STAT 311 - Programming Assignment 1

Data Analyst

2025-04-16

Assignment Overview

This assignment focuses on creating different types of samples from the county dataset in the usdata package: 1. Simple Random Sample (SRS) 2. Stratified Sample by education level 3. Cluster Sample by state

Setup

First, we load the necessary libraries and prepare the dataset:

```
# You will be using the "usdata" package's "county" data
#  for this assignment.
# You will need to run:
#  install.packages("usdata")
# once to install this library - do not include the
#  install.packages command in your submission
#  as it can cause gradescope to fail

## ## ## ## ## DO NOT MODIFY BELOW ## ## ## ## ##
library(usdata)
```

```
## Warning: package 'usdata' was built under R version 4.4.3
```

```
county<-as.data.frame(county)
county<-county[,1:14]
county<-county[rowSums(is.na(county))==0,]
# The set.seed command will ensure your results are consistent
#  each time you run the "source" command
set.seed(311)
## ## ## ## ## DO NOT MODIFY ABOVE ## ## ## ## ##
```

Let's examine the structure of our dataset:

```
dim(county)
```

```
## [1] 3135   14
```

```
head(county, 3)
```

```
##              name   state pop2000 pop2010 pop2017 pop_change poverty
## 1 Autauga County Alabama   43671   54571   55504       1.48    13.7
## 2 Baldwin County Alabama  140415  182265  212628       9.19    11.8
## 3 Barbour County Alabama   29038   27457   25270      -6.22    27.2
##   homeownership multi_unit unemployment_rate metro   median_edu
## 1          77.5        7.2              3.86   yes some_college
## 2          76.7       22.6              3.99   yes some_college
## 3          68.0       11.1              5.90    no   hs_diploma
##   per_capita_income median_hh_income
## 1          27841.70            55317
## 2          27779.85            52562
## 3          17891.73            33368
```

Part 1: Simple Random Sample

Create a simple random sample of n=250 individual counties from all counties in the US.

```
# Treating the "county" dataset as the population of US counties
#  create the dataframe "my.SRS" that represents a simple
#  random sample of n=250 individual counties from all counties in the US.

my.SRS <- county[sample(nrow(county), 250), ]

# Check the dimensions of our sample
dim(my.SRS)
```

```
## [1] 250  14
```

Part 2: Stratified Sample

Create a stratified sample of individual counties based on education level (median_edu).

Let's first check the distribution of education levels in our dataset:

```
table(county$median_edu)
```

```
##
##     below_hs   hs_diploma some_college    bachelors
##            2         1397         1691           45
```

Now we'll create our stratified sample with the specified sample sizes: - 1 county from “below_hs” - 140 from “hs_diploma” - 170 from “some_college” - 4 from “bachelors”

```
# Create empty dataframe to store the stratified sample
my.Stratified <- data.frame()

# Get indices for each stratum
below_hs_indices <- which(county$median_edu == "below_hs")
hs_diploma_indices <- which(county$median_edu == "hs_diploma")
some_college_indices <- which(county$median_edu == "some_college")
bachelors_indices <- which(county$median_edu == "bachelors")

# Sample from each stratum
sample_below_hs <- county[sample(below_hs_indices, 1), ]
sample_hs_diploma <- county[sample(hs_diploma_indices, 140), ]
sample_some_college <- county[sample(some_college_indices, 170), ]
sample_bachelors <- county[sample(bachelors_indices, 4), ]

# Combine all strata samples
my.Stratified <- rbind(sample_below_hs, sample_hs_diploma, sample_some_college, sample_bachelors)

# Verify the distribution in our stratified sample
table(my.Stratified$median_edu)
```

```
##
##     below_hs   hs_diploma some_college    bachelors
##            1          140          170            4
```

Part 3: Cluster Sample

Create a cluster sample of counties, clustered by state, from 5 randomly selected states:

```
# Get all unique states
all_states <- unique(county$state)

# Randomly select 5 states (clusters)
selected_states <- sample(all_states, 5)

# Get all counties from the selected states
my.Clustered <- county[county$state %in% selected_states, ]

# Verify we have exactly 5 states in our sample
unique(my.Clustered$state)
```

```
## [1] Delaware     Kansas       Oregon       Pennsylvania Utah
## 51 Levels: Alabama Alaska Arizona Arkansas California Colorado ... Wyoming
```

```
# Check how many counties we have in our cluster sample
dim(my.Clustered)
```

```
## [1] 240  14
```

Summary

This assignment demonstrates three different sampling methods:

1. **Simple Random Sampling**: Each county has an equal probability of being selected.
2. **Stratified Sampling**: The population is divided into non-overlapping groups (strata) based on education level, and samples are taken from each stratum.
3. **Cluster Sampling**: The population is divided into groups (clusters) based on state, and all counties from randomly selected states are included.

Each sampling method has its advantages depending on the population structure and the goals of the study.