NAMIBIA UNIVERSITY
OF SCIENCE AND TECHNOLOGY

FACULTY OF HEALTH AND APPLIED SCIENCES

DEPARTMENT OF MATHEMATICS AND STATISTICS

QUALIFICATION: Bachelor of Science Honours in Applied Statistics	
QUALIFICATION CODE: O8BSSH	**LEVEL:** 8
COURSE CODE: BIO 801S	**COURSE NAME:** BIOSTATISTICS
SESSION: JULY 2019	**PAPER:** THEORY
DURATION: 3 HOURS	**MARKS:** 100

SUPPLEMENTARY/SECOND OPPORTUNITY EXAMINATION QUESTION PAPER	
EXAMINER	Dr D. NTIRAMPEBA
MODERATOR:	Dr L. PAZVAKAWAMBWA

INSTRUCTIONS
1. Answer ALL the questions in the booklet provided. 2. Show clearly all the steps used in the calculations. 3. All written work must be done in blue or black ink and sketches must be done in pencil.

PERMISSIBLE MATERIALS

1. Non-programmable calculator without a cover.

ATTACHMENTS

1. None

THIS QUESTION PAPER CONSISTS OF 7 PAGES (Including this front page)

Question 1 [25 marks]

1.1 Briefly explain the following terminologies as they are applied to biostatistics.
(a) Censored observation [3]
(b) Survival function [3]
(c) Hazard function [3]

1.2 Graham et al. (1981) study dietary factors in the epidemiology of cancer of the larynx. Interviews were carried out with 338 male patients at Roswell Park Memorial Institute with cancer of the larynx, and with 359 male controls with diseases other than the digestive or respiratory system (and without neoplasme). Table 1 compare vitamin A (IU/month) intake for the cases and the controls:
(a) What are appropriate null and alternative hypotheses for testing association between vitamin A intake and cancer. [2]
(b) Compute and interpret the relative risk (RR) of cancer. [4]

Table 1: Comparison of Vitamin A(IU/month) intake between the Cases and the controls.

Vitamin A	Cases	controls	Total
< 50500	98	78	176
≥ 50500	240	281	521
Total	338	359	697

1.3 In survival analysis, if one wishes to estimate the **percentage of individuals who survive to fixed time or beyond**, two methods, namely the Life table and the Kaplan-Meier method, are commonly used. Briefly discuss these two methods. [10]

Question 2 [25 marks]

2.1 Consider a logistic regression model defined as follows. $logit\,[\pi(X)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2$,1 where $X_1 = 0$ or 1 and $X_2 = 0$ or 1. Find the odds ratio comparing $(X_1 = 1, X_2 = 1)$ to $(X_1 = 0, X_2 = 0)$. [3]

2.2 Consider a single random variable Y whose probability distribution depends on a single parameter θ. The distribution of Y belongs to the **exponential family** if it can be written in the form

$$f(y, \theta) = exp[a(y)b(\theta) + c(\theta) + d(y)],$$

where a, b, c and d are known functions.
Show that

$$Var[a(y)] = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}$$

[13]

2.3 If the random variable Y has the Gamma distribution with a scale parameter θ, which is the parameter of interest, and a known shape parameter φ, then its probability density function is

$$f(y,\theta) = \frac{y^{\varphi-1}\theta^{\varphi}e^{-y\theta}}{\Gamma(\varphi)}.$$

(a) Find variance of Y. [7]
(b) Find the information $\mathcal{I}$ [2]

Question 3 [20 marks]

3.1 The state wildlife biologists want to model how many fish are being caught by fishermen at a state park. Visitors in 250 groups that went to a park were asked whether or not they did have a camper (**camper**), how many people were in the group (**persons**), were there children in the group (**child**) and how many fish were caught (**count**). Some visitors do not fish, but there is no data on whether a person fished or not. Some visitors who did fish did not catch any fish so there are excess zeros in the data because of the people that did not fish. In addition to predicting the number of fish caught, there is interest in predicting the existence of excess zeros, i.e. the zeroes that were not simply a result of bad luck fishing. The variables child, persons, and camper were employed to model counts of fish. A histogram showing the distribution of the variable count is given below.

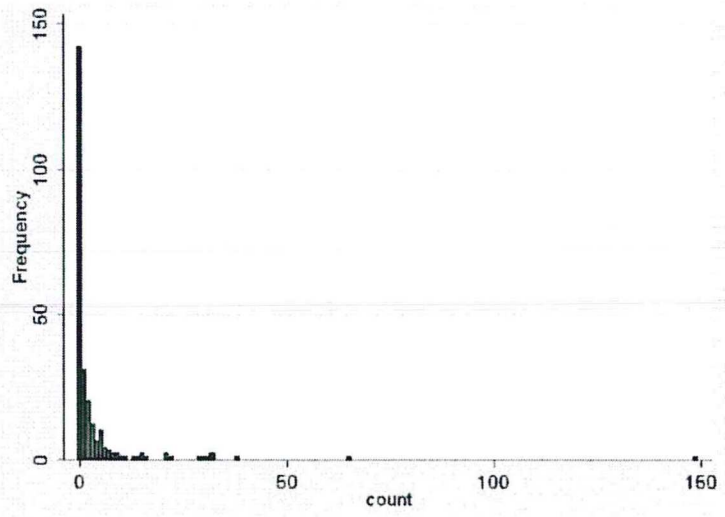

Figure 1: Histogram of number fishes caught

(a) Advise the state wildlife biologists which model is the best among the two fitted models.(Provide reasons) [5]
(b) Find and interpret the rate ratio associated with the variable **child**(in Table 2). [5]

Table 2: Summary of the results of the Poisson model

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-1.98183	0.152263	-13.0158	9.94E-39
child	-1.68996	0.080992	-20.8658	1.09E-96
camper	0.930936	0.089087	10.44979	1.47E-25
persons	1.091262	0.039255	27.79918	4.44E-170
AIC	1682.1			
Overdispersion test:				
alpha	1.81554		2.239	1.26E-02

Table 3: Summary of the results of the Negative binomial model

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-1.62499	0.330416	-4.91801	8.74E-07
child	-1.78052	0.185036	-9.62254	6.42E-22
camper	0.621129	0.2348	2.645353	0.008161
persons	1.0608	0.114401	9.272618	1.82E-20
theta	0.4635			
AIC	820.44			
2 x log-likelihood:	810.44			

3.2 The Indonesia Children Health Study collected information on respiratory infection of 250 Indonesian children (as reported by Sommer, Katz, and Tarwotjo (1984)). The children, all preschoolers, were seen quarterly for up to six quarters (Time). At each examination, the presence (Response=1)or absence of respiratory infection was noted (Response=0). Also information gender (Gender=0 for female and Gender=1 for male), age (in months), and whether a child has deficiency (Vita=1) or does not have a deficiency in Vitamin A (Vita=0). Two models were fitted and their outputs are given below.
(a) Which model among two fitted models is more suitable to these data
(justify your choice). [3]
(b) Which other modeling type would you recommend for these data? [2]
(c) Use the output of model 2 to compute and interpret the odds ratio associated variable **vitamin A**. [5]

```
Model 1:
Call:
glm(formula = RESPONSE ~ TIME + factor(GENDER) + factor(AGE) +factor(VITA)

Deviance Residuals:
Min       1Q       Median    3Q       Max
-1.2452  -0.8848  -0.7232   1.3308   2.1344
```

Table 4: Summary of the binary logistic regression model

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-0.83874	0.171246	-4.8978467	9.69E-07
TIME	0.017145	0.011283	1.51956204	0.12862109
GENDER	-0.57785	0.116083	-4.9779166	6.43E-07
factor(AGE)2	0.198831	0.194019	1.02480073	0.30545722
factor(AGE)3	0.474642	0.200433	2.3680857	0.01788039
factor(AGE)4	-0.14461	0.207698	-0.6962548	0.48626929
factor(AGE)5	0.014511	0.203066	0.07146126	0.94303066
factor(AGE)6	0.105927	0.198663	0.53319821	0.59389638
factor(AGE)7	-1.01027	0.308916	-3.2703576	0.00107412
VITA	0.264827	0.119979	2.20728675	0.02729403

```
Model 2:
 Call:

gee(formula = RESPONSE ~ TIME + factor(GENDER) + factor(AGE) +
    factor(VITA), id = ID, data = data, family = binomial(logit),
    corstr = "AR-M")

Summary of Residuals:
Min          1Q          Median      3Q          Max
-0.5194603 -0.3170443 -0.2321960  0.5861312  0.8966339
```

Table 5: Summary of the binary logistic regression model using gee

	Estimate	Naive S.E.	Naive z	Robust S.E.	Robust z
(Intercept)	-0.87544	0.246343	-3.55373	0.30355538	-2.88395
TIME	0.018008	0.012916	1.394285	0.00905063	1.989719
GENDER	-0.54405	0.175883	-3.09323	0.21694006	-2.50782
factor(AGE)2	0.304589	0.292688	1.040658	0.38338773	0.794466
factor(AGE)3	0.425875	0.305495	1.394047	0.3917259	1.087176
factor(AGE)4	-0.11286	0.314642	-0.35871	0.38066488	-0.29649
factor(AGE)5	0.005532	0.309142	0.017895	0.397365	0.013922
factor(AGE)6	0.115483	0.301842	0.382595	0.37628257	0.306905
factor(AGE)7	-0.99821	0.469195	-2.12748	0.50866707	-1.96239
VITA	0.25732	0.181852	1.414999	0.22572731	1.13996

Table 6: Correlation matrix: autoregressive model order 1

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	1	0.513062	0.263232	0.135055	0.069291	0.035551
[2,]	0.513062	1	0.513062	0.263232	0.135054	0.069291
[3,]	0.263232	0.513062	1	0.513062	0.263232	0.135054
[4,]	0.135054	0.263232	0.513062	1	0.513062	0.263232
[5,]	0.069291	0.135054	0.263232	0.513062	1	0.513062
[6,]	0.035551	0.069291	0.135055	0.263232	0.513062	1

Question 4 [30 marks]

4.1 The survival times (in months from diagnosis of AIDS to death from AIDS or to the end of the study participation) of 23 African-American male participants in San Francisco Men's Health Study (SFMHS) were analysis. The graph below provides a visual comparison between survival experiences of nonsmokers and smokers. Also, the following statistics were obtained. $Chisq = 4.7$ on 1 degree of freedom, $p - value = 0.0299$
Use appropriate test to compare the survival experiences of the two groups at 5% significance level [3]

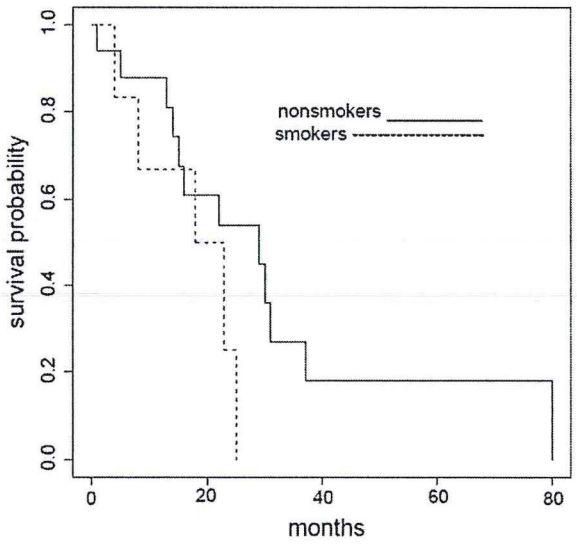

Figure 2: Plots of survival times for SFMHS African-Americans, comparing nonsmokers and smokers.

4.2 As part of clinical trial to evaluate the efficacy of maintenance chemotherapy for sufferers of myelogenous leukemia, patients were randomly assigned to two groups. First group received maintenance chemotherapy and control group did not. The aim of the study was to see if maintenance chemotherapy prolonged the time till relapse. The length of remission for each group are given table below

Table 7: Summary of the length of remission in weeks maintained group and non-maintained group.

Maintained group	161	45^+	18	48	9	13^+	28^+	34	13	31	23	
Non-maintened group	5	5	8	8	12	16^+	23	27	30	33	43	45

Note that the + means that the time till relapse was greater than the reported.

Construct the **Kaplan Meier table for non-maintained group.** [10]

4.3 Let the random variable Y denote the survival time and let $f(y)$ denote its probability density function.
(a) Show that the equation of the hazard function is $h(y) = \frac{f(y)}{s(y)}$,
where $s(y) = P(Y \geq y)$. [7]

(b) Use the equation of the hazard function given in part (a) to show that if Y follows a **Weibull** distribution which has the probability density function

$$f(y, \lambda, \theta) = \frac{\lambda y^{\lambda-1}}{\theta^\lambda} exp\left(-\left[\frac{y}{\theta}\right]^\lambda\right), y \geq 0, \lambda > 0, \theta > 0,$$

where λ and θ determine the shape of the distribution and the scale, respectively, then $h(y) = \lambda\phi y^{\lambda-1}$. (hint:$\theta^{-\lambda} = \phi$) [10]

END OF QUESTION PAPER