Basic Statistical Terminology

Statistics

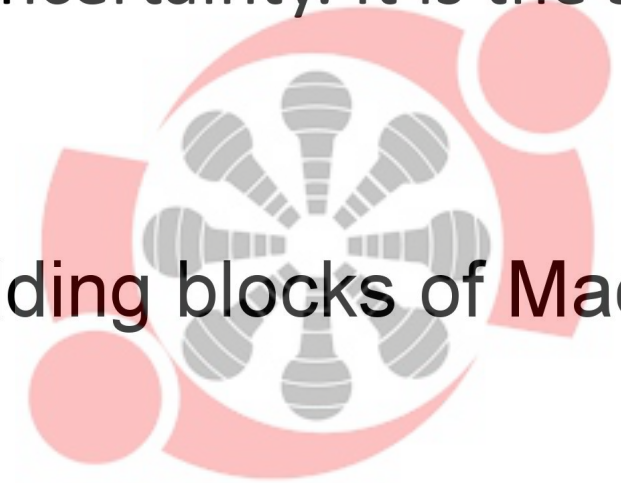
- Statistics is a Mathematical Science pertaining to data collection, analysis, interpretation and presentation.

- Several Statistical functions, principles and algorithms are implemented to analyze raw data, build a statistical model and infer or predict the result.

Statistics more ….

- Statistics is the art and science of using sample data to understand something about the world (or a population) in the context of uncertainty. It is the science of learning from data.

- Stats are the building blocks of Machine Learning algorithms.

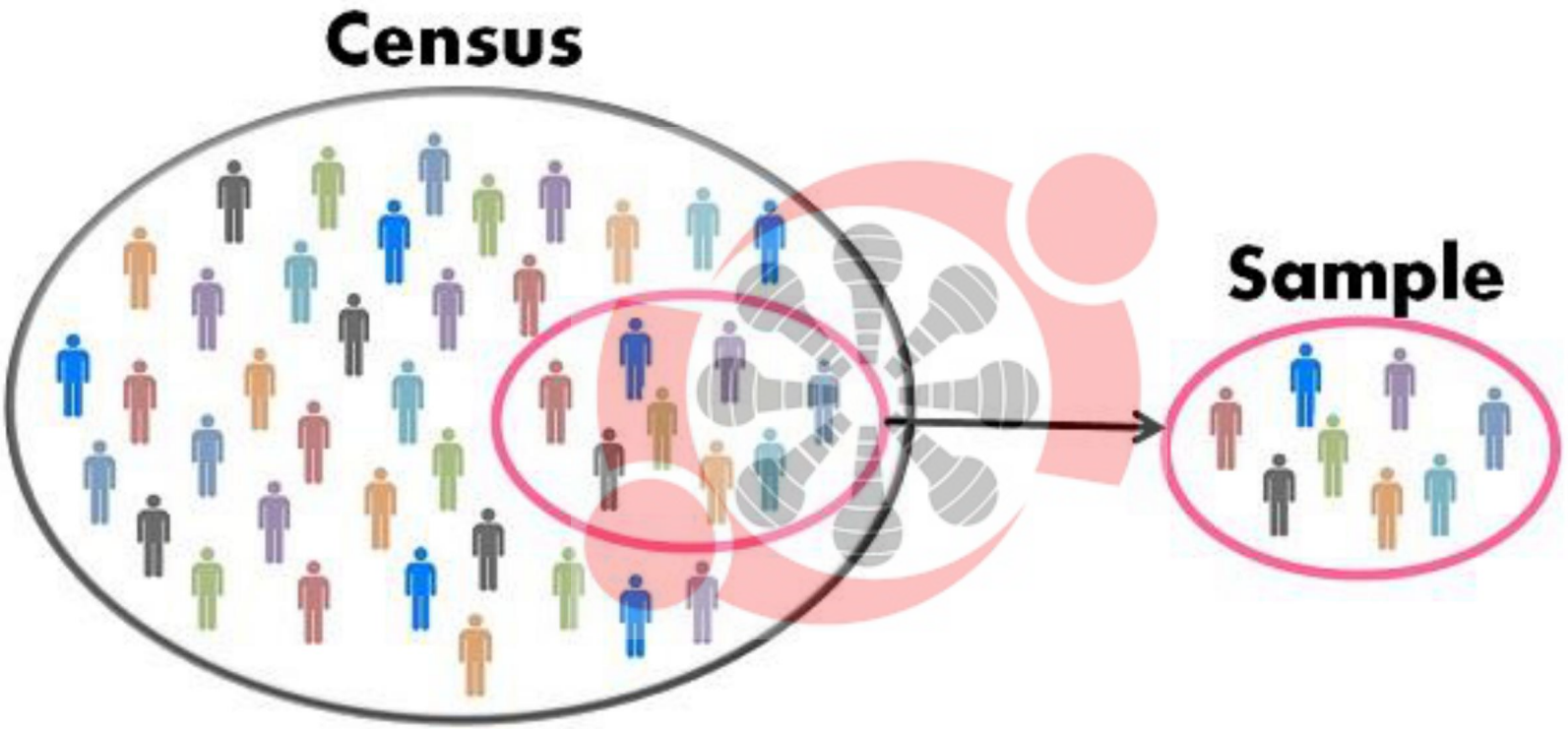


Statistics – Big Picture

Statistics provides a way of organizing data to extract information on a wider and objective basis than relying on personal experience

- First step Data Gathering
- Followed by Data Understanding
- Third Stage Data Analysis and Data Interpretation
- Data Presentation

Population and Sample



Source: https://keydifferences.com/difference-between-census-and-sampling.html

Census and Survey

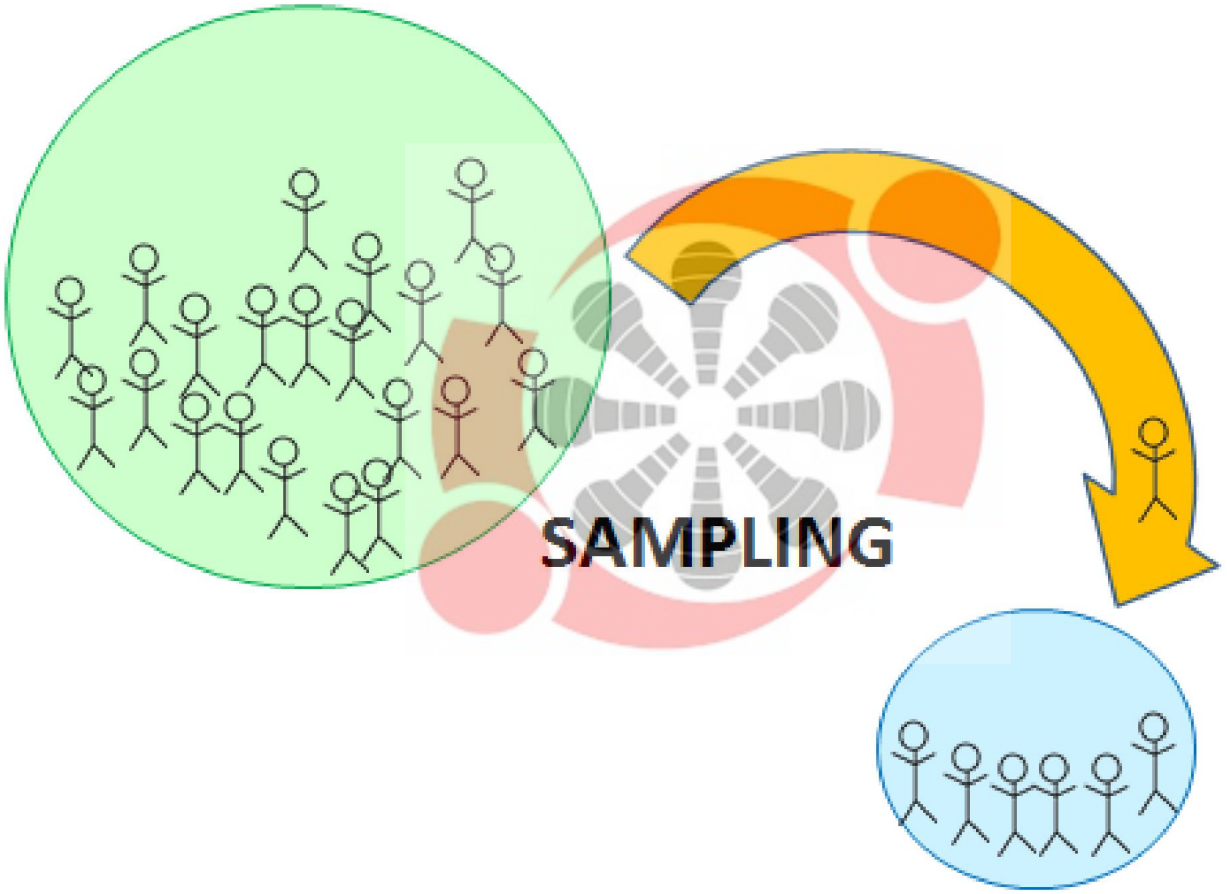
- **Census:** Gathering data from the whole **population** of interest.

Example: General elections, 10-year census, etc.

- **Survey:** Gathering data from the **sample** in order to make conclusions about the population.

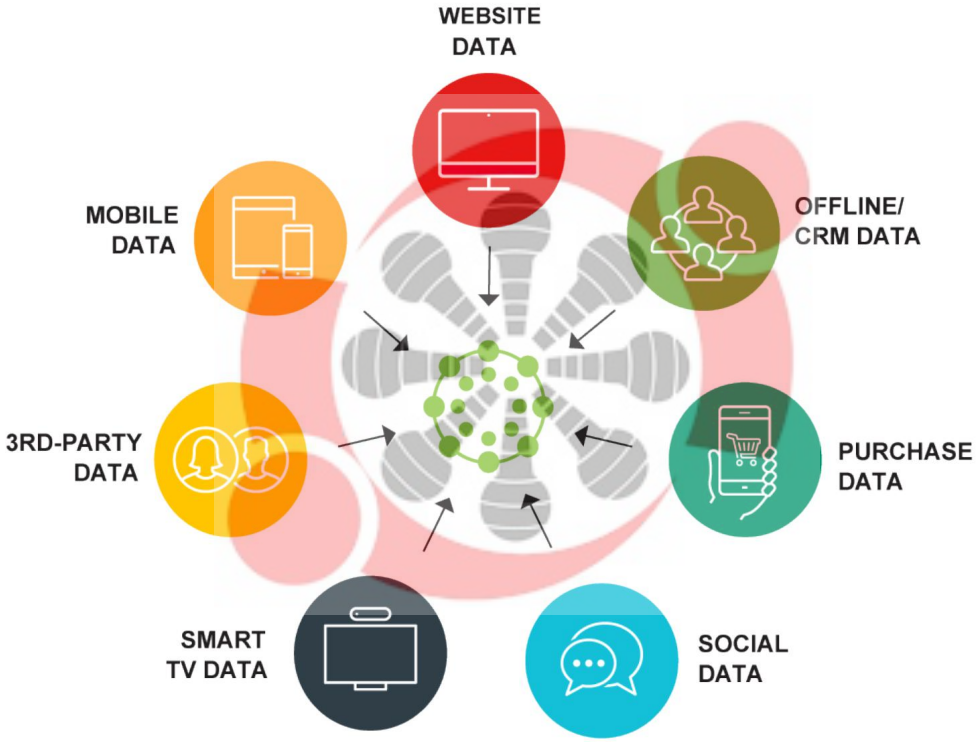
Example: opinion polls, Checking about population check , etc.

Population and Sampling



Data Gathering – Sampling Techniques

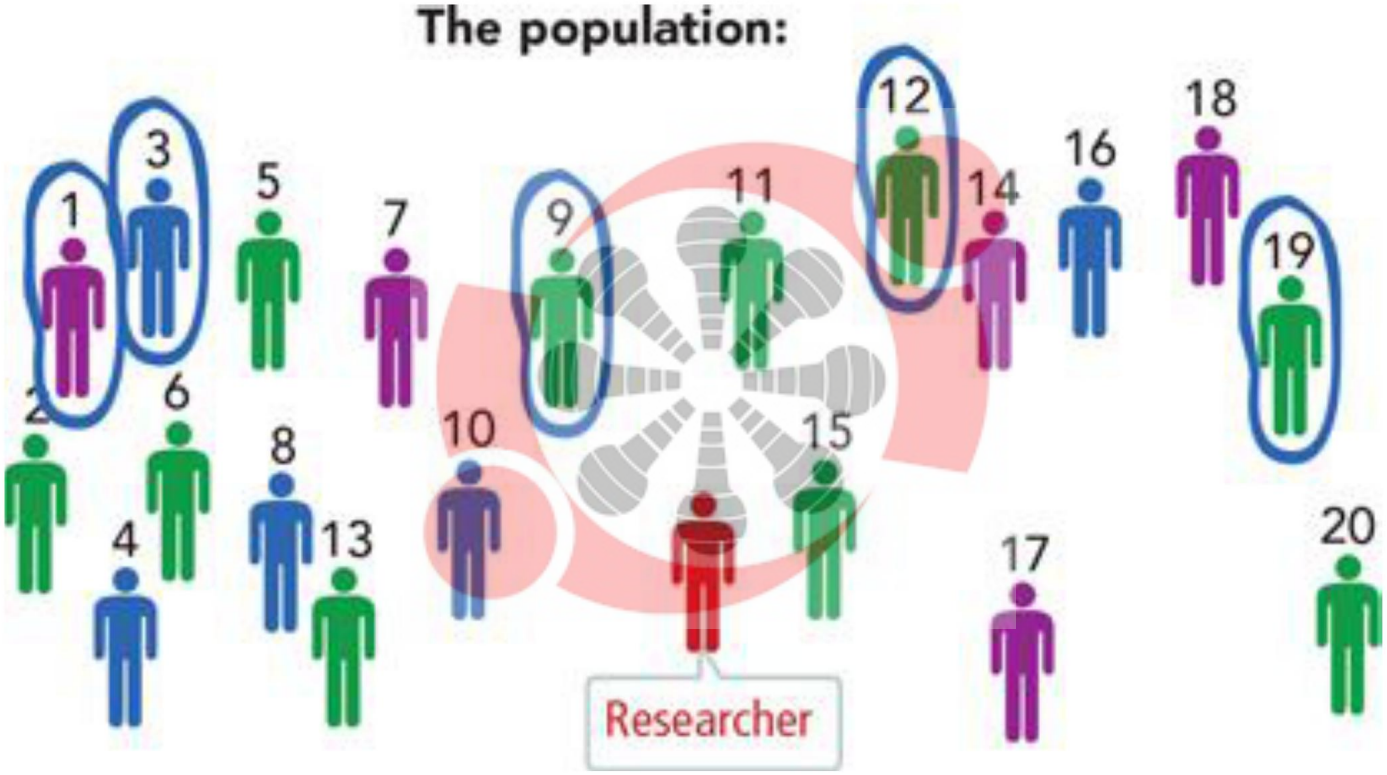
Convenience Sampling



Eg: Online polls, Asking your best friends etc

Data Gathering – Sampling Techniques

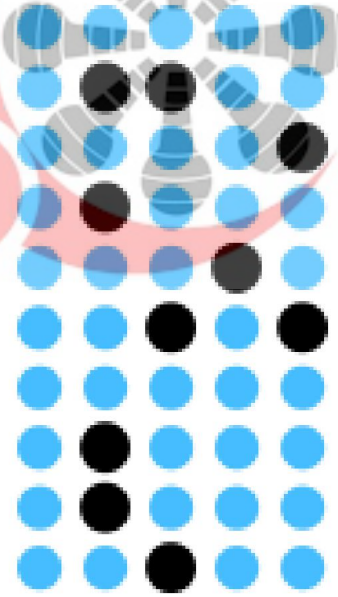
- Random Sampling



- Each member has an equal chance of being selected.

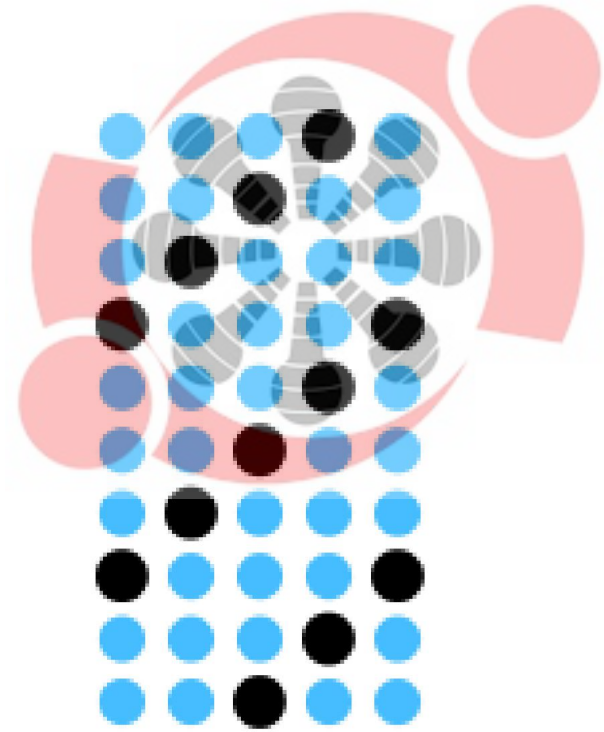
Examples – Random Sample

- Imagine slips of paper each with a person's name, put all the slips into a barrel, mix them up, then dive your hand in and choose some slips of paper.
- But this means you need a **full list of the population** to choose from.



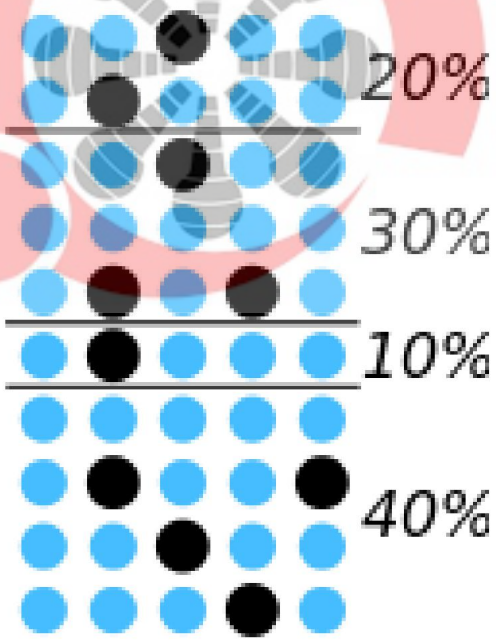
Systematic Sampling

- This is where we follow some system of selection like "every 10th person"



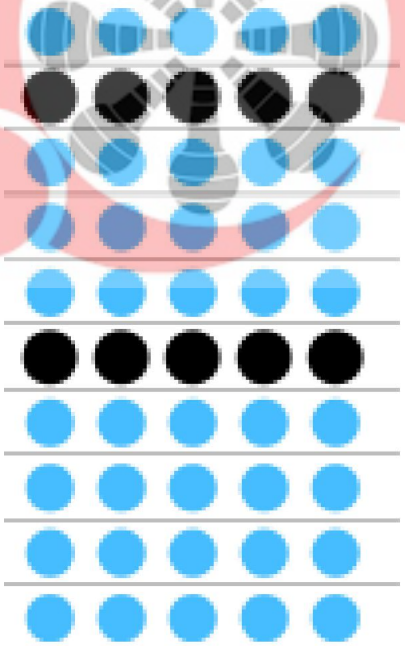
Stratified Sampling

- This is where we divide the population into groups by some characteristic such as **age** or **occupation** or **gender**.
- Then make sure our survey includes people from each group in proportion to how many there are in the whole population.



Clustering Sampling

- We break the population into many groups, then randomly choose whole groups.
- Example: we divide the town into many different zones, then randomly choose 5 zones and survey everyone in those zones.



Parameter and Statistic

Parameter: A descriptive measure of the **population**. For example, population mean, population variance, population standard deviation, etc.

Statistic: A descriptive measure of the **sample**. For example, sample mean, sample variance, sample standard deviation, etc.

Parameter and Statistics

GREEK LETTERS

Mean - μ

Variance – σ^2

Standard Deviation - σ

ROMAN LETTERS

Mean - x

Variance – s^2

Standard Deviation - s

Statistics

Descriptive Statistics

♣ Data gathered about a group to reach conclusion about the same group.

Inferential Statistics

♣ Data gathered from a sample and the statistics generated to reach conclusion about the population from which the sample is taken. Also known as Inductive Statistics

Variables and Data

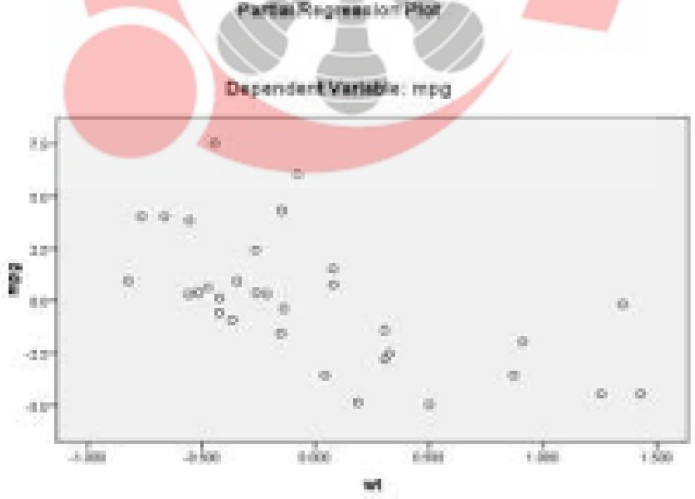
- observation

model	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21	6	160	110	3.9	2.62	16.46	0	1	4	4
Mazda RX4 Wag	21	6	160	110	3.9	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.32	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.44	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.46	20.22	1	0	3	1
Duster 360	14.3	8	360	245	3.21	3.57	15.84	0	0	3	4
Merc 240D	24.4	4	146.7	62	3.69	3.19	20	1	0	4	2
Merc 230	22.8	4	140.8	95	3.92	3.15	22.9	1	0	4	2
Merc 280	19.2	6	167.6	123	3.92	3.44	18.3	1	0	4	4
Merc 280C	17.8	6	167.6	123	3.92	3.44	18.9	1	0	4	4
Merc 450SE	16.4	8	275.8	180	3.07	4.07	17.4	0	0	3	3
Merc 450SL	17.3	8	275.8	180	3.07	3.73	17.6	0	0	3	3
Merc 450SLC	15.2	8	275.8	180	3.07	3.78	18	0	0	3	3
Cadillac Fleetwood	10.4	8	472	205	2.93	5.25	17.98	0	0	3	4
Lincoln Continental	10.4	8	460	215	3	5.424	17.82	0	0	3	4
Chrysler Imperial	14.7	8	440	230	3.23	5.345	17.42	0	0	3	4
Fiat 128	32.4	4	78.7	66	4.08	2.2	19.47	1	1	4	1
Honda Civic	30.4	4	75.7	52	4.93	1.615	18.52	1	1	4	2
Toyota Corolla	33.9	4	71.1	65	4.22	1.835	19.9	1	1	4	1
Toyota Corona	21.5	4	120.1	97	3.7	2.465	20.01	1	0	3	1
Dodge Challenger	15.5	8	318	150	2.76	3.52	16.87	0	0	3	2
AMC Javelin	15.2	8	304	150	3.15	3.435	17.3	0	0	3	2
Camaro Z28	13.3	8	350	245	3.73	3.84	15.41	0	0	3	4
Pontiac Firebird	19.2	8	400	175	3.08	3.845	17.05	0	0	3	2
Fiat X1-9	27.3	4	79	66	4.08	1.935	18.9	1	1	4	1
Porsche 914-2	26	4	120.3	91	4.43	2.14	16.7	0	1	5	2
Lotus Europa	30.4	4	95.1	113	3.77	1.513	16.9	1	1	5	2
Ford Pantera L	15.8	8	351	264	4.22	3.17	14.5	0	1	5	4
Ferrari Dino	19.7	6	145	175	3.62	2.77	15.5	0	1	5	6
Maserati Bora	15	8	301	335	3.54	3.57	14.6	0	1	5	8
Volvo 142E	21.4	4	121	109	4.11	2.78	18.6	1	1	4	2

Variables – Dependent and Independent

- Dependent variables on y-axis and Independent on x-axis.

- Dependent variable also called **Target variable** or Class variable.



Data

Qualitative data (Categorical)

- ♣ is descriptive information (it describes something)

Quantitative data (Numerical)

- ♣ is numerical information (number)

Discrete

Continuous

What do we know about Rohit Sharma ?

- **Qualitative**:
  - His bat color is blue and orange
  - He has short hair
  - He has lots of energy
- **Quantitative**:
  - Discrete:
    - He has 2 legs
    - He has 1 daughter.
  - Continuous:
    - He weighs 70.5 kg
    - He is 175 cm tall

Data – Numeric and Categorical

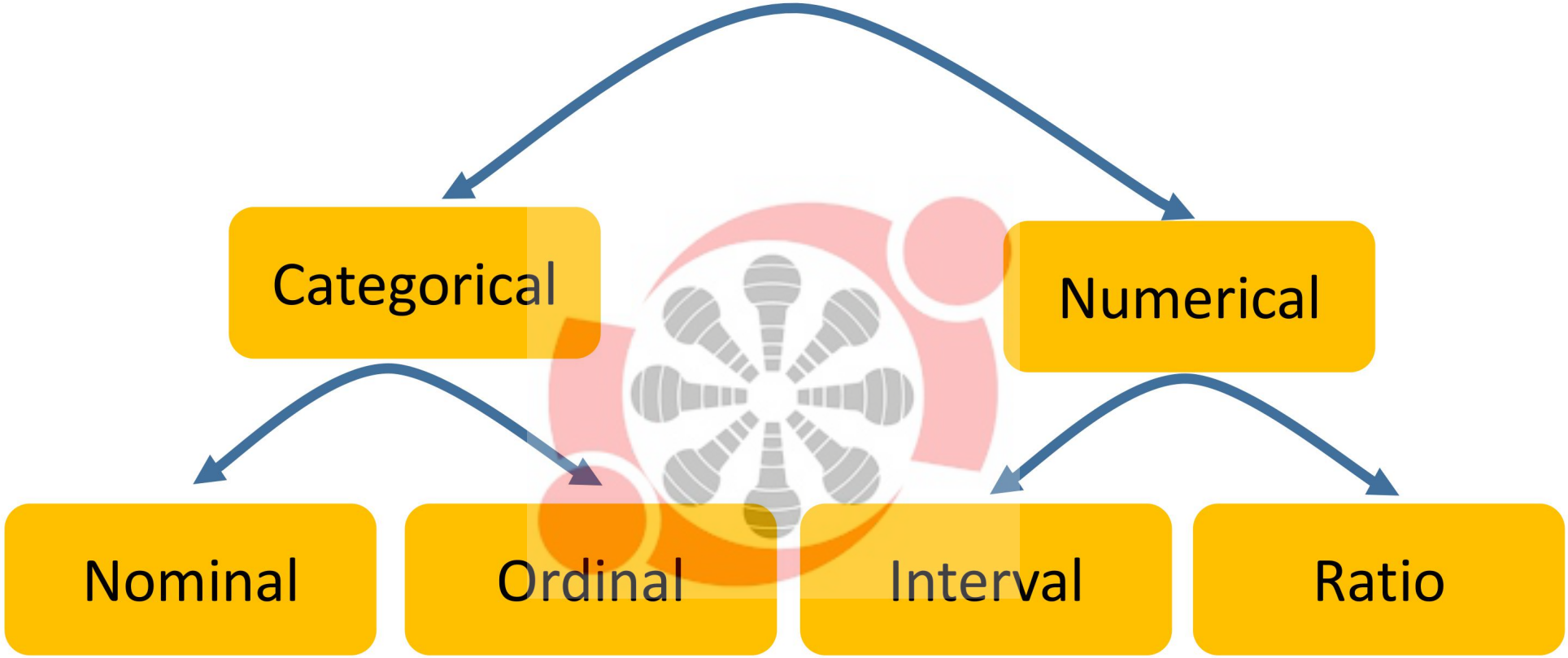
- Count of gold?
- Is it Gold or not ?
- Weight of the gold.

18

500k

g

Data



Nominal Data (Qualitative)

- Nominal means name and count
  - Data are alphabetic or numerical in name
- They are categories without order or direction
- Example:

What is your Gender?

- Male
- Female

What languages do you speak?

- English
- German
- Spanish
- French

Ordinal Data (Qualitative)

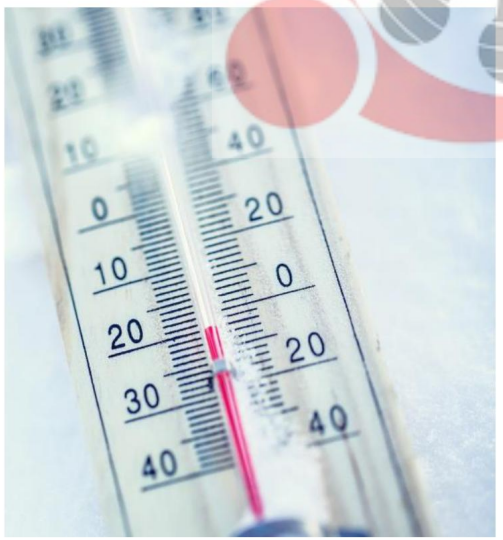
- Ordinal means rank or order
- Data place in order. They are ordered categories like ranking or scaling.
- Has no absolute value
- While there is an order, difference between consecutive levels are not always equal

Example

- Fortune 100 rankings
- Movie ratings , Movie Collections

Quantitative Data - Interval

- Data where ordering is clear
- Difference in data values is meaningful.
- Interval data, also called an integer, is defined as a data type which is measured along a scale, in which each point is placed at equal distance from one another.



Quantitative Data - Ratio

Ratio variables are interval variables, but with the added condition that 0 (zero) of the measurement indicates that there is none of that variable.

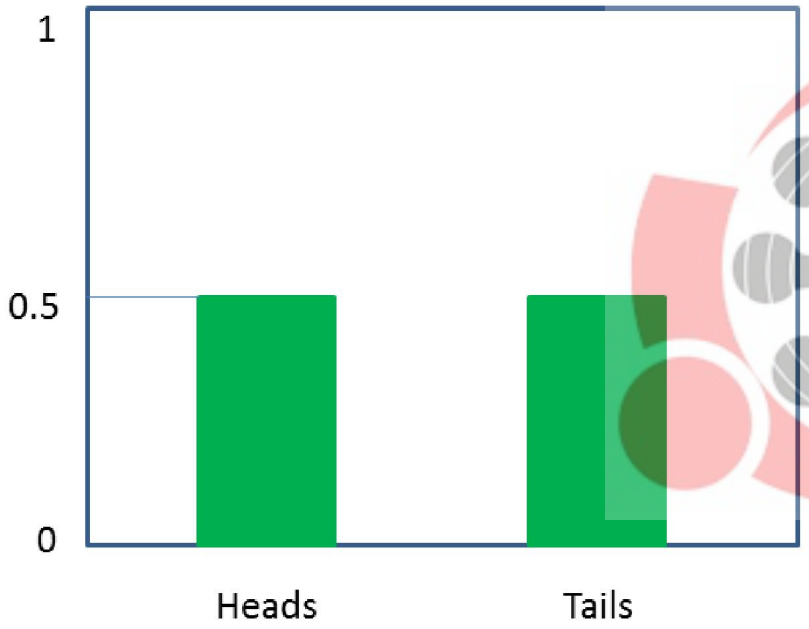
. The name "ratio" reflects the fact that you can use the ratio of measurements. So, for example, a distance of ten meters is twice the distance of 5 meters.

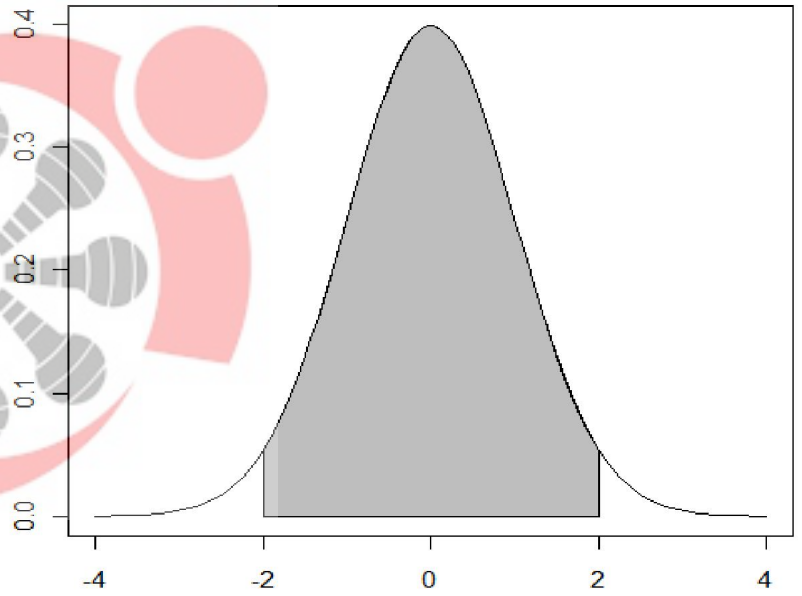
Examples: Weights, Cost of things, Number of correct answers in a exam

Summary of Level of Data Measurement

- Nominal – Categories only
- Ordinal – Categories with some order
- Interval – Meaningful difference, but no zero point
- Ratio – Meaningful difference with a natural starting point.

Discrete and Continuous





Countable

Measurable

Discrete or Continuous

Statement	Discrete / Continuous
Time between customer arrivals at retail outlet	Continuous
Sampling the volume of air in a storage tank	Continuous
Sampling 1 Lakh voters in a exit poll and determining how many voted for the wining candidate	Discrete
Length of newly designed mobile phone	Continuous
No. of customers arriving at a retail outlet during a 1 hour period.	Discrete
No. of defects in a batch of 1000 items	Discrete

Describing Data through Statistics

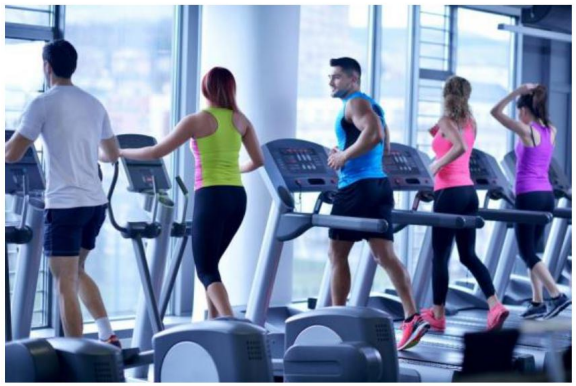
Descriptive Statistics

The Central Tendencies

Virat want to join a health club in a activity that has others in the same age group as him. He is 28 years old. Mean ages for **YOGA, GYM** and **SWIMMING** classes are

22 years

30 years



17 years

The Central Tendencies

Yoga class composition

Age (years)	19	22	23
Frequency, f	1	3	2

$$Mean, \mu = \frac{\sum x}{n} =$$

$$\frac{19 * 1 + 22 * 3 + 23 * 2}{1 + 3 + 2} \approx 22$$

The Central Tendencies

Power workout class composition

Age (years)	20	22	23	90
Frequency, f	4	8	5	1

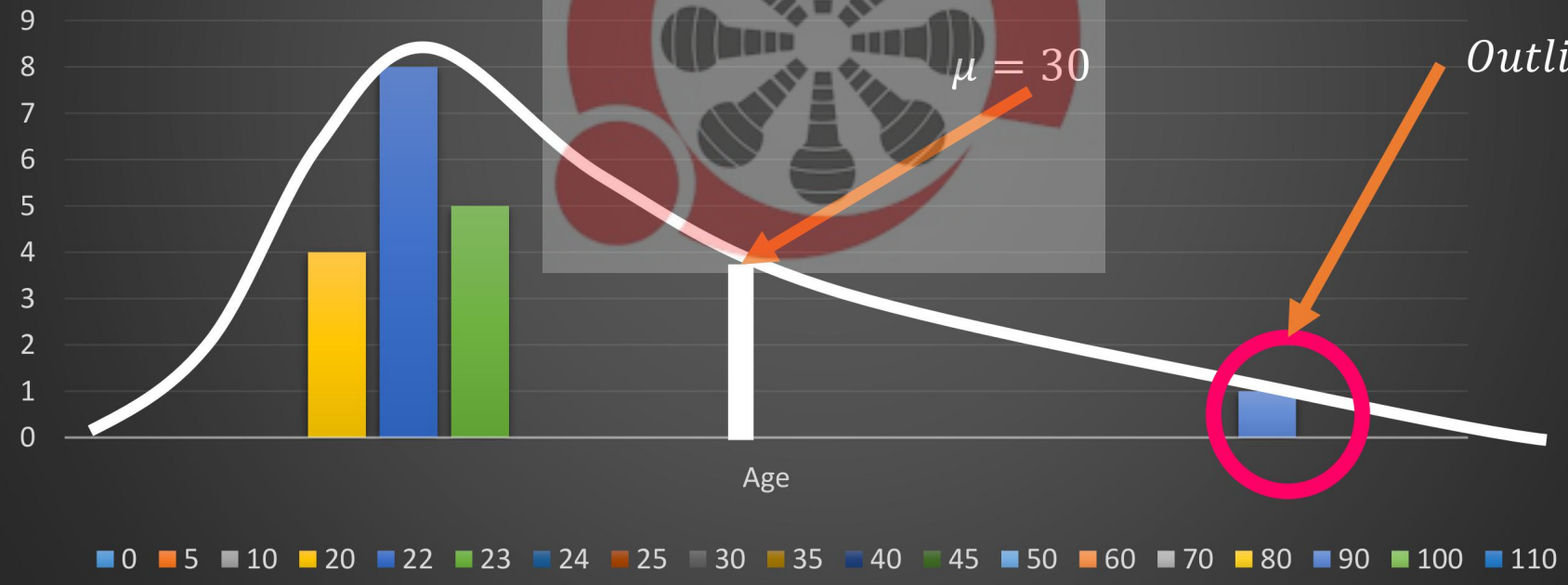
$$Mean, \mu = \frac{\sum x}{n} =$$

$$\frac{20 * 4 + 22 * 8 + 23 * 5 + 90 * 1}{4 + 8 + 5 + 1} = 30$$

The Central Tendencies

Power Workout Class Composition

Age (years)	20	22	23	90
Frequency, f	4	8	5	1



The Central Tendencies – Median

Age (years)	20	22	23	90
Frequency, f	4	8	5	1

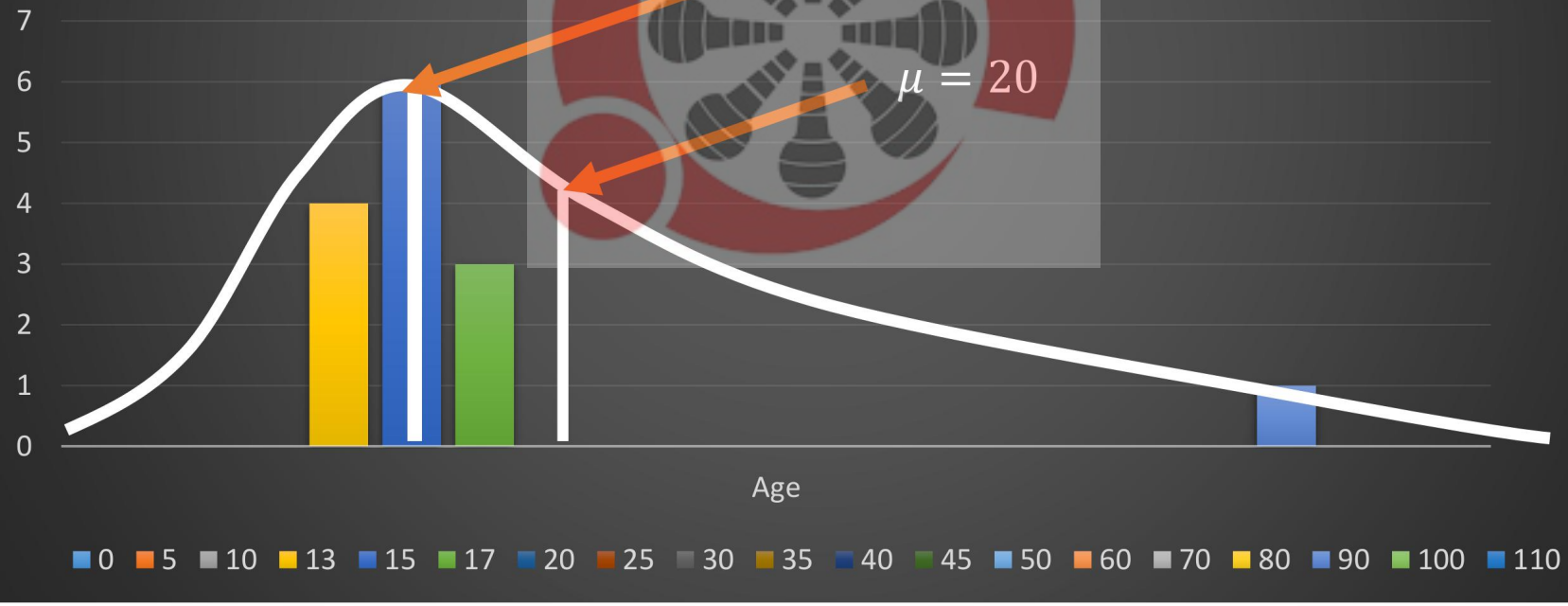
- Data has outlier
- Median - the mid-point

13, 13, 13, 13, 22, 22, 22, 22, 22, 22, 22, 22, 23, 23, 23, 23, 23, 90

The Central Tendencies

Power Workout Class Composition

Age (years)	13	15	17	90
Frequency, f	4	6	3	1

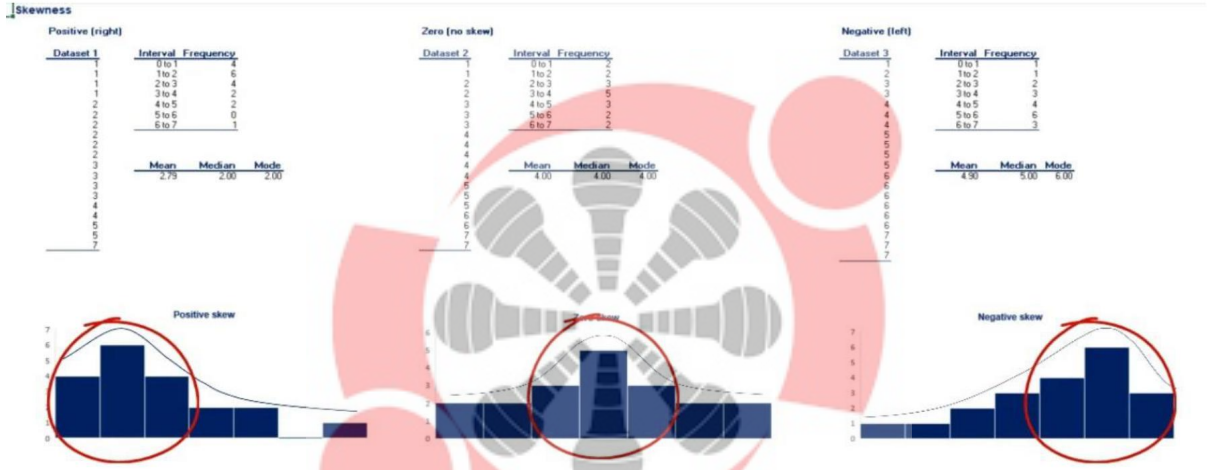


Skewness

- Skewness basically gives the shape of normal distribution of values.
- Skewness is asymmetry in a statistical distribution, in which the curve appears distorted or skewed either to the left or to the right.
- Skewness can be quantified to define the extent to which a distribution differs from a normal distribution.

Skewness

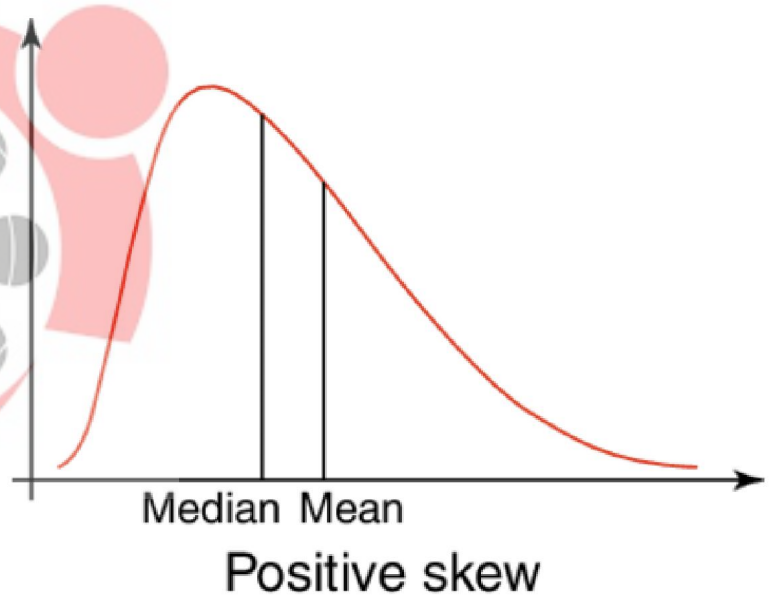
- Skewness tells us a lot about where the data is situated.



- In fact, the mean, median and mode should be used together to get a good understanding of the dataset.
- Measures of asymmetry like skewness are the link between central tendency measures and probability theory.
- This ultimately allows us to get a more complete understanding of the data we are working with.

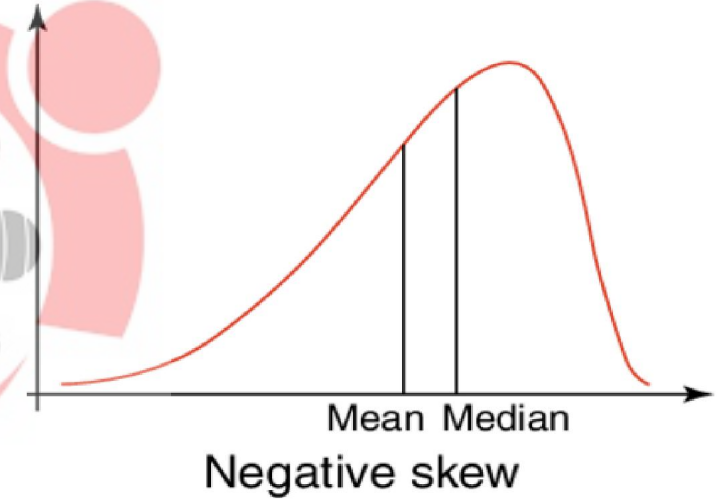
Positive Skewness

A positively skewed distribution means that the extreme data results are larger. This skews the data in that it brings the mean (average) up. The mean will be larger than the median in a Positively skewed distribution.



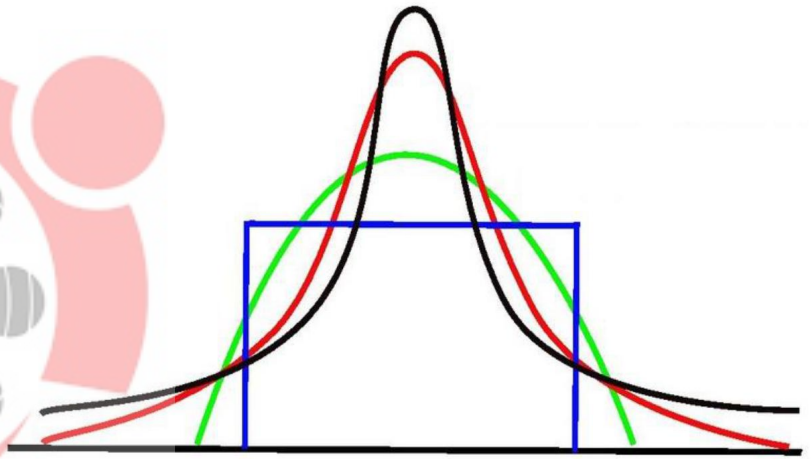
Negative Skewness

A negatively skewed distribution means the opposite: that the extreme data results are smaller. This means that the mean is brought down, and the median is larger than the mean in a negatively skewed distribution.



Kurtosis

The exact interpretation of the measure of Kurtosis used to be disputed, but is now settled. It's about existence of outliers. Kurtosis is a measure of whether the data are heavy-tailed (profusion of outliers) or light-tailed (lack of outliers) relative to a normal distribution.



The Central Tendencies

Sai is disturbed and wants some relaxation. He joins the swimming class where mean age is 17 years. He didn't understand why they were asking where his kid was…

Age (Years)	1	2	3	30	31	32	33
Frequency,f	3	4	3	1	3	2	4

$\mu \approx 17 \ Years$

Median ?

What happens to Median if another kid or adult is added ?

The Central Tendencies

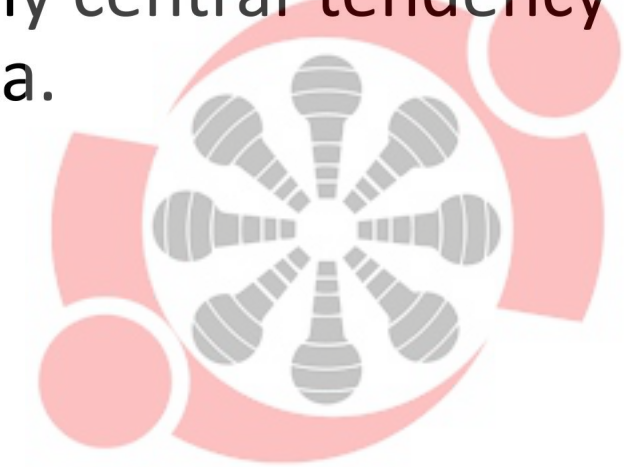
Age (Years)	1	2	3	30	31	32	33
Frequency,f	3	4	3	1	3	2	4

What is the mode – the most frequently occurring data point ?

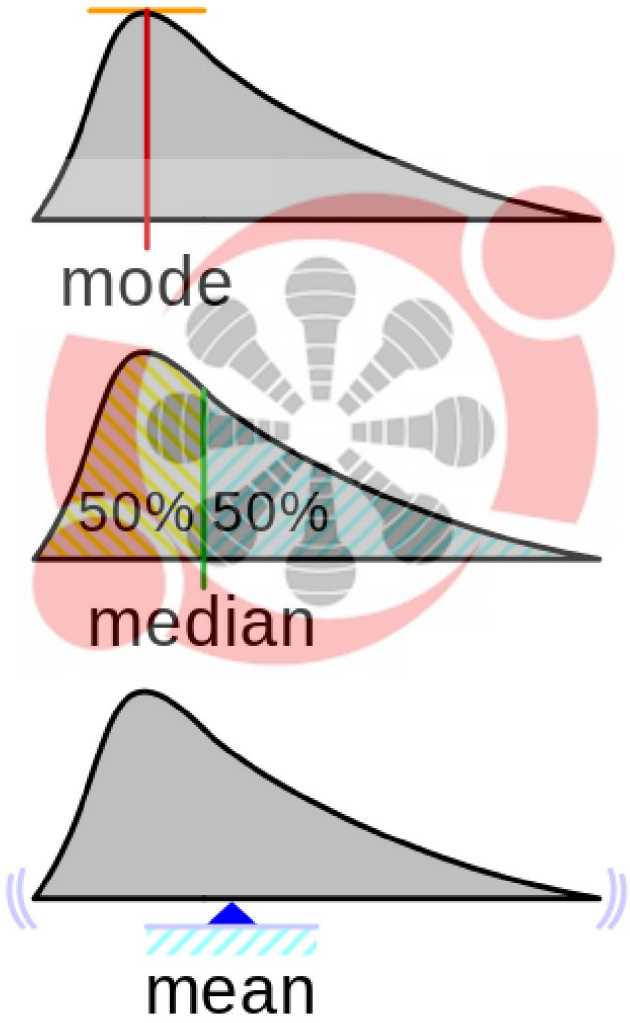
The Central Tendencies

Mean and Median need not be in the dataset but Mode has to be in it.

Mode is also the only central-tendency statistic that works with categorical data.



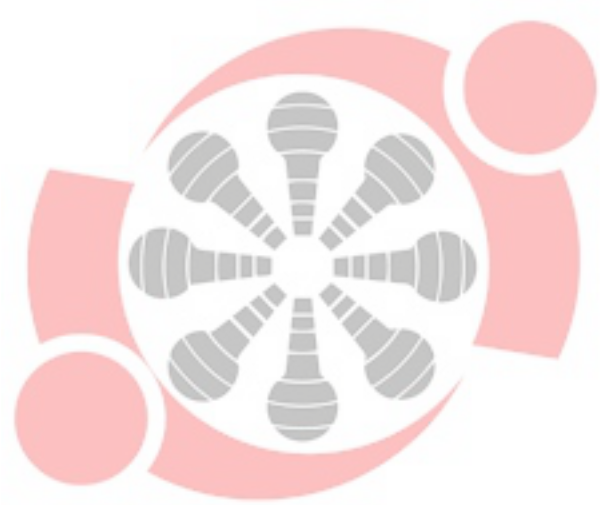
The Central Tendencies



The Central Tendencies

The management of Good Heart Inc. wants to give all its employees a raise. They are unable to decide if they should give a straight Rs. 2000 to everyone or to increase salaries by 10% across the board. The mean salary is Rs. 50,000, the median is Rs. 20,000 and the mode is Rs. 10,000.

How do these central tendencies change in both cases?



Measuring Variability and Spread

Range, Variance, Standard Deviation

Range to differentiate between dataset

- It is quite often, the average only gives part of the picture.
- Averages give us a way of determining where the centre of a set of data is, but they don't tell us how the data varies.
- "The range tells us over how many numbers the data extends, a bit like measuring its width."

Range

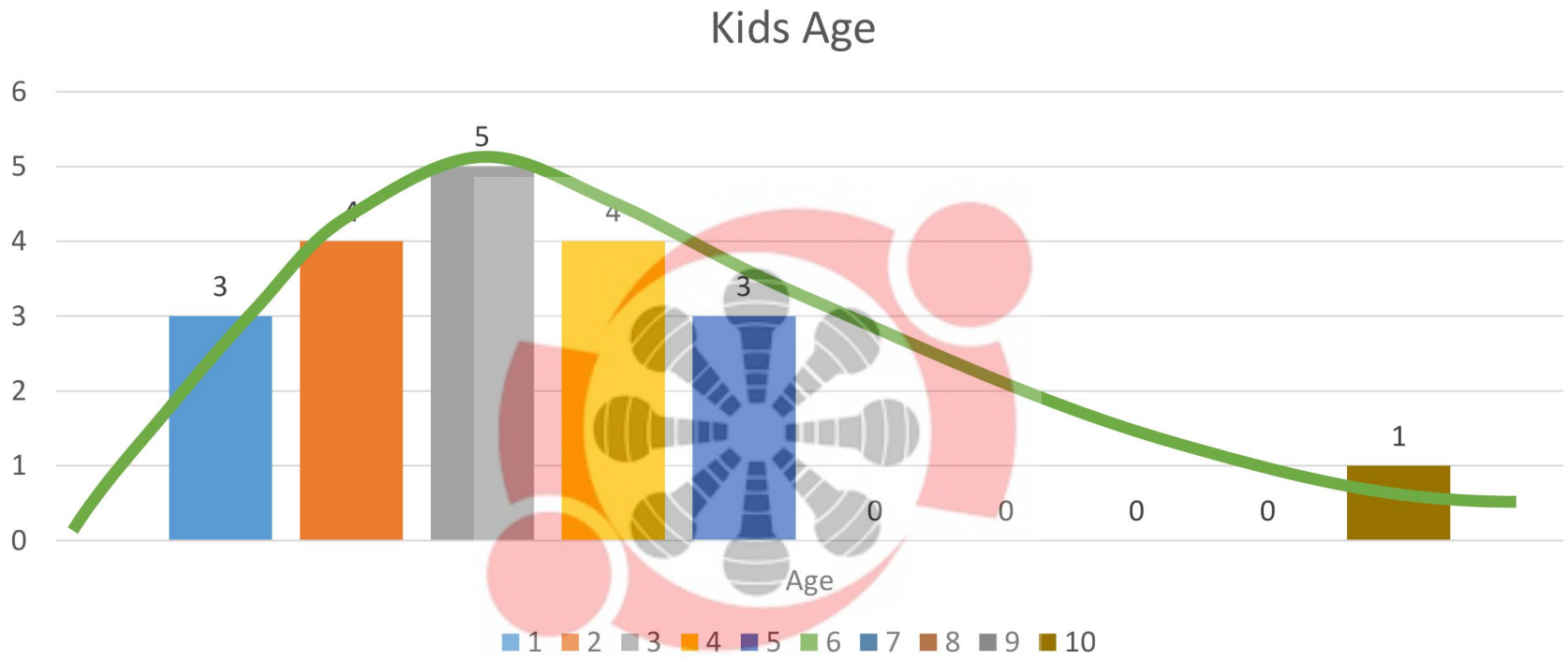
The range is a way of measuring how spread out a set of values are. It's given by Upper bound - Lower bound where the upper bound is the highest value, and the lower bound the lowest.

Range = upper bound - lower bound

= 10 -1

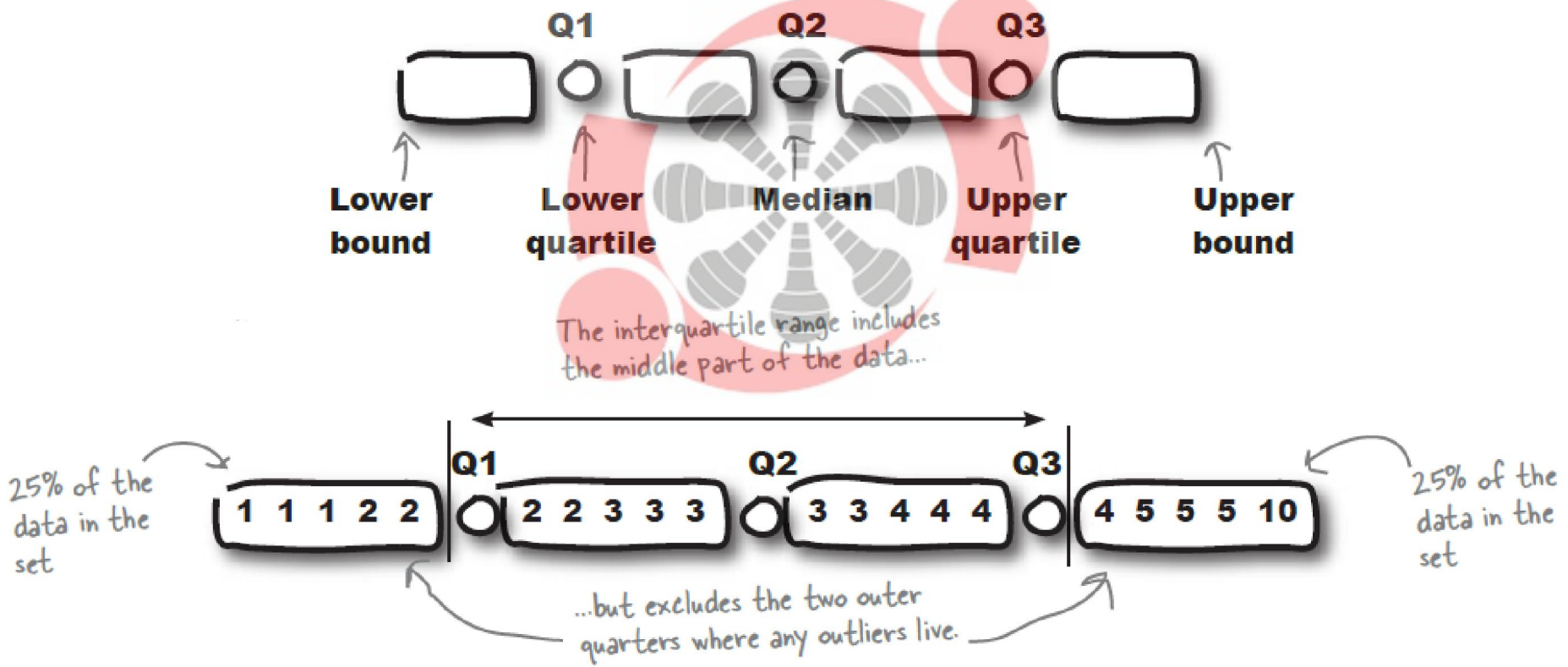
= 9

so, the range is 9


Kids Age
6
5
4
3
2
1
0
3
4
5
4
3
0
0
0
0
1
Age
1 2 3 4 5 6 7 8 9 10

Quartiles will rescue the problem

Quartiles of a set of data is a very similar process to finding the median.



Quartiles

Quartiles : division of the data set into 4 regions If we have *n data-points then the Quartile boundaries are* given by

Lower quartile (25th percentile, Q1) = $\left(\frac{1*(n-1)}{4}+1\right)^{th}$

Middle quartile = Median = $\left(\frac{2*(n-1)}{4}+1\right)^{th} = \frac{(n+1)}{2}$th

Upper quartile (75th percentile, Q3) = = $\left(\frac{3*(n-1)}{4}+1\right)^{th}$

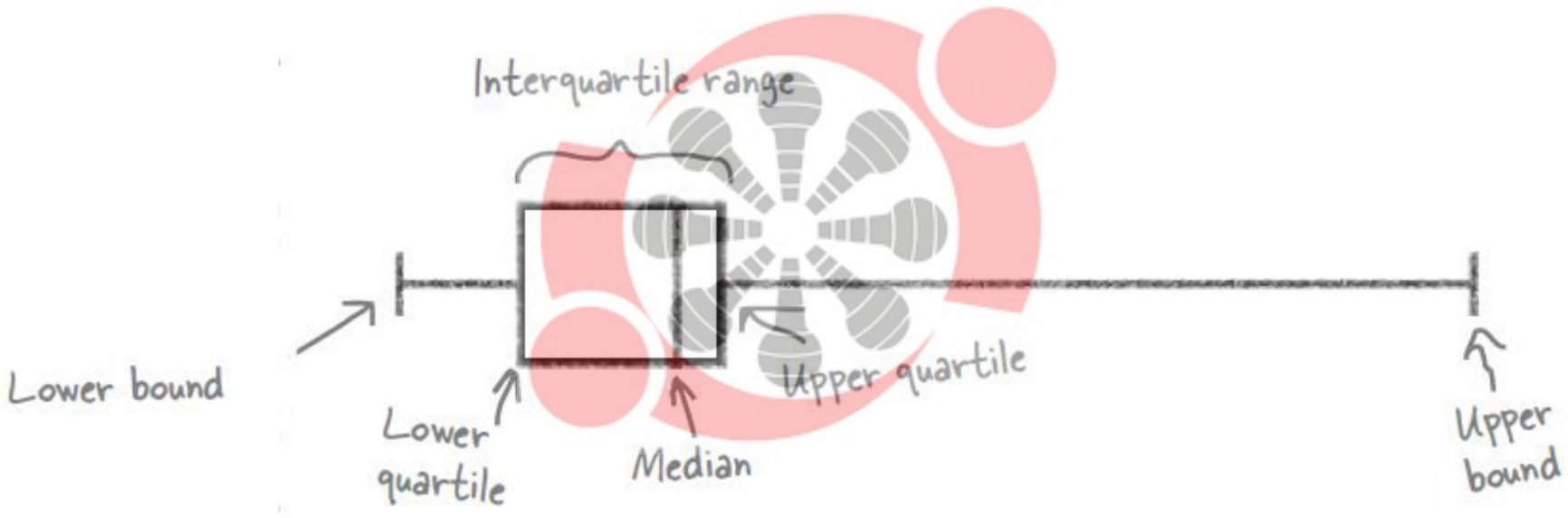
Interquartile range, IQR = Q3 – Q1 (central 50% of data)

Quartiles

Percentile: divide the dataset into 100 regions

$$pth \text{ Percentile} = \left(\frac{p*(n-1)}{100} + 1\right) th$$

Box and Whisker Plot → Quatiles

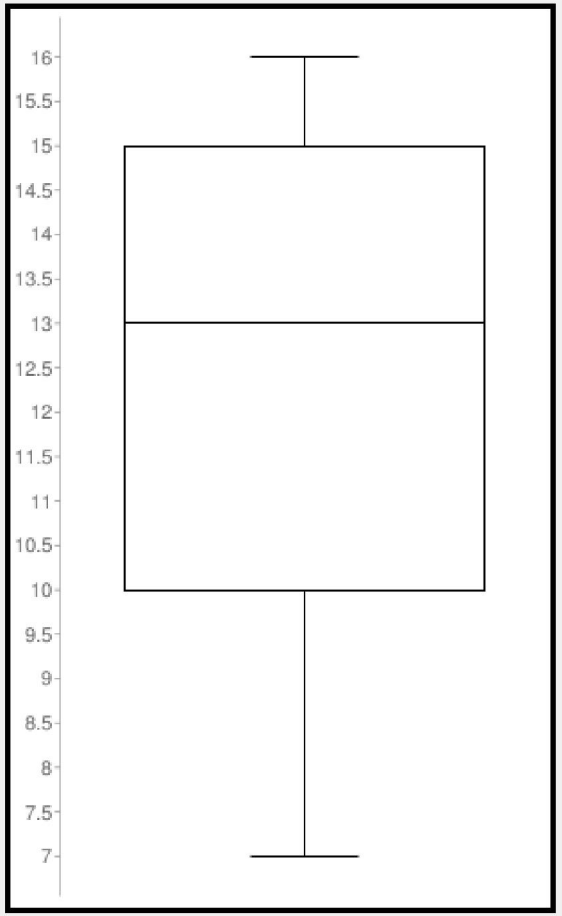


Interpreting Box-whisker plot

Age of the Student in a seminar

Which of the following statements are true?

- All of the students are less than 20 years old
- At least 75% of the students are 10 years old or older
- There is only one 16 year old at the party
- The youngest kid is 7 years old
- Exactly half the kids are older than 13 in a party



Variance

The variance is a way of measuring spread, and it's the average of the distance of values from the mean squared.

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

This is a method of measuring spread

Standard deviation

- Standard deviation is a way of saying how far typical values are from the mean.
- The smaller the standard deviation, the closer values are to the mean.
- The smallest value the standard deviation can take is 0.

$$\sigma = \sqrt{Variance}$$

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

This is a method of measuring spread

Value	Mean	value- Mean	Z	Outlier ?
1	3.26	-2.26	-1.1037	Not outlier
1	3.26	-2.26	-1.1037	Not outlier
1	3.26	-2.26	-1.1037	Not outlier
2	3.26	-1.26	-0.6160	Not outlier
2	3.26	-1.26	-0.6160	Not outlier
2	3.26	-1.26	-0.6160	Not outlier
2	3.26	-1.26	-0.6160	Not outlier
3	3.26	-0.26	-0.6160	Not outlier
3	3.26	-0.26	-0.1283	Not outlier
3	3.26	-0.26	-0.1283	Not outlier
3	3.26	-0.26	-0.1283	Not outlier
3	3.26	-0.26	-0.1283	Not outlier
4	3.26	1.74	0.3593	Not outlier
4	3.26	1.74	0.3593	Not outlier
4	3.26	1.74	0.3593	Not outlier
4	3.26	1.74	0.3593	Not outlier
5	3.26	2.74	0.8470	Not outlier
5	3.26	2.74	0.8470	Not outlier
10	3.26	7.74	3.28	outlier

Action Check`

Basketball coach is in a dilemma choosing between 3 players all having the same average scores.

Points Scored per game	7	8	9	10	11	12	13
Frequeny,f	1	1	2	2	2	1	1

Points Scored per game	7	9	10	11	13
Frequency, f	1	2	4	2	1

Points Scored per game	3	6	7	10	11	13	30
Frequency, f	2	1	2	3	1	1	1

Measuring Variability and Spread

Exclude outliers scientifically – Quartiles

Points Scored per game	3	6	7	10	11	13	30
Frequency, f	2	1	2	3	1	1	1

3 , 3 , 6 , 7 , 7 , 1 0 , 1 0 , 1 0 , 1 1 , 1 3 , 3 0

Median = 10

First Quartile : 3 , 3 , 6 , 7 , 7 , 1 0

Q1 = 6.5

Third Quartile: 1 0 , 1 0 , 1 0 , 1 1 , 1 3 , 3 0

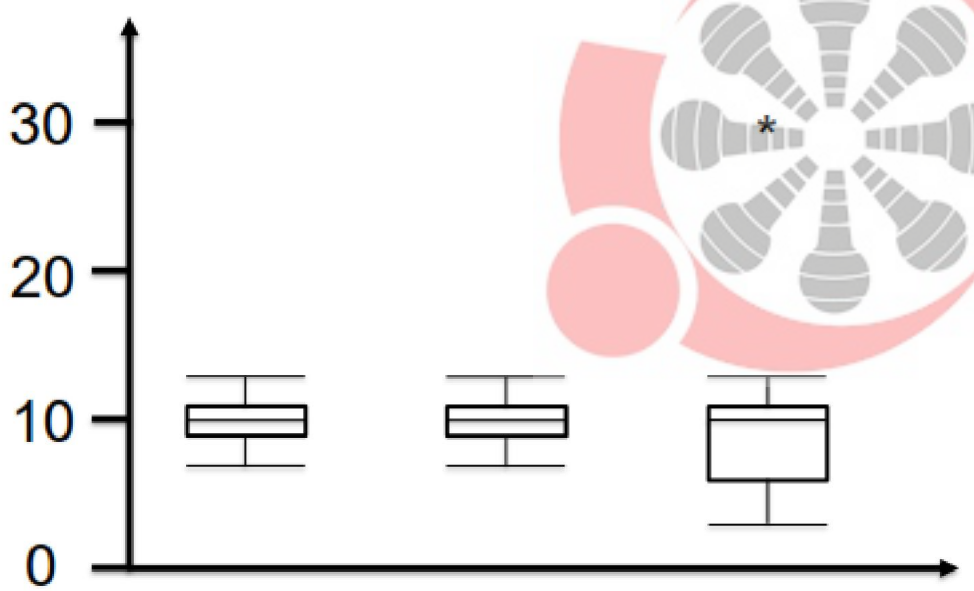
Q3 = 10.5

Box – Whisker Plot

- The Box and Whisker plot allows you to visualize the spread in the data easily
- Steps
  - Compute the Q1, Median and Q3 for the data. Compute IQR=Q3-Q1
  - The Box of the plot is drawn from Q3 to Q1 (50% of data is contained within the box)
  - The Whiskers are a maximum of 1.5*IQR from the top and the bottom of the box.
  - If there are no data points at 1.5*IQR, then pick an actual data point within the range of the Whiskers
  - Points lying outside the 1.5*IQR from the box ends are considered as Outliers.

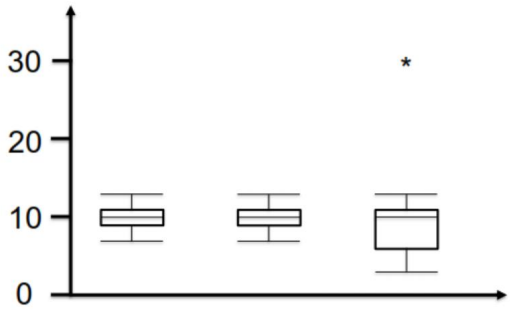
Measuring Variability and Spread

- Exclude outliers scientifically – Quartiles
- Box and whisker diagram or Box plot



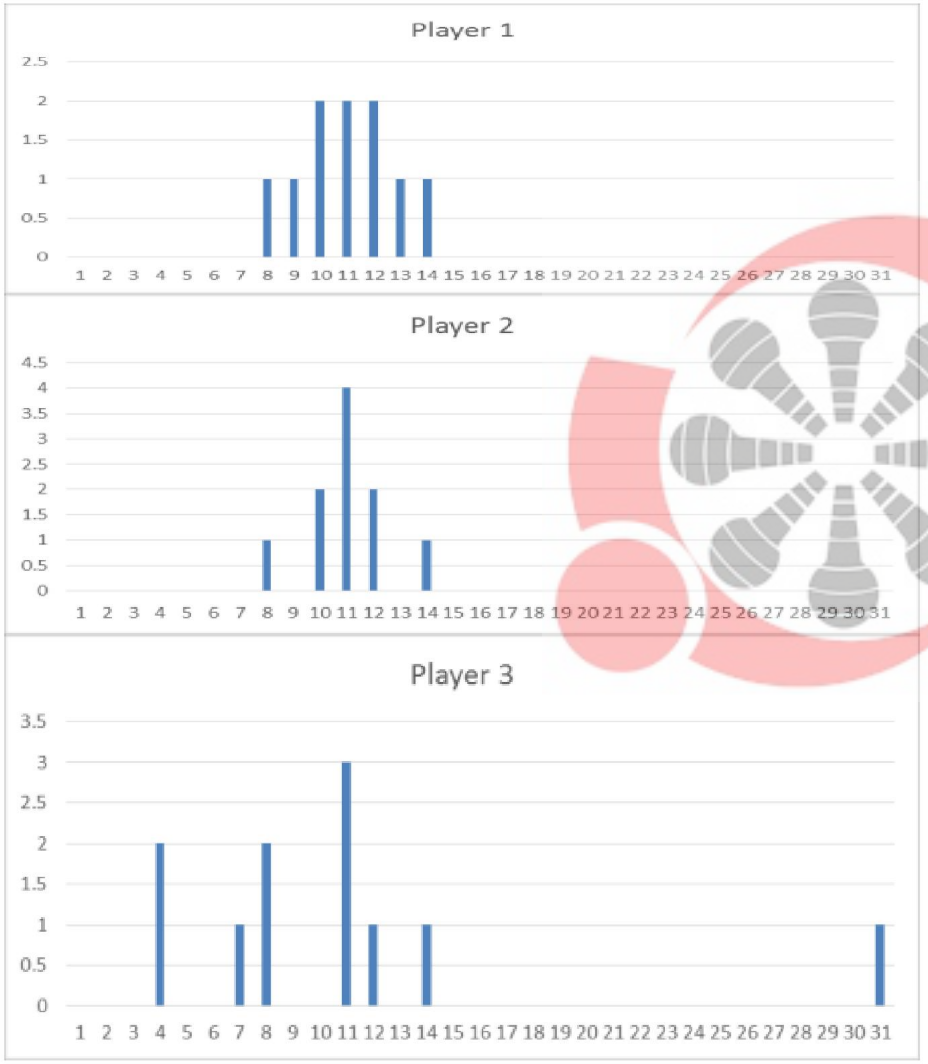
Measuring Variability and Spread

- Exclude outliers scientifically – Quartiles
- Box and whisker diagram or Box plot



Name	Formula	Player 1	Player 2	Player 3
Lower Hinge	Q1 = 1st Quartile	9	9	6.5
Mid Line	Q2 = 2nd Quartile = Median	10	10	10
Upper Hinge	Q3 = 3rd Quartile	11	11	10.5
Body of the box	IQR = Q1 - Q3	2	2	4
Step	1.5* IQR	3	3	6
	Lower Hinge - 1 Step	6	6	0.5
	Upper Hinge + 1 Step	14	14	16.5
Lower Fence	Smallest Actual Data Inside Fence	7	7	3
Upper Fence	Largest Actual Data Inside Fence	13	13	13
Outliers	Value beyond the Fence			30

Attention Check



1.73, 1.48, 7.02

Player 3 is the least reliable.

Measuring Variability and Spread

What happens to Standard Deviation if Good Heart Inc. gave all employees a Rs 2000 raise ?

No Change

What happens to Standard Deviation if Good Heart Inc. gave all employees a 10% raise ?

Increases by 1.1 times

Z - Score

- How far is any given data point from the mean ? (Distance)
  - Z – score can help us answer
- How many standard deviation away (above and below) from the mean is a data point ?
- Units for Z- score is “standard deviation”
- Z – score is measure of distance from mean.

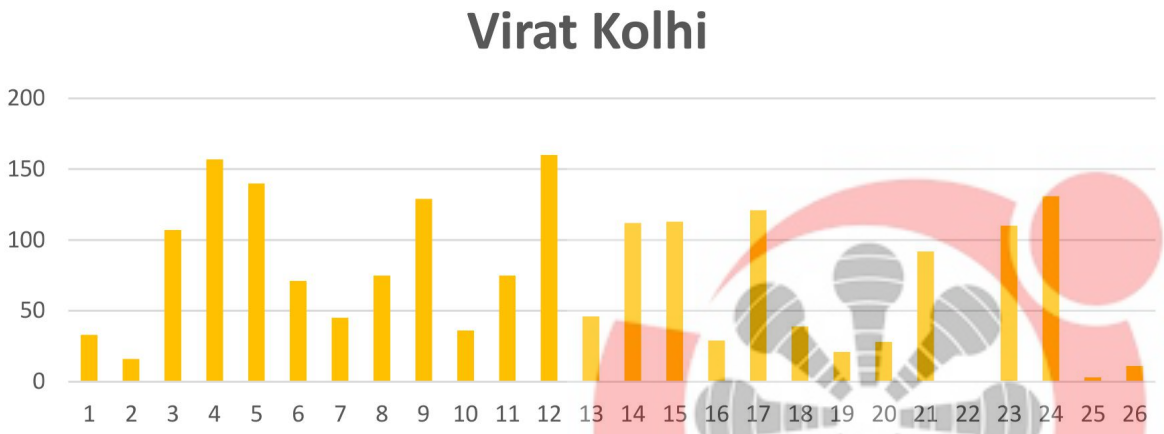
$$Z = \frac{x - \mu}{\sigma}$$

Measuring Variability and Spread

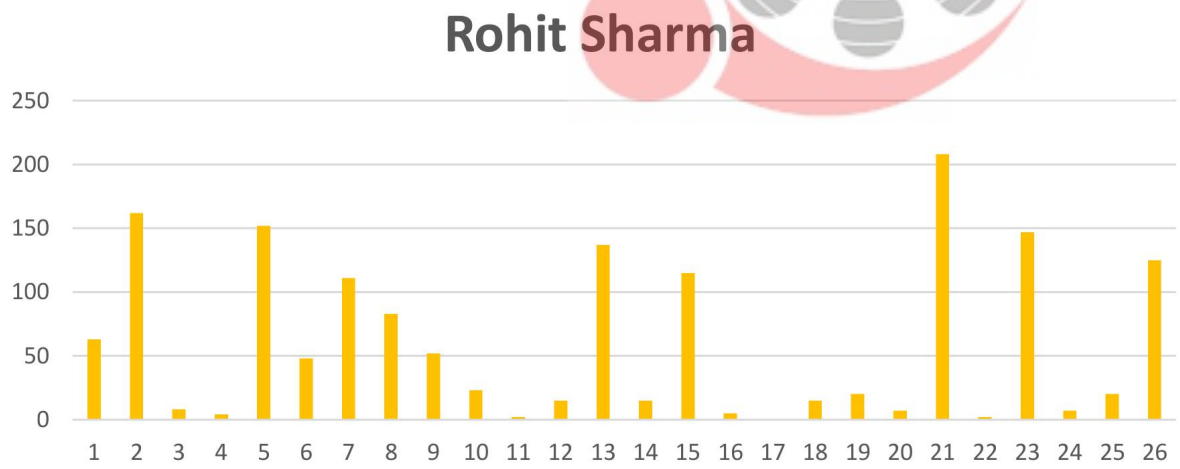
Imagine Virat Kolhi and Rohit Sharma with different abilities: Virat has an average of 73 with 50 stdev and the Rohit has average of 59 with 63 stdev in past 27 matches.

In a particular match session, the Virat scores 85 runs of the time and the Rohit scores 75 Runs. Who did best against their PERSONAL track record?

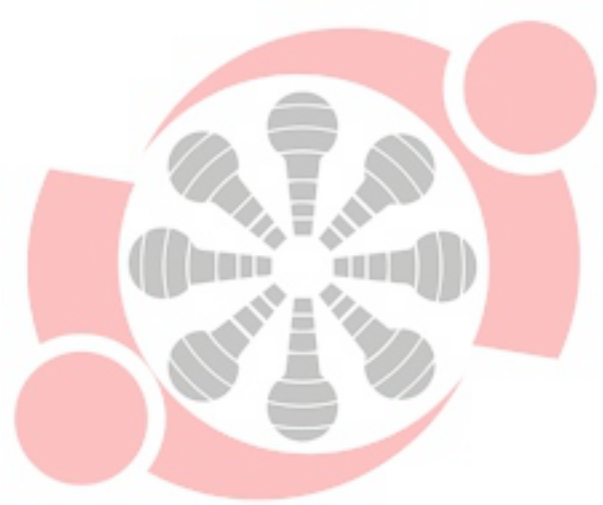
Measuring Variability and Spread



Mean = 73
Std = 50

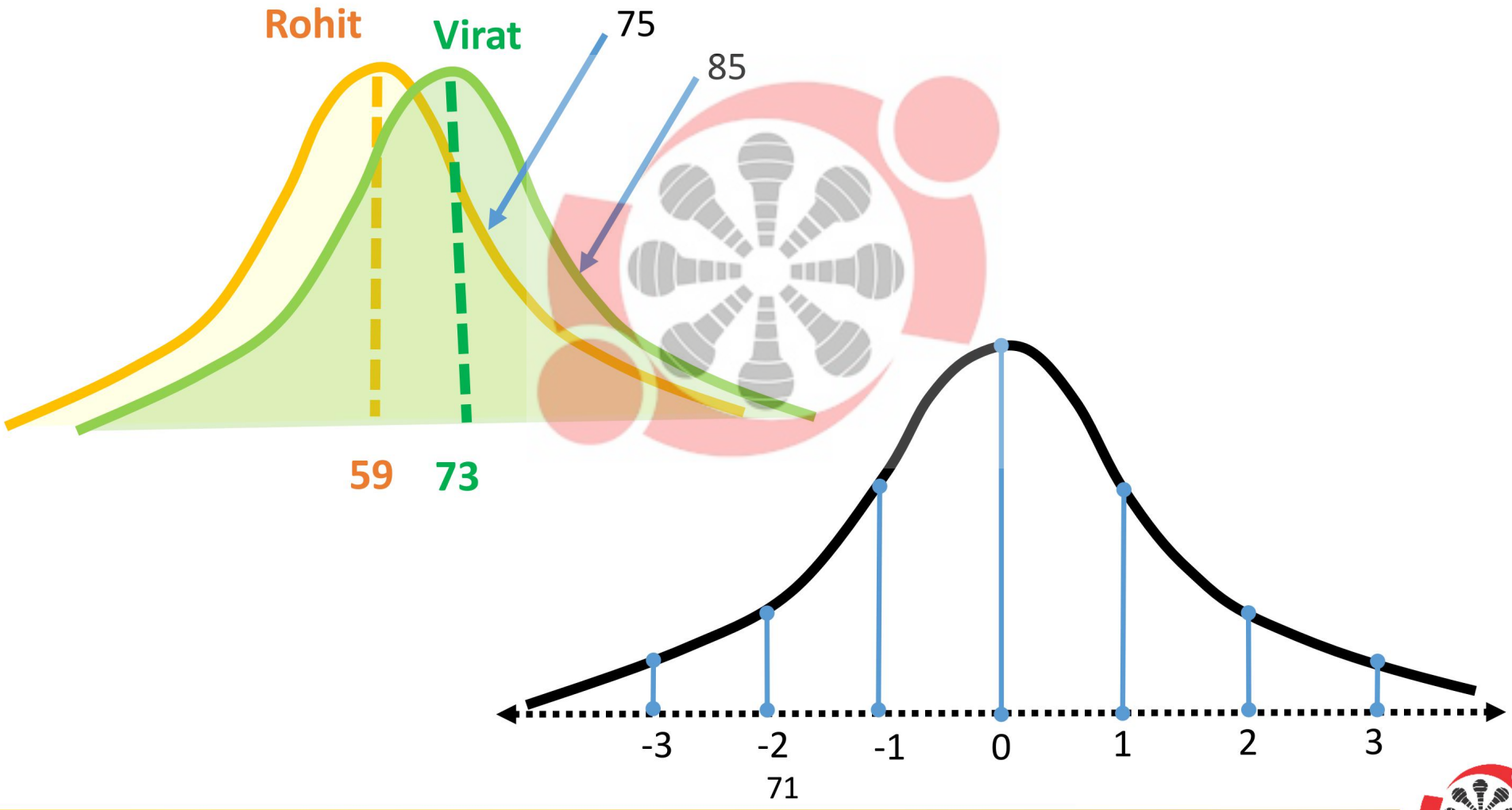


Mean = 59
Std = 63



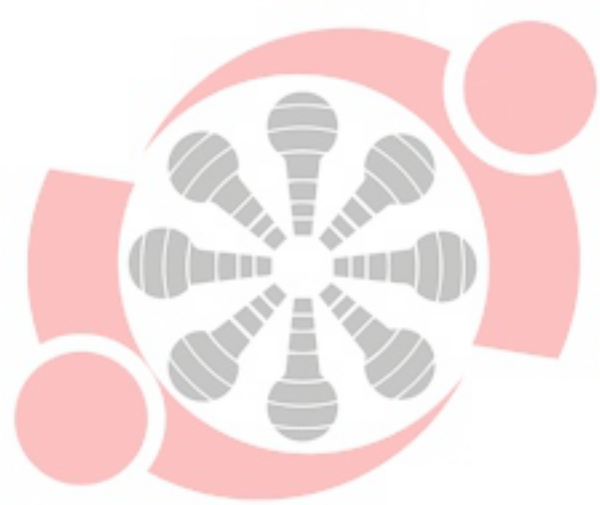
Measuring Variability and Spread

- Standard score, $z = \frac{x - \mu}{\sigma}$, # of stdevs from the mean



Reference

- https://www.mathsisfun.com/data/index.html
- **Head First: Statistics**
- **KhanAcademy**



INNOMATICS RESEARCH LAB