Bios 661: 1 – 5; Bios 673: 2 – 6.

1. C&B 8.12
2. C&B 8.7(a)
3. C&B 8.15
4. [From 2011 master exam] Let $X_1, \ldots, X_n$ be i.i.d. random variables from the pdf

$$f(x|\theta) = (1-\theta) + \frac{\theta}{2\sqrt{x}}, \quad 0 < x < 1, \quad 0 \le \theta \le 1.$$

That is, we have a random sample of size n from the population f. The parameter θ is unknown.

(a) Derive the uniformly most powerful level α test ($0 < \alpha < 1$) for $H_0 : \theta = 0$ against $H_1 : \theta = 1$. Specify the critical region as concisely and as explicitly as possible. Justify your answers. Do not use any approximations.

Solution: Using Neyman-Pearson Lemma, we can derive the UMP test with critical region

$$R = \left\{ \boldsymbol{x} : \frac{f(\boldsymbol{x}|\theta = 1)}{f(\boldsymbol{x}|\theta = 0)} > c \right\},$$

where

$$\frac{f(\boldsymbol{x}|\theta = 1)}{f(\boldsymbol{x}|\theta = 0)} = 2^{-n} \prod_{i=1}^{n} x_i^{-1/2},$$

or, equivalently,

$$R^* = \left\{ \boldsymbol{x} : -\sum_{i=1}^{n} \log x_i > c^* \right\}.$$

Under the null hypothesis ($\theta = 0$), the pdf of $Y_i = -\log X_i$ is

$$f_{Y_i}(y_i|\theta) = e^{-y_i},$$

which is exponential distribution with mean 1. Hence, $-\sum_{i=1}^{n} \log X_i$ follows an Gamma distribution with parameters n and 1, denoted by Gamma$(n, 1)$. To make it a level α test with size α, one can choose $c^* = \Gamma_{n,1,1-\alpha}$, which is the $(1-\alpha)$ quantile of Gamma$(n, 1)$.

(b) For the special case of $n = 5$ and $\alpha = 0.01$, find the critical region (exactly). Also, find the (exact) power of the test.

Solution: For the special case of $n = 5$ and $\alpha = 0.01$, we choose $c^* = \Gamma_{5,1,0.99} = 11.6$. Under the alternative hypothesis $\theta = 1$, the pdf of $Y_i = -\log X_i$ is

$$f_{Y_i}(y_i) = 2^{-1} e^{y_i/2} e^{-y_i} = \frac{1}{2} e^{-y_i/2},$$

which is Exp(2). Then, $-\sum_{i=1}^{5} \log X_i$ follows Gamma(5, 2). The power of the test is

$$P_{\theta=1}\left(-\sum_{i=1}^{5} \log X_i > 11.6\right) = 0.31.$$

(c) An investigator wants to design a study in which the test derived above will be applied. The investigator desires a Type I Error probability of 0.01 and a Type II Error probability of 0.01. Find the minimum required sample size n (exact, or approximate, whichever is easier).

Solution: From (a) and (b), we know

$$0.99 = P_{\theta=1}\left(-\sum_{i=1}^{n} \log X_i > \Gamma_{n,1,0.99}\right) = 1 - F(\Gamma_{n,1,0.99}),$$

where F is the CDF of Gamma$(n, 2)$. The equation above is a function of n, so one would be able to solve n numerically. Here, when $n = 45$, the power is 0.989, and when $n = 46$, the power is 0.9904. One would choose $n = 46$ under exact distribution.

One can also use an approximate distribution to find the sample size. By Central Limit Theorem, under the alternative hypothesis, one can have

$$\sqrt{n}\left(-\sum_{i=1}^{n} \log X_i/n - \mu\right) \to_d N(0, \sigma^2),$$

where $\mu = E(-\log X_1) = 2$ and $\sigma^2 = \text{Var}(-\log X_1) = 4$. That means

$$\begin{aligned} 0.99 &= P_{\theta=1}\left(\frac{\sqrt{n}(-\sum_{i=1}^{n} \log X_i/n - 2)}{\sqrt{4}} > \frac{\sqrt{n}(\Gamma_{n,1,0.99}/n - 2)}{\sqrt{4}}\right) \\ &\approx P\left(Z > \frac{\sqrt{n}(\Gamma_{n,1,0.99}/n - 2)}{\sqrt{4}}\right), \end{aligned}$$

where Z is the standard normal distribution. By this approximation, one can have

$$\frac{\sqrt{n}(\Gamma_{n,1,0.99}/n - 2)}{\sqrt{4}} = -2.33$$

and solve for $n \approx 52$.

(d) This part pertains to the special case of $n = 1$ (sample size = 1). Find $\hat{\theta}$, the maximum likelihood estimator (MLE) of θ. Show that the MLE is biased. Then find constants a and b such that $T(X_1) = a + b\hat{\theta}$ is unbiased for θ. Do you see any potential problems with $T(X_1)$ as an estimator of θ?

Solution: The likelihood function is

$$L(\theta|x) = f(x|\theta) = (1-\theta) + \frac{\theta}{2\sqrt{x}}.$$

The function is differentiable, so one can take the first derivative

$$L'(\theta|x) = \frac{\partial}{\partial\theta}L(\theta|x) = -1 + \frac{1}{2\sqrt{x}}.$$

When $0 < x < 1/4$, $L'(\theta|x) > 0$ and $L(\theta)$ is a monotone increasing function of θ. Therefore, $\hat{\theta} = 1$ if $0 < x < 1/4$. On the other hand, when $1/4 < x < 1$, $L'(\theta|x) < 0$ and $L(\theta)$ is a monotone decreasing function of θ. Therefore, $\hat{\theta} = 0$ if $1/4 < x < 1$. When $x = 1/4$, any point between 0 and 1 can be the MLE. However, since the probability of having $x = 1/4$ is 0, it is little concern to have $\hat{\theta} = 0$ or 1. Formally, we have

$$\hat{\theta} = \begin{cases} 1 & \text{if } 0 < x \le 1/4 \\ 0 & \text{if } 1/4 < x < 1. \end{cases}$$

The expectation of $\hat{\theta}$ is

$$E(\hat{\theta}) = P(0 < X \le 1/4) = \int_0^{1/4} (1-\theta) + \frac{\theta}{2\sqrt{x}} dx = \frac{1}{4} + \frac{1}{4}\theta,$$

which shows $\hat{\theta}$ is biased and $a = b = 1/4$. Finding an unbiased estimator is easy, one can use $4\hat{\theta} - 1$ as the unbiased estimator. Since the estimator is always outside the parameter space, it is not feasible at all.

5. An epidemiologist gathers data (x_i, Y_i) on each of n randomly chosen noncontiguous and demographically similar cities in the United States, where x_i, $i = 1, \ldots, n$, is the

known population size (in millions of people) in city i, and where Y_i is the random variable denoting the number of people in city i with colon cancer. It is reasonable to assume that Y_i, $i = 1, \ldots, n$, has a Poisson distribution with mean $E(Y_i) = \theta x_i$, where $\theta > 0$ is an unknown parameter, and that $Y_1, Y_2, \ldots, Y_n$ are mutually independent random variables.

(a) Use the available data (x_i, Y_i), $i = 1, \ldots, n$, construct a uniformly most powerful (UMP) level α test for $H_0 : \theta = 1$ versus $H_1 : \theta > 1$.

Solution: By Neyman-Pearson Lemma, the uniformly most powerful test has a rejection region as

$$\frac{f(\boldsymbol{y}|\theta_1)}{f(\boldsymbol{y}|\theta = 1)} = \frac{\prod_{i=1}^n (y_i!)^{-1} e^{-\theta_1 x_i} (\theta_1 x_i)^{y_i}}{\prod_{i=1}^n (y_i!)^{-1} e^{-x_i} x_i^{y_i}} = e^{(1-\theta_1)\sum_{i=1}^n x_i} \theta_1^{\sum_{i=1}^n y_i} > c,$$

where $\theta_1 > 1$. Since $\theta_1 > 1$, $f(\boldsymbol{y}|\theta_1)/f(\boldsymbol{y}|\theta = 1) > c$ is equivalent to $\sum_{i=1}^n y_i > c^*$. One hence can establish the test with a critical region $R = \{\boldsymbol{y}; \sum_{i=1}^n y_i > c^*\}$ is the UMP test.

(b) Use the available data (x_i, Y_i), $i = 1, \ldots, n$, construct a uniformly most powerful (UMP) level α test for $H_0 : \theta \leq 1$ versus $H_1 : \theta > 1$. Is this critical region the same as the one used in (a)?

Solution: We intend to use Karlin-Rubin Theorem as the hypothesis is composite vs. composite. We need to prove that $\sum_{i=1}^n Y_i$ is a sufficient statistic, which can be shown by factorizing the pdf as

$$f(\boldsymbol{y}|\theta) = \prod_{i=1}^n (y_i!)^{-1} e^{-\theta x_i} (\theta x_i)^{y_i} = \left(\prod_{i=1}^n (y_i!)^{-1} x_i^{y_i} \right) e^{-\theta \sum_{i=1}^n x_i} \theta^{\sum_{i=1}^n y_i}.$$

We then show that the pdf has the property of maximum likelihood ratio (MLR) in $\sum_{i=1}^n Y_i$. For every $\theta_2 > \theta_1$, one can see the likelihood ratio

$$\frac{f(\boldsymbol{y}|\theta_2)}{f(\boldsymbol{y}|\theta_1)} = e^{(\theta_2 - \theta_1)\sum_{i=1}^n x_i} \left(\frac{\theta_2}{\theta_1} \right)^{\sum_{i=1}^n y_i}$$

is monotone increasing in $S = \sum_{i=1}^n Y_i$. Hence the MLR property stands. By the Karlin-Rubin Theorem the UMP test has a critical region $R = \{\boldsymbol{y}; S = \sum_{i=1}^n y_i > s_0\}$. This critical region is the same as in (a) since they are both established via the same test statistic. As suggested in the following question, $\sum_{i=1}^n Y_i$ follows a Poisson distribution and the c^* and s_0 can be found in satisfaction with the type I error probability.

(c) One can show that $S = \sum_{i=1}^n Y_i$ follows Poisson$(\theta \sum_{i=1}^n x_i)$. If one observes $\sum_{i=1}^n x_i = 0.8$, find c^* in the critical region $\mathcal{R} = \{S : S \geq c^*\}$ with size $\alpha = 0.05$.

(If $X \sim$ Poisson(0.8), then $P(X = 0) = 0.449$, $P(X \leq 1) = 0.808$, $P(X \leq 2) = 0.952$, $P(X \leq 3) = 0.990$, $P(X \leq 4) = 0.999$).

Solution: Given that the type I error probability $\alpha = 0.05$, one has $P(\sum_{i=1}^n Y_i \geq c^*|\theta = 1) \leq 0.05$. Under the null hypothesis ($\theta = 1$) and $\sum_{i=1}^n x_i = 0.8$, $\sum_{i=1}^n Y_i$ follows Poisson(0.8). Therefore,

$$
\begin{aligned}
P(\sum_{i=1}^n Y_i \geq c^*) &= 1 - P(\sum_{i=1}^n Y_i < c^*) \\
&= 1 - P(\sum_{i=1}^n Y_i \leq c^* - 1) \leq 0.05.
\end{aligned}
$$

By the information provided, one should choose $c^* - 1 = 2$. Hence $c^* = 3$.

(d) What is the power when in reality $\theta = 5$, using the critical region in (c) and $\sum_{i=1}^n x_i = 0.8$?

(If $X \sim$ Poisson(4), then $P(X = 0) = 0.018$, $P(X \leq 1) = 0.092$, $P(X \leq 2) = 0.238$, $P(X \leq 3) = 0.433$, $P(X \leq 4) = 0.628$).

Solution: Given that the critical region is $R = \{\boldsymbol{y} : \sum_{i=1}^n y_i \geq 3\}$, the power at $\theta = 5$ is

$$
\begin{aligned}
\beta(5) = P(\sum_{i=1}^n Y_i \geq 3|\theta = 5) &= 1 - P(\sum_{i=1}^n Y_i < 3|\theta = 5) \\
&= 1 - P(\sum_{i=1}^n Y_i \leq 2|\theta = 5) \\
&= 1 - 0.238 = 0.762,
\end{aligned}
$$

since, under $\theta = 5$, $\sum_{i=1}^n Y_i$ follows Poisson$(0.8 \times 5) \equiv$ Poisson (4).

6. Suppose that $Y_1, \ldots, Y_n$, $n > 1$, is a random sample from the pdf

$$
f_Y(y|\theta) = \frac{4}{\sqrt{\pi}} \theta^{-3} y^2 \exp\left(-\frac{y^2}{\theta^2}\right), \quad 0 < y < \infty, \quad 0 < \theta < \infty.
$$

(a) Show that Y_i^2, $i = 1, \ldots, n$, follows a Gamma distribution $\Gamma(3/2, \theta^2)$.

Solution: Let $X = Y^2$. The inverse function is $Y = \sqrt{X}$ with Jacobian $dy = (1/2)x^{-1/2}dx$. The pdf of X is

$$f_X(x) = \frac{2}{\sqrt{\pi}}\theta^{-3}x^{1/2}\exp\left(-\frac{x}{\theta^2}\right) = \frac{1}{\Gamma(\frac{3}{2})\theta^3}x^{1/2}\exp\left(-\frac{x}{\theta^2}\right),$$

which is the pdf of Gamma$(3/2, \theta^2)$.

(b) Derive the uniformly most powerful size α test, $0 < \alpha < 1$, of $H_0 : \theta = 1$ against $H_1 : \theta > 1$. Specify the rejection region as $R = \{\boldsymbol{y} : \sum_{i=1}^n y_i^2 \geq c^*\}$ with some constant c^*.

Solution: According to Neyman-Pearson Lemma, the UMP test has the rejection region $R = \{\boldsymbol{x} : \sum_{i=1}^n x_i \geq c^*\}$. The cutoff c^* can be specified satisfying

$$\alpha = P(\sum_{i=1}^n x_i \geq c^* | \theta = 1).$$

Since X_i follows Gamma$(3/2, 1)$ under the null hypothesis, one can see that $2\sum_{i=1}^n X_i$ follows Gamma$(3n/2, 2)$, which is χ^2_{3n}. One can set $c^* = \chi^2_{3n,1-\alpha}/2$ to satisfy the equation above.

(c) Derive the likelihood ratio test statistic $\lambda(\boldsymbol{y})$ for $H_0 : \theta = 1$ against $H_1 : \theta \neq 1$, and show that the rejection region $R = \{\boldsymbol{y} : \lambda(\boldsymbol{y}) \leq c\}$ is equivalent to $R = \{\boldsymbol{y} : \sum_{i=1}^n y_i^2 \leq c_1^* \text{ or } \sum_{i=1}^n y_i^2 \geq c_2^*\}$. Find the cutoff c_1^* and c_2^* explicitly given size $\alpha = 0.05$.

Solution: The likelihood ratio test statistic can be written as

$$\lambda(x) = \frac{\exp(-n\bar{x})}{(2\bar{x}/3)^{-3n/2}\exp(-n\bar{x}/\theta^2)} \propto \bar{x}^{3/2}\exp\{-n\bar{x}(1-\theta^{-2})\}.$$

Since $1 - \theta^{-2} > 0$, $\lambda(x)$ is a concave function of $\bar{x}$. The equivalent region is $R = \{\boldsymbol{x} : \sum_{i=1}^n x_i \leq c_1^* \text{ or } \sum_{i=1}^n x_i \geq c_2^*\}$, i.e., $R = \{\boldsymbol{y} : \sum_{i=1}^n y_i^2 \leq c_1^* \text{ or } \sum_{i=1}^n y_i^2 \geq c_2^*\}$. To explicitly find c_1^* and c_2^*, we let

$$\alpha_1 = P(\sum_{i=1}^n y_i^2 \leq c_1^* | \theta = 1),$$

and

$$\alpha_2 = P(\sum_{i=1}^{n} y_i^2 \geq c_2^*|\theta = 1),$$

where $\alpha = \alpha_1 + \alpha_2$. One can set $c_1^* = \chi^2_{3n,\alpha_1}/2$ and $c_2^* = \chi^2_{3n,1-\alpha_2}/2$.

(d) Determine whether one rejects the null hypothesis if $n = 25$ and $\hat{\theta} = 1.2$ (observed value of θ) at the 0.05 level.

Solution: Since $\sum_{i=1}^{n} Y_i^2$ follows Gamma$(3n/2, \theta^2)$, the MLE of θ^2 is $\hat{\theta}^2 = (3n/2)^{-1}\sum_{i=1}^{n} Y_i^2$. Given that $n = 25$ and $\hat{\theta} = 1.2$, we can get $\sum_{i=1}^{n} y_i^2 = 54$. Choosing $\alpha_1 = \alpha_2 = 0.025$, the cut-offs are $c_1^* = 26.47$ and $c_2^* = 50.42$. Hence one should reject the null hypothesis.

7. [Bios 673/740 class discussion] Let $X_1, \ldots, X_n$ be a random sample from the discrete uniform distribution on points $1, \ldots, \theta$, where $\theta = 1, 2, \ldots$.

(a) Consider $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$, where $\theta_0 > 0$ is known. Show that

$$\delta^*(X) = \begin{cases} 1 & X_{(n)} > \theta_0 \\ \alpha & X_{(n)} \leq \theta_0 \end{cases}$$

is a UMP size α test.

Solution: One can show that the distribution has a monotone likelihood ratio (MLR) property in $X_{(n)}$. Therefore, by the Karlin-Rubin theorem, the UMP size α test is

$$\delta(X) = \begin{cases} 1 & X_{(n)} > c \\ \gamma & X_{(n)} = c \\ 0 & X_{(n)} < c, \end{cases}$$

where c is an integer and $\gamma \in (0, 1)$.

(b) Consider $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Show that

$$\delta^*(X) = \begin{cases} 1 & X_{(n)} > \theta_0 \ \text{ or } \ X_{(n)} \leq \theta_0\alpha^{1/n} \\ \alpha & \text{otherwise} \end{cases}$$

is a UMP size α test.