Estimation of Average Age and Income Level with Simple Random and Stratified Samlping Methods

Wenxuan Zan [*1], Maggie Ruan [†2], Taniya Cai [‡3], Kevin Yu [§4], and Yimin You [¶5]

Department of Statistics, University of British Columbia
STAT 344: Sample Surveys
Prof. WU, LANG

November 14, 2022

Contribtutions: The group scheduled avaliable time slots and met inperson to do the entire project. After we came up with possible dataset and parameters of interest together, the group decided to split up; with two people, Maggie and Kevin working on one Simple Random Sampling and others, Wenxuan, Taniya and Yimin on Stratified Sampling. Then we cross-checked others' work and give sufficient feedback.

[*1] Group leader. ID: 61336194
[†2] ID: 35003185
[‡3] ID: 95979696
[§4] ID: 18210468
[¶5] ID: 89663447

1 Part I

1.1 Introduction

The current project is interested in estimating key metrics that reflect a country's socioeconomic status. To this end, the **proportion of individuals with annual income greater than 50-thousand dollars per year (denoted as p)** and the **average age (denoted as $\bar{y}$)** of the population is chosen as the parameter of interest. The income level of an individual is coded as a binary variable (income level greater than 50 thousand dollars per year versus income level less or equal to 50 thousand dollars per year) and is chosen to reflect the distribution of wealth within a country. The average age of the population is chosen for it yields insights into the population's age composition. Both metrics provide information on the current demographic and economic composition of society as well as hints at future trends.

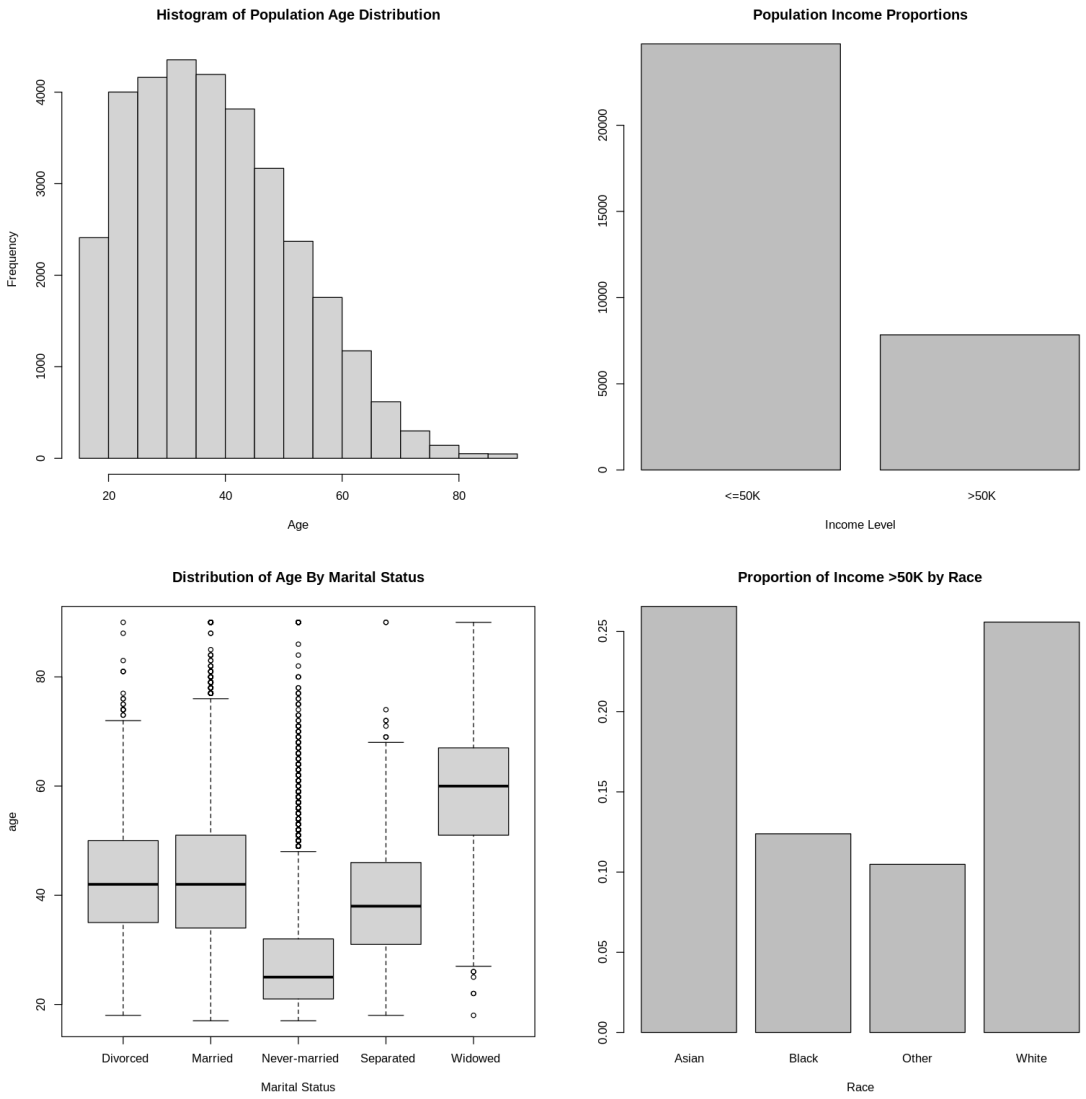
The dataset used in this project is extracted and prepared by Barry Becker based on a 1994 US census database which is publicly available on the UCI Machine Learning Repository. This dataset is composed of 32,561 entries containing many demographic variables such as respondents' age, race, marital status, income level, etc.

The current project utilized the available dataset as the population and obtained estimates of the parameters of interest (average age of individuals and proportion of individuals with income greater than 50K per year) by taking samples from the dataset using simple random sampling and stratified sampling techniques. Throughout the analysis, we will compare the performance of the two sampling techniques and discuss the advantages and disadvantages of both.

1.2 Preliminary Data Analysis

The population means of our two response variables are listed below, and we will be using this to check the accuracy of our two sampling methods later. The code can be found in the appendix.

	Value
Age Population Mean ($\bar{y}_{pop}$)	38.5816
Income Population Proportion (p)	0.2408



The histogram for the population distribution of age is right-skewed with the majority of the population ages being between 20 years old and 50 years old with a mean of around 38.5 years old. If we divide the population by people's marital status we see that there exists a noticeable between stratum differences in age, particularly noticeable between `Never-married`, `Separated`, and `Widowed` sub-populations.

Looking at the bar plot for the population proportion, approximately one-quarter of the population makes more than 50K a year. When we divide the population by people's race, we observe a race difference in the proportion of people earning more than 50-thousand dollars per year. This difference is particularly noticeable when comparing `Asian` and `White` to `Black` and `Other` racial group.

1.3 Analysis and Comparison

1.3.1 Sample Size Determination

Once we load the data into R, we can see that our population size is N = 32561. From here, we can decide on a sample size. We calculate our sample size by using the following formulas:

- $n_0 = \frac{1.96^2 s^2_{guess}}{\delta^2}$, where s^2_{guess} is the guessed variance, and δ is a desired margin of error.
- $n = \frac{n_0}{1+\frac{n_0}{N}}$, where N is the population size. This formula is used to take the finite population correct factor into consideration.

For our continuous response, age, we do not have a most conservative guess of the variance therefore we decide to base our sample size on the calculation for our binary response variable. For our binary data, we assume $\hat{p}$ = 0.5, which provides us with a most conservative guess of the variance and standard error, being 0.25 or 0.5 respectively.

The desired level of margin of error was chosen to be 0.02 (2%). The resulting $n = 2236$ is also a nice and realistic number to use for a data set like the current one. We will use the same sample size for our continuous data analysis and binary data analysis.

1.3.2 Simple Random Sampling

To implement simple random sampling, we treat our dataset as the population and randomly select $n = 2236$ entries without replacement from the dataset to form our sample. Since our sample size is less than 10% of the population size, we can assume that every entry within our dataset had an equal chance of being included in our sample.

To carry out the sampling procedure, we use `sample.int()` function in `R` to randomly draw 2236 indices from 32561 total indices, then use entries with the selected index to form our sample.

Binary Population Recall that one of our response variables is the proportion of people that have an income greater than 50K. Sampling from our population, we obtain a sample proportion of 0.2415 with a standard error of $SE[\hat{p}] = \sqrt{(1-\frac{n}{N})\frac{\hat{p}(1-\hat{p})}{n}} = 0.00873$. The 95% confidence interval is constructed using $\hat{p} \pm 1.96 \times SE[\hat{p}]$, which is [0.22438, 0.25862].

From this, we are 95% confident that the true proportion of people with an income greater than 50K is between 0.22438 and 0.25862.

	Value
Sample Proportion	0.2415
Sample SE	0.00873
Lower CI	0.22438
Upper CI	0.25862

Continous Population For our continuous parameter, we use the sample mean age $\bar{y_s} = \frac{\sum_{i=1}^{n} y_i}{n}$ to estimate the population mean age. The standard error of sample mean age is calculated from

$SE[\bar{y}_s] = \sqrt{(1 - \frac{n}{N})\frac{{s_s}^2}{n}}$, where s_s^2 is the sample variance of individuals' age. As the sample size is large enough, we can utilize the central limit theorem and the asymptotic 95% confidence interval for mean age of of individuals can be constructed by $\bar{y}_s \pm 1.96 \times SE[\bar{y}_s]$

	Value
Sample Mean	38.6556
Sample SE	0.2847
Lower CI	38.0975
Upper CI	39.2137

According to the output, the sample mean is 38.6556 with a standard error of 0.2847. From this, we conclude with 95% confidence that the true population mean age is between 38.0975 and 39.2137.

1.3.3 Stratified Sampling

For stratified sampling method, we first use auxiliary information available to us (the individuals' race and marital status) to divide the population of interest into homogeneous strata. For our binary population, the stratification is based on individuals' `race`. This is because during the exploratory data analysis we noted a between race difference in the proportion of individual with income greater than 50K per year. For this reason, we choose race as our stratification variable in order to maximize between-stratum variation. For our continuous population, we choose the individuals' `marital status` as our base for stratification. This is guided by our intuition as we think people generally move through stages of life, from single to get married, in a similar pace. Therefore, people who have the same marital status should be similar in their age as well.

We listed our assumptions made for the stratified sampling method and our general procedure for binary and continuous analysis.

- Assumptions:

  1. We assume the sampling cost for each stratum is the same.

  2. We assume the variance of each stratum is similar.

  3. We assume each stratum is independent of the other.

- Procedures:

  1. We calculate n_h for each stratum based on stratum size (N_h) such that $n_h = n\frac{N_h}{N}$, where n is the total sample size, and N is the population size.

  2. We divide the population into h strata according to race or marital status and randomly select n_h individuals without replacement from each stratum.

  3. We find the mean and variance of age and proportion of individuals with income greater than 50K for each stratum and combine the estimates to form one stratified estimates $\bar{y}_{str,s}$,$p_{str,s}$ of the population parameters.

Binary Population For the binary population, we divide the population into 4 strata according to individual's race. The four strata are `Asian`, `Black`, `Other`, and `White`.

The stratified estimate for the binary variable is computed as $\hat{p}_{str} = \sum_{h=1}^{H} \frac{N_h}{N} \hat{p}_h$ where $\hat{p}_h$ is the sample proportion of individuals in the h^{th} stratum with income greater than 50K.

The stratified estimate SE of sample proportion is computed as $SE[\hat{p}_{str}] = \sqrt{\hat{V}ar[p_{str}]}$, where $\hat{V}ar[p_{str}] = \sum_{h=1}^{H} (\frac{N_h}{N})^2 (1 - \frac{n_h}{N_h}) \frac{\hat{p}_h(1-\hat{p}_h)}{n_h}$.

The 95% confidence interval is computed as $\hat{p}_{str} \pm 1.96 \times SE[\hat{p}_{str}]$

	Value
Sample Proportion	0.24064
Sample SE	0.00867
Lower CI	0.22364
Upper CI	0.25764

Looking at the above result, the stratified estimate of sample proportion is 0.24064 with a sample SE of 0.00867. We conclude from the above result, with 95% confidence that the true population proportion is between 0.22364 and 0.25764.

Continuous Population For the continuous response, `age`, we divided the population into 5 strata according to individuals marital status, the strata are `Divorced`, `Married`, `Never Married`, `Separated`, `Widowed`.

- The estimate for the mean age is $\bar{y}_{str} = \sum_{h=1}^{H} \frac{N_h}{N} \bar{y}_{s,h}$, when $\bar{y}_{s,h}$ is the sample mean age of h^{th} stratum.
- The standard error for this estimate is $SE_{\bar{y}_{str}} = \sqrt{\sum_{h=1}^{H} (\frac{N_h}{N})^2 SE^2[\bar{y}_{s,h}]}$, where $SE[\bar{y}_{s,h}] = \sqrt{(1 - \frac{n_h}{N_h}) \times \frac{S^2_{s,h}}{n_h}}$ with $S^2_{s,h}$ being the sample variance of the h^{th} stratum.
- The 95% confidence interval for mean age is $\bar{y}_{str} \pm 1.96 \times SE[\bar{y}_{str}]$

	Value
Sample Mean	38.6169
Sample SE	0.2285
Lower CI	38.1690
Upper CI	39.0648

From our result, the estimate for mean age is 38.6169. The standard error for this estimate is 0.2285, Since our sample size sufficiently large (n = 2236), by the central limit theorem, the distribution of our sample mean will follow a normally distributed closely. We conclude that we are 95% confident that the true mean age is between 38.1690 and 39.0648.

1.4 Discussion

Overall, the 95 % confidence intervals constructed from the simple random sampling (SRS) and stratified sampling method captured the true population parameters of interest. We note that it would be infeasible in the real-world setting to compare whether the confidence interval captures the true parameter or not, therefore we focus our comparison of the performance of the two methods on the width of the confidence interval.

To this end, the stratified sampling method yields narrower confidence intervals for both the continuous (age) and binary parameters (proportion of individuals with income greater than 50K) compared to SRS. This is expected because only within-stratum variation contributes to the variance for stratified estimator whereas the variance for simple random estimator has contribution from both between-stratum and within-stratum variation.

In terms of the magnitude of reduction in SE of the estimates, the reduction is most noticeable for the continuous response as the SE of the estimate for mean age reduced to 0.2285 years for the stratified estimate compared to 0.2847 years of the SRS estimate. This is likely because the stratification process based on people's marital status was able to create homogeneous strata in age. Therefore we observe a meaningful reduction in SE of the stratified estimate of mean age compared to the SRS estimate. However, the reduction in SE for the sample proportion is not as meaningful. The SE of the estimated proportion reduced to 0.00867 (SE of the stratified estimate) from 0.00873 (SE of the SRS estimate). This relatively trivial reduction in SE is mainly due to our choice of the variable used to create strata. For the binary response, we used individuals' race to divide the population into strata. This stratification process didn't result in strata that are as different as the ones we have for the continuous response. As a result, the SE of the stratified estimate for the binary response didn't improve much compared to the SRS estimate.

While the stratified sampling method yields an estimate with lower SE, we must take note of the practicality of both techniques. The stratified sampling method requires significant preparation and information about the population before a sample can be taken, such as how to create distinct strata and information about the population and stratum size. Compared to the simple random sample which requires fewer preparations and information about the population. With all the information and preparation needed for a stratified sample, the cost and time required grow immensely which is why a simple random sample is more favoured when it comes to practicality.

One limitation of the project lies in how the strata are selected. We noted that when strata are not distinct enough, then the SE of the stratified estimate would not improve much compared to the SRS estimate such as the case for our binary response discussed above. To address this limitation, we would need to compare the between-stratum variation for multiple ways of creating strata and pick the one that maximizes the between-stratum variation. However, one challenge that arises from this approach is that we would need to know significant information about the population which may or may not be available to us. In addition, we would like to note that this dataset is somewhat dated and the choice of stratum is not optimal in this project. For this reason, we think the general procedure of the current project with improvement can be generalized to a larger population, but not the numeric result.

1.5 Conclusion

In this project, we have conducted two separate sampling techniques, simple random and stratified sampling, to estimate the parameters of interest, the average age of the population and the proportion of people with income greater than 50K. For our conclusion, we will report the estimates obtained using the stratified sampling method because its estimates have a smaller standard error.

Our stratified estimate of the mean population age is 38.62 years of age with a standard error of 0.2285 years of age. We conclude with 95 % confidence that the true population mean age is between 38.17 years and 39.06 years of age. We found that about 24.06 % of individuals within our sample have an annual income greater than 50K with a standard error of 0.867 % and a margin of error of $\pm$ 1.70 %. We conclude with 95 % confidence that between 22.36 % to 25.76 % of individuals within the population have an annual income greater than 50K.

2 Part II

The article by Perlman and Wu concerns the emergence and proliferation of "New Tests" in the field of statistics. The term "New Tests" is used by the authors to describe many of the newly emerged statistical tests with promising properties but often reach conclusions that go against scientific intuition. The proponents of the "New Tests" hold the notion that having a test with the best statistical properties is of paramount importance compared to scientific intuition. The authors of the article refute such a statement and hold a different view on the role of intuition in statistics. As the authors put it, scientific intuition is extremely important in scientific studies in general therefore tests that reach conclusions against scientific intuition are scientifically inappropriate tests. This article reinstated the importance of having good intuition in guiding scientific inquiry in general while reminding us to remain skeptical of statistically sounding yet unintuitive tests.

2.1 Appendix

[1]:
```r
# Data preparation
library(tidyverse)
library(GGally)
adult <- read.csv("adult.data", header = FALSE)
column_names <- c("age",
                  "workclass",
                  "final_weight",
                  "education",
                  "education_num",
                  "marital_status",
                  "occupation",
                  "relationship",
                  "race",
                  "sex",
                  "capital_gain",
                  "capital_loss",
                  "hours_per_week",
                  "native_country",
                  "income")
adult <- setNames(adult, column_names)
adult <- adult %>% mutate_if(is.character, as.factor)
adult <- adult %>%
    mutate(income = as.factor(income))

n = 2236
N = nrow(adult) ## population size

level <- as.character(levels(as.factor(adult$marital_status)))
status <- as.character(adult$marital_status)
for (i in 1:length(status)) {
    ifelse (status[i] == level[2] | status[i] == level[3] | status[i] ==␣
 ↪level[4],
            status[i] <- " Married",
            status[i] <- status[i])
}
adult$status <- as.factor(status)

options(repr.plot.width = 13, repr.plot.height = 13)
par(mfrow=c(2,2))
hist(adult$age,
     main = "Histogram of Population Age Distribution",
     xlab = "Age")
barplot(table(adult$income),
        main = "Population Income Proportions",
        xlab = "Income Level")
```

```r
boxplot(age ~ status,
        data = adult,
        main = "Distribution of Age By Marital Status",
        xlab = "Marital Status")
barplot(income_stratify_data$p,
        main = "Proportion of Income >50K by Race",
        xlab = "Race",
        names.arg = c("Asian", "Black", "Other", "White"))



level <- as.character(levels(as.factor(adult$race)))
races <- as.character(adult$race)
for (i in 1:length(races)) {
    if (races[i] == level[1] | races[i] == level[4]) {
        races[i] <- " Other"
        }
}
adult$race <- as.factor(races)
```

2.2 Simple random sample

2.2.1 Continous Population

[2]:
```r
set.seed(344)

SRS.index <- sample.int(N, n, replace = FALSE)
adultsample <- adult[SRS.index, ]

ybar <- mean(adultsample$age)
s_sd <- sd(adultsample$age)
se_ybar <- sqrt((1-n/N)*(s_sd^2/n))
```

2.2.2 Binary Population

[3]:
```r
set.seed(344)
SRS.index <- sample.int(N, n, replace = FALSE)
adultsample <- adult[SRS.index, ]
income <- adultsample %>%
    group_by(income) %>%
    summarize(n = n())

p_hat <- income[2,2]/n
p_hat = pull(p_hat)
phat_se = sqrt((1-n/N)*p_hat*(1-p_hat)/n)
#round(c(p_hat - 1.96 * phat_se, p_hat + 1.96 * phat_se),5)
```

2.3 Stratified Sampling

2.3.1 Continuous Parameter Estimation

[4]:
```r
## Get proportional stratum sizes
str_continuous <- adult %>%
    group_by(status) %>%
    summarise(N_h = n(), stratum_prop = N_h/N) %>%
    mutate(n_h = round(stratum_prop*n))
#str_continuous
```

[5]:
```r
## Stratified Samples
marital_levels <- levels(adult$status)
nh <- str_continuous$n_h

Divorced_sample <- adult %>%
    filter(status == marital_levels[1]) %>%
    pull(age) %>%
    sample(nh[1])

Married_sample <- adult %>%
    filter(status == marital_levels[2]) %>%
    pull(age) %>%
    sample(nh[2])

NeverMarried_sample <- adult %>%
    filter(status == marital_levels[3]) %>%
    pull(age) %>%
    sample(nh[3])

Separated_sample <- adult %>%
    filter(status == marital_levels[4]) %>%
    pull(age) %>%
    sample(nh[4])

Widowed_sample <- adult %>%
    filter(status == marital_levels[5]) %>%
    pull(age) %>%
    sample(nh[5])
```

[6]:
```
## Estmitate mean age
stratum_prop <- str_continuous$stratum_prop
Divorced_est <- (stratum_prop[1])*mean(Divorced_sample)
Married_est <- (stratum_prop[2])*mean(Married_sample)
NeverMarried_est <- (stratum_prop[3])*mean(NeverMarried_sample)
Separated_est <- (stratum_prop[4])*mean(Separated_sample)
Widowed_est <- (stratum_prop[5])*mean(Widowed_sample)

hat_ybar_str <-␣
 ↪sum(c(Divorced_est,Married_est,NeverMarried_est,Separated_est,Widowed_est))

## estimate SE
Divorced_var <- var(Divorced_sample)

Married_var <- var(Married_sample)

NeverMarried_var <- var(NeverMarried_sample)

Separated_var <- var(Separated_sample)

Widowed_var <- var(Widowed_sample)

vars <- c(Divorced_var,Married_var,NeverMarried_var,Separated_var,Widowed_var)

N_h <- str_continuous$N_h
Divorced_var <- ((stratum_prop[1])^2) * (1-(nh[1]/N_h[1]))*(vars[1]/nh[1])
Married_var <- ((stratum_prop[2])^2) * (1-(nh[2]/N_h[2]))*(vars[2]/nh[2])
NeverMarried_var <- ((stratum_prop[3])^2) * (1-(nh[3]/N_h[3]))*(vars[3]/nh[3])
Separated_var <- ((stratum_prop[4])^2) * (1-(nh[4]/N_h[4]))*(vars[4]/nh[4])
Widowed_var <- ((stratum_prop[5])^2) * (1-(nh[5]/N_h[5]))*(vars[5]/nh[5])

yhatse <-␣
 ↪sqrt(sum(c(Divorced_var,Married_var,NeverMarried_var,Separated_var,Widowed_var)))

result_data <- data.frame(Estimate = hat_ybar_str, SE = yhatse, true_parameter =␣
 ↪mean(adult$age))
#result_data
```

[7]:
```
phat.str.me <-1.96*yhatse
phat.str.confint <-c(hat_ybar_str-phat.str.me,hat_ybar_str+phat.str.me)
```

2.3.2 Binary Parameter Estimation

[8]:
```
level <- as.character(levels(as.factor(adult$race)))
```

[9]:
```
stratum_n <-adult %>%
    group_by(race) %>%
    summarise(stratum_size = n()) %>%
    pull(stratum_size)
```

[10]:
```
income_n_by_stratum <- adult %>%
    group_by(income, race) %>%
    summarise(num_greater_than_50K = n(),
              .groups = "drop")
stratum_size <- c(rep(stratum_n,2))
income_n_by_stratum$stratum_size <- stratum_size
income_stratify_data <- income_n_by_stratum %>%
            mutate(p = num_greater_than_50K/stratum_size,
                   allocation_prop = stratum_size/N,
                   n_h = round(n*allocation_prop)) %>%
            slice(5:8)
```

[11]:
```
binary_stratify_data <- adult %>% select(race, income)
level <- levels(as.factor(binary_stratify_data$race))
Asian <- binary_stratify_data %>% filter(race == level[1]) %>% pull(income)
Black <- binary_stratify_data %>% filter(race == level[2]) %>% pull(income)
Other <- binary_stratify_data %>% filter(race == level[3]) %>% pull(income)
White <- binary_stratify_data %>% filter(race == level[4]) %>% pull(income)
```

[12]:
```
set.seed(123456)
n_h <- income_stratify_data$n_h
asian_sample <- sample(as.character(Asian), n_h[1])
black_sample <- sample(as.character(Black), n_h[2])
other_sample <- sample(as.character(Other), n_h[3])
white_sample <- sample(as.character(White), n_h[4])
```

[14]:
```
asian_prop <- mean(asian_sample == " >50K")
black_prop <- mean(black_sample == " >50K")
other_prop <- mean(other_sample == " >50K")
white_prop <- mean(white_sample == " >50K")
props <- c(asian_prop, black_prop, other_prop, white_prop)
subpopulation_prop <- income_stratify_data$allocation_prop
prop_est <- sum(props*subpopulation_prop)
```

[16]:
```
N_h <- income_stratify_data$stratum_size
FPC = 1 - n_h/N_h
asian_var = FPC[1]*asian_prop*(1-asian_prop)/n_h[1]
black_var = FPC[2]*black_prop*(1-black_prop)/n_h[2]
other_var = FPC[3]*other_prop*(1-other_prop)/n_h[3]
white_var = FPC[4]*white_prop*(1-white_prop)/n_h[4]
prop_var = c(asian_var, black_var, other_var, white_var)
pooled_se = sqrt(sum(prop_var*(subpopulation_prop)^2))
```

[17]:
```
binary_results <- data.frame(name = c("Proportion Estimate", "Proportion SE",␣
↪"Lower Bound", "Upper Bound"),
          value = c(prop_est,round(pooled_se,5),prop_est - 1.
↪96*pooled_se,prop_est + 1.96*pooled_se))
```

[18]:
```
#data visualization
options(repr.plot.width = 13, repr.plot.height = 13)
par(mfrow=c(2,2))
hist(adult$age,
     main = "Histogram of Population Age Distribution",
     xlab = "Age")
barplot(table(adult$income),
        main = "Population Income Proportions",
        xlab = "Income Level")
boxplot(age ~ status,
        data = adult,
        main = "Distribution of Age By Marital Status",
        xlab = "Marital Status")
barplot(income_stratify_data$p,
        main = "Proportion of Income >50K by Race",
        xlab = "Race",
        names.arg = c("Asian", "Black", "Other", "White"))
```