Title - Thomas Boyko - 30191728

1. Let $Y1, Y2, \ldots, Yn$ denote a random sample of size N from a population whose density is given by

$$f_Y(y) = \begin{cases} 3\beta^3 y^{-4} & \beta \leq y, \text{ where } \beta > 0 \text{ is unknown} \\ 0 & \text{otherwise} \end{cases}$$

(a) Derive the bias of the estimator $Y_{(1)} = \hat{\beta}$.
Before we do anything, we find the cdf for Y:

$$\begin{aligned} F_Y(y) &= \int_\beta^y 3\beta^3 y^{-4}\, dy \\ &= -\beta^3 y^{-3}\Big|_\beta^y \\ &= -\left(\frac{\beta}{y}\right)^3 - (-\beta^3\beta^{-3}) \\ &= -\left(\frac{\beta}{y}\right)^3 + 1 \end{aligned}$$

So $F_Y(y) = -\beta^3 y^{-3} + 1$.
Now we can calculate $f_{Y_{(1)}}(y)$.

$$\begin{aligned} f_{Y_{(1)}}(y) &= n\left[1 - F_Y(y)\right]^{n-1} f_Y(y) \\ &= n\left[1 - \left(-\left(\frac{\beta}{y}\right)^3 + 1\right)\right]^{n-1} \frac{3\beta^3}{y^4} \\ &= n\left[\left(\frac{\beta}{y}\right)^3\right]^{n-1} \frac{3\beta^3}{y^4} \\ &= n\left(\frac{\beta}{y}\right)^{3n-3} \frac{3\beta^3}{y^4} \\ &= \frac{3n}{y}\frac{\beta^{3n}}{y^{3n}}. \end{aligned}$$

With support (β, ∞).
We need to find $E[Y_{(1)}]$.

$$\begin{aligned} E[Y_{(1)}] &= \int_\beta^\infty y\frac{3n\beta^{3n}}{y^{3n+1}}\, dy \\ &= \int_\beta^\infty 3n\beta^{3n} y^{-3n}\, dy \\ &= \frac{3n}{1-3n} y^{-3n+1}\beta^{3n}\Big|_\beta^\infty \\ &= \frac{3n}{3n-1}\beta. \end{aligned}$$

Now we find $B(\hat{\beta})$.

$$\begin{aligned} B(\hat{\beta}) &= E[\hat{\beta}] - \beta \\ &= \frac{3n}{3n-1}\beta - \beta \\ &= \frac{\beta}{3n-1}. \end{aligned}$$

(b) Derive $MSE(\hat{\beta})$.

We begin by finding $E[\hat{\beta}^2]$ which we will need to find $V[\hat{\beta}]$.

$$\begin{aligned}
E[\hat{\beta}^2] &= \int_{\beta}^{\infty} y^2 \frac{3n\beta^{3n}}{y^{3n+1}}\, dy \\
&= \int_{\beta}^{\infty} 3n\beta^{3n} y^{-3n+1}\, dy \\
&= \frac{3n}{2-3n}\beta^{3n} y^{2-3n}\Big|_{\beta}^{\infty} \\
&= 0 - \frac{3n}{2-3n}\beta^{3n}\beta^{2-3n} \\
&= \frac{3n}{3n-2}\beta^2.
\end{aligned}$$

And now we can find $MSE[\hat{\beta}]$.

$$\begin{aligned}
MSE(\hat{\beta}) &= B[\hat{\beta}]^2 + V[\hat{\beta}] \\
&= \left(\frac{\beta}{3n-1}\right)^2 + \left(\frac{3n\beta^2}{3n-2}\right) - \left(\frac{3n\beta}{3n-1}\right)^2 \\
&= \frac{\beta^2 - 9n^2\beta^2}{(3n-1)^2} + \frac{3n\beta^2}{3n-2} \\
&= \beta^2\left(\frac{1-9n^2}{(3n-1)^2} + \frac{3n}{3n-2}\right) \\
&= \beta^2\left(\frac{(3n+1)(3n-1)}{(3n-1)^2} + \frac{3n}{3n-2}\right) \\
&= \beta^2\left(\frac{3n+1}{3n-1} + \frac{3n}{3n-2}\right).
\end{aligned}$$

2. Q2: TODO

(a) We begin by finding the CDF for X:

$$F_X(x) = \int_0^{\theta} \frac{\alpha x^{(\alpha-1)}}{\theta^{\alpha}} dx = \frac{x^{\alpha}}{\theta^{\alpha}}$$

Then we use our typical order statistic formula.

$$f_{X_{(n)}}(x) = n\left(\frac{x^{(\alpha n-\alpha)}}{\theta^{(\alpha n-\alpha)}}\right)\left(\frac{\alpha x^{(\alpha-1)}}{\theta^{\alpha}}\right) = \frac{n\alpha x^{(\alpha n-1)}}{\theta^{\alpha n}} I(0<x<\theta) I(\alpha>0)$$

(b) Bias of $X_{(n)}$ is given by:

$$B(X_{(n)}) = \int_0^{\theta} \frac{n\alpha x^{(\alpha n)}}{\theta^{\alpha n}} dx - \theta = \frac{\alpha n\theta}{\alpha n+1} - \theta = \frac{-\theta}{\alpha n+1}$$

Now to find an unbiased estimator for θ: Let $\frac{\alpha n\theta_{ub}}{\alpha n+1} = X_{(n)}$. Then solving for θ_{ub} we get:

$$\theta_{ub} = \frac{X_{(n)}(\alpha n+1)}{\alpha n}.$$

(c)

$$MSE(\theta_{ub}) = E(\theta_{ub}^2) - (E(\theta_{ub}))^2$$

Where:

$$E(\theta_{ub}^2) = (\frac{\alpha * n+1}{\alpha * n})^2 * \int_0^{\theta} \frac{x^2 * n * \alpha * x^{(\alpha * n-1)}}{\theta^{(\alpha * n)}} dx = \frac{\theta^2(\alpha * n+1)^2}{\alpha * n(\alpha * n+2)}$$

So we have

$$MSE(\theta_{ub}) = \frac{\theta^2(\alpha * n+1)^2 - \theta^2\alpha * n(\alpha * n+2)}{\alpha * n(\alpha * n+2)} + \frac{\theta^2}{(\alpha * n+1)^2}$$

3. When estimating a population proportion, we can use $\hat{p}_1 = \frac{x}{n}$ or $\hat{p}_2 = \frac{x+2}{n+4}$. The second version is called the Agresti sample proportion. It does not need to be shown, but $\hat{p}_1$ is an unbiased estimator and $\hat{p}_2$ is a biased estimator. Compare the MSE's in the distributions of these two estimators. For a given sample sizes $n = 1, 5, 10, 16, 100$, which estimator would you suggest 'best' estimates the value of the population proportion, p? Ensure you explain you answer.

   Since $\hat{p}_1$ is unbiased, its MSE is simply equal to the variance $Var(\hat{p}_1)$.

$$\begin{aligned} MSE(\hat{p}_1) = Var(\hat{p}_1) &= Var\left(\frac{x}{n}\right) \\ &= \frac{Var(x)}{n^2} \\ &= \frac{p(1-p)n}{n^2} \\ &= \frac{p(1-p)}{n}. \end{aligned}$$

   Since $\hat{p}_2$ is a biased estimator, we must find its bias before calculating MSE.

$$\begin{aligned} B(\hat{p}_2) &= E[\hat{p}] - p \\ &= \frac{E[x+2]}{n+4} - p \\ &= \frac{E[x]+2}{n+4} - p \\ &= \frac{np+2}{n+4} - p \\ &= \frac{np+2}{n+4} - \frac{p(n+4)}{n+4} \\ &= \frac{2-4p}{n+4}. \end{aligned}$$

   And now we can find MSE

$$\begin{aligned} MSE(\hat{p}_2) &= B(\hat{p}_2)^2 + V(\hat{p}_2) \\ &= \left(\frac{-4p+2}{n+4}\right)^2 + V\left(\frac{x+2}{n+4}\right) \\ &= \left(\frac{-4p+2}{n+4}\right)^2 + V(x)\frac{1}{(n+4)^2} \\ &= \left(\frac{(2-4p)^2}{(n+4)^2}\right) + \frac{np(1-p)}{(n+4)^2} \\ &= \frac{(2-4p)^2 + np(1-p)}{(n+4)^2}. \end{aligned}$$

   Since $\hat{p}_1$ is unbiased we can say it generally estimates the value of p better. However the second estimator $\hat{p}_1$ has consistently less variance than $\hat{p}_1$.

4.

5. Let $X_1, X_2, ..., X_n$ be a random sample from a population of values that has an unknown distribution, in addition to an unknown mean μ and an unknown variance σ^2. Find an unbiased estimator for the variance of $\bar{X}$.

   An unbiased estimator for the variance of $\bar{X}$ is $\frac{S^2}{n}$. Check:

$$B(\frac{S^2}{n}) = E\left[\frac{S^2}{n}\right] - V\left(\bar{X}\right) = \frac{E[S^2]}{n} - \frac{1}{n^2}V\left(\sum_{i=1}^{n} X_i\right) = \frac{\sigma^2}{n} - \frac{1}{n^2}n\sigma^2 = 0$$

6.

7. The .txt in this folder on D2L contains data that resulted from a random sample of $n = 70$ professional hockey players who have contracts with NHL teams. The values are their salary in millions of dollars. Import or Copy and paste the data into R, and answer the following questions.

   (a) Find the sample mean and sample standard deviation, $\bar{X}$ and S.

   We begin by inputting the data into a vector x in R, and use:

```
> mean=mean(x)
> mean
[1] 2.141214
> sd=sd(x)
> sd
[1] 1.911294
```

   So our sample mean is $\bar{X} = 2.131214$ and our sample standard deviation is $S = 1.911294$.

   (b) Find a 95% confidence interval for μ, the mean salary of an NHL player for the season 2013-2014.

   For this we can again use `.est(x,conf.level=.95)`
   in R, which gives us the interval $(1.685482, 2.596946)$.

   Interpret the meaning of this interval in the context of the data.

   We can say with 95% confidence that the true mean salary during this season is between 1.685 million and 2.597 million dollars.

   (c) Find a 95% confidence interval for σ, the standard deviation of the distribution of NHL player salaries for the 2013-2014 season.

   We use DescTools' `VarCI(x,conf.level=.95))` which gives us the interval for variance:

$$(2.685591, 5.259562)$$

   And taking the square root of both yields the 95% interval for σ:

$$(1.638777, 2.293373)$$

   So we can say with 95% confindence that the true standard deviation of NHL player salaries is between 1.639 million and 2.293 million.

8. A recent Angus Reid survey of $n = 1504$ Canadians was taken between November 25 to 28th, 2014. One of the questions posed in the survey was the following: Do you approve or disapprove of proposals to change the criminal code of Canada to allow physicians to assist with the suicide of their patients by prescribing lethal drugs? Here are the findings: Strongly Approve = 556, Moderately Approve = 632, Moderately/Strongly Disapprove = 271, Don't Know/No Opinion = 45 (as reported by Angus Reid.) Find a 95% confidence interval for the proportion of all Canadians who 'approve' of proposals to change the criminal code of Canada to allow physicians to assist with the suicide of their patients by prescribing lethal drugs. Interpret the meaning of this interval in the context of the data.

   Begin by writing our $\hat{p} = \frac{632+556}{1504} = 0.7898936$.

   Then we can build our 95% CI with the formula:

$$\hat{p} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

   And we can do this in R:

```
> phat=(632+556)/n
> c(phat+qnorm(.025)*sqrt(phat*(1-phat)/1054),
phat-qnorm(.025)*sqrt(phat*(1-phat)/1054))
[1] 0.7652995 0.8144878
```

This gives the interval for the true value of p:

$$(0.7652995, 0.8144878)$$

So we can say with 95% confidence that the true proportion of all Canadians who 'approve' of proposals to change the criminal code of Canada to allow physicians to assist with the suicide of their patients by prescribing lethal drugs is between 76.53% and 81.45%.

9. What is the normal body temperature for healthy humans? A random sample of 130 healthy human body temperatures provided by Allen Shoemaker yielded 98.25 degrees and standard deviation 0.73 degrees.

   (a) Give a 99% confidence interval for the average body temperature of healthy people.

   We can find this interval using

   ```
   > zsum.test(98.25,sigma.x =.73,n.x = 130,conf.level = 0.99 )
   ```

   Which gives us a 99 percent confidence interval: $98.08508, 98.41492$

   (b) Does the confidence interval obtained in part (a) contain the value 98.6 degrees, the accepted average temperature cited by physicians and others? What conclusions can you draw?

   The interval contains the accepted average, so we can say that there is no evidence to suggest that the true mean is not 98.6 degrees.

10. Use the data:

$$16, 5, 21, 19, 10, 5, 8, 2, 7, 2, 4, 9$$

to construct a 98% confidence interval estimate for the mean LC50 for DDT. In addition, ensure you (i) state any condition(s) that are necessary for your CI to be valid and (ii) interpret the meaning of your confidence interval, in the context of the data.

Since the size of our sample is quite small, we must suppose it is normally distributed, and since we do not know the true variance, we use `t.test()`

```
> x=c(16, 5, 21, 19, 10, 5, 8, 2, 7, 2, 4, 9)
> t.test(x,conf.level=.98)
```

Which gives our 98% CI:

$$(3.959158, 14.040842)$$

So we can say with 98% confidence that the true mean LC50 for DDT is between 3.96 and 14.04.

11. In a recent survey of $n = 1005$ Canadians between the ages of 18 and 34, the polling company Ipsos found that 723 indicated they "owe it to their parents to keep them comfortable in their retirement."

   (a) Find a 95% confidence interval for the proportion of all Canadians 18 to 34 years of age who hold the same sentiment.

   We begin with $\hat{p} = \frac{723}{1005} = 0.719403$, and $\alpha = .05$ We use our formula for CIs of qualitative data:

$$\hat{p} \pm Z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

   In R:

```
> phat=723/1005
> phat+qnorm(.025)*sqrt(phat*(1-phat)/1005)
[1] 0.6916255
> phat-qnorm(.025)*sqrt(phat*(1-phat)/1005)
[1] 0.7471805
```

   So we can say with 95% confidence that the true proportion of all Canadians 18 to 34 who share this sentiment is between 69.162% and 74.718%.

(b) Polls often come with a 'margin of error'. State what the margin of error is for this poll, and make a statement explaining what this margin of error means.
The population proportion's margin of error is the second piece of our R code:

```
> -qnorm(.025)*sqrt(phat*(1-phat)/1005)
[1] 0.02777747
```

Which tells us that our margin of error is about 2.8%, or the radius around our $\hat{p}$ which we can expect the true population proportion p to lie.

12. The profit of a new car sold by automobile dealer (in thousands of dollars) was recorded for 6 sales.

$$2.1, 3.0, 1.2, 6.2, 4.5, 5.1$$

For this question we assume that profit per car is normally distributed with mean μ and standard deviation σ.

(a) find a 95% confidence interval estimate for μ, the mean profit of a new car sold by automobile dealer.
So we use the R code:

```
> x=c(2.1, 3.0, 1.2, 6.2, 4.5, 5.1)
> xbar=mean(x)
> n=length(x)
> tsum.test(xbar,sd(x),n,conf.level=.95)
```

Which gives the interval for our mean:

$$1.683982, 5.682685$$

(b) a 95% confidence interval for σ, the standard deviation of the profit of a new car sold by an automobile dealer
We can simply use the R code:

```
> VarCI(x,conf.level=.95)
```

Which gives our interval for variance:

$$1.414247, 21.833590$$

Which we convert to standard deviation:

$$1.189221, 4.672643$$

13. The length of time between billing and receipt of payment was recorded for a sample of 100 of a certified public accountant (CPA) firm's clients. The sample mean and standard deviation for the 100 accounts were 39.1 days and 17.3 days, respectively. Find a 99% confidence interval for the mean time between billing and receipt of payment for all of the CPA firm's accounts. State any conditions/assumptions required, and interpret the meaning of your interval in the context of the data.

Since we have summary data and we are looking for the mean without knowing the true standard deviation, we can use `tsum.test(39.1,17.3,100,conf.level=.99)`. This gives us the confidence interval:

$$(34.55632, 43.64368)\,.$$

So we can say with 99% confidence that the true mean time between billing and reciept of payment is between 34.55 and 43.64 days.

14. $X_1, X_2, \ldots, X_n$ represent a random sample of Student ID numbers from University of Calgary students. Assume $Xi \sim Uniform(0, N)$ where N is the total number of U of C students. Use the pivotal quantity $\frac{X_{(n)}}{N}$ to find a 95% lower confidence bound for N . That is, use a sample of student ID's to estimate, with 95% confidence, that the largest Student ID will be at least how large?