NAMIBIA UNIVERSITY
OF SCIENCE AND TECHNOLOGY

FACULTY OF HEALTH, APPLIED SCIENCES AND NATURAL RESOURCES

DEPARTMENT OF MATHEMATICS AND STATISTICS

QUALIFICATION: Bachelor of Science Honours in Applied Statistics	
QUALIFICATION CODE: 08BSHS	**LEVEL:** 8
COURSE CODE: BIO801S	**COURSE NAME:** BIOSTATISTICS
SESSION: JULY 2022	**PAPER:** THEORY
DURATION: 3 HOURS	**MARKS:** 100

SUPPLEMENTARY / SECOND OPPORTUNITY EXAMINATION QUESTION PAPER	
EXAMINER	Dr D. B. GEMECHU
MODERATOR:	Prof L. PAZVAKAWAMBWA

INSTRUCTIONS

1. There are 6 questions, answer ALL the questions by showing all the necessary steps.
2. Write clearly and neatly.
3. Number the answers clearly.
4. Round your answers to at least four decimal places, if applicable.

PERMISSIBLE MATERIALS

1. Non-programmable scientific calculator

THIS QUESTION PAPER CONSISTS OF 9 PAGES (Including this front page)

Question 1 [23 marks]

1.1 Briefly discuss the following study designs (your answer should include definition/uses, advantage and disadvantages).

1.1.1 Ecologic studies [3]

1.1.2 Prospective Cohort study [3]

1.2 Briefly explain the following terminologies as they are applied to Biostatistics.

1.2.1 Right-censored observation [2]

1.2.2 Survival function [2]

1.2.3 Hazard function [2]

1.2.4 Nominal logistic regression. Your explanation should include the model, the type response variable and based on the model stated, show how to compute the predicted probability for the reference category. Assume that there are J categories of the response variable and the first category is the reference category. [6]

1.3 An investigator conducts a study to determine whether there is an association between caffeine intake and Parkinson's disease. He assembles 230 incident cases of PD and samples 455 controls from the general population. After interviewing all subjects, he finds that 64 of the cases had high daily intake of caffeine (exposed) prior to diagnosis and 277 of the controls had low daily intake of caffeine (unexposed) prior to the date of the matched case's diagnosis. The summary of this study is given in table below

	Cases	Control	Total
Exposed	64	178	242
Unexposed	166	277	443
Total	230	455	685

1.3.1 Calculate the odds of being a case among the exposed [2]

1.3.2 Calculate the odds ratio for disease given exposure to high daily intake of caffeine (versus low daily intake of caffeine). [2]

1.3.3 What does the odds ratio indicate? [1]

Question 2 [13 marks]

2.1 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

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

2.1.1 Show that this distribution belongs to the exponential family and find the natural parameter. [4]

2.1.2 Find variance of y. [4]

2.2 Suppose a random sample $y_1, y_2, ..., y_n$ of size n were selected from a Pareto distribution with a parameter θ. The probability density function of y_i is given by

$$f(y_i, \theta) = \theta y_i^{-\theta-1}.$$

Derive the Newton-Raphson approximation estimating equation that will be used obtain the maximum likelihood estimator of θ. [5]

Question 3 [16 marks]

3.1 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]

3.2. Sudden death is an important, lethal cardiovascular endpoint. Most previous studies of risk factors for sudden death have focused on men. Looking at this issue for women is important as well. For this purpose, data were used from the Framingham Heart Study. Several potential risk factors, such as age, blood pressure and cigarette smoking are of interest and need to be controlled for smilutaneously. Therefore a multiple logistic regression was fitted to these data as shown in Table 1. The response is 2-year incidence of sudden death in females without prior coronary heart disease.

Table 1: Model summary for sudden death

Risk Factor	Regression Coefficient (bj)	Standard Error (se(bj))	p-value
Constant	-15.3		
Blood Pressure (mm Hg)	.0019	.0070	.7871
Weight (% of study mean)	-.0060	.0100	.5485
Cholesterol (mg/100 mL)	.0056	.0029	.0536
Glucose (mg/100 mL)	.0066	.0038	.0819
Smoking (cigarettes/day)	.0069	.0199	.7623
Hematocrit (%)	.111	.049	.0235
Vital capacity (centiliters)	-.0098	.0036	.0065
Age (years)	.0686	.0225	.0023

3.2.1 Assess the statistical significance of the individual risk factors. [2]

3.2.2 Give brief interpretations of the age and vital capacity coefficients. [2]

3.2.3 Compute and interpret the odds ratios relating the additional risk of sudden death associated with an increase in consumption of cigarettes by 4 (cigarettes/day) after adjusting for the other risk factors. [2]

3.2.4 Compute and interpret a 95% confidence intervals for the odds ratios relating the additional risk of sudden death associated with an additional year of age after adjusting for the other risk factors. [4]

3.2.5 Predict the probability of sudden death for a 60 year old woman with systolic blood pressure of 110 mmHg, a relative weight of 90% a cholesterol level of 250 mg/100mL, a glucose level of 90 mg/100mL, a hematocrit of 35%, and a vital capacity of 450 centiliters who smokes 10 cigarettes per day. [3]

Question 4 [13 marks]

4. . A researcher conducted a follow-up study of larynx cancer on a group of patients. Refer to the software output provided in the following tables to answer the questions
Variable information:
Stage34: Stage of disease (0=stage 1 or 2 1=stage 3 or 4)
Time: Time to death or on-study time, months
Age50: (Age at diagnosis of larynx cancer-50)
Status: Death indicator (0=alive, 1=dead)

Table 2: Summary of the Cox-Proportial hazards Model 1

	coef	se(coef)	z value	$\Pr(>\lvert z\rvert)$	95% CI
stage34	0.879474	0.286939	3.07	0.002	(0.3170838, 1.441864)
Log likelihood		-192.49913			

Table 3: Summary of the Cox-Proportial hazards Model 2

	coef	se(coef)	z value	$\Pr(>\lvert z\rvert)$	95% CI
stage34	0.8735205	0.2871044	3.04	0.002	(0.3108062, 1.436235)
age50	0.0226812	0.0145471	1.56	0.119	(-0.0058305, 0.051193)
Log likelihood		-191.26058			

Table 4: Summary of the Cox-Proportial hazards Model 3

	coef	se(coef)	z value	$\Pr(>\lvert z\rvert)$	95% CI
Istage341	1.087132	.5725228	1.90	0.058	(-0.0349917, 2.209256)
age50	0.0297464	0.0219454	1.36	0.175	(-0.0132658, 0.0727587)
IstaXage5 1	-0.0127367	0.0293888	-0.43	0.665	(-0.0703378, 0.0448644)
Log likelihood		-191.16652			

4.1 What is the interpretation of the regression coefficient in "Model 1"? Compute and interpret the hazard ratio. Is the effect statistically significant at the 5% level? [4]

4.2 Is there evidence that age confounds the effect of stage34? Justify your response. [2]

4.3 What is the interpretation of the coefficient of age50 in Model 3? [2]

4.4 For patients with stage 3 or 4 cancer, if age increases from 55 to 65, by what multiplicative factor does the fitted Model 3 estimate that their death rate increases? [3]

4.5 Is there evidence that the hazard ratio for stage34 varies by age?

Question 5 [14 marks]

5. A small clinical trial was run to compare two combination treatments in patients with advanced gastric cancer. Twenty participants with stage IV gastric cancer who consent to participate in the trial were randomly assigned to receive chemotherapy before surgery or chemotherapy after surgery. The primary outcome is death and participants were followed for up to 48 months (4 years) following enrollment into the trial. The experiences of participants in each arm of the trial are shown in Table 5.

Table 5: Summary of the experiences of participants in chemotherapy before surgery and chemotherapy after surgery group

Chemotherapy Before Surgery		Chemotherapy After surgery	
Month of Death	Month of Last Contact	Month of Death	Month of Last Contact
8	8	33	48
12	32	28	48
26	20	41	25
14	40		37
21			48
27			25
43			

5.1 Construct life tables for each treatment group using the Kaplan-Meier approach. [6]

5.2 Use Fig.1 to answer the following questions:

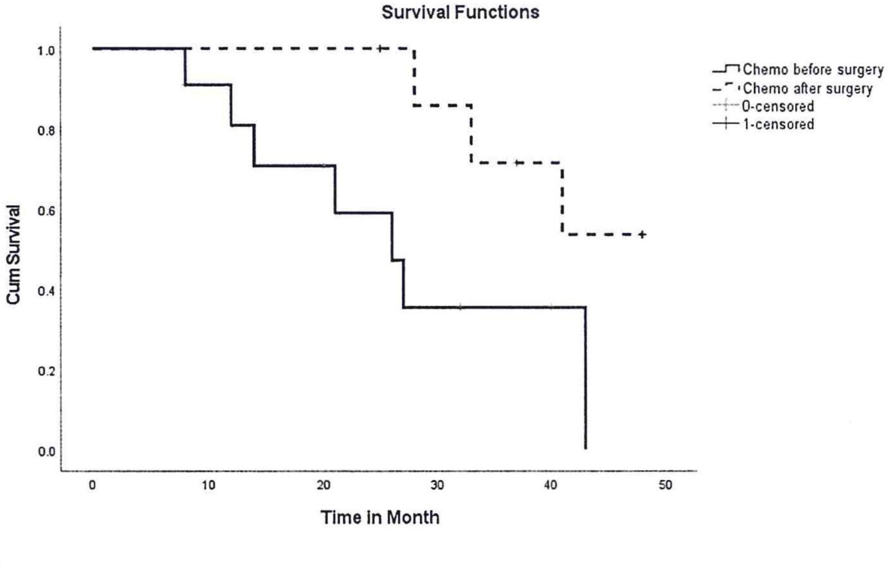

Figure 1: Kaplan-Meier survival curves for Chemotherapy Before Surgery and Chemotherapy after Surgery groups.

5.2.1 Briefly comment on the survival curve. Are the median survival times for the two treatment group the same? (Provide the approximated values for the two medians) [3]

5.2.2 Compare survival between groups using using appropriate test to test, at 5% significance level. Your solution should include the following: state the null and alternative hypothesis; determine the critical value and rejection region; compute the test statistics; write your decision and conclusion based on your result. *Hint: The expected number of deaths in chemotherapy before surgery group and chemotherapy after surgery group were 2.62 and 6.38, respectively.*$\chi^2_{0.05}(1) = 3.8414$ [5]

Question 6 [21 marks]

6. 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. The following are some of descriptive analysis results of the data.

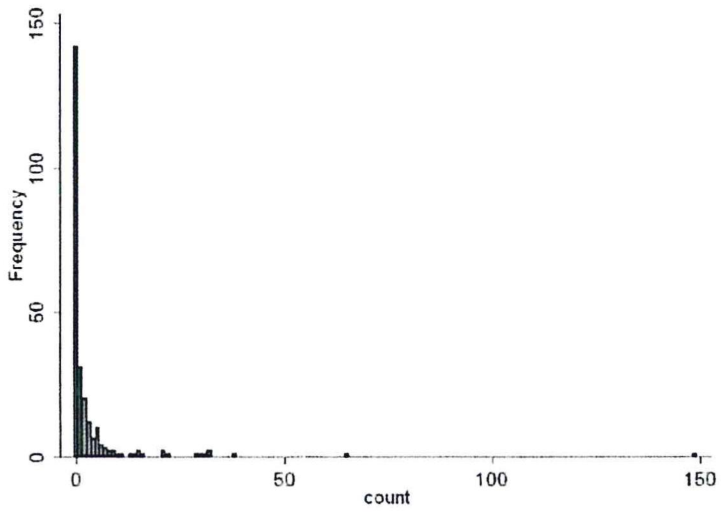

Figure 2: Histogram of number fishes caught

6.1 Use the above descriptive statistics to advise the state wildlife biologists which type of models might be appropriate (state reason(s)). [3]

6.2 Irrespective of your advice, the state wildlife biologists went on fitting the Poisson and negative binomial models. Below is the summary of these fitted models.

6.2.1 Give the assumptions of a Poisson regression model. [2]

Table 6: Some descriptive statistics of explanatory variables used in the study.

child	frequency	Percent	persons	frequency	Percent	camper	frequency	Percent
0	132	52.8	0	57	22.8	0	103	41.2
1	75	30	1	70	28	1	147	58.8
2	33	13.2	2	57	22.8	Tot	250	100
3	10	4	3	66	26.4			
Tot	250	100	Tot	250	100			

Table 7: 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

6.2.2 Use the output provided in the Table 7 or Table 8 to test the overdispersion (Provide the statements of the null and alternative hypotheses). [3]

6.3 The state wildlife biologists went on fitting other four models. The summaries of these models are provided below.

6.3.1 The state wildlife biologists chose model 2 (Table 10) instead model 1 (Table 9). Is their choice justified? (hint use model 2 to justify your answer) [2]

6.3.2 Compute AICs values for the four models (models 1, 2, 3, and 4) and use the obtained values to choose best model. [6]

6.3.3 Compute and interpret the rate ratio and odds ratio associated with variables "**camper**" and "**persons**" in model 1, respectively. (Table 9). [5]

== END OF QUESTION PAPER ==

Total: 100 marks

Table 8: 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			

Table 9: Summary of the results of model 1

	Count model coefficients (truncated Poisson with log link)			
	Estimate	Std.error	z value	Pr(> $\|z\|$)
intercept	1.64668	0.08278	19.892	2.00E-16
child	-0.75918	0.09004	-8.432	2.00E-16
camper	0.75166	0.09112	8.249	2.00E-16
	Zero hurdle model coefficients (binomial with logit link)			
	Estimate	Std.error	z value	Pr(> $\|z\|$)
intercept	-0.7808	0.324	-2.41	1.60E-02
Persons	0.1993	0.1161	1.716	8.62E-02
log-likelihood	-1047			
df	5			

Table 10: Summary of the results of model 2

	Count model coefficients (truncated negative binomial with log link)			
	Estimate	Std.error	z value	Pr(> $\|z\|$)
intercept	-5.8422	37.9602	-0.154	0.8777
child	-0.9122	0.4104	-2.223	0.0262
camper1	0.7861	0.4531	1.735	0.0828
log(thetha)	-8.6573	37.9728	-0.228	0.8197
	Zero hurdle model coefficients (binomial with logit link)			
	Estimate	Std.error	z value	Pr(> $\|z\|$)
intercept	-0.7808	0.324	-2.41	0.016
Persons	0.1993	0.1161	1.716	0.0862
log-likelihood	-445.5			
df	6			

Table 11: Summary of the results of model 3

	Count model coefficients (Poisson with log link)			
	Estimate	Std.error	z value	Pr(> $\|z\|$)
intercept	1.59788	0.08554	18.68	2E-16
child	-1.04286	0.09999	-10.43	2E-16
camper	0.83403	0.09336	8.908	2E-16
	Zero -inflation model coefficients (binomial with logit link)			
	Estimate	Std.error	z value	Pr(> $\|z\|$)
intercept	-1.2975	0.3739	3.471	0.000519
Persons	-0.5644	0.163	-3.463	0.000534
log-likelihood	-1032			
df	5			

Table 12: Summary of the results of model 4

	Count model coefficients (negative binomial wit logit link)			
	Estimate	Std.error	z value	Pr(> $\|z\|$)
intercept	1.371	0.2561	5.353	8.64E-08
child	-1.5153	0.1956	-7.746	9.41E-15
camper	0.8791	0.2693	3.265	0.0011
log(theha)	-0.9854	0.176	-5.6	2.1 e-8
	Zero-inflation model coefficients (binomial with logit link)			
	Estimate	Std.error	z value	Pr(> $\|z\|$)
intercept	1.6031	0.8365	1.916	0.0553
Persons	-1.6666	0.6793	-2.453	0.0142
log-likelihood	-432.5			
df	6			