Instructions: You are required to do questions 1(a), 2(a)(b)(c), 3(a)(b)(c)(e), and 4(a)(b). The question 4(c) is a bonus question worth of 10 points. However, your total score will not be over 100 points if you did really well in other questions. Questions 1(b), 3(d), and 3(f) 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 sample of size n from an uniform distribution with pdf

$$f_X(x|\theta) = \frac{1}{\theta}, \quad 0 < x < \theta,$$

where θ is an unknown parameter. Suppose that a researcher wants to test the null hypothesis $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$.

(a) If $\theta_0 = 1/2$, find the cutoff c such that one rejects the null hypothesis with size 0.05 if $X_{(n)} > c$, where $X_{(n)}$ is the maximum order statistic.

Solution: By the definition of α, one has

$$\begin{aligned} 0.05 &= P(X_{(n)} > c|\theta = 1/2) = 1 - P(X_{(n)} \leq c|\theta = 1/2) \\ &= 1 - \{P(X_1 \leq c|\theta = 1/2)\}^n = 1 - (2c)^n. \end{aligned}$$

Solving the equation, one has $c = 2^{-1}(0.95)^{1/n}$.

(b) [**TAKE HOME**] If the true value of θ is actually 3/4, find the sample size n such that one can detect the difference between $\theta = 1/2$ and $\theta = 3/4$ under 0.8 power.

Solution:

$$\begin{aligned} 0.8 &= P(X_{(n)} > 2^{-1}(0.95)^{1/n}|\theta = 3/4) \\ &= 1 - P(X_{(n)} \leq 2^{-1}(0.95)^{1/n}|\theta = 3/4) \\ &= 1 - \{P(X_1 \leq 2^{-1}(0.95)^{1/n}|\theta = 3/4)\}^n \\ &= 1 - \left(\frac{1}{2}(0.95)^{1/n}\frac{4}{3}\right)^n \\ &= 1 - \left(\frac{2}{3}\right)^n 0.95. \end{aligned}$$

Solving the equation above, one can get $n = 3.84$. We generally have $n = 4$, the nearest integer that reaches more than 0.8 power.

2. It is of interest to know the utilization of primary care in a city, where the rate of the primary care physician visit is expressed as the number of out-patient visits per person-year of community residence. Suppose that n adult residents are randomly selected and are asked for the total number of out-patient visits, denoted by Y_i, and the length of residency in years in the city, denoted by x_i, $i = 1, \ldots, n$. Assume that Y_i follows a Poisson distribution with mean λx_i with x_i considered some constants (not a random variable). A hypothesis test for $H_0 : \lambda = \lambda_0$ versus $H_1 : \lambda \neq \lambda_0$ is proposed by a researcher.

   (a) Find the maximum likelihood estimator of λ, denoted by $\hat{\lambda}$.

   Solution: The likelihood function is

$$L(\lambda) = \prod_{i=1}^{n} \frac{(\lambda x_i)^{y_i} \exp(-\lambda x_i)}{y_i!} = \lambda^{\sum_{i=1}^{n} y_i} \exp(-\lambda \sum_{i=1}^{n} x_i) \prod_{i=1}^{n} x_i^{y_i} / \prod_{i=1}^{n} y_i!,$$

   with the log-likelihood function

$$\ell(\lambda) = \sum_{i=1}^{n} \{y_i(\log \lambda + \log x_i) - \lambda x_i - \log(y_i!)\}.$$

   Taking the first derivative, the score function is

$$U(\lambda) = \sum_{i=1}^{n} (y_i \lambda^{-1} - x_i) = 0.$$

   Solving for the equation, one can find the MLE

$$\hat{\lambda} = \frac{\sum_{i=1}^{n} Y_i}{\sum_{i=1}^{n} x_i},$$

   which can be interpreted as an average number of out-patient visits per year. The second derivative is $-\lambda^{-2} \sum_{i=1}^{n} y_i < 0$, which shows the global maximum occurs at $\hat{\lambda}$.

   (b) Derive the likelihood ratio test (LRT) statistic for the hypothesis, show that the critical region of the likelihood ratio test is equivalent to

$$R = \{\boldsymbol{y} : \ \hat{\lambda} \leq c_1^* \text{ or } \hat{\lambda} \geq c_2^*\},$$

   for some cutoffs c_1^* and c_2^*, and comment on how to find these cutoffs for a *level* α test.

Solution: The likelihood ratio test statistic is defined by

$$\Lambda(\boldsymbol{y}) = \frac{L(\lambda_0|\boldsymbol{y})}{L(\hat{\lambda}|\boldsymbol{y})} = \left(\frac{\lambda_0}{\hat{\lambda}}\right)^{\sum_{i=1}^n y_i} \exp\left\{-(\lambda_0 - \hat{\lambda})\sum_{i=1}^n x_i\right\},$$

which is a concave function of $\hat{\lambda}$. That means, one can see that the critical region $R = \{\boldsymbol{y} : \Lambda(\boldsymbol{y}) \le c\}$ of the likelihood ratio test is equivalent to $R = \{\boldsymbol{y} : \hat{\lambda} \le c_1^* \text{ or } \hat{\lambda} \ge c_2^*\}$. One can find these cutoffs using type-I error probability. That is, find c_1^* and c_2^* such that

$$\alpha_1 = P(\hat{\lambda} \le c_1^* | \lambda = \lambda_0) \text{ and } \alpha_2 = P(\hat{\lambda} \ge c_2^* | \lambda = \lambda_0),$$

where $\alpha_1 + \alpha_2 \le \alpha$. Note that when $\lambda = \lambda_0$, $\sum_{i=1}^n Y_i = \hat{\lambda}\sum_{i=1}^n x_i$ follows a Poisson distribution with mean $\lambda_0 \sum_{i=1}^n x_i$. Hence, one also find the cutoff c_1^* and c_2^* that satisfy

$$P\left(\sum_{i=1}^n Y_i \le c_1^* \sum_{i=1}^n x_i\right) + P\left(\sum_{i=1}^n Y_i \ge c_2^* \sum_{i=1}^n x_i\right) \le \alpha.$$

Note also that given a value of α, it is unlikely we can find these cutoffs that makes $\alpha_1 + \alpha_2 = \alpha$. Hence a *level* α test is more realistic than a *size* α test.

(c) Given that the sample size is large, derive the likelihood ratio test (LRT), score test, and Wald test for the hypothesis, by defining the information number as $I_1(\lambda) = n^{-1} I_n(\lambda)$, an average of expected information from n independent (but not identical) patients. Specify the critical region for each test with size α.

Solution: Following results in (a), we know that the expected information is

$$I_n(\lambda) = E\left(\lambda^{-2}\sum_{i=1}^n Y_i\right) = \lambda^{-2}\lambda\sum_{i=1}^n x_i = \lambda^{-1}\sum_{i=1}^n x_i.$$

Hence, the information number $I_1(\lambda) = \lambda^{-1}\bar{x}$. Under the null hypothesis, we know that the large-sample likelihood ratio test statistic is

$$-2\log\Lambda(\boldsymbol{y}) = -2\{\ell(\lambda_0) - \ell(\hat{\lambda})\} = 2\left\{\sum_{i=1}^n y_i \log\left(\frac{\hat{\lambda}}{\lambda_0}\right) - (\hat{\lambda} - \lambda_0)\sum_{i=1}^n x_i\right\},$$

for which one will reject the null hypothesis if $-2\log\Lambda(\boldsymbol{y}) \ge \chi^2_{1,1-\alpha}$. The critical region for the score test is

$$\left\{\boldsymbol{y} : \left|\frac{n^{-1/2}U(\lambda_0)}{\sqrt{I_1(\lambda_0)}}\right| = \left|\frac{\sqrt{n}\lambda_0^{-1}(\bar{y} - \bar{x})}{\sqrt{\lambda_0^{-1}\bar{x}}}\right| \ge z_{1-\alpha/2}\right\}.$$

The critical region for the Wald test is otherwise

$$\left\{ \boldsymbol{y} : \left| \frac{\sqrt{n}(\hat{\lambda} - \lambda_0)}{\sqrt{I_1(\lambda_0)^{-1}}} \right| = \left| \frac{\sqrt{n}(\hat{\lambda} - \lambda_0)}{\sqrt{\lambda_0 \bar{x}^{-1}}} \right| \geq z_{1-\alpha/2} \right\}.$$

3. Let $X_1, \ldots, X_n$ be a random sample of size n from $N(\mu_0, \sigma^2)$, and let $Y_1, \ldots, Y_m$ be a random sample of size m from $N(\mu_1, \sigma^2)$ with *known* $\sigma^2 > 0$. Assume that two samples are mutually independent and that $\mu_1 = \theta\mu_0$.

(a) Show that $(\bar{X}, \bar{Y})$ are (joint) sufficient statistics for (μ_0, θ), where $\bar{X} = n^{-1}\sum_{i=1}^n X_i$ and $\bar{Y} = m^{-1}\sum_{i=1}^m Y_i$.

Solution: The joint pdf of X and Y is

$$\begin{aligned} f(\boldsymbol{x}, \boldsymbol{y}|\mu_0, \theta) &= (2\pi\sigma^2)^{-(n+m)/2} \exp\left[-\frac{1}{2\sigma^2} \left\{ \sum_{i=1}^n (x_i - \mu_0)^2 + \sum_{i=1}^m (y_i - \theta\mu_0)^2 \right\} \right] \\ &\propto \exp\left(\frac{\mu_0}{\sigma^2} \sum_{i=1}^n x_i + \frac{\theta\mu_0}{\sigma^2} \sum_{i=1}^m y_i \right). \end{aligned}$$

By factorization theorem, $(\sum_{i=1}^n X_i, \sum_{i=1}^m Y_i)$, as well as $(\bar{X}, \bar{Y})$, are sufficient statistics for (μ_0, θ).

(b) Show that the maximum likelihood estimator (MLE) of θ is $\hat{\theta} = \bar{Y}/\bar{X}$.

Solution: The joint pdf $f(\boldsymbol{x}, \boldsymbol{y}|\mu_0, \theta)$ can be factorized into a product of two quadratic functions that are maximized at $\mu_0 = \bar{x}$ and $\theta\mu_0 = \bar{y}$, respectively. One hence can claim the MLE of θ is $\hat{\theta} = \bar{Y}/\bar{X}$.

(c) Letting

$$Q = \frac{\bar{Y} - \theta\bar{X}}{\sigma\sqrt{1/m + \theta^2/n}},$$

show that Q is a pivotal quantity. Use this quantity to find an *approximate* $(1 - \alpha)$ confidence interval for θ, and common on whether one can find an *exact* confidence interval for θ.

Solution: Given the distribution of X and Y, one can see that $\bar{X}$ follows a normal distribution $N(\mu_0, \sigma^2/n)$ and $\bar{Y}$ follows a normal distribution $N(\mu_1, \sigma^2/m)$. Since $\mu_1 = \theta\mu_0$, one can see that

$$\bar{Y} - \theta\bar{X} \sim N(0, \sigma^2(1/m + \theta^2/n)),$$

and Q follows $N(0,1)$ which is free of θ. Hence, one can claim Q is a pivotal quantity and knows that

$$1-\alpha = P(z_{\alpha_1} < Q < z_{1-\alpha_2}),$$

where $\alpha = \alpha_1+\alpha_2$ and z_α is the α percentile of the standard normal distribution. However, since the denominator of Q includes unknown θ, the *exact* confidence interval, while may be solved by a numerical method, cannot easily be computed. Instead, one may replace θ in the denominator of Q by $\hat{\theta}$ and claim that

$$1-\alpha = P\left(z_{\alpha_1} < \frac{\bar{Y}-\theta\bar{X}}{\sigma\sqrt{1/m+\hat{\theta}^2/n}} < z_{1-\alpha_2} \right),$$

when $n \to \infty$. Hence, the $(1-\alpha)$ *approximate* confidence interval is

$$(\bar{X}^{-1}(\bar{Y}-z_{1-\alpha_2}\hat{\eta}), \quad \bar{X}^{-1}(\bar{Y}-z_{\alpha_1}\hat{\eta})),$$

where $\hat{\eta}^2 = \sigma^2(1/m+\hat{\theta}^2/n)$.

(d) [**TAKE HOME**] Show that $\hat{\theta}$ is a consistent estimator of θ, i.e., $\hat{\theta} \to_p \theta$, without using the property of MLE directly, i.e., you are not allowed to prove the consistency by saying $\hat{\theta}$ is an MLE so $\hat{\theta}$ is consistent. Is $\hat{\theta}$ an unbiased estimator of θ?

Solution: The consistency of $\hat{\theta}$ can proved using Weak Law of Large Numbers (WLLN) that shows $\bar{Y} \to_p \mu_1 = \mu_0\theta$ and $\bar{X} \to_p \mu_0$. Hence, one can claim $\bar{X}^{-1} \to_p \mu_0^{-1}$ and $\bar{Y}/\bar{X} \to_p \mu 1/\mu_0 = \theta$. Since $E(\bar{X}^{-1}) \neq \mu_0^{-1}$, one can claim that $\hat{\theta}$ is NOT an unbiased estimator of θ.

(e) Assuming μ_0 is known, find the uniformly most powerful (UMP) test with size α for the hypothesis $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$. Specify the cutoff of the rejection region.

Solution: Using Neyman-Pearson Lemma, one can find the rejection region of the UMP test for $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1 > \theta_0$ is

$$R = \left\{ (\boldsymbol{x},\boldsymbol{y}) : \frac{f(\boldsymbol{x},\boldsymbol{y}|\mu_0,\theta_1)}{f(\boldsymbol{x},\boldsymbol{y}|\mu_0,\theta_0)} > c \right\},$$

with some cutoff constant c. One can have

$$\begin{aligned}\frac{f(\boldsymbol{x},\boldsymbol{y}|\mu_0,\theta_1)}{f(\boldsymbol{x},\boldsymbol{y}|\mu_0,\theta_0)} &= \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^m (y_i-\theta_1\mu_0)^2 - \sum_{i=1}^m (y_i-\theta_0\mu_0)^2\right)\right\} \\ &= \exp\left[-\frac{1}{2\sigma^2}\left\{-2\mu_0(\theta_1-\theta_0)\sum_{i=1}^m y_i + \mu_0^2(m\theta_1^2 - n\theta_0^2)\right\}\right].\end{aligned}$$

Since $\theta_1 > \theta_0$, one can see the ratio is an increasing function of $\sum_{i=1}^m y_i$ or $\bar{y} = \sum_{i=1}^m y_i/m$. Hence the rejection of the UMP test is equivalent to

$$R = \{(\boldsymbol{x},\boldsymbol{y}) : \bar{y} > c^*\}.$$

For finding the cutoff c^*, we let

$$\alpha = P_{\theta_0}\left(\bar{Y} > c^*\right) = P_{\theta_0}\left(\frac{\bar{Y}-\theta_0\mu_0}{\sqrt{\sigma^2/m}} > \frac{c^*-\theta_0\mu_0}{\sqrt{\sigma^2/m}}\right).$$

Since under the null hypothesis, we know $\bar{Y}$ follows $N(\theta_0\mu_0, \sigma^2/m)$. Therefore, one can find the cutoff c^* by assigning

$$z_{1-\alpha} = \frac{c^*-\theta_0\mu_0}{\sqrt{\sigma^2/m}}.$$

Solving the equation, we have $c^* = \theta_0\mu_0 + z_{1-\alpha}\sqrt{\sigma^2/m}$. Since the rejection region does not depend on θ_1, the same rejection region will be applied for every $\theta_1 > \theta_0$, i.e., it is the UMP test for the desired hypothesis.

(f) [**TAKE HOME**] Assuming μ_0 is known, find the uniformly most powerful (UMP) test for the hypothesis $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta > \theta_0$.

Solution: The rejection region of the UMP test under the composite versus composite, one-sided, hypothesis testing is the same as in (e) since $\sum_{i=1}^n Y_i$ is a sufficient statistic and the likelihood function (joint pdf) has the monotone likelihood property (MLR) property.

4. Assuming μ_0 is known in Question 3, answer the following questions when n and m are large, i.e., $n, m \to \infty$.

(a) Show that the MLE of θ under this scenario is $\tilde{\theta} = \bar{Y}/\mu_0$, which has large sample normality as

$$\sqrt{m}(\tilde{\theta}-\theta) \to_d N(0, v_1^2),$$

where v_1^2 is the limiting variance. Find v_1^2.

Solution: The joint pdf of X and Y is

$$f(\boldsymbol{x}, \boldsymbol{y}|\theta) = (2\pi\sigma^2)^{-(n+m)/2} \exp\left[-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{n}(x_i - \mu_0)^2 + \sum_{i=1}^{m}(y_i - \theta\mu_0)^2\right\}\right]$$
$$\propto \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i - \theta\mu_0)^2\right\},$$

which is maximized at $\theta = \bar{y}/\mu_0$. One can claim the MLE of θ is $\tilde{\theta} = \bar{Y}/\mu_0$. For the large sample properties, we know that the score function is

$$U(\theta) = \frac{\partial}{\partial\theta}\log f(\boldsymbol{x}, \boldsymbol{y}|\theta) = -\frac{\partial}{\partial\theta}\frac{1}{2\sigma^2}\sum_{i=1}^{m}(y_i - \theta\mu_0)^2$$
$$= \frac{\mu_0}{\sigma^2}\sum_{i=1}^{m}(y_i - \theta\mu_0),$$

and the observed information is

$$J(\theta) = -\frac{\partial}{\partial\theta}U(\theta) = m\frac{\mu_0^2}{\sigma^2},$$

with expected information $I(\theta) = E\{J(\theta)\} = m\mu_0^2/\sigma^2$. By the large sample property of MLE, one has

$$\sqrt{m}(\tilde{\theta} - \theta) \to_d N(0, I_1^{-1}(\theta)),$$

where $I_1(\theta) = I(\theta)/m = \mu_0^2/\sigma^2$. Accordingly, we have

$$\sqrt{m}(\tilde{\theta} - \theta) \to_d N(0, \sigma^2/\mu_0^2).$$

(b) Despite the asymptotic result in (a), one can actually derive the *exact* distribution of $\tilde{\theta}$, i.e., the distribution of $\tilde{\theta}$ when n is finite. Derive the distribution and use it to construct an *exact* $(1 - \alpha)$ confidence interval for θ.

Solution: Since $\bar{Y}$ follows $N(\theta\mu_0, \sigma^2/m)$, one can see

$$\sqrt{m}(\tilde{\theta} - \theta) \sim N(0, \sigma^2/\mu_0^2).$$

This is a *finite* sample property of $\tilde{\theta}$. Hence, one can construct the *exact* $(1-\alpha)$ confidence interval as

$$1 - \alpha = P\left(z_{\alpha_1} < \frac{\sqrt{m}(\tilde{\theta} - \theta)}{\sigma/\mu_0} < z_{1-\alpha_2}\right),$$

which can be shown as

$$\left(\tilde{\theta} - z_{1-\alpha_2}\frac{\sigma/\mu_0}{\sqrt{m}},\ \ \tilde{\theta} - z_{\alpha_1}\frac{\sigma/\mu_0}{\sqrt{m}}\right),$$

where $\alpha = \alpha_1 + \alpha_2$.

(c) [**BONUS**] An investigator thinks that θ should be positive so a confidence interval that covers a non-positive domain may be hard to interpret. To construct a confidence interval that covers only positive domain, one common approach is to first derive the large sample normality of $\log\tilde{\theta}$, such as

$$\sqrt{m}(\log\tilde{\theta} - \log\theta) \rightarrow_d N(0, v_2^2),$$

with limiting variance v_2^2. Then, construct a $(1-\alpha)$ confidence interval for $\log\theta$ and exponentiate both ends to obtain the confidence interval for θ. Find v_2^2 and construct the $(1-\alpha)$ confidence interval for θ as desired. Compare this interval to 3(c) and 4(b) and comment on your preference.

Solution: By Delta Method, one can obtain

$$\sqrt{m}(\log\tilde{\theta} - \log\theta) \rightarrow_d N(0, v_2^2),$$

where $v_2^2 = \{g'(\theta)\}^2\sigma^2/\mu_0^2 = \sigma^2/(\theta\mu_0)^2$. One hence can claim a $(1-\alpha)$ confidence interval for $\log\theta$ as

$$\left(\log\tilde{\theta} - z_{1-\alpha_2}\frac{\sigma/(\tilde{\theta}\mu_0)}{\sqrt{m}},\ \ \log\tilde{\theta} - z_{\alpha_1}\frac{\sigma/(\tilde{\theta}\mu_0)}{\sqrt{m}}\right),$$

where $\alpha = \alpha_1 + \alpha_2$. Exponentiating both ends, the $(1-\alpha)$ confidence interval for θ is

$$\left(\exp\left(\log\tilde{\theta} - z_{1-\alpha_2}\frac{\sigma/(\tilde{\theta}\mu_0)}{\sqrt{m}}\right),\ \ \exp\left(\log\tilde{\theta} - z_{\alpha_1}\frac{\sigma/(\tilde{\theta}\mu_0)}{\sqrt{m}}\right)\right).$$

I would prefer *exact* interval first since the confidence level is exactly $(1-\alpha)$ even when the sample size is small. However, as a biostatistician, we want to interpret the result in a reasonable way, so the interval covers only positive domain will be reported if it is more reasonable to the investigator.

Another possibility to deal with this question is to report

$$\left(\max\left\{0, \tilde{\theta} - z_{1-\alpha_2}\frac{\sigma/\mu_0}{\sqrt{m}}\right\},\ \ \tilde{\theta} - z_{\alpha_1}\frac{\sigma/\mu_0}{\sqrt{m}}\right).$$

This interval also has $(1-\alpha)$ confidence level. Why?