STAT 512 hw 5

1. Let $Y_1, \dots, Y_n \overset{\text{ind}}{\sim} \text{Exponential}(\lambda)$ and consider estimating λ with $nY_{(1)}$.

   (a) Find the pdf of $Y_{(1)}$ and identify its distribution.

   > The cdf and pdf of the Exponential(λ) distribution are given by $F_Y(y) = 1 - e^{-y/\lambda}$ and $f_Y(y) = (1/\lambda)e^{-y/\lambda}$, so we have
   >
   > $$\begin{aligned} f_{Y_{(1)}}(y_1) &= n[1 - (1 - e^{-y/\lambda})]^{n-1}(1/\lambda)e^{-y/\lambda} \\ &= (n/\lambda)e^{-y(n-1)/\lambda}e^{-y/\lambda} \\ &= \frac{1}{\lambda/n}e^{-y/(\lambda/n)}, \end{aligned}$$
   >
   > so $Y_{(1)} \sim \text{Exponential}(\lambda/n)$.

   (b) Find $\mathbb{E}nY_{(1)}$ and $\text{Var}(nY_{(1)})$.

   > We have $\mathbb{E}nY_{(1)} = n(\lambda/n) = \lambda$ and $\text{Var}(nY_{(1)}) = n^2(\lambda/n)^2 = \lambda^2$.

   (c) Find $\text{MSE}(nY_{(1)})$ as an estimator of λ.

   > We have $\text{MSE}(nY_{(1)}) = \big(\text{Bias}(nY_{(1)})\big)^2 + \text{Var}(nY_{(1)}) = 0 + \lambda^2 = \lambda^2$.

   (d) Find $\text{MSE}\,\bar{Y}_n$ as an estimator of λ.

   > We have $\text{MSE}\,\bar{Y}_n = (\text{Bias}\,\bar{Y}_n)^2 + \text{Var}\,\bar{Y}_n = 0 + \lambda^2/n = \lambda^2/n$.

2. Let $X_1, \dots, X_5$ be a random sample from a distribution with pdf

   $$f_X(x) = \frac{1}{\delta}\mathbf{1}(\delta < x < 2\delta)$$

   for some $\delta > 0$.

   (a) Find the cdf F_X corresponding to the pdf f_X.

   > The cdf is given by
   >
   > $$F_X(x) = \begin{cases} 1, & x \geq 2\delta \\ \dfrac{x - \delta}{\delta}, & \delta \leq x < \delta \\ 0, & x < \delta \end{cases}$$

   (b) Find the pdf of $X_{(2)}$.

For $\delta < x < 2\delta$ We have

$$f_{X_{(2)}}(x) = \frac{5!}{(2-1)!(5-2)!}\left[\frac{x-\delta}{\delta}\right]^{2-1}\left[1-\frac{x-\delta}{\delta}\right]^{5-2}\frac{1}{\delta} = \frac{20}{\delta}\left[\frac{x-\delta}{\delta}\right]\left[1-\frac{x-\delta}{\delta}\right]^{3}.$$

(c) Find the pdf of the $Y = (X_{(2)} - \delta)/\delta$ of $X_{(2)}$ and give the name of the distribution of Y.

We have

$$y = (x-\delta)/\delta = g(x) \iff x = \delta y + \delta = g^{-1}(y) \text{ and } \frac{d}{dy}g^{-1}(y) = \delta.$$

Note that the support of Y is $(0,1)$. So the pdf of Y is given by

$$\begin{aligned} f_Y(y) &= \frac{20}{\delta}y(1-y)^3 \cdot |\delta| \\ &= \frac{\Gamma(2+4)}{\Gamma(2)\Gamma(4)}y^{2-1}(1-y)^{4-1} \text{ for } 0 < y < 1. \end{aligned}$$

So Y has the $\text{Beta}(2,4)$ distribution.

(d) Give $\mathbb{E}Y$ and $\operatorname{Var} Y$.

We have

$$\begin{aligned} \mathbb{E}Y &= \frac{2}{2+4} = 1/3 \\ \operatorname{Var} Y &= \frac{2(4)}{(2+4)^2(2+4+1)} = 2/63. \end{aligned}$$

(e) Find the MSE of $\hat{\delta} = (3/4)X_{(2)}$ when $\hat{\delta}$ is used as an estimator of δ. *Hint:* $X_{(2)} = \delta Y + \delta$.

We have

$$\operatorname{Bias}\hat{\delta} = \mathbb{E}(3/4)X_{(2)} - \delta = (3/4)(\delta\mathbb{E}Y + \delta) - \delta = (3/4)(4/3)\delta - \delta = 0$$

and

$$\operatorname{Var}\hat{\delta} = \operatorname{Var}(3/4)X_{(2)} = (9/16)\operatorname{Var}[\delta Y + \delta] = (9/16)\delta^2 2/63 = \delta^2/56.$$

So

$$\operatorname{MSE}\hat{\delta} = \frac{\delta^2}{56}.$$

3. Let $Y_1, \ldots, Y_n$ be a random sample from the distribution with cdf given by

$$F_Y(y) = \begin{cases} 0, & y < 0 \\ (y/a)^b, & 0 \le y \le a \\ 1, & y > a \end{cases}$$

for some $a, b > 0$, where b is known. Consider the estimator of a given by $\hat{a} = Y_{(n)}$.

(a) Find the pdf of $Y_{(n)}$.

The population pdf is given by

$$f_Y(y) = \frac{b}{a^b} y^{b-1} \quad \text{for } 0 < y < a$$

and the pdf of $Y_{(n)}$ is given by

$$f_{Y_{(n)}}(y) = n \left[\left(\frac{y}{a} \right)^b \right]^{n-1} \frac{b}{a^b} y^{b-1} = \frac{nb}{a^{nb}} y^{nb-1} \quad \text{for } 0 < y < a.$$

(b) Find an expression for $\operatorname{Bias} \hat{a}$

We have

$$\mathbb{E}\hat{a} = \mathbb{E}Y_{(n)} = \int_0^a y \frac{nb}{a^{nb}} y^{nb-1} dy = a \left(\frac{nb}{nb+1} \right),$$

so that

$$\operatorname{Bias} \hat{a} = \mathbb{E}\hat{a} - a = a \left(\frac{nb}{nb+1} \right) - a = a \left[\left(\frac{nb}{nb+1} \right) - 1 \right] = -a \left(\frac{1}{nb+1} \right).$$

(c) Propose a scaled version of $\hat{a}$ which results in an unbiased estimator, $\tilde{a}$, of a.

If we define

$$\tilde{a} = \hat{a} \left(\frac{nb+1}{nb} \right),$$

then

$$\mathbb{E}\tilde{a} = \mathbb{E}\hat{a} \left(\frac{nb+1}{nb} \right) = a \left(\frac{nb}{nb+1} \right) \left(\frac{nb+1}{nb} \right) = a.$$

(d) Find the MSE of $\tilde{a}$.

Since $\tilde{a}$ is unbiased,

$$\operatorname{MSE} \tilde{a} = \operatorname{Var} \tilde{a} = \operatorname{Var} \left[\hat{a} \left(\frac{nb+1}{nb} \right) \right] = \left(\frac{nb+1}{nb} \right)^2 \operatorname{Var} \hat{a},$$

where $\operatorname{Var}\hat{a} = \operatorname{Var} Y_{(n)} = \mathbb{E}Y_{(n)} - (\mathbb{E}Y_n)^2$. We have

$$\mathbb{E}Y_{(n)}^2 = \int_0^a y^2 \frac{nb}{a^{nb}} y^{nb-1} dy = a^2\left(\frac{nb}{nb+2}\right),$$

so that

$$\operatorname{Var} Y_{(n)} = a^2\left(\frac{nb}{nb+2}\right) - \left[a\left(\frac{nb}{nb+1}\right)\right]^2 = a^2\left[\frac{nb}{nb+2} - \left(\frac{nb}{nb+1}\right)^2\right].$$

Finally we have

$$\begin{aligned}\operatorname{MSE}\tilde{a} &= \left(\frac{nb+1}{nb}\right)^2 a^2\left[\frac{nb}{nb+2} - \left(\frac{nb}{nb+1}\right)^2\right]\\ &= a^2\left[\frac{(nb+1)^2}{(nb+2)nb} - 1\right]\\ &= \frac{a^2}{(nb+2)nb}\end{aligned}$$

(e) Find the transformation of a $\text{Uniform}(0,1)$ rv which will result in a realization of Y.

Setting $U = (Y/a)^b$, which has the $\text{Uniform}(0,1)$ distribution by the probability integral transform, and solving for Y gives

$$Y = aU^{1/b},$$

which can be used to generate realizations of the random variable Y.

(f) Run a simulation using R to confirm the formula you obtained for $\operatorname{MSE}\tilde{a}$. Specifically, choose values of a, b, and n and generate $1{,}000$ samples of size n (I recommend choosing $b \leq 5$). On each sample, compute the estimator $\tilde{a}$ and record its value. Then compute the average squared distance of your $\tilde{a}$ values from a over the $1{,}000$ simulated data sets. In addition, compute the value of $\operatorname{MSE}\tilde{a}$ according to your formula from part (d). The numbers should be quite close to each other. You may make use of the following partial code:

```
a.tilde <- numeric()
for(s in 1:S)
{
  U <- runif(n)
  Y <- # your formula for generating Y from U
  a.tilde[s] <- # compute a.tilde on the sample
}

mean( (a.tilde - a)^2 )
# compute also MSE of a.tilde according to your formula
```

Here is what to turn in:

i. Your code.

ii. Your simulated value of $\text{MSE}\,\tilde{a}$ as well as its value according to the formula.

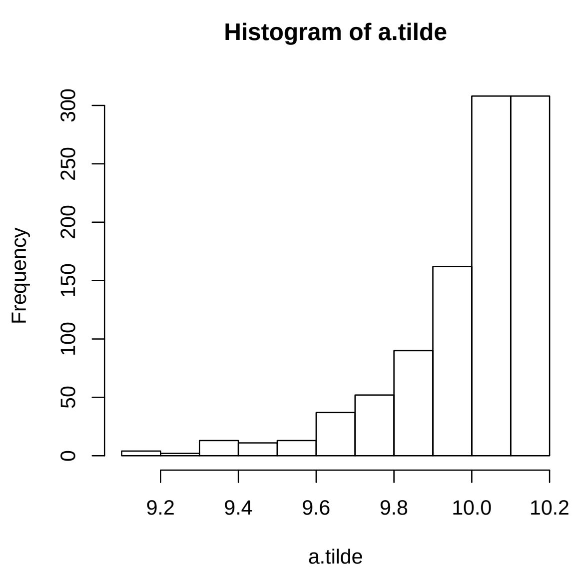
iii. A histogram of your `a.tilde` values.

```
n <- 20
a <- 10
b <- 3
S <- 1000

a.tilde <- numeric()
for(s in 1:S)
{
  U <- runif(n)
  Y <- a*U^(1/b)
  a.tilde[s] <- max(Y) * ((n*b + 1)/(n*b))
}

# compare:
mean( (a.tilde - a)^2 )
a^2/( (n*b + 2)*n*b)
```

Under these settings the simulated value of $\text{MSE}\,\tilde{a}$ was 0.02404551, and the theoretical value is

$$\text{MSE}\,\tilde{a} = \frac{10^2}{(20(3)+2)20(3)} = 0.02688172,$$

so the simulation results make sense. Below is a histogram of the $1{,}000$ values of `a.tilde` from the simulation:



4. Let $X_1, \ldots, X_n \overset{\text{ind}}{\sim} \text{Bernoulli}(p)$ and let $Y = X_1 + \cdots + X_n$. Consider the two estimators of p given by

$$\hat{p} = \frac{Y}{n} \quad \text{and} \quad \tilde{p} = \frac{Y+1}{n+2}.$$

(a) Find $\text{Bias}\,\hat{p}$ and $\text{Bias}\,\tilde{p}$.

We have $\text{Bias}\,\hat{p} = 0$ since the sample mean is always unbiased for the population mean. For $\tilde{p}$ we have

$$\text{Bias}\,\tilde{p} = \frac{\mathbb{E}Y + 1}{n+2} = \frac{np+1}{n+2} - p = \frac{1-2p}{n+2}.$$

(b) Find $\text{Var}\,\hat{p}$ and $\text{Var}\,\tilde{p}$.

We have $\text{Var}\,\hat{p} = p(1-p)/n$, since the sample means is always the population variance divided by the sample size. For $\tilde{p}$ we have

$$\text{Var}\,\tilde{p} = \frac{\text{Var}\,Y}{(n+2)^2} = \frac{np(1-p)}{(n+2)^2} = \left(\frac{n}{n+2}\right)^2 \frac{p(1-p)}{n}.$$

(c) Find $\text{MSE}\,\hat{p}$ and $\text{MSE}\,\tilde{p}$.

Since $\hat{p}$ is unbiased, $\text{MSE}\,\hat{p} = \text{Var}\,\hat{p} = p(1-p)/n$. For $\tilde{p}$ we have

$$\begin{aligned}\text{MSE}\,\tilde{p} &= (\text{Bias}\,\tilde{p})^2 + \text{Var}\,\tilde{p} \\ &= \left(\frac{1-2p}{n+2}\right)^2 + \left(\frac{n}{n+2}\right)^2 \frac{p(1-p)}{n}\end{aligned}$$

(d) If the true value of p is 0.50, which estimator has a lower MSE?

We see that at $p = 0.50$, the bias of $\tilde{p}$ is equal to zero. Since this estimator has a lower variance, that is, since

$$\text{Var}\,\tilde{p} = \left(\frac{n}{n+2}\right)^2 \text{Var}\,\hat{p},$$

we will have $\text{MSE}\,\tilde{p} < \text{MSE}\,\hat{p}$ when $p = 0.50$.

(e) If the true value of p is 0.95, which estimator has a lower MSE?

Plugging $p = 0.95$ into our formulas for $\text{MSE}\,\tilde{p}$ and $\text{MSE}\,\hat{p}$, we find that $\text{MSE}\,\hat{p} < \text{MSE}\,\tilde{p}$ for all $n \geq 1$.

5. Let $X_1, \dots, X_n \overset{\text{ind}}{\sim} \text{Poisson}(\lambda)$. Find a function of $\bar{X}_n$ which is an unbiased estimator of λ^2.
Hint: Begin by finding $\mathbb{E}\bar{X}_n^2$.

We have $\mathbb{E}\bar{X}_n^2 = \operatorname{Var}\bar{X}_n + (\mathbb{E}\bar{X}_n)^2 = \lambda/n + \lambda^2$. We see that the estimator $\tilde{\lambda} = \bar{X}_n^2 - \bar{X}_n/n$ satisfies

$$\mathbb{E}\tilde{\lambda} = \mathbb{E}[\bar{X}_n^2 - \bar{X}_n/n] = \lambda/n + \lambda^2 - \lambda/n = \lambda^2,$$

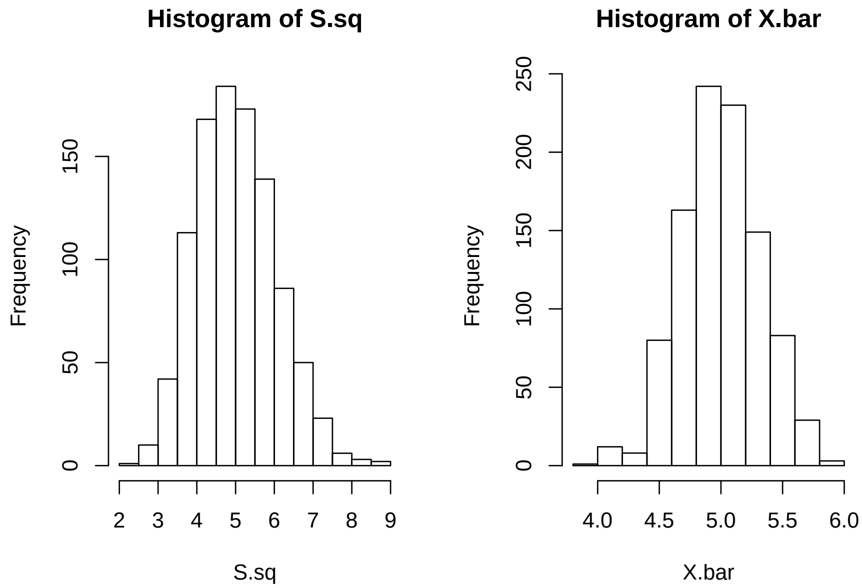
so that it is an unbiased estimator of λ^2.

6. Suppose $X_1, \dots, X_n$ are a random sample from the $\text{Poisson}(\lambda)$ distribution, where λ is unknown.
   (a) Find $\mathbb{E}\bar{X}_n$
   (b) Find $\mathbb{E}S_n^2$.
   (c) Which do you suggest as an estimator for λ? Run a simulation to inform your suggestion: Choose a sample size n and a value of λ and generate $1{,}000$ random samples of size n, computing on each random sample the value of $\bar{X}_n$ and $\bar{S}_n^2$ and storing these. You can do this with a for loop like the following:

```
X.bar <- S.sq <- numeric(S) # S is the number of data sets to simulate
for(s in 1:S)
{
    X <- rpois(n,lambda)
    X.bar[s] <- mean(X)
    S.sq[s] <- var(X)
}
```

   Make histograms of the $1{,}000$ values of $\bar{X}_n$ and $\bar{S}_n^2$ from your simulation and use these to argue for using one or the other as an estimator for λ. Turn in your code and the two histograms.
   Hint: Use `rpois` *to generate the Poisson data.*

These are the histograms resulting the simulation with $n = 50$ and $\lambda = 5$.



We can see that although both histograms are centered at the value of λ, the values of S_n^2 are much more spread out than the values of $\bar{X}_n$, so we would prefer $\bar{X}_n$ as an estimator of λ.