PUSASQF202
Probability &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

A) **5Marks**

If the distribution of number of heads after n independent coin tosses can be approximated normally as N(40, 24),

(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 **2n** coin tosses.

B) **5Marks**

Starting from the PDF of a Poisson distribution and comparing with the form of the exponential family $f(y|\theta,\varphi) = \exp\{(y\theta - b(\theta))/a(\varphi) + c(y,\varphi)\}$, specify the various parameters / functions, i.e. θ, φ, a(), b(), c(). Also suggest an appropriate link functionwith reasons.

C) **5Marks**

The distribution of marks in each subject for any individual student at the AUQAT (Absolutely Unique Quantitative Aptitude Test) can be assumed to be independent and normal but with different parameters as follows:

Subject	μ	σ
English	50	5
Maths	55	10

Find the probability that the average of English marks of a class of 25 students is greater than that of Mathematics for the same class.

Q2. Attempt All Questions

A) **5Marks**

For a quantity following the normal distribution centered at zero, n underlying samples are randomly chosen.

(i) Find an expression for the probability that the sample mean exceeds k times the sample standard deviation, for a positive constant k.

(ii) How would this probability change as n approaches infinity, for a fixed k?

(iii) How would this probability change with k, for a fixed n?

B) **5Marks**

It is considered that the lifetime of a battery follows an exponential distribution with parameter λ. As a Quality Assurance statistician, you are charged with estimating this parameter. However, you only have a maximum available time window of 100 hours beyond which you terminate the test. You start with 10 batteries and record the stopping time for each battery. You observe that eight of the batteries had the following stopping times in hours {7.10, 18.19, 16.12, 18.45, 63.19, 51.96, 3.59, 85.95} while the remaining two were still active after 100 hours. Based on these observations, find the MLE of λ.

C) **5Marks**

It is known that the height (in cm) is distributed normally among the residents of Quantica, with a standard deviation of 15.

(i) Find the minimum sample size required to obtain a 95% confidence interval (CI) with width ±3.

(ii) How many more sample points would be required if we needed to narrow the CI width to ±1?

(iii) Comment on the answers obtained in parts (i) and (ii)

Q3. Attempt All Questions.

A) **5Marks**

The average weight for a sample of 25 randomly chosen students at a university is 70 kg. For that sample, the standard deviation of weights is 16 kg.

(i) Find the 90% confidence interval for the average population weight, assuming the weight to be normally distributed.

(ii) How would answer to part (i) differ if 16 kg was known to be the population standard deviation?

(iii) Comment on the differences in the answers obtained and also on how answers in parts (i) and (ii) would change if the sample size were to tend towards infinity.

B) **5Marks**

The demographics of members of an actuarial society are as follows:

Sex \ Age	Less than 30y	Between 30-45y	Between 45-60y	Above 60y
Male	27	45	52	38
Female	34	48	43	29

Test the hypothesis that the two classification criteria are independent.

C) **5Marks**

If the Spearman rank correlation derived from n data points is 20%, what is the least n for which one would be able to reject the null hypothesis that the underlying correlation rho = 0 against the alternative that rho > 0at 5% level of significance? How would your answer change if it was the Kendall rank correlation (tau) instead?

Spearman: For large n, r~N(0, 1/(n-1)) under H0: $\rho_S r = 0$. *Where* ρ_S *is the Population Spearman's Rank Correlation*

Kendall: For large n, r~N(0, 2(2n+5)/9n(n-1)) under H0: $\tau r = 0$.

Q4. Attempt Q4 A or Q4 B.

A) **15Marks**

You are provided the following summary of a dataset with 42 pairs where X = fire per 1000 houses and Y = theft per 1000 population in the different zip codes of Chicago: $\Sigma y = 1414$, $\Sigma x^2 = 10598.59$, $\Sigma y^2 = 69370$, $\Sigma xy = 22980.90$.

- **i.** Fit a linear regression model.
- **ii.** Test the hypothesis that the slope coefficient of the linear model $\beta = 0$ (i.e. there is no linear relationship).
- **iii.** Find a 95% confidence interval of an individual response y0, corresponding to x0 = 10.
- **iv.** Plotting the data reveals that there is one outlier point (39.7, 147). Find the impact on our linear regression model (fitted parameters) of removing this one point from the data.

B)

You are a sports statistician associated with Quantican Football Association and you are charged with using GLM to model the number of goals expected to be scored by and against each team for a given match in their league. The three teams are called Gaussian Giants, Newtonian Ninjas and Archimedean Aces. Each team (say X) plays every other team (say Y) in two matches – once at their own home ground (home for X; away for Y) and once at the other team's home ground (away for X; home for Y) Here is the score line (number of goals scored by the respective team) in each of their matches:

Match No	Home Team	Away Team	Home team goal	Away team goal
1	GG	NN	2	3
2	GG	AA	2	0
3	NN	AA	1	1
4	NN	GG	1	0
5	AA	GG	3	2
6	AA	NN	1	2

A Poisson distribution is being considered to model the number of goals a team X would score against another team Y in a particular match.

- **i.** Your baseline model is to model the number of goals scored by a team X as only dependent on the team X itself and independent of any other factors (including opposition team Y or home/away). State (with brief reasons) the MLE estimates for the mean μ_X each team X.

Then, you wish to fit an improved model which also accounts for the opposition team.

ii. How would you design the linear predictor? How many parameters would this model need? Describe how the MLE estimates for each of these parameters would be derived.

iii. Describe in detail how you would decide if the improved model is significantly better than the baseline?Why is this important?

You finally want to further improve the model by accounting for a factor which captures whether a team is more likely to score goals at home or away.

iv. How would you further modify the linear predictor? How many additional parameters would this model need?

v. A commentator has suggested that there is no such thing as a home advantage (i.e. a team is not expected to play better or score more goals at home than away). Describe in detail how you'll test this hypothesis within your model.