Stat 138: Introduction to Sampling Designs
Problem Set 5

Anne Christine Amores

May 2, 2025

IPUMS

(a) Select an unequal-probability sample of 10 psus, with probability proportional to number of persons. Take a subsample of 20 persons in each of the selected psus.

```
# Loading the necessary packages
library(readxl)
library(survey)
```

```
# Importing the dataset
ipums <- read_excel("ipums.xlsx", col_names = FALSE)
```

```
# Renaming columns

colnames(ipums) <- c("stratum", "psu", "inctot", "age", "sex", "race",
                     "hispanic", "marstat", "ownershg", "yrsusa", "school",
                     "educrec", "labforce", "occ", "sei", "classwk")
head(ipums)
```

```
## # A tibble: 6 x 16
##   stratum   psu inctot   age   sex  race hispanic marstat ownershg yrsusa
##     <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>    <dbl>   <dbl>    <dbl>  <dbl>
## 1       1     1   4105    18     1     2        0       5        0      0
## 2       1     1   7795    20     1     1        0       5        2      0
## 3       1     1  16985    24     1     1        0       1        1      0
## 4       1     1   7045    21     1     1        0       1        2      0
## 5       1     1   2955    23     1     1        0       5        2      0
## 6       1     1      0    17     1     1        0       5        1      0
## # i 6 more variables: school <dbl>, educrec <dbl>, labforce <dbl>, occ <dbl>,
## #   sei <dbl>, classwk <dbl>
```

```
# Getting PSU sizes
psu_sizes <- ipums %>%
  group_by(psu) %>%
  summarise(M_i = n(), .groups = "drop") # no. of people in each PSU

# Selecting 10 PSUs via PPSWOR using Brewer's Method
set.seed(138)

pik  <- inclusionprobabilities(psu_sizes$M_i, 10)
```

```r
selected_psu_indices <- UPbrewer(pik)

psu_sample <- psu_sizes[selected_psu_indices == 1, ] %>%
  mutate(pi_h = pik[selected_psu_indices == 1])

# Selecting 20 persons per selected PSU via SRSWOR
sampled_people <- ipums %>%
  filter(psu %in% psu_sample$psu) %>%
    group_by(psu) %>%
      slice_sample(n = 20) %>%
        ungroup()

print(sampled_people)
```

```
## # A tibble: 200 x 16
##    stratum   psu inctot   age   sex  race hispanic marstat ownershg yrsusa
##      <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl>    <dbl>   <dbl>    <dbl>  <dbl>
##  1       1     3   6470    71     1     1        0       4        1      0
##  2       1     3   6005    75     2     1        0       4        1      0
##  3       1     3  15840    44     1     1        0       1        1      0
##  4       1     3      0    15     1     1        0       5        1      0
##  5       1     3   5910    62     2     1        0       1        2      0
##  6       1     3      0    55     2     1        0       1        1      0
##  7       1     3   2110    32     2     2        0       5        2      0
##  8       1     3   8005    46     2     1        0       1        1      0
##  9       1     3   6060    29     2     2        0       1        2      0
## 10       1     3  21005    39     1     2        0       3        1      0
## # i 190 more rows
## # i 6 more variables: school <dbl>, educrec <dbl>, labforce <dbl>, occ <dbl>,
## #   sei <dbl>, classwk <dbl>
```

The table above gives a preview of our final sample of 200 people.

(b) Using the sample you selected, estimate the population mean and total of *inctot* and give the standard errors of your estimates.

```r
# Computing weights
sampled_people <- sampled_people %>%
  left_join(psu_sample, by = "psu") %>%
  mutate(weight = (1 / pi_h) * (M_i / 20)) # final weight: stage 1 * stage 2

# Defining survey design
unequalprobdesign <- svydesign(
  id = ~psu,
  weights = ~weight,
  data = sampled_people)

# Estimating population mean and standard error
inctot_mean <- svymean(~inctot, unequalprobdesign)

# Estimating population total and standard error
inctot_total <- svytotal(~inctot, unequalprobdesign)
```

```
inctot_mean
```

```
##          mean     SE
## inctot 7807.2 625.94
```

```
inctot_total
```

```
##            total       SE
## inctot 417380719 33463183
```

Thus, given this sample, **our estimate for the population mean is 7,807.2 and its standard error is 625.94. Our estimate for the population total is 417,380,719 and its standard error is 33,463,183.**