PUSASQF202
Probability and Statistics-2

Time: 2 hours
Total Marks: 60 marks

Note:

1) **The candidate has option to either attempt question 4A or question 4B. Rest all questions are mandatory.**
2) **Numbers to the right indicate full marks.**
3) **The candidates will be provided with the formula sheet and graph papers (if required) for the examination.**
4) **Use of approved scientific calculator is allowed.**

Q1: Attempt All Questions

1A) **5 Marks**

If the distribution of number of heads after n independent coin tosses can be approximated normally as N(60, 36),

(i) Find n and the probability of heads (p).

(ii) Find the Normal Approximation to the Distribution of Number of Tails after n coin tosses.

(iii) How does the answer to part (ii) change if I were to find the distribution of number of tails after 3n coin tosses?

1B) **5 Marks**

Students in a school have the option to take external aptitude tests organised by University of New South Wales. There are 50 students in the class. The distribution of marks out of 100 in each subject for any individual student can be assumed to be independent and normal but with different parameters as follows:

Subject	μ	σ
Environmental sciences (ES)	85	10
World History (WH)	76	10

Students are 25% more likely to opt for environmental science as compared to World history and the participation is in line with this ratio. A total of 45 students opted to sit for the two exams in this ratio.

Find the probability that the average of ES marks is greater than that of WH and comment on the result

1C) **5 Marks**

We have selected n random samples from exponential distribution with parameter λ. The probability distribution function (PDF) is defined as

$$f(x) = \lambda e^{-\lambda x} , 0 < x < \infty$$

You have been asked to

i. Determine the Maximum Likelihood Estimator (MLE) of λ.
ii. Determine p^, the MLE of p, stating the reason for why it is the MLE based on the below information. Comment on your answer.

Motor claims sizes are modelled using an exponential distribution with parameter λ. A random sample of such claims results in the value of the MLE of λ as λ^ = 0.00000314

A large claim is defined as one greater than INR 100,000 and the claims manager is particularly interested in p, the probability that a claim is a large claim.

Q2: Attempt All Questions

2A) **5 Marks**

3 students are waiting to work on their first project and are discussing their approach before the samples arrive. Student 1 wants to extract 40 samples independently from a large portfolio of claims and calculate a 95% confidence interval for the sample mean. Student 2 doesn't agree and wants to independently extract 100 samples and then calculate a 95% confidence interval for the sample mean. Student 3 understood the importance of data validation checks and wants to extract 40 samples independently from the same portfolio but replace outliers and/or any erroneous samples that may skew his/her confidence interval. He/ She then proceeds to calculate a 99% to compensate for the smaller sample size.

Compare in detail without carrying out any calculation how each of the confidence interval will look like include Student 3's interval before and after replacement of outlier.

2B) **5 Marks**

The exponential family is the set of distributions whose probability function, or probability density function (PDF) can be written in the following form:

$f_Y(y\,;\theta,\varphi) = \exp\left[\frac{(y\theta - b(\theta))}{a(\varphi)} + c(y,\varphi)\right]$

Where $a(\varphi)$, $b(\theta)$ and $c(y,\varphi)$ are specific functions.

You are required to rewrite the pdf of the Binomial distribution in the form above, explicitly stating what the 3 specific functions correspond to, the natural parameter and to arrive at the mean and variance of the Binomial distribution using the specific functions stated above.

2C) **5 Marks**

State the definition of a t_k distribution.

Hence, using $\bar{X} \sim N(\mu, \sigma^2/n)$ and $(n-1)S^2/\sigma^2 \sim \chi^2_{n-1}$, Show that:

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

Define the Cauchy distribution and state why it is peculiar.

Q3. Attempt All Questions

3A) **5 Marks**

16 people volunteered to test the new Covid 19 vaccines. 4 of them were given placebo and the others 3 variants of the vaccine. Their reactions to the vaccines measured by antibodies produced were measured and the results shown below:

	Antibodies (y)				**Total**
Baseline	100	100	100	100	400
Variant A	102	108	111	104	425
Variant B	105	107	109	110	431
Variant C	120	118	114	106	458

$\Sigma y = 1714$; $\Sigma y^2 = 184216$

Conduct an analysis of variance test to establish whether the data indicate significant differences amongst the results for the variants at the 1% level of Significance

3B) **5 Marks**

An Insurer has collected data on average alcohol consumption (units of 30 ml per week) and average smoking (no. of Cigarettes per week) in 8 metros across India where they sell Health Insurance in.

Metro City in India	1	2	3	4	5	6	7	8	**Average**
Alcohol consumption (xi)	15	20	12	18	14	16	19	20	**16.75**
Cigarette consumption (yi)	4	8	6	7	2	5	4	9	**5.625**

Calculate the coefficient of correlation 'r' (Pearson's correlation coefficient) between alcohol consumption and cigarette smoking

3C) **5 Marks**

A health insurer believes that the number of claims in a policy year follows a binomial distribution with p = 0.3 and n = 4. He has asked the actuarial department to extract a sample of 200 policies and test his hypothesis using goodness of fit test for this binomial model. The sample reveal the below results

No. of claims	0	1	2	3	4
No. of policies	75	50	60	12	3

Q4. Attempt Q4A or Q4B

4A) **15 Marks**

A sports agency which manages cricket players is eager to understand the relationship between the number of hours spent in extracurricular sports activity in a month (excluding cricket but including weightlifting, football yoga etc.), and the average number of runs scored by opening batsmen of 8 franchise teams after being given a choice to regulate their extracurricular hours themselves. The following data were obtained basis the inputs from players and their performance on field. You may assume all batsmen played equal number of matches all at the same venue using the same pitch i.e. all conditions were same.

No. of hours spent in extracurricular activities (x)	10	15	20	25	30	35	40
Average Runs scored (y)	35	40	42	45	46	48	44

i. Calculate the equation of the least-squares fitted regression line [5]

ii. Calculate a 90% confidence interval for the slope of the underlying regression line. [4]

iii. Using the confidence interval above, test the hypothesis that the slope of the underlying regression line is equal to 1. [2]

iv. Use the fitted line obtained in part (i) to calculate estimates of the runs scored by the opening batsmen when they spend 60 hours in a month on extracurricular activities vs when they spend 37 hours . [2]

v. Comment briefly on the reliability of your results. [2]

Q4B) **15 Marks**

A health claims branch can process 100 claims and has 4 counters. The time taken to process a claim has a mean of 23 mins and a standard deviation of 6 mins. You have been provided a sample of 12 processed claims. The fastest time taken to process a claim is 17 mins

i. State the assumptions you are will be required to make any statistical calculations. [1]

ii. Without any calculation estimate the probability that the sample mean for these 12 sample claims is less than 17 mins (record for fastest claim processing at this branch) [2]

iii. Now for the same situation above, carry out the calculation for this probability. [3]

iv. Compare your estimate with the actual probability by commenting on the results. [2]

v. Calculate the probability that the sample variance to process 12 claims is greater than 8.5 mins (use interpolation if required) [3]

vi. State the definition of a $F_{m,n}$ distribution. [2]

vii. Hence using $(n-1)\, S^2 / \sigma^2 \sim \chi^2_{n-1}$ show that for suitably defined samples; [3]

$$\frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \sim F_{\text{m-1, n-1}}$$