

STA 138: Fall 2006  
Homework 1 Solutions

**Problems:** 1.2, 1.4, 1.6, 1.8, 1.14, 1.16, 1.30

- 1.2 Each of 100 multiple-choice questions on an exam has four possible answers, one of which is correct. For each question, a student guesses by selecting an answer randomly.
- Specify the distribution of the student's number of correct answers.
  - Find the mean and standard deviation of that distribution. Would it be surprising if the student made at least 50 correct responses? Why?
  - Specify the distribution of  $(n_1, n_2, n_3, n_4)$ , where  $n_j$  is the number of times of the student picked choices  $j$ .
  - Find  $E(n_j)$ ,  $var(n_j)$ ,  $cov(n_j, n_k)$ , and  $corr(n_j, n_k)$ .

**Solution:**

- a. The distribution of the student's number of correct answers  $X$  is binomial:  $X \sim B(100, \frac{1}{4})$ . That is

$$P(X = x) = \binom{100}{x} \left(\frac{1}{4}\right)^x \left(1 - \frac{1}{4}\right)^{100-x}, \quad x = 0, 1, \dots, 100.$$

- b. The mean and standard deviation of  $X$  are  $E(X)$  and  $std(X)$  respectively, where

$$E(X) = 100 \times \frac{1}{4} = 25$$

and

$$std(X) = \sqrt{100 \times \frac{1}{4} \times \frac{3}{4}} = 4.33$$

Since  $P(X \geq 50) = 6.63 \times 10^{-8}$ , which is so small, it would be surprising if the student made at least 50 correct responses.

- c. The distribution of  $(n_1, n_2, n_3, n_4)$  is multinomial:  $(n_1, n_2, n_3, n_4) \sim \text{Multi}(100, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$ . That is

$$P(n_1 = x_1, n_2 = x_2, n_3 = x_3, n_4 = x_4) = \frac{100!}{x_1!x_2!x_3!x_4!} \left(\frac{1}{4}\right)^{100}, \quad x_i \geq 0, \sum_{i=1}^4 x_i = 100.$$

- d.  $E(n_j), \text{var}(n_j), \text{cov}(n_j, n_k)$ , and  $\text{corr}(n_j, n_k)$  are computed as follows:

$$\begin{aligned} E(n_j) &= 100 \times \frac{1}{4} = 25. \\ \text{var}(n_j) &= 100 \times \frac{1}{4} \times \frac{3}{4} = 18.75. \\ \text{cov}(n_j, n_k) &= E(n_j n_k) - E(n_j)E(n_k) \\ &= 100 \times 99 \times \frac{1}{4} \times \frac{1}{4} - 25^2 \quad (\text{c.f. 1.14 solution}) \\ &= -6.25. \\ \text{corr}(n_j, n_k) &= \frac{\text{cov}(n_j, n_k)}{\sqrt{\text{var}(n_j)\text{var}(n_k)}} \\ &= -\frac{6.25}{18.75} = -\frac{1}{3}. \end{aligned}$$

1.4 In his autobiography *A Sort of Life*, British author Graham Greene described a period of severe mental depression during which he played Russian Roulette. This “game” consists of putting a bullet in one of the six chambers of a pistol, spinning the chambers to select one at random, and then firing the pistol once at one’s head.

- Greene played this game six times and was lucky that none of them resulted in a bullet firing. Find the probability of this outcome.
- Suppose that he had kept playing this game until the bullet fired. Let  $Y$  denote the number of the game on which it fires. Show the probability mass function for  $Y$ , and justify.

**Solution:**

- $P(\text{fires}) = \frac{1}{6}$ . So,  $P(\text{does not fire}) = \frac{5}{6}$ . He plays the game six times, and from the description we can assume that each game is independent. Hence,  $P(\text{none fire}) = \left(\frac{5}{6}\right)^6 = 0.3349$ .
- $P(Y = y) = \left(\frac{1}{6}\right) \left(\frac{5}{6}\right)^{y-1}$ ,  $y = 1, 2, \dots$ . When the bullet fires for the first time at game  $y$ , then  $y - 1$  games have passed without a bullet firing. This event has probability  $\left(\frac{5}{6}\right)^{y-1}$ . And in the  $y$ ’th game, a bullet fires. This event has probability  $\frac{1}{6}$ . Since the games are independent, we obtain the probability mass function by multiplying the probabilities of the individual events.

1.6 Refer to the vegetarianism example in Section 1.4.3. For testing  $H_0 : \pi = 0.5$  against  $H_a : \pi \neq 0.5$ , show that:

- The likelihood-ratio statistic equals  $2[25 \log(25/12.5)] = 34.7$ .
- The chi-squared form of the score statistic equals 25.0.
- The Wald  $Z$  or chi-squared statistic is infinite.

**Solution:**

- The likelihood-ratio statistic is

$$G^2 = 2 \sum n_j \log(n_j/n\pi_{j0})$$

Since  $n_1 = 0, n_2 = 25, \pi_{10} = \pi_{20} = 0.5$ , then we have

$$G^2 = 2[25 \log(25/12.5)] = 34.7.$$

- The score test statistic is:

$$z_s = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1 - \pi_0)/n}} = -5$$

So the chi-squared form of the score statistic equals 25.0.

- Since  $\hat{\pi} = 0, \pi_0 = 0.5$ , the wald statistic

$$z_w = \frac{\hat{\pi} - \pi_0}{\sqrt{\hat{\pi}(1 - \hat{\pi})/n}} = \infty.$$

Of course, the chi-squared form is also infinite.

1.8 In an experiment on chlorophyll inheritance in maize, for 1103 seedlings of self-fertilized heterozygous green plants, 854 seedlings were green and 249 were yellow. Theory predicts the ratio of green to yellow is 3:1. Test the hypothesis that 3:1 is the true ratio. Report the P-value, and interpret.

**Solution:**

$$\chi^2 = \frac{(854 - 1103 \times 0.75)^2}{1103 \times 0.75} + \frac{(249 - 1103 \times 0.25)^2}{1103 \times 0.25} = 3.46.$$

Degrees of freedom  $df = 1$ , P-value = 0.063. At level  $\alpha=0.05$ , we do not reject the null hypothesis. This result indicates that the 3:1 is the true ratio.

1.14 For the multinomial distribution, show that

$$\text{corr}(n_j, n_k) = -\pi_j \pi_k / \sqrt{\pi_j(1 - \pi_j)\pi_k(1 - \pi_k)}.$$

Show that  $\text{corr}(n_1, n_2) = -1$  when  $c = 2$ .

**Solution:**

First, we have

$$E(n_j) = n\pi_j, \quad \text{var}(n_j) = n\pi_j(1 - \pi_j).$$

$$P(n_j = x_j, n_k = x_k) = \frac{n!}{x_j!x_k!(n - x_j - x_k)!} \pi_j^{x_j} \pi_k^{x_k} (1 - \pi_j - \pi_k)^{n - x_j - x_k}.$$

From this, we have

$$\begin{aligned} E(n_j n_k) &= \sum_{x_j \geq 0, x_k \geq 0, x_j + x_k \leq n} x_j x_k \frac{n!}{x_j!x_k!(n - x_j - x_k)!} \\ &\quad \cdot \pi_j^{x_j} \pi_k^{x_k} (1 - \pi_j - \pi_k)^{n - x_j - x_k} \\ &= n(n - 1)\pi_j \pi_k \sum_{x_j \geq 1, x_k \geq 1, x_j + x_k \leq n - 2} \frac{(n - 2)!}{(x_j - 1)!(x_k - 1)!(n - x_j - x_k)!} \\ &\quad \cdot \pi_j^{x_j - 1} \pi_k^{x_k - 1} (1 - \pi_j - \pi_k)^{n - x_j - x_k} \\ &= n(n - 1)\pi_j \pi_k. \\ \text{cov}(n_j, n_k) &= E(n_j n_k) - E(n_j)E(n_k) \\ &= n(n - 1)\pi_j \pi_k - n^2 \pi_j \pi_k \\ &= -n\pi_j \pi_k. \end{aligned}$$

$$\begin{aligned} \text{Hence } \text{corr}(n_j, n_k) &= \text{cov}(n_j, n_k) / \sqrt{\text{var}(n_j)\text{var}(n_k)} \\ &= -\pi_j \pi_k / \sqrt{\pi_j(1 - \pi_j)\pi_k(1 - \pi_k)}. \end{aligned}$$

If  $c = 2$  then  $\pi_1 + \pi_2 = 1 \implies \pi_2 = 1 - \pi_1$ . Therefore  $\text{corr}(n_1, n_2) = -\pi_1 \pi_2 / \pi_1 \pi_2 = -1$ .

1.16 A likelihood ratio statistic equals  $t_0$ . At the ML estimates, show that the data are  $\exp(t_0/2)$  times more likely under  $H_a$  than under  $H_0$ .

**Solution:**

Let  $l_0$  be the likelihood under  $H_0$  at the ML estimates, and let  $l_1$  be the likelihood under  $H_a$  at the ML estimates (*i.e.* these are the maximum values of the two likelihood functions). We know that

$$t_0 = -2 \log \frac{l_0}{l_1} \implies -t_0/2 = \log \frac{l_0}{l_1} \implies \exp(-t_0/2) = \frac{l_0}{l_1} \implies \exp(-t_0/2)l_1 = l_0$$

So, the data are  $\exp(-t_0/2)$  times more likely under  $H_a$  than under  $H_0$  at the ML estimates.

1.30 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 genotypes.

- Form the log likelihood. Show that  $\hat{\theta} = (2n_1 + n_2)/(2n_1 + 2n_2 + 2n_3)$ .
- Show that  $-\delta^2 L(\theta)/\delta\theta^2 = ((2n_1 + n_2)/\theta^2) + ((n_2 + 2n_3)/(1-\theta)^2)$  and that its expectation is  $2n/\theta(1-\theta)$ . Use this to obtain an asymptotic standard error of  $\hat{\theta}$ .
- Explain how to test whether the probabilities truly have this pattern.

**Solution:**

We know  $(n_1, n_2, n_3)$  is distributed  $Multi(n, \theta^2, 2\theta(1-\theta), (1-\theta)^2)$ . To check that this is a legitimate probability distribution, note that the sum of the probabilities for the three categories is one:  $\theta^2 + 2\theta(1-\theta) + (1-\theta)^2 = (\theta + (1-\theta))^2 = 1$ .

a.

$$\begin{aligned} P(n_1, n_2, n_3) &= \frac{n!}{n_1!n_2!n_3!} (\theta^2)^{n_1} (2\theta(1-\theta))^{n_2} ((1-\theta)^2)^{n_3} \\ &= \frac{n!}{n_1!n_2!n_3!} \theta^{2n_1} 2^{n_2} \theta^{n_2} (1-\theta)^{n_2} (1-\theta)^{2n_3} \\ &= \frac{n!}{n_1!n_2!n_3!} 2^{n_2} \theta^{2n_1+n_2} (1-\theta)^{n_2+2n_3} \end{aligned}$$

We can ignore the first two terms, which do not depend on  $\theta$ , and work with the kernel. So, the log likelihood (kernel only) is:

$$\begin{aligned} L(\theta) &= \log(\theta^{2n_1+n_2} (1-\theta)^{n_2+2n_3}) \\ &= (2n_1 + n_2) \log(\theta) + (n_2 + 2n_3) \log(1-\theta). \end{aligned}$$

To obtain the MLE  $\hat{\theta}$ , take the first derivative of the log likelihood with respect to  $\theta$ , set this equal to zero, and solve for  $\theta$ .

$$\begin{aligned} \delta L(\theta)/\delta\theta &= \frac{2n_1+n_2}{\theta} - \frac{n_2+2n_3}{1-\theta} = 0 \\ \Rightarrow \theta(n_2 + 2n_3) &= (1-\theta)(2n_1 + n_2) \\ \Rightarrow \theta(n_2 + 2n_3) + \theta(2n_1 + n_2) &= (2n_1 + n_2) \\ \Rightarrow (n_2 + 2n_3 + 2n_1 + n_2)\theta &= (2n_1 + n_2) \\ \Rightarrow \hat{\theta} &= (2n_1 + n_2)/(2n_1 + 2n_2 + 2n_3). \end{aligned}$$

- b. Recall that  $E(n_j) = \mu_j = n\pi_j$ , where the  $\pi_j$  are given by the functions of  $\theta$  defined above:  $\mu_1 = n\theta^2$ ,  $\mu_2 = n\theta(1 - \theta)$ ,  $\mu_3 = n(1 - \theta)^2$ .

$$\begin{aligned}
\frac{\delta^2 L(\theta)}{\delta\theta^2} &= -(2n_1 + n_2)\theta^{-2} - (n_2 + 2n_3)(1 - \theta)^{-2} \\
\frac{-\delta^2 L(\theta)}{\delta\theta^2} &= \frac{2n_1 + n_2}{\theta^2} + \frac{n_2 + 2n_3}{(1 - \theta)^2} \\
E\frac{-\delta^2 L(\theta)}{\delta\theta^2} &= \frac{2n\theta^2 + 2n\theta(1 - \theta)}{\theta^2} + \frac{2n\theta(1 - \theta) + 2n(1 - \theta)^2}{(1 - \theta)^2} \\
&= \frac{2n\theta}{\theta^2} + \frac{2n\theta}{(1 - \theta)} + 2n \\
&= 2n \left( \frac{1}{\theta} + \frac{\theta}{(1 - \theta)} + 1 \right) \\
&= 2n \left( \frac{1}{\theta} + \frac{\theta}{(1 - \theta)} + 1 \right) \\
&= \frac{2n}{\theta(1 - \theta)}
\end{aligned}$$

This is the information. The asymptotic variance of  $\hat{\theta}$  is the inverse of the information,  $\theta(1 - \theta)/2n$ . So, the asymptotic standard error of  $\hat{\theta}$  is  $\sqrt{\theta(1 - \theta)/2n}$ .

- c. To test whether the probabilities truly have this pattern, compare the observed  $(n_1, n_2, n_3)$  to the expected counts  $(\mu_1, \mu_2, \mu_3)$  of each genotype with Pearson's chi-squared statistic:

$$X^2 = \sum_j \frac{(n_j - \mu_j)^2}{\mu_j}.$$

The degrees of freedom are  $df = (3 - 1) - 1 = 1$ . The statistic  $X^2$  can be compared to the chi-squared distribution with 1 degree of freedom. One might also perform a likelihood ratio test.