STA2050 - final exam

Chesia Anyika

2024-04-03

# Question 1

**(a)    A state department of agriculture desires to estimate the number of acres under corn plantation within the state. Suggest possible sampling units and frames. (2marks )**

Sampling unit: The sampling unit could be a single corn plantation.

Sampling frame: The sampling frame could be a list of all corn plantations within the state.

**(b)    How do we evaluate the goodness of an estimator? (2marks)**

We can evaluate the goodness of an estimator by taking into account the amount of bias in the chosen estimation method, by examining sampling methods used.

We can also take into account the accuracy of the metric in estimating the population parameter, by taking into account the standard error and the confidence interval.

# Question 2

## Libraries

#for general data manipulation and 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 3.5.0 ✔ 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

(a)    Foresters want to estimate the average age of trees in a stand. Determining age is cumbersome, because one needs to count the tree rings on a core taken from the tree. In general, though, the older the tree, the larger the diameter, and diameter is easy to measure. The foresters measure the diameter of all 1132 trees and find that the population mean equals 10.3. They then randomly select 20 trees for age measurement.

| Tree no | Diameter X | Age y | Tree no | Diameter X | Age y |
| --- | --- | --- | --- | --- | --- |
| 1 | 12 | 125 | 11 | 5.7 | 61 |
| 2 | 11.4 | 119 | 12 | 8 | 80 |
| 3 | 7.9 | 83 | 13 | 10.3 | 114 |
| 4 | 10.5 | 85 | 14 | 12 | 147 |
| 5 | 7.9 | 99 | 15 | 9.2 | 122 |
| 6 | 9 | 117 | 16 | 8.5 | 106 |
| 7 | 7.3 | 69 | 17 | 7.0 | 82 |
| 8 | 10.2 | 133 | 18 | 10.7 | 88 |
| 9 | 11.7 | 154 | 19 | 9.3 | 97 |
| 10 | 11.3 | 168 | 20 | 8.2 | 99 |
|  |  |  |  |  |  |

**a)       Draw a scatterplot of y vs. x.  (3 marks)**

First I defined the two vectors for diameter and age as follows:

#define the diameter vector  
diameterx <- c(12, 11.4, 7.9, 10.5, 7.9, 9, 7.3, 10.2, 11.7, 11.3, 5.7, 8, 10.3, 12, 9.2, 8.5, 7.0, 10.7, 9.3, 8.2)  
  
#check length - should be 20  
l1 <- length(diameterx)  
  
cat('Length of diameter vector:', l1)

## Length of diameter vector: 20

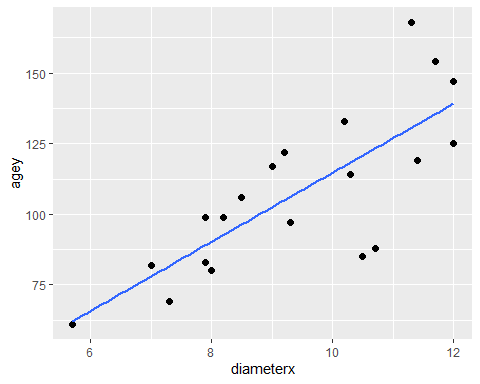
#define the diameter vector  
agey <- c(125, 119, 83, 85, 99, 117, 69, 133, 154, 168, 61, 80, 114, 147, 122, 106, 82, 88, 97, 99)  
  
#check length - should be 20  
l2 <- length(agey)  
  
cat('Length of age vector:', l2)

## Length of age vector: 20

I then plotted a scatterplot of age against diameter as follows:

#create a dataframe  
trees <- data.frame(diameterx, agey)  
   
#plot the scatterplot with a regression line  
ggplot(trees, aes(x=diameterx, y=agey)) +  
 geom\_point(size=2) +  
 geom\_smooth(method="lm", se=FALSE)

## `geom\_smooth()` using formula = 'y ~ x'



The scatter-plot shows an overall positive trend between the age and diameter variables, suggesting a linear relationship between the two. The regression line fitted shows this as well.

**b)      Estimate the population mean age of trees in the stand using ratio estimation and give an approximate standard error for your estimate.  (7marks)**

First I defined my parameters as follows

#define population parameters  
N = 1132  
mu = 10.3  
  
#define sample parameters  
n = 20  
x\_bar <- mean(diameterx)  
y\_bar <- mean(agey)

#compute the ratio estimate as follows  
y\_hat = y\_bar \* (mu / x\_bar)  
  
#view the result   
cat('Ratio estimate: ', y\_hat)

## Ratio estimate: 117.6204

compute the standard deviation

#sample standard deviation   
sd\_y = sd(agey)   
sd\_x = sd(diameterx)   
  
#sample covariance  
cov = cov(diameterx, agey)  
Ratio = y\_bar / mu  
  
MeanSquaredError = ((N-n)/(n\*N)) \* (sd\_y^2 + Ratio^2 \* sd\_x^2 - 2 \* Ratio \* cov)  
  
SE\_1 = sqrt(MeanSquaredError)  
  
#view the standard error  
cat('Standard Error: ', SE\_1)

## Standard Error: 4.038486

**c)        Repeat (b) using regression estimation. (7marks)**

#compute the regression estimate  
beta = cov / sd\_x^2  
y\_reg = y\_bar + beta \* (mu - x\_bar)  
  
cat('Regression estimate: ', y\_reg)

## Regression estimate: 118.3465

# Calculate the approximate standard error for regression estimate  
rho = cor(diameterx, agey) # correlation coefficient  
  
V\_reg = ((N-n)/(n\*N)) \* sd\_y^2 \* (1 - rho^2)  
SE\_2 = sqrt(V\_reg)  
  
cat('Standard Error: ', SE\_2)

## Standard Error: 3.971772

**d)      Label your estimates on your graph. How do they compare?  (3marks)**

# Question 3

## Libraries

#library with the spanish.dat dataset  
library(SDAResources)

##   
## Attaching package: 'SDAResources'

## The following object is masked from 'package:ggplot2':  
##   
## seals

#for the survey design  
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: 'survey'

## The following object is masked from 'package:graphics':  
##   
## dotchart

**a)       A language school owner takes an SRS of 10 of the 72 Introductory Spanish classes offered by the school. Each student in each of the sampled classes is given a vocabulary test and is also asked whether he or she is planning a trip to a Spanish-speaking country in the next year. The data are in file spanish.dat. CLUSTER**

**I.            Estimate the total number of students planning a trip to a Spanish-speaking country in the next year, and give a 95% CI. (6 marks)**

**II.            Estimate the mean vocabulary test score for Introductory Spanish students in the language school, and give a 95%  (6 marks)**

I computed both questions using survey design, as follows:

First I imported the spanish.dat data-set into r and examined it.

data("spanish")  
  
#view the data  
View(spanish)  
head(spanish)

## # A tibble: 6 × 3  
## class score trip  
## <dbl> <dbl> <dbl>  
## 1 34 57 0  
## 2 34 69 0  
## 3 34 61 0  
## 4 34 62 0  
## 5 34 42 0  
## 6 34 45 0

I then defined the sampling weights as follows:

#define sampling weight  
spanish$sampwt <- 72/10  
  
#print the data  
head(spanish)

## # A tibble: 6 × 4  
## class score trip sampwt  
## <dbl> <dbl> <dbl> <dbl>  
## 1 34 57 0 7.2  
## 2 34 69 0 7.2  
## 3 34 61 0 7.2  
## 4 34 62 0 7.2  
## 5 34 42 0 7.2  
## 6 34 45 0 7.2

I then defined the survey design, using the svydesign() function.

#survey design  
design <- svydesign(ids = ~class, weights = ~sampwt, data = spanish)  
  
#estimate the total and mean  
survey\_means <- svytotal(~score + trip, design)  
survey\_means

## total SE  
## score 94262.4 7301.4  
## trip 453.6 120.5

1. Answer(I): The total estimate of students planning a trip to a spanish speaking country is:

with a standard error ( ) of:

1. Answer(II): The mean estimate for vocabulary test scores is:

with a standard error ( ) of:

I then computed the Confidence Interval as per the following formula

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

First I computed the Z-score associated with a 95% confidence interval

#define alpha and p for a two-tailed test  
alpha <- 0.95  
alpha.neg <- 1 - alpha  
alpha.half <- alpha.neg/2  
  
p <- 1-alpha.half  
cat('p: ', p)

## p: 0.975

#perform z-test  
Z\_score <- qnorm(p)  
  
#Print the z\_score  
cat('Zscore: ', Z\_score)

## Zscore: 1.959964

As both standard errors for the total and mean estimates were computed above, I used them to compute their respective Margin of Errors as follows:

### CI for Total Estimate

#define standard of error  
SE\_t = 120.5  
  
#compute margin of error  
ME\_t <- SE\_t\*Z\_score  
  
#view the margin of error:   
cat('Margin of error:', ME\_t )

## Margin of error: 236.1757

#define the point estimate   
PE\_t <- 453.6  
  
#compute the upper and lower bounds of the confidence interval  
lower <- PE\_t - ME\_t  
upper <- PE\_t + ME\_t  
  
#View the confidence interval  
cat("Confidence interval for total estimate: [", lower, ',', upper, ']' )

## Confidence interval for total estimate: [ 217.4243 , 689.7757 ]

**Answer (I):** The confidence interval computed is:

### CI for mean estimate

#define standard of error  
SE\_t = 7301.4  
  
#compute margin of error  
ME\_t <- SE\_t\*Z\_score  
  
#view the margin of error:   
cat('Margin of error:', ME\_t )

## Margin of error: 14310.48

#define the point estimate   
PE\_t <- 94262.4  
  
#compute the upper and lower bounds of the confidence interval  
lower <- PE\_t - ME\_t  
upper <- PE\_t + ME\_t  
  
#View the confidence interval  
cat("Confidence interval for total estimate: [", lower, ',', upper, ']' )

## Confidence interval for total estimate: [ 79951.92 , 108572.9 ]

The confidence interval computed is:

**b)      An industry produces customized truck bodies in 20 plants scattered throughout the country. Only a small number of truck bodies are completed at each plant on any given day. For a particular day, it is desired to estimate the average worker-hours of work for items completed that day in the industry. Discuss the relative merits of cluster sampling (plants as clusters) versus stratified random sampling (plants as strata) for estimating the desired average. Which design do you recommend? (4marks )**