W1-2-60-1-6

JOMO KENYATTA UNIVERSITY OF AGRICULTURE AND TECHNOLOGY

UNIVERSITY EXAMINATIONS 2023/2024

MASTER OF SCIENCE IN EPIDEMIOLOGY AND BIOSTATISTICS

PEH3105: BIOSTATISTICS FOR EPIDEMIOLOGICAL STUDIES

DATE: DECEMBER 2023 **TIME: 3 HOURS**

INSTRUCTIONS: ANSWER ANY FOUR QUESTIONS

1)

(a) Regarding hypothesis testing briefly explain the meaning of the following terms:

(i) Confidence Interval. *(3 Marks)*

(ii) Alternative Hypothesis. *(3 Marks)*

(iii) Significance Level. *(3 Marks)*

(iv) Two-Sided Test. *(3 Marks)*

(b) In Kenya, the distribution of systolic blood pressure (sbp) in people between 40 to 60 years is approximately normal with mean μ= 128mmHg and standard deviation σ=19mmHg. A master's student plans to conduct a study at the KNH Diabetic Clinic to determine if patients in the same age-group newly diagnosed with diabetes have the same mean. If the true mean sbp in these patients is as low as 123mmHg (or as high as 133mmHg) the student wants a risk of only 20% of failing to detect this difference. If the student planned to conduct a two-sided test at the 0.05 significant level what sample size is needed for this study? *(6 Marks)*

$$\mathbf{n} = \frac{2\sigma^2(Z_{\alpha/2} + Z_{1-\beta})^2}{d^2} \;;\; (Z_{\alpha/2} = 1.96 \; and \; Z_{1-\beta} = 0.84)$$

(c) In a human papilloma virus study, 190 randomly selected women attending a pap-smear clinic in Nairobi were asked about their lifetime number of sexual partners. The results of the study indicated an average of 5.5 sexual partners.

(i) Using a population standard deviation of 3.5 lifetime number of sexual partners, construct a 95% confidence interval for the lifetime number of sexual partners in this population. *(3 Marks)*

(ii) Would there be a significant difference if the same study had been done in 290 randomly selected women attending a family planning clinic in Mombasa County where the average lifetime number of sexual partners was 4.5? *(4 Marks)*

$$CI = \bar{x} - z \times \frac{\sigma}{\sqrt{n}} ; \bar{x} + z \times \frac{\sigma}{\sqrt{n}}$$

2)

(a) State the appropriate multivariable regression technique that would be used to analyse each of the following studies:

(i) A cross-sectional study of the association between alcohol consumption and serum cholesterol level in people aged 40 to 60 years. *(2 Marks)*

(ii) A randomised-controlled trial of COVID-19 virus vaccine in 60- to 80-year-old people to prevent cardiovascular deaths. *(2 Marks)*

(iii) A prospective cohort study post-partum women immediately after delivery for the risk factors for post-partum depression. *(2 Marks)*

(iv) A retrospective cohort study to determine if birthweight is associated with the birth order of the baby (1st-born, 2nd-born etc.) and the age of the mother. *(2 Marks)*

(v) A case-control study of the association between three risk factors (low fibre diet, smoking and alcohol consumption) and colon cancer. *(2 Marks)*

(b) In a study to see the effect of alcohol consumption on survival after diagnosis with liver cirrhosis the Cox Proportional Hazards model below was constructed where the effects of sex (1=female, 0=males), age in years and weight (kg) were also examined in a multivariate model. [NB/ follow-up time was in days for 3-years]:

```
. stcox sex alcohol weight age

  No. of subjects =          104                Number of obs    =        104
  No. of failures =           69
  Time at risk    =        68135
                                                LR chi2(4)       =      33.33
  Log likelihood  =   -272.11994                Prob > chi2      =     0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         sex |   2.034673   .6929452     2.09   0.037     1.043764    3.966312
     alcohol |    3.98968   1.416148     3.90   0.000     1.989765    7.999712
      weight |    1.00415   .0090888     0.46   0.647     .9864932    1.022122
         age |   .9888447   .0133477    -0.83   0.406     .9630267    1.015355
------------------------------------------------------------------------------
```

(i) What was the effect of sex on survival after diagnosis with liver cirrhosis? *(4 Marks)*

(ii) What was the effect of alcohol on survival after diagnosis with liver cirrhosis? *(4 Marks)*

(iii) Is there evidence of confounding by weight? *(3 Marks)*

(iv) Outline two other confounding variables which you might consider in this study? *(4 Marks)*

3) The Heart and Oestrogen/progestin Replacement Study (HERS) was a randomized, double-blind, placebo-controlled trial designed to test the efficacy and safety of oestrogen plus progestin therapy for prevention of recurrent coronary heart disease (CHD) events in women. The participants were postmenopausal women with a uterus and with CHD (as evidenced by prior myocardial infarction). Among the risk factors measured were blood glucose levels (mg/dl), weight (kg), systolic blood pressure (sbp), exercise (if the participant exercised or not), age (years) and the average number of alcoholic drinks per week (avgdrpwk). The following

is a STATA output of the simple regression model on the association between weight (predictor) and blood glucose level (outcome) in these women:

- `regress sbp glucose`

```
      Source |       SS           df       MS      Number of obs   =     5,627
-------------+----------------------------------   F(1, 5625)      =     22.30
       Model |  7903.6019         1  7903.6019   Prob > F        =    0.0000
    Residual |    1993187     5,625  354.344356   R-squared       =    0.0039
-------------+----------------------------------   Adj R-squared   =    0.0038
       Total |  2001090.6     5,626  355.686208   Root MSE        =    18.824

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     glucose |    .080238   .0169895     4.72   0.000      .046932    .1135439
       _cons |   126.3466    1.68247    75.10   0.000     123.0483    129.6449
------------------------------------------------------------------------------
```

(a) Briefly describe the relationship between blood glucose (glucose) and systolic blood pressure (sbp) using the results in this model. *(3 Marks)*

(b) Explain the meaning of the following outputs in the model:

(i) Glucose (P>| t | = 0.000) *(2 Marks)*

(ii) Coefficient (_cons = 126.3466) *(2 Marks)*

(iii) Probability (Prob > F = 0.0000) *(2 Marks)*

(c) Which two ways could be used to evaluate the fit of this model? *(4 Marks)*

(d) What assumptions were made in using the method of least squares to estimate the population line in this model? *(4 Marks)*

The following is a STATA output of the multiple regression model for the association between sbp (outcome) and blood glucose and several other predictor variables:

- `regress sbp glucose exercise weight age avgdrpwk`

```
      Source |       SS           df       MS      Number of obs   =     5,570
-------------+----------------------------------   F(5, 5564)      =     46.45
       Model |  79615.6382         5  15923.1276   Prob > F        =    0.0000
    Residual |  1907149.12     5,564  342.765837   R-squared       =    0.0401
-------------+----------------------------------   Adj R-squared   =    0.0392
       Total |  1986764.75     5,569   356.75431   Root MSE        =    18.514

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     glucose |   .0669272    .017215     3.89   0.000     .0331791    .1006754
    exercise |  -2.465935   .5152184    -4.79   0.000    -3.475964   -1.455905
      weight |   .0770876    .018634     4.14   0.000     .0405577    .1136175
         age |   .5154984   .0381471    13.51   0.000     .4407151    .5902817
    avgdrpwk |   .0504105   .0647057     0.78   0.436    -.0764381     .177259
       _cons |   88.61364    3.44207    25.74   0.000     81.86584    95.36145
------------------------------------------------------------------------------
```

(e) Briefly describe the relationship between blood glucose and sbp using the results in this multiple regression model? *(4 Marks)*

(f) Why do you think the coefficient of blood glucose reduced from 0.08 in the simple regression model to -0.067 in the multiple regression model? *(4 Marks)*

4) A study was done in Boston Massachusetts on a sample of 100 low birth weight infants to assess the risk of an infant experiencing germinal matric haemorrhage (bleeding into the brain). The Apgar score at birth (a measure of alertness/fitness at birth – usually from 1 to 10), infant's mother being diagnosed with toxaemia during pregnancy, the gestational age of the infant at birth and the gender of the infant were recorded as predictor variables. The following STATA output represents the results of this study showing the association between these predictors and development of germinal matrix haemorrhage (GMH) in these infants:

```
. logit grmhem apgar5 tox gestage sex

Iteration 0:   log likelihood = -42.270909
Iteration 1:   log likelihood = -38.062857
Iteration 2:   log likelihood = -37.540381
Iteration 3:   log likelihood = -37.535669
Iteration 4:   log likelihood = -37.535667

Logistic regression                               Number of obs   =        100
                                                  LR chi2(4)      =       9.47
                                                  Prob > chi2     =     0.0504
Log likelihood = -37.535667                       Pseudo R2       =     0.1120

------------------------------------------------------------------------------
      grmhem |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      apgar5 |  -.2214754   .1073151    -2.06   0.039     -.431809   -.0111417
         tox |  -1.037503   1.121466    -0.93   0.355    -3.235535    1.160529
     gestage |  -.0665708   .1232046    -0.54   0.589    -.3080474    .1749058
         sex |  -.7886638   .6462168    -1.22   0.222    -2.055225    .4778978
       _cons |   1.868738   3.500318     0.53   0.593    -4.991759    8.729235
------------------------------------------------------------------------------
```

(a) What is the relationship between Apgar score and germinal matrix haemorrhage? *(5 Marks)*

(b) Interpret the z statistic and associated *P-value* for the Apgar Score. *(4 Marks)*

(c) Interpret the likelihood ratio chi-squared test provided (LR chi^2) for the model. *(4 Marks)*

(d) If a particular female infant was born at a gestational age of 32 weeks to a mother who had toxaemia in pregnancy and had an Apgar score of 7 what is the probability that they would develop GMH? [female =0, male=1; toxaemia=1, No Toxaemia=0 *(8 Marks)*

(e) What is the odds ratio of developing GMH for infants born of mothers who had toxaemia during pregnancy? *(4 Marks)*

5) A PhD student conducted a study to evaluate the effect of parasitic infection on the white blood cell differential count (a test which measures the percentage of each type of white blood cell – neutrophils, lymphocytes, basophils, eosinophils etc.). Five hundred and twenty-nine (529) children between 0 to 10 years were enrolled in the study and their blood taken for WBC differential count and a stool sample was taken and analysed for parasitic infection. Background characteristics (e.g., age, gender, SES) were also taken. A two-sided t-test was carried out to determine the difference in mean eosinophil count between those diagnosed to have parasitic infection and those who were negative. The results are shown in the table below:

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
negative |     334     2.73512    .1746729    3.192264    2.391518    3.078721
positive |     195    3.245128    .2543856    3.552302    2.743412    3.746845
---------+--------------------------------------------------------------------
combined |     529    2.923119    .1450123    3.335284    2.638247    3.207991
---------+--------------------------------------------------------------------
    diff |           -.5100085    .3000506               -1.099451    .0794337
------------------------------------------------------------------------------
    diff = mean(negative) - mean(positive)                        t =  -1.6997
Ho: diff = 0                                     degrees of freedom =      527

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0449         Pr(|T| > |t|) = 0.0898          Pr(T > t) = 0.9551
```

(a) Interpret these t-test results. *(8 Marks)*

Linear regression analysis model was created to find out if the effect of parasitic infection on the percentage of eosinophils in blood was influenced by the age (in months) of the child. The model is given here below:

```
      Source |       SS       df       MS              Number of obs =     529
-------------+------------------------------           F(2, 526)     =   11.60
       Model |  248.067957     2  124.033978           Prob > F      =  0.0000
    Residual |  5625.46549   526  10.6948013           R-squared     =  0.0422
-------------+------------------------------           Adj R-squared =  0.0386
       Total |  5873.53345   528  11.1241164           Root MSE      =  3.2703

------------------------------------------------------------------------------
 eosinophils |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   parasites |   .2215738   .3016353     0.73   0.463     -.370984    .8141317
   agemonths |   .0215843   .0048023     4.49   0.000     .0121502    .0310184
       _cons |    1.47037    .333474     4.41   0.000     .8152655    2.125474
------------------------------------------------------------------------------
```

(b) Interpret the results in this model as to the effect of age in months of the children under study on mean percentage of eosinophils in blood (table above). *(7 Marks)*

(c) The significance test in the linear regression above for the association between percentage eosinophils and the age in months uses the *t-distribution with n-2 degrees of freedom*. Why is this and what are the two restrictions? *(4 Marks)*

A t-test was then carried out to further assess the relationship between age in months and parasitic infection (negative or positive) (see table below):

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
negative |     347    58.55908       1.579    29.41351    55.45343    61.66472
positive |     182    72.98352    2.214397    29.87386    68.61416    77.35287
---------+--------------------------------------------------------------------
combined |     529    63.52174    1.318701    30.33013    60.93119    66.11228
---------+--------------------------------------------------------------------
    diff |           -14.42444     2.70654               -19.74137   -9.107508
------------------------------------------------------------------------------
    diff = mean(negative) - mean(positive)                        t =  -5.3295
Ho: diff = 0                                     degrees of freedom =      527

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

(d) Comment on the results of this t-test in relation to the previous two tables (above). *(6 Marks)*

6) A case-control study was done to evaluate the hypothesis that vaginal spermicides, which are used as contraceptives, are associated with the occurrence of Down's syndrome among offspring born to women who used these contraceptive agents. The results, stratified by age, are summarized below:

Crude Data

	Downs Syndrome	
	Present	Absent
Spermicide use	8	218
No spermicide use	24	2290

Stratum Specific Data

Age < 40 years

	Downs Syndrome	
	Present	Absent
Spermicide use	6	208
No spermicide use	18	2118

Age ≥ 40 years

	Downs Syndrome	
	Present	Absent
Spermicide use	2	10
No spermicide use	6	172

(a) Calculate the crude odds ratio and the two stratum-specific odds ratios *(show your work)*. *(9 Marks)*

(b) Calculate the Mantel-Haenszel adjusted odds ratio *(use the formula below and show your work)*. *(5 Marks)*

$$OR_{MH} = \frac{\sum \frac{a_i d_i}{N_i}}{\sum \frac{b_i c_i}{N_i}}$$

The following STATA output is the Mantel-Haenszel adjusted odds ratio model generated from this study's data:

```
. cc   downs spermicide, by(age)

        Age Group |        OR       [95% Conf. Interval]   M-H Weight
------------------+-------------------------------------------------
        <40 years |   3.394231      1.089893    9.041589     1.593191
        ≥40 years |   5.733333      .4962908    37.34844     .3157895
------------------+-------------------------------------------------
            Crude |   3.501529       1.34218    8.170176
     M-H combined |   3.781172      1.666923     8.57704
------------------+-------------------------------------------------
Test of homogeneity (M-H)      chi2(1) =      0.28  Pr>chi2 = 0.5996

                     Test that combined OR = 1:
                                  Mantel-Haenszel chi2(1) =     11.63
                                                  Pr>chi2 =    0.0006
```

(c) Would you conclude that age is an effect modifier in the relationship between spermicide use and Down's syndrome? Why or why not? *(4 Marks)*

(d) Would you conclude that the age is a confounder in the relationship between spermicide use and Down's syndrome? Why or why not? *(4 Marks)*

(e) State in clear terms how age affects the relationship between spermicide use and Down's syndrome. *(3 Marks)*