Question Paper

Exam Date & Time: 07-Jan-2023 (02:30 PM - 05:30 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

VII Semester MakeUp Examination
DATA SCIENCE PART-II (CRA 4061)

DATA SCIENCE - PART II [CRA 4061]

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

Descriptive Questions

Answer all the questions. Section Duration: 180 mins

1) A) Suppose that in an AB test, we test two website design for an online retailer. The first design leads to an average of 10 purchases per day for a sample of 100 days, while the other leads to 11 purchase per day, also for a sample of 100 days. Assuming a common standard deviation of 4 purchases per day. If the groups are independent and the days are independent and identically distributed, calculate the z test statistics. Give the p-value for the test. Do you reject at the 5% level? (5)

B) A slow cooker manufacturer claims that the true average 'Low' temperature setting for their products is 130◦ C. A sample of 9 slow cookers are tested and the average 'Low' temperature setting is 131.08◦ C. If the distribution is normal with standard deviation 1.5◦C , does the data contradict the manufacturer's claim at significance level α = 0.01 ? (3)

C) If more graduates than 0.6 employ statistical inference within the first year of graduation, the dean of a business school wants to know. Imagine that the test statistic was 3.92. What would the conclusion be if the significance level was α = 0.10? (2)

2) A) The figure below is a bootstrap distribution that was generated for a sample mean (5)

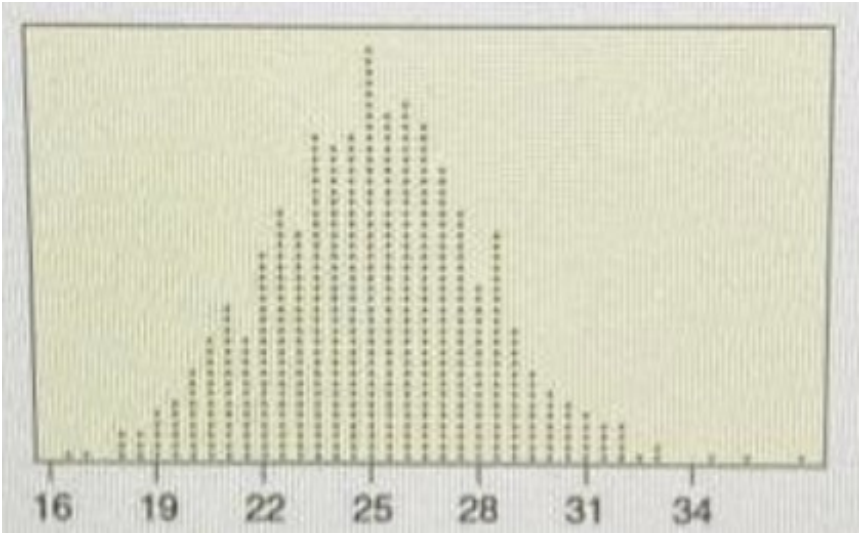

Fig: Bootstrap Distribution

a. Use the above bootstrap distribution to estimate the value of sample statistic

b. Estimate the standard error by estimating the standard deviation of above bootstrap

distribution

c. Use the standard error to construct a 95% confidence interval for the population mean

d. What is the notation (a single letter) that is used for the population mean?

B) Consider the data set with the values: 0,1,2,3,4. If X is a random variable of a random draw from these values and we define the probabilities of each of the outcomes using the probability mass function (PMF) (assuming the probabilities of all the outcomes were the same refer table Q3a). Find the Cumulative Distribution Function of the random variable X (3)

x	0	1	2	3	4
$f(x) = P(X = x)$	1/5	1/5	1/5	1/5	1/5

C) Describe the goals of inference with an example . (2)

3) Discuss the variants used for representing notation of the data. Use suitable examples to cite their application. (5)

A)

B) Assume that the number of daily ad clicks for a company is (approximately) normally distributed with a mean of 1020 and a standard deviation of 50. Find the probability of getting more than 1,160 clicks in a day and write the corresponding R code. (3)

C) Consider the dataset mtcars (name, mpg, cyl, disp, hp, gear, carb) data set. Write R code to find the p-value using z-test. (2)

4) Write R code to generate T-statistic and P-values assuming standard error is already computed and stored in variables sbeta0 and sbeat1. Use Coefficient table for diamond dataset, where diamond is the predictor, and price is the outcome. State the significance of P- value. (5)

A)

B) Write the Rcode to load the dataset SeatBelts(DriversKilled, kms, PetrolPrice). Take the residual for DriversKilled having regressed out kms and an intercept. Take the residual for PetrolPrice having regressed out kms and an intercept. Fit a regression through the origin of the two residuals and state the reason why the results are same as your coefficient for fitting the linear model of driver deaths with kms and PetrolPrice as predictors. (3)

C) Write an R code to generate confidence intervals w.r.t to intercept and slope based on the regression parameters generated below Table 4C. for the mtcars(name, mpg, cyl, disp, hp, gear, carb) dataset with miles per gallon(mpg) as the outcome and horsepower(hp) as the predictor.Center the horsepower first to create the confidence interval for the intercept. (2)

> fit < - lm(hp ~ mpg);

> summary(fit)$coefficients

Table 4C.

	Estimate	Std.Error	t value	Pr(>|t|)
(Intercept)	-259.6	17.32	-14.99	2.523e-19

5) Load the dataset Seatbelts as part of the datasets package via data(Seatbelts). Use as.data.frame to convert the object to a dataframe. Fit a linear model of driver deaths with kms and PetrolPrice as predictor. Add the factor variable law that takes the levels No and Yes. Fit the model with outcome as driver deaths with kms, PetrolPrice and dummy_var as predictor. Add the binary variable, dummy_var and fit the model with outcome as driver deaths with kms, PetrolPrice and dummy_var as predictor . Also discretize the ppwhihc is computes as pp=mean(PetrolPrice)-PetrolPrice/sd(PetrolPrice). PetrolPrice variable can be factored into four levels such as pp>5, pp< (5)

A)

=2,pp< =0,pp

B) Use the least square criteria to find the best fit line for GaltonFamilies (family,parent_height,gender,child_height) data with predictors as parents height and outcome as child height. Write the Rcode to get the best fit line. (3)

C) For the Father.son data, with Father's height(fheight) as predictor and son's height(sheight) as outcome, for the following output of fitting the model, if father's height was 63 inches, write the Rcode to predict the son's height? (2)

-----End-----