Psy 5011: Categorical Data Analysis **(Due: 24/03/20)**

Homework #2

Instructor: Yao, Kaiping Grace *Name:* Hao-Cheng Lo, *Id:* D08227104

Problem 2.1: The PSA blood test is designed to detect prostate cancer. Suppose that of men who have this disease, the test fails to detect prostate cancer in 1 in 4, and of men who do not have it, 1 in 10 have positive test results (so-called false-positive results). Let C ($\bar{C}$) denote the event of having (not having) prostate cancer and let $+$ ($-$) denote a positive (negative) test result.

(a) Which is true: $P(-|C) = 1/4$ or $P(C|-) = 1/4$? $P(\bar{C}|+) = 1/10$ or $P(+|\bar{C}) = 1/10$?

Given men who have this disease, the test fails to detect prostate cancer in 1 in 4:

$$P(-|C) = 1/4 \text{ is correct.}$$

Given men who do not have this disease, 1 in 10 have positive test results:

$$P(+|\bar{C}) = 1/10 \text{ is correct.}$$

(b) Find the sensitivity and specificity of this test.

Sensitivity (Given C, the prob of $+$):

$$P(+|C) = 3/4$$

Specificity (Given $\bar{C}$, the prob of $-$):

$$P(-|\bar{C}) = 9/10$$

(c) Of men who take the PSA test, suppose $P(C) = 0.04$. Find the cell probabilities in the 2×2 table for the joint distribution that cross-classifies $Y =$ diagnosis with $X =$ true disease status.

	Diagnosis (Y)	
True disease status (X)	YES ($+$)	NO ($-$)
YES (C) ($P(C) = 0.04$)	$P(+ \cap C) = P(+\|C)P(C)$ $= 3/4 \times .04 = .03$	$P(- \cap C) = P(-\|C)P(C)$ $= 1/4 \times .04 = .01$
NO ($\bar{C}$) ($P(\bar{C}) = 0.96$)	$P(+ \cap \bar{C}) = P(+\|\bar{C})P(\bar{C})$ $= 1/10 \times .96 = .096$	$P(- \cap \bar{C}) = P(-\|\bar{C})P(\bar{C})$ $= 9/10 \times .96 = .864$

(d) Using (c), find the marginal distribution for the diagnosis and show that $P(C|+) = 0.238$. (In fact, the National Cancer Institute estimates that only about 25% of men who have a slightly elevated PSA level, 4–10 ng/mL, actually have prostate cancer.)

The marginal distribution for the diagnosis:

$$P(+) = .03 + .096 = .126$$
$$P(-) = .01 + .864 = .874$$

And, $P(C|+) = .03/.126 = .238$.

Problem 2.5: Consider the following two studies reported in the *New York Times*:

(a) A British study reported that of smokers who get lung cancer, "women were 1.7 times more vulnerable than men to get small-cell lung cancer." Is 1.7 an odds ratio or a relative risk?

1.7 is relative risk. That is $P(\text{Cancer}|\text{Women}) = 1.7P(\text{Cancer}|\text{Men})$.

(b) A National Cancer Institute study about tamoxifen and breast cancer reported that the women taking the drug were 45% less likely to experience invasive breast cancer compared to the women taking placebo. Find the relative risk for (i) those taking the drug compared to those taking placebo, (ii) those taking placebo compared to those taking the drug.

Let d for the women taking the drug and d^c for the women taking the placebo; c for cancer and c^c for not getting cancer. If $P(c|d^c) = \pi$, then $P(c|d) = (100\% - 45\%)\pi = .55\pi$.

For (i), the relative risk for those taking the drug compared to those taking placebo:

$$P(c|d)/P(c|d^c) = .55\pi/\pi = .55$$

For (ii), the relative risk for those taking placebo compared to those taking the drug:

$$P(c|d^c)/P(c|d) = \pi/.55\pi = 1.82$$

Problem 2.7: For adults who sailed on the Titanic on its fateful voyage, the odds ratio between gender (female, male) and survival (yes, no) was 11.4.:

(a) What is wrong with the interpretation, "The probability of survival for females was 11.4 times that for males?" Give the correct interpretation.

Odds ratio is the ratio of the odds of female (survival to not survival) to the odds of male (survival to not survival). So, the odds of survival for females was 11.4 times the odds of survival for males. Odds are not probabilities, so we cannot apply the odds ratio value to the probability of survival.

(b) The odds of survival for females equaled 2.9. For each gender, find the proportion who survived. Find the value of RR in the interpretation, "The probability of survival for females was RR times that for males."

The odds of survival for males equals $2.9/11.4 = .254$.

Let $P(Y|F) = \pi_f$, then $\pi_f/(1-\pi_f) = 2.9$, $\pi_f = .744$. Hence, for female, the proportion of survival is about 74%.

Let $P(Y|M) = \pi_m$, then $\pi_m/(1-\pi_m) = .25$, $\pi_m = .203$. Hence, for male, the proportion of survival is about 20%.

Therefore, the value of RR is $.744/.203 = 3.67$.

Problem 2.17: Table 2.13 is based on data from the 2016 General Social Survey.:

Table 2.13 Data for Exercise 2.17.

	Political Party Identification		
Race	Democrat	Republican	Independent
White	871	821	336
Black	347	42	83

(a) Test the null hypothesis of independence between political party identification and race. Interpret.

Set $\alpha = .05$, and null and alternative hypotheses:

$$H_O : \pi_{ij} = \pi_{i+}\pi_{+j};\ H_A : \pi_{ij} \neq \pi_{i+}\pi_{+j}\ i = 1, ..., I; j = 1, .., J$$

Test statistic:

$$\text{Under } H_O\text{: } X^2 = \sum_{ij} \frac{(n_{ij} - \hat{\mu_{ij}})^2}{\hat{\mu_{ij}}} \sim \chi^2_{df=1\times 2=2}$$

$$X^2 = \sum_{ij} \frac{(n_{ij} - \hat{\mu_{ij}})^2}{\hat{\mu_{ij}}} = \sum_{ij} \frac{(n_{ij} - n\hat{\pi_{i+}}\hat{\pi_{+j}})^2}{n\hat{\pi_{i+}}\hat{\pi_{+j}}} = 184.323$$

Because $X^2 = 184.323 > \chi^2_{\alpha=.05;\ df=2} = 5.99$, we reject H_O which indicates that the race and political party identification are not independent to each other. That is, for different race, the pattern of political party identification is different, and vice versa.

(b) Use standardized residuals to describe the evidence.

If $|e^*_{ij}| = |\frac{n_{ij} - \hat{\mu_{ij}}}{\sqrt{\hat{\mu_{ij}}(1 - \hat{\pi_{i+}})(1 - \hat{\pi_{+j}})}}| > 2$, then the cell is lacking of fit.

$$|e^*_{11}| = |-11.9|;\ |e^*_{12}| = |+12.9|;\ |e^*_{13}| = |-0.53|$$
$$|e^*_{21}| = |+17.3|;\ |e^*_{22}| = |-48.0|;\ |e^*_{23}| = |+1.11|$$

Observing the std. residuals, we can find that (i) for the Democrat, black people have propensities to identify with Democrat but white people have propensities to not identify with Democrat; (ii) for the Republican, black people have propensities to not identify with Republican but white people have propensities to identify with Democrat; (iii) for Independent, there are no significant differences between the identifications of white and of black.

(c) Partition chi-squared into two components, the first of which compares the races on the (Democrat, Republican) choice. Interpret the quite different results for the two cases.

Here, I partition the table into (i) first 2 cols (the races on the Democrat and Republican choice), which forms a 2×2 table with cell counts of $\begin{bmatrix} 871 & 821 \\ 347 & 42 \end{bmatrix}$ with $df = 1$, and (ii) combining the first two cols and compares them to the third col, which forms a 2×2 table with cell counts of $\begin{bmatrix} 871 + 821 & 336 \\ 347 + 42 & 83 \end{bmatrix} = \begin{bmatrix} 1692 & 336 \\ 389 & 83 \end{bmatrix}$ with $df = 1$.

For (i), there is strong evidence ($X^2 = 185.45$, P-value $< .00001$) of a difference between white and black in the relative numbers in the two categories (i.e. Democrats and Republicans). For (ii), there is no evidence ($X^2 = 0.283$, P-value $= .594$) of a difference between white and black in the relative numbers identifying as Independent or Democrat/Republican.

Problem 2.22: A study considered the effect of prednisolone on severe hypercalcaemia in women with metastatic breast cancer. Of 30 patients, 15 were randomly selected to receive prednisolone and the other 15 formed a control group. Normalization in their level of serum-ionized calcium was achieved by 7 of the 15 prednisolone-treated patients and by 0 of the 15 patients in the control group. Use Fisher's exact test to find a P-value for testing whether results were significantly better for treatment than control. Interpret.

Let X be the prednisolone treatment (x_1) and control group (x_2); Let Y be the normalization in level of serum-ionized calcium (y_1) and not normalization (y_2). Cell counts of X as row and Y as column forms $\begin{bmatrix} 7 & 8 \\ 0 & 15 \end{bmatrix}$. Due to sparse table and fixed marginal sampling, we employ Fisher's exact test and the probability distribution of a cell (representing whole table) follows hypergeometric distribution. For testing whether results were significantly better for treatment than control. Set $\alpha = .05$, and null and alternative hypotheses:

$$H_O : \theta = 1;\ H_A : \theta > 1$$

Compute the odds ratio of the observed table:

$$\hat{\theta} = \infty \implies \tilde{\theta} = 27.35$$

Compute the prob. for the tables where the odds ratios are more than odds ration from the obs. table:

$$p\text{-value} = \sum_{6<i<16} \frac{\binom{15}{i}\binom{15}{7-i}}{\binom{30}{7}} = .003161$$

By right-tail test, the p-value is $< .05$. We reject H_O and claim that true odds ratio is greater than 1, which implies taking treatment has greater odds of normalization in level of serum-ionized calcium than not taking.

Problem 2.28: The expected frequencies in Table 2.16 show a hypothetical relationship among three variables: Y = response, X = drug treatment, and Z = clinic. Show that X and Y are conditionally independent, given Z, but marginally associated. Explain how the marginal XY association can be so different from its conditional association, using the values of the conditional XZ and YZ odds ratios. Explain why it would be misleading to study only the marginal table and conclude that successes are more likely with treatment A than with treatment B.

Table 2.16 Expected frequencies illustrating that conditional independence does not imply marginal independence.

Clinic	Drug Treatment	Response	
		Success	Failure
1	A	18	12
	B	12	8
2	A	2	8
	B	8	32

Fix the level of Z, the conditional XY association given kth level of Z are:

$$\hat{\theta}_{XY(1)} = \frac{18 \times 8}{12 \times 12} = 1; \hat{\theta}_{XY(2)} = \frac{2 \times 32}{8 \times 8} = 1$$

Hence, XY is conditionally independent, given Z.

Sum across K levels of Z yields marginal table; the marginal association is defined by its odds ratio:

$$\hat{\theta}_{XY} = \frac{(18+2) \times (8+32)}{(12+8) \times (12+8)} = 2$$

Hence, XY is marginally associated, cross Z.

Next, I take further investigation on conditional XZ and YZ odds ratio to explain the phenomenon (Simpson's Paradox) above:

$$\hat{\theta}_{ZX(\mathrm{S})} = \frac{18 \times 8}{12 \times 2} = 6; \hat{\theta}_{ZX(\mathrm{F})} = \frac{12 \times 32}{8 \times 8} = 6; \hat{\theta}_{ZY(\mathrm{A})} = \frac{18 \times 8}{12 \times 2} = 6; \hat{\theta}_{ZY(\mathrm{B})} = \frac{12 \times 32}{8 \times 8} = 6$$

We can trivially find that given Y (resp. X), the conditional odds of treatment A (resp. response S) are 6 times higher at clinic 1 than clinic 2. That is, clinic 1 tends to use treatment A and have more success. It would be misleading to use marginal table only and conclude that successes are more likely with treatment A than with treatment B. Subjects within a particular clinic are more homogeneous than overall sample, and response is independent of treatments in each clinic.