Instructions: You are required to do questions 1(a)(b), 2(a)(b)(c), 3(a)(b)(d)(e)(f). Questions 1(c), 2(d) and 3(c) are take-home questions for those who want to get extra credits. However, doing these questions will not move your grade from P to H.

1. Let $X_1, \ldots, X_n$ be a random variables from a uniform distribution on $(-\theta, \theta)$. Let $X_{(1)}$ and $X_{(n)}$ are minimum and maximum order statistics, respectively.

   (a) The range as a random variable can be defined by $R = X_{(n)} - X_{(1)}$. Derive the mean of $R$, and find an unbiased estimator of $\theta$.

   **Solution**: Let $Y = (X + \theta)/(2\theta)$. We know that $Y$ follows $U(0, 1)$. The pdf of $Y$ is $f_Y(y) = 1$ and the cdf of $Y$ is $F_Y(y) = y$. Hence, the pdf of the maximum order statistic $Y_{(n)}$ is

$$f_{Y_{(n)}}(y) = \frac{n!}{(n-1)!1!0!}\{F_Y(y)\}^{n-1} f_Y(y) = ny^{n-1}, \quad 0 < y < 1,$$

   with mean

$$E(Y_{(n)}) = \int_0^1 yny^{n-1}dy = \frac{n}{n+1}.$$

   Similarly, the pdf of the minimum order statistic $Y_{(1)}$ is

$$f_{Y_{(1)}}(y) = \frac{n!}{0!1!(n-1)!} f_Y(y)\{1 - F_Y(y)\}^{n-1} = n(1-y)^{n-1}, \quad 0 < y < 1,$$

   with mean

$$\begin{aligned} E(Y_{(1)}) &= \int_0^1 yn(1-y)^{n-1}dy \\ &= n\frac{\Gamma(2)\Gamma(n)}{\Gamma(2+n)} \int_0^1 \frac{\Gamma(2+n)}{\Gamma(2)\Gamma(n)} y(1-y)^{n-1}dy \\ &= \frac{n!}{(n+1)!} = \frac{1}{n+1}. \end{aligned}$$

   The mean of $R$ can be written as

$$\begin{aligned} E(R) &= E(X_{(n)} - X_{(1)}) \\ &= 2\theta E(Y_{(n)}) - \theta - (2\theta E(Y_{(1)}) - \theta) \\ &= 2\theta \frac{n-1}{n+1}. \end{aligned}$$

   Hence, the unbiased estimator of $\theta$ is $(n+1)R/\{2(n-1)\}$.

(b) Show that $T = n^{-1}\sum_{i=1}^{n}|2X_i|$ is also an unbiased estimator. Without deriving the actual variance of $R$ and $T$, comment on which estimator might have a smaller variance.

**Solution**: One can see that $|X_i|$ follows $U(0,\theta)$, so $E(|X_i|) = \theta/2$ and $E(T) = 2\theta/2 = \theta$, which shows $T$ is an unbiased estimator. Intuitively, $R$ has a smaller variance since it is a function of sufficient statistics. The derivation of the variance demonstrates such intuition.

(c) [**TAKE HOME**] Derive the variance of $R$ and $T$ and comment on which estimator you would prefer.

**Solution**: The variance of $T$ is $\text{Var}(T) = \theta^2/(3n)$, and the variance of $R$ is

$$\text{Var}(R) = \text{Var}(X_{(n)}) + \text{Var}(X_{(1)}) - 2\text{Cov}(X_{(n)}, X_{(1)}),$$

with

$$\begin{aligned}\text{Var}(X_{(n)}) &= 4\theta^2\text{Var}(Y_{(n)}) \\ &= 4\theta^2[E(Y_{(n)}^2) - \{E(Y_{(n)})\}^2] \\ &= 4\theta^2\left(\frac{n}{n+2} - \frac{n^2}{(n+1)^2}\right) \\ &= 4\theta^2\frac{n}{(n+2)(n+1)^2}\end{aligned}$$

where

$$E(Y_{(n)}^2) = \int_0^1 y^2 n y^{n-1} dy = \frac{n}{n+2},$$

and

$$\begin{aligned}\text{Var}(X_{(1)}) &= 4\theta^2\text{Var}(Y_{(1)}) \\ &= 4\theta^2[E(Y_{(1)}^2) - \{E(Y_{(1)})\}^2] \\ &= 4\theta^2\left(\frac{2}{(n+2)(n+1)} - \frac{1}{(n+1)^2}\right) \\ &= 4\theta^2\frac{n}{(n+2)(n+1)^2}\end{aligned}$$

where

$$E(Y_{(1)}^2) = \int_0^1 y^2 n(1-y)^{n-1} dy = n\frac{\Gamma(3)\Gamma(n)}{\Gamma(3+n)} = \frac{2}{(n+2)(n+1)}.$$

The covariance between $X_{(1)}$ and $X_{(n)}$ is ignorable when $n$ is large since they are asymptotically independent. Therefore,

$$\mathrm{Var}\left(\frac{(n+1)}{2(n-1)}R\right) \approx \frac{(n+1)^2}{4(n-1)^2}\frac{n}{(n+2)(n+1)^2}8\theta^2 \approx \frac{2\theta^2}{n^2}.$$

Looking at the variance of $R$ and $T$, we generally would say $R$ is a $n$-convergence estimator and $T$ is a $\sqrt{n}$-convergence estimator. $R$ has a much quicker convergence rate than $T$.

2. In a simple modeling strategy, the time to tumor recurrence after treatment may follow an exponential distribution with probability density function (pdf)

$$f(x|\theta) = \frac{1}{\theta}\exp\left(-\frac{x}{\theta}\right),$$

and survivor function

$$S(x|\theta) = P(X > x|\theta) = \exp\left(-\frac{x}{\theta}\right)$$

Let $X_1, \dots, X_n$ be a random sample of size $n$ from this distribution.

(a) Show that $Q = 2n\bar{X}/\theta$ is a pivotal quantity and use the quantity to find (1-$\alpha$) exact confidence interval for $\theta$.

**Solution**: Since $X_i/\theta$ follows the exponential distribution with mean 1, we know $Q = 2\sum_{i=1}^n X_i/\theta$ follows Gamma$(n, 2)$, which is $\chi^2$ distribution with degree of freedom $2n$, denoted by $\chi^2_{2n}$. Since the distribution is independent of $\theta$, we can conclude $Q$ is a pivotal quantity. To find the (1-$\alpha$) confidence interval for $\theta$, we use

$$1 - \alpha = P\left(\chi^2_{2n,\alpha_1} < \frac{2n\bar{X}}{\theta} < \chi^2_{2n,1-\alpha_2}\right),$$

where $\alpha_1 + \alpha_2 = \alpha$ and get

$$1 - \alpha = P\left(2n\bar{X}/\chi^2_{2n,1-\alpha_2} < \theta < 2n\bar{X}/\chi^2_{2n,\alpha_1}\right),$$

where $\chi^2_{2n,\alpha}$ is $\alpha$ quantile of the $\chi^2_{2n}$ distribution. We now can conclude one can use $(2n\bar{x}/\chi^2_{2n,1-\alpha_2}, 2n\bar{x}/\chi^2_{2n,\alpha_1})$ as the $(1-\alpha)$ confidence interval for $\theta$.

(b) Find the maximum likelihood estimator (MLE) of the survival probability $p = S(x_0|\theta)$ at some point $x_0$ (e.g., 3 months), and derive its large sample distribution.

**Solution**: The MLE of $\theta$ is $\hat{\theta} = \bar{X}$. By the invariance property of the MLE, the MLE of the survival probability $p$ is $\hat{p} = \exp(-x_0/\hat{\theta})$. By Central Limit Theorem (CLT), we know

$$\sqrt{n}(\bar{X} - \theta) \to_d N(0, \theta^2).$$

Applying delta method, one can have

$$\sqrt{n}(g(\bar{X}) - g(\theta)) \to_d N(0, \{g'(\theta)\}^2\theta^2),$$

where $g(\theta) = \exp(-x_0/\theta)$ and $g'(\theta) = x_0\theta^{-2}\exp(-x_0/\theta)$. The limiting variance is $v(\theta) = \{g'(\theta)\}^2\theta^2 = x_0^2\theta^{-2}\exp(-2x_0/\theta)$.

(c) To test whether the survival probability is higher than half at the given point $x_0$, one can postulate a null hypothesis $H_0 : p \leq 0.5$ versus $H_1 : p > 0.5$. To find the uniformly most powerful (UMP) test, a biostatistician decides to make a new random variable $Y_i = I(X_i > x_0)$, where $I(\cdot)$ is the indicator function. Find the UMP test using a random sample $Y_1, \dots, Y_n$ with a significant level $\alpha$.

**Solution**: $Y_1, \dots, Y_n$ follow Ber($p$), which has a MLR property with the sufficient statistic $\sum_{i=1}^n Y_i$. Hence, by Karlin-Rubin Theorem, the rejection region of the UMP test is $R = \{\boldsymbol{y} : \sum_{i=1}^n Y_i > c\}$. To find $c$, we use the type-I error threshold $\alpha$, where

$$\alpha = \sup_{p \leq 0.5} P\left(\sum_{i=1}^n Y_i > c\right).$$

Since $\sum_{i=1}^n Y_i$ follows binomial distribution with parameters $n$ and $p$, the probability in the right hand side of the equation is an increasing function of $p$ (i.e., $\sum_{i=1}^n Y_i$ is more likely larger than $c$ when $p$ is larger). Hence, the supremum occurs at $p = 0.5$. Also, since the distribution of $\sum_{i=1}^n Y_i$ is discrete, we cannot find $c$ that exactly satisfies the equation. Rather, we would find $c$ such that the type-I error probability is the closest but slightly smaller than $\alpha$.

(d) [**TAKE HOME**] One can also use the large sample property in (b) to find an approximate test for the null hypothesis in (c). Derive any approximate test and comment on which test you would prefer.

**Solution**: A straightforward large sample test using (b) is Wald-type test, where the rejection region is

$$R = \left\{\boldsymbol{x} : \left|\frac{\sqrt{n}\{g(\bar{X}) - g(\theta)\}}{g'(\theta)\theta}\right| > z_{1-\alpha/2}\right\}.$$

with $\theta = x_0/\log(2)$. Another option of the Wald-type test is to test $H_0 : \theta = x_0/\log(2)$ versus $H_1 : \theta \neq x_0/\log(2)$. The rejection region of the Wald-type test for this hypothesis testing is

$$R = \left\{ \boldsymbol{x} : \left| \frac{\sqrt{n}(\bar{X} - \theta)}{\theta} \right| > z_{1-\alpha/2} \right\}.$$

Intuitively, the test based on (b) is preferred since $Y_i$ is a dichotomized variable of $X_i$. Say, if we choose $x_0 = \theta$, the mean of $X$, the variance of $\bar{Y} = \sum_{i=1}^n Y_i/n$ is $e^{-1}(1-e^{-1})/n = 0.23/n$, and variance of $g(\bar{X})$ is approximately $e^{-2}/n = 0.13/n$. When $n$ is large, the $g(\bar{X})$ generally has a smaller variance than $\bar{Y}$. However, when $n$ is smaller, the approximation may not be precise and one may prefer an exact test like (c).

3. In the environmental study, the distributions of the concentrations of two air pollutants $X$ and $Y$ can be modeled as follows: the conditional density of $Y$, given $X = x$, can be written as

$$f_Y(y|X = x, \alpha, \beta) = \frac{1}{(\alpha + \beta)x} \exp\left\{ -\frac{y}{(\alpha + \beta)x} \right\}, \quad y > 0, \quad x > 0, \quad \alpha, \beta > 0,$$

and the marginal density of $X$ can be written as

$$f_X(x|\beta) = \frac{1}{\beta} \exp\left( -\frac{x}{\beta} \right), \quad x > 0, \quad \beta > 0.$$

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample of $X$ and $Y$ in $n$ monitoring stations.

(a) Using the fact that the joint probability density function (pdf) $f_{X,Y}(x, y|\alpha, \beta)$ can be written as

$$f_{X,Y}(x, y|\alpha, \beta) = f_Y(y|X = x, \alpha, \beta) f_X(x|\beta),$$

show that two statistics $U_1 = \sum_{i=1}^n X_i$ and $U_2 = \sum_{i=1}^n (Y_i/X_i)$ are joint sufficient statistics for $(\alpha, \beta)$.

**Solution**: The joint pdf can be written as

$$\prod_{i=1}^n f_{X_i,Y_i}(x_i, y_i|\alpha, \beta) = \frac{1}{\beta^n} \exp\left( -\frac{\sum_{i=1}^n x_i}{\beta} \right) \frac{1}{(\alpha + \beta)^n \sum_{i=1}^n x_i} \exp\left\{ -\frac{\sum_{i=1}^n (y_i/x_i)}{(\alpha + \beta)} \right\}.$$

It is not hard to see, by factorization theorem, $U_1 = \sum_{i=1}^n X_i$ and $U_2 = \sum_{i=1}^n (Y_i/X_i)$ are joint sufficient statistics for $(\alpha, \beta)$.

(b) Show that $U_1$ and $U_2$ are uncorrelated, i.e., $\text{Cov}(U_1, U_2) = 0$.

**Solution**: Since $E(X_i) = \beta$ and $E(Y_i|X_i = x_i) = (\alpha + \beta)x_i$, we can have

$$E(Y_i/X_i) = E(X_i^{-1}E(Y_i|X_i)) = E(X_i^{-1}(\alpha + \beta)X_i) = \alpha + \beta,$$

and

$$E(Y_i) = E(E(Y_i|X_i)) = E((\alpha + \beta)X_i) = (\alpha + \beta)\beta.$$

The covariance between $X_i$ and $Y_i/X_i$ can be shown as

$$\begin{aligned}\text{Cov}(X_i, Y_i/X_i) &= E(X_iY_i/X_i) - E(X_i)E(Y_i/X_i) \\ &= (\alpha + \beta)\beta - (\alpha + \beta)\beta \\ &= 0.\end{aligned}$$

Hence, the two statistics $U_1$ and $U_2$ are uncorrelated.

One can also show that $U_1$ and $U_2$ are independent by showing that $V = X$ and $W = Y/X$ are independent via factorizing the joint pdf of $(V, W)$.

(c) **[TAKE HOME]** Derive a 95% confidence interval, either approximate or exact, for the parameter $\gamma = \alpha - \beta$ if $n = 30$, $\hat{\alpha} = 2$, and $\hat{\beta} = 1$, where $\hat{\alpha}$ and $\hat{\beta}$ are maximum likelihood estimators (MLE) of $\alpha$ and $\beta$, respectively.

**Solution**: The easiest approach of deriving the 95% confidence interval is using the large sample property of MLE. Without checking the second derivation, one can find the MLE $\hat{\alpha} = n^{-1}\sum_{i=1}^n (Y_i/X_i) - n^{-1}\sum_{i=1}^n X_i = (U_2 - U_1)/n$ and $\hat{\beta} = n^{-1}\sum_{i=1}^n X_i = U_1/n$. Then, the MLE of $\gamma$ is $\hat{\gamma} = \hat{\alpha} - \hat{\beta} = (U_2 - 2U_1)/n$. Using CLT, we know

$$\sqrt{n}(U_1/n - \beta) \to_d N(0, \beta^2),$$

and

$$\sqrt{n}\{U_2/n - (\alpha + \beta)\} \to_d N(0, (\alpha + \beta)^2),$$

where

$$\begin{aligned}\text{Var}(Y_1/X_1) &= \text{Var}\{E(Y_1/X_1|X_1)\} + E\{\text{Var}(Y_1/X_1|X_1)\} \\ &= 0 + E(X_1^{-2}(\alpha + \beta)^2 X_1^2) \\ &= (\alpha + \beta)^2\end{aligned}$$

Since $U_1$ and $U_2$ are independent, one can show

$$\sqrt{n}\{U_2/n - 2U_1/n - (\alpha - \beta)\} \to_d N(0, (\alpha + \beta)^2 + 4\beta^2),$$

and use the result to construct a $(1-\alpha)$ approximate confidence interval

$$1-\alpha \approx P\left(-z_{1-\alpha/2} < \frac{\sqrt{n}\{(\hat{\alpha}-\hat{\beta})-(\alpha-\beta)\}}{\sqrt{(\hat{\alpha}+\hat{\beta})^2+4\hat{\beta}^2}} < z_{1-\alpha/2}\right)$$

FCL comments: I actually intended to ask for a confidence interval for $\gamma = \alpha+\beta$.

(d) Now, assuming $\beta$ is *known*, derive the explicit expression of the maximum likelihood estimator (MLE) of $\alpha$.

**Solution**: The joint log-likelihood function is proportional to

$$\ell(\alpha) \propto -n\log(\alpha+\beta) - \frac{\sum_{i=1}^n (y_i/x_i)}{(\alpha+\beta)}.$$

Taking the first derivative, we have the score function as

$$U(\alpha) = -n(\alpha+\beta)^{-1} + \sum_{i=1}^n (y_i/x_i)(\alpha+\beta)^{-2}$$

Setting $U(\alpha)=0$, we can solve for MLE of $\alpha$, which can be expressed as

$$\hat{\alpha} = \frac{\partial \ell(\alpha)}{\partial \alpha} = n^{-1}\sum_{i=1}^n (y_i/x_i) - \beta.$$

The observed information is

$$J(\alpha) = -\frac{\partial^2 \ell(\alpha)}{\partial \alpha^2} = -n(\alpha+\beta)^{-2} + 2\sum_{i=1}^n (y_i/x_i)(\alpha+\beta)^{-3}.$$

Since $J(\hat{\alpha}) > 0$, we can claim $\hat{\alpha}$ is the maximizer.

(e) Show that the MLE is an unbiased estimator and its variance reaches the Cramér-Rao Lower Bound (CRLB).

**Solution**: The mean value of $\hat{\alpha}$ is

$$E(\hat{\alpha}) = n^{-1}\sum_{i=1}^n E(Y_i/X_i) - \beta = (\alpha+\beta) - \beta = \alpha,$$

which proves that the MLE is unbiased. Since

$$\begin{aligned}\mathrm{Var}(Y_i/X_i) &= \mathrm{Var}(X_i^{-1}E(Y_i|X_i)) + E(X_i^{-2}\mathrm{Var}(Y_i|X_i))\\ &= \mathrm{Var}(\alpha+\beta) + E((\alpha+\beta)^2)\\ &= (\alpha+\beta)^2,\end{aligned}$$

the variance of $\hat{\alpha}$ is $\mathrm{Var}(\hat{\alpha}) = n^{-2}n\mathrm{Var}(Y_i/X_i) = (\alpha+\beta)^2/n$. Since

$$\begin{aligned}E(J(\alpha)) &= E(-n(\alpha+\beta)^{-2} + 2\sum_{i=1}^{n}(Y_i/X_i)(\alpha+\beta)^{-3})\\ &= -n(\alpha+\beta)^{-2} + 2nE(Y_i/X_i)(\alpha+\beta)^{-3}\\ &= -n(\alpha+\beta)^{-2} + 2n(\alpha+\beta)(\alpha+\beta)^{-3}\\ &= n(\alpha+\beta)^{-2},\end{aligned}$$

we can conclude the variance of MLE reaches the CRLB, which is $(\alpha+\beta)^2/n$.

(f) Again, assuming $\beta$ is *known*, derive the critical regions of the likelihood ratio, score, and Wald-type test to test the null hypothesis $H_0 : \alpha = \beta$ versus $H_1 : \alpha \neq \beta$, when $n$ is large.

**Solution**: The critical region of the likelihood ratio test is

$$R = \{(\boldsymbol{x}, \boldsymbol{y}) : -2\log\lambda(\boldsymbol{x}, \boldsymbol{y}) \geq \chi^2_{1,1-\alpha}\},$$

where $\log\lambda(\boldsymbol{x}, \boldsymbol{y}) = \ell(\beta) - \ell(\hat{\alpha})$. The critical region of the score test is

$$R = \left\{\boldsymbol{y} : \left|\frac{U(\beta)}{\sqrt{I_n(\beta)}}\right| \geq z_{1,1-\alpha/2}\right\},$$

or

$$R = \left\{\boldsymbol{y} : \left|\frac{U(\beta)}{\sqrt{J(\beta)}}\right| \geq z_{1,1-\alpha/2}\right\}.$$

The critical region of the Wald-type test is

$$R = \left\{\boldsymbol{y} : \left|\frac{\sqrt{n}(\hat{\alpha}-\beta)}{\sqrt{I_1^{-1}(\beta)}}\right| \geq z_{1,1-\alpha/2}\right\},$$

where $I_1(\beta) = n^{-1}I_n(\beta) = n^{-1}E(J(\beta))$.