Instructions: You are required to do questions 1(a)(b), 2(a)(b), 3(a)(b) and 4(a)(b)(c)(d). Questions 1(c), 3(c) and 4(e) 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 from a normal distribution $N(0, \sigma^2)$. To test $H_0 : \sigma^2 \leq 2$ versus $H_1 : \sigma^2 > 2$, answer the following questions in order to find the uniformly most powerful (UMP) test.

   (a) Show that $\sum_{i=1}^n X_i^2$ is a sufficient statistic for σ^2 and that the probability density function of X has the monotone likelihood ratio (MLR) property in $\sum_{i=1}^n X_i^2$.

   Solution: Since the normal distribution is an exponential family, one can show that $\sum_{i=1}^n X_i^2$ is a sufficient statistics for σ^2. The likelihood ratio can be written as

$$\begin{aligned}\frac{L(\sigma_2^2|\boldsymbol{x})}{L(\sigma_1^2|\boldsymbol{x})} &= \left(\frac{\sigma_1^2}{\sigma_2^2}\right)^{n/2} \frac{\exp(-\sum_{i=1}^n x_i^2/(2\sigma_2^2))}{\exp(-\sum_{i=1}^n x_i^2/(2\sigma_1^2))} \\ &= \left(\frac{\sigma_1^2}{\sigma_2^2}\right)^{n/2} \exp\left\{-\sum_{i=1}^n x_i^2 \left(\frac{1}{2\sigma_2^2} - \frac{1}{2\sigma_1^2}\right)\right\}.\end{aligned}$$

   Since when $\sigma_2^2 > \sigma_1^2$, the ratio is an increasing function of $\sum_{i=1}^n x_i^2$, one can conclude that the distribution has an MLR property in $\sum_{i=1}^n x_i^2$.

   (b) Based on the proved conditions in (a), show that the critical region of the UMP test can be written as $R = \{\boldsymbol{x} : \sum_{i=1}^n x_i^2 > c\}$. Find c explicitly given type-I error α, using the fact $\sum_{i=1}^n X_i^2/\sigma^2$ follows a χ^2 distribution with degree of freedom n.

   Solution: Based on the proved properties in (a), one can use Karlin-Rubin Theorem to show that the critical region of the UMP test is $R = \{\boldsymbol{x} : \sum_{i=1}^n x_i^2 > c\}$. One can use type-I error α to find c. Specifically,

$$\begin{aligned}\alpha &= \sup_{\sigma^2 \leq 2} P\left(\sum_{i=1}^n X_i^2 > c\right) \\ &= \sup_{\sigma^2 \leq 2} P\left(\frac{\sum_{i=1}^n X_i^2}{\sigma^2} > \frac{c}{\sigma^2}\right) \\ &= P\left(\frac{\sum_{i=1}^n X_i^2}{\sigma^2} > \frac{c}{2}\right).\end{aligned}$$

   Since $\sum_{i=1}^n X_i^2/\sigma^2$ follows a χ^2 distribution with n d.f., one can see that $c/2 = \chi^2_{n,1-\alpha}$. That makes $c = 2\chi^2_{n,1-\alpha}$.

(c) **[TAKE HOME]** Derive the likelihood ratio test (LRT) for $H_0 : \sigma^2 \leq 2$ versus $H_1 : \sigma^2 > 2$, and comment on whether this critical region is different from the UMP test.

2. Let $X_1, \ldots, X_n$ be a random sample from from a density function

$$f_X(x|\theta) = \theta x^{\theta-1}, \quad 0 < x < 1, \quad \theta > 0.$$

If one propose a confidence interval $(X_{(1)}, X_{(n)})$ for the population median ξ, where the population median satisfies $P(X_i > \xi) = 1/2$ and $P(X_i < \xi) = 1/2$,

(a) Derive the expected length of the confidence interval, i.e., $E(X_{(n)} - X_{(1)})$.

Solution: The expected length is $E(X_{(n)} - X_{(1)}) = E(X_{(n)}) - E(X_{(1)})$, where

$$E(X_{(n)}) = \int_0^1 x f_{X_{(n)}}(x) dx = \int_0^1 x n\theta x^{n\theta-1} dx = \frac{n\theta}{n\theta+1},$$

and

$$\begin{aligned} E(X_{(1)}) &= \int_0^1 x f_{X_{(1)}}(x) dx = \int_0^1 x n\theta x^{\theta-1}(1-x^\theta)^{n-1} dx \\ &= \int_0^1 n\theta t(1-t)^{n-1} \frac{1}{\theta} t^{\frac{1}{\theta}-1} dt \quad (\text{let} \quad t = x^\theta) \\ &= n \int_0^1 t^{\frac{1}{\theta}} (1-t)^{n-1} dt \\ &= nB(\frac{1}{\theta}+1, n) \int_0^1 \frac{1}{B(\frac{1}{\theta}+1, n)} t^{\frac{1}{\theta}} (1-t)^{n-1} dt \\ &= \frac{n\Gamma(\frac{1}{\theta}+1)\Gamma(n)}{\Gamma(\frac{1}{\theta}+1+n)}. \end{aligned}$$

(b) Derive the confidence level $(1-\alpha)$, where $1 - \alpha = P(X_{(1)} < \xi < X_{(n)})$.

Solution: One can have

$$\begin{aligned}
P(X_{(1)} < \xi < X_{(n)}) &= P(\{X_{(1)} < \xi\} \cap \{\xi < X_{(n)}\}) \\
&= P(X_{(1)} < \xi) + P(\xi < X_{(n)}) - P(\{X_{(1)} < \xi\} \cup \{\xi < X_{(n)}\}) \\
&= P(X_{(1)} < \xi) + P(\xi < X_{(n)}) - 1 \\
&= 1 - P(X_{(1)} \geq \xi) + 1 - P(X_{(n)} \leq \xi) - 1 \\
&= 1 - \{P(X_1 \geq \xi)\}^n - \{P(X_1 \leq \xi)\}^n \\
&= 1 - \frac{1}{2^{n-1}}.
\end{aligned}$$

3. Let $X_1, \ldots, X_n$ be a random sample from a Poisson distribution with mean λ.

   (a) Show that $\sqrt{n}(\bar{X} - \lambda)$ converges in distribution to $N(0, \lambda)$ and that $\sqrt{n}(\bar{X} - \lambda)/\sqrt{\lambda}$ is a pivotal quantity when n is large.

   Solution: By the Central Limit Theorem,

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

   One can have

   $$\frac{\sqrt{n}(\bar{X} - \lambda)}{\sqrt{\lambda}} \to_d N(0, 1),$$

   which is independent of λ. One can conclude $\sqrt{n}(\bar{X} - \lambda)/\sqrt{\lambda}$ is a pivotal quantity when n is large.

   (b) Using the result in (a), show that

   $$\left(\bar{x} - z_{1-\alpha/2}\sqrt{\frac{\bar{x}}{n}}, \bar{x} + z_{1-\alpha/2}\sqrt{\frac{\bar{x}}{n}}\right)$$

   is a $(1 - \alpha)$ confidence interval for λ, where $z_{1-\alpha/2}$ is the $(1 - \alpha/2)$ quantile of the standard normal distribution. Comment on whether this interval is an *exact* or *approximate* confidence interval.

   Solution: Since $\sqrt{n}(\bar{X} - \lambda)/\sqrt{\lambda}$ is a pivotal quantity when n is large, one can write

$$
\begin{aligned}
1-\alpha &\approx P\left(-z_{1-\alpha/2} \leq \frac{\sqrt{n}(\bar{X}-\lambda)}{\sqrt{\lambda}} \leq z_{1-\alpha/2}\right) \\
&\approx P\left(-z_{1-\alpha/2} \leq \frac{\sqrt{n}(\bar{X}-\lambda)}{\sqrt{\bar{X}}} \leq z_{1-\alpha/2}\right) \\
&\approx P\left(\bar{X}-z_{1-\alpha/2}\sqrt{\bar{X}/n} \leq \lambda \leq \bar{X}+z_{1-\alpha/2}\sqrt{\bar{X}/n}\right)
\end{aligned}
$$

One can conclude

$$
\left(\bar{x}-z_{1-\alpha/2}\sqrt{\bar{x}/n}, \quad \bar{x}+z_{1-\alpha/2}\sqrt{\bar{x}/n}\right)
$$

is a $(1-\alpha)$ confidence interval for λ.

(c) **[TAKE HOME]** Comment on how one can construct a better confidence interval using the fact that

$$
P(|\sqrt{n}(\bar{X}-\lambda)/\sqrt{\lambda}| \leq z_{1-\alpha/2}) = P(\lambda^2-(2\bar{X}+z^2_{1-\alpha/2}/n)\lambda+\bar{X}^2 \leq 0).
$$

4. Let Y_i be the random variable that follows a geometric distribution with success probability θ_i with

$$
f_{Y_i}(y_i) = (1-\theta_i)^{y_i-1}\theta_i, \quad y_i = 1, 2, \ldots, \quad 0 < \theta_i < 1,
$$

for $i = 1, \ldots, n$. This distribution is useful when describing the discrete time to the first event in biostatistics. For example, researchers may want to know how many "weeks" it takes for *P. vivax* malaria to relapse after a certain treatment. A common approach to model the heterogenous θ_i is assumed

$$
\theta_i = \frac{\beta x_i}{1+\beta x_i},
$$

where x_i is a covariate, e.g., patient's age in the malaria relapse. Given the n pairs (Y_i, x_i), $i = 1, \ldots, n$, of data points, the goal of the analysis is to obtain the maximum likelihood estimator (MLE) of β and use the estimator to make statistical inferences.

(a) Given that the likelihood function is

$$
L(\beta|\boldsymbol{y}) = \prod_{i=1}^{n} f_{Y_i}(y_i|\beta) = \prod_{i=1}^{n}\left(1-\frac{\beta x_i}{1+\beta x_i}\right)^{y_i-1}\frac{\beta x_i}{1+\beta x_i},
$$

write down the log-likelihood function, score function and observed information.

Solution: The log-likelihood function is

$$\begin{aligned}\ell(\beta|\boldsymbol{y}) = \log L(\beta|\boldsymbol{y}) &= \sum_{i=1}^{n}\{-(y_i - 1)\log(1+\beta x_i) + \log(\beta x_i) - \log(1+\beta x_i)\} \\ &= -\sum_{i=1}^{n} y_i \log(1+\beta x_i) + \sum_{i=1}^{n}\log(\beta x_i).\end{aligned}$$

The score function is

$$U(\beta) = \frac{\partial}{\partial\beta}\ell(\beta|\boldsymbol{y}) = -\sum_{i=1}^{n} y_i \frac{x_i}{1+\beta x_i} + n\beta^{-1}.$$

The observed information is

$$J(\beta) = -\frac{\partial^2}{\partial\beta^2}\ell(\beta|\boldsymbol{y}) = -\sum_{i=1}^{n} y_i x_i^2 (1+\beta x_i)^{-2} + n\beta^{-2}.$$

(b) Prove that the MLE, denoted by $\hat{\beta}$, satisfies the equation

$$\hat{\beta}^{-1} = n^{-1}\sum_{i=1}^{n} x_i y_i (1+\hat{\beta}x_i)^{-1}.$$

Solution: Set the score function $U(\beta) = 0$ in (a), one can have

$$\beta^{-1} = n^{-1}\sum_{i=1}^{n} x_i y_i (1+\beta x_i)^{-1}.$$

The MLE $\hat{\beta}$ has to satisfy the above equation.

(c) Show that $\sqrt{n}(\hat{\beta} - \beta) \to_d N(0, v(\beta))$, where the asymptotic variance $v(\beta)$ can be consistently estimated by

$$\hat{v}(\hat{\beta}) = \frac{n\hat{\beta}^2}{\sum_{i=1}^{n}(1+\hat{\beta}x_i)^{-1}}.$$

Solution: By the large sample property of the MLE, we know that $\sqrt{n}(\hat{\beta} - \beta) \to_d N(0, I_1^{-1}(\beta))$, where $I_1(\beta) = E(J(\beta|x_i))$. However, since $Y_1, \ldots, Y_n$ are not identical, a better expression for $I_1(\beta)$ is $I_1(\beta) = \lim_{n\to\infty} n^{-1} I(\beta)$, where

$$\begin{aligned}
I(\beta) = E(J(\beta)) =& E(-\sum_{i=1}^{n} Y_i x_i^2 (1+\beta x_i)^{-2} + n\beta^{-2}) \\
&= -\sum_{i=1}^{n} \{x_i^2 (1+\beta x_i)^{-2} E(Y_i) + \beta^{-2}\} \\
&= -\sum_{i=1}^{n} \{x_i^2 (1+\beta x_i)^{-2} \frac{1+\beta x_i}{\beta x_i} + \beta^{-2}\} \\
&= \beta^{-2} \sum_{i=1}^{n} (1+\beta x_i)^{-1}.
\end{aligned}$$

Hence, one can use $n^{-1} I(\hat{\beta})$ to estimate $I_1(\beta)$ since $\hat{\beta}$ is a consistent estimator of β. That results in a consistent estimator for $v(\beta)$ as

$$\hat{v}(\hat{\beta}) = \{n^{-1} I(\hat{\beta})\}^{-1} = \frac{n\hat{\beta}^2}{\sum_{i=1}^{n} (1+\hat{\beta} x_i)^{-1}}.$$

(d) To test the null hypothesis $H_0 : \beta = 1$ versus $H_1 : \beta \neq 1$, derive the critical regions of the likelihood ratio, score, and Wald-type test when n is large.

Solution: The critical region of the likelihood ratio test is

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

where $\log\lambda(\boldsymbol{y}) = \ell(1|\boldsymbol{y}) - \ell(\hat{\beta}|\boldsymbol{y})$. The critical region of the score test is

$$R = \left\{ \boldsymbol{y} : \left| \frac{U(1)}{\sqrt{I(1)}} \right| \geq \chi^2_{1,1-\alpha/2} \right\},$$

or

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

where

$$\begin{aligned}
U(1) &= -\sum_{i=1}^{n} y_i x_i (1+x_i)^{-1} + n, \\
I(1) &= \sum_{i=1}^{n} (1+x_i)^{-1},
\end{aligned}$$

and

$$J(1) = -\sum_{i=1}^{n} y_i x_i^2 (1 + x_i)^{-2} + n.$$

The critical region of the Wald-type test is

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

where

$$I_1(1) = n^{-1} I(1) = n^{-1} \sum_{i=1}^{n} (1 + x_i)^{-1}.$$

(e) **[TAKE HOME]** If the research has no interest to consider $\beta < 1$, she re-writes the hypothesis as $H_0 : \beta = 1$ versus $H_0 : \beta > 1$, comment on how the test regions in (d) should be adjusted.