Indian Statistical Institute, Bangalore
B. Math I, Second Semester, 2022-23
Back Paper Examination,
Intro to Statistics and Computation with Data
Maximum Score 100 Duration: 3 Hours

Values from the normal distribution qnorm(0.9)=1.281552, qnorm(0.995)=2.575829, qnorm(0.95)=1.644854

1. (3+9+6) The length of stay in a hospital after a particular treatment is of interest to patients and hospitals. A random sample of 50 patients who received the treatment was selected. The length of stay, in number of days, is recorded in the table below.

Length of stay	5	6	7	8	9	12	21
Number of patients	4	13	14	11	6	1	1

   (a) Is the distribution symmetric, skewed to the right or skewed to the left?

   (b) Find the median, first and third quartiles.

   (c) Using the 1.5×IQR rule, identify outliers, if any.

2. (2) Researchers are interested in studying whether regular walking reduces cholesterol. A sample of 100 patients with high cholesterol is selected. The patients are advised by the doctor to walk regularly. They are asked to keep a daily record of the number of minutes walked each day. After six months, their cholesterol is tested again and the walking record obtained. Identify and explain the type(s) of bias that can affect this study.

3. (10+2+3) Let $X_1, X_2 \cdots X_n$ be a random sample from the Pareto distribution with pdf given by

$$f(x \mid \theta) = c\theta^c x^{-(c+1)}, \quad \text{if} \quad x > \theta, \tag{1}$$

   where $\theta \in \mathbb{R}^+$ and $c > 2$. Both θ and c are unknown.

   (a) Obtain the method of moments (MoM) estimator for θ.

   (b) Is this the only MoM estimator, or is it possible to obtain other estimators based on the method of moments?

   (c) Is the estimator consistent? Justify your answer.

4. (5+2+3+5)From a bivariate dataset $(x_1, y_1), \cdots, (x_n, y_n)$, we obtain two least squares regression lines 2y=3x+5 and 3y=5x+7. One line is for the regression of Y on X and the other is for the regression of X on Y.

   (a) What are the sample means of X and Y?

   (b) Is the correlation coefficient between X and Y positive or negative? Why?

   (c) What is the value of the correlation coefficient between X and Y?

   (d) Which one of the two lines is the regression line of Y on X? Explain.

5. (5+5) Let $U \sim Unif(0,1)$ and $W = \tan(\pi(U - 1/2)$.

   (a) Find the distribution of W.

   (b) Use this to describe how to generate a random variable from the Cauchy distribution with parameters 10 and 5.

6. (4+2+4) The iris data consists of 4 characters (sepal length, sepal width, petal length, petal width) measured on 50 flowers from each of 3 species (setosa, versicolor, virginica). We run the following command in R.

```
summary(aov(formula = Sepal.Width ~ Species, data = iris))
```

   (a) Complete the table of output.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Species		11.35			<2e-16
Residuals			0.115		

   (b) Explain what the above code is doing.

   (c) Carry out the ANOVA test using the above output stating the null and alternative hypotheses, assumptions and conclusions.

7. (4+4+4+4+4) A medical centre developed a new procedure for surgery of the knee, designed to reduce the recovery time after surgery. To test the effectiveness of the new procedure, a study was conducted in which 210 patients needing knee surgery were randomly assigned to receive either the standard procedure or the new procedure. Summary statistics of the recovery times are given below.

Type of Procedure	Sample Size	Mean recovery Time(days)	Standard deviation of Recovery Time
Standard	110	217	34
New	100	186	29

   Using $\alpha = 0.05$ perform the appropriate test of hypothesis to determine if the mean recovery time is lower for the new procedure. In particular, answer the following questions.

   (a) State the null and alternative hypotheses.

   (b) State the test statistic and find its distribution under the null hypothesis.

   (c) Is the distribution in the previous part exact or approximate? In the latter case, why does the approximation work?

   (d) Find the critical region and compute the value of the test statistic.

   (e) Is the null hypothesis rejected? What is the conclusion regarding the mean recovery time?

8. (10) Suppose you wish to conduct a test of the research hypothesis that the median of a population is greater than 80. You randomly sample 25 measurements from the population and determine that 16 of them exceed 80. Set up and conduct the appropriate test of hypothesis at the 10% level of significance. Be sure to specify all necessary assumptions.