****

**GARISSA UNIVERSITY**

**UNIVERSITY EXAMINATION 2019/2020 ACADEMIC YEAR TWO**

**SECOND SEMESTER EXAMINATION**

**SCHOOL OF SCHOOL OF PURE AND APPLIED SCIENCES**

**FOR THE DEGREE OF BACHELOR OF EDUCATION**

**COURSE CODE: STA 219**

**COURSE TITLE: CATEGORICAL DATA ANALYSIS**

**EXAMINATION DURATION: 2 HOURS**

**DATE: 11/12/2020 TIME: 03.00-05.00 PM**

**INSTRUCTION TO CANDIDATES**

* **The examination has FIVE (5) questions**
* **Question ONE (1) is COMPULSORY**
* **Choose any other TWO (2) questions from the remaining FOUR (4) questions**
* **Use sketch diagrams to illustrate your answer whenever necessary**
* **Do not carry mobile phones or any other written materials in examination room**
* **Do not write on this paper**

**This paper consists of FOUR (4) printed pages *please turn over***

**QUESTION ONE (COMPULSORY)**

1. An experiment studies the number of insects that survive a certain dose of an insecticide, using several batches of insects of size *n* each. The insects are sensitive to factors that vary among batches during the experiment but were not measured, such as temperature level. Explain why the distribution of the number of insects per batch surviving the experiment might show overdispersion relative to a  distribution. **[3 marks]**
2. In the following examples, identify the response variable and the explanatory variable.
3. Attitude towards gun control (favor, oppose), Gender (male, female), mother’s education (high school, college).  **[1 mark]**
4. Heart disease (yes, no), blood pressure, cholesterol level. **[1 mark]**
5. Race (white, nonwhite), Religion (Catholic, Muslim, Jewish, Protestant, Hindu), Vote for president, Annual income. **[1 mark]**
6. Marital status (married, single, divorced, widowed), Quality of life (excellent, good, fair, poor). **[1 mark]**
7. Which measurement scale is most appropriate for the following variables – nominal or ordinal?
8. Political affiliation (Jubilee, NASA, unaffiliated). **[1 mark]**
9. Highest degree obtained (none, high school, bachelor’s, master’s, PhD). **[1 mark]**
10. Patient condition (good, fair, serious, critical).  **[1 mark]**
11. Hospital location (Thika, Mwingi, Garissa, Nairobi). [**1 mark]**
12. Favorite beverage (beer, juice, milk, soft drink, wine, other). **[1 mark]**
13. How often do you feel depressed (never, occasionally, often, always)? [**1 mark**]
14. Based on 2014 murder rates in Ethiopia, an Associated Press story reported that the probability that a newborn child has eventually been a murder victim is 0.0263 for nonwhite males, 0.0049 for white males, 0.0072 for nonwhite females, and 0.0023 for white females.
15. Find the conditional odds ratios between race and whether a murder victim, given the gender. Interpret. Do these variables exhibit homogeneous association? **[3 marks]**
16. Half the newborns are of each gender, for each race. Find the marginal odds ratio between race and whether a murder victim. **[3 marks]**
17. An estimated odds ratio for adult females between the presence of squamous cell carcinoma (Yes, No) and smoking behavior (smoker, nonsmoker) equals 11.7 when the smoker category has subjects whose smoking level *s* is 0 < s< 20 cigarettes per day; it is 26.1 for smokers with  cigarettes per day. Show that the estimated odds ratio between carcinoma (yes, no) and the smoking levels  equals 2.2. **[5 marks]**
18. Each of the 100 multiple-choice questions on an exam has four possible answers but only one correct response. For each question, a student randomly selects one response as the answer.
19. Suggest a distribution for every question in the exam. Explain. **[2 marks]**
20. Specify the distribution of the student’s number of correct answers in the exam. **[2 marks]**
21. Based on the mean and standard deviation of that distribution, would it be surprising if the student made at least 50 correct responses? Explain your reasoning**. [2 marks]**

**QUESTION TWO**

Show that the moment generating function (mgf) for the binomial distribution is  and use it to obtain the first two moments. **[20 marks]**

**QUESTION THREE**

1. For a diagnostic test of a certain disease,  denotes the probability that the diagnosis is positive given that a subject has the disease, and  denotes the probability that the diagnosis is positive given that a subject does not have it. Let  denote the probability that a subject does have the disease.
2. Given that the diagnosis is positive, show that the probability that a subject does have the disease is:

 **[4 marks]**

1. Suppose that a diagnostic test for HIV+ status has both sensitivity and specificity equal to 0.95, and  = 0.005 respectively. Using the joint probabilities relating diagnosis to actual disease status, find the probability that a subject is truly HIV+, given that the diagnostic test is positive. **[4 marks]**
2. The following table is based on records of accidents in 2018 compiled by the Department of Highway Safety and Motor Vehicles in California.



1. Identify the response variable **[1 mark]**
2. Find and interpret the difference of proportions **[3 marks]**
3. Obtain and interpret the relative risk **[3 marks]**
4. Obtain and interpret the odds ratio. **[3 marks]**
5. Why are the relative risk and odds ratio approximately equal? **[2 marks]**

**QUESTION FOUR**

In a crossover trial comparing a new drug to a standard,  denotes the probability that the new one is judged better. It is desired to estimate  and test  against  . In 20 independent observations, the new drug is better each time.

1. Find and sketch the likelihood function. Give the ML estimate of  **[2 marks]**
2. Conduct a Wald test and construct a 95% Wald confidence interval for . Are these sensible? **[3 marks]**
3. Conduct a score test, reporting the P-value. Construct a 95% score confidence interval. Interpret. **[3 marks]**
4. Conduct a likelihood-ratio test and construct a likelihood-based 95% confidence interval. Interpret. **[3 marks]**
5. Construct an exact binomial test and 95% confidence interval. Interpret **[3 marks]**
6. Suppose that researchers wanted a sufficiently large sample to estimate the probability of preferring the new drug to within 0.05, with 0.95 confidence level. If the true probability is 0.90, about how large a sample is needed? **[6 marks]**

**QUESTION FIVE**

The data in the following table refers to a retrospective study of lung cancer and tobacco smoking among patients in several English hospitals. The table compares male lung cancer patients with control patients having other diseases, according to the average number of cigarettes smoked daily over a 10-year period preceding the onset of the disease.



1. Find the sample odds of lung cancer at each smoking level and the five odds ratios that pair each level of smoking with no smoking. As smoking increases, is there a trend? Interpret. **[5 marks]**
2. If the log odds of lung cancer is linearly related to smoking level, the log odds in row *i* satisfies . Show that this implies that the local odds ratios are identical. **[5 marks]**
3. Using these data, can you estimate the probability of lung cancer at each level of smoking? Are the estimated odds ratios in part (a) meaningful? Explain. **[5 marks]**
4. Show that the disease groups are stochastically ordered with respect to their distributions on smoking of cigarettes  **[5 marks]**