Sampling Populations

practical examples of sampling populations in veterinary medicine using R

NCSU: Nicolas Cardenas & Gustavo Machado

2021-10-31

Set some packages and data in R

Simple random sampling

We will a sample of the specified **N** from the elements. here a short example: * how Will give me a cookie ? my options are Felipe, Jason, Abby, Gustavo and Kelsey

I can run one time a random sample selectin just one

```
# set my options to get my cookie
N <- c("Felipe",
       "Jason",
       "Abby",
       "Gustavo",
       "Kelsey" )

# Calculate the sample
sample(N,                # total of the population
       1,                # one person
       replace = F)      # sampletin without replacement
```

```
## [1] "Abby"
```

you get back luck, lets gonna simulate 1000 times

```
# simple random sample repeated 1000 times
result <- sample(N,        # total of the population
          1000,            # one person
          replace = T)     # sample without replacement

# show a table with the results
sort(table(result))
```

```
## result
##    Abby Gustavo  Felipe   Jason  Kelsey
##     192     192     194     207     215
```

Sample Size Calculations in veterinary epidemiology

Here we will calculate the number of animals requited to estimate a prevalence of a disease in a specific population. For this example the expected prevalence for this area is **15%**. And, our question is How many cattle need to be sampled and tested using the **95%** of confidence interval? the total herds is a population `N` of 1000 animals.

```
size <- rsampcalc(N=1000, e=3, ci=95, p=0.15)
print(size)
```

```
## [1] 353
```

Stratified random sampling

For this example we are going to use the `Albania` dataset containing 2017 Albania election that is previously installed in R. First, we are going to explore in detail the variable `qarku` that means county or location

```
sort(table(albania$qarku))
```

```
##
##       Kukes Gjirokaster       Diber       Lezhe       Berat     Shkoder
##         173         235         259         263         305         421
##       Vlore      Durres       Korce     Elbasan        Fier      Tirane
##         447         460         463         547         591        1198
```

```
# Calulate the general sample size
size <- rsampcalc(nrow(albania), e=3, ci=95, p=0.15)
# stratify the data by the variable qarku
stratifiedsample <- ssamp(albania, size, qarku)
sort(table(stratifiedsample$qarku))
```

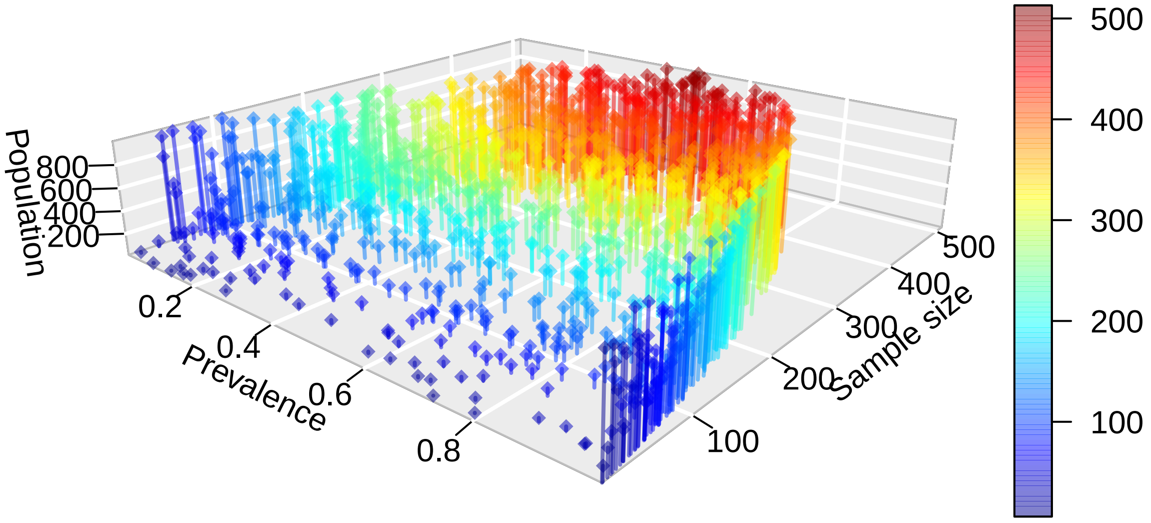
```
##
##       Kukes Gjirokaster       Diber       Lezhe       Berat     Shkoder
##          16          22          24          24          28          39
##       Vlore      Durres       Korce     Elbasan        Fier      Tirane
##          41          42          43          50          55         111
```

Relation among population size adn prevalence

In the next analysis we are going to simulate 1000 sampling designs. For this we are going to consider an initial prevalence of 1% and successively increase it to 100%, in this same way we are going to consider that populations from 10 animals to 100 animals. In the next plot the color reflects the size of the sample where warm colors represent larger sample sizes.

```
#set the number of samples
N <- 1000
myprevalence <- runif(n=N, min=0.01, max=1) # prevalence within 1 to 100%
mypopulation<- runif(n=N, min=10, max=1000) # population withn 10 to 1000
```

```
mysamplesize <- c()
for (i in 1:length(mypopulation)){
  aux <- rsampcalc(mypopulation[i],
                   e=3, ci=95,
                   p=myprevalence[i])
  mysamplesize <- rbind(mysamplesize, aux)
}
#plot the results
mydata <- tibble(myprevalence, mypopulation, mysamplesize= as.numeric(mysamplesize))
scatter3D(mydata$myprevalence,
          mydata$mysamplesize,
          mydata$mypopulation,
          bty = "g", pch = 18,
          lwd = 2, alpha = 0.5,
          expand =0.2,
          phi = 20,
          colvar = mysamplesize,
          ticktype = "detailed",
          type = "h",
          xlab = "Prevalence", ylab = "Sample size", zlab = "Population")
```



Cluster sampling

An aid project has distributed cook stoves in a single province in a resource-poor country. At the end of three years, the donors would like to know what proportion of households are still using their donated stove. A cross-sectional study is planned where villages in a province will be sampled and all households (approximately `75` per village) will be visited to determine if the donated stove is still in use. A pilot study of the prevalence of stove usage in five villages showed that `0.46`of householders were still using their stove and the intracluster correlation coefficient (ICC) for stove use within villages is in the order of `0.20`. If the donor wanted to be 95% confident that the survey estimate of stove usage was within`10%`of the true population value, how many villages (clusters) need to be sampled?

```
epi.ssclus1estb(b = 75,                 # the number of individual listing units in each cluster to be
                Py = 0.46,              # an estimate of the unknown population proportion is this case
                epsilon = 0.10,         # the maxi difference between the estimate and the unknown popu
                error = "relative",     # type of error to be used
                rho = 0.20,             # the intra-cluster correlation
                conf.level = 0.95)$n.psu# IC95%
```

```
## [1] 96
```

Challenge

Data for: Clinical Mastitis in Cows based on Udder Parameter using Internet of Things (IoT) from this study, each represent one animal thus we have a population of n = 1100.

First prepare the data to be analyze, here we will consider the results at `Day = 6`, and we will stratify by the variable `Address`. Then, calculate the number of animals requited to estimate a prevalence of *mastitis* a disease this population with a tolerable margin of error of `3`.For this exercise assume that your expected prevalence for this area **20%**. How many cattle need to be sampled and tested using interval confidence of **95%** ?

the next code filter the data by tyhe day = 6

```
#prepare data for analysis
clinical_mastitis_cows <- clinical_mastitis_cows %>% # indicates the  database
  filter(Day == max(Day))                            # filter by 6 day
```

references

Sample Size Estimation in Veterinary Epidemiologic Research Practical Issues in Calculating the Sample Size for Prevalence Studies