

# Advanced Statistic Techniques and Analytics Notes

---

Akhil S

---

2021MT12054

---

---





# **Session 1(22<sup>nd</sup> January,2022)**

## **Overview of the course & Descriptive Statistics**



---

# **Probability and Statistics for Engineering and the Sciences**

**By**

**JAY L DEVORE**

---

# Overview of the course

---

- ❖ Descriptive Statistics
  - ❖ Probability
  - ❖ Conditional Probability
  - ❖ Random Variables
  - ❖ Probability Distributions – Univariate & Joint
  - ❖ Sampling & Estimation
  - ❖ Testing of Hypothesis – mean , proportions
  - ❖ Regression
  - ❖ Time Series Analysis
-

# Evaluation components

- **Quiz -1** : 5 Marks
- **Mid Semester Exam** : 30 Marks
- **Assignment** : 10 Marks
- **Quiz – 2:** 5 Marks
- **Comprehensive Exam:** 50 Marks

!!!!

- A famous statistician would never travel by airplane, because she had studied air travel and estimated the probability of there being a bomb on any given flight was 1 in a million, and she was not prepared to accept these odds.
- One day a colleague met her at a conference far from home.
- "How did you get here, by train?"
- "No, I flew"
- "What about the possibility of a bomb?"
- "Well, I began thinking that if the odds of one bomb are 1:million, then the odds of TWO bombs are  $(1/1,000,000) \times (1/1,000,000) = 10^{-12}$ . This is a very, very small probability, which I can accept. So, now I bring my own bomb along!"

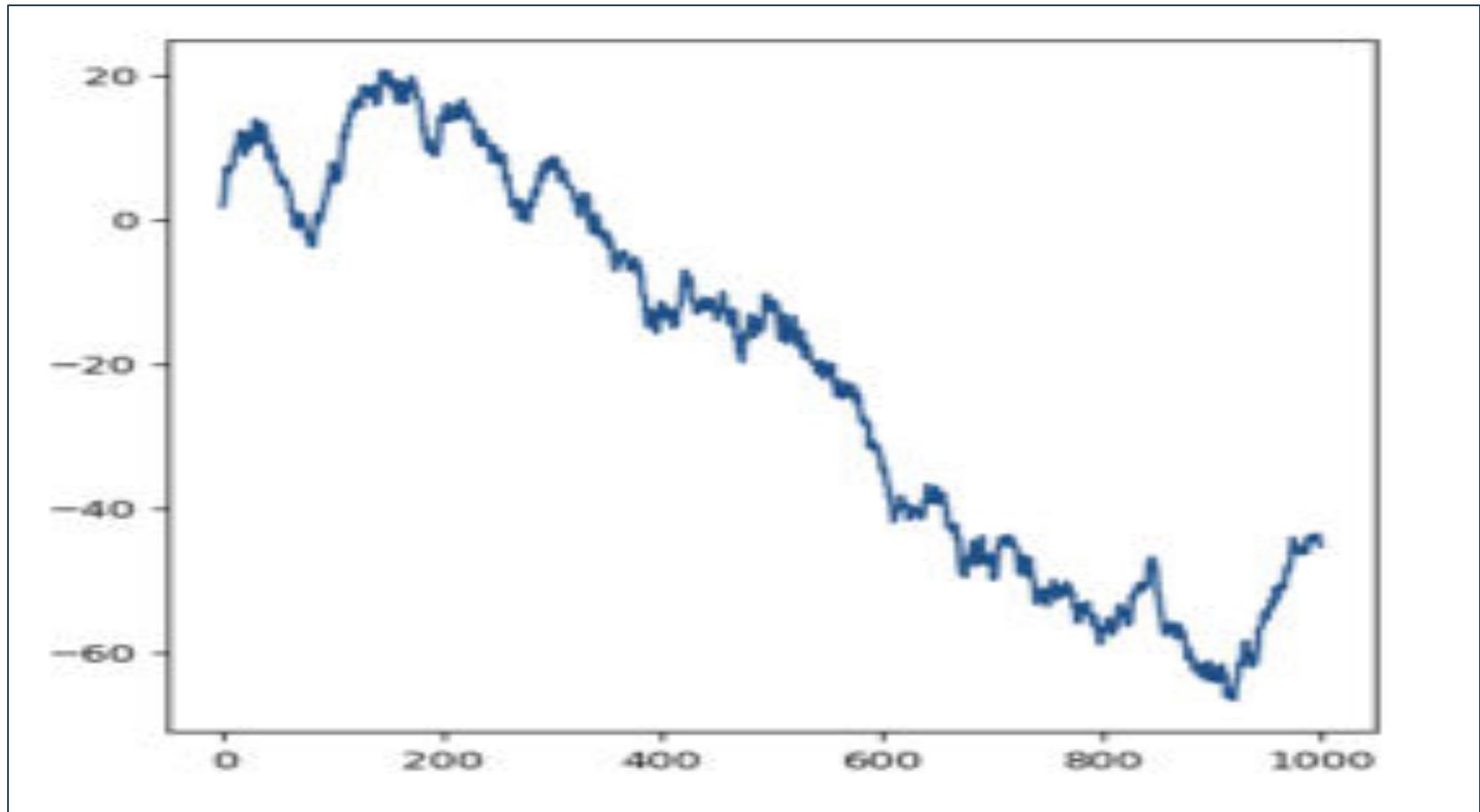
---

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

*H G Wells*

---

# Line charts

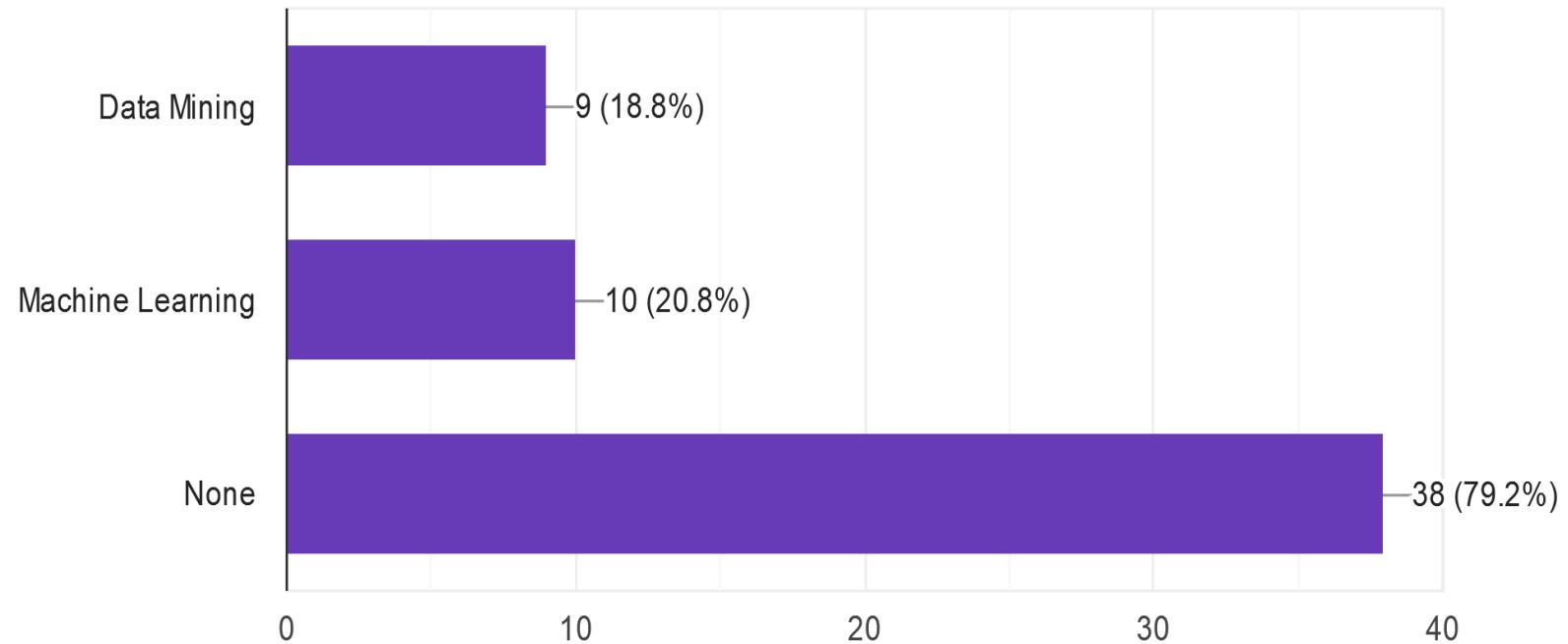




# Bar Chart

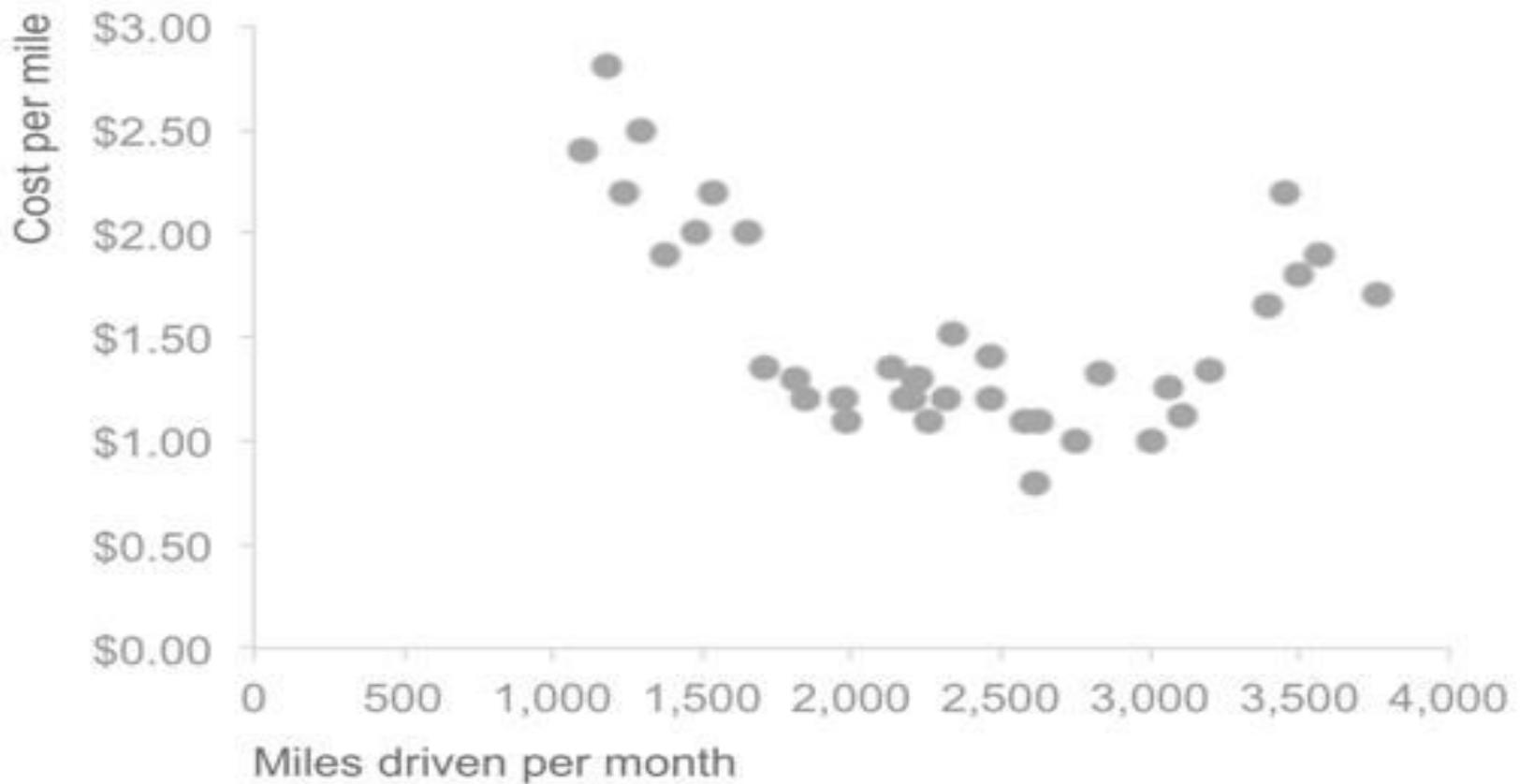
I Completed following courses

48 responses



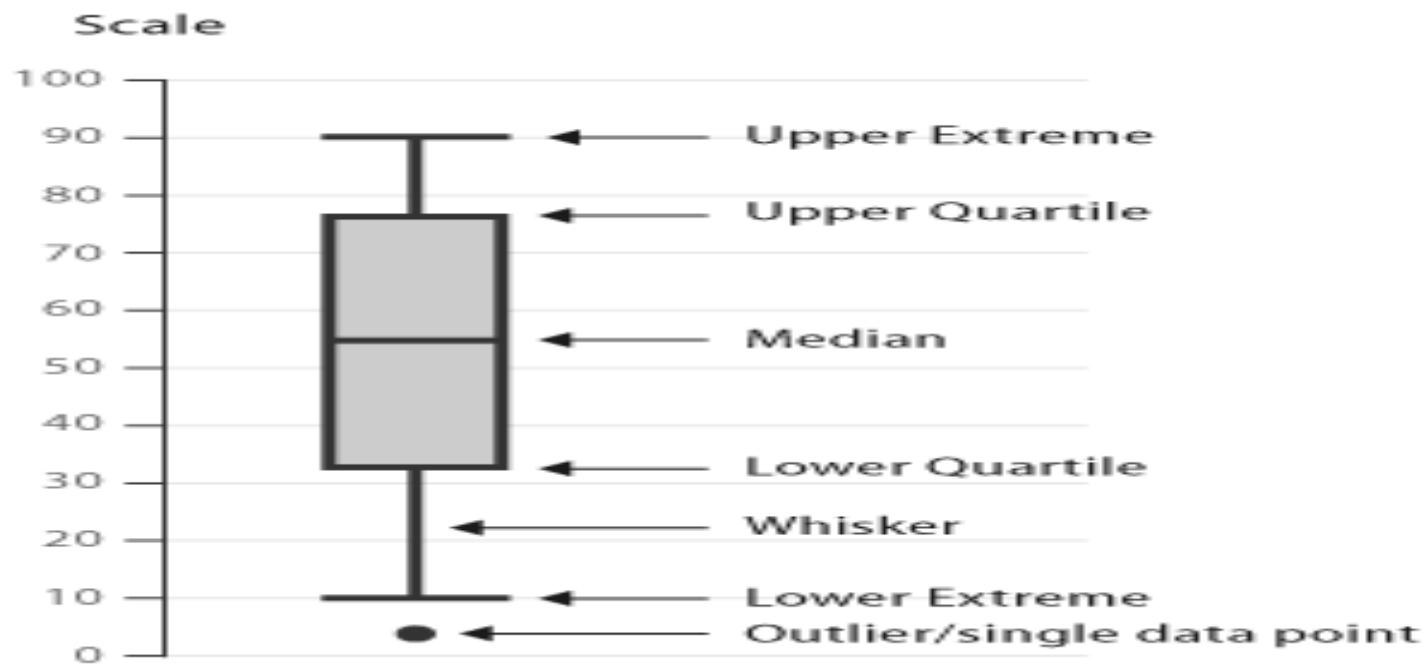
# Scatterplots

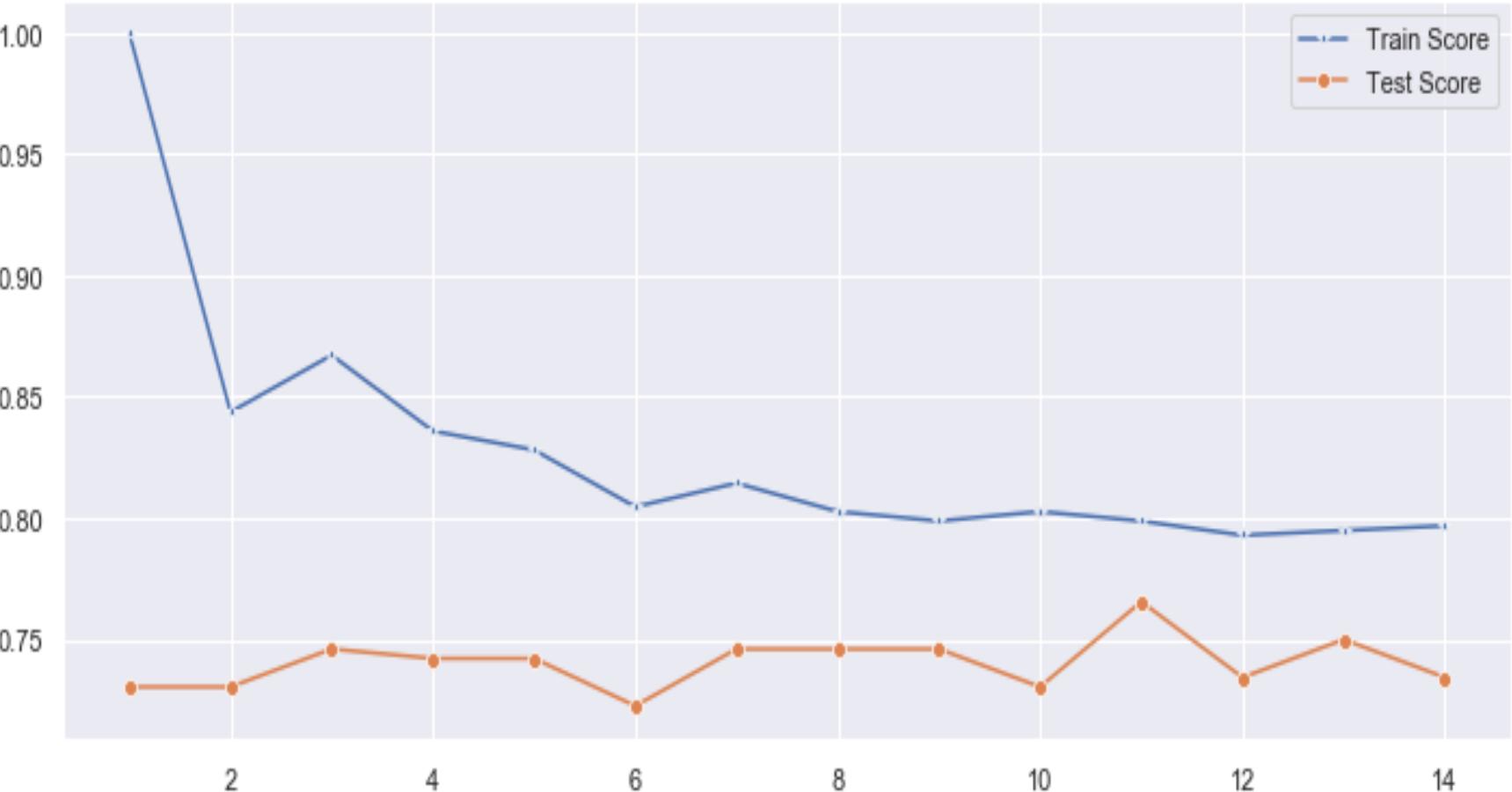
Cost per mile by miles driven

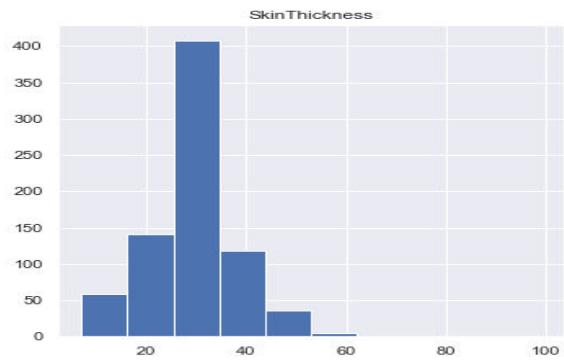
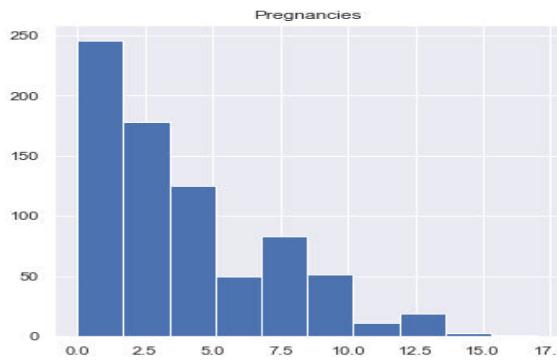
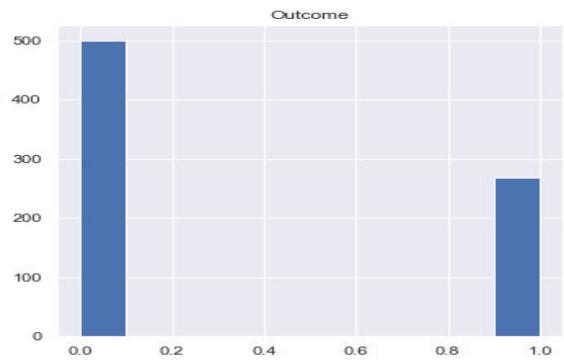
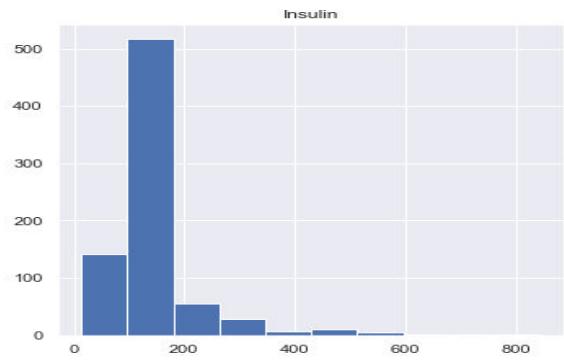
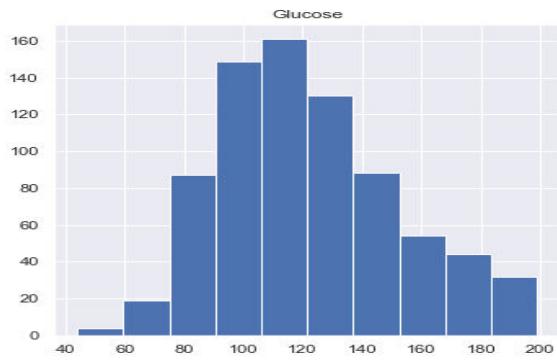
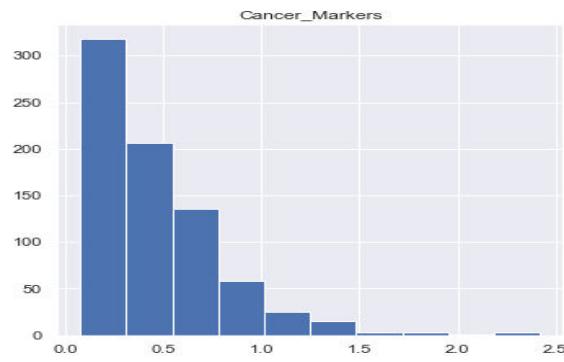
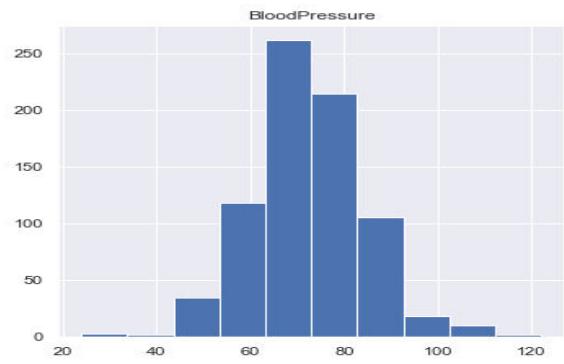
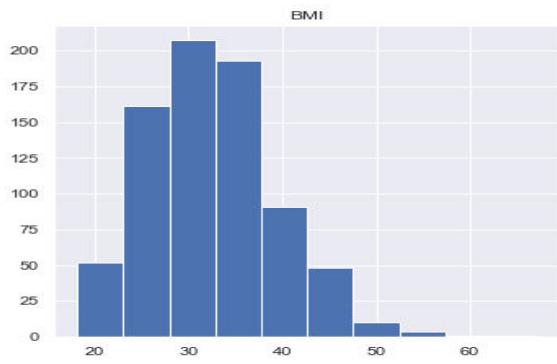
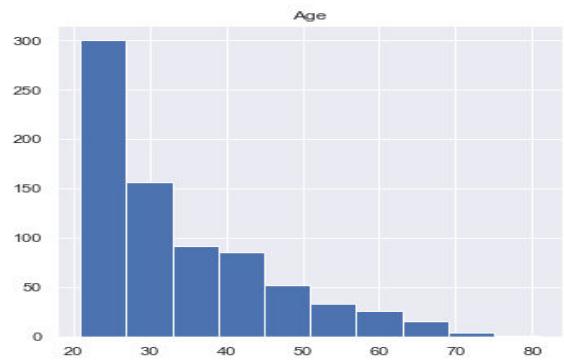


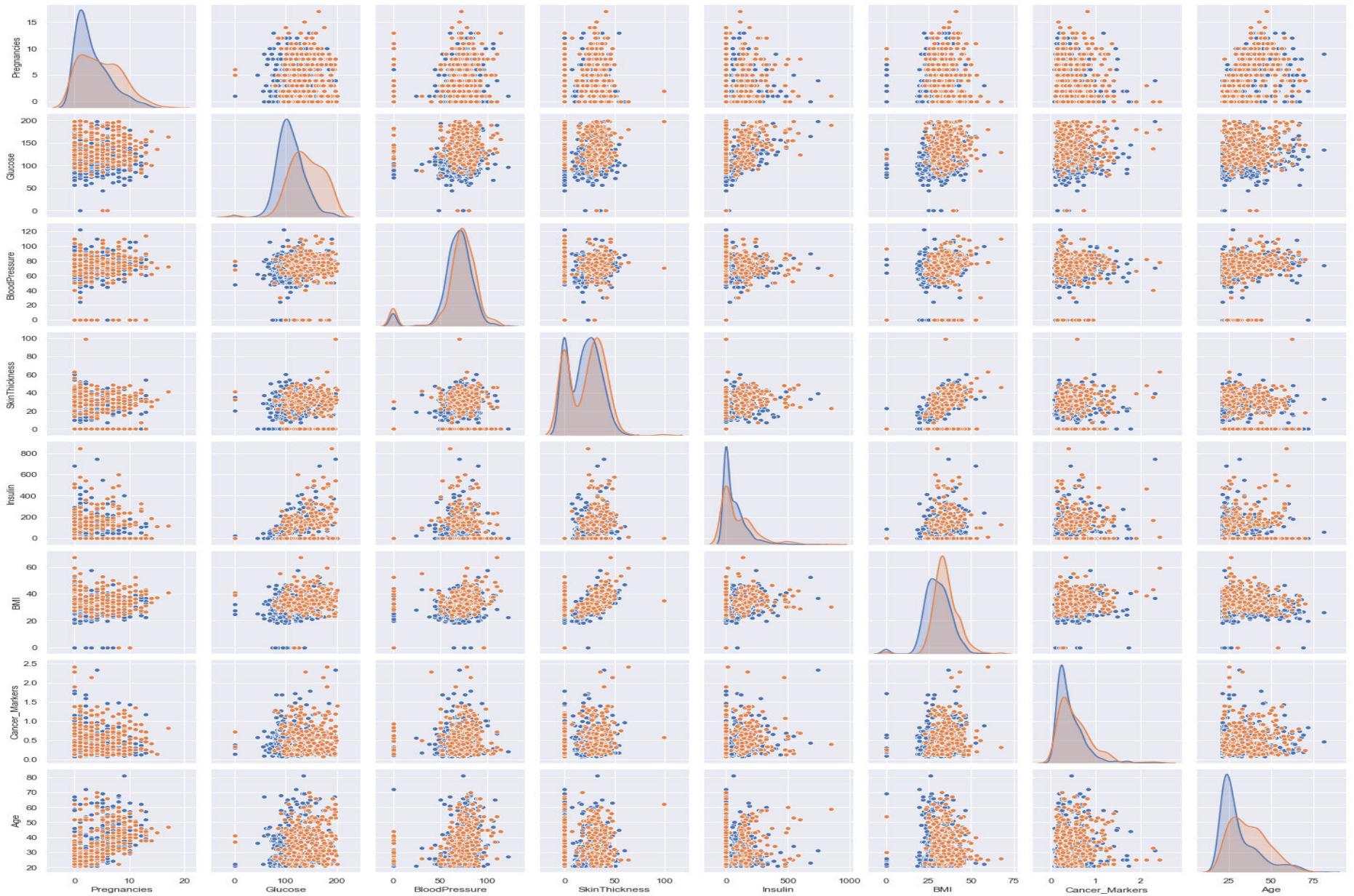
# Box Plot

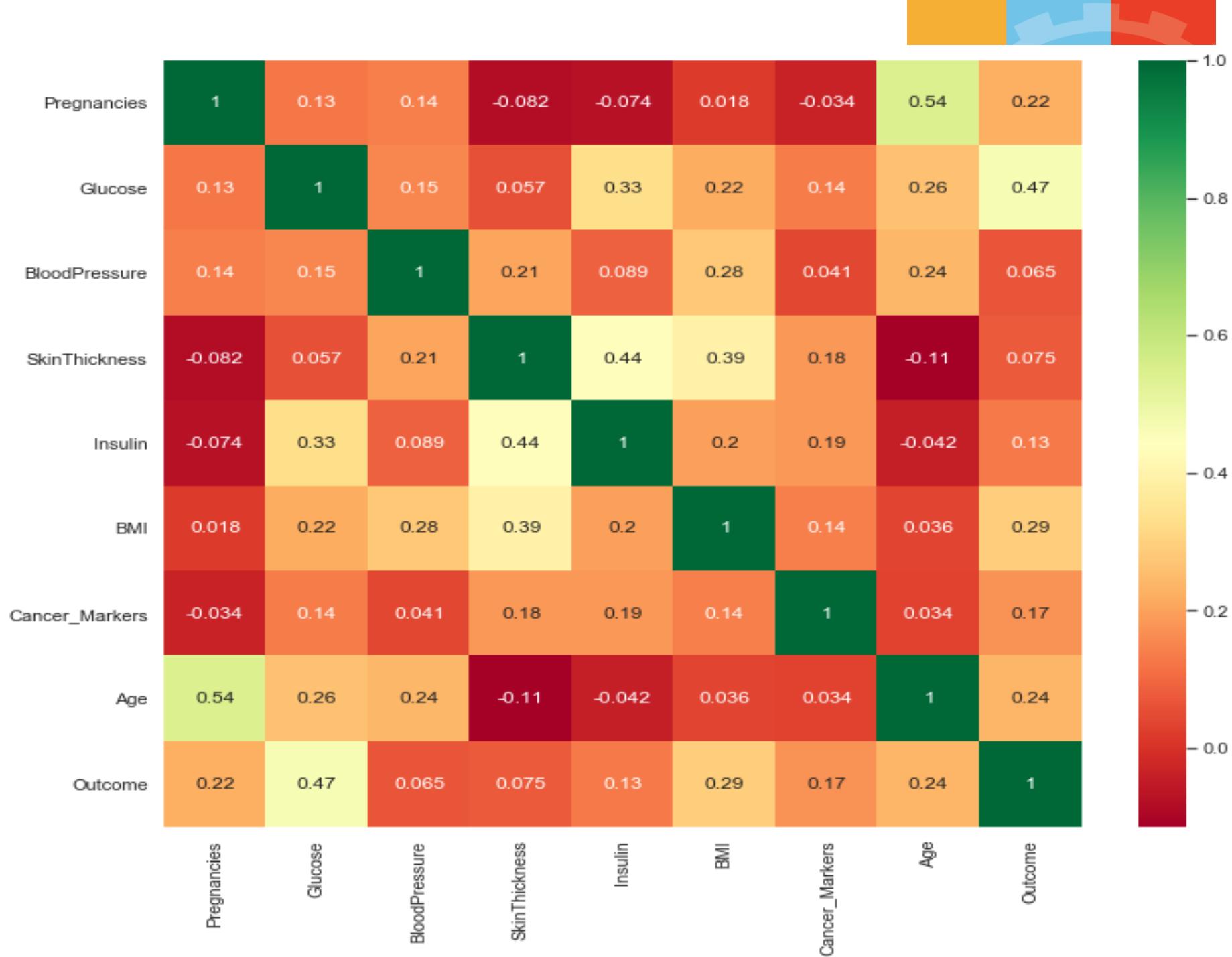
A **Box and Whisker Plot** (or **Box Plot**) is a convenient way of visually displaying the data distribution through their quartiles. The lines extending parallel from the **boxes** are known as the “whiskers”, which are used to indicate variability outside the upper and lower quartiles.



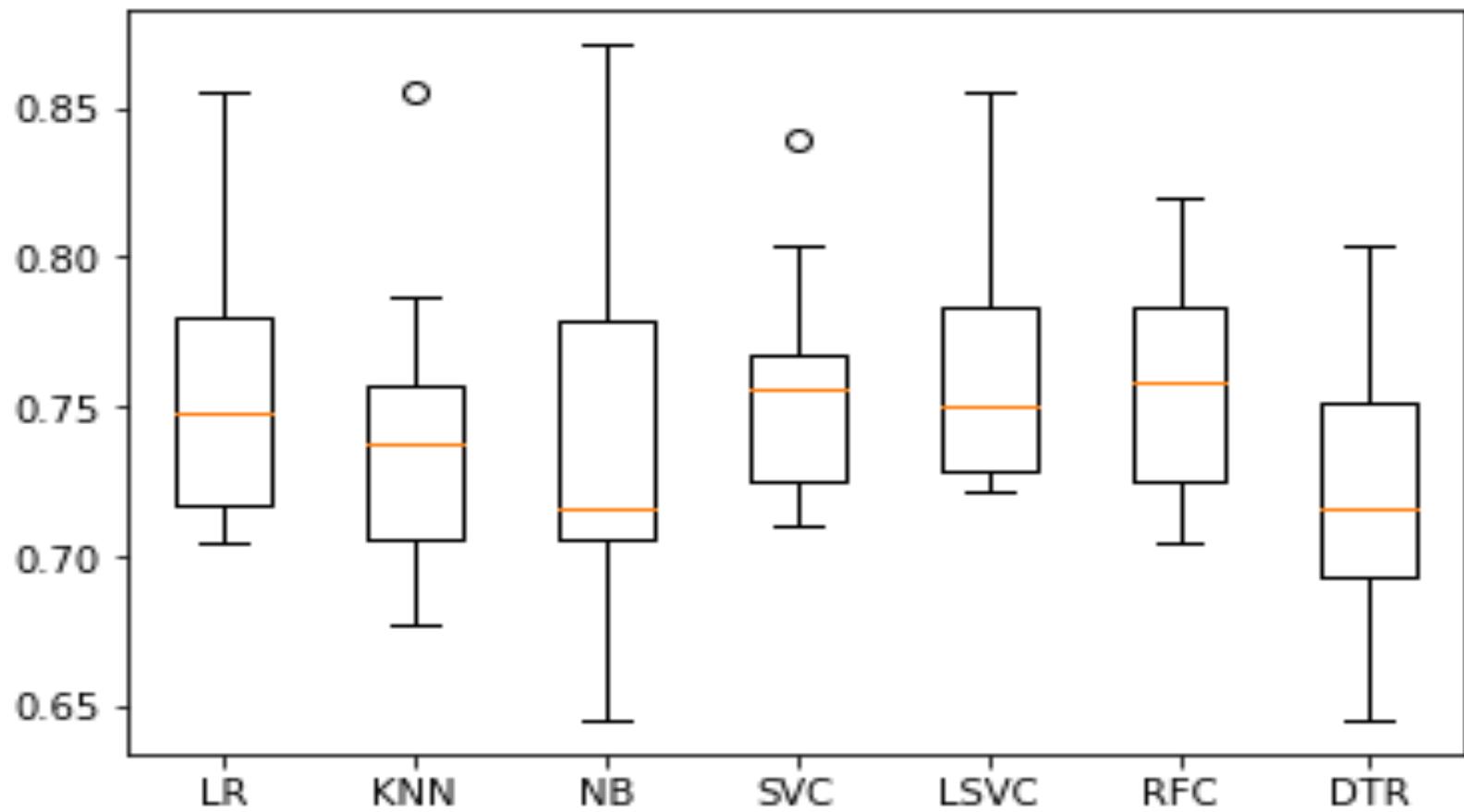








## Algorithm Comparison



# Statistical Summary

Cost	Weight	Weight1	Length	Height	Width	
count	159.000000	159.000000	159.000000	159.000000	159.000000	159.000000
mean	398.326415	26.247170	28.415723	31.227044	8.970994	4.417486
std	357.978317	9.996441	10.716328	11.610246	4.286208	1.685804
min	0.000000	7.500000	8.400000	8.800000	1.728400	1.047600
25%	120.000000	19.050000	21.000000	23.150000	5.944800	3.385650
50%	273.000000	25.200000	27.300000	29.400000	7.786000	4.248500
75%	650.000000	32.700000	35.500000	39.650000	12.365900	5.584500
max	1650.000000	59.000000	63.400000	68.000000	18.957000	8.142000

# **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:

$$\bar{Y} = \frac{\sum Y}{N} = \Sigma Y / N = (Y_1 + Y_2 + Y_3 + \dots + Y_n) / N \quad \text{where}$$

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

- Computation of the Mean for grouped Data

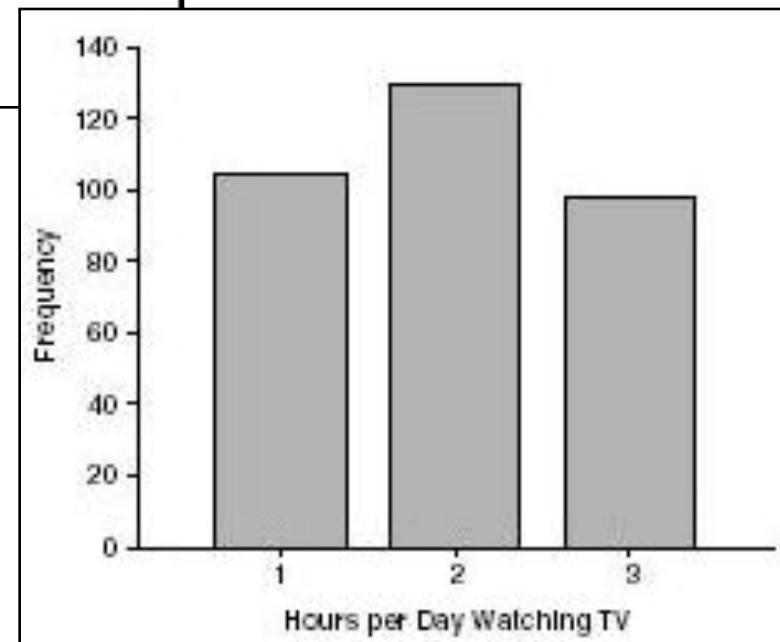
$$\bar{Y} = \frac{\sum f Y}{N} \quad \text{Where } f Y = \text{a score multiplied by its frequency}$$

# Mean: Grouped Scores

Hours Spent Watching TV	Frequency ( <i>f</i> )	<i>fY</i>	Percentage	C%
1	104	104	31.3	31.3
2	130	260	39.2	70.5
3	98	294	29.5	100.0
Total	332	658	100.0	

$$\bar{Y} = \frac{\sum fY}{N} = \frac{658}{332} = 1.98$$

Data of Children watching TV in Bengaluru



# Mean

## Properties

- ✓ It measures stability. Mean is the most stable among other measures of central tendency because every score contributes to the value of the mean.
- ✓ It may easily affected by the extreme scores.
- ✓ The sum of each score's distance from the mean is zero.
- ✓ It can be applied to interval level of measurement
- ✓ It may not be an actual score in the distribution
- ✓ It is very easy to compute.

# 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”**

# Mode

## ➤ 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.
- **When to Use the Median**
  - When the “typical” value is desired.
  - When the data set is measured on a nominal scale

# 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 Median

---

- **Median of Ungrouped Data**

1. Arrange the scores (from lowest to highest or highest to lowest).
  2. Determine the middle most score in a distribution if  $n$  is an odd number (and if  $n$  is an even number, get the average of the two middle most scores)
-

# Median Exercise #1 ( $N$ is odd)

Calculate the median for this hypothetical distribution:

Job Satisfaction	Frequency
Very High	2
High	3
Moderate	5
Low	7
Very Low	4
TOTAL	21

# Median Exercise #2 ( $N$ is even)

- Calculate the median for this hypothetical distribution:

Satisfaction with Health	Frequency
– Very High	5
– High	7
– Moderate	6
– Low	7
– Very Low	3
–	TOTAL 28

# Median in Grouped Data

$$\text{Median} = L + \frac{N(.5) - Cf}{f} \times w$$

- Where:
- L = Lower boundary of the category containing the N/2
- Cf = Cumulative frequency before the median class if the scores are arranged from the lowest to highest value
- w = Size of the class interval
- f = frequency of the median class

# Median in Grouped Data

---

## Steps to solve median for grouped data

1. Complete the table for Cf.
  2. Get  $N/2$  of the scores in the distribution so that you can identify MC
  3. Determine L, w, f and Cf
  4. Solve the median using the given formula
-

# Median

Example: Scores of 40 students in a science class consist of 60 items and they are tabulated below. The highest score is 54 and the lowest score is 10.

X	f	cf<
10 - 14	5	5
15 - 19	2	7
20 - 24	3	10
25 - 29	5	15
30 - 34	2	17 (cfp)
35 - 39	9 (fm)	26
40 - 44	6	32
45 - 49	3	35
50 - 54	5	40
	n = 40	

# Median

## Solution:

- $n/2 = 40/2 = 20$
- The category containing  $n/2$  is (35 – 39)
  - Lower Limit of MC = 35
    - $L = 34.5$
    - $Cf$  (or  $Cfp$ ) = 17
    - $f$  (or  $fm$ ) = 9
      - $w = 5$
  - Median =  $L + (n/2 - Cf) / f * w$ 
 $= 34.5 + (20-17)/9 * 5$ 
 $= 34.5 + 15/9$ 
 $= 36.17$

X	f	cf<
10 - 14	5	5
15 - 19	2	7
20 - 24	3	10
25 - 29	5	15
30 - 34	2	17 (cfp)
35 - 39	9 (fm)	26
40 - 44	6	32
45 - 49	3	35
50 - 54	5	40
	n = 40	

# Median

## Properties

- It may not be an actual observation in the data set.
- Not affected by extreme values because median is a positional measure.
- Can be applied in ordinal level.

## When to Use the Median

- The exact midpoint of the score distribution is desired.
- There are extreme scores in the distribution.

# Percentiles

- A score below which a specific percentage of the distribution falls.
- Finding percentiles in grouped data:

$$25\% = L + \frac{N(.25) - Cf}{f} \times w$$

# Measures of central tendency

- The mean
- the median
- the mode



# 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)
-

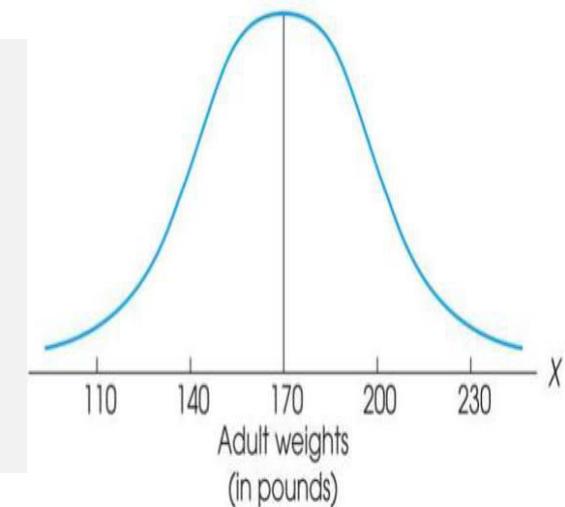
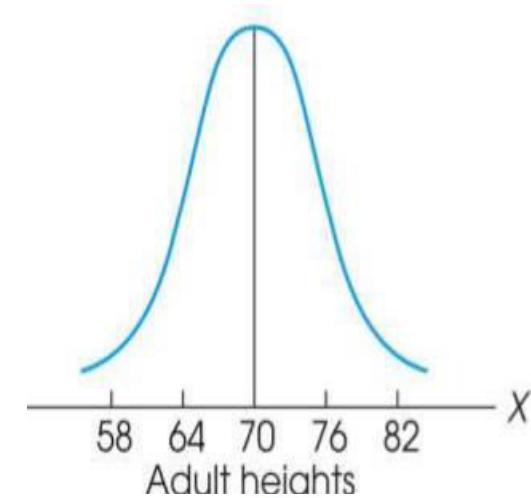
# Measure of Variability

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

$$\text{Range} = \text{URL for } X_{\max} - \text{LRL for } X_{\min}$$

- 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**

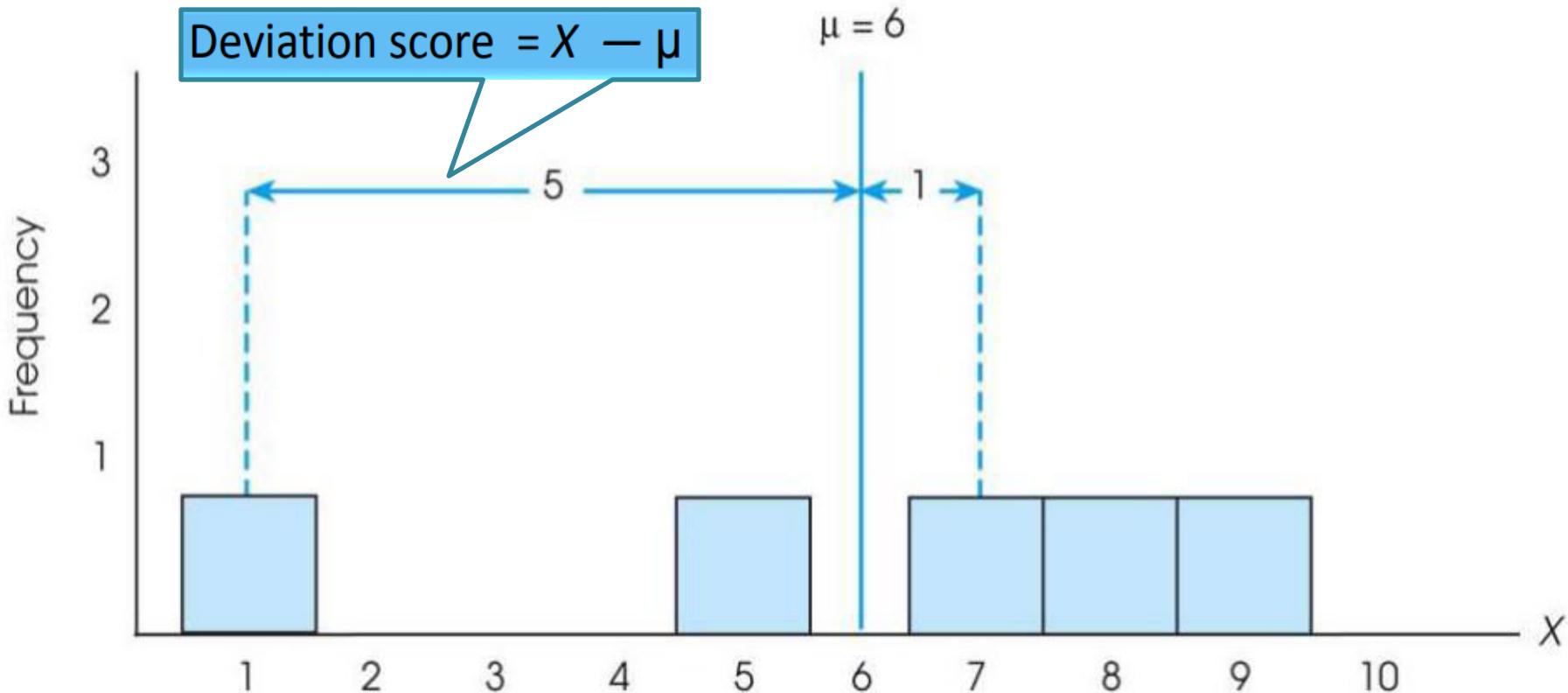
**Range = Highest Score minus Lowest score = 7 - 2 = 5**

# 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



Exercise : Find out the deviations of all the data points with the mean....and then find the 'mean deviation'.

# 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

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

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

# The Population Variance

---

- ❖ Population variance equals mean (average) squared deviation (distance) of the scores from the population mean
  - ❖ Variance is the average of squared deviations, so we identify population variance with a lowercase Greek letter sigma squared:  $\sigma^2$
  - ❖ Standard deviation is the square root of the variance, so we identify it with a lowercase Greek letter sigma:  $\sigma$
-

# Standard Deviation and Variance for a Sample



- Goal of inferential statistics:
  - Draw general conclusions about population
  - 
  - Based on limited information from a sample
- Samples differ from the population
  - Samples have less variability
  - Computing the Variance and Standard Deviation in the same way as for a population would give a biased estimate of the population values

# Sample Standard

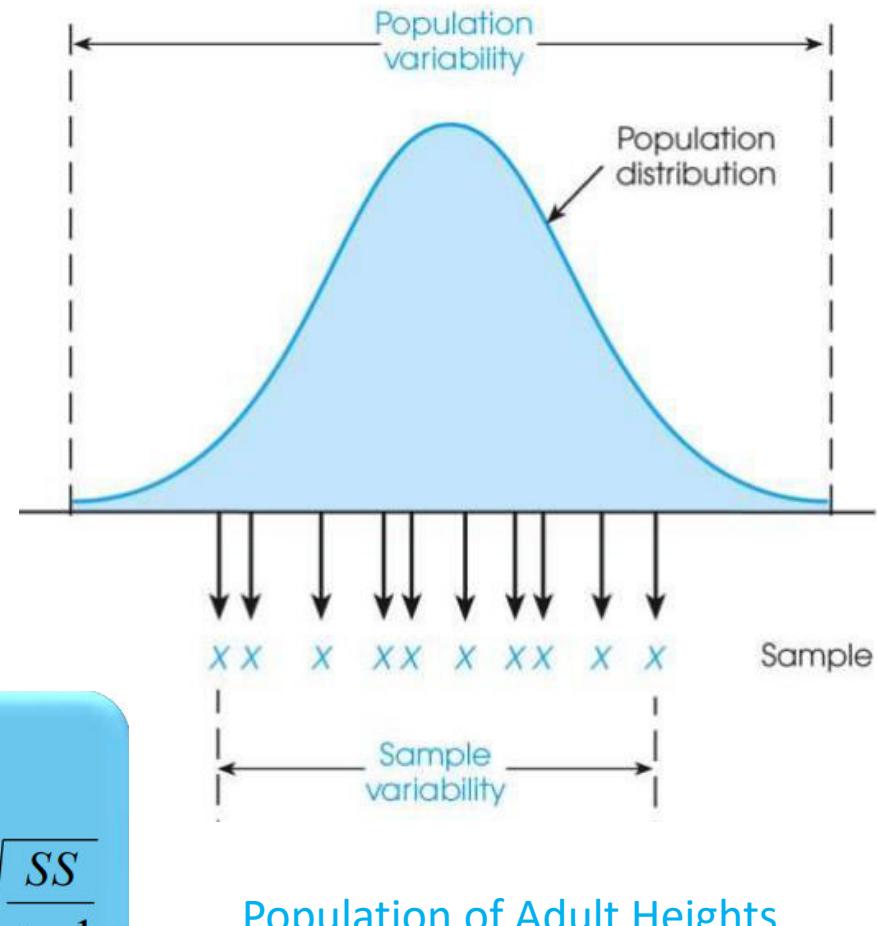


# Deviation and Variance

- Sum of Squares (SS) is computed as before
- Formula for Variance has  $n-1$  rather than  $N$  in the denominator
- Notation uses  $s$  instead of  $\sigma$

$$\text{variance of sample} = s^2 = \frac{SS}{n-1}$$

$$\text{standard deviation of sample} = s = \sqrt{\frac{SS}{n-1}}$$



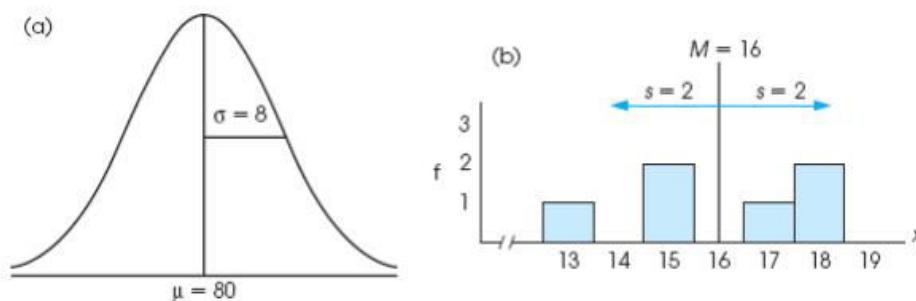
# Degrees of Freedom

---

- Population variance
    - Mean is known
    - Deviations are computed from a known mean
  - Sample variance as estimate of population
    - Population mean is unknown
    - Using sample mean restricts variability
  - Degrees of freedom
    - Number of scores in sample that are independent and free to vary
    - Degrees of freedom (df) =  $n - 1$
-

# Descriptive Statistics

- A standard deviation describes scores in terms of distance from the mean
- Describe an entire distribution with just two numbers ( $M$  and  $s$ )
- Reference to both allows reconstruction of the measurement scale from just these two numbers
- Means and standard deviations together provide extremely useful descriptive statistics for characterizing distributions



# Interquartile range (IQR)

- Measure of Variation
- Also Known as Midspread: Spread in the Middle 50%
- Difference Between Third & First Quartiles:
- Not Affected by Extreme Values
- Interquartile Range =  $Q_3 - Q_1$

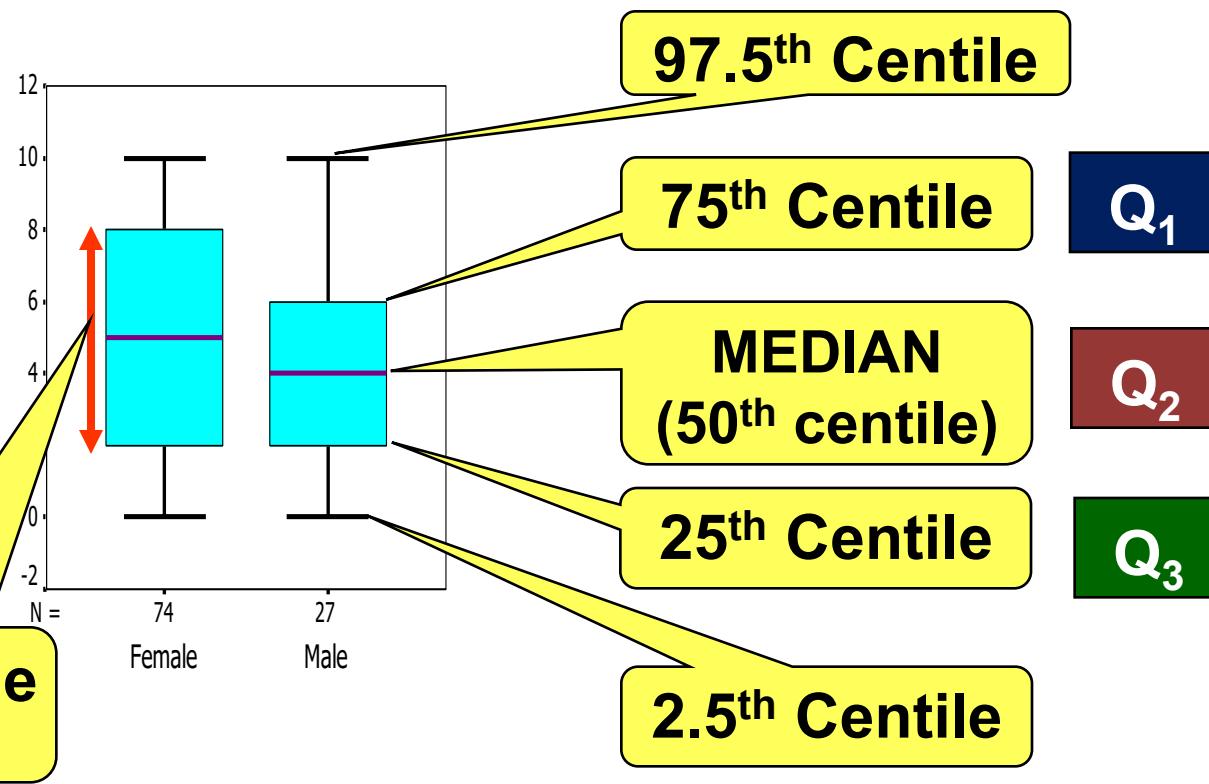
**Data in Ordered Array:** 11 12  $\uparrow$  13 16 16 17 17  $\uparrow$  18 21

$$\text{Position of } Q_1 = \frac{1 \cdot (9 + 1)}{4} = 2.50, \quad Q_1 = 12.5$$

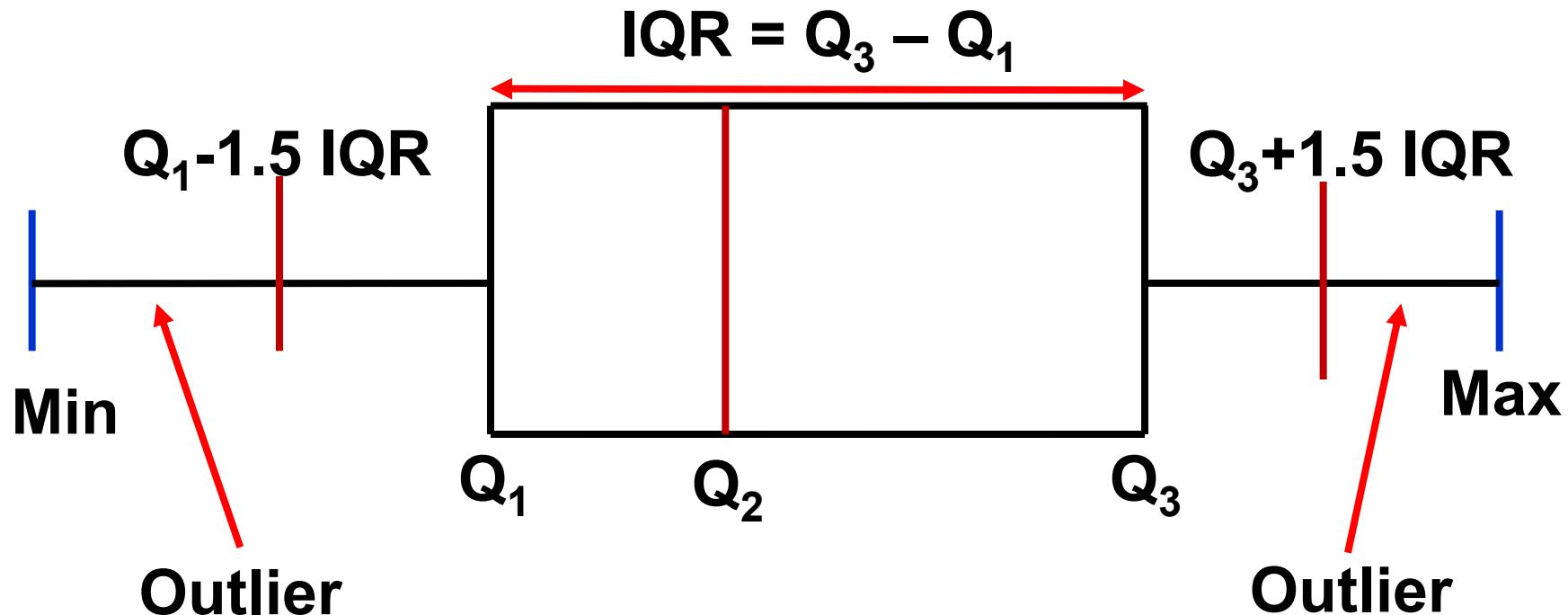
$$\text{Position of } Q_3 = \frac{3 \cdot (9 + 1)}{4} = 7.50, \quad Q_3 = 17.5$$

$$\text{Interquartile Range} = Q_3 - Q_1 = 17.5 - 12.5 = 5$$

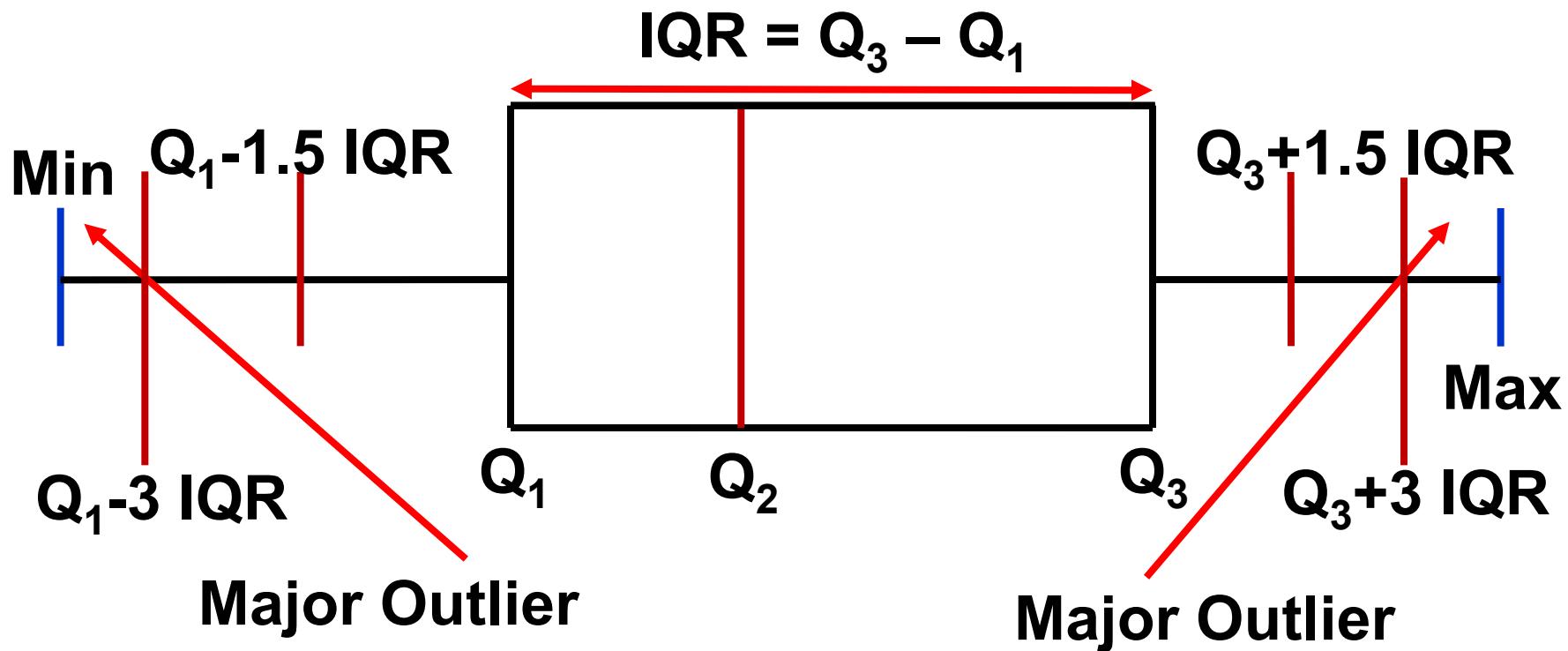
# Box and Whisker plot



# Box and Whisker plot



# Box-and-Whisker plot





# **Session 2**

## **(29<sup>th</sup> January 2022)**

# **Probability & Conditional Probability**

- ❖ Random experiment
- ❖ Sample space
- ❖ Event
- ❖ Types of events
- ❖ Probability

# Example 1

---

- In a certain residential hub, 60% of all households get internet service from the local cable company, 80% get the television service from that company, and 50% get both services from that company.
- If a household is randomly selected, what is the probability that it gets at least one of these two services from the company, and what is the probability that it gets exactly one of these services from the company?

$$P(I) = 60/100 \quad P(T) = 80/100 \quad P(I \cap T) = 50/100 \quad P(I \cup T) = 90/100$$
$$P(\text{EXACTLY ONE}) = \frac{10}{100} + \frac{30}{100}$$

---

# Example 2

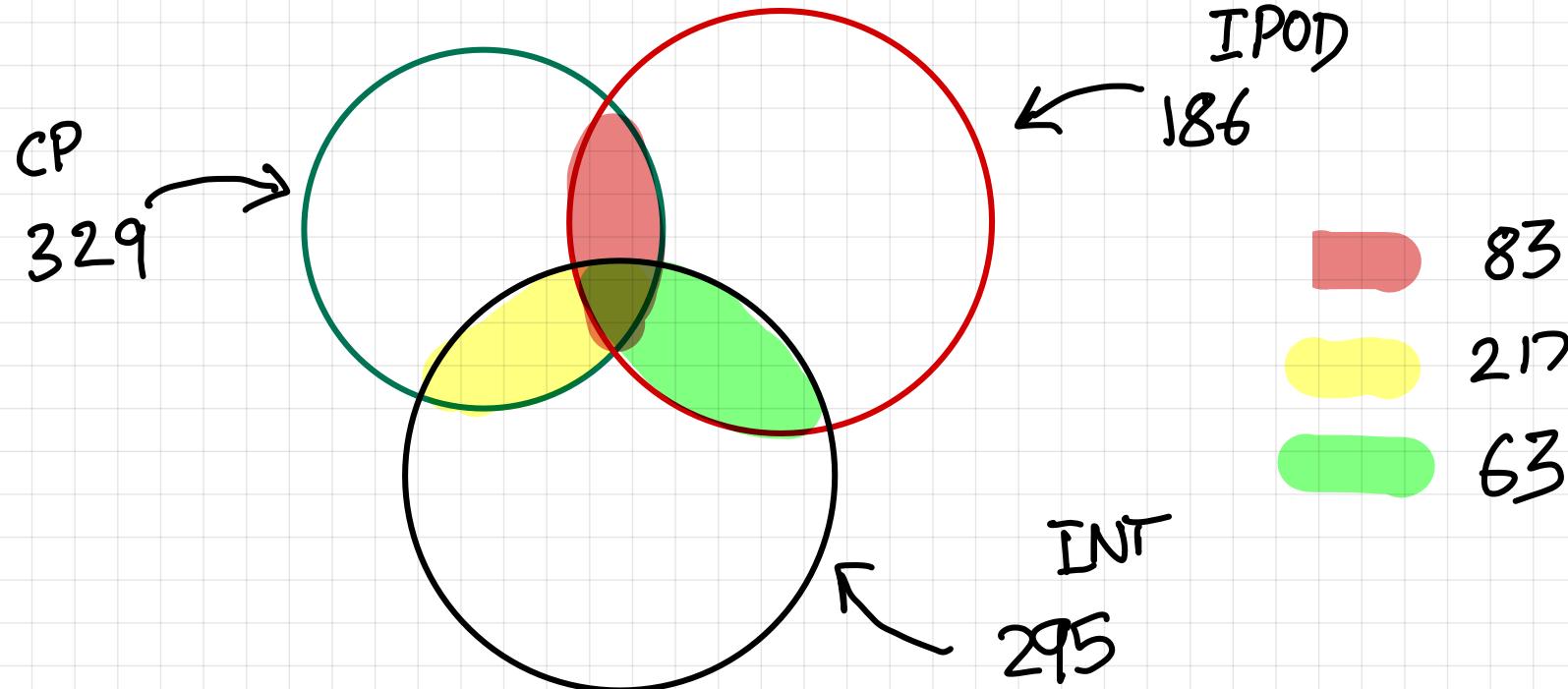
	Blue	Black	Brown	Total
Software prog	35	25	20	80
Project Mgrs	7	8	5	20
Total	42	33	25	100

- If an employee is selected at random , what is the probability that he is a software prog?
- If an employee is selected at random , what is the probability that he is wearing a blue trouser

# Example 3

---

- In a survey of 500 adults were asked the three part question:  
1) Do you own a cell phone 2) Do you own an ipod and 3) Do you have an internet connection? The results of the survey were as follows (no one answered no to all three parts)
  - Cell phone: 329,                      ipod : 186 ,              internet connection:295 ,        Cell phone and ipod:83
  - Cell phone and internet connection: 217,                      ipod and internet connection: 63
  - Find the probabilities of the following events:
-



	83
	217
	63

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

# Ex 3(cont...)

---

- answered yes to all three parts
  - had a cell phone but not an internet connection
  - had an ipod but not a cell phone
  - had an internet connection but not an ipod
  - had a cell phone or an internet connection but not an ipod and
  - had a cell phone but not an ipod or an internet connection.
-

# EXAMPLE 4

---

• A problem in statistics is given to 3 students P, Q and R whose chances of solving it are  $\frac{1}{2}$ ,  $\frac{3}{4}$  and  $\frac{1}{4}$  respectively.

➤ What is the probability that the problem is solved?

$$1 - P(\text{NOT SOLVED}) = 1 - \left(\frac{1}{2} \times \frac{1}{4} \times \frac{3}{4}\right)$$

---

# EXAMPLE 5

---

- A speaks truth in 80% cases and B speaks in 60% cases. What percentage of cases are they likely to contradict each other in stating the same fact.

$$\frac{80}{100} \times \frac{40}{100} + \frac{20}{100} \times \frac{60}{100}$$

---

# EXAMPLE 6

---

- If two dice are thrown , what is the probability that the sum is
    - a) Greater than 8
    - b) Less than 6
    - c) Neither 7 nor 11
-



- If two dice are thrown , what is the probability that the sum is
  - a) Greater than 8
  - b) Less than 6
  - c) Neither 7 nor 11

- If two dice are thrown , what is the probability that the sum is
  - a) Greater than 8
  - b) Less than 6
  - c) Neither 7 nor 11



# Conditional Probability

# Conditional Probability

---

We will use the notation  $P(A | B)$  to represent the **conditional probability of  $A$  given that the event  $B$  has occurred.**  $B$  is the “conditioning event.”

As an example, consider the event  $A$  that a randomly selected student at your university obtained all desired classes during the previous term’s registration cycle. Presumably  $P(A)$  is not very large.

However, suppose the selected student is an athlete who gets special registration priority (the event  $B$ ). Then  $P(A | B)$  should be substantially larger than  $P(A)$ , although perhaps still not close to 1.

---

# The Definition of Conditional Probability

## Definition

For any two events  $A$  and  $B$  with  $P(B) > 0$ , the conditional probability of  $A$  given that  $B$  has occurred is defined by

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad (2.3)$$

# The Multiplication Rule for $P(A \cap B)$

The definition of conditional probability yields the following result, obtained by multiplying both sides of Equation (2.3) by  $P(B)$ .

## The Multiplication Rule

### The Multiplication Rule

$$P(A \cap B) = P(A|B) \cdot P(B)$$

This rule is important because it is often the case that  $P(A \cap B)$  is desired, whereas both  $P(B)$  and  $P(A|B)$  can be specified from the problem description.

Consideration of  $P(B|A)$  as  $P(A \cap B) = P(B|A)$

# Conditional Probability

---

The probability of event B given that event A has occurred  $P(B|A)$  or, the probability of event A given that event B has occurred  $P(A|B)$



# Conditional Probability

---

The probability of event B given that event A has occurred  $P(B|A)$  or, the probability of event A given that event B has occurred  $P(A|B)$



# EXAMPLE 7

---

Toss a six-sided die twice. The sample space consists of all ordered pairs  $(i; j)$  of the numbers  $1; 2; \dots; 6$ , that is,

$$S = \{(1; 1); (1; 2); \dots; (6; 6)\}..$$

Let  $A = \{\text{outcomes match}\}$

and  $B = \{\text{sum of outcomes at least } 8\}$ .

Then find  $P(A), P(B), P(A/B)$  and  $P(B/A)$

---

## EXAMPLE 8

- From a city population , the probability of selecting a male or a smoker is  $7/10$ , a male smoker is  $2/5$ , and a male if a smoker is already selected is  $2/3$ .

- Find the probability of selecting
  - A non – smoker  $1 - P(B)$
  - A male
  - A smoker, if a male is first selected

$$P(A \cup B) = 7/10$$

$$P(A \cap B) = 2/5$$

$$P(A|B) = 2/3$$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

$$P(B) = 3/5$$

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

## Example 9

---

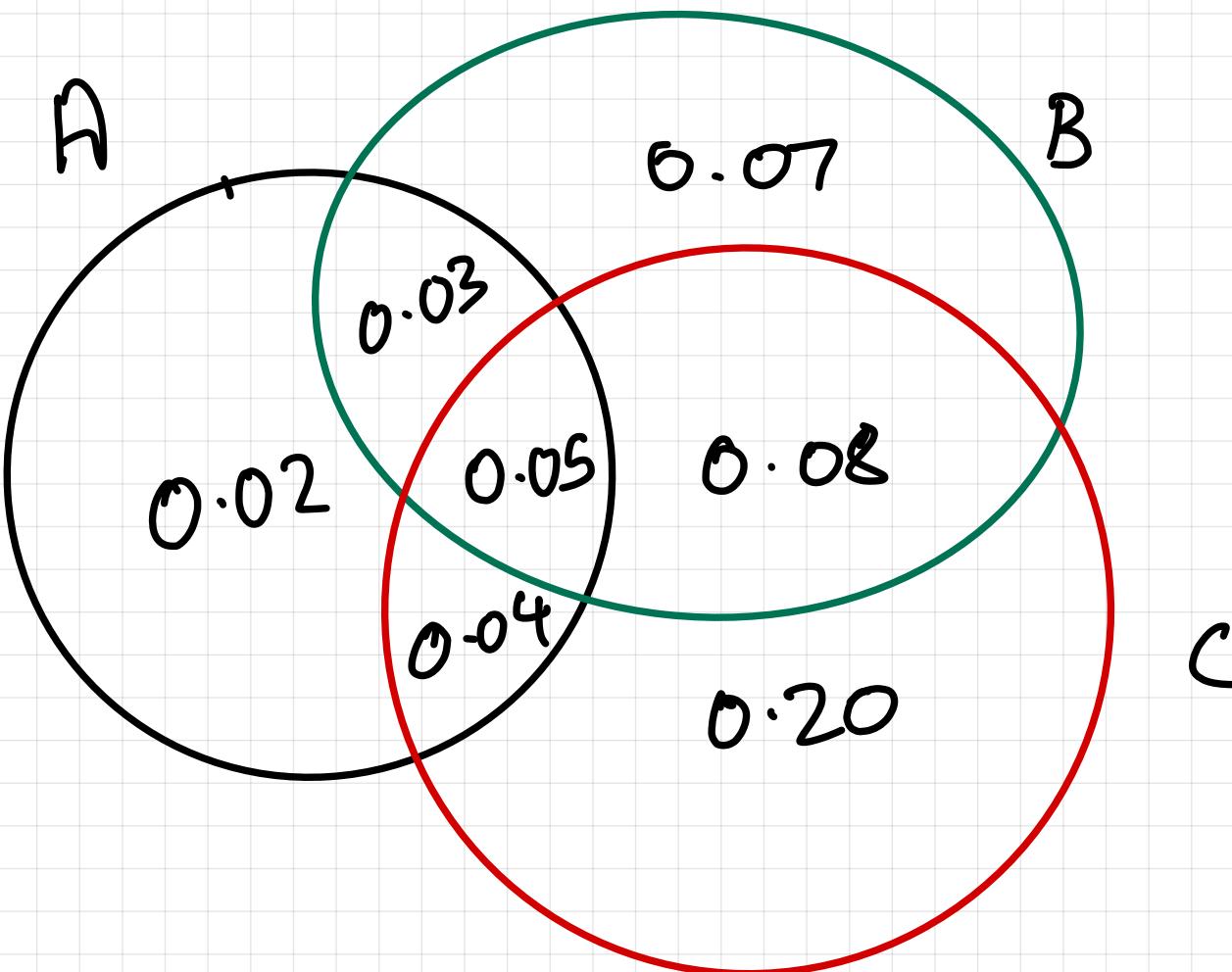
A news magazine publishes three columns entitled “ART”(A), “BOOKS”(B) and “CINEMA:(C). Reading habits of a randomly selected reader with respect to these columns are

- a) Find the probability that he follows ART given that he read BOOKS regularly.
  - b) Find the probability that he follows ART given that he regularly follows at least BOOKS or CINEMA.
  - c) Find the probability that he follows ART given that he regularly follows at least one.
  - d) Find the probability that he follows atleast ART or BOOKS given that he follows cinema regularly.
-

Read Regularly	A	B	C	A & B	A & C	B&C	A&B&C
Probability	0.14	0.23	0.37	0.08	0.09	0.13	0.05

a) Find the probability that he follows ART given that he read BOOKS regularly.

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{0.08}{0.23}$$



$$\begin{aligned}A \cup B \cup C &= 0.37 + 0.07 + 0.03 + 0.02 \\&= 0.49\end{aligned}$$

Read Regularly	A	B	C	A & B	A & C	B&C	A&B&C
Probability	0.14	0.23	0.37	0.08	0.09	0.13	0.05

$$P(B \cup C) = P(B) + P(C) - P(B \cap C) = 0.60 - 0.13 = 0.47$$

$$P(A | B \cup C) = \frac{P(A \cap (B \cup C))}{P(B \cup C)} = \frac{0.12}{0.47}$$

b) Find the probability that he follows ART given that he regularly follows at least BOOKS or CINEMA

c) Find the probability that he follows ART given that he regularly follows at least one.



$$\begin{aligned} P(A | A \cup B \cup C) &= \frac{P(A \cap (A \cup B \cup C))}{P(A \cup B \cup C)} \\ &= \frac{0.14}{0.49} \end{aligned}$$

d) Find the probability that he follows atleast ART or BOOKS given that he follows cinema regularly.

$$P(A \cup B | C) = \frac{P((A \cup B) \cap C)}{P(C)}$$

$$= \frac{0.17}{0.37}$$

## The Law of Total Probability

Let  $A_1, \dots, A_k$  be mutually exclusive and exhaustive events. Then for any other event  $B$ ,

$$\begin{aligned} P(B) &= P(B|A_1)P(A_1) + \cdots + P(B|A_k)P(A_k) \\ &= \sum_{i=1}^k P(B|A_i)P(A_i) \end{aligned} \tag{2.5}$$

# Example 11

---

- Three persons A,B and C are competing for the post of CEO of a company. The chances of they becoming CEO are 0.2,0.3 and 0.4 respectively.
  - The chances of they taking employees beneficial decisions are 0.50,0.45 and 0.6 respectively
  - ❖ What are the chances of having employees beneficial decisions after having new CEO
-

# EXAMPLE 12

---

An individual has 3 different mail accounts. Most of her messages, in fact 70% come into account #1, whereas 20% come into account #2 and the remaining 10% into account #3. Of the messages into account #1, only 1% are spam whereas the corresponding for accounts # 2 and # 3 are 2% and 5% respectively.

What is the probability that a randomly selected message is spam?

---

$$P(S|A) = 1/100$$

$$P(S|B) = 2/100$$

$$P(S|C) = 5/100$$

$$P(S) = P(S|A)P(A) + P(S|B)P(B) + P(S|C)P(C)$$

# SUMMARY





# SUMMARY

## Example 2.30

---

An individual has 3 different email accounts. Most of her messages, in fact 70%, come into account #1, whereas 20% come into account #2 and the remaining 10% into account #3.

Of the messages into account #1, only 1% are spam, whereas the corresponding percentages for accounts #2 and #3 are 2% and 5%, respectively.

What is the probability that a randomly selected spam message is from account #1 ?

---

# Example

In a city patients visits three doctors A,B and C in the ratio 6:5:7.The chance that these doctors refers a case to a specialist are 54%,60% and 55% respectively.

If a patient visits a specialist , what is the probability that he is referred by doctor A

# Bayes' Theorem

## Bayes' Theorem

Let  $A_1, A_2, \dots, A_k$  be a collection of  $k$  mutually exclusive and exhaustive events with *prior* probabilities  $P(A_i)$  ( $i = 1, \dots, k$ ). Then for any other event  $B$  for which  $P(B) > 0$ , the *posterior* probability of  $A_j$  given that  $B$  has occurred is

$$P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^k P(B|A_i) \cdot P(A_i)} \quad j = 1, \dots, k \quad (2.6)$$

# EXAMPLE 1

WEATHER	PLAY	WEATHER	PLAY
SUNNY	YES	SUNNY	YES
OVERCAST	NO	OVERCAST	NO
RAINY	NO	RAINY	NO
SUNNY	YES	SUNNY	YES
SUNNY	NO	RAINY	NO
RAINY	YES	OVERCAST	YES

# Example 2



S. No	Outlook	Temp	Humidity	Windy	Play Tennis
1	Rainy	Hot	High	False	No
2	Rainy	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Sunny	Mild	High	False	Yes
5	Sunny	Cool	Normal	False	Yes
6	Sunny	Cool	Normal	True	No
7	Overcast	Cool	Normal	True	Yes
8	Rainy	Mild	High	False	No
9	Rainy	Cool	Normal	False	Yes
10	Sunny	Mild	Normal	False	Yes
11	Rainy	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Sunny	Mild	High	True	No

today = (Sunny, Hot, Normal, False)

# Example - 3

Magazine Promotion	Watch Promotion	Life Insurance Promotion	Credit Card Insurance	Sex
Yes	No	No	No	Male
Yes	Yes	Yes	Yes	Female
No	No	No	No	Male
Yes	Yes	Yes	Yes	Male
Yes	No	Yes	No	Female
No	No	No	No	Female
Yes	Yes	Yes	Yes	Male
No	No	No	No	Male
Yes	No	No	No	Male
Yes	Yes	Yes	No	Female

New Instance: Magazine Promotion = Yes , Watch Promotion = Yes,  
 Life Insurance Promotion = No, Credit Card Insurance = No then Sex = ?



# Example - 4

	This is my book	Statement
	They are novels	Statement
	Have you read this book	Question
	Who is the author	Question
	What are the characters	Question
	This is how I bought the book	Statement
	I like fictions	Statement
	What is your favourite book	Question





# **Session 3**

## **(05<sup>th</sup> February, 2022)**

# **Bayes Theorem, Random Variables)**

- 
- ❖ Bayes Theorem
  - ❖ Random Variables
  - ❖ Mathematical Expectation
-

# Example 11

---

- Three persons A,B and C are competing for the post of CEO of a company. The chances of they becoming CEO are 0.2,0.3 and 0.4 respectively.
  - The chances of they taking employees beneficial decisions are 0.50,0.45 and 0.6 respectively
  - ❖ What are the chances of having employees beneficial decisions after having new CEO
-

# EXAMPLE 12

---

An individual has 3 different mail accounts. Most of her messages, in fact 70% come into account #1, whereas 20% come into account #2 and the remaining 10% into account #3. Of the messages into account #1, only 1% are spam whereas the corresponding for accounts # 2 and # 3 are 2% and 5% respectively.

What is the probability that a randomly selected message is spam?

---

## Example 2.30

---

An individual has 3 different email accounts. Most of her messages, in fact 70%, come into account #1, whereas 20% come into account #2 and the remaining 10% into account #3.

Of the messages into account #1, only 1% are spam, whereas the corresponding percentages for accounts #2 and #3 are 2% and 5%, respectively.

What is the probability that a randomly selected spam message is from account #1 ?

---

# Example

---

In a city patients visits three doctors A,B and C in the ratio 6:5:7.The chance that these doctors refers a case to a specialist are 54%,60% and 55% respectively.

If a patient visits a specialist , what is the probability that he is referred by doctor A

---

# Bayes' Theorem

## Bayes' Theorem

Let  $A_1, A_2, \dots, A_k$  be a collection of  $k$  mutually exclusive and exhaustive events with *prior* probabilities  $P(A_i)$  ( $i = 1, \dots, k$ ). Then for any other event  $B$  for which  $P(B) > 0$ , the *posterior* probability of  $A_j$  given that  $B$  has occurred is

$$P(A_j|B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^k P(B|A_i) \cdot P(A_i)} \quad j = 1, \dots, k \quad (2.6)$$

# EXAMPLE 1

WEATHER	PLAY	WEATHER	PLAY
SUNNY	YES	SUNNY	YES
OVERCAST	NO	OVERCAST	NO
RAINY	NO	RAINY	NO
SUNNY	YES	SUNNY	YES
SUNNY	NO	RAINY	NO
RAINY	YES	OVERCAST	YES

# Example 2



S. No	Outlook	Temp	Humidity	Windy	Play Tennis
1	Rainy	Hot	High	False	No
2	Rainy	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Sunny	Mild	High	False	Yes
5	Sunny	Cool	Normal	False	Yes
6	Sunny	Cool	Normal	True	No
7	Overcast	Cool	Normal	True	Yes
8	Rainy	Mild	High	False	No
9	Rainy	Cool	Normal	False	Yes
10	Sunny	Mild	Normal	False	Yes
11	Rainy	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Sunny	Mild	High	True	No

today = (Sunny, Hot, Normal, False)

# Example - 3

Magazine Promotion	Watch Promotion	Life Insurance Promotion	Credit Card Insurance	Sex
Yes	No	No	No	Male
Yes	Yes	Yes	Yes	Female
No	No	No	No	Male
Yes	Yes	Yes	Yes	Male
Yes	No	Yes	No	Female
No	No	No	No	Female
Yes	Yes	Yes	Yes	Male
No	No	No	No	Male
Yes	No	No	No	Male
Yes	Yes	Yes	No	Female

New Instance: Magazine Promotion = Yes , Watch Promotion = Yes,  
 Life Insurance Promotion = No, Credit Card Insurance = No then Sex = ?



# Example - 4

	This is my book	Statement
	They are novels	Statement
	Have you read this book	Question
	Who is the author	Question
	What are the characters	Question
	This is how I bought the book	Statement
	I like fictions	Statement
	What is your favourite book	Question





# Random Variables



# Random Variables

---

- A **random variable** is a variable that assumes numerical values associated with the random outcome of an experiment, where one (and only one) numerical value is assigned to each sample point.
-



# Random Variables

---

- A **random variable** is a numerical description of the outcome of an experiment.
  - A random variable can be classified as being either discrete or continuous depending on the numerical values it assumes.
  - A **discrete random variable** may assume either a finite number of values or an infinite sequence of values.
  - A **continuous random variable** may assume any numerical value in an interval or collection of intervals.
-

# Types of random Variables

---

- A **discrete random variable** can assume a countable number of values.
  - ✓ Number of steps to the top of the Eiffel Tower\*
  
- A **continuous random variable** can assume any value along a given interval of a number line.
  - ✓ The time a tourist stays at the top
  - ✓ once s/he gets there



# Types of Random Variables

- **Discrete random variables**
  - Number of sales
  - Number of calls
  - Shares of stock
  - People in line
  - Mistakes per page
- **Continuous random variables**
  - Length
  - Depth
  - Volume
  - Time
  - Weight

# RANDOM VARIABLES

DISCRETE

$$X = 1, 2, 3$$

pmf  $\Rightarrow P(x)$

$$(i) \quad 0 \leq P(x) \leq 1$$

$$(ii) \quad \sum p(x) = 1$$

$$\mu = E(x) = \sum_{i=0}^n x_i p_i$$

CONTINUOUS

$$x \in (-\infty, \infty)$$

pmf  $\Rightarrow f(x)$

$$(i) \quad 0 \leq f(x) \leq 1$$

$$(ii) \quad \int_{-\infty}^{\infty} f(x) dx = 1$$

$$\mu = E(x) = \int x_i f(x_i)$$

# Probability Distributions for Discrete Random Variables



- The **probability distribution** of a discrete random variable is a graph, table or formula that specifies the probability associated with each possible outcome the random variable can assume.
  - $p(x) \geq 0$  for all values of  $x$
  - $\sum p(x) = 1$

# Expected Values of Discrete Random Variables

---

The **mean**, or **expected value**, of a **discrete random variable** is

$$\mu = E(x) = \sum x p(x).$$

# Expected Values of Discrete Random Variables



- The **variance** of a **discrete random variable**  $x$  is

$$\sigma^2 = E[(x - \mu)^2] = \sum (x - \mu)^2 p(x).$$

- The **standard deviation** of a **discrete random variable**  $x$  is

$$\sqrt{\sigma^2} = \sqrt{E[(x - \mu)^2]} = \sqrt{\sum (x - \mu)^2 p(x)}.$$

# Continuous Probability Distributions

---

- **Probability density function**, denoted by  $f(x)$ , which provides the probability for each value of the random variable.
- The required conditions for a discrete probability function are:

$$f(x) \geq 0$$

$$\sum f(x) = 1$$

---





























**Session 4**  
**(12<sup>th</sup> February, 2022)**  
**(Random Variables & Prob.distributions)**



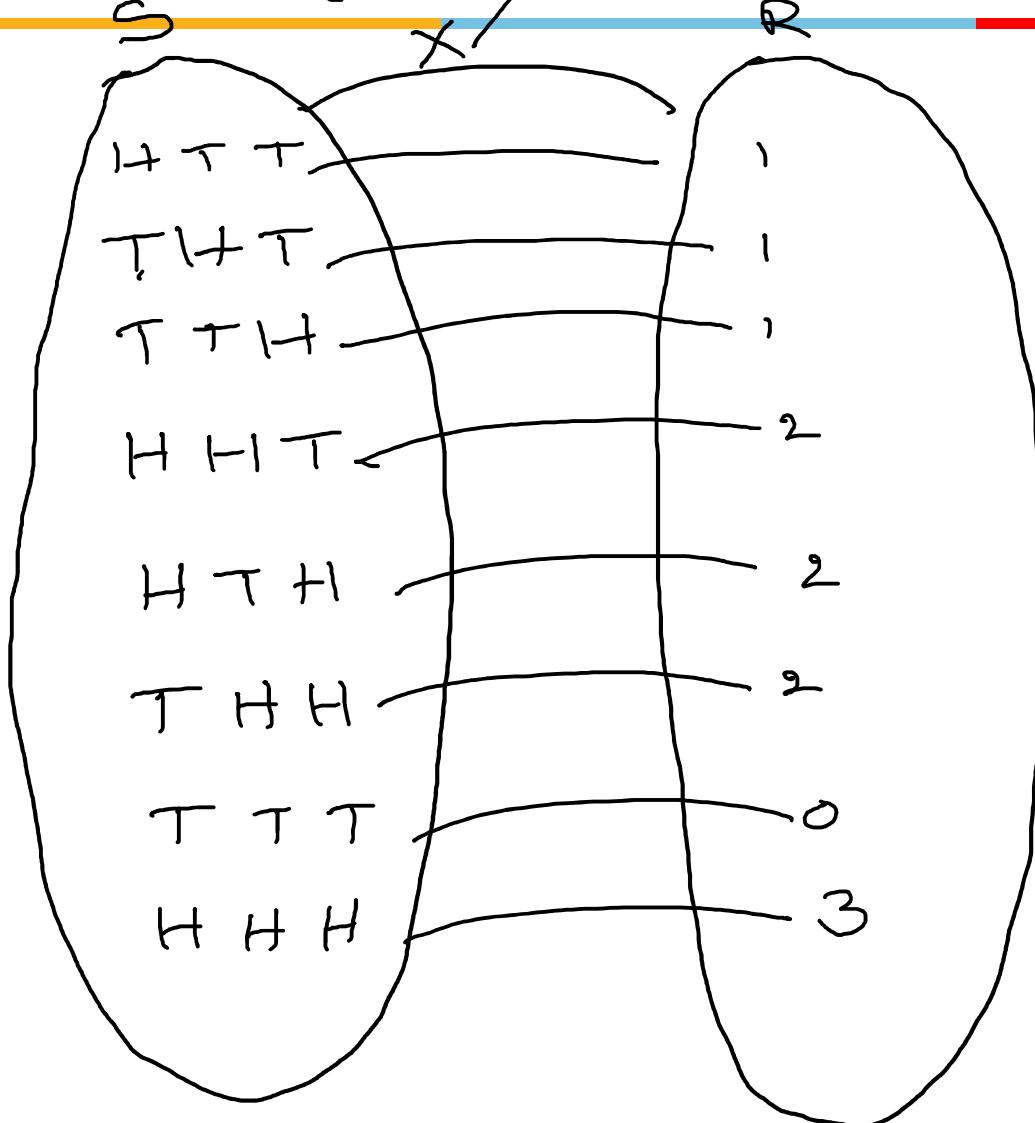
- ❖ Random Variables
- ❖ Mathematical Expectation
- ❖ Mean and Variance
- ❖ Prob.Distributions

Consider

no of Heads



$$S = \{ HHT, THT, TTH, HHT, HTH, THH, TTT, HHH \}$$



$$X : S \rightarrow \mathbb{R}$$

$$\downarrow$$

Sample Space  $\{0, 1, 2, 3\}$

$$X = \{0, 1, 2, 3\}$$



Random  
variable

Consider

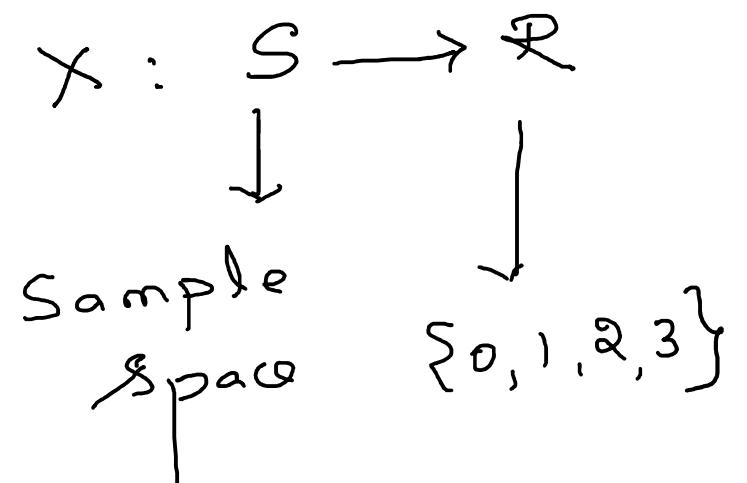
$$S = \{HTT, THT, TTH, HHT, HTH, THH, TTT, HHH\}$$

$X = R.V$ : number of heads

$$X = x, \quad x = 0, 1, 2, 3$$

$X=x$	0	1	2	3
$P(X=x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

probability distribution



$$X = \{0, 1, 2, 3\}$$

Random  
variable  
Discrete

Consider

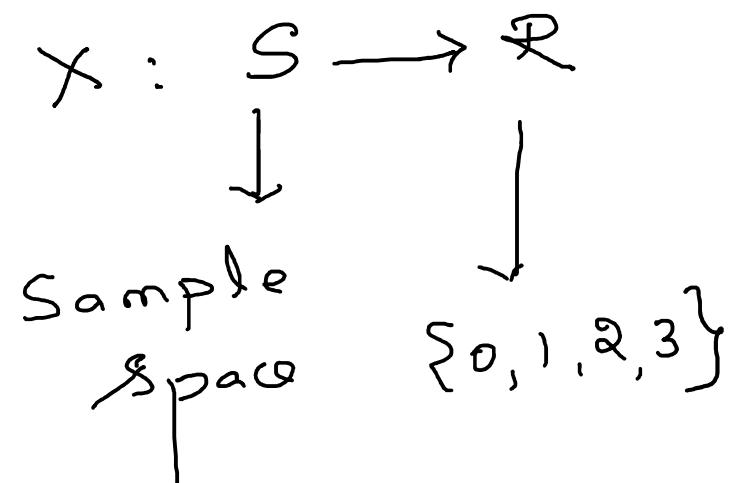
$$S = \{HTT, THT, TTH, HHT, HTH, THH, TTT, HHH\}$$

$X = R.V$ : number of heads

$$X = x, \quad x = 0, 1, 2, 3$$

$X=x$	0	1	2	3
$P(X=x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

probability distribution



$$X = \{0, 1, 2, 3\}$$

Random  
variable  
Discrete



# Probability Distributions for Discrete Random Variables

- The **probability distribution** of a discrete random variable is a graph, table or formula that specifies the probability associated with each possible outcome the random variable can assume.
  - $p(x) \geq 0$  for all values of  $x$
  - $\sum p(x) = 1$



# Expected Values of Discrete Random Variables

The **mean**, or **expected value**, of a **discrete random variable** is

$$\mu = E(x) = \sum x p(x).$$



# Expected Values of Discrete Random Variables

- The **variance** of a **discrete random variable**  $x$  is

$$\sigma^2 = E[(x - \mu)^2] = \sum (x - \mu)^2 p(x).$$

- The **standard deviation** of a **discrete random variable**  $x$  is

$$\sqrt{\sigma^2} = \sqrt{E[(x - \mu)^2]} = \sqrt{\sum (x - \mu)^2 p(x)}.$$

## Random Variables

Discrete

$P(x)$

continuous

$f(x)$

Validation :-

$$1) 0 \leq P(x) \leq 1$$

$$2) \sum P(x) = 1$$

Probability  
distribution  
function

$$1) 0 \leq f(x) \leq 1$$

$$2) \int f(x) dx = 1$$

probability  
density  
function

# Expectation of a random variable

---

$$x_1 \rightarrow P_1$$

then

$$x_2 \rightarrow P_2$$

$$x_1 P_1 + x_2 P_2 + x_3 P_3 + \dots + x_n P_n$$

$$x_3 \rightarrow P_3$$

$$= \sum x_n P_n$$

:

$$x_n \rightarrow P_n$$

$$\text{or } \sum x P(x)$$

is called

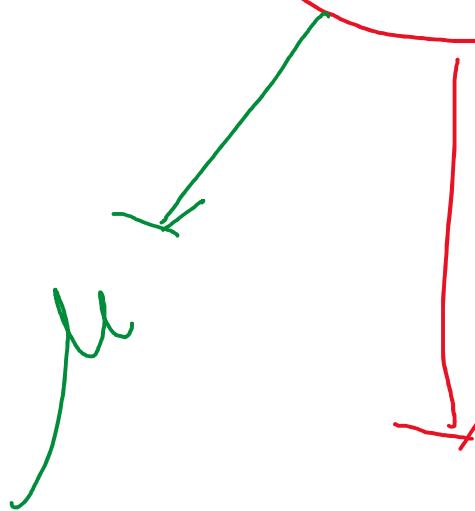
**"Mathematical Expectation" of a random variable  $x$ .**

# Mean of a r.v



Mathematical expectation

$$E(x) = \begin{cases} \sum x p(x) & \text{if } x \text{ is discrete} \\ \int x f(x) dx & \text{if } x \text{ is conti} \end{cases}$$



called as mean of a  
random variable

# Variance of a Random Variable



$$\text{Var}(x) = E(x - \mu)^2$$

$\sigma^2$

$$= E(x^2 - 2\mu x + \mu^2)$$

$$= E(x^2) - 2\mu E(x) + \mu^2$$

$$E(x^2) - [E(x)]^2$$

mean

$$= E(x^2) - 2\mu^2 + \mu^2$$

$$= E(x^2) - \mu^2$$

$$= E(x^2) - [E(x)]^2$$

$$\sum x^2 p(x)$$

discrete

$$\int x^2 f(x) dx$$

continuous

# Example:

$x$	1	2	4	8	16
$P(x)$	0.05	0.10	0.35	0.40	0.10

compute the following

- a)  $E(x)$
- b)  $V(x)$  directly from the definition
- c) standard deviation of  $X$
- d)  $V(x)$  using the shortcut formula

# Example: (Discussion)



$x$	1	2	4	8	16
$P(x)$	0.05	0.10	0.35	0.40	0.10

compute the following

a)  $E(x) = \sum x p(x)$   
 $= 1(0.05) + 2(0.10) + 4(0.35) +$   
 $+ 8(0.40) + 16(0.10)$

$$= 6.45$$

# Example: (Discussion)



$x$	1	2	4	8	16
$P(x)$	0.05	0.10	0.35	0.40	0.10

compute the following

a)  $E(x)$

b)  $V(x)$  directly from the definition

$$\begin{aligned}
 V(x) &= E(x - \mu)^2 = \sum (x - \mu)^2 P(x) \\
 &= (1 - 6.45)^2 (0.05) + (2 - 6.45)^2 (0.10) \\
 &\quad + (4 - 6.45)^2 (0.35) + (8 - 6.45)^2 (0.40) \\
 &\quad + (16 - 6.45)^2 (0.10)
 \end{aligned}$$

# Example:

$x$	1	2	4	8	16
$P(x)$	0.05	0.10	0.35	0.40	0.10

compute the following

$$S.D = \sigma = \sqrt{V(x)} \\ = \sqrt{\text{variance}(x)}$$

c) standard deviation of  $X$

d)  $V(x)$  using the shortcut formula

# Example:

$x$	1	2	4	8	16
$P(x)$	0.05	0.10	0.35	0.40	0.10

compute the following  $\rightarrow$  mean = 6.45

$$\rightarrow V(x) = E(x^2) - [E(x)]^2$$

$$\sum x^2 P(x) = 1^2(0.05) + 2^2(0.10) + 4^2(0.35) \\ + 8^2(0.40) + 16^2(0.10)$$

d)  $V(x)$  using the shortcut formula

# Example

Let  $x$  be a random variable with PDF given by

$$f(x) = \begin{cases} cx^2, & -1 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

- Find constant 'c'.
- Find  $E(x)$  and  $\sigma(x)$
- Find  $P(x \geq y_2)$

# Example

Let  $x$  be a random variable with PDF given by

$$f(x) = \begin{cases} cx^2, & -1 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

a) Find constant 'c'.

$$\int_{-\infty}^{\infty} f(x) dx = 1$$

$$\int_{-1}^1 cx^2 dx = 1 \Rightarrow c \left[ \frac{x^3}{3} \right]_{x=-1}^1 = 1$$

$$\Rightarrow \frac{c}{3} [1^3 - (-1)^3] = 1 \Rightarrow \frac{c}{3} \times 2 = 1$$

ie  $C = 3/2$

# Example

Let  $x$  be a random variable with PDF given by

$$f(x) = \begin{cases} cx^2, & -1 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

$$\int_{-1}^1 x f(x) dx = \int_{-1}^1 x (cx^2) dx = \frac{3}{2} \left( \frac{x^4}{4} \right) \Big|_{x=-1}^{x=1} = \frac{3}{8} (1 - 1) = 0$$

b) Find  $E(x)$  and  $\sigma(x)$

$$E(x^2) - [E(x)]^2$$

$$\downarrow \quad \downarrow$$

$$= \frac{3}{5} - 0 = \frac{3}{5}$$

$$E(x^2) = \int_{-1}^1 x^2 (cx^2) dx$$

$$= \frac{3}{2} \left( \frac{x^5}{5} \right) \Big|_{x=-1}^{x=1}$$

$$= \frac{3}{10} (15 - (-1)^5) = \frac{3}{5}$$

# Example

Let  $x$  be a random variable with PDF given by

$$f(x) = \begin{cases} cx^2, & -1 \leq x \leq 1 \\ 0, & \text{otherwise} \end{cases}$$

$$\begin{aligned}
 \rightarrow &= \int_{y_2}^1 f(x) dx = \frac{3}{2} \int_{y_2}^1 x^2 dx = \frac{3}{2} \left[ \frac{x^3}{3} \right]_{x=y_2}^1 \\
 &= \frac{3}{6} \left[ 1^3 - (y_2)^3 \right] \\
 &= \frac{3}{6} (1 - y_2^3) \\
 &= \frac{3}{6} (1 - \frac{7}{8}) \\
 &= \frac{3}{6} \times \frac{1}{8} \approx \frac{7}{16}
 \end{aligned}$$

# Bernoulli Distribution

## Definition

A random variable 'X' is said to have Bernoulli distribution if its probability mass function is given by

$$p(x) = \begin{cases} p^x q^{1-x}, & x = 0, 1 \\ 0, & \text{elsewhere} \end{cases}$$

# Mean & Variance

$$\text{mean} = \mu = E(x) = \sum x p(x)$$

$$= \sum x p^x q^{1-x}, \quad x = 0, 1$$

$$= 0 \cdot p^0 \cdot q^1 + 1 \cdot p \cdot q^0 = p$$

$$\text{variance } \sigma^2 = E(x^2) - [E(x)]^2$$

$$(E(x^2)) = \sum x^2 p(x) = \sum x^2 \cdot p^x q^{1-x}, \quad x = 0, 1$$

$$= 0 p^0 q^{1-0} + 1^2 \cdot p^1 \cdot q^{1-1} = p$$

$$\sigma^2 = p - p^2 = p(1-p) = pq$$

Mean =  $p$

Variance =  $pq$

# Binomial Distribution

## Binomial trials:

- number of trials ( $n$ ) is fixed
- each trial results in either success or failure
- trials are independent
- probability of success (or failure) remains constant from trial to trial

Total 'n' trials for ex: 10



4) 'x' are successful with prob. p

$(P \cdot P \cdot P \cdot P \dots x \text{ times}) (q \cdot q \cdot q \dots n-x \text{ times})$

$$\begin{matrix} P^x & q^{n-x} \\ \downarrow & \\ mC_x \end{matrix}$$

i.e.

$$mC_x P^x q^{n-x}$$

where  $x = 0, 1, 2, \dots, n$

# Binomial Distribution

## Definition

A random variable 'X' is said to have Binomial distribution if its probability mass function is given by

$$p(x) = \begin{cases} \binom{n}{x} p^x q^{n-x}, & x = 0, 1, 2, 3, \dots, n \\ 0, & \text{elsewhere} \end{cases}$$

# Binomial Distribution

If a coin is tossed 6 times, what is the probability of getting 2 or fewer heads?

$$P(X \leq 2) = \sum p(x) = 0.015625 + 0.09375 + 0.078125 = 0.1875$$

$$P(X = 0) = \binom{6}{0} (0.5)^0 (0.5)^6 = \frac{6!}{6! 0!} (0.5)^6 = 0.015625$$

$$P(X = 1) = \binom{6}{1} (0.5)^1 (0.5)^5 = \frac{6!}{5! 1!} (0.5)^6 = 0.09375$$

$$P(X = 2) = \binom{6}{2} (0.5)^1 (0.5)^4 = \frac{6!}{4! 2!} (0.5)^6 = 0.078125$$

# Example

---

The probability that a man aged 60 years will remain alive till 70 is 0.65.

- ❖ What is the probability that out of 10 such men at least 7 would be alive at 70?

# Solution

$$n = 10, \quad p = 0.65$$

$$q = 1 - p = 0.35$$

$$\begin{aligned}
 P(x \geq 7) &= P(7) + P(8) + P(9) + P(10) \\
 &= {}^{10}C_7 (0.65)^7 (0.35)^3 + {}^{10}C_8 (0.65)^8 (0.35)^2 \\
 &\quad + {}^{10}C_9 (0.65)^9 (0.35) + {}^{10}C_{10} (0.65)^{10} (0.35)^0 \\
 &= 0.515
 \end{aligned}$$

# Example

---

Nationalized bank sanctions short term loan for its customers. The probability of sanctioning loan by the bank is 0.53. If 5 persons are selected at random what is the probability that

- a) none were sanctioned loan?
  - b) between 1 to 4 were sanctioned the loan?
  - c) more than 3 were sanctioned the loan?
-

# Mean & variance

$$P(x) = nC_x P^x q^{n-x}, \quad x = 0, 1, 2, \dots, n$$

$$\text{mean} = \mu = E(x) = \sum x P(x)$$

$$= \sum x nC_x P^x q^{n-x} = \sum x \frac{n!}{x! (n-x)!} P^x q^{n-x}$$

$$= \sum \frac{n!}{(x-1)! (n-x)!} P^x q^{n-x}$$

$$= np \sum_{x=1}^n \frac{(n-1)!}{(x-1)! [(n-1)-(x-1)]!} P^{x-1} q^{(n-1)-(x-1)}$$

$$= np (q+p)^{n-1}$$

$$(q+p)^n = nC_0 q^n p^0 + nC_1 q^{n-1} p^1 + \dots + nC_n q^0 p^n$$

$$= np$$

$$\text{Variance } \sigma^2 = E(x^2) - [E(x)]^2$$



$$\begin{aligned}
 E(x^2) &= \sum x^2 p(x) = \sum x^2 n c_x p^x q^{n-x} \\
 &= \sum [x(x-1) + x] n c_x p^x q^{n-x} \quad \text{mean : } np \\
 &= \sum x(x-1) \cdot n c_x p^x q^{n-x} + \sum x \cancel{n c_x p^x q^{n-x}} \quad \sum x p(x) \\
 &= \sum x(x-1) \frac{n!}{x! (n-x)!} p^x q^{n-x} + np \\
 &= \sum \frac{n(n-1)(n-2)!}{(x-2)![n-(x-2)]!} p^2 p^{(n-2)-(x-2)} q^{n-x} + np \\
 &= n(n-1) p^2 \sum (n-2) c_{x-2} p^x q^{n-x} + np \\
 &= n(n-1) p^2 + np
 \end{aligned}$$

$$\begin{aligned}
 \text{Variance} &= \sigma^2 = n(n-1) p^2 + np - (np)^2 \\
 &= n^2 p^2 - mp^2 + np - n^2 p^2 \\
 &= np(1-p) = npq
 \end{aligned}$$

# Binomial Distribution

All probability distributions are characterized by an expected value and a variance

If  $X$  follows a binomial distribution with parameters  $n$  and  $p$  then we write  $X \sim B(n, p)$

$$\mu = E(X) = np$$

$$\sigma^2 = \text{Var}(X) = npq, q = 1 - p$$

$$\sigma = \text{SD}(X) = \sqrt{npq}$$

Note: the variance will always lie between

$$0*n - 0.25*n$$

$p(1 - p)$  reaches maximum at  $p=0.5$

$$p(1 - p)=.025$$

# Binomial Distribution

## Definition

A random variable 'X' is said to have Binomial distribution if its probability mass function is given by

$$p(x) = \begin{cases} \binom{n}{x} p^x q^{n-x}, & x = 0, 1, 2, 3, \dots, n \\ 0, & \text{elsewhere} \end{cases}$$

# Binomial Distribution

All probability distributions are characterized by an expected value and a variance

If  $X$  follows a binomial distribution with parameters  $n$  and  $p$  then we write  $X \sim B(n, p)$

$$\mu = E(X) = np$$

$$\sigma^2 = \text{Var}(X) = npq, q = 1 - p$$

$$\sigma = \text{SD}(X) = \sqrt{npq}$$

Note: the variance will always lie between

$$0*n - 0.25*n$$

$p(1 - p)$  reaches maximum at  $p=0.5$

$$p(1 - p)=.025$$

# Poisson Distribution

## Definition

A random variable 'X' is said to have Poisson distribution if its probability mass function is given by

$$p(x) = \begin{cases} e^{-\lambda} \frac{\lambda^x}{x!}, & x = 0, 1, 2, \dots \\ 0, & \text{elsewhere} \end{cases}$$

**Example 3.39** Let  $X$  denote the number of creatures of a particular type captured in a trap during a given time period. Suppose that  $X$  has a Poisson distribution with  $\mu = 4.5$ , so on average traps will contain 4.5 creatures. [The article “Dispersal Dynamics of the

Bivalve *Gemma Gemma* in a Patchy Environment” (*Ecological Monographs*, 1995: 1–20) suggests this model; the bivalve *Gemma gemma* is a small clam.] The probability that a trap contains exactly five creatures is

$$P(X = 5) = \frac{e^{-4.5}(4.5)^5}{5!} = .1708$$

The probability that a trap has at most five creatures is

$$P(X \leq 5) = \sum_{x=0}^5 \frac{e^{-4.5}(4.5)^x}{x!} = e^{-4.5} \left[ 1 + 4.5 + \frac{(4.5)^2}{2!} + \cdots + \frac{(4.5)^5}{5!} \right] = .7029 \blacksquare$$



## Problem

Average number of accidents on any day on a national highway is 1.8 . Determine the probability that the number of accidents are

- 1) at least one
- 2) at most one.



# Solution

## Solution

$$\lambda = 1.8$$

$$1) P(X \geq 1) = 1 - P[X = 0]$$

$$= 1 - \frac{e^{-\lambda} \lambda^0}{0!} = 1 - \frac{e^{-1.8} 1.8^0}{0!} = 1 - e^{-1.8}$$

$$2) P[X \leq 1] = P[X = 0] + P[X = 1] = \frac{e^{-1.8} 1.8^0}{0!} + \frac{e^{-1.8} 1.8^1}{1!}$$

# Poisson distribution

$$P(x) = \frac{\bar{e}^\lambda \lambda^x}{x!}, \quad x = 0, 1, 2, \dots$$

$$\text{mean } \mu = E(x) = \sum x P(x)$$

$$= \sum x \frac{\bar{e}^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots$$

$$= \bar{e}^{-\lambda} \sum \frac{\lambda^x}{(x-1)!} = \lambda \bar{e}^{-\lambda} \sum \frac{\lambda^{x-1}}{(x-1)!}$$

$$= \lambda \bar{e}^{-\lambda} \left[ 1 + \lambda + \frac{\lambda^2}{2!} + \dots \right]$$

$$= \lambda \bar{e}^{-\lambda} \cdot \bar{e}^\lambda$$

$$= \lambda$$

$$\therefore \text{mean} = \mu = \lambda$$

$$\text{Variance} = \sigma^2 = E(x^2) - [E(x)]^2$$



$$E(x^2) = \sum x^2 \frac{\bar{e}^\lambda \lambda^x}{x!}$$

$$= \sum [x(x-1) + x] \frac{\bar{e}^\lambda \lambda^x}{x!}$$

$$= \sum x(x-1) \frac{\bar{e}^\lambda \lambda^x}{x!} + \sum x \frac{\bar{e}^\lambda \lambda^x}{x!}$$

$$= \bar{e}^\lambda \sum \frac{\lambda^x}{(x-2)!} + \lambda$$

$$= \bar{e}^\lambda \left[ \lambda^2 + \lambda^3 + \frac{\lambda^4}{2!} + \dots \right] + \lambda$$

$$= \lambda^2 \bar{e}^\lambda \left[ 1 + \lambda + \frac{\lambda^2}{2!} + \dots \right] + \lambda$$

$$= \lambda^2 \bar{e}^\lambda \cdot \bar{e}^\lambda + \lambda = \lambda^2 + \lambda$$

$$\therefore \text{Variance} = \lambda^2 + \lambda - (\lambda)^2 = \lambda$$

$$\boxed{\text{Mean} = \text{Variance} = \lambda}$$

# Example

If the probability of a bad reaction from a certain injection is 0.001.

- ❖ Determine the chance that out of 2000 individuals more than two will get a bad reaction.

# Solution

$$n = 2000 ; P = 0.001$$

$$\lambda = np = 2000 \times 0.001 = 2$$

$$\begin{aligned}
 P(X > 2) &= 1 - P(X \leq 2) \\
 &= 1 - [P(0) + P(1) + P(2)] \\
 &= 1 - \left[ \frac{\bar{e}^\lambda \lambda^0}{0!} + \frac{\bar{e}^\lambda \lambda^1}{1!} + \frac{\bar{e}^\lambda \lambda^2}{2!} \right] \\
 &= 1 - \bar{e}^\lambda \left[ 1 + \lambda + \frac{\lambda^2}{2} \right] \\
 &= 1 - \bar{e}^2 (1 + 2 + 2) \\
 &= 0.3233
 \end{aligned}$$

# Normal Distribution

$$f(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi} \sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where  $-\infty < x < \infty$

Suppose  $x$  is a continuous

$\text{r.v}$  which follows normal distribution

Now, we want to find

$$P(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} f(x) dx$$

$$= \int_{x_1}^{x_2} \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

=

Suppose  $x$  is a continuous

$\text{r.v}$  which follows normal distribution

Now, we want to find

$$P(x_1 \leq x \leq x_2) = \int_{x_1}^{x_2} f(x) dx$$

$$= \int_{x_1}^{x_2} \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Let  $\frac{x-\mu}{\sigma} = z$

$$= \int_{z_1}^{z_2} \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{z^2}{2}} \cdot \sigma dz$$

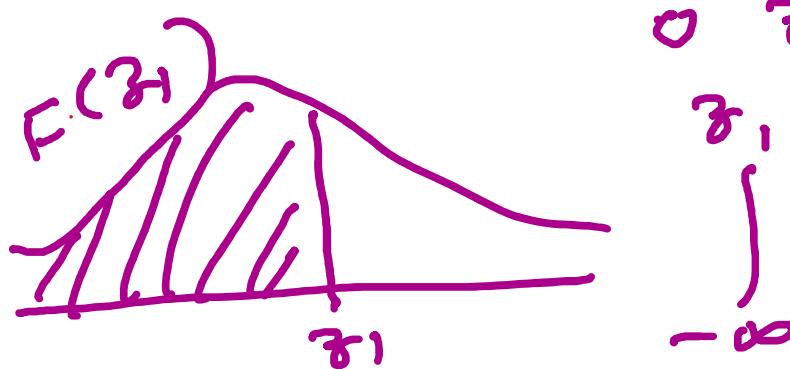
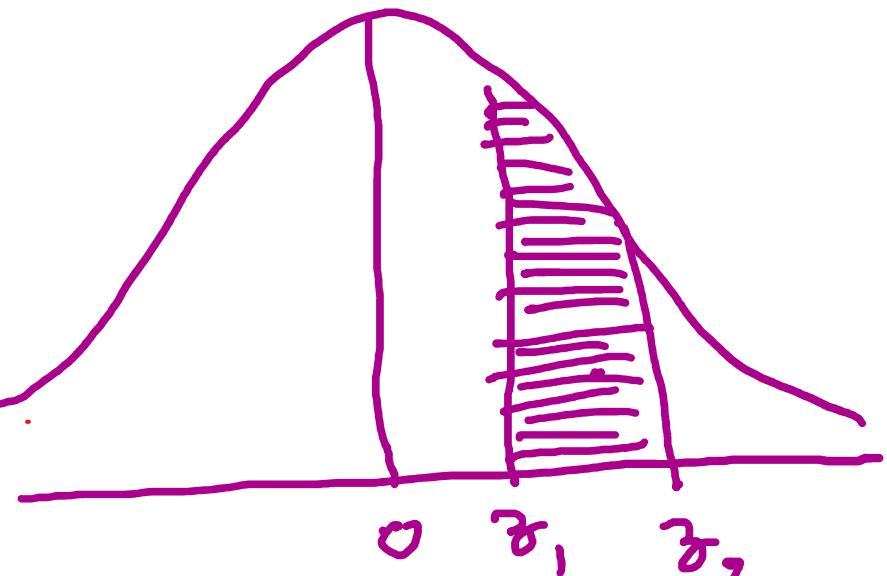
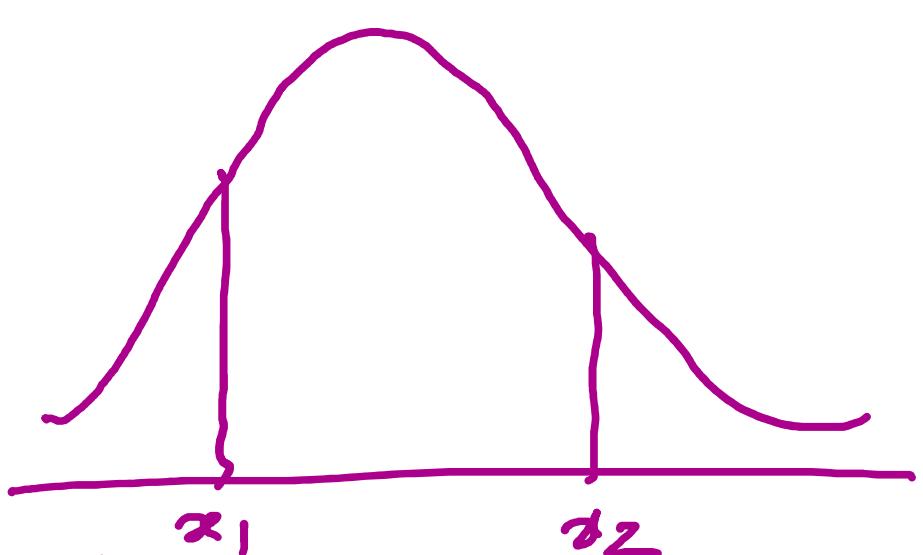
$$x - \mu = \sigma z$$

$$dx = \sigma dz$$

$$= \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

$z_1$        $z_2$

$$= F(z_2) - F(z_1)$$



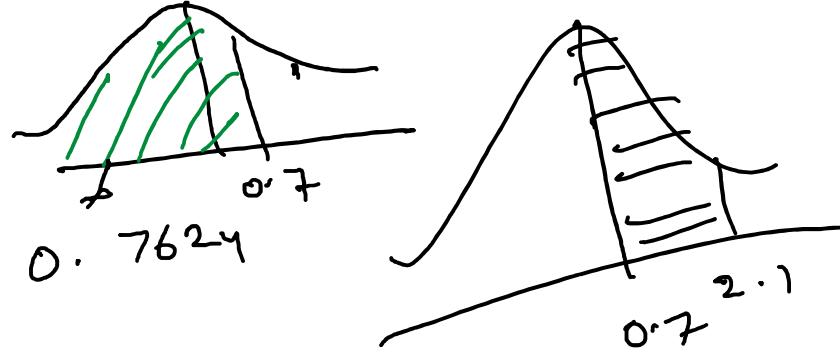
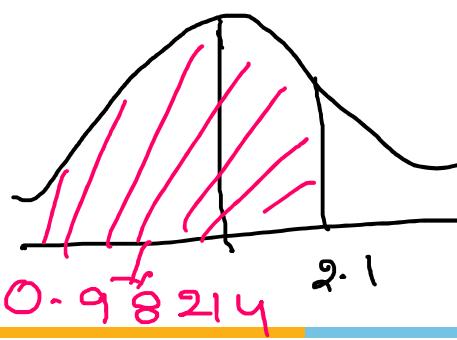
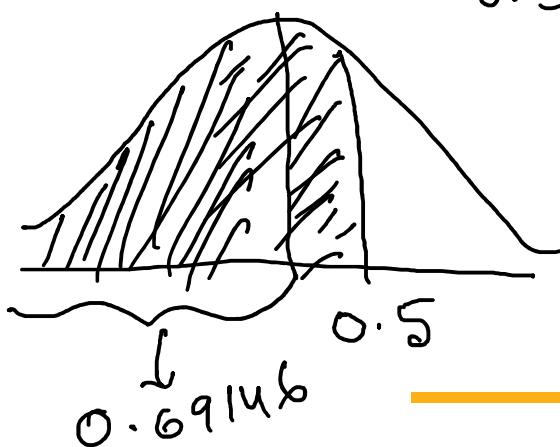


**STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.**

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	<b>.69146</b>	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	<b>.73237</b>	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	<b>.76424</b>	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	<b>.98214</b>	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361

$$0.5 \rightarrow 0.69146$$

$$P(-2.1 \leq z \leq 0.72)$$



## Example :- (from T<sub>1</sub>)



The time it takes a driver to react to the brake lights on a decelerating vehicle is critical in helping to avoid rear-end collisions. It is suggested that reaction time for an in traffic response to a brake signal from a standard brake lights can be modeled with a normal distribution having mean 1.25 sec and S.D of 0.46 sec. What is the probability that reaction time is between 1.00 sec and 1.75 sec?

$$P(1.00 \leq x \leq 1.75)$$

## Example :- (from T<sub>1</sub>)



The time it takes a driver to react to the brake lights on a decelerating vehicle is critical in helping to avoid rear-end collisions. It is suggested that reaction time for an in traffic response to a brake signal from a standard brake lights can be modeled with a normal distribution having mean 1.25 sec and S.D of 0.46 sec. What is the probability that reaction time is between 1.00 sec and 1.75 sec? ie  $P(1.00 \leq X \leq 1.75)$

✓ when  $x=1$ ,  $\therefore \frac{x-\mu}{\sigma} = \frac{1-1.25}{0.46} = -0.54$

✓ when  $x=1.75$   $\therefore \frac{1.75-1.25}{0.46} = 1.09$

$$P(1.00 \leq x \leq 1.75)$$

$$P(-0.54 \leq z \leq 1.09)$$

$$= F(1.09) - F(-0.54)$$

$$= F(1.09) - [1 - F(0.54)]$$

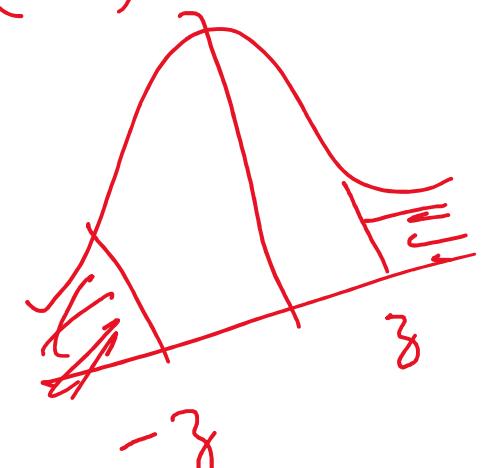
↓

0.8621

0.2946

$$= \underline{\underline{0.5675}}$$

$$R(-z) = 1 - F(z)$$

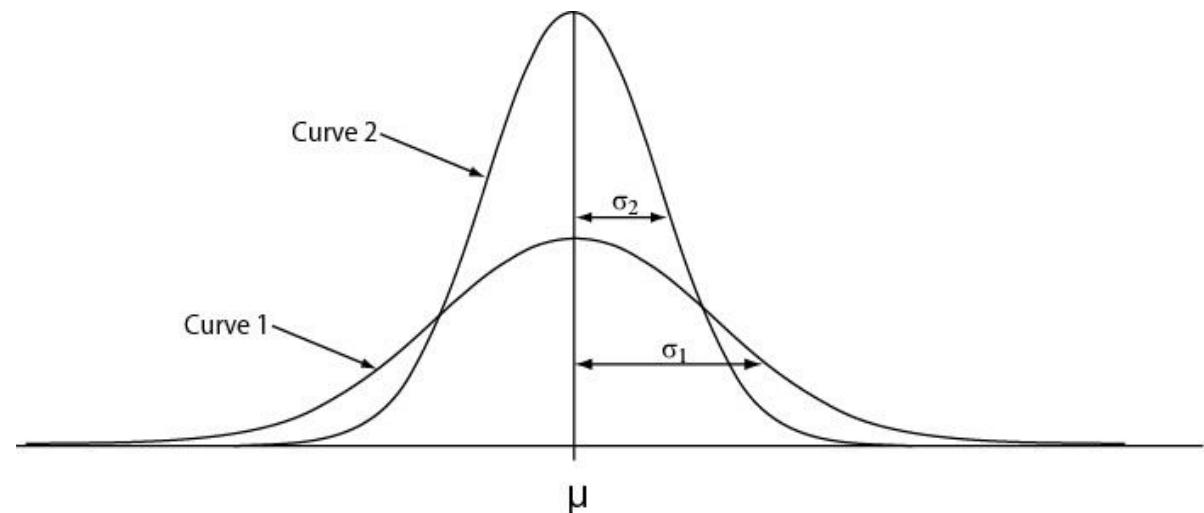
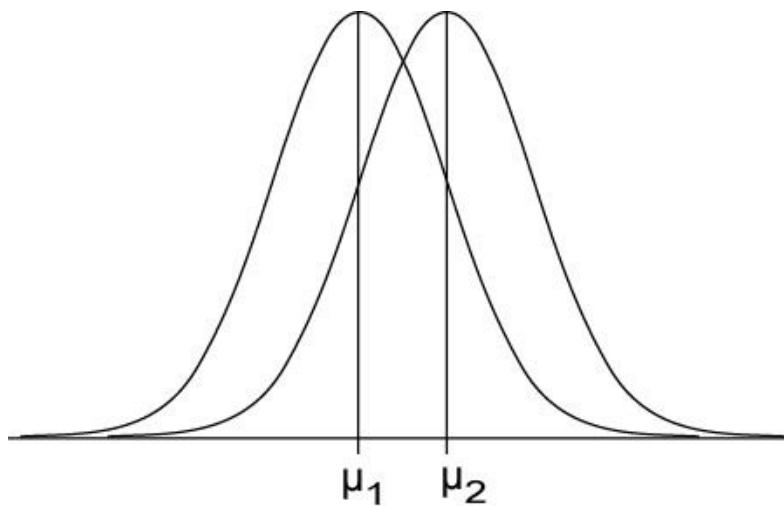


# Parameters $\mu$ and $\sigma$

- Normal *pdfs* have two **parameters**

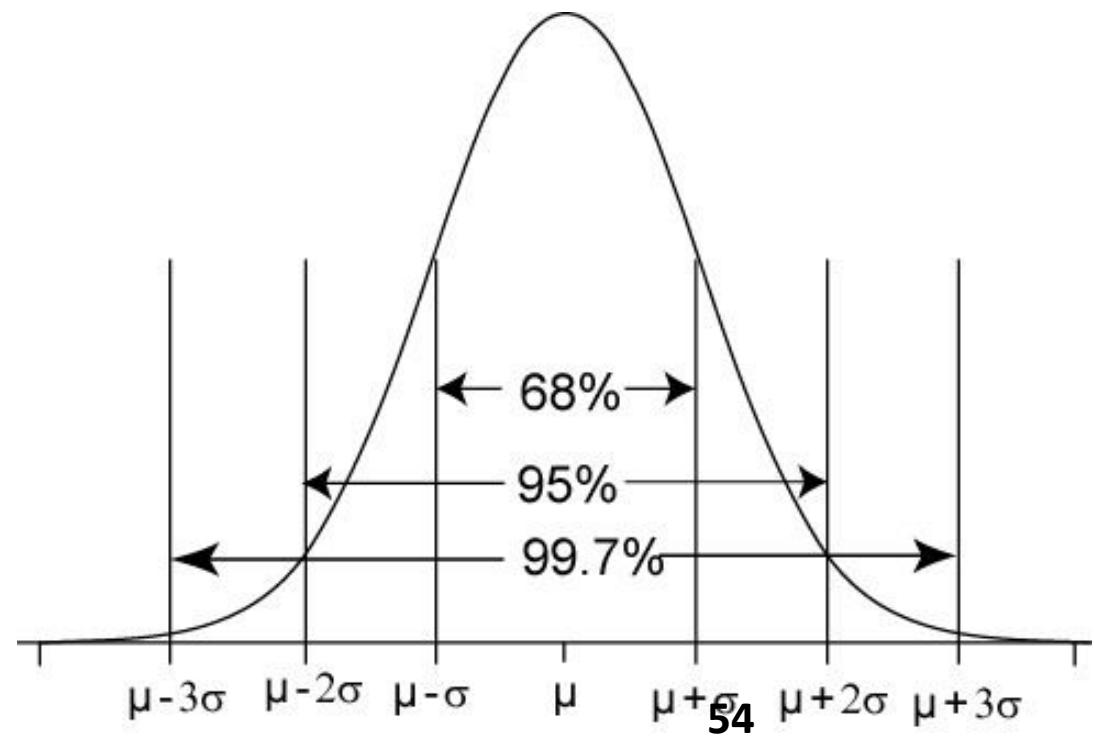
$\mu$  - expected value (mean “mu”)

$\sigma$  - standard deviation (sigma)  
 $\mu$  controls location                     $\sigma$  controls spread



# 68-95-99.7 Rule for Normal Distributions

- **68%** of the AUC within  $\pm 1\sigma$  of  $\mu$
- **95%** of the AUC within  $\pm 2\sigma$  of  $\mu$
- **99.7%** of the AUC within  $\pm 3\sigma$  of  $\mu$



**EXAMPLE: *Finding the Area Under the Standard Normal Curve***

- Find the area under the standard normal curve to the right of  $Z = 1.25$ .

---

**EXAMPLE:** *Finding the Area Under the Standard Normal Curve*

Find the area under the standard normal curve between  $Z = -1.02$  and  $Z = 2.94$ .

$$\begin{aligned} P(-1.02 < x < 2.94) &= P(x < 2.94) - P(x < -1.02) \\ &= 0.9984 - 0.1539 \\ &= 0.8445 \end{aligned}$$

---

# Properties of Normal Distributions

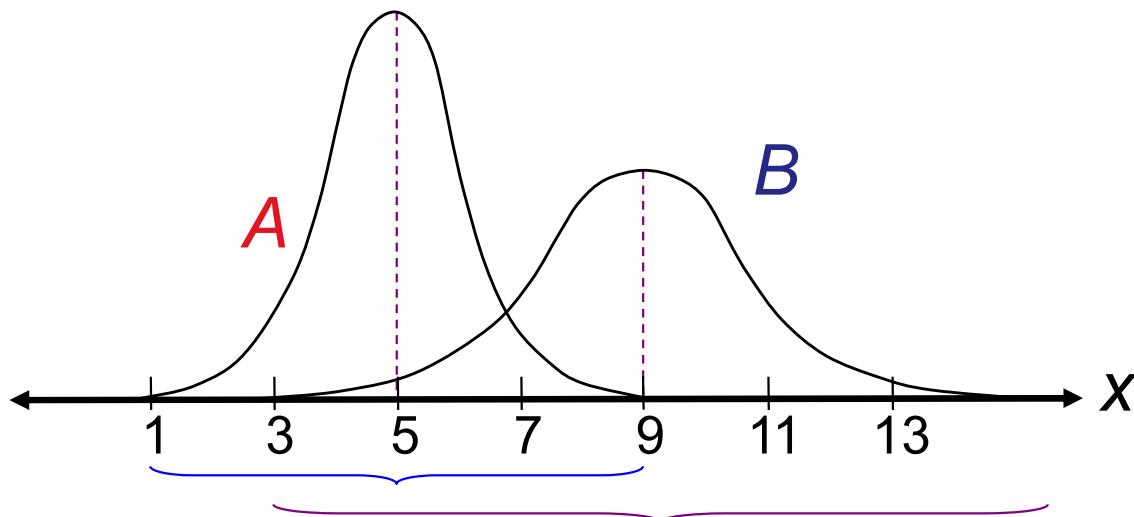
## Properties of a Normal Distribution

1. The mean, median, and mode are equal.
2. The normal curve is bell-shaped and symmetric about the mean.
3. The total area under the curve is equal to one.
4. The normal curve approaches, but never touches the  $x$ -axis as it extends farther and farther away from the mean.

# Means and Standard Deviations

Example:

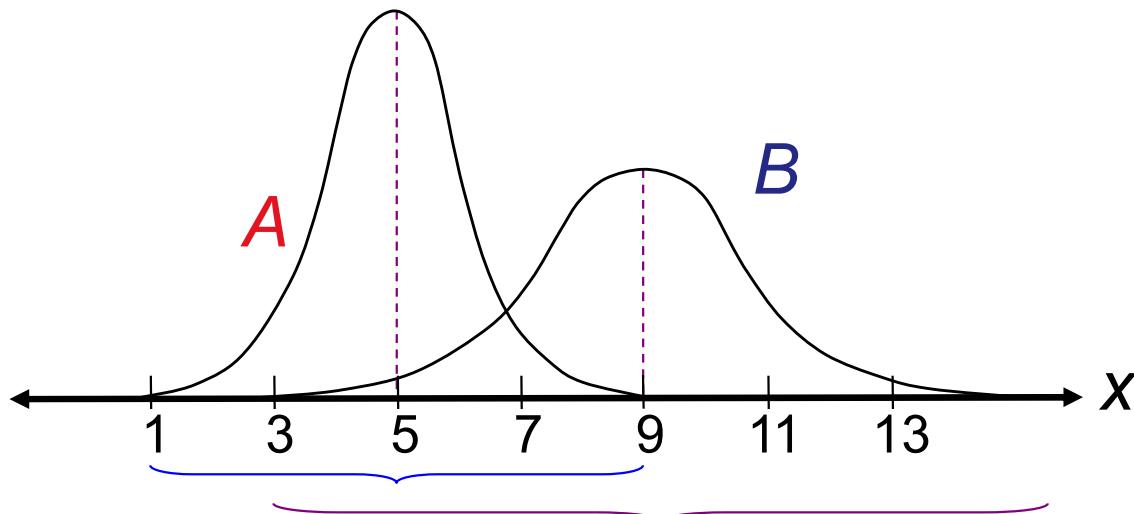
1. Which curve has the greater mean?
2. Which curve has the greater standard deviation?



# Means and Standard Deviations

Example:

1. Which curve has the greater mean?
2. Which curve has the greater standard deviation?



The line of symmetry of curve A occurs at  $x = 5$ . The line of symmetry of curve B occurs at  $x = 9$ . Curve B has the greater mean.

Curve B is more spread out than curve A, so curve B has the greater standard deviation.

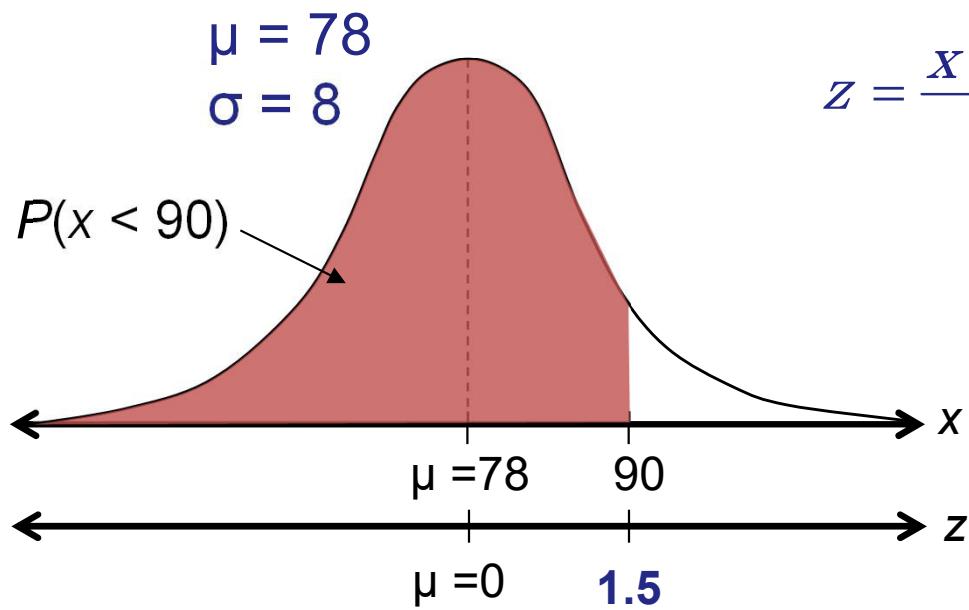


## Example:

The average on a statistics test was 78 with a standard deviation of 8. If the test scores are normally distributed, find the probability that a student receives a test score less than 90.

## Example:

The average on a statistics test was 78 with a standard deviation of 8. If the test scores are normally distributed, find the probability that a student receives a test score less than 90.



$$z = \frac{x - \mu}{\sigma} = \frac{90 - 78}{8} = 1.5$$

The probability that a student receives a test score less than 90 is 0.9332.

$$P(x < 90) = P(z < 1.5) = 0.9332$$

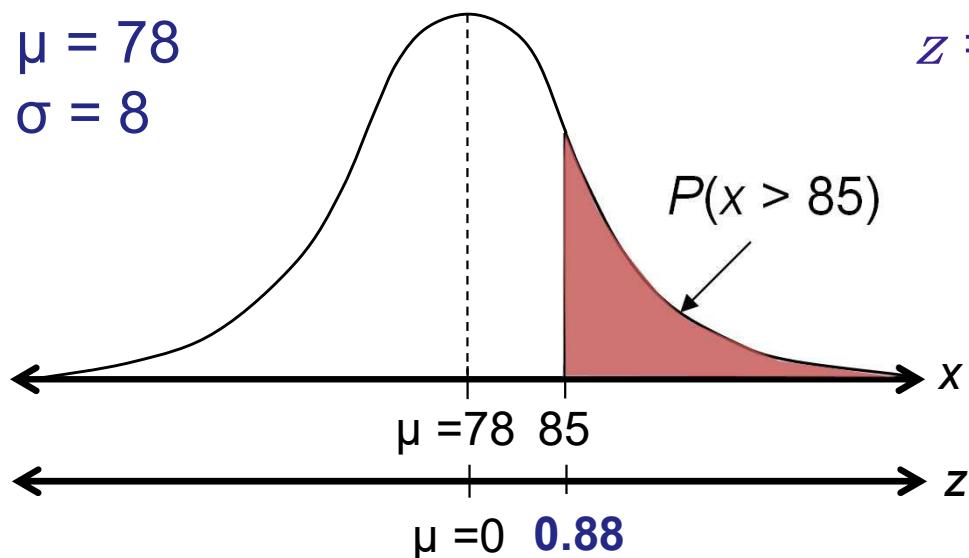
## Example:

The average on a statistics test was 78 with a standard deviation of 8. If the test scores are normally distributed, find the probability that a student receives a test score greater than 85.

## Example:

The average on a statistics test was 78 with a standard deviation of 8. If the test scores are normally distributed, find the probability that a student receives a test score greater than 85.

$$\begin{aligned}\mu &= 78 \\ \sigma &= 8\end{aligned}$$



$$\begin{aligned}z &= \frac{x - \mu}{\sigma} = \frac{85 - 78}{8} \\ &= 0.875 \approx 0.88\end{aligned}$$

The probability that a student receives a test score greater than 85 is 0.1894.

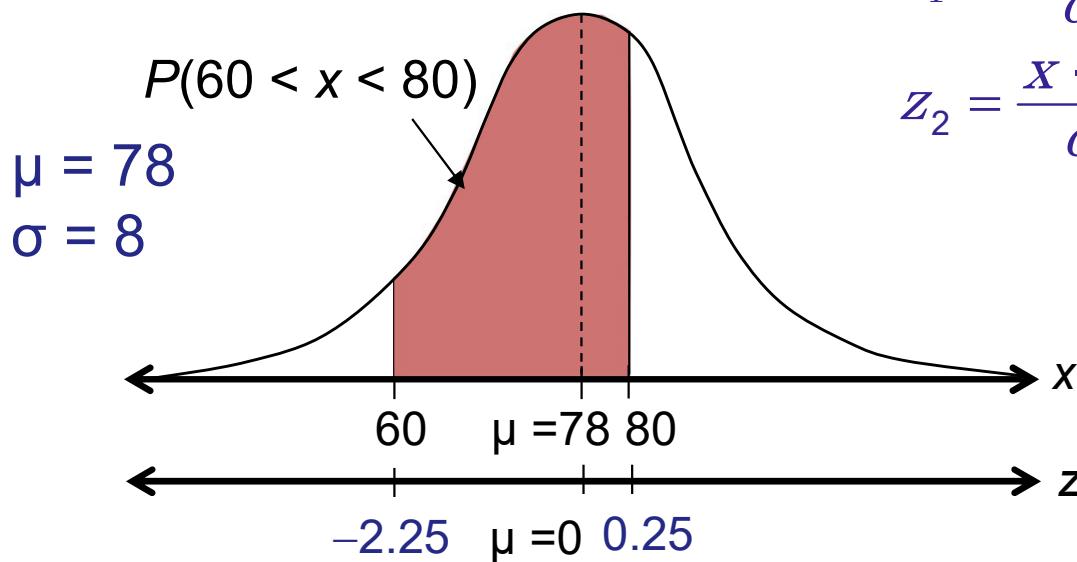
$$P(x > 85) = P(z > 0.88) = 1 - P(z < 0.88) = 1 - 0.8106 = 0.1894$$

## Example:

The average on a statistics test was 78 with a standard deviation of 8. If the test scores are normally distributed, find the probability that a student receives a test score between 60 and 80.

## Example:

The average on a statistics test was 78 with a standard deviation of 8. If the test scores are normally distributed, find the probability that a student receives a test score between 60 and 80.



$$z_1 = \frac{x - \mu}{\sigma} = \frac{60 - 78}{8} = -2.25$$

$$z_2 = \frac{x - \mu}{\sigma} = \frac{80 - 78}{8} = 0.25$$

The probability that a student receives a test score between 60 and 80 is 0.5865.

$$\begin{aligned} P(60 < x < 80) &= P(-2.25 < z < 0.25) = P(z < 0.25) - P(z < -2.25) \\ &= 0.5987 - 0.0122 = 0.5865 \end{aligned}$$

## EXAMPLE:

### *Interpreting the Area Under a Normal Curve*

---

The weights of pennies minted after 1982 are approximately normally distributed with mean 2.46 grams and standard deviation 0.02 grams.

- (a) Shade the region under the normal curve between 2.44 and 2.49 grams.
  - (b) Suppose the area under the normal curve for the shaded region is 0.7745. Provide two interpretations for this area.
-



# TRY

- The customer accounts at a certain departmental store have an average balance of Rs.480 and a SD of Rs.160. Assume that the accounts are normally distributed
  - A) what proportion of the accounts is over Rs 600
  - B) What proportion of accounts is between Rs 240 and Rs 360?



# TRY

- A large flashlight is powered by 5 batteries. Suppose that the life of a battery is normally distributed with mean = 150 hours and S.D = 15 hours.
- The flashlight will cease functioning if one or more of its batteries go dead. Assuming the lives of batteries are independent , what is the probability that flashlight will operate more than 130 hours



# Example

- Of a large group of men, 5% are under 60 inches in height and 40% are between 60 and 65 inches.

Assuming Normal distribution ,find the mean and SD.

$$x_1 = 60$$

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

$$P(x < 60) = P(Z < \frac{60-\mu}{\sigma}) = E\left(\frac{60-\mu}{\sigma}\right) = 0.05 \Rightarrow \frac{60-\mu}{\sigma} = -1.645$$

$$60 + 1.645\sigma = \mu \quad (1)$$

$$\begin{aligned} P(60 < x < 65) &= E\left(\frac{65-\mu}{\sigma}\right) - E\left(\frac{60-\mu}{\sigma}\right) = 0.4 \\ &= E\left(\frac{65-\mu}{\sigma}\right) = 0.45 \end{aligned}$$

$$\frac{65-\mu}{\sigma} = -0.126 \Rightarrow 65 + 0.126\sigma = \mu \quad (2)$$

EQUATING (1) & (2)  $\Rightarrow$

$$65 + 0.126\sigma = 60 + 1.645\sigma$$

$$5 = 1.519\sigma$$

$$\sigma = 5 / 1.519 = 3.291$$

$$\mu = 6$$



# Example

- In a Normal distribution, 7% of the items are under 35 and 89% of the items are under 63.
- Then find Mean and S.D

$$X_1 = 35 \quad X_2 = 63$$

$$P(X < 35) = 7/100 = 0.07 \Rightarrow P\left(Z < \frac{35-\mu}{\sigma}\right) = E\left(\frac{35-\mu}{\sigma}\right) = 0.07$$

$$P(X < 63) = 89/100 = 0.89 \Rightarrow P\left(Z < \frac{63-\mu}{\sigma}\right) = E\left(\frac{63-\mu}{\sigma}\right) = 0.89$$

10.89

0.11

$$\frac{35-\mu}{\sigma} = -1.476$$

$$\mu = 35 + 1.476\sigma \quad \textcircled{1}$$

$$\frac{63-\mu}{\sigma} = 1.227$$

$$\textcircled{2} \quad 63 - 1.227\sigma = \mu$$

$$63 - 35 = (1.227 + 1.476)\sigma$$

$$\sigma = 10.358$$

$$\mu = 35 + 1.476\sigma = 50.288$$

# Example

---

- 1000 light bulbs with a mean life of 120 days are installed in a new factory and their length of life is normally distributed with standard deviation of 20 days
- a) How many bulbs will expire in less than 90 days ?
-

## TITLE STYLE

---

- Suppose that 25% of all students at a large university receive financial aid. Let  $X$  be the number of students in a random sample of size 50 who receive financial aid , so that  $p = 0.25$ .
  - Then find
    - 1) the probability that at most 10 students receive aid
    - 2) Probability that between 5 and 15
-

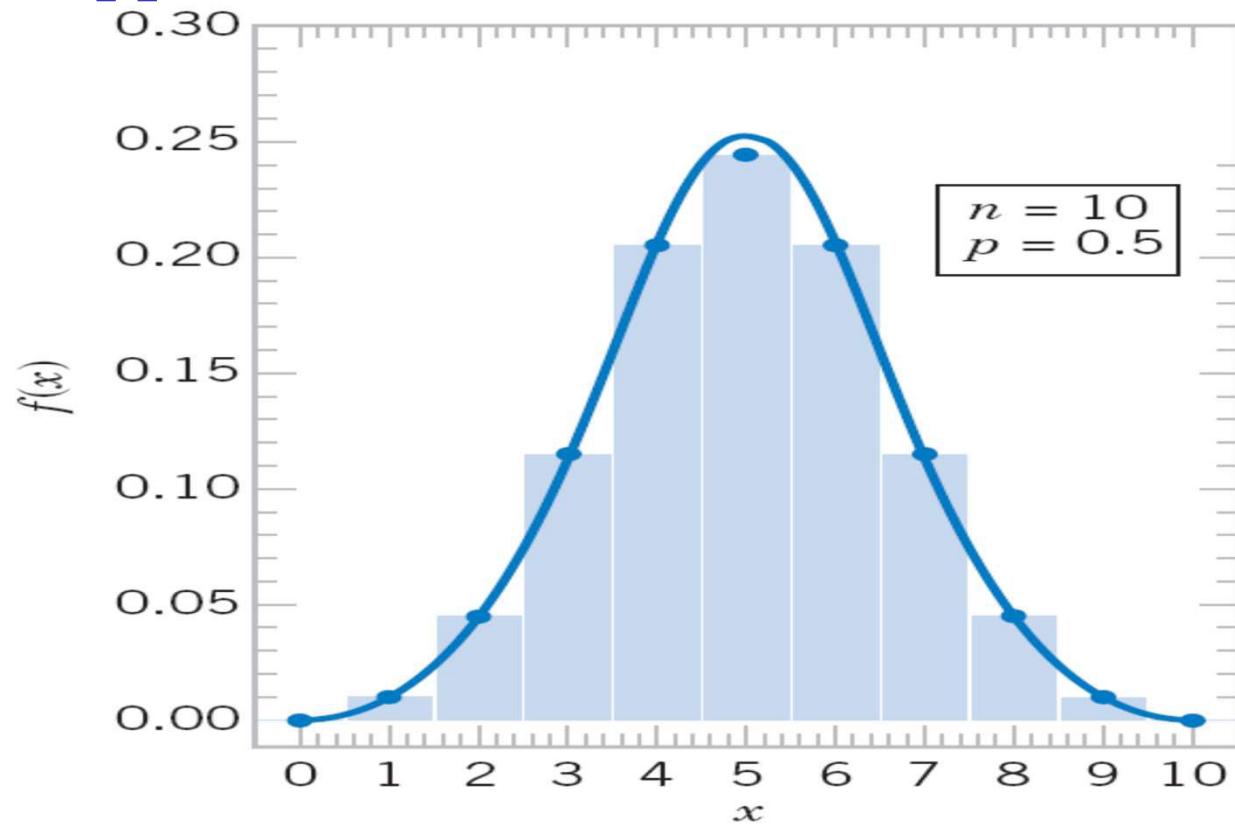
# Normal Approximation to the Binomial

If  $X$  is a binomial random variable,

$$Z = \frac{X - np}{\sqrt{np(1 - p)}} \quad (3-21)$$

is approximately a standard normal random variable. Consequently, probabilities computed from  $Z$  can be used to approximate probabilities for  $X$ .

## Normal Approximation to the Binomial



**Figure 3-36** Normal approximation to the binomial distribution.

## Normal Approximation to the Binomial

### EXAMPLE 3-32

Again consider the transmission of bits in the previous example. To judge how well the normal approximation works, assume that only  $n = 50$  bits are to be transmitted and that the probability of an error is  $p = 0.1$ . The exact probability that 2 or fewer errors occur is

$$P(X \leq 2) = \binom{50}{0} 0.9^{50} + \binom{50}{1} 0.1(0.9^{49}) + \binom{50}{2} 0.1^2(0.9^{48}) = 0.11$$

Based on the normal approximation,

$$P(X \leq 2) = P\left(\frac{X - 5}{\sqrt{50(0.1)(0.9)}} < \frac{2.5 - 5}{\sqrt{50(0.1)(0.9)}}\right) = P(Z < -1.18) = 0.12$$



## Normal Approximation to the Poisson

If  $X$  is a Poisson random variable with  $E(X) = \lambda$  and  $V(X) = \lambda$ ,

$$Z = \frac{X - \lambda}{\sqrt{\lambda}} \quad (3-22)$$

is approximately a standard normal random variable.



# EXAMPLE

- How would you use the normal distribution to find approximate frequency of exactly 5 successes in 100 trials ,the probability of success in each trial being  $p = 0.1$



# **Session 5**

## **(19<sup>th</sup> February, 2022)**

## **(Normal Dist & Sampling)**

- 
- ❖ Random Variables
  - ❖ Mathematical Expectation
  - ❖ Mean and Variance
  - ❖ Prob.Distributions
-



# Normal Distribution







**STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.**

lead

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361





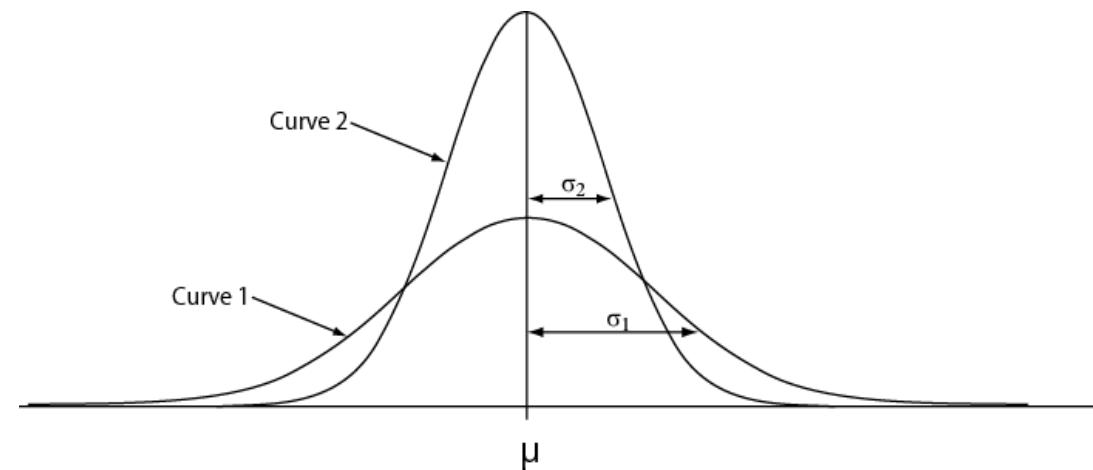
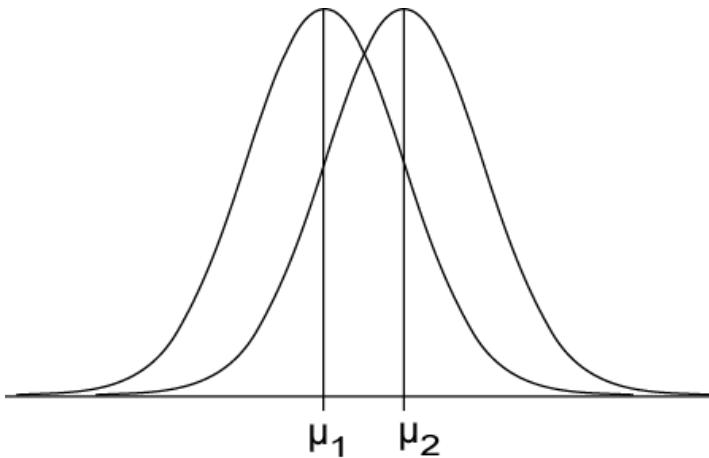


# Parameters $\mu$ and $\sigma$

- Normal *pdfs* have two **parameters**

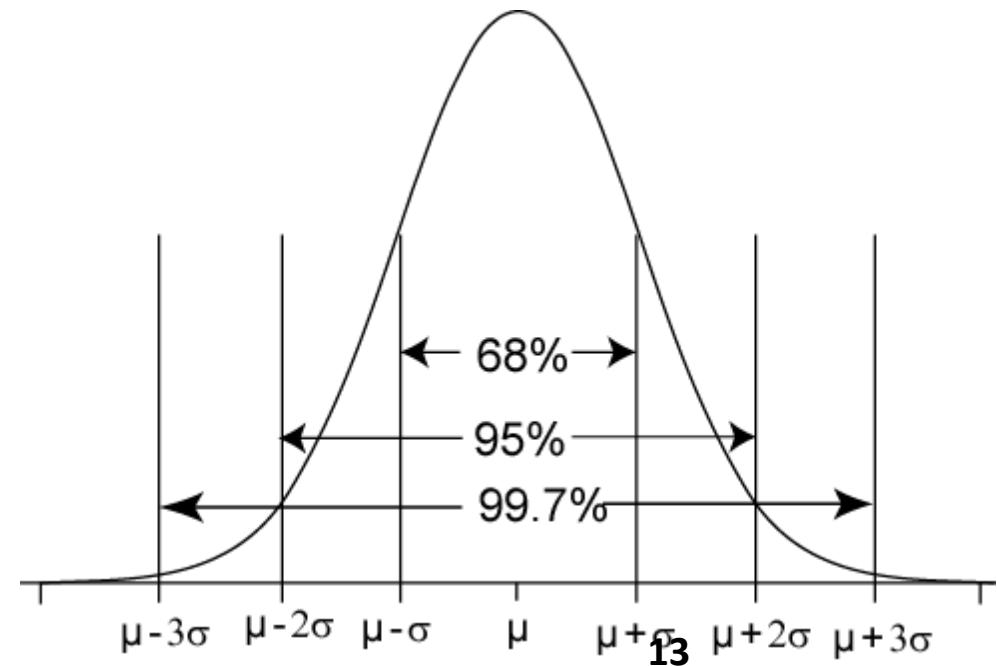
$\mu$  - expected value (mean “mu”)

$\sigma$  - standard deviation (sigma)  
 $\mu$  controls location                     $\sigma$  controls spread



# 68-95-99.7 Rule for Normal Distributions

- 68% of the AUC within  $\pm 1\sigma$  of  $\mu$
- 95% of the AUC within  $\pm 2\sigma$  of  $\mu$
- 99.7% of the AUC within  $\pm 3\sigma$  of  $\mu$



**EXAMPLE:** *Finding the Area Under the Standard Normal Curve*

- Find the area under the standard normal curve to the right of  $Z = 1.25$ .

**EXAMPLE:** *Finding the Area Under the Standard Normal Curve*

Find the area under the standard normal curve between  $Z = -1.02$  and  $Z = 2.94$ .

$$\begin{aligned} P(-1.02 < x < 2.94) &= P(x < 2.94) - P(x < -1.02) \\ &= 0.9984 - 0.1539 \\ &= 0.8445 \end{aligned}$$

# Properties of Normal Distributions

---

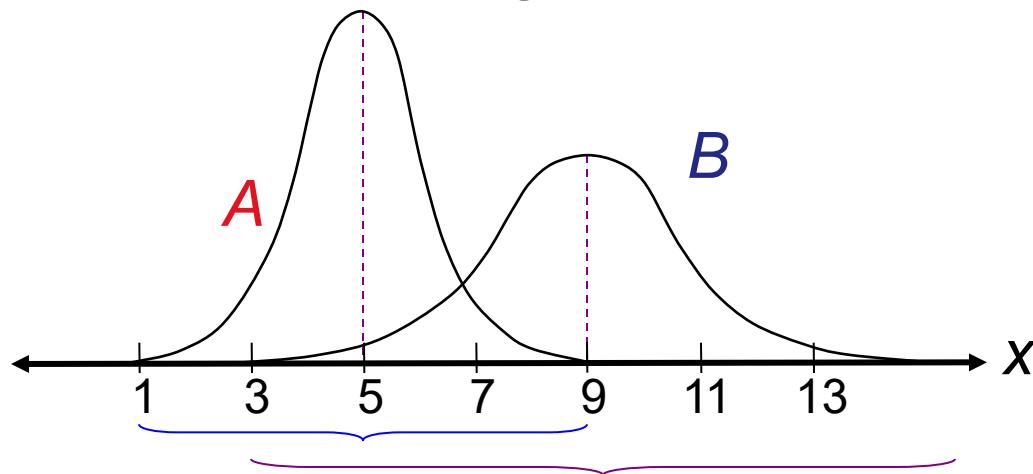
## Properties of a Normal Distribution

1. The mean, median, and mode are equal.
2. The normal curve is bell-shaped and symmetric about the mean.
3. The total area under the curve is equal to one.
4. The normal curve approaches, but never touches the  $x$ -axis as it extends farther and farther away from the mean.

# Means and Standard Deviations

Example:

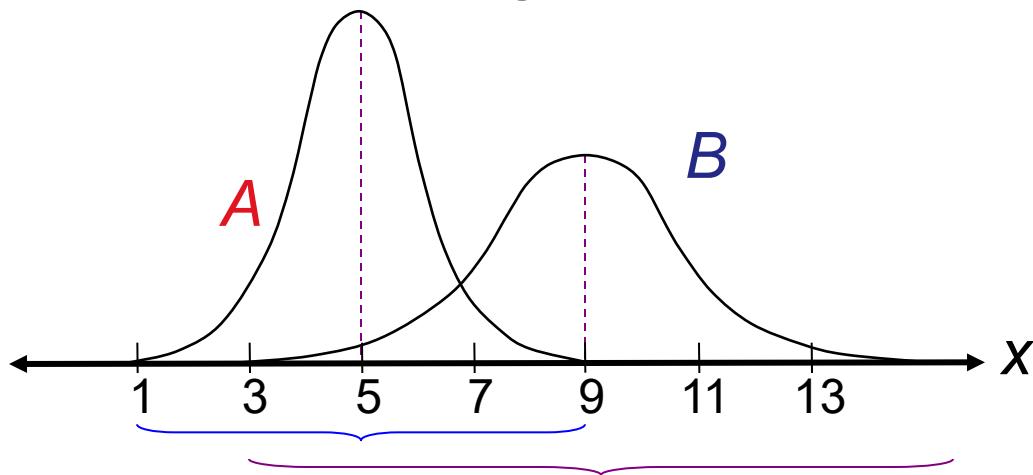
1. Which curve has the greater mean?
2. Which curve has the greater standard deviation?



# Means and Standard Deviations

## Example:

1. Which curve has the greater mean?
2. Which curve has the greater standard deviation?



The line of symmetry of curve  $A$  occurs at  $x = 5$ . The line of symmetry of curve  $B$  occurs at  $x = 9$ . Curve  $B$  has the greater mean.

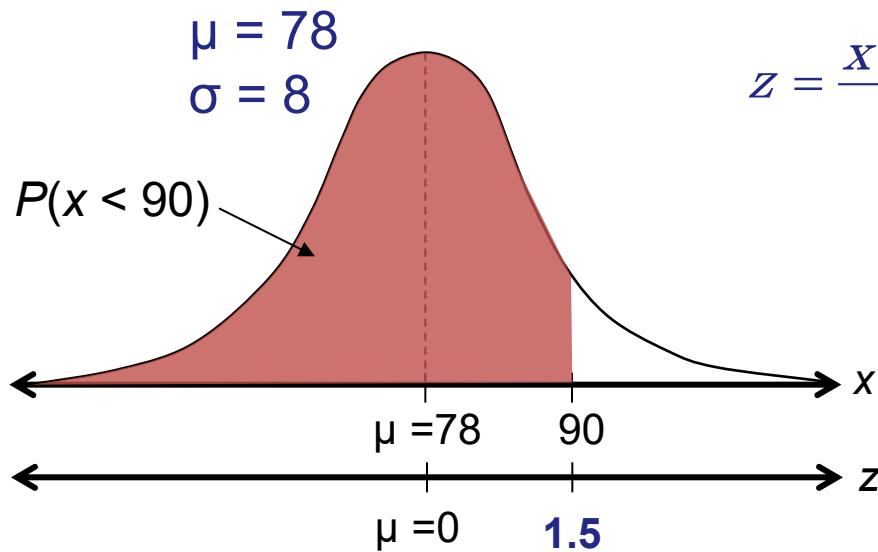
Curve  $B$  is more spread out than curve  $A$ , so curve  $B$  has the greater standard deviation.

## Example:

The average on a statistics test was 78 with a standard deviation of 8. If the test scores are normally distributed, find the probability that a student receives a test score less than 90.

## Example:

The average on a statistics test was 78 with a standard deviation of 8. If the test scores are normally distributed, find the probability that a student receives a test score less than 90.



$$z = \frac{x - \mu}{\sigma} = \frac{90 - 78}{8} = 1.5$$

The probability that a student receives a test score less than 90 is 0.9332.

$$P(x < 90) = P(z < 1.5) = 0.9332$$

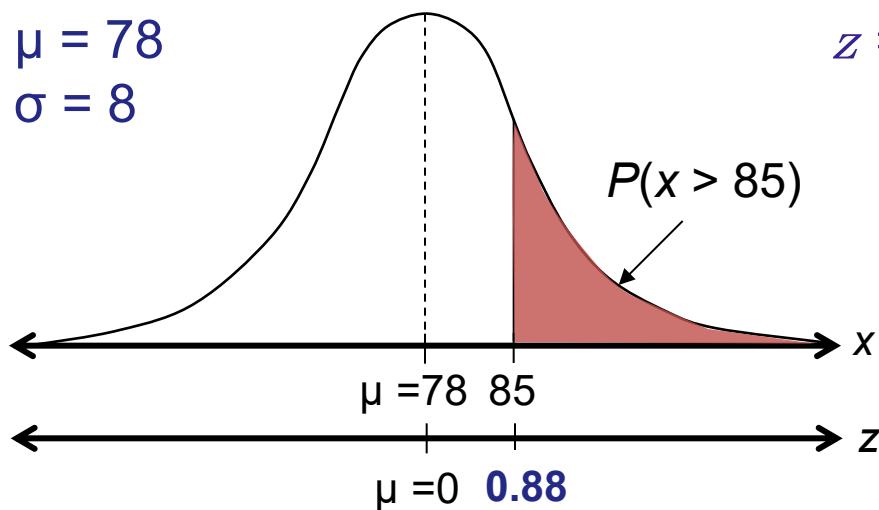
## Example:

The average on a statistics test was 78 with a standard deviation of 8. If the test scores are normally distributed, find the probability that a student receives a test score greater than 85.

## Example:

The average on a statistics test was 78 with a standard deviation of 8. If the test scores are normally distributed, find the probability that a student receives a test score greater than 85.

$$\begin{aligned}\mu &= 78 \\ \sigma &= 8\end{aligned}$$



$$\begin{aligned}z &= \frac{x - \mu}{\sigma} = \frac{85 - 78}{8} \\ &= 0.875 \approx 0.88\end{aligned}$$

The probability that a student receives a test score greater than 85 is 0.1894.

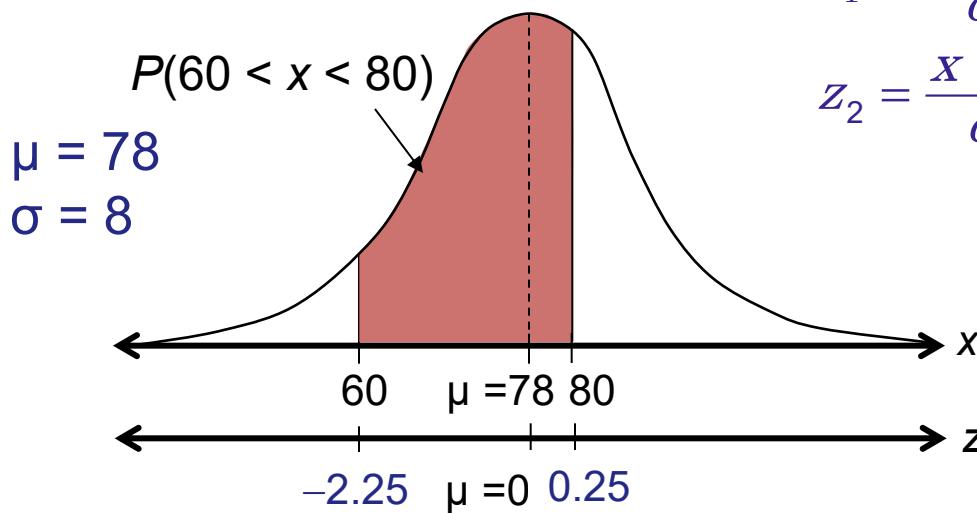
$$P(x > 85) = P(z > 0.88) = 1 - P(z < 0.88) = 1 - 0.8106 = 0.1894$$

## Example:

The average on a statistics test was 78 with a standard deviation of 8. If the test scores are normally distributed, find the probability that a student receives a test score between 60 and 80.

## Example:

The average on a statistics test was 78 with a standard deviation of 8. If the test scores are normally distributed, find the probability that a student receives a test score between 60 and 80.



$$z_1 = \frac{x - \mu}{\sigma} = \frac{60 - 78}{8} = -2.25$$

$$z_2 = \frac{x - \mu}{\sigma} = \frac{80 - 78}{8} = 0.25$$

The probability that a student receives a test score between 60 and 80 is 0.5865.

$$\begin{aligned} P(60 < x < 80) &= P(-2.25 < z < 0.25) = P(z < 0.25) - P(z < -2.25) \\ &= 0.5987 - 0.0122 = 0.5865 \end{aligned}$$

## EXAMPLE:

### *Interpreting the Area Under a Normal Curve*

---

The weights of pennies minted after 1982 are approximately normally distributed with mean 2.46 grams and standard deviation 0.02 grams.

- (a) Shade the region under the normal curve between 2.44 and 2.49 grams.
  - (b) Suppose the area under the normal curve for the shaded region is 0.7745. Provide two interpretations for this area.
-

# TRY

- The customer accounts at a certain departmental store have an average balance of Rs.480 and a SD of Rs.160. Assume that the accounts are normally distributed
  - A) what proportion of the accounts is over Rs 600
  - B) What proportion of accounts is between Rs 240 and Rs 360?

# TRY

- A large flashlight is powered by 5 batteries. Suppose that the life of a battery is normally distributed with mean = 150 hours and S.D = 15 hours.
- The flashlight will cease functioning if one or more of its batteries go dead. Assuming the lives of batteries are independent , what is the probability that flashlight will operate more than 130 hours

# Example

- Of a large group of men, 5% are under 60 inches in height and 40% are between 60 and 65 inches. Assuming Normal distribution ,find the mean and SD.

# Example

- In a Normal distribution, 7% of the items are under 35 and 89% of the items are under 63.
- Then find Mean and S.D

# Example

---

- 1000 light bulbs with a mean life of 120 days are installed in a new factory and their length of life is normally distributed with standard deviation of 20 days
- a) How many bulbs will expire in less than 90 days ?

$$\begin{aligned} \mu &= 120 & \sigma &= 20 & x &= 90 \\ P(x < 90) &= P\left(z < \frac{90-120}{20}\right) \\ &= P(-3/2) = P(-1.5) \\ &\approx 0.066 \end{aligned}$$

$$\begin{aligned} (\text{a}) \text{ Bulbs thatll expire in } < 90 &= 0.066 \times 1000 \\ &= 66 \text{ bulbs} \end{aligned}$$


---

## TITLE STYLE

---

- Suppose that 25% of all students at a large university receive financial aid. Let  $X$  be the number of students in a random sample of size 50 who receive financial aid , so that  $p = 0.25$ .
  - Then find
  - 1) the probability that at most 10 students receive aid
  - 2) Probability that between 5 and 15
-

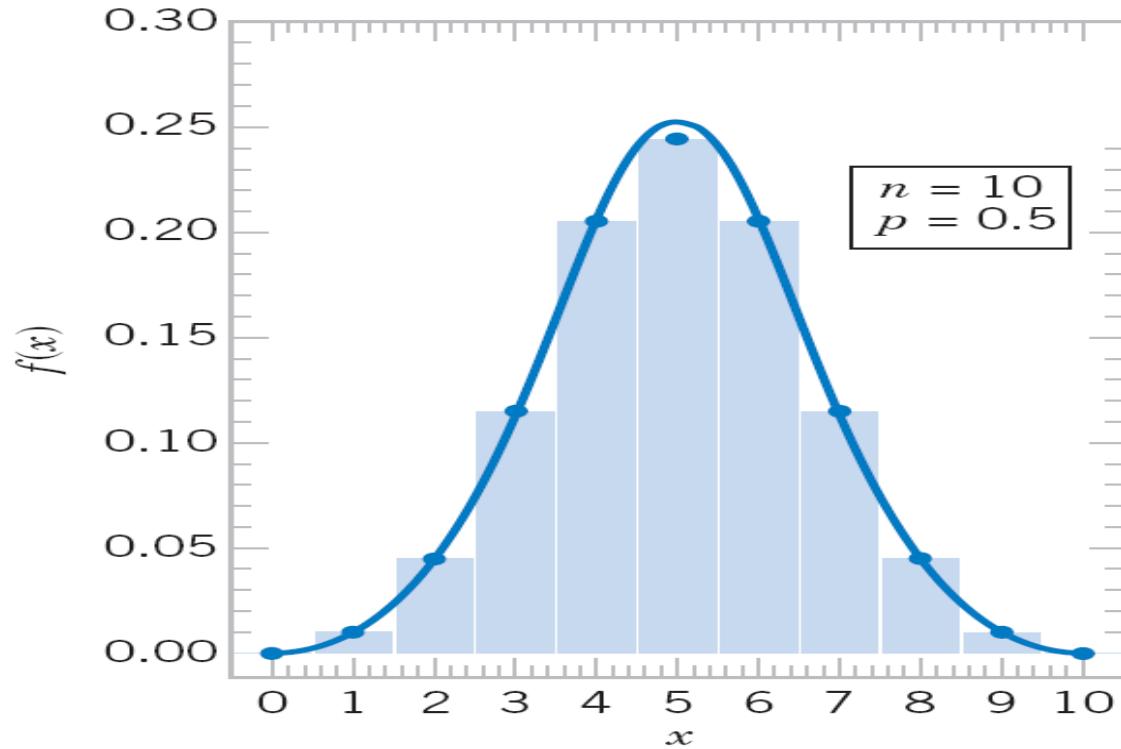
## Normal Approximation to the Binomial

If  $X$  is a binomial random variable,

$$Z = \frac{X - np}{\sqrt{np(1 - p)}} \quad (3-21)$$

is approximately a standard normal random variable. Consequently, probabilities computed from  $Z$  can be used to approximate probabilities for  $X$ .

## Normal Approximation to the Binomial



**Figure 3-36** Normal approximation to the binomial distribution.

## Normal Approximation to the Binomial

### EXAMPLE 3-32

Again consider the transmission of bits in the previous example. To judge how well the normal approximation works, assume that only  $n = 50$  bits are to be transmitted and that the probability of an error is  $p = 0.1$ . The exact probability that 2 or fewer errors occur is

$$P(X \leq 2) = \binom{50}{0} 0.9^{50} + \binom{50}{1} 0.1(0.9^{49}) + \binom{50}{2} 0.1^2(0.9^{48}) = 0.11$$

Based on the normal approximation,

$$P(X \leq 2) = P\left(\frac{X - 5}{\sqrt{50(0.1)(0.9)}} < \frac{2.5 - 5}{\sqrt{50(0.1)(0.9)}}\right) = P(Z < -1.18) = 0.12$$

## Normal Approximation to the Poisson

If  $X$  is a Poisson random variable with  $E(X) = \lambda$  and  $V(X) = \lambda$ ,

$$Z = \frac{X - \lambda}{\sqrt{\lambda}} \quad (3-22)$$

is approximately a standard normal random variable.

# EXAMPLE

- How would you use the normal distribution to find approximate frequency of exactly 5 successes in 100 trials ,the probability of success in each trial being  $p = 0.1$

# Sampling Distribution & Estimation

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20)
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.20,random_state = 42)
```

# Sampling

- Sampling is widely used in business as a means of gathering useful information about a population.
- Data are gathered from samples and conclusions are drawn about the population as a part of the inferential statistics process
- A sample provides a reasonable means for gathering useful decision-making information that might be otherwise unattainable and unaffordable.

# Reasons for Sampling

Taking a sample instead of conducting a census offers several advantages

1. The sample can save money.
2. The sample can save time.
3. For given resources, the sample can broaden the scope of the study.
4. If accessing the population is impossible, the sample is the only option.

# Random Versus Non random Sampling



- In **random** sampling every unit of the population has the same probability of being selected into the sample.
  
- In **nonrandom** sampling not every unit of the population has the same probability of being selected into the sample.

# Stratified Random Sampling

---

- In this the population is divided into non overlapping **subpopulations** called **strata**.
  - The researcher then extracts a random sample from each of the subpopulations.
  - The main reason for using stratified random sampling is that it has the potential for reducing sampling error.
  - With stratified random sampling, the potential to match the sample closely to the population is greater than it is with simple random sampling because portions of the total sample are taken from different population subgroups.
-

# Sampling Error

---

- **Sampling error** occurs when the sample is not representative of the population.
  
  - When random sampling techniques are used to select elements for the sample, sampling error occurs by chance.
-

## Population of Wages of employees of an organization

1861	2495	1000	2497	1865	791	2090	2637	1327	1678
1680	2858	795	2495	2496	2501	1160	1480	1860	2490
2090	2840	2490	2640	659	827	2646	2638	2643	868
1327	1866	1861	2486	2865	3011	2494	1489	1865	2855
2840	2499	2093	2660	1165	2600	2085	2640	2998	1861
2956	2495	2865	1865	3000	3019	1670	2858	2642	1680
3038	3000	1313	596	656	3240	590	2501	2485	3015
2092	1679	3024	2497	2825	2630	2070	2900	1861	2636
2495	2637	2497	1159	2640	3050	870	2896	2500	2638
926	2860	1481	875	2482	1860	2086	934	3200	2490

## Select different samples of varied sizes

### Sample 1

3000    2486    820    1678    2070    2638    2490    1865    1000    2090    596    3200

### Sample 2

2840    2858    3000    2490    2998    3050    2070    2896    3200    2490    3280

### Sample 3

2858    3240    2497    2865    656    2093    934    1861    868    795

### Sample 4

2086    1000    2497    596    656    875    2085    934    1313

### Sample 5

820    1313    3000    2640    596    2640    2600    2495    934    2500

## Select different samples of varied sizes

### Sample 6

2840    2499    1327    1861    2495    3024    3038    2497

### Sample 7

2858    2490    868    1670    1480    2643    1480    1680    2085    2490

### Sample 8

2495    2858    1861    2092    2499    3000    2660    1000    1679    926    2660

### Sample 9

795    791    3200    2085    2638    2497    2486    1159    2640

### Sample 10

3019    3240    3200    3050    3000    3015    2900    2896    2998

**Compute sample mean of these samples**

<b>Sample No.</b>	<b>Sample size</b>	<b>Mean</b>	<b>SD</b>
1	12	1994.42	843.23
2	11	2830.18	349.94
3	10	1866.70	988.57
4	9	1338.00	704.36
5	10	1953.80	920.44
6	8	2447.63	590.64
7	10	1974.40	638.05
8	11	2157.27	715.10
9	9	2032.33	891.53
10	9	3035.33	117.40
<b>Overall</b>	<b>100</b>	<b>2162.24</b>	<b>732.26</b>

# Sampling Variability

---

- The term "sampling variability" refers to the fact that the statistical information from a sample (called a *statistic*) will vary as the random sampling is repeated.
  - **Sampling variability will decrease as the sample size increases.**
  - the samples must be randomly chosen, must be of the same size (not smaller than 30), and the more samples that are used, the more reliable the information gathered will be.
-

Do you consider these sample means and sample SDs as variable ?

If yes, should we not describe the distribution of these variables ?

The distribution of the sample estimates is called sampling distribution

For example the distribution of sample means is called Sampling distribution of mean

# Definition

---

- The probability distribution of a statistic (sample estimate) is called sampling distribution.
  
  - The sampling distribution of a statistic depends on the distribution of the population, the size of the sample, and the method of sample selection.
-

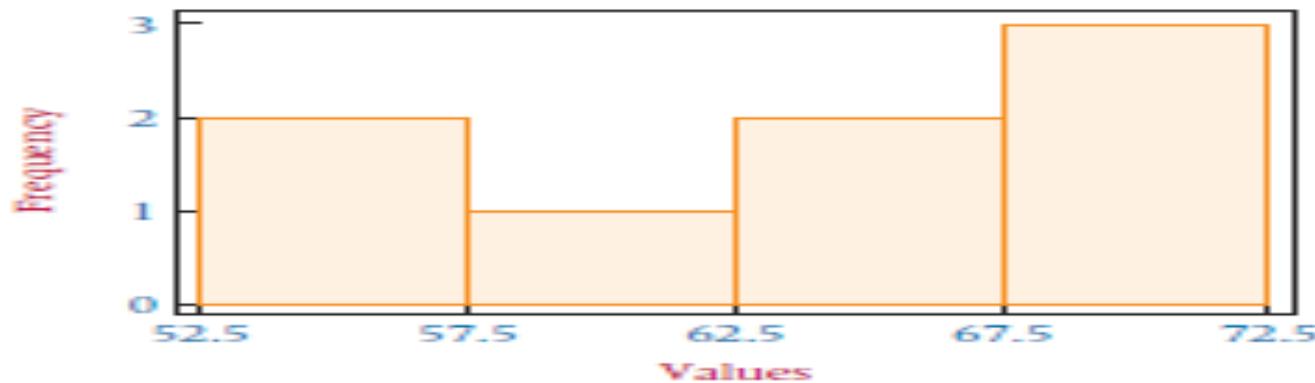
# Sampling Distribution Of $\bar{x}$

---

- The sample mean is one of the more common statistics used in the inferential process.
  - The **distribution** of the values of the sample mean ( $\bar{x}$ ) in repeated **samples** is called the **sampling distribution of  $\bar{x}$**
  - One way to examine the distribution possibilities is to take a population with a particular distribution, randomly select samples of a given size, compute the sample means, and attempt to determine how the means are distributed.
-

# Example

- Suppose a small finite population consists of only  $N = 8$  numbers:
- 54 55 59 63 64 68 69 70
- Using an Excel-produced histogram, we can see the shape of the distribution of this population of data.



- Suppose we take all possible samples of size  $n = 2$  from this population with replacement.

# Example

The result is the following pairs of data.

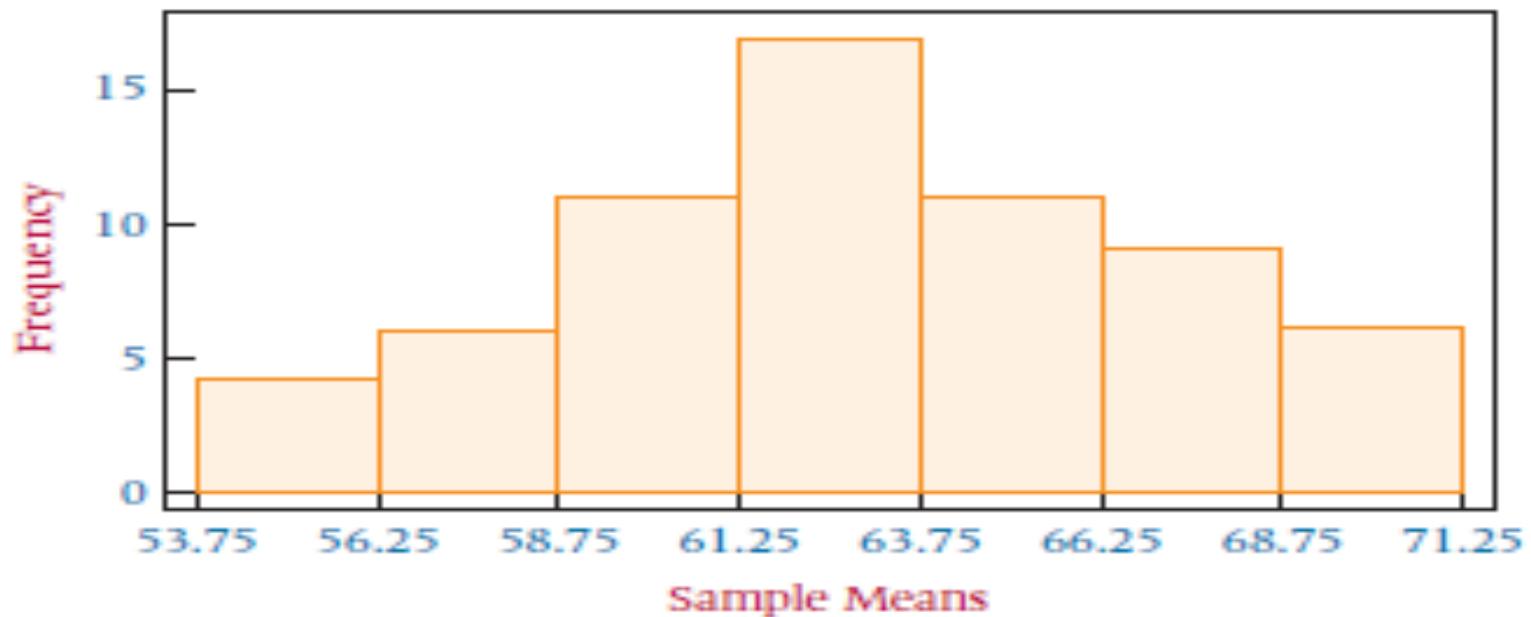
(54,54)	(55,54)	(59,54)	(63,54)
(54,55)	(55,55)	(59,55)	(63,55)
(54,59)	(55,59)	(59,59)	(63,59)
(54,63)	(55,63)	(59,63)	(63,63)
(54,64)	(55,64)	(59,64)	(63,64)
(54,68)	(55,68)	(59,68)	(63,68)
(54,69)	(55,69)	(59,69)	(63,69)
(54,70)	(55,70)	(59,70)	(63,70)
(64,54)	(68,54)	(69,54)	(70,54)
(64,55)	(68,55)	(69,55)	(70,55)
(64,59)	(68,59)	(69,59)	(70,59)
(64,63)	(68,63)	(69,63)	(70,63)
(64,64)	(68,64)	(69,64)	(70,64)
(64,68)	(68,68)	(69,68)	(70,68)
(64,69)	(68,69)	(69,69)	(70,69)
(64,70)	(68,70)	(69,70)	(70,70)

The means of each of these samples follow.

54	54.5	56.5	58.5	59	61	61.5	62
54.5	55	57	59	59.5	61.5	62	62.5
56.5	57	59	61	61.5	63.5	64	64.5
58.5	59	61	63	63.5	65.5	66	66.5
59	59.5	61.5	63.5	64	66	66.5	67
60	61.5	63.5	65.5	66	68	68.5	69
61.5	62	64	66	66.5	68.5	69	69.5
62	62.5	64.5	66.5	67	69	69.5	70

# Example

- Again using an Excel-produced histogram, we can see the shape of the distribution of these sample means.



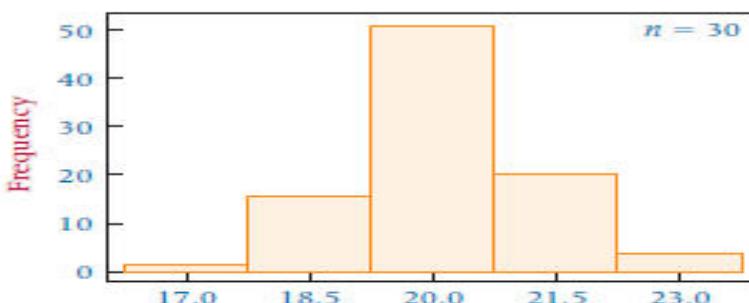
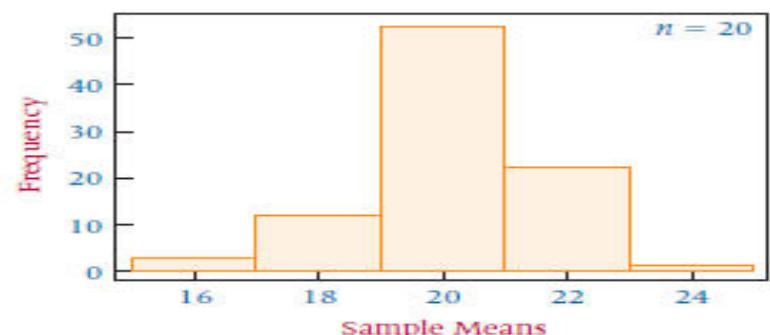
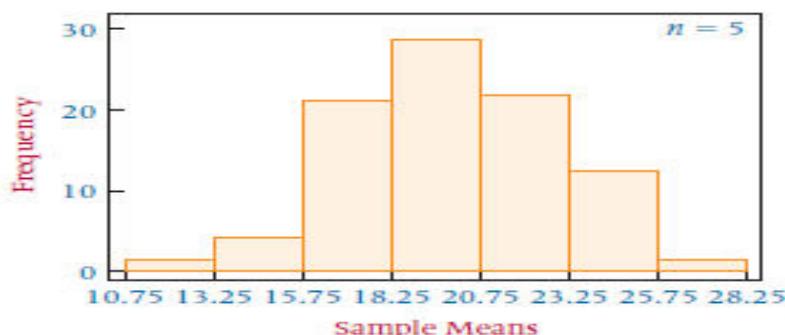
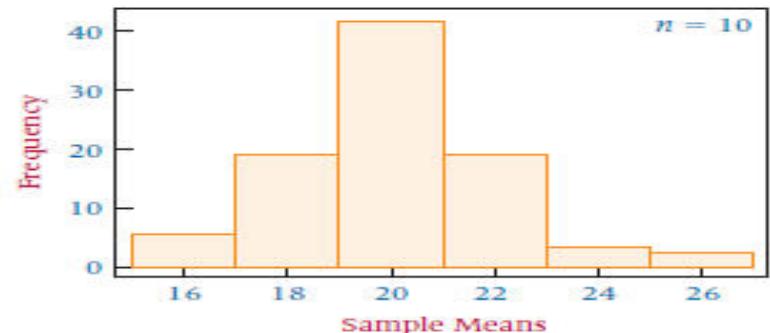
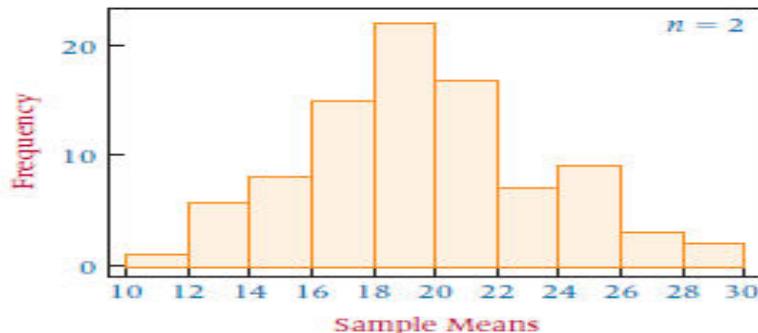
# Conclusions

---

- Notice that the shape of the histogram for sample means is quite unlike the shape of the histogram for the population.
  - The sample means appear to “pile up” toward the middle of the distribution and “tail off” toward the extremes.
  - As sample sizes become much larger, the sample mean distributions begin to approach a **normal distribution** and the variation among the means decreases.
-

## Sample Means from 90 Samples Ranging in Size from $n = 2$ to $n =$

**30 from a Uniformly Distributed Population with  $a = 10$  and  $b = 30$**



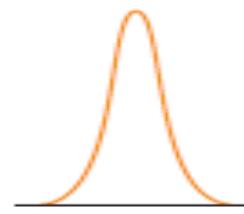
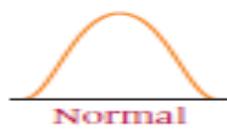
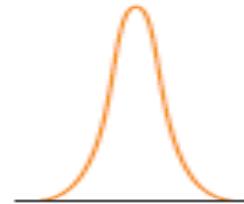
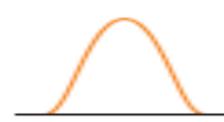
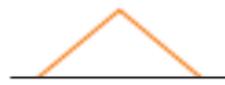
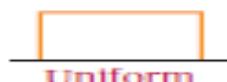
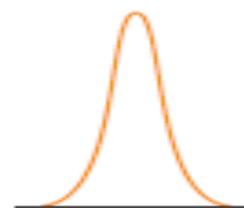
# Shapes of the Distributions of Sample Means

Population Distribution

$n = 2$

$n = 5$

$n = 30$



# Central Limit Theorem

- If samples of size  $n$  are drawn randomly from a population that has a mean of  $\mu$  and a standard deviation of  $\sigma$ , the sample means,  $\bar{x}$ , are approximately normally distributed for sufficiently large sample sizes ( $n \geq 30$ ) regardless of the shape of the population distribution.
- If the population is normally distributed, the sample means are normally distributed for any size sample.
- From mathematical expectation

$$\mu_{\bar{x}} = \mu \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

# Z score for sample means

- The central limit theorem states that sample means are normally distributed regardless of the shape of the population for large samples and for any sample size with normally distributed populations.
- Thus, **sample means** can be **analyzed** by using **z scores**
- The formula to 
$$z = \frac{x - \mu}{\sigma}$$
 z scores for individual values from a normal distribution
- If sample means are normally distributed, the z score formula applied to sample means would be
$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$
- The standard deviation of the statistic of interest is  $\sigma_{\bar{x}}$ , sometimes referred to as the **standard error of the mean**.

# Example

---

Suppose the mean expenditure per customer at a tire store is \$85.00, with a standard deviation of \$9.00.

If a random sample of 40 customers is taken, what is the probability that the sample average expenditure per customer for this sample will be \$87.00 or more?

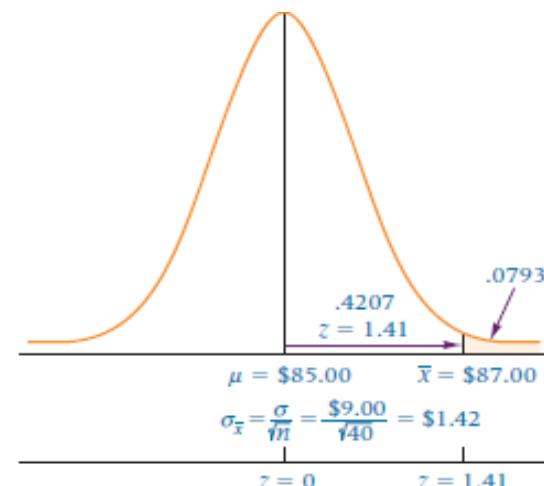
# Solution

Because the sample size is greater than 30, the central limit theorem

Can be used, and the sample means are normally distributed.

$$\mu = \$85 \quad \sigma = \$9$$

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\$87.00 - \$85.00}{\frac{\$9.00}{\sqrt{40}}} = \frac{\$2.00}{\$1.42} = 1.41$$



# Exercise

---

- Suppose that during any hour in a large department store, the average number of shoppers is 448, with a standard deviation of 21 shoppers.
  
  - What is the probability that a random sample of 49 different shopping hours will yield a sample mean between 441 and 446 shoppers?
-

# Solution

For this problem,  $\mu = 448$ ,  $\sigma = 21$ , and  $n = 49$ . The problem is to determine

$$P(441 \leq \bar{x} \leq 446).$$

The following

$$z = \frac{441 - 448}{\frac{21}{\sqrt{49}}} = \frac{-7}{3} = -2.33$$

$$z = \frac{446 - 448}{\frac{21}{\sqrt{49}}} = \frac{-2}{3} = -0.67$$

# Sampling from a Finite Population

---

- The earlier example was based on the assumption that the population was infinitely or extremely large.
- In cases of a finite population, *a statistical adjustment can be made to the z formula for sample means*. The adjustment is called the **finite correction factor**

$$\sqrt{\frac{N-n}{N-1}}.$$

- Following is the z formula for sample means when samples are drawn from finite populations.

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}}$$

# Rules for finite population

---

- As the size of the finite population becomes larger in relation to sample size, the finite correction factor approaches 1.
  
  - In theory, whenever researchers are working with a finite population, they can use the finite correction factor.
  
  - A rough rule of thumb for many researchers is that, if the sample size is **less** than **5%** of the finite population size or  **$n/N < 0.05$** , the finite correction factor does **not** significantly modify the solution.
-

# Exercise

- A production company's 350 hourly employees average 37.6 years of age, with a standard deviation of 8.3 years.
  
- If a random sample of 45 hourly employees is taken, what is the probability that the sample will have an average age of less than 40 years?

# Solution

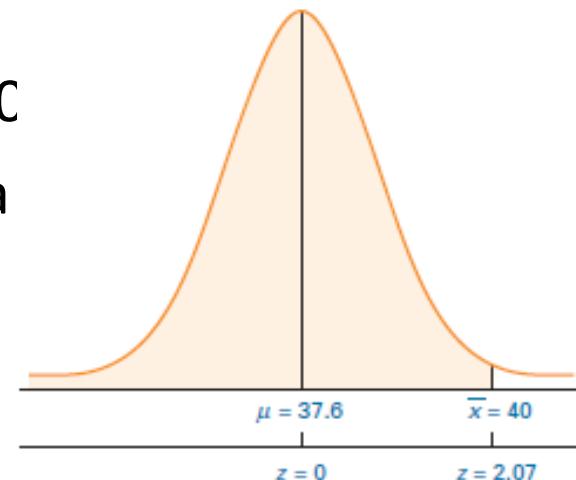
---

- The population mean is 37.6, with a population standard deviation of 8.3.
- The sample size is 45, but it is being drawn from a finite population of 350; that is,  $n = 45$  and  $N = 350$ .
- The sample mean under consideration is 40
- Using the z formula with the finite correction factor gives

$$z = \frac{40 - 37.6}{\frac{8.3}{\sqrt{45}} \sqrt{\frac{350 - 45}{350 - 1}}} = \frac{2.4}{1.157} = 2.07$$

# Solution

- This z value yields a probability of .480
- Therefore, the probability of getting a
- sample average age of less than
- 40 years is **.4808 + .5000 = .9808.**



# Sampling Distribution Of Sample Proportion

---

- If research results in **countable** items such as how many people in a sample have a flexible work schedule, the sample proportion is often the statistic of choice.

## SAMPLE PROPORTION

$$\hat{p} = \frac{x}{n}$$

where

$x$  = number of items in a sample that have the characteristic  
 $n$  = number of items in the sample

# Example

---

- In a sample of 100 factory workers, 30 workers might belong to a union.
  
- The value of sample proportion for this characteristic, union membership, is

$$30/100 = 0.30$$

# How does a researcher use the sample proportion in analysis?

---



- The central limit theorem applies to sample proportions in that the normal distribution approximates the shape of the distribution of sample proportions
- If  $n*p > 5$  and  $n*q > 5$  ( $p$  is the population proportion and  $q = 1 - p$ ).
- The mean of sample proportions for all samples of size  $n$  randomly drawn from a population is  $p$  (the population proportion) and the standard deviation of sample proportions is  $\sqrt{\frac{p \cdot q}{n}}$
- sometimes referred to as the **standard error of the proportion**

# Z Formula For Sample Proportions

---

For  $n * p > 5$  and  $n * q > 5$

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}}$$

where

$\hat{p}$  = sample proportion

$n$  = sample size

$p$  = population proportion

$q = 1 - p$

# Example

---

Suppose 60% of the electrical contractors in a region use a particular brand of wire. What is the probability of taking a random sample of size 120 from these electrical contractors and finding that .50 or less use that brand of wire?

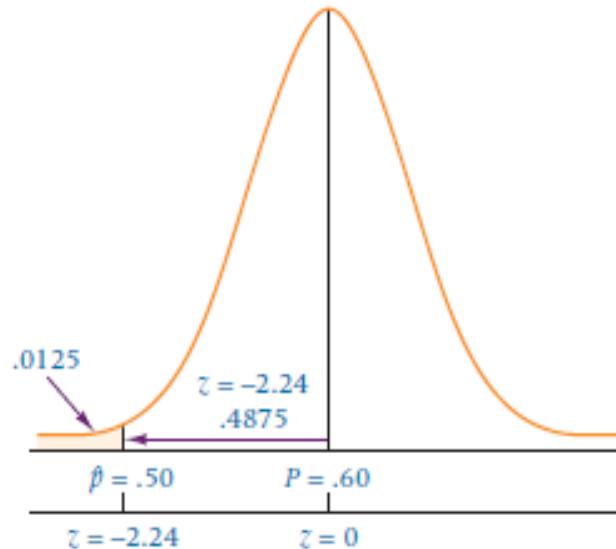
# Solution

$$p = .60 \quad \hat{p} = .50 \quad n = 120$$

The  $z$  formula yields

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{.50 - .60}{\sqrt{\frac{(.60)(.40)}{120}}} = \frac{-10}{.0447} = -2.24$$

ding to



$= -2.24$  is  $.4875$ .

For  $z < -2.24$  (the tail of the distribution), the answer is  $.5000 - .4875 = .0125$ .

# Exercise

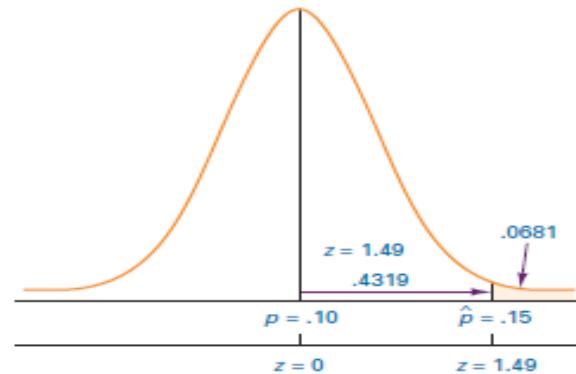
---

If 10% of a population of parts is defective, what is the probability of randomly selecting 80 parts and finding that 12 or more parts are defective?

# Solution

Here,  $p = .10$ ,  $\hat{p} = 12/80 = .15$ , and  $n = 80$ . Entering these values in the z formula yields

$$z = \frac{.15 - .10}{\sqrt{\frac{(.10)(.90)}{80}}} = \frac{.05}{.0335} = 1.49$$



$$P(\hat{p} \geq .15) = .5000 - .4319 = .0681.$$

- The probability of .4319 for a z value of 1.49, which is the area between the sample proportion, .15, and the population proportion, .10. The answer to the question is

# Forms Of Statistical Inference

- ❖ Three forms of statistical inference

- Point estimation
- Interval estimation
- Hypothesis testing

# Point Estimate

---

- A **point estimate** is a statistic taken from a sample that is used to estimate a population parameter.
  - A point estimate is only as good as the representativeness of its sample.
  - If other random samples are taken from the population, the point estimates derived from those samples are likely to vary.
-

# Interval Estimate

- Because of variation in sample statistics, estimating a population parameter with an interval estimate is often preferable to using a point estimate.
- An interval estimate (confidence interval) is a range of values within which the analyst can declare, with some confidence, the population parameter lies.

# Confidence Interval to Estimate $\mu$

100(1 -  $\alpha$ )% CONFIDENCE  
INTERVAL TO ESTIMATE  $\mu$ :

$\sigma$  KNOWN (8.1)

or

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

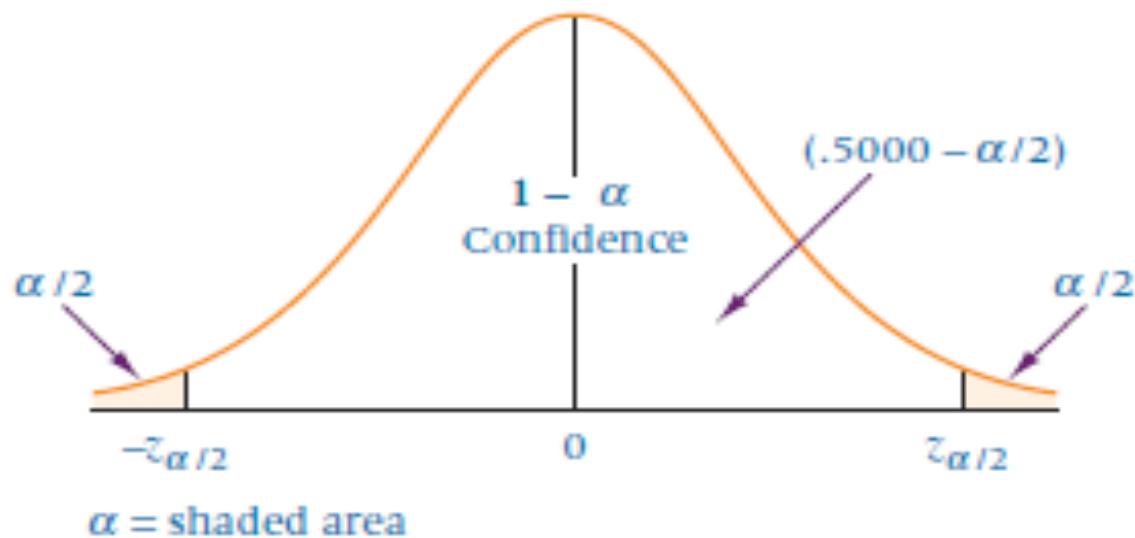
$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where

$\alpha$  = the area under the normal curve outside the confidence interval area

$\alpha/2$  = the area in one end (tail) of the distribution outside the confidence interval

# Confidence Intervals



# Example

---

- In the cellular telephone company problem of estimating the population mean number of minutes called per residential user per month, from the sample of 85 bills it was determined that the sample mean is 510 minutes.
  
  - Suppose past history and similar studies indicate that the population standard deviation is 46 minutes.
  
  - Determine a 95% confidence interval.
-

# Solution

---

The business researcher can now complete the cellular telephone problem. To determine a 95% confidence interval for  $\bar{x} = 510$ ,  $\sigma = 46$ ,  $n = 85$ , and  $z = 1.96$ , the researcher estimates the average call length by including the value of  $z$  in formula 8.1.

$$510 - 1.96 \frac{46}{\sqrt{85}} \leq \mu \leq 510 + 1.96 \frac{46}{\sqrt{85}}$$

$$510 - 9.78 \leq \mu \leq 510 + 9.78$$

$$500.22 \leq \mu \leq 519.78$$

# Solution

---

- The confidence interval is constructed from the point estimate, which in this problem is 510 minutes, and the error of this estimate, which is 9.78 minutes.
  
  - The resulting confidence interval is  $500.22 \leq \mu \leq 519.78$ .
  
  - The cellular telephone company researcher is 95%, confident that the average length of a call for the population is between 500.22 and 519.78 minutes.
-

# Exercise

- A survey was taken of U.S. companies that do business with firms in India. One of the questions on the survey was: Approximately how many years has your company been trading with firms in India?
- A random sample of 44 responses to this question yielded a mean of 10.455 years. Suppose the population standard deviation for this question is 7.7 years.
- Using this information, construct a 90% confidence interval for the mean number of years that a company has been trading in India for the population of U.S. companies trading with firms in India.

# Solution

Here,  $n= 44$ ,  $\bar{x}= 10.455$  and  $\sigma= 7.7$ . To determine the value of  $z_{\alpha/2}$ , divide the

90% confidence in half, or take  $.5000 - \alpha/2 = .5000 - .0500 = 0.45$  where  $\alpha= 10\%$ .

Z tak

$$\bar{x} - z \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z \frac{\sigma}{\sqrt{n}}$$

The  $10.455 - 1.645 \frac{7.7}{\sqrt{44}} \leq \mu \leq 10.455 + 1.645 \frac{7.7}{\sqrt{44}}$

$$10.455 - 1.910 \leq \mu \leq 10.455 + 1.910$$

$$8.545 \leq \mu \leq 12.365$$

# Exercise

---

A study is conducted in a company that employs 800 engineers. A random sample of 50 engineers reveals that the average sample age is 34.3 years. Historically, the population standard deviation of the age of the company's engineers is approximately 8 years.

Construct a 98% confidence interval to estimate the average age of all the engineers in this company.

# Solution

- ❖ This problem has a finite population. The sample size, 50, is greater than 5% of the population, so the finite correction factor may be helpful.
- ❖ In this case  $N = 800$ ,  $n = 50$ ,  $\bar{x} = 34.3$  and  $\sigma = 8$
- ❖ The z value for a 98% confidence interval is 2.33

$$34.30 - 2.33 \frac{8}{\sqrt{50}} \sqrt{\frac{750}{799}} \leq \mu \leq 34.30 + 2.33 \frac{8}{\sqrt{50}} \sqrt{\frac{750}{799}}$$

$$34.30 - 2.55 \leq \mu \leq 34.30 + 2.55$$

$$31.75 \leq \mu \leq 36.85$$

# Estimating The Population Proportion



- Methods similar to those used earlier can be used to estimate the population proportion.
- The central limit theorem for sample proportions led to the following formula

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}}$$

- where  $q = 1 - p$ . Recall that this formula can be applied only when  $n \cdot p$  and  $n \cdot q$  are greater than 5.
- for confidence interval purposes only and for large sample sizes— is substituted for  $p$  in the denominator, yielding

$$z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}}$$

# Confidence Interval To Estimate P



$$\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}} \leq p \leq \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$$

where

$\hat{p}$  = sample proportion

$\hat{q} = 1 - \hat{p}$

$p$  = population proportion

$n$  = sample size

In this formula,  $\hat{p}$  is the point estimate and  $\pm z_{\alpha/2} \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$  is the error of the estimation.

# Example

---

- A study of 87 randomly selected companies with a telemarketing operation revealed that 39% of the sampled companies used telemarketing to assist them in order processing. Using this information, how could a researcher estimate the *population* proportion of telemarketing companies that use their telemarketing operation to assist them in order processing?

# Solution

---

The sample proportion,  $\hat{p} = .39$ , is the *point estimate* of the population proportion,  $p$ . For  $n = 87$  and  $\hat{p} = .39$ , a 95% confidence interval can be computed to determine the interval estimation of  $p$ . The  $z$  value for 95% confidence is 1.96. The value of  $\hat{q} = 1 - \hat{p} = 1 - .39 = .61$ . The confidence interval estimate is

$$.39 - 1.96\sqrt{\frac{(.39)(.61)}{87}} \leq p \leq .39 + 1.96\sqrt{\frac{(.39)(.61)}{87}}$$

$$.39 - .10 \leq p \leq .39 + .10$$

$$.29 \leq p \leq .49$$

# Exercise

---

- Coopers & Lybrand surveyed 210 chief executives of fast-growing small companies. Only 51% of these executives had a management succession plan in place. A spokesperson for Cooper & Lybrand said that many companies do not worry about management succession unless it is an immediate problem. However, the unexpected exit of a corporate leader can disrupt and unfocus a company for long enough to cause it to lose its momentum. Use the data given to compute a 92% confidence interval to estimate the propor

# Solution

The point estimate is the sample proportion given to be .51. It is estimated that .51, or 51% of all fast-growing small companies have a management succession plan. Realizing that the point estimate might change with another sample selection, we calculate a confidence interval.

The value of  $n$  is 210;  $\hat{p}$  is .51, and  $\hat{q} = 1 - \hat{p} = .49$ . Because the level of confidence is 92%, the value of  $z_{.04} = 1.75$ . The confidence interval is computed as

$$.51 - 1.75 \sqrt{\frac{(.51)(.49)}{210}} \leq p \leq .51 + 1.75 \sqrt{\frac{(.51)(.49)}{210}}$$

$$.51 - .06 \leq p \leq .51 + .06$$

$$.45 \leq p \leq .57$$

# Exercise

---

- A clothing company produces men's jeans. The jeans are made and sold with either a regular cut or a boot cut. In an effort to estimate the proportion of their men's jeans market in Oklahoma City that prefers boot-cut jeans, the analyst takes a random sample of 212 jeans sales from the company's two Oklahoma City retail outlets. Only 34 of the sales were for boot-cut jeans. Construct a 90% confidence interval to estimate the proportion of the population in Oklahoma City who prefer boot-cut jeans.

# Solution

The sample size is 212, and the number preferring boot-cut jeans is 34. The sample proportion is  $\hat{p} = 34/212 = .16$ . A point estimate for boot-cut jeans in the population is .16, or 16%. The z value for a 90% level of confidence is 1.645, and the value of  $\hat{q} = 1 - \hat{p} = 1 - .16 = .84$ . The confidence interval estimate is

$$.16 - 1.645 \sqrt{\frac{(.16)(.84)}{212}} \leq p \leq .16 + 1.645 \sqrt{\frac{(.16)(.84)}{212}}$$

$$.16 - .04 \leq p \leq .16 + .04$$

$$.12 \leq p \leq .20$$

# Estimating The Population Variance



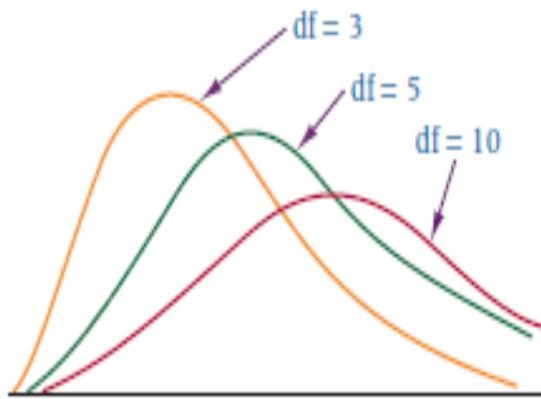
- Sample variance is computed by using the formula  $s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$
- Mathematical adjustment is made in the denominator by using  $n - 1$  to make the sample variance an unbiased estimator of the population variance.
- Suppose a researcher wants to estimate the population variance from the sample variance in a manner that is similar to the estimation of the population mean from a sample mean.
- The relationship of the sample variance to the population variance is captured by the **chi-square distribution**

# Chi-square Statistic

- Although the technique is still rather widely presented as a mechanism for constructing confidence intervals to estimate a population variance, you should proceed with extreme caution and **apply the technique only** in cases where the data are **normally** distributed. The technique lacks robustness.
- $\chi^2$  FORMULA FOR SINGLE VARIANCE (8.5)

$$\chi^2 = \frac{(n - 1)s^2}{\sigma^2}$$

$$df = n - 1$$



Three Chi-Square Distributions



# **Session 6**

## **(26<sup>th</sup> February, 2022)**

## **(Sampling & Testing of Hypothesis)**

- 
- ❖ Sampling
  - ❖ Sampling Distributions
  - ❖ Testing of Hypothesis
-

## Population of Wages of employees of an organization

1861	2495	1000	2497	1865	791	2090	2637	1327	1678
1680	2858	795	2495	2496	2501	1160	1480	1860	2490
2090	2840	2490	2640	659	827	2646	2638	2643	868
1327	1866	1861	2486	2865	3011	2494	1489	1865	2855
2840	2499	2093	2660	1165	2600	2085	2640	2998	1861
2956	2495	2865	1865	3000	3019	1670	2858	2642	1680
3038	3000	1313	596	656	3240	590	2501	2485	3015
2092	1679	3024	2497	2825	2630	2070	2900	1861	2636
2495	2637	2497	1159	2640	3050	870	2896	2500	2638
926	2860	1481	875	2482	1860	2086	934	3200	2490

## Select different samples of varied sizes

### Sample 1

3000    2486    820    1678    2070    2638    2490    1865    1000    2090    596    3200

### Sample 2

2840    2858    3000    2490    2998    3050    2070    2896    3200    2490    3280

### Sample 3

2858    3240    2497    2865    656    2093    934    1861    868    795

### Sample 4

2086    1000    2497    596    656    875    2085    934    1313

### Sample 5

820    1313    3000    2640    596    2640    2600    2495    934    2500

## Select different samples of varied sizes

### Sample 6

2840    2499    1327    1861    2495    3024    3038    2497

### Sample 7

2858    2490    868    1670    1480    2643    1480    1680    2085    2490

### Sample 8

2495    2858    1861    2092    2499    3000    2660    1000    1679    926    2660

### Sample 9

795    791    3200    2085    2638    2497    2486    1159    2640

### Sample 10

3019    3240    3200    3050    3000    3015    2900    2896    2998

**Compute sample mean of these samples**

<b>Sample No.</b>	<b>Sample size</b>	<b>Mean</b>	<b>SD</b>
1	12	1994.42	843.23
2	11	2830.18	349.94
3	10	1866.70	988.57
4	9	1338.00	704.36
5	10	1953.80	920.44
6	8	2447.63	590.64
7	10	1974.40	638.05
8	11	2157.27	715.10
9	9	2032.33	891.53
10	9	3035.33	117.40
<b>Overall</b>	<b>100</b>	<b>2162.24</b>	<b>732.26</b>

# Sampling Variability

---

- The term "sampling variability" refers to the fact that the statistical information from a sample (called a *statistic*) will vary as the random sampling is repeated.
  - **Sampling variability will decrease as the sample size increases.**
  - the samples must be randomly chosen, must be of the same size (not smaller than 30), and the more samples that are used, the more reliable the information gathered will be.
-

Do you consider these sample means and sample SDs as variable ?

If yes, should we not describe the distribution of these variables ?

The distribution of the sample estimates is called sampling distribution

For example the distribution of sample means is called Sampling distribution of mean

# Definition

---

- The probability distribution of a statistic (sample estimate) is called sampling distribution.
  
  - The sampling distribution of a statistic depends on the distribution of the population, the size of the sample, and the method of sample selection.
-

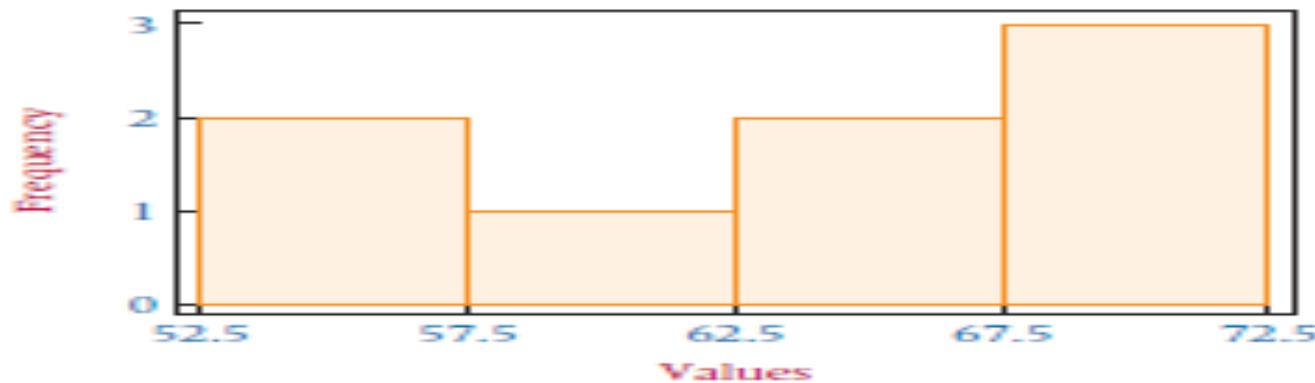
# Sampling Distribution Of $\bar{x}$

---

- The sample mean is one of the more common statistics used in the inferential process.
  - The **distribution** of the values of the sample mean ( $\bar{x}$ ) in repeated **samples** is called the **sampling distribution of  $\bar{x}$**
  - One way to examine the distribution possibilities is to take a population with a particular distribution, randomly select samples of a given size, compute the sample means, and attempt to determine how the means are distributed.
-

# Example

- Suppose a small finite population consists of only  $N = 8$  numbers:
- 54 55 59 63 64 68 69 70
- Using an Excel-produced histogram, we can see the shape of the distribution of this population of data.



- Suppose we take all possible samples of size  $n = 2$  from this population with replacement.

# Example

The result is the following pairs of data.

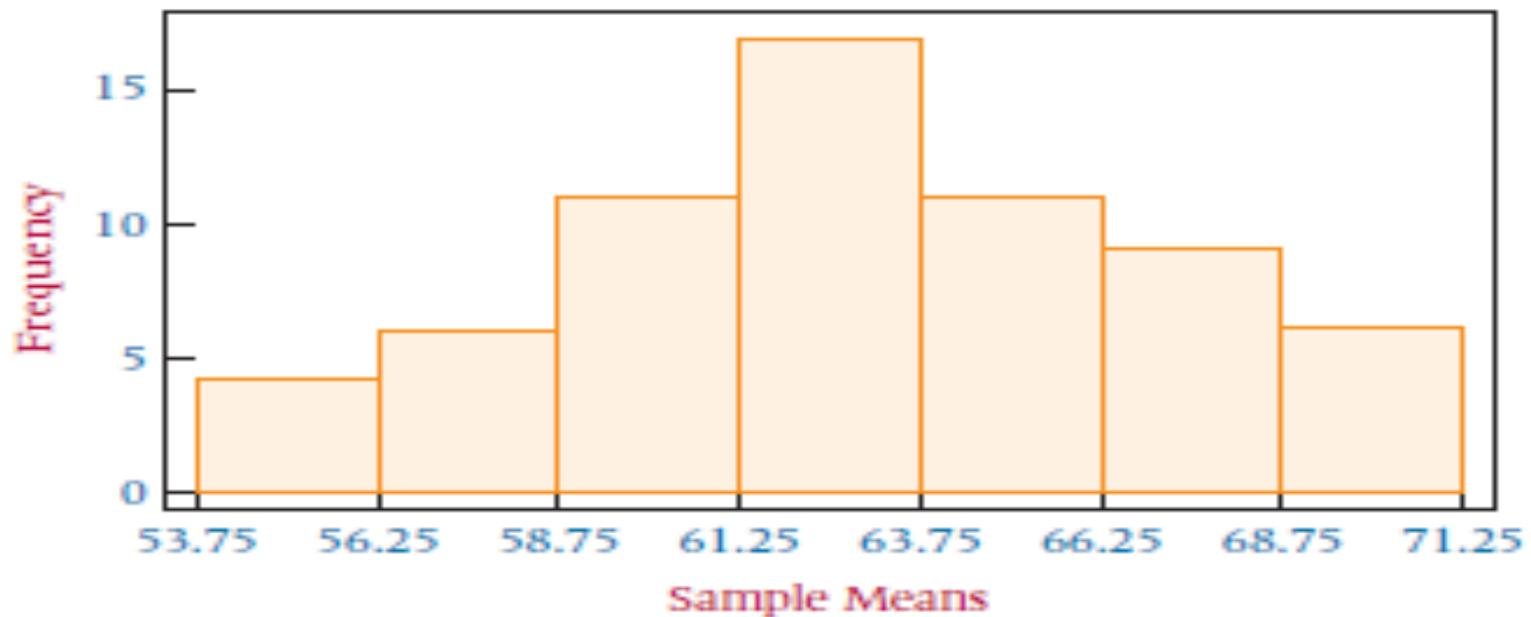
(54,54)	(55,54)	(59,54)	(63,54)
(54,55)	(55,55)	(59,55)	(63,55)
(54,59)	(55,59)	(59,59)	(63,59)
(54,63)	(55,63)	(59,63)	(63,63)
(54,64)	(55,64)	(59,64)	(63,64)
(54,68)	(55,68)	(59,68)	(63,68)
(54,69)	(55,69)	(59,69)	(63,69)
(54,70)	(55,70)	(59,70)	(63,70)
(64,54)	(68,54)	(69,54)	(70,54)
(64,55)	(68,55)	(69,55)	(70,55)
(64,59)	(68,59)	(69,59)	(70,59)
(64,63)	(68,63)	(69,63)	(70,63)
(64,64)	(68,64)	(69,64)	(70,64)
(64,68)	(68,68)	(69,68)	(70,68)
(64,69)	(68,69)	(69,69)	(70,69)
(64,70)	(68,70)	(69,70)	(70,70)

The means of each of these samples follow.

54	54.5	56.5	58.5	59	61	61.5	62
54.5	55	57	59	59.5	61.5	62	62.5
56.5	57	59	61	61.5	63.5	64	64.5
58.5	59	61	63	63.5	65.5	66	66.5
59	59.5	61.5	63.5	64	66	66.5	67
60	61.5	63.5	65.5	66	68	68.5	69
61.5	62	64	66	66.5	68.5	69	69.5
62	62.5	64.5	66.5	67	69	69.5	70

# Example

- Again using an Excel-produced histogram, we can see the shape of the distribution of these sample means.



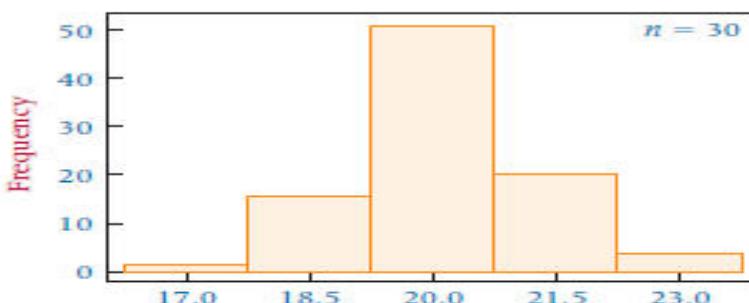
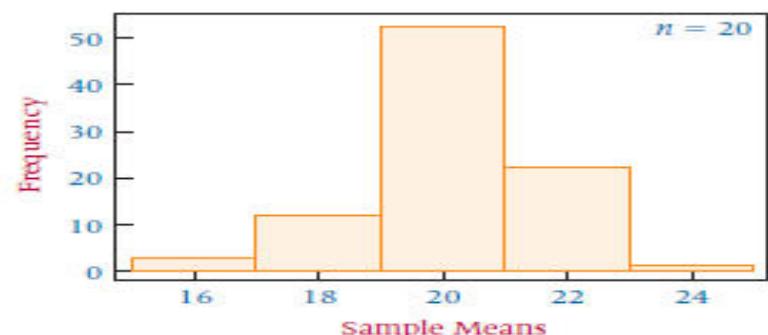
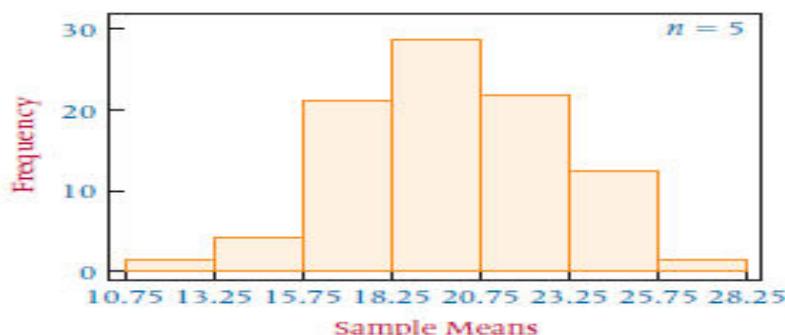
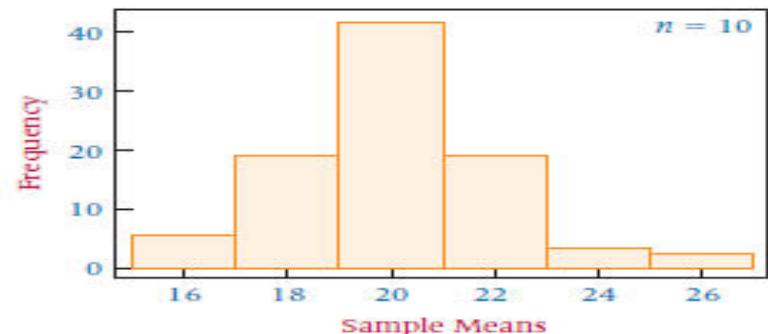
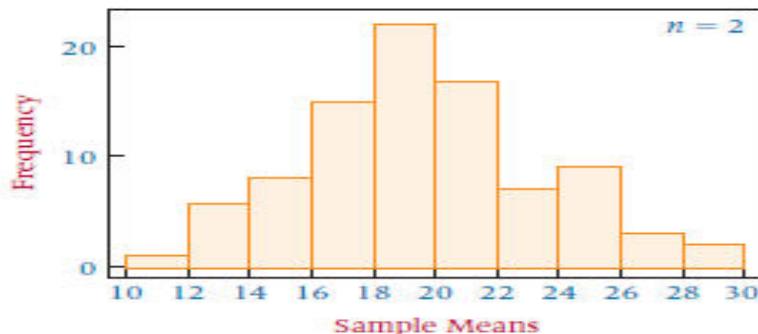
# Conclusions

---

- Notice that the shape of the histogram for sample means is quite unlike the shape of the histogram for the population.
  - The sample means appear to “pile up” toward the middle of the distribution and “tail off” toward the extremes.
  - As sample sizes become much larger, the sample mean distributions begin to approach a **normal distribution** and the variation among the means decreases.
-

## Sample Means from 90 Samples Ranging in Size from $n = 2$ to $n =$

**30 from a Uniformly Distributed Population with  $a = 10$  and  $b = 30$**



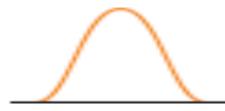
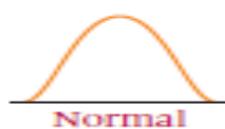
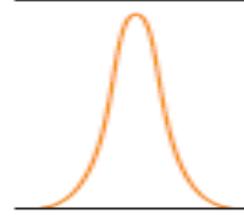
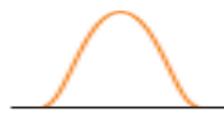
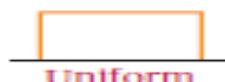
# Shapes of the Distributions of Sample Means

Population Distribution

$n = 2$

$n = 5$

$n = 30$



# Central Limit Theorem

- If samples of size  $n$  are drawn randomly from a population that has a mean of  $\mu$  and a standard deviation of  $\sigma$ , the sample means,  $\bar{x}$ , are approximately normally distributed for sufficiently large sample sizes ( $n \geq 30$ ) regardless of the shape of the population distribution.
- If the population is normally distributed, the sample means are normally distributed for any size sample.
- From mathematical expectation

$$\mu_{\bar{x}} = \mu \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

# Z score for sample means

- The central limit theorem states that sample means are normally distributed regardless of the shape of the population for large samples and for any sample size with normally distributed populations.
- Thus, **sample means** can be **analyzed** by using **z scores**
- The formula to 
$$z = \frac{x - \mu}{\sigma}$$
 z scores for individual values from a normal distribution
- If sample means are normally distributed, the z score formula applied to sample means would be

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}$$

- The standard deviation of the statistic of interest is  $\sigma_{\bar{x}}$ , sometimes referred to as the **standard error of the mean**.

# Example

---

Suppose the mean expenditure per customer at a tire store is \$85.00, with a standard deviation of \$9.00.

If a random sample of 40 customers is taken, what is the probability that the sample average expenditure per customer for this sample will be \$87.00 or more?

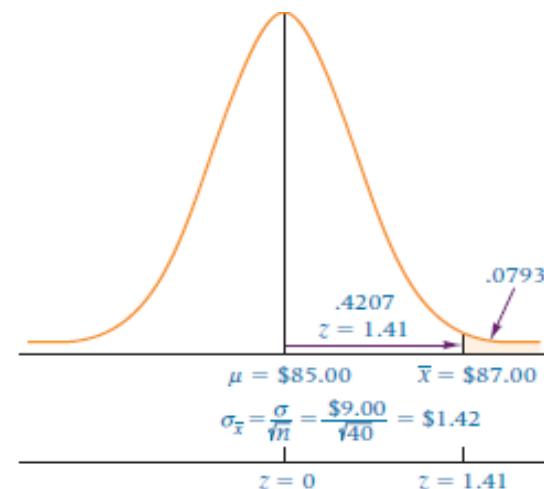
# Solution

Because the sample size is greater than 30, the central limit theorem

Can be used, and the sample means are normally distributed.

$$\mu = \$85 \quad \sigma = \$9$$

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\$87.00 - \$85.00}{\frac{\$9.00}{\sqrt{40}}} = \frac{\$2.00}{\$1.42} = 1.41$$



# Exercise

---

- Suppose that during any hour in a large department store, the average number of shoppers is 448, with a standard deviation of 21 shoppers.
  
  - What is the probability that a random sample of 49 different shopping hours will yield a sample mean between 441 and 446 shoppers?
-

# Solution

For this problem,  $\mu = 448$ ,  $\sigma = 21$ , and  $n = 49$ . The problem is to determine

$$P(441 \leq \bar{x} \leq 446).$$

The following

$$z = \frac{441 - 448}{\frac{21}{\sqrt{49}}} = \frac{-7}{3} = -2.33$$

$$z = \frac{446 - 448}{\frac{21}{\sqrt{49}}} = \frac{-2}{3} = -0.67$$

# Sampling from a Finite Population

---

- The earlier example was based on the assumption that the population was infinitely or extremely large.
- In cases of a finite population, *a statistical adjustment can be made to the z formula for sample means*. The adjustment is called the **finite correction factor**

$$\sqrt{\frac{N-n}{N-1}}.$$

- Following is the z formula for sample means when samples are drawn from finite populations.

$$z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}}$$

# Rules for finite population

---

- As the size of the finite population becomes larger in relation to sample size, the finite correction factor approaches 1.
  
  - In theory, whenever researchers are working with a finite population, they can use the finite correction factor.
  
  - A rough rule of thumb for many researchers is that, if the sample size is **less** than **5%** of the finite population size or  **$n/N < 0.05$** , the finite correction factor does **not** significantly modify the solution.
-

# Exercise

- A production company's 350 hourly employees average 37.6 years of age, with a standard deviation of 8.3 years.
  
- If a random sample of 45 hourly employees is taken, what is the probability that the sample will have an average age of less than 40 years?

# Solution

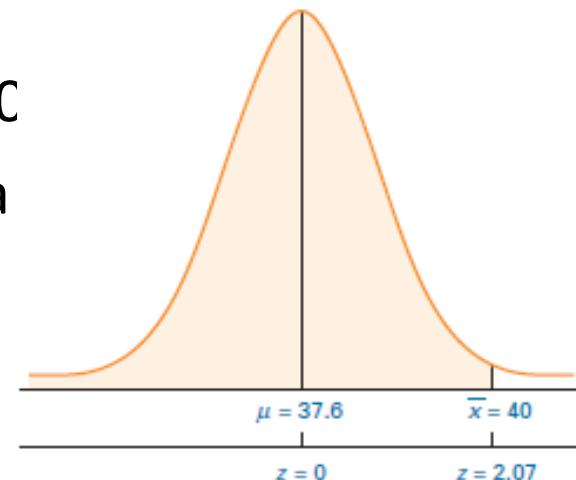
---

- The population mean is 37.6, with a population standard deviation of 8.3.
- The sample size is 45, but it is being drawn from a finite population of 350; that is,  $n = 45$  and  $N = 350$ .
- The sample mean under consideration is 40
- Using the z formula with the finite correction factor gives

$$z = \frac{40 - 37.6}{\frac{8.3}{\sqrt{45}} \sqrt{\frac{350 - 45}{350 - 1}}} = \frac{2.4}{1.157} = 2.07$$

# Solution

- This z value yields a probability of .480
- Therefore, the probability of getting a
- sample average age of less than
- 40 years is **.4808 + .5000 = .9808.**



# Sampling Distribution Of Sample Proportion

---

- If research results in **countable** items such as how many people in a sample have a flexible work schedule, the sample proportion is often the statistic of choice.

## SAMPLE PROPORTION

$$\hat{p} = \frac{x}{n}$$

where

$x$  = number of items in a sample that have the characteristic

$n$  = number of items in the sample

# Example

---

- In a sample of 100 factory workers, 30 workers might belong to a union.
  
- The value of sample proportion for this characteristic, union membership, is

$$30/100 = 0.30$$

# How does a researcher use the sample proportion in analysis?

---



- The central limit theorem applies to sample proportions in that the normal distribution approximates the shape of the distribution of sample proportions
- If  $n*p > 5$  and  $n*q > 5$  ( $p$  is the population proportion and  $q = 1 - p$ ).
- The mean of sample proportions for all samples of size  $n$  randomly drawn from a population is  $p$  (the population proportion) and the standard deviation of sample proportions is  $\sqrt{\frac{p \cdot q}{n}}$
- sometimes referred to as the **standard error of the proportion**

# Z Formula For Sample Proportions

---

For  $n * p > 5$  and  $n * q > 5$

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}}$$

where

$\hat{p}$  = sample proportion

$n$  = sample size

$p$  = population proportion

$q = 1 - p$

# Example

---

Suppose 60% of the electrical contractors in a region use a particular brand of wire. What is the probability of taking a random sample of size 120 from these electrical contractors and finding that .50 or less use that brand of wire?

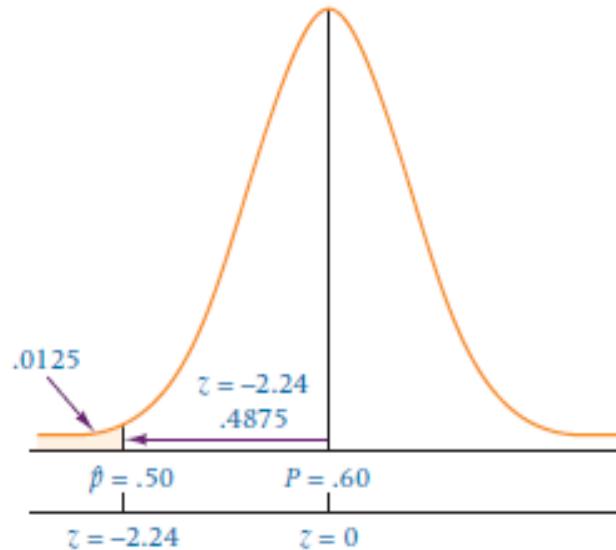
# Solution

$$p = .60 \quad \hat{p} = .50 \quad n = 120$$

The  $z$  formula yields

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{.50 - .60}{\sqrt{\frac{(.60)(.40)}{120}}} = \frac{-10}{.0447} = -2.24$$

ding to



$= -2.24$  is  $.4875$ .

For  $z < -2.24$  (the tail of the distribution), the answer is  $.5000 - .4875 = .0125$ .

# Exercise

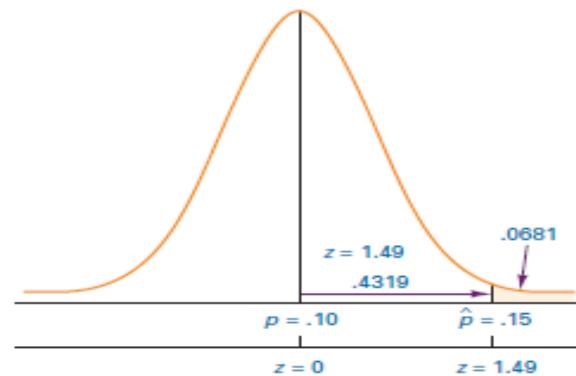
---

If 10% of a population of parts is defective, what is the probability of randomly selecting 80 parts and finding that 12 or more parts are defective?

# Solution

Here,  $p = .10$ ,  $\hat{p} = 12/80 = .15$ , and  $n = 80$ . Entering these values in the z formula yields

$$z = \frac{.15 - .10}{\sqrt{\frac{(.10)(.90)}{80}}} = \frac{.05}{.0335} = 1.49$$



$$P(\hat{p} \geq .15) = .5000 - .4319 = .0681.$$

- The probability of .4319 for a z value of 1.49, which is the area between the sample proportion, .15, and the population proportion, .10. The answer to the question is

# Forms Of Statistical Inference

- ❖ Three forms of statistical inference

- Point estimation
- Interval estimation
- Hypothesis testing

# Point Estimate

- A **point estimate** is a statistic taken from a sample that is used to estimate a population parameter.
- A point estimate is only as good as the representativeness of its sample.
- If other random samples are taken from the population, the point estimates derived from those samples are likely to vary.

# Interval Estimate

- Because of variation in sample statistics, estimating a population parameter with an interval estimate is often preferable to using a point estimate.
- An interval estimate (confidence interval) is a range of values within which the analyst can declare, with some confidence, the population parameter lies.

# Confidence Interval to Estimate $\mu$

100(1 -  $\alpha$ )% CONFIDENCE  
INTERVAL TO ESTIMATE  $\mu$ :

$\sigma$  KNOWN (8.1)

or

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

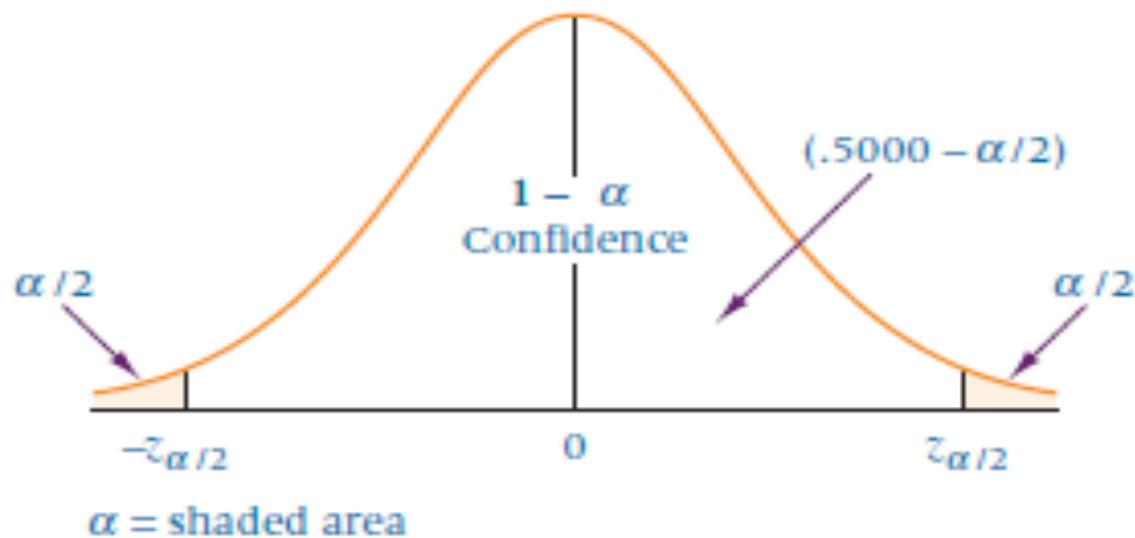
$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where

$\alpha$  = the area under the normal curve outside the confidence interval area

$\alpha/2$  = the area in one end (tail) of the distribution outside the confidence interval

# Confidence Intervals



# Example

---

- In the cellular telephone company problem of estimating the population mean number of minutes called per residential user per month, from the sample of 85 bills it was determined that the sample mean is 510 minutes.
  
  - Suppose past history and similar studies indicate that the population standard deviation is 46 minutes.
  
  - Determine a 95% confidence interval.
-

# Solution

The business researcher can now complete the cellular telephone problem. To determine a 95% confidence interval for  $\bar{x} = 510$ ,  $\sigma = 46$ ,  $n = 85$ , and  $z = 1.96$ , the researcher estimates the average call length by including the value of  $z$  in formula 8.1.

$$510 - 1.96 \frac{46}{\sqrt{85}} \leq \mu \leq 510 + 1.96 \frac{46}{\sqrt{85}}$$

$$510 - 9.78 \leq \mu \leq 510 + 9.78$$

$$500.22 \leq \mu \leq 519.78$$

# Solution

---

- The confidence interval is constructed from the point estimate, which in this problem is 510 minutes, and the error of this estimate, which is 9.78 minutes.
  
  - The resulting confidence interval is  $500.22 \leq \mu \leq 519.78$ .
  
  - The cellular telephone company researcher is 95%, confident that the average length of a call for the population is between 500.22 and 519.78 minutes.
-

# Exercise

- A survey was taken of U.S. companies that do business with firms in India. One of the questions on the survey was: Approximately how many years has your company been trading with firms in India?
- A random sample of 44 responses to this question yielded a mean of 10.455 years. Suppose the population standard deviation for this question is 7.7 years.
- Using this information, construct a 90% confidence interval for the mean number of years that a company has been trading in India for the population of U.S. companies trading with firms in India.

# Solution

Here,  $n= 44$ ,  $\bar{x}= 10.455$  and  $\sigma= 7.7$ . To determine the value of  $z_{\alpha/2}$ , divide the

90% confidence in half, or take  $.5000 - \alpha/2 = .5000 - .0500 = 0.45$  where  $\alpha= 10\%$ .

Z tak

$$\bar{x} - z \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z \frac{\sigma}{\sqrt{n}}$$

The  $10.455 - 1.645 \frac{7.7}{\sqrt{44}} \leq \mu \leq 10.455 + 1.645 \frac{7.7}{\sqrt{44}}$

$$10.455 - 1.910 \leq \mu \leq 10.455 + 1.910$$

$$8.545 \leq \mu \leq 12.365$$

# Exercise

---

A study is conducted in a company that employs 800 engineers. A random sample of 50 engineers reveals that the average sample age is 34.3 years. Historically, the population standard deviation of the age of the company's engineers is approximately 8 years.

Construct a 98% confidence interval to estimate the average age of all the engineers in this company.

# Solution

- ❖ This problem has a finite population. The sample size, 50, is greater than 5% of the population, so the finite correction factor may be helpful.
- ❖ In this case  $N = 800$ ,  $n = 50$ ,  $\bar{x} = 34.3$  and  $\sigma = 8$
- ❖ The z value for a 98% confidence interval is 2.33

$$34.30 - 2.33 \frac{8}{\sqrt{50}} \sqrt{\frac{750}{799}} \leq \mu \leq 34.30 + 2.33 \frac{8}{\sqrt{50}} \sqrt{\frac{750}{799}}$$

$$34.30 - 2.55 \leq \mu \leq 34.30 + 2.55$$

$$31.75 \leq \mu \leq 36.85$$

# Estimating The Population Proportion



- Methods similar to those used earlier can be used to estimate the population proportion.
- The central limit theorem for sample proportions led to the following formula

$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}}$$

- where  $q = 1 - p$ . Recall that this formula can be applied only when  $n \cdot p$  and  $n \cdot q$  are greater than 5.
- for confidence interval purposes only and for large sample sizes— is substituted for  $p$  in the denominator, yielding

$$z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}}$$

# Confidence Interval To Estimate P



$$\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}} \leq p \leq \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$$

where

$\hat{p}$  = sample proportion

$\hat{q} = 1 - \hat{p}$

$p$  = population proportion

$n$  = sample size

In this formula,  $\hat{p}$  is the point estimate and  $\pm z_{\alpha/2} \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$  is the error of the estimation.

# Example

---

- A study of 87 randomly selected companies with a telemarketing operation revealed that 39% of the sampled companies used telemarketing to assist them in order processing. Using this information, how could a researcher estimate the *population* proportion of telemarketing companies that use their telemarketing operation to assist them in order processing?

# Solution

---

The sample proportion,  $\hat{p} = .39$ , is the *point estimate* of the population proportion,  $p$ . For  $n = 87$  and  $\hat{p} = .39$ , a 95% confidence interval can be computed to determine the interval estimation of  $p$ . The  $z$  value for 95% confidence is 1.96. The value of  $\hat{q} = 1 - \hat{p} = 1 - .39 = .61$ . The confidence interval estimate is

$$.39 - 1.96\sqrt{\frac{(.39)(.61)}{87}} \leq p \leq .39 + 1.96\sqrt{\frac{(.39)(.61)}{87}}$$

$$.39 - .10 \leq p \leq .39 + .10$$

$$.29 \leq p \leq .49$$

# Exercise

---

- Coopers & Lybrand surveyed 210 chief executives of fast-growing small companies. Only 51% of these executives had a management succession plan in place. A spokesperson for Cooper & Lybrand said that many companies do not worry about management succession unless it is an immediate problem. However, the unexpected exit of a corporate leader can disrupt and unfocus a company for long enough to cause it to lose its momentum. Use the data given to compute a 92% confidence interval to estimate the propor

# Solution

The point estimate is the sample proportion given to be .51. It is estimated that .51, or 51% of all fast-growing small companies have a management succession plan. Realizing that the point estimate might change with another sample selection, we calculate a confidence interval.

The value of  $n$  is 210;  $\hat{p}$  is .51, and  $\hat{q} = 1 - \hat{p} = .49$ . Because the level of confidence is 92%, the value of  $z_{.04} = 1.75$ . The confidence interval is computed as

$$.51 - 1.75 \sqrt{\frac{(.51)(.49)}{210}} \leq p \leq .51 + 1.75 \sqrt{\frac{(.51)(.49)}{210}}$$

$$.51 - .06 \leq p \leq .51 + .06$$

$$.45 \leq p \leq .57$$

# Exercise

---

- A clothing company produces men's jeans. The jeans are made and sold with either a regular cut or a boot cut. In an effort to estimate the proportion of their men's jeans market in Oklahoma City that prefers boot-cut jeans, the analyst takes a random sample of 212 jeans sales from the company's two Oklahoma City retail outlets. Only 34 of the sales were for boot-cut jeans. Construct a 90% confidence interval to estimate the proportion of the population in Oklahoma City who prefer boot-cut jeans.

# Solution

The sample size is 212, and the number preferring boot-cut jeans is 34. The sample proportion is  $\hat{p} = 34/212 = .16$ . A point estimate for boot-cut jeans in the population is .16, or 16%. The z value for a 90% level of confidence is 1.645, and the value of  $\hat{q} = 1 - \hat{p} = 1 - .16 = .84$ . The confidence interval estimate is

$$.16 - 1.645 \sqrt{\frac{(.16)(.84)}{212}} \leq p \leq .16 + 1.645 \sqrt{\frac{(.16)(.84)}{212}}$$

$$.16 - .04 \leq p \leq .16 + .04$$

$$.12 \leq p \leq .20$$

# Estimating The Population Variance



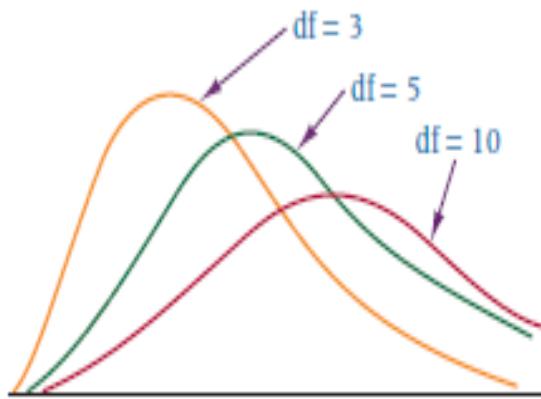
- Sample variance is computed by using the formula  $s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$
- Mathematical adjustment is made in the denominator by using  $n - 1$  to make the sample variance an unbiased estimator of the population variance.
- Suppose a researcher wants to estimate the population variance from the sample variance in a manner that is similar to the estimation of the population mean from a sample mean.
- The relationship of the sample variance to the population variance is captured by the **chi-square distribution**

# Chi-square Statistic

- Although the technique is still rather widely presented as a mechanism for constructing confidence intervals to estimate a population variance, you should proceed with extreme caution and **apply the technique only** in cases where the data are **normally** distributed. The technique lacks robustness.
- $\chi^2$  FORMULA FOR SINGLE VARIANCE (8.5)

$$\chi^2 = \frac{(n - 1)s^2}{\sigma^2}$$

$$df = n - 1$$



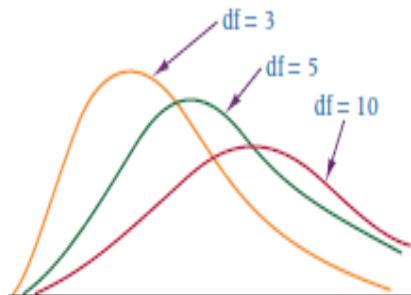
Three Chi-Square Distributions



# **Testing of Hypothesis 2 (Chi Square Test)**

# Chi-square Statistic

- Although the technique is still rather widely presented as a mechanism for constructing confidence intervals to estimate a population variance, you should proceed with extreme caution and **apply the technique only in cases where the population is known to be normally distributed.** We can say that this technique lacks robustness.



Three Chi-Square Distributions

$\chi^2$  FORMULA FOR SINGLE  
VARIANCE (8.5)

$$\chi^2 = \frac{(n - 1)s^2}{\sigma^2}$$

$$df = n - 1$$

## Testing of Hypothesis → Chi-square test: Independence

### Chi-square test

#### Chi-square Test

Independence

Goodness-of-fit

Should be applied ONLY for Frequencies

Not for percentages, ratios, mean etc.

## Testing of Hypothesis → Chi-square test: Independence

Based on attributes used to test

(a) INDEPENDENCE of two different categorical variables

or

(b) GOODNESS OF FIT

Caution:

Should be applied ONLY for FREQUENCIES not for  
percentages, ratios, mean etc

## Testing of Hypothesis → Chi-square test: Independence

### Hypothesis for testing independence

The hypothesis to be tested for independence will be

$H_0$ : The two categorical variables may be independent (may not be associated)

$H_1$ : The two categorical variables may not be independent (may be associated)

## Testing of Hypothesis → Chi-square test: Independence

### Procedure for testing independence

To check the independence (no association) between the two categorical variables, the statistical test used is Chi-square test given by

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}, k = r \times c \text{ # of cells}$$

The test-statistic follows Chi-square distribution with  $(r-1)(c-1)$  degrees of freedom.  $r = \# \text{ of rows}$ ,  $c = \# \text{ of columns}$

## Testing of Hypothesis → Chi-square test: Independence

**Expected frequencies**

$$E_{ij} = \frac{r_i c_j}{n},$$

for  $i=1, 2, \dots, m$ ;  $j=1, 2, \dots, n$

**Chi-square is calculated by**

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i} \approx \chi^2_{[(r-1)(c-1)]}$$

**where  $k = r \times c$  is the total number of cells in the  $r \times c$  contingency table,  $r =$  total no. of rows and  $c$  is total no. of columns.**

## Testing of Hypothesis → Chi-square test: Independence

A study to find the association between smoking and ca. lung has revealed the following data? Find is there any association exists between smoking and ca. lung?

Smoking	Carcinoma of lung		Total
	Present	Absent	
Smokers	69	2431	2500
Non-smokers	24	1476	1500
Total	93	3907	4000

## Testing of Hypothesis → Chi-square test: Independence

### Calculation of expected frequencies

$$E_1 = \frac{c_1 r_1}{n} = \frac{93 * 2500}{4000} = 58.125$$

$$E_2 = \frac{c_2 r_1}{n} = \frac{3907 * 2500}{4000} = 2441.875$$

$$E_3 = \frac{c_1 r_2}{n} = \frac{93 * 1500}{4000} = 34.875$$

$$E_4 = \frac{c_2 r_2}{n} = \frac{3907 * 1500}{4000} = 1465.125$$

## Testing of Hypothesis → Chi-square test: Independence

### Calculation of Chi-square statistic - $\chi^2$

SI No	Observed frequencies ( $O_i$ )	Expected frequencies ( $E_i$ )	$(O_i - E_i)$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
1	69	58.125	10.875	118.266	2.035
2	2431	2441.875	-10.875	118.266	0.048
3	24	34.875	-10.875	118.266	3.391
4	1476	1465.125	10.875	118.266	0.081
Total	4000	4000			$\chi^2 = 5.555$

## Testing of Hypothesis → Chi-square test: Independence

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of $\chi^2$								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21

## Testing of Hypothesis → Chi-square test: Independence

### Calculation of P – value

$$\frac{6.63 - 3.84}{5.56 - 3.84} = \frac{0.05 - 0.01}{P - 0.01}$$

$$P = 0.01 + \frac{(0.05 - 0.01) * (5.56 - 3.84)}{(6.63 - 3.84)} = \mathbf{0.035}$$

## Testing of Hypothesis → Chi-square test: Independence

### Interpretation

$H_0$ : Smoking habit and Cancer of lung may be independent (may not be associated)

$H_1$ : Smoking habit and Cancer of lung may not be independent (may be associated)

$$\chi^2 = 5.555$$

df = 1, Critical value at  $\alpha = 0.05$  is 3.841, P = 0.035

Inference: There may be an association between smoking and Cancer of lung

## Testing of Hypothesis → Chi-square test: Independence

By replacing  $O_1$ ,  $O_2$ ,  $O_3$ , and  $O_4$ , by  $a$ ,  $b$ ,  $c$ , and  $d$  the  $2 \times 2$  contingency table can also be written as

Categorical variable 1	Categorical variable 2		Total
	Response 1	Response 2	
Response 1	$a$	$b$	$r_1$
Response 2	$c$	$d$	$r_2$
Total	$c_1$	$c_2$	$n$

## Testing of Hypothesis → Chi-square test: Independence



To assess the length of hospital stay and the type of insurance, data were taken on 70 individuals



Type of Insurance	Length of Hospital Stay (days)		Total
	≤10	>10	
Type 1	42	3	45
Type 2	18	7	25
Total	60	10	70

Examine whether Chi-square test can be applied to this data to test the independence between type of insurance and length of hospital stay?

## Testing of Hypothesis → Chi-square test: Independence

### Calculation of expected frequencies

$$E_1 = \frac{c_1 r_1}{n} = \frac{60 * 45}{70} = 38.75$$

$$E_2 = \frac{c_2 r_1}{n} = \frac{10 * 45}{70} = 6.43$$

$$E_3 = \frac{c_1 r_2}{n} = \frac{60 * 25}{70} = 21.43$$

$$E_4 = \frac{c_2 r_2}{n} = \frac{10 * 25}{70} = 3.57$$

## Testing of Hypothesis → Chi-square test: Independence

### Calculation of Chi-square statistic - $\chi^2$

SI No	Observed frequencies ( $O_i$ )	Expected frequencies ( $E_i$ )	$(O_i - E_i)$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
1	42	38.57	-	-	-
2	3	6.43	-	-	-
3	18	21.43	-	-	-
4	7	3.57	-	-	-
<b>Total</b>	<b>70</b>	<b>70</b>			<b>???</b>

## Testing of Hypothesis → Chi-square test: Independence

Since the expected frequency in the 4 is less than 5  
Chi-square cannot be applied and hence the Fisher's exact probabilities has to be calculated.

## Testing of Hypothesis → Chi-square test: Independence



To assess the length of hospital stay and the type of insurance, data were taken on 70 individuals



Type of Insurance	Length of Hospital Stay (days)		Total
	≤10	>10	
Type 1	42	3	45
Type 2	13	12	25
Total	55	15	70

Examine whether Chi-square test can be applied to this data to test the independence between type of insurance and length of hospital stay?

## Testing of Hypothesis → Chi-square test: Independence

### Calculation of expected frequencies

$$E_1 = \frac{r_1 c_1}{n} = \frac{45 * 55}{70} = 35.36$$

$$E_2 = \frac{r_1 c_2}{n} = \frac{45 * 15}{70} = 9.64$$

$$E_3 = \frac{r_2 c_1}{n} = \frac{25 * 55}{70} = 19.64$$

$$E_4 = \frac{r_2 c_2}{n} = \frac{25 * 15}{70} = 5.36$$

$$\chi^2 = \frac{(42 - 35.36)^2}{35.36} + \frac{(3 - 9.64)^2}{9.64} + \frac{(13 - 19.64)^2}{19.64} + \frac{(12 - 5.36)^2}{5.36} = 16.307$$

## Testing of Hypothesis → Chi-square test: Independence

- $H_0$ : Duration of hospital stay and type of insurance plan may be independent (not associated)
- $H_1$ : Duration of hospital stay and type of insurance plan may not be independent (Associated)
- $\chi^2 = 16.307$
- $df=1$
- $P < 0.001$
- Inference: Reject  $H_0$ , which shows duration of hospital stay and type of insurance may be associated

## Testing of Hypothesis → Chi-square test: Independence

Hypertension	Non smokers	Moderate smokers	Heavy smokers	Total
A	$O_1$	$O_2$	$O_3$	$r_1$
B	$O_4$	$O_5$	$O_6$	$r_2$
Total	$c_1$	$c_2$	$c_3$	$n$

The expected frequencies and Chi-square statistic are computed by

$$E_1 = \frac{r_1 c_1}{n}, \quad E_2 = \frac{r_1 c_2}{n}, \quad E_3 = \frac{r_1 c_3}{n}, \quad E_4 = \frac{r_2 c_1}{n},$$

$$E_5 = \frac{r_2 c_2}{n}, \quad E_6 = \frac{r_2 c_3}{n}, \quad \text{and} \quad \chi^2 = \sum_{i=1}^6 \frac{(O_i - E_i)^2}{E_i}$$

## Testing of Hypothesis → Chi-square test: Independence

Under the null hypothesis, the observed frequencies and the calculated expected frequencies will be as follows:

Hypertension	Non smokers	Moderate smokers	Heavy smokers	Total
A	$O_1$ $E_1$	$O_2$ $E_2$	$O_3$ $E_3$	$r_1$
B	$O_4$ $E_4$	$O_5$ $E_5$	$O_6$ $E_6$	$r_2$
Total	$c_1$	$c_2$	$c_3$	$n$

## Testing of Hypothesis → Chi-square test: Independence



Three pension plans

Independent of job classification

Use  $\alpha = 0.05$

The opinion of a random sample of 500 employees are shown below

Job Classification	Pension Plan			Total
	1	2	3	
Salaried workers	166	86	68	320
Hourly workers	84	64	32	180
Total	250	150	100	500

$$E_1 = \frac{r_1 c_1}{n} = \frac{320 \times 250}{500} = 106.24$$

$$E_2 = \frac{r_1 c_2}{n} = \frac{320 \times 150}{500} = 96.00$$

$$E_3 = \frac{r_1 c_3}{n} = \frac{320 \times 100}{500} = 64.00$$

$$E_4 = \frac{r_2 c_1}{n} = \frac{180 \times 250}{500} = 90.00$$

$$E_5 = \frac{r_2 c_2}{n} = \frac{180 \times 150}{500} = 54.60$$

$$E_6 = \frac{r_2 c_3}{n} = \frac{180 \times 100}{500} = 36.00$$

## Testing of Hypothesis → Chi-square test: Independence

SI No	$(O_i)$	$(E_i)$	$(O_i - E_i)$	$(O_i - E_i)^2$	$(O_i - E_i)^2 / E_i$
1	166	106.24	59.76	3571.26	33.62
2	86	96.00	-10.00	100.00	1.04
3	68	64.00	4.00	16.00	0.25
4	84	90.00	-6.00	36.00	0.40
5	64	54.60	9.40	88.36	1.62
6	32	36.00	-4.00	16.00	0.44
Total	180	180	Chi-square value	37.37	

## Testing of Hypothesis → Chi-square test: Independence

- $H_0$ : Job satisfaction and pension plan may be independently distributed (not associated)
- $H_1$ : Job satisfaction and pension plan may not be independently distributed (Associated)
- $\chi^2 = 37.37$
- $df=2$
- $P<0.001$
- Inference: Reject  $H_0$ , which shows Job satisfaction and pension plan are associated

# Problem



A sample analysis of examination results of 500 students was made. It was found that 220 students had failed, 170 had secured a third class, 90 were placed in second class and 20 got a first class.

Do these figures commensurate with the general examination result which is in ratio of 4:3:2:1 for the various categories respectively.

$H_0$ : The observed results commensurate with the general examination results.

$H_1$ : The observed results not commensurate with the general examination results.

Expected frequencies are in the ratio of 4:3:2:1

Total frequency = 500

If we divide total frequency 500 in the ratio 4:3:2:1, We get the  
Expected frequencies

As  $200\left(500 \times \frac{4}{10}\right)$ ,  $150\left(500 \times \frac{3}{10}\right)$ ,  $100\left(500 \times \frac{2}{10}\right)$ ,  $50\left(500 \times \frac{1}{10}\right)$

S. NO	(O <sub>i</sub> )	(E <sub>i</sub> )	(O <sub>i</sub> -E <sub>i</sub> )	(O-E) <sup>2</sup>	(O-E) <sup>2</sup> /E <sub>i</sub>
1	220	200	20	400	2
2	170	150	20	400	2.667
3	90	100	10	100	1
4	20	50	30	900	18

## The Chi-square value is

$$\diamond \quad \chi^2 = \sum_1^4 \frac{(O-E)^2}{E} = 2 + 2.667 + 1 + 18 \\ = 23.667$$

Dof is  $4-1 = 3$  and Level of significance  $\alpha = 0.05$

$\chi^2$  tab value is 7.81

$\chi^2$  calculated = 23.667 >  $\chi^2$  tab = 7.81 at  $\alpha = 0.05$  for 3 dof

- ❖ Decision : we reject the Null hypothesis  $H_0$  at  $\alpha = 5\%$  LOS.  
i.e, we accept  $H_1$ .

i.e, The observed results not commensurate with the general examination results

## Example:-

Following is the record of number of accidents took place during the various days of the week.

Monday	Tues day	wednes day	Thurs day	Fri day	Sat day	Sun day
184	148	145	153	150	154	116

Can we conclude that accident is independent of the day in a week?

$$\chi^2 = \sum \frac{(O - e)^2}{e}$$

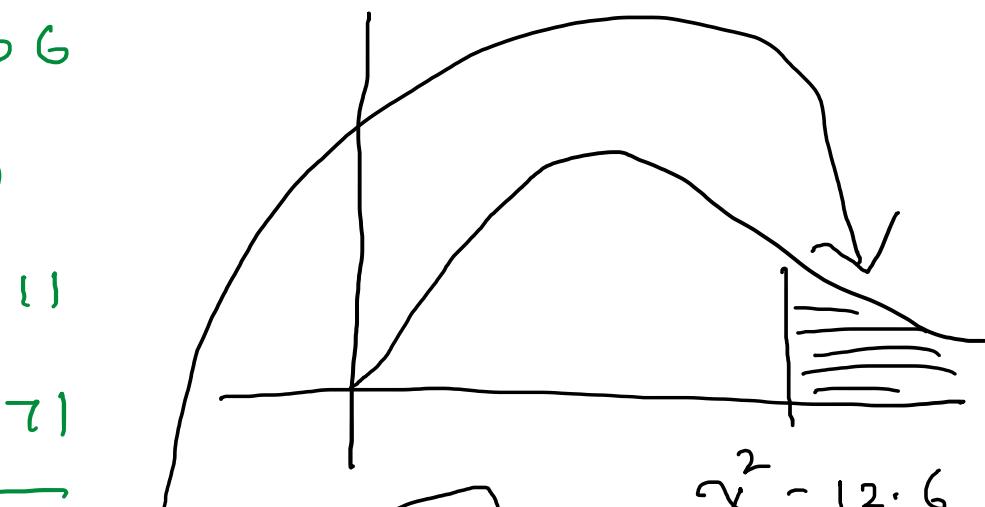
$$1050 \times \frac{1}{7} \\ = 150$$

$$O \quad e \quad \frac{(O - e)^2}{e}$$

			$H_0$ : accident is indept of same day dinner
184	150	7.71	
148	150	0.03	
145	150	0.17	$H_1$ : not
153	150	0.06	
150	150	0	
154	150	0.11	
116	150	7.71	
<hr/>			
Total	1050	$\chi^2 = 15.77$	$\chi^2 = 12.6$

Reject  $H_0$

at 5% loss &  
6 d.o.f





---

# EXAMPLE 24

Example:

	with Cancer	without cancer	total
smokers	400	300	700
Non-smokers	300	500	800
total	700	800	~1500

Can we conclude that

smoking causes cancer?

Discussion,  $H_0$ : smoking causes cancer  
 $H_1$ : no --.  
 $\alpha = 1\%$ .

	with Cancer	without cancer	total
smokers	400 326	300 373	700
Non-smokers	300 313	500 426	800
total	700	800	1500
$\frac{700 \times 700}{1500}$	$\frac{700 \times 800}{1500}$		
$\frac{700 \times 800}{1500}$	$\frac{800 \times 800}{1500}$		

$$\chi^2 = \frac{(O - E)^2}{E} = \frac{(400 - 326)^2}{326} + \frac{(300 - 373)^2}{373} +$$

innovate

achieve

lead

$$+ \frac{(300 - 373)^2}{373} + \frac{(500 - 426)^2}{426} =$$

	with Cancer	without cancer	Total
Smokers	400 326	300 373	700
Non-smokers	300 313	500 426	800
Total	700	800	1500

From tables,

$$\alpha = 5\%$$

$$d.o.f = (2-1)(2-1) = 1 \}$$





---

# EXAMPLE 25

# PROBLEM 1

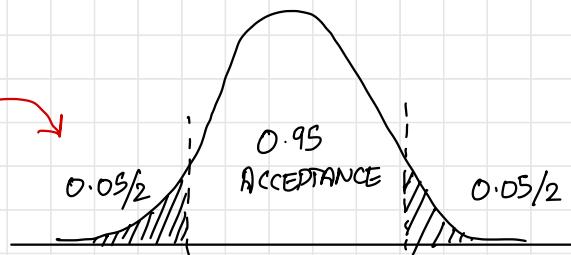
NULL HYPOTHESIS :  $H_0: \mu = 15150$

ALT HYPOTHESIS :  $H_1: \mu \neq 15150$

LEVEL OF SIG =  $\alpha = 5\%$ .

$$\begin{aligned}\mu &= 15150 \\ \sigma &= 1200\end{aligned}$$

2 TAIL TEST



$$E(Z_1) = 0.025 \Rightarrow Z_1 = -1.96$$

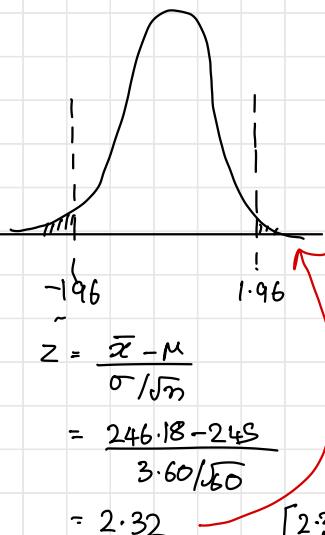
$$E(Z_2) = 0.975 \Rightarrow Z_2 = 1.96$$

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{15200 - 15150}{1200/\sqrt{49}} = 0.2916$$

## PROBLEM 2

$$H_0: \mu = 245$$

$$H_1: \mu \neq 245$$



## PROBLEM 3

$$H_0: \mu_1 - \mu_2 = 0$$

$$H_1: \mu_1 - \mu_2 \neq 0$$

↓  
2 TAILED  
Z-distr

$$Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

$$= \frac{250 - 220}{\sqrt{(40^2 + 55^2)/20}} = 8.82$$

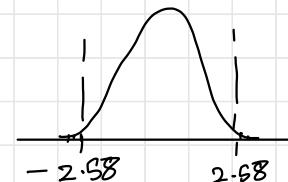
$Z > 2.58$

## PROBLEM 7

$$H_0: p = 40/100$$

$$H_1: p \neq 40/100$$

$$\alpha = 0.01$$



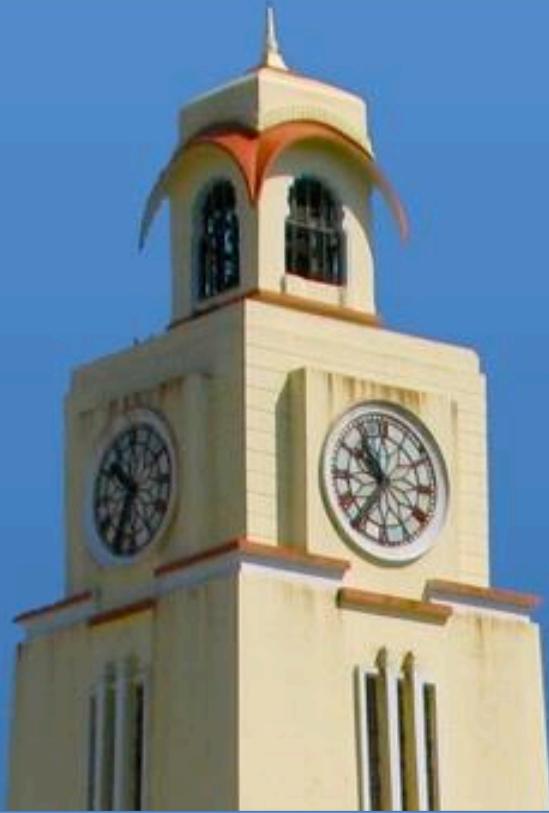
$$Z = \frac{\bar{p} - p}{\sqrt{pq/n}}$$

$$= \frac{82/160 - 40/100}{\sqrt{2400/10000 \times 150}}$$

$$= \frac{0.1466}{0.04} = 3.66$$







# **Session 7(19<sup>th</sup> March 2022)**

## **Testing of Hypothesis**

# The Hypotheses

## Example

Ministry of Human Resource Development (MHRD), Government of India takes an initiative to improve the country's human resources and hence set up **23 IIT's** in the country.

To measure the engineering aptitudes of graduates, MHRD conducts GATE examination for a mark of 1000 in every year. A sample of 300 students who gave GATE examination in 2018 were collected and the mean is observed as 220.

In this context, statistical hypothesis testing is to determine the mean mark of the all GATE-2018 examinee.

The two hypotheses in this context are:

$$H_0: \mu = 220$$

$$H_a: \mu < 220$$

# The Hypotheses

## Note:

- As null hypothesis, we could choose  $H_0: \mu \leq 220$  or  $H_0: \mu \geq 220$
- It is customary to always have the null hypothesis with an equal sign.
- As an alternative hypothesis there are many options available with us.

## Examples:

- I.  $H_a: \mu > 220$
- II.  $H_a: \mu < 220$
- III.  $H_a: \mu \neq 220$

The two hypothesis should be chosen in such a way that they are **exclusive** and **exhaustive**. One or other must be true, but they cannot both be true.

# The Hypotheses

## One-tailed test

- A statistical test in which the alternative hypothesis specifies that the population parameter lies entirely above or below the value specified in  $H_0$  is called a one-sided (or one-tailed) test.

Example.

$$H_0: \mu = 100$$

$$H_a: \mu > 100$$

## Two-tailed test

- An alternative hypothesis that specifies that the parameter can lie on either sides of the value specified by  $H_0$  is called a two-sided (or two-tailed) test.

Example.

$$H_0: \mu = 100$$

$$H_a: \mu <> 100$$

# The Hypotheses

**Note:**

In fact, a 1-tailed test such as:

$$H_0: \mu = 100$$

$$H_a: \mu > 100$$

is same as

$$H_0: \mu \leq 100$$

$$H_a: \mu > 100$$

In essence,  $\mu > 100$ , it does not imply that  $\mu > 80, \mu > 90$ , etc.

# Hypothesis Testing Procedures



- The following **five steps** are followed when testing hypothesis
- Specify  $H_0$  and  $H_1$ , the null and alternate hypothesis, and an **acceptable level of  $\alpha$** .
- Determine an appropriate sample-based test statistics and the **rejection region** for the specified  $H_0$ .
- Collect the sample data and calculate the test statistics.
- Make a decision to either reject or fail to reject  $H_0$ .
- Interpret the result in common language suitable for practitioners.

# Errors in Hypothesis Testing

In hypothesis testing, there are two types of errors.

**Type I error:** A type I error occurs when we incorrectly reject  $H_0$  (i.e., we reject the null hypothesis, when  $H_0$  is true).

**Type II error:** A type II error occurs when we incorrectly fail to reject  $H_0$  (i.e., we accept  $H_0$  when it is not true).

		Observation	
Decision		$H_0$ is true	$H_0$ is false
$H_0$ is accepted	Decision is correct	Type II error	
$H_0$ is rejected	Type I error	Decision is correct	

# Probabilities of Making Errors

## Type I error calculation

$\alpha$ : denotes the probability of making a Type I error

$$\alpha = P(\text{Rejecting } H_0 | H_0 \text{ is true})$$

## Type II error calculation

$\beta$ : denotes the probability of making a Type II error

$$\beta = P(\text{Accepting } H_0 | H_0 \text{ is false})$$

### Note:

- $\alpha$  and  $\beta$  are not independent of each other as one increases, the other decreases
- When the sample size increases, both decrease since sampling error is reduced.
- In general, we focus on Type I error, but Type II error is also important, particularly when sample size is small.

## Errors

		$H_0$ is true	$H_0$ is false
Accept $H_0$	Correct Decision	Type II Error $\rightarrow \beta$	
	Type I Error	Correct Decision	
Reject $H_0$			

$\alpha$

## Testing of Hypothesis

$H_0:$

$\alpha = ?$

$H_1$



Decide

one tailed / two tailed

$H_1: \mu \neq \dots \rightarrow \text{two}$

$H_1: \mu \geq \dots \quad \mu \leq \dots \quad \} \text{ one}$

→ Choose the test

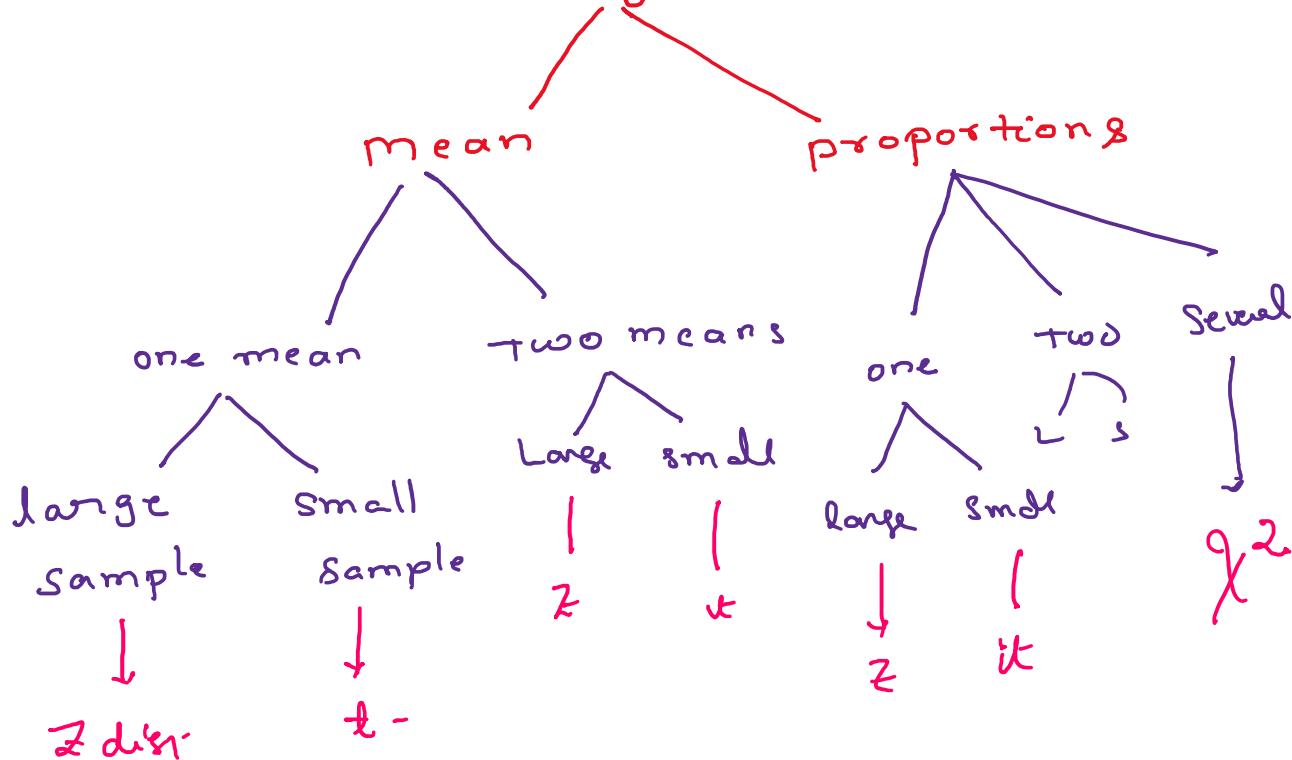
→ calculate  $Z / t$

→ taking decision

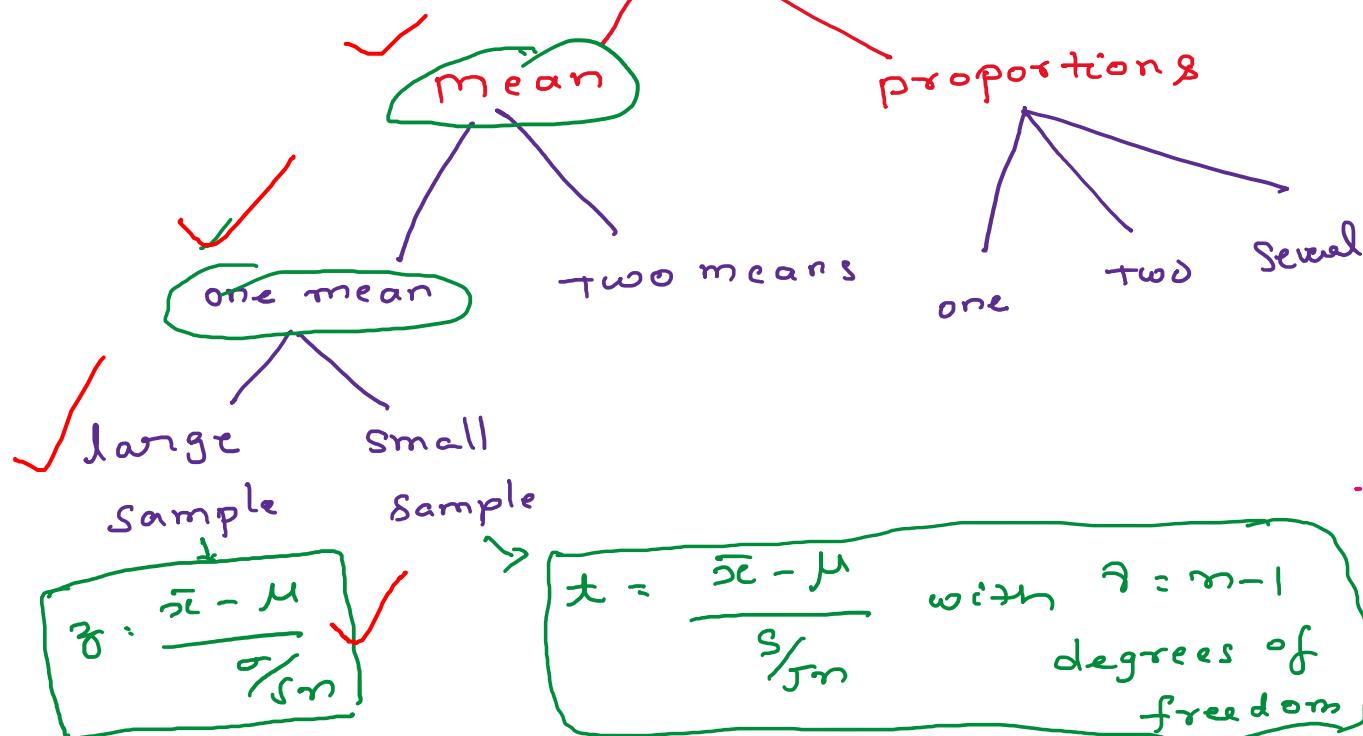
Accept or  
reject  $H_0$



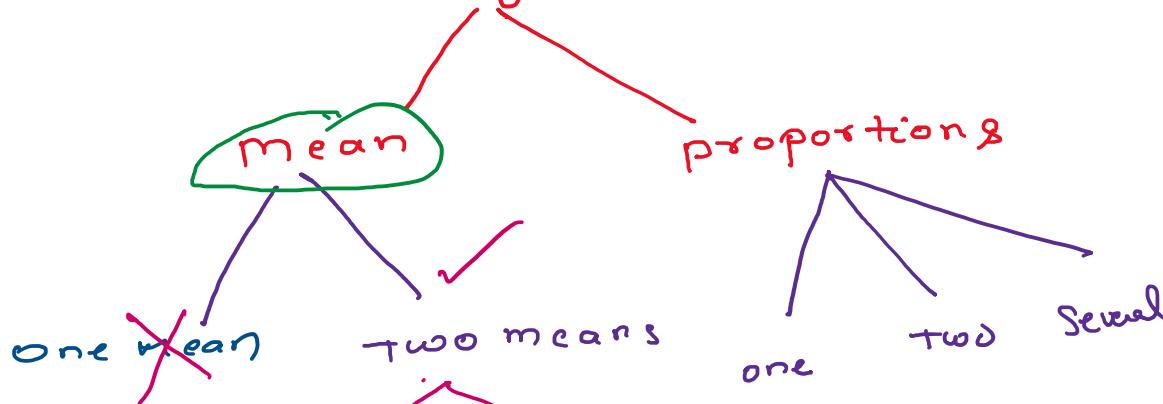
## Testing of Hypothesis



## Testing of Hypothesis



## Testing of Hypothesis



Large sample

$$z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Small sample

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

where  $S^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}$

## Testing of Hypothesis

mean

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

proportions

single

two

several

$$z = \frac{p - P}{\sqrt{\frac{PQ}{n}}}$$

$$z = \frac{(p_1 - p_2) - \delta}{s_p}$$

$$s_p = \sqrt{\hat{p}(1-\hat{p}) \left( \frac{1}{m_1} + \frac{1}{m_2} \right)}$$

$$\hat{p} = \frac{n_1 p_1 + n_2 p_2}{n_1 + n_2}$$

## Testing of Hypothesis

$H_0:$

$\alpha = ?$

$H_1: \mu \neq \dots \rightarrow \text{two}$

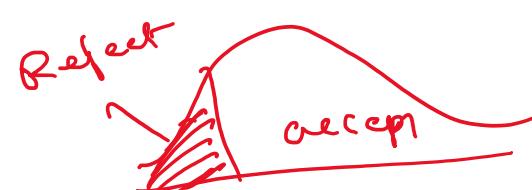
$H_1: \mu \geq \dots \quad \mu \leq \dots \quad \} \text{one}$

→ Choose the test

→ calculate  $Z / t$

→ taking decision  
Accept or  
reject  $H_0$

Decide one tailed / two tailed



# Problem-1

---

It is claimed that a random sample 49 tyres has a mean life of 15200 kms. This sample was drawn from a population whose mean is 15150 kms and a standard deviation of 1200kms. Test the significance at 0.05 level.



# PROBLEM 1

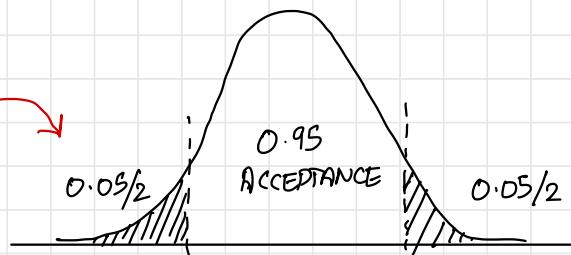
NULL HYPOTHESIS :  $H_0: \mu = 15150$

ALT HYPOTHESIS :  $H_1: \mu \neq 15150$

LEVEL OF SIG =  $\alpha = 5\%$ .

$$\begin{aligned}\mu &= 15150 \\ \sigma &= 1200\end{aligned}$$

2 TAIL TEST



$$E(Z_1) = 0.025 \Rightarrow Z_1 = -1.96$$

$$E(Z_2) = 0.975 \Rightarrow Z_2 = 1.96$$

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{15200 - 15150}{1200/\sqrt{49}} = 0.2916$$

# Solution 1

---

- Let  $\mu$  = true mean life of Tyres
  - Given that  $\mu = 15150$ ,  $n = 49$ ,  $\bar{x} = 15200$  and  $\sigma = 1200$ .
  - Null hypothesis  $H_0 : \mu = 15150$
  - i.e true mean life of tyres is 15150kms
  - Alternate hypothesis  $H_1 : \mu \neq 15150$  (Two tailed test)
  - i.e, true mean life of tyres is not 15200kms
  - Level of significance:  $\alpha = 0.05$
  
  - Here the tailed test is Two tailed test,  $Z_{tab}$
  - Tabulated value of Z at 5% LOS is 1.96 ( ).
-

# Solution 1



- **Test Statistic:**  $Z_{\text{cal}} = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{15200 - 15150}{1200 / \sqrt{49}} = 0.2916$

Here  $|Z_{\text{cal}}| = |0.2916| < Z_{\text{tab}} = 1.96$  at  $\alpha = 5\%$  LOS.

**Decision :** we accept the Null hypothesis  $H_0$  at  $\alpha = 5\%$  LOS.

i.e, true mean life of tyres is 15150kms.

# Problem 2

---

The target thickness for silicon wafers used in a certain type of integrated circuit is 245mm. A sample of 50 wafers is obtained and the thickness of each one is determined, resulting in a sample mean thickness of 246.18mm and a sample standard deviation of 3.60mm.

Does this data suggest that true average wafer thickness is something other than the target value?

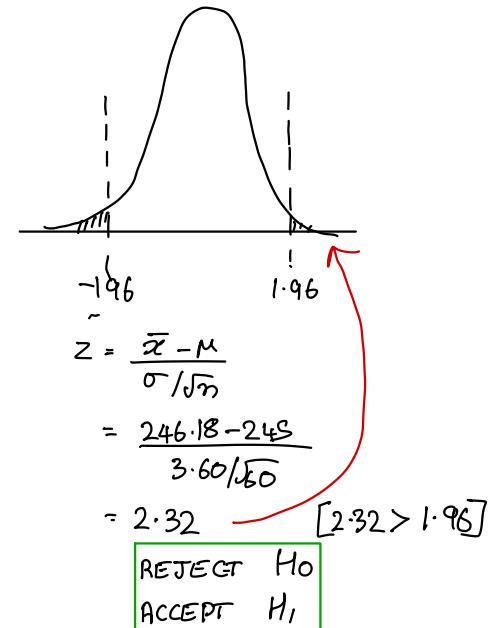
---

# Solution 2

- Let  $\mu$  = true average wafer thickness
- Given that  $\mu = 245$ ,  $n = 50$ ,  $\bar{x} = 246.18$  and  $s = 3.60$ .
- Null hypothesis  $H_0 : \mu = 245$
- i.e true average wafer thickness is same the target value
- 
- Alternate hypothesis  $H_1 : \mu \neq 245$  (Two tailed test)
- true average wafer thickness is something other than the target value
- Level of significance:  $\alpha = 0.05$
- Here the tailed test is Two tailed test,
- Tabulated value of Z at 5% LOS is 1.96 ( $Z_{tab}$ ).

## PROBLEM 2

$$\begin{array}{ll} H_0 & \mu = 245 \\ H_1 & \mu \neq 245 \end{array}$$



## Solution 2

- **Test Statistic:**  $Z_{\text{cal}} = \frac{\bar{x} - \mu}{s / \sqrt{n}} = \frac{246.18 - 245}{3.60 / \sqrt{50}} = 2.32$

Here  $|Z_{\text{cal}}| = |2.32| > Z_{\text{tab}} = 1.96$  at  $\alpha = 5\%$  LOS.

**Decision :** we reject the Null hypothesis  $H_0$  at  $\alpha = 5\%$  LOS.

i.e, we accept  $H_1$ .

i.e, true average wafer thickness is something other than the target value.

# Problem 3

In a survey of buying habits. 400 women shopper's are chosen at random in super Market A located in a certain city. Their average weekly food expenditure is Rs.250 with a SD of Rs.40. For 400 women shopper's are chosen at random in super Market B in another city. The average weekly food expenditure is Rs.220 with a SD of Rs.55.

- Test at 1% LOS whether the average weekly food expenditure of the two populations of shoppers are equal.

# Solution 3

- Given that  $n_1 = 400, n_2 = 400, \bar{x}_1 = 250, \bar{x}_2 = 220, s_1 = 40$  and  $s_2 = 55$ .
- Null hypothesis  $H_0 : \mu_1 = \mu_2$
- i.e, the average weekly food expenditure of the two populations of shoppers are equal.
- Alternate hypothesis  $H_1: \mu_1 \neq \mu_2$  (Two tailed test)
- i.e, the average weekly food expenditure of the two populations of shoppers are not equal.
- Level of significance:  $\alpha = 0.01$
- Here the tailed test is Two tailed test,
- Tabulated value of Z at 1% LOS is  $2.58 (Z_{tab})$ .

## Solution 3

- Test Statistic:  $Z_{\text{cal}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{250 - 220}{\sqrt{\frac{40^2}{400} + \frac{55^2}{400}}} = 8.82$

Here  $|Z_{\text{cal}}| = |8.82| > Z_{\text{tab}} = 2.58$  at  $\alpha = 1\%$  LOS.

Decision : we reject the Null hypothesis  $H_0$  at  $\alpha = 1\%$  LOS.  
i.e, we accept  $H_1$ .

PROBLEM 3

$$\begin{aligned} H_0 &\Rightarrow \mu_1 - \mu_2 = 0 \\ H_1 &\Rightarrow \mu_1 - \mu_2 \neq 0 \end{aligned}$$

2 TAILED  
Z-dist

$$\begin{aligned} Z &= \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \\ &= \frac{250 - 220}{\sqrt{(40^2 + 55^2)/20}} \approx 8.82 \end{aligned}$$

i.e, the average weekly food expenditure of the two populations of shoppers in markets A and B differ significantly.

# Problem-4

---

Two types of new cars produced in INDIA are tested for petrol mileage. One sample consisting of 42 cars averaged 15 kmpl while the other sample consisting of 80 cars averaged 11.5 kmpl with population variance as  $\sigma_1^2 = 2$  and  $\sigma_2^2 = 1.5$  respectively.

Test whether there is any significance difference in the petrol consumption of these two types of cars.

---

## Solution 4

---

- Given that  $n_1 = 42, n_2 = 80, \bar{x}_1 = 15, \bar{x}_2 = 11.5, \sigma_1^2 = 2$  and  $\sigma_2^2 = 1.5$
  - Null hypothesis  $H_0 : \mu_1 = \mu_2$
  - i.e, there is no significance difference in the petrol consumption of these two types of cars.
  - Alternate hypothesis  $H_1 : \mu_1 \neq \mu_2$  (Two tailed test)
  - i.e, there is significance difference in the petrol consumption of these two types of cars.
  - Level of significance:  $\alpha = 0.05$
  - Here the tailed test is Two tailed test,
  - Tabulated value of Z at 5% LOS is 1.96 ( ).  $Z_{tab}$
-

## Solution 4

- **Test Statistic:**  $Z_{cal} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{15 - 11.5}{\sqrt{\frac{2}{42} + \frac{1.5}{80}}} = 13.587$

Here  $|Z_{cal}| = |13.587| > Z_{tab} = 1.96$  at  $\alpha = 5\% \text{ LOS}$ .

**Decision :** we reject the Null hypothesis  $H_0$  at  $\alpha = 5\% \text{ LOS}$ .

i.e, we accept  $H_1$  .

i.e, there is significance difference in the petrol consumption of these two types of cars.

# Problem 5

Studying the flow of traffic at two busy intersections between 4pm and 6pm to determine the possible need for turn signals. It was found that on 40 week days there were on the average 247.3 cars approaching the first intersection from the south which made left turn, while on 30 week days there were on the average 254.1 cars approaching the second intersection from the south made left turns. The corresponding sample SD are 15.2 and 12.

- ❖ Test the significance between the difference of two means at 5% level.

# Problem 6

---

Is there any systematic tendency for part-time college faculty to hold their students to different standards than do full-time faculty? The article “Are There Instructional Differences Between Full-Time and Part-Time Faculty?” (College Teaching, 2009: 23–26) reported that for a sample of 125 courses taught by full-time faculty, the mean course GPA was 2.7186 and the standard deviation was .63342, whereas for a sample of 88 courses taught by part-timers, the mean and standard deviation were 2.8639 and .49241, respectively.

Does it appear that true average course GPA for part-time faculty differs from that for faculty teaching full-time? Test the appropriate hypotheses at significance level 1%.

---

# Problem 7

---

A random sample of 150 recent donations at a certain blood bank reveals that 82 were type A blood. Does this suggest that the actual percentage of type A donations differs from 40%, the percentage of the population having type A blood?

Carry out a test of the appropriate hypotheses using a significance level of .01.

---

# PROPORTION HYPOTHESIS TESTING

- 

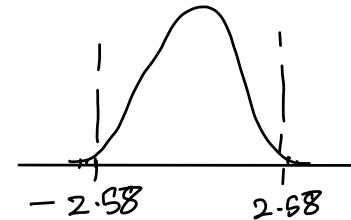
Critical value	Level of significance $\alpha$		
	1%	5%	10%
Two tailed test	$z_{\alpha/2} = 2.58$	$z_{\alpha/2} = 1.96$	$z_{\alpha/2} = 1.645$
One tailed test	$z_{\alpha} = 2.33$	$z_{\alpha} = 1.645$	$z_{\alpha} = 1.28$

PROBLEM 7

$$H_0: P = 40/100$$

$$H_1: P \neq 40/100$$

$$\alpha = 0.01$$



$$\begin{aligned}
 Z &= \frac{\bar{P} - P}{\sqrt{P(1-P)/n}} \\
 &= \frac{\frac{82}{150} - \frac{40}{100}}{\sqrt{\frac{2400}{10000 \times 150}}} \\
 &= \frac{0.1466}{0.04} = 3.66
 \end{aligned}$$

---

# EXAMPLE - 1

- 
- ❖ It is claimed that sports-car owners drive on the average 17000 kms per year. A consumer firm believes that the average milage is probably higher. To check, the consumer firm obtained information from randomly selected 40 sports-car owners that resulted in a sample mean of 17352 kms with a population standard deviation of 1348 kms. At what can be concluded about this claim at
  - ❖ 5% level of significance (Critical value is 1.645)
  - ❖ 1% level of significance (Critical value is 2.331)
-

# Example

$H_0$



The average milage of sports-car as claimed and the sample average milage may be same

$$H_0 : \mu = \mu_0 = 17000$$

$H_1$



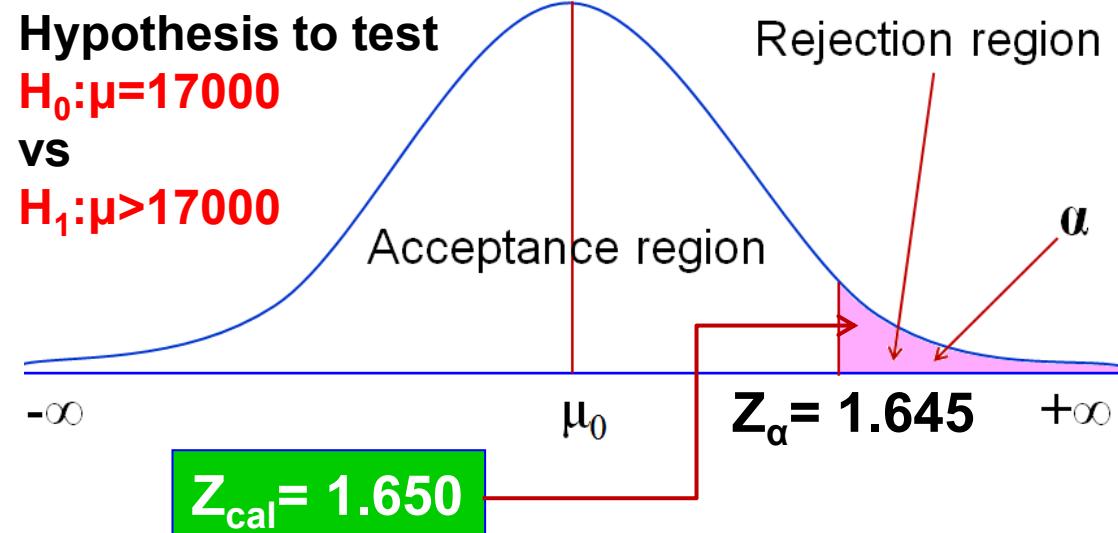
The average milage of sports-car as claimed may be **higher than** the sample average milage

$$H_1 : \mu > \mu_0 = 17000$$

**(a) At 5% level of significance with critical value 1.645**

$$Z = \frac{17352 - 17000}{\sqrt{\frac{1348}{40}}} = 1.650$$

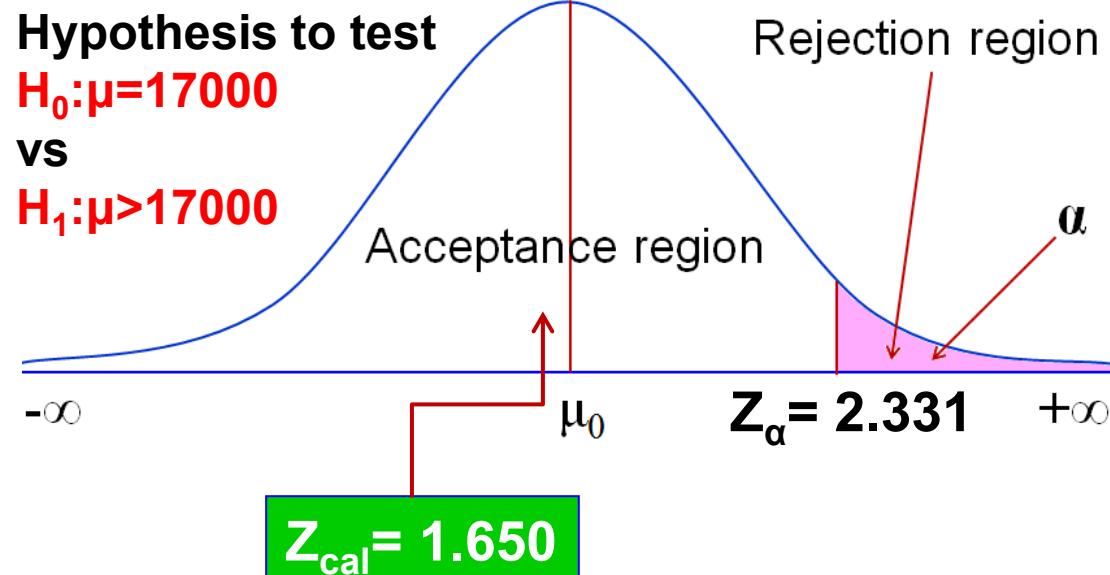
Critical value for  
 $\alpha = 0.05$  is 1.645  
 Since  $Z = 1.650 > 1.645$ , Reject  $H_0$  and Accept  $H_1$



**(b) At 1% level of significance with critical value 2.331**

$$Z = \frac{17352 - 17000}{\frac{1348}{\sqrt{40}}} = 1.650$$

Critical value for  
 $\alpha = 0.01$  is 2.331  
 Since  $Z = 1.624 < 2.330$ , Accept  $H_0$   
 Reject  $H_1$



---

# EXAMPLE - 2

## Example - 2

---

The manager of a courier service believes that packets delivered at the beginning of the month are heavier than those delivered at the end of month. As an experiment, he weighed a random sample of 20 packets at the beginning of the month and found that the mean weight was 5.25 kg. A randomly selected 10 packets at the end of the month had a mean weight of 4.96 kg. It was observed from the past experience that the population variances are 1.20 kg and 1.15 kg. At 5% level of significance, can it be concluded that the packets delivered at the beginning of the month weigh more? Also find P-value and 95% confidence interval for the difference between the means.

---

$H_0$



The mean weight of packets delivered at the early in the month and at the end of month may be same

$$H_0 : \mu_1 = \mu_2$$

$H_1$



The mean weight of packets delivered at the end of the month may be higher than at the early of month

$$H_1 : \mu_1 > \mu_2$$

At 5% (0.05) level of significance with critical value 1.645

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} = \frac{5.25 - 4.96}{\sqrt{\frac{(1.20)^2}{20} + \frac{(1.15)^2}{10}}} = 0.642$$

$$P(Z < 0.642) = 0.7389$$

Hypothesis to test

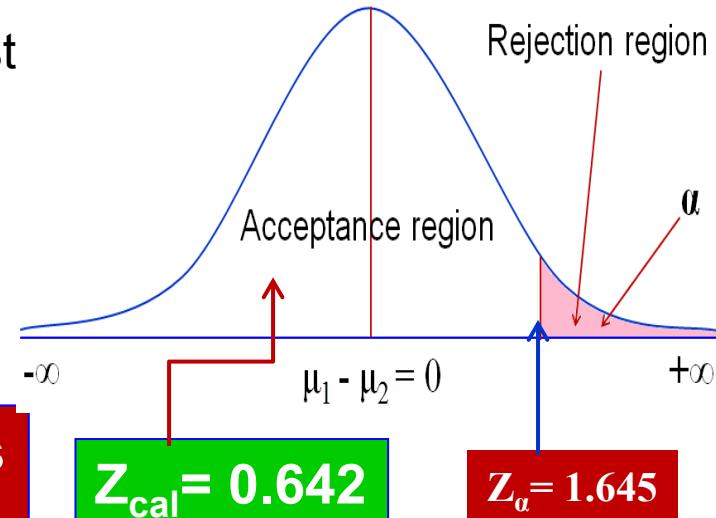
$$H_0: \mu_1 - \mu_2 = 0$$

vs

$$H_1: \mu_1 - \mu_2 > 0$$

???

95% CI for  $\mu$  is  
[- 454, 1.033]



$$Z_{\text{cal}} = 0.642$$

$$Z_{\alpha} = 1.645$$

Critical value for  $\alpha = 0.05$  is 1.645. Since  $Z = 0.642 < 1.645$ , Accept  $H_0$  Reject  $H_1$

## Z-test for proportion of a single population

Z-test

Proportion of a single population (P)

### Assumptions

Assume that the samples are drawn from normal distribution

The sample size should be more than or equal to 30

Subjects should be selected randomly

# Z-test for proportion of a single population

Z-test

Proportion of a single population (P)

## Assumptions

Assume that the samples are drawn from normal distribution

The sample size should be more than or equal to 30

**1 State null and alternative hypothesis**

$H_0 : P = P_0$  vs  $H_1 : P < P_0$   
or  $H_1 : P > P_0$   
or  $H_1 : P \neq P_0$

**2 Specify the level of significance 'α'**

**3 Standard Normal Distribution**

**4 Compute the test statistic**

$$Z = \frac{p - P_0}{\sqrt{\frac{pq}{n}}} \cong N(0, 1)$$

**5 Define the critical region/ rejection criteria**

**6 Conclusion**

**Note: Rejection criteria same as in one sample test or P-value**

---

# EXAMPLE - 3

# Example

---

A builder claims that heat pumps are installed in 70% of all homes being constructed today in the city of Bangalore. Would you agree with this claim if a random sample of new homes in this city shows that 28 out of 55 had heat pumps installed? What P-value and confidence interval are related in this situation?

---

# Z-test for difference between proportions

**1 State null and alternative hypothesis**

$$H_0 : P_1 = P_2 \text{ vs } H_1 : P_1 < P_2 \\ \text{or } H_1 : P_1 > P_2 \\ \text{or } H_1 : P_1 \neq P_2$$

**2 Specify the level of significance ‘α’**

**3 Standard Normal Distribution**

**4 Compute the test statistic**

$$Z = \frac{p_1 - p_2}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \cong N(0, 1)$$

**5 Define the critical region/ rejection criteria**

**6 Conclusion**

**Note: Rejection criteria same as in one sample test or P-value**

## Testing of Hypothesis → Z-test for proportions of two popln.

Based on sample size standard error for proportion may be calculated:

If the sample sizes are equal then  $SE(p_1 - p_2)$  is calculated by

$$SE(p_1 - p_2) = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

Based on sample size standard error for proportion may be calculated:

If the sample sizes are equal then  $SE(p_1 - p_2)$  is calculated by

$$SE(p_1 - p_2) = \sqrt{pq \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}, \quad p = \frac{x_1 + x_2}{n_1 + n_2} \text{ and } q = 1 - p$$

## Testing of Hypothesis → Z-test for proportions of two popln.

A cigarette manufacturing company claims that its brand A cigarettes outsells its brand B cigarettes by 8%. If it is found that 42 out of a random sample of 200 smokers prefer brand A and 18 out of 100 smokers prefer brand B, test at 5% level of significance, whether 8% difference a valid claim. Also construct 95% CI for  $(P_1 - P_2)$  and find P-value.

## Testing of Hypothesis → Z-test for proportions of two popln.

$H_0$

There may be 8% difference in the sale of two brands of cigarettes may be a valid claim

$$H_0 : P_1 - P_2 = 0.08$$

$H_1$

There may be 8% difference in the sale of two brands of cigarettes may not be a valid claim

$$H_1 : P_1 - P_2 \neq 0.08$$

Estimation → Interval estimate

## Finding Confidence Interval for difference in population proportion ( $P_1 - P_2$ )

The 100 (1- $\alpha$ )% confidence interval for difference between two population proportions ( $P_1 - P_2$ )

$$(p_1 - p_2) \pm Z_{\alpha/2} \text{ SE}(p_1 - p_2)$$

**Estimation** → **Interval estimate**

95% confidence interval for difference between  
two population proportions ( $P_1 - P_2$ )

$$(p_1 - p_2) \pm Z_{\alpha/2} \text{ SE } (p_1 - p_2) = (0.21 - 0.18) \pm 1.96 * 0.049 \\ = (-0.066, 0.126)$$

## Testing of Hypothesis → Z-test for proportions of two popln.

At 5% level of significance with critical value  $\pm 1.96$

$$Z = \frac{(0.21 - 0.18) - 0.08}{\sqrt{0.2 * 0.8 \left( \frac{1}{200} + \frac{1}{100} \right)}} = -1.02$$

Critical value for  
 $\alpha = 0.05$  is **-1.96**  
 Since  $Z = -1.02 > -1.96$ , Don't reject  $H_0$  and reject  $H_1$

Hypothesis to test

$$H_0: P_1 - P_2 = 0.08$$

vs

$$H_1: P_1 - P_2 \neq 0.08$$

$$P(Z < -2.074) = 0.1539$$

$$Z_{\text{cal}} = -1.02$$

$-\infty$

$+\infty$

$\alpha/2$

Rejection region

Acceptance region

$\alpha/2$

$$Z_{0.025} = -1.96$$

$$95\% \text{ CI for } P \text{ is } (-0.066, 0.126)$$

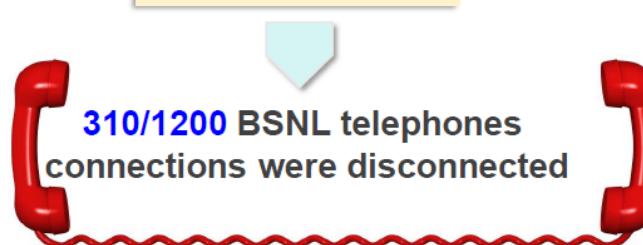
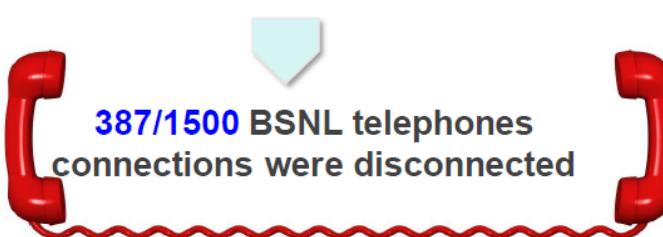
## Testing of Hypothesis → Z-test for proportions of two popln.

Problem on Z-test on difference in proportions

| Use  $\alpha = 1\%$ . |

Bangalore 2019

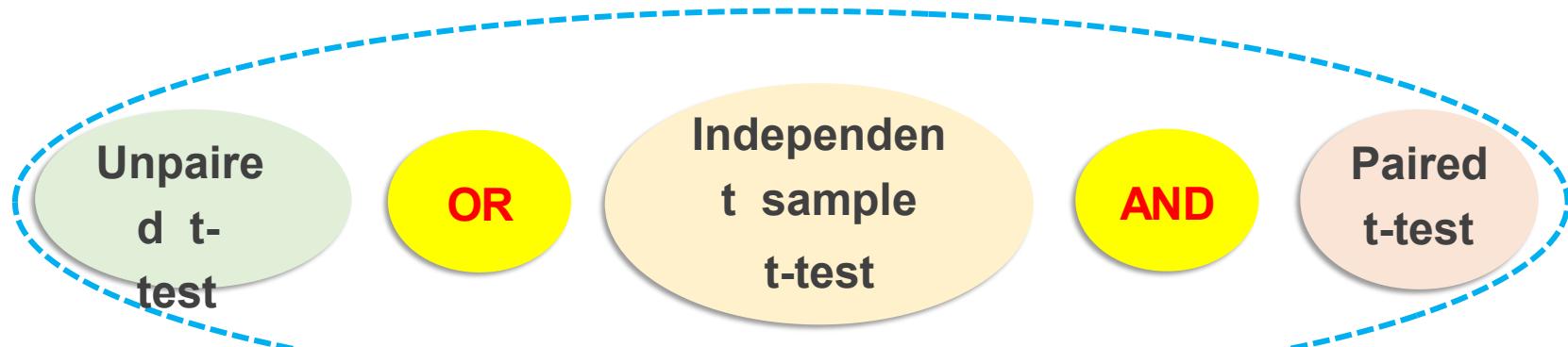
Delhi 2019



- Is there a significant difference between the two proportion of telephones disconnection by BSNL?

Construct 99% confidence interval for difference in proportions

## Testing of Hypothesis → Student's t-test



### Independent Sample t-test (Unpaired t-test)



Testing mean of a single population



Testing difference between means of two populations

**t-test**

## Testing mean of single population ( $\mu$ )

➤ Samples are drawn from normal population

➤ The population variance should be unknown

➤ The sample size should be less than 30 (i.e.,  $n < 30$ )

**t-test**

➤ Sample should be allocated randomly

➤ However even if sample size more than 30 (i.e.,  $n > 30$ ) and population variance unknown, t-test should be continue to apply, because of central limit theorem it approaches normal.

t-test



Degrees of freedom (df): No. of independent observations

## Suppose

$a+b = 20$ . If we assign  $a=9$  then  $b=11$  or vice-versa.  $\therefore df=(2-1)=1$

$a+b+c = 20$ . If we assign  $a=9$  and  $b=6$  then  $c=5$ .  $\therefore df=(3-1)=2$

In general, if there are  $n$  observations  $df = n-1$

# Mean of a single population using t-test

t-test



## Testing mean of a single population ( $\mu$ )

### Assumptions

Assume that the samples are drawn from normal distribution

The population variance may be unknown

The sample size should be less than 30 ( $n < 30$ )

Subjects should be selected randomly

**1 State null and alternative hypothesis**

$$H_0 : \mu = \mu_0 \text{ vs } H_1 : \mu < \mu_0 \\ \text{or } H_1 : \mu > \mu_0 \\ \text{or } H_1 : \mu \neq \mu_0$$

**2 Specify the level of significance 'α'**

**3 Student's t-distribution**

**4 Compute the test statistic**

$$t = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} \cong t_{(\alpha, n-1)}$$

**5 Define the critical region/ rejection criteria**

**6 Conclusion**

**Note:** Rejection criteria may be based on critical value or P-value

## 5 Define the critical region/ rejection criteria

- (i) Reject  $H_0$ , if computed value of  $t$  is less than the critical value, ie.,  $P(t < - t_\alpha)$ , otherwise do not reject  $H_0$
  - (ii) Reject  $H_0$ , if computed value of  $t$  is greater than the critical value, ie.,  $P(t > t_\alpha)$ , otherwise do not reject  $H_0$
-  By combining both (i) and (ii), Reject  $H_0$ , if computed value of  $|t|$  is greater than the critical value, ie.,  $P(|t| > t_\alpha)$ , otherwise do not reject  $H_0$ . Besides  $\alpha$ , the df is also important.

## Conclusion

## 5

### Define the critical region/ rejection criteria

(iii)

Reject  $H_0$ , if computed value of  $t$  is less than or greater than the critical value, ie.,  $P(t < - t_{\alpha/2})$  or  $P(t > t_{\alpha/2})$ , otherwise do not reject  $H_0$



Alternatively, reject  $H_0$ , if computed value of  $|t|$  is greater than the critical value, ie.,  $P(|t| > t_{\alpha/2})$ , otherwise do not reject  $H_0$ . Besides  $\alpha$ , the degrees of freedom is also important.

## Conclusion

---

# EXAMPLE - 4

# Example

---

It is claimed that sports-car owners drive on the average 18580 kms per year. A consumer firm believes that the average milage is probably higher. To check, the consumer firm obtained information from randomly selected 10 sports-car owners that resulted in a sample mean of 17352 kms with a sample standard deviation of 2012 kms. What can be concluded about this claim at

- 5% level of significance
  - 1% level of significance
-

$H_0$



The average milage of sports-car as claimed and the sample average milage may be same

$$H_0 : \mu = \mu_0 = 18580$$

$H_1$



The average milage of sports-car as claimed may be **higher than** the sample average milage

$$H_1 : \mu > \mu_0 = 18580$$

(a) At 5% level of significance with critical value 1.645

$$|t| = \frac{|17352 - 18580|}{\sqrt{\frac{2012}{10}}} = 1.929$$

95% CI for  $\mu$  is

$$\bar{x} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} = [16184.91, 18519.09]$$

P – value = 0.0428

Hypothesis to test

$$H_0: \mu = 18580 \text{ vs } H_1: \mu > 18580$$

Critical value for  $\alpha = 0.05$  is 1.833  
for 9 degree of freedom  
Since  $|t| = 1.929 > 1.833$ , Reject  $H_0$   
and Accept  $H_1$

# Testing the difference between means

**1 State null and alternative hypothesis**

$$H_0 : \mu_1 = \mu_2 \text{ vs } H_1 : \mu_1 < \mu_2 \\ \text{or } H_1 : \mu_1 > \mu_2 \\ \text{or } H_1 : \mu_1 \neq \mu_2$$

**2 Specify the level of significance 'α'**

**3 Standard Normal Distribution**

**4 Compute the test statistic**

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \cong t_{(\alpha, n_1+n_2-2)}$$

**5 Define the critical region/ rejection criteria**

Note: If sample sizes are unequal  
compute pooled SE

**6 Conclusion**

Note: Rejection criteria may be based on critical value or  
P-value

**t-test**

## Difference between means of two populations ( $\mu_1 - \mu_2$ )

- Samples are drawn from normal populations
- The population variances should be unknown
- The sample size should be less than 30 (i.e.,  $n < 30$ )
- The population variances should be equal
- Two groups should be independent
- Subjects should be allocated randomly to both groups
- However even if sample size more than 30 (i.e.,  $n > 30$ ) and population variances unknown, t-test should be continue to apply, because of central limit theorem it approaches normal.

---

# EXAMPLE - 5

# Example

---

Random samples of 15 and 10 were selected from two thermocouples. The sample means were 315, 303 and sample standard deviations were 3.8, 4.9 respectively.

- ❖ Construct 95% CI for difference in means
  - ❖ Test whether there is any significant difference in the means of two thermocouples at 5% level of significance
  - ❖ Find the P-value
-

$H_0$



The mean of two thermocouples may be same

$$H_0 : \mu_1 = \mu_2$$

$H_1$



The mean of two thermocouples may be different

$$H_1 : \mu_1 \neq \mu_2$$

At 5% (0.05) level of significance with critical value

$$|t| = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{315 - 303}{\sqrt{\frac{(3.8)^2}{15} + \frac{(4.9)^2}{10}}} = 3.571$$

Hypothesis to test

$$\begin{aligned} H_0: \mu_1 - \mu_2 &= 0 \\ \text{vs} \\ H_1: \mu_1 - \mu_2 &> 0 \end{aligned}$$

???

**95% CI for  $\mu$  is  
[6.24, 17.76]  
not includes 0**

**95% CI for  $\mu$  is  
[6.24, 17.76]**

Critical value for  $\alpha = 0.05$  is 1.714. Since  $|t| = 3.571 > 1.714$ , Reject  $H_0$  & Accept  $H_1$

## PROBLEM:

---

The manager of a courier service believes that packets delivered at the beginning of the month are heavier than those delivered at the end of month. As an experiment, he weighed a random sample of 15 packets at the beginning of the month and found that the mean weight was 5.25 kg. A randomly selected 10 packets at the end of the month had a mean weight of 4.56 kg. It was observed from the past experience that the sample variances are 1.20 kg and 1.15 kg.

- At 5% level of significance, can it be concluded that the packets delivered at the beginning of the month weigh more?
  - Also find P-value and 95% confidence interval for the difference between the means.
-

$H_0$



**The mean weight of packets delivered at the early in the month and at the end of month may be same**

$$H_0 : \mu_1 = \mu_2$$

$H_1$



**The mean weight of packets delivered at the early in the month may be higher than at the end of month**

$$H_1 : \mu_1 > \mu_2$$

## Estimation → Confidence interval for $(\mu_1 - \mu_2)$ based t-test

Finding Confidence Interval for difference between two population means  $(\mu_1 - \mu_2)$

The 100  $(1-\alpha)\%$  confidence interval for difference between two means

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2} \text{SE} (\bar{X}_1 - \bar{X}_2)$$

$$(\bar{X}_1 - \bar{X}_2) - t_{\alpha/2} \text{SE}(\bar{X}_1 - \bar{X}_2) \leq (\mu_1 - \mu_2) \leq (\bar{X}_1 - \bar{X}_2) + t_{\alpha/2} \text{SE}(\bar{X}_1 - \bar{X}_2)$$

## Estimation → Confidence interval for $(\mu_1 - \mu_2)$ based t-test

95% Confidence Interval for difference between two population means  $(\mu_1 - \mu_2)$

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \text{SE} (\bar{x}_1 - \bar{x}_2) = (5.25 - 4.26) \pm 2.069 * 0.443 \\ = (0.073, 1.907)$$

At 5% (0.05) level of significance with critical value 1.714

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{5.25 - 4.26}{0.443} = 2.233$$

$$0.025 \leq P \leq 0.01$$

$(\mu_1 - \mu_2) = 0$  not included in

Hypothesis to test

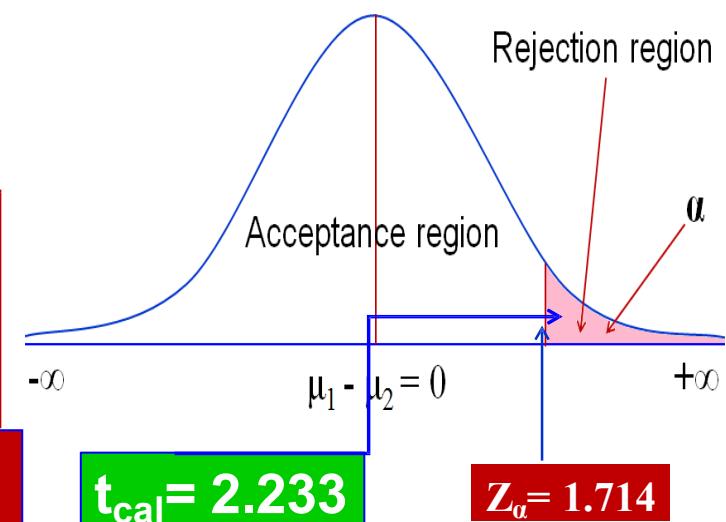
$$H_0: \mu_1 - \mu_2 = 0$$

vs

$$H_1: \mu_1 - \mu_2 > 0$$

???

95% CI for  $\mu_1 - \mu_2$   
is (0.073, 1.907)



Critical value for  $\alpha = 0.05$  is 1.714. Since  $t = 2.233 > 1.714$ , Reject  $H_0$ , Don't reject  $H_1$

# Student's paired t-test

t-test



Testing mean before and after observations  
of a single population ( $\mu_d$ )

## Assumptions

Assume that the difference between before and after observations follow normal distribution

The sample size should be less than 30 ( $n < 30$ )

**1 State null and alternative hypothesis**

$$H_0 : \mu_d = 0 \text{ vs } H_1 : \mu_d < 0 \\ \text{or } H_1 : \mu_d > 0 \\ \text{or } H_1 : \mu_d \neq 0$$

**2 Specify the level of significance 'α'**

**3 Student's t-distribution**

**4 Compute the test statistic**

$$t = \frac{\bar{d} - \mu_d}{\frac{s_d}{\sqrt{n}}} \cong t_{(\alpha, n-1)}$$

**5 Define the critical region/ rejection criteria**

But  $\mu_d$  under  $H_0$  will be 0

**6 Conclusion**

Note: Rejection criteria may be based on critical value or P-value

# Student's paired t-test

- The HRD manager wishes to see if there has been any change in the ability of trainees after a specific training programme.
- The trainees take a aptitude test Before and after training programme.

Subjects	Before (x)	After (y)
1	75	70
2	70	77
3	46	57
4	68	60
5	68	79
6	43	64
7	55	55
8	68	77
9	77	76

Subjects	Before (x)	After (y)	$d = y - x$	$(d - \text{mean})^2$
1	75	70	-5	100
2	70	77	7	4
3	46	57	11	36
4	68	60	-8	169
5	68	79	11	36
6	43	64	21	256
7	55	55	0	25
8	68	77	9	16
9	77	76	-1	36
<b>Total</b>			<b>45</b>	<b>678</b>

$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n} = \frac{45}{9} = 5$$

$$S_d = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1}}$$

$$S_d = \sqrt{\frac{678}{8}} = 9.21$$

At 5% (0.05) level of significance with critical value is 3.31

$$|t| = \frac{\bar{d} - \mu_d}{S_d / \sqrt{n}} = \frac{5 - 0}{9.21 / \sqrt{9}} = 3.07$$

Hypothesis to test

$H_0: \mu_d = 0$   
vs  
 $H_1: \mu_d > 0$

???

**95% CI for  $\mu$  is  
[-2.52, 12.52]  
not includes 0**

**95% CI for  $\mu$  is  
[-2.52, 12.52]**

Critical value for  $\alpha = 0.05$  is 1.895. Since  $|t| = 1.63 < 2.31$ , Accept  $H_0$  & Reject  $H_1$

# Exercise

Diet-modification Program



Ten individuals have participated

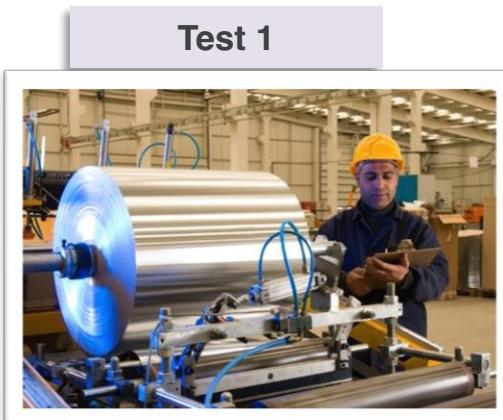


Subject	1	2	3	4	5	6	7	8	9	10
Weight Before	195	213	247	201	187	210	215	246	294	310
Weight After	187	195	221	190	175	197	199	221	278	285

Is there sufficient evidence to support claim that this program is effective in reducing weight?

Use  $\alpha = 0.05$ .

Construct 95% confidence interval for mean difference.



Is there sufficient evidence to conclude that both tests give the same mean impurity level

Specimen	1	2	3	4	5	6	7	8
Test 1	1.2	1.3	1.5	1.4	1.7	1.8	1.4	1.3
Test 2	1.4	1.7	1.5	1.3	2.0	2.1	1.7	1.6

| Using  $\alpha = 0.01$  |

Construct 99% confidence interval for mean difference

---

# EXAMPLE - 6

---

The target thickness for silicon wafers used in a certain type of integrated circuit is 245mm. A sample of 50 wafers is obtained and the thickness of each one is determined, resulting in a sample mean thickness of 246.18mm and a sample standard deviation of 3.60mm.

Does this data suggest that true average wafer thickness is something other than the target value?

---

# Discussion



- Let  $\mu$  = true average wafer thickness
- Given that  $\mu = 245$ ,  $n = 50$ ,  $\bar{x} = 246.18$ ,  $s = 3.60$ .
- Null hypothesis  $H_0 : \mu = 245$
- i.e true average wafer thickness is same the target value
- 
- Alternate hypothesis  $H_1 : \mu \neq 245$  (Two tailed test)
- true average wafer thickness is something other than the target value
- Level of significance:  $\alpha = 0.05$
- Here the tailed test is Two tailed test,
- Tabulated value of Z at 5% LOS is 1.96 ( ).  $Z_{tab}$

**Test Statistic:**  $Z_{\text{cal}} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}} = \frac{246.18 - 245}{\frac{3.60}{\sqrt{50}}} = 2.32$

Here  $|Z_{\text{cal}}| = |2.32| > Z_{\text{tab}} = 1.96$       at  $\alpha = 5\%$  LOS.

**Decision :** we reject the Null hypothesis  $H_0$       at  $\alpha = 5\%$  LOS.  
 i.e, we accept  $H_1$ .

i.e, true average wafer thickness is something other than the target value.

---

# EXAMPLE - 7

# Problem



In a survey of buying habits. 400 women shopper's are chosen at random in super Market A located in a certain city. Their average weekly food expenditure is Rs.250 with a SD of Rs.40. For 400 women shopper's are chosen at random in super Market B in another city. The average weekly food expenditure is Rs.220 with a SD of Rs.55.

Test at 1% LOS whether the average weekly food expenditure of the two populations of shoppers are equal.

- Given that  $n_1 = 400, n_2 = 400, \bar{x}_1 = 250, \bar{x}_2 = 220, s_1 = 40$  and  $s_2 = 55$ .
- Null hypothesis  $H_0 : \mu_1 = \mu_2$
- i.e, the average weekly food expenditure of the two populations of shoppers are equal.
- Alternate hypothesis  $H_1 : \mu_1 \neq \mu_2$  (Two tailed test)
- i.e, the average weekly food expenditure of the two populations of shoppers are not equal.
- Level of significance:  $\alpha = 0.01$
- Here the tailed test is Two tailed test,
- Tabulated value of Z at 1% LOS is 2.58 ( ).

$Z_{tab}$

**Test Statistic:**  $Z_{\text{cal}} = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{250 - 220}{\sqrt{\frac{40^2}{400} + \frac{55^2}{400}}} = 8.82$

Here  $|Z_{\text{cal}}| = |8.82| > Z_{\text{tab}} = 2.58$  at  $\alpha = 1\%$  LOS.

**Decision :** we reject the Null hypothesis  $H_0$  at  $\alpha = 1\%$  LOS.  
 i.e, we accept  $H_1$ .

i.e, the average weekly food expenditure of the two populations of shoppers in markets A and B differ significantly.



# **Session 8(9<sup>th</sup> April 2022)**

## **Correlation and regression**

# Covariance



- Variables may change in relation to each other
- *Covariance* measures how much the movement in one variable predicts the movement in a corresponding variable

# Covariance of $X$ and $Y$

---



$$\text{Cov}(X, Y) =$$

$$= \left[ E(X - \mu_X)(Y - \mu_Y) \right]$$

$$= \sum_x \sum_y (x - \mu_X)(y - \mu_Y) P(x, y)$$

if discrete

$$= \iint (x - \mu_X)(y - \mu_Y) f(x, y) dx dy$$

if continuous

$\text{Cov}(x, y)$

$$= \frac{\sum_{i=1}^n (x - \bar{x})(y - \bar{y})}{n - 1}$$

# Smoking v Lung Capacity Data

$n$	Cigarettes (x )	Lung Capacity (y )
1	0	45
2	5	42
3	10	33
4	15	31
5	20	29

$x$	$y$	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x}) \cdot (y - \bar{y})$
0	45	-10	9	-90
5	42	-5	6	-30
10	33	0	-3	0
15	31	5	-5	-25
20	29	10	-7	-70
$\bar{x} = \frac{\sum x}{n} = 10$		$\sum = -215$		

CO-Variance

$$= \frac{\sum (x - \bar{x})(y - \bar{y})}{n-1}$$

$$= \frac{-215}{5-1}$$

$$= -53.75$$

$$\bar{y} = \frac{\sum y}{n} = 36$$

And also

⇒ Farmer has an impression that  
if he uses more fertilizers, then the  
crop yield increases.

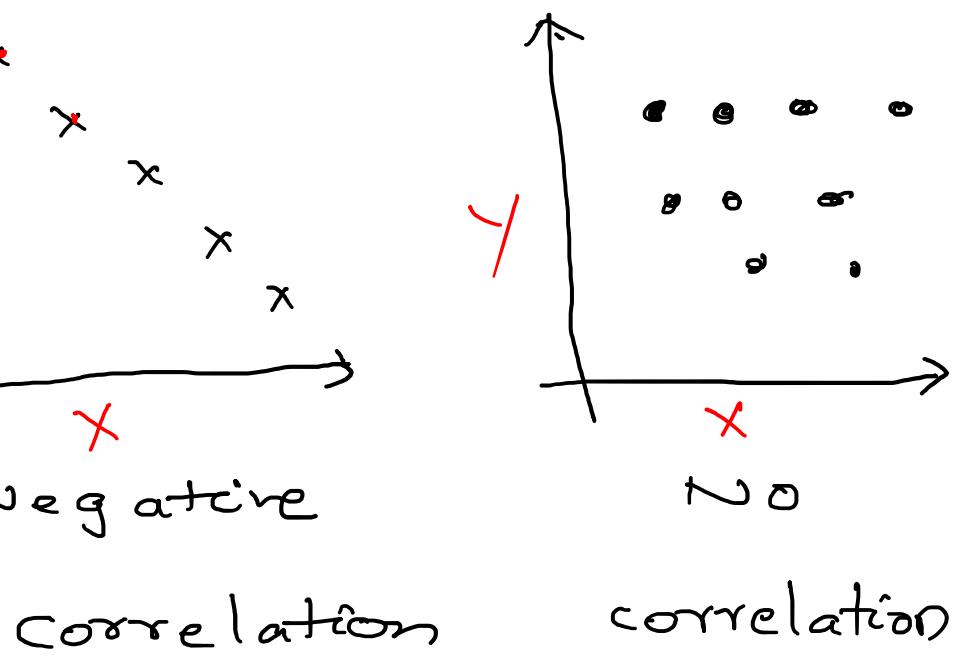
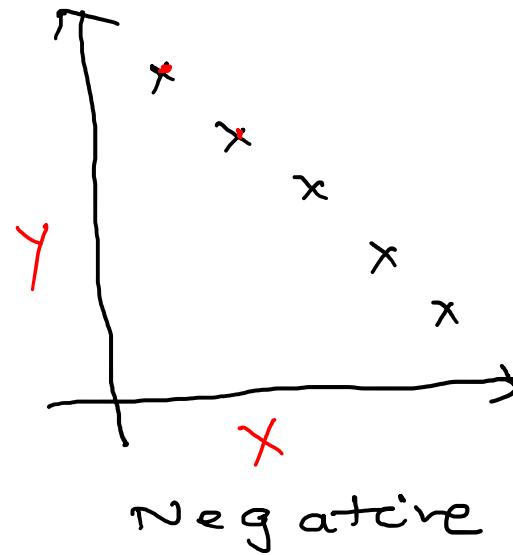
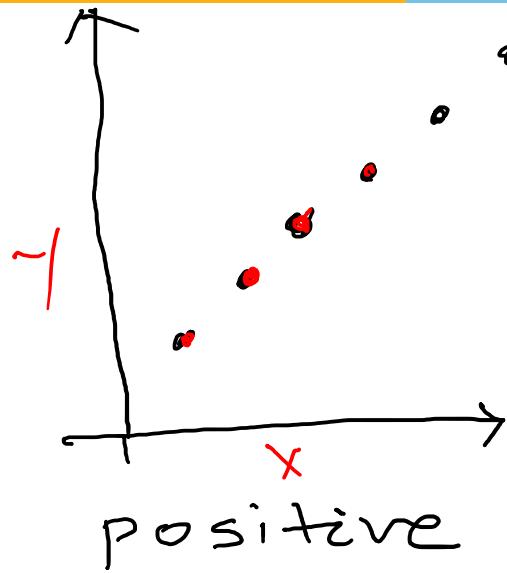
we need to validate this?

How → ?

# Correlation



- Finding the relationship between two quantitative variables without being able to infer causal relationships
  
- Correlation is a statistical technique used to determine the degree to which two variables are related



# Coefficient of correlation:



$$r = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y} = \frac{\sum x \cdot y}{\sqrt{\sum x^2 \cdot \sum y^2}}$$

where  $x = x - \bar{x}$

$$y = y - \bar{y}$$

$$x^2 = (x - \bar{x})^2$$

$$y^2 = (y - \bar{y})^2$$

# Coefficient of Correlation

---

$r = 1 \Rightarrow$  Perfect and positive relation

$r = -1 \Rightarrow$  " , " negative relation

$r = 0 \Rightarrow$  no relation

$0 < r < 1 \Rightarrow$  Partial positive relation

$-1 < r < 0 \Rightarrow$  " negative "

---

# Example - 1

	1	2	3	4	5	6	7	8	9
x	10	11	12	14	13	15	16	12	18
y									

$$\bar{x} = \frac{\sum x}{n} = \frac{45}{9} = 5$$

$$\bar{y} = \frac{\sum y}{n} = \frac{126}{9} = 14$$

$x$	$x =$	$x^2$	$y$	$y = y - 14$	$y^2$	$xy$	$\dots$
1	-4	16	10	-4	16	16	
2	-3	9	11	-3	9	9	
3	-2	4	12	-2	4	4	
4	-1	1	14	0	0	0	
5	0	0	13	-1	1	0	
6	1	1	15	1	1	1	
7	2	4	16	2	4	4	
8	3	9	17	3	9	9	
9	4	16	18	4	16	16	

$$r: \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$$

$$= \frac{59}{\sqrt{60 \times 60}}$$

$$= 0.9833$$

$x$	$x =$	$y$	$y =$	$y^2$	$xy$
1	$x - 5$	16	10	-4	16
2	-3	9	11	-3	9
3	-2	4	12	-2	4
4	-1	1	14	0	0
5	0	0	13	-1	0
6	1	1	15	1	1
7	2	1	16	2	4
8	3	9	17	3	9
9	4	16	18	4	16

$$\text{cov}(x, y)$$

$$= \frac{\sum xy}{n-1}$$

$$= \frac{59}{8}$$

$$= 7.375$$

$$\cdot \pi = 0.9833$$

$$\text{cov}(x, y) = 7.375$$

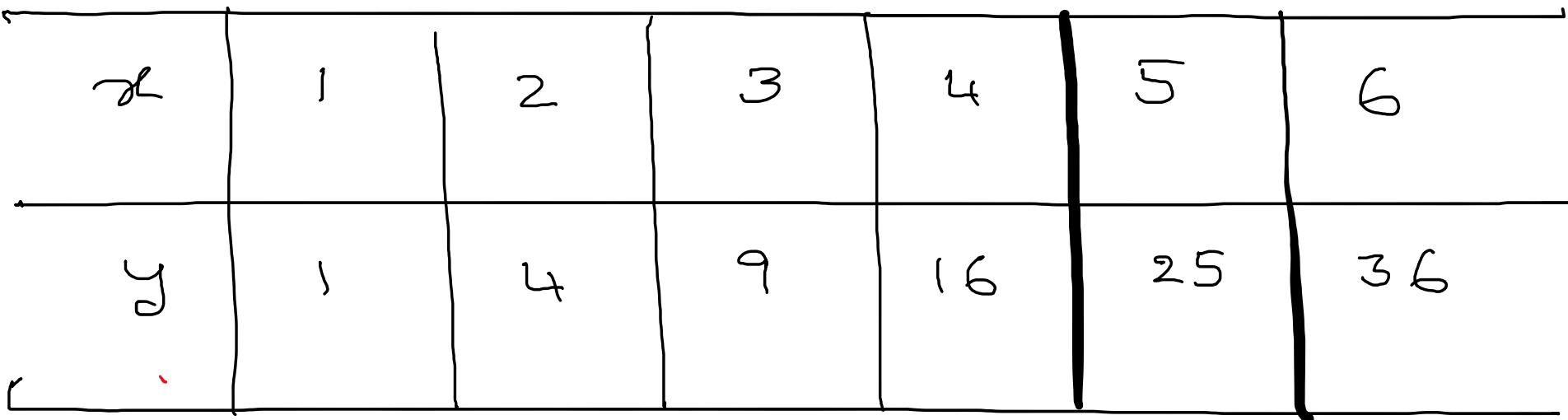
Interpretation

$$\pi^2 = 0.81$$



# Regression

# Regression

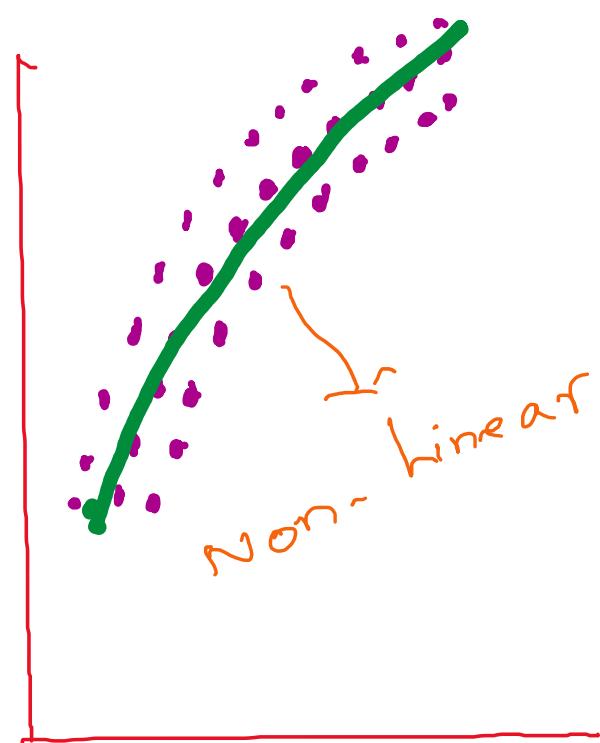
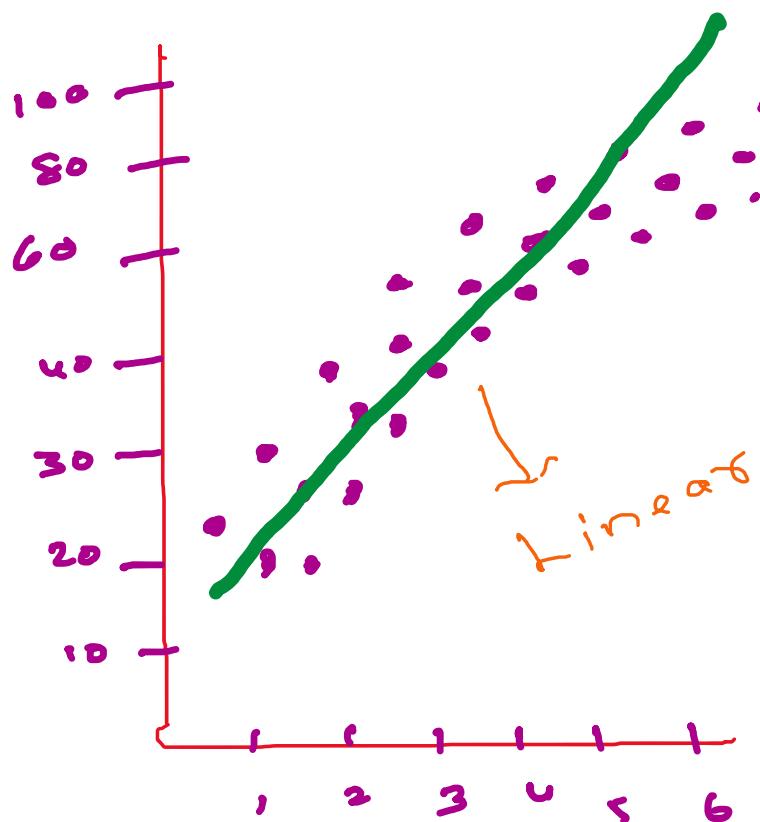


# Regression

x Income (Lakhs)	25	30	20	15	18	22
y Exp (Lakhs)	6	8	5	7	4	5

when  $x = 32$ , then

$$y = ?$$



CORRELATION	REGRESSION
Measuring strength or degree of the relationship between two variables	Having an algebraic equation between two variables
No estimation	Estimation
Both variables are independent	One is dependent variable and the other is independent variable

# Method of Least squares

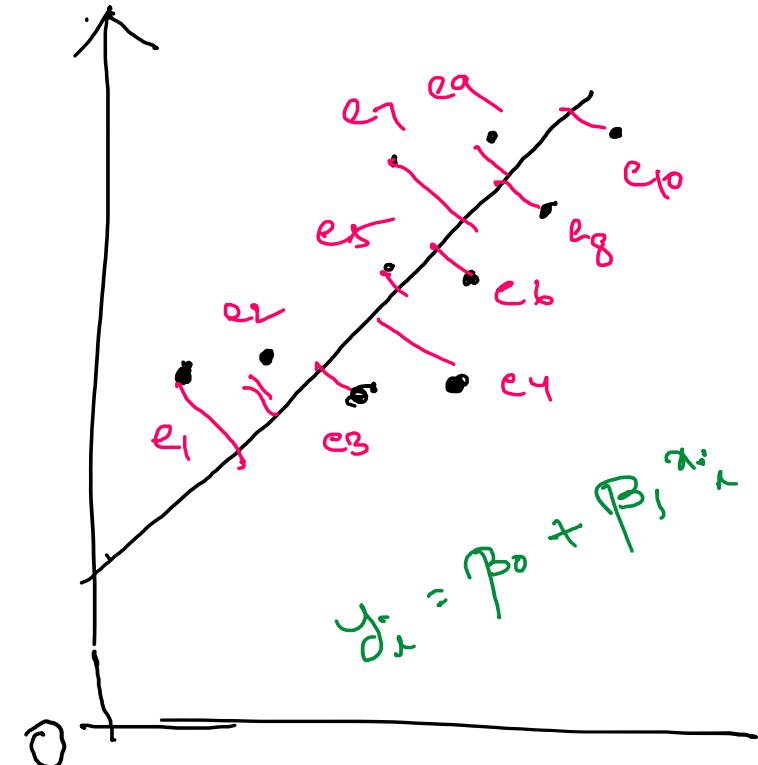


y : Dependent Variable

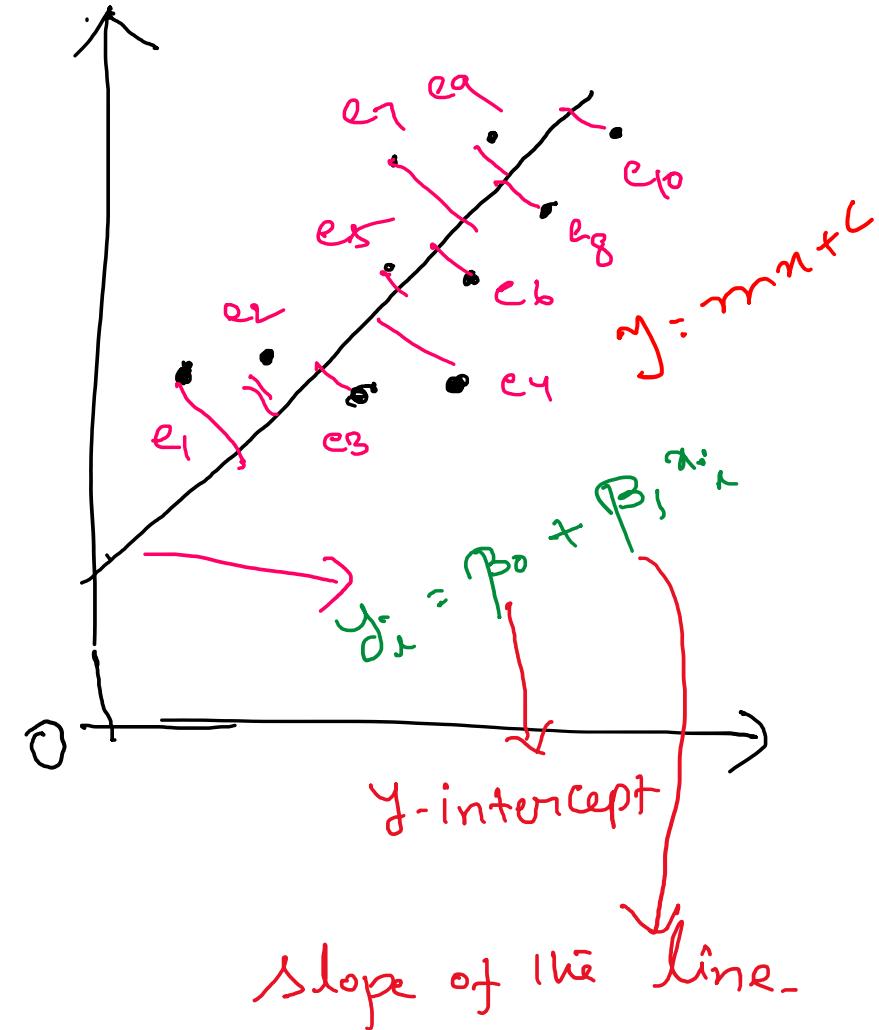
x : Independent Variable

predictor variable

response Variable



# Method of Least squares

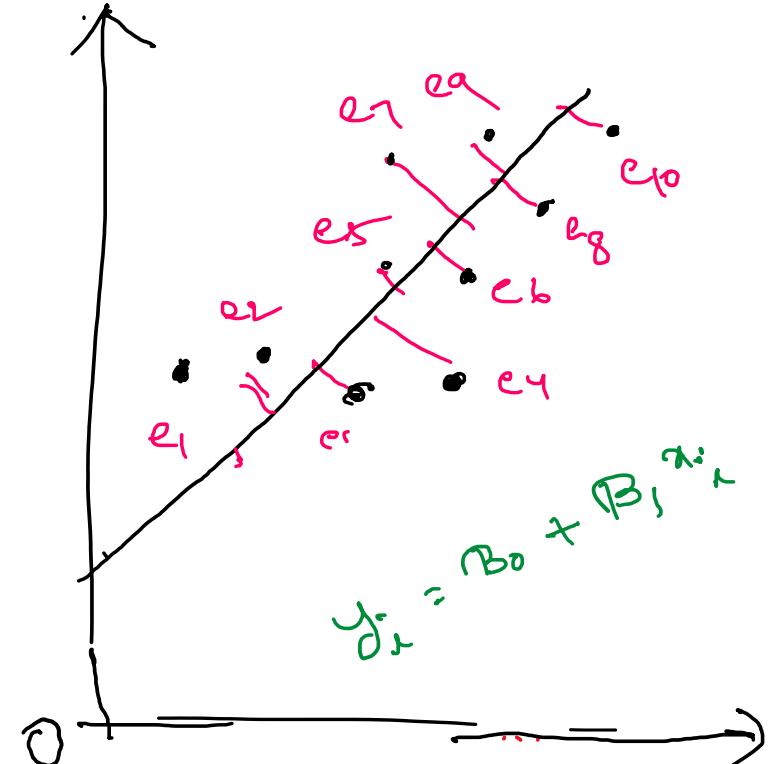


# Method of Least squares

$$S(\beta_0, \beta_1)$$

$$= \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

we need to choose  $\beta_0$  and  $\beta_1$  which minimizes the error.



# Method of Least squares

$$S(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$
$$\frac{\partial S}{\partial \beta_0} = 0 \Rightarrow 2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) (-1)$$
$$\Rightarrow \sum_{i=1}^n y_i = n \beta_0 + \beta_1 \sum_{i=1}^n x_i$$
$$\frac{\partial S}{\partial \beta_1} = 0 \Rightarrow 2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) (2)(-x_i)$$
$$\Rightarrow \sum_{i=1}^n x_i y_i = \beta_0 \sum x_i + \beta_1 \sum x_i^2$$

on solving these, we get  $\beta_0$  &  $\beta_1$   
which minimizes error.

# REGRESSION

LINEAR

$$y = w_0 + w_1 x$$

NON LINEAR

POLYNOMIAL

$$y = w_0 + w_1 x$$

$$y = w_0 + w_1 x + w_2 x^2$$

$$y = w_0 + w_1 x + w_2 x^2 + w_3 x^3$$

TOO VAST

COULD CAUSE  
OVERFITTING

## Linear regression -

$$y = \beta_0 + \beta_1 x$$

$$\begin{aligned}\sum y &= \beta_0 n + \beta_1 \sum x \\ \sum xy &= \beta_0 \sum x + \beta_1 \sum x^2\end{aligned}$$

Normal equations.

$$Y = b_0 + b_1 X$$

$$b_1 = \frac{n \sum xy - \bar{x} \bar{y}}{n \sum x^2 - (\bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

Example :-

company	Advt Expt	Sales Revenue
A	1	1
B	3	2
C	4	2
D	6	4
E	8	6
F	9	8
G	11	8
H	14	9

$$y = a + bx$$

$$\sum y = an + b \sum x$$

$$\sum xy = a \sum x + b \sum x^2$$

Example :-

$$y = \beta_0 + \beta_1 x$$

Sales Revenue y	Advt expt. x	$x^2$	xy
1	1	1	1
2	3	9	6
2	4	16	8
4	6	36	24
6	8	64	48
8	9	81	72
8	11	121	88
9	14	196	126
$\sum 40$		$\sum 56$	$\sum 373$

$$\sum y = n\beta_0 + \beta_1 \sum x$$

$$\sum xy = \beta_0 \sum x + \beta_1 \sum x^2$$

$$\Rightarrow 40 = 8\beta_0 + 56\beta_1$$

$$373 = 56\beta_0 + 524\beta_1$$

on solving

$$\beta_0 = 0.072$$

$$\beta_1 = 0.704$$

i.e.  $y = (0.072) + (0.704)x$

$$\therefore y = (0.072) + (0.704)x$$

when  $x = 0.075$ , then

$$\begin{aligned}y &= (0.072) + (0.704)(0.075) \\&= 0.1248 \quad \approx 12.48\%.\end{aligned}$$

Example:

Consider the following data

$x$	1	2	4	0
$y$	0.5	1	2	0

Fit a linear regression line

Estimate  $y$  when  $x = 5$ .

$x$	$y$	$xy$	$x^2$
1	0.5	0.5	1
2	1	2	4
4	2	8	16
0	0	0	0
$\sum x = 7$		$\sum xy = 3.5$	$\sum x^2 = 21$
$\sum y = 5$			

$$y = \beta_0 + \beta_1 x$$

$$\sum y = n\beta_0 + \beta_1 \sum x$$

$$\sum xy = \beta_0 \sum x + \beta_1 \sum x^2$$

$$3.5 = 7\beta_0 + \beta_1 \quad (?)$$

$$10.5 = -7\beta_0 + 7\beta_1 \quad (?)$$

On solving these

$$\beta_0 = 0$$

$$\beta_1 = 0.5$$

$$\text{i.e. } y = 0 + (0.5)x$$

$$\boxed{\text{When } x = 5, \quad y = (0.5)5 \\ = 0.25}$$

∴ Regression Line :  $y$  on  $x$

$$(y - \bar{y}) = b_{yx} (x - \bar{x})$$

↓  
Regression coeff of  $y$  on  $x$

$$P \frac{\sigma_y}{\sigma_x}$$

coeff of correlation

D) Regression Line :  $x$  on  $y$

$$(x - \bar{x}) = b_{xy} (y - \bar{y})$$



Regression coeff. of  $x$  on  $y$



$$b_{xy} = \frac{\sigma_x}{\sigma_y}$$



coeff of correlation

# Coefficient of Determination

---

$$\underline{SST = SSR + SSE}$$

$$\sum(Y_i - \bar{Y})^2 = \sum(\hat{Y}_i - \bar{Y})^2 + \sum(Y_i - \hat{Y}_i)^2$$

where:

**SST = Sum of Squares due to Total**

**SSR = Sum of Squares due to Regression**

**SSE = Sum of Squares due to Error**



$$r^2 = \frac{SSR}{SST} = \frac{\sum(\hat{Y} - \bar{Y})^2}{\sum(Y - \bar{Y})^2}$$

where:

**SST = Sum of Squares due to Total**

**SSR = Sum of Squares due to Regression**

# Multicollinearity

---

- It refers to the phenomenon of having related predictor variables in the input dataset.
- In simple terms, in a model which has been built using several independent variables, some of these variables might be interrelated, due to which the presence of that variable in the model is redundant. You drop some of these related independent variables as a way of dealing with multicollinearity.
- **Multicollinearity affects:**
  - **Interpretation:**
    - Does “change in Y, when all others are held constant” apply?
  - **Inference:**
    - Coefficients swing wildly, signs can invert
    - p-values are, therefore, not reliable

# Dealing with Multicollinearity

- Two basic ways of dealing with multicollinearity
- Looking at **pairwise correlations**
  - Looking at the correlation between different pairs of independent variables
- Checking the **Variance Inflation Factor (VIF)**
  - Sometimes pairwise correlations aren't enough
  - Instead of just one variable, the independent variable might depend upon a combination of other variables
  - VIF calculates how well one independent variable is explained by all the other independent variables combined

# Variance Inflation Factor (VIF)

- ❖ The VIF is given by:

- 
- 
- 
- where ' $i$ ' refers to the  $i$ -th variable which is being represented as a linear combination of rest of the independent variables.
- 

$$VIF_i = \frac{1}{1 - R_i^2}$$

- ❖ The common heuristic we follow for the VIF values is:

- **> 10:** Definitely high VIF value and the variable should be eliminated.
- **> 5:** Can be okay, but it is worth inspecting.
- **< 5:** Good VIF value. No need to eliminate this variable.

# Linear regression (multiple regression)

Example:-

	size	No of rooms	No of floors	Age of home	price Lakh
1	2000	5	2	45	4000
1	1400	3	1	40	2000
1	1600	3	2	30	3000
1	800	2	1	35	2000

Diagram illustrating the data points and a linear regression line:

- The x-axis represents the independent variables: size ( $x_1$ ), number of rooms ( $x_2$ ), number of floors ( $x_3$ ), and age of home ( $x_4$ ).
- The y-axis represents the dependent variable: price (Lakh) ( $y$ ).
- A black line represents the fitted regression line passing through the data points.
- Red arrows point from the labels  $x_1$ ,  $x_2$ ,  $x_3$ , and  $x_4$  to their respective columns in the table.
- Red arrows point from the label  $y$  to its column in the table.

# Multiple Linear Regression



The data consists of  $n$  observations on a dependent or response variable  $Y$  and  $p$  predictor or explanatory variables

$x_1, x_2, \dots, x_p$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon$$

$\beta_i$ 's are regression coefficients.

# Normal equations

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\sum y = \beta_0 n + \beta_1 \sum x_1 + \beta_2 \sum x_2$$

$$\sum x_1 y = \beta_0 \sum x_1 + \beta_1 \sum x_1^2 + \beta_2 \sum x_1 x_2$$

$$\sum x_2 y = \beta_0 \sum x_2 + \beta_1 \sum x_1 x_2 + \beta_2 \sum x_2^2$$

## Multiple Linear

$$\checkmark y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

$$\sum y = \beta_0 n + \beta_1 \sum x_1 + \beta_2 \sum x_2 + \beta_3 \sum x_3 + \beta_4 \sum x_4$$

$$\sum x_1 y = \beta_0 \sum x_1 + \beta_1 \sum x_1^2 + \beta_2 \sum x_1 x_2 + \beta_3 \sum x_1 x_3 + \beta_4 \sum x_1 x_4$$

$$\sum x_2 y = \beta_0 \sum x_2 + \beta_1 \sum x_1 x_2 + \beta_2 \sum x_2^2 + \beta_3 \sum x_2 x_3 + \beta_4 \sum x_2 x_4$$

$$\sum x_3 y = \beta_0 \sum x_3 + \beta_1 \sum x_1 x_3 + \beta_2 \sum x_2 x_3 + \beta_3 \sum x_3^2 + \beta_4 \sum x_3 x_4$$

$$\sum x_4 y = \beta_0 \sum x_4 + \beta_1 \sum x_1 x_4 + \beta_2 \sum x_2 x_4 + \beta_3 \sum x_3 x_4 + \beta_4 \sum x_4^2$$

so for  $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$

# Normal equations

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\sum y = \beta_0 n + \beta_1 \sum x_1 + \beta_2 \sum x_2$$

$$\sum x_1 y = \beta_0 \sum x_1 + \beta_1 \sum x_1^2 + \beta_2 \sum x_1 x_2$$

$$\sum x_2 y = \beta_0 \sum x_2 + \beta_1 \sum x_1 x_2 + \beta_2 \sum x_2^2$$

# Non - Linear



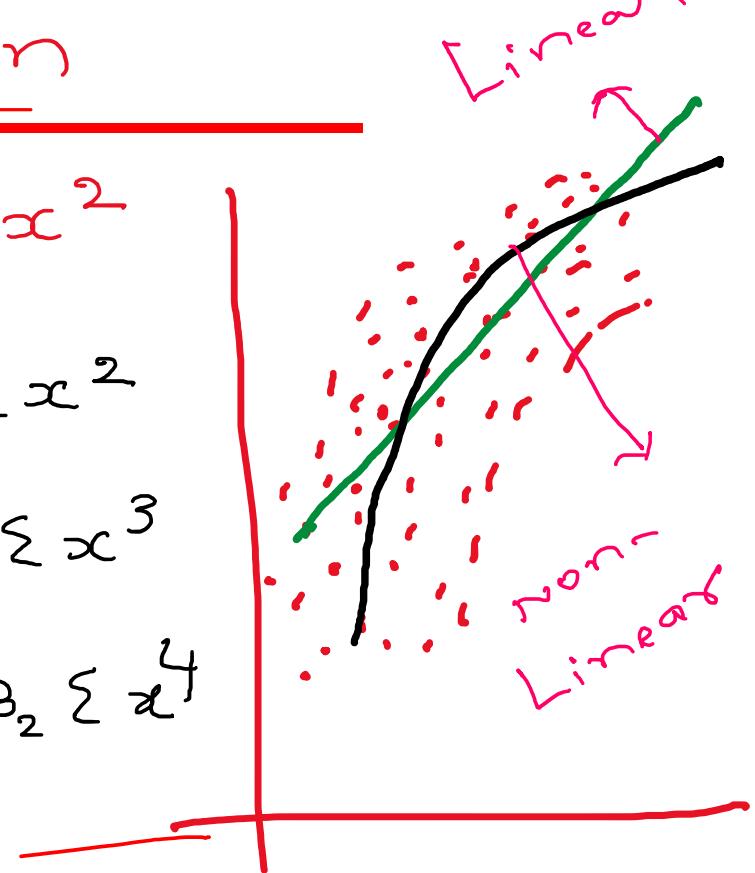
## Regression

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$\sum y = \beta_0 n + \beta_1 \sum x + \beta_2 \sum x^2$$

$$\sum xy = \beta_0 \sum x + \beta_1 \sum x^2 + \beta_2 \sum x^3$$

$$\sum x^2 y = \beta_0 \sum x^2 + \beta_1 \sum x^3 + \beta_2 \sum x^4$$



1)  $y = \alpha x^\beta$

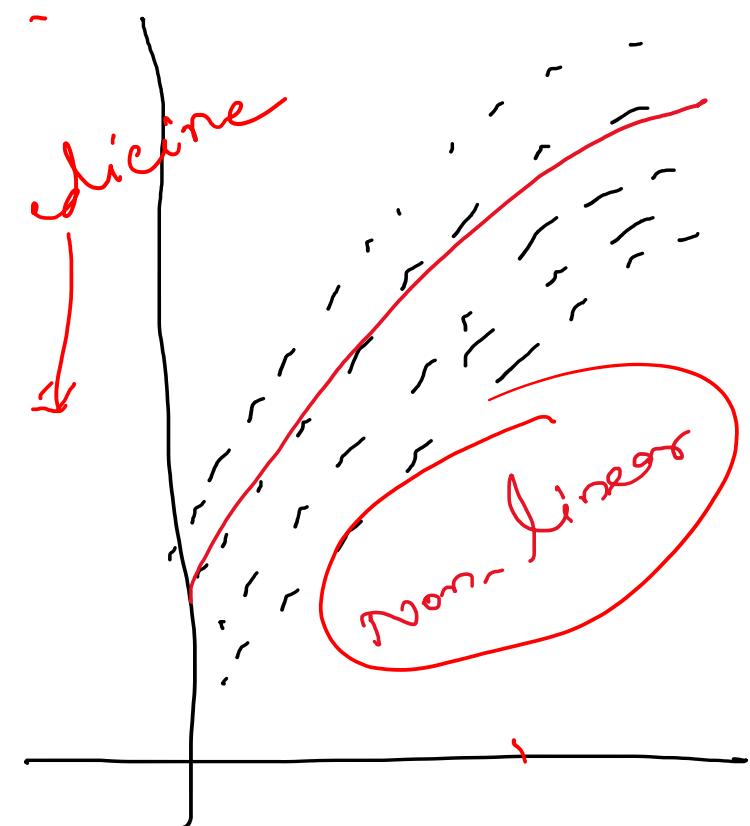
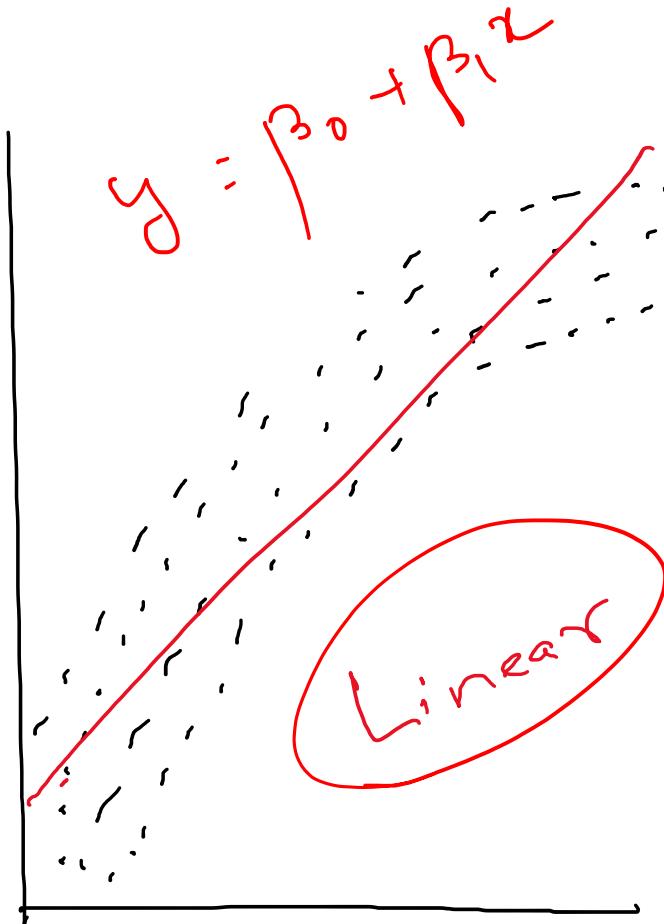
2)  $y = \alpha e^{\beta x}$

3)  $y = \alpha + \beta \log x$

4)  $y = \frac{\alpha}{\alpha x - \beta}$

# Other regressions

just a look



Suppose  $y = ae^{bx}$

$$\log y = \log a + b \log x$$

$\gamma$                        $x$

i.e.

$$\gamma = A + bx$$

$$\sum \gamma = A n + b \sum x \rightarrow \textcircled{1}$$

$$\sum x \cdot \gamma = A \sum x + b \sum x^2 \rightarrow \textcircled{2}$$

Hence, we get  
 $y = ae^{bx}$

Suppose  $y = ax^b$

$$\log(y) = \log(a) + b \log(x)$$

Y      A      X

i.e.  $Y = A + bX$

$$\sum Y = A_n + b \sum X$$

$$\sum XY: A \sum X + b \sum X^2$$

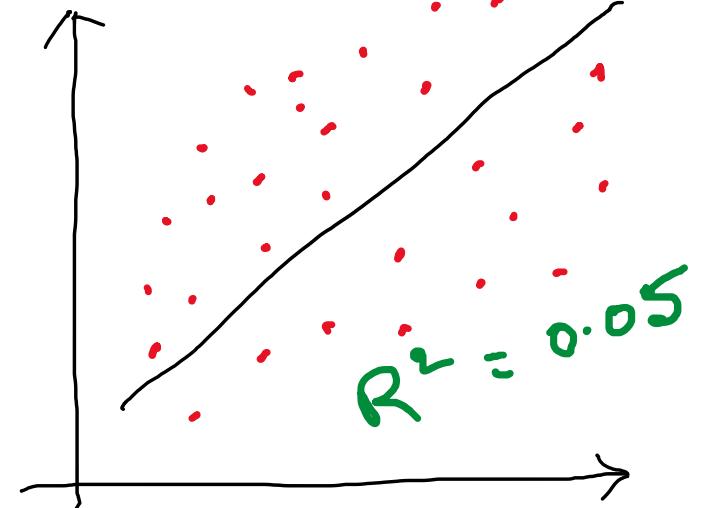
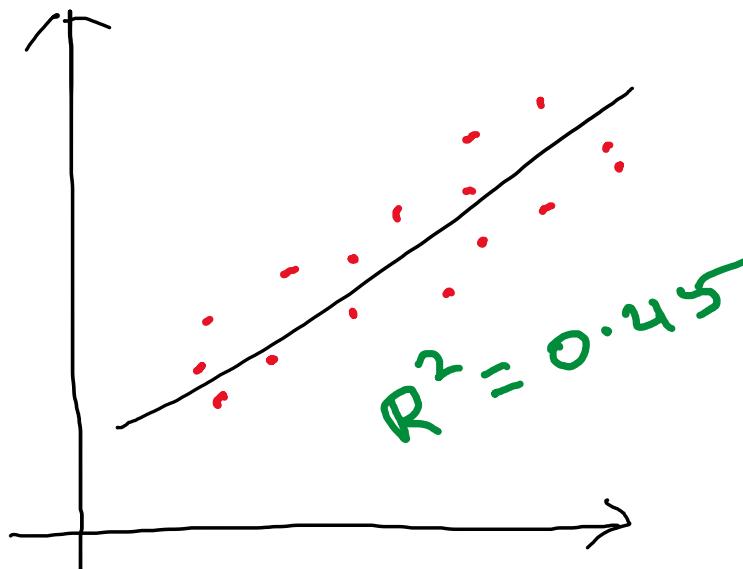
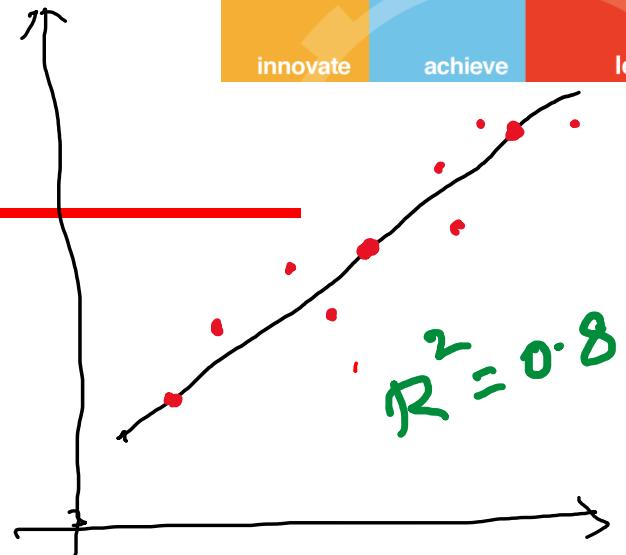
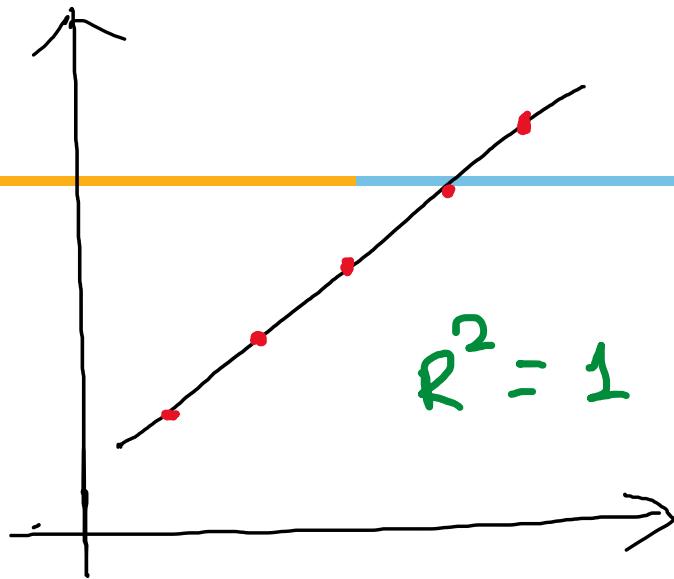
$$y = ax^b$$

RSS → Residual sum of squares

$$= \sum_{i=1}^n (y_i - (\alpha + \beta x_i))^2$$

TSS →  $\sum_{i=1}^n (y_i - \bar{y})^2$  mean of  
respective  
variables

$$R^2 = 1 - \frac{\text{RSS}}{\text{TSS}}$$





**Session 09(16<sup>th</sup> April 2022)**

**ANOVA &**

**Logistic Regression**

# Non - Linear



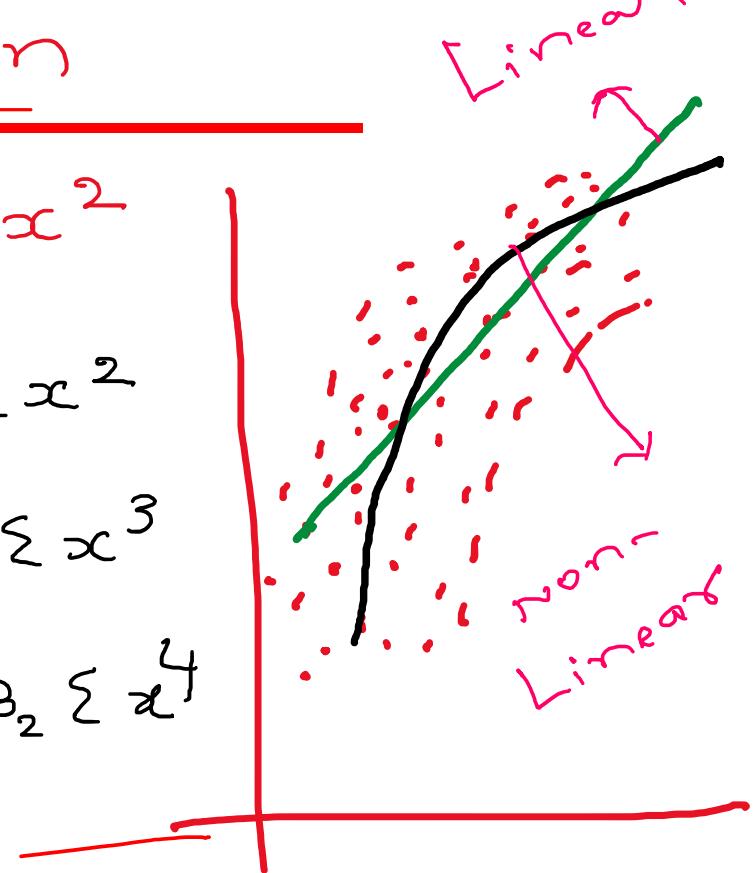
## Regression

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$\sum y = \beta_0 n + \beta_1 \sum x + \beta_2 \sum x^2$$

$$\sum xy = \beta_0 \sum x + \beta_1 \sum x^2 + \beta_2 \sum x^3$$

$$\sum x^2 y = \beta_0 \sum x^2 + \beta_1 \sum x^3 + \beta_2 \sum x^4$$



$$1) Y = \alpha x^{\beta}$$

$$2) Y = \alpha e^{\beta x}$$

$$3) Y = \alpha + \beta \log x$$

$$4) Y = \frac{\alpha}{\alpha x - \beta}$$

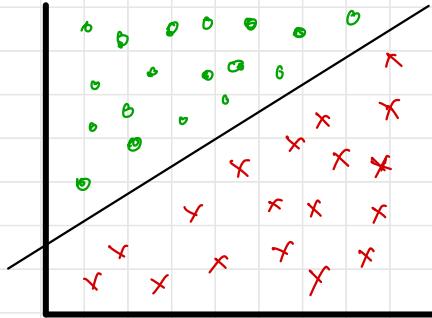
# modelling Qualitative Data

Risk factors for cancer

- age
- gender
- smoking
- diet
- family's medical history

The response variable was the person had cancer ( $y=1$ ) or did not have cancer ( $y=0$ ).

# REGRESSION AS A CLASSIFIER



$y = w_0 + w_1 x$  WE NEED TO TRANSFORM IT TO  $[0, 1]$   
THIS GIVES US A CLASSIFIER

IF  $y \rightarrow [0, 1]$

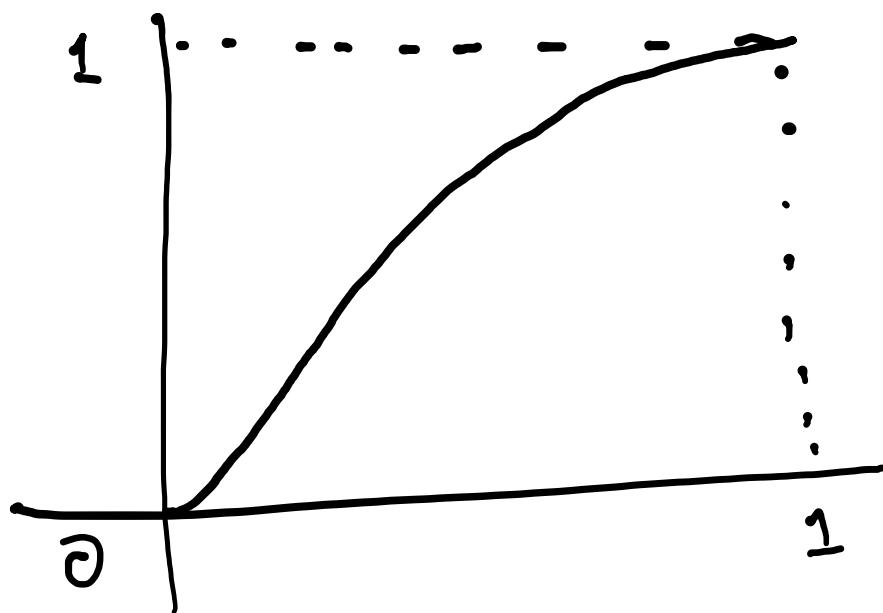
$0 \leftrightarrow 0.5$  |  $0.5 \leftrightarrow 1$   
CLASS 1 | CLASS 2

# Logistic Regression

- ❖ Logistic regression is a supervised classification model.
- ❖ This allows us to make predictions from labelled data ,if the target variable is categorical.
  - Binary classification
  - Examples
    - 1. A customer will default on a loan or not
    - 2. A particular machine will break down in the next month or not
    - 3. Predicting whether an incoming email is spam or not

$$P(Y=1 \mid X=x) = \beta_0 + \beta_1 x$$

 unbounded



logistic  
response  
function

Consider the following data

customer	Age	income lakh	gender	response
X	30	25	M	Yes 1
Y	45	40	F	No 0
Z	50	20	M	Yes 1
A	62	80	F	No 0
B	75	50	M	No 0
C	60	20	F	Yes 1

we want to model this, then

$$y = \beta_0 + \beta_1 (\text{Age}) + \beta_2 (\text{income}) + \\ + \beta_3 (\text{gender})$$

### Issues

1. Errors/residuals are not normally distributed.
2. No guarantee that the target/output/estimation is between 0 & 1.

Let us consider a function

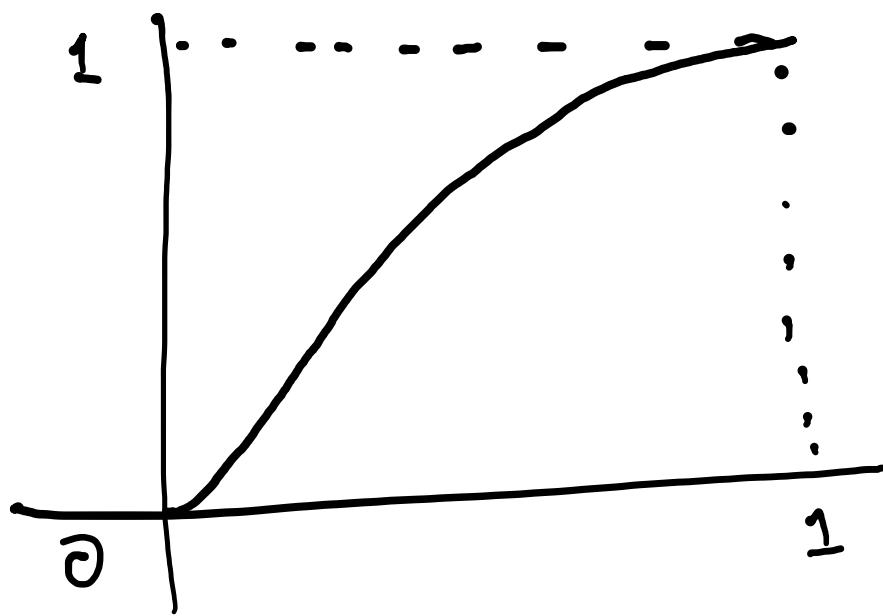
$$f(z) = \frac{1}{1 + e^{-z}}$$

$P(\text{response} = \text{yes} | \text{age}, \text{income}, \text{gender})$

$$= \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{income} + \beta_3 \cdot \text{gender})}}$$

$$P(Y=1 \mid X=x) = \beta_0 + \beta_1 x$$

unbounded



logistic  
response  
function

# Binary Logistic Regression



$Y$  = Binary response

$X$  = Quantitative predictor

$p$  = proportion of 1's (yes, success) at any  $X$

**Equivalent forms of the logistic regression model:**

Logit form

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X \quad , \quad p = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

Probability form

$$= \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}}$$

# Binary Logistic Regression Model



$Y$  = Binary

$X_1, X_2, \dots, X_k$  = Multiple

$\pi$  = proportion of 1's at any  $x_1, x_2, \dots, x_k$

Equivalent forms of the logistic regression model:

**Logit form**  $\log\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$

**Probability form**  $P = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k}}$

# REGRESSION

COST  
FUNCTION

→ SSE  
→ MSE

# LOGISTIC REGRESSION

→ ACCURACY  
→ PRECISION  
→ ENTROPY TEST

# Analysis of Variance

---

## Introduction:



**When there are more than two groups to be compared, it is not correct to compare the groups in pairs, as this type of comparison will not take the within variability into consideration**



**The Analysis procedure used in such comparisons is known as ANALYSIS OF VARIANCE**

## Regression

$$y = w_0 + w_1 x$$
$$\left. \begin{array}{l} w_0 = w_1 = 0 \\ w_0 = w_1 \neq 0 \end{array} \right\}$$

↓  
Testing of  
Hypothesis

Testing

$$\mu_1 = \dots, \quad \mu_1 = \mu_2$$

$$\mu_1 = \mu_2 = \mu_3 = \mu_A$$

$$y = w_0 + w_1 x_1 + w_2 x_2 + \dots + w_6 x_6$$

"ANOVA" ✓

# Case



- Testing the impact of nutrition and exercise on 60 candidates between age 18 and 50. They are grouped with different strategies. Now we need to find the most effective strategy
- Group 1 eats only junk food
- Group 2 eats only healthy food
- Group 3 eats junk food & does cardio exercise every other day
- Group 4 eats healthy food & does cardio .....
- Group 5 eats junk food & does both cardio & strength training every other day
- Group 6 eats healthy food.....

# ANOVA



- Effectiveness of different promotional activities
- Quality of a product produced by different manufacturers in terms of an attribute
- Yield of crop due to varieties of seeds , fertilisers and quality of soil

# ANOVA-analysis of variance



- \* Significance of difference between two sample means

$$H_0 = \mu = \mu_2 = \dots = \mu_k$$

$$H_1 = \mu_1 \neq \mu_2 \neq \dots \neq \mu_k$$

Null Hypothesis

Alternative Hypothesis

# Assumptions



- Each population is normally distributed with mean  $\mu$ ; With equal variances  $\sigma^2$
- Each sample is drawn randomly and independent of other samples

Observations } population

	A	B	C
1	26	18	23
2	25	16	19
3	28	17	26
4	12	18	30
Mean	22.75	17.25	24.50

Example

# ANOVA summary

Source of Variation	Sum of squares	d.o.f	Mean squares	F-value
Between (Samples)	SSTR	n-1	MSTR = $\frac{\text{SSTR}}{n-1}$	$F = \frac{\text{MSTR}}{\text{MSE}}$
within Samples (Error)	SSE	n-n	MSE = $\frac{\text{SSE}}{n-n}$	
Total	SST	n-1		

# Short cut method



$$T = \sum x_1 + \sum x_2 + \dots + \sum x_n$$

$$\text{cal. Fact} : CF = \frac{T^2}{n}, n = n_1 + n_2 + \dots + n_n$$

$$SST = \left[ \sum (x_1^2) + \sum x_2^2 + \dots + \sum x_n^2 \right] - CF$$

$$SSTR = \frac{(\sum x_j)^2}{n_j} - CF$$

$$SSE = SST - SSTR$$

# Example

---

- To test the significance of variation in the retail prices of a commodity in three metro cities, Mumbai, Kolkata and Delhi, four shops are chosen at random and the prices are given below

Mumbai : 16 8 12 14

Kolkata : 14 10 10 6

Delhi : 4 10 8 8

Prices in 3 cities are significantly different ?

---

# Example

---

- A study was conducted to investigate the perception of corporate ethical values among individuals specialising in marketing. Using 0.05 level of significance and the data given below, test for significant differences in perception among three groups.( higher scores indicate higher ethical values)
-

Marketing  
manager

Marketing  
Research

Advertising

6

5

4

5

6

4

5

5

4

4

5

4

6

7

6

5

6

6

$$\pi = 3, m = 18$$

$$T = \sum x_1 + \sum x_2 + \sum x_3 \\ = 30 + 27 + 36 = 93$$

$$CF = \frac{T^2}{m} = \frac{(93)^2}{18} = 480.50$$

$$SST = (\sum x_1^2 + \sum x_2^2 + \sum x_3^2) - CF \\ = 154 + 123 + 218 = 495.00$$

$$SSTR = \left( \frac{\sum x_1^2}{n_1} + \frac{\sum x_2^2}{n_2} + \frac{\sum x_3^2}{n_3} \right) - CF$$

$$= \frac{(30)^2}{6} + \frac{(27)^2}{6} + \frac{(36)^2}{6} - 480.50$$

$$\approx 7$$

$$SSE = SST - SSTR$$

$$= 14.50 - 7 = 7.50$$

ACROSS  
GROUPS

$$MSTR = \frac{SSTR}{df_1} = \frac{7}{2} : 3.5$$

WITHIN  
SAMPLE

$$MSE = \frac{SSE}{df_2} = \frac{7.5}{df_2} : 0.5$$

$$F = \frac{MSTR}{MSE} = \frac{3.5}{0.5} = 7$$

DOF

$N \Rightarrow \delta - 1 = 2$  calculated value: 7  
 $D \Rightarrow n - \delta = 15$  table value: 3.68

7 > 3.68  $\Rightarrow$  Rejected.

# Two way ANOVA

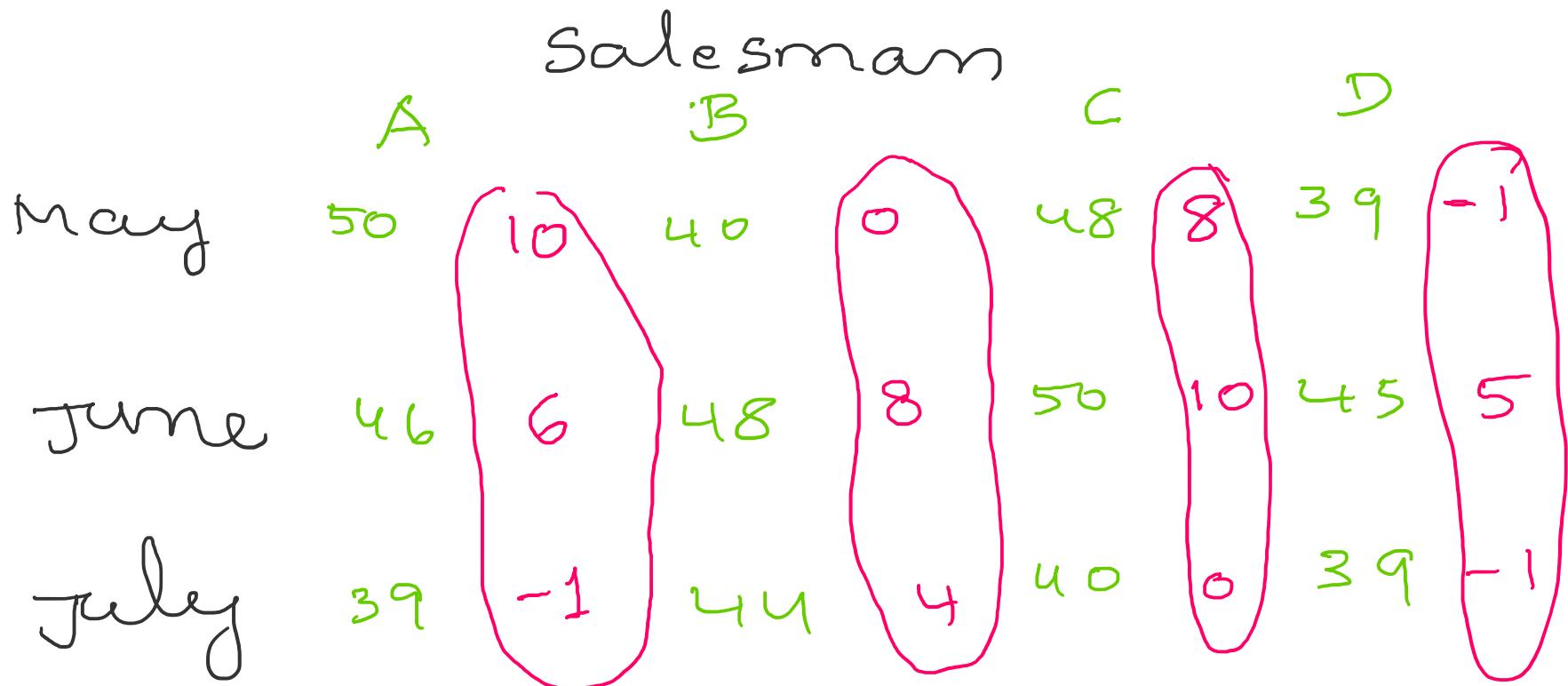
Sources of variation	Sum of square	D.o.F	mean square	test statistic
Between columns }	SSTR	c-1	MSTR = SSTR / c-1	F treatment = MSTR / MSE
Between rows }	SSR	n-1	MSR = SSR / n-1	
Residual { error }	SSE	(c-1) * (n-1)	MSE = SSE / (c-1)(n-1)	F blocks = MSR / MSE
Total	<u>SST</u>	<u>n-1</u>		

# Example (Two way ANOVA)

		Sales			
		A	B	C	D
Month	May	50	40	48	39
	June	46	48	50	45
	July	39	44	49	39

→ Is there any significant difference in the sales made by A, B, C, D

→ Is there a significant diff in the sales made during these months.



TAKE A VALUE AND SUBTRACT  
FROM EVERY VALUE

$$T = 15 + 12 + 18 + 3 = 48$$

$$CF = \frac{T^2}{n} = \frac{(48)^2}{12} = 192$$

↳ Salesman

SSTR = Sum of squares (columns)

$$= \left( \frac{15^2}{3} + \frac{12^2}{3} + \frac{18^2}{3} + \frac{3^2}{3} \right) - 192$$

$$= 42$$

SSR = Sum of squares between  
months (rows)

$$= \left( \frac{17^2}{4} + \frac{29^2}{4} + \frac{22^2}{4} \right) - 192 \\ = 91.5$$

$$SST = \left( \sum x_1^2 + \sum x_2^2 + \sum x_3^2 + \sum x_4^2 \right) - CF \\ = (137 + 80 + 164 + 27) - 192 \\ = 216$$

$$\begin{aligned}SSE &= SST - (SSTR + SSR) \\&= 216 - (42 + 91.5) \\&= 82\end{aligned}$$

$$\begin{aligned}df_c &: 3, \quad df_n : n-1 = 3-1 = 2 \\y_{c-1} & \\df &: (c-1)(n-1) = 3 \times 2 = 6\end{aligned}$$

$$MSTR = \frac{SSTR}{c-1} = \frac{42}{3} = 14$$

$$MSR = \frac{SSR}{n-1} = \frac{91.5}{2} = 45.75$$

$$MSE = \frac{SSE}{(c-1)(n-1)} = \frac{\frac{82.5}{6}}{= 13.75}$$

	Sum of squares	D.o.f	mean squares	Variance ratio
x Between salesmen	SSTR 42.0	c-1 3	MSTR $= \frac{42.0}{3} = 14$	$F_{\text{treatment}}$ $= \frac{14}{13.75}$ $= 1.018$
,	Between months	SSR 91.5	τ-1 2	MSR $= 45.75$
x residual errors	SSE 82.5	$(c-1)(\tau-1)$ 6	MSE 13.75	$F_{\text{block}}$ $= \frac{45.75}{13.75}$ $= 3.327$
Total	216	11		

a)  $F_{\text{treatment}} = 1.018 <$   
 $df_1 = 3, df_2 = 6,$   
 $\alpha = 0.05$

accept

b)  $F_{\text{block}} = 3.327 <$

accept

difference in the sales by ~~salesman~~

difference in sales made during months.

# One - way Analysis of Variance

Three drying formulas for curing glue are studied



Formula A	13	10	8	11	8
Formula B	13	11	14	14	
Formula C	4	1	3	4	2

Test at 5% level of significance whether is any difference in the mean curing time of glue?



Find between which two formulas the mean difference has contributed significantly using least significant difference post-hoc test?

# One - way Analysis of Variance

## **Calculation of sum of squares**

$$1. \text{ Correction factor (CF)} : \frac{G^2}{n} = \frac{150^2}{15} = 1500$$

$$2. TSS = \sum_i \sum_j x_{ij}^2 - CF = 1594 - 1500 = 94$$

$$3. GSS = \sum_{i=1}^k \frac{C_i^2}{n_i} - CF = 1560 - 1500 = 60$$

$$4. WSS = TSS - GSS = 94 - 60 = 34$$

# One - way Analysis of Variance

One-way ANOVA table

Source of variation	df	Sum of squares	Mean sum of squares	F-ratio
Between groups	2	60	30.000	$F=10.588$
Within groups	12	34	2.833	
Total	14	94	$F(4.46) > F_{(0.05; 2, 12)} = 3.885$	

$H_0$  may be rejected and  $H_1$  may be accepted.

# One - way Analysis of Variance

 Least significant difference (LSD) test

 Analysis of Variance provides estimate of Standard error for testing which of the differences between the villages is significant. An estimate of the standard error of the differences between the group means is equal to



$$\sqrt{\frac{2S^2}{k}}$$



Where  $S^2$  is the 'Within groups mean sum of squares and k is the number of observations in each of the group under comparison

# One - way Analysis of Variance

**Post-hoc test: Multiple comparison using Bonferroni's test**

Group (I)	Group (J)	Mean Difference (I-J)	Std. Error	P- value	95% Confidence Interval	
					Lower Bound	Upper Bound
Formula A	Formula B	3.000	1.129	0.063	-0.14	6.14
Formula A	Formula C	2.000	1.019	0.220	-0.83	4.83
<b>Formula B</b>	<b>Formula C</b>	<b>5.000*</b>	<b>1.087</b>	<b>0.002</b>	<b>1.98</b>	<b>8.02</b>

\*The mean difference is significant at the 0.05 level.





# **Session 10(23<sup>rd</sup> April 2022)**

## **Time Series Analysis**

a)  $F_{\text{treatment}} = 1.018 <$   
 $df_1 = 3, df_2 = 6,$   
 $\alpha = 0.05$

accept

b)  $F_{\text{block}} = 3.327 <$

accept

difference in the sales by ~~salesman~~

difference in sales made during months.

# Forecasting



- Predict the next number in the pattern:
  - a) 3.7, 3.7, 3.7, 3.7, 3.7, ?
  - b) 2.5, 4.5, 6.5, 8.5, 10.5, ?
  - c) 5.0, 7.5, 6.0, 4.5, 7.0, 9.5, 8.0, 6.5, ?

# Forecasting

- Predict the next number in the pattern:

a) 3.7, 3.7, 3.7, 3.7, 3.7, **3.7**

b) 2.5, 4.5, 6.5, 8.5, 10.5, **12.5**

c) 5.0, 7.5, 6.0, 4.5, 7.0, 9.5, 8.0, 6.5, **9.0**

# What is a Time Series?

---

- Set of evenly spaced numerical data
  - Obtained by observing response variable at regular time periods
- Forecast based only on past values
  - Assumes that factors influencing past, present, & future will continue
- Example

Year:	1995	1996	1997	1998	1999
Sales:	78.7	63.5	89.7	93.2	92.1

---

# Time series \_ components

---



- Trend
  - Cyclical
  - Seasonality
  - Random / irregularity
-

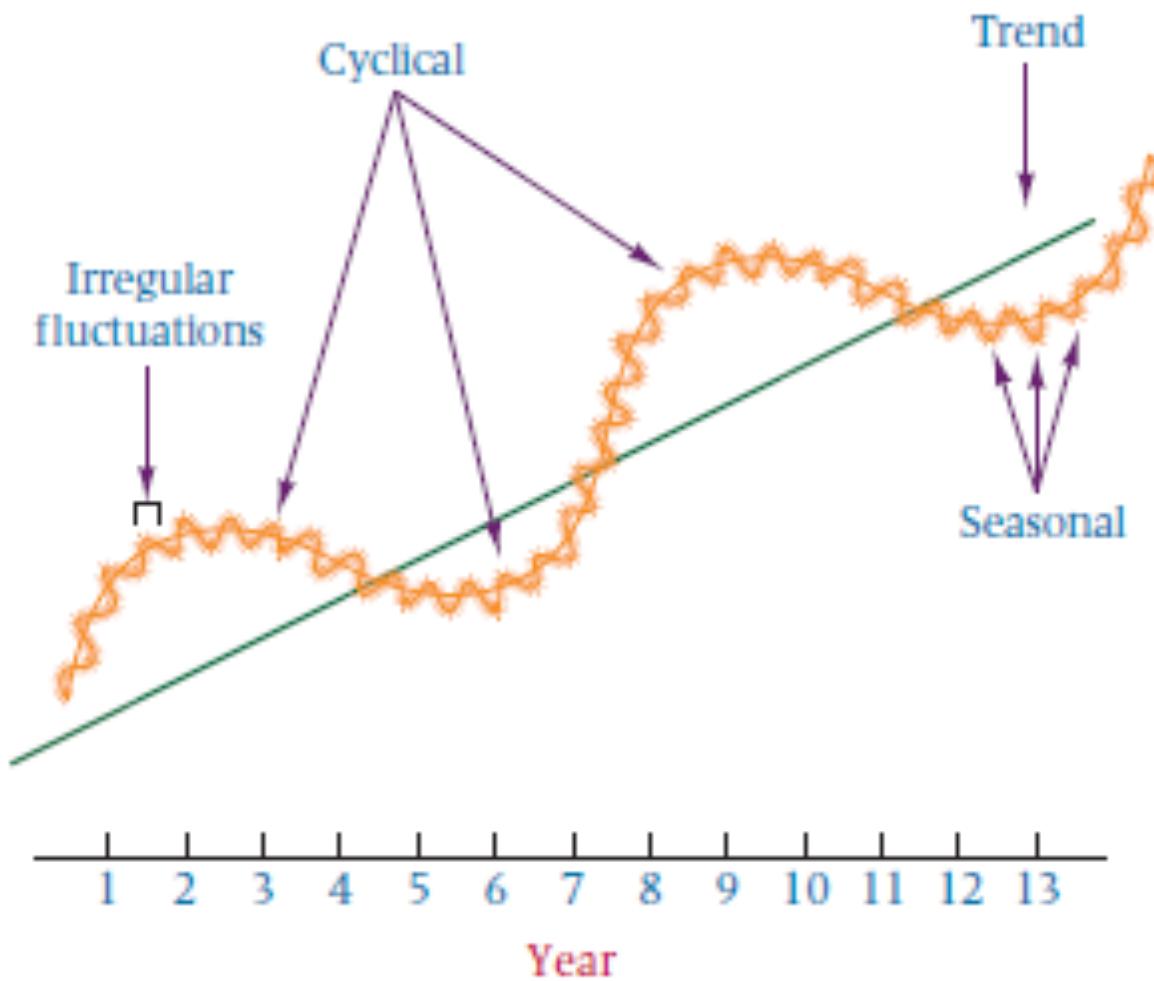
# Time Series Models

- 
- Forecaster looks for data patterns as
    - Data = historic pattern + random variation
  - Historic pattern to be forecasted:
    - Level (long-term average) – data fluctuates around a constant mean
    - Trend – data exhibits an increasing or decreasing pattern
    - Seasonality – any pattern that regularly repeats itself and is of a constant length
    - Cycle – patterns created by economic fluctuations
  - Random Variation cannot be predicted
-

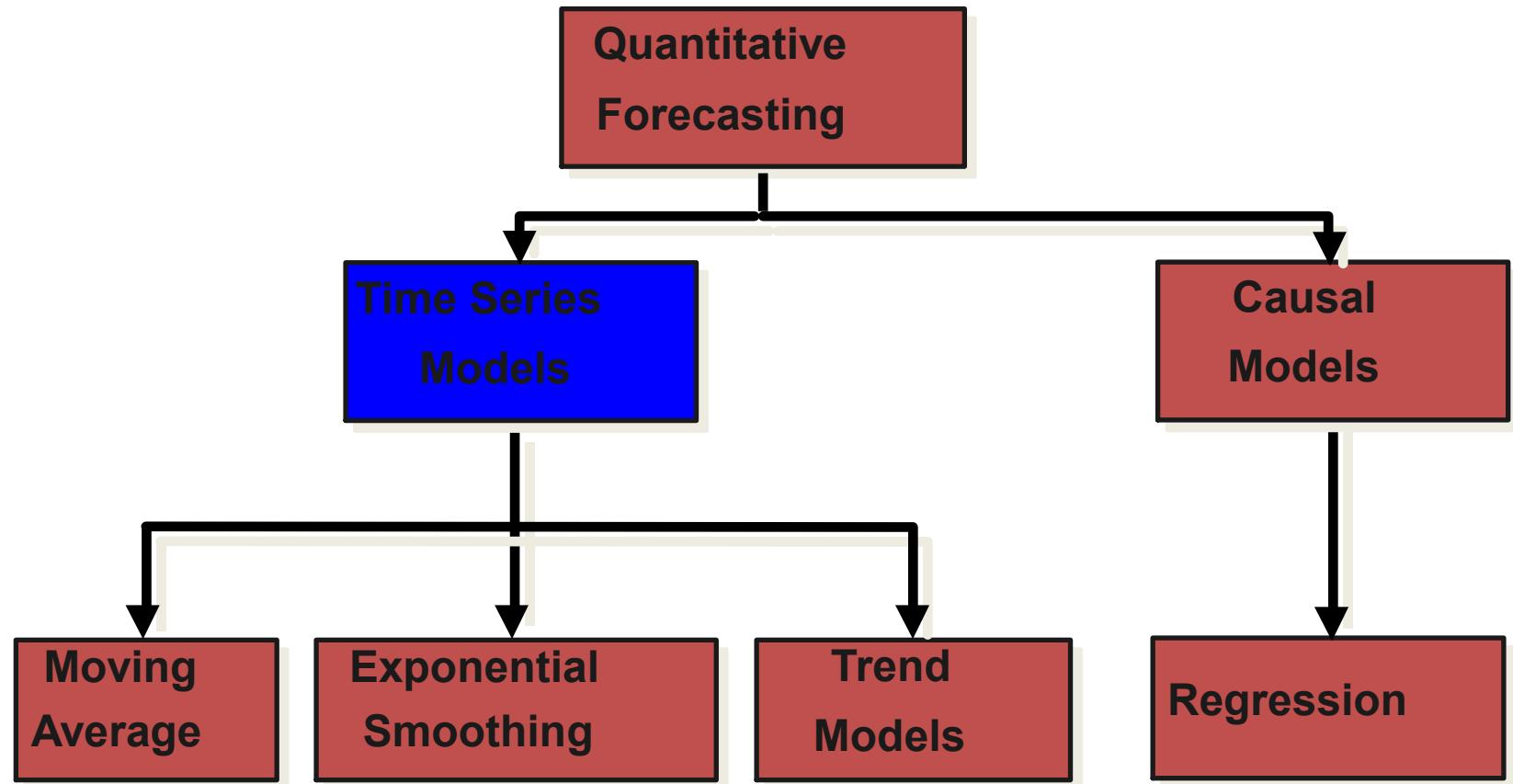
- Time series data that shows no trend, seasonality and cyclicalities is said to be **stationary**.
- The statistical properties of such a time series are independent of time.

i.e

- ✓ The process generating the data has a constant mean
- ✓ The variability is constant over time

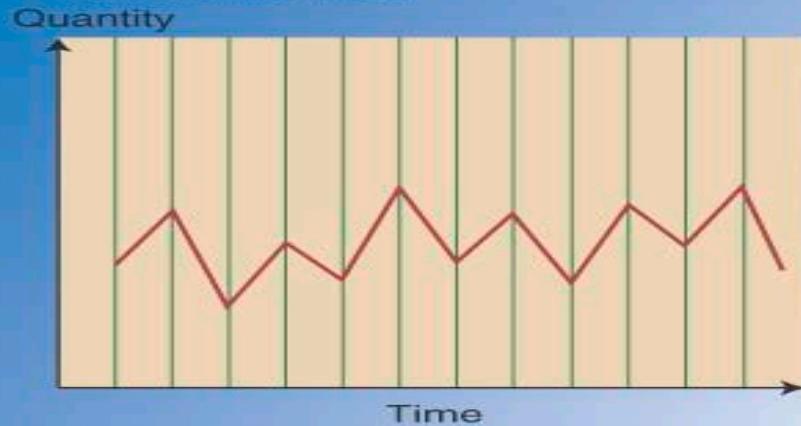


# Quantitative Forecasting Methods



# Time Series Patterns

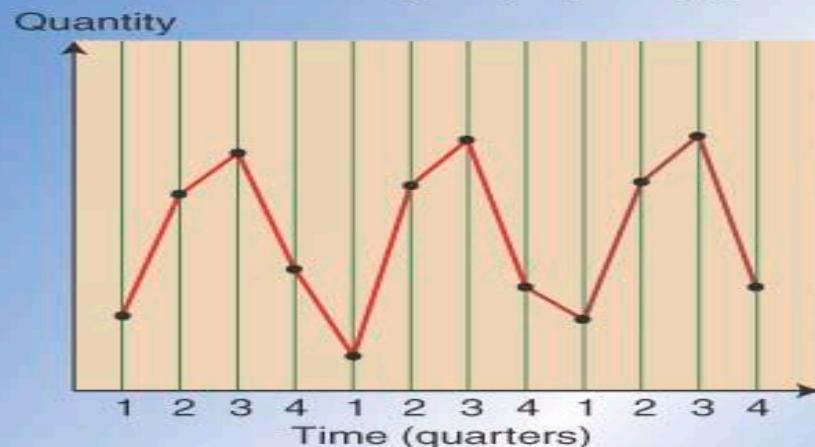
**(a) Level or Horizontal Pattern:**  
Data follows a horizontal pattern around the mean



**(b) Trend Pattern:**  
Data is progressively increasing (shown) or decreasing



**(c) Seasonal Pattern:**  
Data exhibits a regularly repeating pattern



**(d) Cycle:**  
Data increases or decreases over time



# Box – Jenkins Methodology



## 1. Condition data and select a model

- ❖ identify and account for any trends or seasonality in the time series
- ❖ examine the remaining time series and determine a suitable model

## 2. Estimate the model parameters

## 3. Assess the model and return to step 1, if necessary

# Time Series Components

A time series can be described by models based on the following components

$T_t$       **Trend Component**

$S_t$       **Seasonal Component**

$C_t$       **Cyclical Component**

$I_t$       **Irregular Component**

Using these components we can define a time series as the sum of its components or an **additive model**

$$X_t = T_t + S_t + C_t + I_t$$

Alternatively, in other circumstances we might define a time series as the product of its components or a **multiplicative model** – often represented as a logarithmic model

$$X_t = T_t S_t C_t I_t$$



# Smoothing Methods

# Moving Average Models

- Simple Moving Average Forecast

$$F_t = E(Y_t) = \frac{\sum_{i=t-k}^{t-1} Y_i}{k}$$

Weighted Moving Average Forecast

$$F_t = E(Y_t) = \frac{\sum_{i=t-k}^{t-1} w_i Y_i}{k}$$

# Example(Moving averages)

- Use the following data to compute three

YEAR	Saleson (Lakhs)	YEAR	Saleson (Lakhs)
2008	21	2013	22
2009	22	2014	25
2010	23	2015	26
2011	25	2016	27
2012	24	2017	26

Year	Product	3 Year Moving Avg	Error Forecast
2008	21		
2009	22	$\frac{66}{3} = 22.00$	0
2010	23	$\frac{70}{3} = 23.33$	-0.33
2011	25	$\frac{72}{3} = 24.00$	1.00
2012	24	$\frac{71}{3} = 23.67$	0.33
2013	22	$\frac{71}{3} = 23.67$	-1.67
2014	25	$\frac{73}{3} = 24.33$	0.67
2015	26	$\frac{78}{3} = 26.00$	0
2016	27	$\frac{79}{3} = 26.33$	0.67
2017	26	-	-

- **Weighted Moving Average:**

- All weights must add to 100% or 1.00

e.g.  $C_t .5$ ,  $C_{t-1} .3$ ,  $C_{t-2} .2$  (weights add to 1.0)

$$F_{t+1} = \sum C_t A_t$$

- Allows emphasizing one period over others; above indicates more weight on recent data ( $C_t=.5$ )
- Differs from the simple moving average that weighs all periods equally - more responsive to trends

# Example(Weighted moving Averages)

Weights	Month
3	Last month
2	Two months ago
1	Three months ago

Months	1	2	3	4	5	6	7	8	9	10	11	12
Sales	10	12	13	16	19	23	26	30	28	18	16	14

## Example .

### monthly Demand

43 105

44 106

45 110

46 110

47 114

48 121

49 130

50 128

51 137

a) forecast demand  
for month 52  
using 5-month  
moving Avg

b) .. weighted  
moving average  
with weights  
3, 2, 1 - latest  
descending

## Example .

W-5



month	Demand		
43	105	-	a)
44	106	-	$110 + 121 + 130 +$
45	110	109.50	$128 + 137 / 5$
46	110	112.2	$= 126$
47	119	117.0	b)
48	121	120.6	$3 + 137 +$
49	130	126.0	$2 + 128 +$
50	128	-	$1 \times 130$
51	137	-	$\cancel{123}$ units

Handwritten annotations and calculations:

- Month 45: A green vertical line connects demand values 106 and 110, labeled "sum". An arrow points from 110 to the result 109.50.
- Month 46: A green vertical line connects demand values 110 and 119, labeled "sum". An arrow points from 119 to the result 112.2.
- Month 47: A green vertical line connects demand values 119 and 121, labeled "sum". An arrow points from 121 to the result 117.0.
- Month 48: A green vertical line connects demand values 121 and 130, labeled "sum". An arrow points from 130 to the result 120.6.
- Month 49: A green vertical line connects demand values 130 and 128, labeled "sum". An arrow points from 128 to the result 126.0.
- Month 50: A pink vertical line connects demand values 128 and 137, labeled "sum". An arrow points from 137 to the result 126.0.
- Month 51: A pink vertical line connects demand values 137 and 137, labeled "sum". An arrow points from 137 to the result 137.

# Time Series Models

- **Exponential Smoothing:**

Most frequently used time series method because of ease of use and minimal amount of data needed

- Need just three pieces of data to start:
  - Last period's forecast ( $F_t$ )
  - Last period's actual value ( $A_t$ )
  - Select value of smoothing coefficient,  $\alpha$ , between 0 and 1.0
- If no last period forecast is available, average the last few periods or use naive method
- Higher  $\alpha$  values may place too much weight on last period's random variation

$$F_{t+1} = \alpha A_t + (1 - \alpha) F_t$$

$$= F_t + \alpha (A_t - F_t)$$

# Example

- Forecast for the first week of March 1 was 500 units whereas the actual demand is 450 units
  - a) Forecast demand for the next week i. e March 8
  - b) Assume the actual demand during the march is 505 units
- Continue the forecasting assuming that subsequent demands were actually 516, 488, 467, 554 and 510 units

## Example:-

Forecast for the first week of March was 500 units whereas the actual demand is 450 units

a) Forecast demand for the next week i.e. March 8

$$\begin{aligned}
 F_{\text{t+1}} &= F_t + \alpha (A_t - F_t) \\
 &= 500 + 0.1 (450 - 500) \\
 &= \underline{\underline{495}}
 \end{aligned}$$

Week	Demand	$(A_t)$	$(F_t)$ (old)	Forecast	New forecast
March	1	450	500	500	$500 + 0.1(450 - 500) = 495$
	8	505	495	495	$495 + 0.1(505 - 495) = 496$
	15	516	496	496	$496 + 0.1(516 - 496) = 498$
	22	488	498	498	$498 + 0.1(488 - 498) = 497$
April	1	467	497	497	$497 + 0.1(467 - 497) = 494$
	8	554	494	494	$494 + 0.1(554 - 494) = 500$
	15	510	500	500	$500 + 0.1(510 - 500) = 501$

# Stationarity

---

Stationary time series have no trend.  
conditions

1. constant mean
  2. Constant variance
  3. An autocovariance that does not depend on time
-

# Auto Correlation



auto covariance<sub>h</sub>(x<sub>t</sub>)

$$= \text{cov}(x_t, x_{t-h})$$

auto correlation<sub>h</sub>(x<sub>t</sub>)

$$= \frac{\text{Auto cov}_h(x_t)}{\text{std}(x_t) \text{ std}(x_{t-h})}$$

# Auto Correlation Function



auto covariance<sub>h</sub>(x<sub>t</sub>)

$$\gamma_x(h) = \text{cov}(x_t, x_{t-h})$$

$$\text{ACF} = \frac{\gamma_x(h)}{\gamma_x(0)} = \text{Cor}(x_t, x_{t-h})$$

# Stationarity

---

Stationary time series have no trend.  
conditions

1. constant mean
  2. Constant variance
  3. An autocovariance that does not depend on time
-

# iid noise

---

The time series in which there is no trend or seasonal component and the observations are simply independent and identically distributed (iid) random variables with zero mean.

such sequence of random variables  $x_1, x_2 \dots x_n$  as iid noise

---

Differencing is the method of transforming a non-stationary time series into a stationary one.

The first differencing value is the difference b/w the current time period and the previous time period.

How many times - - - ?



# AR Model (auto regressive Model)

# AR model

$y_t$  depends on  $y_{t-1}, y_{t-2}, \dots$  etc

$$y_t = f(y_{t-1}, y_{t-2}, \dots, y_{t-p}, \epsilon_t)$$

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \dots +$$

$$+ \underbrace{\dots + \beta_p y_{t-p}}_{\text{value of time series } t-1} + \epsilon_t$$

constant

$$\text{AR}(0) \Rightarrow y_t = \beta_0$$

$$\text{AR}(1) \Rightarrow y_t = \beta_0 + \beta_1 y_{t-1}$$

# MA model (Moving averages model)

# MA model

$$y_t = f(\epsilon_t, \epsilon_{t-1}, \epsilon_{t-2}, \dots)$$

MA (q)

$$y_t = \theta_0 + \epsilon_t + \bar{\theta}_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \dots + \theta_q \epsilon_{t-q}$$



# ARMA model

# ARMA model – ARMA(p,q)



$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \dots + \beta_p y_{t-p} +$$

$$\theta_0 + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q}$$

✓ If  $p \neq 0$  and  $q=0$ , then AR (P)

✓ If  $p=0$  and  $q \neq 0$ , then MA (Q)



# ARIMA model (Autoregressive Integrated Moving Average Model)

# ARIMA Model



AR : Autoregression - dept relations  
b/w observations & lagged observations

I : Integrated → stationary

MA : moving average  
→ dependency relation b/w  
observations and residual error

# Parameters of the model

---

P : the number of lag observations  
ie lag order

d : degree of differencing

q : order of moving average

---

# In implementation

---

- ✓ Import the data
  - ✓ Convert it into time series
  - ✓ Plot the time series
  - ✓ Plot ACF and PACF
  - ✓ Then fit the model
-



**BITS** Pilani

Pilani|Dubai|Goa|Hyderabad

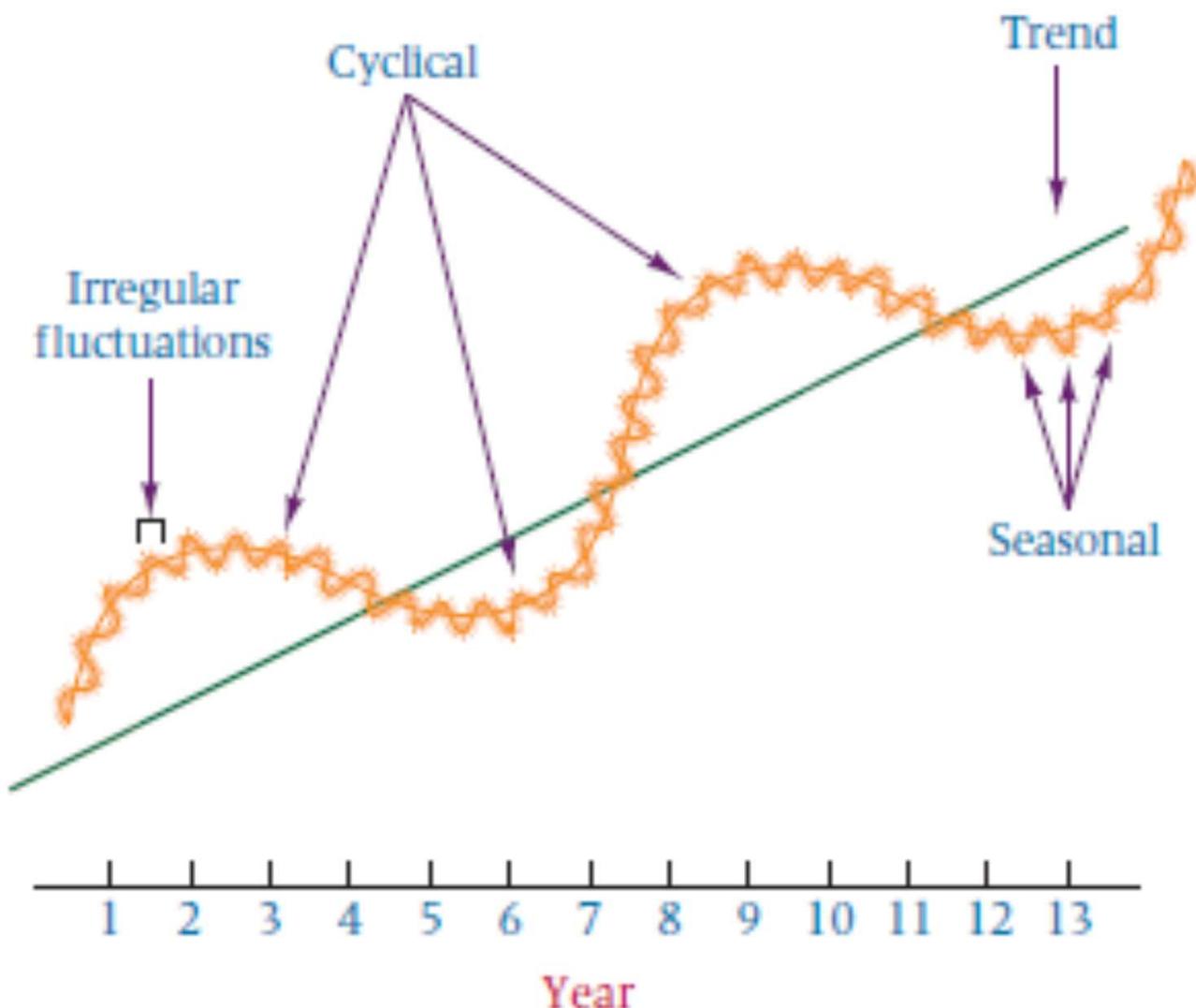
# Advanced Statistical Techniques for Analytics

**Dr Y V K Ravi Kumar, BITS- Pilani**



# **Session 10(23<sup>rd</sup> April 2022)**

## **Time Series Analysis & Multivariate Analysis**

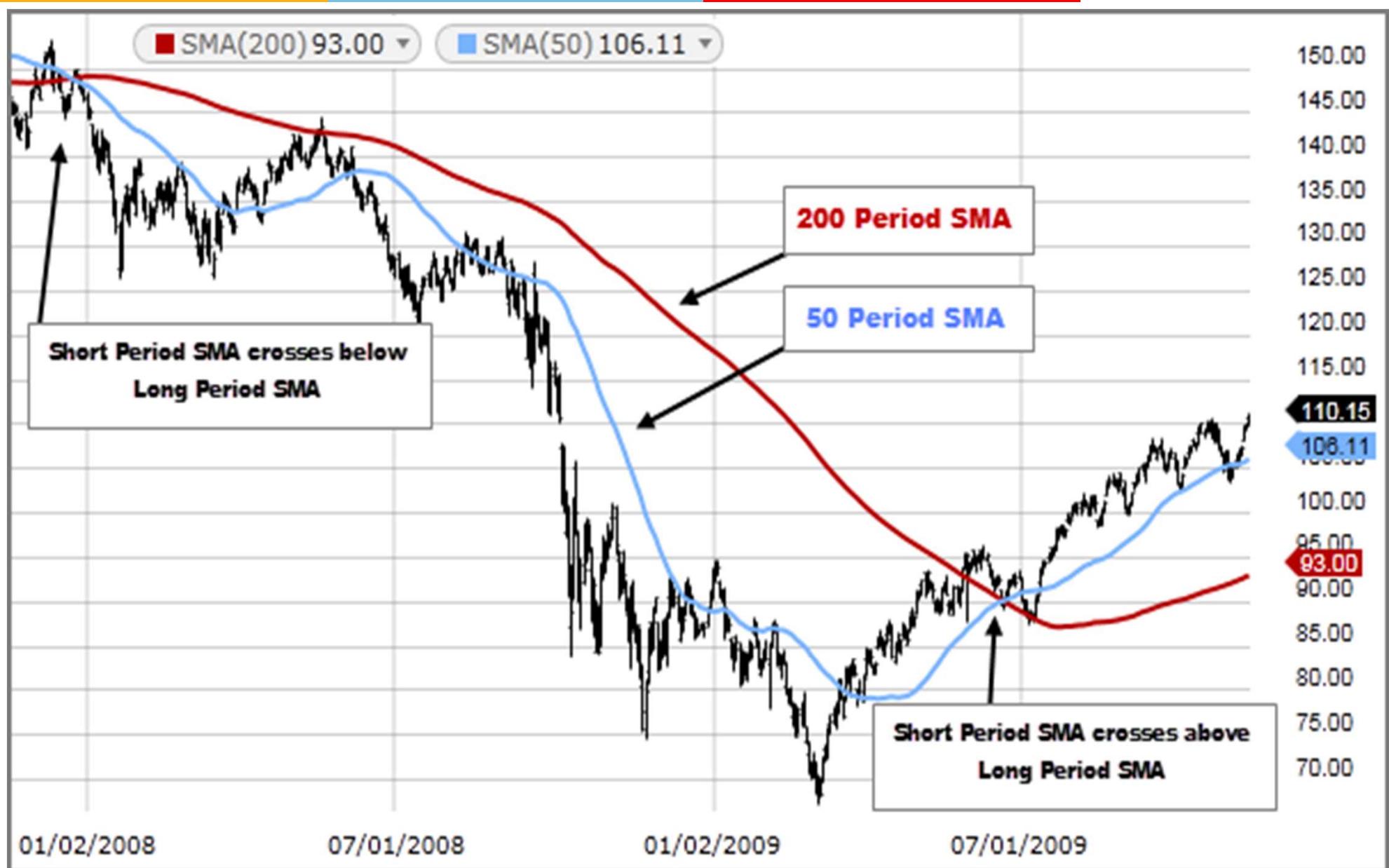




BATS:AAPL, D 318.31 ▼ -0.92 (-0.29%) O:320.25 H:323.33 L:317.52 C:318.31



TradingView





# Time Series Components

A time series can be described by models based on the following components

$T_t$	<b>Trend Component</b>
$S_t$	<b>Seasonal Component</b>
$C_t$	<b>Cyclical Component</b>
$I_t$	<b>Irregular Component</b>

Using these components we can define a time series as the sum of its components or an **additive model**

$$X_t = T_t + S_t + C_t + I_t$$

Alternatively, in other circumstances we might define a time series as the product of its components or a **multiplicative model** – often represented as a logarithmic model

$$X_t = T_t S_t C_t I_t$$



# Smoothing Methods

# Moving Average Models

Simple Moving Average Forecast

$$F_t = E(Y_t) = \frac{\sum_{i=t-k}^{t-1} Y_i}{k}$$

Weighted Moving Average Forecast

$$F_t = E(Y_t) = \frac{\sum_{i=t-k}^{t-1} w_i Y_i}{k}$$

## Example(Moving averages)

Use the following data to compute three year moving average for all available years.

Find the trend and Forecast error

YEAR	Saleson (Lakhs)	YEAR	Saleson (Lakhs)
2008	21	2013	22
2009	22	2014	25
2010	23	2015	26
2011	25	2016	27
2012	24	2017	26



# Time Series Models

- **Weighted Moving Average:**

- All weights must add to 100% or 1.00

e.g.  $C_t .5, C_{t-1} .3, C_{t-2} .2$  (weights add to 1.0)

$$F_{t+1} = \sum C_t A_t$$

- Allows emphasizing one period over others; above indicates more weight on recent data ( $C_t=.5$ )
- Differs from the simple moving average that weighs all periods equally - more responsive to trends

## Example(Weighted moving Averages)

Weights	Month
3	Last month
2	Two months ago
1	Three months ago

Months	1	2	3	4	5	6	7	8	9	10	11	12
Sales	10	12	13	16	19	23	26	30	28	18	16	14

## Example.

month	Demand
43	105
44	106
45	110
46	110
47	114
48	121
49	130
50	128
51	137

- a) forecast demand  
for month 52  
using 5-month  
moving Avg
- b) .. weighted  
moving average  
with weights  
3, 2, 1 - latest to  
descending

# Example.

month

Demand

43

105

-

a)

44

106

-

$$11 \text{ units} + 121 + 130 +$$

45

110 → 109.50

$$128 + 137 - 5$$

46

110 → 112.2

$$= 126$$

47

114 → 117.0

$$3 + 137 +$$

48

121 → 120.6

$$2 + 128 +$$

49

130 → 126.0

$$1 \times 130$$

50

128 -

$$= \cancel{133} \text{ units}$$

51

137 -

# Time Series Models

- Exponential Smoothing:

Most frequently used time series method because of ease of use and minimal amount of data needed

- Need just three pieces of data to start:

- Last period's forecast ( $F_t$ )
- Last periods actual value ( $A_t$ )
- Select value of smoothing coefficient,  $\alpha$ , between 0 and 1.0

$$F_{t+1} = \alpha A_t + (1 - \alpha) F_t$$

$$= F_t + \alpha(A_t - F_t)$$

- If no last period forecast is available, average the last few periods or use naive method
- Higher  $\alpha$  values may place too much weight on last period's random variation

# Example

Forecast for the first week of March 1 was 500 units whereas the actual demand is 450 units

- a) Forecast demand for the next week i. e March 8
- b) Assume the actual demand during the march is 505 units

Continue the forecasting assuming that subsequent demands were actually 516, 488, 467, 554 and 510 units

## Example:-

Forecast for the first week of March was 500 units whereas the actual demand is 450 units

a) Forecast demand for the next week in March

$$\begin{aligned}
 F_{t+1} &= F_t + \alpha (A_t - F_t) \\
 &= 500 + 0.1 (450 - 500) \\
 &= \underline{\underline{495}}
 \end{aligned}$$

Week	Demand	$(A_t)$	$(F_t)$ (old)	New forecast
March 1	450		500	$500 + 0.1 (450 - 500) = 495$
8	505		495	$495 + 0.1 (505 - 495) = 496$
15	516		496	$496 + 0.1 (516 - 496) = 498$
22	488		498	$498 + 0.1 (488 - 498) = 497$
April 1	467		497	$497 + 0.1 (467 - 497) = 494$
8	554		494	$494 + 0.1 (554 - 494) = 500$
15	510		500	$500 + 0.1 (510 - 500) = 501$

# Stationarity

stationary time series have no trend.

conditions

1. constant mean
2. Constant variance
3. An autocovariance that does not depend on time

# Auto Correlation



auto covariance<sub>h</sub>( $x_t$ )

$$= \text{cov}(x_t, x_{t-h})$$

auto correlation<sub>h</sub>( $x_t$ )

$$= \frac{\text{Auto cov}_h(x_t)}{\text{std}(x_t) \text{ std}(x_{t-h})}$$

# Auto Correlation Function

auto covariance  $\gamma_x(h)$

$$\gamma_x(h) = \text{cov}(x_t, x_{t-h})$$

$$\text{ACF} = \frac{\gamma_x(h)}{\gamma_x(0)} = \text{Cor}(x_t, x_{t-h})$$

# Stationarity

stationary time series have no trend.

conditions

1. constant mean
2. Constant variance
3. An autocovariance that does not depend on time

Differencing is the method of transforming a non-stationary time series into a stationary one.

The first differencing value is the difference b/w the current time period and the previous time period.

How many times - - - ?

# AR Model (auto regressive Model)

# AR model

$y_t$  depends on  $y_{t-1}, y_{t-2}, \dots$  etc

$$y_t = f(y_{t-1}, y_{t-2}, \dots, y_{t-p}, \epsilon_t)$$

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \dots + \beta_p y_{t-p} + \epsilon_t$$

constant

value of time series  $t-1$

AR(0)  $\Rightarrow y_t = \beta_0$  if  $\beta_0 = \beta_1 = \beta_2$  IT IS LIKE SMA

AR(1)  $\Rightarrow y_t = \beta_0 + \beta_1 y_{t-1}$

$\beta_0, \beta_1, \beta_2$  can be

like weights as well

# MA model (Moving averages model)

# MA model $\varphi$

$$y_t = f(\epsilon_t, \epsilon_{t-1}, \epsilon_{t-2}, \dots)$$

MA ( $q$ )

$$y_t = \theta_0 + \epsilon_t + \bar{\theta}_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q}$$

# ARMA model

# ARMA model - ARMA(p,q)

$$y_t = \beta_0 + \beta_1 y_{t-1} + \beta_2 y_{t-2} + \dots + \beta_p y_{t-p}$$

$$\theta_0 + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \dots + \theta_q \epsilon_{t-q}$$

✓ If  $p \neq 0$  and  $q=0$ , then AR (P)

✓ If  $p=0$  and  $q \neq 0$ , then MA (Q)

# ARIMA model (Autoregressive Integrated Moving Average Model)

# ARIMA Model

AR : Autoregression - dept relations  
b/w observation & lagged observations

I : Integrated → stationary

MA : moving average  
→ dependency relation b/w  
observation and residual error

## Parameters of the model

p : the number of lag observations  
ie lag order

d : degree of differencing

q : order of moving average

## Probability Distribution

The above probabilities can also be obtained by

$$P(x, y) = \frac{{}^3C_x {}^2C_y {}^4C_{2-x-y}}{{}^9C_2}, \text{ for } x = y = 0, 1, 2 \text{ and } 0 \leq x + y \leq 2$$

x	y		
	0	1	2
0	1/6	2/9	1/36
1	1/3	1/6	0
2	1/12	0	0

To obtain the probability associated with (1, 0), we have  ${}^3C_1 \times {}^2C_0 \times {}^4C_1 = 12$ .

Total number of ways in which two ball are drawn out of nine =  ${}^9C_2 = 36$ .

As these probabilities are equally likely, the probability of the event associated with (1, 0) is  $12/36 = 1/3$ .

## Probability Distribution

$$P(x,y) = \frac{^3c_x^2c_y^4c_{2-x-y}}{^9c_2}, \text{ for } x=y=0,1,2 \text{ and } 0 \leq x+y \leq 2$$

The distribution of both  $(X, Y)$  which occur jointly is called **Joint Discrete Probability Distribution**.

$P(x, y)$  is called **joint probability mass function (jpmf)**

## Joint Probability Distribution

If X and Y are discrete random variables, the joint probability distribution function is given by

$$F(x, y) = P(X \leq x, Y \leq y) = \sum_{s \leq x} \sum_{t \leq y} p(s, t)$$

X	Y		
	0	1	2
0	1/6	2/9	1/36
1	1/3	1/6	0
2	1/12	0	0

Find  $P(X \leq 1, Y \leq 1)$

$$P(X=0, Y=0) + P(X=0, Y=1) +$$

$$P(X=1, Y=0) + P(X=1, Y=1)$$

$$P(X \leq 1, Y \leq 1) = 16/18$$

## Marginal probability distribution

### Discrete Marginal Distributions

If X and Y are random variables, then the marginal distribution of X=x is given by

$$g(x) = \sum_y p(x, y) \quad P_y(x, y)$$

If X and Y are random variables, then the marginal distribution of Y=y is given by

$$h(y) = \sum_x p(x, y) \quad P_x(x, y)$$

# RANDOM VARIABLE

DISCRETE  
 $P(x,y)$

- (i)  $0 \leq P(x,y) \leq 1$
- (ii)  $\sum_x \sum_y P(x,y) = 1$

CONTINUOUS  
 $f(x,y)$

- (i)  $0 \leq f(x,y) \leq 1$
- (ii)  $\iint_{x,y} f(x,y) = 1$

## Marginal probability distribution

Find the marginal probability mass function of X and Y

X	Y		
	0	1	2
0	1/6	2/9	1/36
1	1/3	1/6	0
2	1/12	0	0

$$\sum_{x=0}^2 g(x) = \sum_{y=0}^2 h(y) = 1$$

X	Y			g(x)
	0	1	2	
0	1/6	2/9	1/36	15/36
1	1/3	1/6	0	6/36
2	1/12	0	0	1/12
h(y)	7/12	7/18	1/36	1

## Marginal probability distribution

**Summarizing the results:**

The marginal distribution of  $X=x$  is

x	0	1	2
$g(x)$	$7/12$	$7/18$	$1/36$

The marginal distribution of  $Y=y$  is

h	0	1	2
$h(y)$	$15/36$	$3/6$	$1/12$

## Conditional probability distribution

### Conditional Distributions

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, P(B) > 0$$

If X and Y are random variables, then the conditional distribution of X=x given Y=y is

$$P(X = x | Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p(x,y)}{h(y)}, h(y) > 0$$

Similarly, the conditional distribution of Y=y given X=x is

$$P(Y = y | X = x) = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{p(x,y)}{g(x)}, g(x) > 0$$

## CONDITIONAL DISTRIBUTIONS

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A, B)}{P(B)}$$

$$P(B|A) = \frac{P(A, B)}{P(A)}$$

$$P(X|Y) = \frac{P(X, Y)}{P_Y(X, Y)}$$

$$\rightarrow P(Y|X) = \frac{P(X, Y)}{P_X(X, Y)}$$



$$\begin{aligned} P(X, Y) &= P(X|Y) P(Y) \\ &= P(Y|X) P(X) \end{aligned}$$

# Conditional probability distribution

Find  $P(X=2 | Y=0)$  and  $P(Y=1 | X=1)$

y	X		
	0	1	2
0	1/6	1/3	1/12
1	2/9	1/6	0
2	1/36	0	0

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, P(B) > 0$$

## Conditional probability distribution

Find  $P(X=2 | Y=0)$  and  $P(Y=1 | X=1)$

X	Y			g(x)
	0	1	2	
0	1/6	2/9	1/36	15/3
1	1/3	1/6	0	6 3/6
2	1/12	0	0	1/12
h(y)	7/12	7/18	1/36	1

$$P(X=2 | Y=0) = \frac{P(X=2, Y=0)}{P(Y=0)} = \frac{\frac{1}{12}}{\frac{7}{12}} = \frac{1}{7}$$

$$P(Y=1 | X=1) = \frac{P(X=1, Y=1)}{P(X=1)} = \frac{\frac{1}{6}}{\frac{3}{6}} = \frac{1}{3}$$

$$P(Y=2 | X=0) = \frac{P(X=2, Y=0)}{P(Y=0)} = \frac{\frac{1}{36}}{\frac{15}{36}} = \frac{1}{15}$$

## Conditional probability distribution

**Example 2:** If the joint probability mass function of two discrete random variables is given below, obtain the marginal distribution of X and Y. Also find  $P(X=1|Y=0)$  &  $P(Y=1|X=1)$

X	Y		
	0	1	2
0	3/28	3/14	1/28
1	9/28	3/14	0
2	3/28	0	0

$$P(X=1 | Y=0) = \frac{P(X=1, Y=0)}{P(X=1)} = \frac{\frac{9}{28}}{\frac{15}{28}} = \frac{3}{5}$$

$$P(Y=2 | X=0) = \frac{P(X=0, Y=2)}{P(Y=0)} = \frac{\frac{1}{28}}{\frac{6}{28}} = \frac{1}{6}$$

## Example

The probability distribution of random variables X and Y is given by

$$P(X=0, Y=1) = \frac{1}{3} ; P(X=1, Y=-1) = \frac{1}{3}$$

$$P(X=1, Y=1) = ?$$

a) validate the distribution.

b) Marginal distributions of X and Y

c)  $P(X | Y=1)$

solution:

$y \backslash x$	0	1	
-1	0	$\frac{1}{3}$	$\frac{1}{3}$
1	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{2}{3}$
	$\frac{1}{3}$	$\frac{2}{3}$	1

a) valid

b) Marginal of  $x$ :

$x :$	0	1
$P(x)$	$\frac{1}{3}$	$\frac{2}{3}$

• marginal of  $y$  :

$y :$	-1	1
$P(y)$	$\frac{1}{3}$	$\frac{2}{3}$

$$\begin{aligned}
 c) P(x | y=1) &= \frac{P(x, y=1)}{P(y=1)} \\
 &= \frac{P(x=0, y=1) + P(x=1, y=1)}{P(y=1)} \\
 &= \frac{y_3 + y_3}{2/3} \\
 &= 1
 \end{aligned}$$

## Independence of Random variables

### Independence of random variables

If  $X$  and  $Y$  are two random variables, with joint probability distributions  $p(x, y)$  and the marginal distributions  $g(x)$  and  $h(y)$  respectively, then they are called as statistically independent if and only if

$$p(x, y) = g(x)h(y) \text{ for all } (x, y) \text{ within their range.}$$

**Note:  $X$  and  $Y$  may be discrete or continuous.**

A, B ARE INDEPENDENT IF:

$$P(A \cap B) = P(A) P(B)$$

$$P(X, Y) = P(X) P(Y)$$

WHERE X & Y ARE  
RANDOM VARS

$$P(X | Y) = \frac{P(X, Y)}{P(Y)} = \frac{\cancel{P(X) P(Y)}}{\cancel{P(Y)}} = P(X)$$

# Joint Continuous probability distribution

Let  $X$  and  $Y$  be continuous random variables. A joint Probability density function  $f(x, y)$  for these two variables is a function satisfying  $f(x, y) \geq 0$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$$

Also,

$$P(a \leq X \leq b, c \leq Y \leq d) = \int_a^b \int_c^d f(x, y) dy dx$$

## Joint Continuous probability distribution

### Marginal densities of continuous random variables

Let  $f(x, y)$  be the joint density function of the continuous random variables  $X$  and  $Y$ , then the marginal densities of  $X$  and  $Y$  will be obtained by integrating over the range of  $Y$  for  $X$  and over the range of  $X$  for  $Y$ , i.e.,

$$g(x) = \int_{-\infty}^{\infty} f(x, y) dy \quad \text{and} \quad h(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

# Statistical Methods for Data Science

## Probability Distribution → Joint Continuous probability distribution

---

Two continuous random variables  $X$  and  $Y$  are said to be **independent** if for every pair of  $x$  and  $y$  values

$$f(x, y) = f_X(x) \cdot f_Y(y)$$

If it is not satisfied, then  $X$  and  $Y$  are said to be dependent.

# Statistical Methods for Data Science

Probability Distribution → Joint Continuous probability distribution

---

## Conditional Distributions

Let X and Y be two continuous rv's with joint pdf  $f(x, y)$  and marginal probability distribution of X is  $f_X(x)$ .

Then for any X value of  $x$  for which  $f_X(x) > 0$ , the conditional probability density function of

Y for given X is 
$$f_{Y|X}(y|x) = \frac{f(x,y)}{f_X(x)}$$

# Statistical Methods for Data Science

Probability Distribution → Joint Continuous probability distribution

## Distribution Functions

Let  $(X, Y)$  be a two dimensional random variable then their joint distribution function is given by:

In case of continuous rv's

- $F_{XY}(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^x \int_{-\infty}^y f(x, y) dx dy$

In case of discrete rv's

- $F_{XY}(x, y) = P(X \leq x, Y \leq y) = \sum_{X \leq x} \sum_{Y \leq y} p(x, y)$

# Statistical Methods for Data Science

Probability Distribution → Joint Continuous probability distribution

---

## Marginal density functions

In case of continuous rv's

- $F_X(x) = \int_{-\infty}^x \int_{-\infty}^{\infty} f(x, y) dy dx$
- $F_Y(y) = \int_{-\infty}^y \int_{-\infty}^{\infty} f(x, y) dx dy$

In case of discrete rv's

- $F_X(x) = \sum_{X \leq x, \forall y} P(X \leq x, Y = y)$
- $F_Y(y) = \sum_{Y \leq y, \forall x} P(X = x, Y \leq y)$

# Statistical Methods for Data Science

Probability Distribution → Joint Continuous probability distribution

---

## Covariance

- $\text{Cov}(X, Y) = E(XY) - E(X) \cdot E(Y)$  where
- $E(X) = \int_{-\infty}^{\infty} x f_X(x) dx$
- $E(Y) = \int_{-\infty}^{\infty} y f_Y(y) dy$
- $E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f(x, y) dx dy$
- $V(X) = E(X^2) - [E(X)]^2$
- $V(Y) = E(Y^2) - [E(Y)]^2$

# Example

If  $x$  and  $y$  having joint p.d.f  
 $f(x,y) = \begin{cases} x+y, & 0 < x, y < 1 \\ 0, & \text{otherwise} \end{cases}$

- 1) Find marginal p.d.f
- 2) Are they indept?

If  $x$  and  $y$  having joint p.d.f

$$f(x,y) = \begin{cases} x+y, & 0 < x, y < 1 \\ 0, & \text{otherwise} \end{cases}$$

Marginal of  $x$ ,  $f_x(x,y) = \int_0^1 f(x,y) dy$

$$= \int_0^1 (x+y) dy = \left[ xy + \frac{y^2}{2} \right]_{y=0}^1$$

$$= \left( x + \frac{1}{2} \right) = \frac{2x+1}{2}$$

If  $x$  and  $y$  having joint p.d.f

$$f(x,y) = \begin{cases} x+y, & 0 < x, y < 1 \\ 0, & \text{otherwise} \end{cases}$$

Marginal of  $y$ ,  $f_y(y) = \int_0^1 f(x,y) dx$

$$= \int_0^1 (x+y) dx = \left( \frac{x^2}{2} + xy \right)_{x=0}^1$$

$$= \frac{2y+1}{2}$$

$$f_x(x, y) = \frac{2x+1}{2}$$

$$f_y(x, y) = \frac{2y+1}{2}$$

$$f_x \cdot f_y = \left( \frac{2x+1}{2} \right) \left( \frac{2y+1}{2} \right)$$

$$\neq x + y$$

$$\text{i.e. } f(x, y)$$

i.e.  $x, y$  are not independent.

**CHECK**

$$f(x, y) = f(x) \cdot f(y)$$

## Double integral

$$\int_{y=0}^1 \int_{x=0}^2 xy \, dx \, dy$$

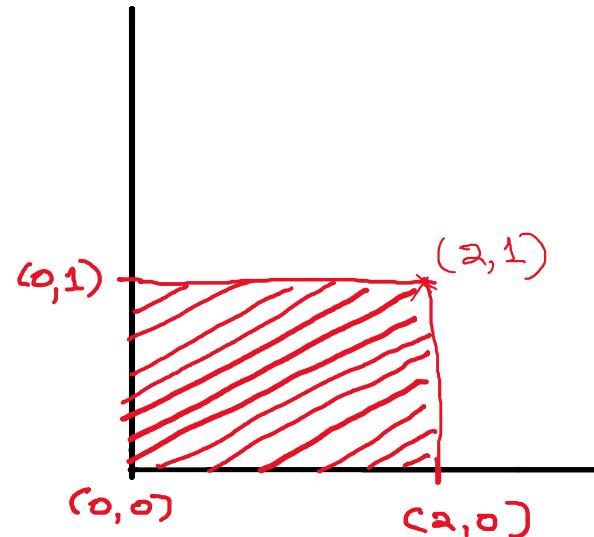
Region is bounded by

$$x = 0, \, x = 2$$

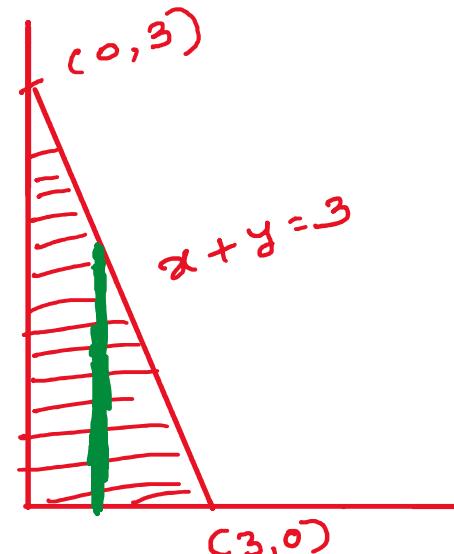
$$y = 0, \, y = 1$$

$$\Rightarrow \int_{y=0}^1 \left( \frac{x^2}{2} \right)^2 y \, dy = 2 \int_0^1 y \, dy \\ = 2 \left[ \frac{y^2}{2} \right]_0^1 = 1$$

$$\Rightarrow \int_{x=0}^2 x \left( \frac{y^2}{2} \right)_0^1 \, dx = \frac{1}{2} \left( \frac{x^2}{2} \right)_0^2 \\ = \frac{1}{2} \left( \frac{4}{2} \right) = 1$$



$$\begin{aligned}
 & \iint_S xy \, dx \, dy : S \text{ is } x + y \leq 3 \\
 & x : 0 \rightarrow 3 \quad \text{or} \quad y : 0 \rightarrow 3 \\
 & y : 0 \rightarrow 3-x \quad x : 0 \rightarrow 3-y \\
 & \int_{x=0}^3 \int_{y=0}^{3-x} xy \, dy \, dx \\
 & = \int_{x=0}^3 x \left( \frac{y^2}{2} \right) \Big|_0^{3-x} \, dx \\
 & = \int_{x=0}^3 \frac{x}{2} (3-x)^2 \, dx = \int_{x=0}^3 \frac{x}{2} (9+x^2-6x) \, dx \\
 & = \frac{1}{2} \int_{x=0}^3 (9x + x^3 - 6x^2) \, dx = \frac{1}{2} \left[ \frac{9x^2}{2} + \frac{x^4}{4} - \frac{6x^3}{3} \right]_0^3 \\
 & = \frac{1}{2} \left[ \frac{81}{2} + \frac{81}{4} - \frac{6 \times 27}{2} \right]
 \end{aligned}$$



## Example :

If  $f(x, y) = \begin{cases} \frac{2}{5}x(2x + 3y), & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0, & \text{Elsewhere} \end{cases}$

(i) Find  $P[X \leq 2/3, Y \leq 1/4]$ ?

## solution

$$P(x \leq \frac{2}{3}, y \leq \frac{1}{4})$$

$$x: 0 \longrightarrow 1$$

$$= \int_{x=0}^{\frac{2}{3}} \int_{y=0}^{\frac{1}{4}} \frac{2}{5}x(2x+3y) dy dx$$

$$y: 0 \longrightarrow 1$$

$$= \int \int \left( \frac{4}{5}x^2 + \frac{6}{5}xy \right) dy dx$$

$$= \frac{4}{5} \left( \frac{x^3}{3} \right) (y) + \frac{6}{5} \left( \frac{x^2}{2} \right) \left( \frac{y^2}{2} \right)$$

$$= \frac{4}{5} \left( \frac{x^3}{3} \right)_0^{\frac{2}{3}} (y)_0^{\frac{1}{4}} + \frac{6}{5} \left( \frac{x^2}{2} \right)_0^{\frac{2}{3}} \left( \frac{y^2}{2} \right)_0^{\frac{1}{4}}$$

$$\iint_S xy \, dx \, dy : S \text{ is } x + y \leq 3$$

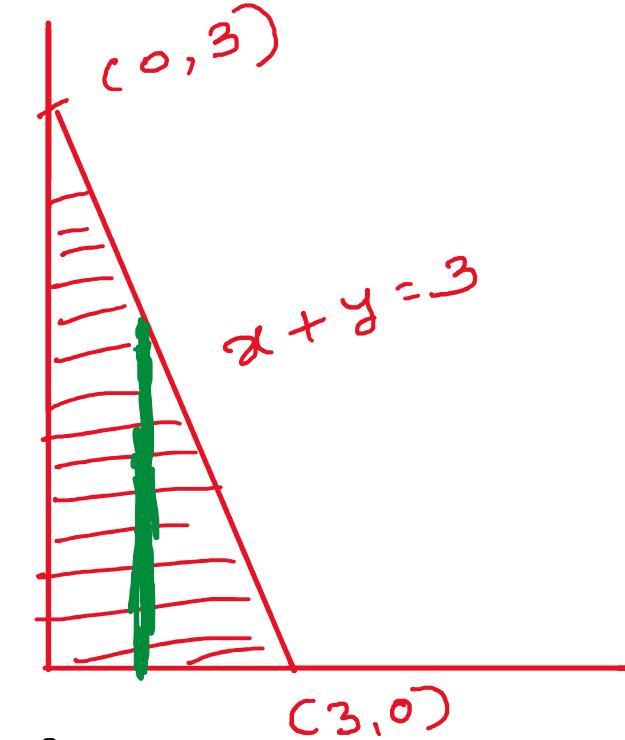
$$x : 0 \rightarrow 3$$

or  $y : 0 \rightarrow 3$

$$x : 0 \rightarrow 3-y$$

$$\int_{x=0}^3 \int_{y=0}^{3-x} xy \, dy \, dx$$

$$= \int_{x=0}^3 x \left( \frac{y^2}{2} \right) \Big|_0^{3-x} \, dx$$



$$= \int_{x=0}^3 \frac{x}{2} (3-x)^2 \, dx = \int_{x=0}^3 \frac{x}{2} (9 + x^2 - 6x) \, dx$$

$$= \frac{1}{2} \int_{x=0}^3 (9x + x^3 - 6x^2) \, dx = \frac{1}{2} \left[ \frac{9x^2}{2} + \frac{x^4}{4} - \frac{6x^3}{3} \right]_0^3$$

$$= \frac{1}{2} \left[ \frac{81}{2} + \frac{81}{4} - \frac{6 \cdot 27}{2} \right]$$

# Example

a)  $P(x_1 < 1, x_2 < 1)$

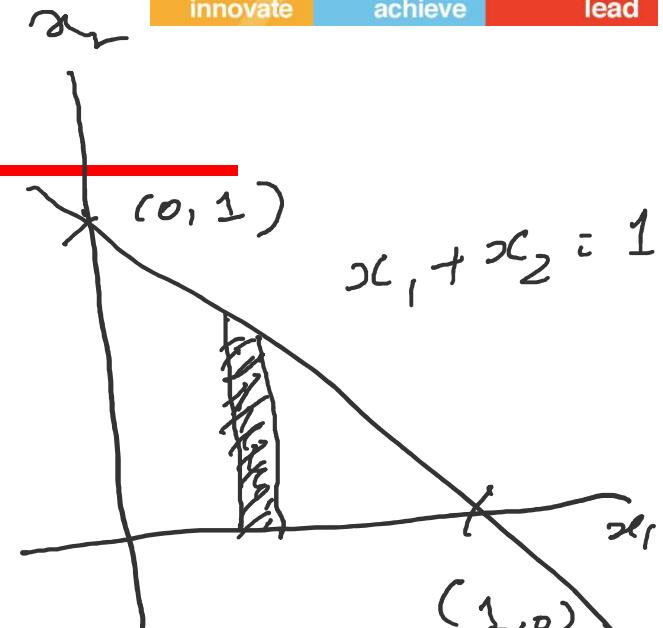
$$\begin{aligned}
 &= \int_{x_1=0}^1 \int_{x_2=0}^1 x_1 x_2 \, dx_1 \, dx_2 \\
 &= \int_{x_1=0}^1 dx_1 \left[ \frac{x_2^2}{2} \right]_0^1 = \frac{1}{2} \left[ \frac{x_1^2}{2} \right]_0^1 \\
 &= \frac{1}{4}
 \end{aligned}$$

b)  $P(x_1 + x_2 < 1)$

$$= \int_{x_1=0}^1 \left[ x_1 x_2 \right]_{x_2=0}^{1-x_1} dx_2 dx_1$$

$$= \int_{x_1=0}^1 x_1 \left[ \frac{x_2^2}{2} \right]_0^{1-x_1} dx_1$$

$$\begin{aligned} &= \int_{x_1=0}^1 \frac{1}{2} x_1 (1-x_1)^2 dx_1 = \frac{1}{2} \int_{x_1=0}^1 x_1 (1+x_1^2 - 2x_1) dx_1 \\ &= \frac{1}{2} \left[ \frac{x_1^2}{2} + \frac{x_1^4}{4} - \frac{2x_1^3}{3} \right]_0^1 \\ &= \frac{1}{2} \left[ \frac{1}{2} + \frac{1}{4} - \frac{2}{3} \right] \end{aligned}$$



$$x_1 : 0 \rightarrow 1$$

$$x_2 : 0 \rightarrow 1 - x_1$$

## Example

Joint P.d.f of  $x$  and  $y$  is given by

$$f(x, y) = 4xy e^{-(x^2 + y^2)} ; x > 0, y > 0.$$

- whether  $x$  and  $y$  are independent?
- conditional density of  $x$  given  $y=y$

# Solution

Two random variables  $x$  and  $y$  are independent if

$$f(x, y) = f_x(x, y) \cdot f_y(x, y)$$

$\downarrow$                      $\downarrow$                      $\downarrow$   
Joint              Marginal      marginal  
p. d. f            of  $x$           of  $y$

marginal of x :-

$$f_x(x, y) = \int_y f(x, y) dy$$

$$= \int_y 4xy e^{-(x^2 + y^2)} dy = 4x e^{-x^2} \int_{y=0}^{\infty} y e^{-y^2} dy$$

$$y^2 = t$$

$$= 4x e^{-x^2} \cdot \frac{1}{2} \int_0^t e^{-t} dt$$

$$2y dy = dt$$

$$y dy = \frac{1}{2} dt$$

$$= 2x e^{-x^2} \left[ \frac{-e^{-t}}{-1} \right]_0^\infty$$

$$= -2x e^{-x^2} (-e^\infty - e^0)$$

$$= 2x e^{-x^2}$$

marginal of Y :-

$$f_y(x, y) = \int_{-\infty}^{\infty} f(x, y) dx$$

$$= \int_{-\infty}^{\infty} 4xy e^{-(x^2+y^2)} dx = 4y e^{-y^2} \int_{y=0}^{\infty} x e^{-x^2} dy$$

$$= 4y e^{-y^2} \cdot \frac{1}{2} \int_0^{\infty} e^{-t} dt$$

$$= 2y e^{-y^2} \left[ \frac{-e^{-t}}{-1} \right]_0^{\infty}$$

$$= -2y e^{-y^2} (-e^{\infty} - e^0)$$

$$= 2y e^{-y^2}$$

$$2x dx = dt$$

$$x dx = \frac{1}{2} dt$$

$$f_x(x, y) \cdot f_y(x, y)$$

$$= (2x e^{-x^2}) (2y e^{-y^2})$$

$$= 4xy e^{-(x^2+y^2)}$$

$$= f(x, y)$$

i.e.  $x$  and  $y$  are independent.

## EXAMPLE

### K VALUE EXAMPLE

The joint density function is given as

$$f(x, y) = \begin{cases} Kxy, & 0 < x < 4, 1 < y < 5 \\ 0, & \text{otherwise} \end{cases}$$

then find

- a) K value
- b)  $P(x \geq 3, y \leq 2)$
- c)  $P(1 < x < 2, 2 < y < 3)$
- d) marginal density functions of  
x and y

# Solution

$$a) \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$$

$$\int_{x=0}^4 \int_{y=1}^5 Kxy dy dx = 1$$

$$\int_{x=0}^4 K \left[ x \frac{y^2}{2} \right]_{y=1}^5 dx = 1$$

$$\therefore \frac{K}{2} \int_{x=0}^4 x (25 - 1) dx = 1$$

$$\therefore \frac{K}{2} \times 24 \times \left[ \frac{x^2}{2} \right]_{x=0}^4 = 1$$

$$\Rightarrow K = \frac{1}{96}$$

b)  $P(x \geq 3, y \leq 2)$

$$= \int_{x=3}^4 \int_{y=1}^2 kxy \, dy \, dx$$

$$= \frac{1}{96} \left[ \frac{x^2}{2} \right]_3^4 \left[ \frac{y^2}{2} \right]_1^2$$

$$= \frac{1}{96} \left[ \frac{16 - 9}{2} \right] \left[ \frac{4 - 1}{2} \right]$$

$$= \frac{1}{96} \times \frac{1}{2} \times \frac{3}{2}$$

$$= \frac{7}{128}$$

c)  $P(1 < x < 2, 2 < y < 3)$

$$= \int_{x=1}^2 \int_{y=2}^3 kxy \, dy \, dx$$

$$= \frac{1}{96} \left[ \frac{x^2}{2} \right]_{x=1}^2 \left[ \frac{y^2}{2} \right]_{y=2}^3$$

$$= \frac{1}{96} \left[ \frac{4 - 1}{2} \right] \left[ \frac{9 - 4}{2} \right]$$

$$= \frac{1}{96} \times \frac{3}{2} \times \frac{5}{2}$$

= 31

$$\frac{5}{128}$$

d) marginal distribution of  $X$ :  $f_x(x, y)$

$$\int_{-\infty}^{\infty} f(x, y) dy = \int_{y=1}^{5} kxy dy$$

$$= \frac{1}{96} x \left[ \frac{y^2}{2} \right]_1^5 = \frac{x}{96} \frac{[25-1]}{2} = \frac{x}{96} \times \frac{24}{2} = \boxed{\frac{x}{8}}$$

e) marginal distribution of  $Y$  :-  $f_y(x, y)$

$$= \int_0^4 kxy dx = \frac{y}{96} \left( \frac{x^2}{2} \right)_0^4$$

$$= \frac{1}{12} y$$

$$f_y(x, y) = \frac{1}{12} y, \quad 1 < y < 5$$



# Thanks

# Advanced statistics methods

---

Akhil Sudhakaran

---

2021MT12054

---

---

---

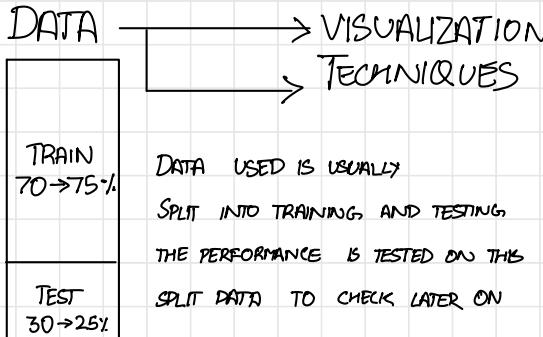


# COURSE OVERVIEW

## EVALUATIVE COMPONENTS

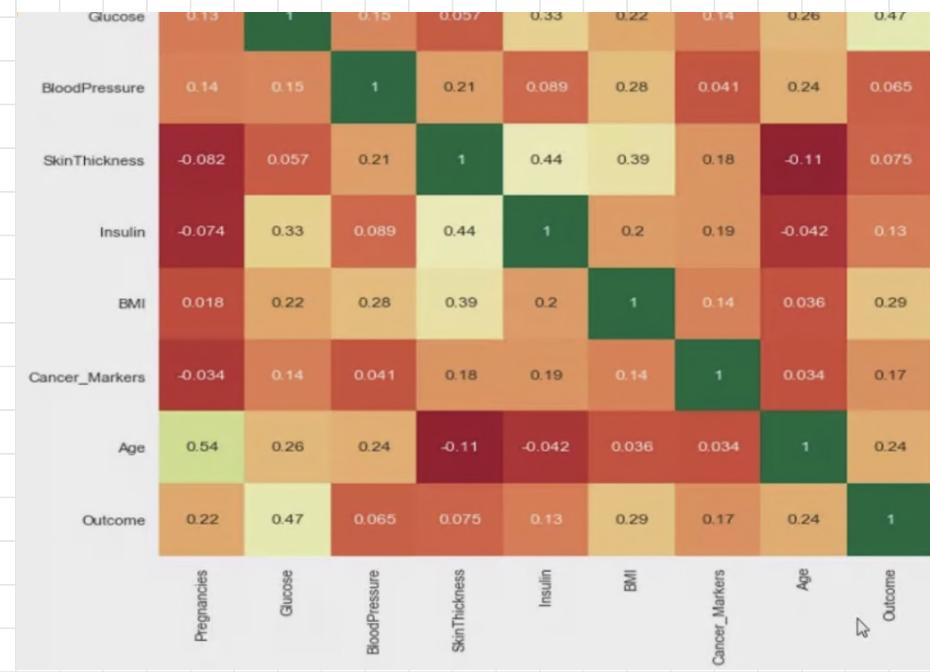
- **Quiz -1** : 5 Marks
- **Mid Semester Exam** : 30 Marks
- **Assignment** : 10 Marks
- **Quiz – 2:** 5 Marks
- **Comprehensive Exam:** 50 Marks

# VISUALIZATIONS



- PIE CHARTS
  - SCATTER PLOTS
  - HISTOGRAMS (DATA IS BALANCED/NOT)
  - HEAT MAPS (CORRELATIONS)
  - BAR CHARTS
  - BOX PLOTS (SPREAD OF DATA/OUTLIER)

## HEAT MAPS



HEAT MAPS SHOW THE CORRELATION BETWEEN TWO VARIABLES. WHEN THE SAME VARIABLE CORRELATES TO ITSELF THEN THE VALUE IS 1.00. IN THE EXAMPLE SHOWN ABOVE WE SEE HOW SEVERAL VARIABLES RELATE TO OTHER VARIABLES.

HEAT MAPS HELP IN DIMENSIONALITY REDUCTION.

LET  $y = f(x_1, x_2, x_3)$

	y
$x_1$	0.2
$x_2$	0.6
$x_3$	0.8

HERE SINCE  $x_1$  CONTRIBUTES LESS TO y WE CAN CONSIDER REMOVING IT FROM THE TRAINING PROCESS

# STATISTICAL TOOLS

- MEAN = AVERAGE = SUM / FREQUENCY  $\frac{\sum x}{n}$  POPULATION MEAN
  - MEDIAN = MIDDLE VALUE IN A SORTED LIST  $\frac{\sum x}{n-1}$  SAMPLE MEAN
  - MODE
  - STANDARD DEVIATION
  - VARIANCE =  $\frac{\sum (x-\mu)^2}{n}$  OR  $\frac{\sum (x-\mu)^2}{n-1}$
- POPULATION VARIANCE SAMPLE VARIANCE

OUTLIERS HOW THEY AFFECT TOOLS  
 MEAN → AFFECTED  
 MEDIAN → NOT AFFECTED

HELPS IN REPLACING MISSING VALUES

## STATISTICAL SUMMARY:

max	1650.000000	59.000000	63.400000	68.000000	18.957000	8.1
-----	-------------	-----------	-----------	-----------	-----------	-----

- 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

# DEGREES OF FREEDOM

Box Plot



# CONDITIONAL PROBABILITY

$$P(A|E) = \frac{P(A \cap E)}{P(E)}$$
$$= \frac{\sum_{x=A,B,C} P(E|x) P(x)}{\sum_{x=A,B,C} P(E|x) P(x)}$$

BAYES THEOREM

Q  
||

$$\begin{aligned} P(S) &= P(A \cap S) + P(B \cap S) + P(C \cap S) \\ &= P(S|A)P(A) + P(S|B)P(B) + P(S|C)P(C) \\ &= (0.01)(0.9) + (0.02)(0.2) + (0.05)0.1 \end{aligned}$$

$$P(A|S) = \frac{(0.01)0.9}{0.016} = 0.5625$$

THE DENOMINATOR IS USUALLY NOT TAKEN INTO CONSIDERATION TO REDUCE COST  
MAX  $P(D|h_x) P(h_x)$  ] MAP HYPOTHESIS

$$P(A) = 6/18$$

$$P(B) = 5/18$$

$$P(C) = 7/18$$

$$P(S|A) = 0.54$$

$$P(S|B) = 0.60$$

$$P(S|C) = 0.55$$

$$\begin{aligned} P(A|S) &= \frac{0.54 \times 6}{0.54 \times 6 + 0.60 \times 5 + 0.55 \times 7} \\ &= \frac{3.24}{10.09} = 0.3211 \end{aligned}$$

# RANDOM VARIABLE

2	1	2	4	8	16
P(x)	0.05	0.10	0.35	0.40	0.10

✓ 6.45

compute the following  $\rightarrow$  mean =

$$V(x) = E(x^2) - [E(x)]^2$$
$$\sum x^2 P(x) = 1(0.05) + 2^2(0.10) + 4^2(0.35) + 8^2(0.40) + 16^2(0.10)$$

d)  $V(x)$  using the shortcut formula

Q  
//

$$(a) \int_{-1}^1 f(x) dx = 1$$

$$\Rightarrow \int_{-1}^1 Cx^2 = 1 \Rightarrow \left[ \frac{Cx^3}{3} \right]_{-1}^1$$

$$\Rightarrow \frac{C}{3} + \frac{C}{3} = 1$$

$$\Rightarrow C = 3/2$$

$$(b) \text{ MEAN } \mu E(x) = \int_{-1}^1 x f(x) dx \\ = \int_{-1}^1 \frac{3}{2} x^2 dx = \frac{3}{2} \frac{x^4}{4} \Big|_{-1}^1 \\ = \frac{3}{8} [1 - 1] = 0$$

$$V = E(x^2) - [E(x)]^2$$

$$= \int x^2 \frac{3}{2} x^2 dx$$

$$= \frac{3}{2} \frac{x^5}{5}$$

$$= \frac{3}{10} [1 - (-1)^5] = \frac{3}{5}$$

# DISTRIBUTIONS

$$P(X) = {}^n C_x p^x q^{n-x} ; \quad x = 0, 1, 2, \dots, n$$

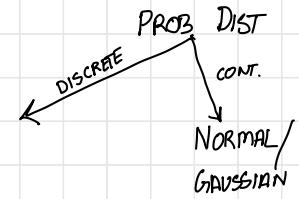
BINOMIAL

$$P(X) = \frac{e^{-\lambda} \lambda^x}{x!} ; \quad x = 0, 1, 2, \dots, \infty$$

POISSON

$$P(X) = p^x q^{1-x} ; \quad x = 0, 1$$

BERNOULLI



BERNOULLI DIST.

$\begin{cases} 1, & \text{if } X=1 \\ 0, & \text{elsewhere} \end{cases}$

~~Q~~

$$P(X=0) = {}^6C_0 (0.5)^0 (0.5)^6$$

$$P(X=1) = {}^6C_1 (0.5)^1 (0.5)^5$$

$$P(X=2) = {}^6C_2 (0.5)^2 (0.5)^4$$

~~Q~~

$$p = 0.65 \quad q = 0.35$$

$$n = 10$$

$$P(X \geq 7) = P(7) + P(8) + P(9) + P(10)$$

$$= 10c_7 (0.65)^7 (0.35)^3 + 10c_8 (0.65)^8 (0.35)^2$$

$$+ 10c_9 (0.65)^9 (0.35)^1 + 10c_{10} (0.65)^{10} (0.35)^0$$

$$= 0.515$$

$$\begin{aligned} p &= 0.53 \quad q = 0.47 \\ n &= 5 \end{aligned}$$

$$P(x=0)$$

$$P(1 < x < 4)$$

$$1 - P(x \leq 3)$$

## MEAN & VARIANCE OF BINOMIAL DIST.

$x: 0, 1, \dots, n$

$$= \sum_{x=0}^n \frac{n!}{(x-1)! (n-x)!} p^x q^{n-x}$$

$$= np \sum_{x=1}^n \frac{(n-1)!}{(x-1)! [(n-1)-(x-1)]!} p^{x-1} q^{(n-1)-(x-1)}$$

$$\begin{aligned} &= np \left( q + p \right)^{n-1} \underbrace{(q+p)^n}_{(q+p)^n = n \cdot q^0 \cdot p^0 + n \cdot q^1 \cdot p^1 + \dots + n \cdot q^{n-1} \cdot p^{n-1}} \\ &= np \end{aligned}$$

$$\begin{aligned}
 &= n^2 p^2 - np^2 + np - n^2 p^2 \\
 &= np(1-p) = (npq)
 \end{aligned}$$

## Poisson's Dist.

$$P(x) = \frac{e^{-\lambda} \lambda^x}{x!} \quad x \rightarrow \text{DISCRETE}$$

MEAN & VARIANCE =  $\lambda$

UNDER SOME CONDITIONS A BINOMIAL DIST. BECOMES A POISSON'S (CONSIDER  $\mu = np$ , WHEN  $n$  IS VERY LARGE AND  $p$  IS VERY SMALL  $\Rightarrow \mu = np = \lambda$ )

# NORMAL DISTRIBUTION



**Example :-** (from T<sub>i</sub>)

The time it takes a driver to react to the brake lights on a decelerating vehicle is critical in helping to avoid rear-end collisions. It is suggested that reaction time for an in traffic response to a brake signal from a standard brake lights can be modeled with a normal distribution having mean 1.25 and S.D of 0.46 sec. What is the probability that reaction time is between 1.00 sec and 1.75 sec?

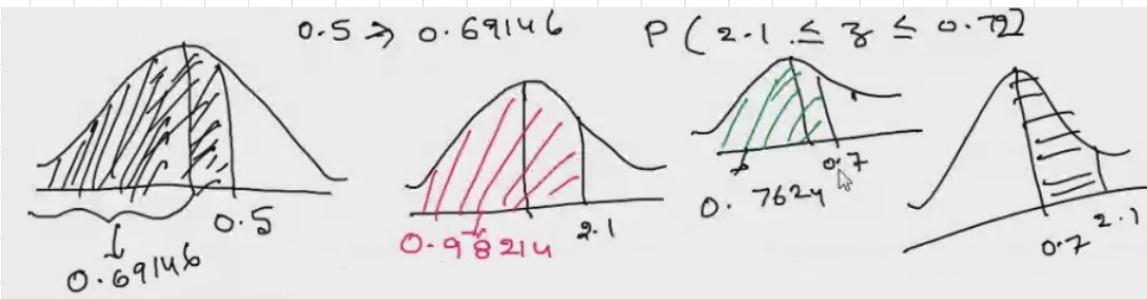
$$\text{ie } P(1.00 \leq x \leq 1.75)$$

$$\checkmark \text{when } x=1, \text{ Z: } \frac{x-\mu}{\sigma} = \frac{1-1.25}{0.46} = -0.54$$

$$\checkmark \text{when } x=1.75, \text{ Z: } \frac{1.75-1.25}{0.46} = 1.09$$

$$\Rightarrow P(-0.54 \leq Z \leq 1.09) \Rightarrow F(1.09) - F(-0.54)$$

NOW LOOK AT THE NORMAL DIST. TABLE.



$$F(1.09) = 0.86214$$

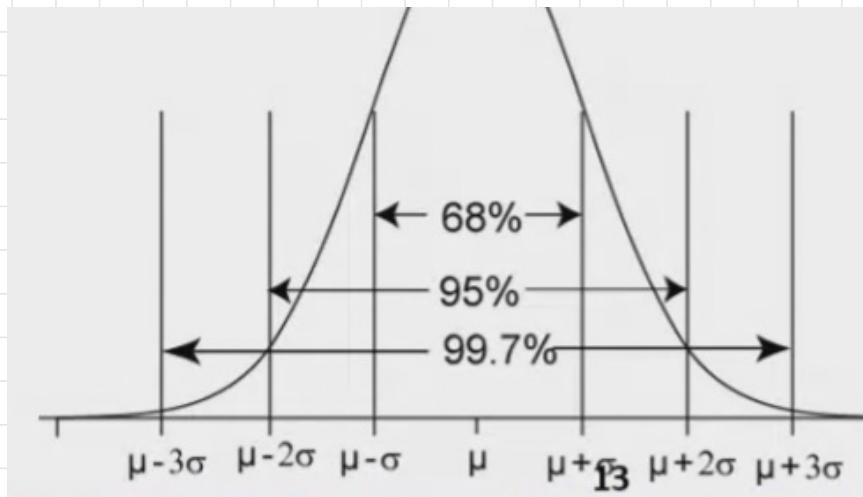
$$F(-0.54) = 1 - F(0.54) = 1 - 0.70540$$

# PARAMETERS $\mu$ AND $\sigma$

[Avg OF DATA]

[SPREAD OF DATA]

68-95-99.7% RULE



Q  
 $F(1.25) = 0.89435$

AREA TO THE RIGHT =  $1 - F(1.25) = 0.10565$

Q

the mean.

3. The total area under the curve is equal to one.
4. The normal curve approaches, but never touches the  $x$ -axis as it extends farther and farther away from the mean.

Q

B

B



$$\begin{aligned}\mu_A &= 5 \\ \sigma_p &= [1-5-9] \\ &= 4\end{aligned}$$

$$\begin{aligned}\mu_B &= 9 \\ \sigma_B &= [3-9-15] \\ &= 6\end{aligned}$$

Q

We NEED  $P(X < 90)$

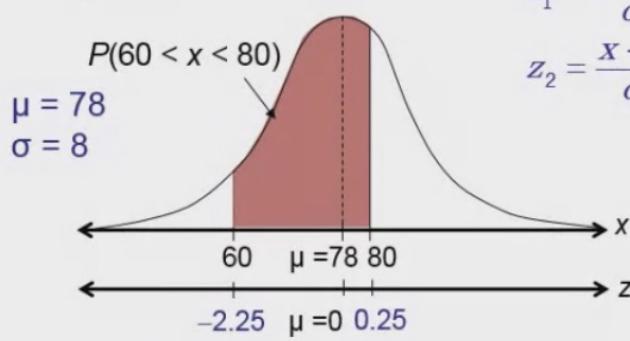
$$Z = \frac{X - \mu}{\sigma} = \frac{90 - 78}{8} = 1.5$$

$$P(X < 90) = P(Z < 1.5)$$

$$P(Z < 1.5) = F(1.5) = 0.9332$$

Q

Q



$$z_1 = \frac{x - \mu}{\sigma} = \frac{60 - 78}{8} =$$
$$z_2 = \frac{x - \mu}{\sigma} = \frac{80 - 78}{8} =$$

$$\begin{aligned} P(60 < x < 80) &= P(-2.25 < z < 0.25) = P(z < 0.25) - P(z < -2.25) \\ &= 0.5987 - 0.0122 = 0.5865 \end{aligned}$$

## PRACTICE QUESTIONS

~~Q~~

$$P(X < 35) = 0.07 = F(z_1) \Rightarrow z_1 = -1.48$$

$$P(X < 63) = 0.89 = F(z_2) \Rightarrow z_2 = 1.23$$

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

$$z_1 = \frac{35 - \mu}{\sigma} \quad z_2 = \frac{63 - \mu}{\sigma}$$

$$-1.48\sigma = 35 - \mu \quad 1.23\sigma = 63 - \mu$$

$$2.69\sigma = 28$$

$$\sigma = \frac{28}{2.69} = 10.408$$

$$\mu = 63 - 1.23\sigma$$

$$= 50.198$$

Q

THIS IS A DISCRETE DIST., SO IT'S A BINOMIAL DIST BUT HERE WE USE NORMAL DIST TO APPROXIMATE VALUES FOR VERY LARGE N

$$p = 0.25 \quad q = 0.75$$

$$\mu = np = 50 \times 0.25$$
$$\sigma = \sqrt{npq} = \sqrt{50 \times 0.25 \times 0.75}$$

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

EXAMPLE:

Based on the normal approximation,

$$P(X \leq 2) = P\left(\frac{X - 5}{\sqrt{50(0.1)(0.9)}} < \frac{2.5 - 5}{\sqrt{50(0.1)(0.9)}}\right) = P(Z < -1.18) = 0.12$$

# SAMPLING DISTRIBUTION & ESTIMATION

POPULATION → UNDERSTAND → SAMPLE

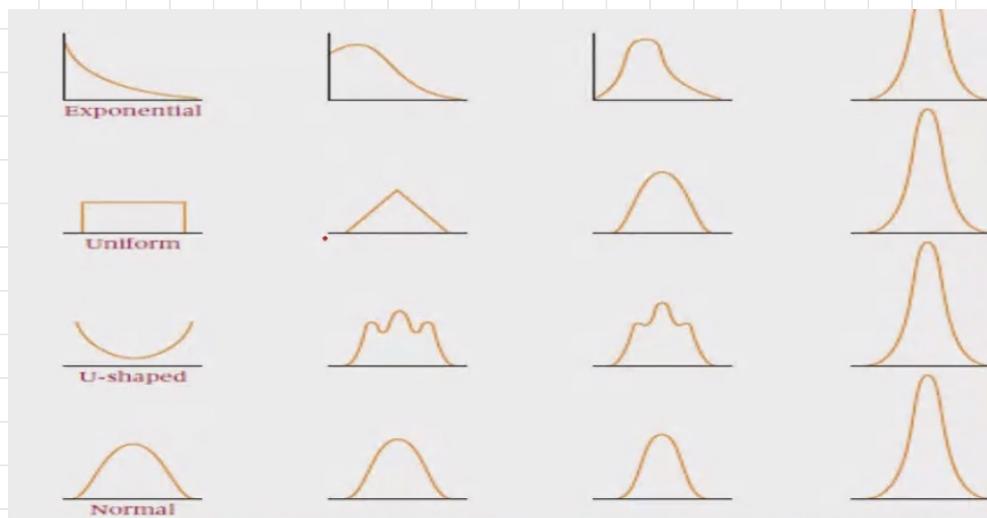
WE CAN CREATE MULTIPLE SAMPLES AND

1661	2495	1000	2497	1005	791	2090	2637	1327	167
1680	2858	795	2495	2496	2501	1160	1480	1860	249
2090	2840	2490	2640	659	827	2646	2638	2643	868
1327	1866	1861	2486	2865	3011	2494	1489	1865	285
2840	2499	2093	2660	1165	2600	2085	2640	2998	186
2956	2495	2865	1865	3000	3019	1670	2858	2642	168
3038	3000	1313	596	656	3240	590	2501	2485	301
2092	1679	3024	2497	2825	2630	2070	2900	1861	263
2495	2637	2497	1159	2640	3050	870	2896	2500	263
926	2860	1481	875	2482	1860	2086	934	3200	249

WHEN TAKING THE SAMPLES GIVEN IN SLIDES WE GET THE FOLLOWING METRICS

## SHAPES OF THE DIST. OF SAMPLE MEANS

$n \Rightarrow$  SAMPLE SIZE



# CENTRAL LIMIT THEOREM CLT

Y

$\mu = 85$        $n = 40 \rightarrow$  LARGE SAMPLE AND FOLLOWS NORMAL DIST.

$$\sigma = 9$$

$$P(\bar{x} \geq 87)$$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \Rightarrow \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{87 - 85}{9/\sqrt{40}} = 1.41$$

$$\begin{aligned} P(\bar{x} \geq 87) &= P(z > 1.41) \\ &= 1 - F(1.41) \\ &= 1 - 0.9207 = 0.0793 \end{aligned}$$

# SAMPLING IN A FINITE POPULATION

Q  
//

$$N = 350 \quad \mu = 37.6 \quad \sigma = 8.3$$

$$n = 45 \quad \text{TBF} \quad P(\bar{x} < 40)$$

$$\begin{aligned} z &= \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}} = \frac{40 - 37.6}{\frac{8.3}{\sqrt{45}} \sqrt{\frac{350-45}{350-1}}} \\ &= 2.07 \end{aligned}$$

$$P(\bar{x} < 40)$$

$$= P(z \leq 2.07)$$

$$= F(2.07) = 0.9808$$

# FORMS OF STATISTICAL INFERENCE :

(MOSTLY USED)

WHEN  $Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$

HERE  $\bar{x} - \mu$  IS CALLED AS SAMPLING ERROR.

WHEN  $n$  IS SMALL ( $< 30$ ) THE SAMPLE FOLLOWS A  $t$  DIST.



population mean number of minutes called per residential user per month, from the sample of 85 bills it was determined that the sample mean is 510 minutes.

- Suppose past history and similar studies indicate that the population standard deviation is 46 minutes.
- Determine a 95% confidence interval.

$$\bar{x} = 510 \quad n=85 \quad \sigma = 46$$

$$510 - \frac{46}{\sqrt{85}} \leq \mu \leq 510 + 1.96 \frac{46}{\sqrt{85}}$$
$$\Rightarrow 500.22 \leq \mu \leq 519.78$$

IN A BINOMIAL DIST

$$\mu = np \quad \& \quad \sigma^2 = npq$$

so

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{x - np}{\sqrt{npq} / \sqrt{n}}$$

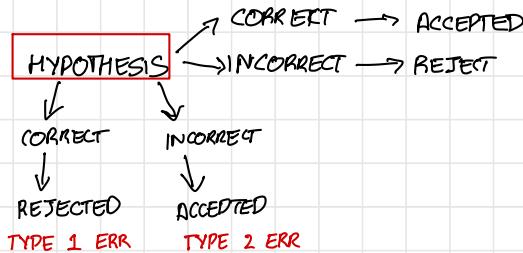
n = sample size

In this formula,  $\hat{p}$  is the point estimate and  $\pm z_{\alpha/2} \sqrt{\frac{\hat{p} \cdot \hat{q}}{n}}$  is the error of the estimation.

# TESTING OF HYPOTHESIS

VALIDATE A GIVEN HYPOTHESIS USING SAMPLING.

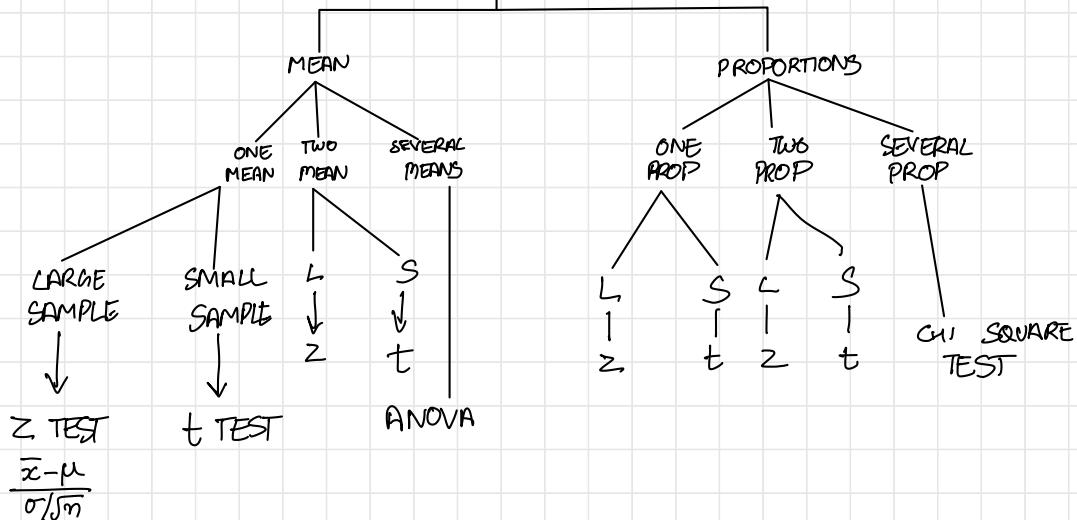
VERIFYING A STATEMENT IS NON PARAMETRIC, BUT NUMERICAL VALUE VERIFICATION IS PARAMETRIC



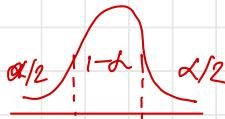
$$P(\text{TYPE 1 ERR}) = \alpha$$

$$P(\text{TYPE 2 ERR}) = \beta$$

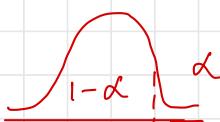
## TESTING OF HYPOTHESIS



TWO TAIL TEST

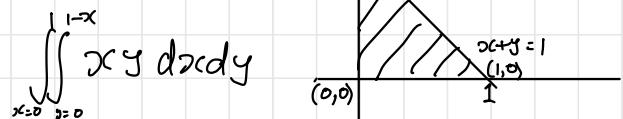


ONE TAIL TEST





$$P(x+y \leq 1)$$



$$\int_{x=0}^{1-x} \int_{y=0}^{1-x} dxdy$$

$$\int_{x=0}^1 \left( \frac{x(1-x)^2}{2} - 0 \right) dx$$

$$\int_{x=0}^1 \frac{x(1-x^2-2x)}{2} dx$$

$$\left[ \frac{x^2}{2} - \frac{x^3}{3} - \frac{2x^2}{2} \right]_0^1$$

$$\frac{1}{2} - \frac{1}{8} - \frac{1}{3}$$

$$\frac{12-3-8}{24}$$

$$=\frac{1}{24}$$

## POST MID

- 1) HYPOTHESIS TESTING → MORE WEIGHTAGE
- 2) CORRELATION & REGRESSION
- 3) TIME SERIES ANALYSIS
- 4) BI VARIATE DISTRIBUTION