UNIVERSITY OF TORONTO SCARBOROUGH
Department of Computer and Mathematical Sciences
Midterm Test, October 2018

STAC51H3 Categorical Data Analysis
Duration: One hour and fifty minutes

Last Name: _____ First Name: _____

Student number: _____

Aids allowed: Solution

- Lecture slides and class notes taken by you
- Textbook (Categorical Data Analysis by Alan Agresti)
- A calculator (No phone calculators are allowed)

No other aids are allowed. For example you are not allowed to have any other textbook or past/sample exams.

All your work must be presented clearly in order to get credit. Answer alone (even though correct) will only qualify for **ZERO** credit. For questions that require numerical answers, you should provide numerical answers to a reasonable degree of accuracy. Just explaining how do them or just coping down the method of solving them from the class notes/book will not qualify for credit. Please show your work in the space provided; you may use the back of the pages, if necessary, but you MUST remain organized. Show your work and answer in the space provided, in ink. Pencil may be used, but then any re-grading will NOT be allowed.

There are 7 questions and 11 pages including this page. Please check to see you have all the pages.

Good Luck!

Question:	1	2	3	4	5	6	7	Total
Points:	12	6	6	11	11	12	13	71
Score:								

$n=10$

1. A random sample of ten employees in a company was given a training to improve their social competence. After the training, their social competence was evaluated and eight of them proved to have increased social competence. $Y=8$

   (a) (4 points) Use the exact binomial test to test the null hypothesis $H_0 : \pi = 0.5$ against the alternative $H_1 : \pi > 0.5$, where π denoted the population proportion of employees that will improve social competence from this training. Use $\alpha = 0.05$.

$$P(Y=8) + P(Y=9) + P(Y=10) = 0.0546875$$

which is greater than 0.05.
the data don't have a sufficient evidence of an improvement.

   (b) (4 points) Use the likelihood ratio test to test the null hypothesis $H_0 : \pi = 0.5$ against the two sided alternative $H_1 : \pi \neq 0.5$. Use $\alpha = 0.05$.

$$L(\pi) = \pi^Y (1-\pi)^{n-Y} = \pi^8 (1-\pi)^2$$

$$\Lambda = \frac{0.5^8 (1-0.5)^2}{0.8^8 (1-0.2)^2} = 0.1455192$$

$$-2\log\Lambda = 3.8548851 > 3.84$$

..So we reject H_0. with $\alpha = 5\%$.

   (c) (4 points) Calcualte Agresti-Coull 95 % confidence interval for π.

$$\tilde{\pi} = \frac{Y + z_{\alpha/2}^2/2}{n + z_{\alpha/2}^2} = \frac{8 + 1.96^2/2}{10 + 1.96^2}$$

$$= 0.7167379$$

$$0.7167379 \pm 1.96\sqrt{\frac{0.7167379(1-0.7167379)}{10+1.96^2}}$$

$$(0.4793618,\ 0.9541142)$$

$$\simeq (0.48,\ 0.95)$$

2. (6 points) Suppose a study looking at the association between smoking (individuals classified as smokers and non-smokers) and lung cancer found that lung cancer is more prevalent among smokers with odds ratio = 2.4. If 1 in 17 non-smokers develop lung cancer during their life time (i.e. the probability of developing lung cancer is 1/17 for non-smokers), what proportion of smokers develop lung cancer during their life time?

$$\frac{\left(\frac{P_S}{1-P_S}\right)}{\left(\frac{P_{NS}}{1-P_{NS}}\right)} = 2.4$$

$P_{NS} = 1/17.$

$$\frac{P_S}{1-P_S} = 2.4 \times \frac{1}{16} = 0.15$$

$$\therefore P_S = \frac{0.15}{1+0.15} = 0.13043478 \approx 0.13$$

3. (6 points) Based on a random sample, a researcher has calculated a 95% Wald confidence interval for the proportion of individuals having a particular disease in a population to be $(0.1608, 0.2392)$. Based on this data, calculate the 95 % score (Wilson) confidence interval for the population proportion.

$$\hat{\pi} = \frac{0.1608 + 0.2392}{2} = 0.2$$

$$SE = \frac{0.2392 - 0.1608}{2 \times 1.96} = 0.02$$

$$SE = \sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} = 0.02$$

$\therefore n = 400$

95% score C.I is:

Mid point: $\left(\frac{n\hat{\pi} + z^2/2}{n + z^2}\right)$

$$= \frac{400 * 0.2 + 1.96^2/2}{400 + 1.96^2}$$

$$= 0.2028538$$

SE: $\frac{\sqrt{n\hat{\pi}(1-\hat{\pi}) + z^2/4}}{n + z^2} = \frac{\sqrt{400 * 0.2 * 0.8 + 1.96^2/4}}{400 + 1.96^2}$

$$= 0.01995783$$

$0.2028538 \pm 1.96(0.01995783)$

$(0.1637365, 0.2419711) \approx (0.164, 0.242)$

4. The number of work-related accidents in any one-month period in an industrial plant, is a random variable (Y), with probability function given by:

$$P(Y = y) = \pi(1 - \pi)^y, y = 0, 1, 2, \ldots$$

The numbers of work-related accidents in the last 10 months are: 5, 4, 6, 4, 2, 5, 2, 3, 7,2.

(a) (5 points) Assuming that the numbers of accidents occurring in different months are independent, find the likelihood function of π.

$$\begin{aligned} L(\pi) &= p(Y_1=5, Y_2=4, \ldots, Y_{10}=2) \\ &= p(Y_1=5) \times \cdots \times (Y_{10}=2) \\ &= \pi(1-\pi)^5 \times \cdots \times \pi(1-\pi)^2 \\ &= \pi^{10}(1-\pi)^{40} \end{aligned}$$

(b) (6 points) Calculate the likelihood function at the values $\pi = 0.10, 0.15, 0.20, 0.25$. Comment on your values. In, particular, based on what you have calculated, can you guess what the maximum likelihood estimator of π will be, at least approximately? Note: As always, likelihood values are small. Please give your answers in the form $x \times 10^{-12}$, giving the values of x to 2 decimal places.

Using R

```
> pi = c(0.1, 0.15, 0.2, 0.25)
> L = pi^10 * (1-pi)^40
> L
```

~~$\pi = 0.2$ provides the~~

The likelihood is increasing till $\pi = 0.2$ and at $\pi = 0.25$, it has decreased. The maximum has occurred at $\pi = 0.2$ or somewhere between 0.2 and 0.25.

5. For years, brand awareness for Big Red chewing gum has been stuck at about 6%, meaning that about 6% of consumers who chew gum say they remember hearing about Big Red gum. The marketing department is planning an advertising campaign to increase brand awareness, in the hope that increased brand awareness will lead to increased sales. After the campaign was running for a few weeks, they wanted to test whether the brand awareness has actually increased. To test this, they interviewed a random sample of 200 gum chewers, and found that twenty had heard of Big Red.

   (a) (2 points) What is the appropriate null hypothesis corresponding to the main question in this study? Use appropriate notation as discussed in class.
   Hint: Brand awareness means the proportion of gum chewers who have heard of this brand (Big Red).

   $H_0: \pi = 0.6$

   (b) (2 points) What is the appropriate alternative hypothesis corresponding to the main question in this study? Use appropriate notation as discussed in class.

   $H_1: \pi > 0.6$

   (c) (7 points) Calculate the Wald test statistic and the corresponding p-value. Interpret your results.
   Test at the 5% level of significance.

$$\hat{\pi} = \frac{20}{200} = 0.1$$

$$Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}} = \frac{0.1 - 0.06}{\sqrt{\frac{0.1(1-0.1)}{200}}} = 1.89$$

p-value ≈ 0.02938 which is less than 5% so reject H_0.
There is a sufficient evidence to suggest that the awareness has increased.

6. A large introductory statistics class has students from all levels. 20% of them are first year students, 40% second year students, 30% third year students and the remaining 10% are fourth year students. We are going to select a random sample of 9 students from this class.

   (a) (4 points) What is mean and the standard deviation of the number of first year students in the sample?

X: # of first year students

$X \sim Bin(9, 0.2)$

$E(X) = 9 \times 0.2 = 1.8$. $SD(X) = \sqrt{9 \times 0.2 \times 0.8} = 1.2$

   (b) (4 points) What is the probability that this sample will have two first year students, two second year students , two third year students and three fourth year students?

$(Y_1, Y_2, Y_3, Y_4) \sim$ Multinomial$(n=9, \pi_1=0.2, \pi_2=0.4, \pi_3=0.3, \pi_4=0.1)$

$$P(Y_1=2, Y_2=2, Y_3=2, Y_4=3) = \frac{9!}{2!2!2!3!} 0.2^2 0.4^2 0.3^2 0.1^3$$

$$= 0.00435456$$

   (c) (4 points) What is the probability that this sample will have two first year students, two second year students and the other five either third or forth year students.

$(Y_1, Y_2, Y_3+Y_4) \sim$ Multinomial$(n=9, \pi_1=0.2, \pi_2=0.4, \pi_3=0.3+0.1)$

So $P(Y_1=2, Y_2=2, Y_3+Y_4=5)$

$$= \frac{9!}{2!2!5!} 0.2^2 0.4^2 0.4^5$$

$$= 0.04954522$$

7. Due to extensive drought conditions in 2014, there were fifty-three serious, 300+ wild fires in California. The contingency table given below displays the data collected on these wild fires. Variables are the season (either early in the year, Jan 1 through June 30, or late in the year, July 1 through Dec 31) and whether or not the wild fire was caused by human activity.

	Human Caused		
Season	Yes	No	
Early	8	9	17
Late	14	22	36
	22	31	53

A critical question that investigators might ask is whether or not the cause of the wild fire is independent from the season of year in which the fire occurred. Use the data above to answer the following questions.

(a) (2 points) What is the estimated joint probability that a wild fire is human-caused and is early in the year?

$8/53 = 0.1509$

(b) (2 points) What is the estimated marginal probability that a wild fire is early in the season?

$\frac{8+9}{53} = \frac{17}{53} \approx 0.3208$

(c) (2 points) Given that the wild fire is early in the season, what is the estimated probability that it is human-caused?

$8/17 = 0.4706$

(d) (7 points) Use the likelihood ratio test to test whether or not the cause of the wild fire is independent from the season of year in which the fire occurred.

$H_0: \theta = 1$ vs $Hi: \theta \neq 1$

$G^2 = 2 \sum_i \sum_j n_{ij} \log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right)$

Question 7 continues on the next page...

You may continue your to answer to question 7 on this page

Season	Human Caused: Yes	Human Caused: No
Early	7.056604	9.943396
Late	14.9434	21.0566

$$2 \times \left[8 \times \log\left(\frac{8}{7.056604}\right) + 9 \times \log\left(\frac{9}{9.943396}\right) + 14 \times \log\left(\frac{14}{14.9434}\right) + 22 \times \log\left(\frac{22}{21.0566}\right) \right]$$

$$\approx 0.316$$

conclusion

$a^2 = 0.316 < \chi^2_1(0.05) = 3.84$

∴ so we do not reject H_0.

No sufficient evidence to conclude that the cause of the wild fire depends on the season of year in which the fire occurred.

END OF TEST