IIITS/M-2018/Mid Sem Exams

Date: Sep 2018

Indian Institute of Information Technology, Sri City, Chittoor

Name of the Exam: Statistics for Data Science

Duration: 1.5 hrs

Max. Marks: 25

Roll No.

Room No.

Seat No.

Name:

Invigilator's Signature

Instructions:

1. Scientific calculators are allowed
2. The exam is **not** open book and student(s) are **not** allowed to bring Text book(s)/ Photocopies / Hand-written notes / laptops.
3. Follow the instructions mentioned in the questions. **Answer any four questions from 1 to 5. Question 6 and 7 are mandatory.** $4 \times 4 + 4 + 5 = 25$.
4. Marks are indicated in [ ]

---

1. a. State univariate Central Limit Theorem. [2]

   b. Classify each of the following as N, nominal; O, ordinal; I, interval; R, ratio data: [2]

   I. letter grade you will receive in this subject

   II. amount of money you have with you

   III. year of your birth (A.D.)

   IV. your IQ (intelligence quotient)

2. a. What is the difference between a parameter and a statistic? State with examples. [2]

   b. What kind of sampling process is performed for each of the following? [2]

   I. To study cancer a researcher divides the population under study into different age groups and takes samples at random from each group in such a manner that all age groups are represented proportionally.

   II. To conduct a political poll a researcher uses an alphabetical list of registered voters and contacts every 100th name starting from the top.

3. Let X_1 and X_2 are two independent random samples taken from a population with mean μ and variance σ^2. Suppose that you have two estimators of μ: [4]

$$\Theta_1 = \frac{X_1 + X_2}{2}$$

$$\Theta_2 = \frac{X_1 + 3X_2}{4}$$

Which estimator between Θ_1 and Θ_2 will you choose and why?

4. a. The scatter plot of a two dimensional data (variables X and Y) exhibit elliptical pattern where the major and minor axis are parallel to X and Y axes respectively. Using this information comment on the characteristics of the concerned variables. [2]
b. Define statistical distance from a multivariate normally distributed random variable X from a population mean vector μ, where the population covariance matrix is Σ. Comment on its distributional property. [2]

5. a. Define Type I error and Type II error. [2]
b. A researcher wishes to test the following hypotheses with α=0.05: [2]

$$H_0 : \mu = 120$$
$$H_1 : \mu > 120$$

He carefully takes a sample of size n = 40 and obtains:

$$sample\ mean, \bar{x} = 128.12$$
$$sample\ standard\ deviation, s = 32.17$$

What is the appropriate conclusion? [information provided: $P(Z > 1.645) = 0.05$ & $P(T_{39} > 1.685) = 0.05$; Z is a standard normal variable and T_ν is a t-distributed random variable with ν degrees of freedom]

6. Suppose that X is a discrete random variable with the following probability mass function: [4]

X	0	1	2	3
P(X)	$\frac{2\theta}{3}$	$\frac{\theta}{3}$	$\frac{2(1-\theta)}{3}$	$\frac{(1-\theta)}{3}$

where $0 \leq \theta \leq 1$ is the parameter. The following 10 independent observations were taken from such a distribution: [3, 0, 2, 1, 3, 2, 1, 0, 2, 1]. Find the maximum likelihood estimate of θ

7. The feeding habits of two species of net-casting spiders are studied. The species, the deinopis and menneus, coexist in eastern Australia. The following data were obtained on the size, in millimeters, of the prey of random samples of the two species: [5]

Deinopis	12.9	10.2	7.4	7.0	10.5	11.9	7.1	9.9	14.4	11.3
Menneus	10.2	6.9	10.9	11.0	10.1	5.3	7.5	10.3	9.2	8.8

The objective of the whole research study is to find out whether there is any difference in the mean size of the prey (of the entire populations) of the two species.
In order to do so, develop a 95% confidence interval for the difference in the mean size of the prey (of the entire populations) of the two species.

Assumptions of the study: a. The measurements are independent.
b. The measurements in each population are normally distributed.
c. The measurements in each population have the same variance σ^2.

[Information provided: $P(Z > 1.96) = 0.025$ & $P(T_{18} > 2.101) = 0.025$; Z is a standard normal variable and T_ν is a t-distributed random variable with ν degrees of freedom]