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: JUNE 2019	**PAPER:** THEORY
DURATION: 3 HOURS	**MARKS:** 100

FIRST 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 8 PAGES (Including this front page)

Question 1 [30 marks]

1.1 Briefly explain the following terminologies as they are applied to biostatistics. In addition, indicate how you will test each terminology.
(a) Confounder [5]
(b) Effect modifier [3]

1.2 Use Fig. 1 to compute and interpret the incidence rate of pneumonia [3]

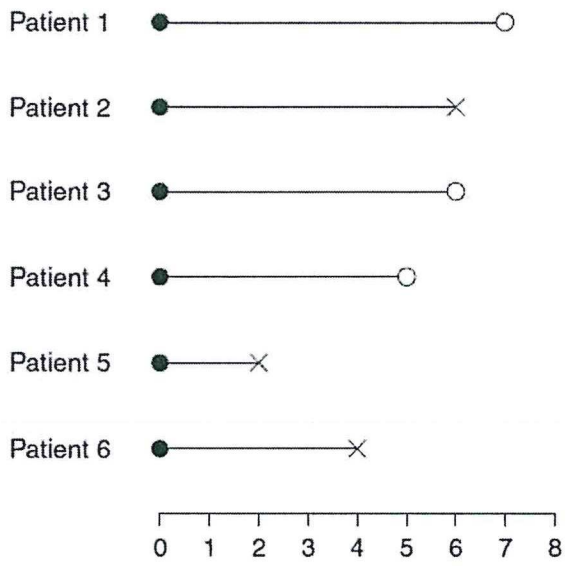

Figure 1: Diagram of individual risk time (**months**) and disease status: The Xs denote pneumonia and the open circles denote no pneumonia.

1.3 Use the information displayed in Table 1 to find and interpret:
(a) Relative risk (RR) of syphilis infection. [3]
(b) Odds ratio (OR) of syphilis infection. [3]

Table 1: Subjects from a cohort study classified to syphilis infection status and the number of sexual partners in preceding 90 days.

Sexual partners	Cases	Non-cases	Total
>= 3	41	58	99
0	10	49	59
Total	51	107	158

1.4 In statistical modelling, one of the important steps is the model checking. Provide **four** aspects of model checking when using a logistic regression model and indicate how you will determine that each aspect is met or not. [13]

Question 2 [20 marks]

2.1 Consider N independent binary random variables $Y_1, \ldots, Y_N$ with $P(Y_i = 1) = \pi_i$ and $P(Y_i = 0) = 1 - \pi_i$. The probability function of Y_i can be written as $\pi^{y_i}(1-\pi)^{1-y_i}$ where $y_i = 1$ or 0.
(a) Show that this probability function belongs to the exponential family of distributions. [5]
(b) Show that the natural parameter is $log(\frac{\pi}{1-\pi})$ [2]
(c) Show that $E(Y_i) = \pi$. [5]

(d) If the link function is $g(\pi) = log(\frac{\pi}{1-\pi}) = X^T\beta$, show that this is equivalent to modelling the probability π as $\pi = \frac{e^{X^T\beta}}{1+e^{X^T\beta}}$ [3]

2.2 Let Y be a normally distributed random variable with mean μ and variance σ^2. In addition, the parameter σ is assumed to be known and fixed. Given below is the probability function of Y

$$f(y,\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} exp[-\frac{1}{2\sigma^2}(y-\mu)^2]$$
$$= exp[y\frac{\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} - \frac{1}{2}log(2\pi\sigma^2) - \frac{y^2}{2\sigma^2}],$$

where $a(y) = y$, $b(\mu) = \frac{\mu}{\sigma^2}$, $c(\mu) = -\frac{\mu^2}{2\sigma^2} - \frac{1}{2}log(2\pi\sigma^2)$, and $d(y) = -\frac{y^2}{2\sigma^2}$
Find the score statistic $(U(y,\mu))$ and the information $(\mathcal{I})$. [5]

Question 3 [27 marks]

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.

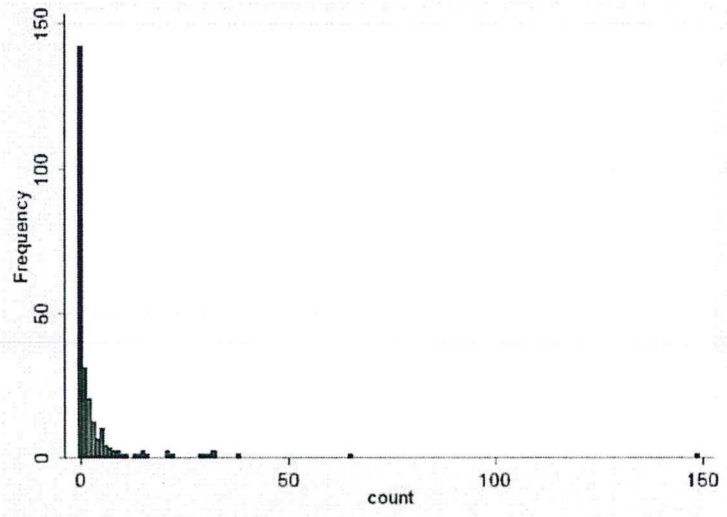

Figure 2: Histogram of number fishes caught

Table 2: Some descriptive statistics of explanatory variables used in the study.

child	frequency	Percent	pesrsons	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			

3.1 Use the above descriptive statistics to advise the state wildlife biologists which type of models might be appropriate (state reason(s)). [5]

3.2 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.

Table 3: 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 4: 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			

(a) Give the assumptions of a Poisson regression model. [3]
(b) Use the output provided in the Table 3 or Table 4 to test the overdispersion (Provide the statements of the null and alternative hypotheses). [6]

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

(a) The state wildlife biologists chose model 2 (Table 6) instead model 1 (Table 5). Is their choice justified? (hint use model 2 to justify your answer) [2]
(b) Compute AICs values for the four models (models 1, 2, 3, and 4) and use the obtained values to choose best model. [6]
(c) Compute and interpret the rate ratio and odds ratio associated with variables "**camper**" and "**persons**" in model 1, respectively. (Table 5). [5]

Table 5: Summary of the results of model 1

Count model cocefficients (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 6: Summary of the results of model 2

Count model cocefficients (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 7: Summary of the results of model 3

Count model cocefficients (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 8: Summary of the results of model 4

	Count model cocefficients (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			

Question 4 [23 marks]

4.1 Find the values representing the letters in Table 9. [5]

Table 9: Life table of 40 patients who received an experimental surgical treatment for a serious disease

Interval$[x, x+1]$	d_x	w_x	n_x	$\acute{n}_x$	q_x	p_x	survival proportion(l_x)
0-1	2	9	40	(a)	(b)	(c)	(d)
1-2	2	6	29	26.000	0.077	0.923	0.871
2-3	4	1	21	20.500	0.195	0.805	0.701
3-4	3	3	(e)	(f)	(g)	(h)	0.556
4-5	2	1	10	9.500	0.211	0.789	0.439
5-6	2	1	7	6.500	0.308	0.692	0.304
6-7	0	0	4	4.000	(j)	(k)	0.304
7-8	1	3	4	2.500	0.400	0.600	0.182

where d_x = the number of deaths reported in the interval, w_x = the number of censored observations in the interval, n_x = number of individuals alive at the start of the interval, $\acute{n}_x$ = the adjusted number of individuals at risk, q_x = the estimated probability of dying, p_x = estimated probability of surviving, and l_x = is the proportion of survivors after x time.

4.2 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 $h(y) = \frac{f(y)}{s(y)}$, where $s(y) = P(Y \geq y)$. [8]

(b) Use the equation of the hazard function given in part (a) to show that if Y follows an exponential distribution with a parameter θ then $h(y) = \theta$. [4]

4.3 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.

(a) Which modelling technique do you think is appropriate to analyse these data? (justify your answer) [3]

(b) It is known that the visual comparison does not formally account for the influence of the variation on estimated values. How else will you compare the survival expriences of the two groups? State the hypotheses. [3]

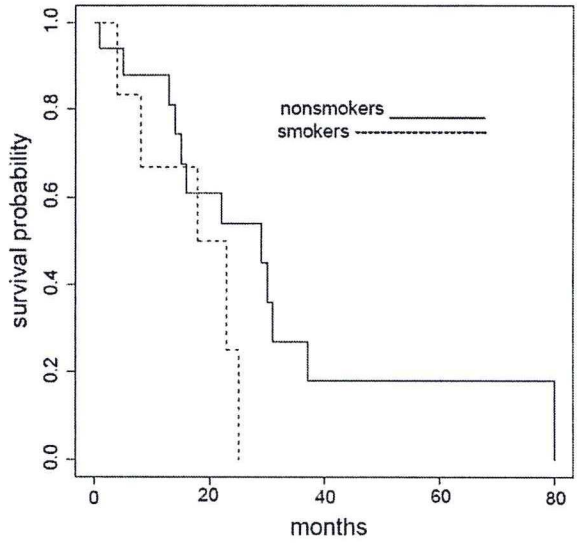

Figure 3: Plots of survival times for SFMHS African-Americans, comparing nonsmokers and smokers.

END OF QUESTION PAPER