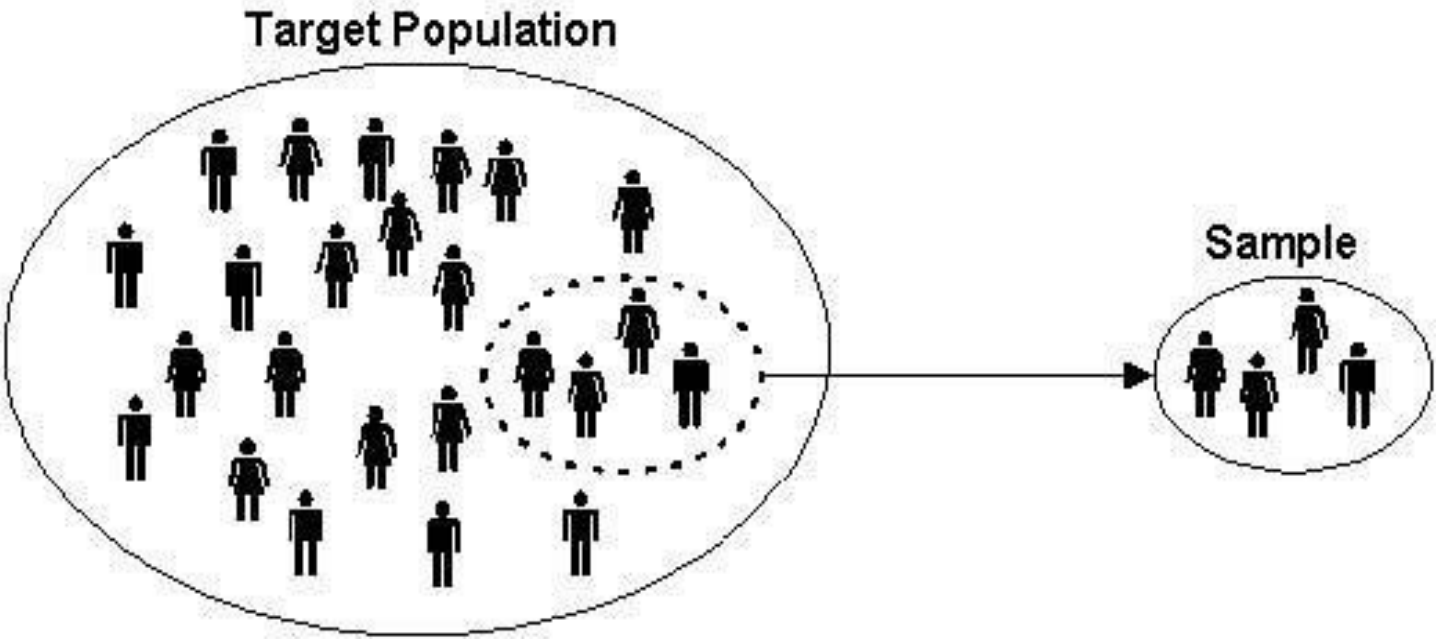
Basic Statistical Terminology

Statistics – Big Picture

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

- Data Gathering
- Data Understanding
- Data Analysis/Interpretation
- Data Presentation

Population and Sample



Source: http://www.snapsurveys.com/blog/wp-content/uploads/2011/08/target-population.jpg

Last accessed: October 7, 2014

Census and Survey

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

  For example, elections, 10-year census, etc.

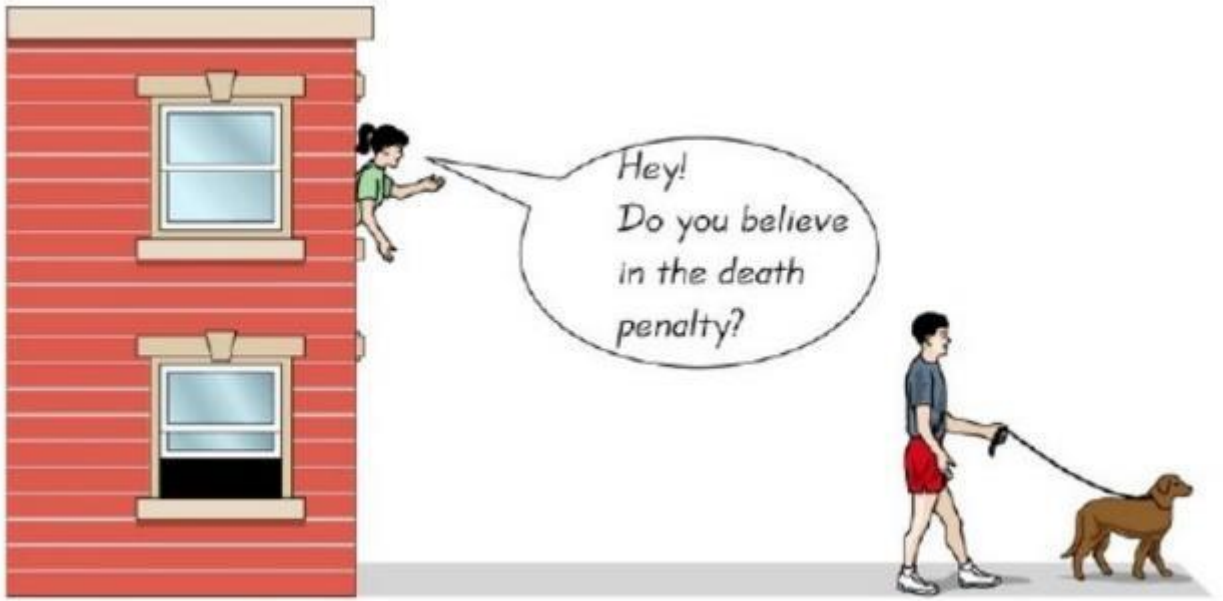
- **Survey:** Gathering data from the **sample** in order to make conclusions about the population.

  For example, opinion polls, quality control checks in manufacturing units, etc.

Statistics is “A telescope that allows us to study the large terrain and make it accessible to our unaided vision”

Data Gathering – Sampling Techniques

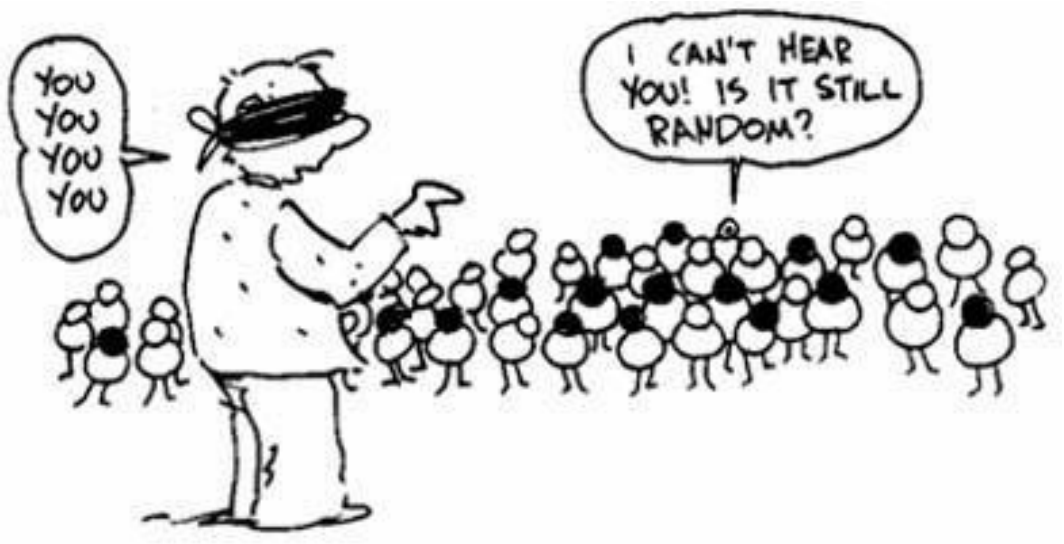
Convenience Sampling



Eg: Online polls, Asking your best friends etc

Data Gathering – Sampling Techniques

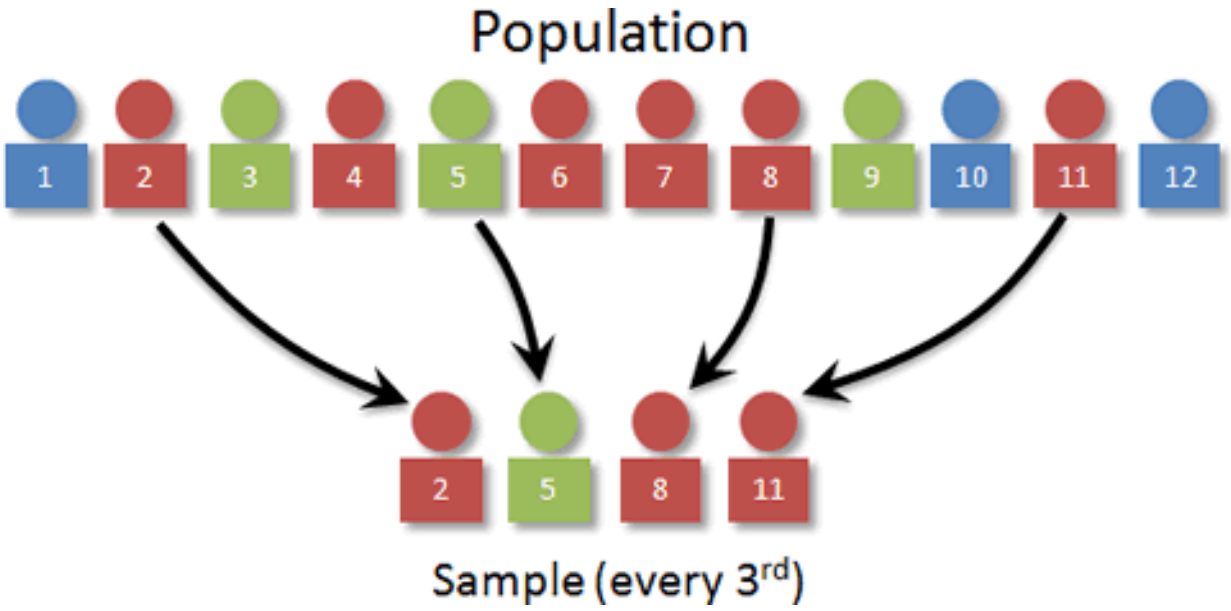
- Random Sampling



- Each member has an equal chance of being selected.

Sampling Techniques

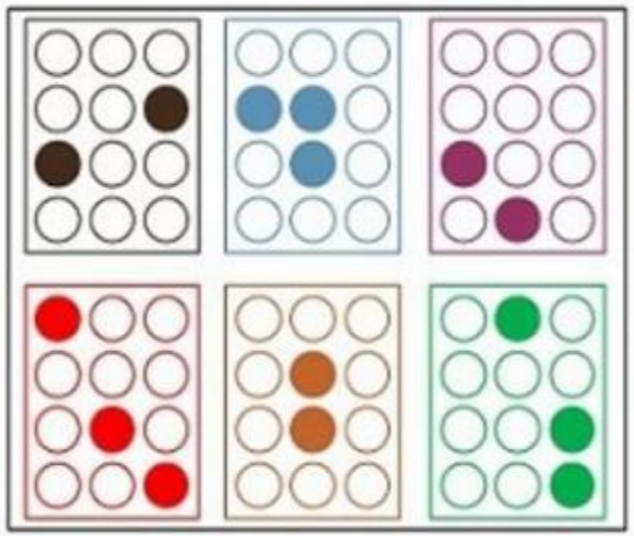
- Systematic Random Sampling



- Example: Supermarket chooses every 10th or 15th customer entering the supermarket and conduct the survey.

Sampling Techniques

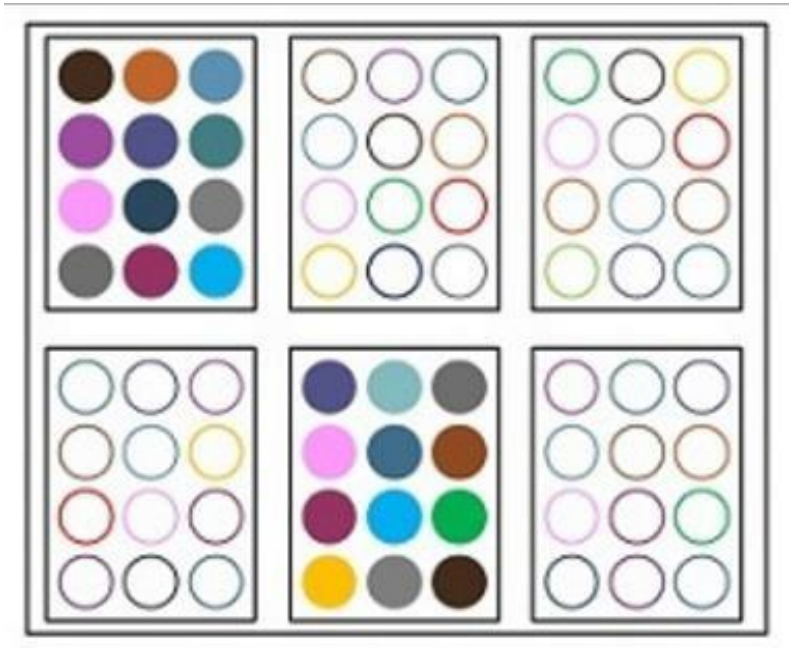
- Stratified Sampling
  - – Divide the data into several relevant strata and then sample from each strata



- Eg: For getting an opinion on demonetization, one choice of strata might be state-wise analysis. We get 20 random volunteers from each and every state.

Sampling Techniques

- Cluster Sampling
  - –Divide the population in to groups or clusters. Then select a one or a few clusters and survey **everyone** from the chosen subset.

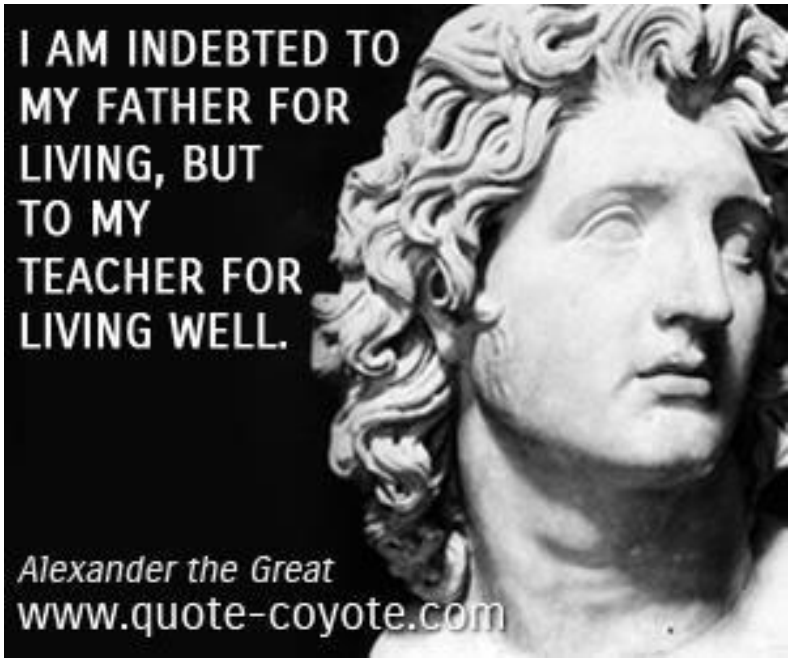


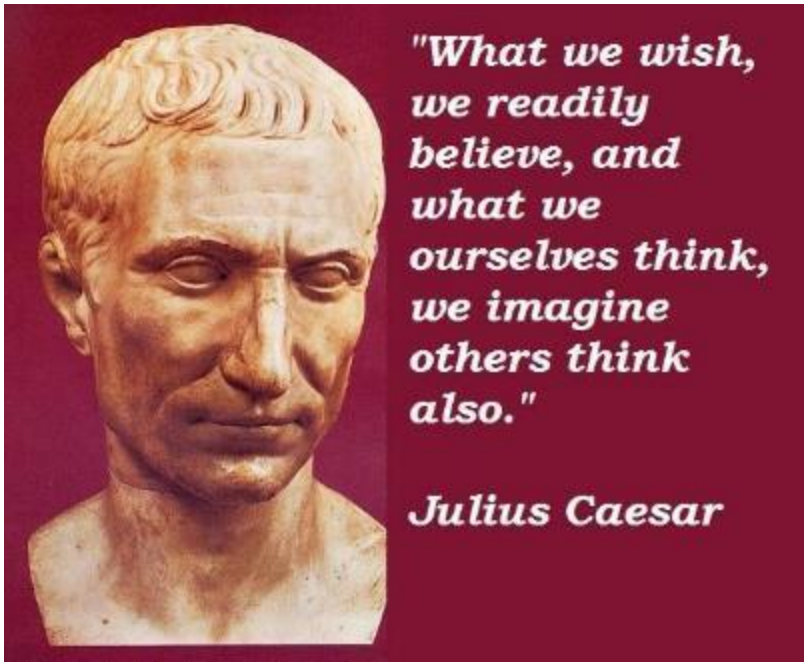
Parameter and Statistic

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 – Population Parameter

- Mean - μ
- Variance – σ^2
- Standard Deviation - σ

Roman – Sample Statistic

- Mean - x
- Variance – s^2
- Standard Deviation - s

Statistics

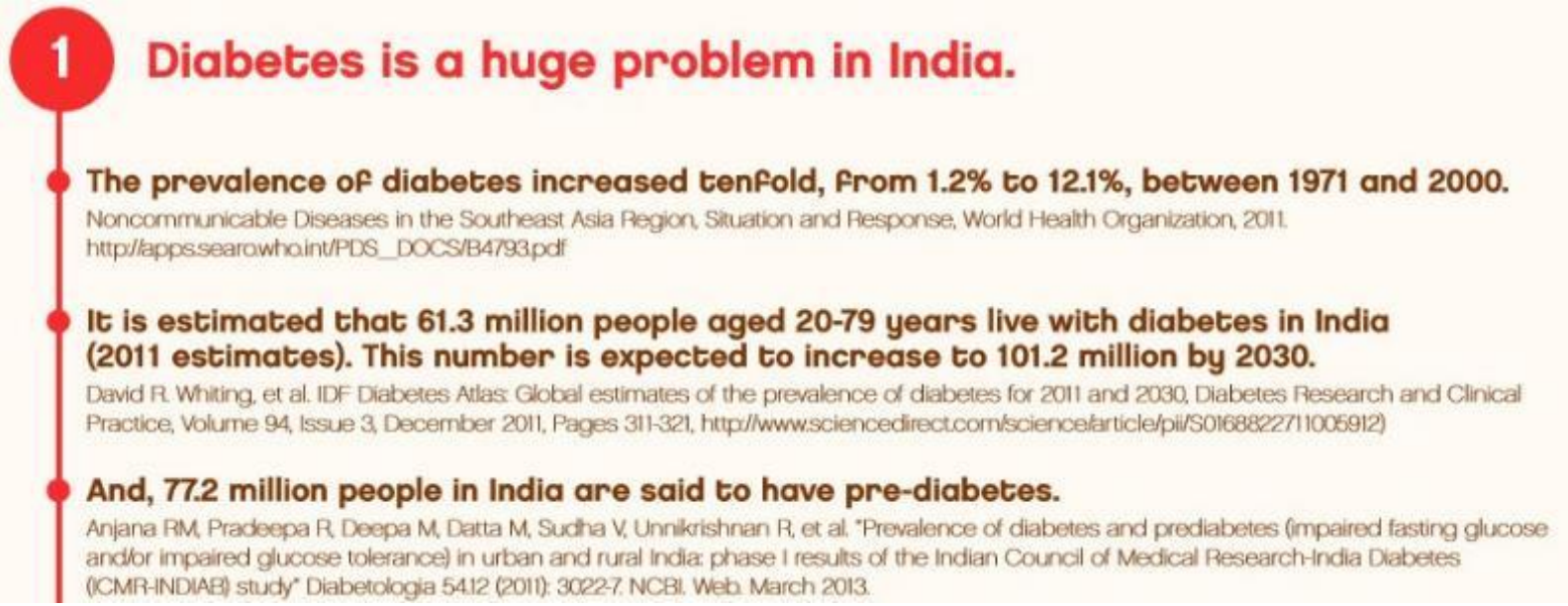
Descriptive Statistics

♣ 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

Descriptive and Inferential Statistics



1 Diabetes is a huge problem in India.

The prevalence of diabetes increased tenfold, from 1.2% to 12.1%, between 1971 and 2000.

Noncommunicable Diseases in the Southeast Asia Region, Situation and Response, World Health Organization, 2011. http://apps.searo.who.int/PDS_DOCS/B4793.pdf

It is estimated that 61.3 million people aged 20-79 years live with diabetes in India (2011 estimates). This number is expected to increase to 101.2 million by 2030.

David R. Whiting, et al. IDF Diabetes Atlas: Global estimates of the prevalence of diabetes for 2011 and 2030, Diabetes Research and Clinical Practice, Volume 94, Issue 3, December 2011, Pages 311-321, http://www.sciencedirect.com/science/article/pii/S0168822711005912)

And, 77.2 million people in India are said to have pre-diabetes.

Anjana RM, Pradeepa R, Deepa M, Datta M, Sudha V, Unnikrishnan R, et al. "Prevalence of diabetes and prediabetes (impaired fasting glucose and/or impaired glucose tolerance) in urban and rural India: phase I results of the Indian Council of Medical Research-India Diabetes (ICMR-INDIAB) study" Diabetologia 54.12 (2011): 3022-7. NCBI. Web. March 2013.

Source: http://www.arogyaworld.org/wp-_content/uploads/2010/10/Arogya_World_IndiaDiabetes_FactSheets_ _CGI2013_web.pdf

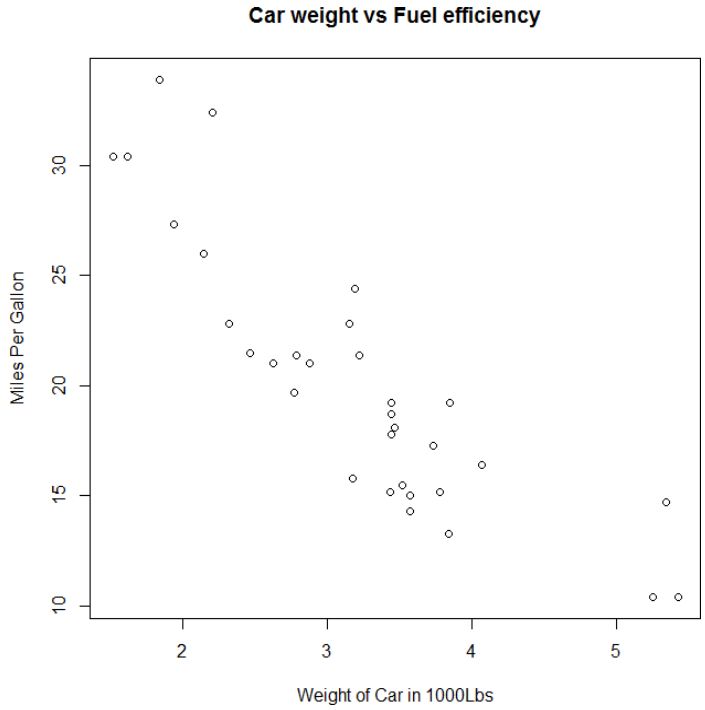
Last accessed: November 25, 2015

Variables and Data

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 Arrow the Dog?

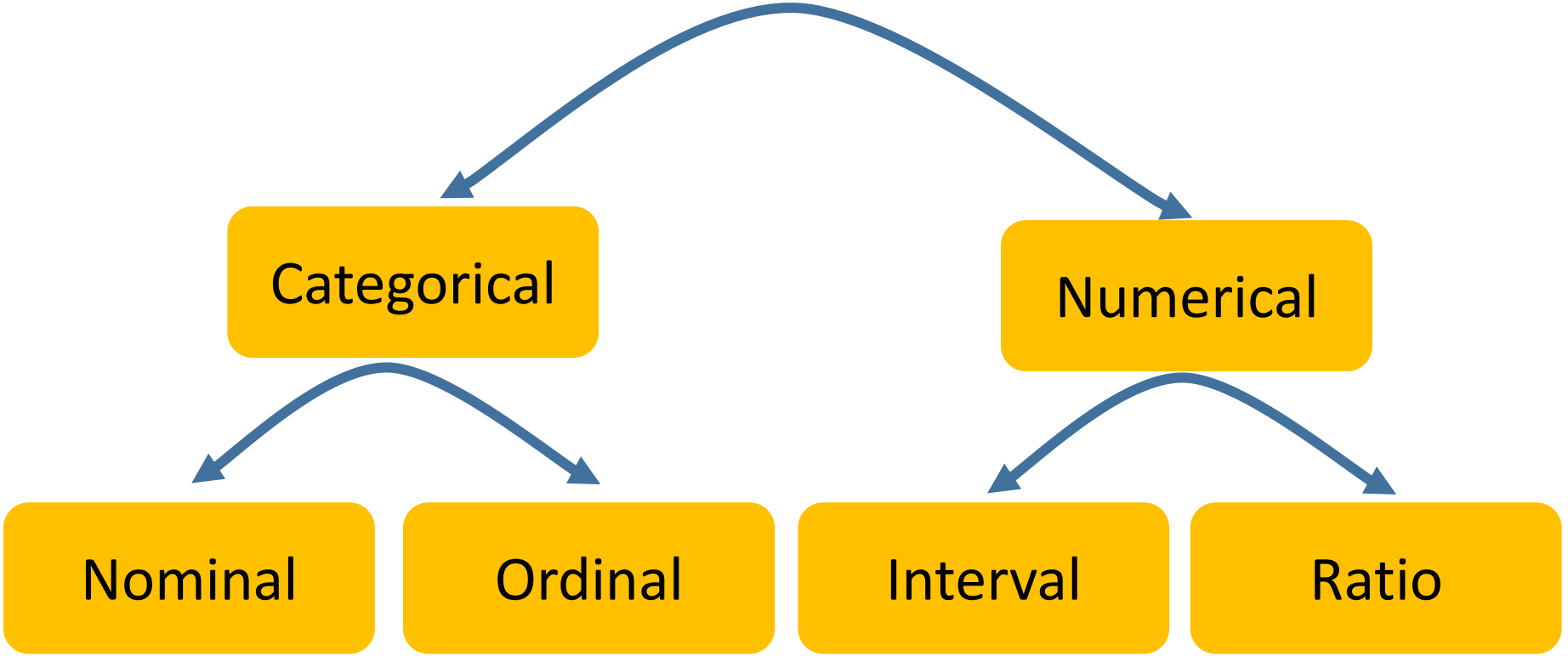
- **Qualitative**:
  - He is brown and black
  - He has long hair
  - He has lots of energy
- **Quantitative**:
  - Discrete:
    - He has 4 legs
    - He has 2 brothers
  - Continuous:
    - He weighs 25.5 kg
    - He is 565 mm tall

Data – Numeric and Categorical

18

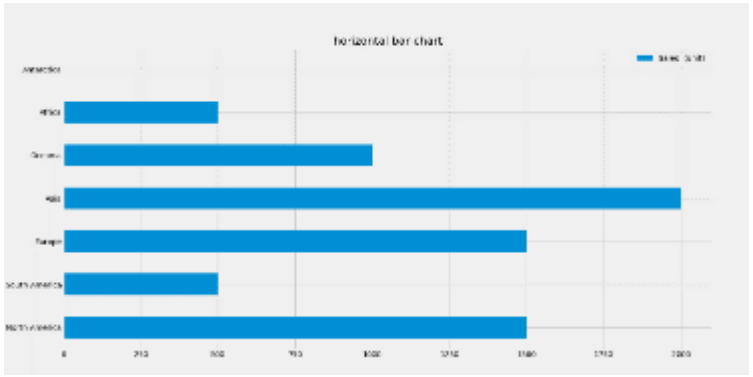
7

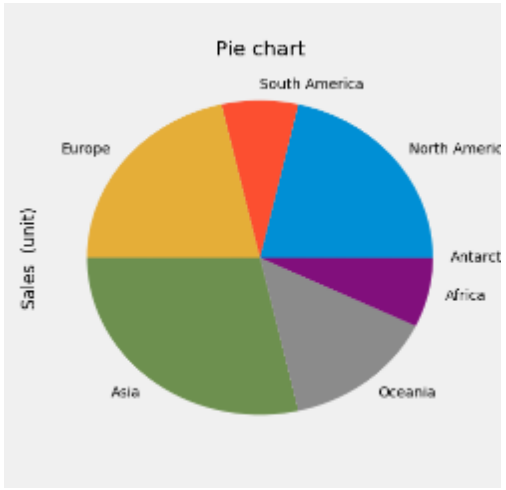
Data



Nominal Data (Qualitative)

- Nominal means name and count
  - Data are alphabetic or numerical in name
- They are categories without order or direction
- They are used to track people , object or event





Continent	Sales (unit)
North America	1,500
South America	500
Europe	1,500
Asia	2,000
Oceania	1,000
Africa	500
Antarctica	1

Ordinal Data (Qualitative)

- Ordinal means rank or order
- Data place in order. They are ordered categories like ranking or scaling.
- Has no absolute value
- More precise comparison are not possible

Categorical Data (Qualitative)

Nominal

- Employee ID
- Gender
- Religion
- Ethnicity
- Pin codes
- Place of birth
- Aadhaar numbers

Ordinal

- Mutual fund risk ratings
- Fortune 50 rankings
- Movie ratings

While there is an order, difference between consecutive levels are not always equal.

Quantitative Data - Interval

Data where ordering is clear and the 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.

- However, there is no natural zero or origin
- Example: Year 1008 vs 2016

Temperature: 14C vs 28C

Quantitative Data - Ratio

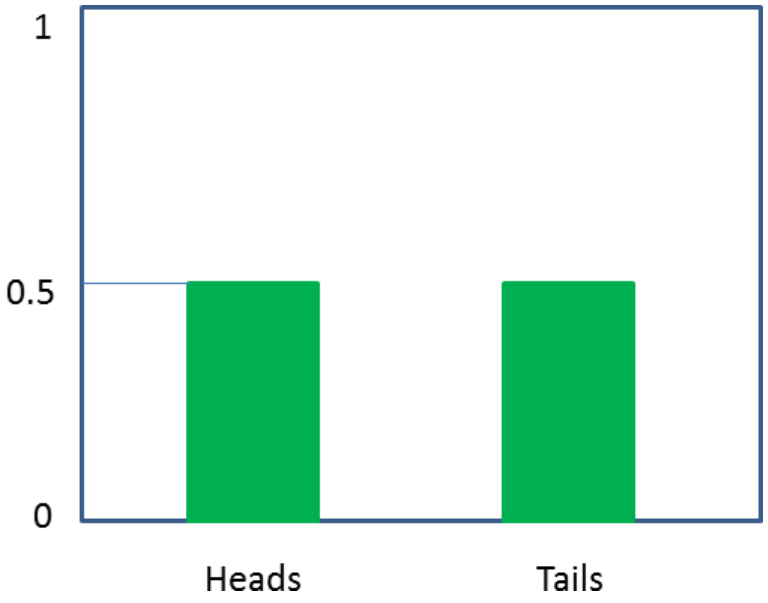
- Ratio level data is similar to Interval level data, with the key difference – there is a natural zero point.

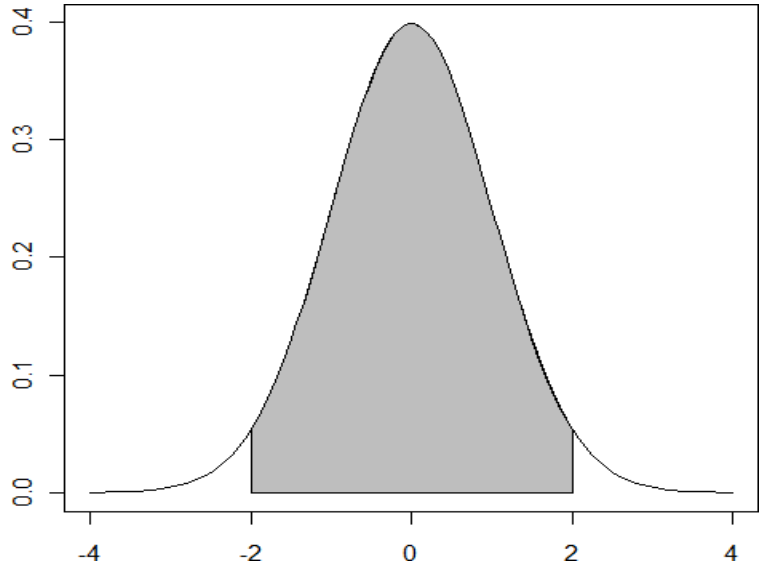
- Examples: Weights, Cost of things, Number of correct answers in a exam

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
Train between customer arrivals at retail outlet	Continuous
Sampling the volume of liquid nitrogen in a storage tank	Continuous
Sampling 100 voters in a exit poll and determining how many voted for the wining candidate	Discrete
Length of newly designed automobiles	Continuous
No. of customers arriving at a retail outlet during a five minute period.	Discrete
No. of defects in a batch of 50 items	Discrete

Describing Data through Statistics

Descriptive Statistics

The Central Tendencies

RaghuRam want to join a health club in a activity that has others in the same age group as him. He is 22 years old. Mean ages for **YOGA**, **POWER WORKOUT** and **SWIMMING** classes are

15 years

20 years

17 years

The Central Tendencies

Yoga class composition

Age (years)	13	15	17
Frequency, f	1	3	2

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

$$\frac{13 * 1 + 15 * 3 + 17 * 2}{1 + 3 + 2}$$

The Central Tendencies

Power workout class composition

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

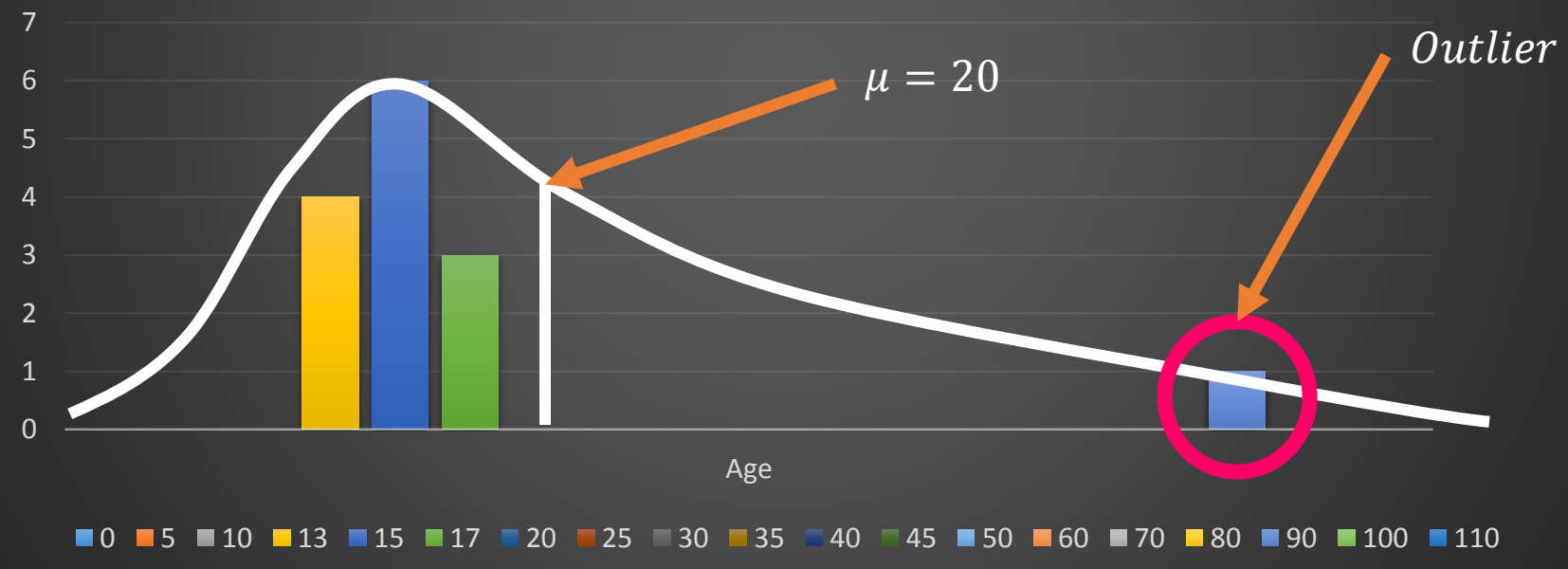
$$Mean, \mu = \frac{\sum x}{n} =$$

$$\frac{13 * 4 + 15 * 6 + 17 * 3 + 90 * 1}{4 + 6 + 3 + 1}$$

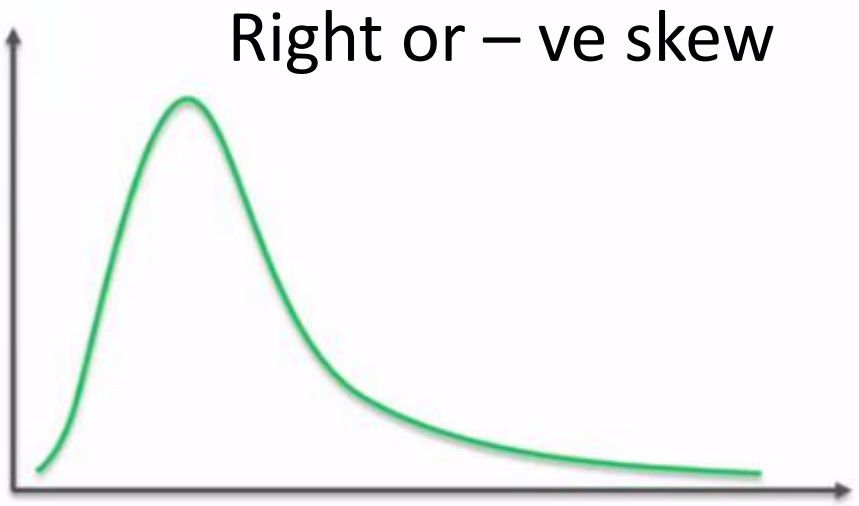
The Central Tendencies

Power Workout Class Composition

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




Left or – ve skew


Right or – ve skew

The Central Tendencies – Median

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

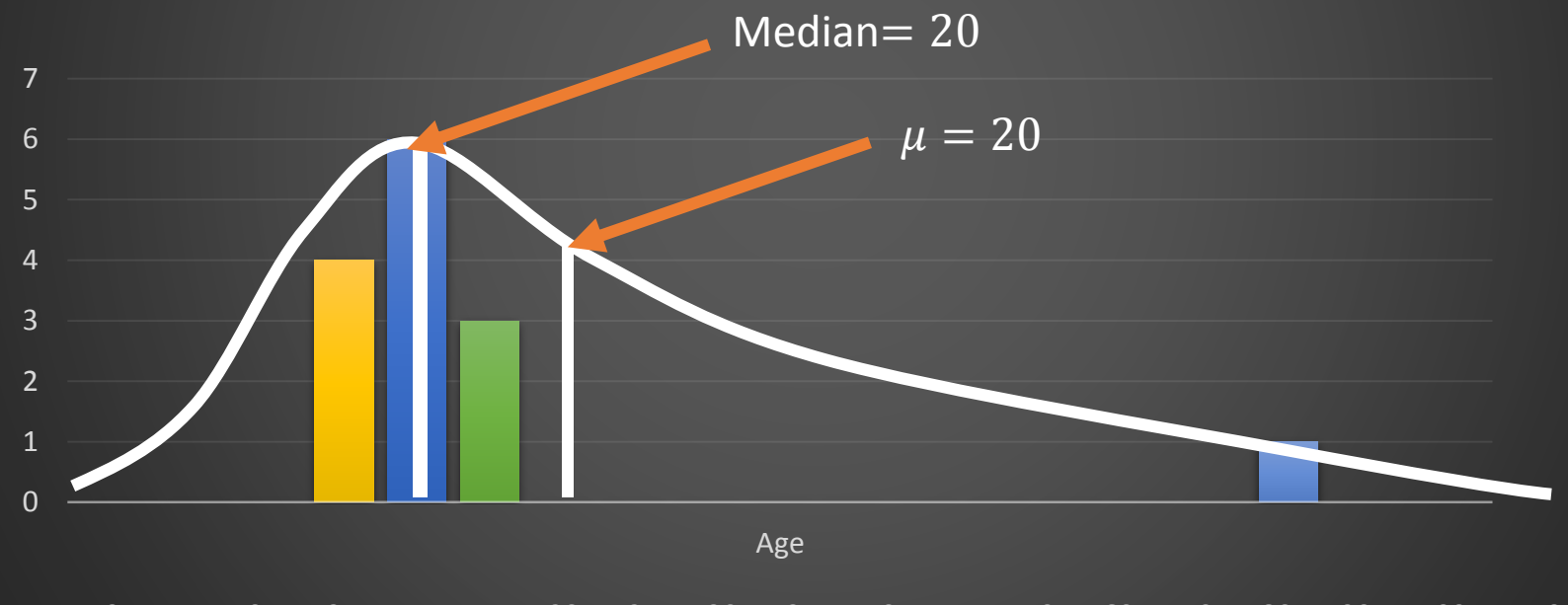
- Data has outlier
- Median - the mid-point

13, 13, 13, 13, 15, 15, 15, 15, 15, 15, 17, 17, 17, 90

The Central Tendencies

Power Workout Class Composition

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



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

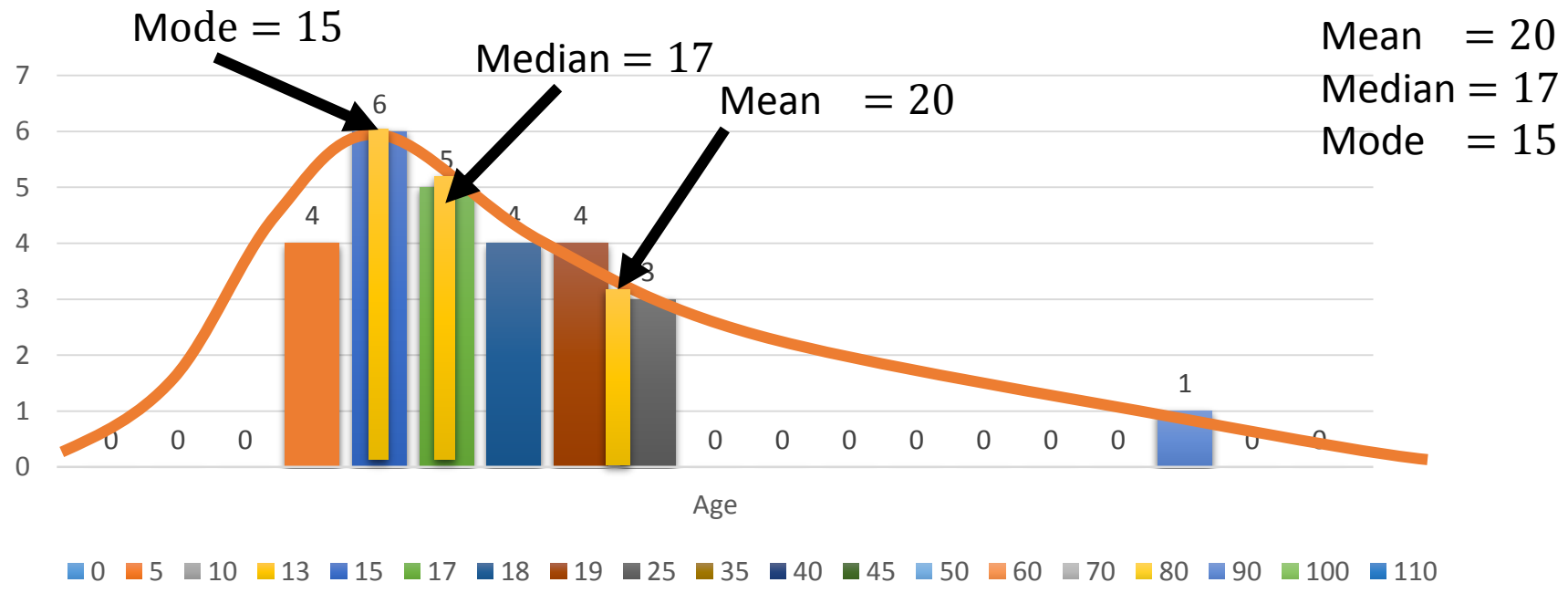
Mean and Median need not be in the dataset but Mode has to be in it.

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

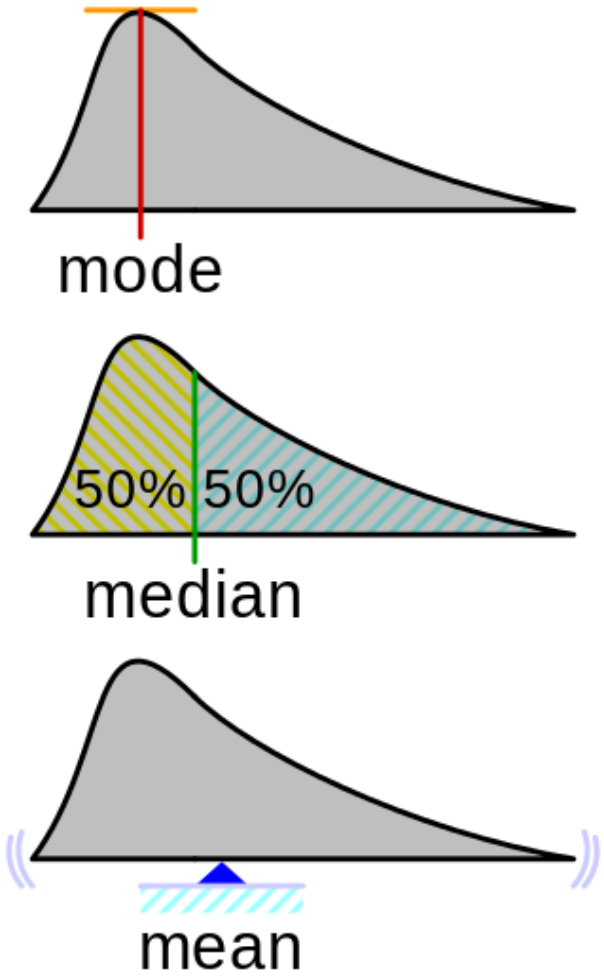
The Central Tendencies

Power Workout Class Composition

Age (years)	13	15	17	18	19	25	90
Frequency, f	4	6	5	4	4	3	1



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

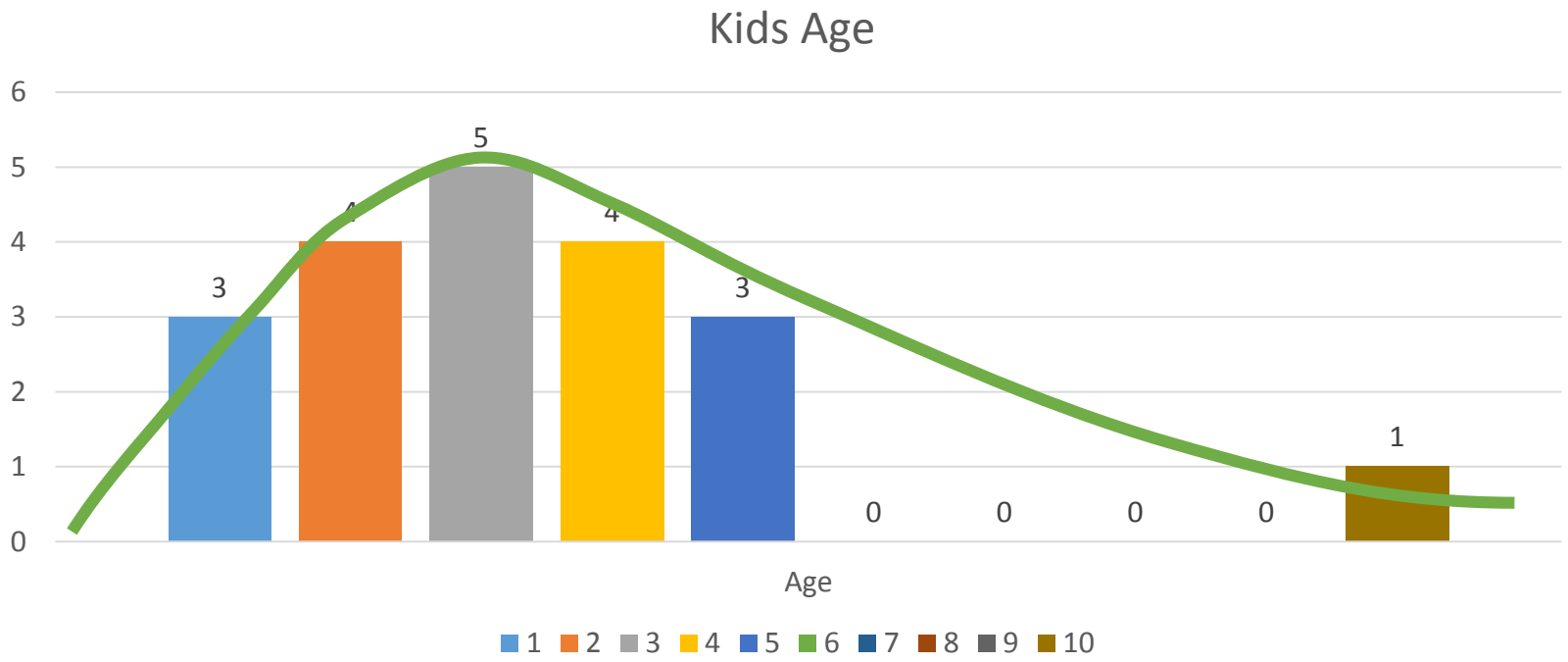
The range is a way of measuring how spread out a set of values are. It's given by Upper bound - Lower bound where the upper bound is the highest value, and the lower bound the lowest.

Range = upper bound - lower bound

= 10 -1

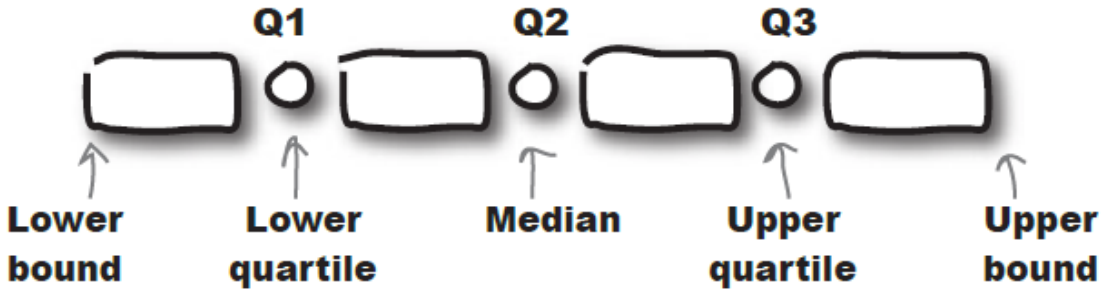
= 9

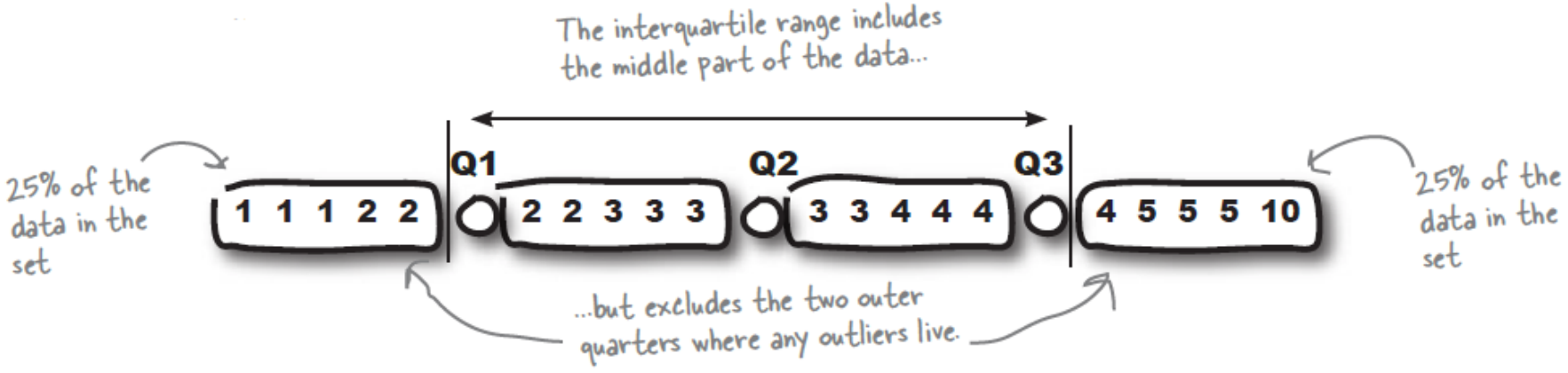
so, the range is 9


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) = = $\left(\frac{3*(n-1)}{4}+1\right)^{th}$

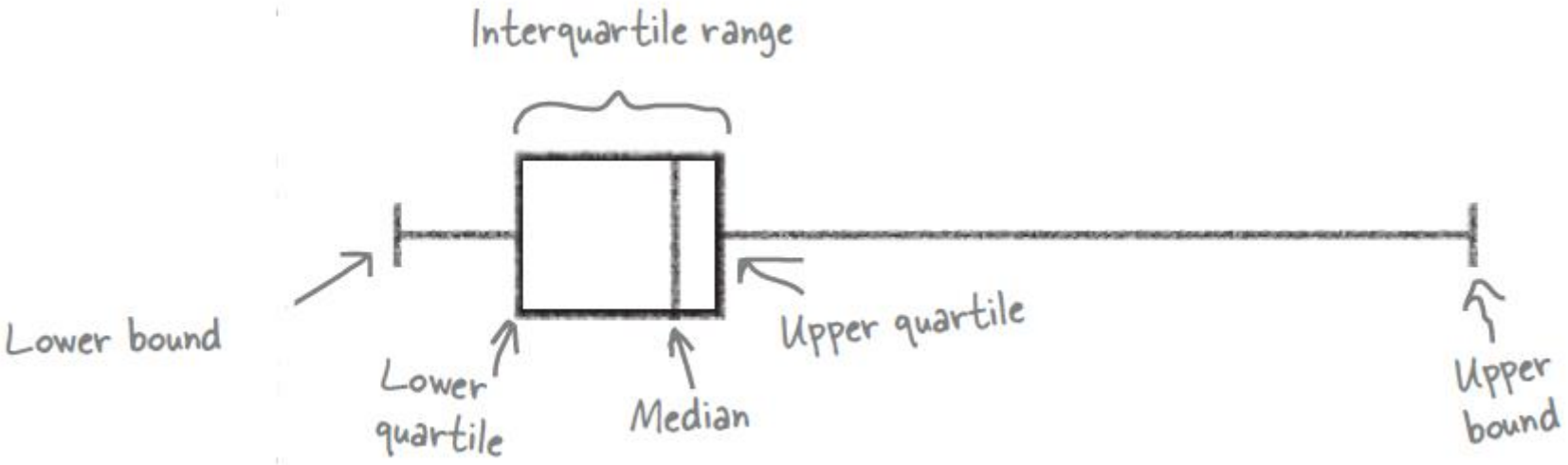
Interquartile range, IQR = Q3 – Q1 (central 50% of data)

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



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 Statson 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

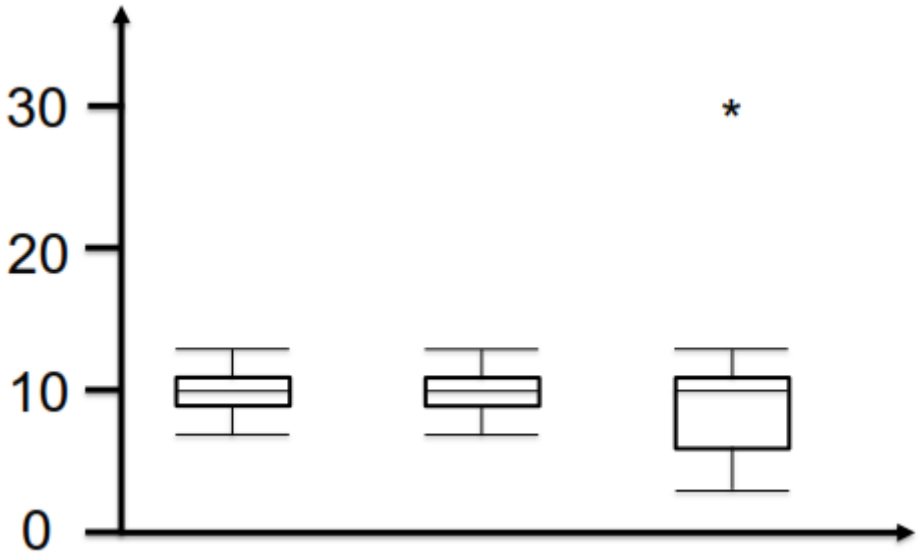
Q3 = 10.5

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.

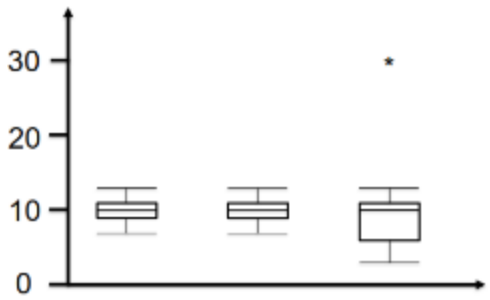
Measuring Variability and Spread

- 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



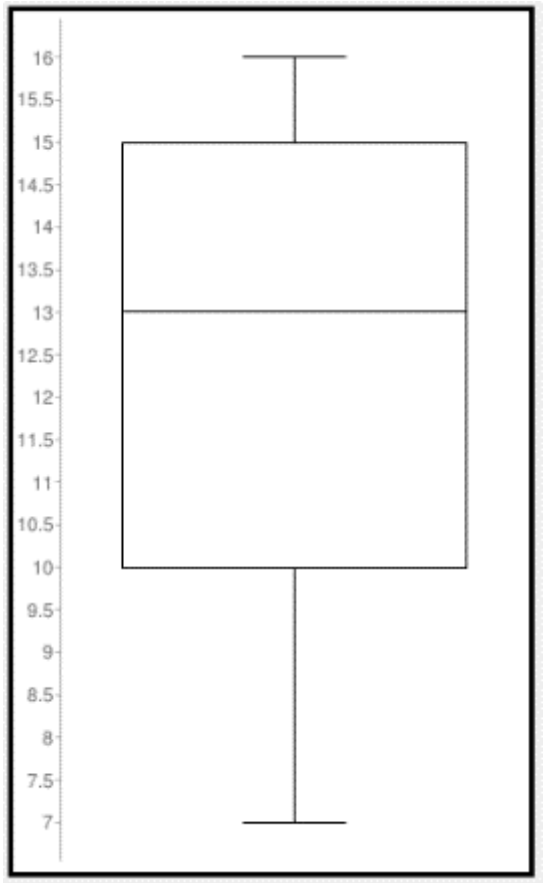
Name	Formula	Player 1	Player 2	Player 3
Lower Hinge	Q1 = 1st Quartile	9	9	6.5
Mid Line	Q2 = 2nd Quartile = Median	10	10	10
Upper Hinge	Q3 = 3rd Quartile	11	11	10.5
Body of the box	IQR = Q1 - Q3	2	2	4
Step	1.5* IQR	3	3	6
	Lower Hinge - 1 Step	6	6	0.5
	Upper Hinge + 1 Step	14	14	16.5
Lower Fence	Smallest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

Interpreting Box-whisker plot

Age of kids in a party

Which of the following statements are true?

- All of the students are less than 17 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



Attention Check

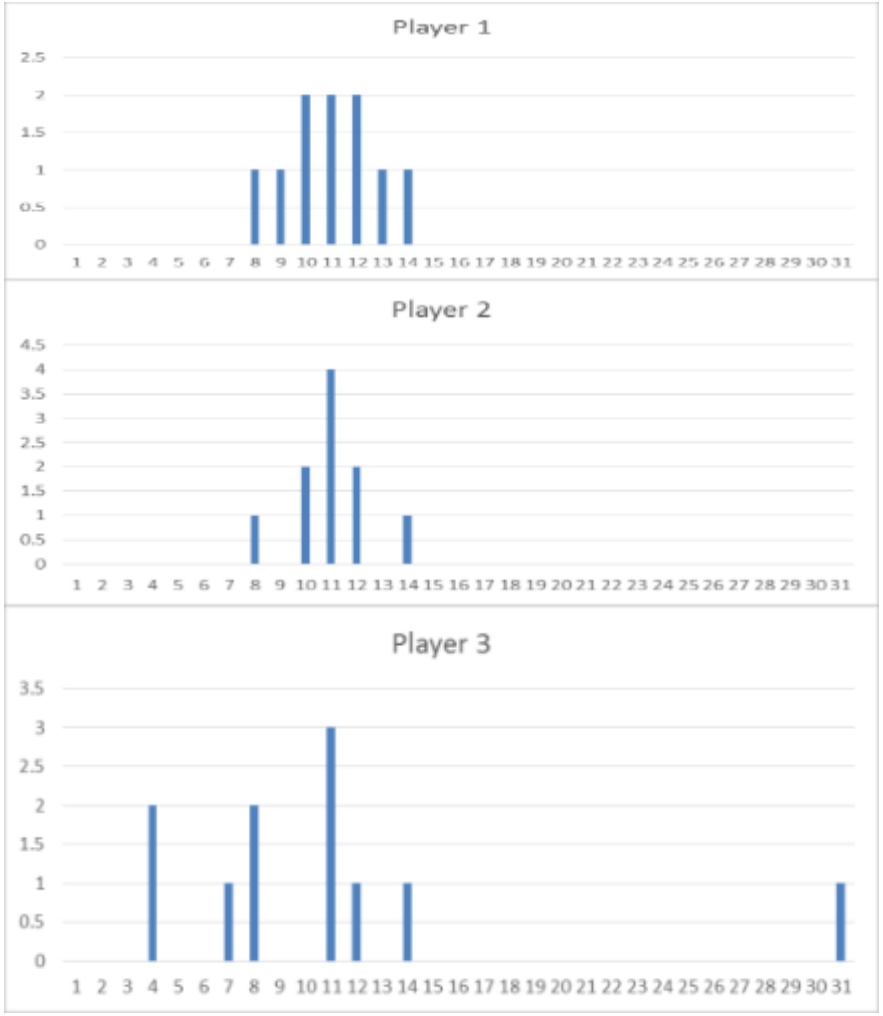
Basketball coach Statson is in a dilemma choosing between 3 players all having the same average scores.

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

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.

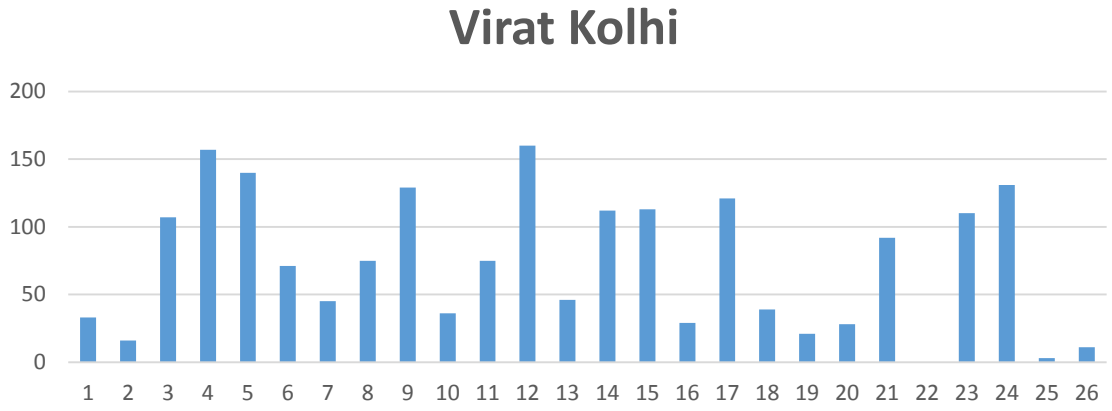
$$Z = \frac{x - \mu}{\sigma}$$

Measuring Variability and Spread

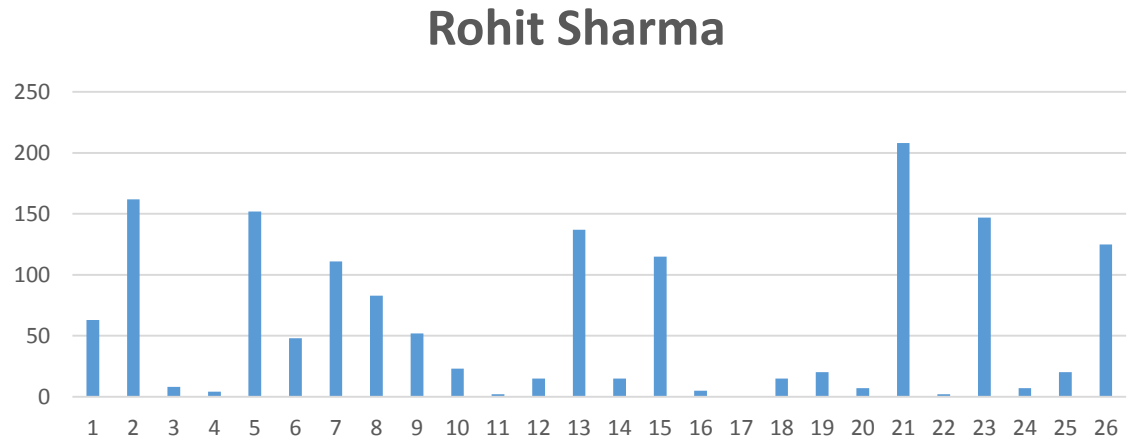
Imagine Virat Kolhi and Rohit Sharma with different abilities: Virat has an average of 73 with 50 stdev and the Rohit has average of 59 with 63 stdev in past 27 matches.

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



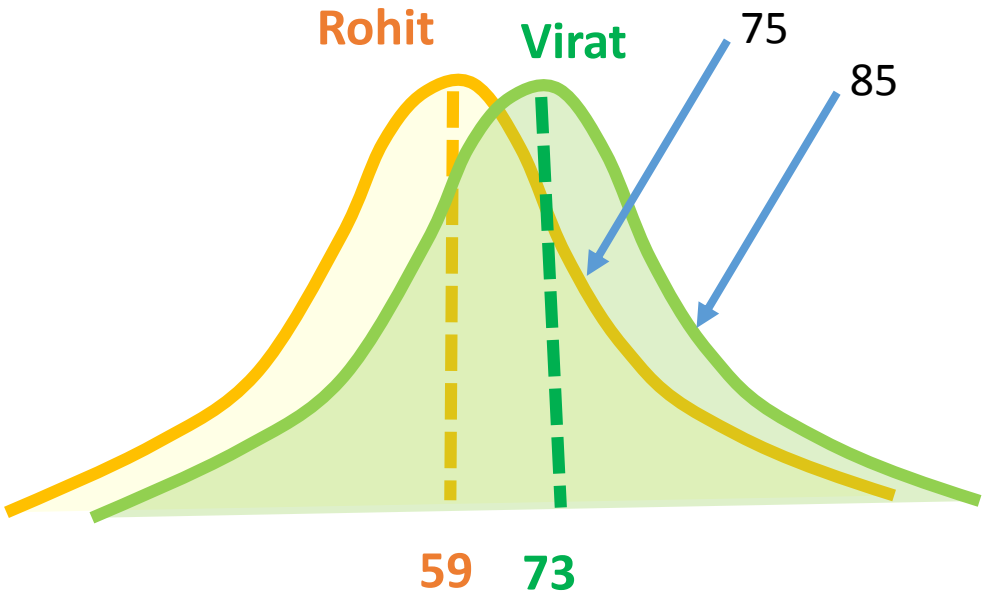
Mean = 73
Std = 50

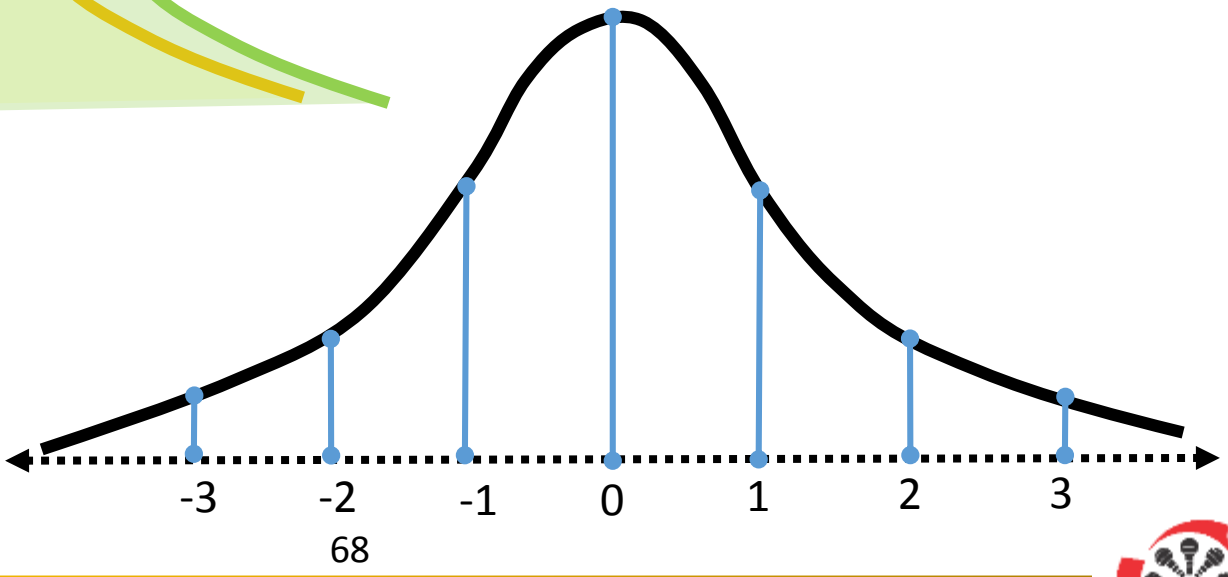


Mean = 59
Std = 63

Measuring Variability and Spread

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





Anscombe's quartet							
I		II		III		IV	
x	y	x	y	x	y	x	y
10	8.04	10	9.1	10	7.46	8	6.6
8	6.95	8	8.1	8	6.77	8	5.8
13	7.58	13	8.7	13	12.7	8	7.7
9	8.81	9	8.8	9	7.11	8	8.8
11	8.33	11	9.3	11	7.81	8	8.5
14	9.96	14	8.1	14	8.84	8	7
6	7.24	6	6.1	6	6.08	8	5.3
4	4.26	4	3.1	4	5.39	19	13
12	10.8	12	9.1	12	8.15	8	5.6
7	4.82	7	7.3	7	6.42	8	7.9
5	5.68	5	4.7	5	5.73	8	6.9

Property	Value
Mean of x in each case	9 (exact)
Sample variance of x in each case	11 (exact)
Mean of y in each case	7.50 (to 2 decimal places)
Sample variance of y in each case	4.122 or 4.127 (to 3 decimal places)
Correlation between x and y in each case	0.816 (to 3 decimal places)
Linear regression line in each case	$y = 3.00 + 0.500x$ (to 2 and 3 decimal places, respectively)