Analysis on effects of college wage premium

STA304 - Fall 2023 -Assignment 1

Ji Hoon Kim - 1005966849

September 28th 2023

Part 1: Designing a survey

Goal

The goal of this survey is to decide whether or not pursuing post-secondary education is worth it in Canada. While everything is subsidized by the government up to secondary school, the cost of pursuing a post-secondary degree is unimaginably expensive today let alone the insane amount of tuition. Some believe that the cost of post-secondary education is undoubtedly worth it in the long run while some believe they are better off investing them elsewhere. This survey will analyze the correlation between education level and current income, giving students a significant insight when it comes to deciding what their career path should look like after high school to have a bright future.

Procedure

The target population is all the adult Canadians under 30, since we are interested in the group of people who got out of their final level of education relatively recently. The frame population would be the list of adults under 30 (Canadians, Temporary/Permanent Residents) with a valid Canadian phone number. We can create a sample population by sending automated texts with a link to the survey, perhaps providing some compensation by giving them a raffle ticket that may be used to collect a prize. This is a non-probability based sampling technique, specifically volunteer-based sampling, which is time efficient and relatively inexpensive, but it may not represent the target population properly. However, by having the frame population as list of people with a valid cell, considering the fact that most people today have cellphones regardless of their education, income or field of occupation, and it won't be geographically biased, it will increase the likelihood of the sample representing the target population well.

Showcasing the survey.

https://www.surveymonkey.com/r/NJ86MFF

Question 1 Categorical

What is the highest level of education you have completed? (Categorical)

1. Less than highschool degree
2. Highschool degree or equivalent (e.g. GED)
3. College Certificate
4. Associate's Degree
5. Bachelor's Degree
6. Master's Degree
7. Doctorate's Degree

The biggest benefit of this question is that, this question will return the people's educational data which can be used to group them when performing statistical analysis which is crucial for claiming the validity of college premium. On the other hand, the drawback is that some may feel uncomfortable answering this question if they feel that their socioeconomic status is being judged by this factor, and they may potentially lie or even opt out of the survey completely.

Question 2 Numerical What was the grade at the end of your final level of education on a scale of 0-100%? Round to 2 decimal places. (e.g. 85.3459 is 85.35)

Coversion Formula: 100*(your grade in your system)/(maximum grade possible on your system)

This question is beneficial because it collects the grade data which is a key numerical data allowing us to calculate important statistics such as mean, median and variance. However, there is a drawback due to same reason as Q1 and because there is some calculation involved in the answer, despite being given the formula, some people might get fatigued by it. There also exists a small chance of calculation errors.

Question 3 Numerical What is your current annual salary of your primary occupation before tax in thousands of Canadian dollar? Indicate 0 if unemployed. Round to 1 decimal places.

This is the most important data in this analysis as our goal is to find a relationship between education and salary. Moreover, salary is a very important factor for anyone today. This data is critical when making any inferences about the population's salary data.

Part 2: Data Analysis

Data

The dataset contains 1000 samples and 6 variables, sampled by non-probability volunteer based sampling. I chose sample size of 1000 because, any sample size lower than that is too small to be representing the whole Canadians that fall in age group 18-30 but at the same time, it's unrealistic to sample any more. Each row represents one person, data about their gender, level of education, name of the institution that issued their degree, number of work experiences and finally, their grade and current salary.

gender, education, work_exp is a categorical variable, the surveyed users has given categories to choose from.

school, which is the column that represents the name of the school, is also a categorical variable because even though surveyed users can input their school name, it can still be grouped into colleges and university.

grade and salary is a continuous numerical variable.

gender, education, work_exp is a categorical variable, and it was simulated by sample() function which collects random value from a given category. education and work_exp was sampled via equal probability, while gender was sampled with 98% of weight on male and female due to be similar to the real world as much as possible.

Grade was sampled from a normal distribution with a mean of 75% and a standard deviation of 7% which correctly represents a typically class average.

Salary was sampled from a gamma distribution of alpha and beta of 5.2 and 14 respectively. And the reason I chose this specific distribution is because this distribution was the most appropriate in fitting the real Canadian Salary distribution. The Canadian median salary was 68.4K and the mean salary slightly above.

```r
survey_data <- data.frame(
  Gender = gender,
  Final_Degree = education,
  School_Name = school,
  Grade = grade, Salary = salary,
  Work_Experience = work_exp)

set.seed(20011005)
for (i in 1:nrow(survey_data)) {
  if (survey_data$Final_Degree[i] %in% c('Doctorate', 'Master', 'Bachelor')) {
    survey_data$School_Name[i] <- sample(c(
  'University of Toronto', 'University of Waterloo', 'University of British Columbia',
  'McGill University', 'University of Alberta', 'University of Montreal',
  'McMaster University', 'Western University', 'University of Calgary',
  'Queen\'s University', 'Simon Fraser University', 'Dalhousie University',
  'University of Ottawa', 'University of Saskatchewan', 'University of Victoria',
  'Laval University', 'Carleton University', 'University of Guelph',
  'York University', 'University of Manitoba', 'Ryerson University',
  'Wilfrid Laurier University'), 1, replace=TRUE,prob=NULL)
  } else if (survey_data$Final_Degree[i] %in% c('College Cert.', 'Associate')) {
    survey_data$School_Name[i] <- sample(c('George Brown College', 'Humber College',
  'Seneca College', 'Sheridan College', 'Centennial College', 'Niagara College',
  'British Columbia Institute of Technology', 'Red River College',
  'Douglas College', 'Vanier College'), 1, replace=TRUE,prob=NULL)
  } else {
    survey_data$School_Name[i] <- 'N/A'
  }
}
```

After sampling each question into a vector 1000 times, I had 5 vectors with a length of 1000, containing all the data for a specific column. Except the school vector, which was just a vector with a string 'School Name' as a dummy variable. I then created a data frame called survey_data which aggregates each vector into each column creating a dataset with 1000 observation and 6 variables. I also renamed the column to more readable words. Finally, using conditional statement and loops, I assigned, random school names to each row. The reason I did this afterwards is because, each school only issues certain degrees. For example, it does not make sense to have Doctorate's Degree from a college or a Associate's degree in a university.

Gender: Gender of a person. (Male, Female, or Non-Binary). Final_Degree: Highest level of degree accomplished by a person. School_Name: Name of the Institution that issued their Highest level of degree. Grade: Final grade that the person graduated with in percentage. Rounded to 2 deicmal places. Salary: Salary of their primary occupation before tax in thousands of Canadian Dollars. Rounded to 1 decimal places. Work_Experience: Number of years they have been working since graduation.

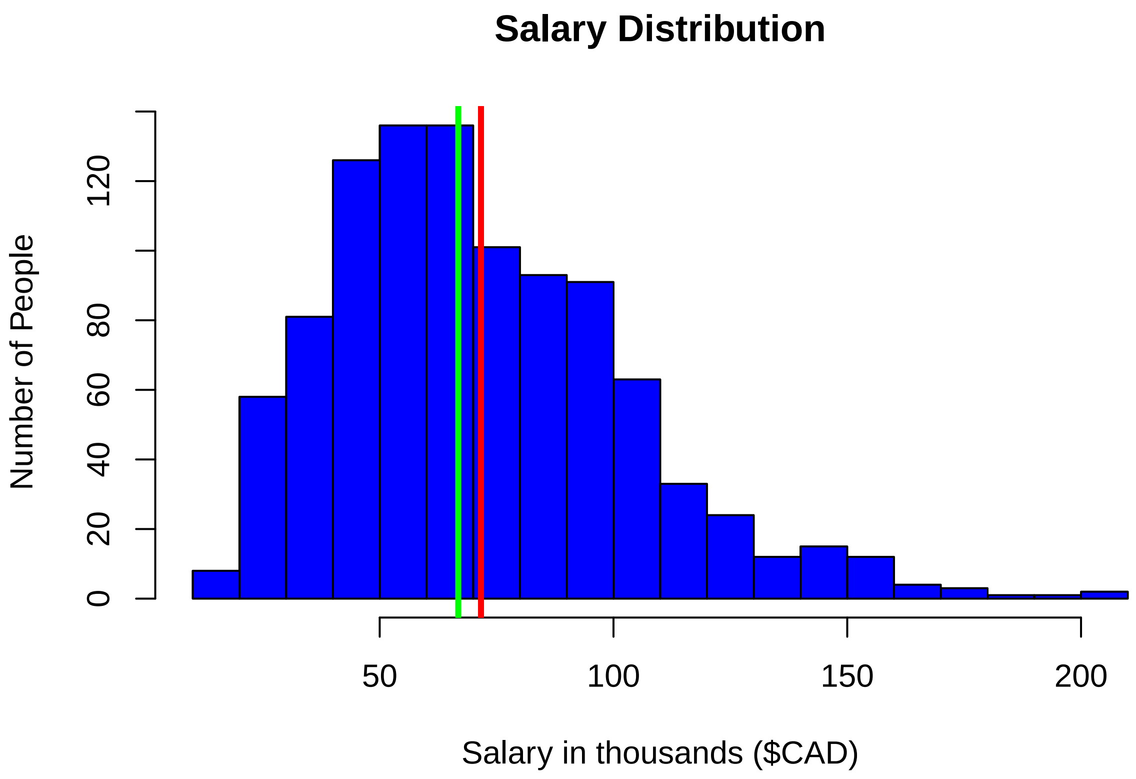
```
summary_table
```

Table 1: Mean, Median, Variance and Higher Order Moments of the Sample Population

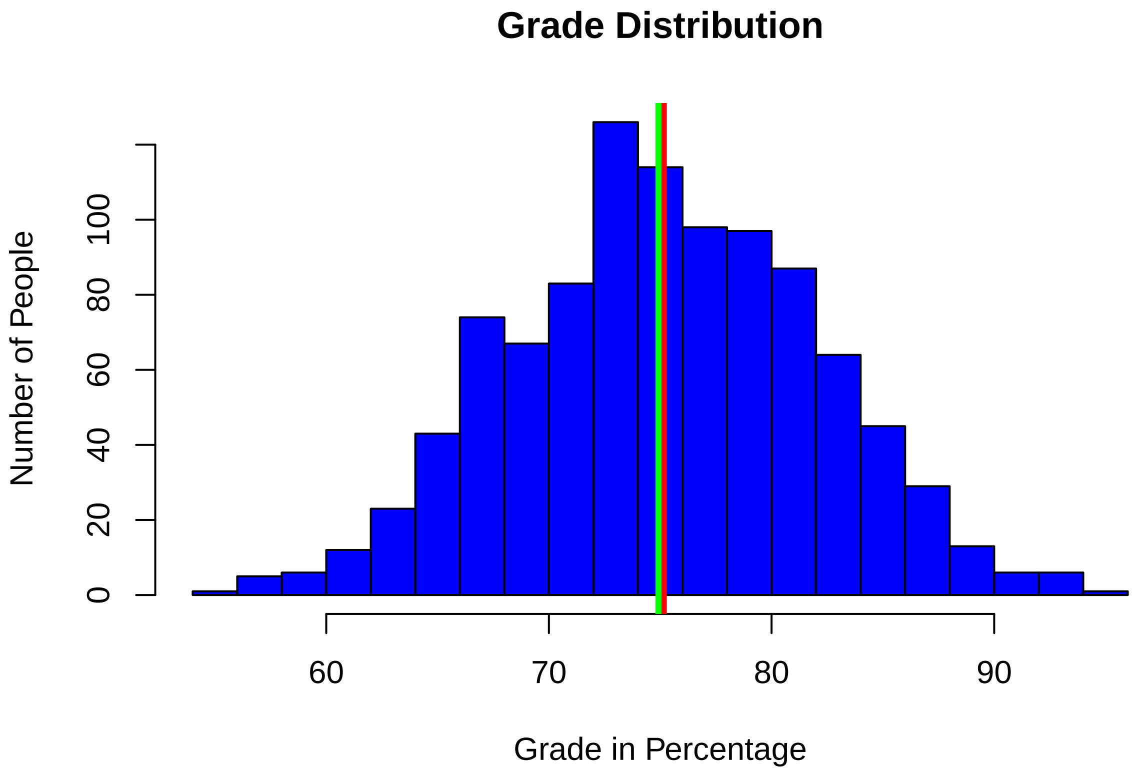
Mean Salary	Median Salary	Variance Salary	Skewness Salary	Kurtosis Salary	Mean Grade	Median Grade	Variance Grade	Skewness Grade	Kurtosis Grade
71.6647	66.8	999.4502	0.8368147	3.832492	75.1587	74.92	46.57486	-0.0077	2.780482

I calculated sample mean, sample median, and sample variance as well as higher order moments such as skewness and kurtosis of the sample population for their current salary and grades. The sample mean, mathematically represented as $\bar{y}$ is the sum of all sample values divided by the sample size (1,000 for our case). The sample variance is the measure of the spread of the data. The sample median is the value of the 50th percentile, in other others, it is the value of the data in the center when the data is sorted in ascending/descending order. In our case, it will be the sum of 500th and 501th value divided by 2. Skewness represents the skewness of the distribution of the sample, if this value is positive, more data is clustered on the left side and vice versa. Because Gamma(5.2,14) is a positively skewed distribution, the value is further from 0 compared to the distribution of the grade which is Normal and shouldn't have a big skew as shown in the values. Finally, the kurtosis represents how crowded the distribution is around the mean. In other words, higher the kurtosis, shorter the tail, meaning less values are far away from the mean.

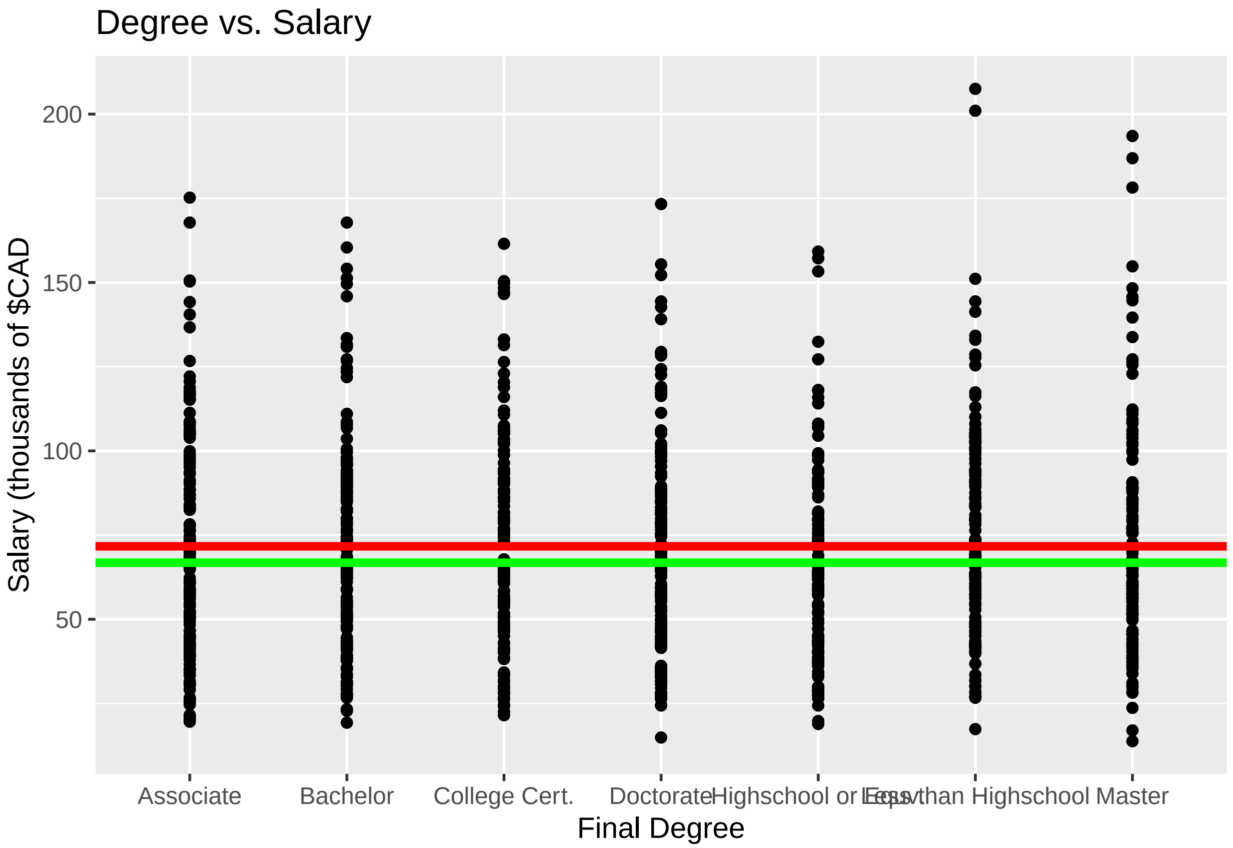
```
hist(survey_data$Salary, main = "Salary Distribution",
     xlab = "Salary in thousands ($CAD)", ylab = "Number of People", col = "blue", breaks=15)
abline(v=mean(salary), col="red", lwd = 3)
abline(v=median(salary), col="green", lwd = 3)
```



```
hist(survey_data$Grade, main = "Grade Distribution",
     xlab = "Grade in Percentage", ylab = "Number of People", col = "blue", breaks=15)
abline(v=mean(grade), col="red", lwd = 3)
abline(v=median(grade), col="green", lwd = 3)
```



```r
library(ggplot2)
ggplot(survey_data, aes(x = Final_Degree, y = Salary)) +
  geom_point() +
  labs(x = "Final Degree", y = "Salary (thousands of $CAD") +
  ggtitle("Degree vs. Salary") +
  geom_hline(yintercept = mean(survey_data$Salary), col = "red", lwd = 1.5) +
  geom_hline(yintercept = median(survey_data$Salary), col = "green", lwd = 1.5)
```



The first histgram is the distribution of the salaries earned by Canadians with age between 18-30. The x-axis represents the salary in thousands of Canadian dollars and the y-axis represents the frequency. The green vertical line represents the median salary and the red vertical line represents the mean salary. Because it was sampled from Gamma(5.2,14), the sample distribution also resembles the population distribution which approximates the true Canadian salary distribution very well.

The second histogram is the distribution of the grades. The x-axis represents the grade in percentage and the y-axis represents the frequency. The green vertical line represents the median grade and the red vertical line represents the mean grade. Because it was sampled from a normal distribution, the sample is also approximately normal, which is a typical distribution found in grades.

Finally, the last plot is a scatter plot of Salary on the y-axis vs Level of Degree on the x-axis. There is no obvious trend between the degree and their income, however this is solely because the sampling was random, meaning that someone with a higher degree has an equal probability of having a high salary assigned to them as to someone with less to no degree. However, if this survey was done in real life, and there indeed is a trend between the level of degree and their income, this plot will clearly reflect that, allowing us to conclude that college wage premium is very real. Similarly, the horizontal green and red line represents the median, and median salary.

Methods

In order to estimate the value of the true parameter, there are numerous different methods we may use. Hypothesis testing and Confidence interval is two out of many. No matter the process, the goal of these methods align. The goal is to use known statistic to make an inference about the true parameter. In this case, our parameter of interest is the true mean salary of Canadians between ages of 18-30 and true mean grade of the same population. My parameter of interest for the Confidence interval will be the mean salary and my parameter of interest for the hypothesis testing will be the mean grade.

Hypothesis Testing

To perform hypothesis testing, we must first define our null hypothesis, H_0 and alternative hypothesis H_A.

$$H_0 : \mu_0 = 75 \text{ vs } H_A : \mu_0 \neq 75$$

When performing a hypothesis test, we may encounter 2 types of error, Type 1 error is rejecting the null hypothesis when null hypothesis is true. We often denote probability of type 1 error occuring to be equal to alpha. Type 2 error is failing to reject null hypothesis when null hypothesis is false. Similarly, probability of type 2 occurring is equal to beta. Realistically, it is impossible to have alpha and beta both equal to 0. However, for a small fixed value of alpha (0.05), we may minimize beta through a likelihood ratio test.

The Likelihood Ratio Test Theorem (LRT) states that for a desired significance level alpha, there exists a some k between 0 and 1 such that,

$$\frac{L(\hat{\Omega}_0)}{L(\hat{\Omega})} \leq k$$

where the numerator is the maximum of the likelihood function with respect to our parameter of interest in the null parameter space (All possible values of null hypothesis). In this case, since our null hypothesis is singleton. we have L(75) in the numerator. The denominator is the maximum of the likelihood function with respect to our parameter of interest in the total paramter space which is the real line. Since our distribution is normal, the value that maximizes the likelihood function is ȳ. We also have a known variance of 49. From this we have,

$$\frac{\left(\frac{1}{\sqrt{98\pi}}\right)^n e^{\frac{-1}{98}\sum_{i=1}^n (y_i-75)^2}}{\left(\frac{1}{\sqrt{98\pi}}\right)^n e^{\frac{-1}{98}\sum_{i=1}^n (y_i-\bar{y})^2}} \leq k$$

$$\Rightarrow \frac{e^{\frac{-1}{98}\sum_{i=1}^n (y_i-75)^2}}{e^{\frac{-1}{98}\sum_{i=1}^n (y_i-\bar{y})^2}} \leq k$$

$$\Rightarrow \frac{e^{\frac{-1}{98}\sum_{i=1}^n (y_i-\bar{y})^2-\frac{n}{98}(\bar{y}-75)^2}}{e^{\frac{-1}{98}\sum_{i=1}^n (y_i-\bar{y})^2}} \leq k$$

$$\Rightarrow e^{-\frac{n}{98}(\bar{y}-75)^2-\frac{1}{98}\sum_{i=1}^n (y_i-\bar{y})^2+\frac{1}{98}\sum_{i=1}^n (y_i-\bar{y})^2} \leq k$$

$$\Rightarrow e^{-\frac{n}{98}(\bar{y}-75)^2} \leq k$$

$$\Rightarrow -\frac{n}{98}(\bar{y}-75)^2 \leq \ln(k) = k_1$$

$$\Rightarrow (\bar{y}-75)^2 \geq (\frac{-98}{n})k_1 = k_2$$

Under H_0, ȳ has a distribution ~ Normal(75,49/n). Therefore,

$$\left(\frac{\bar{y}-75}{\sqrt{\frac{49}{n}}}\right) \sim N(0,1)$$

$$\Rightarrow \left(\frac{\bar{y}-75}{\sqrt{\frac{49}{n}}}\right)^2 \sim \chi^2_{(1)}$$

Now,

$$\left(\frac{\bar{y}-75}{\sqrt{\frac{49}{n}}}\right)^2 \geq \frac{k_2}{\frac{49}{n}} = k^*$$

In order to find k*, we have,

$$\alpha = P\Bigg(\left(\frac{\bar{y}-75}{\sqrt{\frac{49}{n}}}\right)^2 \geq k^*\Bigg)$$

We will now need to find a mathematical decision model.

Results

Since, we set alpha = 0.05, and we have a chi-squared distribution with df=1 in this case, we know, k* must be equal to 0.05th percentile of chi-square, which is approximately 3.84146. we also have n=1000,

Therefore, in conclusion, our rejection region (RR) is defined as,

$$RR = \left\{\left(\frac{\bar{y}-75}{\sqrt{\frac{49}{1000}}}\right)^2 \geq 3.84146\right\}$$

In other words, we reject H_0 if the sample mean falls inside the RR. However, since our sample mean of the grade is equal to 75.1587, it does not fall under the RR. Therefore, we accept H_0 which states that the true population mean is 75.

Confidence Interval

Confidence interval is a range of values relatively small, such that the true value of the parameter of interest lies within the range with a fixed probability, we called this confidence level. In our case, our paramter of interest is the mean salary of Canadians between ages 18-30. Finding a 95% confidence interval means that we want to find a range of salary such that the probability of mean salary of the population lying within that range is 95%. In order to find the confidence interval, we must use the Central Limit Theorem (CLT).

CLT states that, regardless of the distribution of the population, the distribution of the sample mean will always be normal with the same mean, and variance of the original variance divided by the sample size, as long as it has a sufficient sample size (n>=30). In our case, the salary was distributed by Gamma(5.2,14). We also obtained sample mean of 71.6647 thousand $CAD and sample variance of 999.4502. By CLT, we know this sample mean is normally distributed. With the sample mean and the sample variance, we can make an inference about the true population mean. The given formula for a 95% CI on Mean Salary of the population is:

$$\bar{Y} \mp Z_{\frac{\alpha}{2}}\sqrt{\frac{s^2}{n}}$$

where Y bar is the sample mean, and Z is the Z-score of the normal distribution, for the case of 95% confidence interval, we take the Z score of 1.96. Finally, s^2 is the sample variance which is 999.4502 and we have n=1000 for the sample size.

Results

Therefore by computing these values we get that our 95% Confidence Interval is [69.7052749,73.6241251]. This means that, we may say that there is 95% probability that the true mean salary of the population lies between $69,705 and $73,624.

Part 3: Referencing

Generative AI Statement

I used the following generative artificial intelligence (AI) tool: ChatGPT Version available on September 27th, 2023. In order to figure out a function that allows me to calculate higher moments (skewness and kurtosis) of the sample, I prompted chatGPT "how do i compute skewness and kurtosis in R" which gave me an output instructing me to run library(moments) then using the skewness() and kurtosis() function.

I have also prompted ChatGPT to fix my loop used for fixing some logical errors in my sampling (e.g person acquiring Master's Degree when their highest level of education was high school.) The output gave me suggestions and corrections to get the result I needed.

Bibliography

1. Grolemund, G. (2014, July 16) *Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/articles_intro.html. (Last Accessed: April 4, 1991)
2. RStudio Team. (2020). *RStudio: Integrated Development for R*. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.
3. Allaire, J.J., et. el. *References: Introduction to R Markdown*. RStudio. https://rmarkdown.rstudio.com/docs/. (Last Accessed: April 4, 1991)
4. OpenAI. (2023). *ChatGPT (September 13 version) [Large language model]*. https://chat.openai.com/chat (Last Accessed: September 13, 2023)
5. Jeudy, Lucie. "Canada: Total Income Distribution by Income Level." Statista, 15 Feb. 2023, www.statista.com/statistics/484838/income-distribution-in-canada-by-income-level/
6. Government of Canada, Statistics Canada. "Sex at Birth and Gender – 2021 Census Promotional Material." Government of Canada, Statistics Canada, 3 Jan. 2023, www.statcan.gc.ca/en/census/census-engagement/community-supporter/sex-birth-gender

Appendix

Here is a glimpse of the data set simulated/surveyed:

```
## Rows: 1,000
## Columns: 6
## $ Gender          <chr> "M", "F", "F", "F", "F", "F", "F", "M", "F", "M", "F",~
## $ Final_Degree    <chr> "Doctorate", "College Cert.", "College Cert.", "Highsc~
## $ School_Name     <chr> "University of Manitoba", "Humber College", "Douglas C~
## $ Grade           <dbl> 58.44, 80.24, 58.64, 87.07, 74.93, 70.26, 71.55, 68.70~
## $ Salary          <dbl> 74.6, 47.8, 91.8, 89.4, 78.6, 41.3, 75.7, 57.8, 41.4, ~
## $ Work_Experience <chr> "<1", ">10", "1~2", "3~5", ">10", "1~2", "6~10", "<1",~
```