Practical 4

Pranav Gopalkrishna, 1940223

2021-02-05

# Loading necessary packages...

library(samplingbook)

## Loading required package: pps

## Warning: package 'pps' was built under R version 3.6.2

## Loading required package: sampling

## Warning: package 'sampling' was built under R version 3.6.2

## Loading required package: survey

## Warning: package 'survey' was built under R version 3.6.2

## Loading required package: grid

## Loading required package: Matrix

## Loading required package: survival

##   
## Attaching package: 'survival'

## The following objects are masked from 'package:sampling':  
##   
## cluster, strata

##   
## Attaching package: 'survey'

## The following object is masked from 'package:graphics':  
##   
## dotchart

# The data set

head(ToothGrowth)

## len supp dose  
## 1 4.2 VC 0.5  
## 2 11.5 VC 0.5  
## 3 7.3 VC 0.5  
## 4 5.8 VC 0.5  
## 5 6.4 VC 0.5  
## 6 10.0 VC 0.5

The response variable is the length of odontoblasts (cells responsible for tooth growth) in 60 guinea pigs. Each animal received one of three dose levels of vitamin C (0.5, 1, and 2 mg/day) by one of two delivery methods, orange juice or ascorbic acid (a form of vitamin C and coded as VC).

# Population and parameters

Our population is the supplement column of the above data set, which records the supplement used for the corresponding guinea pig.

popl = ToothGrowth$supp  
popl

## [1] VC VC VC VC VC VC VC VC VC VC VC VC VC VC VC VC VC VC VC VC VC VC VC VC VC  
## [26] VC VC VC VC VC OJ OJ OJ OJ OJ OJ OJ OJ OJ OJ OJ OJ OJ OJ OJ OJ OJ OJ OJ OJ  
## [51] OJ OJ OJ OJ OJ OJ OJ OJ OJ OJ  
## Levels: OJ VC

Let P be the proportion of guinea pigs with supplement type VC. This will be the parameter in question. Let N be the population size.

N = length(popl)  
P = sum(popl == "VC") / N  
P

## [1] 0.5

# Sample and estimators

A sample of 8 elements is drawn

set.seed(4)  
sample = sample(popl, 8)  
sample

## [1] OJ VC OJ VC VC OJ OJ OJ  
## Levels: OJ VC

Let p be the estimator of P. Let n be the sample size.

n = length(sample)  
p = sum(sample == "VC") / n  
p

## [1] 0.375

## Demonstrating that p is and unbiased estimator of P

p can be proved to be an unbiased estimator of P. To demonstrate this, consider pMean, which is the mean of sample proportions taken across many samples…

pMean = mean(replicate(10000, sum(sample(popl, 8) == "VC") / length(sample)))

As can be seen, the mean pf sample proportions tends towards the true population proportion as more samples are taken. This indicates unbiasedness.

# Usage of hypergeometric distribution

Hypergeometric distribution defines the probability of there being a specified number of successes in a sample taken without replcaement draw from a finite population with a fixed and known number of successes.

Since p depends on the number of “successes” in the sample i.e. the number of elements in the sample with supplement as VC, the probability distribution of p across multiple samples follows hypergeometric distribution.

We don’t need to specify this in the function to find the confidence interval, of course, but it is good to know, and helps in the interpretation.

# Estimating the confidence interval for sample proportion

To estimate the 95% confidence interval for the value of p, we use to following…

Sprop(y = (sample == "VC"), n = n, N = N, level = 0.95)

##   
## Sprop object: Sample proportion estimate  
## With finite population correction: N= 60   
##   
## Proportion estimate: 0.375   
## Standard error: 0.1703   
##   
## 95% approximate hypergeometric confidence interval:   
## proportion: [0.0411,0.7089]  
## number in population: [3,42]  
## 95% exact hypergeometric confidence interval:   
## proportion: [0.0833,0.75]  
## number in population: [5,45]

# Interpretation

We may say with 95% confidence that the sample proportion will lie between 0.0411 and 0.7089. In other words, out of 100 samples, we may expect 95 samples to have the sample proportion lie between 0.0411 and 0.7089. This means that the estimate for the proportion of guinea pigs with VC supplement can be said to be between 0.0411 and 0.7089, with 95% confidence on this estimate.