Statistical Inference III

Final exam: [+50 points]

2017/1/9 (Mon) 13:30-16:50

Q1
Q2
Q3
Q4

YOUR NAME Jia-Han Shih

NOTE1: Please write down the derivation of your answer very clearly for all questions. The score will be reduced if you only write the answer or if the derivation is not clear. The score will be given even when your answer has a minor mistake but the derivations are clearly stated.

Q1. [+10] Let $X_1,\ldots,X_n \overset{iid}{\sim} p_\theta(x) = f(x-\theta)$, where f is a known symmetric pdf s.t. $Var(X_1) = \sigma_f^2 < \infty$. Consider testing $H_0 : \theta = 0$ under the local alternatives $\theta_n = h/\sqrt{n}$ for a real number h.

(i) [+2] Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\sigma_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$. Derive the asymptotic distribution of the *t*-statistics $t_n = \sqrt{n}\bar{X}_n / \sigma_n$ under the local alternatives.

Answer:

Under the local alternatives, by the CLT, we have

$$\frac{\sqrt{n}(\bar{X}_n - \theta_n)}{\sigma_f} \xrightarrow{d} N(0,1), \quad \text{as} \quad n \to \infty.$$

By the Taylor expansion,

$$\sqrt{n}\theta_n \xrightarrow{p} h, \quad \text{as} \quad n \to \infty,$$

and the fact that sample variance will converge in probability to population variance

$$\frac{\sigma_n}{\sigma_f} \xrightarrow{p} 1, \quad \text{as} \quad n \to \infty.$$

Then by Slusky's theorem,

$$t_n = \frac{\sqrt{n}\bar{X}_n}{\sigma_n} = \left\{ \frac{\sqrt{n}(\bar{X}_n - \theta_n)}{\sigma_f} + \frac{\sqrt{n}\theta_n}{\sigma_f} \right\} \frac{\sigma_f}{\sigma_n} \xrightarrow{d} N\left(\frac{h}{\sigma_f}, 1 \right), \quad \text{as} \quad n \to \infty.$$

Hence we have derived the asymptotic distribution of *t*-statistics t_n under the local alternatives.

(ii) [+4] Derive the asymptotic distribution of the sign statistics $\sqrt{n}S_n^2$ under the local alternatives, where $S_n^2 = \frac{1}{n}\sum_{i=1}^{n}\left\{I(X_i > 0) - \frac{1}{2}\right\}$.

Answer:

Under the local alternatives,

$$E\left\{I(X_i > 0) - \frac{1}{2}\right\} = \Pr(X_i > 0) - \frac{1}{2} = \Pr(X_i - \theta_n > -\theta_n) - \frac{1}{2}$$

$$= 1 - \Pr(X_i - \theta_n \le -\theta_n) - \frac{1}{2} = \frac{1}{2} - F(-\theta_n) = F(0) - F(-\theta_n),$$

where $F(0) = 1/2$ due to that f is a known symmetric probability density function. Then consider the Taylor expansion of $F(-\theta_n)$ around 0,

$$F(-\theta_n) \approx F(0) - \theta_n f(0).$$

Thus, we obtain

$$\sqrt{n}\{F(0) - F(-\theta_n)\} \approx \sqrt{n}\{F(0) - F(0) + \theta_n f(0)\} \xrightarrow{p} hf(0), \quad \text{as } n \to \infty.$$

By the CLT,

$$\sqrt{n}S_n^2 = \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{I(X_i > 0) - \frac{1}{2}\right\} - F(0) + F(-\theta_n)\right] \xrightarrow{d} N\left(0, \frac{1}{4}\right), \quad \text{as } n \to \infty,$$

where variance is derived as

$$\{1 - F(0)\}F(0) = \frac{1}{4}.$$

Therefore, by Slusky's theorem,

$$\sqrt{n}S_n^2 = \sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{I(X_i > 0) - \frac{1}{2}\right\} - F(0) + F(-\theta_n)\right] + \sqrt{n}\{F(0) - F(-\theta_n)\}$$

$$\xrightarrow{d} N\left(hf(0), \frac{1}{4}\right), \quad \text{as } n \to \infty.$$

Hence we have derived the asymptotic distribution of sign test statistics $\sqrt{n}S_n^2$ under the local alternatives.

(iii) [+2] Derive the Pitman ARE $e_{s,t}$

Answer:

The Pitman ARE is

$$e_{s,t} = \lim_{\theta_n \to \theta} \frac{N_t(\alpha, \beta, \theta_n)}{N_s(\alpha, \beta, \theta_n)} = \left(\frac{\mu_s'(0)/\sigma_s}{\mu_t'(0)/\sigma_t} \right)^2 = \left(\frac{2f(0)}{1/\sigma_f} \right)^2 = 4f(0)^2\sigma_f^2 .$$

(iv) [+2] Let f be the pdf of $N(0,1)$.

Compute the Pitman ARE and interpret the value of ARE.

Answer:

If $f \sim N(0,1)$, we have

$$f(0)^2 = \left(\frac{1}{\sqrt{2\pi}} \right)^2 = \frac{1}{2\pi} \qquad \text{and} \qquad \sigma_f^2 = 1.$$

Then the Pitman ARE is

$$e_{s,t} = \frac{4}{2\pi} = \frac{2}{\pi} \approx 0.637 .$$

The test based on t_n is $\pi/2$ more efficient than the test based on S_n^2.

Q2 [+12] Kolmogorov-Smirnov test

To answer the questions below, please define symbols and notations by yourself.

(i) [+1] Define the Kolmogorov-Smirnov statistic (T_n) for testing $H_0 : F = F_0$ vs. $H_1 : F \neq F_0$.

Answer:

Suppose $X_1, X_2, \cdots, X_n \overset{i.i.d.}{\sim} F$, the Kolmogorov-Smirnov (K-S) statistics for testing $H_0 : F = F_0$ vs. $H_1 : F \neq F_0$ is defined as

$$T_n = \sqrt{n} \sup_{t \in \mathrm{R}} | \hat{F}_n(t) - F_0(t) |, \quad \text{where } \hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq t).$$

(ii) [+2] Explain how to compute the critical value ($s_{n,1-\alpha}$) for level α test when F_0 is continuous.

Answer:

When F_0 is continuous, the K-S statistics is distribution free. Therefore, consider the simple uniform distribution ($U_1, U_2, \cdots, U_n \overset{i.i.d.}{\sim} U(0,1)$), the K-S statistics follows

$$T_n = \sqrt{n} \sup_{0<u<1} | \hat{G}_n(u) - u |, \quad \text{where } \hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} I(U_i \leq t).$$

The algorithm of computing the critical value $s_{n,1-\alpha}$ such that $\Pr(T_n > s_{n,1-\alpha}) = \alpha$ is given as follows:

Algorithm (compute the critical value)

Step 1. Generate $U_1^{\ell}, U_2^{\ell}, \cdots, U_n^{\ell} \overset{i.i.d.}{\sim} U(0,1)$.

Step 2. Order the samples $U_{(1)}^{\ell}, U_{(2)}^{\ell}, \cdots, U_{(n)}^{\ell}$.

Step 3. Compute the K-S statistics

$$T_n^{\ell} = \max_{1 \leq i \leq n} \left\{ \frac{i}{n} - U_{(i)}^{\ell}, U_{(i)}^{\ell} - \frac{i-1}{n}, i = 1, 2, \cdots, n \right\}.$$

Step 4. Repeat Step 1 – 3 k times with large k ($\ell = 1, 2, \cdots, k$).

Step 5. Order the K-S statistics $T_n^{(1)}, T_n^{(2)}, \cdots, T_n^{(k)}$.

Step 6. Approximate $s_{n,1-\alpha}$ by $1-\alpha$ percentile of $T_n^{(1)}, T_n^{(2)}, \cdots, T_n^{(k)}$.

(iii) [+3] Show that the test is pointwise consistent in power for $\forall F \neq F_0$.

Answer:

Since the statement $F \neq F_0$ is the same as that there exists t_0 such that $F(t_0) \neq F_0(t_0)$. It is also equivalent to $T_n = \sup_{t \in \mathrm{R}} |F(t) - F_0(t)| > 0$. Therefore, under the alternative hypothesis ($\forall F \neq F_0$), we obtain

$$\begin{aligned}\sqrt{n}\{\hat{F}_n(t_0) - F_0(t_0)\} &= \sqrt{n}\{\hat{F}_n(t_0) - F(t_0)\} + \sqrt{n}\{F(t_0) - F_0(t_0)\} \\ &\xrightarrow{d} N(0, F(t_0)\{1 - F(t_0)\}) + \begin{cases} \infty, & \text{if } F(t_0) > F_0(t_0), \\ -\infty, & \text{if } F(t_0) < F_0(t_0). \end{cases}\end{aligned}$$

Then the power is

$$\begin{aligned}\Pr(T_n > s_{n,1-\alpha}) &= \Pr(\sqrt{n} \sup_{t \in \mathrm{R}} |\hat{F}_n(t) - F_0(t) > s_{n,1-\alpha}) \\ &\geq \Pr(\sqrt{n}\,|\hat{F}_n(t_0) - F_0(t_0) > s_{n,1-\alpha}) \to 1, \quad \text{as } n \to \infty.\end{aligned}$$

Hence we have shown that the K-S test is pointwise consistent in power for $\forall F \neq F_0$.

(iv) [+3] Based on data $X = (2, 4, 6, 8, 10)$, perform a level $\alpha = 0.05$ test $H_0 : F = F_0$ vs. $H_1 : F \neq F_0$ for $F_0 = \mathrm{U}(0, 10)$ using the asymptotic critical value $s_{1-0.05} = 1.4$.

Answer:

We order the data as $X_{(1)} = 2, X_{(2)} = 4, X_{(3)} = 6, X_{(4)} = 8, X_{(5)} = 10$ and the CDF of uniform distribution $U(0, 10)$ is

$$F_0(u) = \frac{u}{10}, \qquad 0 < u < 10.$$

Then the K-S test statistics is derived as follows:

$$\begin{aligned}T_5 &= \max_{1 \leq i \leq 5} \left\{ \frac{i}{5} - F(X_{(i)}), F(X_{(i)}) - \frac{i-1}{5}, i = 1, 2, \cdots, 5 \right\} \\ &= \max \left\{ \frac{1}{5} - \frac{2}{10}, \frac{2}{10} - 0, \frac{2}{5} - \frac{4}{10}, \frac{4}{10} - \frac{1}{5}, \frac{3}{5} - \frac{6}{10}, \frac{6}{10} - \frac{2}{5}, \frac{4}{5} - \frac{8}{10}, \frac{8}{10} - \frac{3}{5}, \right. \\ &\qquad \left. \frac{5}{5} - \frac{10}{10}, \frac{10}{10} - \frac{4}{5} \right\} \\ &= \max \left\{ \frac{2}{10}, \frac{2}{10}, \frac{2}{10}, \frac{2}{10}, \frac{2}{10} \right\} = \frac{2}{10} = 0.2.\end{aligned}$$

Since $T_5 = 0.2 < 1.4 = s_{n,1-\alpha}$, we do not reject the null hypothesis $H_0 : F = F_0$.

(v) [+3] Based on the above data, draw the 95% confidence band for F. Check whether the null distribution $F_0 = \mathrm{U}(0,10)$ is inside the confidence band or not.

Answer:

Yes, Figure 1 reveals that $F_0(u) = u/10$ is inside the 95% confidence band (CB) of the K-S statistics based on the data in **(iv)**.

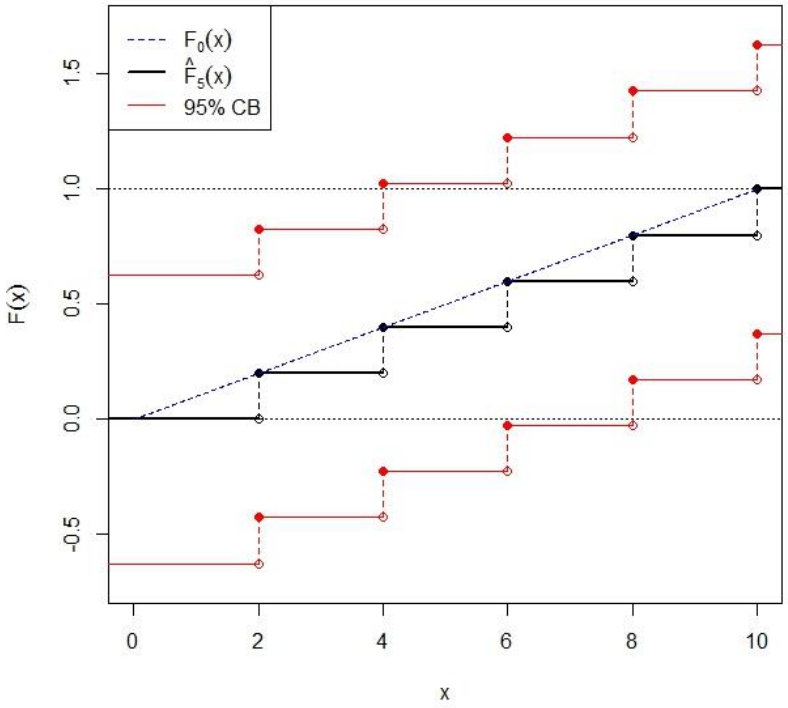

Figure 1. 95% CB of the K-S statistics based on the data in **(iv)**.

Q3[+12] Pearson's χ^2 Test

Let $(Y_1, \ldots, Y_{k+1}) \sim \text{Multi}(n; p_1, \ldots, p_{k+1})$. Consider testing the hypothesis

$$H_0 : p_j = \pi_j, \quad \forall j = 1, \ldots, k+1 \quad \text{vs.} \quad H_1 : p_j \neq \pi_j, \quad \exists j = 1, \ldots, k+1,$$

where π_j's are known.

(1)[+2] Define Pearson's χ^2-test with level α.

Answer:

Pearson's chi-square statistics is defined as

$$Q_n = \sum_{j=1}^{k} \frac{(Y_j - n\pi_j)^2}{n\pi_j}.$$

Then a level α Pearson's chi-square test is

$$\phi = \begin{cases} 1, & \text{if } Q_n > \chi^2_{k,1-\alpha}, \\ 0, & \text{if } Q_n \leq \chi^2_{k,1-\alpha}. \end{cases}$$

(2) [+4] Show that the χ^2-test is pointwise consistent in power.

Answer:

Since the statement $p_j \neq \pi_j$, for some j is equivalent to that there exists ℓ such that $p_\ell \neq \pi_\ell$. Therefore, under the alternative hypothesis (H_1), the power is

$$\begin{aligned} \Pr_{H_1}(Q_n > \chi^2_{k,1-\alpha}) &= \Pr_{H_1}\left(\sum_{j=1}^{k} \frac{(Y_j - n\pi_j)^2}{n\pi_j} > \chi^2_{k,1-\alpha} \right) \geq \Pr_{H_1}\left(\frac{(Y_\ell - n\pi_\ell)^2}{n\pi_\ell} > \chi^2_{k,1-\alpha} \right) \\ &= \Pr_{H_1}\left(\frac{n(Y_\ell/n - \pi_\ell)^2}{\pi_\ell} > \chi^2_{k,1-\alpha} \right). \end{aligned}$$

In addition, we have

$$\lim_{n\to\infty} \frac{n(Y_\ell/n - \pi_\ell)^2}{\pi_\ell} = \lim_{n\to\infty} \frac{n(p_\ell - \pi_\ell)^2}{\pi_\ell} = \infty.$$

Then

$$\Pr_{H_1}(Q_n > \chi^2_{k,1-\alpha}) \geq \Pr_{H_1}\left(\frac{n(Y_\ell/n - \pi_\ell)^2}{\pi_\ell} > \chi^2_{k,1-\alpha} \right) \to 1, \quad \text{as } n \to \infty.$$

Hence we have proven that Pearson's chi-square test is pointwise consistent in power.

(3) [+6] Derive the asymptotic distribution of the χ^2-statistics

Answer:

Under the null hypothesis $H_0 : p_j = \pi_j, \quad \forall j = 1, \ldots, k+1$, we define

$$\mathbf{V}_n = \sqrt{n}\left[\frac{Y_1}{n} - \pi_1 \quad \cdots \quad \frac{Y_k}{n} - \pi_k\right]^{\mathrm{T}}.$$

Since Y_j is the sum of i.i.d. Bernoulli trials and

$$E\left(\frac{Y_j}{n}\right) = p_j, \qquad \operatorname{cov}\left(\frac{Y_i}{n}, \frac{Y_j}{n}\right) = \begin{cases} \pi_i(1-\pi_i), & \text{if } i = j, \\ -\pi_i\pi_j \quad , & \text{if } i \neq j. \end{cases}$$

By multivariate CLT, we obtain

$$\mathbf{V}_n \xrightarrow{d} \mathrm{MVN}_k(\mathbf{0}, \boldsymbol{\Sigma}), \qquad \text{where } \boldsymbol{\Sigma} = \begin{bmatrix} \pi_1 & & 0 \\ & \ddots & \\ 0 & & \pi_k \end{bmatrix} + \begin{bmatrix} \pi_1 \\ \vdots \\ \pi_k \end{bmatrix}\begin{bmatrix} \pi_1 & \cdots & \pi_k \end{bmatrix}.$$

We define

$$\mathbf{D} \equiv \begin{bmatrix} \pi_1 & & 0 \\ & \ddots & \\ 0 & & \pi_k \end{bmatrix} \qquad \text{and} \qquad \mathbf{1} \equiv \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}^{\mathrm{T}}.$$

With these notations, we can rewrite $\boldsymbol{\Sigma}$ as $\boldsymbol{\Sigma} = \mathbf{D} + \mathbf{D}\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathbf{D}$. Then consider

$$\mathbf{V}_n^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{V}_n \xrightarrow{d} \chi_k^2, \qquad \text{as } n \to \infty, \text{ where } \boldsymbol{\Sigma}^{-1} = \mathbf{D}^{-1} + \frac{\mathbf{1}\mathbf{1}^{\mathrm{T}}}{\pi_{k+1}}.$$

Claim:

$$Q_n = \sum_{j=1}^{k} \frac{(Y_j - n\pi_j)^2}{n\pi_j} = \mathbf{V}_n^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{V}_n.$$

It suffices to prove the Claim then we have shown that under the null hypothesis, the asymptotic distribution of Pearson's chi-square statistics follows the chi-square distribution with degree of freedom k.

Proof of Claim:

We have

$$\mathbf{V}_n^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{V}_n = n\left[\frac{Y_1}{n} - \pi_1 \quad \cdots \quad \frac{Y_k}{n} - \pi_k\right]\left(\mathbf{D}^{-1} + \frac{\mathbf{1}\mathbf{1}^{\mathrm{T}}}{\pi_{k+1}}\right)\begin{bmatrix} \frac{Y_1}{n} - \pi_1 \\ \vdots \\ \frac{Y_k}{n} - \pi_k \end{bmatrix}.$$

First we consider

$$n\begin{bmatrix}\frac{Y_1}{n}-\pi_1 & \cdots & \frac{Y_k}{n}-\pi_k\end{bmatrix}\mathbf{D}^{-1}\begin{bmatrix}\frac{Y_1}{n}-\pi_1\\ \vdots\\ \frac{Y_k}{n}-\pi_k\end{bmatrix}$$

$$=n\begin{bmatrix}\frac{Y_1}{n}-\pi_1 & \cdots & \frac{Y_k}{n}-\pi_k\end{bmatrix}\begin{bmatrix}\frac{1}{\pi_1} & & 0\\ & \ddots & \\ 0 & & \frac{1}{\pi_k}\end{bmatrix}\begin{bmatrix}\frac{Y_1}{n}-\pi_1\\ \vdots\\ \frac{Y_k}{n}-\pi_k\end{bmatrix}$$

$$=n\sum_{j=1}^{k}\frac{(Y_j/n-\pi_j)^2}{\pi_j}=\sum_{j=1}^{k}\frac{(Y_j-n\pi_j)^2}{n\pi_j}.$$

Then another term is

$$n\begin{bmatrix}\frac{Y_1}{n}-\pi_1 & \cdots & \frac{Y_k}{n}-\pi_k\end{bmatrix}\frac{\mathbf{1}\mathbf{1}^{\mathrm{T}}}{\pi_{k+1}}\begin{bmatrix}\frac{Y_1}{n}-\pi_1\\ \vdots\\ \frac{Y_k}{n}-\pi_k\end{bmatrix}$$

$$=\frac{n}{\pi_{k+1}}\begin{bmatrix}\frac{Y_1}{n}-\pi_1 & \cdots & \frac{Y_k}{n}-\pi_k\end{bmatrix}\begin{bmatrix}1 & & 1\\ & \ddots & \\ 1 & & 1\end{bmatrix}\begin{bmatrix}\frac{Y_1}{n}-\pi_1\\ \vdots\\ \frac{Y_k}{n}-\pi_k\end{bmatrix}$$

$$=\frac{n}{\pi_{k+1}}\sum_{i=1}^{k}\left(\frac{Y_i}{n}-\pi_i\right)\sum_{j=1}^{k}\left(\frac{Y_j}{n}-\pi_j\right)=\frac{n}{\pi_{k+1}}\left(\sum_{i=1}^{k}\frac{Y_i}{n}-\sum_{i=1}^{k}\pi_i\right)\left(\sum_{j=1}^{k}\frac{Y_j}{n}-\sum_{j=1}^{k}\pi_j\right)$$

$$=\frac{n}{\pi_{k+1}}\left(1-\frac{Y_{k+1}}{n}-1+\pi_{k+1}\right)\left(1-\frac{Y_{k+1}}{n}-1+\pi_{k+1}\right)=\frac{n}{\pi_{k+1}}\left(\pi_{k+1}-\frac{Y_{k+1}}{n}\right)^2$$

$$=\frac{(Y_{k+1}-n\pi_{k+1})^2}{n\pi_{k+1}}.$$

Finally, by combining the results, we obtain

$$\mathbf{V}_n^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{V}_n=\sum_{j=1}^{k}\frac{(Y_j-n\pi_j)^2}{n\pi_j}+\frac{(Y_{k+1}-n\pi_{k+1})^2}{n\pi_{k+1}}=\sum_{j=1}^{k+1}\frac{(Y_j-n\pi_j)^2}{n\pi_j}.$$

Then we have proven the Claim hence derived the asymptotic distribution of Pearson's chi-square statistics under the null hypothesis.

Q4 [+16] Neyman's Smooth Test

Let $p_\theta(x) = C(\theta)\exp\left[\sum_{j=1}^{2}\theta_j T_j(x)\right]$, $x \in (0,1)$, be a density w.r.t. the Lebesgue measure, where $T_1(x) = \sqrt{3}(2x-1)$ and $T_2(x) = \sqrt{5}(6x^2-6x+1)$ be the Legendre polynomials. Draw the graph of $p_\theta(x)$ including the locations of $p_\theta(0)$, $p_\theta(1)$, $\min_{x\in(0,1)} p_\theta(x)$ and $\max_{x\in(0,1)} p_\theta(x)$.

(1)[+4] $\theta_1 = 1$ and $\theta_2 = 0$

Answer:

If $\theta_1 = 1$ and $\theta_2 = 0$, we have

$$
\begin{aligned}
& p_\theta(x) = C(\theta)e^{\{\sqrt{3}(2x-1)\}} \\
\Rightarrow & \int_0^1 p_\theta(x)dx = C(\theta)\int_0^1 e^{\{\sqrt{3}(2x-1)\}}dx \\
\Rightarrow & 1 = C(\theta)e^{-\sqrt{3}}\left(\frac{1}{2\sqrt{3}+1}e^{2\sqrt{3}x}\Big|_0^1\right) \quad . \\
\Rightarrow & C(\theta) = \left\{\frac{e^{-\sqrt{3}}}{2\sqrt{3}+1}(e^{2\sqrt{3}+1}-1)\right\}^{-1} \\
\Rightarrow & C(\theta) = \frac{(2\sqrt{3}+1)e^{\sqrt{3}}}{e^{2\sqrt{3}+1}-1}.
\end{aligned}
$$

Therefore, the density is

$$p_\theta(x) = \frac{(2\sqrt{3}+1)e^{2\sqrt{3}x}}{e^{2\sqrt{3}+1}-1}.$$

Then we can compute

$$p_\theta(0) = \frac{2\sqrt{3}+1}{e^{2\sqrt{3}+1}-1} \qquad \text{and} \qquad p_\theta(1) = \frac{(2\sqrt{3}+1)e^{2\sqrt{3}}}{e^{2\sqrt{3}+1}-1}.$$

Since $p_\theta(x)$ is increasing in x, we obtain

$$\min_{0<x<1} p_\theta(x) = p_\theta(0) \qquad \text{and} \qquad \max_{0<x<1} p_\theta(x) = p_\theta(1).$$

The density is plotted in Figure 2. It shows that the minimum and maximum are attained at $x = 0$ and $x = 1$, respectively.

(2) [+4] $\theta_1 = 0$ and $\theta_2 = 1$ [no need to calculate the numerical value of $C(\theta)$]

Answer:

If $\theta_1 = 0$ and $\theta_2 = 1$, the density is

$$p_\theta(x) = C(\theta) e^{\{\sqrt{5}(6x^2 - 6x + 1)\}}.$$

Then we can compute

$$p_\theta(0) = C(\theta) e^{\sqrt{5}} \quad \text{and} \quad p_\theta(1) = C(\theta) e^{\sqrt{5}}.$$

Since

$$\frac{d}{dx} p_\theta(x) = \sqrt{5}(12x - 6) C(\theta) e^{\{\sqrt{5}(6x^2 - 6x + 1)\}} \overset{\text{set}}{=} 0 \Rightarrow x = \frac{1}{2}$$

and

$$\frac{d^2}{dx^2} p_\theta(x) \Big|_{x=1/2} > 0.$$

Therefore, we obtain

$$\min_{0<x<1} p_\theta(x) = p_\theta(1/2) = C(\theta) e^{-\sqrt{5}/2} \quad \text{and} \quad \max_{0<x<1} p_\theta(x) = p_\theta(1) = C(\theta) e^{\sqrt{5}}.$$

The density is plotted in Figure 3. It shows that the minimum and maximum are attained at $x = 1/2$ and $x = 0, 1$, respectively.

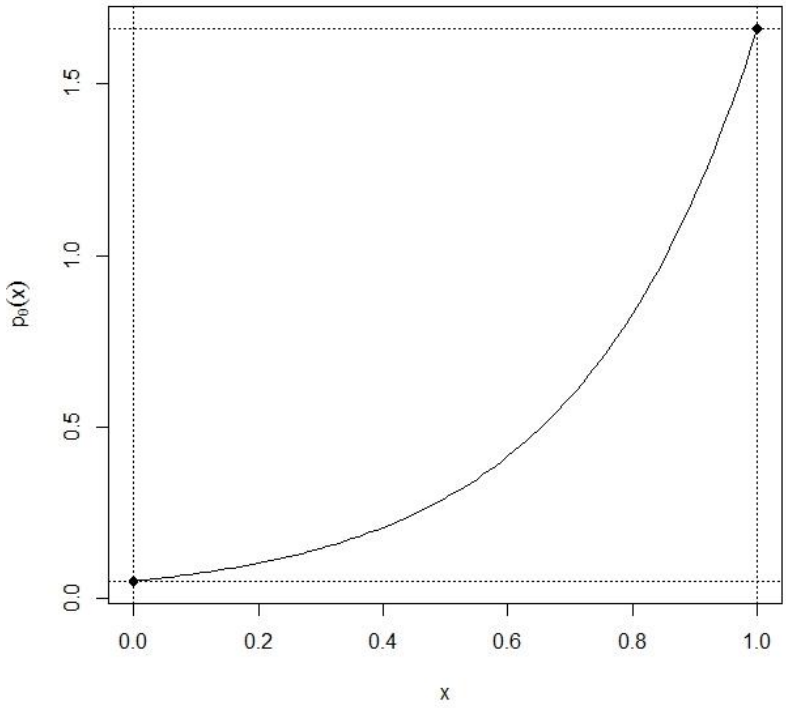

Figure 2. The density $p_\theta(x)$ under $\theta_1 = 1$ and $\theta_2 = 0$ with minimum and maximum attained at $x = 0$ and $x = 1$, respectively.

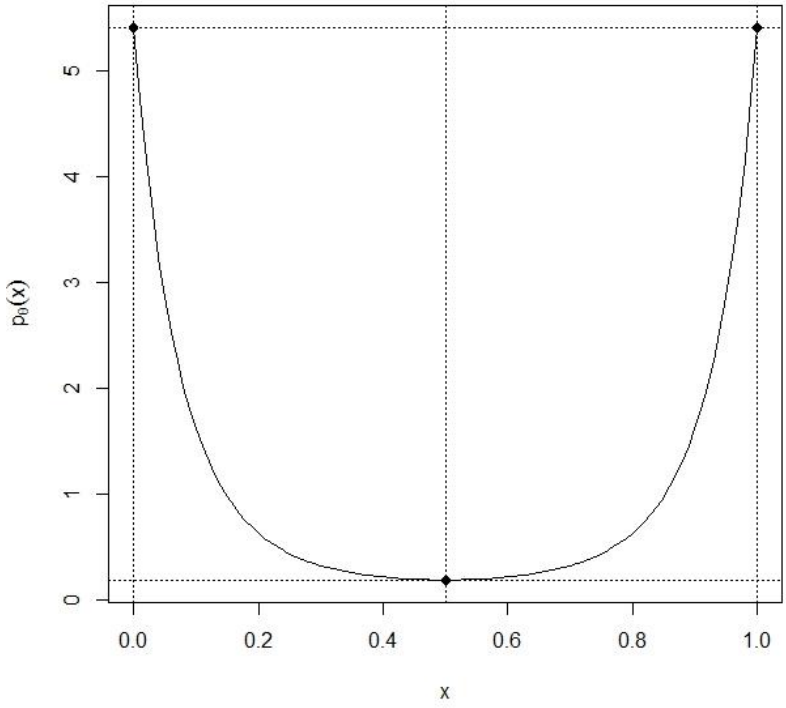

Figure 3. The density $p_\theta(x)$ under $\theta_1 = 0$ and $\theta_2 = 1$ with minimum and maximum attained at $x = 1/2$ and $x = 0, 1$, respectively.

(3) [+4] Derive the Neyman's smooth test statistics.

Answer:

Suppose $X_1, X_2, \cdots, X_n \overset{i.i.d.}{\sim} F$, to test $H_0 : F = U(0,1)$ vs. $H_1 : F \neq U(0,1)$. It is equivalent to test that $H_0 : \theta_j = 0$ for all j vs. $H_1 : \theta_j \neq 0$ for some j. Under the null hypothesis, consider

$$\log p_\theta(x) = \log C(\theta) + \sum_{j=1}^{k} \theta_j T_j(x)$$

$$\Rightarrow \frac{\partial}{\partial \theta_m} \log p_\theta(x) = \frac{\partial}{\partial \theta_m} \log C(\theta) + T_m(x) \overset{\text{set}}{=} 0$$

$$\Rightarrow E\{T_m(X)\} = -\frac{\partial}{\partial \theta_m} \log C(\theta) = 0.$$

The last equality is due to the orthogonality of $T_0(x), T_1(x), \cdots, T_k(x)$. Under the null hypothesis, that is, $H_0 : \theta_j = 0$ for all j, we have

$$\exp\left[\sum_{j=1}^{k} \theta_j T_j(x)\right] = T_0(x) = 1.$$

Therefore,

$$E\{T_m(X)\} = \int_0^1 T_m(x) C(\theta) \exp\left[\sum_{j=1}^{k} \theta_j T_j(x)\right] dx$$

$$= C(\theta) \int_0^1 T_m(x) T_0(x) dx = 0.$$

In addition, the orthogonality of $T_0(x), T_1(x), \cdots, T_k(x)$ also gives the following results

$$\operatorname{cov}\{T_i(x), T_j(x)\} = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases}$$

Thus, we define

$$\mathbf{Z}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[T_1(x_i) \quad \cdots \quad T_k(x_i)\right]^{\mathrm{T}}.$$

By the multivariate CLT, we have

$$\mathbf{Z}_n \overset{d}{\to} \mathrm{MVN}_k(\mathbf{0}, \mathbf{I}), \quad \text{as } n \to \infty.$$

Then the Neyman's smooth test statistics is

$$\mathbf{Z}_n^{\mathrm{T}}\mathbf{I}\mathbf{Z}_n = \mathbf{Z}_n^{\mathrm{T}}\mathbf{Z}_n = \frac{1}{n}\sum_{i=1}^{n}\begin{bmatrix} T_1(x_i) & \cdots & T_k(x_i) \end{bmatrix}\sum_{i=1}^{n}\begin{bmatrix} T_1(x_i) \\ \vdots \\ T_k(x_i) \end{bmatrix}$$

$$= \sum_{j=1}^{k}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}T_j(x_i)\right\}^2 \xrightarrow{d} \chi_k^2, \quad \text{as } n \to \infty.$$

For instance, the Neyman's smooth test with $k=1$ and $k=2$ are

$$\sum_{j=1}^{1}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}T_j(x_i)\right\}^2 = \frac{1}{n}\left\{\sum_{i=1}^{n}\sqrt{3}(2x_i-1)\right\}^2$$

$$= 12n\left\{\frac{1}{n}\sum_{i=1}^{n}x_i - \frac{1}{2}\right\}^2$$

and

$$\sum_{j=1}^{2}\left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}T_j(x_i)\right\}^2 = \frac{1}{n}\left\{\sum_{i=1}^{n}\sqrt{3}(2x_i-1)\right\}^2 + \frac{1}{n}\left\{\sum_{i=1}^{n}\sqrt{5}(6x_i^2-6x_i+1)\right\}^2$$

$$= 12n\left\{\frac{1}{n}\sum_{i=1}^{n}x_i - \frac{1}{2}\right\}^2 + 180n\left\{\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \frac{1}{n}\sum_{i=1}^{n}x_i + \frac{1}{6}\right\}^2,$$

respectively.

(4) [+4] How the Neyman's test statistics is related to the moments of $U(0,1)$.

Answer:

The expressions of the Neyman's smooth test statistics for $k=1$ and $k=2$ are

$$\mathbf{Z}_n^{\mathrm{T}}\mathbf{Z}_n = 12n\left\{\frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{2}\right\}^2$$

and

$$\mathbf{Z}_n^{\mathrm{T}}\mathbf{Z}_n = 12n\left\{\frac{1}{n}\sum_{i=1}^{n} x_i - \frac{1}{2}\right\}^2 + 180n\left\{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{6}\right\}^2,$$

respectively. We reject the null hypothesis if $\mathbf{Z}_n^{\mathrm{T}}\mathbf{Z}_n > \chi^2_{k,1-\alpha}$. Under the null hypothesis, we have

$$\frac{1}{n}\sum_{i=1}^{n} x_i \xrightarrow{p} E_0(X) = \frac{1}{2} \quad \text{which is the first moment of } U(0,1)$$

and

$$\frac{1}{n}\sum_{i=1}^{n} x_i^2 \xrightarrow{p} E_0(X^2) = \frac{1}{3} \quad \text{which is the second moment of } U(0,1).$$

If the data truly follow the uniform distribution $U(0,1)$, the Neyman's smooth test statistics will converge to zero, that is, $\mathbf{Z}_n^{\mathrm{T}}\mathbf{Z}_n \xrightarrow{p} 0$. Then we cannot reject the null hypothesis.