Bios 661: 1 – 5; Bios 673: 2 – 6.

1. C&B 7.40
2. C&B 7.44
3. C&B 8.5(a)(b) [This is a two-parameter case in LRT, using the same principal.]
4. An epidemiologist gathers data (x_i, Y_i) on each of n randomly chosen noncontiguous 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 liver 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 $Y_1, \ldots, Y_n$ constitute a set of mutually independent random variables.

   (a) Find the explicit expression for the MLE $\hat{\theta}$ of θ. Also, find the explicit expressions for $\mathrm{E}(\hat{\theta})$ and $\mathrm{Var}(\hat{\theta})$.

   Solution: The maximum likelihood estimator $\hat{\theta} = \sum_{i=1}^n Y_i / \sum_{i=1}^n x_i$. We have

   $$E(\hat{\theta}) = \sum_{i=1}^n E(Y_i) / \sum_{i=1}^n x_i = \sum_{i=1}^n \theta x_i / \sum_{i=1}^n x_i = \theta,$$

   and

   $$Var(\hat{\theta}) = \sum_{i=1}^n Var(Y_i) / (\sum_{i=1}^n x_i)^2 = \sum_{i=1}^n \theta x_i / (\sum_{i=1}^n x_i)^2 = \theta / \sum_{i=1}^n x_i.$$

   (b) Show that $\hat{\theta}$ is the UMVUE of θ.

   Solution: Sine the joint pdf (likelihood function) can be written as

   $$L(\theta) = h(y) c(\theta) \exp\left(\log\theta \sum_{i=1}^n y_i\right),$$

   we can claim the distribution belongs to exponential family with $\sum_{i=1}^n Y_i$ as a complete and sufficient statistic for θ. Since $\hat{\theta}$ is an unbiased estimator and function of $\sum_{i=1}^n Y_i$, we can claim $\hat{\theta}$ is an UMVUE.

(c) Find the explicit expression for the CRLB for the variance of any unbiased estimator of θ. Comment on if $\text{Var}(\hat{\theta})$ achieves the lower bound.

Solution: The denominator of the CRLB is

$$-E\left\{\frac{\partial^2 \ell(\theta)}{\partial \theta^2}\right\} = -E\left(\frac{\sum_{i=1}^n y_i}{\theta^2}\right) = \frac{\sum_{i=1}^n x_i}{\theta}.$$

The CRLB is $\theta / \sum_{i=1}^n x_i$, which is achieved by $\hat{\theta}$.

5. Let $X_1, \cdots, X_n$ be a random sample from an exponential distribution with pdf

$$f(x|\theta) = \begin{cases} e^{-(x-\theta)} & x \geq \theta \\ 0 & x < \theta, \end{cases}$$

where $-\infty < \theta < \infty$. Consider testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$, where θ_0 is a value specified by the researcher.

(a) Using the definition of likelihood ratio test, find the test statistic $\lambda(\boldsymbol{x})$.

Solution: The maximum of $L(\theta|x)$ under overall parameter space is

$$L(x_{(1)}|\boldsymbol{x}) = e^{-\sum x_i + n x_{(1)}},$$

while the maximum of $L(\theta|x)$ under the null space is

$$L(\theta_0|\boldsymbol{x}) = e^{-\sum x_i + n\theta_0} I(x_{(1)} \geq \theta_0),$$

which makes

$$\lambda(\boldsymbol{x}) = e^{n\theta_0 - n x_{(1)}} I(x_{(1)} \geq \theta_0).$$

(b) The rejection region of the likelihood ratio test is $R = \{\boldsymbol{x} : \lambda(\boldsymbol{x}) \leq c\}$ with some constant cutoff c. Show that this region is equivalent to $R^* = \{\boldsymbol{x} : x_{(1)} \geq c^* \text{ or } x_{(1)} < \theta_0\}$ with another cutoff constant c^*.

Solution: If we draw a graph between $\lambda(\boldsymbol{x})$ and $x_{(1)}$, we have $\lambda(\boldsymbol{x}) = 0$ when $x_{(1)} < \theta_0$ and $\lambda(\boldsymbol{x})$ as a decreasing function of $x_{(1)}$ when $x_{(1)} \geq \theta_0$. Hence, the equivalent region of $R = \{\boldsymbol{x} : \lambda(\boldsymbol{x}) \leq c\}$ is $R^* = \{\boldsymbol{x} : x_{(1)} \geq c^* \text{ or } x_{(1)} < \theta_0\}$.

(c) Find c^* specifically, using the definition of test size:

$$\alpha = \sup_{\theta \in \Theta_0} P(\boldsymbol{X} \in R^* | H_0).$$

Solution: Using the definition, we have

$$\begin{aligned}\alpha &= \sup_{\theta\in\Theta_0} P(X_{(1)} \geq c^* \quad \text{or} \quad X_{(1)} < \theta_0 | H_0) \\ &= P(X_{(1)} \geq c^* \quad \text{or} \quad X_{(1)} < \theta_0 | \theta = \theta_0) \\ &= P(X_{(1)} \geq c^* | \theta = \theta_0) \\ &= 1 - F_{X_{(1)}}(c^* | \theta = \theta_0) \\ &= e^{-n(c^*-\theta_0)}.\end{aligned}$$

The cumulative density function of $X_{(1)}$ is

$$F_{X_{(1)}}(y) = P(X_{(1)} \leq y) = 1 - \{P(X_1 \geq y)\}^n = 1 - e^{-n(y-\theta)},$$

where

$$P(X_1 \geq y) = 1 - P(X_1 \leq y) = 1 - \int_\theta^y e^{-(x-\theta)} dx = e^{-(y-\theta)}.$$

Hence, $c^* = \theta_0 - \log\alpha/n$.

(d) Based on the rejection region R^*, draw the power function over the parameter space $-\infty < \theta < \infty$.

Solution: The power function, by definition, is

$$\begin{aligned}\beta(\theta) &= P(X_{(1)} \geq \theta_0 - \log\alpha/n \quad \text{or} \quad X_{(1)} < \theta_0 | \theta \in \Theta) \\ &= P(X_{(1)} \geq \theta_0 - \log\alpha/n | \theta \in \Theta) + P(X_{(1)} < \theta_0 | \theta \in \Theta).\end{aligned}$$

For $\theta = \theta_0$, we know $\beta(\theta_0) = \alpha$. For $\theta > \theta_0$, we know $P(X_{(1)} < \theta_0 | \theta > \theta_0) = 0$ and

$$P(X_{(1)} \geq \theta_0 - \log\alpha/n | \theta > \theta_0) = e^{-n(\theta_0 - \log\alpha/n - \theta)},$$

which is an increasing function of θ. When $\theta < \theta_0$, we have

$$P(X_{(1)} \geq \theta_0 - \log\alpha/n | \theta < \theta_0) = e^{-n(\theta_0 - \log\alpha/n - \theta)},$$

and

$$P(X_{(1)} < \theta_0 | \theta < \theta_0) = 1 - e^{-n(\theta_0 - \theta)}.$$

Therefore, when $\theta < \theta_0$,

$$\begin{aligned}\beta(\theta) &= e^{-n(\theta_0 - \log\alpha/n - \theta)} + 1 - e^{-n(\theta_0-\theta)} \\ &= -e^{-n(\theta_0-\theta)}(1-\alpha) + 1,\end{aligned}$$

which is a decreasing function of θ. Overall, the power function $\beta(\theta)$ in this test looks like a convex function of θ with $\theta = \theta_0$ at the minimum.

6. Let $X_1, \ldots, X_n$ be a random sample from $N(\mu_x, \sigma^2)$ and let $Y_1, \ldots, Y_m$ be a random sample from $N(\mu_y, \sigma^2)$. Assume that two samples are mutually independent and σ^2 is *unknown*. To test the hypothesis $H_0 : \mu_x = \mu_y$ versus $H_1 : \mu_x \neq \mu_y$: [This is a two-sample case in LRT, resulting in classic two-sample t-test.]

   (a) Derive the likelihood ratio test $\lambda(x, y)$.

   Solution: Under the null parameter space, letting $\mu_0 = \mu_x = \mu_y$, the likelihood function is

$$L(\mu_0, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{\sum_{i=1}^n (x_i - \mu_0)^2}{2\sigma^2}\right\} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^m \exp\left\{-\frac{\sum_{i=1}^m (y_i - \mu_0)^2}{2\sigma^2}\right\}.$$

   Given σ^2, one can solve for the MLE of μ as $\hat{\mu} = \frac{1}{n+m}(\sum_{i=1}^n X_i + \sum_{i=1}^m Y_m)$. Plugging $\hat{\mu}$ back into $L(\mu, \sigma^2)$ and solving for the MLE of σ^2 by maximizing $L(\hat{\mu}, \sigma^2)$, we can have

$$\hat{\sigma}_0^2 = \frac{1}{n+m}\left\{\sum_{i=1}^n (X_i - \hat{\mu}_0)^2 + \sum_{i=1}^m (Y_i - \hat{\mu}_0)^2\right\}.$$

   That makes

$$L(\hat{\mu}_0, \hat{\sigma}_0^2) = \left(\frac{1}{\sqrt{2\pi\hat{\sigma}_0^2}}\right)^{n+m} \exp\left(-\frac{n+m}{2}\right).$$

   One the other hand, under the overall parameter space, the likelihood function is

$$L(\mu_x, \mu_y, \sigma^2) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{\sum_{i=1}^n (x_i - \mu_x)^2}{2\sigma^2}\right\} \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^m \exp\left\{-\frac{\sum_{i=1}^m (y_i - \mu_y)^2}{2\sigma^2}\right\}.$$

   By regular derivations of MLE, we can obtain $\hat{\mu}_x = \bar{X}$, $\hat{\mu}_y = \bar{Y}$, and

$$\hat{\sigma}^2 = \frac{1}{n+m}\left\{\sum_{i=1}^n (X_i - \bar{X})^2 + \sum_{i=1}^m (Y_i - \bar{Y})^2\right\}.$$

   That makes

$$L(\hat{\mu}_x, \hat{\mu}_y, \hat{\sigma}^2) = \left(\frac{1}{\sqrt{2\pi\hat{\sigma}^2}}\right)^{n+m} \exp\left(-\frac{n+m}{2}\right).$$

The likelihood ratio test statistic is

$$\begin{aligned}
\lambda(\boldsymbol{x}, \boldsymbol{y}) &= \frac{L(\hat{\mu}_0, \hat{\sigma}_0^2)}{L(\hat{\mu}_x, \hat{\mu}_y, \hat{\sigma}^2)} \\
&= \left(\frac{\hat{\sigma}_0^2}{\hat{\sigma}^2}\right)^{-\frac{n+m}{2}} \\
&= \left\{1 + \frac{n(\bar{x} - \hat{\mu}_0)^2 + m(\bar{y} - \hat{\mu}_0)^2}{\sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^m (y_i - \bar{y})^2}\right\}^{-\frac{n+m}{2}} \\
&= \left[1 + \frac{1}{m+n-2} \frac{\frac{mn}{m+n}(\bar{x} - \bar{y})^2}{\{\sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^m (y_i - \bar{y})^2\}/(m+n-2)}\right]^{-\frac{n+m}{2}}.
\end{aligned}$$

(b) Show that the rejection region $\lambda(x, y) \leq c$ is equivalent to $|t| \geq c^*$, where

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{(\frac{1}{m} + \frac{1}{n})s_p^2}} \quad \text{and} \quad s_p^2 = \frac{1}{m+n-2}\left\{\sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (y_i - \bar{y})^2\right\}.$$

Solution: Following $\lambda(\boldsymbol{x}, \boldsymbol{y})$ in (a), we have

$$\lambda(\boldsymbol{x}, \boldsymbol{y}) = \left(1 + \frac{1}{n+m-2} t^2\right)^{-\frac{n+m}{2}}.$$

That means having $\lambda(x, y) \leq c$ is equivalent to $|t| \geq c^*$ since $\lambda(\boldsymbol{x}, \boldsymbol{y})$ is a concave function of t.

(c) Find the explicit c^* when $\alpha = 0.05$.

Solution: By the definition of α, we have $\alpha = P(|T| \geq c^* | H_0)$. Since T follows a t distribution with degree of freedom $n + m - 2$. One can choose $c^* = t_{n+m-2, 1-\alpha/2}$.

(d) Given that $n = 14$, $m = 9$, $\bar{x} = 1249.9$, $\bar{y} = 1261.3$, $s_x^2 = n^{-1} \sum_{i=1}^n (x_i - \bar{x})^2 = 549.1$, and $s_y^2 = 156.6$, should one reject the null hypothesis at $\alpha = 0.05$?

Solution: Using previous results, we have $c^* = t_{14+9-2, 0.975} = 2.08$, $s_p^2 = 433.14$ and $t = -1.28$. There is not enough evidence to reject the null hypothesis since $|t| < 2.08$.

7. [Bios 673/740 class discussion, C&B 7.37] Let $X_1, \ldots, X_{n+1}$ be iid Bernoulli(p), and define the function $h(p)$ by

$$h(p) = P\left(\sum_{i=1}^{n} X_i > X_{n+1} | p\right),$$

which is the probability that the first n observations exceed the $(n+1)$st.

(a) Show that

$$T(X_1, \ldots, X_{n+1}) = \begin{cases} 1 & \text{if } \sum_{i=1}^{n} X_i > X_{n+1} \\ 0 & \text{otherwise} \end{cases}$$

is an unbiased estimator of $h(p)$.

(b) Find the best unbiased estimator of $h(p)$.

8. [Bios 673 class discussion] Suppose $X_1, \ldots, X_n$ is a random sample from $N(\mu, \sigma^2)$.

(a) If (μ, σ^2) is unknown, find the UMVUE of the 95th percentile.

Solution: The 95th percentile η shall satisfy

$$0.95 = P(X < \eta) = P\left(\frac{X - \mu}{\sigma} < \frac{\eta - \mu}{\sigma}\right).$$

Hence, one can express $\eta = \mu + 1.64\sigma$. Since the normal distribution belongs to the exponential family, one can show $(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2)$ is a complete sufficient statistic. If one can find $E\{\phi_1(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2)\} = \mu$ and $E\{\phi_2(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2)\} = \sigma$, then one can use Rao-Blackwell-Lehmann-Scheffe theorem to claim the UMVUE as $\hat{\eta} = \hat{\mu} + 1.64\hat{\sigma}$, where $\hat{\mu} = \phi_1(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2)$ and $\hat{\sigma} = \phi_2(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2)$. It is not hard to see that $\phi_1(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2) = \sum_{i=1}^{n} X_i/n = \bar{X}$. As to ϕ_2, one may choose $\phi_2(\sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2) = cS$, where $S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$, and see if we can find the constant c. Since we know that $(n-1)S^2/\sigma^2$ follows a χ^2_{n-1} distribution, we may write

$$\begin{aligned} E\left(\sqrt{\frac{(n-1)S^2}{\sigma^2}}\right) &= \int_0^\infty \sqrt{w} \frac{1}{\Gamma((n-1)/2)2^{(n-1)/2}} w^{(n-1)/2-1} \exp(-w/2) dw \\ &= \frac{\Gamma(n/2)2^{1/2}}{\Gamma((n-1)/2)}. \end{aligned}$$

Hence, the unbiased estimator of σ is cS, where

$$c = \sqrt{\frac{n-1}{2}} \frac{\Gamma((n-1)/2)}{\Gamma(n/2)}.$$

The UMVUE of η is $\hat{\eta} = \bar{X} + 1.96cS$.

(b) If σ^2 is given but μ is unknown, find the UMVUE of $P(X_1 < 1)$.

Solution: The nature unbiased estimator of $P(X_1 < 1)$ is $I(X_1 < 1)$. Now since the σ^2 is given, one can easily see that $\sum_{i=1}^n X_i$ is a complete sufficient statistic. To find the UMVUE, one may apply Lehmann-Scheffe theorem, where the UMVUE of $P(X_1 < 1)$ is

$$\phi\left(\sum_{i=1}^n X_i\right) = E\left(I(X_1 < 1)|\sum_{i=1}^n X_i\right).$$

To find the expectation, one may need to derive the conditional distribution of X_1 given that $\sum_{i=1}^n X_i = t$. Since both X_1 and $\sum_{i=1}^n X_i$ are normally distributed, one may derive the joint distribution first and then derive the conditional distribution based on the property of normality. The joint distribution of $(X_1, \sum_{i=1}^n X_i)$ is

$$\begin{pmatrix} X_1 \\ \sum_{i=1}^n X_i \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ n\mu \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & n \end{pmatrix}\sigma^2\right).$$

That gives the conditional distribution of X_1 given $\sum_{i=1}^n X_i = t$ as

$$X_1|\sum_{i=1}^n X_i = t \sim N\left(\frac{t}{n}, \left(1 - \frac{1}{n}\right)\sigma^2\right),$$

using the fact that the conditional normality of $X|Y = y$ from the joint normality (X, Y) is

$$X|Y = y \sim N\left(\mu_x + \frac{\sigma_x}{\sigma_y}(y - \mu_x), (1 - \rho^2)\sigma_1^2\right).$$

One hence can know that the UMVUE of $P(X_1 < 1)$ is

$$\begin{aligned} \phi\left(\sum_{i=1}^n X_i\right) &= E\left(I(X_1 < 1)|\sum_{i=1}^n X_i\right) \\ &= P\left(X_1 < 1|\sum_{i=1}^n X_i\right) \\ &= \Phi\left(\frac{1 - \sum_{i=1}^n X_i/n}{\sqrt{(1 - 1/n)\sigma^2}}\right), \end{aligned}$$

where $\Phi(\cdot)$ is the cumulative function of the standard normal distribution.