Mid-semester 4

Pranav, 1940223

2021-03-26

# Aim

Determine the required sample size for estimating the population proportion of the variable of your interest by assuming the bound on the error of estimation 0.05 and give the estimates for population proportion and population total with 95% confidence limits by drawing a sample using SRSWOR. Write a report on it.

# Data

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 the proportion of positive rain predictions for the next day.

setwd("~/Documents/Study/computerScience/programming/r/rAssignments")  
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

# Removing NA values from the population

# Making sure the population has no NA values...  
sum(is.na(data$RainTomorrow))

## [1] 3267

# Hence, the population has 3267 NA values.  
# To remove them, we use the following loop...  
popl = c()  
popl\_size = length(data$RainTomorrow)  
for (i in c(1:popl\_size))  
{  
 if(!is.na(data$RainTomorrow[i]))  
 {  
 popl = c(popl, data$RainTomorrow[i])  
 }  
}  
# Checking if NA values are still there...  
sum(is.na(popl))

## [1] 0

The population is now without NA values.

# Population parameters

# Setting population size...

N = length(popl)

# Population total and proportion...

*# Here, 1 => "No", 2 => "Yes"*

A = sum(popl == 2)  
print(A)

## [1] 31877

P = A / N  
print(P)

## [1] 0.2241812

# Population variance...  
V = P\*(1 - P)  
print(V)

## [1] 0.173924

Note that the population total is the number of population units that fulfill certain criteria. In this case, the criterion is that the prediction of tomorrow's rain should be "Yes", or 2. Here, the population total is 31877.

Population proportion is the population mean, i.e. the proportion of units in the population that fulfill the given criteria. Here, the population proportion is 0.2241812.

# Finding the minimum required sample

# Finding the required sample size for the given the bound on error of estimation...  
B = 0.05  
D = B^2 / 4  
n = N \* V / ((N - 1)\*D + V)  
n = round(n)  
print(n)

## [1] 278

Hence, we need a sample size of at least 278 to ensure that we get the bound on the error of estimation as 0.05. Note that the bound on the error of estimation is the bound placed on the difference between the estimated value and the parameter, or in this case, between the estimated population proportion and the actual population proportion.

# Sampling and estimation

Sample proportion is an unbiased estimator of population proportion. Also, population total can be estimated using sample proportion and population size. Hence, we draw a sample and use its statistics.

# Drawing a sample with the required size...  
sample = sample(popl, 278, replace = FALSE)

Finding sample proportion, hence estimated **population proportion**...

n = length(sample)  
a = sum(sample == 2)  
p = a / n  
print(p)

## [1] 0.1870504

Finding estimated **population total**...

N\_est = p \* N  
N\_est

## [1] 26597.25

The estimated population proportion is 0.1870504, and the estimated population proportion is 26597.25.

# Determining 95% confidence intervals

## Checking if sample follows normal distribution shapiro.test(sample)

##   
## Shapiro-Wilk normality test  
##   
## data: sample  
## W = 0.47478, p-value < 2.2e-16

p < 0.05, hence, considering the sample to be representative, we may say that the population does not follow normal distribution.

## Using Chebychev's inequality

Using Chebychev's inequality, we have that

*P((X - mu) ≥ k\*sigma) ≤ 1/k^2* ... (1)

where X is the variable, mu is its mean, sigma is its standard deviation, and k is the number of standard deviations from the mean.

Applying to estimated population proportion and the given bound on error of estimation, we get

*P((p - P) ≥ B) ≤ 1/k^2* ... (2)

We do this because we want to find k, that is, the number of standard deviations required so that the error of estimation i.e. p - P is under B i.e. the bound on error of estimation.

Comparing (1) and (2), we see that

*B = k \* sigma*, where sigma is the standard deviation of the variable.

But the variable is estimated population proportion, i.e. p.Hence, sigma = sqrt(V(p)

Hence, *B = k \* sqrt(V(p))*

Hence, *k = B / sqrt(V(p))*

Variance of p is given by

*V(p) = ((N - n) / N) \* S^2 / n*

where S^2 is the population mean squared.

## Finding the number of standard deviations from the mean required for 95% confidence interval

# S^2 = NP(1 - P) / (N - 1)  
S2 = N\*P\*(1 - P) / (N - 1)  
Vp = ((N - n) / N) \* S2 / n  
print(Vp)

## [1] 0.0006244071

# Finding k  
k = B / sqrt(Vp)  
print(k)

## [1] 2.000949

## The confidence intervals for estimates

### Population proportion

The 95% confidence interval for estimate of P i.e. **population proportion** is

c(p - k\*sqrt(Vp), p + k\*sqrt(Vp))

## [1] 0.1370504 0.2370504

Hence, we can say with 95% confidence that the estimate of the population proportion will lie between 0.1370504 and 0.2370504.

### Population total

Variance of estimated population total is given by population size times variance of proportion, i.e.

*V(N\_est) = N \* V(p)*

Similarly the 95% confidence interval for estimate of A i.e. **population total** is

c(N\_est - k\*N\*sqrt(Vp), N\_est + k\*N\*sqrt(Vp))

## [1] 19487.6 33706.9

Hence, we can say with 95% confidence that the estimate of the population total will lie between 19487.6 and 33706.9.