STA4504 Homework 2

Edgard Maboudou

2.1 An article in the *New York Times* (February 17, 1999) about the PSA blood test for detecting prostate cancer stated that, of men who had this disease, the test fails to detect prostate cancer in 1 in 4 (so called false-negative results), and of men who did not have it, as many as two-thirds receive false-positive results. Let C ($\overline{C}$) 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 $P(C|-) = 1/4$? $P(\overline{C}|+) = 2/3$ or $P(+|\overline{C}) = 2/3$?

b. What is the sensitivity of this test?

c. Of men who take the PSA test, suppose $P(C) = 0.01$. Find the cell probabilities in the 2×2 table for the joint distribution that cross classifies Y = diagnosis $(+, -)$ with X = true disease status $(C, \overline{C})$.

d. Using (c), find the marginal distribution for the diagnosis.

e. Using (c) and (d), find $P(C|+)$, and interpret.

Solution:

a. False-negative result means that the test turns out to be negative when the person has the disease, so $P(-|C) = \frac{1}{4}$. Of men who did not have it, two-thirds receive positive test results, so $P(+|\overline{C}) = \frac{2}{3}$.

b. Given that a subject has the disease, the probability the diagnostic test is positive is called the sensitivity. So

$$P(+|C) = 1 - P(-|C) = 1 - \frac{1}{4} = \frac{3}{4}.$$

c. $P(C, -) = P(-|C) \cdot P(C) = 0.25 \times 0.01 = 0.0025.$
$P(C, +) = P(+|C) \cdot P(C) = 0.75 \times 0.01 = 0.0075.$
$P(\overline{C}, +) = P(+|\overline{C}) \cdot P(\overline{C}) = 0.67 \times (1 - 0.01) = 0.6633.$

$P(\overline{C}, -) = P(-|\overline{C}) \cdot P(\overline{C}) = 0.33 \times (1 - 0.01) = 0.3267.$

$X \setminus Y$	$+$	$-$	
C	.0075	.0025	.01
$\overline{C}$	.6633	.3267	.99
	.6708	.3292	1

d. By above table, the marginal distribution for the diagnosis are $P(+) = 0.6708$ and $P(-) = 0.3292$.

e. Therefore, $P(C|+) = \frac{P(C,+)}{P(+)} = \frac{.0075}{.6708} = 0.0112$. Given that the test result is positive, there is a 1.11% chance that a man has prostate cancer.

□

2.6 In the United States, the estimated annual probability that a woman over the age of 35 dies of lung cancer equals 0.001304 for current smokers and 0.000121 for nonsmokers [M. Pagano and K. Gauvreau, Principles of Biostatistics, Belmont, CA: Duxbury Press (1993), p. 134].

a. Calculate and interpret the difference of proportions and the relative risk. Which is more informative for these data? Why?

b. Calculate and interpret the odds ratio. Explain why the relative risk and odds ratio take similar values.

Solution:

a. (1) Difference of Proportions: $\pi_1 - \pi_2 = 0.001304 - 0.000121 = 0.001183$. The proportion of women smokers over the age of 35 dying of lung cancer exceeds the proportion of women nonsmokers by 0.001183.

(2) Relative Risk: $R.R. = \frac{\pi_1}{\pi_2} = \frac{0.001304}{0.000121} \approx 10.78$. The proportion of women smokers over the age of 35 dying of lung cancer is about 10.8 times the proportion of women nonsmokers.

Relative risk is more informative for these data because the proportions are very small.

b. Odd Ratio: $\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{0.001304/.998696}{0.000121/.999879} \approx 10.79$. Because the proportions are very small, we have $\theta = \frac{\pi_1}{\pi_2} \cdot \frac{1-\pi_2}{1-\pi_1} \approx R.R. \cdot 1$. So the relative risk and odds ratio take similar values.

□

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. (For data, see R. Dawson, J. Statist. Educ. 3, no. 3, 1995.)

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

b. The odds of survival for females equaled 2.9. For each gender, find the proportion who survived.

c. Find the value of R in the interpretation, "The probability of survival for females was R times that for males."

Solution:

a. An odds ratio of 11.4 does not mean that "the probability of survival for females was 11.4 times that for males". That's the interpretation of a relative risk of 11.4, since that measure is a ratio of proportions rather than odds.

Correct interpretation: the odds of survival for females was 11.4 times that for males. It implies that survival is more likely for females than males.

b. Known that $\frac{\pi_1}{1-\pi_1} = 2.9$ and $\theta = \frac{2.9}{\pi_2/(1-\pi_2)} = 11.4$. We obtain $\pi_1 = \frac{2.9}{1+2.9} \approx .744$ for female and $\pi_2 = \frac{2.9}{11.4+2.9} \approx .203$ for male.

c. $R = \frac{\pi_1}{\pi_2} = \frac{.744}{.203} \approx 3.7$.

□

2.20 In an investigation of the relationship between stage of breast cancer at diagnosis (local or advanced) and a woman's living arrangement, of 144 women living alone, 41.0% had an advanced case; of 209 living with spouse, 52.2% were advanced; of 89 living with others, 59.6% were advanced. The authors reported the P-value for the relationship as 0.02. Reconstruct the analysis they performed to obtain this P-value.

Solution: We compute the 3 × 2 contingency table:

$X \setminus Y$	living alone	with spouse	with others	
advanced	59	109	53	221
local	85	100	34	221
	144	209	89	442

To test for independence, let $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$ for all i, j.
Then $\mu_{11} = n\pi_{11} = \frac{n_{1+}n_{+1}}{n} = \frac{211 \times 144}{442} = 72$.
Similarly, $\mu_{12} = 72$, $\mu_{21} = \mu_{22} = 104.5$, and $\mu_{31} = \mu_{32} = 44.5$.

Likelihood Ratio Statistic

$$\begin{aligned} G^2 &= 2\sum_i \sum_j n_{ij} \log\left(\frac{n_{ij}}{\widehat{\mu}_{ij}}\right) \\ &= 2\left(59\log\left(\frac{59}{72}\right) + 85\log\left(\frac{85}{72}\right) + 109\log\left(\frac{109}{104.5}\right) + 100\log\left(\frac{100}{104.5}\right) + 53\log\left(\frac{53}{44.5}\right) + 36\log\left(\frac{36}{44.5}\right)\right) \\ &\approx 8.375 \end{aligned}$$

Since $\chi^2_{.02}(3) \approx 8.375$, P-value is 0.02 .

Reject H_0. At $\alpha = .05$ significant level, there is sufficient evidence to say that the variables are different on each other.

□

2.22 Table 2.15 classifies a sample of psychiatric patients by their diagnosis and by whether their treatment prescribed drugs.

Diagnosis	Drugs	No Drugs	
Schizophrenia	105	8	113
Affective disorder	12	2	14
Neurosis	18	19	37
Personality disorder	47	52	99
Special symptoms	0	13	13
Total	182	94	276

a. Conduct a test of independence, and interpret the P-value.
b. Obtain standardized residuals, and interpret.
c. Partition chi-squared into three components to describe differences and similarities among the diagnoses, by comparing (i) the first two rows, (ii) the third and fourth rows, (iii) the last row to the first and second rows combined and the third and fourth rows combined.

Solution: a. To test for independence, let $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$ for all i, j.
We compute the μ_{ij} table:

Diagnosis	Drugs	No Drugs
Schizophrenia	74.51	38.49
Affective disorder	9.23	4.77
Neurosis	24.40	12.60
Personality disorder	65.28	33.72
Special symptoms	8.57	4.43

Likelihood Ratio Statistic

$$\begin{aligned} G^2 &= 2\sum_i \sum_j n_{ij} \log\left(\frac{n_{ij}}{\widehat{\mu}_{ij}}\right) \\ &= 2\left[105\log\left(\frac{105}{74.51}\right) + \cdots + 52\log\left(\frac{52}{33.72}\right) + 0 + 13\log\left(\frac{13}{4.43}\right)\right] \approx 96.53 \end{aligned}$$

Since $\chi^2_0(4) \approx 96.53$, P-value is < 0.001 .

Note: We can compute Pearson's Chi-Squared Statistic and verify the value of Likelihood Ratio Statistic by using R as show below:

```
> tbl <- matrix( c(105, 8, 12, 2, 18, 19, 47, 52, 0, 13), nrow=5, byrow=T );
> chisq.test(tbl);

        Pearson's Chi-squared test

data:  tbl
X-squared = 84.1885, df = 4, p-value < 2.2e-16

Warning message:
In chisq.test(tbl) : Chi-squared approximation may be incorrect
> tbl <- matrix( c(105, 8, 12, 2, 18, 19, 47, 52, 13), nrow=1, byrow=T);
> mu  <- matrix( c(74.51, 38.49, 9.23, 4.77, 24.40, 12.60, 65.28, 33.72, 4.43), byrow=T)
> G <- 2 * sum(tbl * log(tbl/mu)); G
[1] 96.53448
```

b. The standardized residuals for the (i, j)th cell is

$$\frac{n_{ij} - \widehat{\mu}_{ij}}{\sqrt{\widehat{\mu}_{ij}(1 - p_{i+})(1 - p_{+j})}}.$$

Using this formula, we can get

$$res_{11} = \frac{105 - 74.51}{\sqrt{74.51(1 - \frac{113}{276})(1 - \frac{182}{276})}} \approx 7.876 \ .$$

Similarly, we can compute all the residuals or we can use R as shown below to complete the table:

```
> n <- matrix( c(105, 8, 12, 2, 18, 19, 47, 52, 0, 13), nrow=5, byrow=T );
> mu  <- matrix( c(74.51, 38.49, 9.23, 4.77, 24.40, 12.60, 65.28, 33.72, 8.57, 4.43),
+ nrow=5, byrow=T );
> SE <- matrix( c((1-113/276)*(1-182/276), (1-113/276)*(1-94/276),
+ (1-14/276)*(1-182/276), (1-14/276)*(1-94/276),
+ (1-37/276)*(1-182/276), (1-37/276)*(1-94/276),
+ (1-99/276)*(1-182/276), (1-99/276)*(1-94/276),
+ (1-13/276)*(1-182/276), (1-13/276)*(1-94/276)), nrow=5, byrow=T );
> res <- (n - mu) / sqrt(mu*SE); res
          [,1]      [,2]
[1,]  7.875924 -7.875227
[2,]  1.603516 -1.603036
[3,] -2.385785  2.385993
[4,] -4.841107  4.840823
[5,] -5.138752  5.136585
```

We conclude that the lack of independence due to all cells. Schizophrenia has significantly higher drugs prescribed rate than independence would predict. Special symptoms has significantly lower drugs prescribed rate.

c. Partition G^2 into 3 parts $G^2 = G_1^2 + G_2^2 + G_3^2$.
(i) For G_1^2, we use only the first 2 rows.

Diagnosis	Drugs	No drugs	
Schizophrenia	105	8	113
Affective disorder	12	2	14
	117	10	127

$$G_1^2 = 2\left[105\log\left(\frac{105\times 127}{113\times 117}\right) + \cdots 2\log\left(\frac{2\times 127}{14\times 10}\right)\right] \approx .75,$$

its $df = (2-1)(2-1) = 1$, $\chi^2_{.05}(1) = 3.84 > G_1^2$. So we fail to reject H_0. There is not enough evident to say that the rate of prescribed drugs is different for schizophrenia and affective disorder.

(ii) For G_2^2, we use only the 3rd and 4th rows.

Diagnosis	Drugs	No drugs	
Neurosis	18	19	37
Personality disorder	47	52	99
	65	71	136

$$G_2^2 = 2\left[18\log\left(\frac{18\times 136}{37\times 65}\right) + \cdots 52\log\left(\frac{52\times 136}{99\times 71}\right)\right] \approx .01,$$

its $df = (2-1)(2-1) = 1$, $\chi^2_{.05}(1) = 3.84 > G_2^2$. So we fail to reject H_0. There is not enough evident to say that the rate of prescribed drugs is different for neurosis and personality disorder.

(iii) For G_3^2, we use the last row, 1st and 2nd rows combined, and the 3rd and 4th rows combined.

Diagnosis	Drugs	No drugs	
SA	117	10	127
NP	65	71	136
Special symptoms	0	13	13
	182	94	276

$$G_3^2 = 2\left[117\log\left(\frac{117\times 276}{127\times 182}\right) + \cdots + 0 + 13\log\left(\frac{13\times 276}{13\times 94}\right)\right] \approx 95.77,$$

its $df = (3-1)(2-1) = 2$, $\chi^2_{.05}(2) = 3.84 < G_3^2$. So we reject H_0. Rate of prescribed drugs is different for the special symptoms and SA and NP.

□