Descriptive Statistics:

Descriptive Statistics:

- Frequency Distribution
- Plots on Distributions
- Central Tendency – Mean Median and Mode
- Skewness, Kurtosis
- Quiz check
- Measure of Dispersion – Variability and Spread
- Range, IQR, Percentile, Box plot
- Variance, Standard Deviation z Score
- Scatter plot and Correlation
- Quiz Check

Frequency Distributions:

- A frequency distribution shows us a summarized grouping of data divided into mutually exclusive classes and the number of occurrences(Frequency) in a class.
- It is a way of showing unorganized data notably to show results of an election, income of people for a certain region, sales of a product within a certain period, student loan amounts of graduates, etc.
- Some of the graphs that can be used with frequency distributions are histograms, bar charts, pie charts, etc.
- Frequency distributions are used for both qualitative and quantitative data.

Qualitative - Bar Graph:

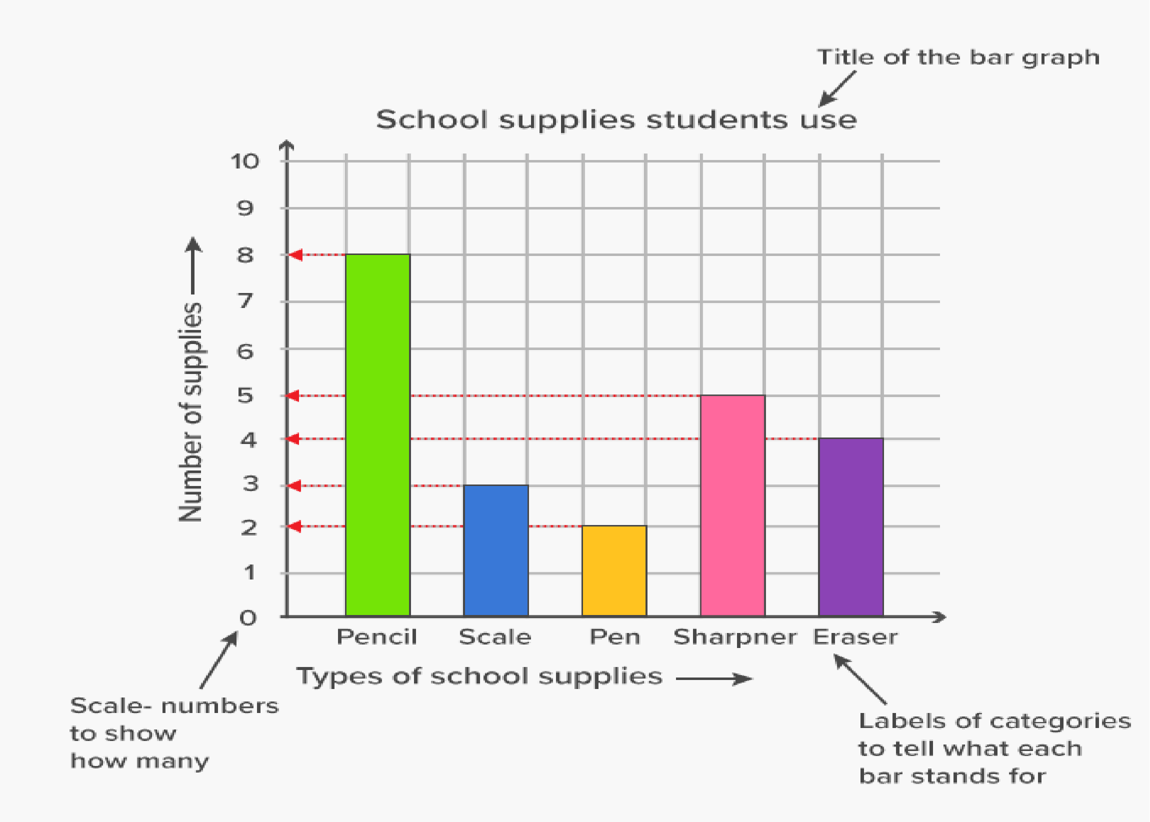
- A **bar chart** or **bar graph** is a chart or graph that presents categorical data with rectangular bars with heights or lengths proportional to the values that they represent. The bars can be plotted vertically or horizontally. A vertical bar chart is sometimes called a **column chart**.

Graphical Representation of Bar:

- A bar graph shows comparisons among discrete categories. One axis (X – axis) of the chart shows the specific categories(Class) being compared, and the other axis (Y – axis) represents a measured value(Frequency).

Class (School Supplies)	Frequency(No.of Suppliers)
Pencil	8
Scale	3
Pen	2
Sharpener	5
Eraser	4

Note: Frequency in Pandas are Value_Counts()



Qualitative - Pie Chart:

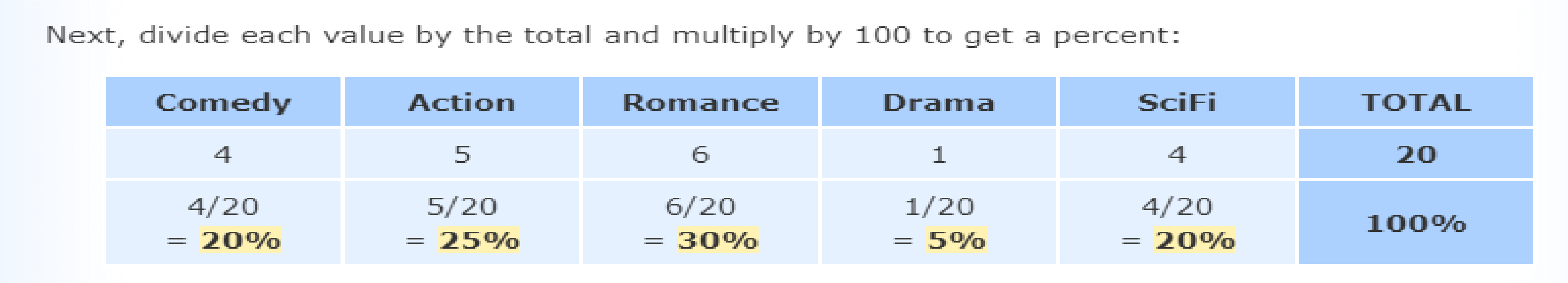
- Pie charts work by splitting your data into distinct groups or categories. The chart consists of a circle split into wedge-shaped slices, and each slice represents a group.

- The size of each slice is proportional to how many are in each group compared with the others.

- The number in a particular group is called the **frequency**

Pie Chart Calculation & Representation:

Comedy	Action	Romance	Drama	SciFi
4	5	6	1	4

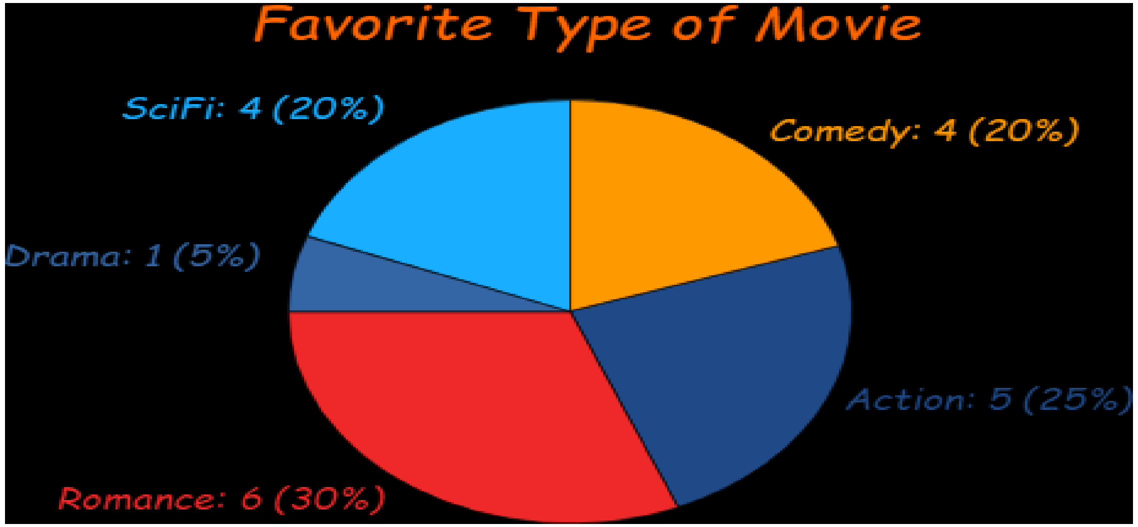


Next, divide each value by the total and multiply by 100 to get a percent:

Comedy	Action	Romance	Drama	SciFi	TOTAL
4	5	6	1	4	**20**
4/20 = **20%**	5/20 = **25%**	6/20 = **30%**	1/20 = **5%**	4/20 = **20%**	**100%**

- The larger the slice, the greater the relative popularity of that group.

- Here Action is relative popularity from others



Quantitative – Histogram:

- Histogram is a graphical representation of the Frequency Distribution of data in which the X-axis represents the classes and the Y-axis represents the frequencies in bars.

- To construct a histogram from a continuous variable you first need to split the data into intervals, called **bins**.

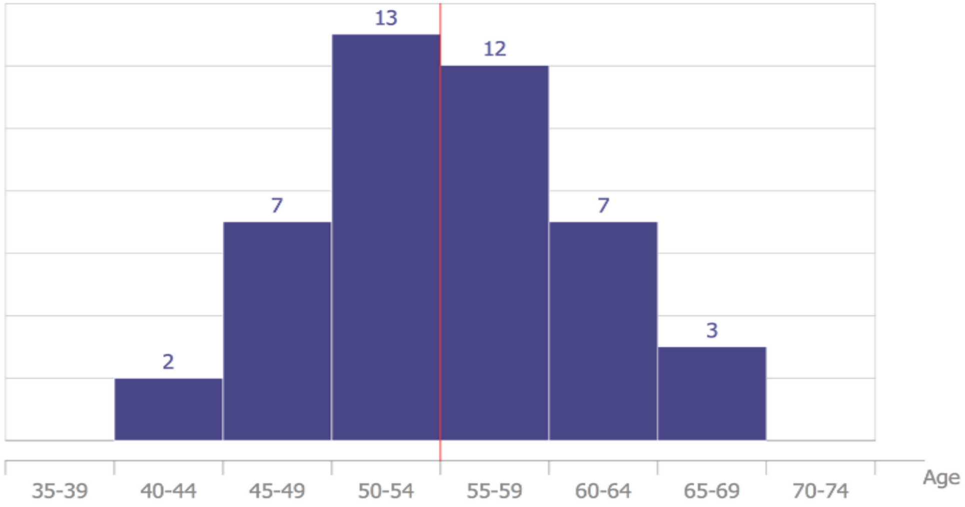
- Lets consider ages of 44 People working in a company below list.

- [41,44, 45,46,46,47,48,49,49,50,50,50,50,51,52,52,53,53,54,55,54,53, 55,55,55,56,56,57,57,57,58,58,59,59,59, 60,61,61,61,62,63,64, 66,68,69]

Visual Representation – Histogram:

- Considering **bin size** equal to 5 you will get distribution as mention below. Also notice the histogram you will get from the distribution shown in the table.

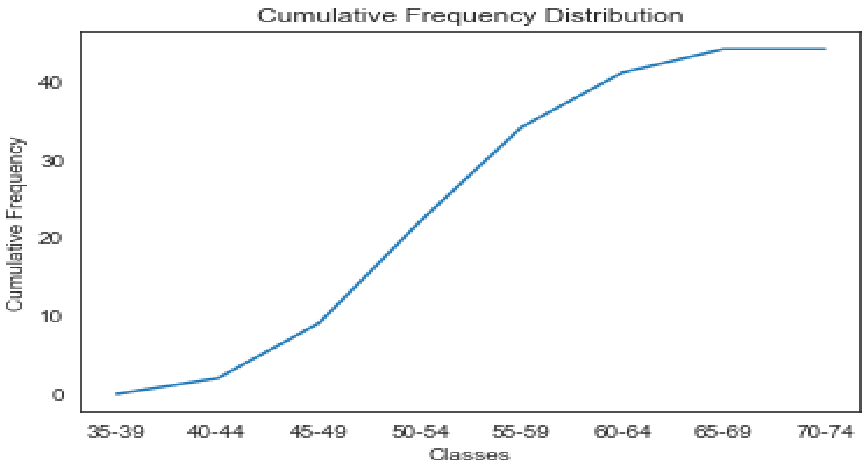
Class	Frequency	Scores included in Bin
35-39	0	
40-44	2	41,44
45-49	7	45,46,46,47,48,49,49
50-54	13	50,50,50,50,51,52,52,53,53,54,55,54,53
55-59	12	55,55,55,56,56,57,57,57,58,58,59,59,59
60-64	7	60,61,61,61,62,63,64
65-69	3	66,68,69
70-74	0	



Cumulative Frequency :

A type of frequency distribution that shows how many observations are above or below the lower boundaries of the classes.

Class	Frequency	Relative Frequency	Cumulative Frequency	Cumulative Relative Frequency
35-39	0	0	0	0
40-44	2	0.045454545	2	0.04545455
45-49	7	0.159090909	9	0.20454545
50-54	13	0.295454545	22	0.5
55-59	12	0.272727273	34	0.77272727
60-64	7	0.159090909	41	0.93181818
65-69	3	0.068181818	44	1
70-74	0	0	44	1
Total	44	1		



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

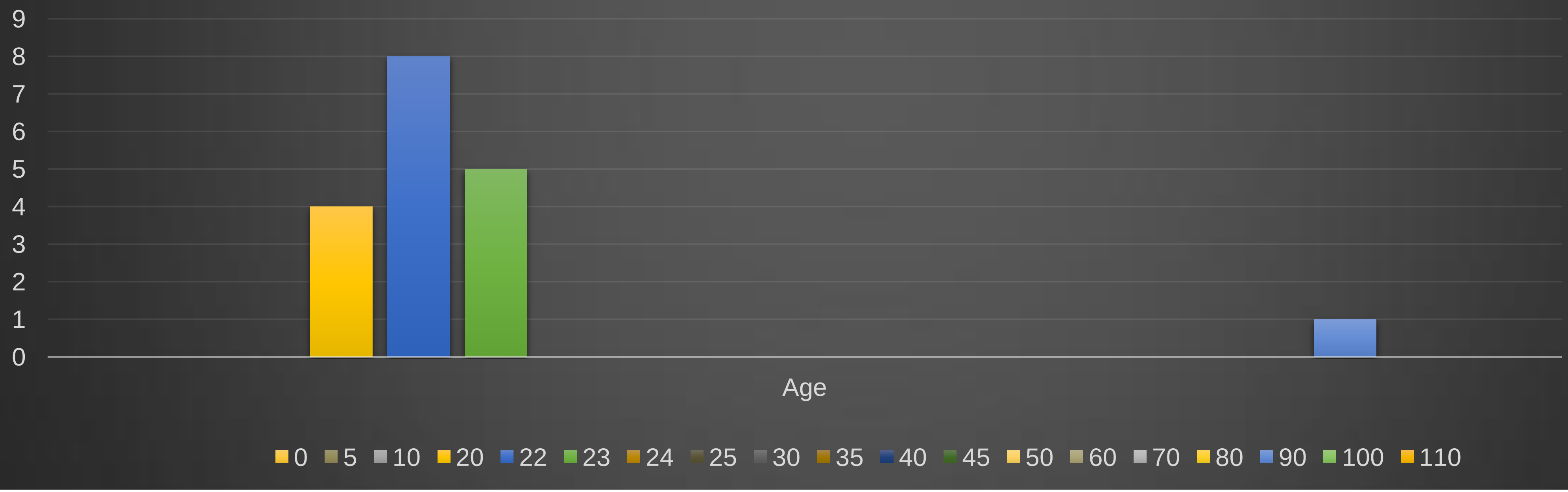
$$Mean, \mu = \frac{\Sigma x}{n} =$$

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

The Central Tendencies:

Power Workout Class Composition

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



Disadvantage of Mean:

- Finding mean is not a good approach as the 'Mean is often affected by Outliers' or in simple words if there are some observations larger or smaller than majority of the other observations then the mean tends to deviate towards these values.

- To generalize it if the distribution of datasets is skewed(troubled by outliers), we do not choose mean. Here we will have to go for Median.

The Central Tendencies – Median the mid-point:

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

- Data has outlier

How to find the median in three steps:

1. Line your numbers up in order, from smallest to largest.
2. If you have an odd number of values, the median is the one in the middle. If you have n numbers, the middle number is at position (n + 1) / 2.
3. If you have an even number of values, get the median by adding the two middle ones together and dividing by 2.

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

The Central Tendencies:

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

Mean

Age	F	Age * F
13	4	52
15	6	90
17	5	85
18	4	72
19	4	76
25	3	75
90	1	90

Sum(F) = 27

Sum(Age*F) **= 540**

Mean **= Sum(Age*F) / Sum(F)**

Mean **= 20**

Median

Age	F	cf
13	4	4
15	6	10
17	5	15
18	4	19
19	4	23
25	3	26
90	1	27

Median = **(N+1) / 2**

Median **= (27+1)/2 = 14th term**

Median = 17

Mode

Most Number of repeated value is Mode. Here 15 is repeated 6 times in the table

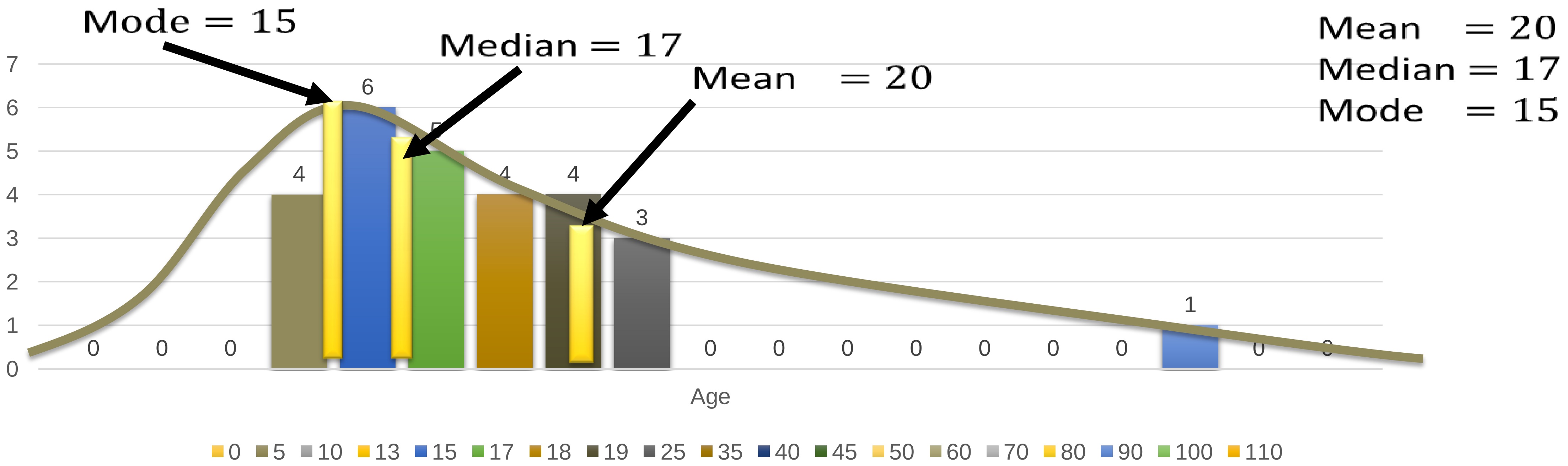
Then,

Mode = 15

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



Mean = 20
Median = 17
Mode = 15

The Central Tendencies:

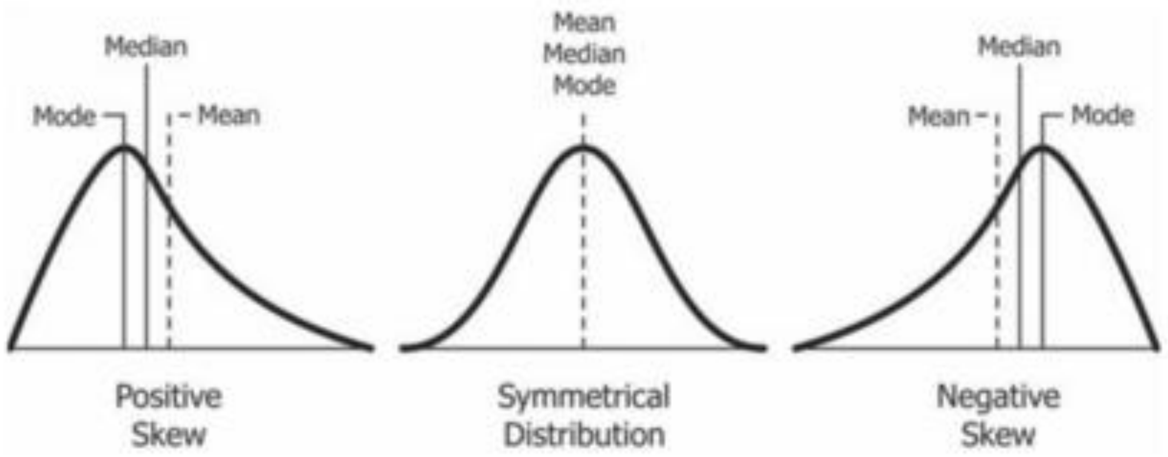
The management of Good Heart Inc. wants to give all its employees a raise. They are unable to decide if they should give a straight Rs. 2000 to everyone or to increase salaries by 10% across the board. The mean salary is Rs. 50,000, the median is Rs. 20,000 and the mode is Rs. 10,000.
How do these central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

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:

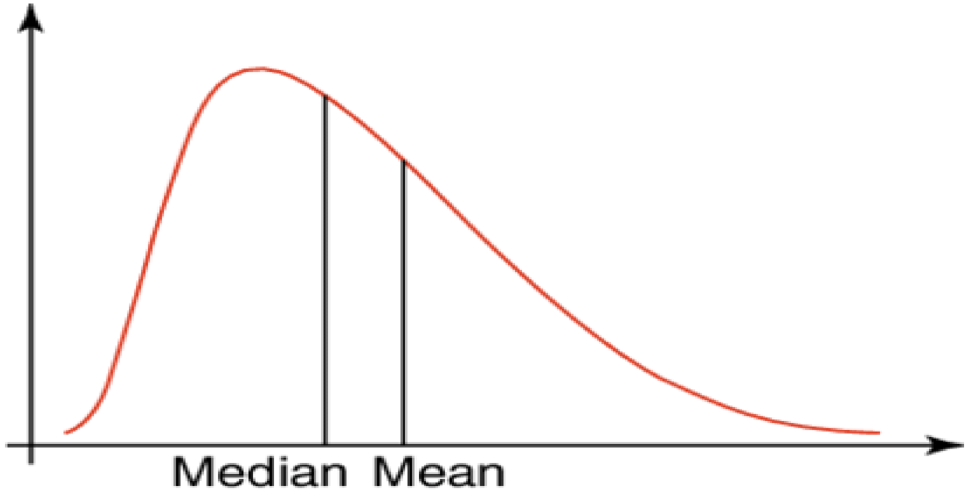
- Skewness tells us a lot about where the data is situated.



- 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:

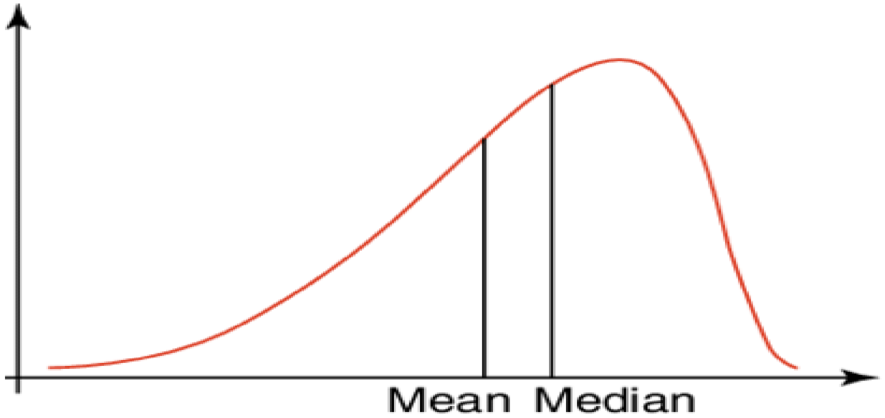
- A positively skewed distribution means that the extreme data results are larger. This skews the data in that it brings the mean (average) up. The mean will be larger than the median in a Positively skewed distribution.



Positive skew

Negative Skewness:

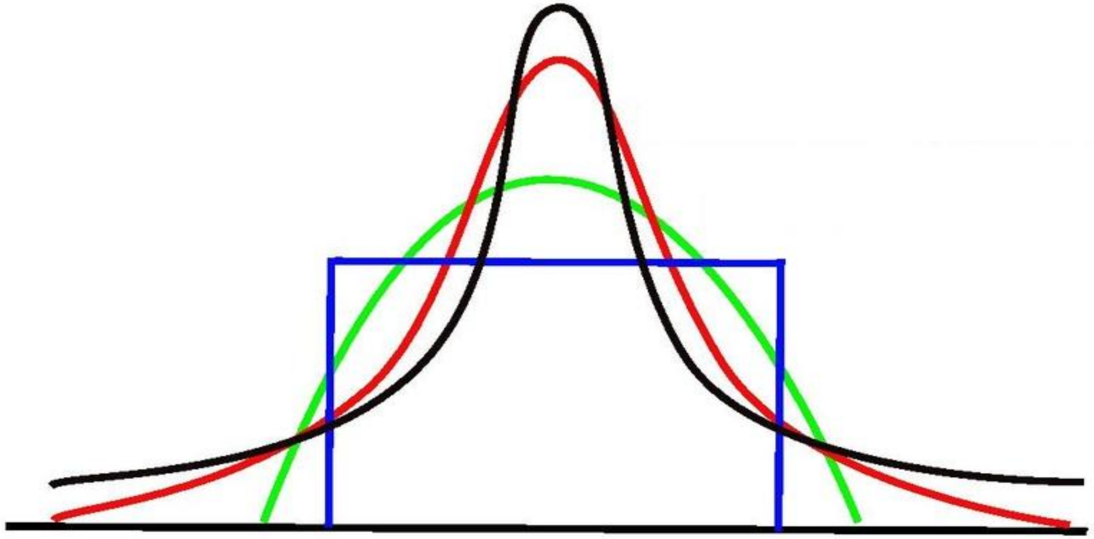
- A negatively skewed distribution means the opposite: that the extreme data results are smaller. This means that the mean is brought down, and the median is larger than the mean in a negatively skewed distribution.



Negative skew

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.



QUIZ on Central Tendency

Q1. If the data is Categorical, which measure of central tendency is most appropriate to use?

Q2. If the data is Numerical which is the suitable Measure to use?

Quiz:

Q1. If the data is Categorical, which measure of central tendency is most appropriate to use?

Mode

Q2. If the data is Numerical which is the suitable Measure to use?

Mean and Median

Quiz:

Q3. Does Mean effect with the Extreme Value?

Q4. Which type of graph displays data in consecutive and equivalent intervals?

Q5. If it is right skewed what are the positions of central tendencies?

Quiz:

Q3. Does Mean effect with the Extreme Value?

Yes

Q4. Which type of graph displays data in consecutive and equivalent intervals?

Histogram

Q5. If it is right skewed what are the positions of central tendencies?

Mode < Median < Mean

Measuring Variability and Spread

Why Measure Of Spread?

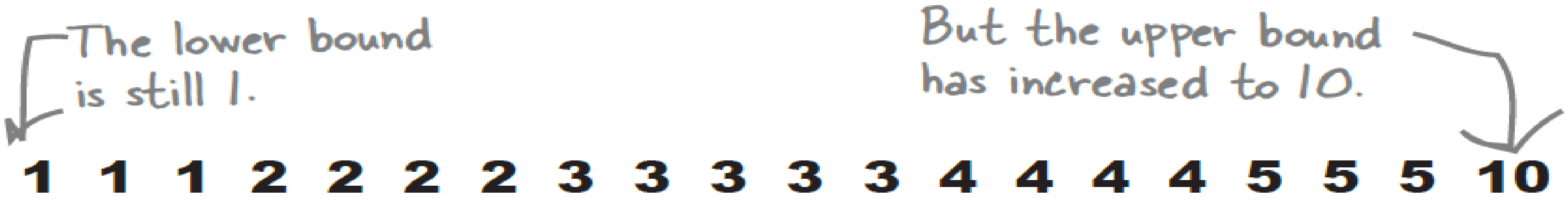
- It is quite often, the average only gives part of the picture.
- Averages give us a way of determining where the centre of a set of data is, but they don’t tell us how the data varies.

Different Measure of Spread:

- Range
- IQR
- Variance
- Standard Deviation

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.

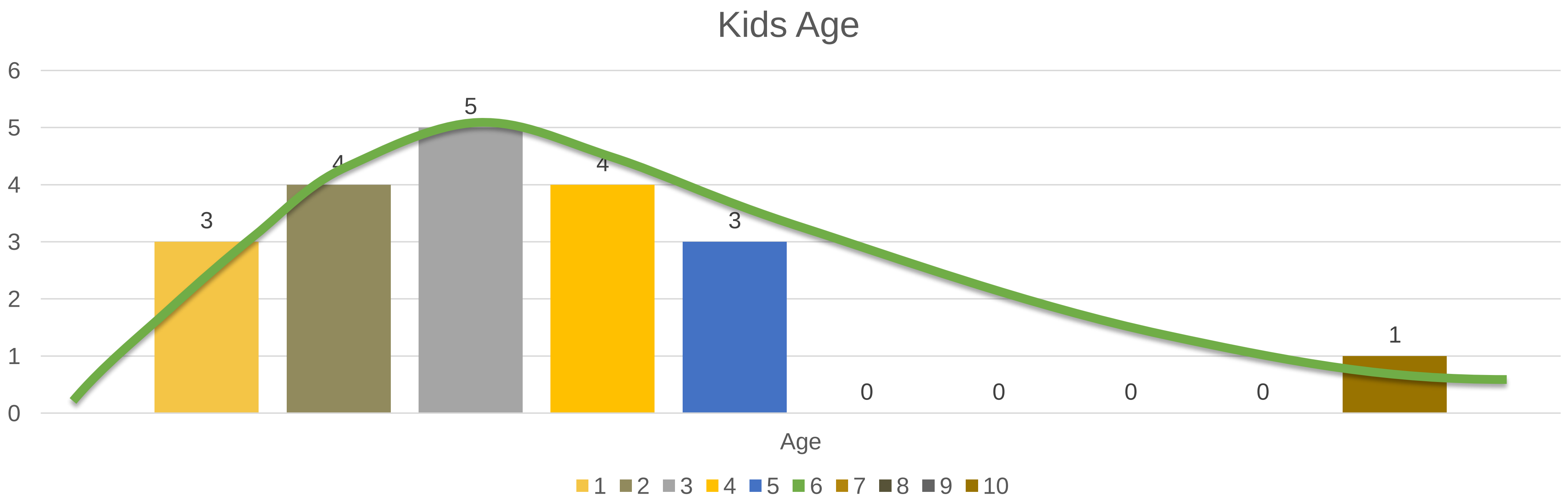

The lower bound is still 1.
But the upper bound has increased to 10.
1 1 1 2 2 2 2 3 3 3 3 3 4 4 4 4 5 5 5 10

Range = upper bound - lower bound

= 10 -1

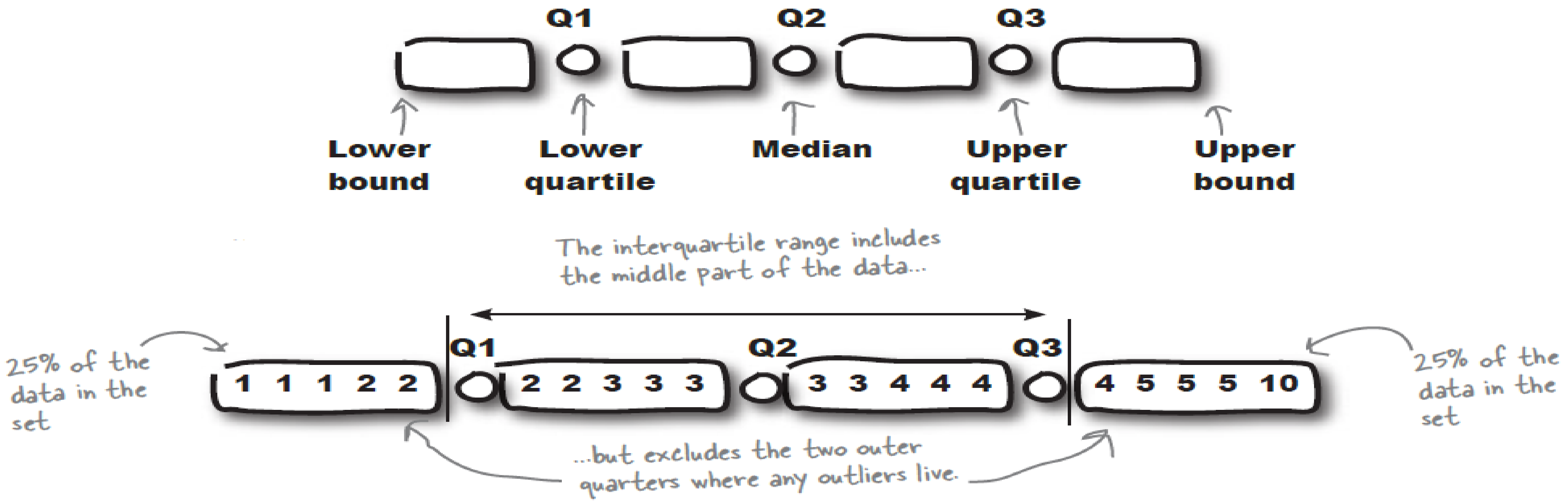
= 9

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.

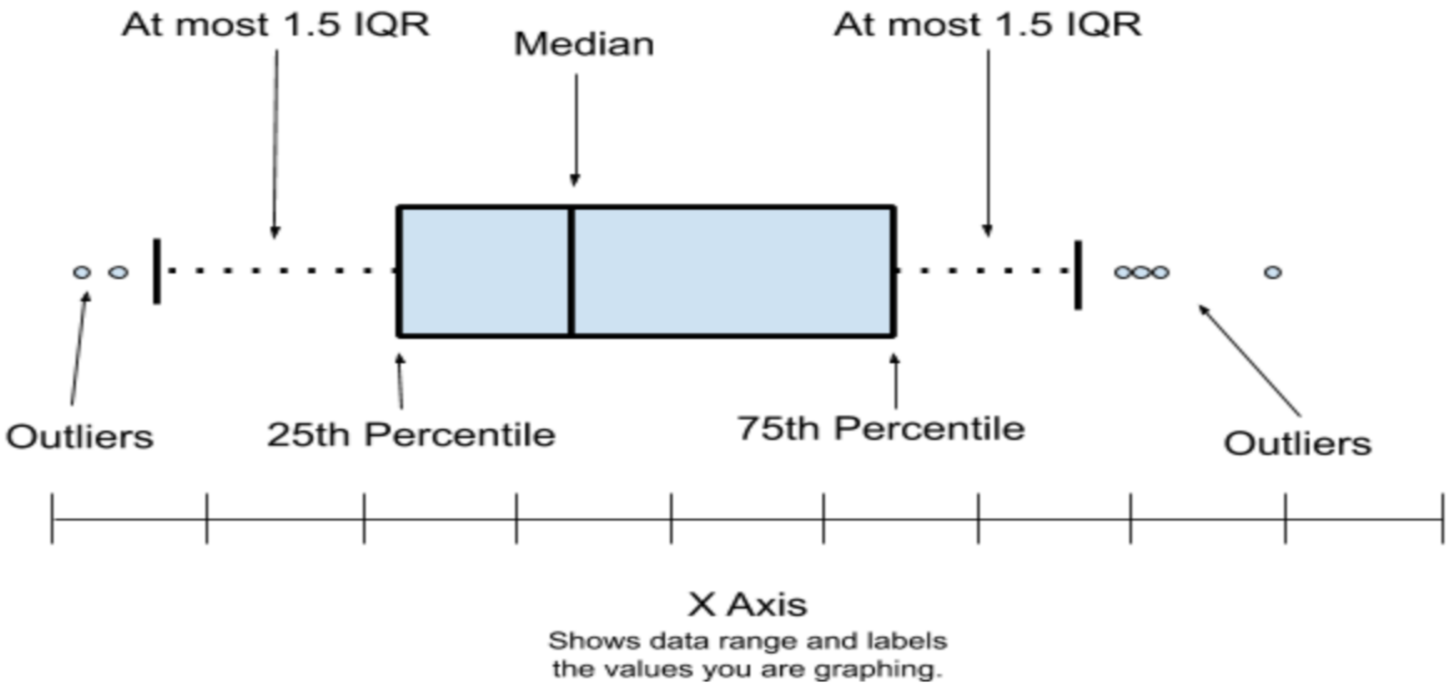

Q1
Q2
Q3
Lower bound
Lower quartile
Median
Upper quartile
Upper bound
The interquartile range includes the middle part of the data...
25% of the data in the set
Q1
Q2
Q3
1 1 1 2 2
2 2 3 3 3
3 3 4 4 4
4 5 5 5 10
25% of the data in the set
...but excludes the two outer quarters where any outliers live.

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.

Box plot Visual Representation:

- Lets Understand the IQR and Identifying the Outliers



Advantage of IQR:

- The main advantage of the IQR is that it is not affected by outliers because it doesn’t take into account observations below Q1 or above Q3.

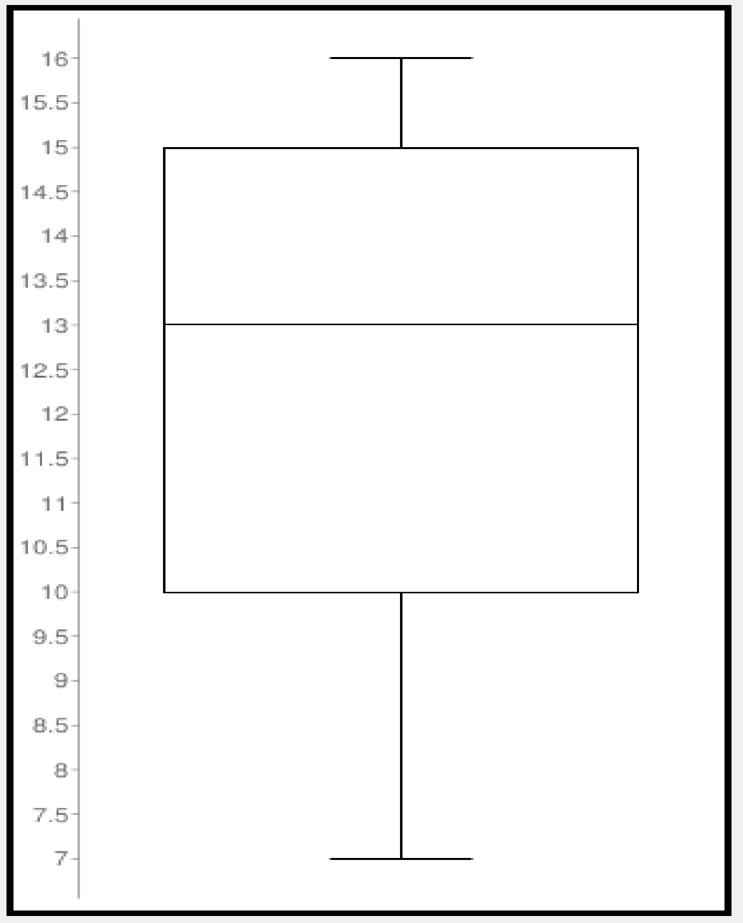
- It might still be useful to look for possible outliers in your study.

- As a rule of thumb, observations can be qualified as outliers when they lie more than 1.5 IQR below the first quartile or 1.5 IQR above the third quartile.

Outliers = Q1 – 1.5* IQR and
= Q3 + 1.5*IQR

Interpreting Box-whisker plot:

- Which of the following statements are true?
  - All of the students are less than 17years 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

Variance – Problem:

x	$(x - \mu)$	$(x - \mu)^2$
4	4 - 6 = - 2	4
6	6 - 6 = 0	0
7	7 - 6 = 1	1
3	3 - 6 = - 3	9
10	10 - 6 = 4	16

$\sum x$ = 30

$\sum(x - \mu)^2$ = 30

$$\mu = \frac{\sum x}{n} = 30 / 5 = 6$$

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

Variance = 30 / 5 = 6

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

Standard Deviation - Problem

x	$(x - \mu)$	$(x - \mu)^2$
4	4 - 6 = - 2	4
6	6 - 6 = 0	0
7	7 - 6 = 1	1
3	3 - 6 = - 3	9
10	10 - 6 = 4	16

$\sum x$ = 30

$\Sigma(x - \mu)^2$ = 30

$$\mu = \frac{\sum x}{n} = 30 / 5 = 6$$

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

$$\sigma = \sqrt{30/5} = 2.43$$

Action Check:

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

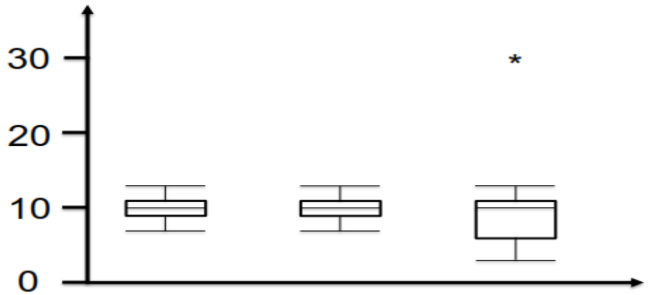
Points Scored per game	7	8	9	10	11	12	13
Frequeny,f	1	1	2	2	2	1	1

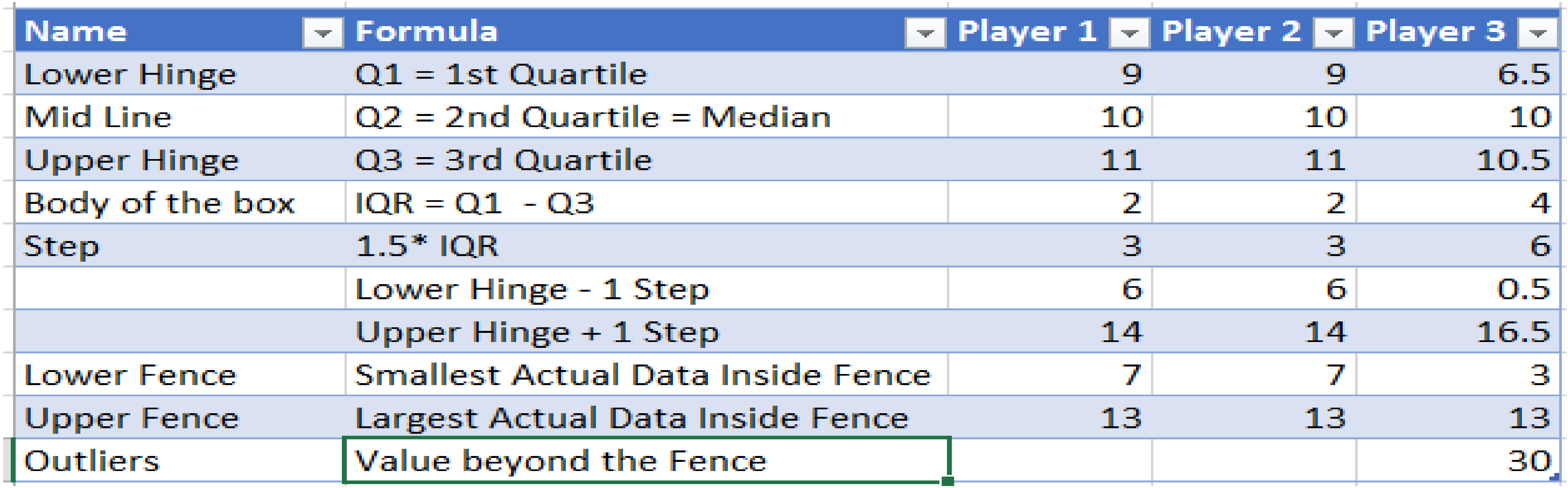
Points Scored per game	7	9	10	11	13
Frequency, f	1	2	4	2	1

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

Measuring Variability and Spread:

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

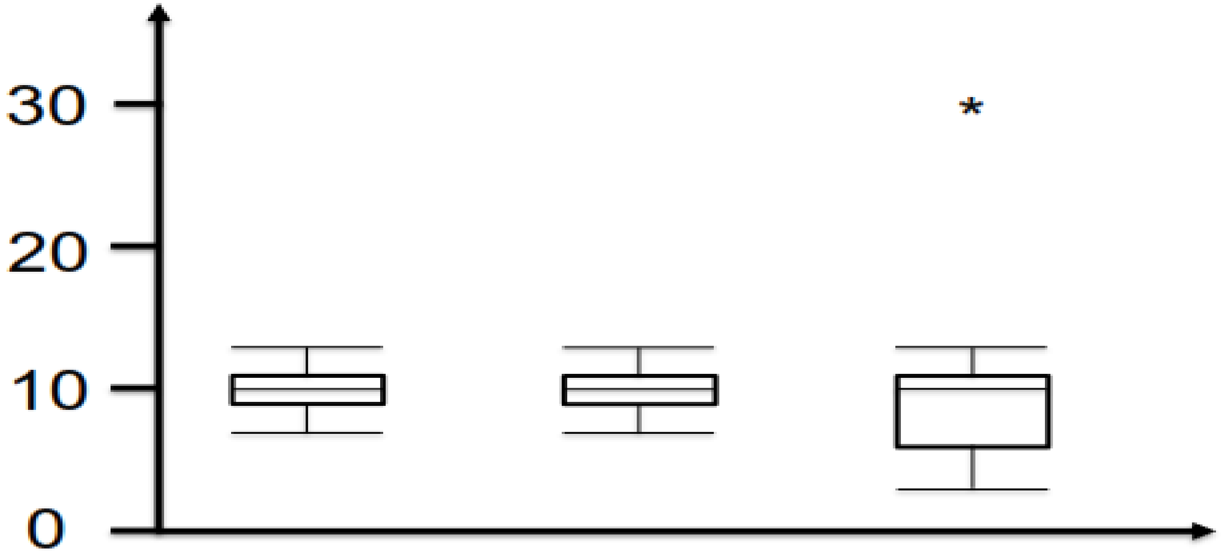




Name	Formula	Player 1	Player 2	Player 3
Lower Hinge	Q1 = 1st Quartile	9	9	6.5
Mid Line	Q2 = 2nd Quartile = Median	10	10	10
Upper Hinge	Q3 = 3rd Quartile	11	11	10.5
Body of the box	IQR = Q1 - Q3	2	2	4
Step	1.5* IQR	3	3	6
	Lower Hinge - 1 Step	6	6	0.5
	Upper Hinge + 1 Step	14	14	16.5
Lower Fence	Smallest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

Measuring Variability and Spread

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



Attention Check:

Player1:

Mean = 10

SD = 0.547

Player2:

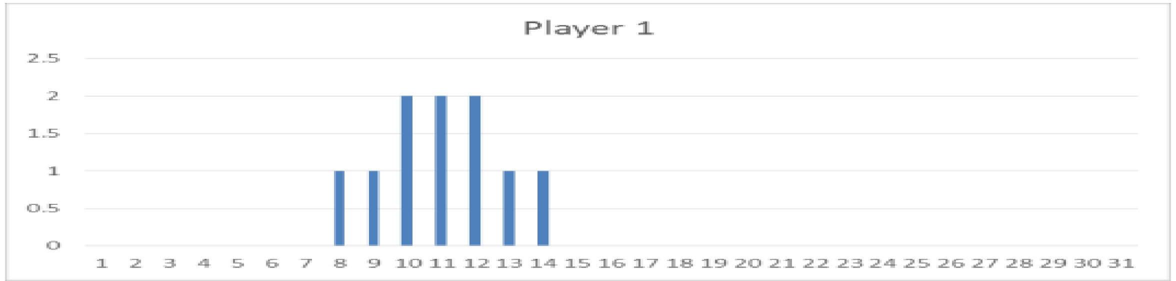
Mean = 10

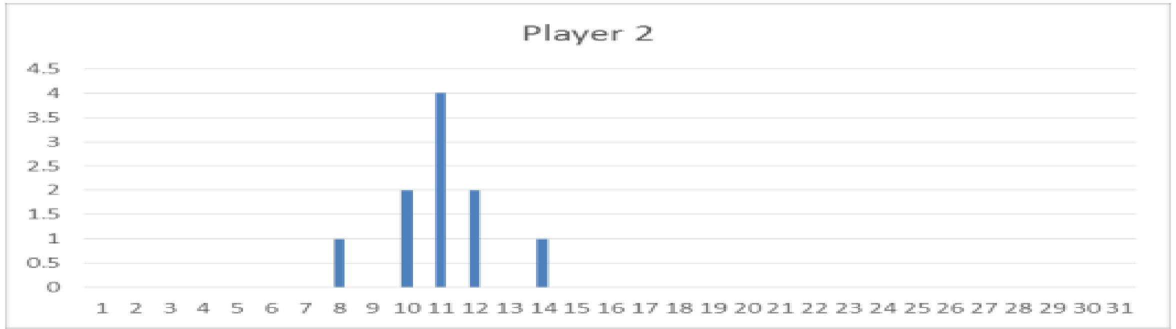
SD = 0.469

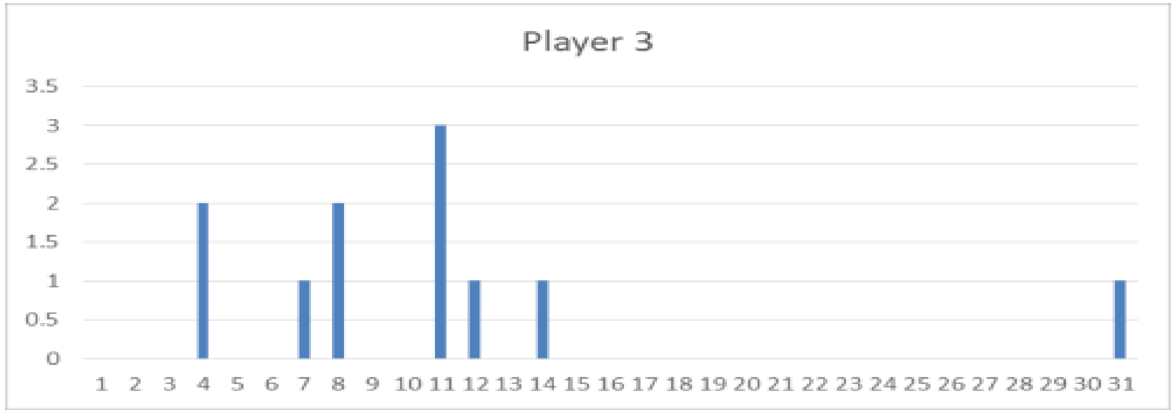
Player3:

Mean = 10

SD = 2.353







Out of 3 Players Who is Reliable and Consistent

?

Player 3 is the least reliable.

Player 2 is consistent.

Good Heart Inc. :

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

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

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

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

Scatter Plot and Correlation.

- Displaying Relationships: Scatter plots
- Interpreting Scatter plots.
- Measuring Linear Association: Correlation.
- Facts About Correlation

Scatter Plots - Bivariate data

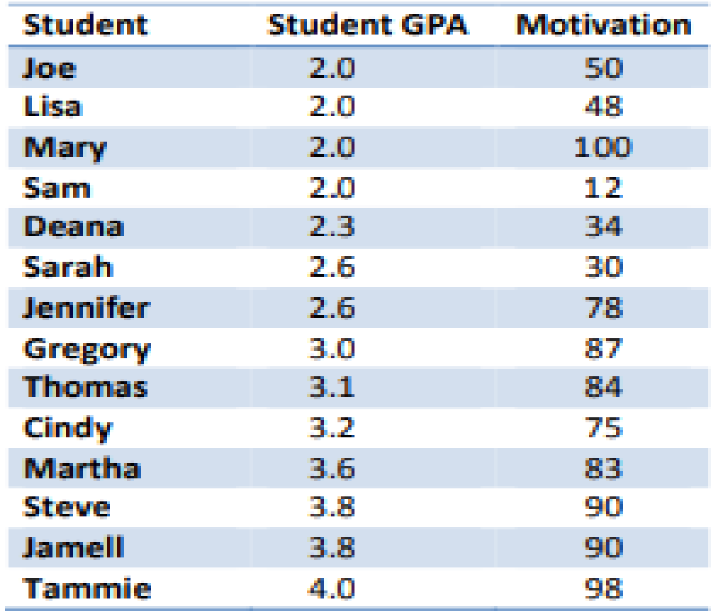
- A scatter plot shows the relationship between two quantitative variables measured for the same individuals.

- The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis.

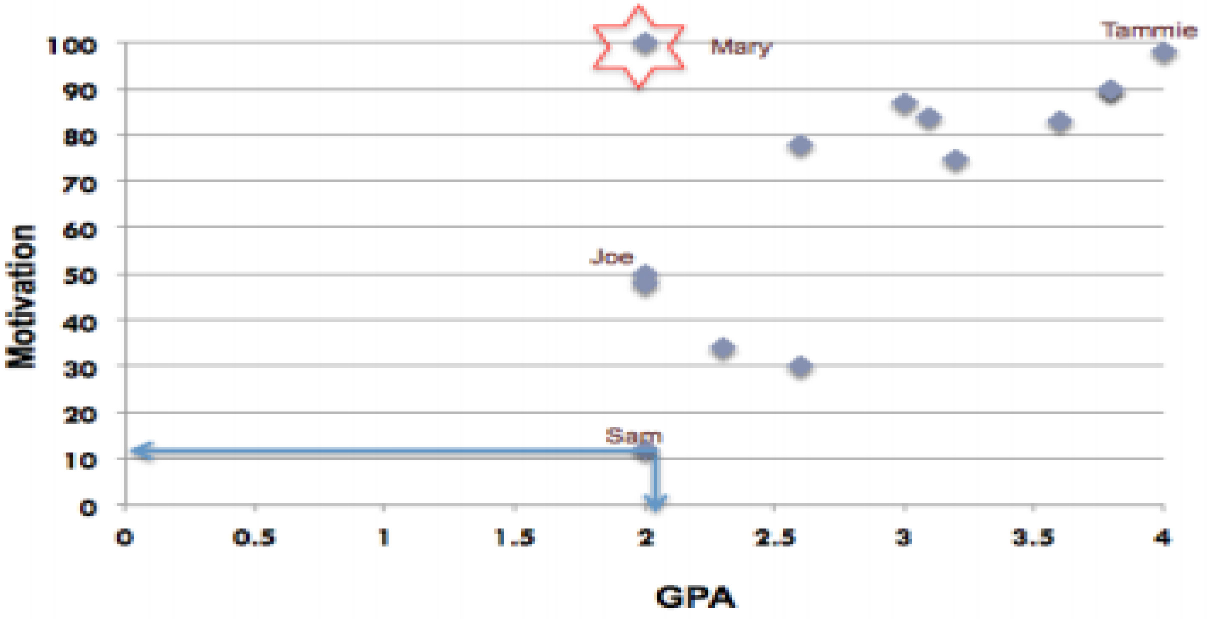
- Each individual in the data appears as a point on the graph.

Scatter plot Example

- What is the relationship between students’ achievement motivation and GPA?
- the relationship between students’ achievement motivation and their GPA is being investigated.



Student	Student GPA	Motivation
Joe	2.0	50
Lisa	2.0	48
Mary	2.0	100
Sam	2.0	12
Deana	2.3	34
Sarah	2.6	30
Jennifer	2.6	78
Gregory	3.0	87
Thomas	3.1	84
Cindy	3.2	75
Martha	3.6	83
Steve	3.8	90
Jamell	3.8	90
Tammie	4.0	98



Interpreting Scatter plots

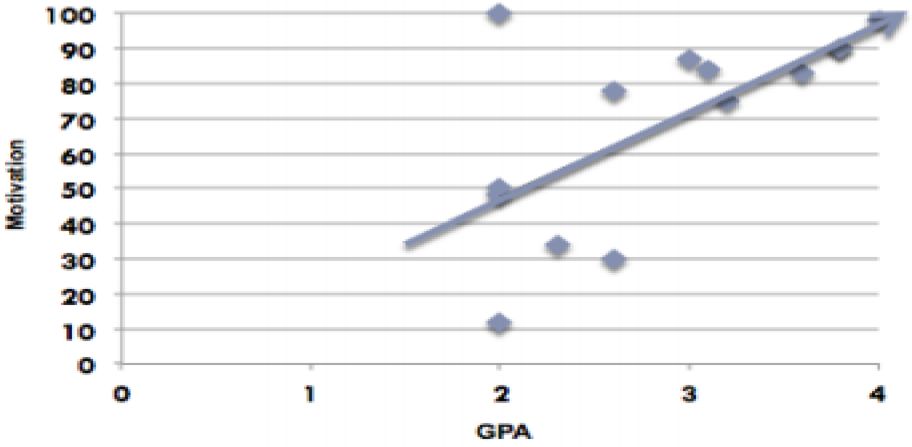
As in any graph of data, look for the overall pattern and for striking departures from that pattern.

- The overall pattern of a scatter plot can be described by the direction, form, and strength of the relationship.

- An important kind of departure is an outlier, an individual value that falls outside the overall pattern of the relationship

Interpreting Scatter plots: Direction

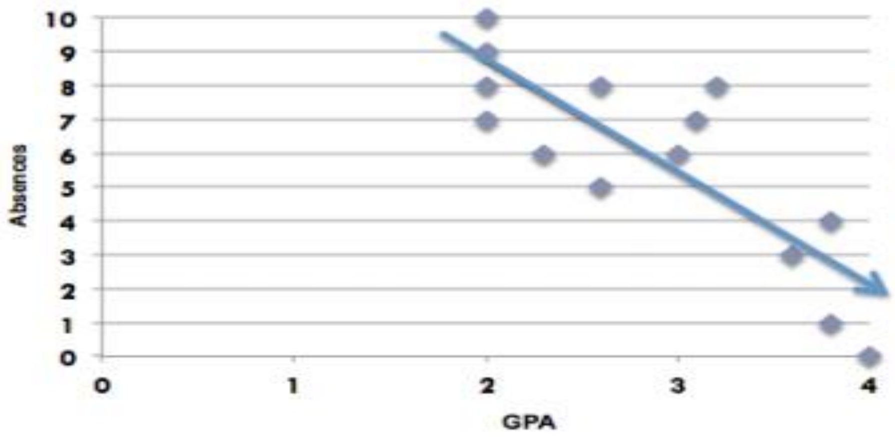
- One important component to a scatter plot is the direction of the relationship between the two variables.
- Two variables have a positive association when above-average values of one tend to accompany above-average values of the other, and when below-average values also tend to occur together.



This example compares students' achievement motivation and their GPA. These two variables have a **positive association** because as GPA increases, so does motivation.

Negative Association

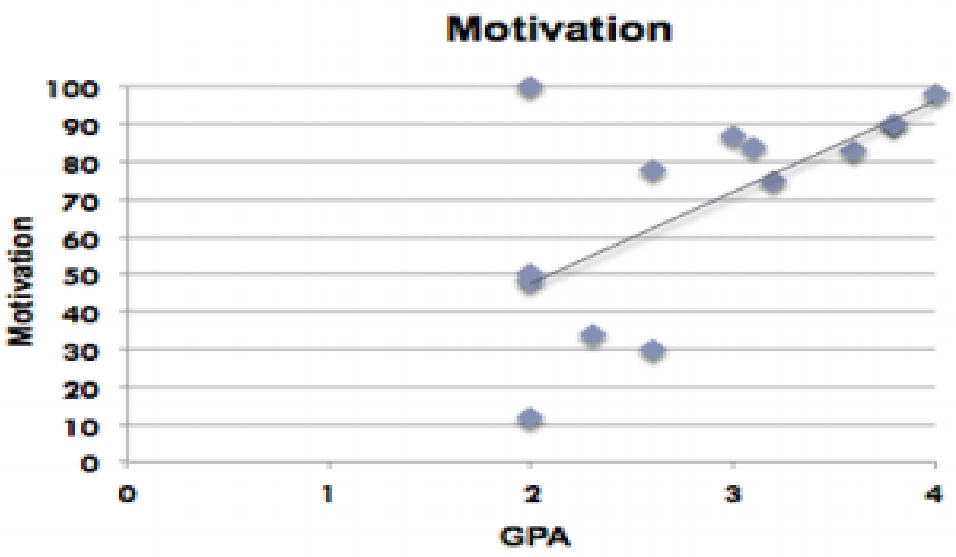
- Two variables have a negative association when above-average values of one tend to accompany below-average values of the other.



This example compares students' GPA and their number of absences. These two variables have a **negative association** because, in general, as a student's number of absences decreases, their GPA increases.

Interpreting Scatter plots: Form

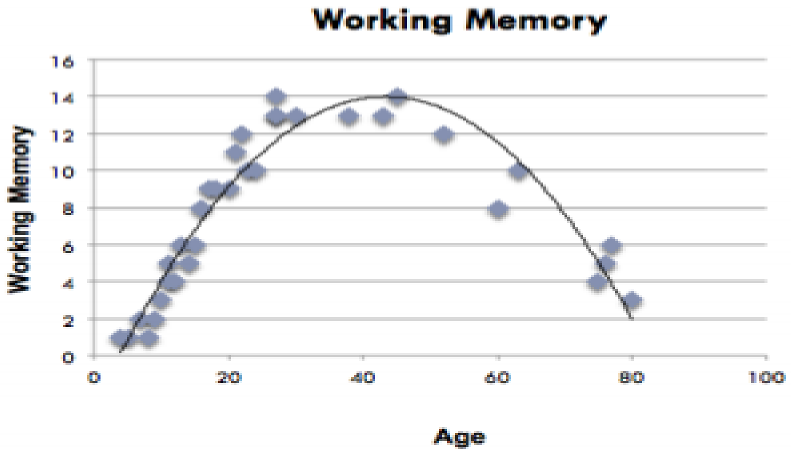
- Another important component to a scatter plot is the form of the relationship between the two variables.
- Linear Relationship



This example illustrates a linear relationship. This means that the points on the scatterplot closely resemble a straight line. A relationship is linear if one variable increases by approximately the same rate as the other variables changes by one unit.

Curvilinear Relationship

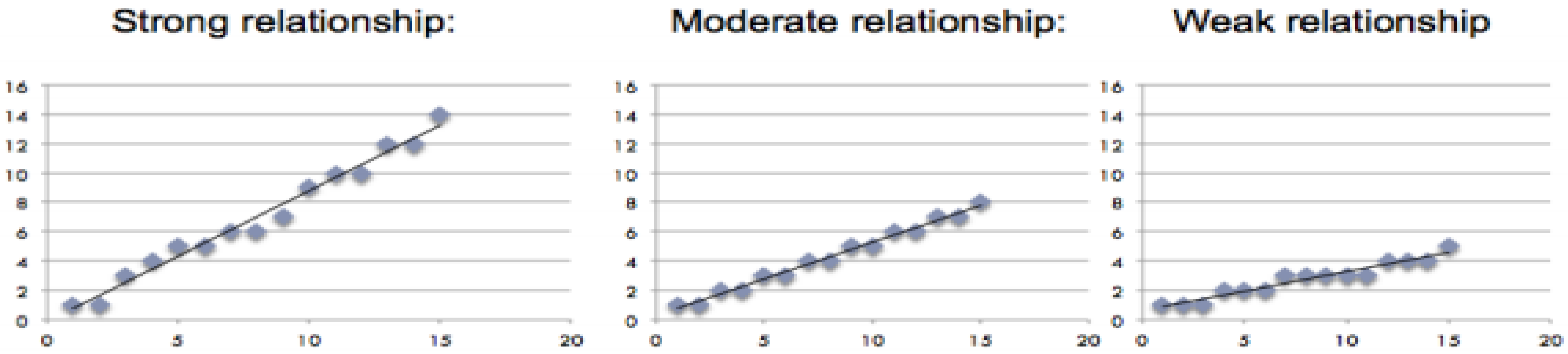
Curvilinear relationship:



This example illustrates a relationship that has the form of a curve, rather than a straight line. This is due to the fact that one variable does not increase at a constant rate and may even start decreasing after a certain point. This example describes a curvilinear relationship between the variable “age” and the variable “working memory.” In this example, working memory increases throughout childhood, remains steady in adulthood, and begins decreasing around age 50.

Interpreting Scatter plots: Strength

- Another important component to a scatter plot is the strength of the relationship between the two variables.
- The slope provides information on the strength of the relationship.



- The strongest linear relationship occurs when the slope is 1. This means that when one variable increases by one, the other variable also increases by the same amount. This line is at a 45 degree angle.

- The strength of the relationship between two variables is a crucial piece of information. Relying on the interpretation of a scatter plot is too subjective. More precise evidence is needed, and this evidence is obtained by computing a coefficient that measures the strength of the relationship under investigation.

Measuring Linear Association

- A scatter plot displays the strength, direction, and form of the relationship between two quantitative variables.

- A correlation coefficient measures the strength of that relationship.

- Calculating a Pearson correlation coefficient requires the assumption that the relationship between the two variables is linear.

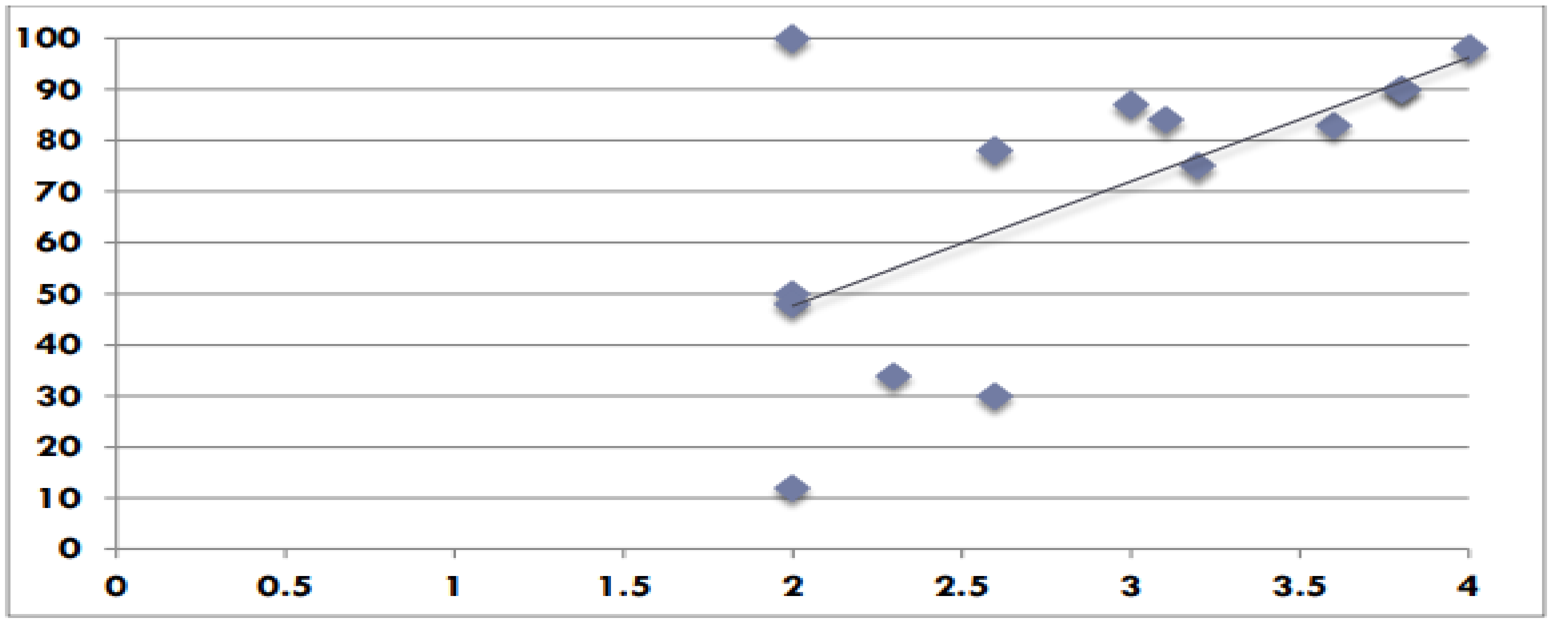
- There is a rule of thumb for interpreting the strength of a relationship based on its r value (use the absolute value of the r value to make all values

Absolute Value of r	Strength of Relationship
$r < 0.3$	None or very weak
$0.3 < r <0.5$	Weak
$0.5 < r < 0.7$	Moderate
$r > 0.7$	Strong

Example – GDP VS Achievement Motivation

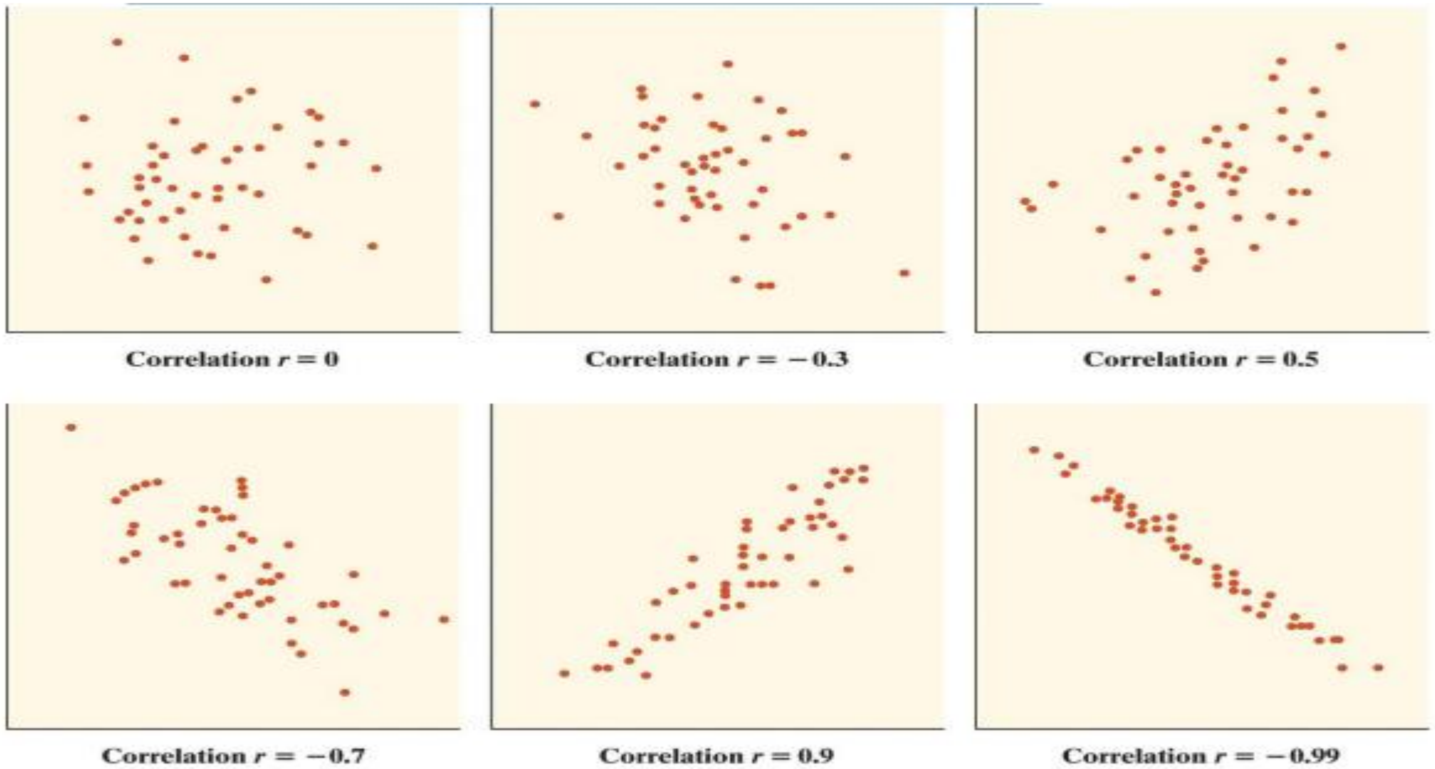
- There is a Moderate, positive, linear Relationship between GPA and achievement Motivation.

r = 0.62



Correlation:

- The images below illustrate what the relationships might look like at different degrees of strength (for different values of r).



Correlation $r = 0$

Correlation $r = -0.3$

Correlation $r = 0.5$

Correlation $r = -0.7$

Correlation $r = 0.9$

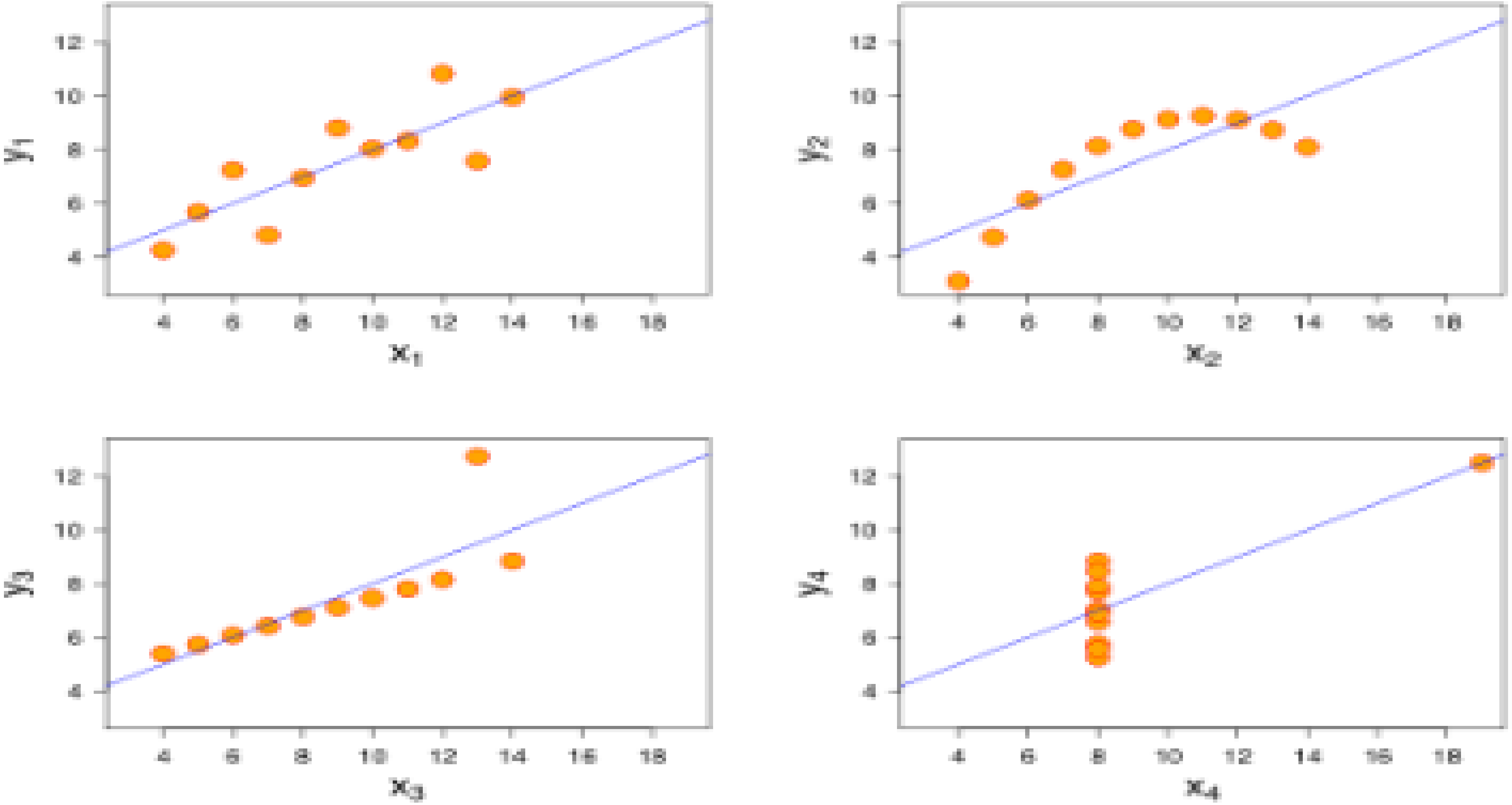
Correlation $r = -0.99$

- For a correlation coefficient of zero, the points have no direction, the shape is almost round, and a line does not fit to the points on the graph.
- As the correlation coefficient increases, the observations group closer together in a linear shape.
- The line is difficult to detect when the relationship is weak (e.g., r = -0.3), but becomes more clear as relationships become stronger (e.g., r = -0.99)

Facts About Correlation

1) The order of variables in a correlation is not important.
2) Correlations provide evidence of association, not causation.
3) r has no units and does not change when the units of measure of x, y, or both are changed.
4) Positive r values indicate positive association between the variables, and negative r values indicate negative associations.
5) The correlation **r is always a number between -1 and 1.**

Pattern in the data



Quiz:

Q1. Formula for calculating IQR

Q2. A value that is much higher or much lower than the other values in a set of data.

Q3. Which Graph is used to show outliers?

Quiz:

Q1. Formula for calculating IQR

IQR = Q3 – Q1

Q2. A value that is much higher or much lower than the other values in a set of data.

Outlier

Q3. Which Graph is used to show outliers?

Box Plot

Quiz:

Q4. The deviation between the observation to its mean is:

Q5. The direction of a correlation can be positive or negative.

Quiz:

Q4. The deviation between the observation to its mean is:

Standard Deviation

Q5. The direction of a correlation can be positive or negative.

True

Quiz:

Q6. The correlation coefficient shows..

Q7. If r=-0.89 is what type of correlation

Quiz:

Q6. The correlation coefficient shows..

The Strength and Direction

Q7. If r=-0.89 is what type of correlation

Strong Negative Correlation

Reference

- https://www.mathsisfun.com/data/index.html
- **Head First: Statistics**
- **KhanAcademy**
- Scatterplots and Correlation **by Dian mindrila Ph.D and Phoebe balentyne, M.Ed based on chapeter 4 of The basic practice of Statistics. (6th ed).**

THAT’S ALL FOLKS