Sampling Populations

practical examples of sampling populations in veterinary medicine using R

NCSU: Nicolas Cardenas & Gustavo Machado

2022-10-30

Set some packages and data in R.

Simple random sampling.

We will take a sample of **N** from a list of target invidious:

- I wanted to sample from a list of friend and calculate who will provide me with my daily cookie?

My options are the following: Felipe, Arthur, Denilson, Jason, Abby, Gustavo and Kelsey.

I can run one time random sample selectin once.

```
# Options to get my cookie
N <- c("Felipe",
       "Jason",
       "Arthur",
       "Denilson",
        "Abby",
       "Gustavo",
       "Kelsey" )

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

```
## [1] "Denilson"
```

To better represent the sample drawing we will simulate this sampling for 1000 times

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

# Present into a table with the results.
sort(table(result))
```

```
## result
##   Felipe  Gustavo     Abby   Arthur Denilson    Jason   Kelsey
##      128      128      135      136      153      153      167
```

Sample size.

Here we will calculate the number of animals needed to estimate disease prevalence in a finite population. For this example the expected prevalence is **15%**. We want to know how the sample size in which a **95%** confidence interval is needed. We know that our total target population is `N` of 1000 animals.

```
size <- rsampcalc(N=1000,   # total of the population
                  e=5,      # tolerable margin of error this case 5%
                  ci=95,    # confidence interval of 95%
                  p=0.15)   # expected prevalence
print(size)
```

```
## [1] 164
```

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 to see how many records are in each region.

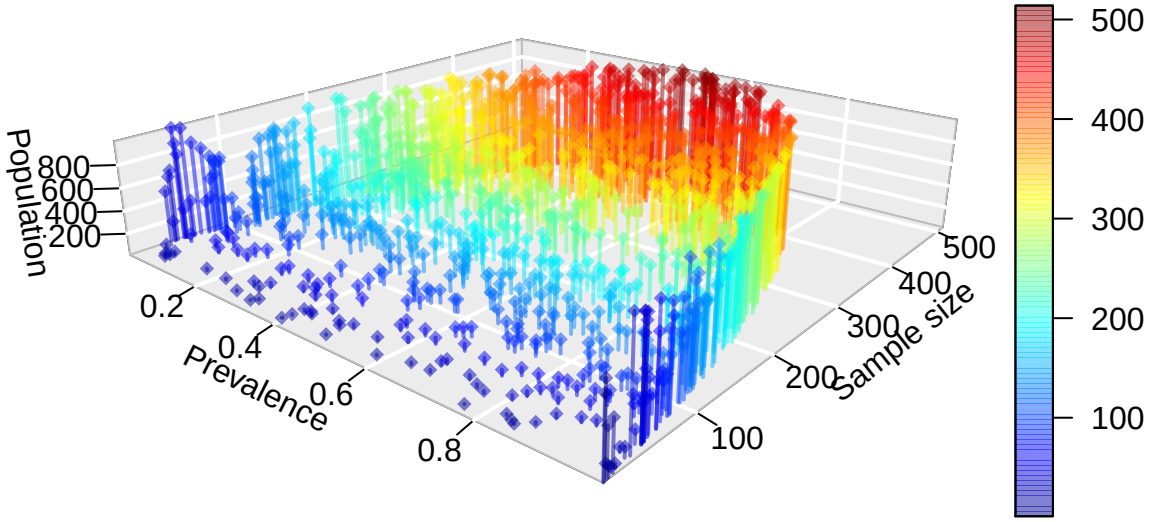
Relation between sample size and prevalence.

In the next analysis, we are going to simulate 1000 sampling designs. Here we are going to consider an initial prevalence of 1% while increasing the expected prevalence all the way to 100%, in this same way we are going to consider that population size from 10 animals to 100 animals. In the next plot, the color reflects the sample size where warm colors represent larger sample sizes.

```
# Set up the number of samples.
N <- 1000
myprevalence <- runif(n=N, min=0.01, max=1) # prevalence between 1 and 100%.
mypopulation<- runif(n=N, min=10, max=1000) # population within 10 and 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")
```



About the dataset

These data are a very small subset of a large mastitis dataset collected by Jens Agger and the Danish Cattle Organization. This dataset contains records from 14,357 test-day observations in 2,178 cows from 40 herds. Milk weights (production records) were collected approximately monthly, and only records from a single lactation for each cow were included in this dataset. Factors that may have affected the somatic cell count (SCC) were also recorded. The major objective of this study was to determine if the relationship between the somatic cell count and milk production varies for cows with different characteristics (age, breed, grazing or not etc).

varibales decription

variable	Description	Codes/units
herdid	herd id	
cowid	cow id	
test	approximate month of lactation	0 to 10
h_size	average herdsize	
c_heifer	parity of the cow	1 = heifer

variable	Description	Codes/units
t_season	season of test day	0 = multiparous 1 = jan-mar 2 = apr-jun 3 = jul-sep 4 = oct-dec
t_dim	days in milk on test-day	days
t_lnscc	log somatic cell count on test day	

```
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 overall sample size.
size <- rsampcalc(nrow(albania),  #  Total number of record
                  e=3,            # tolerable margin of error this case 3%
                  ci=95,          # confidence interval of 95%
                  p=0.5)          # anticipated response distribution

# Stratify the data by the variable 'qarku' which represent the region.
stratifiedsample <- ssamp(albania, size, qarku)
sort(table(stratifiedsample$qarku))
```

```
##
##       Kukes Gjirokaster       Diber       Lezhe       Berat     Shkoder
##          29          39          43          44          51          70
##       Vlore      Durres       Korce     Elbasan        Fier      Tirane
##          74          76          77          91          98         199
```