STA238 - Winter 2021

Assignment 4

Group 192- [redacted] Wei-Han Wang, [redacted]

March 19, 2021

Part 1

Step 1 (Mathematical Justification)

$$L\left(\sigma^2\right) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\sum_{i=1}^n \frac{x_i^2}{2\sigma^2}}$$
$$= \left(2\pi\sigma^2\right)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2}$$

$$\ln L\left(\sigma^2\right) = \frac{n}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^n x_i^2$$
$$\frac{dl}{d\sigma^2} = -\frac{n}{2}\cdot\frac{1}{2\pi\sigma^2}\cdot 2\pi - \frac{1}{2}\left(\sigma^2\right)^{-2}(-1)\cdot\sum_{i=1}^n x_i^2$$

$$= -\frac{n}{2}\cdot\frac{1}{\sigma^2} + \frac{1}{2}\cdot\frac{1}{(\sigma^2)^2}\sum_{i=1}^n x_i^2 = 0$$

$$-\frac{n}{2}\cdot\sigma^2 + \frac{1}{2}\sum_{i=1}^n x_i^2 = 0$$

$$n\sigma^2 = \sum_{i=1}^n x_i^2$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n x_i^2}{n}$$

$$\frac{d^2l}{(d\sigma^2)^2} = -\frac{n}{2}(-1)\left(\sigma^2\right)^{-2} + \frac{1}{2}(-2)\left(\sigma^2\right)^{-3}\sum_{i=1}^n x_i^2$$

$$= \frac{n}{2(\sigma^2)^2} - \frac{\sum_{i=1}^n x_i^2}{(\sigma^2)^3}$$

$$\sigma^2 = \hat{\sigma}^2$$

$$= \frac{n}{2(\hat{\sigma}^2)^2} - \frac{\sum_{i=1}^n x_i^2}{(\hat{\sigma}^2)^3}$$

$$= \frac{n}{2}\cdot\left(\frac{n}{\sum_{i=1}^n x_i^2}\right)^2 - \sum_{i=1}^n x_i^2\cdot\left(\frac{n}{\sum x_i^2}\right)^3$$

$$= -\frac{n^3}{2\left(\sum_{i=1}^n x_i^2\right)^2}$$

$n^3 \geqslant 0 \quad x_i^2 > 0 \quad \sum_{i=1}^{n} x_i^2 \geqslant 0$

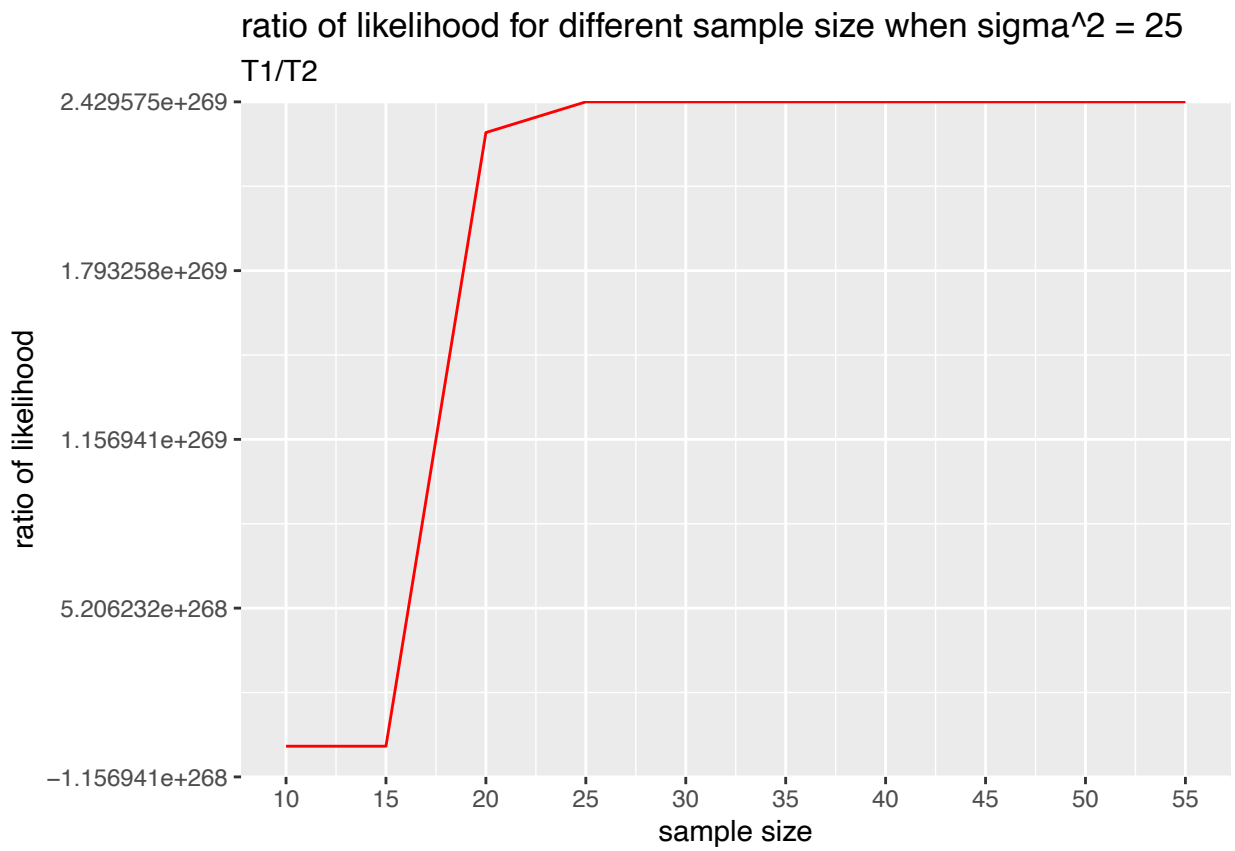
$\frac{-n^3}{2(\sum_{i=1}^{n} x_i^2)^2} \leqslant 0$

$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} x_i^2}{n}$

Therefore, $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} x_i^2}{n}$ is maximum likelihood estimate

Step 2 (Simulation Justification)

```
#population variance
var1<- 25

#pick 10 different sample size
n <- seq(10,55,5)

likelihood <- function(size,sum_x_sq, sigma_sq_hat){
  (2*pi*sigma_sq_hat)^(-size/2)*exp(-1/(2*sigma_sq_hat)*sum_x_sq)
}

ratio_of_likelihood <- rep(NA, length(n))

set.seed(587)

for(i in 1:length(n)){
  sample <- rnorm(n[i], mean=0, sd=sqrt(var1))
  T1 <- 1/(n[i]-1)*sum((sample - mean(sample))^2)
  T2 <- 1/(n[i]*sum((sample - mean(sample)))^2)

  Likelihood_T1 <- likelihood(size=n[i], sum_x_sq=sum(sample^2), sigma_sq_hat = T1)
  Likelihhod_T2 <- likelihood(size=n[i], sum_x_sq=sum(sample^2), sigma_sq_hat = T2)
  ratio_of_likelihood[i] <- Likelihood_T1/Likelihhod_T2



}

tibble(n, ratio_of_likelihood) %>%
  ggplot(aes(x=n)) +
  geom_line(aes(y = ratio_of_likelihood), color = "red") +
  labs(x = "sample size", y = "ratio of likelihood",
       title = "ratio of likelihood for different sample size when sigma^2 = 25",
       subtitle = "T1/T2",
       color = ""
       ) +
scale_x_continuous(breaks = seq(from = 10, to = 55, by = 5 ))
```

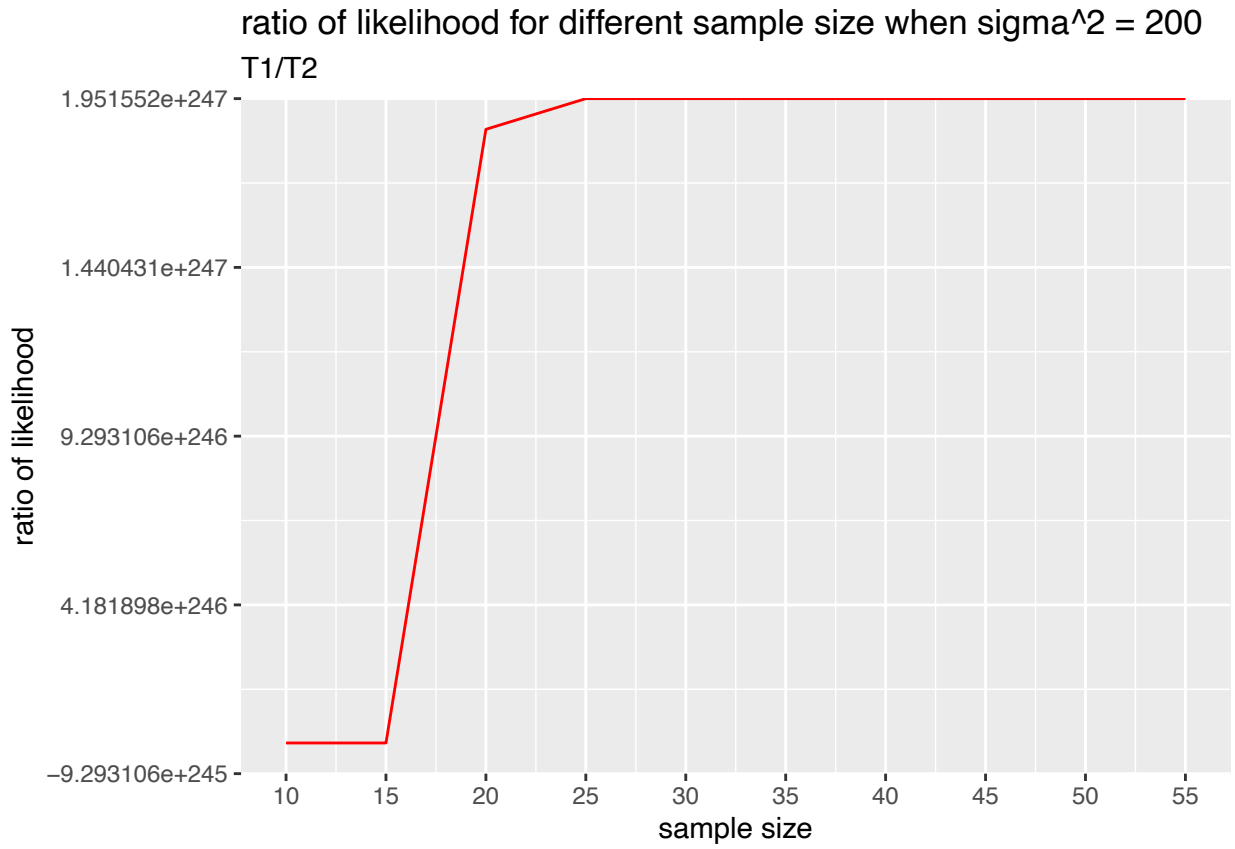


As we know, the bigger the sample size is, the closer the ratio of likelihood approach to 1.

The graph is increasing and ratio of likelihood is larger than one after sample size is equal to 20, therefore we can conclude that T1 over T2 is larger than 1 which mean T1 is larger than T2.

```r
#population variance
var2<- 200

#pick 10 different sample size
n <- seq(10,55,5)

likelihood <- function(size,sum_x_sq, sigma_sq_hat){
  (2*pi*sigma_sq_hat)^(-size/2)*exp(-1/(2*sigma_sq_hat)*sum_x_sq)
}

ratio_of_likelihood <- rep(NA, length(n))

set.seed(587)

for(i in 1:length(n)){
  sample <- rnorm(n[i], mean=0, sd=sqrt(var2))
  T1 <- 1/(n[i]-1)*sum((sample - mean(sample))^2)
  T2 <- 1/(n[i]*sum((sample - mean(sample)))^2)

  Likelihood_T1 <- likelihood(size=n[i], sum_x_sq=sum(sample^2), sigma_sq_hat = T1)
  Likelihhod_T2 <- likelihood(size=n[i], sum_x_sq=sum(sample^2), sigma_sq_hat = T2)
  ratio_of_likelihood[i] <- Likelihood_T1/Likelihhod_T2
```

```
}

tibble(n, ratio_of_likelihood) %>%
  ggplot(aes(x=n)) +
  geom_line(aes(y = ratio_of_likelihood), color = "red") +
  labs(x = "sample size", y = "ratio of likelihood",
       title = "ratio of likelihood for different sample size when sigma^2 = 200",
       subtitle = "T1/T2",
       color = ""
       ) +
scale_x_continuous(breaks = seq(from = 10, to = 55, by = 5 ))
```



The graph has the same pattern with the last graph, which is increasing after sample size equal to 15 and stay constant after sample size is 25. We can make a conclusion that ratio of likelihood is mostly larger than 1, then T1 is larger than T2. In general speaking, neither T1 nor T2 is maximum likelihood estimator. However, as long as we indicate that which log likelihood is higher for T1 and T2, it is cloer to maximum likelihood estimator.

Part 2

Introduction

The terms "deflation" and "inflation" often come across when people watch financial news but we most of the time don't have the concrete idea of how inflation and deflation really affect the economy of the country and the world. Between 1914 and 2021, the world has experienced two economical depression and now we are in the middle of a global pandemic which took a huge toll on the world's economy. In this following report, we will be observing the trend of deflation/inflation of consumer price index from January 1914 to January 2021 in Canada. Consumer Price Index is a number that measures the deflation/inflation of price consumers need to pay to purchase the same product by comparing the current year to the base year price. If the number of CPI increases, it indicates price inflation and vice versa. For the data we've chosen, 2002 is the base year as given in the dataset.

We predict that the average CPI of all items since 1914 will be between 40 to 50 given the boxplot of CPI values frequencies.

Given the spread of boxplot, we assume that the variance of CPI from 1914 to 2021 will be larger than the mean.

In the data section, we collected our data from Statistics Canada Portal and filtered out some data that we are not interested in for this simulation. In the methods section, we discuss the methodology of this simulation and its significance and meanings. In the results section, we present the results of the simulation and finally explain its importance regarding our topic of interest.

Data

This dataset contains the consumer price index, which will be mentioned data from January 1914 to January 2021 of each individual industry, as well as the average CPI of all industries combined. It also contains the data for all the provinces as well as the whole country of Canada, which is an average of all the provinces. The variables that the bootstrap will be performed on is the mean of the CPI and the variance of the CPI.

DATA CLEANING PROCESS:

Step 1- Install and the statcanR package and load the data:

The first step is to install the statcanR package using `devtools::install_github("warint/statcanR")`. Since we are pulling data from the Statistics Canada website, we have to load in the statcanR package. We then call the statcanR package by using `library(statcanR)`.

Step 2- Calling the data:

In this step we start the data cleaning process. We first call our data using `statcan_data("18-10-0004-01", "eng")`. We set the data we called to a new dataframe called mydata.

Step 3- Replacing spaces in column names with underscores:

Once we call the data and take a look at it, we realized that the column names had spaces in them, which would make it dificult to access in functions like `filter()`. This meant that we needed to replace the column names with something such as underscores and we did that using `names(mydata) %<>% stringr::str_replace_all("\\s","_") %>% tolower` .

Step 4- Filtering out all the provinces:

For step 4, we wanted our data to only include all the observations that pertained to the whole country of Canada and not the individual provinces as well. To acheive this we used the filter function where the arguments were the dataset and the condition respectively, and assigned it to `canada_data`.

Step 5- Filtering out data where 1992 is the base period:

In this step, we want all the data where 2002 is the base period, here we filter out all the data where 1992 is the base period to include only the observations in which 2002 is the base period using the `filter()` function, and we save it in `zero_two_data`.

Step 6- Filtering out the CPI for "All-items":

The sixth step is the final step in the data cleaning process. The dataset before we do this step includes the data for all products and groups, however, we just want the CPI for all the products and groups and not each individual product or group. Here, we take the dataset we got in the previous step and and filter the data to include only the "All-items" using the filter function as done in the previous two steps, only changing the used dataset and the condition appropriately, and we save the data in `final_data`.

Important Variables

In the final dataset, we have 16 variables, which is unchanged from the starting dataset. The important variables that we used and their descriptions are as follows:

ref_date: The date when the CPI was calculated and recorded.
geo: The location which the CPI was recorded.
products_and_product_groups: The different products or groups of products of which the CPI was calculated.
uom: This is the base period in which the CPI was calculated.
value: This is the CPI value.

Numerical Summaries Table

Average of the CPI	Standard Deviation of the CPI	Variance of the CPI
43.38569	43.45803	1888.60028

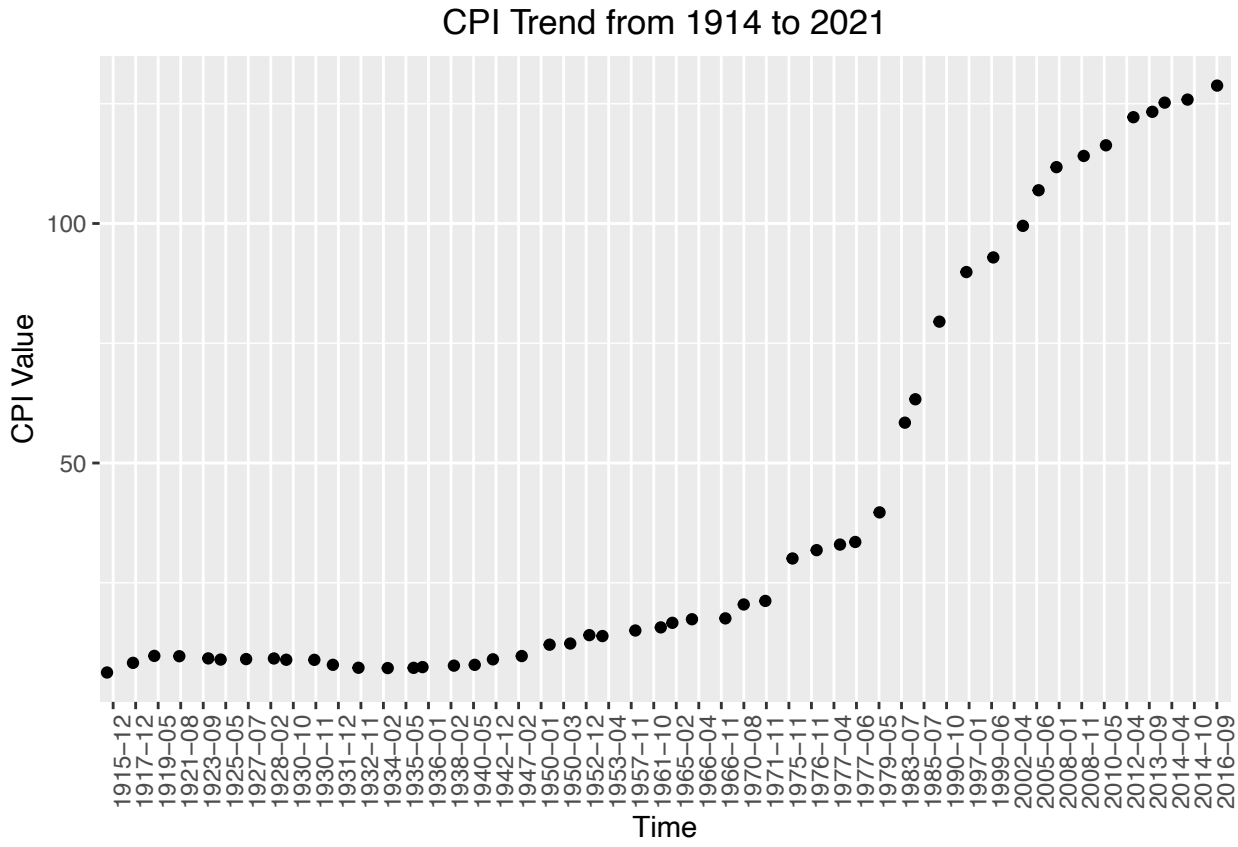
25th Percentile	50th Percentile	75th Percentile
9.2	18.2	85.8

The first numerical summary is the mean of the value variable, which gives us average consumer price index from January 1914 to January 2021. After we do the calculations, we see that the mean CPI was approximately 43.386.

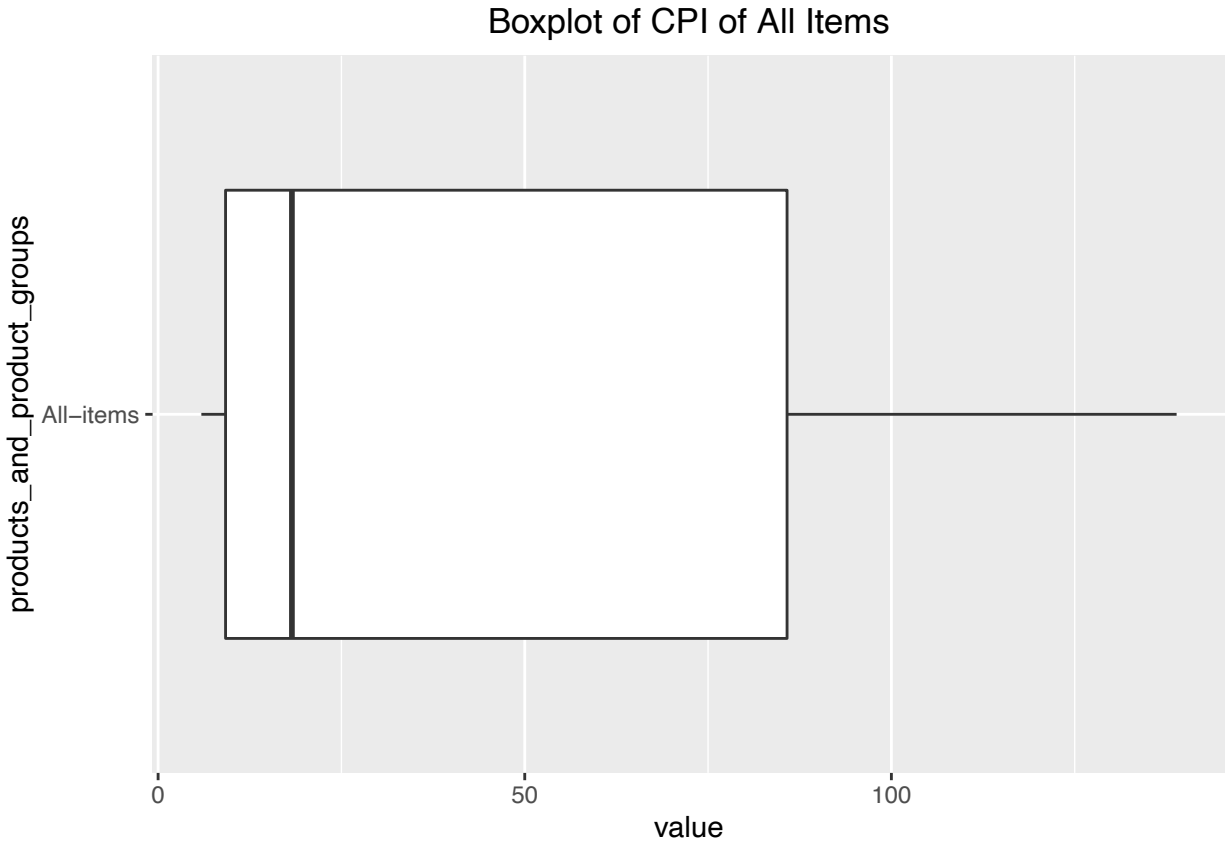
The second numerical summary is the standard deviation of the CPI. This tells us the spread of the graph. We can see that since the standard deviation is almost the same as the mean, which means that it is equally spread out. The standard deviation we calculated was 43.458

The last three numerical summaries that we calculated was the 25th, 50th and 75th percentile of the CPI values for all items in Canada from January 1914 to January 2021. We calculated the 25th percentile to be 9.2, the 50th percentile to be 18.2 and the 75th percentile to be 85.8.

Graphical Summaries



The first graphcal summary that we have is a scatterplot which shows the trend of the CPI for all items in Canada from January 1914 to January 2021. We had to get a sample of our population for this so that the graph can be clear and readable. To do this, we took a sample of 50 observations from our population of 1286 observations and we plotted the sample on a scatter plot. We can see that the trend in this scatterplot is that the CPI is generally increasing as the years go by.



The second graph that we made was a boxplot which where the x-axis was the value, in other words, the consumer price index, and the y-axis was the products and product groups. The bexplot shows the spread of all the CPI values from January 1914 to January 2021 for all items in Canada. We can see that the 50th percentile is around the 20 mark, which corresponds to the 50th percentile we found in our numerical summaries which was 18.2.

Methods

A confidence interval for the mean is how we estimate the true population mean of the data we are analyzing. To find the confidence interval, we will be using the Z distribution to obtain the confidence interval of the true population mean. We chose to use Z distribution instead of t because our sample size is 1286 which is not less than 30 and also we already know the sample variance of the population. The confidence interval for CPI mean is 95% and the confidence interval for CPI variance is 90%.

For the CI of the mean of the CPI, we calculated a 95% confidence interval. A 95% confidence level was the best option here to minimize the error and maximize the accuracy. For the CI of the variance of the CPI, the confidence level we calculated was a 90% confidence interval because if we do a higher confidence level, there is a chance of error since the graph is spread out quite a bit. To keep the chance of errors low for the variance, a 90% confidence level would be the best option.

The parameters of interest for the two confidence intervals simulation are mean and variance. We'll be using empirical bootstrap sampling for the second confidence interval for variance of CPI. Empirical bootstrap methods are used when we don't know the data distribution. We will be sampling the entire data with replacement for 1000 times. Then we take each iteration and calculate the variance.

All analysis for this report was programmed using `R version 4.0.2`. ## Results

Tables of Confidence Intervals

95-percent Confidence Interval of Mean via Critical Value

Lower Bound	Upper Bound
43.39237	47.37901

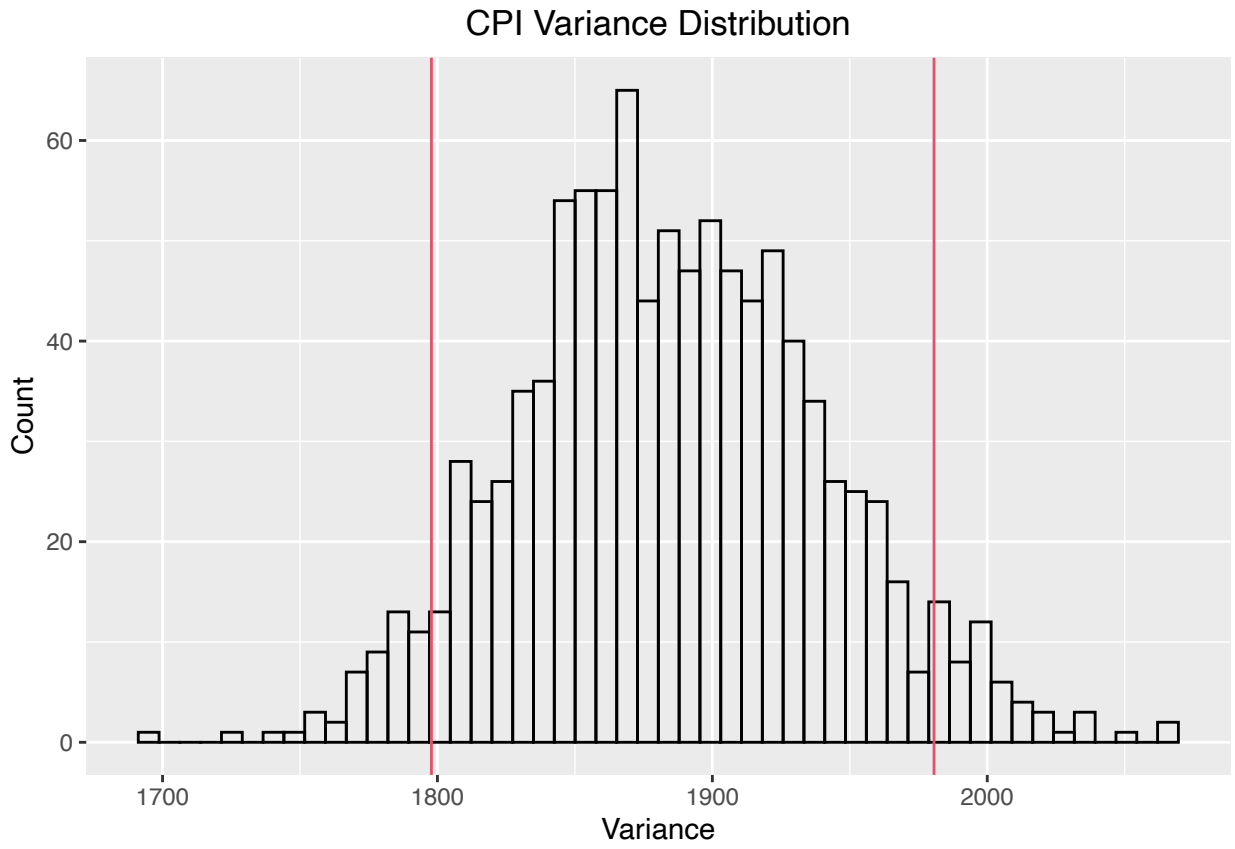
We are 95% confident that the mean of consumer price index from 1914 to 2021 would be between 43.01051 to 47.76088.

90-percent Confidence Interval of Variance via Bootstrap Sampling

5%	95%
1797.804	1980.635

We are 90$ confident that the variance of consumer price index from 1914 to 2021 would be between 1797.804 to 1980.635.

The result of our 95% confidence interval for mean is reasonable because we calculated the population mean of CPI is 45.28569, which is between the interval (43.39237,47.37901). The result of 90% confidence interval of variable is also reasonable because from previous calculation, the population variance is 1888.60028, which is between the interval (1797.804, 1980.635).



The consumer price index variance has a normal distributiion with a peak around 1870. The two red vertical lines indicate the 5% and 95% threshold that forms the 90% confidence for our variance distribution. In other words, the area between the two red lines are the 90% confidence interval.

Conclusions

For our hypotheses, we assumed that the mean of CPI between 1914 and 2021 would be between the value of 40 and 50 and that the variance of this data would be larger than its mean given the wide spread characteristic of the values. For methods, we used Z distribution to find critical values, which then helped us find our mean confidence interval with a 95% confidence level. For the second part of the method section, we used empirical bootstrapping to get variance of CPI.

Our result for the first confidence interval where we calculated the range for the mean of the CPI with a confidence level of 95% was between 43.0105 and 47.7608. This interval ended up being an accurate interval because as we calculated in the Data section, the mean of the CPI is 43.3956.

The second confidence interval that we calculated was for the variance of the CPI. This interval was calculated with a 90% confidence interval and it was between 1797.804 and 1980.635. This interval was also an accurate calculation as the variance was calculated to be 1888.6.

The large variance value suggests a large spread of data which then implies that the consumer price index has been either increasing or decreasing continuously by a lot between the years of 1914 and 2021. But given the knowledge of there is an inflation since 1914, we could assume that there has been a big price inflation since 1914 till today.

One drawback of this simulation we've had is we weren't able to do a comparison between different product groups and also we couldn't find a better way to plot all observations onto a scatterplot without overplotting. For future steps over this simulation, we could use different product groups to create plots so the variance wouldn't be as large. For future simulation ideas, we could do a comparison between the trend of CPI and some other economical characteristics that would support and explain the increase or decreasee of CPI over certain period of time.

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: March 19, 2021)

2. Dekking, F. M., et al. (2005) *A Modern Introduction to Probability and Statistics: Understanding why and how.* Springer Science & Business Media.(Last Accessed: March 19, 2021)

3. Allaire, J.J., et. el. *References: Introduction to R Markdown.* RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: March 19, 2021)

4. Makowski, Dominique (2018, August 31). *How to Cite Packages.* [https://www.r-bloggers.com/2018/08/how-to-cite-packages/] (https://www.r-bloggers.com/2018/08/how-to-cite-packages/). (Last Accessed: March 15, 2021)

5. H. Wickham. *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York, 2016.https://ggplot2.tidyverse.org (Last Accessed: March 19, 2021)

6. Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2021). *dplyr: A Grammar of Data Manipulation.* https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.

7. Thomas Lin Pedersen (2020). *patchwork: The Composer of Plots.* [https://patchwork.data-imaginist.com, https://github.com/thomasp85/patchwork.] (https://patchwork.data-imaginist.com, https://github.com/thomasp85/patchwork)

8. (2020, Nov 16) *"Data Visualization Cheatsheet."* RStudio. [http://https://github.com/rstudio/cheatsheets/raw/master/data-visualization-2.1.pdf] (https://github.com/rstudio/cheatsheets/raw/master/data-visualization-2.1.pdf)

9. (2021, Jan 17) *"Data Transformation Cheatsheet."* RStudio. [http://https://github.com/rstudio/cheatsheets/raw/master/data-transformation.pdf] (https://github.com/rstudio/cheatsheets/raw/master/data-transformation.pdf)

10. Wickham et al., (2019). *Welcome to the tidyverse. Journal of Open Source Software*, 4(43), 1686, https://doi.org/10.21105/joss.01686

11. Mine Çetinkaya-Rundel, David Diez, Andrew Bray, Albert Kim, Ben Baumer, Chester Ismay and Christopher Barr (2020). *openintro: Data Sets and Supplemental Functions from 'OpenIntro' Textbooks and Labs.* R package version 2.0.0. https://github.com/OpenIntroStat/openintro

12. B, Rob, and Tim McMahon. (2015, 18 June) *"Inflation and CPI Consumer Price Index 1930-1939."* InflationData.com. http://inflationdata.com/articles/inflation-cpi-consumer-price-index-1930-1939/(Last Accessed: March 17, 2021)

13. Thierry Warin and Romain Le Duc (2021). *statcanR: Client for Statistics Canada's Open Economic Data.* R package version 0.2.1.9000. https://github.com/warint/statcanR/(Last Accessed: March 19, 2021)