BASICS OF STATISTICS AND PROBABILITY

What is Statistics?

proschool
An IMS Initiative

Statistics is a form of mathematical analysis that uses models, representations and synopsis for a given set of experimental data or real-life studies.

Statistics is “the science of the collection, analysis, interpretation, presentation, and organization of data.”

There are variety of descriptive statistics:

- Measures of central tendency – mean, median, mode
- Measures of dispersion – range, variance, standard deviation
- Measures of shape – skewness, kurtosis

Introduction

In general, statistics summarizes information about data in a meaningful and relevant way.

For example: Describe the population of India?

– population in 2017 is 1,342,512,706
    - That is a statistic – the total sum of all full-time residents of India
- What other statistics can you think of?

“Population density”

“Median Age”

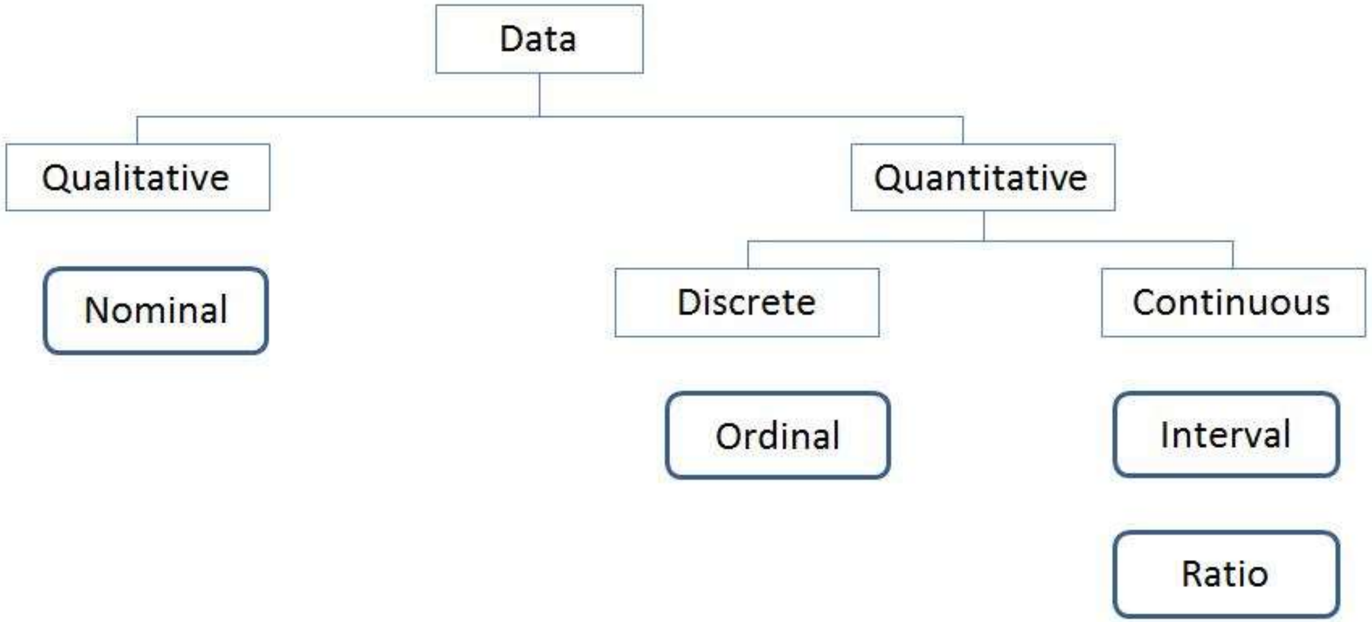
“Distribution by Religion”

“Literacy Rate”

All these statistics summarize information because talking about each data point is impossible.

Data/Attribute Types

- Attribute (or dimensions, features, variables): a data field, representing a characteristic or feature of a data object.
  - E.g., customer _ID, name, address
- Types:
  - Nominal = Categorical
  - Binary
  - Ordinal
  - Numeric: quantitative
    - Interval-scaled
    - Ratio-scaled



Attribute Types

- **Nominal:** categories, states, or "names of things"
  - Hair_color = {auburn, black, blond, brown, grey, red, white}
  - marital status, occupation, ID numbers, zip codes

- **Ordinal**
  - Values have a meaningful order (ranking) but magnitude between successive values is not known.
  - Size = {small, medium, large}, grades, army rankings

- **Binary**
  - Nominal attribute with only 2 states (0 and 1)
  - Symmetric binary: both outcomes equally important
    - e.g., gender
  - Asymmetric binary: outcomes not equally important.
    - e.g., medical test (positive vs. negative)
    - Convention: assign 1 to the important outcome (e.g., HIV positive)

Numeric Attribute Types

- Quantity (integer or real-valued)

- Interval
  - Measured on a scale of equal-sized units
  - Values have order
    - » E.g., *temperature in C˚or F˚, calendar dates*
  - No true zero-point
  - Rare Attribute Type
- Ratio
  - Inherent zero-point
  - We can speak of values as being an order of magnitude larger
    than the unit of measurement (10 K˚ is twice as high as 5 K˚).
    - e.g., temperature in Kelvin, length, counts, monetary quantities

Discrete vs Continuous

- **Discrete Attribute**
  - Has only a finite or countably infinite set of values
    - E.g., zip codes, profession, or the set of words in a collection of documents
  - Sometimes, represented as integer variables
  - Note: Binary attributes are a special case of discrete attributes

- **Continuous Attribute**
  - Has real numbers as attribute values
    - E.g., temperature, height, or weight
  - Practically, real values can only be measured and represented using a finite number of digits
  - Continuous attributes are typically represented as floating-point variables

Analysis of Attribute Types

Scale	Definition	Examples
Nominal	Categorizes but does not rank	Industries, Gender, Occupation
Ordinal	Ranked Categories. Differences between ranks not consistent	Organizational Hierarchy. Star Ratings
Interval	Ranks Data. Differences between ranks equal. No True Zero Point.	Celsius or Fahrenheit Scale. Dates
Ratio	Ranks Data. Differences between ranks equal. Also has a True Zero Point.	Rate of Return, Money

Types of Statistics

There are two types of statistics:

- Descriptive Statistics – Descriptive Statistics deals with analysis and methods related to collection, organization, summarizing and presentation of data.
  Applying the techniques of descriptive statistics, the raw data is collected and transformed into a meaningful form.

- Inferential Statistics - Inferential statistics draws conclusion and makes decision about population using information drawn from a sample.

What is Data Series & Dataset?

- **Data Series:** A row or column of numbers that are plotted in a chart is called a data series.

  Data Series 1:

  19,4,33,2,51,32,2,41,18,2,4,1

- **Dataset :** It is a collection of all related sets of information that is composed of separate elements but can be manipulated as a unit by a computer
- Lets consider a dataset of air quality to summarise all the measures:

Serial no.	Solar Radiation	Wind	Temp	Month	Day
1	190	7.4	67	5	1
2	118	8	72	5	2
3	149	12.6	74	5	3
4	313	11.5	62	5	4
5	299	8.6	65	5	7
6	99	13.8	59	5	8
7	19	20.1	61	5	9
8	194	8.6	69	5	10
9	256	9.7	69	5	12
10	290	9.2	66	5	13

Mean

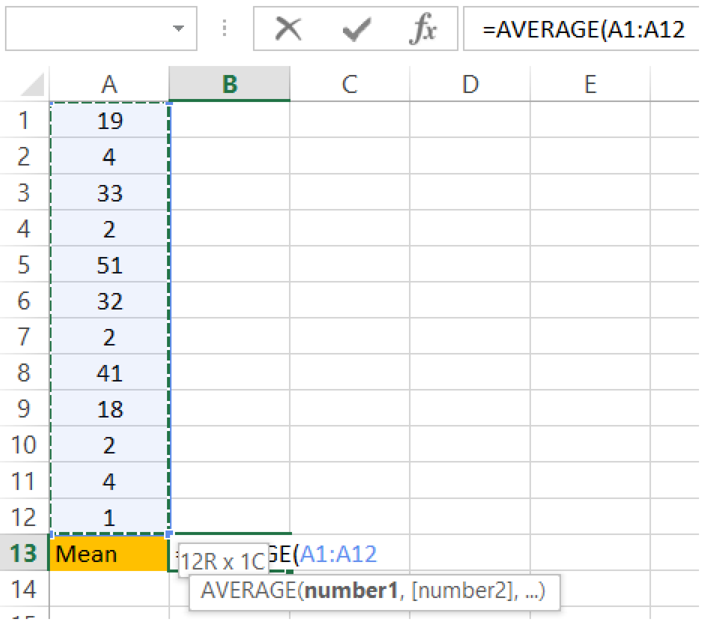
- The mean is the simple mathematical average of a set of two or more numbers
- The mean is the most common measure of the location of a set of points However, the mean is very sensitive to outliers.
- Mean can only be used with numeric data
- In Excel -> It can be computed by Average()
- In R and Python -> mean()

$$\text{Mean}(X) = \frac{1}{m}\sum_{i=1}^{m} Xi$$

Let's continue example 1

Mean = 17.41

(19+4+33+2+51+32+2+41+18+2+4+1)/12



Median

- The middle number; found by ordering all data points and picking out the one in the middle(or if there are two middle numbers, taking the mean of those two numbers).
- It may be thought of as the "middle" value of a data set.

$$Median(X) = \begin{cases} X_{(r+1)} & \text{If } m \text{ is odd, i.e., } r = (m-1)/2 \\ \frac{1}{2}(X_{(r)} + X_{(r+1)}) & \text{If m is even, i.e., } r = m/2 \end{cases}$$

- Where, m = Total number in a dataset

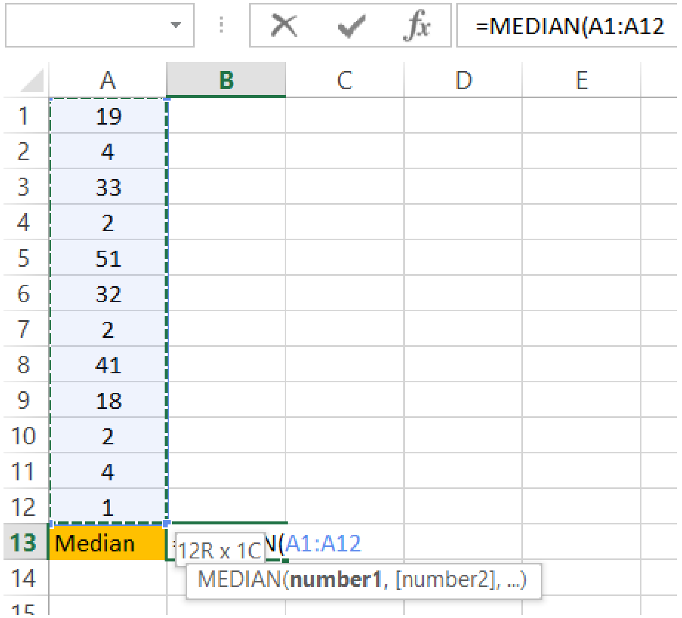
  r = Position of the middle value (In case of even No., Select Nos. those are on same distance from both sides)

Lets continue example 1:

Arrange data in increasing order 1,2,2,2,4,$\boxed{4,18,}$19,32,33,41,51

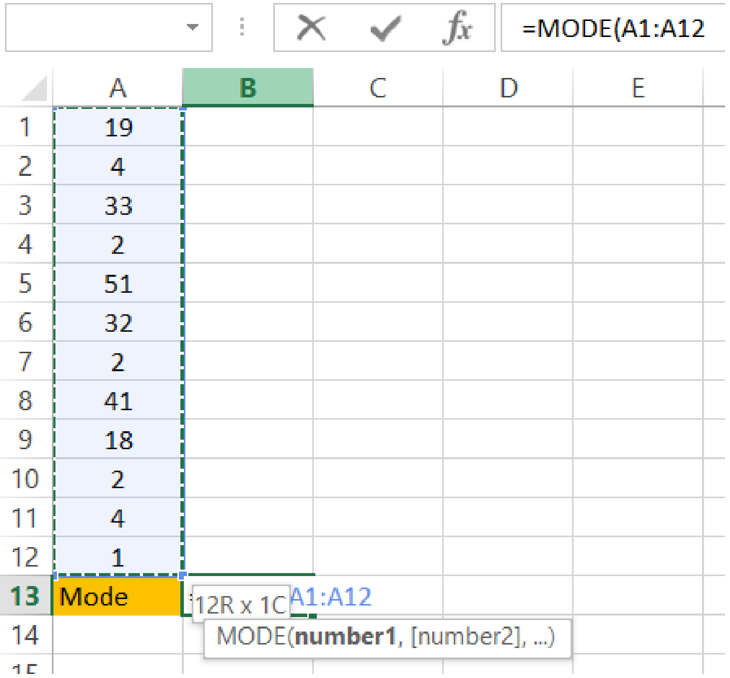
As m is even, take an average of 2 middle numbers

Median = 11 (Calculate i.e. (4+18)/2)



Mode

- The frequency of an attribute value is the numbers of times the value occurs in the data set.
- It is found by collecting and organizing the data in order to count the frequency of each result.
- The mode is the most frequent number—that is, the number that occurs the highest number of times.
- The notions of frequency and mode are typically used with categorical data but it can be used on any data type.
- Lets continue example 1: The dataset is 19,4,33,2,51,32,2,41,18,2,4,1 and mode is 2



Data Distribution

- We can describe the series we looked at in the example 1 as: 19,4,33,2,51,32,2,41,18,2,4,1

"Minimum of 1, Maximum of 51, Average of 17.41."

- Given this description of the data series, what picture do we form of the data? The easiest way to visualize data is to look at its **"distribution".**

- In the next slides we will learn more about the distribution.

Frequency Distribution

A distribution is a visualization of a frequency distribution table:

• Frequency distribution is a representation, either in a graphical or tabular format, that displays the number of observations within a given interval.

•In Frequency distribution, we find the number of counts for a particular observation when the observations are repeated.

Example:

Class(Rs.)	Frequency Students
20-30	5
30-40	8
40-50	9
50-60	10
60-70	6
70-80	2
Total	40

Frequency Distribution

- Steps to find the frequency distribution
  - ✓ Find the range for the given data (Largest Number – Smallest Number)
  - ✓ Determine the width of the class interval

$$\text{Width of the class interval} = \frac{\text{Range}}{\text{Number of class interval}}$$

Types of Frequency

- There are two types of Frequency mentioned below:

✓ **Relative Frequency:**

- To compute relative frequency, one obtains a frequency count for the total population and a frequency count for a subgroup or class interval of the population. .

  Relative Frequency = Frequency of Class interval / Total Observations or Total count.

✓ **Cumulative Frequency:**

- Cumulative frequency for each class interval is the frequency for that class interval added to the preceding cumulative total.

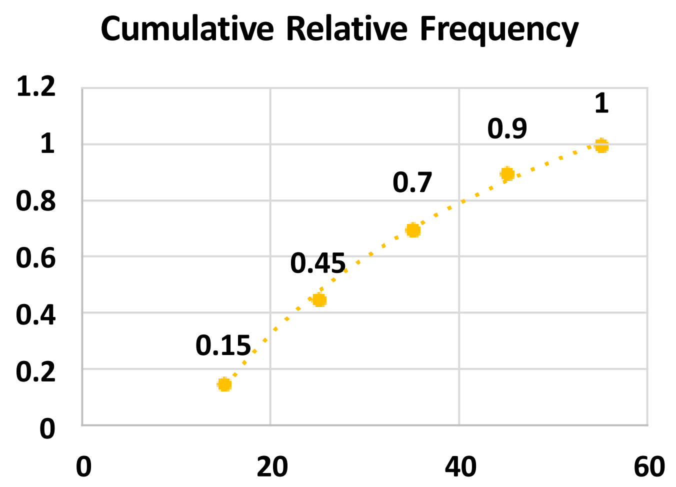
Class(Rs.)	Frequency Students	Relative Frequency	Cumulative Frequency
20-30	5	0.125	0.125
30-40	8	0.2	0.325
40-50	9	0.225	0.55
50-60	10	0.25	0.8
60-70	6	0.15	0.95
70-80	2	0.05	1
Total	40		

Types of Graph

Ogives

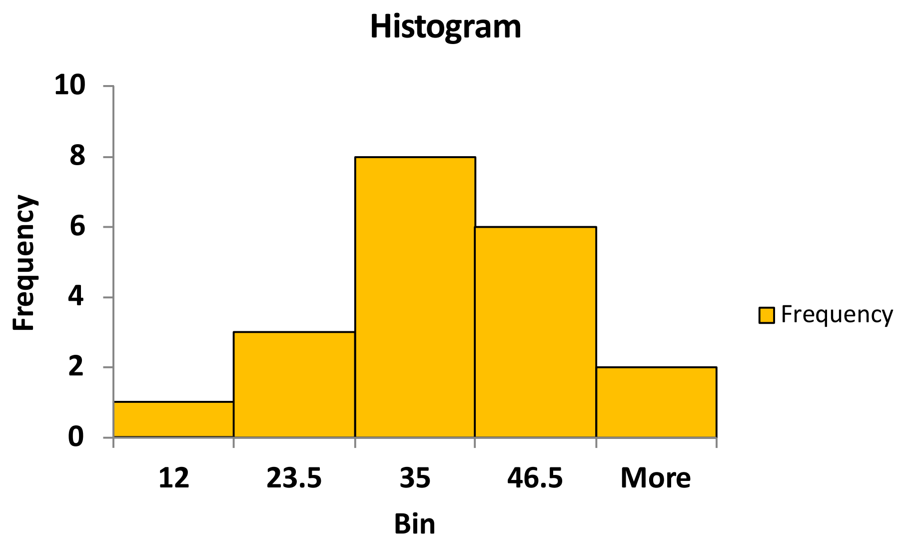
- An ogive is a graph of the cumulative relative frequency from a relative frequency distribution
- Ogives are sometime shown in the same graph as a relative frequency histogram
- Example: 12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58
- Add a cumulative relative frequency column:

Frequency Distribution					
Class	Frequency	Relative Frequency	Class Midpoint	Cumulative Frequency	Cumulative Relative Frequency
10 under 20	3	0.15	15	3	0.15
20 under 30	6	0.3	25	9	0.45
30 under 40	5	0.25	35	14	0.7
40 under 50	4	0.2	45	18	0.9
50 under 60	2	0.1	55	20	1
Total	20	1			



Histogram

- A histogram is a display of statistical information that uses rectangles to show the frequency of data items in successive numerical intervals of equal size.
- The **classes** or **intervals** are shown on the horizontal axis while the **frequency** is measured on the vertical axis.
- Bars of the appropriate heights can be used to represent the number of observations within each class
- Such a graph is called a **histogram**
- Example:
  Data: 12, 13, 17, 21, 24, 24, 26, 27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46, 53, 58

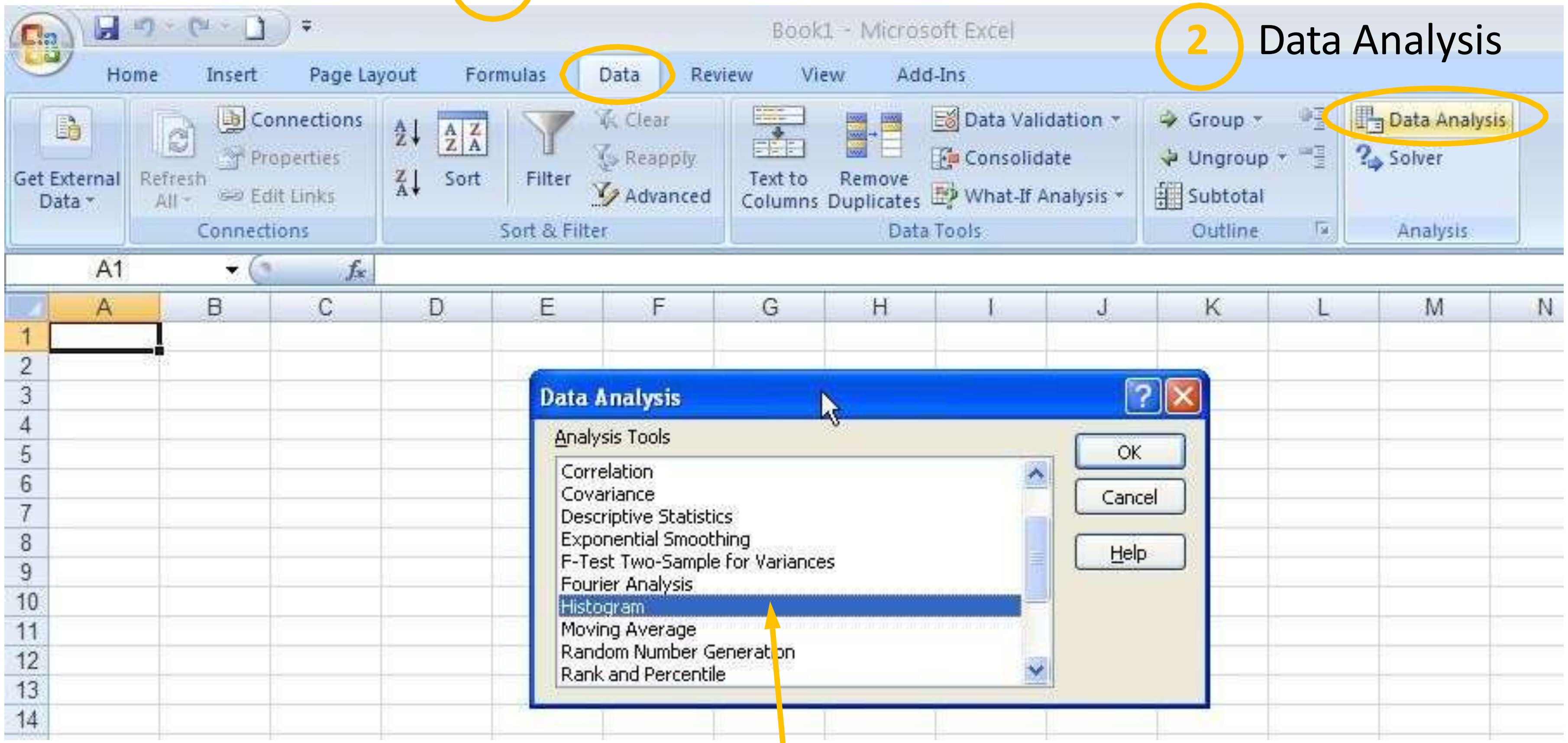


Histogram in Excel

proschool
An IMS Initiative

1 Select “Data” Tab

2 Data Analysis



3 Choose Histogram

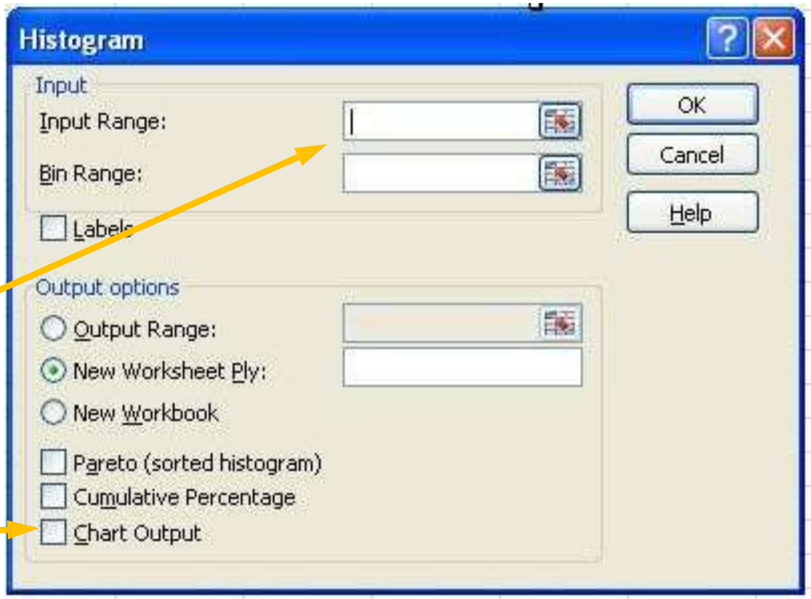
Histogram in Excel (Continued)

- Bin is also known as class interval
- In Excel, we can decide whether to give a bin range or not
- Excel automatically takes the bin range once data is provided

4

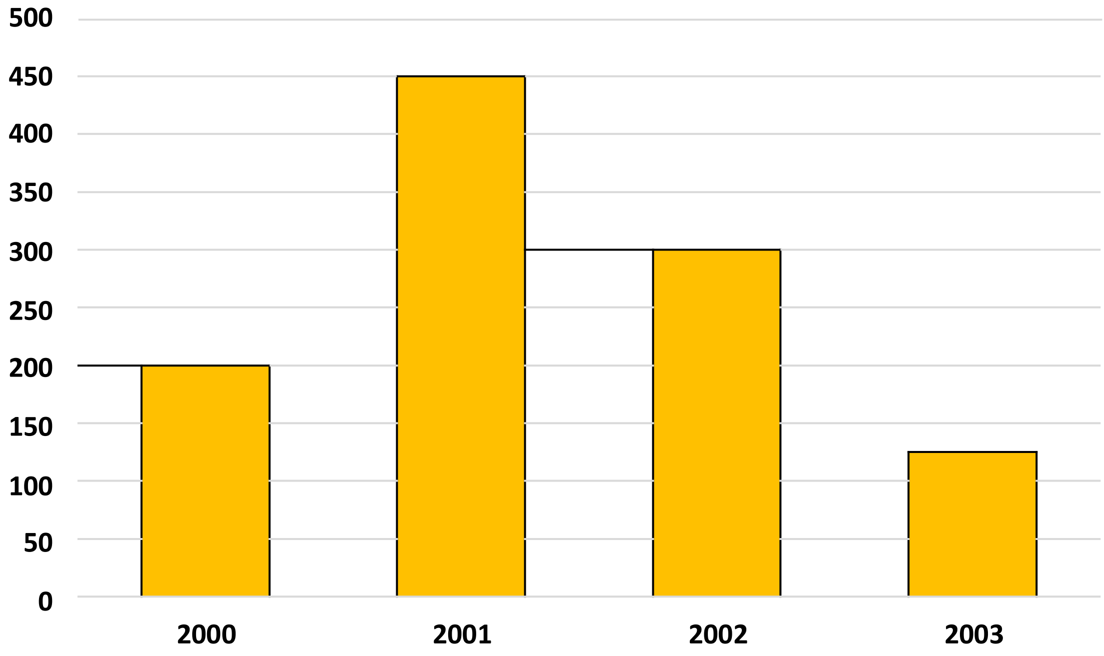
Input data and bin ranges

Select Chart Output



Bar Chart

- Bar charts is often used for qualitative (category) data
- Example:



- (Note that bar charts can also be displayed with horizontal bars)

Pie Chart

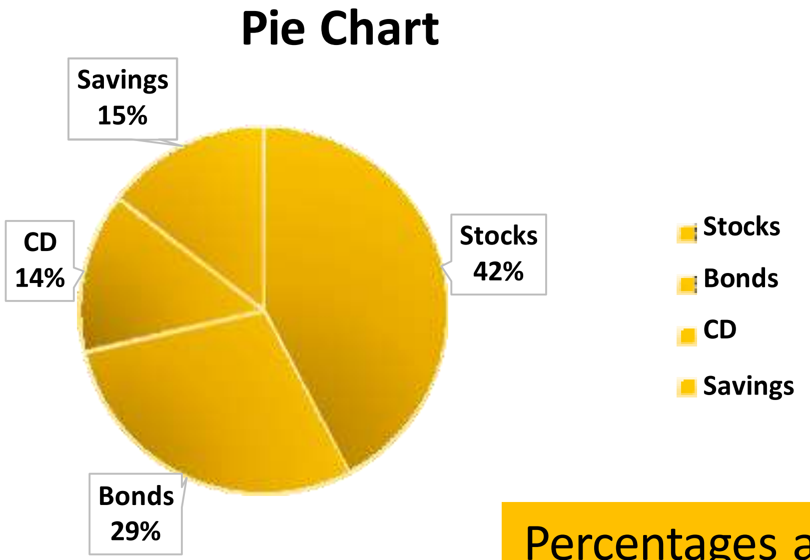
proschool
An IMS Initiative

- A pie chart is a circular statistical graphic which is divided into slices to illustrate numerical proportion.
- In a pie chart, the arc length of each slice (and consequently its central angle and area), is proportional to the quantity it represents.
- Size of pie slice shows the frequency or percentage for each category
- Example:

Current Investment Portfolio

Investment Type	Amount (in thousands $)	Percentage
Stocks	46.5	42.27
Bonds	32	29.09
CD	15.5	14.09
Savings	16	14.55
Total	110	100

(Variables are Qualitative)



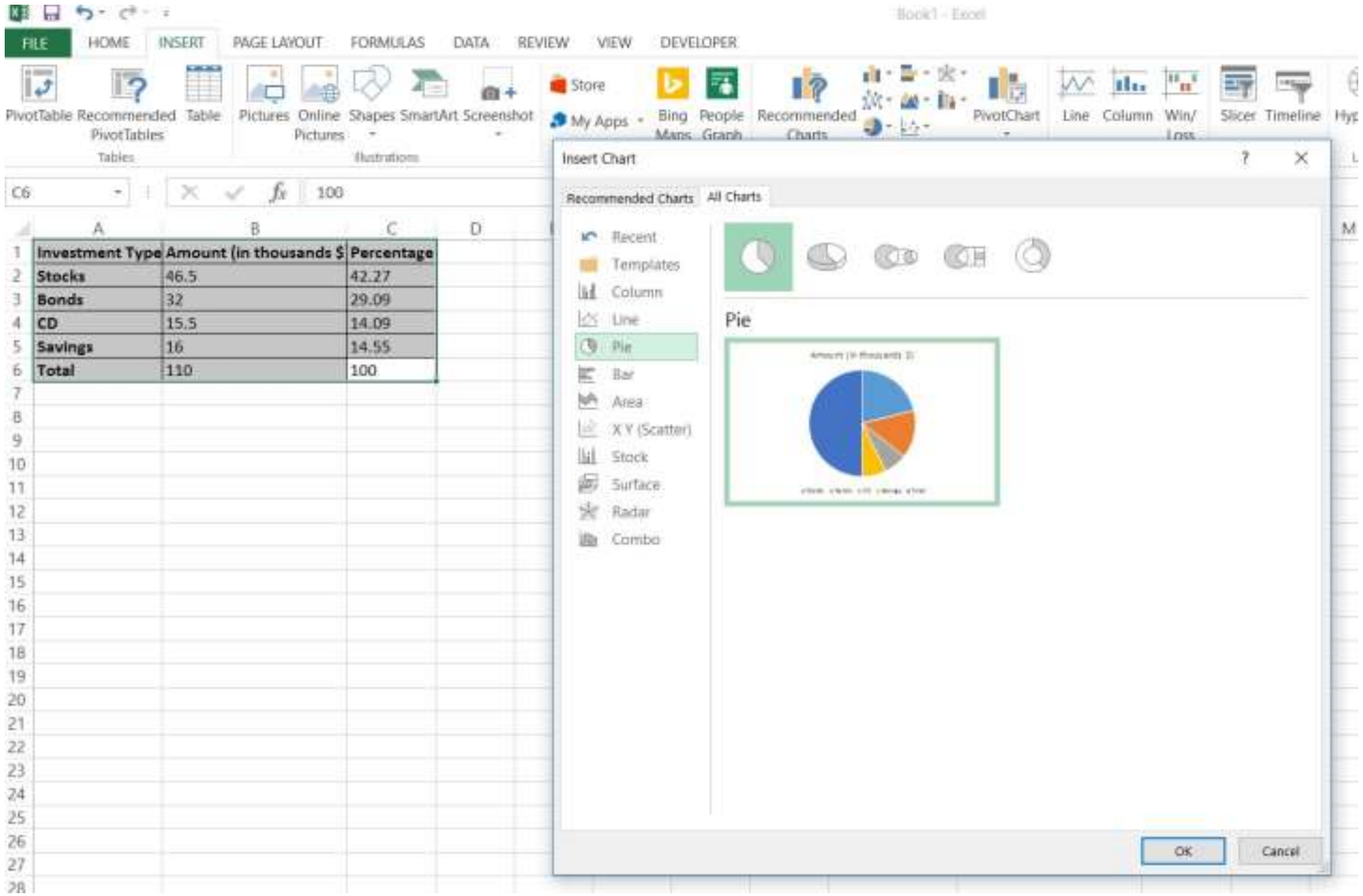
Percentages are rounded to the nearest percent

Steps to Draw a bar chart and pie chart in excel

proschool
An IMS Initiative

1. Select the data.
2. Go to insert and click on recommended chart
3. Go to all chart and select the required chart

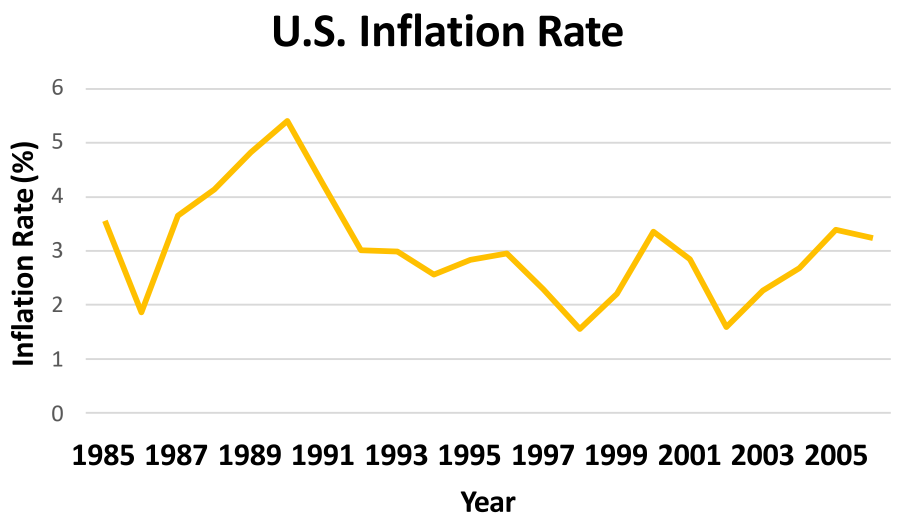
For bar chart same process is applied



Line Chart

proschool
An IMS Initiative

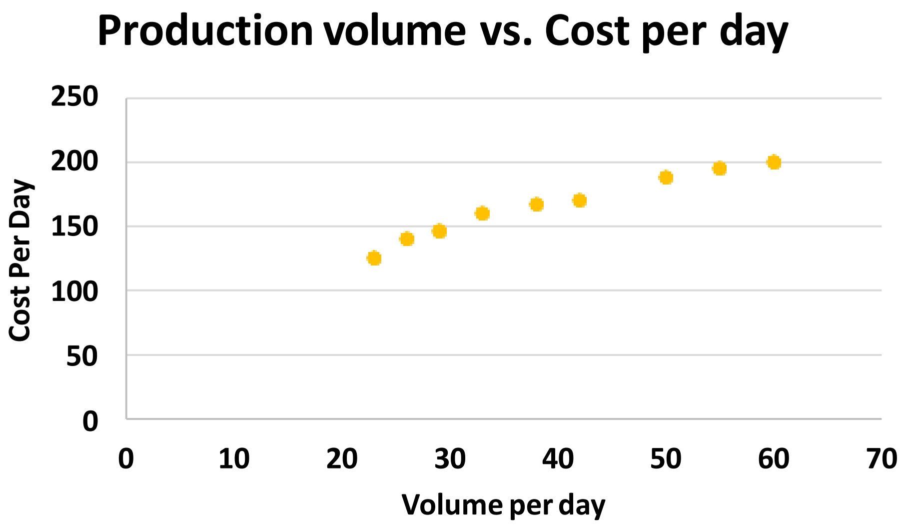
- A line chart or line graph is a type of chart which displays information as a series of data points called 'markers' connected by straight line segments..
- Line charts show values of one variable vs. time
- Time is traditionally shown on the horizontal axis



Year	Inflation Rate
1985	3.56
1986	1.86
1987	3.65
1988	4.14
1989	4.82
1990	5.4
1991	4.21
1992	3.01
1993	2.99
1994	2.56
1995	2.83
1996	2.95
1997	2.29
1998	1.56
1999	2.21
2000	3.36
2001	2.85
2002	1.59
2003	2.27
2004	2.68
2005	3.39
2006	3.24

Scatter Plot

- Scatter Diagrams show points for bivariate data. One variable is measured on the vertical axis and the other variable is measured on the horizontal axis.
- **Purpose:** Scatter plots shows the relationship between two variables.



Volume per day	Cost per day
23	125
26	140
29	146
33	160
38	167
42	170
50	188
55	195
60	200

Measures of Dispersion

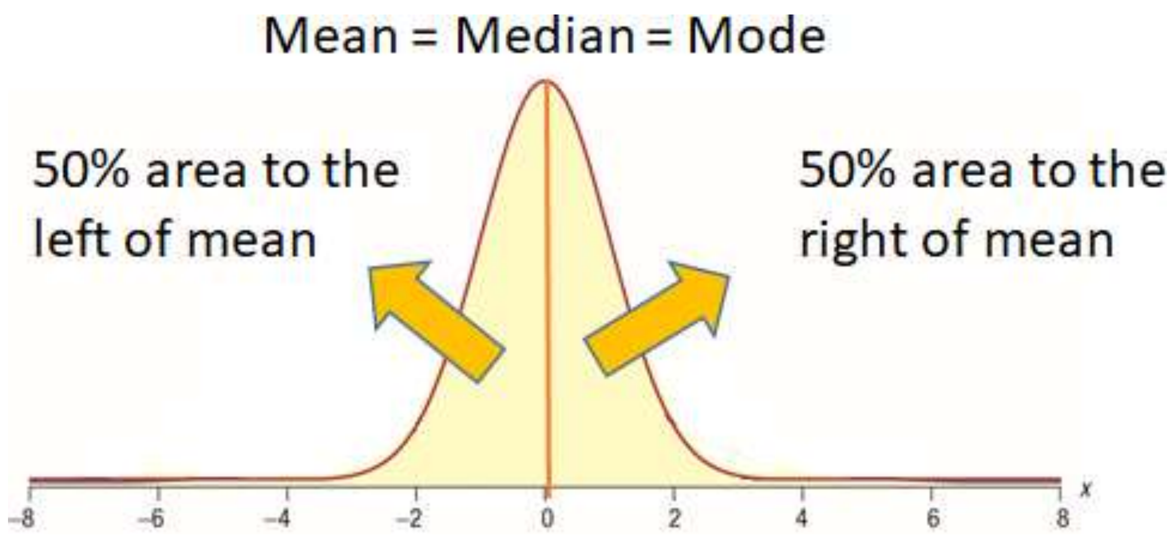
- Dispersion is the extent to which a distribution is stretched or squeezed
- Summary statistics can also be used to understand variation or dispersion in the data

✓ Range
✓ Inter-Quartile Range
✓ Variance
✓ Standard Deviation

Normal Distribution(Bell Shaped Curve)

- A normal distribution is the distribution in which most values cluster in the middle of the range and the rest taper off symmetrically toward either extreme.

- Height is one simple example of something that follows a normal distribution pattern: Most people are of average height, the number of people that are taller and shorter than average are fairly equal and a very small (and still roughly equivalent) number of people are either extremely tall or extremely short.



Range

Range is the difference between a highest and a lowest observation
Range = Highest observation - lowest observation
Example: In **{2, 3, 4, 6, 9, 3, 7, 16, 21 }** the lowest value is 2, and the highest is 21

Range: 21 – 2 = 19

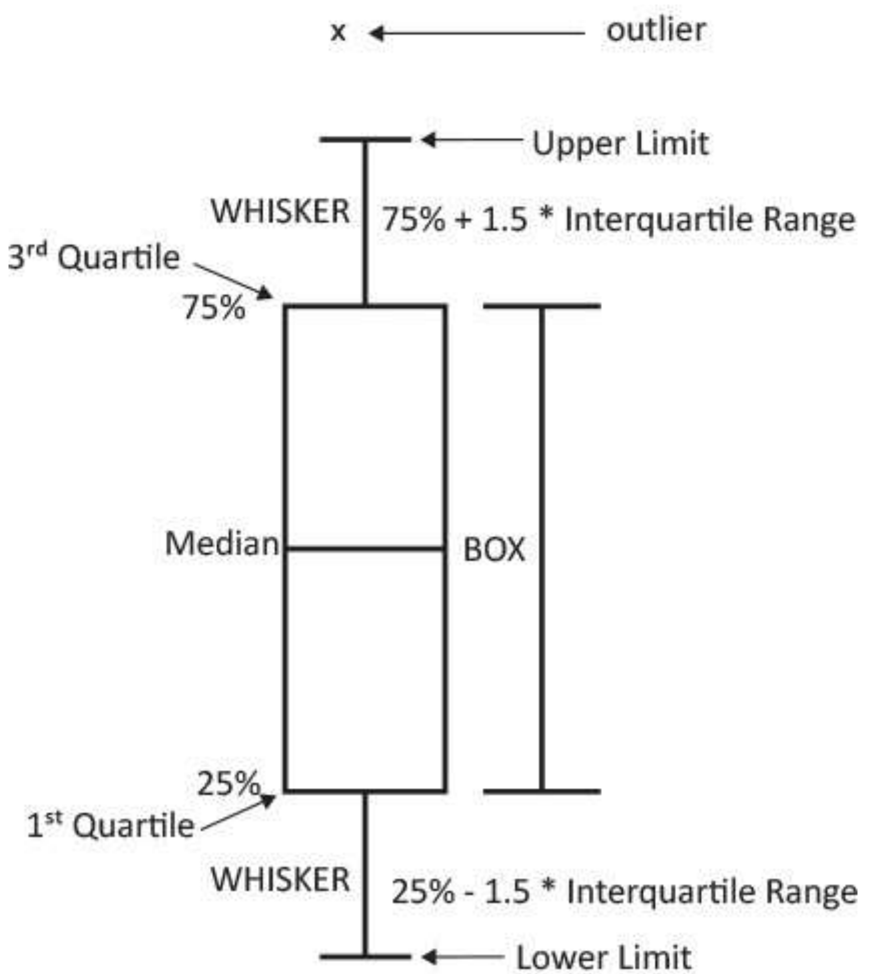
The range can sometimes be misleading when there are extremely high or low values.

Example: In **{8, 11, 5, 9, 7, 6, 19, 58, 45, 90, 4001}**:
the lowest value is 5,
and the highest is 4001,
So the range is 4001 -8 = 3993
The single value of 4001 makes the range large, but most values are below 100.

So we may be better using Box Plot and Standard Deviation

Box and Whisker Plot

- **Box-and-whisker plots** are a handy way to display data broken into four quartiles, each with an equal number of data values. It shows where the middle of the data lies. It's a nice plot to use when analyzing how your data is skewed.
- The median is the middle value of the data where half of the points are above and half are below this value.
- The first quartile represents the point where 25% of the data is below it.
- The third quartile represents the point where 75% of the data is below it.
- The whisker extends up to the highest value of upper limit and down to the lowest value of the lower limit.
- The lowest point of the lower whisker is called the lower limit. It equals Q1 – 1.5 * (Q3-Q1 or interquartile range).
- The highest point of the upper whisker is the called the upper limit. It equals Q3 + 1.5 * (Q3-Q1).
- Outliers are points that fall outside the limits of the whiskers.
- The interquartile is represented by the distance between Q1 and Q3.



Prepare a Box and Whisker Plot

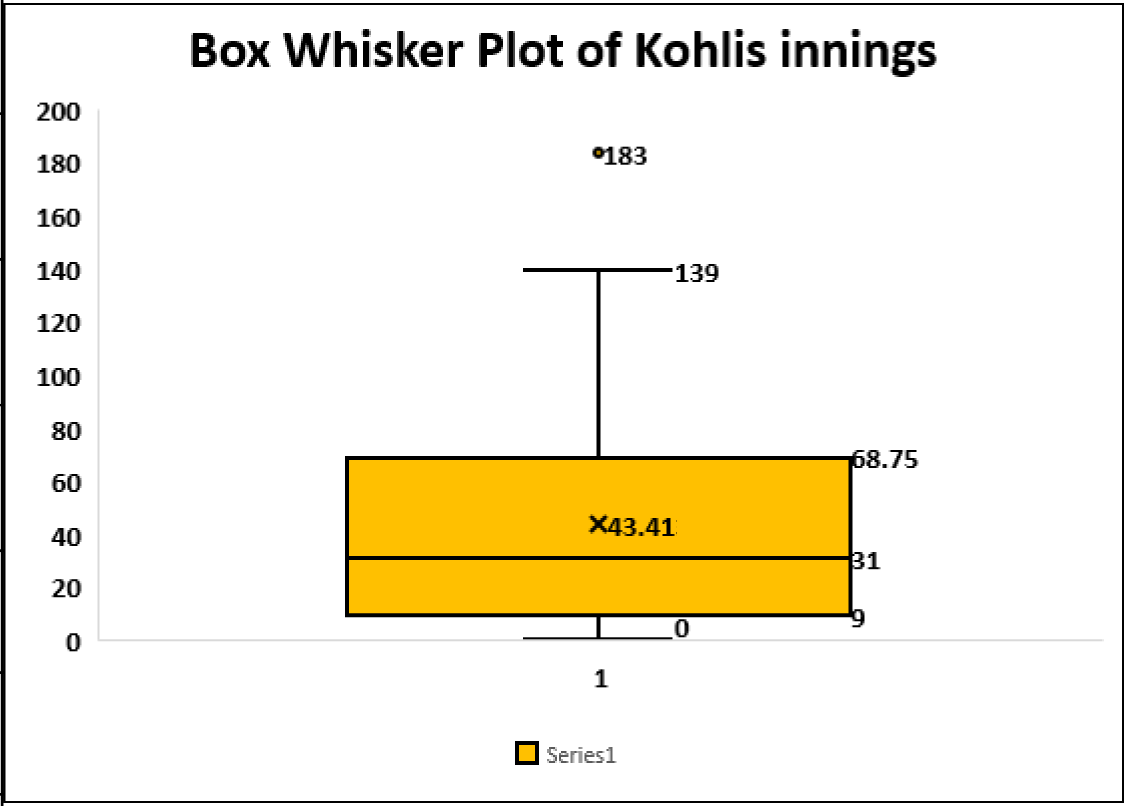
The following are the runs scored by Virat Kohli in 150 innings of One -day international matches

23	66	2	14	37	66	37	1	118	102
1	53	6	115	15	12	107	12	37	71
1	49	78	2	7	15	16	34	8	91
3	22	123	31	6	18	7	8	0	9
38	127	0	102	0	77	9	100	28	107
44	62	31	2	23	31	55	2	10	54
33	2	19	11	128	80	94	87	18	27
33	13	99	43	38	23	22	28	11	10
46	1	86	58	1	20	0	22	68	30
107	40	0	22	106	117	81	54	18	79
8	0	115	22	183	3	2	2	82	16
3	5	68	31	66	0	35	0	0	2
4	48	100	0	108	86	9	63	57	31
9	136	61	26	133	35	24	64	31	54
139	82	68	77	21	112	59	105	2	25

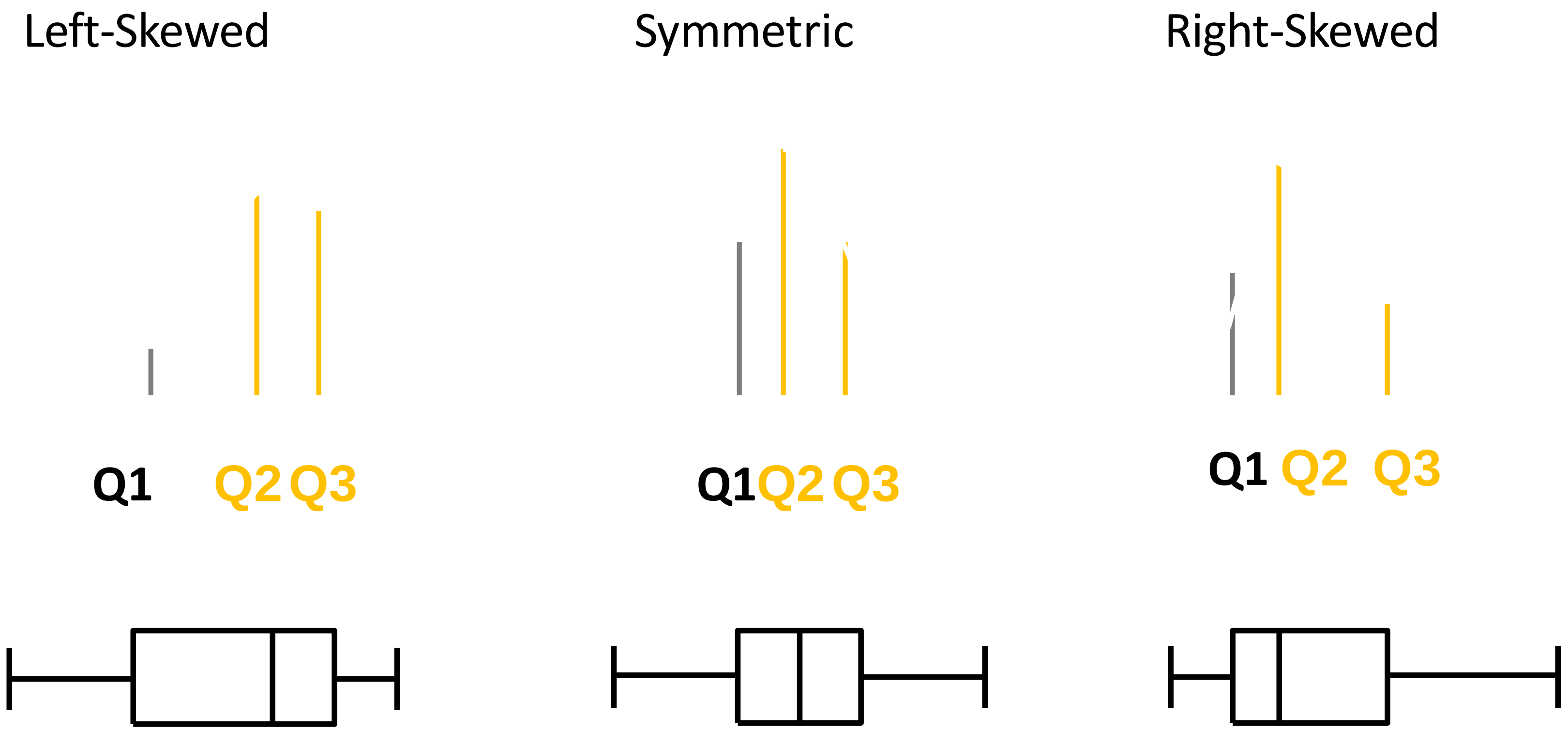
Prepare a Box and whisker plot using “quartile” function in Excel

Box & Whisker Plot of Kohli's innings

Particular	Observations	Formula	Remark
Minimum Value	0	QUARTILE(A$4:A$153,Minimum)	Minimum score in all the innings
1st Quartile	9	QUARTILE(A$4:A$153,25th percentile)	Kohli has scored 9 or less than 9 in his 25% of innings
Median	31	QUARTILE(A$4:A$153,50th percentile)	Kohli has scored approx. 31 or less than 31 in his 50% of innings
3rd Quartile	68	QUARTILE(A$4:A$153,75th percentile)	Kohli has scored 68 or more than 68 in his 75% of innings
Maximum Value	139	QUARTILE(A$4:A$153,Max)	Maximum Scores in all the innings
Interquartile Range	59	68 - 9	
Upper Limit	157	68+ 1.5*59	
Lowe Limit	-80	9 – 1.5*59	
Outliers	183	Innings above 157	Only one score in one inning is more than upper limit



Distribution Shape and Box and Whisker Plot



Percentile

proschool
An IMS Initiative

- A percentile (or a centile) is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall.
- For example, the 20th percentile is the value (or score) below which 20% of the observations may be found.
- The p^{th} percentile in an ordered array of n values is the value in i^{th} position,

Where,

$$i = \frac{p}{100}(n)$$

If i is not an integer, round up to the next higher integer value

Example: Find the 60^{th} percentile in an ordered array of 19 values.

$$i = \frac{p}{100}(n) = \frac{60}{100}(19) = 11.4$$

So use value in the $i = 12^{th}$ position

Variance and Standard Deviation

A firm is starting a delivery service for a new client between 2 points. Since it is a new client, the firm wants to send more consistent delivery boy to deliver the product.

Delivery boy 1 (Time in minutes) – 12,13,17,21,24, 24, 26,27, 27, 30, 32, 35, 37, 38, 41, 43, 44, 46,53,60

Delivery boy 2 (Time in minutes)- 34, 14, 31, 59, 11, 50, 27, 33, 53, 34, 13, 13, 42, 29, 33, 42, 34, 33, 44, 21

The average time taken by both the delivery boys is same i.e. 32.5 minutes

How can the firm arrive on a conclusion?

To find out which delivery boy is more consistent, we need to first understand variance and standard deviation

Variance

- Variance is a measurement of the spread between numbers in a data set.
- It measures how far each number in the set is from the mean.
- If the data is a Sample (a selection taken from a bigger Population), then the calculation changes!
- When you have "N" data values:

✓ The Population: divide by N

✓ A Sample: divide by N-1

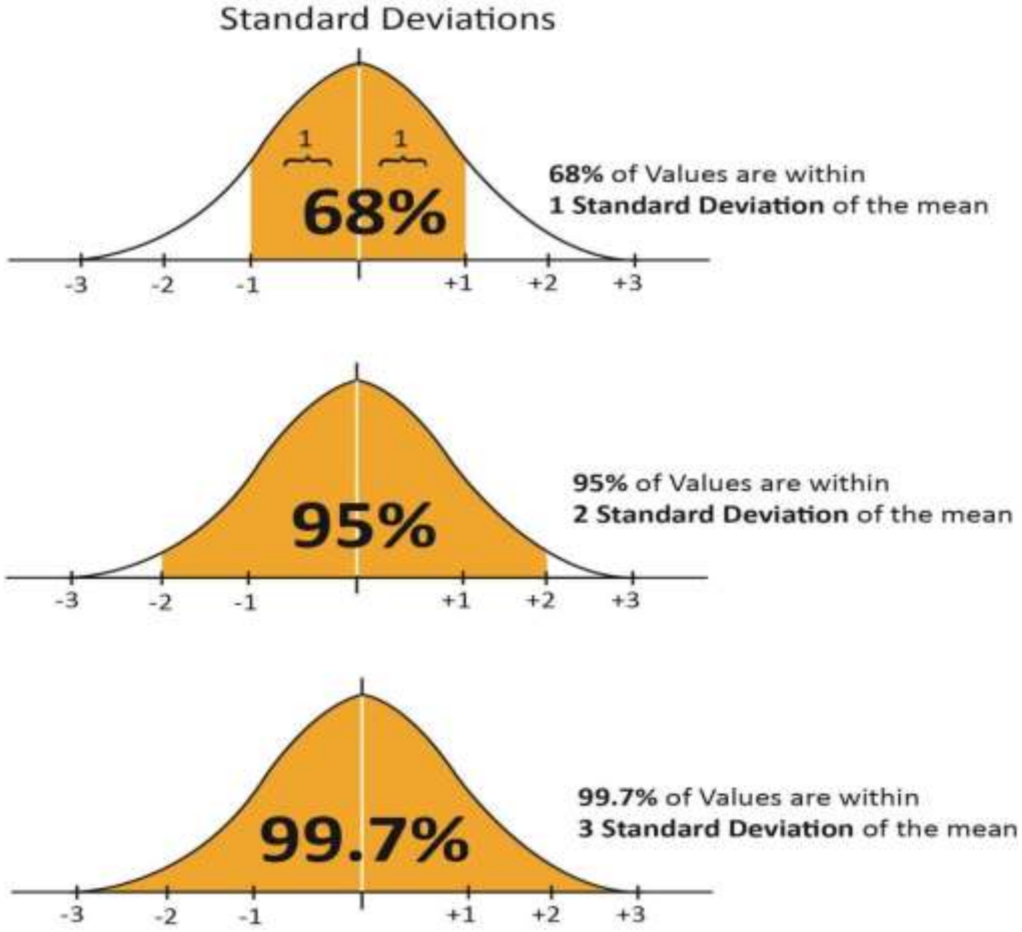
For Sample it is $$s^2 = \frac{1}{(N-1)} \sum_{i=1}^{N} (x_i - \bar{x})^2$$

For Population it is $$\sigma^2 = \frac{1}{N} \sum_{i=1} (x_i - \mu)^2$$

- To calculate the variance follow these steps:
  - Work out the Mean (Simple average of the numbers)
  - Then for each number: subtract the Mean and square the result (the squared difference).
  - Then work out the average of those squared differences.

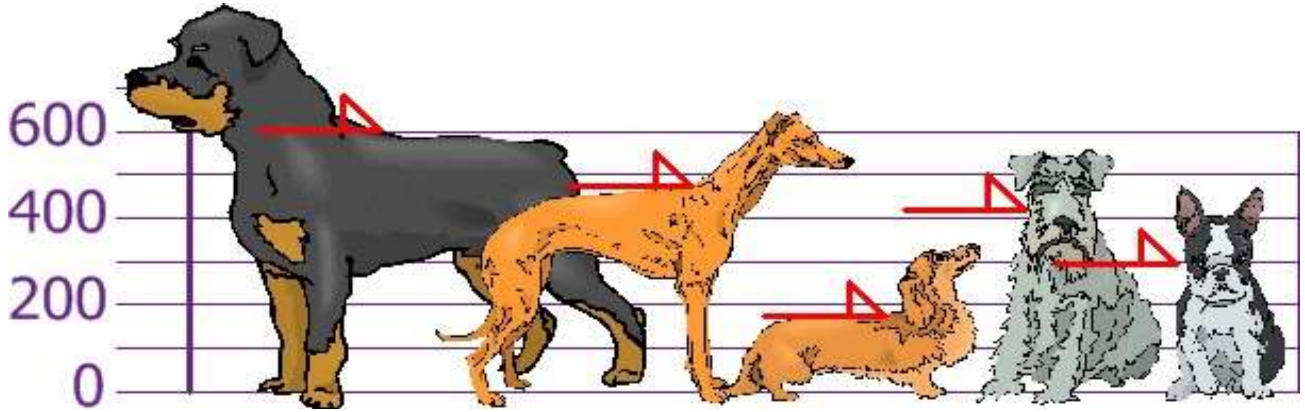
Standard Deviation

- **Standard deviation** is a measure of the dispersion of a set of data from its mean
- **Standard deviation** s (or σ) is just the square root of variance s^2 (or σ^2)
- When we calculate the standard deviation of normal distribution we find that (generally):



Variance: Example 1

- You and your friends have just measured the heights of your dogs (in millimeters):



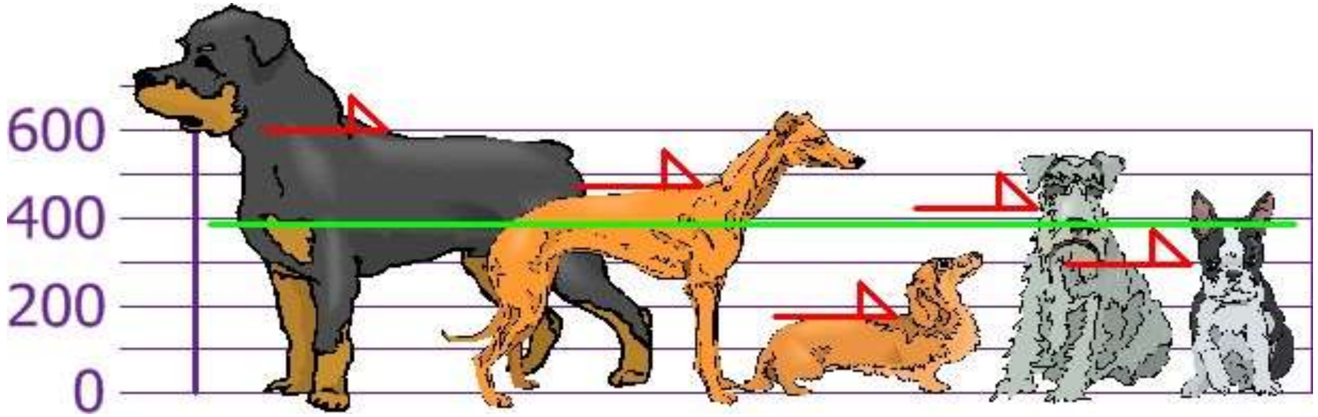
- The heights (at the shoulders) are: 600mm, 470mm, 170mm, 430mm and 300mm.

Variance: Example 1 (continue)

First step is to find the Mean:
Answer:

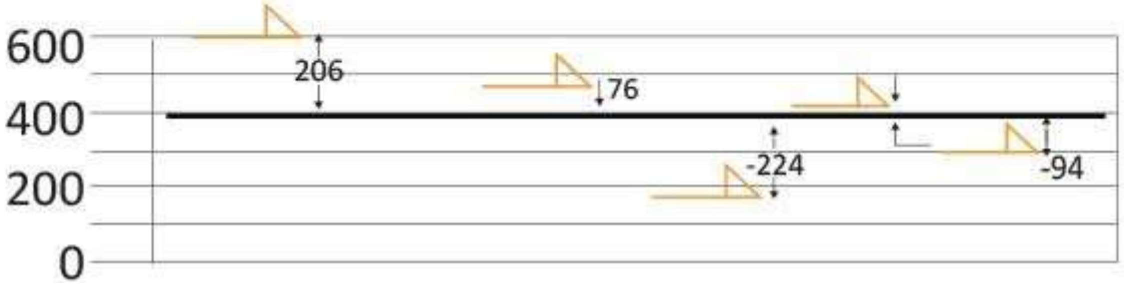
$$\text{Mean} = \frac{600 + 470 + 170 + 430 + 300}{5} = \frac{1970}{5} = 394$$

so the mean (average) height is 394 mm. Let's plot this on the chart:



Variance : Example 1 (continue)

- Now we calculate each dog's difference from the Mean:



- To calculate the Variance, take each difference, square it, and then average the result:

$$\text{Variance}: \ \sigma^2 = \frac{206^2 + 76^2 + (-224)^2 + 36^2 + (-94)^2}{5}$$

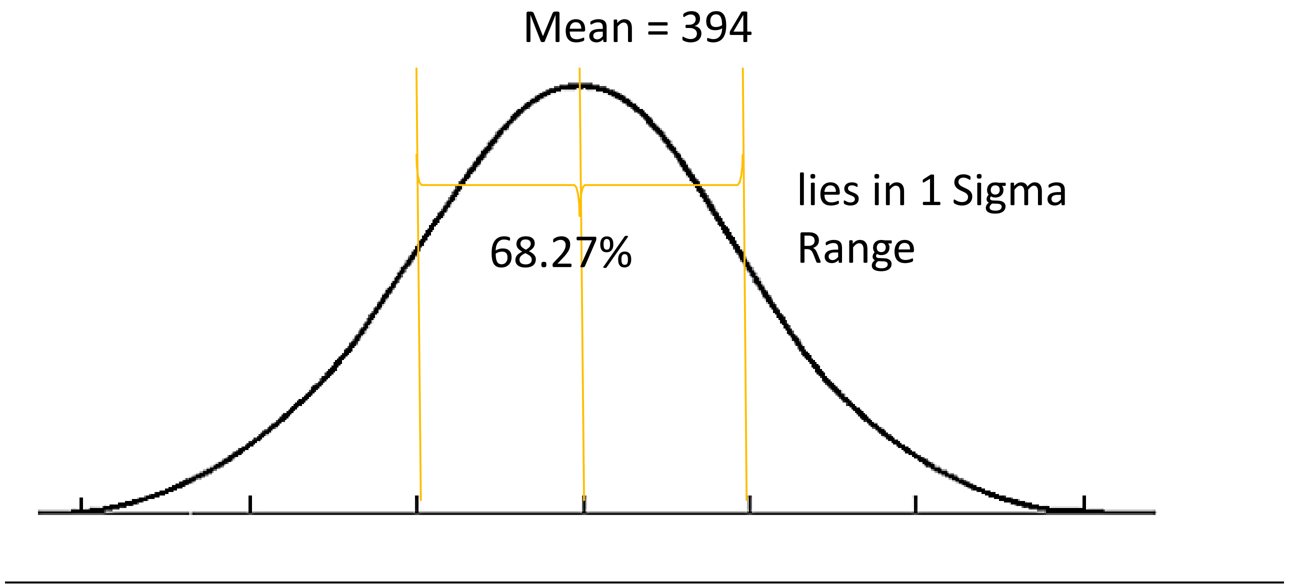
$$= \frac{42436+5776+50176+1296+8836}{5}$$

$= 21704$

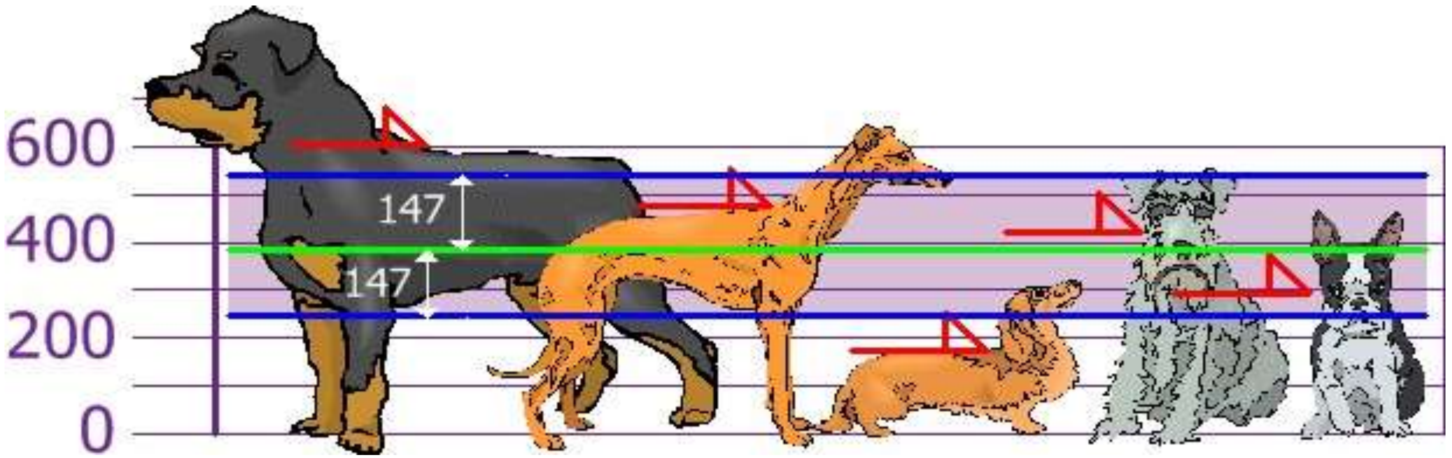
- The variance is 21,704

Standard Deviation

- Lets continue example 2, (where the variance was 21704)
- The Standard Deviation is just the square root of Variance, so:
  Standard Deviation: $\sigma = \sqrt{21,704}$
  $= 147.32... = 147$ (to the nearest mm)
- We can show which heights are within one Standard Deviation (147mm) of the Mean (394mm):



Standard Deviation



So, using the Standard Deviation we have a "standard" way of knowing what is normal, and what is extra large or extra small.
Rottweiler's are tall dogs and Dachshunds are a bit short ... but don't tell them!

You can find the Standard deviation by using excel formula: STDEV

Lets come back to Business Problem

Solution

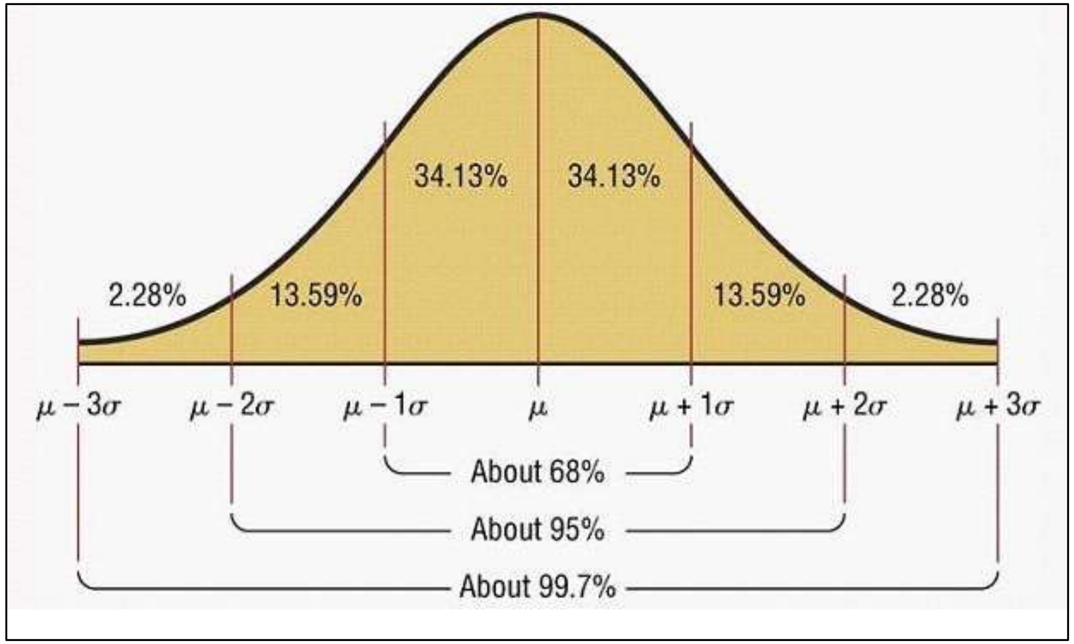
- Excel Formula =stdev(data)

- Standard deviation of the delivery time taken by delivery boy 1: 12.89
- Standard deviation of the delivery time taken by delivery boy 2: 13.55

From the above observations, we can conclude that Delivery boy 1 is more consistent than delivery boy 2. Hence , firm should send delivery boy 1

Empirical Rule

- Empirical rule can be applied for a symmetrical bell shaped frequency distribution
- Empirical rule is known as the three sigma rule
- The range within which approximate percentage of values of the distribution are likely to fall within a given number of standard deviation from the mean is determined below:
    - ✓ Approximately 68.26% of the data is within one standard deviation of the mean.
    - ✓ Approximately 95.44% of the data is within two standard deviations of the mean.
    - ✓ More than 99.72% of the data is within three standard deviations of the mean.



Application of Empirical rule with previous business problem

Delivery boy 1
Mean: 32.5
Standard Deviation: 12.89

Empirical rule 1: Approximately 68.26% of the data is within one standard deviation of the mean = (12.89+32.5) and (32.5 – 12.89) or 45 and 19.

In our dataset, we have got 14 data points (approx. 70% of 20) between 45 and 19.6

Empirical rule 2: Approximately 95.44% of the data is within two standard deviation of the mean = (12.89*2)+32.5 and 32.5 – (12.89*2) or 58 and 6.7

In our dataset, we have got 19 data points (approx. 95% of 20) between 58 and 6.7

Empirical rule 3: Approximately 99.72% of the data is within three standard deviation of the mean = (12.89*3)+32.5 and 32.5 – (12.89*3) or 71 and -6

In our dataset, we have got 20 values (approx. 100% of 20) between 71 and -6

12			
13			
17			
21			
24			
24			
26			
27			
27			
30		2 SD	
32	1 SD		3 SD
35			
37			
38			
41			
43			
44			
46			
53			
60			

Limitations of Empirical Rule

- Limitations of Empirical Rule:
  - ✓ Empirical rule applies only to normally distributed data
  - ✓ It has a wide range of applications, but in cases where distribution is not normal or the shape of distribution is not known, its application is restricted

- Chebyshev's inequality is the best alternative to empirical rule.

Application and Uses of Standard deviation

CHEBYSHEV'S INEQUALITY

For any data set, it can be proved mathematically that –

- At least 75% of all data points will lie within 2 standard deviations of the mean,
- At least 89% of all data points will lie within 3 standard deviations of the mean.
- At least 95% of the data is within 4.5 standard deviations of the mean.

Application of Standard deviation

1) Standard deviation used as a measure or risk

Example:

You are trying to pick stock for investing in the equity market.

- Stock A has an annual returnof 15%, with a standard deviation of 30%
- Stock B has an annual return of 12%, with a standard deviation of 8%
- If you were risk averse, which would you choose?

Measures of Shape

- Measures of shape describes the distribution or pattern of the data in a set

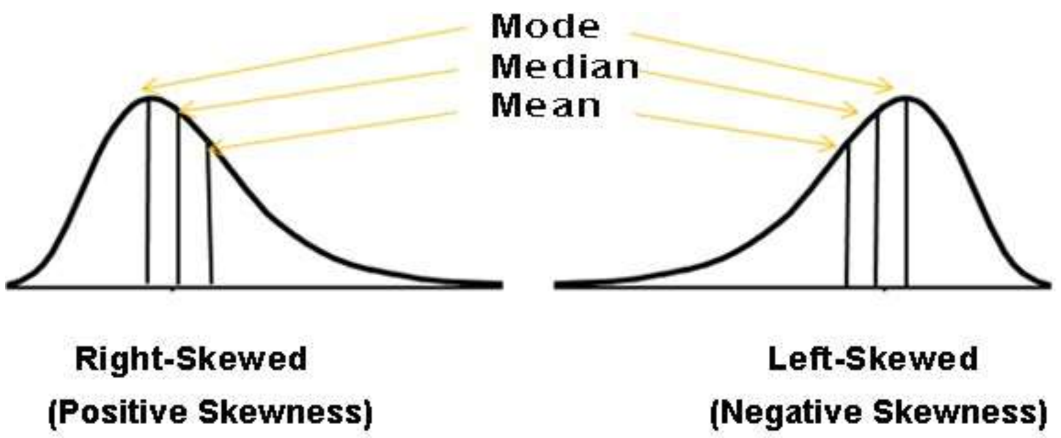
- The distribution shape of the quantitative data can be described as there is a logical order to the values and the low and high end values on the horizontal axis of the histogram

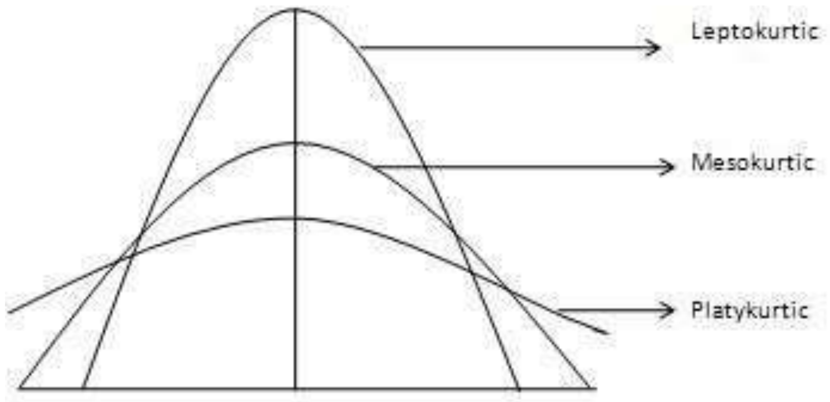
- The distribution shape of the qualitative data cannot be described.

Measures of shape are as follows:

✓ Degree of Skewness

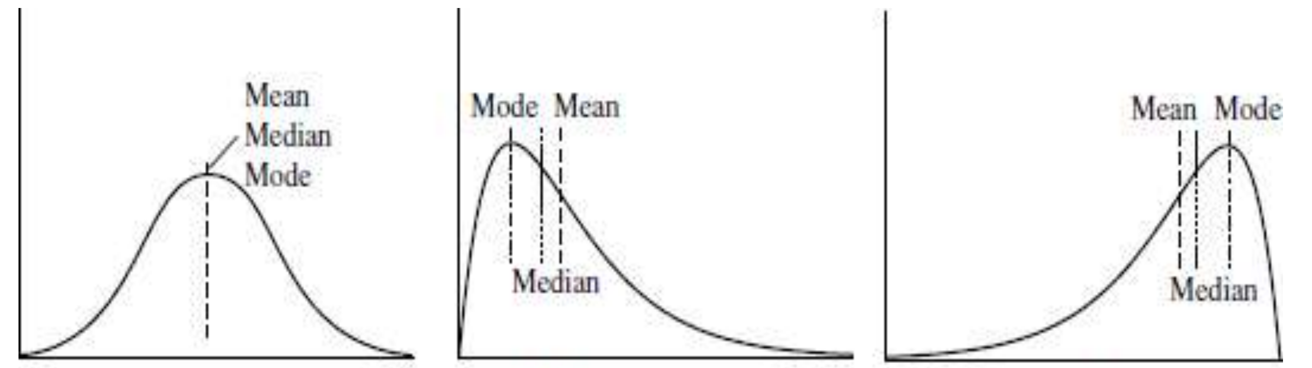
✓ Kurtosis





Degree of Skewness

- Skewness is the tendency for the values to be more frequent around the high or low ends of the x axis
- Skewness is a measure of symmetry
- **Symmetric data –** The data is symmetrically distributed on both side of medium
  - ✓ *mean = median = mode*
- **Positively skewed** -
  - ✓ Tail on the right side is longer than the left side.
  - ✓ *mode < median < mean*
- **Negatively skewed** -
  - ✓ Tail on the left side is longer than the right side.
  - ✓ *mode > median > mean*

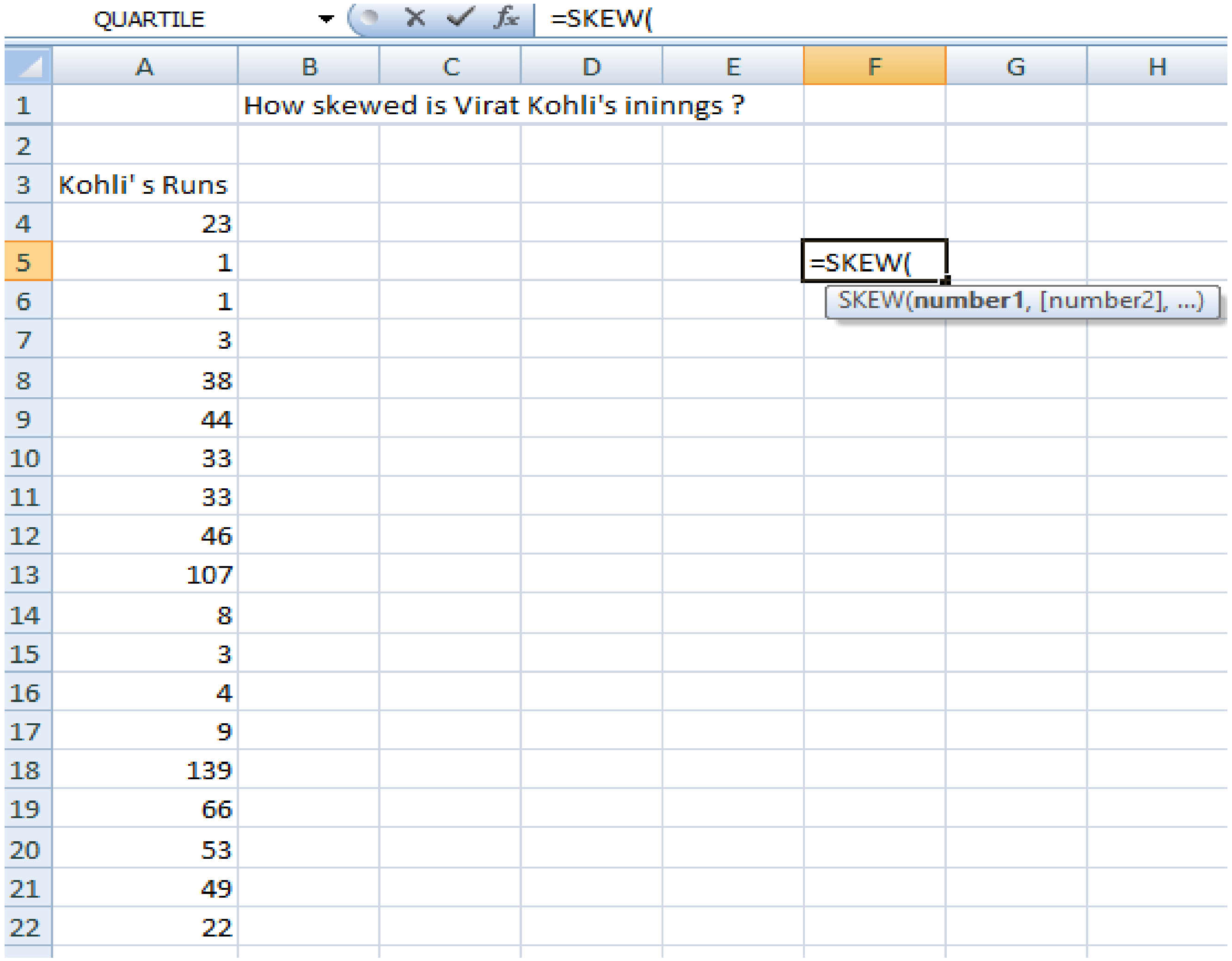


(a) Symmetric data (b) Positively skewed data (c) Negatively skewed data

Skewness Example

How skewed is Virat Kohli's innings ?

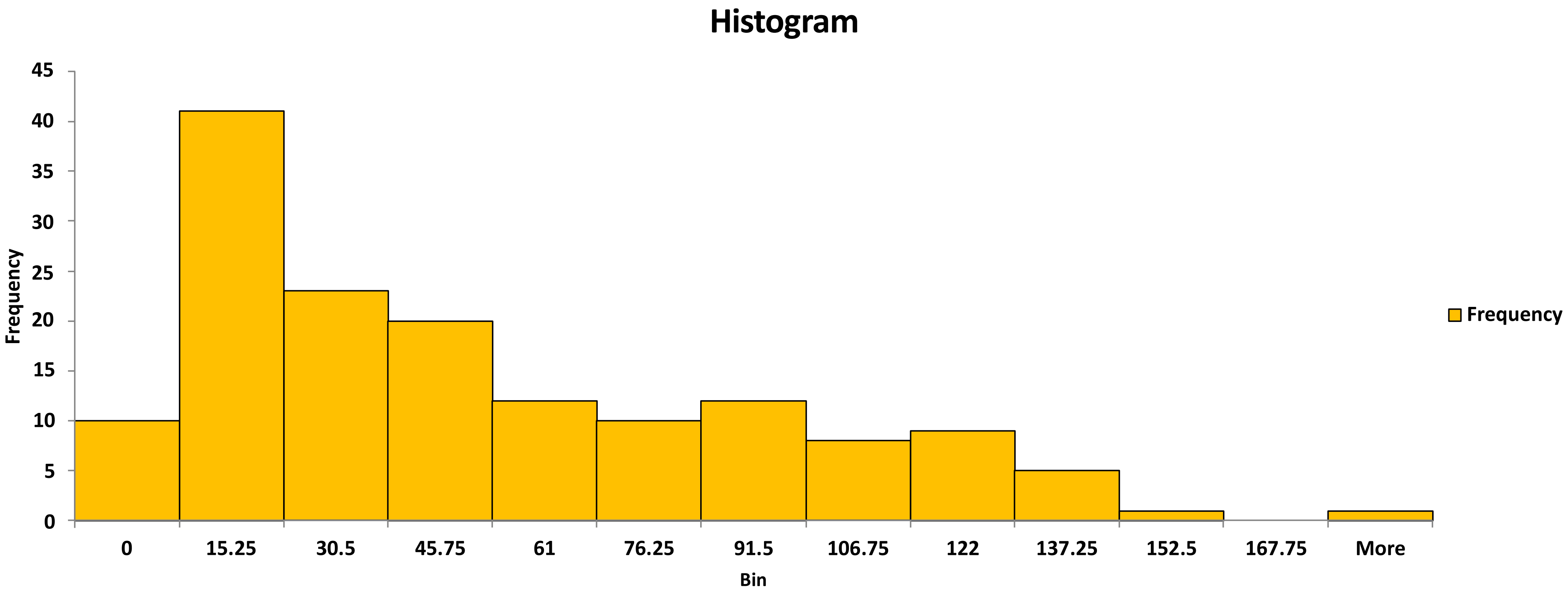
23, 1, 1, 3, 38, 44, 33, 33, 46, 107, 8, 3, 4, 9, 139, 66, 53, 49, 22, 127, 62, 2, 13, 1, 40, 0, 5, 48, 136, 82, 2, 6, 78, 123, 0, 31, 19, 99, 86, 0, 115, 68, 100, 61, 68, 14, 115, 2, 31, 102, 2, 11, 43, 58, 22, 22, 31, 0, 26, 77, 37, 15, 7, 6, 0, 23, 128, 38, 1, 106, 183, 66, 108, 133, 21, 66, 12, 15, 18, 77, 31, 80, 23, 20, 117, 3, 0, 86, 35, 112, 37, 107, 16, 7, 9, 55, 94, 22, 0, 81, 2, 35, 9, 24, 59, 1, 12, 34, 8, 100, 2, 87, 28, 22, 54, 2, 0, 63, 64, 105, 118,37, 8, 0, 28, 10, 18, 11, 68, 18, 82, 0, 57, 31, 2, 102, 71, 91, 9, 107, 54, 27, 10, 30, 79, 16, 2, 31, 54, 25, 37, 12.



Skewness Interpretation

- If skewness is less than -1 or greater than 1, the distribution is highly skewed.
- If skewness is between -1 and -0.5 or between 0.5 and 1, the distribution is moderately skewed.
- If skewness is between -0.5 and 0.5, the distribution is approximately symmetric, close to Normal Distribution.

Skewness Interpretation



The Skewness is 0.91, Mean is 43.17 and Median is 31 which indicates that the data is Positively skewed.

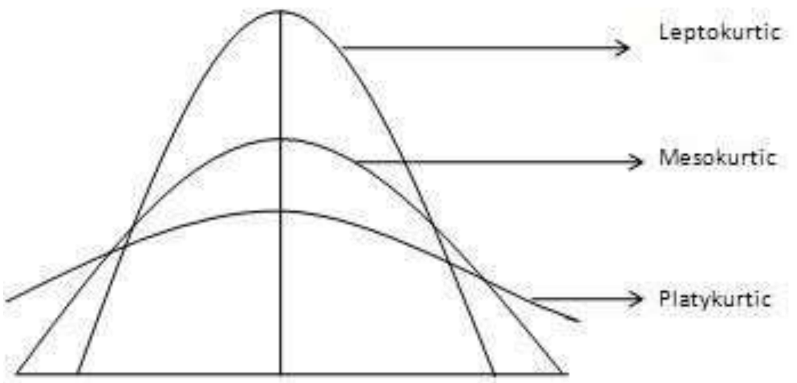
Kurtosis

- Kurtosis is the sharpness of the peak of a frequency-distribution curve.
- It describes the shape of the distribution of the tail's in relation to its shape

Types of Kurtosis

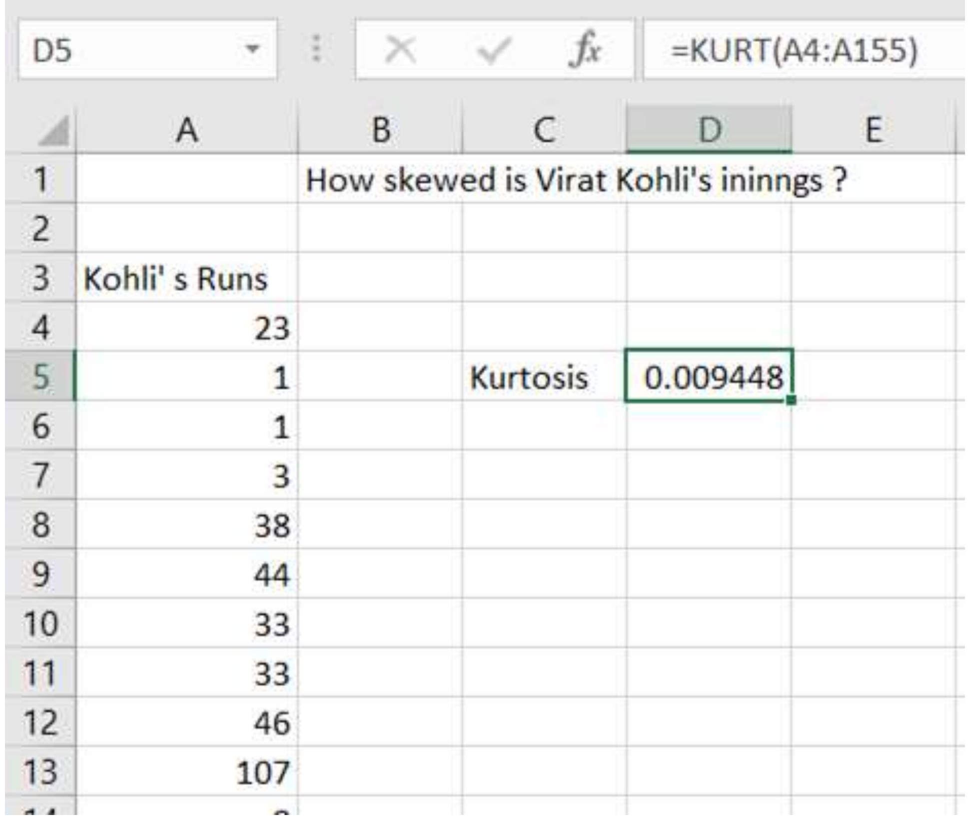
✓ Mesokurtic – It has flatter tail than standard normal distribution and slightly lower peak
✓ Leptokurtic – It has extremely thick tail and a very thin and tall peak
✓ Platykurtic – It has slender tail and a peak that's smaller than Mesokurtic distribution

Kurtosis - Measure of the relative peak of a distribution.
K = 3 indicates a normal "bell-shaped" distribution (mesokurtic).
K < 3 indicates a platykurtic distribution (flatter than a normal distribution with shorter tails).
K > 3 indicates a leptokurtic distribution (more peaked than a normal distribution with longer tails).

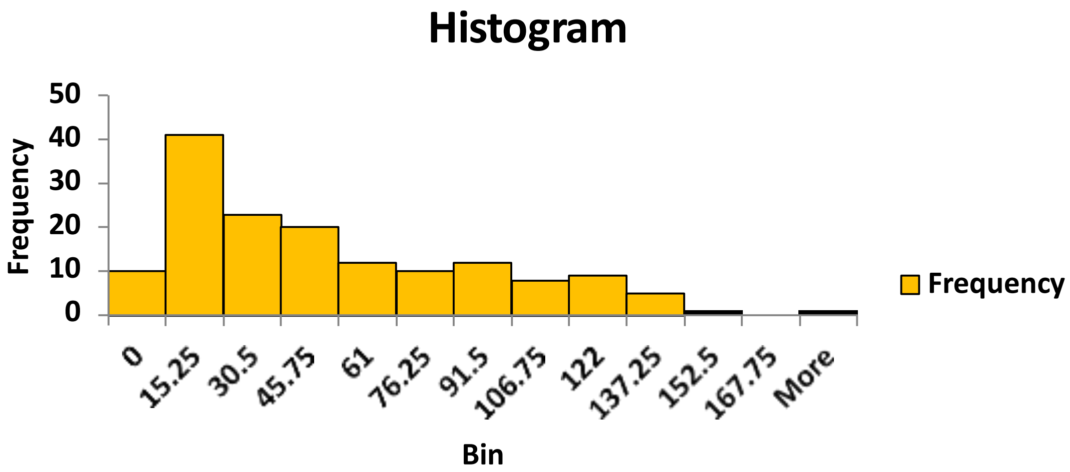


Kurtosis

- Lets consider Kohli’s inning case, the kurtosis is 0.0094. Hence it is Platykurtic as its values is less than 3.



	A	B	C	D	E
1		How skewed is Virat Kohli's ininngs ?			
2					
3	Kohli' s Runs				
4	23				
5	1		Kurtosis	0.009448	
6	1				
7	3				
8	38				
9	44				
10	33				
11	33				
12	46				
13	107				



- Note: There might be certain differences in values when calculate on R.

Descriptive statistic Summary Dataset

Lets find out the descriptive statistics of the students score in a exam from different cities in US

Student ID	State	Gender	Student Status	Country	Student Status	Major	Age	Height (in)	Study hrs	Exam score out of 40
1	California	Female	Graduate	US	Graduate	Politics	30	61	4	30
2	Arizona	Female	Undergraduate	US	Undergraduate	Politics	19	64	2	19
3	New York	Male	Graduate	US	Graduate	Math	26	73	6	26
4	New York	Male	Graduate	US	Graduate	Econ	33	68	3	33
5	Ohio	Male	Graduate	US	Graduate	Econ	37	71	6	37
6	California	Male	Graduate	US	Graduate	Econ	25	67	5	25
7	North Carolina	Male	Graduate	US	Graduate	Politics	39	70	5	39
8	Kansas	Female	Undergraduate	US	Undergraduate	Politics	21	62	5	21
9	California	Female	Undergraduate	US	Undergraduate	Math	18	62	6	18
10	New York	Female	Graduate	US	Graduate	Math	33	66	5	33
11	Mississippi	Male	Undergraduate	US	Undergraduate	Econ	18	67	3	18
12	Virginia	Female	Graduate	US	Graduate	Math	38	59	5	38
13	California	Male	Graduate	US	Graduate	Politics	30	63	4	30
14	New York	Male	Graduate	US	Graduate	Politics	30	75	6	30
15	New York	Female	Undergraduate	US	Undergraduate	Math	21	64	5	21
16	Utah	Female	Graduate	US	Undergraduate	Politics	18	63	2	18
17	New York	Female	Undergraduate	US	Undergraduate	Math	19	60	2	19
18	Pennsylvania	Female	Graduate	US	Graduate	Politics	31	59	4	31
19	Oklahoma	Female	Undergraduate	US	Undergraduate	Math	18	68	4	18
20	New York	Male	Graduate	US	Graduate	Politics	33	63	7	33
21	Ohio	Female	Undergraduate	US	Undergraduate	Econ	19	62	5	19
22	New York	Male	Undergraduate	US	Undergraduate	Econ	21	73	4	21
23	Massachusetts	Female	Graduate	US	Graduate	Politics	25	68	6	25
24	Pennsylvania	Female	Undergraduate	US	Undergraduate	Math	18	65	6	18
25	Ohio	Male	Graduate	US	Undergraduate	Politics	19	64	4	19
26	Minnesota	Male	Graduate	US	Graduate	Econ	28	71	4	28
27	Pennsylvania	Male	Undergraduate	US	Undergraduate	Econ	20	71	5	20
28	Oklahoma	Female	Undergraduate	US	Undergraduate	Econ	20	68	6	20
29	Pennsylvania	Male	Graduate	US	Graduate	Politics	30	72	6	30
30	Ohio	Male	Undergraduate	US	Undergraduate	Econ	19	74	1	19

Descriptive statistic in Excel

proschool An IMS Initiative

Student ID	State	Gender	Student Status	Country	Student Status	Major	Age	Height (in)	Study hrs	Exam score out of 40
1	California	Female	Graduate	US	Graduate	Politics	30	61	4	30
2	Arizona	Female	Undergraduate	US	Undergraduate	Politics	19	64	2	19
3	New York	Male	Graduate	US	Graduate	Math	26	73	6	26
4	New York	Male	Graduate	US	Graduate	Econ	33	68	3	33
5	Ohio	Male	Graduate	US	Graduate	Econ	37	71	6	37
6	California	Male	Graduate	US	Graduate	Econ	25	67	5	25
7	North Carolina	Male	Graduate	US	Graduate	Politics	39	70	5	39
8	Kansas	Female	Undergraduate	US	Undergraduate	Politics	21	62	5	21
9	California	Female	Undergraduate	US	Undergraduate	Math	18	62	6	18
10	New York	Female	Graduate	US	Graduate	Math	33	66	5	33
11	Mississippi	Male	Undergraduate	US	Undergraduate	Econ	18	67	3	18
12	Virginia	Female	Graduate	US	Graduate	Math	38	59	5	38
13	California	Male	Graduate	US	Graduate	Politics	30	63	4	30
14	New York	Male	Graduate	US	Graduate	Politics	30	75	6	30
15	New York	Female	Undergraduate	US	Undergraduate	Math	21	64	5	21
16	Utah	Female	Graduate	US	Undergraduate	Politics	18	63	2	18
17	New York	Female	Undergraduate	US	Undergraduate	Math	19	60	2	19
18	Pennsylvania	Female	Graduate	US	Graduate	Politics	31	59	4	31
19	Oklahoma	Female	Undergraduate	US	Undergraduate	Math	18	68	4	18
20	New York	Male	Graduate	US	Graduate	Politics	33	63	7	33
21	Ohio	Female	Undergraduate	US	Undergraduate	Econ	19	62	5	19
22	New York	Male	Undergraduate	US	Undergraduate	Econ	21	73	4	21
23	Massachusetts	Female	Graduate	US	Graduate	Politics	25	68	6	25
24	Pennsylvania	Female	Undergraduate	US	Undergraduate	Math	18	65	6	18
25	Ohio	Male	Graduate	US	Undergraduate	Politics	19	64	4	19
26	Minnesota	Male	Graduate	US	Graduate	Econ	28	71	4	28
27	Pennsylvania	Male	Undergraduate	US	Undergraduate	Econ	20	71	5	20
28	Oklahoma	Female	Undergraduate	US	Undergraduate	Econ	20	68	6	20
29	Pennsylvania	Male	Graduate	US	Graduate	Politics	30	72	6	30
30	Ohio	Male	Undergraduate	US	Undergraduate	Econ	19	74	1	19
									Min	18
									Max	39
									Mean	25.2
									Median	23
									Mode	19
									Skewness	0.557190515
									Kurtosis	-1.049751548

Descriptive Statistics

Types of questions asked in descriptive statistics

1. What is the average score of students?
2. What is the frequency distribution of the major streams of the student?
3. What is the average age of the students? What is mode?
4. What proportion of students are graduate?
5. What is the distribution of students by heights?

Descriptive Statistics

proschool
An IMS Initiative

1. What is the average score of students?

Average score	25.2

2. What is the frequency distribution of the major streams of the student?

Frequency Distribution of major		
Politics	12	40%
Math	8	27%
Economics	10	33%

3. What is the average age of the students? What is mode?

Average Age	25.2
Mode	19

4. What proportion of students are graduate?

Proportion of student graduate	57%

5. What is the distribution of students by heights?

Average	66.43333
Median	66.5
Mode	68
Standard Deviation	4.658573
variance	21.7023
Skewness	0.171893

Let us discuss Case Study [Classroom case study].

Thank You.