1. Let X and Y be two random variables with the joint probability density function

$$f(x,y) = \begin{cases} 2(x+y), & 0 \le x \le y \le 1 \\ 0, & \text{otherwise} \end{cases}$$

   (a) Let $Z = X + Y$. Show that the joint probability density function $f_{Y,z}(y,z)$ is

$$f_{Y,Z}(y,z) = 2z, \quad 0 \le -y + z \le y \le 1.$$

   Solution: Let $W = Y$. The joint pdf of W and Z is

$$f_{Y,Z}(y,z) = f_{W,Z}(w,z) = f_{X,Y}(z-w, w)|J| = 2z,$$

   where the Jacobian $|J| = 1$ and the domain of (w,z) can be found by the domain transformation from a triangle to another triangle.

   (b) Derive the conditional probability density function of Y given $Z = z$.

   Solution: The conditional pdf can be derived as

$$f(y|z) = \frac{f_{Y,Z}(y,z)}{f_Z(z)} = \begin{cases} 2/z, & \text{if } 0 \le z \le 1 \\ 2/(2-z), & \text{if } 1 \le z \le 2, \end{cases}$$

   $0 \le -y + z \le y \le 1$, where

$$\begin{aligned} f_Z(z) &= \begin{cases} \int_{z/2}^{z} 2z dy, & \text{if } 0 \le z \le 1 \\ \int_{z/2}^{1} 2z dy, & \text{if } 1 \le z \le 2 \end{cases} \\ &= \begin{cases} z^2, & \text{if } 0 \le z \le 1 \\ z(2-z), & \text{if } 1 \le z \le 2 \end{cases} \end{aligned}$$

2. Let $X_1, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$ population and let

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2.$$

   (a) If μ is unknown and σ^2 is known, show that $\bar{X}_n$ is a complete and sufficient statistic and S_n^2 is an ancillary statistic for μ. Hence, $\bar{X}_n$ and S_n^2 are independent by Basu's Theorem.

   Solution: Using the property of exponential family, one can show that $\bar{X}_n$ is a complete and sufficient statistic. Since $(n-1)S_n^2/\sigma^2$ follows χ^2_{n-1} distribution, which is free of μ, one can claim S_n^2 is an ancillary statistic for μ.

(b) Again, if μ is unknown and σ^2 is known, find the constant c such that

$$E\left(c\bar{X}_n\sum_{i=1}^{n}(X_i-\bar{X}_n)^2\right)=\mu.$$

Solution: Since $\bar{X}_n$ and S_n^2 are independent, one can have

$$\begin{aligned}\mu=E\left(c\bar{X}_n\sum_{i=1}^{n}(X_i-\bar{X}_n)^2\right)&=c\sigma^2E(\bar{X}_n)E\left(\frac{(n-1)S_n^2}{\sigma^2}\right)\\&=c\sigma^2\mu(n-1).\end{aligned}$$

Hence, one can see $c=1/\{\sigma^2(n-1)\}$.

(c) Now, if both μ and σ^2 are unknown and X_{n+1} is a new observation, using the fact the $\bar{X}_n$ and S_n^2 are still independent in this case to find the constant k such that

$$\frac{k(\bar{X}_n-X_{n+1})}{S_n}$$

follows a t distribution. Identify the degree of freedom of the t distribution specifically.

Solution: One can see $(\bar{X}_n-X_{n+1})/\sigma_n$ follows $N(0,1)$, where $\sigma_n=\sigma\sqrt{1+1/n}$, and $(n-1)S_n^2/\sigma^2$ follows χ^2_{n-1}. By the definition of t distribution, one can see

$$T=\frac{(\bar{X}_n-X_{n+1})/\sigma_n}{S_n/\sigma}=\frac{(1+1/n)^{-1/2}(\bar{X}_n-X_{n+1})}{S_n}$$

follows t distribution with $n-1$ degrees of freedom. Hence, $k=(1+1/n)^{-1/2}=\sqrt{n/(n+1)}$.

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

$$f_X(x)=\theta e^{-\theta x},\quad x>0,\quad \theta>0,$$

and cdf

$$F_X(x)=1-e^{-\theta x}.$$

(a) Let $X_{(n)}=\max\{X_1,\ldots,X_n\}$ be the maximum order statistic. Show that a new random variable $Z_{(n)}=F_X(X_{(n)})$ has pdf

$$f_{Z_{(n)}}(z)=nz^{n-1},\quad 0<z<1,$$

and $E(Z_{(n)}) = n/(n+1)$.

Solution: We know that the cdf $F(X)$ follows $U(0,1)$ and $Z_{(n)} = F_X(X_{(n)})$ is the maximum order statistic of a random sample of size n from $U(0,1)$. Therefore, the pdf of $Z_{(n)}$ is

$$f_{Z_{(n)}}(z) = \frac{n!}{(n-1)!} z^{n-1} = nz^{n-1}, \quad 0 < z < 1.$$

The $E(Z_{(n)})$ can be derived by

$$E(Z_{(n)}) = \int_0^1 znz^{n-1} dz = n/(n+1).$$

(b) Find the limiting distribution of $Y_n = \theta X_{(n)} - \log(n)$, using the fact that $\lim_{n\to\infty}(1 - x/n)^n = e^{-x}$ for a constant $x > 0$.

Solution: The cdf of Y_n is

$$\begin{aligned}
F_{Y_n}(y) &= P(Y_n \le y) \\
&= P(\theta X_{(n)} - \log(n) \le y) \\
&= P(X_{(n)} \le \theta^{-1}(y + \log(n))) \\
&= \{P(X \le \theta^{-1}(y + \log(n)))\}^n \\
&= \{1 - \exp(-y - \log(n))\}^n \\
&= \{1 - \exp(-y)/n\}^n.
\end{aligned}$$

When $n \to \infty$, $\lim_{n\to\infty} F_{Y_n}(y) = \exp(-\exp(-y))$, $-\infty < y < \infty$, which is a Gumbel distribution that is frequently used in the extreme value theory.

4. In statistics, homogeneity means equal variance between different groups. Therefore, estimation of variance can be of great interest. Say, a random sample of size n, $X_1, \ldots, X_n$, is collected from $N(0, \theta^2)$. One may use $T_n = n^{-1} \sum_{i=1}^n X_i^2$ to estimate the variance θ^2.

   (a) Show that T_n converges in probability to θ^2 and that the limiting distribution of $\sqrt{n}(T_n - \theta^2)$ is $N(0, 2\theta^4)$. Use the result to construct an approximate 95% confidence interval for θ^2. That is, find an interval (L, U) such that $P(L \le \theta^2 \le U) \approx 0.95$.

Solution: By WLLN, T_n converges in probability to $E(X_1^2) = \text{var}(X_1) + E(X_1)^2 = \theta^2$, and by CLT, $\sqrt{n}(T_n - \theta^2) \to_d N(0, 2\theta^4)$ since

$$\text{var}(X_1^2) = \text{var}(Y_1)\theta^4 = 2\theta^4,$$

where $Y_1 = (X_1/\theta)^2$ follows χ_1^2 distribution with variance $\text{var}(Y_1) = 2$. Using the result, we know $\sqrt{n}(T_n - \theta^2)/(\sqrt{2}\theta^2) \to_d N(0,1)$. However, it is easier to construct the confidence interval using the fact that $\sqrt{n}(T_n - \theta^2)/(\sqrt{2}T_n) \to_d N(0,1)$ by Slutsky Theorem. Hence, we can write

$$\begin{aligned} 0.95 &\approx P\left(-1.96 \le \frac{\sqrt{n}(T_n - \theta^2)}{\sqrt{2}T_n} \le 1.96\right) \\ &= P\left(T_n - 1.96\sqrt{2/n}T_n \le \theta^2 \le T_n + 1.96\sqrt{2/n}T_n\right), \end{aligned}$$

and find (L, U) accordingly.

(b) One way to stabilize the variance estimation is to find a transformation function $g(\cdot)$ such that the limiting variance of $g(T_n)$ is free of θ, or even better, $\sqrt{n}\{g(T_n) - g(\theta^2)\}$ converges in distribution to a random variable whose distribution is free of θ. Provide one such transformation function.

Solution: According to delta method, one have

$$\sqrt{n}\{g(T_n) - g(\theta^2)\} \to_d N(0, \{g'(\theta^2)\}^2 2\theta^4).$$

To make the limiting distribution free of θ, one can have $\{g'(\theta^2)\}^2\theta^4 = 1$, which makes

$$g'(\theta^2) = \theta^{-2}.$$

One can see $g(x) = x^{-1}$ and $g(x) = \log(x)$. The liming distribution becomes

$$\sqrt{n}\{\log(T_n) - \log(\theta^2)\} \to_d N(0, 2).$$

(c) Use the $g(T_n)$ you found in (b) to construct another approximate 95% confidence interval for θ^2. Compare it to the one in (a) and comment on which one you would prefer.

Solution: Using the result in (b), one can have

$$
\begin{aligned}
0.95 &\approx P\left(-1.96 \le \frac{\sqrt{n}\{\log(T_n) - \log(\theta^2)\}}{\sqrt{2}} \le 1.96\right) \\
&= P\left(\log(T_n) - 1.96\sqrt{2/n} \le \log(\theta^2) \le \log(T_n) + 1.96\sqrt{2/n}\right) \\
&= P\left(\exp\{\log(T_n) - 1.96\sqrt{2/n}\} \le \theta^2 \le \exp\{\log(T_n) + 1.96\sqrt{2/n}\}\right).
\end{aligned}
$$