# Possible factors that affected mental health and the total number of children of different household income brackets.

Hung Dao, Huynh Vu Minh Hoang 56

October 19, 2020

**Possible factors that affected mental health and the total number of children of different household income brackets**

## Hung Dao and Huynh Vu Minh Hoang

## October 19, 2020

### Abstract

The General Social Survey is a survey conducted by Statistics Canada yearly to create and draw an inference that follows the trend and behaviour of the Canadian population. We randomly sample from the 2017 GSS data to make a logistic model and look for connections between certain variables and make hypothesis upon our data.

### Introduction

With the raw data, we make boxplots between the total number of children, separating each by the income bracket of that family and was able to draw a hypothesis from that. Next, we randomly sampled our data from the size of 20602 observation to 200, while still keeping the core characteristics of the population. Sensing there could be a potential connection between variables age, hours worked and mental health, we decided to run a logistic model to build predict the probability of someone with poor mental health provided how much they worked on average and how old they are.

### Data

General Social Survey or GSS data is that is yearly conducted, with a sample size of 25,000 people. Based on Statistics Canada, the government entity that conducted and provided the data, the data collecting process traditionally have used 2 ways of surveying: Random Digit Dialing (RDD) and Computer Assisted Telephone Interviewing (CATI). Technological advancement and social environment urged for a change in how surveys need to be conducted. The GSS 's sampling frame is now comprised of everyone eligible for telephone surveying as well as self-completed online questionnaires. Some data, such as income, are not collected through the surveys but through drawn from tax records and other administrative files. Statistics Canada said that the purpose of GSS is to monitor changes and behaviour in the conditions and the well being of Canadians to properly provide information on certain social policy issues.

The data we used is the 2017 GSS, comprised of 20,602 observations with 81 different variables.

The target population would be the current population of Canada in 2017, roughly at 36.54 million people, as all the population of Canada is covered by the objective of the study.

The frame population is every one that is covered by the sampling frame, meaning anyone who can be contacted through the means of data collection for the survey.

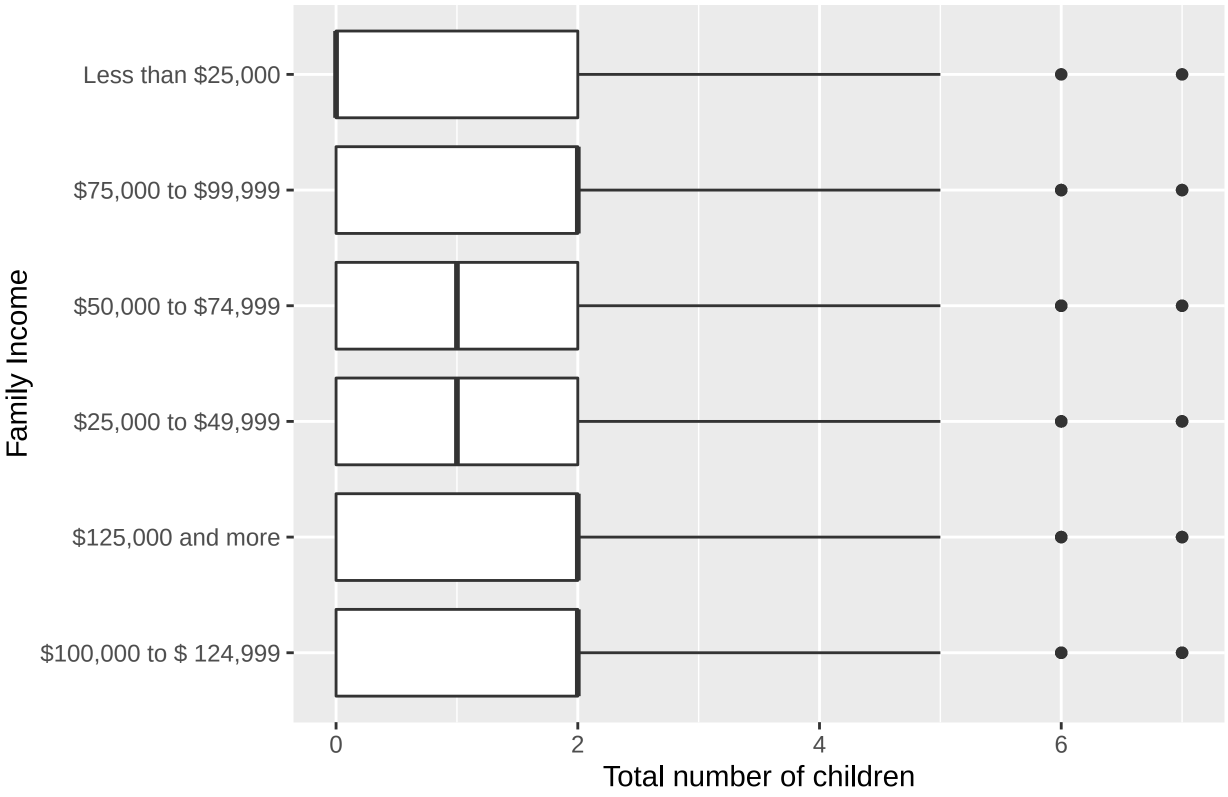
The sampled population is represented by the 20,602 observations in the GSS data.

In this 2017 GSS data, we are interested in a variety of variables such as the number of total children, income in the family, . . . We would like to study in-depth and draw inferences from the possible relationships between interested variables.

Key features of the survey include very detailed demographics data from ages, income, genders to self-rated health and mental health, religion participation. The strength of the data is its attention to detail for variables, beneficial for various cross-sectional studies. A drawback would be a high volume of categorical variables. While it is not impossible to make a model to draw inference from, it would have been easier with a presence of more quantitative variables. Instead of income is divided as brackets, would be easier to make a regression model for income and age if income is a quantitative variable.

we mainly focus on the variable income_family, total_children, age, self_rated_mental_heatlh and average_hours_worked. The variable self_rated_mental_health had been mutated into adjusted_mental_health, which is a binary variable, 0 for poor and 1 for the rest(Excellent, Very Good, Good, Fair, Unsure). The Variable adjusted_hours_worked had been mutated into adjusted_hours_worked, which now comprised of 1,2,3,4 and 5, representing 0 hours worked, 0,1 to 29.9 hours worked, 30 to 40 hours worked, and 50+ hours worked, respectively.

Figure 1. Boxplot of general population



The plot above showed the characteristic of the variable “Total number of children” across different bracket of the variable “Family Income”. We can see the mean for total children in each household across the income bracket, with the mean of the whole sampled population is 1.679. An interesting thing is a mean change from

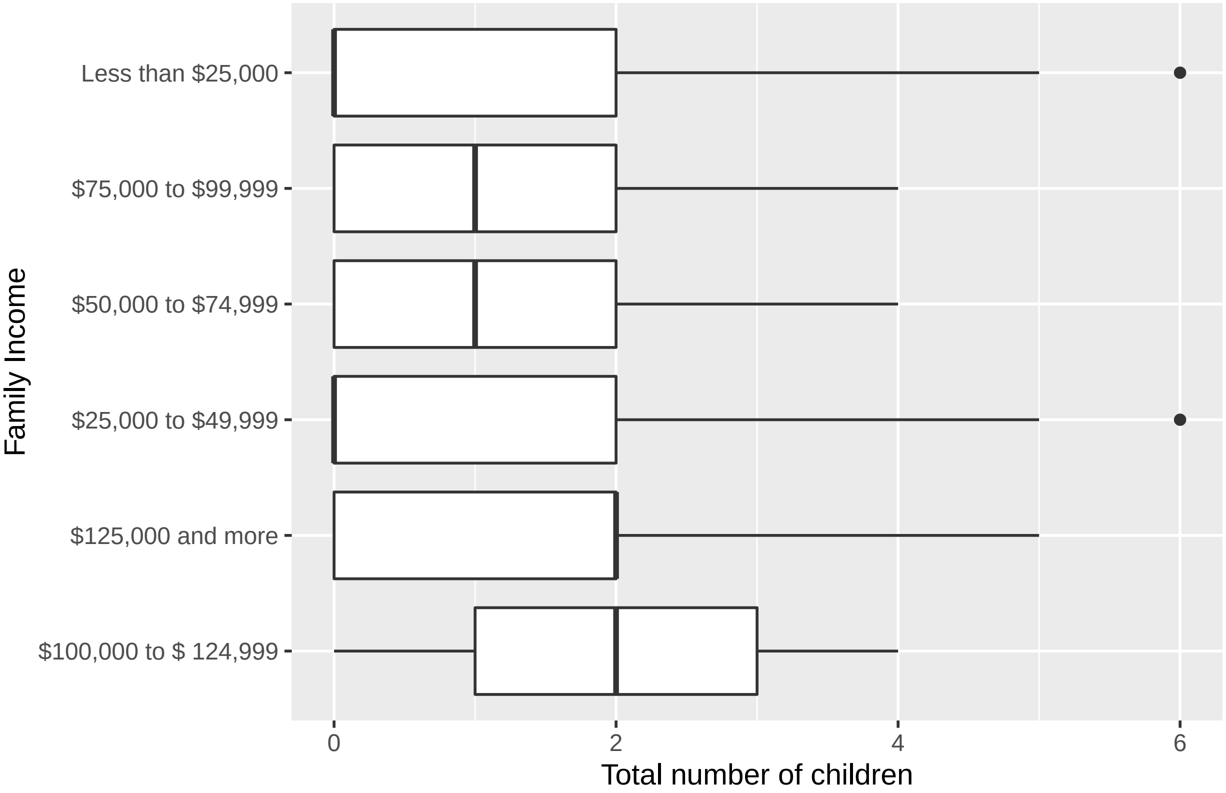
0 to 1 when we move from “less than $25,000” to the next tier “$25,000 to $49,000. the Same thing happens when we move further up from “$50,000 to $74,999” to “$75,000 to $99,000”, the mean of the number of children moved from 1 to 2.

From these changes in the mean, we would like to make a hypothesis that families with a higher income would be more likely to have more children than a family with a much lower income.

## Model

We randomly sampled without replacement 200 samples drawn from the sampled population.

Figure 2. Boxplot between interested variables Sample Data



For this part, we decided to run a logistic regression to see the relationship between mental health, age and hours worked. With mental health is the dependent variable, with both age and hours worked is the independent variables. Mental health is classified as a binary variable, 0 for poor and 1 for all unsure, fair, good, excellent. Age is a quantitative variable and hours worked is a qualitative variable, classifying in term of rank.

rank 1: 0 hours worked

rank 2: 0.1 to 29.9 hours worked

rank 3: 30 to 40 hours worked

rank 4: 41 to 50 hours worked

rank 5: 50+ hours worked

The model is built with the purpose of understanding does more hours worked combined with ages would and how it will affect a person self-rated mental health. Since we are using logistic regression, it’s a purpose is to model the probability of a person's mental health, given the number of hours they worked on average and how old they are.

The null hypothesis would be that ages and hours worked does not have an impact on mental health.

```
##
## Call:
## glm(formula = adjusted_mental_health ~ age + as.factor(adjusted_hours_worked),
##     family = "binomial", data = sample_data)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -3.09028   0.09414   0.13394   0.19816   0.54411
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      1.234e+00  1.341e+00   0.921    0.357
## age                              3.930e-02  3.317e-02   1.185    0.236
## as.factor(adjusted_hours_worked)3 2.007e+00  1.251e+00   1.605    0.109
## as.factor(adjusted_hours_worked)4 1.964e-01  1.279e+00   0.154    0.878
## as.factor(adjusted_hours_worked)5 1.658e+01  2.361e+03   0.007    0.994
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 39.216  on 199  degrees of freedom
## Residual deviance: 33.595  on 195  degrees of freedom
## AIC: 43.595
##
## Number of Fisher Scoring iterations: 18
```

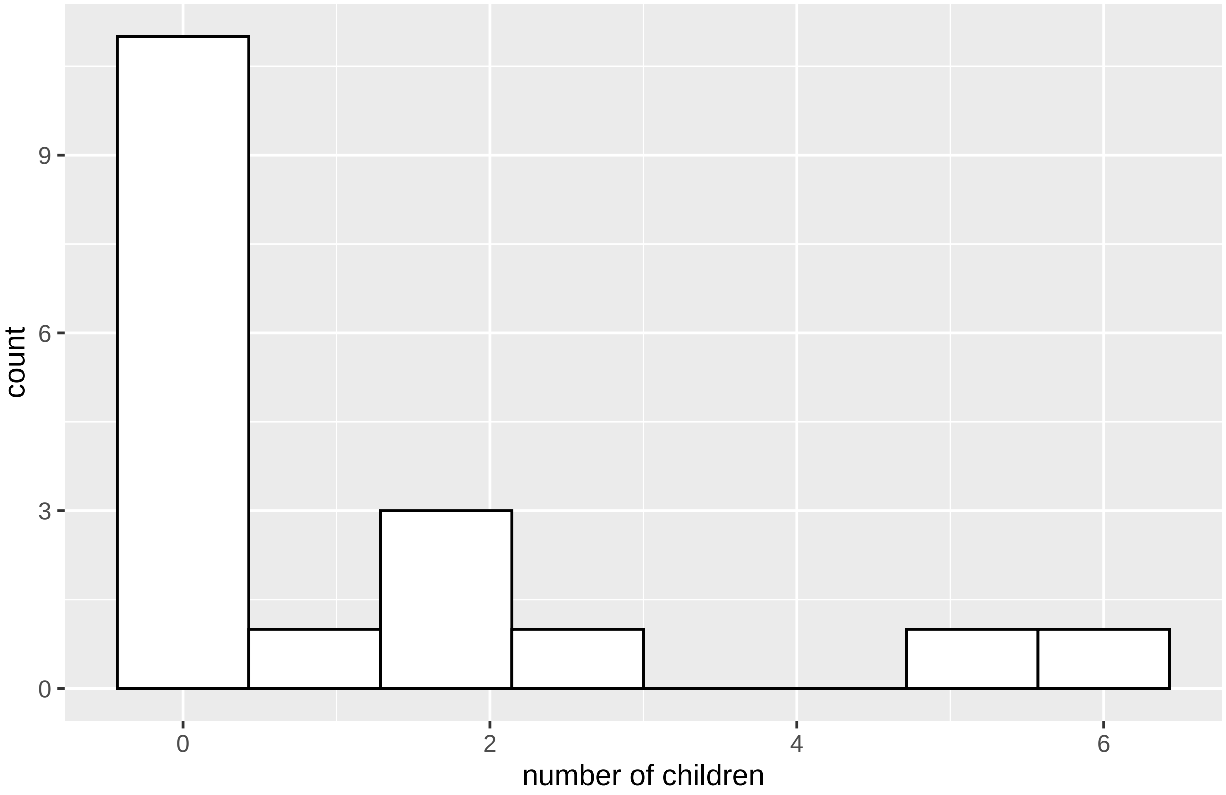
Logistic model:

$$log(\frac{p}{1-p}) = 1.234 + 0.0393 * age + 2.007 * rank_3 + 0.1964 * rank_4 + 16.58 * rank_5$$

## Results

Below are histograms for total children in a family under each income bracket.

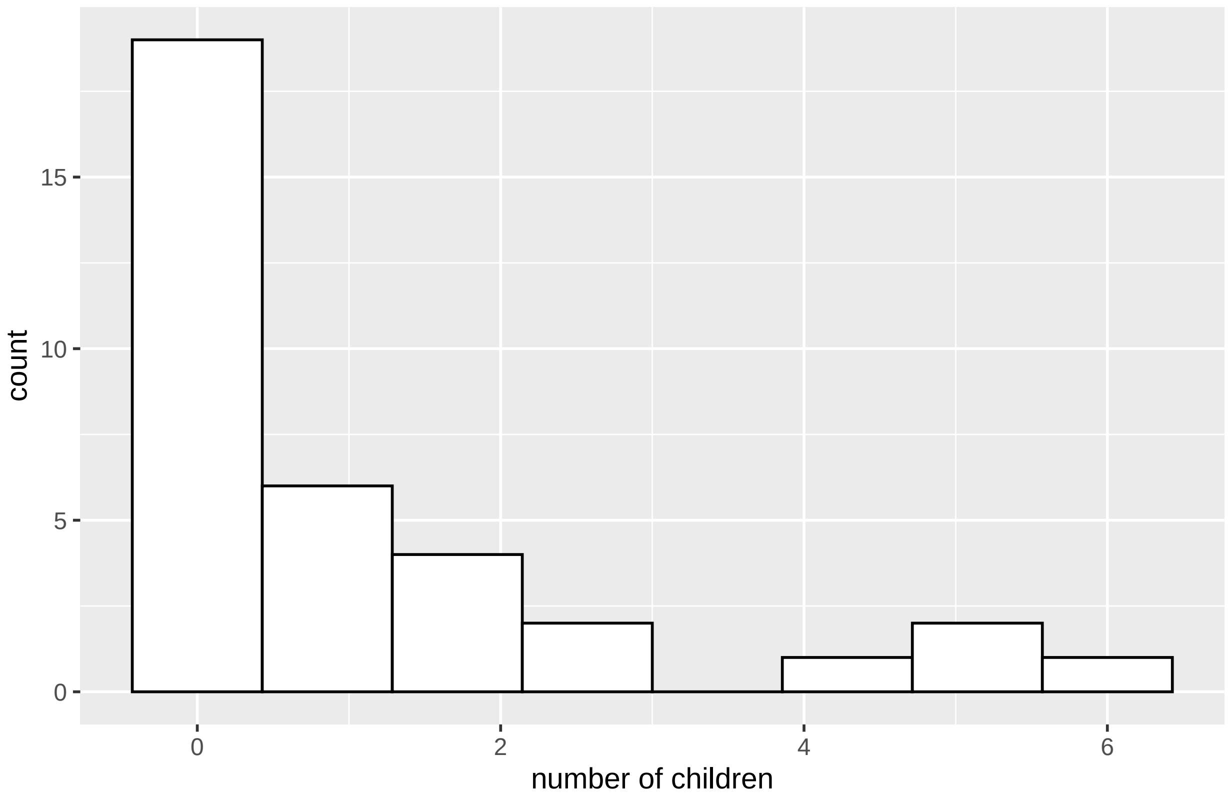
Figure 3. Histogram for family with income below $25,000



```
##  income_family      total_children       age        adjusted_mental_health
##  Length:18          Min.   :0.000   Min.   :17.30   Min.   :1
##  Class :character   1st Qu.:0.000   1st Qu.:28.15   1st Qu.:1
##  Mode  :character   Median :0.000   Median :39.55   Median :1
##                     Mean   :1.167   Mean   :39.91   Mean   :1
##                     3rd Qu.:2.000   3rd Qu.:51.05   3rd Qu.:1
##                     Max.   :6.000   Max.   :70.30   Max.   :1
##  adjusted_hours_worked
##  Min.   :2.000
##  1st Qu.:2.000
##  Median :3.000
##  Mean   :3.167
##  3rd Qu.:3.750
##  Max.   :5.000
```

According to the 5 number summary for this income bracket of “less than $25,000”, the min and the first quantile or the 25th quantile is 0, the mean is 1.167, with the 3rd quantile at 3 and the max is at 6. The histogram appears to be very heavily skewed to the right also. A note is that all of our 18 observations in this subgroup all have more than fair self-rated mental health as well.

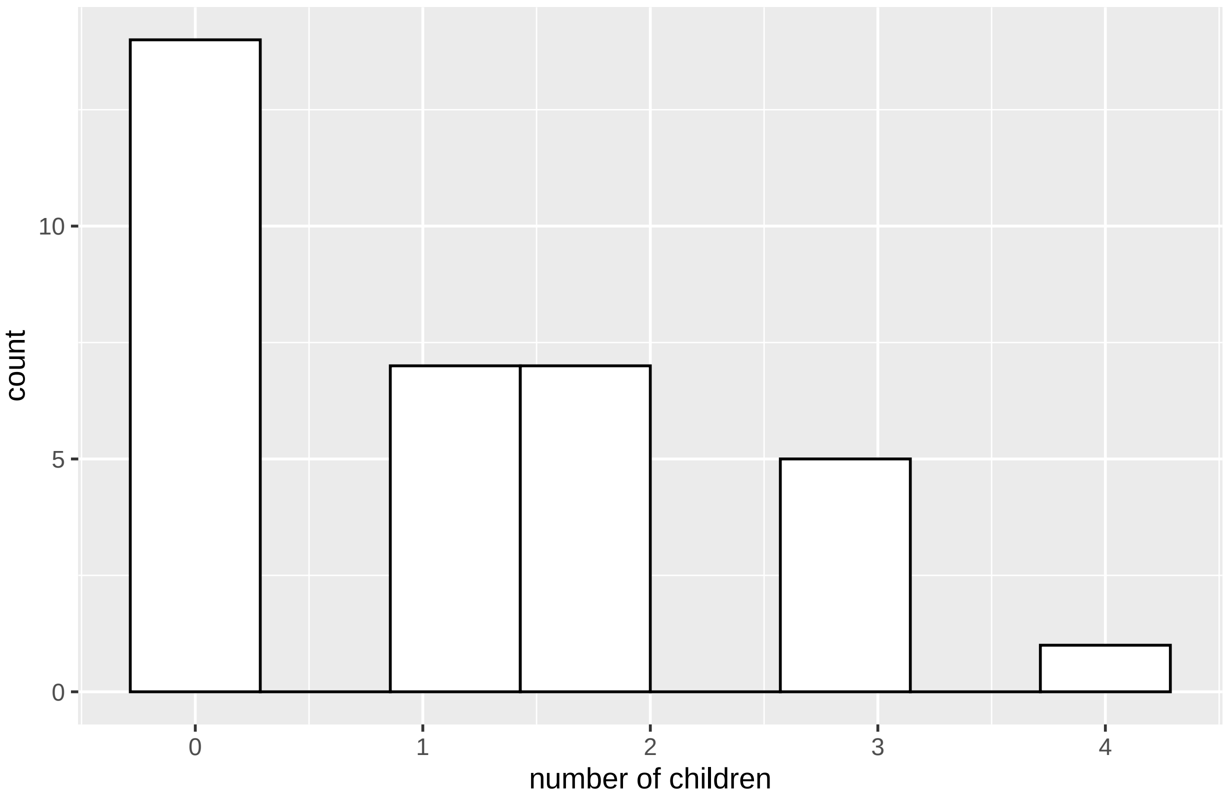
Figure 4. Histogram for family with income between $25,000 to $49,999



```
##  income_family      total_children       age        adjusted_mental_health
##  Length:35          Min.   :0.000   Min.   :15.30   Min.   :0.0000
##  Class :character   1st Qu.:0.000   1st Qu.:26.60   1st Qu.:1.0000
##  Mode  :character   Median :0.000   Median :35.40   Median :1.0000
##                     Mean   :1.143   Mean   :40.72   Mean   :0.9714
##                     3rd Qu.:2.000   3rd Qu.:52.95   3rd Qu.:1.0000
##                     Max.   :6.000   Max.   :72.80   Max.   :1.0000
##  adjusted_hours_worked
##  Min.   :2.000
##  1st Qu.:2.500
##  Median :3.000
##  Mean   :3.029
##  3rd Qu.:3.000
##  Max.   :5.000
```

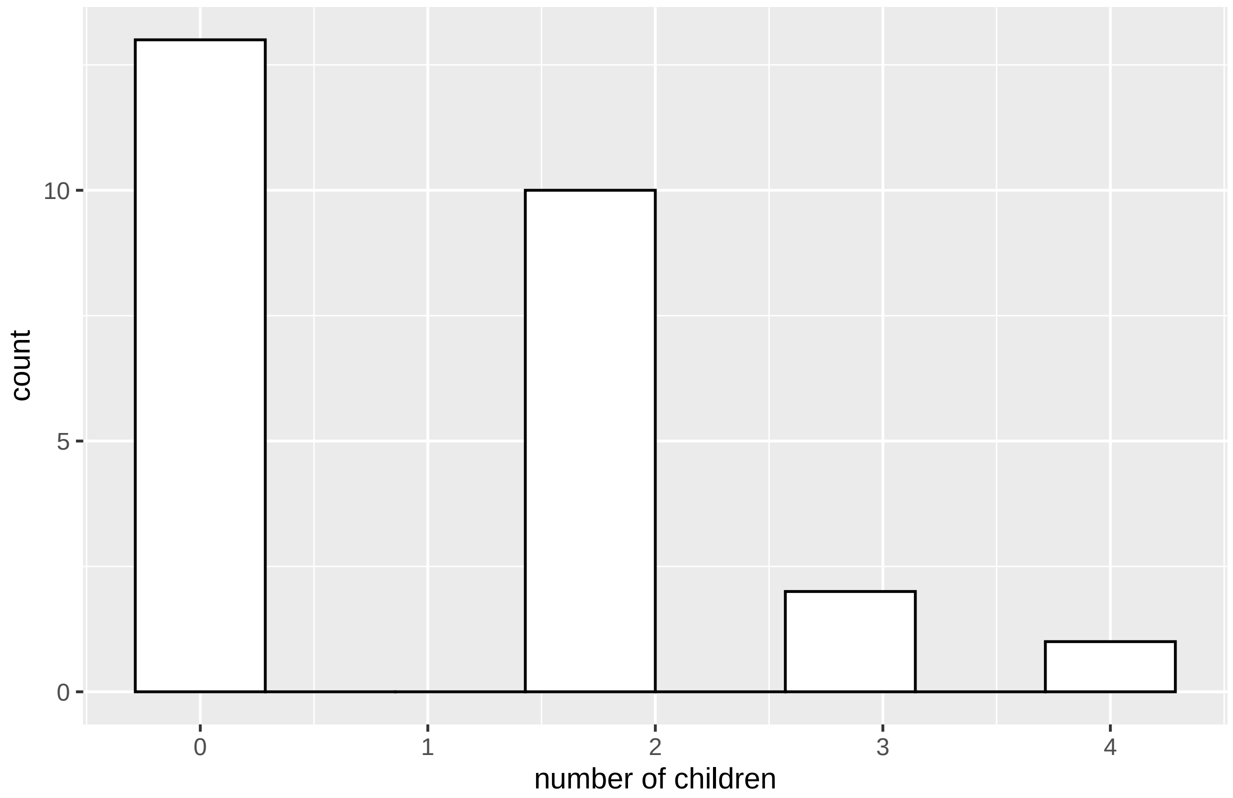
In the income bracket, with 35 observations, there is now 1 observation with poor mental health. The 5 number summary for this bracket is very similar with the first income bracket, with a 0 for both the min and the 1st quantile for the total children have while then mean is now 1.143 with the max at 6.

Figure 5. Histogram for family with income between $50,000 to $74,999



```
##  income_family      total_children       age        adjusted_mental_health
##  Length:34          Min.   :0.000   Min.   :24.50   Min.   :0.0000
##  Class :character   1st Qu.:0.000   1st Qu.:34.52   1st Qu.:1.0000
##  Mode  :character   Median :1.000   Median :49.10   Median :1.0000
##                     Mean   :1.176   Mean   :48.74   Mean   :0.9706
##                     3rd Qu.:2.000   3rd Qu.:62.65   3rd Qu.:1.0000
##                     Max.   :4.000   Max.   :77.20   Max.   :1.0000
##  adjusted_hours_worked
##  Min.   :2.000
##  1st Qu.:3.000
##  Median :3.000
##  Mean   :3.059
##  3rd Qu.:3.000
##  Max.   :5.000
```

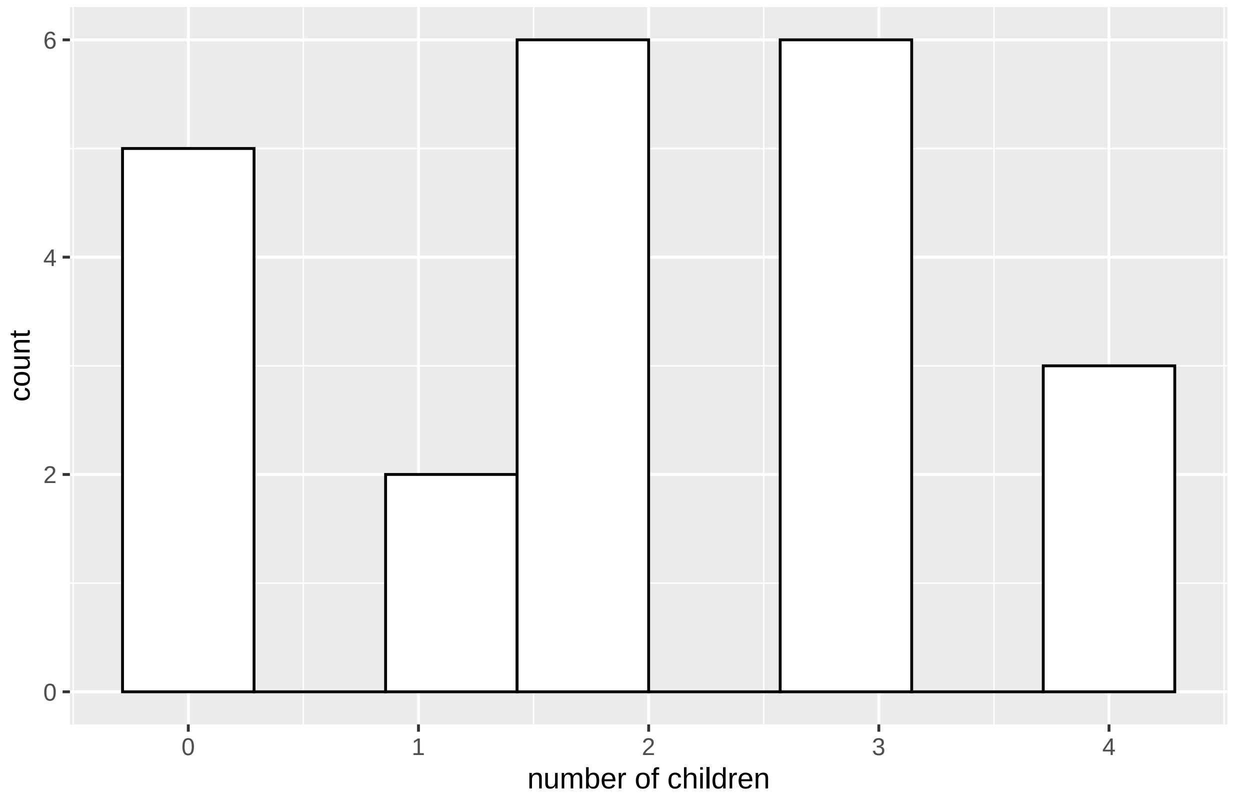
Figure 6. Histogram for family with income between $75,000 to $99,999



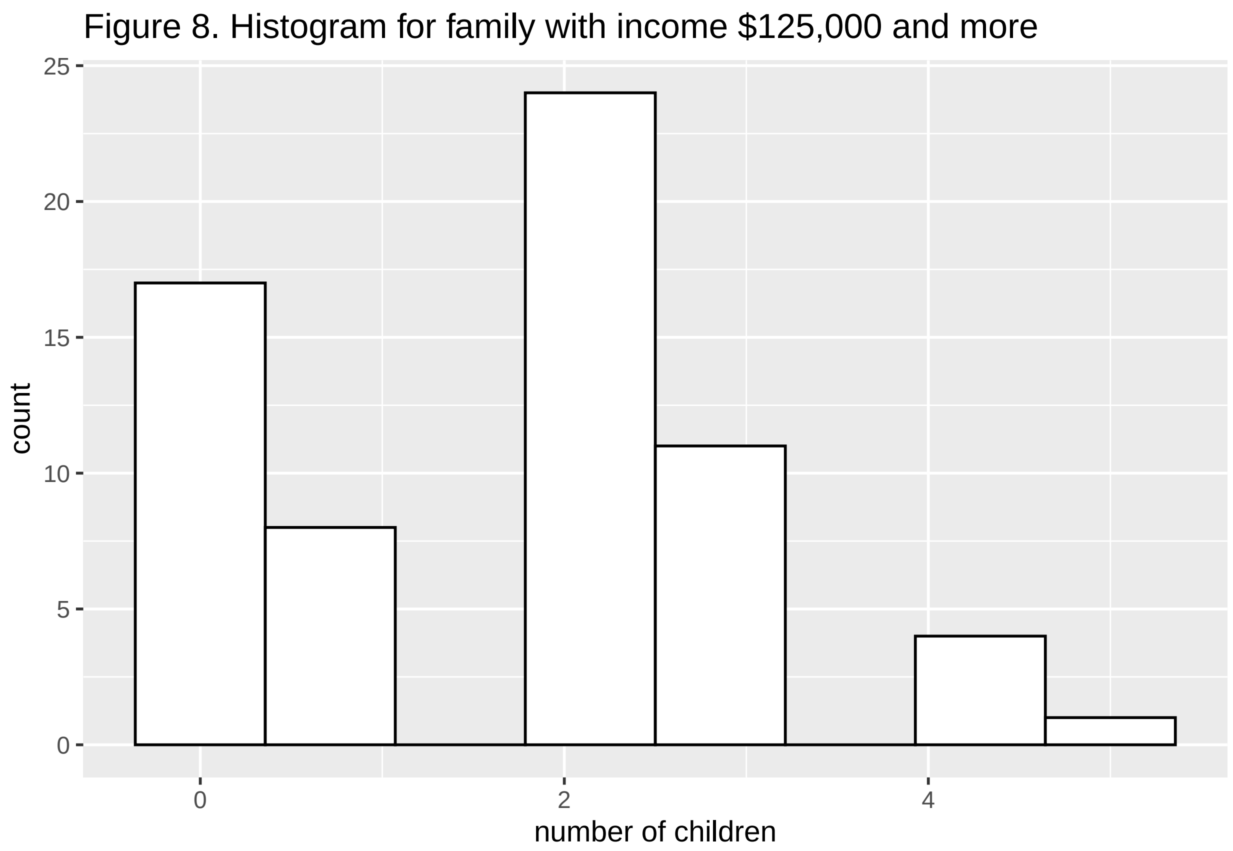
```
##   income_family      total_children       age        adjusted_mental_health
##   Length:26          Min.   :0.000   Min.   :16.40   Min.   :1
##   Class :character   1st Qu.:0.000   1st Qu.:32.15   1st Qu.:1
##   Mode  :character   Median :1.000   Median :38.70   Median :1
##                      Mean   :1.154   Mean   :41.53   Mean   :1
##                      3rd Qu.:2.000   3rd Qu.:52.90   3rd Qu.:1
##                      Max.   :4.000   Max.   :62.80   Max.   :1
##   adjusted_hours_worked
##   Min.   :2.000
##   1st Qu.:3.000
##   Median :3.000
##   Mean   :3.192
##   3rd Qu.:3.750
##   Max.   :5.000
```

While this graph is similar to the first 2 in term of its statistics, it is noted that as income bracket increases, the histogram is less and less right-skewed.

Figure 7. Histogram for family with income between $100,000 to $124,999



```
##  income_family      total_children      age        adjusted_mental_health
##  Length:22          Min.   :0      Min.   :20.80   Min.   :0.0000        
##  Class :character   1st Qu.:1      1st Qu.:35.27   1st Qu.:1.0000        
##  Mode  :character   Median :2      Median :44.00   Median :1.0000        
##                     Mean   :2      Mean   :45.15   Mean   :0.9545        
##                     3rd Qu.:3      3rd Qu.:54.15   3rd Qu.:1.0000        
##                     Max.   :4      Max.   :70.20   Max.   :1.0000        
##  adjusted_hours_worked
##  Min.   :2.000        
##  1st Qu.:3.000        
##  Median :3.000        
##  Mean   :3.045        
##  3rd Qu.:3.000        
##  Max.   :5.000        
```



```
##  income_family      total_children       age        adjusted_mental_health
##  Length:65          Min.   :0.000   Min.   :17.60   Min.   :0.0000
##  Class :character   1st Qu.:0.000   1st Qu.:40.80   1st Qu.:1.0000
##  Mode  :character   Median :2.000   Median :47.70   Median :1.0000
##                     Mean   :1.692   Mean   :46.25   Mean   :0.9846
##                     3rd Qu.:2.000   3rd Qu.:55.20   3rd Qu.:1.0000
##                     Max.   :5.000   Max.   :73.10   Max.   :1.0000
##  adjusted_hours_worked
##  Min.   :2.000
##  1st Qu.:3.000
##  Median :3.000
##  Mean   :3.246
##  3rd Qu.:4.000
##  Max.   :5.000
```

In the 65 observations for the highest income bracket group, the mean of total children is at 1.692, while the mean is at 2, the only income bracket group with the mean lower than the median, indicating there is a left-skewed in the histogram.

Overall, all the graphs made from our randomly sampled without replacement have the mean at 1.1 and are right-skewed, while the highest income bracket has the mean of 1.67 and left-skewed. As for the first 4 histograms, as the income increases, there is a reduction is skewness in each graph.

## Discussion

During our study, there seems to be a weak positive correlation between the total number of children of a family and their household income bracket. According to Statistics Canada, the average household income in 2017 was estimated to be $59,800, so we can see that a greater proportion of families with below the average income level tends to stay childless and this number decrease for more well off household.

Logistic model:

$$log(\frac{p}{1-p}) = 1.234 + 0.0393 * age + 2.007 * rank_3 + 0.1964 * rank_4 + 16.58 * rank_5$$

In which:

rank 1: 0 hours worked

rank 2: 0.1 to 29.9 hours worked

rank 3: 30 to 40 hours worked

rank 4: 41 to 50 hours worked

rank 5: 50+ hours worked

Since our model is built with the response variable, 0 for poor mental health and 1 for fair and above, a probability $p$ close to 1 indicates a high probability of not having poor mental health, while probability $p$ close to 0 suggests that person self-assessment of their mental health be quite low. With the slope for the variable age is 0.0393, smaller comparing to all the other possible slopes for hours worked, we get that age to have a smaller impact than comparing to the hours worked. While we see the slop of hours worked for rank 1 and 2 is 0, which mean hours worked below 30 hours does not impact on mental health, only age does for those who work at maximum 29.9 hours. Rank 3 has a significantly higher slope than rank 4, almost 10 times bigger, and rank 5 has the slope of 16.58, having the greatest impact on mental health. Since the function on the right-hand side is all positive, we can deduce that as the sum of the function on the right-hand side gets bigger, which could mean that on average, they work for a longer period, hence could put more pressure on their mental health, resulting in lower $p$ value.

A drawback for the model would be that the respective value for all of the variables is very high, which would also mean that we would fail to reject the null hypothesis, meaning either that ages and hours worked does not contribute to mental health or we are missing crucial variables that would help us explain the model significantly better.

# Weaknesses

A weakness in the survey process is data collecting through telephone interviews can limit ceratin people in the demographic, as well as the type of data that can be collected. Families that do not have a telephone would be excluded from the CATI sample, while families that are not well versed with the Internet would be excluded from the sample of the online questionnaire. Another problem that surveys and phone interviews generally face is the social desirability bias, which is a type of response bias where the respondent would answer questions not truthfully, but in a way that it would be considered as favoured by others.

Where the data could have been improved upon is going from dividing variables such as hours worked and income into brackets, to have the variables as a quantitative response, clear numbers that can be used to plots and make models more detailed.

The weakness of the study is the lack of multiple models that are needed to get inferences, whether to reject the hypothesis that families with a higher income would be more likely to have more children than a family with a much lower income.

# Next Steps

We could run a hypothesis test for the hypothesis of families with a higher income would be more likely to have more children than a family with a much lower income. For subsequent surveys, a survey that grades a person mental health more in-depth could be beneficial, as well as more details for incomes and hours worked. With a more in-depth mental health score, a subsequent study of how ages, number of children impact on mental health.

## References

CHASS, (2017). GSS17, 2017 (Computing in the Humanities and Social Sciences). Retrieved October 19, 2020, from http://www.chass.utoronto.ca/

Canada, S. (2019, February 26). Canadian Income Survey, 2017 (Canada, Statistics Canada). Retrieved October 19, 2020, from https://www150.statcan.gc.ca/n1/daily-quotidien/190226/dq190226b-eng.htm

Alexander, R. (2020). Telling Stories With Data [Web log post]. Retrieved 2020, from https://www.tellingstorieswithdata.com/