Strathmore
UNIVERSITY

STRATHMORE INSTITUTE OF MATHEMATICAL SCIENCES

MASTER OF SCIENCE IN STATISTICAL SCIENCES

END OF SEMESTER EXAMINATION

STA 8307: CATEGORICAL DATA ANALYSIS

Date: 15TH AUGUST, 2023 **Time: 3 Hours**

Instructions

1. This examination consists of **FIVE** questions.
2. Answer **Question ONE (COMPULSORY)** and any other **TWO** questions.

QUESTION ONE (20 MARKS)

1. Identify the response variables in the following cases: **[2 marks]**
   a. Attitude towards gun control (favor, oppose), Gender (female, male), Mother's education (high school, college).
   b. Heart disease (yes, no), Blood pressure, Cholesterol level.
   c. Gender (male, female), Religion (Catholic, Protestant, Muslim, Jewish), Vote for president, Annual income.
   d. Marital status (married, single, divorced, widowed), Quality of life (excellent. Good. Fair, poor).
2. Determine which scale of measurement, between nominal and ordinal, that is most appropriate for the following variables **[3 marks]**
   a. Political party affiliation
   b. Highest level of education obtained
   c. Patient condition – good, fair, serious, critical.
   d. Hospital location.
   e. Favorite beverage.
   f. How often one feels depressed – never, occasionally, often, always.

3. Each of 100 multiple –choice questions on an exam has four possible answers but one correct response. For each question, a student randomly selects one response as the answer.
   a. Specify the distribution of the student's number of correct answers on the exam. **[1 mark]**
   b. Based on the mean and the standard deviation of the distribution, would it be surprising if the student made at least 50 correct responses? Explain your reasoning. **[2 marks]**
4. Each subject in a sample of 100 men and 100 women is asked to indicate which of the following factors (one or more) are responsible for increase in teenage crime. A – the increasing gap in income between the rich and poor; B – the increase in the percentage of single-parent families; C – insufficient time spent by parents with their children. A cross classification of the responses by gender is:

Gender	A	B	C
Men	60	81	75
Women	75	87	86

   a. Comment on the use of chi-squared test of independence to this table. **[2 marks]**
   b. Explain how this table actually provides information needed to cross classify gender with each of the three variable. **[2 marks]**
   c. Construct the contingency table relating gender to opinion about whether factor *A* is responsible for increase in teenage crime. **[2 marks]**
5. Describe the purpose of the link function of a GLM. Define the identity link, and explain why it is not often used with the binomial parameter. **[6 marks]**

QUESTION TWO (20 MARKS)

---

i. An experiment analyzes imperfection rates for two processes used to fabricate silicon wafers for computer chips. For treatment A applied to 10 wafers, the numbers of imperfections are $8, 7, 6, 6, 3, 4, 7, 2, 3, 4$. Treatment B applied to 10 other wafers has $9, 9, 8, 14, 8, 13, 11, 5, 7, 6$ imperfections. Treat the counts as independent Poisson variates having means μ_A and μ_B. Consider the model $\log \mu = \alpha + \beta x$, where $x = 1$ for treatment B and $x = 0$ for treatment A.
   a. Show that $\beta = \log \mu B - \log \mu A = \log\left(\frac{\mu_B}{\mu_A}\right)$ and $e^{\beta} = \frac{\mu_B}{\mu_A}$. **[2 marks]**
   b. Fit the model. Report the prediction equation and interpret $\hat{\beta}$. **[4 marks]**

c. Test H_0: $\mu_A = \mu_B$ by conducting the Wald or likelihood-ratio test of H_0: $\beta = 0$. Interpret. **[4 marks]**

d. Construct a 95% confidence interval for $\frac{\mu_B}{\mu_A}$. **[3 marks]**

ii. The wafers above are also classified by thickness of silicon ($z = 0, low; z = 1, high$). The first five imperfection counts reported for each treatment refer to $z = 0$ and the last five refer to $z =$ 1. Analyze these data, making inferences about the effects of treatment type and thickness of the coating. **[7 marks]**

QUESTION THREE (20 MARKS)

A study used logistic regression to determine characteristics associated with $Y =$ whether a cancer patient achieved remission ($1 = yes$). The most important explanatory variable was a labeling index (LI) that measures proliferative activity of cells after a patient receives an injection of tritiated thymidine. It represents the percentage of cells that are "labeled." The table below shows the grouped data:

Data on Cancer Remission.

LI	Number of Cases	Number of Remissions	LI	Number of Cases	Number of Remissions	LI	Number of Cases	Number of Remissions
8	2	0	18	1	1	28	1	1
10	2	0	20	3	2	32	1	0
12	3	0	22	2	1	34	1	1
14	3	0	24	1	0	38	3	2
16	3	0	26	1	1			

The corresponding software reports for logistic regression model using LI to predict $\pi = P(Y = 1)$ is as below:

Computer Output

Parameter	Estimate	Standard Error	Likelihood Ratio 95% Conf. Limits		Chi-Square	
Intercept	−3.7771	1.3786	−6.9946	−1.4097	7.51	
Li	0.1449	0.0593	0.0425	0.2846	5.96	

Source	DF	LR Statistic Chi-Square	Pr>ChiSq
li	1	8.30	0.0040

Obs	li	remiss	n	pi_hat	lower	upper
1	8	0	2	0.06797	0.01121	0.31925
2	10	0	2	0.08879	0.01809	0.34010

….

a. Show how software obtained $\hat{\pi} = 0.068$ when $LI = 8$. **[2 marks]**

b. Show that $\hat{\pi} = 0.50$ when $LI = 26.0$. **[2 marks]**

c. Show that the rate of change in the $\hat{\pi}$ is 0.009 when $LI = 8$ and is 0.036 when $LI = 26$. **[5 marks]**

d. The lower quartile and upper quartile for LI are 14 and 28. Show that $\hat{\pi}$ increases by 0.42, from 0.15 to 0.57, between those values. **[2 marks]**

e. When LI increases by 1, show the estimated odds of remission multiply by 1.16. **[2 marks]**

f. Conduct a Wald test for the LI effect and interpret it. **[2 marks]**

g. Conduct a likelihood-ratio test for the LI effect and interpret it. **[2 marks]**

h. Construct the likelihood-ratio confidence interval for the odds ratio and interpret it. **[3 marks]**

QUESTION FOUR (20 MARKS)

The president of a large university recently announced that the school would be switching to hostels that are all single-sex, because, he says, research shows that single-sex hostels reduce the number of student hook-ups. He cites studies showing that, in universities that offer both same-sex and coed housing, students in coed hostels report hooking up more often.

a. What are the cases in the studies cited by the university president? What are the two variables being discussed? Identify each as categorical or quantitative. **[6 marks]**

b. Which is the explanatory variable and which if the response variable? **[2 marks]**

c. According to the second sentence, does there appear to be an association between the variables? **[2 marks]**

d. Use the first sentence to determine whether the university president is assuming a causal relationship between the variables. **[1 mark]**

e. Use the second sentence to determine whether the cited studies appear to be observational studies or experiments. **[1 mark]**

f. Explain a confounding variable that might be influencing the association. **[5 marks]**

g. Can we conclude from the information in the studies that the single-sex hostels reduce the number of student hook-ups? **[1 mark]**

h. What common mistake is the university president making? **[2 marks]**

QUESTION FIVE (20 MARKS)

The table below is from a recent General Social Survey, cross-classifies the degree of fundamentalism of subjects' religious beliefs by their highest degree of education. The table also shows standardized residuals. For these data, $\chi^2 = 69.2$. Write a report of about 200 words, summarizing description and inference for these data. **[20 marks]**

	Religious Beliefs		
Highest Degree	Fundamentalist	Moderate	Liberal
Less than high school	178	138	108
	(4.5)	(−2.6)	(−1.9)
High school or junior college	570	648	442
	(2.6)	(1.3)	(−4.0)
Bachelor or graduate	138	252	252
	(−6.8)	(0.7)	(6.3)