STA2050 - QUIZ 1

Chesia Anyika

ID: 665567

2024-02-07

# Libraries

library(gtools)

# Question 1

**1.      An SRS of size 30 is taken from a population of size 100. The sample values are given below, and in the data file srs30.dat. 8 ,5, 2, 6 ,6 ,3, 8, 6, 10, 7 ,15, 9, 15, 3, 5, 6, 7, 10, 14 3, 4 ,17, 10, 6, 14 ,12, 7, 8 ,12, 9, (page 9)**

**a)      What is the sampling weight for each unit in the sample? (2 marks)**

The sampling weights are equal because this is a simple random sample. It can be computed using the formula:

Where:

is the Population size

is the sample size

I computed this as follows:

N <- 100  
n <- 30   
  
sampling\_weights <- N/n  
  
cat('Sampling Weights: ', sampling\_weights)

## Sampling Weights: 3.333333

The sampling weight is .

**b)      Use the sampling weights to estimate the population total, t. (4marks)**

The formula for estimating population total is:

Where:

is the sampling weights

is the value

First I defined my y values in a vector

srs30 <- c(8 ,5, 2, 6 ,6 ,3, 8, 6, 10, 7 ,15, 9, 15, 3, 5, 6, 7, 10, 14, 3, 4 ,17, 10, 6, 14 ,12, 7, 8 ,12, 9)  
  
#view the vector  
srs30

## [1] 8 5 2 6 6 3 8 6 10 7 15 9 15 3 5 6 7 10 14 3 4 17 10 6 14  
## [26] 12 7 8 12 9

I then multiplied them by the sampling weight, and used the sum() function to get the sum of the product between each value and its sampling weight.

#get estimated population total  
estimated\_total <- sum(srs30 \* sampling\_weights)  
  
#View the result  
cat("Estimated Population Total:", estimated\_total, "\n")

## Estimated Population Total: 823.3333

The Estimated Population Total is .

**c)       Give a 95% CI for t. Does the fpc make a difference for this sample?(4marks)**

The Confidence interval is calculated as follows:

$$ CI = P.E \pm ME \\ ME = Zscore \* SE $$

Where:

is the Point Estimate

is the margin of Error

is the Standard Error

I first computed the Standard Error. To calculate the standard error, we use the Variance of t formula as follows:

Where:

is the population size

is the sample size

is the standard deviation of y

I computed this as follows:

#compute the standard deviation  
sy <- sd(srs30)  
  
#View the standard deviation  
cat('Sample Standard Deviation: ', sy)

## Sample Standard Deviation: 3.997269

#compute the variance of the total  
V\_t <- (N^2)\*(1 - (n/N))\*(((sy)^2)/n)  
  
#View the vatiance of the total  
cat('Variance of the total:', V\_t)

## Variance of the total: 3728.238

The Standard Error is calculated by getting the square root of , as follows:

I computed this as follows:

#compute the standard error  
SE\_t <- sqrt(V\_t)  
  
#view the standard error  
cat('Standard Error: ', SE\_t)

## Standard Error: 61.0593

I then computed the z\_score associated with a 95% confidence level

#define alpha, the critical value  
alpha <- 0.95  
  
#as the test is two tailed, get the diffeence of alpha from 1 and divide by a half  
alpha.neg <- 1-alpha  
alpha.half <- alpha.neg/2  
  
#define p   
p <- 1-alpha.half  
  
#get the zscore using qnorm() function  
zscore <- qnorm(p)   
  
#View the zscore  
cat("Z\_score: ", zscore)

## Z\_score: 1.959964

I then computed the Margin of Error by getting the product of the z\_score and Standard Error, as follows

#compute the margin of error  
ME\_t <- zscore \* SE\_t  
  
#View the margin of error  
cat('Margin of error: ', ME\_t)

## Margin of error: 119.674

I then calculated the Confidence interval by getting the sum and difference of the point estimate and the margin of error.

#compute confidence interval  
lower <- estimated\_total - ME\_t  
upper <- estimated\_total + ME\_t  
  
#view the confidence interval  
cat('Confidence interval : [', lower, ', ', upper, ']')

## Confidence interval : [ 703.6593 , 943.0074 ]

The 95% confidence interval of the total estimate is:

# Question 2

**2.       Suppose we have a population of 5 students enrolled for statistics course and a counsellor wants to find the average amount of time spent by each student in preparing for classes each week. The amount of time (in hours) each student spends per week is given by 7,3,6,10 and 4. If the counsellor takes a sample of three students WOR, obtain the sampling distribution of the sample mean. Compute the population mean and the standard error of the sampling distribution. (12marks)**

First I defined the values given

#Population   
N = 5  
  
#vector of times spent preparing for each student  
times <- c(7, 3, 6, 10, 4)

I then used the combinations() function to get the data-frame of all possible samples of sample size 3, from the vector of population size 5, without replacement

# Set the seed for random number generation  
set.seed(NULL)  
  
# Select an SRS of size n=3 from a population of size N=5 without replacement  
#library(gtools)  
samples1 <- combinations(n = 5, r = 3, v = times, repeats.allowed = FALSE)  
  
samples1

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

I then obtained the sample means of each sample by using the apply() function to apply the mean() function to each row of the data-frame I obtained above.

#apply to each row of samples  
means\_result <- apply(samples1, 1, mean)  
  
#add to the df  
samples1 <- as.data.frame(samples1)  
samples1$means <- means\_result  
  
samples1

## V1 V2 V3 means  
## 1 3 4 6 4.333333  
## 2 3 4 7 4.666667  
## 3 3 4 10 5.666667  
## 4 3 6 7 5.333333  
## 5 3 6 10 6.333333  
## 6 3 7 10 6.666667  
## 7 4 6 7 5.666667  
## 8 4 6 10 6.666667  
## 9 4 7 10 7.000000  
## 10 6 7 10 7.666667

#View the dataframe  
samples1

## V1 V2 V3 means  
## 1 3 4 6 4.333333  
## 2 3 4 7 4.666667  
## 3 3 4 10 5.666667  
## 4 3 6 7 5.333333  
## 5 3 6 10 6.333333  
## 6 3 7 10 6.666667  
## 7 4 6 7 5.666667  
## 8 4 6 10 6.666667  
## 9 4 7 10 7.000000  
## 10 6 7 10 7.666667

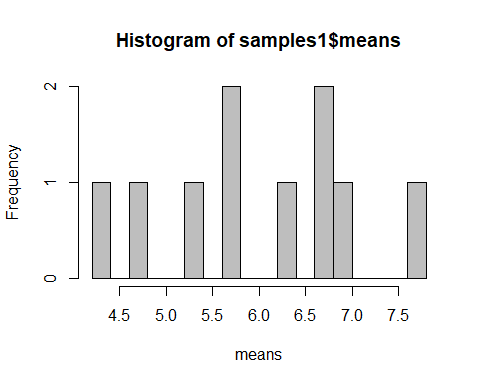
The sampling distribution of the sample means is as follows:

cat('The sampling distribution of the Sample means is: \n', means\_result)

## The sampling distribution of the Sample means is:   
## 4.333333 4.666667 5.666667 5.333333 6.333333 6.666667 5.666667 6.666667 7 7.666667

These means can be represented using a histogram

hist(samples1$means,breaks=20,col="gray",xlab="means")



I then calculated the population mean by using the mean() function on the times vector, which is the population of how much time students spend preparing each week.

#get the population mean  
popmean <- mean(times)  
  
#view the result  
cat('The Population mean is: ', popmean)

## The Population mean is: 6

I then obtained the Standard error as per the formula

#Step 1: calculate the variance  
var <- var(samples1$means)  
  
#define sample and population size  
n = 3  
N = 5  
  
#standard error  
SE <- sqrt(1 - ((n/N)\*(var/n)))  
  
#View the result  
cat('The standard error is: ', SE)

## The standard error is: 0.8819171

The computed standard error is:

# Question 3

**3. John took a stratified sample of New York City food stores. The sampling frame consisted of 1408 food stores with at least 4000 square feet of retail space. The population of stores was stratified into three strata using median household income within the zip code. The prices of a “market basket” of goods were determined for each store; the goal of the survey was to investigate whether prices differ among the three strata. John used the logarithm of total price for the basket as the response y. Results are given in the following table:**

| **Stratum h** |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Low Income** | 190 | 21 | 3.925 | 0.037 |
| **Middle Income** | 407 | 14 | 3.938 | 0.052 |
| **Upper Income** | 811 | 22 | 3.942 | 0.070 |

1. **The planned sample size was 30 in each stratum; this was not achieved because some stores went out of business while the data were being collected. What are the advantages and disadvantages of sampling the same number of stores in each stratum?  (4 marks)**

| Advantages | Disadvantages |
| --- | --- |
| **Lessens Bias** - having the same sample size per stratum lessens the bias in your data, as one stratum is not dominating the analysis as compared to another. | **Differences in variance of samples**: samples may differ in variance, thus having a single sample size for all of them may not yield accurate results. |
| **Allows for Uniformity** - having the same sample size allows for uniformity of interpretations in your analysis, as you do not have to define multiple sample sizes. |  |

b)      Estimate for these data and give a 95% CI. (4marks)

First I created a dataframe with all this information

#creating a data frame with the provided information  
data <- data.frame(  
 Stratum = c("Low Income", "Middle Income", "Upper Income"),  
 N\_h = c(190, 407, 811),  
 n\_h = c(21, 14, 22),  
 mean\_y\_h = c(3.925, 3.938, 3.942),  
 sd\_y\_h = c(0.037, 0.052, 0.070)  
)  
  
data

## Stratum N\_h n\_h mean\_y\_h sd\_y\_h  
## 1 Low Income 190 21 3.925 0.037  
## 2 Middle Income 407 14 3.938 0.052  
## 3 Upper Income 811 22 3.942 0.070

I then computed the y estimate as per the following formula

Where:

is the stratum size

is the population size

is the mean per stratum

I computed this as follows:

#define parameters  
N <- 1408  
  
#compute the y estimate  
y\_str <- sum((data$N\_h/N)\*data$mean\_y\_h)  
  
#View the y estimate  
cat('Y estimate: ', y\_str)

## Y estimate: 3.93855

The computed y estimate is :

To compute the confidence interval, I followed the formulas

$$ CI = PE \pm ME\\ ME = Zscore \* SE $$

Where:

is the point estimate ( )

is the margin of error

is the standard error

First I computed the standard error. For this I needed the variance of the estimate, which i computed as per the formula:

Where:

is the total size per stratum

is the population size

is the standard deviation per stratum

is the sample size per stratum

I computed this as follows

#compute the variance of the mean estimate  
V\_ystr <- sum(((data$N\_h/N)^2)\*((data$sd\_y\_h)^2)/data$n\_h)  
  
#view the variance of the mean estimate  
cat('Variance of the mean estimate: ', V\_ystr)

## Variance of the mean estimate: 9.121966e-05

The standard error of the variance estimate can be gotten as follows:

#compute the standard error  
SE\_ystr <- sqrt(V\_ystr)  
  
#show the standard error  
cat('Standard error: ', SE\_ystr)

## Standard error: 0.009550899

I then computed the Z score for the 95% confidence interval as follows:

#define alpha, the critical value  
alpha <- 0.95  
  
#as the test is two tailed, get the diffeence of alpha from 1 and divide by a half  
alpha.neg <- 1-alpha  
alpha.half <- alpha.neg/2  
  
#define p   
p <- 1-alpha.half  
  
#get the zscore using qnorm() function  
zscore <- qnorm(p)   
  
#View the zscore  
cat("Z\_score: ", zscore)

## Z\_score: 1.959964

I then computed the Margin of Error by multiplying the standard error and the Z score:

#compute the margin of error  
ME\_ystr <- zscore \* SE\_ystr  
  
#view the margin of error  
cat('margin of error: ', ME\_ystr)

## margin of error: 0.01871942

I then computed the confidence interval by getting the sum and difference of the Point estimate, and the margin of error

#compute lower and upper bounds  
lower <- y\_str - ME\_ystr  
upper <- y\_str + ME\_ystr  
  
#view the confidence interval   
cat('Confidence interval : [', lower, ', ', upper, ']')

## Confidence interval : [ 3.91983 , 3.957269 ]

The computed confidence interval is:

c)       Is there evidence that prices are different in the three strata? (2marks)

The narrow confidence interval suggests low sampling variability in the estimates of prices between the three strata. However, it’s crucial to consider that factors beyond sampling error, such as the inability to reach some stores and missing market basket items, may have influenced the results. These non-sampling errors could potentially bias the estimates or affect their generalisability to the population. Therefore, while the confidence interval provides some indication, it’s essential to acknowledge the presence of non-sampling errors when interpreting the results.