Introduction

12 August 2024 23:23

30 ball 60 run

Statistics— approx

Statistics: It is the branch of mathematics of conducting studies to collect, organize, summarize, analyze, and draw a conclusion out of data.

1 Over — 5 Run

1. It helps in making more effective decisions.
2. It is used in many domains like marketing, business, healthcare, sports, etc.
3. We use sats for { **Better Decision Making** }

For example : Sports Analytics: Statistics play a crucial role in sports, helping teams analyze player performance, evaluate strategies, and make informed decisions. For example, batting averages in cricket or shooting percentages in basketball are statistical measures that provide insights into a player's skill and performance.

For example: Suppose you run an online store and want to test whether sending a promotional campaign to certain customers makes them buy more. You decide to send the campaign to some customers and not to others. You then observe whether the customers who received the campaign bought more than those who didn't.

The key questions you need to answer are:

1. How do you choose which customers should receive the campaign?
2. How many customers should be included in the campaign?
3. After the campaign, if the customers who received it bought more, how do you know it's because of the campaign and not because they would have bought more anyway?

To solve this, you would design an experiment:

- **Randomly select** customers to receive the campaign (this helps ensure that any differences in purchasing behavior are due to the campaign and not some other factor).
- Compare the purchasing behavior of those who received the campaign with those who didn't.
- Use statistical methods to determine whether the increase in sales is significant and can be attributed to the campaign, or if it's likely those customers would have bought more regardless of the promotion.

This helps you understand if your campaign is truly effective or if the increased sales would have happened even without it.

100

20

TO ANSWER ALL THESE QUESTIONS WE WILL APPLY THE STATISTICS

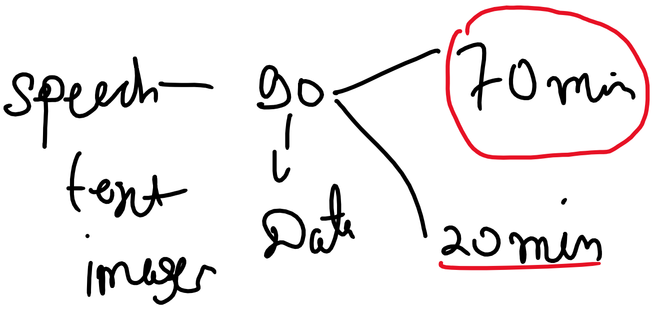
2000

500

1500

Data

12 August 2024 23:26



Data

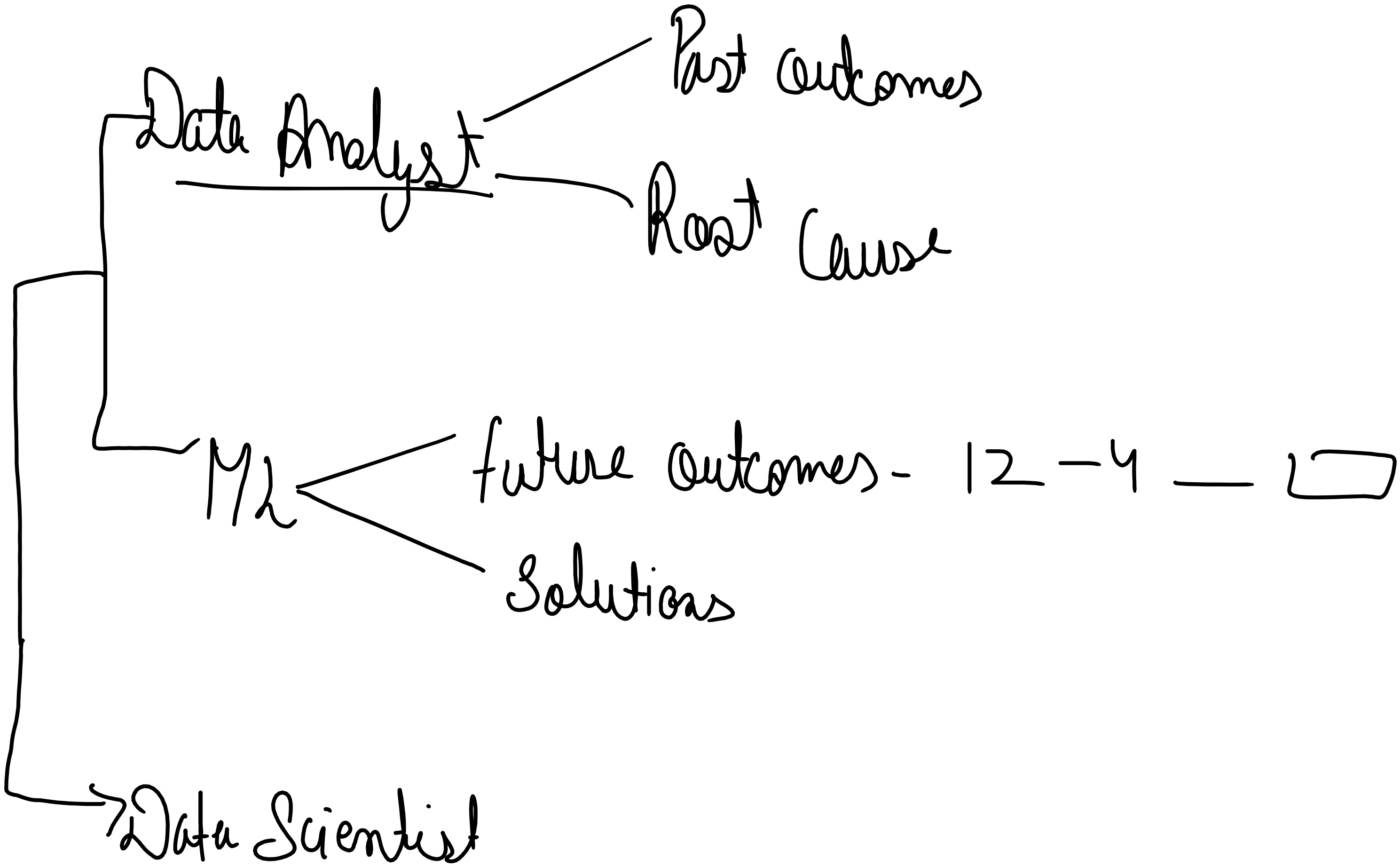
Collection of raw facts are my data.

For ex- Age of students of a class {12 ,24, 19, 15, 14, 20}

Information

useful data are my information. {Insights}

For example : Age of student of a class who are eligible for voting {24 , 19 , 20}.



Types Of Statistics

12 August 2024 23:26

describe

5 Number
count
min
max
mean
Rn

Types of Statistics

1. Descriptive — DA
2. Inferential

Descriptive

It helps us to organize and summarize data using numbers and graphs to look for a pattern in the data set.

Example: You have all the data on how the business is going on, how much inventory you keep, how many ~~customers come to your store, In which~~ month it has been more, at what day of the week it occurs ~~more. Which product~~ is being sold more at what point of time~~, on what~~ hours is your product sold more. What kind of customers come, do male customers come more at a certain point ~~in time, or female customers co~~me then. People with children come more, cigarette buyers come more, or beer buyers come more, or grocery item buyers come more

- Measures of Central tendency: Mean, Median, Mode.
- ✓ The measure of Dispersion:: Standard Deviation, Variance & Range

✓

Inferential

Popln

Small

Inferential statistics is a technique where what uhh do is , uhh take up a sample from a population and using that sample to make prediction about the populations. (because otherwise it is not possible)

- To make an inference or draw a conclusion from the population, sample data is used.
- Using probability to determine how confident we can be that the conclusion we make is correct. (Confidence Interval & margin of error)

Example: Our primary concern is to find out how many people like blue cars in the data set.

Suppose, in a city, 1 lakh people are there. For our analysis, we have taken 100 people from the data set. Out of 100, 20 people like blue cars. i.e., 20/100 means 20% population like blue cars. This 20% is descriptive Statistics.

If we say 20% +/- 2%, i.e., 20% people with 2% margin of error like blue cars. So in this, we are 98% sure that this is correct. This is called inferential.

- **Population**: A collection, of individuals or events whose properties are to be analyzed (**Raw data that we get from the client**).

  Population denoted by (N)

  — N

  CLT

- **Sample**: : A subset of the population. It should be representative of the population.

  Sample denoted by (n)

  — n

- **Things to be careful about which creating samples**
  - ✓ Sample Size
  - ✓ Random
  - ✓ Representative

\ Peru

Measures Of Central Tendency

12 August 2024 23:30

$\mu \approx \bar{x}$

(* measure of Central Tendency- it refers to the measures which is used to determine the centre of the distributed data.

Measures of centeral tendency consists – mean , median , mode

$$\mu = \frac{x_1 + x_2 + x_3 \dots \dots x_n}{N} = \frac{\sum x}{N}$$

So, the population mean is :

$$\bar{x} = \frac{x_1 + x_2 + x_3 \dots \dots x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$

- Here N is the size of the Population data set, x̄ is the sample mean, and x_i is the data points
- ∑ is the summation of the entire data set

Therefore the sample mean is :

1, 2, 3, 3, 4, 5, 100

M = 3

med = 3

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

M = 16.6

med = 3

- Here n is the size of the Population data set, x̄ is the sample mean, and x_i is the data points
- ∑ is the summation of the sample

Example: The systolic blood pressure of seven middle-aged men in:

150, 123, 134, 170, 146, 124 and 113.
The Mean is = (150+123+134+170+146+124+113)/7 = 137.14

Mode and Median

- The median for the sample data arranged in increasing order is defined as :
  1. If"n" is an odd number - Middle value
  2. If "n" is an even number - Midway between the two middle values
- The mode is the most commonly occurring value.
- Mode exists as a data point.
- Useful for qualitative data.

Example – if n is odd

The re-ordered systolic blood pressure data:

113,124,124,132,146,151 and 170.

-> The median here is 132.
-> Two individuals have systolic blood pressure = 124mm Hg, so the Mode is 124.

Example – if n is even

Six men with high cholesterol participated in the study to investigate the effects of diet on cholesterol levels. At the beginning of the study, their cholesterol levels (mg/dl) were as follows:

366, 327, 274, 292, 274 and 230

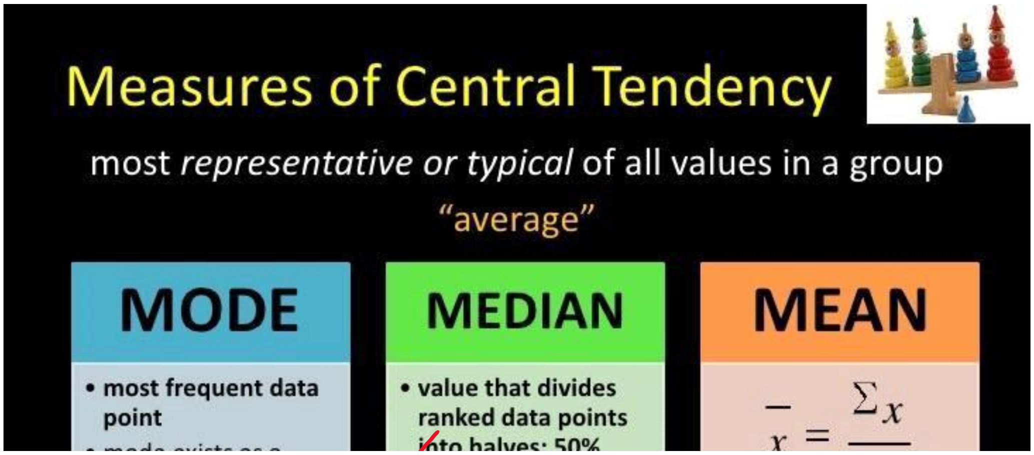
Rearrange the data in ascending order as follows:

230, 274, 274, 292, 327 and 366.

-> The median is 283(average of 274 and 292).

-> The mode between the two men having the same cholesterol level = 274.

Mean, Mode and Median in Brief:



Measures of Central Tendency

most *representative or typical* of all values in a group

"average"

MODE

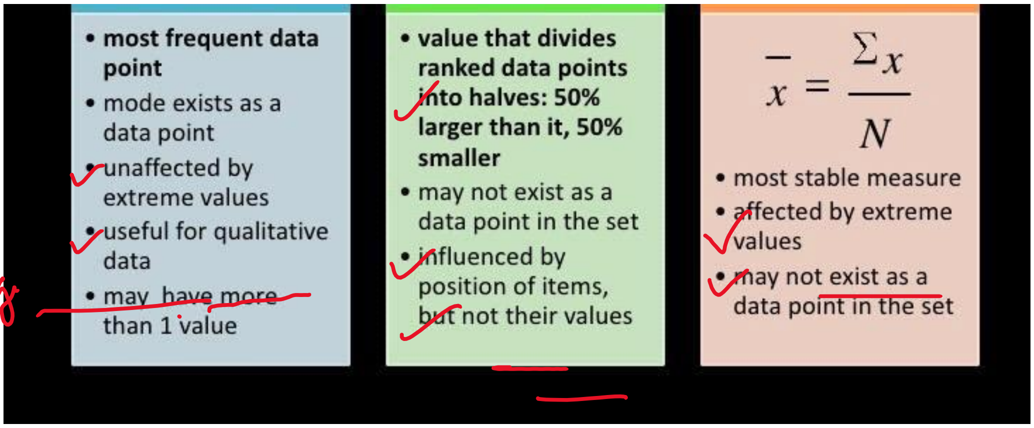
- most frequent data point
- mode exists as a

MEDIAN

- value that divides ranked data points into halves; 50%

MEAN

$\bar{x} = \dfrac{\Sigma x}{}$



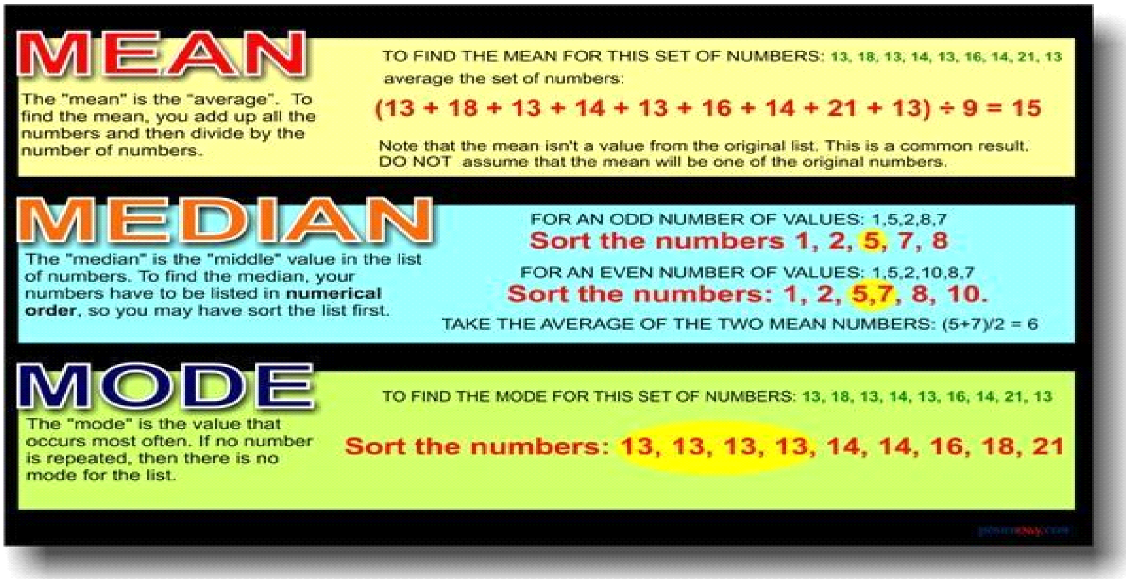
[illegible]

- most frequent data point
- mode exists as a data point
- ✓ unaffected by extreme values
- ✓ useful for qualitative data
- may have more than 1 value

- value that divides ranked data points ✓ into halves: 50% larger than it, 50% smaller
- may not exist as a data point in the set
- ✓ influenced by position of items, ✓ but not their values

$$\bar{x} = \frac{\Sigma x}{N}$$

- most stable measure
- ✓ affected by extreme values
- ✓ may not exist as a data point in the set



MEAN

The "mean" is the "average". To find the mean, you add up all the numbers and then divide by the number of numbers.

TO FIND THE MEAN FOR THIS SET OF NUMBERS: 13, 18, 13, 14, 13, 16, 14, 21, 13

average the set of numbers:

(13 + 18 + 13 + 14 + 13 + 16 + 14 + 21 + 13) ÷ 9 = 15

Note that the mean isn't a value from the original list. This is a common result. DO NOT assume that the mean will be one of the original numbers.

MEDIAN

The "median" is the "middle" value in the list of numbers. To find the median, your numbers have to be listed in **numerical order**, so you may have sort the list first.

FOR AN ODD NUMBER OF VALUES: 1,5,2,8,7

Sort the numbers 1, 2, 5, 7, 8

FOR AN EVEN NUMBER OF VALUES: 1,5,2,10,8,7

Sort the numbers: 1, 2, 5,7, 8, 10.

TAKE THE AVERAGE OF THE TWO MEAN NUMBERS: (5+7)/2 = 6

MODE

The "mode" is the value that occurs most often. If no number is repeated, then there is no mode for the list.

TO FIND THE MODE FOR THIS SET OF NUMBERS: 13, 18, 13, 14, 13, 16, 14, 21, 13

Sort the numbers: 13, 13, 13, 13, 14, 14, 16, 18, 21

✷ **Note:**

- Mean is highly sensitive to outliers
  - Example:
    - 1,2,3,4,5
      - □ -> Mean: 3

        -> Median: 3
    - 1,2,3,4,5,100
      - □ -> Mean: 51.5

        -> Median: 3.5

$2, 3, \;, \underline{7}, 9, 5$

$2, 3, \boxed{\underline{5}}, 7, 9 = \left(\frac{n}{2}+1\right)^{th}$

$\frac{5}{2}+1 = 3$

$1, 2, \overset{3.5}{\boxed{3, 4}}, 5, 6 \Rightarrow \frac{\left(\frac{n}{2}\right)^{th} + \left(\frac{n}{2}+1\right)^{th}}{2}$

3
/
3.5
3.9

$\frac{3+4}{2} = \qquad \frac{(3)^{th} + (4)^{th}}{2}$

Measure Of Dispersion — Spread of data

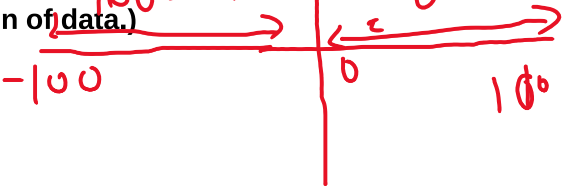
12 August 2024 23:37

$\frac{2000}{3}$ 0

$(-100-0)^2 + (100-0)^2$

$-100 + 100$

$10000+10000$ 0



(* measure of Dispersion- it is a statistical measure that describes the spread of variability of a dataset. it provides information about how the data is distributed around the central tendency (mean, median, mode) of the dataset.

it refers to the measures used to determine the spread or distribution of data.)

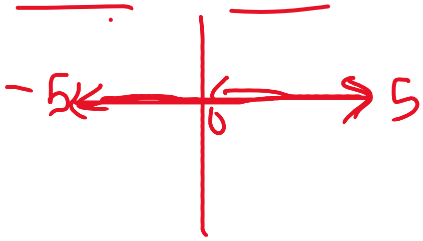
Measure of Dispersion consists: Range, variance, standard deviation

L max − min

Variance: The Variation of the data points around my mean is the variance or we can say that number of data points around my mean.

Variance is measured by first finding the deviation of each element in a data set from the mean and then by squaring it.

Basically , variance is average of all the squared deviations.

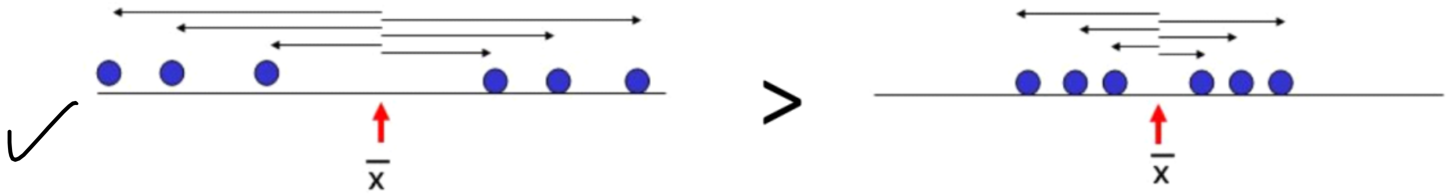


$\frac{-5+5+0}{3} = 0$

$5-(-5) = 10$

It is measured by first finding the Deviation of each element in a data set from the mean, and then by squaring it. Variance is an average of all squared deviations.

The below figure shows that On an avg how far point is distributed from the mean (x̄)



Var ∝ Spread of data

- Here(left side) variance is high because, from the mean(x̄), the points are distributed at a longer distance as compared to the right side, where the distance is a bit smaller.

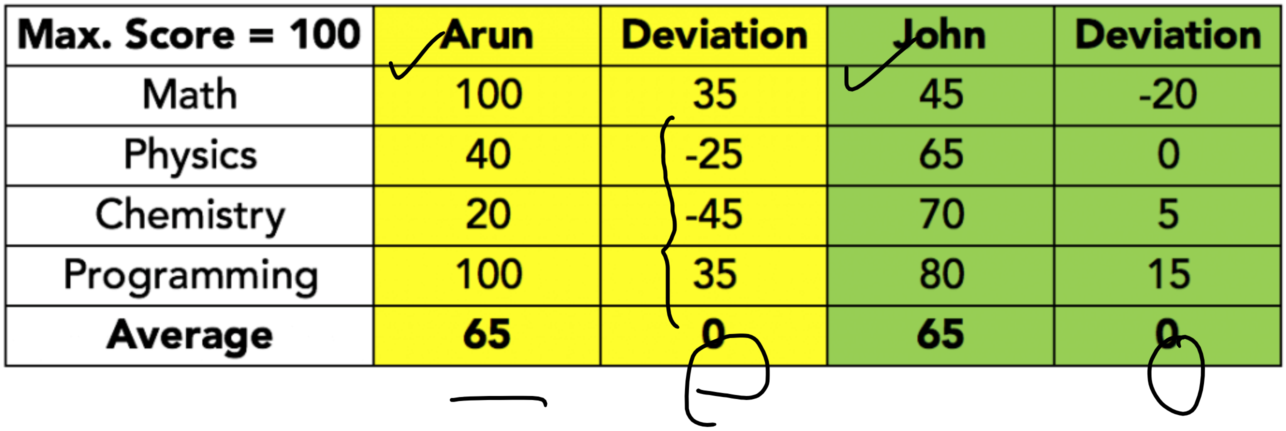
✷ **Variance is Directly Proportional To The Spread Of Data**

Lets take an example for understanding the variation

Dispersion, or spread of data, is measured in terms of how far the data differs from the mean. In other words, if mean is the centre of the data, if we get an idea about how far the individual data points deviate from the mean, we would have an idea about the spread. Therefore, a simple measure of dispersion could be an average of the differences about the mean.
However, this average would have a problem. The actual data points could lie on either side of the mean, and thus, the deviations could be either positive or negative. Now, if we were to compute the average of

these deviations, the negatives and positives would cancel out, and the average would be very low, or even zero.
Take a look at the table below; the deviations are calculated as *Value minus Mean*, and the average of the deviations is calculated:

Take a look at the table below; the deviations are calculated as *Value minus Mean*, and the average of the deviations is calculated:



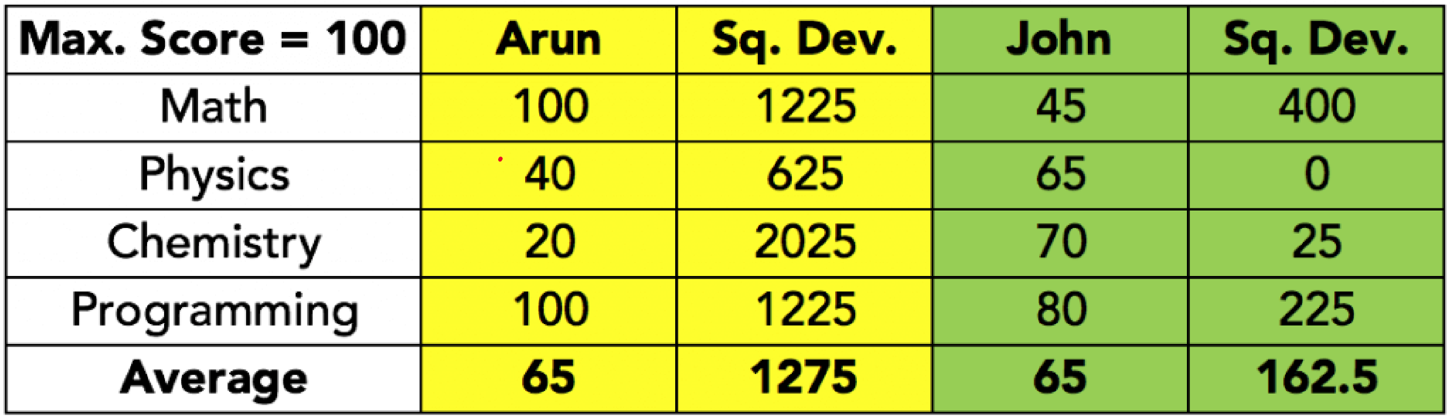
Max. Score = 100	✓ Arun	Deviation	✓ John	Deviation
Math	100	35	45	-20
Physics	40	-25	65	0
Chemistry	20	-45	70	5
Programming	100	35	80	15
Average	**65**	**0**	**65**	**0**

Does this mean that there are no deviations in the data? Clearly, that is untrue.

We need to manipulate the negative values in some way, so as to get rid of the negative sign. This is usually done by squaring the values before computing the average of the squared deviations. The resulting average value is called the variance. In the table below, the deviations have been squared, and then averaged:

`

20 ——————————— 100



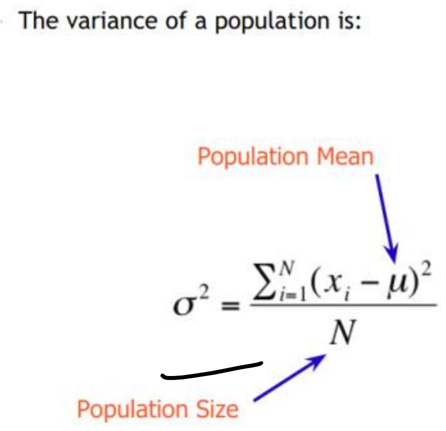
Max. Score = 100	Arun	Sq. Dev.	John	Sq. Dev.
Math	100	1225	45	400
Physics	40	625	65	0
Chemistry	20	2025	70	25
Programming	100	1225	80	225
Average	**65**	**1275** ✓	**65**	**162.5**

45 ——————— 80

This time, the numbers are more in line with what we would expect – Arun's scores have a higher deviation, compared to John's scores.

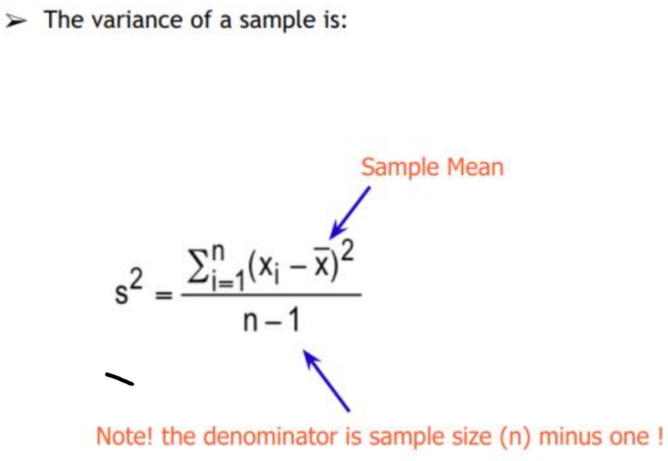
From this we get a overall variation of marks, but not an individual deviation of the marks from the mean

However, in doing this, while the earlier problem of numbers cancelling out is solved, a new problem emerges – what does the 1275 for Arun really mean? Is it indicative of some squared scores? Now, scores are understandable, but what are squared scores? Values with unit dimensions are meaningless when squared.



The variance of a population is:

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



➢ The variance of a sample is:

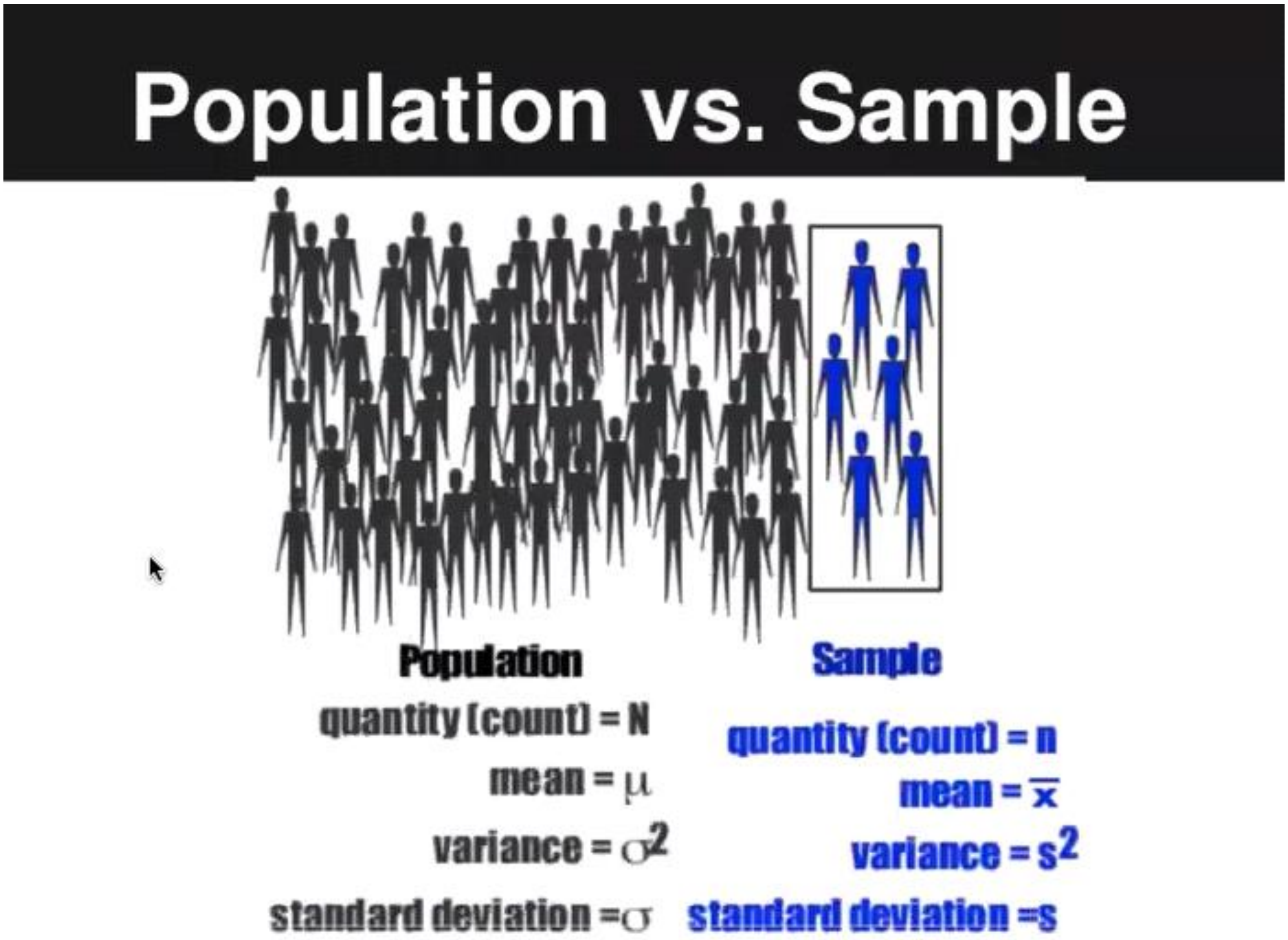
$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

Standard deviation- The distance of the data points from my mean is the standard deviation.

- Standard deviation tells us about the concentration of data around the mean of the data set.
- Standard deviation (S) is the square root of the variance.

✷ Facts about Standard Deviation:

- If the standard deviation is small, the data has little spread (i.e., the majority of points fall very near the mean).
- If standard deviation = 0, there is no spread. This only happens when all data items are the same value.
- The standard deviation is significantly affected by outliers and skewed distributions.



Population vs. Sample

Population	Sample
quantity (count) = N	quantity (count) = n
mean = μ	mean = $\overline{x}$
variance = σ^2	variance = s^2
standard deviation = σ	standard deviation = s

$$\frac{(\underline{-100}\underline{-}0)^2 + (0\underline{-0})^2 + (10\underline{0-}0)^2}{3}$$

$$\frac{\sum \text{Sq devia}}{\text{No devia}} =$$

$$\frac{\sum (x-\mu)^2}{N} = \sigma^2$$

$$\frac{\sum (x-\bar{x})^2}{n-1} = s^2 = \square$$

$$\sigma = \sqrt{\text{Varicun}}$$

$$s = \sqrt{\text{var}}$$

Data Classification

12 August 2024 23:48

Data classified into two types

1. **Numerical (Quantitative Variables)**
   a) Discrete variables
   b) Continuous variables
2. **Categorical (Qualitative Variables)**
   a) Ordinal variables
   b) Nominal variables

1. **Quantitative Data-** Quantitative data are **any data where the data represent amounts** (e.g. height, weight, or age).

Quantitative variable consists: Discrete data & Continuous data

a. **Discrete data** - A discrete data is **whose value is obtained by counting**. Basically Discrete variables are in Whole numbers.
It has a limited number of Possible Values.
For example- Number of childrens are in a family 1, 2 , 3 , 4, 5
Ranking of the Students- 1st, 2nd , 3rd , 4th , 5th ,6th
Age of a Person 12 ,23 ,44 , 95, 56, 78

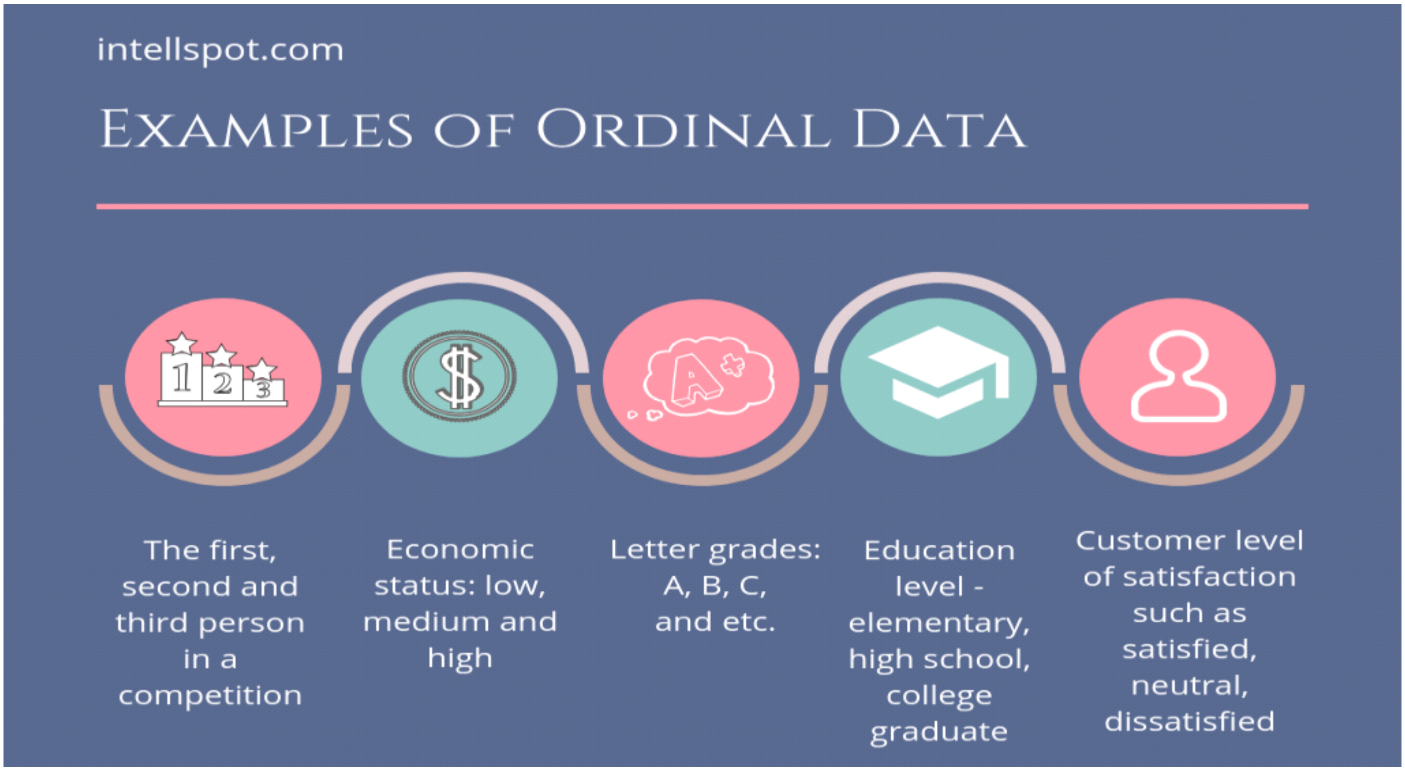
b. **Continuous Data** – Continuous data **is infinite and impossible to count.** Continuous data are in decimals numbers.
For example – Distance = 152.6, 8.25, 205.10, 55.0
How many values le in between 1.0- 2.0

2. **Qualitative Data** - A qualitative data also called a categorical data, is **a data that isn't numerical**. It

describes data that fits into categories. For example: States of india (data include: UP, MP, AP, HP, WB, AP).

Qualitative data consists two types: Ordinal data & Nominal Data

a. **Ordinal Data**- In Ordinal data the order of the values matter.
   For example. In a college dataset
   Highest Rank holder is on the Top of the table
   Lowest Rank Holder is on the Bottom of the table



intellspot.com

EXAMPLES OF ORDINAL DATA

The first, second and third person in a competition

Economic status: low, medium and high

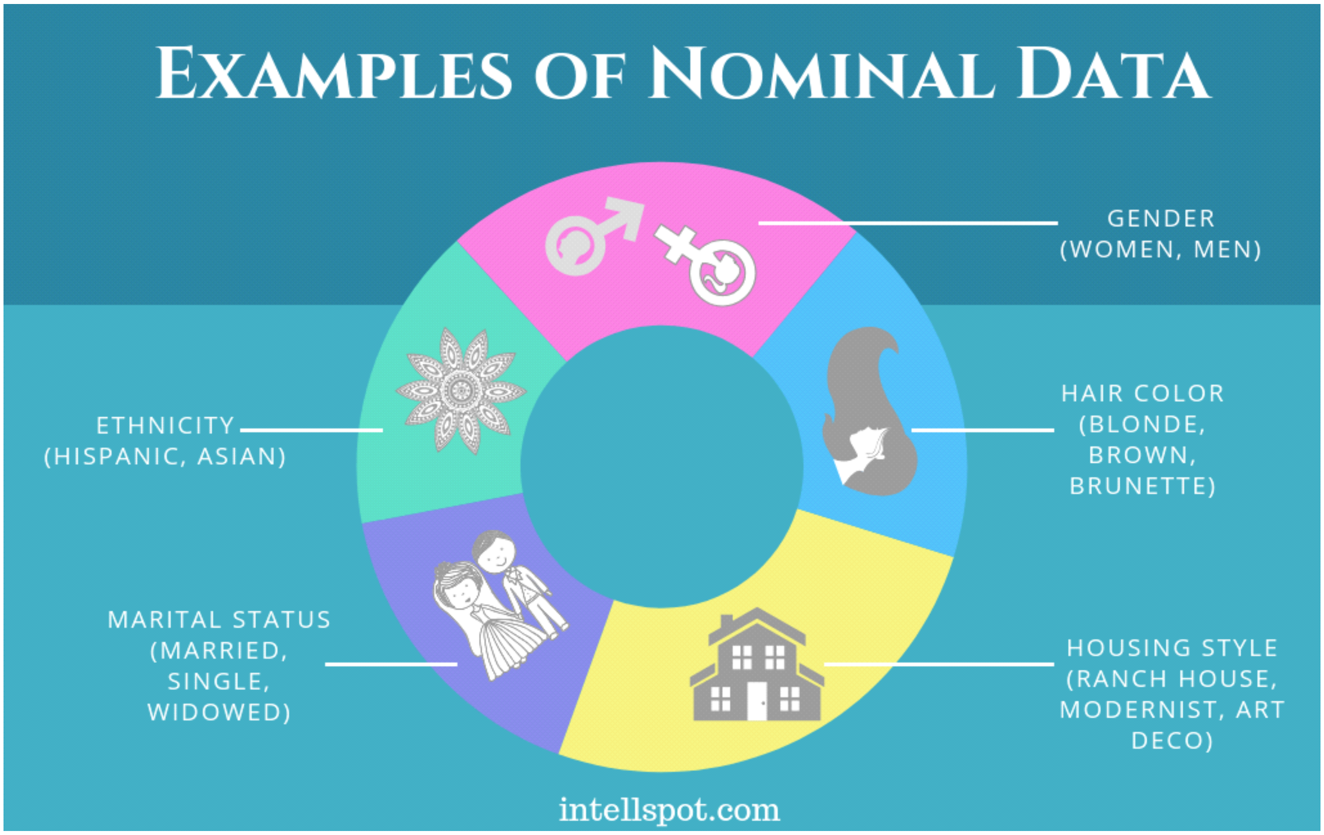
Letter grades: A, B, C, and etc.

Education level - elementary, high school, college graduate

Customer level of satisfaction such as satisfied, neutral, dissatisfied

Nominal data- In Nominal Data are the data than can be split intwo different categories.
For example. Colors= { Red, green, Yellow, Black}
Weather ={ Summer, Winter, Spring , Rainy }
Profession { Student, Teacher, player, engineer, doctor, driver}
Gender { Male , Female }

EXAMPLES OF NOMINAL DATA


EXAMPLES OF NOMINAL DATA
GENDER (WOMEN, MEN)
HAIR COLOR (BLONDE, BROWN, BRUNETTE)
HOUSING STYLE (RANCH HOUSE, MODERNIST, ART DECO)
MARITAL STATUS (MARRIED, SINGLE, WIDOWED)
ETHNICITY (HISPANIC, ASIAN)
intellspot.com

ASSINGMENT

1. What kind of Data Gender is ?
   Nominal data

2. What kind of data length of park is ?
   Continuous data

3. What kind of Data How many Days of the Month is ?
   Discrete data

4. What kind of data Population of the state ?
   Discrete data

5. What kind of data No. of coaches of the Train ?
   Discrete data

6. What kind of data Rank of the College Students ?
   Ordinal data

7. What kind of data Population of the state ?
   Discrete data

8. What kind of data Names of the Airlines ?
   Nominal data

9. What kind of data order of the rainbow color?
   Ordinal data

Percentiles

09 December 2024 11:42

Percentile- A percentile is a value below which a certain percentage of observations lie.
For example :
If say this number is the 25 percentile, this basically say that 25 percentage of entire data or distribution is less than that particular value.

Q1. Calculate the percentile of numbers that odd?
1,2,3,4,5
Percentile= (no.of odd numbers/ no. of numbers) * 100
= (3/5)*100
= 0.6*100
60%
It indicates that 60 % of my entire data have less even numbers
Q2. What is the percentile ranking of 10 ?
2 , 2 , 3 , 4 , 5 , 5 , 5 , 6 , 7 , 8 , 8 , 8 , 8 , 8 , 9 , 9 , 10 , 11 , 11 , 12

Percentile Ranking of 10= (no.of value below 10/ no. of data points)* 100
= (16/20)*100
0.8*100
80 %
Basically it indicates that 80% of the entire distribution, is than 10.
Q3. What is the percentile ranking of 11 ?

Q4. What value exists at percentile Ranking of 80% ?
Value= (percentile/100)* (n+1)
(80/100)*(20+1)
(80/100) * (21)
16.8
Neglect the floating part here 16.8 is the index positon
In our data set the value on 16th index is 10.
So 10 is exist at 80%
Q4. What value exists at percentile Ranking of 85% ?

Quantiles

09 December 2024 11:44

Quantiles: are statistical measures used to divide a set of numerical data into equal-sized groups, with each group containing an equal number of observations.

Quantiles are important measures of variability and can be used to understand distribution of data.

They also can be used to identify outliers

There are several types of quantiles used in statistical analysis, including

- **Quartiles :** Divide the data into four equal parts, Q1 (25th percentile), Q2 (50th percentile or median), and Q3 (75th percentile).
- **Deciles :** Divide the data into ten equal parts, D1 (10th percentile), D2 (20th percentile), ..., D9 (90th percentile).
- **Percentiles :** Divide the data into 100 equal parts, P1 (1st percentile), P2 (2nd percentile), ..., P99 (99th percentile).
- **Quintiles:** Divides the data into 5 equal parts, q1 (20 th percentiles) q2 (40 th percentiles),.......

Things to remember while calculating these measures:

Data should be sorted from low to high

You are basically finding the location of an observation

They are not actual values in the data

All other tiles can be easily derived from Percentiles

Outliers

09 December 2024 11:44

Outliers- Outliers are the extremely high or extremely low values — in a data set that can throw off your stats.

1 , 2 , 2 , 2 , 3 , 3 , 4 , 5 , 5 , 5 , 6 , 6 , 6 , 6 , 7 , 8 , 8 , 9 , 27

Lower limit= The number which is smaller than my lower limit that is outliers.

Lower limit= Q1-1.5*(IQR)

Upper limit - The number which is greater than my upper limit that is outliers.

Upper limit= Q3-1.5*(IQR)

IQR (inter quartile range) = q3-q1

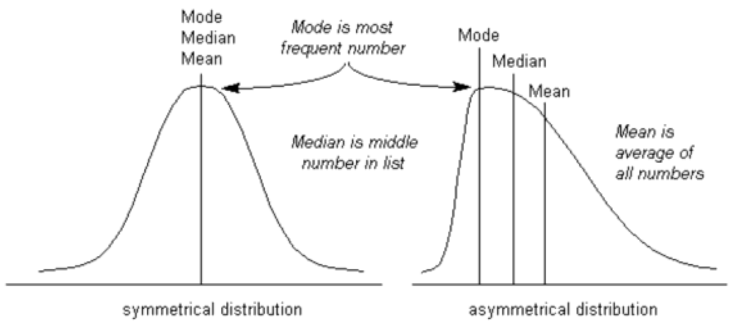
Q3 is my 75 %

Q1 is my 25 %

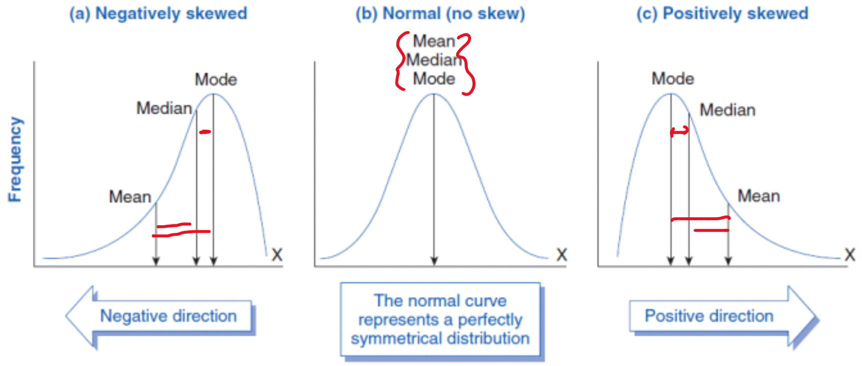
Measures Of Symmetry

16 August 2024 00:06

Distribution- Distribution is simply a collection of data which are arranged in order from smallest to largest.



Data poins 2, 3 , 3 , 3 , 3.2 , 3.1, 4



Normal Distribution (Gaussian Distribution)

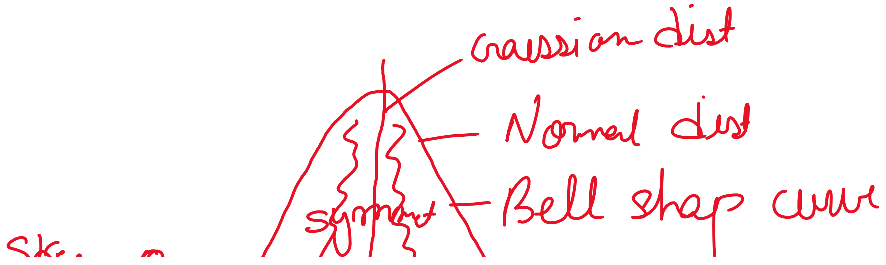
Normal Distribution is one of the most common continuous probability distribution. This type of distribution is important in statistics and is often used to represent random variables whose distribution is not known.

This type of distribution is symmetric, and it's mean, median, and mode are equal.

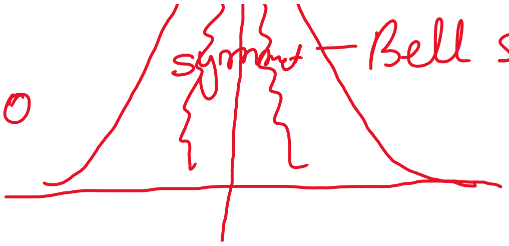
Mathematically, Gaussian Distribution is represented as:

N~(μ, σ2)

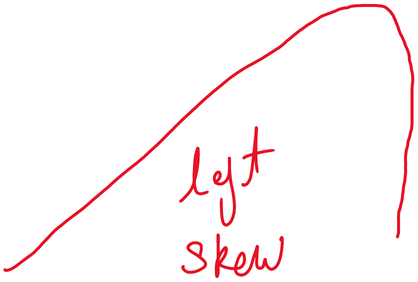
Where N stands for Normal, symbol ~ for distribution, whereas symbol μ stands for mean and σ2 stands for the variance.

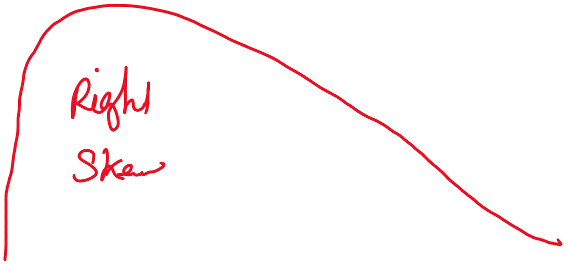


Skew = 0



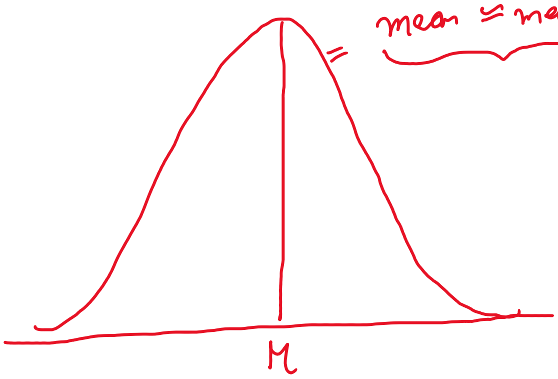
— Bell shap curve





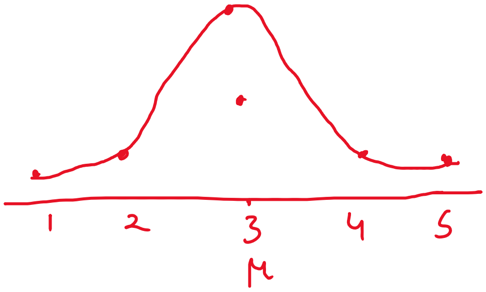
= +ve

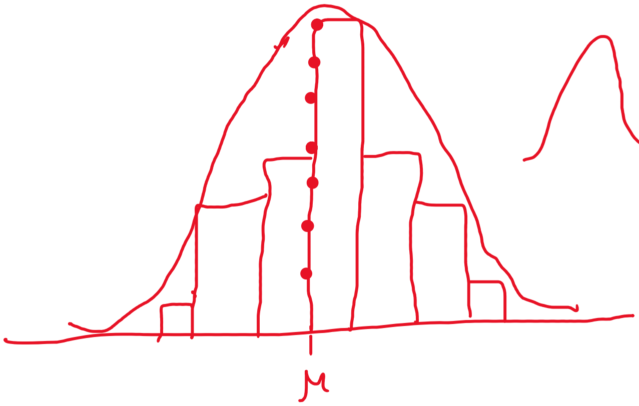
Skew = -ve

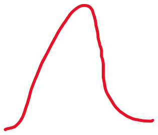


= mean ≈ median ≈ Mode

1, 2, 3, 3, 4, 5







mean
mer
Skew = 0

Normal Distribution (Gaussian Distribution)

16 August 2024 09:49

Normal Distribution (Gaussian Distribution)

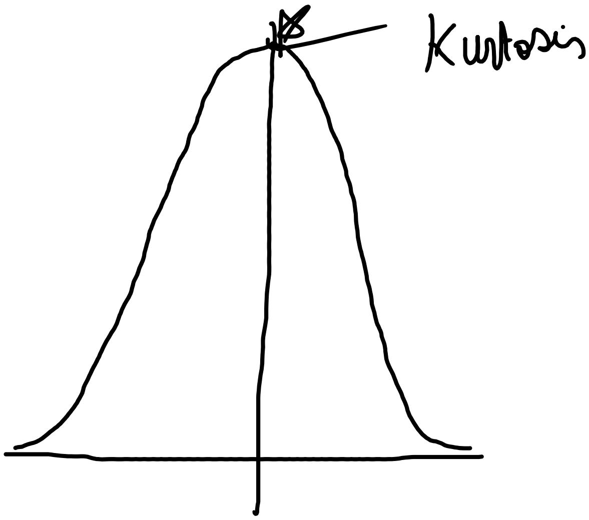
Normal Distribution is one of the most common continuous probability distribution. This type of distribution is important in statistics and is often used to represent random variables whose distribution is not known.

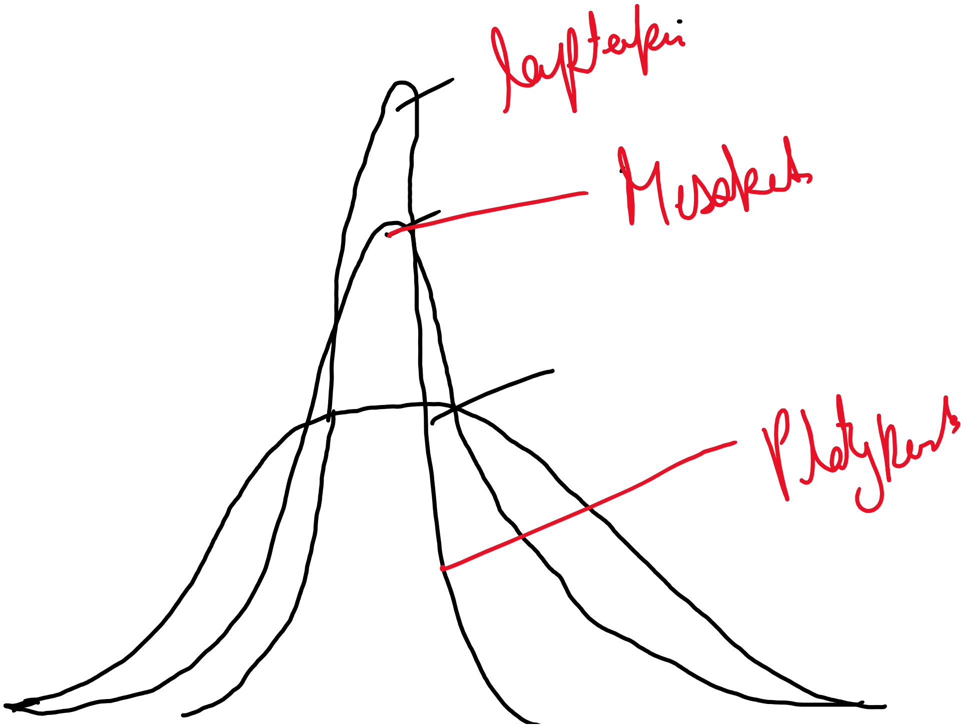
This type of distribution is symmetric, and it's mean, median, and mode are equal.

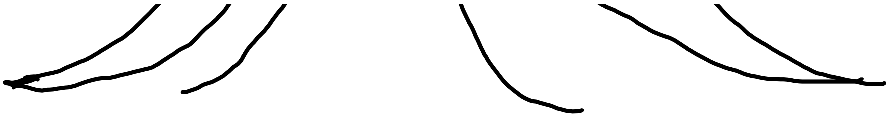
Mathematically, Gaussian Distribution is represented as:

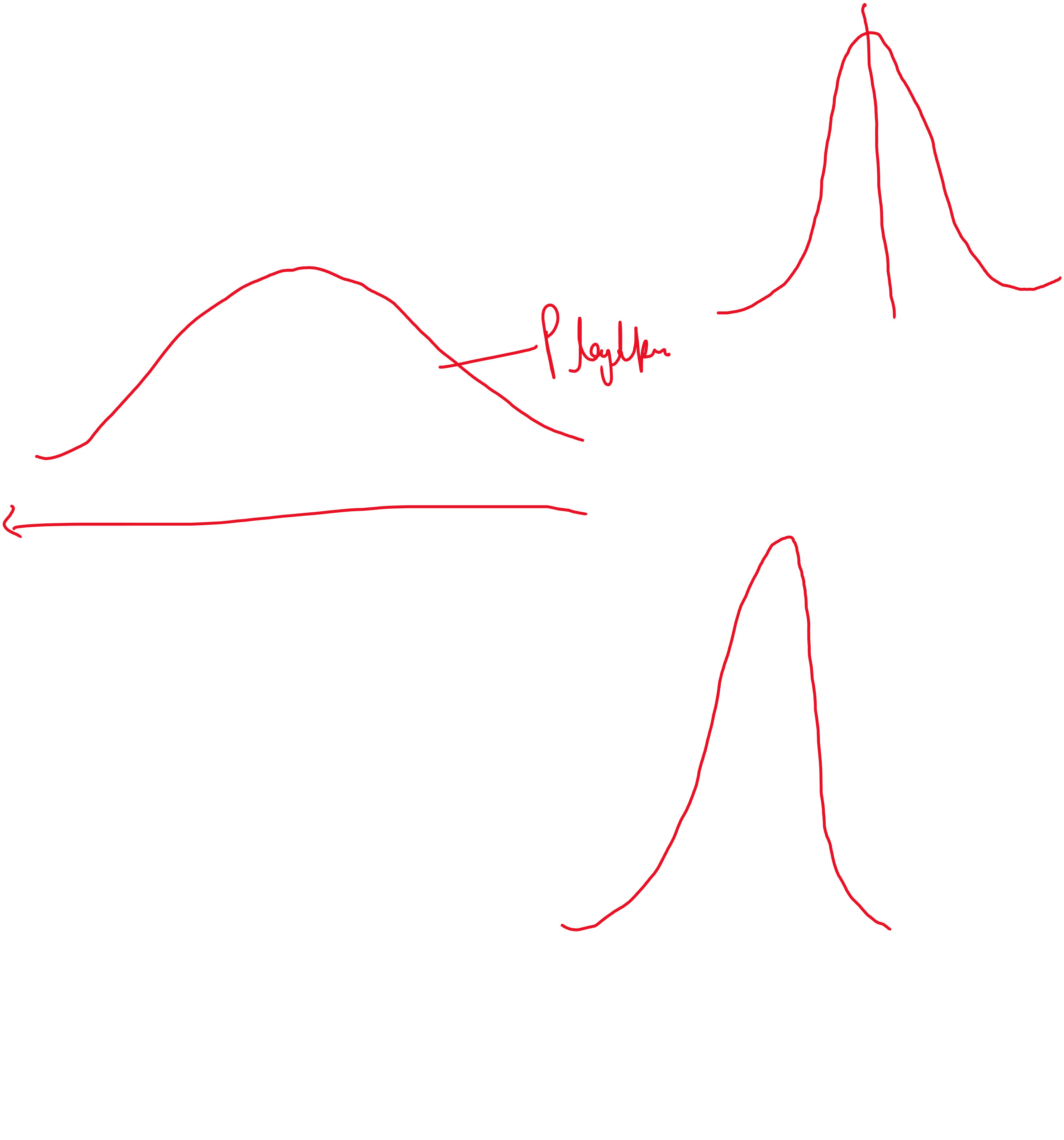
$$N\sim(\mu, \sigma2\ )$$

Where N stands for Normal, symbol ~ for distribution, whereas symbol μ stands for mean and $\sigma2$ stands for the variance.









A (i) distplot

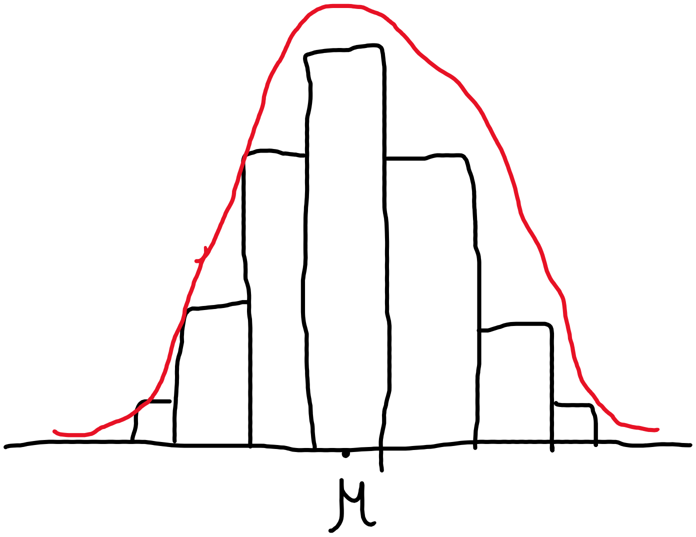
(ii)

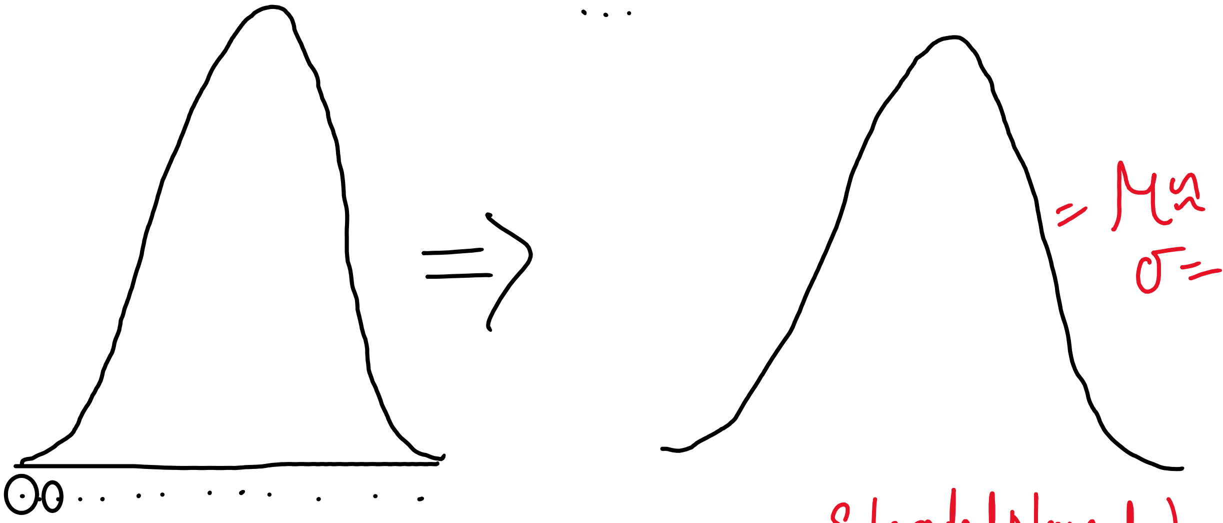
(i) disp

(ii) Skewness = 0

(iii) mean = median

(iv) Q-Q plot





= μ = 0
σ = 1

Standard Normal dist

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

Standardization

Standardization

-3σ to $+3\sigma$

Weight	salary
50kg	50000
60 kg	60000

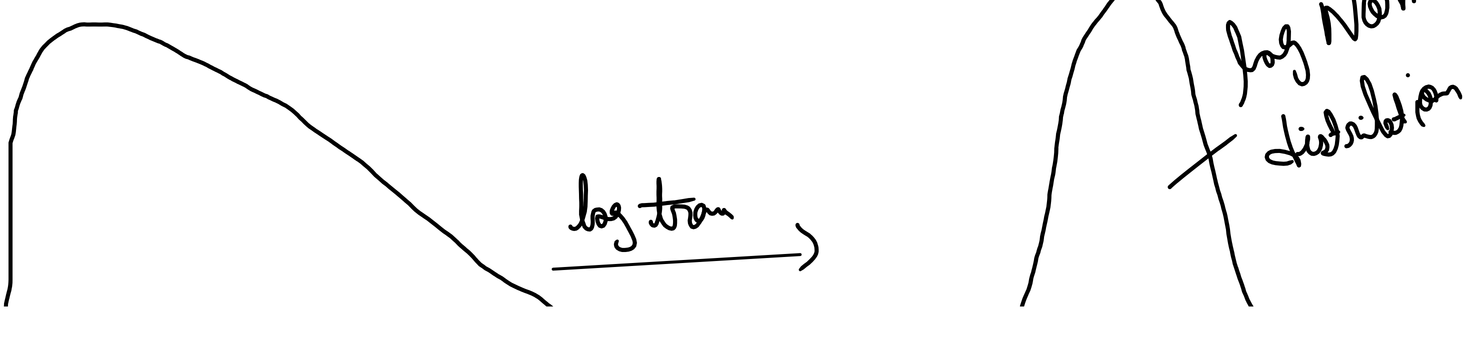
55kg

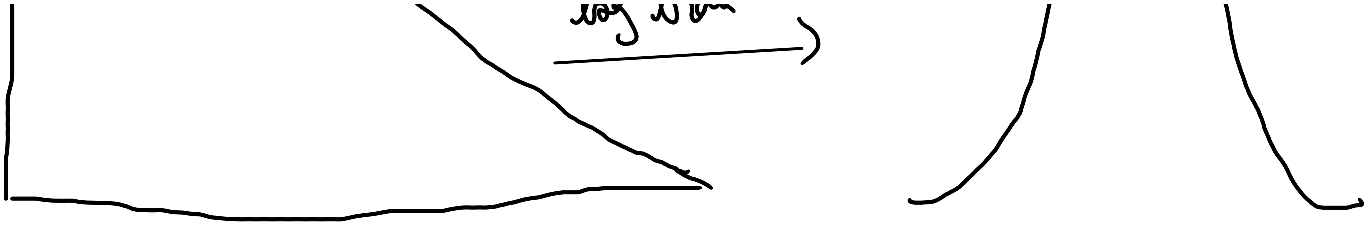
50005

(i) log transformation

(ii) Box cox

(iii) Yeo johnson




log

16 August 2024 09:49

Standard Normal Distribution

Understanding Standardization in the context of statistics. Every distribution can be standardized. Let say if the mean and variance of a variable are μ and σ2, respectively.

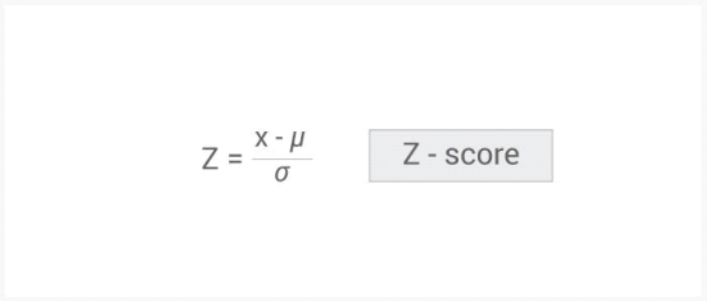
Standardization is the process of transforming the distribution to one with a mean of 0 and a standard deviation of 1.

i.e., ~(μ, σ2) → ~ (0, 1)

When a Normal Distribution is standardized, a result is called a Standard Normal Distribution.

i.e., N~(μ, σ2) → ~ N(0, 1)

We use the following formula for standardization:



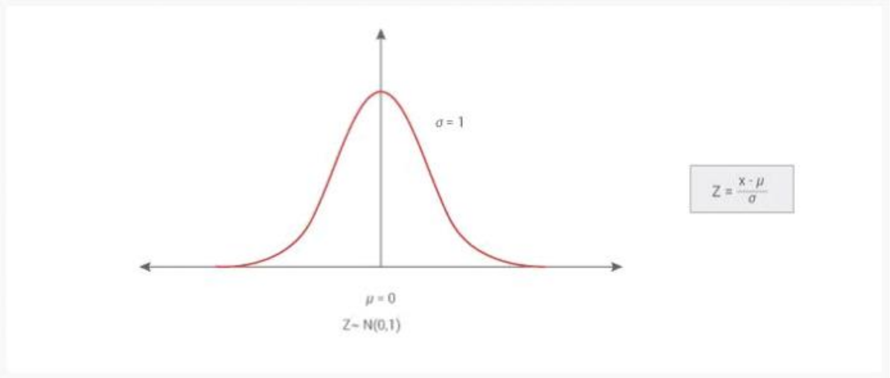
$$Z = \frac{x - \mu}{\sigma}$$

Z - score

Where x is a data element, μ is mean & ‘σ’ is the standard deviation, and Z is used to denote standardization, and Z is known as the z-score.

With the help of Z scores, we can come to know how far a value is from the mean. When you standardize a random variable, its μ becomes 0, and its standard deviation becomes 1.

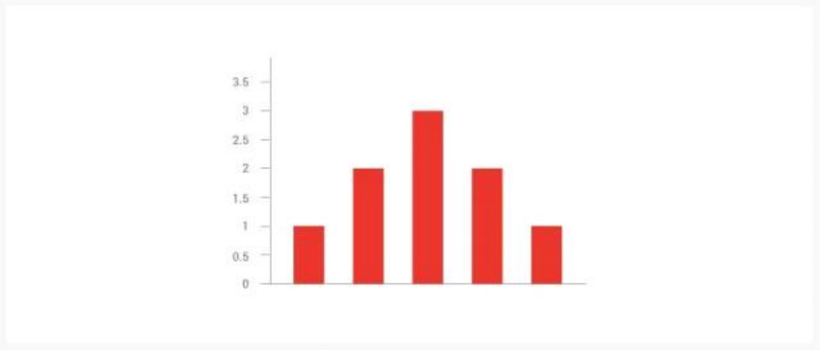
If the Z score of x is 0, then the value of x is equal to the mean.



Let us understand the steps in Standardization with the help of a simple example.

Suppose we have a dataset with elements

X = {1,1,1,2,2,2,3,3,4,4,4,4,5}



We get mean as 3, variance = 1.49 & standard deviation as 1.22 i.e., N ~ (3, 1.49).

Now we will subtract the mean from every the data points, that is, x – μ.

We will get a new data set mentioned below:

X1 = {-2, -1, -1, 0, 0, 1, 1 ,1, 2,2}

μ as 0, but the variance and std dev still as 1.49 and 1.22 respectively

i.e., N ~ (0, 1.49)

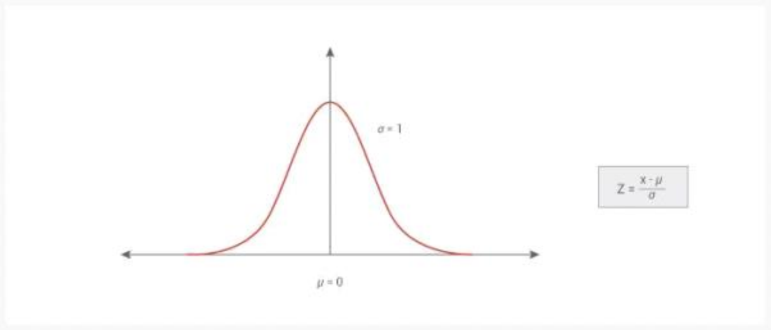
So in the next step of standardization, dividing all data points by the standard deviation, i.e., (x – μ)/σ

Dividing each datapoint by 1.22(standard deviation) we get a new data set as :

X2 = {-1.6, -0.82, 0, 0.82, 0, , 0.82,0,0.82, 0.82, 0 and 1.63.}

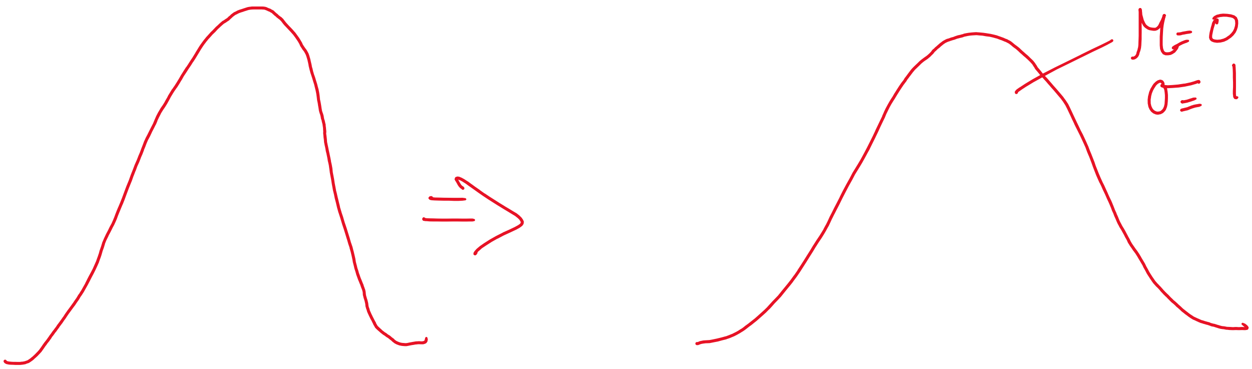
Now if we calculate the mean as 1 i.e., N ~ (0, 1)

Plotting it on a graph :



Using this standardized normal distribution makes inferences & predictions much easier.

$$Z = \frac{x - \mu}{\sigma}$$



5

55 kg

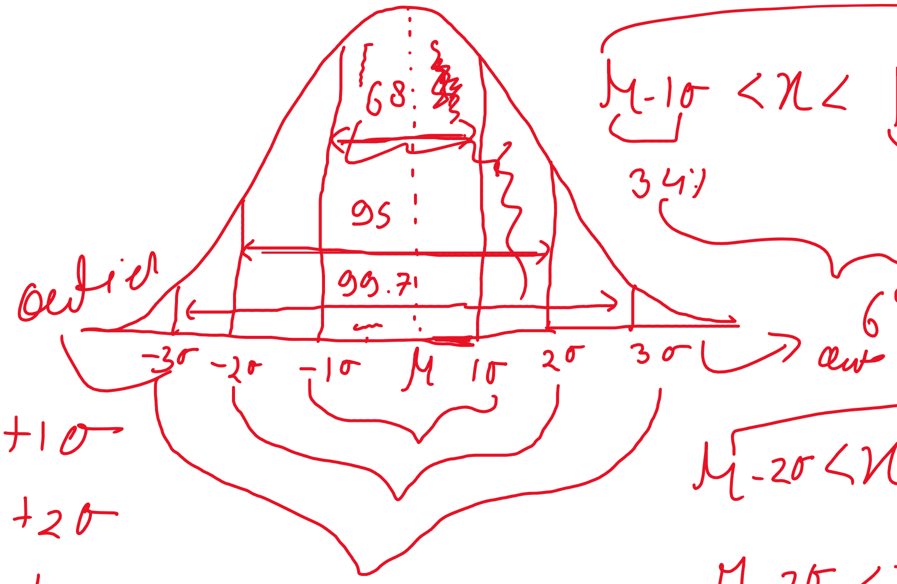
Weight	Salary
50kgs	50k
59kgs	60k
80kgs	80k

50005

100 data paid

[20 30, 5 - - - -

x



$\mu - 1\sigma < x < \mu + 1\sigma$

34% 34%

68%

$\mu - 2\sigma < x < \mu + 2\sigma$

$\mu - 3\sigma < x < \mu + 3\sigma$

-1σ $+1\sigma$

-2σ $+2\sigma$

-3σ $+33$

$$Z = \frac{x - \mu}{\sigma}$$

$Z = 3\sigma$

(i) log transformation

(ii) box cox

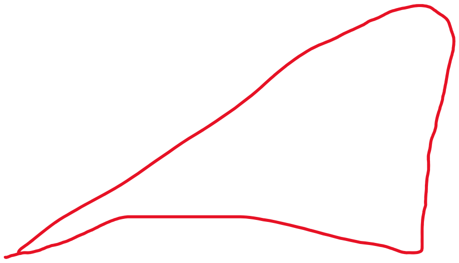
(iii) Yeo johnson

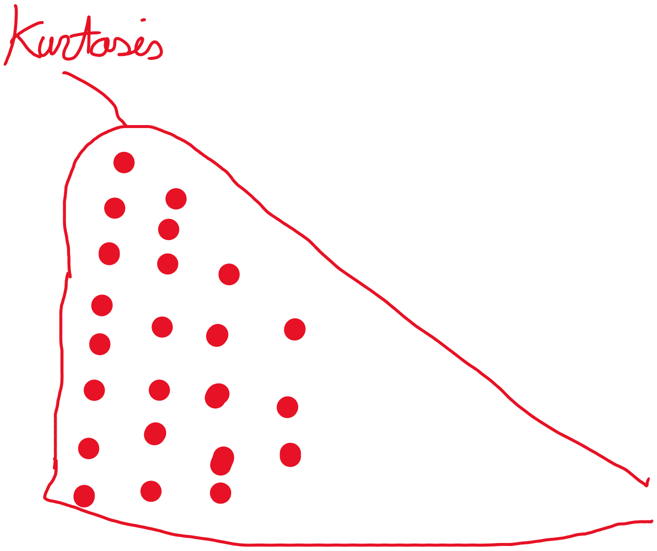
mean $\simeq$ median

Skew $\simeq$ 0

Bell shape

Probplot

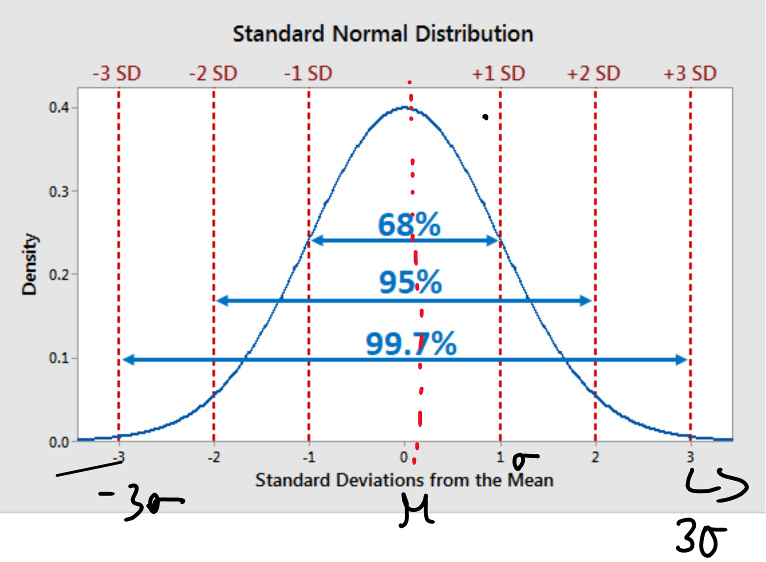




Empirical Rule

21 August 2024 12:49

[. . . 20 . . . .]



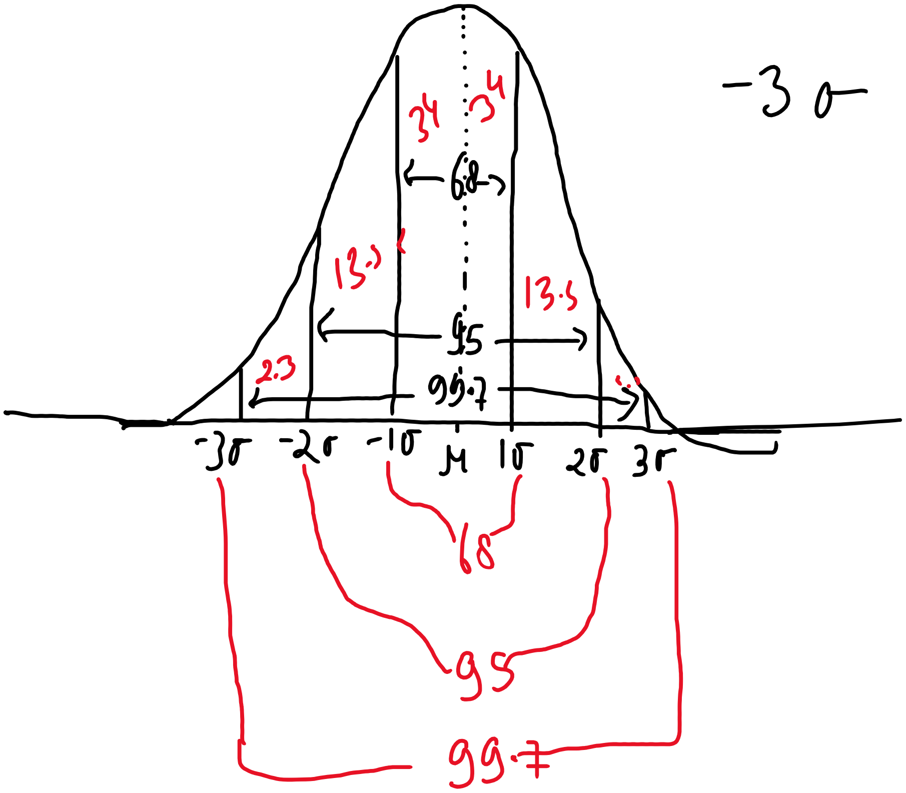
$$\mu - 1\sigma \le x \le \mu + 1\sigma = 68\%$$

$$\mu - 2\sigma \le x \le \mu + 2\sigma = 95\%$$

$$\mu - 3\sigma \le x \le \mu + 3\sigma = 99.7$$

$$Z = \frac{x - \mu}{\sigma}$$

$-3\sigma \quad +3\sigma$



dataset= [11,10,12,14,12,15,14,13,15,102,12,14,17,19,107,
10,13,12,14,12,108,12,11,14,13,15,10,15,12,10,14,13,15,10]

(i) Covariance }
(ii) Correlati } — Causation

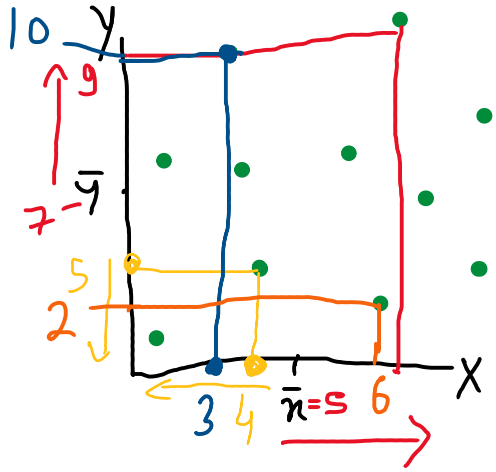
(iii) CLT

$x \uparrow\downarrow \quad Y \uparrow\downarrow = +ve$

$x \uparrow\downarrow \quad Y \downarrow\uparrow = -ve$

Size	Price
1000 Sqft	30 k$
2000 Sqft	50 k$

$-\infty$ ∞



$$\text{cov} = \frac{\sum (x-\bar{x})(y-\bar{y})}{n-1}$$

$$\text{Variance} = \frac{\sum (x-\bar{x})^2}{n-1} = \frac{\sum (x-\bar{x})(x-\bar{x})}{n-1}$$

$x\uparrow\ y\uparrow$ +ve $\quad x > \bar{x},\ y > \bar{y}$ 9

$$\frac{\sum \overbrace{(9-5)}^{+ve}\overbrace{(9-7)}^{+ve}}{100} = +ve$$

$x\downarrow\ y\downarrow$ +ve $\quad x < \bar{x},\ y < \bar{y}$

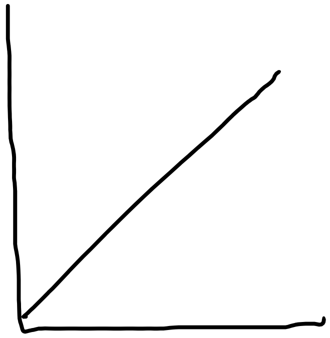
$$\frac{\sum \overbrace{(4-5)}^{-ve}\overbrace{(5-7)}^{-ve}}{100} = +ve$$

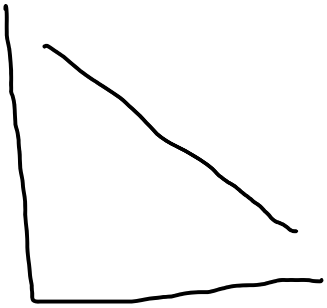
$x\downarrow\ y\uparrow = -ve \quad x < \bar{x},\ y > \bar{y}$

$$\frac{\sum \overbrace{(3-5)}^{-ve}\overbrace{(10-7)}^{+ve}}{100} = -ve$$

$x\uparrow\ y\downarrow = -ve \quad x > \bar{x},\ y < \bar{y}$

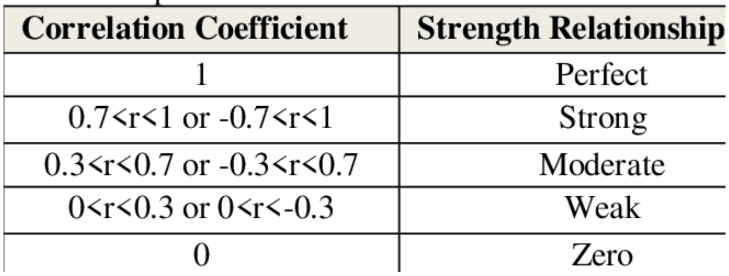
$$\frac{\sum \overbrace{(6-5)}\overbrace{(2-7)}^{-ve}}{100} = -ve$$


+ve


-ve

Correlation < direct, strength

.
Relationship



Correlation Coefficient	Strength Relationship
1	Perfect
0.7<r<1 or -0.7<r<1	Strong
0.3<r<0.7 or -0.3<r<0.7	Moderate
0<r<0.3 or 0<r<-0.3	Weak
0	Zero

CENTRAL LIMIT THEOREM (CLT)

22 August 2024 09:55

Sampling Distribution: Sampling distribution is a probability distribution that describes the statistical properties of a sample statistic (such as the sample mean or sample proportion) computed from multiple independent samples of the same size from a population.

In simple terms i can say that we draw a sample of the same size from the population at multiple times. And whatever the samples you have generated you calculate thee mean of every sample OR variance OR std (any kind of summary statistics) and this set is known as sampling distribution of sample mean,OR sampling distribution of sample variance OR sampling distribution of sample std respectively..

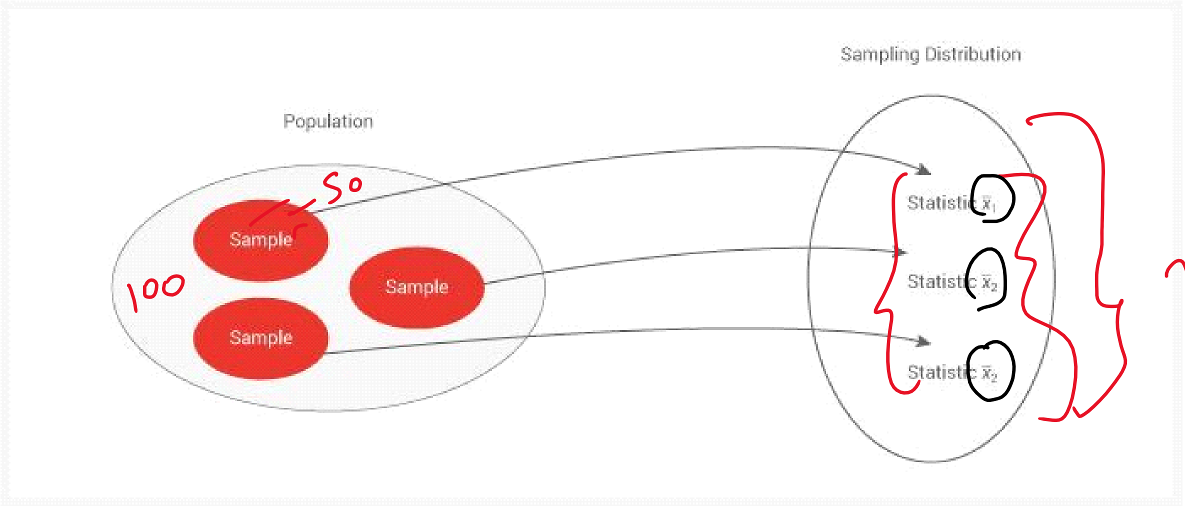
Why Sampling Distribution is important?
Sampling distribution is important in statistics and machine learning because it allows us to estimate the variability of a sample statistic, which is useful for making inferences about the population. By analysing the properties of the sampling distribution, we can compute confidence intervals, perform hypothesis tests, and make predictions about the population based on the sample data.

CLT

22 August 2024 12:02

The Central Limit Theorem (CLT) states that the distribution of the sample means of a large number of independent and identically distributed random variables will approach a normal distribution, regardless of the underlying distribution of the variables.

$M \simeq \bar{x}$



mean

The conditions required for the CLT to hold are:

1. The sample size is large enough, typically greater than or equal to 30.

$M \simeq \bar{x}$

2. The sample is drawn from a finite population or an infinite population with a finite variance.

3. The random variables in the sample are independent and identically distributed.

\ 5 Sem

$$X_1 = [\overbrace{x_1, x_2 \ldots\ldots\ldots x_{50}}^{<50}] = \bar{X}_1$$

$$X_2 = [x_1, x_2 \ldots\ldots\ldots x_{50}] = \bar{X}_2$$

$$X_3 = [x_1, x \ldots\ldots x_{50}] = \bar{X}_3$$

$$\vdots$$

$$X_5 = [x_1, x_2 \ldots\ldots x_{50}] = \bar{X}_5$$

$[\bar{X}_1, \bar{X}_2, \bar{X}_3, \bar{X}_3, \bar{X}_5\}$ – Sampling dist