Practical 8

1940223 Pranav Gopalkrishna

2021-03-20

# Aim

Consider a dataset of your choice as a population. Give the estimates for population mean, population total and their confidence interval by choosing a variable of interest from the population using systematic sampling and write a report on it. Note: sample size should be appropriate enough to give the estimates.

# Data

setwd("~/Documents/Study/computerScience/programming/r/rPrograms/sampling/samplingTechniques")  
data = read.csv("./data/weatherAustralia.csv")  
head(data)

## Date Location MinTemp MaxTemp Rainfall Evaporation Sunshine WindGustDir  
## 1 2008-12-01 Albury 13.4 22.9 0.6 NA NA W  
## 2 2008-12-02 Albury 7.4 25.1 0.0 NA NA WNW  
## 3 2008-12-03 Albury 12.9 25.7 0.0 NA NA WSW  
## 4 2008-12-04 Albury 9.2 28.0 0.0 NA NA NE  
## 5 2008-12-05 Albury 17.5 32.3 1.0 NA NA W  
## 6 2008-12-06 Albury 14.6 29.7 0.2 NA NA WNW  
## WindGustSpeed WindDir9am WindDir3pm WindSpeed9am WindSpeed3pm Humidity9am  
## 1 44 W WNW 20 24 71  
## 2 44 NNW WSW 4 22 44  
## 3 46 W WSW 19 26 38  
## 4 24 SE E 11 9 45  
## 5 41 ENE NW 7 20 82  
## 6 56 W W 19 24 55  
## Humidity3pm Pressure9am Pressure3pm Cloud9am Cloud3pm Temp9am Temp3pm  
## 1 22 1007.7 1007.1 8 NA 16.9 21.8  
## 2 25 1010.6 1007.8 NA NA 17.2 24.3  
## 3 30 1007.6 1008.7 NA 2 21.0 23.2  
## 4 16 1017.6 1012.8 NA NA 18.1 26.5  
## 5 33 1010.8 1006.0 7 8 17.8 29.7  
## 6 23 1009.2 1005.4 NA NA 20.6 28.9  
## RainToday RainTomorrow  
## 1 No No  
## 2 No No  
## 3 No No  
## 4 No No  
## 5 No No  
## 6 No No

The data set contains information about daily weather observations accross Australia, taken for the past 10 years. The target variable is RainTomorrow, or the prediction of whether or not there will be rain tomorrow, and how much. I will be focussing on estimating the mean minimum temperature in Autralia.

# Population parameters

popl = c()  
popl\_size = length(data$MinTemp)  
for (i in c(1:popl\_size))  
 {  
 if(!is.na(data$MinTemp[i]))  
 {  
 popl = c(popl, data$MinTemp[i])  
 }  
}  
  
# Checking if NA values are still there  
sum(is.na(popl))

## [1] 0

# Resetting population size  
popl\_size = length(popl)  
# Population total and mean  
popl\_total = sum(popl)  
popl\_mean = popl\_total/popl\_size  
  
popl\_size / 25

## [1] 5759

# Sampling and sample statistics

Population size is divisible by 25. For our sample, we will pick the 14th element of every group of 20 elements of the population.

sample = c()  
sample\_size = popl\_size / 25 - 1  
for (i in c(0:sample\_size))  
{  
 sample = c(sample, popl[i \* 20 + 14])  
}  
  
# Resetting sample size  
sample\_size = length(sample)  
# Sample mean  
sample\_mean = sum(sample) / sample\_size  
# Estimated population total  
est\_popl\_total = sample\_mean \* popl\_size

Hence, the estimated mean and total of the population is given by

sample\_mean

## [1] 12.02171

est\_popl\_total

## [1] 1730825

Hence, the estimated mean minimum temperature in Australia is 12.02379°C, and the estimated total of minimum temperatures is 1731126°C.

Comparing to population mean and total

popl\_mean

## [1] 12.19403

popl\_total

## [1] 1755636

We can do this because sample mean is an unbiased estimator of population mean. This implies estimated total given by population size times sample mean is an unbiased estimator of population total. As we can see, the estimates are close to the actual values.

# Finding the confidence interval…

Finding the variance of sample mean

# V(y-bar\_r) = ((N - 1)/N) \* S^2 - ((n - 1)/n) \* S^2\_wsy  
# where S^2 is the population mean square, and S^2\_wsy is mean square within systematic sample.  
S2 = var(popl)  
S2\_wsy = var(sample)  
N = popl\_size  
n = sample\_size  
V\_ybar\_r = abs(((N - 1)/N) \* S2 - ((n - 1)/n) \* S2\_wsy)

Bound on error of estimate is

B = 2 \* sqrt(V\_ybar\_r)

Calculating the confidence interval for population mean estimate…

c(sample\_mean - B, sample\_mean + B)

## [1] 9.256596 14.786815

Hence, population mean’s estimate is likely to fall in this interval (75% likelihood, if population is not normal, 95% if population is normal) Calculating the confidence interval for population total estimate…

c(est\_popl\_total - N \* B, est\_popl\_total + N \* B)

## [1] 1332718 2128932

Hence, population total’s estimate is likely to fall in this interval (75% likelihood, if population is not normal, 95% if population is normal)