DS Part I - Descriptive Statistics

Sasken Training, Adyar , Chennai – 600 020

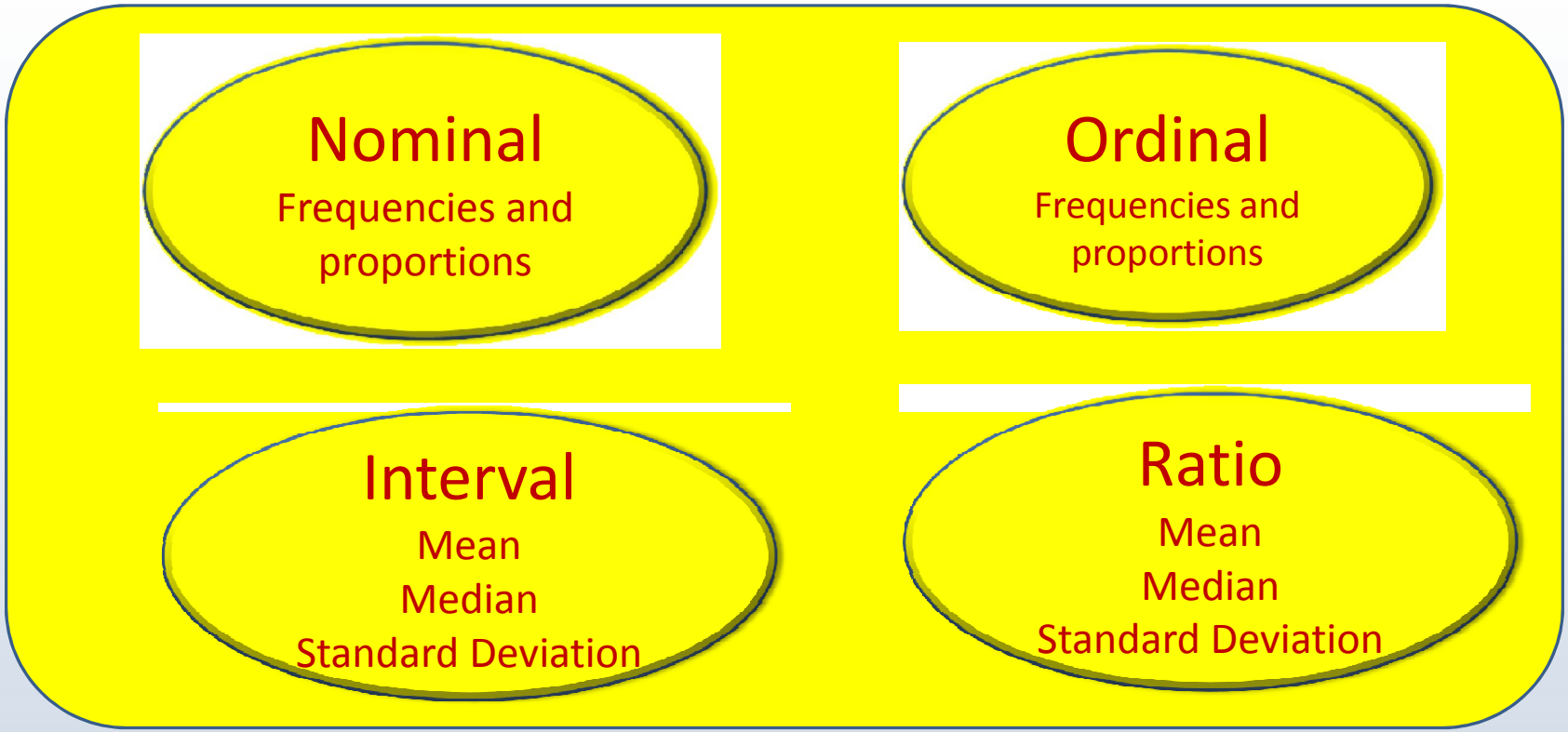
- What is Statistics ?
- What is Descriptive Statistics ?
- Measures of Central Tendency
- Measures of Spread
- Shape Measures

- Introduction to Statistics
- Descriptive Statistics
  - Measuring the Central Tendency
    - Mean, Weighted Mean, Geometric Mean, Median, Mode
  - Measure of Spread
    - Percentiles, Quartiles and Deciles
    - Range, Interquartile, variation
    - Standard Deviation, Coefficient of Variation
  - Shape Measures
    - Skewness and Kurtosis

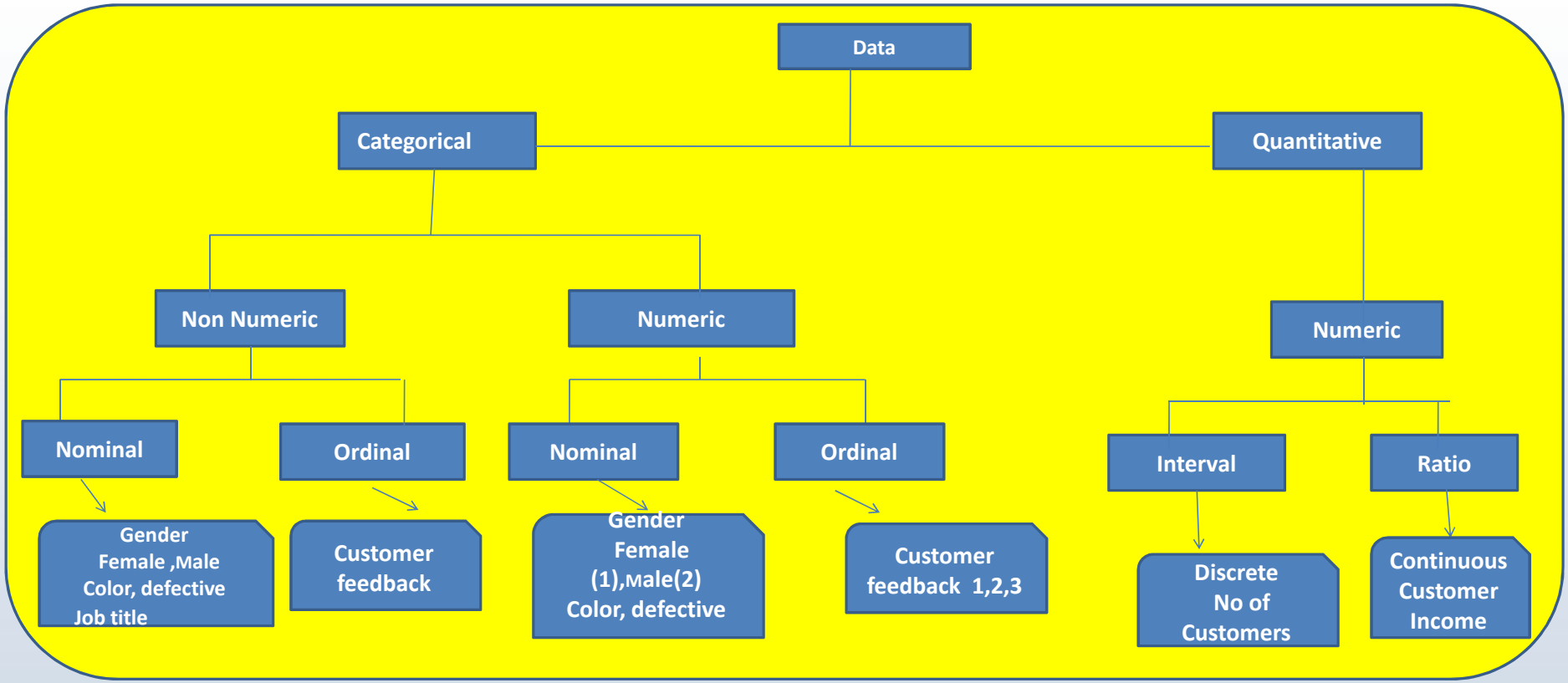
Statistics – What is it ?

It's a science deals with Collection, Classification, Analysis , and Interpretation of numerical facts or data AND possibly use it to predict the future events

Levels of Data measurement - Ordinal, nominal, scale



Levels of Data measurement - Ordinal, nominal, scale



Levels of measurement – Example

Variable	Values	Level of measurement	Discrete or Continuous
Gender	Male (1), Female(2)	Nominal	
Age	23,24,26 etc.,		
Hours spent last week	5.30 hours	Ratio	
Performance rating	1, 2 ,3,4		

Quiz

- What type of data is job title ?
  - Numerical
  - Categorical
  - Text
  - Time series
  - Interval / ratio

- What type of data is email contents ?
  - Numerical
  - Categorical
  - Text
  - Time series
  - Interval / ratio

- Number of emails sent by a person ?
  - Numerical
  - Categorical
  - Text
  - Time series
  - Interval / ratio

Quiz

- What type of data is to/from fields of emails ?
  - Numerical
  - Categorical
  - Text
  - Time series
  - Interval / ratio

Types of Variables

Data set	Variables	Inferences
Univariate	One	Histograms, Descriptive statistics
Bivariate	Two	Scatter Plots, Correlations, simple regression
Multivariate	> 2	Multiple regression, Modeling

Claim Num	Total claim Amount
AQI01	25000
AQI08	18000
BOI09	12000
QCX81	34524

Case	Name	Age	Income	Job	Gender
1	Raj	45	1,25,000	Sr. Mgr	Male
2	Ashok	50	2,56,500	A.Dir	Male
3	Mala	26	12,000	Sw eng	Female
4	Bala	32	8,000	Clerk	Male
5	Sundar	38	2,42,000	A.Dir	Male
6	John	59	81,200	Proj Lead	male

Case	Exp	Salary
1	5	40000
2	8	75000
3	8	70000
4	4	41000

Frequency Distribution

Presenting summary of data in the form of class intervals and frequencies

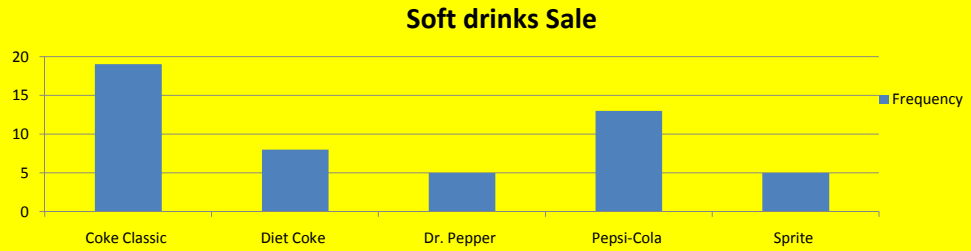
The below is a sample of 50 soft drink purchases

Brand Purchased

Coke Classic	Dr. Pepper	Coke Classic	Diet Coke	Pepsi-Cola
Diet Coke	Sprite	Sprite	Pepsi-Cola	Pepsi-Cola
Pepsi-Cola	Coke Classic	Pepsi-Cola	Coke Classic	Coke Classic
Diet Coke	Diet Coke	Coke Classic	Coke Classic	Dr. Pepper
Coke Classic	Coke Classic	Coke Classic	Coke Classic	Pepsi-Cola
Coke Classic	Coke Classic	Coke Classic	Pepsi-Cola	Sprite
Dr. Pepper	Sprite	Pepsi-Cola	Dr. Pepper	
Diet Coke	Coke Classic	Coke Classic	Coke Classic	
Pepsi-Cola	Diet Coke	Sprite	Diet Coke	
Pepsi-Cola	Coke Classic	Dr. Pepper	Pepsi-Cola	
Coke Classic	Diet Coke	Pepsi-Cola	Pepsi-Cola	

Frequency Distribution

Soft Drink	Frequency	Percentage
Coke Classic	19	38%
Diet Coke	8	16%
Dr. Pepper	5	10%
Pepsi-Cola	13	26%
Sprite	5	10%



- We created a frequency table which was just counting the number of occurrences of each value appeared in the dataset
- This table shows the frequency count of each brand's Soft drink.

Relative Frequency Distribution

- Individual class frequency divided by the total frequency.
  - List how often a value occurs
  - Sum of total relative frequencies should be 1
  - Take each of the frequency value and divide it by no. of samples

Leaves Taken				
4	3	0	3	1
5	2	2	2	0
4	2	1	3	1

Leaves	Frequency	Relative Frequency	Percent frequency	Cum frequency
0	2	.13		.13
1	3	.2		.33
2	4	.27		.50
3	3	.2		
4	2	.13		
5	1	.07		
	15 (sample)	1.00	100	

Cumulative Frequency Distribution

- It’s the running total of frequencies through the classes of frequency distribution

Leaves Taken				
4	3	0	3	1
5	2	2	2	0
4	2	1	3	1

Leaves	Frequency	Relative frequency	Cumulative frequency
0			
1			
2			
3			
4			
5			
	15 (sample)	1.00	

18	19	32	17
19	21	65	19
22	18	28	65
24	18	22	54
28	29	18	29
32	30	17	44
65	21	44	32

18	
19	
21	
22	
24	
28	
29	
30	
32	
44	
65	

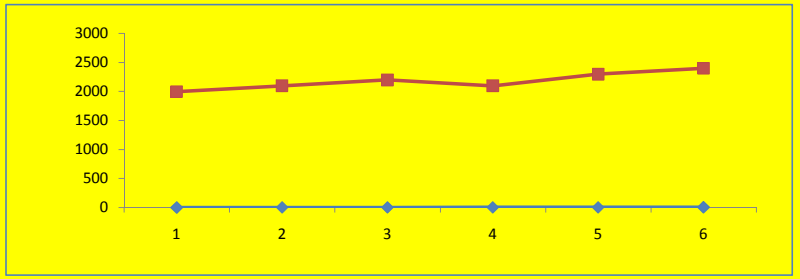
Obtain the frequency of customers who are between 25-30 ?

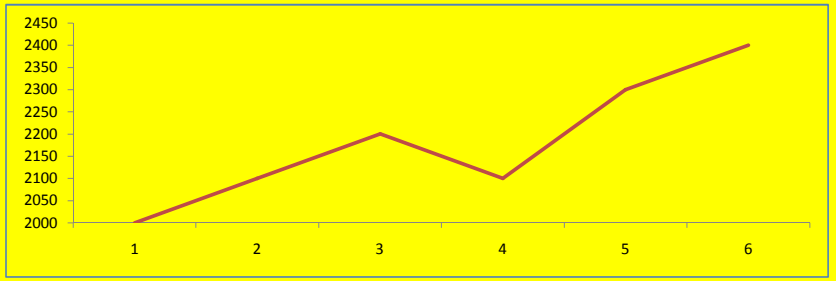
Visual Representation

Data Visualization – Line chart

- Line chart shows trend over time

Months	Revenue
1	2000
2	2100
3	2200
4	2100
5	2300
6	2400





Pie chart

- To compare proportions of different categories or groups

Walmart	6200
Tesco	4600
Costco	5100
Walgreens	2800

Bar charts

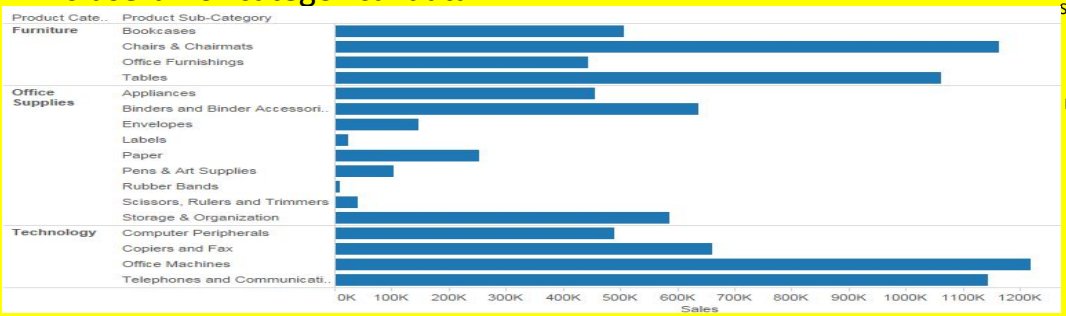
Bar charts allow you to compare relative sizes, but at a higher degree of precision

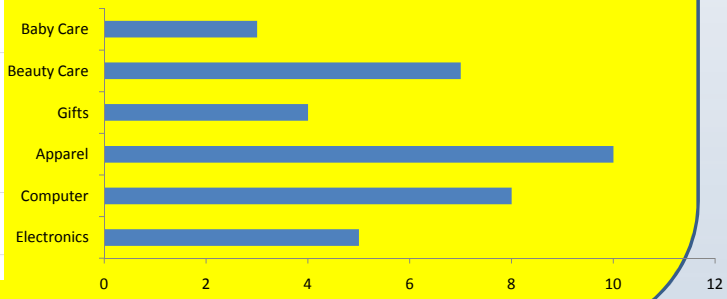
Vertical bar chart

is useful for numerical and category names are short

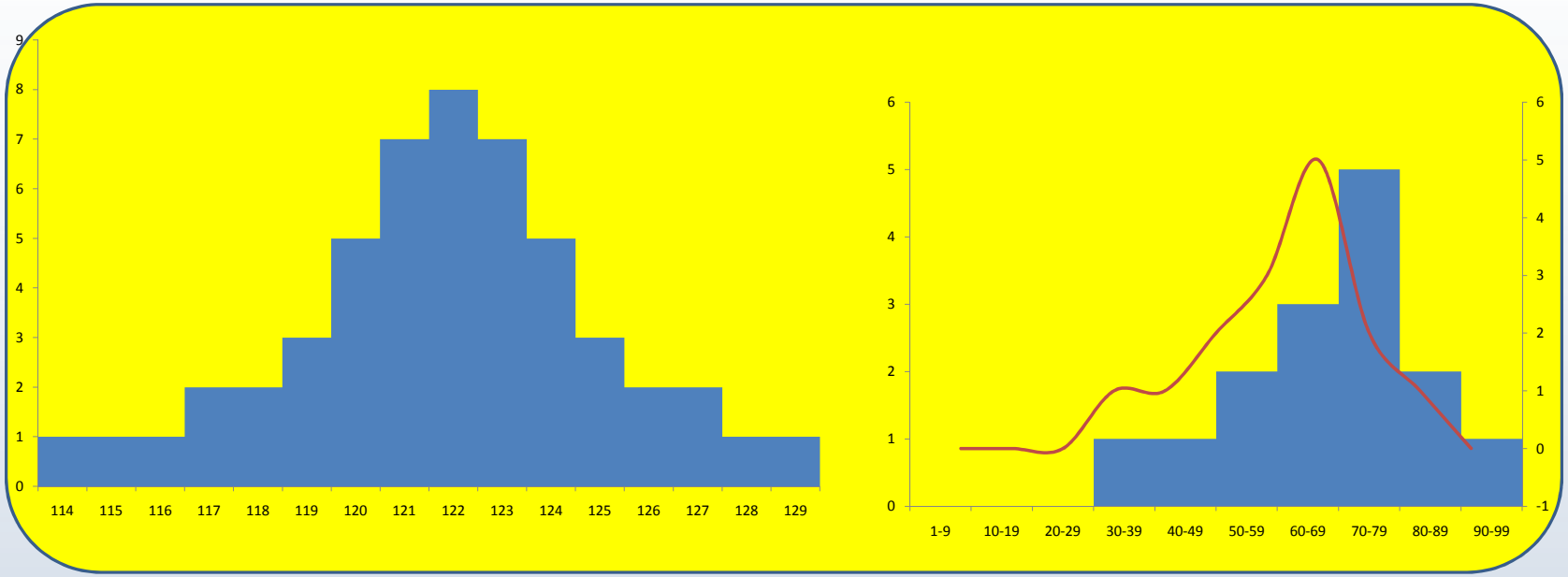
Horizontal bar chart

is useful for categorical data



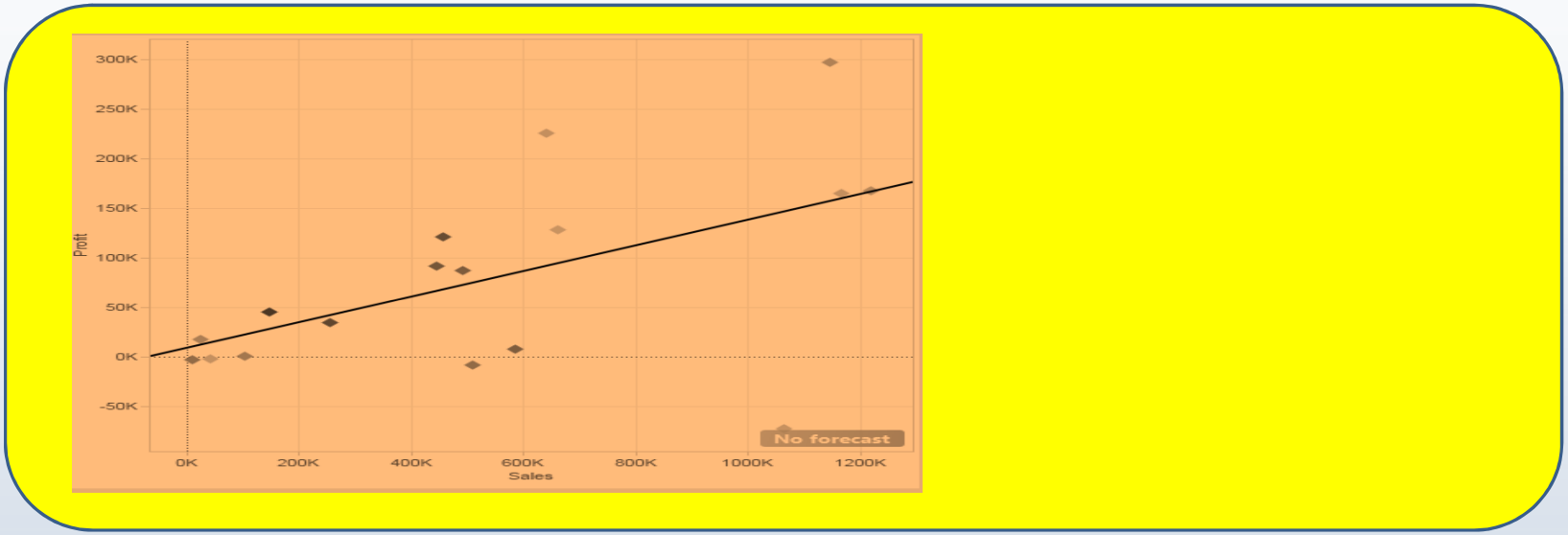


Histogram


9
8
7
6
5
4
3
2
1
0
114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129
6
5
4
3
2
1
0
6
5
4
3
2
1
0
-1
1-9 10-19 20-29 30-39 40-49 50-59 60-69 70-79 80-89 90-99

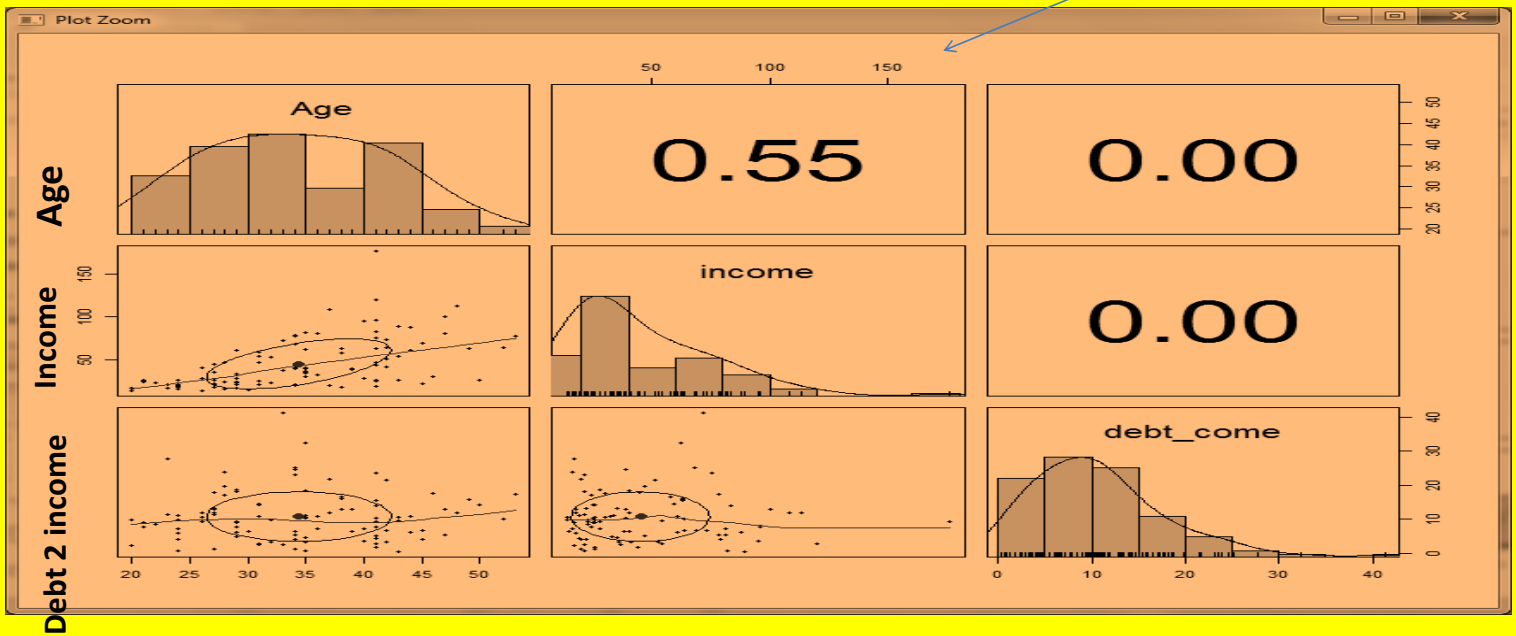
Scatter plot

To graphically represent the relationship between two variables



Matrix Scatter Plot (SPLOM)

Refer the dataset debttrain and the code is SPLOM.R



Descriptive Statistics

Descriptive Summary Statistics – dealing with known's

What is Descriptive Statistics ?

Summarizing or describing the fact known to you Organizes and make sense of Data uses Numerical and Graphical Methods Identifies Patterns in Data Simplifies the information focusing on the items/areas of interest Eliminates undesired information to avoid information overload

Measures of Central Tendency
Mean
Median
Mode
Geometric mean

Measures of Dispersion
Min and max
Range
Standard Deviation

Others
Skewness
Kurtosis
Distribution Plot

Mean

Data Scientist	Data Analyst
58350	$48670
63120	$57320
44640	$38150
56380	$41290
72250	$53160
	$500,000

Average Salary = 48000

X bar = $ 123,098

Summary Descriptive Statistics – using excel

Cust Id	Amount Spent
1	250
2	300
3	280
4	270
5	320
6	290
7	260
8	280
9	240
10	260

No. of Observations = 10
SUM = 2750

MEAN = (2750/10) = 275

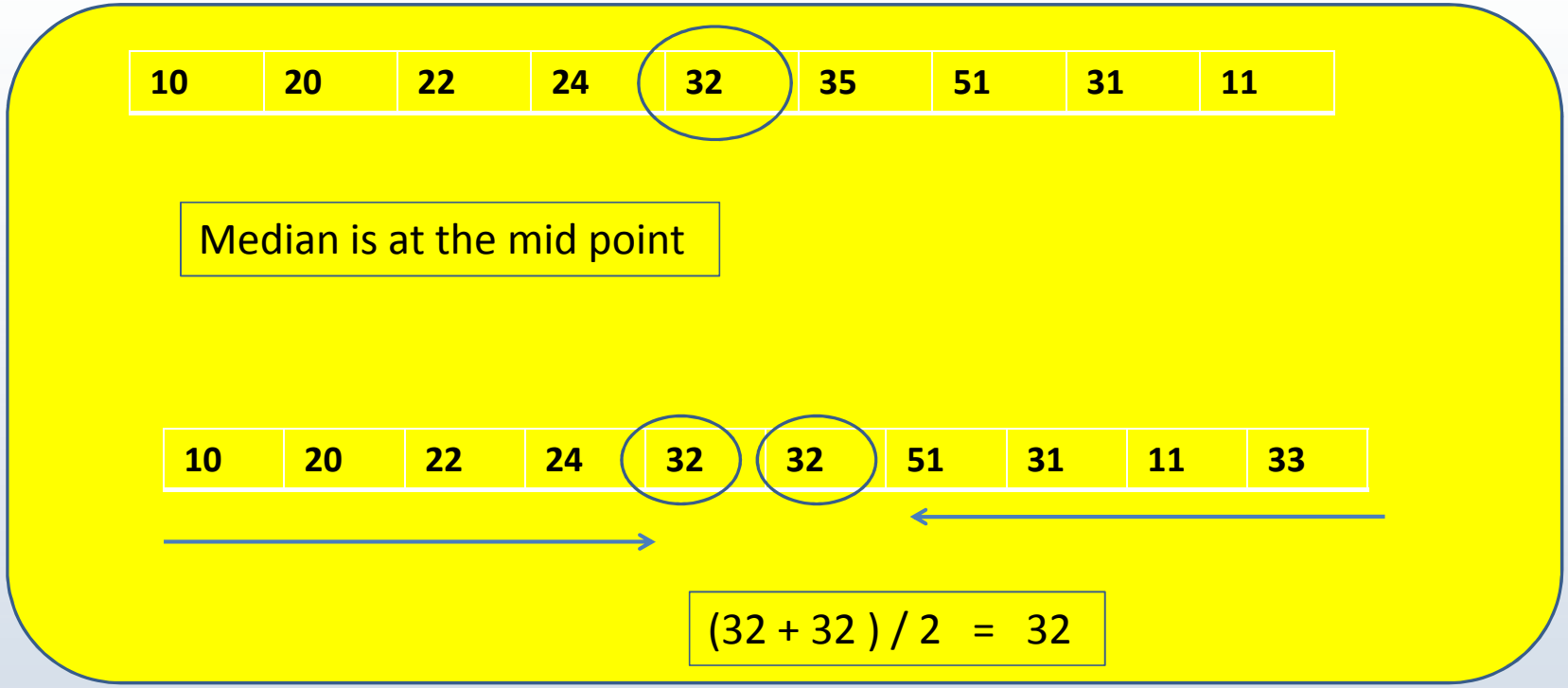
- Works well when data is not heavily skewed
- Easy to compute

Quiz 1 : Mean

What are the properties of the mean ? (put a tick mark)

- All salaries in the distribution affect the mean
- Mean can be described with a formula
- Many samples from the same population will have similar means
- The mean of a sample can be used to make inferences

Median


10 20 22 24 32 35 51 31 11
Median is at the mid point
10 20 22 24 32 32 51 31 11 33
(32 + 32) / 2 = 32

Median

Data Scientist	Data Analyst
58350	$48670
63120	$57320
44640	$38150
56380	$41290
72250	$53160

What is the median value here ?

Median

Data Scientist	Data Analyst
58350	$48670
63120	$57320
44640	$38150
56380	$41290
72250	$53160

What do we have to do to make the Median a useful statistic ?

1. Calculate the average
2. Put the data in order
3. Eliminate outliers
4. Eliminate data values that repeat

Mean Vs Median

Data Scientist	Data Analyst
58350	$48670
63120	$57320
44640	$38150
56380	$41290
72250	$53160
	$500000

Median is 58350

Mean is 47718
Median is 48670

What if the outlier value was $500000

Mean is 123098
Median is 50915

58350

What is the new mean after introducing the outlier ?

Median

Find the median of this data set ?

Data Scientist	Data Analyst
58350	$48670
63120	$57320
44640	$38150
56380	$41290
72250	$53160

Median with Outlier

Find the median of this data set ?

Data Scientist	Data Analyst
58350	$48670
63120	$57320
44640	$38150
56380	$41290
72250	$53160
	$500000

Where do you think the median is ?

1.$ 48, 670
2. $ 53, 160
3. Anywhere in between $48,670 and $53,160
4. Exactly in between $48,670 and $53,160

Mode

Customer Id	Amount Spent (in $)	Amount Spent (bucket)
1	240	
2	280	
3	270	
4	300	
5	277	
6	267	
7	292	
8	2800	
9	260	
10	250	
11	480	

Mins Bucket	No. of Subscribers
< 300	8
300-500	1
> 500	1

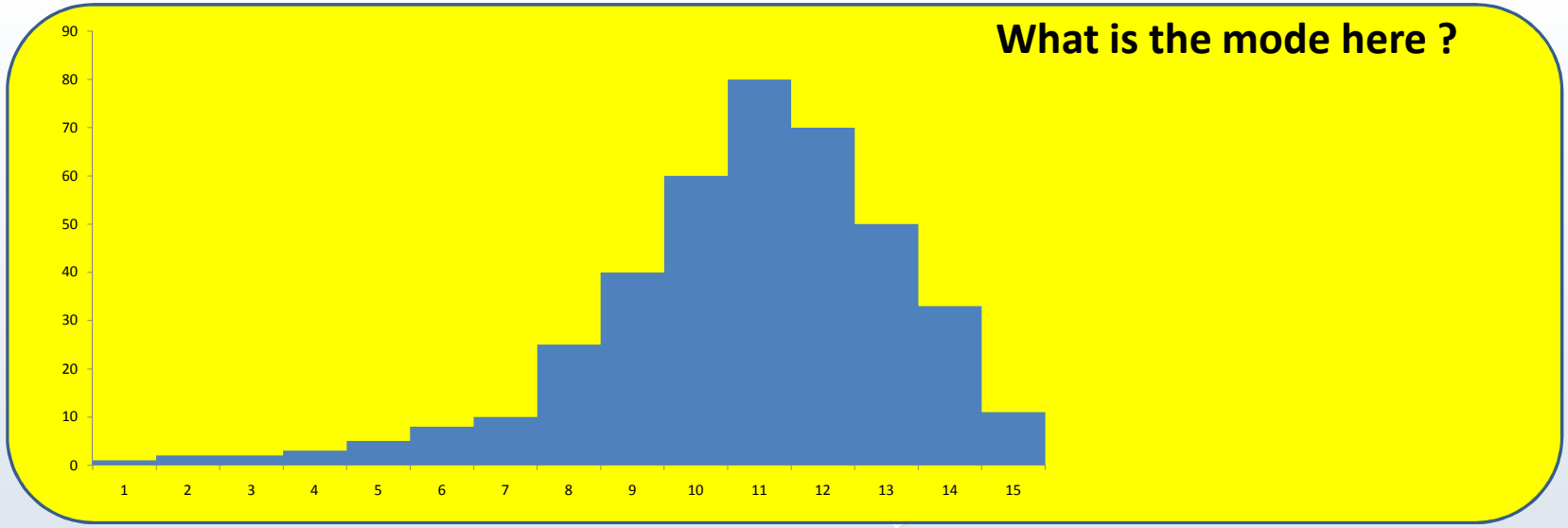
MODE = "<300"

- Works well in "winners take all situations"
- Gives the most popular value
- Easy to Understand

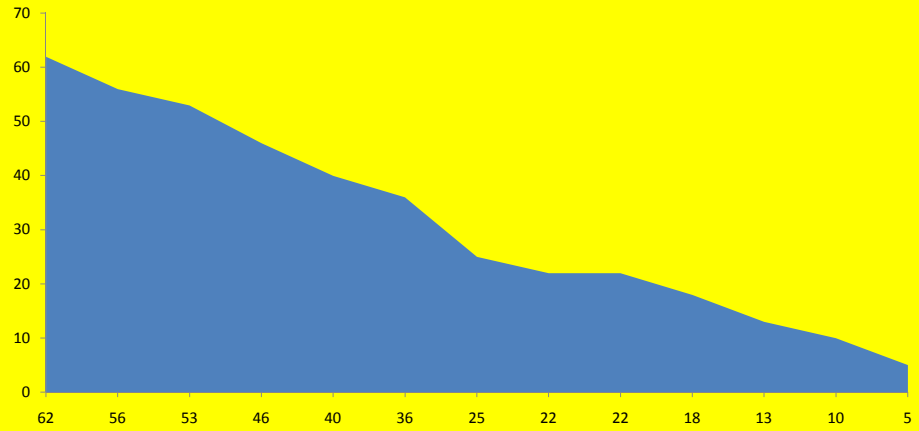
Mode

- One mode: **unimodal**: 1, 2, 3, 3, 4, 5.
- Two: **bimodal**: 1, 1, 2, 3, 4, 4, 5.
- Three: **trimodal**: 1, 1, 2, 3, 3, 4, 5, 5.
- More than one (two, three or more) = **multimodal**.

Quick 1 - Mode for a negatively skewed distribution



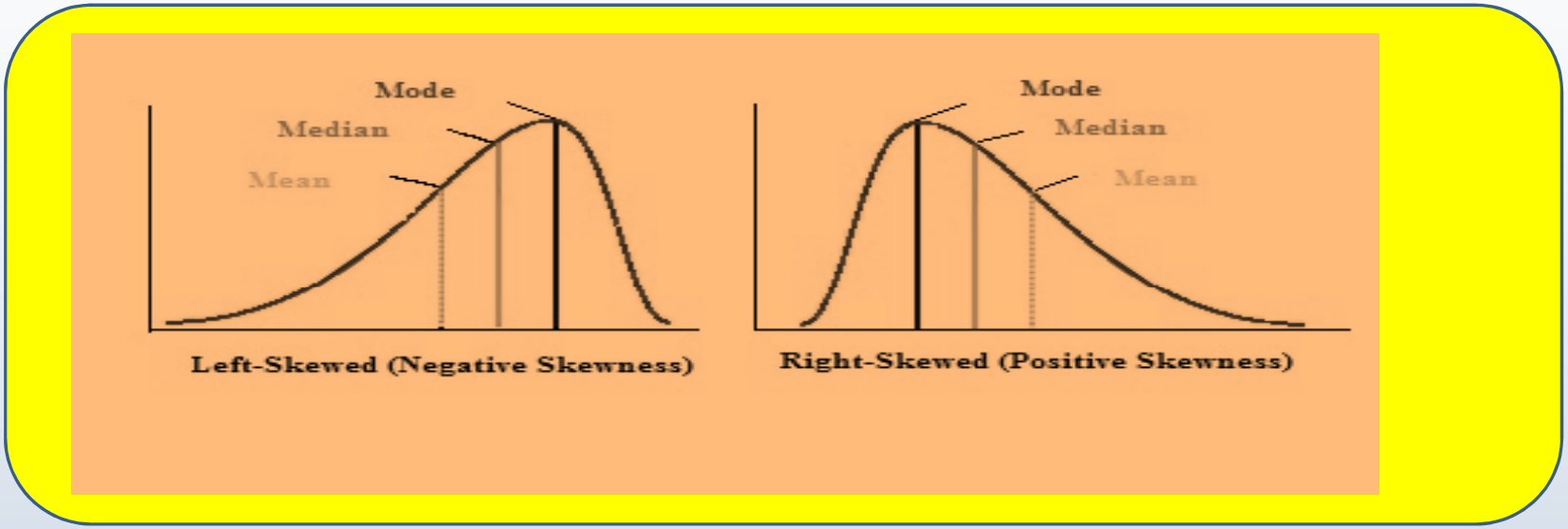
Right Skewed data



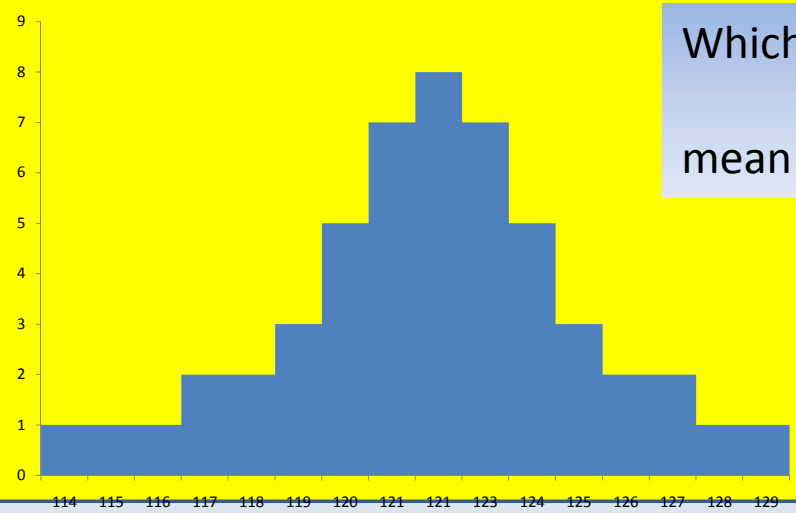
Which one holds true ?

mean < median < mode
median < mode < mean
mode < median < mean
mode < mean < median

Positive and Negative skewed data



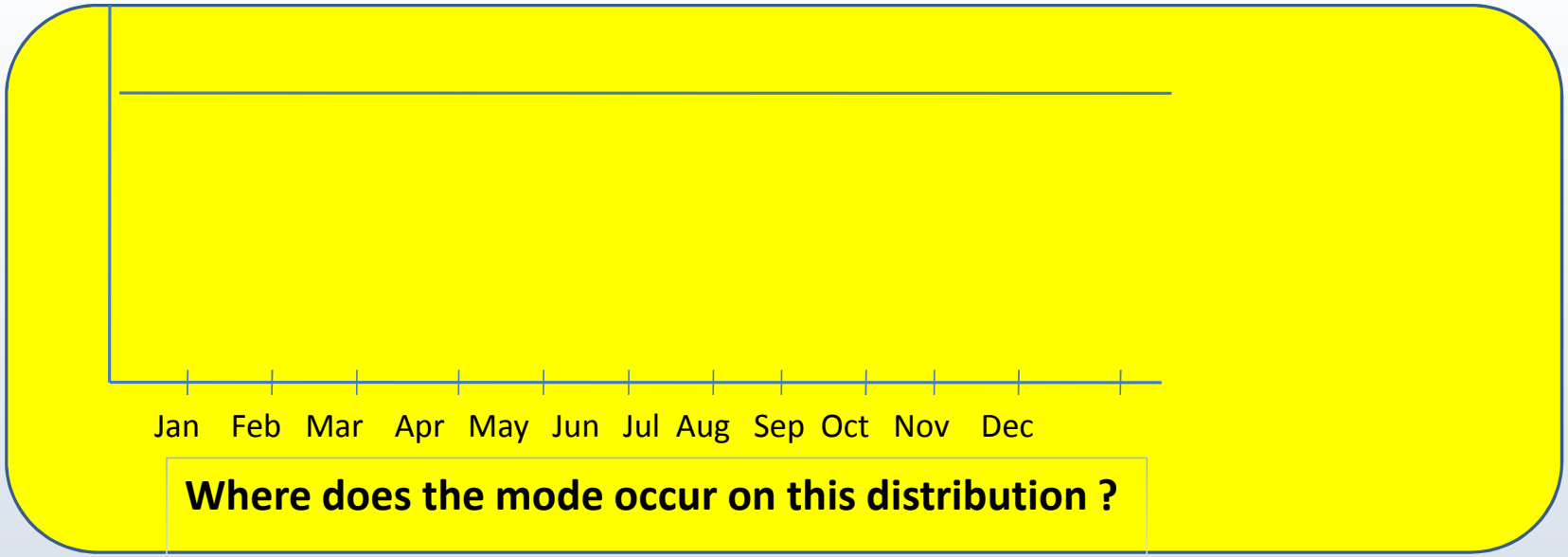
Quiz



Which one is applicable here ?

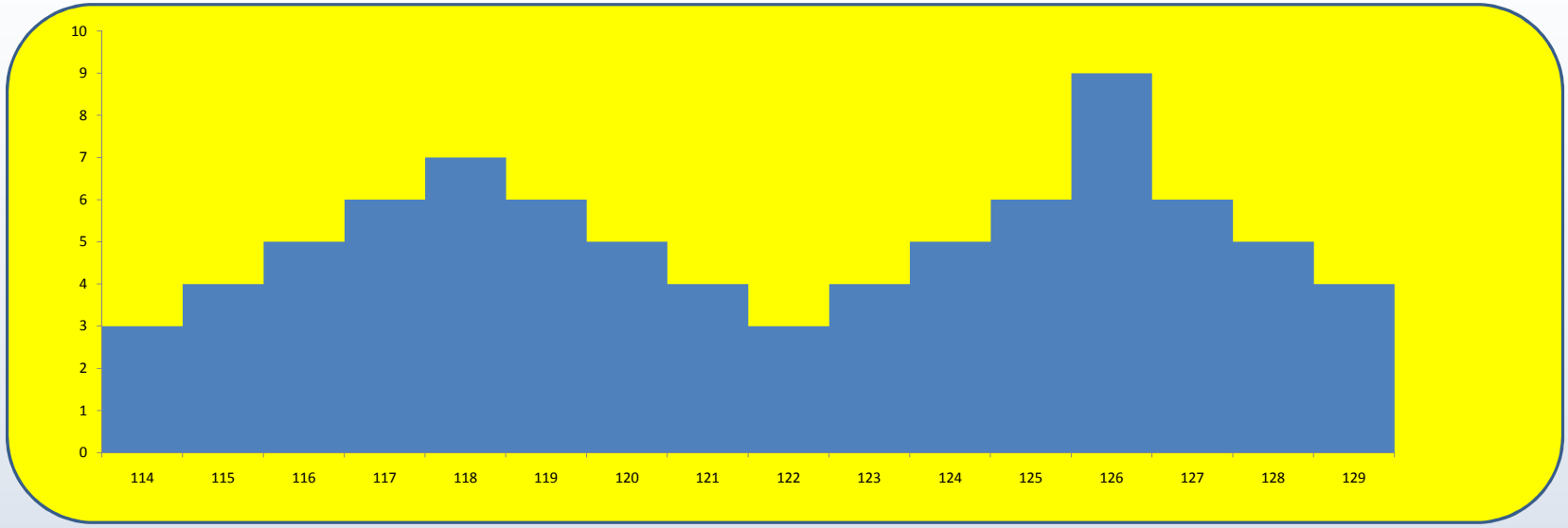
mean ___ median ____mode

Quiz 2 – Mode on a Uniform distribution

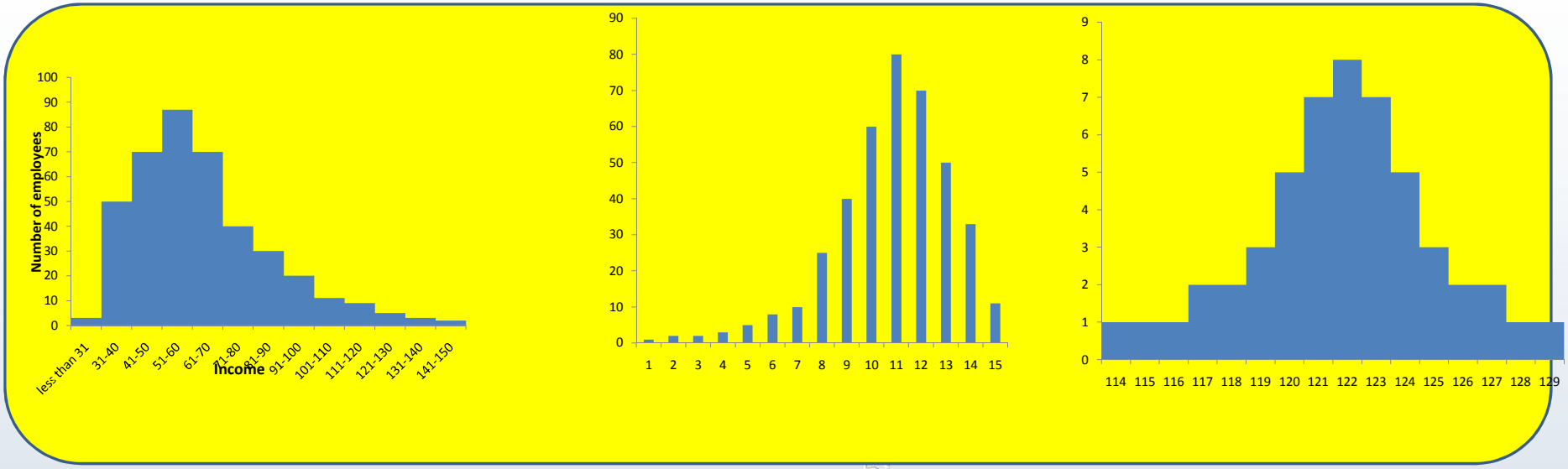


Where does the mode occur on this distribution ?

Quiz 3 – Mode on a bimodal distribution



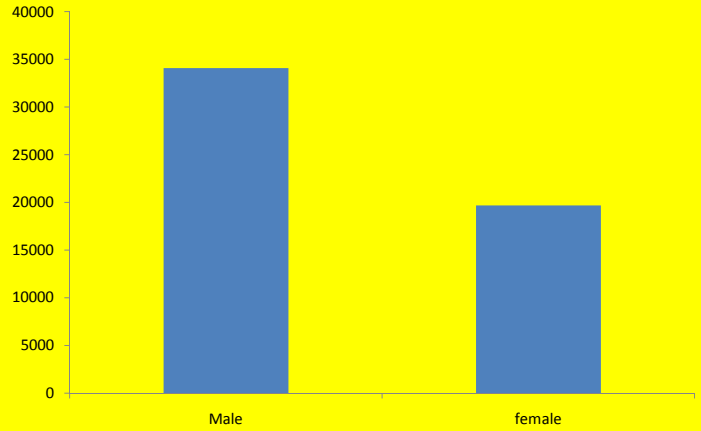
Quiz 4 –Mode of distribution



What is the mode in this case ?

Single number that occurred with the highest frequency
Range that occurred with the highest frequency

Quiz 5 : Mode on Categorical data



What is the mode ?

- 34100
- Male
- 16500
- Female

Quiz 6: Mode

- Using mode we can describe if the data is either categorical or numerical (TRUE/FALSE)
- The expenditures in the below data set affect the mode (TRUE/FALSE)
  - 32, 45 , 32, 25, 28, 32
- There is an equation for the mode (TRUE/FALSE)
- The mode remains same for more than one samples drawn from a same population
- Can a mode change if the bin size in an histogram changes

Geometric Mean

Nth root of the product of n numbers

$[(1+k1)(1+k2)..(1+kn)]^{(1/n)} - 1$

In calculating the GM, the numbers must all be positive. So, add 1 to each value before calculating the GM and subtract 1 from the answer.

Return on Investment from 2 different investments

Investment 1 : 10% -10% 10% -10%

Investment 2 : 50% -50% 50% -50%

Return 1 = (1+10%) *(1-10%)* (1+10%) * (1-10%) = . 98

Average Return – 4 √.98 = .995 or a .5% loss every year

Return2 = (1+50%) *(1-50%) * (1+50%) *(1-50%) = .56

Average Return – 4 √.56 = .87 or a 13% loss every year

Measures of Central Tendency

Measure	Definition	When to use?
Mean	Sum of observations/No. of observations	When data is not very skewed
Median	Middle value or mean of the 2 middle values after sorting the data	When data is highly skewed
Mode	Most popular value	In voting situations or with classes
Geometric mean	Nth root of product of N observations	For averaging rates, areas, volumes etc.

Descriptive Statistics – Measures of Dispersion

- What is Measures of Dispersion?
- Range
- Standard Deviation, Interquartile Range
- Coefficient of Variation

What is measures of Dispersion / Spread ?

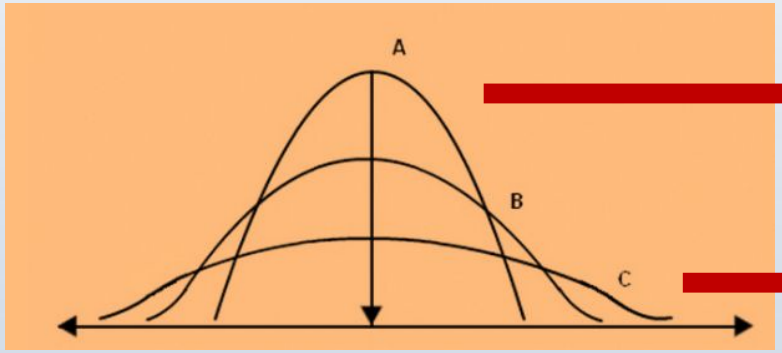
What is measures of dispersion ?

Describes how the data is spreading or the variability

Summary statistics can also be used to understand variation or dispersion in the data.

What is the difference between *Measures of central Tendency* and ***Measures of dispersion*** ?

The difference between measures of central tendency and Measures of dispersion is , central tendency describes the centre of the data ,but it does not tell us anything about the spread of the data We need some measures which show whether the distribution is small or large



Closer spread

Wider spread

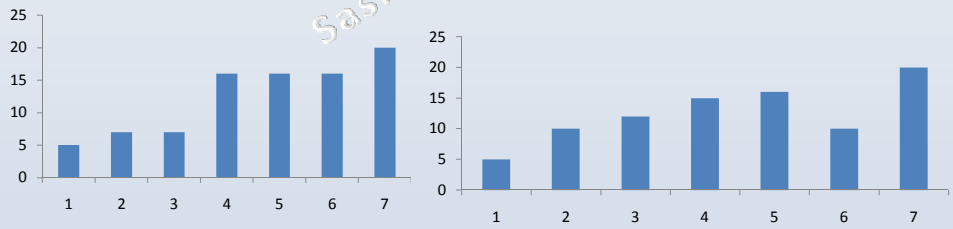
Measures of Dispersion

Range

Measures how spread out a distribution is

- Both the datasets have the similar range , but the values are distributed differently.
- Range shows the width of the data but not how it's dispersed between the bounds.
- Its not the best way of measuring how the data is distributed within the range.
- Range is computed by taking the difference between maximum value and minimum value
- Drawback – It takes into account of outliers

5		5
10		7
12		7
15		16
16		16
10		16
20		20
5		5
20		20

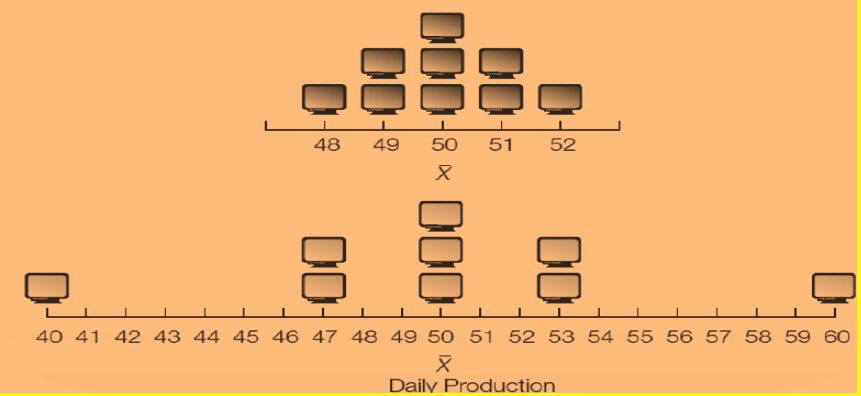


40,89,91,93, 95,100

Application of Range- Case Study

Another reason for examining the dispersion in a set of data is to compare the spread in two or more distributions. For instance, that Sony assembles TVs in North and South Chennai. The arithmetic mean hourly output is both these North and South plant is 40.

With this we can conclude that the distribution of the hourly outputs are identical but this is not correct conclusion



North Chennai production varies from 48 to 52 assemblies per hour. Production at the South plant is more erratic, ranging from 40 to 60 per hour. The range In the hourly production for the South plant is 20 Tvs, found by 60-40

Therefore, the hourly output for North has very less dispersion than the output for the South is more dispersed.

Standard Deviation

Standard deviation is a measure of how close or far are the observations to the mean

Cust id	Amount Spent
1	250
2	345
3	280
4	290
5	175
6	200
7	255
8	150
9	375
10	180

Mean 250

This point is 0 units away from the mean (250-250)

This point is 95 units away from the mean (345-250)

Mean 250

Standard Deviation

Customer Id	Avg. spend (monthly) x	x -	(x -) ^2
1	304	69.2	4788.64
2	50	-184.8	34151.04
3	252	17.2	295.84
4	298	63.2	3994.24
5	234	-0.8	0.64
6	228	-6.8	46.24
7	264	29.2	852.64
8	230	-4.8	23.04
9	228	69.2	4788.64
10	260	-6.8	46.24

Variance = ∑(x –)^2/ N

Standard Deviation – sqrt(sigma^2)

X = Observation

= population mean

N = number of observations in the population

Mean = = 234.8

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

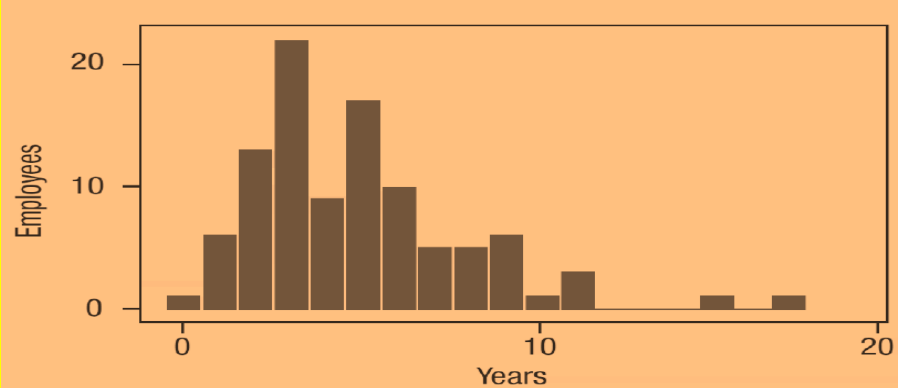
Standard deviation= $\sigma = \sqrt{\frac{\sum(X-\mu)^2}{N}}$

Std Deviation = 22.13

Application of Standard Deviation - Case Study 1

Lets say in a company 1000 employees are organized into a histogram based on the number of years of experience with the company. The mean is 4.9 years, but the spread of the data is from 6 months to 16.8 years.

The mean of 4.9 years is not very representative of all the employees



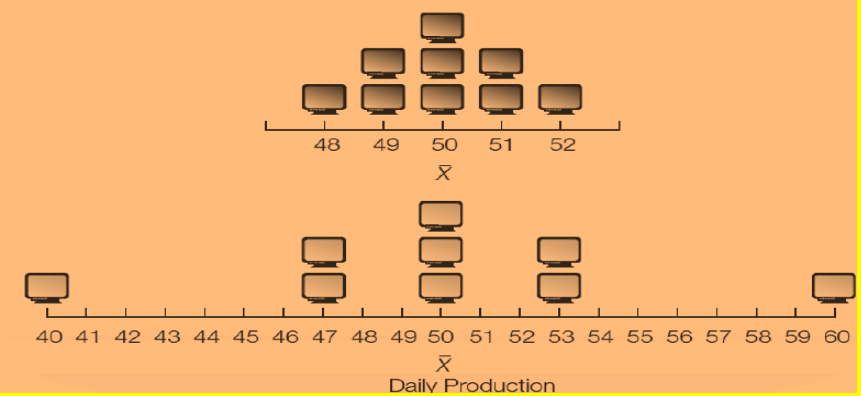
The 100 employees of this company are organized into a histogram based on the number of years of employment with the company.
The mean is 4.9 years, but the spread of the data is from 6 months to 16.8 years.

The mean of 4.9 years is not very representative of all the employees.

Application of Standard Deviation - Case Study 2

Another reason for examining the dispersion in a set of data is to compare the spread in two or more distributions. For instance, that Sony assembles TVs in North and South Chennai. The arithmetic mean hourly output is both these North and South plant is 40.

With this we can conclude that the distribution of the hourly outputs are identical but this is not correct conclusion

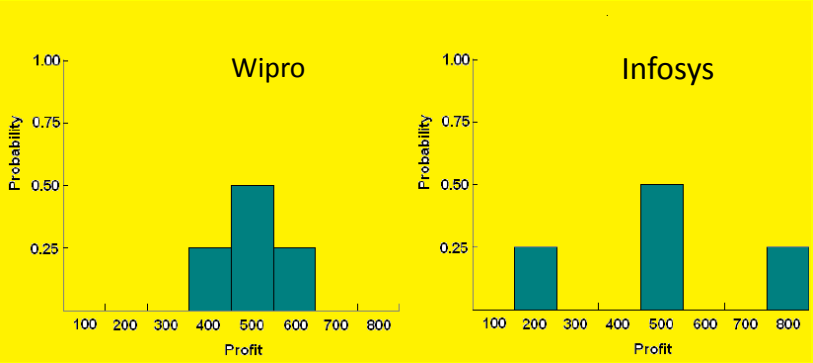


North Chennai production varies from 48 to 52 assemblies per hour. Production at the South plant is more erratic, ranging from 40 to 60 per hour.

Therefore, the hourly output for North is clustered near the mean of 50; the hourly output for the South is more dispersed.

Application of Standard Deviation - Case Study 3

Investment	(1) State of Economy	(2) Probability of Occurrence	(3) Outcome of Investment	(4) Expected Value (2) x (3)
Wipro	Boom	0.25	€600	€150
	Normal	0.50	€500	€250
	Recession	0.25	€400	€100
				€500
Infosys	Boom	0.25	€800	€200
	Normal	0.50	€500	€250
	Recession	0.25	€200	€50
				€500

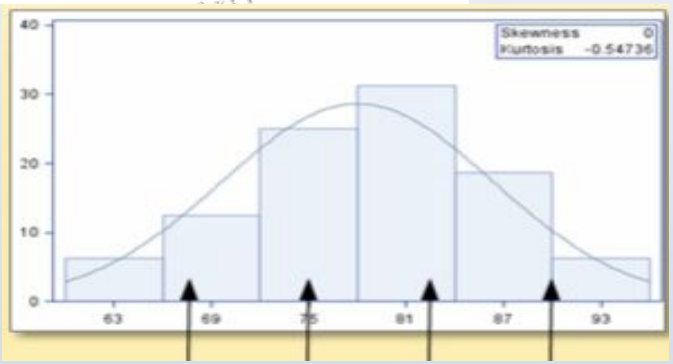


Note that the relationship between the state of the economy and profits is much tighter (i.e. less dispersed) for Wipro than for Infosys. Therefore, Wipro is less risky than Infosys.

Since both investments have the same expected profit, Wipro is preferable to Infosys.

Percentiles

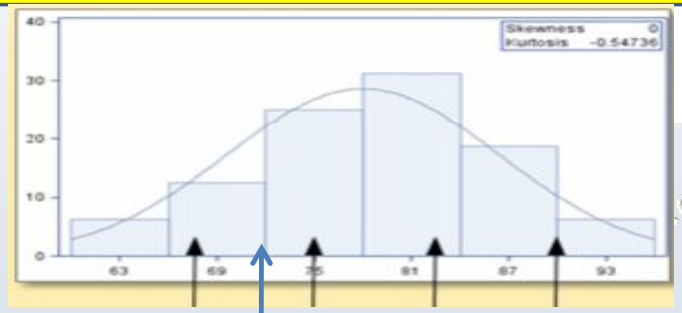
- Percentiles are descriptive statistics that give us reference points in our data.
- A percentile is the value of a variable below which a certain percentage of observations fall.
- The most commonly reported percentiles are quartiles, which break the data up into quarters.



Percentiles

Most commonly reported percentiles are quartiles,which break the data up into quarters

- Sort the data
- We have an even number of data , this means that when we calculate the quartiles , we take the sum of the two values around each quartile and average them

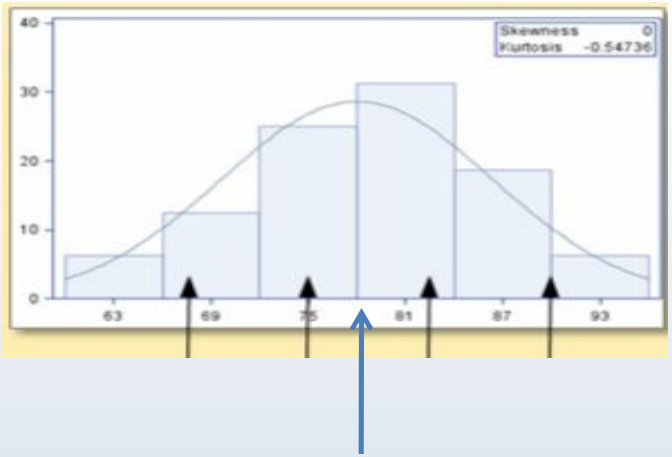


93
89
88
84
83
82
79
78
78
77
74
73
72
68
67
63

25^{th} Percentile = 72.5

25^{th} percentile can also be referred to as 1^{st} quartile, Q1 , or the lower quartile

Percentile



50th Percentile = ?

93
89
88
84
83
82
79
78
78
77
74
73
72
68
67
63

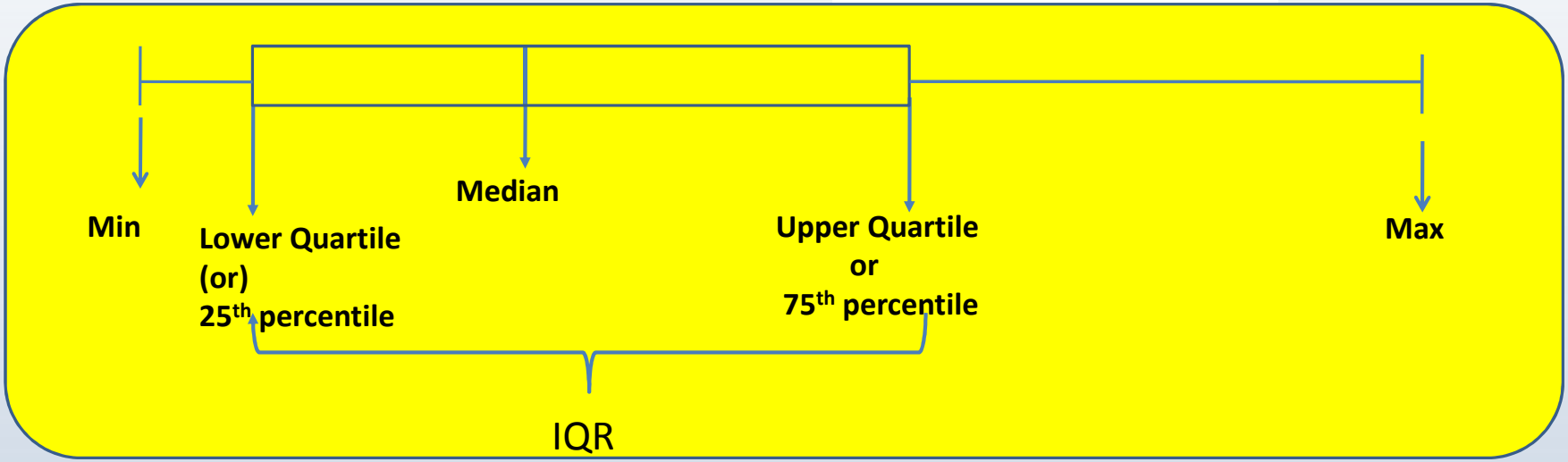
50th percentile can also be referred as Median

83.5, it means that 50% of the data values fall at or below 83.5

Boxplot

What is Boxplot ??

It's a visual representation which helps us to understand how spread the data and to detect the outliers. In order to construct the same, we need min, Q1, median, Q3 and the max value. To determine central tendency, spread, skewness, and the existence of outliers.



Exercise 1

19 19 20 21 22 22 22 23 23 24 25

Q1 Q3

¼ or 25% of the data has a value that is less than or equal to 20

½ or 50% of the data has a value that is less than or equal to 22

¾ or 75% of data that has a value that is less than or equal to 23

½ or 50% of the data lies between 20 and 23

Depends on the context, sometimes
Low percentile = good
High percentile = good

Brain Teaser

A customer Quality Assurance manager , after inspecting the product upto 1 pm , he found 10 defects. Is this is a good sign or bad sign ?

If this is out of 100, it's a good sign
If this is out of 10, it's a bad sign

After getting the update from a sales person about his sales accomplishment over phone , the manager noted down the sales made as 25

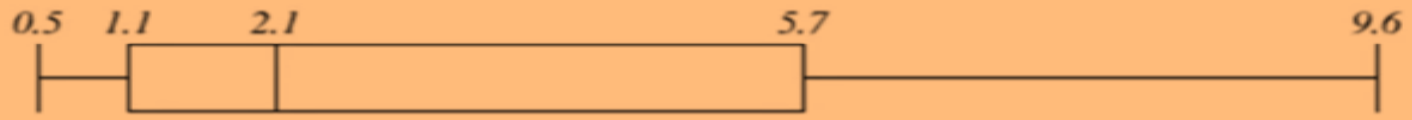
If its made out of 50 customers, its a good sign
If out of 100, is a bad mark

Visualization

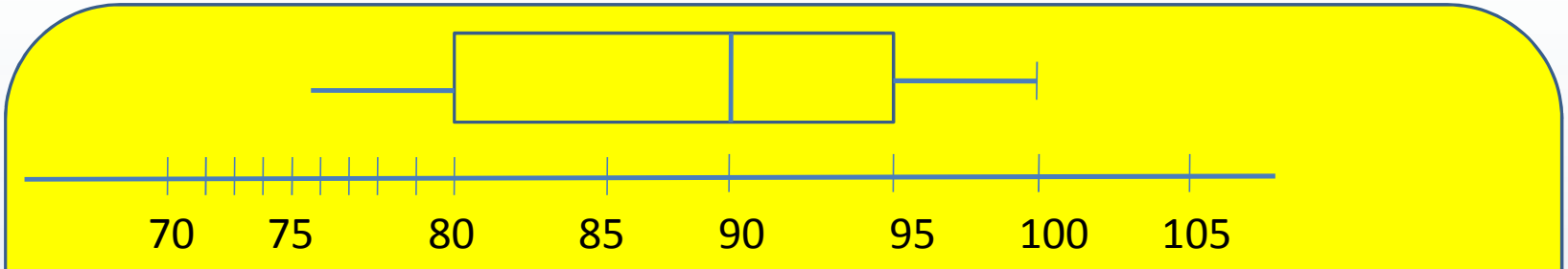
Delivery time in minutes

Heritage Fresh delivery time – South Region

Heritage Fresh delivery time – North Region



Boxplot – Assignment



What was the lowest sales achieved ?
What was the highest sales achieved ?
What was the median Sales achieved ?
The middle 50% of the sales achieved were between which scores ?
The majority of the sales were above 85 , true or false ?
Top 25% of the sales were between which two ranges ? :

Exercise 1

Compute the Lower Quartile, median and upper quartile

Age	19	20	21	22	23	24	25
No. of customers	2	1	1	3	2	1	1

19 19 20 21 22 22 22 23 23 24 25

19 19 20 21 22 22 22 23 23 24 25

A potential outlier is any observation that falls beyond 1.5 times the width of the box on either side, that is, any observation
less than Lower Quar le – 1.5(Upper Quar le– Lower Quar le)
or
greater than Upper Quar le + 1.5(Upper Quar le – Lower Quartile).

Exercise 2

110 110 130 130 140 150 160 160 170 170

110 110 130 130 140 150 160 160 170 170

Quartile 1 = 130
Quartile 2 = 145
Quartile 3 = 160

Quartiles

Sample 1

38946
43420 → Q1
49191
50430
50557 — Q2
52580
53595 → Q3
54135
60181
10,000,000

IQR = Q3 – Q1

Sample 2

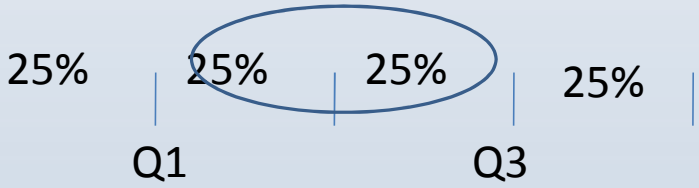
33219
36254 → Q1
38801
46335 — Q2
46840
47596 → Q3
55130
56863
78070
88,830

IQR = Q3 – Q1

About 50% of the data falls within the IQR ?

IQR is affected by every value in the data set

IQR is not affected by outliers



Standard Deviation Vs IQR ?

A = {1,1,1,1,1,1,1} and B = {1,1,1,1,1,1,100000000}.

IRQ for both is 0, but SD is very different.
Which one is is really better ?
this example very nicely illustrates that the IQR tells you where the middle 50% of the data is located while the SD tells you about the spread of the data.

It also shows that the IQR is very resistant to outliers (and to some degree skew) while the SD is not

Shape Measures

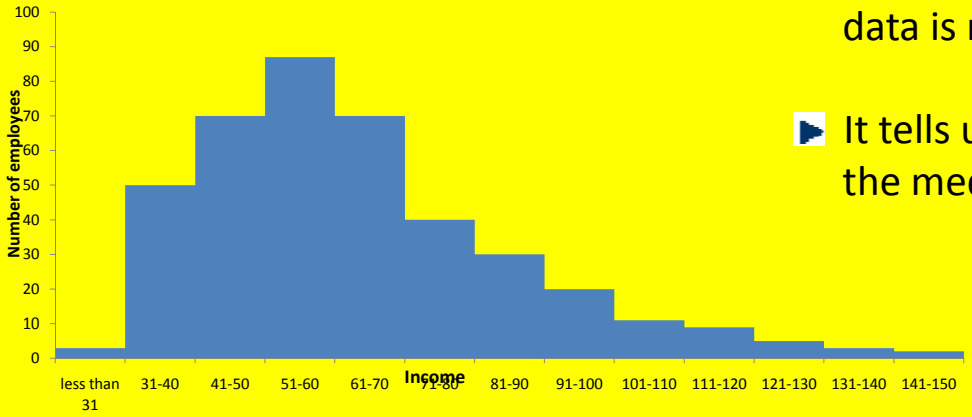
Shape measures

Symmetry

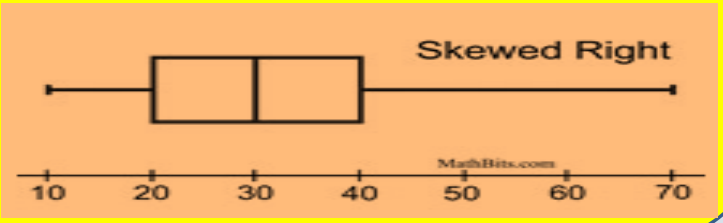
How does data fall on either side of the set mean ? Is it symmetric or skewed ?
Is it possible to have symmetric distributions with more than one peak ?

Skewness : Degree of Asymmetry

Positive/Right Skew

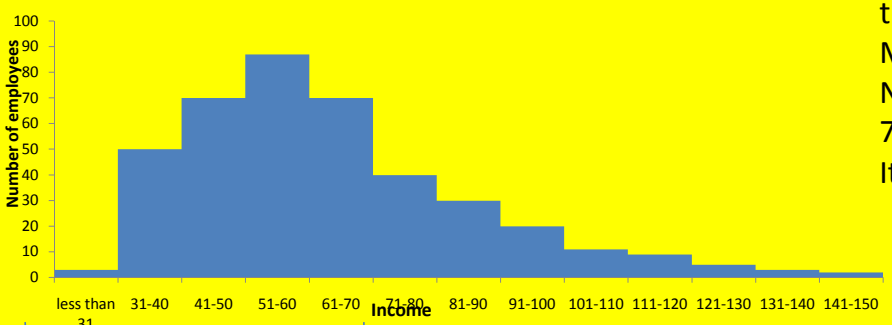


- When the skewness statistic is +ve, the data is right-skewed.
- It tells us that the mean is greater than the median



Positive/Right Skew

When the skewness statistic is +ve, the data is right-skewed.



What can we say about this distribution ?

most employees have an income greater than 131k ?

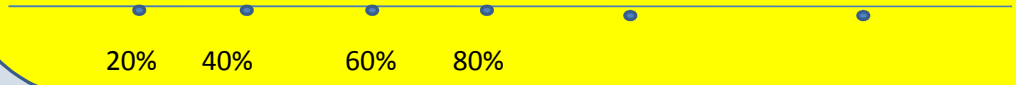
Most common income is less than 60,000

No employee have an income greater than 70,000 ?

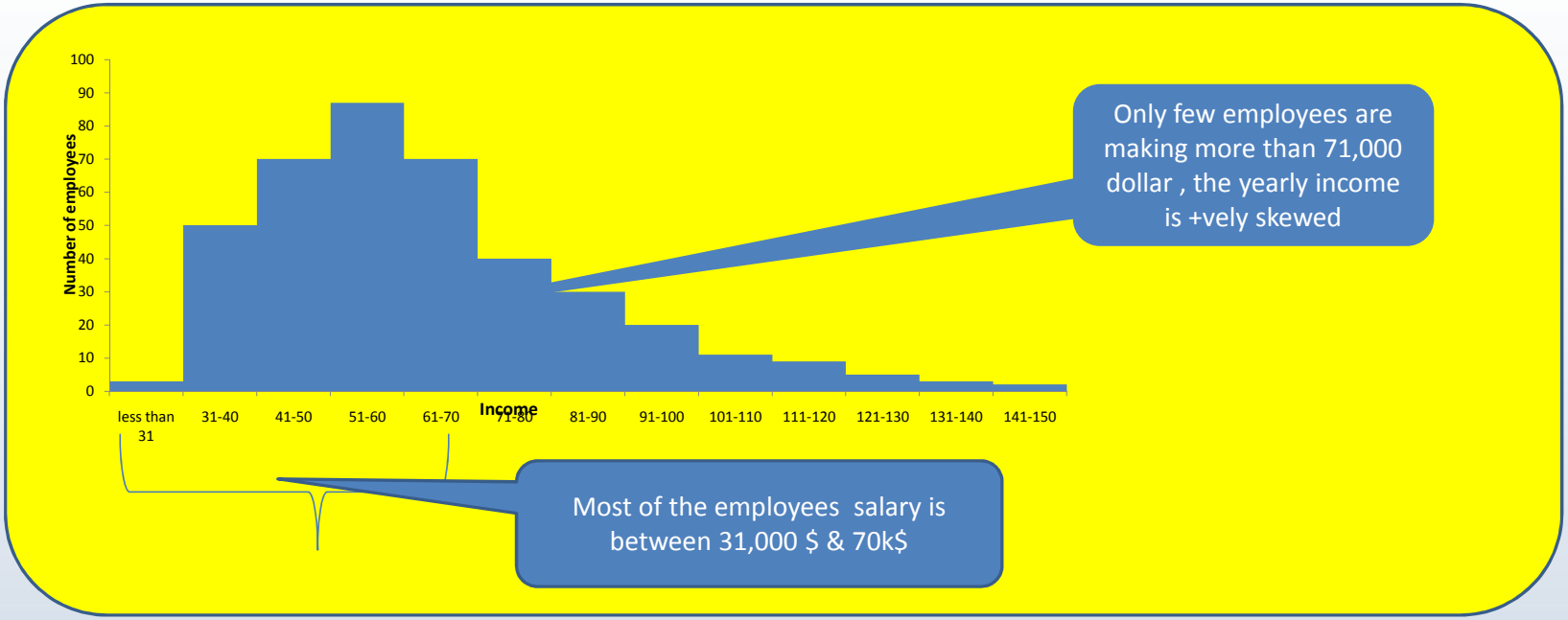
It's a symmetrical distribution

Only few employees are making more than 71,000 dollar , the yearly income is +vely skewed

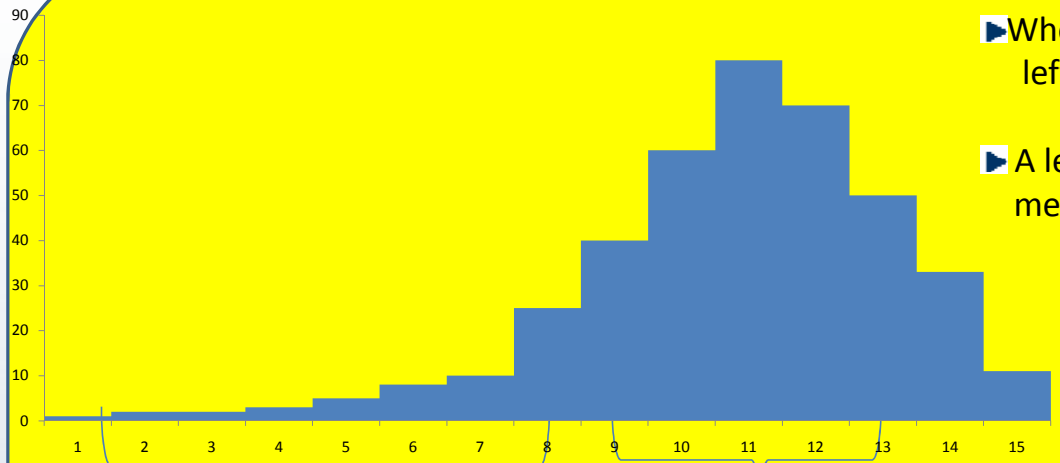
Most of the employees salary is between 31,000 $ & 70k$



Positive Skew



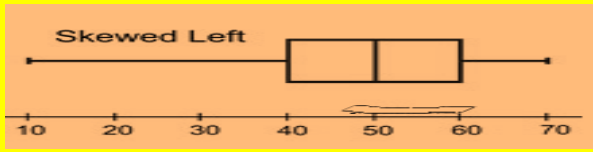
Negative skew or Left-skew - example



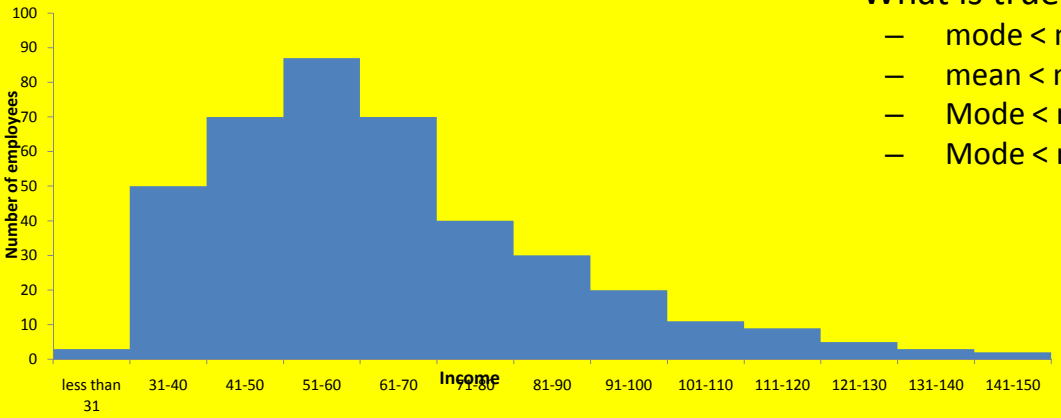
- When the skewness statistic is -ve, the data is left-skewed.
- A left skewed distribution tells us that the mean is less than the median.

Only 56 of the 400 candidates completed less than 9 application

Most of the candidates completed between the 9 & 13 application s



Quiz

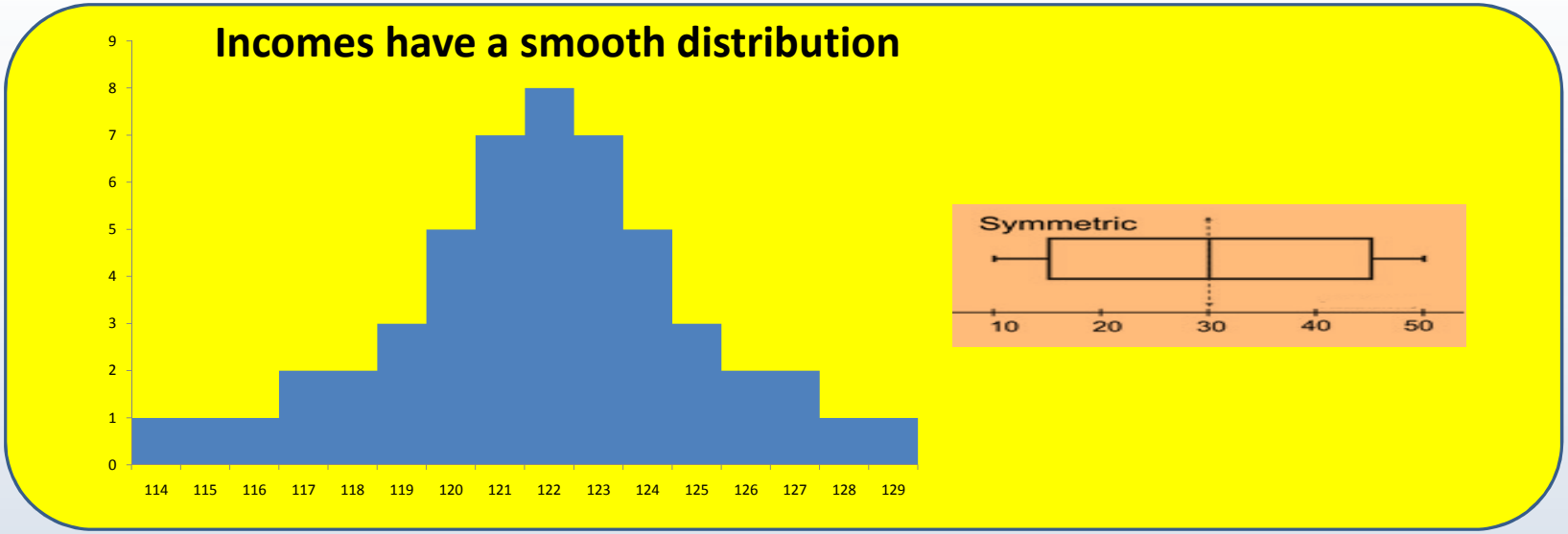


- What is true about this distribution ?
  - mode < mean < median
  - mean < median < mode
  - Mode < median < mean
  - Mode < mean < median

Least

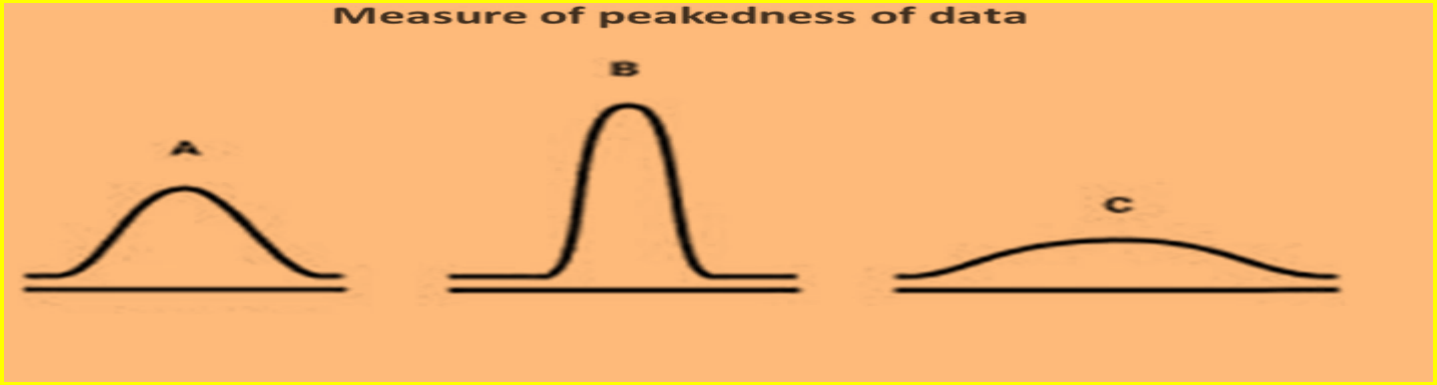
Highest

Normally Distributed data - example

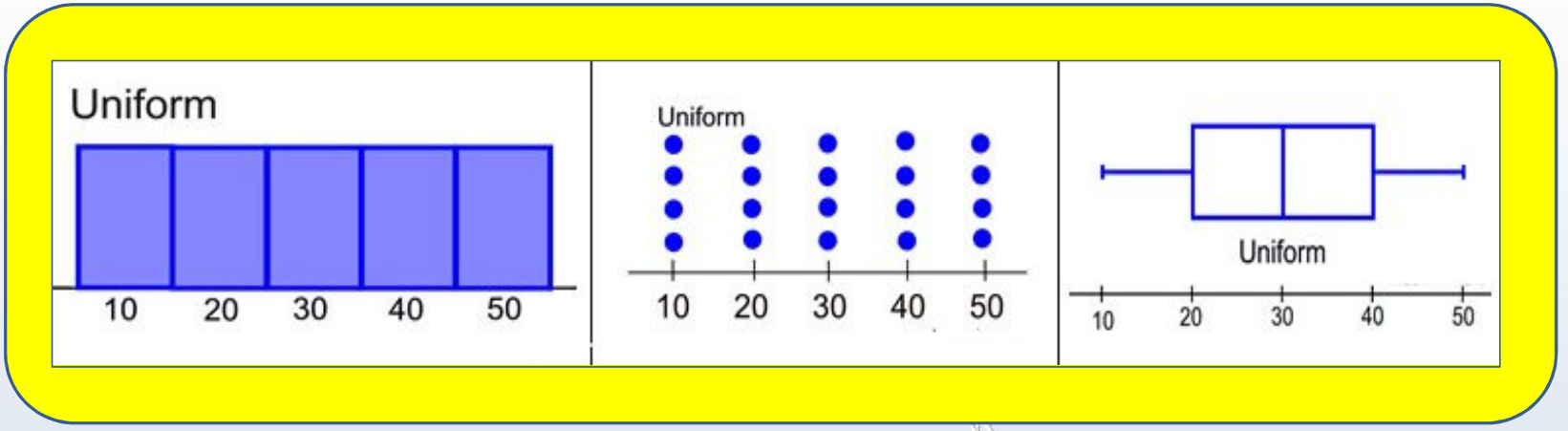


Kurtosis

Measuring the peakedness of data, reflecting the shape of the peak relative to the rest of the distribution



Uniform Distribution



Bi modal distributions

Summary

If the mean and median coincide and the skewness and kurtosis statistics are close to 0, it is likely that your data comes from a normal distribution

Descriptive Statistics – Case Study

Now , we can see how we can use retail data set to apply descriptive statistics to gain insights

Descriptive Statistics

Representative Business questions that can be answered using descriptive statistics:

- What is the distribution of revenue ?
- Are there any difference in products ordered via Telephone and web ?
- What is the frequency distribution of product type ?
- What proportion of product line sold by Category (Product type) ?
- What is the average revenue by product line ?
- Is there any difference in revenue amongst the ordered method Type ? Which mode of Sales generated more revenue ?

```
retail  = pd.read_csv('D:/dat1/retailSales.csv', sep=',')
retail['Revenue'] = retail['unit price'] * retail['Unit Sale']
retail.head()
retail.Revenue.hist()
```