1. Let $X_1, \ldots, X_n$ be a random sample from an exponential distribution with probability density function (pdf)

$$f(x) = \frac{1}{\beta} e^{-x/\beta}$$

and cumulative density function (cdf)

$$F(x) = 1 - e^{-x/\beta}, \quad 0 < x < \infty, \quad 0 < \beta < \infty.$$

(a) Show that $F(X_1), \ldots, F(X_n)$ can be considered as a random sample from a uniform distribution between 0 and 1 by showing that

$$P(F(X_i) \leq x) = x,$$

for $i = 1, \ldots, n$.

Solution: Since F for the exponential distribution is a monotone increasing function, we can write

$$P(F(X_i) \leq x) = P(X_i \leq F^{-1}(x)) = F(F^{-1}(x)) = x.$$

The second equation came from the definition of cdf.

(b) Let $X_{(i)}$ be the order statistics from the random sample $X_1, \ldots, X_n$ and let $Z_i = F(X_{(i)})$. Show that the joint distribution of Z_i and Z_j is

$$f_{Z_i,Z_j}(z_i, z_j) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} z_i^{i-1}(z_j - z_i)^{j-i-1}(1 - z_j)^{n-j},$$

where $i < j$ and $0 < z_i < z_j < 1$.

Solution: Since F is a monotone increasing function, $Z_{(i)} = F(X_{(i)})$ are also order statistics. Since Z_i are uniformly distributed from 0 and 1, we can have $f_Z(z) = 1$ and $F_Z(z) = z$. By the distribution formula of order statistics, we can have

$$f_{Z_i,Z_j}(z_i, z_j) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} z_i^{i-1}(z_j - z_i)^{j-i-1}(1 - z_j)^{n-j},$$

where $i < j$ and $0 < z_i < z_j < 1$.

(c) Let $U = Z_j - Z_i$ and $V = Z_i$. Show that the joint distribution of (U, V) is

$$f_{U,V}(u, v) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} v^{i-1}u^{j-i-1}(1 - u - v)^{n-j}.$$

You need to demonstrate the domain of U and V is $u, v > 0$ and $0 < u + v < 1$.

Solution: The inverse functions are $Z_j = U + V$ and $Z_i = V$. The Jacobian is 1. The joint pdf of (U, V) is

$$f_{U,V}(u,v) = \frac{n!}{(i-1)!(j-i-1)!(n-j)!} v^{i-1} u^{j-i-1} (1-u-v)^{n-j}.$$

The domain of U and V can be derived from transformation from Z_i and Z_j, where $0 < Z_i < Z_j < 1$ (a triangle), to U and V (also a triangle).

(d) Show that the marginal distribution of U is

$$f_U(u) = \frac{\Gamma(n+1)}{\Gamma(j-i)\Gamma(n-j+i+1)} u^{j-i-1}(1-u)^{n-j+i},$$

which is pdf of Beta distribution with $\alpha = j - i$ and $\beta = n - j + i + 1$. [Hint: letting $y = v/(1-u)$ will help on solving the complicated integral.]

Solution: The marginal distribution can be derived by

$$\begin{aligned} f_U(u) &= \int_0^{1-u} f_{U,V}(u,v)dv \\ &= \frac{n!}{(i-1)!(j-i-1)!(n-j)!} \int_0^{1-u} v^{i-1} u^{j-i-1} (1-u-v)^{n-j} dv. \end{aligned}$$

Letting $y = v/(1-u)$, we know $(1-u)dy = dv$ and $0 < y < 1$. The integral becomes

$$u^{j-i-1}(1-u)^{n-j+i} \int_0^1 y^{i-1}(1-y)^{n-j} dy,$$

which equals

$$u^{j-i-1}(1-u)^{n-j+i} \frac{\Gamma(i)\Gamma(n-j+1)}{\Gamma(n-j+i+1)} \int_0^1 \frac{\Gamma(n-j+i+1)}{\Gamma(i)\Gamma(n-j+1)} y^{i-1}(1-y)^{n-j} dy,$$

and can be further simplified to

$$u^{j-i-1}(1-u)^{n-j+i} \frac{\Gamma(i)\Gamma(n-j+1)}{\Gamma(n-j+i+1)},$$

using the property of pdf, which is Beta$(i, n-j+1)$. The marginal pdf can then be written as

$$\begin{aligned} f_U(u) &= \frac{\Gamma(n+1)}{\Gamma(i)\Gamma(j-i)\Gamma(n-j+1)} u^{j-i-1}(1-u)^{n-j+i} \frac{\Gamma(i)\Gamma(n-j+1)}{\Gamma(n-j+i+1)} \\ &= \frac{\Gamma(n+1)}{\Gamma(j-i)\Gamma(n-j+i+1)} u^{j-i-1}(1-u)^{n-j+i}, \end{aligned}$$

which is Beta$(j-i, n-j+i+1)$.

(e) A researcher is eager to find a so-called "tolerance interval" $(X_{(i)}, X_{(j)}]$ that covers at least $(100 \times p)$ percent of the distribution at $(100 \times \gamma)$ level. That is, the interval $(X_{(i)}, X_{(j)}]$ satisfies

$$P(F(X_{(j)}) - F(X_{(i)}) \geq p) = \gamma.$$

Given that $i = 1$ and $j = n$, comment on how one can find the probability γ to show the tolerance level of using range $X_{(n)} - X_{(1)}$ to cover at least 80 percent of the distribution.

Solution: Since $F(X_{(j)}) - F(X_{(i)}) = Z_j - Z_i = U$, and we know the distribution of U, we can have the probability $\gamma = \int_p^1 f_U(u)du$, where $p = 0.8$. Further, since we know U follows Beta$(n-1, 2)$ from the result in (d), we can find the probability using either table or software when we have the sample size n.

2. For a women in a certain high-risk population, suppose that the number of lifetime events of domestic violence involving emergency room treatment is assumed to have the Poisson distribution

$$f_X(x|\lambda) = \lambda^x e^{-\lambda}/x!, \quad x = 0, 1, \ldots, \quad \lambda > 0.$$

Let $X_1, \ldots, X_n$ be iid sample randomly chosen for the high-risk population, and each woman in the random sample is asked to recall the number of lifetime events of domestic violence involving emergency room treatment that she has experienced.

(a) Show that the distribution belongs to an exponential family by identifying $h(x)$, $c(\lambda)$, $w(\lambda)$ and $t(x)$, and show that $Y = \sum_{i=1}^n X_i$ is a complete sufficient statistic for λ

Solution: The pdf can be written as

$$f_X(x|\lambda) = \frac{1}{x!} e^{-\lambda} \exp(x \log \lambda) I(x \in \{0, 1, \ldots\}),$$

we can have $h(x) = (x!)^{-1} I(x \in \{0, 1, \ldots\})$, $c(\lambda) = e^{-\lambda}$, $w(\lambda) = \log \lambda$, and $t(x) = x$. Using the property of exponential family, we can claim $\sum_{i=1}^n X_i$ are complete and sufficient statistic for λ.

(b) Let θ be the probability of a woman ever suffering domestic violence in the past, i.e., $\theta = P(X > 0)$. Show that $\theta = 1 - e^{-\lambda}$ and that $\hat{\theta} = 1 - (1 - 1/n)^Y$ is an unbiased estimator of θ using the fact that Y follows Poisson$(n\lambda)$.

Solution: Since X is Poisson, we can have

$$\theta = P(X > 0) = 1 - P(X = 0) = 1 - e^{-\lambda}.$$

For the unbiasedness, we can write

$$\begin{aligned} E\left\{(1 - 1/n)^Y\right\} &= \sum_{y=0}^{\infty} (1 - 1/n)^y \frac{(n\lambda)^y}{y!} \\ &= e^{-n\lambda} \sum_{y=0}^{\infty} \frac{(n\lambda - \lambda)^y}{y!} \\ &= e^{-n\lambda} e^{n\lambda - \lambda} = e^{-\lambda}. \end{aligned}$$

Hence, we can claim that $\hat{\theta}$ is an unbiased estimator of $\theta = 1 - e^{-\lambda}$.

(c) Due to possible recall bias, a researcher decide to dichotomize X_i into

$$Z_i = \begin{cases} 1 & \text{if } X_i > 0 \\ 0 & \text{if } X_i = 0. \end{cases}$$

Let $\bar{Z} = n^{-1} \sum_{i=1}^n Z_i$. Show that $\bar{Z}$ converges in probability to θ and that

$$\sqrt{n}(\bar{Z} - \theta) \to_d N(0, \theta(1 - \theta)).$$

Solution: By Weak Law of Large Numbers (WLLN), $\bar{Z}$ converges in probability to θ, and by Central Limit Theorem (CLT), we know

$$\sqrt{n}(\bar{Z} - \theta) \to_d N(0, \theta(1 - \theta)),$$

since Z_i follows a Bernoulli distribution with probability $\theta = P(Z_i = 1)$.

(d) In order to estimate λ, the researcher suggests transformation on $\bar{Z}$. Find a function g such that $g(\bar{Z})$ converges in probability to λ and show that

$$\sqrt{n}(g(\bar{Z}) - \lambda) \to_d N(0, e^{\lambda} - 1).$$

Solution: Since $\theta = 1 - e^{-\lambda}$, we can write $\lambda = -\log(1 - \theta)$. We then let $g(\theta) = -\log(1 - \theta)$ and we can show that $g(\bar{Z})$ converges in probability $g(\theta)$ since g is a continuous function. By delta method, we can have

$$\sqrt{n}(g(\bar{Z}) - g(\theta)) \to_d N(0, \{g'(\theta)\}^2 \theta(1 - \theta)).$$

We know $g'(\theta) = (1-\theta)^{-1}$. Then the asymptotic variance becomes $\theta/(1-\theta)$. Plugging in $\theta = 1 - e^{-\lambda}$, we have the variance equal $e^{\lambda} - 1$.

(e) If one would be able to use the original sample $X_1, \ldots, X_n$ to estimate λ, one can show that $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$ also converges in probability to λ and that

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

We now have two consistent estimators $g(\bar{Z})$ and $\bar{X}$ for λ. Which one is preferable? That is, which one has a smaller variance when the sample size is large? Give a heuristic reason why the estimator has a smaller variance.

Solution: Again using WLLN and CLT, one have $\bar{X}$ converges in probability to λ and

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

From (d), the asymptotic variance of $g(\bar{Z}) = -\log(1 - \bar{Z})$ is $e^{\lambda} - 1$, which equals $\lambda + \sum_{k=2}^{\infty} k^{-1} \lambda^k$ and apparently larger than λ. We hence can claim that $g(\bar{Z}) = -\log(1 - \bar{Z})$ has a larger asymptotic variance than $\bar{X}$. This result did make sense since $X_1, \ldots, X_n$ have more information than $Z_1, \ldots, Z_n$ about the original distribution.