STAT 475/675: Midterm Exam 1 (Feb 1, 2018)

Name: *Student ID #:*

Question 1. [6 points] One of the key findings in an NHTSA (National Highway Traffic Safety Administration) data analysis is that 58% of the passenger vehicle occupants in age group 13 to 15 who were killed in traffic crashes were not using restraints – the highest percentage of all age groups in 2014. Assume all the subjects are independent and have the same probability of "*not using restraints*". Answer the following questions.

(i) Give a 95% confidence interval (CI) of the probability that a killed occupant aged 13 to 15 is *"not using restraints"* with the data in Table 1. Does the CI verify statistically that particular key finding of the NHTSA's study?

3

Table 1. **Passenger Vehicle Occupants Killed with Restraint Use**

	Restraint Use (Y)		
Age of Killed Occupant (Years)	Restrained	Unrestrained	Total
Age of 13-15	102	140	242

① $\pi = \text{Prob}(\underbrace{\text{a subject}}_{\text{a killed occupant aged 13-15}} \text{ is not using restraints})$ $\quad \hat{\pi} = \frac{140}{242}$

$\hat{V}(\hat{\pi}) = \frac{1}{242}\hat{\pi}(1-\hat{\pi})$

$n = 242$

$\Rightarrow$ 95% CI $\quad \hat{\pi} \pm (1.96)\sqrt{\hat{V}(\hat{\pi})} : \underline{(0.517, 0.641)}$

② The CI indicates that statistically more likely a killed occupant aged 13-15 was not using restraints. It verifies the study's finding.

(ii) Suppose that another study obtained the same percentage, 58%, of the passenger vehicle occupants in age group 13 to 15 who were killed in traffic crashes were *not using restraints* with the sample size $n = 100$. Did the second study indicate that a killed passenger vehicle occupant aged 13 to 15 is more likely *not using restraints* compared to *using restraints* at the significance level of 0.05? Justify your answer.

3

① With the new data, $n = 100$, $\hat{\pi} = 0.58$

$\Rightarrow$ 95% CI is $(0.48, 0.67)$

The CI doesn't indicate that a killed occupant aged 13 to 15 is more likely not using restraints, since it contains 50%. ②

The discrepancy is due to the smaller sample size.

Question 2. [14 points] Table 2 cross-classifies the NHTSA study data according to gender vs restraint use. Assuming the subjects are independent and with the same distribution with restraint use, answer the following questions based on Table 2.

Table 2. **Passenger Vehicle Occupants Killed with Restraint Use vs Gender**

	Restraint Use (Y)		
Gender (X)	Restrained	Unrestrained	Total
Female	4040	2748	6788
Male	5917	6635	12552
Total	9957	9383	19340

(i.a) Give the MLE of the probability of *"not using restraints"* among females, and the MLE of the probability of *"not using restraints"* among males. (i.b) Give the sample proportion of females with the data. Can you use it to estimate the population proportion of females in the whole nation? Why?

(i.a) $\hat{\pi}_F = 2748/6788 \doteq 0.405$; $\hat{\pi}_M = 6635/12552 \doteq 0.529$

3 (i.b) The sample proportion of females with the data is $\frac{6788}{12552} \doteq 0.351$

(1)

(2) No, I would not. The reason is that very likely the current study sample isn't a random sample from the whole population

(ii) Give the sample odds ratio (OR) of *"not using restraints"* for females comparing to males, and use $\hat{\sigma}_{\log(\widehat{OR})} = 0.0305$, an estimate for the standard error of the log sample OR, to construct a 95% confidence interval (CI) for the OR of *"not using restraints"* for females comparing to males.

(1) $\hat{OR} = (2748)(5917)/(4040)(6635) \doteq 0.607$

3 $\Rightarrow$ An approximate 95% CI of $\log(OR)$ is

$$\log \hat{OR} \pm (1.96)(0.0305)$$

(2) $\therefore$ $(0.571, 0.644)$ is a 95% CI of the OR

(iii) What does the CI in (ii) tell about the difference between females and males in using restraints overall? Explain your finding.

It indicates that females of the killed occupants were less likely compared to the males "not using restraints"

3 This conclusion is based on that the CI doesn't contain 1 (all CI $\subseteq [0,1)$), the odds of females not using restraints is significantly lower compared to the males

(iv) Appendix A is the R outputs of the Pearson's χ^2-test with the data. List the hypotheses of the test, the test statistic, its approximate distribution, and draw a conclusion based on the p-value in the R outputs.

H_0: $X \perp\!\!\!\perp Y$ vs H_1: otherwise

The test statistic is $K = \sum_{\substack{i=1,2 \\ j=1,2}} (N_{ij} - \hat{\mu}_{ij})^2 / \hat{\mu}_{ij}$,

with $\hat{\mu}_{ij} = (N_{i+})(N_{+j})/N$, $N \equiv n$

3

$K \overset{H_0}{\sim} \chi^2(1)$ approximately

$K_{obs} = 270.19$, $p = < 0.001$

independence between "gender" and "restraint use" among the killed occupants.

$\Rightarrow$ The data provide strong evidence against the ↓

(v) Appendix B is the R outputs of the analysis with the data under the simple logistic regression model $logit(\pi(x)) = \alpha + \beta x$: the response Y is the indicator of *"not using restraints"*, the explanatory variable X is gender (the indicator of male), and $\pi(x) = P(Y = 1|X = x)$. Choose to answer one of the two following questions

(v.a) Give a CI of β. What is the connection of the CI to the one in (ii)?

(v.b) Give estimates of $\pi(x)$ with $x = 0$ and $x = 1$ ("female" and "male"). How are they related to the estimates in (i)?

(V.a) ① an approximate 95% CI of β is

$\hat{\beta} \pm (1.96) SE(\hat{\beta}) \Rightarrow 0.4999 \pm (1.96)(0.0305)$

$\therefore$ a CI of β is $(0.44, 0.56)$

(with 95% level)

2

② Since e^{β} is the odds ratio of not using restraints in males comparing to females $\therefore$ $e^{-\beta}$ is the OR considered in (ii)

And $\exp\{-0.44\}$, $\exp\{-0.56\}$ $\Rightarrow$ the upper/lower limits of the CI in (ii)

(V.b) $\hat{\pi}(0) = e^{\hat{\alpha}+\hat{\beta}(0)} / (1+e^{\hat{\alpha}+\hat{\beta}(0)}) \doteq 0.405$

$\hat{\pi}(1) = e^{\hat{\alpha}+\hat{\beta}(1)} / (1+e^{\hat{\alpha}+\hat{\beta}(1)}) \doteq 0.529$

the same

They both estimate the female/male's prob. of not using restraints. (MLE)