"Investing in Africa's Future"

COLLEGE OF HEALTH, AGRICULTURE AND NATURAL SCIENCES

SPH 541: HEALTH STATISTICS

END OF SECOND SEMESTER EXAMINATIONS

MAY/JUNE 2021

LECTURER: MR. E. CHIKAKA

DURATION: 7 HOURS

INSTRUCTIONS

ANSWER **ANY ONE** QUESTION

PLEASE STICK TO THE STANDARD HOUSE STYLE i.e.

- TIMES NEW ROMAN
- FONT SIZE 12
- DOUBLE SPACING
- APA REFERENCING
- SEND YOUR ANSWER AS A PDF DOCUMENT

THE MARK ALLOCATION FOR EACH QUESTION IS INDICATED AT THE END OF THE QUESTION

CREDIT WILL BE GIVEN FOR LOGICAL, SYSTEMATIC AND NEAT PRESENTATION

QUESTION 1

a. What points should one observe while using percentages in statistics? Write a brief note on different types of analysis of data pointing out the significance of each. [5]

b. What is a test of significance? Explain each test pointing out where and when it is used, its merits and demerits. [10]

c. In a study, suppose you collected data that attempted to determine whether high cholesterol levels affect the frequency of heart disease. Two thousand eight hundred and fifty subjects with heart disease were compared to 2974 healthy subjects of similar age, and both groups were asked whether they were diagnosed with "high" cholesterol levels in the preceding 2 years. Among the 2850 heart disease patients, 358 were classified as having "high" cholesterol levels, while 229 of the 2974 disease-free subjects were recorded as having "high" cholesterol levels.

i. Construct the 2 * 2 table for these data making sure to clearly label both axes.[2]

ii. Compute the appropriate measure of association and provide a one sentence interpretation. [5]

iii. Is your measure of association statistically significant? Why or why not? Show formula and calculation. [10]

d. What do you understand by multivariate analysis? Explain how it differs from bivariate analysis. [10]

e. Why is the Sample Size Important? What influences and does not influence sample size calculation for different study designs? [10]

i. We want to estimate the mean systolic blood pressure of Zimbabwean females. The standard deviation is around 20 mmHg. We wish to estimate the true mean to within 10 mmHg with 95% confidence. What is the required sample size? [3]

ii. Two competing therapies for a particular cancer are to be evaluated by the cohort study strategy in a multi-center clinical trial. Patients are randomized to either treatment A or B and are followed for recurrence of disease for 5 years following treatment. How many patients should be studied in each of two arms of the trial in order to be 90% of rejecting Ho: RR=1 in favor of the alternative Ha: RR±1,if the test is to be performed at the α=0.05 level and if it is assumed that p_2=0.35 and RR=1.5[5]

f. What are the assumptions for Linear regression and how do you check them [5]

g. A data set contains information about the hourly wage (in U.S. dollars) and the gender of each of the 534 AU workers surveyed in 2016, as well as information about each worker's age, union membership, and education level (measured in years of education). A linear regression analysis was performed to model hourly wage as a function of worker

sex, number of years of education, and union membership. Below find the results from this regression:

Hourly Wage (in U.S. dollars)	Regression Coefficient	Standard Error of Regression Coefficient
sex (1 = Female, 0 = Male)	-2.5	0.8
union member (1 = yes, 0 = no)	2.9	0.5
years of education	0.76	0.08
Intercept	-0.3	1

a. Write down the regression equation [2]

b. Interpret the regression coefficients for sex, union member, years of education and the intercept. [6]

c. Using the above regression results, estimate the mean hourly wage in 2016 for:

i. male workers with a high school education (13 years of education), who were not union members. [2]

ii. female workers with a first-degree education (17 years of education), who were union members. [2]

d. Give a 95% confidence interval for the mean difference in hourly wages for male workers with a high school education who were union members when compared to male workers with a high school education who were not union members. [4]

e. What do you understand by survival analysis and what is its significance in modern day statistics? [5]

h. Data from 27 patients receiving a non-depleted allergenic bone marrow transplant were examined to see which variables were significantly related to survival time. Backward stepwise Cox regression analysis using diagnosis (CML or not), recipient's age and sex, donor' age and sex, whether the donor had been pregnant, the index of mixed epidermal cell-lymphocyte reactions and whether or not patient developed graft vs host disease (GvHD) to predict survival yields the following model:

Variable	**Regression Coefficient**	**Standard Error**
GvHD (0=No, 1=Yes)	2.3346	0.2318
CML (0=No, 1= Yes)	-1. 808	0.4444

i. What is the interpretation of the opposite signs for the regression coefficients? [2]

ii. Calculate the relative risks of dying (hazard ratios) for the following patients relative to non-GvHD non-CML patients:

(a) With GvHD but not CML

(b) CML but without GvHD

(c) CML and GvHD [6]

iii. Calculate the 90% confidence interval for the hazard ratio associated with GvHD [4]

iv. Comment on the reliability of the Cox regression model in view of the sample size (27) and the number of deaths (18). [2]

QUESTION 2

a Discuss how the sample size of a study can influence the results of inferential statistical tests with regard to clinical significance. [5]

b Describe type I and type II errors, and how they are related to power and significance level of a study. [5]

c. Explain some of the issues to consider when estimating the required sample size for a study. [5]

i. We wish to estimate the proportion of Zimbabwe males who smoke. What sample size do we require to achieve a 95% confidence interval of width ± 5% (that is to be within 5% of the true value)? In a study some years ago that found approximately 30% were smokers. [4]

d. What are the assumptions for Logistic regression and how do you check them. [4]

e. Former kicker for the CHEATERS Football team, Chris Hetler, was very good at making field goals in the 2018 season, but in the 2019 regular season had only made 3 out of 12. The following is the Logistic Regression Output to predict the probability of making a field goal (yes/no), based on how far the kick is (in yards) and the year (2018 or 2019).

Logistic Regression Table

Predictor	Coef	SE Coef	Z	P
Constant	8312.97	3073.50	2.70	0.007
yards	-0.173760	0.0901421	-1.93	0.054
year	-4.14174	1.53141	-2.70	0.007

i. What kind of variables do we have here? [2]
ii. Write down the fitted logistic regression equation [2]
iii. The coefficients of yards and years are both negative. What does that mean? [4]
Find the probability of making a field goal:
iv. from the 30 yd line in 2006 [2]
v. from the 30 yd line in 2005 [2]
vi. from the 40 yd line in 2006 [2]
vii. from the 40 yd line in 2005 [2]

f. What is a Non Parametric Test? [2]

g. When is it used and for which data types? [5]

h. What are the advantages and disadvantages of non-parametric tests? [5]

i. Can you give the parametric equivalence of the following non-parametric tests? [7]

NONPARAMETRIC TEST	PARAMETRIC ALTERNATIVE
1-SAMPLE SIGN TEST	
1-SAMPLE WILCOXON SIGNED RANK TEST	
FRIEDMAN TEST	
KRUSKAL-WALLIS TEST	
MANN-WHITNEY TEST	
MOOD'S MEDIAN TEST	
SPEARMAN RANK CORRELATION	

j. Using STATA package to fit a multiple regression equation with blood pressure as the dependent variable, the independent variables being age, weight, body surface area, duration of hypertension, basal pulse and measure of stress using the data given below:

Y = mean arterial blood pressure (mm Hg)
X_1 = age (years)
X_2 = weight (kg)
X_3 = body surface area (sq m)
X_4 = duration of hypertension (years)
X_5 = basal pulse (beats/min)
X_6 = measure of stress

The following output was obtained.

```
. regress mabp Age Weight bsa Duration basalpulse Stresslevel

      Source |       SS       df       MS              Number of obs =      20
-------------+------------------------------           F(  6,    13) =  560.64
       Model |  557.844135     6  92.9740225            Prob > F      =  0.0000
    Residual |  2.1558651     13  0.165835777           R-squared     =  0.9962
-------------+------------------------------           Adj R-squared =  0.9944
       Total |       560     19  29.4736842            Root MSE      =  .40723

------------------------------------------------------------------------------
        mabp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         Age |   0.7032596   .0496059    14.18   0.000     .5960926    .8104266
      Weight |   0.9699192   .0631086    15.37   0.000     .8335815    1.106257
         Bsa |    3.776502   1.580154     2.39   0.033     .3627878    7.190217
    Duration |   0.0683829   .0484416     1.41   0.182    -.0362687    .1730346
  basalpulse |  -0.0844846   .0516091    -1.64   0.126    -.1959792    .0270101
  Stresslevel |  0.0055715   .0034123     1.63   0.126    -.0018003    .0129433
       _cons |   -12.87047   2.556654    -5.03   0.000    -18.39378   -7.34715
------------------------------------------------------------------------------
```

a) What are the assumptions for ANOVA and multiple linear regression analysis and how do you check for them? [8]
b) Using the output above, which of the variables are significantly related to mean arterial blood pressure and why? [6]
c) Is the model significant? Why? [2]
d) Interpret fully R^2 , Adjusted R^2 and Root MSE [6]
e) What is the coefficient for Age, Weight and Bsa? Interpret each of the coefficients. [6]
f) Which variable contributed the most and least to the model and why? [4]
g) Write a short report of your findings ie objectives, Research Question, Null Hypothesis, Data, Data Analysis, Results and Conclusion. [10]

QUESTION 3

a. Write a brief essay on statistical estimation [4]

b. Distinguish between the following:

(i) Confidence level and significance level; [4]

(ii) Correlation and Regression [4]

(iii) Statistical significance and statistical Inference [4]

(iv) Type 1 and Type 2 errors.

c. Explain some of the issues to consider when estimating the required sample size for a study. [4]

d. In a study conducted by Familiar et al (1989) two drugs for the treatment of peptic ulcer were compared. The percentage of ulcers healed by the drug pirenzepine was 85% and 73% of ulcers were healed by trithiozone.

(i) How many subjects would be required for a randomised trial to have 80% chance of detecting this difference (85% and 73%) at the 0.05 (two-tailed) level of significance? [2]

(ii) How many subjects would be required to increase the power to 90%? [2]

(iii) Taking the healing rate for trithiozone to be 70%, what sample size would be required to detect an increase of 17% in the healing rates due to pirenzepine? (Use 90% power) [2]

(iv) If only 30% of the subjects are to be given trithiozone, what sample size is now required to detect the difference described in part (i) [3]

e. For each of the following stories, determine which type of statistical analysis that would be appropriate to use and why. **Use each type of analysis only once.**

i) ANOVA

ii) Kruskal-Wallis

iii) Paired t test

iv) Wilcoxon Rank-Sum Test

v) Two sample t-test

i. Compare the average number of hours per week spent on Facebook for Freshmen, Sophomore, Juniors and Seniors at UF, based on a random sample of 100 students.

ii. Compare the distribution of the number of hours per week spent on Facebook for Freshmen, Sophomore, Juniors and Seniors at UF, based on random samples of 10 students per group, which had quite different standard deviations.

iii. Compare the average number of hours per week spent on Facebook during the first week in April and the first week in May (finals week) for random students at UF, measured on the same 100 students.

iv. Compare the distribution of the number of hours per week spent on Facebook for male and female students at UF, based on a random sample of 10 students. There was an outlier in one of the groups.

v. Compare the average number of hours per week spent on Facebook for male and female students at UF, based on a random sample of 100 students. [10]

f. For each of the following stories, determine which would be the simplest type of statistical analysis that would be appropriate to use. **Use each type of analysis only once.**

i) Confidence Interval for One Proportion

ii) Contingency Table

iii) Simple Linear Regression

iv) Multiple Regression

v) Logistic Regression

i. Predict the average number of hours per week UF students spend on Facebook, based on their age and gender.

ii. Estimate the fraction of UF students who have Facebook accounts.

iii. Determine if the fraction of UF students who have Facebook accounts is different for Males and Females.

iv. Determine how the probability that a UF student has a Facebbok account changes with the student's age.

v. Predict the average number of hours per week UF students spend on Facebook, based on the student's age. [10]

i. Consider the data presented in the table below. A total of 12 hemophiliacs, all 40 years of age or younger at HIV seroconversion, were followed from the time of primary DIDS diagnosis between 2004 and 2009 until death. In all cases, transmission of HIV had occurred through infected blood products. The time at which the individual contracted AIDS was not known and treatment was not available for most of the patients.

Interval from primary Aids diagnosis until death for a sample of 12 hemophiliac patients at most 40 years of age at HIV seroconversion.

Patient Number	Survival (months)
1	2
2	3
3	3
4	6
5	7
6	9
7	16
8	18
9	18
10	27
11	30
12	35

a. Calculate the survival times using the Product-Limit Method. [8]

b. Construct the survival curve based on the product-limit estimate [5]

c. What is the median survival time for these patients? [2]

j. The results below are from an occupational health study of factors related to the development of byssinosis (a lung disease) in workers employed in a textile industry ***[Higgins and Koch, 1977. Source: Kleinbaum Kupper and Muller. Applied Regression Analysis and Other Multivariate Methods. PWS-KENT Publishing Company, 1988]***. The variables of interest are:

Byssinosis (outcome) 1: yes 0: no
Workplace (type of place worked in) 1: dusty 0: not dusty

Years of employment	Year 1	Year 2
<10	0	0
10-20	1	0
>20	0	1

Smoker 1: yes 0: no

Log Estimates — No of Obs = 5419; Chi2(4) = 278.30; Prob > chi2 = 0.0000; Pseudo R2 = 0.1884

Log Likelihood = -599.44488

Byssinosis	coef.	Std Error	z	P>|z|	95% Conf. Interval	
Wrkplace	3.669849	.1696234	15.74	0.000	2.337393	5.002305
Smoker	1.6193696	.1907348	3.247	0.001	1.2455187	1.9932205
Year1	0.5018791	.2488369	2.017	0.044	0.0141678	0.9895904
Year2	1.6707002	.1813142	3.699	0.000	1.315331	2.02607
_cons	-0.5121706	.2170068	-23.602	0.000	-5.547032	-4.696381

Logistic byssinosis wrkplace year1 year2 [fweight=count]

Log Estimates — No of Obs = 5419; Chi2(4) = 278.30; Prob > chi2 = 0.0000; Pseudo R2 = 0.1884

Log Likelihood = -599.44488

Byssinosis	Odds Ratio	Std Error	z	P>|z|	95% Conf. Interval
Workplace	a	2.448986	15.74	0.000	e
Smoker	b	.3543555	3.247	0.001	f
Year1	c	.4110343	2.017	0.044	g
Year2	d	.3545791	3.699	0.000	h

i. Compute and interpret the ORs a-d [8]
ii. Compute and interpret 95% Confidence Intervals e-h [8]

iii. What variables are statistically significantly associated with development of byssinosis? Why? [8]
iv. What does Pseudo R2= 0.1884 mean? [4]
v. What does the coefficient of 3.67 for the variable “workplace” mean? [4]
vi. Provide an explanation of the interpretation of the 95% confidence for the odds ratio for the variable “smoker”. [4]