Name : Roll number:

If you are caught cheating, even in one question, your final mid term score would be -50 and reporting the matter to academic section for further action. Please refrain from indulging in unfair means.

Unless specified, symbols have their usual meanings

Don't write any rough work in this sheet. Write only the final answers.

Write your name and roll number on top of each sheet of the paper.
Total marks :

Q. 1 a.) – (5 marks)
There is a coin which has the probability of heads as 0.7 and probability of tails as 0.3.

Let an experiment involve tossing a coin 100 times.

Denote by "X" a random variable which counts the number of heads in these 100 tosses.

Denote by "Y", a random variable which counts the number of tails in these 100 tosses.

Let Z = X+2Y

Find E(Z)

$E(Z) = E(X) + 2E(Y) = np + 2nq$

$= 100 \times 0.7 + 2 \times 100 \times 0.3$

$= 70 + 60$

$= 130$

130

Q.1 b.)– (10 marks)
Find the variance of the random variable Z.

21

$Z = X + Y + Y = 100 + Y$

$E(Z) = 100 + E(Y) = 130$

$Var(Z) = Var(Y) = 100 \times 0.7 \times 0.3 = 21$

Name : Roll number:

Q.2 (5 marks) – (Single option correct)
Among mean and median, median is always a better way to summarize the data.

a.) Always True
b.) Always False
c.) Depends on the dataset and the context ✓

Q.3 a.) (10 marks)
The Dean of Student Welfare (DOSW) receives a lot of emails on a daily basis. On any day, the average number of emails received by DOSW is 200. It can be assumed to follow a Poisson distribution.

Answering an email takes some time. Some emails can be responded quickly , while some emails take longer. Assume that time required to respond to a email follows a uniform distribution with minimum possible value as 2 minute and maximum possible value as 5 minutes.

Find the expected time spent by DOSW in answering to emails.

$a = 2$
$b = 5$
$E(X) = \frac{a+b}{2} = 3.5$
$E(X) = 200 \times 3.5 = 700$

700

Q.3b.) (15 marks)
Let T denote the random variable which represents the time required by DOSW to respond to emails.

Find the VARIANCE of random variable T.

2600

$N \sim Pois(200)$ $\quad E(N) = 200, \ Var(N) = 200$

$X \sim Unif(2, 5)$. $\quad E(X) = 3.5, \ Var(X) = \frac{9}{12}$

Variance $= E(N) \cdot Var(X) + (E(X))^2 \cdot Var(N)$

Name : Roll number:

Q.4 (5 marks)
A data of heights of 1000 randomly selected individuals in Roorkee is collected, the range of the observations are from 145 cm to 187cm. From this dataset, to estimate the average height of an individual in Roorkee, a 90% Confidence Interval, 95% confidence interval and 99% confidence interval is built.
Which of the three would have the highest width?

a.) 90% CI
b.) 95% CI
c.) 99% CI
d.) All have same width

Q.5 (5+5+5+5+5 marks)
Let X be a random variable that follows a normal distribution with mean = 100 and variance = 36.

a.) Find P(X = 100) 0

b.) Find P(X<100) 0.5

c.) Find P(X<105) 0.7976

d.) Find P(X>110) 0.04779

e.) Find P(95<X<105) 0.5953

Q.6 a.) [15 marks]
Amul wants to launch "Kaju Katli" as a new product to give competition to local sweets vendor. The objective of Amul is to capture atleast 30% of the market.

In a pilot study, Amul surveyed 1000 customers who were regular buyers of Kaju Katli from local sweets vendor, out of which 320 said that they would switch to Amul Kaju Katli. From this data, what is the symmetric confidence interval for the true proportion of customers who would switch to Amul.

0.3498, 0.2911

$\hat{p} = \frac{320}{1000} = 0.32$

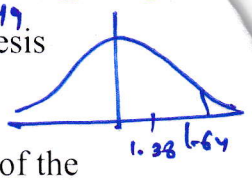
$H_0: p < 0.3$

$H_1: p \geq 0.3$

Name : $\sigma_{\bar{p}} = \sqrt{\frac{0.3 \times 0.7}{1000}} = 0.01449$ Roll number:

Q.6b.) [15 marks]
Using the data, Roger Federer, CEO of Amul would like to test the hypothesis that Amul's Kaju Katli would be able to capture 30 % of the market.

$\bar{p} = \frac{320}{1000} = 0.32 \Rightarrow z = \frac{0.32 - 0.3}{0.01449} = 1.38$



Taking Amul's objective as the alternative hypothesis, what is the p-value of the hypothesis test.

$= 0.9162$

$= 0.0838$

0.0838

Q.6c.) [10 marks]
By looking at the p-value in part b.) of the problem, should the alternative hypothesis be accepted? Answer briefly in less than 20 words.
[Assume level of significance as 0.05]

$0.0838 > 0.05$ (Reject H_1)

Q.7 [8+8 marks]
Consider a scenario where it is thought that a variable Y is dependent on two variables X1 and X2.

Usually, the hypothesised linear regression model is -

E(Y) = alpha + beta_1*X1 + beta_2*X2

Mr. Ajit Mohan from IIT Kanpur believes that life is not constant and wants to propose a new relationship between Y, X1 and X2 without the constant term.

The hypothesised relationship as per Mr. Mohan is

E(Y) = beta_1*X1 + beta_2*X2

After building this model, Mr. Mohan collects the data for 5 observations and wants to give the best estimate of beta_1 and beta_2 using the same methodology as is done in the usual linear regression model.

Y	X1	X2
100	50	30
150	60	50
180	150	15
200	80	45
240	100	65

minimize: $\sum_{i=1}^{5} (y_i - \beta_1 X_{1i} - \beta_2 X_{2i})^2$

Take partial derivative w.r.t β_1 & β_2. Solve 2 eqn's.

Name : Roll number:

What is the best estimate for beta_1 and beta_2 given by Mr. Mohan?

beta_1 : ~~1.015~~ 1.015
beta_2 : ~~2.095~~ 2.095

Q.8 [15 marks]

Use CLT, $T = X_1 + X_2 + \dots + X_{100000}$
$T \sim N(80000, 80000 \times 0.8 \times 0.2)$

A coin having probability of heads as 0.8 is tossed 100000 times. What is the approximate probability that observed number of heads is less than 80200?

0.943664 (or 0.96145) — using CLT

Q.9 [10+5+10+10 marks]

In the file shared with you, build a multiple linear regression model in which "CPI" is the dependent variable and "Unemployment%" and "Literacy Rate%" are the independent variables. Ignore the other two columns of GDP and Per capita income.
Run the multiple linear regression analysis using Python/Excel and answer the following questions.

a.) What is the multiple regression model ?

$4.9148 - 0.1508\,U + 0.4886\,LR$

b.) By looking at the p-value of the F-test, should the proposed linear relationship between these variables considered or not?

$6.19 \times 10^{-8} < 0.05$ (Reject H_0) There is lin rel^n

c.) Give the interpretation of coefficient of "literacy rate%" in less than two lines.

For a unit increase in literacy rate, CPI increases by 0.4886 on an average, keeping

Name :

Roll number:

d.) What is the p-value for the two independent variables considered? Can we be confident in giving the interpretation of the coefficient of the unemployment(%)?? Explain briefly.

Unemployement → $p = 0.144 > 0.05$ (Accept H_0)

There is no linear rel^n.

Literacy rate → $p = 4.5 \times 10^{-6}$

No, because p-value > 0.05

Name : Roll number:

Q.10 [10+5+15 marks]

A sincere IIT-Roorkee student named Rancho is concerned with the time spent in gossiping per day. Rancho has 40 close friends. Rancho goes and tries to meet all of them separately, however not all of them are available daily. He estimates that there is a 20% chance that the close friend he tries to meet is available. The time spent in a single gossip session is 20 minutes on an average. Making suitable assumptions, calculate -

a.) Expected number of close friends that Rancho meets every day

Binomial RV) $= 40 \times 0.2 = 8$

b.) Expected time spent in gossiping per day (in minutes)

$8 \times 20 = 160$ minutes

c.) After this calculation, Rancho is worried about he is spending too much time in gossiping. He makes a new rule for himself (apart from the ones given in IIT-R rulebook) that he will meet no more than 5 of his friends every day for gossiping. Thus, if 5 or less friends are available, he will meet and gossip with all of them. If the number of available friends is greater than 5, he would choose any 5 from them and meet only those 5. Making suitable assumptions calculate -

c.) Expected number of close friends that Rancho meets every day with this new rule. (Write correct answer upto 3 decimal places)

4.886

$0 \cdot p_0 + 1 \cdot p_1 + \ldots + 5 \cdot p_5$

AF	Prob
0	p_0
1	p_1
2	p_2
3	p_3
4	p_4
≥ 5	p_5

$= 1 - (p_0 + p_1 + \ldots + p_4)$