Question Paper

Exam Date & Time: 30-Nov-2022 (02:00 PM - 05:00 PM)

MANIPAL ACADEMY OF HIGHER EDUCATION

VII SEMESTER B.TECH END SEMESTER EXAMINATIONS, NOV 2022

DATA SCIENCE - PART II [CRA 4061]

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

Answer all the questions.

Instructions to Candidates: Answer ALL questions Missing data may be suitably assumed
T-distribution Table can referred from the uploaded formula book

1)		A random sample of 10 boys had the following I.Q's 70, 120, 110, 101, 88, 83, 95, 98, 107, 100. Do these data support the assumption of a population mean I.Q of 100 ?. Find a reasonable range in which most of the mean I.Q values of samples of 10 boys lie. (Consider the level of significance 5%)	(5)
	A)		
	B)	A sample of 100 men yielded an average PSA level of 3.0 with a σ of 1.1. What are the complete set of values that a 5% two sided Z test of H0 : μ = μ0 would fail to reject the null hypothesis for? Identify the lower and upper values?	(3)
	C)	Suppose a friend has 8 children, 7 of which are girls and none are twins. If each gender has an independent 50% probability for each birth, what's the probability of getting 7 or more girls out of 8 births?	(2)
2)		Suppose that 18 obese subjects were randomized, 9 each, to a new diet pill and a placebo. Subjects' body mass indices (BMIs) were measured at a baseline and again after having received the treatment or placebo for four weeks. The average difference from follow-up to the baseline (follow-up - baseline) was 3 kg/m^2 for the treated group and 1 kg/m2 for the placebo group. The corresponding standard deviations of the differences was 1.5 kg/m2 for the treatment group and 1.8 kg/m2 for the placebo group. The study aims to answer whether the change in BMI over the four-week period appear to differ between the treated and placebo groups. What is the pooled variance estimate? (to 2 decimal places)	(5)
	A)		
	B)	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?	(3)
	C)	Write the R code to find the probability of 2 heads in 10 coin flips where probability of heads is 0.3. .	(2)
3)		Formalize the definition of conditional probability of an event *A* given that *B* has occurred. Suppose 5% of housing projects have issues with asbestos. The sensitivity of a test for asbestos is 96% and the specificity is 89%. What is the probability that a housing project has no asbestos given a negative test expressed as a percentage to the nearest percentage point?	(5)
	A)		
	B)	Illustrate with example the conditions required for the Probability mass function to be valid	(3)

	C)	"For any two events the probability that at least one occurs is the sum of their probabilities minus their intersection." Justify the following statement with example.	(2)
4)		Analyze Wald interval's coverage using confidence intervals and investigate their frequency performance over repeated realizations of the experiment. Consider different values of p. Write the R code for investigating Wald interval coverage.	(5)
	A)		
	B)	Write R code to generate Estimate, standard Error, T statistic value using Coefficient table for diamond dataset, where diamond is the predictor, and price is the outcome.	(3)
	C)	Write the Rcode to best fit the linear regression model for the dataset student. For the given output, write the interpreted output, with marks(x) as predictor and performance(y) as outcome.	(2)
5)		Write the Rcode to do the following: load the SeatBelts(PetrolPrice,DriversKilled) dataset package via data(SeatBelts). Convert the object to dataframe. Assume new petrol price denoted as PP to be computed using mean(PetrolPrice)/Sd(PetrolPrice) and new kms, to be denoted as new_kms=mean(km)*1000. Fit the linear model with new values where DriversKilled is the outcome and predictors are PP and new_kms. Also predict the number of driver deaths at the mean kms and average PetrolPrice levels.	(5)
	A)		
	B)	Write the Rcode to get the best fit line based on least square criteria using GaltonFamilies data with child's height as predictor and parents height as outcome. Compute the slope after re-centering the data. Also mention the code to verify the output using linear model formula.	(3)
	C)	Write the Rcode to show that the normalizing the variable results in the slope being correlation. Use Galton data with parents height as predictor and child heights as outcome.	(2)

-----End-----