W1-2-60-1-6

JOMO KENYATTA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY

UNIVERSITY EXAMINATIONS 2021/2022

EXAMINATIONS

TID 3103/PEH 3102: BIOSTATISTICS AND DEMOGRAPHY/BIOSTATISTICS

DATE: AUGUST 2022 **TIME: 3 HOURS**

INSTRUCTIONS: ANSWER ANY FOUR (4) QUESTIONS

QUESTION ONE: 25 MARKS

(a) For each statement, answer true or false and give an explanation for your answer. (2 marks each)

(i) For measurements of a large data set with values that are approximately of the same magnitude except for a few measurements that are exceptionally large; the mean would be larger than the median and the histogram would be skewed with a long left tail.

(ii) The shape of the distribution of all possiblevaluesof a random variable can be described using a probability distribution

(iii) A random sample of size nine is drawn from a population that follows normal distribution with mean equal to 10 and variance equal to 36. The sampling distribution of the sample mean is normal distribution with mean of 10 and variance of 4.

(iv) If there is sufficient evidence to reject a null hypothesis at 0.01 level of significance, then there is sufficient evidence to reject it at 0.05 level of significance.

(v) If a representative sample gives a 95% confidence interval for population mean μ as (2.34,5.98), then roughly 95% of the time in repetitions of the experiment, this interval will capture μ

(b) The frequency distribution beow shows the heights in inches of200 students:

Interval of measurement	No of students
60- 64	15
65-69	54
70-74	98
75-79	24
80-84	9

(i) Which category does the median height measurement fall in?

(1 mark)

(iii) From the grouped frequency data, the average height of students is calculated as 70.95inches. What do you conclude from the mean, median and the frequencies shown in the table above; about the shape of the distribution that these data came from? Give reasons for your answer

(4 marks)

(c) A health agency reports that t 71% of female university students use contraceptives regularly. A researcher believes the percentage is higher than reported value. To this end, he randomly selects 1250 female students and found that 962 use contraceptives regularly.

(i) Identify the parameter and statistic of interest?

(2 marks)

(ii) State the null and alternative hypotheses of the test.

(2 marks)

(iii) State the type I and type II errors

(4 marks)

(iv) If the p-value of the test carried out is < 0.0001, what conclusiondid the resarcher make?

(2 marks)

QUESTION TWO: 25 MARKS

(a) Give an example of study where each of the the following graphical tools is appropriate for descriptive analysis

(i) Mosaic plots

(ii) Multiple bar charts

(iii) Box plots

(iv) doughnut charts

(vi) Scatter plots

(10 marks)

(b) (i) In an actuarial study involving a random sample of 25 people, the average age at the time of their death was 74.32 years and standard deviation of 18.549 years. Using this information is the average age at death significantly different from 77 years?

(4 marks)

(iii) A non-parametric test is carried out to detrmine if the median age at time of death is significantly different from 77 years. Part of the results were as follows:

```
sign            obs         sum ranks expected
---------------------------------------------
positive      12       109      162.5
negative    13       216      162.5
---------------------------------------------
 test statistic: z = -1.440 p-value=0.149
```

At 0.05 level of significance, what conclusion can be made?

(2 marks)

(iii) Compare the results of the tests in part(i) and part(ii).

(2 marks)

(c) A health research center reported that for a given population, the proportion of those who have health insurance is 32%. Another research group claims the proportion is higher and seeks to verify their claim. A random sample of 1500 individuals is selected and 34.8% are foundto have health insurance.

(i) what is the sampling distribution of sample proportion? Explain why this distribution is valid.

(3 marks)

(ii) Construct the confidence interval for the test applied to verify the claim.

(2 marks)

(iii) What conclusion is drawn by the resarch group?

(2 marks)

QUESTION THREE: 25 MARKS

(a) An weight management group introduces an intervention program to help individuals lose weight progressively. To determine if it is effective, fifteen individuals considered to be overweight are selected and enrolled into the program. Their weights(in pounds) are recorded at four time intervals: before beginning the program, one month after beginning, one month later and finally at the end of the program(program takes three months). Repeated measures ANOVA technique is used to analyze the recorded weights and part of the results were as follows:

```
                 Analysis of Variance
Source             SS        df        MS          F     Prob > F
---------------------------------------------------------------
Subject       17111.433     14       1222.25
Interval       3099.067      3       1033.022    137.595   0.0000
Error         1924.433      42         80.185
---------------------------------------------------------------
Mauchy's Sphericity test: p-value = 0.0843
```

(i) Comment on the assumptions made for the analysis that resulted in results above and if they are met.

(5 marks)

(ii) Interpret the results obtained

(2 marks)

(b) Scientists conducted an experiment to test the effects of five different diets in turkeys.They randomly assigned six turkeys to each of the five diet groups and fed them for a fixed period of time. The weight gains were recorded and analyzed using non parametric techniques. Part of the results of the analysis were as follows:

```
Group    1    2    3    4    5
Obs       6    6     6    6    6
Ranksum 21 61.5 106.5 111 165
chi-squared with ties = 25.536   p-value=0.00004
```

(i) What test was used for analysis? Why do you think was the reason for using this test?

(3 marks)

(ii) Interpret the output given

(2 marks)

(c) A study of 100 patients is performed to determine if cholesterol levels are lowered after three months of taking a new drug. Cholesterol levels are measured on each individual at the beginning of the study and three months later. On average the cholesterol levels among these 100 patients decreased by 15.0 and the standard deviation of the changes in cholesterol was 40.

(i) Compute the test statistic.

(2 marks)

(ii) If critical value is -1.645, what conclusion can be made?

(2 marks)

(iii) If the assumptions made for this test are not met, what is the alternative test? Statethe hypotheses test in this case.

(3 marks)

(d) One hundred and sixty three heterosexual couples, at least one of whom was HIV-infected: were enrolled in an HIV transmission study whose aim was to promote condom use as a means of reducing HIV infection. 68 couples indicated use of condoms before aund after intervention campaign while 52 couples did not use condoms at all(neither before or after intervention). In addition 38 couples who were not not using condom before noe began using after the intervention while the rest stopped using condoms after campaign.

(i) Represent this information in a two way contingency table

(2 marks)

(iii) Which statistical test for used for analysis? Give reason for your answer

(2 marks)

(iii) If the p-value of the test is 0.0236; interpret the findingsof the study.

(2 marks)

QUESTION FOUR: 25 MARKS

(a) According to an article in a medical journal, the type of cancer suffered by individuals in a certain county is the same for males and females.A research team set out to test this claim and hence selected two samples of 200 men and 200 women diagnosed as having cancer. The results of analysis were as follows:

```
       |          cancer   |
gender | lung other stomach | Total
-------------------------------------------

female | 60     68     72    | 200
male   | 76     55     69    | 200
-------------------------------------------

Total | 136    123    141   | 400
Pearson chi2 statistic = 3.3201 p-value=0.1901
```

(i) State the appropriate test to used for analysis in this case and hypotheses to be tested.

(3 mark)

(ii) Interpret the results.

(2 marks)

(b) A community-based prospective cohort study of randomly selected HIV discordant and concordant negative couples was carried out to determine the association between incident HIV infection and pregnancy intention (= 0 if did not intend pregnancy, =1 if intend pregnancy). The researcher felt that consistent use of condoms with partner can be treated as third variable for this investigation. A stratified Analysis was carried out and results obtained were as follows: Mantel Haenszel odds ratio estimate = 1.0472 95% confidence interval of odds ratio= [0.653; 1.68] Breslow Day Test: test statistic = 4.564 p-value=0.102 Interpret the result.

(6 marks)

(c) A cross-sectional analysis of singleton live births in a developed nations was carried out to determine the impact of maternal body mass index on external cephalic version success. Part of the results of the analysis were as follows:

BMI	<18.5	$18.5- < 25$	$25- < 30$	$30-< 35$	$35-< 40$	≥ 40
Successful	1229	16938	8269	4510	2010	1244
Failure	603	8412	4305	2350	1134	884
Total	1832	25350	12574	6860	3144	2128

Given that the test statistic value is 50.833,what conclusions would you make?(use critical value =3.84)

(7 marks)

(d) A medical study was done to evaluate the effectiveness of a cholesterol lowering drug. The treatment group had 125 subjects and the control group had 125 subjects. The cholesterol levels of the subjects were measured both before and after the treatment. The results were that 97 of the subjects in the treatment group had a reduced cholesterol level while 78 of the control subjects had a reduced cholesterol level. The population parameter is proportionof treatment-proportion of control

(i) Construct the corresponding 95% confidence interval for the test and interpret it.

(5 marks)

(ii) Can you conclude that the drug is effective at 0.05 level of significance?

(2 marks)

QUESTION FIVE: 25 MARKS

(a) A weight management group introduces an intervention program to help individuals lose weight progressively. To determine if it is effective, ten individuals considered to be overweight are selected and enrolled into the program. Their weights(in pounds) are recorded at four time intervals: before beginning the program(1), one month after beginning(2), one month later(3) and finally at the end of the program(4).
A non parametric test was used to analyze the data and part of the results were results were as below:

```
---------------------------------------
drug          1      2      3     4
---------------------------------------
Rank Sum    31.5 26.5 25.0 17.0
---------------------------------------
test statistic = 6.51 probability = 0.089
```

(i) What test was used for analysis? Give reason why it was used. statistic? (3 marks)

(ii) Interpret the output given. (2 marks)

(b) A medical researcher wanted to compare two different methods, I and II, for measuring cardiac output. Data are measured on 26 subjects using two methods.The results of the comparison were as follows:

```
----------------------------------------------------------------
Variable    Obs Mean    Std. Err. Std. Dev. [95\% Conf. Interval]
----------------------------------------------------------------
methodI     26    5.5054   0.29175   1.48764
methodII    26    5.0881   0.33715   1.71182
----------------------------------------------------------------
 diff     26   0.4173    0.11729   0.59805    0.17575 0.65886
----------------------------------------------------------------
Ho: mean(diff) = 0   mean(diff) = mean(method I - method II) t = 3.558
Ha: mean(diff) < 0   Ha: mean(diff) != 0          Ha: mean(diff) > 0
Pr(T < t) = 0.9992   Pr(|T|> |t|) = 0.001526    Pr(T > t) = 0.000763
```

(i) State assumptionsmade for the test carried out and how to validate them.

(4 marks)

(ii) Interpret the results

(4 marks)

(c) In a clinical study, the aim was to examine the effect of a low carbohydrate diet on the level of LDL(low diet lipoprotein) cholesterol in blood(mmol/litre). The participants in the study were women aged between 18 and 30 years of age, and had a BMI of between 24.5 and 27.5. For each of them the effect of diet on LDL level was calculated as the difference between the LDL after and before the diet. A negative value means LDL level is lower after the diet than before the diet. A non parametric test was used to analysis the data and results yielded a p-value of 0.291.

(i) State two possible scenarios that led the researcher to use a non parametric test.

(2 marks)

(ii) State the null and alternative hypothesis for the test .

(2 mark)

(iii) Interpret theresults.

(2 marks)

(d) Researchers studied the association between birth mothers' smoking habits and the birth weights(in kgs) of their babies. Group 1 consisted of non-smokers.Group 2 comprised smokers who smoked less than one pack of cigarettes per day, Group 3 comprised of mothers who smoked between one and two packs per day and Group 4 smoked more than two packs per day. 15 mothers from each group were randomly selected and birth weight of infants recorded. Analysis of the comparison of the average birth weight of the four different categories of smoking status yielded the following results:

```
Shapiro Wilks test:
Group      1          2          3.          4
p-value   0.42      0.258      0.139      0.075
************************************************************
Analysis of Variance
Source              SS          df     MS            F       Prob > F
--------------------------------------------------------------
Between groups 13553106.4   3     4517702.15 145.25 0.0000
Within groups 1741787.73   56     31103.3524
--------------------------------------------------------------
Bartlett's test for equal variances:chi2(3) =6.6403 Prob>chi2 = 0.0843
```

(ii) Were the assumptions made for this test met? Give reasonfor your answer.
(4 marks)

(iii) Interpret theresults.
(2 marks)

QUESTION SIX: 25 MARKS

(a) Distinguish between

(i) Correlation Analysis and Linear regression analysis

(ii) Pearson's correlation coefficient and Spearman's rank correlation

(3 marks)

(b) A study is carried out to determine the relationship between age and EEG(electrocephalography). Data were collected on 20 subjects between ages 20 and 60 years. The investigator wishes to know if it can be concluded that this particular EEG output is inversely correlated with age.
The results of the test carried out were as follows:
Shapiro test for normality:

variable	p-value
age	0.422
EEG	0.023

Test for correlation: p-value = 0.0002

(i) Which test for used for analysis in this case? give reason for answer.

(2 marks)

(ii) What conulsion did the investigator make?

(2 marks)

(c) As part of a study carried out among males suffering from hypertension; an investigator sort to determine if there was a linear relationship between cholesterol level and: duration of exrcise(in hours per week), diet inventory score and age(<45 or ≥ 45) of men. A summary of results of the fit is:

Variable	Estimate	p-value
Intercept	8.445	
duration	-0.539	< 0.0001
≥ 45	0.439	0.022

(i) Interpret the parameter estimates

(4 marks)

(ii) Comment on the significance of each predictor

(4 marks)

(d) As part of a research study to examine the association between developing coronary heart disease(CHD) and: obesity(=1 if obese, =0, otherwise) and smoking (=1, if smoker and =0, if non-smoker), a binary logistic regression model was fitted with systolic blood pressure (=1, if sbp < 130;=0 otherwise) as an extraneous variable. The results of the fit were as follows:

Variable	exp(estimate)	p-value
obesity	1.767	0.002
smoking	1.57	< 0.0001
sbp	2.43	< 0.0001
sbp×obesity	0.453	0.0027

Interpret the results .

(10 marks)