BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
WORK INTEGRATED LEARNING PROGRAMMES

M Tech(AIML)

Course Handout(w.e.f. October 2022)

Part A: Content Design

Course Title	**Introduction to Statistical Methods**
Course No(s)	AIML ZC418
Credit Units	4
Credit Model	2 - 0.5 - 1.5. 2units for class room hours, 0.5 unit for Tutorial, 1.5 units for Student preparation. 1 unit = 32 hours
Content Authors	Dr YVK Ravi Kumar
Date	October ,2022

Course Description

Basic probability concepts, Conditional probability, Bayes Theorem, Probability distributions, Continuous and discrete distributions, Transformation of random variables, estimating mean, variance, covariance, Hypothesis Testing, Maximum likelihood, ANOVA – single factor, dual factor, time series analysis: AR, MA, ARIMA, SARIMA, sampling based on distribution, statistical significance, Gaussian Mixture Model, Expectation Maximization.

Course Objectives

CO1	Introducing basic concepts of probability and statistics to students
CO2	Students will be able to apply statistical techniques to understand the dats
CO3	Students will be able to do statistical analysis of the model / algorithm

Text Books

No	Author(s), Title, Edition, Publishing House
T1	Statistics for Data Scientsists, An introduction to probability ,statistics and Data Analysis,MauritsKaptein 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.

Reference Books

No	Author(s), Title, Edition, Publishing House
R1	Miller and Freund's Probability and statistics for Engineers, 8th Edition,PHI
R2	Statistics for Business and Economics by Anderson, Sweeney and Wiliams, CENAGE learning

Modular Content Structure

Module 1:Basic Probability &Statisitcs

1.1.Measures of Central Tendency

1.2.Measures of Variability

1.3. Basic Probability concepts

1.3.1.Axioms of Probability

1.3.2.Definition of Probability

1.3.3.Mutually exclusive and independent events

Module 2: Conditional Probability & Bayes theorem

2.1. Conditional Probability

2.2. Conditonal Probability of Independent events

2.3. Bayes Theorem

2.4. Introduction to Naïve Bayes concept

Module 3: Probability Distributions

3.1. Random Variables

3.1.1. Discrete random variable – Single and two variables

3.1.2. Discrete random variable – Single and two variables

3.1.3. Mean, Variance, Co – Variance of Random variables

3.1.3. Transformation of random variables

3.2. Probability Distributions

3.2.1. Bernoulli Distribution

3.2.2. Binomial Distribution

3.2.3. Poisson Distribution

3.2.4. Normal(Gaussian) distribution

3.2.5. Introduction of t – distribution , F – distribution , Chi Sqaure distribution.

Module 4: Hypothesis Testing

4.1. Sampling – random sampling and Stratified sampling

4.2. Sampling distribution – Cental Limit theorem

4.3. Estimation – Interval Estimation,Confidence level

4.4. Testing of Hypothesis

4.4.1.Mean based

4.4.2. Proportions related

4.4.3. ANOVA – Single and dual factor

4.5. Maximum likelihood

Module 5:Prediction & Forecasting

5.1. Correlation

5.2. Regression

5.3. Time Series Analysis

5.3.1. Introduction, Components of time series data

5.3.2.MA model – basic and weighted MA model

5.3.3.Time series models

5.3.3.1.AR Model

5.3.3.2. ARIMA Model

5.3.3.3. SARIMA,SARIMAX,VAR,VARMAX

5.3.3.4. Simple exponential smoothing model

Module 6: Gaussian Mixture model & Expectation Maximization

Learning Outcomes:

No	Learning Outcomes
LO1	Clear understanding of the various statistical models to model the data
LO2	Drawing conclusions from the models selected to understand the data

Part B: Course Handout

Academic Term	I semester ,2022 – 23
Course Title	**Introduction to Statistical Methods**

Course No	AIML ZC418 / DSECT ZC418

Course Contents

Contact Session 1: Module 1(Module 1:Basic Probability &Statisitcs)

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
HW	Problems on data symmetry & outlier identification	T1& T2
Lab	Display of statistical Data & Understanding the statistical summary	Lab 1

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

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		

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		

Contact Session 4: Module 2(Conditional Probability & Bayes theorem)

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

Contact Session 5: Module 3: Probability Distributions

Contact Session	List of Topic Title	Reference
CS - 5	Random variables - Discrete & continuous Expectaion of a random variable,mean and variance of a random variable – Sinlge random random variable& Joint distributions	T1& T2
HW	Problems on random variables	T1& T2
Lab	Probability Distributions & Sampling	Lab 3

Contact Session 6: Module 3: Probability Distributions

Contact Session	List of Topic Title	Reference
CS - 6	Bernoulli,Binomial, Poisson and Norma distributions. Inroduction to t – distribution , F – Distribution and Chi Sqaure distributions	T1& T2
HW	Problems on Probability distributions	T1& T2
Lab	Probability Distributions & Sampling	Lab 3

Contact Session 7: Module 4: Hypothesis Testing

Contact Session	List of Topic Title	Reference
CS - 7	Sampling – random sampling and Stratified sampling,Sampling distribution – Cental Limit theorem,Estimation– Interval Estimation,Confidence level	T1& T2

HW	Problems on Interval estimation	T1& T2
Lab		

Contact Session 8:

Contact Session	List of Topic Title	Reference
CS - 8	REVISION OF THE TOPICS COVERED	
HW		
Lab		

MID SEMESTER EXAMINATION

Contact Session 9 :Module 4: Hypothesis Testing

Contact Session	List of Topic Title	Reference
CS –9	Testing of Hypothesis -mean and proportions related models (one mean, two mean, one proportion and Several proportions with small and big samples wherever applicable)	T1& T2
HW	Problems on Testing of Hypothesis	T1& T2
Lab		

Contact Session10: Module 4: Hypothesis Testing

Contact Session	List of Topic Title	Reference
CS – 10	Maximum likelihood ,ANOVA – Single and dual factor	T1& T2
HW	Problems on ANOVA	T1& T2
Lab	ANOVA	Lab 4

Contact Session 11: Module 5: Prediction & Forecasting

Contact Session	List of Topic Title	Reference
CS - 11	Correlation & regression	T1& T2
HW	Problems on correlation and regression	T1& T2
Lab	Regression – Analysis of model summary	Lab 5

Contact Session 12: :Module 5: Prediction & Forecasting

Contact Session	List of Topic Title	Reference
CS - 12	Time Series Analysis - Introduction, Components of time series data,Moving Averages and weighted moving averges model	T3
HW	Problems on Time series	T3
Lab	Time series	Lab 6

Contact Session 13: Module 5: Prediction & Forecasting

Contact Session	List of Topic Title	Reference
CS - 13	Time series models - AR Model,ARMAModel,ARIMA	T3
HW	Problems on Time Series	T3
Lab	Time series	Lab 6

Contact Session 14: Module 5: Prediction & Forecasting

Contact Session	List of Topic Title	Reference
CS - 14	Time Series Models - SARIMA,SARIMAX,VAR,VARMAX and Simple exponential smoothing model	T3

HW	Problems on Time Series	
Lab	Time series	Lab 6

Contact Session 15: Module 6: Gaussian Mixture model & Expectation Maximization

Contact Session	List of Topic Title	Reference
CS - 15	Gaussian Mixture model & Expectation Maximization	Class Notes
HW	Problems on GMM & EM	
Lab		

Contact Session 16:

Contact Session	List of Topic Title	Reference
CS - 16	**REVISION OF THE SYLLABUS**	
HW		
Lab		

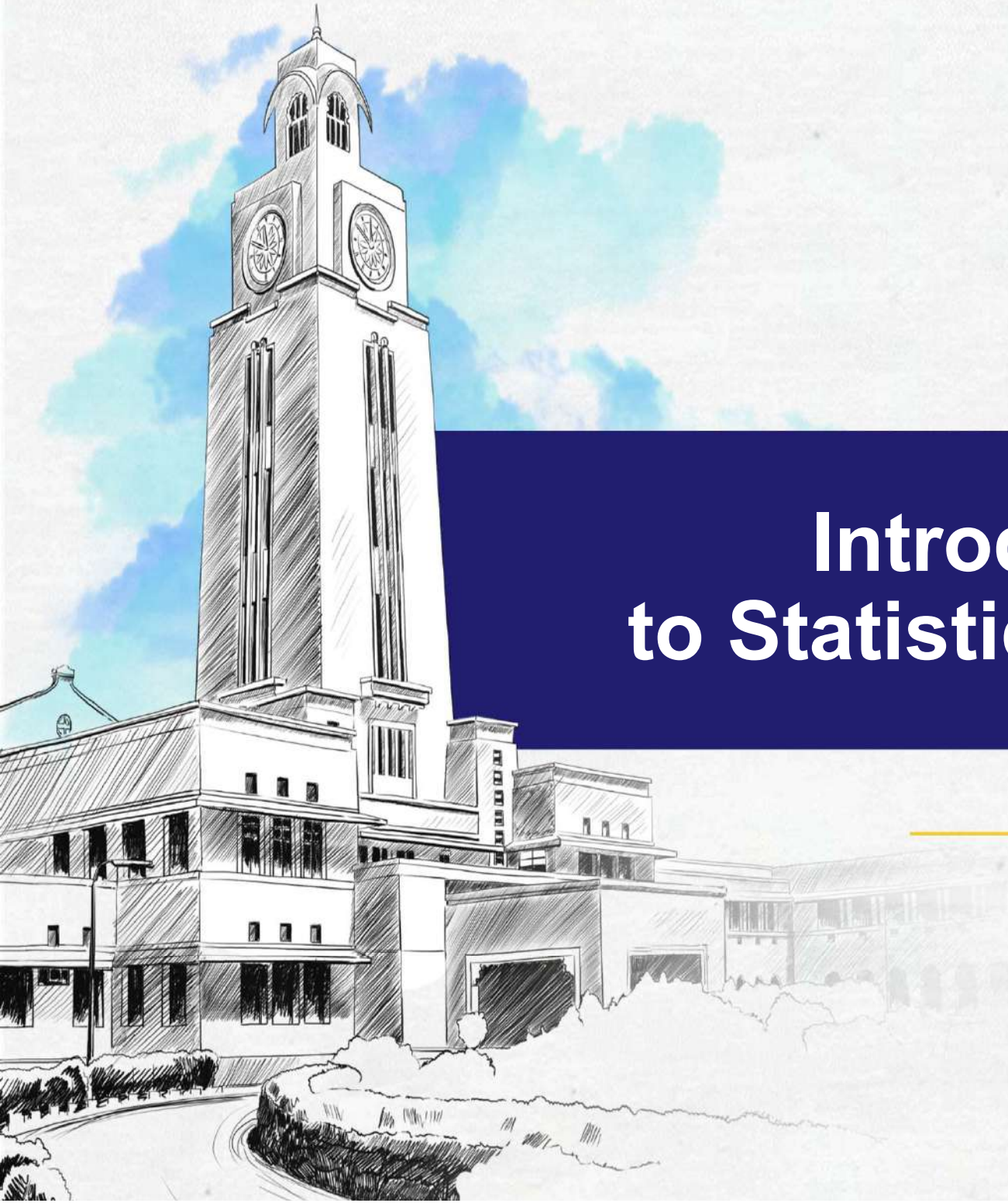
Detailed Plan for Lab work

Lab No.	Lab Objective	Lab Sheet Access URL	Session Reference
1	Display of statistical Data& Understanding the statistical summary		1
2	Bayes theorem & Naïve Bayes Concept		4
3	Probability Distributions& Sampling		5 & 6
4	ANOVA		10
5	Regression – Analysis of model summary		11
6	Time series		12 - 14

Evaluation Scheme:
Legend: EC = Evaluation Component; AN = After Noon Session; FN = Fore Noon Session

No	Name	Type	Duration	Weight	Day, Date, Session, Time
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%	

Note:
Syllabus for Mid-Semester Test (Closed Book): Topics in Session Nos. 1 to 8
Syllabus for Comprehensive Exam (Open Book): All topics (Session Nos. 1 to 16)



BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Statistical Methods

ISM Team

innovate achieve lead

BITS Pilani

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

(Session 1: 20th /21st May,2023)

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 & 3 (Best two will be considered)	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

innovate achieve lead

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

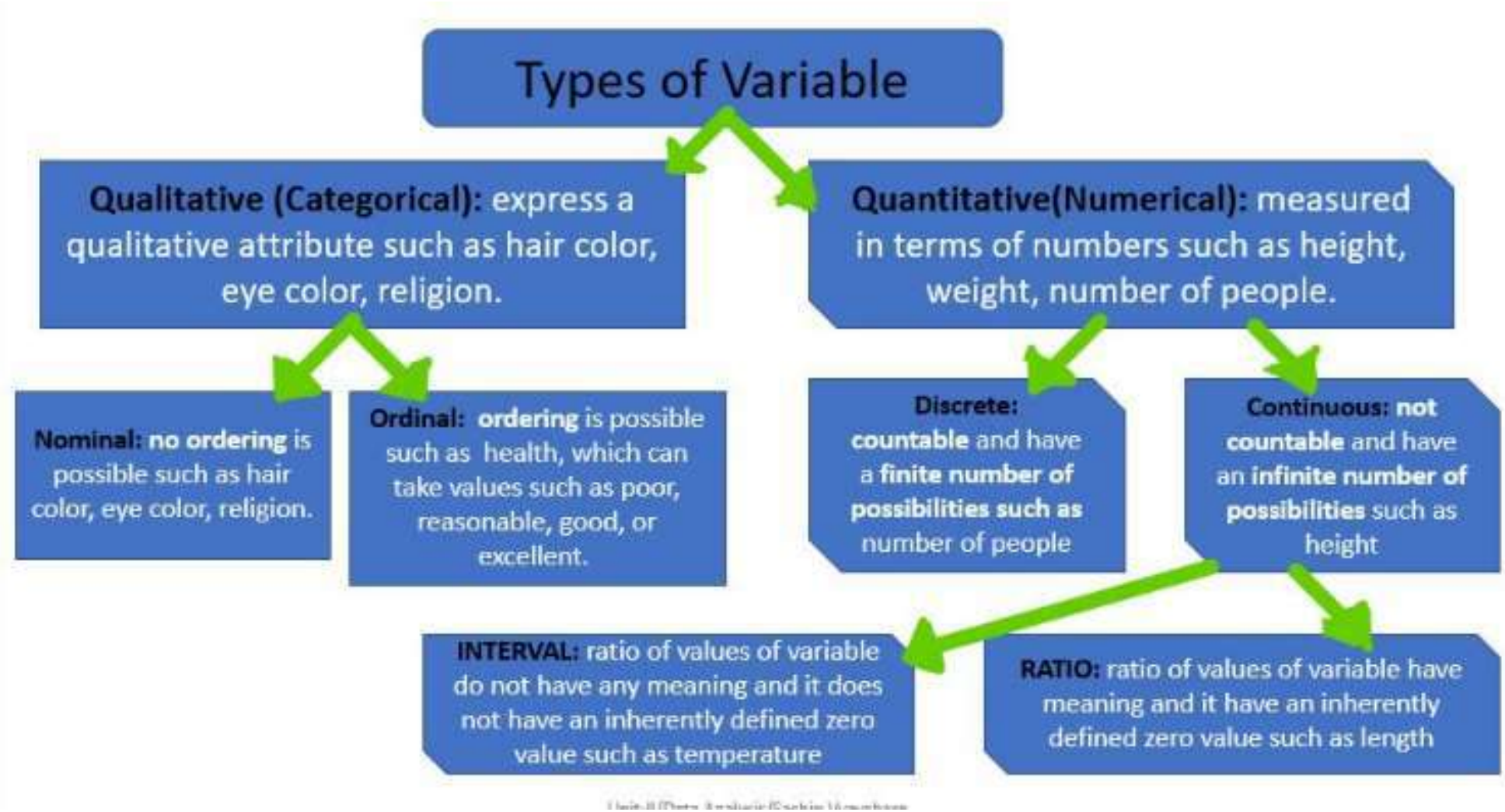
H G Wells

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:

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

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

- Computation of the Mean for grouped Data

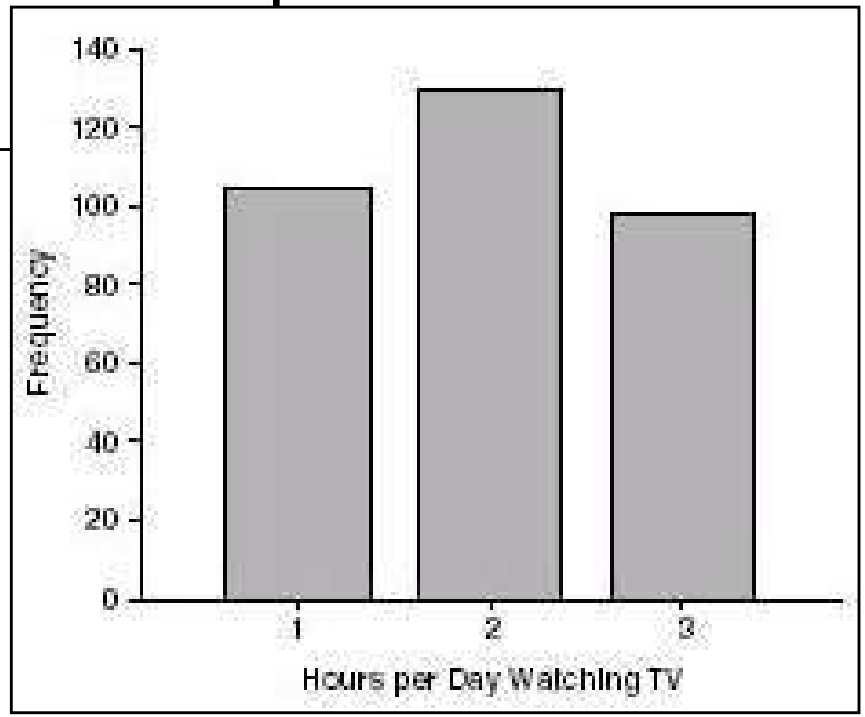
$$\bar{Y} = \frac{\sum f Y}{N}$$ Where f Y = a score multiplied by its frequency

Mean: Grouped Scores

Hours Spent Watching TV	Frequency (f)	fY	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

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

μ

$n-1$ $\bar{x}$

Drawback?

- the median

10, 15, 20, 25, 26

10, 15, 20, 25, 28, 32

Average of these

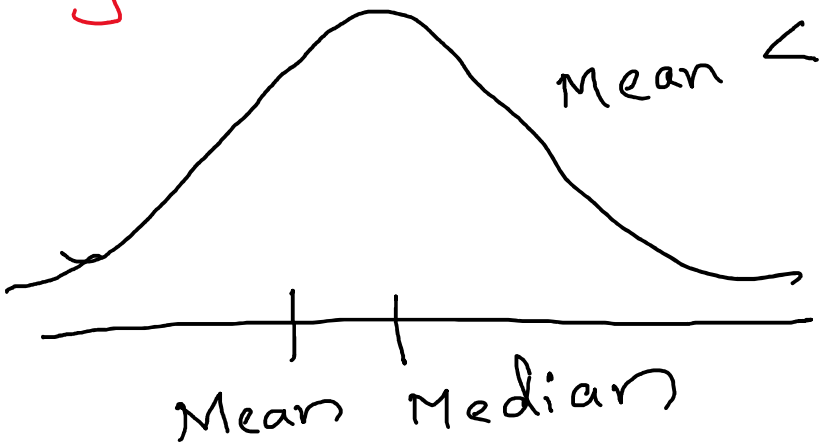
- the mode

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

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

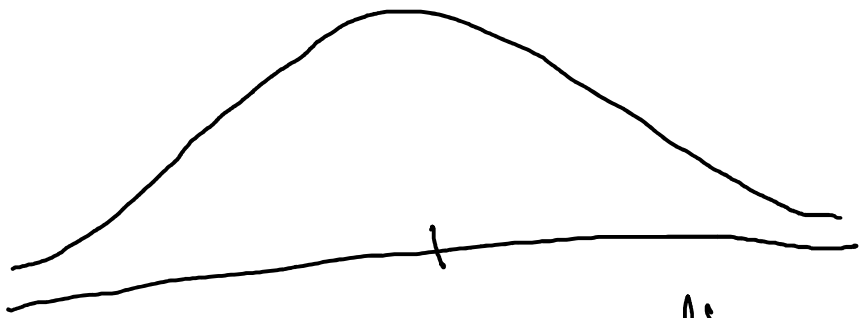
Data : Symmetrical and Asymmetrical

Negative Skew

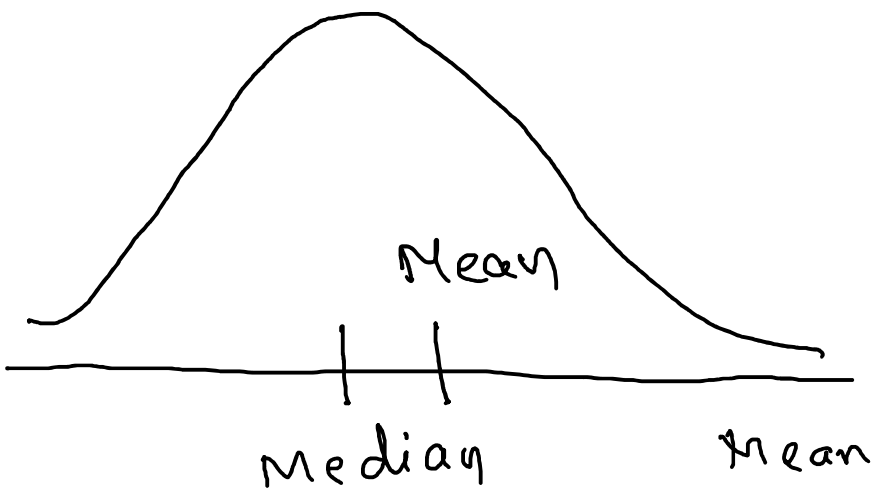


Mean < Median

symmetric



Mean = Median



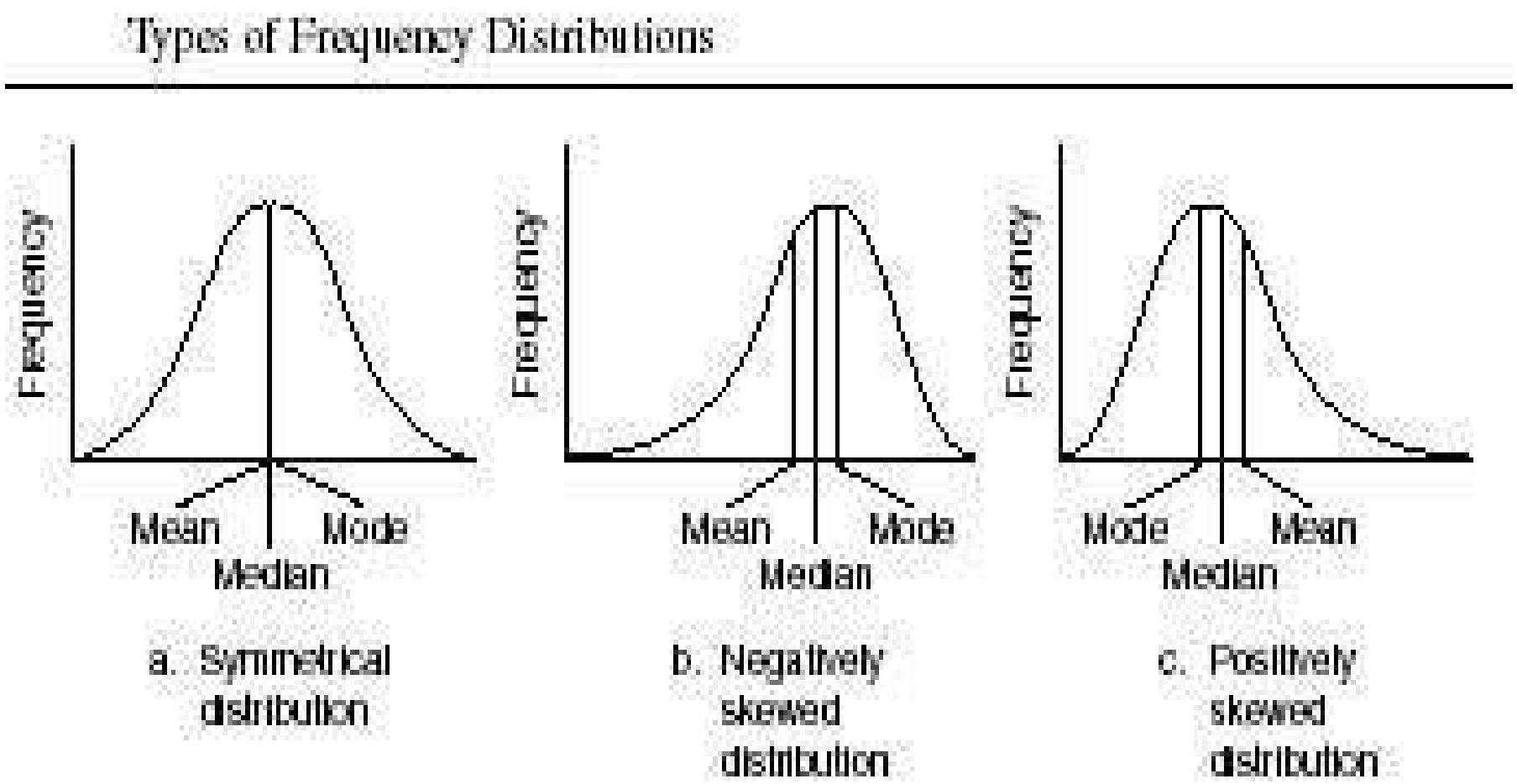
Positive Skew

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



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

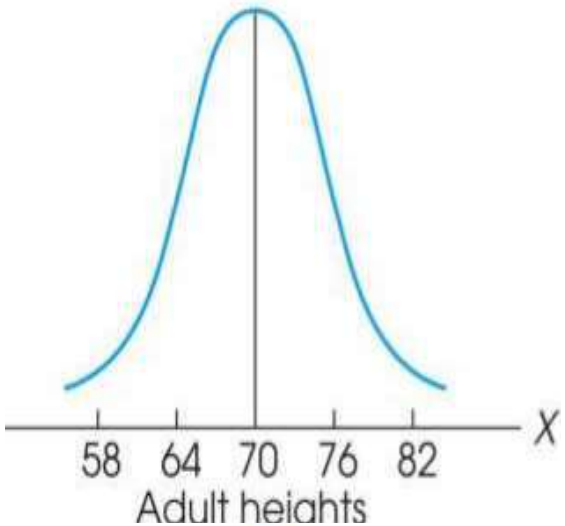
Three Measures of Variability:

- The Range
- The Variance
- The Standard Deviations

Measure of Variability

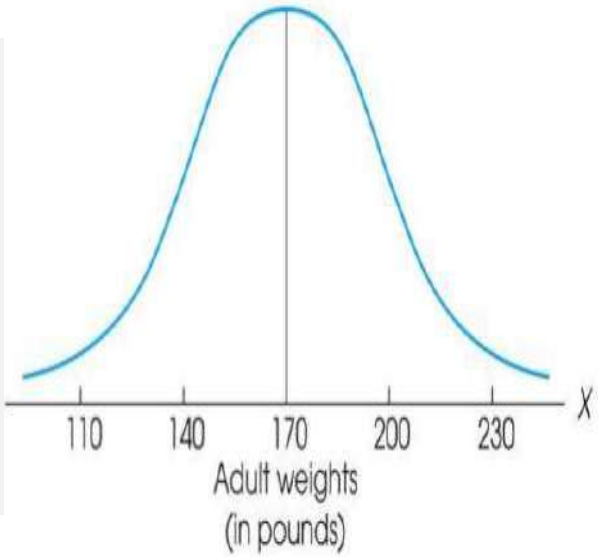
Variability can be defined several ways:

- 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

Range = URL for X_{max} - LRL for X_{min}

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

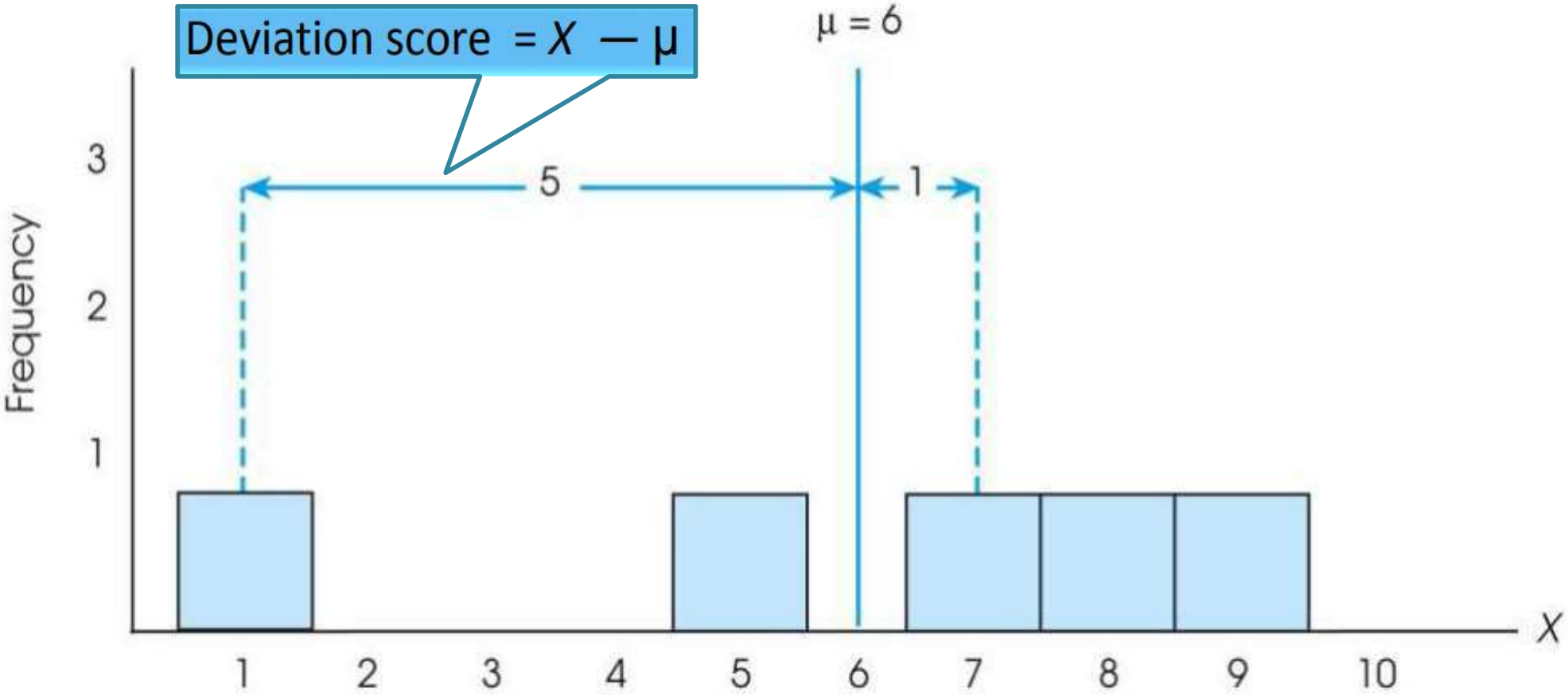
Example: For a set of scores: 7, 2, 7, 6, 5, 6, 2

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

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{45}{5} = 5$$

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

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

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

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{45}{5} = 5$$

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

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

Standard Deviation and Variance for a Sample

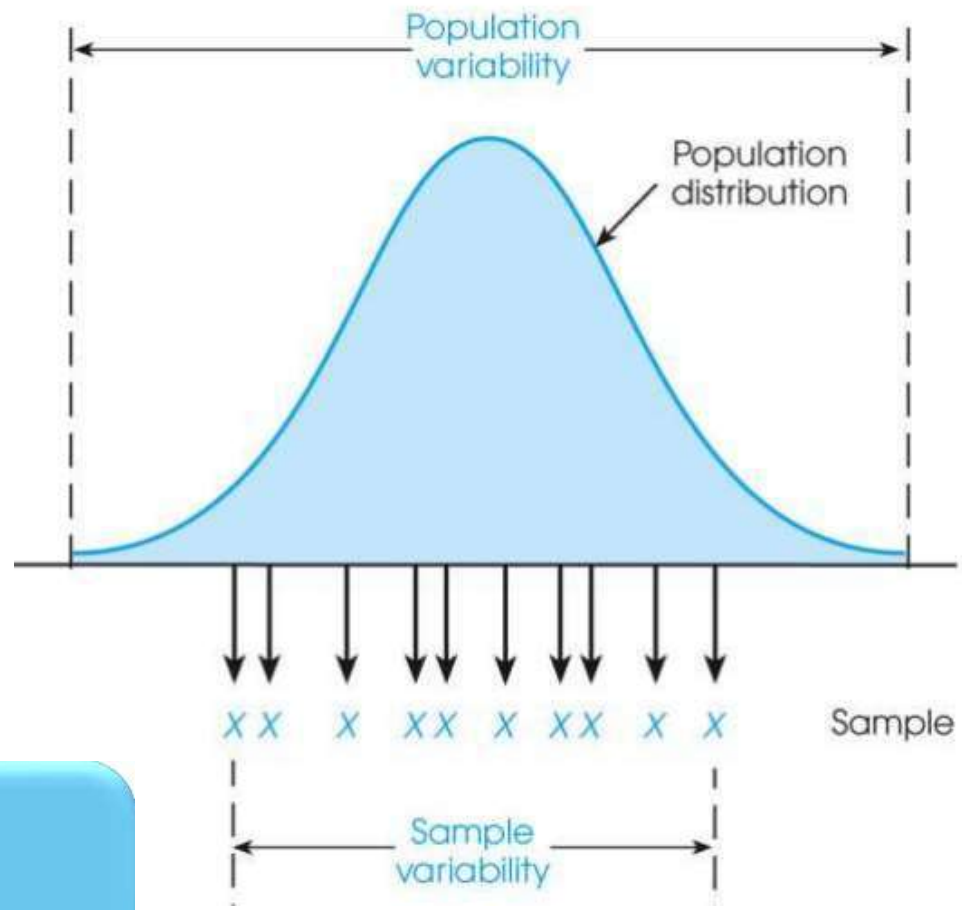
- Goal of inferential statistics:
  - Draw general conclusions about population
  -
  - Based on limited information from a sample

- 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



Population of Adult Heights

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

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

Degrees of Freedom

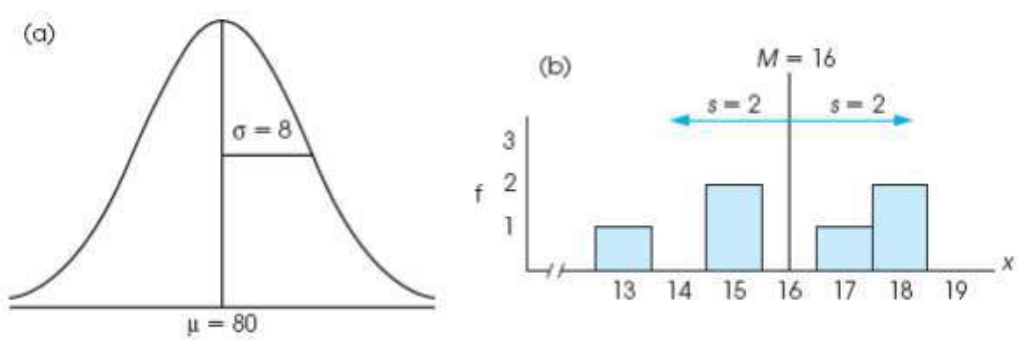
- Population variance
  - Mean is known
  - Deviations are computed from a known mean

- Sample variance as estimate of population
  - Population mean is unknown
  - Using sample mean restricts variability

- Degrees of freedom
  - Number of scores in sample that are independent and free to vary
  - Degrees of freedom (df) = n – 1

Descriptive Statistics

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



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

Interquartile Range = IQR = $Q_3 - Q_1$

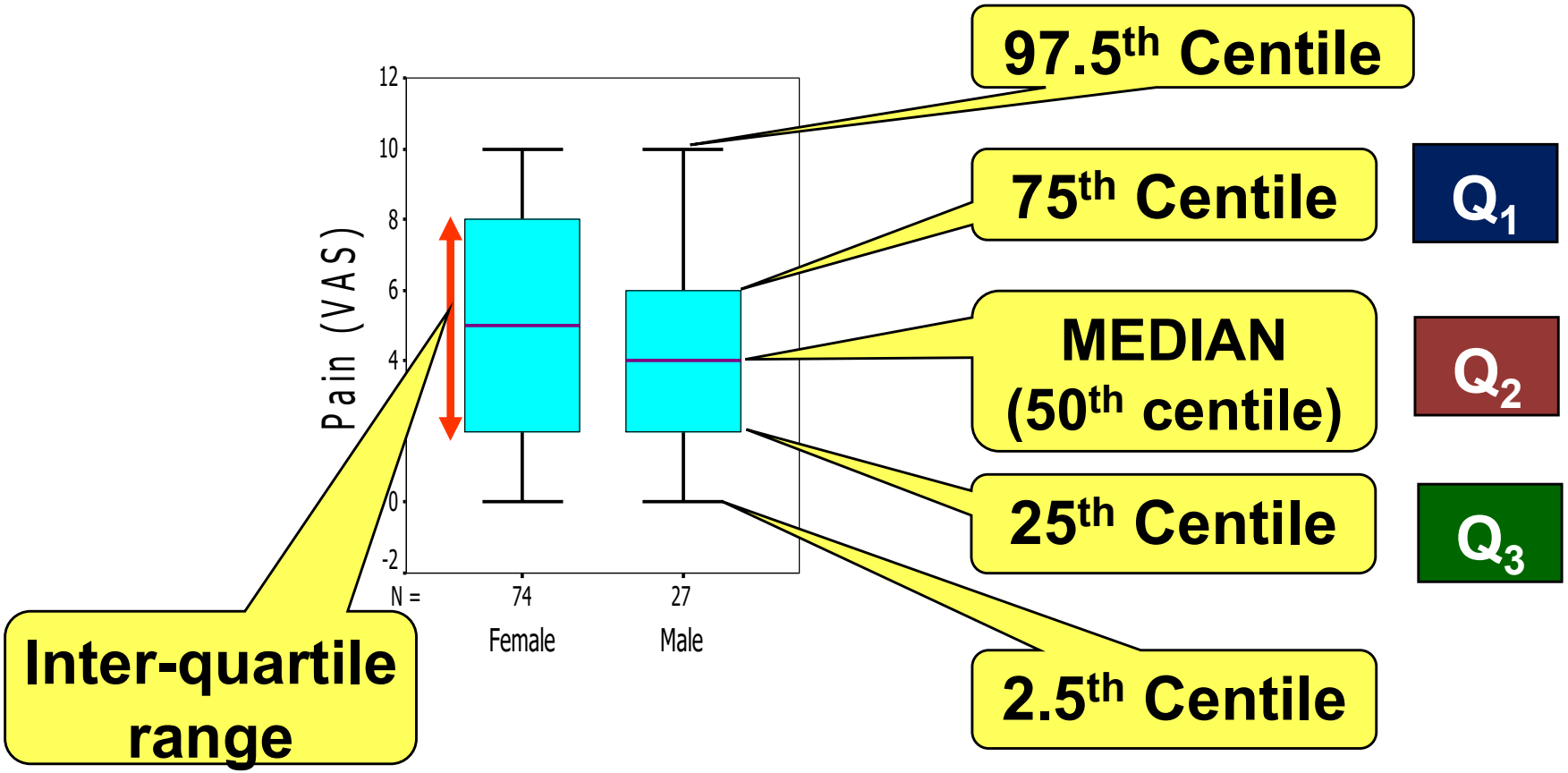
Data in Ordered Array: 11 12 13 16 16 17 17 18 21

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

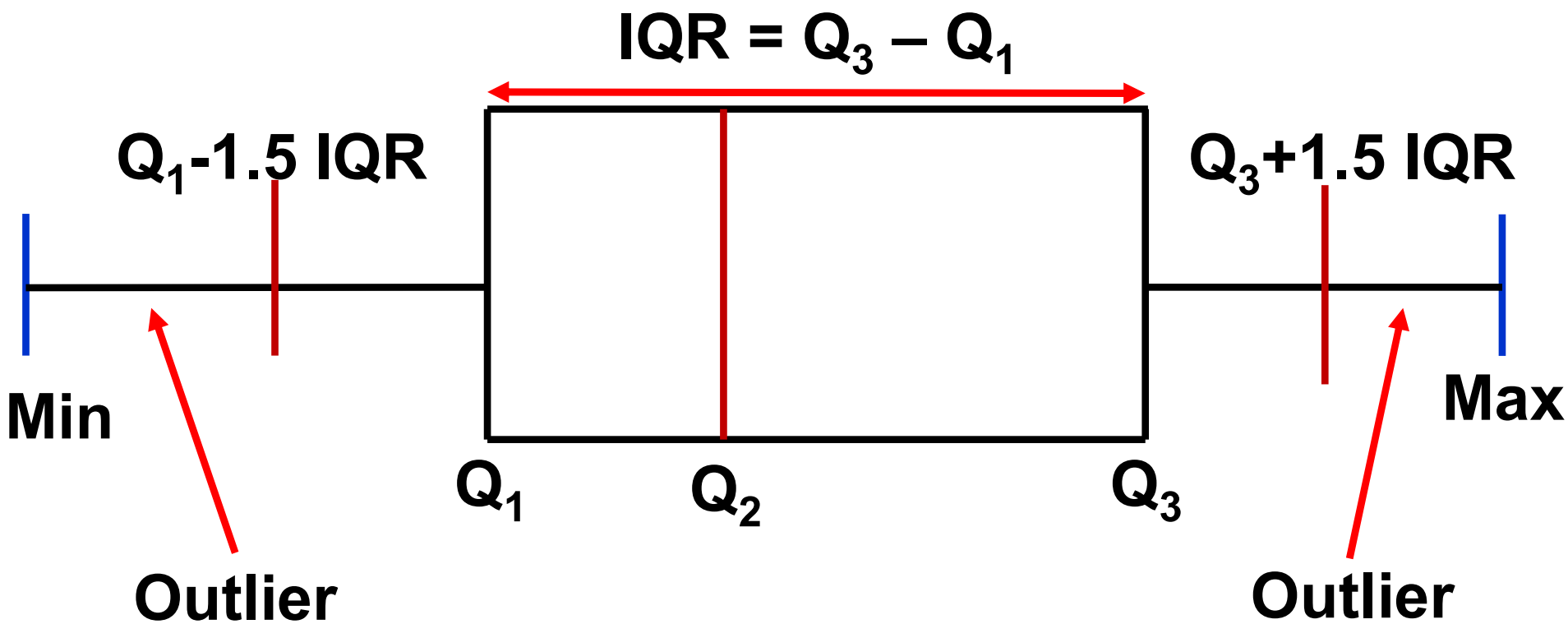
Position of $Q_3 = \frac{3 \cdot (9 + 1)}{4} = 7.50$, $Q_3 = 17.5$

Interquartile Range = IQR= $Q_3 - Q_1 = 17.5 - 12.5 = 5$

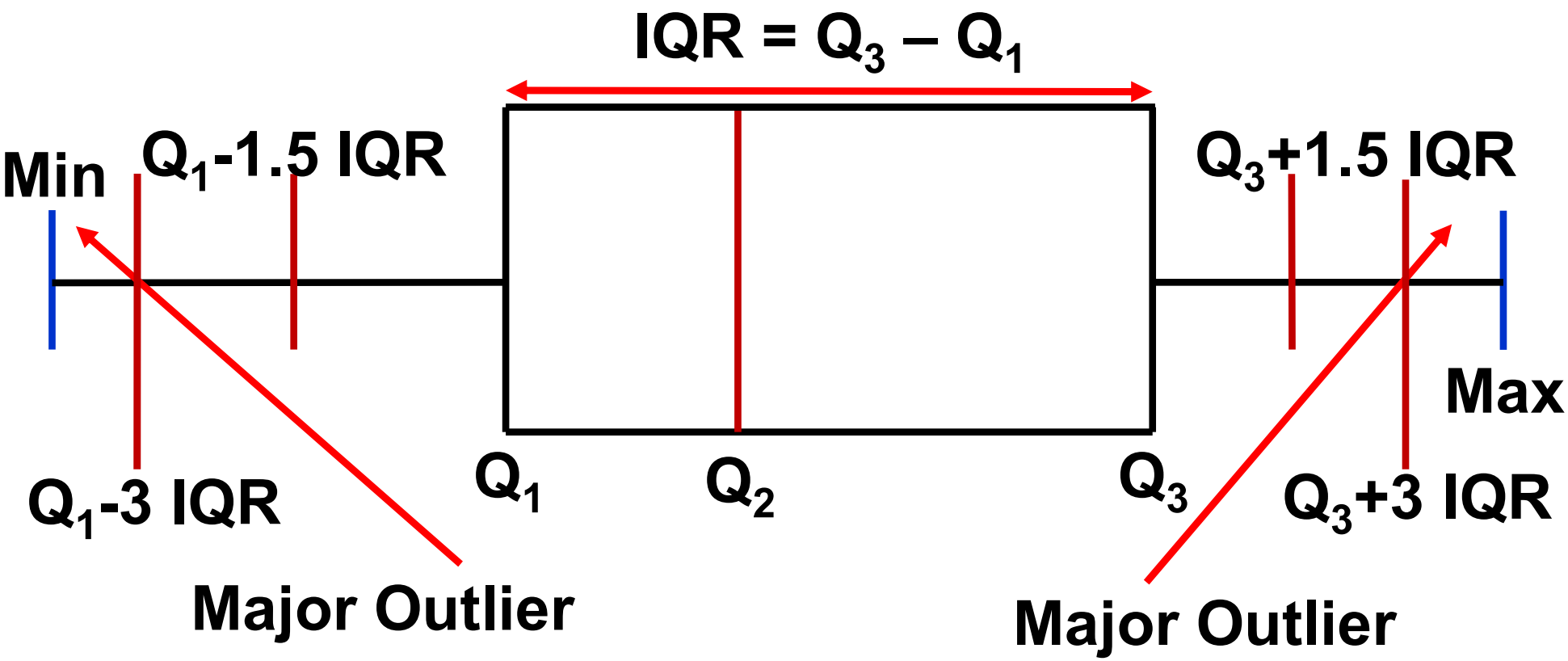
Box-and-Whisker plot



Box-and-Whisker plot



Box-and-Whisker plot



Potential outliers

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

  Lower limit = Q_1 - 1.5 x IQR

  Upper limit = Q_3 + 1.5 x 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.

the experiment.

Sample spaces and events

- 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

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

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

Introduction to Statistical Methods

ISM Team

IMP Note to Self

innovate achieve lead

BITS Pilani

Session No 2

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

(Session 2: 27th/28th MAY 2023)

Contact Session 2

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

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 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, 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.

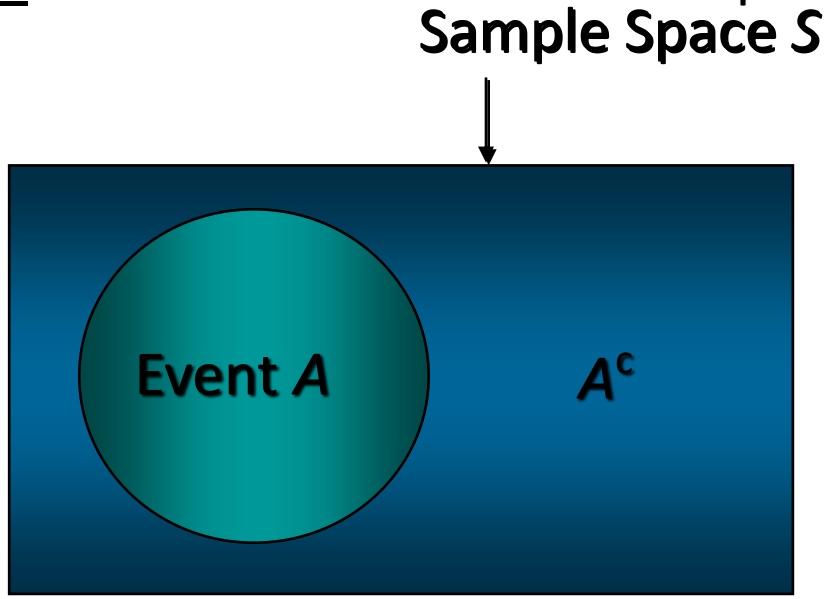
Event

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*.

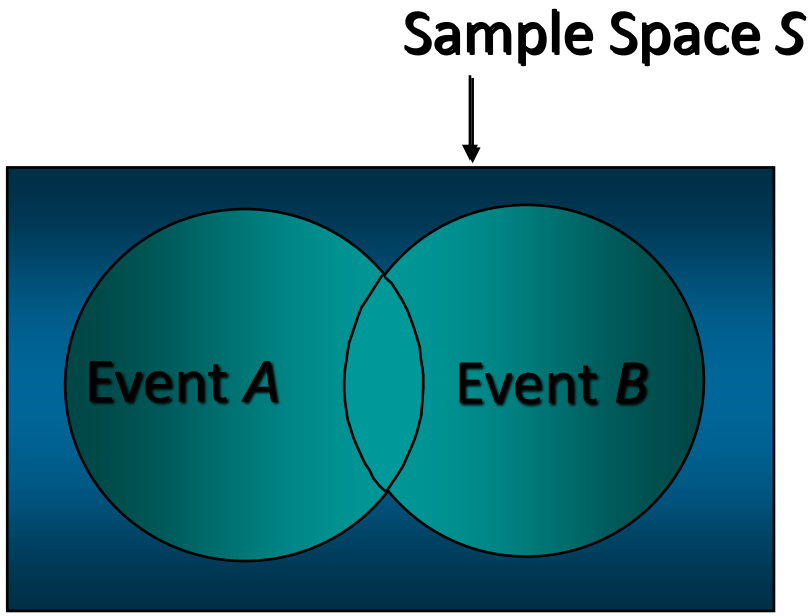
Complement of an Event

- The <u>complement</u> 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 <u>Venn diagram</u> below illustrates the concept of a complement.



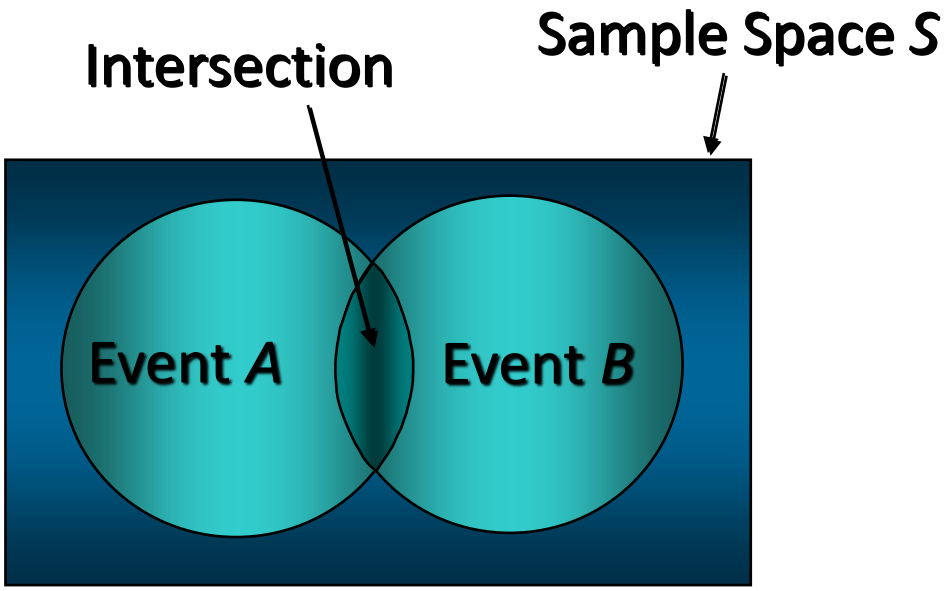
Union of Two Events

- The <u>union</u> 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* U *B*
- The union of *A* and *B* is illustrated below.



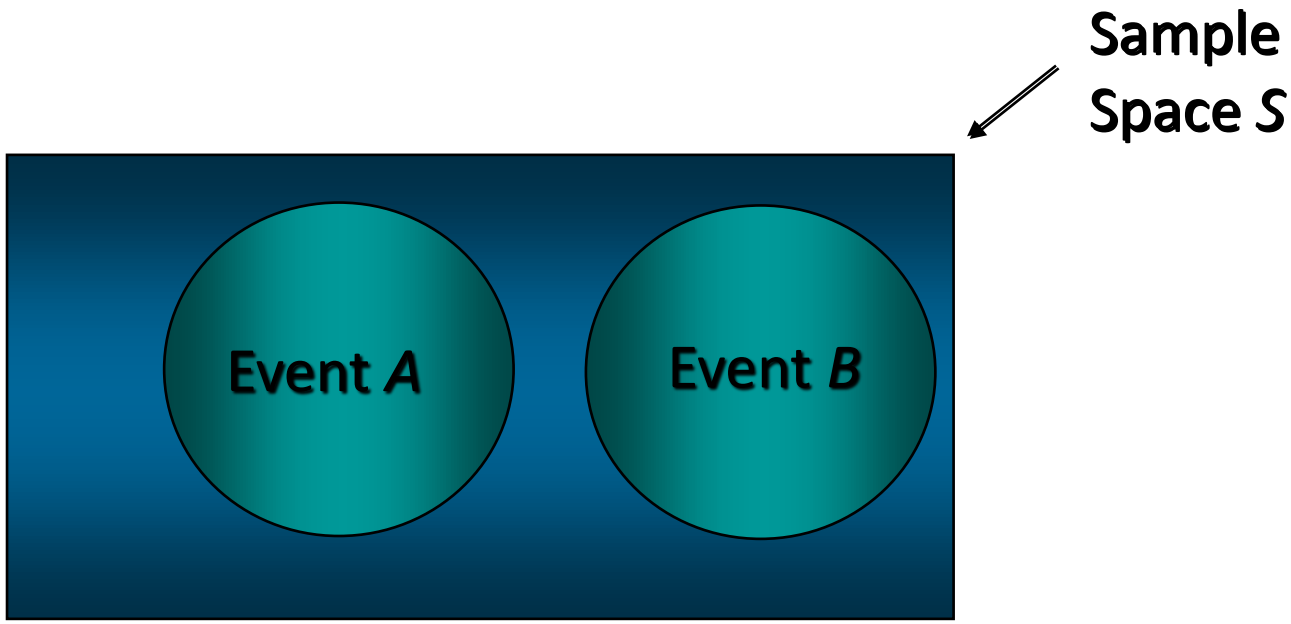
Intersection of Two Events

- The <u>intersection</u> 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 <u>mutually exclusive</u> 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 $E_1 \cap E_2 = \varnothing$

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

Probability as a Numerical Measure of the Likelihood of Occurrence

Increasing Likelihood of Occurrence

Probability: 0 .5 1

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

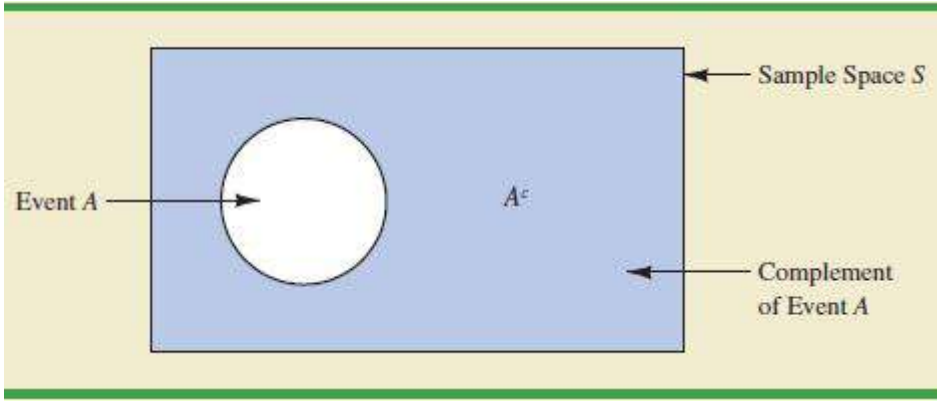
  $$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

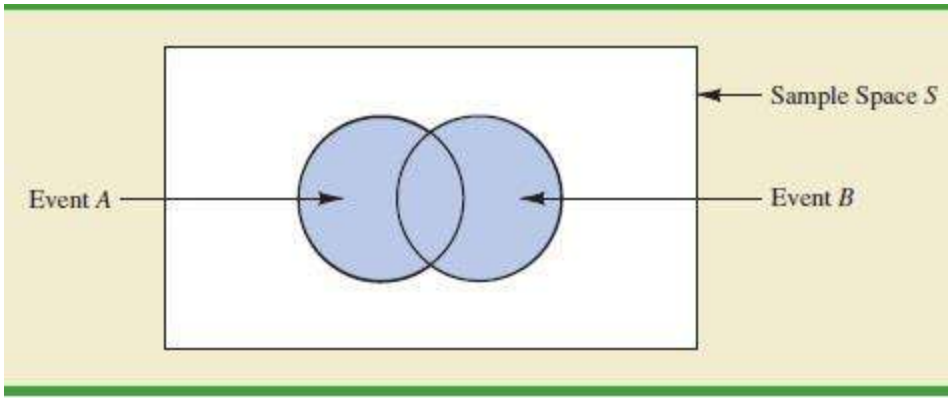
- If events *A* and *B* are **mutually exclusive**, then the rule can be simplified to

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

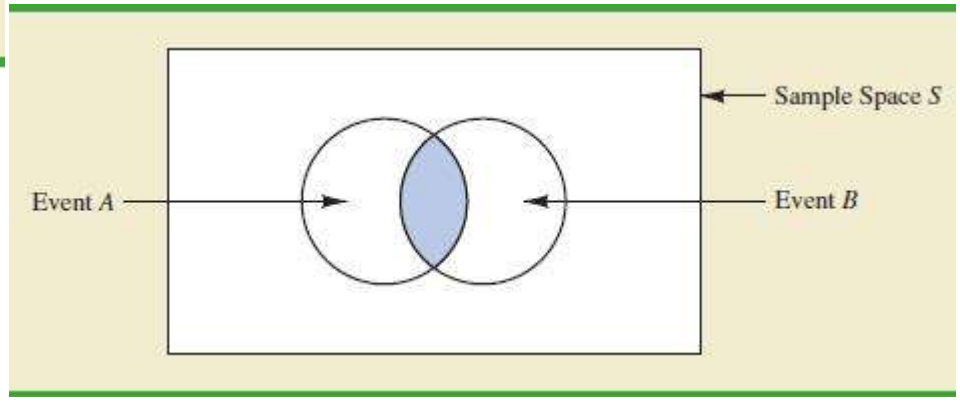
Probability and Venn Diagram

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





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



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: 1

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

Solution:

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$P(\text{neither } 7 \text{ nor } 11)$

$= 1 - [P(7) + P(11)]$

$= 1 - \left[\frac{6}{36} + \frac{2}{36}\right]$

$= 28/36 = 7/9$

Example:2

The probability that a student passes in statistics examination is 2/3 and the probability that he /she will not pass in mathematics examination is 5/9. The probability that he/she will pass in at least one of the examination is 4/5. Find the probability that he /she will pass in both the examinations

Solution:

The probability of passing in statistics P(S)=2/3

The probability of passing in Mathematics P(M)=1-5/9 =4/9

Probability of Passing at least one of these examination

= P(S∪M) = 4/5

$$\boldsymbol{P(S \cup M) = P(S) + P(M) - P(S \cap M).}$$

$$\frac{4}{5} = \frac{2}{3} + \frac{4}{9} - P(S \cap M).$$

$$P(S \cap M) = \frac{2}{3} + \frac{4}{9} - \frac{4}{5} = \frac{14}{45}.$$

The probability of passing both examinations $P(S \cap M) = \frac{14}{45}$

Example: 3

Suppose that 75% of all investors invest in traditional annuities and 45% of them invest in the stock market. If 85% invest in the stock market and/or traditional annuities, what percentage invest in both?

Solution:

Let A be the event that a randomly selected investor invests in traditional annuities.

Let B be the event that he or she invests in the stock market.

Then $P(A) = 0.75$, $P(B) = 0.45$, and $P(A \cup B) = 0.85$

Since,

$$P(AB) = P(A) + P(B) - P(A \cup B)$$
$$= 0.75 + 0.45 - 0.85$$
$$= 0.35.$$

Example: 4

Suppose the manufacturer's specifications for the length of a certain type of computer cable are 2000 ± 10 millimeters. In this industry, it is known that small cable is just as likely to be defective (not meeting specifications) as large cable. That is, the probability of randomly producing a cable with length exceeding 2010 millimeters is equal to the probability of producing a cable with length smaller than 1990 millimeters. The probability that the production procedure meets specifications is known to be 0.99.

(a) What is the probability that a cable selected randomly is too large?

(b) What is the probability that a randomly selected cable is larger than 1990 millimeters?

Solution:

***Solution*:** Let M be the event that a cable meets specifications. Let S and L be the events that the cable is too small and too large, respectively. Then

(a) $P(M) = 0.99$ and $P(S) = P(L) = (1 - 0.99)/2 = 0.005$.

(b) Denoting by X the length of a randomly selected cable, we have

$$P(1990 \leq X \leq 2010) = P(M) = 0.99.$$

Since $P(X \geq 2010) = P(L) = 0.005$,

$$P(X \geq 1990) = P(M) + P(L) = 0.995.$$

This also can be solved by using Theorem 2.9:

$$P(X \geq 1990) + P(X < 1990) = 1.$$

Thus, $P(X \geq 1990) = 1 - P(S) = 1 - 0.005 = 0.995$.

Exercise problems:

Q 1: 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?

Exercise problems

Q 2. 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.

Q 3. 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?

Exercise problems

Q 4: 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.

a) One motor will operate satisfactorily and other will not.

Exercise problems

Q 5. 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?

Exercise problems

Q 6. 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?.

Q7: If A and B are events with $P(A \cup B) = 7/8$, $P(A \cap B) = ¼$ and $P(A') = 5/8$, find $P(A)$, $P(B)$ and $P(A \cap B')$**).**

Thanks

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Statistical Methods

ISM Team

innovate achieve lead

BITS Pilani
Pilani Campus

Session 3: Conditional Probability, Independent events and, Total Probability

(Session 3: 3rd/4th June 2023)

IMP Note to Self

RECAPTIULATION:

- Introduction to Probability
- Review of Set Theory
- Counting Principles
- Definition of Probability
- Axioms of Probability
- Addition Rule of Probability

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,indepents events, Total probability	T1 & T2
HW	Problems on conditional probability	T1 & T2
Lab		

CONTENTS:

- Conditional Probability
- Independent Events
- Total Probability

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

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

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

It can also written as $P(A \cap B) = P(B)\, P(A/B) \qquad P(B) \neq 0$

$= P(A)\, P(B/A) \qquad P(A) \neq 0$

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

CONDITIONAL PROBABILITY

Multiplication Rule

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

$$P(A_1 \cap A_2 \cap \ldots \cap A_n) = P(A_1)\, P(A_2/A_1) P(A_3/A_1 \cap A_2) \ldots\ldots P(A_n/A_1 \cap A_2 \cap \ldots\ldots \cap A_{n-1})$$

Examples:1

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?

$$P(M) = 0.25, \quad P(C) = 0.15, \quad P(M \cap C) = 0.1$$

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

b. If the 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?

Examples:2

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

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

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

Examples: 3

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 $\boldsymbol{P}\left(\frac{\boldsymbol{A}}{\overline{\boldsymbol{D}}}\right)$

Ans:

Let D and A be the events that the flight departs and arrives on time respectively. Then, $\boldsymbol{P}\ (\boldsymbol{D}) = \mathbf{0.83}, \boldsymbol{P}(\boldsymbol{A}) = \mathbf{0.82}$ and $\boldsymbol{P}\ (\boldsymbol{D} \cap \boldsymbol{A}) = \mathbf{0.78}$

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

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

(ii) Probability that the plane departed on time given that it has arrived on time is $\boldsymbol{P}\left(\frac{\boldsymbol{D}}{\boldsymbol{A}}\right) = \frac{\boldsymbol{P}(\boldsymbol{D} \cap \boldsymbol{A})}{\boldsymbol{P}(\boldsymbol{A})} = \frac{\mathbf{0.78}}{\mathbf{0.82}} = \mathbf{0.9512}$

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

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

Example: 4

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

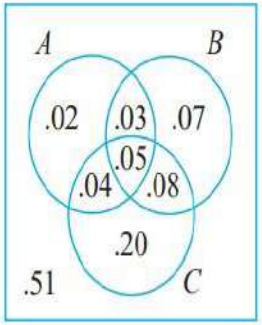
$$P\left(\frac{B}{A}\right) = \frac{P(B \cap A)}{P(A)} = \frac{0.60}{0.70} = 0.8571$$

BITS Pilani, Pilani Campus

Example: 5

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

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



Find

A). P(A/B)

B). P(A/BUC)

C). P(A/ reads at least one)

Solution

We thus have

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

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

$$P(A|\text{reads at least one}) = P(A|A \cup B \cup C) = \frac{P(A \cap (A \cup B \cup C))}{P(A \cup B \cup C)}$$

$$= \frac{P(A)}{P(A \cup B \cup C)} = \frac{.14}{.49} = .286$$

and

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

INDEPENDENT EVENTS

We can deduce an important result from the conditional probability:

If B has no effect on A, then, $\boldsymbol{P\left(\frac{A}{B}\right) = P(A)}$ Also $\boldsymbol{P\left(\frac{B}{A}\right) = P(B)}$

and we say the events are independent.

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

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

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

or $$\boxed{P(A \cap B) = P(A) \times P(B)}$$

Example: 1

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(AUBUC) = P(A) + P(B) + P(C) - P(A∩B) - P(B∩C) - P(A∩C) + P(A ∩ B ∩ C)

Using complement of the event i.e the problem is not solved . Therefore P(problem solved is)=1-P(not solved)

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

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

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

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

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

Example: 2

A box contains 20 fuses of which 5 are defective. If two fuses are choosen at random one after the other. What is probability that both the fuses are defective if

(i) the first fuse is replaced,

(ii) (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

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

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

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

Miscellaneous examples

1. A box A contains 2 white balls and 4 black balls. Another box B contains 5 white balls and 7 black balls. A ball is transferred from box A to box B, than a ball is drawn from box B. What is the probability that it is white.

2. A person X can hit a target three times in five shots, a second person Y twice in five shots and a third person Z thrice in four shots. They all fire once simultaneously. Find the probability that (i) atleast two shots hit, (ii) exactly two shots hit.

Continued

3. If A and B are two events with P(A)= ½, P(B)= 1/3 and $P(A \cap B) = 1/4$
Find

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

4.Three students X,Y,Z write an examination. Their chances of passing are $\frac{1}{2}$, $\frac{1}{3}$ and $\frac{1}{4}$ respectively. Find the probability that

(i) all of them pass,

(ii) at least one of them passes,

(iii) at least two of them pass.

Continued

5. The probability that Ram will hit a target is 0.56 and the probability that shyam will hit the target is 0.45. The probability that both will hit the target is 0.25. find the probability that

(i) at least one will hit the target

(ii) exactly one will hit the target.

6. If A and B are two events such that

$$P(A \cap B) = 0.05, P(B) = 0.7, P(A \cap \overline{B}) = 0.02$$

Find (i) $P(A \cup B)$, (ii) $P(A \cup \overline{B})$

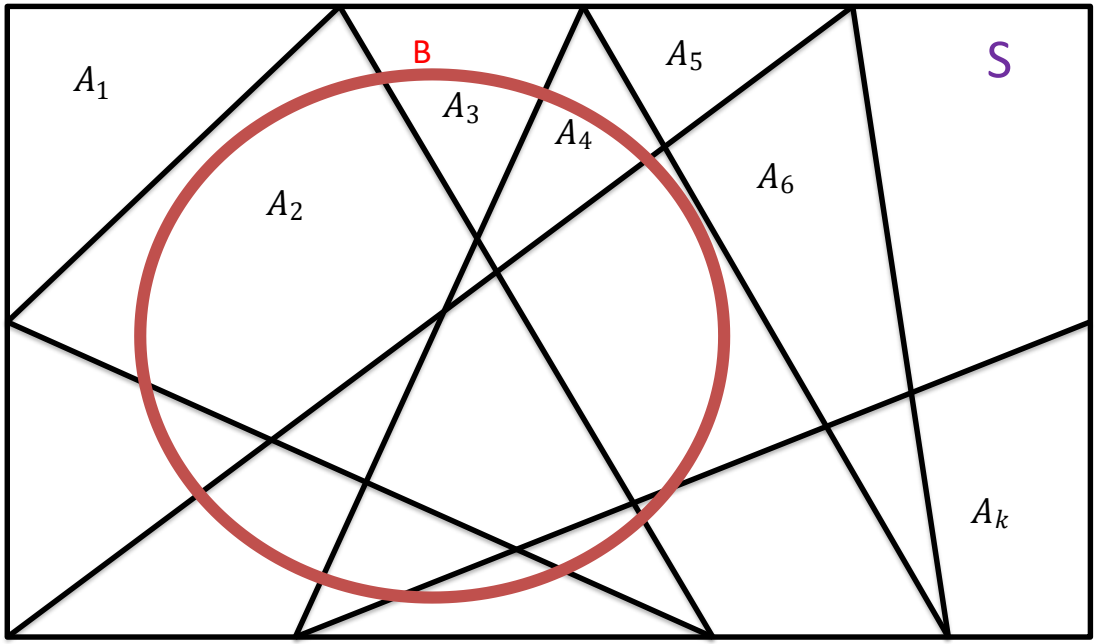
Total Probability

The Law of Total Probability

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

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

$S = A_1 \cup A_2 \cup \ldots \cup A_n$ and $A_1 \cap A_2 \cap \ldots \cap A_n = \emptyset$



$$\therefore \boldsymbol{B} = \boldsymbol{B} \cap \boldsymbol{S} = \boldsymbol{B} \cap \{\boldsymbol{A}_1 \cup \boldsymbol{A}_2 \cup \boldsymbol{A}_3, \ . \quad . \quad . \cup \boldsymbol{A}_n\}$$

Proof:

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

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

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,

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

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

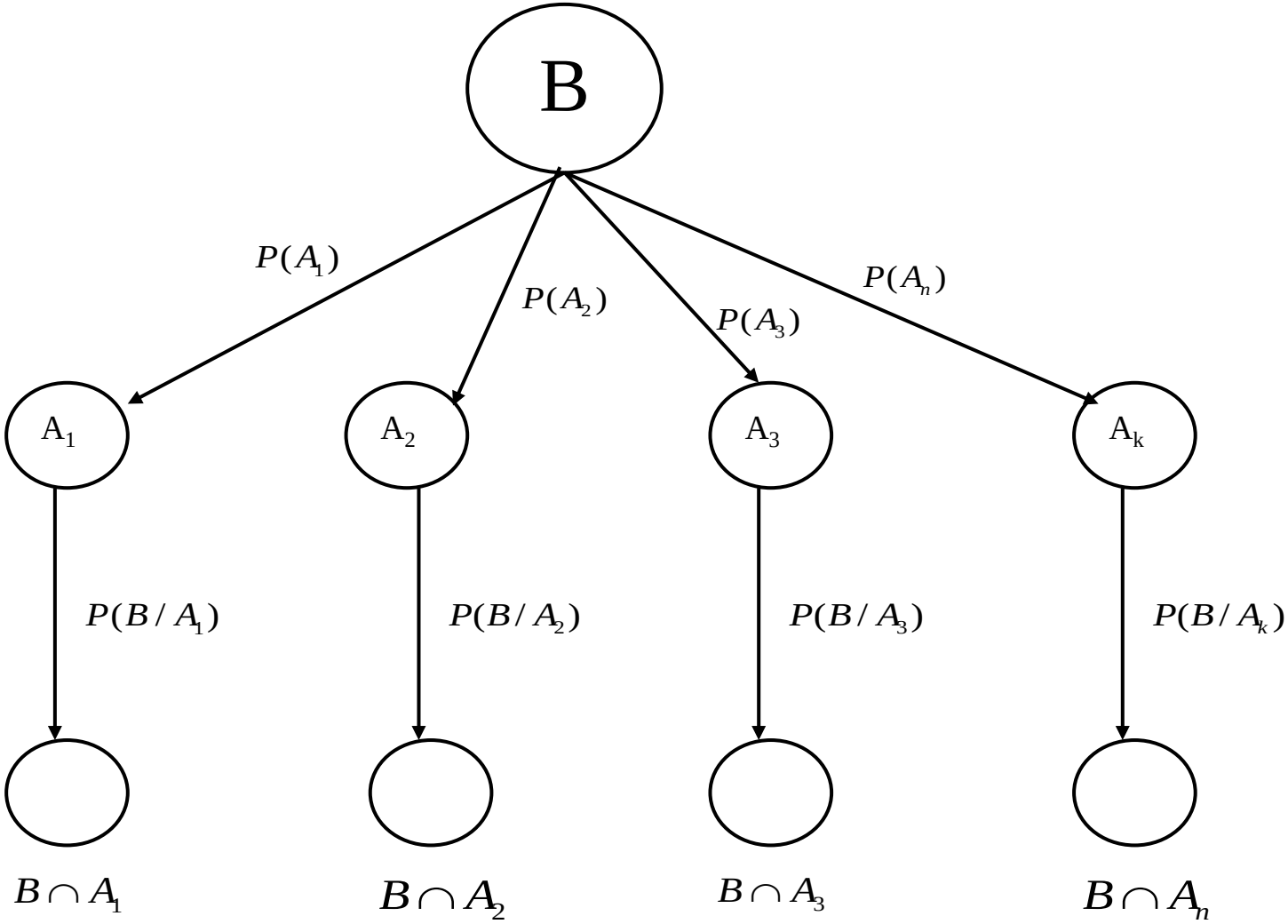
Using multiplication rule on each term on R.H.S, namely

$$P(B \cap A_i) = P(A_i) \cdot P(B \mid A_i) \qquad \text{(A)}$$

$$P(B) = \sum_{i=1}^{i=n} P(A_i) P(B \mid A_i)$$

(B)-Total theorem on Probability

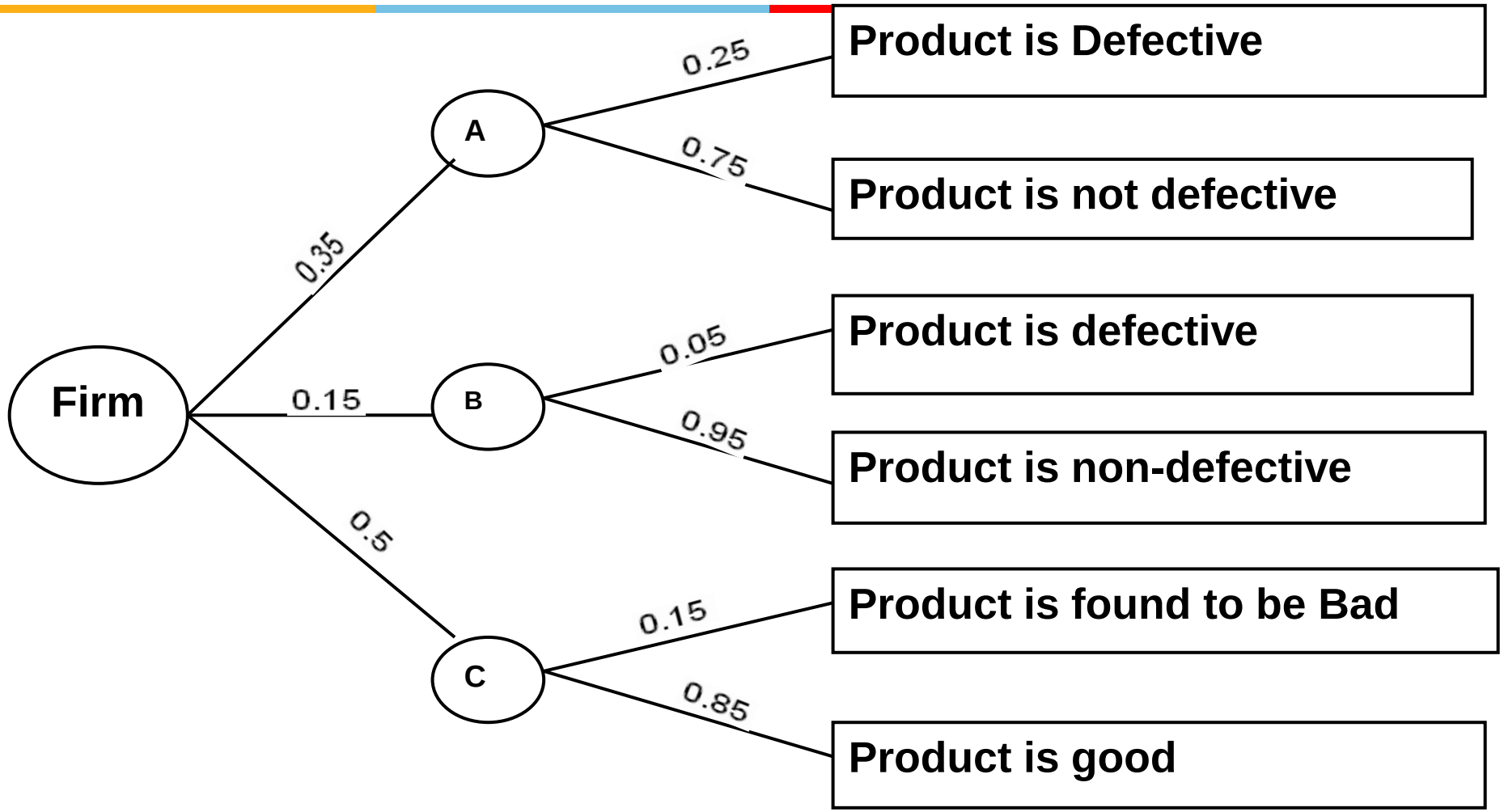
The Theorem of Total Probability
(tree diagram)



Examples :1

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".

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

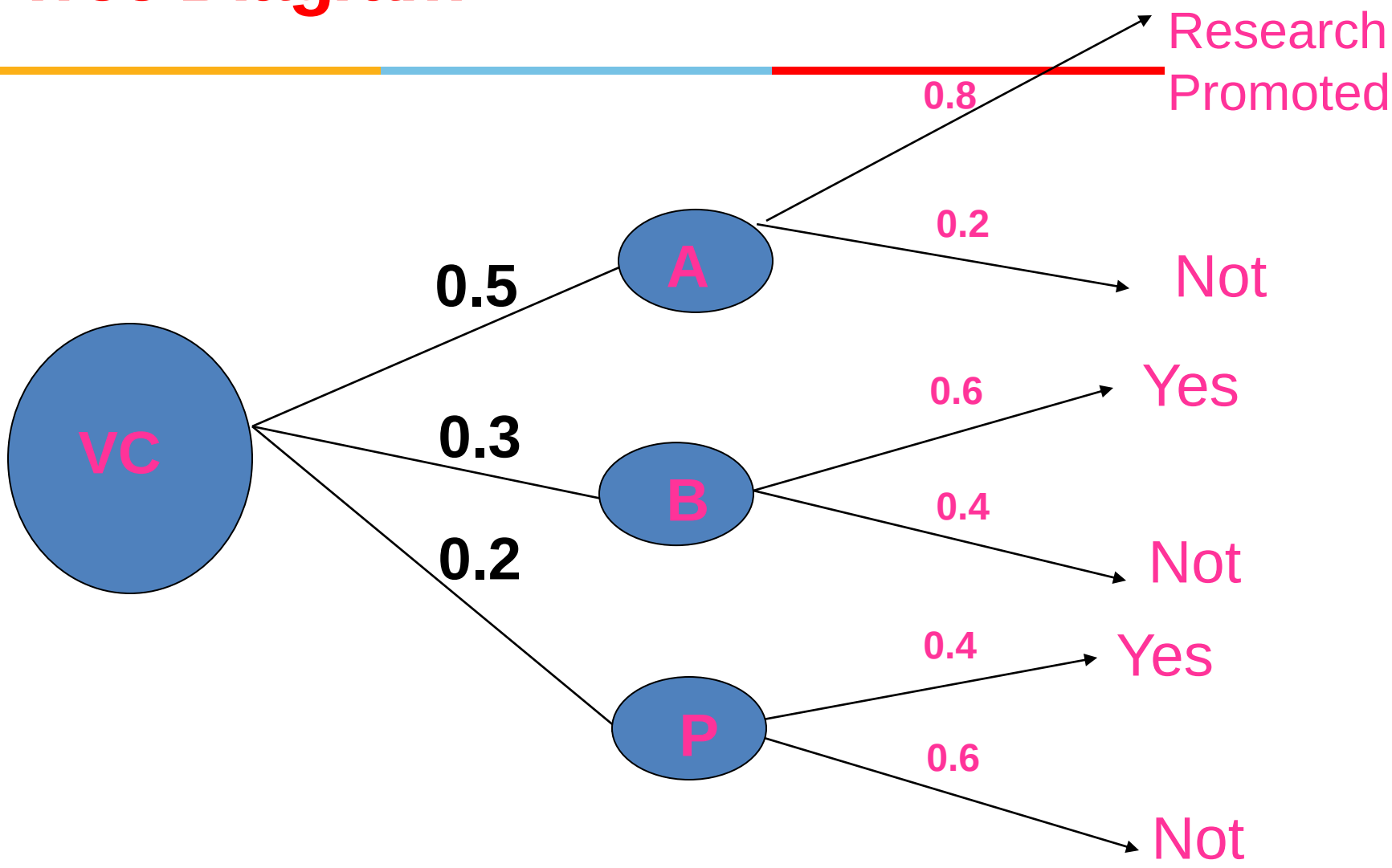
Therefore, the chance that product is manufactured by the plant C is

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

Example 2

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



Solution

Let **X: " Research work is promoted"**

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

Now to find P[" An Academician is VC" / "Research work is promoted i.e. event X"] $= \frac{0.5 \times 0.8}{0.66} = 0.6061$

Practice problems

Example 1.

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 2.

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?

Continued

Example 3.

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?

Example: 4

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

Example: 5

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?

IMP Note to Self

Thank you

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Statistical Methods

ISM Team

innovate achieve lead

BITS Pilani

Pilani Campus

Session 4: Bayes theorem and Naïve Bayes theorem

(Session 4: 10th/11th June 2023)

IMP Note to Self

RECAPTIULATION:

- Introduction to Probability
- Review of Set Theory
- Counting Principles
- Definition of Probability,
- Axioms of Probability,
- Addition Rule of Probability
- Conditional Probability,
- Independent events and,
- Total Probability

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 Scientsists, 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

BAYES THEOREM:

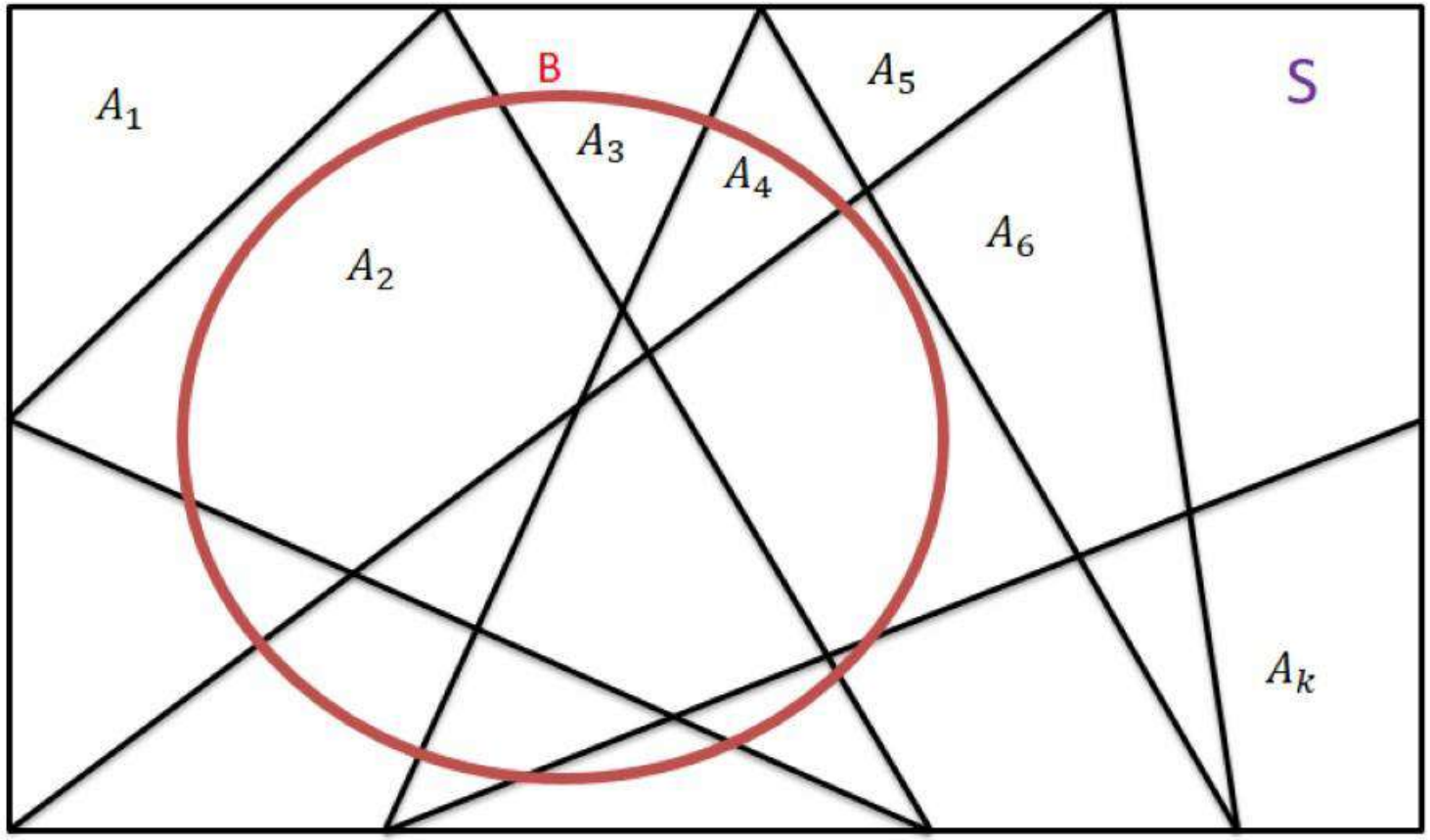
Let $P = \{A_1,\ A_2,\ A_3,\ \ldots\ , A_k\}$ be a set of exhaustive and mutually exclusive events of a sample space S with $P(A_i) \neq 0$ For each i. If B is any other event associated with A_i with $P(B) \neq 0$, then

$$P\left(\frac{A_i}{B}\right) = \frac{P(A_i)P\left(\frac{B}{A_i}\right)}{P(B)}$$

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

Geometrical Representation of Bayes theorem

$S = A_1 \cup A_2 \cup \ldots \cup A_n$ and $A_1 \cap A_2 \cap \ldots \cap A_n = \emptyset$



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

Proof:

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

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

But we know that

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

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

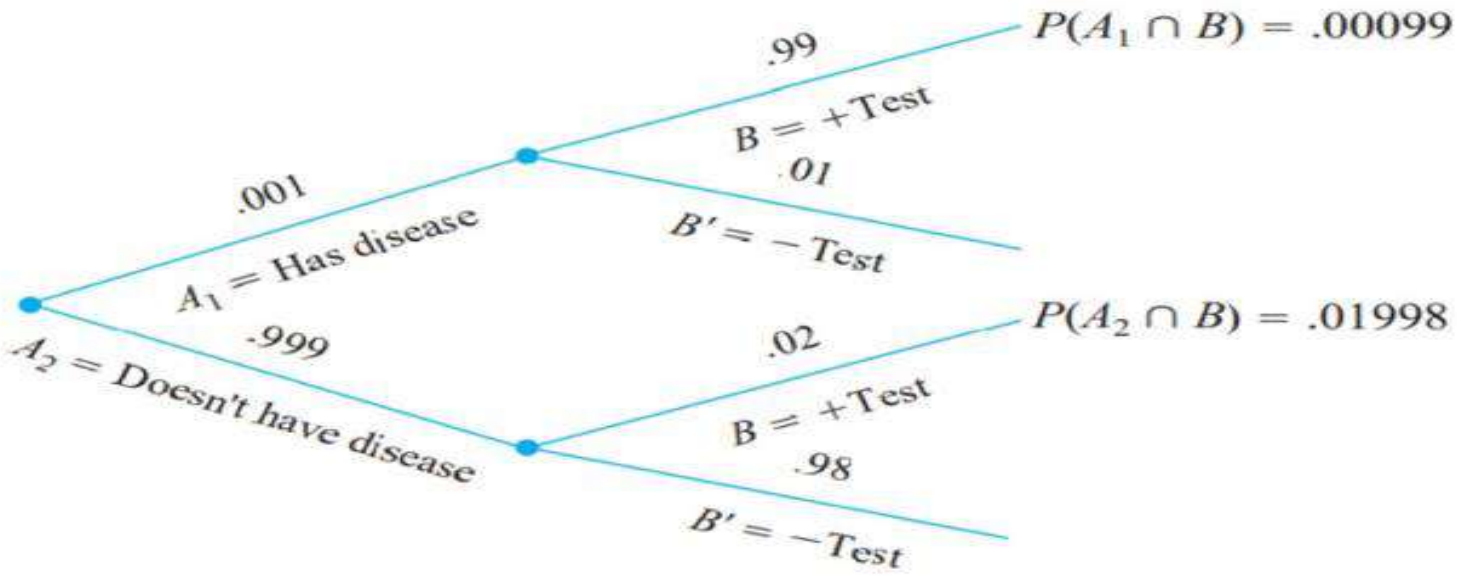
Hence the theorem proved

Example: 1

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?

Solution:

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



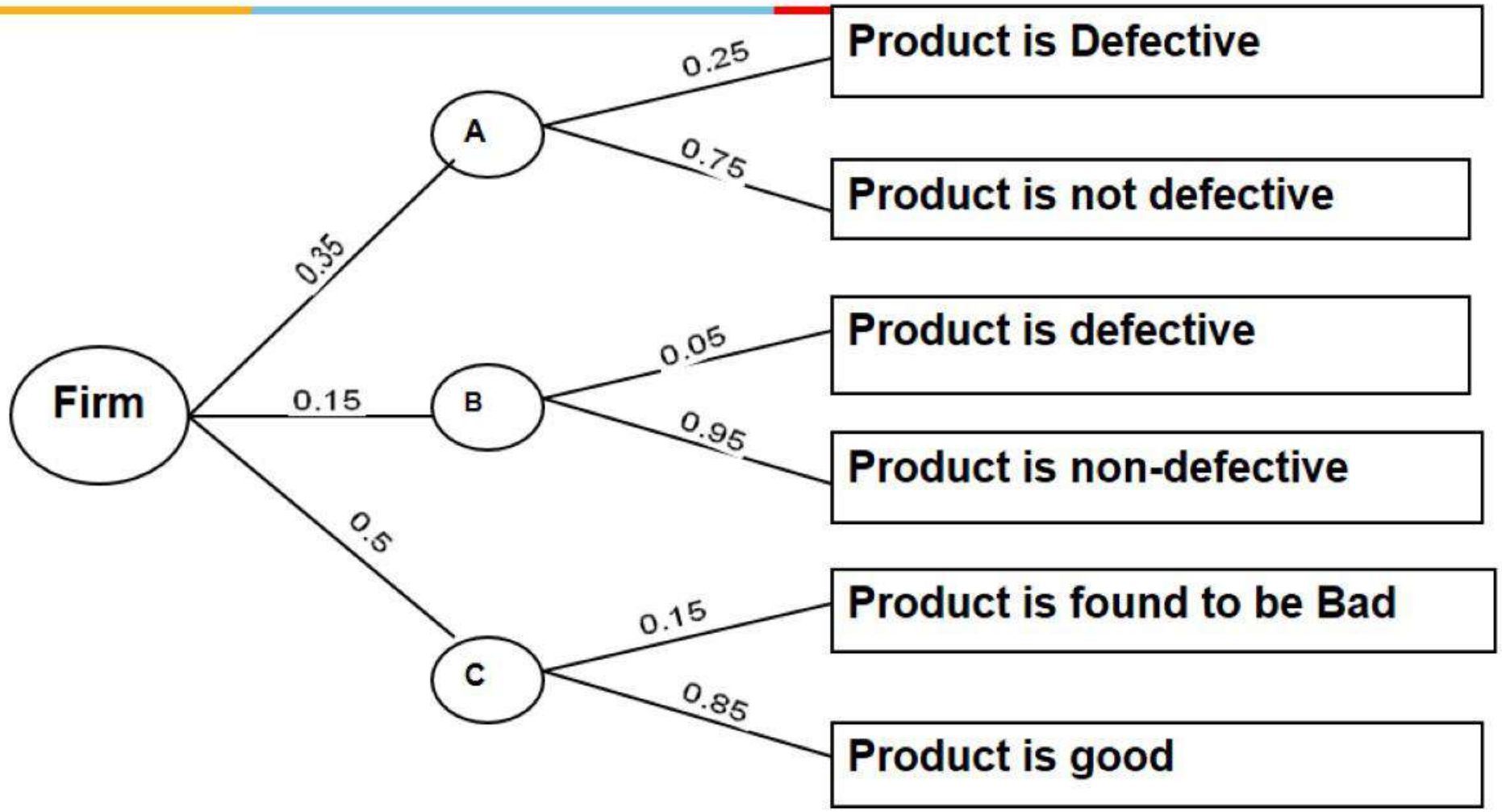
Next to each branch corresponding to a positive test result, the multiplication rule yields the recorded probabilities. Therefore, $P(B) = .00099 + .01998 = .02097$, from which we have

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

Example: 2

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".

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

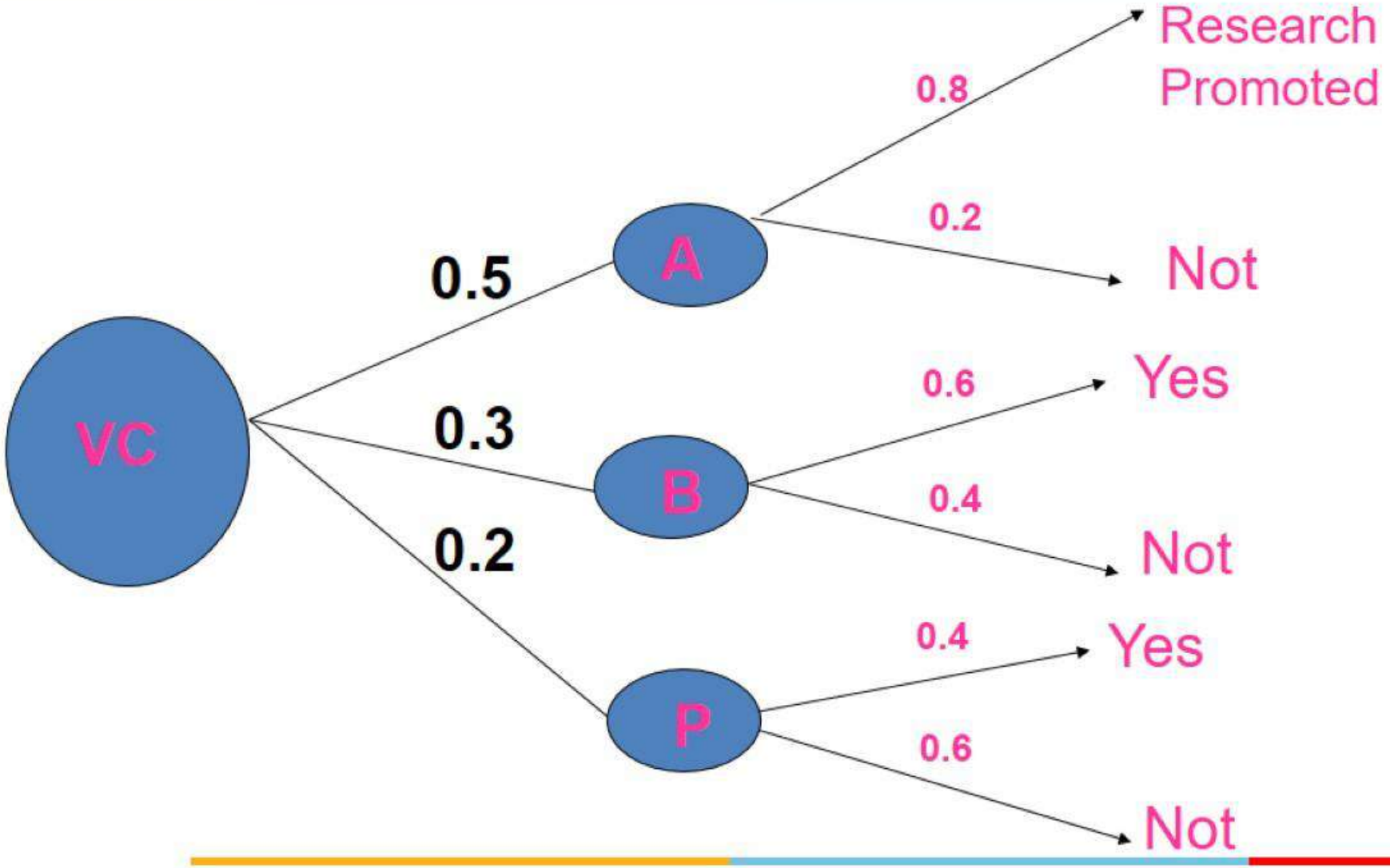
Therefore, the chance that product is manufactured by the plant C is

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

Example: 2

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



Let **X: " Research work is promoted"**

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

Now to find P[" An Academician is VC" / "Research work is promoted i.e. event X"] $= \frac{0.5 \times 0.8}{0.66} = 0.6061$

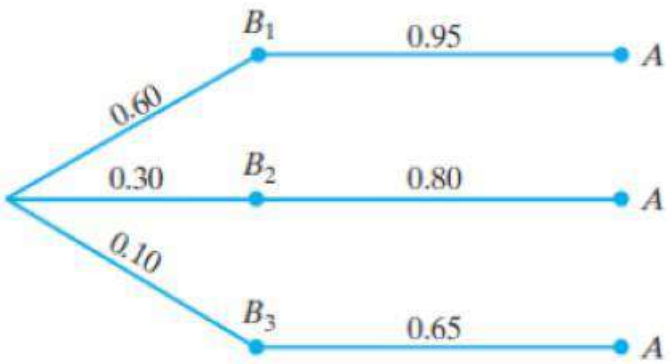
Example: A manufacturer of tablets receives its LED screens from three different suppliers, 60% from supplier $B1$, 30% from supplier $B2$, and 10% from supplier $B3$. In other words, the probabilities that any one LED screens received by the plant comes from these three suppliers are 0.60, 0.30, and 0.10. Also suppose that 95% of the LED screens from $B1$, 80% of those from $B2$, and 65% of those from $B3$ 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 $B3$.

If A denotes the event that a LED screen received by the plant performs according to specifications, and $B1$, $B2$, and $B3$ are the events that it comes from the respective suppliers

$$A = A \cap [B1 \cup B2 \cup B3]$$
$$= (A \cap B1) \cup (A \cap B2) \cup (A \cap B3)$$
$$= P(A \cap B1) + P(A \cap B2) + P(A \cap B3)$$
$$P(A) = P(B1) \cdot P(A \mid B1) + P(B2) \cdot P(A \mid B2) + P(B3) \cdot P(A \mid B3)$$
$$P(A) = (0.60)(0.95) + (0.30)(0.80) + (0.10)(0.65) = 0.875$$



$$P(B3 \mid A) = \frac{(0.10)(0.65)}{(0.60)(0.95) + (0.30)(0.80) + (0.10)(0.65)} = 0.074$$

Example: 3

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 ¾ that it will get the job provided firm W does not bid. The probability is ¾ that W will bid, and if it does, the probability that V will get the job is only1/3. (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?

Solution:

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

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

$$P(V) = P(V/W)\,P(W) + P(V/W^1)\,P(W^1) = 11/18$$

(b) $P(W^1/V) = P(V/W^1)\,P(W^1) / P(V) = 9/11$

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?

Example:

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

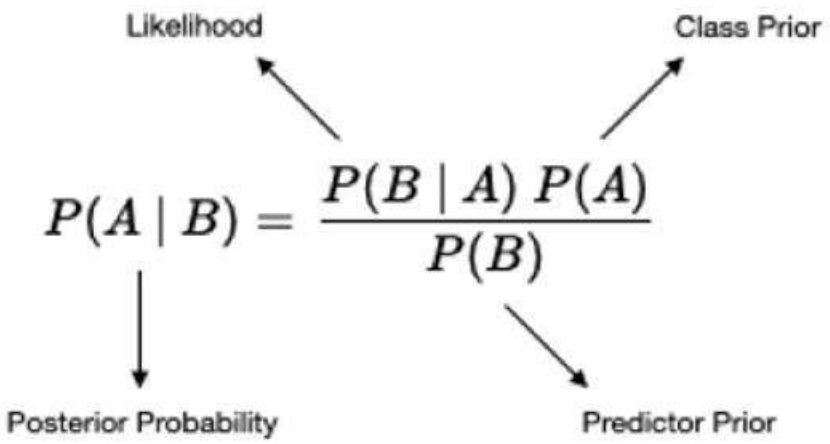
- $P(h)$ = prior probability of hypothesis h, before seeing the training data
- $P(D)$ = prior probability of training data D
- $P(h|D)$ = probability of h given D
- $P(D|h)$ = probability of D given h

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

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=ck,

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



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

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

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

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

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

$$h_{MAP} = argmax_{h \in H} P(h|D)$$

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

$$= argmax_{h \in H} P(D|h)P(h)$$

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(hi) = p(hj)

- 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} = 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:

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

Applying conditional independence

Naïve Bayes assumes X_i are conditionally independent given Y

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

$$P(X_1,X_2|Y) = P(X_1|X_2,Y)P(X_2|Y)$$
$$= P(X_1|Y)P(X_2|Y)$$

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

How many parameters to describe $P(X_1,\cdots,X_n|Y)$? $P(\mathrm{Y})$?

Without conditional independence assumption?

With conditional independence assumption?

Naïve Bayes Independence assumption

Assumption:

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

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

Naïve Bayes classifier

- **Bayes rule:**

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

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

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

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

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

Prior Probability

MLE

NAÏVE BAYES CLASSIFIER

- Assume independence among attributes X_i when class is given:
  - $P(X_1, X_2, \ldots, X_d \mid Y_j) = P(X_1 \mid Y_j)\, P(X_2 \mid Y_j) \ldots P(X_d \mid Y_j)$
  - Now we can estimate $P(X_i \mid Y_j)$ for all X_i and Y_j combinations from the training data
  - New point is classified to Y_j if $P(Y_j)\, \Pi\, P(X_i \mid Y_j)$ 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(Yes/Sunny) and P(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

	Play		
	no	yes	Row Total
weather			
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

sunny = 5/14 , rainy = 5/14

Overcast = 4/14

Play probabilities

no = 5/14

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

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

Weather	No	Yes	
Rainy	3	2	$5/14 = 0.36$
Sunny	2	3	$5/14 = 0.36$
overcast	0	4	$4/14 = 0.29$
Total	5	9	
	$5/14 = 0.36$	$9/14 = 0.64$	

Now $P(\text{Yes} \mid \text{sunny}) = \dfrac{P(\text{Sunny} / \text{Yes})\, P(\text{Yes})}{P(\text{sunny})}$

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

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

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

today = (Sunny, Hot, Normal, False)

$X \quad X_1 \quad X_2 \quad X_3 \quad X_4$

yes
no

$$P(yes \mid X) = \frac{P(X \mid yes)\, P(yes)}{P(X)}$$

$$\cancel{P(no \mid X)}$$

$$= \frac{P(X_1, X_2, X_3, X_4 \mid yes)\, yes}{P(X_1, X_2, X_3, X_4)} = \frac{P(X_1 \mid Y)\, P(X_2 \mid Y)\, P(X_3 \mid Y)\, P(X_4 \mid Y)}{P(X_1)\, P(X_2)\, P(X_3)\, P(X_4)}\, P(y)$$

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 = ?

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

h_i = 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)	$4/6$	$3/4$	$2/6$	$2/4$	$2/6$	$3/4$	$2/6$	$1/4$
NO	$2/6$	$1/4$	$4/6$	$2/4$	$4/6$	$1/4$	$4/6$	$3/4$

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

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

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

$$= \frac{0.0593}{P(E)}$$

YES

YES

NO

NO

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

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

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

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

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

$0.0593 > 0.0281$

ie 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

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

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

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

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

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

=> Mammals

Issues with Naïve Bayes Classifier

Naïve Bayes Classifier:

P(Refund = Yes | No) = 3/7
P(Refund = No | No) = 4/7
P(Refund = Yes | Yes) = 0
P(Refund = No | Yes) = 1
P(Marital Status = Single | No) = 2/7
P(Marital Status = Divorced | No) = 1/7
P(Marital Status = Married | No) = 4/7
P(Marital Status = Single | Yes) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
P(Marital Status = Married | Yes) = 0

For Taxable Income:
If class = No: sample mean = 110
sample variance = 2975
If class = Yes: sample mean = 90
sample variance = 25

- P(Yes) = 3/10
  P(No) = 7/10

- P(Yes | Married) = 0 x 3/10 / P(Married)
  P(No | Married) = 4/7 x 7/10 / P(Married)

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
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Naïve Bayes Classifier:

P(Refund = Yes | No) = 2/6
P(Refund = No | No) = 4/6
P(Refund = Yes | Yes) = 0
P(Refund = No | Yes) = 1
P(Marital Status = Single | No) = 2/6
P(Marital Status = Divorced | No) = 0
P(Marital Status = Married | No) = 4/6
P(Marital Status = Single | Yes) = 2/3
P(Marital Status = Divorced | Yes) = 1/3
P(Marital Status = Married | Yes) = 0/3
For Taxable Income:
If class = No: sample mean = 91
sample variance = 685
If class = No: sample mean = 90
sample variance = 25

Given X = (Refund = Yes, Divorced, 120K)

P(X | No) = 2/6 X 0 X 0.0083 = 0

P(X | Yes) = 0 X 1/3 X 1.2 X 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

Laplace Smoothing

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

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(sports| *A very close game*)

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

Using Bayes Theorem:

P(sports| *A very close game*)

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

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(close | sports) = 0, which makes product 0

A Simple Example

Draw | Naive Bayes

Text classification

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(sports| *A very close game*)

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

Using Bayes Theorem:

P(sports| *A very close game*)

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

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

P(A very close game) = P(A) P(very) P(close) P(game)

P(A very closed game | sports) = P(a | sports) P(very | sports) P(close | sports) P(game | sports)

"close" doesn't appear in sentences of sports tag, So P(close | sports) = 0, which makes product 0

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

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

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

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

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	_POSITIVE
2	I HATED THE MOVIE	_ NEGATIVE
3	A GREAT MOVIE ,GOOD MOVIE	_POSITIVE
4	POOR ACTING	_NEGATIVE
5	GREAT ACTING , A GOOD MOVIE	_POSITIVE
NEW	I HATED THE POOR ACTING	_????

$P(c / x)$

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

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

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

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

ie $P(c_1 \mid x) = \dfrac{P(x \mid c_1)\, P(c_1)}{\cancel{P(x)}}$

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

$\dfrac{P(I, +)}{P(+)}$ $\qquad$ $\dfrac{P(\text{hated}, +)}{P(+)}$

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

$$P(I \mid +) = \frac{1+1}{14+10}$$

$$P(I \mid -) = \frac{1+1}{6+10}$$

"I hated the poor acting"

word	positive	negative
I	$\frac{1+1}{14+10} = 0.0833$	$\frac{1+1}{6+10} = 0.125$
hated	$\frac{0+1}{14+10} = 0.0417$	$\frac{1+1}{6+10} = 0.125$
the	$\frac{1+1}{14+10} = 0.0833$	$\frac{1+1}{6+10} = 0.125$
poor	$\frac{0+1}{14+10} = 0.0417$	$\frac{1+1}{6+10} = 0.125$
acting	$\frac{1+1}{14+10} = 0.0833$	$\frac{1+1}{6+10} = 0.125$

x : I hate the Poor acting

$P(+/x)$

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

$\downarrow$

$3/5$

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

$P(-/x)$

$= (\quad)(\quad)(\quad)(\quad)(\quad) P(-1)$:

$\downarrow 2/5$

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

$\therefore$ 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 (Normal/ Dear, Friend) and P(Spam/ 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 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

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 word	normal	spam	Row Total
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

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

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

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 word	normal	spam	Row Total
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

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

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

Steps to Apply Bayes Theorem

Play word	normal	spam	Row Total
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

Step 4 - Apply probabilities from frequency table

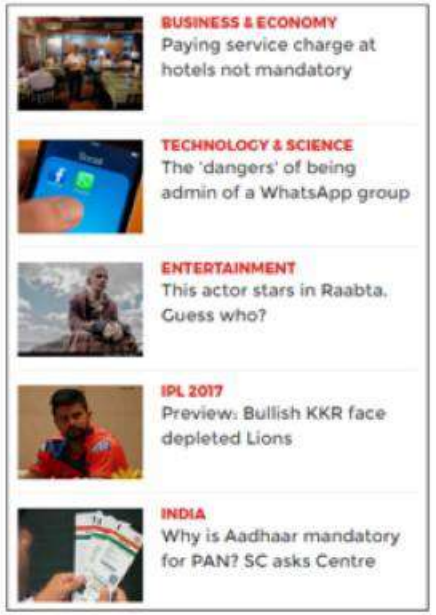
$$P(N).\ P(L/N).\ P(M/N)^4 = \left(\frac{21}{32}\right).\left(\frac{4}{21}\right).\left(\frac{2}{21}\right)^4 = 0.00001$$

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

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

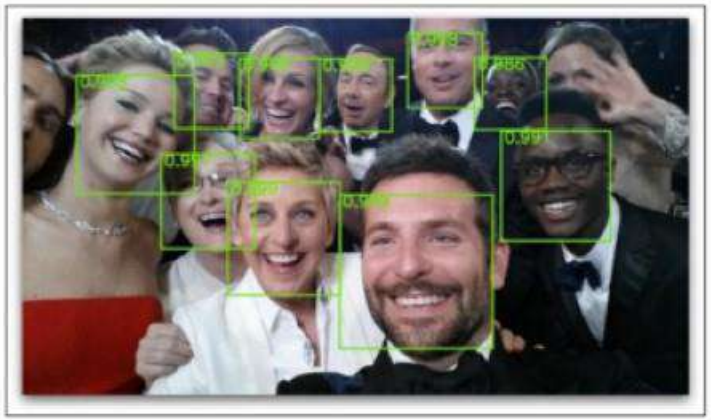
Naïve Bayes Classifier Applications

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

aardvark	0
about	2
all	2
Africa	1
apple	0
anxious	0
...	
gas	
...	
oil	
…	
Zaire	

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 Expectaion of a random variable,mean and variance of a random variable – Sinlge 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 <u>discrete random variable</u> may assume either finite or countably infinite number of values
- A <u>continuous random variable</u> 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:

$$f(x) \geq 0$$

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

- 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 <u>expected value</u>, or mean, of a random variable is a measure of its central location.
  - The mean or Expected value of a discrete random variable:

$$E(x) = \mu = \Sigma x f(x)$$

- The <u>variance</u> summarizes the variability in the values of a random variable.
  - Variance of a discrete random variable:

$$\mathrm{Var}(x) = \sigma^2 = \Sigma(x - \mu)^2 f(x)$$

- The <u>standard deviation</u>, 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, $V(aX) = a^2 . V(X)$
- 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

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) $\sum P(x) = 1$

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

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

probability distribution function

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

a) $E(X)$

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

c) standard deviation of X

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

Example

Let x be a random variable with PDF given by

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

a) Find constant 'c'.

b) Find $E(x)$ and $V(x)$

c) Find $P(x \geq 1/2)$

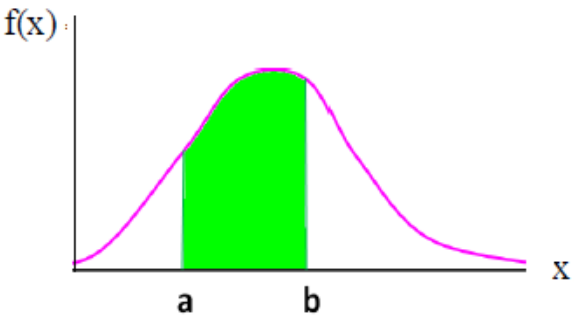
Continuous Probability Distributions

- A <u>continuous random variable</u> 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 <u>area under the graph</u> of the <u>probability density function</u> 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 x1 to x2 is defined to be the area under the graph of the **probability density function** between x1 and x2



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) $f(x) \geq 0$

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

(3) $P(a \leq X \leq b) = \int_{a}^{b} f(x)\,dx =$ area under $f(x)$ from a to b for any a and b (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

$$\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

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

$$\int kf(u)du = k\int f(u)du$$

$$\int u^n du = \frac{u^{n+1}}{n+1}$$

$$\int e^u du = e^u$$

$$\int \sin u \, du = -\cos u$$

$$\int \cos u \, du = \sin u$$

$$\int [f(u) \pm g(u)]du = \int f(u)du \pm \int g(u)du$$

$$\int u dv = uv - \int v du$$

Continuous Random Variables

EXAMPLE 1 **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

(a) between 1 and 3;

(b) greater than 0.5.

Solution Evaluating the necessary integrals, we get

(a)
$$\int_1^3 2e^{-2x}\,dx = e^{-2} - e^{-6} = 0.133$$

(b)
$$\int_{0.5}^{\infty} 2e^{-2x}\,dx = e^{-1} = 0.368$$

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

$$F(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ \displaystyle\int_0^x 2e^{-2t}\,dt = 1 - e^{-2x} & \text{for } x > 0 \end{cases}$$

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

$$\mu = \int_{-\infty}^{\infty} x f(x)\,dx = \int_{0}^{\infty} x \cdot 2e^{-2x}\,dx = \frac{1}{2}$$

Alternatively, the expectation of x is $E(X) = 0.5$

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)\,dx = \int_{0}^{\infty} \left(x - \frac{1}{2}\right)^2 \cdot 2e^{-2x}\,dx = \frac{1}{4}$$

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 Exer cise 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 $f(x) = \dfrac{k}{1+x^2}$ 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

$$f(x) = \begin{cases} \dfrac{1}{4.5} e^{-x/4.5} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

Find the probabilities that one of these laptops will pro vide 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

$$f(x) = \begin{cases} \dfrac{1}{4.5} e^{-x/4.5} & \text{for } x > 0 \\ 0 & \text{for } x \leq 0 \end{cases}$$

(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

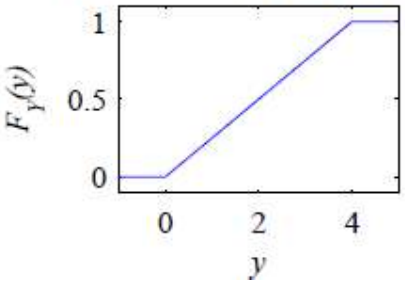
$$F_Y(y) = \begin{cases} 0 & y < 0, \\ y/4 & 0 \le y \le 4, \\ 1 & y > 4. \end{cases}$$

Sketch the CDF of Y and calculate the following probabilities:

(1) $P[Y \le -1]$ *(2)* $P[Y \le 1]$

(3) $P[2 < Y \le 3]$ *(4)* $P[Y > 1.5]$

The CDF of Y is



From the CDF $F_Y(y)$, we can calculate the probabilities:

(1) $P[Y \le -1] = F_Y(-1) = 0$

(2) $P[Y \le 1] = F_Y(1) = 1/4$

(3) $P[2 < Y \le 3] = F_Y(3) - F_Y(2) = 3/4 - 2/4 = 1/4$

(4) $P[Y > 1.5] = 1 - P[Y \le 1.5] = 1 - F_Y(1.5) = 1 - (1.5)/4 = 5/8$

The probability density function of the random variable Y is

$$f_Y(y) = \begin{cases} 3y^2/2 & -1 \le y \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

Sketch the PDF and find the following:

(1) the expected value $E[Y]$

(2) the second moment $E[Y^2]$

(3) the variance $\mathrm{Var}[Y]$

(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

$$F(x) = \begin{cases} 1 - \dfrac{4}{x^2} & \text{for } x > 2 \\ 0 & \text{for } x \le 2 \end{cases}$$

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

$$f_X(x) = \begin{cases} cx & 0 \le x \le 2, \\ 0 & \text{otherwise.} \end{cases}$$

Use the PDF to find

(a) the constant c,

(b) $P[0 \le X \le 1]$,

(c) $P[-1/2 \le X \le 1/2]$,

(d) the CDF $F_X(x)$.

4.4.4 The probability density function of random variable Y is

$$f_Y(y) = \begin{cases} y/2 & 0 \le y < 2, \\ 0 & \text{otherwise.} \end{cases}$$

What are $E[Y]$ and $Var[Y]$?

4.4.5 The cumulative distribution function of the random variable Y is

$$F_Y(y) = \begin{cases} 0 & y < -1, \\ (y+1)/2 & -1 \le y \le 1, \\ 1 & y > 1. \end{cases}$$

What are $E[Y]$ and $Var[Y]$?

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:

$$(i)\ J_{ij} \geq 0 \quad (ii) \sum_{i=1}^{m} \sum_{j=1}^{n} J_{ij} = 1$$

- 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	Sum
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}	$f(x_m)$
Sum	$g(y_1)$	$g(y_2)$	$\dots$	$g(y_n)$	$Total = 1$

- 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): / Education (X)		Age : 25-35 30	Age: 35-55 45	Age: 55-85 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: $p(x) = \Sigma_y\, p(x,y)$
$p(x=1) = .03 + .06 + .10 = .19$

Education (X)	Age (Y):	Age : 25-35 30	Age: 35-55 45	Age: 55-85 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): / Education (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, $P[(X,Y) \in A] = \sum\sum_A f(x,y)$.

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) $P(X \geq 1, Y \leq 2)$.

X \ Y	0	1	2	3
0	0	1/8	1/4	1/8
1	1/8	1/4	1/8	0

Solution:

(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

$X + Y = 0 + 2 \; or \; 0 + 3 \; or \; 1 + 1 \; or \; 1 + 2 \; or \; 1 + 3$

$$P(X + Y \geq 2) = P(0,2) + P(0,3) + P(1,1) + P(1,2) + P(1,3)$$

$$= \frac{1}{4} + \frac{1}{8} + \frac{1}{4} + \frac{1}{8} + 0 = \frac{3}{4}$$

(v) If $X \geq 1, X = \{1\}$. If $Y \leq 2, Y = \{0,1,2\}$

$$P(X \geq 1, Y \leq 2) = P(1,0) + P(1,1) + P(1,2)$$

$$= \frac{1}{8} + \frac{1}{4} + \frac{1}{8} = \frac{1}{2}$$

Problem:

- Two ballpoint pens are selected at random form 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 $0 \le x + y \le 2$.

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,

$x \backslash y$	-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

$$f(x,y)=\begin{cases}\frac{2}{5}(2x+3y), & 0\le x\le 1, 0\le y\le 1\\ 0, & \text{elsewhere}\end{cases}$$

a) Verify whether

b) Find P[(X,Y) ∈ A], where A is the region {(x,y) | 0<x<½,¼<y<½}

c) Find g(x) and h(y) for the joint density function.

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,\ y)dxdy = 1$$

Example 9 – Solution

a)

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = \int_{0}^{1}\int_{0}^{1} \frac{2}{5}(2x+3y)\,dx\,dy$$

$$= \int_{0}^{1} \left.\frac{2x^2}{5} + \frac{6xy}{5}\right|_{x=0}^{x=1} dy$$

$$= \int_{0}^{1} \left(\frac{2}{5} + \frac{6y}{5}\right) dy = \left.\frac{2y}{5} + \frac{3y^2}{5}\right|_{0}^{1}$$

$$= \frac{2}{5} + \frac{3}{5} = 1$$

Example – Solution

b)

$$P[(X,Y)\in A] = P(0 < X < \tfrac{1}{2}, \tfrac{1}{4} < Y < \tfrac{1}{2})$$

$$= \int_{\frac{1}{4}}^{\frac{1}{2}}\int_{0}^{\frac{1}{2}} \frac{2}{5}(2x+3y)\,dx\,dy$$

$$= \int_{\frac{1}{4}}^{\frac{1}{2}} \left.\frac{2x^2}{5} + \frac{6xy}{5}\right|_{x=0}^{x=\frac{1}{2}} dy$$

$$= \int_{\frac{1}{4}}^{\frac{1}{2}} \left(\frac{1}{10} + \frac{3y}{5}\right) dy = \left.\frac{y}{10} + \frac{3y^2}{10}\right|_{\frac{1}{4}}^{\frac{1}{2}}$$

$$= \frac{1}{10}\left[\left(\frac{1}{2} + \frac{3}{4}\right) - \left(\frac{1}{4} + \frac{3}{16}\right)\right] = \frac{13}{160}$$

Example – Solution

By definition,

$$g(x) = \int_{-\infty}^{\infty} f(x,y)dy = \int_{0}^{1} \frac{2}{5}(2x+3y)dy = \frac{4xy}{5} + \frac{6y^2}{10}\Bigg|_{y=0}^{y=1} = \frac{4x+3}{5}$$

For $0 \le x \le 1$, 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 $0 \le y \le 1$, and $h(y)=0$ elsewhere.

Problem:

Find 'K' if the joint probability density function of a bivariate random variable (X,Y) is given by

$$f(x,y)=\begin{cases} K(1-x)(1-y) & if\ 0<x<4 \\ 0 & otherwise \end{cases}$$

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:** $1 = \Sigma f(x,y) = 89c$

$c = \frac{1}{89}$

(ii) $x = 0, y = \{-1, 2\}$

$P(x = 0, y \leq 2)$

$= P(0,-1) + P(0,2)$

$= c + 4c = 5c$

$= 5/89$

(iii) $x = \{-1, 0, 1\},\ y = \{3\}$

$P(x \leq 1, y > 2)$

$= P(-1,3) + P(0,3) + P(1,3)$

$= 10c + 9c + 10c$

$= 29c = 29/89$

Cont.

By data, $X = \{-1, 0, 1, 3\}$ and $Y = \{-1, 2, 3\}$

(iv) $P(x \geq 2 - y) = P(x + y \geq 2)$

$$= P(-1, 3) + P(0, 2) + P(0, 3) + P(1, 2) + P(1, 3) + P(3, -1) + P(3, 2) + P(3, 3)$$

$$= 10c + 4c + 9c + 5c + 10c + 10c + 13c + 18c$$

$$= 79c = 79/89$$

Thank You !

BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Statistical Methods

ISM Team

innovate achieve lead

BITS Pilani

Session No 6
Probability Distributions

(24th/25th June 2023)

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 Bernouli 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

Binomial Distribution

Definition

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

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

Binomial Distribution

- Binomial distribution is a discrete probability distribution.

- Binomial distribution will be applied under the following experimental conditions

1) The number of trails (n) is finite

2) The trails are independent of each other

3) The probability of success p is constant for each trail.

4) Each trail results in two mutually exclusive events known as success and failure.

Mean & variance

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

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

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

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

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

$${}^{(n-1)}C_{x-1}$$

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

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

$$= np$$

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

$$E(x^2) = \sum x^2 p(x) = \sum x^2 \, {}^nC_x p^x q^{n-x}$$

$$= \sum [x(x-1) + x] \, {}^nC_x p^x q^{n-x}$$

$\sum x p(x) \rightarrow$ mean $= np$

$$= \sum x(x-1) \cdot {}^nC_x p^x q^{n-x} + \sum x \, {}^nC_x p^x q^{n-x}$$

$$= \sum x(x-1) \frac{n!}{x!\,(n-x)!} p^x q^{n-x} + np$$

$$= \sum \frac{n(n-1)(n-2)!}{(x-2)!\,[(n-2)-(x-2)]!} p^2 p^{(x-2)} \cdot q^{(n-2)-(x-2)} + np$$

$$= n(n-1)p^2 \sum {}^{(n-2)}C_{x-2} \, p^{x-2} q^{n-x} + np$$

$$= n(n-1)p^2 + np$$

Variance $= \sigma^2 = n(n-1)p^2 + np - (np)^2$

$$= \cancel{n^2p^2} - np^2 + np - \cancel{n^2p^2}$$

$$= np(1-p) = npq$$

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

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. $p(\text{at most } 2) = p(0) + p(1) + p(2) = 0.0687 + 0.2062 + 0.2835$

$= 0.\mathbf{5584}$

D. $p(\text{at least } 2) = p(2) + p(3) \ldots + p(12)$

$= 1 - [p(0) + p(1)] = 1 - .0687 - 0.2062 = 0.\mathbf{7251}$

Understandings

Phrase	*Math Symbol*
"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

2) At least two will hit the target

Example

An unbiased dice is thrown 5 times and occurrence of 1 or 6 considered as success find, .

1) Probability of exactly one success
2) Probability of at least 4 success
3) Probability of at least one success
4) Mean and variance

Example

- Find the binomial distribution if the mean is 4 and variance is 3

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;

$$\text{mean} = \mu = 2, \text{ but } \mu = np \Rightarrow 2 = 20p \therefore p = 0.1$$

$$\therefore P(x) = 20c_x (0.1)^x (0.9)^{20-x}$$

- 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*1000=323 samples

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

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

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

- 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…

- 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 $n \to \infty$ and $p \to 0$ such that $\lambda = np$ = constant

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

Example

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(at most 2 calls) = P(x ≤ 2) = P(0)+P(1)+P(2)=0.2381

(ii) P(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.

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

$$\lambda = 0.6 \text{ and } t = 1 \therefore \quad p(x, \lambda t) = \frac{e^{-\lambda t}(\lambda t)^x}{x!}$$

b)

$$p(x, \lambda t) = \frac{e^{-0.6}(0.6)^x}{x!}$$

$$P(X \geq 1) = 1 - P(X < 1) = 1 - P(0)$$

$$= 1 - e^{-0.6} = 0.4512$$

$$\lambda = 0.6 \text{ and } t = 4 \therefore \quad p(x, \lambda t) = \frac{e^{-2.4}(2.4)^x}{x!}$$

$$P(X \geq 3) = 1 - P(X < 3) = 1 - [P(0) + P(1) + P(2)]$$

$$= 0.4303$$

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.

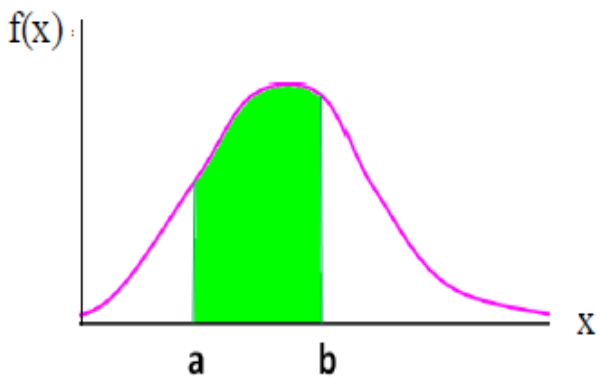
Exercise:

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 x1 to x2 is defined to be the area under the graph of the **probability density function** between x1 and x2



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:

<u>Normal Distribution:</u>

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

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

$$-\infty<x<\infty$$

$$-\infty<\mu<\infty,\sigma>0$$

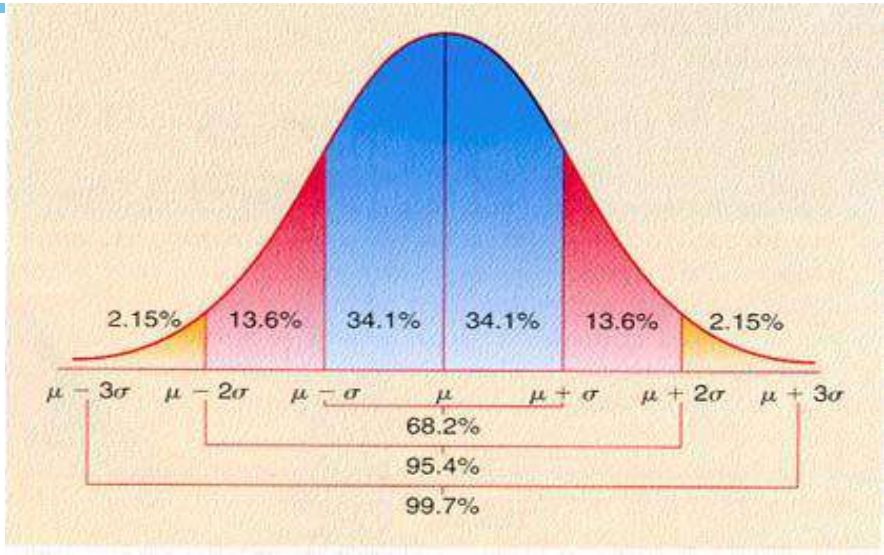
Normal Distribution

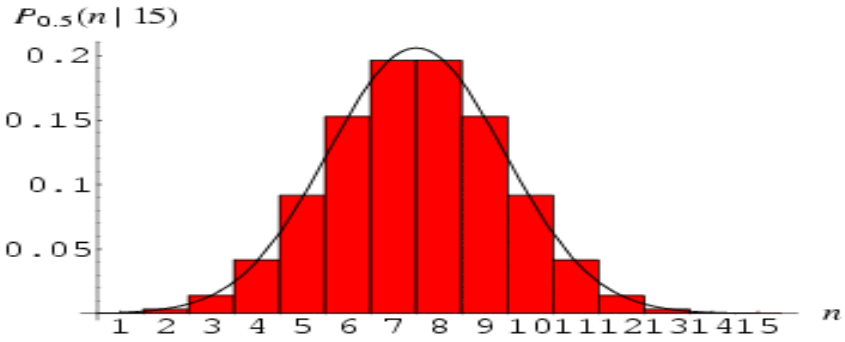
Change in Mean determines the shift in the distribution
Change in the deviation determines the spread of the data points

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

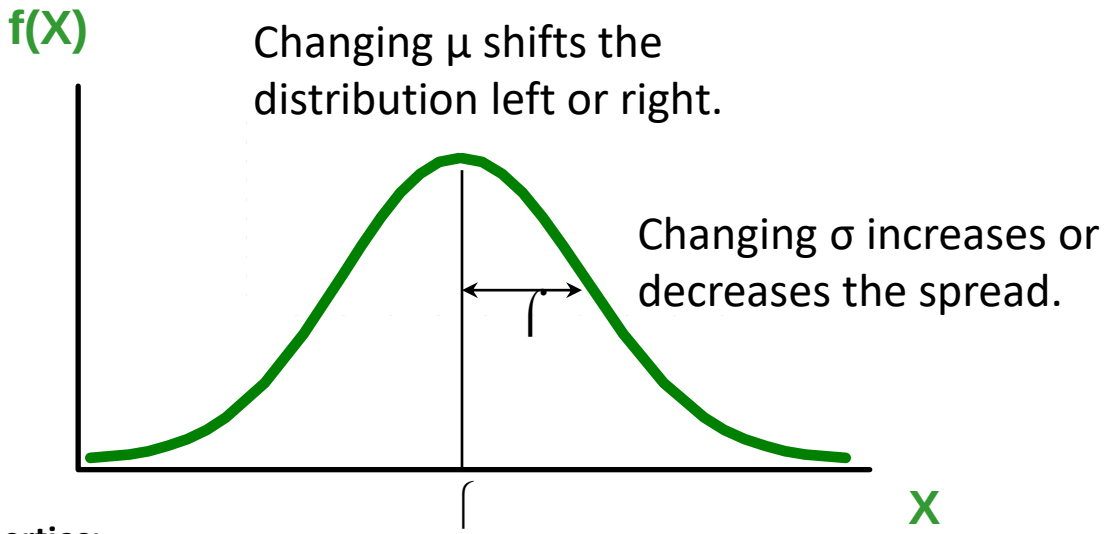
μ = mean
σ = standard deviation
π = 3.14159
e = 2.71828

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)=μ = $$\int_{-\infty}^{+\infty} x \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx$$

Var(X)=σ^2 = $$\int_{-\infty}^{+\infty} x^2 \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} dx) - \mu^2$$

Standard Deviation(X)=σ

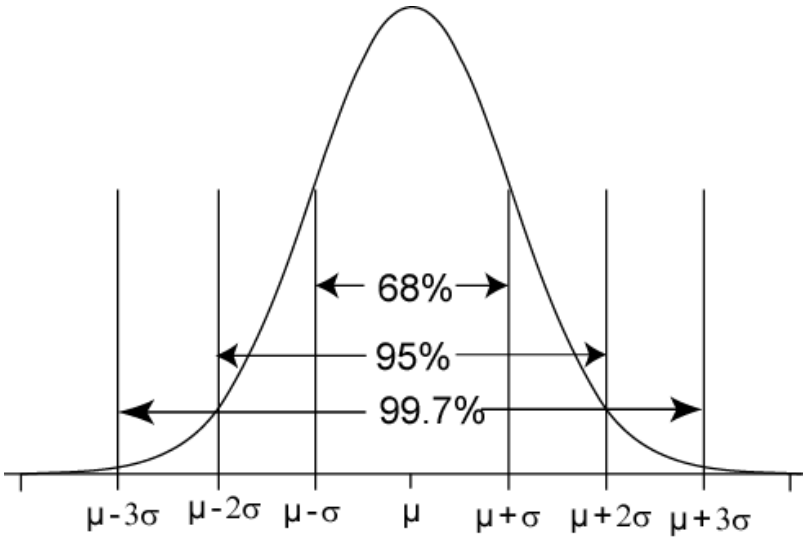
No matter what μ and σ are, the area between μ-σ and μ+σ is about 68%; the area between μ-2σ and μ+2σ is about 95%; and the area between μ-3σ and μ+3σ is about 99.7%. Almost all values fall within 3 standard deviations.

68-95-99.7 Rule for Normal Distributions

68% of the AUC(Area under curve) within ±1σ of μ

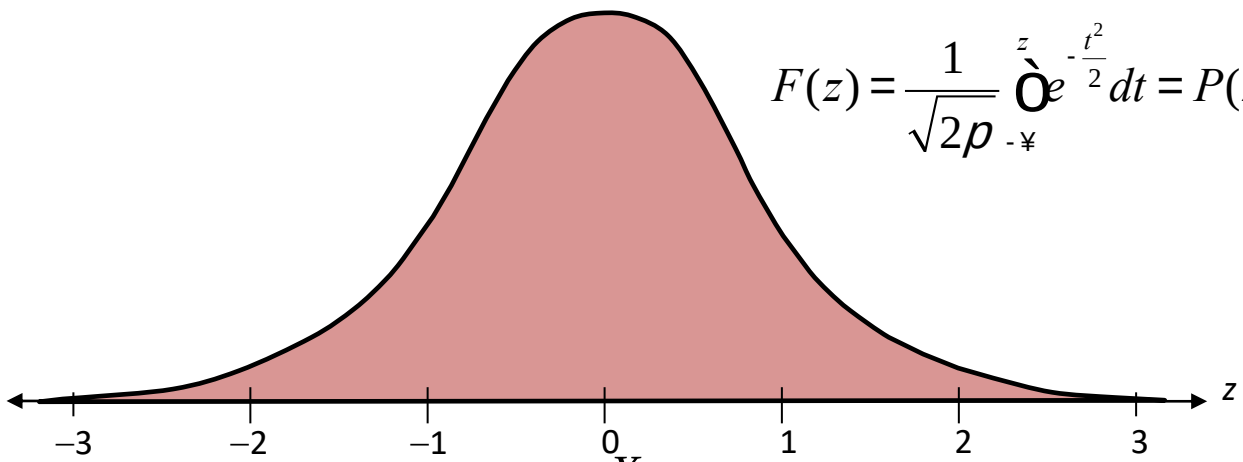
95% of the AUC within ±2σ of μ

99.7% of the AUC within ±3σ of μ



Standard Normal Distribution

The **standard normal distribution** is a normal distribution with a mean of 0 and a standard deviation of 1.



$$F(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt = P(Z \le z)$$

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

All normal distributions can be converted into the standard normal curve by subtracting the mean and dividing by the standard deviation

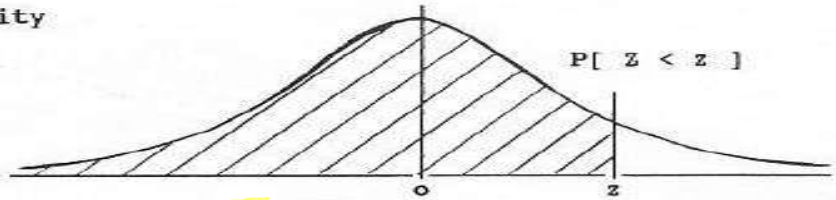
STANDARD STATISTICAL TABLES

1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value z

i.e.

$$P[\ Z < z\ ] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}Z^2)\, dZ$$



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 = $P(Z > 2.06) = 1 - F(2.06) = 0.0197$

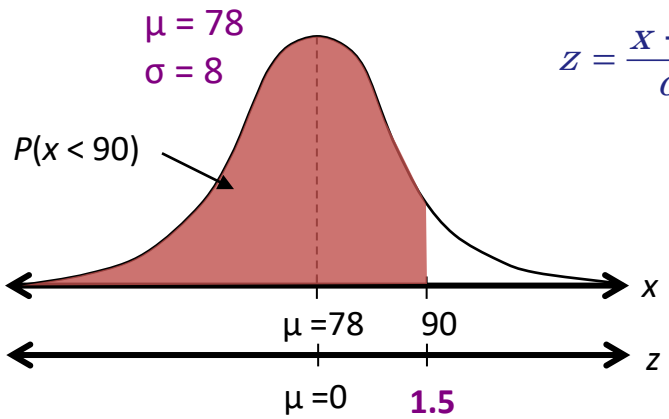
d)greater than -1.82 = $P(Z > -1.82) = 1 - F(-1.82) = 1 - 0.0344 = 0.9656$

$P(Z > -1.82) = 1 - F(-1.82) = 1 - (1 - F(1.82) = F(1.82) = 0.9656$

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.



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

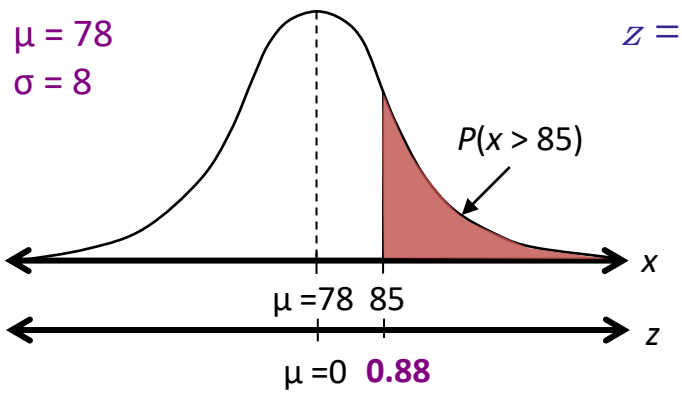
The probability that a student receives a test score less than 90 is 0.9332.

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

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.

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

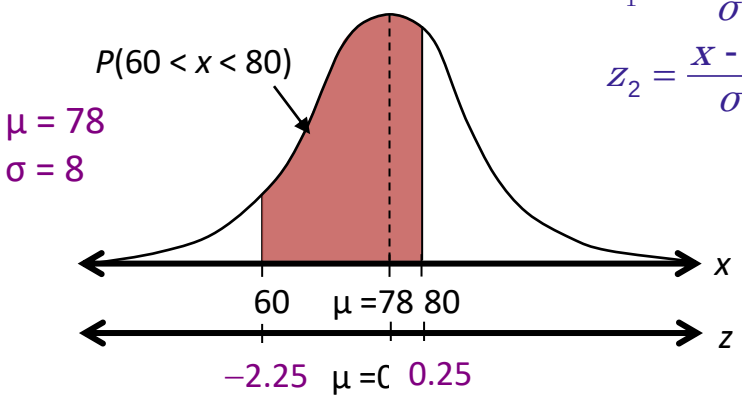
Probability and Normal Distributions

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.

$P(60 < x < 80) = P(-2.25 < z < 0.25) = F(0.25) - F(-2.25) = 0.5987 - 0.0122 = 0.5865$

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 μ = 20 seconds, P(X > 31.5) = 0.25

σ = ?

$$z = \frac{x - m}{s} = \frac{31.5 - 20}{s}$$

$$s = \frac{11.5}{z} = \frac{11.5}{0.675} = 17.04$$

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

$P(X > 450) = P[(450 – 496) /25] = P(Z > -1.84) = 1 – F(-1.84) = 1- 00329 = .9671$

b. between 445.5 and 522C_v

$P(445.5 < X < 522) = P(-2.02 < Z < 1.04) = F(1.04) – F(-2.02)= 0.8508 – 0.0217 = 0.8291$

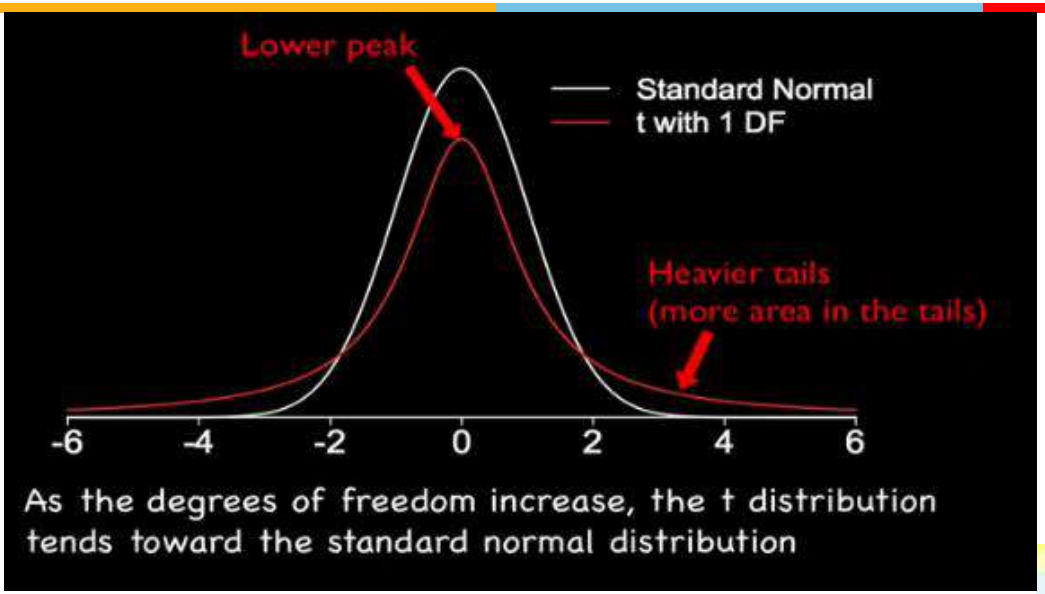
Normal Approximation to Binomial Distribution:

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.

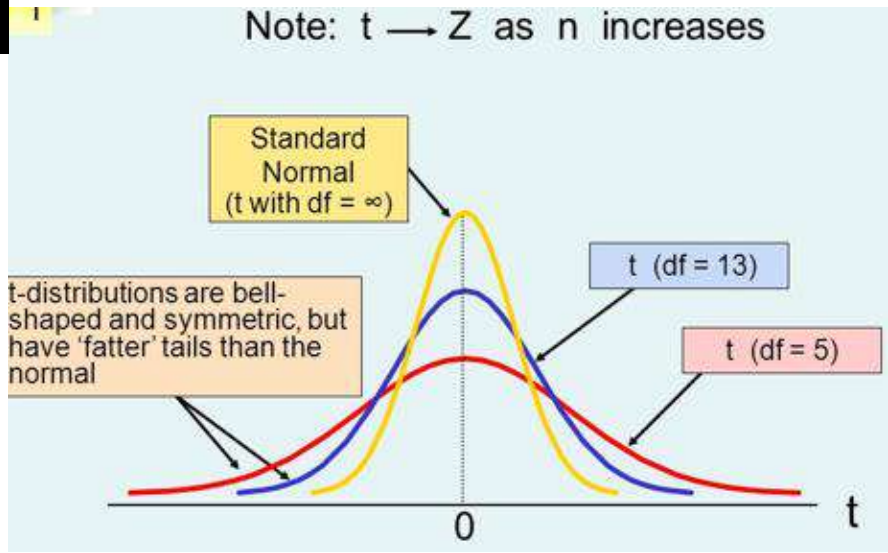
Introduction to t-distribution



The pdf of the t distribution with ν degrees of freedom:

$$f(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\,\Gamma(\frac{\nu}{2})}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

for $-\infty < t < \infty$



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 $n < 30$.

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} \quad with\ (n-1)\ d.o.f.$$

Where sample S.D. is calculated by formula

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(x_i - \bar{x} \right)^2$$

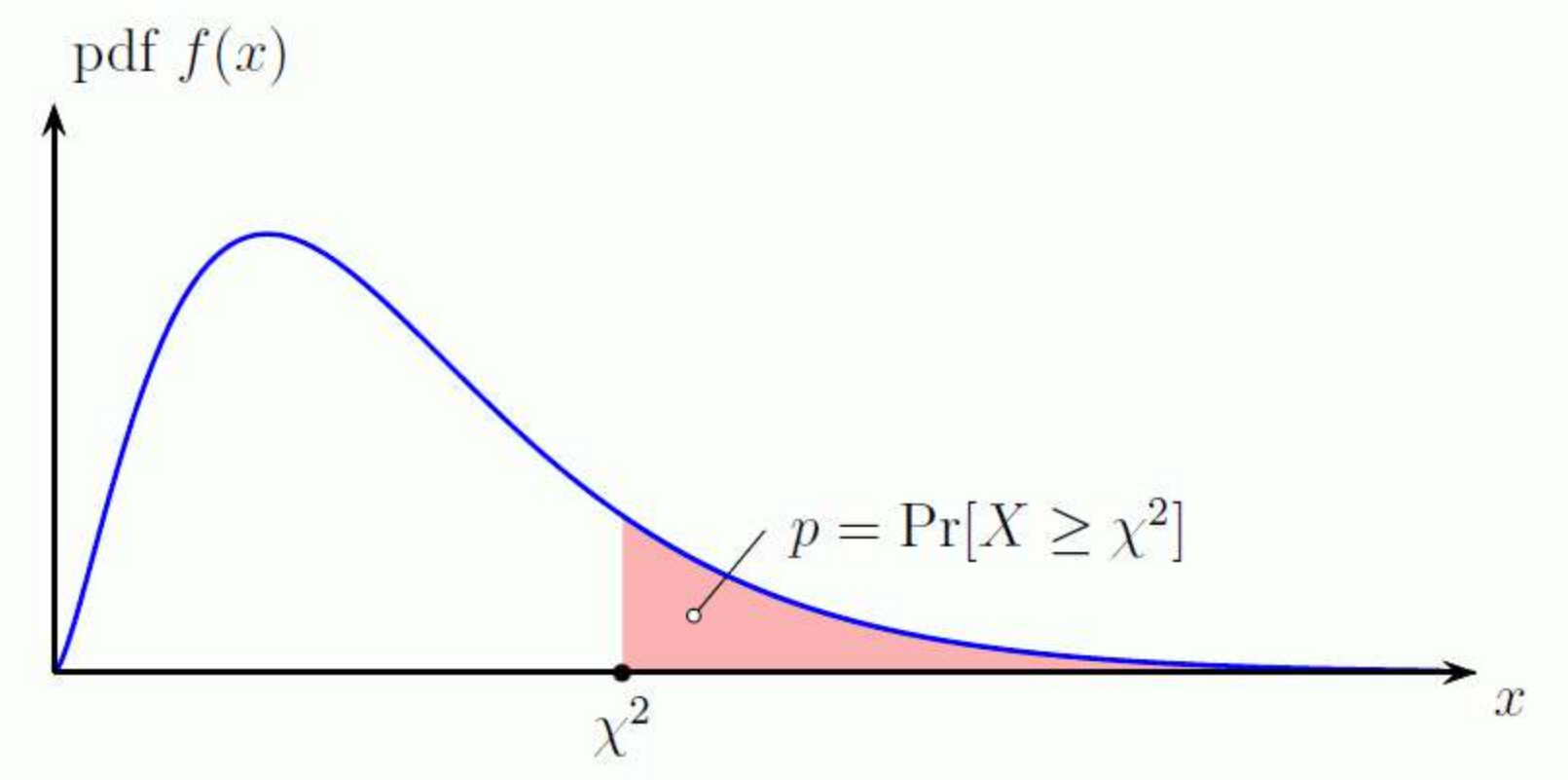
Introduction to Chi- Square(χ^2) distribution

- Chi-Square distribution pdf is given by

$$f(x) = \begin{cases} \dfrac{1}{2^{\vartheta/2}\Gamma(\frac{\vartheta}{2})} x^{\frac{\vartheta}{2}-1} e^{-\frac{x}{2}} & for\ x > 0 \\ 0 & otherwise \end{cases}$$

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

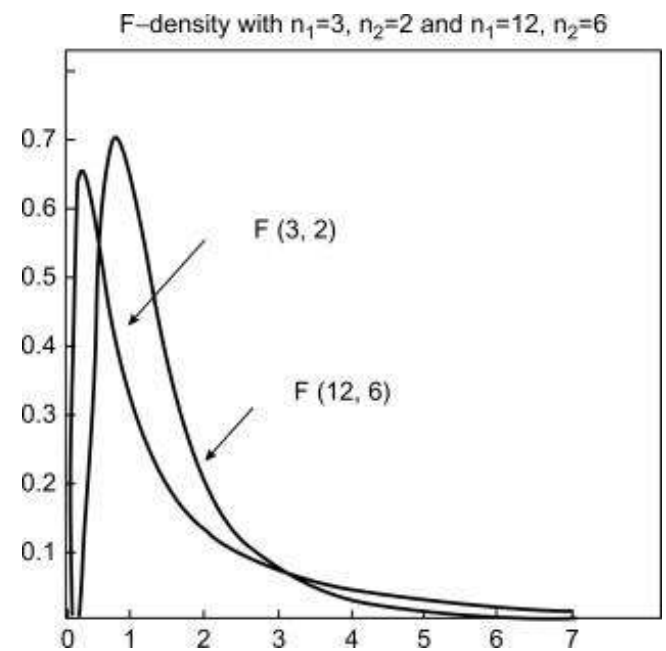
- Chi-Square distribution curve is not symmetrical, and hence not a normal curve.
- Chi-Square varies from 0 to ∞(Curve lies entirely in first quadrant).
- It depends only on degrees of freedom ϑ.
- It is very important in estimation and hypothesis testing.
- It is used in sampling distributions of the 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

$$f(x) = \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}} x^{\frac{n_1}{2}-1}\left(1+\frac{n_1}{n_2}x\right)^{-\frac{n_1+n_2}{2}} \quad , x > 0$$

- 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.



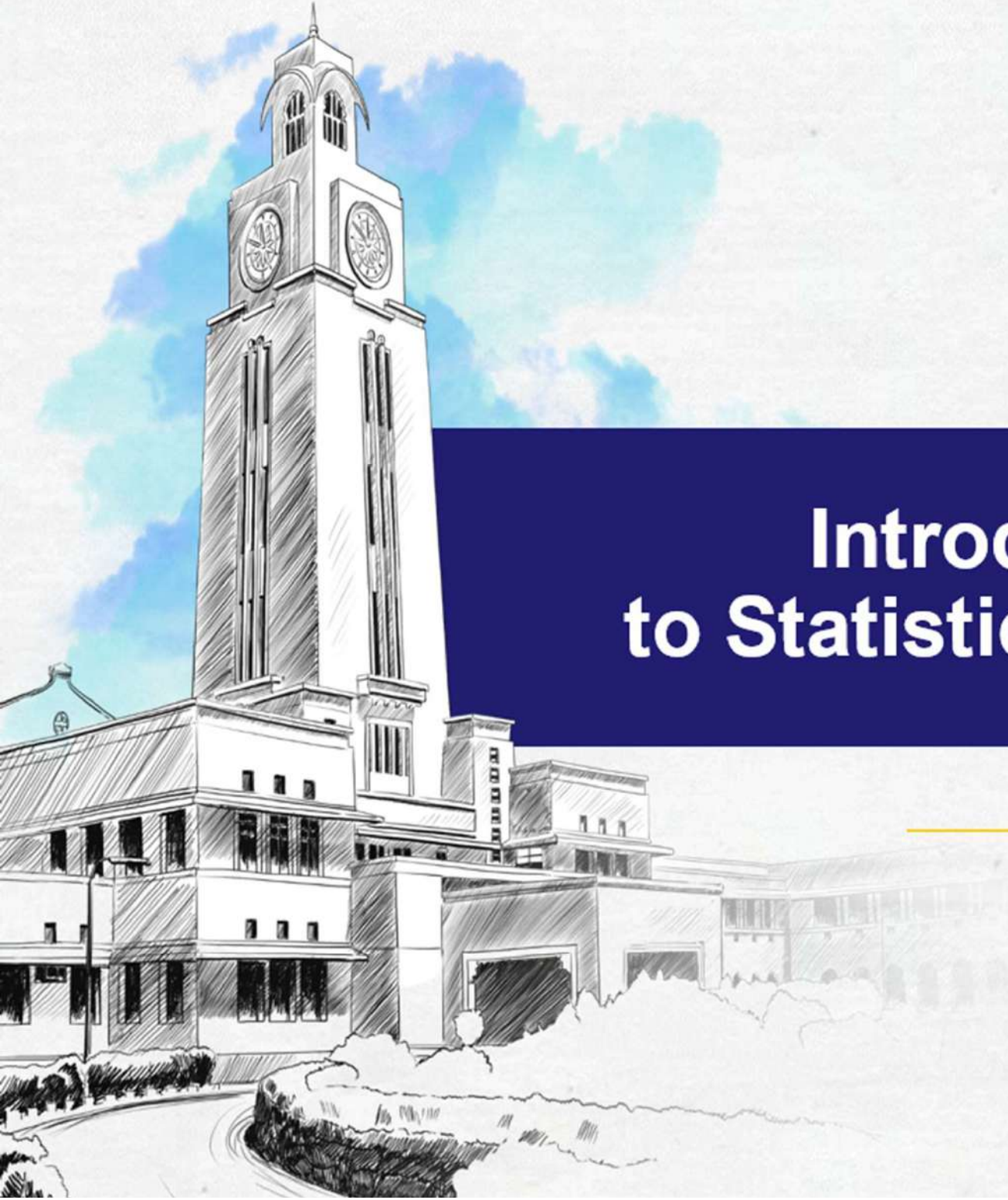
Introduction to F-distribution
(Sampling distribution of the ratio of two sample variances)

- 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

ISM Team

innovate achieve lead

BITS Pilani

Session No 7

Testing of Hypothesis

($1^{st}/2^{nd}$ July, 2023)

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: Terminology

POPULATION (Big data or Huge or Massive data)

A population is the set of all elements of interest in a particular study. The process of conducting a survey to collect data for the entire population is called a census.

SAMPLE (Small data)

A sample is a subset of the population. The process of conducting a survey to collect data for a sample is called a sample survey.

As one of its major contributions, statistics uses data from a sample to make estimates and test hypotheses about the characteristics of a population in the point of view of cost constraint. This part of statistics is called inferential statistics.

Terminology

Parameter

The various statistical characteristics or constants of a population such as Population mean, Population SD, Population Proportion etc., are called parameters.

Statistic

The various statistical functions of a sample such as sample mean, sample sd, sample proportion etc., are called statistics.

A statistic is a function of sample observations and is used to estimate its corresponding parameter. It is also known as estimator and its calculated value is called estimate.

A statistic is a variable and becomes a random variable if the sample is random sample.

Sampling

- Sampling is a method or process of selecting samples from populations.
- 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.

Types of Sampling

Sampling are of two types:

- Random Sampling
- Non-Random Sampling

Random Sampling	Non –Random Sampling
Simple Random Sampling	Judgemental Sampling
Stratified Sampling	Quota Sampling
Systematic Sampling	Convenience Sampling
Clustered Sampling	Snowball Sampling

Random Versus Non random Sampling

- In **random(probability)**sampling every unit of the population has the same probability of being selected into the sample.

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

Simple Random Sampling

In a simple random sample, every member of the population has an equal chance of being selected. The sampling frame should include the whole population.

To conduct this type of sampling, we can use tools like random number generators or other techniques that are based entirely on chance.

Example:
Suppose we want to select a simple random sample of 1000 employees of a social media marketing company. You assign a number to every employee in the company database from 1 to 1000, and use a random number generator to select 100 numbers.

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.

Example:

- A survey about timekeeping might divide the population by time zone, then take 100 random samples per zone.

- A study on tax reform might stratify a population according to income, then take random samples from each stratum.

Non-probability sampling

Non-probability sampling is a sampling method that uses non-random criteria like the availability, geographical proximity, or expert knowledge of the individuals you want to research in order to answer a research question.

Non-probability sampling is used when the population parameters are either unknown or not possible to individually identify.

Example:

A visitors to a website that doesn't require users to create an account could form part of a non-probability sample.

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

12

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

Sample 2

11

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 $\bar{x}$

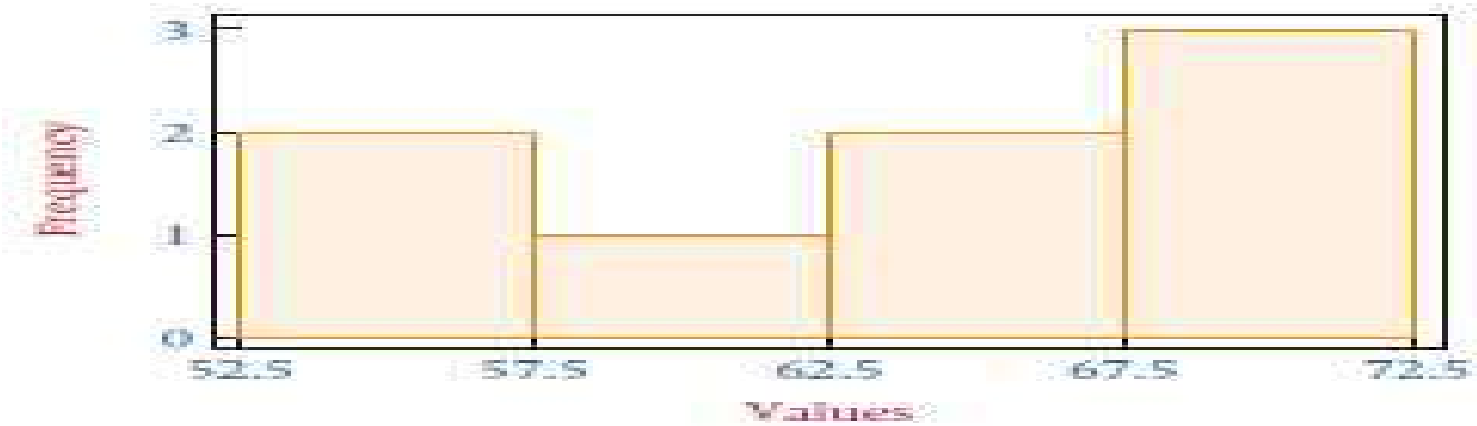
- The sample mean is one of the more common statistics used in the inferential process.

- 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.

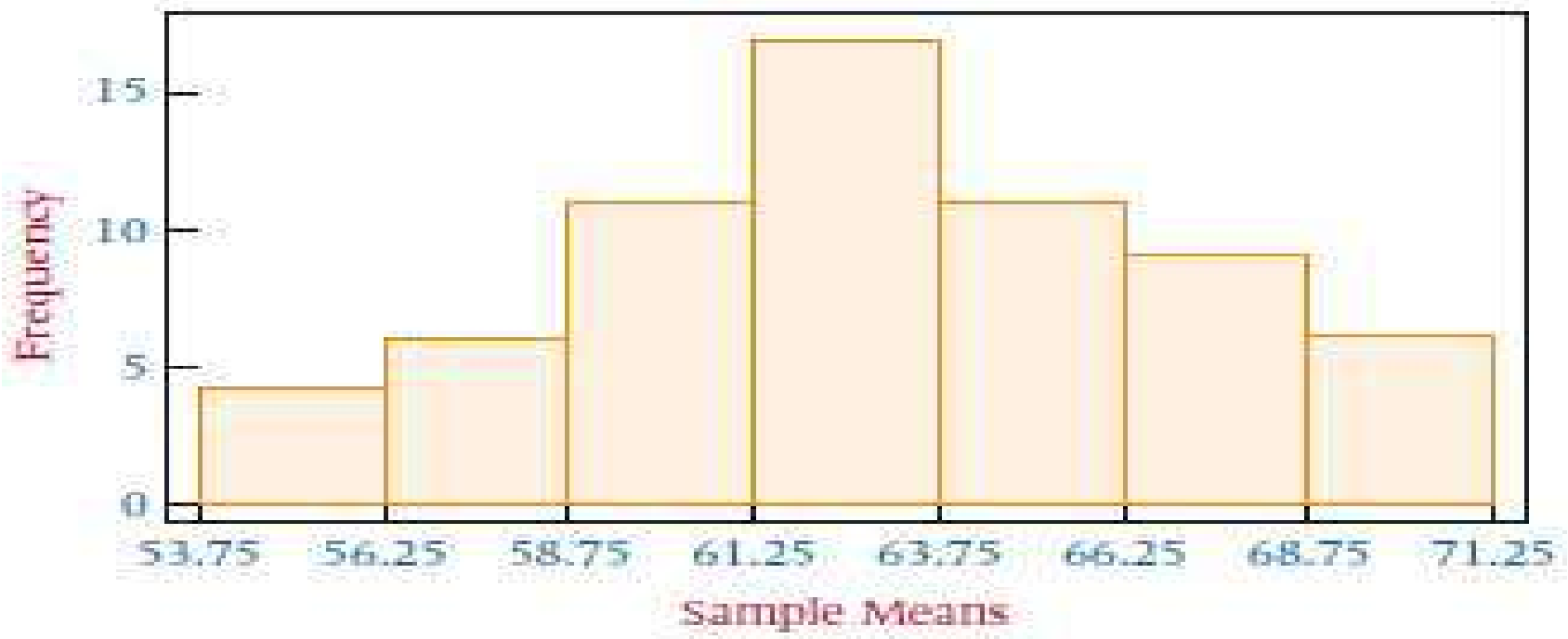
(54,54)	(55,54)	(59,54)	(63,54)
(54,55)	(55,55)	(59,55)	(63,55)
(54,59)	(55,59)	(59,59)	(63,59)
(54,63)	(55,63)	(59,63)	(63,63)
(54,64)	(55,64)	(59,64)	(63,64)
(54,68)	(55,68)	(59,68)	(63,68)
(54,69)	(55,69)	(59,69)	(63,69)
(54,70)	(55,70)	(59,70)	(63,70)
(64,54)	(68,54)	(69,54)	(70,54)
(64,55)	(68,55)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



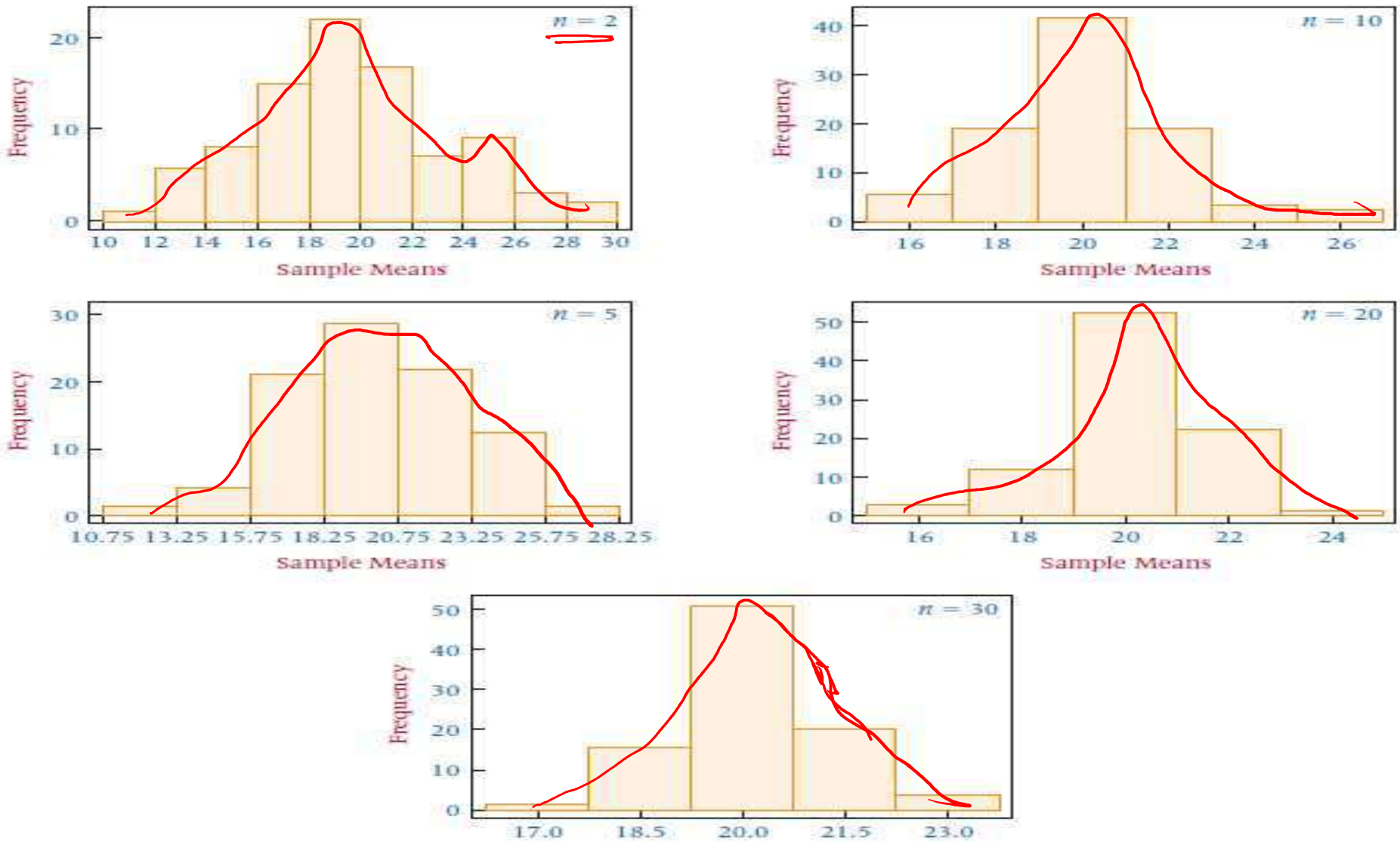
Conclusions

- Notice that the shape of the histogram for sample means is quite unlike the shape of the histogram for the population.

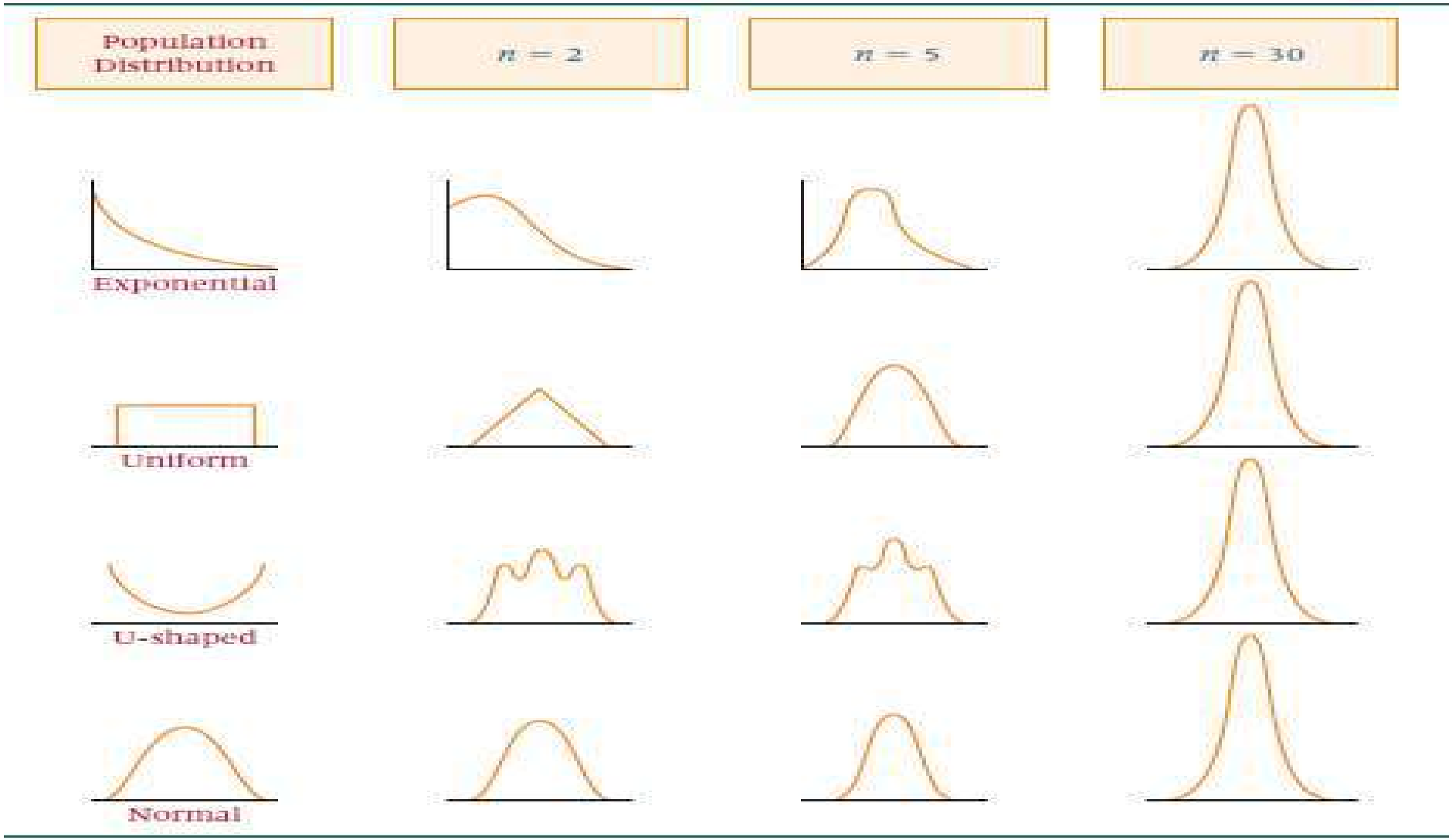
- The sample means appear to “pile up” toward the middle of the distribution and “tail off” toward the extremes.

- 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



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* >= 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

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

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?

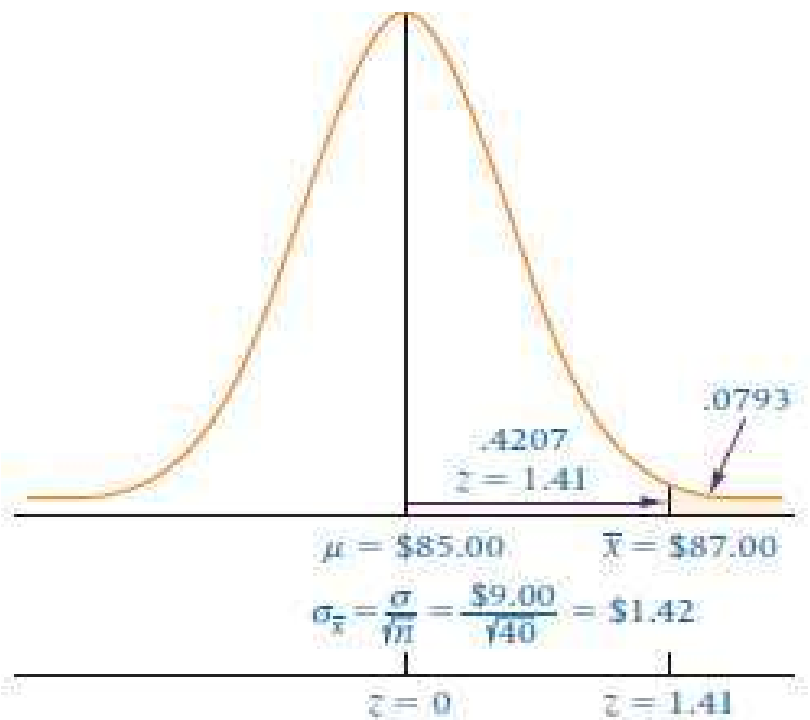
Solution

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

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

μ= \$85 σ= \$9

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



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

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

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

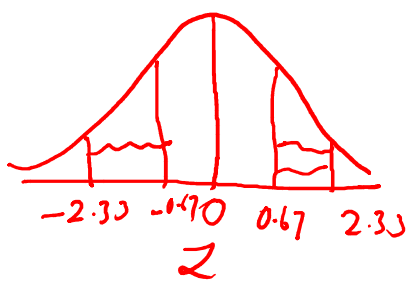
For this problem, μ= 448, σ= 21, and n = 49. The problem is to determine

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

The following

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

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



$$P(-2.33 < Z < -0.67)$$

$$= P(0.67 < Z < 2.33)$$

$$= F(2.33) - F(0.67)$$

$$= 0.9901 - 0.74857$$

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

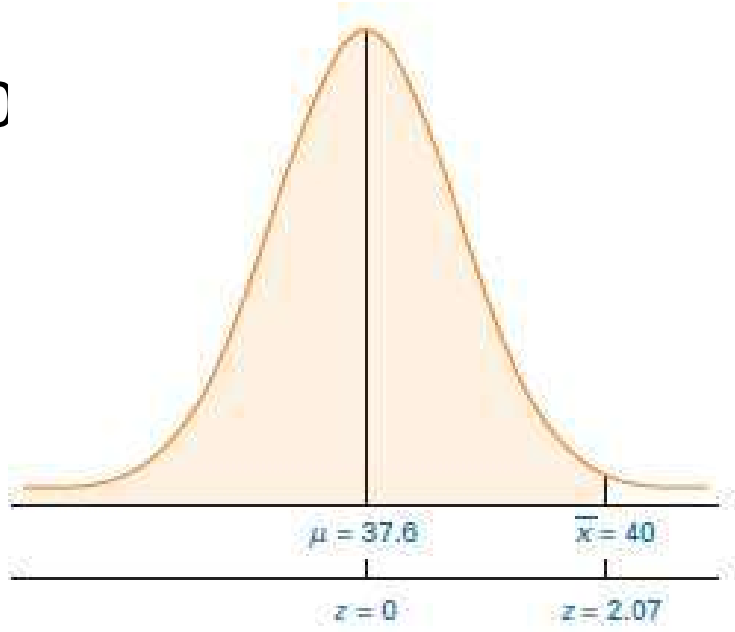
- The population mean is 37.6, with a population standard deviation of 8.3.
- The sample size is 45, but it is being drawn from a finite population of 350; that is, n = 45 and N = 350.
- The sample mean under consideration is 40
- Using the z formula with the finite correction factor gives

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

...solution

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

0.98077
≃ 0.9808



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

$$30/100 = 0.30 = 30\%$$

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 $n*p > 15$ and $n*q > 15$(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 $n*p > 1$ 5 and $n*q > 15$

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

where

$\hat{p}$ = sample proportion
n = sample size
p = population proportion
$q = 1 - p$

Example

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

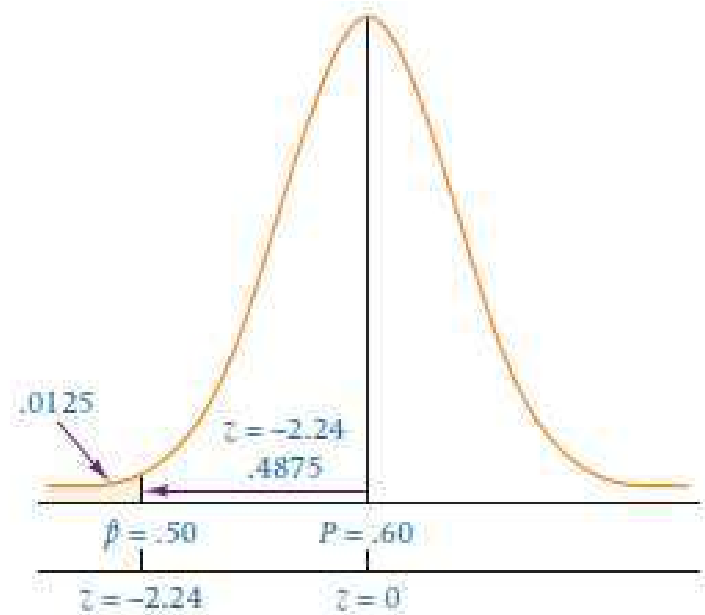
Solution

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

The z formula yields

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

= -2.24 is .4875.



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

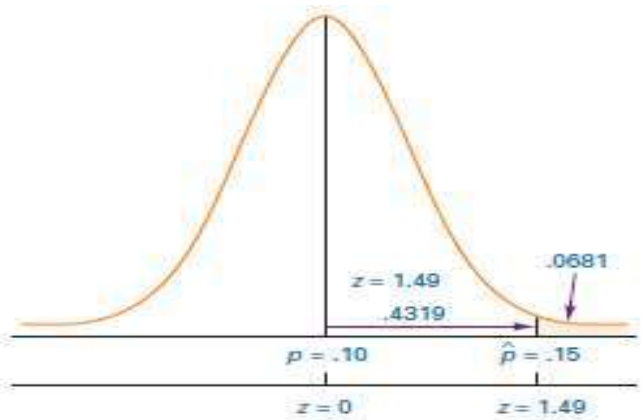
Example

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 between the sample proportion, .15, and the population proportion, .10. The answer to the question is

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

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)

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

or

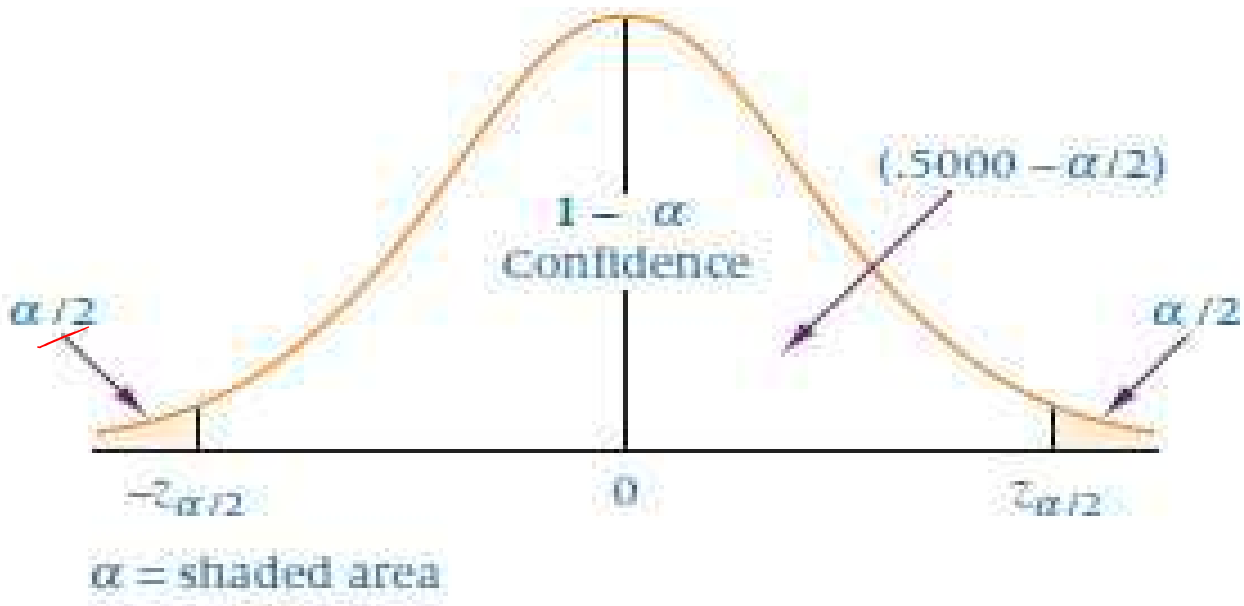
$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

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



Standard normal variable value $z_{\alpha/2}$ (Table Value)	Level of significance $\propto$		
	1%=0.01	**5%=0.05**	**10%=0.1**
$\propto/2$	0.01/2 =0.005	0.05/2=0.025	0.1/2=0.05
F($z_{\alpha/2}$)=1-$\propto/2$ P(Z<=$z_{\alpha/2}$)= 1-$\propto/2$	P(Z<=2.58)= 0.995 then $z_{\alpha/2}$= **2.58**	P(Z<=1.96)= 0.975 then $z_{\alpha/2}$= **1.96**	P(Z<=1.645)= 0.95 then $z_{\alpha/2}$= **1.645**

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 <= μ <= 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 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 σ= 8

- The z value for a 98% confidence interval is 2.33

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \qquad \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

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

$$34.30 - 2.55 \le \mu \le 34.30 + 2.55$$

$$31.75 \le \mu \le 36.85$$

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*p and n*q are **greater** than **5**.

- for confidence interval purposes only and for large sample sizes— is substituted for p in the denominator, yielding

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

Confidence Interval To Estimate P

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

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

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

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

$$.29 \le p \le .49$$

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

$$\hat{p} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \le p \le \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

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

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

$$.45 \le p \le .57$$

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 frm 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 $\boldsymbol{\sigma} = 5.1$.

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 $n = 50$, $\overline{x} = 305.58$ nm, and $s2 = 1366.86$(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

Exercise

A survey was taken of U.S. companies that do business with firms in India. One of the questions on the survey was: Approximately how many years has your company been trading with firms in India?

A random sample of 44 responses to this question yielded a mean of 10.455 years. Suppose the population standard deviation for this question is 7.7 years.

Using this information, construct a 90% confidence interval for the mean number of years that a company has been trading in India for the population of U.S. companies trading with firms in India.

Solution

Here, n= 44, $\bar{x}$= 10.455 and σ= 7.7. To determine the value of $z_{\alpha/2}$, divide the 90% confidence in half, or take .5000 - α/2= .5000 - .0500= 0.45 where α= 10%.

Z table yields a z value of 1.645 for the area of .45

The confidence interval is

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

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

$$10.455 - 1.910 \le \mu \le 10.455 + 1.910$$

$$8.545 \le \mu \le 12.365$$

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

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

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

$$.12 \le p \le .20$$

innovate achieve lead

Thank You

Introduction to Statistical Methods

BITS Pilani

Pilani Campus

innovate achieve lead

BITS Pilani
Pilani Campus

Course No: DSECL ZC413 / AIML ZC418
Course Title: ISM
WEBINAR 1 : 06.06.2023

Topics - Webinar

- Descriptive Statistics
  - Measures of Central Tendency
  - Measures of Variability
- Probability
  - Probability – Introduction and Basics
  - Conditional probability

Measures of Central Tendency

A psychologist wrote a computer program to simulate the way a person responds to a standard IQ test. To test the program, he gave the computer 15 different forms of a popular IQ test and computed its IQ from each form

IQ Values:

134	136	137	138	138	143	144	144	145	146	146	146
	147	148	153								

Find the following Statistical measures:

i. Mean, median, and mode

ii. Rang, Variance and standard deviation

iii. Five point summary of the data.

iv. The interquartile range.

v. Identify potential outliers, if any.

vi. Construct and interpret a boxplot

Measures of Central Tendency

2 3 4 5
Average.

Arranging the data in ascending order:

M

134, 136, 137, 138, 138, 143, 144, 144, 145, 146, 146, 146, 147, 148, 153

N=15

N=

S.No	Formula	Solution
Mean	$\mu = \frac{\sum_{i=1}^{n} x_i}{n}$	μ = 143
Median	$p = \frac{n+1}{2}$ th $= \frac{15+1}{2} = 8^{th}$ item	Median = 144
Mode	The mode is the value or values that occur most frequently in the data set. A data set can have more than one mode, and it can also have no mode	Mode = 146

x	f
134	1
136	1
⋮	
144	2
146	3

Mode

144 3
146 3
148 3

Unimodal

147 3

146, 147

Bimodal.

Multimodal

Mode, Bimodal, and Multimodal

A given set of data may have one or more than one Mode. A set of numbers with one Mode is unimodal, a set of numbers having two Modes is bimodal, a set of numbers having three Modes is trimodal, and any set of numbers having four or more than four Modes is known as multimodal.

A set of values having a single unique mode is said to be unimodal, one with two modes is called bimodal, and one with three modes is called trimodal.

Bimodal Mode – A set of data including two modes is identified as a bimodal model. This indicates that there are two data values that possess the highest frequencies. For example, the mode of data set B = { 8, 12, 12, 14, 15, 19, 17, 19} is 12 and 19 as both 12 and 19 are repeated twice in the given set.

No mode:
If no number in a set of numbers occurs more than once, that set has no mode: 3, 6, 9, 16, 27, 37, 48.

Population $\mu =$

Sample $\bar{x}$

Measures of Central Tendency

Range	$\text{Range} = x_n - x_1$	Minimum = 134 Maximum =153 Range R = 19
Variance	For a Population $\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$ For a Sample $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$	Variance = 26 $x \mid (x-\mu)^2$ Σx $\mu =$
Standard deviation	For a Population $\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$ For a Sample $s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$	Standard Deviation = 5.09901951

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

$$= \frac{\sum_{i=1}^{N}(X^2 - 2\mu X + \mu^2)}{N}$$

$$= \frac{\sum X^2}{N} - \frac{2\mu\sum X}{N} + \frac{N\mu^2}{N}$$

$$= \frac{\sum X^2}{N} - 2\mu^2 + \mu^2$$

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

$\sum_{i=1}^{N} \mu^2$

cons.

Σx_i^2

$$= \frac{\Sigma x^2}{N} - \left(\frac{\Sigma x}{N}\right)^2$$

x	x^2
Σx	Σx^2

Measures of Central Tendency

Min

134 136 137 138 138 143 144 144 145 146 146 146 147

148 153

Max

Q_1 M Q_3

- **Minimum**.

$\frac{1}{4}$ N^{th}

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

Measures of Central Tendency

Quartiles	Quartiles separate a data set into four sections. The median is the second quartile Q_2. It divides the ordered data set into higher and lower halves. The first quartile, Q_1, is the median of the lower half not including Q_2. The third quartile, Q_3, is the median of the higher half not including Q_2.	Quartiles: Q_1 --> 138 Q_2 — Mead --> 144 Q_3 --> 146
Interquartile range	IQR=Q3−Q1	Interquartile Range IQR =8
Potential outliers, if any.	$\text{Upper Fence} = \overset{146}{Q_3} + \boxed{1.5 \times \overset{8}{IQR}} = 158$ $\text{Lower Fence} = Q_1 - \boxed{1.5 \times IQR} = 126$	none

Major outliers .. $Q_3 + 3(IQR)$

Measures of Central Tendency – Box plot

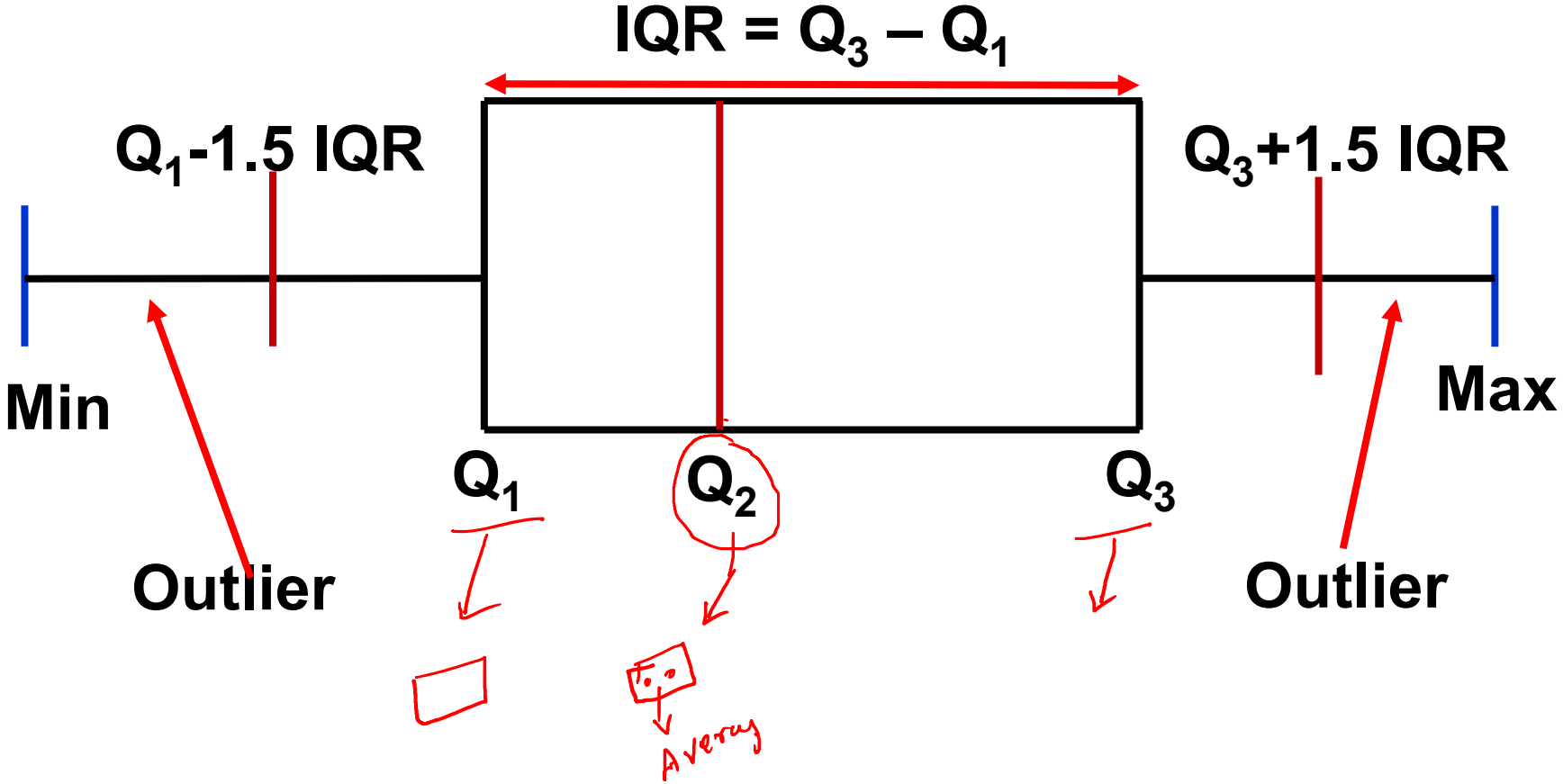
Population size: 15
Median: 144
Minimum: 134
Maximum: 153
First quartile: 138
Third quartile: 146
Interquartile Range: 8
Outliers: none

(126, 158)

M = middle $\frac{N+1}{2}$ th.

$Q_1 = \frac{1}{4}(N+1)$

$Q_3 = \frac{3}{4}$



Q1 , Median (Q2) and Q3

Example of Quartiles (n is odd)

Suppose the distribution of math scores in a class of 19 students in ascending order is:

- 59, 60, 65, 65, 68, 69, 70, 72, 75, **75**, 76, 77, 81, 82, 84, 87, 90, 95, 98

First, mark down the median, Q2, which in this case is the 10th value: 75.

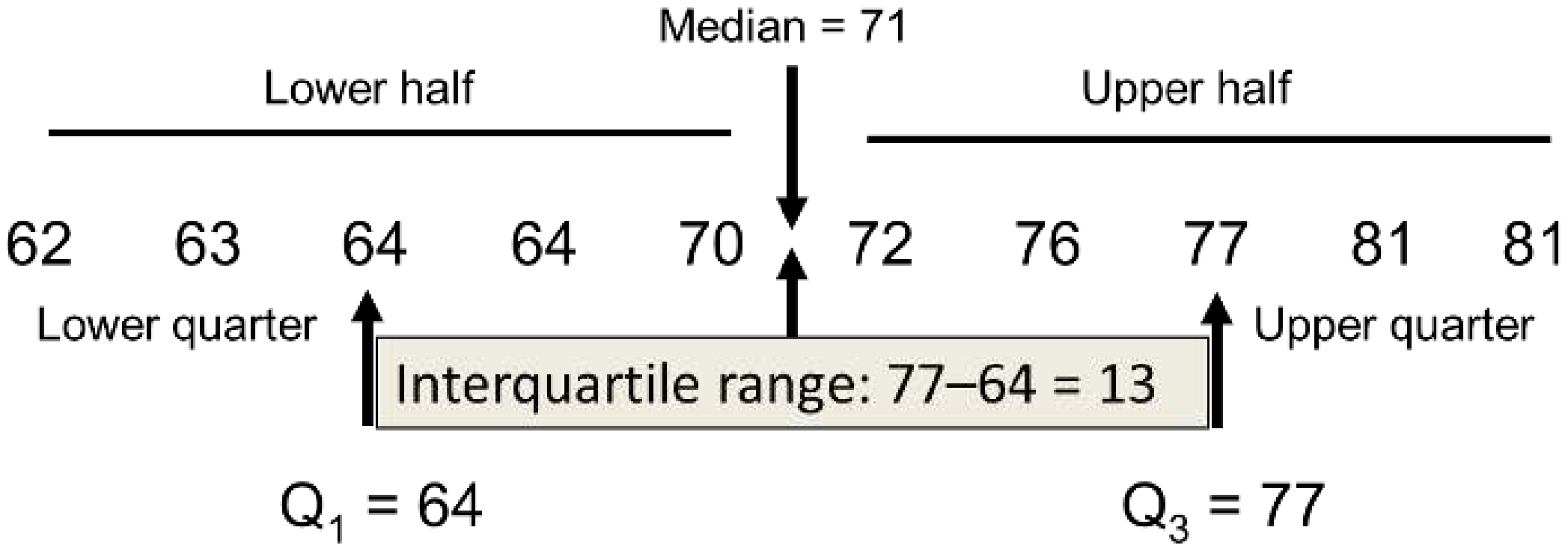
Q1 is the central point between the smallest score and the median. In this case, Q1 falls between the first and fifth score: 68. (Note that the median can also be included when calculating Q1 or Q3 for an odd set of values. If we were to include the median on either side of the middle point, then Q1 will be the middle value between the first and 10th score, which is the average of the fifth and sixth score—(fifth + sixth)/2 = (68 + 69)/2 = 68.5).

Q3 is the middle value between Q2 and the highest score: 84. (Or if you include the median, Q3 = (82 + 84)/2 = 83).

Now that we have our quartiles, let's interpret their numbers. A score of 68 (Q1) represents the first quartile and is the 25th percentile. 68 is the median of the lower half of the score set in the available data—that is, the median of the scores from 59 to 75.

Q1 tells us that 25% of the scores are less than 68 and 75% of the class scores are greater. Q2 (the median) is the 50th percentile and shows that 50% of the scores are less than 75, and 50% of the scores are above 75. Finally, Q3, the 75th percentile, reveals that 25% of the scores are greater and 75% are less than 84.

Q1 , Median (Q2) and Q3 (n is even)



Consider the following statistical summary of a dataset. Write at least three useful observations as a part of data pre – processing. **[Midsem Sep 2023]**

	Nr	Cells	QValue	Fat	Protein
count	969.000000	969.000000	969.000000	969.000000	969.000000
mean	7.074303	358.284830	90.016584	3.620279	3.300196
std	4.759793	344.324223	4.998924	0.349956	0.136071
min	1.000000	0.000000	62.650000	2.240000	2.750000
25% Q1	3.000000	157.000000	87.570000	3.410000	3.220000
50%	6.000000	283.000000	90.750000	3.610000	3.310000
75% Q3	10.000000	476.000000	93.400000	3.820000	3.380000
max	20.000000	5226.000000	100.000000	5.420000	3.840000

I. Direct Observation

- Mea

Range

- Coeff. Var = Sd/Mean → Protein
- Range → Protein
- $Q_3 - Q_1$

Contin….

Observation 1: The variation in Protein is found very low when compared with re maining using coefficient of variation = sd/mean

Observation 2: The range in Protein is found very low when compared with remain ing using range = max-Min

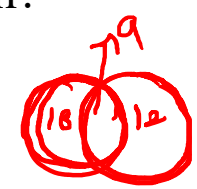
Observation 3: The middle range in Protein is found very low when compared with remaining using quartile range = Q3(75%)-Q1(25%)

Etc……

Basic Probability

In a sample of 100 nails, 18 of them have ~~only~~ defective heads (A), 12 contain defective tail ends (B) and 9 contain both (A∩B) defectives.

What is the probability that a nail selected at random is either a defective head or a defective tail? (A ∪ B)



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

$$= \frac{18}{100} + \frac{12}{100} - \frac{9}{100}$$

$$= 0.21$$

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

Not

Basic Probability

If $P(A) = 1/2$, $P(B) = 1/3$ and $P(A \cap B) = 1/5$ then find

a). $P(A \cup B)$ b). $P(A^c \cap B)$ c). $P(A \cap B^c)$ $\rightarrow P(A) - P(A \cap B)$

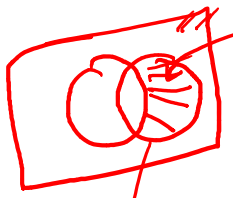
d). $P(A^c \cap B^c)$ e). $P(A^c \cup B^c)$ f). $P((A \cup B)^c)$

$P(A \cup B)^c$

$= 1 - P(A \cup B)$

$P(A^c \cap B) = P(B - (A \cap B)) = P(B) - P(A \cap B)$

$P(A^c)P(B)$



Basic Probability

If $P(A) = 1/2$, $P(B) = 1/3$ and $P(A \cap B) = 1/5$ then find

a). $P(A \cup B)$ b). $P(A^c \cap B)$ c). $P(A \cap B^c)$

d). $P(A^c \cap B^c)$ e). $P(A^c \cup B^c)$ f). $P((A \cup B)^c)$

If $P(A) = 1/2$, $P(B) = 1/3$ and $P(A \cap B) = 1/5$ then find

a). $P(A \cup B)$ b). $P(A^c \cap B)$ c). $P(A \cap B^c)$

d). $P(A^c \cap B^c)$ e). $P(A^c \cup B^c)$ f). $P((A \cup B)^c)$

Solution

a) $P(A \cup B) = P(A) + P(B) - P(A \cap B) = 1/2 + 1/3 - 1/5 = 19/30 = 0.6333$

b) $P(A^c \cap B) = P(B - A) = P(B - (A \cap B)) = P(B) - P(A \cap B) = 1/3 - 1/5 = 2/15 = 0.1333$

c) $P(A \cap B^c) = P(A - B) = P(A - (A \cap B)) = P(A) - P(A \cap B) = 1/2 - 1/5 = 3/10 = 0.3$

d) $P(A^c \cap B^c) = P((A \cup B)^c) = 1 - P(A \cup B) = 1 - 19/30 = 11/30 = 0.3667$

e) $P(A^c \cup B^c) = P((A \cap B)^c) = 1 - P(A \cap B) = 1 - 1/5 = 4/5 = 0.8$

f) $P((A \mid\mid B)^c) = 1 - P(A \mid\mid B) = 1 - 19/30 = 11/30 = 0.3667$

Basic Probability

Consider the following data from the industry. Let D be the set of software developers in a company . Some subsets of D and some of probability that a random employee will belong to the set are represented below. Find the missing term in the table with justification.

Try to identify inconsistencies also (if exists and validate them). **[Midsem Jan 2023]**

Set in words	**Symbol**	**Probability**
All software developers	D	?
All software testing engineers	T	?
All software developers ~~and~~ / or software testing engineers	$D \cup T$	0.25
All software developers who are not software testing engineers	$D - T$	0.05
All software testing engineers who are not software developers	$T - D$	0.10
All software developers and software testing engineers	$T \cap D$	?

$D - (D \cap T)$

$D \cup T^c$

$P(D - T) =$

$P(D - (D \cap T)) = 0.05$

$P(D) - P(D \cap T) = 0.05$

$P(T) - P(D \cap T) = 0.10$

$[P(D) + P(T) - P(D \cap T)] - P(D \cap T) = 0.15$

$0.25 - P(D \cap T) = 0.15$

$P(D \cap T) = 0.1$

$P(D) = 0.15$

$P(T) = 0.2$

Basic Probability

A political leader has submitted his nomination to compete in two different electoral constituencies namely A1 and A2. The probability of wining in constituency A1 and A2 is 0.80 and 0.65 respectively. The probability of losing at least one of the constituencies is 0.35. What will be the probability that he will win in one of the constituencies?

A B

- A –

$P(A) = 0.8$, A^c

$P(B) = 0.65$ B^c

$P(A^c \cup B^c) = 0.35$

$P((A \cap B)^c) = 0.35$

$\boxed{P(A \cap B) = 0.65}$

$A \cup B$

$P(A^c \cup B^c) = (A \cap B)^c$

$A^c \cup B^c = (A \cap B)^c$

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

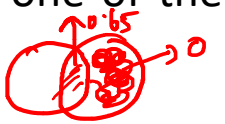
Required $= P(A \cup B) - P(A \cap B)$

$= 0.8 - 0.65$

$= 0.15$ //

Basic Probability

A political leader has submitted his nomination to compete in two different electoral constituencies namely A1 and A2. The probability of wining in constituency A1 and A2 is 0.80 and 0.65 respectively. The probability of losing at least one of the constituencies is 0.35. What will be the probability that he will win in one of the constituencies?



Assume that A, B be the events defined as follows:

A : "Winning in constituency A1"

B : "Winning in constituency A2"

Given:

P(A) = 0.80, P(B) = 0.65

and $P(\bar{A} \cup \bar{B}) = 0.35$

Now, $\because P(\bar{A} \cup \bar{B}) = 0.35$

$\therefore P(\overline{A \cap B}) = 0.35$

$\Rightarrow 1 - P(A \cap B) = 0.35$

$\Rightarrow \boxed{P(A \cap B) = 0.65}$

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

= 0.80 + 0.65 – 0.65

$\therefore P(A \cup B) = 0.80$

Then, $P(He\ will\ win\ in\ one\ of\ the\ constituencies) = P(A \cup B) - P(A \cap B)$

$= 0.80 - 0.65$

$\therefore P(He\ will\ win\ in\ one\ of\ the\ constituencies) = 0.15$

$P(He\ will\ win\ in\ constituency\ A1\ ONLY) = P(A) - P(A \cap B)$

= 0.80 – 0.65

= 0.15

$P(He\ will\ win\ in\ constituency\ A2\ ONLY) = P(B) - P(A \cap B)$

= 0.65 – 0.65

= 0

Independent events

Comment on the statement:

"If two events are mutually exclusive, then they are independent also and vice versa"

$P(A \cap B) = 0$

$\downarrow$

$P(A) \times P(B) = 0$

either $P(A) = 0$ or $P(B) = 0$

Two independent events cannot be mutually exclusive events - unless one or both events have a probability of zero (meaning one of the events is impossible).

Independent events

Let A and B be the two possible outcomes of an experiment and suppose P(A) = 0.4, P(B) = p and P(AUB) = 0.7

(i) For what choice of ‘p’ are A and B mutually exclusive?

(ii) For what choice of ‘p’ are A and B independent?

Independent events

No common element

Let A and B be the two possible outcomes of an experiment and suppose P(A) = 0.4, P(B) = p and P(AUB) = 0.7

(i) For what choice of 'p' are A and B mutually exclusive?

element

(ii) For what choice of 'p' are A and B independent?

Prob

P(A

(i) If A and B are mutually exclusive then $P(A \cap B) = 0$

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

$0.7 = 0.4 + P(B) - 0$

$\therefore P(B) = 0.7 - 0.4 = 0.3$

(ii) If A and B are independent then $P(A \cap B) = P(A).P(B)$

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

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

$$\begin{aligned} 0.7 &= 0.4 + P - 0.4 \cdot P \\ &= 0.4 + P(1 - 0.4) \\ &= 0.4 + 0.6P \\ 0.6P &= 0.7 - 0.4 = 0.3 \\ P &= \frac{0.3}{0.6} = \frac{1}{2} = 0.5 \end{aligned}$$

$\therefore P(B) = 0.5$

Answer. (i) p=0.3, (ii) p=0.5

Dependent events – Conditional Probability

If P (A) =1/2, P (B) =1/3, P (A/B) = 1/6 find i). P(B/A) and ii). P(AUB / A)

A, B

$P(B) > 0$

$P(A/B)$

$\{P(\text{pass}/\text{Teacher}) = 1.00$

$P(\text{pass}) = 0.75$

A & B depend.

- $P(A/B) = \dfrac{P(A \cap B)}{P(B)}$

- A & B independent

$P(A/B) = P(A)$

- $\boxed{P(A)}$

$P(B/A) = \dfrac{P(B \cap A)}{P(A)}$

$P(A/B) = \dfrac{P(A \cap B)}{P(B)} \Rightarrow P(A \cap B) = P(A/B)\,P(B)$

$P(B \cap A) = \frac{1}{6} \times \frac{1}{3}$

$P[\text{pass}/\text{height}] =$

$P(\text{pass})$

$P(A \cup B/A) = \dfrac{P[(A \cup B) \cap A]}{P(A)} = \dfrac{P(A)}{P(A)} = 1$

Conditional Probability

Consider the following data. **[Midsem Jan 2023]**

A: the event that a patient selected is with high B.P

B: the event that a patient selected is diabetic

C: the event that a patient selected is with cancer

Events	A	B	C	A and B	B and C	A and C	(A and B and C)
Probability	0.35	0.40	0.45	0.22	0.25	0.22	0.15

i) If a patient is selected at random, what is the probability that he suffering from at least one of these health issues?

$P(A \cup B \cup C) =$

i) What is the probability that the patient selected is with B.P given that he is with either of the remaining two diseases?

$P(A / B \cup C)$

iii) What is the probability that the patient is diabetic given that he is suffering from at least one of the diseases mentioned?

$P(B / A \cup B \cup C)$

$P(A/B)$ given already occured

Conditional Probability

F A

In an online shopping survey, 30% of persons made shopping in Flipkart, 45% of persons made shopping in Amazon and 5% made purchases in both. If a person is selected at random, find

i) the probability that he makes shopping in at least one of two companies $P(F \cup A)$

ii).the probability that he makes shopping in Amazon given that he already made shopping in Flipkart. $P(A/F)$

iii).the probability that the person will not make shopping in Flipkart given that he already made purchase in Amazon.

$P[F^c/A]$

Conditional Probability

In an online shopping survey, 30% of persons made shopping in Flipkart, 45% of persons made shopping in Amazon and 5% made purchases in both. If a person is selected at random, find

i) the probability that he makes shopping in at least one of two companies

3rd law. 0.25

ii).the probability that he makes shopping in Amazon given that he already made shopping in Flipkart.

iii).the probability that the person will not make shopping in Flipkart given that he already made purchase in Amazon.

Solution: Given $P(F) = 30\% = 0.30$

$P(A) = 45\% = 0.45$

$P(F \cap A) = 5\% = 0.05$

i) $P(F \cup A) = P(F) + P(A) - P(F \cap A)$

$= 0.30 + 0.45 - 0.05 = 0.7$

ii) $P(A \mid F) = \frac{P(A \cap F)}{P(F)}$

$= \frac{0.05}{0.30} = 0.167$

iii) $P(F' \mid A) = \frac{P(F' \cap A)}{P(A)}$

$P(F' \cap A) = P(A) - P(A \cap F)$

$= 0.45 - 0.05$

$= 0.40$

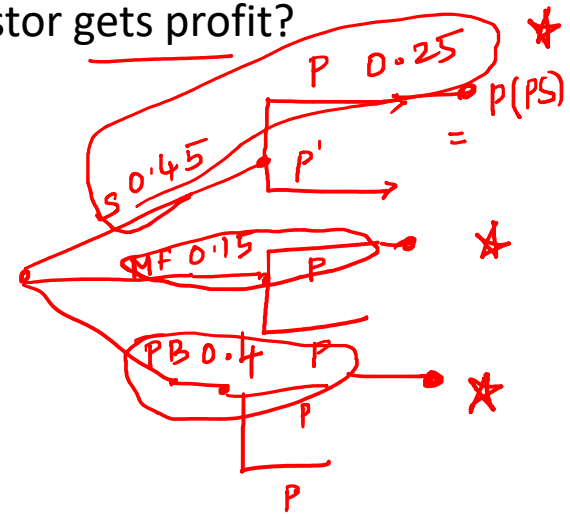
$P(F' \mid A) = \frac{0.40}{0.45} = 0.88$

100 %

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

Total Probability

An investor made 45% of investment in stocks, 15% in mutual funds and the rest in his personal business. There are 25%, 10% and 20% of chances of obtaining profit in stocks, mutual funds and personal business respectively. What is the probability that the investor gets profit?



$P(P/S) = 0.25$

$P(P/MF) = 0.10$

$P(P/PB) = 0.2$

$P(\text{Profit})$

$= P(PS \text{ or } PMF \text{ or } PPB)$

$= P(PS) \ldots$

$P(S) = 0.45$

$P(MF) = 0.15$

$P(PB) = 0.4$

$P(PS) = P(S)\,P(P/S)$

An investor made 45% of investment in stocks, 15% in mutual funds and the rest in his personal business. There are 25%, 10% and 20% of chances of obtaining profit in stocks, mutual funds and personal business respectively.

a) What is the probability that the investor gets profit.

b) Given that the investor has got some profit, what are the probabilities that he got it from stocks, mutual funds and business.

Solution Q2:

Given $P(S) = 45\% = 0.45$, $P(M) = 15\% = 0.15$, $P(B) = 0.4$

$P(P/S) = 0.25$, $P(P/M) = 0.1$, $P(P/B) = 0.2$

a) $P(P) = P(S)P(P/s)+P(M)P(P/M)+P(B)P(P/B) = 0.2075$

b) $P(S/P) = P(S)P(P/s)/P(P) = 0.542169$

$P(M/P) = P(M)P(P/M)/P(P) = 0.072289$

$P(B/P) = P(B)P(P/B)/P(P) = 0.385542$

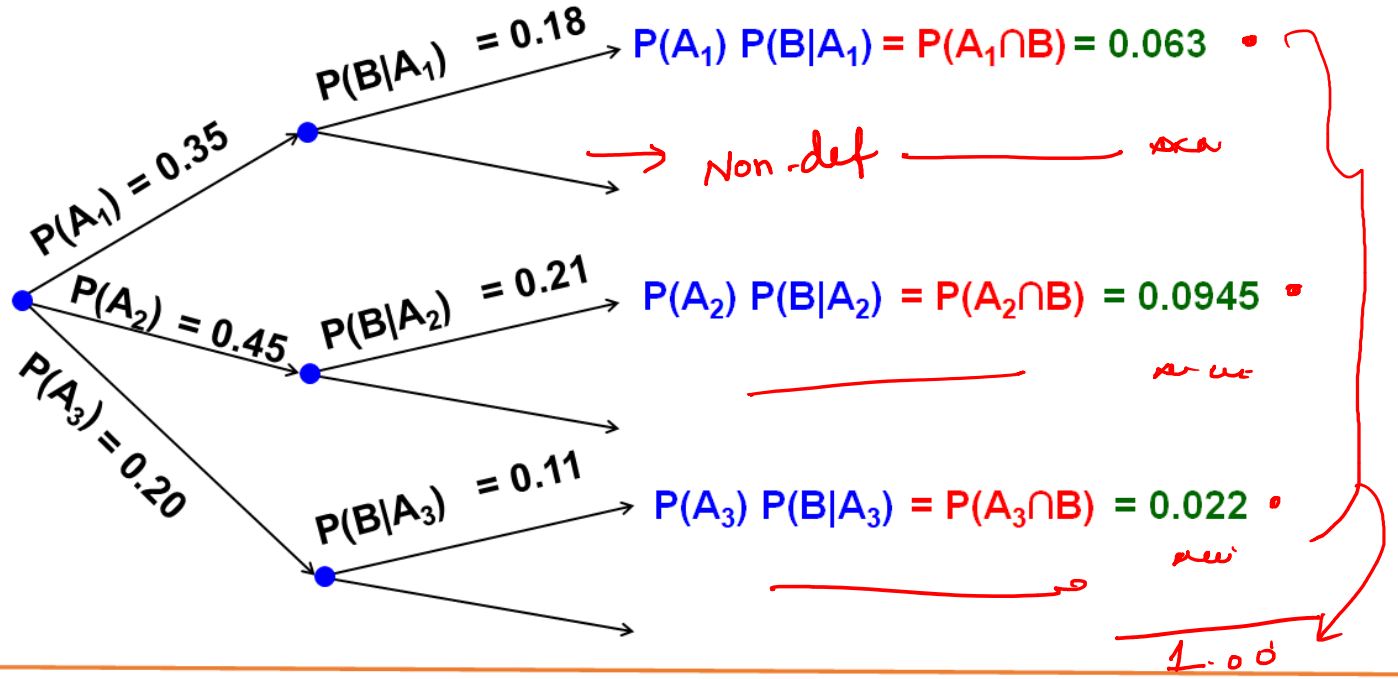
$P(X \cap Y) = P(X/Y)P(Y)$

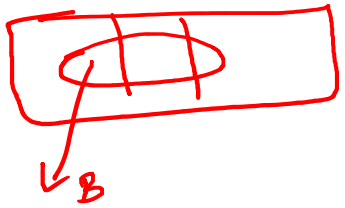
$= P(Y/X)P(X)$

$P(X/Y) = \frac{P(X \cap Y)}{P(Y)}$

Total Probability

Of 1000 car parts produced, it is known that 350 are produced in one plant, 450 parts in a second plant, and 200 parts in a third plant. Also, it is known that the probabilities are 0.18, 0.21, and 0.11 that the parts will be defective if they are produced in the first, second and third plants respectively. What is the probability that a randomly picked part from this batch is defective?





$P(B) = P(BA_1 \text{ or } BA_2 \text{ or } BA_3)$

$= P(BA_1) + \cdots$

$= P(A_1) \times P(B/A_1)$

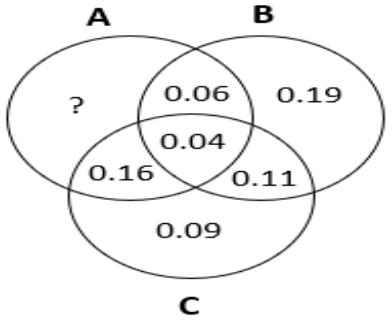
$=$

$$P(B) = P(A_1)\,P(B|A_1) + P(A_2)\,P(B|A_2) + P(A_3)\,P(B|A_3) = P(A_1 \cap B) + P(A_2 \cap B) + P(A_2 \cap B) = 0.1795$$

An insurance company insured 1,000 taxi drivers, 2,000 car drivers and 3,000 truck drivers. The probability of their accident is 0.05, 0.15 and 0.1 respectively. One of the insured persons meets with an accident. What is the probability that he is a truck driver?

Exercise

1. If P (A) =1/2, P (B) =1/3, P (A/B) = 1/6 find i). P(B/A) ii). P(B/A') iii).P(AUB / A) iv).P(B/A) .
2. There are three events A, B and C. The probability of occurrence of at least one of them is 0.23. Using the probabilities given in the following Venn diagram, find the probability of the event A.



3. A manufacturing company produces certain types of output by 4 machines i.e A, B, C and D. Machine A produces 15%, Machine B produces 30 % and Machine C produces 30% of daily production. Based on experience it is observed that 1% of the output by Machine A is defective. Similarly, the defectives by other machines are 2% ,3% and 4% respectively. An item is drawn at random and found to be defective. Is it possible to find the defective item is produced bv which Machine? If so, find it.

Exercise

4. A manufacturer has three machine operators A, B and C. The first operator A produce 1% defective items, whereas the other two operators B and C produce 5% and 7% defective items respectively. A is on the job for 50% of the time, B is on the job for 30% of the time. A defective item is produced, what is the probability that it was produced by A, B, C? Based on this write your observations.

5. Consider the following data related to the employees, who are on travel. 40% check work email, 20% use cell phone to stay connected to work, 25% bring laptop with them, 23% check both work email and use cell phone to stay connected, and 50% neither check work email nor use a cell phone to stay connected nor bring a laptop. In addition, 88 out of every 100 who bring a laptop also check work email, and 70 out of every 100 who use a cell phone to stay connected also bring a laptop.

i) What is the probability that a randomly selected traveller who checks work email also uses a cell phone to stay connected?

ii) What is the probability that someone who brings a laptop on vacation also uses a cell phone to stay connected?

iii) If the randomly selected traveler checked work email and brought a laptop, what is the probability that he/she uses a cell phone to stay connected?

THANK YOU!

Introduction to Statistics

BITS Pilani
Pilani Campus

innovate achieve lead

BITS Pilani
Pilani Campus

Course No: DSECL ZC413
Course Title: ISM
WEBINAR 2

Topics - Webinar

- Bayes' and Naïve Bayes Classifiers
- Discrete Random Variable
- Continuous Random Variable
- Probability distribution and density function
- Mean Variance and Standard deviation
- Joint Probability function
- Binomial Distribution

Topics - Webinar

- Poisson Distribution
- Normal Distribution
- Uniform Distribution

Conditional Probability

20 $\quad 50$ $\quad {}^{50}C_1 = \frac{50}{50} = 1$

Example: Suppose three coins tossed simultaneously. Let E and F be two events which are defined as follows:

E- At least Two heads appear.

F-First Coin shows tail.

a) Find $P(E)$. b) If it is given that the first coin shows tail then what is $P(E)$?

Solution Let S be a sample space.

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

$n(S) = 8$,

$E = \{HHH, HHT, THH, HTH\}$

$F = \{TTT, THH, THT, TTH\}$

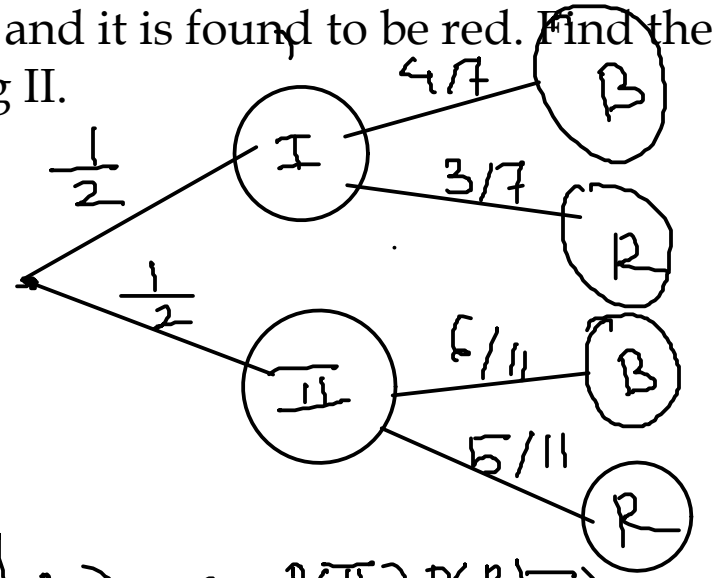
(a) $P(E) = \frac{n(E)}{n(S)} = \frac{4}{8} = \frac{1}{2}$.

(b) $P(E|F) = \frac{n(E \cap F)}{n(F)} = \frac{P(E \cap F)}{P(F)} = \frac{1}{4}$.

Bayes' Theorem

Example: Bag I contains 3 red and 4 black balls and while another Bag II contains 5 red and 6 black balls. One ball is drawn at random from one of the bags and it is found to be red. Find the probability that it was drawn from Bag II.

Solution

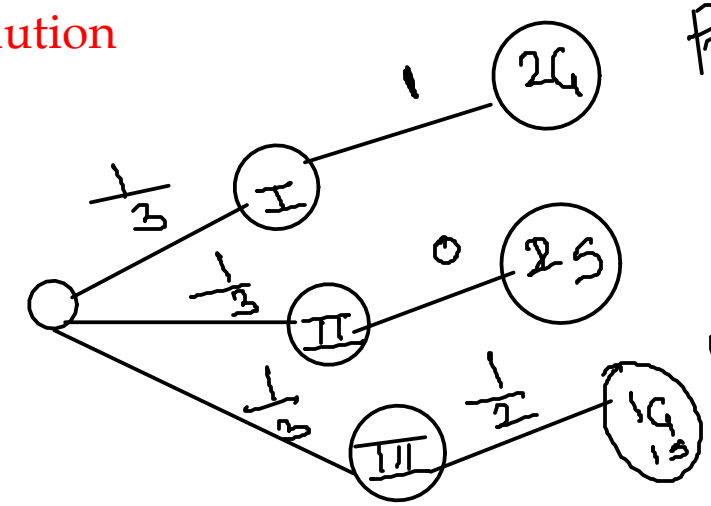


$$P(II \mid R) = \frac{P(II)\,P(R \mid II)}{P(R)} = \frac{\frac{1}{2} \times \frac{5}{11}}{\frac{1}{2} \times \frac{3}{7} + \frac{1}{2} \times \frac{5}{11}}$$

Bayes' Theorem

Example: Given three identical boxes I,II and III, each containing two coins . In box I both coins are gold coins, in box II, both are silver coins and in the box III there is one gold and one silver coin. A person chooses a box at random and takes out a coin. If the coin is gold, what is the probability that the other coin in the box is also of gold.

Solution



Prob of selecting gold coin

$$= P(G) = \frac{1}{3}\times 1 + \frac{1}{3}\times 0 + \frac{1}{3}\times\frac{1}{2}$$

$$= \frac{1}{3} + 0 + \frac{1}{6} = \frac{3}{6} = \frac{1}{2}.$$

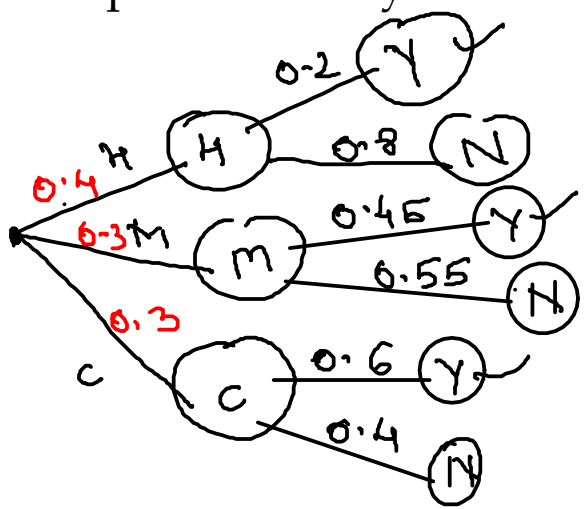
From getting other coin also gold, we have to select first coin from bag I.

$$\therefore \text{Required prob} = P(I|G) = \frac{P(I)\,P(G|I)}{P(G)} = \frac{\frac{1}{3}\times 1}{1/2} = \frac{2}{3}.$$

Bayes' Theorem

Example : From the past record it is known that the chance of hot temperature , mild and cool temperature on a particular day is 0.4,0.3 and 0.3 respectively. The chance of playing golf is depending on temperature of that day. If temperature is hot then chance of playing golf is 0.2, for mild temperature chance is 0.45 and for cool temperature chance is 0.6. If the team plays golf on a particular day then find the probability that the day was hot.

Solution



$0.08 + 0.135 + 0.18$

Prob of playing golf game

$$P(Y) = (0.4)\times(0.2) + (0.3)(0.45) + (0.3)(0.6) = 0.395$$

$$P(H|Y) = \frac{P(H)\,P(Y|H)}{P(Y)} = \frac{0.4\times 0.2}{0.395} = \frac{0.08}{0.395}$$

Naive Bayes Classifiers

Consider a fictional dataset that describes the weather conditions for playing a game of golf. Given the weather conditions, each tuple classifies the conditions as fit("Yes") or unfit("No") for playing golf.

	Outlook	Temperature	Humidity	Windy	Play Golf
0	Rainy	Hot	High	False	No
1	Rainy	Hot	High	True	No
2	Overcast	Hot	High	False	Yes
3	Sunny	Mild	High	False	Yes
4	Sunny	Cool	Normal	False	Yes
5	Sunny	Cool	Normal	True	No
6	Overcast	Cool	Normal	True	Yes
7	Rainy	Mild	High	False	No
8	Rainy	Cool	Normal	False	Yes
9	Sunny	Mild	Normal	False	Yes
10	Rainy	Mild	Normal	True	Yes
11	Overcast	Mild	High	True	Yes
12	Overcast	Hot	Normal	False	Yes
13	Sunny	Mild	High	True	No

Naive Bayes Classifiers

If Given that the weather conditions are "Rainy outlook", "Temperature is hot", "high humidity" and "no wind" then predict that the golf match will be played or not under such condition.

Solution

$$P(Y / R, H, HH, F)$$

$$= P(R/Y)\, P(H/Y)\, P(HH/Y)\, P(F/Y) \cdot P(Y)$$

$$P(N / R, H, HH, F) = P(R/N)\, P(H/N) \cdots P(N)$$

Naive Bayes Classifiers

$P(H/Y) = \frac{P(H \cap Y)}{9} = \frac{2}{9}$

$E = \{2, 4, 6\}$

$F = \{1, 3, 5\}$

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 $\frac{2}{9}$	2/5
Total	9	5	100%	100%

$= \frac{P(R \cap Y)}{9}$

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%

$E \cap F = \phi.$

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%

$\{HH, HT, TH, TT\}$

$E = \{HH\}$

$F = \{TT\}$

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

$E \cap F = \phi.$

Naive Bayes Classifiers

Suppose today = (Sunny, Hot, Normal, False)

So, probability of playing golf is given by:

$$P(Yes|today) = \frac{P(SunnyOutlook|Yes)P(HotTemperature|Yes)P(NormalHumidity|Yes)P(NoWind|Yes)P(Yes)}{P(today)}$$

and probability to not play golf is given by:

$$P(No|today) = \frac{P(SunnyOutlook|No)P(HotTemperature|No)P(NormalHumidity|No)P(NoWind|No)P(No)}{P(today)}$$

Since, P(today) is common in both probabilities, we can ignore P(today) and find proportional probabilities as:

$$P(Yes|today) \propto \frac{2}{9}.\frac{2}{9}.\frac{6}{9}.\frac{6}{9}.\frac{9}{14} \approx 0.0141$$

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

And

$$P(No|today) \propto \frac{3}{5}.\frac{2}{5}.\frac{1}{5}.\frac{2}{5}.\frac{5}{14} \approx 0.0068$$

Now, since

$$P(Yes|today) + P(No|today) = 1$$

These numbers can be converted Into a probability by making the sum equal to 1 (normalization):

$$P(Yes|today) = \frac{0.0141}{0.0141+0.0068} = 0.67 \quad \text{and} \quad P(No|today) = \frac{0.0068}{0.0141+0.0068} = 0.33$$

Since $P(Yes|today) > P(No|today)$.

So, prediction that golf would be played is 'Yes'.

Discrete Random Variable & their Probability Distribution

1. Suppose a random variable X takes on the values $-4,2,3,7$ with respective probabilities $\frac{k+2}{10}, \frac{2k-3}{10}, \frac{3k-4}{10}, \frac{k+1}{10}$. Find the distribution and expected value of X.

$\Rightarrow$ Here $f(-4) = \frac{k+2}{10}$, $f(2) = \frac{2k-3}{10}$, $f(3) = \frac{3k-4}{10}$, $f(7) = \frac{k+1}{10}$

since $f(x)$ is prob mass function

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

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

$$\frac{7k-4}{10} = 1 \Rightarrow k = 2.$$

$\therefore$ Distribution is $f(-4) = \frac{4}{10}$, $f(2) = \frac{1}{10}$, $f(3) = \frac{2}{10}$, $f(7) = \frac{3}{10}$

The expected value of $X = \sum x f(x) = -4 \times \frac{4}{10} + 2 \cdot \frac{1}{10} + 3 \times \frac{2}{10} + 7 \times \frac{3}{10}$

$$= \frac{-16+2+6+21}{10} = \frac{13}{10} = 1.3$$

Discrete Random Variable & their Probability Distribution

2. Let X be a random variable with following distribution

x	1	3	4	5
$f(x)$	0.4	0.1	0.2	0.3

Find i) the mean ii) variance iii) standard deviation iv) $E(3X+2)$ v) $E(X^2)$ vi) $E(2^x)$

$\Rightarrow$ i) the mean $= \sum x f(x) = 1\times\frac{4}{10} + 3\times\frac{1}{10} + 4\times\frac{2}{10} + 5\times\frac{3}{10}$

$= \frac{4+3+8+15}{10} = \frac{30}{10} = 3.$

ii) variance $= \sigma^2 = E(x^2) - (E(x))^2$ —— (1)

$E(x^2) = \sum x^2 f(x) = 1\times\frac{4}{10} + 9\times\frac{1}{10} + 16\times\frac{2}{10} + 25\times\frac{3}{10}$

$= \frac{4+9+32+75}{10} = 12$

$\therefore \sigma^2 = 12 - (3)^2 = 12 - 9 = 3$

iii) $S.D = \sqrt{3}$.

iv) $E(3x+2) = 3E(x)+2 = 3\times3+2 = 11.$

v) $E(x^2) = 12$

vi) $E(2^x) = \sum 2^x f(x) = 2^1\times\frac{4}{10} + 2^3\times\frac{1}{10} + 2^4\times\frac{2}{10} + 2^5\times\frac{3}{10}$

$= \frac{8}{10} + \frac{8}{10} + \frac{32}{10} + \frac{96}{10} = \frac{144}{10}.$

$32 \times 3 = 96$, $96 + 48 = 144$

Discrete Random Variable & their Probability Distribution

3. A fair coin is tossed until a head or five tails occur. Find the expected number of tosses of the coin.

$\Rightarrow$ Sample Space $= \{H, TH, TTH, TTTH, TTTTH, TTTTT\}$

X be a random which counts number of tosses

$\therefore$ X assigns $\{1, 2, 3, 4, 5\}$ values to sample space.

The pmf $f(x)$ is, $f(1) = \frac{1}{2}$, $f(2) = \frac{1}{4}$, $f(3) = \frac{1}{8}$, $f(4) = \frac{1}{16}$

$$f(5) = \frac{1}{32} + \frac{1}{32} = \frac{1}{16}.$$

$\therefore$ Expected number of tosses $= \sum x f(x)$

$$= 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} + 3 \cdot \frac{1}{8} + 4 \cdot \frac{1}{16} + 5 \cdot \frac{1}{16}$$

$$= \frac{8 + 8 + 6 + 4 + 5}{16} = \frac{31}{16} \approx 2$$

Discrete Random Variable & their Probability Distribution

4. A box contains 8 light bulbs of which 3 are defective. A bulb is selected from the box and tested. If it is defective, another bulb is selected and tested, until a non-defective bulb is chosen. Find the expected number of bulbs chosen.

$\Rightarrow$ If we write D for defective and N for non-defective then sample space S has four elements.

$$S = \{N, DN, DDN, DDDN\}$$

Let x be a random variable which represent the number of bulbs chosen.

$\therefore$ x assigns $\{1, 2, 3, 4\}$ values to sample points.

and pmf is, $f(1) = \frac{5}{8}$, $f(2) = \frac{3}{8} \cdot \frac{5}{7} = \frac{15}{56}$, $f(3) = \frac{3}{8} \cdot \frac{2}{7} \times \frac{5}{6} = \frac{5}{56}$

$$f(4) = \frac{3}{8} \cdot \frac{2}{7} \cdot \frac{1}{6} \cdot \frac{5}{5} = \frac{1}{56}$$

$$\therefore E(x) = \sum x f(x) = 1 \cdot \frac{5}{8} + 2 \cdot \frac{15}{56} + 3 \cdot \frac{5}{56} + 4 \cdot \frac{1}{56} = \frac{3}{2}$$

Discrete Random Variable & their Probability Distribution

5. A game is played by two person A and B as follows:

"A throws two dice. If the sum is 7 he wins \$3 from B. If the sum is 8 he loses \$2 and if sum is 3 he loses \$4, otherwise no money changes hand. Is this a fair game?"

$\Rightarrow$ Let x be a random variable which represent the gain by person A.

$\therefore x$ assigns $\{0, 3, -2, -4\}$ to sample space.

Let f be prob distribution function.

$\therefore f(3) = P\{$getting sum 7 after throwing two dice$\}$

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

$f(-2) = P\{$getting sum 8$\}$

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

$f(-4) = P\{$getting sum 3$\}$

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

$f(0) = P\{$getting sum other than 7, 8, and 3$\} = \frac{23}{36}$.

$\therefore E(x) = 0 \times \frac{23}{36} + 3 \times \frac{6}{36} - 2 \times \frac{5}{36} - 4 \times \frac{2}{36} = 0$

$\therefore$ Game is fair.

Discrete Random Variable & their Probability Distribution

6. A linear array EMPLOYEE has n elements. Suppose NAME appears randomly in the array, and there is a linear search to find the location K of NAME, that is, to find K such that EMPLOYEE[K]=NAME. Let X denote the number of comparisons in the linear search. Find the expected value of X.

$\Rightarrow$ Let X denote number of comparisons. Since NAME appears randomly in any position in the array with the same prob $\frac{1}{n}$. we have $X = 1, 2, 3, \ldots n$ each probability with $1/n$.

Hence $E(x) = 1 \cdot \frac{1}{n} + 2 \cdot \frac{1}{n} + \cdots + n \cdot \frac{1}{n} = \frac{n+1}{2}$

Continuous Random Variable & their Probability Distribution

7. Let X be random variable with probability density function $f(x) = \frac{1}{2}e^{-|x|}$ for all $x \in \mathbb{R}$. If $Y = X^2$ then find cumulative distribution function of Y.

$$\Rightarrow \quad F(Y) = P(Y \leq y) = P(X^2 \leq y) = P(-\sqrt{y} \leq x \leq \sqrt{y})$$

$$= \int_{-\sqrt{y}}^{\sqrt{y}} \frac{1}{2} e^{-|x|} dx$$

But $e^{-|x|}$ is an even function

$$\therefore F(Y) = 2\int_0^{\sqrt{y}} \frac{1}{2} e^{-x} dx = \int_0^{\sqrt{y}} e^{-x} dx = -e^{-x}\Big|_0^{\sqrt{y}}$$

$$= 1 - e^{-\sqrt{y}}$$

$$\therefore F(Y) = \begin{cases} 1 - e^{-\sqrt{y}} & y \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

Continuous Random Variable & their Probability Distribution

8. Let X be random variable with probability density function $f(x) =$

$$\begin{cases} x^2\left(2x+\frac{3}{2}\right) & 0 < x \le 1 \\ 0 & otherwise \end{cases}$$

If $y = \frac{2}{x} + 3$ then find $var(y)$.

$$\Rightarrow \quad Var(Y) = 4\,Var\left(\frac{1}{x}\right) = 4\left\{E(1/x^2) - (E(1/x))^2\right\}$$

$$E(1/x^2) = \int_{-\infty}^{\infty} \frac{1}{x^2} f(x)\,dx = \int_0^1 \frac{1}{\cancel{x^2}} \cancel{x^2}\left(2x+\frac{3}{2}\right)dx$$

$$= \frac{2x^2}{2} + \frac{3}{2}x \Big|_0^1 = \frac{5}{2}.$$

$$E(1/x) = \int_0^1 \frac{1}{x} x^2\left(2x+\frac{3}{2}\right)dx$$

$$= \int_0^1 \left(2x^2 + \frac{3}{2}x\right)dx = \frac{2x^3}{3} + \frac{3}{2}\frac{x^2}{2}\Big|_0^1$$

$$= \frac{2}{3} + \frac{3}{4}\Big|_0^1 = \frac{14}{12} = \frac{7}{6}$$

$$\therefore Var(Y) = 4\left\{\frac{5}{2} - \left(\frac{7}{6}\right)^2\right\} = 4\left\{\frac{5}{2} - \frac{49}{36}\right\}.$$

Continuous Random Variable & their Probability Distribution

9. The distribution of amount of gravel (in tons) sold by a particular construction supply company in a given week is a continuous random variable X with probability density function

$f(x) = \begin{cases} \frac{3}{2}(1-x^2) & 0 \le x \le 1 \\ 0 & otherwise \end{cases}$. What is the cumulative distributive function of sales for any X? How do you use this to find the probability that $X < 0.25$? What about the probability that X is greater than 0.75? What about $p(0.25 < X < 0.75)$.

$\Rightarrow$ Let $F(x)$ be cdf, $\therefore F(x) = P(X \le x) = \int_{-\infty}^{x} f(x)\,dx$

If $x < 0$, then $F(x) = \int_{-\infty}^{x} f(x)\,dx = \int_{-\infty}^{x} 0\,dx = 0.$

If $0 \le x < 1$ then $F(x) = \int_{-\infty}^{x} f(x)\,dx = \int_{-\infty}^{0} 0\,dx + \int_{0}^{x} \frac{3}{2}(1-x^2)\,dx$

$= \frac{3}{2}\left(x - \frac{x^3}{3}\Big|_0^x\right) = \frac{3}{2}\left(x - \frac{x^3}{3}\right)$ —— (1)

If $x \geq 1$ then $F(x) = \int_{-\infty}^{x} f(x)\,dx = \int_{-\infty}^{0} 0\,dx + \int_{0}^{1} \frac{3}{2}(1-x^2)\,dx + \int_{1}^{x} 0\,dx$

$= 1.$

(b) $P(X < 0.25) = F(0.25) = \frac{3}{2}\left(0.25 - \frac{(0.25)^3}{3}\right)$ from (1)

(c) $P(X > 0.75) = \int_{0.75}^{\infty} f(x)\,dx = \int_{3/4}^{1} \frac{3}{2}(1-x^2)\,dx$

$$= \frac{3}{2}\left(x - \frac{x^3}{3}\Big|_{3/4}^{1}\right)$$

$$= \frac{3}{2}\left\{1 - \frac{1}{3} - \frac{3}{4} + \frac{(3/4)^3}{3}\right\}$$

$$= \frac{3}{2}\left\{\frac{2}{3} - \frac{3}{4} + \frac{27}{64 \times 3}\right\}.$$

(d) $P(0.25 < X < 0.75) = \int_{0.25}^{0.75} f(x)\,dx = \int_{1/4}^{3/4} \frac{3}{2}(1-x^2)\,dx$

$$= \frac{3}{2}\left(x - \frac{x^3}{3}\Big|_{1/4}^{3/4}\right) = \frac{3}{2}\left\{\frac{3}{4} - \frac{1}{3}\left(\frac{3}{4}\right)^3 - \frac{1}{4} + \frac{1}{3}\left(\frac{1}{4}\right)^3\right\}$$

$$= \frac{3}{2}\left\{\frac{3}{4} - \frac{9}{64} - \frac{1}{4} + \frac{1}{64 \times 3}\right\}$$

Continuous Random Variable & their Probability Distribution

10. If cumulative distribution function is $F(x) = \begin{cases} 0 & x \le 0 \\ (1-e^{-x})^2 & x > 0 \end{cases}$. Find probability density function and $p(1 < X < 2)$.

$\Rightarrow$ Let $f(x)$ be density function of x.

$$\therefore f(x) = \frac{d}{dx} F(x) = \begin{cases} 0 & x \le 0 \\ 2(1-e^{-x})e^{-x} & x > 0 \end{cases}$$

$$= \begin{cases} 0 & x \le 0 \\ 2(e^{-x} - e^{-2x}) & x > 0 \end{cases}$$

$$\therefore p(1 < x < 2) = 2\int_1^2 (e^{-x} - e^{-2x})\,dx = 2\left\{-e^{-x} + \frac{e^{-2x}}{2}\Big|_1^2\right\}$$

$$= 2\left\{-e^{-2} + \frac{e^{-4}}{2} + e^{-1} - \frac{e^{-2}}{2}\right\}$$

$$= 2\left\{-\frac{3e^{-2}}{2} + \frac{e^{-4}}{2} + e^{-1}\right\}$$

Joint Probability Distribution

11. Let X denote the number of times a photocopy machine will malfunction: 0, 1, 2, or 3 times, on any given month. Let Y denote the number of times a technician is called on an emergency call. The joint p.m.f. $p(x, y)$ is presented in the table below:

X

Y

	0	1	2	3
0	0.15	0.30	0.05	0
1	0.05	0.15	0.05	0.05
2	0	0.05	0.10	0.05

Find a) $P(Y > X)$ b) Marginal probability function of X and Y. c) $P(X \leq 2, Y \geq 1)$.

(a) $p(Y>X) = p(0,1) + p(0,2) + p(1,2)$

$= 0.05 + 0 + 0.05 = 0.1$

(b) Marginal prob function of $X = f_1(x) = \sum_y f(x,y)$

$\therefore f_1(0) = f(0,0) + f(0,1) + f(0,2) = 0.15 + 0.05 + 0 = 0.2$

$f_1(1) = f(1,0) + f(1,1) + f(1,2) = 0.30 + 0.15 + 0.05 = 0.5$

$f_1(2) = f(2,0) + f(2,1) + f(2,2) = 0.05 + 0.05 + 0.10 = 0.2$

$f_1(3) = f(3,0) + f(3,1) + f(3,2) = 0 + 0.05 + 0.05 = 0.1$

Marginal prob function of $Y = f_2(y) = \sum_x f(x,y)$.

$f_2(0) = f(0,0) + f(1,0) + f(2,0) + f(3,0) = 0.15 + 0.30 + 0.05 + 0 = 0.5$

$f_2(1) = f(0,1) + f(1,1) + f(2,1) + f(3,1) = 0.05 + 0.15 + 0.05 + 0.05 = 0.3$

$f_2(2) = f(0,2) + f(1,2) + f(2,2) + f(3,2) = 0 + 0.05 + 0.10 + 0.05 = 0.2$

(c) $P(X \leq 2, Y \geq 1) = f(0,1) + f(0,2) + f(1,1) + f(1,2) + f(2,1) + f(2,2)$

$= 0.30 + 0.05 + 0.15 + 0.05 + 0.05 + 0.10$

$= 0.7$

Joint Probability Distribution

12. Let the joint density function for (X, Y) be $f(x, y) = \begin{cases} \frac{c(x+y)}{3} & 0 < x < 2, 0 < y < 1 \\ 0 & ohrtwise \end{cases}$

Find a) the constant c b) $P(X > Y)$ c) Marginal density function of X and Y.

d) Are X and Y are independent?

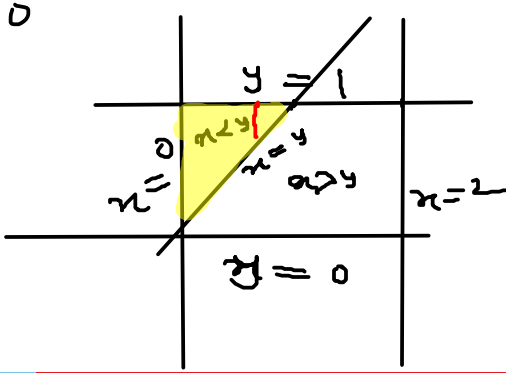
(a) since $f(x,y)$ is a density function

$$\therefore \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx = 1 \;;\; \int_0^2\int_0^1 \frac{c(x+y)}{3}\,dy\,dx = 1$$

$$\therefore \frac{c}{3}\int_0^2 \left.xy + \frac{y^2}{2}\right|_0^1 dx = 1 \Rightarrow \frac{c}{3}\int_0^2 \left(x + \frac{1}{2}\right) dx = 1$$

$$\Rightarrow \frac{c}{3}\left\{\frac{x^2}{2} + \frac{x}{2}\right|_0^2\right\} = 1$$

$$\Rightarrow \frac{c}{3}\left\{\frac{4}{2} + \frac{2}{2}\right\} = 1 \Rightarrow c = 1.$$



(b) $P(X > Y) = 1 - P(X \le Y)$

$$= 1 - \int_0^1\int_x^1 \frac{x+y}{3}\,dy\,dx$$

$$= 1 - \frac{1}{3}\int_0^1 xy + \frac{y^2}{2}\Big|_x^1 dx = 1 - \frac{1}{3}\int_0^1 x + \frac{1}{2} - \frac{3x^2}{2}\, dx$$

$$= 1 - \frac{1}{3}\left\{ \frac{x^2}{2} + \frac{x}{2} - \frac{3x^3}{6} \Big|_0^1 \right\}$$

$$= 1 - \frac{1}{3}\left\{ \frac{1}{2} + \cancel{\frac{1}{2}} - \cancel{\frac{1}{2}} \right\}$$

$$= 1 - \frac{1}{6} = \frac{5}{6}.$$

marginal density function of $X = f_1(x) = \int_{-\infty}^{\infty} f(x,y)\, dy$

$$= \int_0^1 \frac{x+y}{3}\, dy$$

$$= \frac{1}{3}\left\{ xy + \frac{y^2}{2} \Big|_0^1 \right\} = \begin{cases} \frac{1}{3}\left(x + \frac{1}{2}\right) & 1 \le x \le 2 \\ 0 & \text{otherwise} \end{cases}$$

marginal density function of $Y = f_2(y)$

$$= \int_{-\infty}^{\infty} f(x,y)\, dx = \frac{1}{3}\int_0^2 x + y\, dx$$

$$= \frac{1}{3}\left\{ \frac{x^2}{2} + xy \Big|_0^2 \right\} = \begin{cases} \frac{1}{3}(2 + 2y) & 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

(d) consider $f_1(x) \cdot f_2(y) = \frac{1}{3}\left(\frac{2x+1}{2}\right) \cdot \frac{(2+2y)}{3}$

$$\neq \frac{x+y}{3} = f(x,y)$$

$\therefore$ X and Y are not independent.

Joint Probability Distribution

13. Dick and Jane have agreed to meet for lunch between noon (0:00 p.m.) and 1:00 p.m. Denote Jane's arrival time by X, Dick's by Y, and suppose X and Y are independent with probability density functions $f_X(x) = \begin{cases} 3x^2 & 0 \le x \le 1 \\ 0 & otherwise \end{cases}$, $f_Y(y) = \begin{cases} 2y & 0 \le y \le 1 \\ 0 & otherwise \end{cases}$

Find the probability that Jane arrives before Dick.

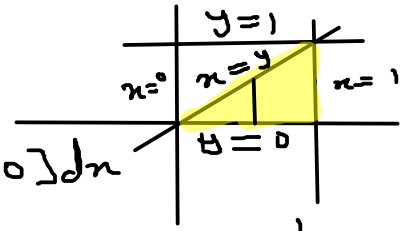
$\Rightarrow$ Here it is given that X and Y are independent

$$\therefore f(x,y) = f_X(x) f_Y(y)$$

$$= \begin{cases} 6x^2y & 0 \le x \le 1,\ 0 \le y \le 1 \\ 0 & \text{otherwise} \end{cases}$$

$$P(\text{Jane arrives before Dick}) = P(X<Y)$$

$$= \int_0^1 \int_0^x 6x^2y \, dy \, dx$$



$$= \int_0^1 \frac{6x^2y^2}{2} \Big|_0^x dx = 3\int_0^1 x^2 \cdot [x^2 - 0] dx$$

$$= 3\int_0^1 x^4 dx = 3\left\{ \frac{x^5}{5} \Big|_0^1 \right\} = \frac{3}{5}.$$

Binomial Distribution

14. If from six to seven in the evening one telephone line in every five is engaged in a conversation: what is the probability that when 10 telephone numbers are chosen at random, only two are in use?

$\Rightarrow$ Let x be a random variables which represant the number of telephone lines in use.

This is Binomally distributed random variable because telephone line either in use or not in use.

Let the telephone line in use will be success.

$\therefore$ Prob (success) $= \frac{1}{5}$.

$n = 10$, $p(x=2) = \binom{10}{2}\left(\frac{1}{5}\right)^2\left(\frac{4}{5}\right)^8$.

Binomial Distribution

15. It has been determined that 5% of drivers checked at a road stop show traces of alcohol and 10% of drivers checked do not wear seat belts. In addition, it has been observed that the two infractions are independent from one another. If an officer stops five drivers at random:

a) Calculate the probability that exactly three of the drivers have committed any one of the two offenses.

b) Calculate the probability that at least one of the drivers checked has committed at least one of the two offenses.

$\Rightarrow$ Let A be an event that driver traces of alcohol.

B be an event that driver do not wear seat bell.

$P(A \cup B) = P(A) + P(B) - P(A \cap B) = (0.05) + 0.01 - (0.05)(0.01)$

$= 0.145.$

Here success is driver commited any one the two offenses.

$\therefore\ p(\text{success}) = 0.145.$

(a). $n = 5,\ p = 0.145,\ q = 0.855$

$$p(x=3) = {}^5C_3 (0.145)^3 (0.855)^2 = 0.0223$$

(b) $p(x \geq 1) = 1 - p(x=0) = 1 - {}^5C_0 (0.145)^0 (0.855)^5.$

$$= 0.543.$$

Binomial Distribution

16. A student takes an 18-question multiple choice exam, with four choices per question. Suppose one of the choice is obviously incorrect and the students make an "educated" guess of the remaining choices. Find the expected number E of correct answers and the standard deviation σ.

$\Rightarrow$ Let x be a random variable which count the number of correct answers.

Here success is giving correct answer to the question.

$p(\text{success}) = \frac{1}{3}$ {because one of the choice is obviously incorrect}.

As x is binomally distributed

$\therefore$ Expected value of $x = E(x) = np = 18 \times \frac{1}{3} = 6$

and $S.D = \sqrt{npq} = \sqrt{18 \times \frac{1}{3} \times \frac{2}{3}} = \sqrt{4} = 2$.

Poisson Distribution

17. Suppose 300 misprints are distributed randomly throughout a book of 500 pages. Find the probability P that a given page contains: (a) Exactly 2 misprints (b) 2 or more misprints.

$\Rightarrow$ We view the number of misprints on one page as the number of success in a seq of Bernoullis trials.

Here $n = 300$, prob of getting a mistake on a given page $= \frac{1}{500}$.

$\therefore \lambda = np = \frac{300}{500} = 0.6$.

As number of misprints are more, so we can solve this problem by Poisson distribution.

(a) $p(x=2) = \frac{(0.6)^2 e^{-0.6}}{2!} = 0.0988$.

(b) $p(x \geq 2) = 1 - \{p(0) + p(1)\}$

$= 1 - \left\{\frac{(0.6)^0 e^{-0.6}}{0!} + \frac{(0.6)^1 e^{-0.6}}{1!}\right\}$

$= 0.122$.

Poisson Distribution

$$f(x) = \frac{e^{-\lambda}\lambda^{x}}{x!}$$

18. The number of industrial injuries per working week in a particular factory is known to follow a Poisson distribution with mean 0.5. Find the probability that (a) in a particular week there will be: (i) less than 2 accidents, (ii) more than 2 accidents; (b) in a three-week period there will be no accidents.

$\Rightarrow$ Here $\lambda = 0.5$

Let x be a random variable which count number of accidents in industry which is Poisson ally distr

① $p(x<2) = p(x=0) + p(x=1)$

$$= \frac{e^{-0.5}(0.5)^0}{0!} + \frac{e^{-0.5}(0.5)^1}{1!}$$

② $p(x>2) = 1 - \{p(0) + p(1) + p(2)\}$

$$= 1 - \left\{\frac{e^{-0.5}(0.5)^0}{0!} + \frac{e^{-0.5}(0.5)^1}{1!} + \frac{e^{-0.5}(0.5)^2}{2!}\right\}$$

③ In a three week period, the avg of number of accidents $= 3 \times \lambda$

$$\therefore p(x=0) = \frac{e^{-(3\times\lambda)}(1.5)^0}{0!} = (e^{-\lambda})^3 = (e^{-0.5})^3.$$

Normal Distribution

19. Suppose the weights of 2000 male students are normally distributed with $\mu = 155lb$ and standard deviation $\sigma = 20\ lb$. Find the number of students with weights: a) not more than 100lb b) between 150 and 175 lb (inclusive) c) greater than or equal to 200lb.

$\Longrightarrow$ Let X be random variables which measure wt of students is normally distributed with

$\mu = 155, \quad \sigma = 20.$

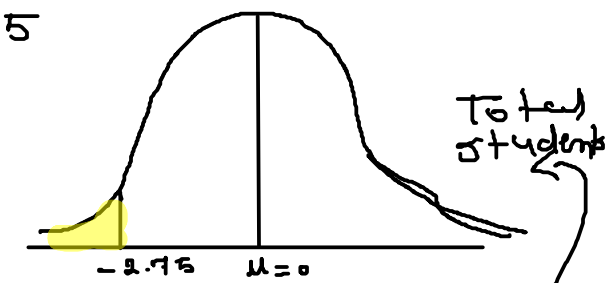
(a) $p(X \leq 100) = ?$

Let $z = \dfrac{100 - 155}{20} = \dfrac{-55}{20} = -2.75$

$\therefore p(X \leq 100) = p(z \leq -2.75)$

$= 0.5 - \phi(2.75)$

$= 0.5 - 0.4970 = 0.03$



$\therefore$ num of students ht not more than 100 lb $= 0.03 \times 2000$
$= 60.$

(b) $p(150 \leq X \leq 175)$

std unit of $150 = \dfrac{150 - 155}{20} = \dfrac{-5}{20} = -0.25.$

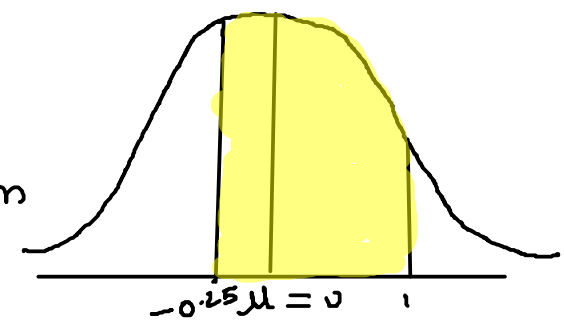
std unit of 175 $= \frac{175-155}{20} = 1$.

$\therefore P(150 \le x \le 175) = P(-0.25 \le z \le 1)$

$= \phi(0.25) + \phi(1)$

$= 0.0987 + 0.3413$

$= 0.44$.



$\therefore$ number of students with ht between 150 and 175 lb $= 0.44 \times 2000$

$= 880$.

(c) $P(x \geq 200) = ?$

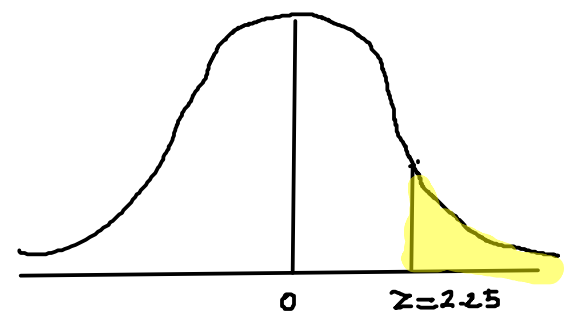
std unit of 200 $= \frac{200-155}{20} = \frac{45}{20} = 2.25$.

$P(x \geq 200) = P(z \geq 2.25)$

$= 0.5 - \phi(2.25)$

$= 0.5 - 0.4878$

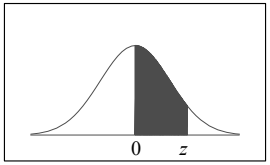
$= 0.0122$.



$\therefore$ Number of students with ht more than 200 lb $= 2000 \times 0.0122$

$= 24.4 \approx 24$

Standard Normal Distribution Table



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995

Gilles Cazelais. Typeset with LaTeX on April 20, 2006.

Normal Distribution

20. A fair die is tossed 720 times. Find the probability that the face 6 will occur: Between 100 and 125 times b) more than 130 times c)less than 110 times.

$\Rightarrow$ This is problem of Binomial distribution with large number of trials; and

mean $= \mu = np = 720 \times \frac{1}{6} = 120$, $\sigma = \sqrt{npq} = 10$

Here np is also large, so we can solve this problem by using normal distribution.

(a) $BP(100 \le x \le 125) = NP(99.5 \le x \le 125.5)$.

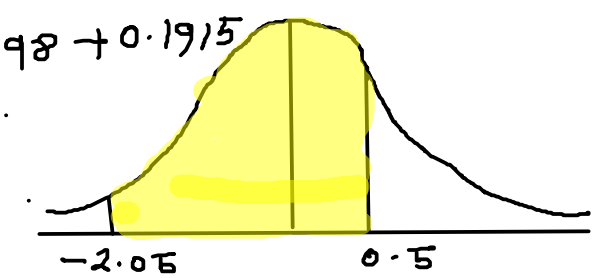
$\therefore$ std unit of $99.5 = \frac{99.5 - 120}{10} = -2.05$

std unit of $125 = \frac{125 - 120}{10} = 0.5$

$\therefore NP(99.5 \le x \le 125.5) = NP(-2.05 \le z \le 0.5)$

$= \phi(2.05) + \phi(0.5) = 0.4798 + 0.1915$

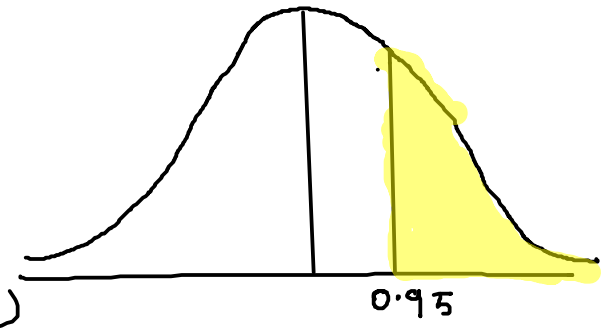
$= 0.6713$.



(b) $BP(X \geq 130) = NP(X \geq 129.5)$

std unit of $129.5 = \dfrac{129.5 - 120}{10}$

$= 0.95$



$NP(X \geq 129.5) = NP(Z \geq 0.95)$

$= 0.5 - \phi(0.95)$

$= 0.5 - 0.3289 = 0.1711.$

(c) $BP(X \leq 110) = NP(X \leq 110.5)$

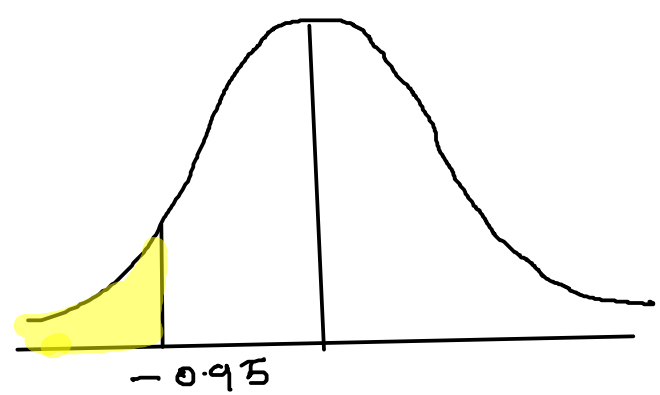
std unit of $110.5 = \dfrac{110.5 - 120}{10} = -0.95.$

$NP(X \leq 110.5) = NP(Z \leq -0.95)$

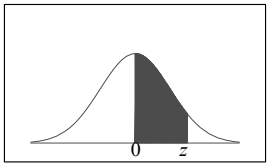
$= 0.5 - \phi(0.95)$

$= 0.5 - 0.3289$

$= 0.1711.$



Standard Normal Distribution Table



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4997	.4997

Gilles Cazelais. Typeset with LaTeX on April 20, 2006.

Thank you

IMP Note to Self

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
Work Integrated Learning Programmes Division

Cluster Programme - M. Tech(AIML)
I Semester , 2022 – 23(January,2023)
Mid semester Examination (**Regular**)

Course No : AIMLCL ZC418
Course Title : Introduction to Statistical Methods
Nature of Exam. : Open Book (Online)
Weightage : 30 Marks
Duration : 120 minutes
Date : 8th January,2023

Number of questions:5
Number of Pages: 2

Q1. .a). Consider the following data from the industry. Let D be the set of software developers in a company .Some subsets of D and some of probability that a random employee will belong to the set are represented below. Find the missing term in the table with justification. **[3M]**
Try to identify inconsistencies also (if exists and validate them).

Set in words	**Symbol**	**Probability**
All software developers	D	
All software testing engineers	T	
All software developers and / or software testing engineers		0.25
All software developers who are not software testing engineers		0.05
All software testing engineers who are not software developers		0.10
All software developers and software testing engineers		

b). A biased coin with probability of falling as heads 0.6 is tossed 20 times. Obtain the prob of getting 5 heads using Binomial , Poisson assumptions and compare. **[3M]**

Q2.a) Consider the following data. **[3M]**

A: the event that a patient selected is with high B.P

B: the event that a patient selected is diabetic

C: the event that a patient selected is with cancer

Events	A	B	C	A and B	B and C	A and C	(A and B and C)
Probability	0.35	0.40	0.45	0.22	0.25	0.22	0.15

i).If a patient is selected at random, what is the probability that he suffering from at least one of these health issues?

ii).What is the probability that the patient selected is with B.P given that he is with either of the remaining two diseases?

iii).What is the probability that the patient is diabetic given that he is suffering from at least one of the diseases mentioned?

b).A manufacturer has four machine operators A, B C and D. The first operator A produce 1% defective items, whereas the other two operators B, C and D produce 5%, 6% and 7% defective items respectively. A is on the job for 30% of the time, B is on the job for 30% of the time is on the job for 20% of the time and D is on the job for 20% of the time. A defective item is produced, what is the probability that it is produced by A? Suppose if B spends 15% of the time (instead of 30% as planned) on the job then what is the probability that the defective item is produced by him? **[3M]**

3. a).A manufacturer knows that on an average of 2% of the electric toasters that he makes will require repairs within 90 days after they are sold. Use the normal approximation to the binomial distribution to determine the normal approximation to the binomial distribution to determine the probability that among 1200 of these toasters at least 30 will require repairs within the first 90 days after they sold. **[3M]**

b). If the probability density function of a random variable X is given by

$$f(x) = k(1-x^2) \quad \text{for } 0 \le x \le 2$$

Find (i) the value of *k*

(ii) Probability P ($0.5 \le X \le 2$)

(iii) Probability at x = 2.

Q4.a). Suppose the random variable X and Y have joint density function defined by **[3M]**

$$f(x,y) = \begin{cases} k(2x+y), 2 < x < 5, 0 < y < 4 \\ 0 \; otherwise \end{cases}$$

Find (i) k (ii) P(X > 2,Y > 3) (iii) P(X > 4) (iv) P((X+Y) < 3)

b). The joint probability distribution of two random variables X and Y is given by **[3M]**

p(x,y)		x	
		2	4
y	1	0.1	0.15
	3	0.2	0.3

	5	0.1	0.15

i) Evaluate the marginal distributions of X and Y

ii) Find P(Y=3/X=4)

iii) Are they independent random variables?

Q5. a). Out of 1000 families with 5 children each how many families would be expected to have (i) 2 boys (ii) at least 1 boy (iii) at most 3 girls (iv) children of both sexes. Assuming equal probabilities for boys and girls **[3M]**

b).A sample of 100 battery cells tested to find the length of life produced. Assuming the data to be normally distributed with mean 12hrs and standard deviation 3hrs, what percentage of battery cells are expected to have life length of life **[3M]**

i).More than 15hrs ii) Less than 6 hrs iii) between 10 and 14 hrs

XXXXXX

BIRLA INSTITUTE OF TECHNOLOGY & SCIENCE, PILANI
Work Integrated Learning Programmes Division
Cluster Programme - M. Tech(AIML)
I Semester , 2022 – 23(January,2023)
Mid semester Examination (**Makeup**)

Course No : AIMLCZC418
Course Title : Introduction to Statistical Methods
Nature of Exam. : Open Book (Online)
Weightage : 30 Marks
Duration : 120 minutes
Date : 28th January,2023_FN

Number of questions:5
Number of Pages: 2

Q.1.a). Consider the following summary of a data. **[4M]**
Write important 4 observations from this summary which helps us to understand the data as a part of pre – processing.

	count	mean	std	min	25%	50%	75%	max
symboling	205.0	0.834146	1.245307	-2.00	0.00	1.00	2.00	3.00
wheel_base	205.0	98.756585	6.021776	86.60	94.50	97.00	102.40	120.90
length	205.0	174.049268	12.337289	141.10	166.30	173.20	183.10	208.10
width	205.0	65.907805	2.145204	60.30	64.10	65.50	66.90	72.30
height	205.0	53.724878	2.443522	47.80	52.00	54.10	55.50	59.80
curb_weight	205.0	2555.565854	520.680204	1488.00	2145.00	2414.00	2935.00	4066.00
engine_size	205.0	126.907317	41.642693	61.00	97.00	120.00	141.00	326.00
bore	205.0	3.329366	0.270858	2.54	3.15	3.31	3.58	3.94
stroke	205.0	3.256098	0.313634	2.07	3.11	3.29	3.41	4.17
compression_ratio	205.0	10.142537	3.972040	7.00	8.60	9.00	9.40	23.00
horsepower	205.0	104.165854	39.529733	48.00	70.00	95.00	116.00	288.00
peak_rpm	205.0	5126.097561	477.035772	4150.00	4800.00	5200.00	5500.00	6600.00
city_mpg	205.0	25.219512	6.542142	13.00	19.00	24.00	30.00	49.00
highway_mpg	205.0	30.751220	6.886443	16.00	25.00	30.00	34.00	54.00

b). Validate the following statement. Justify it. **[3M]**

"If two events are mutually exclusive, then they are independent also and vice versa"

Q.2.a).If P (A) =1/2, P (B) =1/3, P (A/B) = 1/6 find i). P(B/A) ii). P(B/A') iii).P(AUB / A) iv).P(B/A) v). $P\left(A \cap \overline{B}\right)$ vi). P $\left(\overline{A} \cap \overline{B}\right)$ vii). P $\left(\overline{A} \cap B\right)$ **[4M]**

b).Consider the following information related to road accidents.

Authorities identified that 15% of the road accidents are due to rash driving by self(A), 25%, 20%, 40% of road accidents are *due to bad roads(B)*, *road indiscipline by others(C)* and *various other reasons(D)* respectively. Out of these accidents (A, B, C and D) 1%, 2%, 3% and 4% died after treatment. Find the percentage of deaths due to accidents. **[3M]**

Q.3.a).A random variable X has the density function **[3M]**

$$f(x) = \begin{cases} ke^{-0.5\,x}, x \geq 0 \\ 0\ elsewhere \end{cases},$$

i).Determine k that renders f(x) as a valid density function.

ii).Find P (-1.5 < X < 2.5)

iii).Find P ([X] =1), where [X] is the greatest integer less than or equal to X

b). Consider the following joint distribution of x and y. **[4M]**

p(x,y)		x	
		1	2
y	1	0.15	0.15
	2	0.20	0.20
	3	0.05	0.10
	4	0.05	0.10

i) Validate the distribution

ii) Find Marginal distributions of x,y

iii) Find P(Y = 2/ X < 2)

iv) Are they independent random variables?

Q.4.a). If possible, find the probability distribution function of a binomial distribution with mean 4 and variance 4/3. **[3M]**

b). The life of a certain kind of electronic device has a mean of 250 hours and a standard deviation of 20 hours. Assuming that he distribution of life times which are measured to the nearest hour can be approximated closely with a normal curve, **[3M]**

(i) Find the probability that any one of these devices will have a lifetime of more than 200 hours.

(ii) What percentage will have life time from 250 to 300 hours.

Q.5. It is observed that waiting time of patients in a hospital is 15 minutes with a Standard deviation of 2 minutes. Sampling is used to validate this observation. Find the probability that the sampling mean lies between 12 to 17 minutes with sample size of 35**.** **[3M]**

XXXXXX