STA 532 Final Exam

Jae Hyun Lee, jl914

29 April, 2020

Final Exam for STA-532

1.

(a)

$$
\begin{aligned}
E(f(W)g(Z) \mid W=w) &= \int f(W)g(Z)p(Z \mid W=w)dZ \\
&= \int f(w)g(z)p(z \mid w)dz \\
&= f(w)\int g(z)p(z \mid w)dz \\
&= f(w)\int g(Z)p(Z \mid W=w)dz \\
&= f(w)E(g(Z) \mid W=w)
\end{aligned}
$$

(b)

$$
\begin{aligned}
E(Z) &= \int\int zp(w,z)dzdw \\
&= \int\int zp(z \mid w)p(w)dzdw \\
&= \int p(w)\left(\int zp(z \mid w)dz\right)dw \\
&= \int p(w)E(Z \mid W)dw \\
&= E(E(Z \mid W))
\end{aligned}
$$

(c)

$$
\begin{aligned}
Var(Z) &= E(Z^2) - E(Z)^2 \\
&= E(E(Z^2 \mid W)) - E(E(Z \mid W))^2 \\
&= E(E(Z^2 \mid W)) - E(E(Z \mid W)^2) + E(E(Z \mid W)^2) - E(E(Z \mid W))^2 \\
&= E[E(Z^2 \mid W) - E(Z \mid W)^2] + Var(E(Z \mid W)) \\
&= E(Var(Z \mid W)) + Var(E(Z \mid W))
\end{aligned}
$$

2.

(a)

For positive random variable $Y \in (0,\infty)$ and $c > 0$, $Pr(Y > c) \leq E(Y)/c$.

PF)

$$\begin{aligned} E(Y) = \int_0^\infty yp(y)dy &= \int_0^c yp(y)dy + \int_c^\infty yp(y)dy \\ &\geq \int_c^\infty yp(y)dy \\ &\geq \int_c^\infty cp(y)dy = cPr(Y > c) \end{aligned}$$

which is proof for markovian inequality. Above equation can be applied in this case. Now $Y =\mid X - a \mid \geq 0$ and $c = \epsilon$. Then

$$Pr(\mid X - a \mid \geq \epsilon) = Pr((X-a)^2 \geq \epsilon^2) \leq E((X-a)^2)/\epsilon^2$$

(b)

$$\begin{aligned} E[(X-a)^2] &= E[(X - E(X) + E(X) - a)^2] \\ &= E[(X - E(X))^2] + E[(E(X) - a)^2] + 2E[(X - E(X))(E(X) - a)] \\ &= Var(X) + (E(X) - a)^2 \\ because\ \ 2E[(X-E(X))(E(X)-a)] &= 2(E(X) - a) \times [E(X) - E(X)] = 0 \end{aligned}$$

(c)

$$\begin{aligned} Pr(\mid \hat{\theta} - \theta \mid \geq \epsilon) &\leq E[(\hat{\theta} - \theta)^2]/\epsilon^2 \quad from\ result\ of\ (a) \\ &= \frac{1}{\epsilon^2}[Var(\hat{\theta}) + (E(\hat{\theta}) - \theta)^2] \quad from\ result\ of\ (b) \\ &= \frac{1}{\epsilon^2}[Var(\hat{\theta}) + Bias(\hat{\theta})^2] \quad where\ Bias(\hat{\theta}) = E(\hat{\theta}) - \theta \end{aligned}$$

(d)

$\lim_{n\to\infty} Pr(\mid \ \theta_n - \theta \mid \geq \epsilon) = 0$ if $\hat{\theta}_n$ is consistent estimator of θ. Since $Var(\hat{\theta}_n), Bias^2(\hat{\theta}) \geq 0$. $\lim_{n\to\infty} Var(\hat{\theta}_n) = 0$ and $\lim_{n\to\infty} Bias(\hat{\theta}_n) = 0$ is condition for $\hat{\theta}_n$ to be consistent estimator of θ.

3.

(a)

$$\begin{aligned} \hat{\theta} = \frac{1}{n}\sum \frac{Y_i}{X_i} &\to E(\hat{\theta}) = E(\frac{1}{n}\sum \frac{Y_i}{X_i}) = \sum \frac{1}{n} E(\frac{Y_i}{X_i}) \\ &\to V(\hat{\theta}) = V(\frac{1}{n}\sum \frac{Y_i}{X_i}) = \sum \frac{1}{n^2} V(\frac{Y_i}{X_i}) \end{aligned}$$

$$\begin{aligned}
E(\frac{Y_i}{X_i}) &= \int\int \frac{Y_i}{X_i} p(X_i, Y_i) dX_i dY_i \\
&= \int\int \frac{Y_i}{X_i} p(Y_i \mid X_i) p(X_i) dX_i dY_i \\
&= \int \frac{1}{X_i} p(X_i) \left(\int Y_i P(Y_i \mid X_i) dY_i \right) dX_i \\
&= \int \frac{1}{X_i} p(X_i) \times \theta X_i dX_i = \theta
\end{aligned}$$

$$\begin{aligned}
V(\frac{Y_i}{X_i}) &= E\left[(\frac{Y_i}{X_i} - E(\frac{Y_i}{X_i}))^2 \right] \\
&= E[(\frac{Y_i}{X_i})^2] - E(\frac{Y_i}{X_i})^2 \\
&= E[(\frac{Y_i}{X_i})^2] - \theta^2 \\
E[(\frac{Y_i}{X_i})^2] &= \int\int (\frac{Y_i}{X_i})^2 p(X_i, Y_i) dY_i dX_i \\
&= \int\int (\frac{Y_i}{X_i})^2 p(Y_i \mid X_i) p(X_i) dY_i X_i \\
&= \int \frac{1}{X_i^2} \left(\int Y_i^2 p(Y_i \mid X_i) dY_i \right) p(X_i) dX_i \\
&= \int \frac{1}{X_i^2} E(Y_i^2 \mid X_i) p(X_i) dX_i \\
&= \int \frac{1}{X_i^2} (X_i^2 \theta^2 + X_i \theta) p(X_i) dX_i \\
&= \theta^2 + \theta E(\frac{1}{X_i})
\end{aligned}$$

Thus

$$\begin{aligned}
E(\hat{\theta}) &= \frac{1}{n} n\theta = \theta \\
V(\hat{\theta}) &= \frac{1}{n^2} n(\theta E(\frac{1}{X_i})) = \frac{\theta}{n} E(\frac{1}{X_i})
\end{aligned}$$

(b)

$$\begin{aligned}
P(\mid \hat{\theta} - \theta \mid \geq \epsilon) &\leq V(\hat{\theta})/\epsilon^2 \quad \textit{by chebyshev inequality} \\
&= \frac{\theta}{n\epsilon^2} E(\frac{1}{X_i}) \\
&= \frac{\theta}{n\epsilon^2} E(X_i^{-1}) \to 0 \; as \; n \to \infty \\
&because \; E(\mid X_i \mid^r) < \infty \; for \; all \; r
\end{aligned}$$

(c)

$$
\begin{aligned}
&Let \ \frac{\frac{Y_i}{X_i}-\theta}{\sqrt{\theta E(X_i^{-1})}} \ be \ z_i \sim^{iid} P_{x,y} \ then \ E(z_i)=0 \ V(z_i)=1 \\
&and\sqrt{n}(\hat{\theta}-\theta)/\sqrt{\theta E(X^{-1})} \ be \ \bar{z}=\frac{1}{\sqrt{n}}\sum z_i
\end{aligned}
$$

Then MGF of $\bar{z}$ is as follow:

$$
\bar{z}=E(e^{t\bar{z}})=E(e^{\frac{t}{\sqrt{n}}\sum z_i})=\prod E(e^{\frac{t}{\sqrt{n}}z_i})=M_z(\frac{t}{\sqrt{n}})^n
$$

By taylor approximation

$$
\begin{aligned}
M_z(t)&=1+M_z'(0)t+\frac{M_z''(0)}{2}t^2 \quad M_z'(0)=0 \ M_z''(0)=1 \\
&=1+\frac{1}{2}t^2 \\
&\to M_{\bar{z}}(t)=(1+\frac{1}{2n}t^2)^n \sim e^{\frac{t^2}{2}} \quad which \ is \ MGF \ of \ z\sim N(0,1) \\
&\to \bar{z}\sim N(0,1)\to\sqrt{n}(\hat{\theta}-\theta)\sim N(0,\theta E(X^{-1})) \quad which \ uses \ proof \ of \ CLT.
\end{aligned}
$$

Therefore $V=\theta E(X^{-1})$.

4.

(a)

$$
\begin{aligned}
P(X_i,Y_i)=P(Y_i\mid X_i)P(X_i)&=\frac{1}{Y_i!}(\theta X_i)^{Y_i}e^{-\theta X_i}\times P(X_i) \\
\to likelihood&=\prod\frac{1}{Y_i!}(\theta X_i)^{Y_i}e^{-\theta X_i}\times P(X_i) \\
\to loglik&=logC+\sum Y_i log(\theta X_i)-\theta\sum X_i+l_X \\
&=logC+\sum Y_i log(X_i)+\sum Y_i log(\theta)-\theta\sum X_i+l_X \\
\to\frac{d}{d\theta}loglik&=\frac{1}{\theta}\sum Y_i-\sum X_i\to\hat{\theta}_{mle}=\frac{\sum Y_i}{\sum X_i}
\end{aligned}
$$

(b)

Expectation of $\hat{\theta}_{mle}$ is

$$
\begin{aligned}
E(\hat{\theta}_{mle})=E(\frac{\sum Y_i}{\sum X_i})&=E\left[\frac{1}{\sum X_i}E(\sum Y_i\mid\sum X_i)\right] \\
&=E\left[\frac{1}{\sum X_i}E(\sum Y_i\mid X_1\cdots X_n)\right] \\
&=E\left[\frac{1}{\sum X_i}\sum E(Y_i\mid X_i)\right] \\
&=E\left[\frac{1}{\sum X_i}\sum\theta X_i\right]=\theta
\end{aligned}
$$

Variance of $\hat{\theta}_{mle}$

$$\begin{aligned} V(\hat{\theta}_{mle}) = V(\frac{\sum Y_i}{\sum X_i}) &= V(E(\frac{\sum Y_i}{\sum X_i} \mid X_1 \cdots X_n)) + E(V(\frac{\sum Y_i}{\sum X_i} \mid X_1 \cdots X_n)) \\ &= V(\theta) + E(V(\frac{\sum Y_i}{\sum X_i} \mid X_1 \cdots X_n)) \end{aligned}$$

and

$$V(\frac{\sum Y_i}{\sum X_i} \mid X_1 \cdots X_n) = E\left[\left(\frac{\sum Y_i}{\sum X_i}\right)^2 \mid X_1 \cdots X_n\right] - E\left[\frac{\sum Y_i}{\sum X_i} \mid X_1 \cdots X_n\right]^2$$

After that,

$$\begin{aligned} &(\sum Y_i)^2 = \sum Y_i^2 + \sum_{i \neq j} Y_i Y_j \\ &E(Y_i^2 \mid X_1 \cdots X_n) = E(Y_i^2 \mid X_i) = \theta^2 X_i^2 + \theta X_i \\ &E(Y_i Y_j \mid X_1 \cdots X_n) = E(Y_i Y_j \mid X_i, X_j) = E(Y_i \mid X_i) E(Y_j \mid X_j) = \theta^2 X_i X_j \\ &\to E((\sum Y_i)^2 \mid X_1 \cdots X_n) = \sum (\theta^2 X_i^2 + \theta X_i) + \sum_{i \neq j} \theta^2 X_i X_j = \theta^2 (\sum X_i)^2 + \theta \sum X_i \\ &\to E\left[\left(\frac{\sum Y_i}{\sum X_i}\right)^2 \mid X_1 \cdots X_n\right] = \theta^2 + \theta \frac{1}{\sum X_i} \end{aligned}$$

As a result,

$$\begin{aligned} V(\frac{\sum Y_i}{\sum X_i} \mid X_1 \cdots X_n) &= E\left[\left(\frac{\sum Y_i}{\sum X_i}\right)^2 \mid X_1 \cdots X_n\right] - E\left[\frac{\sum Y_i}{\sum X_i} \mid X_1 \cdots X_n\right]^2 \\ &= \theta^2 + \theta \frac{1}{\sum X_i} - \theta^2 = \theta \frac{1}{\sum X_i} \end{aligned}$$

$$V(\hat{\theta}_{mle}) = V(\theta) + E(\theta \frac{1}{\sum X_i}) = \theta E(\frac{1}{\sum X_i})$$

(c)

$V(\hat{\theta}_{mle}) = \theta E(\frac{1}{\sum X_i})$. Since X_i is positive r.v, $lim_{n \to \infty} \sum X_i = \infty \to lim_{n \to \infty} \frac{1}{\sum X_i} = 0$. As a result, $lim_{n \to \infty} V(\hat{\theta}) \to^p 0$. Then by chebyshev inequality,

$$\begin{aligned} Pr(\mid \hat{\theta}_{mle} - \theta \mid > \epsilon) &\leq E((\hat{\theta}_{mle} - \theta)^2)/\epsilon^2 \\ &= [V(\hat{\theta}_{mle}) + Bias(\hat{\theta}_{mle})^2]/\epsilon^2 \to 0 \end{aligned}$$

because $\hat{\theta}_{mle}$ is unbiased estimator for θ. Consequently, $\hat{\theta}_{mle}$ is consistent estimator of θ.

(d)

$$\hat{\theta}_{mle} - \theta \to^d N(0, V(\hat{\theta}_{mle})) = N(0, \theta\frac{1}{n}E(\frac{1}{\bar{X}}))$$
$$\to \sqrt{n}(\hat{\theta}_{mle} - \theta) \to^d N(0, \theta E(\frac{1}{\bar{X}})) \to V_{mle} = \theta E(\frac{1}{\bar{X}})$$

Now, for concave function f(X) = 1/X, by using Jensen's inequality,

$$V = \theta E(X^{-1}), V_{mle} = \theta E(\frac{1}{\bar{X}}) = \theta\frac{1}{E(X)}$$
$$V = E(f(X)) \geq f(E(X)) = V_{mle}$$

Therefore, we can conclude that V is weakly larger than V_{mle}

5.

(a)

I think the asymptotic normality and efficiency of MLE were the most interesting concept which I learned in this course. The MLE is an estimator for parameters used in most statistical situation. Sample mean which is also widely used in statistical inference is usually revealed to be MLE. Therefore, knowing that MLE asymptotically follows normal distribution is very helpful for most of statistical inference. Moreover, its variance is usually smaller than other estiamtors and this fact gives theoretical support to use MLE over other estimators. Before I took this course, I have learned this concept, but it was not fully understandable. Especially, I could not understand the purpose of Fisher's information function before. Now I have learned that second order derivative's of likelihood gives curvature of likelihood function which indicates how the likelihood function is concentrated nearby MLE.

(b)

asymptotic normality of MLE

$$l'(\theta) = l'(\theta_0) + (\theta - \theta_0)l''(\theta_0) \quad by\ first\ order\ approximation$$
$$\to 0 = l'(\hat{\theta}_{mle}) = l'(\theta_0) + (\hat{\theta}_{mle} - \theta_0)l''(\theta_0)$$
$$\to \hat{\theta}_{mle} = \theta_0 - \frac{l'(\theta_0)}{l''(\theta_0)}$$

If we look at the parts:

$$l'(\theta_0) = \sum S(\theta_0, y_i)\ where\ S\ is\ score\ functon$$
$$l'(\theta_0)/n = \frac{1}{n}\sum S(\theta_0, y_i) = \bar{S}$$
$$\bar{S} \sim N(E(S), V(S)/n) \quad by\ CLT\ and\ E(S) = 0$$
$$\to l'(\theta_0)/n \sim N(0, V(S)/n)$$
$$\to l'(\theta_0)/n \to^d Z\sqrt{\frac{V(S)}{n}} \quad where\ Z \sim N(0,1)$$
$$\to l'(\theta_0) \to^d Z\sqrt{nV(S)}$$

From above two result,

$$
\begin{aligned}
&\hat{\theta}_{mle} \rightarrow^{d} \theta_0 - \frac{Z\sqrt{nV(S)}}{l''(\theta_0)} = \theta_0 - Z\frac{\sqrt{V(S)/n}}{l''(\theta_0)/n} \\
&V(S) = I(\theta_0), l''(\theta_0)/n = \frac{1}{n}\sum \frac{d^2}{d\theta^2} logf(y_i \mid \theta) = -I(\theta_0) \\
&\hat{\theta}_{mle} \rightarrow^{d} \theta_0 + \frac{Z}{\sqrt{nI(\theta_0)}} \\
&\rightarrow \hat{\theta}_{mle} \sim N(\theta_0, \frac{1}{nI(\theta_0)})
\end{aligned}
$$

From this theory, it was able to find that MLE asymtotically follows normal distribution and its variance is Fisher information. As a result, we can utilize useful properties of normal distribution on MLE.

Efficiency of MLE

For any unbiased estimator $\hat{\theta}$

$$
\begin{aligned}
Cor(l'(\theta_0), \hat{\theta})^2 \leq 1 \rightarrow \frac{Cov(l'(\theta_0), \hat{\theta})^2}{V(l'(\theta_0))V(\hat{\theta})} \leq 1 \rightarrow \frac{Cov(l'(\theta_0), \hat{\theta})^2}{V(l'(\theta_0))} &\leq V(\hat{\theta}) \\
Cov(l'(\theta_0, \hat{\theta})) = E(l'(\theta_0)\hat{\theta}) - E(l'(\theta_0))E(\hat{\theta}) &= E(l'(\theta_0)\hat{\theta}) \\
&= \int \hat{\theta}\left[\frac{d}{d\theta} log \prod f(y_i \mid \theta)\right] \prod f(y_i \mid \theta) dy_1 \cdots dy_n \\
&= \int \hat{\theta} \frac{\frac{d}{d\theta}\prod f(y_i \mid \theta)}{\prod f(y_i \mid \theta)} \prod f(y_i \mid \theta) dy_1 \cdots dy_n \\
&= \int \hat{\theta} \frac{d}{d\theta} \prod f(y_i \mid \theta) dy_1 \cdots dy_n \\
&= \frac{d}{d\theta} \int \hat{\theta} \prod f(y_i \mid \theta) dy_1 \cdots y_n \\
&= \frac{d}{d\theta} E(\hat{\theta}) = 1 \\
\rightarrow V_{mle} &= \frac{1}{nI(\theta_0)} \leq V(\hat{\theta})
\end{aligned}
$$

Not only MLE can use useful properties of normal distribution but also its variance is smaller than other estimators. From this theories, it is understandable why MLE is widely used in various statistical inference over other estimators.

(c)

At 4.(a) $Y_i \sim Possion(\theta X_i)$ and I have found that its loglikelihood function is $logC + \sum Y_i log(X_i) + \sum Y_i log(\theta) - \theta \sum X_i + l_{X_i}$ and mle is $\theta_{mle} = \frac{\sum Y_i}{\sum X_i}$. Moreover, we have specified asymptotic distribution of $\hat{\theta}_{mle} - \theta \rightarrow^{d} N(0, V(\hat{\theta}_{mle})) = N(0, \theta\frac{1}{n}E(\frac{1}{X}))$ which indicates that MLE is normally distributed even though Y actually does not follow normal distribution. Moreover, as shown at 4.(d), its variance is smaller than sample mean of statistics Y_i/X_i. Therefore, even though Y_i does not follow normal distribution, MLE makes it possible to use nice properties of normal distributio keeping its variance is smaller than variance of nonparametric estimation.

(d)

In traditional statistical setting, it has used parametric inference rather than nonparametric inference. As we have shown in 4.(d), if our specification of parameter is accurate, it is able to find superiority of MLE over naive sample mean in their precison. The variance of MLE is smaller than widely used nonparametric estimator sample mean. In addition, theories decribed at 5.(a), 5.(b) gives support for conventions in statistical inference. Therefore, it was very impressive for me to learn these difference between parametric estimator MLE and nonparametric estiamtor sample mean.