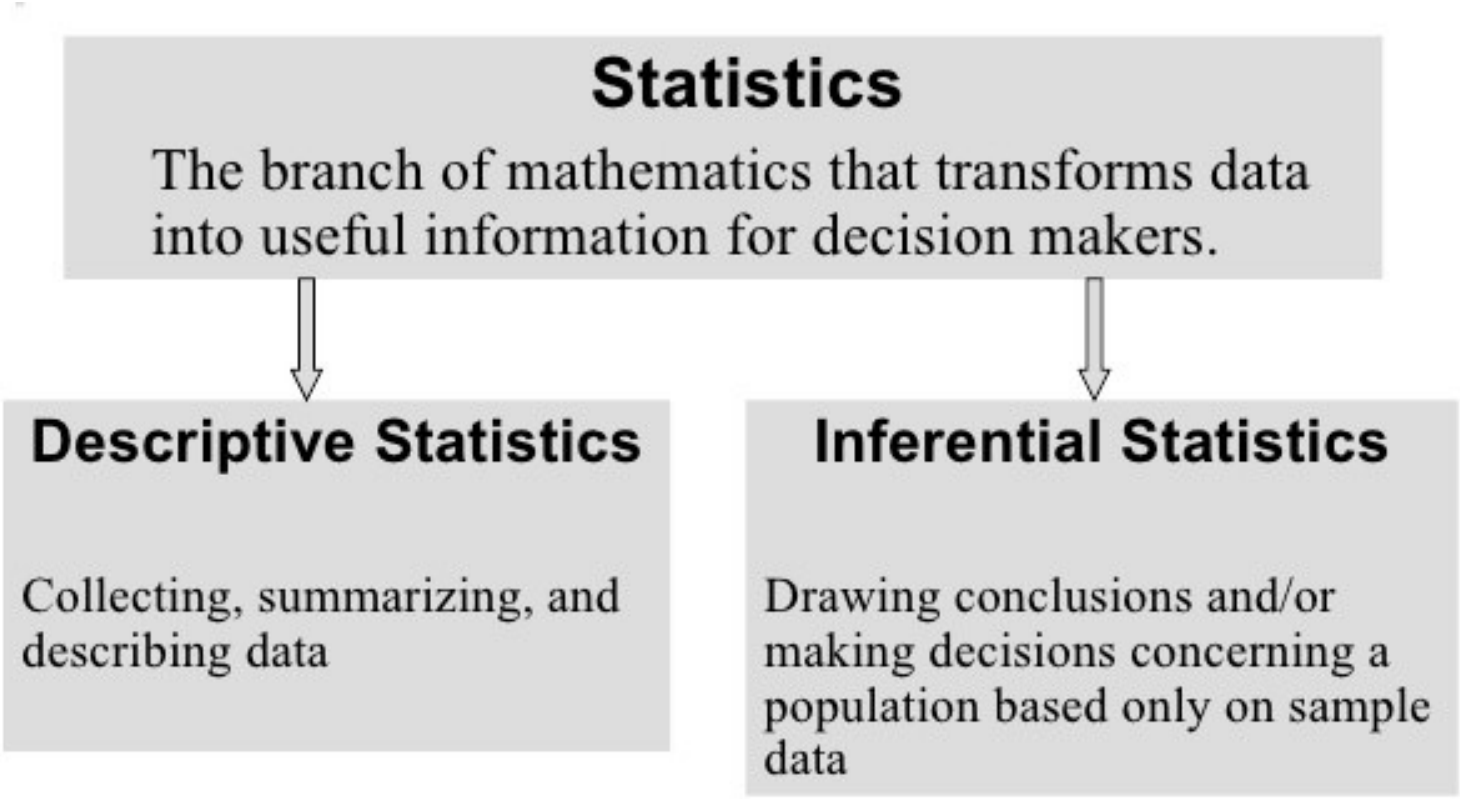
Overview of the course & Descriptive Statistics

Statistics


Statistics
The branch of mathematics that transforms data into useful information for decision makers.
Descriptive Statistics
Collecting, summarizing, and describing data
Inferential Statistics
Drawing conclusions and/or making decisions concerning a population based only on sample data

“Statistical thinking will be one day as necessary for efficient citizenship as the ability to read and write”

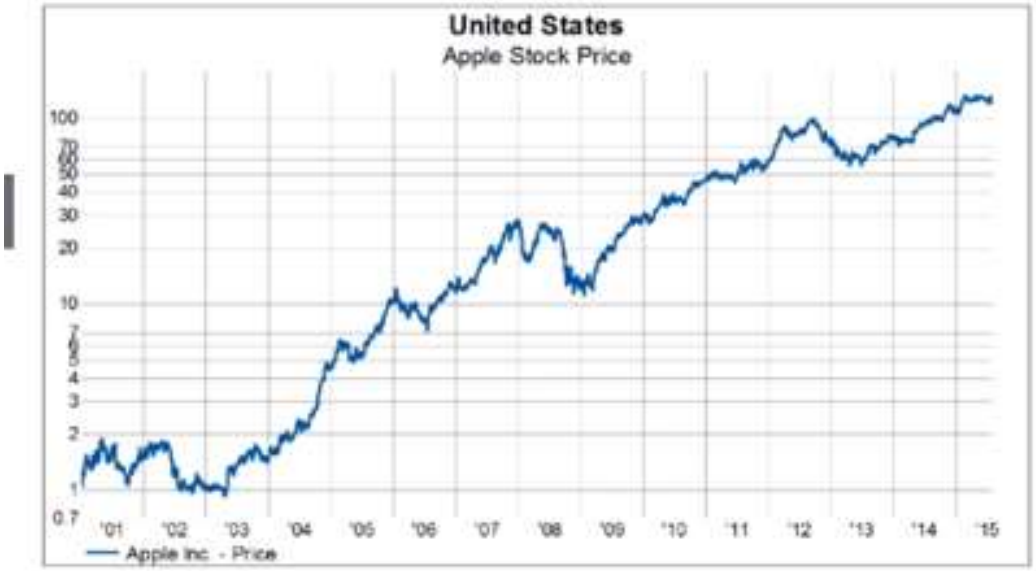
H G Wells

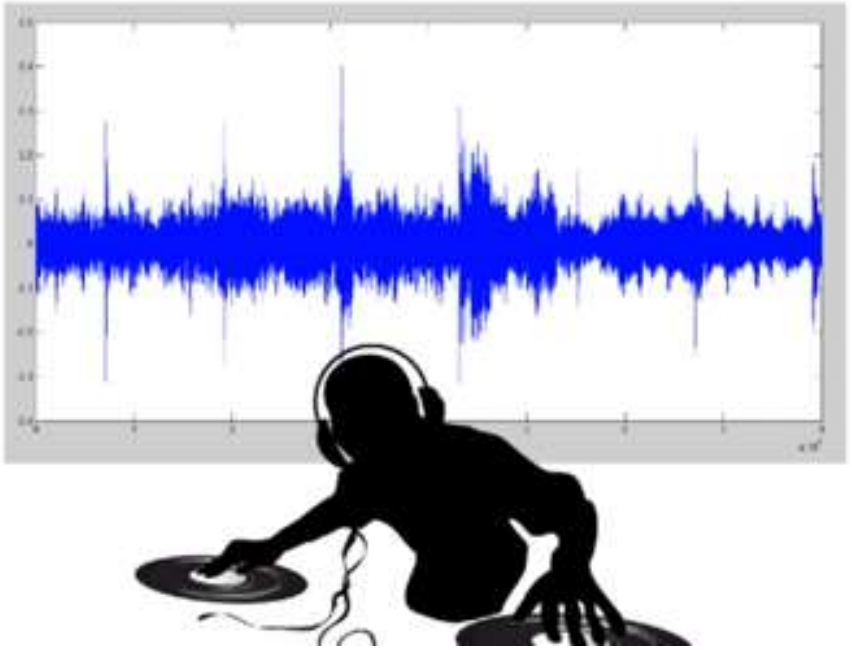
Statistical Visualization

A picture is worth a thousand words!

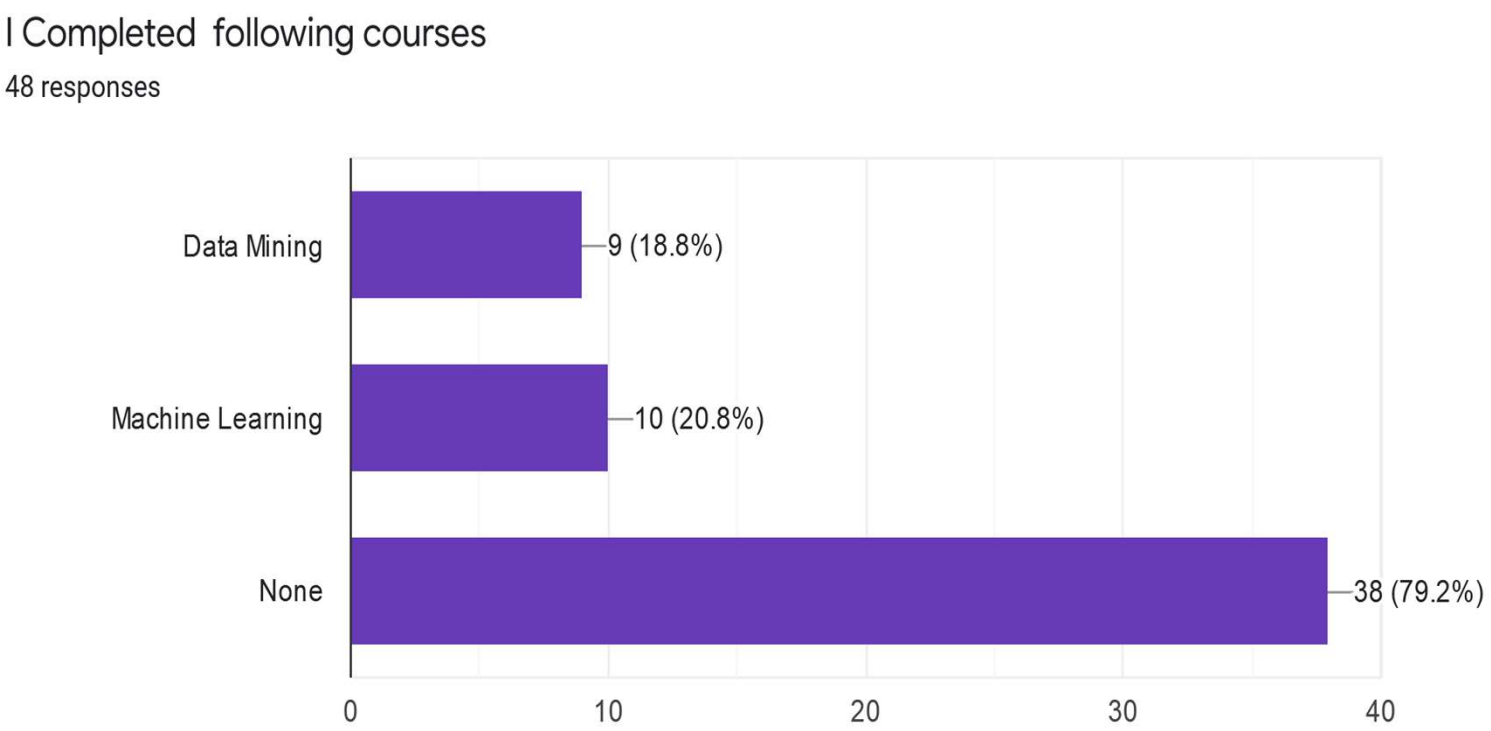
- Bar chart / graph
- Histogram
- Box plot
- Pie chart
- Density plot
- Line chart
- Frequency polygons
- Scatter plots

Line charts

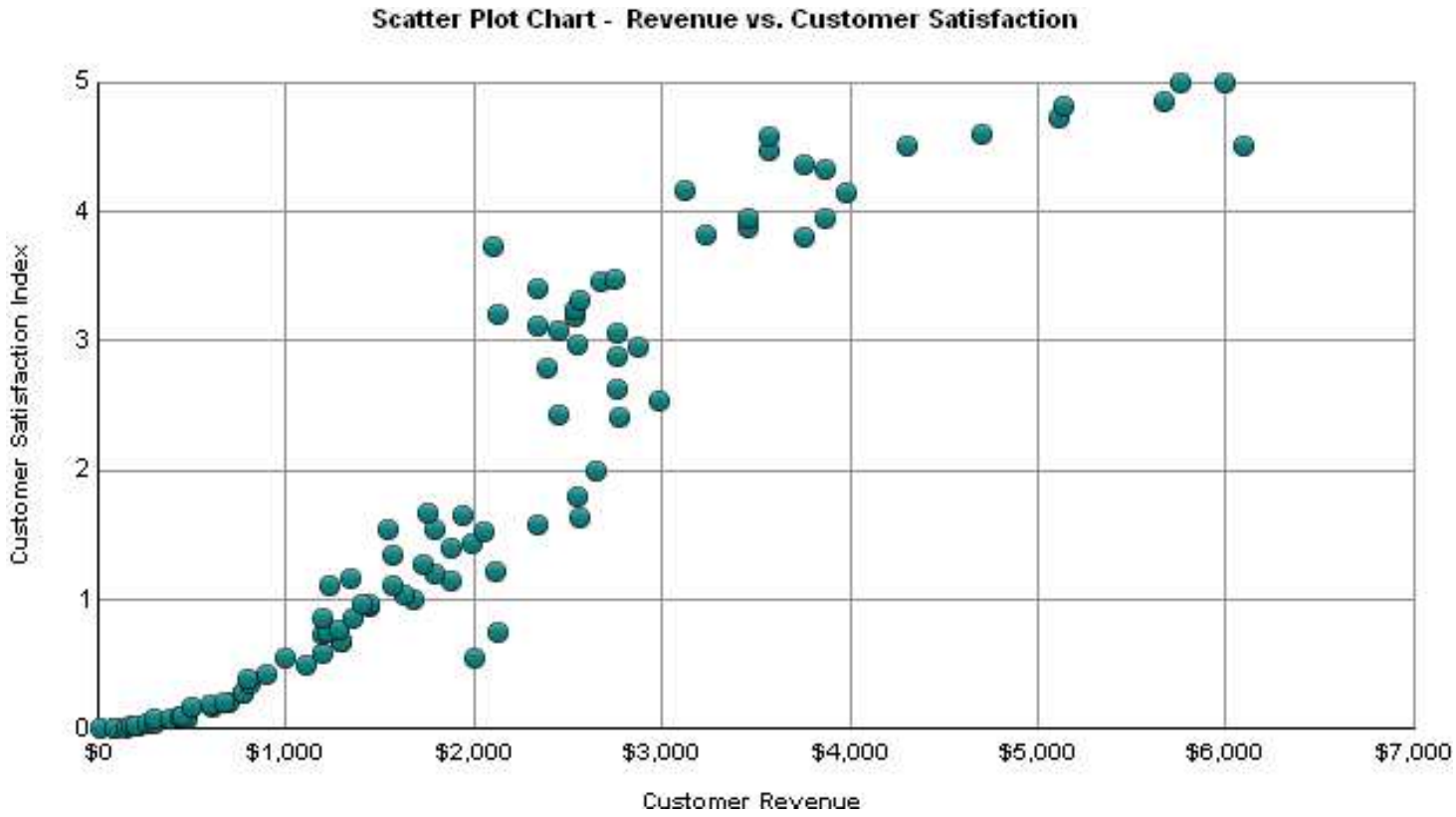

United States
Apple Stock Price
100
70
60
50
40
30
20
10
7
6
5
4
3
2
1
0.7
'01
'02
'03
'04
'05
'06
'07
'08
'09
'10
'11
'12
'13
'14
'15
Apple Inc. - Price



Bar Chart


I Completed following courses
48 responses
Data Mining
9 (18.8%)
Machine Learning
10 (20.8%)
None
38 (79.2%)
0
10
20
30
40

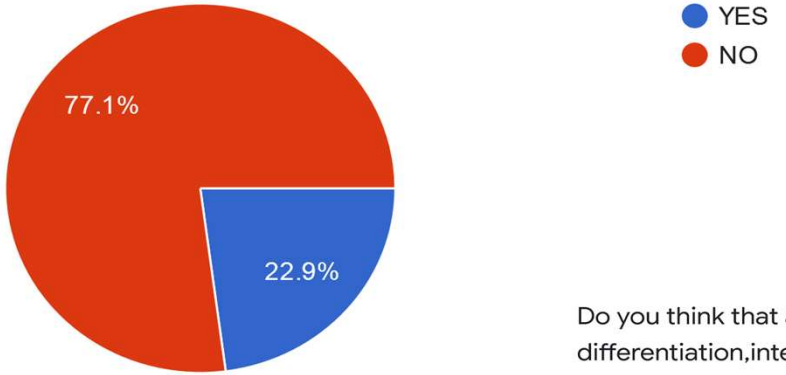
Scatterplots



Pie Charts

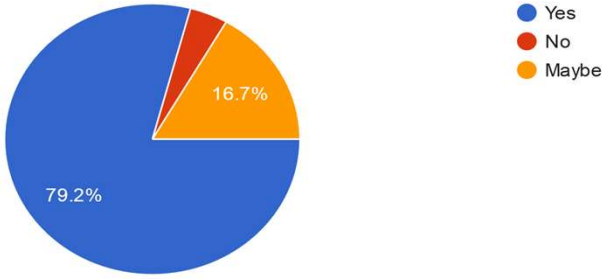
Work profile is related to Analytics(i.e Already into Machine learning / Analytics at workplace)

48 responses



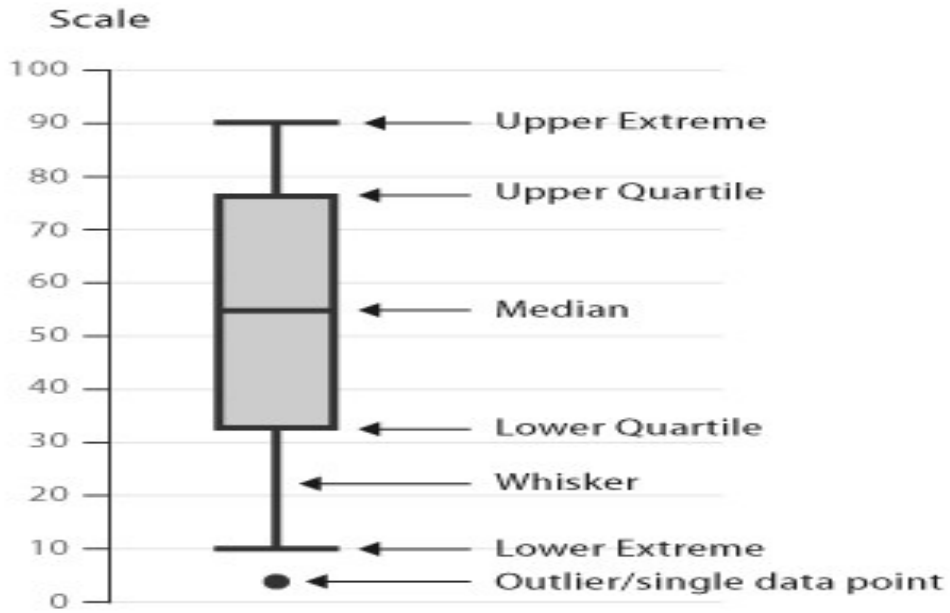
Do you think that a session is required to revise basics required like set theory,basic differentiation,integration etc

48 responses



Box Plot

A **Box and Whisker Plot** (or **Box Plot**) is a convenient way of visually displaying the data distribution through their quartiles. The lines extending parallel from the **boxes** are known as the “whiskers”, which are used to indicate variability outside the upper and lower quartiles.



Measures of Central Tendency

Measures of Variability

Measures of Central Tendency

- Measure of central tendency provides a very convenient way of describing a set of scores with a single number that describes the **PERFORMANCE** of the group.

- Also defined as a single value that is used to describe the "**center**" of the data.

- Three commonly used measures of central tendency:
  1. Mean
  2. Median
  3. Mode

Mean

- Also referred as the "**arithmetic average**"
- The most commonly used measure of the center of data
- Numbers that describe what is average or typical of the distribution
- Computation of Sample Mean:

$$\bar{Y} = \frac{\sum Y}{N}$$ = Σ Y/ N = (Y1 + Y2 + Y3 + … Yn) / N where

"Y bar" equals the sum of all the scores, Y, divided by the number of scores, N.

- Computation of the Mean for grouped Data

$$\bar{Y} = \frac{\sum f Y}{N}$$ Where f Y = a score multiplied by its frequency

The Median

- The score that divides the distribution into two equal parts, so that half the cases are above it and half below it.

- The median is the middle score, or average of middle scores in a distribution.
  - Fifty percent (50%) lies below the median value and 50% lies above the median value.
  - It is also known as the middle score or the 50th percentile.

The Mode

- The category or score with the largest frequency (or percentage) in the distribution.
- The mode can be calculated for variables with levels of measurement that are: nominal, ordinal, or interval-ratio.

Example:

- Number of Votes for Candidates for Lok Sabha MP. The mode, in this case, gives you the "central" response of the voters: the most popular candidate.

  - Candidate A – 11,769 votes
  - Candidate B – 39,443 votes
  - Candidate C – 78,331 votes

The Mode: "Candidate C"

Properties

- It can be used when the data are qualitative as well as quantitative.
- It may not be unique.
- It is affected by extreme values.
- It may not exist.

Nominal, Ordinal, Interval Scales

Offers:	Nominal	Ordinal	Interval
The sequence of variables is established	–	Yes	Yes
Mode	Yes	Yes	Yes
Median	–	Yes	Yes
Mean	–	–	Yes
Difference between variables can be evaluated	–	–	Yes
Addition and Subtraction of variables	–	–	Yes

Where do you live?

- 1- Suburbs
- 2- City
- 3- Town

How satisfied are you with our services?

- Very Unsatisfied – 1
- Unsatisfied – 2
- Neutral – 3
- Satisfied – 4
- Very Satisfied – 5

Examples of interval variables include:

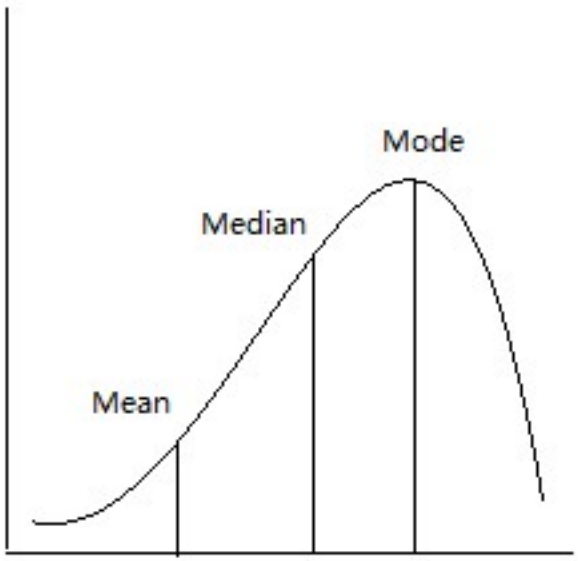
- temperature (Farenheit), temperature (Celcius), pH, SAT score (200-800), credit score (300-850).

Understanding the data

Symmetric OR **not**

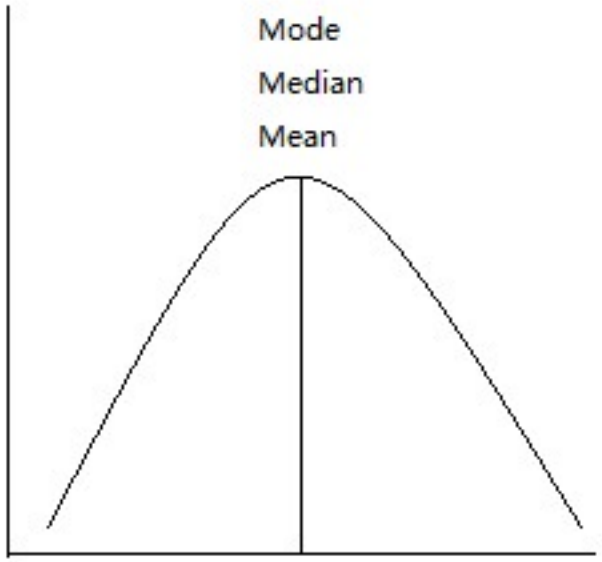
If mean = median = mode ,then the data is symmetric

Skewness of data



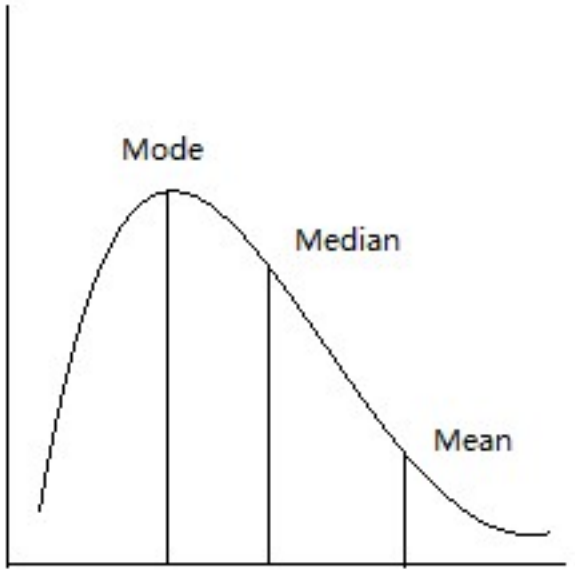
Left skew

Negative Skew – Mean < Median



Normal Distribution

Symmetrical



Right skew

Positive Skew – Mean> Median

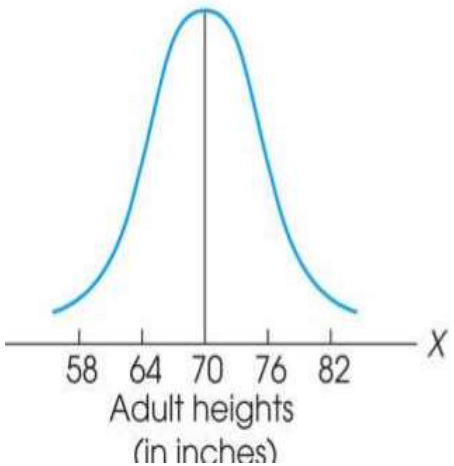
Shape of the Distribution

- **Symmetrical : mean is about equal to median**
- **Skewed**
  - **Negatively : mean < median**
  - **Positively : mean > median**
- **Bimodal : has two distinct modes**
- **Multi-modal : has more than 2 distinct modes)**

Measure of Variability

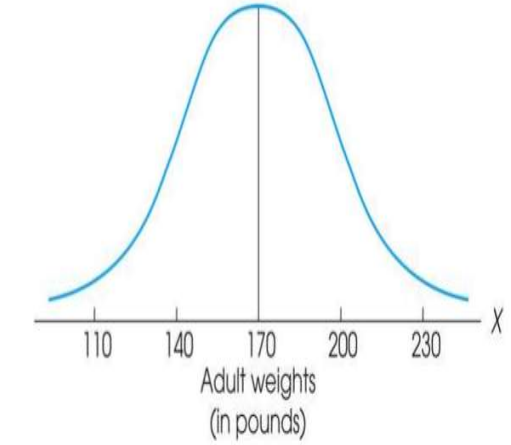
Variability can be defined several ways:

- A quantitative distance measure based on the differences between scores
- Describes distance of the spread of scores or distance of a score from the mean



Purposes of Measure of Variability:

- Describe the distribution
- Measure how well an individual score represents the distribution



The Three Measures

Three Measures of Variability:

- The Range
- The Variance
- The Standard Deviations

The Ranges

- The distance covered by the scores in a distribution – From smallest value to highest value
- For continuous data, real limits are used

$$\text{Range} = X_{max} - X_{min}$$

- Based on two scores, not all the data – An imprecise, unreliable measure of variability

Example: For a set of scores: 7, 2, 7, 6, 5, 6, 2

Range = Highest Score minus Lowest score = 7 - 2 = 5

The Standard Deviation

- Most common and most important measure of variability is the standard deviation
    - A measure of the standard, or average, distance from the mean
    - Describes whether the scores are clustered closely around the mean or are widely scattered
- Calculation differs for population and samples
- Variance is a necessary *companion concept* to standard deviation but *not the same* concept

The Standard Deviation

New Strategy :

a) First square each deviation score
b) Then sum the Squared Deviations (SS)
c) Average the squared deviations

- Mean Squared Deviation is known as **"Variance"**
- Variability is now measured in squared units

$$Standard\ Deviation = \sqrt{Variance}$$

The Variance

- Variance equals mean (average) squared deviation (distance) of the scores from the mean

$$\text{Variance} = \frac{\text{sum of squared deviations}}{\text{number of scores}}$$

where $SS = \sum (X - \mu)^2$

Random Experiment

Term "**random experiment**" is used to describe any action whose outcome is not known in advance. Here are some examples of experiments dealing with statistical data:

- Tossing a coin
- Counting how many times a certain word or a combination of words appears in the text of the “King Lear”
- Counting occurrences of a certain combination of amino acids in a protein database.
- Pulling a card from the deck

Sample spaces and events

The ***sample space*** of a random experiment is a set S that includes all possible outcomes of the experiment.

For example, if the experiment is to throw a die and record the outcome, the sample space is S = { 1,2,3,4,5,6}

- **Discrete sample spaces --- S ={1,2,3,4}**
- **Continuous sample spaces ---- S = (1,3)**

A **discrete sample space** S is either finite or countably infinite.
A **continuous sample space** S is uncountably infinite.

Sample Space of Rolling a Pair of Dice

Sum ↓						
2	(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
3	(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
4	(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
5	(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
6	(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
7	(6, 1)	(6, 2)	(6, 3)	(6, 4)	(6, 5)	(6, 6)
8	9	10	11	12		

Event

An event is a subset of the sample space of a random experiment.

An event is a set of outcomes of the experiment. This includes the *null* (empty) set of outcomes and the set of *all* outcomes. Each time the experiment is run, a given event *A* either *occurs*, if the outcome of the experiment is an element of *A*, or *does not occur*, if the outcome of the experiment is not an element of *A*.

- If a single face is considered when a die is rolled, then it will be **simple event.**
  - For example suppose getting 5 or 6 or 3 or 2 etc… on the die when it is thrown, is called as simple event.
- If the event is any even number on the die, then the event is consisting of points {2, 4, 6}, which is known as **compound event.**
  - That compound event is consisting of three simple events i.e., {2}, {4} and {6}

Simple & Compound Event

A **simple** (or single) **event** is an **event** with a single outcome (only one "answer"). ... A **compound event** is the combination of two or more **simple events** (with two or more outcomes)

- If a single face is considered when a die is rolled, then it will be **simple event.**
  - For example suppose getting 5 or 6 or 3 or 2 etc… on the die when it is thrown, is called as simple event.
- If the event is any even number on the die, then the event is consisting of points {2, 4, 6}, which is known as **compound event.**
  - That compound event is consisting of three simple events i.e., {2}, {4} and {6}

Mutually Exclusive Events

Two events are said to be mutually exclusive if the events have no sample points in common. That is, two events are mutually exclusive if, when one event occurs, the other cannot occur.

The two possible outcomes of a coin flip are mutually exclusive; when you flip a coin, it cannot land both heads and tails simultaneously.

Questions:

1) **Rolling a number divisible by 2** or **rolling a number divisible by 3**
2) **Rolling a number divisible by 2** or **rolling a number that is a multiple of 5**
3) **Rolling a prime number** or **rolling an even number**
4) **Rolling a non-prime number or rolling an odd number**

Answers:
1) Non-mutually exclusive (you could a roll a 6, which is divisible by both 2 and 3)
2) Mutually exclusive (you cannot roll a 2,4, or 6 at the same time as you roll a 5)
3) Non-mutually exclusive (you could roll a 2, which is an even prime number)
4) Mutually exclusive (the only non-prime numbers on the die are 4 and 6, which are not odd)

Answers:
1) Non-mutually exclusive (you could a roll a 6, which is divisible by both 2 and 3)
2) Mutually exclusive (you cannot roll a 2,4, or 6 at the same time as you roll a 5)
3) Non-mutually exclusive (you could roll a 2, which is an even prime number)
4) Mutually exclusive (the only non-prime numbers on the die are 4 and 6, which are not odd)

Axioms of Probability

Probability is a measure of the likelihood of an event to occur. Many events cannot be predicted with total certainty. We can predict only the chance of an event to occur i.e. how likely they are to happen, using it.

Probability is a number that is assigned to each member of a collection of events from a random experiment that satisfies the following properties:

If S is the sample space and E is any event in a random experiment,

(1) $P(S) = 1$

(2) $0 \leq P(E) \leq 1$

(3) For two events E_1 and E_2 with $E_1 \cap E_2 = \varnothing$

$$P(E_1 \cup E_2) = P(E_1) + P(E_2)$$

Probability

Probability of event to happen P(E) = $\dfrac{\text{Number of favorable outcomes}}{\text{Total Number of outcomes}}$

1) There are 6 pillows in a bed, 3 are red, 2 are yellow and 1 is blue. What is the probability of picking a yellow pillow?

2) There is a container full of colored bottles, red, blue, green and orange. Some of the bottles are picked out and displaced. Sumit did this 1000 times and got the following results:

- No. of blue bottles picked out: 300
- No. of red bottles: 200
- No. of green bottles: 450
- No. of orange bottles: 50

a) What is the probability that Sumit will pick a green bottle?

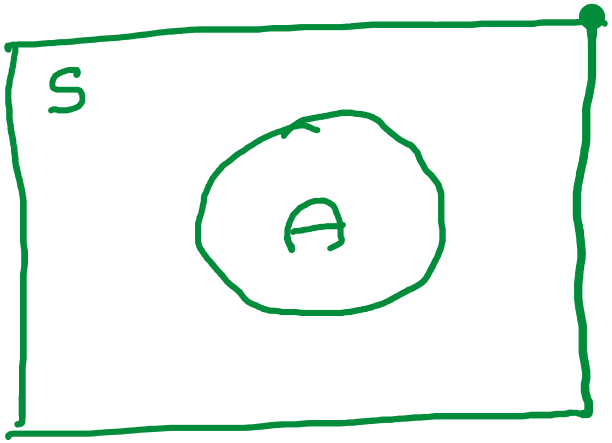
b) If there are 100 bottles in the container, how many of them are likely to be green?

Probability

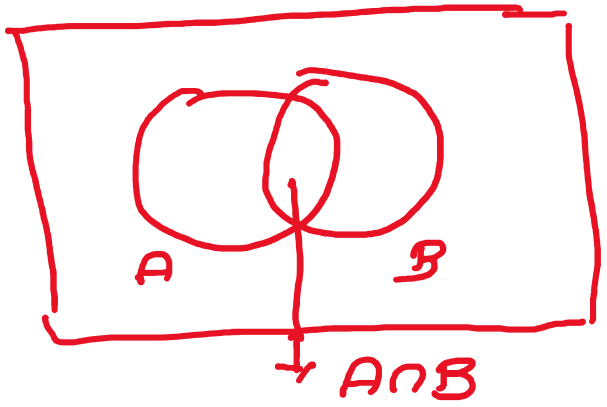
$A \cup \bar{A} = S$

$P(A \cup \bar{A}) = P(S)$

$P(A) + P(\bar{A}) = 1$ ✓



$P(A \cup B) = P(A) + P(B) - P(A \cap B)$



Thanks