AFRICA UNIVERSITY

A United Methodist-Related Institution

"Investing in Africa's future"

COLLEGE OF HEALTH AGRICULTURE & NATURAL SCIENCE (CHANS)

DEPARTMENT OF PUBLIC HEALTH AND NURSING [DPHN]

MASTERS OF PUBLIC HEALTH [MPH]

NSPH 541: HEALTH STATISTICS

END OF SECOND SEMESTER FINAL EXAMINATIONS

NOVEMBER – DECEMBER 2023

LECTURER: DR Z. M. ZINGONI

DURATION: 3 HRS

INSTRUCTIONS

Answer **ALL** Questions in **Section A** and **ANY 3** questions from **Section B**

The mark allocation for each question is indicated at the end of the question

Credit will be given for logical, systematic and neat presentations.

Answer ALL questions: 40 marks

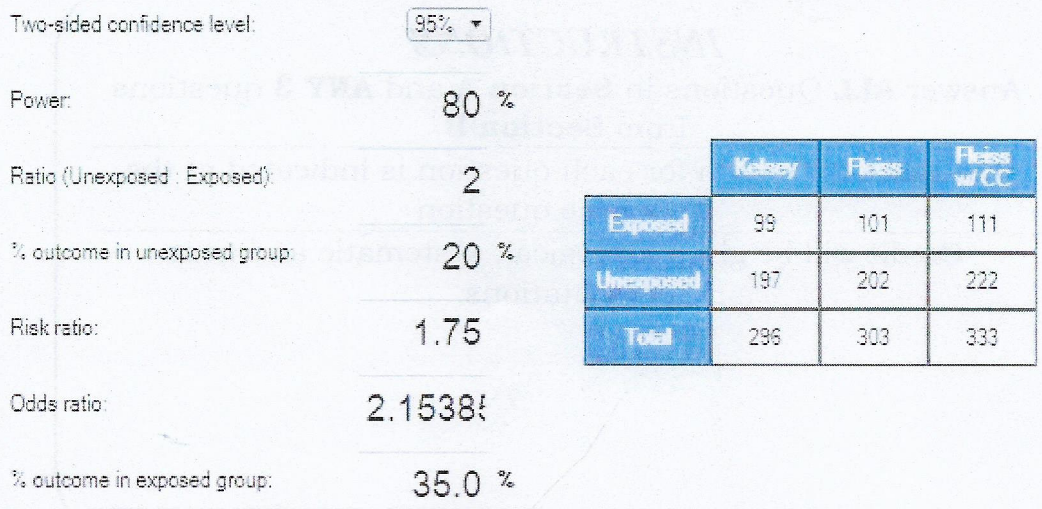
1. State the 3 approaches to hypothesis testing and their respective decision rule [6]

Approaches	Decision rule

2. In hypothesis testing, the decision made can either be correct or incorrect. Fill in the corresponding probabilities and terms associated with making correct or incorrect decisions. [4]

The Decision	The truth	
	Null is true	Null is false
Reject the null	A	B
Fail to reject the null	C	D

3. A researcher planned to conduct a study and asked an MPH student to help with sample size calculation. The student used Epi Info and the following output was produced.



```
Two-sided confidence level:           95% ▾
Power:                                80 %
Ratio (Unexposed : Exposed):          2
% outcome in unexposed group:         20 %
Risk ratio:                           1.75
Odds ratio:                           2.1538!
% outcome in exposed group:           35.0 %
```

	Kelsey	Fleiss	Fleiss w/ CC
Exposed	99	101	111
Unexposed	197	202	222
Total	296	303	333

i. State the alpha level used in this study [2]

ii. State the final sample size the researcher would use [1]

iii. If the power was changed from 80% to 90%, will the sample size increase or decrease? [2]

iv. Using the information provided in this output, re-calculate the sample size manually. [5]

v. If an attrition/non-response of 20% is factored into this calculation, what is the final sample size? [3]

4. Linear regression is one of the approaches used to determine how one variable influence another. Define each of the parameters in the linear regression equation [5]

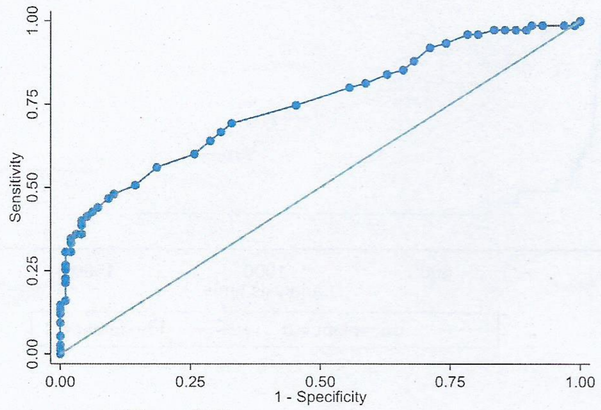
$$y= □_0 + □_1x +έ$$

5. Logistic regression is another type of model used in research. Describe the differences between linear regression and logistic regression based on the following: [4]

	Linear regression	Logistic regression
Dependent variable		
Measure of association		
The hypothesized value of the measure of association		
Model fitness		

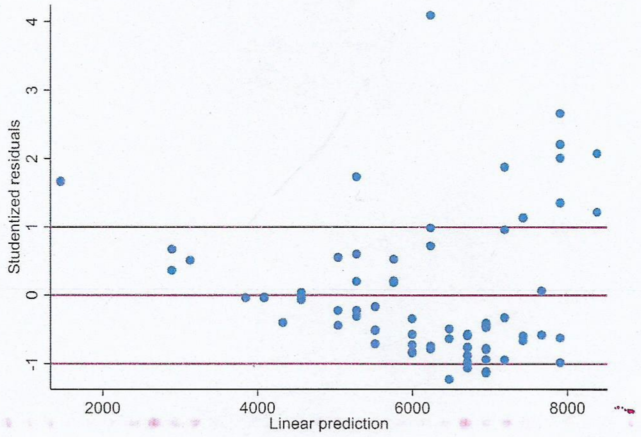
6. Comment on each of the plots below: what the plot is used for and what it means [8]

Plot A

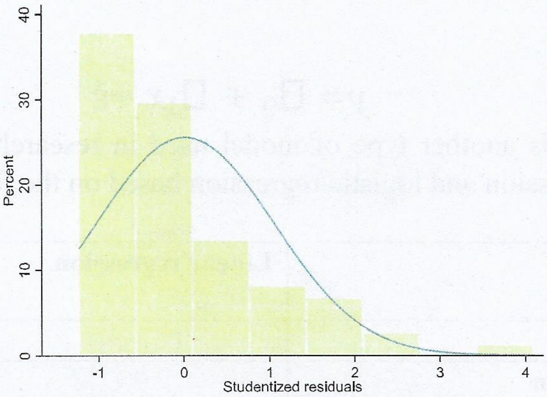


Area under ROC curve = 0.7492

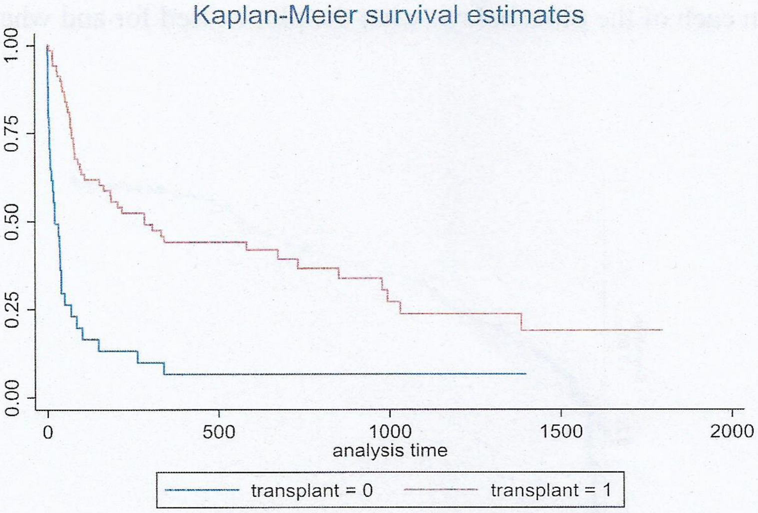
Plot B:



Plot C:


Percent
0
10
20
30
40
-1
0
1
2
3
4
Studentized residuals

Plot D:


Kaplan-Meier survival estimates
0.00
0.25
0.50
0.75
1.00
0
500
1000
1500
2000
analysis time
transplant = 0
transplant = 1

SECTION B

Answer ONLY 3 questions: 20 marks each

Question 7: 20 marks

a) Research can be conducted using primary data or secondary data. State **two national surveys** that are normally considered for secondary data analysis by researchers. [2]

b) Though secondary data analysis is a cheap and readily available source of data. Discuss the data quality issues one should take into account before using secondary data. [10]

An MPH student at Africa University set to determine the effect of two drugs for the treatment of peptic ulcers. Previous studies reported that the percentage of ulcers healed by the pirenzepine drug was 55% while the trithiozone drug healed 35% of ulcers.

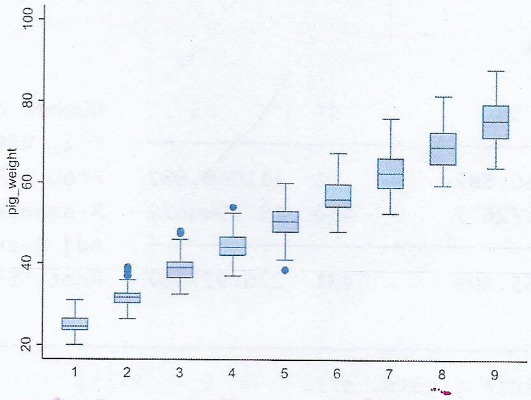
c) How many participants would be required for a randomised trial to have 80% power of detecting this difference between the two drugs (55% and 35%) at the 0.05 two-tailed level of significance? [5]

d) How many participants would be required to increase the power to 90%? [3]

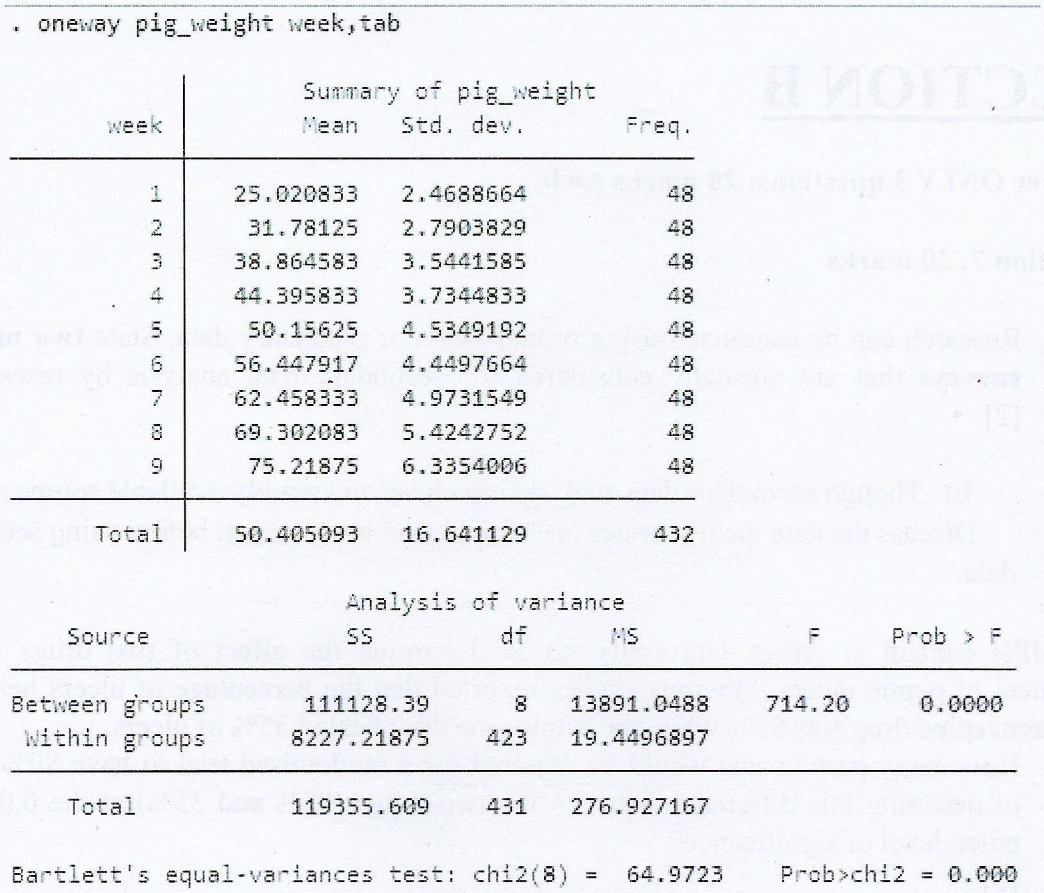
Question 8: 20 marks

An animal scientist set to assess the weight gain of the pig project over 9 weeks while feeding them on a new animal feed product X.

a) Comment on the descriptive results in box and whisker chart below [2]



b) For the next analysis, the scientist performed the ANOVA test to see differences between weeks and the results are shown below.



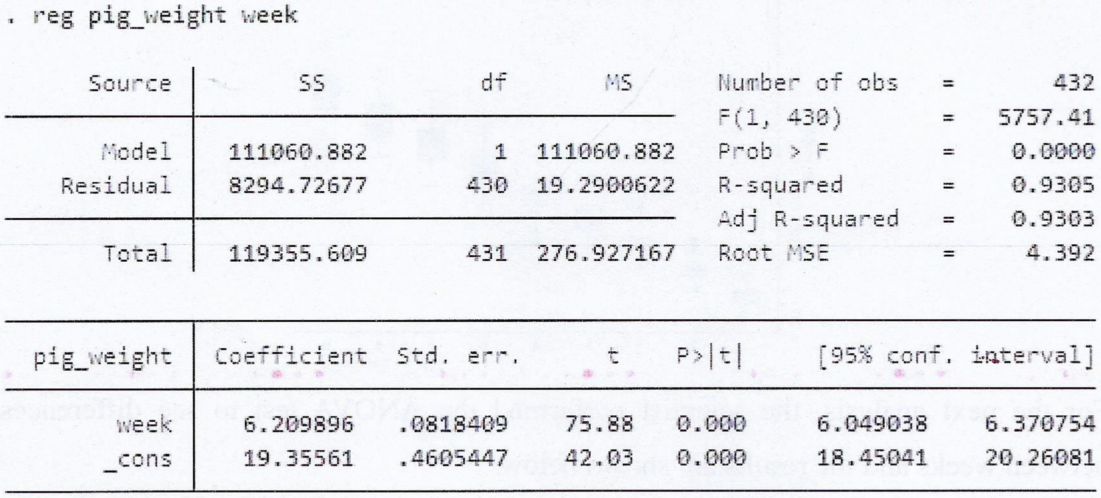
```
. oneway pig_weight week,tab

            |          Summary of pig_weight
       week |        Mean   Std. dev.       Freq.
------------+------------------------------------
          1 |   25.020833   2.4688664          48
          2 |    31.78125   2.7903829          48
          3 |   38.864583   3.5441585          48
          4 |   44.395833   3.7344833          48
          5 |    50.15625   4.5349192          48
          6 |   56.447917   4.4497664          48
          7 |   62.458333   4.9731549          48
          8 |   69.302083   5.4242752          48
          9 |    75.21875   6.3354006          48
------------+------------------------------------
      Total |   50.405093   16.641129         432

                        Analysis of variance
    Source              SS         df      MS            F     Prob > F
------------------------------------------------------------------------
Between groups      111128.39      8   13891.0488    714.20     0.0000
 Within groups     8227.21875    423   19.4496897
------------------------------------------------------------------------
    Total          119355.609    431   276.927167

Bartlett's equal-variances test: chi2(8) =  64.9723    Prob>chi2 = 0.000
```

i. Comment on the general pattern of the mean pig weight as shown on the results output. [1]

ii. Was there a significant mean difference in the pig means between weeks? Show all the hypothesis testing steps [5]

iii. One of the assumptions for ANOVA is the homogeneity of variance. Assess if the assumption was valid or not valid as per the Bartlett test's results [2]

c) The scientist fitted the linear regression model and the results are shown below:



```
. reg pig_weight week

      Source |       SS           df       MS      Number of obs   =       432
-------------+----------------------------------   F(1, 430)       =   5757.41
       Model |  111060.882         1  111060.882   Prob > F        =    0.0000
    Residual |  8294.72677       430  19.2900622   R-squared       =    0.9305
-------------+----------------------------------   Adj R-squared   =    0.9303
       Total |  119355.609       431  276.927167   Root MSE        =     4.392

------------------------------------------------------------------------------
  pig_weight | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        week |   6.209896   .0818409    75.88   0.000     6.049038    6.370754
       _cons |   19.35561   .4605447    42.03   0.000     18.45041    20.26081
------------------------------------------------------------------------------
```

i. Comment on the variability of time (weeks) in explaining the pig weight [2]

ii. Interpret the simple linear regression model intercept. [2]

iii. Interpret the simple linear regression slope [3]

iv. What is the predicted pig's weight at 7 weeks? [3]

QUESTION 9: 20 marks

a) A field epidemiologist was informed of a typhoid outbreak in Chiredzi. The researcher collected data from the clinic and developed a research question to identify the source of typhoid in Chiredzi. A sample size of 140 patients were interviewed about their source of water for drinking. The summary data is shown in the following 2x2 table:

	Typhoid cases		
Open well	**Yes**	**No**	**Totals**
Yes	a	30	b
No	25	c	45
Totals	d	50	140

i. Fill in the missing values in the results table [2]

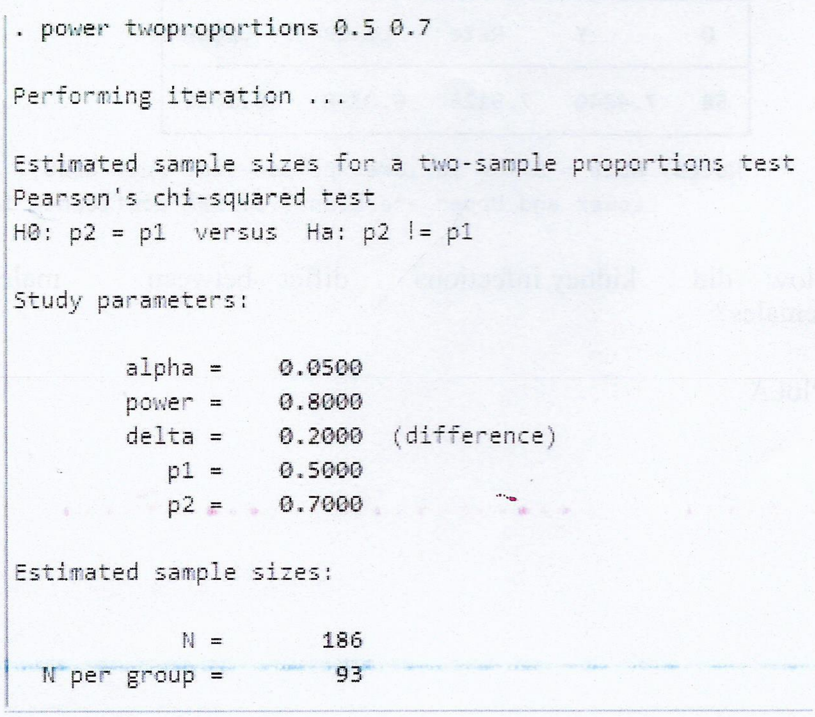
ii. State the appropriate measure of association for this study [1]

iii. Calculate the measure of association [2]

iv. Using a level of significance (α) of 0.05, determine if drinking water from an open source was a possible risk factor for typhoid in this community [5]

v. What recommendations can the researcher give based on the results obtained? [2]

The researcher calculated the sample size using Stata and obtained the following output

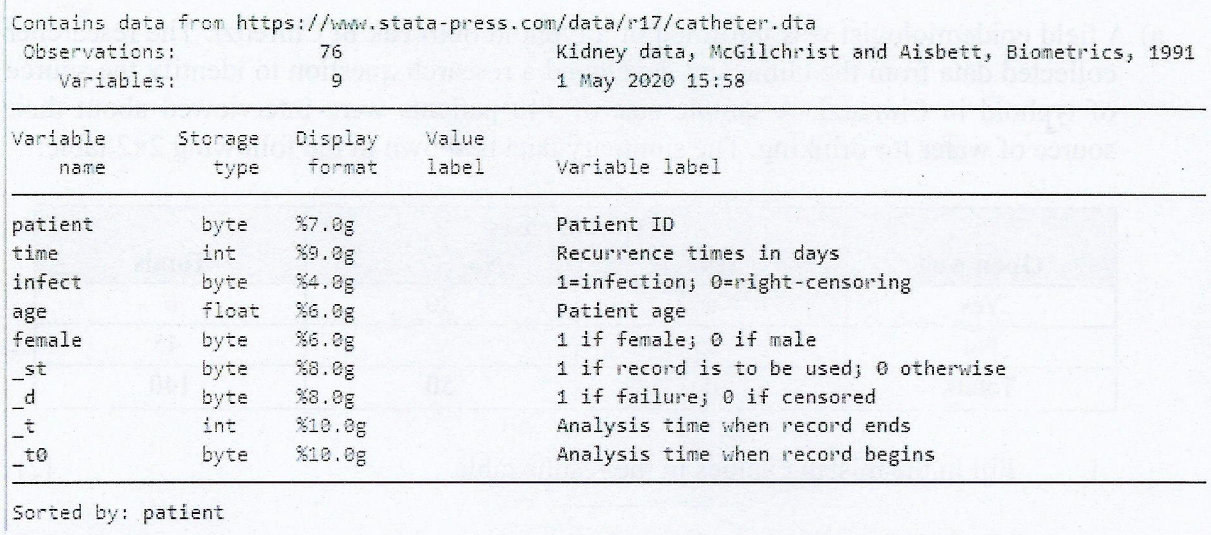


```
. power twoproportions 0.5 0.7

Performing iteration ...

Estimated sample sizes for a two-sample proportions test
Pearson's chi-squared test
H0: p2 = p1  versus  Ha: p2 != p1

Study parameters:

        alpha =    0.0500
        power =    0.8000
        delta =    0.2000  (difference)
           p1 =    0.5000
           p2 =    0.7000

Estimated sample sizes:

            N =       186
  N per group =        93
```

vi. Recalculate the sample size manually [8]

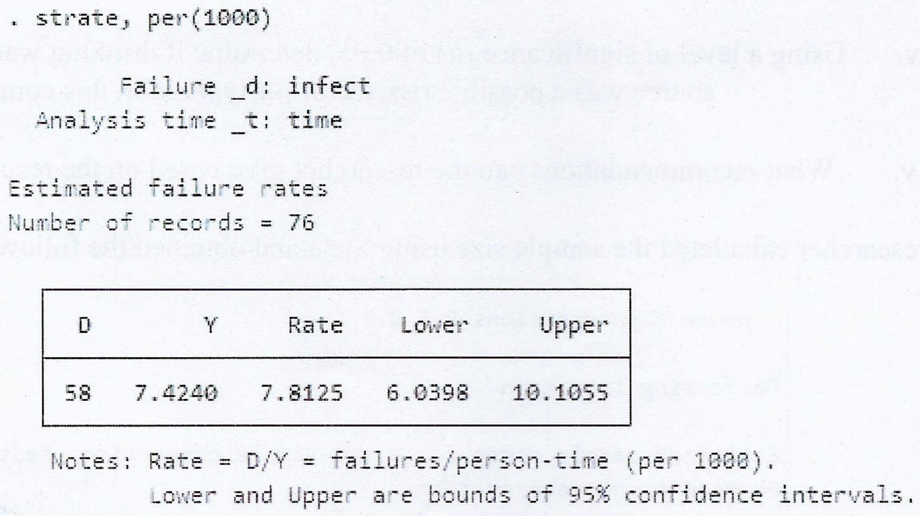
QUESTION 10: 20 marks

To compare the rate of kidney infection among patients, the following data was collected.



```
Contains data from https://www.stata-press.com/data/r17/catheter.dta
 Observations:            76              Kidney data, McGilchrist and Aisbett, Biometrics, 1991
    Variables:             9              1 May 2020 15:58
--------------------------------------------------------------------------------------------------
Variable      Storage   Display    Value
    name         type    format    label      Variable label
--------------------------------------------------------------------------------------------------
patient         byte    %7.0g                 Patient ID
time            int     %9.0g                 Recurrence times in days
infect          byte    %4.0g                 1=infection; 0=right-censoring
age             float   %6.0g                 Patient age
female          byte    %6.0g                 1 if female; 0 if male
_st             byte    %8.0g                 1 if record is to be used; 0 otherwise
_d              byte    %8.0g                 1 if failure; 0 if censored
_t              int     %10.0g                Analysis time when record ends
_t0             byte    %10.0g                Analysis time when record begins
--------------------------------------------------------------------------------------------------
Sorted by: patient
```

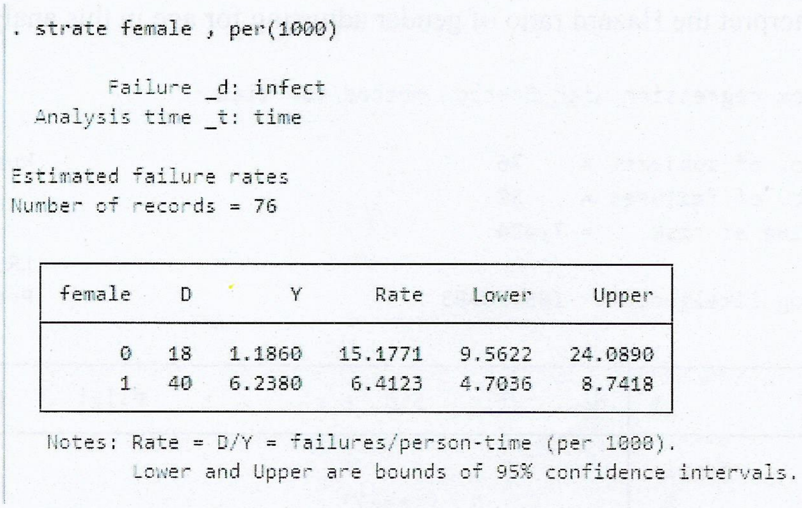
i) What was the rate of kidney infection in this study? [2]



```
. strate, per(1000)

        Failure _d: infect
  Analysis time _t: time

Estimated failure rates
Number of records = 76

+---------------------------------------------------+
|  D        Y     Rate     Lower      Upper         |
|---------------------------------------------------|
| 58   7.4240   7.8125    6.0398    10.1055         |
+---------------------------------------------------+

   Notes: Rate = D/Y = failures/person-time (per 1000).
          Lower and Upper are bounds of 95% confidence intervals.
```

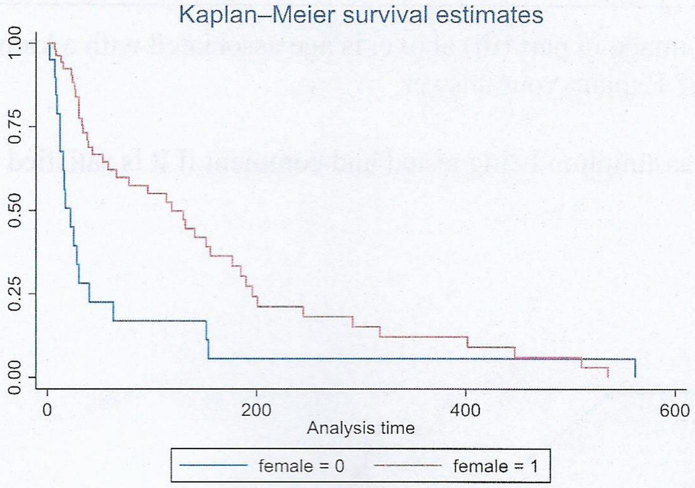
ii) How did kidney infections differ between males and females? [3+3+4]

Plot A

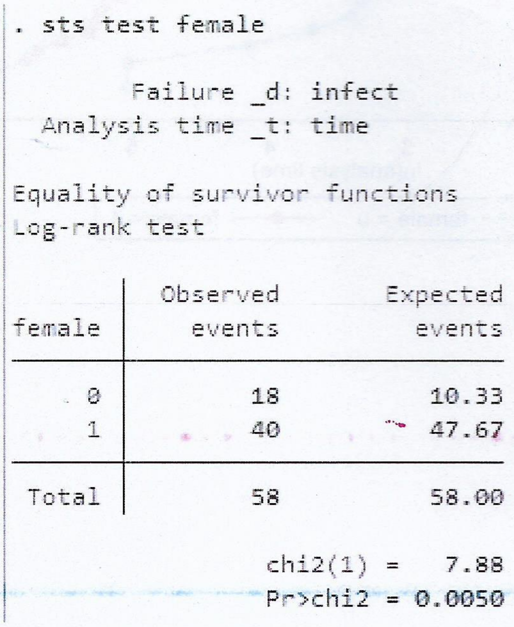


```
. strate female , per(1000)

        Failure _d: infect
  Analysis time _t: time

Estimated failure rates
Number of records = 76

  +------------------------------------------------------+
  | female    D         Y       Rate     Lower     Upper |
  |------------------------------------------------------|
  |      0   18    1.1860    15.1771    9.5622   24.0890 |
  |      1   40    6.2380     6.4123    4.7036    8.7418 |
  +------------------------------------------------------+

   Notes: Rate = D/Y = failures/person-time (per 1000).
          Lower and Upper are bounds of 95% confidence intervals.
```

Plot B

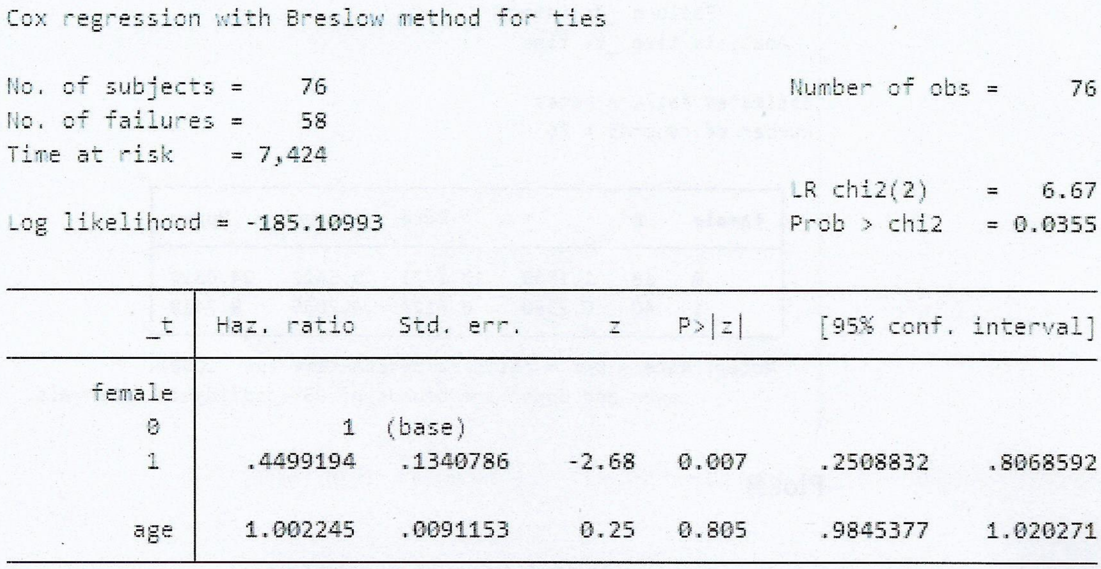


Plot C



```
. sts test female

        Failure _d: infect
  Analysis time _t: time

Equality of survivor functions
Log-rank test

       |  Observed       Expected
female |    events         events
-------+-------------------------
     0 |        18          10.33
     1 |        40          47.67
-------+-------------------------
 Total |        58          58.00

               chi2(1) =    7.88
               Pr>chi2 = 0.0050
```

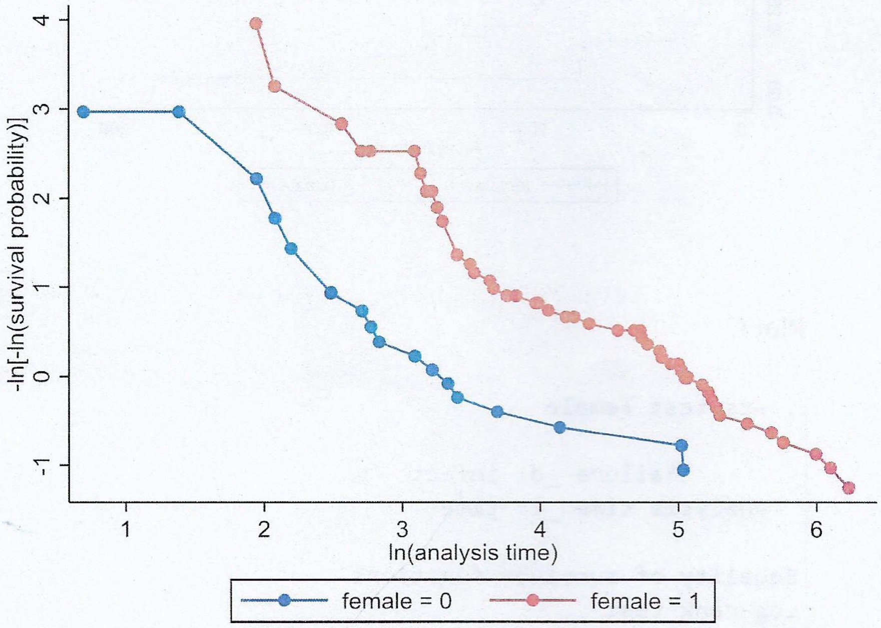
iii) Interpret the Hazard ratio of gender adjusting for age in this analysis [3]



```
Cox regression with Breslow method for ties

No. of subjects =     76                              Number of obs =     76
No. of failures =     58
Time at risk    = 7,424
                                                      LR chi2(2)    =   6.67
Log likelihood = -185.10993                           Prob > chi2   = 0.0355

------------------------------------------------------------------------------
          _t | Haz. ratio   Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
      female |
           0 |          1  (base)
           1 |   .4499194   .1340786    -2.68   0.007     .2508832    .8068592
             |
         age |   1.002245   .0091153     0.25   0.805     .9845377    1.020271
------------------------------------------------------------------------------
```

iv) From the image in part (iii) above, is age associated with a kidney infection in this study? Explain your answer [3]

v) State the assumption being tested and comment if it is satisfied or not [2]


-ln[-ln(survival probability)]
ln(analysis time)
female = 0
female = 1