STA4030: Categorical Data Analysis

Assignment 2

Due Date and Time: October 24, 2020 (Saturday), 10:00PM

INSTRUCTION:

- Please scan your answers in **one single .pdf file** and submit via Blackboard System.
- **Late submissions** will receive a mark of zero.
- Students may discuss set problems with others, but your final submissions must be your own work.
- All these questions should be answered using a pen, paper, calculator (good practice for your midterm and final).
- You may use any software you like, e.g., R, Python, Excel, etc., to find the percentiles regarding relative distributions (for example, to find p-values).
- Show and write down your solutions in detail and clearly.

Problem Set 2:

1. Table 1 is from a report on the relationship between aspirin use and myocardial infarction (heart attacks) by the Physicians' Health Study Research Group at Harvard Medical School. Find the P-value for testing that the incidence of heart attacks is independent of aspirin intake using the chi-square test. Interpret your results.

	Myocardial Infarction	
Group	Yes	No
Placebo	158	10321
Aspirin	71	10410

Table 1: Heart Attack Data.

2. An analysis of campus accident data was made to determine the distribution of numbers of fatal accidents for automobiles of two sizes. The data for 16 accidents are given in Table 2. Do the data indicate that the frequency of fatal accidents is independent on the size of automobiles? Choose a test, justify your choice, perform it and interpret the results.

	Size of Auto		
	Small	Large	Total
Fatal	2	6	8
Non-fatal	4	4	8
Total	6	10	16

Table 2: Campus Accident Data.

3. Give a "real world" example of three variables X, Y, and Z, for which you expect X and Y to be marginally associated but conditionally independent, controlling for Z.

4. Table 3 is based on records of accidents in 1988 compiled by the Department of Highway Safety and Motor Vehicles in Florida.

	Injury	
Safety Equipment in Use	Fatal	Non-fatal
None	1598	162526
Seat Belt	502	412360

Table 3: Highway Safety Data.

   (a). Find and interpret the difference of proportions, relative risk, and odds ratio. Why are the relative risk and odds ratio approximately equal?

   (b). Construct 95% confidence intervals for the difference of proportions, the relative risk, and the odds ratio. Interpret.

5. Table 4 refers to applicants to graduate school at the University of California, Berkeley for the fall 1973 session. Admissions decisions are presented by gender of applicant, for the three largest graduate departments. Denote the three variables by A = whether admitted, G = gender, and D = department.

	Whether Admitted			
	Male		Female	
Department	Yes	No	Yes	No
1	478	302	80	23
2	365	199	16	7
3	117	203	204	385
4	133	276	127	250
5	53	138	94	299
6	22	351	24	317

Table 4: Berkeley Data.

   (a). Find the sample AG conditional odds ratios, and compare them with the sample AG marginal odds ratio. Why are they so different?

(b). Conduct the Cochran-Mantel-Haenszel test. Specify the hypothesis and interpret. Comment on the applicability of this method to this data.

6. Table 5 is a three-way table which summarizes the data obtained from 300 couples. Note that, $Z = 1$ represents that Husband's Age < 50, and $Z = 2$ represents that Husband's Age > 50.

Z		$X \backslash Y$	Blood Pressure (Wife) Abnormal	Normal
1	Blood Pressure (Husband)	Abnormal	8	25
		Normal	28	95
2	Blood Pressure (Husband)	Abnormal	20	11
		Normal	16	97

Table 5: Blood Pressure Data.

(a). Test the hypothesis that X is marginally independent of Y.

(b). Test the hypothesis that given Z, X is independent of Y.

THE END