iris

HImanshu Pareek

August 27, 2017

## Comparing population and sample mean and variance

filename is "IRIS.xlsx" Reading data into R from excel file using readxl package

filename <- "IRIS.xlsx"  
  
library(readxl)

## Warning: package 'readxl' was built under R version 3.3.3

iris <- read\_excel (  
 filename,  
 na = c("Not Available", "not available", "NOT AVAILABLE", "na", "NA", "null", "NULL", "-", "\_", "?"))  
head (iris)

## # A tibble: 6 x 6  
## Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm Species  
## <dbl> <dbl> <dbl> <dbl> <dbl> <chr>  
## 1 1 5.1 3.5 1.4 0.2 Iris-setosa  
## 2 2 4.9 3.0 1.4 0.2 Iris-setosa  
## 3 3 4.7 3.2 1.3 0.2 Iris-setosa  
## 4 4 4.6 3.1 1.5 0.2 Iris-setosa  
## 5 5 5.0 3.6 1.4 0.2 Iris-setosa  
## 6 6 5.4 3.9 1.7 0.4 Iris-setosa

Removing na values (if any)

iris <- na.omit(iris)  
head (iris)

## # A tibble: 6 x 6  
## Id SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm Species  
## <dbl> <dbl> <dbl> <dbl> <dbl> <chr>  
## 1 1 5.1 3.5 1.4 0.2 Iris-setosa  
## 2 2 4.9 3.0 1.4 0.2 Iris-setosa  
## 3 3 4.7 3.2 1.3 0.2 Iris-setosa  
## 4 4 4.6 3.1 1.5 0.2 Iris-setosa  
## 5 5 5.0 3.6 1.4 0.2 Iris-setosa  
## 6 6 5.4 3.9 1.7 0.4 Iris-setosa

Deleting id column (because it is duplicate in our data)

iris <- subset(iris, select = -Id)  
head (iris)

## # A tibble: 6 x 5  
## SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm Species  
## <dbl> <dbl> <dbl> <dbl> <chr>  
## 1 5.1 3.5 1.4 0.2 Iris-setosa  
## 2 4.9 3.0 1.4 0.2 Iris-setosa  
## 3 4.7 3.2 1.3 0.2 Iris-setosa  
## 4 4.6 3.1 1.5 0.2 Iris-setosa  
## 5 5.0 3.6 1.4 0.2 Iris-setosa  
## 6 5.4 3.9 1.7 0.4 Iris-setosa

Creating random indices for sample

sample\_index <- sample.int(nrow(iris), 50)  
head (sample\_index)

## [1] 29 56 66 26 83 45

Select sample from iris data

sample\_iris <- iris[sample\_index, ]  
head (sample\_iris)

## # A tibble: 6 x 5  
## SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm Species  
## <dbl> <dbl> <dbl> <dbl> <chr>  
## 1 5.2 3.4 1.4 0.2 Iris-setosa  
## 2 5.7 2.8 4.5 1.3 Iris-versicolor  
## 3 6.7 3.1 4.4 1.4 Iris-versicolor  
## 4 5.0 3.0 1.6 0.2 Iris-setosa  
## 5 5.8 2.7 3.9 1.2 Iris-versicolor  
## 6 5.1 3.8 1.9 0.4 Iris-setosa

Grouping sample\_iris by Species

library(dplyr)

## Warning: package 'dplyr' was built under R version 3.3.3

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

sample\_iris <- group\_by (sample\_iris, Species)  
head (sample\_iris)

## # A tibble: 6 x 5  
## # Groups: Species [2]  
## SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm Species  
## <dbl> <dbl> <dbl> <dbl> <chr>  
## 1 5.2 3.4 1.4 0.2 Iris-setosa  
## 2 5.7 2.8 4.5 1.3 Iris-versicolor  
## 3 6.7 3.1 4.4 1.4 Iris-versicolor  
## 4 5.0 3.0 1.6 0.2 Iris-setosa  
## 5 5.8 2.7 3.9 1.2 Iris-versicolor  
## 6 5.1 3.8 1.9 0.4 Iris-setosa

Summarizing data to get mean and variances

sample\_iris\_summary <- summarize(sample\_iris,   
 sl\_mean = mean(SepalLengthCm),  
 sw\_mean = mean(SepalWidthCm),  
 pl\_mean = mean(PetalLengthCm),  
 pw\_mean = mean(PetalWidthCm),  
 sl\_var = var(SepalLengthCm),  
 sw\_var = var(SepalWidthCm),  
 pl\_var = var(PetalLengthCm),  
 pw\_var = var(PetalWidthCm)  
 )

Grouping original iris data

iris\_group <- group\_by(iris, Species)  
head (iris\_group)

## # A tibble: 6 x 5  
## # Groups: Species [1]  
## SepalLengthCm SepalWidthCm PetalLengthCm PetalWidthCm Species  
## <dbl> <dbl> <dbl> <dbl> <chr>  
## 1 5.1 3.5 1.4 0.2 Iris-setosa  
## 2 4.9 3.0 1.4 0.2 Iris-setosa  
## 3 4.7 3.2 1.3 0.2 Iris-setosa  
## 4 4.6 3.1 1.5 0.2 Iris-setosa  
## 5 5.0 3.6 1.4 0.2 Iris-setosa  
## 6 5.4 3.9 1.7 0.4 Iris-setosa

Summarizing original data

iris\_summary <- summarize(iris\_group,   
 sl\_mean = mean(SepalLengthCm),  
 sw\_mean = mean(SepalWidthCm),  
 pl\_mean = mean(PetalLengthCm),  
 pw\_mean = mean(PetalWidthCm),  
 sl\_var = var(SepalLengthCm),  
 sw\_var = var(SepalWidthCm),  
 pl\_var = var(PetalLengthCm),  
 pw\_var = var(PetalWidthCm)  
)

Printing the summary

sample\_iris\_summary

## # A tibble: 3 x 9  
## Species sl\_mean sw\_mean pl\_mean pw\_mean sl\_var sw\_var  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Iris-setosa 5.118750 3.506250 1.475000 0.243750 0.1229583 0.16062500  
## 2 Iris-versicolor 5.820000 2.770000 4.190000 1.310000 0.3311579 0.08326316  
## 3 Iris-virginica 6.371429 2.964286 5.435714 2.021429 0.2621978 0.07324176  
## # ... with 2 more variables: pl\_var <dbl>, pw\_var <dbl>

iris\_summary

## # A tibble: 3 x 9  
## Species sl\_mean sw\_mean pl\_mean pw\_mean sl\_var sw\_var  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Iris-setosa 5.006 3.418 1.464 0.244 0.1242490 0.14517959  
## 2 Iris-versicolor 5.936 2.770 4.260 1.326 0.2664327 0.09846939  
## 3 Iris-virginica 6.588 2.974 5.552 2.026 0.4043429 0.10400408  
## # ... with 2 more variables: pl\_var <dbl>, pw\_var <dbl>

Here we can see that mean and variance of sample iris data (n = 50) and original iris data (n = 150) are very much similar. But if we look closer we see that Variance of PetalWidth for population is lesser than that for sample baut all other variances are somewhat greater for population data.