MA 324 — End-semester Examination
Statistical Inference and Multivariate Analysis — 14:00–17:00 IST
IIT GUWAHATI — May 05, 2022

Model Answers of End-semester Examination

1. (5 points) Let $X_1, X_2, \ldots, X_n$ be a random sample from a population with probability density function

$$f(x;\,\alpha) = \begin{cases} \alpha x^{\alpha-1} e^{-x^\alpha} & \text{if } x > 0 \\ 0 & \text{otherwise,} \end{cases}$$

where $\alpha > 0$ is unknown parameter. Show that the maximum likelihood estimator of α exists and is unique.

Solution: The log-likelihood function of α is

$$l(\alpha) = n \ln \alpha + (\alpha - 1) \sum_{i=1}^{n} \ln x_i - \sum_{i=1}^{n} x_i^\alpha \qquad \text{for } \alpha > 0.$$

We need to check if the maximum of $l(\alpha)$ exists and unique on $(0,\,\infty)$. Now,

$$l'(\alpha) = \frac{n}{\alpha} + \sum_{i=1}^{n} \ln x_i - \sum_{i=1}^{n} x_i^\alpha \ln x_i.$$

Clearly, for $\alpha > 0$, $l'(\alpha)$ is a continuous function. Moreover,

$$\lim_{\alpha \to 0+} l'(\alpha) = \infty \qquad \text{and} \qquad \lim_{\alpha \to \infty} l'(\alpha) < 0,$$

as

$$\begin{aligned} \lim_{\alpha \to \infty} l'(\alpha) &= \sum_{i=1}^{n} \ln x_i - \lim_{\alpha \to \infty} x_{(n)}^\alpha \sum_{i=1}^{n} \left(\frac{x_i}{x_{(n)}} \right)^\alpha \ln x_i \\ &= \begin{cases} -\infty & \text{if } x_{(n)} > 1 \\ \sum_{i=1}^{n} \ln x_i & \text{if } x_{(n)} \le 1. \end{cases} \end{aligned}$$

Therefore, by intermediate value property of continuous function, there exists at least one $\alpha > 0$ such that $l'(\alpha) = 0$. Now,

$$l''(\alpha) = -\frac{n}{\alpha^2} - \sum_{i=1}^{n} x_i^\alpha \left(\ln x_i \right)^2 < 0 \quad \text{for all } \alpha > 0.$$

Thus, $l'(\alpha)$ is a strictly decreasing function. Therefore, $l'(\alpha) = 0$ has exactly one solution in $(0,\,\infty)$. Also, this solution maximizes $l(\alpha)$. Hence proved.

2. (5 points) Let X_1 and X_2 be a random sample of size two from the following probability mass function

$$P(X = k) = \begin{cases} (1-p)p^k & \text{if } k = 0,\, 1,\, 2, \ldots \\ 0 & \text{otherwise,} \end{cases}$$

where $p \in (0,\,1)$. To test the hypotheses $H_0 : p = 0.50$ against $H_1 : p = 0.25$, show that the most powerful level α test is given by the test function

$$\Psi(x_1,\,x_2) = \begin{cases} 1 & \text{if } x_1 + x_2 < k_0 \\ \frac{\alpha - 1 + (k_0+2)(0.5)^{k_0+1}}{(k_0+1)(0.5)^{k_0+2}} & \text{if } x_1 + x_2 = k_0 \\ 0 & \text{if } x_1 + x_2 > k_0, \end{cases}$$

where the integer k_0 is such that $1 - (k_0+2)(0.5)^{k_0+1} \le \alpha < 1 - (k_0+3)(0.5)^{k_0+2}$.

Solution: For $p \in (0,\, 1)$, the likelihood function of p is

$$L(p) = (1-p)^2\, p^{x_1+x_2}.$$

Therefore,

$$\frac{L\left(\frac{1}{4}\right)}{L\left(\frac{1}{2}\right)} = 9\left(\frac{1}{2}\right)^{x_1+x_2+2}.$$

Using NP lemma, the critical function of level α MP test is given by

$$\Psi(x_1,\, x_2) = \begin{cases} 1 & \text{if } \frac{L(0.25)}{L(0.50)} > k \iff x_1 + x_2 < k_0 \\ \gamma & \text{if } \frac{L(0.25)}{L(0.50)} = k \iff x_1 + x_2 = k_0 \\ 1 & \text{if } \frac{L(0.25)}{L(0.50)} < k \iff x_1 + x_2 > k_0, \end{cases}$$

where γ and k are such that

$$P_{H_0}\left(X_1 + X_2 < k_0\right) + \gamma P_{H_0}\left(X_1 + X_2 = k_0\right) = \alpha$$

Here, under H_0 and for $k = 0,\, 1,\, 2,\, \ldots$

$$P\left(X_1 + X_2 = k\right) = \sum_{i=0}^{k}\left(\frac{1}{2}\right)^{k+2} = (k+1)\left(\frac{1}{2}\right)^{k+2}.$$

Now, we need to find k_0 such that

$$P_{H_0}\left(X_1 + X_2 < k_0\right) \le \alpha < P_{H_0}\left(X_1 + X_2 = k_0\right) \iff 1 - (k_0+2)\left(\frac{1}{2}\right)^{k_0+1} \le \alpha < 1 - (k_0+3)\left(\frac{1}{2}\right)^{k_0+2},$$

and

$$\gamma = \frac{\alpha - 1 + (k_0+2)\left(\frac{1}{2}\right)^{k_0+2}}{(k_0+1)\left(\frac{1}{2}\right)^{k_0+2}}.$$

Hence proved.

3. (a) (5 points) Let $X_1,\, X_2,\, \ldots,\, X_{10}$ be a random sample of size 10 from a population with probability density function

$$f(x;\theta_1) = \begin{cases} e^{-(x-\theta_1)} & \text{if } x > \theta_1 \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta_1 \in \mathbb{R}$ is unknown parameter. Also, assume that $Y_1,\, Y_2,\, \ldots,\, Y_{10}$ is a random sample of size 10 from a population with probability density function

$$g(x;\theta_2) = \begin{cases} e^{-(x-\theta_2)} & \text{if } x > \theta_2 \\ 0 & \text{otherwise,} \end{cases}$$

where $\theta_2 \in \mathbb{R}$ is unknown parameter. Further assume that two random samples are independent. Construct a 95% confidence interval for $\theta_1 - \theta_2$ based on minimal sufficient statistics of random samples.

Solution: Note that $X_{(1)}$ and $Y_{(1)}$ are minimal sufficient statistics for θ_1 and θ_2, respectively. Moreover, $X_{(1)}$ and $Y_{(1)}$ are independent. Here sample size $n = 10$. Now for $x > 0$,

$$P\left(X_{(1)} > x\right) = e^{-nx} \qquad \text{and} \qquad P\left(Y_{(1)} > x\right) = e^{-nx}.$$

Now, consider the random variable $Z = Z_2 - Z_1$, where $Z_1 = X_{(1)} - \theta_1$ and $Z_2 = Y_{(1)} - \theta_2$. Then for

$x \in \mathbb{R}$,

$$\begin{aligned} P(Z > x) &= P(Z_2 - Z_1 > x) \\ &= \int_0^\infty P(Z_2 - Z_1 > x | Z_1 = y) f_{Z_1}(y) dy \\ &= \int_0^\infty P(Z_2 > x + y) n e^{-ny} dy \\ &= \begin{cases} \int_0^{-x} n e^{-ny} dy + \int_{-x}^\infty n e^{-n(x+y)} e^{-ny} dy & \text{if } x < 0 \\ \int_0^\infty n e^{-n(x+y)} e^{-ny} dy & \text{if } x \geq 0 \end{cases} \\ &= \begin{cases} 1 - \frac{1}{2} e^{nx} & \text{if } x < 0 \\ \frac{1}{2} e^{-nx} & \text{if } x \geq 0. \end{cases} \end{aligned}$$

This shows that the distribution of Z does not depend on θ_1 and θ_2. Thus, Z is a pivot. Now, we need to find a and b such that

$$P(Z \leq a) = 0.025 \quad \text{and} \quad P(X \geq b) = 0.025 \quad \Longrightarrow \quad a = \frac{1}{n} \ln 0.05 \approx -0.3 \quad \text{and} \quad b = -\frac{1}{n} \ln 0.05 \approx 0.3.$$

Therefore, a 95% confidence interval for $\theta_1 - \theta_2$ is $\left[X_{(1)} - Y_{(1)} + \frac{1}{n} \ln 0.05,\ X_{(1)} - Y_{(1)} - \frac{1}{n} \ln 0.05\right]$.

(b) (2 points) If

(3.45, 3.62, 3.77, 8.55, 4.40, 5.36, 4.34, 3.10, 3.26, 3.97)

and

(0.57, 0.55, 0.72, 1.21, 0.68, 1.08, 0.71, 1.59, 1.78, 0.65)

are realizations of the first and second random samples, respectively, compute the confidence interval that you obtain in part (a). You need to provide numerical limits of the confidence interval.

Solution: For the given data, the observed values of $X_{(1)}$ and $Y_{(1)}$ are 3.10 and 0.55, respectively. Therefore, based on the samples, a 95% confidence interval for $\theta_1 - \theta_2$ is [2.255, 2.85].

4. A researcher is investigating the use of a windmill to generate electricity. The researcher collected data on the DC output (y) of a windmill and the corresponding average wind velocity (x) in miles per hour for 5 consecutive days. The data is given in the following table. The preliminary aim of the researcher is to fit a linear regression model considering DC output as response and average wind velocity as regressor.

x	y
5.00	1.58
6.00	1.82
3.40	1.05
2.70	0.50
10.00	2.23

For all the parts in this question, please write the steps clearly mentioning statistical modeling and expressions. No need to derive any estimator or tests.

(a) (3 points) Compute the coefficients of a linear regression using the above data.

Solution: Here $\overline{x} = 5.42$, $\overline{y} = 1.436$, $S_{xy} = 7.1244$, and $S_{xx} = 32.968$. Therefore, $\widehat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \approx 0.2161$ and $\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} \approx 0.2647$.

(b) (3 points) Determine the coefficient of determination. Interpret the result.

Solution: Here, $SS_T \approx 1.8237$. Therefore, $R^2 = \frac{SS_{Reg}}{SS_T} = \frac{\widehat{\beta}_1 S_{xy}}{SS_T} \approx 0.8442$. As the value of R^2 is large, it shows that the model fits the data quite well.

(c) (2 points) Test, at level 0.05, the significance of the linear regression, by stating null and alternative hypotheses clearly.

Solution: Here we want to test $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$. The test statistics is

$$t = \frac{\widehat{\beta}_1}{\sqrt{MS_{Res}/S_{xx}}},$$

which follows a t-distribution with $n-2=3$ degrees of freedom. Now the observed value of t is 4.032 and $t_{3;0.025}=4.54$. Thus, observed value of t is less than $t_{3;0.025}$. Hence, the null hypothesis is accepted. Therefore, this regression is not significant.

(d) (2 points) Suppose that the weather forecast says that the average wind speed for tomorrow will be 7 miles per hour. Find the 99% prediction interval of DC output for tomorrow.

Solution: A 99% prediction interval for DC output for wind speed 7 miles per hour is

$$\left[\widehat{y}_0 \mp t_{3;0.005}\sqrt{MS_{Res}\left(1+\frac{1}{5}+\frac{(7-5.42)^2}{32.968}\right)}\right] \approx [-0.253, 3.808].$$

5. (3 points) Let $\boldsymbol{X}=(X_1,\ldots,X_p)'$ be a p-dimensional random vector with variance-covariance matrix Σ and $\boldsymbol{Y}=(Y_1,\ldots,Y_p)'$ be the corresponding principal components. Show that correlation coefficient between Y_i and X_k is $\frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}$, where σ_{kk} is the kth diagonal entry of Σ, λ_i is the ith largest eigenvalue of Σ, and e_{ik} is the ith entry of the eigenvector of Σ corresponding to the eigenvalue λ_i.

Solution:

$$\begin{aligned}
\rho_{Y_i,X_k} &= \frac{Cov\,(Y_i,\,X_k)}{\sqrt{Var(Y_i)Var(X_k)}} \\
&= \frac{Cov\,(\boldsymbol{e}_i'\boldsymbol{X},\,\boldsymbol{a}'\boldsymbol{X})}{\sqrt{\lambda_i\sigma_{kk}}} \quad \text{where, } \boldsymbol{a} \text{ is the } k\text{-th unit vector in } \mathbb{R}^n. \\
&= \frac{\boldsymbol{a}'\Sigma\boldsymbol{e}_i}{\sqrt{\lambda_i\sigma_{kk}}} \\
&= \frac{\boldsymbol{a}'\lambda_i\boldsymbol{e}_i}{\sqrt{\lambda_i\sigma_{kk}}} \\
&= \frac{eik\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}}.
\end{aligned}$$

Hence proved.

6. Let U_1 and U_2 be two independent uniform random variables on the interval $(0,\,1)$. Suppose that $\boldsymbol{X}=(X_1,\,X_2,\,X_3)'$, where $X_1=U_1$, $X_2=U_2$, and $X_3=U_1+U_2$.

(a) (2 points) Compute the correlation matrix $\boldsymbol{\rho}$ of $\boldsymbol{X}$.

Solution: The correlation matrix is

$$\boldsymbol{\rho}=\begin{bmatrix} 1 & 0 & \frac{1}{\sqrt{2}} \\ 0 & 1 & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 1 \end{bmatrix}.$$

(b) (2 points) Find the variance of first principal component based on the correlation matrix that you find in part(a).

Solution: The characteristic equation of $\boldsymbol{\rho}$ is $\lambda(\lambda-1)(\lambda-1)=0$. Therefore, the eigenvalues are 0, 1, and 2. Thus, the variance of the first principal component, which is the largest eigenvalue, is 2.

7. Prove or disprove the following statements. Note that you need to provide a sequence of logic to prove a statement. To disprove a statement, a counter example must be given.

(a) (3 points) Let Θ_1 and Θ_2 be two parametric spaces such that $\Theta_1 \subset \theta_2$. Then an unbiased estimator of $\theta \in \Theta_1$ is also unbiased estimator of $\theta \in \Theta_2$.

Solution: Let $X \sim N(\theta,\,1)$. Take $\Theta_1=\{0\}$ and $\Theta_2=\mathbb{R}$. Clearly, $\Theta_1 \subset \Theta_2$. Take $T=0$. Then, $E(T)=0$ for all $\theta \in \mathbb{R}$. Thus, $E(T)=\theta$ for all $\theta \in \Theta_1$. But, $E(T) \neq \theta$ for some $\theta \in \Theta_2$. Thus, T is unbiased for $\theta \in \Theta_1$, but biased for $\theta \in \Theta_2$. Hence, the given statement is not true.

(b) (3 points) A minimal sufficient statistic is complete.

Solution: Let $X_1, X_2, \ldots, X_n$ be random sample from $N\left(\theta, \theta^2\right)$, where $\theta > 0$ and $n > 1$. Then $\boldsymbol{T} = \left(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2\right)$ is minimal sufficient statistic. Now,

$$E\left[\frac{1}{2n}\sum_{i=1}^n X_i^2 - \frac{1}{n(n+1)}\left(\sum_{i=1}^n X_i\right)^2\right] = 0.$$

However,

$$P\left(\frac{1}{2n}\sum_{i=1}^n X_i^2 - \frac{1}{n(n+1)}\left(\sum_{i=1}^n X_i\right)^2\right) = 0 \neq 1.$$

Therefore, $\boldsymbol{T}$ is not complete. Thus, the given statement is not true.

Useful Information

The following table gives the values of upper α-points of a t distribution with f degrees of freedom.

	$\alpha = 0.005$	$\alpha = 0.025$	$\alpha = 0.010$	$\alpha = 0.050$
$f = 3$	5.84	4.54	3.18	2.35
$f = 4$	4.60	3.74	2.78	2.13
$f = 5$	4.03	3.36	2.57	2.01