F19 STA 100 A01 Discussion 03

Yishan Huang

2019/10/15

Discussion Time: Tuesday 8:00 – 8:50 am, Haring Hall 1204.

Notes: https://github.com/Hahahuo-13316/sta100-a01-fall19

Office hour: Tuesday 12:00 – 1:00 pm, Mathematical Sciences Building 1117.

Email: yishuang@ucdavis.edu

More about R

We use an inbuild dataset 'iris' to illustrate the methods of data analysis using R.

```
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
```

```
summary(iris)
```

```
##   Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
##  Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
##  1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
##  Median :5.800   Median :3.000   Median :4.350   Median :1.300
##  Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
##  3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
##  Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
##        Species
##  setosa    :50
##  versicolor:50
##  virginica :50
##
##
##
```

```
pl <- iris$Petal.Length
pl_1 <- pl[iris$Species == "setosa"]
pl_2 <- pl[iris$Species == "versicolor"]
pl_3 <- pl[iris$Species == "virginica"]

c(mean(pl_1), mean(pl_2), mean(pl_3))
```

```
## [1] 1.462 4.260 5.552
```

```
c(var(pl_1), var(pl_2), var(pl_3))
```

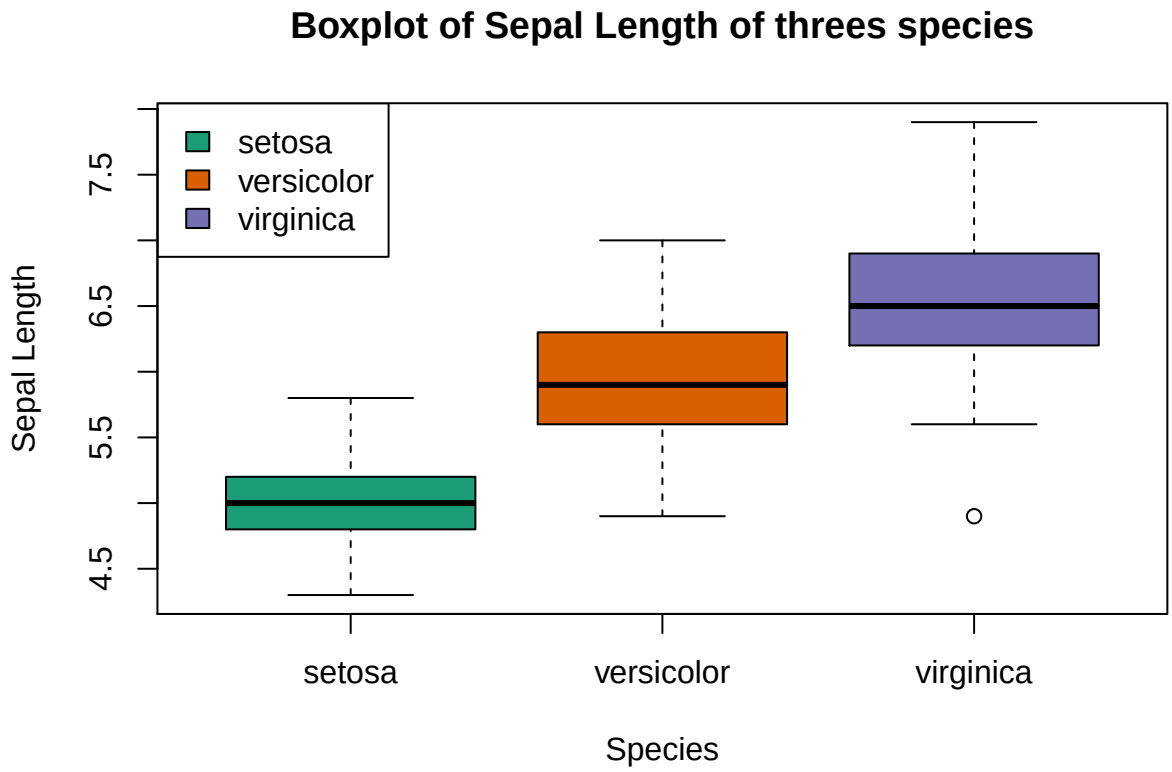
```
## [1] 0.03015918 0.22081633 0.30458776
```

```r
summary(pl_1)
```

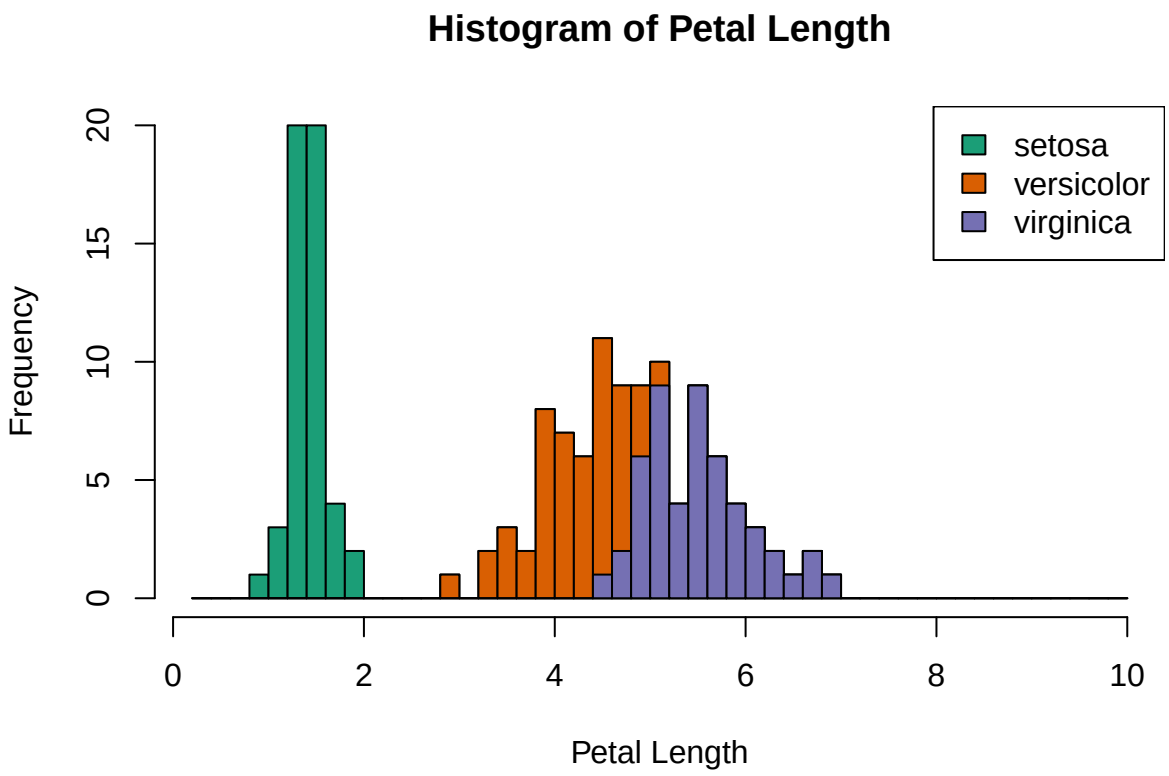
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   1.400   1.500   1.462   1.575   1.900
```

Now let's make some figures!

```r
library(RColorBrewer)
pal <- brewer.pal(8, "Dark2")
boxplot(Sepal.Length ~ Species, data = iris, col = pal[1:3],
        ylab = "Sepal Length", main = "Boxplot of Sepal Length of threes species")
legend("topleft", c("setosa", "versicolor", "virginica"), fill = pal[1:3])
```




```r
hist(iris$Petal.Length, breaks=c(1:50/5), col = pal[1],
     main = "Histogram of Petal Length", xlab = "Petal Length")
hist(iris$Petal.Length[iris$Species == "versicolor" | iris$Species == "virginica"],
     breaks=c(1:50/5), col = pal[2], add=TRUE)
hist(iris$Petal.Length[iris$Species == "virginica"], breaks=c(1:50/5),
     col = pal[3], add=TRUE)
legend("topright", c("setosa", "versicolor", "virginica"), fill = pal[1:3])
```

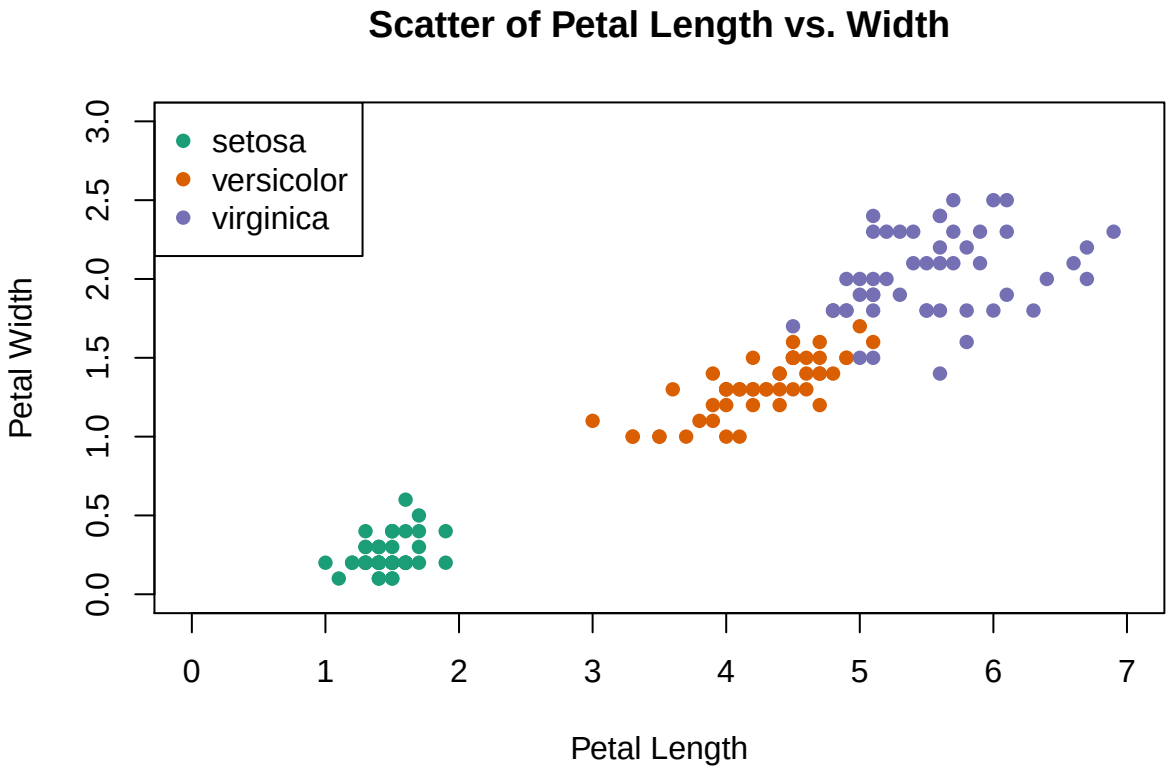



```
plot(iris$Petal.Length[iris$Species == "setosa"],
     iris$Petal.Width[iris$Species == "setosa"],
     col = pal[1], pch = 16, xlim = c(0, 7), ylim = c(0, 3),
     xlab = "Petal Length", ylab = "Petal Width",
     main = "Scatter of Petal Length vs. Width")

points(iris$Petal.Length[iris$Species == "versicolor"],
       iris$Petal.Width[iris$Species == "versicolor"],
       col = pal[2], pch = 16)

points(iris$Petal.Length[iris$Species == "virginica"],
       iris$Petal.Width[iris$Species == "virginica"],
       col = pal[3], pch = 16)

legend("topleft", c("setosa", "versicolor", "virginica"),
       col = pal[1:3], pch = 16)
```



Problems

- (2.4.1) Here are the data from Exercise 2.3.10 on the number of virus-resistant bacteria in each of 10 aliquots: $[14, 15, 13, 21, 15, 14, 26, 16, 20, 13]$
  - Determine the median and the quartiles.
  - Determine the interquartile range.
  - How large would an observation in this data set have to be in order to be an outlier?
- (2.6.6) Ten patients with high blood pressure participated in a study to evaluate the effectiveness of the drug Timolol in reducing their blood pressure. The accompanying table shows systolic blood pressure measurements taken before and after 2 weeks of treatment with Timolol. Calculate the mean and SD of the change in blood pressure (note that some values are negative). (Note: The changes are $[-13, -29, -7, 2, -10, -43, 4, 15, -13, -30]$.)
- (2.6.11) Listed in increasing order are the serum creatine phosphokinase (CK) levels (U/l) of 36 healthy men. (They are 25, 42, 48, 57, 58, 60, 62, 64, 67, 68, 70, 78, 82, 83, 84, 92, 93, 94 ,95, 95, 100, 101, 104, 110, 110, 113, 118, 119, 121, 123, 139, 145, 151, 163, 201, 203.) The sample mean CK level is 98.3 U/l and the SD is 40.4 U/l. What percentage of the observations are within
  - 1 SD of the mean?
  - 2 SDs of the mean?
  - 3 SDs of the mean?
- (2.7.2) The mean and SD of a set of 47 body temperature measurements were as follows: $y = 36.497\,^{\circ}\text{C}$, $s = 0.172\,^{\circ}\text{C}$. If the 47 measurements were converted to °F,
  - What would be the new mean and SD?
  - What would be the new coefficient of variation?