MS-C1620 Statistical Inference

Exercise 12

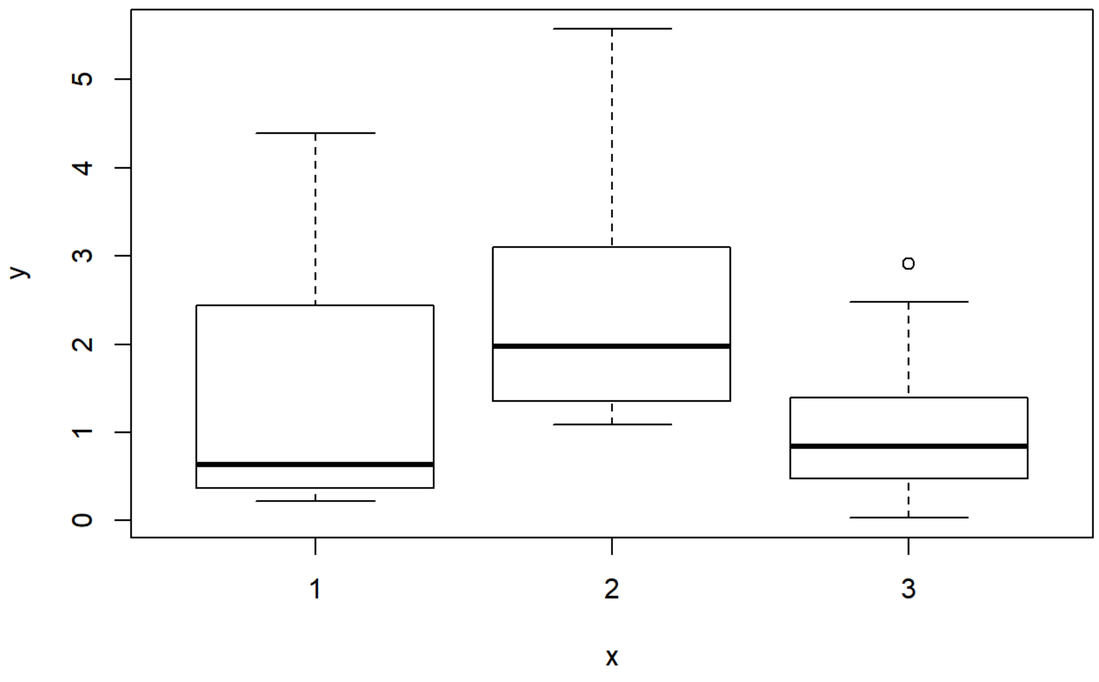
Homework exercise

To be solved at home before the exercise session.

---

1. Consider a data set with measurements of the variable `y` for three groups (`x`). Each group has sample size 15. Below are shown boxplots of the groups, along with outputs given by ANOVA and the Kruskal-Wallis test for the data.
    a. What are the conclusions of the two tests?
    b. Which test (if either) would you trust and why?
    c. How would you continue the analysis?

```
boxplot(y ~ x, data = my_data)
```




```
summary(aov(y ~ x, data = my_data))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## x            1   1.13   1.129   0.586  0.448
## Residuals   43  82.89   1.928
```

```
kruskal.test(y ~ x, data = my_data)
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  y by x
## Kruskal-Wallis chi-squared = 10.185, df = 2, p-value = 0.006142
```

Class exercise

To be solved at the exercise session.

---

1. A botanist wants to test the hypothesis that the three iris species have equal expected value of `Sepal.Width`.
    a. Visualize the data.
    b. Conduct an analysis of variance.
    c. Are the assumptions of ANOVA satisfied?
    d. If the assumptions are fulfilled, conduct pairwise comparisons using the Bonferroni correction.
    e. State your conclusions.

---

2. The data set `mtcars` has measurements for 32 cars. We investigate the relationship between `mpg` (miles/gallon, the response) and `hp` and `am` (horsepowers and transmission type, the explanatory variables) through an *analysis of covariance*.
    a. Find a suitable visualization for the data.
    b. Using the function `lm`, fit a regression model with the covariates `hp`, `am` and `hp:am` (the final one is an interaction effect, the product of the two covariates).
    c. Interpret the fitted model (homework problem 10.1.a might prove helpful).

---

3. **(Optional)** Consider still the `mtcars` data set but replace the variable `am` with the variable `gear` (and make sure its type is `factor`). Fit the linear regression model `mpg ~ hp + gear` and find out how the function `anova` can be used to test whether all regression coefficients related to `gear` are equal to zero **simultaneously**. *Note that the situation is different from problem 2 as `gear` has three classes (i.e., two coefficients) and thus the p-values from the model only relate to the hypotheses whether the two coefficients can be set to zero **individually**.*
