JARAMOGI OGINGA ODINGA UNIVERSITY OF SCIENCE AND TECHNOLOGY

SCHOOL OF BIOLOGICAL, PHYSICAL, MATHEMATICS AND ACTUARIAL SCIENCES

UNIVERSITY EXAMINATION FOR DEGREE OF BACHELOR OF SCIENCE IN ACTUARIAL SCIENCE

SPECIAL RESIT 2020/2021

MAIN REGULAR

COURSE CODE: SAS 205

COURSE TITLE: STATISTICAL COMPUTING I

EXAM VENUE: **STREAM: (BSc Actuarial Science)**

DATE: **EXAM SESSION:**

TIME: 3.00 HOURS

Instructions:

1. **Answer questions one and any other two.**
2. **Candidates are advised not to write on the question paper.**
3. **Candidates must hand in their answer booklets to the invigilator while in the examination room.**
4. **All computations and data analysis to be done with R statistical software.**

Examination dataset

In this series of questions, we examine data from a study of 158 infants who visited st. Mary's hospital, Langata in Nairobi for a vitamin K shot. Assume that the infants in the study are a representative sample from all infants in Kenya.

Nurses administered vitamin K shot to each infant. Infants were randomized to two different protocols to study how to reduce pain experienced by the infants due to the shot. The infants were divided into two groups – the control group, where standard protocol for handling the infants was used; and an intervention group, where mothers held their infants prior to, during, and after administration of the shot. Pain was measured using the Neonatal infant pain score (NIPS) (Lawrence et al. 1993). The variables in the dataset are described below:

id - unique identifier
group – 1 if intervention group, 0 if control
pain0 – NIPS score 0 seconds after shot
pain30 – NIPS score30 seconds after shot
pain60 – NIPS score60 seconds after shot
pain120 – NIPS score120 seconds after shot
crytime – total time that the infant cried in seconds

use the babies.dta or babies.csv dataset to answer all the questions below:

QUESTION ONE (30 Marks)

Exploratory Analysis

Before analyzing the babies.dta or babies.csv dataset, explore the dataset using summary statistics and graphical analyses.

a) Make a boxplot of cry time by group. According to the boxplot, which group has more variability in cry time? (2marks)

b) Using the central limit theorem, construct a 95% confidence interval for the average total cry time for infants in the control group and infants in the intervention group. For this question only, assume that the standard deviation of cry time within each group is known and is equal to 22 seconds.

i) Control

Lower Bound (2marks)

Upper Bound (2marks)

ii) Intervention

Lower Bound (2marks)

Upper Bound (2marks)

Paired data

c) To help measure standard infant reactions to the shot without the intervention, restrict analysis to the control group. Among the controls, examine whether NIPS scores decreased within 30 seconds of receiving the shot. Restrict your analysis in the following questions to the control group.

Generate a new covariate reflecting change in NIPS score over the 30 second interval (NIPS score at 30 seconds minus NIPS score at time of shot). In R, you can use the command “babies_ controls = babies[which(babies$group ==0),]".

First, examine some descriptive statistics.

i) What is the mean change in NIPS score over the 30 second interval? (2marks)

ii) What is the median change in NIPS score over the 30 second interval? (2marks)

iii) Is the distribution of change in NIPS score over the 30 second interval symmetric? (2marks)

iv) Test the null hypothesis that the average change in NIPS score over the 30 seconds is equal to 0, versus the alternative that the average change in NIPS score is different from 0. Conduct the test at the 0.05 level of significance. Assume that the change in NIPS score is normally distributed. What is the absolute value of the test statistic? (2marks)

v) What is the distribution of the test statistic under the null? (2marks)

vi) What is your p-value? (2marks)

vii) What is your conclusion? (2marks)

d) Now, conduct an analogous non-parametric test, testing the null hypothesis that the median change in NIPS score is equal to 0, versus the alternative that the median change in NIPS score is different from 0. Conduct the test at the 0.05 level of significance. Use the most powerful non-parametric test that you have available. (2marks)

i) What is the p-value? (2marks)

ii) What is your conclusion? (2marks)

iii) Examine the distribution of change in NIPS score. Compare the results of the parametric test that assumes normality versus the non-parametric test. Are you surprised that these two tests above gave somewhat similar conclusions? Why? (2marks)

QUESTION TWO (20 Marks)

Parametric test for infants experiencing severe pain

a) In this question, examine average cry time by group among infants who initially experienced severe pain. Restrict analysis to infants with a NIPS score of 7 immediately after receiving the shot.

Make sure that you start this question with the full babies.dta or babies.csv examination dataset. Then, use the covariate pain0 to construct the relevant subset. In R, you can use an appropriate command to "drop if pain0 < 7" to restrict to the appropriate subset.

Assume that, within group, cry time among infants who initially experienced severe pain follows a normal distribution.

b) Among infants with initial severe pain, estimate the average cry time in each group, with a corresponding 95% confidence interval.

i) Control

Estimate (2marks)

Lower Bound (2marks)

Upper Bound (2marks)

ii) Intervention

Estimate (2marks)

Lower Bound (2marks)

Upper Bound (2marks)

c) Among infants with initial severe pain, conduct a test of the null hypothesis that the two groups have equal means versus the alternative hypothesis that the means are not equal at the 0.05 level of significance. Assume that the variances within each group are equal.

i) What is the absolute value of your test statistic? (1mark)

ii) How many degrees of freedom does the test statistic have under the null? (1mark)
iii) What is the p-value? (1mark)

iv) What do you conclude from this test? (1mark)

d) Suppose instead we wanted to examine the change in NIPS score from time 0 to time 120 among infants within initial severe pain. Construct a new covariate that represents this change in NIPS score. Use graphical summaries to examine the distribution of this covariate. (4marks)

QUESTION THREE (20 Marks)

Two-sample Non-parametric Test

Now, examine the relationship between cry time and group among infants in Nairobi Kenya

a) Suppose we wish to perform a two-sample test, but do not want to make any normality (or other strong parametric) assumptions. Conduct an appropriate non-parametric test to test whether the distribution of cry time is the same in both groups at the 0.05 level of significance.

i) What is the p-value? (2marks)

ii) Your conclusion from the test? (3marks)

b) Assuming randomization was successful and all participants complied with their assigned exposure, explain whether each of the following affects the result and how?

i) Confounding by sex of the infant correct (3marks)

ii) Confounding by the amount of pain experienced by the infant (3marks)

iii) Effect modification by sex of the infant (3marks)

iv) Misclassification of the exposure status of the infant (3marks)

c) Provide detailed comparison between the result in (b) above to a parametric equivalence (3marks)

QUESTION FOUR (20 Marks)

Linear Regression

In the babies.dta full dataset, generate a covariate called painind defined as 1 if the infant experienced severe pain upon receiving the shot (pain0 = 7) and as 0 otherwise. In R, use an appropriate command to create the new variable:

Fit a linear regression model with total cry time as the outcome; and with group and painind (the severe pain indicator) as covariates. The regression model is:

$y_i = \beta_0 + \beta_1 group_i + \beta_2 painind_i + \epsilon_i$

where

$$\epsilon_i \sim N(0, \sigma^2)$$

.

a) Using the notation from the model above, what are your estimates of the regression coefficients and residual standard deviation, interpret the parameters?

β_0 (2marks)

β_1 (2marks)

β_2 (2marks)

σ (2marks)

b) Using the fitted regression model, estimate the average change in cry time for infants with severe pain versus those without severe pain, holding group constant. Provide a 95% confidence interval for this estimate.

Estimate: (2marks)

95% Confidence interval Lower Bound: (2marks)

95% Confidence interval Upper Bound: (2marks)

c) Again, use the notation above for the regression model. The correct interpretation for is:

β_1 (2marks)

d) Using the regression model, estimate the average cry time in the following groups:

i) Control group infants with severe pain upon receiving the shot (2marks)

ii) Control group infants without severe pain upon receiving the shot (2marks)

QUESTION FIVE (20 Marks)

a) Without using the regression model, estimate the mean cry time in the following groups:

i) Control group infants with severe pain upon receiving the shot (2marks)

ii) Control group infants without severe pain upon receiving the shot (2marks)

iii) Intervention group infants with severe pain upon receiving the shot (2marks)

iv) Intervention group infants without severe pain upon receiving the shot (2marks)

We refer to these as 'non-parametric' estimates, because they do not rely on modeling assumptions (whereas the estimates in question 4 are based on the linear regression model).

b) Compare estimates of the group-specific means from the regression model to the "non-parametric" estimates above. In large sample sizes, would you expect the "non-parametric" estimates or the regression based estimates to have less bias (e.g. be closer to the true group-specific means in the population)? (2marks)

c) With continuous covariates, we cannot estimate the means using the non-parametric method as above due to the "curse of dimensionality." This is because: explain. (2marks)

d) Suppose that sex is an effect modifier of the association between group and cry time. Which of the following is a correct way to analyze the data? (2marks)

e) Construct a linear regression model with sex as a covariate providing fitted model including estimated parameters, standard errors, 95% confidence interval, and coefficient of determination. (2marks)

f) Construct two separate linear regression models: one among male infants and one among female infants. Provide the two separate regression models including estimated parameters, standard errors, 95% confidence interval, and coefficient of determination. (2marks)

g) Construct a linear regression model, but do not control for sex as a covariate because this is a randomized clinical trial. Give the regression models obtained including estimated parameters, standard errors, 95% confidence interval, and coefficient of determination. (2marks)