Solution for CDA2015

- **Solution for 1.6.** (a). Note that $-2(L(\pi_0)-L(\hat{\pi})) = 2(y\log\frac{y}{n\pi_0}+(n-y)\log\frac{n-y}{n-n\pi_0})$.

  When $\pi_0 = 1/2, n = 25$ and $y = 0$, we get $2(25\log\frac{25}{12.5}) = 34.7$.

  (b). Note that $u(\pi) = \frac{\partial}{\partial\pi}\log f(y) = \frac{\partial}{\partial\pi}\big(y\log\pi+(n-y)\log(1-\pi)\big) = \frac{y}{\pi}-\frac{n-y}{1-\pi}$.

  So, $u(\pi_0) = -25/(1-1/2) = -50$.

  On the other hand, $\iota(\pi) = -E[\frac{\partial u(\pi)}{\partial\pi}] = -E[-\frac{y}{\pi^2}-\frac{n-y}{(1-\pi)^2}] = \frac{n}{\pi}+\frac{n}{1-\pi} = \frac{n}{\pi(1-\pi)}$.

  So, $\iota(\pi_0) = 25/(0.5(1-0.5)) = 100$ and $z_S^2 = \frac{u^2(\pi_0)}{\iota(\pi_0)} = 25$.

  (c). Since $\hat{\pi} = y/n = 0$ and $\mathrm{Var}(\hat{\pi}) = \pi(1-\pi)/n$, it follows that $\widehat{SE(\hat{\pi})} = \sqrt{0(1-0)/25} = 0$ and $z_W = \frac{\hat{\pi}-\pi_0}{\widehat{SE(\hat{\pi})}} = \infty$.

- **Solution for 1.8.** The null hypothesis is $H_0 : \pi_{10} = 0.75$ and $\pi_{20} = 0.25$.

  From the statement, we know $n = 1103$, $n_1 = 854$ and $n_2 = 249$. The expected numbers of observations under H_0 are: $\mu_1 = 1103\times0.75 = 827.25$ and $n_2 = 1103\times0.25 = 275.75$. So the Pearson test statistic is

  $$\begin{aligned} X^2 &= \sum_{i=1}^{2}\frac{(n_i-\mu_i)^2}{\mu_i} \\ &= \frac{(854-827.25)^2}{827.25}+\frac{(249-275.75)^2}{275.75} \approx 3.46. \end{aligned}$$

  Note that $X^2 \sim \chi^2_{df}$ with $df = 2-1 = 1$. So

  $$p = P(\chi^2_{(1)} > 3.46) = 0.06287 \quad (> 0.05)$$

  and we have no reason to reject H_0, which implies there is statistical evidence (but not strong) to support that the ratio of green to yellow is 3:1.

  Remark: Wald test and likelihood ratio test also work.

- **Solution for 1.30.** (a). Note that the log likelihood function is

  $$L(\theta) = \log\Big\{\theta^{2n_1}[2\theta(1-\theta)]^{n_2}(1-\theta)^{2n_3}\Big\} = n_2\log 2+(2n_1+n_2)\log\theta+(n_2+2n_3)\log(1-\theta)$$

  and that

  $$\frac{\partial L}{\partial\theta} = \frac{2n_1+n_2}{\theta}-\frac{n_2+2n_3}{1-\theta}.$$

  By setting $\partial L/\partial\theta = 0$, we have $\hat{\theta} = (2n_1+n_2)/(2n_1+2n_2+2n_3)$.

  (b). Let $T = -\partial^2 L/\partial\theta^2$. Obviously,

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

  Then

  $$\begin{aligned} ET &= \frac{2En_1+En_2}{\theta^2}+\frac{En_2+2En_3}{(1-\theta)^2} \\ &= \frac{2\theta^2 n+2\theta(1-\theta)n}{\theta^2}+\frac{2\theta(1-\theta)n+2(1-\theta)^2 n}{(1-\theta)^2} = \frac{2n}{\theta(1-\theta)}. \end{aligned}$$

So, the asymptotic standard error of $\hat{\theta}$ is

$$SE(\hat{\theta}) = \Big(- E(\frac{\partial^2 L}{\partial \theta^2})\Big)^{-1/2} = \sqrt{\frac{\theta(1-\theta)}{2n}}.$$

(c). There exist two methods to test whether the probabilities truly have this pattern. One is the Pearson test and the other is the likelihood ratio test.

For Pearson test,

- step 1: Let $\hat{\mu}_1 = n\hat{\theta}^2$, $\hat{\mu}_2 = 2n\hat{\theta}(1-\hat{\theta})$ and $\hat{\mu}_3 = n(1-\hat{\theta})^2$.
- step 2: Define the Person statistic $X^2 = \sum_{i=1}^{3} \frac{(n_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$.
- step 3: If $X^2 \geq \chi^2_d(\alpha)$, then reject H_0. Otherwise, accept H_0.
  Here $\chi^2_d(\alpha)$ denotes the $100 \times (1-\alpha)$-percent quantile of a chi-squared random variable with $df = d$ and $d = (c-1) - p = (3-1) - 1 = 1$.

For likelihood ratio test,

- step 1: Let $L_0 = L(\hat{\theta})$ and $L_1 = \sum_{i=1}^{3} n_i \log(n_i/n)$.
- step 2: Define the likelihood ratio statistic $\Lambda = -2(L_0 - L_1)$.
- step 3: If $\Lambda \geq \chi^2_d(\alpha)$, then reject H_0. Otherwise, accept H_0.

- **Solution for 2.8.** (a). The odds of survival for female is 11.4 times that of male. When the probabilities of survival for both female and male are close to zero, the quoted interpretation is approximately correct.

  (b). $\hat{\Omega}_f = 2.9$, $\hat{\Omega}_m = \left(\frac{\hat{\theta}}{\hat{\Omega}_f}\right)^{-1} = (11.4/2.9)^{-1} \approx 0.254$. So $\hat{\pi}_f = \frac{\hat{\Omega}_f}{1+\hat{\Omega}_f} \approx 74.36\%$ and $\hat{\pi}_m \approx 20.29\%$.

- **Solution for 2.12.**

  $$\hat{\theta}_{AG}(1) = \frac{512 \times 19}{313 \times 89} = 0.35, \quad \hat{\theta}_{AG}(2) = \frac{353 \times 8}{207 \times 17} = 0.80, \quad \hat{\theta}_{AG}(3) = \frac{120 \times 391}{205 \times 202} = 1.13,$$

  $$\hat{\theta}_{AG}(4) = \frac{138 \times 244}{279 \times 131} = 0.92, \quad \hat{\theta}_{AG}(5) = \frac{53 \times 299}{138 \times 94} = 1.22, \quad \hat{\theta}_{AG}(6) = \frac{22 \times 317}{351 \times 24} = 0.83$$

  and

  $$\hat{\theta}_{AG} = \frac{1198 \times 1278}{1493 \times 557} = 1.84.$$

  According to the results above, male students have more possibility to be admitted than female students (since $\hat{\theta}_{AG} = 1.84 > 1$). But for different departments, this conclusion may not be correct. For department A, B and F, female students have more possibility while in department C and E, male students have more possibility.

- **Solution for 2.18.** (a).

Daily Average Number of Cigarettes	Disease Group Lung Cancer Patients	 Control Patients	Sample Odds	Odd Ratio vs. None	Estimated Probability
None	7	61	0.1147541		0.1029412
< 5	55	129	0.4263566	3.7153931	0.2989130
5 - 14	489	570	0.8578947	7.4759399	0.4617564
15 - 24	475	431	1.1020882	9.6039112	0.5242826
25 - 49	293	154	1.9025974	16.5797774	0.6554810
50 +	38	12	3.1666667	27.5952381	0.7600000

$\Omega_1 = 7/61 \approx 0.115$, $\Omega_2 = 55/129 \approx 0.426$, $\Omega_3 = 489/570 \approx 0.858$,

$\Omega_4 = 475/431 \approx 1.102$, $\Omega_5 = 293/154 \approx 1.903$, $\Omega_6 = 38/12 \approx 3.167$. So,

$\theta_1 = \Omega_1/\Omega_2 \approx 0.269$, $\theta_2 = \Omega_1/\Omega_3 \approx 0.134$, $\theta_3 = \Omega_1/\Omega_4 \approx 0.104$,

$\theta_4 = \Omega_1/\Omega_5 \approx 0.060$, $\theta_5 = \Omega_1/\Omega_6 \approx 0.036$.

It is easy to see that as smoking increases, the odds ratio also increases and hence there is a positive trend.

(b). Local odds ratio: $\theta = \frac{\pi_{i1}\pi_{(i+1)2}}{\pi_{i2}\pi_{(i+1)1}} = \frac{\Omega_i}{\Omega_{i+1}} = \frac{e^{\alpha+\beta i}}{e^{\alpha+\beta(i+1)}} = e^{-\beta}$.

(c). No. (see textbook line -14 page 42) ”*Using a retrospective sample, we cannot estimate the probability of lung cancer at each category of smoking behavior. For Table 2.5 we do not know the population prevalence of lung cancer, and the patients suffering it were probably sampled at a rate far in excess of their occurrence in the general population.*”

(d). Note that if $P(X \leq t) \geq P(Y \leq t)$, for all $t \in (-\infty, +\infty)$, then X is called stochastically smaller than Y. By empirical distribution, we have

Daily Average Number of Cigarettes	Lung Cancer Patients	Accumulated Proportion	Control Patients	Accumulated Proportion
None	7	0.0051584	61	0.0449521
< 5	55	0.0456890	129	0.1400147
5 - 14	489	0.4060427	570	0.5600590
15 - 24	475	0.7560796	431	0.8776713
25 - 49	293	0.9719971	154	0.9911570
50 +	38	1.0000000	12	1.0000000

Since the third column is smaller than the fifth column, ”Control Patients” is stochastically smaller than ”Lung Cancer Patients”.