Sampling Populations

practical examples of sampling populations in veterinary medicine using R

NCSU: Nicolas Cardenas & Gustavo Machado

2021-11-01

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, Jason, Abby, Gustavo and Kelsey.

> I can run one time random sample selectin once.

```
# 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)      # sampling without replacement
```

```
## [1] "Felipe"
```

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
##    Abby  Kelsey Gustavo   Jason  Felipe
##     181     202     203     205     209
```

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=3,      # tolerable margin of error this case 3%
                  ci=95,    # confidence interval of 95%
                  p=0.15)   # expected prevalence
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 to see how many records are in each region.

```
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))
```

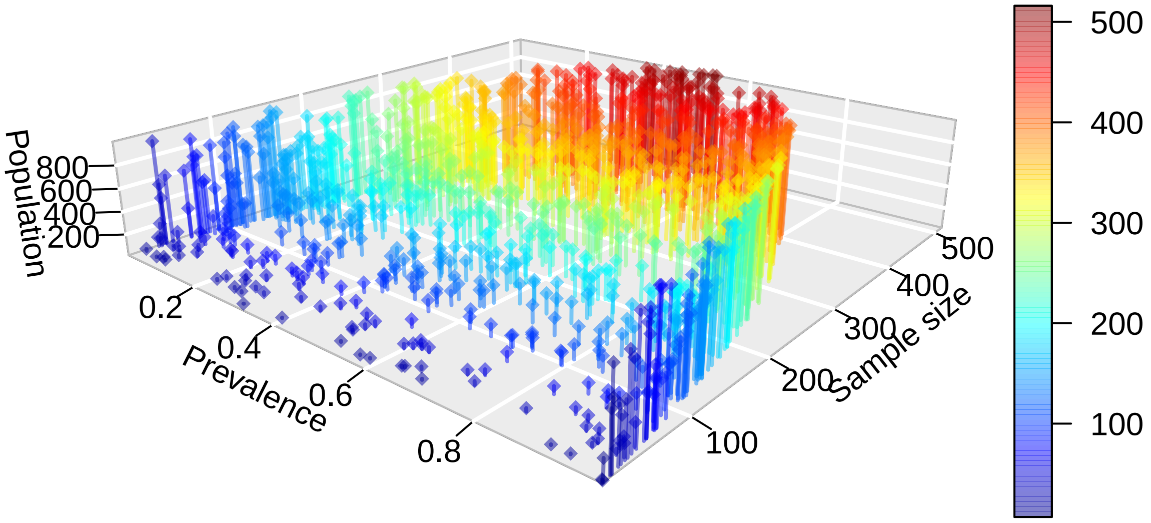
```
##
##       Kukes Gjirokaster       Diber       Lezhe       Berat     Shkoder
##          29          39          43          44          51          70
##       Vlore      Durres       Korce     Elbasan        Fier      Tirane
##          74          76          77          91          98      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

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")
```



# 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 was 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 0.46of householders were still using their stove and the intra-cluster 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 within10%of the true population value, how many villages (clusters) need to be sampled?

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

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

Assignment.

Data for: Clinical Mastitis in cows based on Udder Parameter using Internet of Things (IoT) from this study, each row represents one animal, therefore, 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*" 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 confidence interval of **95%** ?

> This code will filter the data as needed (day 6).

```
#prepare data for analysis
clinical_mastitis_cows <- clinical_mastitis_cows %>% # indicates thedatabase
  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