



BITS Pilani

Pilani | Dubai | Goa | Hyderabad

Introduction to Statistical Methods

ISM Team



Overview of the course & Basic Probability & Statistics (CS -1)

(Session 1: 12th /13th Nov 2022)

Overview of the course

- ❖ M 1 : Basic Probability & Statistics
- ❖ M 2 : Conditional Probability & Bayes' theorem
- ❖ M 3 : Probability Distributions
- ❖ M 4 : Hypothesis Testing
- ❖ M 5 : Prediction & Forecasting
- ❖ M 6 : Prediction & Forecasting Gaussian Mixture model & Expectation Maximization

TEXT BOOKS

T1 : Statistics for Data Scientists, An introduction to probability
statistics and Data Analysis, Maurits Kaptein et al, Springer 2022

T2 : Probability and Statistics for Engineering and Sciences,
8th Edition, Jay L Devore, Cengage Learning

T3 : Introduction to Time Series and Forecasting, Second Edition,
Peter J Brockwell, Richard A Davis, Springer.

Evaluation Components

No	Name	Type	Weight
EC-1(a)	Quizzes – 1 & 2	Online	10%
EC-1(b)	Assignments - 2	Online	20%
EC-2	Mid-Semester Test	Closed Book	30%
EC-3	Comprehensive Exam	Open Book	40%

Module 1: (Basic Probability & Statistics)

Contact Session	List of Topic Title	Reference
CS - 1	Measures of Central Tendency & Measures of Variability, Data – Symmetric & Asymmetric, outlier detection, 5 point summary, Introduction to probability	T1 & T2

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

H G Wells

Statistics

Statistics may be defined as science that is employed to

- Collect the data
- Present and organize the data in a systematic manner
- Analyse the data
- Infer about the data
- Take decision from the data.

In other words, Statistics can also be defined as numerical data with a view to analyse it.

Types of Variable

Qualitative (Categorical): express a qualitative attribute such as hair color, eye color, religion.

Quantitative(Numerical): measured in terms of numbers such as height, weight, number of people.

Nominal: no ordering is possible such as hair color, eye color, religion.

Ordinal: ordering is possible such as health, which can take values such as poor, reasonable, good, or excellent.

Discrete: countable and have a finite number of possibilities such as number of people

Continuous: not countable and have an infinite number of possibilities such as height

INTERVAL: ratio of values of variable do not have any meaning and it does not have an inherently defined zero value such as temperature

RATIO: ratio of values of variable have meaning and it have an inherently defined zero value such as length

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

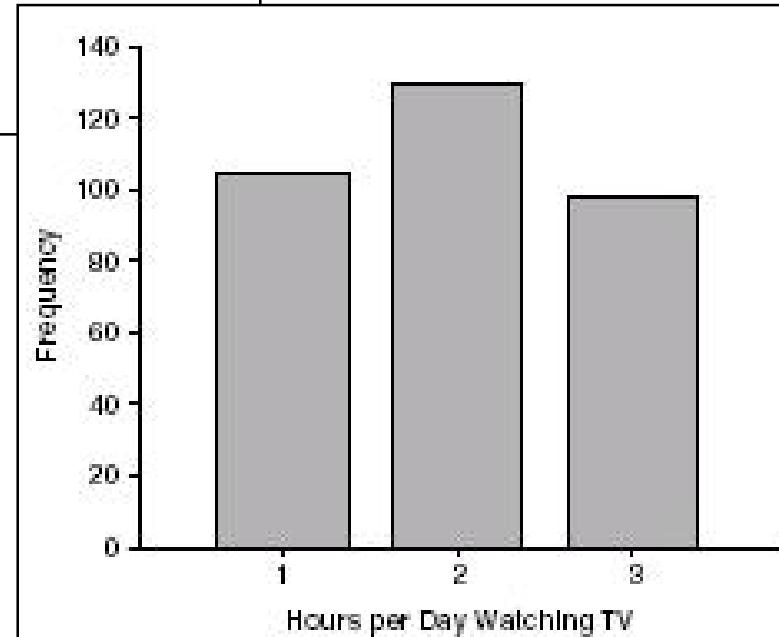
$$\bar{Y} = \frac{\sum f Y}{N} \quad \text{Where } f Y = \text{a score multiplied by its frequency}$$

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.
-

Mean

When to Use the Mean

- Sampling stability is desired.
- Other measures are to be computed such as standard deviation, coefficient of variation and skewness

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 not affected by extreme values.

When to Use the Mode

- 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.
-

Measures of central tendency

➤ The mean

Draw Back?

$$\frac{x_1 + x_2 + x_3 + \dots + x_n}{n}$$

μ

$n-1$ \bar{x}

➤ the median

$$10, 15, 20, 25, 26$$

$$10, 15, 20, 25, 28, 32$$

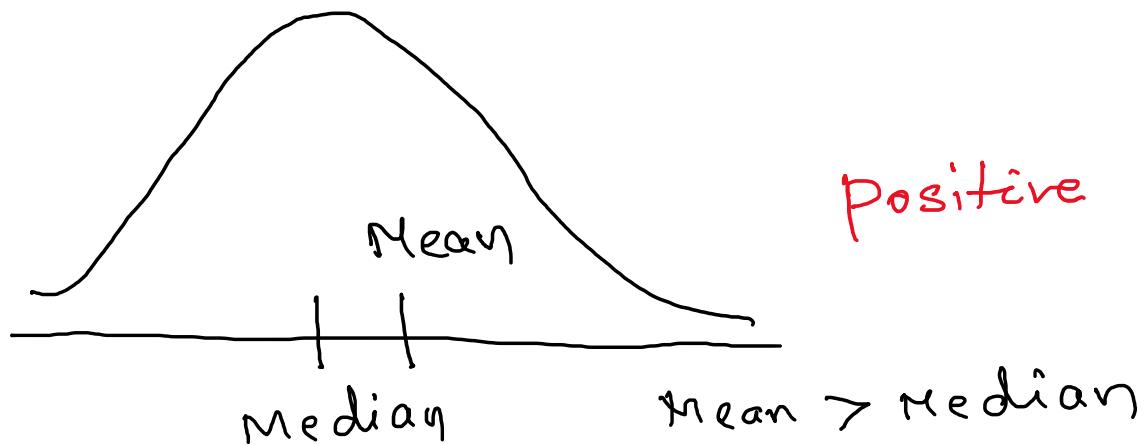
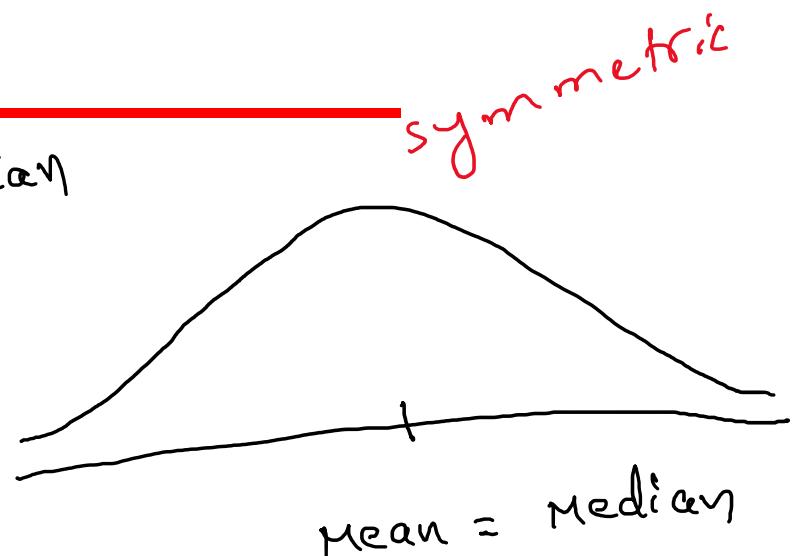
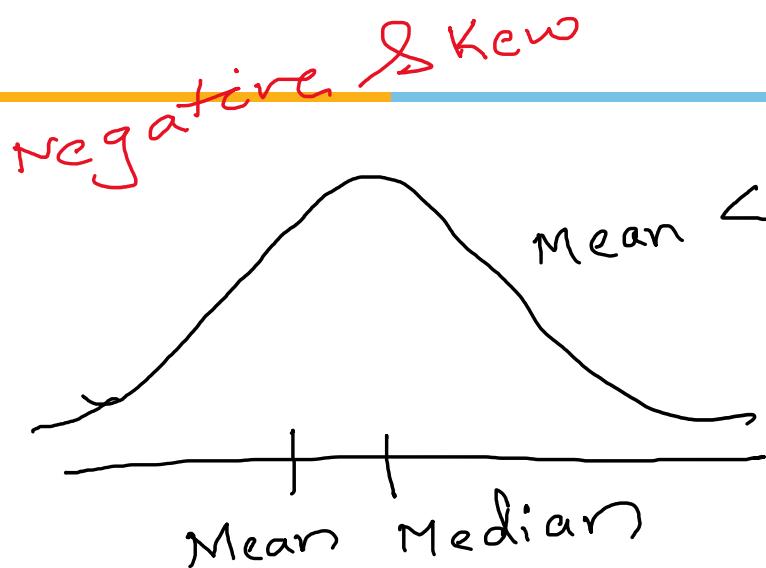
Average of
these

➤ the mode

$$2, 5, 5, 2, 3, 2, 2, 2$$

$$5, 2, 5, 5, 2, 3, 2, 2, 2, 5, 5$$

Data : Symmetrical and Asymmetrical



positive skew

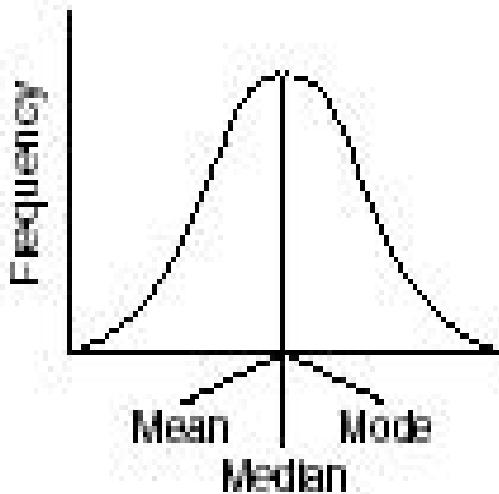
mean > median

Shape of the distribution of data

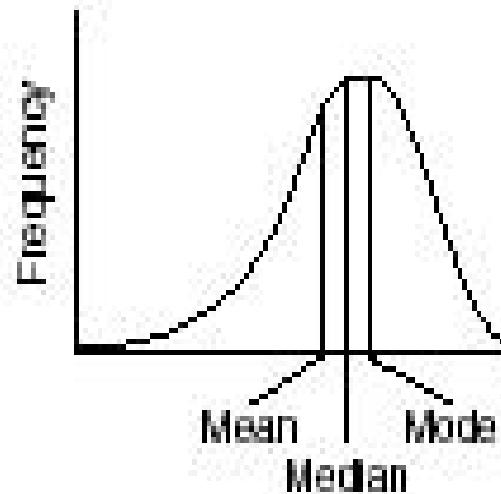
- Symmetrical : Mean is equal to median
 - Skewed
 - Negatively : mean < median
 - Positively : mean > median
 - Bimodal : has two distinct modes
 - Multi-modal : has more than 2 distinct modes
-

Distribution Shape

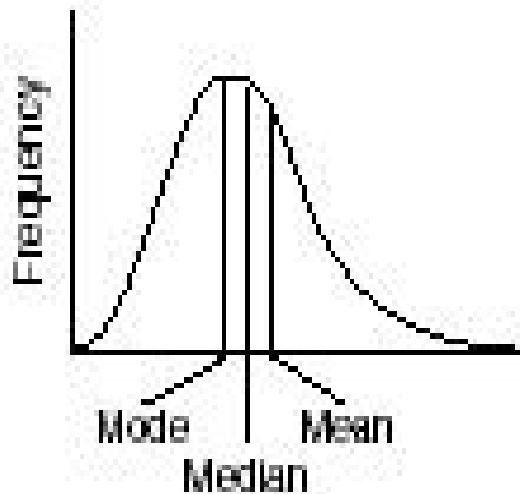
Types of Frequency Distributions



a. Symmetrical distribution



b. Negatively skewed distribution



c. Positively skewed distribution

Sl. No.	X_1	X_2
1	2	1
2	8	15
3	5	5
4	3	5
5	7	6
6	8	3
7	5	5
8	2	2
9	5	3
Total	45	45

Sl. No.	X_1
1	2
2	8
3	5
4	3
5	7
6	8
7	5
8	2
9	5
Total	45

Statistical measures	Group 1
Mean	5
Median	5
Mode	5

Sl. No.	X_2
1	1
2	15
3	5
4	5
5	6
6	3
7	5
8	2
9	3
Total	45

Statistical measures	Group 2
Mean	5
Median	5
Mode	5

Sl. No.	X_1	X_2
1	2	1
2	8	15
3	5	5
4	3	5
5	7	6
6	8	3
7	5	5
8	2	2
9	5	3
Total	45	45

Statistical
measures

Group
1 & 2

Mean

5

Median

5

Mode

5



Do we need any other measure?

Answer: Yes

Measures of variability

Three Measures of Variability:

- The Range
 - The Variance
 - The Standard Deviations
-

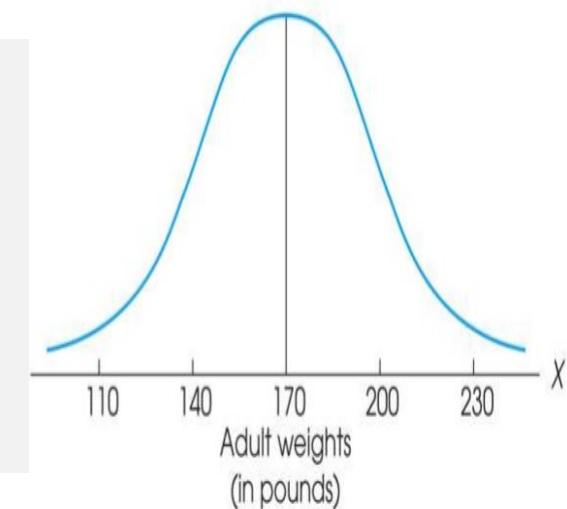
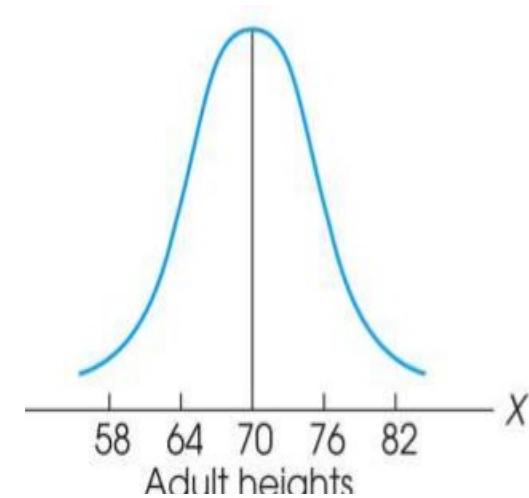
Measure of Variability

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

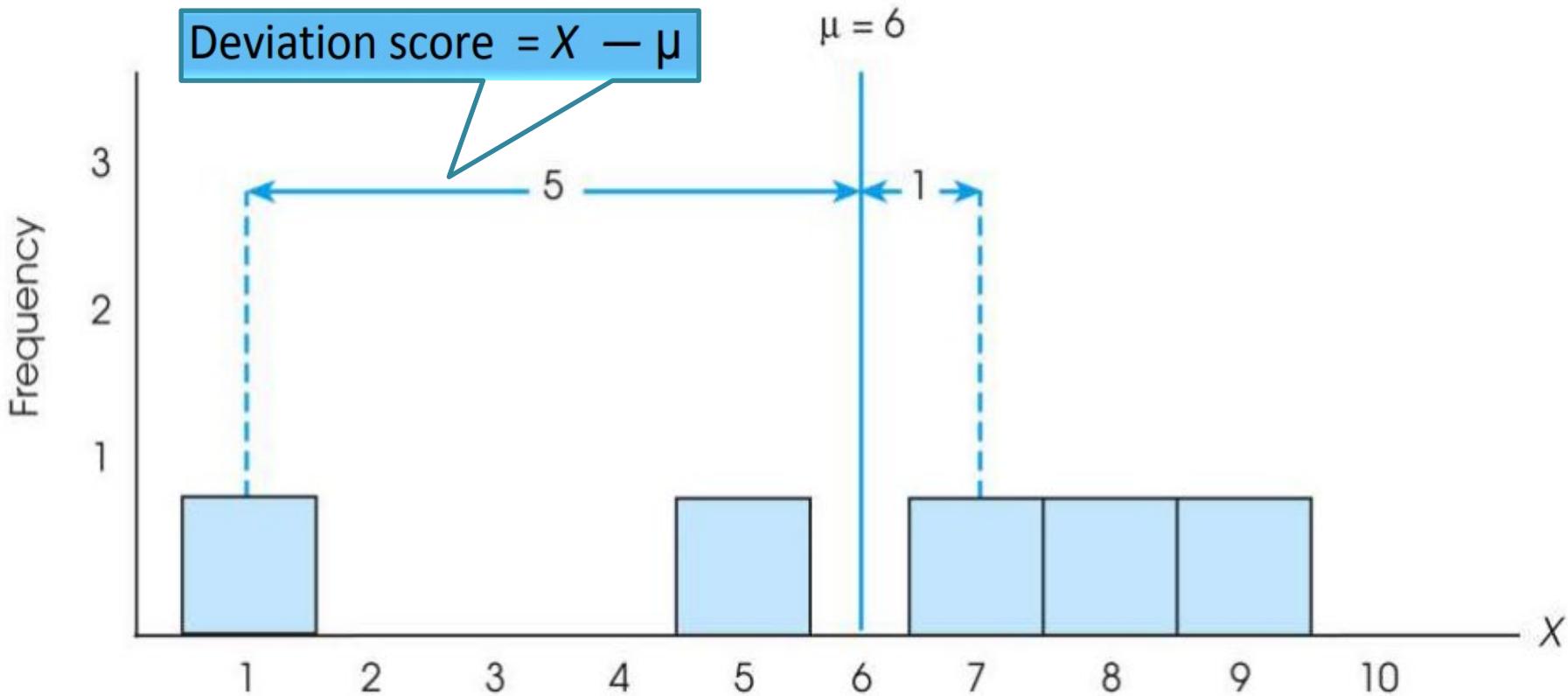
Example: For a set of scores: 7, 2, 7, 6, 5, 6, 2

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

- Mean deviations will always be ‘zero’ !
(because Mean is a balance point)

Then, how do you find ‘Standard Deviation’ ?



Need a new strategy

The Standard Deviation

New Strategy :

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

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

Standard Deviation = $\sqrt{Variance}$

The Variance

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

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

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

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: σ^2
- ❖ Standard deviation is the square root of the variance, so we identify it with a lowercase Greek letter sigma: σ

Sl. No.	X_1
1	2
2	8
3	5
4	3
5	7
6	8
7	5
8	2
9	5
Total	45

Statistical measures	Group 1
Mean	5
Median	5
Mode	5

Sl. No.	X_1
1	2
2	8
3	5
4	3
5	7
6	8
7	5
8	2
9	5
Total	45

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{45}{5} = 5$$

$$S = \sqrt{\frac{\sum (X - \bar{X})^2}{n-1}}$$

$$S = \sqrt{\frac{44}{8}} = 2.345$$

Sl. No.	X_2
1	1
2	15
3	5
4	5
5	6
6	3
7	5
8	2
9	3
Total	45

$$\bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{45}{5} = 5$$

$$S = \sqrt{\frac{\sum (X - \bar{X})^2}{n - 1}}$$

$$S = \sqrt{\frac{134}{8}} = 4.093$$

Learning Check

- a) If all the scores in a data set are the same, the Standard Deviation is equal to 1.00

True / False
?

Select the correct option

- b) The standard deviation measures ...
- (1) Sum of squared deviation scores
 - (2) Standard distance of a score from the mean
 - (3) Average deviation of a score from the mean
 - (4) Average squared distance of a score from the mean
-

Solution

- a) If all the scores in a data set are the same, they are equal to the mean and hence the deviation from mean = 0 therefore, Standard Deviation is equal to **zero**

False

- b) The standard deviation measures ...
- (1) Sum of squared deviation scores
 - (2) Standard distance of a score from the mean
 - (3) Average deviation of a score from the mean
 - (4) Average squared distance of a score from the mean
-

Standard Deviation and Variance for a Sample



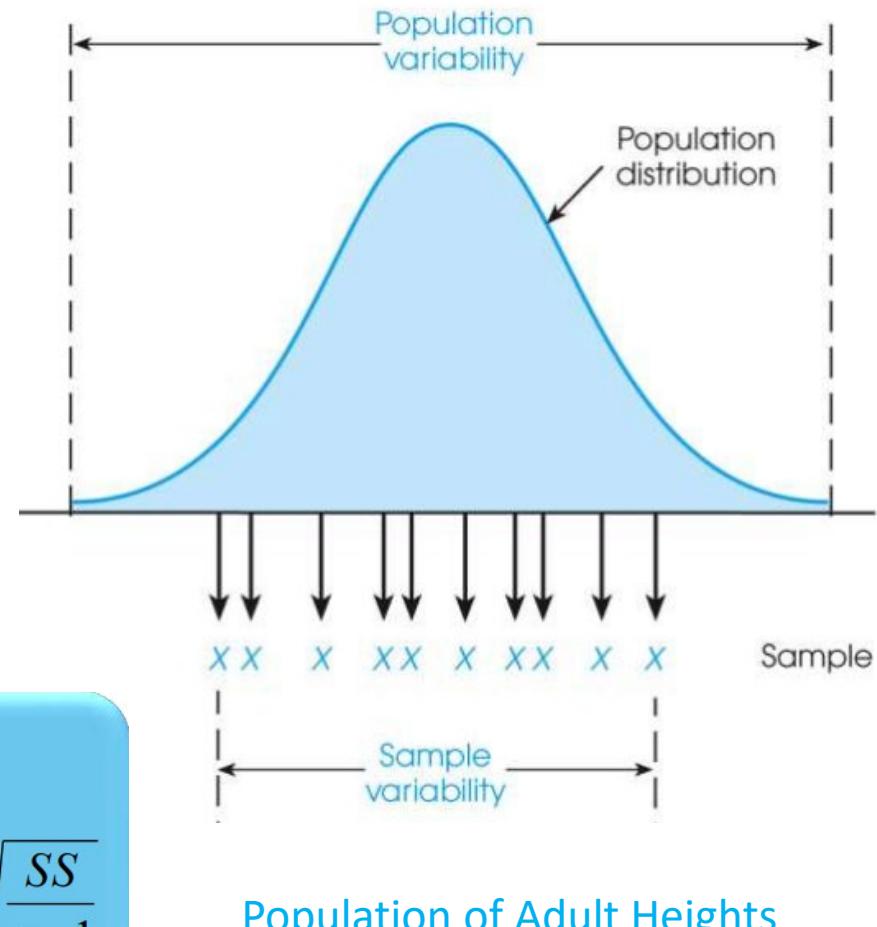
- Goal of inferential statistics:
 - Draw general conclusions about population
 -
 - Based on limited information from a sample
- Samples differ from the population
 - Samples have less variability
 - Computing the Variance and Standard Deviation in the same way as for a population would give a biased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

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

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



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) = $n - 1$
-

Learning Check

Select the correct option

a) A sample of four scores has $SS = 24$. What is the variance?

- (1) The variance is 6
- (2) The variance is 7
- (3) The variance is 8
- (4) The variance is 12

b) A sample systematically has less variability than a population

c) The standard deviation is the distance from the Mean to the farthest point on the distribution curve

True / False
?

True / False
?

Solution

Select the correct option

- a) A sample of four scores has $SS = 24$. What is the variance?
 - (1) The variance is 6
 - (2) The variance is 7
 - (3) The variance is 8
 - (4) The variance is 12

- b) Extreme scores affect variability, but are less likely to be included in a sample

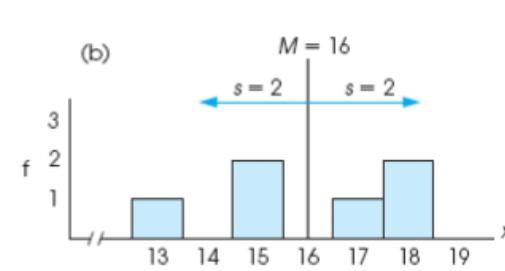
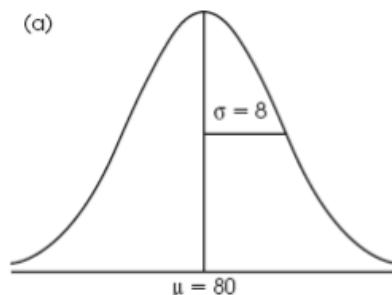
- c) The standard deviation extends from the mean approximately halfway to the most extreme score

True

False

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



Five point summary of Data

The five number summary of data includes 5 items:

- ❖ **Minimum.**
 - ❖ **Q1** (the first quartile, or the 25% mark).
 - ❖ **Median.**
 - ❖ **Q3** (the third quartile, or the 75% mark).
 - ❖ **Maximum.**
-

Interquartile range (IQR)

- It is measure of Variation
- Also Known as Midspread : Spread in the Middle 50%
- Difference Between Third & First Quartiles:
- Not Affected by Extreme Values

$$\text{Interquartile Range} = \text{IQR} = Q_3 - Q_1$$

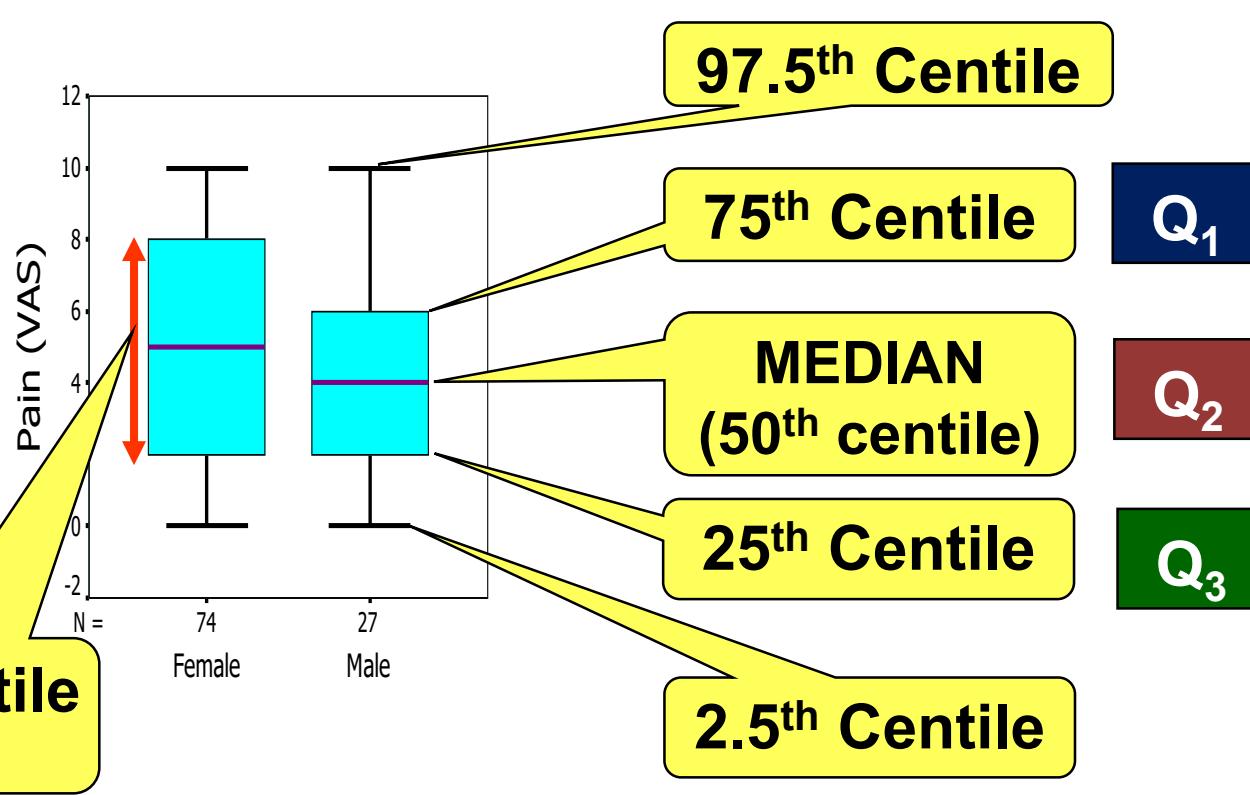
Data in Ordered Array: 11 12 13 16 16 17 17 18 21

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

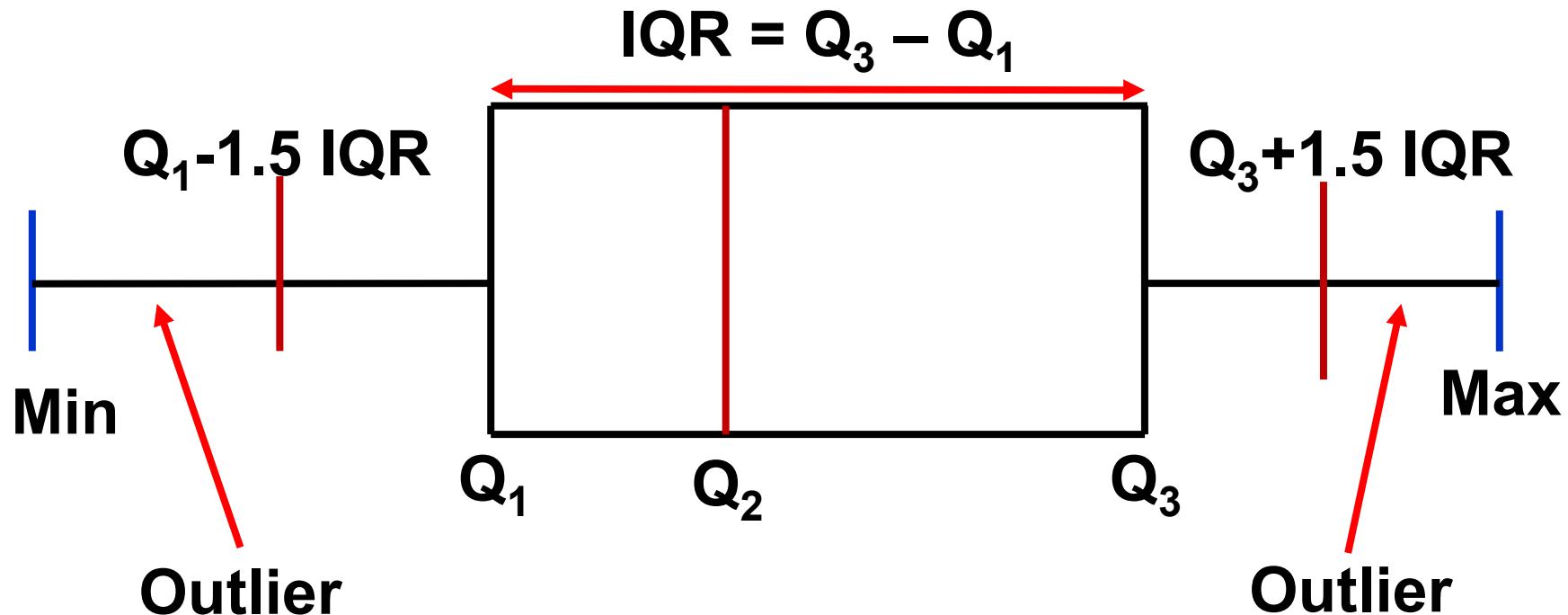
$$\text{Position of } Q_3 = \frac{3 \cdot (9 + 1)}{4} = 7.50, \quad Q_3 = 17.5$$

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

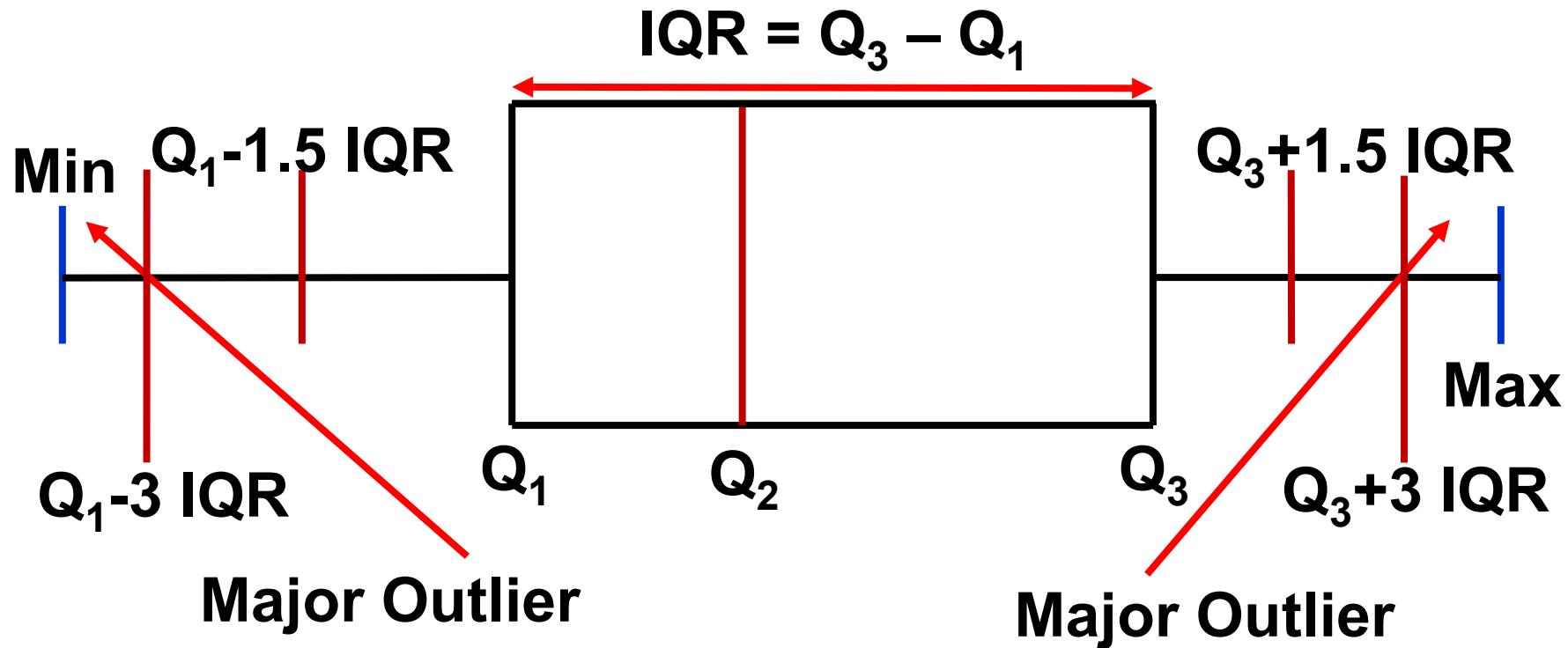
Box and Whisker plot



Box and Whisker plot



Box-and-Whisker plot



Potential outliers

- ❖ The lower limit and upper limit of a data set are given by:

$$\text{Lower limit} = Q_1 - 1.5 \times \text{IQR}$$

$$\text{Upper limit} = Q_3 + 1.5 \times \text{IQR}$$

- ❖ Data points that lie below the lower limit or above the upper limit are **potential outliers**.
-

HW problem :

For the data set below:

82	45	64	80	82	74	79	80	80	78	80	80	48	73	80	79	81	70	78	73
----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----	----

- (a.)** Obtain and interpret the quartiles.
 - (b.)** Determine and interpret the interquartile range.
 - (c.)** Find and interpret the five-number(point) summary.
 - (d.)** Identify potential outliers, if any.
 - (e.)** Construct and interpret a boxplot.
-

HW problem :

Human measurements provide a rich area of application for statistical methods. The article "A Longitudinal Study of the Development of Elementary School Children's Private Speech" (*Merrill-Palmer Q.*, 1990: 443–463) reported on a study of children talking to themselves (private speech). It was thought that private speech would be related to IQ, because IQ is supposed to measure mental maturity, and it was known that private speech decreases as students progress through the primary grades. The study included 33 students whose first-grade IQ scores are given here:

82	96	99	102	103	103	106	107	108	108	108	108
109	110	110	111	113	113	113	113	115	115	118	118
119	121	122	122	127	132	136	140	146			

Describe the data and comment on any interesting features.



Introduction to probability

Random Experiment :

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

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

Sample spaces and events

Sample space :

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

If the experiment is to throw a die and record the outcome, then,

sample space is $S = \{ 1,2,3,4,5,6\}$

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

Definition of probability

Classical approach :

CLASSICAL PROBABILITY

Probability of an event = $\frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}}$

Empirical approach :

Empirical or **relative frequency** is the second type of objective probability. It is based on the number of times an event occurs as a proportion of a known number of trials.

EMPIRICAL PROBABILITY The probability of an event happening is the fraction of the time similar events happened in the past.

In terms of a formula:

$$\text{Empirical probability} = \frac{\text{Number of times the event occurs}}{\text{Total number of observations}}$$

The empirical approach to probability is based on what is called the law of large numbers. The key to establishing probabilities empirically is that more observations will provide a more accurate estimate of the probability.

LAW OF LARGE NUMBERS Over a large number of trials, the empirical probability of an event will approach its true probability.

Axiomatic approach :

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

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

- (1) $P(S) = 1$
- (2) $0 \leq P(E) \leq 1$
- (3) For two events E_1 and E_2 with $E_1 \cap E_2 = \emptyset$

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



Thank You



BITS Pilani

Pilani | Dubai | Goa | Hyderabad

M.Tech.(Data Science & Engineering)

Introduction to Statistical Methods

Team ISM



Session No 2

**Axioms of Probability, Probability basics, mutually exclusive and
independent events,**

(Session 2: 19th/20th Nov 2022)

Contact Session 2

Contact Session 2: Module 1(Module 1:Basic Probability & Statistics)

Contact Session	List of Topic Title	Reference
CS - 2	Axioms of Probability,Mutually exclusive and independent events,Problem solving to understand basic probability concepts	T1 & T2
HW	Problems on probability	T1 & T2
Lab		

Agenda

- Experiments, assignment of probabilities
- Events and their probability
- Some basic relationships of probability
- Basic problem solving

RECALL: Random Experiment

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

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

Sample spaces, sample sets and events

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

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

-
- Discrete sample spaces.
 - Continuous sample spaces
-

Discrete Random Variables

- A discrete random variable is one which may take on only a countable number of distinct values such as 0, 1, 2, 3, 4,....
- • Discrete random variables are usually (but not necessarily) counts.

Examples:

- ❖ number of children in a family
 - ❖ the Friday night attendance at a cinema
 - ❖ the number of patients a doctor sees in one day
 - ❖ the number of defective light bulbs in a box of ten
 - ❖ the number of “heads” flipped in 3 trials
-

Continuous Random Variable

- ❖ A continuous random variable is one which takes an infinite number of possible values.

 - ❖ Examples:
 - ✓ height
 - ✓ weight
 - ✓ the amount of sugar in an orange
 - ✓ the time required to run a mile.
-

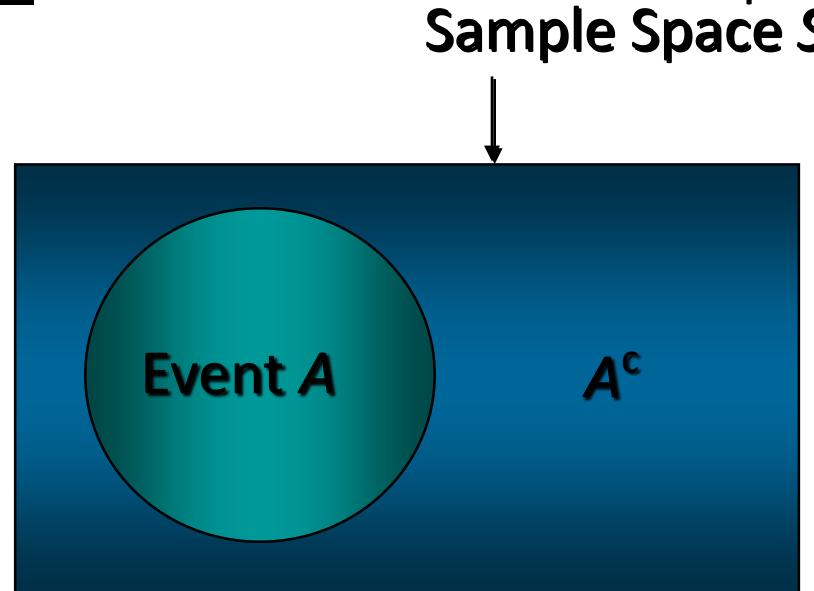
Event

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

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

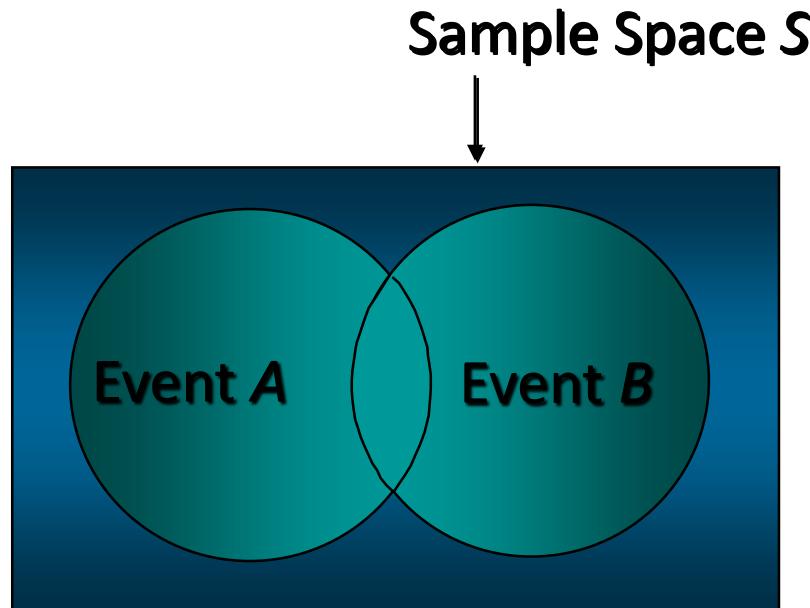
Complement of an Event

- The complement of event A is defined to be the event consisting of all sample points that are not in A .
- The complement of A is denoted by A^c .
- The Venn diagram below illustrates the concept of a complement.



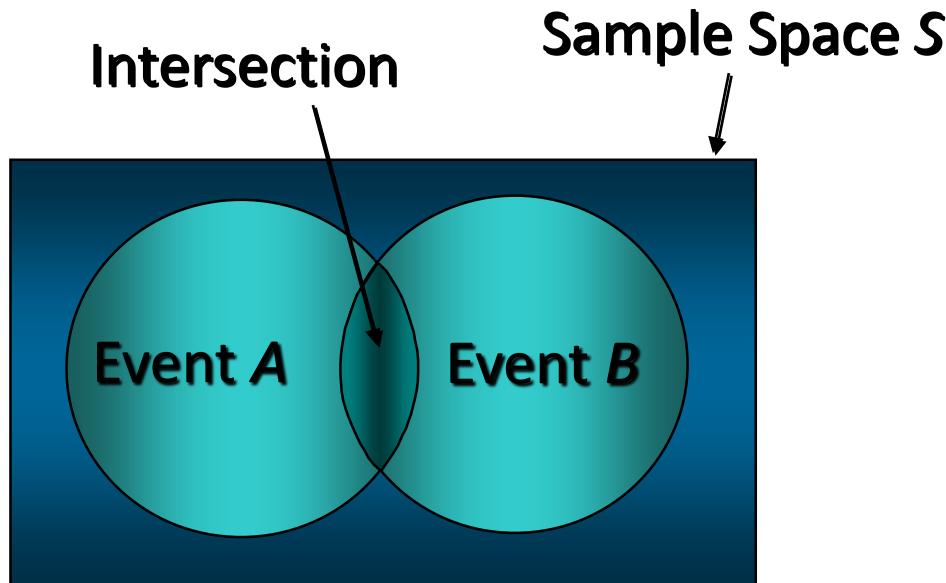
Union of Two Events

- The union of events A and B is the event containing all sample points that are in A or B or both.
- The union is denoted by $A \cup B$
- The union of A and B is illustrated below.



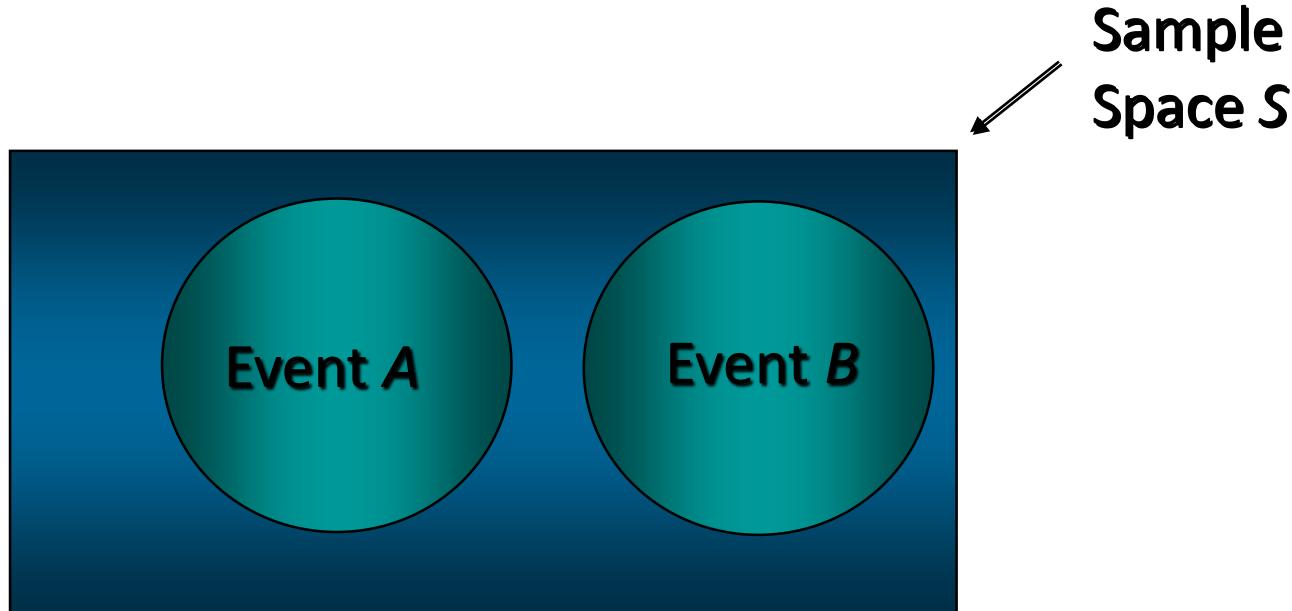
Intersection of Two Events

- The intersection of events A and B is the set of all sample points that are in both A and B .
- The intersection of A and B is the area of overlap in the illustration below.



Mutually Exclusive Events

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



Axioms of Probability



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

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

- (1) $P(S) = 1$
- (2) $0 \leq P(E) \leq 1$
- (3) For two events E_1 and E_2 with $E_1 \cap E_2 = \emptyset$

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

Probability as a Numerical Measure of the Likelihood of Occurrence

Increasing Likelihood of Occurrence



The occurrence of the event is
just as likely as it is unlikely.

THE ADDITION RULE

- The probability that event A or B will occur is given by

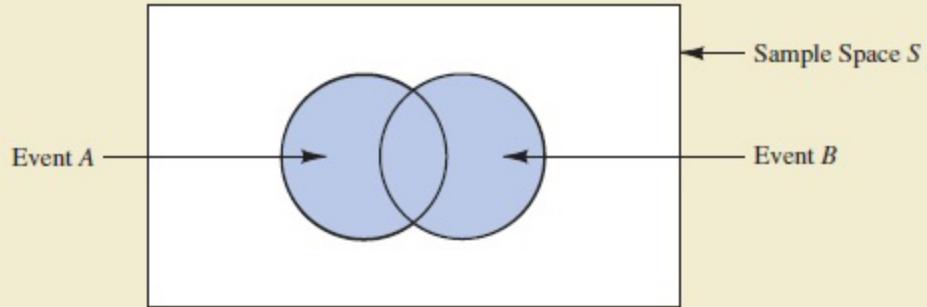
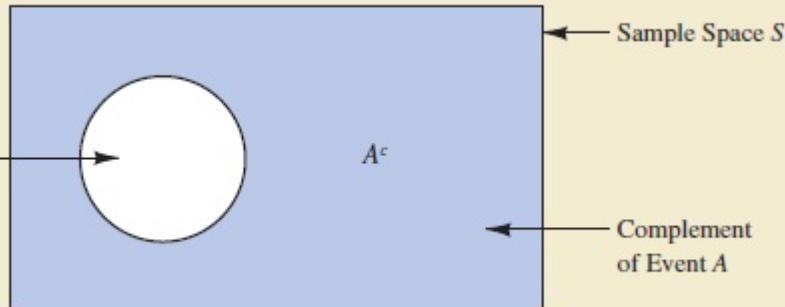
$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

- If events A and B are **mutually exclusive**, then the rule can be simplified to

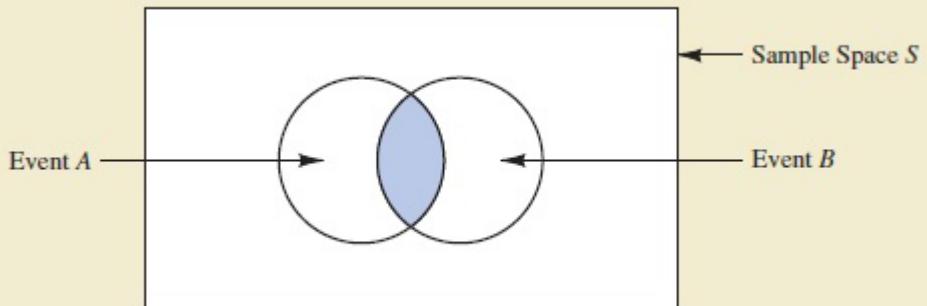
$$P(A \cup B) = P(A) + P(B).$$

Probability and Venn Diagram

$$P(A) = 1 - P(A^c)$$



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



Independent & Dependent

Events are either

- *Independent* (the occurrence of one event has no effect on the probability of occurrence of the other) or
 - *Dependent* (the occurrence of one event gives information about the occurrence of the other)
-

Example

An experiment has the four possible mutually exclusive outcomes A, B, C, D. Check whether the following assignments of probability are permissible:

- (a) $P(A)= 0.38, P(B) = 0.16, P(C) =0.11, P(D) = 0.35$ Permissible
- (b) $P(A)= 0.31, P(B) = 0.27, P(C) =0.28, P(D) = 0.16$ NOT
- (c) $P(A)= 0.32, P(B) = 0.27, P(C) = -0.06, P(D) = 0.47$ NOT
- (d) $P(A)= 1/2, P(B) = 1/4, P(C) = 1/8, P(D) = 1/16$ NOT
- (e) $P(A)= 5/8, P(B) = 1/6, P(C) = 1/3, P(D) = 2/9$ NOT

EXAMPLE

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
-

Example

$$P(X = 2) = P\{(1, 1)\} = \frac{1}{36}$$

$$P(X = 3) = P\{(1, 2), (2, 1)\} = \frac{2}{36}$$

$$P(X = 4) = P\{(1, 3), (2, 2), (3, 1)\} = \frac{3}{36}$$

$$P(X = 5) = P\{(1, 4), (2, 3), (3, 2), (4, 1)\} = \frac{4}{36}$$

$$P(X = 6) = P\{(1, 5), (2, 4), (3, 3), (4, 2), (5, 1)\} = \frac{5}{36}$$

$$P(X = 7) = P\{(1, 6), (2, 5), (3, 4), (4, 3), (5, 2), (6, 1)\} = \frac{6}{36}$$

$$P(X = 8) = P\{(2, 6), (3, 5), (4, 4), (5, 3), (6, 2)\} = \frac{5}{36}$$

$$P(X = 9) = P\{(3, 6), (4, 5), (5, 4), (6, 3)\} = \frac{4}{36}$$

$$P(X = 10) = P\{(4, 6), (5, 5), (6, 4)\} = \frac{3}{36}$$

$$P(X = 11) = P\{(5, 6), (6, 5)\} = \frac{2}{36}$$

$$P(X = 12) = P\{(6, 6)\} = \frac{1}{36}$$

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

\rightarrow a) $P(\text{sum} > 8)$

$$= P(9) + P(10) + P(11) + P(12)$$

$$= \frac{4}{36} + \frac{3}{36} + \frac{2}{36} + \frac{1}{36} = \frac{10}{36}$$

\rightarrow b) $P(\text{sum} < 6)$

$$= P(5) + P(4) + P(3) + P(2)$$

$$= \frac{1}{36} + \frac{2}{36} + \frac{3}{36} + \frac{4}{36} = \frac{10}{36}$$

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

$$\begin{aligned}
 & P(\text{neither 7 nor 11}) \\
 &= 1 - [P(7) + P(11)] \\
 &= 1 - \left[\frac{6}{36} + \frac{2}{36} \right] \\
 &= \frac{28}{36} = \frac{7}{9}
 \end{aligned}$$

Example

Consider the following table.

	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

A Survey conducted by a bank revealed that 40% of the accounts are savings accounts and 35% of the accounts are current accounts and the balance are loan accounts.

- What is the probability that an account taken at random is a loan account ?
 - What is the probability that an account taken at random is **NOT** savings account ?
 - What is the probability that an account taken at random is **NOT** a current account
 - What is the probability that an account taken at random is a current account or a loan account?
-

Example

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?

Example

The sales manager of an e-commerce company says that 80% of those who visit their website for the first time do not buy any mobile. If a new customer visits the website, what is the probability that the customer would buy mobile

:

EXAMPLE

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.

Example

The next generation of miniaturised wireless capsules with active locomotion will require two miniature electric motors to manoeuvre each capsule. Suppose 10 motors have been fabricated but that, in spite of test performed on the individual motors 2 will not operate satisfactorily when placed into capsule, to fabricate a new capsule, 2 motors will be randomly selected(that is, each pair of motors has the same chance of being selected) find the probability that

- a) Both motors will operate satisfactorily in the capsule.
 - b) One motor will operate satisfactorily and other will not.
-

HW problems

Q) Consider randomly selecting a student at a certain university, and let A denote the event that the selected individual has a Visa credit card and B be the analogous event for a MasterCard.

Suppose that $P(A)= 0.5$, $P(B)= 0.4$, $P(A \cap B) = 0.3$.

- i) Compute the probability that the selected individual has at least one of the two types of cards (i.e., the probability of the event $A \cup B$).
 - ii) What is the probability that the selected individual has neither type of card?
 - iii) Describe, in terms of A and B , the event that the selected student has a Visa card but not a MasterCard, and then calculate the probability of this event.
-

Q) In a certain residential suburb, 60% of all households get Internet service from the local cable company, 80% get television service from that company, and 50% get both services from that company.

If a household is randomly selected,

- i) What is the probability that it gets at least one of these two services from the company, and
 - ii) What is the probability that it gets exactly one of these services from the company?
-

Q) Suppose that 55% of all adults regularly consume coffee, 45% regularly consume carbonated soda, and 70% regularly consume at least one of these two products.

- i) What is the probability that a randomly selected adult regularly consumes both coffee and soda?

 - ii) What is the probability that a randomly selected adult doesn't regularly consume at least one of these two products?
-

Q) The probability that 'A' will be alive 10 years hence is $\frac{5}{8}$ and that B will be alive is $\frac{3}{4}$. Find the probability that

- a) At least one is alive
 - b) Exactly one is alive
 - c) None are alive
-

Q) Suppose a student is selected at random from 80 students where 30 are taking mathematics, 20 are taking chemistry and 10 are taking both. Find the probability 'p' that the student is taking Mathematics or chemistry?.

3) If A and B are events with $P(A \cup B) = 7/8$, $P(A \cap B) = 1/4$ and $P(A') = 5/8$, find $P(A)$, $P(B)$ and $P(A \cap B')$.

-
- Q) The probability that a new airport will get an award for its design is 0.16, the probability that it will get an award for the efficient use of materials is 0.24 and the probability that it will get both award is 0.11
- a) What is probability that it will get at least one of the two awards?
 - b) What is probability that it will get only one of the two awards?
-



Thanks





BITS Pilani

Pilani|Dubai|Goa|Hyderabad

M.Tech.(AIML)
Introduction to Statistical Methods

Team ISM



Session No 3

**Introduction to Conditional Probability, independent events, Total
Probability**

(Session 3: 26th/27th Nov 2022)

Contact Session 3



Contact Session 3: Module 2(Conditional Probability & Bayes theorem)

Contact Session	List of Topic Title	Reference
CS - 3	Introduction to conditional probability,independents events, Total probability	T1 & T2
HW	Problems on conditional probability	T1 & T2
Lab		



Agenda

- Conditional Probability
- Independent events
- Total Probability

Text Books

No	Author(s), Title, Edition, Publishing House
T1	Statistics for Data Scientists, An introduction to probability, statistics and Data Analysis, Maurits Kaptein et al, Springer 2022
T2	Probability and Statistics for Engineering and Sciences, 8 th Edition, Jay L Devore, Cengage Learning

CONDITIONAL PROBABILITY

The probabilities assigned to various events depend on what is known about the experimental situation when the assignment is made. Subsequent to the initial assignment, partial information relevant to the outcome of the experiment may become available. Such information may cause us to revise some of our probability assignments

CONDITIONAL PROBABILITY

We examine how the information “an event B has occurred” affects the probability assigned to A. For example, A might refer to an individual having a particular disease in the presence of certain symptoms. If a blood test is performed on the individual and the result is negative , then the probability of having the disease will change (it should decrease, but not usually to zero, since blood tests are not infallible). We will use the notation to represent the conditional probability of A given that the event B has occurred. B is the “conditioning event.”

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

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

Conditional probabilities in a 2×2 contingency table

	A	A^c	
B	$\Pr(A \cap B) = \Pr(A B) \Pr(B)$	$\Pr(A^c \cap B) = \Pr(A^c B) \Pr(B)$	$\Pr(B)$
B^c	$\Pr(A \cap B^c) = \Pr(A B^c) \Pr(B^c)$	$\Pr(A^c \cap B^c) = \Pr(A^c B^c) \Pr(B^c)$	$\Pr(B^c)$
	$\Pr(A)$	$\Pr(A^c)$	1

Examples on Conditional Probability

Example: In a housing colony, 70% of the houses are well planned and 60% of the houses are well planned and well built. Find the probability that an arbitrarily chosen house in this colony is well built given that it is well planned.

Solution: Let A be the event that the house is well planned

B be the event that the house is well built

therefore $P(A) = 0.70$, $P(A \cap B) = 0.60$

$$P\left(\frac{B}{A}\right) = \frac{P(B \cap A)}{P(A)} = \frac{0.60}{0.70} = 0.8571$$

CONDITIONAL PROBABILITY

Example:

If the probability that a communication system will have high fidelity is 0.81 and the probability that it will have high fidelity and high selectivity is 0.18, what is probability that a system with high fidelity will also have high selectivity?

Solution:

If A is the event that a communication system has high selectivity and B is the event that it has high fidelity, we have $P(B)=0.81$ and $P(A \cap B)=0.18$, and substitution into the formula yields

$$P(A | B) = \frac{0.18}{0.81} = \frac{2}{9}$$

CONDITIONAL PROBABILITY



Example:

Suppose that of all individuals buying a certain digital camera, 60% include an optional memory card in their purchase, 40% include an extra battery, and 30% include both a card and battery.

- Given that the randomly selected individual purchased an extra battery, find the probability that an optional card was also purchased.
- Given that the randomly selected individual purchased an optional card, find the probability that an extra battery was also purchased.

Solution:

let $A = \{\text{memory card purchased}\}$ and $B = \{\text{battery purchased}\}$. Then $P(A) = .60$, $P(B) = .40$, and $P(\text{both purchased}) = P(A \cap B) = .30$.

$$a) \quad P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{.30}{.40} = .75$$

$$b) \quad P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{.30}{.60} = .50$$

Notice that $P(A|B) \neq P(A)$ and $P(B|A) \neq P(B)$.

Example:

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

<i>Read regularly</i>	A	B	C	$A \cap B$	$A \cap C$	$B \cap C$	$A \cap B \cap C$
<i>Probability</i>	.14	.23	.37	.08	.09	.13	.05

Find $P(A|B)$, $P(A|B \cup C)$, $P(A|$ reads at least one)

Solution: We thus have

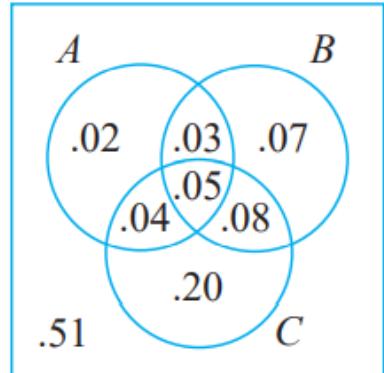
$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{.08}{.23} = .348$$

$$P(A|B \cup C) = \frac{P(A \cap (B \cup C))}{P(B \cup C)} = \frac{.04 + .05 + .03}{.47} = \frac{.12}{.47} = .255$$

$$\begin{aligned} P(A|\text{reads at least one}) &= P(A|A \cup B \cup C) = \frac{P(A \cap (A \cup B \cup C))}{P(A \cup B \cup C)} \\ &= \frac{P(A)}{P(A \cup B \cup C)} = \frac{.14}{.49} = .286 \end{aligned}$$

and

$$P(A \cup B|C) = \frac{P((A \cup B) \cap C)}{P(C)} = \frac{.04 + .05 + .08}{.37} = .459$$



Examples on Conditional Probability

Example: In a certain college, 25% of the students failed Maths, 15% of the students failed chemistry and 10% of the students failed both maths and chemistry. A student is selected at random, find

- If he failed chemistry, what is the probability that he failed Maths? $P(M) = 0.25$, $P(C) = 0.15$, $P(M \cap C) = 0.1$

$$P\left(\frac{M}{C}\right) = \frac{P(M \cap C)}{P(C)} = \frac{0.1}{0.15} = 0.6667$$

- If he failed maths, what is the probability he failed chemistry?

$$P\left(\frac{C}{M}\right) = \frac{P(C \cap M)}{P(M)} = \frac{0.1}{0.25} = 0.4$$

- What is the probability that the student failed in Maths or chemistry?

Examples on Conditional Probability

Example : The probabilities of a regularly scheduled flight departs on time is 0.83, arrives on time is 0.82 & it departs and arrives on time is 0.78. Find the probability that a plane (i) arrives on time given that it departed on time, (ii) departed on time given that it has arrived on time and (iii) find $P\left(\frac{A}{D}\right)$

Ans: Let D and A be the events that the flight departs and arrives on time respectively. Then,

$$P(D) = 0.83, P(A) = 0.82 \text{ and } P(D \cap A) = 0.78$$

(i) Probability that the plane arrives on time given that it departed on time is

$$P\left(\frac{A}{D}\right) = \frac{P(D \cap A)}{P(D)} = \frac{0.78}{0.83} = 0.9398$$

(ii) Probability that the plane departed on time given that it has arrived on time is

$$P\left(\frac{D}{A}\right) = \frac{P(D \cap A)}{P(A)} = \frac{0.78}{0.82} = 0.9512$$

$$(iii) P\left(\frac{A}{\bar{D}}\right) = \frac{P(A \cap \bar{D})}{P(\bar{D})} = \frac{0.82 - 0.78}{1 - 0.83} = 0.24$$

This is the probability that the flight arrives on time given that it did not depart on time

CONDITIONAL PROBABILITY



Multiplication Rule

Let A and B be two events in sample space.

The **conditional probability** that event A occurs given that event B has occurred and it is denoted by

$$P\left(\frac{A}{B}\right) = \frac{P(A \text{ and } B)}{P(B)} \quad \text{OR} \quad P\left(\frac{A}{B}\right) = \frac{P(A \cap B)}{P(B)}$$

$$\begin{aligned} \text{It can also be written as } P(A \cap B) &= P(B) P(A|B) & P(B) \neq 0 \\ &= P(A) P(B|A) & P(A) \neq 0 \end{aligned}$$

Let A, B and C be three events in a sample space S,

then $P(A \cap B \cap C) = P(A) P(B|A) P(C|A \cap B)$ and it is called **Multiplication Rule**

Multiplication Rule

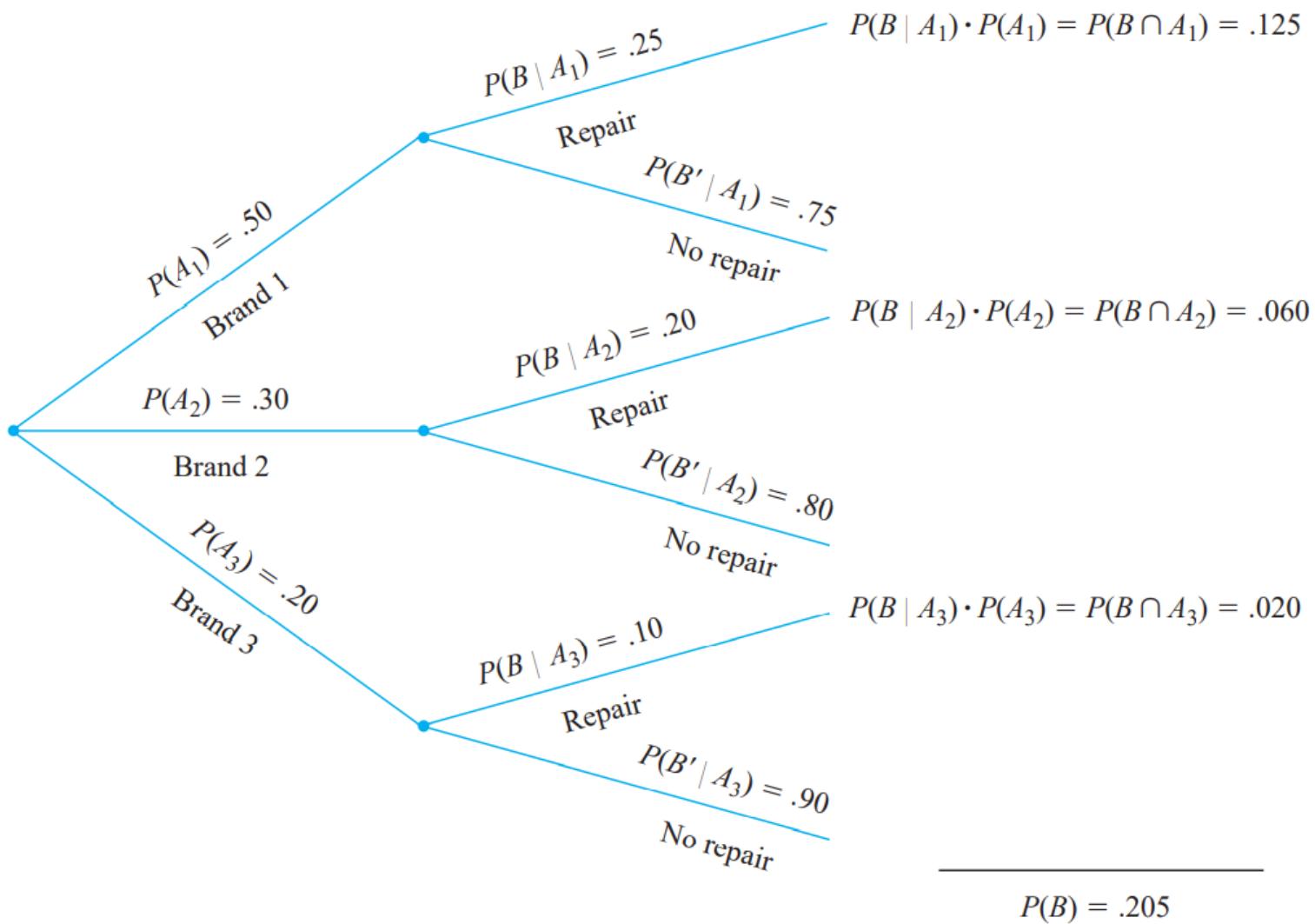
In general, A_1, A_2, \dots, A_n are events in S , then

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) P(A_2 / A_1) P(A_3 / A_1 \cap A_2)$$

$$\dots \dots P(A_n / A_1 \cap A_2 \cap \dots \cap A_{n-1})$$

A chain of video stores sells three different brands of DVD players. Of its DVD player sales, 50% are brand 1 (the least expensive), 30% are brand 2, and 20% are brand 3. Each manufacturer offers a 1-year warranty on parts and labor. It is known that 25% of brand 1's DVD players require warranty repair work, whereas the corresponding percentages for brands 2 and 3 are 20% and 10%, respectively.

1. What is the probability that a randomly selected purchaser has bought a brand 1 DVD player that will need repair while under warranty?
2. What is the probability that a randomly selected purchaser has a DVD player that will need repair while under warranty?
3. If a customer returns to the store with a DVD player that needs warranty repair work, what is the probability that it is a brand 1 DVD player? A brand 2 DVD player? A brand 3 DVD player?



1. What is the probability that a randomly selected purchaser has bought a brand 1 DVD player that will need repair while under warranty?

$$P(A_1 \cap B) = P(B|A_1) \cdot P(A_1) = .125$$

2. What is the probability that a randomly selected purchaser has a DVD player that will need repair while under warranty?

$$\begin{aligned}P(B) &= P[(\text{brand 1 and repair}) \text{ or } (\text{brand 2 and repair}) \text{ or } (\text{brand 3 and repair})] \\&= P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B) \\&= .125 + .060 + .020 = .205\end{aligned}$$

3. If a customer returns to the store with a DVD player that needs warranty repair work, what is the probability that it is a brand 1 DVD player? A brand 2 DVD player? A brand 3 DVD player?

$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{.125}{.205} = .61$$

$$P(A_2|B) = \frac{P(A_2 \cap B)}{P(B)} = \frac{.060}{.205} = .29$$

$$P(A_3|B) = 1 - P(A_1|B) - P(A_2|B) = .10$$

INDEPENDENT EVENTS

We can deduce an important result from the conditional probability:

If B has no effect on A, then, $P\left(\frac{A}{B}\right) = P(A)$ Also $P\left(\frac{B}{A}\right) = P(B)$ and we say the events are independent.

i.e., The probability of A does not depend on B.

so,
$$P\left(\frac{A}{B}\right) = \frac{P(A \cap B)}{P(B)}$$

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

or
$$P(A \cap B) = P(A) \times P(B)$$

Examples on Independent Events

A box contains 20 fuses of which 5 are defective. If two fuses are chosen at random one after the other. What is probability that both the fuses are defective if (i) the first fuse is replaced, (ii) the first fuse is not replaced.

Solution: Let A be the event that the first fuse is defective and
B be the event that the second fuse is defective

(i) When the first fuse is replaced, the events are independent hence

$$P(A \cap B) = P(A) \times P(B) = \frac{5C_1}{20C_1} \times \frac{5C_1}{20C_1} = \frac{1}{4} \times \frac{1}{4} = \frac{1}{16}$$

(ii) When first fuse is not replaced, the events are not independent then

$$P(B \cap A) = P(A) \times P\left(\frac{B}{A}\right) = \frac{5C_1}{20C_1} \times \frac{4C_1}{19C_1} = \frac{1}{19}$$

Examples on Independent Events

A problem in statistics is given to 3 students A,B,C. Their chances of solving it are $1/2, 1/3, 1/4$. Find the probability that the problem is solved.

Solution: Problem can be solved by either A or B or C

Therefore we have to use

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

Using complement of the event i.e the problem is not solved .

Therefore $P(\text{problem solved is }) = 1 - P(\text{not solved})$

$$= 1 - P(\bar{A} \cap \bar{B} \cap \bar{C})$$

$$= 1 - P(\bar{A}) P(\bar{B}) P(\bar{C})$$

$$= 1 - \left[1 - \frac{1}{2}\right] \left[1 - \frac{1}{3}\right] \left[1 - \frac{1}{4}\right]$$

$$= \frac{1}{4}$$

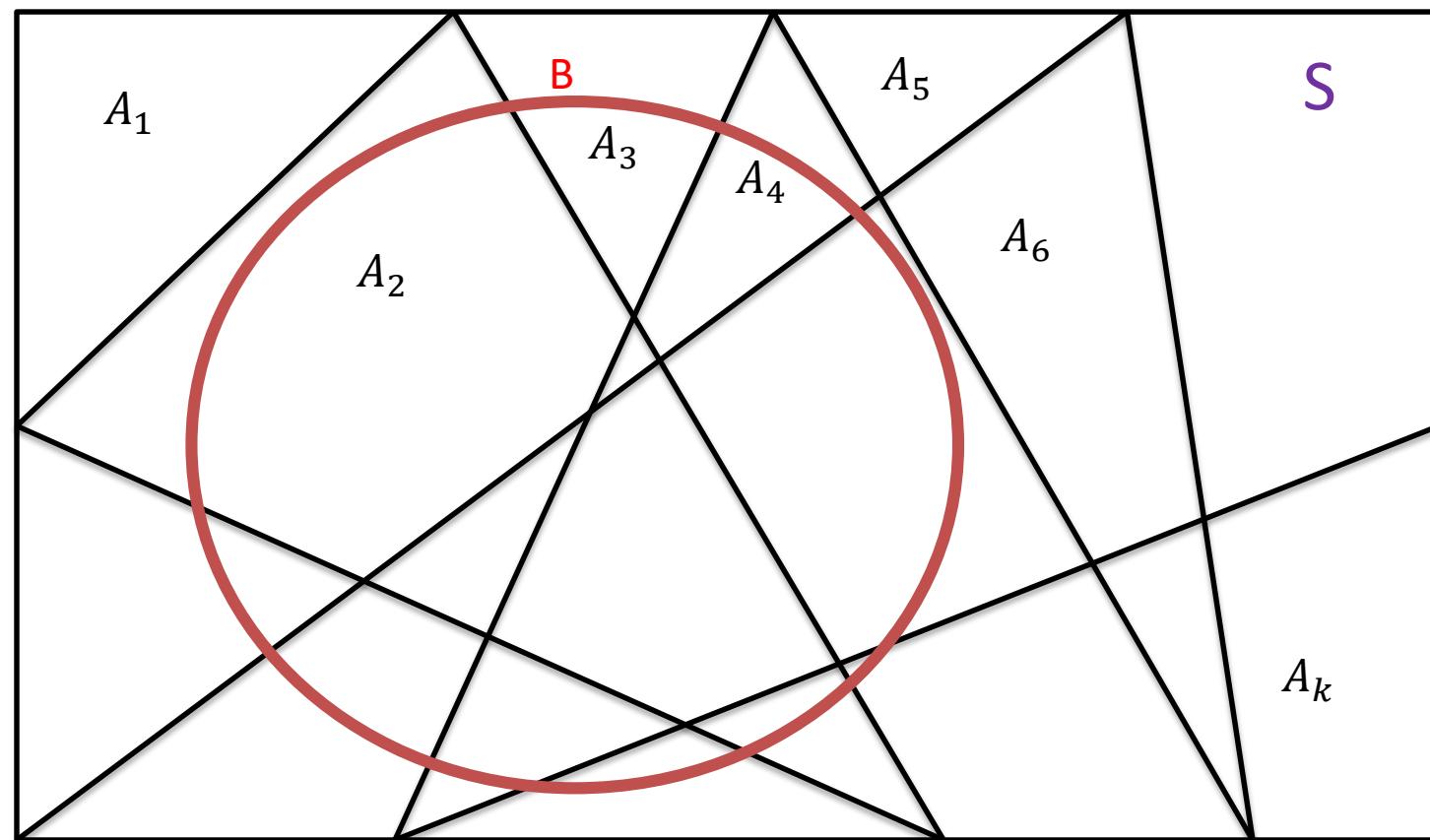
**Note: If A, B,C are independent
then $\bar{A}, \bar{B}, \bar{C}$ are also independent.**

The Law of Total Probability

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

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

$$S = A_1 \cup A_2 \cup \dots \cup A_n \quad \text{and} \quad A_1 \cap A_2 \cap \dots \cap A_n = \emptyset$$



$$\therefore B = B \cap S = B \cap \{A_1 \cup A_2 \cup A_3, \dots, \cup A_n\}$$

Proof:

We have $S = \{A_1 \cup A_2 \cup A_3 \cup \dots \cup A_n\}$ and $A \subset S$

$$\therefore B = B \cap S = B \cap \{A_1 \cup A_2 \cup A_3, \dots \cup A_n\}$$

Using distributive law in the R.H.S, we get

Since $B \cap A_i$ ($i = 1$ to n) are mutually exclusive, we have by applying addition rule of probability,

$$P(B) = P(B \cap A_1) + P(B \cap A_2) + \dots + P(B \cap A_n)$$

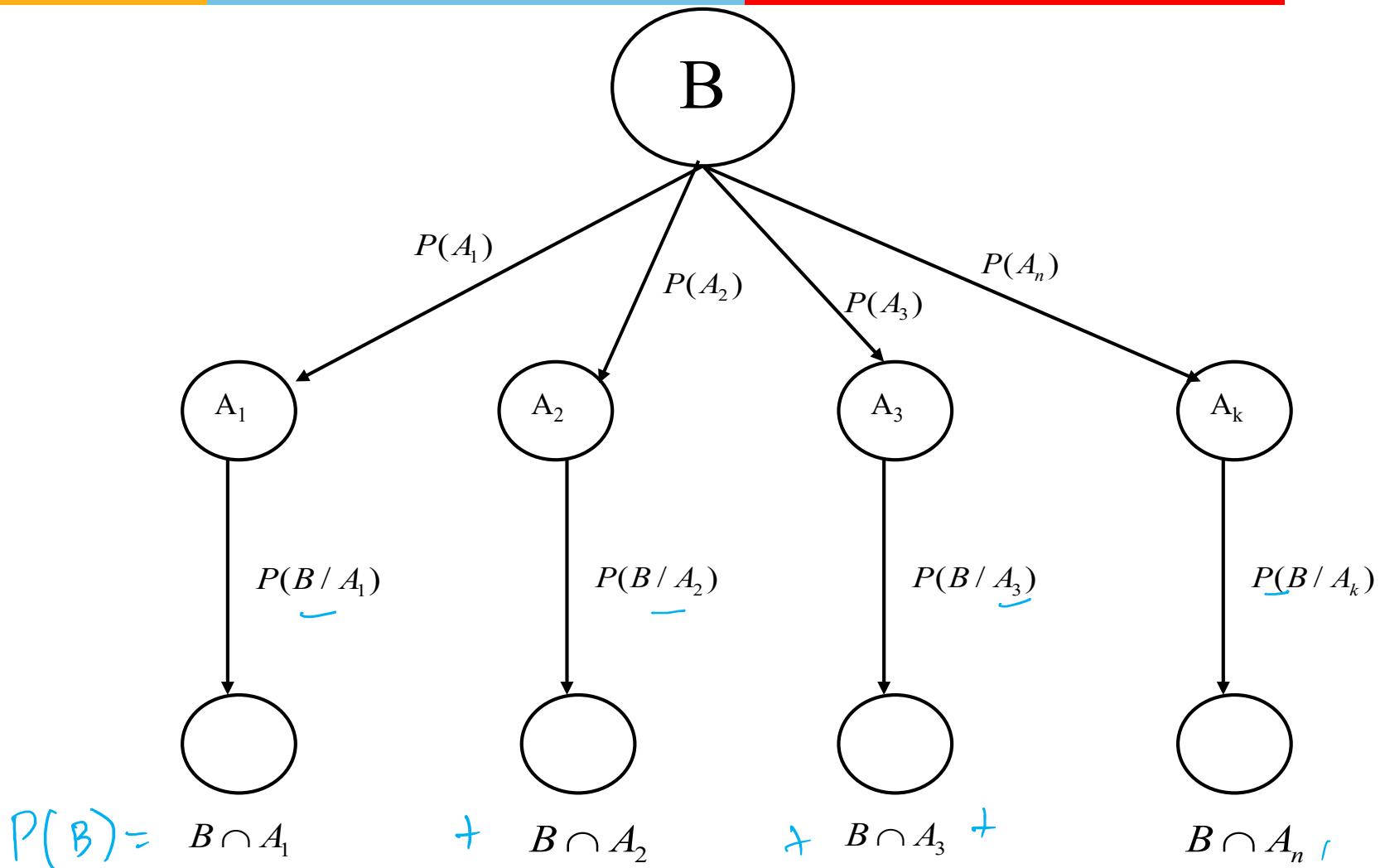
i.e., $P(B) = \sum_{i=1}^{i=n} P(B \cap A_i)$

Using multiplication rule on each term on R.H.S, namely

$$P(B \cap A_i) = P(A_i) \cdot P(B | A_i) \quad (1)$$

$$P(B) = \sum_{i=1}^{i=n} P(A_i) P(B | A_i) \quad (2)-\text{Total theorem on Probability}$$

The Theorem of Total Probability (tree diagram)



Law of Total Probability

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 message is spam?

$A_i = \{\text{message is from account } \# i\}$ for $i = 1, 2, 3$, $B = \{\text{message is spam}\}$
Then the given percentages imply that

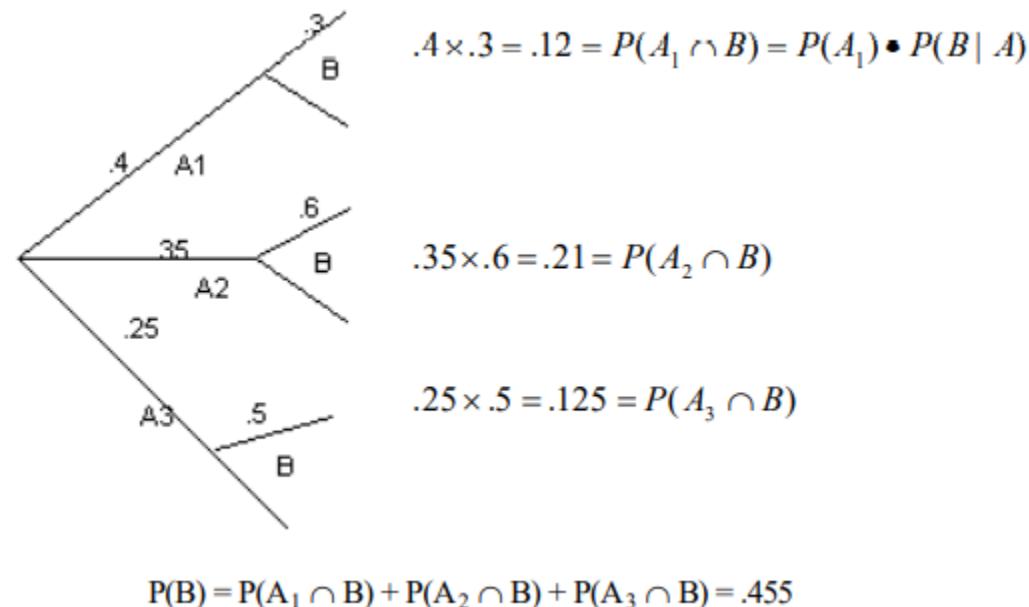
$$P(A_1) = .70, P(A_2) = .20, P(A_3) = .10$$

$$P(B|A_1) = .01, P(B|A_2) = .02, P(B|A_3) = .05$$

Now it is simply a matter of substituting into the equation for the law of total probability: $P(B) = (.01)(.70) + (.02)(.20) + (.05)(.10) = .016$

Law of Total Probability

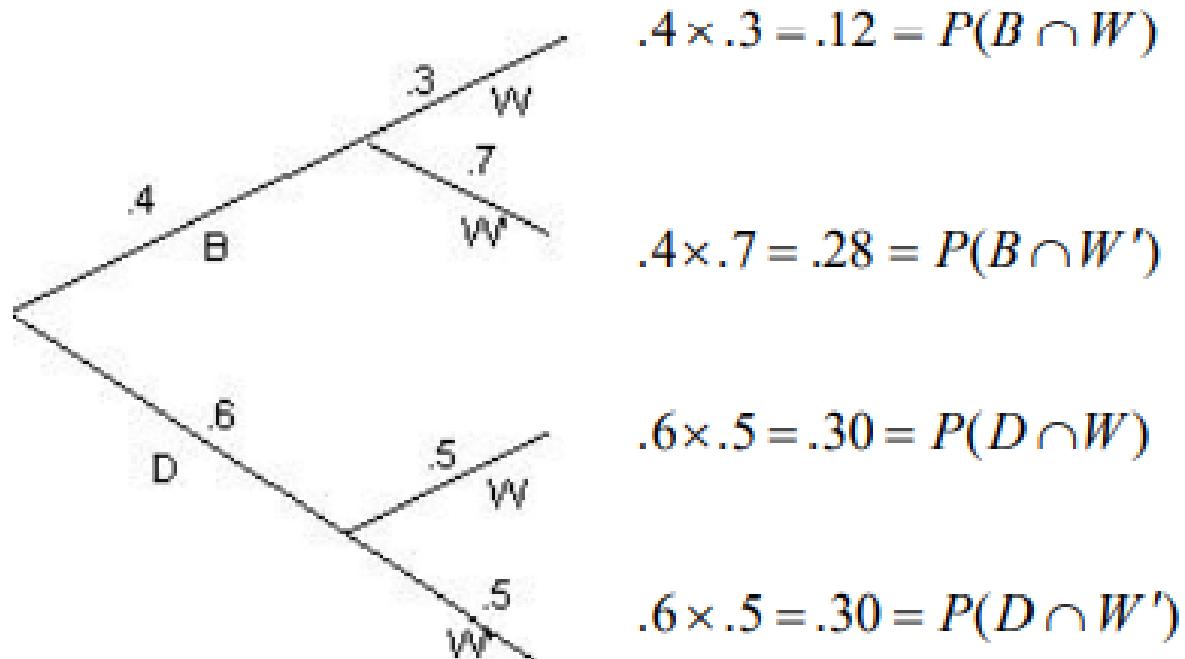
At a certain gas station, 40% of the customers use regular gas (A1), 35% use plus gas (A2), and 25% use premium (A3). Of those customers using regular gas, only 30% fill their tanks (event B). Of those customers using plus, 60% fill their tanks, whereas of those using premium, 50% fill their tanks. What is the probability that the next customer fills the tank?



Law of Total Probability

A company that manufactures video cameras produces a basic model and a deluxe model. Over the past year, 40% of the cameras sold have been of the basic model. Of those buying the basic model, 30% purchase an extended warranty, whereas 50% of all deluxe purchasers do so. What is the probability that that a randomly selected purchaser has an extended warranty?

Using a tree diagram, B = basic, D = deluxe, W = warranty purchase, W' = no warranty



We want $P(W) = .30 + .12 = .42$

HW: Exercise

The population of a particular country consists of three ethnic groups. Each individual belongs to one of the four major blood groups. The accompanying *joint probability table* gives the proportions of individuals in the various ethnic group–blood group combinations.

		Blood Group			
		O	A	B	AB
Ethnic Group	1	.082	.106	.008	.004
	2	.135	.141	.018	.006
	3	.215	.200	.065	.020

Suppose that an individual is randomly selected from the population, and define events by $A = \{\text{type A selected}\}$, $B = \{\text{type B selected}\}$, and $C = \{\text{ethnic group 3 selected}\}$.

- Calculate $P(A)$, $P(C)$, and $P(A \cap C)$.
- Calculate both $P(A|C)$ and $P(C|A)$, and explain in context what each of these probabilities represents.
- If the selected individual does not have type B blood, what is the probability that he or she is from ethnic group 1?

HW: Exercise

If A and B are two events with $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{3}$ and $P(A \cap B) = \frac{1}{4}$

Find

$$P\left(\frac{A}{B}\right), P\left(\frac{B}{A}\right), P\left(\frac{\bar{A}}{\bar{B}}\right), P\left(\frac{\bar{B}}{\bar{A}}\right), P\left(\frac{A}{\bar{B}}\right)$$

One card is randomly collected from the deck of 52 cards.

- What is the probability that this card is a heart?
- What is the probability that this card is not a heart?
- What is the probability that it is a heart and a king?
- What is the probability that the card is a heart or a king?
- Are the events that the card is a heart and is a king independent?

A card is randomly drawn from an incomplete deck of cards from which the ace of diamonds is missing.

1. What is the probability that the card is “clubs”?
2. What is the probability that the card is a “queen”?
3. Are the events “clubs” and “queen” independent?

In a group of children from primary school there are 18 girls and 15 boys. Of the girls, 9 have had measles. Of the boys, 6 have had measles.

1. What is the probability that a randomly chosen child from this group has had measles?
2. If we randomly choose one person from the group of 18 girls, what is the probability that this girl has had measles?
3. Are the events “boy” and “measles” in this example independent?

In a Japanese cohort study, 5,322 male non-smokers and 7,019 male smokers were followed for four years. Of these men, 16 non-smokers and 77 smokers developed lung cancer.

1. What is the probability that a randomly chosen non-smoker from this group developed lung cancer?
2. What is the probability that a randomly chosen smoker from this group developed lung cancer?
3. Are the events “smoking” and “lung cancer” in this example independent?
4. What is the conditional probability that the patient is a smoker if he has developed lung cancer?



Thanks

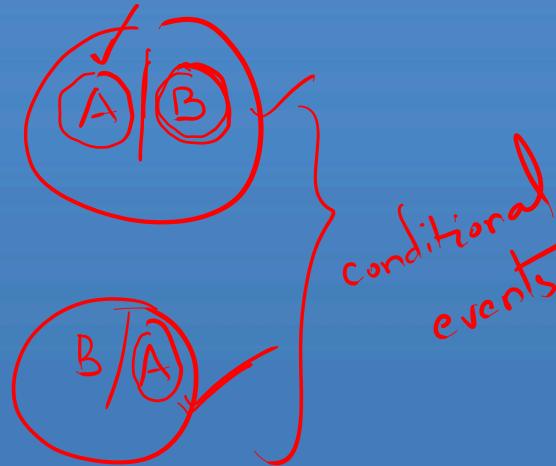


BITS Pilani

Pilani|Dubai|Goa|Hyderabad

M.Tech.(Data Science & Engineering) Introduction to Statistical Methods

Team ISM



A, C, B, D

C/A ✓

B/(A,C) ✓

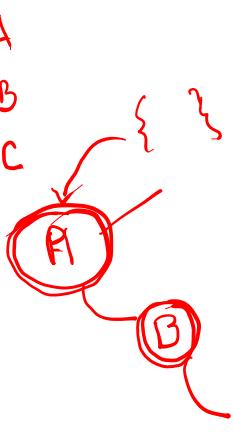
D/(A,B,C) ✓

B/(A,C)

D/(A,B,C)

Session No 3

Introduction to Conditional Probability, independent events, Total



B/A

C/A,B

C/A \cap B

Probability

(Session 3: 26th/27th Nov 2022)

Buying ^{new} / Amazon
visit website

B/D ✓

A → buying mobile phone ✓
'B/A' →

Contact Session 3

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

$$P(A|\underline{B}) = P(A)$$

$= P(A) \rightarrow$ independent even

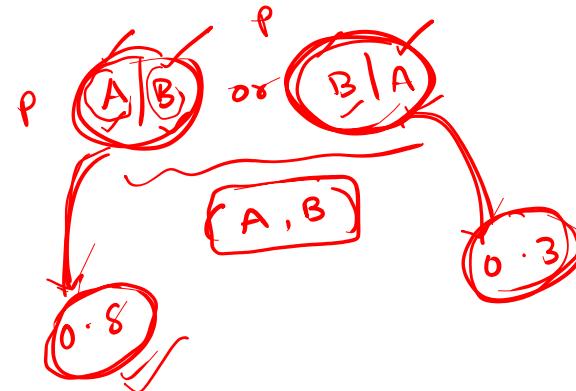
$$P(B|\underline{A}) = P(B) \rightarrow$$
 indep
events

Contact Session 3: Module 2(Conditional Probability & Bayes theorem)

Contact Session	List of Topic Title	Reference
CS - 3	Introduction to conditional probability, indepents events, Total probability	T1 & T2
HW	Problems on conditional probability	T1 & T2
Lab		

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

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



Agenda

-
- Conditional Probability
 - Independent events
 - Total Probability

Text Books

No	Author(s), Title, Edition, Publishing House
T1	Statistics for Data Scientists, An introduction to probability, statistics and Data Analysis, Maurits Kaptein et al, Springer 2022
T2	Probability and Statistics for Engineering and Sciences, 8 th Edition, Jay L Devore, Cengage Learning

CONDITIONAL PROBABILITY

The probabilities assigned to various events depend on what is known about the experimental situation when the assignment is made. Subsequent to the initial assignment, partial information relevant to the outcome of the experiment may become available. Such information may cause us to revise some of our probability assignments

CONDITIONAL PROBABILITY

We examine how the information “an event B has occurred” affects the probability assigned to A. For example, A might refer to an individual having a particular disease in the presence of certain symptoms. If a blood test is performed on the individual and the result is negative , then the probability of having the disease will change (it should decrease, but not usually to zero, since blood tests are not infallible). We will use the notation to represent the conditional probability of A given that the event B has occurred. B is the “conditioning event.”

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

$$P(B) \neq 0$$

$$P(B) = 0$$

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

Conditional probabilities in a 2×2 contingency table

	A	A^c	
B	$\Pr(A \cap B) = \Pr(A B) \Pr(B)$	$\Pr(A^c \cap B) = \Pr(A^c B) \Pr(B)$	$\Pr(B)$
B^c	$\Pr(A \cap B^c) = \Pr(A B^c) \Pr(B^c)$	$\Pr(A^c \cap B^c) = \Pr(A^c B^c) \Pr(B^c)$	$\Pr(B^c)$
	$\Pr(A)$	$\Pr(A^c)$	1

Examples on Conditional Probability

A. well planned
B. well built

Example: In a housing colony, 70% of the houses are well planned and 60% of the houses are well planned and well built. Find the probability that an arbitrarily chosen house in this colony is well built given that it is well planned.

Solution: Let A be the event that the house is well planned

$P(A \cap B) = \frac{60}{100}$ B be the event that the house is well built

therefore $P(A) = 0.70$, $P(A \cap B) = 0.60$

$$P\left(\frac{B}{A}\right) = \frac{P(B \cap A)}{P(A)} = \frac{0.60}{0.70} = 0.8571$$

CONDITIONAL PROBABILITY

Example:

If the probability that a communication system will have high fidelity is 0.81 and the probability that it will have high fidelity and high selectivity is 0.18, what is probability that a system with high fidelity will also have high selectivity?

Solution:

If A is the event that a communication system has high selectivity and B is the event that it has high fidelity, we have $P(B)=0.81$ and $P(A \cap B)=0.18$, and substitution into the formula yields

$$P(B) = 0.81$$

$$P(A|B) = \frac{0.18}{0.81} = \frac{2}{9}$$

CONDITIONAL PROBABILITY



Example:

Verify whether the events A & B are independent or not
 $P(A \cap B) = P(A)P(B)$

(Suppose that of all individuals buying a certain digital camera, 60% include an optional memory card in their purchase, 40% include an extra battery, and 30% include both a card and battery.)

- Given that the randomly selected individual purchased an extra battery, find the probability that an optional card was also purchased.
- Given that the randomly selected individual purchased an optional card, find the probability that an extra battery was also purchased.

Solution:

let $A = \{ \text{memory card purchased} \}$ and $B = \{ \text{battery purchased} \}$. Then $P(A) = .60$, $P(B) = .40$, and $P(\text{both purchased}) = P(A \cap B) = .30$.

$$a) P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{.30}{.40} = .75 \quad P(A)$$

$$b) P(B|A) = \frac{P(A \cap B)}{P(A)} = \frac{.30}{.60} = .50 \quad P(B)$$

Notice that $P(A|B) \neq P(A)$ and $P(B|A) \neq P(B)$.

Example:

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

<i>Read regularly</i>	A	B	C	$A \cap B$	$A \cap C$	$B \cap C$	$A \cap B \cap C$
<i>Probability</i>	.14	.23	.37	.08	.09	.13	.05

Find $P(A|B)$, $P(A|B \cup C)$, $P(A|$ reads at least one)

Solution: We thus have

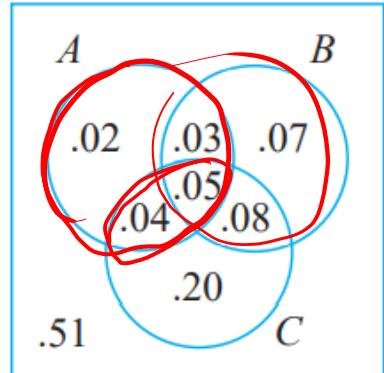
$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{.08}{.23} = .348$$

$$P(A|B \cup C) = \frac{P(A \cap (B \cup C))}{P(B \cup C)} = \frac{.04 + .05 + .03}{.47} = \frac{.12}{.47} = .255$$

$$\begin{aligned} P(A|\text{reads at least one}) &= P(A|A \cup B \cup C) = \frac{P(A \cap (A \cup B \cup C))}{P(A \cup B \cup C)} \\ &= \frac{P(A)}{P(A \cup B \cup C)} = \frac{.14}{.49} = .286 \end{aligned}$$

and

$$P(A \cup B|C) = \frac{P((A \cup B) \cap C)}{P(C)} = \frac{.04 + .05 + .08}{.37} = .459$$



Examples on Conditional Probability

Example: In a certain college, 25% of the students failed Maths, 15% of the students failed chemistry and 10% of the students failed both maths and chemistry. A student is selected at random, find

- a. If he failed chemistry, what is the probability that he failed Maths? $P(M) = 0.25, P(C) = 0.15, P(M \cap C) = 0.1$

$$P\left(\frac{M}{C}\right) = \frac{P(M \cap C)}{P(C)} = \frac{0.1}{0.15} = 0.6667$$

- b. If he failed maths, what is the probability he failed chemistry?

$$P\left(\frac{C}{M}\right) = \frac{P(C \cap M)}{P(M)} = \frac{0.1}{0.25} = 0.4$$

- c. What is the probability that the student failed in Maths or chemistry?

$$P(M \cup C) = P(M) + P(C) - P(M \cap C)$$

Examples on Conditional Probability

$$\checkmark \quad P(\bar{A}) = 1 - P(A) \quad \checkmark$$

$$\checkmark \quad P(\bar{A}/B) = 1 - P(A/B) \quad \checkmark \quad \text{LHS} \quad \frac{P(\bar{A} \cap B)}{P(B)} \quad \checkmark$$

Example : The probabilities of a regularly scheduled flight departs on time is 0.83, arrives on time is 0.82 & it departs and arrives on time is 0.78. Find the probability that a plane (i) arrives on time given that it departed on time, (ii) departed on time given that it has arrived on time and (iii) find $P\left(\frac{A}{D}\right)$

$$\text{RHS} = 1 - \frac{P(A \cap B)}{P(B)}$$

Ans: Let D and A be the events that the flight departs and arrives on time respectively. Then,

$$P(D) = 0.83, P(A) = 0.82 \text{ and } P(D \cap A) = 0.78$$

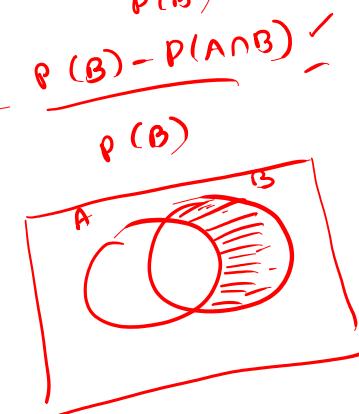
(i) Probability that the plane arrives on time given that it departed on time is

$$P\left(\frac{A}{D}\right) = \frac{P(D \cap A)}{P(D)} = \frac{0.78}{0.83} = 0.9398$$

(ii) Probability that the plane departed on time given that it has arrived on time is

$$P\left(\frac{D}{A}\right) = \frac{P(D \cap A)}{P(A)} = \frac{0.78}{0.82} = 0.9512$$

$$(iii) P\left(\frac{A}{\bar{D}}\right) = \frac{P(A \cap \bar{D})}{P(\bar{D})} = \frac{0.82 - 0.78}{1 - 0.83} = 0.24$$



This is the probability that the flight arrives on time given that it did not depart on time

CONDITIONAL PROBABILITY



Multiplication Rule

Let A and B be two events in sample space.

The **conditional probability** that event A occurs given that event B has occurred and it is denoted by

$$P\left(\frac{A}{B}\right) = \frac{P(A \text{ and } B)}{P(B)}$$
 OR $P\left(\frac{B}{A}\right) = \frac{P(A \cap B)}{P(A)}$

It can also be written as $P(A \cap B) = P(B) P(A/B)$
 $= P(A) P(B/A)$

$P(B) \neq 0$
 $P(A) \neq 0$

$$\begin{aligned} P(A \cap B) &= P(B|A) P(A) \\ &= P(A|B) P(B) \\ P(x,y) &= P(Y|x) P(x) \\ &= P(x|Y) P(Y) \end{aligned}$$

Let A, B and C be three events in a sample space S, rule
multiplication rule

then $P(A \cap B \cap C) = P(A) P(B/A) P(C/A \cap B)$ and it is called **Multiplication Rule**
 $P(x,y,z) = P(z/x,y) \cdot P(y/x)$

Multiplication Rule

In general, A_1, A_2, \dots, A_n are events in S , then

$$P(A_1 \cap A_2 \cap \dots \cap A_n) = P(A_1) P(A_2 / A_1) P(A_3 / A_1 \cap A_2)$$

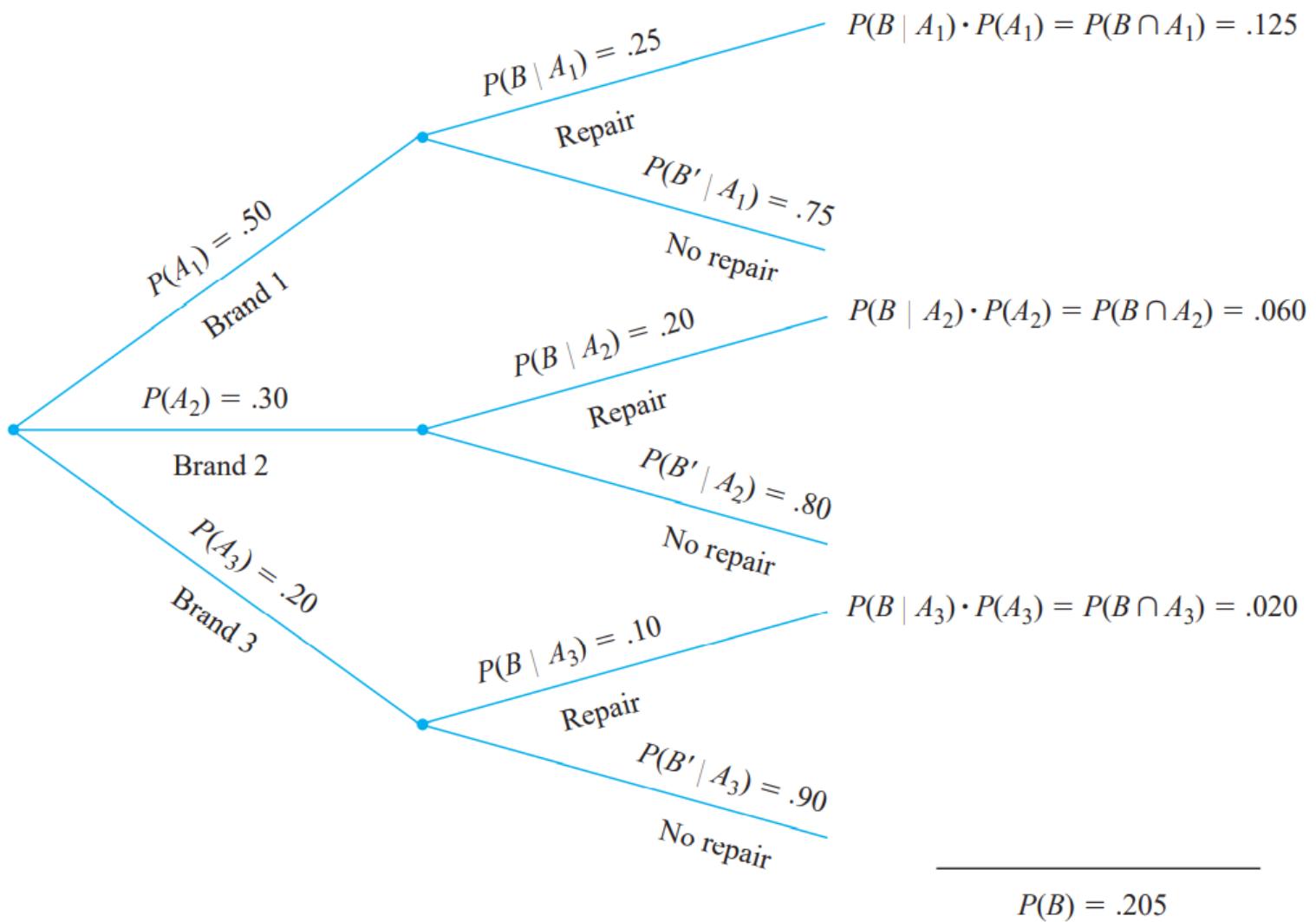


$$\dots \dots P(A_n / A_1 \cap A_2 \cap \dots \cap A_{n-1})$$

Postpone

A chain of video stores sells three different brands of DVD players. Of its DVD player sales, 50% are brand 1 (the least expensive), 30% are brand 2, and 20% are brand 3. Each manufacturer offers a 1-year warranty on parts and labor. It is known that 25% of brand 1's DVD players require warranty repair work, whereas the corresponding percentages for brands 2 and 3 are 20% and 10%, respectively.

1. What is the probability that a randomly selected purchaser has bought a brand 1 DVD player that will need repair while under warranty?
2. What is the probability that a randomly selected purchaser has a DVD player that will need repair while under warranty?
3. If a customer returns to the store with a DVD player that needs warranty repair work, what is the probability that it is a brand 1 DVD player? A brand 2 DVD player? A brand 3 DVD player?



1. What is the probability that a randomly selected purchaser has bought a brand 1 DVD player that will need repair while under warranty?

$$P(A_1 \cap B) = P(B|A_1) \cdot P(A_1) = .125$$

2. What is the probability that a randomly selected purchaser has a DVD player that will need repair while under warranty?

$$\begin{aligned}P(B) &= P[(\text{brand 1 and repair}) \text{ or } (\text{brand 2 and repair}) \text{ or } (\text{brand 3 and repair})] \\&= P(A_1 \cap B) + P(A_2 \cap B) + P(A_3 \cap B) \\&= .125 + .060 + .020 = .205\end{aligned}$$

3. If a customer returns to the store with a DVD player that needs warranty repair work, what is the probability that it is a brand 1 DVD player? A brand 2 DVD player? A brand 3 DVD player?

$$P(A_1|B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{.125}{.205} = .61$$

$$P(A_2|B) = \frac{P(A_2 \cap B)}{P(B)} = \frac{.060}{.205} = .29$$

$$P(A_3|B) = 1 - P(A_1|B) - P(A_2|B) = .10$$

INDEPENDENT EVENTS

We can deduce an important result from the conditional probability:

If B has no effect on A, then, $P\left(\frac{A}{B}\right) = P(A)$ Also $P\left(\frac{B}{A}\right) = P(B)$ and we say the events are independent.

i.e., The probability of A does not depend on B.

so,
$$P\left(\frac{A}{B}\right) = \frac{P(A \cap B)}{P(B)}$$

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

or
$$P(A \cap B) = P(A) \times P(B)$$

Examples on Independent Events

A box contains 20 fuses of which 5 are defective. If two fuses are chosen at random one after the other. What is probability that both the fuses are defective if (i) the first fuse is replaced, (ii) the first fuse is not replaced.

Solution: Let A be the event that the first fuse is defective and
B be the event that the second fuse is defective

(i) When the first fuse is replaced, the events are independent hence

$$P(A \cap B) = P(A) \times P(B) = \frac{5C_1}{20C_1} \times \frac{5C_1}{20C_1} = \frac{1}{4} \times \frac{1}{4} = \frac{1}{16}$$

(ii) When first fuse is not replaced, the events are not independent then

$$P(B \cap A) = P(A) \times P\left(\frac{B}{A}\right) = \frac{5C_1}{20C_1} \times \frac{4C_1}{19C_1} = \frac{1}{19}$$

Examples on Independent Events

A problem in statistics is given to 3 students A,B,C. Their chances of solving it are $1/2, 1/3, 1/4$. Find the probability that the problem is solved.

Solution: Problem can be solved by either A or B or C

Therefore we have to use

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

Using complement of the event i.e the problem is not solved .

Therefore $P(\text{problem solved is }) = 1 - P(\text{not solved})$

$$= 1 - P(\bar{A} \cap \bar{B} \cap \bar{C})$$

$$= 1 - P(\bar{A}) P(\bar{B}) P(\bar{C})$$

$$= 1 - \left[1 - \frac{1}{2}\right] \left[1 - \frac{1}{3}\right] \left[1 - \frac{1}{4}\right]$$

$$= \frac{1}{4}$$

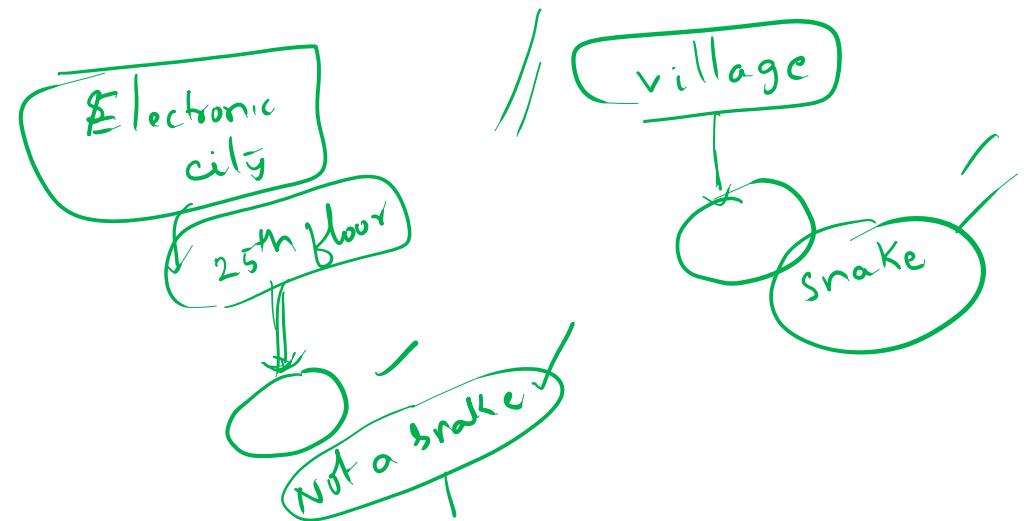
**Note: If A, B,C are independent
then $\bar{A}, \bar{B}, \bar{C}$ are also independent.**

The Law of Total Probability

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

$$P(B) = P(B|A_1)P(A_1) + \dots + P(B|A_k)P(A_k)$$

$$= \sum_{i=1}^k P(B|A_i)P(A_i)$$

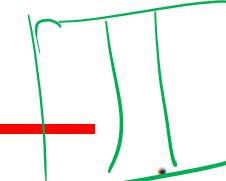


Law of Total Probability

$$P(A) = \frac{70}{100} \checkmark \quad P(S|A) = \frac{1}{100}$$

$$P(B) = \frac{20}{100} \checkmark \quad P(S|B) = \frac{2}{100}$$

$$P(C) = \frac{10}{100} \checkmark \quad P(S|C) = \frac{5}{100}$$



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 message is spam?

$$A_i = \{\text{message is from account } \# i\} \text{ for } i = 1, 2, 3, \checkmark \quad B = \{\text{message is spam}\}$$

Then the given percentages imply that

$$\begin{aligned} S &= (S \cap A) \cup (S \cap B) \cup (S \cap C) \\ P(S) &= P(S \cap A) + P(S \cap B) + P(S \cap C) \\ &= P(S|A)P(A) + P(S|B)P(B) \\ &\quad + P(S|C)P(C) \end{aligned}$$

Now it is simply a matter of substituting into the equation for the law of total probability:

$$P(B) = (.01)(.70) + (.02)(.20) + (.05)(.10) = .016 = \frac{70}{100} + \frac{1}{100} + \frac{20}{100} \cdot \frac{2}{100} + \frac{10}{100} \cdot \frac{5}{100} \checkmark$$

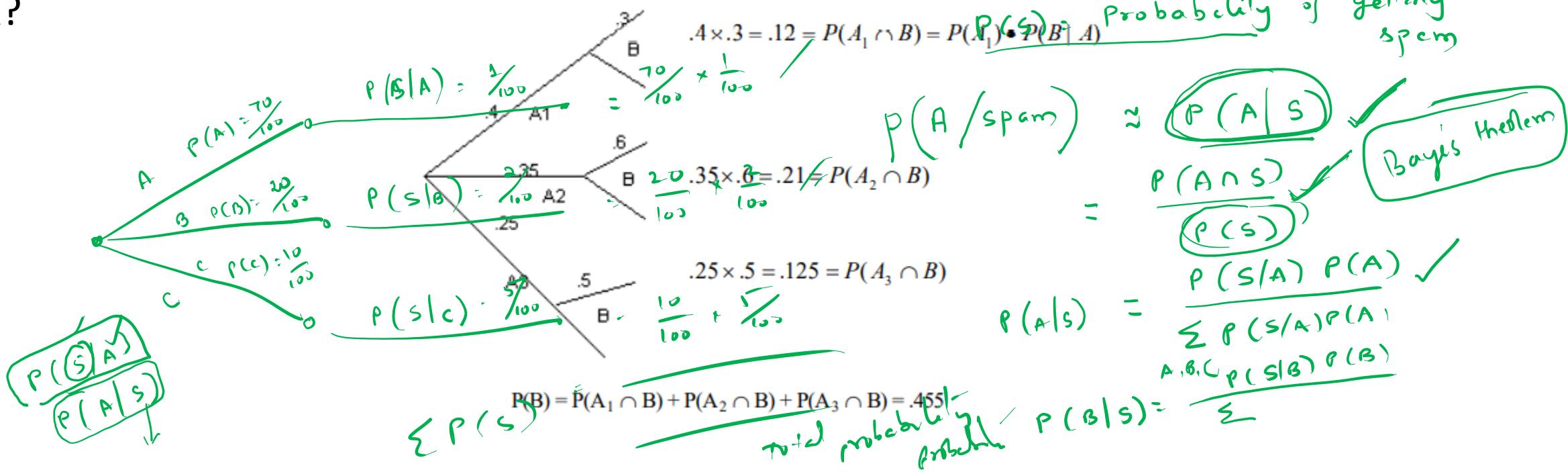
Law of Total Probability

$$P(A|S) = \frac{P(S|A) P(A)}{\sum}$$

✓ B/A ✓
S/A,B,C -> Bayes theorem

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

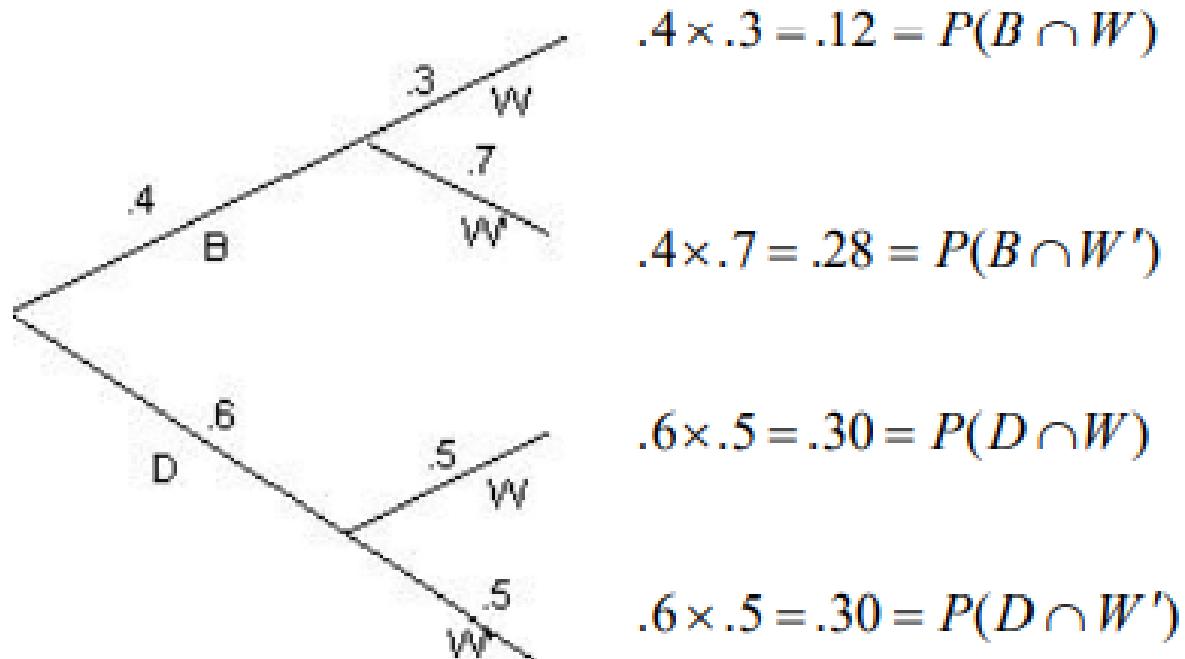
At a certain gas station, 40% of the customers use regular gas (A1), 35% use plus gas (A2), and 25% use premium (A3). Of those customers using regular gas, only 30% fill their tanks (event B). Of those customers using plus, 60% fill their tanks, whereas of those using premium, 50% fill their tanks. What is the probability that the next customer fills the tank?



Law of Total Probability

A company that manufactures video cameras produces a basic model and a deluxe model. Over the past year, 40% of the cameras sold have been of the basic model. Of those buying the basic model, 30% purchase an extended warranty, whereas 50% of all deluxe purchasers do so. What is the probability that that a randomly selected purchaser has an extended warranty? ✓

Using a tree diagram, B = basic, D = deluxe, W = warranty purchase, W' = no warranty



We want $P(W) = .30 + .12 = .42$

HW: Exercise

The population of a particular country consists of three ethnic groups. Each individual belongs to one of the four major blood groups. The accompanying *joint probability table* gives the proportions of individuals in the various ethnic group–blood group combinations.

		Blood Group			
		O	A	B	AB
Ethnic Group	1	.082	.106	.008	.004
	2	.135	.141	.018	.006
	3	.215	.200	.065	.020

Suppose that an individual is randomly selected from the population, and define events by $A = \{\text{type A selected}\}$, $B = \{\text{type B selected}\}$, and $C = \{\text{ethnic group 3 selected}\}$.

- Calculate $P(A)$, $P(C)$, and $P(A \cap C)$.
- Calculate both $P(A|C)$ and $P(C|A)$, and explain in context what each of these probabilities represents.
- If the selected individual does not have type B blood, what is the probability that he or she is from ethnic group 1?

HW: Exercise

If A and B are two events with $P(A) = \frac{1}{2}$, $P(B) = \frac{1}{3}$ and $P(A \cap B) = \frac{1}{4}$

Find

$$P\left(\frac{A}{B}\right), P\left(\frac{B}{A}\right), P\left(\frac{\bar{A}}{\bar{B}}\right), P\left(\frac{\bar{B}}{\bar{A}}\right), P\left(\frac{A}{\bar{B}}\right)$$

One card is randomly collected from the deck of 52 cards.

- What is the probability that this card is a heart?
- What is the probability that this card is not a heart?
- What is the probability that it is a heart and a king?
- What is the probability that the card is a heart or a king?
- Are the events that the card is a heart and is a king independent?

A card is randomly drawn from an incomplete deck of cards from which the ace of diamonds is missing.

1. What is the probability that the card is “clubs”?
2. What is the probability that the card is a “queen”?
3. Are the events “clubs” and “queen” independent?

In a group of children from primary school there are 18 girls and 15 boys. Of the girls, 9 have had measles. Of the boys, 6 have had measles.

1. What is the probability that a randomly chosen child from this group has had measles?
2. If we randomly choose one person from the group of 18 girls, what is the probability that this girl has had measles?
3. Are the events “boy” and “measles” in this example independent?

In a Japanese cohort study, 5,322 male non-smokers and 7,019 male smokers were followed for four years. Of these men, 16 non-smokers and 77 smokers developed lung cancer.

1. What is the probability that a randomly chosen non-smoker from this group developed lung cancer?
2. What is the probability that a randomly chosen smoker from this group developed lung cancer?
3. Are the events “smoking” and “lung cancer” in this example independent?
4. What is the conditional probability that the patient is a smoker if he has developed lung cancer?



Thanks



BITS Pilani

Pilani|Dubai|Goa|Hyderabad

M.Tech.(AIML)
Introduction to Statistical Methods

Team ISM



Session No 4

Bayes theorem &

Introduction to Naïve Bayes concept

(Session 4: 3rd/4th Dec 2022)

Contact Session 4

Contact Session	List of Topic Title	Reference
CS - 4	Bayes theorem(with proof),Introduction to Naïve Bayes concept.	T1 & T2
HW	Problems on Bayes theorem	T1 & T2
Lab	Bayes theorem & Naïve Bayes Concept	Lab 2

Agenda

-
- Bayes Theorem
 - Introduction to Naïve Bayes concept

Text Books

No	Author(s), Title, Edition, Publishing House
T1	Statistics for Data Scientists, An introduction to probability, statistics and Data Analysis, Maurits Kaptein et al, Springer 2022
T2	Probability and Statistics for Engineering and Sciences, 8 th Edition, Jay L Devore, Cengage Learning

BAYES' THEOREM

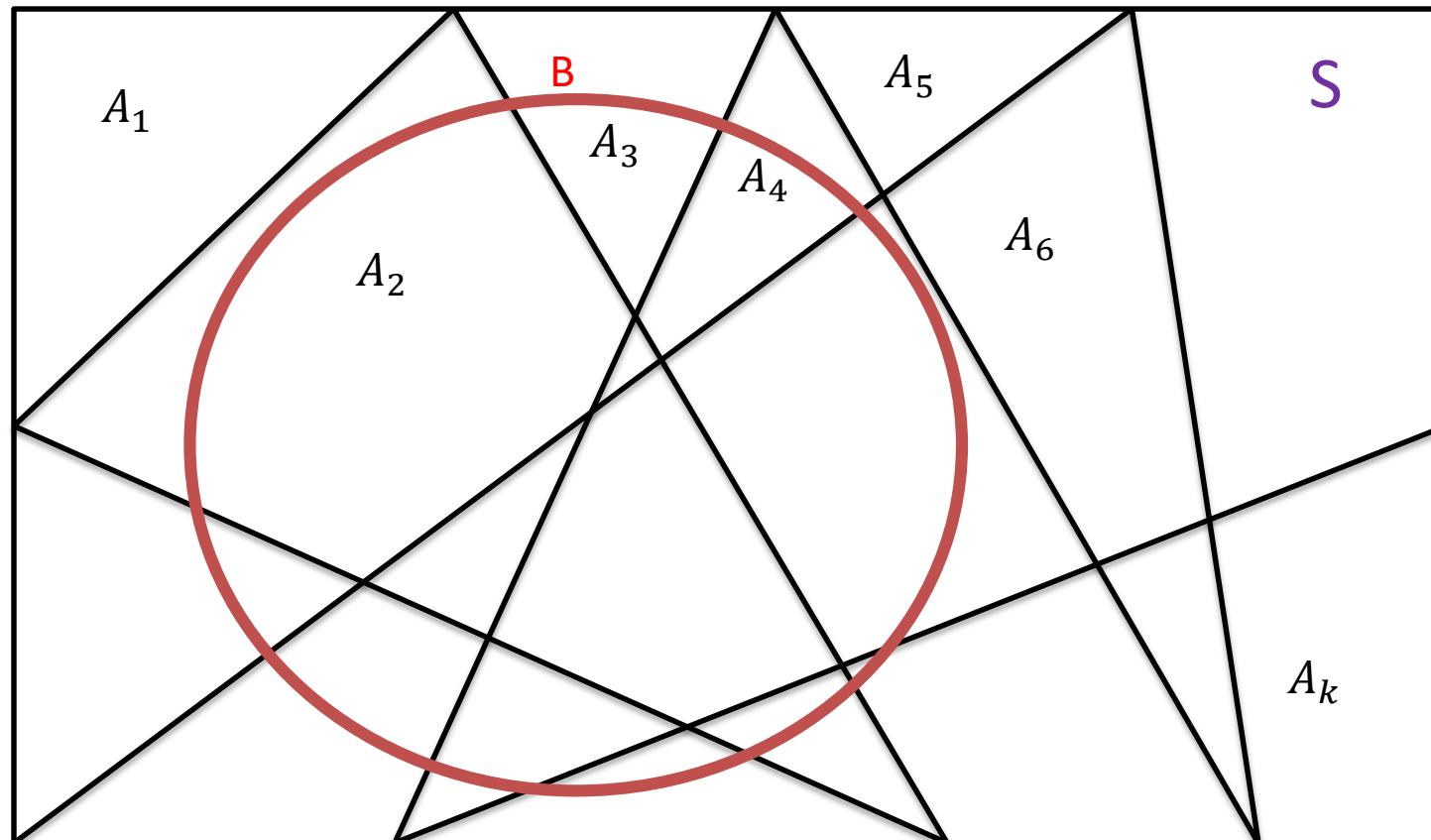
Let $P = \{A_1, A_2, A_3, \dots, A_n\}$ be a set of exhaustive and mutually exclusive events of a sample space S with $P(A_i) \neq 0$

For each i . If B is any other event associated with A_i with $P(B) \neq 0$, then

$$P\left(\frac{A_i}{B}\right) = \frac{P(A_i)P\left(\frac{B}{A_i}\right)}{P(B)}$$
$$= \frac{P(A_i)P\left(\frac{B}{A_i}\right)}{\sum_{i=1}^n P(A_i)P\left(\frac{B}{A_i}\right)}$$

Geometrical Representation of Bayes theorem

$$S = A_1 \cup A_2 \cup \dots \cup A_n \quad \text{and} \quad A_1 \cap A_2 \cap \dots \cap A_n = \emptyset$$



$$\therefore B = B \cap S = B \cap \{A_1 \cup A_2 \cup A_3, \dots \cup A_n\}$$

Proof:

The conditional Probability of A_i for any i given B is given as

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

But we know that

$$P(A_i \cap B) = P(A_i) \cdot P(B | A_i) \text{ & } P(B) = \sum_{i=1}^{i=n} P(A_i) P(B | A_i)$$

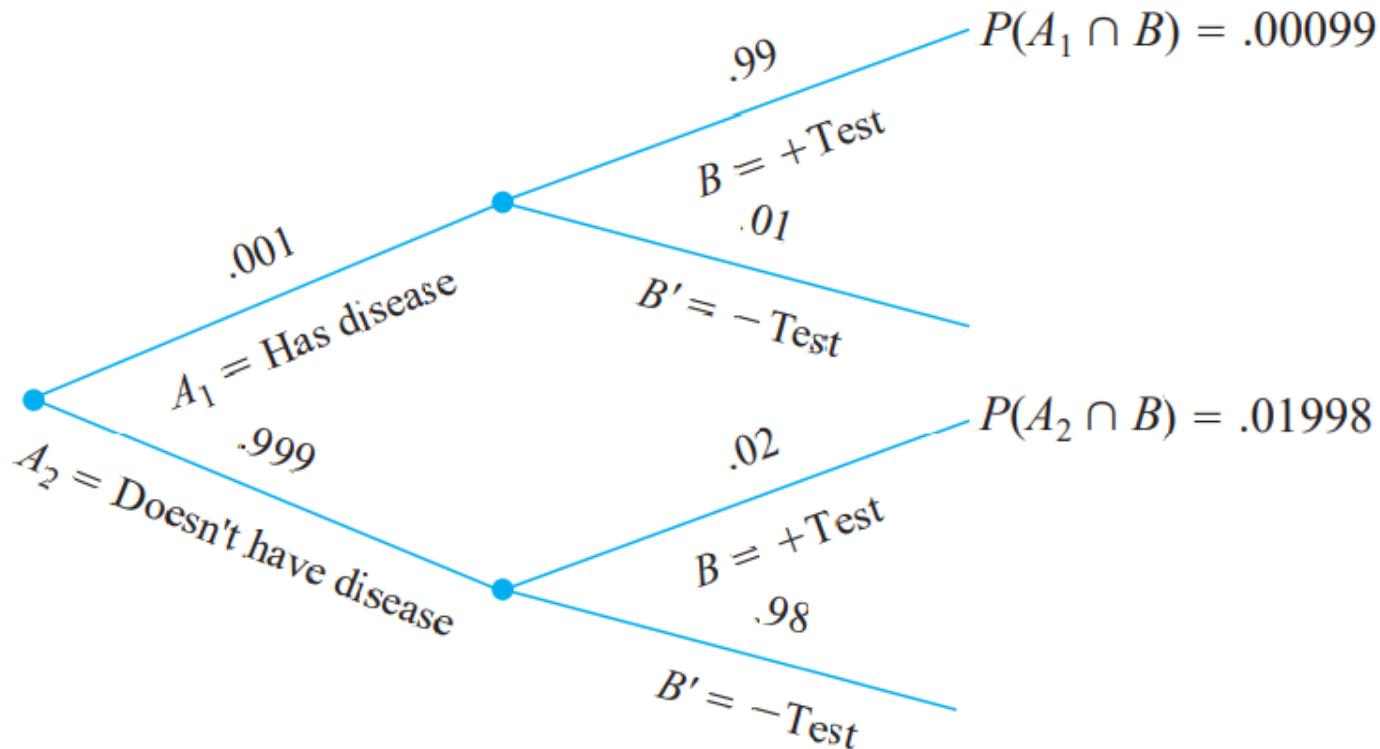
$$P(A_i | B) = \frac{P(A_i) \cdot P(B | A_i)}{\sum_{i=1}^{i=n} P(A_i) P(B | A_i)}$$

Hence the theorem proved

Example:

Only 1 in 1000 adults is afflicted with a rare disease for which a diagnostic test has been developed. The test is such that when an individual actually has the disease, a positive result will occur 99% of the time, whereas an individual without the disease will show a positive test result only 2% of the time. If a randomly selected individual is tested and the result is positive, what is the probability that the individual has the disease?

To use Bayes' theorem, let A_1 = individual has the disease, A_2 = individual does not have the disease, and B = positive test result. Then $P(A_1) = .001$, $P(A_2) = .999$, $P(B|A_1) = .99$, and $P(B|A_2) = .02$. The tree diagram for this problem is in Figure



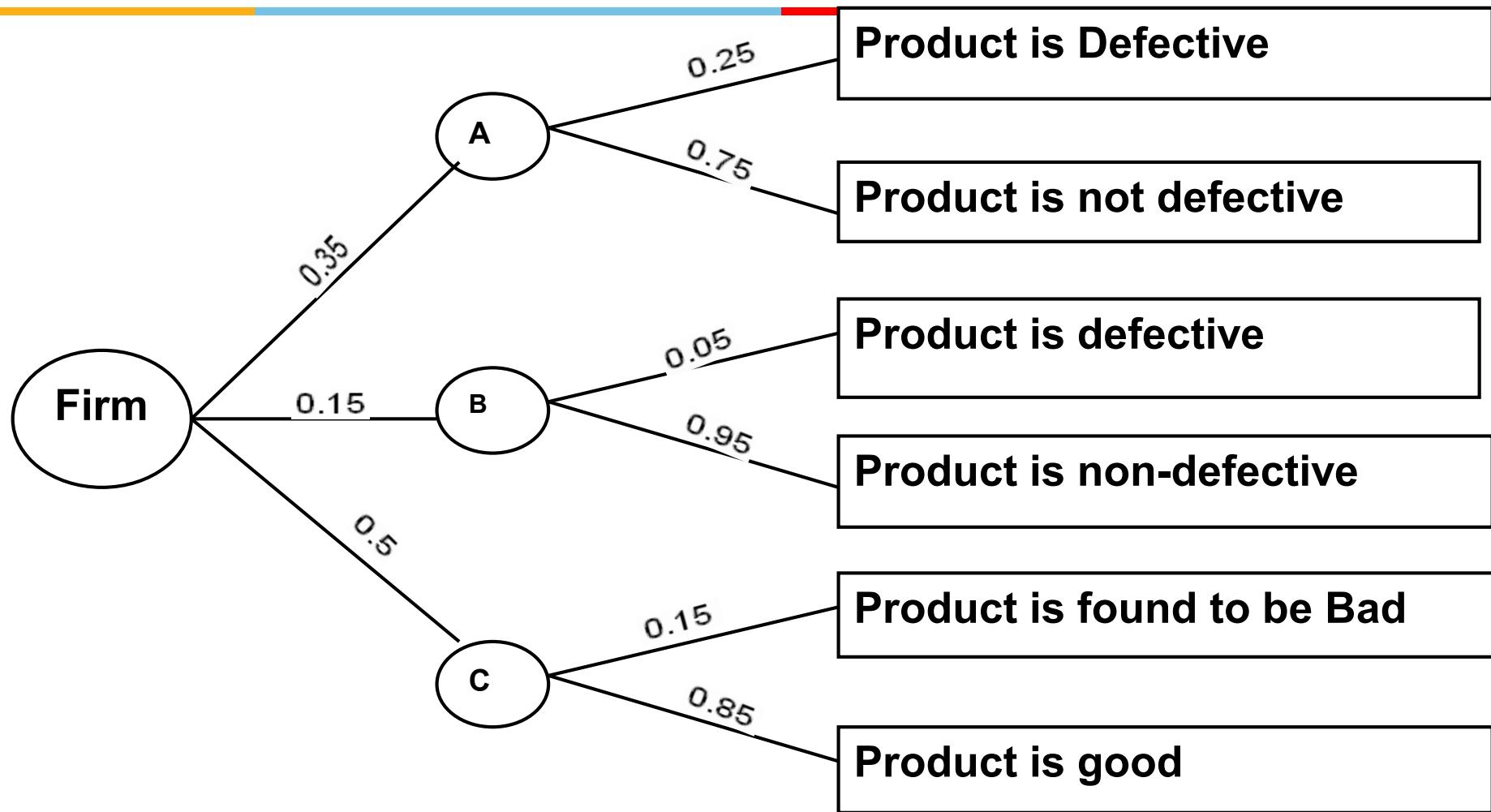
Next to each branch corresponding to a positive test result, the multiplication rule yields the recorded probabilities. Therefore, $P(B) = .00099 + .01998 = .02097$, from which we have

$$P(A_1 | B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{.00099}{.02097} = .047$$

Example:

A certain firm has plants A, B, C producing, respectively 35%, 15% and 50% of the total output. The probabilities of a non – defective product are, respectively, 0.75, 0.95 and 0.85. A Customer receives a bad product, what is the Chance that product came from the plant C?

Tree Diagram



Solution

Let X : “Customer receives a defective product”.

$$\begin{aligned}\text{Clearly, } P(X) &= P(A)P\left[\frac{X}{A}\right] + P(B)P\left[\frac{X}{B}\right] + P(C)P\left[\frac{X}{C}\right] \\ &= 0.17\end{aligned}$$

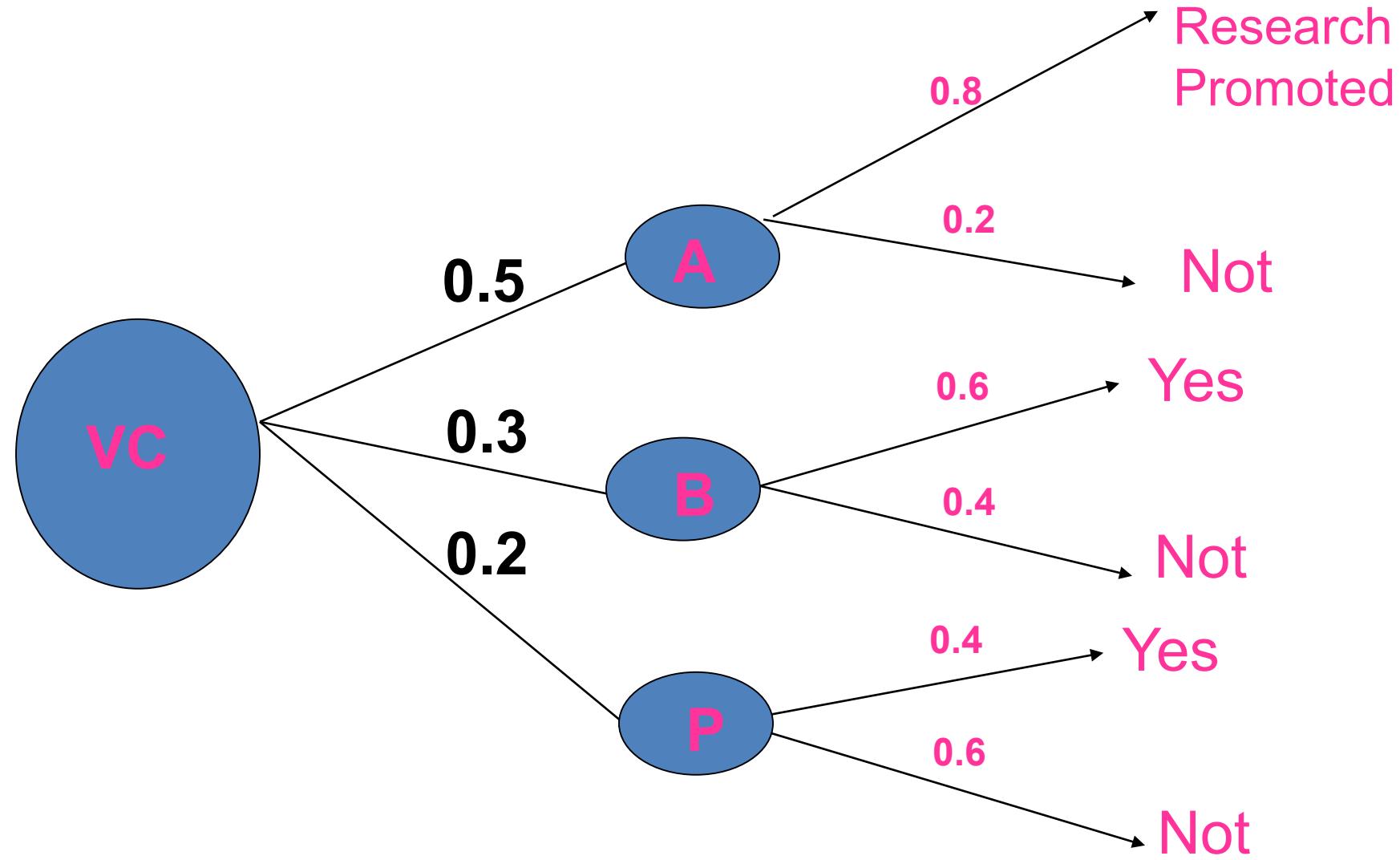
Therefore, the chance that product is manufactured by the plant C is

$$P(C | X) = \frac{P(C \cap X)}{P(X)} = \frac{0.5 \cdot 0.15}{0.17} = 0.4412$$

Example

The chances that an academician, a business man and a politician becoming Vice Chancellor of an university are 0.5, 0.3 and 0.2 respectively. The probability that research work will be promoted in the university by these 3 gentlemen are respectively are 0.8, 0.6 and 0.4. It is found Research work has been promoted by the university. What is the chance that an academician has become the VC?

Tree Diagram



Example

The chances that an academician, a business man and a politician becoming Vice Chancellor of an university are 0.5, 0.3 and 0.2 respectively. The probability that research work will be promoted in the university by these 3 gentlemen are respectively are 0.8, 0.6 and 0.4. It is found Research work has been promoted by the university. What is the chance that an academician has become the VC?

Let X : “Research work is promoted”

$$\text{Clearly, } P(X) = 0.5 \times 0.8 + 0.3 \times 0.6 + 0.2 \times 0.4 = 0.66$$

Now to find $P[\text{“An Academician is VC”} / \text{“Research work is promoted i.e. event } X\text{”}]$

$$= \frac{0.5 \times 0.8}{0.66} = 0.6061$$

Example: A manufacturer of tablets receives its LED screens from three different suppliers, 60% from supplier B_1 , 30% from supplier B_2 , and 10% from supplier B_3 . In other words, the probabilities that any one LED screens received by the plant comes from these three suppliers are 0.60, 0.30, and 0.10. Also suppose that 95% of the LED screens from B_1 , 80% of those from B_2 , and 65% of those from B_3 perform according to specifications.

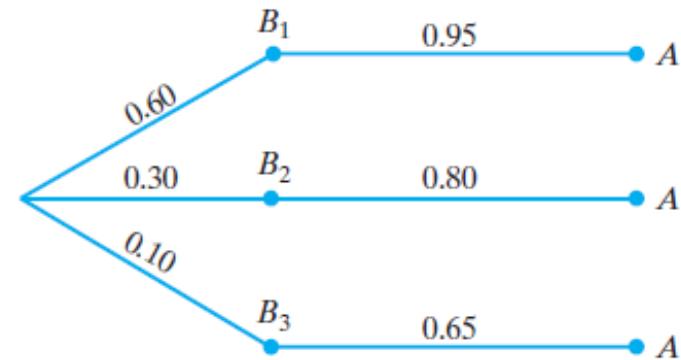
- 1) What is the probability that any one LED screen received by the plant will perform according to specifications?
- 2) Determine the probability that a particular LED screen, which is known to perform according to specifications, came from supplier B_3 .

If A denotes the event that a LED screen received by the plant performs according to specifications, and B_1 , B_2 , and B_3 are the events that it comes from the respective suppliers

$$\begin{aligned} A &= A \cap [B_1 \cup B_2 \cup B_3] \\ &= (A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3) \\ &= P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3) \end{aligned}$$

$$\begin{aligned} P(A) &= P(B_1) \cdot P(A | B_1) + \\ &\quad P(B_2) \cdot P(A | B_2) \\ &\quad + P(B_3) \cdot P(A | B_3) \end{aligned}$$

$$\begin{aligned} P(A) &= (0.60)(0.95) + (0.30)(0.80) + (0.10)(0.65) \\ &= 0.875 \end{aligned}$$



$$P(B_3 | A) = \frac{(0.10)(0.65)}{(0.60)(0.95) + (0.30)(0.80) + (0.10)(0.65)} = 0.074$$

Suggested Problems

Example: Two firms V and W consider bidding on a road-building job, which may or may not be awarded depending on the amounts of the bids. Firm V submits a bid and the probability is $\frac{3}{4}$ that it will get the job provided firm W does not bid. The probability is $\frac{3}{4}$ that W will bid, and if it does, the probability that V will get the job is only $\frac{1}{3}$.
(a) what is the probability that V will get the job? (b) If V gets the job, what is the probability that W did not bid?

Answer: Given $P(V/W^1) = \frac{3}{4}$, $P(W) = \frac{1}{3}$, $P(V/W) = \frac{1}{3}$ $P(W^1) = \frac{2}{3}$

$$(a) V = (V \cap W) \cup (V \cap W^1) \Rightarrow P(V) = P(V \cap W) + P(V \cap W^1)$$

$$P(V) = P(V/W) P(W) + P(V/W^1) P(W^1) = \frac{11}{18}$$

$$(b) P(W^1/V) = P(V/W^1) P(W^1) / P(V) = \frac{9}{11}$$

Suggested Problems

Example. An office has 4 secretaries handling respectively 20%, 60%, 15% and 5% of the files of all government reports. The probability that they misfile such reports are respectively 0.05, 0.1, 0.1 and 0.05. Find the probability that the misfiled report can be blamed on the first secretary.

Example . In a class 70% are boys and 30% are girls. 5% of boys and 3% of girls are irregular to the classes. What is the probability of a student selected at random is irregular to the classes and what is the probability that the irregular student is a girl?

Suggested Problems

Example 5. Three machines A, B and C produce respectively 60%, 30% and 10% of the total number of items of a factory. The percentage of defective outputs of these machines are 2%, 3% and 4%. An item is selected at random and is found to be defective. (i) Find the probability that the item was produced by machine C? (ii) What is the probability that the item was produced by machine C or B?

Bayesian Learning

- Naive Bayes is a set of simple and efficient machine learning algorithms for solving a variety of classification and regression problems.
- Naive Bayes assumes conditional independence where Bayes theorem does not. This means the relationship between all input features are independent.
- Bayesian learning algorithms that calculate explicit probabilities for hypotheses, such as the naive Bayes classifier, are among the most practical approaches to certain types of learning problems
- For example: Problem of learning to classify text documents such as electronic news articles.
- For such learning tasks, the naive Bayes classifier is among the most effective algorithms known

Features of Bayesian learning

- Prior knowledge can be combined with observed data to determine the final probability of a hypothesis.
 - Prior knowledge is provided by asserting
 - ❖ prior probability for each candidate hypothesis, and
 - ❖ probability distribution over observed data for each possible hypothesis.
 - New instances can be classified by combining the predictions of multiple hypotheses, weighted by their probabilities.
-

Bayes Theorem

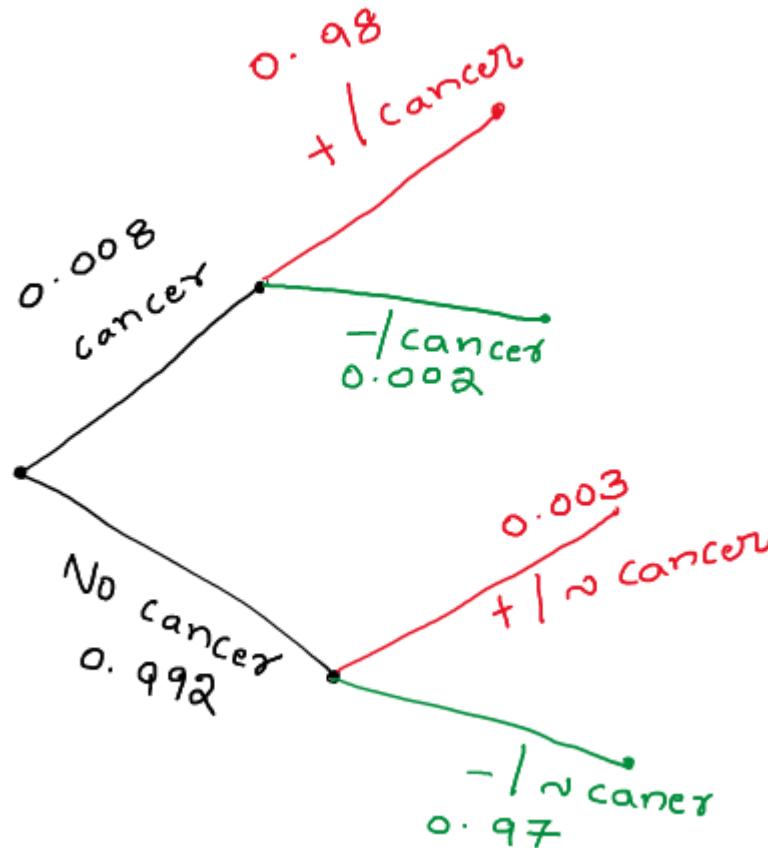
- $P(h)$ = prior probability of hypothesis h , before seeing the training data
- $P(D)$ = prior probability of training data D
- $P(h|D)$ = probability of h given D
- $P(D|h)$ = probability of D given h

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Bayes Theorem: Example

- Consider a medical diagnosis problem in which there are two alternative hypotheses:
 - H1:That a patient has a particular form of cancer
 - H2:That the patient does not
 - The available data is from a particular laboratory test with two possible outcomes
 - + Positive
 - Negative
 - Over the entire population of people only 0.008 have this disease. The test returns a corrective positive result in only 98% of the cases in which the disease is actually present and a correct negative result in only 97% of the cases in which the disease is not present.
 - How does $P(\text{cancer}/+)$ compare to $P(\sim \text{cancer}/+)$?
-

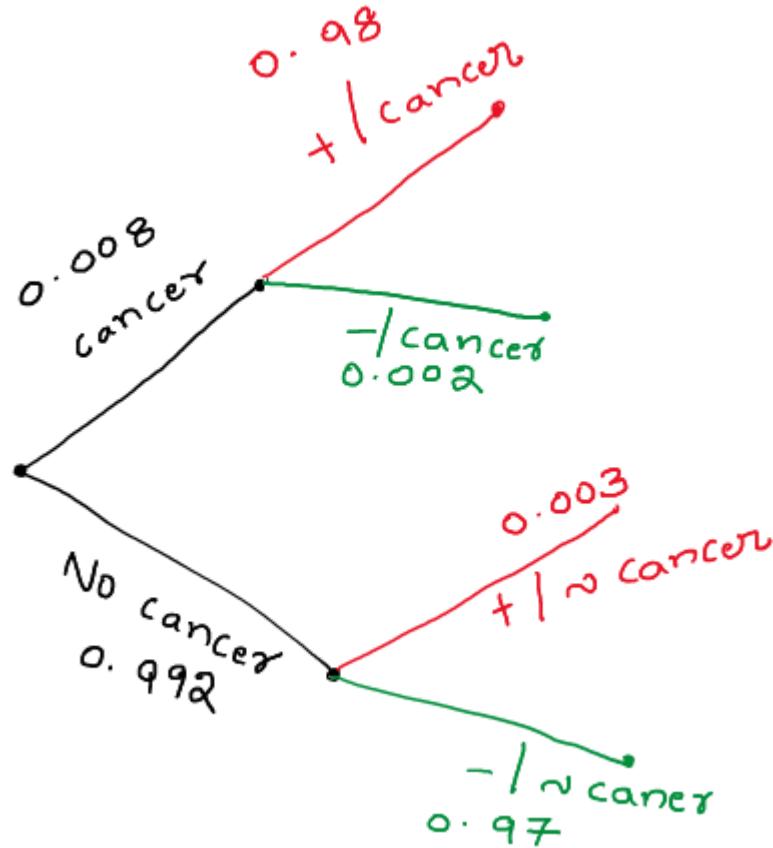
Bayes Theorem: Example



$$\begin{aligned}
 P(\text{cancer}/+) &= \frac{P(+/\text{cancer})P(\text{cancer})}{P(+)} \\
 &= \frac{0.98 \times 0.008}{0.98 \times 0.008 + 0.003 \times 0.992} \\
 &= 0.72485
 \end{aligned}$$

$$\begin{aligned}
 P(\sim\text{cancer}/+) &= \frac{P(+/\sim\text{cancer})P(\sim\text{cancer})}{P(+)} \\
 &= \frac{0.003 \times 0.992}{0.98 \times 0.008 + 0.003 \times 0.992} \\
 &= 0.27515
 \end{aligned}$$

Bayes Theorem: Example

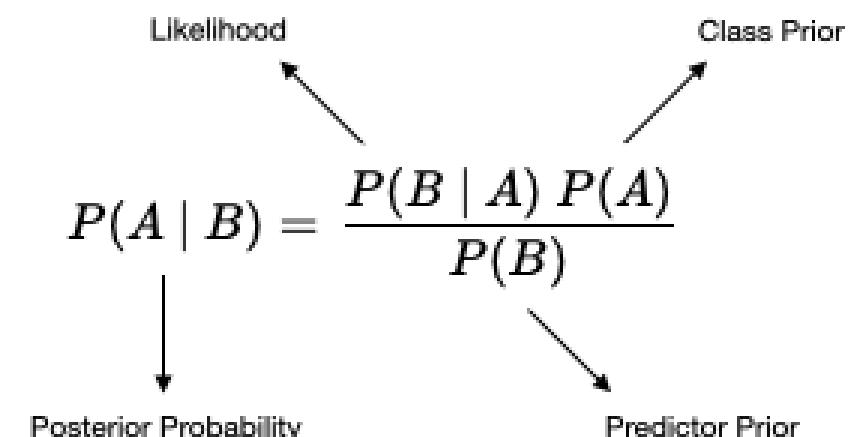


$$\begin{aligned}
 P(\text{cancer} | -) &= \frac{P(- | \text{cancer}) P(\text{cancer})}{P(-)} \\
 &= \frac{(0.02)(0.008)}{?} = \\
 P(\sim \text{cancer} | -) &= \frac{P(- | \sim \text{cancer}) P(\sim \text{cancer})}{P(-)} \\
 &= \frac{(0.97)(0.992)}{P(-)} =
 \end{aligned}$$

Machine Learning

- Generative models
 - Build model to estimate the posterior probability $P(Y|X)$ by estimating
 - likelihood of data given target (hypothesis) $P(X|Y)$
 - Prior probabilities over target $P(Y)$
 - In general, for a specific class $Y=c_k$,

$$P(Y = c_k|X) = \frac{P(X|Y = c_k)*P(Y=c_k)}{P(X)}$$



Choosing Hypotheses

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- Generally want the most probable hypothesis given the training data
- *Maximum a posteriori* hypothesis h_{MAP} :
$$\begin{aligned} h_{MAP} &= \arg \max_{h \in H} P(h|D) \\ &= \arg \max_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\ &= \arg \max_{h \in H} P(D|h)P(h) \end{aligned}$$
- If assume $P(h_i) = P(h_j)$ then can further simplify, and choose the *Maximum likelihood* (ML) hypothesis

$$h_{ML} = \arg \max_{h_i \in H} P(D|h_i)$$

Brute Force MAP Hypothesis

- For each hypothesis h in H , calculate the posterior probability

$$P(h | D) = \frac{P(D | h)P(h)}{P(D)}$$

- Output the hypothesis h_{MAP} with the highest posterior probability

$$h_{MAP} = \operatorname{argmax}_{h \in H} P(h | D)$$

MAP Hypothesis

- Using Bayes theorem, we compute the MAP hypothesis for all probable hypothesis (or all unique class labels)
- Identify the best hypothesis describing the data as

$$\begin{aligned} h_{MAP} &= \operatorname{argmax}_{h \in H} P(h|D) \\ &= \operatorname{argmax}_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\ &= \operatorname{argmax}_{h \in H} P(D|h)P(h) \end{aligned}$$

H: set of all hypothesis

P(D) is independent of h and is same for all hypothesis, therefore dropped

Maximum Likelihood Estimation



- When no prior information is available, all hypothesis are equally likely i.e $p(h_i) = p(h_j)$
- This is also true for a balanced class problem where all the classes are equally likely
- This is known as uniform prior
- MAP hypothesis further simplified to
- $h_{ML} = \operatorname{argmax} P(D/h)$ (where h belongs to H)

Conditional independence

- **Definition:** X is conditionally independent of Y given Z , if the probability distribution governing X is independent of the value of Y , given the value of Z

$$(\forall i, j, k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z_k)$$

$$P(X|Y, Z) = P(X|Z)$$

Example:

$$P(\text{Thunder}|\text{Rain, Lightning}) = P(\text{Thunder}|\text{Lightning})$$

Applying conditional independence

Naïve Bayes assumes X_i are conditionally independent given Y

e.g., $P(X_1|X_2, Y) = P(X_1|Y)$

$$\begin{aligned}P(X_1, X_2|Y) &= P(X_1|X_2, Y)P(X_2|Y) \\&= P(X_1|Y)P(X_2|Y)\end{aligned}$$

General form: $P(X_1, \dots, X_n|Y) = \prod_{j=1}^n P(X_j|Y)$

How many parameters to describe $P(X_1, \dots, X_n|Y)$? $P(Y)$?

Without conditional independence assumption?

With conditional independence assumption?

Naïve Bayes Independence assumption

Assumption:

$$P(X_1, \dots, X_n | Y) = \prod_{j=1}^n P(X_j | Y)$$

i.e., X_i and X_j are conditionally independent
given Y for $i \neq j$

Naïve Bayes classifier

- **Bayes rule:**

$$P(Y = y_k | X_1, \dots, X_n) = \frac{P(Y = y_k)P(X_1, \dots, X_n | Y = y_k)}{\sum_j P(Y = y_j)P(X_1, \dots, X_n | Y = y_j)}$$

- **Assume conditional independence among X_i 's:**

$$P(Y = y_k | X_1, \dots, X_n) = \frac{P(Y = y_k)\Pi_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j)\Pi_i P(X_i | Y = y_j)}$$

- **Pick the most probable (MAP) Y**

$$\hat{Y} \leftarrow \operatorname{argmax}_{y_k} P(Y = y_k)\Pi_i P(X_i | Y = y_k)$$

↑
Prior
Probability
↑
MLE

NAÏVE BAYES CLASSIFIER

- Assume independence among attributes X_i when class is given:
 - ❖ $P(X_1, X_2, \dots, X_d | Y_j) = P(X_1 | Y_j) P(X_2 | Y_j) \dots P(X_d | Y_j)$
 - ❖ Now we can estimate $P(X_i | Y_j)$ for all X_i and Y_j combinations from the training data
 - ❖ New point is classified to Y_j if $P(Y_j) \prod P(X_i | Y_j)$ is maximal.

Example 1:

If the weather is sunny,
then the player will play
or not?

i.e. Play/ Sunny = Yes or No

Note if we know $P(\text{Yes/Sunny})$ and
 $P(\text{No/Sunny})$ then we can answer the
question asked

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Steps to Apply Bayes Theorem



Step 1- View or collect “raw” data.

Step 2 - Convert long data to a frequency table

weather	Play		Row Total
	no	yes	
Sunny	2	3	5
Overcast	0	4	4
Rainy	3	2	5
Column Total	5	9	14

Step 3 - Row and column sums to get probabilities

Weather probabilities

$$\text{sunny} = 5/14, \text{rainy} = 5/14$$

$$\text{Overcast} = 4/14$$

Play probabilities

$$\text{no} = 5/14$$

$$\text{yes} = 9/14$$

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Steps to Apply Bayes Theorem



		Play		
		no	yes	Row Total
weather				
Sunny		2	3	5
				P(Sunny)= 5/14
Overcast		0	4	4
				P(Overcast) = 4/14
Rainy		3	2	5
				P(Rainy)=5/14
Column Total		5	9	14
				P(no)=5/14 P(yes)=9/14

Step 4 - Apply probabilities from frequency table to Bayes theorem

$$P(\text{yes} \mid \text{sunny}) = \frac{P(\text{sunny} \mid \text{yes}) P(\text{yes})}{P(\text{sunny})}$$

weather	no	yes
Rainy	3	2
sunny	2	3
overcast	0	4
Total	5	9

$$\frac{5}{14} = 0.36$$

$$\frac{9}{14} = 0.64$$

$$\frac{4}{14} = 0.29$$

$$\frac{5}{14} = 0.36 \quad \frac{9}{14} = 0.64$$

$$\text{Now } P(\text{Yes} | \text{sunny}) = \frac{P(\text{sunny} | \text{yes}) P(\text{yes})}{P(\text{sunny})}$$

$$= \frac{(3/9)(9/14)}{5/14} = \underline{\underline{0.60}} \quad \checkmark$$

$$P(\text{no} | \text{sunny}) = \frac{(2/5)(5/14)}{5/14} = \underline{\underline{0.40}}$$

Example 2:

If the features of
today = (Outlook is Sunny, Temp is Hot, Humidity is Normal, Windy is False),
 then the player will play or not?

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

x x_1 x_2 x_3 x_4

 today = (Sunny, Hot, Normal, False)

$$\begin{aligned}
 P(\neg y_n | x) &= \frac{P(x | \text{yes}) P(\text{yes})}{P(x)} \\
 &= \frac{P(x_1, x_2, x_3, x_4 | \text{yes}) \text{yes}}{P(x_1, x_2, x_3, x_4)} = \frac{P(x_1 | y) P(x_2 | y) P(x_3 | y) P(x_4 | y)}{P(x_1) P(x_2) P(x_3) P(x_4)}
 \end{aligned}$$

Outlook

	Yes	No	P(yes)	P(no)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0/5
Rainy	3	2	3/9	2/5
Total	9	5	100%	100%

Temperature

	Yes	No	P(yes)	P(no)
Hot	2	2	2/9	2/5
Mild	4	2	4/9	2/5
Cool	3	1	3/9	1/5
Total	9	5	100%	100%

Humidity

	Yes	No	P(yes)	P(no)
High	3	4	3/9	4/5
Normal	6	1	6/9	1/5
Total	9	5	100%	100%

Wind

	Yes	No	P(yes)	P(no)
False	6	2	6/9	2/5
True	3	3	3/9	3/5
Total	9	5	100%	100%

Play

Play		P(Yes)/P(No)
Yes	9	9/14
No	5	5/14
Total	14	100%

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 = ?

$D = \{ \text{magazine promotion, watch promotion, Life insurance promotion, credit card insurance} \}$

$h_i = \text{male or Female}$

$$P(\underline{\text{male}} / \text{Yes, yes, no, no}) = ?$$

$$P(\underline{\text{Female}} / \text{yes, Yes, No, no}) = ?$$

		magazine promotion		watch promotion		L.I promotion		credit Card promotion	
		Male	Female	Male	Female	Male	F	M	F
YES		4	3	2	2	2	3	2	1
	NO	2	1	4	2	4	1	4	3
Ratios (YES)		$\frac{4}{6}$	$\frac{3}{4}$	$\frac{2}{6}$	$\frac{2}{4}$	$\frac{2}{6}$	$\frac{3}{4}$	$\frac{2}{6}$	$\frac{1}{4}$
		$\frac{2}{6}$	$\frac{1}{4}$	$\frac{4}{6}$	$\frac{2}{4}$	$\frac{4}{6}$	$\frac{1}{4}$	$\frac{4}{6}$	$\frac{3}{4}$

$$P(\text{male} | E)$$

$$= \frac{P(E | \text{male}) P(\text{male})}{P(E)}$$

$$= \frac{\left(\frac{4}{6} \cdot \frac{2}{6} \cdot \frac{4}{6} \cdot \frac{4}{6} \right) \left(\frac{3}{5} \right)}{P(E)}$$

$$= \frac{0 \cdot 0593}{P(E)}$$

YES

YES

NO

NO

$$\frac{6}{10} \cdot \frac{3}{5}$$

$$P(\text{Female} | E)$$

$$= \frac{P(E | \text{Female}) P(\text{Female})}{P(E)}$$

$$= \frac{\left(\frac{3}{4}\right)\left(\frac{2}{4}\right)\left(\frac{1}{4}\right)\left(\frac{3}{4}\right) \cdot \frac{2}{5}}{P(E)}$$

$$= \frac{\left(\frac{9}{128}\right)\left(\frac{2}{5}\right)}{P(E)} = \frac{0.0281}{P(E)}$$

$$0.0593 > 0.0281$$

is male ✓

Example 4:

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

A: attributes

M: mammals

N: non-mammals

$$P(A|M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A|N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A|M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A|N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

$$P(A|M)P(M) > P(A|N)P(N)$$

=> Mammals

Issues with Naïve Bayes Classifier

Naïve Bayes Classifier:

$$P(\text{Refund} = \text{Yes} | \text{No}) = 3/7$$

$$P(\text{Refund} = \text{No} | \text{No}) = 4/7$$

$$P(\text{Refund} = \text{Yes} | \text{Yes}) = 0$$

$$P(\text{Refund} = \text{No} | \text{Yes}) = 1$$

$$P(\text{Marital Status} = \text{Single} | \text{No}) = 2/7$$

$$P(\text{Marital Status} = \text{Divorced} | \text{No}) = 1/7$$

$$P(\text{Marital Status} = \text{Married} | \text{No}) = 4/7$$

$$P(\text{Marital Status} = \text{Single} | \text{Yes}) = 2/3$$

$$P(\text{Marital Status} = \text{Divorced} | \text{Yes}) = 1/3$$

$$P(\text{Marital Status} = \text{Married} | \text{Yes}) = 0$$

- | $P(\text{Yes}) = 3/10$

- | $P(\text{No}) = 7/10$

- | $P(\text{Yes} | \text{Married}) = 0 \times 3/10 / P(\text{Married})$

- | $P(\text{No} | \text{Married}) = 4/7 \times 7/10 / P(\text{Married})$

For Taxable Income:

If class = No: sample mean = 110

sample variance = 2975

If class = Yes: sample mean = 90

sample variance = 25

Issues with Naïve Bayes Classifier

Consider the table with Tid = 7 deleted

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	120K	Yes
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Naïve Bayes Classifier:

$$P(\text{Refund} = \text{Yes} | \text{No}) = 2/6$$

$$P(\text{Refund} = \text{No} | \text{No}) = 4/6$$

→ $P(\text{Refund} = \text{Yes} | \text{Yes}) = 0$

$$P(\text{Refund} = \text{No} | \text{Yes}) = 1$$

$$P(\text{Marital Status} = \text{Single} | \text{No}) = 2/6$$

→ $P(\text{Marital Status} = \text{Divorced} | \text{No}) = 0$

$$P(\text{Marital Status} = \text{Married} | \text{No}) = 4/6$$

$$P(\text{Marital Status} = \text{Single} | \text{Yes}) = 2/3$$

$$P(\text{Marital Status} = \text{Divorced} | \text{Yes}) = 1/3$$

$$P(\text{Marital Status} = \text{Married} | \text{Yes}) = 0/3$$

For Taxable Income:

If class = No: sample mean = 91

sample variance = 685

If class = Yes: sample mean = 90

sample variance = 25

Given $X = (\text{Refund} = \text{Yes}, \text{Divorced}, 120\text{K})$

$$P(X | \text{No}) = 2/6 \times 0 \times 0.0083 = 0$$

$$P(X | \text{Yes}) = 0 \times 1/3 \times 1.2 \times 10^{-9} = 0$$

**Naïve Bayes will not be able to
classify X as Yes or No!**

Naïve Bayes for Text Classification



- Naïve Bayes is commonly used for **text classification**
- For a document with k terms $d = (t_1, \dots, t_k)$

Fraction of documents in c

$$P(c|d) = P(c)P(d|c) = P(c) \prod_{t_i \in d} P(t_i|c)$$

- $P(t_i|c)$ = Fraction of terms from **all documents** in c that are t_i

Number of times t_i appears in some document in c

$$P(t_i|c) = \frac{N_{ic} + 1}{N_c + T}$$

Laplace Smoothing

Total number of terms in all documents in c

Number of unique words (vocabulary size)

- Easy to implement and works relatively well
- **Limitation:** Hard to incorporate **additional features** (beyond words).
 - E.g., number of adjectives used.

A Simple Example:

Text	Tag
“A great game”	Sports
“The election was over”	Not sports
“Very clean match”	Sports
“A clean but forgettable game”	Sports
“It was a close election”	Not sports

Which tag does the sentence *A very close game* belong to? i.e. $P(\text{sports} \mid A \text{ very close game})$

Feature Engineering: Bag of words i.e use word frequencies without considering order

Using Bayes Theorem:

$$P(\text{sports} \mid A \text{ very close game})$$

$$= P(A \text{ very close game} \mid \text{sports}) P(\text{sports})$$

$$P(A \text{ very close game})$$

We assume that every word in a sentence is **independent** of the other ones

“close” doesn’t appear in sentences of sports tag, So $P(\text{close} \mid \text{sports}) = 0$, which makes product 0

A Simple Example

Draw | Naive Bayes
Text classification

Text	Tag	
"A great game"	Sports	Which tag does the sentence "A very close game" belong to? i.e. $P(\text{sports} \mid \text{A very close game})$
"The election was over"	Not sports	Feature Engineering: Bag of words i.e use word frequencies without considering order
"Very clean match"	Sports	Using Bayes Theorem:
"A clean but forgettable game"	Sports	$P(\text{sports} \mid \text{A very close game})$ = $\frac{P(\text{A very close game} \mid \text{sports}) P(\text{sports})}{P(\text{A very close game})}$
"It was a close election"	Not sports	

We assume that every word in a sentence is **independent** of the other ones

$$P(\text{A very close game}) = P(A) P(\text{very}) P(\text{close}) P(\text{game})$$

$$P(\text{A very closed game} \mid \text{sports}) = P(\text{a} \mid \text{sports}) P(\text{very} \mid \text{sports})$$

$$P(\text{close} \mid \text{sports}) P(\text{game} \mid \text{sports})$$

"close" doesn't appear in sentences of sports tag, So $P(\text{close} \mid \text{sports}) = 0$, which makes product 0

$$P(\text{sports} \mid \text{A very close game}) = \frac{P(\text{A very close game} \mid \text{sports}) \cdot P(\text{sports})}{P(\text{A very close game})}$$

$$P(\text{not sports} \mid \text{A very close game})$$

Laplace smoothing

- Laplace smoothing: we add 1 or in general constant k to every count so it's never zero.
 - To balance this, we add the number of possible words to the divisor, so the division will never be greater than 1
 - In our case, the possible words are ['a', 'great', 'very', 'over', 'it', 'but', 'game', 'election', 'clean', 'close', 'the', 'was', 'forgettable', 'match'].
-

Apply Laplace Smoothing

Word	P(word Sports)	P(word Not Sports)
a	2+1 / 11+14	1+1 / 9+14
very	1+1 / 11+14	0+1 / 9+14
close	0+1 / 11+14	1+1 / 9+14
game	2+1 / 11+14	0+1 / 9+14

$$\begin{aligned}
 & P(a|Sports) \times P(very|Sports) \times P(close|Sports) \times P(game|Sports) \times \\
 & P(Sports) \\
 & = 2.76 \times 10^{-5} \\
 & = 0.0000276
 \end{aligned}$$

$$\begin{aligned}
 & P(a|Not\ Sports) \times P(very|Not\ Sports) \times P(close|Not\ Sports) \times \\
 & P(game|Not\ Sports) \times P(Not\ Sports) \\
 & = 0.572 \times 10^{-5} \\
 & = 0.00000572
 \end{aligned}$$

Example :

Doc No	Text
1	I LOVED THE MOVIE
2	I HATED THE MOVIE
3	A GREAT MOVIE ,GOOD MOVIE
4	POOR ACTING
5	GREAT ACTING , A GOOD MOVIE
NEW	I HATED THE POOR ACTING

Example :

Doc No	Text	
1	I LOVED THE MOVIE	<u>POSITIVE</u>
2	I HATED THE MOVIE	<u>NEGATIVE</u>
3	A GREAT MOVIE ,GOOD MOVIE	<u>POSITIVE</u>
4	POOR ACTING	<u>NEGATIVE</u>
5	GREAT ACTING , A GOOD MOVIE	<u>POSITIVE</u>
NEW	I HATED THE POOR ACTING	<u>????</u>

$$P(c/x)$$

$$P(+ / \text{I hated the poor acting}) =$$

$$P(- / \text{I hated the poor acting}) =$$

Based on these probabilities, we can decide the class which the new text belongs

$P(+ | \text{I hated the acting})$

i.e. $P(c_1 | x) = \frac{P(x | c_1) P(c_1)}{P(x)}$

$$= P(\text{I} | +) P(\text{hated} | +) P(\text{the} | +) P(\text{acting} | +) P(+)$$

$$\frac{P(\text{I}, +)}{P(+)}$$

$$\frac{P(\text{hated}, +)}{P(+)}$$

words	positive	negative
I	1	1
loved	1	0
the	1	1
movie	4	1
hated	0	1
a	2	0
great	2	0
Poor	0	1
acting	1	1
good	2	0

$$P(\pm | +)$$

$$= \frac{1 + 1}{14 + 10}$$

$$P(I | -)$$

$$= \frac{1 + 1}{6 + 10}$$

"I hated the poor
acting"

word

I

$$\frac{1+1}{14+10} = 0.0833$$

hated

$$\frac{0+1}{14+10} = 0.0417$$

the

$$\frac{1+1}{14+10} = 0.0833$$

poor

$$\frac{0+1}{14+10} = 0.0417$$

acting

$$\frac{1+1}{14+10} = 0.0833$$

positive

negative

$$\frac{1+1}{6+10} = 0.125$$

$x : I$ hate the Poor
acting

$$P(+|x)$$

$$= () () () () () \times P(+)$$

\downarrow
 $3/5$

$$= 6.03 \times 10^{-7}$$

$$P(-|x)$$

$$= () () () () () () P(-) :$$

\downarrow
 $2/8$

$$= 1.22 \times 10^{-5}$$

\therefore negative class

Example:

Suppose we got the new message with the words '**Dear Friend**', Decide whether this new message is a normal or spam message?

i.e. Normal/ Dear, Friend = Yes or No

Note if we know $P(\text{Normal} / \text{Dear, Friend})$ and $P(\text{Spam} / \text{Dear, Friend})$ then we can answer the question asked

Email word	Spam	Email word	Spam
Dear	Yes	Friend	No
Friend	No	Friend	Yes
Dear	No	Dear	No
Dear	No	Lunch	No
Dear	No	Friend	No
Friend	No	Dear	No
Lunch	No	Dear	No
Friend	No	Dear	No
Lunch	No	Dear	No
Dear	Yes	Money	Yes
Money	Yes	Money	No
Money	Yes	Money	Yes

Steps to Apply Bayes Theorem



Step 1- View or collect “raw” data.

Step 2 - Convert long data to a frequency table

	Play		Row Total
word	normal	spam	
Dear	8	2	10
Friend	5	1	6
Lunch	3	0	3
Money	1	4	5
Column Total	17	7	24

Step 3 - Row and column sums to get probabilities

Email word	Spam	Email word	Spam
Dear	Yes	Friend	No
Friend	No	Friend	Yes
Dear	No	Dear	No
Dear	No	Lunch	No
Dear	No	Friend	No
Friend	No	Dear	No
Lunch	No	Dear	No
Friend	No	Dear	No
Lunch	No	Dear	No
Dear	Yes	Money	Yes
Money	Yes	Money	No
Money	Yes	Money	Yes

Steps to Apply Bayes Theorem



Play

	normal	spam	Row Total
word			
Dear	8	2	10
Friend	5	1	6
Lunch	3	0	3
Money	1	4	5
Column Total	17	7	24

As $P(N/D, F) > P(S/D, F)$, we can decide that Dear Friend is Normal message.

Step 4 - Apply probabilities from frequency table to Bayes theorem

$$P(N/D, F) = \frac{P(D, F/N) \cdot P(N)}{P(D, F)} = \frac{P(D/N) \cdot P(F/N) \cdot P(N)}{P(D) \cdot P(F)} = \frac{\left(\frac{8}{17}\right) \cdot \left(\frac{5}{17}\right) \cdot \left(\frac{17}{24}\right)}{\left(\frac{10}{24}\right) \cdot \left(\frac{6}{24}\right)} = \frac{0.098}{0.104} = 0.9423$$

$$P(S/D, F) = \frac{P(D, F/S) \cdot P(S)}{P(D, F)} = \frac{P(D/S) \cdot P(F/S) \cdot P(S)}{P(D) \cdot P(F)} = \frac{\left(\frac{2}{7}\right) \cdot \left(\frac{1}{7}\right) \cdot \left(\frac{7}{24}\right)}{\left(\frac{10}{24}\right) \cdot \left(\frac{6}{24}\right)} = \frac{0.012}{0.104} = 0.1153$$

Example continued:

Suppose we got the new message contains the word '**Lunch Money Money Money Money**' , Decide whether this new message is a normal or spam message?

Email word	Spam	Email word	Spam
Dear	Yes	Friend	No
Friend	No	Friend	Yes
Dear	No	Dear	No
Dear	No	Lunch	No
Dear	No	Friend	No
Friend	No	Dear	No
Lunch	No	Dear	No
Friend	No	Dear	No
Lunch	No	Dear	No
Dear	Yes	Money	Yes
Money	Yes	Money	No
Money	Yes	Money	Yes

Steps to Apply Bayes Theorem



Play

	normal	spam	Row Total
word			
Dear	8	2	10
Friend	5	1	6
Lunch	3	0	3
Money	1	4	5
Column Total	17	7	24

We can observe that we have to classify any message with Lunch as Normal message, no matter how many times we see the word Money and that's the problem.

To work around this problem add 1 count to the frequency table to each word(Laplace smoothing)

Step 4 - Apply probabilities from frequency table

$$P(N) \cdot P(L/N) \cdot P(M/N)^4 = \left(\frac{17}{24}\right) \cdot \left(\frac{3}{17}\right) \cdot \left(\frac{1}{17}\right)^4 = 0.0000015$$

$$P(S) \cdot P(L/S) \cdot P(M/S)^4 = \left(\frac{7}{24}\right) \cdot \left(\frac{0}{7}\right) \cdot \left(\frac{4}{7}\right)^4 = 0$$

Steps to Apply Bayes Theorem



Play

	normal	spam	Row Total
word			
Dear	8+1	2+1	12
Friend	5+1	1+1	8
Lunch	3+1	0+1	5
Money	1+1	4+1	7
Column Total	21	11	32

As $P(S/L, M^4) > P(N/L, M^4)$, we can decide that
Lunch Money Money Money Money is Spam
message.

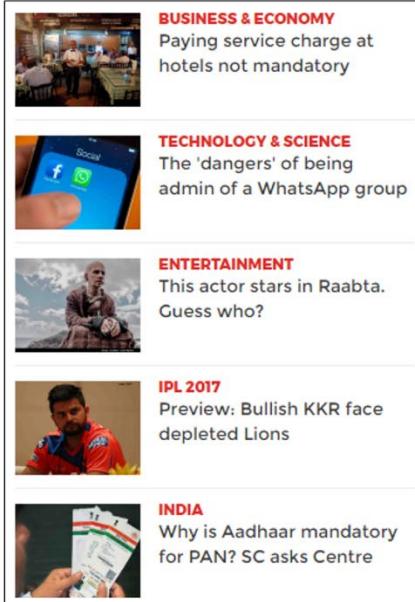
Step 4 - Apply probabilities from frequency table

$$P(N) \cdot P(L/N) \cdot P(M/N)^4 = \left(\frac{21}{32}\right) \cdot \left(\frac{4}{21}\right) \cdot \left(\frac{2}{21}\right)^4 = 0.00001$$

$$P(S) \cdot P(L/S) \cdot P(M/S)^4 = \left(\frac{11}{32}\right) \cdot \left(\frac{1}{11}\right) \cdot \left(\frac{5}{11}\right)^4 = 0.00133$$

Naïve Bayes Classifier Applications

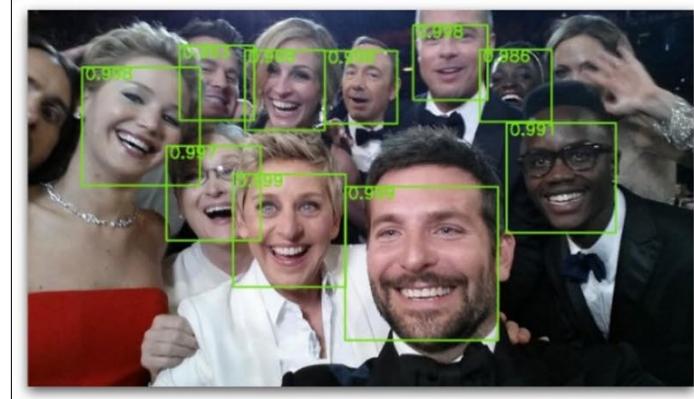
Categorizing News



Email Spam Detection



Face Recognition



Sentiment Analysis



Naive Bayes Classifier

- ✓ Along with decision trees, neural networks, one of the most practical learning methods.
 - ✓ When to use
 - ✓ Moderate or large training set available
 - ✓ Attributes that describe instances are conditionally independent given classification
 - ✓ Successful applications:
 - ✓ Diagnosis
 - ✓ Classifying text documents
-

Learning to Classify Text

- Why?
 - ❖ Learn which news articles are of interest
 - ❖ Learn to classify web pages by topic
 - Naive Bayes is among most effective algorithms
 - What attributes shall we use to represent text documents??
-

Baseline: Bag of Words Approach



Practical Issues of Bayesian learning



- Require initial knowledge of many probabilities
 - Often estimated based on background knowledge, previously available data, and assumptions about the form of the underlying distributions.
- Significant computational cost required to determine the Bayes optimal hypothesis in the general case (linear in the number of candidate hypotheses)

HW: Exercise

Consider the car theft problem with attributes Color, Type, Origin, and the target, Stolen can be either Yes or No.

we need to classify whether the car is stolen, given the features of the car.

Given the Red color Domestic SUV car Find the probability of whether the car is stolen?

Color	Type	Origin	Stolen?
Red	Sports	Domestic	Yes
Red	Sports	Domestic	No
Red	Sports	Domestic	Yes
Yellow	Sports	Domestic	No
Yellow	Sports	Imported	Yes
Yellow	SUV	Imported	No
Yellow	SUV	Imported	Yes
Yellow	SUV	Domestic	No
Red	SUV	Imported	No
Red	Sports	Imported	Yes

HW: Exercise

If the weather is Snowy,
then the player will play
or not?

weather	Player play
Sunny	yes
Rainy	no
Cloudy	yes
Sunny	no
Sunny	yes
snowy	no
Rainy	yes
Cloudy	no
Cloudy	yes
Sunny	yes
snowy	no
Cloudy	yes
Rainy	no
snowy	no
snowy	yes





Thanks



BITS Pilani

Pilani|Dubai|Goa|Hyderabad

M.Tech.(Data Science & Engineering) Introduction to Statistical Methods

Team ISM



Session No 4

Bayes theorem &

Introduction to Naïve Bayes concept

(Session 4: 3rd/4th Dec 2022)

Contact Session 4

Contact Session	List of Topic Title	Reference
CS - 4	Bayes theorem(with proof),Introduction to Naïve Bayes concept.	T1 & T2
HW	Problems on Bayes theorem	T1 & T2
Lab	Bayes theorem & Naïve Bayes Concept	Lab 2

Agenda

-
- Bayes Theorem
 - Introduction to Naïve Bayes concept

Text Books

No	Author(s), Title, Edition, Publishing House
T1	Statistics for Data Scientists, An introduction to probability, statistics and Data Analysis, Maurits Kaptein et al, Springer 2022
T2	Probability and Statistics for Engineering and Sciences, 8 th Edition, Jay L Devore, Cengage Learning

BAYES' THEOREM

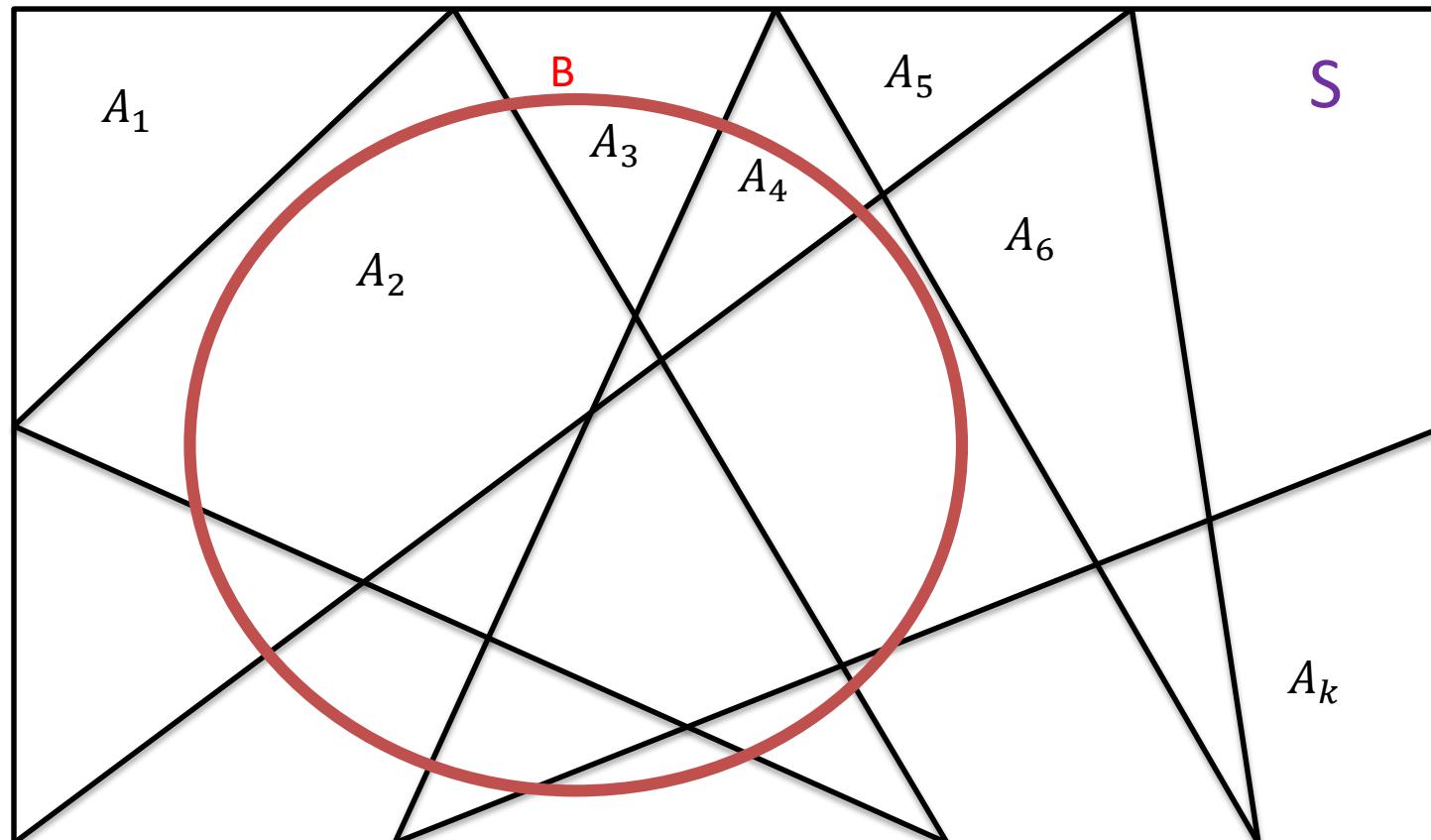
Let $P = \{A_1, A_2, A_3, \dots, A_n\}$ be a set of exhaustive and mutually exclusive events of a sample space S with $P(A_i) \neq 0$

For each i . If B is any other event associated with A_i with $P(B) \neq 0$, then

$$P\left(\frac{A_i}{B}\right) = \frac{P(A_i)P\left(\frac{B}{A_i}\right)}{P(B)}$$
$$= \frac{P(A_i)P\left(\frac{B}{A_i}\right)}{\sum_{i=1}^n P(A_i)P\left(\frac{B}{A_i}\right)}$$

Geometrical Representation of Bayes theorem

$$S = A_1 \cup A_2 \cup \dots \cup A_n \quad \text{and} \quad A_1 \cap A_2 \cap \dots \cap A_n = \emptyset$$



$$\therefore B = B \cap S = B \cap \{A_1 \cup A_2 \cup A_3, \dots \cup A_n\}$$

Proof:

The conditional Probability of A_i for any i given B is given as

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

But we know that

$$P(A_i \cap B) = P(A_i) \cdot P(B | A_i) \text{ & } P(B) = \sum_{i=1}^{i=n} P(A_i) P(B | A_i)$$

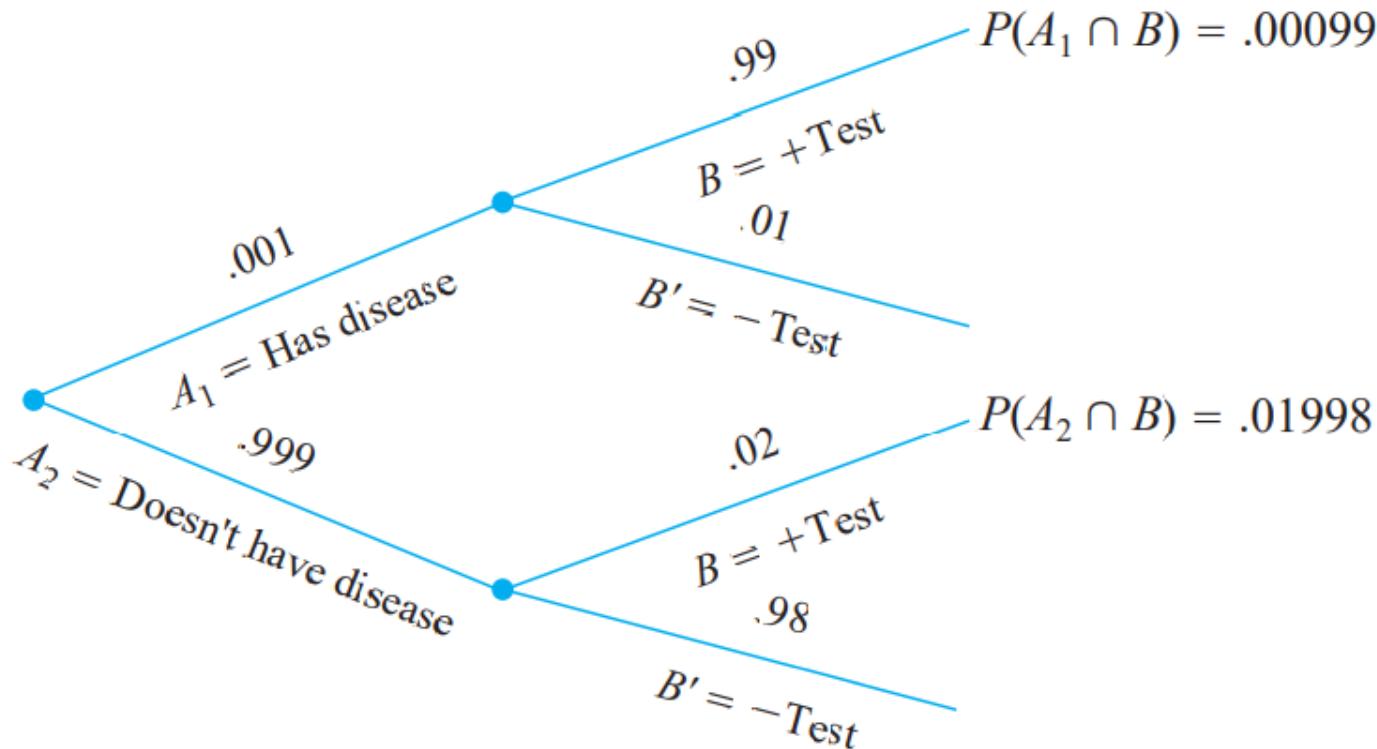
$$P(A_i | B) = \frac{P(A_i) \cdot P(B | A_i)}{\sum_{i=1}^{i=n} P(A_i) P(B | A_i)}$$

Hence the theorem proved

Example:

Only 1 in 1000 adults is afflicted with a rare disease for which a diagnostic test has been developed. The test is such that when an individual actually has the disease, a positive result will occur 99% of the time, whereas an individual without the disease will show a positive test result only 2% of the time. If a randomly selected individual is tested and the result is positive, what is the probability that the individual has the disease?

To use Bayes' theorem, let A_1 = individual has the disease, A_2 = individual does not have the disease, and B = positive test result. Then $P(A_1) = .001$, $P(A_2) = .999$, $P(B|A_1) = .99$, and $P(B|A_2) = .02$. The tree diagram for this problem is in Figure



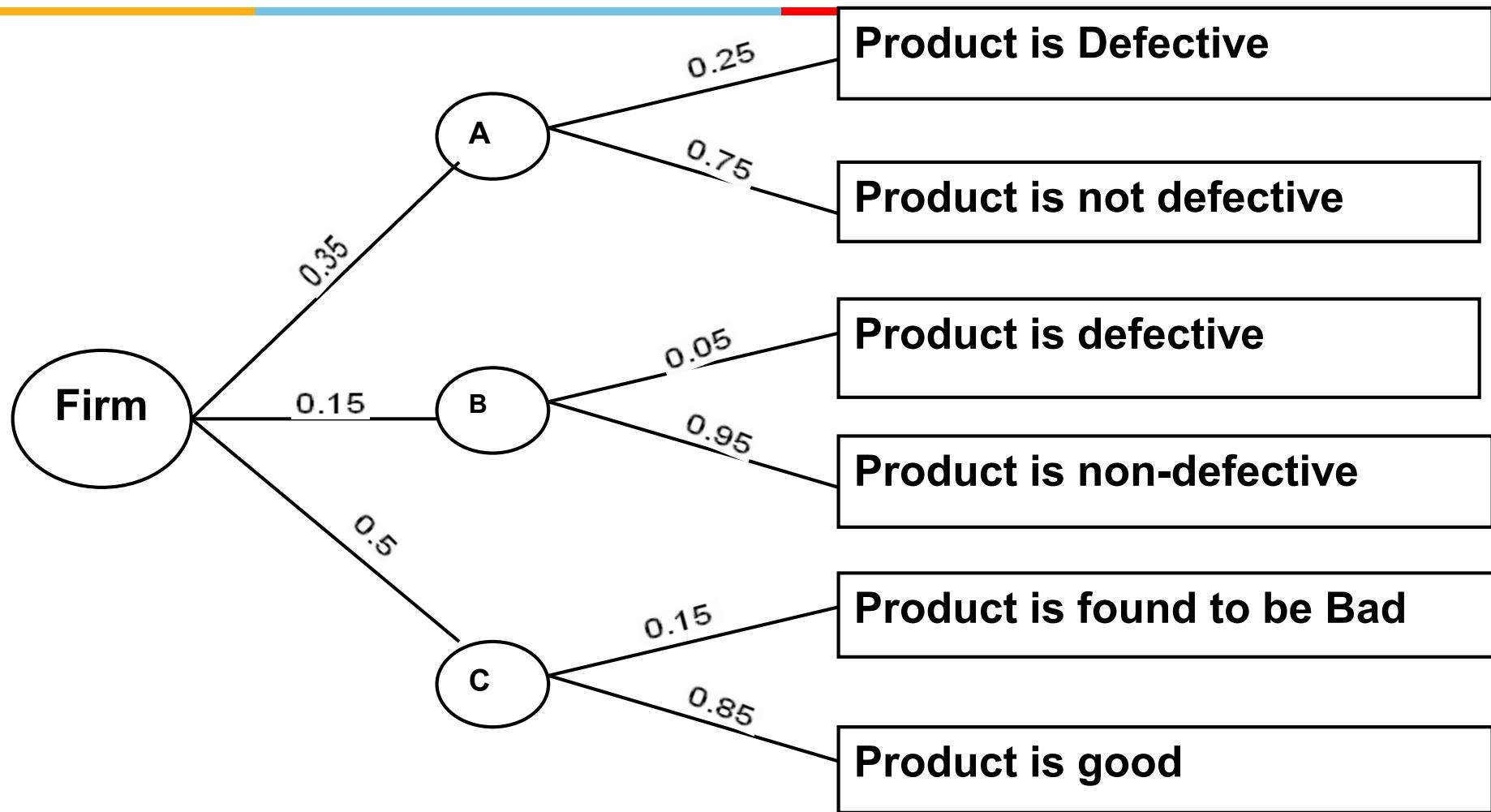
Next to each branch corresponding to a positive test result, the multiplication rule yields the recorded probabilities. Therefore, $P(B) = .00099 + .01998 = .02097$, from which we have

$$P(A_1 | B) = \frac{P(A_1 \cap B)}{P(B)} = \frac{.00099}{.02097} = .047$$

Example:

A certain firm has plants A, B, C producing, respectively 35%, 15% and 50% of the total output. The probabilities of a non – defective product are, respectively, 0.75, 0.95 and 0.85. A Customer receives a bad product, what is the Chance that product came from the plant C?

Tree Diagram



Solution

Let X : “Customer receives a defective product”.

$$\begin{aligned}\text{Clearly, } P(X) &= P(A)P\left[\frac{X}{A}\right] + P(B)P\left[\frac{X}{B}\right] + P(C)P\left[\frac{X}{C}\right] \\ &= 0.17\end{aligned}$$

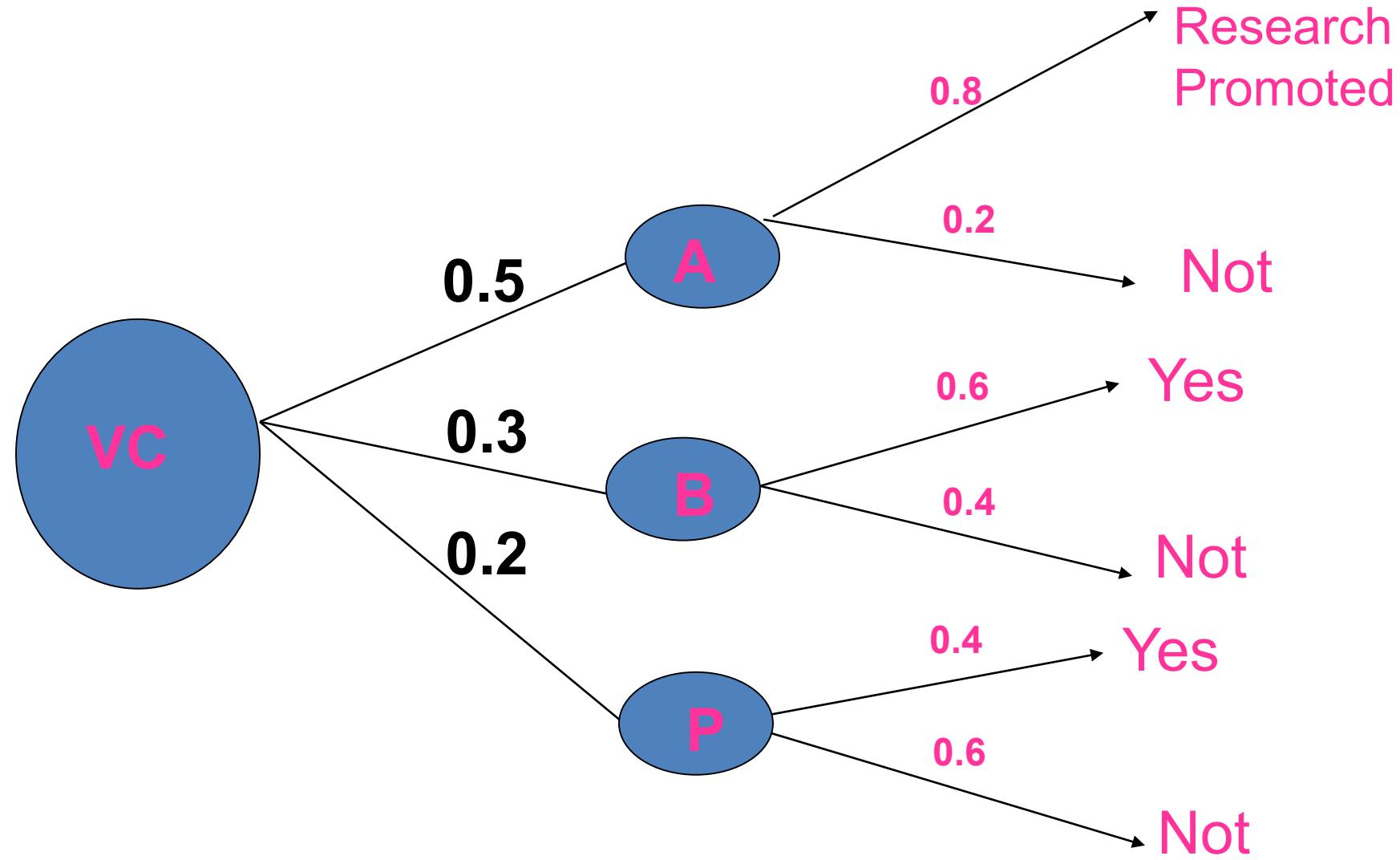
Therefore, the chance that product is manufactured by the plant C is

$$P(C | X) = \frac{P(C \cap X)}{P(X)} = \frac{0.5 \cdot 0.15}{0.17} = 0.4412$$

Example

The chances that an academician, a business man and a politician becoming Vice Chancellor of an university are 0.5, 0.3 and 0.2 respectively. The probability that research work will be promoted in the university by these 3 gentlemen are respectively are 0.8, 0.6 and 0.4. It is found Research work has been promoted by the university. What is the chance that an academician has become the VC?

Tree Diagram



Example

The chances that an academician, a business man and a politician becoming Vice Chancellor of an university are 0.5, 0.3 and 0.2 respectively. The probability that research work will be promoted in the university by these 3 gentlemen are respectively are 0.8, 0.6 and 0.4. It is found Research work has been promoted by the university. What is the chance that an academician has become the VC?

Let X : “Research work is promoted”

$$\text{Clearly, } P(X) = 0.5 \times 0.8 + 0.3 \times 0.6 + 0.2 \times 0.4 = 0.66$$

Now to find $P[\text{“An Academician is VC”} / \text{“Research work is promoted i.e. event } X\text{”}]$

$$= \frac{0.5 \times 0.8}{0.66} = 0.6061$$

Example: A manufacturer of tablets receives its LED screens from three different suppliers, 60% from supplier B_1 , 30% from supplier B_2 , and 10% from supplier B_3 . In other words, the probabilities that any one LED screens received by the plant comes from these three suppliers are 0.60, 0.30, and 0.10. Also suppose that 95% of the LED screens from B_1 , 80% of those from B_2 , and 65% of those from B_3 perform according to specifications.

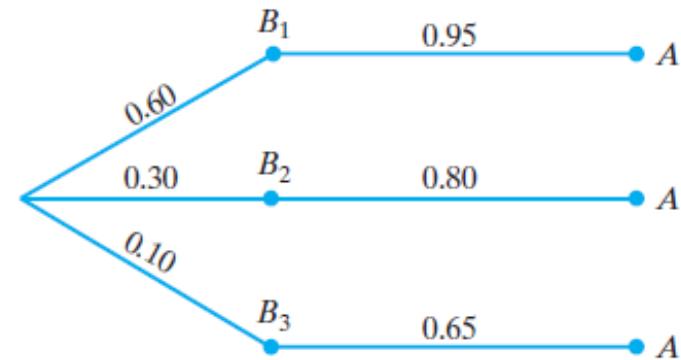
- 1) What is the probability that any one LED screen received by the plant will perform according to specifications?
- 2) Determine the probability that a particular LED screen, which is known to perform according to specifications, came from supplier B_3 .

If A denotes the event that a LED screen received by the plant performs according to specifications, and B_1 , B_2 , and B_3 are the events that it comes from the respective suppliers

$$\begin{aligned} A &= A \cap [B_1 \cup B_2 \cup B_3] \\ &= (A \cap B_1) \cup (A \cap B_2) \cup (A \cap B_3) \\ &= P(A \cap B_1) + P(A \cap B_2) + P(A \cap B_3) \end{aligned}$$

$$\begin{aligned} P(A) &= P(B_1) \cdot P(A | B_1) + \\ &\quad P(B_2) \cdot P(A | B_2) \\ &\quad + P(B_3) \cdot P(A | B_3) \end{aligned}$$

$$\begin{aligned} P(A) &= (0.60)(0.95) + (0.30)(0.80) + (0.10)(0.65) \\ &= 0.875 \end{aligned}$$



$$P(B_3 | A) = \frac{(0.10)(0.65)}{(0.60)(0.95) + (0.30)(0.80) + (0.10)(0.65)} = 0.074$$

Suggested Problems

Example: Two firms V and W consider bidding on a road-building job, which may or may not be awarded depending on the amounts of the bids. Firm V submits a bid and the probability is $\frac{3}{4}$ that it will get the job provided firm W does not bid. The probability is $\frac{3}{4}$ that W will bid, and if it does, the probability that V will get the job is only $\frac{1}{3}$.
(a) what is the probability that V will get the job? (b) If V gets the job, what is the probability that W did not bid?

Answer: Given $P(V/W^1) = \frac{3}{4}$, $P(W) = \frac{1}{3}$, $P(V/W) = \frac{1}{3}$ $P(W^1) = \frac{2}{3}$

$$(a) V = (V \cap W) \cup (V \cap W^1) \Rightarrow P(V) = P(V \cap W) + P(V \cap W^1)$$

$$P(V) = P(V/W) P(W) + P(V/W^1) P(W^1) = \frac{11}{18}$$

$$(b) P(W^1/V) = P(V/W^1) P(W^1) / P(V) = \frac{9}{11}$$

Suggested Problems

Example. An office has 4 secretaries handling respectively 20%, 60%, 15% and 5% of the files of all government reports. The probability that they misfile such reports are respectively 0.05, 0.1, 0.1 and 0.05. Find the probability that the misfiled report can be blamed on the first secretary.

Example . In a class 70% are boys and 30% are girls. 5% of boys and 3% of girls are irregular to the classes. What is the probability of a student selected at random is irregular to the classes and what is the probability that the irregular student is a girl?

Suggested Problems

Example 5. Three machines A, B and C produce respectively 60%, 30% and 10% of the total number of items of a factory. The percentage of defective outputs of these machines are 2%, 3% and 4%. An item is selected at random and is found to be defective. (i) Find the probability that the item was produced by machine C? (ii) What is the probability that the item was produced by machine C or B?

Bayesian Learning

- Naive Bayes is a set of simple and efficient machine learning algorithms for solving a variety of classification and regression problems.
- Naive Bayes assumes conditional independence where Bayes theorem does not. This means the relationship between all input features are independent.
- Bayesian learning algorithms that calculate explicit probabilities for hypotheses, such as the naive Bayes classifier, are among the most practical approaches to certain types of learning problems
- For example: Problem of learning to classify text documents such as electronic news articles.
- For such learning tasks, the naive Bayes classifier is among the most effective algorithms known

Features of Bayesian learning

- Prior knowledge can be combined with observed data to determine the final probability of a hypothesis.
 - Prior knowledge is provided by asserting
 - ❖ prior probability for each candidate hypothesis, and
 - ❖ probability distribution over observed data for each possible hypothesis.
 - New instances can be classified by combining the predictions of multiple hypotheses, weighted by their probabilities.
-

Bayes Theorem

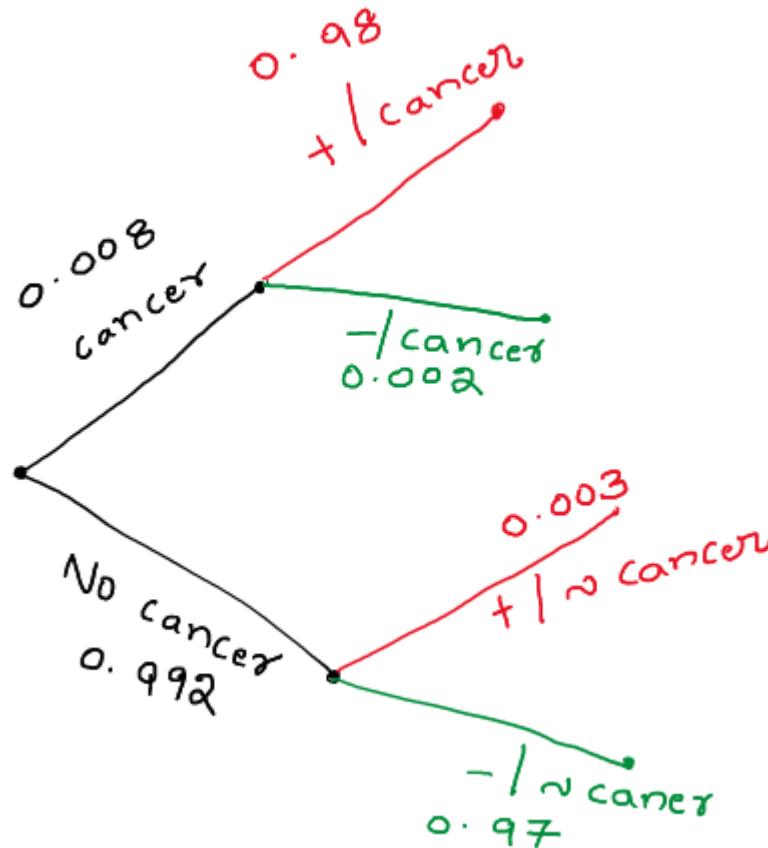
- $P(h)$ = prior probability of hypothesis h , before seeing the training data
- $P(D)$ = prior probability of training data D
- $P(h|D)$ = probability of h given D
- $P(D|h)$ = probability of D given h

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Bayes Theorem: Example

- Consider a medical diagnosis problem in which there are two alternative hypotheses:
 - H1:That a patient has a particular form of cancer
 - H2:That the patient does not
 - The available data is from a particular laboratory test with two possible outcomes
 - + Positive
 - Negative
 - Over the entire population of people only 0.008 have this disease. The test returns a corrective positive result in only 98% of the cases in which the disease is actually present and a correct negative result in only 97% of the cases in which the disease is not present.
 - How does $P(\text{cancer}/+)$ compare to $P(\sim \text{cancer}/+)$?
-

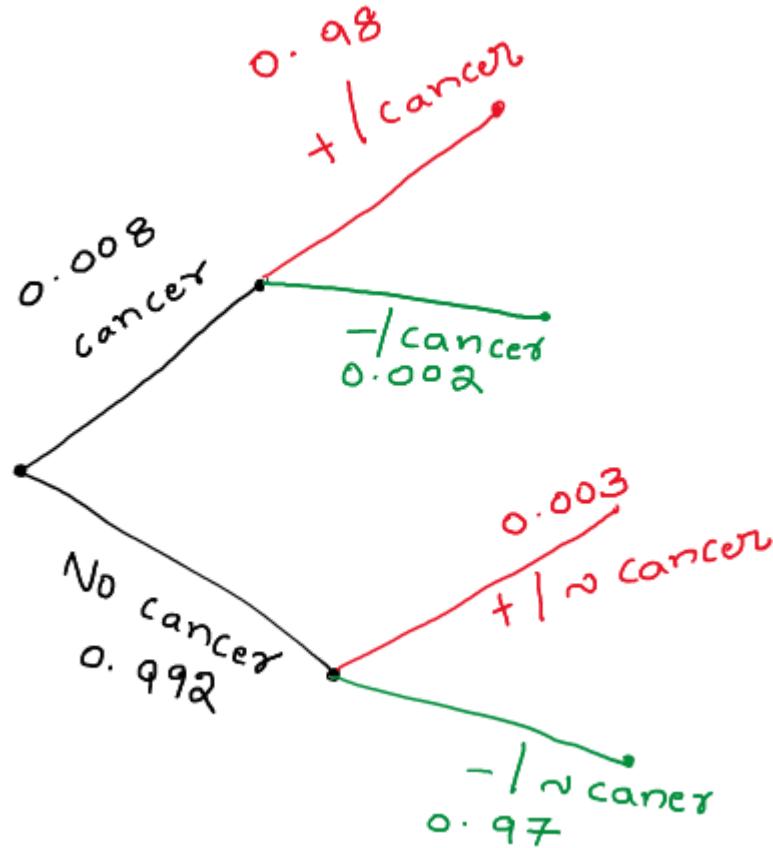
Bayes Theorem: Example



$$\begin{aligned}
 P(\text{cancer}/+) &= \frac{P(+/\text{cancer})P(\text{cancer})}{P(+)} \\
 &= \frac{0.98 \times 0.008}{0.98 \times 0.008 + 0.003 \times 0.992} \\
 &= 0.72485
 \end{aligned}$$

$$\begin{aligned}
 P(\sim\text{cancer}/+) &= \frac{P(+/\sim\text{cancer})P(\sim\text{cancer})}{P(+)} \\
 &= \frac{0.003 \times 0.992}{0.98 \times 0.008 + 0.003 \times 0.992} \\
 &= 0.27515
 \end{aligned}$$

Bayes Theorem: Example

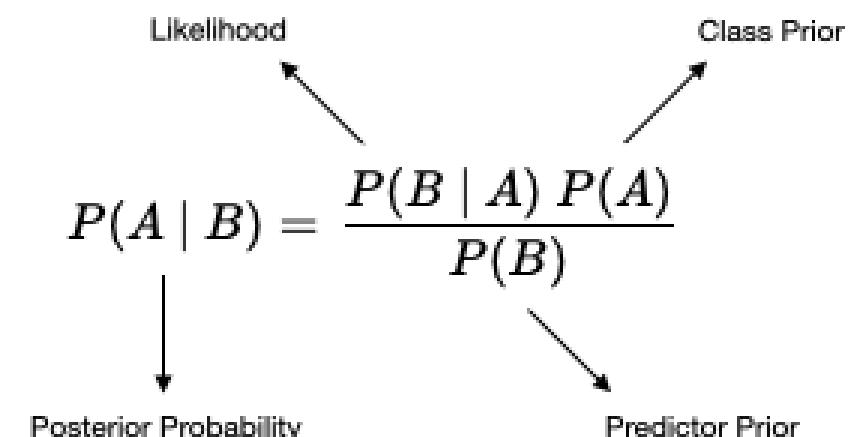


$$\begin{aligned}
 P(\text{cancer} | -) &= \frac{P(- | \text{cancer}) P(\text{cancer})}{P(-)} \\
 &= \frac{(0.02)(0.008)}{?} = \\
 P(\sim \text{cancer} | -) &= \frac{P(- | \sim \text{cancer}) P(\sim \text{cancer})}{P(-)} \\
 &= \frac{(0.97)(0.992)}{P(-)} =
 \end{aligned}$$

Machine Learning

- Generative models
 - Build model to estimate the posterior probability $P(Y|X)$ by estimating
 - likelihood of data given target (hypothesis) $P(X|Y)$
 - Prior probabilities over target $P(Y)$
 - In general, for a specific class $Y=c_k$,

$$P(Y = c_k | X) = \frac{P(X|Y = c_k) * P(Y=c_k)}{P(X)}$$



Choosing Hypotheses

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- Generally want the most probable hypothesis given the training data
- *Maximum a posteriori* hypothesis h_{MAP} :
$$\begin{aligned} h_{MAP} &= \arg \max_{h \in H} P(h|D) \\ &= \arg \max_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\ &= \arg \max_{h \in H} P(D|h)P(h) \end{aligned}$$
- If assume $P(h_i) = P(h_j)$ then can further simplify, and choose the *Maximum likelihood* (ML) hypothesis

$$h_{ML} = \arg \max_{h_i \in H} P(D|h_i)$$

Brute Force MAP Hypothesis

- For each hypothesis h in H , calculate the posterior probability

$$P(h | D) = \frac{P(D | h)P(h)}{P(D)}$$

- Output the hypothesis h_{MAP} with the highest posterior probability

$$h_{MAP} = \operatorname{argmax}_{h \in H} P(h | D)$$

MAP Hypothesis

- Using Bayes theorem, we compute the MAP hypothesis for all probable hypothesis (or all unique class labels)
- Identify the best hypothesis describing the data as

$$\begin{aligned} h_{MAP} &= \operatorname{argmax}_{h \in H} P(h|D) \\ &= \operatorname{argmax}_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\ &= \operatorname{argmax}_{h \in H} P(D|h)P(h) \end{aligned}$$

H: set of all hypothesis

P(D) is independent of h and is same for all hypothesis, therefore dropped

Maximum Likelihood Estimation



- When no prior information is available, all hypothesis are equally likely i.e $p(h_i) = p(h_j)$
- This is also true for a balanced class problem where all the classes are equally likely
- This is known as uniform prior
- MAP hypothesis further simplified to
- $h_{ML} = \operatorname{argmax} P(D/h)$ (where h belongs to H)

Conditional independence

- **Definition:** X is conditionally independent of Y given Z , if the probability distribution governing X is independent of the value of Y , given the value of Z

$$(\forall i, j, k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z_k)$$

$$P(X|Y, Z) = P(X|Z)$$

Example:

$$P(\text{Thunder}|\text{Rain, Lightning}) = P(\text{Thunder}|\text{Lightning})$$

Applying conditional independence

Naïve Bayes assumes X_i are conditionally independent given Y

e.g., $P(X_1|X_2, Y) = P(X_1|Y)$

$$\begin{aligned}P(X_1, X_2|Y) &= P(X_1|X_2, Y)P(X_2|Y) \\&= P(X_1|Y)P(X_2|Y)\end{aligned}$$

General form: $P(X_1, \dots, X_n|Y) = \prod_{j=1}^n P(X_j|Y)$

How many parameters to describe $P(X_1, \dots, X_n|Y)$? $P(Y)$?

Without conditional independence assumption?

With conditional independence assumption?

Naïve Bayes Independence assumption

Assumption:

$$P(X_1, \dots, X_n | Y) = \prod_{j=1}^n P(X_j | Y)$$

i.e., X_i and X_j are conditionally independent
given Y for $i \neq j$

Naïve Bayes classifier

- Bayes rule:

$$P(Y = y_k | X_1, \dots, X_n) = \frac{P(Y = y_k)P(X_1, \dots, X_n | Y = y_k)}{\sum_j P(Y = y_j)P(X_1, \dots, X_n | Y = y_j)}$$

- Assume conditional independence among X_i 's:

$$P(Y = y_k | X_1, \dots, X_n) = \frac{P(Y = y_k)\prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j)\prod_i P(X_i | Y = y_j)}$$

- Pick the most probable (MAP) Y

$$\hat{Y} \leftarrow \operatorname{argmax}_{y_k} P(Y = y_k)\prod_i P(X_i | Y = y_k)$$

↑
Prior
Probability
↑
MLE

NAÏVE BAYES CLASSIFIER

- Assume independence among attributes X_i when class is given:
 - ❖ $P(X_1, X_2, \dots, X_d | Y_j) = P(X_1 | Y_j) P(X_2 | Y_j) \dots P(X_d | Y_j)$
 - ❖ Now we can estimate $P(X_i | Y_j)$ for all X_i and Y_j combinations from the training data
 - ❖ New point is classified to Y_j if $P(Y_j) \prod P(X_i | Y_j)$ is maximal.

Example 1:

If the weather is sunny,
then the player will play
or not?

i.e. Play/ Sunny = Yes or No

Note if we know $P(\text{Yes/Sunny})$ and
 $P(\text{No/Sunny})$ then we can answer the
question asked

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Steps to Apply Bayes Theorem



Step 1- View or collect “raw” data.

Step 2 - Convert long data to a frequency table

weather	Play		Row Total
	no	yes	
Sunny	2	3	5
Overcast	0	4	4
Rainy	3	2	5
Column Total	5	9	14

Step 3 - Row and column sums to get probabilities

Weather probabilities

$$\text{sunny} = 5/14, \text{rainy} = 5/14$$

$$\text{Overcast} = 4/14$$

Play probabilities

$$\text{no} = 5/14$$

$$\text{yes} = 9/14$$

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Steps to Apply Bayes Theorem



		Play		
		no	yes	Row Total
weather				
Sunny		2	3	5
				P(Sunny)= 5/14
Overcast		0	4	4
				P(Overcast) = 4/14
Rainy		3	2	5
				P(Rainy)=5/14
Column Total		5	9	14
				P(no)=5/14 P(yes)=9/14

Step 4 - Apply probabilities from frequency table to Bayes theorem

$$P(\text{yes} \mid \text{sunny}) = \frac{P(\text{sunny} \mid \text{yes}) P(\text{yes})}{P(\text{sunny})}$$

weather	no	yes
Rainy	3	2
sunny	2	3
overcast	0	4
Total	5	9

$$\frac{5}{14} = 0.36$$

$$\frac{9}{14} = 0.64$$

$$\frac{4}{14} = 0.29$$

$$\frac{5}{14} = 0.36 \quad \frac{9}{14} = 0.64$$

$$\text{Now } P(\text{Yes} | \text{sunny}) = \frac{P(\text{sunny} | \text{yes}) P(\text{yes})}{P(\text{sunny})}$$

$$= \frac{(3/9)(9/14)}{5/14} = \underline{\underline{0.60}} \quad \checkmark$$

$$P(\text{no} | \text{sunny}) = \frac{(2/5)(5/14)}{5/14} = \underline{\underline{0.40}}$$

Example 2:

If the features of
today = (Outlook is Sunny, Temp is Hot, Humidity is Normal, Windy is False),
 then the player will play or not?

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

$$P(Y_{\text{Yes}} | X) = \frac{P(X | Y_{\text{Yes}}) P(Y_{\text{Yes}})}{P(X)}$$

$$P(Y_{\text{Yes}} | x_1, x_2, x_3, x_4) = \frac{P(x_1, x_2, x_3, x_4 | Y_{\text{Yes}}) P(Y_{\text{Yes}})}{P(x_1, x_2, x_3, x_4)}$$

$$= \frac{P(x_1 | Y_{\text{Yes}}) P(x_2 | Y_{\text{Yes}}) P(x_3 | Y_{\text{Yes}}) P(x_4 | Y_{\text{Yes}}) P(Y_{\text{Yes}})}{P(x_1) P(x_2) P(x_3) P(x_4)}$$

$$P(x_1, x_2, x_3, x_4) = P(x_1 \cap x_2 \cap x_3 \cap x_4)$$

$$P(x_1) =$$

$$P(x_2) =$$

$$P(x_3) =$$

$$P(x_4) =$$

$$P(x_1 | Y_{\text{Yes}})$$

$$P(x_2 | Y_{\text{Yes}})$$

$$P(x_3 | Y_{\text{Yes}})$$

$$P(x_4 | Y_{\text{Yes}})$$

today = (Sunny, Hot, Normal, False)

x_1, x_2 are independent

$$P(x_1 \cap x_2) = P(x_1) P(x_2)$$

Naive Bayes Classifier

Indep

Indep

Bayes

Naive

Outlook

	Yes	No	P(yes)	P(no)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0/5
Rainy	3	2	3/9	2/5
Total	9	5	100%	100%

Temperature

	Yes	No	P(yes)	P(no)
Hot	2	2	2/9	2/5
Mild	4	2	4/9	2/5
Cool	3	1	3/9	1/5
Total	9	5	100%	100%

Humidity

	Yes	No	P(yes)	P(no)
High	3	4	3/9	4/5
Normal	6	1	6/9	1/5
Total	9	5	100%	100%

Wind

	Yes	No	P(yes)	P(no)
False	6	2	6/9	2/5
True	3	3	3/9	3/5
Total	9	5	100%	100%

Play		P(Yes)/P(No)
Yes	9	9/14
No	5	5/14
Total	14	100%

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 = ?

$D = \{ \text{magazine promotion}, \text{watch}$
 $\text{promotion}, \text{Life Insurance}$
 $\text{promotion}, \text{credit card insurance} \}$

$h_i = \text{male or Female}$

$$P(\text{male} / \text{Yes, Yes, No, No}) = ?$$

$$P(\text{Female} / \text{Yes, Yes, No, No}) = ?$$

$$\frac{P(\text{Yes/Male}) \cdot P(\text{W.P.YN/Male}) \cdot P(\text{LI=No/Male}) \cdot P(\text{CC/No})}{P(\text{YN}) P(\text{YN}) P(\text{No}) P(\text{No})}$$

		magazine promotion		watch promotion		L.I Promotion		credit Card Promotion	
		Male	Female	Male	Female	Male	F	M	F
YES		4	3	2	2	2	3	2	1
	NO	2	1	4	2	4	1	4	3
Ratios (YES)		$\frac{4}{6}$	$\frac{3}{4}$	$\frac{2}{6}$	$\frac{2}{4}$	$\frac{2}{6}$	$\frac{3}{4}$	$\frac{2}{6}$	$\frac{1}{4}$
		$\frac{2}{6}$	$\frac{1}{4}$	$\frac{4}{6}$	$\frac{2}{4}$	$\frac{4}{6}$	$\frac{1}{4}$	$\frac{4}{6}$	$\frac{3}{4}$

$$P(\text{male} | E)$$

$$= \frac{P(E | \text{male}) P(\text{male})}{P(E)}$$

$$= \left(\frac{4}{6} \cdot \frac{2}{6} \cdot \frac{4}{6} \cdot \frac{4}{6} \right) \left(\frac{3}{5} \right)$$

$$P(E)$$

$$\frac{0 \cdot 0593}{P(E)}$$

YES

YES

NO

NO

$$\frac{6}{10} \cdot \frac{3}{5}$$

$$P(\text{Female} | E)$$

$$= \frac{P(E | \text{Female}) P(\text{Female})}{P(E)}$$

$$= \frac{\left(\frac{3}{4}\right)\left(\frac{2}{4}\right)\left(\frac{1}{4}\right)\left(\frac{3}{4}\right) \cdot \frac{2}{5}}{P(E)}$$

$$= \frac{\left(\frac{9}{128}\right)\left(\frac{2}{5}\right)}{P(E)} = \frac{0.0281}{P(E)}$$

$$0.0593 > 0.0281$$

is male ✓

Example 4:

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Give Birth	Can Fly	Live in Water	Have Legs	Class
yes	no	yes	no	?

A: attributes

M: mammals

N: non-mammals

$$P(A|M) = \frac{6}{7} \times \frac{6}{7} \times \frac{2}{7} \times \frac{2}{7} = 0.06$$

$$P(A|N) = \frac{1}{13} \times \frac{10}{13} \times \frac{3}{13} \times \frac{4}{13} = 0.0042$$

$$P(A|M)P(M) = 0.06 \times \frac{7}{20} = 0.021$$

$$P(A|N)P(N) = 0.004 \times \frac{13}{20} = 0.0027$$

P(A|M)P(M) > P(A|N)P(N)

=> Mammals

Issues with Naïve Bayes Classifier

Naïve Bayes Classifier:

$$P(\text{Refund} = \text{Yes} | \text{No}) = 3/7$$

$$P(\text{Refund} = \text{No} | \text{No}) = 4/7$$

$$P(\text{Refund} = \text{Yes} | \text{Yes}) = 0$$

$$P(\text{Refund} = \text{No} | \text{Yes}) = 1$$

$$P(\text{Marital Status} = \text{Single} | \text{No}) = 2/7$$

$$P(\text{Marital Status} = \text{Divorced} | \text{No}) = 1/7$$

$$P(\text{Marital Status} = \text{Married} | \text{No}) = 4/7$$

$$P(\text{Marital Status} = \text{Single} | \text{Yes}) = 2/3$$

$$P(\text{Marital Status} = \text{Divorced} | \text{Yes}) = 1/3$$

$$P(\text{Marital Status} = \text{Married} | \text{Yes}) = 0$$

- | $P(\text{Yes}) = 3/10$

- | $P(\text{No}) = 7/10$

- | $P(\text{Yes} | \text{Married}) = 0 \times 3/10 / P(\text{Married})$

- | $P(\text{No} | \text{Married}) = 4/7 \times 7/10 / P(\text{Married})$

For Taxable Income:

If class = No: sample mean = 110

sample variance = 2975

If class = Yes: sample mean = 90

sample variance = 25

Issues with Naïve Bayes Classifier

Consider the table with Tid = 7 deleted

Tid	Refund	Marital Status	Taxable Income	Evade
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	120K	Yes
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Naïve Bayes Classifier:

$$P(\text{Refund} = \text{Yes} | \text{No}) = 2/6$$

$$P(\text{Refund} = \text{No} | \text{No}) = 4/6$$

→ $P(\text{Refund} = \text{Yes} | \text{Yes}) = 0$

$$P(\text{Refund} = \text{No} | \text{Yes}) = 1$$

$$P(\text{Marital Status} = \text{Single} | \text{No}) = 2/6$$

→ $P(\text{Marital Status} = \text{Divorced} | \text{No}) = 0$

$$P(\text{Marital Status} = \text{Married} | \text{No}) = 4/6$$

$$P(\text{Marital Status} = \text{Single} | \text{Yes}) = 2/3$$

$$P(\text{Marital Status} = \text{Divorced} | \text{Yes}) = 1/3$$

$$P(\text{Marital Status} = \text{Married} | \text{Yes}) = 0/3$$

For Taxable Income:

If class = No: sample mean = 91

sample variance = 685

If class = Yes: sample mean = 90

sample variance = 25

Given $X = (\text{Refund} = \text{Yes}, \text{Divorced}, 120\text{K})$

$$P(X | \text{No}) = 2/6 \times 0 \times 0.0083 = 0$$

$$P(X | \text{Yes}) = 0 \times 1/3 \times 1.2 \times 10^{-9} = 0$$

**Naïve Bayes will not be able to
classify X as Yes or No!**

Naïve Bayes for Text Classification



- Naïve Bayes is commonly used for **text classification**
- For a document with k terms $d = (t_1, \dots, t_k)$

Fraction of documents in c

$$P(c|d) = P(c)P(d|c) = P(c) \prod_{t_i \in d} P(t_i|c)$$

- $P(t_i|c)$ = Fraction of terms from **all documents** in c that are t_i

Number of times t_i appears in some document in c

$$P(t_i|c) = \frac{N_{ic} + 1}{N_c + T}$$

Laplace Smoothing

Total number of terms in all documents in c

Number of unique words (vocabulary size)

- Easy to implement and works relatively well
- **Limitation:** Hard to incorporate **additional features** (beyond words).
 - E.g., number of adjectives used.

A Simple Example:

Text	Tag
"A great game"	Sports ✓
"The election was over"	Not sports ✓
"Very clean match"	Sports ✓
"A clean but forgettable game"	Sports ✓
"It was a close election"	Not sports ✓

Which tag does the sentence *A very close game* belong to? i.e. $P(\text{sports} | \text{A very close game})$

Feature Engineering: Bag of words i.e use word frequencies without considering order

Using Bayes Theorem:

$$P(\text{sports} | \text{A very close game}) \\ = P(\text{A very close game} | \text{sports}) P(\text{sports})$$

$P(\text{A very close game})$

$$= P(\text{A} | \text{sports}) \cdot P(\text{very} | \text{sports}) \\ P(\text{close} | \text{sports}) \\ P(\text{game} | \text{sports})$$

(0) (0) (0)

We assume that every word in a sentence is **independent** of the other ones

"close" doesn't appear in sentences of sports tag, So $P(\text{close} | \text{sports}) = 0$, which makes product 0

A Simple Example

Text	Tag	
"A great game"	Sports	Which tag does the sentence <i>A very close game</i> belong to? i.e. $P(\text{sports} \mid A \text{ very close game})$
"The election was over"	Not sports	Feature Engineering: Bag of words i.e use word frequencies without considering order
"Very clean match"	Sports	Using Bayes Theorem:
"A clean but forgettable game"	Sports	$P(\text{sports} \mid A \text{ very close game})$ = $\frac{P(A \text{ very close game} \mid \text{sports}) P(\text{sports})}{P(A \text{ very close game})}$
"It was a close election"	Not sports	

We assume that every word in a sentence is **independent** of the other ones

$$P(A \text{ very close game}) = P(A) P(\text{very}) P(\text{close}) P(\text{game})$$

$$P(A \text{ very close game} \mid \text{sports}) = \frac{P(a \mid \text{sports}) P(\text{very} \mid \text{sports})}{P(\text{close} \mid \text{sports}) P(\text{game} \mid \text{sports})}$$

"close" doesn't appear in sentences of sports tag, So $P(\text{close} \mid \text{sports}) = 0$, which makes product 0

Laplace smoothing

Naive Bayes with
Laplace Smoothing



- Laplace smoothing: we add 1 or in general constant k to every count so it's never zero.
- To balance this, we add the number of possible words to the divisor, so the division will never be greater than 1
- In our case, the possible words are ['a', 'great', 'very', 'over', 'it', 'but', 'game', 'election', 'clean', 'close', 'the', 'was', 'forgettable', 'match'].

Apply Laplace Smoothing

Word	P(word Sports)	P(word Not Sports)
a	2+1 / 11+14	1+1 / 9+14
very	1+1 / 11+14	0+1 / 9+14
close	0+1 / 11+14	1+1 / 9+14
game	2+1 / 11+14	0+1 / 9+14

$$\begin{aligned}
 & P(a|Sports) \times P(very|Sports) \times P(close|Sports) \times P(game|Sports) \times \\
 & P(Sports) \\
 & = 2.76 \times 10^{-5} \\
 & = 0.0000276
 \end{aligned}$$

$$\begin{aligned}
 & P(a|Not\ Sports) \times P(very|Not\ Sports) \times P(close|Not\ Sports) \times \\
 & P(game|Not\ Sports) \times P(Not\ Sports) \\
 & = 0.572 \times 10^{-5} \\
 & = 0.00000572
 \end{aligned}$$

Example :

Doc No	Text
1	I LOVED THE MOVIE
2	I HATED THE MOVIE
3	A GREAT MOVIE ,GOOD MOVIE
4	POOR ACTING
5	GREAT ACTING , A GOOD MOVIE
NEW	I HATED THE POOR ACTING

Example :

Doc No	Text	
1	I LOVED THE MOVIE	<u>POSITIVE</u>
2	I HATED THE MOVIE	<u>NEGATIVE</u>
3	A GREAT MOVIE ,GOOD MOVIE	<u>POSITIVE</u>
4	POOR ACTING	<u>NEGATIVE</u>
5	GREAT ACTING , A GOOD MOVIE	<u>POSITIVE</u>
NEW	I HATED THE POOR ACTING	<u>????</u>

$$P(c/x)$$

$$P(+ / \text{I hated the poor acting}) =$$

$$P(- / \text{I hated the poor acting}) =$$

Based on these probabilities, we can decide the class which the new text belongs

$P(+ | \text{I hated the acting})$

i.e. $P(c_1 | x) = \frac{P(x | c_1) P(c_1)}{P(x)}$

$$= P(\text{I} | +) P(\text{hated} | +) P(\text{the} | +) P(\text{acting} | +) P(+)$$

$$\frac{P(\text{I}, +)}{P(+)}$$

$$\frac{P(\text{hated}, +)}{P(+)}$$

words	positive	negative
I	1	1
loved	1	0
the	1	1
movie	4	1
hated	0	1
a	2	0
great	2	0
Poor	0	1
acting	1	1
good	2	0

$$P(\pm | +)$$

$$= \frac{1 + 1}{14 + 10}$$

$$P(I | -)$$

$$= \frac{1 + 1}{6 + 10}$$

"I hated the poor
acting"

word

I

$$\frac{1+1}{14+10} = 0.0833$$

hated

$$\frac{0+1}{14+10} = 0.0417$$

the

$$\frac{1+1}{14+10} = 0.0833$$

poor

$$\frac{0+1}{14+10} = 0.0417$$

acting

$$\frac{1+1}{14+10} = 0.0833$$

positive

negative

$$\frac{1+1}{6+10} = 0.125$$

$x : I$ hate the Poor
acting

$$P(+|x)$$

$$= () () () () () \times P(+)$$

\downarrow
 $3/5$

$$= 6.03 \times 10^{-7}$$

$$P(-|x)$$

$$= () () () () () () P(-) :$$

\downarrow
 $2/8$

$$= 1.22 \times 10^{-5}$$

\therefore negative class

Example:

Suppose we got the new message with the words '**Dear Friend**', Decide whether this new message is a normal or spam message?

i.e. Normal/ Dear, Friend = Yes or No

Note if we know $P(\text{Normal} / \text{Dear, Friend})$ and $P(\text{Spam} / \text{Dear, Friend})$ then we can answer the question asked

Email word	Spam	Email word	Spam
Dear	Yes	Friend	No
Friend	No	Friend	Yes
Dear	No	Dear	No
Dear	No	Lunch	No
Dear	No	Friend	No
Friend	No	Dear	No
Lunch	No	Dear	No
Friend	No	Dear	No
Lunch	No	Dear	No
Dear	Yes	Money	Yes
Money	Yes	Money	No
Money	Yes	Money	Yes

Steps to Apply Bayes Theorem



Step 1- View or collect “raw” data.

Step 2 - Convert long data to a frequency table

	Play		Row Total
word	normal	spam	
Dear	8	2	10
Friend	5	1	6
Lunch	3	0	3
Money	1	4	5
Column Total	17	7	24

Step 3 - Row and column sums to get probabilities

Email word	Spam	Email word	Spam
Dear	Yes	Friend	No
Friend	No	Friend	Yes
Dear	No	Dear	No
Dear	No	Lunch	No
Dear	No	Friend	No
Friend	No	Dear	No
Lunch	No	Dear	No
Friend	No	Dear	No
Lunch	No	Dear	No
Dear	Yes	Money	Yes
Money	Yes	Money	No
Money	Yes	Money	Yes

Steps to Apply Bayes Theorem



Play

	normal	spam	Row Total
word			
Dear	8	2	10
Friend	5	1	6
Lunch	3	0	3
Money	1	4	5
Column Total	17	7	24

As $P(N/D, F) > P(S/D, F)$, we can decide that Dear Friend is Normal message.

Step 4 - Apply probabilities from frequency table to Bayes theorem

$$P(N/D, F) = \frac{P(D, F/N) \cdot P(N)}{P(D, F)} = \frac{P(D/N) \cdot P(F/N) \cdot P(N)}{P(D) \cdot P(F)} = \frac{\left(\frac{8}{17}\right) \cdot \left(\frac{5}{17}\right) \cdot \left(\frac{17}{24}\right)}{\left(\frac{10}{24}\right) \cdot \left(\frac{6}{24}\right)} = \frac{0.098}{0.104} = 0.9423$$

$$P(S/D, F) = \frac{P(D, F/S) \cdot P(S)}{P(D, F)} = \frac{P(D/S) \cdot P(F/S) \cdot P(S)}{P(D) \cdot P(F)} = \frac{\left(\frac{2}{7}\right) \cdot \left(\frac{1}{7}\right) \cdot \left(\frac{7}{24}\right)}{\left(\frac{10}{24}\right) \cdot \left(\frac{6}{24}\right)} = \frac{0.012}{0.104} = 0.1153$$

Example continued:

Suppose we got the new message contains the word '**Lunch Money Money Money Money**' , Decide whether this new message is a normal or spam message?

Email word	Spam	Email word	Spam
Dear	Yes	Friend	No
Friend	No	Friend	Yes
Dear	No	Dear	No
Dear	No	Lunch	No
Dear	No	Friend	No
Friend	No	Dear	No
Lunch	No	Dear	No
Friend	No	Dear	No
Lunch	No	Dear	No
Dear	Yes	Money	Yes
Money	Yes	Money	No
Money	Yes	Money	Yes

Steps to Apply Bayes Theorem



Play

	normal	spam	Row Total
word			
Dear	8	2	10
Friend	5	1	6
Lunch	3	0	3
Money	1	4	5
Column Total	17	7	24

We can observe that we have to classify any message with Lunch as Normal message, no matter how many times we see the word Money and that's the problem.

To work around this problem add 1 count to the frequency table to each word(Laplace smoothing)

Step 4 - Apply probabilities from frequency table

$$P(N) \cdot P(L/N) \cdot P(M/N)^4 = \left(\frac{17}{24}\right) \cdot \left(\frac{3}{17}\right) \cdot \left(\frac{1}{17}\right)^4 = 0.0000015$$

$$P(S) \cdot P(L/S) \cdot P(M/S)^4 = \left(\frac{7}{24}\right) \cdot \left(\frac{0}{7}\right) \cdot \left(\frac{4}{7}\right)^4 = 0$$

Steps to Apply Bayes Theorem



Play

	normal	spam	Row Total
word			
Dear	8+1	2+1	12
Friend	5+1	1+1	8
Lunch	3+1	0+1	5
Money	1+1	4+1	7
Column Total	21	11	32

As $P(S/L, M^4) > P(N/L, M^4)$, we can decide that
Lunch Money Money Money Money is Spam
message.

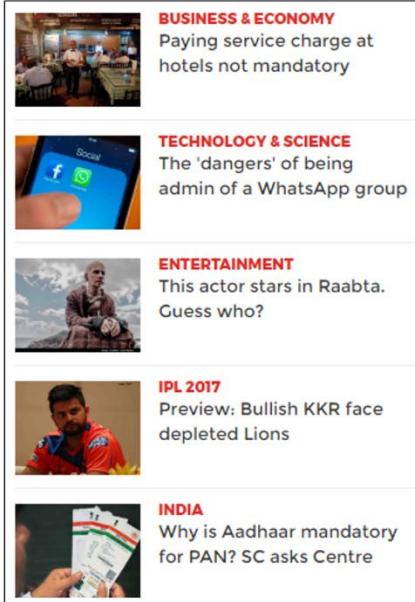
Step 4 - Apply probabilities from frequency table

$$P(N) \cdot P(L/N) \cdot P(M/N)^4 = \left(\frac{21}{32}\right) \cdot \left(\frac{4}{21}\right) \cdot \left(\frac{2}{21}\right)^4 = 0.00001$$

$$P(S) \cdot P(L/S) \cdot P(M/S)^4 = \left(\frac{11}{32}\right) \cdot \left(\frac{1}{11}\right) \cdot \left(\frac{5}{11}\right)^4 = 0.00133$$

Naïve Bayes Classifier Applications

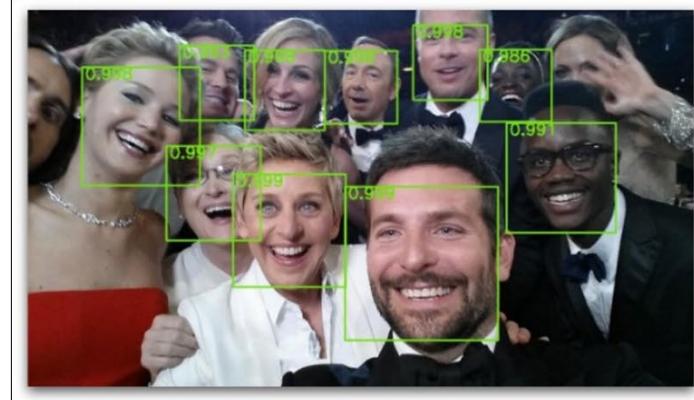
Categorizing News



Email Spam Detection



Face Recognition



Sentiment Analysis



Naive Bayes Classifier

- ✓ Along with decision trees, neural networks, one of the most practical learning methods.
 - ✓ When to use
 - ✓ Moderate or large training set available
 - ✓ Attributes that describe instances are conditionally independent given classification
 - ✓ Successful applications:
 - ✓ Diagnosis
 - ✓ Classifying text documents
-

Learning to Classify Text

- Why?
 - ❖ Learn which news articles are of interest
 - ❖ Learn to classify web pages by topic
 - Naive Bayes is among most effective algorithms
 - What attributes shall we use to represent text documents??
-

Baseline: Bag of Words Approach



Practical Issues of Bayesian learning



- Require initial knowledge of many probabilities
 - Often estimated based on background knowledge, previously available data, and assumptions about the form of the underlying distributions.
- Significant computational cost required to determine the Bayes optimal hypothesis in the general case (linear in the number of candidate hypotheses)

HW: Exercise

Consider the car theft problem with attributes Color, Type, Origin, and the target, Stolen can be either Yes or No.

we need to classify whether the car is stolen, given the features of the car.

Given the Red color Domestic SUV car Find the probability of whether the car is stolen?

Color	Type	Origin	Stolen?
Red	Sports	Domestic	Yes
Red	Sports	Domestic	No
Red	Sports	Domestic	Yes
Yellow	Sports	Domestic	No
Yellow	Sports	Imported	Yes
Yellow	SUV	Imported	No
Yellow	SUV	Imported	Yes
Yellow	SUV	Domestic	No
Red	SUV	Imported	No
Red	Sports	Imported	Yes

HW: Exercise

If the weather is Snowy,
then the player will play
or not?

weather	Player play
Sunny	yes
Rainy	no
Cloudy	yes
Sunny	no
Sunny	yes
snowy	no
Rainy	yes
Cloudy	no
Cloudy	yes
Sunny	yes
snowy	no
Cloudy	yes
Rainy	no
snowy	no
snowy	yes





Thanks



BITS Pilani
Pilani Campus

M.Tech. (AIML)

Session-5 (Random Variables)

Team ISM

Session-5 Agenda

Random variables –

Discrete & continuous Expectation of a random variable,
mean and variance of a random variable –

Single random random variable &

Joint distributions

 Contact Session 5: Module 3: Probability Distributions

Contact Session	List of Topic Title	Reference
CS - 5	Random variables - Discrete & continuous Expectation of a random variable, mean and variance of a random variable – Single random random variable & Joint distributions	T1 & T2
HW	Problems on random variables	T1 & T2
Lab	Probability Distributions & Sampling	Lab 3

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.
 - In mathematical language, a random variable is a function whose domain is the sample space and whose range is the set of real numbers.
-

Random Variables

- 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 finite or countably infinite number of values
- A continuous random variable may assume any numerical value in an interval or collection of intervals.
- Continuous random variables are generated in experiments where things are “measured” as opposed to “counted”.
- Experimental outcomes based on measurement of time, distance, weight, volume etc. generate continuous RV.

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

Two 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

Discrete Probability Distributions

- The probability distribution for a random variable describes how probabilities are distributed over the values of the random variable.
- The probability distribution is defined by a probability 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:

$$f(x) \geq 0$$

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

- We can describe a discrete probability distribution with a **table, graph, or equation**.
 - Advantage: once the probability distribution is known, it is relatively easy to determine the probability of a variety of events that may be of interest to the decision maker.
-

Probability Distributions for Discrete Random Variables

- Say a random variable x follows this pattern: $p(x) = (.3)(.7)^{x-1}$ for $x > 0$.
 - This table gives the probabilities (rounded to two digits) for x between 1 and 10.

x	$P(x)$
1	.30
2	.21
3	.15
4	.11
5	.07
6	.05
7	.04
8	.02
9	.02
10	.01

Expected Value and Variance

- The expected value, or mean, of a random variable is a measure of its central location.
 - The mean or Expected value of a discrete random variable:

$$E(x) = \mu = \sum xf(x)$$

- The variance summarizes the variability in the values of a random variable.
 - Variance of a discrete random variable:

$$\text{Var}(x) = \sigma^2 = \sum (x - \mu)^2 f(x)$$

- The standard deviation, σ , is defined as the positive square root of the variance.
-

Rules of Expected Value

- Multiplying RV by a constant a , $E(aX) = a.E(X)$
 - Adding a constant b , $E(X+b) = E(X) + b$
 - Therefore, $E(aX + b) = ?$
-

Variability 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

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

Rules of variability

- Multiplying RV by a constant a , $V(aX) = a^2 \cdot V(X)$
 - Adding a constant b , $V(X+b) = V(X)$
 - $\sigma_{aX} = |a| \cdot \sigma_X, \quad \sigma_{X+b} = \sigma_X$
-

Example:

At a shooting range, a shooter is able to hit a target in either 1, 2 or 3 shots. Let x be a random variable indicating the number of shots fired to hit the target. The following probability function was proposed.

$$f(x) = x/6$$

Is this probability function valid?

Identify the r.v to be discrete or continuous?

EXPERIMENT	Random Variable (x)
Audit 50 tax returns	Number of returns that contains error
Operate a restaurant for one day	Number of customers
Observe an employee's work	No. of productive hours in an 8-hour workday

Example: JSL Appliances

- Discrete random variable with a finite number of values
 - Let x = number of TV sets sold at the store in one day
where x can take on 5 values (0, 1, 2, 3, 4)

 - Discrete random variable with an infinite sequence of values
 - Let x = number of customers arriving in one day
where x can take on the values 0, 1, 2, ...
 - We can count the customers arriving, but there is no finite upper limit on the number that might arrive.
-

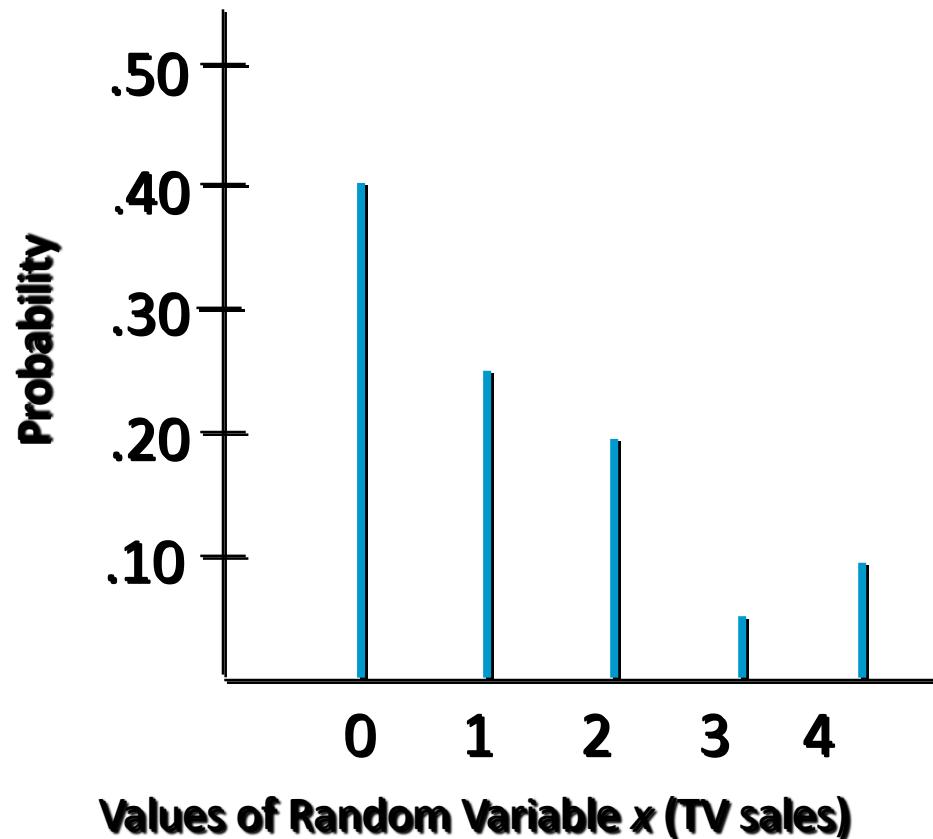
Example : JSL Appliances

Using past data on TV sales (below left), a tabular representation of the probability distribution for TV sales (below right) was developed.

Units Sold	No of days	x	f(x)
0	80	0	0.4
1	50	1	0.25
2	40	2	0.2
3	10	3	0.05
4	20	4	0.1
Total	200	1	

Example: JSL Appliances

- Graphical Representation of the Probability Distribution



Example: JSL Appliances

Expected Value of a Discrete Random Variable

x	f(x)	xf(x)
0	0.4	0.00
1	0.25	0.25
2	0.2	0.40
3	0.05	0.15
4	0.1	0.40

$$E(x) = 1.20$$

The expected number of TV sets sold in a day is 1.2

Example: JSL Appliances

- Variance and Standard Deviation of a Discrete Random Variable

x	$x - \mu$	$(x - \mu)^2$	$f(x)$	$(x - \mu)^2 f(x)$
0	-1.2	1.44	.40	.576
1	-0.2	0.04	.25	.010
2	0.8	0.64	.20	.128
3	1.8	3.24	.05	.162
4	2.8	7.84	.10	<u>.784</u>
				1.660 = σ^2

- The variance of daily sales is 1.66 TV sets squared.
- The standard deviation of sales is 1.2884 TV sets.

Discrete Uniform Probability Distribution

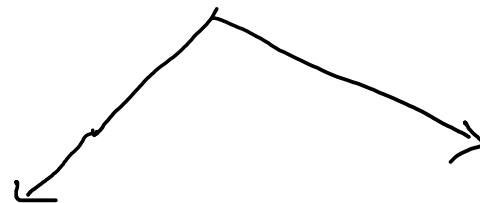
- The discrete uniform probability distribution is the simplest example of a discrete probability distribution given by a formula.
- The discrete uniform probability function is

$$f(x) = 1/n$$

where:

n = the number of values the random variable may assume
Note that the values of the random variable are equally likely.

Random Variables



Discrete

$$X = \{1, 2, 3, 4\}$$

→ No of students

→ No of bits
transmitted

continuous

$$X \in (1, 3)$$

- ↓
→ pressure
→ Temperature
→ Time
→ Voltage

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

Example::

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

compute the following

- $E(x)$
- $V(x)$ directly from the definition
- standard deviation of X
- $V(x)$ 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 $P(x \geq y_2)$

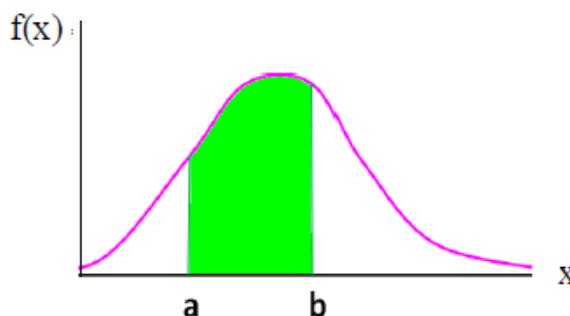
Continuous Probability Distributions

- A continuous random variable can assume any value in an interval on the real line or in a collection of intervals.
 - It is not possible to talk about the probability of the random variable assuming a particular value.
 - Instead, we talk about the probability of the random variable assuming a value within a given interval.
 - The probability of the random variable assuming a value within some given interval from x_1 to x_2 is defined to be the area under the graph of the probability density function between x_1 and x_2 .
-

Continuous Random Variables

A continuous random variable can assume any value in an interval on the real line or in a collection of intervals.

The probability of the random variable assuming a value within some given interval from x_1 to x_2 is defined to be the area under the graph of the **probability density function** between x_1 and x_2



Example:

- Height of students in a class
- Amount of ice tea in a glass
- Change in temperature throughout a day
- Price of a car in next year

Continuous Random Variables

Probability Density Function

For a continuous random variable X , a **probability density function** is a function such that

$$(1) \quad f(x) \geq 0$$

$$(2) \quad \int_{-\infty}^{\infty} f(x) dx = 1$$

$$(3) \quad P(a \leq X \leq b) = \int_a^b f(x) dx = \text{area under } f(x) \text{ from } a \text{ to } b \text{ for any } a \text{ and } b \quad (4.1)$$

Continuous Random Variables

Cumulative Distribution Function

The **cumulative distribution function** of a continuous random variable X is

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(u) du$$

for $-\infty < x < \infty$.

Probability Density Function from the Cumulative Distribution Function

Given $F(x)$,

$$f(x) = \frac{dF(x)}{dx}$$

as long as the derivative exists.

Continuous Random Variables

Mean and Variance

Suppose that X is a continuous random variable with probability density function $f(x)$. The **mean or expected value** of X , denoted as μ or $E(X)$, is

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

The **variance** of X , denoted as $V(X)$ or σ^2 , is

$$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2$$

The **standard deviation** of X is $\sigma = \sqrt{\sigma^2}$.

Integration Formulas

$$\int kf(u)du = k \int f(u)du$$

$$\int u^n du = \frac{u^{n+1}}{n+1}$$

$$\int e^u du = e^u$$

$$\int \sin u du = -\cos u$$

$$\int \cos u du = \sin u$$

$$\int [f(u) \pm g(u)] du = \int f(u)du \pm \int g(u)du$$

$$\int udv = uv - \int vdu$$

Continuous Random Variables

EXAMPLE I

Calculating probabilities from the probability density function

If a random variable has the probability density

$$f(x) = \begin{cases} 2e^{-2x} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

find the probabilities that it will take on a value

- between 1 and 3;
- greater than 0.5.

Solution Evaluating the necessary integrals, we get

$$(a) \quad \int_1^3 2e^{-2x} dx = e^{-2} - e^{-6} = 0.133$$

$$(b) \quad \int_{0.5}^{\infty} 2e^{-2x} dx = e^{-1} = 0.368$$



With reference to the preceding example, find the distribution function and use it to determine the probability that the random variable will take on a value less than or equal to 1.

Performing the necessary integrations, we get

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \int_0^x 2e^{-2t} dt = 1 - e^{-2x} & \text{for } x > 0 \end{cases}$$

and substitution of $x = 1$ yields

$$F(1) = 1 - e^{-2} = 0.865$$



Determining the mean and variance using the probability density function

With reference to Example 1, find the mean and the variance of the given probability density.

Performing the necessary integrations, using integrations by parts, we get

$$\mu = \int_{-\infty}^{\infty} xf(x) dx = \int_0^{\infty} x \cdot 2 e^{-2x} dx = \frac{1}{2}$$

Alternatively, the expectation of x is $E(X) = 0.5$

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = \int_0^{\infty} \left(x - \frac{1}{2}\right)^2 \cdot 2 e^{-2x} dx = \frac{1}{4}$$



A probability density function assigns probability one to $(-\infty, \infty)$

Find k so that the following can serve as the probability density of a random variable:

$$f(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ kxe^{-4x^2} & \text{for } x > 0 \end{cases}$$

Solution

$$\int_{-\infty}^{\infty} f(x) dx = \int_0^{\infty} kxe^{-4x^2} dx = \int_0^{\infty} \frac{k}{8} \cdot e^{-u} du = \frac{k}{8} = 1$$

so that $k = 8$.



Continuous Random Variables

5.4 If the probability density of a random variable is given by

$$f(x) = \begin{cases} x & \text{for } 0 < x < 1 \\ 2 - x & \text{for } 1 \leq x < 2 \\ 0 & \text{elsewhere} \end{cases}$$

find the probabilities that a random variable having this probability density will take on a value

- (a) between 0.2 and 0.8; (b) between 0.6 and 1.2.

5.14 Find μ and σ^2 for the probability density of Exercise 5.4.

Continuous Random Variables

$$f(x) = \begin{cases} x & \text{for } 0 < x < 1 \\ 2 - x & \text{for } 1 \leq x < 2 \\ 0 & \text{elsewhere} \end{cases}$$

Find μ and σ^2 for the probability density

Continuous Random Variables

5.6 Given the probability density $f(x) = \frac{k}{1+x^2}$ for $-\infty < x < \infty$, find k.

Continuous Random Variables



- 5.10 The length of satisfactory service (years) provided by a certain model of laptop computer is a random variable having the probability density

$$f(x) = \begin{cases} \frac{1}{4.5} e^{-x/4.5} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

Find the probabilities that one of these laptops will provide satisfactory service for

- (a) at most 2.5 years; (b) anywhere from 4 to 6 years; (c) at least 6.75 years.

Continuous Random Variables

$$f(x) = \begin{cases} \frac{1}{4.5} e^{-x/4.5} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

- (a) at most 2.5 years; (b) anywhere from 4 to 6 years; (c) at least 6.75 years.

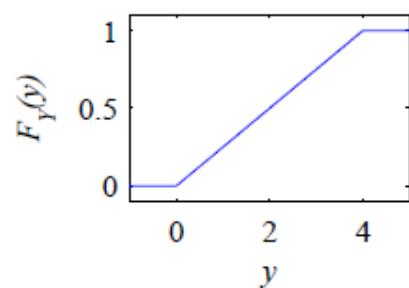
The cumulative distribution function of the random variable Y is

$$F_Y(y) = \begin{cases} 0 & y < 0, \\ y/4 & 0 \leq y \leq 4, \\ 1 & y > 4. \end{cases}$$

Sketch the CDF of Y and calculate the following probabilities:

- | | |
|-----------------------|-------------------|
| (1) $P[Y \leq -1]$ | (2) $P[Y \leq 1]$ |
| (3) $P[2 < Y \leq 3]$ | (4) $P[Y > 1.5]$ |

The CDF of Y is



From the CDF $F_Y(y)$, we can calculate the probabilities:

- (1) $P[Y \leq -1] = F_Y(-1) = 0$
- (2) $P[Y \leq 1] = F_Y(1) = 1/4$
- (3) $P[2 < Y \leq 3] = F_Y(3) - F_Y(2) = 3/4 - 2/4 = 1/4$
- (4) $P[Y > 1.5] = 1 - P[Y \leq 1.5] = 1 - F_Y(1.5) = 1 - (1.5)/4 = 5/8$

The probability density function of the random variable Y is

$$f_Y(y) = \begin{cases} 3y^2/2 & -1 \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Sketch the PDF and find the following:

- | | |
|----------------------------------|---------------------------------------|
| (1) the expected value $E[Y]$ | (2) the second moment $E[Y^2]$ |
| (3) the variance $\text{Var}[Y]$ | (4) the standard deviation σ_Y |

Recall - Continuous Random Variables

Properties:

$$f(x) \geq 0$$

$$\int_a^b f(x) dx = 1$$

$$P(a \leq X \leq b) = \int_a^b f(x) dx = F(b) - F(a)$$

$$f(x) = P(a \leq X \leq b) = \int_a^b f(x) dx$$

$$f(c) = P(X = c) = P(c \leq X \leq c) = \int_c^c f(x) dx = 0$$

$$E(x) = \int_a^b x f(x) dx$$

$$E(x^2) = \int_a^b x^2 f(x) dx$$

$$V(X) = E(X^2) - ((E(X))^2)$$

Exercise - Continuous Random Variables

- 5.7** If the distribution function of a random variable is given by

$$F(x) = \begin{cases} 1 - \frac{4}{x^2} & \text{for } x > 2 \\ 0 & \text{for } x \leq 2 \end{cases}$$

find the probabilities that this random variable will take on a value

- (a) less than 3; (b) between 4 and 5.

- 5.9** Let the phase error in a tracking device have probability density

$$f(x) = \begin{cases} \cos x & 0 < x < \pi/2 \\ 0 & \text{elsewhere} \end{cases}$$

Find the probability that the phase error is

- (a) between 0 and $\pi/4$; (b) greater than $\pi/3$.

Find μ and σ for the distribution of the phase error

- 4.3.1** The random variable X has probability density function

$$f_X(x) = \begin{cases} cx & 0 \leq x \leq 2, \\ 0 & \text{otherwise.} \end{cases}$$

Use the PDF to find

- (a) the constant c ,
- (b) $P[0 \leq X \leq 1]$,
- (c) $P[-1/2 \leq X \leq 1/2]$,
- (d) the CDF $F_X(x)$.

- 4.4.4** The probability density function of random variable Y is

$$f_Y(y) = \begin{cases} y/2 & 0 \leq y < 2, \\ 0 & \text{otherwise.} \end{cases}$$

What are $E[Y]$ and $\text{Var}[Y]$?

- 4.4.5** The cumulative distribution function of the random variable Y is

$$F_Y(y) = \begin{cases} 0 & y < -1, \\ (y+1)/2 & -1 \leq y \leq 1, \\ 1 & y > 1. \end{cases}$$

What are $E[Y]$ and $\text{Var}[Y]$?

5.1 Joint Probability distribution

Introduction:

- ❖ Let $X = \{x_1, x_2, \dots, x_m\}$ and $Y = \{y_1, y_2, \dots, y_n\}$ be two discrete random variables. Then $P(x, y) = J_{ij}$ is called joint probability function of X and Y if it satisfies the conditions:
 - (i) $J_{ij} \geq 0$
 - (ii) $\sum_{i=1}^m \sum_{j=1}^n J_{ij} = 1$
- ❖ Set of values of this joint probability function J_{ij} is called joint probability distribution of X and Y.

	y_1	y_2	\dots	y_n	<i>Sum</i>
x_1	J_{11}	J_{12}	\dots	J_{1n}	$f(x_1)$
x_2	J_{21}	J_{22}	\dots	J_{2n}	$f(x_2)$
\dots	\dots	\dots	\dots	\dots	\dots
x_m	J_{m1}	J_{m2}	\dots	J_{mn}	$f(x_m)$
<i>Sum</i>	$g(y_1)$	$g(y_2)$	\dots	$g(y_n)$	$Total = 1$

- So far we have been talking about the probability of a single variable, or a variable conditional on another.
- We often want to determine the joint probability of two variables, such as **X** and **Y**. Suppose we are able to determine the following information for education (X) and age (Y) for all Indian citizens based on the census.

Age (Y):		Age : 25-35	Age: 35-55	Age: 55-85
Education (X)		30	45	70
None	0	.01	.02	.05
Primary	1	.03	.06	.10
Secondary	2	.18	.21	.15
College	3	.07	.08	.04

Each cell is the relative frequency (f/N).

We can define the joint probability distribution as:
 $p(x, y) = \Pr(X = x \text{ and } Y = y)$

Example: what is the probability of getting a 30 year old college graduate?

$$p(x,y) = \Pr(X=3 \text{ and } Y=30) = .07$$

We can see that: $p(x) = \sum_y p(x,y)$
 $p(x=1) = .03 + .06 + .10 = .19$

Education (X)	Age (Y): 25-35	Age : 30	Age: 45	Age: 70
None	0	.01	.02	.05
Primary	1	.03	.06	.10
Secondary	2	.18	.21	.15
College	3	.07	.08	.04

Marginal Probability

- We call this the **marginal probability** because it is calculated by summing across rows or columns and is thus reported in the margins of the table.

We can calculate this for our entire table.

Age (Y):\nEducation (X)	30	45	70	$p(x)$
None: 0	.01	.02	.05	.08
Primary: 1	.03	.06	.10	.19
Secondary: 2	.18	.21	.15	.54
College: 3	.07	.08	.04	.19
$p(y)$.29	.37	.34	1

If X and Y are discrete random variables, the joint probability distribution of X and Y is a description of the set of points (x,y) in the range of (X,Y) along with the probability of each point.

The joint probability distribution of two discrete random variables is sometimes referred to as the **bivariate probability distribution** or **bivariate distribution**.

Thus we can describe the joint probability distribution of two discrete random variables is through a **joint probability mass function**

$$f(x,y) = P(X=x, Y=y)$$

Joint Probability Mass Function



- : The function $f(x, y)$ is a joint probability distribution or probability mass function of the discrete random variables X and Y if
1. $f(x, y) \geq 0$ for all (x, y) ,
 2. $\sum_x \sum_y f(x, y) = 1$,
 3. $P(X = x, Y = y) = f(x, y)$.

For any region A in the xy plane, $P[(X, Y) \in A] = \sum_A \sum_A f(x, y)$.

Joint Density Function

When X and Y are continuous random variables, the **joint density function** $f(x, y)$ is a surface lying above the xy plane, and $P[(X, Y) \in A]$, where A is any region in the xy plane, is equal to the volume of the right cylinder bounded by the base A and the surface.

The function $f(x, y)$ is a **joint density function** of the continuous random variables X and Y if

1. $f(x, y) \geq 0$, for all (x, y) ,
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$,
3. $P[(X, Y) \in A] = \int \int_A f(x, y) dx dy$, for any region A in the xy plane.

Marginal Distributions

The marginal distributions of X alone and of Y alone are

$$g(x) = \sum_y f(x, y) \quad \text{and} \quad h(y) = \sum_x f(x, y)$$

for the discrete case, and

$$g(x) = \int_{-\infty}^{\infty} f(x, y) dy \quad \text{and} \quad h(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

for the continuous case.

Consider the joint distribution of X and Y .

Compute the following probabilities:

- (i) $P(X = 1, Y = 2)$ (ii) $P(X \geq 1, Y \geq 2)$
- (iii) $P(X \leq 1, Y \leq 2)$ (iv) $P(X + Y \geq 2)$ (v) $P(X \geq 1, Y \leq 2)$.

Solution:

$X \backslash Y$	0	1	2	3
0	0	$1/8$	$1/4$	$1/8$
1	$1/8$	$1/4$	$1/8$	0

(i) $X = \{0, 1\}, Y = \{0, 1, 2, 3, 4\}$

$$P(X = 1, Y = 2) = P(1, 2) = \frac{1}{8}$$

(ii) If $X \geq 1, X = \{1\}$. If $Y \geq 2, Y = \{2, 3\}$

$$P(X \geq 1, Y \geq 2) = P(1, 2) + P(1, 3) = \frac{1}{8} + 0 = \frac{1}{8}$$

(iii) If $X \leq 1, X = \{0, 1\}$. If $Y \leq 2, Y = \{0, 1, 2\}$

$$P(X \leq 1, Y \leq 2) = P(0, 0) + P(0, 1) + P(0, 2) + P(1, 0) + P(1, 1) + P(1, 2)$$

$$= 0 + \frac{1}{4} + \frac{1}{8} + \frac{1}{8} + \frac{1}{4} + \frac{1}{8} = \frac{7}{8}$$

Cont.

(iv) If $X + Y \geq 2$ then

$$X + Y = 0 + 2 \text{ or } 0 + 3 \text{ or } 1 + 1 \text{ or } 1 + 2 \text{ or } 1 + 3$$

$$\begin{aligned}P(X + Y \geq 2) &= P(0, 2) + P(0, 3) + P(1, 1) + P(1, 2) + P(1, 3) \\&= \frac{1}{4} + \frac{1}{8} + \frac{1}{4} + \frac{1}{8} + 0 = \frac{3}{4}\end{aligned}$$

(v) If $X \geq 1, X = \{1\}$. If $Y \leq 2, Y = \{0, 1, 2\}$

$$\begin{aligned}P(X \geq 1, Y \leq 2) &= P(1, 0) + P(1, 1) + P(1, 2) \\&= \frac{1}{8} + \frac{1}{4} + \frac{1}{8} = \frac{1}{2}\end{aligned}$$

Problem:

- Two ballpoint pens are selected at random from a box that contains blue pens, 2 red pens and 3 green pens. If X is the number of blue pens selected and Y is the number of red pens selected, find the joint probability function $f(x,y)$

- **Solution:**

The possible pairs of values (x,y) are $(0,0), (0,1), (1,0), (1,1), (0,2), (2,0)$

The joint probability distribution can be represented by the formula

$$f(x,y) = \frac{\binom{3}{x} \binom{2}{y} \binom{3}{2-x-y}}{\binom{8}{2}},$$

for $x = 0, 1, 2; y = 0, 1, 2;$ and $0 \leq x + y \leq 2.$

Joint distribution

$f(x,y)$		X			Rows Total
		0	1	2	
Y	0	3/28	9/28	3/28	15/28
	1	3/14	3/14	0	3/7
	2	1/28	0	0	1/28
Columns Total		5/14	15/28	3/28	1

3. Find the joint distribution of X and Y which are the independent random variables with the following respective distributions.

x_i	1	2
$f(x_i)$	0.7	0.3

y_j	-2	5	8
$g(y_j)$	0.3	0.5	0.2

Solution:

Since X and Y are independent random variables,

$$J_{ij} = f(x_i)g(y_j)$$

Therefore,

$x \setminus y$	-2	5	8	$f(x)$
1	0.21	0.35	0.14	0.7
2	0.09	0.15	0.06	0.3
$g(y)$	0.3	0.5	0.2	Total = 1

-
6. The joint probability distribution of two discrete random variables X and Y is given by $f(x, y) = k(2x + y)$ for $0 \leq x \leq 2$, $0 \leq y \leq 3$. (i) Find the value of k . (ii) The marginal distribution of X and Y (iii) Show that X and Y are dependent.
-

Q.9 A candy company distributed boxes of chocolates with a mixture of creams, toffees, and nuts coated in both light and dark chocolate. For a randomly selected box, let X and Y , respectively, be the proportions of the light and dark chocolates that are creams and suppose that the joint density function is

$$f(x, y) = \begin{cases} \frac{2}{5}(2x + 3y), & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0, & \text{elsewhere} \end{cases}$$

- a) Verify whether
- b) Find $P[(X, Y) \in A]$, where A is the region $\{(x, y) \mid 0 < x < \frac{1}{2}, \frac{1}{4} < y < \frac{1}{2}\}$
- c) Find $g(x)$ and $h(y)$ for the joint density function.

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

Example 9 – Solution

a)

$$\begin{aligned}
 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy &= \int_0^1 \int_0^1 \frac{2}{5} (2x + 3y) dx dy \\
 &= \int_0^1 \left[\frac{2x^2}{5} + \frac{6xy}{5} \right]_{x=0}^{x=1} dy \\
 &= \int_0^1 \left(\frac{2}{5} + \frac{6y}{5} \right) dy = \left[\frac{2y}{5} + \frac{3y^2}{5} \right]_0^1 \\
 &= \frac{2}{5} + \frac{3}{5} = 1
 \end{aligned}$$

Example – Solution

b)

$$\begin{aligned}
 P[(X, Y) \in A] &= P(0 < X < \frac{1}{2}, \frac{1}{4} < Y < \frac{1}{2}) \\
 &= \int_{\frac{1}{4}}^{\frac{1}{2}} \int_0^{\frac{1}{2}} \frac{2}{5} (2x + 3y) dx dy \\
 &= \int_{\frac{1}{4}}^{\frac{1}{2}} \left[\frac{2x^2}{5} + \frac{6xy}{5} \right]_{x=0}^{x=\frac{1}{2}} dy \\
 &= \int_{\frac{1}{4}}^{\frac{1}{2}} \left(\frac{1}{10} + \frac{3y}{5} \right) dy = \left[\frac{y}{10} + \frac{3y^2}{10} \right]_{\frac{1}{4}}^{\frac{1}{2}} \\
 &= \frac{1}{10} \left[\left(\frac{1}{2} + \frac{3}{4} \right) - \left(\frac{1}{4} + \frac{3}{16} \right) \right] = \frac{13}{160}
 \end{aligned}$$

Example – Solution



By definition,

$$g(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_0^1 \frac{2}{5} (2x + 3y) dy = \left. \frac{4xy}{5} + \frac{6y^2}{10} \right|_{y=0}^{y=1} = \frac{4x + 3}{5}$$

For $0 \leq x \leq 1$, and $g(x)=0$ elsewhere.

Similarly,
$$h(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_0^1 \frac{2}{5} (2x + 3y) dx = \frac{4(1 + 3y)}{5}$$

For $0 \leq y \leq 1$, and $h(y)=0$ elsewhere.

Problem:

Find 'K' if the joint probability density function of a bivariate random variable (X,Y) is given by

$$f(x, y) = \begin{cases} K(1-x)(1-y) & \text{if } 0 < x < 4 \\ 0 & \text{otherwise} \end{cases}$$

7. The joint probability distribution of X and Y is given by $f(x, y) = c(x^2 + y^2)$ for $x = -1, 0, 1, 3$ and $y = -1, 2, 3$. (i) Find the value of c . (ii) $P(x = 0, y \leq 2)$ (iii) $P(x \leq 1, y > 2)$ (iv) $P(x \geq 2 - y)$

Solution:

By data, $X = \{-1, 0, 1, 3\}$ and $Y = \{-1, 2, 3\}$

$$f(x, y) = c(x^2 + y^2)$$

The joint probability distribution of X and Y :

X \ Y	-1	2	3	$f(X)$
-1	$2c$	$5c$	$10c$	$17c$
0	c	$4c$	$9c$	$14c$
1	$2c$	$5c$	$10c$	$17c$
3	$10c$	$13c$	$18c$	$41c$
$g(Y)$	$15c$	$27c$	$47c$	$89c$

- (i) **Find c :** $1 = \sum f(x, y) = 89c$

$$c = \frac{1}{89}$$

$$(ii) \quad x = 0, y = \{-1, 2\}$$

$$\begin{aligned} P(x = 0, y \leq 2) \\ &= P(0, -1) + P(0, 2) \\ &= c + 4c = 5c \\ &= 5/89 \end{aligned}$$

$$(iii) \quad x = \{-1, 0, 1\}, y = \{3\}$$

$$\begin{aligned} P(x \leq 1, y > 2) \\ &= P(-1, 3) + P(0, 3) + P(1, 3) \\ &= 10c + 9c + 10c \\ &= 29c = 29/89 \end{aligned}$$

Cont.

(iv) By data, $X = \{-1, 0, 1, 3\}$ and $Y = \{-1, 2, 3\}$

$$\begin{aligned} P(x \geq 2 - y) &= P(x + y \geq 2) \\ &= P(-1, 3) + P(0, 2) + P(0, 3) + P(1, 2) + P(1, 3) + P(3, -1) + P(3, 2) + P(3, 3) \\ &= 10c + 4c + 9c + 5c + 10c + 10c + 13c + 18c \\ &= 79c = 79/89 \end{aligned}$$



Thank You !



BITS Pilani
Pilani Campus

M.Tech. (AIML)

Session-5 (Random Variables)

Team ISM

Session-5 Agenda

Random variables –

Discrete & continuous Expectation of a random variable,
mean and variance of a random variable –

Single random random variable &

Joint distributions

 Contact Session 5: Module 3: Probability Distributions

Contact Session	List of Topic Title	Reference
CS - 5	Random variables - Discrete & continuous Expectation of a random variable, mean and variance of a random variable – Single random random variable & Joint distributions	T1 & T2
HW	Problems on random variables	T1 & T2
Lab	Probability Distributions & Sampling	Lab 3

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.
 - In mathematical language, a random variable is a function whose domain is the sample space and whose range is the set of real numbers.
-

Random Variables

- 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 finite or countably infinite number of values
- A continuous random variable may assume any numerical value in an interval or collection of intervals.
- Continuous random variables are generated in experiments where things are “measured” as opposed to “counted”.
- Experimental outcomes based on measurement of time, distance, weight, volume etc. generate continuous RV.

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

Two 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

Discrete Probability Distributions

- The probability distribution for a random variable describes how probabilities are distributed over the values of the random variable.
- The probability distribution is defined by a probability 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:

$$f(x) \geq 0$$

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

- We can describe a discrete probability distribution with a **table, graph, or equation**.
 - Advantage: once the probability distribution is known, it is relatively easy to determine the probability of a variety of events that may be of interest to the decision maker.
-

Probability Distributions for Discrete Random Variables

- Say a random variable x follows this pattern: $p(x) = (.3)(.7)^{x-1}$ for $x > 0$.
 - This table gives the probabilities (rounded to two digits) for x between 1 and 10.

x	$P(x)$
1	.30
2	.21
3	.15
4	.11
5	.07
6	.05
7	.04
8	.02
9	.02
10	.01

Expected Value and Variance

- The expected value, or mean, of a random variable is a measure of its central location.
 - The mean or Expected value of a discrete random variable:

$$E(x) = \mu = \sum xf(x)$$

- The variance summarizes the variability in the values of a random variable.
 - Variance of a discrete random variable:

$$\text{Var}(x) = \sigma^2 = \sum (x - \mu)^2 f(x)$$

- The standard deviation, σ , is defined as the positive square root of the variance.
-

Rules of Expected Value

- Multiplying RV by a constant a , $E(aX) = a.E(X)$
 - Adding a constant b , $E(X+b) = E(X) + b$
 - Therefore, $E(aX + b) = ?$
-

Variability 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

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

Rules of variability

- Multiplying RV by a constant a , $V(aX) = a^2 \cdot V(X)$
 - Adding a constant b , $V(X+b) = V(X)$
 - $\sigma_{aX} = |a| \cdot \sigma_X, \quad \sigma_{X+b} = \sigma_X$
-

Example:

At a shooting range, a shooter is able to hit a target in either 1, 2 or 3 shots. Let x be a random variable indicating the number of shots fired to hit the target. The following probability function was proposed.

$$f(x) = x/6$$

Is this probability function valid?

Identify the r.v to be discrete or continuous?

EXPERIMENT	Random Variable (x)
Audit 50 tax returns	Number of returns that contains error
Operate a restaurant for one day	Number of customers
Observe an employee's work	No. of productive hours in an 8-hour workday

Example: JSL Appliances

- Discrete random variable with a finite number of values
 - Let x = number of TV sets sold at the store in one day
where x can take on 5 values (0, 1, 2, 3, 4)

 - Discrete random variable with an infinite sequence of values
 - Let x = number of customers arriving in one day
where x can take on the values 0, 1, 2, ...
 - We can count the customers arriving, but there is no finite upper limit on the number that might arrive.
-

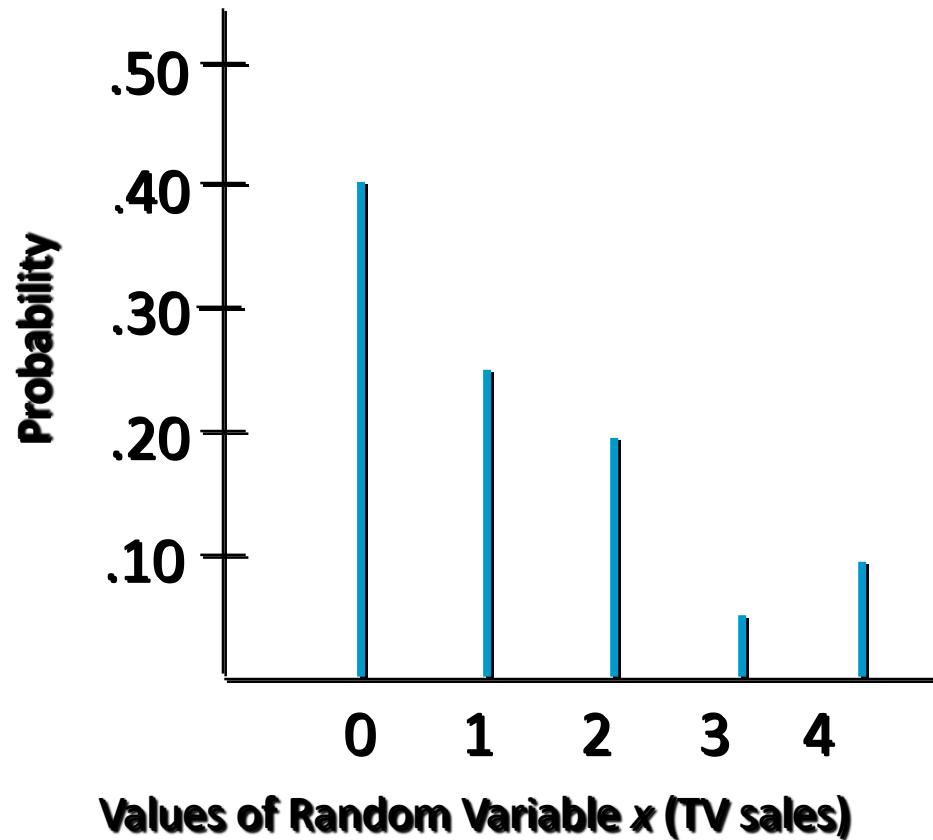
Example : JSL Appliances

Using past data on TV sales (below left), a tabular representation of the probability distribution for TV sales (below right) was developed.

Units Sold	No of days	x	f(x)
0	80	0	0.4
1	50	1	0.25
2	40	2	0.2
3	10	3	0.05
4	20	4	0.1
Total	200	1	

Example: JSL Appliances

- Graphical Representation of the Probability Distribution



Example: JSL Appliances

Expected Value of a Discrete Random Variable

x	f(x)	xf(x)
0	0.4	0.00
1	0.25	0.25
2	0.2	0.40
3	0.05	0.15
4	0.1	0.40

$$E(x) = 1.20$$

The expected number of TV sets sold in a day is 1.2

Example: JSL Appliances

- Variance and Standard Deviation of a Discrete Random Variable

x	$x - \mu$	$(x - \mu)^2$	$f(x)$	$(x - \mu)^2 f(x)$
0	-1.2	1.44	.40	.576
1	-0.2	0.04	.25	.010
2	0.8	0.64	.20	.128
3	1.8	3.24	.05	.162
4	2.8	7.84	.10	<u>.784</u>
				1.660 = σ^2

- The variance of daily sales is 1.66 TV sets squared.
- The standard deviation of sales is 1.2884 TV sets.

Discrete Uniform Probability Distribution

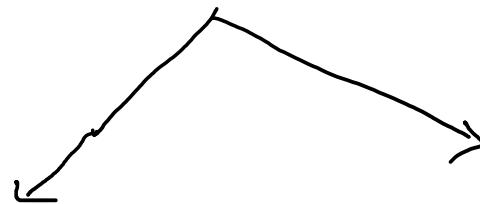
- The discrete uniform probability distribution is the simplest example of a discrete probability distribution given by a formula.
- The discrete uniform probability function is

$$f(x) = 1/n$$

where:

n = the number of values the random variable may assume
Note that the values of the random variable are equally likely.

Random Variables



Discrete

$$X = \{1, 2, 3, 4\}$$

→ No of students

→ No of bits
transmitted

continuous

$$X \in (1, 3)$$

- ↓
→ pressure
→ Temperature
→ Time
→ Voltage

Quiz - 1 ✓ (3) → 5 min } f(10)

one attempt
Best - 2
10 & 60 minutes

Random Variables

Discrete

$$P(x)$$

continuous

$$f(x)$$

Validation :-

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

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

$= \sum (x - \mu)^2 p(x)$ Probability distribution function

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

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

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

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

probability density function

$$\text{mean} = E(x) = \int x f(x) dx$$

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

Mean
variance

Example: *Try*

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

compute the following

- mean* a) $E(x)$
- Variance* b) $V(x)$ directly from the definition
- s.d.* c) standard deviation of X
- d) $V(x)$ using the shortcut formula
($E(x^2)$ - $(E(x))^2$)

Example

$$\int x^n dx = \frac{x^{n+1}}{n+1}$$

$$\int u v du = u \int v - \int u v'$$

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 .

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

$$\int_{-1}^1 Cx^2 dx = 1$$

b) Find $E(x)$ and $V(x)$

$$C \left[\frac{x^3}{3} \right]_{-1}^1 = 1$$

c) Find $P(x \geq y_2)$

$$E(x) = \int_{-1}^1 x Cx^2 dx = C \int_{-1}^1 x^3 dx$$

$$= \frac{3}{2} \left[\frac{x^4}{4} \right]_{-1}^1 = 0$$

$$\frac{C}{3} [1 - (-1)] = 1$$

$$\frac{2}{3} C = 1$$

$$C = \frac{3}{2}$$

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

↓
mean

$$\int x^2 f(x) dx = \int_{-1}^1 x^2 c x^2 dx = \frac{3}{2} \left[\frac{x^5}{5} \right]_{-1}^1$$

$$= \frac{3}{10} \{ 1 - (-1) \} - \frac{3}{10} \times 2 = \frac{3}{5}$$

$$= \frac{3}{5} - 0^2 \Rightarrow \frac{3}{5} \quad \checkmark$$

$$P(X \geq Y_2) = \int_{Y_2}^1 f(x) dx = \int_{Y_2}^1 c x^2 dx = \frac{3}{2} \left[\frac{x^3}{3} \right]_{Y_2}^1$$

$$= \frac{3}{6} [1^3 - (Y_2)^3]$$

=

$$\frac{d}{dx} P(x) = f(x) \quad \checkmark$$

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



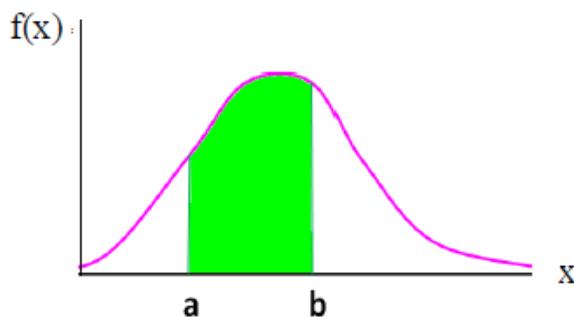
Continuous Probability Distributions

- A continuous random variable can assume any value in an interval on the real line or in a collection of intervals.
 - It is not possible to talk about the probability of the random variable assuming a particular value.
 - Instead, we talk about the probability of the random variable assuming a value within a given interval.
 - The probability of the random variable assuming a value within some given interval from x_1 to x_2 is defined to be the area under the graph of the probability density function between x_1 and x_2 .
-

Continuous Random Variables

A continuous random variable can assume any value in an interval on the real line or in a collection of intervals.

The probability of the random variable assuming a value within some given interval from x_1 to x_2 is defined to be the area under the graph of the **probability density function** between x_1 and x_2



Example:

- Height of students in a class
- Amount of ice tea in a glass
- Change in temperature throughout a day
- Price of a car in next year

Continuous Random Variables

Probability Density Function

For a continuous random variable X , a **probability density function** is a function such that

$$(1) \quad f(x) \geq 0$$

$$(2) \quad \int_{-\infty}^{\infty} f(x) dx = 1$$

$$(3) \quad P(a \leq X \leq b) = \int_a^b f(x) dx = \text{area under } f(x) \text{ from } a \text{ to } b \text{ for any } a \text{ and } b \quad (4.1)$$

Continuous Random Variables

Cumulative Distribution Function

The **cumulative distribution function** of a continuous random variable X is

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(u) du$$

for $-\infty < x < \infty$.

Probability Density Function from the Cumulative Distribution Function

Given $F(x)$,

$$f(x) = \frac{dF(x)}{dx}$$

as long as the derivative exists.

Continuous Random Variables

Mean and Variance

Suppose that X is a continuous random variable with probability density function $f(x)$. The **mean or expected value** of X , denoted as μ or $E(X)$, is

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

The **variance** of X , denoted as $V(X)$ or σ^2 , is

$$\sigma^2 = V(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2$$

The **standard deviation** of X is $\sigma = \sqrt{\sigma^2}$.

Integration Formulas

$$\int kf(u)du = k \int f(u)du$$

$$\int u^n du = \frac{u^{n+1}}{n+1}$$

$$\int e^u du = e^u$$

$$\int \sin u du = -\cos u$$

$$\int \cos u du = \sin u$$

$$\int [f(u) \pm g(u)] du = \int f(u)du \pm \int g(u)du$$

$$\int udv = uv - \int vdu$$

Continuous Random Variables

EXAMPLE I

Calculating probabilities from the probability density function

If a random variable has the probability density

$$f(x) = \begin{cases} 2e^{-2x} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

find the probabilities that it will take on a value

- between 1 and 3;
- greater than 0.5.

Solution Evaluating the necessary integrals, we get

$$(a) \quad \int_1^3 2e^{-2x} dx = e^{-2} - e^{-6} = 0.133$$

$$(b) \quad \int_{0.5}^{\infty} 2e^{-2x} dx = e^{-1} = 0.368$$



With reference to the preceding example, find the distribution function and use it to determine the probability that the random variable will take on a value less than or equal to 1.

Performing the necessary integrations, we get

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \int_0^x 2e^{-2t} dt = 1 - e^{-2x} & \text{for } x > 0 \end{cases}$$

and substitution of $x = 1$ yields

$$F(1) = 1 - e^{-2} = 0.865$$



Determining the mean and variance using the probability density function

With reference to Example 1, find the mean and the variance of the given probability density.

Performing the necessary integrations, using integrations by parts, we get

$$\mu = \int_{-\infty}^{\infty} xf(x) dx = \int_0^{\infty} x \cdot 2 e^{-2x} dx = \frac{1}{2}$$

Alternatively, the expectation of x is $E(X) = 0.5$

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx = \int_0^{\infty} \left(x - \frac{1}{2}\right)^2 \cdot 2 e^{-2x} dx = \frac{1}{4}$$



A probability density function assigns probability one to $(-\infty, \infty)$

Find k so that the following can serve as the probability density of a random variable:

$$f(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ kxe^{-4x^2} & \text{for } x > 0 \end{cases}$$

Solution

$$\int_{-\infty}^{\infty} f(x) dx = \int_0^{\infty} kxe^{-4x^2} dx = \int_0^{\infty} \frac{k}{8} \cdot e^{-u} du = \frac{k}{8} = 1$$

so that $k = 8$.



Continuous Random Variables

5.4 If the probability density of a random variable is given by

$$f(x) = \begin{cases} x & \text{for } 0 < x < 1 \\ 2 - x & \text{for } 1 \leq x < 2 \\ 0 & \text{elsewhere} \end{cases}$$

find the probabilities that a random variable having this probability density will take on a value

- (a) between 0.2 and 0.8; (b) between 0.6 and 1.2.

5.14 Find μ and σ^2 for the probability density of Exercise 5.4.

Continuous Random Variables

$$f(x) = \begin{cases} x & \text{for } 0 < x < 1 \\ 2 - x & \text{for } 1 \leq x < 2 \\ 0 & \text{elsewhere} \end{cases}$$

Find μ and σ^2 for the probability density

Continuous Random Variables

5.6 Given the probability density $f(x) = \frac{k}{1+x^2}$ for $-\infty < x < \infty$, find k.

Continuous Random Variables



- 5.10 The length of satisfactory service (years) provided by a certain model of laptop computer is a random variable having the probability density

$$f(x) = \begin{cases} \frac{1}{4.5} e^{-x/4.5} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

Find the probabilities that one of these laptops will provide satisfactory service for

- (a) at most 2.5 years; (b) anywhere from 4 to 6 years; (c) at least 6.75 years.

Continuous Random Variables

$$f(x) = \begin{cases} \frac{1}{4.5} e^{-x/4.5} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

- (a) at most 2.5 years; (b) anywhere from 4 to 6 years; (c) at least 6.75 years.

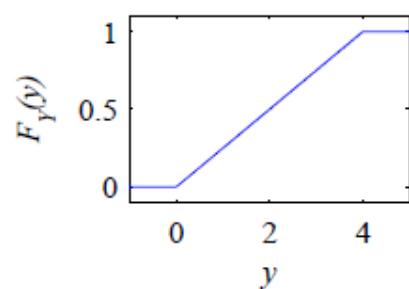
The cumulative distribution function of the random variable Y is

$$F_Y(y) = \begin{cases} 0 & y < 0, \\ y/4 & 0 \leq y \leq 4, \\ 1 & y > 4. \end{cases}$$

Sketch the CDF of Y and calculate the following probabilities:

- | | |
|-----------------------|-------------------|
| (1) $P[Y \leq -1]$ | (2) $P[Y \leq 1]$ |
| (3) $P[2 < Y \leq 3]$ | (4) $P[Y > 1.5]$ |

The CDF of Y is



From the CDF $F_Y(y)$, we can calculate the probabilities:

- (1) $P[Y \leq -1] = F_Y(-1) = 0$
- (2) $P[Y \leq 1] = F_Y(1) = 1/4$
- (3) $P[2 < Y \leq 3] = F_Y(3) - F_Y(2) = 3/4 - 2/4 = 1/4$
- (4) $P[Y > 1.5] = 1 - P[Y \leq 1.5] = 1 - F_Y(1.5) = 1 - (1.5)/4 = 5/8$

The probability density function of the random variable Y is

$$f_Y(y) = \begin{cases} 3y^2/2 & -1 \leq y \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Sketch the PDF and find the following:

- | | |
|----------------------------------|---------------------------------------|
| (1) the expected value $E[Y]$ | (2) the second moment $E[Y^2]$ |
| (3) the variance $\text{Var}[Y]$ | (4) the standard deviation σ_Y |

Recall - Continuous Random Variables

Properties:

$$f(x) \geq 0$$

$$\int_a^b f(x) dx = 1$$

$$P(a \leq X \leq b) = \int_a^b f(x) dx = F(b) - F(a)$$

$$f(x) = P(a \leq X \leq b) = \int_a^b f(x) dx$$

$$f(c) = P(X = c) = P(c \leq X \leq c) = \int_c^c f(x) dx = 0$$

$$E(x) = \int_a^b x f(x) dx$$

$$E(x^2) = \int_a^b x^2 f(x) dx$$

$$V(X) = E(X^2) - ((E(X))^2)$$

Exercise - Continuous Random Variables

- 5.7** If the distribution function of a random variable is given by

$$F(x) = \begin{cases} 1 - \frac{4}{x^2} & \text{for } x > 2 \\ 0 & \text{for } x \leq 2 \end{cases}$$

find the probabilities that this random variable will take on a value

- (a) less than 3; (b) between 4 and 5.

- 5.9** Let the phase error in a tracking device have probability density

$$f(x) = \begin{cases} \cos x & 0 < x < \pi/2 \\ 0 & \text{elsewhere} \end{cases}$$

Find the probability that the phase error is

- (a) between 0 and $\pi/4$; (b) greater than $\pi/3$.

Find μ and σ for the distribution of the phase error

- 4.3.1** The random variable X has probability density function

$$f_X(x) = \begin{cases} cx & 0 \leq x \leq 2, \\ 0 & \text{otherwise.} \end{cases}$$

Use the PDF to find

- (a) the constant c ,
- (b) $P[0 \leq X \leq 1]$,
- (c) $P[-1/2 \leq X \leq 1/2]$,
- (d) the CDF $F_X(x)$.

- 4.4.4** The probability density function of random variable Y is

$$f_Y(y) = \begin{cases} y/2 & 0 \leq y < 2, \\ 0 & \text{otherwise.} \end{cases}$$

What are $E[Y]$ and $\text{Var}[Y]$?

- 4.4.5** The cumulative distribution function of the random variable Y is

$$F_Y(y) = \begin{cases} 0 & y < -1, \\ (y+1)/2 & -1 \leq y \leq 1, \\ 1 & y > 1. \end{cases}$$

What are $E[Y]$ and $\text{Var}[Y]$?

5.1 Joint Probability distribution

Introduction:

- ❖ Let $X = \{x_1, x_2, \dots, x_m\}$ and $Y = \{y_1, y_2, \dots, y_n\}$ be two discrete random variables. Then $P(x, y) = J_{ij}$ is called joint probability function of X and Y if it satisfies the conditions:
 - (i) $J_{ij} \geq 0$
 - (ii) $\sum_{i=1}^m \sum_{j=1}^n J_{ij} = 1$
- ❖ Set of values of this joint probability function J_{ij} is called joint probability distribution of X and Y.

	y_1	y_2	\dots	y_n	<i>Sum</i>
x_1	J_{11}	J_{12}	\dots	J_{1n}	$f(x_1)$
x_2	J_{21}	J_{22}	\dots	J_{2n}	$f(x_2)$
\dots	\dots	\dots	\dots	\dots	\dots
x_m	J_{m1}	J_{m2}	\dots	J_{mn}	$f(x_m)$
<i>Sum</i>	$g(y_1)$	$g(y_2)$	\dots	$g(y_n)$	$Total = 1$

- So far we have been talking about the probability of a single variable, or a variable conditional on another.
- We often want to determine the joint probability of two variables, such as **X** and **Y**. Suppose we are able to determine the following information for education (X) and age (Y) for all Indian citizens based on the census.

Age (Y):		Age : 25-35	Age: 35-55	Age: 55-85
Education (X)		30	45	70
None	0	.01	.02	.05
Primary	1	.03	.06	.10
Secondary	2	.18	.21	.15
College	3	.07	.08	.04

Each cell is the relative frequency (f/N).

We can define the joint probability distribution as:
 $p(x, y) = \Pr(X = x \text{ and } Y = y)$

Example: what is the probability of getting a 30 year old college graduate?

$$p(x,y) = \Pr(X=3 \text{ and } Y=30) = .07$$

We can see that: $p(x) = \sum_y p(x,y)$
 $p(x=1) = .03 + .06 + .10 = .19$

Education (X)	Age (Y): 25-35	Age : 30	Age: 45	Age: 70
None	0	.01	.02	.05
Primary	1	.03	.06	.10
Secondary	2	.18	.21	.15
College	3	.07	.08	.04

Marginal Probability

- We call this the **marginal probability** because it is calculated by summing across rows or columns and is thus reported in the margins of the table.

We can calculate this for our entire table.

Age (Y):\nEducation (X)	30	45	70	$p(x)$
None: 0	.01	.02	.05	.08
Primary: 1	.03	.06	.10	.19
Secondary: 2	.18	.21	.15	.54
College: 3	.07	.08	.04	.19
$p(y)$.29	.37	.34	1

If X and Y are discrete random variables, the joint probability distribution of X and Y is a description of the set of points (x,y) in the range of (X,Y) along with the probability of each point.

The joint probability distribution of two discrete random variables is sometimes referred to as the **bivariate probability distribution** or **bivariate distribution**.

Thus we can describe the joint probability distribution of two discrete random variables is through a **joint probability mass function**

$$f(x,y) = P(X=x, Y=y)$$

Joint Probability Mass Function



- : The function $f(x, y)$ is a joint probability distribution or probability mass function of the discrete random variables X and Y if
1. $f(x, y) \geq 0$ for all (x, y) ,
 2. $\sum_x \sum_y f(x, y) = 1$,
 3. $P(X = x, Y = y) = f(x, y)$.

For any region A in the xy plane, $P[(X, Y) \in A] = \sum_A \sum_A f(x, y)$.

Joint Density Function

When X and Y are continuous random variables, the **joint density function** $f(x, y)$ is a surface lying above the xy plane, and $P[(X, Y) \in A]$, where A is any region in the xy plane, is equal to the volume of the right cylinder bounded by the base A and the surface.

The function $f(x, y)$ is a **joint density function** of the continuous random variables X and Y if

1. $f(x, y) \geq 0$, for all (x, y) ,
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy = 1$,
3. $P[(X, Y) \in A] = \int \int_A f(x, y) dx dy$, for any region A in the xy plane.

Marginal Distributions

The marginal distributions of X alone and of Y alone are

$$g(x) = \sum_y f(x, y) \quad \text{and} \quad h(y) = \sum_x f(x, y)$$

for the discrete case, and

$$g(x) = \int_{-\infty}^{\infty} f(x, y) dy \quad \text{and} \quad h(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

for the continuous case.

Consider the joint distribution of X and Y .

Compute the following probabilities:

- (i) $P(X = 1, Y = 2)$ (ii) $P(X \geq 1, Y \geq 2)$
- (iii) $P(X \leq 1, Y \leq 2)$ (iv) $P(X + Y \geq 2)$ (v) $P(X \geq 1, Y \leq 2)$.

Solution:

$X \backslash Y$	0	1	2	3
0	0	$1/8$	$1/4$	$1/8$
1	$1/8$	$1/4$	$1/8$	0

(i) $X = \{0, 1\}, Y = \{0, 1, 2, 3, 4\}$

$$P(X = 1, Y = 2) = P(1, 2) = \frac{1}{8}$$

(ii) If $X \geq 1, X = \{1\}$. If $Y \geq 2, Y = \{2, 3\}$

$$P(X \geq 1, Y \geq 2) = P(1, 2) + P(1, 3) = \frac{1}{8} + 0 = \frac{1}{8}$$

(iii) If $X \leq 1, X = \{0, 1\}$. If $Y \leq 2, Y = \{0, 1, 2\}$

$$P(X \leq 1, Y \leq 2) = P(0, 0) + P(0, 1) + P(0, 2) + P(1, 0) + P(1, 1) + P(1, 2)$$

$$= 0 + \frac{1}{4} + \frac{1}{8} + \frac{1}{8} + \frac{1}{4} + \frac{1}{8} = \frac{7}{8}$$

Cont.

(iv) If $X + Y \geq 2$ then

$$X + Y = 0 + 2 \text{ or } 0 + 3 \text{ or } 1 + 1 \text{ or } 1 + 2 \text{ or } 1 + 3$$

$$\begin{aligned}P(X + Y \geq 2) &= P(0, 2) + P(0, 3) + P(1, 1) + P(1, 2) + P(1, 3) \\&= \frac{1}{4} + \frac{1}{8} + \frac{1}{4} + \frac{1}{8} + 0 = \frac{3}{4}\end{aligned}$$

(v) If $X \geq 1, X = \{1\}$. If $Y \leq 2, Y = \{0, 1, 2\}$

$$\begin{aligned}P(X \geq 1, Y \leq 2) &= P(1, 0) + P(1, 1) + P(1, 2) \\&= \frac{1}{8} + \frac{1}{4} + \frac{1}{8} = \frac{1}{2}\end{aligned}$$

Problem:

- Two ballpoint pens are selected at random from a box that contains blue pens, 2 red pens and 3 green pens. If X is the number of blue pens selected and Y is the number of red pens selected, find the joint probability function $f(x,y)$

- **Solution:**

The possible pairs of values (x,y) are $(0,0), (0,1), (1,0), (1,1), (0,2), (2,0)$

The joint probability distribution can be represented by the formula

$$f(x,y) = \frac{\binom{3}{x} \binom{2}{y} \binom{3}{2-x-y}}{\binom{8}{2}},$$

for $x = 0, 1, 2; y = 0, 1, 2;$ and $0 \leq x + y \leq 2.$

Joint distribution

$f(x,y)$		X			Rows Total
		0	1	2	
Y	0	3/28	9/28	3/28	15/28
	1	3/14	3/14	0	3/7
	2	1/28	0	0	1/28
Columns Total		5/14	15/28	3/28	1

3. Find the joint distribution of X and Y which are the independent random variables with the following respective distributions.

x_i	1	2
$f(x_i)$	0.7	0.3

y_j	-2	5	8
$g(y_j)$	0.3	0.5	0.2

Solution:

Since X and Y are independent random variables,

$$J_{ij} = f(x_i)g(y_j)$$

Therefore,

$x \setminus y$	-2	5	8	$f(x)$
1	0.21	0.35	0.14	0.7
2	0.09	0.15	0.06	0.3
$g(y)$	0.3	0.5	0.2	Total = 1

-
6. The joint probability distribution of two discrete random variables X and Y is given by $f(x, y) = k(2x + y)$ for $0 \leq x \leq 2$, $0 \leq y \leq 3$. (i) Find the value of k . (ii) The marginal distribution of X and Y (iii) Show that X and Y are dependent.
-

Q.9 A candy company distributed boxes of chocolates with a mixture of creams, toffees, and nuts coated in both light and dark chocolate. For a randomly selected box, let X and Y , respectively, be the proportions of the light and dark chocolates that are creams and suppose that the joint density function is

$$f(x, y) = \begin{cases} \frac{2}{5}(2x + 3y), & 0 \leq x \leq 1, 0 \leq y \leq 1 \\ 0, & \text{elsewhere} \end{cases}$$

- a) Verify whether
- b) Find $P[(X, Y) \in A]$, where A is the region $\{(x, y) \mid 0 < x < \frac{1}{2}, \frac{1}{4} < y < \frac{1}{2}\}$
- c) Find $g(x)$ and $h(y)$ for the joint density function.

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

Example 9 – Solution

a)

$$\begin{aligned}
 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx dy &= \int_0^1 \int_0^1 \frac{2}{5} (2x + 3y) dx dy \\
 &= \int_0^1 \left[\frac{2x^2}{5} + \frac{6xy}{5} \right]_{x=0}^{x=1} dy \\
 &= \int_0^1 \left(\frac{2}{5} + \frac{6y}{5} \right) dy = \left[\frac{2y}{5} + \frac{3y^2}{5} \right]_0^1 \\
 &= \frac{2}{5} + \frac{3}{5} = 1
 \end{aligned}$$

Example – Solution

b)

$$\begin{aligned}
 P[(X, Y) \in A] &= P(0 < X < \frac{1}{2}, \frac{1}{4} < Y < \frac{1}{2}) \\
 &= \int_{\frac{1}{4}}^{\frac{1}{2}} \int_0^{\frac{1}{2}} \frac{2}{5} (2x + 3y) dx dy \\
 &= \int_{\frac{1}{4}}^{\frac{1}{2}} \left[\frac{2x^2}{5} + \frac{6xy}{5} \right]_{x=0}^{x=\frac{1}{2}} dy \\
 &= \int_{\frac{1}{4}}^{\frac{1}{2}} \left(\frac{1}{10} + \frac{3y}{5} \right) dy = \left[\frac{y}{10} + \frac{3y^2}{10} \right]_{\frac{1}{4}}^{\frac{1}{2}} \\
 &= \frac{1}{10} \left[\left(\frac{1}{2} + \frac{3}{4} \right) - \left(\frac{1}{4} + \frac{3}{16} \right) \right] = \frac{13}{160}
 \end{aligned}$$

Example – Solution



By definition,

$$g(x) = \int_{-\infty}^{\infty} f(x, y) dy = \int_0^1 \frac{2}{5} (2x + 3y) dy = \left. \frac{4xy}{5} + \frac{6y^2}{10} \right|_{y=0}^{y=1} = \frac{4x + 3}{5}$$

For $0 \leq x \leq 1$, and $g(x)=0$ elsewhere.

Similarly,
$$h(y) = \int_{-\infty}^{\infty} f(x, y) dx = \int_0^1 \frac{2}{5} (2x + 3y) dx = \frac{4(1 + 3y)}{5}$$

For $0 \leq y \leq 1$, and $h(y)=0$ elsewhere.

Problem:

Find 'K' if the joint probability density function of a bivariate random variable (X,Y) is given by

$$f(x, y) = \begin{cases} K(1-x)(1-y) & \text{if } 0 < x < 4 \\ 0 & \text{otherwise} \end{cases}$$

7. The joint probability distribution of X and Y is given by $f(x, y) = c(x^2 + y^2)$ for $x = -1, 0, 1, 3$ and $y = -1, 2, 3$. (i) Find the value of c . (ii) $P(x = 0, y \leq 2)$ (iii) $P(x \leq 1, y > 2)$ (iv) $P(x \geq 2 - y)$

Solution:

By data, $X = \{-1, 0, 1, 3\}$ and $Y = \{-1, 2, 3\}$

$$f(x, y) = c(x^2 + y^2)$$

The joint probability distribution of X and Y:

X \ Y	-1	2	3	$f(X)$
-1	$2c$	$5c$	$10c$	$17c$
0	c	$4c$	$9c$	$14c$
1	$2c$	$5c$	$10c$	$17c$
3	$10c$	$13c$	$18c$	$41c$
$g(Y)$	$15c$	$27c$	$47c$	$89c$

- (i) **Find c :** $1 = \sum f(x, y) = 89c$

$$c = \frac{1}{89}$$

$$(ii) \quad x = 0, y = \{-1, 2\}$$

$$\begin{aligned} P(x = 0, y \leq 2) \\ &= P(0, -1) + P(0, 2) \\ &= c + 4c = 5c \\ &= 5/89 \end{aligned}$$

$$(iii) \quad x = \{-1, 0, 1\}, y = \{3\}$$

$$\begin{aligned} P(x \leq 1, y > 2) \\ &= P(-1, 3) + P(0, 3) + P(1, 3) \\ &= 10c + 9c + 10c \\ &= 29c = 29/89 \end{aligned}$$

Cont.

(iv) By data, $X = \{-1, 0, 1, 3\}$ and $Y = \{-1, 2, 3\}$

$$\begin{aligned} P(x \geq 2 - y) &= P(x + y \geq 2) \\ &= P(-1, 3) + P(0, 2) + P(0, 3) + P(1, 2) + P(1, 3) + P(3, -1) + P(3, 2) + P(3, 3) \\ &= 10c + 4c + 9c + 5c + 10c + 10c + 13c + 18c \\ &= 79c = 79/89 \end{aligned}$$



Thank You !



M.Tech. (AIML)

Session : 6 (Probability Distribution)

Team AIML

BITS Pilani
Pilani Campus

Session : 6 Agenda

Distributions:-

- ❖ Bernoulli
 - ❖ Binomial
 - ❖ Poisson
 - ❖ Normal distributions.
 - ❖ Introduction to t-distribution
 - ❖ F-Distribution and
 - ❖ Chi Square distributions
-

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

$$\begin{aligned}
 \text{mean} = \mu = E(x) &= \sum x p(x) \\
 &= \sum x p^x q^{1-x}, \quad x=0,1 \\
 &= 0 \cdot p^0 \cdot q^1 + 1 \cdot p \cdot q^0 = p
 \end{aligned}$$

$$\begin{aligned}
 \text{variance } \sigma^2 &\approx E(x^2) - [E(x)]^2 \\
 E(x^2) &= \sum x^2 p(x) = \sum x^2 \cdot p^x q^{1-x}, \quad x=0,1 \\
 &= 0^2 p^0 q^1 + 1^2 \cdot p \cdot q^0 = p
 \end{aligned}$$

→ $\sigma^2 = p - p^2 = p(1-p) = pq$

Mean = p

Variance = pq

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

- Binomial distribution is a discrete probability distribution.
- Binomial distribution will be applied under the following experimental conditions
 - 1) The number of trials (n) is finite
 - 2) The trials are independent of each other
 - 3) The probability of success p is constant for each trial.
 - 4) Each trial results in two mutually exclusive events known as success and failure.

Mean & variance

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

$$\begin{aligned}
 \text{mean} = \mu &= E(x) = \sum x P(x) \\
 &= \sum x nC_x P^x q^{n-x} = \sum x \frac{n!}{x! (n-x)!} P^x q^{n-x} \\
 &= \sum \frac{n!}{(x-1)! (n-x)!} P^x q^{n-x} \\
 &= np \sum_{x=1}^n \frac{(n-1)!}{(x-1)! [(n-1)-(x-1)]!} P^{x-1} q^{(n-1)-(x-1)} \\
 &= np \left(q + p \right)^{n-1} \quad \underbrace{(q+p)^n}_{(q+p)^n = nC_0 q^n \cdot p^0 + nC_1 q^{n-1} \cdot p^1 + \dots + nC_n q^0 \cdot p^n} \\
 &= np
 \end{aligned}$$

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

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

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

Binomial Distribution

4.11 Which conditions for the binomial distribution, if any, fail to hold in the following situations?

(a) The number of persons having a cold at a family reunion attended by 30 persons.

Ans: Getting cold is not independent. So it is not binomial

(b) Among 8 projectors in the department office, 2 do not work properly but are not marked defective. Two are selected and the number that do not work properly will be recorded

Ans: Probability of success is not same in each trial. So this is not binomial

Binomial Distribution Thinking Challenge



You're a telemarketer selling service contracts for Macy's. You've sold 20 in your last 100 calls ($p = .20$). If you call **12** people tonight, what's the probability of

- A. No sales?
- B. Exactly 2 sales?
- C. At most 2 sales?
- D. At least 2 sales?

Binomial Distribution Solution*

$n = 12, p = .20$

A. $p(0) = .0687$

B. $p(2) = .2835$

C. $p(\text{at most } 2) = p(0) + p(1) + p(2)$
= $.0687 + .2062 + .2835$
= **.5584**

D. $p(\text{at least } 2) = p(2) + p(3) \dots + p(12)$
= $1 - [p(0) + p(1)]$
= $1 - .0687 - .2062$
= **.7251**

Understandings

Phrase	<i>Math Symbol</i>
“at least”	\geq
“more than” or “greater than”	$>$
“fewer than” or “less than”	$<$
“no more than”	\leq
“exactly”	$=$

Binomial Distribution



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

$$P(x \leq 2) = \sum p(x) = P(0) + P(1) + P(2)$$

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

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

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

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

Example

The probability that a bomb dropped from a plane will strike the target is $1/5$. if six such bombs are dropped.



Find the probability that

1) exactly two bombs hit the target

$$p = \frac{1}{5} \text{ and } q = \frac{4}{5}$$

2) At least two will hit the target

$$n=6$$

$$\begin{aligned} i) P(X=2) &= {}^n C_r \cdot p^r \cdot q^{n-r} \\ &= {}^6 C_2 \cdot \left(\frac{1}{5}\right)^2 \cdot \left(\frac{4}{5}\right)^4 \end{aligned}$$

$$\begin{aligned} ii) P(X \leq 2) &= P(X=0) + P(X=1) + P(X=2) \\ &= 1 - [P(X=0) + P(X=1)] \\ &= \end{aligned}$$

Example

An unbiased dice is thrown 5 times and occurrence of 1 or 6 considered as success find,

1) Probability of exactly one success

$$p = \text{getting 1 or 6}$$

$$p = 2/6$$

2) Probability of at least 4 success

$$q = 4/6$$

3) Probability of at least one success

$$n = 5$$

4) Mean and variance

(i) $P(X=1) = P(X=\text{Exactly one success})$

$$= {}^5C_1 \cdot \left(\frac{1}{3}\right)^1 \cdot \left(\frac{2}{3}\right)^4 = 5 \cdot \frac{1}{3} \times \frac{16}{81}$$

$\frac{5!}{1!4!} = 5$

$${}^nC_r \cdot p^r \cdot q^{n-r}$$

$$\text{mean} = np = 5 \times \frac{2}{6} = \frac{10}{6}$$

$$\text{Var.} = npq = 5 \times \frac{2}{6} \times \frac{4}{6} = \frac{40}{36}$$

(ii) $P(X \geq 4) = P(X=4) + P(X=5) + \dots$

$$= {}^5C_4 \cdot \left(\frac{1}{3}\right)^4 \cdot \left(\frac{2}{3}\right)^1 + {}^5C_5 \cdot \left(\frac{1}{3}\right)^5 \cdot \left(\frac{2}{3}\right)^0$$

(iii) $P(X \geq 1) = P(X=1) + P(X=2) + \dots$

$${}^5C_0 = \frac{5!}{0!5!} = 1$$

$${}^5C_4 = \frac{5!}{4!1!} = 5$$

$${}^5C_5 = \frac{5!}{5!0!} = 1$$

$$= 1 - P(X=0)$$

Example

Seven dice are thrown 729 times. How many times do you expect at least four dice to show three or five?


Seven dice are thrown
 $n = 7$
 $r = 4 \text{ or } 5 \text{ or } 6 \text{ or } 7$
 $p = 2/6 = 1/3$

$$729 \times [P(X=4) + P(X=5) + P(X=6) + P(X=7)]$$

$$N \times [n_C_r \cdot p^r \cdot q^{n-r}] = 729 \left[{}^7C_4 \cdot \left(\frac{1}{3}\right)^4 \cdot \left(\frac{2}{3}\right)^3 + {}^7C_5 \cdot \left(\frac{1}{3}\right)^5 \cdot \left(\frac{2}{3}\right)^2 + {}^7C_6 \cdot \left(\frac{1}{3}\right)^6 \cdot \left(\frac{2}{3}\right)^1 + {}^7C_7 \cdot \left(\frac{1}{3}\right)^7 \cdot \left(\frac{2}{3}\right)^0 \right]$$

Ans.
δ<?

Example

Find the binomial distribution if the mean is 4 and variance is 3

B.D.

$$np = 4 \rightarrow ①$$

$$npq = 3 \rightarrow ②$$

$$n = \frac{4}{\frac{1}{4}}$$

$$n = 16$$

$$4q = 3 \Rightarrow q = 3/4$$

then $p = 1/4$

Additional problem

In a large number of parts manufactured by a machine , the mean number of defectives in a sample of 20 is 2. Out of 1000 such samples , how many would be expected to contain atleast 3 defective parts.

Sol: given $n=20$;

$$\text{mean} = \mu = 2, \text{ but } \mu = np \Rightarrow 2 = 20p \therefore p = 0.1$$

$$\therefore P(X \geq 3) = {}^{20}C_3 (0.1)^x (0.9)^{20-x} = [P(x=3) + P(x=4) + \dots + P(x=20)]$$

Therefore the probability that there are atleast 3 defective parts in a sample of 20 is

$$P(X \geq 3) = 1 - P(X < 3) = 1 - [P(0) + P(1) + P(2)]$$

$$\therefore P(X \geq 3) = 0.323$$

Hence out of 1000 samples we get $0.323 \times 1000 = 323$ samples

Discrete Probability Distribution

Discrete Probability Distributions

Binomial

Hypergeometric

Poisson

Binomial Distribution Recall

A **binomial experiment** is a probability experiment that satisfies the following conditions.

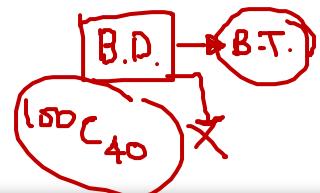
1. The experiment is **repeated** for a fixed number of trials, where each trial is **independent** of other trials.
2. There are only **two possible outcomes** of interest for each trial. The outcomes can be classified as a success (S) or as a failure (F).
3. The probability of a success $P(S)$ is the same for each trial.
4. The random variable x counts the number of successful trials.

Poisson Distribution

Definition

\downarrow
 $\{[n \rightarrow \infty; p \rightarrow 0]\}$

*n is very large
p is very small (success)*



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

$$P(X=r) = \frac{\bar{e}^d \cdot d^r}{r!}$$

\downarrow
 R.V. $d = np$

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

Poisson distribution is the discrete probability distribution of a discrete random variable X, which may not have upper bound. It is defined for non negative values of x as follows

Cont...

"P.D. is a particular case of B.D."

No fixation for
~~Upper Case~~
~~150000~~

- Poisson distribution is suitable for rare events for which the probability of occurrence 'p' is very small and the number of trials 'n' is very large.
 - Also binomial distribution can be approximated by Poisson distribution when $n \rightarrow \infty$ and $p \rightarrow 0$ such that $\lambda = np = \text{constant}$
 - ① Telephone Calls →
 - ② Patients in Hospital
 - ③ Arrival pattern of def. Vehicles in Work shop.
 - ④ Demand pattern of def. Vehicles in Workshop
 - ⑤ Emission of Radioactive particles
 - ⑥ No. of death from M.A. or Cancer
 - ⑦ No. of accidents reported in a particular city
 - ⑧ No. of car accidents in one unit of time
- $n \rightarrow \infty$
 $p \rightarrow 0$

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

Poisson Distribution

Police records show that number of accident victims died in road traffic accidents is 0.1%. What is the probability that among 500 randomly selected accident victims

- (i) none have died?
- (ii) at least 3 have died
- (iii) between 2 and 6 have died

Example:

Example 6: A hospital switch board receives an average of 4 emergency calls in a 10 minute interval. What is the probability that

- (i) there are at most 2 emergency calls in 10 minute interval
- (ii) there are exactly 3 emergency calls in a 10 minute interval

Ans: Given Mean = 4calls. We have

$$P(x) = \frac{(4)^x e^{-4}}{x!}$$

- (i) $P(\text{at most 2 calls}) = P(x \leq 2) = P(0)+P(1)+P(2) = 0.2381$
- (ii) $P(\text{Exactly 3 calls}) = P(x = 3) = 0.1954$

Example:

Example 2: The mean number of power outages in the city of Bangalore follows a Poisson variate and is 4 per year. Find the probability that in a given year,

- a) there are exactly 3 outages, b) there are more than 3 outages.

Poisson Mean and Variance

Mean:

$$\mu = \lambda$$

For a Poisson random variable, the variance and mean are the same!

Variance and Standard Deviation

$$\sigma^2 = \lambda$$

$$\sigma = \sqrt{\lambda}$$

where λ = expected number of occurrences in a given experiment.

Example

- 4.55 In a factory, 8% of all machines break down at least once a year. Use the Poisson approximation to the binomial distribution to determine the probabilities that among 25 machines (randomly chosen in the factory):
- (a) 5 will break down at least once a year;
 - (b) at least 4 will break down once a year;
 - (c) anywhere from 3 to 8, inclusive, will break down at least once a year.

Suggested Problems

Given that a switch board of a consultant's office receives on the average 0.6 calls per minute, find the probabilities that

- (a) in a given minute there will be at least 1 call;
 - (b) in a 4-minute interval there will be at least 3 calls.
-

Solution:

a) here

$$\lambda = 0.6 \text{ and } t = 1 \therefore p(x, \lambda t) = \frac{e^{-\lambda t} (\lambda t)^x}{x!}$$

$$p(x, \lambda t) = \frac{e^{-0.6} (0.6)^x}{x!}$$

b)

$$P(X \geq 1) = 1 - P(X < 1) = 1 - P(0)$$

$$= 1 - e^{-0.6} = 0.4512$$

$$\lambda = 0.6 \text{ and } t = 4 \therefore p(x, \lambda t) = \frac{e^{-2.4} (2.4)^x}{x!}$$

$$P(X \geq 3) = 1 - P(X < 3) = 1 - [P(0) + P(1) + P(2)]$$

$$= 0.4303$$

Exercise:

The number of customers arriving at a cafeteria at an average rate of 0.3 per minute.

- (a) Find the probability that exactly 2 customers arrive in a 10-minute span.
- (b) Find the probability that 2 or more customers arrive in a 10-minute span.
- (c) Find the probability that exactly one customer arrives in a 5-minute span and one customer arrives in the next 5-minute span.

Exercise:

The average number of trucks arriving on any one day at a truck depot in a certain city is known to be 12. what is the probability that on a given day fewer than nine trucks will arrive at this depot?

Exercise:

A certain kind of sheet metal has on the average, five defects per 10 square feet. If we assume a Poisson distribution, what is the probability that a 15-square-foot sheet of the metal will have at least six defects?

Exercise:

The number of flaws in a fiber optic cable follows a poisson process with an average of 0.6 per 100 feet.

- a) Find the probability of exactly 2 flaws in 200 foot cable.
- b) Find the probability of exactly 1 flaw in first 100 feet and exactly 1 flaw in the second 100 feet.

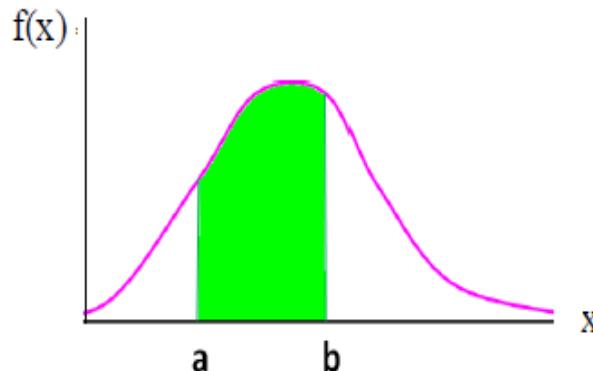
Exercise:

Suppose that on an average 1 out of 10000 houses catch fire in a year in a district. If there are 2000 houses in the district, find the probability that exactly 5 house will catch fire during that year.

RECALL: Continuous Random Variables

A continuous random variable can assume any value in an interval on the real line or in a collection of intervals.

The probability of the random variable assuming a value within some given interval from x_1 to x_2 is defined to be the area under the graph of the **probability density function** between x_1 and x_2



Example:

- Height of students in a class
- Amount of ice tea in a glass
- Change in temperature throughout a day
- Price of a car in next year

Continuous Probability Distributions:

Normal Distribution:

Normal distribution is a general distribution. Any information such as heights, weights, lengths, widths, sizes, profits, prices, wages, time etc. containing sufficiently large data with mean and variance follows normal distribution. The distribution of errors of repeated measurements follows normal distribution.

Normal Distribution

A continuous random variable X which assumes all possible values in the entire real space, i.e., $-\infty < x < \infty$ is said to follow normal distribution with two parameters such as mean μ and variance σ^2 if it has the probability density function given by

$$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad -\infty < x < \infty$$
$$\quad \quad \quad -\infty < \mu < \infty, \sigma > 0$$

Normal Distribution

Change in Mean determines the shift in the distribution

Change in the deviation determines the spread of the data points

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

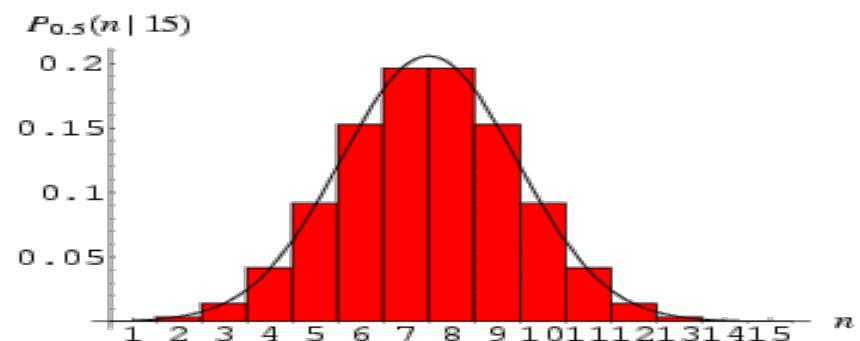
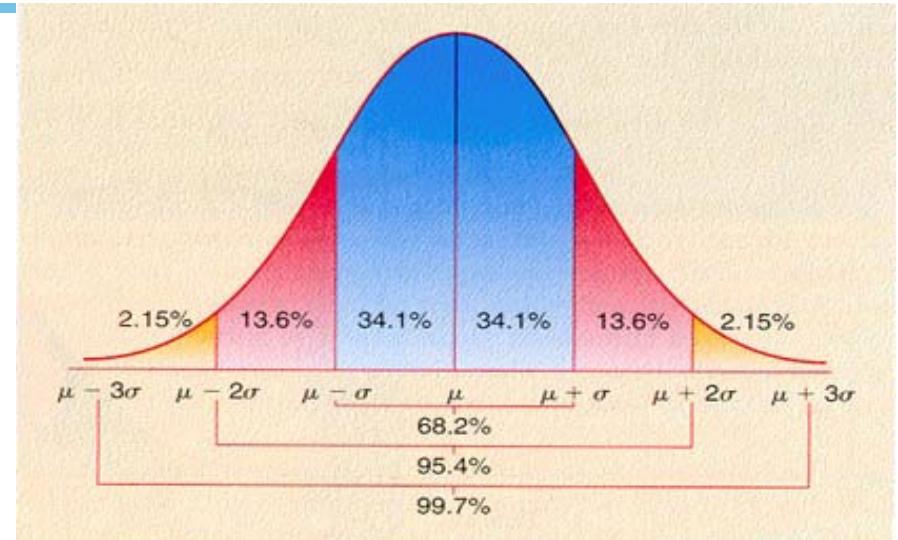
μ = mean

σ = standard deviation

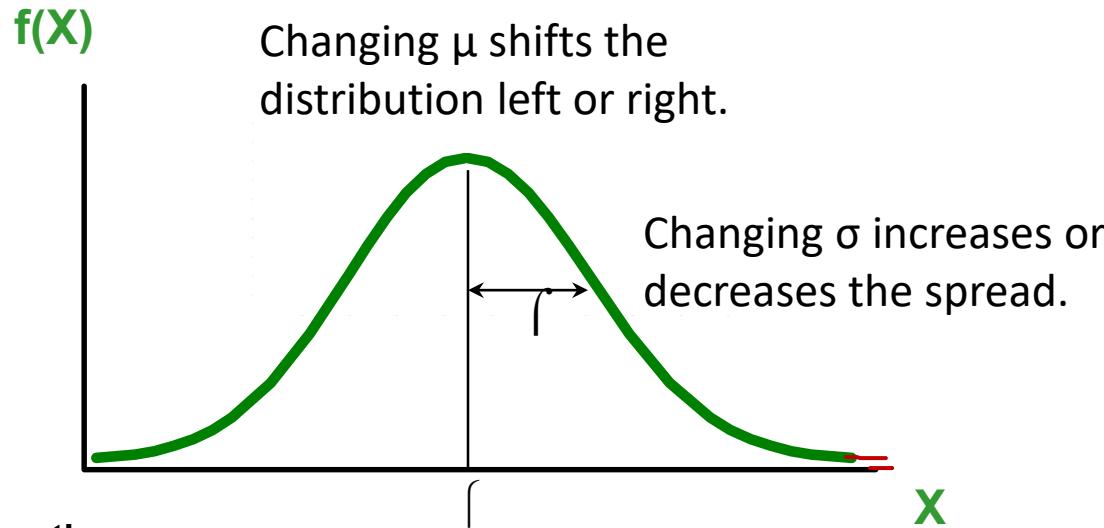
$\pi = 3.14159$

$e = 2.71828$

This is a bell shaped curve with different centers and spreads depending on μ and σ



Normal Distribution



Properties:

1. Normal curve is bell shaped and symmetric about the mean
2. Mean = Mode = Median
3. Total area under normal curve is equal to 1
4. Normal curve approaches but never touches the x axis as it extends farther and farther away from the mean

Normal Distribution

$$E(X) = \mu = \int_{-\infty}^{+\infty} x \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

$$\text{Var}(X) = \sigma^2 = \int_{-\infty}^{+\infty} x^2 \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx - \mu^2$$

Standard Deviation(X) = σ

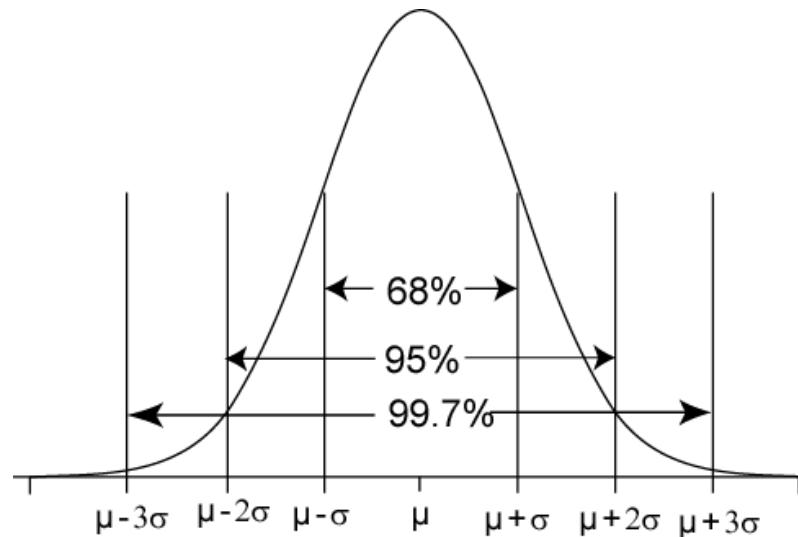
No matter what μ and σ are, the area between $\mu-\sigma$ and $\mu+\sigma$ is about 68%; the area between $\mu-2\sigma$ and $\mu+2\sigma$ is about 95%; and the area between $\mu-3\sigma$ and $\mu+3\sigma$ is about 99.7%. Almost all values fall within 3 standard deviations.

68-95-99.7 Rule for Normal Distributions

68% of the AUC(Area under curve) within $\pm 1\sigma$ of μ

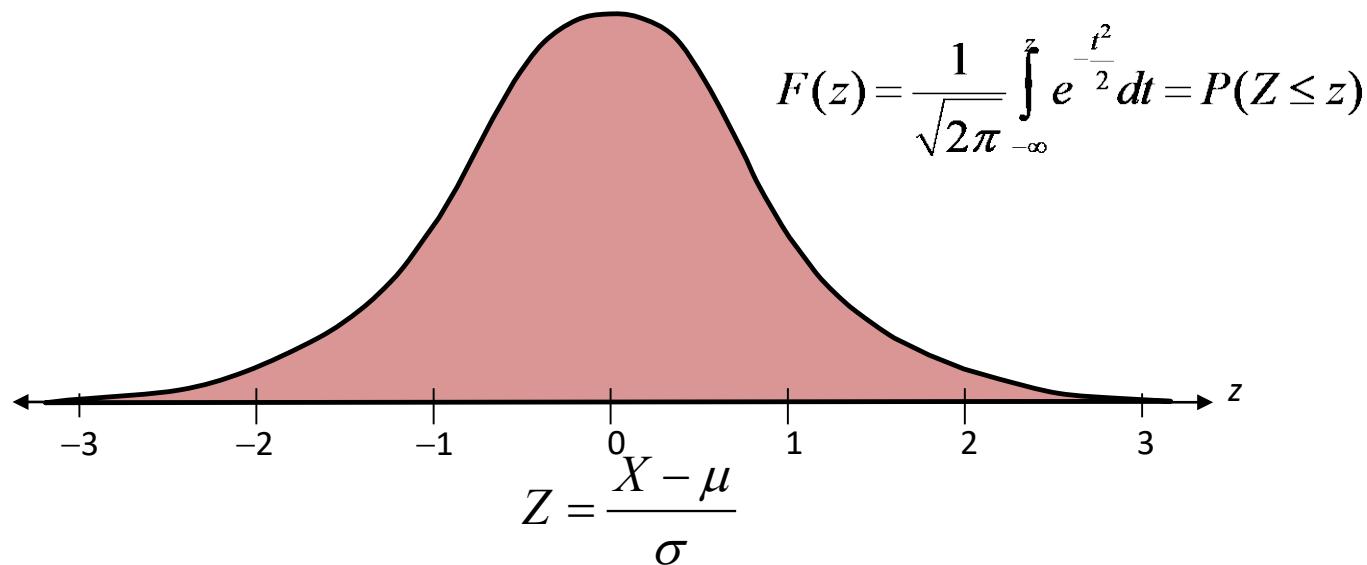
95% of the AUC within $\pm 2\sigma$ of μ

99.7% of the AUC within $\pm 3\sigma$ of μ



Standard Normal Distribution

The **standard normal distribution** is a normal distribution with a mean of 0 and a standard deviation of 1.



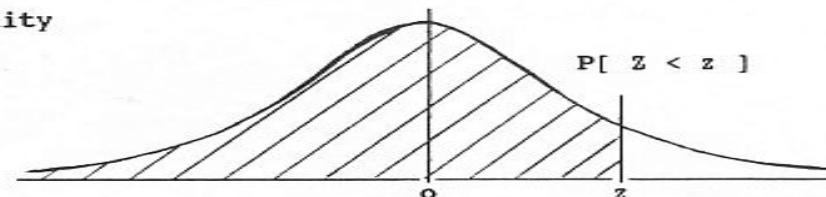
All normal distributions can be converted into the standard normal curve by subtracting the mean and dividing by the standard deviation

STANDARD STATISTICAL TABLES

1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value z
i.e.

$$P[Z < z] = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}z^2) dz$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
z	3.00	3.10	3.20	3.30	3.40	3.50	3.60	3.70	3.80	3.90
P	0.9986	0.9990	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1.0000

Problem

If a random variable has the standard normal distribution, find the probability
that it will take on a value

a) less than 1.75 = $P(Z < 1.75) = F(1.75) = 0.9599$

b) less than -1.25 = $P(Z < -1.25) = F(-1.25) = 1 - F(1.25) = 0.1056$

c) greater than 2.06 = $P(Z > 2.06) = 1 - F(2.06) = 0.0197$

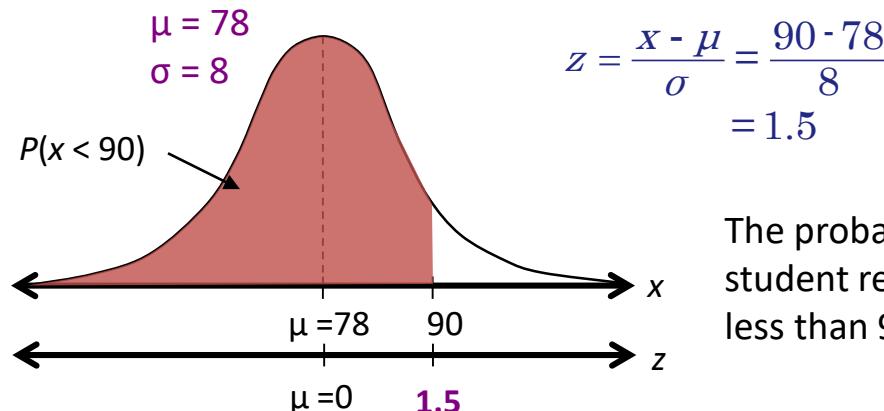
d) greater than -1.82 = $P(Z > -1.82) = 1 - F(-1.82) = 1 - 0.0344 = 0.9656$

$$P(Z > -1.82) = 1 - F(-1.82) = 1 - (1 - F(1.82)) = F(1.82) = 0.9656$$

Probability and Normal Distributions

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.



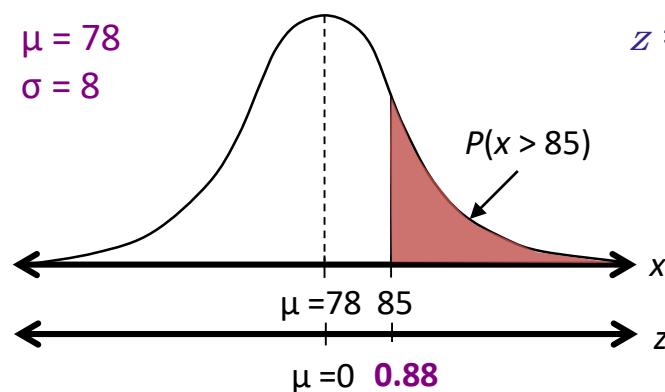
The probability that a student receives a test score less than 90 is **0.9332**.

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

Probability and Normal Distributions

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.



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

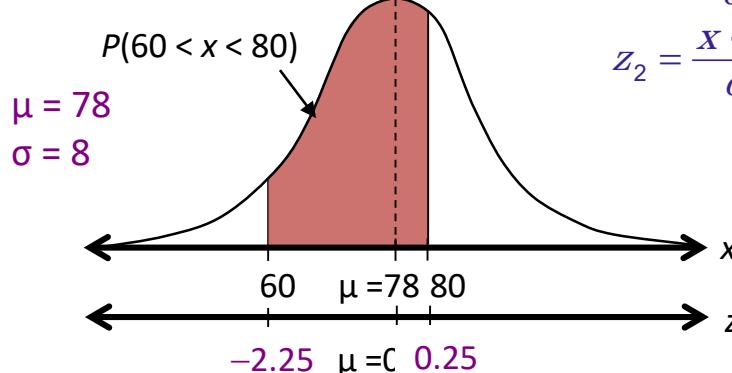
The probability that a student receives a test score greater than 85 is 0.1894.

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

Probability and Normal Distributions

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.

$$P(60 < x < 80) = P(-2.25 < z < 0.25) = F(0.25) - F(-2.25) = 0.5987 - 0.0122 = 0.5865$$

Problems

If $Z \sim N(0,1)$, to find constants a and b such that $P(Z \leq a) = 0.9147$ and $P(Z \geq b) = 0.0526$

$$a = 1.37 \text{ and } b = 1.62$$

If $X \sim N(25, 36)$, find a constant c such that $P(|X - 25| \leq c) = 0.9544$.

Solution :

$$P(|X - 25| \leq c) = 0.9544$$

$$\Rightarrow P\left(\frac{-c}{6} \leq \frac{X - 25}{6} \leq \frac{c}{6}\right) = 0.9544 \Rightarrow \Phi\left(\frac{c}{6}\right) - [1 - \Phi\left(\frac{c}{6}\right)] = 0.9544$$

$$\Rightarrow \Phi\left(\frac{c}{6}\right) = 0.9772 \Rightarrow \frac{c}{6} = 2 \Rightarrow c = 12$$

Problem

The time for oil to percolate to all parts of an engine can be treated as a random variable having a normal distribution with mean 20 seconds. Find its standard deviation if the probability is 0.25 that will take a value greater than 31.5 seconds.

Given : Mean $\mu = 20$ seconds, $P(X > 31.5) = 0.25$

$\sigma = ?$

$$z = \frac{x - \mu}{\sigma} = \frac{31.5 - 20}{\sigma}$$
$$\sigma = \frac{11.5}{z} = \frac{11.5}{0.675} = 17.04$$

Ex.

Monthly Salary X in a big organisation is normally distributed with mean Rs 3000 and standard deviation of Rs.250 what should be the minimum salary of a worker in this organisation so that the probability that he belongs to top 5%workers?

Problem

Butterfly-style valves used in heating and ventilating industries have a high flow coefficient. Flow coefficient can be modeled by a normal distribution with mean $496 C_v$ and standard deviation $25C_v$. Find the probability that a valve will have a flow coefficient of

- a) atleast $450C_v$

$$P(X > 450) = P[(450 - 496) / 25] = P(Z > -1.84) = 1 - F(-1.84) = 1 - 0.0329 = .9671$$

- b. between 445.5 and $522C_v$

$$P(445.5 < X < 522) = P(-2.02 < Z < 1.04) = F(1.04) - F(-2.02) = 0.8508 - 0.0217 = 0.8291$$

Normal Approximation to Binomial Distribution:

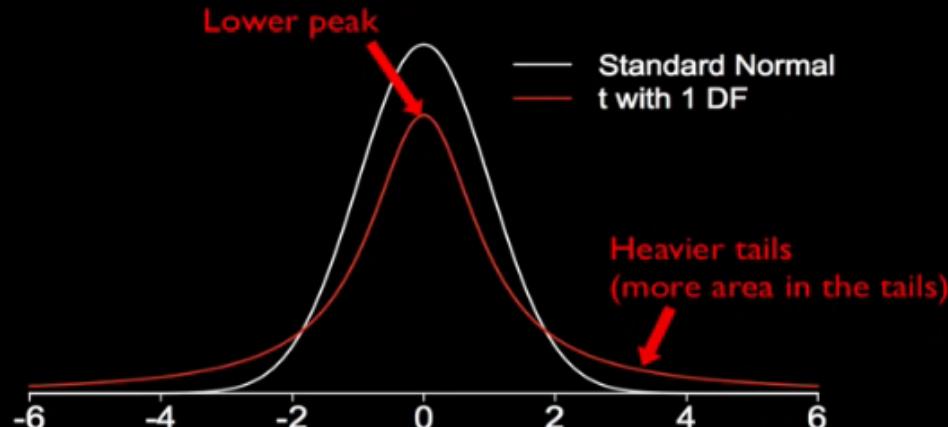


If n tends to infinity (preferably $n>30$) and neither p nor $1-p$ is so small, that is p and $1-p$ both are large preferably $np>15$ and $n(1-p)>15$ then binomial probabilities can be approximated by normal probabilities using continuity correction and taking

$$Z = \frac{X - \mu}{\sigma} = \frac{X - np}{\sqrt{np(1 - p)}}$$

Where X is binomial random variable approximated as normal random variable.

Introduction to t-distribution



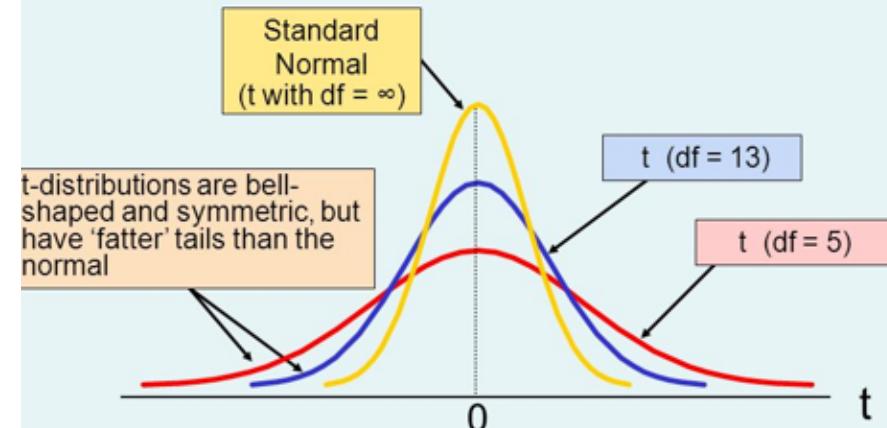
As the degrees of freedom increase, the t distribution tends toward the standard normal distribution

The pdf of the t distribution with ν degrees of freedom:

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} (1 + \frac{t^2}{\nu})^{-\frac{\nu+1}{2}}$$

for $-\infty < t < \infty$

Note: $t \rightarrow Z$ as n increases



Introduction to t-distribution (Sampling distribution of mean when σ unknown)

The t-distribution, also known as the **Student's t-distribution**, is a type of probability distribution that is similar to the normal distribution with its bell shape but has heavier tails.

It is used for estimating population parameters for **small sample sizes when**

t - Distribution When σ is unknown and $n < 30$.

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} \text{ with } (n-1) \text{ d.o.f.}$$

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

Where sample S.D. is calculated by formula

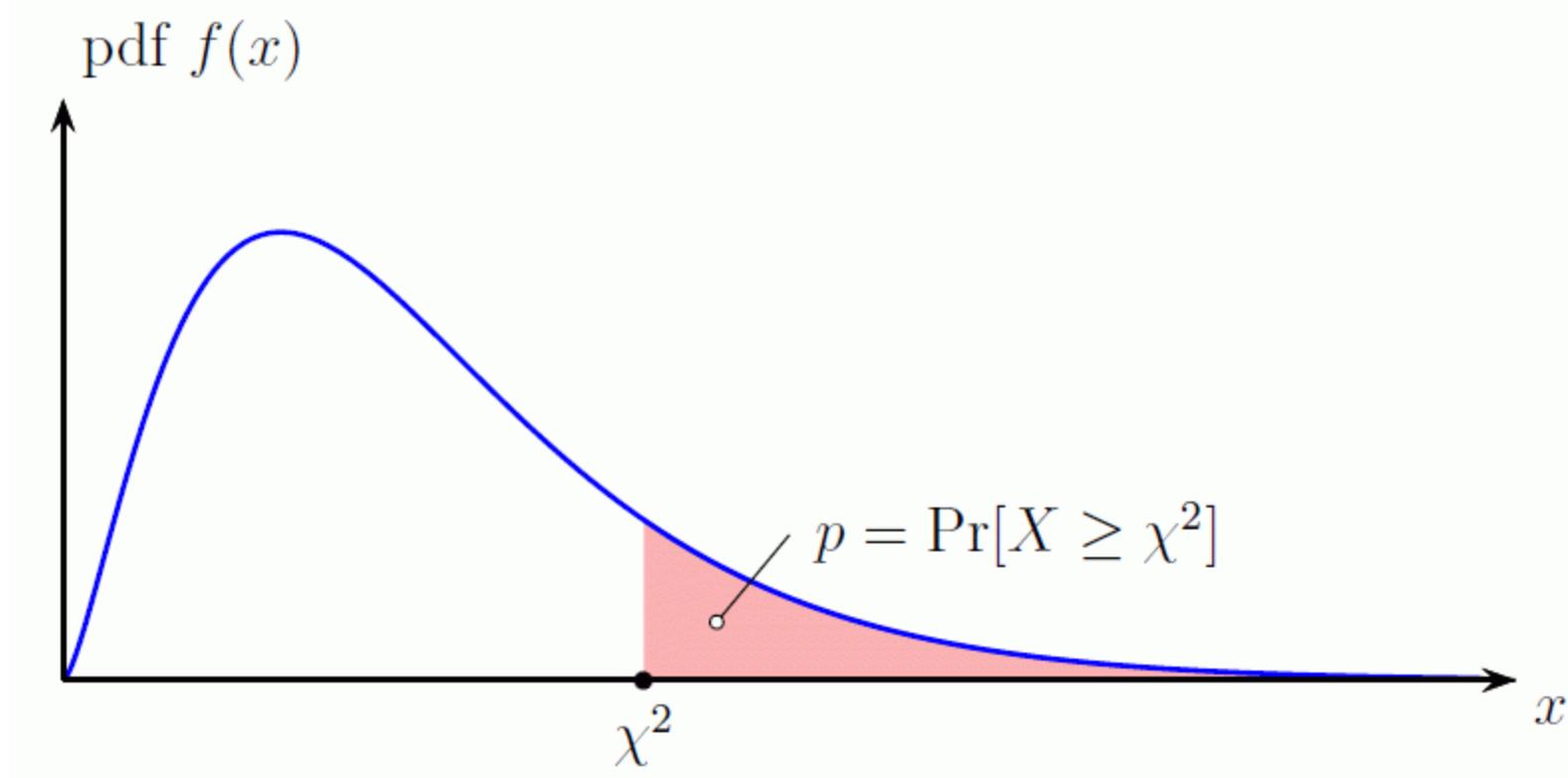
Introduction to Chi-Square (χ^2) distribution

- Chi-Square distribution pdf is given by

$$f(x) = \begin{cases} \frac{1}{2^{\vartheta/2} \Gamma(\frac{\vartheta}{2})} x^{\frac{\vartheta}{2}-1} e^{-\frac{x}{2}} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

where ϑ is the parameter of distribution, also known as degrees of freedom.

- Introduction to Chi-Square(χ^2) distribution



Introduction to Chi-Square (χ^2) distribution

- ❖ Chi-Square distribution curve is not symmetrical, and hence not a normal curve.
 - ❖ Chi-Square varies from 0 to ∞ (Curve lies entirely in first quadrant).
 - ❖ It depends only on degrees of freedom ϑ .
 - ❖ It is very important in estimation and hypothesis testing.
 - ❖ It is used in sampling distributions of the sample variance, analysis of variance.
 - ❖ It is used as a measure of goodness of fit.
-

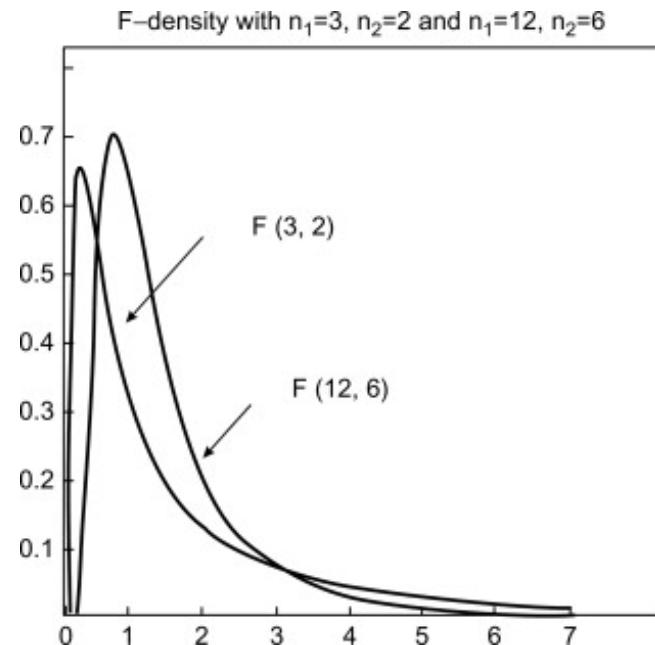
Introduction to F-distribution

(Sampling distribution of the ratio of two sample variances)

- The pdf for a random variable $X \sim F(n_1, n_2)$ is given by

$$f(x) = \frac{\Gamma\left(\frac{n_1 + n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} x^{\frac{n_1}{2}-1} \left(1 + \frac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}}, x > 0$$

- F- distribution curve lies entirely in first quadrant.
- The F- curve depends not only on the two parameters ϑ_1 and ϑ_2 but also on the order in which they are stated.



Introduction to F-distribution

(Sampling distribution of the ratio of two sample variances)

- ❖ The F-distribution was developed by Fisher to study the behavior of two variances from random samples taken from two independent normal populations.
- ❖ In applied problems we may be interested in knowing whether the population variances are equal or not, based on the response of the random samples.
- ❖ If S_1^2 and S_2^2 are variances of independent random sample of size n_1 and n_2 from a normal populations with variances σ_1^2 and σ_2^2 , then

$$F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2}$$

which follows F –distribution with $\vartheta_1 = n_1 - 1$ and $\vartheta_2 = n_2 - 1$ d.o.f.



Thanks



BITS Pilani

Pilani|Dubai|Goa|Hyderabad

Introduction to Statistical Methods

Team ISM



Session No 7

Testing of Hypothesis

(24/25 December, 2022)

Contact Session 7: Module 4: Hypothesis Testing

Contact Session	List of Topic Title	Reference
CS - 7	Sampling – random sampling and Stratified sampling, Sampling distribution – Central Limit theorem, Estimation– Interval Estimation, Confidence level	T1 & T2

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

Sample No.	Sample size	Mean	SD
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
Overall	100	2162.24	732.26

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.

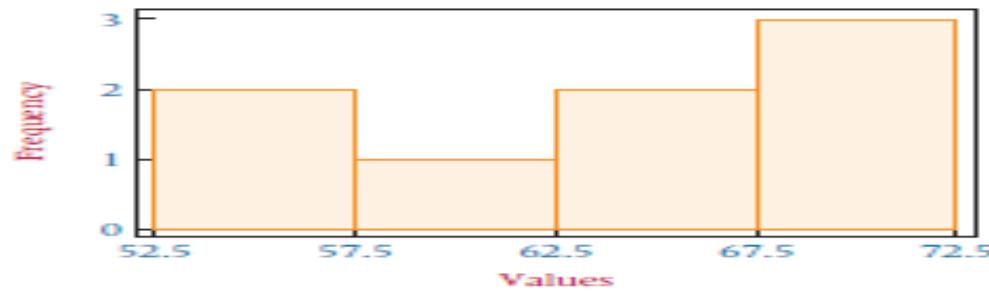
 - The sampling distribution of a statistic depends on the distribution of the population, the size of the sample, and the method of sample selection.
-

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.

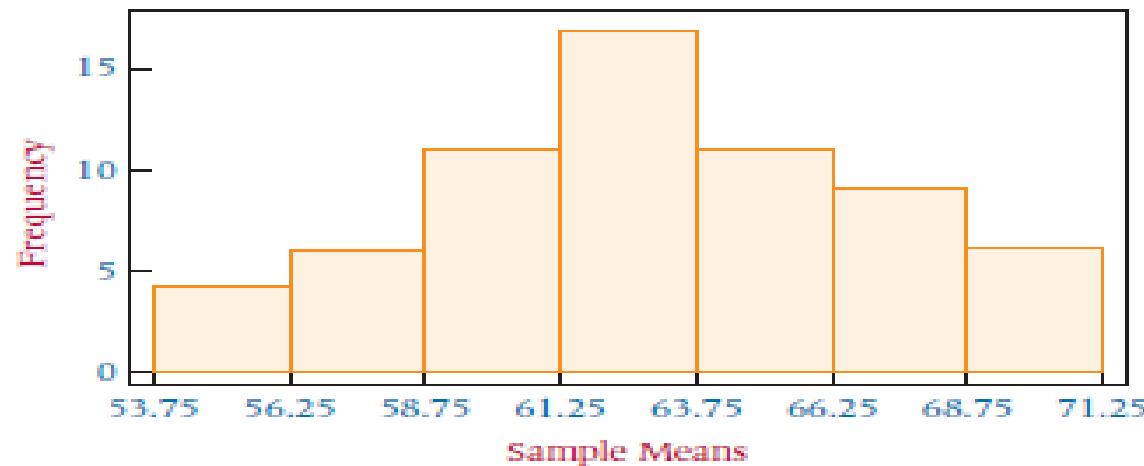
(54,54)	(55,54)	(59,54)	(63,54)
(54,55)	(55,55)	(59,55)	(63,55)
(54,59)	(55,59)	(59,59)	(63,59)
(54,63)	(55,63)	(59,63)	(63,63)
(54,64)	(55,64)	(59,64)	(63,64)
(54,68)	(55,68)	(59,68)	(63,68)
(54,69)	(55,69)	(59,69)	(63,69)
(54,70)	(55,70)	(59,70)	(63,70)
(64,54)	(68,54)	(69,54)	(70,54)
(64,55)	(68,55)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

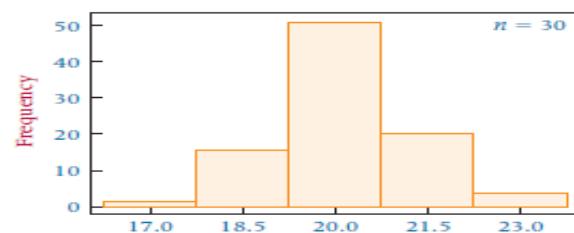
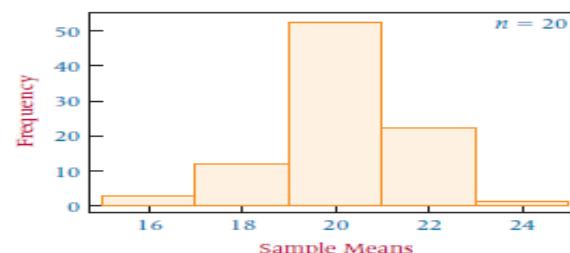
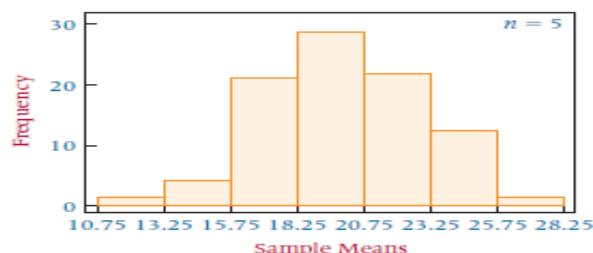
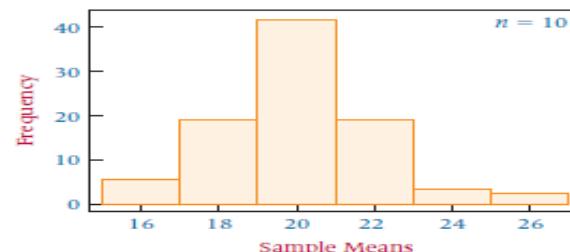
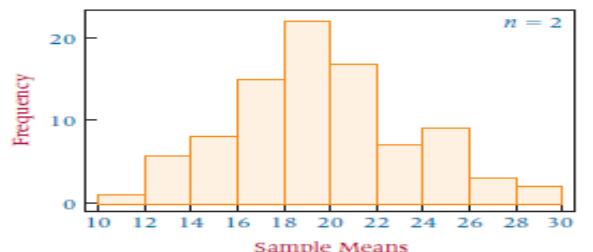
- Again using an Excel-produced histogram, we can see the shape of the distribution of these sample means.



Conclusions

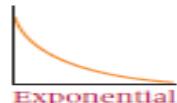
- Notice that the shape of the histogram for sample means is quite unlike the shape of the histogram for the population.
- The sample means appear to “pile up” toward the middle of the distribution and “tail off” toward the extremes.
- As sample sizes become much larger, the sample mean distributions 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 $b = 30$

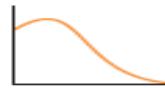


Shapes of the Distributions of Sample Means

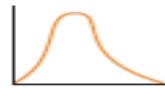
Population Distribution



$n = 2$



$n = 5$



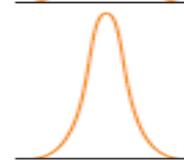
$n = 30$



Uniform



Normal



Central Limit Theorem

- If samples of size n are drawn randomly from a population that has a mean of μ and a standard deviation of σ , 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 determine z scores for individual values from a normal distribution:

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

- 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.

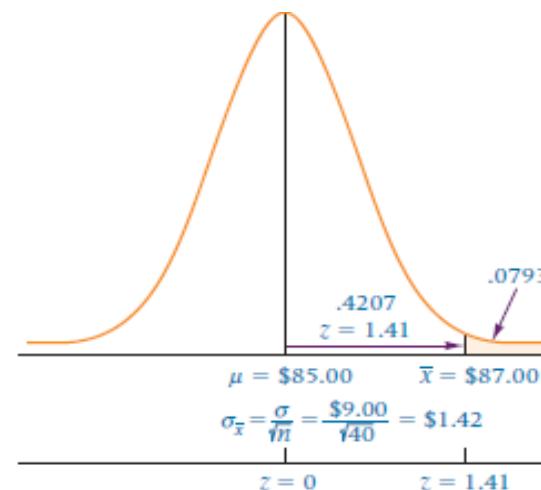
If a random sample of 40 customers is taken, what is the probability that the sample average expenditure per customer for this sample will be \$87.00 or more?

Solution

Because the sample size is greater than 30, the central limit theorem
Can be used, and the sample means are normally distributed.

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

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



Example

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.

Example

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?

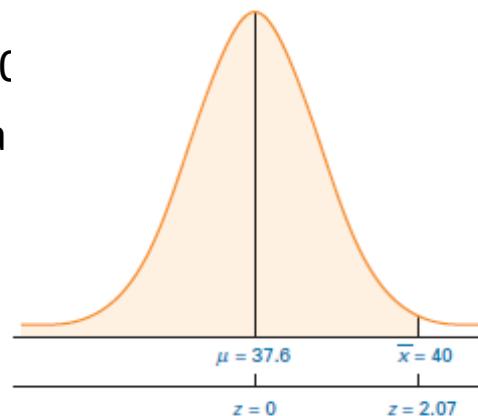
Solution

- 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 .4808.
- 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?

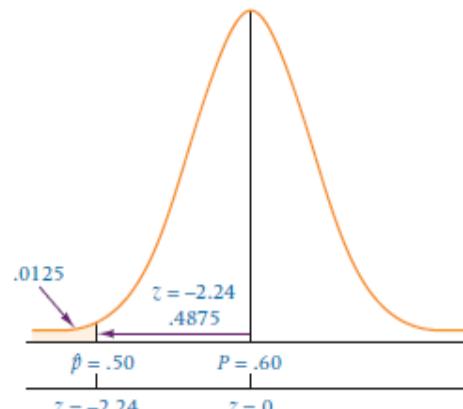
Solution

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

The z formula yields

The probability

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



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

Example

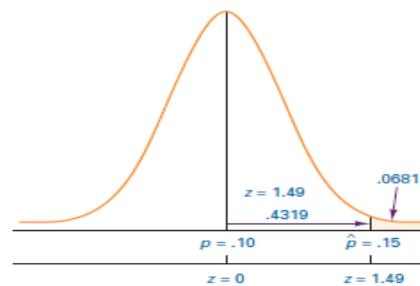
If 10% of a population of parts is defective, what is the probability of randomly selecting 80 parts and finding that 12 or more parts are defective?

Solution

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

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

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



on, .15, and

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

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 μ

100(1 - α)% CONFIDENCE
INTERVAL TO ESTIMATE μ :
 σ KNOWN (8.1)

or

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

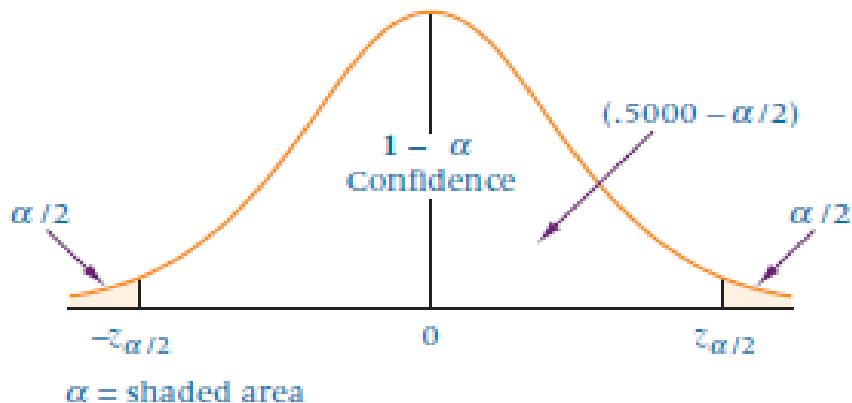
$$\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where

α = 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. □

Example

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 table yields a z value of 1.645 for the area of .45

The confidence interval is

$$\begin{aligned} \bar{x} - z \frac{\sigma}{\sqrt{n}} &\leq \mu \leq \bar{x} + z \frac{\sigma}{\sqrt{n}} \\ 10.455 - 1.645 \frac{7.7}{\sqrt{44}} &\leq \mu \leq 10.455 + 1.645 \frac{7.7}{\sqrt{44}} \\ 10.455 - 1.910 &\leq \mu \leq 10.455 + 1.910 \\ 8.545 &\leq \mu \leq 12.365 \end{aligned}$$

Example

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
- where $q = 1 - p$. Recall that this formula can be applied only when $n \cdot p$ and $n \cdot q$ are **greater than 5**.
$$z = \frac{\hat{p} - p}{\sqrt{\frac{p \cdot q}{n}}}$$
- for confidence interval purposes ...— 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

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

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

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

Example

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 proportions

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

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

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

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

Home Work Problems

Question :

Car mufflers are constructed by nearly automatic machine. One manufacturer finds that, for any type of car muffler, the time for a person to set up and complete a production run has a normal distribution with mean 1.82 hours and standard deviation 1.20.

What is the probability that the sample mean of the next 40 runs will be from 1.65 to 2.04 hours ?

Question :

Engine bearings depend on a film of oil to keep shaft and bearing surfaces separated. Insufficient lubrication causes bearings to be overloaded. The insufficient lubrication can be modeled as a random variable having a mean 0.6520 ml and standard deviation 0.0125 ml.

The sample mean of insufficient lubrication will be obtained from a random sample of 60 bearings.

What is the probability that sample mean \bar{x} will be between 0.600 ml and 0.640 ml ?

Question :

A random sample size of $n = 100$ is taken from a population with $\sigma = 5.1$.

Given that the sample mean is $\bar{x} = 2.16$,

construct a 95% confidence interval for the population mean μ .

Question :

With reference to the data in section 2.1 (of R1) , we have

$n = 50$, $\bar{x} = 305.58$ nm, and $s^2 = 1366.86$ (hence, $s=36.97$ nm),

Construct a 99% confidence interval for the population mean of all nanopillars.

*

245	333	296	304	276	336	289	234	253	292
366	323	309	284	310	338	297	314	305	330
266	391	315	305	290	300	292	311	272	312
315	355	346	337	303	265	278	276	373	271
308	276	364	390	298	290	308	221	274	343





Thank You



BITS Pilani

Pilani|Dubai|Goa|Hyderabad

Introduction to Statistical Methods

Team ISM





Session No 9

Testing of Hypothesis

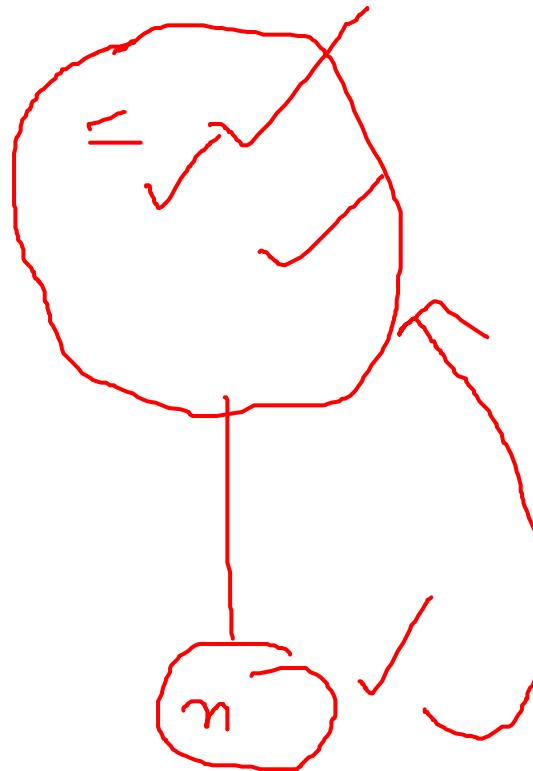
(4th /5th FEB ,2023)

Contact Session	List of Topic Title	Reference
CS - 9	Testing of Hypothesis - Type I & II errors, Critical region, t – test, Chi – Square test and F – test(Introduce and discuss these tests)	T1:Chapter 7 ,8,9 & 10
HW	Problems on Testing of Hypothesis	T1:Chapters 7 to 10

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 μ

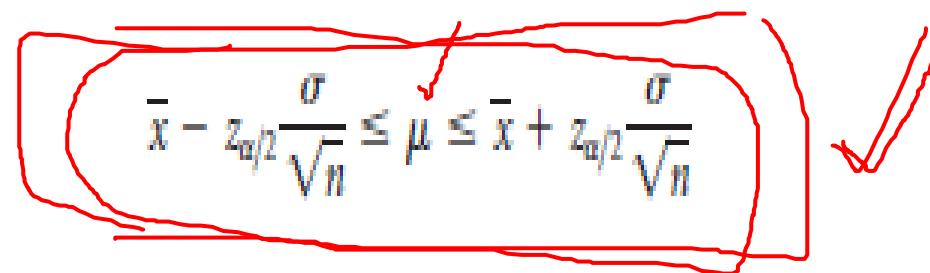
$$-\bar{z}_{\alpha/2} \leq \underline{\underline{z}} \leq \bar{z}_{\alpha/2}$$

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

100(1 - α)% CONFIDENCE
INTERVAL TO ESTIMATE μ :
 σ KNOWN (8.1)

or

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



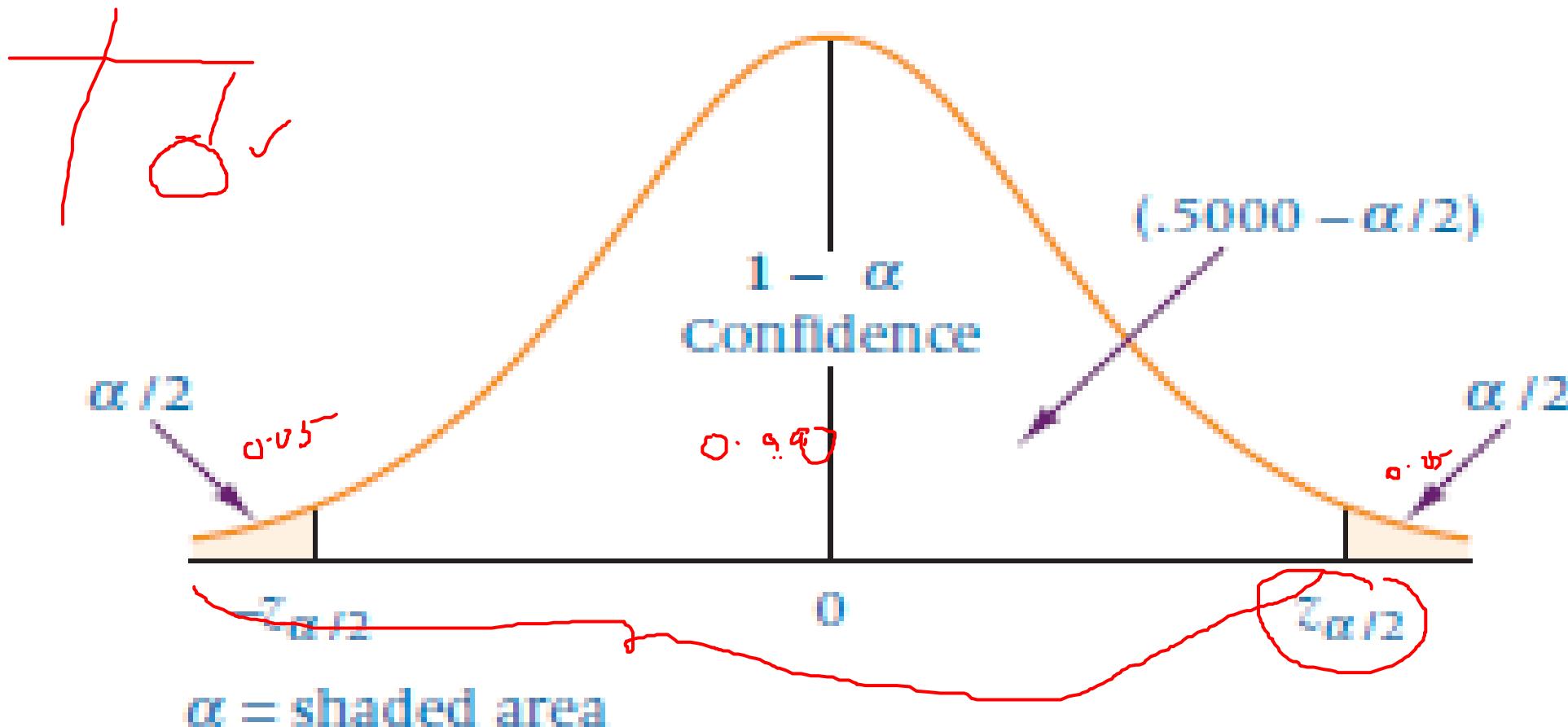
where

α = 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.

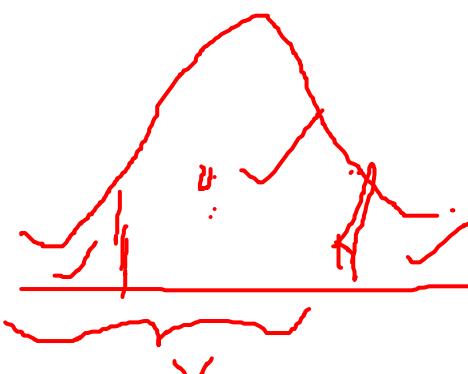
Solution

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.

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

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

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



Example

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 table yields a z value of 1.645 for the area of .45

The confidence interval is



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

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

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

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

Example

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.


$$\frac{\sigma}{\sqrt{n}}$$

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

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

$$\begin{aligned} z &= \frac{x - \mu}{\sigma} \\ &= \frac{x - np}{\sqrt{npq}} \end{aligned}$$

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

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

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

$$\boxed{.29 \leq p \leq .49}$$

Example

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 proportions

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

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

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

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

Home Work Problems



Question :

Car mufflers are constructed by nearly automatic machine. One manufacturer finds that, for any type of car muffler, the time for a person to set up and complete a production run has a normal distribution with mean 1.82 hours and standard deviation 1.20.

What is the probability that the sample mean of the next 40 runs will be from 1.65 to 2.04 hours ?

Question :

Engine bearings depend on a film of oil to keep shaft and bearing surfaces separated. Insufficient lubrication causes bearings to be overloaded. The insufficient lubrication can be modeled as a random variable having a mean 0.6520 ml and standard deviation 0.0125 ml.

The sample mean of insufficient lubrication will be obtained from a random sample of 60 bearings.

What is the probability that sample mean \bar{x} will be between 0.600 ml and 0.640 ml ?

Question :

A random sample size of $n = 100$ is taken from a population with $\sigma = 5.1$.

Given that the sample mean is $x = 2.16$,

construct a 95% confidence interval for the population mean μ .

Question :

With reference to the data in section 2.1 (of R1) , we have_

$n = 50$, $x = 305.58$ nm, and $s^2 = 1366.86$ (hence, $s=36.97$ nm),

Construct a 99% confidence interval for the population mean of all nanopillars.

*

245	333	296	304	276	336	289	234	253	292
366	323	309	284	310	338	297	314	305	330
266	391	315	305	290	300	292	311	272	312
315	355	346	337	303	265	278	276	373	271
308	276	364	390	298	290	308	221	274	343

Need for testing of hypothesis

Often the decisions are made based on samples estimates to generalize on population parameter (as described in sampling and estimation).

In this process, there may be a difference between the estimate and the parameter which needs to be examined.

The following possibilities might arise due to sampling

$$|\text{Estimate-Parameter}| = \begin{cases} 0 \\ \text{Small} \\ \text{Large} \end{cases}$$



Need for testing of hypothesis

Case(i):

If the difference is zero, it is called unbiased

Case(ii):

If the difference is small, it may due to chance or sampling error (improper sampling technique used leads to sampling error)

Case(iii):

If the difference is large, it may a real one or due to sampling error
(improper sampling technique used leads to sampling error)

Hence, there is a need to test what type of difference is between estimate and parameter.

Hypothesis

A statement which is yet to be proved/ established or a statement on the parameter(s) of the Probability distribution to be tested

Null Hypothesis

~~Hypothesis of no difference or neutral or may be due to Sampling variation~~

Alternative Hypothesis

~~Hypothesis of difference which is yet to be proved/ established~~

Hypothesis

Hypothesis testing (Non-statistical)



A suspected criminal is produced before jury.
The Jury has to decide whether the defendant
is innocent or guilty.



Jury must decide between two hypotheses

The null hypothesis



H_0 : The defendant may be innocent

The alternative hypothesis



H_1 : The defendant may be guilty

Hypothesis

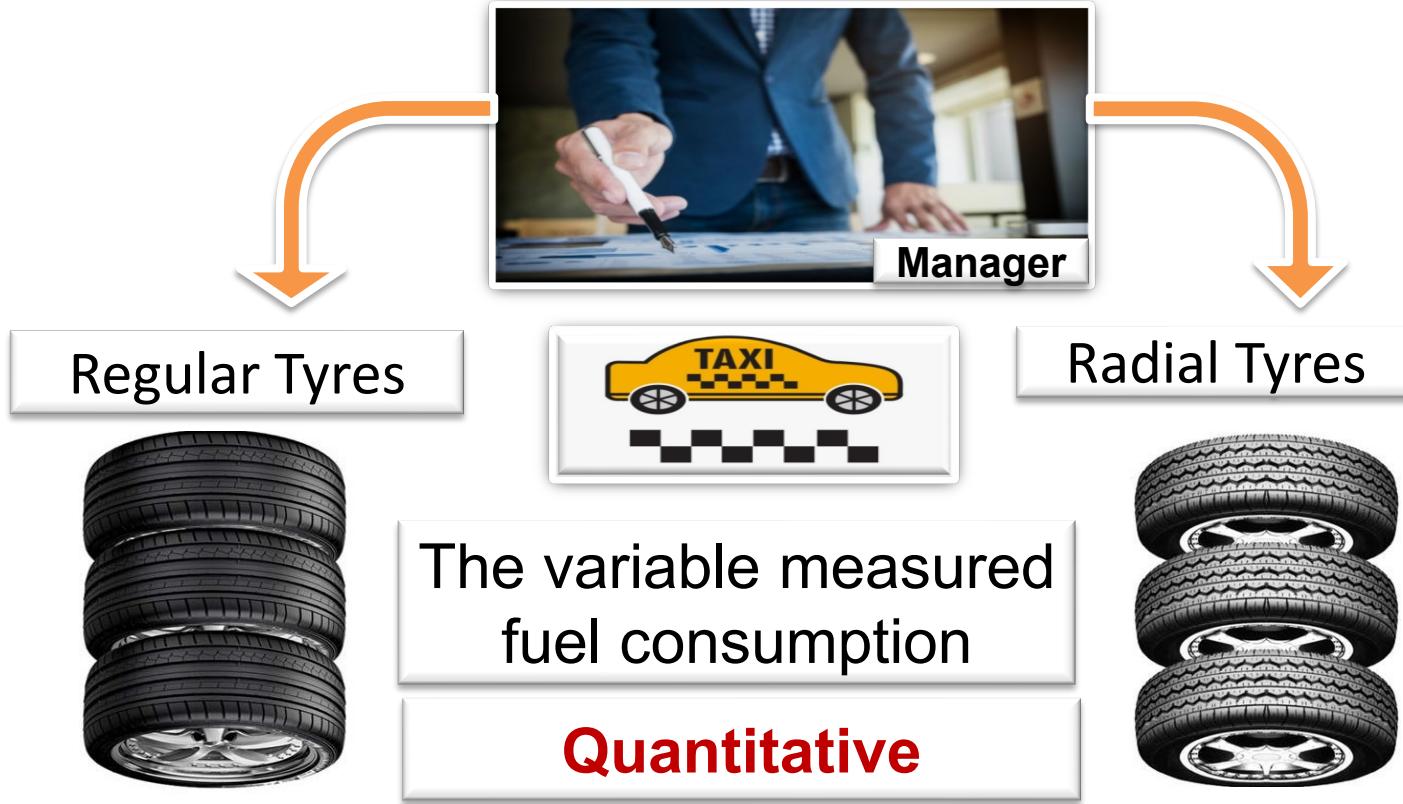
The jury do not know which hypothesis is true.

The jury should make a decision on the basis of evidence presented before them by the advocates .

Hypothesis - Formulation

- A taxi company manager is trying to decide whether the use of radial tires or regular belted tires improves fuel economy.
 - The variable measured is **quantitative**, therefore
-

Hypothesis - Formulation



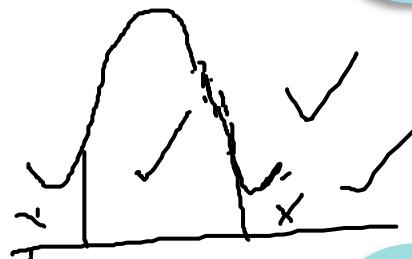
Hypothesis - Formulation

$$\mu = 40 \checkmark$$

$$\mu \neq 40$$

 H_0

The **mean** fuel consumption in cars fitted with radial tyres and regular belted tires will be same



$$H_0 : \mu_1 = \mu_2 \checkmark$$

Note: H_0 can also be stated as **one-tailed**

 H_1

The **mean** fuel consumption in cars fitted with radial tyres may be inferior to regular belted tires



$$H_1 : \mu_1 < \mu_2$$

4 \neq \checkmark



Hypothesis - Formulation

H_1



The **mean** fuel consumption in cars fitted with radial tires may be better than regular belted tires

$$H_1 : \mu_1 > \mu_2$$

H_1



The **mean** fuel consumption in cars fitted with radial tires and regular belted tires may be different

$$H_1 : \mu_1 \neq \mu_2$$

Hypothesis - Formulation

Two judges have to judge independently whether the defendant is innocent or guilty on the basis of evidence. Lack of sufficient evidence may lead to erroneous decisions like false positive or false negative. Suppose based on evidences, if we are interested in finding proportion of false positivity in the judgement, then the hypothesis to be tested, if variable measured is **qualitative** will be

Hypothesis - Formulation



Judge 1



Judge 2

Suppose based on evidences, if we are interested in finding **proportion of false positivity** in the judgment of two Judges

Formulate the hypotheses

???

Hypothesis - Formulation

H_0

The proportion of false positive judgement between Judges may be same

$$H_0 : P_1 = P_2$$

H_1

The proportion of false positive judgement by Judge 1 may be lower than proportion of false positive judgement by Judge 2

$$H_1 : P_1 < P_2$$

Hypothesis - Formulation

H_1



The proportion of false positive judgement by Judge 1 may be more than proportion of false positive judgement by Judge 2

$$H_1 : P_1 > P_2$$

H_1



The proportion of false positive judgement between both Judges may be different

$$H_1 : P_1 \neq P_2$$

Test

Test = $\begin{cases} \mu_1 < \mu_2 \Rightarrow \text{One-tailed test} \\ \mu_1 > \mu_2 \Rightarrow \text{One-tailed test} \\ \mu_1 \neq \mu_2 \Rightarrow \text{Two-tailed test} \end{cases}$

Test

Test is a statistical rule which decides whether to accept the null hypothesis or not ?

Warning

Decision is made based on the sample not on the population



Leads to possibility of **error** between the decision made and the reality

Types of test

A statistical rule which decides whether to accept or reject the null hypothesis on the basis of data

Parametric tests

Based on the assumption of some probability distribution

Non-parametric tests

Not based on any assumption of probability distribution

Parametric tests

It is assumed that the data do follow some probability distribution which is characterized by any parameters.

Large Sample Test

$n \geq 30$

Standard Normal Test

Z-Test

Small Sample Test

$n < 30$

Student's t-test

Unpaired t-Test

Paired t-Test

Analysis of Variance

ANOVA

Rm ANOVA

Non - Parametric tests

It is assumed that the data do not follow any probability distribution which is not characterized by any parameters.

Distribution - free tests

Chi - Square Test

Fisher's exact probabilities

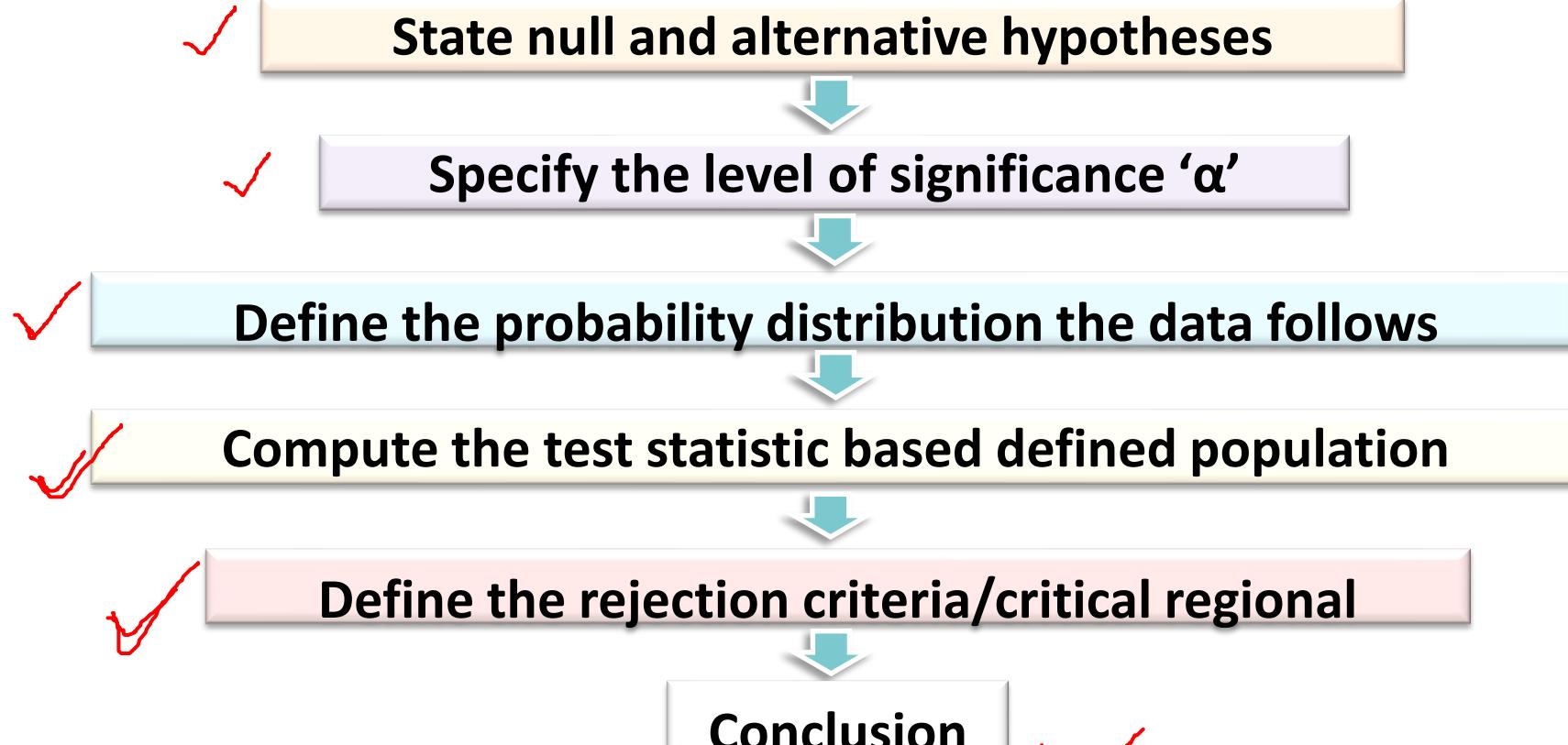
Mann – Whitney U test

Wilcoxon Signed Rank Test

Kruskal - WallisTest

Friedman'sTest

Steps involved in Testing of Hypothesis

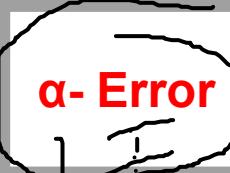


Errors in decision making

Any example based on data		
Statistical Example		
Decision	Null Hypothesis (H_0)	
	True	False
Accept		
Reject		

Any example based on data		
Statistical Example		
Decision	Null Hypothesis (H_0)	
	True	False
Accept		Type – II Error
Reject	Type – I Error	

Errors in decision making

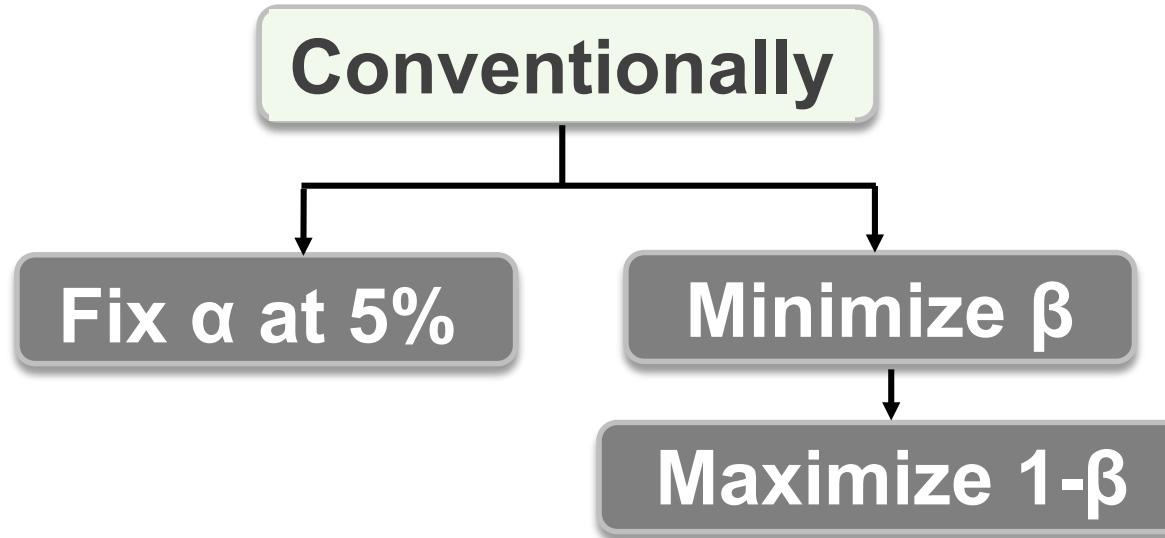
Any example based on data	
Statistical Example	
Decision	Null Hypothesis (H_0)
Accept	True  
Reject	 α - Error 

$\alpha = 0.5$

$(1 - \alpha)$

Any example based on data	
Statistical Example	
Decision	Null Hypothesis (H_0)
Accept	True  False
Reject	Confidence Level $(1-\alpha)$  Power $(1-\beta)$

Decision on α -error and β - error



Parametric tests

Z-test

This is a test based on Standard Normal Distribution

Used for testing the

Mean of a single population (μ)

Difference between means of two populations ($\mu_1 - \mu_2$)

Proportion of a single population (P)

Difference between proportions of two populations ($P_1 - P_2$)

Assumptions on Z-test

- Samples are drawn from normal distribution
- The population variances should be known

Z-test

- Two groups should be independent

- Subjects should be allocated randomly to both groups

- The sample size should be more than 30 (i.e., $n \geq 30$)

Testing mean of a single population

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 Standard Normal Distribution

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \cong N(0, 1)$$

4 Compute the test statistic

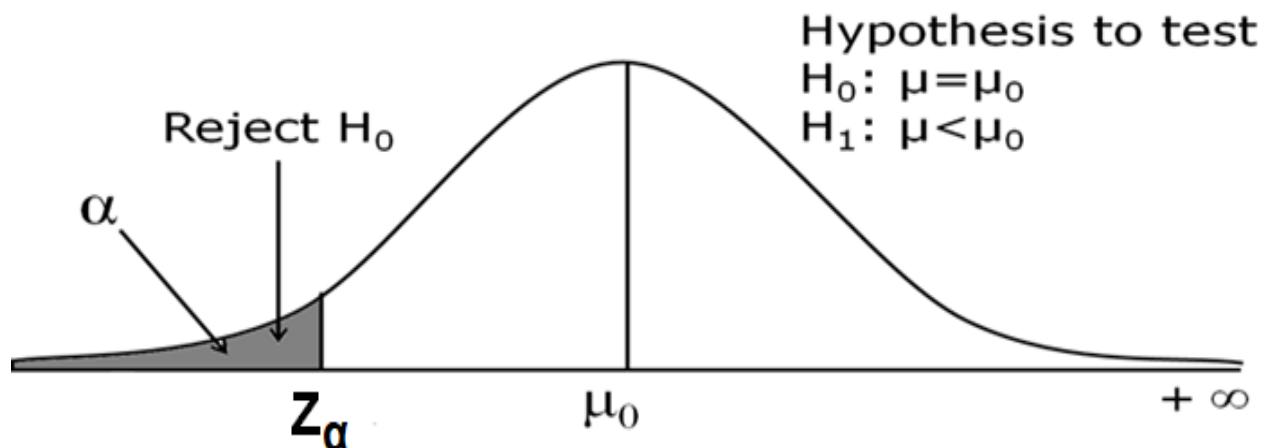
5 Define the critical region/ rejection criteria

6 Conclusion

Rejection criteria

5 Define the critical region/ rejection criteria

(i) Reject H_0 if computed value of Z is less than the critical value, ie., $P(Z < -z_\alpha)$, otherwise do not reject H_0



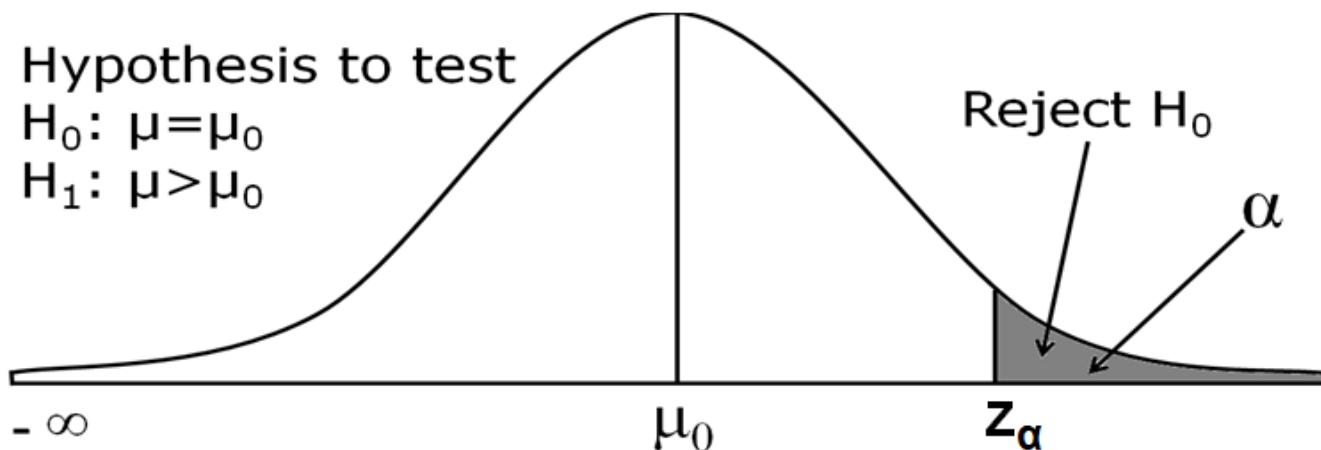
6 Conclusion

Rejection criteria

5 Define the critical region/ rejection criteria

(ii)

Reject H_0 if computed value of Z is greater than the critical value, ie., $P(Z > z_\alpha)$, otherwise do not reject H_0

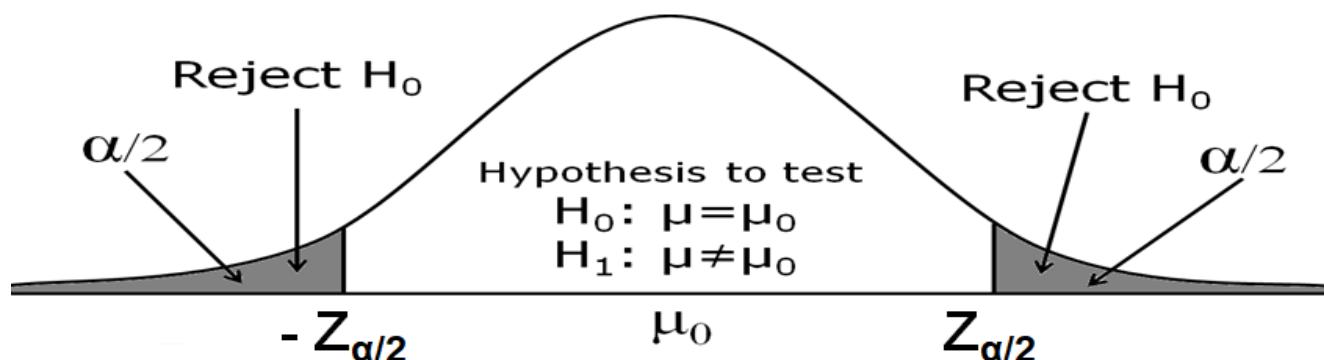


6 Conclusion

Rejection criteria

5 Define the critical region/ rejection criteria

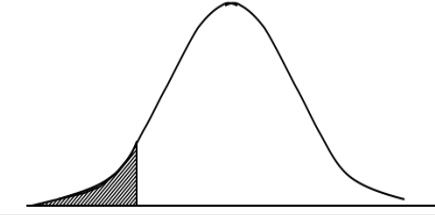
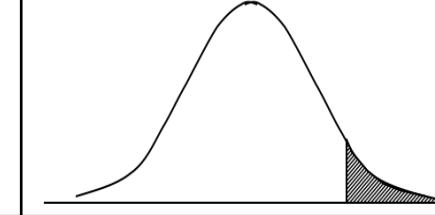
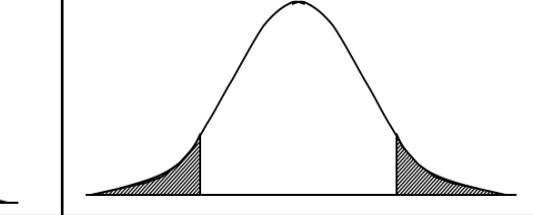
(iii) Reject H_0 if computed value of Z is less than or greater than the critical value, ie., $P(Z < - z_{\alpha/2})$ or $P(Z > z_{\alpha/2})$, otherwise do not reject H_0



6 Conclusion

Rejection criteria

Summary of One- and Two-Tail Tests

One-Tail Test (left tail)	One-Tail Test (right tail)	Two-Tail Test (Either left or right tail)
$H_0: \mu = \mu_0$ vs $H_1: \mu < \mu_0$	$H_0: \mu = \mu_0$ vs $H_1: \mu > \mu_0$	$H_0: \mu = \mu_0$ vs $H_1: \mu \neq \mu_0$
		

P - value

In hypothesis testing, the choice of the value of α is somewhat arbitrary. For the same data, if the test is based on two different values of α , the conclusion could be different. Many Statisticians prefer to compute the so called P-value, which is calculated based on the observed test statistic. For computing the P-value, it is not necessary to specify a value of α . We can use the given value data to obtain the P-value.

P – value: The strength of the evidence against the null hypothesis that the true difference in the population is zero

In other words

Corresponding to an observed value of a test statistic, the P-value (or attained level of significance) is the lowest level of significance at which the null hypothesis would have been rejected.

P-value

P - value



Possibility that the observed differences were a chance event



Entire population need to be studied to know that a difference is really present with certainty



Research community and statisticians had to pick a level of uncertainty at which they could live

P-value

If the P-value is less than 1% (< 0.01),

Overwhelming evidence that supports the alternative hypothesis

If the P-value is between 5% and 10%,

Weak evidence that supports the alternative hypothesis

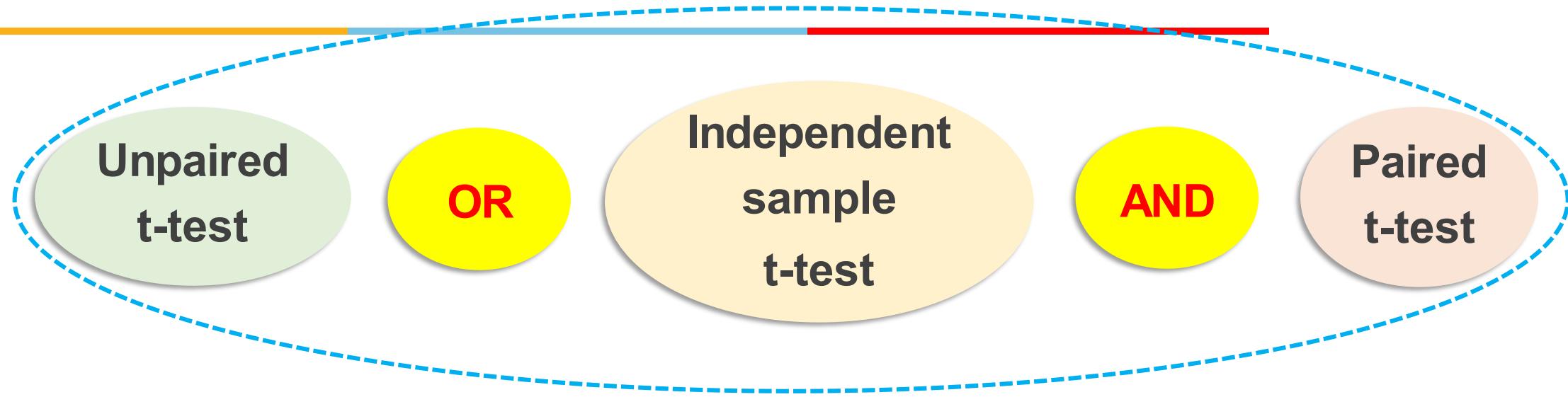
If the P-value is between 1% and 5%,

Strong evidence that supports the alternative hypothesis

If the P-value exceeds 10%,

No evidence that supports the alternative hypothesis.

Testing of Hypothesis → Student's t-test



Independent Sample t-test (Unpaired t-test)



Testing mean of a single population

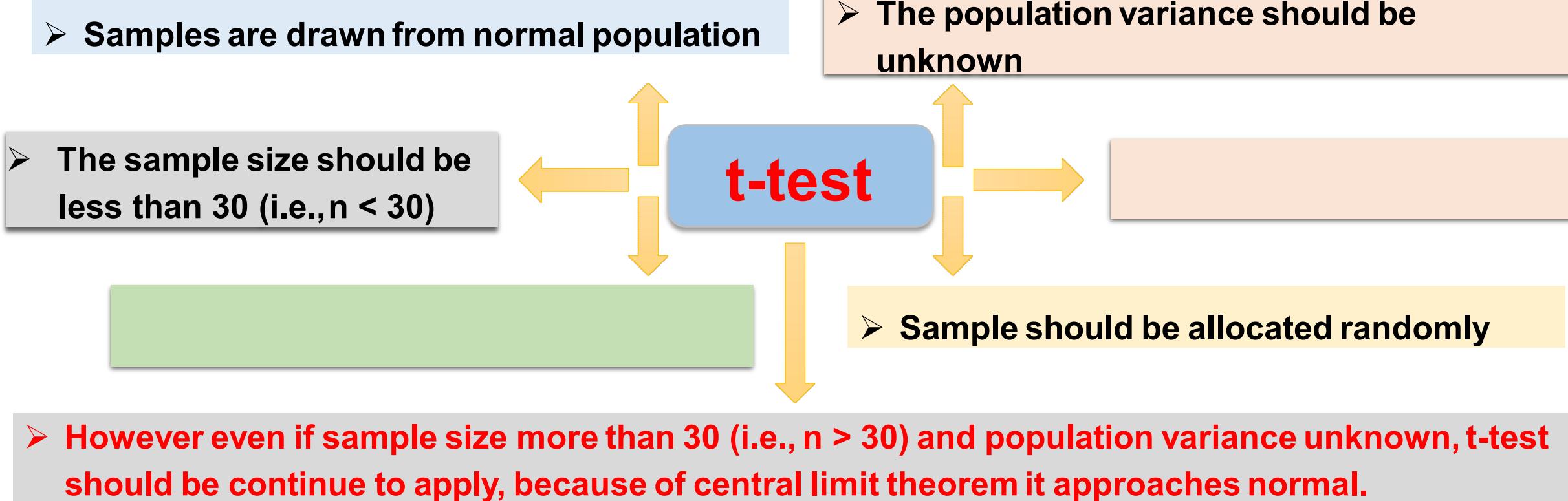


Testing difference between means of two populations

t-test



Testing mean of single population (μ)



t-test



Degrees of freedom (df): No. of independent observations

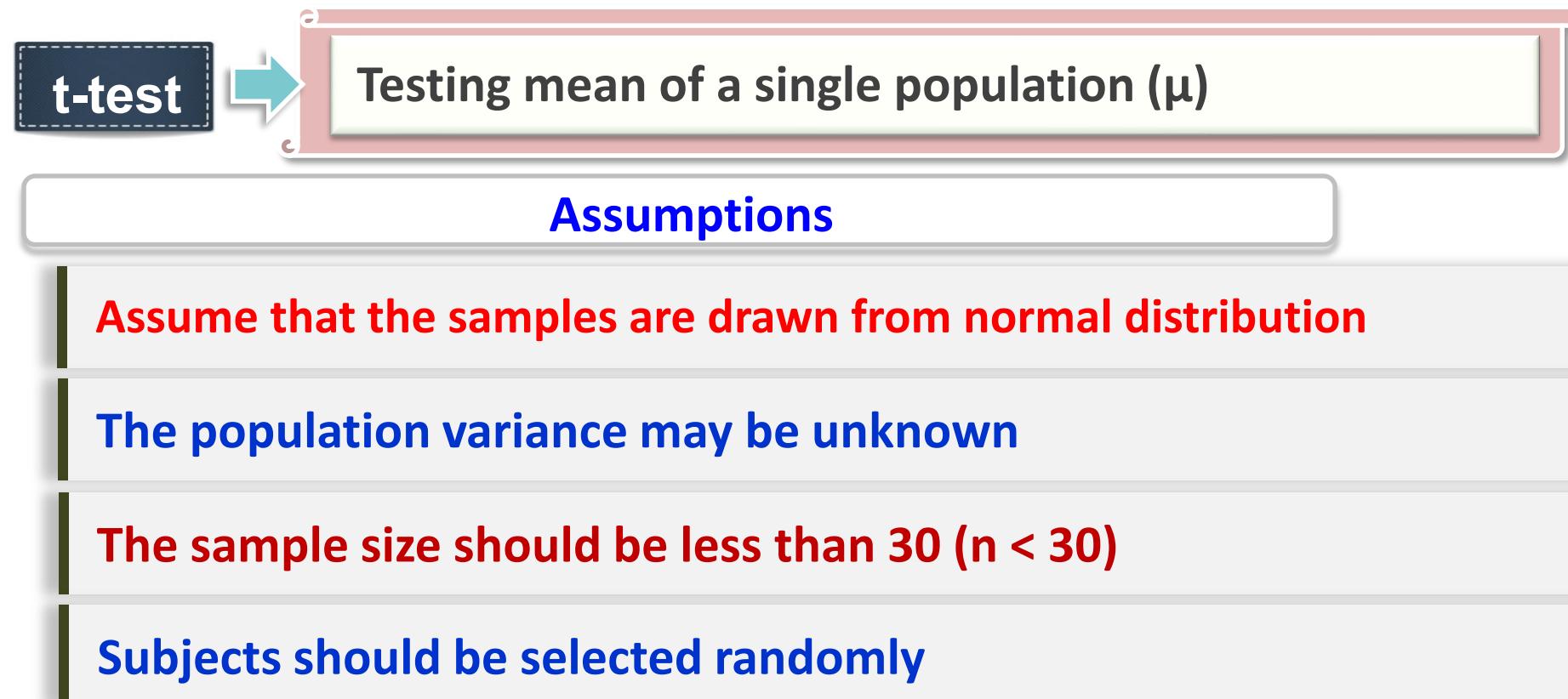
Suppose

$a+b = 20$. 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



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 α , 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

x

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 α , the degrees of freedom is also important.

Conclusion

Example 1

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

$$\begin{array}{ll} n \geq 30 & \rightarrow Z \\ < 30 & \rightarrow t \end{array}$$

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

$$|t| = \frac{|17352 - 18580|}{\sqrt{10}} = 1.929$$

95% CI for μ is

$$\bar{x} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}} = [16184.91, 18519.09]$$

P – value = 0.0428

Hypothesis to test

$$H_0: \mu = 18580 \text{ vs } H_1: \mu > 18580$$

$$\alpha = 0.05$$

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

EXAMPLE 2

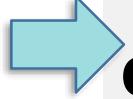
The management of a local health club claims that its members lose on the average 7 kgs or more within 3 months after joining the club. To check this claim, a consumer agency took a random sample of 15 members of this health club and found that they lost an average of 6.26 kgs within the first three months of membership. The sample standard deviation 1.91 kgs.

- Test at 1% level of significance whether the claim made by management of a local health club is acceptable or not?
- Also find the P-value of this test.

H_0 

The average weight loss as claimed by the health club management is 7

$$H_0 : \mu = \mu_0 \geq 7$$

 H_1 

The average weight loss as claimed by the health club management of 7 may be **higher than the sample average weight loss**

$$H_1 : \mu = \mu_0 < 7$$

At 1% (0.01) level of significance with critical value 2.624

$$|t| = \frac{|\bar{x} - \mu_0|}{\frac{s}{\sqrt{n}}} = \frac{|6.26 - 7|}{\frac{1.91}{\sqrt{15}}} = 1.501$$

P-value is

$$0.05 < P < 0.1$$

Hypothesis to test

$$H_0: \mu = \mu_0 \geq 7$$

vs

$$H_1: \mu = \mu_0 < 7$$

95% CI for μ

$$[5.203, 7.317]$$

includes $\mu_0 = 7$

**99% CI for μ is
(4.792, 7.728)**

Critical value for $\alpha = 0.01$ is 2.624 for df=14. $|t| = 1.501 < 2.624$, accept H_0 & Reject 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

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \cong t_{(\alpha, n_1+n_2-2)}$$

4 Compute the test statistic

5 Define the critical region/ rejection criteria

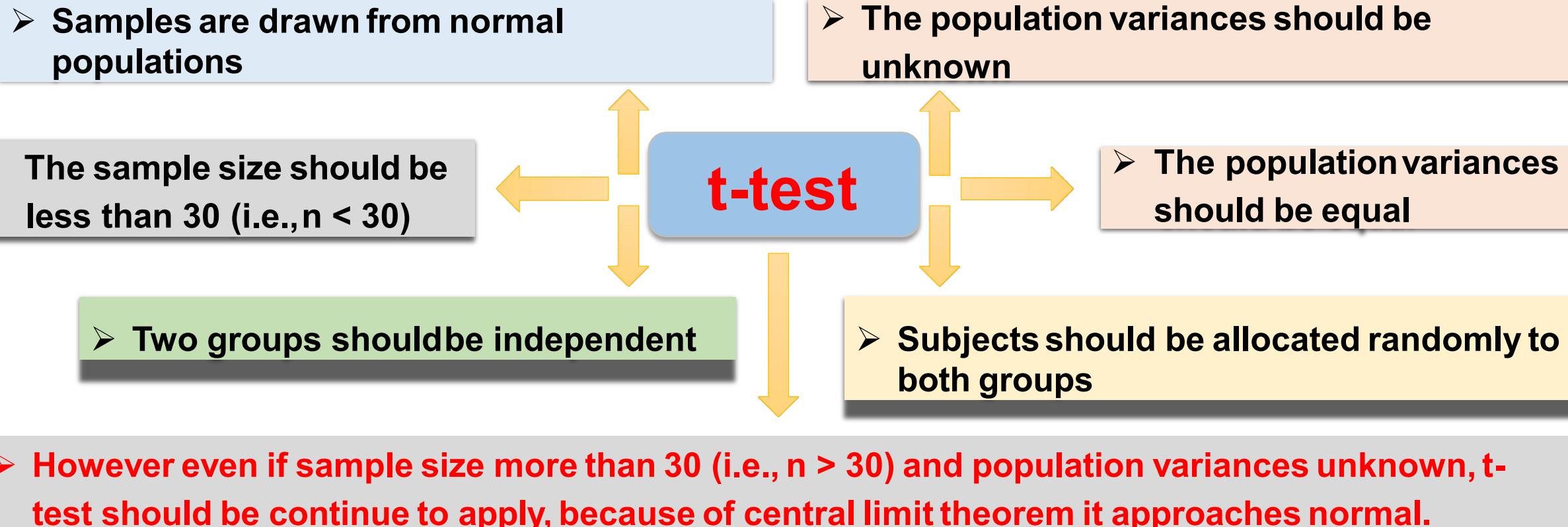
Note: If sample sizes are unequal compute
pooled SE

6 Conclusion

Note: Rejection criteria may be based on critical value or P-value

t-test

Difference between means of two populations ($\mu_1 - \mu_2$)

- 
- The diagram illustrates the requirements for a t-test. A central blue box labeled "t-test" is connected by double-headed orange arrows to six surrounding boxes, each containing a requirement. A single-headed orange arrow points downwards from the central box to a final box at the bottom.
- 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 3

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 μ is
[6.24, 17.76] not
includes 0

95% CI for μ 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

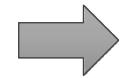
$$H_0 : \mu_1 = \mu_2$$

 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

$$(\bar{x}_1 - \bar{x}_2) \pm t_{\alpha/2} \text{SE}(\bar{x}_1 - \bar{x}_2)$$

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2} \text{SE}(\bar{x}_1 - \bar{x}_2) \leq (\mu_1 - \mu_2) \leq (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2} \text{SE}(\bar{x}_1 - \bar{x}_2)$$

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

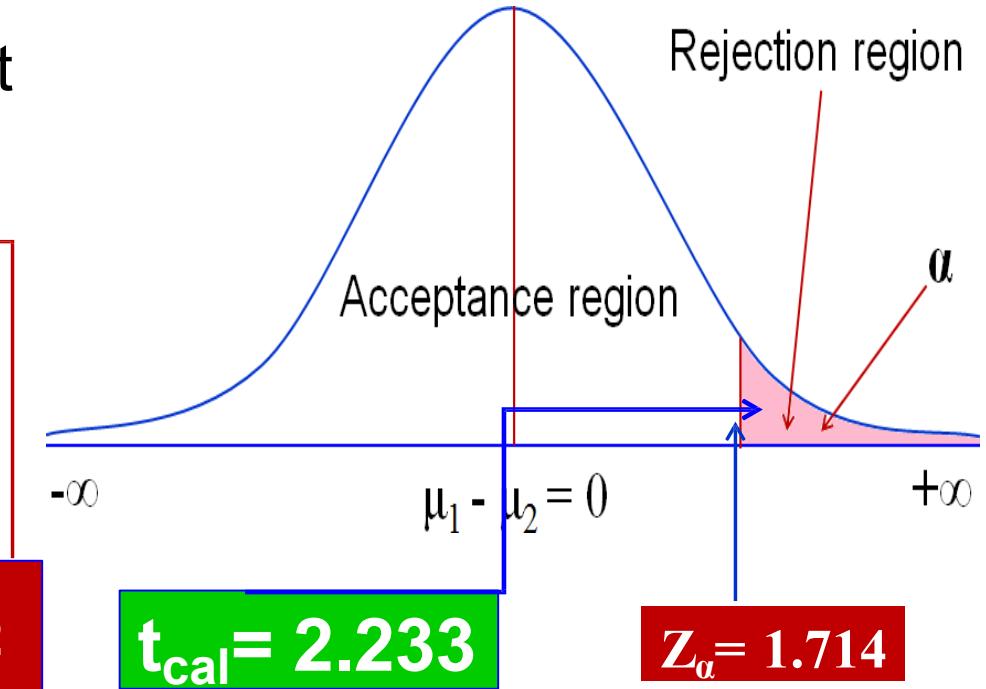
$$H_0: \mu_1 - \mu_2 = 0$$

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

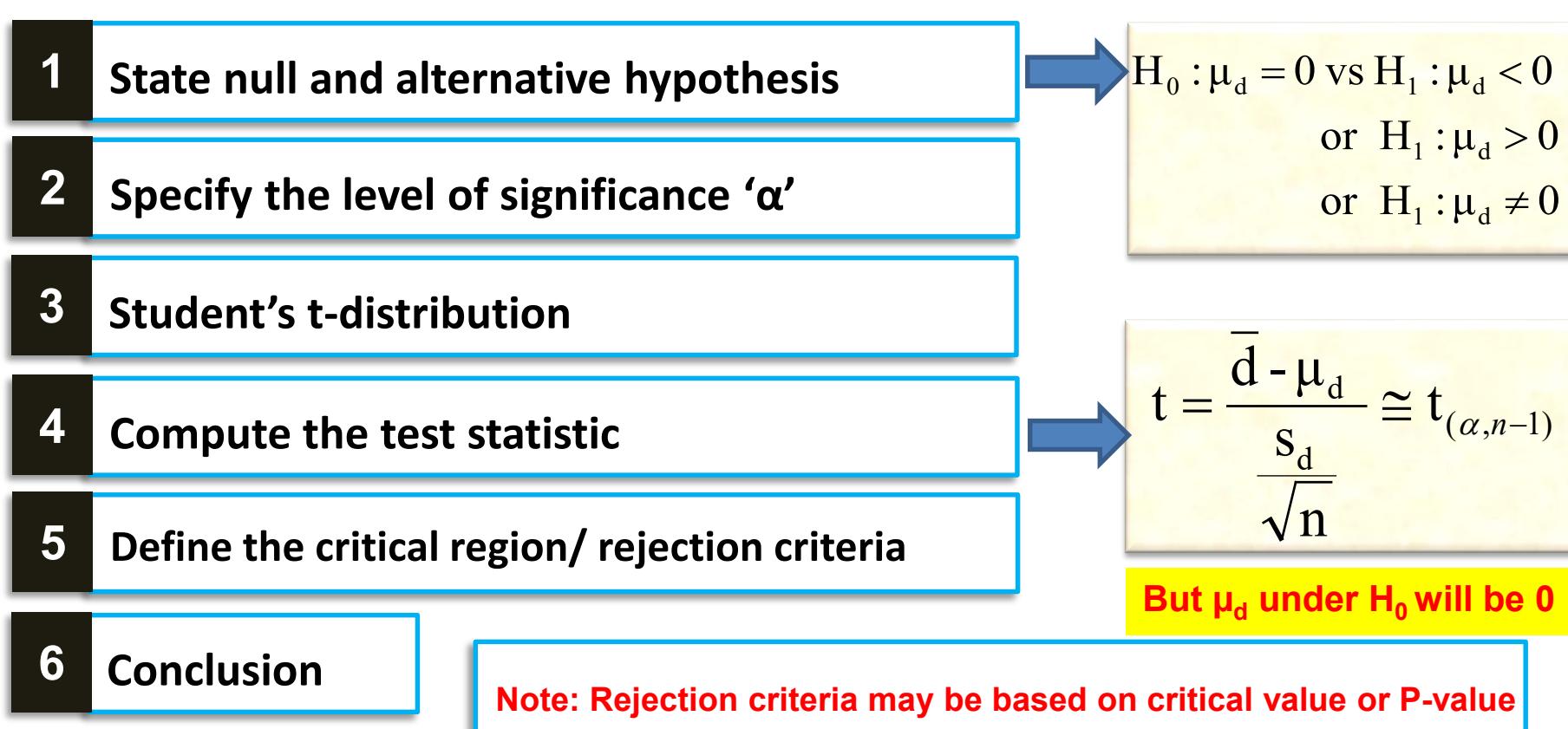


Testing mean before and after observations of a single population (μ_d)

Assumptions

Assume that the difference between before and after observations follow normal distribution

The sample size should be less than 30 ($n < 30$)



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
Total			45	678

$$\bar{d} = \frac{\sum_{i=1}^n d_i}{n} = \frac{45}{9} = 5$$

$$S_d = \sqrt{\frac{\sum_{i=1}^n (d_i - \bar{d})^2}{n-1}}$$

$$S_d = \sqrt{\frac{678}{8}} = 9.21$$

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 μ is
[-2.52, 12.52]
not includes 0**

**95% CI for μ 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



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.

Test 1

Test 2


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

Errors in Hypothesis Testing



Type I error calculation

α : denotes the probability of making a Type I error

$$\alpha = P(\text{Rejecting } H_0 | H_0 \text{ is true})$$

Type II error calculation

β : denotes the probability of making a Type II error

$$\beta = P(\text{Accepting } H_0 | H_0 \text{ is false})$$

Note:

α and β 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 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	
	Type I error	Decision is correct	

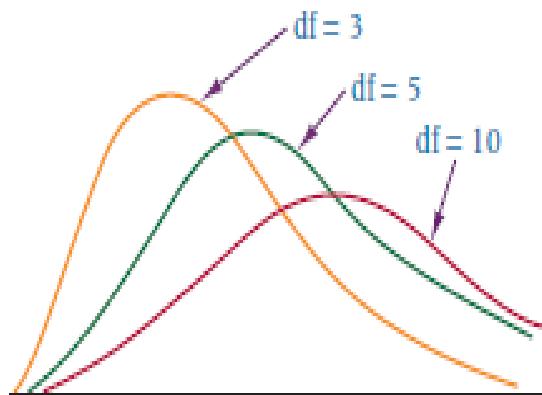
Chi-square Statistic

- 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.

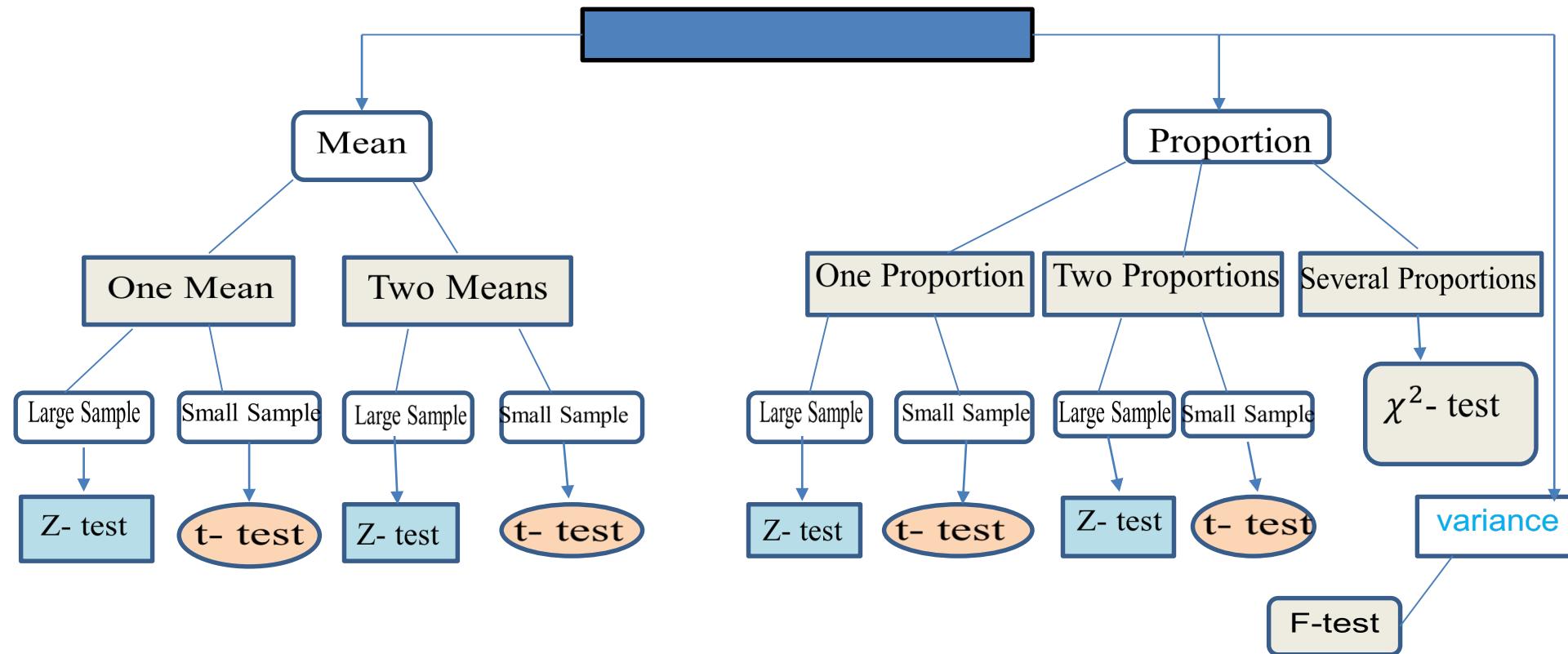
χ^2 FORMULA FOR SINGLE
VARIANCE (8.5)

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

$$df = n - 1$$



Three Chi-Square Distributions



Critical value	Level of significance α		
	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

The Edison Electric Institute has published figures on the annual number of kilowatt-hours expended by various home appliances. It is claimed that a vacuum cleaner expends an average of 46 kilowatt-hours per year. If a random sample of 20 homes included in a planned study indicates that vacuum cleaners expend an average of 42 kilowatt-hours per year with a standard deviation of 11.9 kilowatt-hours.

- Does this suggest at the 0.05 level of significance that vacuum cleaners expend, on average less than 46 kilowatt-hours annually?
- Assume population of kilowatt-hours to be normal.

PROBLEM

The manager of a courier service believes that packets delivered at the end of the month are heavier than those delivered early in the month.

As an experiment, he weighed a random sample of 10 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. The respective sample standard deviations are 1.20 kg and 1.15 kg.

- At 5% level of significance, can it be concluded that the packets delivered at the end of the month weigh more?
- Also find P- value and 95% confidence interval for the difference between the means.

Problem:

- . Suppose μ_1 and μ_2 are true mean stopping distances at 50 mph for cars of a certain type equipped with two different types of braking systems.
 - Use the two-sample t test at significance level .01 to test $H_0: \mu_1 - \mu_2 = -10$ versus $H_0: \mu_1 - \mu_2 < -10$ for the following data: $m=6, \bar{x} = 115.7, s_1 = 5.03, n = 6, \bar{y} = 129.3$ and $s_2 = 5.38$
-



Thanks



BITS Pilani

Pilani|Dubai|Goa|Hyderabad

Introduction to Statistical methods

ISM Team



Session 7

Testing of hypothesis

Tests based variances

Chi-square – test for single variance

Testing of Hypothesis

One sample Variance (χ^2 – test)

One sample F - test

χ^2 -test

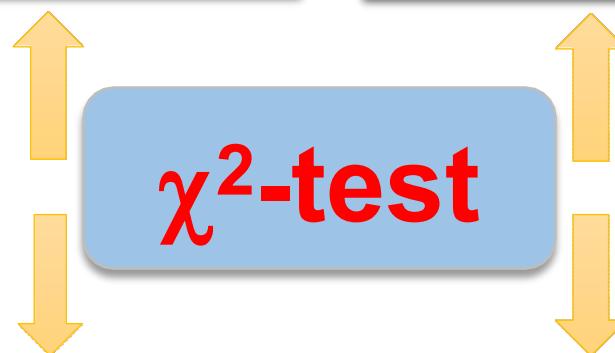


One sample variance distributed as
Chi-square distribution with $n-1$ degrees of freedom

➤ Samples are drawn from normal distribution

➤ The population variance should be known

χ^2 -test



➤ The sample size should be
less than 30 (i.e., $n < 30$)

➤ Subjects should be selected randomly



1

State null and alternative hypothesis



$H_0 : \sigma^2 = \sigma_0^2$ vs $H_1 : \sigma^2 < \sigma_0^2$
or $H_1 : \sigma^2 > \sigma_0^2$
or $H_1 : \sigma^2 \neq \sigma_0^2$

2

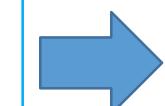
Specify the level of significance ‘ α

3

Chi-square - Distribution

4

Compute the test statistic



$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \cong \chi^2_{(\alpha, n-1)}$$

5

Define the critical region/ rejection criteria

6

Conclusion

Testing of Hypothesis → One sample Variance (χ^2 – test)

A manufacturer of car batteries claim that the life of his batteries is approximately normally distributed with a standard deviation equal to 0.9 year. If a random sample of 10 batteries has a standard deviation of 1.2 years, do you think that $\sigma > 0.9$ year? Use a 0.05 level of significance.

Testing of Hypothesis → One sample Variance (χ^2 – test)

At 5% (0.05) level of significance with critical value is 16.919 for 9 degrees of freedom

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} = \frac{9 \times 1.44}{0.81} = 16.00$$

100(1 - α)% CI for σ^2 is

$$\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}, n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}, n-1}}$$

Hypothesis to test

$$H_0: \sigma^2 = \sigma_0^2 = 0.81$$

vs

$$H_1: \sigma_1^2 > \sigma_2^2 > 0.81$$

???

Critical value for $\alpha = 0.05$ is 16.919. Since $\chi^2 = 16.00 < 16.919$, Accept H_0 & Reject H_1

Fisher's F – test for ratio of variances

Two sample F – test : Distributed as Fisher's F-distribution

F-test

Ratio of two population variances: $\sigma_{12}^2/\sigma_{22}^2$

➤ Samples are drawn from normal distribution

➤ The population variances should be equal

➤ The sample size should be less than 30 (i.e., $n < 30$)

➤ Two groups should be independent

F-test

➤ Subjects should be allocated randomly to both groups

Testing of Hypothesis → Ratio of two Variance (F – test)

1 State null and alternative hypothesis

$$H_0: \sigma_1^2 = \sigma_2^2 \text{ vs } H_1: \sigma_1^2 < \sigma_2^2$$

or $H_1: \sigma_1^2 > \sigma_2^2$
or $H_1: \sigma_1^2 \neq \sigma_2^2$

2 Specify the level of significance ‘ α ’

3 Fisher’s F - Distribution

4 Compute the test statistic

5 Define the critical region/ rejection criteria

6 Conclusion

$$F = \frac{S_1^2}{S_2^2} \cong F_{(\alpha, n_1-1, n_2-1)}$$

Testing of Hypothesis → Ratio of two Variance (F – test)

The variability in the amount of impurities present in a batch of chemicals used for a particular process depends on the length of time that the process is in operation.

Suppose a sample of size 25 is drawn from the normal process which is to be compared to a sample of a new process that has been developed to reduce the variability of impurities. Test at 5%, whether the variability in the new process is less as compared to the original process.

	Sample 1	Sample 2
n	25	25
S^2	1.04	0.51

Testing of Hypothesis → Ratio of two Variance (F – test)

At 5% (0.05) level of significance with critical value is 1.98 for (24, 24) degrees of freedom

$$F = \frac{S_1^2}{S_2^2} = \frac{1.04}{0.51} = 2.04$$

100(1 – α)% CI for $\frac{\sigma_1^2}{\sigma_2^2}$ is

$$\frac{S_1^2}{S_2^2} f_{1-\frac{\alpha}{2}, n_2-1, n_1-1} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S_1^2}{S_2^2} f_{\frac{\alpha}{2}, n_2-1, n_1-1}$$

Hypothesis to test

$$\begin{aligned} H_0: \sigma_1^2 &= \sigma_2^2 \\ \text{vs} \\ H_1: \sigma_1^2 &> \sigma_2^2 \end{aligned}$$

???

Critical value for α = 0.05 is 1.98. Since F = 2.04 > 1.98, Reject H_0 & Accept H_1

Testing of Hypothesis → **Ratio of two Variance (F – test)**

A company manufactures impellers for use in jet-turbine engines. One of the operations involves grinding a particular surface finish of a titanium alloy component. Two different grinding processes can be used and both processes can produce parts at identical mean surface roughness. The manufacturing engineer would like to select the process having the least variability in surface roughness. A random sample of $n_1 = 12$ parts from the first process results in a sample standard deviation of $s_1 = 5.1$ microinches of $n_2 = 15$ parts from the second process results in sample standard deviation of $s_2 = 4.7$ microinches. Test at 5%, is there a sufficient evidence that the first process vary more than the second process?

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

Assumptions of Chi-square test

If the expected cell frequencies is < 5

Yate's correction should be applied for continuity

In a **2 x 2 contingency table**, if one or more of the cell has the expected cell frequencies is < 5 ,

Fisher's exact probabilities should be computed

For the use of **Chi-square test**

The sample size should not be less than 20.

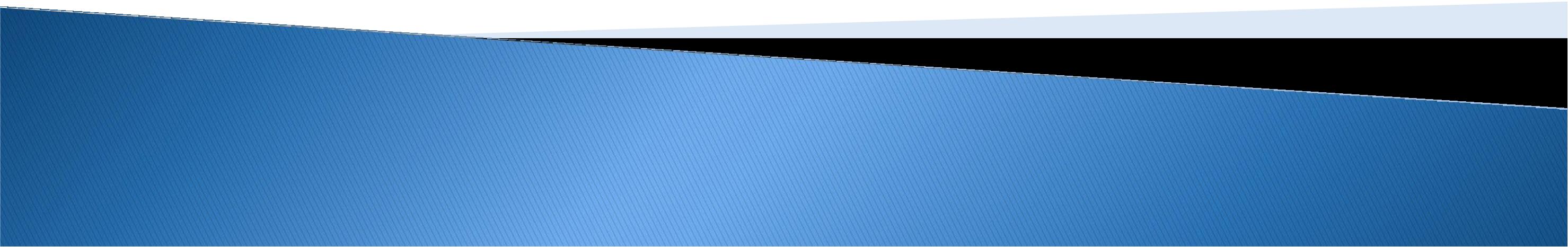
The Fisher's exact Probability



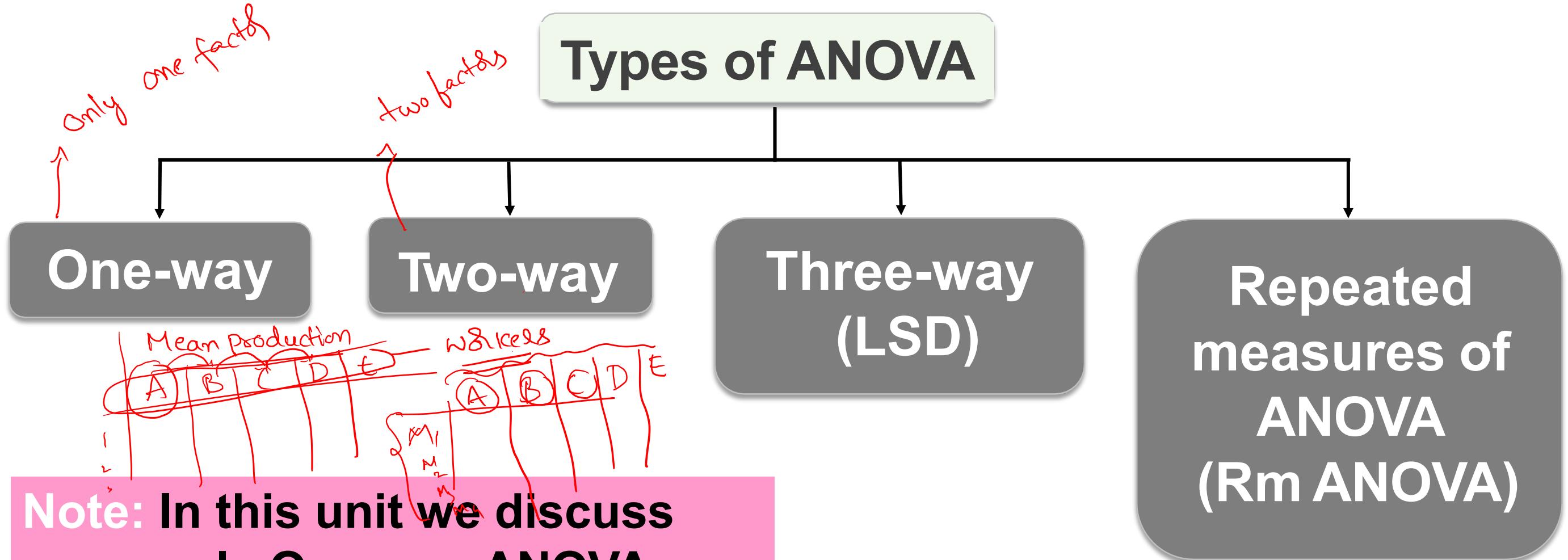
$$P = \frac{1}{n!} \frac{r_1!}{a!} \frac{r_2!}{b!} \frac{c_1!}{c!} \frac{c_2!}{d!}$$

For an $r \times c$ table, if the expected frequencies in any cells are < 5 , merge the rows or columns meaningfully

Analysis of Variance (ANOVA)



Testing of Hypothesis → Analysis of Variance (ANOVA)



Testing of Hypothesis → Why Analysis of Variance (ANOVA)

Student's t-test cannot be applied



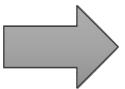
No. of groups are more than two (say k) and are independent



If t-test is applied, the type-I error will increase

ANOVA Used to test equality of more than
two population means against not equal

Q



ANOVA



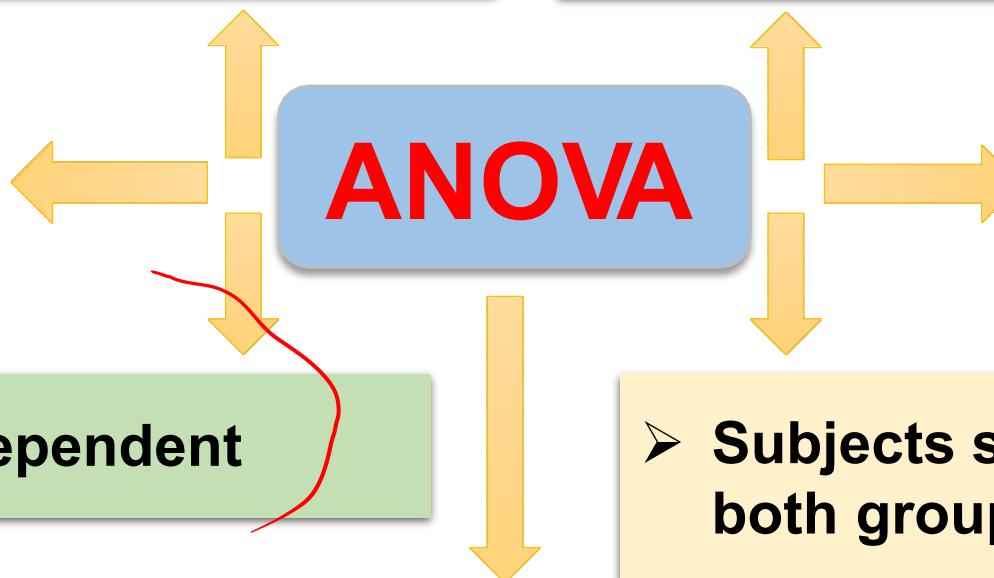
Testing equality of k group means against not equal

➤ Samples are drawn from normal population

➤ The population variances should be equal

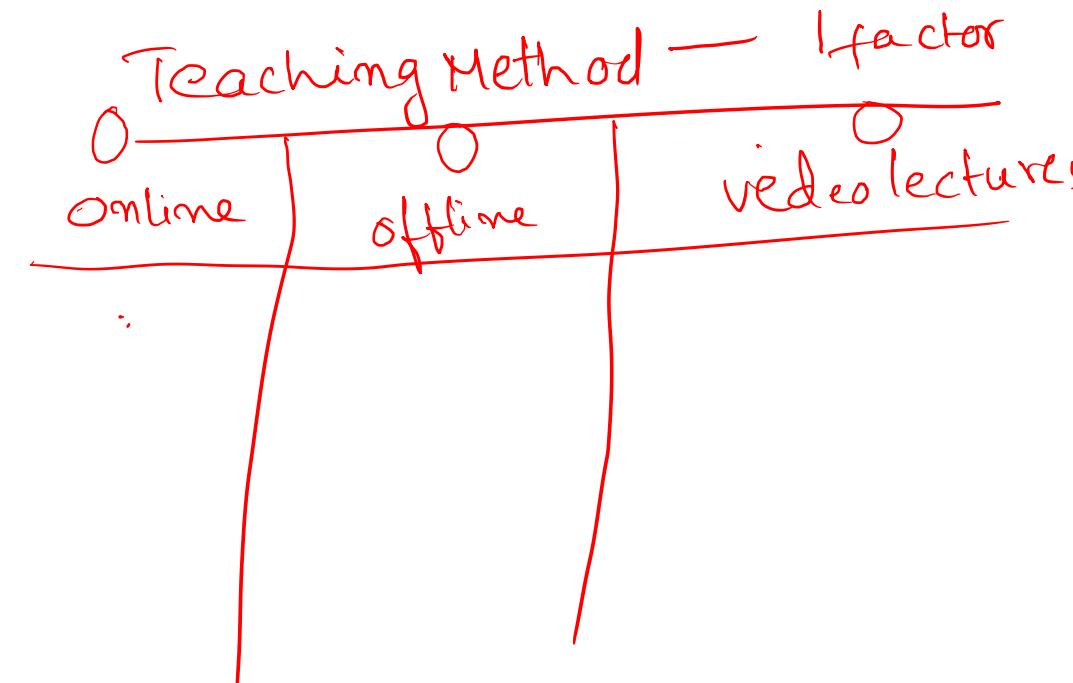
➤ The sample size should be less than 30 (i.e., $n < 30$)

➤ Groups should be independent

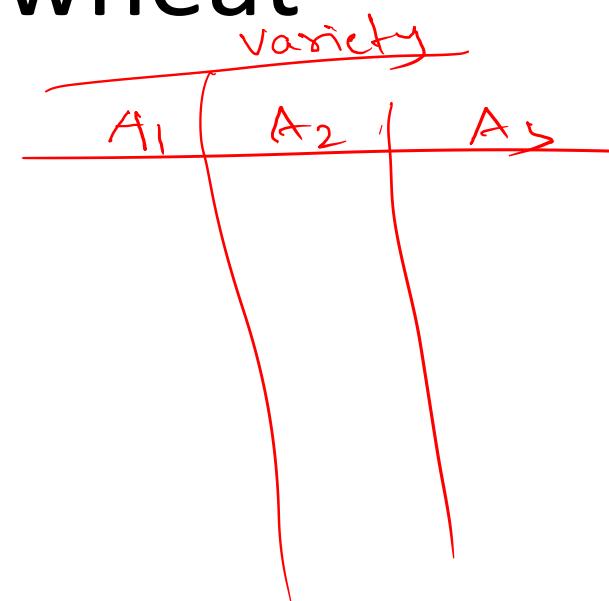


➤ However even if sample size more than 30 (i.e., $n > 30$) ANOVA should be continue to apply, because of central limit theorem it approaches normal.

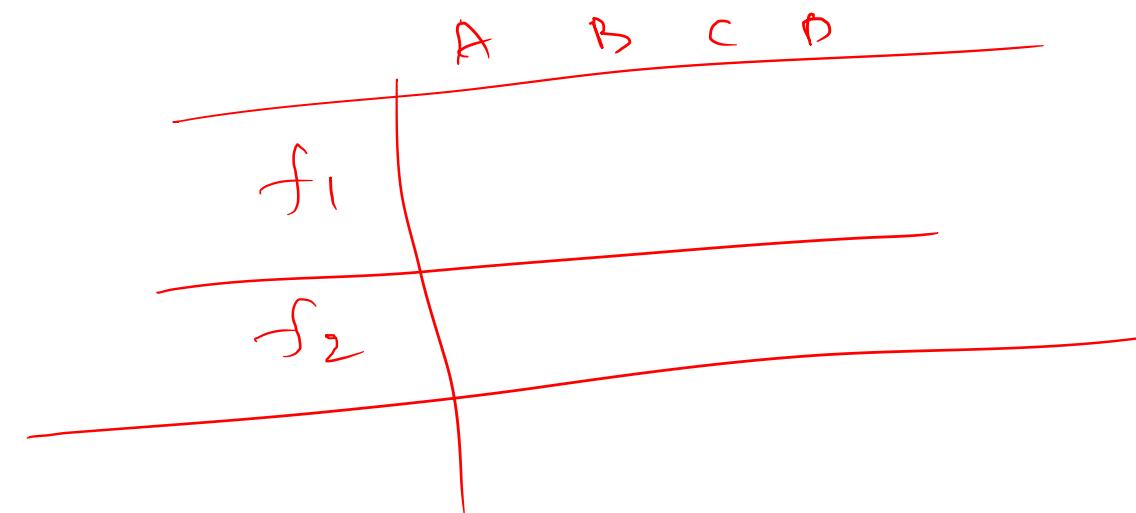
➤ Consider an experimental design of the teaching method to evaluate the strength of the teaching to the students via the online mode, offline mode and video lecture based mode.



- An agricultural experiment was conducted to compare yields of three varieties of seeds used to produce wheat



- An agricultural experiment was conducted to compare the yields of 4 varieties of rice applied by two types of fertilizers



G_1	G_2	G_3	.	G_k
X_{11}	X_{21}	X_{31}	.	X_{k1}
X_{12}	X_{22}	X_{32}	.	X_{k2}
X_{13}	X_{23}	X_{33}	.	X_{k3}
.
6	4	1	.	.
X_{1n_1}	X_{1n_2}	X_{1n_3}	.	X_{1n_k}
^{total} C_1	C_2	C_3	.	C_k

✓ $n_1 + n_2 + n_3 + \dots + n_k = n$

$C_1 + C_2 + C_3 + \dots + C_k = G$

The hypotheses to
be tested are

$H_0: \underline{\mu_1 = \mu_2 = \dots = \mu_k}$

VS

$H_1: \underline{\mu_1 \neq \mu_2 \neq \dots \neq \mu_k}$

G_1	G_2	G_3	.	G_k
X_{11}	X_{21}	X_{31}	.	X_{k1}
X_{12}	X_{22}	X_{32}	.	X_{k2}
X_{13}	X_{23}	X_{33}	.	X_{k3}
.
X_{1n_1}	X_{1n_2}	X_{1n_3}	.	X_{1n_k}
^{Total} C_1	C_2	C_3	.	C_k
^{Mean} \bar{x}_1	\bar{x}_2	\bar{x}_3	.	\bar{x}_k

Between groups }
With in groups }

Total observation

$$\frac{n_1 + n_2 + \dots + n_k = n}{G = c_1 + c_2 + c_3 + \dots + c_k}$$

Overall mean

$$= \frac{\sum x_{ij}}{n} = \frac{c_1 + c_2 + c_3 + \dots + c_k}{n}$$

$$= \frac{G}{n}$$

$$\underline{GSS} = \sum \frac{c_i^2}{n_i} - C \cdot F$$

$$= \frac{c_1^2}{n_1} + \frac{c_2^2}{n_2} + \dots + \frac{c_k^2}{n_k} - C \cdot F$$

$$C \cdot F =$$

Testing of Hypothesis

One-way Analysis of Variance

One-way ANOVA table

Source of variation	df	Sum of squares	Mean sum of squares	F-ratio
Between groups G_1, G_2, \dots, G_K	$k - 1$	<u>GSS</u>	$\frac{MGSS}{df} = \frac{GSS}{K-1}$	F
Within groups	$n - k$	WSS	$\frac{MWSS}{df} = \frac{WSS}{n-k}$	
Total	$n - 1$	TSS		

Testing of Hypothesis

One-way Analysis of Variance

1 State null and alternative hypothesis ✓

2 Specify the level of significance 'α' ✓

3

4 Compute the test statistic ✓

5 Define the critical region/ rejection criteria ✓

6 Conclusion

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

vs

$$H_1 : \mu_1 \neq \mu_2 \neq \dots \neq \mu_k$$

$$F = \frac{MGSS}{MWSS} \approx F_{(k-1, n-k)}$$

MGSS- Mean group sum of squares

MWSS- Mean within group sum of squares

Testing of Hypothesis → One-way Analysis of Variance

Calculation of sum of squares

1. Correction factor (CF) : $\frac{G^2}{n}$

$$GSS \quad CF = \frac{C_1 + C_2 + C_3 + \dots + C_K}{n}$$

WSS

TSS

$$TSS = GSS + WSS$$

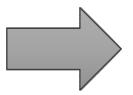
2. Total sum of squares (TSS) : $\sum_{i} \sum_{j} x_{ij}^2 - CF$

$$WSS = TSS - GSS$$

3. Between group sum of squares (GSS) : $\sum_{i=1}^k \frac{C_i^2}{n_i} - CF$

4. Within group sum of squares (WSS) : $TSS - GSS$

Testing of Hypothesis



One-way Analysis of Variance

One-way ANOVA table

Source of variation	df	Sum of squares	Mean sum of squares	F-ratio
Between groups	$k - 1$	GSS	$MGSS = \frac{GSS}{k - 1}$	$F = \frac{MGSS}{MWSS}$
Within groups	$n - k$	WSS	$MWSS = \frac{WSS}{n - k}$	
Total	$n - 1$	TSS	$F \approx F$ - distribution with $k - 1$ and $n - k$ df	

Problem 1

A completely randomised design experiment with 10 plots and 3 treatments gave the following result.

Plot No:	1	2	3	4	5	6	7	8	9	10
Treatment	A	B	C	A	C	C	A	B	A	B
Yield	5	4	3	7	5	1	3	4	1	7

Analyse the result for treatment effects

Solution:

Null hypothesis H_0 : There is no significant difference among the average yields in the 3 treatment.

Treatments	Yields from plots			
A	5	7	3	1
B	4	4	7	-
C	3	5	1	-

$$\checkmark \text{B/n groups} - GSS = \sum \frac{x_i^2}{n_i} - C.F$$

With Groups = TSS - GSS

~~$$\checkmark \text{Total} = \sum x_{ij}^2 - C.F$$~~

$$GSS = \frac{G^2}{n_1} + \frac{C_2^2}{n_2} + \frac{C_3^2}{n_3} - C.F$$

$$= \frac{(16)^2}{4} + \frac{(15)^2}{3} + \frac{9^2}{3} - 160$$

$$= 6$$

$$\checkmark C.F = \frac{G^2}{n} = \frac{\text{Grand total}}{\text{no. of observation}}$$

$$G = \underline{\underline{C_1 + C_2 + C_3 +}}$$

		Treatment A	Treatment B	Treatment C	
		$\sum X_1^2$	$\sum X_2^2$	$\sum X_3^2$	
\checkmark	X_1	5	4	3	
\checkmark		25	16	9	
\checkmark	7	49	16	25	
\checkmark	3	9	49	1	
\checkmark	1	1	-	-	0
$C_1 = 16$	$\sum X_1$	84	$C_2 = 15$	81	$C_3 = 9$
		$\sum X_1^2$	$\sum X_2$	$\sum X_3^2$	$\sum X_2^3$

$$\begin{aligned} TSS &= \sum x_{ij}^2 - C.F \\ &= 16^2 + 15^2 + 9^2 - 160 \\ &= 40 \end{aligned}$$

$$\begin{aligned} WSS &= TSS - GSS \\ &= 40 - 6 \\ &= 34 \end{aligned}$$

$$C.F = 160$$

- To find MGSS;

$$\checkmark G = \text{sum of all items} = \sum X_1 + \sum X_2 + \sum X_3 = 16 + 15 + 9 = 40$$

$$C.F = \text{correction factor} = G^2/n = (40)^2/10 = 160$$

$$G = 16 + 15 + 9 = 40$$

$$C.F = \frac{G^2}{n} = \frac{(40)^2}{10} = \frac{1600}{10} = 160$$

GSS = Sum of squares between samples

$$= \frac{\sum x_1^2}{n_1} + \frac{\sum x_2^2}{n_2} + \frac{\sum x_3^2}{n_3} - C \cdot F$$

$$= \frac{(16)^2}{4} + \frac{(15)^2}{3} + \frac{9^2}{3} - 160$$

$$= 64 + 75 + 27 - 160$$

$$= 6$$

* we can find MGSS (Mean squares b/w samples)

for finding MWSS (Mean squares within samples)

$$WSS = TSS - GSS$$

Total sum of squares - Sum of squares
b/w samples

$$TSS = \sum x_1^2 + \sum x_2^2 + \sum x_3^2 - C.F$$

$$= 84 + 81 + 35 - 160$$

$$= 40$$

$$\therefore WSS = 40 - 6 = 34$$

Source of variation	df	Sum of squares	Mean sum of squares	F-ratio
Groups (3)	$3-1 = 2$ K-1	GSS	$MGSS = \frac{6}{2} = 3$	$F = \frac{MGSS}{MWSS}$
Within groups	$10 - 3 = 7$ n - k	WSS	$MWSS = \frac{34}{7} = 4.86$	$= \frac{3}{4.86} = 0.617$
Total	n - 1	TSS		

$$F < F_\alpha$$

$$F(v_1, v_2) = F(2, 7) = \underline{\underline{4.617}}$$

Two way ANOVA

The following table gives monthly sales (in thousand rupees) of a certain film in three states by its four salesmen.

	Salesmen			
States	I	II	III	IV
A	6	5	3	8
B	8	9	6	5
C	10	7	8	7

Setup the analysis of variance table and test whether there is any significant difference (i) between sales by the film salesmen and (ii) between sales in the three states.

Solution :

Null Hypothesis

- ✓(1) H_0 : There is no significant difference between sales by four salesmen (columns)
- ✓(2) H_0 : There is no significant difference between in all 3 states. (rows)

Salesmen					
States	I	II	III	IV	Total ↓
A	6	5	3	8	22
B	8	9	6	5	28
C	10	7	8	7	32
Total →	24	21	17	20	82 (T)
\bar{X}_j	8	7	5.67	6.67	

Step 1: Grand Total $\boxed{T = 82}$

Step 2: Correction factor $C.F = \frac{T^2}{N} = \frac{(82)^2}{12} = \underline{\underline{560.33}}$

Step 3: \checkmark SSC: Sum of Squares b/n columns
(Salesmen)

$$\frac{\sum c_i^2}{n} = \frac{\sum x_1^2}{n} + \frac{\sum x_2^2}{n} + \frac{\sum x_3^2}{n} + \frac{\sum x_4^2}{n} - C.F$$

$$= \frac{24^2}{3} + \frac{(27)^2}{3} + \frac{(17)^2}{3} + \frac{(20)^2}{3} - 560.33$$

$$= 8.334 \checkmark$$

degrees of freedom $\nu_1 = K - 1$ (K-columns)
 $= 4 - 1$
 $= 3$

Step 4: $\underline{SSR} = \text{Sum of squares b/w rows}$ (states)

$$= \frac{1}{4} ((22)^2 + (28)^2 + (32)^2) - \underline{\underline{C.F}}$$

$$= \frac{1}{4} (2292) - 560 \cdot 333$$

$$= 12 \cdot 667$$

$d.f = \nu_2 = \infty - 1 = 3 - 1 = 2$

TSS = Total sum of squares

= Sum of squares of each values - C.F

$$= 6^2 + 8^2 + 10^2 + \dots + (\bar{5})^2 + (\bar{7})^2 - 560.33$$

$$= 602 - 560.333$$

$$= \underline{\underline{41.667}} \checkmark$$

* SSE = Residual

$$\begin{aligned}
 \cdot SSE &= \text{Residual} \\
 &= \underline{\text{Total sum of squares}} - \left(\underline{\text{SSC}} + \underline{\text{SSR}} \right) \\
 &= 41 \cdot 667 - 8 \cdot 334 - 12 \cdot 667 \\
 &= 20 \cdot 667^{\checkmark} \\
 \text{Degrees of freedom for Residual} \\
 &= \underline{\Sigma}_{\underline{g}} = (K-1)(\alpha-1) = 3 \times 2 = \boxed{6}
 \end{aligned}$$

$$F_C = \frac{r}{v_1 v_2} = \boxed{6,3}$$

Source of variation	df	Sum of squares	Mean sum of squares	F-ratio
Between columns	K-1	SSC	$\underline{\underline{MS}_C} = \frac{SSC}{d.f}$	$F_C = \frac{MS_E}{MS_C} = 0.81$
Between rows	r-1	SSR	$\underline{\underline{MS}_R} = \frac{SSR}{d.f}$	$F_R = \frac{MS_R}{MS_E} = 1.84$
Residual	<u>(r-1).(k-1)</u>	SSE	$\underline{\underline{MS}_E} = \frac{SSE}{d.f}$	

For $v_1 = 6$ $v_2 = 3$

$$F_{v_1, v_2} = 8.94 \text{ at } 5\% \text{ level}$$

of significance.

$$F_{\text{tab}} = 8.94$$

$$F_c < F_{\text{tab}}$$

\therefore There is no significant difference
in sales at 5% level of significance

For $v_1 = 2, v_2 = 6$ F_{f}

$$F_{\text{c}} = 5.14 \quad \text{at } \alpha = 0.05$$

$$F_{2,6}$$

$$F_c < F_{\text{tab}}$$

\therefore There is no significant
difference in states at
5% level of significance



BITS Pilani

Pilani|Dubai|Goa|Hyderabad

Introduction to Statistical Methods

Team ISM



Session – 12

Correlation & Regression

x₁ x₂

Relation

Prediction

x₁ x₂

relation ✓

In case of bivariate and multivariate populations, usually, we have to answer two types of questions:

Q1: Does there exist correlation (i.e., association) between two (or more) variables?

If yes, of **what degree?**

Q2: Is there any cause and effect relationship between the two variables (in case of bivariate population) or one variable in one side and two or more variables on the other side (in case of multivariate population)?

If yes, of **what degree and in which direction?**

To find solutions to the above questions, two approaches are known.

Correlation Analysis

Regression Analysis



Correlation Analysis

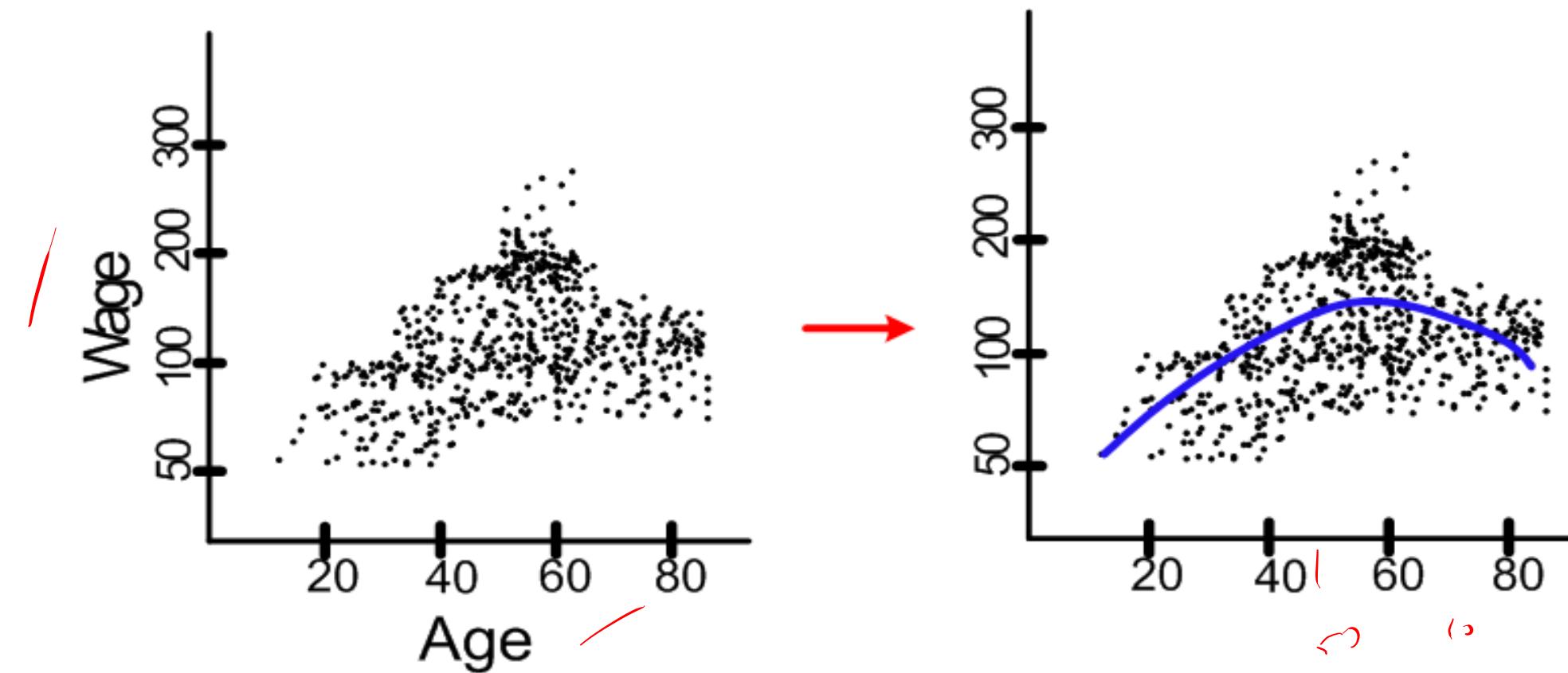
Example: Wage Data

A large data regarding the wages for a group of employees from the western region of a country is given.

In particular, we wish to understand the following relationships:

- *Employee's age and wage*: How wages vary with ages?
Explanatory Variable
- *Calendar year and wage*: How wages vary with time?
Relational Variable
- *Employee's age and education*: Whether wages are anyway related with employees' education levels?

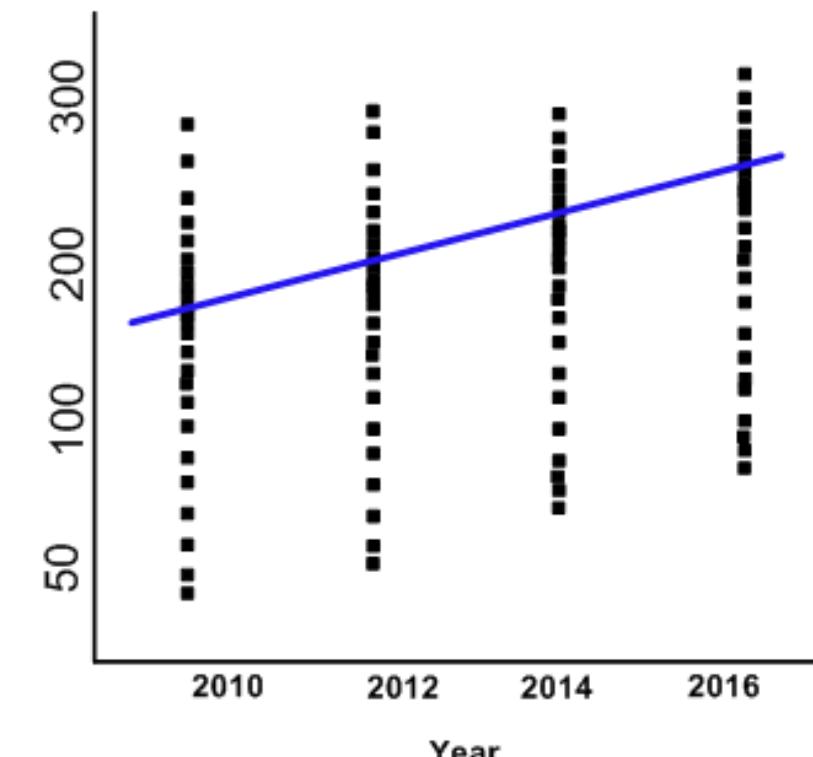
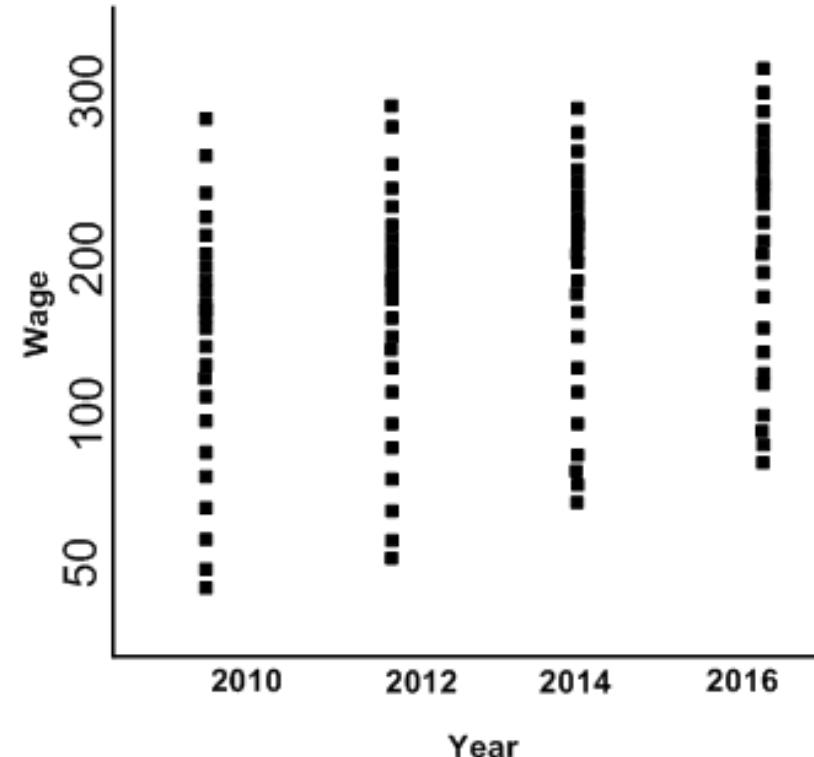
- Case I. Wage versus Age
- *Employee's age and wage:* How wages vary with ages?



Interpretation: On the average, wage increases with age until about 60 years of age, at which point it begins to decline.

➤ Case II. Wage versus Calendar

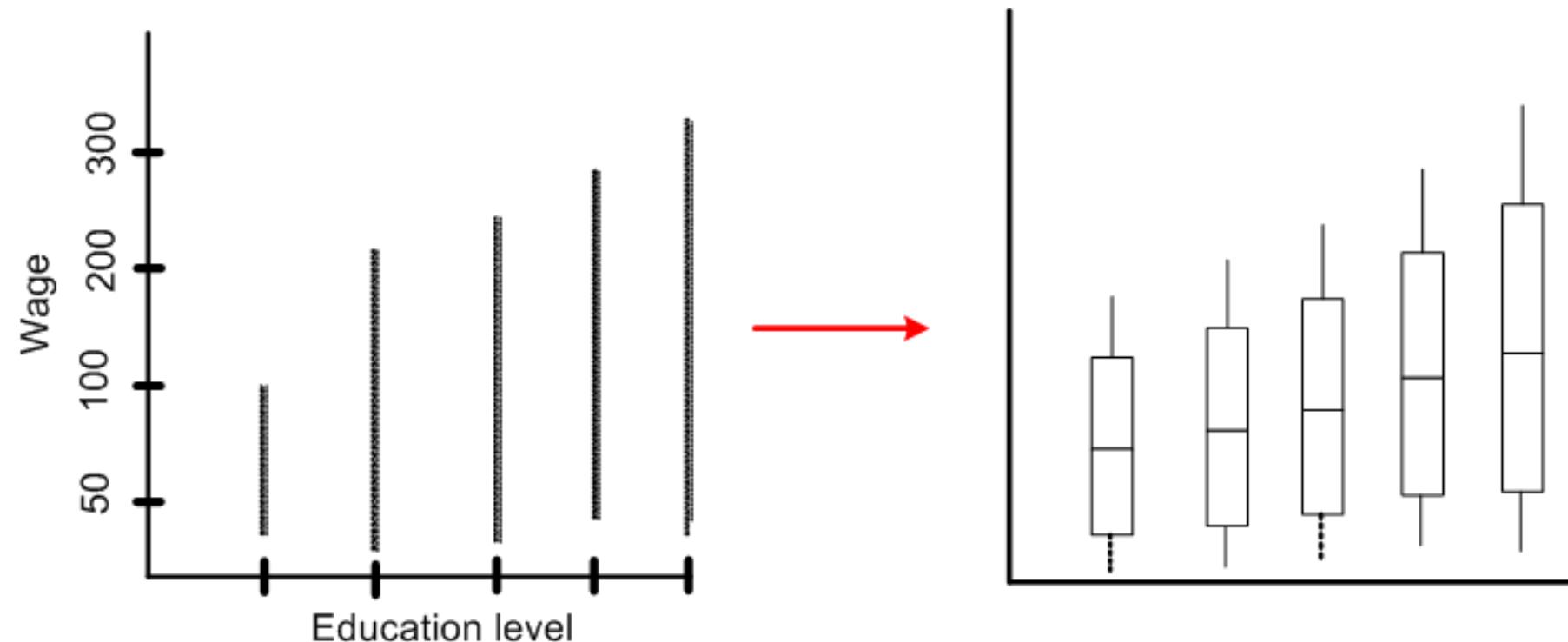
➤ *Wage and calendar year:* How wages vary with years?



Interpretation: ?

➤ Case III. Wage versus Age

Wage and education level: Whether wages vary with employees' education levels?



Interpretation: ?

Covariance ✓

The statistical technique used for studying the existence and extent of relationship among 2 or more variables is through Covariance

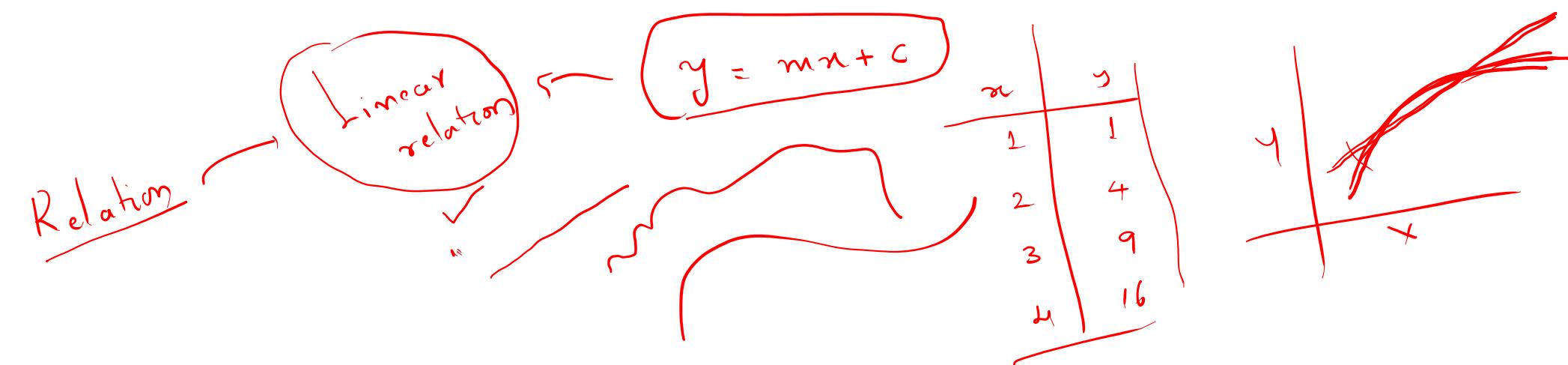
$$\text{Covariance}(X, Y) = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

Variance:

Mean +

Mean of

- So in bivariate and multivariate case, the statistical technique used for studying the existence of relationship between all variables is through **Covariance**.
- Covariance signifies the direction of the linear relationship between the two variables.
- By direction we mean if the *variables* are directly proportional or inversely proportional to each other. (Increasing the value of one variable might have a positive or a negative effect on the value of the other variable).



- The values of covariance can be any number between the two opposite infinities. Also, it's important to mention that covariance only measures how two variables change together, not the dependency of one variable on another one.
- The upper and lower limits for the covariance depend on the variances of the variables involved. These variances, in turn, can vary with the scaling of the variables. Even a change in the units of measurement can change the covariance. Thus, covariance is only useful to find the direction of the relationship between two variables and not the magnitude.

If X_1 and X_2 are two variables, the covariance between X_1 and X_2 is defined as

$$\text{Cov}(X_1, X_2) = \frac{\sum_{i=1}^n (X_1 - \bar{X}_1)(X_2 - \bar{X}_2)}{n-1}$$

If X_1, X_2, \dots, X_n are n variables, then the Covariance between these n variables is the Variance – Covariance matrix given by

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdot & \cdot & \cdot & \sigma_{12} \\ \sigma_{21} & \sigma_{22} & \cdot & \cdot & \cdot & \sigma_{2n} \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \sigma_{n1} & \sigma_{n2} & \cdot & \cdot & \cdot & \sigma_{n1} \end{bmatrix}$$

where $\sigma_{ij} = \begin{cases} \text{Variance, for } i = j \\ \text{Covariance, for } i \neq j \end{cases}$

PCA

Variance-Covariance Matrix

All diagonal elements in the variance-covariance matrix are variances and off diagonal elements are covariances.

Σ is a symmetric matrix.

- Even though the covariance explains the relationship between two or more variables, the units the measurement of variable is attached with the value of the covariance. If they have different unit of measurements, based on the sign it is possible to say whether they are related or not, but the degree (strength) of relationship is not possible to decide based on the value of covariance due to different units of measurement.

Example 1 :

Find the covariance between age (months) and weight (kgs) for the following data ✓

Age (years)	7	6	8	5	6	9
Weight (kgs)	12	8	12	10	11	13

Solution -1

innovate

achieve

lead

Age (yrs) (X)	Weight (kgs) (Y)	$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X})(Y - \bar{Y})$
5 ✓	20 ✓	-2	-3 ✓	6 ✓
4 ✓	18	-3	-5	15
7 L	23	0	0	0
10 -	32	3	9	27
8 -	26	1	3	3
8 ✓	25	1	2	2
$\sum X = 42$ ✓	$\sum Y = 144$			$\sum (X - \bar{X})(Y - \bar{Y}) = 53$

$$\text{Cov}(X, Y) = \frac{53}{5} = 10.6 \text{ years.kgs} \rightarrow \text{Interpret this answer !}$$

+ → positive
- → negative

$n-1$
 ↑↑ → +ve
 ↑↓ → -ve
 no relation

Covariance based on Probability distribution.

If X and Y are two random variables with probability mass function $p(x)$ / probability density function $f(x)$, then the covariance between X and Y

$$\begin{aligned} \text{Var}(x) &= E(x^2) - [E(x)]^2 \\ \text{Cov}(x, y) &= E((x-\bar{x})(y-\bar{y})) \\ &= \sum \sum (x-\bar{x})(y-\bar{y}) \cdot p(x, y) \end{aligned}$$

$$\underline{\text{Cov}(X, Y) = E[(X-\mu_X)(Y-\mu_Y)] = E(XY) - E(X)E(Y)}$$

where $E(X) = \mu_X$ and $E(Y) = \mu_Y$ are the means of X and Y.

➤ Covariance(x, x) = Var(x) ✓

➤ Covariance(x, y) = Covariance(y, x) ✓

$$\sum \frac{(x - \bar{x})(y - \bar{y})}{n-1}$$

$$\sum \frac{(y - \bar{y})(x - \bar{x})}{n-1}$$

Var-Cov matrix
Symmetric Matr

$$\begin{matrix} x & x_1 & x_2 & x_3 \\ p & p_1 & p_2 & p_3 \end{matrix}$$

$$E(x) = \sum x_i p_i$$

mean

Application of Covariance.

1. Multivariate linear regression:

- ❖ If there is a dependent variable and n independent variables, to find out the relationship between these (n+1) variables, a variance – covariance matrix will be helpful.

2. Time series:

- ❖ In the given series of data over the time, it helps in finding the autocorrelation at different time lags. For example, if $X_{t1}, X_{t2}, \dots, X_{tn}$ denote the time series data over a given period, then the autocovariance

ACF

Application of Covariance. at

lag h is given by

$$\gamma_X(t+h, t) = \text{Cov} (X_{\underline{t+h}}, X_{\underline{t}})$$

For example, if $X_{t_1}, X_{t_2}, \dots, X_{t_n}$ denote the time series data over a given period, then the auto-covariance.

The auto-covariance at lag h is then given by

$$\hat{\gamma}(h) = \frac{1}{n} \sum_{i=1}^{n-|h|} (X_{t+|h|} - \bar{X})(X_t - \bar{X}), -n < h < n.$$

Application of Covariance.

With the help of auto-covariance, the autocorrelation is given by

$$r_k = \frac{\sum_{t=k+1}^T (y_t - \bar{y})(y_{t-k} - \bar{y})}{\sum_{t=1}^T (y_t - \bar{y})^2}$$

Karl Pearson's Correlation coefficient

- ❖ To overcome this, Karl Pearson's suggestion was to divide the covariance of the variables by their respective standard deviations. The ratio of covariance to the product of standard deviations is called Karl Pearson's product moment correlation, which is free of unit of measurement.
- ❖ Karl Pearson's product moment correlation returns a numerical value, popularly known as **Coefficient of Correlation**, denoted by 'r'
- ❖ Correlation coefficient is a statistic that assesses the **strength** and **direction of relationship** of two continuous variables

Formula for computing Karl Pearson's correlation coefficient (r) is
standardized covariance (unit less)

$$r = \frac{\text{Covariance}(x, y)}{\sqrt{\text{Var}(x)} \sqrt{\text{Var}(y)}}$$

σ_x σ_y

-

$[-1, 1]$

$$r = \frac{\frac{\sum_{i=1}^n (X - \bar{X})(Y - \bar{Y})}{n - 1}}{\sqrt{\frac{\sum_{i=1}^n (X - \bar{X})^2}{n - 1}} \sqrt{\frac{\sum_{i=1}^n (Y - \bar{Y})^2}{n - 1}}} = \frac{\sum_{i=1}^n (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum_{i=1}^n (X - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y - \bar{Y})^2}}$$

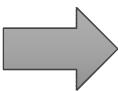
Karl Pearson's Correlation coefficient

Or **alternatively** we can use

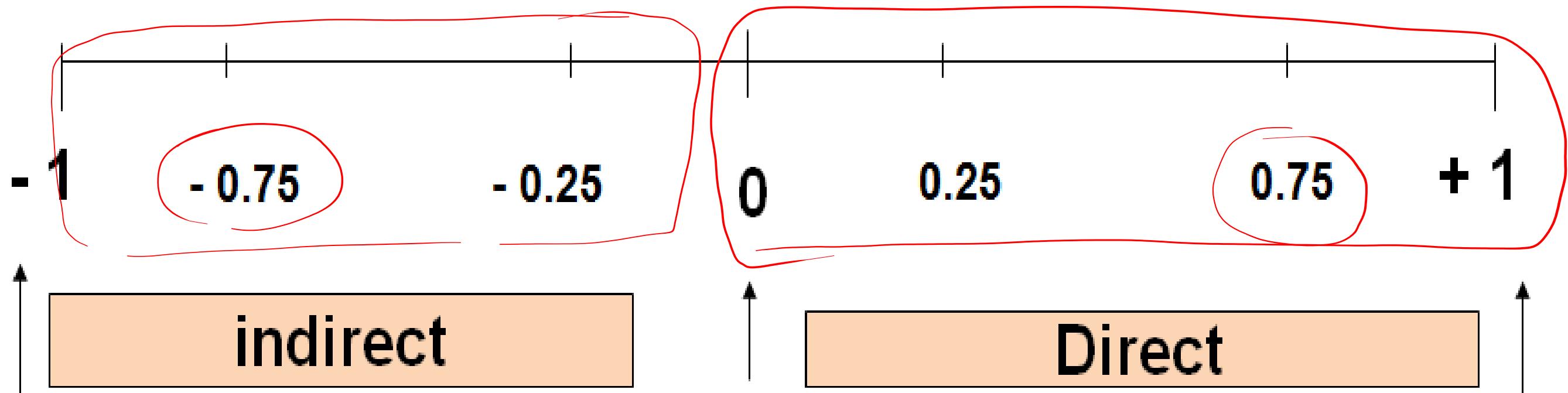
$$r = \frac{\sum xy}{\sqrt{\sum x^2} \sqrt{\sum y^2}}, \quad x = (X - \bar{X}), \quad y = (Y - \bar{Y})$$

Simple Correlation coefficient

- If the sign is +ve this means the relation is direct (an increase in one variable is associated with an increase in the other variable and a decrease in one variable is associated with a decrease in the other variable).
- If the sign is -ve this means an inverse or indirect relationship (an increase in one variable is associated with a decrease in the other).
- The value of r ranges between (-1) and (+1), i.e., $-1 \leq r \leq +1$
- The value of r denotes the strength of the association as illustrated by the following diagram.



Strong Intermediate Weak Weak Intermediate Strong



**Perfect -ve
correlation**

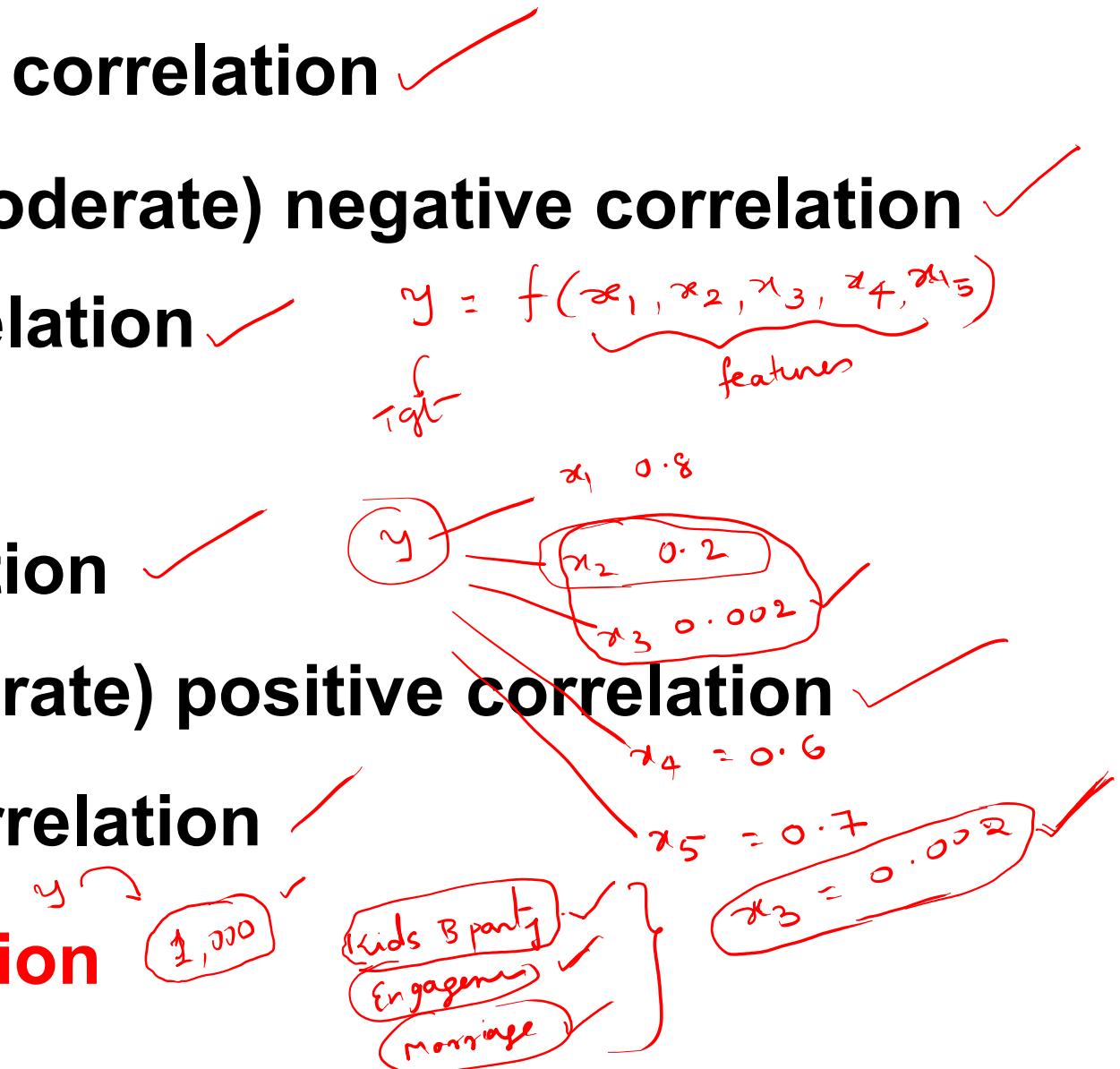


No relation

**Perfect +ve
correlation**

Simple Correlation coefficient

- $r = -1 \Rightarrow$ Perfect negative correlation
- -0.99 to -0.76 \Rightarrow Strong negative correlation ✓
- -0.75 to -0.26 \Rightarrow Intermediate (Moderate) negative correlation ✓
- -0.25 to 0 \Rightarrow Weak negative correlation ✓
- $r = 0 \Rightarrow$ Zero correlation
- 0 to 0.25 \Rightarrow Weak positive correlation ✓
- 0.26 to 0.75 \Rightarrow Intermediate (Moderate) positive correlation ✓
- 0.76 to 0.99 \Rightarrow Strong positive correlation
- $r = 1 \Rightarrow$ Perfect positive correlation



Example

Serial No	Age (yrs)	Weight (Kg)
1	7	12
2	6	8
3	8	12
4	5	10
5	6	11
6	9	13

$$\sum \frac{(x - \bar{x})(y - \bar{y})}{\sigma_x \sigma_y}$$

Solution

Sl no.	Age (yrs) (X)	Wt (Kg) (Y)	XY	X ²	Y ²
1	7	12	84	49	144
2	6	8	48	36	64
3	8	12	96	64	144
4	5	10	50	25	100
5	6	11	66	36	121
6	9	13	117	81	169
Total	$\sum X=41$	$\sum Y=66$	$\sum XY= 461$	$\sum X^2= 291$	$\sum Y^2= 742$

Karl Pearson's Correlation coefficient

$$r = \frac{n\sum XY - (\sum X)(\sum Y)}{\sqrt{n\sum X^2 - (\sum X)^2} \sqrt{n\sum Y^2 - (\sum Y)^2}}$$

$$r = \frac{6*461 - 41*66}{\sqrt{6*291 - (41)^2} \sqrt{6*742 - (66)^2}}$$

r = 0.759

(Positive (Direct) strong correlation)

Karl Pearson's Correlation coefficient

Example: Relationship between anxiety and test Scores

Anxiety (X)	Test score (Y)	X ²	Y ²	XY
10	2	100	4	20
8	3	64	9	24
2	9	4	81	18
1	7	1	49	7
5	6	25	36	30
6	5	36	25	30
$\Sigma X = 32$	$\Sigma Y = 32$	$\Sigma X^2 = 230$	$\Sigma Y^2 = 204$	$\Sigma XY = 129$

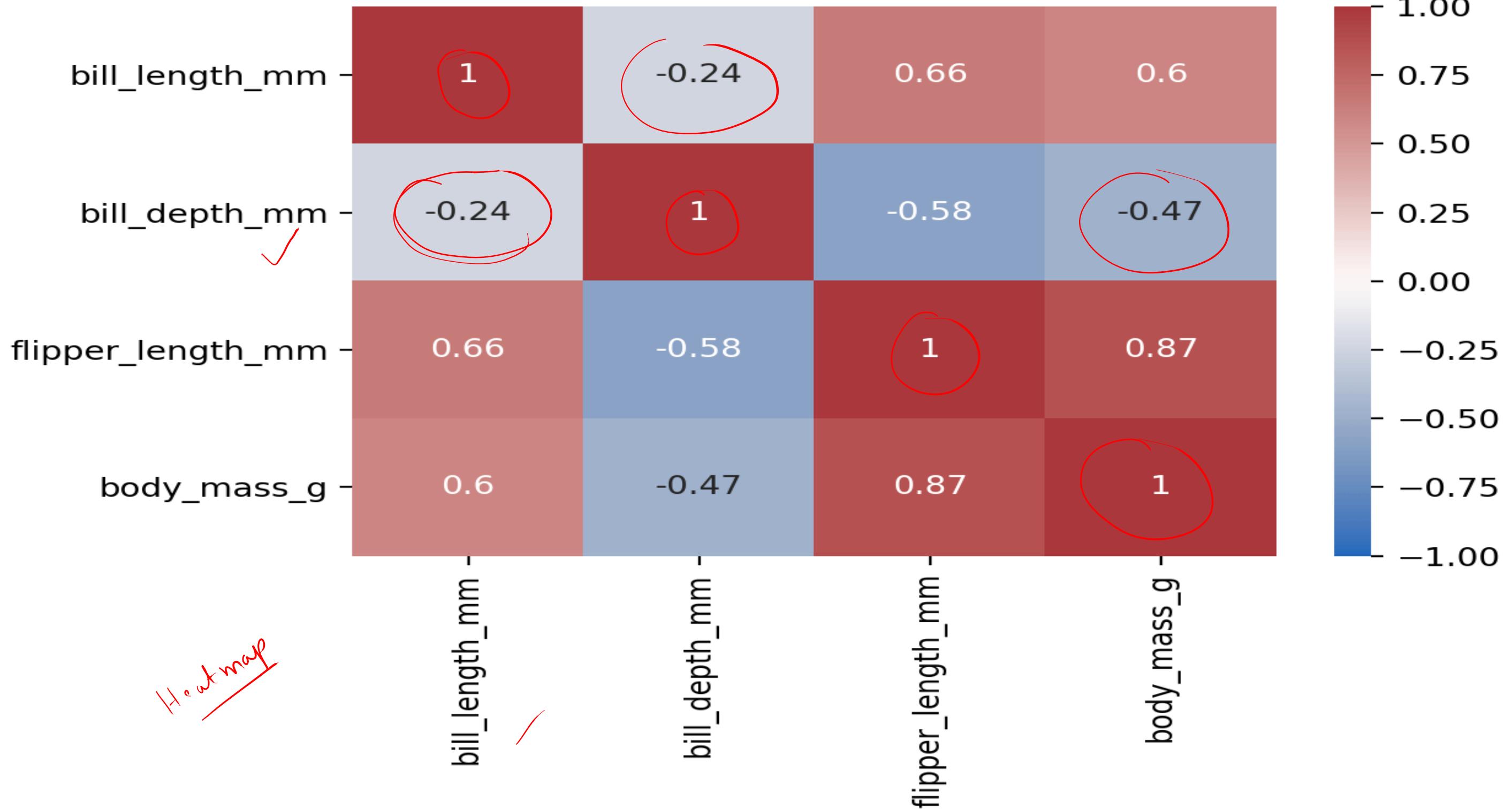
Correlational Analysis → Karl Pearson's Correlation coefficient

$$r = \frac{n \sum XY - (\sum X)(\sum Y)}{\sqrt{n \sum X^2 - (\sum X)^2} \sqrt{n \sum Y^2 - (\sum Y)^2}}$$

$$r = \frac{6*129 - 32*32}{\sqrt{(6*230 - (32)^2)(6*204 - (32)^2)}} = \frac{774 - 1024}{\sqrt{(356)(200)}} = -0.94$$

r = - 0.94 (Negative (Indirect) strong correlation)

A Correlation Matrix as a Heat Map



Example : A small study is conducted involving 17 infants to investigate the association between gestational age at birth, measured in weeks, and birth weight, measured in grams.

Infant ID #	Gestational Age (wks)	Birth Weight (gm)
1	34.7	1895
2	36.0	2030
3	29.3	1440
4	40.1	2835
5	35.7	3090
6	42.4	3827
7	40.3	3260
8	37.3	2690
9	40.9	3285
10	38.3	2920
11	38.5	3430
12	41.4	3657
13	39.7	3685
14	39.7	3345
15	41.1	3260
16	38.0	2680
17	38.7	2005

Solution :

In this example, birth weight is the dependent variable and gestational age is the independent variable. Thus Y = birth weight and X = gestational age.

For the given data, it can be shown the following

$$\bar{X} = \frac{\sum X}{n} = \frac{652.1}{17} = 38.4.$$

$$\bar{Y} = \frac{\sum Y}{n} = \frac{49,334}{17} = 2902.$$

$$s_x^2 = \frac{\sum (X - \bar{X})^2}{n-1} = \frac{159.45}{16} = 10.0.$$

$$s_y^2 = \frac{\sum (Y - \bar{Y})^2}{n-1} = \frac{7,767,660}{16} = 485,578.8.$$

$$r^* = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{s_x \cdot s_y} = 0.82$$

Conclusion: The sample's correlation coefficient indicates a **strong positive correlation** between Gestational Age and Birth Weight.

Homework Problems

innovate

achieve

lead

Q1) Consider the following data.

Expenditure (in 1000 Rs)	30	70	40	20	10	40	10	20
Actual Sales(in 1000 Rs)	10	16	19	14	17	16	13	18

For the above data,

- i) Find coefficient of correlation between expenditure and actual sales.
- ii) Find rank correlation between expenditure and actual sales.

[Ans = 0.8108]

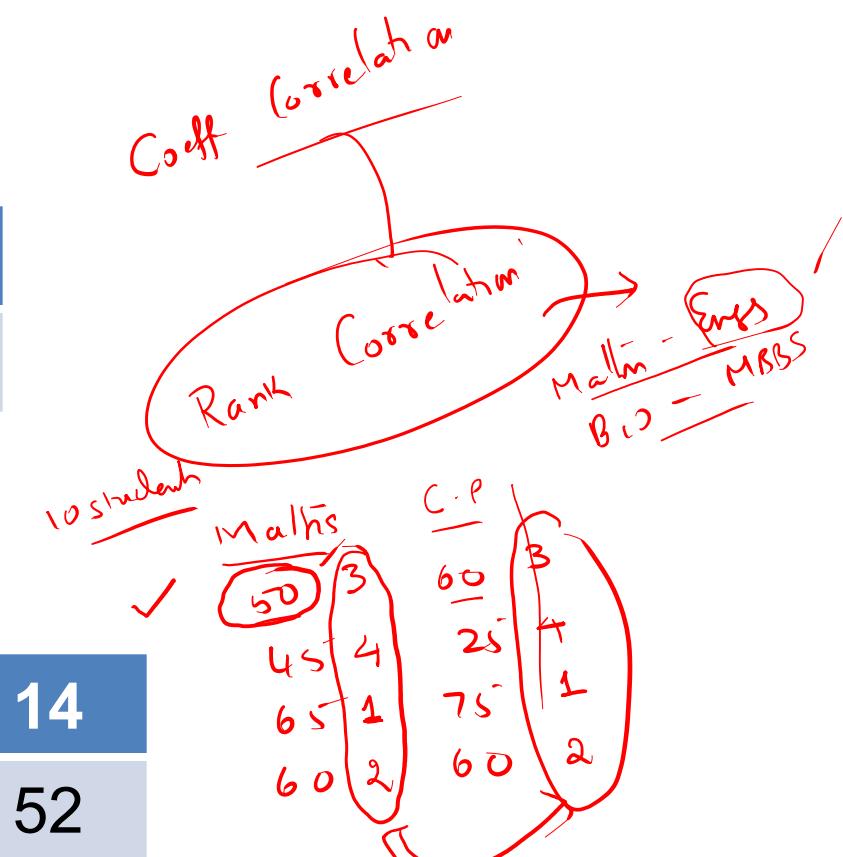
[Ans = -0.2424]

Q2) Find coefficient of correlation between X & Y.

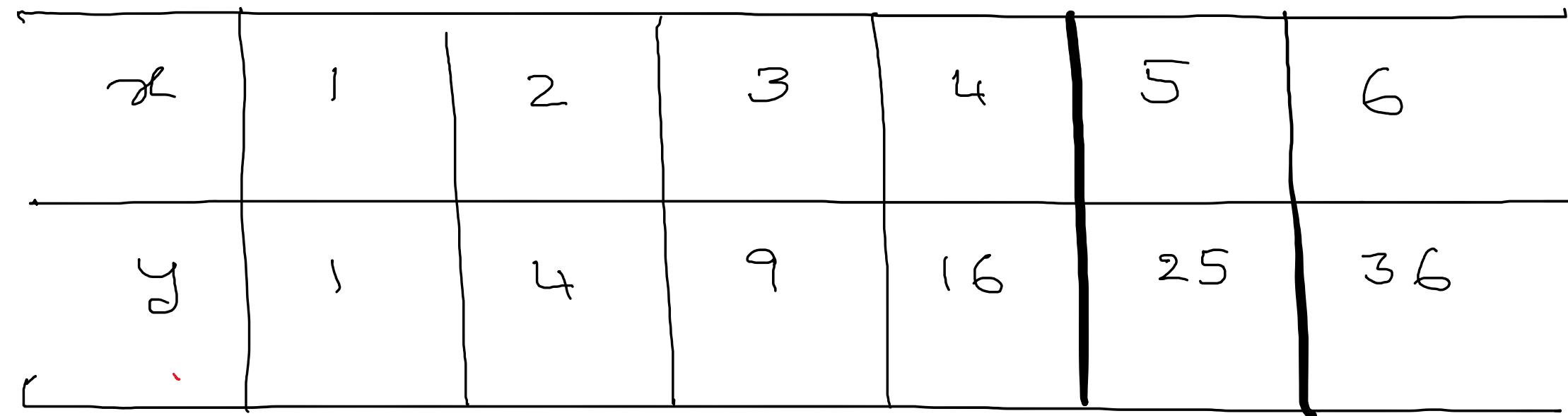
X	79	29	45	61	24	38	33	52	65	63	82	50
Y	15	12	12	6	6	7	3	10	12	13	13	14

Q3) Check whether X and Y are having any relation.

X	12	14	15	17	17	16	16	15	14	14
Y	52	52	57	62	67	67	67	62	62	52



Regression



prediction

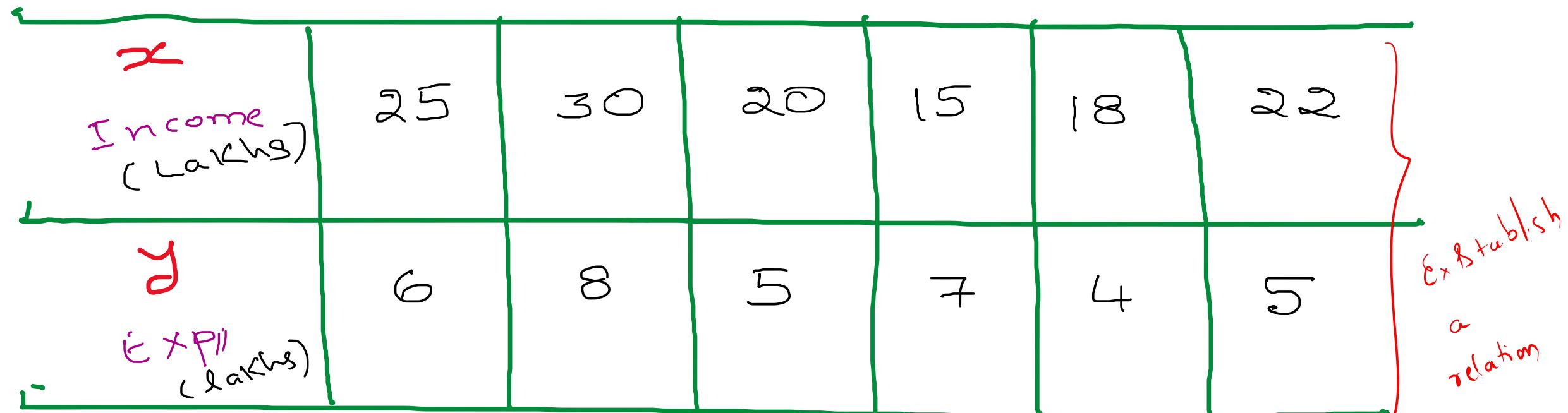
$$x = 8$$

$$y = ?$$

$$y = x^2$$

$$\begin{aligned} y &= 8 \\ y &= 64 \end{aligned}$$

Regression

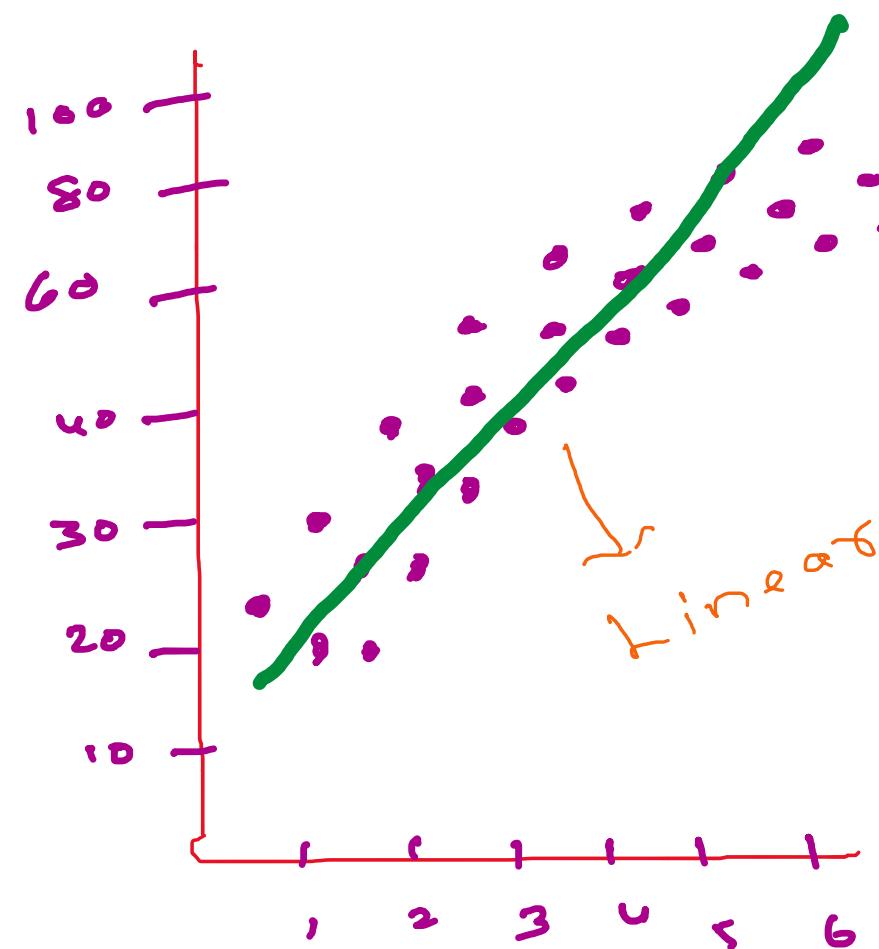


when $x = 32$, Then

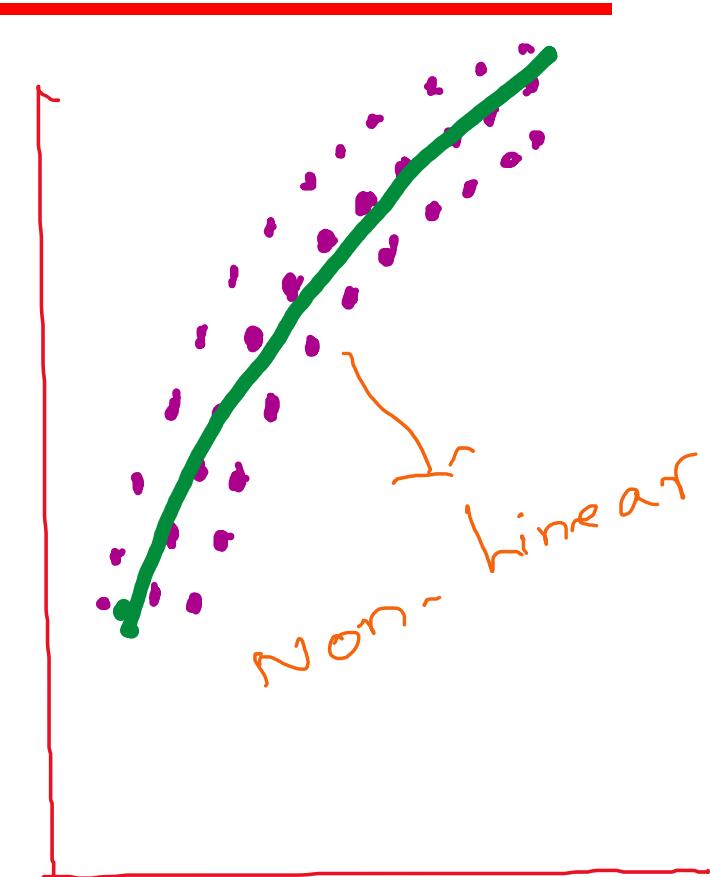
$$y = ?$$

Linear regression

$$y = mx + c$$
$$y = w_0 + w_1 x$$

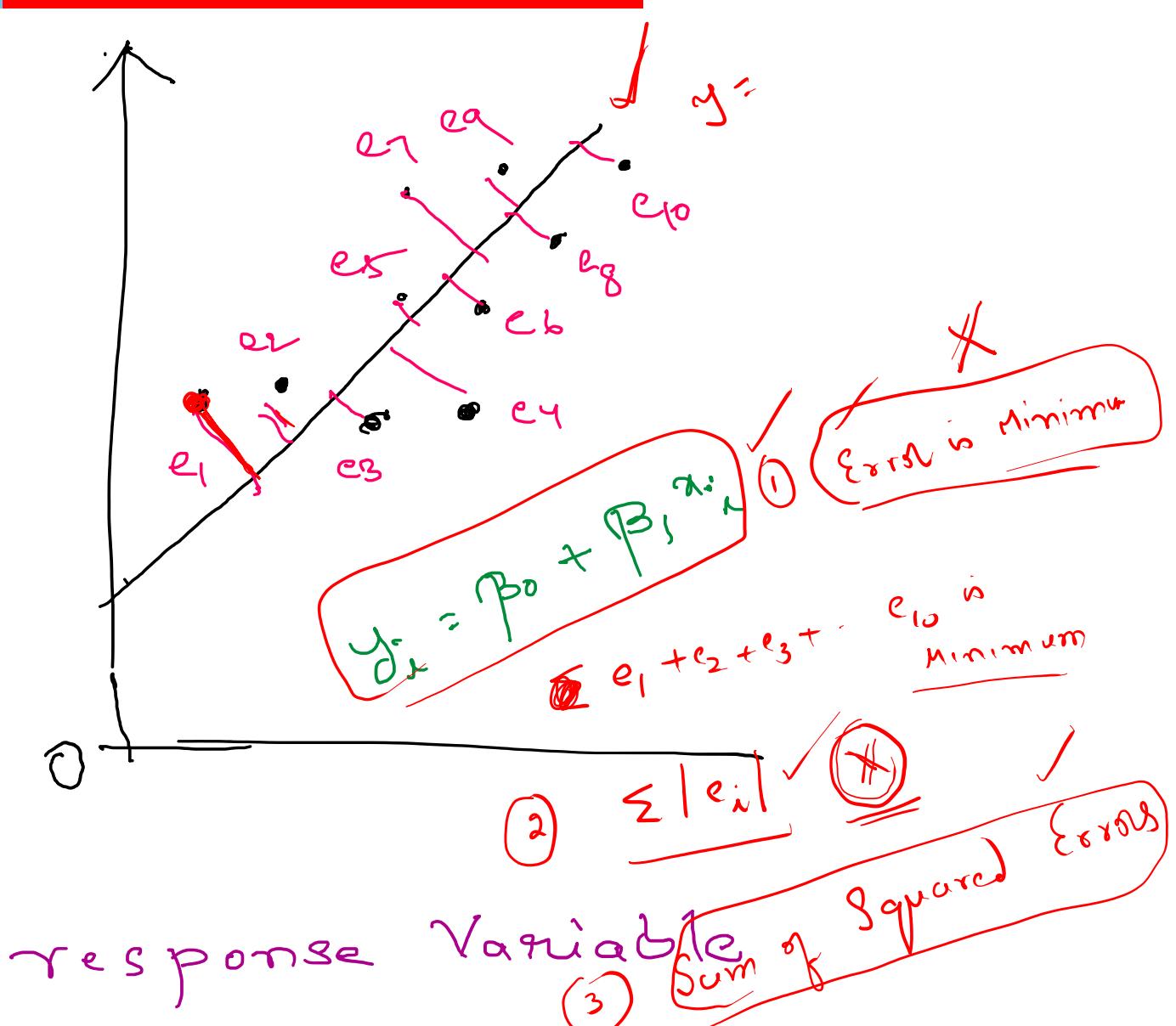
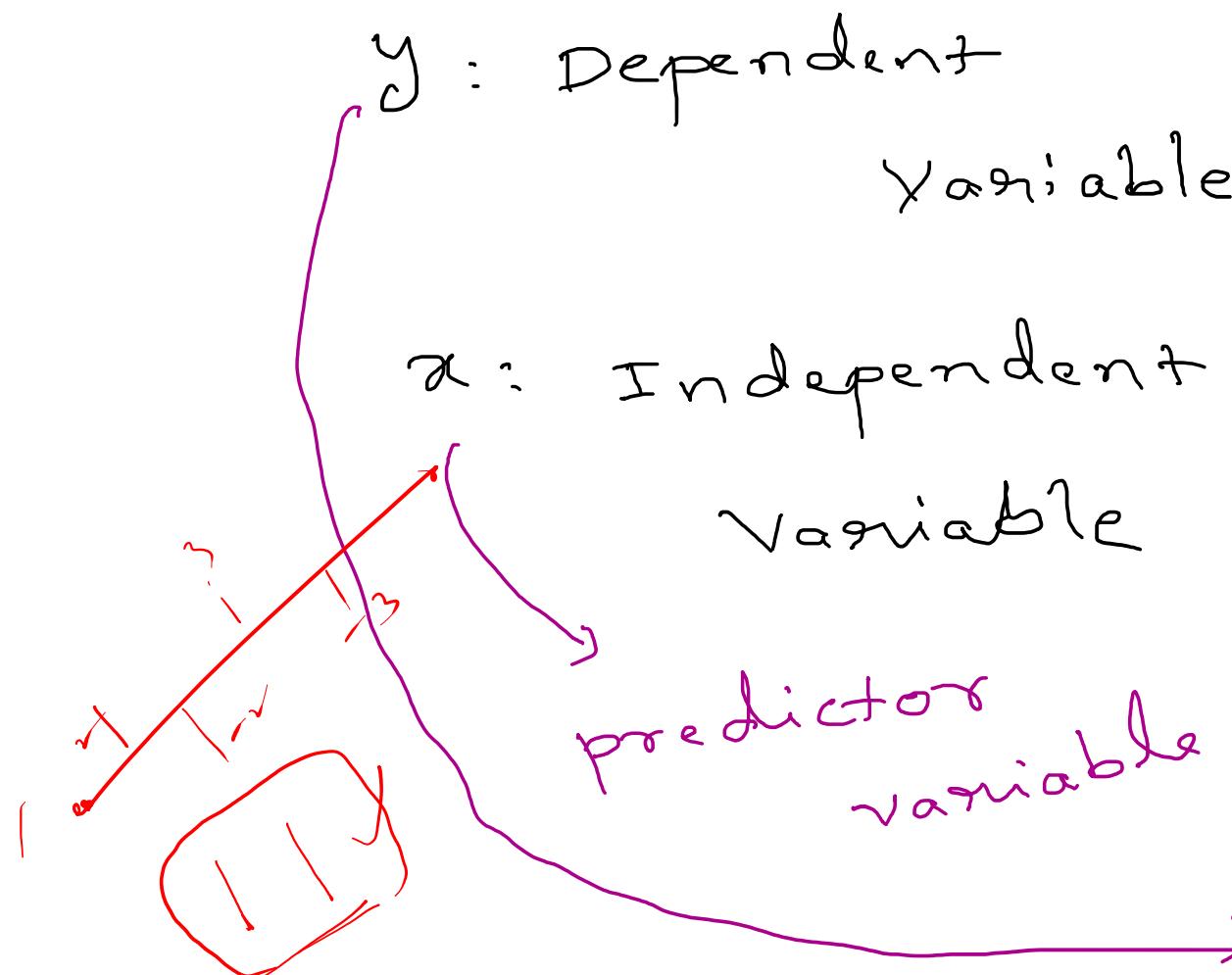


Linear

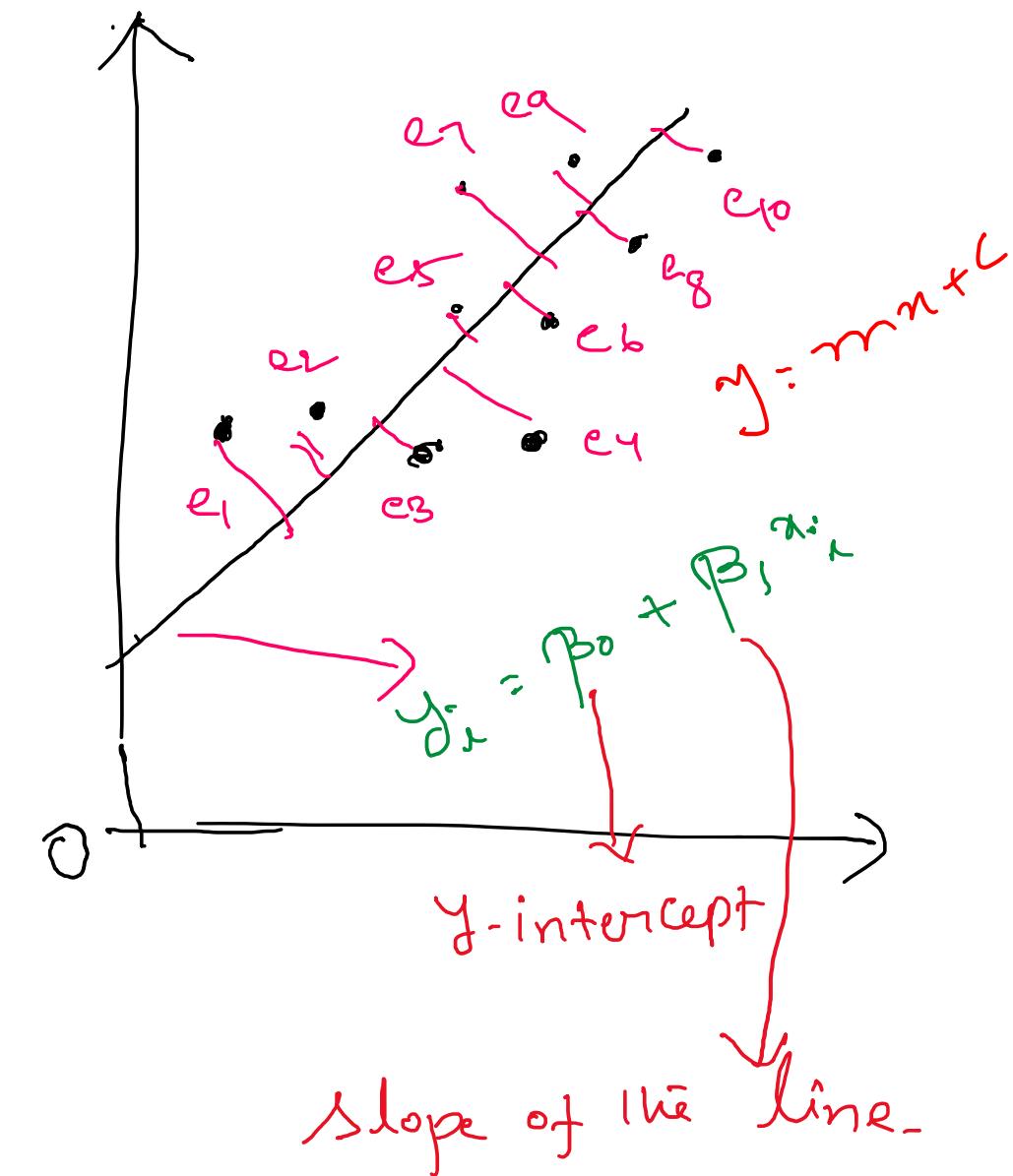


CORRELATION	REGRESSION
Measuring strength or degree of the relationship between two variables	Having an algebraic equation between two variables
No estimation	Estimation
Both variables are independent	One is dependent variable and the other is independent variable

Method of Least Squares



Method of Least squares



Method of Least squares

$$S(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - (\hat{y}_i))^2$$

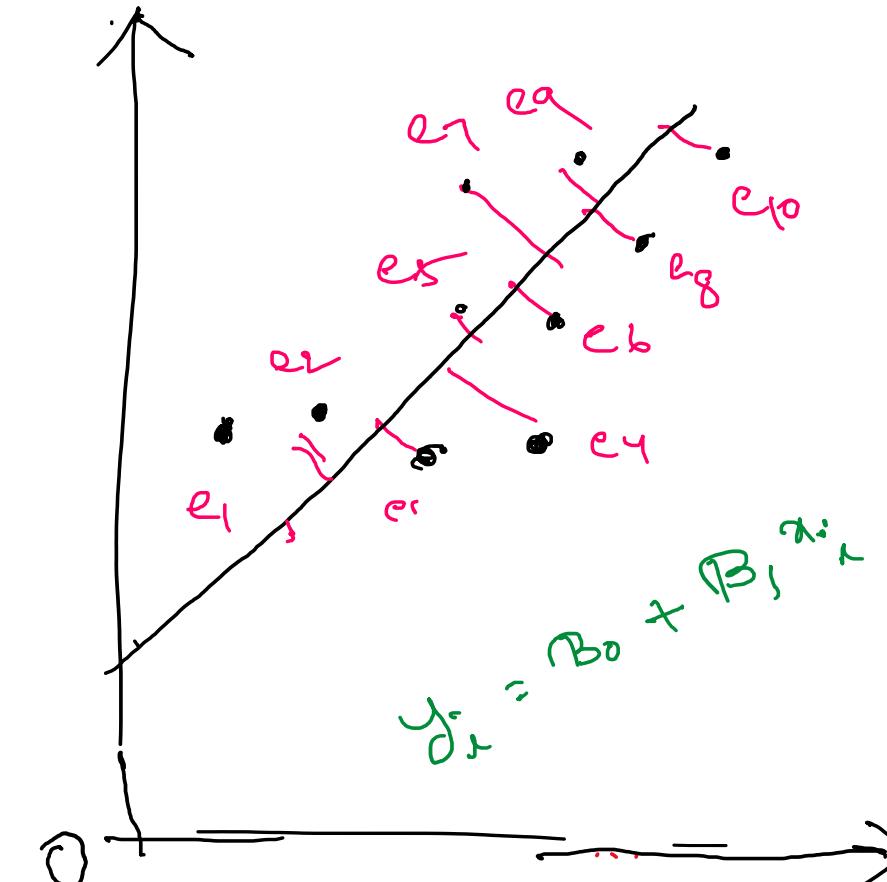
Cost / Loss

we need to choose
 β_0 and β_1 , which

minimizes the
 error.

Convex function

SSE



Method of Least squares

$$S(\beta_0, \beta_1) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

$$\frac{\partial S}{\partial \beta_0} = 0$$

$$\Rightarrow 2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) (-1) = 0$$

$$\sum_{i=1}^n y_i = n\beta_0 + \beta_1 \sum_{i=1}^n x_i$$

$$\frac{\partial S}{\partial \beta_1} = 0 \Rightarrow 2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) (-x_i) = 0$$

$$\Rightarrow \sum_{i=1}^n x_i y_i = \beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2$$

On solving these, we get β_0 & β_1
 which minimizes error.

Linear regression -

$$y = \beta_0 + \beta_1 x$$

$$\sum y = \beta_0 n + \beta_1 \sum x$$

$$\sum xy = \beta_0 \sum x + \beta_1 \sum x^2$$

Normal equations

$$y = \beta_0 + \beta_1 x$$

$$\begin{cases} \beta_0 = ? \\ \beta_1 = ? \end{cases}$$

$$Y = \beta_0 + \beta_1 X$$

$$\beta_1 = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$\beta_0 = \frac{\sum y - \beta_1 \sum x}{n}$$

Example :-

company	Advt	Sales
	Expt x	Revenue y
A	1	1
B	3	2
C	4	2
D	6	4
E	8	6
F	9	8
G	11	8
H	14	9

$$y = a + b\boxed{x}$$

$$\sum y = \underline{an} + b\underline{x}$$

$$\sum xy = a \underline{x} + b \underline{x^2}$$

Example :-

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
$\sum 60$		$\sum 524$	$\sum 373$

$$y = \beta_0 + \beta_1 x$$

$$\sum y = n\beta_0 + \beta_1 \sum x$$

$$\sum xy = \beta_0 \sum x + \beta_1 \sum x^2$$

$$\Rightarrow 40 = 8\beta_0 + 56\beta_1$$

$$373 = 56\beta_0 + 524\beta_1$$

on solving

$$\beta_0 = 0.072$$

$$\beta_1 = 0.704$$

i.e. $y = (0.072) + (0.704)x$

$$i.e \quad y = (0.072) + (0.704)x$$

when $x = \underline{0.075}$, then

$$\begin{aligned}y &= (0.072) + (0.704)(0.075) \\&= \textcircled{0.1248} \approx 12.48\%.\end{aligned}$$

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
$\Sigma = 7$	$\Sigma 3.5$	$\Sigma 10.5$	$\Sigma 21$

$$y = \beta_0 + \beta_1 x \quad \checkmark$$

$$\sum y = n\beta_0 + \beta_1 \sum x$$

$$\sum xy = \beta_0 \sum x_1 + \beta_1 \sum x^2$$

$$3.5 = 4\beta_0 + \beta_1 (7)$$

$$10.5 = 7\beta_0 + \beta_1 (21)$$

On solving these

$$\beta_0 = 0 \quad \checkmark$$

$$\beta_1 = 0.5 \quad \checkmark$$

$$\text{i.e. } y = 0 + (0.5)x$$

When $x = 5$, $y = (0.5)5$
 $= 0.25$

Assumptions of the Linear Regression

Assumptions about the Error

- $E(\varepsilon_i) = 0$ for $i = 1, 2, \dots, n.$
- $\sigma(\varepsilon_i) = \sigma_\varepsilon$ where σ_ε is unknown.
- The errors are independent, that is, the error in the i th observation is independent of the error observed in the j th observation.
- The ε_i are normally distributed (with mean 0 and standard deviation σ_ε).
MSE SSE

$$y_i = \hat{y}_i + (y_i - \hat{y}_i)$$

↓ ↓ ↓
 observed Fit Deviation from fit

subtracting \bar{y} from both sides

$$(y_i - \bar{y}) = (\hat{y} - \bar{y}) + (y_i - \hat{y}_i)$$

↓ ↓ ↓
 Deviation Deviation Residual
 from due to fit
 mean

RSS → Residual sum of squares

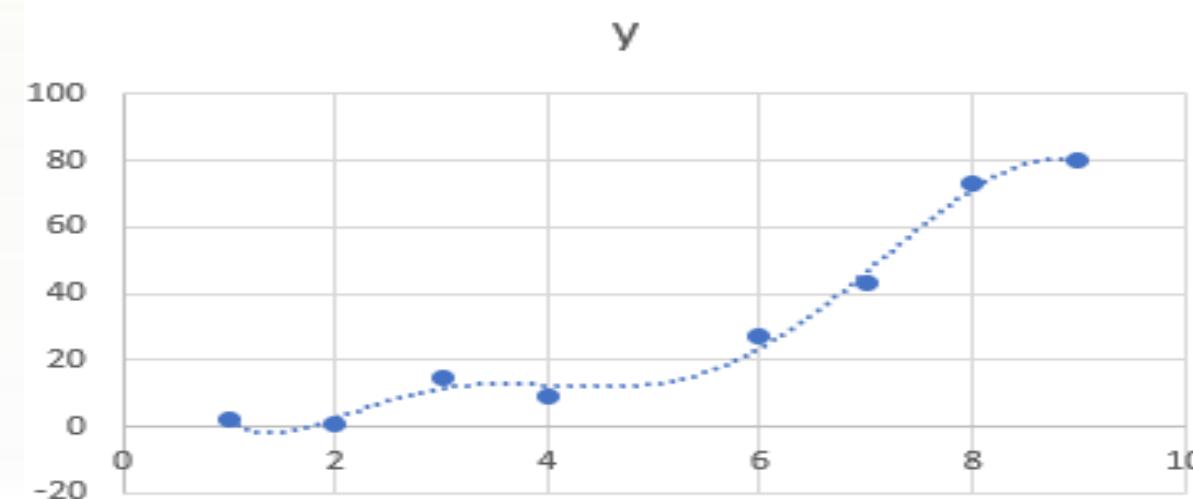
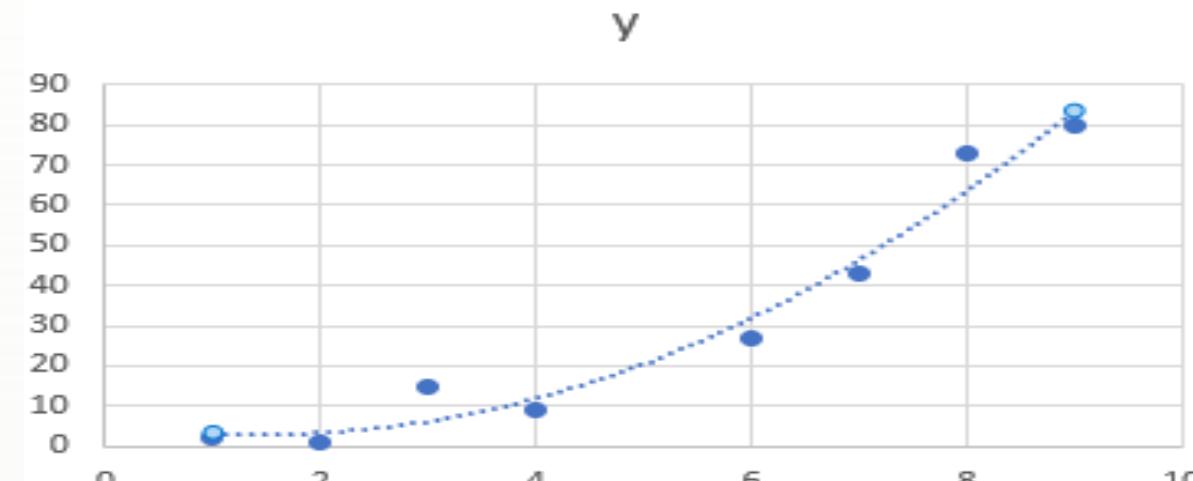
$$= \sum_{i=1}^n (y_i - (\alpha + \beta x_i))^2$$

TSS → $\sum_{i=1}^n (y_i - \bar{y})^2$ mean of
respective
variables

$$R^2 = 1 - \frac{RSS}{TSS}$$

Overfitting

Which of these two models would be a better fit to the data?



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.

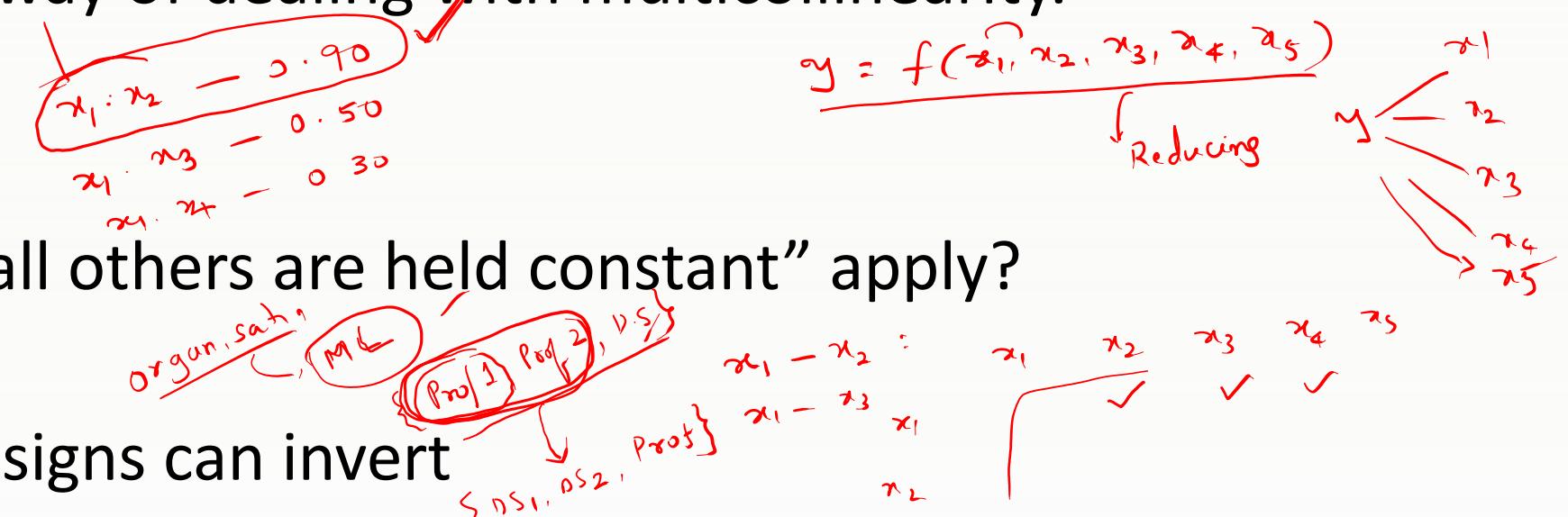
Multicollinearity affects:

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:

$$VIF_i = \frac{1}{1 - R_i^2}$$



where 'i' refers to the i-th variable which is being represented as a linear combination of rest of the independent variables.

- ❖ 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.

$$y = f(x_1, x_2) \quad y = w_0 + w_1 x_1 + w_2 x_2$$

Example:-

x_0	size	No of rooms	No of floors	Age of home	Price Lakh
0	2000	5	2	45	4000
1	1400	3	1	40	2000
1	1600	3	2	30	3000
1	800	2	1	35	2000

$\downarrow x_1 \quad \downarrow x_2 \quad \downarrow x_3$

w_0, w_1, w_2

$A^T \cdot A$
 $A^T \cdot B$
 $(\cdot)^{-1}$
 $O.L.S.$

$$y = w_0 + w_1 x$$

$$\text{SSE} = \sum (y - \hat{y})^2$$

$$= \sum (y - (w_0 + w_1 x))^2$$

$$\sum y = w_0 n + w_1 \sum x$$

$$\sum xy = w_0 \sum x + w_1 \sum x^2$$

$$\sum (y - (w_0 + w_1 x_1 + w_2 x_2))^2$$

$$\sum y = w_0 n + w_1 \sum x_1 + w_2 \sum x_2$$

$$\sum xy = w_0 \sum x_1 + w_1 \sum x_1^2 + w_2 \sum x_1 x_2$$

$$\sum x_1 y = w_0 \sum x_3 + w_1 \sum x_1 x_2 + w_2 \sum x_2^2$$

Multiple Linear Regression

The data consists of n observations on a dependent or response variable Y and p predictor or explanatory variables

$$x_1, x_2, \dots, x_p$$

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \epsilon$$

β_i 's are regression coefficients.

Normal equations

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\sum y = \beta_0 n + \beta_1 \sum x_1 + \beta_2 \sum x_2$$

$$\sum x_1 y = \beta_0 \sum x_1 + \beta_1 \sum x_1^2 + \beta_2 \sum x_1 x_2$$

$$\sum x_2 y = \beta_0 \sum x_2 + \beta_1 \sum x_1 x_2 + \beta_2 \sum x_2^2$$

Multiple Linear

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

$$\sum y = \beta_0 n + \beta_1 \sum x_1 + \beta_2 \sum x_2 + \beta_3 \sum x_3 + \beta_4 \sum x_4$$

$$\sum x_1 y = \beta_0 \sum x_1 + \beta_1 \sum x_1^2 + \beta_2 \sum x_1 x_2 + \beta_3 \sum x_1 x_3 + \beta_4 \sum x_1 x_4$$

$$\sum x_2 y = \beta_0 \sum x_2 + \beta_1 \sum x_1 x_2 + \beta_2 \sum x_2^2 + \beta_3 \sum x_2 x_3 + \beta_4 \sum x_2 x_4$$

$$\sum x_3 y = \beta_0 \sum x_3 + \beta_1 \sum x_1 x_3 + \beta_2 \sum x_2 x_3 + \beta_3 \sum x_3^2 + \beta_4 \sum x_3 x_4$$

$$\sum x_4 y = \beta_0 \sum x_4 + \beta_1 \sum x_1 x_4 + \beta_2 \sum x_2 x_4 + \beta_3 \sum x_3 x_4 + \beta_4 \sum x_4^2$$

separately $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$

Normal equations

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

$$\sum y = \beta_0 n + \beta_1 \sum x_1 + \beta_2 \sum x_2$$

$$\sum x_1 y = \beta_0 \sum x_1 + \beta_1 \sum x_1^2 + \beta_2 \sum x_1 x_2$$

$$\sum x_2 y = \beta_0 \sum x_2 + \beta_1 \sum x_1 x_2 + \beta_2 \sum x_2^2$$

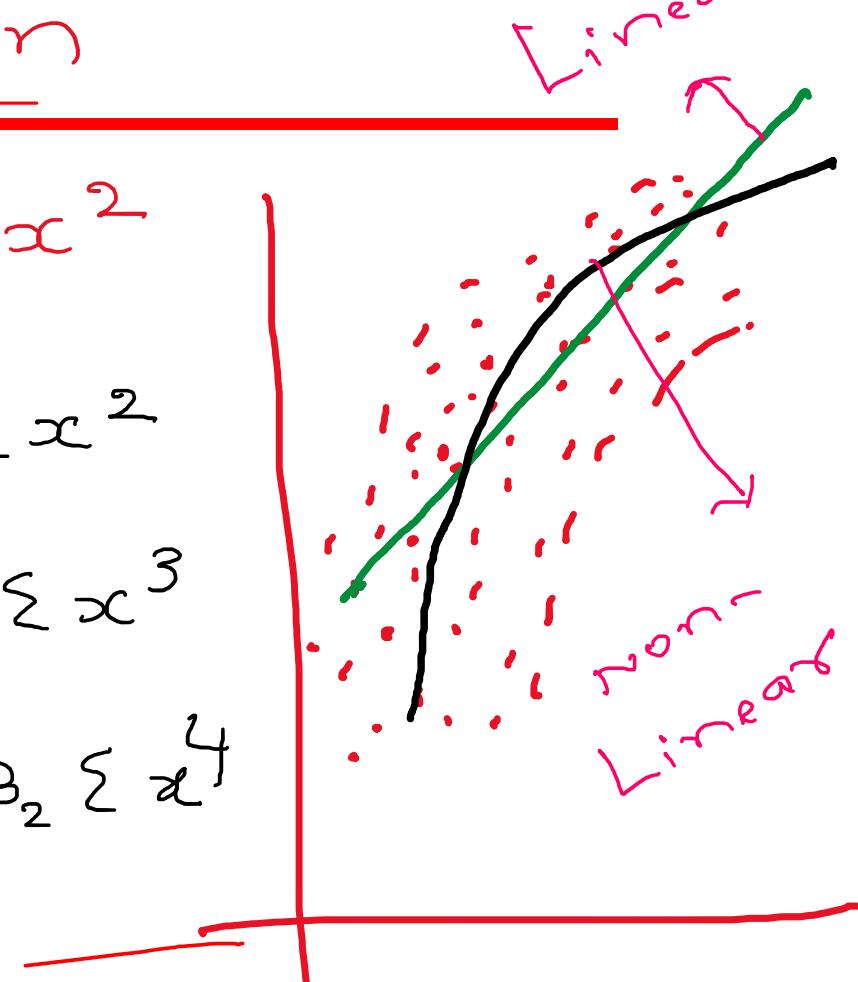
Non - Linear Regression

$$y = \beta_0 + \beta_1 x + \beta_2 x^2$$

$$\sum y = \beta_0 n + \beta_1 \sum x + \beta_2 \sum x^2$$

$$\sum xy = \beta_0 \sum x + \beta_1 \sum x^2 + \beta_2 \sum x^3$$

$$\sum x^2 y = \beta_0 \sum x^2 + \beta_1 \sum x^3 + \beta_2 \sum x^4$$



$$1) Y = \alpha x^{\beta}$$

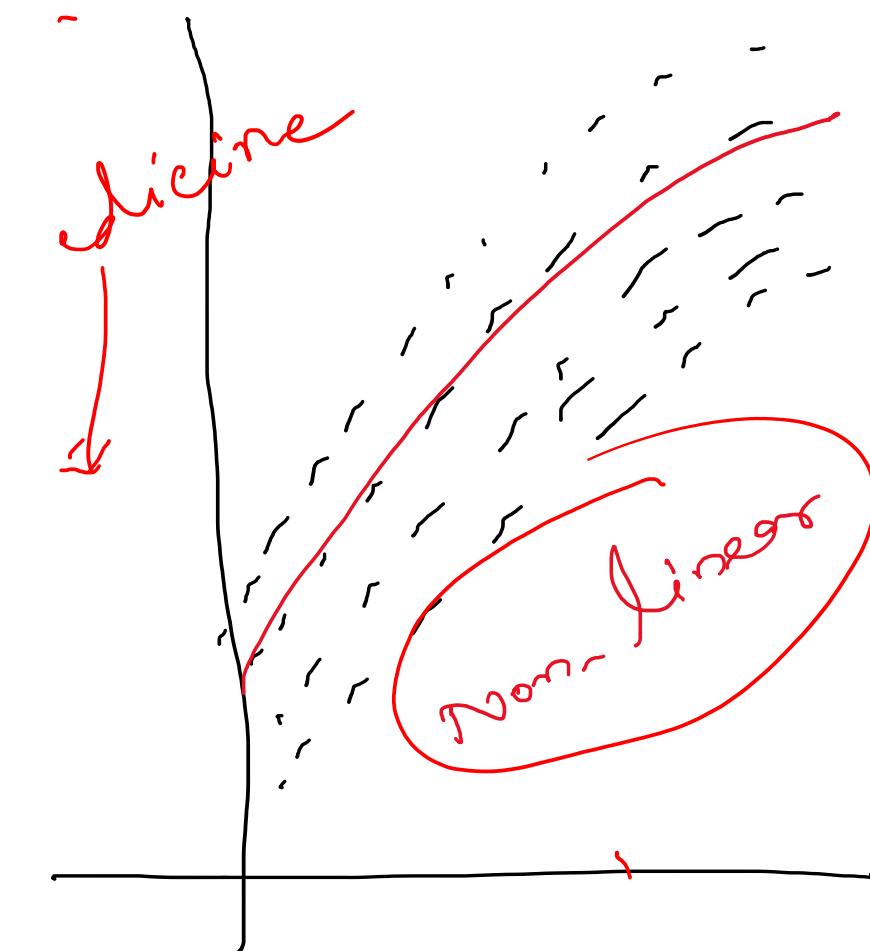
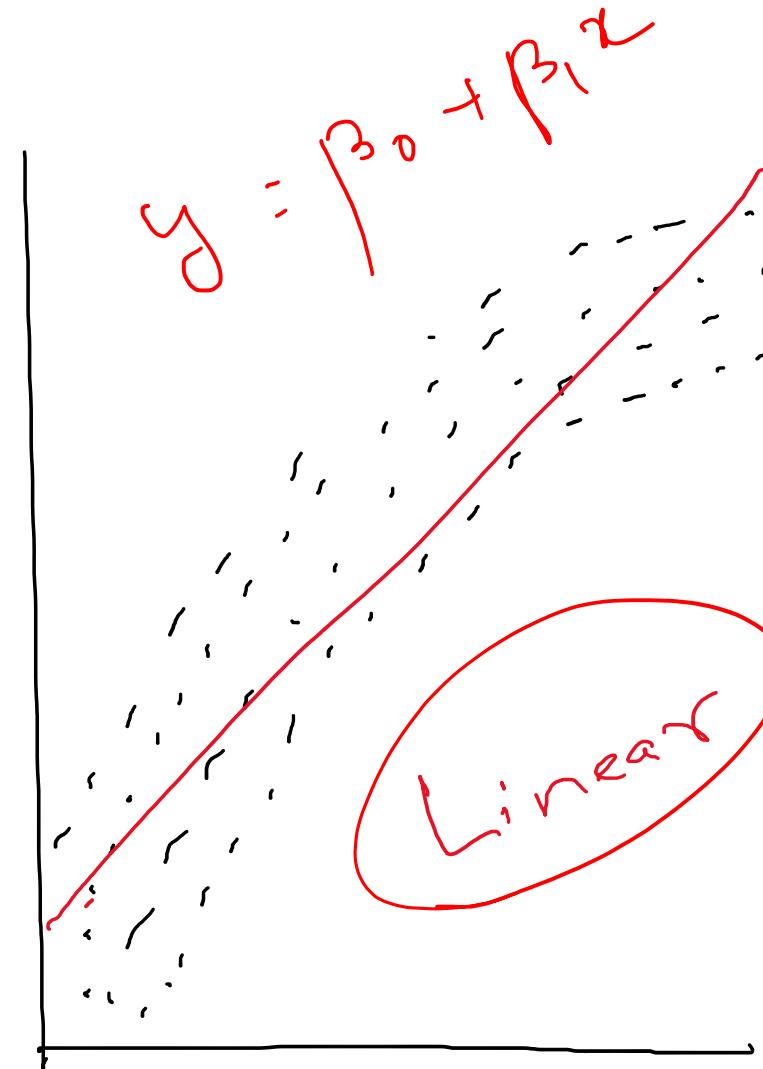
$$2) Y = \alpha e^{\beta x}$$

$$3) Y = \alpha + \beta \log x$$

$$4) Y = \frac{\alpha}{\alpha x - \beta}$$

Other regressions

just a look



Suppose $y = a e^{bx}$

$$\log y = \log a + b \log x$$

γ x

A

ie

$$\gamma = A + b x$$

$$\sum \gamma = A n + b \sum x \rightarrow 1$$

$$\sum x \cdot \gamma = A \sum x + b \sum x^2 \rightarrow 2$$

Hence, we get
 $b n$

$$y = a e^{bx}$$

R – Squared vs Adjusted R - Squared

- In multiple regression, adjusted R – squared is better metric than R – squared asses the goodness of fit of the model

Accuracy
SLR

$\hat{R}^2 = 0.80$

MLR

- R – squared always increases if additional variables are added into model , even if they are not related to the dependent variable

Lasso & Ridge

"Apollo" - THE BEST

GLS

Medical shop

"Colony"

SSE

The Best ✓

OLS

$\hat{Y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$

Census Data: 0.5

Medical data: 0.5 ✓

• OLS estimation:

• LASSO estimation:

• Ridge regression estimation:

Gradient Descent Method

W₀, W₁

Error

W₀, W₁

Error - 51

W₀, W₁

Error - 21

Disadvantage

Programming

Control over the error

$$\min SSE = \sum (Y - \hat{Y})^2$$

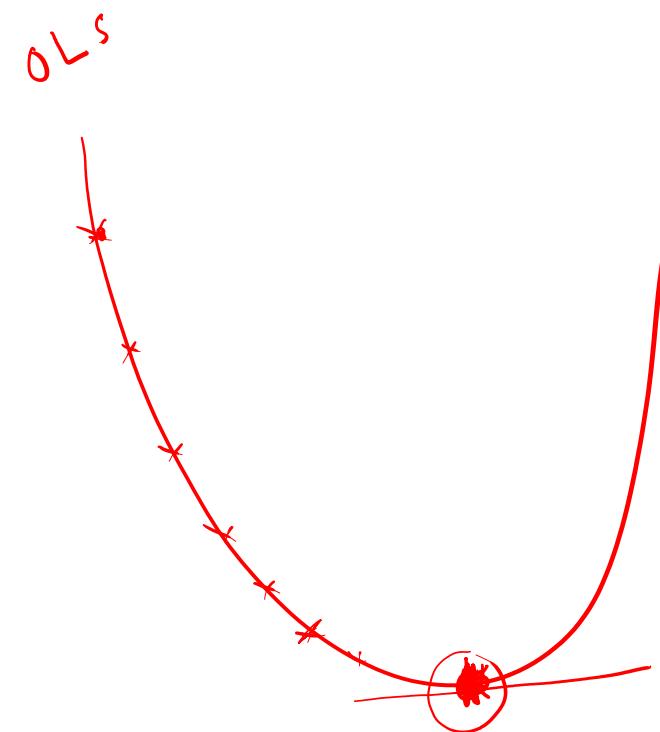
$$\min SSE = \sum_{i=1}^n (Y - \hat{Y})^2 + \lambda \sum_{j=1}^p |\beta_j|$$

$$\min SSE = \sum_{i=1}^n (Y - \hat{Y})^2 + \lambda \sum_{j=1}^p |\beta_j|^2$$

$$w_{n+1} := w_n - \alpha \left(\frac{\partial K}{\partial w_i} \right)$$

A hand-drawn diagram illustrating the update rule for gradient descent. A horizontal line is divided into three colored segments: yellow on the left, blue in the middle, and red on the right. A red arrow points from the yellow segment towards the blue segment. Above the line, the update equation is written in red. To the right of the equation, a red circle contains the Greek letter α . A curved red arrow originates from this circle and points downwards and to the right, indicating the direction of the weight update. Below the line, there is a small red sketch of a hand holding a pen.

Thanks





BITS Pilani

Pilani|Dubai|Goa|Hyderabad

Introduction to Statistical Methods

Team ISM

T.S



Maximum Likelihood Estimation (MLE)

Estimation is the process of estimating unknown true values of population parameters using their corresponding best sample statistics (good estimators) in an optimum manner.

An estimator is said to be a good if it is

- unbiased,
- consistent,
- efficient and
- sufficient while estimating its parameter.

MAP

$$\begin{aligned}
 p(\hat{A}|E) &= \frac{P(E|A)p(A)}{\sum p(E|\theta)p(\theta)} \\
 p(\hat{B}|E) &= \frac{P(E|B)p(B)}{\sum p(E|\theta)p(\theta)} \\
 \max(p(E|A), p(E|B)) &\quad \text{(MAP)} \\
 \arg \max_{\theta} P(E|\theta) &= \frac{P(A)}{P(A) + P(B)} \\
 \arg \max_{\theta} P(E|\theta) &= \frac{P(A)}{P(A) + P(B)} = \frac{p(A)}{p(A) + p(B)} \\
 \text{ML Estimate} &
 \end{aligned}$$

Maximum Likelihood Estimation (MLE)

- ❖ Method of Maximum Likelihood Estimation is the best and most popular one among all methods to obtain an almost good or best estimator for a population parameter.
- ❖ It is a method of obtaining an estimator which most (maximum) likely estimates the true value of the parameter i.e., finding an estimator that can give most likely nearer value for the unknown true value of parameter.
- ❖ The corresponding estimator is called maximum likelihood estimator (MLE).

Maximum Likelihood Estimation (MLE)

Suppose we have a random sample x_1, x_2, \dots, x_n whose assumed probability distribution depends on some unknown parameter θ .

Ex:

- 1) For Binomial unknown parameters are n, p . ✓
- 2) For Poisson unknown parameter is λ . ✓
- 3) For Normal unknown parameters are μ and σ^2 . ✓

Our goal is to find good estimate of θ (population parameter) using sample and which can be done with the help of MLE.

Maximum Likelihood

- ❖ It is observed that a good estimate of unknown parameter θ would be the value of θ that maximizes the probability
- ❖ i.e. the likelihood of getting the data we observed (this is reason, why we called as likelihood function)

Maximum Likelihood function

- ❖ Let x_1, x_2, \dots, x_n be i.i.d. random variables drawn from some probability distribution that depends on some unknown parameter θ .
- ❖ The goal of MLE to maximize likelihood function

$$L(\theta) = f(x_1, x_2, \dots, x_n | \theta)$$

$$= f(x_1 | \theta) * f(x_2 | \theta) \dots f(x_n | \theta)$$

$$L(\theta) = \prod_{i=1}^n f(x_i / \theta)$$

Maximum Likelihood Estimation (MLE)

- ❖ The maximum likelihood estimate (MLE) of θ is that value of θ that maximizes $\text{likelihood}(\theta)$.

It is defined as

$$L(\theta) = \prod_{i=1}^n f(x_i / \theta)$$

$$\log L(\theta) = \sum_{i=1}^n \log f(x_i / \theta)$$

For maximization,
we have

$$\frac{dL}{d\theta} = 0 \quad ; \quad \frac{d^2L}{d\theta^2} < 0$$

Maximum Likelihood Estimation (MLE)

If L and $\log_e L$ are not differentiable or integrable or principle of maxima-minima fails then in such case direct method of finding the estimator of the parameter which maximizes L or $\log_e L$ is applied using order statistic principle empirically.

Maximum Likelihood Estimation (MLE)

MLEs are:

- ❖ Consistent
 - ❖ Efficient
 - ❖ Sufficient
 - ❖ MLEs May (or may not) be unbiased
 - ❖ MLEs are Asymptotically normally distributed
 - ❖ Asymptotically tend to have least variance.
-

Example: An unfair coin is flipped 100 times, and 61 heads are observed. The coin either has probability $\frac{1}{3}$, $\frac{1}{2}$, or $\frac{2}{3}$ of flipping a head each time it is flipped. Which of the three is the MLE?

Example: An unfair coin is flipped 100 times, and 61 heads are observed. The coin either has probability $\frac{1}{3}$, $\frac{1}{2}$, or $\frac{2}{3}$ of flipping a head each time it is flipped. Which of the three is the MLE?

Solution: Here the distribution is the binomial distribution with $n = 100$.

$$P(H = 61 | p = \frac{1}{3}) = \binom{100}{61} \left(\frac{1}{3}\right)^{61} \left(\frac{2}{3}\right)^{39} \approx 9.6 \times 10^{-9}$$

$$P(H = 61 | p = \frac{1}{2}) = \binom{100}{61} \left(\frac{1}{2}\right)^{61} \left(\frac{1}{2}\right)^{39} = 0.007$$

$$P(H = 61 | p = \frac{2}{3}) = \binom{100}{61} \left(\frac{2}{3}\right)^{61} \left(\frac{1}{3}\right)^{39} = 0.040$$

p.m.f.

$$P(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

$$0 \leq p \leq 1$$

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

Since $P(H = 61 | p = \frac{2}{3})$ is maximum and hence MLE is $p = \frac{2}{3}$



Example: An unfair coin is flipped 100 times, and 61 heads are observed. What is the MLE when nothing is previously known about the coin?

$$\int \frac{df}{dx} = 0$$

Example: An unfair coin is flipped 100 times, and 61 heads are observed. What is the MLE when nothing is previously known about the coin?

Solution: Since the distribution follows is Binomial distribution, with parameter p .

Here $n = 100$. The likelihood function (MLE) is

$$P(H = 61|p) = \binom{100}{61} p^{61}(1-p)^{39}$$

Maximization

For maximization

$$\frac{d}{dp} P(H = 61|p) = 0$$

$$\Rightarrow \binom{100}{61} [61p^{60}(1-p)^{39} - 39p^{61}(1-p)^{38}] = 0$$

$$\Rightarrow p^{60}(1-p)^{38}(61 - 100p) = 0$$

$$\Rightarrow p = 0, \frac{61}{100}, 1$$

max func
100 - 61 = 39

Thus, the likelihoods are

$$P(H = 61|p = 0) = 0$$

$$P(H = 61|p = \frac{61}{100}) = \binom{100}{61} \left(\frac{61}{100}\right)^{61} \left(\frac{39}{100}\right)^{39}$$

$$P(H = 61|p = 1) = 0$$

Since $P(H = 61|p = \frac{61}{100})$ is maximum
and hence $p = \frac{61}{100}$ is the MLE.

min p? max p?

Maximum Likelihood for a Binomial distribution

- ❖ Suppose we wish to find the maximum likelihood estimate (MLE) of θ for a Binomial distribution,

$$p_k(k, \theta) = nC_k \theta^k (1 - \theta)^{n-k}$$

$$\frac{d}{d\theta} (\quad) = 0$$

$$\log p_k(k, \theta) = \log(nC_k) + k \log(\theta) + (n - k) \log((1 - \theta))$$

$$\frac{\partial \log p_k(k, \theta)}{\partial \theta} = 0 \Rightarrow 0 + \frac{k}{\theta} - \frac{n - k}{1 - \theta} = 0$$

$$k - k\theta = n\theta - k\theta \Rightarrow \theta = \frac{k}{n}$$

Maximum Likelihood Example 3:

Consider a sample 0,1,0,0,1,0 from a binomial distribution, with the form $P[X=0]=(1-p)$, $P[X=1]=p$. Find the maximum likelihood estimate of p.

$$\underline{p(x)}$$

Soln :



Maximum Likelihood Example 3:

Consider a sample $0,1,0,0,1,0$ from a binomial distribution, with the form $P[X=0]=(1-p)$, $P[X=1]=p$. Find the maximum likelihood estimate of p .

Soln :

$$\begin{aligned} L(p) &= P[X=0] P[X=1] P[X=0] P[X=0] P[X=1] P[X=0] \\ &= (1-p) p (1-p) (1-p) p (1-p) \\ &= (1-p)^4 p^2. \end{aligned}$$

$$\text{Log } L(p) = \log[(1-p)^4 p^2.] = \log[(1-p)^4] + \log[(p^2).] = 4\log(1-p) + 2\log p$$

$$\frac{\partial \text{Log } L(p)}{\partial p} = 0 \text{ means,}$$

$$\frac{-4}{1-p} + \frac{2}{p} = 0 \Rightarrow \frac{-4p + 2 - 2p}{p(1-p)} = 0 \Rightarrow p = \frac{1}{3}$$

That is , there is $1/3$ chance to observe this sample if we believe the population to be Binomial distributed .

MLE for Binomial distribution parameter P

Let $X_1, X_2, \dots, X_N \in \mathbb{R}$ be samples obtained from a Binomially Distribution.

Binomial Distribution is used to model 'x' successes in 'n' Bernoulli trials. Its p.d.f. is given by:

$$f(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}$$

The likelihood function $L(p)$ is given by:

$$L(p) = \prod_{i=1}^n f(x_i) = \prod_{i=1}^N \left| \frac{n!}{x_i!(n-x_i)!} p^{x_i} (1-p)^{n-x_i} \right|$$

The log-likelihood is:

$$\ln L(p) = \sum_{i=1}^N \ln(n!) - \sum_{i=1}^N \ln(x_i!) - \sum_{i=1}^N \ln(n-x_i!) + \sum_{i=1}^N x_i \cdot \ln(p) + \left(n - \sum_{i=1}^N x_i \right) \cdot \ln(1-p)$$

Setting its derivative with respect to p to zero,

$$\frac{d}{dp} \ln L(p) = \frac{1}{p} \cdot \sum_{i=1}^N xi - \frac{1}{1-p} \sum_{i=1}^N (n - xi) = 0$$

which implies,

$$\frac{1}{p} \cdot \sum_{i=1}^N xi = \left(\frac{1}{1-p}\right)(N \cdot n - \sum_{i=1}^N xi)$$

giving,

$$\hat{p} = \frac{1}{N} \left(\frac{\sum_{i=1}^N x_i}{n} \right) = \frac{1}{N} \left(\frac{x_1}{n} + \frac{x_2}{n} + \dots + \frac{x_N}{n} \right) = \bar{x}$$

which is the maximum likelihood estimate.

MLE for Poisson Distribution Parameter

Let $X_1, X_2, \dots, X_n \in \mathbb{R}$ be a random sample from a Poisson distribution

The p.d.f. of a Poisson Distribution is :

$$f(x) = \frac{\lambda^x e^{-\lambda}}{x!}; \text{ where } x = 0, 1, 2, \dots$$

The likelihood function is:

$$L(\lambda) = \prod_{i=1}^n \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = e^{-\lambda n} \left| \frac{\lambda^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i} \right|$$

The log-likelihood is:

$$\ln L(\lambda) = -\lambda n + \sum_{i=1}^n x_i \cdot \ln(\lambda) - \ln(\prod_{i=1}^n x_i)$$

Setting its derivative with respect to λ to zero, we have:

$$\frac{d}{d\lambda} \ln L(\lambda) = -n + \sum_{i=1}^n xi \cdot \frac{1}{\lambda} = 0$$

giving,

$$\hat{\lambda} = \frac{\sum_{i=1}^n x_i}{n} = \bar{x}$$

which is the maximum likelihood estimate

MLEs for Normal Distribution Parameters

Let X_1, X_2, \dots, X_n be a random sample from a normal distribution with unknown mean μ and variance σ^2 . Find maximum likelihood estimators of mean μ and variance σ^2 .

In finding the estimators, the first thing we'll do is write the probability density function as a function of $\theta_1 = \mu$ and $\theta_2 = \sigma^2$:

$$f(x_i; \theta_1, \theta_2) = \frac{1}{\sqrt{\theta_2} \sqrt{2\pi}} \exp \left[-\frac{(x_i - \theta_1)^2}{2\theta_2} \right]$$

for $-\infty < \theta_1 < \infty$ and $0 < \theta_2 < \infty$. We do this so as not to cause confusion when taking the derivative of the likelihood with respect to σ^2 . Now, that makes the likelihood function:

$$L(\theta_1, \theta_2) = \prod_{i=1}^n f(x_i; \theta_1, \theta_2) = \theta_2^{-n/2} (2\pi)^{-n/2} \exp \left[-\frac{1}{2\theta_2} \sum_{i=1}^n (x_i - \theta_1)^2 \right]$$

and therefore the log of the likelihood function:

$$\log L(\theta_1, \theta_2) = -\frac{n}{2} \log \theta_2 - \frac{n}{2} \log(2\pi) - \frac{\sum (x_i - \theta_1)^2}{2\theta_2}$$

Now, upon taking the partial derivative of the log likelihood with respect to θ_1 , and setting to 0, we see that a few things cancel each other out, leaving us with:

$$\frac{\partial \log L(\theta_1, \theta_2)}{\partial \theta_1} = \frac{-\cancel{2} \sum (x_i - \theta_1) \cancel{(-1)}}{\cancel{2} \theta_2} \stackrel{\text{SET}}{\equiv} 0$$

Now, multiplying through by θ_2 , and distributing the summation, we get:

$$\sum x_i - n\theta_1 = 0 \quad \checkmark$$

Now, solving for θ_1 , and putting on its hat, we have shown that the maximum likelihood estimate of θ_1 is:

$$\hat{\theta}_1 = \hat{\mu} = \frac{\sum x_i}{n} = \bar{x}$$

Now for θ_2 . Taking the partial derivative of the log likelihood with respect to θ_2 , and setting to 0, we get:

$$\frac{\partial \log L(\theta_1, \theta_2)}{\partial \theta_2} = -\frac{n}{2\theta_2} + \frac{\sum (x_i - \theta_1)^2}{2\theta_2^2} \stackrel{\text{SET}}{\equiv} 0$$

Multiplying through by $2\theta_2^2$:

$$\frac{\partial \log L(\theta_1, \theta_2)}{\partial \theta_1} = \left[-\frac{n}{2\theta_2} + \frac{\sum (x_i - \theta_1)^2}{2\theta_2^2} \stackrel{\text{set}}{\equiv} 0 \right] \times 2\theta_2^2 \quad \checkmark$$

Click to edit Master title style

Hence the MLEs for the parameters mean and variance of the normal model are respectively:

we get:

$$-n\theta_2 + \sum(x_i - \theta_1)^2 = 0$$

And, solving for θ_2 , and putting on its hat, we have shown that the maximum likelihood estimate of θ_2 is:

$$\hat{\theta}_2 = \hat{\sigma}^2 = \frac{\sum(x_i - \bar{x})^2}{n}$$

$$\hat{\mu} = \frac{\sum X_i}{n} = \bar{X} \text{ and } \hat{\sigma}^2 = \frac{\sum(X_i - \bar{X})^2}{n}$$

MLEs for the parameters of Uniform distribution



Let x_1, x_2, \dots, x_n be a random sample drawn from a Uniform population with probability function

$$f(X, a, b) = 1/(b-a) \text{ for } a \leq X \leq b.$$

The likelihood function is expressed as follows:

$$L = f(x_1, a, b) f(x_2, a, b) \dots f(x_n, a, b)$$

$$= 1/(b-a) \times 1/(b-a) \times \dots \times 1/(b-a)$$

$$= 1/(b-a)^n = (b-a)^{-n}$$

$$\log L = -n \log(b-a)$$

$$d \log L / da = 0$$

$$n/(b-a) = 0$$

$n = 0$ (Contradiction)

Also, $d \log L / db = 0$

$$-n/(b-a) = 0$$

$-n = 0$ (Contradiction)

Therefore, principle of maxima fails to give MLEs.

Now, we use order statistic principle. By ordering the sample observations, we obtain

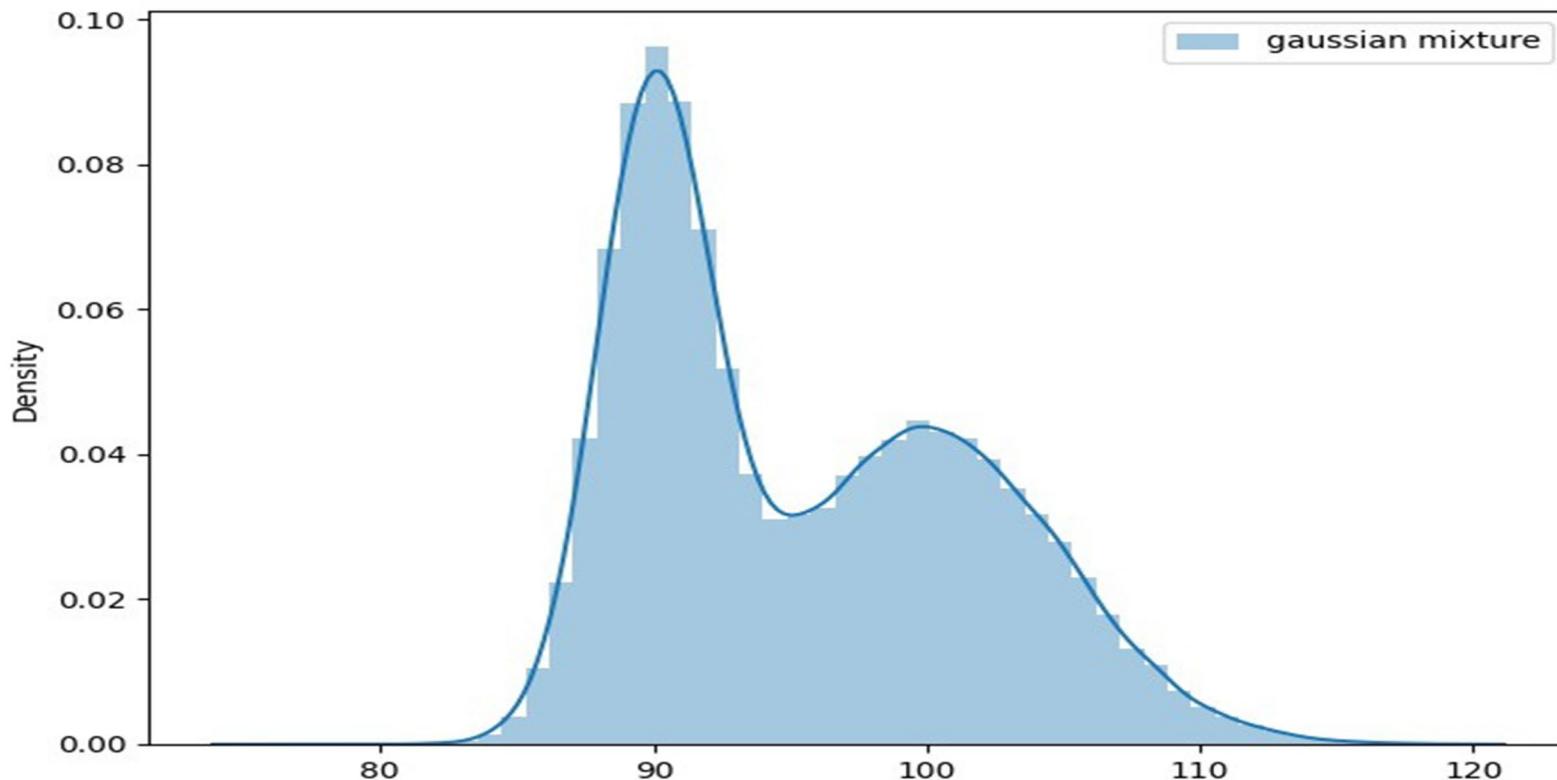
$$a \leq x_{(1)} < x_{(2)} < \dots < x_{(n)} \leq b \text{ since } a \leq X \leq b$$

Where $x_{(1)} = \text{Min}\{x_1, x_2, \dots, x_n\}$ and $x_{(n)} = \text{Max}\{x_1, x_2, \dots, x_n\}$

As $x_{(1)}$ and $x_{(n)}$ fall nearest to a and b respectively in the inside interval $[a, b]$ then $x_{(1)}$ and $x_{(n)}$ are considered as MLEs for the respective parameters a and b using order statistic principle.

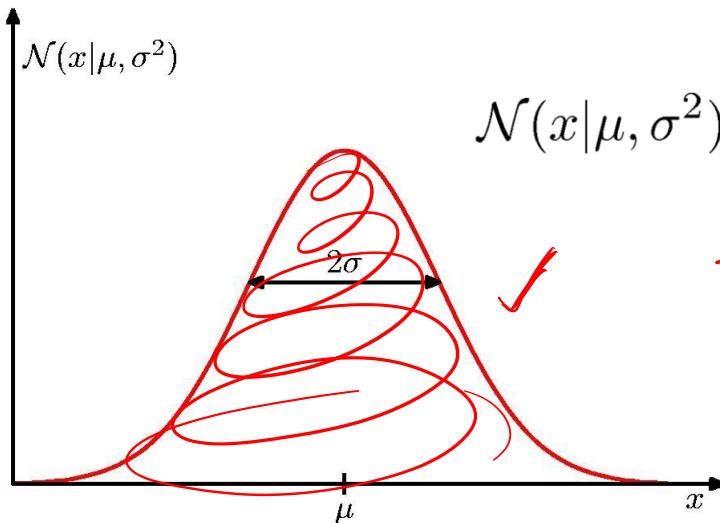
Gaussian Mixture Model

- ❖ Suppose Company A share price is normally distributed with mean price Rs. 100 and standard deviation of price Rs.2 with 1000 sample points
- ❖ Company B share price is normally distributed with mean price Rs. 100 and standard deviation of price Rs.2 with 800 sample points
- ❖ Now, both samples are mixed, then we obtain Normal (Gaussian) mixture model.



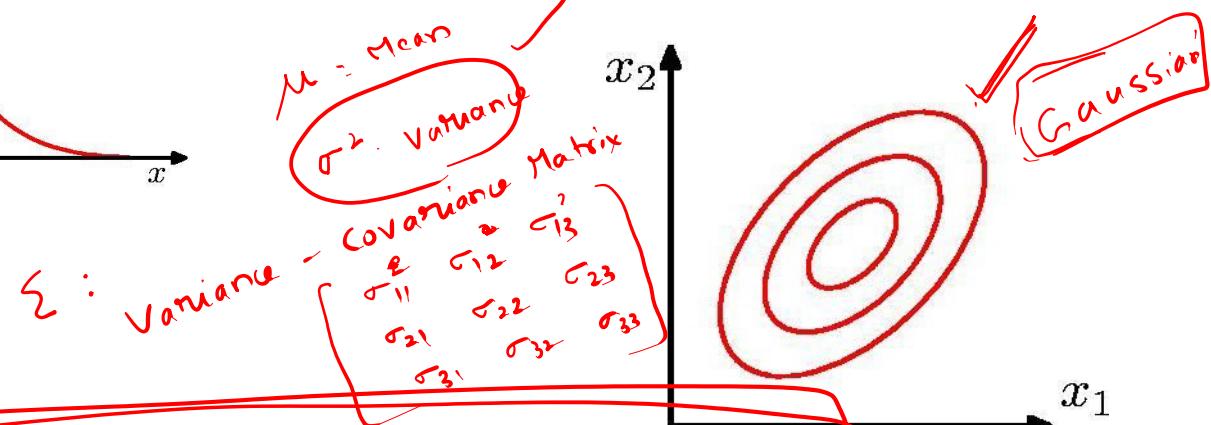
So after mixing the processes together, we have the dataset that we see on the plot. We can notice 2 peaks: around 90 and 100, but for many of the points in the middle of the peaks it is ambiguous to which distribution they were drawn from. So how should we approach this problem?

Gaussian Distribution



$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{1}{2\sigma^2}(x-\mu)^2 \right\}$$

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



$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

Parameter

Mixtures of Gaussians

- Combine simple models into a complex model:

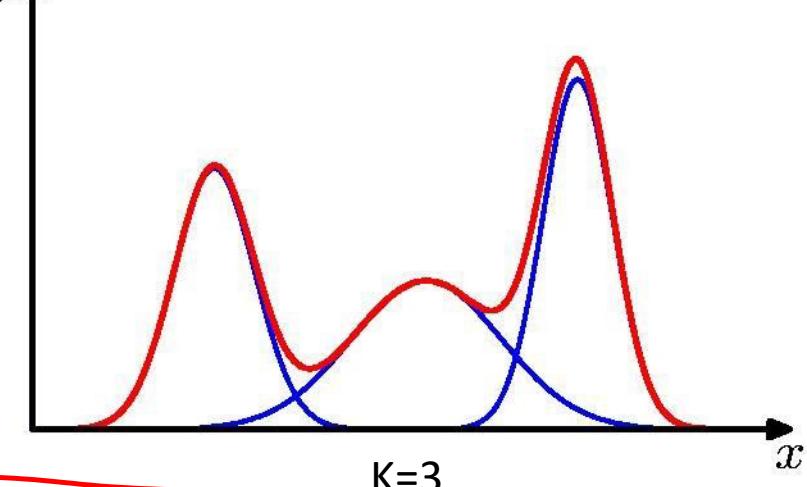
$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Component
 Mixing coefficient

$$\forall k : \pi_k \geq 0$$

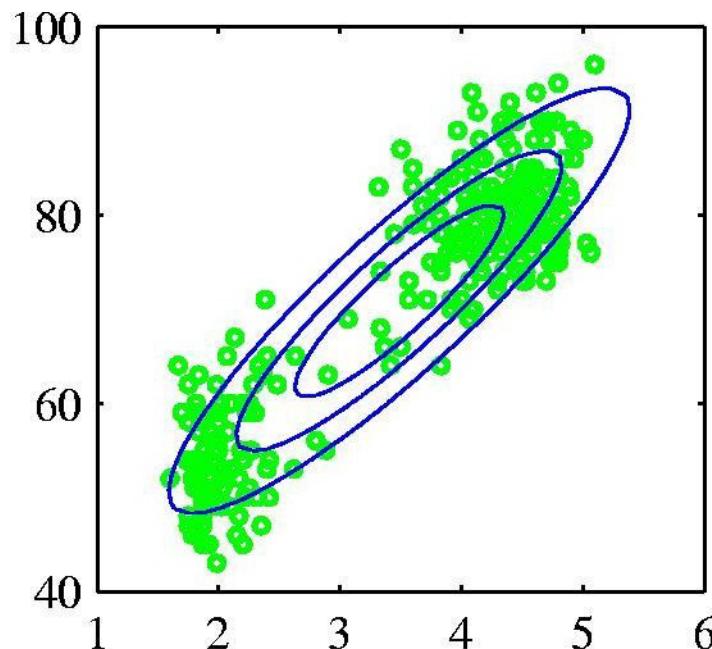
$$\sum_{k=1}^K \pi_k = 1$$

$$p(x)$$

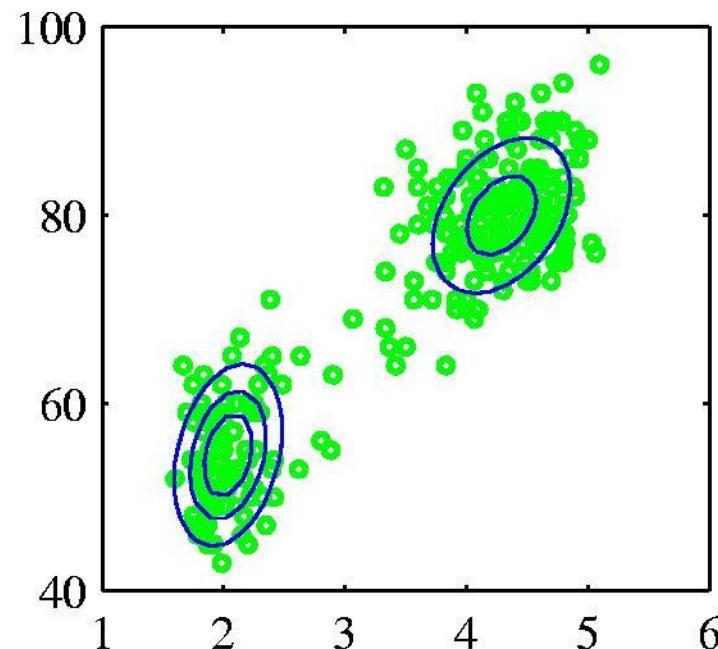


- Find parameters through EM (Expectation Maximization) algorithm

Probabilistic version: Mixtures of Gaussians



Single Gaussian

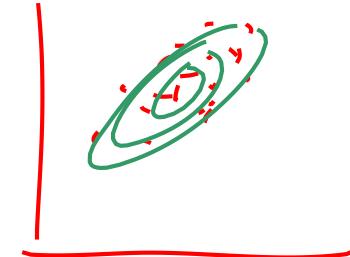


Mixture of two
Gaussians

Gaussian Mixture Model

- Data set

$$D = \{\mathbf{x}_n\} \quad n = 1, \dots, N$$



- Consider first a single Gaussian
- Assume observed data points generated independently

$$p(D|\boldsymbol{\mu}, \Sigma) = \prod_{n=1}^N \mathcal{N}(\mathbf{x}_n|\boldsymbol{\mu}, \Sigma)$$

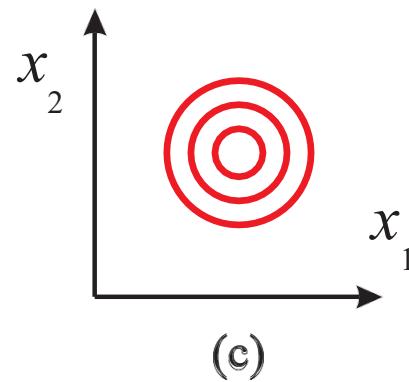
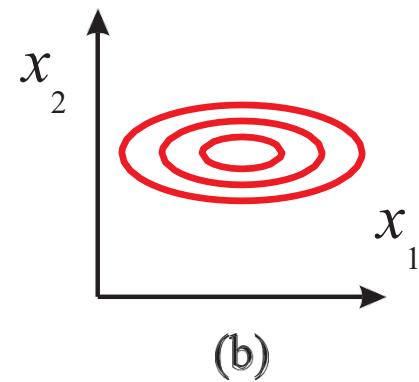
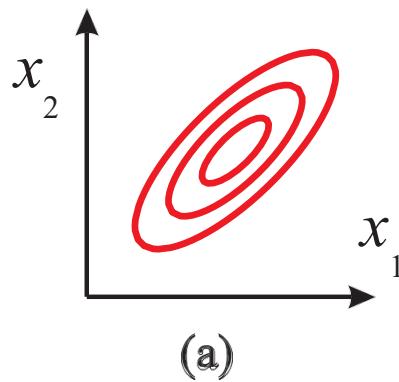
- Viewed as a function of the parameters, this is known as the *likelihood function*

The Gaussian Distribution

- Multivariate Gaussian

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$


The diagram shows two blue arrows pointing upwards from the labels "mean" and "covariance" to the corresponding terms in the equation. The arrow from "mean" points to the term $\boldsymbol{\mu}$, and the arrow from "covariance" points to the term $\boldsymbol{\Sigma}$.



Gaussian Mixture Model

- K-dimensional binary random variable z having a 1-of-K representation in which a particular element z_k is equal to 1 and all other elements are equal to 0.
 - The values of z_k therefore satisfy $z_k \in \{0,1\}$
 - K possible states for the vector z according to which element is nonzero.
 - Joint distribution $p(x,z)$ in terms of a marginal distribution $p(z)$ and a conditional distribution $p(x|z)$,
 - Marginal distribution over z is specified in terms of the mixing coefficients π_k , such that $p(z_k = 1) = \pi_k$
-

Fitting the Gaussian Mixture

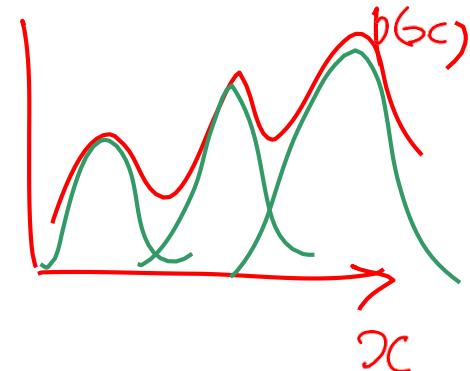
- We wish to invert this process – given the data set, find the corresponding parameters:
 - mixing coefficients ✓
 - means ✓
 - covariances ✓
- If we knew which component generated each data point, the maximum likelihood solution would involve fitting each component to the corresponding cluster
- Problem: the data set is unlabelled
- We shall refer to the labels as *latent* (= hidden) variables

Gaussian Mixture Model



- Linear super-position of Gaussians

$$p(\mathbf{x}) = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$



- Normalization and positivity require

$$\sum_{k=1}^K \pi_k = 1 \quad 0 \leq \pi_k \leq 1$$

- Can interpret the mixing coefficients as prior probabilities

$$p(\mathbf{x}) = \sum_{k=1}^K p(k)p(\mathbf{x}|k)$$

Gaussian Mixture Model

- z uses a 1-of-K representation, we can also write this distribution in the form

$$p(\mathbf{z}) = \prod_{k=1}^K \pi_k^{z_k}$$

- Conditional distribution of x given a particular value for z is a Gaussian

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^K \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}$$

- Joint distribution is given by $p(z)p(x|z)$, and the marginal distribution of x is then obtained by summing the joint distribution over all possible states of z to give

$$p(\mathbf{x}) = \sum_{\mathbf{z}} p(\mathbf{z}) \underbrace{p(\mathbf{x}|\mathbf{z})}_{\text{circled}} = \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Gaussian Mixture Model

- Conditional probability of z given x
- use $\gamma(z_k)$ to denote $p(z_k = 1 | x)$, whose value can be found using Bayes' theorem

$$\begin{aligned}
 \underbrace{\gamma(z_k) \equiv p(z_k = 1 | x)} &= \frac{p(z_k = 1)p(x|z_k = 1)}{\sum_{j=1}^K p(z_j = 1)p(x|z_j = 1)} \\
 &= \frac{\pi_k \mathcal{N}(x|\mu_k, \Sigma_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(x|\mu_j, \Sigma_j)}.
 \end{aligned}$$

✓

- π_k as the prior probability of $z_k = 1$, and the quantity $\gamma(z_k)$ as the corresponding posterior probability once we have observed x .

Maximum Likelihood

Log of likelihood function:

$$\ln p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$

- Maximizing the log likelihood function for a Gaussian mixture model turns out to be a more complex problem than for the case of a single Gaussian.
- The difficulty arises from the presence of the summation over k that appears inside the logarithm, so that the logarithm function no longer acts directly on the Gaussian.

GMM Problems and Solutions

- How to maximize the log likelihood
 - ❖ solved by expectation-maximization (EM) algorithm
 - How to avoid singularities in the likelihood function
 - ❖ solved by a Bayesian treatment
 - How to choose number K of components
 - ❖ also solved by a Bayesian treatment
-

Expectation Maximization (EM) Algorithm



- Conditions for MLE: Setting the derivatives of $\ln p(\mathbf{X} | \pi, \mu, \Sigma)$ in with respect to the means μ_k of the Gaussian components to zero, we obtain

$$0 = - \sum_{n=1}^N \underbrace{\frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}} \boldsymbol{\Sigma}_k (\mathbf{x}_n - \boldsymbol{\mu}_k)$$
$$\gamma(z_{nk})$$

rearranging we obtain:

$$\boldsymbol{\mu}_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}_n$$

where we have defined

$$N_k = \sum_{n=1}^N \gamma(z_{nk})$$

Expectation Maximization (EM) Algorithm

- μ_k for the kth Gaussian component is obtained by taking a weighted mean of all of the points in the data set
- Weighting factor for data point x_n is given by the posterior probability $\gamma(z_{nk})$ that component k was responsible for generating x_n
- If we set the derivative of $\ln p(X|\pi, \mu, \Sigma)$ with respect to Σ_k to 0, and follow a similar line of reasoning, making use of the result for the maximum likelihood solution for the covariance matrix of a single Gaussian

$$\Sigma_k = \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk})(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T$$

Expectation Maximization (EM) Algorithm

- Maximize $\ln p(\mathbf{X}|\pi, \mu, \Sigma)$ with respect to the mixing coefficients π_k with constraint

$$\sum_{k=1}^K \pi_k = 1 \quad 0 \leq \pi_k \leq 1$$

- Using a Lagrange multiplier and maximizing the following quantity

$$\ln p(\mathbf{X}|\pi, \mu, \Sigma) + \lambda \left(\sum_{k=1}^K \pi_k - 1 \right)$$

which gives

$$0 = \sum_{n=1}^N \frac{\mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_j \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)} + \lambda$$

- If we now multiply both sides by π_k and sum over k , we find $\lambda = -N$.
- Rearranging we obtain

$$\pi_k = \frac{N_k}{N}$$

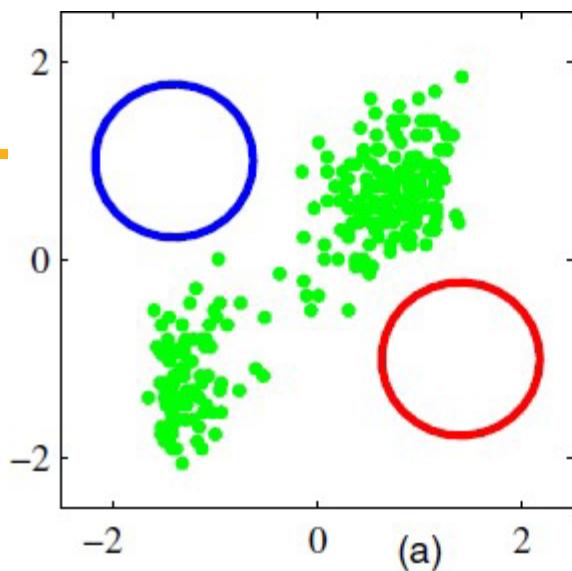
Expectation Maximization (EM) Algorithm

- We first choose some initial values for the means, covariances, and mixing coefficients.
 - Then we alternate between the following two updates that we shall call the E step and the M step
 - In the expectation step, or E step, we use the current values for the parameters to evaluate the posterior probabilities,
 - We then use these probabilities in the maximization step, or M step, to re-estimate the means, covariances, and mixing
 - In practice, the algorithm is deemed to have converged when the change in the log likelihood function, or alternatively in the parameters, falls below some threshold
-

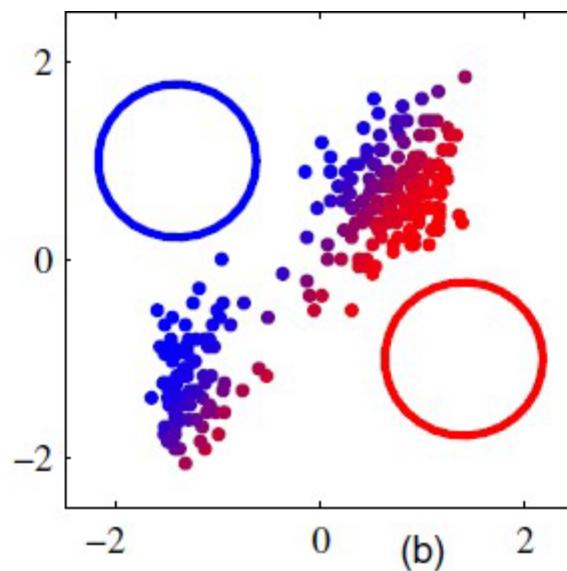
EM Algorithm – Informal Derivation

- The solutions are not closed form since they are coupled
- Suggests an iterative scheme for solving them:
 - make initial guesses for the parameters
 - alternate between the following two stages:
 - E-step: evaluate responsibilities
 - M-step: update parameters using ML results
- Each EM cycle guaranteed not to decrease the likelihood

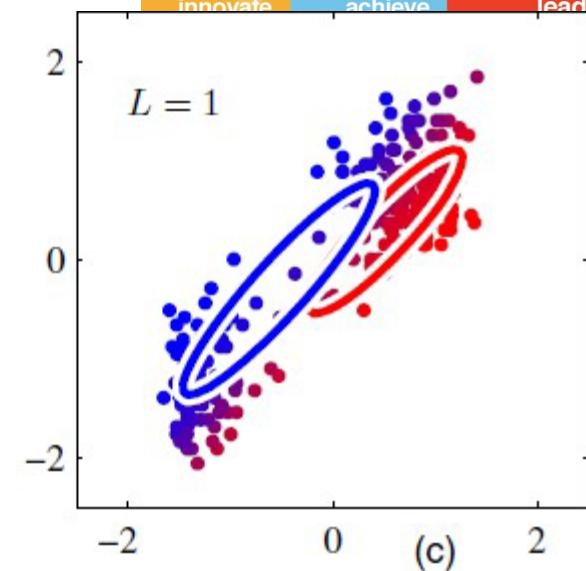
Initialization



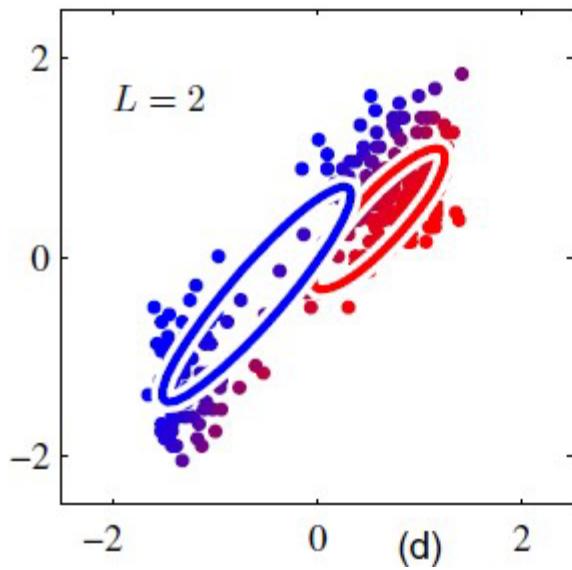
E step



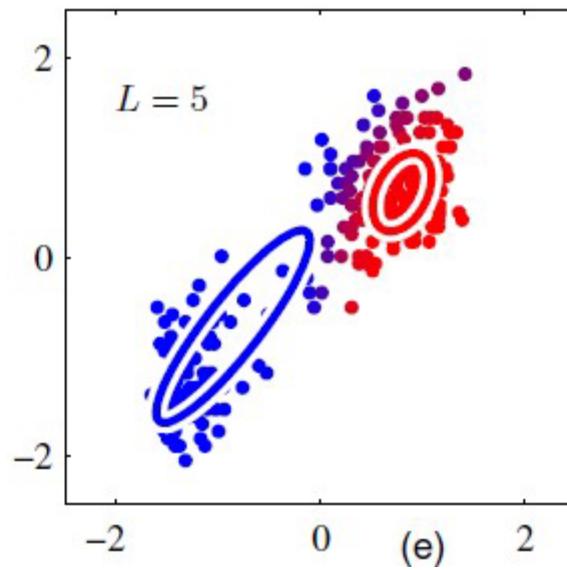
M step



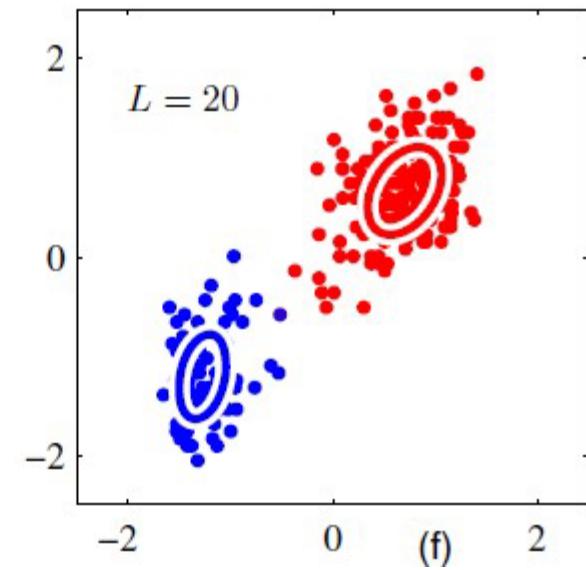
$L = 2$



$L = 5$



$L = 20$



EM algorithm for GMM

1. Initialize the means μ_k , covariances Σ_k and mixing coefficients π_k , and evaluate the initial value of the log likelihood.
2. ~~E step:~~ Evaluate the responsibilities using the current parameter values



$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{j=1}^K \pi_j \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)}$$

Exponential
smoothin
 $F_{t+1} = F_t + (x_t - F_t)^2$

3rd
1
2
3

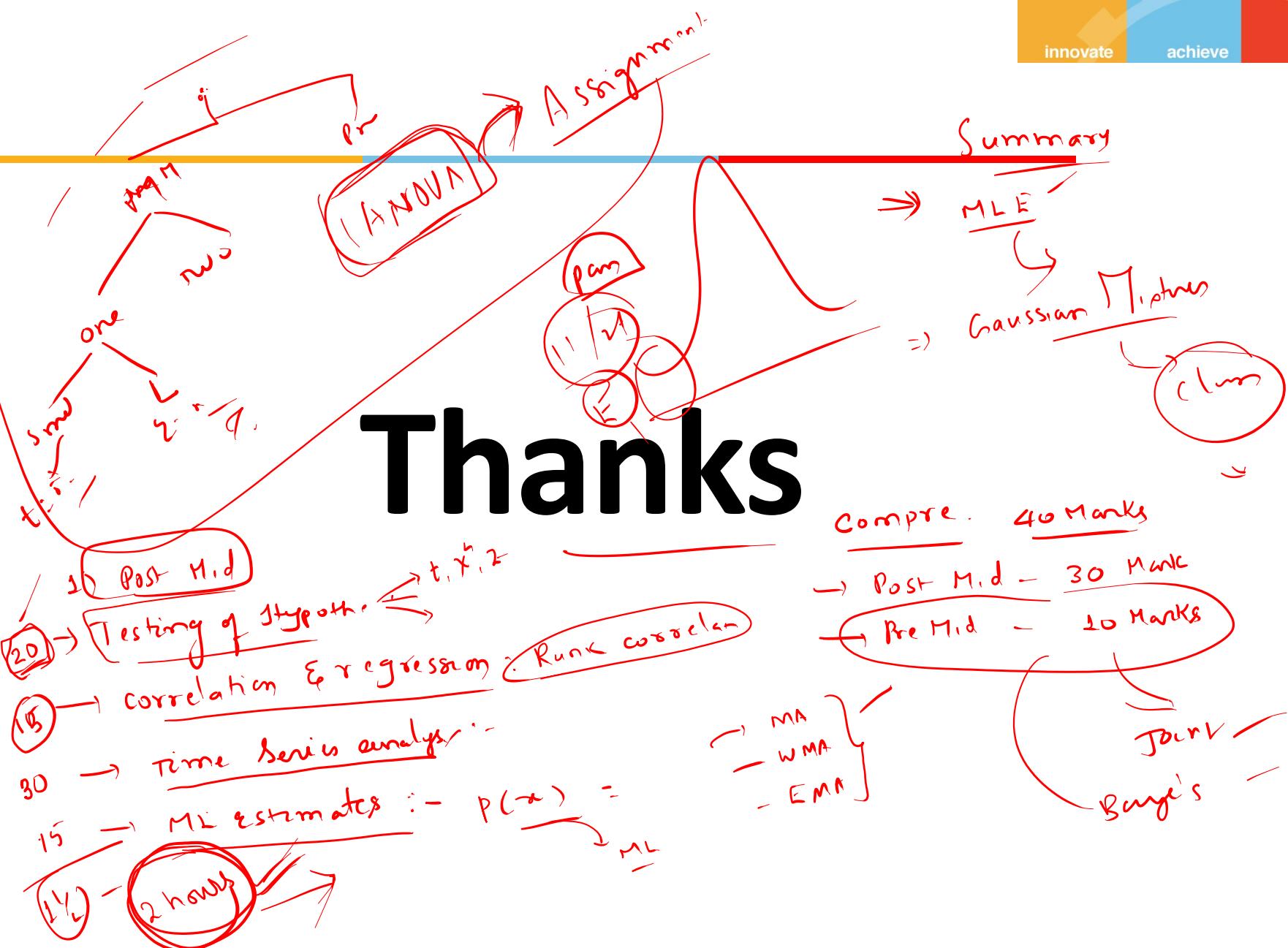
EM algorithm for GMM

3. **M step:** Re-estimate the parameters using the current responsibilities

$$\begin{aligned}
 \mu_k^{\text{new}} &= \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) \mathbf{x}_n \\
 \Sigma_k^{\text{new}} &= \frac{1}{N_k} \sum_{n=1}^N \gamma(z_{nk}) (\mathbf{x}_n - \mu_k^{\text{new}}) (\mathbf{x}_n - \mu_k^{\text{new}})^T \\
 \pi_k^{\text{new}} &= \frac{N_k}{N} \quad \text{where } N_k = \sum_{n=1}^N \gamma(z_{nk})
 \end{aligned}$$

4. Evaluate the log likelihood

$$\ln p(\mathbf{X} | \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\pi}) = \sum_{n=1}^N \ln \left\{ \sum_{k=1}^K \pi_k \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \right\}$$





BITS Pilani

Pilani | Dubai | Goa | Hyderabad

Statistical Methods for Data Science

ISM Team

Dr.Gangaboraiah, PhD (Stats)

- Former Professor of Statistics, KIMS, Bangalore
 - Work Experience
 - Kempegowda Institute of Medical Sciences (KIMS), Bangalore (35 years)
 - Govt. Homeopathy Medical College, Bangalore (5 years)
 - SJC Institute of Technology, Chickballapur (13 years, Visiting Professor)
 - Manipal University, Bangalore Centre (Since 2008, Visiting Professor)
 - MS (Computer Science), MS (Computer Network)
 - Data Science
 - BITS (Since 2013, Visiting Professor)
 - MTech (Data Science)
 - WIPRO and Aricent (2019)
 - Consultant Medical Statistician
 - Public Health Foundation of India
 - Microlabs
- Visiting other University/ Institutions
- Universities
 - Rajiv Gandhi University of Health Sciences, B'llore
 - Dayanand Sagar University, Bangalore
 - Institutions
 - Acharya B M Reddy Institute of Pharmacy, B'llore
 - Al-Ameen College of Pharmacy, Bangalore
 - Krupanidhi College of Nursing, Bangalore
 - RV College of Physiotherapy, B'angalore

Agenda ➔ Here is what you learn in the entire session

1 Definition of Statistics

2 Types of variables, Types of data

3 Scales of measurement

4 Measures of central tendency

5 Measures of dispersion/ variable



Session 7

Testing of hypothesis

Tests based variances

Chi-square – test for single variance

Testing of Hypothesis → One sample Variance (χ^2 – test)

One sample F - test

χ^2 -test



One sample variance distributed as
Chi-square distribution with $n-1$ degrees of freedom

➤ Samples are drawn from normal distribution

➤ The population variance should be known

χ^2 -test

➤ The sample size should be
less than 30 (i.e., $n < 30$)

➤ Subjects should be selected randomly

Testing of Hypothesis → One sample Variance (χ^2 – test)

1 State null and alternative hypothesis

$$\begin{aligned} H_0: \sigma^2 = \sigma_0^2 &\text{ vs } H_1: \sigma^2 < \sigma_0^2 \\ \text{or } H_1: \sigma^2 > \sigma_0^2 \\ \text{or } H_1: \sigma^2 \neq \sigma_0^2 \end{aligned}$$

2 Specify the level of significance ‘ α ’

3 Chi-square - Distribution

4 Compute the test statistic

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} \cong \chi^2_{(\alpha, n-1)}$$

5 Define the critical region/ rejection criteria

6 Conclusion

Testing of Hypothesis → One sample Variance (χ^2 – test)

A manufacturer of car batteries claim that the life of his batteries is approximately normally distributed with a standard deviation equal to 0.9 year. If a random sample of 10 batteries has a standard deviation of 1.2 years, do you think that $\sigma > 0.9$ year? Use a 0.05 level of significance.

Testing of Hypothesis → One sample Variance (χ^2 – test)

At 5% (0.05) level of significance with critical value is 16.919 for 9 degrees of freedom

$$\chi^2 = \frac{(n-1)S^2}{\sigma_0^2} = \frac{9 \times 1.44}{0.81} = 16.00$$

100(1 – α)% CI for σ^2 is

$$\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}, n-1}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}, n-1}}$$

Hypothesis to test

$$H_0: \sigma^2 = \sigma_0^2 = 0.81$$

vs

$$H_1: \sigma_1^2 > \sigma_2^2 > 0.81$$

???

Critical value for $\alpha = 0.05$ is 16.919. Since $\chi^2 = 16.00 < 16.919$, Accept H_0 & Reject H_1

Fisher's F – test for ratio of variances

Testing of Hypothesis → Ratio of two Variance (F – test)

Two sample F – test : Distributed as Fisher's F-distribution

F-test



Ratio of two population variances: σ_1^2/σ_2^2

➤ Samples are drawn from normal distribution

➤ The population variances should be equal

➤ The sample size should be less than 30 (i.e., $n < 30$)



➤ Two groups should be independent

➤ Subjects should be allocated randomly to both groups

Testing of Hypothesis → Ratio of two Variance (F – test)

1 State null and alternative hypothesis

$$\begin{aligned} H_0: \sigma_1^2 &= \sigma_2^2 \text{ vs } H_1: \sigma_1^2 < \sigma_2^2 \\ \text{or } H_1: \sigma_1^2 &> \sigma_2^2 \\ \text{or } H_1: \sigma_1^2 &\neq \sigma_2^2 \end{aligned}$$

2 Specify the level of significance ‘ α ’

3 Fisher’s F - Distribution

4 Compute the test statistic

5 Define the critical region/ rejection criteria

6 Conclusion

$$F = \frac{S_1^2}{S_2^2} \cong F_{(\alpha, n_1-1, n_2-1)}$$

Testing of Hypothesis

Ratio of two Variance (F – test)

The variability in the amount of impurities present in a batch of chemicals used for a particular process depends on the length of time that the process is in operation.

Suppose a sample of size 25 is drawn from the normal process which is to be compared to a sample of a new process that has been developed to reduce the variability of impurities. Test at 5%, whether the variability in the new process is less as compared to the original process.

	Sample 1	Sample 2
n	25	25
S^2	1.04	0.51

Testing of Hypothesis → Ratio of two Variance (F – test)

At 5% (0.05) level of significance with critical value is 1.98 for (24, 24) degrees of freedom

$$F = \frac{S_1^2}{S_2^2} = \frac{1.04}{0.51} = 2.04$$

100(1 – α)% CI for $\frac{\sigma_1^2}{\sigma_2^2}$ is

$$\frac{S_1^2}{S_2^2} f_{1-\frac{\alpha}{2}, n_2-1, n_1-1} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{S_1^2}{S_2^2} f_{\frac{\alpha}{2}, n_2-1, n_1-1}$$

Hypothesis to test

$$\begin{aligned} H_0: \sigma_1^2 &= \sigma_2^2 \\ \text{vs} \\ H_1: \sigma_1^2 &> \sigma_2^2 \end{aligned}$$

???

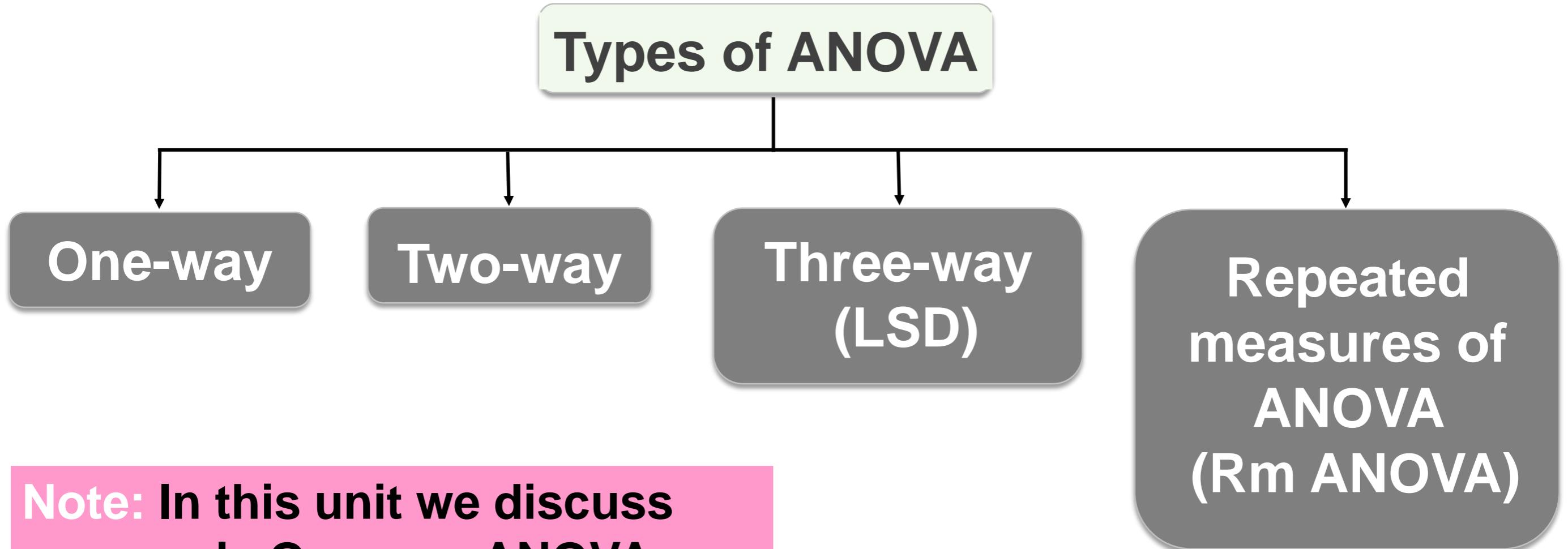
Critical value for $\alpha = 0.05$ is 1.98. Since $F = 2.04 > 1.98$, Reject H_0 & Accept H_1

Testing of Hypothesis → Ratio of two Variance (F – test)

A company manufactures impellers for use in jet-turbine engines. One of the operations involves grinding a particular surface finish of a titanium alloy component. Two different grinding processes can be used and both processes can produce parts at identical mean surface roughness. The manufacturing engineer would like to select the process having the least variability in surface roughness. A random sample of $n_1 = 12$ parts from the first process results in a sample standard deviation of $s_1 = 5.1$ microinches of $n_2 = 15$ parts from the second process results in sample standard deviation of $s_2 = 4.7$ microinches. Test at 5%, is there a sufficient evidence that the first process vary more than the second process?

Analysis of Variance (ANOVA)

Testing of Hypothesis → Analysis of Variance (ANOVA)



Note: In this unit we discuss
only One-way ANOVA
and Two ANOVA in detail

Testing of Hypothesis → Why Analysis of Variance (ANOVA)

Student's t-test cannot be applied



No. of groups are more than two (say k) and are independent



If t-test is applied, the type-I error will increase

ANOVA Used to test equality of more than two population means against not equal

Testing of Hypothesis → One-way Analysis of Variance

ANOVA

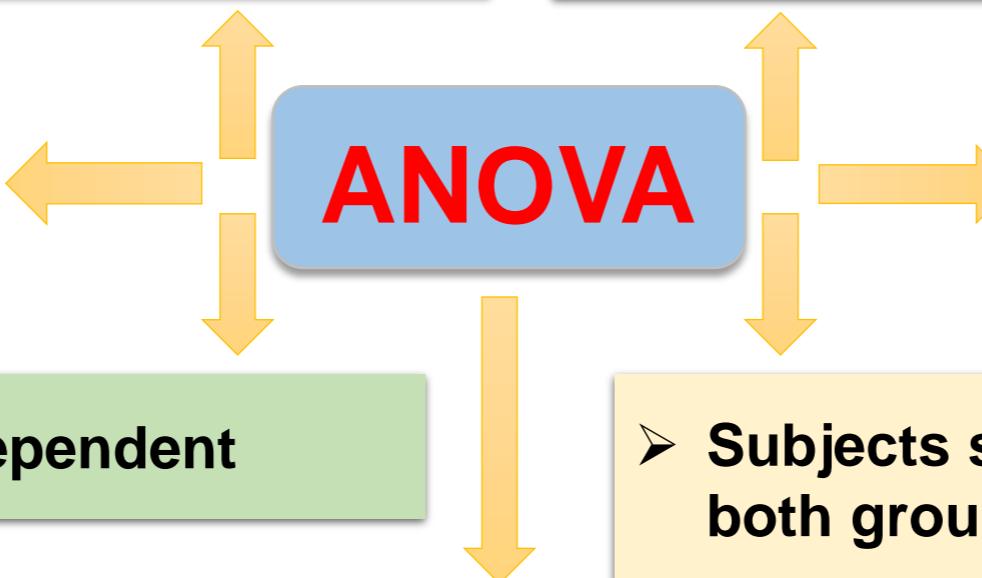


Testing equality of k group means against not equal

➤ Samples are drawn from normal population

➤ The population variances should be equal

➤ The sample size should be less than 30 (i.e., $n < 30$)



➤ Groups should be independent

➤ Subjects should be allocated randomly to both groups

➤ However even if sample size more than 30 (i.e., $n > 30$) ANOVA should be continue to apply, because of central limit theorem it approaches normal.

Testing of Hypothesis → One-way Analysis of Variance

G_1	G_2	G_3	.	.	G_k
X_{11}	X_{21}	X_{31}	.	.	X_{k1}
X_{12}	X_{22}	X_{32}	.	.	X_{k2}
X_{13}	X_{23}	X_{33}	.	.	X_{k3}
.
.
.
X_{1n_1}	X_{1n_2}	X_{1n_3}	.	.	X_{1n_k}
C_1	C_2	C_3	.	.	C_k

$$n_1 + n_2 + n_3 + \dots + n_k = n$$

$$C_1 + C_2 + C_3 + \dots + C_k = G$$

The hypotheses to
be tested are

$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

vs

$$H_1: \mu_1 \neq \mu_2 \neq \dots \neq \mu_k$$

Testing of Hypothesis → One-way Analysis of Variance

1 State null and alternative hypothesis

$$H_0 : \mu_1 = \mu_2 = \dots = \mu_k$$

vs

2 Specify the level of significance ‘ α ’

$$H_1 : \mu_1 \neq \mu_2 \neq \dots \neq \mu_k$$

3 Student’s t-distribution

4 Compute the test statistic

$$F = \frac{\text{MGSS}}{\text{MWSS}} \approx F_{(k-1, n-k)}$$

5 Define the critical region/ rejection criteria

MGSS- Mean group sum of squares

6 Conclusion

MWSS- Mean within group sum of squares

Testing of Hypothesis → One-way Analysis of Variance

Calculation of sum of squares

1. Correction factor (CF) : $\frac{G^2}{n}$

2. Total sum of squares (TSS) : $\sum_i \sum_j x_{ij}^2 - CF$

3. Between group sum of squares (GSS) : $\sum_{i=1}^k \frac{C_i^2}{n_i} - CF$

4. Within group sum of squares (WSS) : TSS - GSS

Testing of Hypothesis → One-way Analysis of Variance

One-way ANOVA table

Source of variation	df	Sum of squares	Mean sum of squares	F-ratio
Between groups	$k - 1$	GSS	$MGSS = \frac{GSS}{k - 1}$	$F = \frac{MGSS}{MWSS}$
Within groups	$n - k$	WSS	$MWSS = \frac{WSS}{n - k}$	
Total	$n - 1$	TSS	$F \approx F$ - distribution with $k - 1$ and $n - k$ df	

Testing of Hypothesis → One-way Analysis of Variance

χ Reject H_0 , if $F > F_{(\alpha, k-1, n-k)}$

χ If H_0 is rejected, then further mean differences between any two groups should be tested using Post-hoc tests. Following are different Post-hoc tests.

φ Bonferroni's test

φ Tucky's HSD test

φ Schaeffe's test

φ Duncan's test

φ Least Significant Difference (LSD) test

Testing of Hypothesis → 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

φ

$$\sqrt{S^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)}, i, j = 1, 2, \dots, k$$

φ

Where S^2 is the 'Within groups mean sum of squares and n_i and n_j are the number of observations in i^{th} and j^{th} groups under comparison, k is the number of groups

Testing of Hypothesis → One-way Analysis of Variance

Test at 5% level of significance
is there any significant difference in mean iron intake among four groups of patients?
Also test if significant, which group means have contributed to the difference in means using LSD test

A clinical trial Iron intake of four groups of patients (mg)

Group 1	Group 2	Group 3	Group 4
11.5	19.5	18.5	30.0
12.5	18.5	16.5	26.5
18.5	16.0	24.5	27.0
21.0	22.0	30.0	34.0
28.0	30.0	28.5	20.0
26.0	24.5	14.0	22.5
14.0	19.0	19.0	28.0
22.0	24.0	17.0	32.0
20.0	19.5	18.0	27.0
22.0	15.0	29.0	25.5

Testing of Hypothesis → One-way Analysis of Variance

Calculation of sum of squares

$$1. \text{ Correction factor (CF)} : \frac{G^2}{n} = \frac{981^2}{40} = 19847.03$$

$$2. TSS = \sum_i \sum_j x_{ij}^2 - CF = 21119.5 - 19847.03 = 1272.47$$

$$3. GSS = \sum_{i=1}^k \frac{C_i^2}{n_i} - CF = 20196.6 - 19847.03 = 349.52$$

$$4. WSS = TSS - GSS = 1272.47 - 349.52 = 922.95$$

Testing of Hypothesis → One-way Analysis of Variance

One-way ANOVA table

Source of variation	df	Sum of squares	Mean sum of squares	F-ratio
Between groups	3	349.52	116.51	F=4.455
Within groups	36	922.95	25.64	
Total	39	1272.47	$F(4.46) > F_{(0.05; 3, 36)} = 4.38$	

H_0 may be rejected and H_1 may be accepted, and continue with post-hoc test.

Testing of Hypothesis → One-way Analysis of Variance

χ

S^2 = Within Groups Mean sum of squares
= 25.6375

χ

k = Number of observations in each Group = 10 df of
within villages = 36

φ

t - Statistic value corresponding to $\alpha = 0.05$ for 36 df is
2.03

φ

The LSD is

$$\left\{ t_{0.05} \sqrt{S^2 \left(\frac{1}{n_i} + \frac{1}{n_j} \right)} \right\}$$

Testing of Hypothesis → One-way Analysis of Variance

x

$$\left\{ 2.03 \sqrt{25.64 \left(\frac{1}{10} + \frac{1}{10} \right)} \right\} = 4.597$$

x

Group of patients			
1	2	3	4
Mean iron intake (mg)			
20	21	22	28

φ

It can be seen that only the difference between Group 4 is different from other groups as only this difference is more than 4.597. Hence, Group 4 makes the difference in significance.

Testing of Hypothesis → 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	17	14	13	10	11

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?

Testing of Hypothesis → One-way Analysis of Variance

Calculation of sum of squares

$$1. \text{ Correction factor (CF)} : \frac{G^2}{n} = \frac{179^2}{15} = 2136.067$$

$$2. TSS = \sum_i \sum_j x_{ij}^2 - CF = 2219.5 - 2136.067 = 82.933$$

$$3. GSS = \sum_{i=1}^k \frac{C_i^2}{n_i} - CF = 2164.167 - 2136.067 = 28.1$$

$$4. WSS = TSS - GSS = 82.933 - 28.1 = 54.833$$

Testing of Hypothesis → 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	28.1	14.05	F=3.078
Within groups	12	54.83	4.569	
Total	14	82.93	$F(3.078) < F_{(0.05; 2, 12)} = 3.885$	

H_0 may be not rejected and H_1 may be rejected (no need for post-hoc test)

P – value = 0.084

Testing of Hypothesis → 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?

Testing of Hypothesis → One-way Analysis of Variance

Calculation of sum of squares

$$1. \text{ Correction factor (CF)} : \frac{G^2}{n} = \frac{140^2}{15} = 1306.667$$

$$2. TSS = \sum_i \sum_j x_{ij}^2 - CF = 1454 - 1306.667 = 147.333$$

$$3. GSS = \sum_{i=1}^k \frac{C_i^2}{n_i} - CF = 1416.667 - 1306.667 = 110$$

$$4. WSS = TSS - GSS = 147.333 - 110 = 37.333$$

Testing of Hypothesis → 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	110	55.000	F=17.679
Within groups	12	37.333	3.111	
Total	14	147.667	$F_{obs}(17.679) > F_{(0.05; 2, 12)} = 3.885$	

H_0 may be rejected and H_1 may be accepted.

P – value = 0.0003

Testing of Hypothesis → One-way Analysis of Variance

Post-hoc test: Multiple comparison using Bonferroni's test (SPSS output)

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	1.183	0.026*	0.42	5.58
Formula A	Formula C	3.667	1.068	0.005*	1.34	5.99
Formula B	Formula C	6.667	1.139	< 0.001*	4.19	9.15

*The mean difference is significant at the 0.05 level.

Testing of Hypothesis → Two-way Analysis of Variance

A clinical trial on Iron intake of four groups of patients (mg). Test at 5% level of significance is there any significant difference in mean iron intake among ten groups of patients as well as between three trimesters? Also test if significant, which group means have contributed to the difference in means using LSD test

Trimester	Group means for iron intake										Row Total
	G ₁	G ₂	G ₃	G ₄	G ₅	G ₆	G ₇	G ₈	G ₉	G ₁₀	
I	11.5	19.5	18.5	12.5	18.5	16.5	26.5	18.5	16.0	24.5	182.5
II	27.0	28.0	22.0	21.0	15.0	19.5	20.0	26.0	30.0	28.5	237.0
III	28.0	30.0	26.0	30.0	24.5	28.5	26.0	30.0	27.0	25.5	275.5
Column Total	66.5	77.5	66.5	63.5	58.0	64.5	72.5	74.5	73.0	78.5	695

Testing of Hypothesis → Two-way Analysis of Variance

SUMMARY	Average	SD	Variance
Trimester 1	18.25	4.66	21.74
Trimester 2	23.70	4.88	23.84
Trimester 3	27.55	2.05	4.19
Group 1	22.17	9.25	85.58
Group 2	25.83	5.58	31.08
Group 3	22.17	3.75	14.08
Group 4	21.17	8.75	76.58
Group 5	19.33	4.80	23.08
Group 6	21.50	6.24	39.00
Group 7	24.17	3.62	13.08
Group 8	24.83	5.84	34.08
Group 9	24.33	7.37	54.33
Group 10	26.17	2.08	4.33

Source of Variation	df	SS	MS	F	P-value	F crit
Trimester	2	436.72	218.36	12.53	0.0003	3.55
Group	9	134.17	14.91	0.86	0.58	2.46
Error	18	313.78	17.43			
Total	29	884.67				

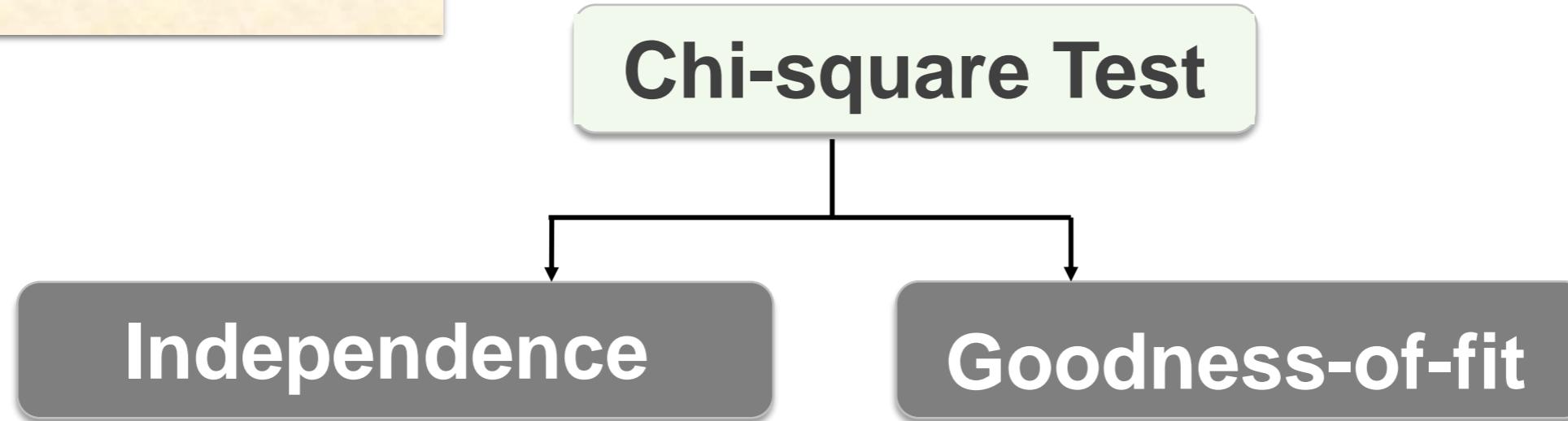
Φ

The mean difference of trimester varies significantly ($F=12.53 > F_{(0.05; 2, 18)} = 3.55$) and the mean group difference is not significant ($F=0.86 > F_{(0.05; 9, 18)} = 2.46$) . Carry out Post-hoc test for trimester

Chi-square test

Testing of Hypothesis → Chi-square test: Independence

Chi-square test



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

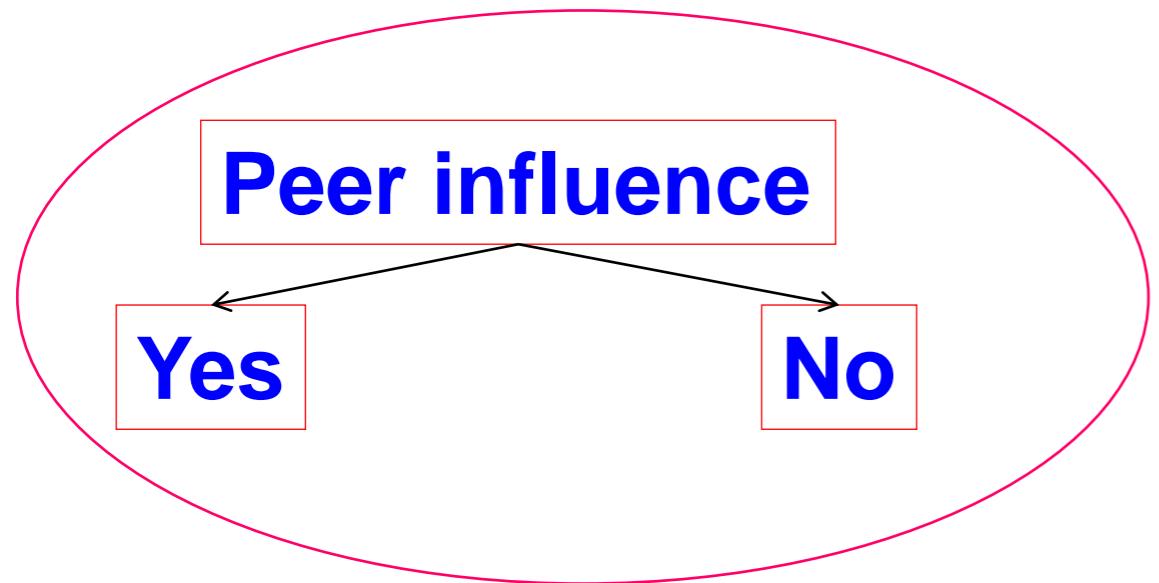
Caution:

Should be applied **ONLY** for **FREQUENCIES** not for

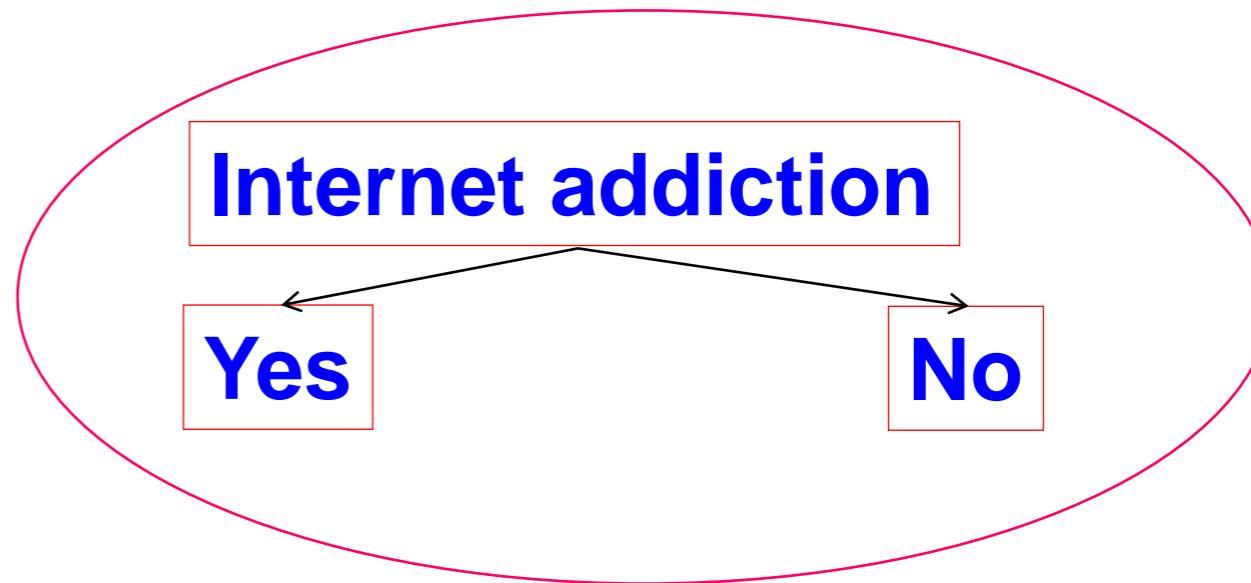
percentages, ratios, mean etc

Testing of Hypothesis → Chi-square test: Independence

Categorical Variable 1



Categorical Variable 2



Chi-square is the right choice to test whether the
two variables are related or not

Testing of Hypothesis → Chi-square test: Independence

Categorical Variable 1	Categorical Variable 1		Row Total
	Response 1	Response 2	
Response 1	O_1	O_2	r_1
Response 2	O_3	O_4	r_2
Column Total	C_1	C_2	n

Testing of Hypothesis → Chi-square test: Independence

A study to find the independence (not associated) between smoking and ca. lung has revealed the following data. Find is there any association exists between smoking and ca. lung?

Testing of Hypothesis → Chi-square test: Independence

Smoking	Cancer of lung		Row Total
	Present	Absent	
Yes	69	2431	2500
No	24	1476	1500
Column Total	93	3907	4000

Testing of Hypothesis → Chi-square test: Independence

2 x 2 Contingency Table for testing independence

Categorical Variable 1	Categorical variable 2		Total
	Present	Absent	
Present	O_1 E_1	O_2 E_2	r_1
Absent	O_3 E_3	O_4 E_4	r_2
Total	c_1	c_2	n

Calculation
of expected
frequencies

$$E_1 = \frac{r_1 c_1}{n}$$

$$E_3 = \frac{r_2 c_1}{n}$$

$$E_2 = \frac{r_1 c_2}{n}$$

$$E_4 = \frac{r_2 c_2}{n}$$

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

Assumptions of Chi-square test

If the expected cell frequencies is < 5

Yate's correction should be applied for continuity

In a **2 x 2 contingency table**, if one or more of the cell has the expected cell frequencies is < 5 ,

Fisher's exact probabilities should be computed

For the use of **Chi-square test**

The sample size should not be **less than 20**.

The Fisher's exact Probability



$$P = \frac{1}{n!} \frac{r_1!}{a!} \frac{r_2!}{b!} \frac{c_1!}{c!} \frac{c_2!}{d!}$$

For an $r \times c$ table, if the expected frequencies in any cells are < 5 , merge the rows or columns meaningfully

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 - χ^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 χ^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

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

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

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 x 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

Alternate formula for calculation of chi-square statistic

When the expected frequencies in all the cells are more than 5, alternatively Chi-square statistic can be calculated using the formula for 2 x 2 table only by

$$\chi^2 = n \frac{(ad - bc)^2}{r_1 r_2 c_1 c_2}$$

Where a, b, c, d are cell frequencies; r_1 and r_2 are row totals; c_1 and c_2 are column totals

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 - χ^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	-	-	-
Total	70	70			???

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

H_0 : Type of insurance plan and length of hospital stay may be independent

H_1 : Type of insurance plan and length of hospital stay may be associated

Computed Probability

$$P = \frac{1}{70!} \frac{45!}{42!} \frac{25!}{3!} \frac{60!}{18!} \frac{10!}{7!}$$

One tailed: P=0.0201

Two tailed: P=0.0282

Conclusion: H_0 may be rejected and hence the type of insurance plan and length at may not be independent (may be associated)

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

$$E_1 = \frac{r_1 c_1}{n} ,$$

$$E_2 = \frac{r_1 c_2}{n} ,$$

$$E_3 = \frac{r_1 c_3}{n} ,$$

$$E_4 = \frac{r_2 c_1}{n} ,$$

$$E_5 = \frac{r_2 c_2}{n} ,$$

$$E_6 = \frac{r_2 c_3}{n} , \text{ and}$$

$$\chi^2 = \sum_{i=1}^6 \frac{(O_i - E_i)^2}{E_i}$$

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

Testing of Hypothesis → Chi-square test: Goodness-of-fit

A powerful test for testing the significance of the discrepancy between theory and experiment was given by Karl Pearson known as “Chi-square test for Goodness – of – fit”. It enables to find the deviation of the experiment from theory is just by Chance or is it really due to the inadequacy of the theory to fit the observed data.

Testing of Hypothesis → Chi-square test: Goodness-of-fit

If O_i ($i = 1, 2, \dots, n$) is a set of observed (experimental) Frequencies and E_i ($i = 1, 2, \dots, n$) is the corresponding set of expected (theoretical or hypothetical) frequencies, then the Chi-square test statistic is given by

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} \approx \chi^2_{(n-1)}$$

follows Chi-square distribution with $n - 1$ degree of freedom.

Testing of Hypothesis → Chi-square test: Goodness-of-fit

Example

The following data shows the distribution of digits in numbers chosen at random from a telephone directory. The digits are:

Digits	0	1	2	3	4	5	6	7	8	9	Total
f	1026	1107	997	966	1075	933	1107	972	954	853	10000

H_0 : The digits occur uniformly frequently in the directory

H_1 : The digits do not occur uniformly frequently in the directory

Testing of Hypothesis → Chi-square test: Goodness-of-fit

Under the null hypothesis, the expected frequency for each of the digits 0, 1, ..., 9 is $10000 \div 10 = 1000$. The Chi-square value is

$$\chi^2 = \sum_{i=1}^n \frac{(O_i - E_i)^2}{E_i} = \frac{(1026 - 1000)^2}{1000} + \dots + \frac{(853 - 1000)^2}{1000}$$

$$\chi^2 = 58.542$$

Since $\chi^2 = 58.542 > 16.919$ (critical value, $df = 9$), it can be infer that the digits are not uniformly distributed.

Testing of Hypothesis → Chi-square test: Goodness-of-fit

Jaswant is interested in breeding flowers of a certain species. The experimental breeding can result in four possible types of flowers

- (a) Megenta flowers with green stigma (MG)
- (b) Megenta flowers with red stigma (MR)
- (c) Red flowers with green stigma (RG)
- (d) Red flowers with red stigma (RR)

Testing of Hypothesis → Chi-square test: Goodness-of-fit

According to Mendel's law, these four kinds of flowers should come out in the ratio of 9 : 3 : 3 : 1. Jaswant found that under her experiment, out of 160 flowers that bloomed the number of flowers with types MG, MR, RG, and RR were 84, 35, 28, and 13. She wants to find out, whether these data are compatible with Mendel's law. Use $\alpha = 0.05$.

Testing of Hypothesis → Chi-square test: Goodness-of-fit

Solution:

$$p_1 = \frac{9}{16}$$

$$p_2 = \frac{3}{16}$$

$$p_3 = \frac{3}{16}$$

$$p_4 = \frac{1}{16}$$

H₀: The distribution of flowers may follow multinomial distribution

H₁: The distribution of flowers may not follow multinomial distrn.

Flowers type	O _i	p _i	E _i	$\frac{(O_i - E_i)^2}{E_i}$
MG	84	0.5625	90	0.4000
MR	35	0.1875	30	0.8333
RG	28	0.1875	30	0.1333
RR	13	0.0625	10	0.9000
$\chi^2 =$				2.2667

Since $\chi^2 = 2.267 > 7.81$ (critical value at df = 3), it can be inferred that her experiment may be compatible with Mendel's law.

Testing of Hypothesis → Chi-square test: Goodness-of-fit

A consultant was employed by a city council to study the pattern of bus arrival and departure at a very busy interstate bus terminus. She collected data from the arrival of 200 buses. Based on the data, the average arrival time was found to be $\lambda = 2.96$. She divided the arrivals into 6 categories. Assuming that the arrivals follow Poisson distribution test whether the arrival distribution follows Poisson law. Use $\alpha = 0.01$.

Testing of Hypothesis → Chi-square test: Goodness-of-fit

The probabilities are to be calculated using Poisson distribution with $\lambda = 2.96$ and $x = 0, 1, 2, 3, 4$, and ≥ 5 . The results are as follows:

No. of arrivals	O_i	p_i	E_i	$\frac{(O_i - E_i)^2}{E_i}$
0	10	0.0524	10.48	0.0220
1	13	0.1545	30.90	10.3693
2	45	0.2277	45.54	0.0064
3	49	0.2238	44.76	0.4016
4	32	0.1651	33.02	0.0315
≥ 5	41	0.1765	35.30	0.9204
$\chi^2 =$				11.7512

Since $\chi^2 = 3.402 < 13.27$ (critical value at $df = k-2 = 4$), it can be infer that the arrivals and departures follow Poisson law.

Testing of Hypothesis → Chi-square test: Goodness-of-fit

A chemical company wishes to know if its sales of a liquid chemical are normally distributed. This information will help them in planning and controlling the inventory. The sales record for a random sample of 200 days are as follow: Using a 5% level test whether the company's sales normally distributed. The sample mean and sample standard deviation are 40 and 2.5 respectively.

Sales in '000 liters	< 34	34.0 - 35.5	35.5 - 37.0	37.0 - 38.5	38.5 - 40.0	40.0 - 41.5	41.5 - 43.0	43.0 - 44.5	44.5 - 46.0	> 46
No. of days	2	13	20	35	43	51	27	10	1	0

Testing of Hypothesis → Chi-square test: Goodness-of-fit

Sales in '000 litres	No. of days (O_i)
< 34	2
34.0 - 35.5	13
35.5 - 37.0	20
37.0 - 38.5	35
38.5 - 40.0	43
40.0 - 41.5	51
41.5 - 43.0	27
43.0 - 44.5	10
44.5 - 46.0	1
> 46	0

Testing of Hypothesis → Chi-square test: Goodness-of-fit

Solution:

To compute expected frequencies, find Z-value and identify the corresponding probability from Z-table.

For example, in case of first class interval $Z = \frac{X - \mu}{\sigma} = \frac{34 - 40}{2.5} = -2.4$

$$P(34 \leq X) P(-2.4 \leq Z \leq 0) = P(0 \leq Z \leq 2.4) = 0.4918$$

$$p_1 = 0.5 - 0.4918 = 0.0082$$

$$E_1 = 200 \times 0.0082 = 1.64$$

Testing of Hypothesis → Chi-square test: Goodness-of-fit

Sales in '000 litres	No. of days (O _i)	Class probabilities (p _i)	Exp. Freq. (E _i)	$\frac{(O_i - E_i)^2}{E_i}$	$\frac{(O_i - E_i)^2}{E_i}$ adjusted
< 34	2	0.0082	1.64	0.0790	4.7176
34.0 - 35.5	13	0.0277	5.54	10.0454	1.0925
35.5 - 37.0	20	0.0792	15.84	0.3136	0.3136
37.0 - 38.5	35	0.1592	31.84	0.1015	0.1015
38.5 - 40.0	43	0.2257	45.14	0.7607	0.7607
40.0 - 41.5	51	0.2257	45.14	0.7357	0.7357
41.5 - 43.0	27	0.1592	31.84	2.1531	2.1531
43.0 - 44.5	10	0.0792	15.84	3.7205	5.3193
44.5 - 46.0	1	0.0277	5.54		
> 46	0	0.0082	1.64		

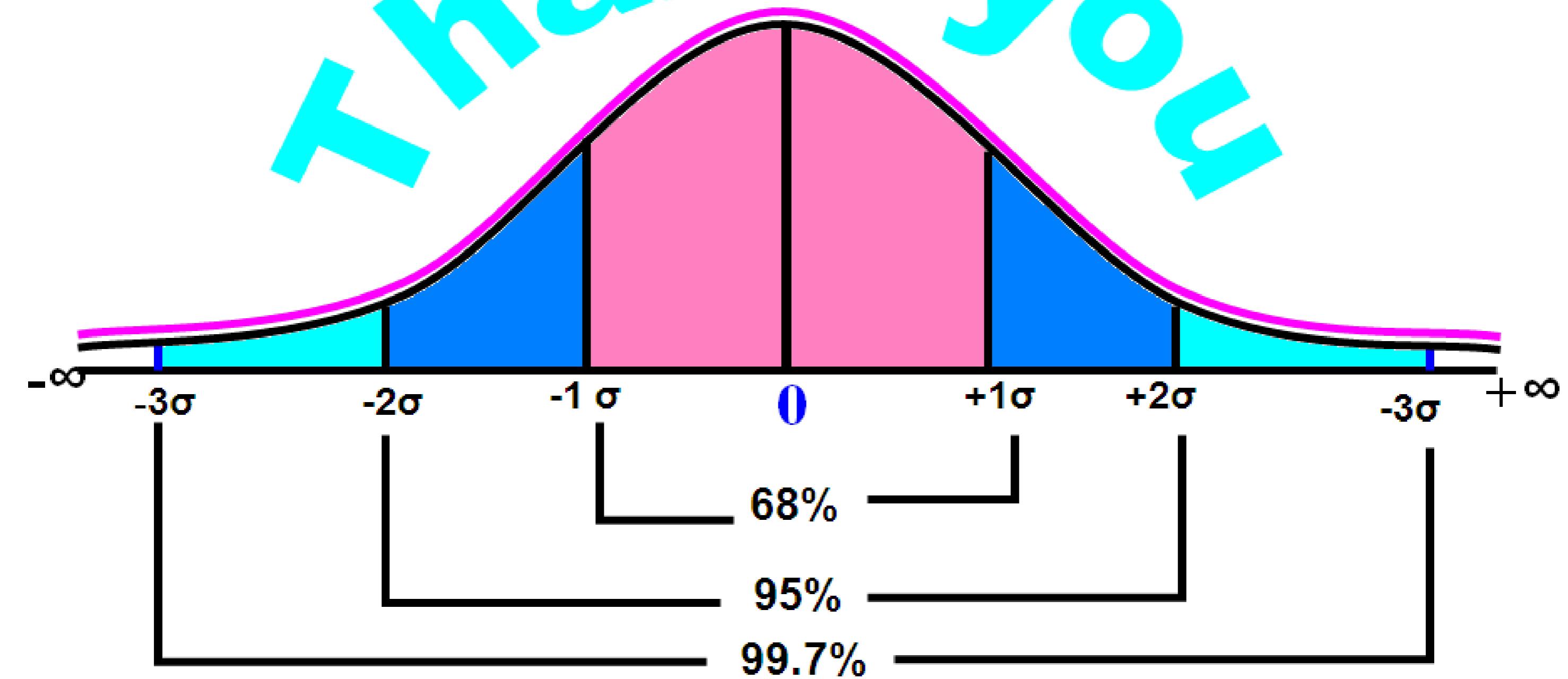
Testing of Hypothesis → Chi-square test: Goodness-of-fit

In the first and last class intervals, the expected frequencies are < 5 and hence, the class observed frequencies are to be merged to get expected frequency ≥ 5 to incorporate continuity correction for Chi-square.

The χ^2_{obs} - value is 15.194 and the critical value is 11.07. The degrees of freedom is $(k-1)-2 = 5$

Since $\chi^2_{\text{obs}} (15.194) > \chi^2_{(0.05, 5)} (11.07)$, H_0 may be rejected and the data may not follow normal distribution fit.

Thank you



Tables Appendix

- Table 1** Binomial Distribution Cumulative Probabilities T-2
- Table 2** Poisson Distribution Cumulative Probabilities T-4
- Table 3** Standard Normal Distribution Cumulative Probabilities T-7
- Table 4** Standardized Normal Scores T-9
- Table 5** Critical Values for the *t* Distribution T-10
- Table 6** Critical Values for the Chi-Square Distribution T-11
- Table 7** Critical Values for the *F* Distribution T-13
- Table 8** Critical Values for the Studentized Range Distribution T-16
- Table 9** Critical Values for the Wilcoxon Signed-Rank Statistic T-19
- Table 10** Critical Values for the Wilcoxon Rank-Sum Statistic T-22
- Table 11** Critical Values for the Runs Test T-25
- Table 12** Greek Alphabet T-27

T-2 Tables Appendix

Table 1 Binomial Distribution Cumulative Probabilities

Let X be a binomial random variable with parameters n and p : $X \sim B(n, p)$. This table contains cumulative probabilities:

$$P(X \leq x) = \sum_{k=0}^x P(X = k) = P(X = 0) + P(X = 1) + P(X = 2) + \cdots + P(X = x).$$

$n = 5$		p														
x		0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99
0	0.9510	0.7738	0.5905	0.3277	0.2373	0.1681	0.0778	0.0313	0.0102	0.0024	0.0010	0.0003	0.0000			
1	0.9990	0.9774	0.9185	0.7373	0.6328	0.5282	0.3370	0.1875	0.0870	0.0308	0.0156	0.0067	0.0005	0.0000		
2	1.0000	0.9988	0.9914	0.9421	0.8965	0.8369	0.6826	0.5000	0.3174	0.1631	0.1035	0.0579	0.0086	0.0012	0.0000	
3		1.0000	0.9995	0.9933	0.9844	0.9692	0.9130	0.8125	0.6630	0.4718	0.3672	0.2627	0.0815	0.0226	0.0010	
4			1.0000	0.9997	0.9990	0.9976	0.9898	0.9688	0.9222	0.8319	0.7627	0.6723	0.4095	0.2262	0.0490	

$n = 10$		p															
x		0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99	
0	0.9044	0.5987	0.3487	0.1074	0.0563	0.0282	0.0060	0.0010	0.0001	0.0000							
1	0.9957	0.9139	0.7361	0.3758	0.2440	0.1493	0.0464	0.0107	0.0017	0.0001	0.0000	0.0000					
2	0.9999	0.9885	0.9298	0.6778	0.5256	0.3828	0.1673	0.0547	0.0123	0.0016	0.0004	0.0001	0.0000				
3	1.0000	0.9990	0.9872	0.8791	0.7759	0.6496	0.3823	0.1719	0.0548	0.0106	0.0035	0.0009	0.0000				
4		0.9999	0.9984	0.9672	0.9219	0.8497	0.6331	0.3770	0.1662	0.0473	0.0197	0.0064	0.0001	0.0000			
5			1.0000	0.9999	0.9936	0.9803	0.9527	0.8338	0.6230	0.3669	0.1503	0.0781	0.0328	0.0016	0.0001		
6				1.0000	0.9991	0.9965	0.9894	0.9452	0.8281	0.6177	0.3504	0.2241	0.1209	0.0128	0.0010	0.0000	
7					0.9999	0.9996	0.9984	0.9877	0.9453	0.8327	0.6172	0.4744	0.3222	0.0702	0.0115	0.0001	
8						1.0000	1.0000	0.9999	0.9983	0.9893	0.9536	0.8507	0.7560	0.6242	0.2639	0.0861	0.0043
9							1.0000	0.9999	0.9990	0.9940	0.9718	0.9437	0.8926	0.6513	0.4013	0.0956	

$n = 15$		p																	
x		0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99			
0	0.8601	0.4633	0.2059	0.0352	0.0134	0.0047	0.0005	0.0000											
1	0.9904	0.8290	0.5490	0.1671	0.0802	0.0353	0.0052	0.0005	0.0000										
2	0.9996	0.9638	0.8159	0.3980	0.2361	0.1268	0.0271	0.0037	0.0003	0.0000									
3	1.0000	0.9945	0.9444	0.6482	0.4613	0.2969	0.0905	0.0176	0.0019	0.0001	0.0000								
4		0.9994	0.9873	0.8358	0.6865	0.5155	0.2173	0.0592	0.0093	0.0007	0.0001	0.0000							
5			0.9999	0.9978	0.9389	0.8516	0.7216	0.4032	0.1509	0.0338	0.0037	0.0008	0.0001						
6				1.0000	0.9997	0.9819	0.9434	0.8689	0.6098	0.3036	0.0950	0.0152	0.0042	0.0008					
7					1.0000	0.9958	0.9827	0.9500	0.7869	0.5000	0.2131	0.0500	0.0173	0.0042	0.0000				
8						0.9992	0.9958	0.9848	0.9050	0.6964	0.3902	0.1311	0.0566	0.0181	0.0003	0.0000			
9							0.9999	0.9992	0.9963	0.9662	0.8491	0.5968	0.2784	0.1484	0.0611	0.0022	0.0001		
10							1.0000	0.9999	0.9993	0.9907	0.9408	0.7827	0.4845	0.3135	0.1642	0.0127	0.0006		
11								1.0000	0.9999	0.9981	0.9824	0.9095	0.7031	0.5387	0.3518	0.0556	0.0055	0.0000	
12									1.0000	0.9997	0.9963	0.9729	0.8732	0.7639	0.6020	0.1841	0.0362	0.0004	
13										1.0000	0.9995	0.9948	0.9647	0.9198	0.8329	0.4510	0.1710	0.0096	
14											1.0000	0.9995	0.9953	0.9866	0.9648	0.7941	0.5367	0.1399	

Table 1 Binomial Distribution Cumulative Probabilities (Continued)

<i>n</i> = 20		<i>p</i>																					
<i>x</i>		0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99							
0	0.8179	0.3585	0.1216	0.0115	0.0032	0.0008	0.0000																
1	0.9831	0.7358	0.3917	0.0692	0.0243	0.0076	0.0005	0.0000															
2	0.9990	0.9245	0.6769	0.2061	0.0913	0.0355	0.0036	0.0002															
3	1.0000	0.9841	0.8670	0.4114	0.2252	0.1071	0.0160	0.0013	0.0000														
4		0.9974	0.9568	0.6296	0.4148	0.2375	0.0510	0.0059	0.0003														
5			0.9997	0.9887	0.8042	0.6172	0.4164	0.1256	0.0207	0.0016	0.0000												
6				1.0000	0.9976	0.9133	0.7858	0.6080	0.2500	0.0577	0.0065	0.0003	0.0000										
7					0.9996	0.9679	0.8982	0.7723	0.4159	0.1316	0.0210	0.0013	0.0002	0.0000									
8						0.9999	0.9900	0.9591	0.8867	0.5956	0.2517	0.0565	0.0051	0.0009	0.0001								
9							1.0000	0.9974	0.9861	0.9520	0.7553	0.4119	0.1275	0.0171	0.0039	0.0006							
10								0.9994	0.9961	0.9829	0.8725	0.5881	0.2447	0.0480	0.0139	0.0026	0.0000						
11								0.9999	0.9991	0.9949	0.9435	0.7483	0.4044	0.1133	0.0409	0.0100	0.0001						
12									1.0000	0.9998	0.9987	0.9790	0.8684	0.5841	0.2277	0.1018	0.0321	0.0004					
13										1.0000	0.9997	0.9935	0.9423	0.7500	0.3920	0.2142	0.0867	0.0024	0.0000				
14											1.0000	0.9984	0.9793	0.8744	0.5836	0.3828	0.1958	0.0113	0.0003				
15												0.9997	0.9941	0.9490	0.7625	0.5852	0.3704	0.0432	0.0026				
16													1.0000	0.9987	0.9840	0.8929	0.7748	0.5886	0.1330	0.0159	0.0000		
17													0.9998	0.9964	0.9645	0.9087	0.7939	0.3231	0.0755	0.0010			
18														1.0000	0.9995	0.9924	0.9757	0.9308	0.6083	0.2642	0.0169		
19															1.0000	0.9992	0.9968	0.9885	0.8784	0.6415	0.1821		

<i>n</i> = 25		<i>p</i>																							
<i>x</i>		0.01	0.05	0.10	0.20	0.25	0.30	0.40	0.50	0.60	0.70	0.75	0.80	0.90	0.95	0.99									
0	0.7778	0.2774	0.0718	0.0038	0.0008	0.0001	0.0000																		
1	0.9742	0.6424	0.2712	0.0274	0.0070	0.0016	0.0001																		
2	0.9980	0.8729	0.5371	0.0982	0.0321	0.0090	0.0004	0.0000																	
3	0.9999	0.9659	0.7636	0.2340	0.0962	0.0332	0.0024	0.0001																	
4	1.0000	0.9928	0.9020	0.4207	0.2137	0.0905	0.0095	0.0005	0.0000																
5			0.9988	0.9666	0.6167	0.3783	0.1935	0.0294	0.0020	0.0001															
6				0.9998	0.9905	0.7800	0.5611	0.3407	0.0736	0.0073	0.0003														
7					1.0000	0.9977	0.8909	0.7265	0.5118	0.1536	0.0216	0.0012	0.0000												
8						0.9995	0.9532	0.8506	0.6769	0.2735	0.0539	0.0043	0.0001												
9							0.9999	0.9827	0.9287	0.8106	0.4246	0.1148	0.0132	0.0005	0.0000										
10								1.0000	0.9944	0.9703	0.9022	0.5858	0.2122	0.0344	0.0018	0.0002	0.0000								
11									0.9985	0.9893	0.9558	0.7323	0.3450	0.0778	0.0060	0.0009	0.0001								
12										0.9996	0.9966	0.9825	0.8462	0.5000	0.1538	0.0175	0.0034	0.0004							
13											0.9999	0.9991	0.9940	0.9222	0.6550	0.2677	0.0442	0.0107	0.0015						
14												1.0000	0.9998	0.9982	0.9656	0.7878	0.4142	0.0978	0.0297	0.0056					
15													1.0000	0.9995	0.9868	0.8852	0.5754	0.1894	0.0713	0.0173	0.0001				
16													0.9999	0.9957	0.9461	0.7265	0.3231	0.1494	0.0468	0.0005					
17														1.0000	0.9988	0.9784	0.8464	0.4882	0.2735	0.1091	0.0023	0.0000			
18														0.9997	0.9927	0.9264	0.6593	0.4389	0.2200	0.0095	0.0002				
19														0.9999	0.9980	0.9706	0.8065	0.6217	0.3833	0.0334	0.0012				
20															1.0000	0.9995	0.9905	0.9095	0.7863	0.5793	0.0980	0.0072	0.0000		
21															0.9999	0.9976	0.9668	0.9038	0.7660	0.2364	0.0341	0.0001			
22																1.0000	0.9996	0.9910	0.9679	0.9018	0.4629	0.1271	0.0020		
23																0.9999	0.9984	0.9930	0.9726	0.7288	0.3576	0.0258			
24																	1.0000	0.9999	0.9992	0.9962	0.9282	0.7226	0.2222		

Table 2 Poisson Distribution Cumulative Probabilities

Let X be a Poisson random variable with parameter λ . This table contains cumulative probabilities:

$$P(X \leq x) = \sum_{k=0}^x P(X = k) = P(X = 0) + P(X = 1) + \cdots + P(X = x).$$

x	λ									
	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50
0	0.9512	0.9048	0.8607	0.8187	0.7788	0.7408	0.7047	0.6703	0.6376	0.6065
1	0.9988	0.9953	0.9898	0.9825	0.9735	0.9631	0.9513	0.9384	0.9246	0.9098
2	1.0000	0.9998	0.9995	0.9989	0.9978	0.9964	0.9945	0.9921	0.9891	0.9856
3		1.0000	1.0000	0.9999	0.9999	0.9997	0.9995	0.9992	0.9988	0.9982
4			1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9998
5							1.0000	1.0000	1.0000	1.0000

x	λ									
	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95	1.00
0	0.5769	0.5488	0.5220	0.4966	0.4724	0.4493	0.4274	0.4066	0.3867	0.3679
1	0.8943	0.8781	0.8614	0.8442	0.8266	0.8088	0.7907	0.7725	0.7541	0.7358
2	0.9815	0.9769	0.9717	0.9659	0.9595	0.9526	0.9451	0.9371	0.9287	0.9197
3	0.9975	0.9966	0.9956	0.9942	0.9927	0.9909	0.9889	0.9865	0.9839	0.9810
4	0.9997	0.9996	0.9994	0.9992	0.9989	0.9986	0.9982	0.9977	0.9971	0.9963
5	1.0000	1.0000	0.9999	0.9999	0.9999	0.9998	0.9997	0.9997	0.9995	0.9994
6		1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999
7								1.0000	1.0000	1.0000

x	λ									
	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0	0.3329	0.3012	0.2725	0.2466	0.2231	0.2019	0.1827	0.1653	0.1496	0.1353
1	0.6990	0.6626	0.6268	0.5918	0.5578	0.5249	0.4932	0.4628	0.4337	0.4060
2	0.9004	0.8795	0.8571	0.8335	0.8088	0.7834	0.7572	0.7306	0.7037	0.6767
3	0.9743	0.9662	0.9569	0.9463	0.9344	0.9212	0.9068	0.8913	0.8747	0.8571
4	0.9946	0.9923	0.9893	0.9857	0.9814	0.9763	0.9704	0.9636	0.9559	0.9473
5	0.9990	0.9985	0.9978	0.9968	0.9955	0.9940	0.9920	0.9896	0.9868	0.9834
6	0.9999	0.9997	0.9996	0.9994	0.9991	0.9987	0.9981	0.9974	0.9966	0.9955
7	1.0000	1.0000	0.9999	0.9999	0.9998	0.9997	0.9996	0.9994	0.9992	0.9989
8			1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9998	0.9998
9						1.0000	1.0000	1.0000	1.0000	1.0000

Table 2 Poisson Distribution Cumulative Probabilities (Continued)

x	λ									
	2.1	2.2	2.3	2.4	2.5	2.6	2.7	2.8	2.9	3.0
0	0.1225	0.1108	0.1003	0.0907	0.0821	0.0743	0.0672	0.0608	0.0550	0.0498
1	0.3796	0.3546	0.3309	0.3084	0.2873	0.2674	0.2487	0.2311	0.2146	0.1991
2	0.6496	0.6227	0.5960	0.5697	0.5438	0.5184	0.4936	0.4695	0.4460	0.4232
3	0.8386	0.8194	0.7993	0.7787	0.7576	0.7360	0.7141	0.6919	0.6696	0.6472
4	0.9379	0.9275	0.9162	0.9041	0.8912	0.8774	0.8629	0.8477	0.8318	0.8153
5	0.9796	0.9751	0.9700	0.9643	0.9580	0.9510	0.9433	0.9349	0.9258	0.9161
6	0.9941	0.9925	0.9906	0.9884	0.9858	0.9828	0.9794	0.9756	0.9713	0.9665
7	0.9985	0.9980	0.9974	0.9967	0.9958	0.9947	0.9934	0.9919	0.9901	0.9881
8	0.9997	0.9995	0.9994	0.9991	0.9989	0.9985	0.9981	0.9976	0.9969	0.9962
9	0.9999	0.9999	0.9999	0.9998	0.9997	0.9996	0.9995	0.9993	0.9991	0.9989
10	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999	0.9999	0.9998	0.9998	0.9997
11					1.0000	1.0000	1.0000	1.0000	0.9999	0.9999
12									1.0000	1.0000

x	λ									
	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.0
0	0.0450	0.0408	0.0369	0.0334	0.0302	0.0273	0.0247	0.0224	0.0202	0.0183
1	0.1847	0.1712	0.1586	0.1468	0.1359	0.1257	0.1162	0.1074	0.0992	0.0916
2	0.4012	0.3799	0.3594	0.3397	0.3208	0.3027	0.2854	0.2689	0.2531	0.2381
3	0.6248	0.6025	0.5803	0.5584	0.5366	0.5152	0.4942	0.4735	0.4532	0.4335
4	0.7982	0.7806	0.7626	0.7442	0.7254	0.7064	0.6872	0.6678	0.6484	0.6288
5	0.9057	0.8946	0.8829	0.8705	0.8576	0.8441	0.8301	0.8156	0.8006	0.7851
6	0.9612	0.9554	0.9490	0.9421	0.9347	0.9267	0.9182	0.9091	0.8995	0.8893
7	0.9858	0.9832	0.9802	0.9769	0.9733	0.9692	0.9648	0.9599	0.9546	0.9489
8	0.9953	0.9943	0.9931	0.9917	0.9901	0.9883	0.9863	0.9840	0.9815	0.9786
9	0.9986	0.9982	0.9978	0.9973	0.9967	0.9960	0.9952	0.9942	0.9931	0.9919
10	0.9996	0.9995	0.9994	0.9992	0.9990	0.9987	0.9984	0.9981	0.9977	0.9972
11	0.9999	0.9999	0.9998	0.9998	0.9997	0.9996	0.9995	0.9994	0.9993	0.9991
12	1.0000	1.0000	1.0000	0.9999	0.9999	0.9999	0.9999	0.9998	0.9998	0.9997
13				1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999
14									1.0000	1.0000

Table 2 Poisson Distribution Cumulative Probabilities (Continued)

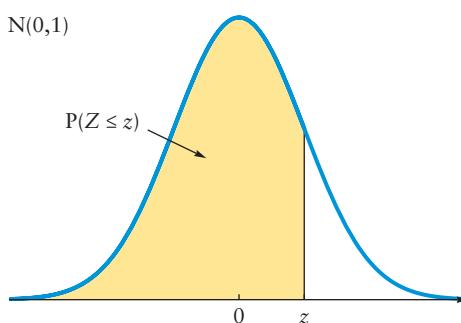
x	λ									
	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9	5.0
0	0.0166	0.0150	0.0136	0.0123	0.0111	0.0101	0.0091	0.0082	0.0074	0.0067
1	0.0845	0.0780	0.0719	0.0663	0.0611	0.0563	0.0518	0.0477	0.0439	0.0404
2	0.2238	0.2102	0.1974	0.1851	0.1736	0.1626	0.1523	0.1425	0.1333	0.1247
3	0.4142	0.3954	0.3772	0.3594	0.3423	0.3257	0.3097	0.2942	0.2793	0.2650
4	0.6093	0.5898	0.5704	0.5512	0.5321	0.5132	0.4946	0.4763	0.4582	0.4405
5	0.7693	0.7531	0.7367	0.7199	0.7029	0.6858	0.6684	0.6510	0.6335	0.6160
6	0.8786	0.8675	0.8558	0.8436	0.8311	0.8180	0.8046	0.7908	0.7767	0.7622
7	0.9427	0.9361	0.9290	0.9214	0.9134	0.9049	0.8960	0.8867	0.8769	0.8666
8	0.9755	0.9721	0.9683	0.9642	0.9597	0.9549	0.9497	0.9442	0.9382	0.9319
9	0.9905	0.9889	0.9871	0.9851	0.9829	0.9805	0.9778	0.9749	0.9717	0.9682
10	0.9966	0.9959	0.9952	0.9943	0.9933	0.9922	0.9910	0.9896	0.9880	0.9863
11	0.9989	0.9986	0.9983	0.9980	0.9976	0.9971	0.9966	0.9960	0.9953	0.9945
12	0.9997	0.9996	0.9995	0.9993	0.9992	0.9990	0.9988	0.9986	0.9983	0.9980
14	0.9999	0.9999	0.9998	0.9998	0.9997	0.9997	0.9996	0.9995	0.9994	0.9993
15	1.0000	1.0000	1.0000	0.9999	0.9999	0.9999	0.9999	0.9999	0.9998	0.9998
16				1.0000	1.0000	1.0000	1.0000	1.0000	0.9999	0.9999
17								1.0000	1.0000	

x	λ									
	5.5	6.0	6.5	7.0	7.5	8.0	8.5	9.0	9.5	10.0
0	0.0041	0.0025	0.0015	0.0009	0.0006	0.0003	0.0002	0.0001	0.0001	0.0000
1	0.0266	0.0174	0.0113	0.0073	0.0047	0.0030	0.0019	0.0012	0.0008	0.0005
2	0.0884	0.0620	0.0430	0.0296	0.0203	0.0138	0.0093	0.0062	0.0042	0.0028
3	0.2017	0.1512	0.1118	0.0818	0.0591	0.0424	0.0301	0.0212	0.0149	0.0103
4	0.3575	0.2851	0.2237	0.1730	0.1321	0.0996	0.0744	0.0550	0.0403	0.0293
5	0.5289	0.4457	0.3690	0.3007	0.2414	0.1912	0.1496	0.1157	0.0885	0.0671
6	0.6860	0.6063	0.5265	0.4497	0.3782	0.3134	0.2562	0.2068	0.1649	0.1301
7	0.8095	0.7440	0.6728	0.5987	0.5246	0.4530	0.3856	0.3239	0.2687	0.2202
8	0.8944	0.8472	0.7916	0.7291	0.6620	0.5925	0.5231	0.4557	0.3918	0.3328
9	0.9462	0.9161	0.8774	0.8305	0.7764	0.7166	0.6530	0.5874	0.5218	0.4579
10	0.9747	0.9574	0.9332	0.9015	0.8622	0.8159	0.7634	0.7060	0.6453	0.5830
11	0.9890	0.9799	0.9661	0.9467	0.9208	0.8881	0.8487	0.8030	0.7520	0.6968
12	0.9955	0.9912	0.9840	0.9730	0.9573	0.9362	0.9091	0.8758	0.8364	0.7916
13	0.9983	0.9964	0.9929	0.9872	0.9784	0.9658	0.9486	0.9261	0.8981	0.8645
14	0.9994	0.9986	0.9970	0.9943	0.9897	0.9827	0.9726	0.9585	0.9400	0.9165
15	0.9998	0.9995	0.9988	0.9976	0.9954	0.9918	0.9862	0.9780	0.9665	0.9513
16	0.9999	0.9998	0.9996	0.9990	0.9980	0.9963	0.9934	0.9889	0.9823	0.9730
17	1.0000	0.9999	0.9998	0.9996	0.9992	0.9984	0.9970	0.9947	0.9911	0.9857
18		1.0000	0.9999	0.9999	0.9997	0.9993	0.9987	0.9976	0.9957	0.9928
19			1.0000	1.0000	0.9999	0.9997	0.9995	0.9989	0.9980	0.9965
20				1.0000	0.9999	0.9998	0.9996	0.9991	0.9984	
21					1.0000	0.9999	0.9998	0.9996	0.9993	
22						1.0000	0.9999	0.9999	0.9997	
23							1.0000	0.9999	0.9999	
24								1.0000	1.0000	

Table 3 Standard Normal Distribution Cumulative Probabilities

Let Z be a standard normal random variable: $\mu = 0$ and $\sigma = 1$.

This table contains cumulative probabilities: $P(Z \leq z)$.



<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
-3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
-3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
-3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
-3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

Table 3 Standard Normal Distribution Cumulative Probabilities (Continued)

<i>z</i>	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Special critical values: $P(Z \geq z_\alpha) = \alpha$

α	0.10	0.05	0.025	0.01	0.005	0.001	0.0005	0.0001
z_α	1.2816	1.6449	1.9600	2.3263	2.5758	3.0902	3.2905	3.7190

α	0.00009	0.00008	0.00007	0.00006	0.00005	0.00004	0.00003	0.00002	0.00001
z_α	3.7455	3.7750	3.8082	3.8461	3.8906	3.9444	4.0128	4.1075	4.2649

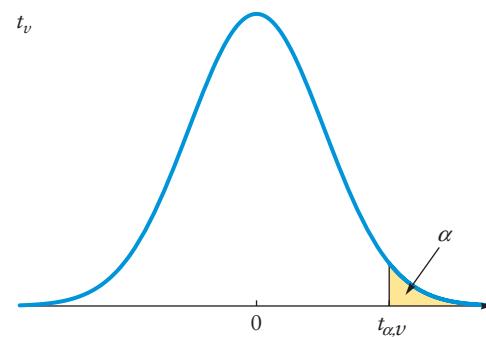
Table 4 Standardized Normal Scores

This table contains the standardized normal scores, z_i , for selected values of n .

<i>i</i>	<i>n</i>					
	10	20	25	30	40	50
1	-1.55	-1.87	-1.96	-2.04	-2.16	-2.24
2	-1.00	-1.40	-1.52	-1.61	-1.75	-1.85
3	-0.66	-1.13	-1.26	-1.36	-1.51	-1.62
4	-0.38	-0.92	-1.06	-1.18	-1.34	-1.46
5	-0.12	-0.74	-0.90	-1.02	-1.20	-1.33
6	0.12	-0.59	-0.76	-0.89	-1.08	-1.22
7	0.38	-0.45	-0.64	-0.78	-0.98	-1.12
8	0.66	-0.31	-0.52	-0.67	-0.88	-1.03
9	1.00	-0.19	-0.41	-0.57	-0.79	-0.95
10	1.55	-0.06	-0.30	-0.47	-0.71	-0.87
11		0.06	-0.20	-0.38	-0.63	-0.80
12		0.19	-0.10	-0.29	-0.56	-0.73
13		0.31	0.00	-0.21	-0.49	-0.67
14		0.45	0.10	-0.12	-0.42	-0.61
15		0.59	0.20	-0.04	-0.35	-0.55
16		0.74	0.30	0.04	-0.28	-0.49
17		0.92	0.41	0.12	-0.22	-0.44
18		1.13	0.52	0.21	-0.16	-0.38
19		1.40	0.64	0.29	-0.09	-0.33
20		1.87	0.76	0.38	-0.03	-0.28
21			0.90	0.47	0.03	-0.23
22			1.06	0.57	0.09	-0.18
23			1.26	0.67	0.16	-0.13
24			1.52	0.78	0.22	-0.07
25			1.96	0.89	0.28	-0.02
26				1.02	0.35	0.02
27				1.18	0.42	0.07
28				1.36	0.49	0.13
29				1.61	0.56	0.18
30				2.04	0.63	0.23
31					0.71	0.28
32					0.79	0.33
33					0.88	0.38
34					0.98	0.44
35					1.08	0.49
36					1.20	0.55
37					1.34	0.61
38					1.51	0.67
39					1.75	0.73
40					2.16	0.80
41						0.87
42						0.95
43						1.03
44						1.12
45						1.22
46						1.33
47						1.46
48						1.62
49						1.85
50						2.24

Table 5 Critical Values for the t Distribution

This table contains critical values associated with the t distribution, $t_{\alpha,\nu}$, defined by α and the degrees of freedom, ν .



ν	α								
	0.20	0.10	0.05	0.025	0.01	0.005	0.001	0.0005	0.0001
1	1.3764	3.0777	6.3138	12.7062	31.8205	63.6567	318.3088	636.6192	3183.0988
2	1.0607	1.8856	2.9200	4.3027	6.9646	9.9248	22.3271	31.5991	70.7001
3	0.9785	1.6377	2.3534	3.1824	4.5407	5.8409	10.2145	12.9240	22.2037
4	0.9410	1.5332	2.1318	2.7764	3.7469	4.6041	7.1732	8.6103	13.0337
5	0.9195	1.4759	2.0150	2.5706	3.3649	4.0321	5.8934	6.8688	9.6776
6	0.9057	1.4398	1.9432	2.4469	3.1427	3.7074	5.2076	5.9588	8.0248
7	0.8960	1.4149	1.8946	2.3646	2.9980	3.4995	4.7853	5.4079	7.0634
8	0.8889	1.3968	1.8595	2.3060	2.8965	3.3554	4.5008	5.0413	6.4420
9	0.8834	1.3830	1.8331	2.2622	2.8214	3.2498	4.2968	4.7809	6.0101
10	0.8791	1.3722	1.8125	2.2281	2.7638	3.1693	4.1437	4.5869	5.6938
11	0.8755	1.3634	1.7959	2.2010	2.7181	3.1058	4.0247	4.4370	5.4528
12	0.8726	1.3562	1.7823	2.1788	2.6810	3.0545	3.9296	4.3178	5.2633
13	0.8702	1.3502	1.7709	2.1604	2.6503	3.0123	3.8520	4.2208	5.1106
14	0.8681	1.3450	1.7613	2.1448	2.6245	2.9768	3.7874	4.1405	4.9850
15	0.8662	1.3406	1.7531	2.1314	2.6025	2.9467	3.7328	4.0728	4.8800
16	0.8647	1.3368	1.7459	2.1199	2.5835	2.9208	3.6862	4.0150	4.7909
17	0.8633	1.3334	1.7396	2.1098	2.5669	2.8982	3.6458	3.9651	4.7144
18	0.8620	1.3304	1.7341	2.1009	2.5524	2.8784	3.6105	3.9216	4.6480
19	0.8610	1.3277	1.7291	2.0930	2.5395	2.8609	3.5794	3.8834	4.5899
20	0.8600	1.3253	1.7247	2.0860	2.5280	2.8453	3.5518	3.8495	4.5385
21	0.8591	1.3232	1.7207	2.0796	2.5176	2.8314	3.5272	3.8193	4.4929
22	0.8583	1.3212	1.7171	2.0739	2.5083	2.8188	3.5050	3.7921	4.4520
23	0.8575	1.3195	1.7139	2.0687	2.4999	2.8073	3.4850	3.7676	4.4152
24	0.8569	1.3178	1.7109	2.0639	2.4922	2.7969	3.4668	3.7454	4.3819
25	0.8562	1.3163	1.7081	2.0595	2.4851	2.7874	3.4502	3.7251	4.3517
26	0.8557	1.3150	1.7056	2.0555	2.4786	2.7787	3.4350	3.7066	4.3240
27	0.8551	1.3137	1.7033	2.0518	2.4727	2.7707	3.4210	3.6896	4.2987
28	0.8546	1.3125	1.7011	2.0484	2.4671	2.7633	3.4082	3.6739	4.2754
29	0.8542	1.3114	1.6991	2.0452	2.4620	2.7564	3.3962	3.6594	4.2539
30	0.8538	1.3104	1.6973	2.0423	2.4573	2.7500	3.3852	3.6460	4.2340
40	0.8507	1.3031	1.6839	2.0211	2.4233	2.7045	3.3069	3.5510	4.0942
50	0.8489	1.2987	1.6759	2.0086	2.4033	2.6778	3.2614	3.4960	4.0140
60	0.8477	1.2958	1.6706	2.0003	2.3901	2.6603	3.2317	3.4602	3.9621
70	0.8468	1.2938	1.6669	1.9944	2.3808	2.6479	3.2108	3.4350	3.9257
80	0.8461	1.2922	1.6641	1.9901	2.3739	2.6387	3.1953	3.4163	3.8988
90	0.8456	1.2910	1.6620	1.9867	2.3685	2.6316	3.1833	3.4019	3.8780
100	0.8452	1.2901	1.6602	1.9840	2.3642	2.6259	3.1737	3.3905	3.8616
200	0.8434	1.2858	1.6525	1.9719	2.3451	2.6006	3.1315	3.3398	3.7891
500	0.8423	1.2832	1.6479	1.9647	2.3338	2.5857	3.1066	3.3101	3.7468
∞	0.8416	1.2816	1.6449	1.9600	2.3263	2.5758	3.0902	3.2905	3.7190

Table 6 Critical Values for the Chi-Square Distribution

This table contains critical values associated with the chi-square distribution, $\chi_{\alpha,\nu}^2$, defined by α and the degrees of freedom, ν .

ν	α							
	0.9999	0.9995	0.999	0.995	0.99	0.975	0.95	0.90
1	0.07157	0.06393	0.05157	0.04393	0.0002	0.0010	0.0039	0.0158
2	0.0002	0.0010	0.0020	0.0100	0.0201	0.0506	0.1026	0.2107
3	0.0052	0.0153	0.0243	0.0717	0.1148	0.2158	0.3518	0.5844
4	0.0284	0.0639	0.0908	0.2070	0.2971	0.4844	0.7107	1.0636
5	0.0822	0.1581	0.2102	0.4117	0.5543	0.8312	1.1455	1.6103
6	0.1724	0.2994	0.3811	0.6757	0.8721	1.2373	1.6354	2.2041
7	0.3000	0.4849	0.5985	0.9893	1.2390	1.6899	2.1673	2.8331
8	0.4636	0.7104	0.8571	1.3444	1.6465	2.1797	2.7326	3.4895
9	0.6608	0.9717	1.1519	1.7349	2.0879	2.7004	3.3251	4.1682
10	0.8889	1.2650	1.4787	2.1559	2.5582	3.2470	3.9403	4.8652
11	1.1453	1.5868	1.8339	2.6032	3.0535	3.8157	4.5748	5.5778
12	1.4275	1.9344	2.2142	3.0738	3.5706	4.4038	5.2260	6.3038
13	1.7333	2.3051	2.6172	3.5650	4.1069	5.0088	5.8919	7.0415
14	2.0608	2.6967	3.0407	4.0747	4.6604	5.6287	6.5706	7.7895
15	2.4082	3.1075	3.4827	4.6009	5.2293	6.2621	7.2609	8.5468
16	2.7739	3.5358	3.9416	5.1422	5.8122	6.9077	7.9616	9.3122
17	3.1567	3.9802	4.4161	5.6972	6.4078	7.5642	8.6718	10.0852
18	3.5552	4.4394	4.9048	6.2648	7.0149	8.2307	9.3905	10.8649
19	3.9683	4.9123	5.4068	6.8440	7.6327	8.9065	10.1170	11.6509
20	4.3952	5.3981	5.9210	7.4338	8.2604	9.5908	10.8508	12.4426
21	4.8348	5.8957	6.4467	8.0337	8.8972	10.2829	11.5913	13.2396
22	5.2865	6.4045	6.9830	8.6427	9.5425	10.9823	12.3380	14.0415
23	5.7494	6.9237	7.5292	9.2604	10.1957	11.6886	13.0905	14.8480
24	6.2230	7.4527	8.0849	9.8862	10.8564	12.4012	13.8484	15.6587
25	6.7066	7.9910	8.6493	10.5197	11.5240	13.1197	14.6114	16.4734
26	7.1998	8.5379	9.2221	11.1602	12.1981	13.8439	15.3792	17.2919
27	7.7019	9.0932	9.8028	11.8076	12.8785	14.5734	16.1514	18.1139
28	8.2126	9.6563	10.3909	12.4613	13.5647	15.3079	16.9279	18.9392
29	8.7315	10.2268	10.9861	13.1211	14.2565	16.0471	17.7084	19.7677
30	9.2581	10.8044	11.5880	13.7867	14.9535	16.7908	18.4927	20.5992
31	9.7921	11.3887	12.1963	14.4578	15.6555	17.5387	19.2806	21.4336
32	10.3331	11.9794	12.8107	15.1340	16.3622	18.2908	20.0719	22.2706
33	10.8810	12.5763	13.4309	15.8153	17.0735	19.0467	20.8665	23.1102
34	11.4352	13.1791	14.0567	16.5013	17.7891	19.8063	21.6643	23.9523
35	11.9957	13.7875	14.6878	17.1918	18.5089	20.5694	22.4650	24.7967
36	12.5622	14.4012	15.3241	17.8867	19.2327	21.3359	23.2686	25.6433
37	13.1343	15.0202	15.9653	18.5858	19.9602	22.1056	24.0749	26.4921
38	13.7120	15.6441	16.6112	19.2889	20.6914	22.8785	24.8839	27.3430
39	14.2950	16.2729	17.2616	19.9959	21.4262	23.6543	25.6954	28.1958
40	14.8831	16.9062	17.9164	20.7065	22.1643	24.4330	26.5093	29.0505
50	21.0093	23.4610	24.6739	27.9907	29.7067	32.3574	34.7643	37.6886
60	27.4969	30.3405	31.7383	35.5345	37.4849	40.4817	43.1880	46.4589
70	34.2607	37.4674	39.0364	43.2752	45.4417	48.7576	51.7393	55.3289
80	41.2445	44.7910	46.5199	51.1719	53.5401	57.1532	60.3915	64.2778
90	48.4087	52.2758	54.1552	59.1963	61.7541	65.6466	69.1260	73.2911
100	55.7246	59.8957	61.9179	67.3276	70.0649	74.2219	77.9295	82.3581

Table 6 Critical Values for the Chi-Square Distribution (Continued)

ν	α							
	0.10	0.05	0.025	0.01	0.005	0.001	0.0005	0.0001
1	2.7055	3.8415	5.0239	6.6349	7.8794	10.8276	12.1157	15.1367
2	4.6052	5.9915	7.3778	9.2103	10.5966	13.8155	15.2018	18.4207
3	6.2514	7.8147	9.3484	11.3449	12.8382	16.2662	17.7300	21.1075
4	7.7794	9.4877	11.1433	13.2767	14.8603	18.4668	19.9974	23.5127
5	9.2364	11.0705	12.8325	15.0863	16.7496	20.5150	22.1053	25.7448
6	10.6446	12.5916	14.4494	16.8119	18.5476	22.4577	24.1028	27.8563
7	12.0170	14.0671	16.0128	18.4753	20.2777	24.3219	26.0178	29.8775
8	13.3616	15.5073	17.5345	20.0902	21.9550	26.1245	27.8680	31.8276
9	14.6837	16.9190	19.0228	21.6660	23.5894	27.8772	29.6658	33.7199
10	15.9872	18.3070	20.4832	23.2093	25.1882	29.5883	31.4198	35.5640
11	17.2750	19.6751	21.9200	24.7250	26.7568	31.2641	33.1366	37.3670
12	18.5493	21.0261	23.3367	26.2170	28.2995	32.9095	34.8213	39.1344
13	19.8119	22.3620	24.7356	27.6882	29.8195	34.5282	36.4778	40.8707
14	21.0641	23.6848	26.1189	29.1412	31.3193	36.1233	38.1094	42.5793
15	22.3071	24.9958	27.4884	30.5779	32.8013	37.6973	39.7188	44.2632
16	23.5418	26.2962	28.8454	31.9999	34.2672	39.2524	41.3081	45.9249
17	24.7690	27.5871	30.1910	33.4087	35.7185	40.7902	42.8792	47.5664
18	25.9894	28.8693	31.5264	34.8053	37.1565	42.3124	44.4338	49.1894
19	27.2036	30.1435	32.8523	36.1909	38.5823	43.8202	45.9731	50.7955
20	28.4120	31.4104	34.1696	37.5662	39.9968	45.3147	47.4985	52.3860
21	29.6151	32.6706	35.4789	38.9322	41.4011	46.7970	49.0108	53.9620
22	30.8133	33.9244	36.7807	40.2894	42.7957	48.2679	50.5111	55.5246
23	32.0069	35.1725	38.0756	41.6384	44.1813	49.7282	52.0002	57.0746
24	33.1962	36.4150	39.3641	42.9798	45.5585	51.1786	53.4788	58.6130
25	34.3816	37.6525	40.6465	44.3141	46.9279	52.6197	54.9475	60.1403
26	35.5632	38.8851	41.9232	45.6417	48.2899	54.0520	56.4069	61.6573
27	36.7412	40.1133	43.1945	46.9629	49.6449	55.4760	57.8576	63.1645
28	37.9159	41.3371	44.4608	48.2782	50.9934	56.8923	59.3000	64.6624
29	39.0875	42.5570	45.7223	49.5879	52.3356	58.3012	60.7346	66.1517
30	40.2560	43.7730	46.9792	50.8922	53.6720	59.7031	62.1619	67.6326
31	41.4217	44.9853	48.2319	52.1914	55.0027	61.0983	63.5820	69.1057
32	42.5847	46.1943	49.4804	53.4858	56.3281	62.4872	64.9955	70.5712
33	43.7452	47.3999	50.7251	54.7755	57.6484	63.8701	66.4025	72.0296
34	44.9032	48.6024	51.9660	56.0609	58.9639	65.2472	67.8035	73.4812
35	46.0588	49.8018	53.2033	57.3421	60.2748	66.6188	69.1986	74.9262
36	47.2122	50.9985	54.4373	58.6192	61.5812	67.9852	70.5881	76.3650
37	48.3634	52.1923	55.6680	59.8925	62.8833	69.3465	71.9722	77.7977
38	49.5126	53.3835	56.8955	61.1621	64.1814	70.7029	73.3512	79.2247
39	50.6598	54.5722	58.1201	62.4281	65.4756	72.0547	74.7253	80.6462
40	51.8051	55.7585	59.3417	63.6907	66.7660	73.4020	76.0946	82.0623
50	63.1671	67.5048	71.4202	76.1539	79.4900	86.6608	89.5605	95.9687
60	74.3970	79.0819	83.2977	88.3794	91.9517	99.6072	102.6948	109.5029
70	85.5270	90.5312	95.0232	100.4252	104.2149	112.3169	115.5776	122.7547
80	96.5782	101.8795	106.6286	112.3288	116.3211	124.8392	128.2613	135.7825
90	107.5650	113.1453	118.1359	124.1163	128.2989	137.2084	140.7823	148.6273
100	118.4980	124.3421	129.5612	135.8067	140.1695	149.4493	153.1670	161.3187

Table 7 Critical Values for the F Distribution

This table contains critical values associated with the *F* distribution, F_{α, ν_1, ν_2} , defined by α and the degrees of freedom ν_1 and ν_2 .

$\alpha = 0.05$	ν_2	ν_1									
		1	2	3	4	5	6	7	8	9	10
1	161.45	199.50	215.71	224.58	230.16	233.99	236.77	238.88	240.54	241.88	245.95
2	18.51	19.00	19.16	19.25	19.30	19.35	19.37	19.38	19.40	19.43	19.46
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.81	8.79	8.70	8.66
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.86
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.62
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	3.94
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.51
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.22
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.01
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.85
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.72
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.62
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.53
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.46
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.40
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.35
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.31
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.27
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.23
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.20
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.18
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.15
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.13
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.11
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.09
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.01
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	1.92
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07	2.03	1.87
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.84
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97	1.93	1.77

Table 7 Critical Values for the F Distribution (Continued)

$\alpha = 0.01$		ν_1										ν_2					
ν_2	1	2	3	4	5	6	7	8	9	10	15	20	30	40	50	60	100
2	98.50	99.00	99.17	99.25	99.30	99.33	99.37	99.40	99.43	99.45	99.47	99.48	99.48	99.48	99.49	99.49	
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23	26.87	26.50	26.41	26.35	26.32	26.24	
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55	14.20	14.02	13.84	13.75	13.69	13.58	
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05	9.72	9.55	9.38	9.29	9.24	9.20	
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.56	7.40	7.23	7.14	7.09	7.06	
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.31	6.16	5.99	5.91	5.86	5.82	
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.52	5.36	5.20	5.12	5.07	5.03	
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	4.96	4.81	4.65	4.57	4.52	4.48	
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.56	4.41	4.25	4.17	4.12	4.08	
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.25	4.10	3.94	3.86	3.81	3.78	
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.01	3.86	3.70	3.62	3.57	3.54	
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.82	3.66	3.51	3.43	3.38	3.34	
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.66	3.51	3.35	3.27	3.22	3.18	
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.52	3.37	3.21	3.13	3.08	3.05	
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.41	3.26	3.10	3.02	2.97	2.93	
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.31	3.16	3.00	2.92	2.87	2.83	
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.23	3.08	2.92	2.84	2.78	2.75	
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.15	3.00	2.84	2.76	2.71	2.67	
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.09	2.94	2.78	2.69	2.64	2.61	
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.03	2.88	2.72	2.64	2.58	2.55	
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	2.98	2.83	2.67	2.58	2.53	2.50	
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	2.93	2.78	2.62	2.54	2.48	2.45	
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	2.89	2.74	2.58	2.49	2.44	2.40	
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.85	2.70	2.54	2.45	2.40	2.36	
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.70	2.55	2.39	2.30	2.25	2.21	
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.52	2.37	2.20	2.11	2.06	2.02	
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78	2.70	2.42	2.27	2.10	2.01	1.95	1.91	
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.35	2.20	2.03	1.94	1.88	1.84	
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59	2.50	2.22	2.07	1.89	1.80	1.74	1.69	

Table 7 Critical Values for the F Distribution (Continued)

$\alpha = 0.001$		ν_1															
ν_2	1	2	3	4	5	6	7	8	9	10	15	20	30	40	50	60	100
2	998.50	999.00	999.17	999.25	999.30	999.33	999.36	999.37	999.39	999.40	999.43	999.45	999.47	999.48	999.48	999.49	999.49
3	167.03	148.50	141.11	137.10	134.58	132.85	131.58	130.62	129.86	129.25	127.37	126.42	125.45	124.96	124.66	124.47	124.07
4	74.14	61.25	56.18	53.44	51.71	50.53	49.66	49.00	48.47	48.05	46.76	46.10	45.43	45.09	44.88	44.75	44.47
5	47.18	37.12	33.20	31.09	29.75	28.83	28.16	27.65	27.24	26.92	25.91	25.39	24.87	24.60	24.44	24.33	24.12
6	35.51	27.00	23.70	21.92	20.80	20.03	19.46	19.03	18.69	18.41	17.56	17.12	16.67	16.44	16.31	16.21	16.03
7	29.25	21.69	18.77	17.20	16.21	15.52	15.02	14.63	14.33	14.08	13.32	12.93	12.53	12.33	12.20	12.12	11.95
8	25.41	18.49	15.83	14.39	13.48	12.86	12.40	12.05	11.77	11.54	10.84	10.48	10.11	9.92	9.80	9.73	9.57
9	22.86	16.39	13.90	12.56	11.71	11.13	10.70	10.37	10.11	9.89	9.24	8.90	8.55	8.37	8.26	8.19	8.04
10	21.04	14.91	12.55	11.28	10.48	9.93	9.52	9.20	8.96	8.75	8.13	7.80	7.47	7.30	7.19	7.12	6.98
11	19.69	13.81	11.56	10.35	9.58	9.05	8.66	8.35	8.12	7.92	7.32	7.01	6.68	6.52	6.42	6.35	6.21
12	18.64	12.97	10.80	9.63	8.89	8.38	8.00	7.71	7.48	7.29	6.71	6.40	6.09	5.93	5.83	5.76	5.63
13	17.82	12.31	10.21	9.07	8.35	7.86	7.49	7.21	6.98	6.80	6.23	5.93	5.63	5.47	5.37	5.30	5.17
14	17.14	11.78	9.73	8.62	7.92	7.44	7.08	6.80	6.58	6.40	5.85	5.56	5.25	5.10	5.00	4.94	4.81
15	16.59	11.34	9.34	8.25	7.57	7.09	6.74	6.47	6.26	6.08	5.54	5.25	4.95	4.80	4.70	4.64	4.51
16	16.12	10.97	9.01	7.94	7.27	6.80	6.46	6.19	5.98	5.81	5.27	4.99	4.70	4.54	4.45	4.39	4.26
17	15.72	10.66	8.73	7.68	7.02	6.56	6.22	5.96	5.75	5.58	5.05	4.78	4.48	4.33	4.24	4.18	4.05
18	15.38	10.39	8.49	7.46	6.81	6.35	6.02	5.76	5.56	5.39	4.87	4.59	4.30	4.15	4.06	4.00	3.87
19	15.08	10.16	8.28	7.27	6.62	6.18	5.85	5.59	5.39	5.22	4.70	4.43	4.14	3.99	3.90	3.84	3.71
20	14.82	9.95	8.10	7.10	6.46	6.02	5.69	5.44	5.24	5.08	4.56	4.29	4.00	3.86	3.77	3.70	3.58
21	14.59	9.77	7.94	6.95	6.32	5.88	5.56	5.31	5.11	4.95	4.44	4.17	3.88	3.74	3.64	3.58	3.46
22	14.38	9.61	7.80	6.81	6.19	5.76	5.44	5.19	4.99	4.83	4.33	4.06	3.78	3.63	3.54	3.48	3.35
23	14.20	9.47	7.67	6.70	6.08	5.65	5.33	5.09	4.89	4.73	4.23	3.96	3.68	3.53	3.44	3.38	3.25
24	14.03	9.34	7.55	6.59	5.98	5.55	5.23	4.99	4.80	4.64	4.14	3.87	3.59	3.45	3.36	3.29	3.17
25	13.88	9.22	7.45	6.49	5.89	5.46	5.15	4.91	4.71	4.56	4.06	3.79	3.52	3.37	3.28	3.22	3.09
30	13.29	8.77	7.05	6.12	5.53	5.12	4.82	4.58	4.39	4.24	3.75	3.49	3.22	3.07	2.98	2.92	2.79
40	12.61	8.25	6.59	5.70	5.13	4.73	4.44	4.21	4.02	3.87	3.40	3.14	2.87	2.73	2.64	2.57	2.44
50	12.22	7.96	6.34	5.46	4.90	4.51	4.22	4.00	3.82	3.67	3.20	2.95	2.68	2.53	2.44	2.38	2.25
60	11.97	7.77	6.17	5.31	4.76	4.37	4.09	3.86	3.69	3.54	3.08	2.83	2.55	2.41	2.32	2.25	2.12
100	11.50	7.41	5.86	5.02	4.48	4.11	3.83	3.61	3.44	3.30	2.84	2.59	2.32	2.17	2.08	2.01	1.87

Table 8 Critical Values for the Studentized Range Distribution

This table contains critical values associated with the Studentized range distribution, $Q_{\alpha, k, \nu}$, defined by α , and the degrees of freedom k and ν , where k is the number of degrees of freedom in the numerator (the number of treatment groups) and ν is the number of degrees of freedom in the denominator.

$\alpha = 0.05$	k																		
ν	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	6.085	8.331	9.798	10.881	11.734	12.434	13.027	13.538	13.987	14.387	14.747	15.076	15.375	15.650	15.905	16.143	16.365	16.573	16.769
3	4.501	5.910	6.825	7.502	8.037	8.478	8.852	9.177	9.462	9.717	9.946	10.155	10.346	10.522	10.686	10.838	11.090	11.114	11.240
4	3.926	5.040	5.757	6.287	6.706	7.053	7.347	7.602	7.826	8.027	8.208	8.373	8.524	8.664	8.793	8.914	9.027	9.133	9.233
5	3.635	4.602	5.218	5.673	6.033	6.330	6.582	6.801	6.995	7.167	7.324	7.465	7.596	7.716	7.828	7.932	8.030	8.122	8.208
6	3.460	4.339	4.896	5.305	5.629	5.895	6.122	6.319	6.493	6.649	6.789	6.917	7.034	7.143	7.244	7.338	7.426	7.509	7.587
7	3.344	4.165	4.681	5.060	5.359	5.606	5.815	5.997	6.158	6.302	6.431	6.550	6.658	6.759	6.852	6.939	7.020	7.097	7.169
8	3.261	4.041	4.529	4.886	5.167	5.399	5.596	5.767	5.918	6.053	6.175	6.287	6.389	6.483	6.571	6.653	6.729	6.801	6.870
9	3.199	3.948	4.415	4.755	5.024	5.244	5.432	5.595	5.738	5.867	5.983	6.089	6.186	6.276	6.359	6.437	6.510	6.579	6.644
10	3.151	3.877	4.327	4.654	4.912	5.124	5.304	5.460	5.598	5.722	5.833	5.935	6.028	6.114	6.194	6.269	6.339	6.405	6.467
11	3.113	3.820	4.256	4.574	4.823	5.028	5.202	5.353	5.486	5.605	5.713	5.811	5.901	5.984	6.062	6.134	6.202	6.265	6.325
12	3.081	3.773	4.199	4.508	4.750	4.950	5.119	5.265	5.395	5.510	5.615	5.710	5.797	5.878	5.953	6.023	6.089	6.151	6.209
13	3.055	3.734	4.151	4.453	4.690	4.884	5.049	5.192	5.318	5.431	5.533	5.625	5.711	5.789	5.862	5.931	5.995	6.055	6.112
14	3.033	3.701	4.111	4.407	4.639	4.829	4.990	5.130	5.253	5.363	5.463	5.554	5.637	5.714	5.785	5.852	5.915	5.973	6.029
15	3.014	3.673	4.076	4.367	4.595	4.782	4.940	5.077	5.198	5.306	5.403	5.492	5.574	5.649	5.719	5.785	5.846	5.904	5.958
16	2.998	3.649	4.046	4.333	4.557	4.741	4.896	5.031	5.150	5.256	5.352	5.439	5.519	5.593	5.662	5.726	5.786	5.843	5.896
17	2.984	3.628	4.020	4.303	4.524	4.705	4.858	4.991	5.108	5.212	5.306	5.392	5.471	5.544	5.612	5.675	5.734	5.790	5.842
18	2.971	3.609	3.997	4.276	4.494	4.673	4.824	4.955	5.071	5.173	5.266	5.351	5.429	5.501	5.567	5.629	5.688	5.743	5.794
19	2.960	3.593	3.977	4.253	4.468	4.645	4.794	4.924	5.037	5.139	5.231	5.314	5.391	5.462	5.528	5.589	5.647	5.701	5.752
20	2.950	3.578	3.958	4.232	4.445	4.620	4.768	4.895	5.008	5.108	5.199	5.282	5.357	5.427	5.492	5.553	5.610	5.663	5.714
25	2.913	3.523	3.890	4.153	4.358	4.526	4.667	4.789	4.897	4.993	5.079	5.158	5.230	5.297	5.359	5.417	5.471	5.522	5.570
30	2.888	3.487	3.845	4.102	4.301	4.464	4.601	4.720	4.824	4.917	5.001	5.077	5.147	5.211	5.271	5.327	5.379	5.429	5.475
40	2.858	3.442	3.791	4.039	4.232	4.388	4.521	4.634	4.735	4.824	4.904	4.977	5.044	5.106	5.163	5.216	5.266	5.313	5.358
50	2.841	3.416	3.758	4.002	4.190	4.344	4.473	4.584	4.681	4.768	4.847	4.918	4.983	5.043	5.098	5.150	5.199	5.245	5.288
100	2.806	3.365	3.695	3.929	4.109	4.256	4.379	4.484	4.577	4.659	4.733	4.800	4.862	4.918	4.971	5.020	5.066	5.108	5.149
200	2.789	3.339	3.664	3.893	4.069	4.212	4.332	4.435	4.525	4.605	4.677	4.742	4.802	4.857	4.908	4.955	4.999	5.041	5.080
300	2.783	3.331	3.654	3.881	4.056	4.198	4.317	4.419	4.508	4.587	4.659	4.723	4.782	4.837	4.887	4.934	4.978	5.019	5.057
400	2.780	3.327	3.649	3.875	4.050	4.191	4.309	4.411	4.500	4.578	4.649	4.714	4.772	4.826	4.876	4.923	4.967	5.007	5.046
500	2.779	3.324	3.645	3.872	4.046	4.187	4.305	4.406	4.494	4.573	4.644	4.708	4.766	4.820	4.870	4.917	4.960	5.001	5.039

Table 8 Critical Values for the Studentized Range Distribution (Continued)

$\alpha = 0.01$	k																		
ν	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	14.035	19.019	22.293	24.717	26.628	28.199	29.528	30.677	31.687	32.585	33.395	34.129	34.802	35.421	35.995	36.529	37.028	37.496	37.937
3	8.260	10.616	12.169	13.324	14.240	14.997	15.640	16.198	16.689	17.128	17.524	17.884	18.214	18.519	18.802	19.065	19.311	19.543	19.761
4	6.511	8.118	9.173	9.958	10.582	11.099	11.539	11.925	12.264	12.566	12.840	13.089	13.318	13.530	13.726	13.909	14.081	14.242	14.394
5	5.702	6.976	7.806	8.421	8.913	9.321	9.669	9.971	10.239	10.479	10.695	10.893	11.075	11.243	11.399	11.544	11.681	11.809	11.930
6	5.243	6.331	7.033	7.556	7.974	8.318	8.611	8.869	9.097	9.300	9.485	9.653	9.808	9.951	10.084	10.208	10.325	10.434	10.538
7	4.948	5.919	6.543	7.006	7.373	7.678	7.940	8.167	8.368	8.548	8.711	8.859	8.996	9.124	9.242	9.353	9.456	9.553	9.645
8	4.745	5.635	6.204	6.625	6.960	7.238	7.475	7.681	7.864	8.028	8.177	8.312	8.437	8.552	8.659	8.760	8.854	8.942	9.026
9	4.595	5.428	5.957	6.347	6.658	6.915	7.134	7.326	7.495	7.647	7.785	7.910	8.026	8.133	8.233	8.326	8.413	8.495	8.573
10	4.482	5.270	5.769	6.136	6.428	6.669	6.875	7.055	7.214	7.356	7.485	7.603	7.712	7.813	7.906	7.994	8.076	8.153	8.226
11	4.392	5.146	5.621	5.970	6.247	6.476	6.671	6.842	6.992	7.127	7.250	7.362	7.465	7.560	7.649	7.732	7.810	7.883	7.952
12	4.320	5.046	5.502	5.836	6.101	6.321	6.507	6.670	6.814	6.943	7.060	7.167	7.265	7.356	7.441	7.520	7.594	7.664	7.731
13	4.261	4.964	5.404	5.727	5.981	6.192	6.372	6.528	6.666	6.791	6.903	7.006	7.100	7.188	7.269	7.345	7.417	7.484	7.548
14	4.210	4.895	5.322	5.634	5.881	6.085	6.258	6.409	6.543	6.664	6.772	6.871	6.962	7.047	7.125	7.199	7.268	7.333	7.394
15	4.167	4.836	5.252	5.556	5.796	5.994	6.162	6.309	6.438	6.555	6.660	6.757	6.845	6.927	7.003	7.074	7.141	7.204	7.264
16	4.131	4.786	5.192	5.488	5.722	5.915	6.079	6.222	6.348	6.461	6.564	6.658	6.743	6.824	6.897	6.967	7.032	7.093	7.151
17	4.099	4.742	5.140	5.430	5.659	5.847	6.007	6.147	6.270	6.380	6.480	6.572	6.656	6.733	6.806	6.873	6.937	6.997	7.053
18	4.071	4.703	5.094	5.379	5.603	5.787	5.944	6.081	6.201	6.309	6.407	6.496	6.579	6.655	6.725	6.791	6.854	6.912	6.967
19	4.046	4.669	5.054	5.333	5.553	5.735	5.888	6.022	6.141	6.246	6.342	6.430	6.510	6.585	6.654	6.719	6.780	6.837	6.891
20	4.024	4.639	5.018	5.293	5.509	5.687	5.839	5.970	6.086	6.190	6.285	6.370	6.449	6.523	6.591	6.654	6.714	6.770	6.823
25	3.942	4.527	4.884	5.143	5.346	5.513	5.654	5.777	5.885	5.982	6.070	6.150	6.223	6.291	6.355	6.414	6.469	6.521	6.571
30	3.889	4.454	4.799	5.048	5.242	5.401	5.536	5.653	5.756	5.848	5.932	6.008	6.078	6.142	6.202	6.258	6.311	6.360	6.407
40	3.825	4.367	4.695	4.931	5.114	5.265	5.392	5.502	5.599	5.685	5.764	5.835	5.900	5.961	6.017	6.069	6.118	6.165	6.208
50	3.787	4.316	4.634	4.863	5.040	5.185	5.308	5.414	5.507	5.590	5.665	5.734	5.796	5.854	5.908	5.958	6.005	6.050	6.092
100	3.714	4.216	4.516	4.730	4.896	5.031	5.144	5.242	5.328	5.405	5.474	5.537	5.594	5.648	5.697	5.743	5.786	5.826	5.864
200	3.714	4.216	4.516	4.730	4.896	5.031	5.144	5.242	5.328	5.405	5.474	5.537	5.594	5.648	5.697	5.743	5.786	5.826	5.864
300	3.666	4.152	4.440	4.645	4.803	4.931	5.039	5.132	5.213	5.286	5.351	5.410	5.464	5.514	5.560	5.603	5.644	5.682	5.717
400	3.661	4.144	4.431	4.634	4.791	4.919	5.026	5.118	5.199	5.271	5.335	5.394	5.448	5.543	5.586	5.626	5.664	5.699	5.734
500	3.657	4.139	4.425	4.628	4.784	4.911	5.018	5.110	5.190	5.262	5.327	5.385	5.438	5.488	5.533	5.576	5.616	5.653	5.688

Table 8 Critical Values for the Studentized Range Distribution (Continued)

$\alpha = 0.001$	ν	k																	
2		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2	44.666	60.323	70.586	78.162	84.127	89.022	93.650	97.285	100.480	103.325	105.886	108.211	110.340	112.300	114.115	115.805	117.385	118.867	120.263
3	18.275	23.298	26.609	29.075	31.030	32.645	34.016	35.327	36.389	37.338	38.194	38.974	39.688	40.347	40.959	41.529	42.062	42.564	43.036
4	12.174	14.965	16.798	18.225	19.333	20.253	21.037	21.719	22.323	22.862	23.350	23.795	24.204	24.581	24.932	25.259	25.566	25.854	26.126
5	9.710	11.671	12.959	13.924	14.695	15.335	15.882	16.358	16.780	17.158	17.500	17.811	18.098	18.402	18.651	18.884	19.102	19.307	19.501
6	8.431	9.955	10.965	11.719	12.322	12.824	13.254	13.629	13.961	14.260	14.530	14.777	15.004	15.215	15.411	15.593	15.765	15.927	16.079
7	7.649	8.933	9.761	10.388	11.316	11.674	11.988	12.265	12.515	12.742	12.949	13.139	13.316	13.480	13.634	13.778	13.914	14.043	
8	7.130	8.252	8.980	9.523	9.948	10.317	10.625	10.894	11.133	11.347	11.559	11.740	11.906	12.060	12.203	12.337	12.463	12.582	12.694
9	7.130	8.252	8.980	9.523	9.948	10.317	10.625	10.894	11.133	11.347	11.559	11.740	11.906	12.060	12.203	12.337	12.463	12.582	12.694
10	6.486	7.411	8.007	8.451	8.805	9.100	9.353	9.574	9.770	9.954	10.106	10.245	10.387	10.512	10.629	10.737	10.840	10.936	11.027
11	6.274	7.137	7.688	8.099	8.427	8.700	8.934	9.138	9.320	9.483	9.631	9.767	9.892	10.017	10.121	10.218	10.309	10.394	10.475
12	6.106	6.917	7.442	7.820	8.128	8.383	8.602	8.793	8.963	9.116	9.254	9.381	9.498	9.607	9.708	9.803	9.892	9.976	10.055
13	5.969	6.740	7.234	7.595	7.885	8.126	8.333	8.513	8.674	8.818	8.949	9.068	9.179	9.281	9.377	9.466	9.550	9.630	9.705
14	5.855	6.593	7.070	7.410	7.692	7.914	8.111	8.282	8.434	8.571	8.696	8.810	8.915	9.012	9.103	9.188	9.268	9.343	9.414
15	5.760	6.470	6.920	7.257	7.517	7.742	7.924	8.088	8.234	8.365	8.483	8.592	8.693	8.786	8.873	8.954	9.030	9.102	9.170
16	5.678	6.365	6.799	7.125	7.377	7.585	7.769	7.923	8.063	8.189	8.303	8.407	8.504	8.593	8.676	8.754	8.828	8.897	8.963
17	5.614	6.274	6.695	7.010	7.254	7.457	7.629	7.783	7.921	8.037	8.147	8.248	8.341	8.427	8.508	8.583	8.654	8.720	8.783
18	5.550	6.201	6.609	6.909	7.147	7.343	7.511	7.658	7.781	7.908	8.017	8.116	8.199	8.283	8.361	8.433	8.502	8.566	8.628
19	5.493	6.129	6.527	6.820	7.051	7.243	7.407	7.550	7.676	7.790	7.894	7.990	8.079	8.162	8.238	8.302	8.369	8.431	8.491
20	5.444	6.065	6.455	6.741	6.967	7.154	7.314	7.453	7.577	7.687	7.788	7.880	7.966	8.046	8.121	8.190	8.256	8.318	8.376
25	5.264	5.840	6.196	6.456	6.662	6.831	6.976	7.102	7.213	7.314	7.404	7.487	7.558	7.629	7.696	7.758	7.816	7.871	7.924
30	5.154	5.698	6.033	6.277	6.469	6.628	6.763	6.880	6.984	7.077	7.161	7.238	7.309	7.375	7.436	7.494	7.548	7.598	7.646
40	5.022	5.527	5.837	6.062	6.239	6.385	6.508	6.615	6.710	6.795	6.872	6.942	7.006	7.066	7.121	7.173	7.222	7.268	7.312
50	4.946	5.426	5.725	5.939	6.107	6.245	6.361	6.463	6.552	6.632	6.705	6.771	6.832	6.888	6.940	6.989	7.035	7.078	7.119
100	4.795	5.244	5.512	5.706	5.855	5.978	6.083	6.173	6.252	6.323	6.387	6.445	6.499	6.548	6.594	6.637	6.678	6.715	6.751
200	4.723	5.151	5.408	5.596	5.738	5.854	5.952	6.038	6.110	6.178	6.237	6.292	6.342	6.388	6.431	6.471	6.509	6.544	6.577
300	4.700	5.122	5.375	5.556	5.696	5.814	5.910	5.993	6.066	6.131	6.189	6.244	6.291	6.335	6.379	6.418	6.455	6.489	6.522
400	4.688	5.107	5.358	5.538	5.677	5.791	5.890	5.972	6.044	6.108	6.166	6.219	6.267	6.312	6.355	6.393	6.427	6.460	6.494
500	4.681	5.098	5.348	5.527	5.665	5.778	5.874	5.959	6.031	6.095	6.152	6.205	6.253	6.297	6.338	6.376	6.412	6.448	6.479

Table 9 Critical Values for the Wilcoxon Signed-Rank Statistic

This table contains critical values and probabilities for the Wilcoxon signed-rank statistic T_+ : n is the sample size, and c_1 and c_2 are defined by $P(T_+ \leq c_1) = \alpha$ and $P(T_+ \geq c_2) = \alpha$, respectively.

n	c_1	c_2	α	n	c_1	c_2	α	n	c_1	c_2	α	n	c_1	c_2	α	n	c_1	c_2	α	
1	0	1	0.5000	10	0	55	0.0010	12	0	78	0.0002	13	0	91	0.0001	14	0	105	0.0001	
2	0	3	0.2500		1	54	0.0020		1	77	0.0005		1	90	0.0002		1	104	0.0001	
3	0	6	0.1250		2	53	0.0029		2	76	0.0007		2	89	0.0004		2	103	0.0002	
4	0	10	0.0625		3	52	0.0049		3	75	0.0012		3	88	0.0006		3	102	0.0003	
	1	9	0.1250		4	51	0.0068		4	74	0.0017		4	87	0.0009		4	101	0.0004	
5	0	15	0.0313		5	50	0.0098		5	73	0.0024		5	86	0.0012		5	100	0.0006	
	1	14	0.0625		6	49	0.0137		6	72	0.0034		6	85	0.0017		6	99	0.0009	
	2	13	0.0938		7	48	0.0186		7	71	0.0046		7	84	0.0023		7	98	0.0012	
	3	12	0.1563		8	47	0.0244		8	70	0.0061		8	83	0.0031		8	97	0.0015	
	10	45	0.0420		9	46	0.0322		9	69	0.0081		9	82	0.0040		9	96	0.0020	
6	0	21	0.0156		10	45	0.0420		10	68	0.0105		10	81	0.0052		10	95	0.0026	
	1	20	0.0313		11	44	0.0527		11	67	0.0134		11	80	0.0067		11	94	0.0034	
	2	19	0.0469		12	43	0.0654		12	66	0.0171		12	79	0.0085		12	93	0.0043	
	3	18	0.0781		13	42	0.0801		13	65	0.0212		13	78	0.0107		13	92	0.0054	
	4	17	0.1094		14	41	0.0967		14	64	0.0261		14	77	0.0133		14	91	0.0067	
	5	16	0.1563		15	40	0.1162		15	63	0.0320		15	76	0.0164		15	90	0.0083	
	16	39	0.1377		17	39	0.1377		16	62	0.0386		16	75	0.0199		16	89	0.0101	
7	0	28	0.0078	11	0	66	0.0005		17	61	0.0461		17	74	0.0239		17	88	0.0123	
	1	27	0.0156		1	65	0.0010		18	60	0.0549		18	73	0.0287		18	87	0.0148	
	2	26	0.0234		2	64	0.0015		19	59	0.0647		19	72	0.0341		19	86	0.0176	
	3	25	0.0391		3	63	0.0024		20	58	0.0757		20	71	0.0402		20	85	0.0209	
	4	24	0.0547		4	62	0.0034		21	57	0.0881		21	70	0.0471		21	84	0.0247	
	5	23	0.0781		5	61	0.0049		22	56	0.1018		22	69	0.0549		22	83	0.0290	
	6	22	0.1094		6	60	0.0068		23	55	0.1167		23	68	0.0636		23	82	0.0338	
	7	21	0.1484		7	59	0.0093		24	54	0.1331		24	67	0.0732		24	81	0.0392	
8	0	36	0.0039		8	58	0.0122		25	53	0.1506		25	66	0.0839		25	80	0.0453	
	1	35	0.0078		9	57	0.0161			26	65	0.0955		26	79	0.0520				
	2	34	0.0117		10	56	0.0210			27	64	0.1082		27	78	0.0594				
	3	33	0.0195		11	55	0.0269			28	63	0.1219		28	77	0.0676				
	4	32	0.0273		12	54	0.0337			29	62	0.1367		29	76	0.0765				
	5	31	0.0391		13	53	0.0415			30	61	0.1527		30	75	0.0863				
	6	30	0.0547		14	52	0.0508			31	74	0.0969								
	7	29	0.0742		15	51	0.0615			32	73	0.1083								
	8	28	0.0977		16	50	0.0737			33	72	0.1206								
	9	27	0.1250		17	49	0.0874			34	71	0.1338								
9	0	45	0.0020		18	48	0.1030			35	70	0.1479								
	1	44	0.0039		19	47	0.1201			36	69	0.1629								
	2	43	0.0059		20	46	0.1392													
	3	42	0.0098																	
	4	41	0.0137																	
	5	40	0.0195																	
	6	39	0.0273																	
	7	38	0.0371																	
	8	37	0.0488																	
	9	36	0.0645																	
	10	35	0.0820																	
	11	34	0.1016																	
	12	33	0.1250																	

Table 9 Critical Values for the Wilcoxon Signed-Rank Statistic (Continued)

<i>n</i>	<i>c</i> ₁	<i>c</i> ₂	α	<i>n</i>	<i>c</i> ₁	<i>c</i> ₂	α	<i>n</i>	<i>c</i> ₁	<i>c</i> ₂	α	<i>n</i>	<i>c</i> ₁	<i>c</i> ₂	α
15	0	120	0.0000	16	0	136	0.0000	17	0	153	0.0000	18	0	171	0.0000
1	119	0.0001		1	135	0.0000		1	152	0.0000		1	170	0.0000	
2	118	0.0001		2	134	0.0000		2	151	0.0000		2	169	0.0000	
3	117	0.0002		3	133	0.0001		3	150	0.0000		3	168	0.0000	
4	116	0.0002		4	132	0.0001		4	149	0.0001		4	167	0.0000	
5	115	0.0003		5	131	0.0002		5	148	0.0001		5	166	0.0000	
6	114	0.0004		6	130	0.0002		6	147	0.0001		6	165	0.0001	
7	113	0.0006		7	129	0.0003		7	146	0.0001		7	164	0.0001	
8	112	0.0008		8	128	0.0004		8	145	0.0002		8	163	0.0001	
9	111	0.0010		9	127	0.0005		9	144	0.0003		9	162	0.0001	
10	110	0.0013		10	126	0.0007		10	143	0.0003		10	161	0.0002	
11	109	0.0017		11	125	0.0008		11	142	0.0004		11	160	0.0002	
12	108	0.0021		12	124	0.0011		12	141	0.0005		12	159	0.0003	
13	107	0.0027		13	123	0.0013		13	140	0.0007		13	158	0.0003	
14	106	0.0034		14	122	0.0017		14	139	0.0008		14	157	0.0004	
15	105	0.0042		15	121	0.0021		15	138	0.0010		15	156	0.0005	
16	104	0.0051		16	120	0.0026		16	137	0.0013		16	155	0.0006	
17	103	0.0062		17	119	0.0031		17	136	0.0016		17	154	0.0008	
18	102	0.0075		18	118	0.0038		18	135	0.0019		19	152	0.0012	
19	101	0.0090		19	117	0.0046		19	134	0.0023		20	151	0.0014	
20	100	0.0108		20	116	0.0055		20	133	0.0028		21	150	0.0017	
21	99	0.0128		21	115	0.0065		21	132	0.0033		22	149	0.0020	
22	98	0.0151		22	114	0.0078		22	131	0.0040		23	148	0.0024	
23	97	0.0177		23	113	0.0091		23	130	0.0047		24	147	0.0028	
24	96	0.0206		24	112	0.0107		24	129	0.0055		25	146	0.0033	
25	95	0.0240		25	111	0.0125		25	128	0.0064		26	145	0.0038	
26	94	0.0277		26	110	0.0145		26	127	0.0075		27	144	0.0045	
27	93	0.0319		27	109	0.0168		27	126	0.0087		28	143	0.0052	
28	92	0.0365		28	108	0.0193		28	125	0.0101		29	142	0.0060	
29	91	0.0416		29	107	0.0222		29	124	0.0116		30	141	0.0069	
30	90	0.0473		30	106	0.0253		30	123	0.0133		31	140	0.0080	
31	89	0.0535		31	105	0.0288		31	122	0.0153		32	139	0.0091	
32	88	0.0603		32	104	0.0327		32	121	0.0174		33	138	0.0104	
33	87	0.0677		33	103	0.0370		33	120	0.0198		34	137	0.0118	
34	86	0.0757		34	102	0.0416		34	119	0.0224		35	136	0.0134	
35	85	0.0844		35	101	0.0467		35	118	0.0253		36	135	0.0152	
36	84	0.0938		36	100	0.0523		36	117	0.0284		37	134	0.0171	
37	83	0.1039		37	99	0.0583		37	116	0.0319		38	133	0.0192	
38	82	0.1147		38	98	0.0649		38	115	0.0357		39	132	0.0216	
39	81	0.1262		39	97	0.0719		39	114	0.0398		40	131	0.0241	
40	80	0.1384		40	96	0.0795		40	113	0.0443		41	130	0.0269	
41	79	0.1514		41	95	0.0877		41	112	0.0492		42	129	0.0300	
				42	94	0.0964		42	111	0.0544		43	128	0.0333	
				43	93	0.1057		43	110	0.0601		44	127	0.0368	
				44	92	0.1156		44	109	0.0662		45	126	0.0407	
				45	91	0.1261		45	108	0.0727		46	125	0.0449	
				46	90	0.1372		46	107	0.0797		47	124	0.0494	
				47	89	0.1489		47	106	0.0871		48	123	0.0542	
								48	105	0.0950		49	122	0.0594	
								49	104	0.1034		50	121	0.0649	
								50	103	0.1123		51	120	0.0708	
								51	102	0.1217		52	119	0.0770	
								52	101	0.1317		53	118	0.0837	
								53	100	0.1421		54	117	0.0907	
								54	99	0.1530		55	116	0.0982	
												56	115	0.1061	
												57	114	0.1144	
												58	113	0.1231	
												59	112	0.1323	
												60	111	0.1419	
												61	110	0.1519	

Table 9 Critical Values for the Wilcoxon Signed-Rank Statistic (Continued)

<i>n</i>	<i>c</i> ₁	<i>c</i> ₂	α	<i>n</i>	<i>c</i> ₁	<i>c</i> ₂	α	<i>n</i>	<i>c</i> ₁	<i>c</i> ₂	α	<i>n</i>	<i>c</i> ₁	<i>c</i> ₂	α
19	0	190	0.0000	19	41	149	0.0145	20	0	210	0.0000	20	41	169	0.0077
	1	189	0.0000		42	148	0.0162		1	209	0.0000		42	168	0.0086
	2	188	0.0000		43	147	0.0180		2	208	0.0000		43	167	0.0096
	3	187	0.0000		44	146	0.0201		3	207	0.0000		44	166	0.0107
	4	186	0.0000		45	145	0.0223		4	206	0.0000		45	165	0.0120
	5	185	0.0000		46	144	0.0247		5	205	0.0000		46	164	0.0133
	6	184	0.0000		47	143	0.0273		6	204	0.0000		47	163	0.0148
	7	183	0.0000		48	142	0.0301		7	203	0.0000		48	162	0.0164
	8	182	0.0000		49	141	0.0331		8	202	0.0000		49	161	0.0181
	9	181	0.0001		50	140	0.0364		9	201	0.0000		50	160	0.0200
	10	180	0.0001		51	139	0.0399		10	200	0.0000		51	159	0.0220
	11	179	0.0001		52	138	0.0437		11	199	0.0001		52	158	0.0242
	12	178	0.0001		53	137	0.0478		12	198	0.0001		53	157	0.0266
	13	177	0.0002		54	136	0.0521		13	197	0.0001		54	156	0.0291
	14	176	0.0002		55	135	0.0567		14	196	0.0001		55	155	0.0319
	15	175	0.0003		56	134	0.0616		15	195	0.0001		56	154	0.0348
	16	174	0.0003		57	133	0.0668		16	194	0.0002		57	153	0.0379
	17	173	0.0004		58	132	0.0723		17	193	0.0002		58	152	0.0413
	18	172	0.0005		59	131	0.0782		18	192	0.0002		59	151	0.0448
	19	171	0.0006		60	130	0.0844		19	191	0.0003		60	150	0.0487
	20	170	0.0007		61	129	0.0909		20	190	0.0004		61	149	0.0527
	21	169	0.0008		62	128	0.0978		21	189	0.0004		62	148	0.0570
	22	168	0.0010		63	127	0.1051		22	188	0.0005		63	147	0.0615
	23	167	0.0012		64	126	0.1127		23	187	0.0006		64	146	0.0664
	24	166	0.0014		65	125	0.1206		24	186	0.0007		65	145	0.0715
	25	165	0.0017		66	124	0.1290		25	185	0.0008		66	144	0.0768
	26	164	0.0020		67	123	0.1377		26	184	0.0010		67	143	0.0825
	27	163	0.0023		68	122	0.1467		27	183	0.0012		68	142	0.0884
	28	162	0.0027		69	121	0.1562		28	182	0.0014		69	141	0.0947
	29	161	0.0031		70	120	0.1660		29	181	0.0016		70	140	0.1012
	30	160	0.0036						30	180	0.0018		71	139	0.1081
	31	159	0.0041						31	179	0.0021		72	138	0.1153
	32	158	0.0047						32	178	0.0024		73	137	0.1227
	33	157	0.0054						33	177	0.0028		74	136	0.1305
	34	156	0.0062						34	176	0.0032		75	135	0.1387
	35	155	0.0070						35	175	0.0036		76	134	0.1471
	36	154	0.0080						36	174	0.0042		77	133	0.1559
	37	153	0.0090						37	173	0.0047				
	38	152	0.0102						38	172	0.0053				
	39	151	0.0115						39	171	0.0060				
	40	150	0.0129						40	170	0.0068				

Table 10 Critical Values for the Wilcoxon Rank-Sum Statistic

This table contains critical values and probabilities for the Wilcoxon rank-sum statistic W = the sum of the ranks of the m observations in the smaller sample: m and n are the sample sizes, and c_1 and c_2 are defined by $P(W \leq c_1) = \alpha$ and $P(W \geq c_2) = \alpha$, respectively.

m	n	c_1	c_2	α	m	n	c_1	c_2	α	m	n	c_1	c_2	α	m	n	c_1	c_2	α
2	3	3	9	0.1000	3	8	6	30	0.0061	4	7	10	38	0.0030	5	5	15	40	0.0040
2	4	3	11	0.0667			7	29	0.0121			11	37	0.0061			16	39	0.0079
		4	10	0.1333			8	28	0.0242			12	36	0.0121			17	38	0.0159
2	5	3	13	0.0476			9	27	0.0424			13	35	0.0212			18	37	0.0278
		4	12	0.0952			10	26	0.0667			14	34	0.0364			19	36	0.0476
2	6	3	15	0.0357			11	25	0.0970			15	33	0.0545			20	35	0.0754
		4	14	0.0714	3	9	12	24	0.1394			16	32	0.0818			21	34	0.1111
		5	13	0.1429								17	31	0.1152			22	33	0.1548
2	7	3	17	0.0278			8	31	0.0182	4	8	10	42	0.0020	5	6	15	45	0.0022
		4	16	0.0556			9	30	0.0318			11	41	0.0040			16	44	0.0043
		5	15	0.1111			10	29	0.0500			12	40	0.0081			17	43	0.0087
2	8	3	19	0.0222			11	28	0.0727			13	39	0.0141			18	42	0.0152
		4	18	0.0444			12	27	0.1045			14	38	0.0242			19	41	0.0260
		5	17	0.0889	3	10	13	26	0.1409			15	37	0.0364			20	40	0.0411
		6	16	0.1333			6	36	0.0035			16	36	0.0545			21	39	0.0628
2	9	3	21	0.0182			7	35	0.0070			17	35	0.0768	5	7	15	50	0.0013
		4	20	0.0364			8	34	0.0140			18	34	0.1071			16	49	0.0025
		5	19	0.0727			9	33	0.0245			19	33	0.1414			17	48	0.0051
		6	18	0.1091			10	32	0.0385	4	9	10	46	0.0014			18	47	0.0088
2	10	3	23	0.0152			11	31	0.0559			11	45	0.0028			19	46	0.0152
		4	22	0.0303			12	30	0.0804			12	44	0.0056			20	45	0.0240
		5	21	0.0606	4	4	13	29	0.1084			13	43	0.0098			21	44	0.0366
		6	20	0.0909			14	28	0.1434			14	42	0.0168			22	43	0.0530
		7	19	0.1364								15	41	0.0252			23	42	0.0745
3	3	6	15	0.0500			13	23	0.1000			16	40	0.0378			24	41	0.1010
		7	14	0.1000	4	5	10	30	0.0079			17	39	0.0531			25	40	0.1338
3	4	6	18	0.0286			11	29	0.0159			18	38	0.0741	5	8	15	55	0.0008
		7	17	0.0571			12	28	0.0317			19	37	0.0993			16	54	0.0016
		8	16	0.1143			13	27	0.0556	4	10	10	50	0.0010			17	53	0.0031
3	5	6	21	0.0179			14	26	0.0952			11	49	0.0020			18	52	0.0054
		7	20	0.0357			15	25	0.1429			12	48	0.0040			19	51	0.0093
		8	19	0.0714	4	6	10	34	0.0048			13	47	0.0070			20	50	0.0148
		9	18	0.1250			11	33	0.0095			14	46	0.0120			21	49	0.0225
3	6	6	24	0.0119			12	32	0.0190			15	45	0.0180			22	48	0.0326
		7	23	0.0238			13	31	0.0333			16	44	0.0270			23	47	0.0466
		8	22	0.0476			14	30	0.0571			17	43	0.0380			24	46	0.0637
		9	21	0.0833			15	29	0.0857			18	42	0.0529			25	45	0.0855
		10	20	0.1310			16	28	0.1286			19	41	0.0709			26	44	0.1111
3	7	6	27	0.0083						20	40	0.0939			27	43	0.1422		
		7	26	0.0167						21	39	0.1199							
		8	25	0.0333						22	38	0.1518							
		9	24	0.0583															
		10	23	0.0917															
		11	22	0.1333															

Table 10 Critical Values for the Wilcoxon Rank-Sum Statistic (Continued)

<i>m</i>	<i>n</i>	<i>c</i> ₁	<i>c</i> ₂	α	<i>m</i>	<i>n</i>	<i>c</i> ₁	<i>c</i> ₂	α	<i>m</i>	<i>n</i>	<i>c</i> ₁	<i>c</i> ₂	α	<i>m</i>	<i>n</i>	<i>c</i> ₁	<i>c</i> ₂	α
5	9	15	60	0.0005	6	7	21	63	0.0006	6	10	21	81	0.0001	7	8	28	84	0.0002
		16	59	0.0010			22	62	0.0012			22	80	0.0002			29	83	0.0003
		17	58	0.0020			23	61	0.0023			23	79	0.0005			30	82	0.0006
		18	57	0.0035			24	60	0.0041			24	78	0.0009			31	81	0.0011
		19	56	0.0060			25	59	0.0070			25	77	0.0015			32	80	0.0019
		20	55	0.0095			26	58	0.0111			26	76	0.0024			33	79	0.0030
		21	54	0.0145			27	57	0.0175			27	75	0.0037			34	78	0.0047
		22	53	0.0210			28	56	0.0256			28	74	0.0055			35	77	0.0070
		23	52	0.0300			29	55	0.0367			29	73	0.0080			36	76	0.0103
		24	51	0.0415			30	54	0.0507			30	72	0.0112			37	75	0.0145
		25	50	0.0559			31	53	0.0688			31	71	0.0156			38	74	0.0200
		26	49	0.0734			32	52	0.0903			32	70	0.0210			39	73	0.0270
		27	48	0.0949			33	51	0.1171			33	69	0.0280			40	72	0.0361
		28	47	0.1199			34	50	0.1474			34	68	0.0363			41	71	0.0469
		29	46	0.1489	6	8	21	69	0.0003			35	67	0.0467			42	70	0.0603
5	10	15	65	0.0003			22	68	0.0007			36	66	0.0589			43	69	0.0760
		16	64	0.0007			23	67	0.0013			37	65	0.0736			44	68	0.0946
		17	63	0.0013			24	66	0.0023			38	64	0.0903			45	67	0.1159
		18	62	0.0023			25	65	0.0040			39	63	0.1099			46	66	0.1405
		19	61	0.0040			26	64	0.0063			40	62	0.1317	7	9	28	91	0.0001
		20	60	0.0063			27	63	0.0100			41	61	0.1566			29	90	0.0002
		21	59	0.0097			28	62	0.0147	7	7	28	77	0.0003			30	89	0.0003
		22	58	0.0140			29	61	0.0213			29	76	0.0006			31	88	0.0006
		23	57	0.0200			30	60	0.0296			30	75	0.0012			32	87	0.0010
		24	56	0.0276			31	59	0.0406			31	74	0.0020			33	86	0.0017
		25	55	0.0376			32	58	0.0539			32	73	0.0035			34	85	0.0026
		26	54	0.0496			33	57	0.0709			33	72	0.0055			35	84	0.0039
		27	53	0.0646			34	56	0.0906			34	71	0.0087			36	83	0.0058
		28	52	0.0823			35	55	0.1142			35	70	0.0131			37	82	0.0082
		29	51	0.1032			36	54	0.1412			36	69	0.0189			38	81	0.0115
		30	50	0.1272	6	9	21	75	0.0002			37	68	0.0265			39	80	0.0156
		31	49	0.1548			22	74	0.0004			38	67	0.0364			40	79	0.0209
6	6	21	57	0.0011			23	73	0.0008			39	66	0.0487			41	78	0.0274
		22	56	0.0022			24	72	0.0014			40	65	0.0641			42	77	0.0356
		23	55	0.0043			25	71	0.0024			41	64	0.0825			43	76	0.0454
		24	54	0.0076			26	70	0.0038			42	63	0.1043			44	75	0.0571
		25	53	0.0130			27	69	0.0060			43	62	0.1297			45	74	0.0708
		26	52	0.0206			28	68	0.0088			44	61	0.1588			46	73	0.0869
		27	51	0.0325			29	67	0.0128							47	72	0.1052	
		28	50	0.0465			30	66	0.0180							48	71	0.1261	
		29	49	0.0660			31	65	0.0248							49	70	0.1496	
		30	48	0.0898			32	64	0.0332										
		31	47	0.1201			33	63	0.0440										
		32	46	0.1548			34	62	0.0567										
							35	61	0.0723										
							36	60	0.0905										
							37	59	0.1119										
							38	58	0.1361										

Table 10 Critical Values for the Wilcoxon Rank-Sum Statistic (Continued)

<i>m</i>	<i>n</i>	<i>c</i> ₁	<i>c</i> ₂	α	<i>m</i>	<i>n</i>	<i>c</i> ₁	<i>c</i> ₂	α	<i>m</i>	<i>n</i>	<i>c</i> ₁	<i>c</i> ₂	α	<i>m</i>	<i>n</i>	<i>c</i> ₁	<i>c</i> ₂	α
7	10	28	98	0.0001	8	9	36	108	0.0000	9	9	45	126	0.0000	10	10	55	155	0.0000
		29	97	0.0001			37	107	0.0001			46	125	0.0000			56	154	0.0000
7	11	30	96	0.0002	8		38	106	0.0002	9		47	124	0.0001	10		57	153	0.0000
		31	95	0.0004			39	105	0.0003			48	123	0.0001			58	152	0.0000
7	12	32	94	0.0006	8		40	104	0.0005	9		49	122	0.0002	10		59	151	0.0001
		33	93	0.0010			41	103	0.0008			50	121	0.0004			60	150	0.0001
7	13	34	92	0.0015	8		42	102	0.0012	9		51	120	0.0006	10		61	149	0.0002
		35	91	0.0023			43	101	0.0019			52	119	0.0009			62	148	0.0002
7	14	36	90	0.0034	8		44	100	0.0028	9		53	118	0.0014	10		63	147	0.0004
		37	89	0.0048			45	99	0.0039			54	117	0.0020			64	146	0.0005
7	15	38	88	0.0068	8		46	98	0.0056	9		55	116	0.0028	10		65	145	0.0008
		39	87	0.0093			47	97	0.0076			56	115	0.0039			66	144	0.0010
7	16	40	86	0.0125	8		48	96	0.0103	9		57	114	0.0053	10		67	143	0.0014
		41	85	0.0165			49	95	0.0137			58	113	0.0071			68	142	0.0019
7	17	42	84	0.0215	8		50	94	0.0180	9		59	112	0.0094	10		69	141	0.0026
		43	83	0.0277			51	93	0.0232			60	111	0.0122			70	140	0.0034
7	18	44	82	0.0351	8		52	92	0.0296	9		61	110	0.0157	10		71	139	0.0045
		45	81	0.0439			53	91	0.0372			62	109	0.0200			72	138	0.0057
7	19	46	80	0.0544	8		54	90	0.0464	9		63	108	0.0252	10		73	137	0.0073
		47	79	0.0665			55	89	0.0570			64	107	0.0313			74	136	0.0093
7	20	48	78	0.0806	8		56	88	0.0694	9		65	106	0.0385	10		75	135	0.0116
		49	77	0.0966			57	87	0.0836			66	105	0.0470			76	134	0.0144
7	21	50	76	0.1148	8		58	86	0.0998	9		67	104	0.0567	10		77	133	0.0177
		51	75	0.1349			59	85	0.1179			68	103	0.0680			78	132	0.0216
7	22	52	74	0.1574	8		60	84	0.1383	9		69	102	0.0807	10		79	131	0.0262
															10	130	0.0315		
8	8	36	100	0.0001	8	10	36	116	0.0000	9	10	70	101	0.0951	10		80	130	0.0315
		37	99	0.0002			37	115	0.0000			71	100	0.1112			81	129	0.0376
8	9	38	98	0.0003	8		38	114	0.0001	9		72	99	0.1290	10		82	128	0.0446
		39	97	0.0005			39	113	0.0002			9	10	45	135	0.0000		83	127
8	10	40	96	0.0009	8		40	112	0.0003	9		46	134	0.0000	10		84	126	0.0615
		41	95	0.0015			41	111	0.0004			47	133	0.0000			85	125	0.0716
8	11	42	94	0.0023	8		42	110	0.0007	9		48	132	0.0001	10		86	124	0.0827
		43	93	0.0035			43	109	0.0010			49	131	0.0001			87	123	0.0952
8	12	44	92	0.0052	8		44	108	0.0015	9		50	130	0.0002	10		88	122	0.1088
		45	91	0.0074			45	107	0.0022			51	129	0.0003			89	121	0.1237
8	13	46	90	0.0103	8		46	106	0.0031	9		52	128	0.0005	10		90	120	0.1399
		47	89	0.0141			47	105	0.0043			53	127	0.0007			91	119	0.1575
8	14	48	88	0.0190	8		48	104	0.0058	9		54	126	0.0011	10		92	118	0.1758
		49	87	0.0249			49	103	0.0078			55	125	0.0015			93	117	0.1943
8	15	50	86	0.0325	8		50	102	0.0103	9		56	124	0.0021	10		94	116	0.2130
		51	85	0.0415			51	101	0.0133			57	123	0.0028			95	115	0.2315
8	16	52	84	0.0524	8		52	100	0.0171	9		58	122	0.0038	10		96	114	0.2500
		53	83	0.0652			53	99	0.0217			59	121	0.0051			97	113	0.2685
8	17	54	82	0.0803	8		54	98	0.0273	9		60	120	0.0066	10		98	112	0.2870
		55	81	0.0974			55	97	0.0338			61	119	0.0086			99	111	0.3055
8	18	56	80	0.1172	8		56	96	0.0416	9		62	118	0.0110	10		100	110	0.3240
		57	79	0.1393			57	95	0.0506			63	117	0.0140			101	109	0.3425
8	19				8		58	94	0.0610	9		64	116	0.0175	10		102	108	0.3610
							59	93	0.0729			65	115	0.0217			103	107	0.3795
8	20				8		60	92	0.0864	9		66	114	0.0267	10		104	106	0.3980
							61	91	0.1015			67	113	0.0326			105	105	0.4165
8	21				8		62	90	0.1185			68	112	0.0394	10		106	104	0.4350
							63	89	0.1371			69	111	0.0474			107	103	0.4535
8	22				8		64	88	0.1577			70	110	0.0564	10		108	102	0.4720
											71	109	0.0667		109	101	0.4905		
8	23				8						72	108	0.0782		110	99	0.5090		

Table 11 Critical Values for the Runs Test

This table contains cumulative probabilities associated with the runs test. Let m be the number of observations in one category, n be the number of observations in the other category ($m \leq n$), and V be the number of runs. The values in this table are the probabilities $P(V \leq v)$ if the order of observations is random.

m	n	2	3	4	5	v	6	7	8	9
2	2	0.33333	0.66667	1.0000						
2	3	0.20000	0.50000	0.90000	1.00000					
2	4	0.13333	0.40000	0.80000	1.00000					
2	5	0.0952	0.33333	0.7143	1.0000					
2	6	0.0714	0.2857	0.6429	1.0000					
2	7	0.0556	0.2500	0.5833	1.0000					
2	8	0.0444	0.22222	0.5333	1.0000					
2	9	0.0364	0.20000	0.4909	1.0000					
2	10	0.0303	0.1818	0.4545	1.0000					
3	3	0.10000	0.30000	0.7000	0.9000	1.0000				
3	4	0.0571	0.20000	0.5429	0.8000	0.9714	1.0000			
3	5	0.0357	0.1429	0.4286	0.7143	0.9286	1.0000			
3	6	0.0238	0.1071	0.3452	0.6429	0.8810	1.0000			
3	7	0.0167	0.0833	0.2833	0.5833	0.8333	1.0000			
3	8	0.0121	0.0667	0.2364	0.5333	0.7879	1.0000			
3	9	0.0091	0.0545	0.2000	0.4909	0.7455	1.0000			
3	10	0.0070	0.0455	0.1713	0.4545	0.7063	1.0000			
4	4	0.0286	0.1143	0.3714	0.6286	0.8857	0.9714	1.0000		
4	5	0.0159	0.0714	0.2619	0.5000	0.7857	0.9286	0.9921	1.0000	
4	6	0.0095	0.0476	0.1905	0.4048	0.6905	0.8810	0.9762	1.0000	
4	7	0.0061	0.0333	0.1424	0.3333	0.6061	0.8333	0.9545	1.0000	
4	8	0.0040	0.0242	0.1091	0.2788	0.5333	0.7879	0.9293	1.0000	
4	9	0.0028	0.0182	0.0853	0.2364	0.4713	0.7455	0.9021	1.0000	
4	10	0.0020	0.0140	0.0679	0.2028	0.4186	0.7063	0.8741	1.0000	

Table 11 Critical Values for the Runs Test (Continued)

m	n	v									
		2	3	4	5	6	7	8	9	10	11
5	5	0.0079	0.0397	0.1667	0.3571	0.6429	0.8333	0.9603	0.9921	1.0000	
5	6	0.0043	0.0238	0.1104	0.2619	0.5216	0.7381	0.9113	0.9762	0.9978	1.0000
5	7	0.0025	0.0152	0.0758	0.1970	0.4242	0.6515	0.8335	0.9545	0.9924	1.0000
5	8	0.0016	0.0101	0.0536	0.1515	0.3473	0.5758	0.7933	0.9293	0.9837	1.0000
5	9	0.0010	0.0070	0.0390	0.1189	0.2867	0.5105	0.7343	0.9021	0.9720	1.0000
5	10	0.0007	0.0050	0.0290	0.0949	0.2388	0.4545	0.6783	0.8741	0.9580	1.0000
6	6	0.0022	0.0130	0.0671	0.1753	0.3918	0.6082	0.8247	0.9329	0.9870	0.9978
6	7	0.0012	0.0076	0.0425	0.1212	0.2960	0.5000	0.7331	0.8788	0.9662	0.9924
6	8	0.0007	0.0047	0.0280	0.0862	0.2261	0.4126	0.6457	0.8205	0.9371	0.9837
6	9	0.0004	0.0030	0.0190	0.0629	0.1748	0.3427	0.5664	0.7622	0.9021	0.9720
6	10	0.0002	0.0020	0.0132	0.0470	0.1369	0.2867	0.4965	0.7063	0.8636	0.9580
7	7	0.0006	0.0041	0.0251	0.0775	0.2086	0.3834	0.6166	0.7914	0.9225	0.9749
7	8	0.0003	0.0023	0.0154	0.0513	0.1492	0.2960	0.5136	0.7040	0.8671	0.9487
7	9	0.0002	0.0014	0.0098	0.0350	0.1084	0.2308	0.4266	0.6224	0.8059	0.9161
7	10	0.0001	0.0009	0.0064	0.0245	0.0800	0.1818	0.3546	0.5490	0.7433	0.8794
8	8	0.0002	0.0012	0.0089	0.0317	0.1002	0.2145	0.4048	0.5952	0.7855	0.8998
8	9	0.0001	0.0007	0.0053	0.0203	0.0687	0.1573	0.3186	0.5000	0.7016	0.8427
8	10	0.0000	0.0004	0.0033	0.0134	0.0479	0.1170	0.2514	0.4194	0.6209	0.7822
9	9	0.0000	0.0004	0.0030	0.0122	0.0445	0.1090	0.2380	0.3992	0.6008	0.7620
9	10	0.0000	0.0002	0.0018	0.0076	0.0294	0.0767	0.1786	0.3186	0.5095	0.6814
10	10	0.0000	0.0001	0.0010	0.0045	0.0185	0.0513	0.1276	0.2422	0.4141	0.5859

Table 12 Greek Alphabet

This table contains the Greek alphabet: the letter name, the lowercase letter, the variant of the lowercase letter where applicable, and the uppercase letter.

Name	Lowercase letter	Lowercase variant	Uppercase letter
Alpha	α		A
Beta	β		B
Gamma	γ		Γ
Delta	δ		Δ
Epsilon	ε	ε	E
Zeta	ζ		Z
Eta	η		H
Theta	θ	ϑ	Θ
Iota	ι		I
Kappa	κ		K
Lambda	λ		Λ
Mu	μ		M
Nu	ν		N
Xi	ξ		Ξ
Omicron	\o		O
Pi	π	ϖ	Π
Rho	ρ	ϱ	R
Sigma	σ	ς	Σ
Tau	τ		T
Upsilon	υ		Υ
Phi	ϕ	φ	Φ
Chi	χ		X
Psi	ψ		Ψ
Omega	ω		Ω

