Confidence intervals

SISS - Applied Statistics - Chiara Seghieri and Costanza Tortù

2023-11-13

Preliminaries

Recall packages

Import Data

The iris dataset is a built-in dataset in R that contains measurements on 4 different attributes (in centimeters) for 150 flowers from 3 different species. Iris, introduced by Ronald Fisher in his 1936 paper The use of multiple measurements in taxonomic problems, contains three plant species (setosa, virginica, versicolor) and four features measured for each sample. These quantify the morphologic variation of the iris flower in its three species, all measurements given in centimeters.

```
rm(list=ls())
data("iris")
```

Have a first look at data

```
dim(iris)# units x variables
```

```
## [1] 150   5
```

```
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         0.2  setosa
## 5          5.0         3.6          1.4         0.2  setosa
## 6          5.4         3.9          1.7         0.4  setosa
```

Inspect variables

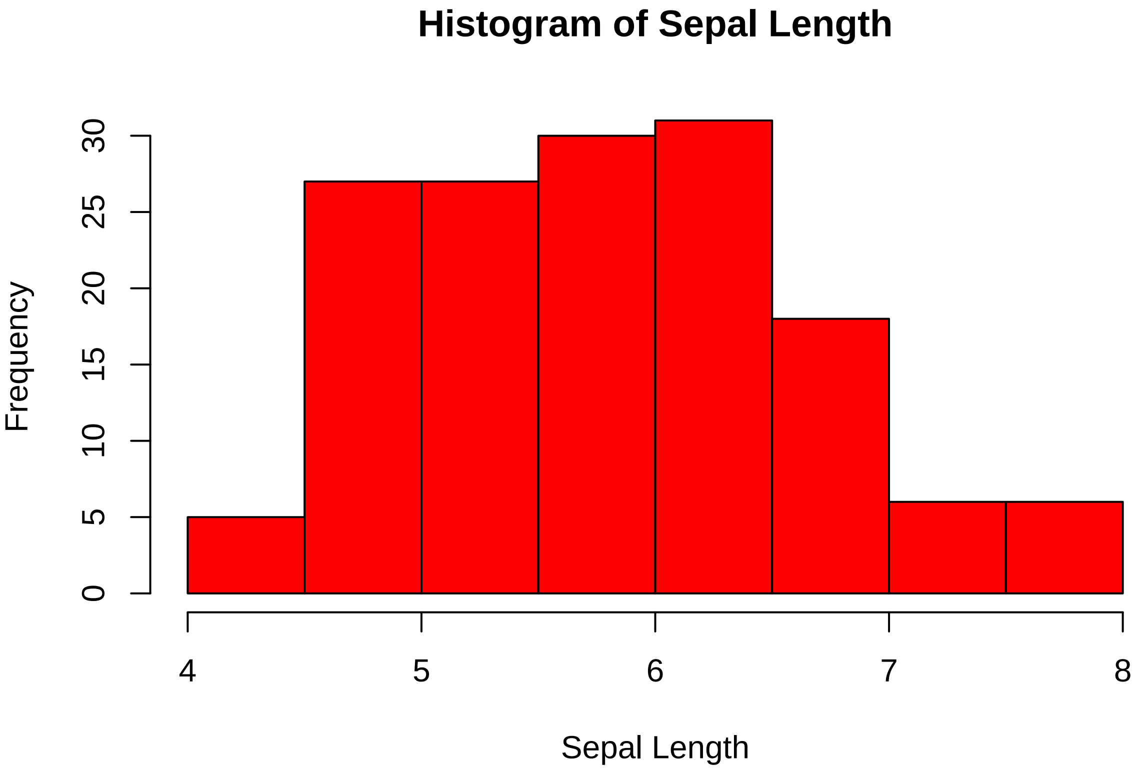
```
colnames(iris)
```

```
## [1] "Sepal.Length" "Sepal.Width"  "Petal.Length" "Petal.Width"  "Species"
```

```
quantitative_variables <- c("Sepal.Length", "Sepal.Width", "Petal.Length", "Petal.Width")
qualitative_variables <- c("Species")
```

Look at the distribution of earnings in 1978

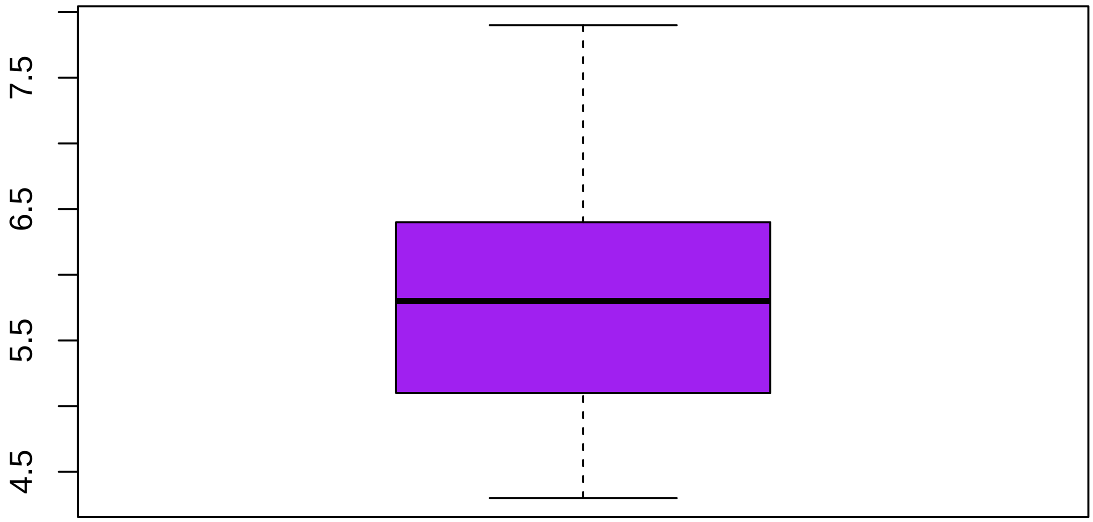
```
hist(iris$Sepal.Length,
     main ="Histogram of Sepal Length",
     col = "red",
     xlab = "Sepal Length")
```



```
summary(iris$Sepal.Length)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   4.300   5.100   5.800   5.843   6.400   7.900
```

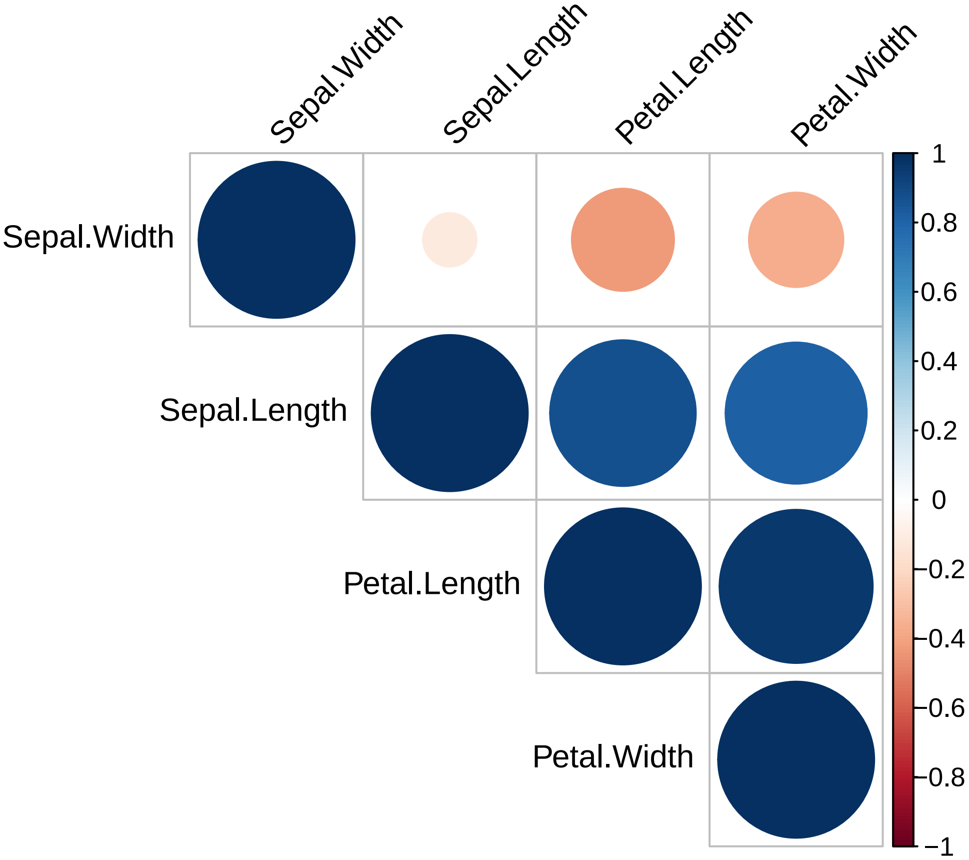
```
boxplot(iris$Sepal.Length,
        col="purple")
```



Let's have a look at the correlation among continuous variables

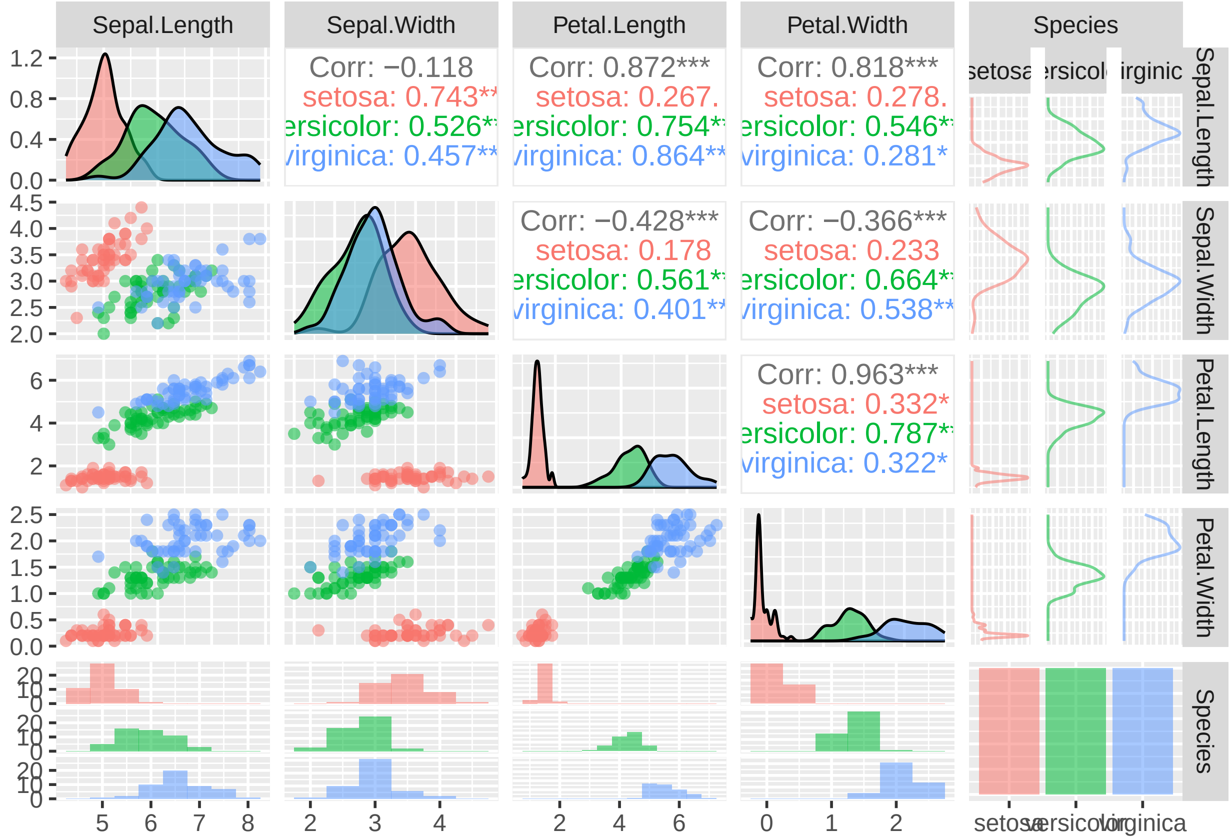
```
correlation_matrix <- cor(iris[, c(quantitative_variables)])

corrplot(correlation_matrix, type = "upper",
         order = "hclust",
         tl.col = "black", tl.srt = 45)
```



Let's explore the relationships among variables, with respect to the flower species

```
ggpairs(iris, aes(color = Species, alpha = 0.5),
        progress = FALSE,
        upper = list(combo = "facetdensity"),
        lower = list(combo=wrap("facethist",
        binwidth=0.5)))
```



Compute confidence intervals for the average Sepal length

Compute sample mean

```
  sample.mean <- mean(iris$Sepal.Length)
  print(sample.mean)
```

```
## [1] 5.843333
```

Compute sample variance

```
sample.n <- length(iris$Sepal.Length)
sample.sd <- sd(iris$Sepal.Length)
sample.se <- sample.sd/sqrt(sample.n)
print(sample.se)
```

```
## [1] 0.06761132
```

Find the t-score

```r
alpha = 0.05
degrees.freedom = sample.n - 1
t.score = qt(p=alpha/2, df=degrees.freedom,lower.tail=F)
print(t.score)
```

```
## [1] 1.976013
```

Compute margin of error

```r
margin.error <- t.score * sample.se
print(margin.error)
```

```
## [1] 0.1336009
```

Now we are ready to compute the confidence interval

```r
lower.bound <- sample.mean - margin.error
upper.bound <- sample.mean + margin.error
print(c(lower.bound,upper.bound))
```

```
## [1] 5.709732 5.976934
```

Compute confidence intervals for the variance of Sepal Length

Compute sample variance

```r
sample.n <- length(iris$Sepal.Length)
sample.var <- var(iris$Sepal.Length)
print(sample.var)
```

```
## [1] 0.6856935
```

Find the chi-scores

```r
alpha = 0.05
degrees.freedom = sample.n - 1
chi.scores = qchisq(c(1-alpha/2, alpha/2), df =  degrees.freedom)
print(chi.scores)
```

```
## [1] 184.687 117.098
```

Now we are ready to compute the confidence interval

```r
lower.bound <- degrees.freedom*sample.var/chi.scores[1]
upper.bound <- degrees.freedom*sample.var/chi.scores[2]

print(c(lower.bound,upper.bound))
```

```
## [1] 0.5531973 0.8725029
```