Tufts University
Department of Mathematics
Spring 2022

MA 166: Statistics

Solutions to Midterm exam (v1.0) [1]

16 March 2022

1. **(25 points) Let $X_1 = k_1, X_2 = k_2, \ldots, X_n = k_n$ be n independent random samples from the geometric distribution $p_X(k;\theta) = (1-\theta)\theta^k$ where $k \in \{0, 1, \ldots, \infty\}$. Find the maximum likelihood estimator $\hat{\theta}$ for the parameter θ.**

   We have

   $$L(\theta) = \prod_{j=1}^{n} p_X(k_j;\theta) = \prod_{j=1}^{n} (1-\theta)\theta^{k_j} = (1-\theta)^n \theta^{n\overline{k}},$$

   where we have defined $\overline{k} := \frac{1}{n}\sum_{j=1}^{n} k_j$. It follows that

   $$\log L(\theta) = n\log(1-\theta) + n\overline{k}\log\theta.$$

   To find the maximum of this function, we set

   $$0 = \frac{\partial}{\partial\theta}\log L(\theta) = -\frac{n}{1-\theta} + \frac{n\overline{k}}{\theta}.$$

   It follows that the estimate for θ, namely θ_e, is given by the solution to

   $$\frac{\theta_e}{1-\theta_e} = \overline{k},$$

   which is

   $$\theta_e = \frac{\overline{k}}{1+\overline{k}},$$

   and so the desired estimator is

   $$\boxed{\hat{\theta}(\vec{k}) = \frac{\frac{1}{n}\sum_{j=1}^{n} k_j}{1+\frac{1}{n}\sum_{j=1}^{n} k_j}.}$$

2. **(40 points) Let $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ be n independent random samples from a pdf of the general form**

   $$f_X(x;\theta) = \exp\left[K(x)p(\theta) + S(x) + q(\theta)\right],$$

   where θ is a parameter. You may also assume that the support of f does not depend on θ.

(a) **(25 points) Show that the estimator**

$$\hat{\boldsymbol{\theta}}(\vec{\boldsymbol{x}}) := \sum_{\boldsymbol{j=1}}^{\boldsymbol{n}} \boldsymbol{K}(\boldsymbol{x_j}),$$

which may or may not be unbiased for $\boldsymbol{\theta}$, is a sufficient estimator.
We form the likelihood function

$$\begin{aligned} L(\theta) &= \prod_{j=1}^{n} f_X(x_j;\theta) \\ &= \prod_{j=1}^{n} \exp\left[K(x_j)p(\theta) + S(x_j) + q(\theta)\right] \\ &= \exp\left[\left(\sum_{j=1}^{n} K(x_j)\right) p(\theta) + \left(\sum_{j=1}^{n} S(x_j)\right) + nq(\theta)\right] \\ &= \exp\left[\hat{\theta}\left(\vec{x}\right) p(\theta) + nq(\theta)\right] \exp\left(\sum_{j=1}^{n} S(x_j)\right) \end{aligned}$$

This is of the form in the second factorization criterion,

$$L(\theta) = g\left[\hat{\theta}(x_1,\ldots,x_n)\,;\,\theta\right]\; b(x_1,\ldots,x_n),$$

where we identify

$$g\left[\hat{\theta}(x_1,\ldots,x_n)\,;\,\theta\right] = \exp\left[\hat{\theta}\left(\vec{x}\right) p(\theta) + nq(\theta)\right],$$

and

$$b(x_1,\ldots,x_n) = \exp\left(\sum_{j=1}^{n} S(x_j)\right).$$

Hence, by the second factorization criterion, $\hat{\theta}$ is sufficient.

(b) **(15 points) Show that the exponential distribution**

$$\boldsymbol{f_X(x;\theta)} = \begin{cases} \boldsymbol{\theta e^{-\theta x}} & \textbf{if } \boldsymbol{x \geq 0} \\ \boldsymbol{0} & \textbf{otherwise} \end{cases}$$

is of the general form described in this problem, and identify the corresponding sufficient estimator.
The exponential distribution has support on $x \geq 0$, which does not depend on the parameter θ. In its region of support, we can write it in the form

$$f_X(x;\theta) = \theta e^{-\theta x} = e^{x(-\theta)+\log\theta}.$$

This is of the form

$$f_X(x;\theta) = \exp\left[K(x)p(\theta) + S(x) + q(\theta)\right],$$

where $K(x) = x$, $p(\theta) = -\theta$, $S(x) = 0$, and $q(\theta) = \log\theta$. Hence the estimator

$$\hat{\theta} = \sum_{j=1}^{n} K(x_j)$$

or

$$\boxed{\hat{\theta} = \sum_{j=1}^{n} x_j}$$

is sufficient.

3. **(35 points) You have reason to believe that X is a normal random variable, but you do not know either its mean or its variance, and you are able to take only three samples of X in your laboratory. Those samples turned out to be $X_1 = x_1 = 1$, $X_2 = x_2 = 2$ and $X_3 = x_3 = 3$.**

   (a) **(7 points) Find the sample mean, $\overline{x}$.**
   The sample mean is

$$\boxed{\overline{x} = \frac{x_1 + x_2 + x_3}{3} = \frac{1+2+3}{3} = 2.}$$

   (b) **(7 points) Find the sample variance, s^2.**
   The sample variance is

$$\boxed{s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + (x_3 - \overline{x})^2}{3-1} = \frac{(-1)^2 + 0^2 + (+1)^2}{3} = 1,}$$

   and hence the sample standard deviation is $s = 1$.

   (c) **(7 points) Find the 90% confidence interval for the actual mean μ, based on the three measurements that you took. You may express your answer in terms of quantities given in the "Some Useful Information" section, if you wish.**
   This confidence interval is

$$\left[\overline{x} - t_{\alpha/2,3-1}\frac{s}{\sqrt{3}}, \overline{x} + t_{\alpha/2,3-1}\frac{s}{\sqrt{3}}\right],$$

   or, since $\overline{x} = 2$, $s = 1$ and $\alpha = 0.1$, the 90% confidence interval is

$$\boxed{\left[2 - \frac{t_{0.05,2}}{\sqrt{3}}, 2 + \frac{t_{0.05,2}}{\sqrt{3}}\right].}$$

   (d) **(7 points) Suppose that your purpose in taking your three samples was to test $H_0 : \mu = 1$ against $H_1 : \mu > 1$ at the $100(1-\alpha)\%$ level of confidence. Give a criterion involving α for accepting or rejecting H_0.**

We want the probability of our making a Type I error to be α. That is, we demand

$$P\left(\text{we reject } H_0 \mid \text{given } H_0 \text{ is true}\right) = \alpha.$$

We form $t = \frac{\overline{x}-\mu_0}{s/\sqrt{n}} = \frac{2-1}{1/\sqrt{3}} = \sqrt{3}$. We reject H_0 if $t \geq t_{\alpha,3-1}$, or

$$\boxed{\text{We reject } H_0 \text{ if } t_{\alpha,2} \leq \sqrt{3}.}$$

(e) **(7 points) Repeat part (d) with the null hypothesis $\boldsymbol{H_0 : \mu = 2}$, and the alternative hypothesis $\boldsymbol{H_1 : \mu \neq 2}$. This time, you should be able to give and justify a definite answer as to whether or not $\boldsymbol{H_0}$ should be rejected, given your three data points.**

This time, our t statistic is $t = \frac{\overline{x}-\mu_0}{s/\sqrt{n}} = \frac{2-2}{1/\sqrt{3}} = 0$. We reject H_0 if either (a) $t \leq -t_{\alpha/2,3-1}$, or (b) $t \geq +t_{\alpha/2,3-1}$. In other words, we reject H_0 if either (a) $0 \leq -t_{\alpha/2,2}$ or $0 \geq +t_{\alpha/2,2}$. Both of these criteria are seen to be the same thing, namely $t_{\alpha/2,2} \leq 0$. This will be the case only if $\alpha/2 \geq 1/2$, or $\alpha \geq 1$. Since this does not make sense, we conclude that we would never reject H_0 in this scenario, no matter what the value of α.