

Assignment 2 Solutions

1. Question 1

(a) Log likelihood

The pmf for $X \sim \text{Bin}(n, p)$ is

$$f(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x} \quad x = 0, 1, \dots, n.$$

Hence the log likelihood function based on observed value x is

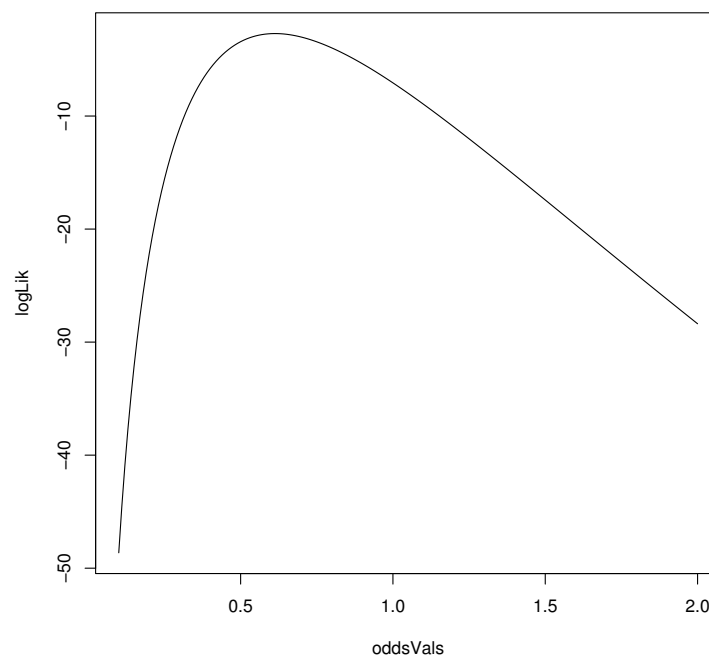
$$l(p; x) = \log f(x; n, p) = \log \binom{n}{x} + x \log(p) + (n-x) \log(1-p).$$

Plugging in $p = \frac{\theta}{1+\theta}$, the log likelihood for θ is

$$l(\theta; x) = \log f(x; n, p) = \log \binom{n}{x} + x \log \left(\frac{\theta}{1+\theta} \right) + (n-x) \log \left(\frac{1}{1+\theta} \right).$$

In the sleep apnea example, $n = 150$ and the number of positive responses $x = 57$. Below is the plot for log likelihood.

Figure 1: Log likelihood for sleep apnea example



(b) **Deriving MLE**

In order to maximize the log likelihood, we can re-organize it as

$$l(\theta; x) = \log \binom{n}{x} + x \log(\theta) + n \log \left(\frac{1}{\theta + 1} \right).$$

We let its first derivative equal to 0, i.e.

$$\frac{\partial l(\theta; x)}{\partial \theta} = 0 + \frac{x}{\theta} - \frac{n}{\theta + 1} = 0.$$

So the MLE for θ can be obtained by solving the above equation, i.e. the MLE is

$$\hat{\theta} = \frac{x}{n - x},$$

as it is not hard to verify that $\frac{\partial^2 l(\theta; x)}{\partial \theta^2} |_{\theta=\hat{\theta}} < 0$.

Note: it can also be derived by plugging in the MLE of p , according to the invariance property of MLE.

(c) **Expected Fisher information**

Based on the formula of Fisher information,

$$I(\theta) = -\frac{\partial^2 l(\theta; x)}{\partial \theta^2} = \frac{x}{\theta^2} - \frac{n}{(\theta + 1)^2}.$$

Hence the expected Fisher information for θ is

$$E \left[\frac{x}{\theta^2} - \frac{n}{(\theta + 1)^2} \right] = \frac{n}{\theta(\theta + 1)^2}.$$

So the estimated variance for $\hat{\theta}$ is

$$\widehat{Var}[\hat{\theta}] = \frac{1}{I(\hat{\theta})} = \frac{\hat{\theta}(\hat{\theta} + 1)^2}{n}.$$

Due to the asymptotic properties of MLE, a large sample 95% C.I. for θ can be approximated by

$$\hat{\theta} \pm Z_{0.025} \sqrt{\frac{\hat{\theta}(\hat{\theta} + 1)^2}{n}},$$

where $\hat{\theta} = \frac{x}{n-x}$ is the MLE.

For the odds of a positive sleep apnea diagnosis, plugging in $\hat{\theta} = \frac{57}{150-57} = 0.613$, $n = 150$ and $Z_{0.025} = 1.96$, we can obtain the large sample C.I. as

$$(0.410, 0.815).$$

2. Question 2

(a) **Approximate C.I. for $a\mu$**

The large sample C.I. for the mean μ for each group can be constructed based on

$$X \sim N(\mu, \mu) \quad \text{approximately.}$$

Based on the additivity of the normal distribution, for any real value a ,

$$aX \sim N(a\mu, a^2\mu) \quad \text{approximately.}$$

Approximating the μ in the variance with $\hat{\mu} = X$, a 95% approximate C.I. for $a\mu$ is

$$aX \pm Z_{0.025} \sqrt{a^2 X}$$

For the OC(yes) group, we have $X = 9$ and $a = 1000/2935$. By plugging in the values, we can obtain the C.I. as

$$(1.063, 5.070).$$

For the OC(no) group, we have $X = 239$ and $a = 1000/135130$. By plugging in the values, we can obtain the C.I. as

$$(1.544, 1.993).$$

(b) **Approximate C.I. for relative risk**

Denote μ_1, μ_2 as the means for the OC(yes) and OC(no) groups respectively, and T_1, T_2 as the corresponding total number of person years. Then the ratio for each group is $\lambda_i = \mu_i/T_i, i = 1, 2$. The relative risk (RR) is then

$$RR = \frac{\lambda_1}{\lambda_2} = \frac{\mu_1 T_2}{\mu_2 T_1}.$$

In order to construct a C.I. for RR , we consider the estimator

$$\widehat{\log(RR)} = \log\left(\frac{X_1}{X_2}\right) - \log\left(\frac{T_1}{T_2}\right),$$

where X_1 and X_2 are number of cases for the 2 groups.

From the results in the class,

$$\widehat{\log(RR)} \sim N\left(\log(RR), \frac{1}{\mu_1} + \frac{1}{\mu_2}\right) \quad \text{approximately.}$$

Hence a 95% approximate C.I. for $\log(RR)$ is

$$\log\left(\frac{X_1}{X_2}\right) - \log\left(\frac{T_1}{T_2}\right) \pm Z_{0.025} \sqrt{\frac{1}{X_1} + \frac{1}{X_2}}.$$

After exponentially transforming back, a 95% approximate C.I. for RR is

$$\left(\frac{X_1 T_2}{X_2 T_1} e^{-Z_{0.025} \sqrt{\frac{1}{X_1} + \frac{1}{X_2}}}, \frac{X_1 T_2}{X_2 T_1} e^{Z_{0.025} \sqrt{\frac{1}{X_1} + \frac{1}{X_2}}} \right).$$

In our example, $X_1 = 9$, $X_2 = 239$, $T_1 = 2935$ and $T_2 = 135130$. Plugging in the numbers, we can obtain the estimated RR as

$$\widehat{\log(RR)} = 1.734,$$

and an approximate 95% C.I. for RR as

$$(0.891, 3.373).$$

3. Question 3

(a) Solutions to text book questions

Exercise 2.2

Assume that

$$X \sim N(\mu, \sigma^2),$$

where both μ and σ^2 are unknown. The exact 95% C.I. for μ is given by

$$\bar{X} \pm t_{0.025}^{(n-1)} \frac{s}{\sqrt{n}}.$$

Using R we can obtain the 95% C.I. for μ as

$$(1.313, 3.062)$$

Exercise 4.11

When the sample size is large, normal approximation can be used for Poisson distributions, i.e. for $Poisson(\mu)$ an approximate 95% C.I. for μ is

$$\bar{X} \pm Z_{0.025} \sqrt{\frac{\bar{X}}{n}}$$

Since the sample size in this example is only 12, a large sample interval is not adequate here.

Using the data, we can obtain the large C.I. as

$$(1.463, 2.912),$$

which is noticeably narrower than the one using a Gaussian assumption.

(b) Goodness of fit

Pearson approach

The test statistic for Pearson approach is

$$\chi^2 = \sum_{i=1}^K \frac{(n_i - e_i)^2}{e_i},$$

where n_i and e_i are observed and expected cell counts corresponding to i th possible outcome with probability p_i .

From the observed counts, we have $K = 6$ and $\mathbf{n} = (n_0, n_1, n_3, n_4, n_5) = (3, 2, 6, 1, 2, 2)$, where n_i is number of decades with i earthquakes.

For $X \sim Poisson(\mu)$, we have

$$p_i = P(X = i) = \frac{e^{-\mu} \mu^i}{i!}, \quad i = 0, 1, \dots, 4$$

and

$$p_5 = 1 - \sum_{l=0}^4 p_l.$$

Hence, we have the expected cell counts as

$$e_i = np_i, \quad i = 0, 1, \dots, 5$$

where $n = 16$ is the total number of decades in our sample.

Plugging in these values, we have

$$\chi^2(6 - 1 - 1) = 4.582326,$$

which indicates a p -value of 0.33. So there is no statistical evidence against using of Poisson distribution for this data set. The model has good fit.

likelihood approach

The test statistic for likelihood approach is

$$G^2 = 2 \sum_{i=1}^K n_i \log \left(\frac{n_i}{e_i} \right),$$

where K , n_i and e_i are the same as the Pearson approach.

Plugging in the values, we have

$$\chi^2(6 - 1 - 1) = 4.985033,$$

which indicates a p -value of 0.29. So there is no statistical evidence against using of Poisson distribution for this data set. The result is consistent with that of the Pearson approach.