INDIAN STATISTICAL INSTITUTE

POST-GRADUATE DIPLOMA IN BUSINESS ANALYTICS (PGDBA): 2024-26

Course: INFERENCE

Assignment 1

Deadline: Sunday, November 24, 2024 (11.59 p.m.)

- *Use R to solve the problems.*
- *All relevant R programming code should be submitted for evaluation, together with a properly-formulated report* **(as a pdf document),** *containing intermediate and final outputs (including plots, if any),* ***with explanation wherever required, and comments if asked for.***
- *The data used should also be submitted as a .txt or .csv file.*
- *Submission should be made only via the Google Classroom for the course.*
- *Extra credit may be given for outstanding individual effort.*
- *There will be a penalty for copying if detected. The decision of the instructor will be final.*

PART I

From resources available on the internet (e.g., the UCI Machine Learning Repository), identify a dataset consisting of approximately 500 (raw) measurements (taken on the same set of sampling units) on two discrete-valued and two continuous-valued variables. This dataset may be used in future assignments too.

1. Write a short paragraph describing the dataset, including the significance of the individual variables. Do not forget to mention the source of the data, providing appropriate links.

2. For one of the discrete variables (say, X) as well as one of the continuous variables (say, Y), carry out the following exercises:
   a. Provide the frequency distribution in a tabular form.
   b. Plot the histogram, clearly mentioning what (in-built) method you have used for determining the bin-width.
   c. Generate the box-and-whisker plots for data on the two variables.
   d. Comment on the properties of the distribution of the two variables on the basis of your observations from the outputs in (b)-(d).

3. Consider the following 7 datasets:

   a. Dataset corresponding to the variable X in problem no. 2.
   b. Dataset corresponding to the variable Y in problem no. 2.

c. Datasets of size 500 simulated from the following probability distributions:

i. A Poisson distribution with mean $\lambda = 5$.
ii. A continuous uniform distribution over the interval [3,8].
iii. An exponential distribution with pdf $f(x) = 7e\exp[-7x]$, $x > 0$.
iv. A Beta distribution with parameters $\alpha = 2$ and $\beta = 3$.
v. A log-normal distribution with parameters $\mu = 0$ and $\sigma = 1$.

For each dataset,

- Draw 100 samples independently, for three different values of the sample size, namely, $n = 10, 50, 100$;
- compute the means of the 100 samples for each value of n;
- plot the histograms of the original dataset as well as and the histograms of the means in the form of a 7×4 array in which the rows correspond to the 7 datasets listed above and columns 2-4 correspond to the three values of n while the first column contains the histograms of the original dataset.

Based on these plots, comment on the behaviour of the distribution of sample means in each case.

4. One way to understand missing data mechanisms is to generate hypothetical complete data and then create missing values by specific mechanisms, so that the deleted values may be retained and used to compare methods.

a. Generate 500 independent observations each on $X \sim N(0,1)$ and $Y \sim N(6,4)$ to create a bivariate dataset.
b. Using appropriate functions from an R package like *missMethods,*

i. delete 20% of the observations on X randomly by each of the three missing-data mechanisms MCAR, MAR and NMAR; provide a plot of the (X, Y) observations using a different colour for data points with missing X-values;
ii. impute missing values in the resulting bivariate dataset by implementing any two methods discussed in class;
iii. evaluate the performance of the imputation methods used in (ii) using the RMSE criterion.
iv. Repeat steps (i)-(iii) after randomly deleting 30% of the observations on X in step (i).

c. Provide a short description of the functions used by you and present all the results in a suitable tabular form.
d. Comment on the results.

PART II

Denote the four datasets selected by you in PART I by A, B, C and D respectively, where A and B contain observations on discrete variables, and the other two contain continuous-valued data.

5. Consider the dataset labeled A.
   a. Compute the median of the data. Label it as a.
   b. Select a simple random sample of size 100 from the dataset.
   c. Let Z be a random variable representing the number of observations in the sample that are less than a. Let z denote the observed value of Z.
   d. Let $p = P[X \leq a]$. At the 5% level of significance, test the hypothesis that p is larger than 0.5.
   e. Provide an approximate 95% confidence interval for p.

6. Fit a normal distribution to the dataset B and conduct a statistical test to assess whether the fit is good.

7. Divide the dataset B randomly into two subsets whose sizes are in the approximate ratio 3:2. Label these subsets as B1 and B2.

   a. At the 1% level of significance, based on the inference made in problem no. 6, conduct an appropriate exact/ approximate test of equality of the population variances corresponding to the two subsets.

   b. Based on the inference made in problem no. 6 AND part (a) of this problem, conduct an appropriate exact/ approximate test of size 0.01 for the equality of the population means corresponding to the two subsets B1 and B2.

   c. Provide an exact/approximate 99% confidence interval for the difference of the means.