Statistics: Homework 2

Due on Aug 6, 2014

Instructor: Rados Radoicic 6:00 pm

Weiyi Chen

Problem 1

Problem T1: "Exponential"

Distribution of T

Since the pdf of exponential distribution is

$$f_\alpha(x) = \alpha e^{-\alpha x} \tag{1}$$

when $x \geq 0$. Recall that the pdf of gamma distribution is

$$g(x;\alpha,\beta) = \frac{\beta^\alpha x^{\alpha-1} e^{-x\beta}}{\Gamma(\alpha)} \quad \text{for } x \geq 0 \text{ and } \alpha, \beta > 0 \tag{2}$$

Therefore the observations from the exponential distribution $E(\alpha)$ satisfies $X \sim \Gamma(1,\alpha)$. Then, the sum of them

$$T = \sum_{i=1}^{n} X_i \sim \Gamma(n,\alpha) \tag{3}$$

according to the summation property of gamma distribution.

Distribution of S

For $S = 2\alpha T = \sum_{i=1}^{n} 2\alpha X_i$, let $Y_i = 2\alpha X_i$, then

$$F_\alpha(y) = Pr(Y < y) = Pr(2\alpha X < y) = Pr(X < \frac{y}{2\alpha}) = 1 - e^{-y/2} \tag{4}$$

which implies Y_i are observations from the exponential distribution $E(1/2)$, therefore the sum of them

$$S = \sum_{i=1}^{n} Y_i \sim \Gamma(n, \frac{1}{2}) \sim \chi^2_{2n} \tag{5}$$

since the pdf of Chi-squared distribution is

$$f(x;\, k) = \frac{x^{(k/2-1)} e^{-x/2}}{2^{k/2}\Gamma\left(\frac{k}{2}\right)} \tag{6}$$

Confidence interval

To derive the 95% confidence interval for the mean $1/\alpha$, we first find the asymptotic normality, the fisher information is

$$I(\alpha_0) = -E_{\alpha_0}\left[(\frac{\partial^2}{\partial \alpha^2} l_\alpha(x))|_{\alpha=\alpha_0} \right] = \frac{1}{\alpha_0^2} \tag{7}$$

So the asymptotic normality gives:

$$\sqrt{n}(\frac{1}{\overline{X}} - \alpha_0) \to N(0, \alpha_0^2) \tag{8}$$

as $n \to \infty$. Therefore we know,

$$Pr(-c < \sqrt{n}(\frac{1}{\overline{X}} - \alpha)/\alpha < c) = 0.95 \tag{9}$$

where $c = 1.96$ for normal distribution.
For the 95% confidence interval of $1/\alpha$, we just need to rewrite the inequality in the probability bracket above and derive

$$Pr(\overline{X}(1 - \frac{c}{\sqrt{n}}) < \frac{1}{\alpha} < \overline{X}(1 + \frac{c}{\sqrt{n}})) = 0.95 \tag{10}$$

Therefore the interval is

$$\overline{X}(1 - \frac{c}{\sqrt{n}}) < \frac{1}{\alpha} < \overline{X}(1 + \frac{c}{\sqrt{n}}) \tag{11}$$

Problem 2

Problem T3: "Poisson"

Minimize $\alpha_1(\delta) + \alpha_2(\delta)$

Recall the theorem: let δ^* denote a test procedure such that the hypothesis H_0 is not rejected if $af_0(x) > bf_1(x)$ and the hypothesis H_0 is rejected if $af_0(x) < bf_1(x)$. The null hypothesis H_0 can be either rejected or not if $af_0(x) = bf_1(x)$. Then for every other test procedure δ,

$$a\alpha(\delta^*) + b\beta(\delta^*) \leq a\alpha(\delta) + b\beta(\delta) \tag{12}$$

Now we use this theorem for the values of $a = b = 1$, the optimal procedure to reject H_0 if

$$\frac{f_1(X)}{f_0(X)} > 1 \tag{13}$$

for Poisson distribution,

$$f_i(X) = \frac{\exp(-n\lambda_i)\lambda_i^{\sum_i x_i}}{\prod_{i=1}^n (x_i!)} \tag{14}$$

Now we take the ratio of $f_2(X)/f_1(X)$ and then take log on both sides,

$$\log \frac{f_2(X)}{f_1(X)} = \log \frac{\exp(-n\lambda_2)\lambda_2^{\sum_i x_i}}{\exp(-n\lambda_1)\lambda_1^{\sum_i x_i}} = \sum_i x_i \log(\frac{\lambda_2}{\lambda_1}) - n(\lambda_2 - \lambda_1) \tag{15}$$

Since $\lambda_2 > \lambda_1$, it follows that $\frac{f_2(X)}{f_1(X)} > 1$ if and only if $\overline{X}_n > c$.

Find the value of c

Solving the equation above as

$$\sum_i x_i \log(\frac{\lambda_2}{\lambda_1}) - n(\lambda_2 - \lambda_1) > 0 \tag{16}$$

then

$$\overline{X}_n > \frac{\lambda_2 - \lambda_1}{\log(\lambda_2/\lambda_1)} \tag{17}$$

Therefore the value of c is

$$c = \frac{\lambda_2 - \lambda_1}{\log(\lambda_2/\lambda_1)} \tag{18}$$

Determine the minimum value of $\alpha_1(\delta) + \alpha_2(\delta)$

Now if H_i is true then $Y = \sum_i X_i$ will have a Poisson distribution with mean $n\lambda_i$. From last part we have

$$y = n\frac{\lambda_2 - \lambda_1}{\log(\lambda_2/\lambda_1)} = 7.213 \tag{19}$$

The poisson mean is

$$n\lambda_1 = 5 \tag{20}$$

So we find the $\alpha(\delta)$ for $n = 20, \lambda_1 = 0.25$,

$$\alpha_1(\delta) = Pr(Y > 7.213|H_1) = 0.1333 \tag{21}$$

In the same way,

$$\alpha_2(\delta) = Pr(Y \leq 7.213|H_2) = 0.2203 \tag{22}$$

Therefore minimum of $\alpha_1(\delta) + \alpha_2(\delta)$ is

$$\alpha_1(\delta) + \alpha_2(\delta) = 0.3536 \tag{23}$$

Problem 3

Problem T4: ”Goodness of height”

Answer

Recall that the height of the certain large city men follows normal distribution for which the mean is 68 inches and the standard deviation is 1 inch. Let the heights of the 500 men who exist in a certain neighborhood of the city be X, following normal distribution. Let Z be the random variable following normal distribution, the distribution is shown as follows:

Probability of X	Probability of Z	Required probability
Pr(X<66)	Pr(Z<-2)	0.02275
Pr(66<X<67.5)	Pr(-2<Z<-0.5)	0.2858
Pr(67.5<X<68.5)	Pr(-0.5<Z<0.5)	0.3829
Pr(68.5<X<69)	Pr(0.5<Z<2)	0.2858
Pr(X>69)	Pr(Z>2)	0.02275

By using the above table, we express the null hypothesis in the above situation as follows:

$$H_0 : p_i = p_i^0 \tag{24}$$

for all heights come from normal distribution. Against is the following alternative hypothesis:

$$H_1 : p_i \neq p_i^0 \tag{25}$$

at least one height not come from normal distribution. Now we observed following values between the illustrated intervals:
Now we compute the chi-square test statistics as follows:

$$Q = \sum_k \frac{(N_i - np_i^0)^2}{np_i^0} = 27.50 \tag{26}$$

Interval	Value(N_i)	np_i^0
X<66	18	500×0.02275
66<X<67.5	177	500×0.2858
67.5<X<68.5	198	500×0.3829
68.5<X<69	102	500×0.2858
X>69	5	500×0.02275

Then we compute the p-value for the observed test statistic. The decision criterion is: reject null-hypothesis if p-value is less than α. As per the definition, the p-value for the given alternative test statistic would be

$$Pr(\chi_4^2 \leq 27.50) = 1.5749 \times 10^{-5} \tag{27}$$

Hence we have strong evidence that H_0 is false. Thus we can conclude that at least one height not come from the normal distribution.

Problem 4

Problem T5: "NBA"

Answer

We are interested to test the null hypothesis that observed $n = 200$ values follow the binomial distribution. The probability are as follows:

$$\pi_0(\theta) = P_0 = (1-\theta)^4 \tag{28}$$
$$\pi_1(\theta) = P_1 = 4\theta(1-\theta)^3 \tag{29}$$
$$\pi_2(\theta) = P_2 = 6\theta^2(1-\theta)^2 \tag{30}$$
$$\pi_3(\theta) = P_3 = 4\theta^3(1-\theta) \tag{31}$$
$$\pi_4(\theta) = P_4 = \theta^4 \tag{32}$$

The observed values are given as $N_{i=0:4} = 33, 67, 66, 15, 19$ respectively. Consider the following hypothesis for the above situation:

$$H_0 : \text{Observed values follow binomial distribution} \tag{33}$$

Against

$$H_1 : \text{An observed value does not follow binomial distribution} \tag{34}$$

To test the hypothesis, the likelihood function $L(\theta)$ for the observed numbers $N_0, \dots, N_4$ will be

$$L(\theta) = \prod_{i=0}^{4} [\pi_i(\theta)]^{N_i} \tag{35}$$

Taking the logarithm of both sides we get

$$l(\theta) = (N_1 + 2N_2 + 3N_3 + 4N_4)\log\theta + (4N_0 + 3N_1 + 2N_2 + N_3)\log(1-\theta) \tag{36}$$

Now taking the differentiation with respect to θ and let it be 0, we obtain MLE of θ as:

$$\hat{\theta} = \frac{N_1 + 2N_2 + 3N_3 + 4N_4}{4n} = 0.4 \tag{37}$$

Similar to the last problem, by using the MLE of θ and the binomial distribution we compute the probabilities as follows:

Games	N_i	$n\pi_i(\hat{\theta})$
0	33	25.92
1	67	17.28
2	66	11.52
3	15	7.68
4	19	5.12

Now we compute the chi-square test statistics as follows:

$$Q = \sum_k \frac{(N_i - n\pi_i(\hat{\theta}))^2}{n\pi_i(\hat{\theta})} = 47.81 \tag{38}$$

The tail area corresponding to 47.81 is computed by using the chi-square distribution with degree of freedom as $5 - 1 - 1 = 3$, and the p-value is 2.3373×10^{-10}. The p-value is very small therefore we reject the null hypothesis and conclude that the observed value does not follow binomial distribution.

Problem 5

Problem P1: Chapter 7 R-lab.
The package 'fEcofin' is not available currently, we are not able to derive the data.