Sta532 Hw7

Xu Chen 3/6/2020

Exercise 1

(a) The moment generating function of $X \sim \mathsf{N}(\mu, \sigma^2)$ is

$$\begin{aligned}
M_X(t) = \mathsf{E}[e^{tX}] &= \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2 + tx\right) \mathrm{d}x \\
&= \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x^2 - 2(\mu+\sigma^2 t)x + \mu^2)\right) \mathrm{d}x \\
&= \exp\left(-\frac{1}{2\sigma^2}(\mu^2 - (\mu+\sigma^2 t)^2)\right) \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(x^2 - 2(\mu+\sigma^2 t)x + (\mu+\sigma^2 t)^2)\right) \mathrm{d}x \\
&= \exp\left(\mu t + \sigma^2 t^2/2\right)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathsf{E}[Y_i] &= M_X(1) = \exp(\mu + \sigma^2/2) \\
\mathsf{Var}[Y_i] &= \mathsf{E}[Y_i^2] - \mathsf{E}[Y_i]^2 = M_X(2) - M_X^2(1) = \exp(2\mu + 2\sigma^2) - \exp(2\mu + \sigma^2) \\
\mathsf{E}[\bar{Y}] &= \exp(\mu + \sigma^2/2) \\
\mathsf{Var}[\bar{Y}] &= \frac{1}{n}\mathsf{Var}[Y_i] = \frac{1}{n}\left[\exp(2\mu + 2\sigma^2) - \exp(2\mu + \sigma^2)\right]
\end{aligned}$$

(b) The likelihood function of (μ, σ^2) is

$$L(\mu, \sigma^2) = (\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu)^2\right)$$

and hence the log-likelihood is

$$\ell(\mu, \sigma^2) = -\frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (X_i - \mu)^2$$

We solve for the MLE by setting the partial derivatives to be 0.

$$\begin{aligned}
\sum_{i=1}^n (X_i - \mu) &= 0 \\
-\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^n (X_i - \mu)^2 &= 0
\end{aligned}$$

The MLE for (μ, σ^2) is given by $(\hat{\mu}, \hat{\sigma}^2) = (\bar{X}, \sum_{i=1}^n (X_i - \bar{X})^2/n)$. According to the invariance property of MLE (homework 6), the MLE of $\phi = \exp(\mu + \sigma^2/2)$ is

$$\hat{\phi} = \exp(\hat{\mu} + \hat{\sigma}^2/2) = \exp\left(\sum_{i=1}^n \log Y_i/n + \sum_{i=1}^n (\log Y_i - \sum_{i=1}^n \log Y_i/n)^2/(2n)\right)$$

It can be cumbersome to calculate the variance of $\hat{\phi}$ directly. As it is a function of $(\hat{\mu}, \hat{\sigma}^2)$ whose asymptotic covariance matrix is easier to calculate, we try to approximate the asymptotic variance of $\hat{\phi}$ by using Delta's method. The asymptotic theory of MLE tells us

$$\sqrt{n}\left(\begin{pmatrix} \hat{\mu} \\ \hat{\sigma}^2 \end{pmatrix} - \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix}\right) \xrightarrow{\text{d}} \mathsf{N}(0, I(\mu, \sigma^2)^{-1})$$

where $I(\mu, \sigma^2)^{-1} = \begin{pmatrix} \sigma^2 & 0 \\ 0 & 2\sigma^4 \end{pmatrix}$ (shown in the lecture notes).

Therefore, using Delta's method, we have

$$\sqrt{n}(\hat{\phi} - \phi) = \sqrt{n}(f(\hat{\mu}, \hat{\sigma}^2) - f(\mu, \sigma^2)) \xrightarrow{\text{d}} \mathsf{N}(0, \nabla f(\mu, \sigma^2)^\top I(\mu, \sigma^2)^{-1} \nabla f(\mu, \sigma^2))$$

where $f(x, y) = \exp(x + y^2/2)$ and hence $\nabla f(x, y) = (e^{x+y/2}, e^{x+y/2}/2)^\top$. Therefore, we have

$$\sqrt{n}(\hat{\phi} - \phi) \xrightarrow{\text{d}} \mathsf{N}\left(0, \sigma^2 e^{2\mu+\sigma^2}(1 + \sigma^2/2)\right)$$

Therefore,

$$\mathsf{Var}[\hat{\phi}] \approx \frac{\sigma^2 e^{2\mu+\sigma^2}(1 + \sigma^2/2)}{n}$$

Given the expression of $\mathsf{Var}[\bar{Y}]$, we are essentially comparing $\exp(\sigma^2) - 1$ and $\sigma^2 + \sigma^4/2$. According to the Taylor's expansion of $e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$, we have

$$e^x > 1 + x + \frac{x^2}{2}$$

for all $x > 0$. Therefore, we have $\exp(\sigma^2) - 1 > \sigma^2 + \sigma^4/2$ and hence $\mathsf{Var}[\bar{Y}] > \mathsf{Var}[\hat{\phi}]$.

(c) We set $\mu = 0$ and investigate how the biases, variances and MSEs of two estimators change for different values of σ^2. We let σ vary from 2 to 6. For each value of σ, we simulate 10^4 datasets with each dataset containing 10^3 samples. We calculate three measures and provide the result in the following figure.

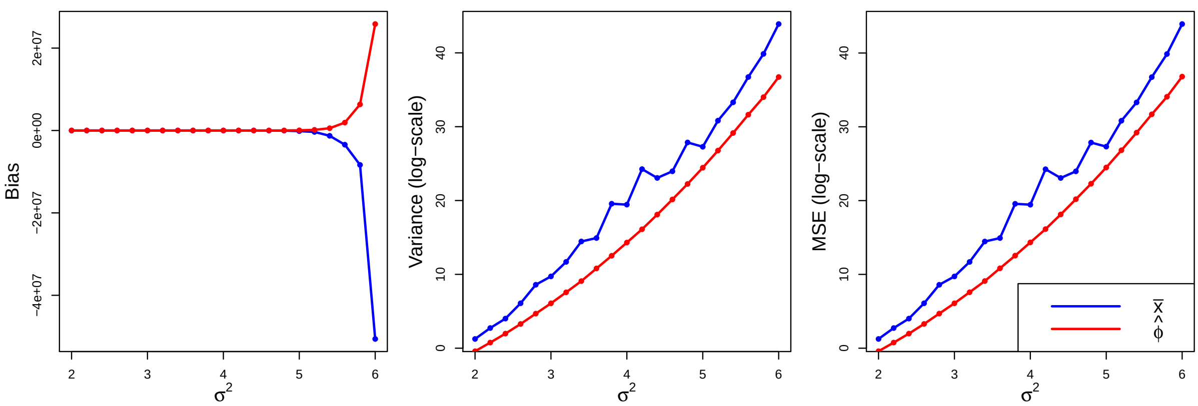

Figure 1: Simulation results.

Biases, variances and MSEs all increase as σ^2 increases. These results are consistent with the theoretical results we derived in part **(a)** and **(b)**. Particularly, both variance and MSE of $\hat{\phi}$ are smaller than those of $\bar{Y}$. We should expect that with a larger number of datasets, the curves should become smoother as the results are closer to the theoretical analysis.

Exercise 2

(a) The log-likelihood is

$$\ell_n(a,b) = na\log b - n\log\Gamma(a) + (a-1)\sum_{i=1}^n \log Y_i - b\sum_{i=1}^n Y_i$$

and hence the likelihood equations are

$$\begin{aligned}
\frac{\partial}{\partial a}\ell_n(a,b) &= n\log b - n\Gamma'(a)/\Gamma(a) + \sum_{i=1}^n \log Y_i = 0\\
\frac{\partial}{\partial b}\ell_n(a,b) &= na/b - \sum_{i=1}^n Y_i = 0
\end{aligned}$$

(b) The second order derivatives are

$$\begin{aligned}
\frac{\partial^2}{\partial a^2}\ell(a,b) &= -(\log\Gamma(a))'' = -(\Gamma''(a)\Gamma(a) - (\Gamma'(a))^2)/\Gamma^2(a)\\
\frac{\partial^2}{\partial b^2}\ell(a,b) &= -a/b^2\\
\frac{\partial^2}{\partial a\partial b}\ell(a,b) &= 1/b
\end{aligned}$$

Therefore,

$$I(a,b) = -\mathsf{E}\left(\frac{\partial^2}{\partial(a,b)^2}\ell(a,b)\right) = \begin{pmatrix} (\log\Gamma(a))'' & -1/b \\ -1/b & a/b^2 \end{pmatrix}$$

Therefore, we have

$$\sqrt{n}\left(\begin{pmatrix}\hat{a}\\ \hat{b}\end{pmatrix} - \begin{pmatrix}a\\ b\end{pmatrix}\right) \xrightarrow{\text{d}} \mathsf{N}(0, I(a,b)^{-1}).$$

(c) By Delta's method we have

$$\sqrt{n}(\hat{\mu}-\mu) = \sqrt{n}(f(\hat{a},\hat{b}) - f(a,b)) \xrightarrow{\text{d}} \mathsf{N}(0, \nabla f(a,b)^\top I(a,b)^{-1}\nabla f(a,b))$$

where $f(x,y) = x/y$ and hence $\nabla f(x,y) = (1/y, -x/y^2)^\top$. Therefore,

$$\sqrt{n}(\hat{\mu}-\mu) \xrightarrow{\text{d}} \mathsf{N}(0, \nabla f(a,b)^\top I(a,b)^{-1}\nabla f(a,b))$$

We have

$$I(a,b)^{-1} = \frac{1}{\frac{a}{b^2}(\log\Gamma(a))'' - \frac{1}{b^2}} \begin{pmatrix} a/b^2 & 1/b \\ 1/b & (\log\Gamma(a))'' \end{pmatrix}$$

and hence

$$\nabla f(a,b)^\top I(a,b)^{-1} \nabla f(a,b) = a/b^2$$

Therefore,

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{\mathrm{d}} \mathsf{N}(0, a/b^2)$$

Remark 1. Instead of using Delta's method to figure out the asymptotic distribution of $\hat{\mu}$, we can try directly finding out $\hat{\mu}$ in terms of samples as follows. Let us reparametrize Gamma distribution in terms of (a,μ) where $\mu = a/b$. Under this parametrization, the pdf of $\mathsf{Gamma}(a,\mu)$ is

$$f(x) = \frac{(a/\mu)^a}{\Gamma(a)} x^{a-1} \exp(-ax/\mu)$$

So here if $Y \sim \mathsf{Gamma}(a,\mu)$, then we have $\mathsf{E}[Y] = \mu$. The log-likelihood with n samples is

$$\ell(a,\mu) = na(\log a - \log \mu) - n \log\Gamma(a) + (a-1)\sum_{i=1}^n \log Y_i - a/\mu \sum_{i=1}^n Y_i$$

The MLE of (a,μ) can be solved by

$$\begin{aligned} &\frac{\partial}{\partial \mu}\ell(a,\mu) = -\frac{na}{\mu} + \frac{a\sum_{i=1}^n Y_i}{\mu^2} = 0 \\ &\frac{\partial}{\partial a}\ell(a,\mu) = 0 \end{aligned}$$

Regardless of the second equation, we notice that the solution for the first equation is $\hat{\mu} = \sum_{i=1}^n Y_i/n$. Therefore, the MLE of μ is sample mean. Then, by CLT, we immediately have

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow{\mathrm{d}} \mathsf{N}(0, a/b^2)$$

Since we actually proved that $\hat{\mu} = \bar{Y}$, their variances are exactly same.

Exercise 3

(a) The log-likelihood is

$$\ell(\mu,\sigma^2) = -\frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(Y-\mu)^2$$

and hence the expected log-likelihood is

$$\begin{aligned}
\mathsf{E}[\ell(\mu,\sigma^2)] &= -\frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\mathsf{E}[(Y-\mu)^2] \\
&= -\frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\mathsf{E}[Y^2+\mu^2-2\mu Y] \\
&= -\frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(\mu^2 - 2\mu a/b + (a/b)^2 + a/b^2\right)
\end{aligned}$$

Setting the partial derivatives to be 0,

$$\begin{aligned}
\frac{\partial}{\partial\mu}\mathsf{E}[\ell(\mu,\sigma^2)] &= -\frac{1}{2\sigma^2}(2\mu - 2a/b) = 0 \\
\frac{\partial}{\partial\sigma^2}\mathsf{E}[\ell(\mu,\sigma^2)] &= -\frac{1}{2\sigma^2} + \frac{1}{2\sigma^4}(\mu^2 - 2\mu a/b + (a/b)^2 + a/b^2) = 0
\end{aligned}$$

Therefore, $(\mu,\sigma^2) = (a/b, a/b^2)$ will maximize the expected log-likelihood.

(b) Since we are assuming a normal model, the MLE is given by

$$\hat\mu = \bar Y, \quad \hat\sigma^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - \bar Y)^2$$

Since $Y_i \overset{\text{iid}}{\sim} \mathsf{Gamma}(a,b)$, its fourth order moment is finite. Therefore, by WLLN, we have

$$\begin{aligned}
\hat\mu = \frac{1}{n}\sum_{i=1}^n Y_i &\to \mathsf{E}[Y_i] = \frac{a}{b} \\
\frac{1}{n}\sum_{i=1}^n (Y_i - \mathsf{E}[Y_1])^2 &\to \mathsf{E}[(Y_i - \mathsf{E}[Y_1])^2] = \frac{a}{b^2}
\end{aligned}$$

in probability. Therefore,

$$\begin{aligned}
\hat\sigma^2 &= \frac{1}{n}\sum_{i=1}^n \left(Y_i - \mathsf{E}[Y_i] + \mathsf{E}[Y_i] - \bar Y\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^n \left(Y_i - \mathsf{E}[Y_i])^2 - (\mathsf{E}[Y_i] - \bar Y\right)^2
\end{aligned}$$

Since $\bar Y \to \mathsf{E}[Y_i]$ in probability and hence by continuous mapping theorem, we have $(\bar Y - \mathsf{E}[Y_i])^2 \to 0$ in probability. Therefore, we have $\hat\sigma^2 \to a/b^2$ in probability.

(c) If assuming a normal likelihood, using the CLT shown in the lecture notes, we have

$$\frac{\sqrt{n}(\hat\mu - \mu)}{\hat\sigma} \xrightarrow{\text{d}} \mathsf{N}(0,1)$$

and hence the standard error we use is $\hat\sigma/\sqrt{n}$. If we correctly assumes the Gamma model, then from the previous problem, we have

$$\frac{\sqrt{n}(\hat\mu - \mu)}{\sqrt{a/b^2}} \xrightarrow{\text{d}} \mathsf{N}(0,1)$$

and from part **(b)** we know that $\hat{\sigma}^2 \to a/b^2$ in probability. Therefore, we have

$$\frac{\sqrt{n}(\hat{\mu}-\mu)}{\sqrt{a/b^2}}\frac{\sqrt{a/b^2}}{\hat{\sigma}} \xrightarrow{\text{d}} \mathsf{N}(0,1)$$
$$\frac{\sqrt{n}(\hat{\mu}-\mu)}{\hat{\sigma}} \xrightarrow{\text{d}} \mathsf{N}(0,1)$$

Hence, we also use $\hat{\sigma}/\sqrt{n}$ as the standard error of $\hat{\mu}$ if we correctly specify the model. By a similar argument, an alternative estimate of standard error of $\hat{\mu}$ under the Gamma model is $\sqrt{\hat{a}}/\sqrt{n\hat{b}^2}$.

(d) According to part **(b)**, we know that

$$\hat{\sigma} \to \sqrt{\frac{a}{b^2}}$$

in probability under the true Gamma model. We also know that if we specified the Gamma model correctly,

$$\sqrt{\frac{\hat{a}}{\hat{b}^2}} \to \sqrt{\frac{a}{b^2}}$$

in probability. Therefore, we conclude that the standard errors we use under model misspecification or correct specification converge to the same limit as sample size increases. Therefore, the effect of model misspecification is negligible when sample size is large.

Exercise 4

(a) For researcher A,

$$L_n^A(\theta) = \prod_{i=1}^{n} f(X_i \mid \theta)$$

For researcher B, using change of variable, the pdf for Y_i is

$$f(g^{-1}(Y_i) \mid \theta) \cdot \left|(g^{-1})'(Y_i)\right|$$

Therefore, the likelihood is

$$L_n^B(\theta) = \prod_{i=1}^{n} f(g^{-1}(Y_i) \mid \theta) \cdot \left|(g^{-1})'(Y_i)\right|$$

(b) Since $\prod_{i=1}^{n} |(g^{-1})'(Y_i)|$ is a constant with respect to θ and $g^{-1}(Y_i) = X_i$, we know that the θ maximizes $L_n^A(\theta)$ must also maximizes $L_n^B(\theta)$. Therefore, the MLE under two likelihoods are the same. We denote the MLE as $\hat{\theta}$.

According to the asymptotic theory of MLE, we have

$$\sqrt{n}(\hat{\theta}-\theta) \xrightarrow{\text{d}} \mathsf{N}(0, I_A(\theta)^{-1})$$

where

$$
\begin{aligned}
I_A(\theta) &= -\mathsf{E}_{X\sim f(\cdot\mid\theta)}\left[\frac{\partial^2}{\partial\theta^2}\log f(x\mid\theta)\right] \\
&= -\int\left(\frac{\partial^2}{\partial\theta^2}\log f(x\mid\theta)\right)f(x\mid\theta)\mathrm{d}x \\
&= -\int\left(\frac{\partial^2}{\partial\theta^2}\log f(g^{-1}(y)\mid\theta)\right)f(g^{-1}(y)\mid\theta)(g^{-1})'(y)\mathrm{d}y \\
&= -\int\left(\frac{\partial^2}{\partial\theta^2}\left\{\log f(g^{-1}(y)\mid\theta)+\log|(g^{-1})'(y)|\right\}\right)f(g^{-1}(y)\mid\theta)(g^{-1})'(y)\mathrm{d}y \\
&= -\mathsf{E}_{Y\sim f(g^{-1}(\cdot)|\theta)|(g^{-1}(y))'|}\left[\frac{\partial^2}{\partial\theta^2}\log\left\{f(g^{-1}(y)\mid\theta)|(g^{-1})'(y)|\right\}\right] \\
&= I_B(\theta)
\end{aligned}
$$

Therefore, both MLEs and their asymptotic variances are exactly same. It means that the inference of model parameter is invariant with respect to monotonic transformation of data.

Exercise 5

(a) Since $\mathsf{E}[\bar{Y}] = \theta/2$, we may use $\hat{\theta}_1 = 2\bar{Y}$ as an estimator of θ (method of moments). Its variance is

$$
\mathsf{Var}[\hat{\theta}_1] = \mathsf{Var}[2\bar{Y}] = \frac{\theta^2}{3n}
$$

(b) The likelihood is

$$
L_n(\theta) = \prod_{i=1}^{n}\frac{\mathbb{1}(0 < Y_i < \theta)}{\theta}
$$

It is maximized at $\hat{\theta}_2 = Y_{(n)} := \max\{Y_1, \ldots, Y_n\}$.

(c) For $0 < y < \theta$, the CDF of $\hat{\theta}_2$ is

$$
\begin{aligned}
F(y) = \mathsf{Pr}(Y_{(n)} \leqslant y) &= \mathsf{Pr}(\{Y_i \leqslant y \text{ for all } i = 1, \ldots, n\}) \\
&= \prod_{i=1}^{n}\frac{y}{\theta} = \frac{y^n}{\theta^n}
\end{aligned}
$$

Therefore, the CDF of $\hat{\theta}_2$ is

$$
F(y) = \begin{cases} 0 & y \leqslant 0 \\ y^n/\theta^n & 0 < y < \theta \\ 1 & y \geqslant \theta \end{cases}
$$

(d)

$$
\begin{aligned}
\mathsf{E}[\hat{\theta}_2] &= \int_0^\theta n y^n/\theta^n \mathrm{d}y = \frac{n}{n+1}\theta \\
\mathsf{Var}[\hat{\theta}_2] &= \mathsf{E}[\hat{\theta}_2^2] - \mathsf{E}[\hat{\theta}_2]^2 = \frac{n}{n+2}\theta^2 - \frac{n^2}{(n+1)^2}\theta^2 \\
&= \frac{n\theta^2}{(n+2)(n+1)^2}
\end{aligned}
$$

Therefore, $\hat{\theta}_2$ is a biased estimator while $\hat{\theta}_1$ is unbiased. For the variances, it suffices to compare $3n^2$ and $(n+2)(n+1)^2$. We note that $n+2 \geqslant 3$ and $(n+1)^2 > n^2$ for any $n \geqslant 1$. Therefore, $\hat{\theta}_2$ has a smaller variance than $\hat{\theta}_1$.

The MSE of $\hat{\theta}_2$ is

$$
\begin{aligned}
\mathsf{MSE}[\hat{\theta}_2] &= \frac{\theta^2}{(n+1)^2} + \frac{n\theta^2}{(n+2)(n+1)^2} \\
&= \frac{2\theta^2}{(n+2)(n+1)}
\end{aligned}
$$

The MSE of $\hat{\theta}_1$ is $\theta^2/(3n)$. When sample size is large, the MSE of $\hat{\theta}_2$ is much smaller than that of $\hat{\theta}_1$.

(e) The answer is no. In this case, the Fisher information is

$$
I(\theta) = \mathsf{E}[-1/\theta^2] = -\frac{1}{\theta^2}, \quad \theta > Y_{(n)}
$$

The information is negative and definitely cannot be used as an approximate variance of the MLE. The properties of MLE we discussed in the lectures break down here because this model does not satisfy the regular conditions. Particularly, the support of pdf ($f(y \mid \theta)$) is $(0, \theta)$ which depends on θ and the likelihood is not a differentiable (not even continuous at $Y_{(n)}$) function of θ on the whole positive real line.

(f) Since $\mathsf{E}[\hat{\theta}_2] = \frac{n}{n+1}\theta$, an unbiased estimator based on MLE would be $\hat{\theta}_3 = \frac{n+1}{n}Y_{(n)}$. The MSE of $\hat{\theta}_3$ is

$$
\mathsf{MSE}[\hat{\theta}_3] = \frac{(n+1)^2}{n^2}\mathsf{Var}[\hat{\theta}_2] = \frac{\theta^2}{n(n+2)}
$$

and we have

$$
\mathsf{MSE}[\hat{\theta}_3] \leqslant \mathsf{MSE}[\hat{\theta}_2]
$$

for all $n \geqslant 1$ with the equality holds only when $n = 1$. We note that $\hat{\theta}_3$ is the UMVUE (i.e., an unbiased estimator with smaller variance than any other unbiased estimator for all possible values of the parameter) of θ by Lehmann-Scheffé theorem.