Homework #6

See Canvas for the HW #6 assignment and due date. Complete all of the following problems. Ideally, the theoretical problems should be answered in a Markdown cell directly underneath the question. If you don't know LaTex/Markdown, you may submit separate handwritten solutions to the theoretical problems, but please see the class scanning policy. Please do not turn in messy work. Computational problems should be completed in this notebook (using the R kernel). Computational questions may require code, plots, analysis, interpretation, etc. Working in small groups is allowed, but it is important that you make an effort to master the material and hand in your own work.

A. Theoretical Problems

A.1 [10 points] Approximate Confidence Interval for Proportions

Recall from an earlier assignment that if $np > 5$ and $n(1-p) > 5$, then $X \sim Bin(n, p)$ is well-approximated by $Y \sim N\left(np, np(1-p)\right)$. Use this approximation in the question below. In particular, an approximate $(1-\alpha) \times 100\%$ confidence interval for a population proportion p is given by

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}},$$

where $\hat{p}$ is the sample proportion.

Suppose that 12 people in a sample of 95 are members of the Green Party. Calculate an approximate 90% confidence interval for the true proportion of Green Party members in the population. Interpret this interval.

$$\text{Given: } \hat{p} = \frac{12}{95}$$

The critical value for a 90% confidence interval can be obtained from a standard normal distribution or a t-distribution table. For a normal distribution, $z_{\alpha/2}$ is approximately 1.645.

The formula for the confidence interval is:

$$\text{Confidence Interval} = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Substituting the values:

$$\text{Confidence Interval} = \frac{12}{95} \pm 1.645\sqrt{\frac{\frac{12}{95}\left(1-\frac{12}{95}\right)}{95}}$$

In [ ]:
```
lower_bound <- (12/95) - 1.645 * sqrt(((12/95) * (1 - (12/95))) / 95)
upper_bound <- (12/95) + 1.645 * sqrt(((12/95) * (1 - (12/95))) / 95)

lower_bound
upper_bound
```

0.0702484222092599
0.182383156738109

Calculating the values in R gives the interval:

$$\text{Confidence Interval} \approx (0.070248, 0.182383)$$

Interpretation: We are 90% confident that the true proportion of Green Party members in the population is within the calculated interval $(0.070248, 0.182383)$. $$

A.2 [10 points] Speed of Light

In 1881 Michelson and Newcomb measured the time light took to travel a distance of 7400 meters. From a study of their experimental setup and a descriptive study of their 64 measurements, we conclude that the data can be assumed to be i.i.d. These measurements yield the following sample quantities in microseconds for the sample mean $\bar{x}$ and sample standard deviation s:

$$\bar{x} = 27.75, s = 5.08$$

Construct a 95% confidence interval for the time light takes to travel 7400 meters.

Given sample statistics: $\bar{x} = 27.75$, $s = 5.08$, $n = 64$

The formula for a 95% confidence interval for the population mean (μ) is given by:

$$\text{Confidence Interval} = \bar{x} \pm z_{\alpha/2}\left(\frac{s}{\sqrt{n}}\right)$$

Substituting the values:

$$\text{Confidence Interval} = 27.75 \pm 1.96 \left(\frac{5.08}{\sqrt{64}} \right)$$

In [ ]:
```r
lower_bound <- 27.75 - 1.96 * (5.08 / sqrt(64))
upper_bound <- 27.75 + 1.96 * (5.08 / sqrt(64))

lower_bound
upper_bound
```

26.5054
28.9946

$$\text{Confidence Interval} \approx (26.5054, 28.9946)$$

A.3 A Change on Confidence

A journal article reports that a sample of size $n = 5$ was used as a basis for calculating a 95% CI for the true average natural frequency (Hz) of delaminated beams of a certain type. The resulting interval was $(229.764, 233.504)$. You decide that a confidence level of 99% is more appropriate than the 95% level used.

A.3 [14 points] (a) What are the limits of the 99% interval?

$$\text{Given interval for a } 95\% \text{ confidence level: } (229.764, 233.504)$$

To find the limits of the 99% confidence interval, we use the formula for the confidence interval:

$$\text{Confidence Interval} = \bar{x} \pm t_{\alpha/2} \left(\frac{s}{\sqrt{n}} \right)$$

In this case, since the sample size is small ($n = 5$), we use the t-distribution. For a 99% confidence interval with 4 degrees of freedom (5-1), the critical value $t_{\alpha/2}$ is approximately 4.604.

Substituting the values into the formula:

$$\text{Confidence Interval} = \bar{x} \pm 4.604 \left(\frac{s}{\sqrt{n}} \right)$$

The 99% confidence interval is then:

$$\text{Confidence Interval} = \bar{x} \pm 4.604 \left(\frac{s}{\sqrt{5}} \right)$$

Given that the 95% confidence interval is $(229.764, 233.504)$, we can find $\bar{x}$ by taking the midpoint of the interval:

$$\bar{x} = \frac{229.764 + 233.504}{2} = 231.634$$

Now, use the formula to find s:

$$233.504 = 231.634 + 4.604 \left(\frac{s}{\sqrt{5}} \right)$$

Solving for s:

$$s = \frac{233.504 - 231.634}{4.604/\sqrt{5}}$$

In [ ]:
```
# Calculate s
s= (233.504 - 231.634) / (4.604 / sqrt(5))

s
```

0.908220486082671

Now, substitute this value along with s and n into the formula to find the limits of the 99% confidence interval.

In [ ]:
```
lower_bound <- 231.634 - 4.604 * (s / sqrt(5))
upper_bound <- 231.634 + 4.604 * (s / sqrt(5))

lower_bound
upper_bound
```

229.764

233.504

$$\text{Confidence Interval} \approx (229.764, 233.504)$$

A.4 MLEs

Suppose that $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$, where σ is known, and we are ultimately interested in an estimator for $\theta = \mu^2$.

A.4 (a) [12 points] First, find the MLE of μ. *(Hint: you may need to look back at the*

Unit 2 notes for the PDF of the normal distribution).

The likelihood function for the sample is the product of the individual PDFs:

$$L(\mu; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)$$

Taking the natural logarithm of the likelihood function simplifies the calculations:

$$\ln L(\mu) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

To find the maximum likelihood estimator (MLE) for μ, differentiate $\ln L(\mu)$ with respect to μ and set it equal to zero to find maximum:

$$\frac{d}{d\mu}\ln L(\mu) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i - \mu) = 0$$

Solving gives the MLE for μ, which is just the sample mean:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$

A.4 (b) [4 points] Find the maximum likelihood estimator (MLE) for θ, denoted $\hat{\theta}$.

This should be easy!

$$\text{Given: } \theta = \mu^2$$

The maximum likelihood estimator (MLE) for μ is:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Now, substitute the MLE for μ into the expression for θ:

$$\hat{\theta} = \left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)^2$$

A.4(c) [10 points] Compute the bias of $\hat{\theta}$, denoted $Bias(\hat{\theta})$. Recall that $Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$.

$$\text{Given: } \hat{\theta} = \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)^2, \quad \theta = \mu^2$$

Compute the expected value of $\hat{\theta}$:

$$\begin{aligned} E(\hat{\theta}) &= E\left[\left(\frac{1}{n} \sum_{i=1}^{n} x_i \right)^2 \right] \\ &= \left(\frac{1}{n} \right)^2 \sum_{i=1}^{n} E(x_i)^2 \\ &= \left(\frac{1}{n} \right)^2 \sum_{i=1}^{n} \mu^2 \\ &= \frac{1}{n} \mu^2 \end{aligned}$$

Now, find the bias:

$$\begin{aligned} Bias(\hat{\theta}) &= E(\hat{\theta}) - \theta \\ &= \frac{1}{n} \mu^2 - \mu^2 \end{aligned}$$

$$

B. Computational Problems

B.1 Hubble Data

Load `hubble.csv` into `R`. A description of the variables can be obtained from page 73 of https://cran.r-project.org/web/packages/gamair/gamair.pdf.

In [ ]:
```
hubble = read.csv("hubble.csv")

head(hubble)
summary(hubble)
```

A data.frame: 6 x 4

	X	Galaxy	y	x
	<int>	<chr>	<int>	<dbl>
1	1	NGC0300	133	2.00
2	2	NGC0925	664	9.16
3	3	NGC1326A	1794	16.14
4	4	NGC1365	1594	17.95
5	5	NGC1425	1473	21.88
6	6	NGC2403	278	3.22

```
       X           Galaxy                y               x        
 Min.   : 1.00   Length:24          Min.   :  80.0   Min.   : 2.00  
 1st Qu.: 6.75   Class :character   1st Qu.: 616.5   1st Qu.: 8.53  
 Median :12.50   Mode  :character   Median : 827.0   Median :13.08  
 Mean   :12.50                      Mean   : 924.4   Mean   :12.05  
 3rd Qu.:18.25                      3rd Qu.:1423.2   3rd Qu.:15.87  
 Max.   :24.00                      Max.   :1794.0   Max.   :21.98  
```

B.1 (a) [20 points] Calculate the 85% confidence interval for the mean of a galaxy's distance from Earth in megaparsecs in `R` by doing the computation explicitly.

In [ ]:
```r
distance = sqrt(hubble$y^2 + hubble$x^2)

n <- length(distance)
x_bar <- sum(distance) / n
s <- sqrt(sum((distance - x_bar)^2) / (n - 1))

confidence_level <- 0.85
df <- n - 1 #degrees of freedom

SE <- s / sqrt(n) #standard error
t_value <- qt((1 + confidence_level) / 2, df)
ME <- t_value * SE #margin of error

confidence_interval1 <- c(x_bar - ME, x_bar + ME)
confidence_interval1
```

768.561179335184 · 1080.36090261526

B.1 (b) [10 points] Find a built-in `R` function that does this computation automatically and verify that the built in `R` function does the same thing as the confidence interval formula used in part (a).

In [ ]:
```r
confidence_interval2 <- t.test(distance,conf.level = 0.85)$conf.int

confidence_interval2
```

768.561179335184 · 1080.36090261526

B.1(c) [10 points] Interpret the confidence interval.

Interpretation: We are 85% confident that the the distance a galaxy is from Earth is within the calculated interval $(768.561179335184, 1080.36090261526)$ megaparsecs
$$