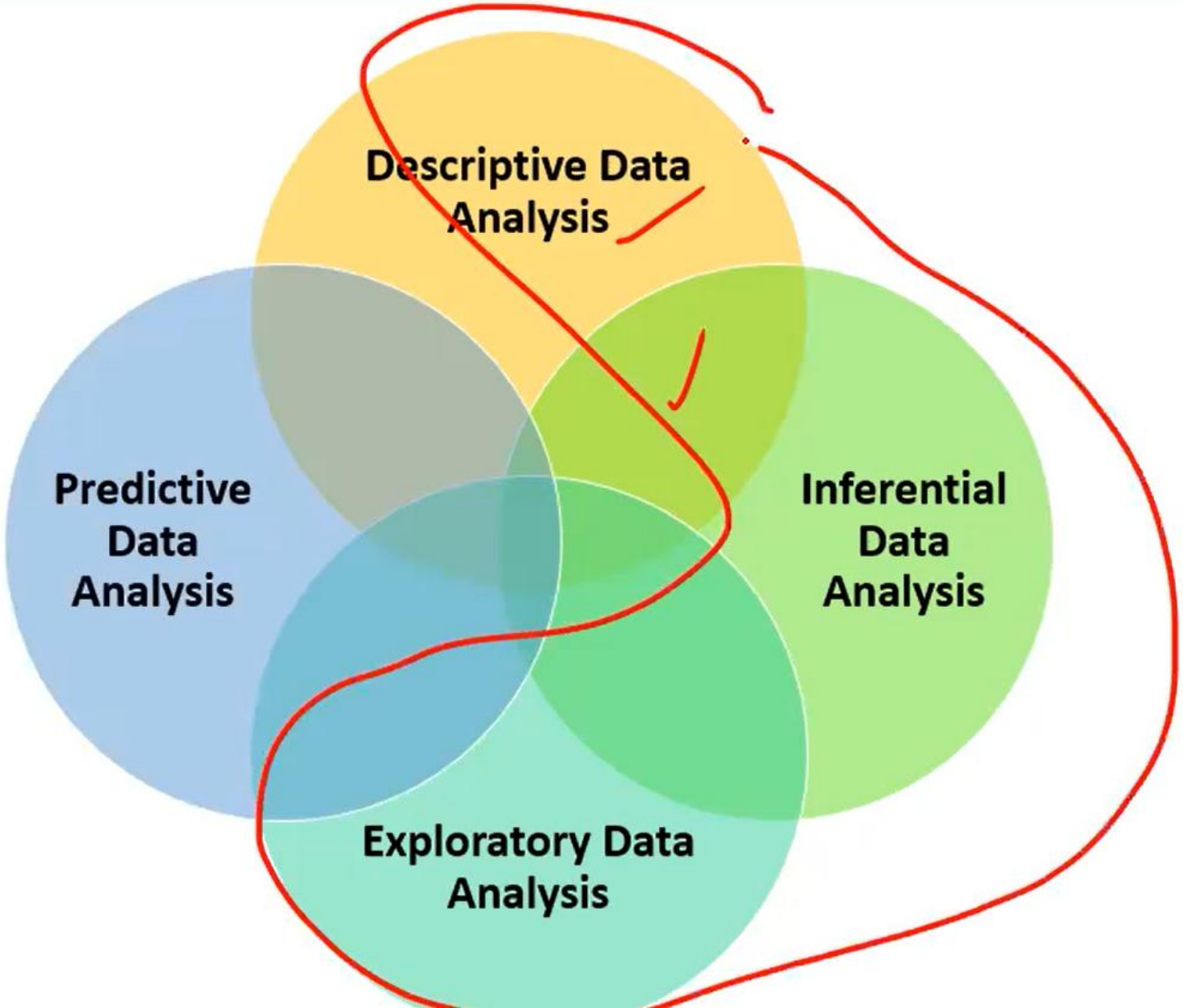
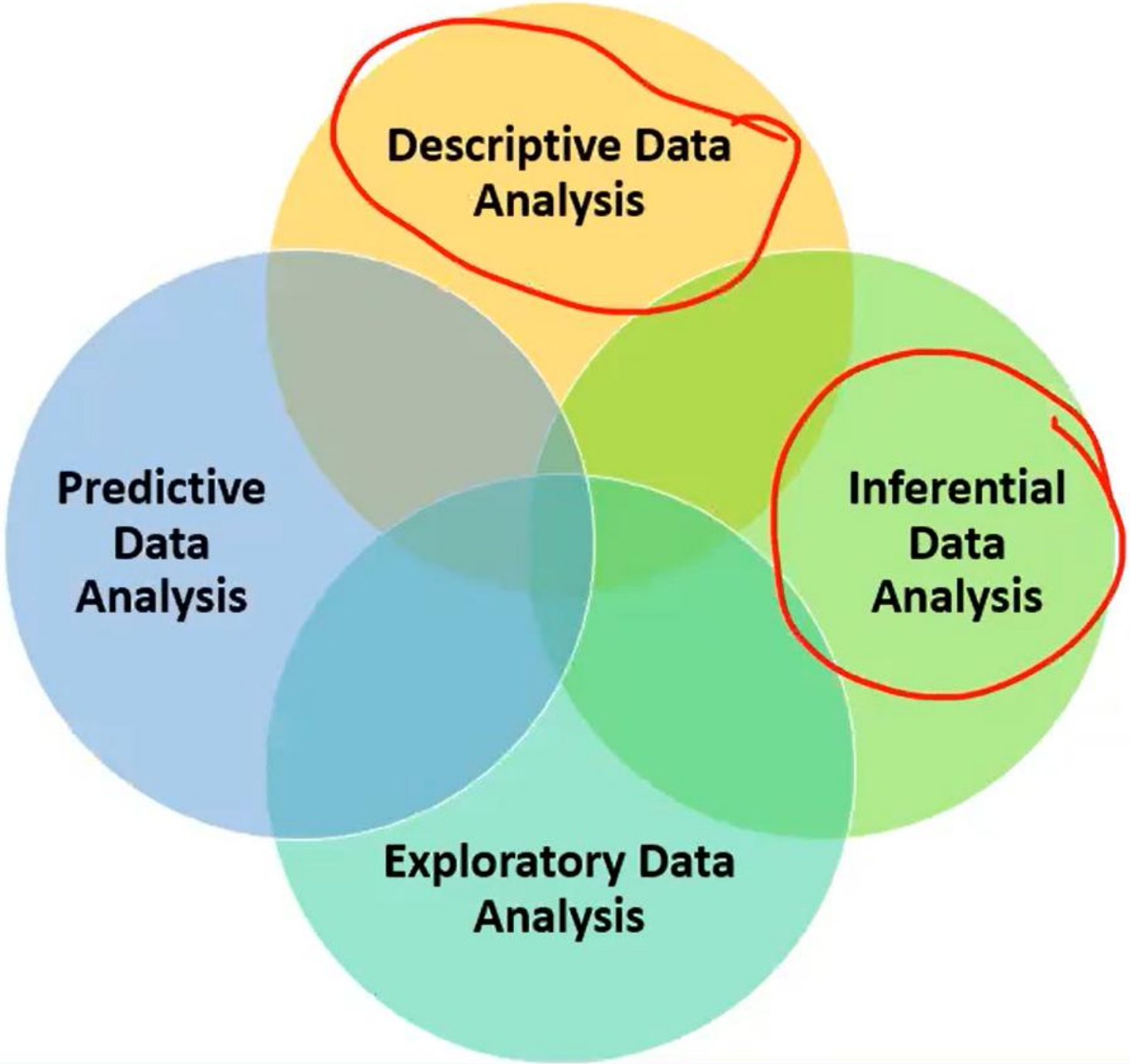
Data Science

Sub Domain Under Data Science




Descriptive Data Analysis
Predictive Data Analysis
Inferential Data Analysis
Exploratory Data Analysis

Data science

Got it 👍 let's keep it very **simple**:
Descriptive analysis = just **describes what you see** in the data.
👉 Example: *"The average exam score in our sample is 75."*
Inferential analysis = goes a step further and **uses the sample to say something about the whole population**.
👉 Example: *"Since the sample average is 75, we can estimate that the average score of all students in the school is likely between 72 and 78."*
It's called *inferential* because you are **inferring** (guessing with evidence) about the bigger group using only a part of it.

Example: Tea Preference 🧉
You want to know: *"Do most people in my office prefer green tea or coffee?"*
But you can't ask **all 500 employees**, so you take a **sample** of 50 people.
👉 From the sample:
30 people said **coffee**
20 people said **green tea**

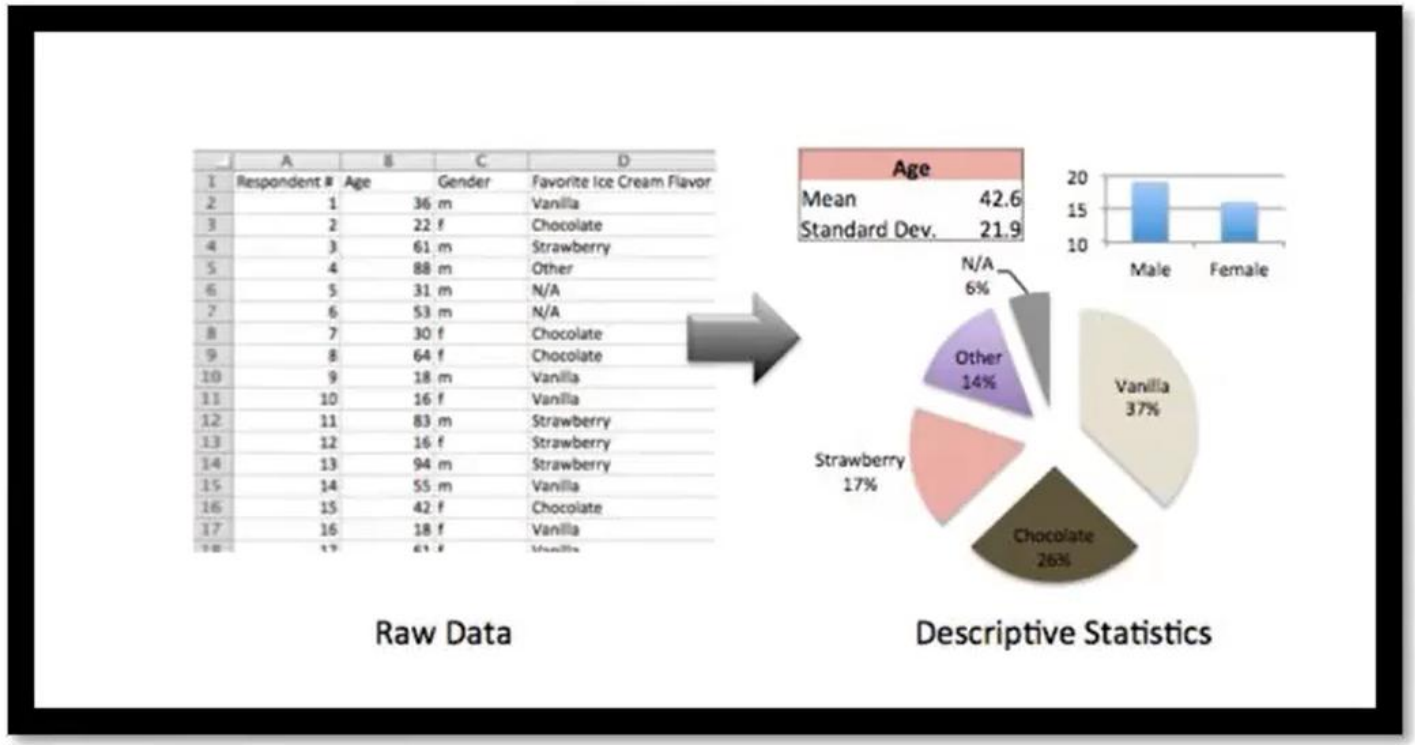
Descriptive analysis:
"In my sample, 60% prefer coffee and 40% prefer green tea."
Inferential analysis:
*"Based on my sample, I can infer that in the **entire office** population, the majority probably prefer coffee."*
You might also calculate a **confidence interval** (e.g., *"I'm 95% confident that between 55% and 65% of the whole office prefers coffee."*)

So **descriptive = sample facts**, **inferential = population conclusions** ✅

Ramisha Rani K, AI Trainer

TYPES OF DATA ANALYSIS

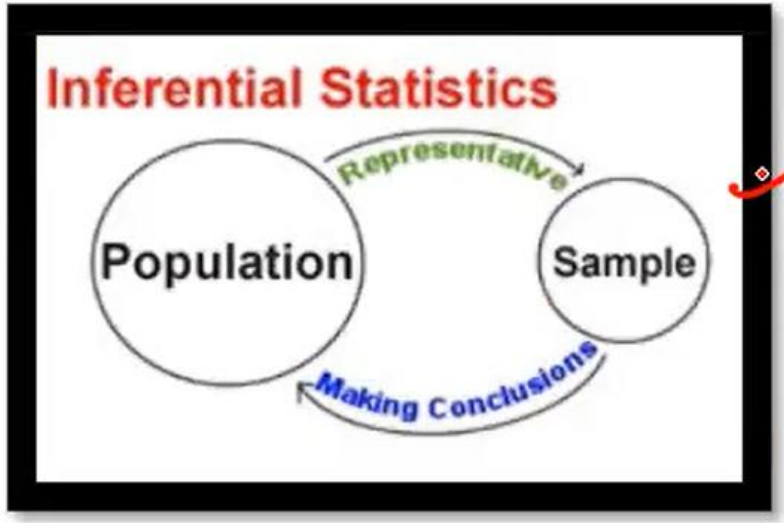
Descriptive Analysis



ainer

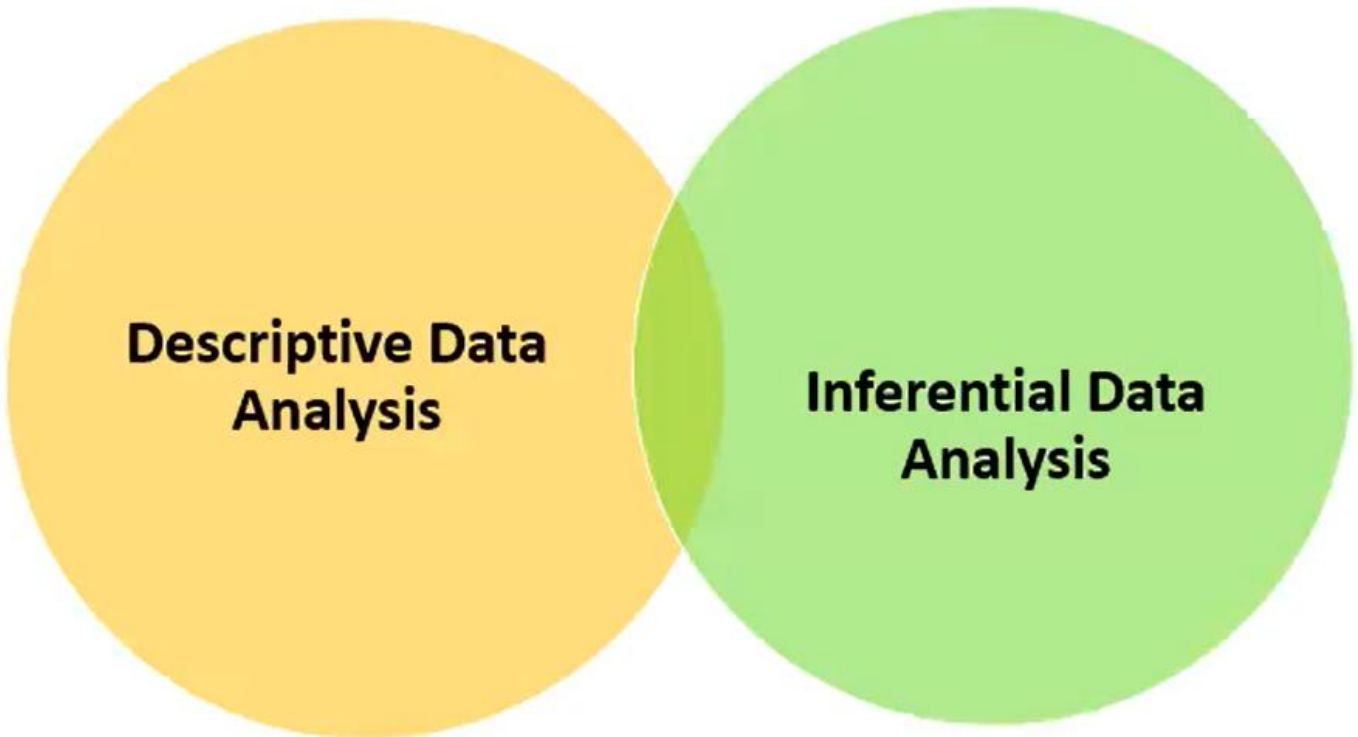
TYPES OF DATA ANALYSIS

Inferential Data Analysis



TYPES OF DATA ANALYSIS

Exploratory Data Analysis



Ramisha Rani K, AI Trainer

APPLICATION OF DATA SCIENCE

Statistics Concept	Application Field	Example
Binomial Probability Distribution $P(r) = {}_nC_r\,(p)^r(1-p)^{n-r}$ Mean $\mu = np$ Standard Deviation $\sigma = \sqrt{np(1-p)}$ p- probability of success r- number of successes n- number of trials		Sampling Inspection

amisha Rani K, AI Trainer

APPLICATION OF DATA SCIENCE

Statistics Concept	Application Field	Example
Correlation Analysis		Cross Market Analysis

Types Of Data Analysis

EXPLORATORY DATA ANALYSIS

ainer

TYPES OF EXPLORATORY DATA ANALYSIS

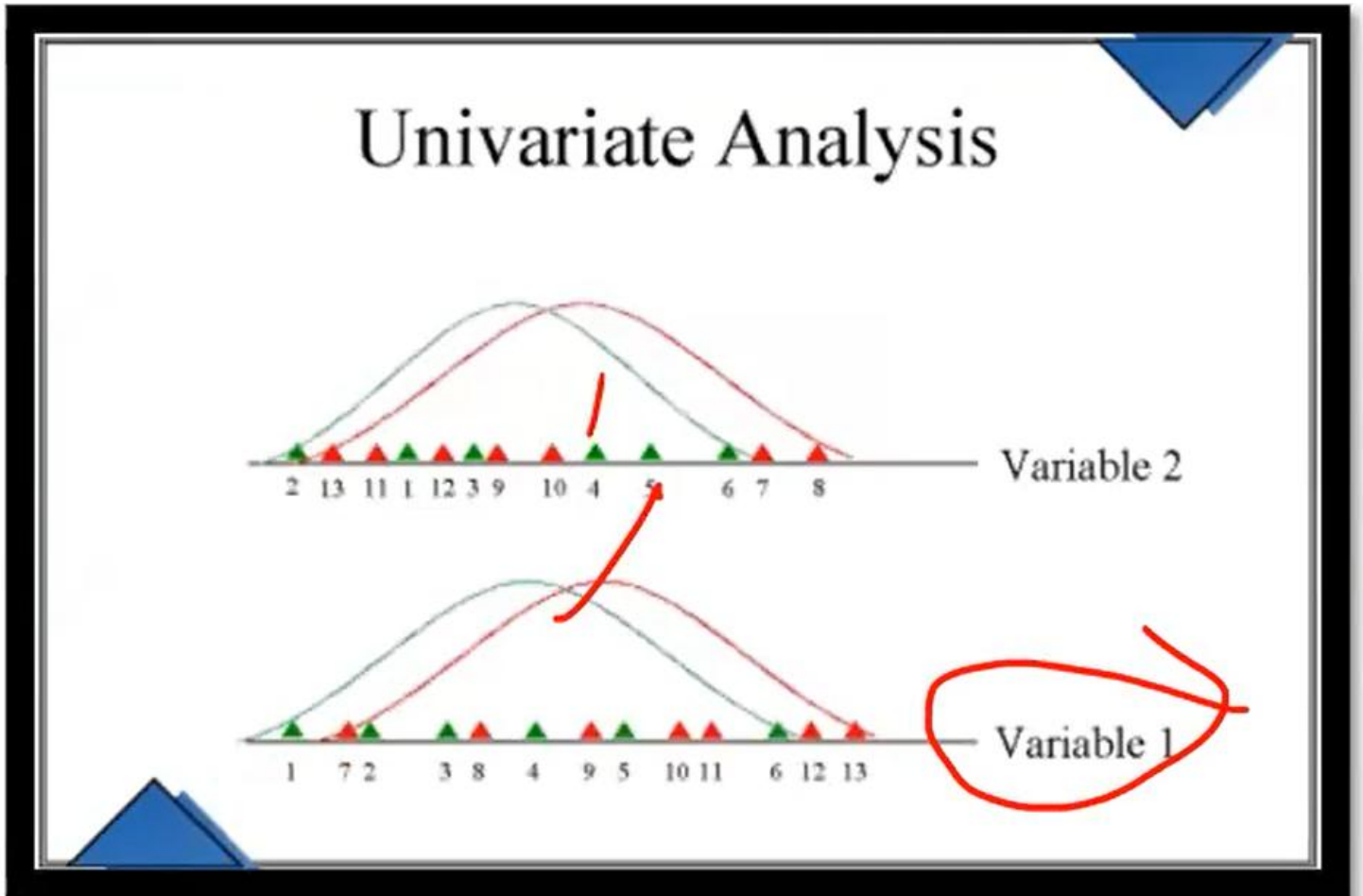
Univariate Data	Bivariate Data	Multivariate Data
One Variable Mean Average ..	Two Variable Cause and Effect Relationship	More Than Two Variable Cause and Effect of All the variable

TYPES OF EXPLORATORY DATA ANALYSIS

Univariate Data

TYPES OF EXPLORATORY DATA ANALYSIS

One Variable

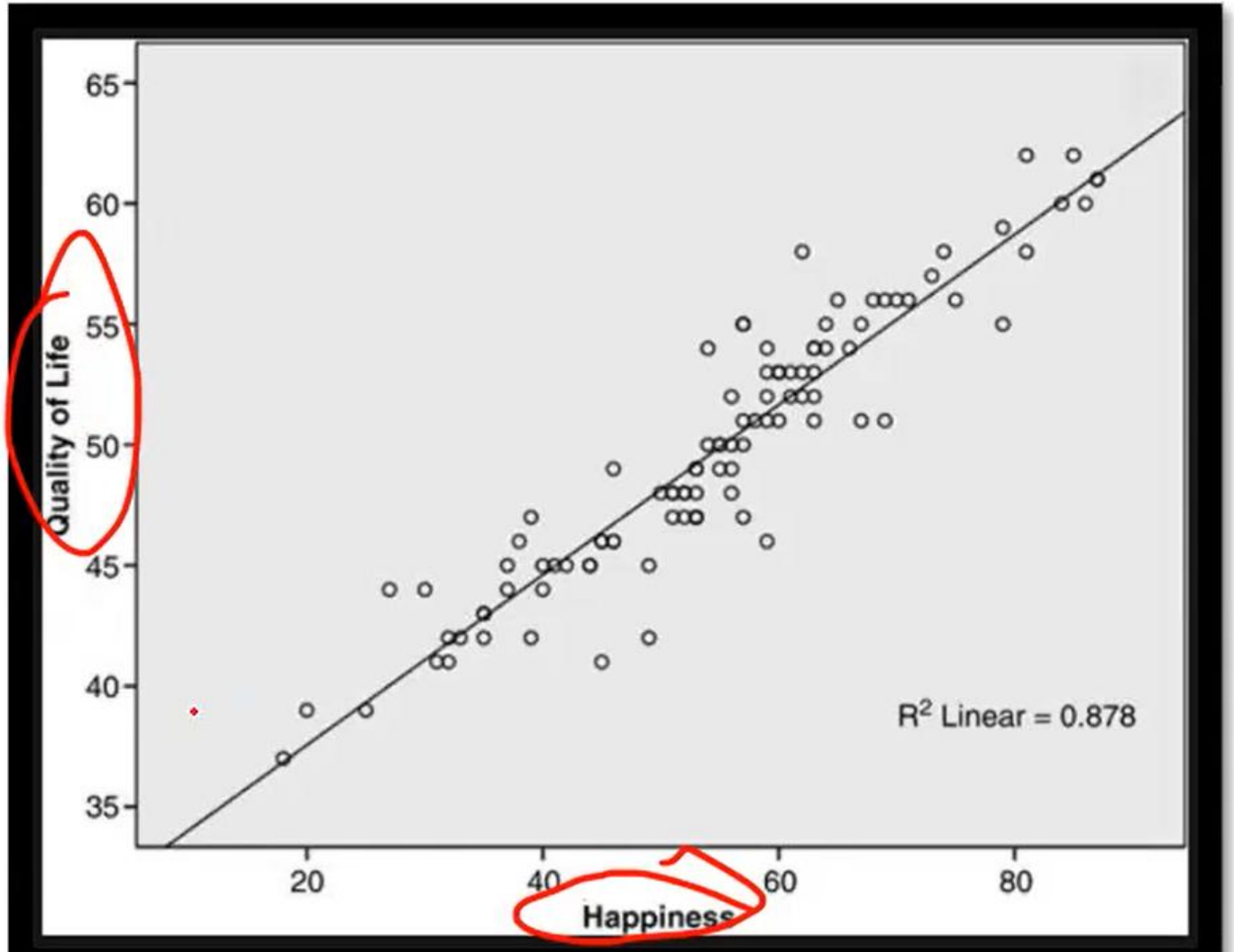


TYPES OF EXPLORATORY DATA ANALYSIS

Bivariate Data

TYPES OF EXPLORATORY DATA ANALYSIS

Two Variable
Cause and Effect



Rani K, AI Trainer

TYPES OF VARIABLE

Continuous Value

•Values that can be **measured** and can take **any value (including decimals)** within a range.

•Example:

- Height (170.2 cm, 170.25 cm, etc.).
- Weight (65.5 kg, 65.55 kg, …).
- Temperature (22.1°C, 22.15°C).

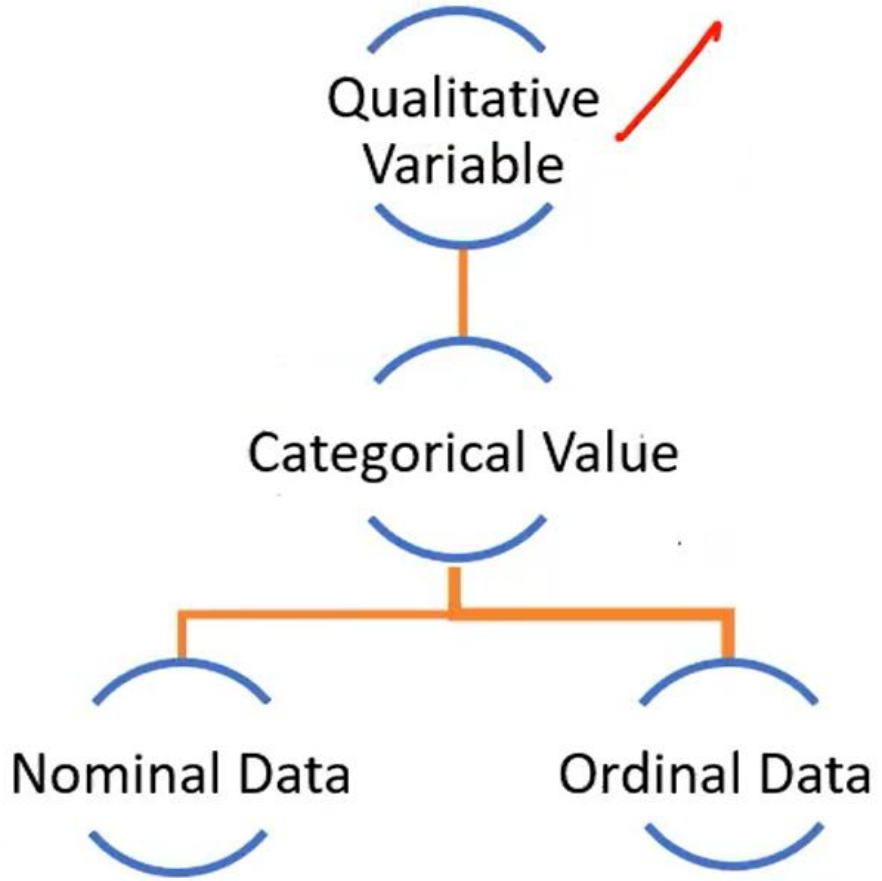
👉 Think: **measurable, can have fractions/decimals**.

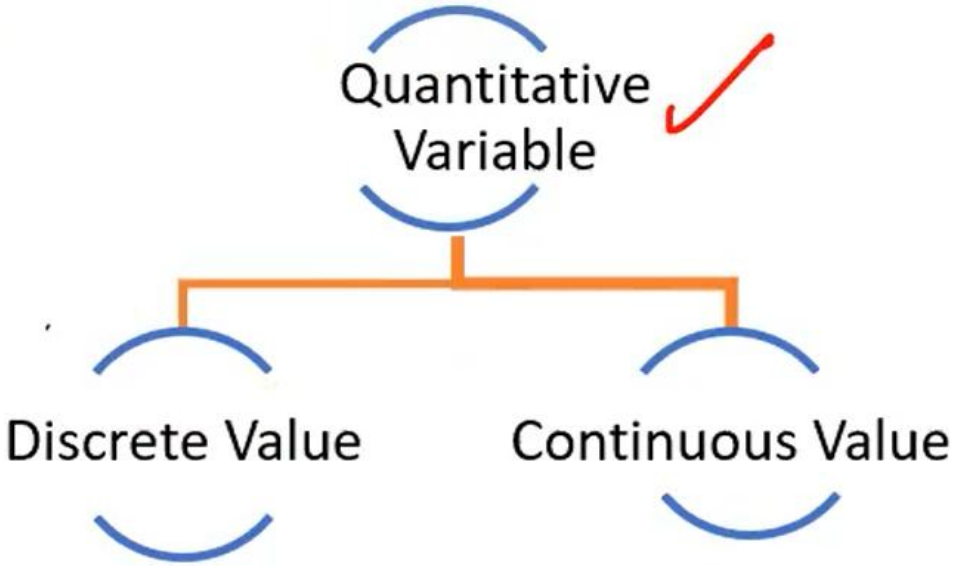
🔹 **Discrete Value**

•Values that can be **counted** (separate, fixed numbers).

•Example:

- Number of students in a class (25, 26, 27… not 25.5).
- Number of cars in a parking lot.





Person	Number of Pets 🐈 (Discrete)	Height (cm) 📏 (Continuous)
A	2	168.5
B	0	172.3

TYPES OF VARIABLE

Qualitative Variable — Categorical Value

MIP_Sample_Table Dataset - Sheet 1 [4]

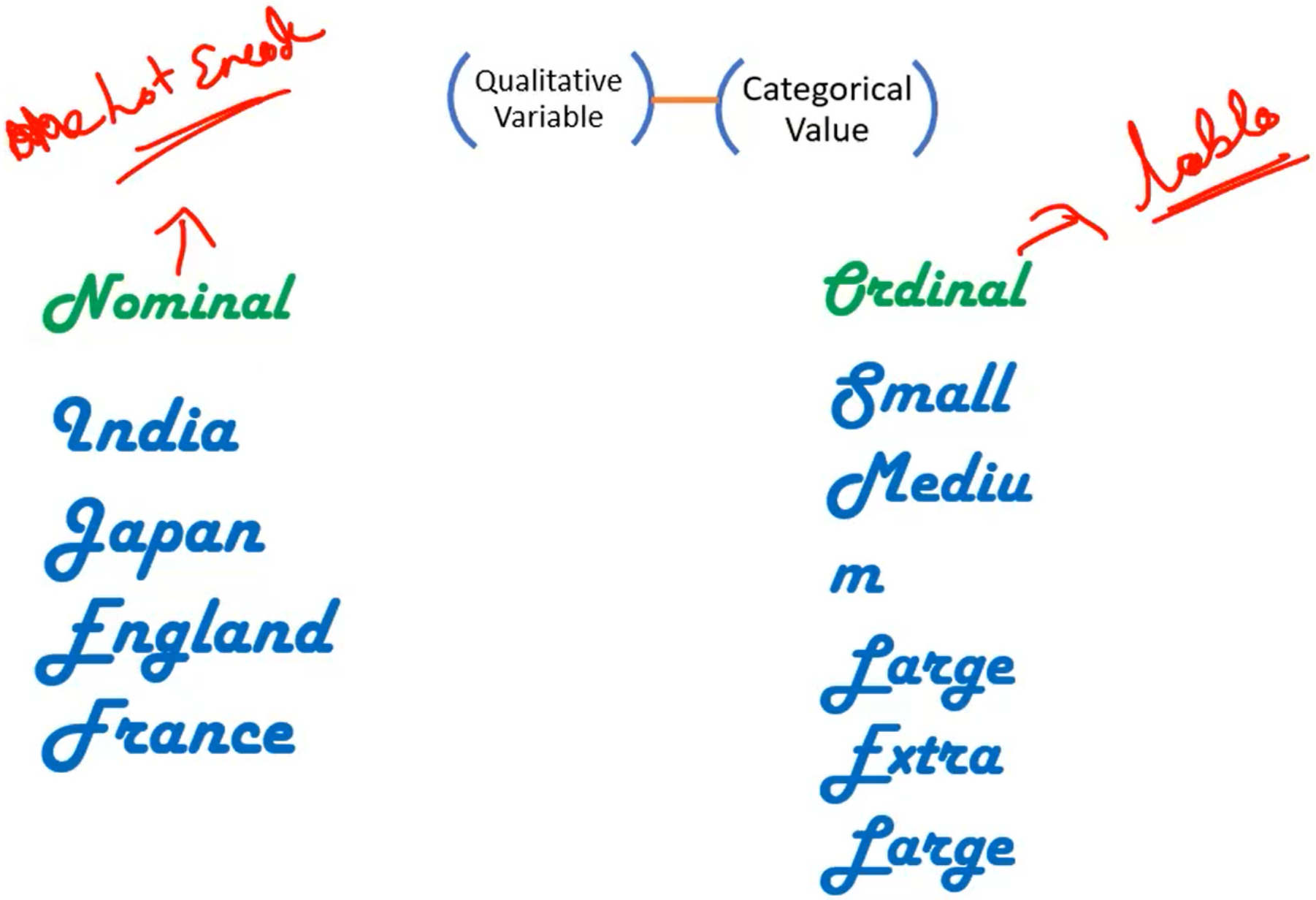
File Edit Select DNA-Seq Genotype Numeric RNA-Seq Plot Scripts Help

All: 42 x 11
Active: 42 x 11

Unsort		G 1	C 2	G 3		I 5	B 6	G 7
Map	Samples	Plate Name	Well	Sample Source	Sample Type	Internal Pico (ng/uL)	CV%	Experiment N
1	yw440_01	Example_GH	A01	FFPE_HCI	Normal	50	0	57
2	yw440_02	Example_GH	A02	FFPE_HCI	Normal	50	0	58
3	yw440_03	Example_GH	A03	FFPE_HCI	Tumor	50	0	59
4	yw440_04	Example_GH	A04	FFPE_HCI	Tumor	50	0	60
5	yw440_05	Example_GH	A05	FFPE_HCI	Tumor	50	0	61
6	yw440_06	Example_GH	A06	FFPE_HCI	Normal	50	0	62
7	yw440_07	Example_GH	A07	FFPE_HCI	Tumor	50	0	63
8	yw440_08	Example_GH	A08	FFPE_HCI	Normal	50	0	64
9	yw440_09	Example_GH	A09	FFPE_HCI	Tumor	50	0	65
10	yw440_10	Example_GH	A10	FFPE_HCI	Tumor	50	0	66
11	yw440_11	Example_GH	A11	FFPE_HCI	Tumor	50	0	67
12	yw440_12	Example_GH	B01	FFPE_HCI	Normal	50	0	68
13	yw440_13	Example_GH	B02	FFPE_HCI		50	0	69

MIP_Sample_Table Dataset - Sheet 1

TYPES OF VARIABLE



TYPES OF VARIABLE

Discrete Value

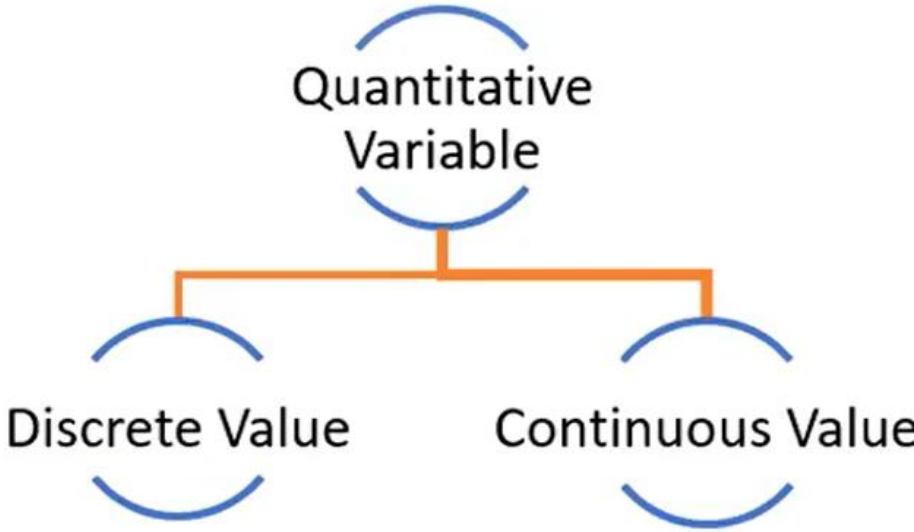
MIP_Copy_Number Dataset - Sheet 1 [7]

File Edit Select DNA-Seq Genotype Numeric RNA-Seq Plot Scripts Help

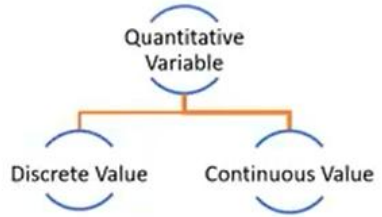
All: 330,000 x 174
Active: 330,000 x 174

Unsort		R 7	R 8	R 9	R 10	R
Map	Assay Name	yw440_01_Copy_A	yw440_01_Copy_B	yw440_01_CopyNumber	yw440_01_AlleleRatio	yw440_(
1	amp184660	2.2903	-0.001	2.2893	-0.0004	
2	amp53303	2.2717	-0.0012	2.2706	-0.0005	
3	amp364886	1.6909	-0.0071	1.6837	-0.0042	
4	amp121615	0.8527	1.0442	1.8969	0.8166	
5	amp288732	2.0576	-0.0081	2.0495	-0.0039	
6	amp358123	1.6147	-0.009	1.6057	-0.0056	
7	amp70940	2.0292	-0.0059	2.0233	-0.0029	
8	amp358884	1.6335	0.0084	1.6418	0.0051	
9	amp948	1.9056	-0.0117	1.8939	-0.0061	
10	amp336335	1.8854	-0.0031	1.8822	-0.0017	
11	amp126788	1.6479	-0.0025	1.6454	-0.0015	
12	amp145662	1.7346	-0.0013	1.7333	-0.0008	
13	amp329597	2.4909	-0.0013	2.4895	-0.0005	

MIP_Copy_Number Dataset - Sheet 1



TYPES OF VARIABLE



Continuous Value

MIP_Copy_Number Dataset - Sheet 1 [7]

File Edit Select DNA-Seq Genotype Numeric RNA-Seq Plot Scripts Help

All: 330,000 x 174
Active: 330,000 x 174

Unsort		R 7	R 8	R 9	R 10	R
Map	Assay Name	yw440_01_Copy_A	yw440_01_Copy_B	yw440_01_CopyNumber	yw440_01_AlleleRatio	yw440_(
1	amp184660	2.2903	-0.001	2.2893	-0.0004	
2	amp53303	2.2717	-0.0012	2.2706	-0.0005	
3	amp364886	1.6909	-0.0071	1.6837	-0.0042	
4	amp121615	0.8527	1.0442	1.8969	0.8166	
5	amp288732	2.0576	-0.0081	2.0495	-0.0039	
6	amp358123	1.6147	-0.009	1.6057	-0.0056	
7	amp70940	2.0292	-0.0059	2.0233	-0.0029	
8	amp358884	1.6335	0.0084	1.6418	0.0051	
9	amp948	1.9056	-0.0117	1.8939	-0.0061	
10	amp336335	1.8854	-0.0031	1.8822	-0.0017	
11	amp126788	1.6479	-0.0025	1.6454	-0.0015	
12	amp145662	1.7346	-0.0013	1.7333	-0.0008	
13	amp329597	2.4909	-0.0013	2.4895	-0.0005	

MIP_Copy_Number Dataset - Sheet 1

Discrete Value

MIP_Sample_Table Dataset - Sheet 1 [4]

File Edit Select DNA-Seq Genotype Numeric RNA-Seq Plot Scripts Help

All: 42 x 11
Active: 42 x 11

Unsort		G 1	C 2	G 3	C 4	I 5	B 6	G 7
Map	Samples	Plate Name	Well	Sample Source	Sample Type	Internal	CV%	Experiment N
1	yw440_01	Example_GH	A01	FFPE_HCI	Normal	50	0	57
2	yw440_02	Example_GH	A02	FFPE_HCI	Normal	50	0	58
3	yw440_03	Example_GH	A03	FFPE_HCI	Tumor	50	0	59
4	yw440_04	Example_GH	A04	FFPE_HCI	Tumor	50	0	60
5	yw440_05	Example_GH	A05	FFPE_HCI	Tumor	50	0	61
6	yw440_06	Example_GH	A06	FFPE_HCI	Normal	50	0	62
7	yw440_07	Example_GH	A07	FFPE_HCI	Tumor	50	0	63
8	yw440_08	Example_GH	A08	FFPE_HCI	Normal	50	0	64
9	yw440_09	Example_GH	A09	FFPE_HCI	Tumor	50	0	65
10	yw440_10	Example_GH	A10	FFPE_HCI	Tumor	50	0	66
11	yw440_11	Example_GH	A11	FFPE_HCI	Tumor	50	0	67
12	yw440_12	Example_GH	B01	FFPE_HCI	Normal	50	0	68
13	yw440_13	Example_GH	B02	FFPE_HCI	Normal	50	0	69

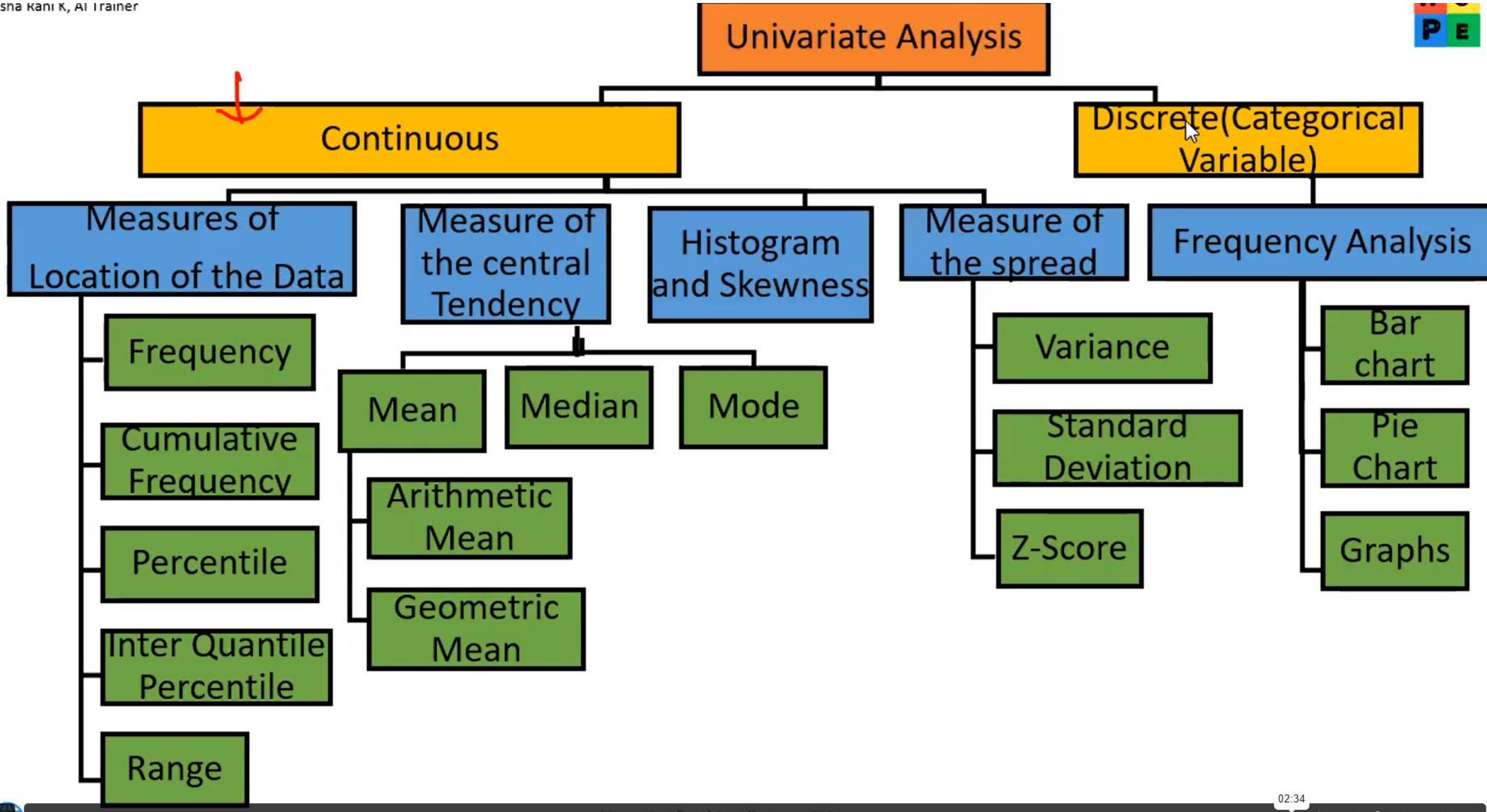
MIP_Sample_Table Dataset - Sheet 1

Null –nothing
Non null- something present
NaN=Nothing

Datatype

Int64 = whole number no decmeical point
Float64= decimal –continusou value
Object=categorical data
Char= categorical data

Quan- int, float
Qual=object,chara

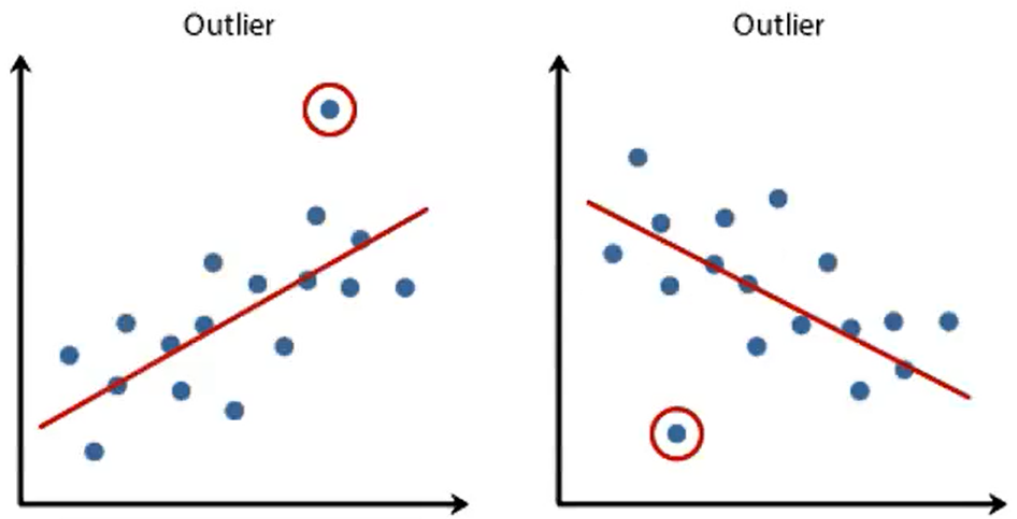

Univariate Analysis
Continuous
Discrete(Categorical Variable)
Measures of Location of the Data
Measure of the central Tendency
Histogram and Skewness
Measure of the spread
Frequency Analysis
Frequency
Cumulative Frequency
Percentile
Inter Quantile Percentile
Range
Mean
Median
Mode
Arithmetic Mean
Geometric Mean
Variance
Standard Deviation
Z-Score
Bar chart
Pie Chart
Graphs

- **Outlier = an “odd” data point that is very different from the rest.**

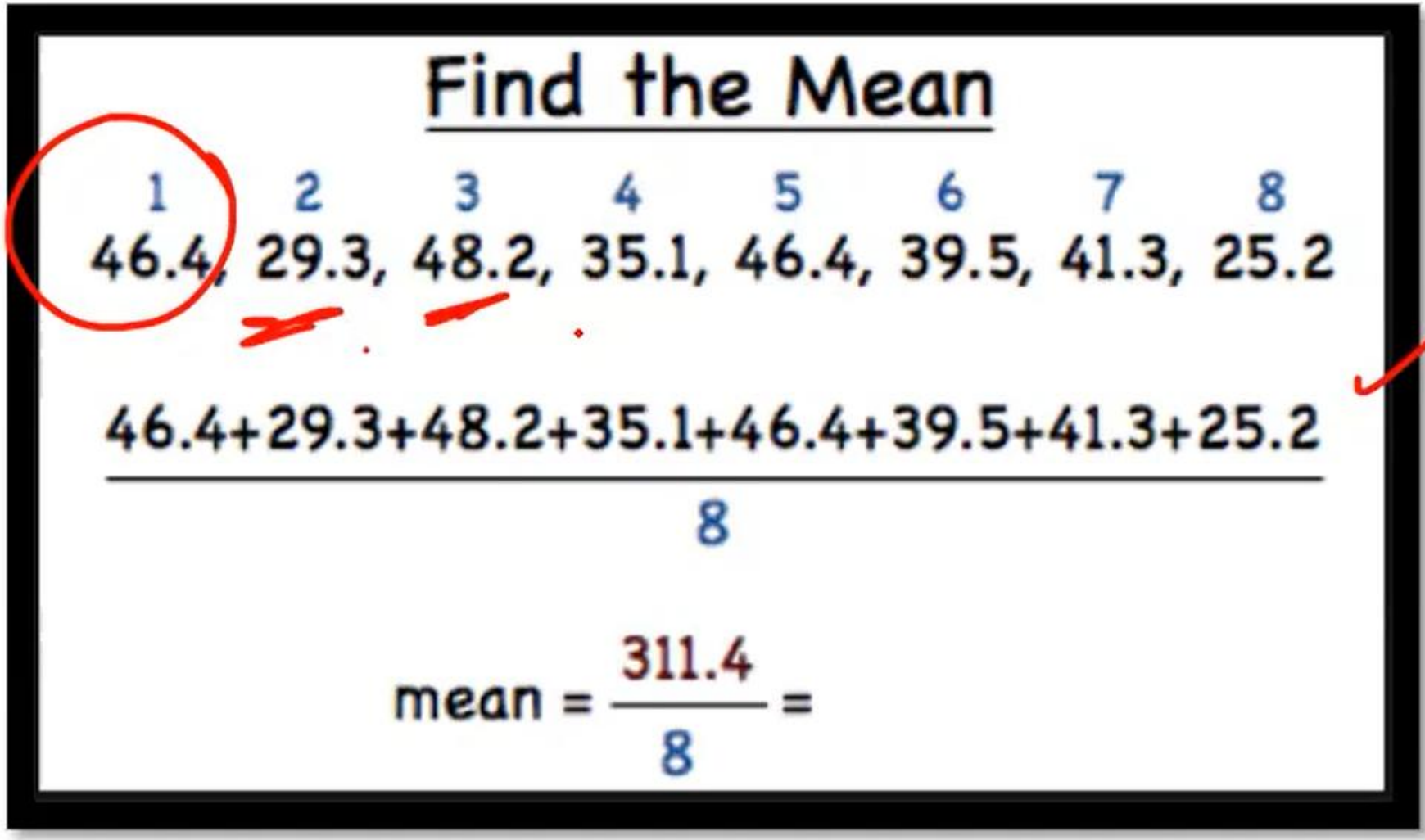
•It’s much **higher** or **lower** than most of the other values.

•Can happen due to error (wrong entry) or genuine unusual cases.

What Is Outlier?



Central Tendency: Mean |Arithmetic Mean



Central Tendency: Mean |Arithmetic Mean

What is Mean?

Average which gives overall center value.

When to use mean?

When you are have to report about overall performance of a task.

Example,

Overall performance of a student in a class

To fill missing value in data preprocessing

Formula:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + \lambda + X_n}{n}$$

Types of Mean:

Population Mean	Sample Mean
$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
N = number of items in the population	n = number of items in the sample

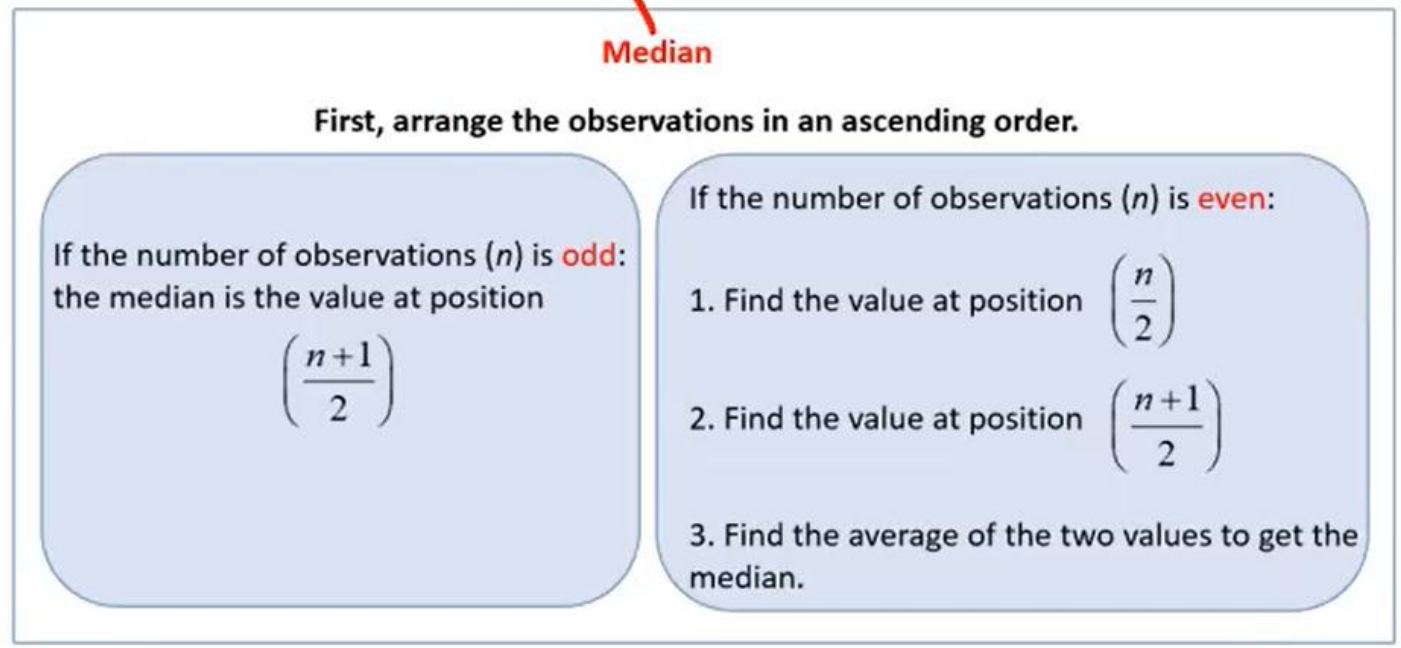
Central Tendency: Median

What is Median?

Midpoint of data.

When to use Median?

- When outlier exists in nominal , interval and ratio variable,
- we can use median instead of mean.



25,30,45,50,200

Mean **Outlier not omit**

Human error
Data behavior

25+30+45+50+200/5= 70

Median

salary

Outlier will omit

25,30,45,50,200=45

Outlier = an “odd” data point that is very different from the rest.

•It’s much **higher** or **lower** than most of the other values.

Mean (average)

•Add up all the numbers, divide by how many numbers.

•Example: [2, 3, 4, 5, 6]

$$\text{Mean} = (2 + 3 + 4 + 5 + 6)/5 = 20/5 = 4$$

👉 Mean is **affected much by outliers**

Example: [2, 3, 4, 5, 100] → Mean = 22.8 (not really “typical”).

🔹 **Median (middle value)**

•Arrange numbers in order, pick the **middle one**.

•Example: [2, 3, 4, 5, 6] → Median = 4

•If even count, take the average of the two middle numbers.

Example: [2, 3, 4, 5] → Median = (3+4)/2 = 3.5

👉 Median is **not affected much by outliers**.

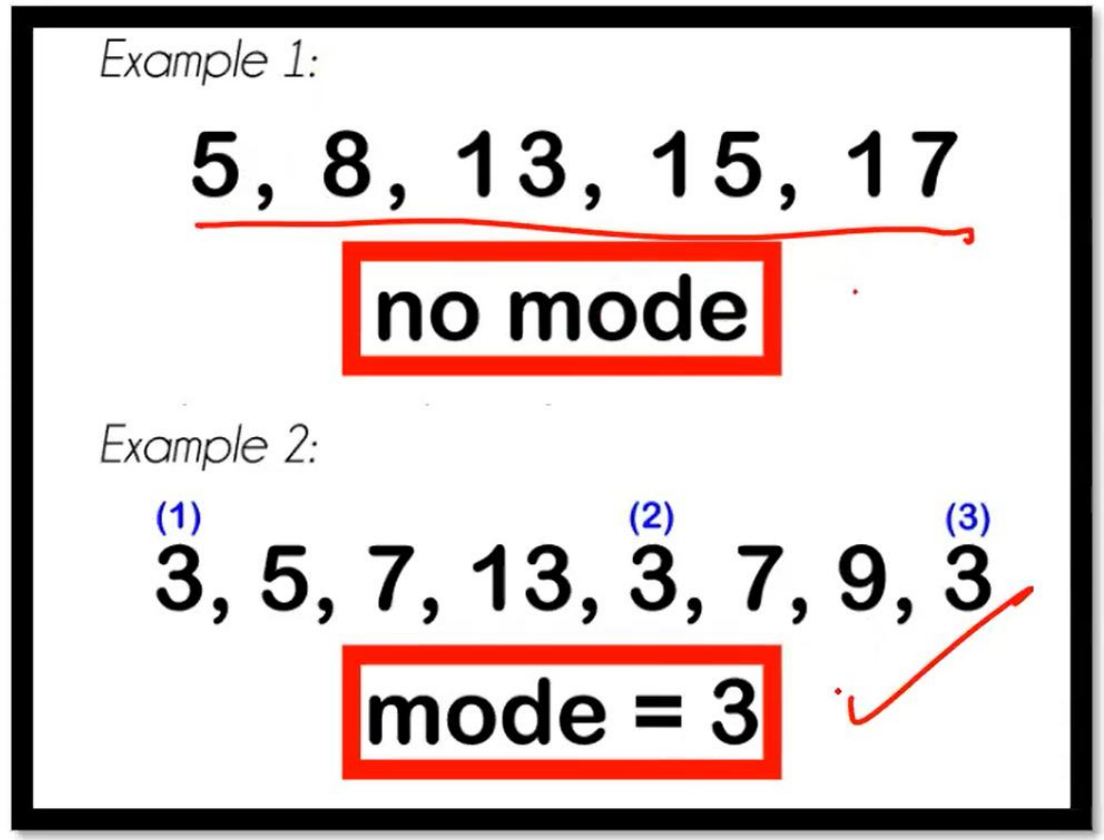
Example: [2, 3, 4, 5, 100] → Median = 4 (still typical).

Central Tendency: Mode

What is mode?
Most repeated data point.

When to use mode?
To find out most repeated value of a dataset.

Example,
To finalize the recognized face , most repeated detected face can be concluded as detected face



Central Tendency: Mode

Example 1:

5, 8, 13, 15, 17

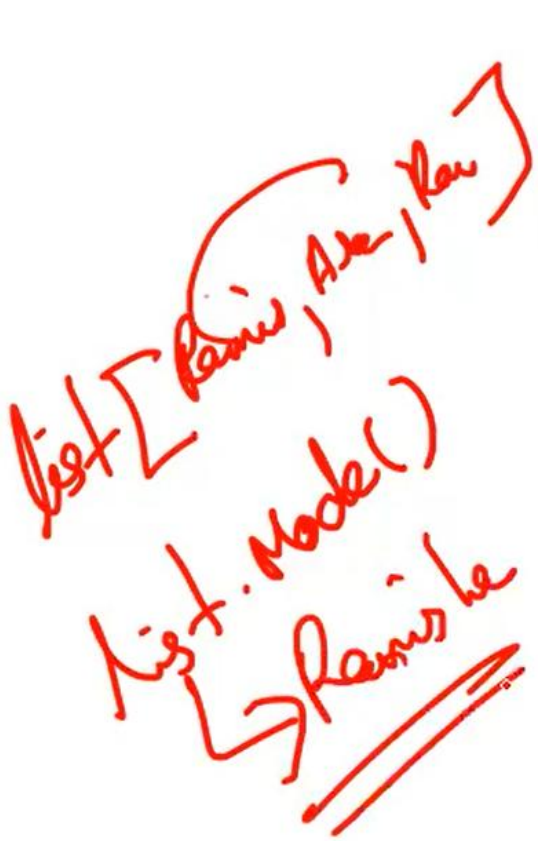
no mode

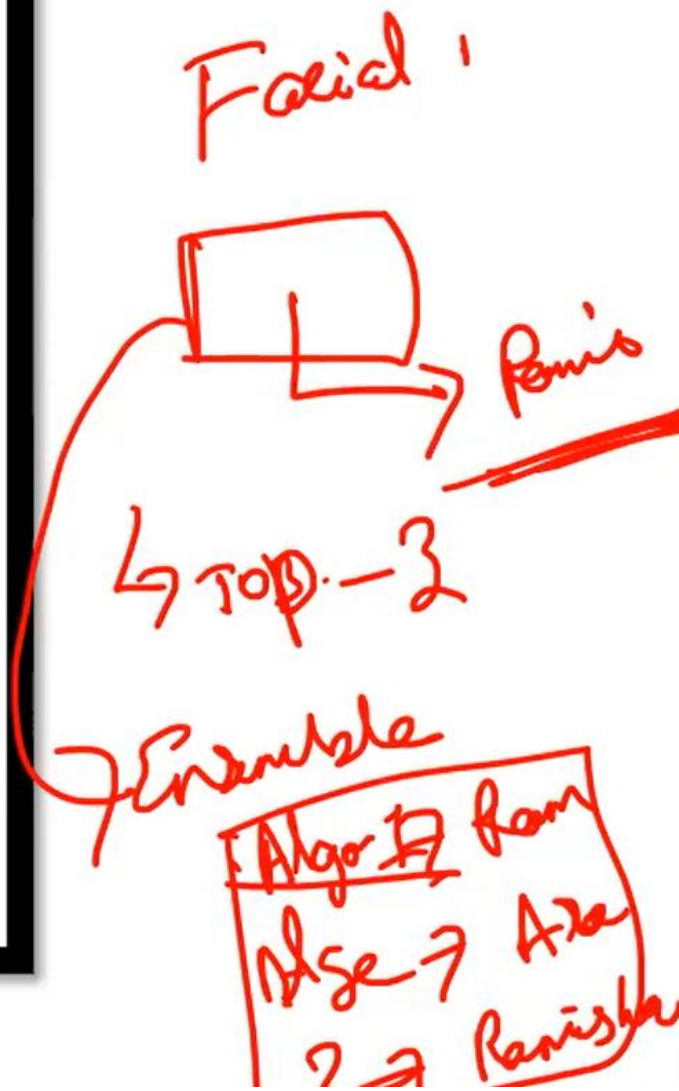
Example 2:

(1) (2) (3)

3, 5, 7, 13, 3, 7, 9, 3

mode = 3



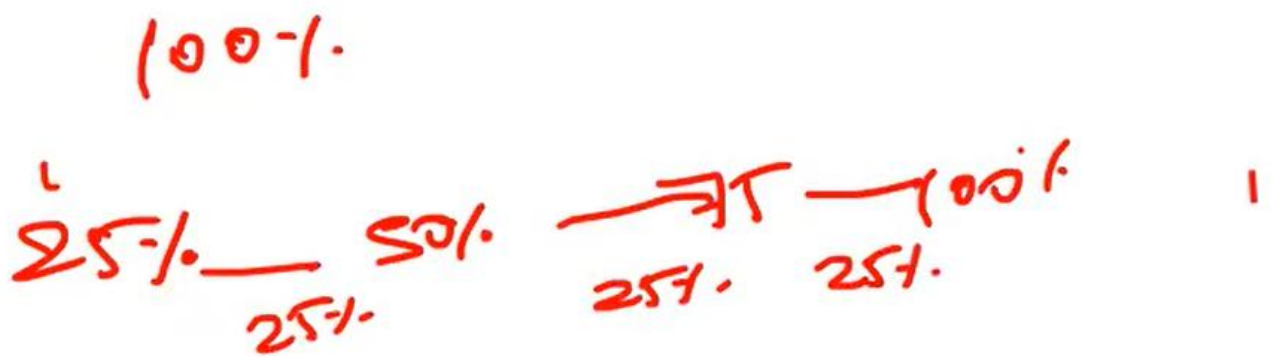


MEASURE OF LOCATION OF THE DATA

What is Percentile?

Percentile tells about the value exist within the range

Dividing whole dataset into four parts 25^{th},50^{th},75^{th} ,100^{th} in terms of percentage. Each part of the percentile boundary will be calculated.



MEASURE OF LOCATION OF THE DATA

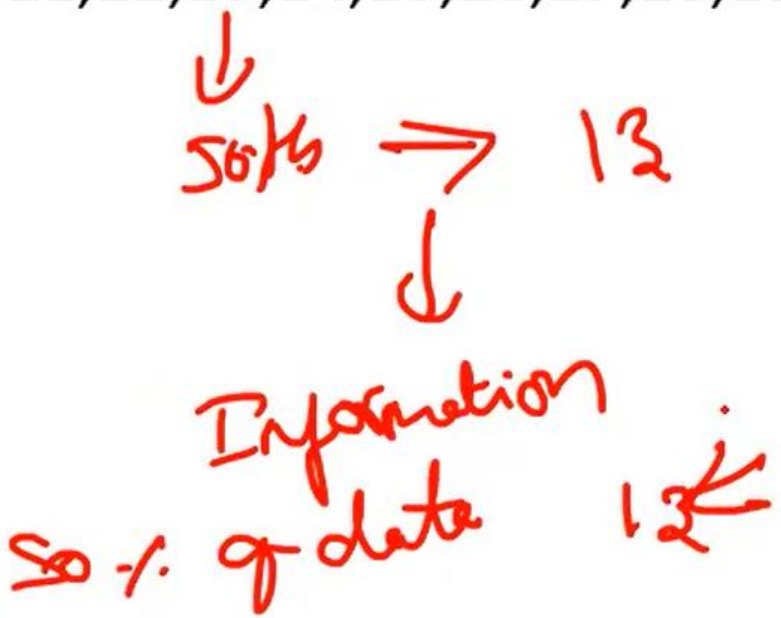
What is Percentile?

Dividing whole dataset into four parts 25^{th},50^{th},75^{th} ,100^{th} in terms of percentage. For each part percentile boundary will be calculated.

MEASURE OF LOCATION OF THE DATA

Original Dataset=[1,4,5,6, 2,3,7,8, 11,12, 9,10,13,14, 17,18,19,20,15,16]

Sorted Dataset=[1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,18,19,20]



MEASURE OF LOCATION OF THE DATA

Find the value for X^{th} percentile?

A Formula for Finding the *k*th Percentile

If you were to do a little research, you would find several formulas for calculating the k^{th} percentile. Here is one of them.

k = the k^{th} percentile. It may or may not be part of the data.

i = the index (ranking or position of a data value)

n = the total number of data points, or observations

- Order the data from smallest to largest.
- Calculate $i = \frac{k}{100}(n+1)$
- If i is an integer, then the k^{th} percentile is the data value in the i^{th} position in the ordered set of data.
- If i is not an integer, then round i up and round i down to the nearest integers. Average the two data values in these two positions in the ordered data set. This is easier to understand in an example.

MEASURE OF LOCATION OF THE DATA

Find the value for X^{th} percentile?

Listed are 29 ages for Academy Award winning best actors *in order from smallest to largest.*

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Find the 70^{th} percentile.

b. Find the 83^{rd} percentile.

70% of data value exist between 18 to 64

25% of data value exist before 29

Descriptive Analysis of Student Performance and Salary
Calculation of Central Tendency (Mean, Median, Mode) and Percentiles (Q1, Q2, Q3, Q4)

Python – Code

```
import pandas as pd
import numpy as np
#data collection
dataset=pd.read_csv("Placement.csv")
#collect Quan & Qual data
quan=[x for x in dataset if dataset[x].dtype!=object]
qual=[x for x in dataset if dataset[x].dtype==object]
#remove sl_no
quan.remove("sl_no")

#create New Table for central Tendency & descriptive
descriptive=pd.DataFrame(index=["Mean","Median","Mode",
 ↪"Q0-Min-0%","Q1-25%","Q2-50%","Q3-75%","Q4-Max-100%",
                                "min-max","0-25%","25%-50%","50%-75%","75%-100%"]
                         ,columns=quan)
for x in quan:
    #print(x)
    descriptive.loc["Mean", x] = round(dataset[x].mean(), 2)
    descriptive.loc["Median", x] = dataset[x].median()
    descriptive.loc["Mode", x] = round(dataset[x].mode()[0], 2)
    descriptive.loc["Q0-Min-0%",x]=dataset.describe().loc['min',x]
    descriptive.loc["Q1-25%",x]=dataset.describe().loc['25%',x]
    descriptive.loc["Q2-50%",x]=dataset.describe().loc['50%',x]
    descriptive.loc["Q3-75%",x]=dataset.describe().loc['75%',x]
    descriptive.loc["Q4-Max-100%",x]=dataset.describe().loc['max',x]
    descriptive.loc["min-max",x]=dataset.describe().loc['max',x]-dataset.
 ↪describe().loc['min',x]
    descriptive.loc["0-25%",x]=dataset.describe().loc['25%',x]-dataset.
 ↪describe().loc['min',x]
    descriptive.loc["25%-50%",x]=dataset.describe().loc['50%',x]-dataset.
 ↪describe().loc['25%',x]
    descriptive.loc["50%-75%",x]=dataset.describe().loc['75%',x]-dataset.
 ↪describe().loc['50%',x]
    descriptive.loc["75%-100%",x]=dataset.describe().loc['max',x]-dataset.
 ↪describe().loc['75%',x]
descriptive
```

Python – Result

	ssc_p	hsc_p	degree_p	etest_p	mba_p	salary
Mean	67.3	66.33	66.37	72.1	62.28	288655.41
Median	67.0	65.0	66.0	71.0	62.0	265000.0
Mode	62.0	63.0	65.0	60.0	56.7	300000.0
Q0-Min-0%	40.89	37.0	50.0	50.0	51.21	200000.0
Q1-25%	60.6	60.9	61.0	60.0	57.945	240000.0
Q2-50%	67.0	65.0	66.0	71.0	62.0	265000.0
Q3-75%	75.7	73.0	72.0	83.5	66.255	300000.0
Q4-Max-100%	89.4	97.7	91.0	98.0	77.89	940000.0
min-max	48.51	60.7	41.0	48.0	26.68	740000.0
0-25%	19.71	23.9	11.0	10.0	6.735	40000.0
25%-50%	6.4	4.1	5.0	11.0	4.055	25000.0
50%-75%	8.7	8.0	6.0	12.5	4.255	35000.0
75%-100%	13.7	24.7	19.0	14.5	11.635	640000.0

```
outliers = data[(data < lower_bound) | (data > upper_bound)]
```

Python – Result
Central tendency (mean, median, Mode)
Percentile (Q1,Q2,Q3,Q4)

Measure	ssc_p	hsc_p	degree_p	etest_p	mba_p	salary
Mean	67.30	66.33	66.37	72.10	62.28	288655.41
Median	67.00	65.00	66.00	71.00	62.00	265000.00
Mode	62.00	63.00	65.00	60.00	56.70	300000.00
Q0 (Min-0%)	40.89	37.00	50.00	50.00	51.21	200000.00
Q1 (25%)	60.60	60.90	61.00	60.00	57.95	240000.00
Q2 (50%)	67.00	65.00	66.00	71.00	62.00	265000.00
Q3 (75%)	75.70	73.00	72.00	83.50	66.26	300000.00
Q4 (Max-100%)	89.40	97.70	91.00	98.00	77.89	940000.00

Whole Class Summary of Result (Mark Based)

Whole Class summary	10th	12th	Degree	Enterance test	MBA Degree
Average class percentage (Mean):	67.3	66.33	66.37	72.1	62.28
Middle class percentage (Median):	67	65	66	71	62
Most common class percentage (Mode):	62	63	65	60	56.7
Minium class percentage(Q0,min-0%):	40.89	37	50	50	51.21
25% percentage of class percentage(Q1,25%):	60.6	60.9	61	60	57.95
50% percentage of class percentage(Q2, 50%):	67	65	66	71	62
75% percentage of class percentage(Q3, 75%):	75.7	73	72	83.5	66.26
Maximum class percentage(Q4, max-100%)	89.4	97.7	91	98	77.89
From Minimum – to maximum percentage different:	48.51	60.7	41	48	26.68
0-25% class percentage different	19.71	23.9	11	10	6.74
25%-50% class percentage different:	6.4	4.1	5	11	4.06
50%-75% class percentage different:	8.7	8	6	12.5	4.26
75%-100% class percentage different :	13.7	24.7	19	14.5	11.64

Whole Class Summary of Result (Salary Based)

Whole Class salary summary	Salary
Average class salary (Mean):	288655.41
Middle salary of class (Median):	265000
Most common class salary (Mode):	300000
Minium class salary (Q0,min-0%):	200000
25% percentage of class salary (Q1,25%):	240000
50% percentage of class salary (Q2, 50%):	265000
75% percentage of class salary (Q3, 75%):	300000
Maximum class salary (Q4, max-100%)	940000
From Minimum – to maximum salary different:	740000
0-25% class salary different	40000
25%-50% class salary different:	25000
50%-75% class salary different:	35000
75%-100% class salary different :	640000

Lesser Outlier & Greater Outlier

Lesser Outlier

5,45,60,65,75,78,87,90

5,45,60,65,75,78,87,90

Greater Outlier

45,60,65,75,78,87,90,400

MEASURE OF LOCATION OF THE DATA

Interquartile Range(IQR)

- What is the purpose of IQR?

  To know the outlier range present in the dataset.

100

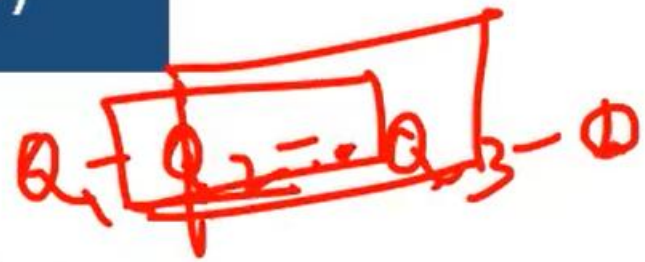
75% - 25%

IQR = Q3-Q1

Interquartile Range(IQR)

- What is the purpose of IQR?

To know the outlier range present in the dataset.



IQR = Q3-Q1

Lesser Outlier

Less Than

Outlier range= Q1- 1.5*IQR

Greater Outlier

Greater Than

Outlier range= Q3+ 1.5*IQR

outlier coming reason-
Human typo error
nature coming (normal behaviour)

, 25, 30, 35, 40, 45, 90])
and IQR
)
)

Reason for 1.5 × IQR rule

- The **IQR (Q3 − Q1)** covers the **middle 50%** of the data.
- Multiplying by **1.5** extends this range to about **99% of a normal (bell-shaped) distribution**.
- This makes the rule a **balance**:
  - Not too strict (so normal values aren't wrongly flagged).
  - Not too loose (so extreme values are still caught).

📊 **Intuition**

- Values within **Q1 − 1.5×IQR** and **Q3 + 1.5×IQR** are considered **reasonable spread**.
- Anything outside is **too far away from the bulk of the data** → called an **outlier**.

🔹 **Variations**

- **Mild outliers:** Outside 1.5 × IQR
- **Extreme outliers:** Outside **3 × IQR**

- So, **1.5** is not magic, it's a **commonly accepted statistical cutoff** that works well in practice (especially in boxplots).

Frequency | Relative Frequency | Cumulative Relative Frequency

Twenty students were asked how many hours they worked per day. Their responses, in hours, are as follows: 56332475235654435253.

Data value	Frequency
2	3
3	5
4	3
5	6
6	2
7	1

How many students have spent 2 hours per day?

Frequency | Relative Frequency | Cumulative Relative Frequency

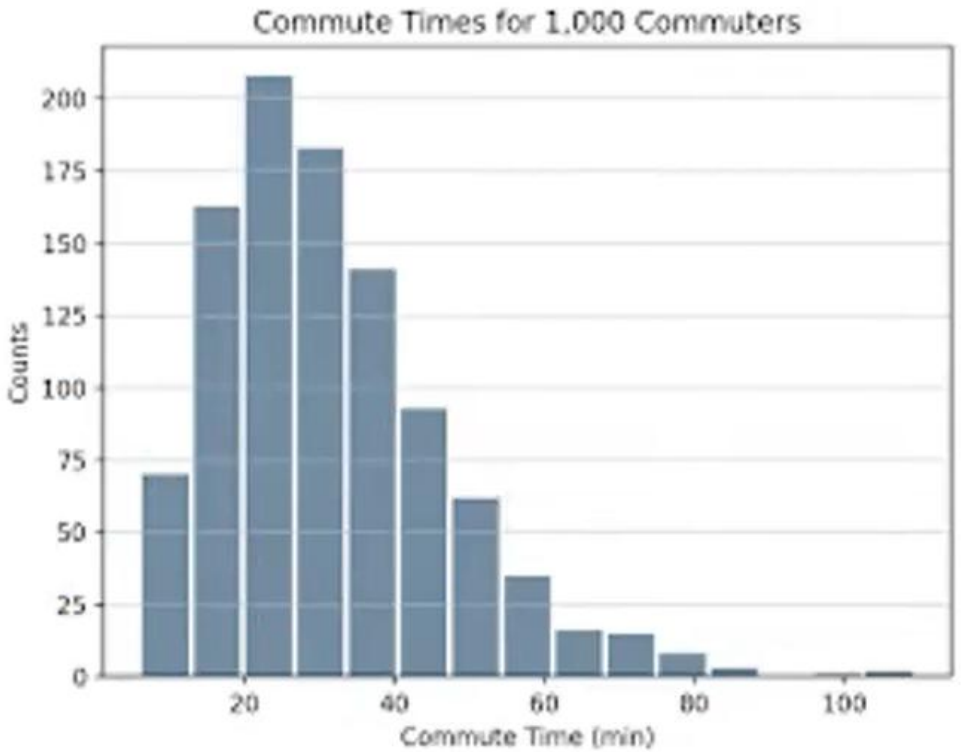
Twenty students were asked how many hours they worked per day. Their responses, in hours, are as follows: 56332475235654435253.

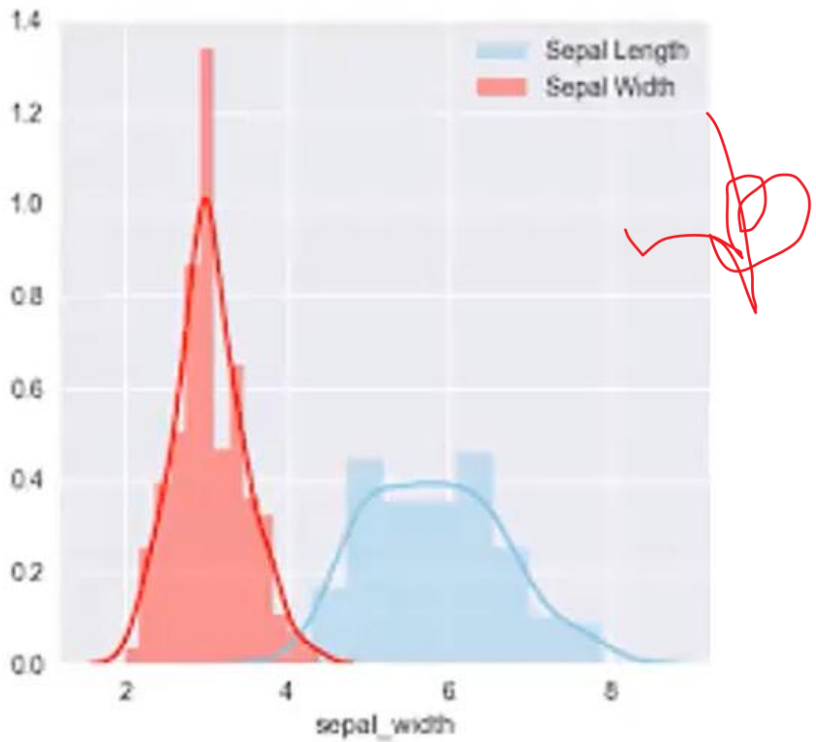
Data value	Frequency	Relative frequency	Cumulative relative frequency
2	3	$\frac{3}{20}$ or 0.15	0.15
3	5	$\frac{5}{20}$ or 0.25	0.15 + 0.25 = 0.40
4	3	$\frac{3}{20}$ or 0.15	0.40 + 0.15 = 0.55
5	6	$\frac{6}{20}$ or 0.30	0.55 + 0.30 = 0.85
6	2	$\frac{2}{20}$ or 0.10	0.85 + 0.10 = 0.95
7	1	$\frac{1}{20}$ or 0.05	0.95 + 0.05 = 1.00

What is the Cumulative Frequency of 5 hours?

Histogram

Freq Visual



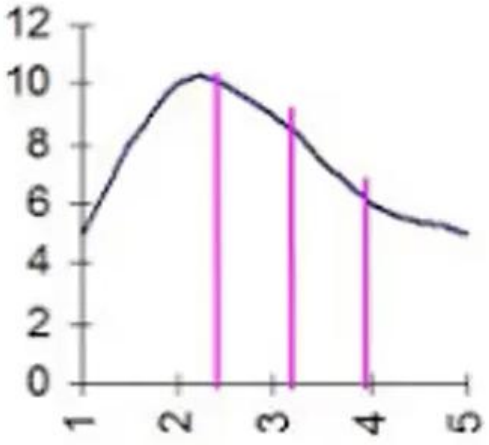


Rani K, AI Trainer

Skewness

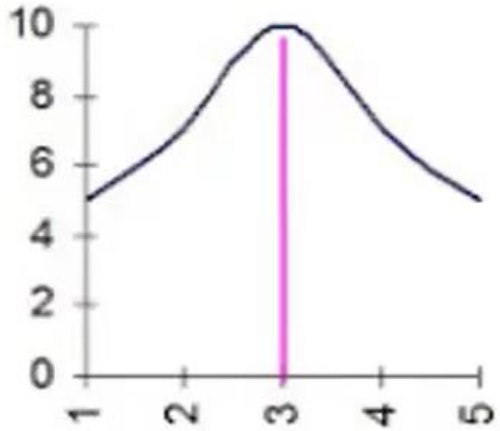
Peekness of mean median mode

Skewness is a measure of symmetry, or more precisely, the lack of symmetry.



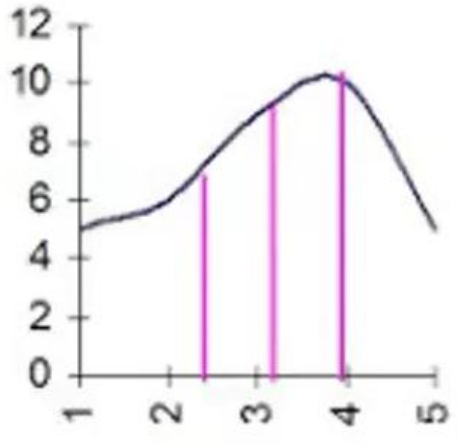
Mode >Med> Mean

positive



Mean = Median = Mode

Normal



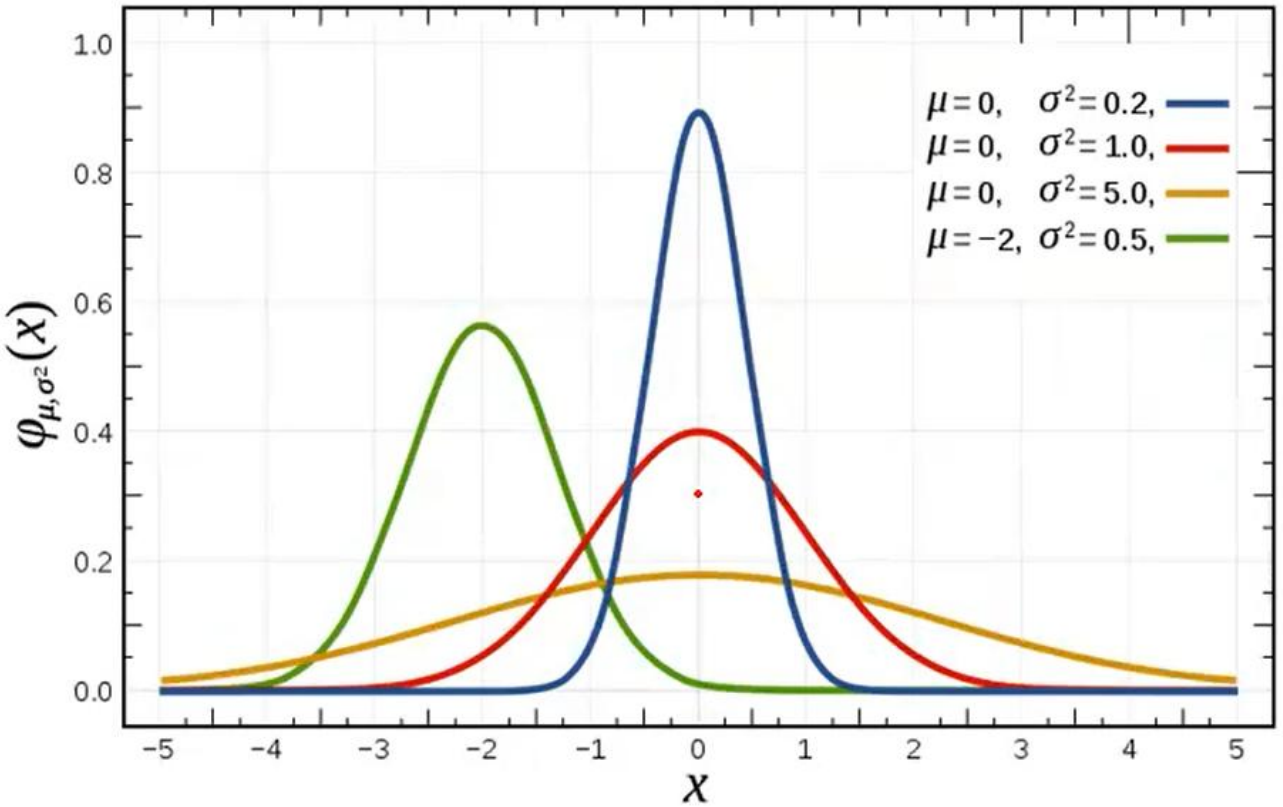
Mean< Med<Mode

Negative

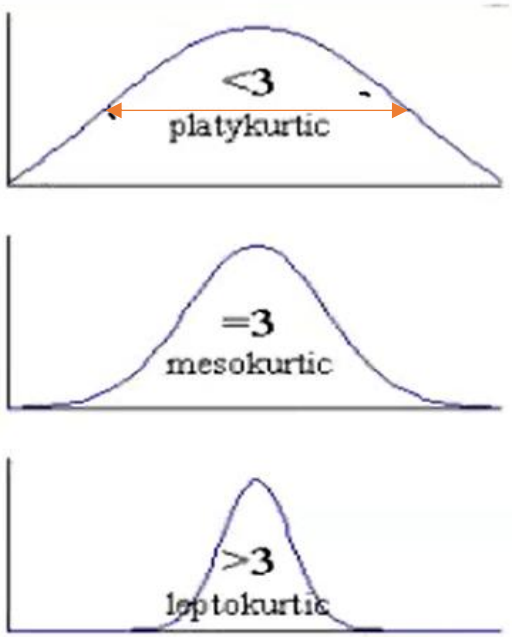
$$Skew = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{(X_i - \bar{X})}{\sigma}\right]^3$$

Kurtosis

A measure of the peakness or convexity of a curve is known as Kurtosis.



Happennig of peakness length



$$Kurt = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{(X_i - \bar{X})}{\sigma}\right]^4$$

Measure of spread

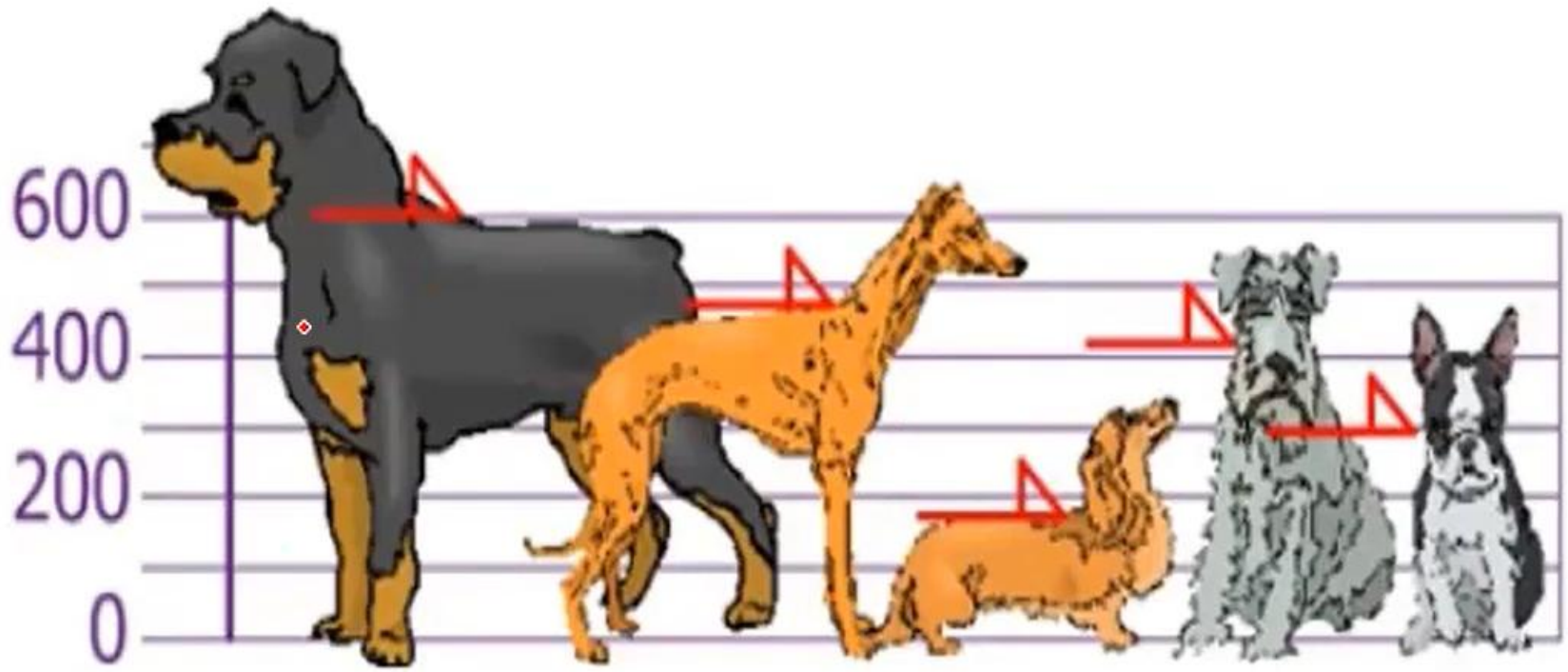
Variance

Standard Deviation

Mean → sum
Stand → Bevim

Measure of spread |Standard Deviation

Find out the Mean, the Variance, and the Standard Deviation



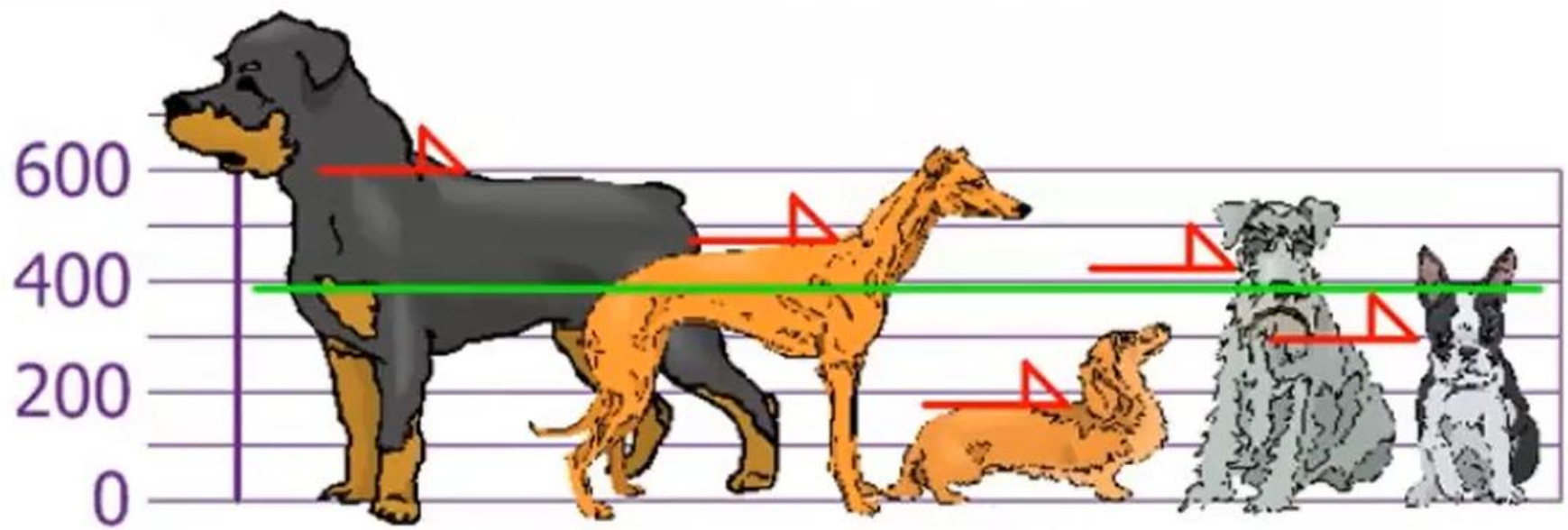
The heights (at the shoulders) are:
600mm, 470mm, 170mm, 430mm and 300mm

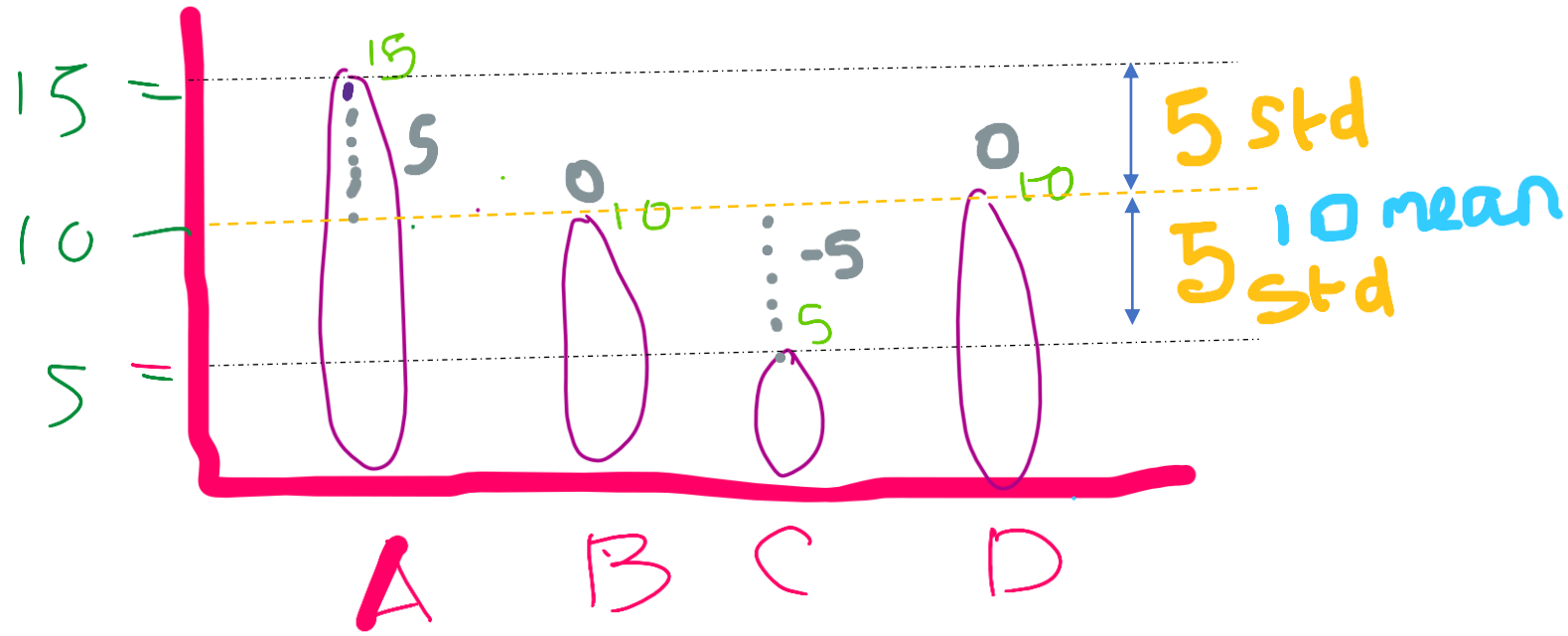
Measure of spread |Standard Deviation

Find MEAN

$$\text{Mean} = \frac{600 + 470 + 170 + 430 + 300}{5} = \frac{1970}{5} = 394$$

So Mean (average) height is 394 mm





$$\text{Mean} = \frac{15+10+5+10}{4} = 10$$

$$\text{Variant} = \frac{5^2+0^2+(-5)^2+(0)^2}{4} = 25$$

$$\text{Std} = \sqrt{25} = 5$$

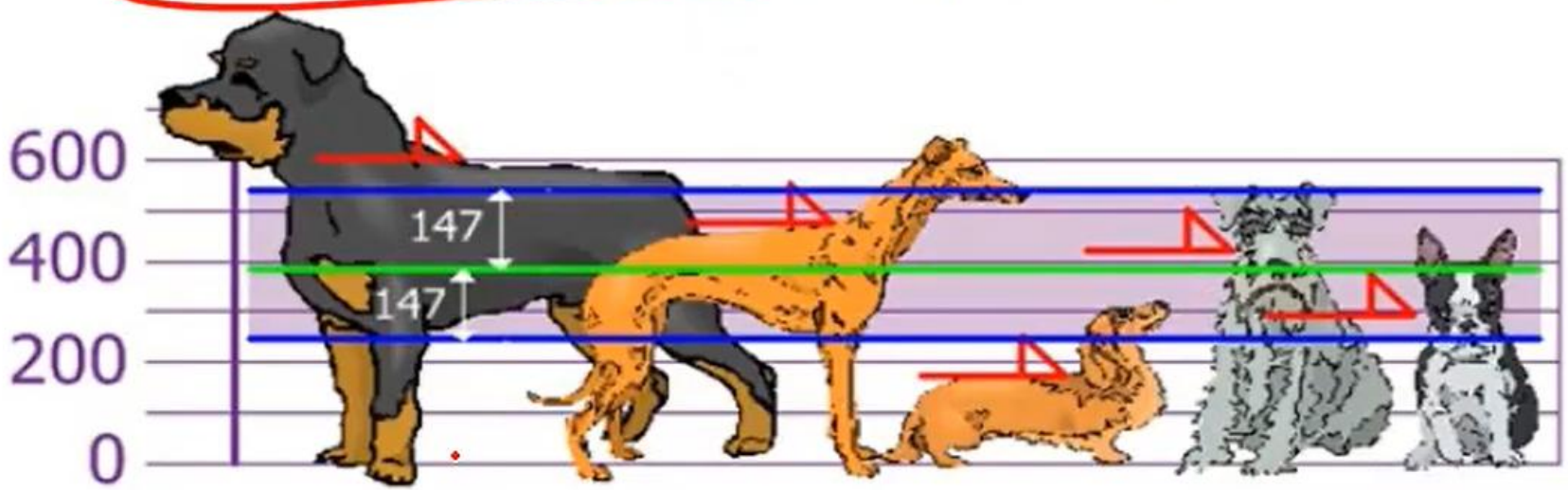
Measure of spread |Standard Deviation

STANDARD DEVIATION (σ)

Standard Deviation= square root of Variance

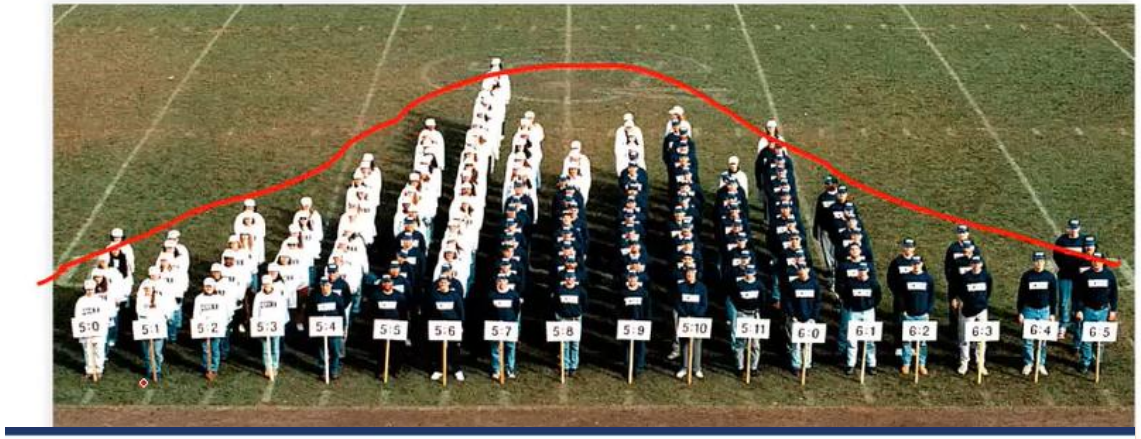
σ = √21,704 = 147.32... = 147mm

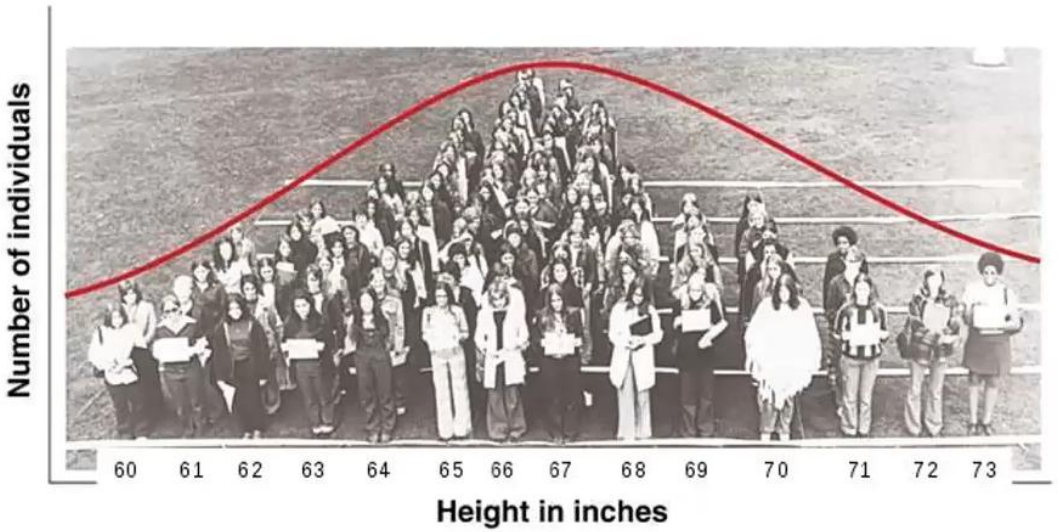
Standard Deviation is useful. Now we can show which heights are within one Standard Deviation (147mm) of the Mean:

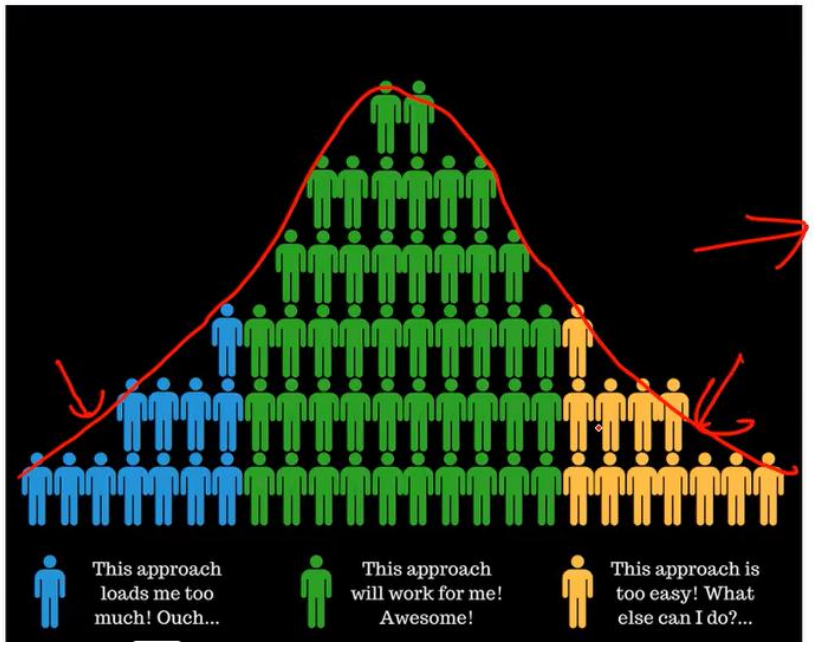


Rottweilers **are** tall dogs and Dachshunds **are** a bit short

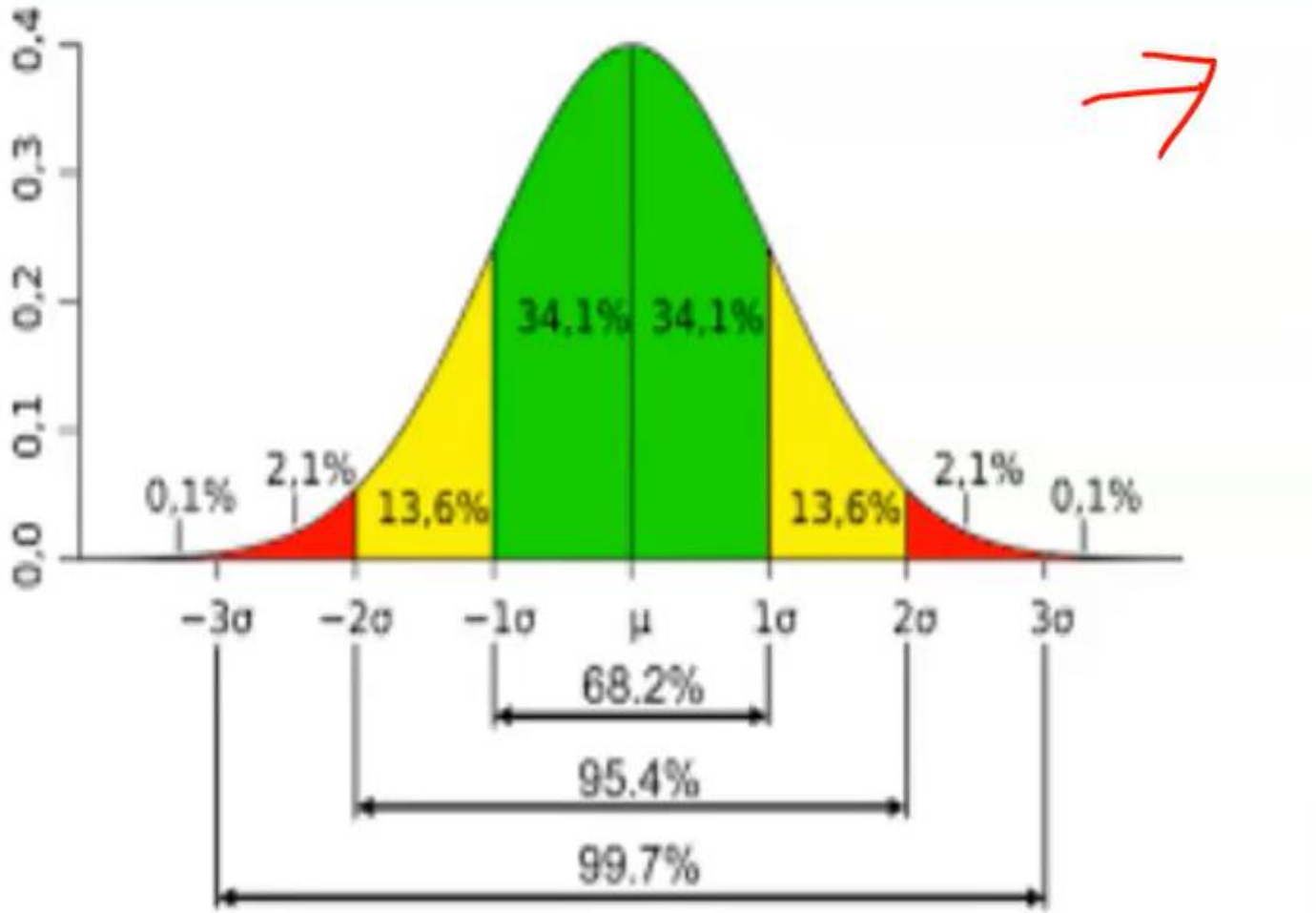
Normal Distribution







Normal Distribution



Measure of Spread| Normal Distribution

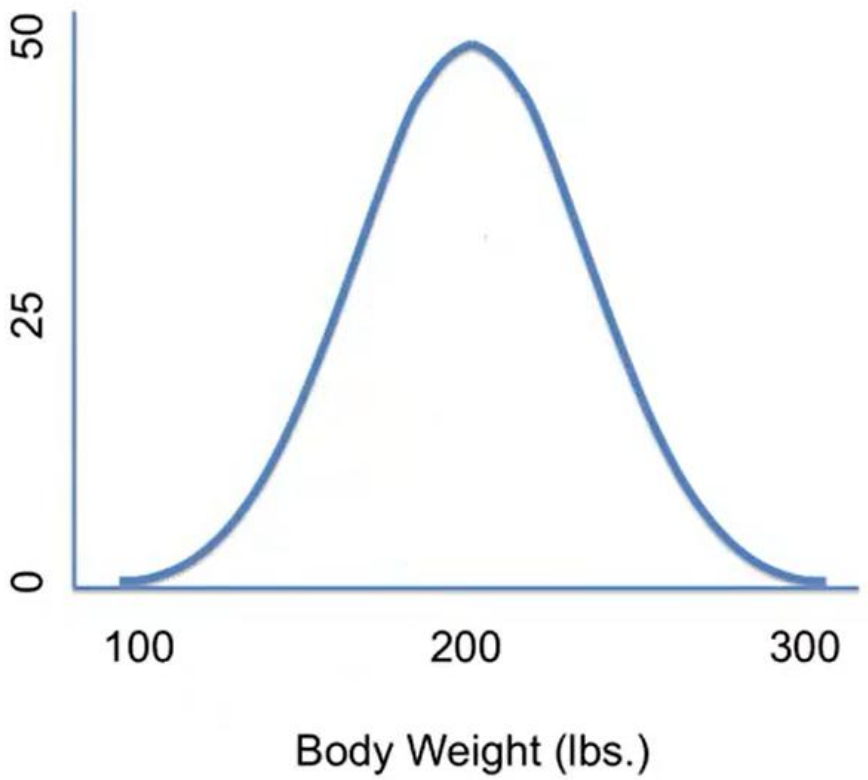
- Approximately 68% of the data is within one standard deviation of the mean.
- Approximately 95% of the data is within two standard deviations of the mean.
- More than 99% of the data is within three standard deviations of the mean.
- This is known as the Empirical Rule.
- It is important to note that this rule only applies when the shape of the distribution of the data is bell-shaped and symmetric. We will learn more about this when studying the "Normal" or "Gaussian" probability distribution in later chapters.

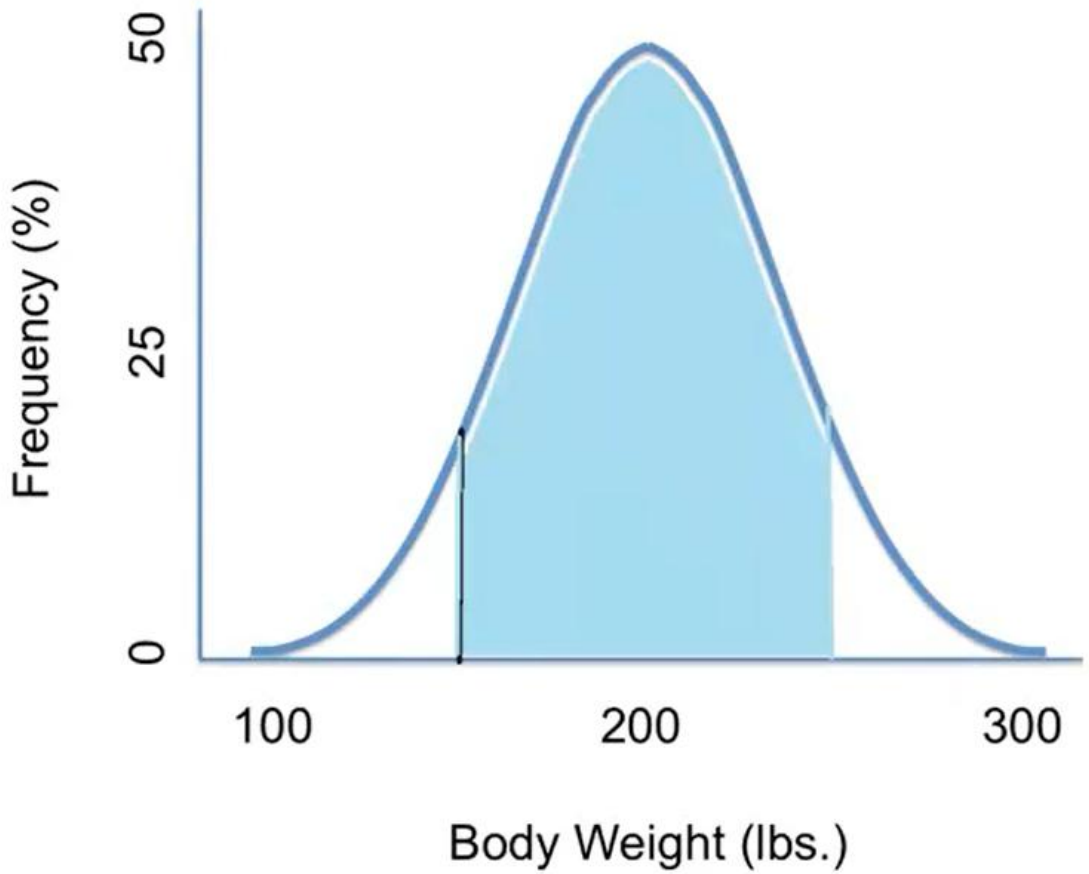
Probability Density Funtion

What is the probability that a person will weigh between 150 and 250 ?

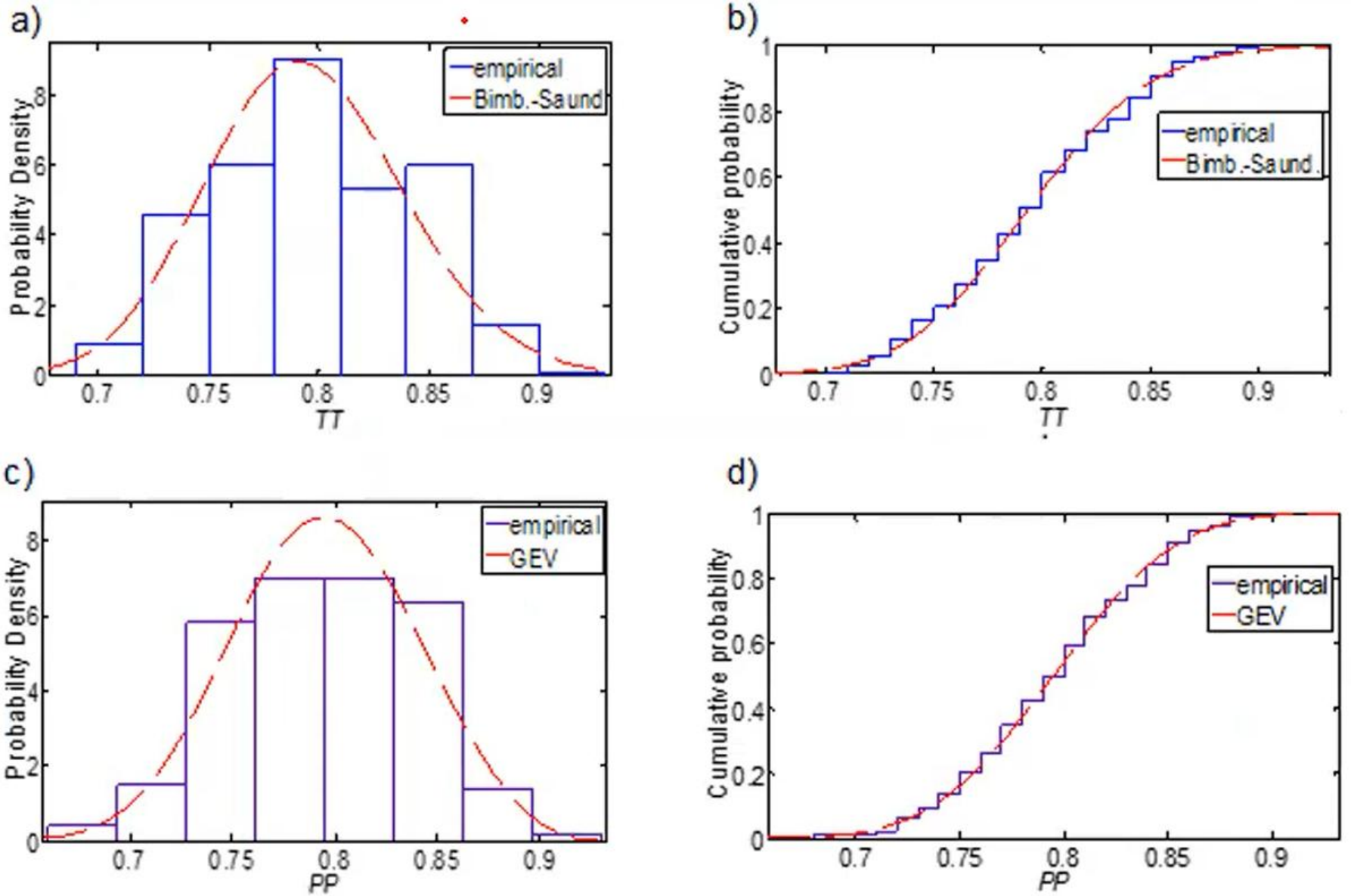
Written in notation, the question becomes:

$P(150 < Y < 250)$



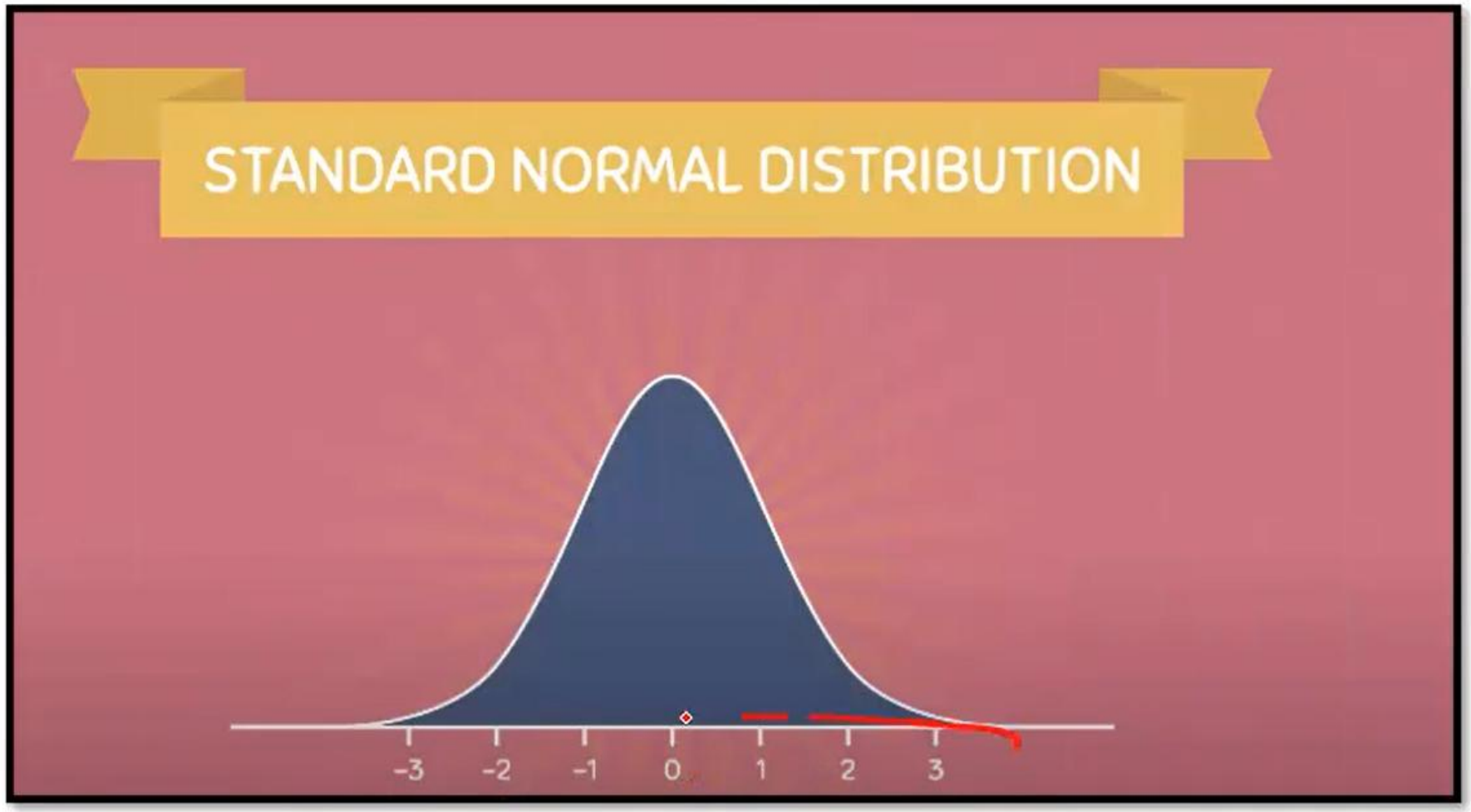


Cumulative Density Function



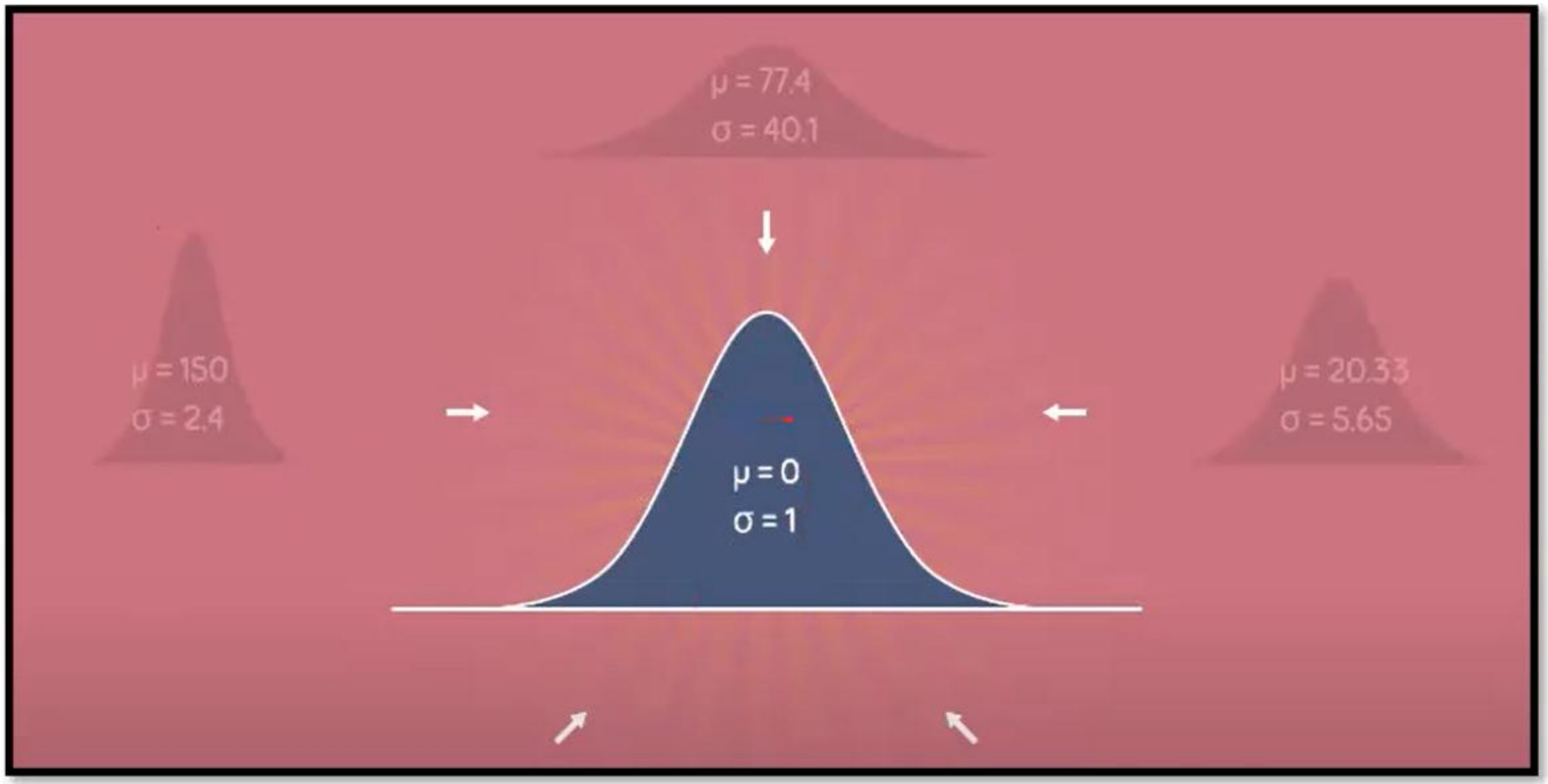
Standard Normal Distribution

Converting the normal distribution to standard normal distribution



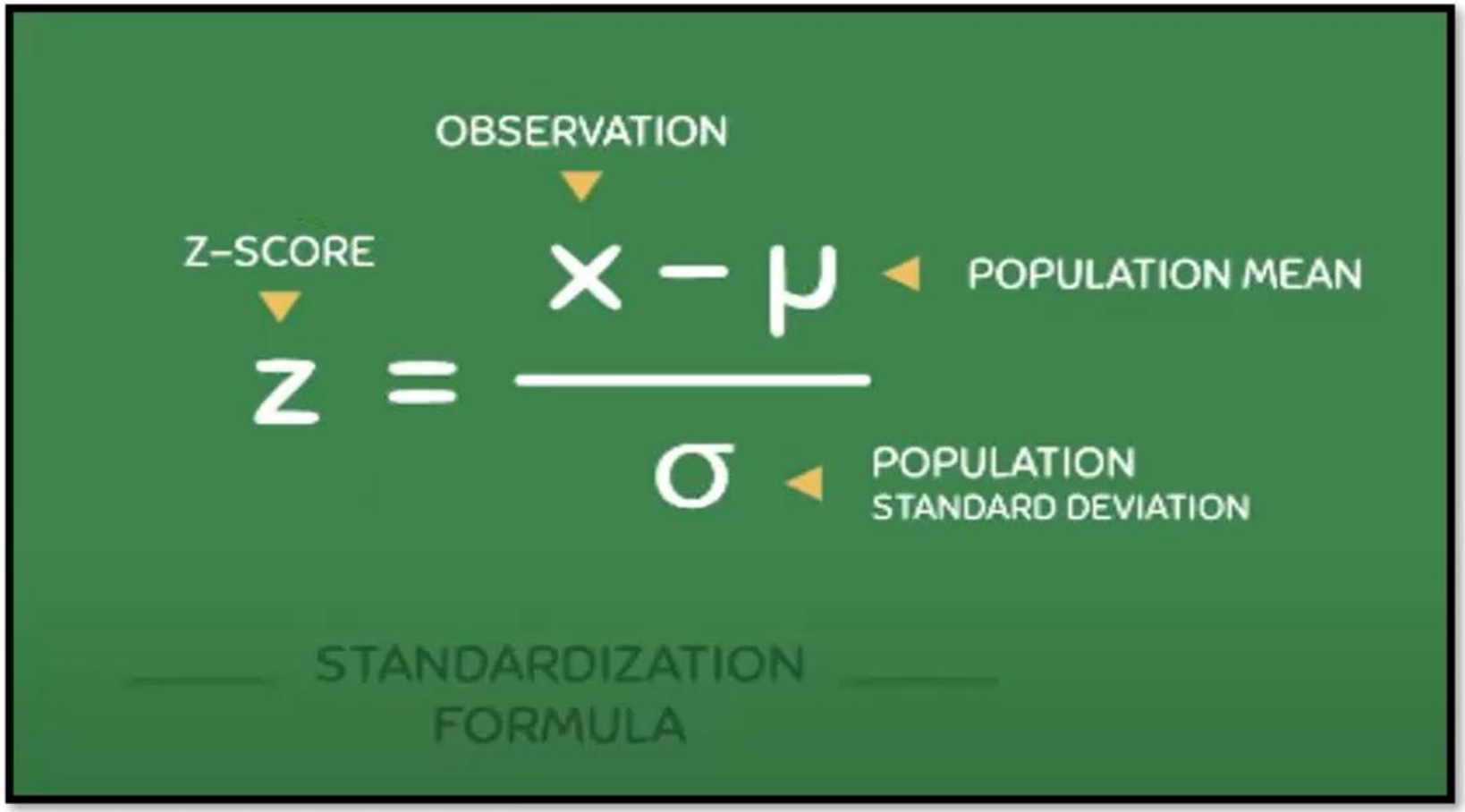
Standard Normal Distribution

Converting the normal distribution to standard normal distribution



Standard Normal Distribution

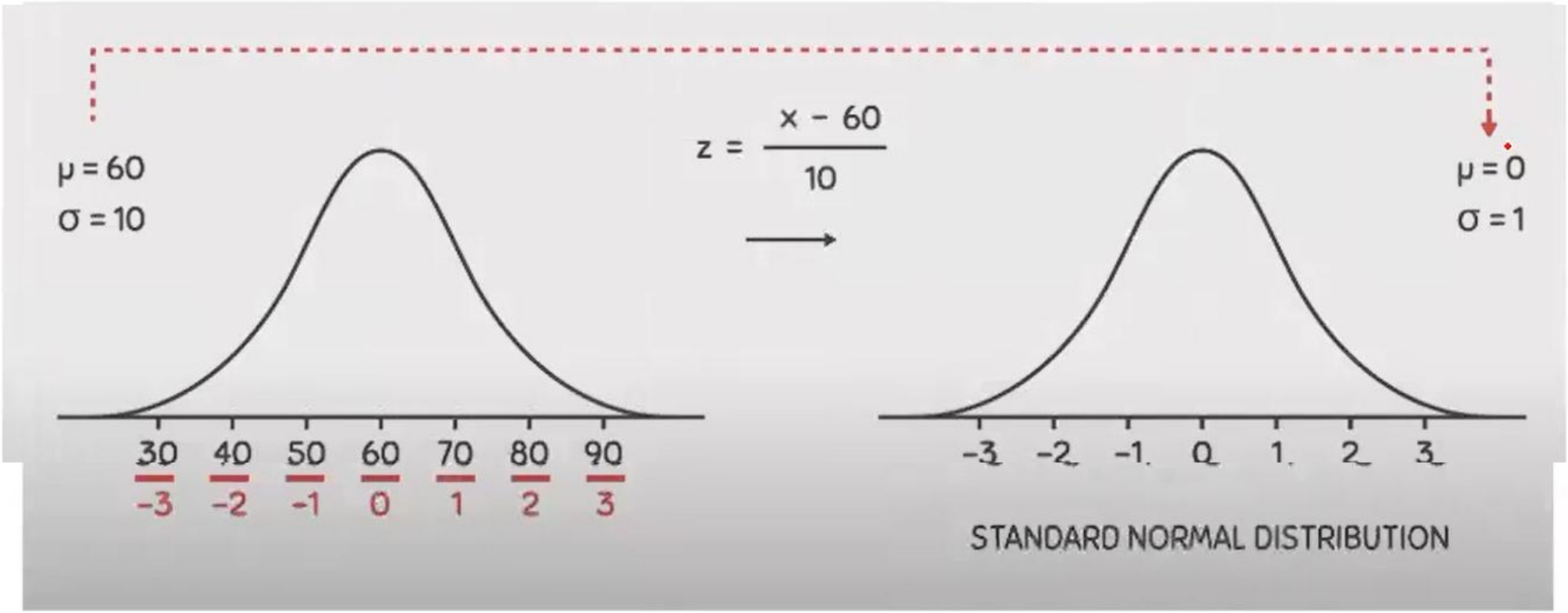
Converting the normal distribution to standard normal distribution



Standard Normal Distribution

Converting the normal distribution to standard normal distribution

EXAMPLE | Suppose that we gathered data from last year's final chemistry exam and found that it followed a normal distribution with a mean of 60 and a standard deviation of 10.



Measure of spread |Z-Score

Comparing Values from Different Data Sets

Two students, John and Ali, from different high schools, wanted to find out who had the highest GPA when compared to his school. Which student had the highest GPA when compared to his school?

Student	GPA	School mean GPA	School standard deviation
John	2.85	3.0	0.7
Ali	77	80	10

For John, $z = \#ofSTDEVs = \frac{2.85-3.0}{0.7} = -0.21$

For Ali, $z = \#ofSTDEVs = \frac{77-80}{10} = -0.3$

Measure of spread |Z-Score

Comparing Values from Different Data Sets

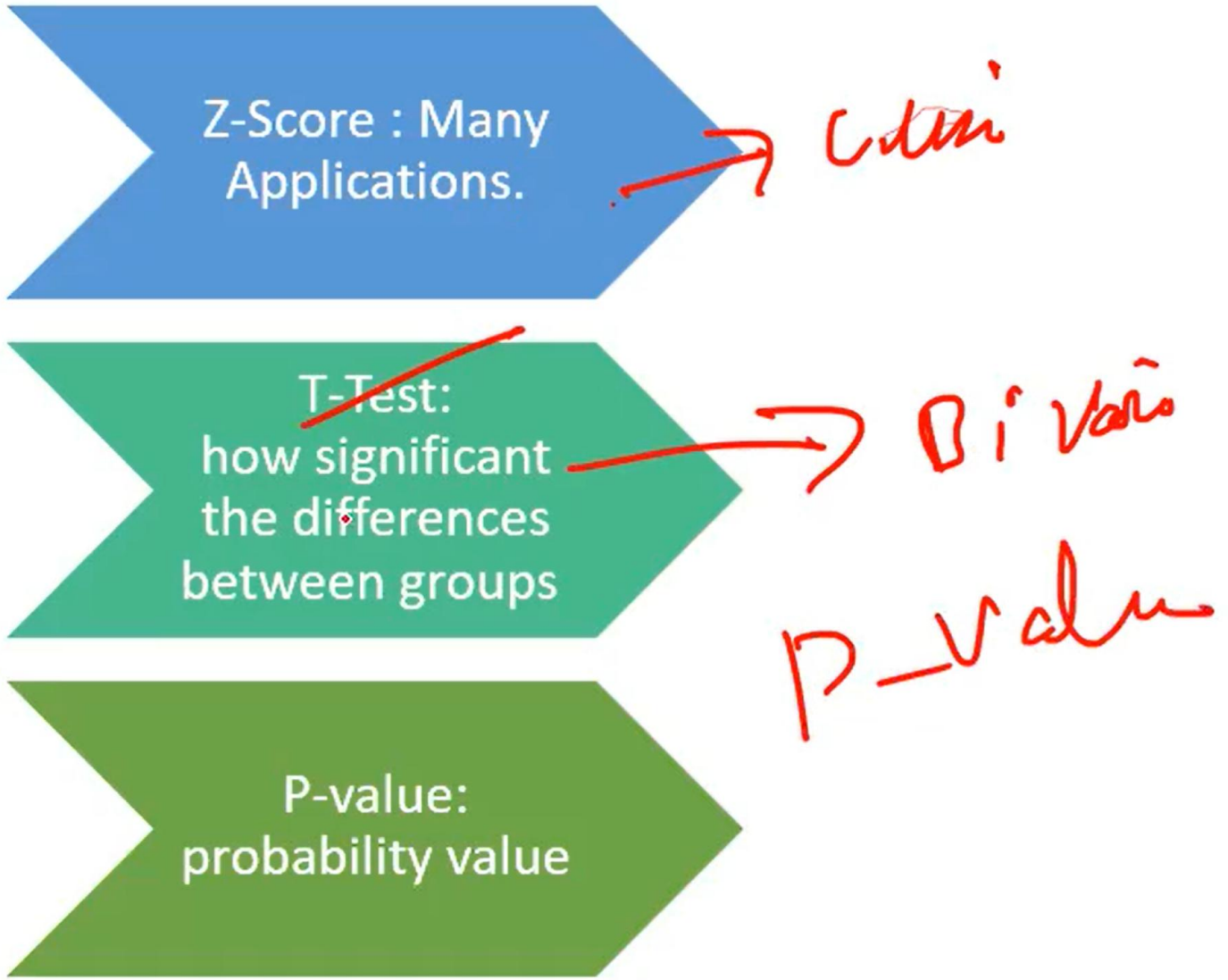
For John, $z = \#ofSTDEVs = \frac{2.85-3.0}{0.7} = -0.21$

For Ali, $z = \#ofSTDEVs = \frac{77-80}{10} = -0.3$

John has the better GPA when compared to his school because his GPA is 0.21 standard deviations **below** his school's mean while Ali's GPA is 0.3 standard deviations **below** his school's mean.

John's z-score of –0.21 is higher than Ali's z-score of –0.3. For GPA, higher values are better, so we conclude that John has the better GPA when compared to his school.

TYPES OF TEST



Z - SCORE

Z-Score : converting any std deviation to standard deviation