1. Let $X_1, \ldots, X_n$ be a random sample from an exponential distribution with pdf

$$f_X(x) = \frac{1}{\theta} e^{-x/\theta}, \quad 0 < x < \infty, \quad 0 < \theta < \infty.$$

A researcher is interested in deriving the distribution of a random variable $U = X_1 / \sum_{i=1}^n X_i$, which is a ratio of any given variable to the summation of n variables.

(a) To derive the distribution, one statistician suggests to create another random variable $V = \sum_{i=1}^n X_i$ and write $V = X_1 + Y_1$, where $Y_1 = \sum_{i=2}^n X_i$. That is, one can have $U = X_1/(X_1 + Y_1)$ and $V = X_1 + Y_1$. To use the transformation method to find the distribution of U, find the inverse function of X_1 and Y_1 as a function of U and V and derive the determinant of the Jacobian matrix.

Solution: The inverse functions are $X_1 = UV$ and $Y_1 = V - UV$. The determinant of the Jacobian matrix is $|J| = v$.

(b) Show that the joint pdf of U and V is

$$f_{U,V}(u, v) = \frac{1}{\Gamma(n-1)\theta^n} (1-u)^{n-2} v^{n-1} e^{-v/\theta},$$

using the factor that Y_1, as a summation of $(n-1)$ random variables with an exponential distribution, follows a Gamma distribution with pdf

$$f_{Y_1}(y) = \frac{1}{\Gamma(n-1)\theta^{n-1}} y^{n-2} e^{-y/\theta}, \quad 0 < y < \infty.$$

Solution: The joint pdf of U and V is

$$\begin{aligned} f_{U,V}(u, v) &= f_{X_1,Y_1}(uv, v - uv)|J| \\ &= \frac{1}{\theta} e^{-uv/\theta} \frac{1}{\Gamma(n-1)\theta^{n-1}} (v - uv)^{n-2} e^{-(v-uv)/\theta} v \\ &= \frac{1}{\Gamma(n-1)\theta^n} (1-u)^{n-2} v^{n-1} e^{-v/\theta}. \end{aligned}$$

(c) Make an argument that U and V are independent and derive the marginal distributions of U and V.

Solution: The joint pdf of U and V can be written as

$$f_{U,V}(u, v) = \frac{\Gamma(n)}{\Gamma(n-1)} (1-u)^{n-2} \frac{1}{\Gamma(n)\theta^n} v^{n-1} e^{-v/\theta}.$$

Since the joint pdf can be written as a product of two respective functions of u and v, one can conclude U and V are independent. Specifically, one can see U follows Beta$(1, n-1)$ and V follows Gamma(n, θ). This distribution of V makes sense since $V = \sum_{i=1}^n X_i$.

(d) Show that $V = \sum_{i=1}^n X_i$ is a complete and sufficient statistic and that $U = X_1 / \sum_{i=1}^n X_i$ is an ancillary statistic of θ.

Solution: Since the pdf of the exponential distribution can be written as

$$f_X(x) = h(x)c(\theta)\exp\{w(\theta)t(x)\},$$

where $h(x) = I(0 < x < \infty)$, $c(\theta) = 1/\theta$, $w(\theta) = -1/\theta$, and $t(x) = x$. One can claim that $\sum_{i=1}^n X_i$ is a complete and sufficient statistic. Since the distribution of U is independent of θ, one can claim U is an ancillary statistic.

(e) Let an indicator function $\delta(X_1)$ be defined by

$$\delta(X_1) = \begin{cases} 1 & \text{if } \ X_1 > c, \\ 0 & \text{otherwise}, \end{cases}$$

where c is a constant. Show that

$$E\{\delta(X_1)|\sum_{i=1}^n X_i = t\} = (1 - c/t)^{n-1},$$

and

$$E(X_1|\sum_{i=1}^n X_i = t) = t/n,$$

using the fact that $E\{\delta(X_1)\} = P(X_1 > c)$ and Basu's Theorem.

Solution: One can have

$$\begin{aligned} E\{\delta(X_1)|\sum_{i=1}^n X_i = t\} &= P(X_1 > c|\sum_{i=1}^n X_i = t) \\ &= P\left(\frac{X_1}{\sum_{i=1}^n X_i} > \frac{c}{t}|\sum_{i=1}^n X_i = t\right) \\ &= P\left(\frac{X_1}{\sum_{i=1}^n X_i} > \frac{c}{t}\right) \quad \text{(Basu)} \\ &= \int_{\frac{c}{t}}^1 (n-1)(1-x)^{n-2}dx \\ &= (1 - c/t)^{n-1}, \end{aligned}$$

and

$$\begin{aligned} E\left(X_1 | \sum_{i=1}^n X_i = t\right) &= E\left(\frac{X_1}{\sum_{i=1}^n X_i} t | \sum_{i=1}^n X_i = t\right) \\ &= tE\left(\frac{X_1}{\sum_{i=1}^n X_i}\right) \quad \text{(Basu)} \\ &= \frac{t}{n}. \end{aligned}$$

2. An event occurrence, e.g., mortality or re-hospitalization, can be considered as an *end point* in a clinical trial. A biostatistician tends to use $X_1, \ldots, X_n$ to represent the event occurrence of a random sample of size n in the *control* group and assumes that they follow a Bernoulli distribution with mean θ_1, $0 < \theta_1 < 1$. Similarly, one can let $Y_1, \ldots, Y_n$ represent the event occurrence of a random sample of size n in the *treatment* group and assume they are from a Bernoulli distribution with mean θ_2, $0 < \theta_2 < 1$. One common quantity a biomedical researcher is interested for the comparison between control and treatment groups is called *odds ratio*, which is a ratio of two odds. Answer the following questions and ultimately derive the large sample distribution of the odds ratio estimator.

   (a) Given that $X_1, \ldots, X_n$ follow Bernoulli(θ_1) and that $Y_1, \ldots, Y_n$ follow Bernoulli(θ_2), derive the limiting (asymptotic) distribution of $\bar{X} = n^{-1}\sum_{i=1}^n X_i$ and $\bar{Y} = n^{-1}\sum_{i=1}^n Y_i$ using Central Limit Theorem (CLT).

   Solution: Using the Central Limit Theorem, one can have

   $$\sqrt{n}(\bar{X} - \theta_1) \to_d N(0, \theta_1(1-\theta_1)),$$

   and

   $$\sqrt{n}(\bar{Y} - \theta_2) \to_d N(0, \theta_2(1-\theta_2)).$$

   (b) The odds, defined by $\gamma_1 = \theta_1/(1-\theta_1)$, can be used to describe how large θ_1 is, likewise for $\gamma_2 = \theta_2/(1-\theta_2)$. However, due to a limited range of γ_1 and γ_2, a biostatistician tends to work on log-odds, which is defined by $\log(\gamma_1)$ and $\log(\gamma_2)$, respectively, for control and treatment groups. If one uses $\log(\hat{\gamma}_1) = \log(\bar{X}/(1-\bar{X}))$ and $\log(\hat{\gamma}_2) = \log(\bar{Y}/(1-\bar{Y}))$ to estimate $\log(\gamma_1)$ and $\log(\gamma_2)$, respectively, derive the limiting (asymptotic) distributions of the two log-odds estimators.

   Solution: According to Delta Method,

   $$\sqrt{n}[\log\{\bar{X}/(1-\bar{X})\} - \log(\theta_1/(1-\theta_1))] \to_d N(0, g'(\theta_1)^2\theta_1(1-\theta_1)),$$

where $g(\theta_1) = \log\{\theta_1/(1-\theta_1)\}$ and $g'(\theta_1) = \{\theta_1/(1-\theta_1)\}^{-1}$. Hence, the limiting variance of $\log\{\bar{X}/(1-\bar{X})\}$ equals $g'(\theta_1)^2\theta_1(1-\theta_1) = \{\theta_1/(1-\theta_1)\}^{-1}$. Likewise,

$$\sqrt{n}[\log\{\bar{Y}/(1-\bar{Y})\} - \log(\theta_2/(1-\theta_2))] \to_d N(0, \{\theta_2(1-\theta_2)\}^{-1}).$$

(c) The logarithm of the odds ratio, which is defined by $\log(\gamma_1/\gamma_2)$, can then be estimated by the difference of two log-odds, i.e., $\log(\hat{\gamma}_1) - \log(\hat{\gamma}_2)$. Assuming X and Y are independent, show that the limiting (asymptotic) distribution of $\log(\hat{\gamma}_1/\hat{\gamma}_2)$ is

$$\sqrt{n}\{\log(\hat{\gamma}_1/\hat{\gamma}_2) - \log(\gamma_1/\gamma_2)\} \to_d N(0, \sigma^2),$$

with σ^2 as a function of θ_1 and θ_2.

[Hint: If $X_n \to_d X$ and $Y_n \to_d Y$, then $X_n + Y_n \to_d X + Y$ when X_n and Y_n are independent for each n.]

Solution: According to the result in (b), we can get

$$\sqrt{n}\{\log(\hat{\gamma}_1/\hat{\gamma}_2) - \log(\gamma_1/\gamma_2)\} \to_d N(0, \sigma^2),$$

where $\sigma^2 = \{\theta_1(1-\theta_1)\}^{-1} + \{\theta_2(1-\theta_2)\}^{-1}$.

3. Let $X_1, \ldots, X_n$ be a random sample from a uniform distribution with pdf

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

and cdf

$$F_X(x) = \frac{x}{\theta}, \quad 0 < x < \theta.$$

(a) Show that

$$P(X_{(n)} \leq x) = \left(\frac{x}{\theta}\right)^n,$$

where $X_{(n)}$ is the maximum order statistic, and that, for any $\epsilon \in (0, \theta)$,

$$P(|X_{(n)} - \theta| \leq \epsilon) = 1 - \left(1 - \frac{\epsilon}{\theta}\right)^n,$$

and that $X_{(n)}$ convergence in probability to θ.

Solution: The distribution function of $X_{(n)}$ is

$$F_{X_{(n)}}(x) = P(X_{(n)} \leq x) = P(X_1 \leq x, \ldots, X_n \leq x) = \left(\frac{x}{\theta}\right)^n.$$

Since we have

$$\begin{aligned}
P(|X_{(n)} - \theta| \le \epsilon) &= P(-\epsilon \le X_{(n)} - \theta \le \epsilon) \\
&= P(-\epsilon \le X_{(n)} - \theta \le 0) \\
&= P(\theta - \epsilon \le X_{(n)} \le \theta) \\
&= 1 - \left(1 - \frac{\epsilon}{\theta}\right)^n,
\end{aligned}$$

we can claim $\lim_{n\to\infty} P(|X_{(n)} - \theta| \le \epsilon) = 1$ for any $\epsilon \in (0, \theta)$. That shows $X_{(n)}$ convergence in probability to θ by definition.

(b) Show that $Z_n = n(\theta - X_{(n)})$ converges in distribution to an exponential distribution with mean θ, using the fact that $\lim_{n\to\infty}(1 - x/n)^n = e^{-x}$ for some $x \in (0, n)$.

Solution: By the definition of the cdf of Z_n, we can have

$$\begin{aligned}
F_{Z_n}(z) &= P(Z_n \le z) \\
&= P(n(\theta - X_{(n)}) \le z) \\
&= P(X_{(n)} \ge \theta - z/n) \\
&= 1 - P(X_{(n)} \le \theta - z/n) \\
&= 1 - P(X_1 \le \theta - z/n) \cdots P(X_n \le \theta - z/n) \\
&= 1 - \{(\theta - z/n)/\theta\}^n \\
&= 1 - \{1 - (z/\theta)/n\}^n.
\end{aligned}$$

When $n \to \infty$, $\lim_{n\to\infty} F_{Z_n}(z) = 1 - e^{-z/\theta} = F_Z(z)$, where Z follows an exponential distribution with mean θ.

(c) [Bonus] Show that $Y_n = n\{1 - F_X(X_{(n)})\}$ converges in distribution to an exponential distribution with mean 1.

Solution: The distribution function of Y_n is

$$\begin{aligned}
F_{Y_n}(y) &= P(Y_n \leq y) \\
&= P(n\{1 - F_X(X_{(n)})\} \leq y) \\
&= 1 - P(F_X(X_{(n)}) \leq 1 - y/n) \\
&= 1 - P(X_{(n)} \leq F_X^{-1}(1 - y/n)) \\
&= 1 - P(X_1 \leq F_X^{-1}(1 - y/n), \ldots, X_n \leq F_X^{-1}(1 - y/n)) \\
&= 1 - P(F_X(X_1) \leq 1 - y/n, \ldots, F_X(X_n) \leq 1 - y/n) \\
&= 1 - P(X_1 \leq \theta(1 - y/n), \ldots, X_n \leq \theta(1 - y/n)) \\
&= 1 - \{P(X_1 \leq \theta(1 - y/n))\}^n \\
&= 1 - (1 - y/n)^n.
\end{aligned}$$

The limiting distribution of Y_n is $\lim_{n\to\infty} F_{Y_n}(y) = 1 - e^{-y}$, which is a distribution function of an exponential distribution with mean 1.

We can also express $Y_n = n\{1 - F_X(X_{(n)})\} = n(1 - X_{(n)}/n)$ and use a similar approach in (b). That is,

$$\begin{aligned}
F_{Y_n}(y) &= P(Y_n \leq y) \\
&= P(n(1 - X_{(n)}/\theta) \leq y) \\
&= P(X_{(n)} \geq \theta(1 - y/n)) \\
&= 1 - P(X_{(n)} \leq \theta(1 - y/n)) \\
&= 1 - P(X_1 \leq \theta(1 - y/n)) \cdots P(X_n \leq \theta(1 - y/n)) \\
&= 1 - (1 - y/n)^n,
\end{aligned}$$

and $Y_n \to_d Y$, where Y is an exponential distribution with mean 1.