Question Paper

Exam Date & Time: 05-Jan-2024 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

DATA SCIENCE - PART II [CRA 4061]

Marks: 50 **Duration: 180 mins.**

Descriptive

Answer all the questions.

1A) For the salaries data set given below, write suitable R Code to address the following (5)

Table 1: First twelve rows of Salaries

rank	discipline	yrs.since.phd	yrs.service	sex	salary
Prof	B	19	18	Male	139750
Prof	B	20	16	Male	173200
AsstProf	B	4	3	Male	79750
Prof	B	45	39	Male	115000
Prof	B	40	41	Male	141500
AssocProf	B	6	6	Male	97000
Prof	B	30	23	Male	175000
Prof	B	45	45	Male	147765
Prof	B	21	20	Male	119250
Prof	B	18	18	Female	129000
AssocProf	B	12	8	Male	119800
AsstProf	B	7	2	Male	79800

a. Visualize the difference in salary between male and female instructors

b. Fit the model with salary as the dependent variable and sex as the independent and calculate the regression coefficients

c. Examine the relationship between Sex, Salary, and Years of Service and compute the equation of the fitted regression line for the relationship

d. Compute the residuals and regression to the mean for the above data set.

e. Visualize the relationship between fitted values and residuals.

1B) Write suitable R code to perform the following operations for the data str given below. (3)

'data.frame': 192 obs. of 10 variables:

$ Year : Time-Series from 1969 to 1985: 1969 1969 1969 1969 1969 ...

$ Month : Factor w/ 12 levels "Jan","Feb","Mar",..: 1 2 3 4 5 6 7 8 9 10 ...

$ DriversKilled: num 107 97 102 87 119 106 110 106 107 134 ...

$ drivers : num 1687 1508 1507 1385 1632 ...

$ front : num 867 825 806 814 991 ...

$ rear : num 269 265 319 407 454 427 522 536 405 437 ...

$ kms : num 9059 7685 9963 10955 11823 ...

$ PetrolPrice : num 0.103 0.102 0.102 0.101 0.101 ...

$ VanKilled : num 12 6 12 8 10 13 11 6 10 16 ...

$ law : num 0 0 0 0 0 0 0 0 0 0 ...

a. Load the dataset Seatbelts. Fit a linear model of driver deaths with # kms and PetrolPrice as predictors. Add the variable law as a predictor and interpret the results.

b. Predict the number of driver deaths at the average kms and petrol levels.

1C) What are the relationships/differences between Bias, Variance, and Residuals? (2)

2A) Write a code in R to generate the data set given below. Assess the relationship between tumor size and cancer type and model it graphically. Justify why might linear regression not be the best fit for modeling this relationship. (5)

Patient ID	Age	Tumor Size (cm)	Cancer Type (Dependent Variable)
1	45	2.3	Benign
2	60	3.1	Malignant
3	38	1.8	Benign
4	55	2.9	Malignant
5	50	2.5	Benign
6	42	2.0	Benign
7	48	3.5	Malignant
8	52	2.2	Benign
9	58	3.2	Malignant
10	47	2.4	Benign

2B) You ask a collection of husbands and wives to guess how many jellybeans are in a jar. The correlation is 0.2. The standard deviation for the husbands is 10 beans while the standard deviation for wives is 8 beans. Assume that the data were centered so that 0 is the mean for each. The centered guess for a husband was 30 beans (above the mean). What would be your best estimate of the wife's guess? (3)

2C) Explore and discuss the strengths and limitations of employing the Poisson distribution emphasizing its utility in modeling real-world phenomena where the occurrence of events is tied to a specific time frame or proportion. (2)

3A) Consider this data set from the 2000 United States presidential election in the state of Florida. It records the number of votes each candidate received by county. We wish to investigate the relationship between the number of votes for Bush against the number of votes for Buchanan. Demonstrate the same for this example using R programming. Compute the slope after re-centering the data. Also, estimate the residual variation. Further, justify with this example how the least squares method finds the best fit for a set of data points. (5)

county < - c("County1", "County2", "County3", "County4", "County5")

bush_votes < - c(5000, 6000, 7000, 5500, 8000)

buchanan_votes < - c(1000, 1200, 1500, 1100, 1300)

3B) What is P-Value? What is the R Command used to find the probability of obtaining a t-statistic as (3)

large as 2.5 with 15 degrees of freedom when

$H_0 : \mu = \mu_0$ versus $H_a : \mu > \mu_0$?

3C) You randomly draw one card from a deck of 52 cards. (2)

A: The card is a face card (Jack, Queen, or King).

B: The card is a spade.

Explain what a conditional probability of 0.25 for P(B|A) signifies in this scenario?

4A) a) Provide insights into how understanding variance contributes to the analysis and interpretation of datasets. (5)

b) Present a specific scenario where knowledge of variance would be particularly valuable.

c)Write a R command to calculate the variance of a numeric vector or a column in a data frame. Include the necessary parameters, and if applicable, use a sample dataset or vector to demonstrate the application of the command.

d) What is the distribution of the sample variance of a random sample from a population is centered at?

e) What does the sample variance estimate when derived from a random sample taken from a population?

4B) Imagine a clinical trial is being designed to test the effectiveness of a new medication in reducing blood pressure. The researchers plan to compare the mean blood pressure levels before and after the treatment. In the absence of the treatment effect, the standard deviation of the blood pressure is known to be 10 mmHg. They aim to detect a mean reduction of at least 5 mmHg with a sample size of 30 participants. The significance level (α) is set at 0.05. (3)

a) Write a R command to perform a power calculation for this study. What are the different components of the command?

b) Interpret the result of the power calculation which is 0.80.

4C) Explain how you anticipate the relationship between statistical power and sample size when all other conditions are held constant. Provide a reasoned description of the expected impact on power as the sample size increases. (2)

5A) Write pseudocode to outline the bootstrap procedure for calculating a confidence interval for the median from a dataset with n observations. (5)

5B) Write a R command to simulate the averages of 10 standard normal random variables and explain the purpose of this command. Additionally, discuss how the number of simulations (nosim) influences the results and the interpretation of the simulated averages. (3)

5C) A sample of 12 men yielded a sample average brain volume of 1,150cc and a standard deviation of 40cc. What is the complete set of values of μ_0 that a test of H_0:$\mu=\mu_0$ would fail to reject the null hypothesis in a two-sided 5% Student's t-test? Provide the necessary R commands to solve this. (2)

-----End-----