KARPAGAM ACADEMY OF HIGHER EDUCATION

(Deemed to be University)

(Established under Section 3 of UGC Act, 1956)

Pollachi Main Road, Coimbatore – 641 021, Tamil Nadu

Department of Mathematics

19BPU202	**Business Analytics**	**L T P C**
Semester – II		**5 2 0 5**

Course Objectives

To make the students

1. To understand the concept of matrices
2. To acquire the knowledge of differential calculus
3. To know the concepts of central tendency and dispersion
4. To understand the correlation and regression concepts
5. To be aware of the index numbers and trend analysis

Course Outcomes

Learners should be able to

1. Utilize the concept of matrices, differential calculus to solve business problems
2. Calculate and apply the measure of central tendency and dispersion in decision making.
3. Evaluate the relationship and association between variables to formulate the strategy in business.
4. Apply the concept of index numbers and trend analysis in business decisions.
5. Demonstrate capabilities as problem-solving, critical thinking, and communication skills related to the discipline of statistics.

Unit – I: Introduction to data science

Concepts of measurement, scales of measurement, Different types and scales of data (ratio, interval, nominal and ordinal);

Design of data collection formats with illustration, data quality and issues with date collection systems with examples from business, cleaning and treatment of missing data, principles of data visualization.

Data summarization and visualization methods : Histograms, Frequency distributions, Relative frequency, measures of central tendency and dispersion; Tables, Graphs, Charts, Box Plot; Chebychev's Inequality.

Unit – II: Probability and Sampling Estimation

Basic probability concepts, Conditional probability, Bayes Theorem, Probability distributions, Continuous and discrete distributions, Binomial Distribution, Uniform Distribution, Exponential Distribution, Normal distribution, Central Limit Theorem, Sequential decision making, Decision tree

Sampling and estimation: Estimation problems, Point and interval estimates, Confidence Intervals

Unit – III: Linear Algebra

Linear equations and matrices, matrix operations, solving system of linear equations, Gauss-Jordan method, Concept & Computation of determinant and inverse of matrix, Eigen values and eigen vectors, Illustrations of the methods, Positive semi definite and position definite matrices, illustrations.

Unit – IV: Hypothesis testing:

Constructing a hypothesis test; Null and alternate hypotheses; Test Statistic; Type I and Type II Error; Z test, t test, two sample t tests; Level of significance, Power of a test, ANOVA, Test for goodness of fit, Non-parametric tests.

Unit – V: Regression

Problem definition, Data pre-processing; model building; Diagnostics and Validation
Simple linear regression: Coefficient of determination, Significance tests for predictor variables, Residual analysis, Confidence and Prediction intervals

Suggested Readings:

1. U Dinesh Kumar (2017), Business Analytics: The Science of Data - Driven Decision Making, Wiley, New Delhi.
2. R. Evans James (2017), Business Analytics, 2nd edition, Pearson Education, New Delhi.
3. S. Christian Albright , Wayne L. Winston (2015), Business Analytics: Data Analysis and Decision Making, 5th edition, Cengage Publications
4. Howard Anton (Author), Chris Rorres (Author) (2016) , Elementary Linear Algebra with Supplemental Applications, 11 edition, Wiley, India.
5. Friedberg / Insel / Spence (2015), Linear Algebra, 4th edition, Pearson Education, New Delhi.

KARPAGAM ACADEMY OF HIGHER EDUCATION
(Deemed to be University)
(Established under Section 3 of UGC Act, 1956)
Pollachi Main Road, Coimbatore – 641 021, Tamil Nadu

Department of Mathematics

Subject : Business Analytics | **Semester II** | **L T P C**
Subject Code : 19BPU202 | **Class : I B.Com BPS** | **5 2 0 5**

Glossary of Statistical Terms

2 X 5 factorial design	A factorial design with one variable having two levels and the other having five levels.
Alpha	The probability of a Type I error.
Abscissa	Horizontal axis.
Additive law of probability	The rule giving the probability of the occurrence of one or more mutually exclusive events.
Adjacent values	Actual data points that are no more extreme than the inner fences.
Alternative hypothesis (H_1)	The hypothesis that is adopted when H_0 is rejected. Usually the same as the research hypothesis.
ß (Beta)	The probability of a Type II error.
Categorical data	Data representing counts or number of observations in each category.
Cell	The combination of a particular row and column ‹the set of observations obtained under identical treatment conditions.
Central limit theorem	The theorem that specifies the nature of the sampling distribution of the mean.
Chi-square test	A statistical test often used for analyzing categorical data.
Conditional probability	The probability of one event *given* the occurrence of some other event.
Confidence interval	An interval, with limits at either end, with a specified probability of including the parameter being estimated.
Confidence limits	An interval, with limits at either end, with a specified probability of including the parameter being estimated.
Constant	A number that does not change in value in a given situation.
Continuous variables	Variables that take on *any* value.
Correlation	Relationship between variables.
Correlation coefficient	A measure of the relationship between variables.
Count data	Data representing counts or number of observations in each category.
Covariance	A statistic representing the degree to which two variables vary together.
Criterion variable	The variable to be predicted.
Critical value	The value of a test statistic at or beyond which we will reject H0 .
Decision making	A procedure for making logical decisions on the basis of sample data.
Degrees of freedom (*df*)	The number of independent pieces of information remaining after estimating one or more parameters.

Density	Height of the curve for a given value of X- closely related to the probability of an observation in an interval around X.
Dependent variables	The variable being measured. The data or score.
Discrete variables	Variables that take on a small set of possible values.
Dispersion	The degree to which individual data points are distributed around the mean.
Distribution free tests	Statistical tests that do not rely on parameter estimation or precise distributional assumptions.
Effect size	The difference between two population means divided by the standard deviation of either population.
Efficiency	The degree to which repeated values for a statistic cluster around the parameter.
Event	The outcome of a trial.
Exhaustive	A set of events that represents all possible outcomes.
Expected value	The average value calculated for a statistic over an infinite number of samples.
Expected frequencies	The expected value for the number of observations in a cell if H_0 is true.
Experimental hypothesis	Another name for the research hypothesis.
Exploratory data analysis (EDA)	A set of techniques developed by Tukey for presenting data in visually meaningful ways.
External Validity	The ability to generalize the results from this experiment to a larger population.
Frequency distribution	A distribution in which the values of the dependent variable are tabled or plotted against their frequency of occurrence.
Frequency data	Data representing counts or number of observations in each category.
Goodness of fit test	A test for comparing observed frequencies with theoretically predicted frequencies.
Grand total ($\sum X$)	The sum of all of the observations.
Hypothesis testing	A process by which decisions are made concerning the values of parameters.
Independent variables	Those variables controlled by the experimenter.
Independent events	Events are independent when the occurrence of one has no effect on the probability of the occurrence of the other.
Interaction	A situation in a factorial design in which the effects of one independent variable depend upon the level of another independent variable.
Intercept	The value of Y when X is 0.
Interval scale	Scale on which equal intervals between objects represent equal differences ‹differences are meaningful.
Interval estimate	A range of values estimated to include the parameter.
Joint probability	The probability of the co-occurrence of two or more events.
Kurtosis	A measure of the peakedness of a distribution.
Leptokurtic	A distribution that has relatively more scores in the center and in the tails.
Linear	A situation in which the best-fitting regression line is a straight line.

relationship	
Linear regression	Regression in which the relationship is linear.
Marginal totals	Totals for the levels of one variable summed across the levels of the other variable.
Matched samples	An experimental design in which the same subject is observed under more than one treatment.
Mean absolute deviation (m.a.d.)	Mean of the absolute deviations about the mean.
Mean	The sum of the scores divided by the number of scores.
Measurement	The assignment of numbers to objects.
Measurement data	Data obtained by measuring objects or events.
Measures of central tendency	Numerical values referring to the center of the distribution.
Median location	The location of the median in an ordered series.
Median (Med)	The score corresponding to the point having 50% of the observations below it when observations are arranged in numerical order.
Mesokurtic	A distribution with a neutral degree of kurtosis.
Midpoints	Center of interval -- average of upper and lower limits.
Mode (Mo)	The most commonly occurring score.
Monotonic relationship	A relationship represented by a regression line that is continually increasing (or decreasing), but perhaps not in a straight line.
Multiplicative law of probability	The rule giving the probability of the joint occurrence of independent events.
Mutually exclusive	Two events are mutually exclusive when the occurrence of one precludes the occurrence of the other.
Negative relationship	A relationship in which increases in one variable are associated with decreases in the other.
Negatively skewed	A distribution that trails off to the left.
Nominal scale	Numbers used only to distinguish among objects.
normal distribution	A specific distribution having a characteristic bell-shaped form.
Ordinal scale	Numbers used only to place objects in order.
Ordinate	Vertical axis.
Outlier	An extreme point that stands out from the rest of the distribution.
p level	The probability that a particular result would occur by chance if H_0 is true. The exact probability of a Type I error.
Parameters	Numerical values summarizing population data.
Parametric tests	Statistical tests that involve assumptions about, or estimation of, population parameters.
Pearson product-moment correlation coefficient (*r*)	The most common correlation coefficient.
Percentile	The point below which a specified percentage of the observations fall.

Phi	The correlation coefficient when both of the variables are measured as dichotomies.
Platykurtic	A distribution that is relatively thick in the "shoulders."
Pooled variance	A weighted average of the separate sample variances.
Population variance	Variance of the population ‹usually estimated, rarely computed.
Population	Complete set of events in which you are interested.
Positively skewed	A distribution that trails off to the right.
Power	The probability of correctly rejecting a false H_0 .
Predictor variable	The variable from which a prediction is made.
Protected *t*	A technique in which we run *t* tests between pairs of means only if the analysis of variance was significant.
Quantitative data	Data obtained by measuring objects or events.
Random sample	A sample in which each member of the population has an equal chance of inclusion.
Random Assignment	Assigning participants to groups or cells on a random basis.
Range	The distance from the lowest to the highest score.
Range restrictions	Refers to cases in which the range over which X or Y varies is artificially limited.
Ranked data	Data for which the observations have been replaced by their numerical ranks from lowest to highest.
Rank - randomization tests	A class of nonparametric tests based on the theoretical distribution of randomly assigned ranks.
Ratio scale	A scale with a true zero point -- ratios are meaningful.
Real lower limit	The points halfway between the top of one interval and the bottom of the next.
Real upper limit	The points halfway between the top of one interval and the bottom of the next.
Rectangular distribution	A distribution in which all outcomes are equally likely.
Regression	The prediction of one variable from knowledge of one or more other variables.
Regression equation	The equation that predicts Y from X.
Regression coefficients	The general name given to the slope and the intercept ‹most often refers just to the slope.
Rejection region	The set of outcomes of an experiment that will lead to rejection of H_0 .
Rejection level	The probability with which we are willing to reject H0 when it is in fact correct.
Related samples	An experimental design in which the same subject is observed under more than one treatment.
Relative frequency view	Definition of probability in terms of past performance.
Research	The hypothesis that the experiment was designed to investigate.

hypothesis	
Sample	Set of actual observations. Subset of the population.
Sample statistics	Statistics calculated from a sample and used primarily to describe the sample.
Sample variance (s^2)	Sum of the squared deviations about the mean divided by N - 1.
Sample with replacement	Sampling in which the item drawn on trial N is replaced before the drawing on trial N + 1.
Sampling distribution of differences between means	The distribution of the differences between means over repeated sampling from the same population(s).
Sampling distribution of the mean	The distribution of sample means over repeated sampling from one population.
Sampling distributions	The distribution of a statistic over repeated sampling from a specified population.
Sampling error	Variability of a statistic from sample to sample due to chance.
Scales of measurement	Characteristics of relations among numbers assigned to objects.
Scatter plot	A figure in which the individual data points are plotted in two-dimensional space.
Scatter diagram	A figure in which the individual data points are plotted in two-dimensional space.
Scattergram	A figure in which the individual data points are plotted in two-dimensional space.
Sigma	Symbol indicating summation.
Significance level	The probability with which we are willing to reject H_0 when it is in fact correct.
Simple effect	The effect of one independent variable at one level of another independent variable.
Skewness	A measure of the degree to which a distribution is asymmetrical.
Slope	The amount of change in Y for a one unit change in X.
Spearman's correlation coefficient for ranked data (r_s)	A correlation coefficient on ranked data.
Standard deviation	Square root of the variance.
Standard error	The standard deviation of a sampling distribution.
Standard error of differences between means	The standard deviation of the sampling distribution of the differences between means.
Standard error of estimate	The average of the squared deviations about the regression line.
Standard scores	Scores with a predetermined mean and standard deviation.
Standard normal distribution	A normal distribution with a mean equal to 0 and variance equal to 1. Denoted N (0, 1).

Statistics	Numerical values summarizing sample data.
Student's *t* distribution	The sampling distribution of the *t* statistic.
Subjective probability	Definition of probability in terms of personal subjective belief in the likelihood of an outcome.
Sufficient statistic	A statistic that uses all of the information in a sample.
Sums of squares	The sum of the squared deviations around some point (usually a mean or predicted value).
Symmetric	Having the same shape on both sides of the center.
T scores	A set of scores with a mean of 50 and a standard deviation of 10.
Test statistics	The results of a statistical test.
Type I error	The error of rejecting H_0 when it is true.
Type II error	The error of not rejecting H_0 when it is false.
Unconditional probability	The probability of one event *ignoring* the occurrence or nonoccurrence of some other event.
Unimodal	A distribution having one distinct peak.
Variables	Properties of objects that can take on different values.
Weighted average	The mean of the form: $(a_1X_1 + a_2X_2)/(a_1 + a_2)$ where a_1 and a_2 are weighting factors and X_1 and X_2 are the values to be average.

KARPAGAM ACADEMY OF HIGHER EDUCATION
(Deemed to be University)
(Established under Section 3 of UGC Act, 1956)
Pollachi Main Road, Coimbatore – 641 021, Tamil Nadu

Department of Mathematics

19BPU202 **Business Analytics** **L T P C**
Semester – II **5 2 0 5**

Unit – I: Introduction to data science

Concepts of measurement, scales of measurement, Different types and scales of data (ratio, interval, nominal and ordinal);

Design of data collection formats with illustration, data quality and issues with date collection systems with examples from business, cleaning and treatment of missing data, principles of data visualization.

Data summarization and visualization methods : Histograms, Frequency distributions, Relative frequency, measures of central tendency and dispersion; Tables, Graphs, Charts, Box Plot; Chebychev's Inequality.

Suggested Readings:

1. U Dinesh Kumar (2017), Business Analytics: The Science of Data - Driven Decision Making, Wiley, New Delhi.
2. R. Evans James (2017), Business Analytics, 2nd edition, Pearson Education, New Delhi.
3. S. Christian Albright , Wayne L. Winston (2015), Business Analytics: Data Analysis and Decision Making, 5th edition, Cengage Publications
4. Howard Anton (Author), Chris Rorres (Author) (2016) , Elementary Linear Algebra with Supplemental Applications, 11 edition, Wiley, India.
5. Friedberg / Insel / Spence (2015), Linear Algebra, 4th edition, Pearson Education, New Delhi.

Unit – I

Introduction

Statistical tools are found useful in progressively increasing of disciplines. In ancient times the statistics or the data regarding the human force and wealth available in their land had been collected by the rulers. Now-a-days the fundamental concepts of statistics are considered by many to be essential part of their knowledge.

Origin and Growth

The origin of the word 'statistics' has been traced to the Latin word 'status', the Italian word 'statista' , the French word 'statistique' and the German word 'statistik'. All these words mean political state.

Meaning

The word 'statistics' is used in two different meanings. As a plural word it means data or numerical statements. As a singular word it means the science of statistics and statistical methods. The word 'statistics' is also used currently as singular to mean data.

Definitions

Statistics is "the science of collection, organization, presentation, analysis and interpretation of numerical data". – Dr S.P.Gupta.

Statistics are numerical statement of facts in any department of enquiry, placed in relation to each other". – Dr.A.L.Bowley.

Functions

The following are the important functions of statistics.

- σ Collection
- σ Numerical Presentation
- σ Diagrammatic Presentation
- σ Condensation
- σ Comparison
- σ Forecasting
- σ Policy Making
- σ Effect Measuring
- σ Estimation
- σ Tests of significance.

Characteristics

* Statistics is a Quantitative Science.
* It never considers a single item.
* The values should be different.
* Inductive logic is applied.
* Statistical results are true on the average.
* Statistics is liable to be misused.

Samples vs. Populations

Population: A complete set of observations or measurements about which conclusions are to be drawn.

Sample: A subset or part of a population.

Not necessarily random

Statistics vs. Parameters

Parameter: A summary characteristic of a population.

Summary of Central tendency, variability, shape, correlation

E.g., Population mean, Population Standard Deviation, Population Median, Proportion of population of registered voters voting for Bush, Population correlation between Systolic & Diastolic BP

Statistic: A summary characteristic of a sample. Any of the above computed from a sample taken from the population.

E.g., Sample mean, Sample Standard Deviation, median, correlation coefficient

MEASURES OF CENTRAL TENDENCY

Introduction

In this chapter we are going to deal with Measures of central tendency and about the measures of dispersion. The measures of central tendency concentrate about the values in the central part of the distribution. Plainly speaking an average of a statistical series is the value of the variable which is the representative of the entire distribution. If we know the average alone we cannot form a complete idea about the distribution so for the completeness of the idea we use Measures of dispersion.

Measures of Central Tendency

According to Professor Bowley the measures of central tendency are "statistical constants which enable us to comprehend in a single effort the significance of the whole "
The following three are the basic measures of central tendency in this chapter we deal with

- Arithmetic Mean or simply Mean
- Median
- Mode

Arithmetic Mean or Mean

Arithmetic Mean or simply Mean is the total values of the item divided by their number of the items. It is usually denoted by $\bar{X}$

Individual series

$\bar{X} = \Sigma X / N$

Example 1

The expenditure of ten families are given below .Calculate arithmetic mean.

30, 70, 10, 75, 500, 8, 42, 250, 40, 36

Solution

Here N=10

$\Sigma X = 30 + 70 + 10 + 75 + 500 + 8 + 42 + 250 + 40 + 36 = 1061$

$\bar{X} = 1061 / 10 = 106.1$

Discrete series

$\bar{X} = \Sigma f X / \Sigma f$

Example 2

Calculate the mean number of person per house.

No. of person : 2 3 4 5 6

No. of house : 10 25 30 25 10

Solution

X	f	f X
2	10	20
3	25	75
4	30	120
5	25	125
6	10	60
	$\Sigma f = 100$	$\Sigma f X = 400$

$\bar{X}$

$\bar{X} = 400 / 100 = 4$.

Continuous series

$\bar{X}$

$\bar{X} = \Sigma f m / \Sigma f$ where m represents the mid value

Mid-value = (upper boundary + lower boundary) / 2.

Example 3

Calculate the mean for the following.

Marks : 20-30 30-40 40-50 50-60 60-70 70-80

No. of student : 5 8 12 15 6 4

Solution

C.I	f	m	f m
20-30	5	25	125
30-40	8	35	280
40-50	12	45	540
50-60	15	55	825
60-70	6	65	390
70-80	4	75	300
	$\Sigma f = 50$		$\Sigma f m = 2460$

$\bar{X}$

$\bar{X} = 2460 / 50 = 49.2$.

Median

The median is the value for the middle most items when all the items are in the order of magnitude. It is denoted by M or Me.

Individual series

For odd number of items Median Position = (N+1) / 2

For even number of item

Position of the Median = [(N / 2)+((N/2)+1)] / 2

Example 1

Calculate median for the following.

22 10 6 7 12 8 5

Solution

Here N =7

Arrange in ascending order or descending order.

5 6 7 8 10 12 22

(N+1) / 2= (7+1) /2 = 4 th item = 8

Discrete series

Position of the median = (N+1) / 2 th item.

Example 2

Find the median for the following.

X : 10 15 17 18 21

F: 4 16 12 5 3

Solution

X	f	c.f
10	4	4
15	16	20
17	12	32
18	5	37
21	3	40

N= 40

(N+1) /2 = (40+1) / 2 = 20.5th item

= (20 th item +21st item) /2 =(15+17) /2

= 16.

Continuous series

$$M = L+\frac{[((N/2) - c.f) \times i]}{f.}$$

Where L - lower boundary, f - frequency, I - size of class interval and c.f - cumulative frequency.

Example 3

Calculate the median height (Ht) for the No .of Students (NoS) given below.

Ht: 145-150 150-155 155-160 160-165 165-170 170-175

NoS: 2 5 10 8 4 1

Solution

Height	No. of student	c.f
145-150	2	2
150-155	5	7
155-160	10	17
160-165	8	25
165-170	4	29
170-175	1	30
	Σ f = 30	

Position of the median = N/2 th item = 30 / 2 =15.

$$M = L+\frac{[((N/2) - c.f) \times i]}{f.}$$

$$= 155+ \frac{[(15-7)x5]}{10} = 155+(40/10) = 159.$$

Mode

Mode is the value which has the greatest frequency density. Mode is usually denoted by Z.

Individual series

In a set of observations the value which occur more number of time is known as Mode. In other way the most frequented value in a set of value is Mode.

Example 1

Determine the mode for the set of Individual observations given as follows 32, 35, 42, 32, 42, 32.

Solution

Mode = 32 Uni-model

Discrete series

Determine the Mode

Size of dress	**No. of set**
18	55
20	120
22	108
24	45

Here the mode represents highest frequency ie. I 120.

So, Mode = 20

Continuous series

$Z = L + [i(f_1 - f_0) / (2f_1 - f_0 - f_2)]$

Where L- lower boundary , f_1-frequency of the modal class, f_0 – frequency of the preceding modal class, f_2- frequency of the succeeding modal class, i-size of class interval , c.f- cumulative frequency.

Example

Determine the mode

Marks	: 0-10	10-20	20-30	30-40	40-50
No.of student	: 5	20	35	20	12

Solution

Marks	No. of student
0-10	5
10-20	20
20-30	35
30-40	20
40-50	12

$Z = L + [i(f_1-f_0) /(2f_1 –f_0 –f_2)]$

$= 20+[10(35-20)/(2(35)-20-20)] = 20+5 = 25.$

Empirical relation

- Mode = 3 median -2 mean.

MEASURES OF DISPERSION

Measure of dispersion deals mainly with the following three measures

- Range
- Standard deviation
- Coefficient of variation

Range

Range is the difference between the greatest and the smallest value.

- Range = L – S , where L-largest value & S-Smallest value
- Coefficient of range = (L-S) /(L+S)

Individual series

Example

Find the value of range and its coefficient of range for the following data.

8 , 10, 5, 9, 12, 11

Solution

Range = L – S

= 12- 5 =7

Coefficient of range = (L-S) /(L+S)

= (12-5) / (12+5)

= 7 /17

= 0.4118

Continuous series

Range = L – S, where L-Mid-value of largest boundary & S-Mid-value of smallest boundary

Calculate the range for the following continuous frequency distribution

Marks : 20-30 30-40 40-50 50-60 60-70 70-80

No.of student : 5 8 12 15 6 4

Solution

C.I	f	m
20-30	5	25
30-40	8	35
40-50	12	45
50-60	15	55
60-70	6	65
70-80	4	75

Here L=75 & S=25

Range = L – S = 75-25 = 50

Quartile Deviation

Quartile Deviation is half of the difference between the first and the third quartiles. Hence it is called Semi Inter Quartile Range.

Coefficient of Quartile Deviation

A relative measure of dispersion based on the quartile deviation is called the coefficient of quartile deviation. It is defined as

$$= \frac{\frac{Q_3 - Q_1}{2}}{\frac{Q_3 + Q_1}{2}} = \frac{Q_3 - Q_1}{Q_3 + Q_1}$$

It is pure number free of any units of measurement. It can be used for comparing the dispersion in two or more than two sets of data.

Example

The wheat production (in Kg) of 20 acres is given as: 1120, 1240, 1320, 1040, 1080, 1200, 1440, 1360, 1680, 1730, 1785, 1342, 1960, 1880, 1755, 1720, 1600, 1470, 1750, and 1885. Find the quartile deviation and coefficient of quartile deviation.

Solution

After arranging the observations in ascending order, we get

1040, 1080, 1120, 1200, 1240, 1320, 1342, 1360, 1440, 1470, 1600, 1680, 1720, 1730, 1750, 1755, 1785, 1880, 1885, 1960.

$$Q_1 = \text{Value of } \left(\frac{n+1}{4}\right)th \text{ item}$$

$$= \text{Value of } \left(\frac{20+1}{4}\right)th \text{ item}$$

$$= \text{Value of } (5.25)th \text{ item}$$

$$= 5th \text{ item} + 0.25(6th \text{ item} - 5th \text{ item}) = 1240 + 0.25(1320 - 1240)$$

$$Q_1 = 1240 + 20 = 1260$$

$$Q_3 = \text{Value of } \frac{3(n+1)}{4}th \text{ item}$$

$$= \text{Value of } \frac{3(20+1)}{4}th \text{ item}$$

$$= \text{Value of } (15.75)th \text{ item}$$

$$= 15th \text{ item} + 0.75(16th \text{ item} - 15th \text{ item}) = 1750 + 0.75(1755 - 1750)$$

$$Q_3 = 1750 + 3.75 = 1753.75$$

Quartile Deviation (Q.D)

$$= \frac{Q_3 - Q_1}{2} = \frac{1753.75 - 1260}{2} = \frac{492.75}{2} = 246.875$$

Coefficient of Quartile Deviation

$$= \frac{Q_3 - Q_1}{Q_3 + Q_1} = \frac{1753.75 - 1260}{1753.75 + 1260} = 0.164$$

Standard deviation

The standard deviation is the root mean square deviation of the values from the arithmetic mean .It is a positive square root of variants. It is also called root mean square deviation. This is usually denoted by σ .

Individual series

$$\sigma = \sqrt{(\Sigma x^2 / N) - (\Sigma x / N)^2}$$

Example 1

Calculate standard deviation for the following data.

40,41,45,49,50,51,55,59,60,60.

Solution

X	X^2
40	1600
41	1681
45	2025
49	2401
50	2500
51	2601
55	3025
59	3481
60	3600
60	3600

510 $\Sigma x^2 = 26504$

$$\sigma = \sqrt{(\Sigma x^2 / N) - (\Sigma x / N)^2}$$

$$= \sqrt{(26514/10) - (510/10)^2}$$

$$= 7.09$$

Discrete series

$$\sigma = \sqrt{(\Sigma fx^2 / \Sigma f) - (\Sigma fx / \Sigma f)^2}$$

Example 2

Calculate standard deviation for the following data.

X : 0 1 2 3 4 5

F : 1 2 4 3 0 2

Solution

X	**f**	**fx**	**x^2**	**fx^2**
0	1	0	0	0
1	2	2	1	2
2	4	8	4	16
3	3	9	9	27
4	0	0	16	0
5	2	10	25	50
	$\Sigma f = 12$	$\Sigma fx = 29$		$\Sigma fx^2 = 95$

$$\sigma = \sqrt{(\Sigma fx^2 / \Sigma f) - (\Sigma fx / \Sigma f)^2}$$

$= \sqrt{(95/12) - (29/12)^2} = 1.44$

Continuous series

$$\sigma = \sqrt{(\Sigma fm^2 / \Sigma f) - (\Sigma fm / \Sigma f)^2}$$

Example 3

C.I : 0-10 10-20 20-30 30-40 40-50

F : 2 5 9 3 1

Solution

C.I	f	m	fm	m^2	fm^2
0-10	2	5	10	25	50
10-20	5	15	75	225	1125
20-30	9	25	225	625	5625
30-40	3	35	105	1225	3675
40-50	1	45	5	2025	2025
	20		460		12500

$$\sigma = \sqrt{(\Sigma fm^2 / \Sigma f) - (\Sigma fm / \Sigma f)^2}$$

$= \sqrt{(12500/20) - (460/20)^2}$

$= 9.79$

Coefficient of variation

Coefficient of variation = [standard deviation / arithmetic mean] x100

Example 1

Calculate the coefficient of variation.

Mean= 51, standard deviation = 7.09

Solution

Coefficient of variation = [standard deviation / arithmetic mean] x100

= (7.09 /51) x100

= 13.9

Questions

1) Calculate the Mean for the following.

X	20	30	35	15	10
f	2	3	4	3	2

2) Define Median and give Example.

3) Calculate the Range and its Coefficient for the following data.

X	:	12	14	16	18	20
f	:	1	3	5	3	1

4) What do you mean by Bimodal?

5) Calculate the Median for the following data.

80 100 50 90 120 110

6) Write the relation between Standard Deviation and Variance.

7) Calculate the Average number of students per class for the following data.

26 46 33 25 36 27 34 29

8) Find Median and Mode for the following data.

13 16 17 15 18 14 19 15 12

9) Define the term Range.

10) Find the Arithmetic Mean for the following data.

70 60 75 50 42 95 46

11) Calculate the Range and its Coefficient for the following data.

17 10 56 19 12 11 18 14

12) Define the term Quartile Deviation.

13) Find the median for 57, 58, 61, 42, 38, 65, 72, and 66

14) Write the empirical relation for Mode.

Exercise

1. Calculate the Mode for the following Continuous Frequency Distribution.

Salary (in Rs. 1000s)	0 -19	20-39	40 - 59	60-79	80-99
No. of Employees	5	20	35	20	12

2. Find the Mean and the Standard Deviation for the given below data set.

10	14	20	12	21	16	19	17	14	25

3. Calculate the Standard Deviation and Coefficient of Variance (CV) for the following data.

X	0 – 10	10 - 20	20 - 30	30 – 40	40 - 50
f	2	5	9	3	1

4. Calculate the Median for the following Continuous Frequency Distribution.

Wages (in Rs.)	0 - 19	20 - 39	40 - 59	60 – 79	80 - 99
No. of Workers	5	20	35	20	12

5. Calculate the Coefficient of Variation for the following data.

X	6	9	12	15	18
f	7	12	13	10	8

6. Calculate the Median for the following.

Hourly Wages (in Rs.)	40 - 50	50 – 60	60 - 70	70 - 80	80 - 90	90 - 100
Number of Employees	10	20	15	30	15	10

7. The following data give the details about salaries (in thousands of rupees) of seven employees randomly selected from a Pharmaceutical Company.

Serial No.	1	2	3	4	5	6	7
Salary per Annum ('000)	89	57	104	73	26	121	81

Calculate the Standard Deviation and Coefficient of variance of the given data.

8. Calculate the Arithmetic Mean for the following data.

Height (cms) : 160 161 162 163 164 165 166

No. of Persons : 27 36 43 78 65 48 28

9. Calculate the Coefficient of Variance for the following data. 77 73 75 70 72 76 75 72 74 7

Question	Option 1	Option 2	Option 3	Option 4	Answer
A measure of central tendency helps to get a single representative value for a set of …….. values.	sum	difference	equal	unequal	total
…….. is the total of the values of the items divided by their number	arithmetic mean	arithmetic median	arithmetic mode	arithmetic range	arithmetic mean
Arithmetic Mean is the ……...of the values of the items divided by their number	sum	difference	product	total	total
Arithmetic Mean is the total of the values of the items ……. by their number	sum	difference	product	divided	divided
The …….. of the deviations of the values from their arithemetic mean is zero.	sum	difference	product	division	sum
The sum of the deviations of the values from their …….is zero.	arithmetic mean	arithmetic median	arithmetic mode	arithmetic range	arithmetic mean
The sum of the deviations of the values from their arithemetic mean is …….....	zero	one	two	three	zero
……... is the value of the middle most item when all the items are in order of magnitude.	Mean	Median	Mode	Range	Median
Median is the value of the ……... most item when all the items are in order of magnitude.	initial	final	middle	higher	middle
…….. is the value which has the greatest frequency density.	Mean	Median	Mode	Range	Mode
Mode is the value which has the …….frequency density.	smallest	greatest	initial	final	greatest
……. mean is the appropriate root of the product of the values of the items.	arithmetic	geometric	harmonic	standard	geometric
Geometric mean is the appropriate…….of the product of the values of the items.	sum	difference	root	quotient	root
Geometric mean is the appropriate root of the …….of the values of the items.	sum	difference	product	divided	product
……... is the reciprocal of the mean of reciprocals of the values of the items	arithmetic	geometric	harmonic	standard	harmonic

Harmonic mean is the ……..of the mean of reciprocals of the values of the items	sum	difference	root	reciprocal	reciprocal
Harmonic mean is the reciprocal of the ……. of reciprocals of the values of the items	Mean	Median	Mode	Range	Mean
In symmetical distributions the relation is ……	mean=median=mode	mean≠median=mode	mean=median≠mode	mean≠median≠mode	mean=median=mode
The relation between the means is ……	A.M< G.M< H.M	A.M = G.M =H.M	A.M > G.M > H.M	A.M≠ G.M≠ H.M	A.M > G.M > H.M
……...are positional values.	Relative	absolute	possibility	finite	Relative
……divide the total frequency into ten equal parts and hence their name.	quartile	mean	median	standard	quartile
Deciles divide the …….frequency into ten equal parts and hence their name.	quartile	deciles	percentiles	mean	percentiles
Deciles divide the total frequency into …….. equal parts and hence their name.	ten	twenty	fifty	hundred	hundred
……. divide the total frequency into hundred equal parts and hence their name.	Relative	absolute	possibility	finite	Relative
Percentiles divide the total frequency into …….. parts and hence their name.	real numbers	pure numbers	complex numbers	imaginary numbers	pure numbers
……..measures give pure numbers which are free form the units of measurements of data.	scale	value	units	range	units
Relative measures give ……...which are free form the units of measurements of data.	real numbers	pure numbers	complex numbers	imaginary numbers	pure numbers
Relative measures give pure numbers which are free form the ……. of measurements of data.	scale	value	units	range	units
……..and ……measures are two kinds of measures of dispersion.	absolute and possibility	finite and infinite	non relative and relative	absolute and relative	absolute and relative
……... is the difference between the greatest and smallest of the values.	Median	Mean	Range	Mode	Range
Range is the ……... between the greatest and smallest of the values.	sum	difference	product	quotient	difference

Range is the difference between the …….. of the values.	smallest and greatest	greatest and smallest	finite and infinite	greatest and infinite	greatest and smallest
……. is used in statistical quality control.	Median	Mean	Range	Mode	Range
Range is used in statistical …….. control.	units	constant	quality	value	quality
…….deviation is half of the difference between first and third quartiles.	quartile	mean	median	standard	quartile
Quartile deviation is ……..of the difference between first and third quartiles.	one fourth	half	one third	three fourth	half
Quartile deviation is half of the difference between ……..quartiles.	first and third	first and two	two and third	third and fourth	first and third
There are …….. kinds of mean deviations	one	two	three	four	three
Standard deviation the……..deviation of the values from their arithmetic mean	root mean square	root median square	root mode square	root range square	root mean square
Standard deviation the root mean square deviation of the values from their arithmetic ……	mean	median	mode	standard deviation	mean
……... deviationof the values from the arithmetic mean is known as variance.	Mean square	root mean square	range square	standard deviation	Mean square
Mean square deviationof the values from the arithmetic mean is known as variance.	arithmetic range	arithmetic mode	arithmetic median	arithmetic mean	arithmetic mean
Mean square deviationof the values from the arithmetic mean is known as …….	mean	median	variance	standard deviation	variance
…….. is the positive square root of variance.	mean	median	variance	standard deviation	standard deviation
Standarad deviation is the positive ……...of variance.	square root	cubic root	fourth root	fifth root	square root
Standarad deviation is the positive square root of …….	mean	median	variance	standard deviation	variance

KARPAGAM ACADEMY OF HIGHER EDUCATION
(Deemed to be University)
(Established under Section 3 of UGC Act, 1956)
Pollachi Main Road, Coimbatore – 641 021, Tamil Nadu

Department of Mathematics

19BPU202 **Business Analytics** **L T P C**
Semester – II **5 2 0 5**

Unit – II: Probability and Sampling Estimation

Basic probability concepts, Conditional probability, Bayes Theorem, Probability distributions, Continuous and discrete distributions, Binomial Distribution, Uniform Distribution, Exponential Distribution, Normal distribution, Central Limit Theorem, Sequential decision making, Decision tree

Sampling and estimation: Estimation problems, Point and interval estimates, Confidence Intervals

Suggested Readings:

1. U Dinesh Kumar (2017), Business Analytics: The Science of Data - Driven Decision Making, Wiley, New Delhi.
2. R. Evans James (2017), Business Analytics, 2nd edition, Pearson Education, New Delhi.
3. S. Christian Albright , Wayne L. Winston (2015), Business Analytics: Data Analysis and Decision Making, 5th edition, Cengage Publications
4. Howard Anton (Author), Chris Rorres (Author) (2016) , Elementary Linear Algebra with Supplemental Applications, 11 edition, Wiley, India.
5. Friedberg / Insel / Spence (2015), Linear Algebra, 4th edition, Pearson Education, New Delhi.

Probability

Definition: A probability experiment is an action, or trial, through which specific results (counts, measurements or responses) are obtained. The result of a single trial in a probability experiment is an outcome. The set of all possible outcomes of a probability experiment is the sample space. An event consists of one or more outcomes and is a subset of the sample space.

Example 1 The experiment consists of tossing a coin then rolling a die. the sample space consists of

H						T					
1	2	3	4	5	6	1	2	3	4	5	6
H1	H2	H3	H4	H5	H6	T1	T2	T3	T4	T5	T6

How many outcomes are there? Do you agree, disagree, or have no opinion, and what is your gender?

An event that consists of a single outcome is called a simple event

DEFINITION: Classical (or theoretical) is used when each outcome in a sample space is equally likely to occur. The Classical probability of an event E is given by:

$$P(E) = \frac{Number\ of\ outcomes\ in\ E}{Total\ number\ of\ outcomes\ in\ sample\ space}$$

Example 3 Roll a die: What is the sample space? {1,2,3,4,5,6}
Event A: rolling a 3, p = 1/6 = 0.157. Note this is a simple event.
Event C: rolling < 5, p =4/6 = 0.667. Note this is not a simple event.

DEFINITION Empirical (or statistical) probability is based on observations obtained from probability experiments. The empirical probability of an event E is the relative frequency of event E:

$$P(E) = \frac{Frequency\ of\ event\ E}{Total\ \ frequency} = \frac{f}{n}$$

Example: Finding Empirical Probabilities . Each fish (Bluegill, Redgill, and Crappy) is equally likely to get caught. You catch and release the following.

Fish Type	Number of times caught, f
Bluegill	13
Redgill	17
Crappy	10
	$\Sigma f = 40$

Probability of catching a bluegill = 13/40 = 0.325

Law of Large Numbers (p 114): As an experiment is repeated over and over, the empirical probability of the event approaches the theoretical (actual) probability of the event.

Basic Concepts of Probability

In statistics, an experiment is a process leading to at least two possible outcomes with uncertainty as to which will occur.

The set of all possible outcomes of an experiment is called the sample space (S). Each outcome in S is called a sample point.

Example 1

Three items are selected at random from a manufacturing process. Each item is inspected and classified defective (D) or non-defective (N).

An event is a subset of a sample space, it consists of one or more outcomes with a common characteristic.

Example 2

The event that the number of defectives in above example is greater than 1.

The null space or empty space is a subset of the sample space that contains no outcomes (∅).

The intersection of two events A and B denoted by (A∩B) is the event containing all outcomes that are common to A and B.

Events are mutually exclusive if they have no elements in common.

The union of two events A and B denoted by (A∪B) is the event containing all the elements that belong to A or to B or to both.

Events are collectively exhaustive if no other outcome is possible for a given experiment.

The complement of an event A with respect to S is the set of outcomes of S that are not in A denoted by (A' or A^c or $\bar{A}$).

Probability of an Event

Notation : P(A) The probability of an event A

Probability postulates :

1. P(S) = 1
2. P(∅) = 0
3. $0 \le P(A) \le 1$

Methods of Assigning Probabilities

1. The classical approach

   If an experiment can result in any one of N different equally likely outcomes, and if exactly n of these outcomes correspond to event A, then

$$P(A) = \frac{n}{N}$$

2. The relative frequency approach

   If some number N of experiments are conducted and the event A occurs in N_A of them, then

$$P(A) = \lim_{N \to \infty} \frac{N_A}{N}$$

3. The subjective approach

   Subjective probability is a personal assessment of the likelihood of an event.

Principle of Counting - Permutation and Combination

Counting Sample Points

Fundamental principle of counting:

If an operation can be performed in N_1 ways, a second operation can be performed in N_2 ways, and so forth, then the sequence of k operations can be performed in

$$N_1 \ N_2 \ N_3 \ \dots \ N_k \quad \text{ways}$$

Example 3

Suppose a licence plates containing two letters following by three digits with the first digit not zero. How many different licence plates can be printed?

A permutation is an arrangement of all or part of a set of objects.

Example 4

The possible permutations from 3 letters A, B, C

The number of permutations of n distinct objects is n!.

The number of permutations of n distinct objects taken r at a time is

$$nPr = \frac{n!}{(n-r)!}$$

Example 5

In how many ways can 10 people be seated on a bench if only 4 seats are available?

The combination is a collection of n objects taken r at a time in any selections of r objects where order does not count. The number of combinations of n objects taken r at a time is

$$nCr = \frac{n!}{r!(n-r)!}$$

Example 6

A box contains 8 eggs, 3 of which are rotten. Three eggs are picked at random. Find the probabilities of the following events.

a) Exactly two eggs are rotten.
b) All eggs are rotten.
c) No egg is rotten.

Addition Rule and Complimentary Rule

Addition Rule

1. For events that are not mutually exclusive

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

Example 7

A card is drawn from a complete deck of playing cards. What is the probability that the card is a heart or an ace?

For mutually exclusive events

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

Complementary Rule

If A and A′ are complementary events then

$$P(A') = 1 - P(A)$$

Conditional Probability, Statistically Independence and Multiplication Rule

Conditional Probability

Let A and B be two events. The conditional probability of event A, given event B, denoted by P(A|B) is defined as

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$ provided that P(B) > 0.

Similarly, the conditional probability of B given A is defined as

$$P(B|A) = \frac{P(A \cap B)}{P(A)}$$ provided that P(A) > 0.

Statistically Independence

Two events are independent when the occurrence or non-occurrence of one event has no effect on the probability of occurrence of the other event.

Definition : Two events A and B are independent if and only if

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

Multiplication Rule

1. For dependent events

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

   $= P(B)P(A|B)$

2. For independent events

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

Theorem of Total Probability

If the events $B_1, B_2, \ldots, B_k$ constitute a partition of the sample space S such that $P(B_i) \neq 0$ (i = 1, …, k) then for any event A of S

$$P(A) = P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + \ldots + P(B_k)P(A|B_k)$$

Baye's Theorem

If $B_1, B_2, \ldots, B_k$ are mutually exclusive events such that $B_1 \cup B_2 \cup \ldots \cup B_k$ contains all sample points of S, then for any event A of S with $P(A) \neq 0$,

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{P(B_1)P(A|B_1) + P(B_2)P(A|B_2) + \ldots + P(B_k)P(A|B_k)}$$

for i = 1, 2, 3, …, k

Possible Questions

Unit II

PART-B

1. Explain the functions of Random variable by an example.
2. Write a short note about the following terms:
   i. Conditional Probability
   ii. Independent and Dependent Event
   iii. Mutually Exclusive Event
3. Explain the Characteristics of Random variable with an example.
4. Write a short note about the following terms:
   a) Event and Mutually Exclusive Event
   b) Exclusive and Exhaustive Events
   c) Dependent and Independent Events
   d) Simple and Compound Events
5. (i) Explain the of axioms of the theory probability.
   (ii) State and prove Bayes theorem
6. State and prove the Addition and Multiplication theorems of Probability
7. Write a short note about the following terms:
   i) Random Event and Independent Event
   ii) Differentiate between Permutation and Combination.

PART-C

8. Write a short note about the following terms:
   a) Conditional Probability
   b) Bayes' Theorem
   c) Event and Mutually Exclusive Event
   d) Exclusive and Exhaustive Events
   e) Dependent and Independent Events

Question	Option 1	Option 2	Option 3	Option 4	Answer
The probability of drawing a card of King from a pack of cards is……	1/4	1/11	1/12	1/13	1/13
In tossing a coin, the probability of getting head is ……..	1/2	1/3	2	0	1/2
The probability that a leap year selected at random contain 53 Sundays is…….	1/7	2/7	3/7	1/53	2/7
A bag contains 7 red and 8 black balls. The probability of drawing a red ball is …….	7/15	8/15	1/15	14/15	7/15
The probability of drawing a card of clubs from a pack of 52 cards is ……..	0	(1/3)	2/4	1/4	1/4
The probability of drawing an ace or queen card from a pack of 52 cards is …	1/13	1/4	2/13	1/52	1/13
The total probability is is always equal to……..	0.5	2	1	0	1
A variable whose value is a number determined by the outcome of a random experiment is called a……	Sample	Random variable	Outcome	Event	Random variable
If a random variable takes only a finite or a countable number of values, it is called ……..	Finite random space	Continous random variable	Discrete random variable	Infinite random variable	Discrete random variable
A continuous random variable is a random variable X which can take any value between….	Interval	Limits	Finite values	Infinite values	Interval
Suppose that X be a discrete or continuous random variable, then distribution function is a……function of x.	Non-decreasing	Decreasing	Neither increasing nor decreasing	Can be increasing and decreasing	Non-decreasing
The function f(x) = 5x^4 , 0 < x < 1 can be a…….. of a random variable X.	Probability mass function	Probability density function	Distribution function	Exponential function	Probability density function

If F(x) is the cumulative distribution function of a continuous random variable X with p.d.f f(x) then……	F'(x) = f(x)	F'(x) not equal to f(x)	F'(x) < f(x)	F'(x) >f(x)	F'(x) = f(x)
If X is a continuous random variable with p.d.f f(x), then F(b)-F(a)=	P(a>X>b)	P(a<X>b)	P(b<X<a)	P(a<X<b)	P(a<X<b)
Which one of the following represents the best estimate of the population mean?	The sample mean	The mean of several sample means	The mode of several sample means	The median of several sample means	The mean of several sample means
Which of the following statements are true?	Parameters describe samples and statistics describe populations	Statistics describe samples and populations	Parameters describe populations and statistics describe samples	Both (a) and (b) above	Parameters describe populations and statistics describe samples
The narrower the confidence intervals:	The more confidence you can place in your results	The less you can rely on your results	The greater the chance that your results were due to sampling error	Correlation between the two scores	The more confidence you can place in your results
Statistical significance:	Is directly equivalent to psychological importance	Does not necessarily mean that results are psychologically important	Depends on sample size	Both (b) and (c) above	Both (b) and (c) above

All other things being equal:	The more sample size increases, the more power decreases	The more sample size increases, the more power increases	Sample size has no relationship to power	The more sample size increases, the more indeterminate the power	The more sample size increases, the more power increases
Find probability of drawing diamond and a heart card from a pack of 52 cards?	13/102	1/4	2/13	7/16	13/102
The probability of drawing king and queen card from a pack of 52 cards is	13/102	1/4	2/13	8/663	8/663
Two coins are tossed five times, find the probability of getting an even number of heads ?	0.25	1	0.4	0.25	0.25
All other things being equal, the more powerful the statistical test:	The wider the confidence intervals	The more likely the confidence interval will include zero	The narrower the confidence interval	The smaller the sample size	The narrower the confidence interval
Power can be calculated by a knowledge of:	The statistical test, the type of design and the effect size	The statistical test, the criterion significance level and the effect size	The criterion significance level, the effect size and the type of design	The criterion significance level, the effect size and the sample size	The criterion significance level, the effect size and the sample size

Which of the following constitute continuous variables?	Anxiety rated on a scale of 1 to 5 where 1 equals not anxious, 3 equals moderately anxious and 5 equals highly anxious	Gender	Temperature	Intelligence	Temperature
A continuous variable can be described as:	Able to take only certain discrete values within a range of scores	Able to take any value within a range of scores	Being made up of categories	Being made up of variables	Able to take any value within a range of scores
Which one of the following represents the best estimate of the population mean?	The sample mean	The mean of several sample means	The mode of several sample means	The median of several sample means	The mean of several sample means
Which one of the following represents the best estimate of the population mean?	The sample mean	The mean of several sample means	The mode of several sample means	The median of several sample means	The mean of several sample means
The narrower the confidence intervals:	The more confidence you can place in your results	The less you can rely on your results	The greater the chance that your results were due to sampling error	Correlation between the two scores	The more confidence you can place in your results
Which of the following could be considered as categorical variables?	Gender	Brand of baked beans	Hair colour	All of the above	All of the above

One card is drawn at random from a well-shuffled pack of 52 cards. What is the probability that it will be a diamond ?	1/13	1/4	1/52	1/15	1/4
Which of the following is a continous probability distribution?	Normal	Poisson	Binomial	Uniform	Normal
For which distribution, mean,meadian and mode coincides?	Poisson	F	Chi square	Normal	Normal
The range of standard normal variate is …….	$-\infty$ to $+\infty$	0 to 1	0 to ∞	1 to ∞	$-\infty$ to $+\infty$

KARPAGAM ACADEMY OF HIGHER EDUCATION
(Deemed to be University)
(Established under Section 3 of UGC Act, 1956)
Pollachi Main Road, Coimbatore – 641 021, Tamil Nadu

Department of Mathematics

19BPU202	Business Analytics	L T P C
Semester – II		5 2 0 5

Unit – III: Linear Algebra

Linear equations and matrices, matrix operations, solving system of linear equations, Gauss-Jordan method, Concept & Computation of determinant and inverse of matrix, Eigen values and eigen vectors, Illustrations of the methods, Positive semi definite and position definite matrices, illustrations.

Suggested Readings:

1. U Dinesh Kumar (2017), Business Analytics: The Science of Data - Driven Decision Making, Wiley, New Delhi.
2. R. Evans James (2017), Business Analytics, 2nd edition, Pearson Education, New Delhi.
3. S. Christian Albright , Wayne L. Winston (2015), Business Analytics: Data Analysis and Decision Making, 5th edition, Cengage Publications
4. Howard Anton (Author), Chris Rorres (Author) (2016) , Elementary Linear Algebra with Supplemental Applications, 11 edition, Wiley, India.
5. Friedberg / Insel / Spence (2015), Linear Algebra, 4th edition, Pearson Education, New Delhi.

Linear Algebra Basics

Introduction

The name – from **mat**rix **lab**oratory – says it all: MATLAB was designed from the get-go to work with matrices. It even treats individual numbers as a special type of matrix termed a *scalar*. Consequently, MATLAB is demonstrably faster at working with matrix-based models than comparable technical computing programs such as Mathematica or Maple. Indeed, the ease with which MATLAB works with matrices is a main reason for making MATLAB the basis for the laboratory component of the Introductory Mathematical Biology course.

Because of the intimate association between MATLAB and matrices, your ability to use MATLAB effectively will be significantly enhanced if you develop some degree of familiarity with a few concepts and techniques from linear algebra, including:

- diagonal and identity matrices
- matrix addition and subtraction
- the matrix transpose
- matrix multiplication
- matrix division
- powers of matrices

Other important topics from linear algebra – the eigenvalue problem, in particular – are beyond the scope of this handout, and we'll deal with those as the need arises next semester.

When you're working with matrices, keep in mind that matrices and the techniques used to work with them were developed to facilitate solving systems of linear equations such as:

$$\begin{aligned} 2x_1 + 3x_2 + 5x_3 + x_4 &= 27 \\ 0.4x_1 + 0.9x_2 + 3.3x_3 - 1.4x_4 &= 6.5 \\ -7x_1 + 12x_2 + 4x_3 - 2x_4 &= 21 \\ x_1 - 4x_3 + 3x_4 &= 1 \end{aligned}$$

In fact, solving for the values of x_1, x_2, x_3 and x_4 in such systems is what the basic techniques of linear algebra were tailor-made to do, but subsequent conceptual development of the field has led to a number of techniques that we will use extensively in our study of mathematical biology.

You should familiarize yourself with the material in this handout prior to the start of our course. I encourage you to have pencil and paper at hand and apply them liberally as you work through the derivations, examples, and sample problems for yourself. Learning math is *not* a passive endeavor!

Matrices, vectors, and scalars

A matrix is a rectangular (or square) array of numbers or other data such as strings of text arranged in rows and columns. In fact, the term array is commonly used interchangeably with matrix, although programming languages and programmers tend to use array, while mathematicians and biologists tend to use matrix. Any set of data that you'd enter into a spreadsheet can be represented with a matrix. The entries in a matrix are termed the *elements* of the matrix. A general representation of a matrix would be:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1j} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2j} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3j} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{i1} & a_{i2} & a_{i3} & \cdots & a_{ij} \end{bmatrix} = \begin{bmatrix} a_{ij} \end{bmatrix}$$

Matrices are usually represented with a bold font capital letter (although you will see other representations), and the elements of a matrix are enclosed in brackets, as above, or parentheses. The subscripts *i* and *j* respectively represent the *row* and *column* number, and the *dimension* of a matrix is the number of rows and columns. Examples of matrices include:

$\begin{bmatrix} 1 & -3 \\ 2 & 7.2 \end{bmatrix}$	a 2×2 matrix; an example of a **square matrix**.	$\begin{bmatrix} 14.3 \end{bmatrix}$	a 1×1 matrix; termed a **scalar**
$\begin{bmatrix} 6 & -3 & 1 \\ 2 & 0 & 9 \end{bmatrix}$	a 2×3 **rectangular matrix**	$\begin{bmatrix} 8 & 2.5 \\ 0 & -1.1 \\ -5 & -3 \end{bmatrix}$	A 3×2 rectangular matrix
$\begin{bmatrix} -8 & 0 & 1.4 & -1.3 \end{bmatrix}$	a 1×4 matrix; an example of a **row vector**.	$\begin{bmatrix} 9.1 \\ 2.8 \\ -0.7 \end{bmatrix}$	A 3×1 matrix; an example of a **column vector**.
$\begin{bmatrix} M & F \\ 2 & 7.2 \end{bmatrix}$	a 2×2 matrix containing both numeric and text elements	$\begin{bmatrix} 0 & 4 & 8 \\ .3 & 0 & 0 \\ 0 & .9 & 0 \end{bmatrix}$	A **Leslie matrix** model of population dynamics.

Matrices

Students have been working with matrices since Grade 9 and have been performing various operations with number arrays. However, a review of the terminology involved with matrices would be beneficial for the students.

A matrix is a rectangular array of numbers enclosed by parenthesis. Matrices are usually named using upper case letters. Some examples of matrices are:

$$A = \begin{pmatrix} -4 & 3 & -6 & 9 \end{pmatrix} \qquad B = \begin{pmatrix} 5 & -2 \\ 3 & -7 \end{pmatrix} \qquad C = \begin{pmatrix} 4 & -1 & 6 \\ 2 & 7 & -4 \\ 0 & 1 & -3 \end{pmatrix}$$

- The individual numbers in a matrix are called the elements.
- A horizontal arrangement of the numbers in a matrix is called a row.
- A vertical arrangement of the numbers in a matrix is called a column.
- The number of rows and the number of columns in a matrix is called the dimensions of a matrix.
- Any matrix that has the same number of rows as it has columns is called a square matrix.

In the following matrix $B = \begin{pmatrix} 5 & -2 \\ 3 & -7 \end{pmatrix}$ the elements are 5, -2, 3, -7. The elements in row one are 5 and -2 and those in row two are 3 and -7. The elements in column one are 5 and 3 and those in column two are -2 and -7.The dimensions of B are 2 rows by 2 columns. The dimensions are written as 2×2 and read as "two by two". Since the number of rows and the number of columns are the same, this is also a square matrix.

The operations of addition, subtraction, scalar multiplication and multiplication can be performed using matrices. Addition and subtraction can only be done if the matrices being used are of the same dimensions. These operations are done with the corresponding elements of the matrices involved. Scalar multiplication can be done on any matrix since it is simply applying the distributive property – multiply each element of the matrix by the number before the matrix. Multiplication can only be performed if the number of columns of the first matrix equals the number of rows of the second matrix. Then the process is carried out by doing row by column multiplication. Here are some examples of these operations done with matrices.

Some Special Matrices

Symmetric Matrices

The elements of a square matrix for which *i* = *j*, e.g., $a_{11,}\ a_{22,}\ a_{33}\ldots$, comprise the *diagonal* of the matrix and are termed the *diagonal elements* of the matrix, while the other elements, a_{ij} , $i \neq j$, are referred to as the *off-diagonal elements*. Thus, in the matrix

$$\mathbf{A} = \begin{bmatrix} 9 & 3 & 6 \\ 1 & 7 & 9 \\ 0 & 7 & 2 \end{bmatrix}$$

the diagonal elements are 9, 7, and 2. If the off-diagonal elements on each side of the diagonal are 'mirror images' of each other (i.e., $a_{ij} = a_{ji}$ for all *i* and *j*), as in

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 4 \\ 1 & 7 & 5 \\ 4 & 5 & 2 \end{bmatrix}$$

then the matrix is termed *symmetric*.

Diagonal Matrices

A diagonal matrix is a symmetric matrix in which the off-diagonal elements equal zero:

$$\mathbf{A} = \begin{bmatrix} \sqrt{7} & 0 & 0 \\ 0 & 14 & 0 \\ 0 & 0 & \pi \end{bmatrix}.$$

Diagonal matrices are encountered in areas such as microarray analysis, Markovian models of molecular evolution, and models of population dynamics. Most biologically relevant matrices can be *diagonalized* (converted to a diagonal form), allowing us to take advantage of some of their special properties.

The Identity Matrix

One especially important diagonal matrix is the *identity matrix*. The diagonal elements of the identity matrix are all ones, while its off-diagonal elements are all zeros. A 3×3 identity matrix thus looks like this:

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The term identity matrix is appropriate because multiplying a matrix (we'll see how to do that later) by its corresponding identity matrix yields the same matrix. I.e.,

$$\mathbf{AI} = \mathbf{A}$$

The identity matrix is thus the functional equivalent of the number one.

Working With Matrices I – Basic Matrix Manipulations

The Transpose of a Matrix

The transpose of a matrix is an important construct that you will use frequently when working with MATLAB. The transpose of matrix **A** is represented variously by $\mathbf{A}^T$, **A**′, $\mathbf{A}^{tr}$, ${}^t\mathbf{A}$, or, rarely, $\widetilde{\mathbf{A}}$. Most linear algebra texts use $\mathbf{A}^T$, while MATLAB and most research journals use **A**′, which we'll therefore generally use in this course.

Operationally, the transpose of a matrix is created by taking each of its rows and 'converting' them into the corresponding columns of the transpose matrix, meaning the first row of a matrix becomes the first column of its transpose, the second row becomes the second column, and so on. Thus, if

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix},$$

the transpose of **A** is

$$\mathbf{A}' = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}' = \begin{bmatrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{bmatrix}$$

This is handy for example, when manually entering data into a MATLAB program. While it's easier to enter data into MATLAB as row vectors, a program you're written may require data in column vector form – or you may more simply be more comfortable working with column vectors. In either case, all you do is enter the data in row vector form, then take the transpose to get your column vector.

Those of you that have previously worked with arrays in languages such as C/C++ or Java are no doubt beginning to appreciate how easy array manipulations are with MATLAB!

Practice Problems:

Calculate the transpose of each of the following:

$$\mathbf{A} = \begin{bmatrix} 6 & 11 & 0 \\ -1 & \pi & 11 \\ 0.5 & \sqrt{5} & -3 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} 4 & 1 & 3 \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} 9 & 0 & -4 \end{bmatrix}$$

$$\mathbf{D} = \begin{bmatrix} .4 & -6 & 3.1 & 2 \\ 8 & -1 & 4 & 0 \end{bmatrix}$$

$$\mathbf{E} = \begin{bmatrix} 6 & 11 & 0 \\ 11 & \pi & \sqrt{5} \\ 0 & \sqrt{5} & -3 \end{bmatrix}$$ (what kind of matrix is **E**?)

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

solutions

Concatenation of Matrices

You're no doubt accustomed to entering data into columns of an Excel™ spreadsheet, hitting the <Enter> key after each entry to drop down to a cell in the next row. In other words, you create column vectors when you enter data into most spreadsheets. In contrast, when entering data in MATLAB you will probably find that data are more readily entered as row vectors. This brings us to an important MATLAB technique: *concatenation*. Concatenation basically means joining two arrays to produce one.

Before we actually discussion concatenation, however, we need to mention another use of the term "dimension". Once you start working with MATLAB, you will soon encounter the terms "column dimension" and "row dimension". These terms are not standard linear algebra jargon, and their use in MATLAB is perhaps unfortunate, because of the usage of dimension described above. Nevertheless, the terms are useful in matrix manipulations, which can be seen by applying the MATLAB **sum** command to the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

Simply running the MATLAB command **sum(A)** will yield:

```
ans =

   12    15    18
```

Note that the **sum(A)** command returned a row vector containing the sum of the terms in each column, as a result of applying the sum operator along the "row dimension" (= dimension 1 in MATLAB-speak). To generate the row sums instead, you run the command **sum(A,2)** which causes MATLAB to sum across the "column dimension" (= dimension 2) and yields

```
ans =

    6
   15
   24
```

Note that when we ran the **sum(A)** command without specifying a dimension, MATLAB defaulted to summing along the row dimension, meaning the result we obtained was the same as would have been returned had we entered the command **sum(A,1)**. This is a general rule: MATLAB commands default to the row dimension unless you specify otherwise. During the upcoming semester, we will work with a number of other MATLAB commands that let you specify the dimension along which you wish the operation to proceed.

Ok, back to concatenation. You will find that it will help considerably when you're writing and debugging programs in MATLAB if you train yourself to think of matrices as concatenated column or row vectors. Thus, the matrix

$$\mathbf{A} = \begin{bmatrix} 3 & 4 & 9 \\ 7 & 8 & 9 \\ 4 & 2 & 0 \end{bmatrix}$$

can be thought of three row vectors stacked one above the other (= concatenated along the *row dimension*):

$$\boldsymbol{a}_1 = \begin{bmatrix} 3 & 4 & 9 \end{bmatrix}, \boldsymbol{a}_2 = \begin{bmatrix} 7 & 8 & 9 \end{bmatrix} \text{and } \boldsymbol{a}_3 = \begin{bmatrix} 4 & 2 & 0 \end{bmatrix}.$$

or as three column vectors

$$\boldsymbol{a}_1 = \begin{bmatrix} 3 \\ 7 \\ 4 \end{bmatrix}, \boldsymbol{a}_2 = \begin{bmatrix} 4 \\ 8 \\ 2 \end{bmatrix} \text{ and } \boldsymbol{a}_3 = \begin{bmatrix} 9 \\ 9 \\ 0 \end{bmatrix},$$

placed side-by-side (= concatenated along the *column dimension*). Concatenation is readily accomplished in MATLAB, and is typically used to merge complementary sets of data from different sources, say, when loading data from different Excel™ files, or when combining results of calculations that have been performed by different programs or parts of programs.

To see how concatenation works, and to get you started thinking like a MATLAB programmer, let

$$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}.$$

The MATLAB command **cat(**<dimension>**,A,B)** concatenates the two 2×2 matrices **A** and **B** along the dimension, row or column, that you specify. (unlike the **sum()** command, you *must* specify a dimension when using **cat(1,A,B)**.) Thus, **cat(1,A,B)** causes concatenation along the row dimension and yields the matrix

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \\ 7 & 8 \end{bmatrix},$$

while **cat(2,A,B)** concatenates along the column dimension and yields

$$\begin{bmatrix} 1 & 2 & 5 & 6 \\ 3 & 4 & 7 & 8 \end{bmatrix}.$$

You will have many occasions to use concatenation as you progress through our study of MATLAB and its various applications.

Working With Matrices II – Matrix Arithmetic

Matrix Addition:

Matrix addition (and subtraction) is straightforward…you just add or subtract the corresponding elements in each matrix. The only requirement is that the dimensions of each matrix must the same. Let's illustrate this by calculating the sum and difference of two 3x3 matrices:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \text{ and } \mathbf{B} = \begin{bmatrix} 4 & 5 & 6 \\ 7 & 8 & 9 \\ 1 & 2 & 3 \end{bmatrix},$$

then

$$\mathbf{C} = \mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11}+b_{11} & a_{12}+b_{12} & a_{13}+b_{13} \\ a_{21}+b_{21} & a_{22}+b_{22} & a_{23}+b_{23} \\ a_{31}+b_{31} & a_{32}+b_{32} & a_{33}+b_{33} \end{bmatrix} = \begin{bmatrix} 1+4 & 2+5 & 3+6 \\ 4+7 & 5+8 & 6+9 \\ 7+1 & 8+2 & 9+3 \end{bmatrix} = \begin{bmatrix} 5 & 7 & 9 \\ 11 & 13 & 15 \\ 8 & 10 & 12 \end{bmatrix}$$

and

$$\mathbf{D} = \mathbf{A} - \mathbf{B} = \begin{bmatrix} a_{11}-b_{11} & a_{12}-b_{12} & a_{13}-b_{13} \\ a_{21}-b_{21} & a_{22}-b_{22} & a_{23}-b_{23} \\ a_{31}-b_{31} & a_{32}-b_{32} & a_{33}-b_{33} \end{bmatrix} = \begin{bmatrix} 1-4 & 2-5 & 3-6 \\ 4-7 & 5-8 & 6-9 \\ 7-1 & 8-2 & 9-3 \end{bmatrix} = \begin{bmatrix} -3 & -3 & -3 \\ -3 & -3 & -3 \\ 6 & 6 & 6 \end{bmatrix}$$

Practice Problems

Calculate $\mathbf{A}+\mathbf{B}$ and $\mathbf{A}-\mathbf{B}$ for the following:

$$\mathbf{A} = \begin{bmatrix} 7 & 8 & 2 \\ 9 & 2 & 3 \\ 1 & 6 & 4 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 8 & 4 & 2 \\ 8 & 4 & 9 \\ 3 & 1 & 6 \end{bmatrix}$$

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

solutions

Important note: matrix addition and subtraction are examples of what MATLAB refers to as *element-wise operations*. That is, the addition or subtraction operator is applied element-by-element, and the result is a third matrix whose dimensions are identical with those of the original matrices. Other element-wise operators employed by MATLAB include:

- Scalar multiplication: $2 \cdot \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 2\cdot 1 & 2\cdot 2 \\ 2\cdot 3 & 2\cdot 4 \end{bmatrix} = \begin{bmatrix} 2 & 4 \\ 6 & 8 \end{bmatrix}$
- Matrix exponentiation: $\exp\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} e^1 & e^2 \\ e^3 & e^4 \end{bmatrix} = \begin{bmatrix} 2.7183 & 7.3891 \\ 20.0855 & 54.5982 \end{bmatrix}$
- Trigonometric functions of a matrix, e.g.: $\sin\begin{bmatrix} 0 & \pi/2 \\ \pi & 2\pi \end{bmatrix} = \begin{bmatrix} \sin 0 & \sin(\pi/2) \\ \sin\pi & \sin 2\pi \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$
- Logarithm of a matrix: $\log\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} \log 1 & \log 2 \\ \log 3 & \log 4 \end{bmatrix} = \begin{bmatrix} 0 & 0.6931 \\ 1.0986 & 1.3863 \end{bmatrix}$

Note in the last example that the log operator returned the *Naperian* or *natural logarithm* (base e = $\log_e$ where e = 2.71838...) of the matrix elements, rather than the *common* logarithm (base 10 = $\log_{10}$) that you may have been expecting. Powers of e and Naperian logarithms arise naturally from the structure of many biological models, while base 10 logarithms never do. Consequently, we will have little, if any, occasion to use base 10 logarithms and, unless otherwise specified, $\log x$, $\log y$, etc. will refer to the natural logarithm, in keeping with contemporary mathematical convention. Also note that MATLAB expects angles to be entered as radians (1 radian = $180^\circ/\pi \cong 57.3^\circ$); thus π radians equals 180°, $\pi/2$ radians equals 90°, and so on.

Matrix Multiplication:

Matrix multiplication is more complex than matrix addition or subtraction, and is carried out in accordance with a strict rule that stems directly from the fact that the technique was developed to facilitate solution of systems of linear equations. If **C** is a matrix resulting from the multiplication of two matrices, **A** and **B**, then the elements c_{ij} of **C** are given by:

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj} \,, \qquad \text{Equation 1}$$

where *n* is the number of columns in **A** *and* the number of rows in **B**. Look carefully at the subscripts of *a* and *b*, and note that Equation 1 requires that *the number of columns in the left-hand matrix* (= *i*) must be the same as *the number of rows in the right-hand matrix* (also equal to *i*). Note also that Equation 1 means that matrix multiplication is *not* an element-wise operation. That is

$$\mathbf{C} = \mathbf{AB} \neq \begin{bmatrix} a_{11}b_{11} & a_{12}b_{12} & a_{13}b_{13} \\ a_{21}b_{21} & a_{22}b_{22} & a_{23}b_{23} \\ a_{31}b_{31} & a_{32}b_{32} & a_{33}b_{33} \end{bmatrix}$$

Let's apply Equation 1 to a pair of 3×3 matrices. Let

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \text{ and } \mathbf{B} = \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix},$$

The product of **A** and **B** is then defined by Equation 1 as:

$$\mathbf{AB} = \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} & a_{11}b_{13} + a_{12}b_{23} + a_{13}b_{33} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} & a_{21}b_{13} + a_{22}b_{23} + a_{23}b_{33} \\ a_{31}b_{11} + a_{32}b_{21} + a_{33}b_{31} & a_{31}b_{12} + a_{32}b_{22} + a_{33}b_{32} & a_{31}b_{13} + a_{32}b_{23} + a_{33}b_{33} \end{bmatrix}$$

In other words, you multiply each of the elements of a *row* in the *left-hand* matrix by the corresponding elements of a *column* in the *right-hand* matrix (that's why the number of elements in the row and the column must be equal), and then sum the resulting *n* products. The choice of the row and column to be used in the multiplication and summation is based on which element of the product matrix (**C**) that you wish to calculate. Specifically:

- The left-hand matrix row you work with is the same as the row of the product matrix element you wish to calculate,
- The right-hand matrix column you work with is the same as the column of the product matrix element you wish to calculate.

Example. Calculate the product of (in this case) two 3×3 matrices, **A** and **B**:

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix} = \mathbf{AB} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \times \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix}$$

To calculate the value of c_{11},

$$\begin{bmatrix} \textcircled{c_{11}} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix}$$

you proceed as follows:

$$c_{11} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \times \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} \Rightarrow c_{11} = a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31}$$

Similarly, for c_{23}

$$\begin{bmatrix} c_{11} & c_{12} & c_{13} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{33} \end{bmatrix}$$

you do this:

$$c_{23} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \times \begin{bmatrix} b_{11} & b_{12} & b_{13} \\ b_{21} & b_{22} & b_{23} \\ b_{31} & b_{32} & b_{33} \end{bmatrix} \Rightarrow c_{23} = a_{21}b_{13} + a_{22}b_{23} + a_{23}b_{33}$$

Important: the two matrices being multiplied needn't be square, or even of the same dimension. All that's required is that the number of columns in the left-hand matrix be the same as the number of rows in the right-hand matrix.

Example: Find the product of two matrices, **A** and **B**, where

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \text{ and } \mathbf{B} = \begin{bmatrix} 4 & 5 & 6 \\ 7 & 8 & 9 \\ 1 & 2 & 3 \end{bmatrix}$$

We first note that, in accordance with Equation 1, multiplication of **A** by **B** is allowed because the number of columns in **A** equals the number of rows in **B**. The product **C** = **AB** is then calculated as follows:

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \begin{bmatrix} 4 & 5 & 6 \\ 7 & 8 & 9 \\ 1 & 2 & 3 \end{bmatrix}$$

$$= \begin{bmatrix} 1*4+2*7+3*1 & 1*5+2*8+3*2 & 1*6+2*9+3*3 \\ 4*4+5*7+6*1 & 4*5+5*8+6*2 & 4*6+5*9+6*3 \\ 7*4+8*7+9*1 & 7*5+8*8+9*2 & 7*6+8*9+9*3 \end{bmatrix}$$

$$= \begin{bmatrix} 21 & 27 & 33 \\ 57 & 72 & 87 \\ 93 & 117 & 141 \end{bmatrix}$$

Practice Problems:

For each of the following, determine whether the indicated multiplication is allowed; if so, calculate the corresponding product matrix:

a. $\begin{bmatrix} 2 & 1 & 3 \\ 1 & 8 & 3 \\ 5 & 7 & 1 \end{bmatrix}\begin{bmatrix} 1 & 6 & 1 \\ -3 & 5 & 3 \\ 2 & 4 & -9 \end{bmatrix}$

b. $\begin{bmatrix} 4 & 3 & 1 \\ 2 & 2 & 2 \\ 1 & 4 & 3 \end{bmatrix}\begin{bmatrix} 5 \\ 4 \\ 3 \end{bmatrix}$

c. $\begin{bmatrix} 5 \\ 4 \\ 3 \end{bmatrix}\begin{bmatrix} 4 & 3 & 1 \\ 2 & 2 & 2 \\ 1 & 4 & 3 \end{bmatrix}$

d. $\begin{bmatrix} 2 & 3 & 4 \\ 2 & 1 & 2 \\ 1 & 4 & 5 \end{bmatrix}\begin{bmatrix} 1.5 & -0.5 & -1 \\ 4 & -3 & -2 \\ -3.5 & 2.5 & 2 \end{bmatrix}$

e. $\begin{bmatrix} 1 & \pi & 2 & 6 \\ 2 & 5 & 1.3 & 5 \end{bmatrix}\begin{bmatrix} 5 & 3 \\ 5 & 1 \end{bmatrix}$

f. $\begin{bmatrix} 1 & 2 & 3 \end{bmatrix}\begin{bmatrix} 5 & 3 & 2 \end{bmatrix}$, $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}\begin{bmatrix} 5 \\ 3 \\ 2 \end{bmatrix}$, $\begin{bmatrix} 1 & 2 & 3 \end{bmatrix}\begin{bmatrix} 5 \\ 3 \\ 2 \end{bmatrix}$, $\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}\begin{bmatrix} 5 & 3 & 2 \end{bmatrix}$

solutions

IMPORTANT

Commutivity of Matrix Operations

This may strike you but as a bit arcane – ok, as a *lot* arcane – but it's essential that you keep in mind the concept of *commutivity*. A mathematical operation numbers is said to be commutative if the order in which the operation is carried out doesn't matter. Thus:

$3 + 4 = 4 + 3 = 7$

$3 - 4 = -4 + 3 = -1$

$3 \times 4 = 4 \times 3 = 12$

$(1/3) \times (4) = (4) \times (1/3) = 4/3$

and so on. Vector addition, subtraction, and multiplication are also commutative (as long as you use the transpose of the right-hand vector), as are matrix addition and subtraction. However – and this has important ramifications for your work with MATLAB – matrix multiplication is generally *not* commutative. In other words, in contrast with numbers or variables, if **A** and **B** are two matrices, in general $\mathbf{AB} \neq \mathbf{BA}$. Check the non-commutativity of matrix multiplication for yourself by calculating

$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 5 & 6 \\ 7 & 8 \end{bmatrix}$

$\mathbf{AB} = \mathbf{BA}$

There *are* situations in which matrix multiplication is commutative, but setup of any problem involving matrix multiplication is crucial to avoid ending up with invalid results.

Multiplication of Vectors

Example: Suppose a budding young field biologist needs to quantify the total reproductive success in a population of American kestrels (a type of falcon), and has obtained data of the number of young birds (fledglings) leaving the nest for a sample of 33 nests:

Number of Fledglings	Number of Nests
2	3
3	17
4	12
5	1

We calculate the total number of fledglings produced by the population by multiplying each entry in the "Number of Fledglings" column by the corresponding value in the "Number of Nests" column and summing the resulting four terms. The combination of multiplication of two numbers followed by summing of products hints at matrix multiplication, so let's see if we can set up the calculation as a vector multiplication problem, making it considerably easier to accomplish with MATLAB. The data can be represented by two column vectors:

$$\mathbf{F} = \begin{bmatrix} 2 \\ 3 \\ 4 \\ 5 \end{bmatrix}, \text{ and } \mathbf{N} = \begin{bmatrix} 3 \\ 17 \\ 12 \\ 1 \end{bmatrix}$$

and a simple matrix multiplication should give us our answer. But, how to set it up? he obvious approach of multiplying the two column vectors directly:

$$\mathbf{S} = \begin{bmatrix} 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} \begin{bmatrix} 3 \\ 17 \\ 12 \\ 1 \end{bmatrix}$$

won't work because of the row-column number mismatch. Can we rectify the situation? A moment's consideration suggests that taking the transpose of one or the other might work. I.e.:

$$\mathbf{S} = \begin{bmatrix} 2 & 3 & 4 & 5 \end{bmatrix} \begin{bmatrix} 3 \\ 17 \\ 12 \\ 1 \end{bmatrix} \text{ or } \mathbf{S} = \begin{bmatrix} 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} \begin{bmatrix} 3 & 17 & 12 & 1 \end{bmatrix}$$

Both are allowed by Equation 1, since in each case, the number of rows in the left-hand vector equals the number of columns in the right-hand vector. Does it matter which of the two setups we choose? Let's see:

$$\mathbf{S} = \begin{bmatrix} 2 & 3 & 4 & 5 \end{bmatrix} \begin{bmatrix} 3 \\ 17 \\ 12 \\ 1 \end{bmatrix} = 2 \cdot 3 + 3 \cdot 17 + 4 \cdot 12 + 5 \cdot 1 = 110$$

Success! And, note what we did here: we took *the transpose of the left-hand vector* in order to get it 'in the proper form' to carry out the desired multiplication.

But, what about the other combination, formed by taking the transpose of the right-hand vector…will that work, too? Equation 1 tells us it will, so let's carry out the multiplication and see if we get the desired result:

$$\mathbf{S} = \begin{bmatrix} 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} \begin{bmatrix} 3 & 17 & 12 & 1 \end{bmatrix} = \begin{bmatrix} 6 & 34 & 24 & 2 \\ 9 & 51 & 36 & 3 \\ 12 & 68 & 48 & 4 \\ 15 & 85 & 60 & 5 \end{bmatrix}$$

So, the multiplication 'works', but the result is clearly not what we wanted! (and not very useful to our budding field biologist, either)

Is there a way to tell in advance how to set up a problem involving vector multiplication? It turns out that there is, but to choose correctly, you need to know the form – scalar or matrix – that your result should take. Recall from Equation 1 that multiplication of two matrices requires that the number of rows in the left-hand matrix must equal the number of columns in the right-hand matrix. While not apparent from Equation 1 is that the dimensions of the product will be given by the 'other' to dimensions, the number of columns in the left-hand matrix and the number of rows in the right-hand matrix. In general terms:

$$\mathbf{A}_{l\times m}\mathbf{B}_{m\times n} = \mathbf{C}_{l\times n}$$

Thus, in the present case, setting the problem up as $\mathbf{F}_{1\times 4}\mathbf{N}_{4\times 1} = \mathbf{S}_{1\times 1}$ gives us the desired result (a scalar), while $\mathbf{F}_{4\times 1}\mathbf{N}_{1\times 4} = \mathbf{S}_{4\times 4}$ does not.

By the way, this was not a trivial exercise, because it illustrates the extremely important point that you need to 'think your way through' *any* problem involving matrices, to be certain that the way you have set up the problem is actually going to give you a result of the form you want. Indeed, you should get in the habit of working through *any* modeling problem in advance. An especially powerful way of doing this is to keep track of units associated with your variables as your run them through your model.

Practice Problem:

1. Suppose your assistant had entered the American kestrel data as row vectors, instead of column vectors. How would you have to set up the vector multiplication problem in order to obtain the correct result?

solution

Multiplication By the Identity Matrix

Earlier, I asserted that the identity matrix was the matrix equivalent of the number one because the result of multiplication of any matrix by its corresponding identity matrix is simply the matrix itself. I.e., for any matrix **A**,

$$\mathbf{AI} = \mathbf{A}$$

Check this assertion by multiplying each of the following matrices by its identity matrix:

$$\mathbf{A} = \begin{bmatrix} 1 & 6 & 7 \\ 2 & 5 & 8 \\ 3 & 4 & 9 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} 8 & \sqrt{\pi} & -0.4 \\ -1 & 0 & 1 \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} 8 & -1 \\ \sqrt{\pi} & 0 \\ -0.4 & 1 \end{bmatrix}$$ (question: how are **B** and **C** related?)

solutions

Now, practice your matrix multiplication by checking to see if multiplication by the identity matrix is commutative by comparing the products **AI** with **IA**, **BI** with **IB**, and **CI** with **IC**. Hint: refer to Equation 1 to see if you need to use identity matrices of different dimensions to calculate the product **IA**, **IB**, and **IC**.

Matrix Division

In terms of the way we usually think of division of numbers or functions, i.e., $\frac{9.6}{24} = 0.4$ or $\frac{x^2 + 2x + 1}{x+1} = x + 1$, matrix division ($\frac{\mathbf{A}}{\mathbf{B}}$) isn't defined. In fact, quite often you won't even find the word "division" in the index of a linear algebra textbook. To see why, try dividing one 5×5 matrix by another (as with multiplication, matrix division is *not* element-wise):

$$\mathbf{C} = \frac{\mathbf{A}}{\mathbf{B}} = \frac{\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} & a_{15} \\ a_{21} & a_{22} & a_{23} & a_{24} & a_{25} \\ a_{31} & a_{32} & a_{33} & a_{34} & a_{35} \\ a_{41} & a_{42} & a_{43} & a_{44} & a_{45} \\ a_{51} & a_{52} & a_{53} & a_{54} & a_{55} \end{bmatrix}}{\begin{bmatrix} b_{11} & b_{12} & b_{13} & b_{14} & b_{15} \\ b_{21} & b_{22} & b_{23} & b_{24} & b_{25} \\ b_{31} & b_{32} & b_{33} & b_{34} & b_{35} \\ b_{41} & b_{42} & b_{43} & b_{44} & b_{45} \\ b_{51} & b_{52} & b_{53} & b_{54} & b_{55} \end{bmatrix}}$$

I think you'll agree that it's not at all obvious how to proceed! Fortunately, the functional equivalent of matrix division *is* defined. Of course, it's even less straightforward than matrix multiplication and, as we'll see, it's not even always possible to carry out.

First recall that division of numbers or variables may be represented in three ways. Thus, division of 7 by 3 can be represented as follows:

$$\frac{7}{3} = 7\left(\frac{1}{3}\right) = 7 \cdot 3^{-1}$$

Likewise for variables:

$$\frac{x}{y} = x\left(\frac{1}{y}\right) = xy^{-1}$$

And therein lies the clue about how to proceed. With $\frac{\mathbf{A}}{\mathbf{B}}$ *per se* undefined, we take advantage of the fact that matrix multiplication *is* defined and recast the matrix division problem as a matrix multiplication problem. That is, we write the problem as:

$$\mathbf{C} = \frac{\mathbf{A}}{\mathbf{B}} = \mathbf{A}\left(\frac{1}{\mathbf{B}}\right) = \mathbf{A}\mathbf{B}^{-1}$$

where $\mathbf{B}^{-1}$ represents the *inverse* of matrix **B**. In other words, to carry out matrix 'division' we take the inverse of the divisor matrix and multiply it by the dividend matrix. Since we know how to multiply two matrices, all we need to carry out matrix division is a matrix inverse of the proper dimensions (cf. Equation 1).

Let's illustrate this with an example. Let:

$$\mathbf{A} = \begin{bmatrix} 9 & 2 \\ 3 & 7 \end{bmatrix} \text{ and } \mathbf{B} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

You don't need to know how to calculate the inverse of **B** for this course (calculating inverses of matrices larger than 2x2 is tedious, and, believe me, downright unpleasant for anything larger than 3x3), but if you want to see how the inverse of a 2x2 matrix is calculated, click here. If you're willing to accept my word for it, the inverse of **B** is:

$$\mathbf{B}^{-1} = \begin{bmatrix} -2 & 1 \\ 1.5 & -0.5 \end{bmatrix}$$

With that result in hand, straightforward matrix multiplication gives us the answer:

$$\mathbf{C} = \frac{\mathbf{A}}{\mathbf{B}} = \mathbf{A}\mathbf{B}^{-1} = \begin{bmatrix} 9 & 2 \\ 3 & 7 \end{bmatrix}\begin{bmatrix} -2 & 1 \\ 1.5 & -0.5 \end{bmatrix} = \begin{bmatrix} -15 & 8 \\ 4.5 & -0.5 \end{bmatrix}$$

Of course, we should cultivate the habit of checking our result, which we do as follows:

$$\mathbf{C} = \frac{\mathbf{A}}{\mathbf{B}} \Leftrightarrow \mathbf{C}\mathbf{B} = \mathbf{A} \Rightarrow \mathbf{C}\mathbf{B} = \begin{bmatrix} -15 & 8 \\ 4.5 & -0.5 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 9 & 2 \\ 3 & 7 \end{bmatrix} = \mathbf{A} \qquad \textit{QED.}$$

Note also that $\mathbf{B}\mathbf{C} = \begin{bmatrix} -6 & 7 \\ -27 & 22 \end{bmatrix} \neq \mathbf{A}$ (because matrix multiplication isn't commutative).

Singular Matrices

Matrix division as outlined above is readily extended to matrices of any size…'all' we need is the inverse of the divisor matrix. But, there's a catch: *not all matrices are invertible*, so not all matrices can be used as divisors. This can be illustrated by carrying out the following division:

$$\mathbf{C} = \frac{\mathbf{B}}{\mathbf{A}} = \frac{\begin{bmatrix} 9 & 2 \\ 4 & 5. \end{bmatrix}}{\begin{bmatrix} 1 & 3 \\ 1.8 & 5.4 \end{bmatrix}}$$

If we calculate the inverse of matrix **A** the result is

$$\mathbf{A}^{-1} = \frac{1}{\det \mathbf{A}}\begin{bmatrix} 5.4 & -3 \\ -1.8 & 1 \end{bmatrix} = \frac{1}{5.4 \cdot 1 - (-1.8)\cdot(-3)}\begin{bmatrix} 5.4 & -3 \\ -1.8 & 1 \end{bmatrix}$$

$$= \frac{1}{0}\begin{bmatrix} 5.4 & -3 \\ -1.8 & 1 \end{bmatrix} = \infty\begin{bmatrix} 5.4 & -3 \\ -1.8 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \infty & \infty \\ \infty & \infty \end{bmatrix}$$

where $\det \mathbf{A}$ refers to the *determinant* of matrix **A**, which is defined for a 2x2 matrix as $\det \mathbf{A} = a_{11}a_{22} - a_{21}a_{12}$. The fact that the determinant of **A** equals zero renders the product $\mathbf{C} = \mathbf{B}\mathbf{A}^{-1}$ meaningless, because all elements of **C** will equal infinity, no matter what the values of the elements of matrix **B** might be. A matrix whose inverse consists entirely of infinite elements is said to be *singular*, and the following statements are equivalent:

- Matrix **A** is singular
- $\det \mathbf{A} = 0$.
- Matrix **A** cannot be inverted.
- $\mathbf{A}^{-1}$ does not exist.
- Division by $\mathbf{A}$ is undefined.

If you tell MATLAB to divide **B** by **A** where both are dimension 2x2 and **A** is singular, you will get the following message:

```
Warning: Matrix is singular to working precision.

ans =

   Inf   Inf
   Inf   Inf
```

indicating that you instructed MATLAB to divide one matrix (**B**) by a singular matrix (**A**) whose inverse doesn't exist, yielding a 2x2 matrix whose elements are all infinite. The take-home message is that you must be wary when working with models that involve matrix division. Matrices that are close to singular will 'work' as divisors, and if your model generates a nearly singular matrix that it subsequently use as a divisor, untoward – and highly undesirable – things may happen. We'll treat this topic in great detail next semester.

Powers of Matrices

The need to take powers of a matrix (as opposed to powers of the individual elements of a matrix) arises frequently in biological models such as models of population dynamics and Markov processes. We will confine ourselves to the situation where the power is a positive integer. First recall that the n^{th} power of a number or a variable is simply the number multiplied by itself n - 1 times. Thus, $3^2 = 3 \cdot 3 = 9$, $3^3 = 3 \cdot 3 \cdot 3 = 27$, $x^2 = x \cdot x$, $x^3 = x \cdot x \cdot x$, and so on.. It's pretty much the same for positive powers of a matrix: *if* **A** *is a square matrix*, then

$$\mathbf{A}^2 = \mathbf{A} \cdot \mathbf{A}$$
$$\mathbf{A}^3 = \mathbf{A} \cdot \mathbf{A} \cdot \mathbf{A}$$

and so on. (can you see why you can't calculate powers of a rectangular matrix?) Also, as with numbers and variables:

- $\mathbf{A}^m \mathbf{A}^n = \mathbf{A}^{m+n}$
- $\left(\mathbf{A}^{\mathrm{m}}\right)^n = \mathbf{A}^{\mathrm{mn}}$
- $e^{\log(A)} = \mathbf{A}$.

Sources for more information:

If you want to learn more about linear algebra, or just would like an alternative presentation of the above topics, here are some useful URLs:

- http://www.sosmath.com/matrix/matrix.html
- http://www.math.ucdavis.edu/~daddel/linear_algebra_appl/OTHER_PAGES/other_pages.html
- http://pax.st.usm.edu/cmi/mat-linalg_html/linal.html
- A good downloadable text: ftp://joshua.smcvt.edu/pub/hefferon/book/book.pdf

- Another good online text: http://www.numbertheory.org/book/
- Wikipedia (www.wikipedia.com) is another excellent source for information on all sorts of math-related topics, although the authors of the articles typically expect of their readers a certain degree of mathematical sophistication …

Solutions to Practice Problems

Matrix transpose

$$\mathbf{A}' = \begin{bmatrix} 6 & -1 & 0.5 \\ 11 & \pi & \sqrt{5} \\ 0 & 11 & -3 \end{bmatrix} = \begin{bmatrix} 6 & 11 & 0 \\ -1 & \pi & 11 \\ 0.5 & \sqrt{5} & -3 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} 4 \\ 1 \\ 3 \end{bmatrix}$$

$$\mathbf{C}' = \begin{bmatrix} 9 \\ 0 \\ -4 \end{bmatrix}$$

$$\mathbf{D}' = \begin{bmatrix} 4 & 8 \\ -6 & -1 \\ 3.1 & 4 \\ 2 & 0 \end{bmatrix}$$

$$\mathbf{E}' = \begin{bmatrix} 6 & 11 & 0 \\ 11 & \pi & \sqrt{5} \\ 0 & \sqrt{5} & -3 \end{bmatrix}$$

$$\mathbf{I}' = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

return

Matrix addition & subtraction

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} 15 & 12 & 4 \\ 17 & 6 & 12 \\ 4 & 7 & 19 \end{bmatrix}$$

$$\mathbf{A} - \mathbf{B} = \begin{bmatrix} -1 & 4 & 0 \\ 1 & -2 & -6 \\ -2 & 5 & -2 \end{bmatrix}$$

$$\mathbf{A} + \mathbf{B} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

$$\mathbf{A}-\mathbf{B}=\begin{bmatrix}0&0&0\\0&0&0\\0&0&0\end{bmatrix}$$

return

Matrix Multiplication

a. Multiplication is allowed:

$$\text{answer}=\begin{bmatrix}5&29&-22\\-17&58&-2\\-14&69&17\end{bmatrix}$$

b. Multiplication is allowed:

$$\text{answer}=\begin{bmatrix}35\\24\\30\end{bmatrix}$$

c. Multiplication is not allowed.

d. Multiplication is allowed:

$$\text{answer}=\begin{bmatrix}1&0&0\\0&1&0\\0&0&1\end{bmatrix}$$

(what kind of matrix is this? What does this result suggest about the relationship between the two matrices?)

e. Multiplication is not allowed.

f. Multiplication is not allowed, multiplication is not allowed, 17, $\begin{bmatrix}5&3&2\\10&6&4\\15&9&6\end{bmatrix}$

return

Matrix multiplication not commutative

$$\mathbf{AB}=\begin{bmatrix}1&2\\3&4\end{bmatrix}\begin{bmatrix}5&6\\7&8\end{bmatrix}=\begin{bmatrix}19&22\\43&50\end{bmatrix}$$

$$\mathbf{BA}=\begin{bmatrix}5&6\\7&8\end{bmatrix}\begin{bmatrix}1&2\\3&4\end{bmatrix}=\begin{bmatrix}23&34\\31&46\end{bmatrix}\neq\mathbf{AB}$$

return

Multiplication of row vectors

In the case of two row vectors:

$$\mathbf{N}=\begin{bmatrix}2&3&4&5\end{bmatrix},\quad \mathbf{F}=\begin{bmatrix}3&17&12&1\end{bmatrix}$$

where the result must be a scalar, the correct setup for the problem involves taking the transpose of **F** and calculating $\mathbf{NF}'$:

$$\begin{bmatrix}2&3&4&5\end{bmatrix}\begin{bmatrix}3\\17\\12\\1\end{bmatrix}=2\cdot3+3\cdot17+4\cdot12+5\cdot1=110$$

return

Multiplication by the identity matrix

$$\mathbf{AI} = \begin{bmatrix} 1 & 6 & 7 \\ 2 & 5 & 8 \\ 3 & 4 & 9 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 6 & 7 \\ 2 & 5 & 8 \\ 3 & 4 & 9 \end{bmatrix}$$

$$\mathbf{BI} = \begin{bmatrix} 8 & \sqrt{\pi} & -0.4 \\ -1 & 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 8 & \sqrt{\pi} & -0.4 \\ -1 & 0 & 1 \end{bmatrix}$$

$$\mathbf{CI} = \begin{bmatrix} 8 & -1 \\ \sqrt{\pi} & 0 \\ -0.4 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 8 & -1 \\ \sqrt{\pi} & 0 \\ -0.4 & 1 \end{bmatrix}$$

return

Examples: Perform the indicated operations:

1. $(A)+(B)$ Given: $A = \begin{pmatrix} 4 & -3 & 6 \\ -8 & 5 & -9 \end{pmatrix}$ $B = \begin{pmatrix} -5 & 6 & -2 \\ 3 & 7 & -4 \end{pmatrix}$

Solution: $A+B = \begin{pmatrix} 4+-5 & -3+6 & 6+-2 \\ -8+3 & 5+7 & -9+-4 \end{pmatrix} \Rightarrow A+B = \begin{pmatrix} -1 & 3 & 4 \\ -5 & 12 & -13 \end{pmatrix}$

2. $(A)-(B)$ Given: $A = \begin{pmatrix} 6 & -7 \\ -4 & 5 \\ -3 & 2 \end{pmatrix}$ $B = \begin{pmatrix} -8 & 3 \\ 3 & -1 \\ 2 & -8 \end{pmatrix}$

Solution: $A-B = \begin{pmatrix} 6--8 & -7-3 \\ -4-3 & 5--1 \\ -3-2 & 2--8 \end{pmatrix} \Rightarrow A-B = \begin{pmatrix} 14 & -10 \\ -7 & 6 \\ -5 & 10 \end{pmatrix}$

3. $5\begin{pmatrix} -4 & 3 \\ 6 & -2 \end{pmatrix} \Rightarrow$ Solution: $\begin{pmatrix} -20 & 15 \\ 30 & -10 \end{pmatrix}$

4. $(A)(B)$ Given: $A = \begin{pmatrix} 6 & -2 & 3 \\ -4 & 2 & 5 \end{pmatrix}$ $B = \begin{pmatrix} 2 & -3 \\ 4 & -5 \\ 1 & -6 \end{pmatrix}$

Solution: $(A)(B) \Rightarrow \begin{pmatrix} 6\times 2+-2\times 4+3\times 1 & 6\times -3+-2\times -5+3\times -6 \\ -4\times 2+2\times 4+5\times 1 & -4\times -3+2\times -5+5\times -6 \end{pmatrix} \Rightarrow \begin{pmatrix} 7 & -26 \\ 5 & -28 \end{pmatrix}$

This review of operations with matrices should be sufficient to enable the students to recall this prior knowledge.

Associated with every square matrix is a value called the determinant. This value for a 2×2 matrix is the number that results from the difference between the products of the numbers in each diagonal of the matrix. If $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ represents any 2×2 matrix, then the determinant is the result of **ad – bc.**

The determinant of **A** can be represented as detA or as $|A|$.

Notation: $\det A = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$

Examples: Find the determinant of each of the following matrices:

1. $\begin{pmatrix} 5 & 3 \\ 7 & 2 \end{pmatrix}$ 2. $\begin{pmatrix} 6 & -4 \\ 2 & -3 \end{pmatrix}$ 3. $\begin{pmatrix} 4 & -2 \\ 3 & 5 \end{pmatrix}$ 4. $\begin{pmatrix} -3 & 8 \\ -2 & -4 \end{pmatrix}$

$(5)(2)-(7)(3)$	$(6)(-3)-(2)(-4)$	$(4)(5)-(3)(-2)$	$(-3)(-4)-(-2)(8)$
$10-21$	$-18--8$	$20--6$	$12--16$
-11	-10	26	28

The inverse of a 2×2 matrix can also be determined. If $A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$ represents any 2×2 matrix, then the inverse of **A**, written as $\mathbf{A^{-1}}$, is found by: $A^{-1} = \begin{pmatrix} \frac{d}{\det A} & \frac{-b}{\det A} \\ \frac{-c}{\det A} & \frac{a}{\det A} \end{pmatrix}$

When a 2×2 matrix is multiplied by its inverse, the result is the identity matrix $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$.

Examples: Find the inverse of each of the following matrices:

1. $A = \begin{pmatrix} 3 & -2 \\ 5 & -4 \end{pmatrix}$

Solution: The first step is to find the determinant of (A).

$|A| = -12 - -10$

$|A| = -2$

$$A^{-1}=\begin{pmatrix}\frac{-4}{-2} & \frac{2}{-2}\\ \frac{-5}{-2} & \frac{3}{-2}\end{pmatrix}$$

$$A^{-1}=\begin{pmatrix}\frac{4}{2} & \frac{-2}{2}\\ \frac{5}{2} & \frac{-3}{2}\end{pmatrix} or \begin{pmatrix}2 & -1\\ \frac{5}{2} & \frac{-3}{2}\end{pmatrix}$$

Do not leave a negative sign in the denominator of any fraction. Apply the rule for division of integers and place the resulting sign in the numerator of the fraction.
When using the inverse matrix to solve a system of linear equations, it is best to leave all the elements in fraction form. To confirm that the answer is correct, multiply the matrix by its inverse.

Check: $\begin{pmatrix}3 & -2\\ 5 & -4\end{pmatrix}\begin{pmatrix}2 & -1\\ \frac{5}{2} & \frac{-3}{2}\end{pmatrix}\Rightarrow\begin{pmatrix}6+\frac{-10}{2} & -3+\frac{6}{2}\\ 10+\frac{-20}{2} & -5+\frac{12}{2}\end{pmatrix}\Rightarrow\begin{pmatrix}6-5 & -3+3\\ 10-10 & -5+6\end{pmatrix}\Rightarrow\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$ The product of $(A)\times(A)^{-1}$ is the identity matrix. Therefore the inverse matrix is correct.

2. $A=\begin{pmatrix}2 & -6\\ 1 & 3\end{pmatrix}$

Solution: The first step is to find the determinant of (A).

$$|A|=6--6$$
$$|A|=12$$

$$A^{-1}=\begin{pmatrix}\frac{3}{12} & \frac{6}{12}\\ \frac{-1}{12} & \frac{2}{12}\end{pmatrix} \quad or A^{-1}=\begin{pmatrix}\frac{1}{4} & \frac{1}{2}\\ \frac{-1}{12} & \frac{1}{6}\end{pmatrix}$$

When using the inverse matrix to solve a system of linear equations, it is best to leave all of the elements in fraction form with the same common denominator as shown in the first inverse rather than as reduced fractions as shown in the second inverse.

Check: $\begin{pmatrix}2 & -6\\ 1 & 3\end{pmatrix}\begin{pmatrix}\frac{3}{12} & \frac{6}{12}\\ \frac{-1}{12} & \frac{2}{12}\end{pmatrix}\Rightarrow\begin{pmatrix}\frac{6}{12}+\frac{6}{12} & \frac{12}{12}-\frac{12}{12}\\ \frac{3}{12}-\frac{3}{12} & \frac{6}{12}+\frac{6}{12}\end{pmatrix}\Rightarrow\begin{pmatrix}\frac{12}{12} & \frac{0}{12}\\ \frac{0}{12} & \frac{12}{12}\end{pmatrix}\Rightarrow\begin{pmatrix}1 & 0\\ 0 & 1\end{pmatrix}$

The product of $(A)\times(A)^{-1}$ is the identity matrix. Therefore the inverse matrix is correct.

The multiplication above was done using A^{-1} in which all the elements had the common denominator 12. Therefore, the resulting products could be manipulated in order to produce the identity matrix.

3. $A = \begin{pmatrix} -10 & 4 \\ 5 & 2 \end{pmatrix}$

Solution: The first step is to find the determinant of (A).

$$|A| = -20 - 20$$
$$|A| = -40$$
$$A^{-1} = \begin{pmatrix} \frac{2}{-40} & \frac{-4}{-40} \\ \frac{-5}{-40} & \frac{-10}{-40} \end{pmatrix}$$
$$A^{-1} = \begin{pmatrix} \frac{-2}{40} & \frac{4}{40} \\ \frac{5}{40} & \frac{10}{40} \end{pmatrix} \quad or A^{-1} = \begin{pmatrix} \frac{-1}{20} & \frac{1}{10} \\ \frac{1}{8} & \frac{1}{4} \end{pmatrix}$$

Check: $\begin{pmatrix} -10 & 4 \\ 5 & 2 \end{pmatrix}\begin{pmatrix} \frac{-2}{40} & \frac{4}{40} \\ \frac{5}{40} & \frac{10}{40} \end{pmatrix} \Rightarrow \begin{pmatrix} \frac{20}{40}+\frac{20}{40} & \frac{-40}{40}+\frac{40}{40} \\ \frac{-10}{40}+\frac{10}{40} & \frac{20}{40}+\frac{20}{40} \end{pmatrix} \Rightarrow \begin{pmatrix} \frac{40}{40} & \frac{0}{40} \\ \frac{0}{40} & \frac{40}{40} \end{pmatrix} \Rightarrow \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

The product of $(A)\times(A)^{-1}$ is the identity matrix. Therefore the inverse matrix is correct.

4. $A = \begin{pmatrix} 6 & -3 \\ 5 & -2 \end{pmatrix}$

Solution: The first step is to find the determinant of (A).

$$|A| = -12 - -15$$
$$|A| = 3$$
$$A^{-1} = \begin{pmatrix} \frac{-2}{3} & \frac{3}{3} \\ \frac{-5}{3} & \frac{6}{3} \end{pmatrix} \quad or A^{-1} = \begin{pmatrix} \frac{-2}{3} & 1 \\ \frac{-5}{3} & 2 \end{pmatrix}$$

Check: $\begin{pmatrix} 6 & -3 \\ 5 & -2 \end{pmatrix}\begin{pmatrix} \frac{-2}{3} & \frac{3}{3} \\ \frac{-5}{3} & \frac{6}{3} \end{pmatrix} \Rightarrow \begin{pmatrix} \frac{-12}{3}+\frac{15}{3} & \frac{18}{3}-\frac{18}{3} \\ \frac{-10}{3}+\frac{10}{3} & \frac{15}{3}-\frac{12}{3} \end{pmatrix} \Rightarrow \begin{pmatrix} \frac{3}{3} & \frac{0}{3} \\ \frac{0}{3} & \frac{3}{3} \end{pmatrix} \Rightarrow \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

The product of $(A)\times(A)^{-1}$ is the identity matrix. Therefore the inverse matrix is correct.

Solving a 2×2 system of linear equations by using the inverse matrix method

A system of linear equations can be solved by using our knowledge of inverse matrices.
The steps to follow are:

- Express the linear system of equations as a matrix equation.
- Determine the inverse of the coefficient matrix.
- Multiply both sides of the matrix equation by the inverse matrix. In order to multiply the matrices on the right side of the equation, the inverse matrix must appear in front of the answer matrix.(the number of columns in the first matrix must equal the number of rows in the second matrix).
- Complete the multiplication. The solution will appear as:

$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$ where c_1 *and* c_2 are the solutions.

Examples: Solve the following system of linear equations by using the inverse matrix method:

1. $\left\{\begin{matrix} 2x+9y=-1 \\ 4x+y=15 \end{matrix}\right\}$

Solution: $\begin{pmatrix} 2 & 9 \\ 4 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -1 \\ 15 \end{pmatrix}$ This is the matrix equation that represents the system.

If $A = \begin{pmatrix} 2 & 9 \\ 4 & 1 \end{pmatrix}$ then $\begin{matrix} |A| = 2-36 \\ |A| = -34 \end{matrix}$

$A^{-1} = \begin{pmatrix} \frac{1}{-34} & \frac{-9}{-34} \\ \frac{-4}{-34} & \frac{2}{-34} \end{pmatrix}$ $A^{-1} = \begin{pmatrix} \frac{-1}{34} & \frac{9}{34} \\ \frac{4}{34} & \frac{-2}{34} \end{pmatrix}$

This is the determinant and the inverse of the coefficient matrix.

$$\begin{pmatrix} \frac{-1}{34} & \frac{9}{34} \\ \frac{4}{34} & \frac{-2}{34} \end{pmatrix}\begin{pmatrix} 2 & 9 \\ 4 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{-1}{34} & \frac{9}{34} \\ \frac{4}{34} & \frac{-2}{34} \end{pmatrix}\begin{pmatrix} -1 \\ 15 \end{pmatrix}$$

$$\begin{pmatrix} \frac{-2}{34}+\frac{36}{34} & \frac{-9}{34}+\frac{9}{34} \\ \frac{8}{34}+\frac{-8}{34} & \frac{36}{34}+\frac{-2}{34} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{1}{34}+\frac{135}{34} \\ \frac{-4}{34}+\frac{-30}{34} \end{pmatrix}$$

$$\begin{pmatrix} \frac{34}{34} & \frac{0}{34} \\ \frac{0}{34} & \frac{34}{34} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{136}{34} \\ \frac{-34}{34} \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 4 \\ -1 \end{pmatrix}$$ The common point or solution is (4, -1).

This is the result of multiplying the matrix equation by the inverse of the coefficient matrix.

2. $\begin{Bmatrix} 3x-6y=45 \\ 9x-5y=-8 \end{Bmatrix}$

Solution: $\begin{pmatrix} 3 & -6 \\ 9 & -5 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 45 \\ -8 \end{pmatrix}$

If $A = \begin{pmatrix} 3 & -6 \\ 9 & -5 \end{pmatrix}$ then $|A| = -15 - -54$

$|A| = 39$

$$A^{-1} = \begin{pmatrix} \frac{-5}{39} & \frac{6}{39} \\ \frac{-9}{39} & \frac{3}{39} \end{pmatrix}$$

$$\begin{pmatrix} \frac{-5}{39} & \frac{6}{39} \\ \frac{-9}{39} & \frac{3}{39} \end{pmatrix}\begin{pmatrix} 3 & -6 \\ 9 & -5 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{-5}{39} & \frac{6}{39} \\ \frac{-9}{39} & \frac{3}{39} \end{pmatrix}\begin{pmatrix} 45 \\ -8 \end{pmatrix}$$

$$\begin{pmatrix} \frac{-15}{39}+\frac{54}{39} & \frac{30}{39}+\frac{-30}{39} \\ \frac{-27}{39}+\frac{27}{39} & \frac{54}{39}+\frac{-15}{39} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}=\begin{pmatrix} \frac{-225}{39}+\frac{-48}{39} \\ \frac{-405}{39}+\frac{-24}{39} \end{pmatrix}$$

$$\begin{pmatrix} \frac{39}{39} & \frac{0}{39} \\ \frac{0}{39} & \frac{39}{39} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}=\begin{pmatrix} \frac{-273}{39} \\ \frac{-429}{39} \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}=\begin{pmatrix} -7 \\ -11 \end{pmatrix}$$ The common point or solution is (-7, -11).

In the next example, the products will be written over the common denominator instead of being written as two separate fractions.

3. $\left\{\begin{matrix} 4x+y=-13 \\ -6x-5y=37 \end{matrix}\right\}$

Solution: $\begin{pmatrix} 4 & 1 \\ -6 & -5 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}=\begin{pmatrix} -13 \\ 37 \end{pmatrix}$ If $A=\begin{pmatrix} 4 & 1 \\ -6 & -5 \end{pmatrix}$ then $|A|=-20--6$

$$|A|=-14$$

$$A^{-1}=\begin{pmatrix} \frac{-5}{-14} & \frac{-1}{-14} \\ \frac{6}{-14} & \frac{4}{-14} \end{pmatrix} \qquad A^{-1}=\begin{pmatrix} \frac{5}{14} & \frac{1}{14} \\ \frac{-6}{14} & \frac{-4}{14} \end{pmatrix}$$

$$\begin{pmatrix} \frac{5}{14} & \frac{1}{14} \\ \frac{-6}{14} & \frac{-4}{14} \end{pmatrix}\begin{pmatrix} 4 & 1 \\ -6 & -5 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}=\begin{pmatrix} \frac{5}{14} & \frac{1}{14} \\ \frac{-6}{14} & \frac{-4}{14} \end{pmatrix}\begin{pmatrix} -13 \\ 37 \end{pmatrix}$$

$$\begin{pmatrix} \frac{20+-6}{14} & \frac{5+-5}{14} \\ \frac{-24+24}{14} & \frac{-6+20}{14} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}=\begin{pmatrix} \frac{-65+37}{14} \\ \frac{78+-148}{14} \end{pmatrix}$$

$$\begin{pmatrix} \frac{14}{14} & \frac{0}{14} \\ \frac{0}{14} & \frac{14}{14} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{-28}{14} \\ \frac{-70}{14} \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -2 \\ -5 \end{pmatrix}$$

The common point or solution is (-2, -5).

4. $\left\{\begin{matrix} 3x - y = -11 \\ x + 2y = 8 \end{matrix}\right\}$

Solution: $\begin{pmatrix} 3 & -1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -11 \\ 8 \end{pmatrix}$

If $A = \begin{pmatrix} 3 & -1 \\ 1 & 2 \end{pmatrix}$ then $|A| = 6 - -1$, $|A| = 7$ $\qquad A^{-1} = \begin{pmatrix} \frac{2}{7} & \frac{1}{7} \\ \frac{-1}{7} & \frac{3}{7} \end{pmatrix}$

$$\begin{pmatrix} \frac{2}{7} & \frac{1}{7} \\ \frac{-1}{7} & \frac{3}{7} \end{pmatrix}\begin{pmatrix} 3 & -1 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{2}{7} & \frac{1}{7} \\ \frac{-1}{7} & \frac{3}{7} \end{pmatrix}\begin{pmatrix} -11 \\ 8 \end{pmatrix}$$

$$\begin{pmatrix} \frac{6+1}{7} & \frac{-2+2}{7} \\ \frac{-3+3}{7} & \frac{1+6}{7} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{-22+8}{7} \\ \frac{11+24}{7} \end{pmatrix}$$

$$\begin{pmatrix} \frac{7}{7} & \frac{0}{7} \\ \frac{0}{7} & \frac{7}{7} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{-14}{7} \\ \frac{35}{7} \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -2 \\ 5 \end{pmatrix}$$

The common point or solution is (-2, 5)

Exercises:

1. Find the determinant of each of the following matrices:

a) $\begin{pmatrix} 2 & -3 \\ -4 & -10 \end{pmatrix}$ b) $\begin{pmatrix} 7 & 3 \\ -4 & -2 \end{pmatrix}$ c) $\begin{pmatrix} 5 & -2 \\ -3 & 2 \end{pmatrix}$ d) $\begin{pmatrix} -8 & 5 \\ -4 & 3 \end{pmatrix}$

2. Find the inverse of each of the following 2×2 matrices. Check your solution by performing the operation of $(A) \times (A)^{-1}$.

a) $A = \begin{pmatrix} 4 & 3 \\ -1 & -2 \end{pmatrix}$ b) $A = \begin{pmatrix} 6 & -3 \\ 7 & -4 \end{pmatrix}$ c) $A = \begin{pmatrix} 5 & 0 \\ -4 & 2 \end{pmatrix}$ d) $A = \begin{pmatrix} -9 & 3 \\ 5 & -2 \end{pmatrix}$

Answers:

Finding the determinant

1. a) $\begin{pmatrix} 2 & -3 \\ -4 & -10 \end{pmatrix}$ b) $\begin{pmatrix} 7 & 3 \\ -4 & -2 \end{pmatrix}$ c) $\begin{pmatrix} 5 & -2 \\ -3 & 2 \end{pmatrix}$ d) $\begin{pmatrix} -8 & 5 \\ -4 & 3 \end{pmatrix}$

det = -20- + 12 det = -14- -12 det = 10- + 6 det -24- - 20

det = -32 det = -2 det = 4 det = -4

Finding the inverse matrix

1. a) $A = \begin{pmatrix} 4 & 3 \\ -1 & -2 \end{pmatrix}$ b) $A = \begin{pmatrix} 6 & -3 \\ 7 & -4 \end{pmatrix}$ c) $A = \begin{pmatrix} 5 & 0 \\ -4 & 2 \end{pmatrix}$ d) $A = \begin{pmatrix} -9 & 3 \\ 5 & -2 \end{pmatrix}$

$|A| = -8 - -3$ $|A| = -24 - -21$ $|A| = 10 - 0$ $|A| = 18 - 15$

$|A| = -5$ $|A| = -3$ $|A| = 10$ $|A| = 3$

$A^{-1} = \begin{pmatrix} \frac{-2}{-5} & \frac{-3}{-5} \\ \frac{1}{-5} & \frac{4}{-5} \end{pmatrix}$ $A^{-1} = \begin{pmatrix} \frac{-4}{-3} & \frac{3}{-3} \\ \frac{-7}{-3} & \frac{6}{-3} \end{pmatrix}$ $A^{-1} = \begin{pmatrix} \frac{2}{10} & \frac{0}{10} \\ \frac{4}{10} & \frac{5}{10} \end{pmatrix}$ $A^{-1} = \begin{pmatrix} \frac{-2}{3} & \frac{-3}{3} \\ \frac{-5}{3} & \frac{-9}{3} \end{pmatrix}$

$A^{-1} = \begin{pmatrix} \frac{2}{5} & \frac{3}{5} \\ \frac{-1}{5} & \frac{-4}{5} \end{pmatrix}$ $A^{-1} = \begin{pmatrix} \frac{4}{3} & \frac{-3}{3} \\ \frac{7}{3} & \frac{-6}{3} \end{pmatrix}$ $A^{-1} = \begin{pmatrix} \frac{1}{5} & 0 \\ \frac{2}{5} & \frac{1}{2} \end{pmatrix}$ $A^{-1} = \begin{pmatrix} \frac{-2}{3} & -1 \\ \frac{-5}{3} & -3 \end{pmatrix}$

$A^{-1} = \begin{pmatrix} 1 & -1 \\ \frac{7}{3} & -2 \end{pmatrix}$

Exercises: Solve the following systems of linear equations by using the inverse matrix method:

1. $\begin{Bmatrix} -5x+3y=21 \\ -2x+7y=-21 \end{Bmatrix}$ 2. $\begin{Bmatrix} 2x+3y=48 \\ 3x+2y=42 \end{Bmatrix}$ 3. $\begin{Bmatrix} 2x-6y=3 \\ 4x-3y=5 \end{Bmatrix}$ 4. $\begin{Bmatrix} -x+y=1 \\ -4x+2y=8 \end{Bmatrix}$

Answers:

Solving systems of linear equations using the inverse matrix method

1. $\begin{Bmatrix} -5x+3y=21 \\ -2x+7y=-21 \end{Bmatrix}$ If $A=\begin{pmatrix} -5 & 3 \\ -2 & 7 \end{pmatrix}$ then $|A|=-35--6$

$$|A|=-29$$

$$A^{-1}=\begin{pmatrix} \frac{7}{-29} & \frac{-3}{-29} \\ \frac{2}{-29} & \frac{-5}{-29} \end{pmatrix} \Rightarrow A^{-1}=\begin{pmatrix} \frac{-7}{29} & \frac{3}{29} \\ \frac{-2}{29} & \frac{5}{29} \end{pmatrix}$$

$$\begin{pmatrix} -5 & 3 \\ -2 & 7 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}=\begin{pmatrix} 21 \\ -21 \end{pmatrix}$$

$$\begin{pmatrix} \frac{-7}{29} & \frac{3}{29} \\ \frac{-2}{29} & \frac{5}{29} \end{pmatrix}\begin{pmatrix} -5 & 3 \\ -2 & 7 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}=\begin{pmatrix} \frac{-7}{29} & \frac{3}{29} \\ \frac{-2}{29} & \frac{5}{29} \end{pmatrix}\begin{pmatrix} 21 \\ -21 \end{pmatrix}$$

$$\begin{pmatrix} \frac{35+-6}{29} & \frac{-21+21}{29} \\ \frac{10-10}{29} & \frac{-6+35}{29} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}=\begin{pmatrix} \frac{-147+-63}{29} \\ \frac{-42+-105}{29} \end{pmatrix}$$

$$\begin{pmatrix} \frac{29}{29} & \frac{0}{29} \\ \frac{0}{29} & \frac{29}{29} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}=\begin{pmatrix} \frac{-210}{29} \\ \frac{-147}{29} \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix}=\begin{pmatrix} -7.24 \\ -5.07 \end{pmatrix}$$

2. $\left.\begin{matrix}2x+3y=48\\3x+2y=42\end{matrix}\right\}$ If $A=\begin{pmatrix}2 & 3\\3 & 2\end{pmatrix}$ then $|A|=4--+9$

$$|A|=-5$$

$$A^{-1}=\begin{pmatrix}\frac{2}{-5} & \frac{-3}{-5}\\ \frac{-3}{-5} & \frac{2}{-5}\end{pmatrix}\Rightarrow A^{-1}=\begin{pmatrix}\frac{-2}{5} & \frac{3}{5}\\ \frac{3}{5} & \frac{-2}{5}\end{pmatrix}$$

$$\begin{pmatrix}2 & 3\\3 & 2\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}48\\42\end{pmatrix}$$

$$\begin{pmatrix}\frac{-2}{5} & \frac{3}{5}\\ \frac{3}{5} & \frac{-2}{5}\end{pmatrix}\begin{pmatrix}2 & 3\\3 & 2\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}\frac{-2}{5} & \frac{3}{5}\\ \frac{3}{5} & \frac{-2}{5}\end{pmatrix}\begin{pmatrix}48\\42\end{pmatrix}$$

$$\begin{pmatrix}\frac{-4+9}{5} & \frac{-6+6}{5}\\ \frac{6+-6}{5} & \frac{9+-4}{5}\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}\frac{-96+126}{5}\\ \frac{144+-84}{5}\end{pmatrix}$$

$$\begin{pmatrix}\frac{5}{5} & \frac{0}{5}\\ \frac{0}{5} & \frac{5}{5}\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}\frac{30}{5}\\ \frac{60}{5}\end{pmatrix}$$

$$\begin{pmatrix}1 & 0\\0 & 1\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}=\begin{pmatrix}6\\12\end{pmatrix}$$

3. $\left.\begin{matrix}2x-6y=3\\4x-3y=5\end{matrix}\right\}$ If $A=\begin{pmatrix}2 & -6\\4 & -3\end{pmatrix}$ then $|A|=-6--24$

$$|A|=18$$

$$A^{-1}=\begin{pmatrix}\frac{-3}{18} & \frac{6}{18}\\ \frac{-4}{18} & \frac{2}{18}\end{pmatrix}$$

$$\begin{pmatrix} 2 & -6 \\ 4 & -3 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 3 \\ 5 \end{pmatrix}$$

$$\begin{pmatrix} \frac{-3}{18} & \frac{6}{18} \\ \frac{-4}{18} & \frac{2}{18} \end{pmatrix}\begin{pmatrix} 2 & -6 \\ 4 & -3 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{-3}{18} & \frac{6}{18} \\ \frac{-4}{18} & \frac{2}{18} \end{pmatrix}\begin{pmatrix} 3 \\ 5 \end{pmatrix}$$

$$\begin{pmatrix} \frac{-6+24}{18} & \frac{18+-18}{18} \\ \frac{-8+8}{18} & \frac{24+-6}{18} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{-9+30}{18} \\ \frac{-12+10}{18} \end{pmatrix}$$

$$\begin{pmatrix} \frac{18}{18} & \frac{0}{18} \\ \frac{0}{18} & \frac{18}{18} \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{21}{18} \\ \frac{-2}{18} \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1.1\overline{6} \\ -.1\overline{1} \end{pmatrix}$$

4. $\left\{\begin{matrix} -x+y=1 \\ -4x+2y=8 \end{matrix}\right\}$ If $A = \begin{pmatrix} -1 & 1 \\ -4 & 2 \end{pmatrix}$ then $|A| = -2 - -4$

$$|A| = 2$$

$$A^{-1} = \begin{pmatrix} \frac{2}{2} & \frac{-1}{2} \\ \frac{4}{2} & \frac{-1}{2} \end{pmatrix}$$

$$\begin{pmatrix} -1 & 1 \\ -4 & 2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 1 \\ 8 \end{pmatrix}$$

$$\begin{pmatrix} \frac{2}{2} & \frac{-1}{2} \\ \frac{4}{2} & \frac{-1}{2} \end{pmatrix}\begin{pmatrix} -1 & 1 \\ -4 & 2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{2}{2} & \frac{-1}{2} \\ \frac{4}{2} & \frac{-1}{2} \end{pmatrix}\begin{pmatrix} 1 \\ 8 \end{pmatrix}$$

$$\begin{pmatrix} \frac{-2+4}{2} & \frac{2+-2}{2} \\ \frac{-4+4}{2} & \frac{4+-2}{2} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{2+-8}{2} \\ \frac{4+-8}{2} \end{pmatrix}$$

$$\begin{pmatrix} \frac{2}{2} & \frac{0}{2} \\ \frac{0}{2} & \frac{2}{2} \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \frac{-6}{2} \\ \frac{-4}{2} \end{pmatrix}$$

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} -3 \\ -2 \end{pmatrix}$$

1. Modified form of Gauss Jacobi method is method.
   a) Gauss Jordan **b) Gauss Seidal**
   c) Gauss Jacobi d) Crout's
2. Forward difference operator is denoted by the symbol ------
   (a) Δ (b) ∇ (c) $\sum$ (d) $\prod$
3. If A is non singular ,its inverse is ----------
a) Adj A / |A| b) |A| / Adj A c) 1/Adj A d) 1/|A|
4. A rectangular arrangement of numbers in rows and columns is called ----------.
a) set **b) matrix** c) order d) sub matrix
5. A square matrix such that A' = A is called----------
a) symmetric b) skew symmetric c) hermitan d) scalar
6. A rectangular arrangement of numbers in rows and columns is called ----------.
a) set **b) matrix** c) order d) sub matrix
7. The addition of two matrices are possible only when they are ----------.
a) of same order b) of any order c) scalar matrices d) unit matrices
8. The number of elements in an m x n matrix is ----------
a) mn b) n c) m^2 d) n^2
9. A matrix A is said to be Orthogonal matrix if ----------.
a) $AA^T = I$ b) $AA^T = 0$ c) $A = A^T$ d) $A = 1/A$
10. A square matrix is said to be singular if its determinant is ----------
a) 0 b) 1 c) 2 d) -1
11. 2. $(A^T)^T$ = ----------.
a) A^{-1} b) A^T **c) A** d) 1 / A
12. A diagonal matrix in which all the diagonal elements are equal is called --------.
a) diagonal matrix **b) scalar matrix** c) unit matrix d) null matrix
13. [3 8 9 -2] is a row matrix of order----------
a) 4 x 1 **b) 1x 4** c) 1x 1 d) 4 x 4
14. In Iteration method if the convergence is ---------- then the convergence is of order one.
    a. a) cubic b) quadratic **c) linear** d) zero
15. When Gauss Jordan method is used to solve AX = B, A is transformed into -----
    a) scalar matrix **b) diagonal matrix**
    c) upper triangular matrix d) lower triangular matrix
16. The augment matrix is the combination of ----------.
    a) Coefficient matrix and constant matrix
    b) Unknown matrix and constant matrix
    c) Coefficient matrix and Unknown matrix
    d) Coefficient matrix, constant matrix and Unknown matrix
17. Crout's method is a ------ method.
    a. **a) direct** b) indirect c) iterative d) interpolation
18. Modified form of Gauss Jacobi method is ---------- method.
    a. a) Gauss Jordan **b) Gauss Seidal** c) Gauss Jacobi d) Crout's
19. Forward difference operator is denoted by the symbol --------
    a. **a) Δ** b) ∇ c) $\sum$ d) $\prod$
20. Shifting operator is also known as ---------- operator.
    a) translation b) central c) forward d) backward.
21. Relation between E and ∇ is ∇ = ----------

a. a) E – 1 **b) $1 - E^{-1}$** c) $1 + E^{-1}$ d) $1 * E^{-1}$

22. ---------- Method is also called as Bolzano method or interval having method.
    a. **a) Bisection** b) False position c) Newton Rapson d) Euler
23. The convergence of iteration method is ----------.
    a. a) zero b) polynomial c) quadratic **d) linear**
24. The order of convergence of Regula falsi method may be assumed to --------.
    a. a) 1.513 **b) 1.618** c) 1.234 d)1.638
25. The Newton Rapson method fails if ----------.
    a. **a) f' (x) = 0** b) f (x) = 0 c) f (x) =1 d) $f(x) \neq 0$
26. Gauss Jordan method is a ----------.
    a. **a) Direct method** b) indirect method c) iterative method d) convergent
27. The direct method fails if any one of the pivot elements become ----------.
    a. b) one **b) zero** c) two d) negative
28. In difference, f(x+h) – f(x) = --------
    a. **a) $\Delta f(x)$** b) $\nabla f(x)$ c) $\Delta^2 f(x)$ d) h(x)
29. The operators are distributive over ----------
    a) subtraction b) multiplication c) division **d) addition**
30. Relation between Δ and E is Δ = ----------
    a) E – 1 b) E + 1 c) E * 1 d) 1 – E
31. The modification of Gauss – Elimination method is called ----------.
    a. **a) Gauss Jordan** b) Gauss Seidal c) Gauss Jacobbi d) Crout's
32. Method of triangularization is also a --------- method
    a) indirect **b) direct** c) iterative d) root.
33. Example for iterative method ----------.
    a. a) Gauss elimination **b) Gauss Seidal** c) Gauss Jordon d) Bisection
34. The difference operator is denoted by --------.
    a. **a) D** b) δ c) ∇ d) Δ

KARPAGAM ACADEMY OF HIGHER EDUCATION
(Deemed to be University)
(Established under Section 3 of UGC Act, 1956)
Pollachi Main Road, Coimbatore – 641 021, Tamil Nadu

Department of Mathematics

Subject : Business Analytics	**Semester II**	**L T P C**
Subject Code : 19BPU202	**Class : I B.Com (BPS)**	**5 2 0 5**

UNIT IV –Hypothesis Testing

Constructing a hypothesis test; Null and alternate hypotheses; Test Statistic; Type I and Type II Error; Z test, t test, two sample t tests; Level of significance, Power of a test, ANOVA, Test for goodness of fit, Non-parametric tests

Suggested Readings:

1. U Dinesh Kumar (2017), Business Analytics: The Science of Data - Driven Decision Making, Wiley, New Delhi.
2. R. Evans James (2017), Business Analytics, 2nd edition, Pearson Education, New Delhi.
3. S. Christian Albright , Wayne L. Winston (2015), Business Analytics: Data Analysis and Decision Making, 5th edition, Cengage Publications
4. Howard Anton (Author), Chris Rorres (Author) (2016) , Elementary Linear Algebra with Supplemental Applications, 11 edition, Wiley, India.
5. Friedberg / Insel / Spence (2015), Linear Algebra, 4th edition, Pearson Education, New Delhi.

Hypothesis:

A statistical hypothesis is an assumption that we make about a population parameter, which may or may not be true concerning one or more variables.

According to Prof. Morris Hamburg "A hypothesis in statistics is simply a quantitative statement about a population".

Hypothesis testing:

Hypothesis testing is to test some hypothesis about parent population from which the sample is drawn.

Example:

A coin may be tossed 200 times and we may get heads 80 times and tails 120 times, we may now be interested in testing the hypothesis that the coin is unbiased.

To take another example we may study the average weight of the 100 students of a particular college and may get the result as 110lb. We may now be interested in testing the hypothesis that the sample has been drawn from a population with average weight 115lb.

Hypotheses are two types

1. Null Hypothesis
2. Alternative hypothesis

Null hypothesis:

The hypothesis under verification is known as *null hypothesis* and is denoted by H_0 and is always set up for possible rejection under the assumption that it is true.

For example, if we want to find out whether extra coaching has benefited the students or not, we shall set up a null hypothesis that *"extra coaching has not benefited the students".* Similarly, if we want to find out whether a particular drug is effective in curing malaria we will take the null hypothesis that *"the drug is not effective in curing malaria*".

Alternative hypothesis:

The rival hypothesis or hypothesis which is likely to be accepted in the event of rejection of the null hypothesis H_0 is called alternative hypothesis and is denoted by H_1 or H_a.

For example, if a psychologist who wishes to test whether or not a certain class of people have a mean I.Q. 100, then the following null and alternative hypothesis can be established.

The null hypothesis would be

$$H_0 : \mu = 100$$

Then the alternative hypothesis could be any one of the statements.

$$H_1 : \mu \neq 100$$
$$(or)\, H_1 : \mu > 100$$
$$(or)\, H_1 : \mu < 100$$

Errors in testing of hypothesis:

After applying a test, a decision is taken about the acceptance or rejection of null hypothesis against an alternative hypothesis. The decisions may be four types.

1) The hypothesis is true but our test rejects it.(type-I error)
2) The hypothesis is false but our test accepts it. .(type-II error)
3) The hypothesis is true and our test accepts it.(correct)
4) The hypothesis is false and our test rejects it.(correct)

The first two decisions are called errors in testing of hypothesis.

i.e.1) Type-I error

2) Type-II error

1) Type-I error: The type-I error is said to be committed if the null hypothesis (H_0) is true but our test rejects it.

2) Type-II error: The type-II error is said to be committed if the null hypothesis (H_0) is false but our test accepts it.

Level of significance:

The maximum probability of committing type-I error is called level of significance and is denoted by α .

α = P (Committing Type-I error)

= P (H_0 is rejected when it is true)

This can be measured in terms of percentage i.e. 5%, 1%, 10% etc…….

Power of the test:

The probability of rejecting a false hypothesis is called power of the test and is denoted by $1-\beta$.

Power of the test =P (H_0 is rejected when it is false)

= 1- P (H_0 is accepted when it is false)

= 1- P (Committing Type-II error)

= $1-\beta$

- A test for which both α and β are small and kept at minimum level is considered desirable.
- The only way to reduce both α and β simultaneously is by increasing sample size.
- The type-II error is more dangerous than type-I error.

Critical region:

A statistic is used to test the hypothesis H_0. The test statistic follows a known distribution. In a test, the area under the probability density curve is divided into two regions i.e. the region of acceptance and the region of rejection. The region of rejection is the region in which H_0 is rejected. It indicates that if the value of test statistic lies in this region, H_0 will be rejected. This region is called critical region. The area of the critical region is equal to the level of significance α. The critical region is always on the tail of the distribution curve. It may be on both sides or on one side depending upon the alternative hypothesis.

One tailed and two tailed tests:

A test with the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta \neq \theta_0$, it is called a two tailed test. In this case the critical region is located on both the tails of the distribution.

A test with the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta > \theta_0$ (right tailed alternative) or $H_1 : \theta < \theta_0$ (left tailed alternative) is called one tailed test. In this case the critical region is located on one tail of the distribution.

$H_0 : \theta = \theta_0$ against $H_1 : \theta > \theta_0$ ------- right tailed test

$H_0 : \theta = \theta_0$ against $H_1 : \theta < \theta_0$ ------- left tailed test

Sampling distribution:

Suppose we have a population of size 'N' and we are interested to draw a sample of size 'n' from the population. In different time if we draw the sample of size n, we get different samples of different observations i.e. we can get ${}^{N}c_{n}$ possible samples. If we calculate some particular statistic from each of the ${}^{N}c_{n}$ samples, the distribution of sample statistic is called sampling distribution of the statistic. For example if we consider the mean as the statistic, then the distribution of all possible means of the samples is a distribution of the sample mean and it is called sampling distribution of the mean.

Standard error:

Standard deviation of the sampling distribution of the statistic t is called standard error of t.

$$\text{i.e.} \quad \text{S.E (t)} = \sqrt{Var(t)}$$

Utility of standard error:

1. It is a useful instrument in the testing of hypothesis. If we are testing a hypothesis at 5% l.o.s and if the test statistic i.e. $|Z| = \left|\frac{t - E(t)}{S.E(t)}\right| > 1.96$ then the null hypothesis is rejected at 5% l.o.s otherwise it is accepted.
2. With the help of the S.E we can determine the limits with in which the parameter value expected to lie.
3. S.E provides an idea about the precision of the sample. If S.E increases the precision decreases and vice-versa. The reciprocal of the S.E i.e. $\frac{1}{S.E}$ is a measure of precision of a sample.
4. It is used to determine the size of the sample.

Test statistic:

The test statistic is defined as the difference between the sample statistic value and the hypothetical value, divided by the standard error of the statistic.

$$\text{i.e. test statistic } Z = \frac{t - E(t)}{S.E(t)}$$

Procedure for testing of hypothesis:

1. Set up a null hypothesis i.e. $H_0 : \theta = \theta_0$.
2. Set up a alternative hypothesis i.e. $H_1 : \theta \neq \theta_0$ or $H_1 : \theta > \theta_0$ or $H_1 : \theta < \theta_0$
3. Choose the level of significance i.e. α.
4. Select appropriate test statistic Z.
5. Select a random sample and compute the test statistic.
6. Calculate the tabulated value of Z at α % l.o.s i.e. Z_α.
7. Compare the test statistic value with the tabulated value at α % l.o.s. and make a decision whether to accept or to reject the null hypothesis.

Large sample tests:

The sample size which is greater than or equal to 30 is called as large sample and the test depending on large sample is called large sample test.

The assumption made while dealing with the problems relating to large samples are

Assumption-1: The random sampling distribution of the statistic is approximately normal.

Assumption-2: Values given by the sample are sufficiently closed to the population value and can be used on its place for calculating the standard error of the statistic.

Large sample test for single mean (or) test for significance of single mean:

For this test

The null hypothesis is $H_0 : \mu = \mu_0$

against the two sided alternative $H_1 : \mu \neq \mu_0$

where μ is population mean

μ_0 is the value of μ

Let $x_1, x_2, x_3, \ldots\ldots\ldots\ldots, x_n$ be a random sample from a normal population with mean μ and variance σ^2

i.e. if $X \sim N(\mu, \sigma^2)$ then $\bar{x} \sim N\left(\mu, \sigma^2/n\right)$, Where $\bar{x}$ be the sample mean

Now the test statistic $Z = \dfrac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{\bar{x} - E(\bar{x})}{S.E(\bar{x})} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H_0

If $|Z| < Z_\alpha$, accept the null hypothesis H_0

Note: if the population standard deviation is unknown then we can use its estimate s, which will be calculated from the sample. $s = \sqrt{\dfrac{1}{n-1}\sum (x - \bar{x})^2}$.

Large sample test for difference between two means:

If two random samples of size n_1 and n_2 are drawn from two normal populations with means μ_1 and μ_2, variances σ_1^2 and σ_2^2 respectively

Let $\bar{x}_1$ and $\bar{x}_2$ be the sample means for the first and second populations respectively

Then $\bar{x}_1 \sim N\left(\mu_1, {\sigma_1}^2 / n_1\right)$ and $\bar{x}_2 \sim N\left(\mu_2, {\sigma_2}^2 / n_2\right)$

Therefore $\bar{x}_1 - \bar{x}_2 \sim N\left(\mu_1 - \mu_2, \frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}\right)$

For this test

The null hypothesis is $H_0 : \mu_1 = \mu_2 \Rightarrow \mu_1 - \mu_2 = 0$

against the two sided alternative $H_1 : \mu_1 \neq \mu_2$

Now the test statistic $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{(\bar{x}_1 - \bar{x}_2) - E(\bar{x}_1 - \bar{x}_2)}{S.E(\bar{x}_1 - \bar{x}_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{S.E(\bar{x}_1 - \bar{x}_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{{\sigma_1}^2}{n_1} + \frac{{\sigma_2}^2}{n_2}}} \sim N(0,1) \qquad [\text{since } \mu_1 - \mu_2 = 0 \text{ from } H_0]$$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H_0

If $|Z| < Z_\alpha$, accept the null hypothesis H_0

Note: If ${\sigma_1}^2$ and ${\sigma_2}^2$ are unknown then we can consider ${S_1}^2$ and ${S_2}^2$ as the estimate value of ${\sigma_1}^2$ and ${\sigma_2}^2$ respectively..

Large sample test for single standard deviation (or) test for significance of standard deviation:

Let $x_1, x_2, x_3, \ldots\ldots\ldots\ldots, x_n$ be a random sample of size n drawn from a normal population with mean μ and variance σ^2,

for large sample, sample standard deviation s follows a normal distribution with mean σ and variance $\sigma^2/2n$ i.e. $s \sim N\left(\sigma, \sigma^2/2n\right)$

For this test

The null hypothesis is $H_0 : \sigma = \sigma_0$

against the two sided alternative $H_1 : \sigma \neq \sigma_0$

Now the test statistic $Z = \dfrac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{s - E(s)}{S.E(s)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{s - \sigma}{\sigma/\sqrt{2n}} \sim N(0,1)$$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H_0

If $|Z| < Z_\alpha$, accept the null hypothesis H_0

Large sample test for difference between two standard deviations:

If two random samples of size n_1 and n_2 are drawn from two normal populations with means μ_1 and μ_2, variances σ_1^2 and σ_2^2 respectively

Let s_1 and s_2 be the sample standard deviations for the first and second populations respectively

Then $s_1 \sim N\left(\sigma_1, {\sigma_1}^2 / 2n_1\right)$ and $\bar{x}_2 \sim N\left(\sigma_2, {\sigma_2}^2 / 2n_2\right)$

Therefore $s_1 - s_2 \sim N\left(\sigma_1 - \sigma_2, \frac{{\sigma_1}^2}{2n_1} + \frac{{\sigma_2}^2}{2n_2}\right)$

For this test

The null hypothesis is $H_0 : \sigma_1 = \sigma_2 \Rightarrow \sigma_1 - \sigma_2 = 0$

against the two sided alternative $H_1 : \sigma_1 \neq \sigma_2$

Now the test statistic $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{(s_1 - s_2) - E(s_1 - s_2)}{S.E(s_1 - s_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{(s_1 - s_2) - (\sigma_1 - \sigma_2)}{S.E(s_1 - s_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{(s_1 - s_2)}{\sqrt{\frac{{\sigma_1}^2}{2n_1} + \frac{{\sigma_2}^2}{2n_2}}} \sim N(0,1) \qquad [\text{since } \sigma_1 - \sigma_2 = 0 \text{ from } H_0]$$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H_0

If $|Z| < Z_\alpha$, accept the null hypothesis H_0

Large sample test for single proportion (or) test for significance of proportion:

Let x is number of success in n independent trails with constant probability p, then x follows a binomial distribution with mean np and variance npq.

In a sample of size n let x be the number of persons processing a given attribute then the sample proportion is given by $\hat{p} = \frac{x}{n}$

Then $E(\hat{p}) = E\left(\frac{x}{n}\right) = \frac{1}{n}E(x) = \frac{1}{n}np = p$

And $V(\hat{p}) = V\left(\frac{x}{n}\right) = \frac{1}{n^2}V(x) = \frac{1}{n^2}npq = \frac{pq}{n}$

$S.E(\hat{p}) = \sqrt{\frac{pq}{n}}$

For this test

The null hypothesis is $H_0 : p = p_0$

against the two sided alternative $H_1 : p \neq p_0$

Now the test statistic $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{\hat{p} - E(\hat{p})}{S.E(\hat{p})} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}} \sim N(0,1)$$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H_0

If $|Z| < Z_\alpha$, accept the null hypothesis H_0

Large sample test for single proportion (or) test for significance of proportion:

let x_1 and x_2 be the number of persons processing a given attribute in a random sample of size n_1 and n_2 then the sample proportions are given by $\hat{p}_1 = \frac{x_1}{n_1}$ and $\hat{p}_2 = \frac{x_2}{n_2}$

Then $E(\hat{p}_1) = p_1$ and $E(\hat{p}_2) = p_2 \Rightarrow E(\hat{p}_1 - \hat{p}_2) = p_1 - p_2$

And $V(\hat{p}_1) = \frac{p_1 q_1}{n_1}$ and $V(\hat{p}_2) = \frac{p_2 q_2}{n_2} \Rightarrow V(\hat{p}_1 - \hat{p}_2) = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$

$$S.E(\hat{p}_1) = \sqrt{\frac{p_1 q_1}{n_1}} \text{ and } S.E(\hat{p}_2) = \sqrt{\frac{p_2 q_2}{n_2}} \Rightarrow S.E(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$$

For this test

The null hypothesis is $H_0 : p_1 = p_2$

against the two sided alternative $H_1 : p_1 \neq p_2$

Now the test statistic $Z = \frac{t - E(t)}{S.E(t)} \sim N(0,1)$

$$= \frac{\hat{p}_1 - \hat{p}_2 - E(\hat{p}_1 - \hat{p}_2)}{S.E(\hat{p}_1 - \hat{p}_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{S.E(\hat{p}_1 - \hat{p}_2)} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \sim N(0,1)$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{pq}{n_1} + \frac{pq}{n_2}}} \sim N(0,1) \quad \text{Since } p_1 = p_2 \text{ from } H_0$$

$$\Rightarrow Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \sim N(0,1)$$

When p is not known p can be calculated by $p = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$ and $q = 1 - p$

Now calculate $|Z|$

Find out the tabulated value of Z at α % l.o.s i.e. Z_α

If $|Z| > Z_\alpha$, reject the null hypothesis H_0

If $|Z| < Z_\alpha$, accept the null hypothesis H_0

- **As σ is unknown,**

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} = \left[\bar{x} - z_{\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2} \frac{s}{\sqrt{n}} \right]$$

Step 2: If μ_0 falls into the above confidence intervals, then

do *not* reject H_0. Otherwise, reject H_0.

Example 1:

The average starting salary of a college graduate is $19000 according to government's report. The average salary of a random sample of 100 graduates is $18800. The standard error is 800.

(a) Is the government's report reliable as the level of significance is 0.05.
(b) Find the p-value and test the hypothesis in (a) with the level of significance $\alpha = 0.01$.
(c) The other report by some institute indicates that the average salary is $18900. Construct a 95% confidence interval and test if this report is reliable.

[solutions:]

(a)

$$H_0 : \mu = \mu_0 = 19000 \text{ vs. } H_a : \mu \neq \mu_0 = 19000,$$
$$n = 100, \bar{x} = 18800, s = 800, \alpha = 0.05$$

Then,

$$|z| = \left| \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \right| = \left| \frac{18800 - 19000}{800/\sqrt{100}} \right| = |-2.5| = 2.5 > z_{\alpha/2} = z_{0.025} = 1.96.$$

Therefore, reject H_0.

(b)

$$\text{p-value} = P(|Z| > |z|) = P(|Z| > 2.5) = 2 \cdot P(Z > 2.5) = 0.0124 > 0.01$$

Therefore, ***not*** reject H_0.

(c)

$$H_0 : \mu = \mu_0 = 18900 \text{ vs } H_a : \mu \neq \mu_0 = 18900,$$

A 95% confidence interval is

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} = 18800 \pm z_{0.025} \frac{800}{\sqrt{100}} = 18800 \pm 1.96 \cdot 80 = [18643.2, 18956.8].$$

Since $\mu_0 = 18900 \in [18643.2, 18956.8]$, Therefore, ***not*** reject H_0 **.**

Example 2:

A sample of 49 provides a sample mean of 38 and a sample standard deviation of 7. Let $\alpha = 0.05$. Please test the hypothesis

$$H_0 : u = 40 \text{ vs. } H_a : u \neq 40.$$

based on

(a) classical hypothesis test
(b) p-value
(c) confidence interval.
[solution:]

$$\bar{x} = 38,\ s = 7,\ u_0 = 40,\ n = 49,\ z = \frac{\bar{x} - u_0}{s/\sqrt{n}} = \frac{38 - 40}{7/\sqrt{49}} = -2.$$

(a)

$$|z| = 2 > 1.96 = z_{0.025}$$

we reject H_0.

(b)

$$p - value = P(|Z| > |z|) = P(|Z| > 2) = 2 * (1 - 0.9772) = 0.0456 < 0.05 = \alpha$$

we reject H_0.

(c)

$100\times(1-\alpha)\%=95\%$ confidence interval is

$$\bar{x}\pm z_{\alpha/2}\frac{s}{\sqrt{n}}=38\pm z_{0.025}\frac{7}{\sqrt{49}}=38\pm1.96=\left[36.04,39.96\right].$$

Since $40\notin\left[36.04,39.96\right]$, we reject H_0.

Hypothesis Testing for the Mean (Small Samples)

For samples of size less than 30 and when σ is unknown, if the population has a normal, or nearly normal, distribution, the *t*-distribution is used to test for the mean μ.

Using the t-Test for a Mean μ when the sample is small		
Procedure	**Equations**	**Example 4**
State the claim mathematically and verbally. Identify the null and alternative hypotheses	State H_0 and H_a	$H_0:\mu\geq16500$ $H_a:\mu<16500$ $n=14,\bar{x}=15700,s=1250$
Specify the level of significance	Specify α	$\alpha=0.05$
Identify the degrees of freedom and sketch the sampling distribution	$d.f=n-1$	$d.f.=13$
Determine any critical values. If test is left tailed, use One tail, α column with a negative sign. If test is right tailed, use One tail, α column with a positive sign. If test is two tailed, use Two tails, α column with a negative and positive sign.	Table 5 (*t*-distribution) in appendix B	The test is left-tailed. Since test is left tailed and $d.f=13$, the critical value is $t_0=-1.771$
Determine the rejection regions.	The rejection region is $t<t_0$	The rejection region is $t<-1.771$
Find the standardized test statistic	$t=\frac{\bar{x}-\mu}{\sigma_{\bar{x}}}\approx\frac{\bar{x}-\mu}{s/\sqrt{n}}$	$t=\frac{15700-16500}{1250/\sqrt{14}}\approx-2.39$
Make a decision to reject or fail to reject the null hypothesis	If t is in the rejection region, reject H_0, Otherwise do not reject H_0	Since $-2.39<-1.771$, reject H_0

Interpret the decision in the context of the original claim.		Reject claim that mean is at least 16500.

Chi-Square Tests and the F-Distribution

Goodness of Fit

DEFINITION A **chi-square goodness-of-fit test is** used to test whether a frequency distribution fits an expected distribution.

The test is used in a multinomial experiment to determine whether the number of results in each category fits the null hypothesis:

H_0 : The distribution fits the proposed proportions

H_1 : The distribution differs from the claimed distribution.

To calculate the test statistic for the chi-square goodness-of-fit test, you can use observed frequencies and expected frequencies.

DEFINITION The **observed frequency O** of a category is the frequency for the category observed in the sample data.

The **expected frequency E** of a category is the calculated frequency for the category. Expected frequencies are obtained assuming the specified (or hypothesized) distribution. The expected frequency for the *i*th category is

$$E_i = np_i$$

where *n* is the number of trials (the sample size) and p_i is the assumed probability of the *i*th category.

The Chi-square Goodness of Fit Test: The sampling distribution for the goodness-of-fit test is a chi-square distribution with $k-1$ degrees of freedom where k is the number of categories. The test statistic is

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where *O* represents the observed frequency of each category and *E* represents the expected frequency of each category. To use the chi-square goodness of fit test, *the following must be true*

.

1. The observed frequencies must be obtained using a random sample.
2. The expected frequencies must be ≥ 5.

Performing the Chi-Square Goodness-of-Fit Test (p 496)		
Procedure	Equations	Example **(p 497)**
Identify the claim. State the null and alternative hypothesis.	State H_0 and H_1	H_0: Classical 4% Country 36% Gospel 11% Oldies 2% Pop 18% Rock 29%
Specify the significance level	Specify α	$\alpha = 0.01$
Determine the degrees of freedom	d.f. = #categories - 1	$d.f. = 6-1 = 5$
Find the critical value	χ^2_α: Obtain from Table 6 Appendix B	$\varphi^2_{0.01}(d.f = 5) = 15.086$
Identify the rejection region	$\chi^2 \geq \chi^2_\alpha$	$\chi^2 \geq 15.086$
Calculate the test statistic	$\chi^2 = \sum \frac{(O-E)^2}{E}$	Survey results, n = 500 Classical O= 8 E = .04*500 = 20 Country O = 210 E = .36*500 = 180 Gospel O = 7 E = .11*500 = 55 Oldies O = 10 E = .02*500 = 10 Pop O = 75 E = .18*500 = 90 Rock O= 125 E = .29*500 = 145 Substituting $\chi^2 = 22.713$
Make the decision to reject or fail to reject the null hypothesis	Reject if χ^2 is in the rejection region Equivalently, we reject if the P-value (the probability of getting as extreme a value or more extreme) is $\leq \alpha$	Since 22.713 > 15.086 we reject the null hypothesis Equivalently $P(X \geq 22.713) < 0.01$ so reject the null hypothesis. (Note Table 6 of Appendix B doesn't have a value less than 0.005.)
Interpret the decision in the context of the original claim		Music preferences differ from the radio station's claim.

Using Minitab to perform the Chi-Square Goodness-of-Fit Test (Manual p 237)

The data from the example above (Example 2 p 497) will be used.

Enter Three columns: Music Type: Classical, etc, Observed: 8 etc, Distribution 0.04, etc. (Note the names of the columns 'Music Types', 'Observed' and 'Distribution' are entered in the gray row at the top.)

Select **Calc->Calculator**, **Store the results in** C4, and calculate the **Expression** C3*500, click **OK**, Name C4 'Expected' since it now contains the expected frequencies

Music Type	Observed	Distribution	Expected
Classical	8	0.04	20
Country	210	0.36	180
Gospel	72	0.11	55
Oldies	10	0.02	10
Pop	75	0.18	90
Rock	125	0.29	145

Next calculate the chi-square statistic, $(O-E)^2/E$ as follows: Click **Calc->Calculator**. **Store the results in** C5 and calculate the **Expression** (C2-C4)**2/C4. Click on **OK** and C5 should contain the calculated values.

7.2000
5.0000
41.8909
0.0000
2.5000
2.7586

Next add up the values in C5 and the sum is the test statistic as follows: Click on **Calc->Column Statistics**. Select **Sum** and enter C5 for the **Input Variable**. Click OK. The chi-square statistic is displayed in the session window as follows:

Sum of C5

```
Sum of C5 = 22.7132
```

Next calculate the P-value: Click on **Calc->Probability Distributions->Chi-square**. Select **Cumulative Probability** and enter 5 **Degrees of Freedom** Enter the value of the test statistic `22.7132` for the **Input Constant**. Click **OK**.

The following is displayed on the Session Window.

Cumulative Distribution Function

```
Chi-Square with 5 DF

      x   P( X <= x )
22.7132     0.999617
```

$P(X \leq 22.7132) = 0.999617$ So the P-value = 1 – 0.999617 = 0.000383. This is less that α = 0.01 so we reject the null hypothesis.

Instead of calculating the P-value, we could have found the critical value from the Chi-Square table (Table 6 Appendix B) for 5 degrees of freedom as we did above. The value is 15.086, and since our test statistic is 22.7132, we reject the null hypothesis.

Chi-Square with M&M's

H_0: Brown: 13%, Yellow: 14%, Red: 13%, Orange: 20%, Green 16%, Blue 24%
Significance level: α = 0.05
Degrees of freedom: number of categories – 1 = 5
Critical Value: $\chi^2_{0.05}(d.f.=5)=11.071$
Rejection Region: $\chi^2 \geq 11.071$
Test Statistic: $\chi^2 = \sum \frac{(O-E)^2}{E}$, where *O* is the actual number of M&M's of each color in the bag and *E* is the proportions specified under H_0 times the total number.
Reject H_0 if the test statistic is greater than the critical value (1.145)

Section 10.2 Independence

This section describes the chi-square test for independence which tests whether two random variables are independent of each other.

DEFINTION An ***r x c* contingency table** shows the observed frequencies for the two variables. The observed frequencies are arranged in r rows and c columns. The intersection of a row and a column is called a **cell.**

The following is a contingency table for two variables A and B where f_{ij} is the frequency that A equals A_i and B equals B_j.

	A_1	A_2	A_3	A_4	A
B_1	f_{11}	f_{12}	f_{13}	f_{14}	$f_{1.}$
B_2	f_{21}	f_{22}	f_{23}	f_{24}	$f_{2.}$
B_3	f_{31}	f_{32}	f_{33}	f_{34}	$f_{3.}$
B	$f_{.1}$	$f_{.2}$	$f_{.3}$	$f_{.4}$	f

If A and B are independent, we'd expect

$$f_{ij} = prob(A = A_i) * prob(B = B_j) * f = \left(\frac{f_{i.}}{f}\right)\left(\frac{f_{.j}}{f}\right) f = \frac{(f_{i.})(f_{.j})}{f}$$

$$\frac{(sum of\ row\ i) * (sum of\ column\ j)}{sample\ size}$$ (

Example 1 Determining the expected frequencies of CEO's ages as a function of company size under the assumption that age is independent of company size.

	<= 39	40 - 49	50 - 59	60 - 69	>= 70	Total
Small/midsize	42	69	108	60	21	300
Large	5	18	85	120	22	250
Total	47	87	193	180	43	550

	<= 39	40 - 49	50 - 59	60 - 69	>= 70	Total
Small/midsize	$\frac{300*47}{550}$ ≈ 25.64	$\frac{300*87}{550}$ ≈ 47.45	$\frac{300*193}{550}$ ≈ 105.27	$\frac{300*180}{550}$ ≈ 98.18	$\frac{300*43}{550}$ ≈ 23.45	300

Large	$\frac{250*47}{550} \approx 21.36$	$\frac{250*87}{550} \approx 39.55$	$\frac{250*193}{550} \approx 87.73$	$\frac{250*180}{550} \approx 81.82$	$\frac{250*43}{550} \approx 19.55$	250
Total	47	87	193	180	43	550

After finding the expected frequencies under the assumption that the variables are independent, you can test whether they are independent using the chi-square independence test.

DEFINITION A **chi-square independence test** is used to test the independence of two random variables. Using a chi-square test, you can determine whether the occurrence of one variable affects the probability of occurrence of the other variable.

To use the test,

1. The observed frequencies must be obtained from a random sample
2. Each expected frequency must be ≥ 5

The sampling distribution for the test is a chi-square distribution with

$(r-1)(c-1)$

degrees of freedom, where r and c are the number of rows and columns, respectively, of the contingency table. The test statistic for the chi-square independence test is

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

where O represents the observed frequencies and E represents the expected frequencies.

To begin the test we state the null hypothesis that the variables are independent and the alternative hypothesis that they are dependent.

Performing a Chi-Square Test for Independence (p 507)		
Procedure	Equations	Example2 **(p 507)**
Identify the claim. State the null and alternative hypotheses.	State H_0 and H_1	H_0: CEO's ages are independent of company size H_1: CEO's ages are dependent on company size.
Specify the level of significance	Specify α	$\alpha = 0.01$
Determine the degrees of freedom	$d.f. = (r-1)(c-1)$	$d.f. = (2-1)(5-1) = 4$

Find the critical value.	χ^2_α :Obtain from Table 6, Appendix B	$\chi^2_\alpha \geq 13.277$
Identify the rejection region	$\chi^2 \geq \chi^2_\alpha$	$\chi^2 \geq 13.277$
Calculate the test statistic	$\chi^2 = \sum \frac{(O-E)^2}{E}$	$\sum \frac{(O-E)^2}{E} \approx 77.9$ Note that O is in the table of actual CEO's ages above, and E is in the table of Expected CEO's ages (if independent of size) above
Make a decision to reject or fail to reject the null hypothesis	Reject if χ^2 is in the rejection region. Equivalently, we reject if the P-value (the probability of getting as extreme a value or more extreme) is $\leq \alpha$	Since 77.9 > 13.277 we reject the null hypothesis Equivalently $P(X \geq 77.0) < \alpha$ so reject the null hypothesis. (Note Table 6 of Appendix B doesn't have a value less than 0.005.)
Interpret the decision in the context of the original claim		CEO's ages and company size are dependent.

The test statistic (77.887) is greater than the critical value obtained from Table 6, Appendix B (13.277) so the null hypothesis is rejected. (Alternatively the P-Value (0.000) is less than the level of significance, α (0.01) so the null hypothesis is rejected.)

3. An urban geographer randomly samples 20 new residents of a neighborhood to determine their ratings of local bus service. The scale used is as follows: 0–very dissatisfied, 1–dissatisfied, 2– neutral, 3–satisfied, 4–very satisfied. The 20 responses are 0,4,3, 2,2,1,1,2,1,0,01,2,1,3,4,2,0,4,1. Use the sign test to see whether the population median is 2.

Solution:

There are 5 observations above the hypothesized median. Because the sample size is larger than 10, we test using the sample proportion $p = 5/20 = 0.25$. Using the PROB-VALUE method the steps in this test are:

1) H_0: $\pi = 0.5$ and H_A: π ¹ 0.5
2) We will use the *Z*-distribution

3) We will use the 5%-level, thus $\alpha = 0.05$
4) The test statistic is $z = (0.25 - 0.5)/\sqrt{0.25/20} = -2.24$
5) Table A-4 shows that $P(|Z| > 2.24) \approx 0.025$.
6) Because PROB-VALUE $< \alpha$, we reject H_0. We conclude π is different than 0.5, and thus the median is different than 2.

4. A course in statistical methods was team-taught by two instructors, Professor Jovita Fontanez and Professor Clarence Old. Professor Fontanzez used many active learning techniques, whereas Old employed traditional methods. As part of the course evaluation, students were asked to indicate their instructor preference. There was reason to think students would prefer Fontanez, and the sample obtained was consistent with that idea: of the 12 students surveyed, 8 preferred Professor Fontanez and 2 preferred Professor Old. The remaining students were unable to express a preference. Test the hypothesis that the students prefer Fontanez. (*Hint:* Use the sign test.)

Solution:

Although the sample is large enough for a normal approximation, we will use the binomial distribution to illustrate its application. Of the 12 observations, 8 preferred Prof. Fontanez, thus we need the probability of observing 8 or more successes in 12 trials of a Bernoulli process with the probability of success equal to 0.5. From Table A-1, we get

$$P(X \geq 8) = P(X = 8) + P(X = 9) + P(X = 10) + P(X = 11) + P(X = 12)$$
$$P(X \geq 8) = 0.1208 + 0.0537 + 0.0161 + 0.0029 + 0.002 = 0.1937$$

Adopting the 5% uncertainty level, we see that PROB-VALUE $> \alpha$. Thus we fail to reject H_0. We cannot conclude students prefer Fontanez.

5. Use the data in Table 10-8 to perform two Mann–Whitney tests: (a) compare uncontrolled intersections and intersections with yield signs, and (b) compare uncontrolled intersections and intersections with stop signs.

Solution:

(a) The rank sums are 119.5 and 90.5 for the yield-signed and uncontrolled intersections respectively. Given the small sample size, we use an exact test rather than the normal approximation. The associated PROB-VALUE is

0.272. Adopting a 5% level of uncertainty, we fail to reject the hypothesis of no difference. We cannot conclude the samples were drawn from different populations.

(b) The rank sums are 130.5 and 59.5 for the stop-signed and uncontrolled intersections respectively. Given the small sample size, we use an exact test rather than the normal approximation. The associated PROB-VALUE is 0.013. Adopting a 5% level of uncertainty, we reject the hypothesis of no difference. We conclude the samples were drawn from different populations.

6. Solid-waste generation rates measured in metric tons per household per year are collected randomly in selected areas of a township. The areas are classified as high-density, low density, or sparsely settled. It is thought that generation rates probably differ because of differences in waste collection and opportunities for on-site storage. Do the following data support this hypothesis?

High Density	Low Density	Sparsely Settled
1.84	2.04	1.07
3.06	2.28	2.31
3.62	4.01	0.91
4.91	1.86	3.28
3.49	1.42	1.31

Solution:

We will use the multi-sample Kruskal-Wallis test with an uncertainly level $\alpha = 0.1$. The null hypothesis is that all samples have been drawn from the same population. The rank sums are 55, 39 and 26 for the high density, low density, and sparsely settled samples respectively. The Kruskal-Wallis statistic is

$$H = \frac{12}{15(15+1)}\left(\frac{55^2}{5}+\frac{39^2}{5}+\frac{26^2}{5}\right) - 3(15+1) = 4.22$$

Using the $\chi 2$ distribution with $3-1=2$ degrees of freedom, the associated PROB-VALUE is 0.121. We fail to reject the null hypothesis. The sample does not support the hypothesis of differing waste generation rates.

7. The distances travelled to work by a random sample of 12 people to their places of work in 1996 and again in 2006 are shown in the following table.

	Distance (km)			Distance (km)	
Person	1996	2006	Person	1996	2006
1	8.6	8.8	7	7.7	6.5
2	7.7	7.1	8	9.1	9
3	7.7	7.6	9	8	7.1
4	6.8	6.4	10	8.1	8.8
5	9.6	9.1	11	8.7	7.2
6	7.2	7.2	12	7.3	6.4

Has the length of the journey to work changed over the decade?

Solution:

The sample can be considered as twelve paired observations. By taking differences between paired values, we get measures of the change for each individual. If the median change for the population is zero, we expect a sample to have a median difference near zero. Thus we will do a sign test for the median difference with a hypothesized value of zero. In other words, the hypotheses are $H_0 : \eta = 0$ and $H_A : \eta \neq 0$. We denote samples values whose distance decreased with a minus sign. Sample values with a positive difference get a plus sign. The sample becomes

S = {-,+,+,+,+,0,+,-,+,-,+,+}

Ignoring the tie, this is a sample of size 11 with 8 values above the hypothesized median. We are using Format (C) of Table 10-1, thus the PROB-VALUE is $2P(X \geq 8)$ where X is a binomial variable with $\pi = 0.5$. From the equation for the binomial, the PROB-VALUE is found to be 0.113. At the $\alpha = 10\%$ level, we fail the reject the null hypothesis. We cannot conclude there has been a change in distance.

8. One hundred randomly sampled residents of a city subject to periodic flooding are classified according to whether they are on the floodplain of the major river bisecting the

city or off the floodplain. These households are then surveyed to determine whether they currently have flood insurance of any kind. The survey results are as follows:

	On the Floodplain	Off the Floodplain
Insured	50	10
No Insurance	15	25

Test a relevant hypothesis.

Solution:

We will do a χ^2 test for a relationship between insurance and house location. The null hypothesis is no relationship (independence). Augmenting the data with expected frequencies, we have:

	On the Floodplain	Off the Floodplain
Insured	50 (39)	10 (21)
No Insurance	15 (26)	25 (14)

The corresponding χ^2 value is 22.16. Table A-8 shows that with 1 degree of freedom, $P(\chi^2 > 20)$ is zero to 3 decimal places. Thus for any reasonable level of uncertainty (any $\alpha < 0.0005$), we can reject the null hypothesis.

9. The occurrence of sunshine over a 30-day period was calculated as the percentage of time the sun was visible (i.e., not obscured by clouds). The daily percentages were:

Day	Percentage of sunshine	Day	Percentage of sunshine	Day	Percentage of sunshine
1	75	11	21	21	77
2	95	12	96	22	100
3	89	13	90	23	90
4	80	14	10	24	98
5	7	15	100	25	60

6	84	16	90	26	90
7	90	17	6	27	100
8	18	18	0	28	90
9	90	19	22	29	58
10	100	20	44	30	0

If we define a sunny day as one with over 50% sunshine, determine whether the pattern of occurrence of sunny days is random.

Solution:

For this we can use the number-of-runs test. Rather than calculate runs across two samples, here we will simply note if a day has 50% or more sunshine. The sample becomes

S={+,+,+,+,-,+,+,-,+,+,-,+,+,-,+,+,-,-,-,-,+,+,+,+,+,+,+,+,-}

We see that the sample consists of 12 runs. There are $n_x = 21$ sunny days, and $n_y = 9$ cloudy days. Because $n_x < 20$, we cannot use the normal approximation given in Table 10-5. Instead the probability is computed using combinatorial rules, and is approximately 0.4. This is far too large for rejection of the randomness hypothesis. We cannot conclude the pattern is non-random.

10. Test the normality of the DO data (a) using the Kolmogorov–Smirnov test with the ungrouped data of Table 2-4 and (b) using the χ^2 test with $k = 6$ classes of Table 2-6.

Solution:

(a) We will take the mean and standard deviation as known rather than estimated from the sample. Doing so results in calculated PROB-VALUES that are smaller than the true values (i.e., we are more likely to reject the null hypothesis). For the DO data the mean and standard deviation are 5.58 and 0.39 respectively. We sort the data, and then find the differences between the observed and expected cumulative distributions. The table below shows the results for a few of the 50 observations:

x_i	$S(x_i)$	$F(x_i)$	$\lvert S(x_i)-F(x_i)\rvert$
4.2	0.020	0.015	0.005
4.3	0.040	0.023	0.017

4.4	0.060	0.032	0.028
…	…	…	…
5.9	0.780	0.692	0.088
…	…	...	…
6.7	0.960	0.960	0.000
6.8	0.980	0.972	0.008
6.9	1.000	0.981	0.019

The maximum difference is 0.088. Table A-9 shows that with 50 degrees of freedom, the corresponding PROB-VALUE is about 0.6. We obviously cannot reject the hypothesis of normality.

(b) Here we will take the mean and standard deviation as unknown, to be estimated from the sample. In other words, we estimate two parameters from the sample. In building the χ^2 table, we combine the first two and the last two categories in Table 2-6 to ensure at least 2 expected frequencies per cell. This reduces the number of categories 4, as seen in the table below:

Group	Minimum	Maximum	O_j	E_j	$(O_j-E_j)^2/E_j$
1	4.000	4.990	9	3.3	10.13
2	5.000	5.490	10	17.0	2.89
3	5.500	5.990	20	21.7	0.14
4	6.000	6.990	11	7.0	2.24

The observed Chi-square value is 15.4. With $k - p - 1 = 4 - 2 - 1 = 1$ degrees of freedom, Table A-8 shows that the PROB-VALUE is less than 0.0005. We therefore reject the null hypothesis.

Note that with only 4 classes, we can obtain only a rough idea of the distribution of DO. The 4 classes given in Table 2-6 do not yield a distribution that is at all similar to the normal distribution. In practice one would need many classes (and observations) for the χ^2 test to be reliable.

Possible Questions

PART-B

1. According to the IQ level and the economic conditions of their homes 1000 students at a college were graded. Use χ^2 test to find out whether there is any association between economic condition at home and IQ.

Economic Conditions	IQ		Total
	High	Low	
Rich	**460**	**140**	600
Poor	**240**	**160**	400
Total	700	300	1000

(Note: The level of significance is 0.05 and table value is 3.84).

2. Test Median class size for Math is larger than the median class size for English for the following data using Mann – Whitney U test.

Class size (Math, M)	23	45	34	78	34	66	62	95	81
Class size (English, E)	30	47	18	34	44	61	54	28	40

3. Mr. Gowtham, Personal Manager is concerned about absenteeism. He decides to sample the records to determine if absenteeism is distributed evenly throughout the six-day work-week. The null hypothesis to be tested is: absenteeism is distributed evenly throughout the week. The sample results are as follows:

Day	**Number of Absentees**
Monday	12
Tuesday	9
Wednesday	11
Thursday	10
Friday	9
Saturday	9

i. Using χ^2 test of significance, compute χ^2 value.
ii. Is the null hypothesis rejected?
iii. Specifically, what does this indicate to the Personal Manager?
(Note: The level of significance is 0.01 and table value is 15.086).

4. How Z-test is used for testing significance of proportions?
In a referendum submitted to the student body and 850 men and 550 women voted. Out of these, 530 of men and 310 of women voted 'yes'. Does this indicate a significant difference in opinion on matter between men and women students? (Use $\alpha = 5\%$ and $Z_{(0.05)} = 1.96$)

Question	Option 1	Option 2	Option 3	Option 4	Answer
Accept Ho when it is true leads ----------.	Type I error	Type II error	correct decision	either (a) or (b)	correct decision
An assumption of t – test is population of the sample is ----------.	Binomial	Poisson	normal	exponential	normal
Any hypothesis concerning a population is called a ----------.	Sample	population	statistical measure	statistical hypothesis	statistical hypothesis
Degree of freedom for statistic chi-square incase of contingency table of order 2x2 is----------	3	4	2	1	1
If $S_1^2 > S_2^2$, then the F – statistic is ----------.	S_1 / S_2	S_2 / S_1	S_1^2 / S_2^2	S_1^3 / S_2^3	S_1^2 / S_2^2
If the data is given in the form of a series of variables, then the DOF is --------.	n	n-1	n+1	(r – 1)(c – 1)	n-1
If the sample size is less than 30, then the sample is called ----------.	Large sample	small sample	population	alternative hypothesis	small sample
In -------, the variance of population from which samples are drawn are equal	t-test	Chi-Square test	Z-test	F-test	F-test
In F – test, the variance of population from which samples are drawn are ----------.	equal	not equal	small	large	equal
In sampling distribution the standard error is ----------.	np	pq	npq	sqrt(npq)	sqrt(npq)
Larger group from which the sample is drawn is called ----------.	Sample	sampling	universe	parameter	universe
Rejecting Ho when it is true leads ----------.	Type I error	Type II error	correct decision	either (a) or (b)	Type I error
Student's t-test is applicable only when ----------.	The variate values are independent	the variable is distributed normally	The sample is not large	all the above	all the above
The characteristic of the chi–square test is ----------.	DOF	LOS	ANOVA	independence of attributes	independence of attributes
The correct decision is ----------.	Reject Ho when it is true	Accept Ho when it is false	Reject Ho when it is false	none of these	Reject Ho when it is false
The degrees of freedom for two samples in t – test is ---------.	$n_1 + n_2 + 1$	$n_1 + n_2 – 2$	$n_1 + n_2 + 2$	$n_1 + n_2 – 1$	$n_1 + n_2 – 2$
The degrees of freedom of chi – square test is ----------.	(r – 1)(c – 1)	(r + 1)(c + 1)	(r + 1)(c – 1)	(r – 1)(c + 1)	(r – 1)(c – 1)
The expected frequency of chi – square test can be calculated as ----------.	(RT + CT) / GT	(RT – CT) / GT	(RT * CT) / GT	(RT*CT)	(RT * CT) / GT
The formula for c^2 is ----------	$å(O–E)^2/E$	$å(E+O)^2/E$	å(O-E) /E	$å(O-E)^2/O$	$å(O–E)^2/E$
Type II error occurs only if ----------.	Reject Ho when it is true	Accept Ho when it is false	Accept Ho when it is true	reject Ho when it is false	Accept Ho when it is false
Z – test is applicable only when the sample size is ----------.	zero	one	small	large	large
If the computed value is less than the critical value, then ----------.	Null hypothesis is accepted	Null hypothesis is rejected	Alternative hypothesis is accepted	population	Null hypothesis is accepted
Small sample test is also known as ----------.	Exact test	t – test	normal test	F-test	t – test
The distribution used to test goodness of fit is----------	F distribution	c^2distribution	t distribution	Z distribution	c^2distribution
_____ are the values that mark the boundaries of the confidence interval.	Confidence intervals	Confidence limits	Levels of confidence	Margin of error	Confidence limits
A 95% confidence interval for the mean of a population is such that:	It contains 95% of the values in the population	There is a 95% chance that it contains all the values in the population.	There is a 95% chance that it contains the standard deviation of the population	There is a 95% chance that it contains the mean of population	There is a 95% chance that it contains the mean of population
A confidence interval will be widened if:	The confidence level is increased and the sample size is reduced	The confidence level is increased and the sample size is increased	The confidence level is decreased and the sample size is decreased	The confidence level is decreased and the sample size is increased	The confidence level is increased and the sample size is reduced
A good way to get a small standard error is to use a ……..	Repeated sampling	Small sample	Large Sample	Large Population	Large Sample
A hypothesis may be classified as ……...	simple and composite	composite only	null only	total population and null	simple and composite
An estimator is a random variable because it varies from	Population to population	Population to sample	Sample to population	Sample to sample	Sample to sample
Analysis of variance utilizes:	t-test	Chi-Square test	Z-test	F-test	F-test
Degree of freedom for statistic chi-square incase of contingency table of order 2x2 is…….	4	3	2	1	1
Difference between value of parameter of population and value of unbiased estimator point is classified as…….	Sampling error	Marginal error	Confidence error	Population error	Sampling error
For the interval estimation of μ when σ is known and the sample is large, the proper distribution to use is……	t distribution with n +1 degrees of freedom	t distribution with n-1 degrees of freedom	t distribution with n degrees of freedom	normal distribution	normal distribution
If a standard error of a statistic is less than that of another then what is the former is said to be……..	efficient	unbiased	consistent	sufficient	efficient
If a statistic 't' follows student's t distribution with n degrees of freedom then t^2 follows -----	c^2distribution with (n-1) degrees of freedom	c^2distribution with n degrees of freedom	c^2distribution with n^2 degrees of freedom	c^2distribution with (n+1) degrees of freedom	c^2distribution with (n-1) degrees of freedom
If the calculated value is less than the table value then we accept the ---------- hypothesis.	Alternative	null	both	sample	null
If the calculated value is less than the table value, then we accept the ---------- hypothesis.	Alternative	Null	Statistics	Sample	Null
If the computed value is greater than the critical value, then ----------.	Null hypothesis is accepted	Null hypothesis is rejected	Alternative hypothesis is accepted	small sample	Null hypothesis is rejected
If true value of population parameter is 10 and estimated value of population parameter is 15 then error of estimation is…	5	25	0.667	150	5
In chi – square test, if the values of expected frequency are less than 5, then they are combined together with the neighbouring frequencies. This is known as ----------.	Goodness of fit	DOF	LOS	pooling	pooling
In chi – square test, if the values of expected frequency are less than 5, then they are combined together with the neighbouring frequencies	Goodness of fit	Degrees of freedom	Level of significance	Pooling	Pooling

In confidence interval estimation, formula of calculating	Point estimate ± margin of error	Point estimate - margin of error	Point estimate x margin of error	Point estimate + margin of error	Point estimate ± margin of error
In F – test, the variance of population from which samples …….	Equal	Different	Large	Small	Equal
Interval estimate is associated with	Probability	Non-probability	Range of values	Number of parameters	Range of values
Rejecting null hypothesis when it is true leads to……...	Type I error	Type II error	Type III error	Correct decision	Type I error
Small sample test is also known as ……	Z-test	t-test	Exact test	Normal test	t-test
Student's t-test is applicable in case of ----------.	Small samples	for sample of size between 5 and 30	Large samples	none of the above	Small samples
Student's t-test is applicable in case of ……...	Small samples	for sample of size between 5 and 25	Large samples	for sample of size of more than 100	Small samples
The maximum probability of committing type I error, which we specified in a test is known as ----------.	Null hypothesis	alternative hypothesis	DOF	level of significance	level of significance
The mean of Chi - distribution with *n* degrees of freedom ……...	n	n+1	2n	0	n
The technique of analysis of variance refered to as ------	ANOVA	F – test	Z – test	Chi- square test	ANOVA
The term STATISTIC refers to the statistical measures ……..	Population	Hypothesis	Sample	Parameter	Sample
The two variations, variation within the samples and variations between the samples are tested for their significance by -------	Chi- square test	F – test	t-test	Z – test	F – test
The value of Z test at 5% level of significance is	0.96	3.95	1.96	2.56	1.96
The value of Z test at 5% level of significance is ----------.	3.96	2.96	1.96	0.96	1.96
Under --------- classification , the influence of two attribute or factors is considered	two way	three way	one way	many	two way
When S is used to estimate σ, the margin of error is computed by using…….	normal distribution	t distribution	sample mean	population mean	t distribution
Which of the following is a non-parametric test	Chi square	F	t	Z	Chi square
Which one of the following refers the term Correlation?	Relationship between two values	Relationship between two variables	Average relationship between two variables	Relationship between two things	Relationship between two variables
Z – test is applicable only when the sample size is …….	Zero	2	Small	Large	Large

KARPAGAM ACADEMY OF HIGHER EDUCATION
(Deemed to be University)
(Established under Section 3 of UGC Act, 1956)
Pollachi Main Road, Coimbatore – 641 021, Tamil Nadu

Department of Mathematics

Subject : Business Analytics	**Semester II**	**L T P C**
Subject Code : 19BPU202	**Class : I B.Com (BPS)**	**5 2 0 5**

Unit – V: Regression

Problem definition, Data pre-processing; model building; Diagnostics and Validation
Simple linear regression: Coefficient of determination, Significance tests for predictor variables, Residual analysis, Confidence and Prediction intervals

Suggested Readings:

1. U Dinesh Kumar (2017), Business Analytics: The Science of Data - Driven Decision Making, Wiley, New Delhi.
2. R. Evans James (2017), Business Analytics, 2nd edition, Pearson Education, New Delhi.
3. S. Christian Albright , Wayne L. Winston (2015), Business Analytics: Data Analysis and Decision Making, 5th edition, Cengage Publications
4. Howard Anton (Author), Chris Rorres (Author) (2016) , Elementary Linear Algebra with Supplemental Applications, 11 edition, Wiley, India.
5. Friedberg / Insel / Spence (2015), Linear Algebra, 4th edition, Pearson Education, New Delhi.

CORRELATION AND REGRESSION ANALYSIS

Simple Linear Correlation

The term Correlation refers to the relationship between the variables. Simple correlation refers to the relationship between two variables. Various types of correlation are considered.

Type of Correlation

Positive or Negative when the values of two variables change in the same direction, their positive correlation between the two variables.

Example

X:	50	60	70	95	100	105	34	25	18	10	7
Y:	23	32	37	41	46	50	51	49	42	33	19

Simple or Partial or Multiple

When only two variables are considered as under positive or negative correlation above the correlation between them is called Simple correlation. When more than two variables as considered the correlation between two of them when all other variables are held constant, i.e., when the linear effects of all other variables on them are removed is called partial correlation. When more than two variables are considered the correlation between one of them and its estimate based on the group consisting of the other variables is called multiple correlation.

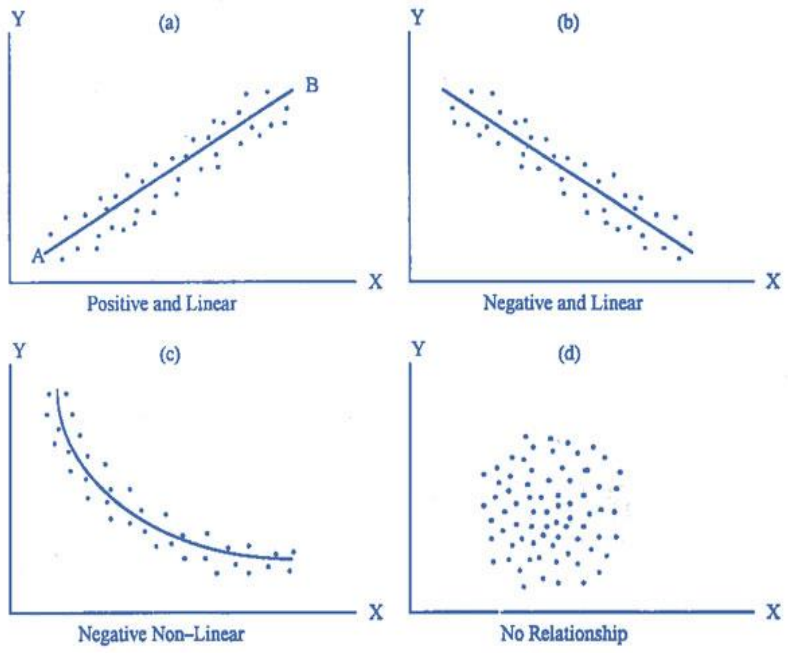
Methods of Finding Correlation Coefficient

The following four methods are available under simple linear correlation and among them; product moment method is the best one.

- Scatter Diagram
- Karl Pearson's correlation coefficient or product moment correlation coefficient (r)
- Spearman's rank correlation coefficient (ρ)
- Correlation coefficient by concurrent deviation method (r_c).

Scatter Diagram

Scatter diagram is a graphic picture of the sample data. Suppose a random sample of n pairs of observations has the values $(X_1, Y_1), (X_2, Y_2), (X_3, Y_3), \ldots, (X_n, Y_n)$. These points are plotted on a rectangular co-ordinate system taking independent variable on *X*-axis and the dependent variable on *Y*-axis. Whatever be the name of the independent variable, it is to be taken on *X*-axis. Suppose the plotted points are as shown in figure (a). Such a diagram is called scatter diagram. In this figure, we see that when *X* has a small value, *Y* is also small and when *X* takes a large value, *Y* also takes a large value. This is called direct or positive relationship between *X* and *Y*. The plotted points cluster around a straight line. It appears that if a straight line is drawn passing through the points, the line will be a good approximation for representing the original data. Suppose we draw a line ***AB*** to represent the scattered points. The line ***AB*** rises from left to the right and has positive slope. This line can be used to establish an approximate relation between the random variable *Y* and the independent variable *X*. It is nonmathematical method in the sense that different persons may draw different lines. This line is called the regression line obtained by inspection or judgment.



Making a scatter diagram and drawing a line or curve is the primary investigation to assess the type of relationship between the variables. The knowledge gained from the scatter diagram can be used for further analysis of the data. In most of the cases the diagrams are not as simple as in figure (a). There are quite complicated diagrams and it is difficult to choose a proper

mathematical model for representing the original data. The scatter diagram gives an indication of the appropriate model which should be used for further analysis with the help of method of least squares. Figure (b) shows that the points in the scatter diagram are falling from the top left corner to the right. This is a relation called inverse or indirect. The points are in the neighborhood of a certain line called the regression line.

As long as the scattered points show closeness to a straight line of some direction, we draw a straight line to represent the sample data. But when the points do not lie around a straight line, we do not draw the regression line. Figure (c) shows that the plotted points have a tendency to fall from left to right in the form of a curve. This is a relation called non-linear or curvilinear. Figure (d) shows the points which apparently do not follow any pattern. If X takes a small value, Y may take a small or large value. There seems to be no sympathy between X and Y. Such a diagram suggests that there is no relationship between the two variables.

Karl Pearson's Correlation Coefficient

Karl Pearson's Product-Moment Correlation Coefficient or simply Pearson's Correlation Coefficient for short, is one of the important methods used in Statistics to measure Correlation between two variables.

A few words about Karl Pearson: Karl Pearson was a British mathematician, statistician, lawyer and a eugenicist. He established the discipline of mathematical statistics. He founded the world's first statistics department In the University of London in the year 1911. He along with his colleagues Weldon and Galton founded the journal "Biometrika" whose object was the development of statistical theory.

The Correlation between two variables X and Y, which are measured using Pearson's Coefficient, give the values between +1 and -1. When measured in population the Pearson's Coefficient is designated the value of Greek letter rho (ρ). But, when studying a sample, it is designated the letter r. It is therefore sometimes called Pearson's r. Pearson's coefficient reflects the linear relationship between two variables. As mentioned above if the correlation coefficient is +1 then there is a perfect positive linear relationship between variables, and if it is -1 then there is a perfect negative linear relationship between the variables. And 0 denotes that there is no relationship between the two variables.

The degrees -1, +1 and 0 are theoretical results and are not generally found in normal circumstances. That means the results cannot be more than -1, +1. These are the upper and the lower limits.

Pearson's Coefficient computational formula

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}}$$

Sample question: compute the value of the correlation coefficient from the following table:

Subject	Age x	Weight Level y
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

Step 1: Make a chart. Use the given data, and add three more columns: xy, x2, and y2.

Subject	Age x	Weight Level y	xy	x^2	y^2
1	43	99			
2	21	65			
3	25	79			
4	42	75			
5	57	87			
6	59	81			

Step 2: Multiply x and y together to fill the xy column. For example, row 1 would be 43 × 99 = 4,257

Step 3: Take the square of the numbers in the x column, and put the result in the x^2 column.

Subject	Age x	Weight Level y	xy	x^2	y^2
1	43	99	4257	1849	
2	21	65	1365	441	
3	25	79	1975	625	
4	42	75	3150	1764	
5	57	87	4959	3249	
6	59	81	4779	3481	

Step 4: Take the square of the numbers in the y column, and put the result in the y^2 column.

Step 5: Add up all of the numbers in the columns and put the result at the bottom.2 column. The Greek letter sigma (Σ) is a short way of saying "sum of."

Subject	Age X	Weight Y	XY	X^2	Y^2
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
Σ	247	486	20485	11409	40022

Step 6: Use the following formula to work out the correlation coefficient.

The answer is: 1.3787 × 10-4 the range of the correlation coefficient is from -1 to 1. Since our result is 1.3787 × 10-4, a tiny positive amount, we can't draw any conclusions one way or another.

Spearman's Rank Correlation Coefficient

The Spearman correlation coefficient is often thought of as being the Pearson correlation coefficient between the ranked variables. In practice, however, a simpler procedure is normally used to calculate ρ. The n raw scores X_i, Y_i are converted to ranks x_i, y_i, and the differences $d_i = x_i - y_i$ between the ranks of each observation on the two variables are calculated.

If there are no tied ranks, then ρ is given by

$$\rho = 1 - \left(\frac{6 \sum d^2}{N(N^2 - 1)} \right)$$

If tied ranks exist, Pearson's correlation coefficients between ranks should be used for the calculation:

One has to assign the same rank to each of the equal values. It is an average of their positions in the ascending order of the values.

Example 1

X : 21 36 42 37 25

Y : 47 40 37 42 43

For the data given above, calculate the rank correlation coefficient.

Solution

		Rank of X and Y			
X	Y	R(X)	R(Y)	d	d^2
21	47	5	1	4	16
36	40	3	4	-1	1
42	37	1	5	-4	16
37	42	2	3	-1	1
25	43	4	2	2	4
			Total	$\sum d = 0$	$\sum d^2 = 38$

$$\rho = 1 - \left(\frac{6 \sum d^2}{N(N^2 - 1)} \right)$$

$$= 1 - \left(\frac{6 \times 38}{5 (5^2 - 1)} \right)$$

$$= 1 - 1.9 = -0.9$$

Tied Ranks

When one or more values are repeated the two aspects- ranks of the repeated values and changes in the formula are to be considered.

Example 2

Find the rank correlation coefficient for the percentage of marks secured by a group of 8 students in Economics and Statistics.

Marks in Eco: 50 60 65 70 75 40 70 80

Marks in Stat: 80 71 60 75 90 82 70 50

Solution

Let X be Marks in Economics and Y be Marks in Statistics

		Rank of X and Y			
X	Y	X	Y	d	d^2
50	80	7	3	4	16
60	71	6	5	1	1
65	60	5	7	-2	4
70	75	3.5	4	-0.5	0.25
75	90	2	1	1	1
40	82	8	2	6	36
70	70	3.5	6	-2.5	6.25
80	50	1	8	-7	49
			Total	$\sum d = 0$	$\sum d^2$ =113.5

$$\rho = 1 - \left(\frac{6\{ \sum d^2 + m(m^2-1)/12\}}{N(N^2-1)} \right)$$

When $m=2$, $m(m^2-1)/12 = 0.5$

Therefore $\rho = 1-\left\{6\{113.5+0.5\}/8(8^2-1)\right\}$

$= 1- 1.3571 = -0.3571$

Simple Linear Regression

The line which gives the average relationship between the two variables is known as the regression equation. The regression equation is also called estimating equation.

Uses

1. Regression analysis is used in statistics and other displines.
2. Regression analysis is of practical use in determining demand curve, supply curve, consumption function, etc from market survey.
3. In Economics and Business, there are many groups of interrelated variables.
4. In social resarch, the relation between variables may not known; the relation may differ from place to place.
5. The value of dependent variable is estimated corresponding to any value of the independent variable using the appropriate regression equation.

Method of Least Squares

From a scatter diagram, there is virtually no limit as to the number of lines that can be drawn to make a linear relationship between the 2 variables

- the objective is to create a BEST FIT line to the data concerned
- the criterion is the called the method of least squares
- i.e. the *sum of squares* of the *vertical deviations* from the points to the line be a minimum (based on the fact that the dependent variable is drawn on the vertical axis)
- the linear relationship between the dependent variable (Y) and the independent variable(x) can be written as $Y = a + bX$, where a and b are parameters describing the vertical intercept and the slope of the regression.
- Similarly the linear relationship between the dependent variable (XY) and the independent variable(Y) can be written as $X = a' + b'Y$, where a and b are parameters describing the vertical intercept and the slope of the regression.

Calculating the coefficients a and b

The values of a and b for the given pairs of values of (xi,yi) I = 1, 2, 3 ….. are determined

Using the normal equations as

$\sum y = Na + b\sum x$

$\sum xy = a\sum x + b\sum x^2$

Similarly, the values of a' and b' for the given pairs of values of (xi,yi) I = 1, 2, 3 ….. are determined,

Using the normal equations as,

$\sum x = Na' + b'\sum y$

$\sum xy = a'\sum y + b'\sum y^2$

Methods of forming the regression equations

- Regression equations based on normal equations.
- Regression equations based on X and Y and b_{YX} and b_{XY}.

Example 1

From the following data, obtain the two regression equations.

X	6	2	10	4	8
Y	9	11	5	8	7

use normal equations

Solution

X	Y	XY	X^2	Y^2
6	9	54	36	81
2	11	22	4	121
10	5	50	100	25
4	8	32	16	64
8	7	56	64	49
$\sum x=0$	$\sum y=0$	$\sum xy=214$	$\sum x^2=220$	$\sum y^2=340$

Let the regression equation Y on X is Y = a + bX

The normal equations are,

$\sum y = Na + b\sum x$

$\sum xy = a\sum x + b\sum x^2$

By substituting the values from the table, we get

5a+30b = 40 ------- 1

30a + 180b = 214 -------- 2

Solving these two equations we get,

a=11.90 and b= -0.65

Therefore the regression Y on X is Y = 11.90-0.65X.

Let the regression equation X on Y is X = a' + b'Y

The normal equations are,

$\sum x = Na + b\sum y$

$\sum xy = a\sum y + b\sum y^2$

By substituting the values from the table, we get

5a'+40'b = 30 ------- 3

40a' + 340b' = 214 -------- 4

Solving these two equations we get,

a' = 16.40 and b= -1.30

Therefore the regression equation X on Y is X = 16.40-1.30Y

Example 2

From the data given below, find

(i) the two regression equations

(ii) The correlation coefficient between the variables X and Y

(iii) The value of Y when X= 30

X : 25 28 35 32 31 36 29 38 34 32

Y : 43 46 49 41 36 32 31 30 33 39

Solution

X	Y	x= X- X`	Y= Y-Y`	xy	x^2	y^2
25	43	-7	5	-35	49	25
28	46	-4	8	-32	16	64
35	49	3	11	33	9	121
32	41	0	3	0	0	9
31	36	-1	-2	2	1	4
36	32	4	-6	-24	16	36
29	31	-3	-7	21	9	49

38	30	6	-8	-48	36	64
34	33	2	-5	-10	4	25
32	39	0	1	0	0	1
320	380	0	0	-93	140	398

$X` = 32$, $Y` = 38$, $b_{xy} = \sum xy / \sum y^2 = -0.2337$, $b_{yx} = \sum xy / x^2 = -0.6643$

iv) Regression equation of Y on X ,$(Y - Y`)$

$= b_{yx} (X-X`) (Y – 38) = -0.6643(X-32) \Rightarrow$

$Y = 59.26-0.6643X$

(ii) Regression equation of X on Y , $(X - X`)$

$= b_{xy} (Y-Y`) (X – 32) = -0.2337Y +8.88 \Rightarrow$

$X = 40.88 - 0.233\ Y$

(iii) $r = + \sqrt{b_{yx} b_{xy}} = -0.3940$

(iv) $Y = 59.26-0.6643x30= 39$

Properties of Regression coefficients

1. The two regression equations are generally different and are not to be interchanged in their usage. ` `
2. The two regression lines intersect at (X, Y).
3. Correlation coefficient is the geometric mean of two regression coefficients.
4. The two regression coefficients and the correlation coefficient have the same sign.
5. Both the regression coefficients and the correlation coefficient cannot be greater than one numerically and simultaneously.
6. Regression coefficients are independent of change of origin but are affected by the change of scale.
7. Each regression coefficient is in the unit of the measurement of the dependent variable.
8. Each regression coefficient indicates the quantum of change in the dependent variable corresponding to unit increase in the independent variable.

Questions

1) What are the types of Correlation?
2) Write any two properties of Correlation.
3) What is the range of Correlation Coefficient?
4) Define Positive Correlation.
5) What is meant by Regression?
6) What are the formulae for Regression co-efficients?
7) Distinguish between Correlation and Regression.
8) Write the formula for Rank Correlation, when more than one rank is repeated.
9) If $b_{xy} = -0.2337$ and $b_{yx} = -0.6643$ then find the Correlation Coefficient.
10) What is Negative Correlation? Give an example?
11) Write down the formula for Karl Pearson's Coefficient of Correlation.
12) Define Scatter Diagram.
13) What is Simple Correlation?
14) Define Regression Equation.
15) When X = 40, Y = 60, $\sigma_x = 10$, $\sigma_Y = 15$ and r = 0.7 find the Regression Equation of Y on X.

Exercise

1) Calculate the Correlation Coefficient from the following variables.

Sales in ('0000)	57	58	59	59	60	61	62	64
Advertisement Expenditure ('000)	17	16	15	18	12	14	19	11

2) Marks obtained by 8 students in Accountancy (X) and Statistics (Y) are given below. Compute Rank Correlation Coefficient.

X	25	20	28	22	40	60	20
Y	40	30	50	30	20	10	30

3) Calculate the two Regression Equations from the following data.

X	10	12	13	12	16	15
Y	40	38	43	45	37	43

4) Calculate Karl Pearson's Coefficient of Correlation from the following data.

Wages	100	101	102	102	100	99	97	98
Cost of Living	98	99	99	97	95	92	95	94

5) From the data given find two Regression Equations.

X	10	12	13	12	16	15
Y	20	28	23	25	27	30.

i) Estimate Y when X = 20.
ii) Estimate X when Y = 35.

6) A comparison of the undergraduate Grade Point Averages of 10 corporate employees with their scores in a managerial trainee examination produced the results shown *in* the following table.

Exam Score	89	83	79	91	95	82	69	66	75	80
GPA	2.4	3.1	2.5	3.5	3.6	2.5	2.0	2.2	2.6	2.7

Measure the Correlation Coefficient between Exam scores and GPA by using Rank Method and also interpret the data given with the help of Scatter Diagram.

7) Develop the Regression Equation that best fit the data given below using annual income as an independent variable and amount of life insurance as dependent variable.

Annual Income (Rs. in 000's)	62	78	41	53	85	34
Amount of Life Insurance (Rs. in 00's)	25	30	10	15	50	7

8) The ranks of ten students in Economics and Statistics subjects are as follows.

Economics	3	5	8	4	7	10	2	1	6	9
Statistics	6	4	9	8	1	2	3	10	5	7

Calculate Spearman's Rank Correlation Coefficient.

9) You are given the following data:

	X	Y
Arithmetic Mean	36	85
Standard Deviation	11	8

Correlation coefficient between X and Y = 0.66

Find the two Regression Equations. And also find Correlation Coefficient.

Question	Opt 1	Opt 2	Opt 3	Opt 4	Answer
The regression line cut each other at the point of-------	Average of X only	Average of Y only	Average of X and Y	the median of X on Y	Average of X and Y
Given the coefficient of correlation being 0.8, the coefficient of determination will be	0.98	0.64	0.66	0.54	0.64
Given the coefficient of correlation being 0.9, the coefficient of determination will be	0.98	0.81	0.66	0.54	0.81
If the coefficient of determination being 0.49, what is the coefficient of correlation	0.7	0.8	0.9	0.6	0.7
Given the coefficient of determination being 0.36, the coefficient of correlation will be	0.3	0.4	0.6	0.5	0.6
Which one of the following refers the term Correlation?	Relationship between two values	Relationship between two variables	Average relationship between two variables	Relationship between two things	Relationship between two variables
If r = +1, then the relationship between the given two variables is	perfectly positive	perfectly negative	no correlation	high positive	perfectly positive
If r = - 1, then the relationship between the given two variables is	perfectly positive	perfectly negative	no correlation	low Positive	perfectly negative
If r = 0, then the relationship between the given two variables is	Perfectly positive	perfectly negative	no correlation	both positive and negative	no correlation
Coefficient of correlation value lies between	1 and –1	0 and 1	0 and ∞	0 and –1.	1 and –1
While drawing a scatter diagram if all points appear to form a straight line getting Downward from left to right, then it is inferred that there is	Perfect positive correlation	simple positive correlation	Perfect negative correlation	no correlation	Perfect negative correlation
The range of the rank correlation coefficient is	0 to 1	–1 to 1	0 to ∞	– ∞ to ∞	–1 to 1
If r =1, then the angle between two lines of regression is	Zero degree	sixty degree	ninety degree	thirty degree	ninety degree

Regression coefficient is independent of	Origin	scale	both origin and scale	neither origin nor scale.	Origin
If the correlation coefficient between two variables X and Y is negative, then the Regression coefficient of Y on X is	Positive	negative	not certain	zero	negative
If the correlation coefficient between two variables X and Y is positive, then the Regression coefficient of X on Y is	Positive	negative	not certain	zero	Positive
There will be only one regression line in case of two variables if	r =0	r = +1	r = –1	r is either +1 or –1	r =0
The regression line cut each other at the point of	Average of X only	Average of Y only	Average of X and Y	the median of X on Y	Average of X and Y
If b_{xy} and b_{yx} represent regression coefficients and if $b_{yx} > 1$ then b_{xy} is	Less than one	greater than one	equal to one	equal to zero	Less than one
Rank correlation was discovered by	R.A.Fisher	Sir Francis Galton	Karl Pearson	Spearman	Spearman
Formula for Rank correlation is	$1- (6\Sigma d^2 /(n(n2-1)))$	$1- (6\Sigma d^2 /(n(n2+1)))$	$1+ (6\Sigma d^2 /(n(n2+1)))$	$1 /(n(n2-1))$	$1- (6\Sigma d^2 /(n(n2-1)))$
With b_{xy}=0.5, r = 0.8 and the variance of Y=16, the standard deviation of X=	6.4	2.5	10	25.6	2.5
The coefficient of correlation r =	$(b_{xy.} b_{yx})^{1/4}$	$(b_{xy .} b_{yx})^{-1/2}$	$(b_{xy .} b_{yx})^{1/3}$	$(b_{xy.} b_{yx})^{1/2}$	$(b_{xy.} b_{yx})^{1/2}$
If two regression coefficients are positive then the coefficient of correlation must be	Zero	negative	positive	one	positive
If two-regression coefficients are negative then the coefficient of correlation must be	Positive	negative	zero	one	Positive
The regression equation of X on Y is	X = a + bY	X = a + bX	X= a - bY	Y= a + bX	X = a + bY
The regression equation of Y on X is	X = a + bY	X = a + bX	X= a - bY	Y= a + bX	Y= a + bX
The given two variables are perfectly positive, if	r = +1	r = -1	r = 0	$r \neq +1$	r = +1
The relationship between two variables by plotting the values on a chart, known as-	coefficient of correlation	Scatter diagram	Correlogram	rank correlation	Scatter diagram

If x and y are independent variables then,	cov(x,y)≠ 0	cov(x,y)= 1	cov(x,y)= 0	cov(x,y) >1	cov(x,y)= 0
Correlation coefficient is the ---------- of the two regression coefficients.	Mode	Geometric mean	Arithmetic mean	median	Geometric mean
b_{xy} = 0.4, b_{yx} = 0.9 then r =	0.6	0.3	0.1	-0.6	0.6
b_{xy}=1/5, r=8/15, s_x = 5 then s_y=	40/13	13/40	40/3	3	40/3
The geometric mean of the two regression coefficients.	Correlation coefficient	regression coefficients	coefficient of range	coefficient of variation	Correlation coefficient
If two variables are uncorrelated, then the lines of regression	Do not exist	coincide	Parallel to each other	perpendicular to each other	perpendicular to each other
If the given two variables are correlated perfectly negative, then	r = +1	r = -1	r = 0	r ≠ +1	r = -1
If the given two variables have no correlation, then	r = +1	r = -1	r = 0	r ≠ +1	r = 0
If the correlation coefficient between two variables X and Y is ---------, the Regression coefficient of Y on X is positive	Negative	positive	not certain	zero	positive
If the correlation coefficient between two variables X and Y is ---------, the Regression coefficient of Y on X is negative	Negative	positive	not certain	zero	Negative
---------- is independent of origin and scale.	Correlation coefficient	regression coefficients	coefficient of range	coefficient of variation	Correlation coefficient
The angle between two lines of regression is ninety degree, if ----------	r = 2	r = 0	r = 1	r = -1	r = 1
--------- is used to measure closeness of relationship between variables.	Regression	mean	Rank correlation	correlation	correlation
If r is either +1 or –1, then there will be only one -------- line in case of two variables	Correlation	regression	rank correlation	mean	regression
When b_{xy}=0.85 and b_{yx}= 0.89, then correlation coefficient r =	0.98	0.5	0.68	0.87	0.87

If b_{xy} and b_{yx} represent regression coefficients and if $b_{xy} < 1$, then b_{yx} is	less than 1	greater than one	equal to one	equal to zero	greater than one
While drawing a scatter diagram if all points appear to form a straight line getting Downward from left to right, then it is inferred that there is--------	Perfect positive correlation	simple positive correlation	Perfect negative correlation	no correlation	Perfect negative correlation
If r =1, the angle between two lines of regression is----------	Zero degree	sixty degree	ninety degree	thirty degree	ninety degree
Regression coefficient is independent of----------	Origin	scale	both origin and scale	neither origin nor scale.	Origin
There will be only one regression line in case of two variables if----------	r =0	r = +1	r = –1	r is either +1 or –1	r =0