Assignment 2

Mel Miller
9/16/2020

Table of Contents

Load Relevant Libraries

I will first load the libraries I'll need to complete this assignment.

```
library (tidyverse)
library (ggplot2)

person_data <- read_csv("people.csv")
attach(person_data)
```

Distribution of Continuous Variables

My dataset, filtered for older adults (age 65+) in Massachusetts, contains the following 4 continuous variables:

1. Age (AGEP)
2. Total income, past 12 months (PINCP)
3. Social Security Income, past 12 months (SSP)

4. Number of persons in respective household (NP)

For each of these variables, I will include the following statistics:

- Sample mean
- Standard deviation
- 95-percent confidence interval for the population mean
- The interquartile range
- A histogram to detail the variable's distribution

Age

Mean, standard deviation, and 95-percent confidence interval

The mean of the age variable in my data set is 74.55, with a standard deviation of 7.85, and a 95-percent confidence interval with a lower limit of 74.42 and upper limit of 74.69 From this point forward, I'll report this data as: *mean* (SD=*SD*, 95% CI: *CI lower limit*, *CI upper limit*). In other words, the sample mean for age is 74.55 (SD=7.85, 95% CI: 74.42, 74.69).

```
mean(AGEP)

## [1] 74.55433

sd(AGEP)

## [1] 7.852848

conf_int <- t.test(person_data$AGEP)
conf_int$conf.int[1]
```

```
## [1] 74.42341
```

```r
conf_int$conf.int[2]
```

```
## [1] 74.68526
```

Interquartile range

The interquartile range for age is 68 to 79. The median age in this sample is 72.

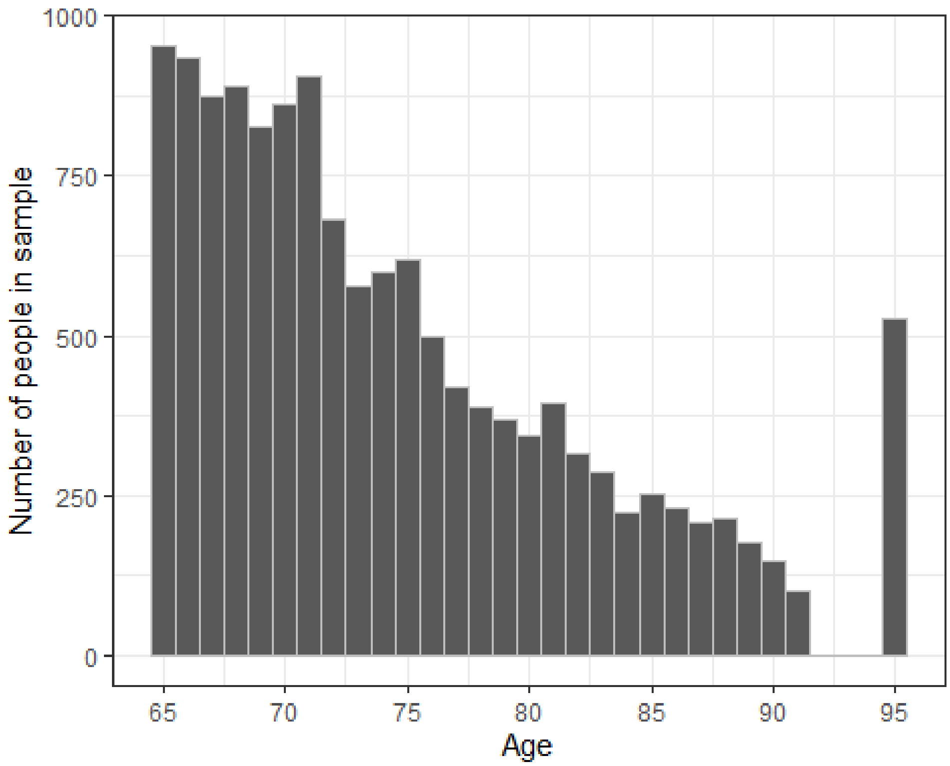
```r
quantile(AGEP)
```

```
##   0%  25%  50%  75% 100% 
##   65   68   72   79   95
```

Histogram

The histogram for age distribution in my sample is skewed right, meaning that more people in this sample are closer to age 65 than they are to age 95. The unimodal distribution peaks at 75, and there appears to outliers for age 95. Given this gap between age 91 and 95, however, it seems that all folks who were over the age of 91 were coded as 95 years old, probably because the low frequencies of ages over 91.

```r
ggplot(person_data, aes(x = AGEP)) +
  geom_histogram(color = "gray",
                 binwidth = 1) +
  theme_bw() +
  scale_x_continuous(name = "Age",
                     breaks = breaks <- seq(60, 100, by = 5)) +
  scale_y_continuous(name = "Number of people in sample")
```



Total income

Mean, standard deviation, and 95-percent confidence interval

The sample mean for total income is $46,439.75 (SD=$72,222.61, 95% CI: $45,235.62, $47,643.88).

```
mean(PINCP)

## [1] 46439.75

sd(PINCP)

## [1] 72222.61

conf_int <- t.test(person_data$PINCP)
conf_int$conf.int[1]

## [1] 45235.62

conf_int$conf.int[2]

## [1] 47643.88
```

Interquartile range

The interquartile range for total income is $12,900 to $52,000. The median income in this sample is $25,475.

```r
quantile(PINCP)
```

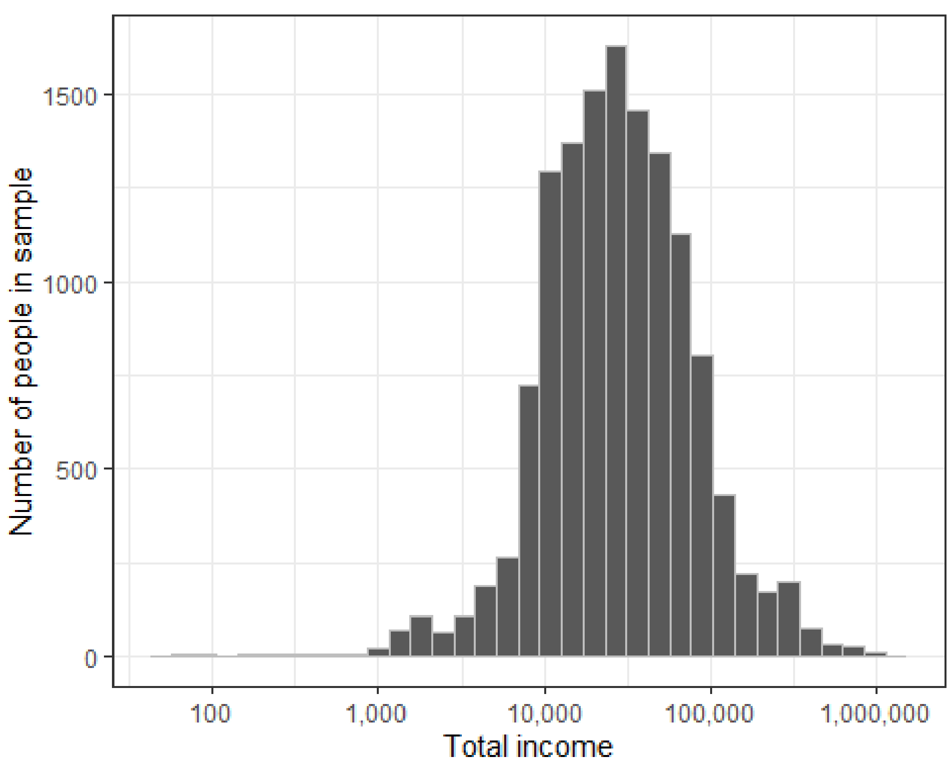
```
##       0%      25%      50%      75%     100%
##    -6100    12900    25475    52000 1353400
```

Histogram

I first ran this histogram with a linear x-axis. The shape was hard to distinguish, and the x-axis was not legible. Placing this histogram on a log x-axis, it is clear that the distribution for total income has a normal distribution, making this a log-normal distribution.

```r
ggplot(person_data, aes(x = PINCP)) +
  geom_histogram(color = "gray",
                 binwidth=.3) +
  theme_bw() +
  scale_x_continuous(name = "Total income",
                     trans = "log",
                     breaks = c (10, 100, 1000, 10000, 100000, 1000000),
                     labels = c ("10", "100", "1,000", "10,000", "100,000",
"1,000,000")) +
  scale_y_continuous(name = "Number of people in sample")
```

```
## Warning in self$trans$transform(x): NaNs produced

## Warning: Transformation introduced infinite values in continuous x-axis

## Warning: Removed 558 rows containing non-finite values (stat_bin).
```



Social Security Income

Mean, standard deviation, and 95-percent confidence interval

The sample mean for social security income is $11,936.87 (SD=$9,532.18, 95% CI: $11,777.94, $12,095.79).

```
mean(SSP)

## [1] 11936.87

sd(SSP)

## [1] 9532.176

conf_int <- t.test(person_data$SSP)
conf_int$conf.int[1]

## [1] 11777.94

conf_int$conf.int[2]

## [1] 12095.79
```

Interquartile range

The interquartile range for social security income is $2,400 to $18,500. The median social security income is $12,000.

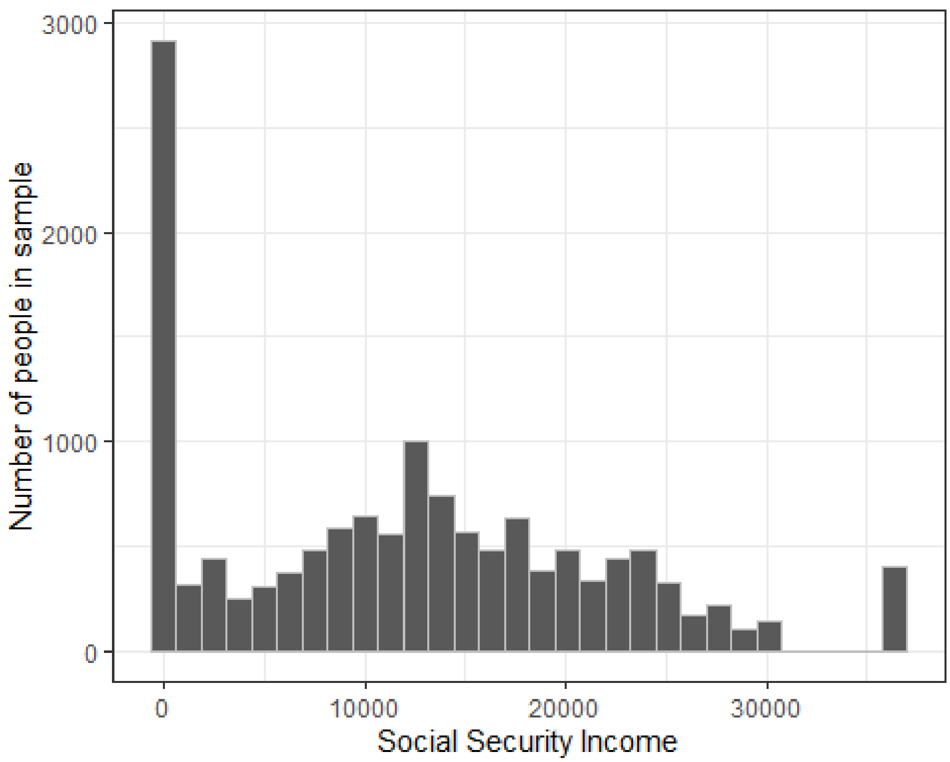
```
quantile(SSP)

##    0%   25%   50%   75%  100%
##     0  2400 12000 18500 36400
```

Histogram

The histogram for social security income appears to be random (i.e., it does not have an apparent pattern). The distribution is greatly impacted by the initial peak at $0. Because of this peak, I considered removing all individuals who do not receive Social Security Income from my sample. However, I am curious to learn more about these folks and their relationships with other variables, and how these relationships differ from those who receive this income. I therefore decided to leave this variable as is for now.

There also seems to be outliers at $36,400. Similar to the outliers on the age histogram, I wonder if the number of social security incomes over $30,000 is so low that anyone who reported a social security income over this $30,000 had this variable coded as $36,400.

```
ggplot(person_data, aes(x = SSP)) +
  geom_histogram(color = "gray")+
  theme_bw() +
  scale_x_continuous(name = "Social Security Income") +
  scale_y_continuous(name = "Number of people in sample")
```



Household size

Mean, standard deviation, and 95-percent confidence interval

The sample mean for household size is 2.03 (SD=1.16, 95% CI: 2.01, 2.05).

```
mean(NP)

## [1] 2.025756

sd(NP)

## [1] 1.15529

conf_int <- t.test(person_data$NP)
conf_int$conf.int[1]

## [1] 2.006494

conf_int$conf.int[2]

## [1] 2.045018
```

Interquartile range

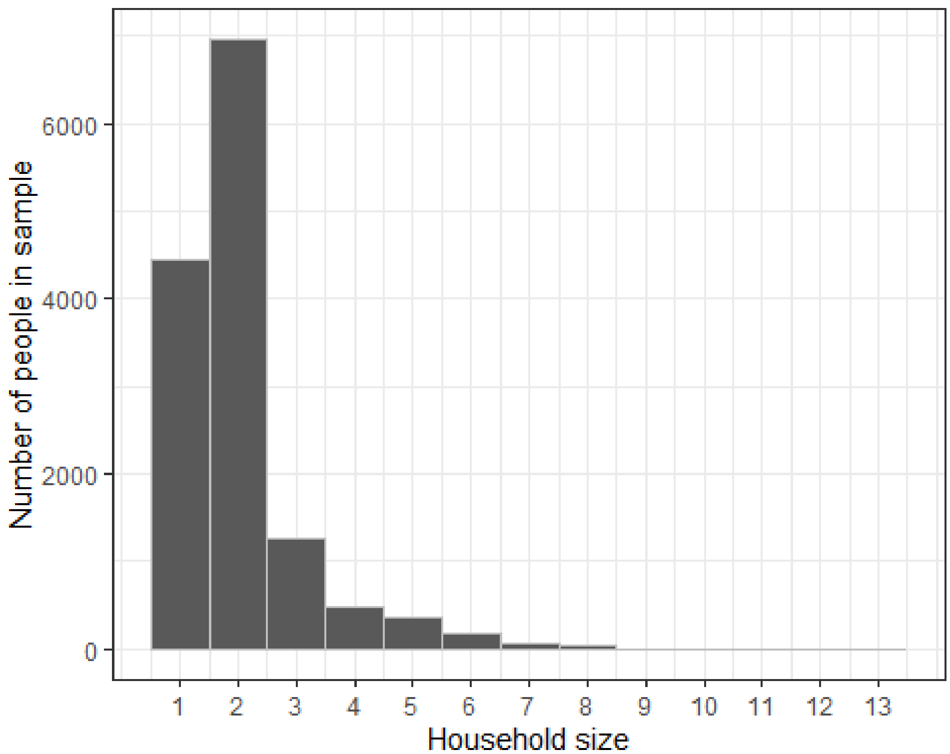
The interquartile range for household size is 1 to 2. The median for household size in this sample is also 2.

```
quantile(NP)

##   0%  25%  50%  75% 100%
##    1    1    2    2   13
```

Histogram

This unimodal Poisson distribution is skewed right, with its peak at 2.

```
ggplot(person_data, aes(x = NP)) +
  geom_histogram(color = "gray",
                 binwidth = 1) +
  theme_bw() +
  scale_x_continuous(name = "Household size",
                     breaks = seq(1, 13, by = 1))+
  scale_y_continuous(name = "Number of people in sample")
```



Distribution of Categorical Variables

My dataset contains the following 8 categorical variables, all filtered for older adults (age 65+) in Massachusets:

1. Sex (SEX)
2. Race (RAC1P)
3. Marital status (MAR)

4. Housing tenure (TEN)
5. Employment status (ESR)
6. When last worked (WKL)
7. Independent living difficulty (DOUT)
8. Grandparent living with children (GCL)

For each of these variables, I will include the 95-percent interval for the proportion of the population in each category. At the end of this file, I'll create one table (in Microsoft Word) that details this data for all eight variables.

```
unique(person_data$SEX_label)

## [1] "Female" "Male"

unique(person_data$RAC1P_label)

## [1] "White alone"
## [2] "Black or African American alone"
## [3] "Some Other Race alone"
## [4] "Two or More Races"
## [5] "Asian alone"
## [6] "American Indian alone"
## [7] "American Indian and Alaska Native tribes specified; or American
Indian or Alaska Native, not specified and no other races"

unique(person_data$MAR_label)

## [1] "Married"                             "Widowed"
## [3] "Never married or under 15 years old" "Divorced"
## [5] "Separated"

unique(person_data$TEN_label)

## [1] "N/A (GQ/vacant)"
## [2] "Owned free and clear"
## [3] "Rented"
## [4] "Owned with mortgage or loan (include home equity loans)"
## [5] "Occupied without payment of rent"

unique(person_data$ESR_label)

## [1] "Not in labor force"
## [2] "Civilian employed, at work"
## [3] "Civilian employed, with a job but not at work"
## [4] "Unemployed"

unique(person_data$WKL_label)

## [1] "Over 5 years ago or never worked" "Within the past 12 months"
## [3] "1-5 years ago"

unique(person_data$DOUT_label)
```

```
## [1] "Yes" "No"

unique(person_data$GCL_label)

## [1] "No"  "Yes"
```

Sex

```
t.test(person_data$SEX_label=="Female")

##
##  One Sample t-test
##
## data:  person_data$SEX_label == "Female"
## t = 133.13, df = 13821, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.5535854 0.5701304
## sample estimates:
## mean of x
## 0.5618579

t.test(person_data$SEX_label=="Male")

##
##  One Sample t-test
##
## data:  person_data$SEX_label == "Male"
## t = 103.82, df = 13821, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.4298696 0.4464146
## sample estimates:
## mean of x
## 0.4381421

table (person_data$SEX_label)/sum (table(person_data$SEX_label))

##
##    Female      Male
## 0.5618579 0.4381421
```

Race

```
t.test (person_data$RAC1P_label=="White alone")

##
##  One Sample t-test
##
## data:  person_data$RAC1P_label == "White alone"
## t = 380.07, df = 13821, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
```

```
##  0.9079685 0.9173824
## sample estimates:
## mean of x
## 0.9126754
```

```r
t.test (person_data$RAC1P_label=="Black or African American alone")
```

```
##
##  One Sample t-test
##
## data:  person_data$RAC1P_label == "Black or African American alone"
## t = 23.058, df = 13821, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.03389342 0.04019137
## sample estimates:
## mean of x
## 0.0370424
```

```r
t.test (person_data$RAC1P_label=="Some Other Race alone")
```

```
##
##  One Sample t-test
##
## data:  person_data$RAC1P_label == "Some Other Race alone"
## t = 12.438, df = 13821, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.009324856 0.012813764
## sample estimates:
##  mean of x
## 0.01106931
```

```r
t.test (person_data$RAC1P_label=="Two or More Races")
```

```
##
##  One Sample t-test
##
## data:  person_data$RAC1P_label == "Two or More Races"
## t = 9.6234, df = 13821, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.005300321 0.008011790
## sample estimates:
##   mean of x
## 0.006656056
```

```r
t.test (person_data$RAC1P_label=="Asian alone")
```

```
##
##  One Sample t-test
##
```

```
## data:  person_data$RAC1P_label == "Asian alone"
## t = 21.066, df = 13821, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.02821513 0.03400452
## sample estimates:
##  mean of x
## 0.03110982
```

```
t.test (person_data$RAC1P_label=="American Indian alone")
```

```
##
##  One Sample t-test
##
## data:  person_data$RAC1P_label == "American Indian alone"
## t = 4.0022, df = 13821, p-value = 6.309e-05
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.0005906318 0.0017245179
## sample estimates:
##   mean of x
## 0.001157575
```

```
t.test (person_data$RAC1P_label=="American Indian and Alaska Native tribes
specified; or American Indian or Alaska Native, not specified and no other
races")
```

```
##
##  One Sample t-test
##
## data:  person_data$RAC1P_label == "American Indian and Alaska Native
tribes specified; or American Indian or Alaska Native, not specified and no
other races"
## t = 2.0002, df = 13821, p-value = 0.0455
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  5.799031e-06 5.729884e-04
## sample estimates:
##    mean of x
## 0.0002893937
```

```
table (person_data$RAC1P_label)/sum (table(person_data$RAC1P_label))
```

```
##
##
American Indian alone
##
0.0011575749
## American Indian and Alaska Native tribes specified; or American Indian or
Alaska Native, not specified and no other races
##
```

```
0.0002893937
##
Asian alone
##
0.0311098249
##
Black or African American alone
##
0.0370423962
##
Some Other Race alone
##
0.0110693098
##
Two or More Races
##
0.0066560556
##
White alone
##
0.9126754449
```

Marital status

```
t.test (person_data$MAR_label=="Married")

##
##  One Sample t-test
##
## data:  person_data$MAR_label == "Married"
## t = 131.23, df = 13821, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.5464814 0.5630542
## sample estimates:
## mean of x
## 0.5547678

t.test (person_data$MAR_label=="Widowed")

##
##  One Sample t-test
##
## data:  person_data$MAR_label == "Widowed"
## t = 60.023, df = 13821, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.2000194 0.2135243
## sample estimates:
## mean of x
## 0.2067718
```

```r
t.test (person_data$MAR_label=="Never married or under 15 years old")
```

```
##
##  One Sample t-test
##
## data:  person_data$MAR_label == "Never married or under 15 years old"
## t = 37.509, df = 13821, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.08756084 0.09721705
## sample estimates:
##  mean of x
## 0.09238895
```

```r
t.test (person_data$MAR_label=="Divorced")
```

```
##
##  One Sample t-test
##
## data:  person_data$MAR_label == "Divorced"
## t = 46.473, df = 13821, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.1294467 0.1408471
## sample estimates:
## mean of x
## 0.1351469
```

```r
t.test (person_data$MAR_label=="Separated")
```

```
##
##  One Sample t-test
##
## data:  person_data$MAR_label == "Separated"
## t = 12.355, df = 13821, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.009191471 0.012657755
## sample estimates:
##  mean of x
## 0.01092461
```

```r
table (person_data$MAR_label)/sum (table(person_data$MAR_label))
```

```
##
##                            Divorced                             Married
##                          0.13514687                          0.55476776
## Never married or under 15 years old                           Separated
##                          0.09238895                          0.01092461
##                             Widowed
##                          0.20677181
```

Housing tenure

```
t.test (person_data$TEN_label=="N/A (GQ/vacant)")

##
##  One Sample t-test
##
## data:  person_data$TEN_label == "N/A (GQ/vacant)"
## t = 32.476, df = 13821, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.06662214 0.07518078
## sample estimates:
##  mean of x
## 0.07090146

t.test (person_data$TEN_label=="Owned free and clear")

##
##  One Sample t-test
##
## data:  person_data$TEN_label == "Owned free and clear"
## t = 100.5, df = 13821, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.4139904 0.4304605
## sample estimates:
## mean of x
## 0.4222254

t.test (person_data$TEN_label=="Rented")

##
##  One Sample t-test
##
## data:  person_data$TEN_label == "Rented"
## t = 51.729, df = 13821, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.1560588 0.1683515
## sample estimates:
## mean of x
## 0.1622052

t.test (person_data$TEN_label=="Owned with mortgage or loan (include home
equity loans)")

##
##  One Sample t-test
##
## data:  person_data$TEN_label == "Owned with mortgage or loan (include home
equity loans)"
```

```
## t = 82.706, df = 13821, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.3232201 0.3389127
## sample estimates:
## mean of x
## 0.3310664
```

```r
t.test (person_data$TEN_label=="Occupied without payment of rent")
```

```
##
##  One Sample t-test
##
## data:  person_data$TEN_label == "Occupied without payment of rent"
## t = 13.805, df = 13821, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.01167027 0.01553274
## sample estimates:
## mean of x
## 0.0136015
```

```r
table (person_data$TEN_label)/sum (table(person_data$TEN_label))
```

```
##
##                                         N/A (GQ/vacant)
##                                              0.07090146
##                        Occupied without payment of rent
##                                              0.01360150
##                                     Owned free and clear
##                                              0.42222544
## Owned with mortgage or loan (include home equity loans)
##                                              0.33106642
##                                                  Rented
##                                              0.16220518
```

Employment status

```r
t.test (person_data$ESR_label=="Not in labor force")
```

```
##
##  One Sample t-test
##
## data:  person_data$ESR_label == "Not in labor force"
## t = 216.52, df = 13821, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.7653279 0.7793111
## sample estimates:
## mean of x
## 0.7723195
```

```r
t.test (person_data$ESR_label=="Civilian employed, at work")

##
##  One Sample t-test
##
## data:  person_data$ESR_label == "Civilian employed, at work"
## t = 61.595, df = 13821, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.2085272 0.2222354
## sample estimates:
## mean of x
## 0.2153813

t.test (person_data$ESR_label=="Civilian employed, with a job but not at
work")

##
##  One Sample t-test
##
## data:  person_data$ESR_label == "Civilian employed, with a job but not at
work"
## t = 9.1376, df = 13821, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.004716783 0.007293057
## sample estimates:
##  mean of x
## 0.00600492

t.test (person_data$ESR_label=="Unemployed")

##
##  One Sample t-test
##
## data:  person_data$ESR_label == "Unemployed"
## t = 9.3565, df = 13821, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.004975694 0.007612933
## sample estimates:
##   mean of x
## 0.006294313

table (person_data$ESR_label)/sum (table(person_data$ESR_label))

##
##                      Civilian employed, at work
##                                     0.215381276
## Civilian employed, with a job but not at work
##                                     0.006004920
```

```
##                              Not in labor force
##                                     0.772319491
##                                      Unemployed
##                                     0.006294313
```

When last worked

```
t.test (person_data$WKL_label=="Over 5 years ago or never worked")

##
##  One Sample t-test
##
## data:  person_data$WKL_label == "Over 5 years ago or never worked"
## t = 142.49, df = 13821, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.5868088 0.6031782
## sample estimates:
## mean of x
## 0.5949935

t.test (person_data$WKL_label=="Within the past 12 months")

##
##  One Sample t-test
##
## data:  person_data$WKL_label == "Within the past 12 months"
## t = 71.459, df = 13821, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.2623870 0.2771876
## sample estimates:
## mean of x
## 0.2697873

t.test (person_data$WKL_label=="1-5 years ago")

##
##  One Sample t-test
##
## data:  person_data$WKL_label == "1-5 years ago"
## t = 46.487, df = 13821, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.1295177 0.1409207
## sample estimates:
## mean of x
## 0.1352192

table (person_data$WKL_label)/sum (table(person_data$WKL_label))
```

```
##
##                    1-5 years ago Over 5 years ago or never worked
##                        0.1352192                        0.5949935
##        Within the past 12 months
##                        0.2697873
```

Independent living difficulty

```
t.test (person_data$DOUT_label=="Yes")

##
##  One Sample t-test
##
## data:  person_data$DOUT_label == "Yes"
## t = 52.457, df = 13821, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.1598353 0.1722440
## sample estimates:
## mean of x
## 0.1660396

t.test (person_data$DOUT_label=="No")

##
##  One Sample t-test
##
## data:  person_data$DOUT_label == "No"
## t = 263.47, df = 13821, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.8277560 0.8401647
## sample estimates:
## mean of x
## 0.8339604

table (person_data$DOUT_label)/sum (table(person_data$DOUT_label))

##
##        No       Yes
## 0.8339604 0.1660396
```

Grandparent living with children

```
t.test (person_data$GCL_label=="Yes")

##
##  One Sample t-test
##
## data:  person_data$GCL_label == "Yes"
## t = 24.626, df = 13821, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
```

```
##  0.03868868 0.04538019
## sample estimates:
##  mean of x
## 0.04203444
```

```r
t.test (person_data$GCL_label=="No")
```

```
##
##  One Sample t-test
##
## data:  person_data$GCL_label == "No"
## t = 561.23, df = 13821, p-value < 2.2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.9546198 0.9613113
## sample estimates:
## mean of x
## 0.9579656
```

```r
table (person_data$GCL_label)/sum (table(person_data$GCL_label))
```

```
##
##         No        Yes
## 0.95796556 0.04203444
```

Displaying the Distribution for All Categorical Variables

This is the end of my R Markdown file. The following table displaying the distribution for all categorical variables was created in Microsoft Word.

		Sample proportion (%)	95-percent confidence interval (%)
Sex	Female	56.2	55.4 to 57.0
	Male	43.8	43.0 to 44.6
Race	White	91.3	90.8 to 91.7
	Black or African American	3.7	3.4 to 4.0
	Some Other Race	1.1	0.9 to 1.3
	Two or More Races	0.1	0.1 to 0.1
	Asian alone	3.1	2.8 to 3.4
	American Indian	0.1	0.1 to 0.2
	American Indian and Alaska Native*	0.0	0.0 to 0.0
Marital status	Married	55.5	54.6 to 56.3
	Widowed	20.7	20.0 to 21.4
	Never married	9.2	8.8 to 9.7
	Divorced	13.5	12.9 to 14.1
	Separated	1.1	0.9 to 1.3
Housing tenure	N/A	7.1	6.7 to 7.5
	Owned	42.2	41.4 to 43.0
	Rented	16.2	15.6 to 16.8
	Owned with mortgage	33.1	32.3 to 33.9
	Occupied without rent	1.4	1.2 to 1.6
Employment status	Not in labor force	77.2	76.5 to 77.9
	Employed, at work	21.5	20.9 to 22.2
	Employed, with a job but not at work	0.6	0.5 to 0.7
	Unemployed	0.6	0.5 to 0.8
When last worked	Over 5 years ago or never worked	59.5	58.7 to 60.3
	Within the past 12 months	27.0	26.2 to 27.7
	1-5 years ago	13.5	13.0 to 14.1
Independent living difficulty	Yes	16.6	16.0 to 17.2
	No	83.4	82.7 to 84.0
Living with grandchildren	Yes	4.2	3.9 to 4.5
	No	95.8	95.5 to 96.1

**Original variable name is: "American Indian and Alaska Native tribes specified; or American Indian or Alaska Native, not specified and no other races"*