Week 4

PSTAT 122

4/18/2022

1. Review

Models for the data:

$$y_{ij} = \mu + \tau_i + \epsilon_{ij}$$
$$\epsilon_{ij} \sim N(0, \sigma^2)$$

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{..})^2, \quad df = n - 1$$
$$SS_A = n_i \sum_{i=1}^{a}(\bar{y}_{i.} - \bar{y}_{..})^2, \quad df = a - 1$$
$$SS_e = SS_T - SS_A, \quad df = n - a$$

where $n_1 + n_2 + \cdots + n_a = n$. And we can compute these terms according to the following formulas:

$$SS_T = \sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{..}^2}{n}$$
$$SS_A = \sum_{i=1}^{a}\frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{n}$$
$$SS_e = SS_T - SS_A$$

We are interested in testing the hypothesis:

$$H_0: \quad \tau_1 = \tau_2 = \cdots = \tau_a = 0$$
$$H_1: \quad \tau_i \neq 0$$

The test statistic is $F = \frac{SS_A/(a-1)}{SS_e/(n-a)} = \frac{MS_A}{MS_e} \overset{H_0}{\sim} F_{a-1,n-a}$. So we reject H_0 if $F > F_{a-1,n-a}(\alpha)$.

Supplement

- $SS_e = SS_T - SS_A$, whether H_0 holds or not.
- $\frac{SS_e}{\sigma^2} \sim \chi^2_{n-a}$, whether H_0 holds or not.
- $\frac{SS_A}{\sigma^2} \overset{H_0}{\sim} \chi^2_{a-1}$

Factor (Level)	y_{ij}			$y_{i.}$	$y_{i.}^2$
A	12	18		30	900
B	14	12	13	39	1521
C	19	17	21	57	3249
D	24	30		54	2916

2. Analysis

Example The table below gives a small data set, along with some summary statistics.

$n_1 = n_4 = 2$, $n_2 = n_3 = 3$, $n = n_1 + n_2 + n_3 + n_4 = 10$.

$$y_{..} = \sum_{i,j} y_{i,j} = 180$$

$$SS_A = \sum_{i=1}^{4} \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{n} = 258$$

$$SS_T = \sum_{i=1}^{4} \sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{..}^2}{n} = 304$$

$$SS_e = SS_T - SS_A = 46$$

Analyzing sources of variance, we have $F = \frac{SS_A/(4-1)}{SS_e/(10-4)} = 11.21739 > F_{0.05}(3, 6)$.

```
x<-c("A","A","B","B","B","C","C","C","D","D")
y<-c(12,18,14,12,13,19,17,21,24,30)
data = data.frame(x, y)
aov <- aov(y~x, data)
summary(aov)
```

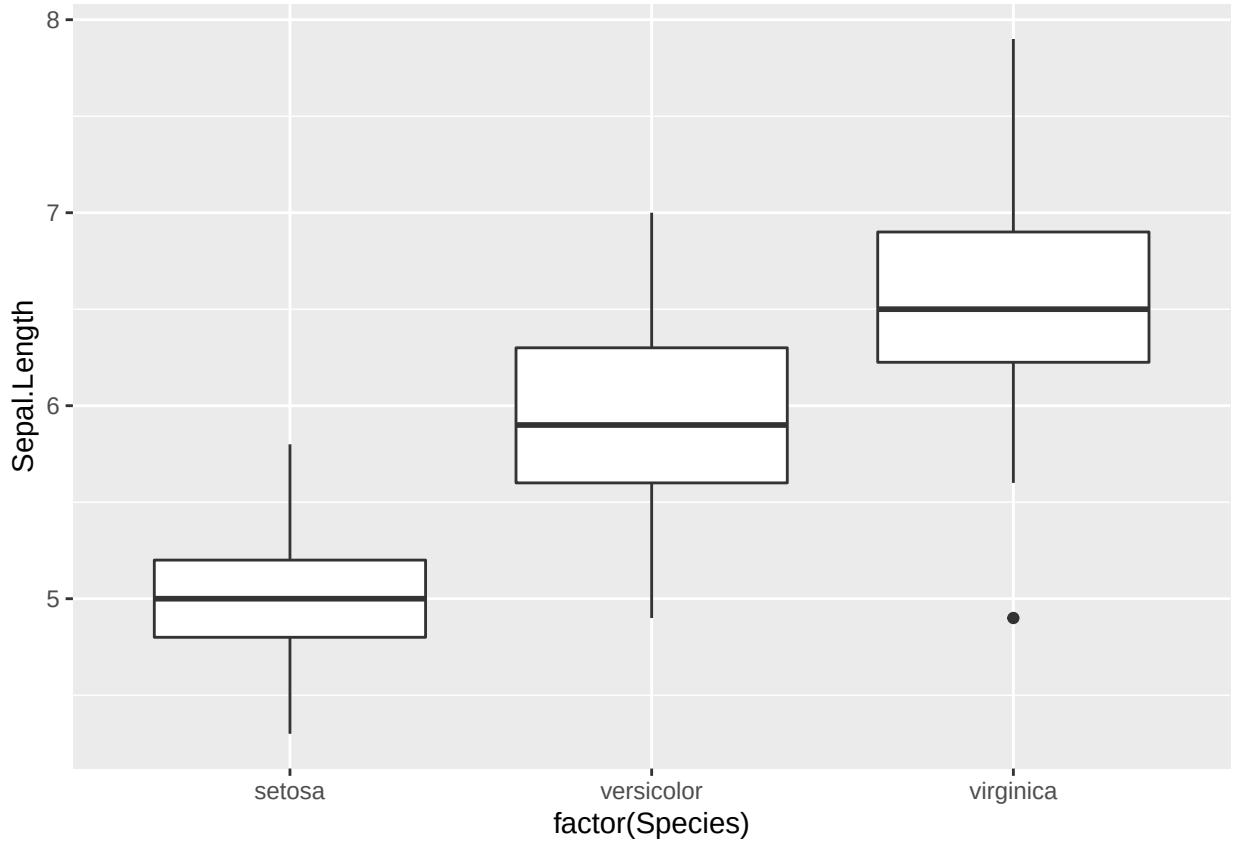
```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## x            3    258   86.00   11.22 0.00713 **
## Residuals    6     46    7.67
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
data(iris)
head(iris)
```

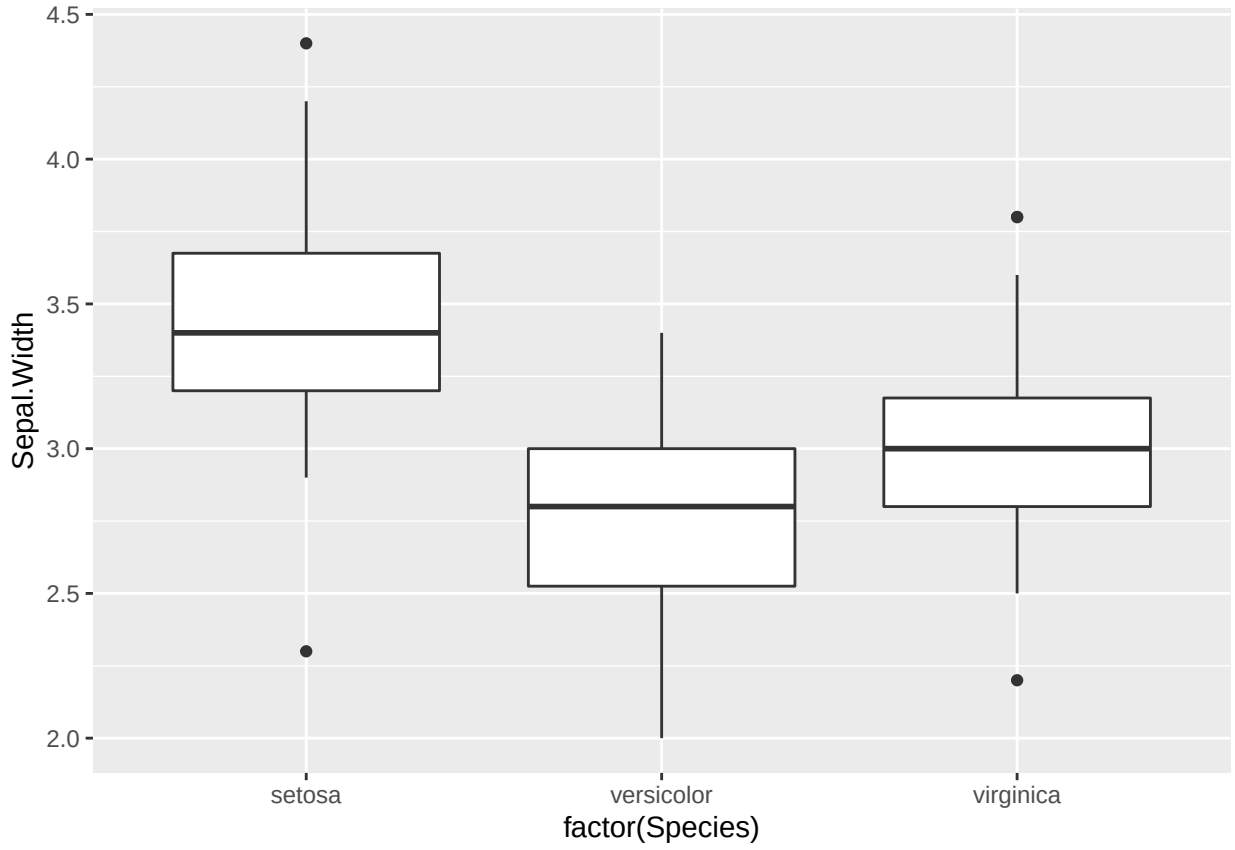
Real data

```
##   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
```

```
library(ggplot2)
ggplot(iris, aes(x=factor(Species), y=Sepal.Length)) + geom_boxplot()
```




```
ggplot(iris, aes(x=factor(Species), y=Sepal.Width)) + geom_boxplot()
```




```
aov1 <- aov(Petal.Length ~ Species, iris)
summary(aov1)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Species       2  437.1  218.55    1180 <2e-16 ***
## Residuals   147   27.2    0.19
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
aov2 <- aov(Sepal.Length ~ Species, iris)
summary(aov2)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Species       2  63.21  31.606   119.3 <2e-16 ***
## Residuals   147  38.96   0.265
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```