Midterm 6620/8620

Total: 110 pts + 15 pts

1. (20 pts) In the lecture, we saw an example with a sample of 25 students and none of them were vegetarian. Denote π the population vegetarian proportion. Consider $H_0 : \pi = \pi_0$ vs $H_a : \pi \neq \pi_0$

(1) (2 pts) Write down the likelihood function using π
(2) (3 pts) Find L_0, the maximum likelihood under $H_0 : \pi = \pi_0$? (Hint: it should be expressed in terms of π_0)
(3) (3 pts) Find L_{max}, the maximum likelihood over all possible π (Hint: you have to find the MLE first and then the corresponding ML)
(4) (3 pts) For $H_0 : \pi = 0.5$ vs $H_a : \pi \neq 0.5$, show the likelihood-ratio test statistic and report the p-value. What's your conclusion?
(5) (3 pts) For $H_0 : \pi = 0.003$ vs $H_a : \pi \neq 0.003$, show the likelihood-ratio test statistic and report the p-value. What's your conclusion?
(6) (3 pts) Suppose now we observe 0 success out of n samples. Show that the $(1-\alpha)100\%$ likelihood ratio confidence interval for π is $(0, 1 - \exp(-\chi^2_{1,\alpha}/2n))$ where $P(X > \chi^2_{1,\alpha}) = \alpha, X \sim \chi^2_1$
(7) (3 pts) Suppose now we observe 0 success out of n samples. Show that the $(1-\alpha)100\%$ score confidence interval is $(0, z^2_{\alpha/2}/(n + z^2_{\alpha/2}))$

2. (10 pts) The table below is based on records of accidents in 1988 compiled by the Department of Highway Safety and Motor Vehicles in Florida. Identify the response variable, and find and interpret the difference of proportions, relative risk, and odds ratio and their corresponding 95% confidence interval. Why are the relative risk and odds ratio approximately equal?

Safety Equipment in Use	Fatal Injury	Nonfatal Injury
None	1601	162527
Seat Belt	510	412368

3. (15 pts) The table below presents admissions decisions by gender for the six largest graduate departments. Denote the 3 variables by A = whether admitted, G=gender, and D=department. Find the sample AG conditional odds ratios and the marginal odds ratio. Interpret, and explain why they give such different indications of the AG association.

	Whether Admitted			
	Male		Female	
Department	Yes	No	Yes	No
A	512	313	89	19
B	353	207	17	8
C	120	205	202	391
D	138	279	131	244
E	53	138	94	299
F	22	351	24	317
Total	1198	1493	557	1278

4. (15 pts) A study on educational aspirations of high school students measured aspirations with the scale. The counts for each category are shown as below, where FI denotes family income

Aspirations	FI Low	FI middle	FI high
Some high school	11	9	9
High school graduate	52	44	41
Some college	23	13	12
College graduate	22	10	27

(1) Test independence of educational aspirations and family income using X^2 and G^2, respectively. Show me how you find the expected count for μ_{22}

(2) Find the standardized Pearson residuals.

5. (50 + 5 pts) The horseshoe crab dataset is in `horseshoe crab.txt` with following variables:

Var	Description
`y`	whether a female crab has a satellite (1=yes, 0=no)
`weight`	weight in grams
`width`	width in centimeters
`color`	2=median light, 3=medium, 4=medium dark and 5=dark

(1) (10 pts) Read in dataset using the following command in `R` and divide `weight` by 1000 and now the weight is measured in `kg`

```
dat = read.table("horseshoe crab.txt",header=TRUE)
dat$weight = dat$weight/1000
```

Fit a logistic regression model using `width` and `color`. Treat `color` as **qualitative** and use color 5 as the reference level using the below command

```
dat$color = factor(dat$color)
dat$color = relevel(dat$color, ref="5")
```

Interpret effects.

(2) (3 pts) Construct 95% confidence intervals for each parameter in part (1)
(3) (5 pts) What is the average width? Predict the probability for each color at the average width.
(4) (5 pts) Test if color contributes significantly to the model.
(5) (5 pts) Test if there is significant difference between the slope parameters for medium-light and dark crabs. (Bonus: can you construct 95% CI for the difference? 5 pts)
(6) (5 pts) Does the model adding `weight` provide an improved fit? Interpret. (Hint: model comparison)
(7) (12 pts) Using models that treat color as **quantitative**, repeat the analyses in parts (1) to (3). You might need the following code to change the color to quantitative data:

```
dat$color = as.numeric(dat$color)
```

Bonus

1. (10 pts) Genotypes AA, Aa, and aa occur with probabilities $[\theta^2, 2\theta(1-\theta), (1-\theta)^2]$. A multinomial sample of size n has frequencies (n_1, n_2, n_3) of these three genotypes.

(1) Form the log likelihood L. Find $\hat{\theta}_{MLE}$
(2) Show that

$$-\frac{\partial^2 l(\theta)}{\partial \theta^2} = \frac{2n_1 + n_2}{\theta^2} + \frac{n_2 + 2n_3}{(1-\theta)^2}$$

and its expectation is

$$\frac{2n}{\theta(1-\theta)}.$$

Can you use this to obtain an asymptotic standard error of $\hat{\theta}_{MLE}$?