STA355H1 - Assignment 2

1. (a) From the slides of lecture 4, the sample median $M = \begin{cases} (X_{(n/2)} + X_{(n/2+1)})/2 & \text{if } n \text{ is even} \\ X_{((n+1)/2)} & \text{if } n \text{ is odd} \end{cases}$ is an estimator of $F^{-1}(1/2)$. Since $n = 2m$ is even, we have $\hat{\theta}(F) = (X_{(m)} + X_{(m+1)})/2$.

   (b) Note that $\hat{\theta}_{-i} = X_{(m+1)}$ for $i \in \{1, \ldots, m\}$ and $\hat{\theta}_{-i} = X_{(m)}$ for $i \in \{m+1, \ldots, n\}$. Then, $\hat{\theta}_{\bullet} = \frac{1}{n}\Sigma_{i=1}^{n}\hat{\theta}_{-i} = \frac{1}{n}(mX_{(m)} + mX_{(m+1)}) = \frac{1}{2}(X_{(m)} + X_{(m+1)})$. By the compact form of the jackknife estimator,

$$\begin{aligned}
\widehat{\text{Var}}(\hat{\theta}(F)) &= \frac{n-1}{n}\Sigma_{i=1}^{n}(\hat{\theta}_{-i} - \hat{\theta}_{\bullet})^2 = \frac{n-1}{n}\Sigma_{i=1}^{n}(\hat{\theta}_{-i} - \frac{1}{2}(X_{(m)} + X_{(m+1)}))^2 \\
&= \frac{n-1}{n}\left[\Sigma_{i=1}^{n}(\hat{\theta}_{-i})^2 - (X_{(m)} + X_{(m+1)})\Sigma_{i=1}^{n}\hat{\theta}_{-i} + \frac{1}{4}\Sigma_{i=1}^{n}(X_{(m)} + X_{(m+1)})^2\right] \\
&= \frac{n-1}{n}\left[m(X_{(m)}^2 + X_{(m+1)}^2) - m(X_{(m)} + X_{(m+1)})^2 + \frac{m}{2}(X_{(m)} + X_{(m+1)})^2\right] \\
&= \frac{n-1}{n}\left[\frac{m}{2}X_{(m)}^2 - mX_{(m)}X_{(m+1)} + \frac{m}{2}X_{(m+1)}^2\right] = \frac{n(n-1)}{n}(\frac{X_{(m+1)} - X_{(m)}}{2})^2.
\end{aligned}$$

   Thus, we have that $n\widehat{\text{Var}}(\hat{\theta}(F)) = n(n-1)[(X_{(m+1)} - X_{(m)})/2]^2$.

   (c) First, define $U_i \sim \text{Unif}(0,1)$ such that $X_i = g(U_i)$ for all i. By the representation of uniform order statistics using $E_i \sim \text{Exp}(1)$ for all i,

$$\begin{aligned}
n(U_{(m+1)} - U_{(m)}) &\overset{d}{=} n(\frac{E_1 + \ldots + E_{(m+1)}}{E_1 + \ldots + E_{(n+1)}} - \frac{E_1 + \ldots + E_{(m)}}{E_1 + \ldots + E_{(n+1)}}) \\
&= \frac{nE_{(m+1)}}{E_1 + \ldots + E_{(n+1)}} \xrightarrow{d} E_{(m+1)} \equiv W
\end{aligned}$$

   since $(E_1 + \ldots + E_{(n+1)})/n \xrightarrow{p} 1$ by the WLLN. Then, since g has a continuous derivative in a neighborhood of $1/2$, by the mean value theorem,

$$\begin{aligned}
n(X_{(m+1)} - X_{(m)}) &= \frac{g(U_{(m+1)}) - g(U_{(m)})}{U_{(m+1)} - U_{(m)}} n(U_{(m+1)} - U_{(m)}) \\
&= g'(U_{(m)} + c(U_{(m+1)} - U_{(m)}))n(U_{(m+1)} - U_{(m)})
\end{aligned}$$

   where $c \in [0,1]$. Next, $U_{(m)} \xrightarrow{p} 1/2$ by lemma 1 and $c(U_{(m+1)} - U_{(m)}) = cn(U_{(m+1)} - U_{(m)})/n \xrightarrow{d} cW/n \to 0$, so by applying Slutsky's theorem two times we have

$$\begin{aligned}
&g'(U_{(m)} + c(U_{(m+1)} - U_{(m)})) \xrightarrow{d} g'(1/2) \\
\implies &n(X_{(m+1)} - X_{(m)}) \xrightarrow{d} g'(1/2)W.
\end{aligned}$$

   Using the result from (b),

$$\begin{aligned}
n\widehat{\text{Var}}(\hat{\theta}(F)) &= n(n-1)(\frac{X_{(m+1)} - X_{(m)}}{2})^2 = \frac{n-1}{4n}(n(X_{(m+1)} - X_{(m)}))^2 \\
&\xrightarrow{d} \frac{1}{4}(g'(1/2)W)^2 \\
&= \frac{1}{4f^2(F^{-1}(1/2))}\left(\frac{\chi_2^2}{2}\right)^2
\end{aligned}$$

where the second line follows from the continuous mapping theorem and $(n-1)/n \to 1$, and the last line follows from lemmas 2 and 3.

Lemma 1: Claim: $U_{(m)} \xrightarrow{p} \frac{1}{2}$. Proof: Define $S_m = \Sigma_{i=1}^m E_i$ and $S_{n+1} = \Sigma_{i=1}^{n+1} E_i$. Notice that for all $\varepsilon > 0$,

$$\mathbb{P}(U_{(m)} \leq \frac{1}{2} + \varepsilon) = \mathbb{P}(\frac{S_m}{S_{n+1}} \leq \frac{1}{2} + \varepsilon) = \mathbb{P}(\frac{S_m}{m}\frac{n+1}{S_{n+1}}\frac{m}{n+1} \leq \frac{1}{2} + \varepsilon) \to \mathbb{P}(\frac{n}{2n+2} \leq \frac{1}{2} + \varepsilon) = 1$$

as $n \to \infty$ since S_m/m and $(n+1)/S_{n+1}$ both $\to 1$ a.s. by the SLLN. Similarly, it can be shown that $\mathbb{P}(U_{(m)} \leq \frac{1}{2} - \varepsilon) \to 0$ as $n \to \infty$, so

$$\mathbb{P}(|U_{(m)} - \frac{1}{2}| \leq \varepsilon) = \mathbb{P}(U_{(m)} \leq \frac{1}{2} + \varepsilon) - \mathbb{P}(U_{(m)} < \frac{1}{2} - \varepsilon) \to 1 - 0 = 1.$$

Lemma 2: Claim: $g'(1/2) = 1/f(F^{-1}(1/2))$. Proof: By assumption, g has a continuous derivative in a neighborhood of $1/2$, and furthermore assume that this derivative > 0. By the inverse function theorem, $g'(1/2) = (F^{-1})'(1/2) = 1/F'(F^{-1}(1/2)) = 1/f(F^{-1}(1/2))$.

Lemma 3: Claim: $W \overset{d}{=} \chi_2^2/2$. Proof: Define $Z \sim \chi_2^2$, so $f_Z(z) = e^{-z/2}/2$ for $z \geq 0$. Since $w = z/2$ is invertible, by the change of variables technique, $f_W(w) = f_Z(2w)\frac{d}{dw}2w = \frac{2}{2}e^{-2w/2} = e^{-w}$ as expected.

(d) Define S_m and S_{n+1} as in lemma 1. For $U_{(m)}$, we have

$$\sqrt{n}(U_{(m)} - \frac{1}{2}) \overset{d}{=} \sqrt{n}(\frac{S_m}{S_{n+1}} - \frac{1}{2}) = \sqrt{n}(\frac{S_m - S_{n+1}/2}{S_{n+1}}) \xrightarrow{p} \frac{1}{\sqrt{n}}(S_m - \frac{1}{2}S_{n+1})$$

since $\frac{S_{n+1}}{n} = \frac{S_{n+1}}{n+1}\frac{n+1}{n} \xrightarrow{p} 1$ by the WLLN. Since $S_m - S_{n+1}/2 = \Sigma_{i=1}^m E_i/2 - \Sigma_{i=m+1}^{n+1} E_i/2$,

$$\mathbb{E}[\frac{1}{\sqrt{n}}(S_m - \frac{1}{2}S_{n+1})] = \frac{1}{\sqrt{n}}(\frac{m}{2} - \frac{n-m+1}{2}) = \frac{1}{\sqrt{n}}(m - \frac{n}{2} - \frac{1}{2}) = -\frac{1}{2\sqrt{n}} \to 0$$
$$\mathrm{Var}[\frac{1}{\sqrt{n}}(S_m - \frac{1}{2}S_{n+1})] = \frac{1}{n}(\frac{m}{4} + \frac{n-m+1}{4}) = \frac{1}{n}(\frac{n}{4} + \frac{1}{4}) = \frac{1}{4} + \frac{1}{4n} \to \frac{1}{4}$$

by independence of the E_i's and the fact that $\mathbb{E}(E_i) = \mathrm{Var}(E_i) = 1$ for all i. Similarly for $U_{(m+1)}$, we have

$$\sqrt{n}(U_{(m+1)} - \frac{1}{2}) \overset{d}{=} \sqrt{n}(\frac{S_{m+1}}{S_{n+1}} - \frac{1}{2}) = \sqrt{n}(\frac{S_{m+1} - S_{n+1}/2}{S_{n+1}}) \xrightarrow{p} \frac{1}{\sqrt{n}}(S_{m+1} - \frac{1}{2}S_{n+1}),$$

and since $S_{m+1} - S_{n+1}/2 = \Sigma_{i=1}^{m+1} E_i/2 - \Sigma_{i=m+2}^{n+1} E_i/2$,

$$\mathbb{E}[\frac{1}{\sqrt{n}}(S_{m+1} - \frac{1}{2}S_{n+1})] = \frac{1}{\sqrt{n}}(\frac{m}{2} - \frac{n-m}{2}) = \frac{1}{\sqrt{n}}(m - m) = 0$$
$$\mathrm{Var}[\frac{1}{\sqrt{n}}(S_{m+1} - \frac{1}{2}S_{n+1})] = \frac{1}{n}(\frac{m}{4} + \frac{n-m}{4}) = \frac{1}{4}.$$

Thus, by a CLT for weighted sums, $\sqrt{n}(U_{(m)} - \frac{1}{2})$ and $\sqrt{n}(U_{(m+1)} - \frac{1}{2})$ both $\xrightarrow{d} \mathcal{N}(0, 1/4)$. Next,

$$\begin{aligned}
\sqrt{n}(\hat{\theta}(F) - F^{-1}(\frac{1}{2})) &= \sqrt{n}(\frac{X_{(m)} + X_{m+1}}{2} - F^{-1}(\frac{1}{2})) \\
&= \frac{\sqrt{n}}{2}(X_{(m)} - F^{-1}(\frac{1}{2})) + \frac{\sqrt{n}}{2}(X_{(m+1)} - F^{-1}(\frac{1}{2})) \\
&\overset{d}{=} \frac{\sqrt{n}}{2}(F^{-1}(U_{(m)}) - F^{-1}(1/2)) + \frac{\sqrt{n}}{2}(F^{-1}(U_{(m+1)}) - F^{-1}(1/2)) \\
&\xrightarrow{d} \mathcal{N}\Big[0, \frac{1}{4}\frac{1/4}{f^2(F^{-1}(1/2))}\Big] + \mathcal{N}\Big[0, \frac{1}{4}\frac{1/4}{f^2(F^{-1}(1/2))}\Big] \equiv W_1 + W_2
\end{aligned}$$

by the delta method and since $(F^{-1})' = 1/f^2(F^{-1}(1/2))$ as in lemma 2. Notice that

$$\begin{aligned}
\mathbb{E}(W_1 + W_2) &= 0 \\
\text{Var}(W_1 + W_2) &= \text{Var}(W_1) + 2\text{Cov}(W_1, W_2) + \text{Var}(W_2) = 4\text{Var}(W_1)
\end{aligned}$$

since $\text{Cov}(W_1, W_2) = \text{Var}(W_1) = \text{Var}(W_2)$, so we finally have

$$\sqrt{n}(\hat{\theta}(F) - F^{-1}(\frac{1}{2})) \xrightarrow{d} \mathcal{N}\Big[0, \frac{1/4}{f^2(F^{-1}(1/2))}\Big].$$

Assuming that $\mathbb{E}[|X|^r] < \infty$ for $r \geq 2$,

$$n\text{Var}(\hat{\theta}(F)) = \text{Var}\Big[\sqrt{n}(\hat{\theta}(F) - F^{-1}(\frac{1}{2}))\Big] \to \frac{1/4}{f^2(F^{-1}(1/2))}.$$

where the first equality follows since $F^{-1}(1/2)$ is a fixed quantity with no variance.

(e) No. The jackknife estimator has an extra $(\chi^2_2/2)^2$ factor, and since $\mathbb{P}((\chi^2_2/2)^2 \leq 1) = \mathbb{P}(W^2 \leq 1) = \mathbb{P}(W \leq 1) = 1 - e^{-1} < 1 = \mathbb{P}(1 \leq 1)$, we can conclude that $(\chi^2_2/2)^2 \not\to 1$ in distribution. By the contrapositive of convergence in probability implying convergence in distribution, we have $\frac{1/4}{f^2(F^{-1}(1/2))}(\chi^2_2/2)^2 \not\to \frac{1/4}{f^2(F^{-1}(1/2))}$ in probability. Hence, the estimator is not consistent.

2. (a) See the attached code and numerical result. The estimated coverage probability is somewhat low since we expect that an overwhelming majority of confidence intervals of the sample median would contain the true median. Thus, the jackknife estimate seems to have subpar performance, which aligns with my results from part 1 (e).

   (b) See the attached code and numerical result.

   (c) The bootstrap technique outperforms the other since it has a higher estimated coverage probability, which suggests that the bootstrap standard error is relatively more accurate.

3. (a) See the attached code and plot. The plot shows that the variability of the extreme order statistics is greater than that of the middle order statistics.

   (b) See the attached code and plot. Similar to in part (a), the plot shows that the variability of the extreme order statistics is greater than that of the middle order statistics. Additionally, this variability is greater compared to in part (a), suggesting a difference between the normal and logistic distributions.

   (c) The logistic distribution has heavier tails than the normal distribution, which would explain the greater variability in the extreme order statistics of (b). Thus, I conclude that quantile-quantile plots are more useful for distributions with lighter tails.

4. (a) See the attached code, plots, and numerical results. I cannot conclusively determine that the incomes follow a log-normal distribution or not.

   - The QQ plot shows that the central order statistics are quite close to the theoretical normal quantiles. Although the extreme order statistics noticeably deviate from the quantiles, we know from part 1 (a) that they have inherently more variability. Thus, the plot suggests that log(incomes) is likely normally distributed.

   - The Shapiro-Wilk test yields a p-value of 0.00144, which strongly suggests that log(incomes) is not normally distributed.

   - The boxplot of log(incomes) appears somewhat different from the boxplot of a normal sample. In particular, the former boxplot is skewed upwards, with its median and interquartile range closer to its maximum value. Nevertheless, the interquartile range is symmetric around the median, and the skewness can be explained by the variability in the extreme order statistics. Thus, the boxplots suggest that log(incomes) is likely normally distributed.

   - In the line-up of the QQ plots, the true QQ plot looks similar to the other normally-distributed plots, and I was not able to correctly determine the true position even after close inspection. By the line-up protocol, this means that the distribution of log(incomes) is not significantly different from a log-normal distribution.

   (b) See the attached code, plot, and numerical results. The Gini index is closer to 0 than 1, indicating that income inequality is low. This is consistent with the Lorenz plot where the curve is quite close to the line, which implies that the shaded region's area (one-half of the Gini index) is quite small. Finally, the estimated standard errors are relatively low compared to the index, meaning that we are reasonably confident that our index estimate is close to the true index of the population. Furthermore, the estimators are very similar with only a $\sim 2.5\%$ relative error.