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: JUNE 2022	**PAPER:** THEORY
DURATION: 3 HOURS	**MARKS:** 100

FIRST OPPORTUNITY EXAMINATION QUESTION PAPER	
EXAMINER	Dr D. B. GEMECHU
MODERATOR:	Prof L. PAZVAKAWAMBWA

INSTRUCTIONS

1. There are 5 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 6 PAGES (Including this front page)

Question 1 [22 marks]

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

1.1.1 Cross-sectional studies [3]

1.1.2 Case-Control Studies [3]

1.2 Wilkinson et al. (2021) studied the secondary attack rate of COVID-19 in household contacts in the Winnipeg Health Region, Canada. In their study, the authors included 381 individuals from 102 unique households (102 primary cases and 279 household contacts). A total of 41 contacts from 25 households developed COVID-19 symptom in the 14 days since last unprotected exposure to the primary case. Calculate the secondary attack rate of COVID-19. [2]

1.3 In January of 1990, 410 young adults offered to participate in a 10-year prospective study to determine their risk of Type-I diabetes. This group underwent an initial blood test to determine whether they were diabetic, and eligible subjects were re-tested yearly for the next 10 years. Assume that persons with new diagnoses of Type-I diabetes and those lost to follow-up were disease-free for quarter (1/4) of the year. Among the group that offered to join the study:

- There were 2 individuals who were found to have diabetes on the initial blood screening; these 2 people were referred for treatment and were not enrolled in the study
- 6 individuals developed diabetes during the course of the study at the times indicated in Table 1.
- 2 individuals who were initially disease-free were lost to follow-up during the study at the times indicated Table 1.

Table 1: Individual risk time and health status. **Key:** 0 = Lost to follow-up, + = Blood test positive for diabetes, —— = Continued disease-free follow-up and X =Year of death

Subject	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
1	——	——	+							
2	——	——	——	——	——	——	——	——	——	+
3	——	——	——	——	——	——	——	+	x	
4	——	+	x							
5	——	——	——	——	+					
6	——	——	+							
7	——	——	——	——	0					
8	——	——	——	——	——	0				

Answer the following questions based on information displayed in Table 1 to calculate

1.3.1 Incidence rate of Type-I diabetes. [4]

1.3.2 Point prevalence of Type-I diabetes in the year 1993. [2]

1.4 Within 10 days after attending a wedding, an outbreak of disease occurred among attendees. Of the 83 guests and wedding party members, 79 were interviewed; 54 of the 79 met the case definition. Table 2 shows consumption of wedding cake and illness status. Use the information displayed in Table 2 to answer the following questions.

Table 2: Consumption of wedding cake and illness status

Ate wedding cake?	Ill	Well	Total
Yes	50	3	53
No	4	22	26
Total	54	25	79

1.4.1 Compute Incidence proportion (food-specific attack rate) [2]

1.4.2 Calculate the relative risk of the disease. [2]

1.4.3 Calculate and interpret the attributable proportion for wedding cake. [2]

1.4.4 Calculate and interpret the odds ratio of the disease. [2]

Question 2 [30 marks]

2.1 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

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

[11]

2.2 If the random variable Y has a Weibull distribution with a parameter θ with pdf

$$f(y;\theta) = \frac{2y}{\theta^2}e^{-(y/\theta)^2}$$

2.2.1 Show that this distribution belongs to the exponential family and find the natural parameter. [3]

2.2.2 Find the score statistics U. [3]

2.2.3 Find variance of $a(y)$. [3]

2.3 Consider the N independent random variables $Y_1, Y_2, ..., Y_N$ corresponding to the numbers of successes in N different subgroups or strata. If $Y_i \sim Bin(n_i, \pi_i)$ and

$$\pi_i = \frac{exp(\beta_1 + \beta_2 x_i)}{1 + exp(\beta_1 + \beta_2 x_i)}$$

derive the information matrix. [10]

Question 3 [20 marks]

3. Table 3 provides a nominal logistic regression model for the relationship between the level of back pain during work (0=no pain, 1= mild pain and 2 =sever pain) and factors such as age categories (0=18-35 years and 1 = above 35 years) and smoking status (0 = never smoked, 1= ex-smoker and 2=current smoker) of the workers. Answer the questions based the result presented.

Table 3: Model summary for level of back pain during work

Parameter	Estimate (std. error)	Odds ratio (95% CI)
$Log(\pi_2/\pi_1)$: mild pain vs. no pain (ref)		
Intercept	-3.3128 (0.1909)	
Age (older):	0.5380 (0.1713)	1.71 (,)
Smoking status (ex-smoker):	0.7881 (0.2588)	2.20 (0.2809, 1.2954)
Smoking status (current smoker):	0.8319 (0.2140)	2.30(0.4126, 1.2513)
$Log(\pi_3/\pi_1)$: sever pain vs. no pain (ref)		
Intercept:	-5.1447 (0.4073)	
Age (older):	1.3785 (0.2855)	3.97 (0.8189, 1.9381)
Smoking status (ex-smoker):	0.8223 (0.5031)	2.28 (-0.1638, 1.8084)
Smoking status (current smoker):	1.3465 (0.4164)	3.84 (0.5304, 2.1626)
log-likelihood function for the fitted model: -791.3756 (df=8)		
log-likelihood function for the null model: -19.77502 (df=2)		

3.1 Express the fitted model using appropriate expression and describe its components. [3]

3.2 Test the overall importance of the explanatory variables using likelihood ratio test. [4]

3.3 Construct a 95% confidence limit for the odds ratio of older age in the first model. [2]

3.4 Assess the statistical significance of the individual explanatory variables. [3]

3.5 Comment on the odds ratio of the variable age. [3]

3.6 Compute the estimated probability by considering younger age worker who was never smoker. [5]

Question 4 [14 marks]

4.1 The following R-package output shows the survival analysis of cancer data which was conducted to study survival in 228 patients with lung cancer. The variables included in the model are:
time: Survival time in days
status: censoring status 1=censored, 2=dead
age: Age in years
sex: Male=1 Female=2

```
Model 1:
Call:
coxph(formula = Surv(time, status) ~ sex, data = lung)
```

Table 4: Summary of the Cox-Proportial hazards model 1

	coef	exp(coef)	se(coef)	z value	Pr(> $|z|$)
sex(Female)	-0.5310	0.5880	0.1672	-3.176	0.00149

```
log-likelihood: 'log Lik.' -744.593 (df=1)

Call:
coxph(formula = Surv(time, status) ~ sex + age, data = lung)
```

Table 5: Summary of the Cox-Proportial hazards model

	coef	Hazard Ratio	se(coef)	z value	Pr(> $|z|$)
sex(Female)	-0.513219		0.167458	-3.065	0.00218
age	0.017045		0.009223	1.848	0.06459

```
log-likelihood: 'log Lik.' -742.8482 (df=2)
```

4.1.1 Compute the hazard ratio for both predictors included in the model and interpret your result. [4]

4.1.2 Briefly comment on the $p-value$ of the variable "sex". Your discussion should include the null and alternative hypothesis. [4]

4.2 Let the random variable Y denote the survival time and let $f(y)$ denote its probability density function. Show that the equation of the hazard function is $h(y) = \frac{f(y)}{s(y)}$, where $s(y) = P(Y \geq y)$. [6]

Question 5 [14 marks]

5 Adams et al. (2020) published an article on "Modeling COVID-19 Cases in Nigeria Using Some Selected Count Data Regression Models." Two of the models the authors fitted was the Poisson Regression (PR) and Negative Binomial Regression (NBR) models. The authors considered the daily cumulative deaths as dependent variable and three predictors (Active cases, Critical cases and Confirmed cases). The summary results of their fitted model are given in Tables 6 and 7, respectively. Answer Questions 5.1 and 5.2 based these results.

Table 6: Summary of the results of the Poisson regression model

	Estimate	Std. Error	z value	Pr(> $\|z\|$)
(Intercept)	3.168602	0.02082	152.18	< 0.001
Active cases	0.000616	0.00002	32.46	< 0.001
Critical cases	0.110749	0.00389	28.51	< 0.001
Confirmed cases	-0.00027	0.02982	-25.80	< 0.001
AIC	4289.529			
Deviance	4289.529 (df=128)			
log-likelihood:	-2140.7646			

Table 7: Summary of the results of the Negative binomial model

	Estimate	Std. Error	z value	Pr(> $\|z\|$)
(Intercept)	1.599214	0.16575	9.65	< 0.001
Active cases	0.001466	0.00029	5.01	< 0.001
Critical cases	0.305694	0.07497	4.08	< 0.001
Confirmed cases	-0.00077	0.00016	-4.86	< 0.001
AIC	1296.721			
log-likelihood:	-643.36058			

5.1 Referring to result **(Poisson regression) presented in Table 6,**

5.1.1 Compute the baseline rate value. [3]

5.1.2 Find and interpret the rate ratio associated with the variable "Critical cases". [4]

5.1.3 Find and interpret the rate ratio for a 100 unit increase in active cases of COVID-19. [4]

5.2 Which model is the best among the two fitted models? (Provide reasons) [3]

== END OF QUESTION PAPER ==
Total: 100 marks