Statistical Analysis of Household Types and Incomes

Nick Hepler

February 20, 2015

1 Project Summary

This source data set contains a sample of actual responses to the American Community Survey Public Use Microdata Sample (PUMS) from residents in New York State. Each observation represents a response from a single housing unit and the sample data represents responses from approximately one percent of the United States population.

Utilizing this data set, a comparative analysis of household income was conducted based upon households with no parent present in which the household was headed by a grandparent or not headed by a grandparent.

2 Methods

The source data was obtained through the American Community Survey Public Use Microdata Sample (PUMS) furnished by the United States Census Bureau, American Community Survey Office. The original data can be found on the Census FTP Site.

The R programming language was employed to perform the statistical analysis to include measures of central tendency and dispersion. In order to compare the two data samples, the data was subset using the NPP variable in the source data which denotes if a household with no parent present is headed by a grandparent. Missing data in the FINCP variable denoting annual household income was ignored, but negative income was included in the calculations. Calculations were rounded to the nearest hundredth.

3 Data Considerations

The raw data set contains 92,810 observations. Of these observations, the NPP variable was missing in 18,659 observations. Based on the rationale for this report, these observations were excluded leaving 74,151 observations remaining.

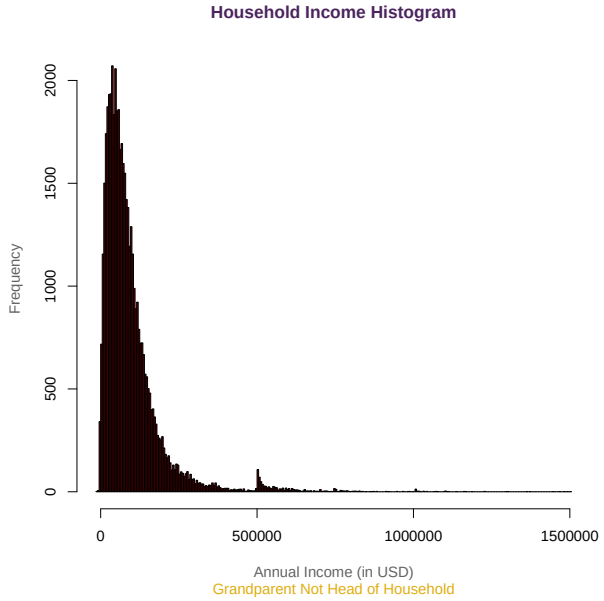
Of the 74,151 observations remaining, 25,732 had a missing value for the FINCP variable leaving 48,419 observations available for analysis. Of these observations, 48,156 observations were for households not headed by a grandparent with no parent present and the 263 remaining observations were associated with households headed by a grandparent with no parent present.

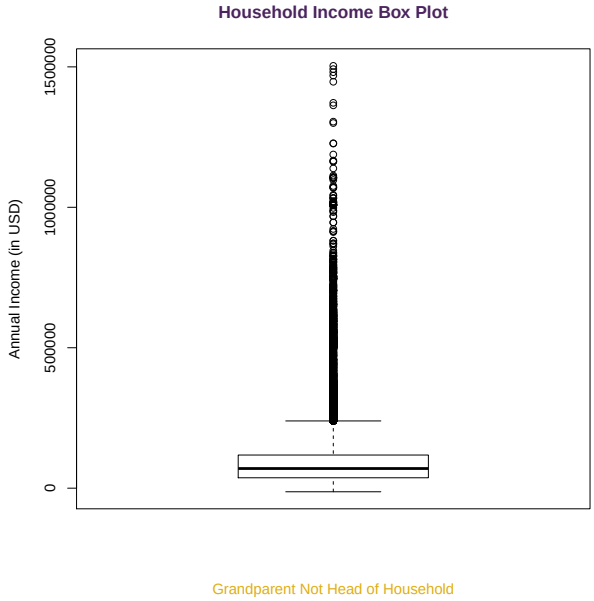
4 Summary Statistics

4.1 Not Grandparent Headed Household

Measuring the central tendency of the data for households not grandparent headed, we find the average (n = 48,156) annual household income reported was $95,153.96. The median household income was $70,000.00. The minimum reported income was $-12,800.00 and the maximum reported income was $1,503,600.00. The range of the dispersion for income is $1,516,400.00.

Measuring the variability of the sample data for this population we find the interquartile range of household income was $80,990.00. The standard deviation of household income was $101,243.48.

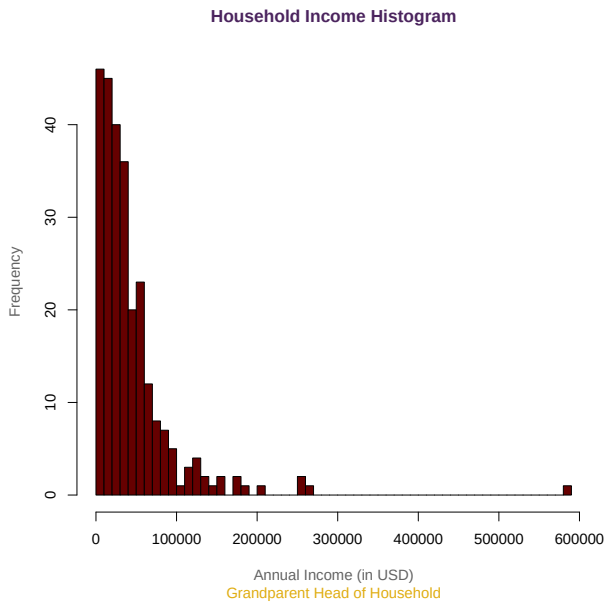


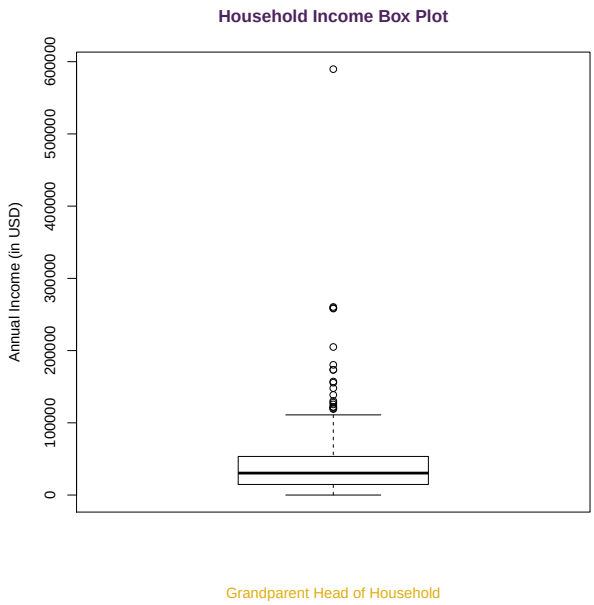


4.2 Grandparent Headed Household

Measuring the central tendency of the data for grandparent headed households, we find the average (n = 263) annual household income reported was $44,123.11. The median household income was $30,400.00 The minimum reported income was $0 and the maximum reported income was $589,700.00.

Measuring the variability of the sample data for this population we find the interquartile range of household income was $38,790.00. The standard deviation of household income was $54,516.05.


Household Income Histogram
Frequency
0
10
20
30
40
0
100000
200000
300000
400000
500000
600000
Annual Income (in USD)
Grandparent Head of Household


Household Income Box Plot
Annual Income (in USD)
0
100000
200000
300000
400000
500000
600000
Grandparent Head of Household

A Source Code

The following code can be run using `source("acs_pums_analysis.R")` from R. This source code downloads the raw data file from the source.

```
# acs_pums_analysis.R
# Statistical analysis of household types and incomes.
#
# Copyright (C) 2015  Nick Hepler
#
#  Version 0.9.0
#
# This program is free software; you can redistribute it and/or modify
# it under the terms of the GNU General Public License as published by
# the Free Software Foundation; either version 2 of the License, or
# (at your option) any later version.

# This program is distributed in the hope that it will be useful,
# but WITHOUT ANY WARRANTY; without even the implied warranty of
# MERCHANTABILITY or FITNESS FOR A PARTICULAR PURPOSE.  See the
# GNU General Public License for more details.

# You should have received a copy of the GNU General Public License along
# with this program; if not, write to the Free Software Foundation, Inc.,
# 51 Franklin Street, Fifth Floor, Boston, MA 02110-1301 USA.

require(dplyr)

# Download raw data and read into global envrionment.
temp <- tempfile()
download.file("http://www2.census.gov/acs2012_1yr/pums/csv_hny.zip",temp)
raw <- read.csv(unz(temp, "ss12hny.csv"))


#-------------------------------------------------------------#
# Data Management                                             #
# Converts the raw data into a clean data set suitable    #
# for statistical analysis                                    #
#-------------------------------------------------------------#

# Create a transitional data set, rename variables to lower case, and select
# only the fincp & npp variables for analysis
transition <- raw
names(transition) <- tolower(names(transition))
transition <- select(transition, fincp, npp)

# Convert npp variable to a factor, rename the variables.
transition$household <- NA
transition$household[transition$npp == 0] <-
```

```
  "Not a grandparent headed household with no parent present"
transition$household[transition$npp == 1] <-
  "Grandparent headed household with no parent present"
transition <- select(transition, -(npp))
transition$household <- factor(transition$household, levels=c(
  "Not a grandparent headed household with no parent present",
  "Grandparent headed household with no parent present"))
names(transition)[1] <- "income"
transition$income <- as.numeric(transition$income)

# Subset the samples
nghh <- filter(transition,
  household == "Not a grandparent headed household with no parent present")
ghh <- filter(transition,
  household == "Grandparent headed household with no parent present")

# Identify missing values
missing.household <- sum(is.na(transition$household))
missing.income <- sum(is.na(transition$income))
missing.income.nghh <- sum(is.na(nghh$income))
missing.income.ghh <- sum(is.na(ghh$income))

# Calculate number of observations with valid values.
valid.income <- length(transition$income) - missing.income
valid.household <- length(transition$household) - missing.household
valid.income.nghh <- length(nghh$income) - missing.income.nghh
valid.income.ghh <- length(ghh$income) - missing.income.ghh
valid.income.total <- valid.income.nghh + valid.income.ghh

# Group data using dplyr by household type.
by_household <- group_by(transition, household)

#-----------------------------------------------------------#
# Statistics                                                #
# Converts the raw data into a clean data set suitable    #
# for statistical analysis                                  #
#-----------------------------------------------------------#

# Mean/Median for each sample.
summarise(by_household, mean(income, na.rm=TRUE))
summarise(by_household, median(income, na.rm=TRUE))

# Min/Max for each sample.
summarise(by_household, min(income, na.rm=TRUE))
summarise(by_household, max(income, na.rm=TRUE))

# Calculate range for each sample.
diff(range(nghh$income, na.rm=TRUE))
```

```
diff(range(ghh$income, na.rm=TRUE))

# Calculate interquartile range for each sample.
quantile.nghh <- quantile(nghh$income, na.rm=TRUE)
quantile.ghh <- quantile(ghh$income, na.rm=TRUE)
quantile.nghh[4] - quantile.nghh[2]
quantile.ghh[4] - quantile.ghh[2]


# Calculate standard deviation for each sample.
summarise(by_household, sd(income, na.rm=TRUE))

# Create histogram for each sample.
par(cex.lab=.90)
options(scipen=999)

pdf("./report/hist_nghh.pdf")
hist(nghh$income,
  breaks="FD",
  col="#660000",
  xlab="Annual Income (in USD)", col.lab="#636363",
  main="Household Income Histogram", col.main="#4B245E",
  sub="Grandparent Not Head of Household", col.sub="#E0AD12")
dev.off()

pdf("./report/hist_ghh.pdf")
hist(ghh$income,
  breaks="FD",
  col="#660000",
  xlab="Annual Income (in USD)", col.lab="#636363",
  main="Household Income Histogram", col.main="#4B245E",
  sub="Grandparent Head of Household", col.sub="#E0AD12")
dev.off()

pdf("./report/box_nghh.pdf")
boxplot(nghh$income,
  sub="Grandparent Not Head of Household", col.sub="#E0AD12",
  ylab="Annual Income (in USD)",
  main="Household Income Box Plot", col.main="#4B245E")
dev.off()

pdf("./report/box_ghh.pdf")
boxplot(ghh$income,
  sub="Grandparent Head of Household", col.sub="#E0AD12",
  ylab="Annual Income (in USD)",
  main="Household Income Box Plot", col.main="#4B245E")
dev.off()
```

```
par(cex.lab=1)
options(scipen=0)

# Clean up global envrionment.
rm(list=ls())

# Remove temporary file
unlink(temp)
```