deviation is small,
the curve is tall and narrow.

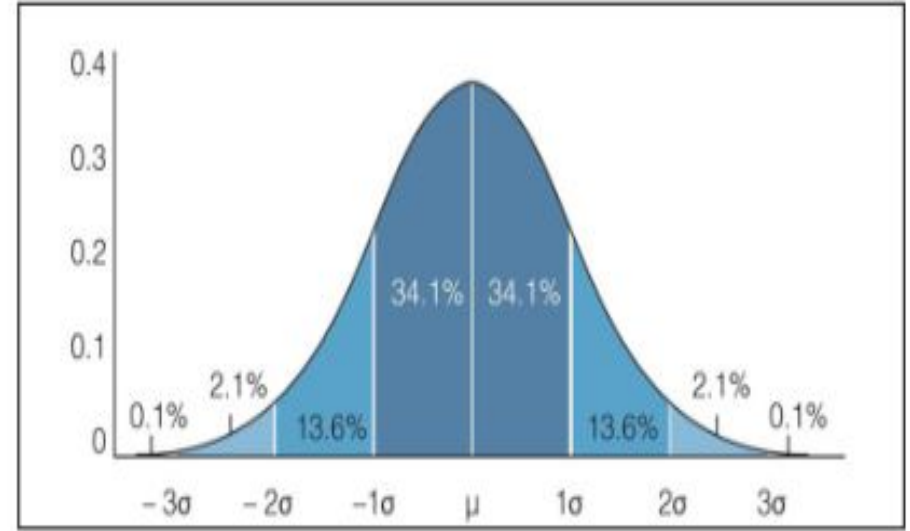
Normal distribution

- The most common and naturally occurring distribution is Normal Distribution. It is also known as Gaussian Distribution.
- The mean, median and mode are exactly the same.
- The distribution is symmetric about the mean—half the values fall below the mean and half above the mean.
- The distribution can be described by two values: the mean and the standard deviation.

Normal distribution

The Empirical rule states that we can expect a certain amount of data to live between sets of standard deviations. Specifically, the Empirical rule states for data that is distributed normally:

- about 68% of the data fall within 1 standard deviation
- about 95% of the data fall within 2 standard deviations
- about 99.7% of the data fall within 3 standard deviations



The shape of the distribution resembles that of a classic bell, and hence it is called a bell-shaped curve.

The normal distribution is written as

$$N(x|\mu, \sigma) \sim \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

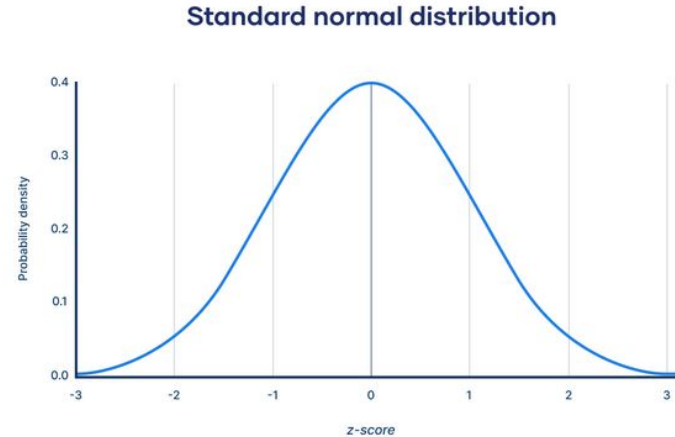
x is a random variable

Parameter μ is the mean

Parameter σ is the standard deviation

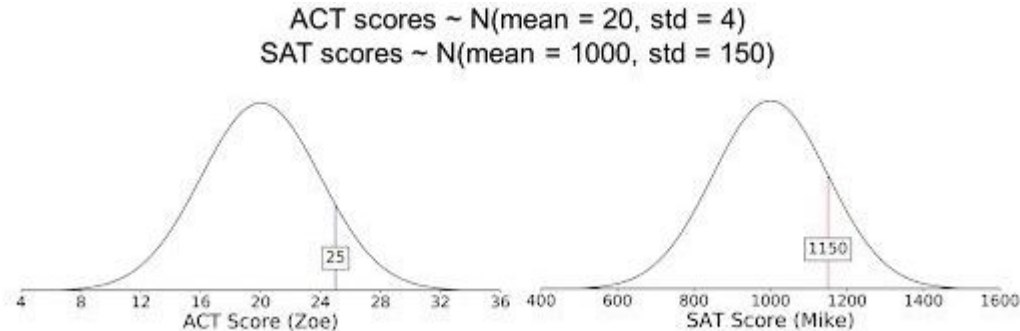
Standard Normal Distribution

- The simplest case of a normal distribution is known as the standard normal distribution or unit normal distribution.
- This is a special case when $\mu = 0$ and $\sigma = 1$, and it is described by this probability density function (or density):



Standard Normal Distribution

- Imagine a scenario where we compare the standardized test results from two students. Let's call them Zoe and Mike. Zoe took the ACT and scored a 25, while Mike took the SAT and scored 1150. Which of the test takers scored better? And what proportion of people scored worse than Zoe and Mike?



Standard Normal Distribution

- Since Zoe has a higher z-score than Mike, Zoe performed better on her test.

ACT Score (Zoe)

$$z = \frac{25 - 20}{4}$$

$$z = 1.25$$

SAT Score (Mike)

$$z = \frac{1150 - 1000}{150}$$

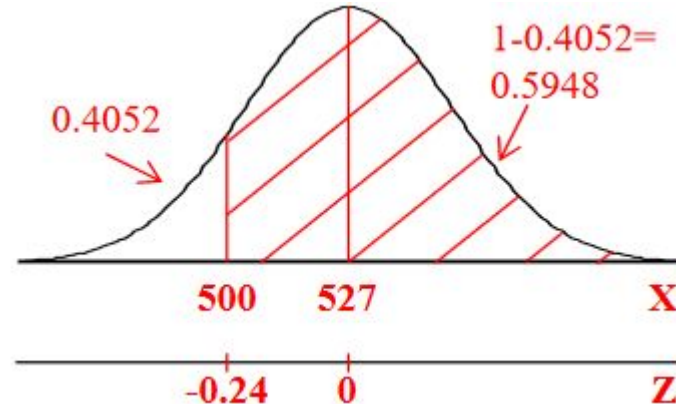
$$z = 1.0$$

Link to Z-table: https://drive.google.com/file/d/1PNRYpMnZc89rXW16IG2w0XJvE_TzG8bJ/view?usp=sharing

- The value in the Z table for 1.25 is .8944 which is the probability. Roughly 89.44 percent of people scored worse than Zoe on the ACT.
- The value in the Z table for 1.0 is .8413, which is the probability. Roughly 84.13 percent of people scored worse than Mike on the SAT.

Standard Normal Distribution

- Most graduate schools of business require applicants for admission to take the Graduate Management Admission Council's GMAT examination. Scores on the GMAT are roughly normally distributed with a mean of 527 and a standard deviation of 112. What is the probability of an individual scoring above 500 on the GMAT?



Standard Normal Distribution

Normal Distribution

$$Z = \frac{500 - 527}{112} = -0.24107$$

$$\mu = 527$$

$$\sigma = 112$$

$$\Pr\{X > 500\} = \Pr\{Z > -0.24\} = 1 - 0.4052 = \boxed{0.5948}$$

- The average number of acres burned by forest in a large New Mexico country is 4,300 acres per year, with a standard deviation of 750 acres. The distribution of the number of acres burned is normal. What is the probability that between 2,500 and 4,200 acres will be burned in any given year?

Normal Distribution

$\mu = 4300$

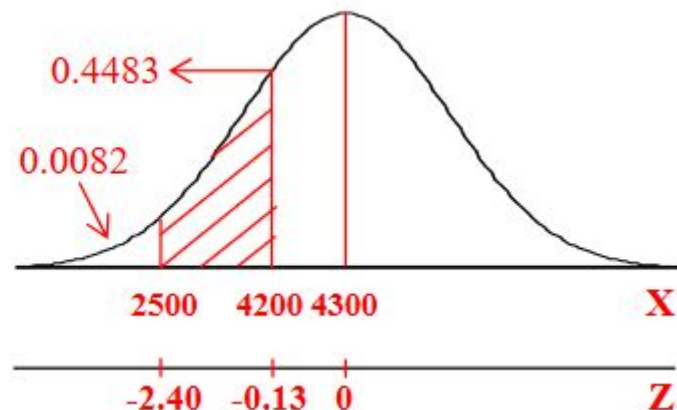
$\sigma = 750$

$$P(2500 < X < 4200) = P(-2.40 < Z < -0.13)$$

$$P(-2.40 < Z < -0.13) = P(Z < -0.13) - P(Z < -2.40)$$

$$P(-2.40 < Z < -0.13) = 0.4483 - 0.0082 = \boxed{0.4401}$$

$$Z = \frac{2500 - 4300}{750} = -2.40$$
$$Z = \frac{4200 - 4300}{750} = -0.13333$$



Introduction to Data science

The word science has been used since the 14th century in the sense of "the state of knowing".

It was borrowed from the Latin word scientia, meaning "knowledge, awareness, understanding".

Data is a collection of discrete values that convey information, describing quantity, quality, fact, statistics, other basic units of meaning

The word "data" was first used to mean "transmissible and storable computer information" in 1946.

----- Wikipedia

Introduction to Data science

Data science is the art and science of acquiring knowledge through data.

Data science is all about how we take data, use it to acquire knowledge, and then use that knowledge to do the following:

- Make decisions
- Predict the future
- Understand the past/present
- Create new industries/products

Data science definitions

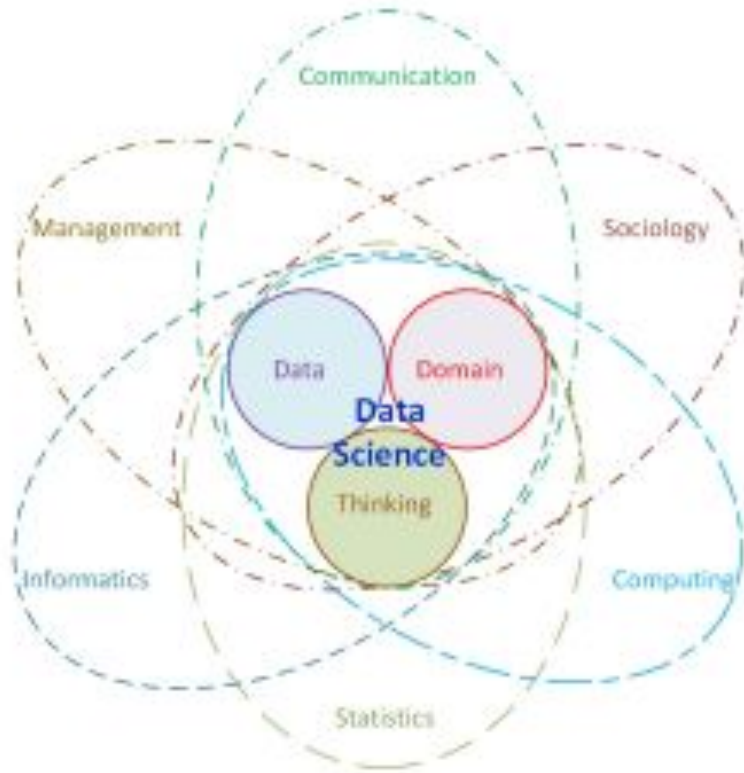
Definiton1: “data science is the science of data” or “data science is the study of data.”

Definition2: Data science is a new trans-disciplinary field that builds on and synthesizes a number of relevant disciplines and bodies of knowledge, such as statistics, informatics, computing, communication, management and sociology, to study data and its domain following a data science thinking.

A discipline-based data science formula is given as follows: data science = statistics + informatics + computing + communication + sociology + management | data + environment + thinking, where “|” means “conditional on.”

Longbing Cao. 2017. Data Science: A Comprehensive Overview. ACM Comput. Surv. 50, 3, Article 43 (May 2018), 42 pages. <https://doi.org/10.1145/3076253>

Trans-disciplinary data science



Longbing Cao. 2017. Data science: challenges and directions. *Commun. ACM* 60, 8 (August 2017), 59–68. <https://doi.org/10.1145/3015456>

Data science versus Applied data science

In applied data science, the art of researching is added on with data science.

Applied data science focus on researching new applications, new algorithms, making conventional algorithms faster by optimizing mathematical functions.

Data products

The outputs of data science are data products. It can be a discovery, prediction, service, recommendation, decision-making insight, thinking, model, mode, paradigm, tool, or system. The ultimate data products of value are knowledge, intelligence, wisdom, and decision.

Longbing Cao. 2017. Data Science: A Comprehensive Overview. ACM Comput. Surv. 50, 3, Article 43 (May 2018), 42 pages. <https://doi.org/10.1145/3076253>

Knowledge is the awareness of facts.

Intelligence is the ability to acquire and apply knowledge.

Wisdom is the ability to act productively using knowledge obtained through experience.

Data Scientist

A **data scientist** is someone who creates programming code and combines it with statistical knowledge to create insights from data.

Data scientists are responsible for breaking down the data into usable information and creating software and algorithms that help companies and organizations determine optimal operations.

-Wikipedia

Example

Ben Runkle, CEO, Sigma Technologies, is trying to resolve a huge problem.

The company is consistently losing long-time customers. He does not know why they are leaving, but he must do something fast.

He is convinced that in order to reduce his churn, he must create new products and features, and consolidate existing technologies.

To be safe, he calls in his chief data scientist, Dr. Jessie Hughan.

However, she is not convinced that new products and features alone will save the company. Instead, she turns to the transcripts of recent customer service tickets.

- "... Not sure how to export this; are you?"
- "Where is the button that makes a new list?"
- "Wait, do you even know where the slider is?"
- "If I can't figure this out today, it's a real problem..."

Example

It is clear that customers were having problems with the existing UI/UX (user interface/user experience).

We tend to call people like Runkle, a driver. He wants to make all decisions quickly and iterate over solutions until something works.

Dr. Haghun is much more analytical. She wants to solve the problem just as much as Runkle, but she turns to user-generated data instead of her gut feeling for answers.

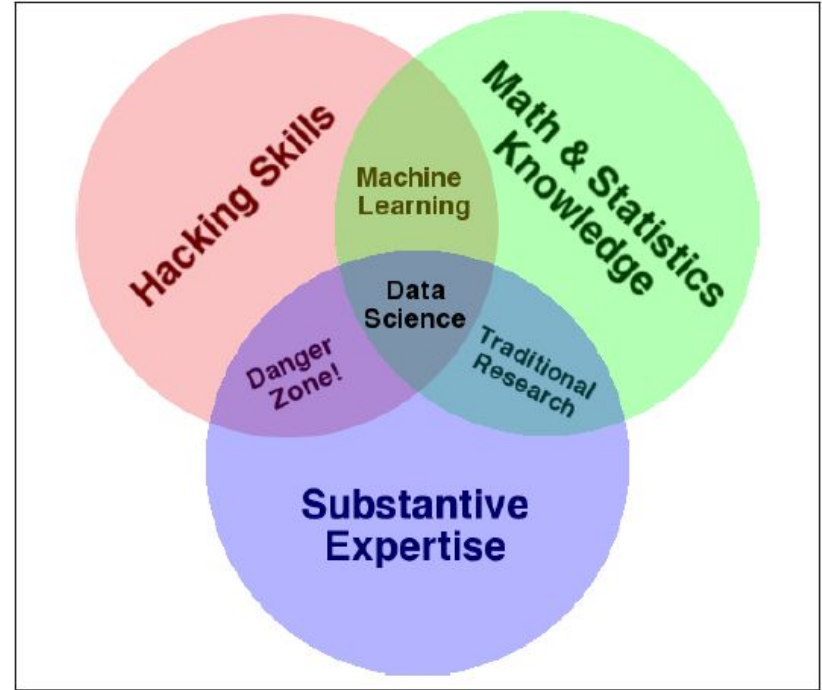
Data science is about applying the skills of the analytical mind and using them as a driver would - using data generated by the company as her source of information rather than just picking up a solution and going with it.

Data science Venn diagram

Understanding data science begins with three basic areas:

- Math/statistics: This is the use of equations and formulas to perform analysis
- Computer programming: This is the ability to use code to create outcomes on the computer
- Domain knowledge: This refers to understanding the problem domain (medicine, finance, social science, and so on)

- This is called as Drew Conway's venn diagram <http://drewconway.com/zia/?p=2378>



- Data Science is the intersection of the three key areas.
- In order to gain knowledge from data, we must be able to utilize computer programming to access the data, understand the mathematics behind the models we derive, and above all, understand our analyses' place in the domain we are in.

The Math

The Math needed for data science is specifically statistics and probability.

These subdomains of mathematics are used to create what are called models.

A data model refers to an organized and formal relationship between elements of data, usually meant to simulate a real-world phenomenon.

Example: In Biology, spawner-recruit model is used to judge the biological health of a species.

It is a basic relationship between the number of healthy parental units of a species and the number of new units in the group of animals.

The following graph was formed to visualize the relationship between the solman spawners and recruits. There is some sort of positive relationship.

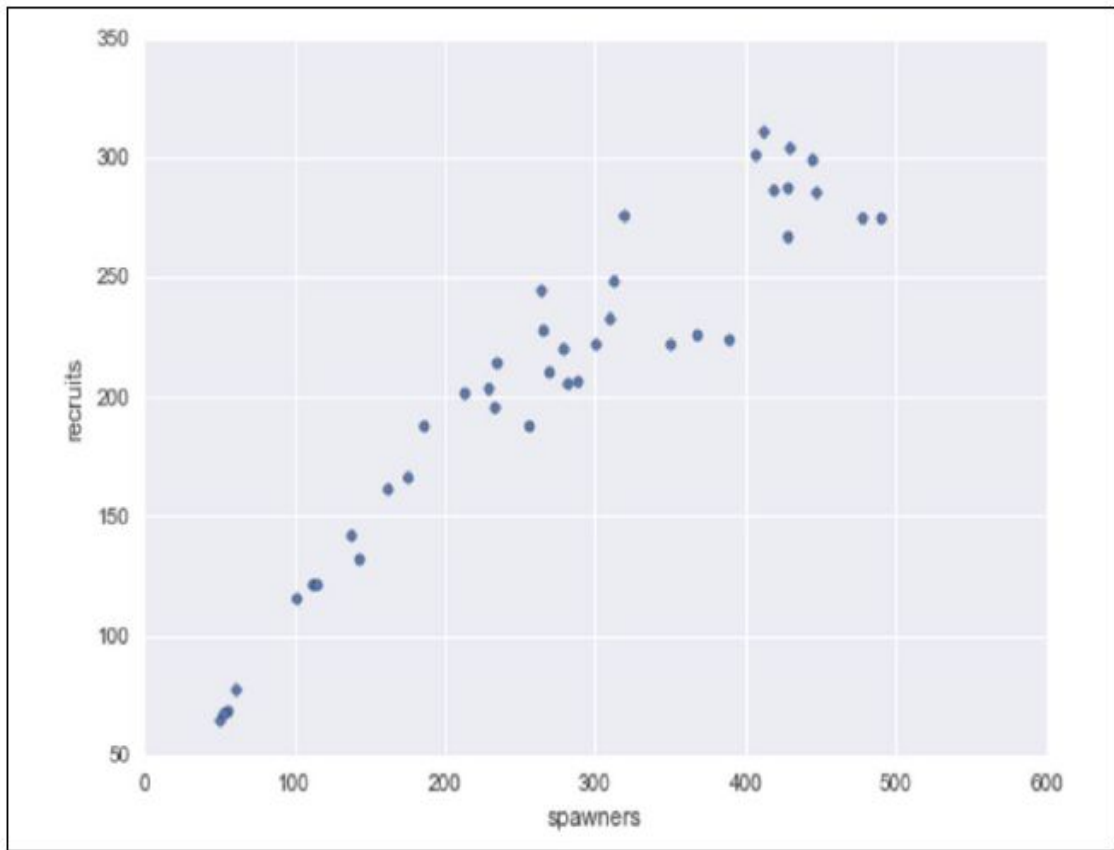
How to formalize this relationship?

Models allow us to plug in one variable to get the other. Consider the following example:

$$\text{Recruits} = 0.5 * \text{Spawners} + 60$$

In this example, let's say we knew that a group of salmons had 1.15 (in thousands) of spawners. Then, we would have the following:

$$\text{Recruits} = 0.5 * 1.15 + 60$$



Computer Programming

Computer languages are how we communicate with the machine.

A computer speaks many languages and, like a book, can be written in many languages; similarly, data science can also be done in many languages. Python, Julia, and R are some of the many languages available to us.

Domain Knowledge

It focuses mainly on having knowledge about the particular topic you are working on.

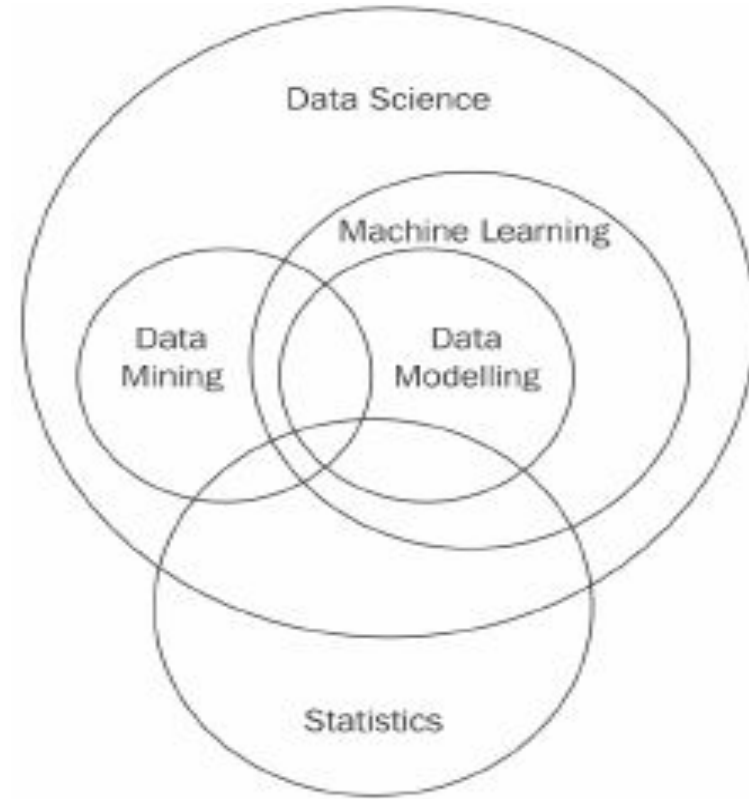
For example, if you are a financial analyst working on stock market data, you have a lot of domain knowledge.

Does that mean that if you're not a doctor, you can't work with medical data?

Data scientists can apply their skills to any area, even if they aren't fluent in it. Data scientists can adapt to the field and contribute meaningfully when their analysis is complete.

The state of data science

- Data mining is the part of data science where we try to find relationships between variables.
- Machine learning refers to giving computers the ability to learn from data without explicit "rules" being given by a programmer.
- Statistics is a mathematical body of science that pertains to the collection, analysis, interpretation or explanation, and presentation of data.
- A probability model is a mathematical representation of a random phenomenon. It is defined by its sample space, events within the sample space, and probabilities associated with each event.
- Probability models deals with predicting the likelihood of future events, while statistical models involves the analysis of the frequency of past events.



Data versus Big Data

- “Big data is the data characterized by three attributes volume velocity and variety.” -----IBM
- “Big data is the data characterized by four attribute volume velocity variety and value.”-----Oracle
- “Big Data is the frontier of a firm’s ability to store, process, and access all the data it needs to operate effectively, make decisions, reduce risks, and serve customers.” --- Forrester
- “Big Data in general is defined as high volume, velocity and variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.” -- Gartner
- “Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or doesn't fit the structures of your database architectures. To gain value from this data, you must choose an alternative way to process it.” -- O'Reilly
- **Big Data is data that is too large to be processed by a single machine.**

Modeling

Data model - representation of data by the database managers.

Architectural model - Architects capture attributes of buildings through blueprints and three-dimensional models.

Biological model - biologists capture protein structure with three-dimensional visualizations of the connections between amino acids.

Statistical model - Mathematical representation of observed data. Statisticians and data scientists capture the uncertainty and randomness of data-generating processes with mathematical functions that express the shape and structure of the data itself.

A model is our attempt to understand and represent the nature of reality.

Statistical modeling

- Probability distributions are the foundation of statistical models.
- Probability distribution is the mathematical function that gives the probabilities of occurrence of different possible outcomes for an experiment.
- An experiment is any procedure that can be infinitely repeated and has a well-defined set of possible outcomes known as sample space.
- Example: Flip a coin twice. The sample space is $\{(H, T), (T, H), (T, T), (H, H)\}$ where "H" means "heads" and "T" means "tails". Note that each of (H, T), (T, H), ... are possible outcomes of the experiment.
- A significant part of Data science is about understanding the behaviours and properties of variables, and this is not possible without knowing what distributions they belong to.

Fitting a model

Fitting a model means that you estimate the parameters of the model using the observed data.

You are making your algorithm learn the relationship between predictors and outcome so that you can predict the future values of the outcome.

Once you fit the model, you actually can write it as $y = 7.2 + 4.5x$, for example, which means this equation expresses the relationship between your two variables.

Overfitting is the term used to mean that you used a dataset to estimate the parameters of your model, but your model isn't that good at capturing reality beyond your sampled data. This happens when the sampled data size is too small.

[Example](#)

Exploratory data analysis (EDA)

EDA is a method of systematically going through the data, plotting distributions of all variables (using box plots), plotting time series of data, transforming variables, looking at all pairwise relationships between variables using scatterplot matrices, and generating summary statistics for all of them.

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques.

Employee Dataset: [Click here](#)

It contains 8 columns namely – First Name, Gender, Start Date, Last Login, Salary, Bonus%, Senior Management, and Team.

[EDA in python](#)

EDA: Exercise

Find the below dataset

https://docs.google.com/spreadsheets/d/1R6yUd64DDcuDzB3evLmrDVPbDHKT23P4tlzfclCR_M/edit?usp=sharing

Survival of patients who had undergone surgery for breast cancer.

Attributes:

1. Age of the patient
2. Patients year of operation
3. Number of lymph nodes detected
4. Survival status =1 (the patient survived 5 years or longer)
2 (the patient died within 5 years)

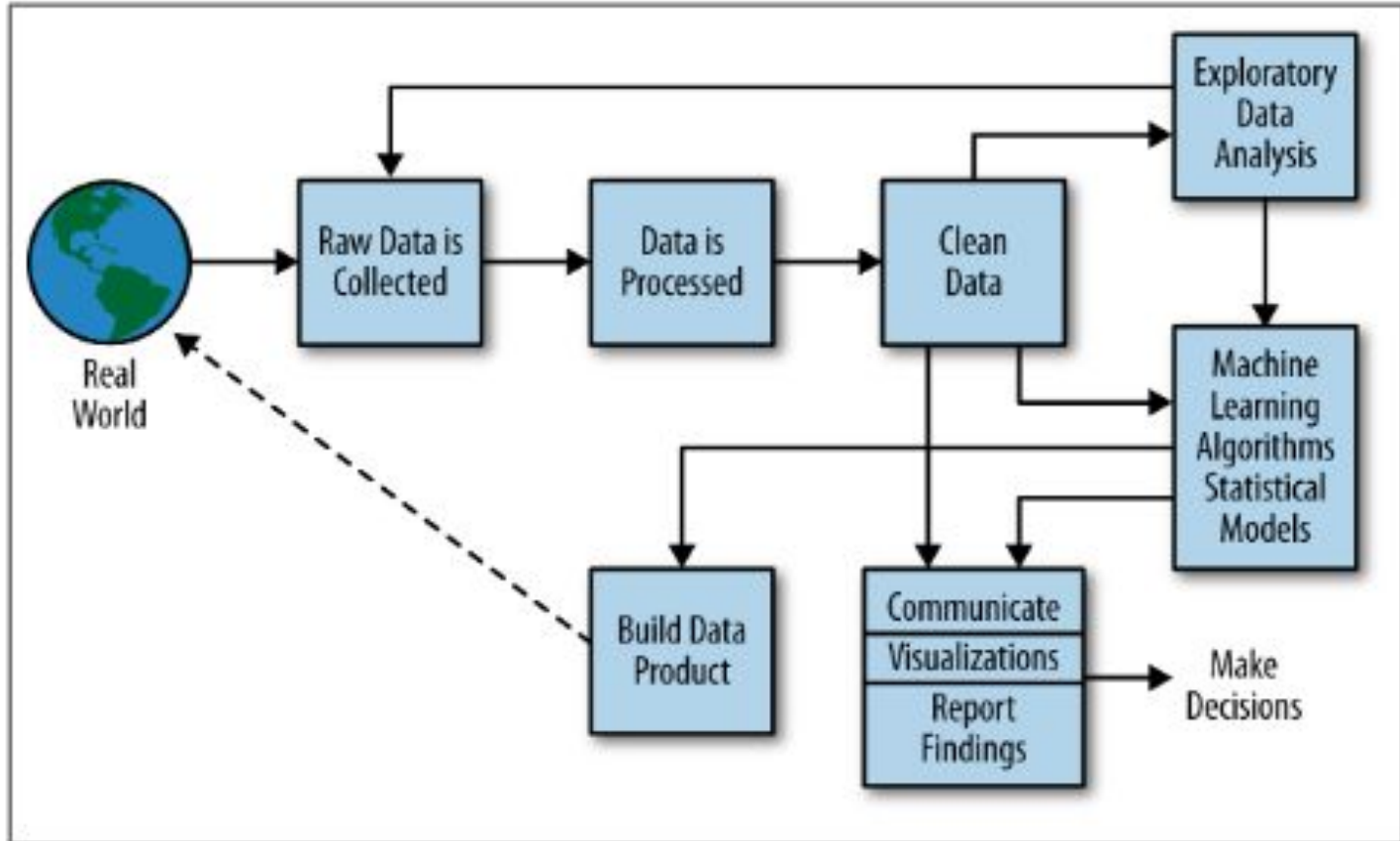
EDA: Exercise

1. What is the range of patient age?
2. The operation is performed between the years ?
3. The maximum number of lymph nodes found ?
4. Do you have any missing values?
5. Most of the operations are performed in which year?
6. Is the data balanced or not?
7. Identify whether outlier exists?

EDA: Exercise

1. What is the range of patient age? 30-83 years
2. The operation is performed between the years ? 1958-1969
3. The maximum number of lymph nodes found ? 52
4. Do you have any missing values? No
5. Most of the operations are performed in which year? 1958
6. Is the data balanced or not? Imbalanced

Data science process



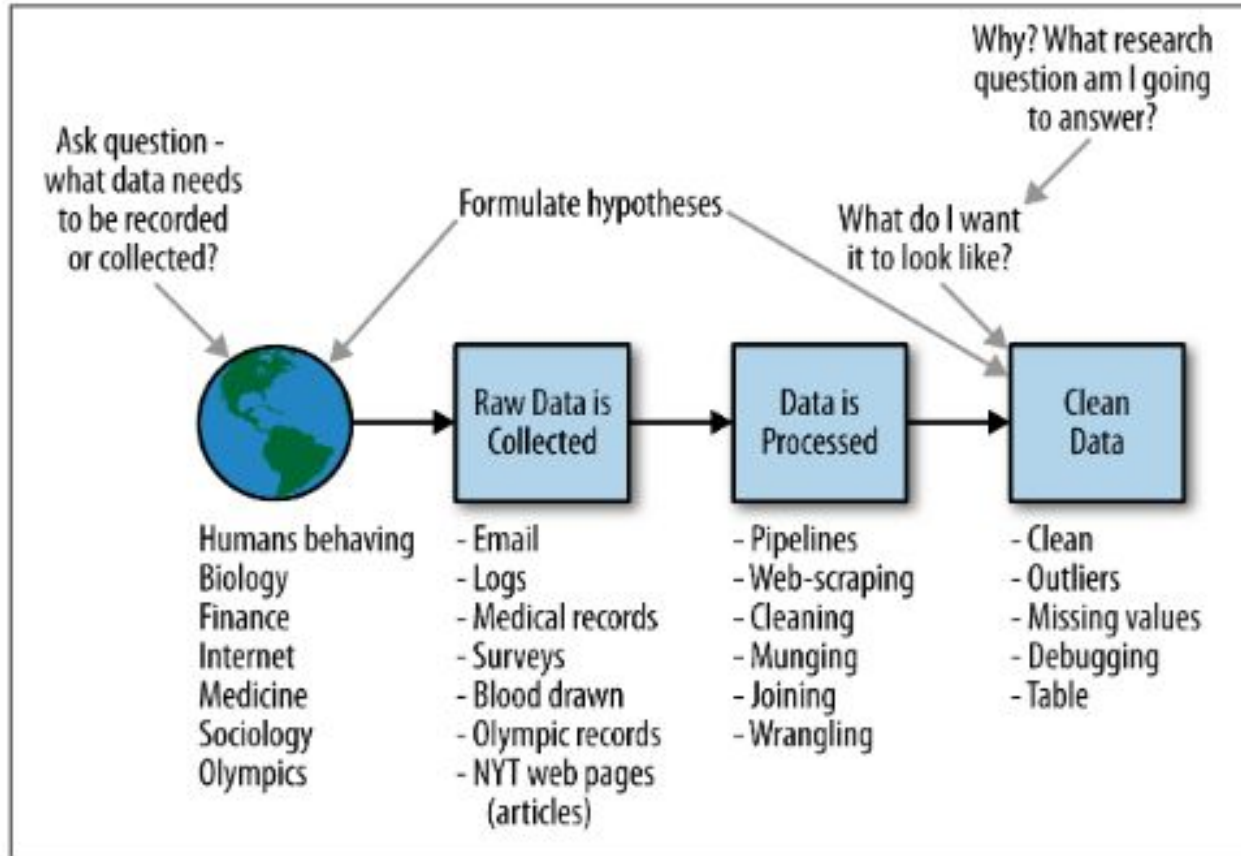
Data science process

- Collect the raw data and process it to make clean for analysis.
- Data cleaning and munging/wrangling refers to removing the erroneous data and transforming the data into a usable form.
- To do cleaning and munging, the tools such as Python and R can be used.
- Once we have this clean dataset, we should be doing some kind of EDA.
- In the course of doing EDA, we may realize that it isn't actually clean because of duplicates, missing values and outliers.
- Now, go back to collect more data, or spend more time cleaning the dataset.
- Design a machine learning model, and the model depends on the type of the problem that we are trying to solve.

Data science process

- Finally communicate the findings.
- Alternatively, our goal may be to build or prototype a “data product”; e.g., a spam classifier, or a search ranking algorithm, or a recommendation system.
- Data product gets incorporated back into the real world, and users interact with that product, and that generates more data, which creates a feedback loop.
- Take this loop into account in any analysis you do by adjusting for any biases your model caused.
- This makes data science different from the statistics.

A Data scientist role in this process



A Data scientist role in this process

- Ask a question.
- Do background research.
- Construct a hypothesis.
- Test your hypothesis by doing an experiment.
- Analyze your data and draw a conclusion.
- Communicate your results.