STAT 135 CONCEPTS OF STATISTICS
HOMEWORK 3

Assigned July 15, 2021, due July 22, 2021

This homework pertains to materials covered in Lecture 6, 7 and 8. The assignment can be typed or handwritten, with your name on the document, and **with properly labeled input code and computer output for those problems that require it**. To obtain full credit, please write clearly and show your reasoning. If you choose to collaborate, the write-up should be your own. Please show your work! Upload the file to the Week 3 Assignment on bCourses.

Note in this homework, we use the following abbreviations: Cramer-Rao bound (CR bound), Uniformly minimum-variance unbiased estimator (UMVUE), probability density/mass function (pdf/pmf), Maximum likelihood estimators (MLE) and likelihood ratio test (LRT).

Problem 1. Consider the i.i.d random variables $X_1, \ldots, X_n$ from Poission(λ).

(1) Show that both sample mean $\bar{X}_n$ and sample variance $S^2 = (n-1)^{-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ are both unbiased estimators of λ.
(2) Calculate the CR lower bound for any unbiased estimator of λ.
(3) Is either $\bar{X}_n$ or S^2 a UMVUE of λ?
(*Hint*: For i.i.d random variables, $\mathrm{Var}(S^2) = \frac{E(X-\mu)^4}{n} - \frac{\sigma^4(n-3)}{n(n-1)}$. For Poission($\lambda$), the 4th central moment is $E(X-\lambda)^4 = \lambda(1+3\lambda)$.)

Solution.

(1) Since for Poisson random variable $X \sim$ Poission(λ), $E(X) = \mathrm{Var}(X) = \lambda$, we know sample mean $\bar{X}_n$ and sample variance S^2 are both unbiased estimators of λ.
(2) First we calculate the Fisher information for λ:

$$I(\lambda) = -E_\lambda\left(\frac{\partial^2}{\partial\lambda^2}\log f\right) = E_\lambda\left(\frac{X}{\lambda^2}\right) = \frac{1}{\lambda}.$$

Then the CR lower bound should be

$$\mathrm{CR\_bound}(\lambda) = \frac{1}{nI(\lambda)} = \frac{\lambda}{n}.$$

(3) It is easy to calculate $\mathrm{Var}(\bar{X}_n) = \frac{\lambda}{n}$, which attains the CR lower bound. Therefore, the sample mean is a UMVUE.

For the sample variance, we know from the hint that

$$\mathrm{Var}(S^2) = \frac{\lambda(1+3\lambda)}{n} - \frac{\lambda^2(n-3)}{n(n-1)} = \frac{\lambda}{n} + \frac{2\lambda^2}{n} + \frac{2\lambda^2}{n(n-1)} > \frac{\lambda}{n}.$$

Therefore, S^2 is not a UMVUE.

Problem 2. Consider the i.i.d random variables $X_1, \ldots, X_n$ from some population $f(x|\theta)$ with the true parameter being θ_0, which is sufficiently smooth to satisfy

$$\int_{\mathcal{X}} \frac{\partial}{\partial\theta} f(x|\theta)dx = \frac{\partial}{\partial\theta}\int_{\mathcal{X}} f(x|\theta)dx,$$
$$\int_{\mathcal{X}} \frac{\partial^2}{\partial^2\theta} f(x|\theta)dx = \frac{\partial^2}{\partial\theta^2}\int_{\mathcal{X}} f(x|\theta)dx,$$

in which $\mathcal{X}$ is the set of all possible values of the population.

(1) Suppose $X \sim f(x|\theta_0)$. Show that

$$E_{\theta_0}\left[\frac{\partial}{\partial\theta}\log f(X|\theta_0)\right] = 0.$$

(2) Denote $f(x_1, \ldots, x_n|\theta) = \prod_{i=1}^n f(x_i|\theta)$ as the joint density. Prove

$$E_{\theta_0}\left[\frac{\partial}{\partial\theta}\log f(X_1, \ldots, X_n|\theta_0)\right] = 0.$$

(3) Prove

$$E_{\theta_0}\left[\frac{\partial}{\partial\theta}\log f(X_1, \ldots, X_n|\theta_0)\right]^2 = nI(\theta_0).$$

Solution.

(1) By definition of expected value,

$$\begin{aligned}
E_{\theta_0}\left[\frac{\partial}{\partial\theta}\log f(X|\theta_0)\right] &= \int_{\mathcal{X}} \frac{\partial}{\partial\theta}\log f(x|\theta_0) \times f(x|\theta_0)dx \\
&= \int_{\mathcal{X}} \frac{\frac{\partial}{\partial\theta} f(x|\theta_0)}{f(x|\theta_0)} f(x|\theta_0)dx = \int_{\mathcal{X}} \frac{\partial}{\partial\theta} f(x|\theta_0)dx \\
&\overset{\text{Exchangeability}}{=} \left[\frac{\partial}{\partial\theta}\int_{\mathcal{X}} f(x|\theta)\right]\Bigg|_{\theta=\theta_0} = 0.
\end{aligned}$$

(2) Similarly,

$$
\begin{aligned}
E_{\theta_0}\left[\frac{\partial}{\partial\theta}\log f(X_1,\ldots,X_n|\theta_0)\right] &\overset{\text{i.i.d}}{=} \int_{\mathcal{X}^n}\frac{\partial}{\partial\theta}\log\left\{\prod_{i=1}^n f(x_i|\theta_0)\right\}\times\prod_{i=1}^n f(x_i|\theta_0)d\mathbf{x}_n \\
&= \int_{\mathcal{X}^n}\frac{\partial}{\partial\theta}\left\{\prod_{i=1}^n f(x_i|\theta_0)\right\}\, d\mathbf{x}_n \\
&= \int_{\mathcal{X}^n}\sum_{j=1}^n\left\{\prod_{i\neq j} f(x_i|\theta_0)\times\frac{\partial}{\partial\theta}f(x_j|\theta_0)\right\}\, d\mathbf{x}_n \\
&= \sum_{j=1}^n\left\{\prod_{i\neq j}\int_{\mathcal{X}} f(x_i|\theta_0)dx_i\right\}\times\int_{\mathcal{X}}\frac{\partial}{\partial\theta}f(x_j|\theta_0)dx_j = 0.
\end{aligned}
$$

(3) First note that

$$
\begin{aligned}
\left[\frac{\partial}{\partial\theta}\log f(X_1,\ldots,X_n|\theta_0)\right]^2 &= \left[\sum_{i=1}^n\frac{\partial}{\partial\theta}\log f(X_i|\theta_0)\right]^2 \\
&= \sum_{i=1}^n\left\{\frac{\partial}{\partial\theta}\log f(X_i|\theta_0)\right\}^2 + \sum_{i\neq j}\frac{\partial}{\partial\theta}\log f(X_i|\theta_0)\times\frac{\partial}{\partial\theta}\log f(X_j|\theta_0).
\end{aligned}
$$

Then by the i.i.d assumption,

$$
\begin{aligned}
&E_{\theta_0}\left[\frac{\partial}{\partial\theta}\log f(X_1,\ldots,X_n|\theta_0)\right]^2 \\
=&nE_{\theta_0}\left\{\frac{\partial}{\partial\theta}\log f(X|\theta_0)\right\}^2 + n(n-1)E_{\theta_0}\left\{\frac{\partial}{\partial\theta}\log f(X_1|\theta_0)\times\frac{\partial}{\partial\theta}\log f(X_2|\theta_0)\right\} \\
\overset{\text{independence}}{=}&nE_{\theta_0}\left\{\frac{\partial}{\partial\theta}\log f(X|\theta_0)\right\}^2 + n(n-1)E_{\theta_0}\left\{\frac{\partial}{\partial\theta}\log f(X_1|\theta_0)\right\}\times E_{\theta_0}\left\{\frac{\partial}{\partial\theta}\log f(X_2|\theta_0)\right\} \\
\overset{(1)}{=}&nE_{\theta_0}\left\{\frac{\partial}{\partial\theta}\log f(X|\theta_0)\right\}^2 \overset{\text{Definition}}{=} nI(\theta_0).
\end{aligned}
$$

□

Problem 3. Suppose $X_1,\ldots,X_n$ are independently sampled from a distribution with pdf

$$
f(x|\theta) = \begin{cases} 1/\theta, & \text{if } x\in(0,\theta), \\ 0, & \text{otherwise.}\end{cases}
$$

(1) Find the MLE, $\hat\theta_{MLE}$, of θ. Is it a sufficient statistic for θ?
(2) Calculate the Fisher information $I(\theta)$.
(3) Can we apply Theorem D of Lecture 5 to obtain the asymptotic normality $\sqrt{n}(\hat\theta_{MLE}-\theta)\xrightarrow{d} N(0,1/I(\theta))$? Explain why.

(4) If you answered no to the previous question, derive the exact sampling distribution of $\hat{\theta}_{MLE}$ and see why it can not be approximated by a Normal distribution.
(*Hint*: The density of the sample maximum $X_{(n)}$ is $f_{X_{(n)}}(x) = nf(x)F^{n-1}(x)$.)
(5) Show that $Y = \frac{n+1}{n}X_{(n)}$ is an unbiased estimator of θ. Calculate $\mathrm{Var}(Y)$, and show that it is smaller than the CR bound $\frac{1}{nI(\theta)}$. Is this result contradictory to the CR inequality in Theorem E?

Solution.

(1) The joint likelihood function for this sample can be written as

$$L(\theta|\mathbf{X}_n) = \left(\frac{1}{\theta}\right)^n \mathbb{1}\{\theta > X_{(n)}\}.$$

Since $L(\theta|\mathbf{X}_n)$ is monotonically decreasing function of θ, $\hat{\theta}_{MLE} = X_{(n)}$. Meanwhile, by the Fisher-Neyman factorization theorem, we know $X_{(n)}$ is a sufficient statistic for θ.

(2) By definition,

$$I(\theta) = E_\theta\left(\frac{\partial}{\partial\theta}\log f\right)^2 = \frac{1}{\theta^2}.$$

(3) No, we cannot. There are two assumptions violated that are needed to ensure asymptotic normality.

- Exchangeability assumption:

$$\frac{\partial}{\partial\theta}\int_0^\theta f(x|\theta)dx = 0 \neq \int_0^\theta \frac{\partial}{\partial\theta} f(x|\theta)dx = \int_0^\theta -\frac{1}{\theta^2}dx = -\frac{1}{\theta}.$$

- Support assumption:

$\{x : f(x|\theta) > 0\} = \{x : x \in (0,\theta)\}$ is dependent on θ.

(4) By the hint, the density function for $\hat{\theta}_{MLE}$ is

$$f_{X_{(n)}}(x) = n\frac{1}{\theta}\left(\frac{x}{\theta}\right)^{n-1} = \frac{nx^{n-1}}{\theta^n},$$

which is not a symmetric distribution – this density converges to ∞ when $x \searrow 0$ and converges to 0 when $x \nearrow \infty$. Therefore, the sampling distribution of $\hat{\theta}_{MLE}$ clearly does not converge in distribution to a Normal distribution as $n \to \infty$.

(5) From (4),

$$E(X_{(n)}) = \int_0^\theta \frac{nx^n}{\theta^n}dx = \frac{n\theta}{n+1}.$$

Thus, $E(Y) = \theta$ is an unbiased estimator of θ.

To obtain the variance of Y, we note

$$E(X_{(n)}^2) = \int_0^\theta \frac{nx^{n+1}}{\theta^n}dx = \frac{n\theta^2}{n+2}, \quad \mathrm{Var}(X_{(n)}) = \frac{n\theta^2}{n+2} - \left(\frac{n\theta}{n+1}\right)^2 = \frac{n\theta^2}{(n+2)(n+1)^2},$$

from which we get

$$\mathrm{Var}(Y) = \left(\frac{n+1}{n}\right)^2 \frac{n\theta^2}{(n+1)(n+2)^2} = \frac{\theta^2}{n(n+2)} < \frac{1}{nI(\theta)} = \frac{\theta^2}{n}.$$

That is, the variance of Y is always strictly less than the CR lower bound, which seems contradictory to Theorem E. However, since the exchangeability assumption and the independence of the support from the model parameter do not hold, **CR inequality is not applicable to this pdf** in the first place.

Problem 4. Let's say we observed 14 i.i.d $N(\mu, \sigma^2)$ samples with both μ and σ^2 being unknown:

$$\{-1.398,\ 8.061,\ 13.609,\ 4.325,\ 12.140,\ -4.611,\ 6.669,$$
$$4.340,\ 1.776,\ 7.355,\ -3.100,\ -3.784,\ 9.962,\ -4.150\}.$$

(1) Find 99% exact confidence intervals for μ and σ^2.
(2) Find 99% bootstrap confidence intervals for μ and σ^2.
(3) Compare the two sets of CIs. Which one has more accuracy?
(4) How much larger a sample do you think you would need to halve the length of the confidence interval for μ?

Solution. First we calculate the sample mean and sample variance:

$$\bar{X}_n = 3.657,\ \hat{\sigma}_n^2 = 36.660 \text{ and } S^2 = 39.479.$$

(1) Since $n = 14$ and $t_{13}(0.01/2) = 3.012$, the 99% exact confidence intervals for μ is

$$\bar{X}_n \pm t_{13}(0.01/2)\frac{S}{\sqrt{n}} = 3.657 \pm 3.012\frac{\sqrt{39.479}}{\sqrt{14}} = [\mathbf{-1.401}, \mathbf{8.715}].$$

Since $\chi^2_{13}(0.01/2) = 29.819$ and $\chi^2_{13}(1 - 0.01/2) = 3.565$, the 99% exact confidence intervals for σ^2 is

$$\left[\frac{n\hat{\sigma}_n^2}{\chi^2_{13}(0.01/2)}, \frac{n\hat{\sigma}_n^2}{\chi^2_{13}(1-0.01/2)}\right] = \left[\frac{14 * 36.660}{29.819}, \frac{14 * 36.660}{3.565}\right] = [\mathbf{17.212}, \mathbf{143.966}].$$

(2) Since $z_{0.01/4} = 2.807$, the 99% bootstrap confidence intervals for μ is

$$\bar{X}_n \pm z_{0.01/4}\frac{\hat{\sigma}_n}{\sqrt{n}} = 3.657 \pm 2.807\frac{\sqrt{36.660}}{\sqrt{14}} = [\mathbf{-1.057}, \mathbf{8.371}].$$

The 99% bootstrap confidence intervals for σ is

$$\hat{\sigma}_n \pm z_{0.01/4}\frac{\hat{\sigma}_n}{\sqrt{2n}} = \sqrt{36.660} \pm 2.807\frac{\sqrt{36.660}}{\sqrt{2 * 14}} = [\mathbf{2.843}, \mathbf{9.267}],$$

and thus the 99% bootstrap confidence intervals for σ^2 is $[\mathbf{2.843}, \mathbf{9.267}]^2 = [\mathbf{8.082}, \mathbf{85.870}]$.

(3) The bootstrap CIs have narrower widths. However, that does not mean that the bootstrap CIs are more accurate than the exact CIs. Since the exact CIs are obtained from the exact sampling distribution, narrow CIs will cover less than 99% AUC (area under the curve) of the sampling distribution. That is, the actual coverage probability of the bootstrap CIs is less than the advertised confidence level.

(4) By the Normal approximation, the length of the CIs is proportional to $1/\sqrt{n}$. Therefore, to halve the length, we solve

$$\frac{1}{\sqrt{n}} \leq 0.5 * \frac{1}{\sqrt{14}},$$

which gives $n \geq 14/0.5^2 = 56$.

Problem 5. Let $X_1, \dots, X_n$ be i.i.d samples from a Bernoulli(p) population. It is desired to test

$$H_0 : p = 0.59, \qquad \text{versus} \qquad H_1 : p = 0.45.$$

(1) Derive the rejection region of the LRT.

(2) Use the central limit theorem to determine, approximately, the sample size needed so that type I error and type II error are both at most 0.01.

Solution.

(1) Similarly to Example 2 of lecture 7, we first calculate the likelihood ratio

$$\begin{aligned}
\lambda(\mathbf{X}_n) =& \frac{\sup_{p=0.59} L(p|\mathbf{X}_n)}{\sup_{p=0.59 \text{ or } 0.45} L(p|\mathbf{X}_n)} \\
=& \frac{0.59^{\sum_{i=1}^n X_i} 0.41^{n-\sum_{i=1}^n X_i}}{\max\{0.59^{\sum_{i=1}^n X_i} 0.41^{n-\sum_{i=1}^n X_i}, 0.45^{\sum_{i=1}^n X_i} 0.55^{n-\sum_{i=1}^n X_i}\}} \\
=& \min\left\{1, \left(\frac{0.59}{0.45}\right)^{\sum_{i=1}^n X_i} \left(\frac{0.41}{0.55}\right)^{n-\sum_{i=1}^n X_i}\right\} \\
=& \min\left\{1, \left(\frac{0.59 * 0.55}{0.45 * 0.41}\right)^{\sum_{i=1}^n X_i} \left(\frac{0.41}{0.55}\right)^n\right\} = \min\left\{1, (1.7588)^{\sum_{i=1}^n X_i} \left(\frac{0.41}{0.55}\right)^n\right\}.
\end{aligned}$$

Thus, the rejection region of the LRT should be

$$\begin{aligned}
R = \{\lambda(\mathbf{X}_n) \leq c\} &\overset{c\in[0,1]}{=} \left\{(1.7588)^{\sum_{i=1}^n X_i} \left(\frac{0.41}{0.55}\right)^n \leq c\right\} \\
&= \left\{\bar{X}_n \leq c'\right\}, \text{ where } c' = \frac{\log\{c(0.55/0.41)^n\}}{n \log 1.7588} = 0.5203 + \frac{\log c}{n \log 1.7588}.
\end{aligned}$$

(2) Under H_0, $\sqrt{n}(\bar{X}_n - 0.59) \xrightarrow{d} N(0, 0.59 * 0.41)$. Therefore,

$$\begin{aligned}
P(\text{Type I error}) =& P(\mathbf{X}_n \in R | H_0) = P(\bar{X}_n \leq c' | H_0) \\
=& P\left[\frac{\sqrt{n}(\bar{X}_n - 0.59)}{\sqrt{0.59 * 0.41}} \leq \frac{\sqrt{n}(c' - 0.59)}{\sqrt{0.59 * 0.41}} \middle| H_0\right] \\
=& \Phi\left[\frac{\sqrt{n}(c' - 0.59)}{\sqrt{0.59 * 0.41}}\right].
\end{aligned}$$

To ensure $P(\text{Type I error}) \leq 0.01$, we need

$$\frac{\sqrt{n}(c' - 0.59)}{\sqrt{0.59 * 0.41}} \leq -z_{0.01} = -2.3263.$$

Note $c' < 0.59$ when n is sufficiently large. Therefore, the inequality above can be transformed into

$$n \geq \left(\frac{2.3263}{0.59 - c'}\right)^2 * 0.59 * 0.41 \approx \left(\frac{2.3263}{0.59 - 0.5203}\right)^2 * 0.59 * 0.41 = 269.4646.$$

On the other hand, $\sqrt{n}(\bar{X}_n - 0.45) \xrightarrow{d} N(0, 0.45 * 0.55)$ under H_1. Therefore,

$$\begin{aligned}
P(\text{Type II error}) =& P(\mathbf{X}_n \notin R | H_1) = P(\bar{X}_n > c' | H_1) \\
=& 1 - \Phi\left[\frac{\sqrt{n}(c' - 0.45)}{\sqrt{0.45 * 0.55}}\right].
\end{aligned}$$

To ensure $P(\text{Type I error}) \leq 0.01$, we need

$$\frac{\sqrt{n}(c' - 0.45)}{\sqrt{0.45 * 0.55}} \geq z_{0.01} = 2.3263,$$

which leads to

$$n \geq \left(\frac{2.3263}{c' - 0.45}\right)^2 * 0.45 * 0.55 \geq \left(\frac{2.3263}{0.5203 - 0.45}\right)^2 * 0.45 * 0.55 = 271.0167.$$

To sum up, the sample size need to be at least $\max\{269.4646, 271.0167\} = 271.0167$ to control both type I & II errors at 0.01.

Problem 6. Suppose $X_1, \ldots, X_n$ are independently sampled from Pareto(θ, ν) with pdf

$$f(x|\theta) = \begin{cases} \frac{\theta \nu^\theta}{x^{\theta+1}}, & \text{if } x \geq \nu, \\ 0, & \text{otherwise.} \end{cases}$$

in which $\theta > 0$ and $\nu > 0$.

(1) Find the MLEs of θ and ν.
(2) Show that the LRT of

$$H_0 : \theta = 1,\ \nu \text{ unknown}, \qquad \text{versus} \qquad H_1 : \theta \neq 1,\ \nu \text{ unknown}$$

has rejection region of the form $\{T(\mathbf{X}_n) \le c_1 \text{ or } T(\mathbf{X}_n) \ge c_2\}$ where $0 < c_1 < c_2$ and

$$T(\mathbf{X}_n) = \log\left[\frac{\prod_{i=1}^n X_i}{X_{(1)}^n}\right],$$

in which $X_{(1)}$ is the sample minimum.

Solution.

(1) First write out the joint likelihood

$$L(\theta,\nu|\mathbf{X}_n) = \frac{\theta^n \nu^{n\theta}}{\left(\prod_{i=1}^n X_i\right)^{\theta+1}} \mathbb{1}\{X_{(1)} \ge \nu\}.$$

The log-likelihood is

$$l(\theta,\nu|\mathbf{X}_n) = \begin{cases} n\log\theta + n\theta\log\nu - (\theta+1)\sum_{i=1}^n \log X_i, & \text{if } X_{(1)} \ge \nu, \\ -\infty, & \text{otherwise.} \end{cases}$$

Since $\theta > 0$, $l(\theta,\nu|\mathbf{X}_n)$ is a strictly increasing function of ν for any fixed θ. Thus, $\hat{\nu}_{MLE} = X_{(1)}$.

Take derivation with respect to θ, and we get

$$\frac{\partial l}{\partial \theta} = \frac{n}{\theta} + n\log\nu - \sum_{i=1}^n \log X_i = 0$$

$$\Rightarrow \hat{\theta}_{MLE} = \frac{1}{n^{-1}\sum_{i=1}^n \log\left(X_i/\hat{\nu}_{MLE}\right)} = \frac{1}{n^{-1}\log\prod_{i=1}^n\left(X_i/X_{(1)}\right)}.$$

(2) Note that $\Theta = \{\theta > 0, \nu > 0\}$ and $\Theta_0 = \{\theta = 1, \nu > 0\}$. Thus,

$$\sup_{\Theta} L(\theta,\nu|\mathbf{X}_n) = L(\hat{\theta}_{MLE},\hat{\nu}_{MLE}|\mathbf{X}_n) = \hat{\theta}_{MLE}^n \exp\left\{\hat{\theta}_{MLE}\log\prod_{i=1}^n(\hat{\nu}_{MLE}/X_i)\right\} \Big/ \prod_{i=1}^n X_i$$

$$= \frac{n^n}{\left\{\log\prod_{i=1}^n (X_i/X_{(n)})\right\}^n}\exp\{-n\} \Big/ \prod_{i=1}^n X_i.$$

To maximize $L(\hat{\theta}_{MLE},\hat{\nu}_{MLE}|\mathbf{X}_n)$ in Θ_0, ν still has to take a value of $X_{(1)}$. Thus,

$$\sup_{\Theta_0} L(\theta,\nu|\mathbf{X}_n) = L(1, X_{(1)}|\mathbf{X}_n) = \exp\left\{\log\prod_{i=1}^n (X_{(1)}/X_i)\right\} \Big/ \prod_{i=1}^n X_i.$$

Denote $T(\mathbf{X}_n) = \log\prod_{i=1}^n (X_i/X_{(1)})$, and the likelihood ratio is

$$\lambda(\mathbf{X}_n) = \frac{T^n(\mathbf{X}_n)}{n^n}\exp\{n - T(\mathbf{X}_n)\} = n^{-n}e^n\frac{T^n(\mathbf{X}_n)}{e^{T(\mathbf{X}_n)}}.$$

Since the function $f(t) = t^n/e^t$ monotonically increases from 0 to n and then decreases from n to ∞ (see Figure 1), the rejection region can be simplified as the following

$$R = \{\lambda(\mathbf{X}_n) \le c\} = \{T(\mathbf{X}_n) \le c_1 \text{ or } T(\mathbf{X}_n) \ge c_2\}.$$

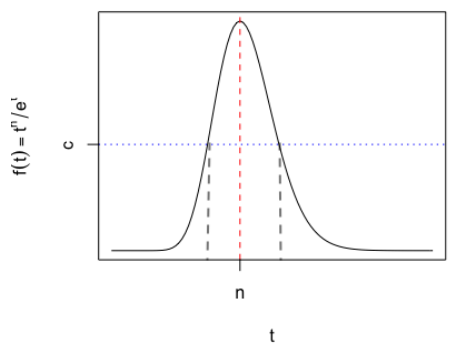

FIGURE 1. The curve of $f(t) = t^n/e^t$ which attains maximum at $t = n$. To ensure $f(t) \leq c$, we need $t \leq c_1$ and $t \geq c_2$ in which $c_1 < c_2$ are the two roots of $f(t) = c$.

Problem 7. Let $X_1, \ldots, X_n$ be i.i.d with beta$(\mu, 1)$ pdf and $Y_1, \ldots, Y_m$ be i.i.d with beta$(\theta, 1)$. Also assume that X's and Y's are independent of each other.

(1) Find an LRT of $H_0 : \theta = \mu$ versus $H_1 : \theta \neq \mu$, and show that the LRT statistic is

$$T(\mathbf{X}_n, \mathbf{Y}_m) = \frac{\sum_{i=1}^n \log X_i}{\sum_{i=1}^n \log X_i + \sum_{i=1}^m \log Y_i}.$$

(2) Find the distribution of T when H_0 is true, and show how to get a test of size $\alpha = 0.05$.
(*Hint*: $-\log X_i \sim \text{Gamma}(1, 1/\mu)$ and $-\log Y_i \sim \text{Gamma}(1, 1/\theta)$. Also, $W/(W+V) \sim \text{Beta}(n, m)$ if $W \sim \text{Gamma}(n, 1/\mu)$, $V \sim \text{Gamma}(m, 1/\mu)$ and W is independent of V.)

(3) Suppose $n = 13$ and $m = 17$. The samples collected are

$$\begin{aligned}\mathbf{X}_{13} = \{&0.610,\ 0.344,\ 0.289,\ 0.700,\ 0.710,\ 0.266,\ 0.244,\ 0.919,\\ &0.022,\ 0.006,\ 0.073,\ 0.849,\ 0.773\},\\ \mathbf{Y}_{17} = \{&0.781,\ 0.479,\ 0.821,\ 0.766,\ 0.444,\ 0.443,\ 0.290,\ 0.862,\\ &0.684,\ 0.151,\ 0.931,\ 0.753,\ 0.694,\ 0.121,\ 0.264,\ 0.731,\ 0.575\}.\end{aligned}$$

Will you reject the null hypothesis? Report the p-value.

Solution.

(1) The joint likelihood can be written as

$$L(\mu, \theta | \mathbf{X}_n, \mathbf{Y}_m) = \mu^n \prod_{i=1}^n X_i^{\mu-1} \theta^m \prod_{i=1}^m Y_i^{\theta-1}.$$

In this case, $\Theta = \{\theta > 0, \mu > 0\}$ and $\Theta_0 = \{\theta = \mu, \theta > 0, \mu > 0\}$.

Since there is independence between $\mathbf{X}_n$ and $\mathbf{Y}_m$, it is easy to verify that $\hat{\mu}_{MLE} = n / \sum_{i=1}^n \log(1/X_i)$, $\hat{\theta}_{MLE} = m / \sum_{i=1}^m \log(1/Y_i)$ and

$$\begin{aligned}\sup_{\Theta} L(\mu, \theta | \mathbf{X}_n, \mathbf{Y}_m) &= L(\hat{\mu}_{MLE}, \hat{\theta}_{MLE} | \mathbf{X}_n, \mathbf{Y}_m) \\ &= \left[\frac{n}{\sum_{i=1}^n \log(1/X_i)}\right]^n \Big/ \prod_{i=1}^n X_i \times \left[\frac{m}{\sum_{i=1}^m \log(1/Y_i)}\right]^m \Big/ \prod_{i=1}^m Y_i.\end{aligned}$$

Similarly,

$$\sup_{\Theta_0} L(\mu, \theta | \mathbf{X}_n, \mathbf{Y}_m) = \left[\frac{n+m}{\sum_{i=1}^n \log(1/X_i) + \sum_{i=1}^m \log(1/Y_i)}\right]^{n+m} \Big/ \left(\prod_{i=1}^n X_i \times \prod_{i=1}^m Y_i\right).$$

Using the definition $T(\mathbf{X}_n, \mathbf{Y}_m)$, the likelihood ratio can be written as

$$\begin{aligned}\lambda(\mathbf{X}_n, \mathbf{Y}_m) &= \frac{(n+m)^{n+m}}{n^n m^m} \left[\frac{\sum_{i=1}^n \log X_i}{\sum_{i=1}^n \log X_i + \sum_{i=1}^m \log Y_i}\right]^n \left[\frac{\sum_{i=1}^m \log Y_i}{\sum_{i=1}^n \log X_i + \sum_{i=1}^m \log Y_i}\right]^m \\ &= \frac{(n+m)^{n+m}}{n^n m^m} T^n(\mathbf{X}_n, \mathbf{Y}_m) \left[1 - T(\mathbf{X}_n, \mathbf{Y}_m)\right]^m.\end{aligned}$$

Similarly to Problem 6, this is a uni-modal function of $T(\mathbf{X}_n, \mathbf{Y}_m)$ which attains its maximum at $t = \frac{n}{n+m}$. Therefore, the rejection region is only dependent on $T(\mathbf{X}_n, \mathbf{Y}_m)$, which becomes the LRT statistic – rejecting H_0 if $\lambda(\mathbf{X}_n, \mathbf{Y}_m) \leq c$ is equivalent to rejecting if $T \leq c_1$ or $T \geq c_2$, where $c_1 < \frac{n}{n+m} < c_2$.

(2) Under H_0, we know from the hint that

$$T(\mathbf{X}_n, \mathbf{Y}_m) \sim \text{Beta}(n, m).$$

(3) Let's plug in the data value and calculate the LRT statistic to get

$$T(\mathbf{X}_n, \mathbf{Y}_m) = 0.6086678.$$

Recall that p-value is the probability of observing a statistic as extreme as 0.6086678 under H_0. Since $0.6086678 > n/(m+n) = 0.4333$, the observed statistic is larger than the mode of the unimodal function in (1). Therefore, being as extreme as 0.6086678 means (also see Figure 2)

$$\begin{aligned}p\text{-value} =& 2 * P\left\{T(\mathbf{X}_n, \mathbf{Y}_m) \geq 0.6086678 \mid H_0\right\} \\ =& \texttt{2*pbeta(0.6086678, 13, 17, lower.tail=FALSE)} \\ =& 0.0526 > 0.05.\end{aligned}$$

Therefore, we fail to reject H_0 at significance level 0.05, and conclude that there is not enough evidence to support the alternative hypothesis $\theta \neq \mu$.

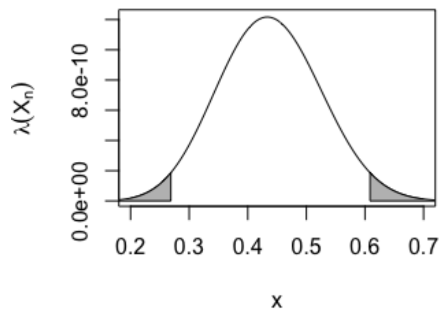

FIGURE 2. What it means to be as extreme as 0.6086678 under H_0.

Problem 8. In a study of the effect of cigarette smoking on the carbon monoxide diffusing capacity (DL) of the lung, researchers found that current smokers had DL readings significantly lower than those of either ex-smokers or non-smokers. The carbon monoxide diffusing capacities for a random sample of n = 20 current smokers are listed here:

$$\{103.768,\ 92.295,\ 100.615,\ 102.754,\ 88.602,\ 61.675,\ 88.017,\ 108.579,\ 73.003,\ 90.677,\ 71.210,\ 73.154,\ 123.086,\ 84.023,\ 82.115,\ 106.755,\ 91.052,\ 76.014,\ 89.222,\ 90.479\}$$

(1) Compute the sample mean and sample standard deviation of the above data.
(2) Do these data indicate that the mean DL reading for current smokers is significantly lower than 100, which is the average for nonsmokers? Use a one-sided hypothesis test, with $\alpha = 0.01$. Since $n < 30$, you will need to use exact Student's t distribution to find the rejection regions of the test.

Solution.

(1) It is easy to obtain
$$\bar{X}_n = 89.855, \quad S = 14.904.$$
(2) We want to test whether
$$H_0 : \mu = 100 \text{ or } H_1 : \mu < 100.$$
From Lecture 9 Example 5, the rejection region for this test should be
$$R = \left\{ \frac{\sqrt{n}(\bar{X}_n - 100)}{S} \leq -t_{n-1}(\alpha) \right\}.$$

In this case,

$$-t_{n-1}(\alpha) = -t_{19}(0.01) = -2.539,$$

and

$$\frac{\sqrt{n}(\bar{X}_n - 100)}{S} = -3.044.$$

Therefore, this sample is in the rejection region – we reject the null hypothesis and conclude that the mean DL reading for current smokers is **significantly lower** than 100, which is the average for non-smokers.

Problem 9. Suppose X is a random variable whose pmf under H_0 and H_1 is given by

x	1	2	3	4	5	6	7
$f(x\|H_0)$	0.01	0.01	0.01	0.01	0.02	0.01	0.93
$f(x\|H_1)$	0.06	0.05	0.04	0.03	0.02	0.01	0.79

Use the Neyman-Pearson Lemma to find the most powerful test for H_0 versus H_1 with $\alpha = 0.04$. Compute the probability of Type II error for this test.

Solution. Since H_0 and H_1 can both be regarded as singletons, Neyman-Pearson Lemma is applicable. Let's now first derive the likelihood ratio

X	1	2	3	4	5	6	7
$\lambda(X)$	1/6	1/5	1/4	1/3	1	1	1.177

The LRT rejection region should be $R = \{\lambda(X) \le c\}$. To find c such that the significance level $\alpha = 0.04$, we try to solve

$$0.04 = P(\lambda(X) \le c|H_0).$$

Note that $P(\lambda(X) \le 1/3|H_0) = P(X = 1, 2, 3, 4|H_0) = 0.4$ and $P(\lambda(X) \le 1|H_0) = P(X = 1, 2, 3, 4, 5, 6|H_0) = 0.07$. Therefore,

$$R = \{\lambda(X) \le 1/3\}.$$

Then the type II error can be calculated as

$$P(\text{type II error}) = P(\lambda(X) > 1/3|H_1) = P(X = 5, 6, 7|H_1) = 0.82.$$

Problem 10. Let X be a single observation from $f(x|\theta) = \theta x^{\theta-1}$, $0 < x < 1$.

(1) Find the most powerful test using significance level $\alpha = 0.05$ for testing the hypotheses $H_0 : \theta = 1$ versus $H_1 : \theta = 2$ (sketch the densities $f(x|H_0)$ and $f(x|H_1)$ for the two hypotheses).
(2) What is the power of the test?
(3) What is the p-value of $X = .8$?
(4) For fixed $\alpha = 0.05$, is the test uniformly most powerful against the alternative hypothesis $H_1 : \theta > 1$?

Solution.

(1) Since there is only one sample,
$$f(X|H_0) = 1, \quad f(X|H_1) = 2X.$$
The likelihood ratio is
$$\lambda(X) = \frac{f(X|H_0)}{\max\{f(X|H_0), f(X|H_1)\}} = \min\left\{1, \frac{1}{2X}\right\}.$$
The rejection region is
$$R = \{\lambda(X) \leq c\} = \{1/(2X) \leq c\} = \{X \geq 1/(2c)\}.$$
To make the significance level $\alpha = 0.05$,
$$0.05 = P(X \geq 1/(2c)|H_0) = \int_{1/(2c)}^{1} 1dx = 1 - 1/(2c).$$
Thus, $c = 0.5263$. By the Neyman-Pearson Lemma, the UMP test for these two hypotheses has a rejection region $R = \{1/(2X) \leq 0.5263\}$.

(2) The power of the test is
$$\beta = 1 - P(\text{type II error}) = \int_{1/(2c)}^{1} 2xdx = 1 - 1/(2c)^2 = 0.0974.$$

(3) The p-value is
$$p\text{-value} = P(X \geq 0.8|H_0) = \int_{0.8}^{1} 1dx = 0.2.$$

(4) By Karlin-Rubin theorem, we know this test is UMP against the alternative hypothesis $H_1 : \theta > 1$.