STAT 545 Homework 2

Quan Zhou*

September 16, 2015

Problem 1

Assume that $y_1, y_2, \cdots, y_n$ are independent from a Possion distribution.

(a) (1 point) Obtain the likelihood funcrion. Show that the ML estimator $\hat{\mu} = \bar{y}$.

The likelihood function and log-likelihood function is :

$$\begin{aligned} l\left(\mu \mid \underline{y}\right) &= \prod_{i=1}^{n} \exp\left(-\mu\right) \frac{\mu^{y_i}}{y_i!} \\ L\left(\mu \mid \underline{y}\right) &= \sum_{i=1}^{n} \left(-\mu + y_i \log \mu - \log y_i!\right) \\ &= n\bar{y} \log \mu - n\mu - \sum_{i=1}^{n} \log y_i! \end{aligned}$$

Take the first derivative, set it to zero and solve for μ.

$$\frac{dL}{d\mu} = \frac{n\bar{y}}{\mu} - n$$

Then we have the ML estimator for μ or

$$\hat{\mu}_{ML} = \bar{y}$$

(b) (1 point each) Construct a large-sample test statistics for $H_0 : \mu = \mu_0$ using (i) the Wald method, (ii) the score method, (iii) the likelihood-ratio method.

*an adapted version from Quan's original homework.

1. Wald test:
   Wald test is defined as
   $$W = \frac{(\hat{\mu} - \mu_0)^2}{Var(\hat{\mu})}$$

   We estimate the variance of the $\hat{\mu}$ by calculating the inverse of the fisher information evaluated at $\hat{\mu}$ or
   $$I(\mu)\,|_{\mu=\hat{\mu}} = -E\left[\frac{d^2L}{d\mu^2} \mid \mu\right]_{\mu=\hat{\mu}} = E\left[n\bar{y}\mu^{-2} \mid \mu\right]_{\mu=\hat{\mu}} = \frac{n}{\hat{\mu}}$$

   Therefore, the Wald statistics is
   $$W = \frac{n(\bar{y} - \mu_0)^2}{\bar{y}}$$

   We reject the H_0 when $W > \chi^2_{1,\alpha}$

2. Score test:
   $$S(\mu_0) = \frac{\left(\frac{n\bar{y}}{\mu_0} - n\right)^2}{n/\mu_0} = \frac{n(\bar{y} - \mu_0)^2}{\mu_0}$$

   We reject the H_0 when $S > \chi^2_{1,\alpha}$

3. Likelihood ratio test:
   $$\begin{aligned} \sup_{\Theta_0} L &= n\bar{y}\log\mu_0 - n\mu_0 - \sum_{i=1}^{n}\log y_i! \\ \sup_{\Theta_1} L &= n\bar{y}\log\bar{y} - n\bar{y} - \sum_{i=1}^{n}\log y_i! \end{aligned}$$

   Hence
   $$\begin{aligned} -2\log\Lambda &= 2(n\bar{y}\log\bar{y} - n\bar{y} - n\bar{y}\log\mu_0 + n\mu_0) \\ &= 2n(\bar{y}\log\frac{\bar{y}}{\mu_0} + \mu_0 - \bar{y}) \end{aligned}$$

   Reject when $-2\log\Lambda > \chi^2_{1,\alpha}$.

(c) (1 point each) Construct a large-sample confidence interval for μ using (i) the Wald method, (ii) the score method, (iii) the likelihood-ratio method.

1. We simply invert the test to get the confidence interval. For Wald test,

$$\begin{aligned} CI &= \{\mu : z_W = \frac{\sqrt{n}|\bar{y} - \mu|}{\sqrt{\bar{y}}} < z_{\alpha/2}\} \\ &= \{\bar{y} - \sqrt{\bar{y}/n} z_{\alpha/2} < \mu < \bar{y} + \sqrt{\bar{y}/n} z_{\alpha/2}\} \end{aligned}$$

2. For score test,

$$\begin{aligned} CI &= \{\mu : z_S = \frac{\sqrt{n}|\bar{y} - \mu|}{\sqrt{\mu}} < z_{\alpha/2}\} \\ &= \{\frac{2n\bar{y} + z_{\alpha/2}^2 - z_{\alpha/2}\sqrt{z_{\alpha/2}^2 + 4n\bar{y}}}{2n} < \mu < \frac{2n\bar{y} + z_{\alpha/2}^2 + z_{\alpha/2}\sqrt{z_{\alpha/2}^2 + 4n\bar{y}}}{2n}\} \end{aligned}$$

3. For LRT:

$$CI = \{\mu : 2n(\bar{y} \log \frac{\bar{y}}{\mu} + \mu - \bar{y}) < \chi_{1,\alpha}^2\}$$

   No closed-form expressions for the endpoints are available.

Problem 2

(2 point) An investigator wants to estimate the proportion of patients who respond to a new cancer treatment, and he wants his response rate estimate to be reasonably precise. Specifically, he would like to have at least 95 confidence that the precision of the response rate estimate is ± 0.1. On the other hand, he has only sufficient funds to recruit and treat 75 patients in his study. Is 75 patients sufficient to achieve his goal? What alternative information you might want from the investigator in order to answer his question in a potentially more efficient manner?

Let π be the response rate (binomial parameter). $n = 75$ is a fairly large sample so we consider the asymptotic tests. The Fisher information is $\frac{n}{\pi(1-\pi)}$. Hence, the standard error of the MLE estimator is

$$SE = \sqrt{\frac{\pi(1-\pi)}{n}}$$

A 95% confidence interval of is then

$$\pi \in (\hat{\pi} - 1.96\sqrt{\frac{\pi(1-\pi)}{n}}, \hat{\pi} + 1.96\sqrt{\frac{\pi(1-\pi)}{n}})$$

To answer whether ± 0.1 can be a 95% confidence interval, let's try different values of π first. The Figure 1 shows that we cannot give a definite answer at present.

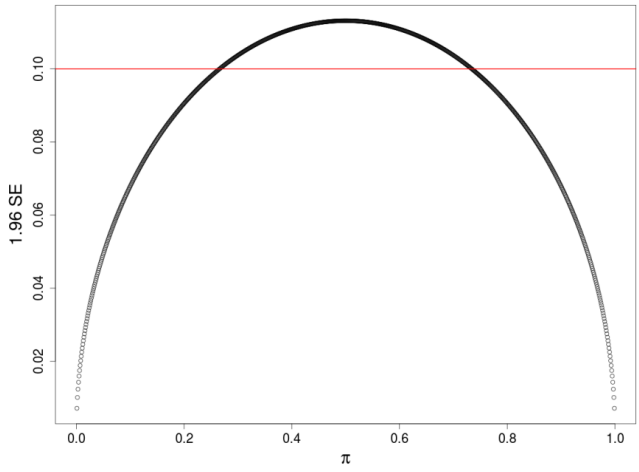

Figure 1: Length of Confidence Interval with respect to π.

What we lack here is an estimate of π. If the investigator can give us an estimate π_0, we can plug π_0 into the expressions above. It seems if $\pi_0 < 0.265$ or $\pi_0 > 0.735$, the 75 sample size is enough for his goal. Therefore we would like to know the range of the true response rate.

Problem 3

Genotypes AA, Aa and aa occur with probabilities $\left[\theta^2, 2\theta\left(1-\theta\right), \left(1-\theta\right)^2\right]$. A multinomial sample of size n has frequencies (n_1, n_2, n_3) of those genotypes.

(a) (2 point) Form the log likelihood. Show that $\hat{\theta} = \left(2n_1 + n_2\right) / \left(2n_1 + 2n_2 + 2n_3\right)$.

The likelihood function is

$$l(\theta) = \theta^{2n_1}[2\theta(1-\theta)]^{n_2}(1-\theta)^{2n_3}$$

Hence, the log-likelihood is

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

Solve the equation:

$$\frac{dL}{d\theta} = \frac{2n_1 + n_2}{\theta} - \frac{2n_3 + n_2}{1-\theta} = 0$$

We obtain

$$\hat{\theta}_{ML} = \frac{2n_1 + n_2}{2n_1 + 2n_2 + 2n_3}$$

(b) (2 point) Show that $-\partial^2 L(\theta)/\partial\theta^2 = [(2n_1 + n_2)/\theta^2] + \left[(n_1 + 2n_3)/(1-\theta)^2\right]$ and that its expectation is $2n/[\theta(1-\theta)]$. Use this to obtain an asymptotic standard error of $\hat{\theta}$.

By differentiating $\frac{\partial L}{\partial \theta}$, we get

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

So,

$$\begin{aligned} I(\theta) &= E[-\frac{\partial^2 L}{\partial \theta^2}] \\ &= \frac{2n\theta^2 + 2n\theta(1-\theta)}{\theta^2} + \frac{2n(1-\theta)^2 + 2n\theta(1-\theta)}{(1-\theta)^2} \\ &= 4n + \frac{2n(1-\theta)}{\theta} + \frac{2n\theta}{1-\theta} \\ &= 2n(2 + \frac{1-\theta}{\theta} + \frac{\theta}{1-\theta}) \\ &= \frac{2n}{\theta(1-\theta)} \end{aligned}$$

By the asymptotic normality of MLE estimator, its standard error is given by

$$SE_{asym}\left(\hat{\theta}\right) = I(\hat{\theta})^{-1/2} = \sqrt{\frac{\hat{\theta}(1-\hat{\theta})}{2n}}$$

(c) (2 point) Explain how to test whether the probabilities truly have this pattern.

The question statement is a little bit ambiguous. Let us consider the null hypothesis H_0 : $p_{AA} = \theta^2, p_{Aa} = 2\theta(1-\theta).p_{aa} = (1-\theta)^2$ with known θ. Then we can quickly compute a Pearson statistic which follows χ^2_2. We can also perform a Wald test and the asymptotic standard error is already given in part (b).

However, I guess the real interest is in the independence of two alleles A and a. That is, we do not know θ, but we want to test whether p_{AA}, p_{Aa}, p_{aa} have such a relationship, which is called Hardy-Weinberg equilibrium. In this case, we estimate θ by $\hat{\theta}_{ML}$ and still compute a Pearson statistic. The degree of freedom is only 1 since we have two constraints on the

three probabilities: 1) the sum is 1 or Hardy-Weinberg equilibrium 2) the usage of ML for estimating θ. The Pearson statistic is

$$X^2 = \sum_{i=1}^{3} \frac{(n_i - \hat{n_i}^2)}{\hat{n}_i} \sim \chi_1^2$$

where $\hat{n_1} = n\hat{\theta}^2, \hat{n_2} = 2n\hat{\theta}(1-\hat{\theta}), \hat{n_3} = n(1-\hat{\theta})^2$. Reject when $X^2 > \chi_1^2(\alpha)$.

Likelihood Ratio test Statistics G

$$G^2 = 2\left(n_1 \log\left(n_1/\left(n\hat{\theta}^2\right)\right) + n_2 \log\left(n_2/\left[2n\hat{\theta}\left(1-\hat{\theta}\right)\right]\right) + n_3 \log\left(n_3/\left(n\left(1-\hat{\theta}\right)^2\right)\right)\right) \sim \chi_1^2$$

where we reject H_0 when $G^2 > \chi_1^2(\alpha)$.