Instructions: You are required to do questions 1(a)(b)(c), 2(a)(b)(c), 3(a)(b)(c). Questions 1(d), 2(d) and 3(b) 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. Pareto distribution remains of interest to healthcare policy researchers. A famous 80/20 principal may well explain the US healthcare expenditures data. The probability density function of the Pareto distribution can be defined as

$$f_Y(y) = a\theta^a y^{-(a+1)}, \quad 0 < \theta < y < \infty, \quad 0 < a < \infty,$$

where θ is the minimum possible value of Y and a is the scale parameter. Let $Y_1, \cdots, Y_n$ be a random sample from $f_Y(y)$.

(a) With a assumed <u>known</u>, show that the maximum likelihood estimator (MLE) of θ is $\hat{\theta} = Y_{(1)}$.

Solution: The MLE, $\hat{\theta}$, can be found to maximize the likelihood function

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f_Y(y_i|\theta) \\ &= a^n \theta^{an} \prod_{i=1}^{n} y_i^{-(a+1)} I(\theta < y_{(1)}). \end{aligned}$$

By drawing the graph of the function, one can easily see that the function is maximized at $\theta = y_{(1)}$. Hence, one can write $\hat{\theta} = Y_{(1)}$.

(b) To do inference on θ, one biostatistician suggests deriving a likelihood ratio test (LRT) to test the null hypothesis $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$. Again, with a assumed <u>known</u>, show that the rejection region of the LRT with size α can be written as $R = \{\boldsymbol{y} : y_{(1)} < \theta_0 \text{ or } y_{(1)} > \theta_0 \alpha^{-1/(an)}\}$, where $y_{(1)}$ is observed minimum value.

Solution: The LRT can be written as

$$\begin{aligned} \lambda(\boldsymbol{y}) &= \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in \Theta} L(\theta)} \\ &= \frac{a^n \theta_0^{an} \prod_{i=1}^{n} y_i^{-(a+1)}}{a^n y_{(1)}^{an} \prod_{i=1}^{n} y_i^{-(a+1)}} I(\theta_0 < y_{(1)}) \\ &= \left(\frac{\theta_0}{y_{(1)}}\right)^{an} I(\theta_0 < y_{(1)}). \end{aligned}$$

Drawing the $\lambda(\boldsymbol{y})$ versus $y_{(1)}$ graph, one can easily see that the rejection region of the LRT $R = \{\boldsymbol{y} : \lambda(\boldsymbol{y}) < c\}$ is equivalent to $R = \{\boldsymbol{y} : y_{(1)} < \theta_0 \text{ or } y_{(1)} > c\}$ for some cutoff c. To find the size equaling α, one can have

$$\begin{aligned}
\alpha &= P(y_{(1)} < \theta_0 \text{ or } y_{(1)} > c|\theta = \theta_0) \\
&= P(Y_{(1)} < \theta_0|\theta = \theta_0) + P(Y_{(1)} > c|\theta = \theta_0) \\
&= P(Y_{(1)} > c|\theta = \theta_0) \\
&= \{P(Y_1 > c|\theta = \theta_0)\}^n \\
&= (\theta_0/c)^{an},
\end{aligned}$$

since

$$P(Y_1 > y|\theta = \theta_0) = 1 - F_Y(y|\theta) = (\theta/y)^a.$$

Hence, $c = \theta_0\alpha^{-1/(an)}$ and the rejection region is $R = \{\boldsymbol{y} : y_{(1)} < \theta_0 \text{ or } y_{(1)} > \theta_0\alpha^{-1/(an)}\}$

(c) Derive the cumulative density function (CDF) of $Y_{(1)}$ and use the CDF to obtain a $(1-\alpha)$ confidence interval for θ.

Solution: The CDF of $Y_{(1)}$, as derived in the solution for (b), is

$$F_{Y_{(1)}}(y|\theta) = 1 - (\theta/y)^{an}.$$

Treating the CDF as a pivotal quantity, one can have

$$\begin{aligned}
1 - \alpha &= P(\alpha_1 < F_{Y_{(1)}}(y|\theta) < 1 - \alpha_2) \\
&= P(\alpha_1 < 1 - (\theta/Y_{(1)})^{an} < 1 - \alpha_2) \\
&= P(\alpha_2 < (\theta/Y_{(1)})^{an} < 1 - \alpha_1) \\
&= P(Y_{(1)}\alpha_2^{1/(an)} < \theta < Y_{(1)}(1 - \alpha_1)^{1/(an)})
\end{aligned}$$

(d) [**TAKE HOME**] Convert the rejection region of the LRT in (b) to obtain a $(1-\alpha)$ confidence interval and compare the interval to the one obtained in (c). Which confidence interval would you prefer?

2. (Continued) A healthcare researcher collected healthcare expenditure data $y_1, \cdots, y_n$, and aims to test the 80/20 principal, which states 20% of patients are responsible for 80% of healthcare expenditures. In the following questions, we assume θ is *known*.

(a) Show that the maximum likelihood estimator (MLE) of a is

$$\hat{a} = \left(n^{-1} \sum_{i=1}^{n} \log(Y_i) - \log(\theta) \right)^{-1}.$$

Solution: The log-likelihood function of a is

$$\ell(a) = n \log(a) + an \log(\theta) - (a+1) \sum_{i=1}^{n} \log(y_i),$$

with score function

$$U(\alpha) = \frac{\partial}{\partial a} \ell(a) = na^{-1} + n \log(\theta) - \sum_{i=1}^{n} \log(y_i).$$

Setting $U(\alpha) = 0$, one can have

$$\hat{a} = \left(n^{-1} \sum_{i=1}^{n} \log(Y_i) - \log(\theta) \right)^{-1}.$$

The second derivative is

$$\frac{\partial^2}{\partial a^2} \ell(a) = -na^{-2} < 0.$$

One can claim $\hat{a}$ is the maximizer.

(b) Find the uniformly most powerful (UMP) test for the null hypothesis $H_0 : a \leq a_0$ versus $H_1 : a > a_0$ with test size α. Specify the cutoff values in your rejection region.

Solution: To apply Karlin-Rubin Theorem, one has to show the distribution has an MLR property. First, one need to find the sufficient statistic for a. Write the single pdf as an exponential family

$$f_{Y_i}(y_i|a) = h(y_i) c(a) \exp\{w(a) t(y_i)\},$$

where $h(y_i) = y_i^{-1} I(y_i > \theta)$, $c(a) = \theta^a$, $w(a) = -a$, and $t(y_i) = \log(y_i)$. By the property of the exponential family, we can claim $T(\boldsymbol{y}) = \sum_{i=1}^{n} \log(Y_i)$ is a sufficient statistic. Now, the likelihood ratio can be written as

$$\frac{L(a_2)}{L(a_1)} = \left(\frac{a_2}{a_1} \right)^n \theta^{n(a_2 - a_1)} \left(\prod_{i=1}^{n} y_i \right)^{-(a_2 - a_1)}.$$

When $a_2 > a_1$, the ratio is a monotone decreasing function of $T(\boldsymbol{y}) = \sum_{i=1}^n \log(Y_i)$ (or monotone increasing function of $T(\boldsymbol{y}) = -\sum_{i=1}^n \log(Y_i)$). Therefore, the pdf has the MLR property, and the rejection region of the UMP test can be written as $R = \{\boldsymbol{y} : \sum_{i=1}^n \log(Y_i) < c\}$.

Let $W = \log(Y) - \log(\theta)$. Using transformation method, it is not hard to show that W follows an exponential distribution with probability density function

$$f_W(w) = a\exp(-aw), \quad 0 < w < \infty, \quad 0 < a < \infty,$$

The inverse function is $Y = \theta\exp(W)$. The probability density function of W is

$$f_W(w) = f_Y(\theta e^w)\theta e^w = a\theta^a\theta^{-(a+1)}e^{-w(a+1)}\theta e^w = ae^{-aw}.$$

Since W_i follows the exponential distribution with mean a^{-1}, we know $\sum_{i=1}^n W_i$ follows Gamma(n, a^{-1}). To find the cutoff c in the UMP test, one have

$$\begin{aligned}\alpha &= \sup_{a\le a_0} P\left(\sum_{i=1}^n \log(Y_i) < c|a\right) \\ &= P\left(\sum_{i=1}^n W_i < c - n\log(\theta)|a = a_0\right).\end{aligned}$$

That makes one to write $c - n\log(\theta) = \Gamma_{n,a_0^{-1},\alpha}$, which is the α quantile of the distribution Gamma(n, a^{-1}). One can choose the cutoff $c = n\log(\theta) + \Gamma_{n,a_0^{-1},\alpha}$.

(c) Let $\pi_{0.8}$ denote the expenditure whereas 20% of the expenditures are higher than that value (i.e., 80% quantile). If the 80/20 principal actually holds, then the summation of expenditures higher than $\pi_{0.8}$ is about 80% of the total expenditures (i.e., heavy upper tail). To estimate $\pi_{0.8}$, one biostatistician suggests using maximum likelihood estimation because, statistically, one can write

$$\int_\theta^{\pi_{0.8}} f_Y(y|a)dy = 0.8.$$

Use the formula above to find the maximum likelihood estimator $\hat{\pi}_{0.8}$ of $\pi_{0.8}$ and derive its large sample distribution.

Solution: The left-hand side of the formula is a cdf cumulative to $\pi_{0.8}$. From 1(b), one can write

$$F_Y(\pi_{0.8}|a) = 1 - (\theta/\pi_{0.8})^a = 0.8,$$

and

$$\pi_{0.8} = \theta(0.2)^{-1/a}.$$

Hence, by invariance property of MLE, the MLE of $\pi_{0.8}$ is

$$\hat{\pi}_{0.8} = \theta(0.2)^{-1/\hat{a}} = \theta 5^{\bar{W}}.$$

Since

$$\sqrt{n}(\bar{W} - a^{-1}) \to_d N(0, a^{-2}),$$

one can have, by delta method,

$$\sqrt{n}\{\hat{\pi}_{0.8} - \pi_{0.8}\} \to_d N(0, \{g'(a^{-1})\}^2 a^{-2}),$$

where $g(x) = \theta 5^x$ and $g'(x) = \theta 5^x \log(5)$. One can claim the large sample distribution of $\hat{\pi}_{0.8}$ follows

$$\sqrt{n}\{\hat{\pi}_{0.8} - \pi_{0.8}\} \to_d N(0, \theta^2 5^{2/a} \log(5)^2 a^{-2}).$$

Since $\log(\hat{\pi}_{0.8}) = \log\theta + \bar{W}\log(5)$, one can also write

$$\sqrt{n}\{\log(\hat{\pi}_{0.8}) - \log(\pi_{0.8})\} \to_d N(0, \log(5)^2 a^{-2}).$$

(d) [**TAKE HOME**] Using the large sample property in (c), derive 95% approximate confidence interval for $\pi_{0.8}$. Comment on how this interval may help interpret the 80/20 principal.

3. (Continued) Other than Pareto distribution, Feenberg and Skinner (1994) use a log-normal distribution with pdf

$$f(y|\mu, \sigma^2) = \frac{1}{y\sqrt{2\pi\sigma^2}} \exp\left\{\frac{-(\log y - \mu)^2}{2\sigma^2}\right\}, \quad 0 < y < \infty, \quad \sigma^2 > 0,$$

to describe the upper tail of the distribution of healthcare cost data. If one define $X = \log(Y)$, it is quite easy to show that the random variable follows $N(\mu, \sigma^2)$.

(a) Show that $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ are complete and sufficient statistics for (μ, σ^2).

Solution: Since X_i follows $N(\mu, \sigma^2)$, one can write the pdf as an exponential family

$$f(x_i|\mu, \sigma^2) = h(x)c(\mu, \sigma^2)\exp\{w_1(\mu, \sigma^2)t_1(x_i) + w_2(\mu, \sigma^2)t_2(x_i)\}.$$

It is not difficult to see $(\sum_{i=1}^n X_i, \sum_{i=1}^n X_i^2)$ is a complete and sufficient statistic.

(b) Find a constant c such that $E(cS) = \sigma$, where

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

[Hint: You may use the fact that $(n-1)S^2/\sigma^2$ follows a chi-square distribution with degree of freedom $(n-1)$.]

Solution: Using the hint, one can derive

$$E(\sqrt{Y}) = \int_0^\infty \sqrt{y} f_Y(y)dy = \frac{\Gamma(n/2)}{\Gamma((n-1)/2)}\sqrt{2},$$

where $Y = (n-1)S^2/\sigma^2$ follows a chi-square distribution with degree of freedom $(n-1)$. Therefore, we have

$$E\left(\frac{\sqrt{n-1}S}{\sigma}\right) = \frac{\Gamma(n/2)}{\Gamma((n-1)/2)}\sqrt{2},$$

and it is not hard to see

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

(c) Let $\eta_{0.8}$ denote the 80% quantile of the distribution of $X = \log(Y)$. Fine the uniformly minimum variance unbiased estimator (UMVUE) of $\eta_{0.8}$, which equals $\log(\pi_{0.8})$.

Solution: Note that, this $\pi_{0.8}$ is different from $\pi_{0.8}$ in the question 2 since the distribution is different. Since $\eta_{0.8}$ is the 80% quantile, we can have

$$P(X > \eta_{0.8}) = P\left(\frac{X-\mu}{\sigma} > \frac{\eta_{0.8}-\mu}{\sigma}\right) = 0.2.$$

Hence, we can write

$$z_{0.8} = \frac{\eta_{0.8}-\mu}{\sigma}$$

and

$$\eta_{0.8} = \mu + z_{0.8}\sigma.$$

By Lehmann-Scheffe theorem, one can claim

$$\eta_{0.8} = \bar{X} + z_{0.8}cS$$

is the UMVUE since $E(\bar{X}) = \mu$ and $E(cS) = \sigma$.

4. In the transmission of polyclonal malaria from human to mosquitos, a historical hypothesis is that the transmission is mediated by a non-random selection process. This is called "bottleneck" in the research of malaria. Assume a human subject contains two unique haplotypes CAM1 and CAM2 with proportions p_0 and $1-p_0$, respectively. After multiple mosquitos were infected by the human subject, frequencies of two haplotypes were collected from each mosquito. Let $n_1, n_2, \ldots, n_m$ are total readings of mosquitos $i = 1, \ldots, m$, and $x_1, \ldots, x_m$ are frequencies of haplotype CAM1. One can consider X_i follows a binomial distribution $B(n_i, p)$ for $i = 1, \ldots, m$.

   (a) To test the bottleneck (a haplotype diminishing/dominating), a biostatistics graduate Jeremy Saxe suggests to use large sample testing to test the null hypothesis $H_0 : p = p_0$ versus $H_1 : p \neq p_0$. Assuming $X_1, \ldots, X_m$ are independent, derive the critical regions of the likelihood ratio, score, and Wald-type test when $n = \sum_{i=1}^m n_i$ is large.

   Solution: The likelihood function of p can be written as

$$L(p) = \prod_{i=1}^m f(x_i|p) = \prod_{i=1}^m \left(\begin{array}{c} n_i \\ x_i \end{array} \right) p^{x_i}(1-p)^{n_i - x_i}$$

   with log-likelihood function as

$$\ell(p) = \sum_{i=1}^m \log f(x_i|p) \propto \log(p) \sum_{i=1}^m x_i + \log(1-p) \sum_{i=1}^m (n_i - x_i).$$

   To find the MLE, one can derive the score function

$$\begin{aligned} U(p) = \partial \ell(p)/\partial p &= p^{-1} \sum_{i=1}^m x_i - (1-p)^{-1} \sum_{i=1}^m (n_i - x_i) = 0 \\ &= (1-p) \sum_{i=1}^m x_i - p \sum_{i=1}^m (n_i - x_i) = 0. \end{aligned}$$

   It is not hard to see that $\hat{p} = \sum_{i=1}^m x_i / \sum_{i=1}^m n_i$. Further, one can derive the observed information function as

$$J(p) = -\partial^2 \ell(p)/\partial p^2 = p^{-2} \sum_{i=1}^m x_i + (1-p)^{-2} \sum_{i=1}^m (n_i - x_i),$$

   which is by any means positive. Therefore, one can claim that $\hat{p}$ is the maximizer. Deriving the score and information function helps establish the large sample

testing. First, one can write the large sample likelihood ratio as

$$\begin{aligned} -2\log\lambda(\boldsymbol{x}) &= 2\{\ell(\hat{p}) - \ell(p_0)\} \\ &= 2\left\{\log\left(\frac{\hat{p}}{p_0}\right)\sum_{i=1}^{m} x_i + \log\left(\frac{1-\hat{p}}{1-p_0}\right)\sum_{i=1}^{m}(n_i - x_i)\right\}, \end{aligned}$$

and the rejection region of the large sample likelihood ratio test with size α can be written as

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

With expected information

$$I_n(p) = E\{J(p)\} = p^{-1}\sum_{i=1}^{m} n_i + (1-p)^{-1}\sum_{i=1}^{m} n_i,$$

one can write the rejection region of the score test as

$$R = \left\{\boldsymbol{x} : \left|\frac{U(p_0)}{\sqrt{I_n(p_0)}}\right| > z_{1-\alpha/2}\right\}.$$

Since the sample is not iid, one can also write the rejection region as

$$R = \left\{\boldsymbol{x} : \left|\frac{n^{-1/2}U(p_0)}{\sqrt{I_1(p_0)}}\right| > z_{1-\alpha/2}\right\},$$

where $I_1(p_0) = n^{-1}I_n(p_0)$. Finally, the rejection region of Wald-test is

$$R = \left\{\boldsymbol{x} : \left|\frac{\hat{p} - p_0}{\sqrt{I_n(p_0)^{-1}}}\right| > z_{1-\alpha/2}\right\}.$$

One can also use

$$R = \left\{\boldsymbol{x} : \left|\frac{\hat{p} - p_0}{\sqrt{I_n(\hat{p})^{-1}}}\right| > z_{1-\alpha/2}\right\},$$

or

$$R = \left\{\boldsymbol{x} : \left|\frac{\hat{p} - p_0}{\sqrt{J(\hat{p})^{-1}}}\right| > z_{1-\alpha/2}\right\}.$$