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 $\hat{\theta} = Y_{(1)}$.

   (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.

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

   (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 <u>*known*</u>.

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

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

   (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.

   (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.

   (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).

(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 $(n-1)$.]

(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 $\log(\pi_{0.8})$.

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.