Terms & Conditions for Buying the Online Soft Copy

- The User must Read & Accept the Terms and Conditions (T&C) carefully before clicking on the accept option for Buying the Online Soft Copy of E-books / Assignment Solution Guides / Sample Question Papers (Question Bank) / Projects etc. based on IGNOU and other Universities / Boards / Institutes. Under this Particular Facility you may buy only the Online Soft Copy of E-books / Assignment Solution Guides / Sample Question Papers (Question Bank) / Projects based on IGNOU and other Universities / Boards / Institutes, no Hard Copy or Printed Copy shall be provided under this facility.
- The products which are for Online Reading i.e. E-Books, Sample Papers, Projects etc. are valid for 365 days only (From the Date of Purchase) and no kind of Downloading, Printing, Copying etc. are allowed in this facility as these products are just for Online Reading and References in your Mobile / Tablet / Computers.
- The Downloading facility may only be available for purchase made regarding Assignment Solution Guide (Solved Assignment) on a Special Request under special scheme offered by studybadshah.com time to time and all the other purchases made are just for Online Reading and no kind of Downloading, Printing, Copying etc. are allowed in this facility as these products are just for Online Reading and References in your Mobile / Tablet / Computers.
- All the online soft copy products E-books / Assignment Solution Guides / Sample Question Papers (Question Bank) / Projects etc. given in this website shall contain a diffused watermark on nearly every page to protect the material from being pirated / copy / misused etc.
- In these E-books / Assignment Solution Guides / Sample Question Papers (Question Bank) / Projects based on IGNOU and other Universities / Boards, Solutions of Only the Selected Questions are provided, the answers of all the Questions are not Provided.
- In these E-Books / Sample Assignment Solution Guides / Sample Question Papers (Question Bank) / Projects based on IGNOU and other Universities / Boards / Institutes, only Minimum Requirement of the Assignments Questions / Sample Papers have been answered, Like if 5 Questions are given and it has been asked that Answer any 3 out of them, then only 3 selected questions answers shall be given, each & every question shall not been answered.
- This is a Chargeable Facility / Provision to Buy the Online Soft Copy of E-books / Assignment Solution Guides / Sample Question Papers (Question Bank) / Projects based on IGNOU and other Universities / Boards / Institutes available online through our Website Which a Subscriber / Buyer may Read Online (whichever facility is offered by the website time to time) on his or her Mobile / Tablet / Computer. The E-books / Assignment Solution Guides / Sample Question Papers (Question Bank) / Projects based on IGNOU and other Universities / Boards and their answer given in these Soft Copy provides you just the approximate pattern of the actual Answer. However, the actual Content / Study Material / Assignments / Question Papers / Projects might somewhat vary in its contents, distribution of marks and their level of difficulty.
- These Sample Answers/Solutions are prepared by the author for the help, guidance and reference of the student to get an idea of how he/she can answer the questions. Sample answers may be Seen as the Guide/Reference Material only. Neither the publisher nor the author or seller will be responsible for any damage or loss due to any mistake, error or discrepancy as we do not claim the Accuracy of these solution / Answers. Any Omission or Error is highly regretted though every care has been taken while preparing these Sample

Answers/Solutions. Any mistake, error or discrepancy noted may be brought to the publishers notice which shall be taken care of in the next edition. Please consult your Teacher/Tutor or refer to the prescribed & recommended study material of the university / board / institute / Govt of India Publication or notification if you have any doubts or confusions before you appear in the exam or Prepare your Assignments before submitting to the University/Board/Institute.

- Study Badshah shall remain the custodian of the Contents right / Copy Right of the Content of these reference E-books / Assignment Solution Guides / Sample Question Papers (Question Bank) / Projects based on IGNOU and other Universities / Boards / Institutes given / being offered at the website www.studybadshah.com.
- The User agrees Not to reproduce, duplicate, copy, sell, resell or exploit for any commercial purposes, any portion of these Services / Facilities, use of the Service / Facility, or access to the Service / Facility.
- The Price of these E-books / Assignment Solution Guides / Sample Question Papers (Question Bank) / Projects based on IGNOU and other Universities / Boards may be Revised / Changed without any Prior Notice.
- The time duration of providing this online reading facility of 365 days may be alter or change by studybadshah.com without any Prior Notice.
- The Right to accept the order or reject the order of any E-books / Assignment Solution Guides / Sample Question Papers (Question Bank) / Projects made by any customer is reserved with www.studybadshah.com only.
- All material prewritten or custom written is intended for the sole purpose of research and exemplary purposes only. We encourage you to use our material as a research and study aid only. Plagiarism is a crime, and we condone such behaviour. Please use our material responsibly.
- In any Dispute What so ever Maximum Anyone can Claim is the Cost of a particular E-book / Assignment Solution Guides / Sample Question Papers (Question Bank) / Projects based on IGNOU and other Universities / Boards which he had paid to Study Badshah company / website.
- If In case any Reader/Student has paid for any E-Book/ Sample Papers/ Project / Assignment etc and is unable to Access the same at our Website for Online Reading Due to any Technical Error/ Web Admin Issue / Server Blockage at our Website www.studybadshah.com then He will be send a New Link for that Particular E-Book/ Sample Papers/ Project / Assignment to Access the same and if Still the Issue is Not Resolved Because of Technical Error/ Web Admin Issue / Server Blockage at our website then His Amount for that Particular Purchase will be refunded by our website via PayTM.
- All the Terms, Matters & Disputes are Subjected to "Delhi" Jurisdiction Only.

QUESTION PAPER

(June – 2016)

(Solved)

STATISTICAL TECHNIQUES

Time: 2 hours] *[Maximum Marks: 50*

Note: *(i) Attempt both Sections, i.e. Section A and Section B.*
*(ii) Attempt any **four** questions from Section A.*
*(iii) Attempt any **three** questions from Section B.*
(iv) Non-scientific calculator is allowed.

SECTION-A

Q. 1. The mean and standard deviation of 20 items is found to be 10 and 2, respectively. At the time of checking it was found that one noted item with value 8 was incorrect. Calculate the mean and standard deviation, if the wrong item is deleted.

Sol. Let the variable x denote items. Then we are given

$$\bar{X} = \frac{\Sigma X}{20} = 10$$

$$\Rightarrow \quad \Sigma x = 200$$

$$\text{Corrected } \Sigma x = 200 - (\text{Wrong item value})$$

$$= 200 - 8$$

$$= 192$$

$$\therefore \text{ Corrected mean item} = \frac{192}{19} = 10.10$$

$$\text{Standard deviation on } \sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - (\bar{x})^2}$$

$$\Rightarrow \quad 2 = \sqrt{\frac{1}{20}\text{Incorrect}\sum_{i=1}^{n} x_i^2 - (10)^2}$$

$$\Rightarrow \quad 2 = \frac{1}{20}\text{Incorrect}\sum_{i=1}^{n} x_i^2 - 100$$

$$\Rightarrow \text{ Incorrect } \sum_{i=1}^{n} x_i^2 = 2080$$

$$\therefore \text{ Correct } \sum_{i=1}^{n} x_i^2 = \text{ Incorrect } \sum_{i=1}^{n} x_i^2 - (8)^2$$

$$= 2080 - 64$$

$$= 2016$$

$$\therefore \text{ Correct Standard deviation} = \sqrt{\frac{\text{correct}\sum x_i^2}{n} - (\text{correct Mean})^2}$$

$$= \sqrt{\frac{2016}{19} - (10.1)^2}$$

$$= \sqrt{106.1 - 102.1}$$

$$= \sqrt{4.09}$$

$$= 202$$

Q. 2. Let x_1 and x_2 be two independent random variables with variances Var $(x_1) = k$, Var $(x_2) = 2$. If the variance of $y = 3x_2 - x_1$ is 25, then find k.

Sol. Var $(y) = 3x_2 - x_1 = 25$

$\Rightarrow$ 3 var (x_2) – Var $(x_1) = 25$

$\Rightarrow \quad 3k - 2 = 25$

$\Rightarrow \quad 3\text{ k} = 27$

$$\boxed{k = 9}$$

Q. 3. *(a)* State and prove the Addition theorem of probability.

Sol. Prove the formula for general additional rule of three events

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

Now

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

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

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

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

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

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

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

***(b)* Suppose that *A* and *B* are two independent events, associated with a random experiment. The probability of occurrence of event *A* or *B* is 0.8, while the probability of occurrence of event A is 0.5. Determine the occurrence of probability of event *B*.**

Sol. If events *A* and *B* are associated

$$P(A \text{ or } B) = P(A) + P(B)$$

$$\cdot 8 = \cdot 5 + P(B)$$

$$P(B) = \cdot 8 - \cdot 5$$

$$\boxed{P(B = \cdot 3)}$$

Q. 4. *(a)* What do you understand by a random variable? Define the types of random variables.

Ans. Ref.: See Chapter 3, Page No. 40, "Random variable", Page No. 42, "Discrete Random Variable" Page No. 43, "Continuous Random Variable".

***(b)* A bag contains 10 white and 3 black balls. Balls are drawn one by one without replacement till all the black balls are drawn. Find the probability that all black balls are drawn by the 6th draw.**

Sol. 1st Draw to select 1 balck ball then

$$= 13\, c_1 = 13$$

2nd Draw to select 2 black ball then

$$12\, c_1 = 12$$

3rd Draw to select 3rd black ball then

$$= 11 c_1 = 11$$

Next 4th, 5th and 6th Draw back ball are finished. So

Total black ball probability $= 13c_1 \times 12c_1 \times 11c_1$

$$= \frac{13.12!}{12!} \times \frac{12.11!}{11!\cdot} \times \frac{11.10!}{10!}$$

$$= 13 \times 12 \times 11$$

$$= 1716$$

Q. 5. A survey of 64 medical labs revealed that the mean price charged for a certain test was ₹ 120, with a standard deviation of ₹ 60. Test whether the data indicates that the mean price of this test is more than ₹ 100 at 5% level of significance.

Sol. $H_0 : \mu = 120$ *i.e.* average mean price charged for a certain test.

$H_1 : \mu < 120$

Now

$$Z_0 = \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma}$$

Where

$$n = 64$$

$$\bar{X} = 100$$

$$\mu = 120$$

$$\sigma = 60$$

The critical region at 5% level of significance for this medical test is:

$W : Z_0 < -1.96$

$$\therefore \quad Z_0 = \frac{\sqrt{64}(100 - 120)}{60}$$

$$= \frac{8 \times -20}{60}$$

$$= -2.66.$$

At this does not tall under the critical region, H_o is accepted.

SECTION B

Q. 6. Describe the following tests in detail:

***(a)* Paired t-test**

Sol. Paired sample t-test is a statistical technique that is used to compare two population means in the case of two samples that are correlated.

Steps:

1. Set up hypothesis: We set up two hypotheses. The first is the null hypothesis, which assumes that the mean of two paired samples are equal. The second hypothesis will be an alternative hypothesis, which assumes that the means of two paired samples are not equal.

2. Select the level of significance: After making the hypothesis, we choose the level of significance. In most of the cases, significance level is 5%, (in medicine, the significance level is set at 1%).

3. Calculate the parameter: To calculate the parameter we will use the following formula:

$$t = \frac{\bar{d}}{\sqrt{s^2 / n}}$$

Where d bar is the mean difference between two samples, s^2 is the sample variance, n is the sample size and t is a paired sample t-test with n-1 degrees of freedom. An alternate formula for paired sample t-test is:

$$t = \frac{\sum d}{\sqrt{\frac{n\left(\sum d^2\right) - \left(\sum d\right)^2}{n-1}}}$$

4. Testing of hypothesis or decision making: After calculating the parameter, we will compare the calculated value with the table value. If the calculated value is greater than the table value, then we will reject the null hypothesis for the paired sample t-test. If the calculated value is less than the table value, then we will accept the null hypothesis and say that there is no significant mean difference between the two paired samples.

(b) **Chi-Square test for independence of Attributes.**

Ans. Ref.: See Chapter 7, Page No. 87, "Test of Independence".

Q. 7. Differentiate between any two of the following:

(a) **Simple Random Sampling with Replacement and Simple Random Sampling Without Replacement.**

Ans. when we sample with replacement, the two sample values are independent. Practically, this means that what we get on the first one doesn't affect what we get on the second. Mathematically, this means that the covariance between the two is zero.

In sampling without replacement, the two sample values aren't independent. Practically, this means that what we got on the for the first one affects what we can get for the second one. Mathematically, this means that the covariance between the two isn't zero. That complicates the computations. In particular, if we have a SRS (Simple Random Sample) without replacement, from a population with variance σ^2, then the covariance of two of the different sample values is $\frac{-\sigma^2}{N-1}$, where N is the population size.

When we sample without replacement, and get a non-zero covariance, the covariance depends on the population size. If the population is very large, this covariance is very close to zero. In that case, sampling with replacement isn't much different from sampling without replacement. In some discussions, people describe this difference as sampling from an infinite population (sampling with replacement) versus sampling from a finite population (without replacement).

(b) **Probability (Random) Sampling and Non-Random Sampling.**

Ans. Survey designs based on random sampling are designed to select sampling units from the population with known probabilities. This means that the sampling properties of estimators of population quantities can be determined, such as whether or not the estimator is unbiased (i.e., does it on average give the right answer?) and what is its precision (i.e., how do we calculate its variance or its standard error).

This is the objective scientific approach to sampling and the only one for which sampling properties of estimators are known. Other methods may provide good information but there is no guarantee that they will, and their sampling properties are unknown. However even with random sampling, if response rates are poor then the possibility of non-response bias will compromise the estimation.

Non-random sampling, is any other kind of sampling. Such methods are often used for speed and convenience, and also they do not require a sampling frame. Their big disadvantage is that sampling error cannot reliably be quantified, as the sampling properties of any estimators used are not known (since the probability of choosing any one individual or sample cannot be determined).

Convenience or accessibility sampling involves asking a sample of people to respond to a survey. An example is distributing survey questionnaires at a meeting of a local beekeeping association or at a beekeepers' convention. However these people may not be representative of the whole target population of beekeepers, for example due to local weather conditions in the first case, or the fact that attendees at a convention may be real enthusiasts whose bee husbandry practices are not typical of the general beekeeping population. A small convenience sample may be very useful for a pilot survey but is not recommended more generally.

***(c)* One-Sample Test and Two-Sample Test**

Ans. One-sample Test: This Test null hypothesis: the population mean for the treatment group is not significantly different from known or standard value c. This is stated succinctly as

$$H_0: \mu = c$$

The alternative hypothesis: the population mean is not equal to *c* or,

$$H_1: \mu \neq c$$

Example: The speed of light in a vacuum *c* is a well-known constant of nature. (In fact it is the same everywhere in the universe.) Measurements of the speed of light in water are taken. Test the null hypothesis that the speed of light in water is not significantly different than the speed of light in a vacuum.

Two-sample Test: Use a paired sample test when there is a natural one-to-one pairing between the subjects in two treatment groups. In this case, the difference scores $d_i = x^2_i - x_{1i}$ can be computed and a one-sample test performed using the null hypothesis that the mean of the difference is not significantly different than zero:

$$H_0: \mu_{diff} = 0$$

The alternative hypothesis is

$$H_0: \mu_{diff} \neq 0$$

Example: A sample of houses is chosen. For each house, a section is painted with a new paint and a section is painted with a standard paint. For each house, measure the difference between the lifetimes of the new paint minus the lifetime of the standard paint. Test the null hypothesis that differences are not significantly different than zero.

Q. 8. The following table show the sample values of 3 independent normal random variables. Test whether they have the same mean [use ANOVA]. Given $F_{0.05}$ (2,9) = 4.26.

X_1	:	**13**	**11**	**16**	**22**
X_2	:	**16**	**08**	**21**	**11**
X_3	:	**15**	**12**	**25**	**10**

Ans.

X_1	:	13	11	16	22
X_2	:	16	08	21	11
X_3	:	15	12	25	10
N	:	3	3	3	3 = 12
Total	:	44	31	62	43 = 180
Avg	:	14.6	10.3	20.6	14.3 = 59.8

$$CF = \frac{(\text{Total})^2}{N} = \frac{(180)^2}{12} = 2700$$

$$\sum C_1^2 = 13^2 + 16^2 + 15^2 = 169 + 256 + 225 = 680$$

$$\sum C_2^2 = 11^2 + 8^2 + 12^2 = 121 + 64 + 144 = 329$$

$$\sum C_3^2 = 16^2 + 21^2 + 25^2 = 256 + 441 + 625 = 1322$$

$$\sum C_4^2 = 22^2 + 11^2 + 10^2 = 484 + 121 + 100 = 705$$

Catalyst of Variable

$$= \frac{(680)^2 + (329)^2 + (1322)^2 + (705)^2}{3} - CF$$

$$= \frac{462400 + 108241 + 1747684 + 497025}{3} - CF$$

$$= \frac{2,815,350}{3} - CF$$

$$= 938,450 - 2700$$

$$= 935, 750$$

Error sum of square in sample

= Tss – Catalyst of variable

= 3.036 – 935750

= – 932, 714

ANOVA Table

SV	DF	SS	MS	F
Catalyst	3	935750	311916.6	
Error				
(With in Sample)	9	– 932714	– 103634.8	

$$F(2, 9) = -\frac{311916.6}{103634.8} = -3.009$$

So $F_{TAB} > F_{cae}$, Since difference in significant and we conclude that data donot suggest the difference 3 independent normal random variable.

Q. 9. ***(a)*** **Discuss the following:**

(i) **Control Chart for variables**

Ans. Variables control charts plot continuous measurement process data, such as length or pressure, in a time-ordered sequence. In contrast, attribute control charts plot count data, such as the number of defects or defective units. Variables control charts, like all control charts, help you identify causes of variation to investigate, so that you can adjust your process without over-controlling it.

(ii) **Control Chart for attributes**

Ans. Attribute control charts that plot nonconformities (defects) or nonconforming units (defectives). A nonconformity refers to a quality characteristic and a nonconforming unit refers to the overall product. A unit may have many nonconformities, but the unit itself is either conforming or nonconforming. For example, a scratch on a metal panel is a nonconformity. If several scratches exist, the entire panel may be considered nonconforming.

(b) **Describe control chart for $\bar{X}$ and *R* in detail. Also suggest when R-chart and S-chart can be used.**

Ans. An $\bar{X}$-*R* chart plots the process mean ($\bar{X}$ chart) and process range (*R* chart) over time for variables data in subgroups. This combination control chart is widely used to examine the stability of processes in many industries.

For example, you can use $\bar{X}$-*R* charts to monitor the process mean and variation for subgroups of part lengths, call times, or hospital patients' blood pressure over time.

The $\bar{X}$ chart and the *R* chart are displayed together because you should interpret both charts to determine whether your process is stable. Examine the R chart first because the process variation must be in control to correctly interpret the $\bar{X}$ chart. The control limits of the $\bar{X}$ chart are calculated considering both process spread and center. If the *R* chart is out of control, then the control limits on the $\bar{X}$ chart may be inaccurate and may falsely indicate an out-of-control condition or fail to detect one.

An *R* chart plots the process range over time for variables data in subgroups. This control chart is widely used to examine the stability of processes in many industries.

For example, you can use *R* charts to examine process variation for subgroups of part lengths, call times, or hospital patients' blood pressure over time.

An *S* chart plots the process standard deviation over time for variables data in subgroups. This control chart is widely used to examine the stability of processes in many industries. For example, you can use *S* charts to examine process variation for subgroups of part lengths, call times, or hospital patients' blood pressure over time.

■■

QUESTION PAPER

(June – 2015)

(Solved)

STATISTICAL TECHNIQUES

Time: 2 hours] ***[Maximum Marks: 50***

Note: *(i) Attempt both Sections, A and B.*
(ii) Attempt any ***four*** *questions from Section A.*
(iii) Attempt any ***three*** *questions from Section B.*

SECTION-A

Q. 1. ***(a)*** **In order to find the correlation coefficient between two variables X and Y from 20 pairs of observations, the following calculations were made:**

$\sum x = 15,\ \sum y = -6,\ \sum xy = 50$

$\sum x^2 = 61$ **and** $\sum y^2 = 90.$

Calculate the correlation coefficient and the slope of the regression line of Y on X.

Sol. Given by $\sum x = 15,\ \sum y = -6,$

$\sum xy = 50$

$\sum x^2 = 61,\ \sum y^2 = 90$

$n = 20$

$$\bar{x} = \frac{\sum x}{n} = \frac{15}{20} = 0.75$$

$$\bar{y} = \frac{\sum y}{n} = -\frac{6}{20} = -0.3$$

$$b_{yx} = \frac{n\sum xy - \sum x \sum y}{n\sum x^2 - \left(\sum x\right)^2}$$

$$= \frac{20 \times 50 - 15 \times (-6)}{20 \times 61 - 15 \times 15}$$

$$= \frac{1000 + 90}{1220 - 225} = \frac{1090}{995} = 1.09$$

Regression line of y on x is given by

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

$$y + 0.3 = 1.09\,(x - 0.75)$$

$$y = 1.09x - 0.82 - 0.3$$

$$\boxed{y = 1.09\,x - 0.52}$$ **Ans.**

Coefficient of Correlation

$$r^2 = b_{xy} \cdot b_{xy}$$

$$b_{xy} = \frac{n\sum xy - \sum x \sum y}{n\sum y^2 - (\sum y)^2}$$

$$= \frac{20 \times 50 - 15 + (-6)}{20 \times 90 - (-6)^2}$$

$$= \frac{1000 + 90}{1800 - 36} = \frac{1090}{1764} = 0.61$$

$$b_{xyx} = 1.09$$

$$r^2 = 1.09 \times 0.61$$

$$r^2 = 0.673$$

$$r = \sqrt{0.673} = 0.818$$ **Ans.**

Q. 2. Suppose 2% of the items made in a factory are defective. Find the probability that there are

(i) **3 Defectives in a sample of 100,**

(ii) **no defectives in a sample of 50.**

Sol. Suppose 2% of the item produced by a factory are defective. Using the poisson approximation to the binomial.

(i) The probability that there are 3 defective items in a sample of 100 items.

$$K = np = 100\,(.0_2) = 2$$

$$P(X=3) = \frac{e^{-2}\,2^3}{3!} = \frac{0.135\times 8}{6} = \frac{1.08}{6}$$

$$= 0.180 \textbf{ Ans.}$$

(ii) $K = nb = 50\,(.02) = 1$

$$P(X=0) = \frac{e^{-1}\,1^0}{1!} = e^{-1} = 0.367 \textbf{ Ans.}$$

Q. 3. Telephone Directories have telephone numbers which are the combinations of the digits 0 to 9. The observer notes the frequency of occurrence of these digits and wants to test whether the digits occur with same frequency or not ($\alpha = 0.05$). The data are given below:

Digits	Frequency
0	99
1	100
2	82
3	65
4	50
5	77
6	88
7	57
8	82
9	30

(Given that $\chi^2_9\,(0.05) = 16.918$)

Sol. $E = \frac{99+100+82+65+50+77+88+57+82+20}{10} = \frac{730}{10} = 73$

Digit	Frequency (Θ)	Expected Frequency (E)	O – E	$(O-E)^2$	$\frac{(0-E)^2}{E}$
0	99	73	26	676	9.260273973
1	100	73	27	729	9.9863
2	82	73	9	81	1.109589041
3	65	73	–8	64	0.876712328
4	50	73	–23	529	7.246575342
5	77	73	4	16	0.219178082
6	88	73	15	225	3.082191781
7	57	73	–16	256	3.506849315
8	82	73	9	81	1.109589041
9	30	73	–43	1849	25.32876712
				4506	$\sum\frac{(O-E)^2}{E} = 61.7260274$

Since the calculate value is more than the given value i.e. $x^2_9\,(0.5). < 61.7260274$. So H_0 is rejected i.e. digits are not occur with same frequency.

Q. 4. Fit a linear trend $y = a + b$ * Demand, to the data collected in a unit manufacturing umbrellas given in the following table:

Month	1	2	3	4	5	6
Demand	46	56	54	43	57	56

Ans.

Month (x)	1	2	3	4	5	6
Demand (y)	46	56	54	43	57	56

Let the straight line of best fit be

$$y = a + bx \qquad \text{— (1)}$$

Normal eqn. are

$$\sum y = ma + b\sum x \qquad \text{— (2)}$$

$$\sum xy = a\sum x + b\sum x^2 \quad \text{— (3)}$$

Here, $m = 6$

x	y	xy	x^2
1	46	46	1
2	56	112	4
3	54	162	9
4	43	172	16
5	57	285	25
6	56	336	36
$\Sigma x = 21$	$\Sigma y = 312$	$\Sigma xy = 1113$	$\Sigma x^2 = 91$

Substituting the value in eqn. 2 and 3, we get.

$312 = 6a + 21b$ — (4) × 21

$1113 = 21a + 91b$ — (5) × 6

$6552 = 126a + 441b$

$6678 = 126a + 546b$

$-126 = -105b$

$$b = \frac{126}{105} = 1.2$$

Put the value in eqn. (4)

$374.4 = 6a + 25.2$

$374.4 - 25.2 = 6a,$

$$\frac{349.2}{6} = a,\ a = 58.2,$$

Hence, required line $\boxed{y = 58.2 + 1.2x}$ **Ans.**

Q. 5. The mean weekly sales of soap bars in different departmental stores was 146.3 bars per store. After an advertisement campaign the mean weekly sales of 22 stores for a typical week increased to 153.7 and showed a standard deviation of 17.2. Was the advertisement campaign successful at 5% level of significance? (Given t_{21} (0.05) = 2.08)

Ans. We are given $n = 22$ $\bar{X} = 153.7$ $S = 17.2$

Null Hypothesis H_0: $\mu = 146.3$, the deviation between $\bar{x}$ and μ is just due to fluctuation of sampling. In other words, advertisement is not effective. Alternate Hypothesis H_1: $\mu > 146.3$

Test static: Under the rule hypothesis H0, the test static is

$$t = \frac{\bar{x} - \mu}{5/\sqrt{n}} = \frac{\bar{x} - \mu}{5/\sqrt{n-1}} = t_{n-1} = t_{21}$$

$$t = \frac{153.7 - 146.3}{17.2/\sqrt{21}}$$

$$\frac{7.4 \times 4.28}{17.2} = \frac{33.892}{17.2}$$

$t = 1.970$

Tabulated value of T for 21 d.f at 5% level of significance for single (right) tail test is 2.08. Since calculated value of T is lest than the tabulated value. it is significant.

Q. 6. Write two merits and two demerits of Median An incomplete frequency distribution is given as follows:

C.I.	Frequency
10 – 20	**12**
20 – 30	**30**
30 – 40	**?**
40 – 60	**65**
50 – 60	**?**
60 – 70	**25**
70 – 80	**18**

Given that median value of 200 observations is 46, determine the missing frequencies using the median formula.

Ans.

CI	Frequency = f	Cf → Cummulative frequency
10 – 20	12	12
20 – 30	30	42
30 – 40	x	$42 + n$
40 – 50	65	$107 + x$
50 – 60	y	$107 + x + y$
60 – 70	25	$132 + x + y$
70 – 80	18	$150 + x + y$
	$N = \Sigma f = 200 = 150 + x + y$ $\Rightarrow x + y = 50$ — (1)	

Median is $\frac{200}{2} = 100$th term which lies in 40 – 50th group.

$l_1 = 40,\ l_2 = 50, f_1 = 65,\ C = 42 + x$

$m = \frac{200}{2} = 100$

Use Median $= m = l_1 + \frac{l_2 - l_1}{f_1}(m - c)$

$\Rightarrow \quad 46 = 40 + \frac{50 - 40}{65}(100 - 42 - x)$

$\Rightarrow \quad x = 19$ (after solving)

$\Rightarrow \quad y = 31$ use eqn. (1)

SECTION - B

Q. 7. A Chemical firm wants to determine how four catalysts differ in yield. The firm runs the experiment in three of its plants, types A, B, C. In each plant, the yield is measured with each catalyst. The yield (in quintals) are as follows:

Plant	Catalyst			
	1	2	3	4
A	2	1	2	4
B	3	2	1	3
C	1	3	3	1

Perform an ANOVA and comment whether the yield due to a particular catalyst is significant or not at 5% level of significance. Given $F_{3,6} = 4.76$.

Ans.

Plant	Catalyst			
	1	2	3	4
A	2	1	2	4
B	3	2	1	3
C	1	3	3	1

N	3	3	3	3	= 12
Total	6	6	6	8	= 26
Avg	2	2	2	2.66	= 8.66
Σy_{ij}^2	14	14	14	26	= 68

$CF = \frac{(26)^2}{12} = 56.33$

$SSS = \Sigma C_1^2 = 14,\ \Sigma C_2^2 = 14,\ \Sigma C_3^2 = 14,$

$\Sigma C_4^2 = 26$

Where C_1 = Catalyst 1, C_2 = Catalyst 2, C_3 = Catalyst 3, C_4 = Catalyst 4.

TSS = 68

Catalyst sum of square (SS) between Sample.

$= \frac{(14)^2 + (14)^2 + (14)^2 + (26)^2}{3} - CF$

$= \frac{196 + 196 + 196 + 676}{3} - 56.33$

$= 421.33 - 56.33$

$= 365.00$

Error sum of square (ESS) with in sample

= TSS – Catalyst sum of square

= 68 – 365

= – 297

ANOVA TABLE

SV	DF	SS	MS	F
Catalyst (Between Sample)	3	365	121.67	$F_{3,8} = \frac{121.67}{37.125}$
Error (With in Sample)	8	– 297	– 37.125	$F_{3,8} = -3.27$

Conclusion: $F_{TAB} > F_{Cae}$, Since difference in significant and we conclude that data do not suggest the difference in 4 catalyst in 3 plant.

Q. 8. Find and plot the regression line of *y* on *x* on scatter diagram for the data given below:

Speed km/hr	30	40	50	60
Stopping distance in feet	160	240	330	435

Ans.

Speed km/hr	Stopping distance in fect		
x	*y*	x^2	*xy*
30	160	900	4800
40	240	1600	9600

x	y	x^2	xy
50	330	2500	16500
60	435	3600	26100
$\Sigma x = 180$	$\Sigma y = 1165$	$\Sigma x^2 = 8600$	$\Sigma xy = 5700$

The line of regression of y on x is given by

$$y - \bar{y} = b_{yx}(x - \bar{x})$$

where b_{yx} is the coefficient of regression given by

$$b_{yx} = \frac{x\,\Sigma xy - \Sigma x\,\Sigma y}{x\,\Sigma x^2 - (\Sigma x)^2}$$

Here $n = 4$,

$$b_{yx} = \frac{4 \times 5700 - 180 \times 1165}{4 \times 8600 - (180)^2}$$

$$b_{yx} = \frac{135300 - 209700}{34400 - 32400} = \frac{-74400}{2000}$$

$$\boxed{b_{yx} = -37.2}$$

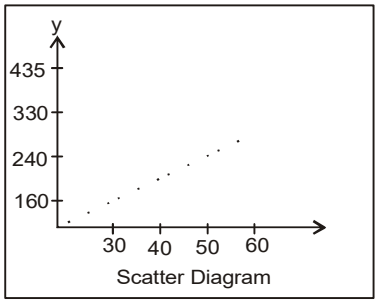
$$\bar{y} = \Sigma y / n = \frac{1165}{4} = 291.25$$

$$\bar{x} = \Sigma x / n = \frac{180}{4} = 45$$

$$y - 291.25 = -37.2\,(x - 45)$$

$$y = -37.2\,x + 1674 + 291.25$$

$$\boxed{y + 37.2\,x = 1965.25}$$



Q. 9. In an air pollution study, a random sample of 200 households was selected from each of 2 communities. The respondent in each house was asked whether or not anyone in the house was bothered by air pollution. The responses are tabulated below (Given χ_1^2 (0.05) = 3.841) :

(Community	Yes	No	Total
I	43	157	200
II	81	119	200
Total	124	276	400

Can the researchers conclude that the 2 communities are bothered differently by air pollution? (α = 0.05)

$$E(43) = \frac{124 \times 200}{400} = 62$$ **Ans.**

The complete table of expected frequency is

Table of Expected Frequencies

62	200 – 62 = 138	200
124 – 62 = 62	276 – 138 = 138	200
124	276	400

Computation of χ^2

O	E	O –E	(O – E)²
43	62	–19	361
81	62	19	361
157	138	19	361
119	138	– 19	361
400	**400**	**00**	**1444**

$$\therefore \quad X^2 = \sum\left[\frac{(O - E)^2}{E}\right]$$

$$X^2 = 361\left[\frac{1}{62} + \frac{1}{62} + \frac{1}{138} + \frac{1}{138}\right]$$

$$= 361\ [0.0161 + 0.0161 + 0.0072 + 0.0072]$$

$$= 361 \times 0.04664$$

$$= 16.8393$$

$$d.f = (z - 1)(2 - 1) = 1$$

Tabulated $\chi^2_{0.05}$ for 1 $d.f$ = 3.841

Since the calculated value of x^2, 16.83 is much greater that the tabulated value of x^2 at 5% level of

significance; the value of x^2 is highly significant and null hypothesis is rejected.

Hence, we conclude that two community (I& II) differ significantly as regards the consumption of air pollution among them.

Q. 10. The Police plans to enforce speed limits by using radar traps at 4 different locations within the city limits. The radar traps at each of the location L_1 , L_2, L_3 and L_4 are operated 40%, 30%, 20% and 30% of the time. If a person who is speeding on his way to work has probabilities of 0.2, 0.1, 0.5 and 0.2 respectively, of passing through these locations, what is the probability that he will receive a speeding ticket? Find also the probability that he will receive a speeding ticket at locations L_1, L_2, L_3 and L_4.

Sol. Let L_1, L_2, L_3 and L_4 the radar traps at 4 different locations.

Then we have

L_i	L_1	L_2	L_3	L_4	Total
$P(L_i)$	0.40	0.30	0.20	0.30	
$P(L/L_i)$	0.2	0.1	0.5	0.2	
$P(L \cap /L_i)$	0.08	0.03	0.1	0.06	$P(L) = 0.27$

According to Bayer's rule as

$$P\left(\frac{L_1}{L}\right) = \frac{P(L_1)\, P\left(\frac{L}{L_1}\right)}{\Sigma P(L_i)\, P\left(\frac{L}{L_1}\right)} = \frac{0.08}{0.27} = 0.296$$

Similarly we get

$$P\left(\frac{L_2}{L}\right) = \frac{0.03}{0.27} = 0.111$$

$$P\left(\frac{L_3}{L}\right) = \frac{0.1}{0.27} = 0.370$$

$$P\left(\frac{L_4}{L}\right) = \frac{0.06}{0.27} = 0.222$$ **Ans.**

■ ■

QUESTION PAPER

(June – 2014)

(Solved)

STATISTICAL TECHNIQUES

Time: 2 hours] ***[Maximum Marks: 50***

Note: *(i) Attempt both* ***Sections*** *i.e.* ***Section A*** *and* ***Section B.*** *(ii) Attempt* ***any four*** *questions from* ***Section A.*** *(iii) Attempt* ***any three*** *questions from* ***Section B.*** *(iv) Use of* ***Non-scientific*** *calculator is* ***allowed.***

SECTION–A

Q. 1. With the help of an suitable example, describe the term "Probability Distribution". How the Binomial Distribution differes from the Poissons Distribution?

Ans. A probability distribution is a table or an equation that links each outcome of a statistical experiment with its probability of occurrence.

Probability Distribution Prerequisites

To understand probability distribution, it is important to understand variables, random variables, and some notations.

- A **variable** is a symbol (A, B, *x*, *y*, etc.) that can take on any of a specified set of values.
- When the value of a variable is the outcome of a statistical experiment, that variable is a **random variable**.

Generally, statisticians use a capital letter to represent a random variable and a lower-case letter, to represent one of its values. For example,

- X represents the random variable X.
- P(X) represents the probability of X.
- $P(X = x)$ refers to the probability that the random variable X is equal to a particular value, denoted by *x*. As an example, $P(X = 1)$ refers to the probability that the random variable X is equal to 1.

An example will make clear the relationship between random variables and probability distributions. Suppose you flip a coin two times. This simple statistical experiment can have four possible outcomes: HH, HT, TH, and TT. Now, let the variable X represent the number of Heads that result from this experiment. The variable X can take on the values 0, 1, or 2. In this example, X is a random variable; because its value is determined by the outcome of a statistical experiment.

Suppose a die is tossed. What is the probability that the die will land on 6?

Solution: When a die is tossed, there are 6 possible outcomes represented by S = { 1, 2, 3, 4, 5, 6 }. Each possible outcome is a random variable (X), and each outcome is equally likely to occur. Thus, we have a uniform distribution. Therefore, the $P(X = 6) = 1/6$.

Binomial Distribution differs from the Poissons Distribution

Binomial Distribution	Poisson Distribution
• Fixed number of trials *(n)* [10 pie throws]	Infinite number of trials.
• Only 2 possible outcomes [hit or miss]	Unlimited number of outcomes possible
• Probability of success is constant *(p)* [0.4 success rate]	Mean of the distribution is the same for all intervals *(mu)*
• Each trial is independent [throw 1 has no effect on throw 2]	Number of occurrences in any given interval independent of others
• Predicts number of successes within a Set number of trials	Predicts number of occurrences per unit time, space, ...
• Can be used to test for independence	Can be used to test for independence.

Q. 2. Suppose A and B are two independent events, associated with an random experiment. If the probability of occurrence of either A or B equals 0.6; while probability that only A occurs equals 0.4; then determine the probability of occurrence of event B.

Sol. If events A and B are associated

$$P(A \text{ or } B) = P(A) + P(B)$$
$$.6 = .4 + P(B)$$
$$P(B) = .6 - .4$$
$$P(B) = .2$$

Q. 3. A sample of size $n = 50$, is drawn from the population of 200 observations. If standard deviation of the data is 22, then find the standard error.

Sol. $n = 50$

Standard deviation $(\sigma) = 22$

$$\text{SE (Standard Error)} = \frac{\sigma}{\sqrt{n}}$$
$$= \frac{22}{\sqrt{50}}$$
$$= \frac{22}{7.07}$$
$$= 3.11$$

Q. 4. Construct Model ANOVA table for one-way classification.

Ans. The *t*-test is commonly used to test the equality of two population means when the data are composed of two random samples. We wish to extend this procedure so that the equality of $r \geq 2$ population means can be tested using r independent samples. Thus the hypothesis and the alternatives are

$H_0 : \mu_0 = \mu_2 = ... = \mu_r$

H_1 : at least two means are not equal

where μ_j, J = 1, 2, ..., r is the mean of the j^{th} population.

It is not hard to imagine situations in which it is of interest to compare a number of means. For example, 5 varieties of corn are available, and it is to be determined whether or not the average yield from each variety is the same; a company is testing 3 brands of bicycle tires and wants to know if the average life of each brand is the same; 4 teaching methods are being investigated for their effectiveness; an automotive company wants to determine which of 4 seat-belt designs would provide the best protection in the event of a head-on collision; a drug company would like to compare the effectiveness of 6 different drugs for treating diabetes.

In designing an experiment for a one-way classification, units are assigned at random to any one of the r treatments under investigation. For this reason, the one-way classification is sometimes referred to as a completely randomized design.

Notataion

Samples from each of the r populations are collected.

X_{ij} = the i^{th} observation receiving treatment j, $i = 1, 2,\ldots, n_j$; $j = 1, 2,\ldots, r$

$$\overline{X}_{.j} = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij} \text{ mean of sample } j$$

$$s_j^2 = \frac{1}{n_j - 1}\sum_{j=1}^{n_j}\left(X_{ij} - \overline{X}_{.j}\right)^2$$

variance of sample j

$$\overline{X}_{..} = \frac{1}{N_j}\sum_{i=1}^{r}\sum_{i=1}^{n_j} X_{ij}, N\sum_{i=1}^{r} n_j$$

$$s^2 = \frac{1}{N-1}\sum_{j=1}^{r}\sum_{i=1}^{n_j}\left(X_{ij} - \overline{X}_{..}\right)^2$$

= Variance of all N observations

μ_j and σ_j^2, $j = 1, 2, ..., r$, denote the mean and variance of population j

Here's one way the data can be arranged once it is collected:

Treatments

	1	**2**	**3**	...	**r**	
Observations	X_{11}	X_{12}	X_{13}		X_{1r}	
	X_{21}	X_{22}	X_{23}		X_{2r}	
	.	.	.		.	
	.	.	.		.	
	.	.	.		.	
	$X_{n_1 1}$		$X_{n_3 3}$			
		$X_{n_2 2}$			$X_{n_1 r}$	
Means	$\overline{X}_{.1}$	$\overline{X}_{.2}$	$\overline{X}_{.3}$		$\overline{X}_{.r}$	$\overline{X}_{..}$
Variances	s_1^2	s_2^2	s_3^2		s_r^2	s^2

Q. 5. Write short notes on (*any two*):

***(a) t* - test for Mean**

Ans. Ref.: See Chapter 4, Page No. 55 '*t*-distribution'

***(b)* F - test for equality of two variances**

Ans. Ref.: See Chapter 4, Page No. 56 'F-distribution'

***(c)* Chi-square - test for independence of Attributes.**

Ans. Ref.: See Chapter 4, Page No. 55 'Chi-square Distribution'

SECTION–B

Q. 6. Using the regression line $\hat{Y} = 90 + 50\ X$, fill up the values in the table below:

Sample No. *(i)*	12	21	15	1	24
x_i	0.96	1.28	1.65	1.84	2.35
y_i	138	160	178	190	210
$\hat{y}_i$	138	-	-	-	-
e_i	0	-	-	-	-

After filling the table, compute the parameters R and R^2, finally interpret the correlation between X and Y.

Sol. Regression line $\hat{Y} = 90 + 50\ X$

As we know that

$$R = \frac{S_{xy}}{\sqrt{S_{xx}\ S_{yy}}}$$

$$= \frac{138.076}{\sqrt{2.6929 \times 7839.93}}$$

$$= 09504$$

$$R^2 = .9033$$

Q. 7. What do you understand by the term forecasting? With the help of a suitable example discuss the relation between forecasting and future planning. Briefly discuss both forecasting model.

Ans. Forecasting means predicting something that will happen in the future or estimating the probability that it will happen. Predicting weather conditions is a common application of forecasting. Meteorologists link their past experiences with weather conditions and scientific information about weather patterns to predict what the weather will be like in the future. Weather forecasts often are correct; and of course, they frequently are mistaken, either in terms of the intensity or timing of weather conditions. Weather forecasts are used to plan activities, such as when to go on a picnic, when to mow the lawn, or what clothing to wear on a particular day. Sometimes forecasts are wrong and the picnic gets rained out, the lawn gets too dry, or the outfit selected is too cool for the colder-than-expected temperatures.

People use forecasting and planning in many other aspects of their personal and professional lives:

- They forecast that their jobs are secure and their income will be sufficient to purchase a house.
- They estimate that they can complete a task at work within a given amount of time and, therefore, can also accept responsibility for an additional assignment.
- They think it is likely that a business meeting will be long, and they forego dinner plans.
- They predict that additional training and education will make them more marketable and, therefore, invest in education.

Forecasting for juvenile corrections is similar to these examples. It involves examining the past and present for quantitative information and trends and qualitative patterns. These are applied toward making predictions of numbers and tendencies based on experience and logical assumptions of what will happen in the future. Plans are then based on the probability that those forecasts will be reasonably accurate.

Forecasts often miss the mark of 100 per cent accuracy. However, they provide the best basis for planning available. Considering many variables is vital when developing forecasts. The social, economic, and political contexts always must be considered when forecasts are used. For example, in a largely industrial setting, one must consider the possible effects if a major manufacturer were to close operations. Employment

and economic conditions in a community or state could change drastically. Many prognosticators believe changes in welfare will increase the Nation's population of poor children, at least temporarily. Political tides and public opinions often vacillate between liberal and conservative viewpoints, and these changes frequently prescribe different responses to delinquency. Some of these consequences are reasonably predictable, while others are unexpected and probably cannot be incorporated in jurisdictional or program forecasts and plans. Neither can the actions of particular individuals necessarily be predicted. For example, a judge who believes only youth who commit a second violent offense should spend some time in confinement, or a probation officer who takes youth back to court for any and all violations of conditions of probation will both affect related juvenile corrections programs.

Relation between Future Planning and Forecasting: Planning and forecasting are inextricably intertwined each other. Planning is concerned with future which is highly uncertain. Therefore planners have to make assumptions about the future events. In order to make correct an assumption prediction of future events is essential. Forecasting is the primary source of planning premises which serve as the foundation for building the superstructure of planning. The information generated through forecasting service is an input to planning. Forecasting is an integral part of the planning processes to the extent that it provides the necessary basis for charting out the future course of action.

Forecasting is prerequisite to planning. Forecasting indicate the probable course of future events, plans decide how to prepare for these events. Without forecasting will be a futile mental exercise and the organization would be at the mercy of future events. For example, a business enterprise predicts competitive technological, social and political conditions likely to prevail over the next five years.

On the basis of these forecasts, it determines objectives; strategies and policies concerning sales grow with, product range, market coverage, competitive initiative, profitability, etc. Planning and forecasting both are concerned with future. However, there is some difference between the two and difference lies in the scope of the two processes. Planning is more comprehensive including many elements and sub-processes to arrive at decisions concerning what is to be done, how to be done, and when to be done. Forecasting involves estimates of future events and provides parameters to planning. Planning result in the commitment of resources; whereas no such commitment is involved in forecasting.

A large number of persons are involved in the planning processes tough major decisions are at the top level. Forecasting is normally done at middle or lower level. It may be entrusted to staff specialties whose decisions do forecasting does not involve decision making but helps in decision making by providing clues about what is likely to happen future. Fourthly forecasting involves what the future is likely to be and is likely to behave. Planning, on the other hand indicates what the future is desired to be and how to make it a reality.

In fact, forecasting is the essence of planning because the future course of action is determined in the light of the forecast made.

Q. 8. Differentiate between following (any two):

(a) **Linear and circular systematic sampling**

(b) **Z-test and *t*-test**

(c) **Correlation and Regression**

Ans. Ref.: See Chapter 12, Page No. 149 'Linear Systematic Sampling' and 'Circular Systematic Sampling'

(b) **Z-test and *t*-test:** A *z*-test is used for testing the mean of a population versus a standard, or comparing the means of two populations, with large samples whether you know the population standard deviation or not. It is also used for testing the proportion of some characteristic versus a standard proportion, or comparing the proportions of two populations.

Examples: *(i)* Comparing the average engineering salaries of men versus women.

(ii) Comparing the fraction defectives from 2 production lines.

A *t*-test is used for testing the mean of one population against a standard or comparing the means of two populations if you do not know the populations' standard deviation and when you have a limited sample ($n < 30$). If you know the populations' standard deviation, you may use a Z-test.

Example: Measuring the average diameter of shafts from a certain machine when you have a small sample.

(c) **Correlation and Regression:** Correlation and linear regression are the most commonly used techniques for investigating the relationship between two quantitative variables.

The goal of a correlation analysis is to see whether two measurement variables vary, and to quantify the strength of the relationship between the variables, whereas regression expresses the relationship in the form of an equation.

For example, in students taking a Math and English test, we could use correlation to determine whether students who are good at Math end to be good at English as well, and regression to determine whether the marks in English can be predicted for given marks in Maths.

In regression analysis, we have to find out the relationship between the dependent variable (response) and the (explanatory) independent variable.

The analysis consists of putting an appropriate model, done by the method of least squares, with a view to exploiting the relationship between the variables to estimate the expected response for a given value of the independent variable. For example, if we are interested in the effect of age on height, then by fitting a regression line, we can predict the height for a given age.

Q. 9. ***(a)*** **Compare and contrast Random Sampling with Non-Random Sampling. Briefly discuss the methods involved in the selection of any simple random sample.**

Ans. Random with Non-Random Sampling Methods: Although random sampling is generally the preferred survey method, few people doing surveys use it because of prohibitive costs; i.e., the method requires numbering each member of the survey population, whereas non-random sampling involves taking every *n*th member. Findings indicate that as long as the attribute being sampled is randomly distributed among the population, the two methods give essentially the same results. If the attribute is not randomly distributed, the two methods give radically different results. In some instances the non-random methods yield much better inferences about the population; in other instances, its inferences are much worse.

It is possible to have both random selection and assignment in a study. If we draw a random sample of 100 clients from a population of 1000 current clients of an organization. That is random sampling. If we randomly assign 50 of these clients to get some new additional treatment, that will be called random assignment.

It is also possible to have only one of these (random selection or random assignment) but not the other in a study. For instance, if you do not randomly draw the 100 cases from your list of 1000 but instead just take the first 100 on the list, you do not have random selection. But you could still randomly assign this non-random sample to treatment versus control. Or, you could randomly select 100 from your list of 1000 and then non-randomly (haphazardly) assign them to treatment or control.

And, it's possible to have neither random selection nor random assignment. In a typical non-equivalent groups design in education you might non-randomly choose two 5th grade classes to be in your study. This is non-random selection. Then, you could arbitrarily assign one to get the new educational program and the other to be the control. This is non-random (or non-equivalent) assignment.

Random selection is related to sampling. Therefore it is most related to the external validity (or generalizability) of your results. After all, we would randomly sample so that our research participants better represent the larger group from which they're drawn. Random assignment is most related to design. In fact, when we randomly assign

participants to treatments we have, by definition, an experimental design. Therefore, random assignment is most related to internal validity. After all, we randomly assign in order to help assure that our treatment groups are similar to each other (i.e., equivalent) prior to the treatment.

(b) **Calculate an estimate of Median for following data:**

Class	0-24.9	25-49.9	50-74.9	75-99.9	100-124.9	125-149.0
Frequency	6	11	14	16	13	10

Sol. L = 74.9
(the lower class limit of the 75–99.9)

$h = 70$
$Cf_b = 6 + 11 + 14$
$= 31$
$f_m = 16$
$w = 6$

$$\text{Median} = L + \frac{h/2 - Cf_b}{f_m} \times w$$

$$= 74.9 + \frac{35 - 31}{16} \times 6$$

$$= 74.9 + 1.5$$

$$= 76.4$$

■■

QUESTION PAPER

(June – 2013)

(Solved)

STATISTICAL TECHNIQUES

Time: 2 hours] ***[Maximum Marks: 50***

Note: *(i) Attempt both* ***sections A*** *and* ***Section B.***
(ii) Attempt ***any four*** *questions from* ***Section A.***
(iii) Attempt ***any three*** *questions from* ***Section B.***
(iv) Use of ***Non-scientific*** *calculator is* ***allowed.***

SECTION–A

Q. 1. Define the term Random Experiment and Random Variable. Briefly discuss the types of Random Variables, with suitable examples.

Ans. Random Experiments: Probability theory is based on the paradigm of a random experiment; that is, an experiment whose outcome cannot be predicted with certainty, before the experiment is run. We usually assume that the experiment can be repeated indefinitely under essentially the same conditions. This assumption is important because probability theory is concerned with the long-term behaviour as the experiment is replicated. Naturally, a complete definition of a random experiment requires a careful definition of precisely what information about the experiment is being recorded, that is, a careful definition of what constitutes an outcome.

Examples and Simulations: The *dice experiment* consists of rolling a pair of (distinct) dice and recording the number spots showing on each die.

Buffon's coin experiment consists of tossing a coin with radius $r < 1$ on a floor covered with square tiles of side length 1. The coordinates of the center of the coin are recorded, relative to axes through the center of the square, parallel to the sides.

1. Run the simulation of the dice experiment 100 times and observe the results.

2. Run the simulation of Buffon's coin experiment 100 times and observe the results.

Random Variables: Suppose again that we have a random experiment with sample space S. A function X from S into another set T is called a (T-valued) *random variable*. Probability has its own notation, very different from other branches of mathematics. As a case in point, random variables are usually denoted by capital letters near the end of the alphabet.

A random variable X as a measurement of interest in the context of the random experiment. A random variable X is random in the sense that its value depends on the outcome of the experiment, which cannot be predicted with certainty before the experiment is run. Each time the experiment is run, an outcome $s \in S$ occurs, and a given random variable X takes on the value X(s). In general, as you will see, the notation of probability suppresses references to the sample space. Indeed, sometimes the sample space is hidden in the sense that we don't know what it is.

Random variables can be discrete or continuous.

● **Discrete:** Within a range of numbers, discrete variables can take on only certain values. Suppose, for example, that we flip a coin and count the number of heads. The number of heads will be a value between zero and plus infinity. Within that range, though, the number of heads can be only certain values. For example, the number of heads can only be a whole number, not a fraction. Therefore, the number of heads is a discrete variable. And because the number of heads results from a random process - flipping a coin - it is a discrete random variable.

● **Continuous:** Continuous variables, in contrast, can take on any value within a range of values. For example, suppose we randomly select an individual from a population. Then, we measure the age of that person. In theory, his/her age can take on any value between zero and plus infinity, so age is a continuous variable. In this example, the age of the person

selected is determined by a chance event; so, in this example, age is a continuous random variable.

Q. 2. The Probability that atleast one of the two Independent events occur is 0.5. Probability that first event occurs but not the second is (3/25). Also the probability that the second event occurs but not the first is (8/25). Find the probability that none of the two event occurs.

Sol. $P(A) = \frac{3}{25}$

$P(B) = \frac{8}{25}$

$P(A \cap B) = .5$

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

$= \frac{3}{25} + \frac{8}{25} - .5$

$= .12 + .32 - .5$

$= .06$

Q. 3. Marks of six students are tabulated below:

Name	Raj	Anil	Amit	Om	Rita	Renu
Marks	54	50	52	48	50	52

From the population, tabulated above, you are suppose to choose a sample of size two.

(a) **Determine, how many samples of size two are possible.**

(b) **Construct sampling distribution of means by taking samples of size 2 and organize the data.**

Ans. *(a)* 2

(b) $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

$\sigma = \frac{2}{\sqrt{6}} = \frac{2}{2.44} = .81$

$\sigma_{\bar{x}} = \frac{.81}{\sqrt{2}} = \frac{.81}{1.41} = .57$

Q. 4. Expand the term ANOVA. Briefly discuss the utility of ANOVA, with suitable examples.

Ans. Ref.: See Chapter-8, Page No. 97 'Analysis of Variance: Basic Concepts' and Page No. 102 Example

Q. 5. List the advantages and disadvantages of using a sampling approach instead of a census approach for studying the characteristics of data.

Ans. Advantages of using sampling approach instead of a census approach are:

(i) **Speed or Saving in Time:** Under this technique a statistical investigation is carried out speedily and consequently a lot of time and energy is saved not only in the collection of the data, but also in the processing, editing and analyzing these data. This is because in sample enquiry, only a part of the universe is contacted and studied instead of each and every unit of the universe. As such this technique is more preferred to census technique where the results of the investigation are needed most urgently and quickly.

(ii) **Economy or Saving in Time:** Under this technique a lot of expenses are saved both in terms of money and energy not only in the collection of data, but also in the administration, transport and training etc. This is because in sample technique only an action of the population is studied and examined to arrive at the desired conclusion, thus, economy is maintained in all the phases of the enquiry conducted under this technique.

(iii) **Adaptability:** The sample technique of data collection, unlike census technique is very much adaptable to the changes in the circumstances of the universe. This means that the size of the sample can be increased or decreased according to the size of the universe, availability of resources and the degree of accuracy desired.

Similarly, the nature of the sample technique, also, can be changed according to the nature, object and scope of the enquiry. Thus the sample technique of data collection is very much flexible and adaptable to the changes in conditions of the enquiry.

(iv) **Scientific Approach:** The sample technique of data collection is very much scientific in its approach. Particularly, the techniques of random sampling are based on the Theory of Probability which is a mathematical concept.

Besides, this technique is based on certain important laws and principles viz.

(i) Law of Statistical Regularity.
(ii) Law of Inertia of Large Numbers,
(iii) Principle of Persistence,
(iv) Principle of Optimization,
(v) Principle of validity,
(vi) It is also possible to ascertain the extent of sampling error and degree of reliability of the results under this technique.

***(v)* Administrative Convenience:** Under this technique of data collection we find administrative convenience in as much as the number of units to be studied here is usually very limited and the number both of field and administrative staff to be maintained, therefore, is very less.

There is little botheration in the recruitment, training control and supervision of the various staff required under this method of enquiry their number will be very few.

***(vi)* Dependability or Reliability:** The result of enquiry derived under the sample technique is more dependable than those derived under the census technique. This is because under the sample technique it is always possible to determine the extent sampling errors and the degree of reliability of the results in terms of probability.

***(vii)* Indispensability:** The sample technique of data collection is found indispensable in certain types of universe viz. infinite universe, hypothetical universes and universe liable to be destroyed through testing. In case of these universes technique can never be applied.

Disadvantages of using Sampling Approach instead of a Census Approach

***(i)* Inaccuracy in Results:** The sample technique may lead to inaccurate results, if the sample are not selected properly and the personnel conducting the survey are not properly trained and have bias.

***(ii)* Expensive:** The sample technique may be expensive in terms of time and energy, if the size of the sample is considerably large and the procedure adopted in the technique is complicated. Moreover, it needs the services of qualified, skilled, experienced and trained personnel which entail heavy amount of expenditure.

***(iii)* Unsuitability:** In certain cases of statistical study, where, information needed from each and every unit of the universe and the universe comprises of the of heterogeneous nature, the sample technique of data collection is not at all suitable. Moreover, if the field of enquiry is very small, this technique of data collection is suitable.

***(iv)* Inherent Defects in the Methods:** Each of the various methods of the sample technique suffers from certain inherent defects which can never be rooted out. Hence, the results obtained through the sample technique can never be flawless or free from all defects.

Q. 6. Given the following sample of 10 numbers

12 41 48 58 14 43 50 59 15 79

Compute Mean deviation and Standard deviation for the data given above.

Sol. Mean

$$\bar{x} = \frac{12+41+48+58+14+43+50+59+15+79}{10}$$

$$\bar{x} = \frac{419}{10}$$

$$\bar{x} = 41.9$$

Sample Standard Deviation $S = \sqrt{\frac{\Sigma\left(x_i - \bar{x}\right)^2}{n-1}}$

$$\Sigma\left(x_i - \bar{x}\right)^2 = (12–41.9)^2+(41.0–41.9)^2$$
$$+ (48–41.9)^2+(58–41.9)^2+(14–41.9)^2$$
$$+ (43–41.9)^2+(50–41.9)^2+(59–41.9)^2$$
$$+(15–41.9)^2+(79–41.9)^2$$
$$= (–29.9)^2+(–.09)^2+(6.1)^2+(16.1)^2+(–27.9)^2$$
$$+(1.1)^2 +(8.1)^2+(17.1)^2+(–26.9)^2+(37.1)^2$$
$$= 894.01+.81+37.21+259.21+778.41+1.21$$
$$+65.61+292.41+723.61+1376.41$$
$$= 4428.9$$

$$S = \frac{\sqrt{4428.9}}{9} = \sqrt{492.1} = 22.18$$

SECTION–B

Q. 7. Explain any two of the following with the help of an example each:

***(a)* Godness of fit test**

***(b)* Test of Independence**

***(c)* Criteria for a good estimator**

Ans. *(a)* Goodness of Fit Test: The test is applied when you have one categorical variable from a single population. It is used to determine whether sample data are consistent with a hypothesized distribution.

For example, suppose a company printed baseball cards. It claimed that 30% of its cards were rookies; 60%, veterans; and 10%, All-Stars. We could gather a random sample of baseball cards and use a chi-square goodness of fit test to see whether our sample distribution differed significantly from the distribution claimed by the company.

When to use the Chi-Square Goodness of Fit Test?

The chi-square goodness of fit test is appropriate when the following conditions are met:

- The sampling method is simple random sampling.
- The population is at least 10 times as large as the sample.
- The variable under study is categorical.
- The expected value of the number of sample observations in each level of the variable is at least 5.

(b) **Test for Independence:** The test is applied when you have two categorical variables from a single population. It is used to determine whether there is a significant association between the two variables.

For example, in an election survey, voters might be classified by gender (male or female) and voting preference (Democrat, Republican, or Independent). We could use a chi-square test for independence to determine whether gender is related to voting preference. The sample problem at the end of the lesson considers this example.

When to Use Chi-Square Test for Independence?

The test procedure described is appropriate when the following conditions are met:

- The sampling method is simple random sampling.
- Each population is at least 10 times as large as its respective sample.
- The variables under study are each categorical.
- If sample data are displayed in a contingency table, the expected frequency count for each cell of the table is at least 5.

(c) **Criteria for a Good Estimator:** An estimator is a statistical parameter that provides an estimation of a population parameter.

Example: The mean of the age of men attending a show is $28 \leq \text{age} \leq 36$.

There are two criteria used to establish the endpoints of an interval estimator:

(1) **The level of precision:** How sure you want to be about its values.

(2) **The credibility:** How believable is the estimator.

Both precision and credibility of interval estimators improves with the increasing quality and quantity of the sample.

An **efficient estimator** considers the reliability of the estimator in terms of its tendency to have a smaller standard error for the same sample size when compared each other.

Examples: The median is an unbiased estimator of μ when the sample distribution is normally distributed; but standard error is 1.25 greater than that of the sample mean, so the sample mean is a more efficient estimator than the median.

Q. 8. Explain the term "Time Series". Briefly discuss any two categories of time series analysis.

Ans. Time Series: An ordered sequence of values of a variable at equally spaced time intervals.

Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. **Time series forecasting** is the use of a model to predict future values based on previously observed values. While regression analysis is often employed in such a way as to test theories that the current values of one or more independent time series affect the current value of another time series, this type of analysis of time series is not called "time series analysis", which focuses on comparing values of a single time series or multiple dependent time series at different points in time.

Time series analysis can be used to accomplish different goals:

(1) Descriptive analysis determines what trends and patterns a time series has by plotting or using more complex techniques. The most basic approach is to graph the time series and look at:

- Overall trends (increase, decrease, etc.)
- Cyclic patterns (seasonal effects, etc.)
- Outliers - points of data that may be erroneous
- Turning points - different trends within a data series.

(2) Spectral analysis is carried out to describe how variation in a time series may be accounted for by cyclic components. This may also be referred to as "Frequency Domain". With this an estimate of the spectrum over a range of frequencies can be obtained and periodic components in a noisy environment can be separated out.

Example: What is seen in the ocean as random waves may actually be a number of different frequencies and amplitudes that are quite stable and predictable. Spectral analysis is used on the wave

height vs. time to determine which frequencies are most responsible for the patterns that are there, but can't be readily seen without analysis.

(3) Forecasting can do just that - if a time series has behaved a certain way in the past, the future behaviour can be predicted within certain confidence limits by building models.

Example: Tidal charts are predictions based upon tidal heights in the past. The known components of the tides (e.g., positions of the moon and sun and their weighted values) are built into models that can be employed to predict future values of the tidal heights.

(4) Intervention analysis can explain if there is a certain event that occurs that changes a time series. This technique is used a lot of the time in planned experimental analysis. In other words–Is there a change in a time series before and after a certain event?

Examples:

1. If a plant's growth rate before changing the amount of light it gets is different from that afterwards, an intervention has occurred - the change in light is the intervention.

2. When a community of goats changes its behavior after a bear shows up in the area, then there may be an intervention.

(5) Explanative Analysis (Cross Correlation): Using one or more variable time series, a mechanism that results in a dependent time series can be estimated. A common question to be answered with this analysis would be "What relationship is there between two time series data sets?"

Categories of Time Series Analysis

1. Trend Analysis: There are no proven "automatic" techniques to identify trend components in the time series data; however, as long as the trend is monotonous (consistently increasing or decreasing) that part of data analysis is typically not very difficult. If the time series data contain considerable error, then the first step in the process of trend identification is smoothing.

Smoothing: Smoothing always involves some form of local averaging of data such that the non-systematic components of individual observations cancel each other. The most common technique is moving average smoothing which replaces each element of the series by either the simple or weighted average of n surrounding elements, where n is the width of the smoothing "window".

Medians can be used instead of means. The main advantage of median as compared to moving average smoothing is that its results are less biased by outliers (within the smoothing window). Thus, if there are outliers in the data (e.g., due to measurement errors), median smoothing typically produces smoother or at least more "reliable" curves than moving average based on the same window width. The main disadvantage of median smoothing is that in the absence of clear outliers it may produce more "jagged" curves than moving average and it does not allow for weighting.

In the relatively less common cases (in time series data), when the measurement error is very large, the distance weighted least squares smoothing or negative exponentially weighted smoothing techniques can be used. All those methods will filter out the noise and convert the data into a smooth curve that is relatively unbiased by outliers. Series with relatively few and systematically distributed points can be smoothed with bicubic splines.

Fitting a Function: Many monotonous time series data can be adequately approximated by a linear function; if there is a clear monotonous non-linear component, the data first need to be transformed to remove the non-linearity. Usually a logarithmic, exponential, or (less often) polynomial function can be used.

2. Analysis of Seasonality: Seasonal dependency (seasonality) is another general component of the time series pattern. This concept is formally defined as correlational dependency of order k between each i'th element of the series and the (i-k)th element and measured by autocorrelation (i.e., a correlation between the two terms); k is usually called the lag. If the measurement error is not too large, seasonality can be visually identified in the series as a pattern that repeats every k elements.

Auto correlation correlogram: Seasonal patterns of time series can be examined via correlograms. The correlogram (autocorrelogram) displays graphically and numerically the autocorrelation function (ACF), that is, serial correlation coefficients (and their standard errors) for consecutive lags in a specified range of lags (e.g., 1 through 30). Ranges of two standard errors for each lag are usually marked in correlograms but typically the size of auto correlation is of more interest than its

reliability because we are usually interested only in very strong autocorrelations.

Q. 9. Explain any two of the follwing:

***(a)* Cluster sampling**

***(b)* Stratified random sampling**

***(c)* Systematic sampling**

Ans. *(a)* Cluster sampling refers to a sampling method that has the following properties:

- The population is divided into N groups, called clusters.
- The researcher randomly selects *n* clusters to include in the sample.
- The number of observations within each cluster M_1 is known, and $M = M_1 + M_2 + M_3 + ... + M_{N-1} + M_N$.
- Each element of the population can be assigned to one, and only one, cluster.

Two Types of Cluster Sampling Methods

- **One-stage Sampling:** All of the elements within selected clusters are included in the sample.
- **Two-stage Sampling:** A subset of elements within selected clusters are randomly selected for inclusion in the sample.

Advantages and Disadvantages: Assuming the sample size is constant across sampling methods, cluster sampling generally provides less precision than either simple random sampling or stratified sampling. This is the main disadvantage of cluster sampling.

Sometimes, the cost per sample point is less for cluster sampling than for other sampling methods. Given a fixed budget, the researcher may be able to use a bigger sample with cluster sampling than with the other methods. When the increased sample size is sufficient to offset the loss in precision, cluster sampling may be the best choice.

***(b)* Stratified random sampling** refers to a sampling method that has the following properties:

- The population consists of N elements.
- The population is divided into H groups, called **strata.**
- Each element of the population can be assigned to one, and only one, stratum.
- The number of observations within each stratum N_h is known, and $N = N_1 + N_2 + N_3 + ... + N_{H-1} + N_H$.
- The researcher obtains a probability sample from each stratum.

Advantages and Disadvantages

Stratified sampling offers several advantages over simple random sampling:

- A stratified sample can provide greater precision than a simple random sample of the same size.
- Because it provides greater precision, a stratified sample often requires a smaller sample, which saves money.
- A stratified sample can guard against an "unrepresentative" sample (e.g., an all-male sample from a mixed-gender population).
- We can ensure that we obtain sufficient sample points to support a separate analysis of any subgroup.

The main disadvantage of a stratified sample is that it may require more administrative effort than a simple random sample.

***(c)* Systematic sampling** is a random sampling technique which is frequently chosen by researchers for its simplicity and its periodic quality.

In systematic random sampling, the researcher first randomly picks the first item or subject from the population. Then, the researcher will select each *n*th subject from the list.

The procedure involved in systematic random sampling is very easy and can be done manually. The results are representative of the population unless certain characteristics of the population are repeated for every *n*th individual, which is highly unlikely.

The process of obtaining the systematic sample is much like an arithmetic progression.

1. Starting Number: The researcher selects an integer that must be less than the total number of individuals in the population. This integer will correspond to the first subject.

2. Interval: The researcher picks another integer which will serve as the constant difference between any two consecutive numbers in the progression.

The integer is typically selected so that the researcher obtains the correct sample size

For example, the researcher has a population total of 100 individuals and need 12 subjects. He first picks his starting number, 5.

Then the researcher picks his interval, 8. The members of his sample will be individuals 5, 13, 21, 29, 37, 45, 53, 61, 69, 77, 85, 93.

Advantages of Systematic Sampling

- The main advantage of using systematic sampling over simple random sampling is its simplicity. It allows the researcher to add a degree of system or process into the random selection of subjects.
- Another advantage of systematic random sampling over simple random sampling is the assurance that the population will be evenly sampled. There exists a chance in simple random sampling that allows a clustered selection of subjects. This is systematically eliminated in systematic sampling.

Disadvantage of Systematic Sampling

- The process of selection can interact with a hidden periodic trait within the population. If the sampling technique coincides with the periodicity of the trait, the sampling technique will no longer be random and representativeness of the sample is compromised.

Q. 10. A comany wants to estimate, how its monthly costs are related to its monthly output rate. The data for a sample of nine months is tabulated below:

Output (Tons)	1	2	4	8	6	5	8	9	7
Cost (Lakhs)	2	3	4	7	6	5	8	8	6

Using the data given above, perform following tasks:

***(a)* Calculate the best linear regression line, where the monthly output is the dependent variable and monthly cost is the independent variable.**

***(b)* Use the regression line to predict the company's monthly cost, if they decide to produce 4 tons per month.**

Sol. ***(a)***

x	y	x^2	y^2	xy
1	2	1	4	2
2	3	4	9	6
4	4	16	16	16
8	7	64	49	56
6	6	36	36	36
5	5	25	25	25
8	8	64	64	64
9	8	81	64	72
7	6	49	36	42
$\Sigma x=50$	$\Sigma x=49$	$\Sigma x^2=340$	$\Sigma y^2=303$	$\Sigma xy=319$

Regression 4 on x is

$$\left(\hat{y}-\bar{y}\right) = b_{yx}\left(x-\bar{x}\right)$$

$$\bar{y} = \frac{\Sigma y}{N} = \frac{49}{9} = 5.4$$

$$\bar{x} = \frac{\Sigma x}{N} = \frac{50}{9} = 5.5$$

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

$$= \frac{9(319) - (50)(49)}{9(340) - (50)^2} = \frac{2871 - 2450}{3060 - 2500}$$

$$\frac{421}{560} = .75$$

$$\hat{y} - 5.4 = .75\ (x - 5.5)$$

$$\hat{y} = .75X = .75*5.5 + 5.4$$

$$\hat{y} = .75X = 4.125 + 5.4$$

$$\hat{y} = .75X = -9.52$$

(b) When $x = 4$

$$\hat{y} = .75*4 - 9.52$$

$$\hat{y} = 3 - 9.52$$

$$\hat{y} = -6.52$$

■■

QUESTION PAPER

(December – 2013)

(Solved)

STATISTICAL TECHNIQUES

Time: 2 hours] *[Maximum Marks: 50*

Note: *(i) Attempt both **Sections A** and **Section B.***
*(ii) Attempt **any four** question from **Section A.***
*(iii) Attempt **any three** question from **Section B.***
*(iv) Use of **Non-scientific** calculator is **allowed.***

SECTION–A

Q. 1. "Explain the term probability distribution." How Binomial distribution differs from poisson distribution?

Ans. A probability distribution is a table or an equation that links each outcome of a statistical experiment with its probability of occurrence.

Probability Distribution Prerequisites

To understand probability distribution, it is important to understand variables, random variables, and some notations.

- A **variable** is a symbol (A, B, *x*, *y*, etc.) that can take on any of a specified set of values.
- When the value of a variable is the outcome of a statistical experiment, that variable is a **random variable**.

Generally, statisticians use a capital letter to represent a random variable and a lower-case letter, to represent one of its values. For example,

- X represents the random variable X.
- P(X) represents the probability of X.
- P(X = *x*) refers to the probability that the random variable X is equal to a particular value, denoted by *x*. As an example, P(X = 1) refers to the probability that the random variable X is equal to 1.

An example will make clear the relationship between random variables and probability distributions. Suppose you flip a coin two times. This simple statistical experiment can have four possible outcomes: HH, HT, TH, and TT. Now, let the variable X represent the number of Heads that result from this experiment. The variable X can take on the values 0, 1, or 2. In this example, X is a random variable; because its value is determined by the outcome of a statistical experiment.

Suppose a die is tossed. What is the probability that the die will land on 6?

Solution: When a die is tossed, there are 6 possible outcomes represented by S = { 1, 2, 3, 4, 5, 6 }. Each possible outcome is a random variable (X), and each outcome is equally likely to occur. Thus, we have a uniform distribution. Therefore, the P(X = 6) = 1/6.

Binomial Distribution differs from the Poissons Distribution

Binomial Distribution	Poisson Distribution
● Fixed number of trials (*n*) [10 pie throws]	Infinite number of trials
● Only 2 possible outcomes [hit or miss]	Unlimited number of outcomes possible
● Probability of success is constant (*p*) [0.4 success rate]	Mean of the distribution is the same for all intervals (mu)
● Each trial is independent [throw 1 has no effect on throw 2]	Number of occurrences in any given interval independent of others
● Predicts number of successes within a set number of trials	Predicts number of occurrences per unit time, space, ...
● Can be used to test for independence	Can be used to test for independence.

Q. 2. Suppose that A and B are two independent events, associated with a random experiment. If the probability that A or B occurs equals 0.6; while probability that A occurs equals 0.4. Determine the probability that B occurs.

Sol. If events A and B are associated

$$P(A \text{ or } B) = P(A) + P(B)$$
$$.6 = .4 + P(B)$$
$$P(B) = .6 - .4$$
$$P(B) = .2$$

Q. 3. Construct Model ANOVA table for one-way classification.

Ans. The *t*-test is commonly used to test the equality of two population means when the data are composed of two random samples. We wish to extend this procedure so that the equality of $r \geq 2$ population means can be tested using r independent samples. Thus the hypothesis and the alternatives are

$H_0 : \mu_0 = \mu_2 = ... = \mu_r$

H_1 : at least two means are not equal

where μ_j, J = 1, 2, ..., r is the mean of the j^{th} population.

It is not hard to imagine situations in which it is of interest to compare a number of means. For example, 5 varieties of corn are available, and it is to be determined whether or not the average yield from each variety is the same; a company is testing 3 brands of bicycle tires and wants to know if the average life of each brand is the same; 4 teaching methods are being investigated for their effectiveness; an automotive company wants to determine which of 4 seat-belt designs would provide the best protection in the event of a head-on collision; a drug company would like to compare the effectiveness of 6 different drugs for treating diabetes.

In designing an experiment for a one-way classification, units are assigned at random to any one of the r treatments under investigation. For this reason, the one-way classification is sometimes referred to as a completely randomized design.

Notataion

Samples from each of the r populations are collected.

X_{ij} = the i^{th} observation receiving treatment j, $i = 1, 2, \ldots, n_j$; $j = 1, 2, \ldots, r$

$$\overline{X}_{.j} = \frac{1}{n_j}\sum_{i=1}^{n_j} X_{ij} \text{ mean of sample } j$$

$$s_j^2 = \frac{1}{n_j - 1}\sum_{j=1}^{n_j}\left(X_{ij} - \overline{X}_{.j}\right)^2$$

variance of sample j

$$\overline{X}_{..} = \frac{1}{N_j}\sum_{i=1}^{r}\sum_{i=1}^{n_j} X_{ij}, N\sum_{i=1}^{r} n_j$$

$$s^2 = \frac{1}{N-1}\sum_{j=1}^{r}\sum_{i=1}^{n_j}\left(X_{ij} - \overline{X}_{..}\right)^2$$

= Variance of all N observations

μ_j and σ_j^2, $j = 1, 2, ..., r$, denote the mean and variance of population j.

Here's one way the data can be arranged once it is collected.

Treatments

	1	**2**	**3**	**...**	***r***	
Observations	X_{11}	X_{12}	X_{13}		X_{1r}	
	X_{21}	X_{22}	X_{23}		X_{2r}	
	.	.	.		.	
	.	.	.		.	
	.	.	.		.	
	$X_{n_1 1}$		$X_{n_3 3}$			
		$X_{n_2 2}$			$X_{n_1 r}$	
Means	$\overline{X}_{.1}$	$\overline{X}_{.2}$	$\overline{X}_{.3}$		$\overline{X}_{.r}$	$\overline{X}_{..}$
Variances	s_1^2	s_2^2	s_3^2		s_r^2	s^2

Q. 4. From a population of 200 observations, a sample of n = 50 is selected. Calculate the standard error; if the population standard deviation equals 22.

Sol. $n = 50$

Standard deviation (σ) = 22

$$\text{SE (Standard Error)} = \frac{\sigma}{\sqrt{n}}$$

$$= \frac{22}{\sqrt{50}}$$

$$= \frac{22}{7.07}$$

$$= 3.11$$

Q. 5. Compare and contrast Random Sampling with Non-Random Sampling. Briefly discuss the methods involved in selection of any sample random sample.

Ans. Random with Non-Random Sampling Methods: Although random sampling is generally the preferred survey method, few people doing surveys use it because of prohibitive costs; i.e., the method requires numbering each member of the survey population, whereas non-random sampling involves taking every *n*th member. Findings indicate that as long as the attribute being sampled is randomly distributed among the population, the two methods give essentially the same results. If the attribute is not randomly distributed, the two methods give radically different results. In some instances the non-random methods yield much better inferences about the population; in other instances, its inferences are much worse.

It is possible to have both random selection and assignment in a study. If we draw a random sample of 100 clients from a population list of 1000 current clients of an organization. That is random sampling. If we randomly assign 50 of these clients to get some new additional treatment that will be called random assignment.

It is also possible to have only one of these (random selection or random assignment) but not the other in a study. For instance, if you do not randomly draw the 100 cases from your list of 1000 but instead just take the first 100 on the list, you do not have random selection. But you could still randomly assign this non-random sample to treatment versus control. Or, you could randomly select 100 from your list of 1000 and then non randomly (haphazardly) assign them to treatment or control.

And, it's possible to have neither random selection nor random assignment. In a typical non-equivalent groups design in education you might non randomly choose two 5th grade classes to be in your study. This is nonrandom selection. Then, you could arbitrarily assign one to get the new educational program and the other to be the control. This is non-random (or non-equivalent) assignment.

Random selection is related to sampling. Therefore it is most related to the external validity (or generalizability) of your results. After all, we would randomly sample so that our research participants better represent the larger group from which they're drawn. Random assignment is most related to design. In fact, when we randomly assign participants to treatments we have, by definition, an experimental design. Therefore, random assignment is most related to internal validity. After all, we randomly assign in order to help assure that our treatment groups are similar to each other (i.e., equivalent) prior to the treatment.

Q. 6. Calculate an estimate of median for the following data:

CLASS		FREQUENCY
0 – 24.9	–	6
25 – 49.9	–	11
50 – 74.9	–	14
75 – 99.9	–	16
100 – 124.9	–	13
125 – 149.9	–	10

Sol. L = 74.9
(the lower class limit of the 75–99.9)
h = 70
Cf_b = 6 + 11 + 14
= 31
f_m = 16
w = 6

$$\text{Median} = L + \frac{h/2 - Cf_b}{f_m} \times w$$

$$= 74.9 + \frac{35 - 31}{16} \times 6$$

$$= 74.9 + 1.5$$

$$= 76.4$$

SECTION–B

Q. 7. Explain any two of the following:

***(a) t* - distribution**

Ans. Ref.: See Chapter 4, Page No. 55 '*t*-distribution'.

***(b)* F - distribution**

Ans. Ref.: See Chapter 4, Page No. 56 'F-distribution'.

***(c)* CHI - SQUARE distribution**

Ans. Ref.: See Chapter 4, Page No. 55, 'Chi-square Distribution'

Q. 8. Using the Regression line $y = 90 + 50x$, fill up the values in the table below:

SAMPLE No. *(i)*	12	21	15	1	24
x_i	0.96	1.28	1.65	1.84	2.35
y_i	138	160	178	190	210
$\hat{y}_i$	138	–	–	–	–
$\hat{e}_i$	0	–	–	–	–

After filling the table, compute the parameters of Goodness to fit i.e. R and R^2. Based on the result of R and R^2, interpret the correlation between variable *x* and *y*.

Sol. Regression Line $\hat{Y} = 90 + 50\ X$

As we know that

$$R = \frac{S_{xy}}{\sqrt{S_{xx}\ S_{yy}}}$$

$$= \frac{138.076}{\sqrt{2.6929 \times 7839.93}}$$

$$= 09504$$

$$R^2 = .9033$$

Q. 9. What is forecasting? How forecasting is related to future planning, give suitable example in support of your answer? Briefly discuss any forecasting model.

Ans. Forecasting means predicting something that will happen in the future or estimating the probability that it will happen. Predicting weather conditions is a common application of forecasting. Meteorologists link their past experiences with weather conditions and scientific information about weather patterns to predict what the weather will be like in the future. Weather forecasts often are correct; and of course, they frequently are mistaken, either in terms of the intensity or timing of weather conditions. Weather forecasts are used to plan activities, such as when to go on a picnic, when to mow the lawn, or what clothing to wear on a particular day. Sometimes forecasts are wrong and the picnic gets rained out, the lawn gets too dry, or the outfit selected is too cool for the colder-than-expected temperatures.

People use forecasting and planning in many other aspects of their personal and professional lives:

- They forecast that their jobs are secure and their income will be sufficient to purchase a house.
- They estimate that they can complete a task at work within a given amount of time and, therefore, can also accept responsibility for an additional assignment.
- They think it is likely that a business meeting will be long, and they forego dinner plans.
- They predict that additional training and education will make them more marketable and, therefore, invest in education.

Forecasting for juvenile corrections is similar to these examples. It involves examining the past and present for quantitative information and trends and qualitative patterns. These are applied toward making predictions of numbers and tendencies based on experience and logical assumptions of what will happen in the future. Plans are then based on the probability that those forecasts will be reasonably accurate.

Forecasts often miss the mark of 100 per cent accuracy. However, they provide the best basis for planning available. Considering many variables is vital when developing forecasts. The social, economic, and political contexts always must be considered when forecasts are used. For example, in a largely industrial setting, one must consider the possible effects if a major manufacturer were to close operations. Employment and economic conditions in a community or state could

change drastically. Many prognosticators believe changes in welfare will increase the Nation's population of poor children, at least temporarily. Political tides and public opinions often vacillate between liberal and conservative viewpoints, and these changes frequently prescribe different responses to delinquency. Some of these consequences are reasonably predictable, while others are unexpected and probably cannot be incorporated in jurisdictional or program forecasts and plans. Neither can the actions of particular individuals necessarily be predicted. For example, a judge who believes only youth who commit a second violent offense should spend some time in confinement, or a probation officer who takes youth back to court for any and all violations of conditions of probation will both affect related juvenile corrections programs.

Relation between Future Planning and Forecasting: Planning and forecasting are inextricably intertwined each other. Planning is concerned with future which is highly uncertain. Therefore planners have to make assumptions about the future events. In order to make correct an assumption prediction of future events is essential. Forecasting is the primary source of planning premises which serve as the foundation for building the super structure of planning. The information generated through forecasting service is an input to planning. Forecasting is an integral part of the planning processes to the extent that it provides the necessary basis for charting out the future course of action.

Forecasting is prerequisite to planning. Forecasting indicate the probable course of future events, plans decide how to prepare for these events. Without forecasting will be a futile mental exercise and the organization would be at the mercy of future events. For example, a business enterprise predicts competitive technological, social and political conditions likely to prevail over the next five years.

On the basis of these forecasts, it determines objectives; strategies and policies concerning sales grow with, product range, market coverage, competitive initiative, profitability, etc. Planning and forecasting both are concerned with future. However, there is some difference between the two and difference lies in the scope of the two processes. Planning is more comprehensive including many elements and sub-processes to arrive at decisions concerning what is to be done, how to be done, and when to be done. Forecasting involves estimates of future events and provides parameters to planning. Planning result in the commitment of resources whereas no such commitment is involved in forecasting.

A large number of people are involved in the planning processes though major decisions are taken at the top level. Forecasting is normally done at middle or lower level. It may be entrusted to staff specialties whose decisions do forecasting does not involve decision making but helps in decision making by providing clues about what is likely to happen future. Fourthly forecasting involves what the future is likely to be and is likely to behave. Planning, on the other hand indicates what the future is desired to be and how to make it a reality.

In fact, forecasting is the essence of planning because the future course of action is determined in the light of the forecast made.

Q. 10. Differentiate between the following *(any two)*:

***(a)* Linear systematic sampling and circular systematic sampling.**

Ans. Ref.: See Chapter 12, Page No. 149 'Linear Systematic Sampling' and 'Circular Systematic Sampling'

***(b)* Z-test and T-test:** A Z-test is used for testing the mean of a population versus a standard, or comparing the means of two populations, with large samples whether you know the population standard deviation or not. It is also used for testing the proportion of some characteristic versus a standard proportion, or comparing the proportions of two populations.

Examples: *(i)* Comparing the average engineering salaries of men versus women.

(ii) Comparing the fraction defectives from 2 production lines.

A T-test is used for testing the mean of one population against a standard or comparing the means of two populations if you do not know the populations' standard deviation and when you have a limited sample

($n < 30$). If you know the populations' standard deviation, you may use a Z-test.

Example: Measuring the average diameter of shafts from a certain machine when you have a small sample.

***(c)* Correlation and Regression:** Correlation and linear regression are the most commonly used techniques for investigating the relationship between two quantitative variables.

The goal of a correlation analysis is to see whether two measurement variables vary, and to quantify the strength of the relationship between the variables, whereas regression expresses the relationship in the form of an equation.

For example, in students taking a Math and English test, we could use correlation to determine whether students who are good at Math and to be good at English as well, and regression to determine whether the marks in English can be predicted for given marks in Math.

In regression analysis, we find out the nature of the relationship between the dependent variable (response) and the (explanatory) independent variable.

The analysis consists of choosing and fitting an appropriate model, done by the method of least squares, with a view to exploiting the relationship between the variables to estimate the expected response for a given value of the independent variable. For example, if we are interested in the effect of age on height, then by fitting a regression line, we can predict the height for a given age.

■■