STA2050 - Midsem

Chesia Anyika

2024-02-21

# Libraries

#for combination computation  
library(gtools)  
  
#for golfsrs.dat  
library(SDAResources)  
  
#for histogram plotting  
library(tidyverse)

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.0 ✔ readr 2.1.4  
## ✔ forcats 1.0.0 ✔ stringr 1.5.0  
## ✔ ggplot2 3.4.1 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.2 ✔ tidyr 1.3.0  
## ✔ purrr 1.0.1   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the ]8;;http://conflicted.r-lib.org/conflicted package]8;; to force all conflicts to become errors

library(sampling)  
library(survey)

## Loading required package: grid  
## Loading required package: Matrix  
##   
## Attaching package: 'Matrix'  
##   
## The following objects are masked from 'package:tidyr':  
##   
## expand, pack, unpack  
##   
## Loading required package: survival  
##   
## Attaching package: 'survival'  
##   
## The following objects are masked from 'package:sampling':  
##   
## cluster, strata  
##   
##   
## Attaching package: 'survey'  
##   
## The following object is masked from 'package:graphics':  
##   
## dotchart

# Question 1

### Part a)

For the survey given below describe the target population, sampling frame, sampling unit, and observation unit. Discuss any possible sources of selection bias or inaccuracy of responses **(4 marks)**

*A survey is conducted to find the average weight of cows in a region. A list of all farms is available for the region, and 50 farms are selected at random. Then the weight of each cow at the 50 selected farms is recorded.*

The **Target Population** is all cows in that specific region.

The **Sampling Frame** is the list of all the farms in the specific region.

The **Sampling Unit** is one farm.

The **Observation Unit** is one cow.

**Bias**

There is no discernible reason for bias in the information given above.

### Part b)

(i). List all possible simple random samples of size that can be selected from the population . (**3marks)**

To compute this I used the combination() function from the gtools library as follows:

#define parameters  
v1 <- c(0,1,2,3,4)  
N = length(v1)  
n = 2  
  
#get the samples of n=2 of v1 #library(gtools)  
samples1 <- combinations(n = N, r = n, v = v1, repeats.allowed = FALSE)  
  
#View the samples  
samples1

## [,1] [,2]  
## [1,] 0 1  
## [2,] 0 2  
## [3,] 0 3  
## [4,] 0 4  
## [5,] 1 2  
## [6,] 1 3  
## [7,] 1 4  
## [8,] 2 3  
## [9,] 2 4  
## [10,] 3 4

1. Calculate for the population and for the sample **(5marks)**

First I computed the population variance, using the var() function on the population vector

Var.pop <- var(v1)  
cat('Population variance: ', Var.pop)

## Population variance: 2.5

I then obtained the mean estimates of the samples as per the formula

var.est = (1 - n/N) \* Var.pop/n  
cat('variance estimate:', var.est)

## variance estimate: 0.75

### Part c)

The Internet site www.golfcourse.com listed 14,938 golf courses by state. It gave a variety of information about each course, including greens fees, course rating, par for the course, and facilities. Data from an SRS of 120 of the golf courses is in file golfsrs.dat.

**i.** Display the data in a histogram for the weekday greens fees for nine holes of golf. How would you describe the shape of the data? **(3marks)**

First I imported the data into r

data("golfsrs")  
  
names(golfsrs)

## [1] "RN" "state" "holes" "type" "yearblt" "wkday18" "wkday9"   
## [8] "wkend18" "wkend9" "backtee" "rating" "par" "cart18" "cart9"   
## [15] "caddy" "pro"

The variable associated with **weekday greens fees for nine holes** is wkday9. First I checked the datatype of the variable.

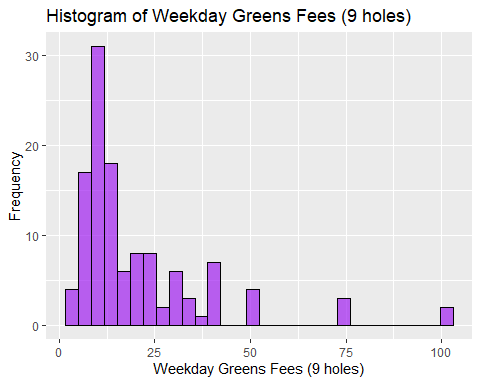
class(golfsrs$wkday9)

## [1] "numeric"

I then plotted the histogram using geom\_histogram() from the ggplot2 library.

#plot histogram #library(tidyverse)  
ggplot(golfsrs, aes(x = wkday9)) +  
 geom\_histogram(fill = "purple", color = "black", alpha = 0.7) +  
 labs(x = "Weekday Greens Fees (9 holes)", y = "Frequency") +  
 ggtitle("Histogram of Weekday Greens Fees (9 holes)")

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



The histogram is right skewed, implying that the mode of the variable is less than the median and mean. Thus weekday greens fees for 9 holes tend to be on the lower end.

**ii.**Find the average weekday greens fee to play nine holes of golf, and give the SE for your estimate. **(5marks)**

# Question 2

### Part a)

Suppose that a city has 90,000 dwelling units, of which 35,000 are houses, 45,000 are apartments, and 10,000 are condominiums. You believe that the mean electricity usage is about twice as much for houses as for apartments or condominiums, and that the standard deviation is proportional to the mean so that S1 = 2S2 = 2S3.

1. How would you allocate a stratified sample of 900 observations if you wanted to estimate the mean electricity consumption for all households in the city? **(4 marks)**

To do this i created a data-frame with all the observations, and calculated the allocations in the column nh.

# define parameters  
stratum <- c("Houses", "Apartments", "Condos")  
Nh <- c(35000, 45000, 10000)   
Rh <- c(2,1,1)   
  
#NhRh for each stratum  
NhRh <-Nh\*Rh  
  
#nh for each stratum  
nh <- round(900\*NhRh/sum(NhRh))  
  
#create sums vector  
Sum <- c('Sum', sum(Nh), sum(Rh), sum(NhRh), sum(nh))  
  
#create a data frame  
dwelling.df <- data.frame(Stratum=stratum,Nh=Nh,Rh=Rh,NhRh=NhRh,nh=nh)  
  
#row bind the sums  
dwelling.df <- rbind(dwelling.df, Sum)  
  
#Print the data frame  
dwelling.df

## Stratum Nh Rh NhRh nh  
## 1 Houses 35000 2 70000 504  
## 2 Apartments 45000 1 45000 324  
## 3 Condos 10000 1 10000 72  
## 4 Sum 90000 4 125000 900

The nh column represents the allocations, which are:

$$
Houses = 504\\Apartments = 324\\Condos = 72
$$

1. Explain an advantage of using a stratified sample rather than a simple random sample: **(2 marks)**

Ensuring diversity of your sample : Stratified random sampling allows one to be protected from the possibility of obtaining a really bad sample, by ensuring each stratum is represented in the sample.

### Part b)

The data below  are the number of pepper standards for the randomly selected villages in each of the three strata of Karen Zone

| **Stratum** | Number of villages selected from the stratum | Total number of villages in the stratum |
| --- | --- | --- |
| **1** | 41, 116, 19, 15, 144, 159, 212, 57, 28, 76, 119 |  |
| **2** | 39, 70, 38, 37, 161, 38, 27, 119, 36, 128, 30, 208 |  |
| **3** | 252, 386, 192, 296, 115, 159, 120 |  |

Estimate the total number of pepper standards along with its standard error in Karen Zone. **(10 marks)**

First I created vectors with the above information

#create vectors of numbers of villages sampled  
v1 <- c(41, 116, 19, 15, 144, 159, 212, 57, 28, 76, 119)  
v2 <- c(39, 70, 38, 37, 161, 38, 27, 119, 36, 128, 30, 208)  
v3 <- c(252, 386, 192, 296, 115, 159, 120)  
  
#check lengths  
l1 =length(v1)  
l2 = length(v2)  
l3 = length(v3)  
  
l1

## [1] 11

l2

## [1] 12

l3

## [1] 7

Using the lengths above, I created a stratum variable and created a dataframe with this and the vectors

#define variables  
stratum <- c(rep(1, l1), rep(2, l2), rep(3, l3))  
Villages <- c(v1, v2, v3)  
  
#create dataframe  
df <- data.frame(stratum, Villages)  
  
#view dataframe  
df

## stratum Villages  
## 1 1 41  
## 2 1 116  
## 3 1 19  
## 4 1 15  
## 5 1 144  
## 6 1 159  
## 7 1 212  
## 8 1 57  
## 9 1 28  
## 10 1 76  
## 11 1 119  
## 12 2 39  
## 13 2 70  
## 14 2 38  
## 15 2 37  
## 16 2 161  
## 17 2 38  
## 18 2 27  
## 19 2 119  
## 20 2 36  
## 21 2 128  
## 22 2 30  
## 23 2 208  
## 24 3 252  
## 25 3 386  
## 26 3 192  
## 27 3 296  
## 28 3 115  
## 29 3 159  
## 30 3 120

To get the total number of pepper standards, we use the formula

#get sample means   
y1 <- sum(v1)/l1  
y2 <- sum(v2)/l2  
y3 <- sum(v3)/l3  
  
#define parameter vectors  
Ni <- c(441, 405, 103)  
yi <- c(y1, y2, y3)  
  
#run formula  
T.est <- sum(Ni\*yi)  
  
#estimator of population total  
cat('Total Estimate:', T.est)

## Total Estimate: 93316.6

### Part c)

1. John is looking at the beats per minute of tracks in his music player. He has 1200 tracks. He decides to take a systematic sample of 25% of his tracks. Determine the tracks that should be chosen. Using linear systematic sampling with program R  to select n  **(4 marks)**