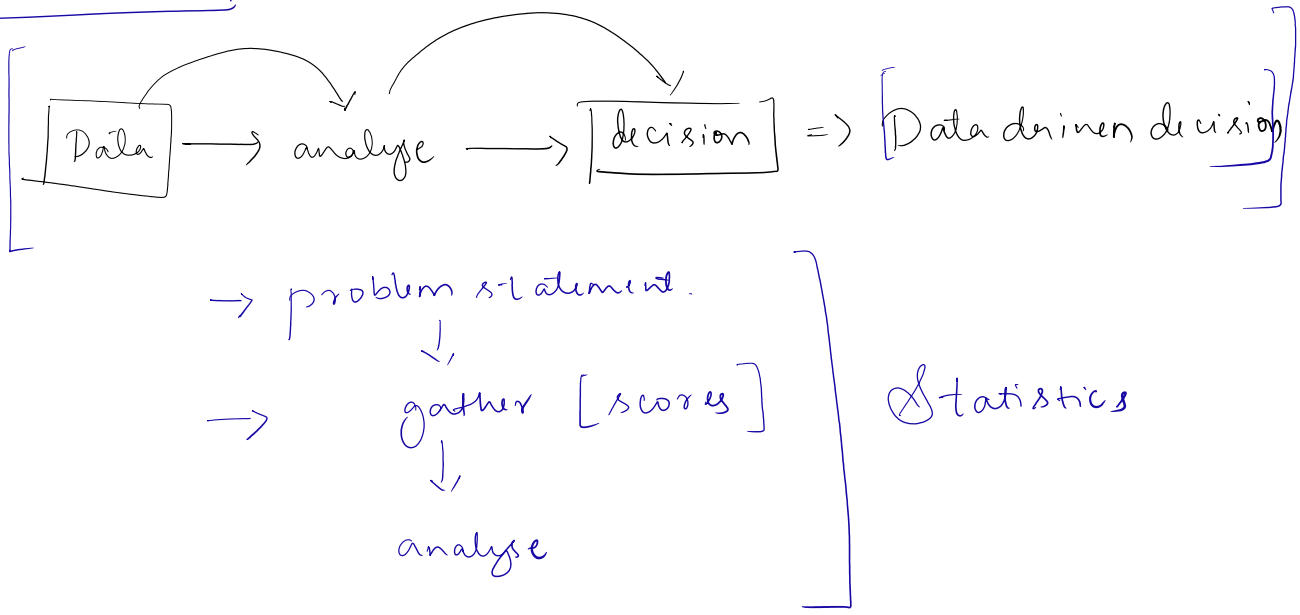
17 February 2025 20:38

Statistics :



STATISTICS

Statistics is a field of science concerned with collecting , analyzing , interpreting and presenting data.

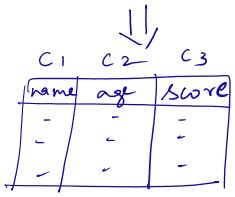
Eg: If you wish to compute average height , weight or grade points of a students at college. We collect sample from the data and we analyse.
Eg2:In presidential election , opinion poll samples of 1000-2000 people are taken. The opinion poll represent view of the people in the country.

Datamites

TYPES OF STATISTICS

- DESCRIPTIVE STATISTICS
- INFERENTIAL STATISTICS

→ 1. describe / description
→ summary stats
→ Graphs
→ Tables



Σ avg

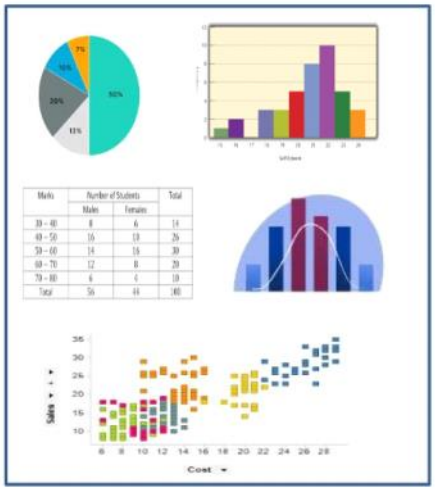
DESCRIPTIVE STATISTICS

Descriptive statistics aims to describe chunk of raw data using summary statistics ,graphs and tables. ∈

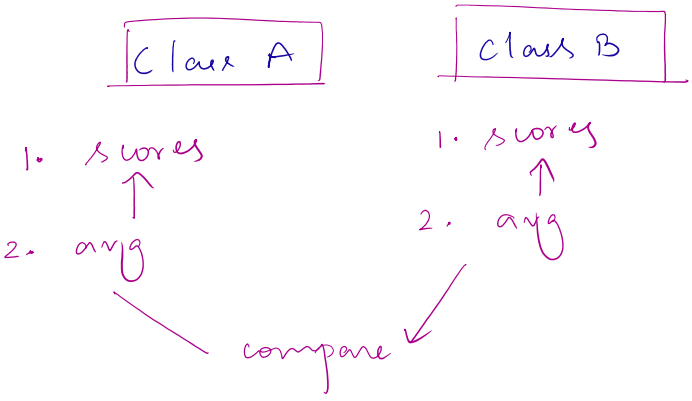
Eg: Suppose let us say we have raw data shows information about test scores of students at a particular school.

How do we analyze?

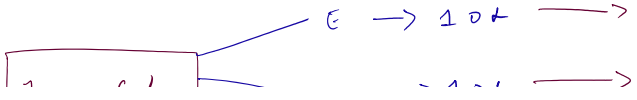
We use descriptive statistics to find the average score and create a graph that helps us to visualize the distribution of scores

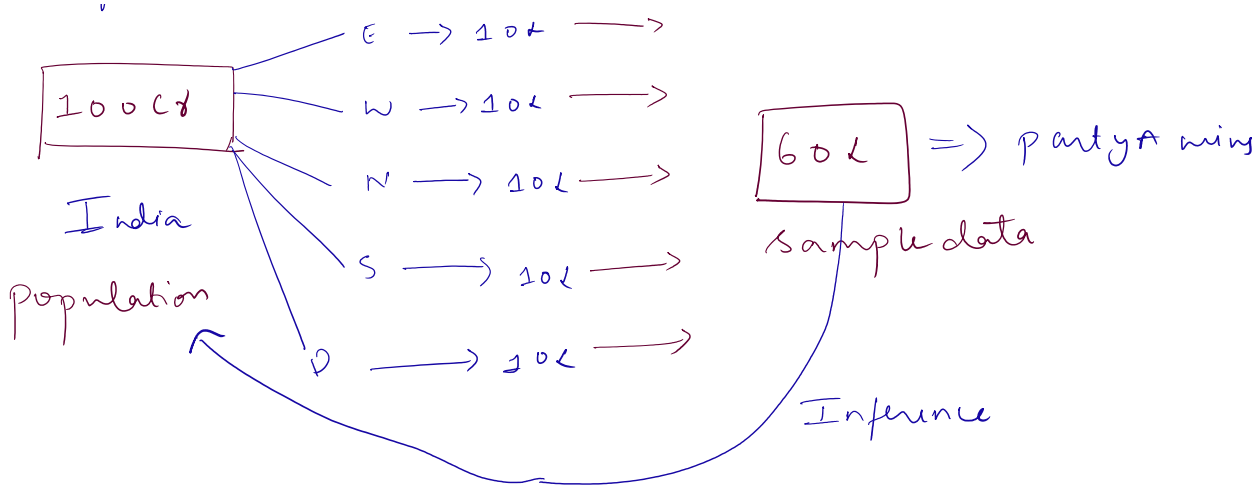


Datamites

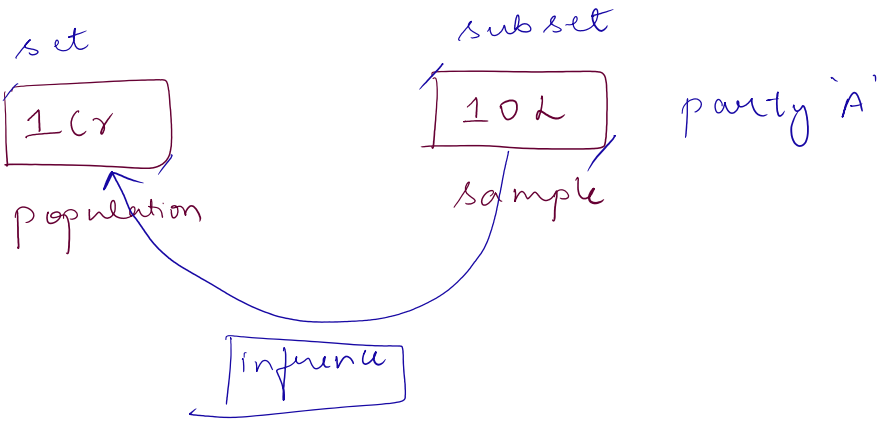


2. Inferential statistics





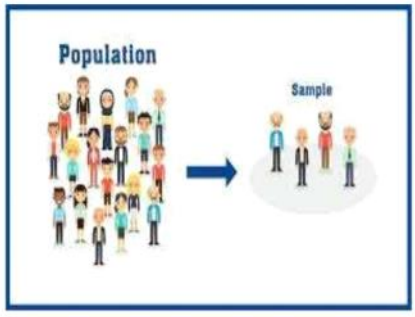
Inferential Stats



INFERENTIAL STATISTICS

Inferential statistics uses a sample of data from a population to draw inference/conclusion about the larger population.

Eg: We might be interested in understanding political preference of millions of people in country. However it would be expensive to actually survey every individual in the country. Thus, we take smaller survey of say 1000 people ,then use the results of samples to draw inference about the population as whole.



Datamites

BASIC TERMINOLOGIES

set

POPULATION

Every possible individual element that we are interested in measuring.

subset

SAMPLE

A portion /subset of population

Datamites

BASIC TERMINOLOGIES

=> PARAMETER

A parameter is a number that describes some characteristics of a population.

Eg: Let us say we need to measure mean height of all the trees in India

STATISTIC <=

A statistic is a number that describes some characteristics of sample.

Eg: Let us say we need to measure mean height of 1000 trees in India

mean

=> [1 2 3 4 5 6 7 8 9 10] → pop

→ with replacems → s1 (1 2 3), s2 (3 4 5)

→ without replacement → [(1, 2, 3)] [(4 5 6)]

s1 mean, s2 mean

Experiment

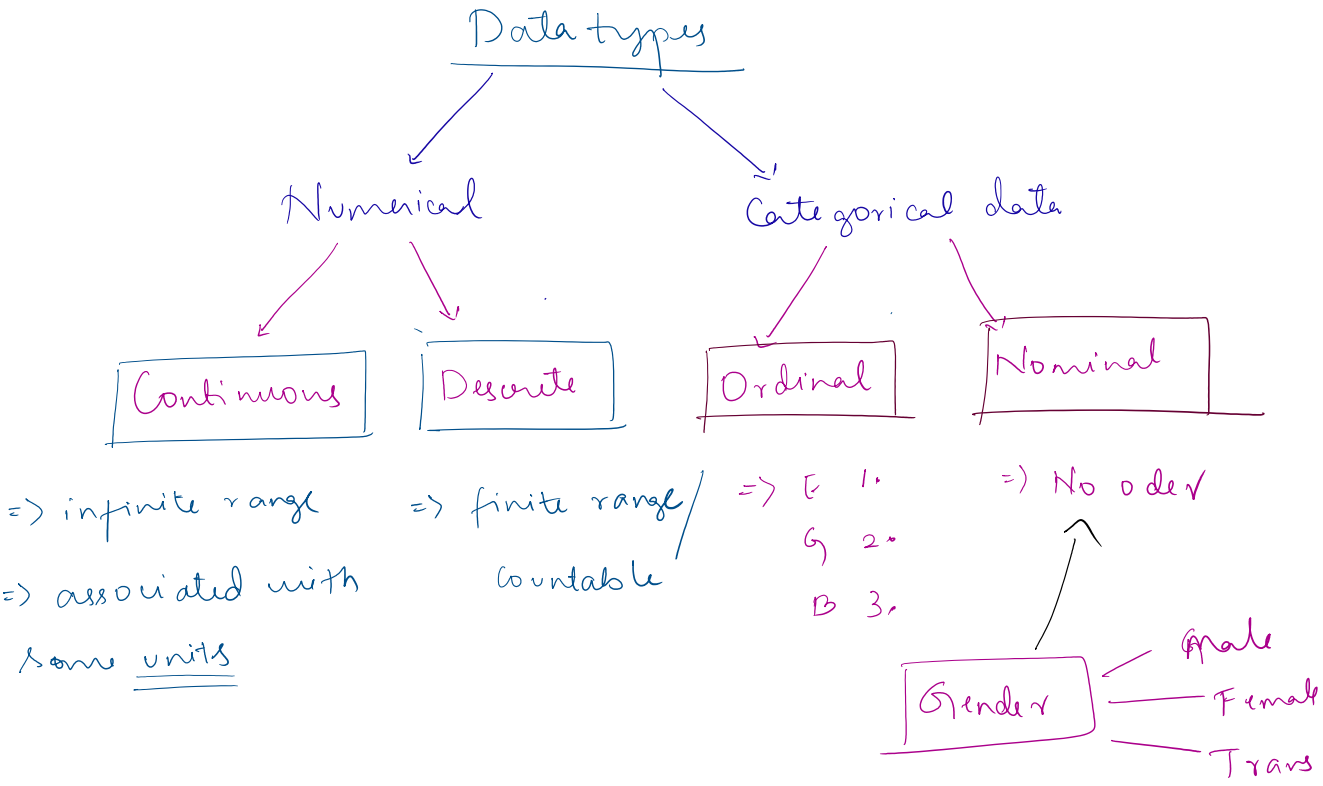
A planned activity whose results yield a set of data.

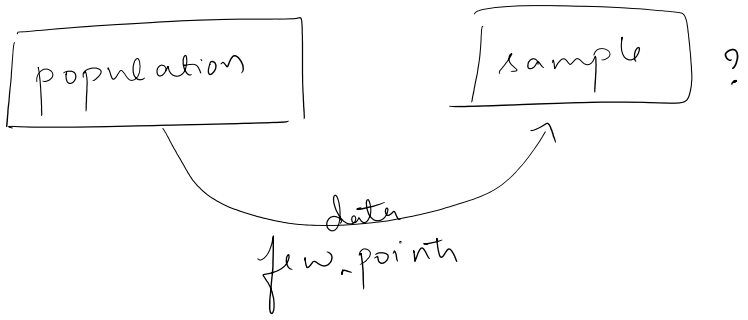
Variable

A variable is any characteristics, number or quantity that can be measured or counted.

variable

=> The avg age of class A students is 18


Data types
Numerical
Categorical data
Continuous
Discrete
Ordinal
Nominal
=> infinite range
=> associated with some units
=> finite range / countable
=> No order
Gender
Female
Trans


population
sample
?
data
few points

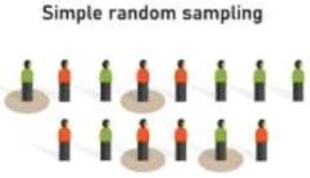
SAMPLING METHODS

- Simple random sample
- Stratified random sample
- Cluster random sample
- Systematic random sample
- Convenience sample

1. Simple Random sampling

SIMPLE RANDOM SAMPLE

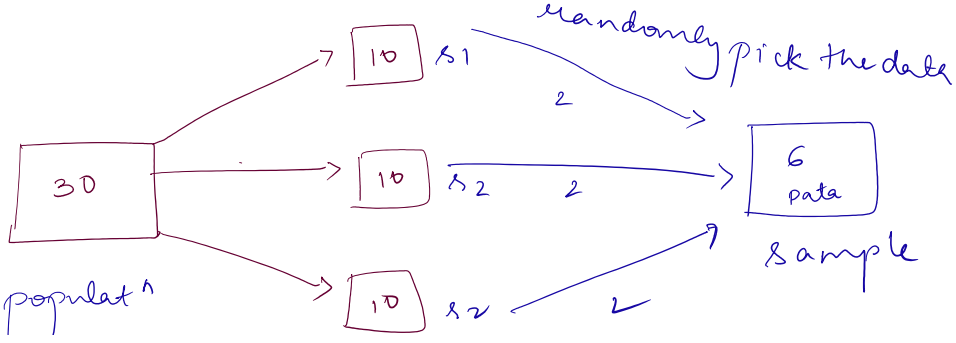
Every member of a population has an equal chance of being selected to be in a sample. Randomly selecting members by means of random selection.

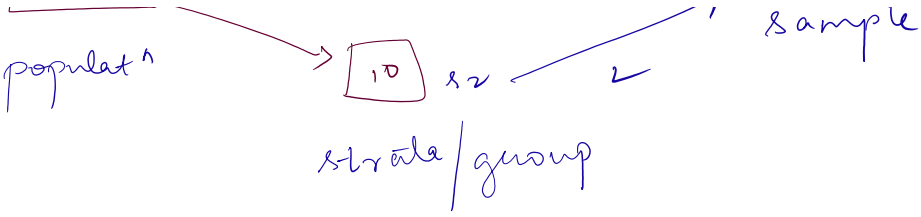


pop 60 randomly sample 10

Stratified RS

sample = 6

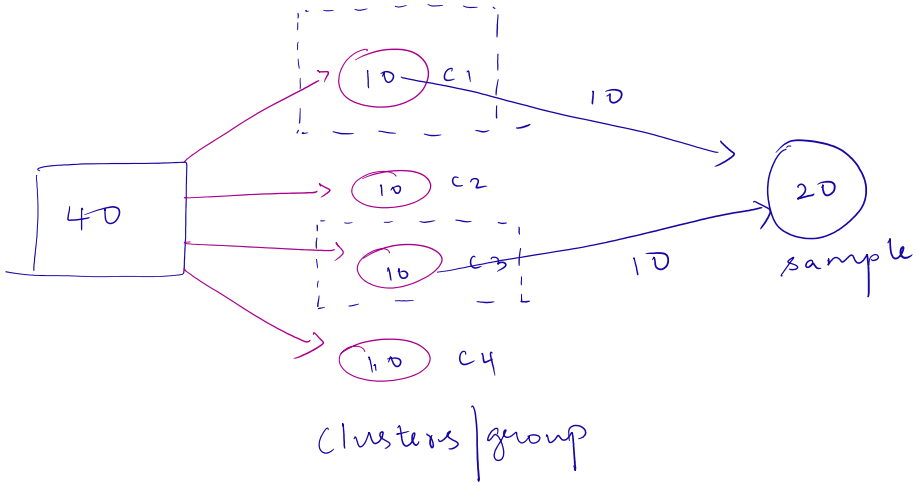




STRATIFIED RANDOM SAMPLE

Split entire population into groups. Randomly select some members from each group to be in a sample.

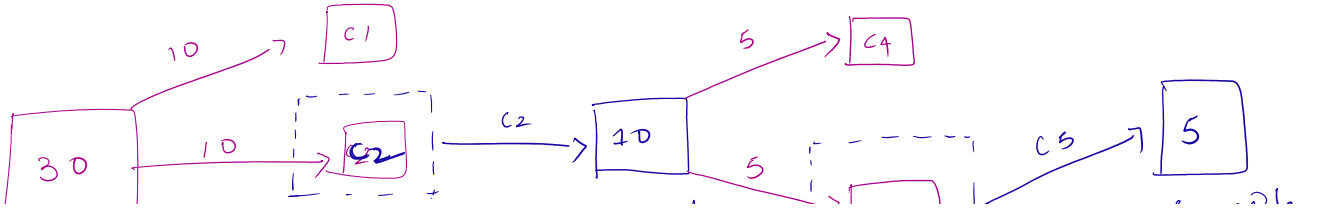
Cluster RS :-

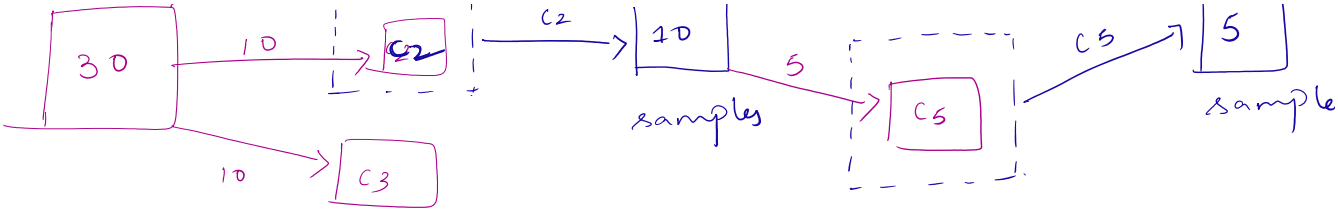


Multistage RS :-

=> Performing cluster RS in multiple stages

sample : 5





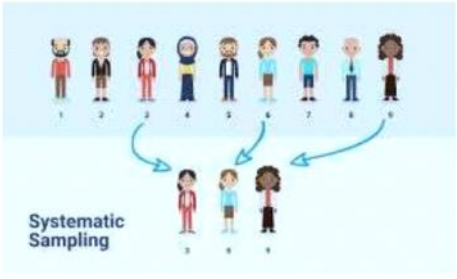
10 point in each

SYSTEMATIC RANDOM SAMPLE

⇒ 1 2 3 4 5 6 7 8 9 10

2 4 6 8 10

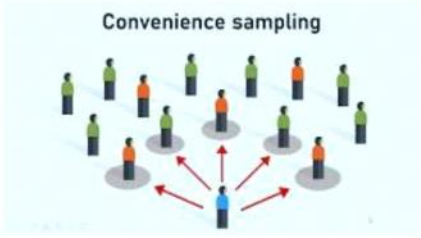
⇒ Put every member of a population into some order. Choosing random starting point and select every nth member to be in the sample.



Eg:
A teacher puts students in alphabetical order according to their last name, randomly chooses a starting point and picks every 3rd student to be in the sample.

CONVENIENCE SAMPLE

Choose members of a population that are conveniently or readily available to be included in the sample.



Eg:
A researcher stands in front of a library during the day and polls people that happen to walk by.

Originally, the Pepsi Challenge was a blind taste test conducted at shopping malls, stores, and other public venues. Participants taste unmarked cups containing Coca-Cola and Pepsi and then indicate their preference.

Datamites

MEASURE OF CENTRAL TENDENCY

A **measure of central tendency** is single value that represents center point of a dataset. This is referred to as " **the central location**" of the dataset.

Three types of Measure of central tendency

- Mean
- Median
- Mode

$\Rightarrow$ 1 2 3 4 5 $= \frac{1+2+3+4+5}{5} = \frac{15}{5} = 3$

MEAN

Mean is the average of the numbers or the data

How to find mean ?
It is calculated by summing all the numbers and dividing it by total number of values.

population mean, $\mu = \frac{\sum X_i}{N}$

sample mean $\bar{x} = \frac{\sum x_i}{N}$

$\Rightarrow$ the summary of the entire data.

$\Rightarrow$ 2 5 6 8 1 4 3

MEDIAN

Median is the middle value in the group of values when values are arranged in either ascending or descending order .

Note:

- It is a number that separates ordered data into halves.
- Half the values are the same number or smaller than the median, and half the values are the same number or larger.

1) 1 2 3 **4** 5 6 8 — $N = 7$ (odd)

lower: 1 2 3 | upper: 5 6 8

median : 4

2) 1 2 3 4 5 6 8 10 — $N = 8$ (even)

median : $\frac{4+5}{2} = 4.5$

$\Rightarrow$ data into 2 equal halves!

MODE

Mode is defined as the most frequently occurring, or repetitive value in a dataset. Use this when working with categorical data.

Eg: You conduct survey about people favorite colors and you want to know which color occurs most frequently.

1] 1 2 3 4 => no mode

2] 1 1 2 3 4 => 1

3] 1 1 2 2 3 4 5 =>

1, 2

Bimodal

4] 1 1 2 2 3 3 4 5

1, 2, 3 => multimodal

Numerical → mean
→ median

Categorical → mode

WHEN TO USE MEAN, MEDIAN AND MODE?

MEAN
Used to find the average value in a dataset.
MEDIAN
Used to find the middle value in a dataset.
MODE
Used to find the most frequently occurring value in a dataset.

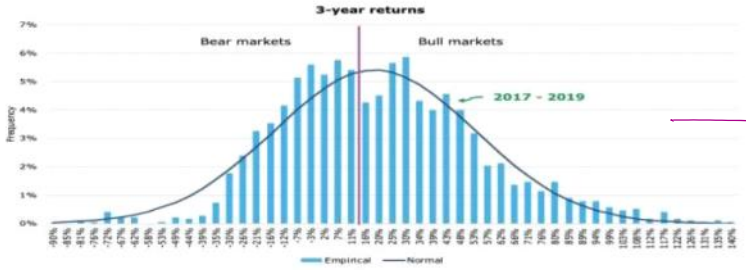
[1 2 3 4 5 6 7 8 9 10] → A Normal data

[1 2 3 4 5 6 7 8 9 10, 1000, 2000, -100] → B

DISTRIBUTIONS

A distribution is simply a collection of data, or scores, on a variable. Usually, these scores are arranged in order from smallest to largest and then they can be presented graphically.

The graphical representation of all observations is known as distribution

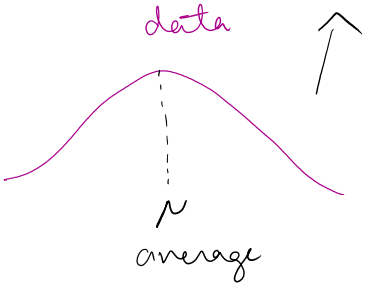


→ histogram

Datamites

Data ⇒ particular range ⇒ Normally distributed data

Data → histogram ⇒



MEASURE OF DISPERSION / Data Variability

How "**Spread out**" the values are. We measure "**spread**" using measure of dispersion such as

- **Range**
- **Variance**
- **Standard deviation**
- **Interquartile range**

Datamites

RANGE

Range in statistics is the difference between the highest and lowest values.

Range = max - min

Piyush check your connection!

John take 7 statistics tests over the course of a semester and scores are 94,88,73,84,91,87 and 79. What is the range of scores?

Range = 94 - 73

Range = 21

38 - 20

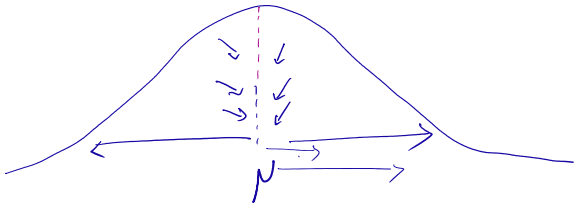
18

Dataset 1	Dataset 2
20	11
21	16
22	19
25	23
26	25
29	32
33	39
34	46
38	52

Datam

VARIANCE

Variance measures how far the data is spread out from the mean. More the value of variance , the data is more scattered from its mean and if the variance is low, then it is less scattered from mean.



Data → var() => low value

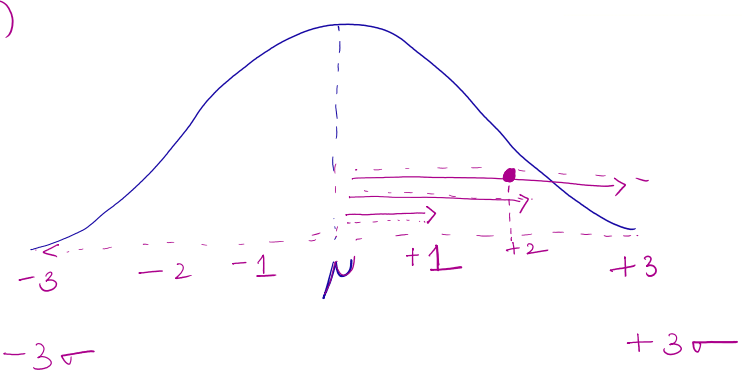
=> all the datapoints are near to the mean.

Data → var → high => abnormal

STANDARD DEVIATION

Standard Deviation (SD) is a measure that is used to quantify the amount of variation or dispersion of a set of data values. It is square root of variance.

The standard deviation is small when data are all concentrated close to the mean, which exhibits little variation or spread. If standard deviation is high then data has more spread out from the mean, exhibiting more variation.



$N = 5$ $x = 17$

1. if the data is normal, then the whole data is spread in the range of $(-3, +3)$

Variance : $\sigma^2 = \frac{\sum (x-\mu)^2}{N}$ $[1, 2, 3, 4, 5]$ $\mu = 3$

$$= \frac{\sum (1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2}{5}$$

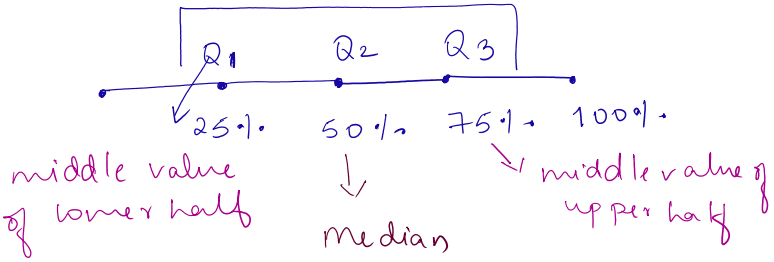
$\sigma^2 =$ ☐

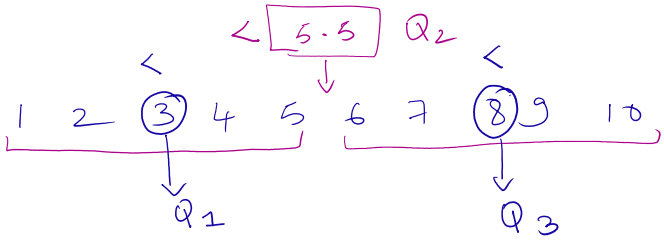
$$Std() = \sqrt{\text{variance}} = \sqrt{\frac{(x-\mu)^2}{N}} = \sigma$$

QUARTILES

Quartiles are the values that divide entire data into quarters and each occupies $\frac{1}{4}^{th}$ of the data.

NOTE:
Q1 → lower quartile
25% of the data below the lower quartile
Q2 → middle quartile(median)
50% of the data lie below it and other 50% lie above median value. Q3 → upper quartile
75% of the data below it and other 25% of the data above upper quartile





$Q_2 : 5.5$

$Q_1 : 3$

PERCENTILES

$Q_1 \rightarrow 25^{th}$ $Q_3 \rightarrow 75^{th}$

$Q_2 \rightarrow 50^{th}$

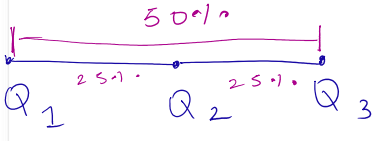
Percentile is defined as the value below which a given percentage falls under. Percentiles are the values that separate the data into 100 parts.

Q1 is same as 25th percentile
Q2 is same as 50th percentile
Q3 is same as 75th percentile

Eg: Universities and colleges use percentiles when SAT/GRE results are used to determine minimum testing score that will be used as an acceptance factor.

IQR Inter Quartile range

IQR is used to measures the spread of the middle half of the data. It is the range for the middle 50% of the data which falls between Q1 and Q3. It is the best measure to know the spread of data when data a skewed.
Note:
Used to identify outliers and to compare distributions of two datasets.



IQR is the difference the upper and lower quartile

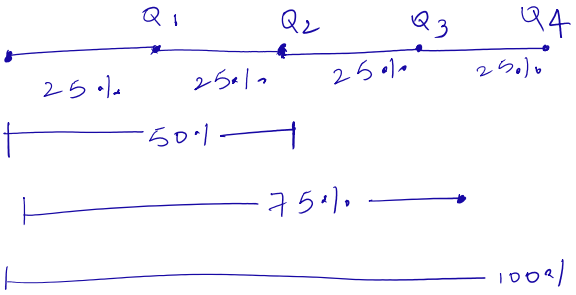
$$IQR = Q3 - Q1$$

Datamites

IQR => high => middle 50% => very scattered

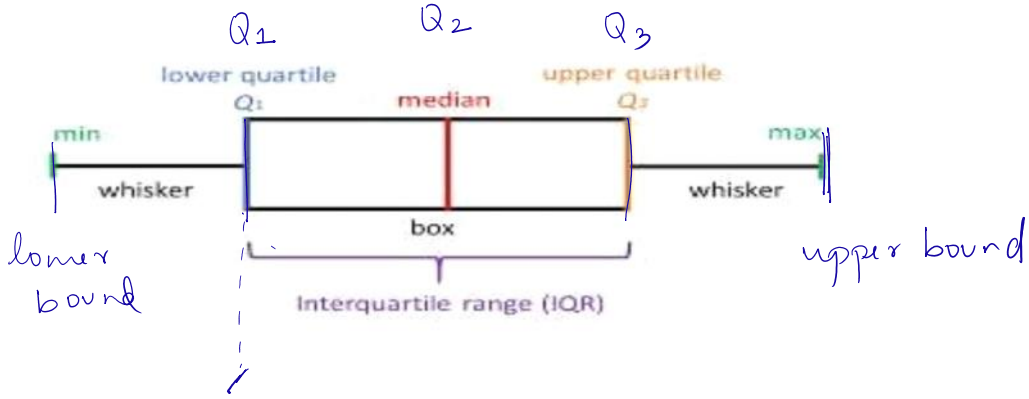
IQR => high => middle 50% => very scattered

=> low => ——"—— => less ——"——



IQR USING BOXPLOT

=> normal / Not



☆ Outliers

=> age: [20 22 26 28 29 30 120 200]

outliers

=> an observation i.e falling far away from the original set of data

=> abnormal value!

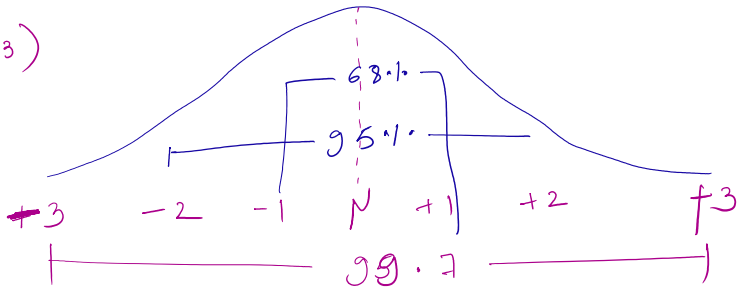
=> Outlier is not good for our analysis

=) Outlier is not good for our analysis

→ Handle — drop
→ Handle — replace with mean / median

If the data is normal

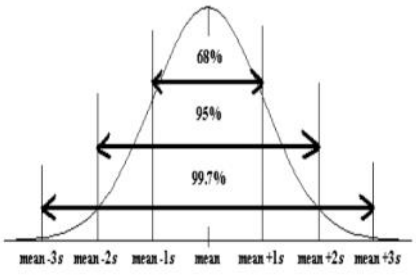
99.7% => (-3 +3)



EMPERICAL RULE

The empirical rule states that for a normal distribution, nearly all of the data will fall within three standard deviations of the mean. The empirical rule can be broken down into three parts:

- About 68% of the x values lie between -1σ and +1σ of the mean μ. (one standard deviation)
- About 95% of the x values lie between -2σ and +2σ of the mean μ. (two standard deviation)
- About 99.7% of the x values lie between -3σ and +3σ of the mean μ. (three standard deviation, note that entire data lie within this range)



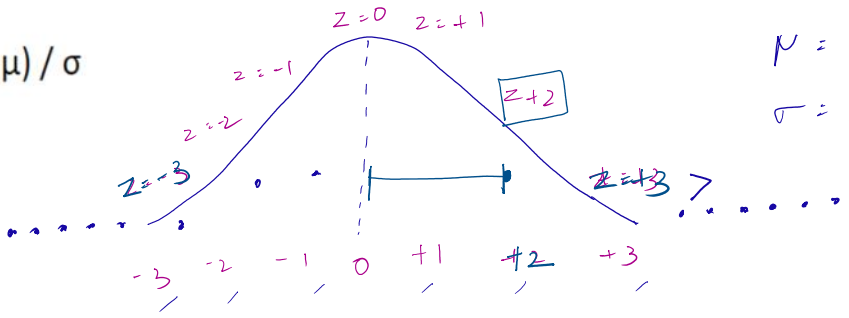
Datamites

Z- Score / Z-Value/ Standard Score

Datamit
Global Institute for Data

- A z-score (aka, a standard score) indicates how many standard deviations an element is above or below from the mean. A z-score can be calculated from the following formula.

- $z = (X - \mu) / \sigma$



$\mu = 5$

$\sigma = 6$

$x = 17$

$= \frac{17-5}{6}$

$= \frac{12}{6}$

$z = +2$

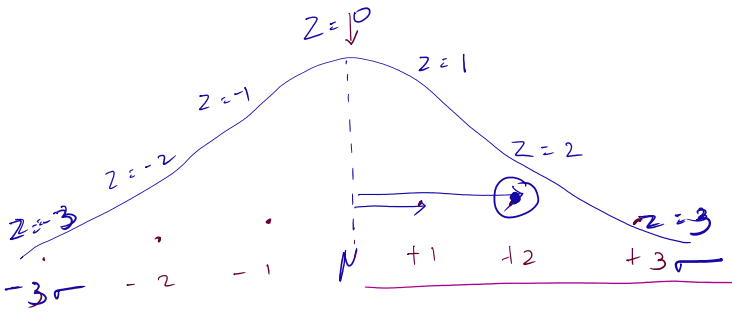
$\therefore$ the input data $x = 17$ is at $+2\sigma$ from the mean

range : $(-3, +3)$ i.e z value for an x < -3 or $> +3$ $\Rightarrow$ out liers

Outliers

Datamites
Global Institute for Data Science

- - Z-scores generally range from -3.0 to +3.0.
- - For bell shaped distributions, the empirical rule says 99.7% of all the data values have z-scores between -3.0 and +3.0.
- - We consider any z-score that is either less than -3.0 or greater than +3.0 to be an **outlier**.



x $\quad x = 17$, $N = 5$, $\sigma = 6$

$$z\ score = \frac{x - \mu}{\sigma}$$

$$\boxed{z\ score = +2}$$

[$\therefore$ $x = 17$ is at the $+2\sigma$ from mean]

$\therefore$ the z score $(-3, +3)$

x —— z score $\rightarrow +3$

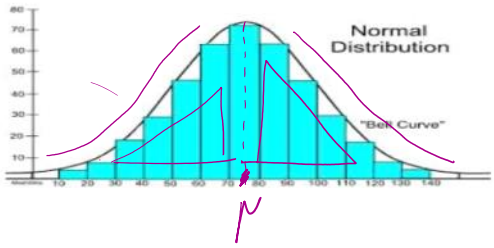
$\rightarrow < -3$

$x \Rightarrow z\ score = 4.2 \Rightarrow$ outlier

$\longrightarrow$ normal distribution

NORMAL DISTRIBUTION

Normal distribution, also known as the Gaussian distribution, is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean.



PROPERTIES

mean ≈ median for a normally / without outliers

1. The mean, mode and median are all equal.
2. The curve is symmetric at the center (i.e. around the mean, μ).
3. Exactly half of the values are to the left of center and exactly half the values are to the right.
- The total area under the curve is 1.

are almost similar

1 2 (3) 4 5 median, Q_2 = 3

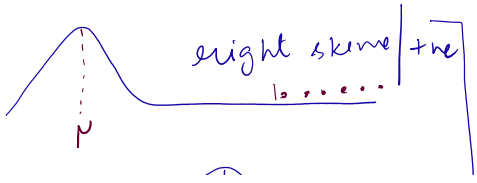
mean = 3

STANDARD NORMAL DISTRIBUTION

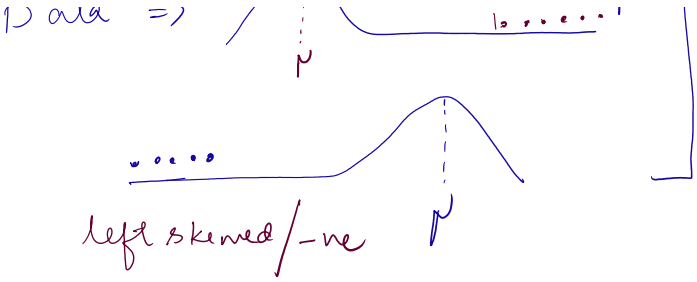
The standard normal distribution is a special case of the normal distribution. It is the distribution that occurs when a normal random variable has a mean of zero and a standard deviation of one.

Standard normal distribution : $\mu = 0$ *and* $\sigma = 1$

=> mean = 0, std() = 1 => std normal distribut^n

Data => 

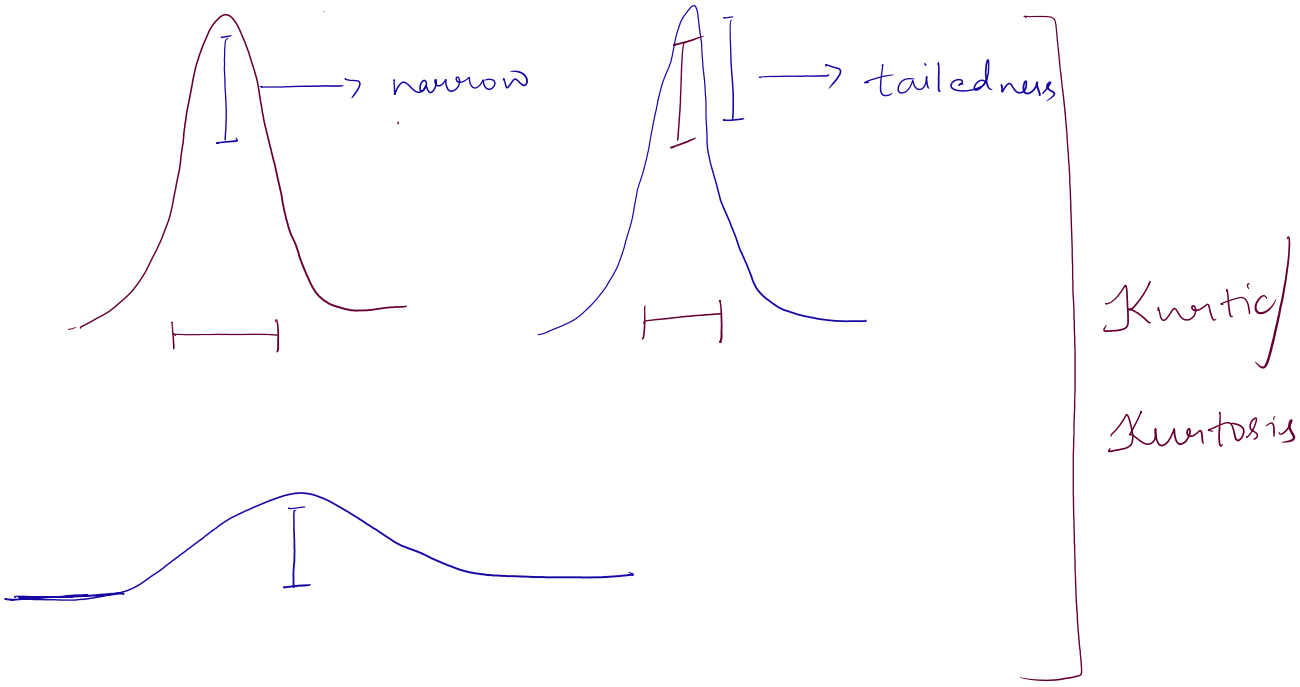
skewed distribution



=> If we have skewed data, the probability of outliers will be high

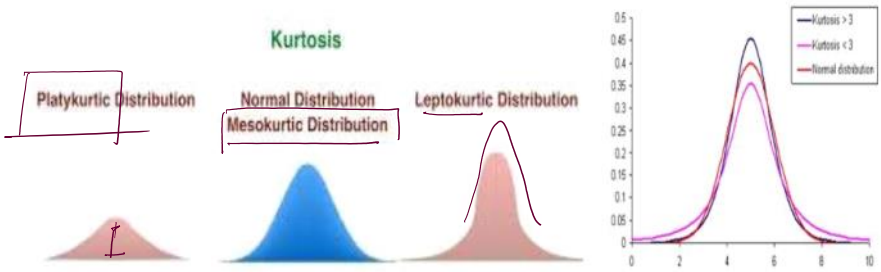
Why Skewness?

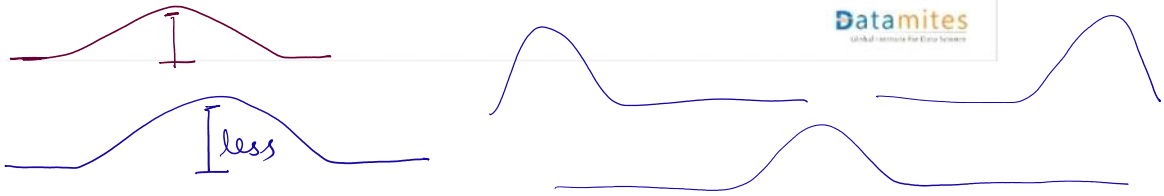
Skewness tells us the direction of outliers. In a positive skew, the tail of a distribution curve is longer on the right side. This means the outliers of the distribution curve are further out towards the right and closer to the mean on the left. Skewness does not inform on the number of outliers; it only communicates the direction of outliers.



KURTOSIS

In probability theory and statistics, kurtosis is a measure of the "tailedness" of the probability distribution of a real-valued random variable.





SKEWNESS AND KURTOSIS RANGE

- If the skewness is between -0.5 and 0.5, the data are fairly symmetrical.If the skewness is between -1 and – 0.5 or between 0.5 and 1, the data is moderately skewed.
- If the skewness is greater than 1or less than -1, the data is highly skewed.
- A standard normal distribution has kurtosis of 3 and is recognized as mesokurtic. An increased kurtosis (>3) can be visualized as a thin "bell" with a high peak whereas a decreased kurtosis corresponds to a broadening of the peak and "thickening" of the tails.

$(-1, +1)$

1. $(-0.5, 0.5)$ -> normally distributed

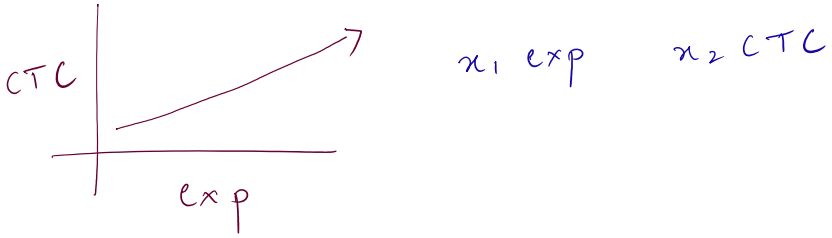
2. $(-1, -0.5)$ => left

3. $(0.5, +1)$ => right

$K = 3$, normally meso

$K > 3$ lepto

$K < 3$ meso



COVARIANCE

Covariance is a measure of the relationship between two random variables. The metric evaluates how much – to what extent – the variables change together. In other words, it is essentially a measure of the variance between two variables.

⇒ x_1 x_2 ⟶ if the 2 variables are related or not

$$cov(X,Y) = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{N}$$

X, Y = population mean

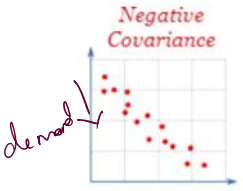
n = no of observations in sample

$\bar{x}$: mean of variable x

$\bar{y}$: —"— variable y

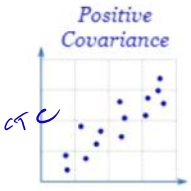
N : total sample

- **Positive covariance:** Indicates that two variables tend to move in the same direction.
- **Negative covariance:** Reveals that two variables tend to move in inverse directions.
- **Zero covariance :** Two variables are independent.

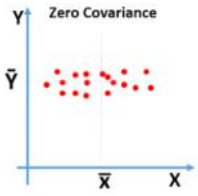


demand ↓

price ↑'s



exp



Datami

CORRELATION

(−1, 1)

Correlation is used to test relationships between quantitative variables. In other words, it's a measure of how things are related. The study of how variables are correlated is called **correlation analysis.**

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \sum(y-\bar{y})^2}}$$

1. x_1 x_2 ⇒ corr() ⇒ 0.9

90% similarity

⇒ strongly related

2. x_1 x_2 ⇒ corr() ⇒ 0.4

40% similar, weakly

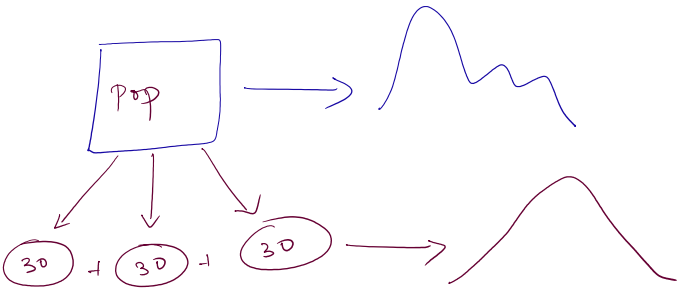
−0.8

CENTRAL LIMIT THEOREM

The central limit theorem states that the distribution of sample means approximates a normal distribution as the sample size gets larger (assuming that all samples are identical in size), regardless of population distribution shape.

CLT in one sentence "Even if I'm not normal, the average is normal"

When collecting means of the samples from any distribution, the no of samples taken for calculating the mean should be greater or equal to 30.

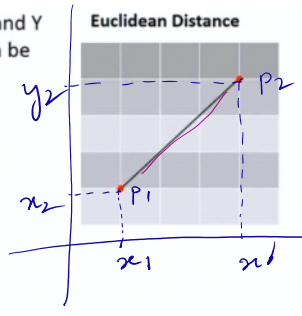


Even if populatn is not normal, the avg of samples => normally distributed

Similarity of data points is calculated by distance

Euclidean Distance

It is a classical method to calculate the distance between two objects X and Y in the Euclidean space (1- or 2- or n- dimension space). This distance can be calculated by traveling along the line, connecting the points.

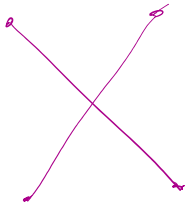


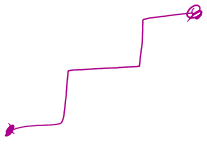
You can use the Pythagorean Theorem to compute this distance:

$$dist = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2}$$

$$dist = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Datamites

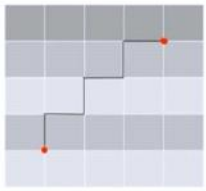




Manhattan Distance

It is similar to Euclidean Distance, but the distance (for example, two points, separated by building blocks in a city) is calculated by traversing vertical and horizontal lines in the grid-based system.

Manhattan Distance



You can use the following formula to compute this distance:

$$d_t = |x_2 - x_1| + |y_2 - y_1|$$

Datamites

Minkowski Distance

It is a metric on the Euclidean space and can be considered as a generalization of both the Euclidean and Manhattan distances.

You can use the following formula to compute this distance:

$$dist = \left(\sum_{k=1}^{n} |p_k - q_k|^r\right)^{\frac{1}{r}}$$

→ 1 → Manhattan ✓
→ 2 → Euclidean ✓
Task

When r =1; it computes the Manhattan distance.

When r =2; it computes the Euclidean distance.

When r =∞; it computes Supremum.

1/2 = √

case;- $r = 1$

$dist = \left(|p_k - q_k|^r\right)^{1/r}$

$= \left(|p_k - q_k|^1\right)^{1/1}$

$= |p_k - q_k|^1$

$= |p_k - q_k| \Rightarrow |x_2 - x_1| + |y_2 - y_1|$ ⇒ Manhattan distance

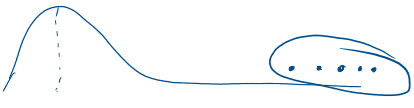
Case $r = 2$

$dist: \left(|p_k - q_k|^r\right)^{1/r}$

$= \left((p_k - q_k)^2\right)^{1/2}$

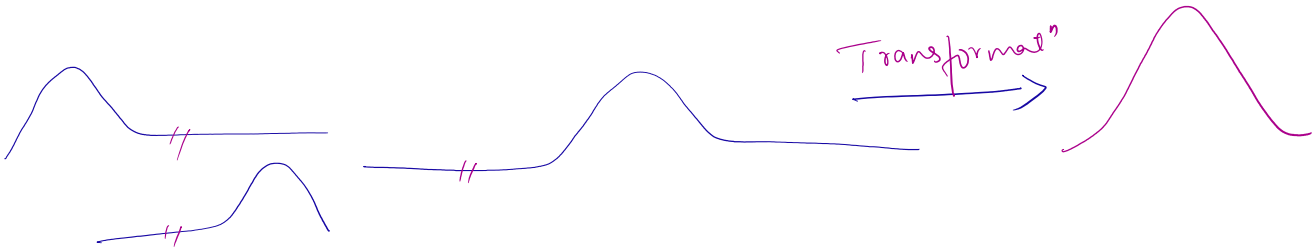
$= \sqrt{(p_k - q_k)^2} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$ Euclidean

★ Transformation



⇒ process where skewed data is converted into normally

=> process where skewed data is converted into normally distributed data (near to normal)



1. log
2. Exponential sqrt, cube
3. Reciprocal
4. Box - Cox