MS-C1620 Statistical Inference

Exercise 12

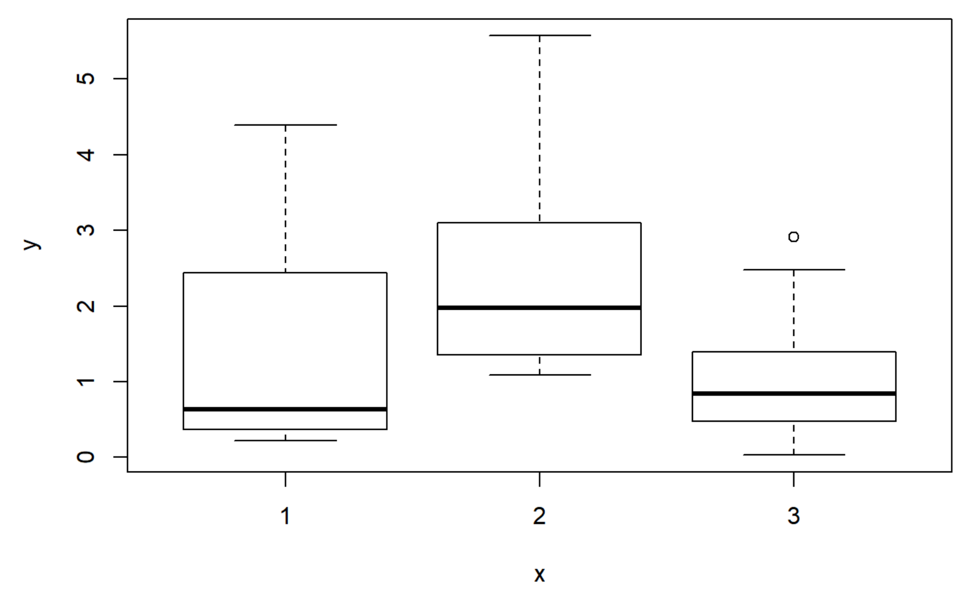
Homework exercise

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
```

```
# a.
# ANOVA claims that there is no evidence that the expected values of the groups differ.
# K-W claims that the medians of the groups are not all same.

# b.
# The boxplots show evidence that the groups have positively skew distributions, meaning that the normality assumption of AN
OVA is unlikely to hold. Moreover, the "replacement" of the normality assumption with a large enough sample size is question
able here as we have only 15 obs. per group -> Cannot trust ANOVA.
# K-W requires that the group distributions have the same shape, meaning that the boxplots should look otherwise similar but
have possibly different locations in the y-axis. Based on the plot, this seems plausible -> We can trust K-W and thus conclu
de that the group medians differ.

# c.
# The analysis could be continued with pair-wise testing using e.g. the two-sample rank test to find out which pairs of grou
ps have differing medians (accompanied with a suitable correction, such as the Bonferroni correction).
```

Class exercise

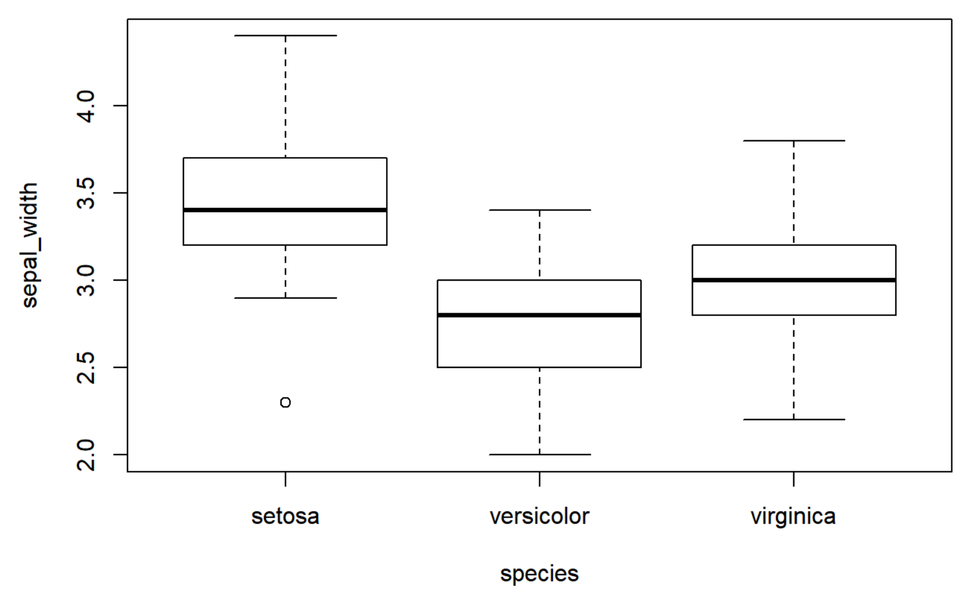
To be solved at the exercise session.

---

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.

```
# a.
flowers <- data.frame(sepal_width = iris[, 2], species = iris[, 5])

# The boxplots show that at least the group "Setosa" seems to differ from the others
boxplot(sepal_width ~ species, data = flowers)
```



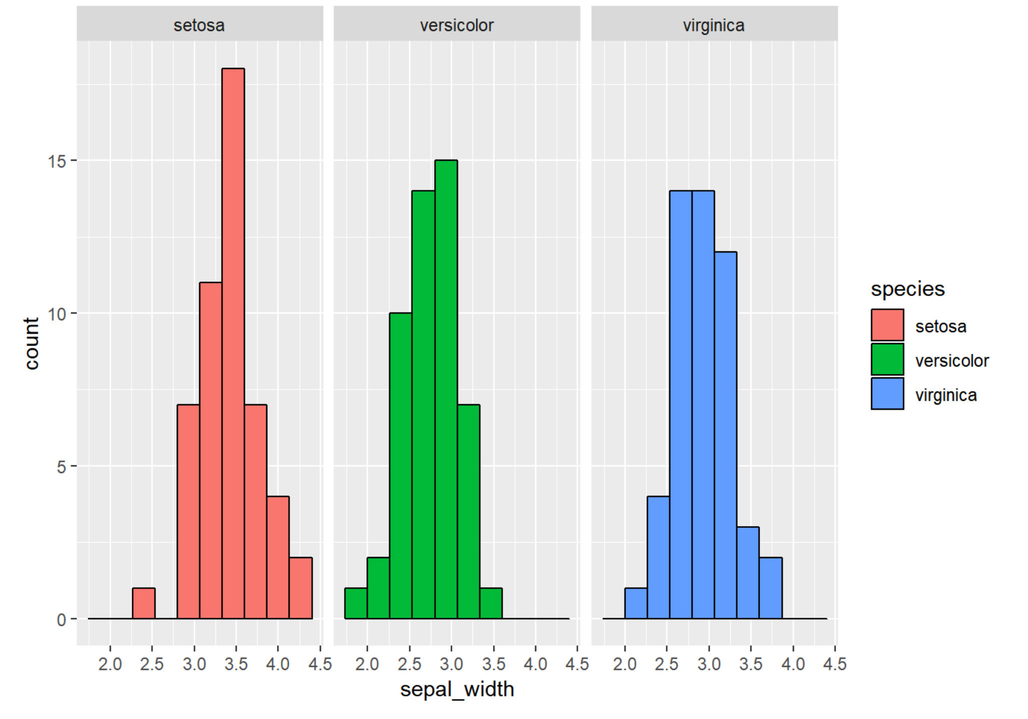

```
# b.
# ANOVA finds differences in the group expected values, p-value < 0.05
# -> Not plausible that the expected values are the same (given that the assumptions of ANOVA hold).
flowers_aov <- aov(sepal_width ~ species, data = flowers)
summary(flowers_aov)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## species       2  11.35   5.672   49.16 <2e-16 ***
## Residuals   147  16.96   0.115
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# c.
# Bartlett's test shows no evidence that the variances would differ.
bartlett.test(sepal_width ~ species, data = flowers)
```

```
##
##  Bartlett test of homogeneity of variances
##
## data:  sepal_width by species
## Bartlett's K-squared = 2.0911, df = 2, p-value = 0.3515
```

```
# Based on the histograms, the group-wise normality assumption seems plausible.
library(ggplot2)
ggplot(flowers, aes(x = sepal_width, fill = species)) +
  geom_histogram(bins = 10, col = "black") +
  facet_wrap(. ~ species)
```




```
# -> we conclude that the assumptions of ANOVA are fulfilled.


# d. & e.
# Pairwise testing with the Bonferroni correction reveals that all three groups differ
# from each other in terms of expected values.
pairwise.t.test(flowers$sepal_width, flowers$species, p.adjust.method = "bonferroni")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  flowers$sepal_width and flowers$species
##
##            setosa  versicolor
## versicolor < 2e-16 -
## virginica  1.4e-09 0.0094
##
## P value adjustment method: bonferroni
```

---

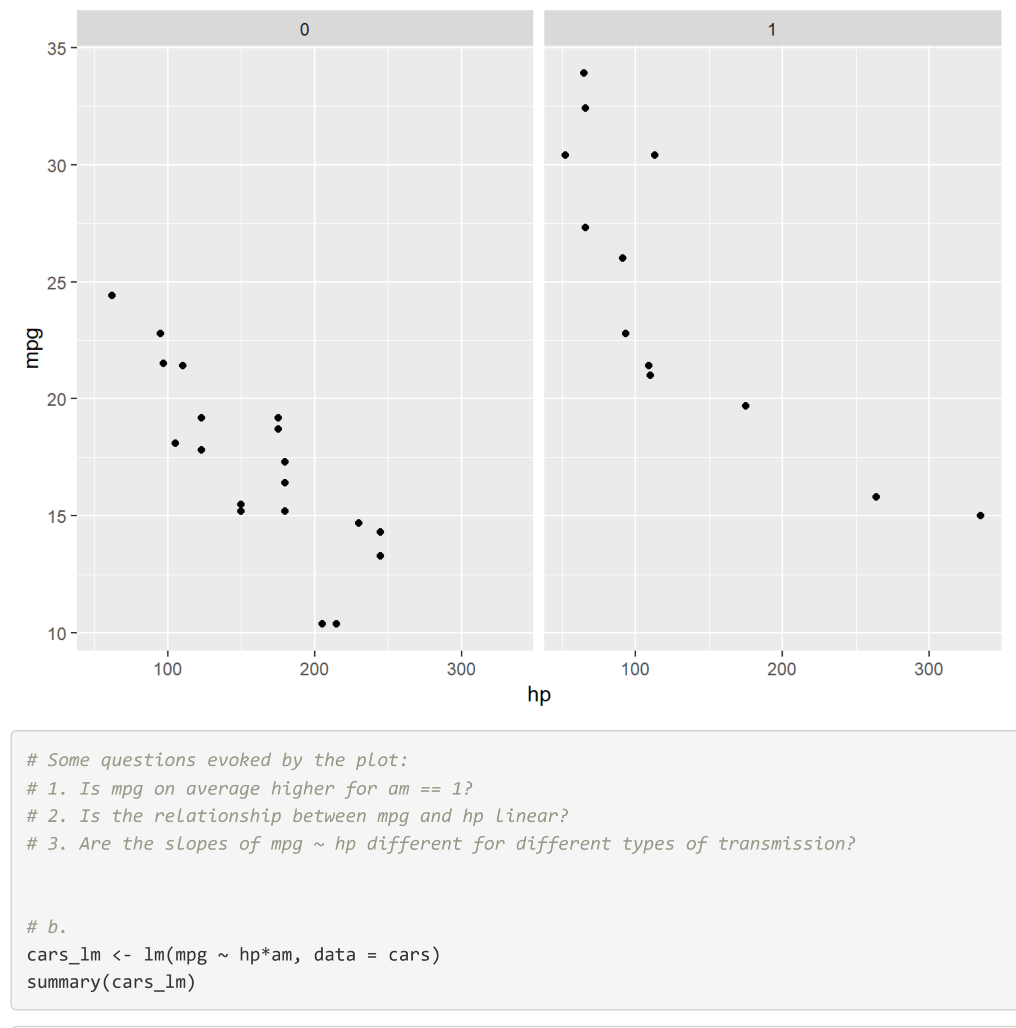
2. The data set `mtcars` has measurements for 32 cars. We investigate the relationship between `mpg` (miles/gallon, the response) and `hp` and `am` (horsepowers and transmission type, the explanatory variables) through an *analysis of covariance*.
    a. Find a suitable visualization for the data.
    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).

```
#a.
cars <- data.frame(mpg = mtcars$mpg, hp = mtcars$hp, am = mtcars$am)

# Scatter plots of mpg and hp for each level of am
# plot(mpg ~ hp, data = cars[cars$am == "0", ])
# plot(mpg ~ hp, data = cars[cars$am == "1", ])

# Or more succinctly in ggplot
library(ggplot2)

ggplot(cars, aes(x = hp, y = mpg)) +
  geom_point() +
  facet_wrap(. ~ am)
```




```
# Some questions evoked by the plot:
# 1. Is mpg on average higher for am == 1?
# 2. Is the relationship between mpg and hp linear?
# 3. Are the slopes of mpg ~ hp different for different types of transmission?


# b.
cars_lm <- lm(mpg ~ hp*am, data = cars)
summary(cars_lm)
```

```
##
## Call:
## lm(formula = mpg ~ hp * am, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3818 -2.2696  0.1344  1.7058  5.8752
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 26.6248479  2.1829432  12.197 1.01e-12 ***
## hp          -0.0591370  0.0129449  -4.568 9.02e-05 ***
## am           5.2176534  2.6650931   1.958   0.0603 .
## hp:am        0.0004029  0.0164602   0.024   0.9806
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.961 on 28 degrees of freedom
## Multiple R-squared:  0.782,  Adjusted R-squared:  0.7587
## F-statistic: 33.49 on 3 and 28 DF,  p-value: 2.112e-09
```

```
# c.
# As in homework problem 10.1.a, we write the model separately for am == 0 and am == 1:
#
# am == 0:
# E[mpg] = b0 + b1*hp
#
# am == 1:
# E[mpg] = b0 + b1*hp + b2 + b3*hp = (b0 + b2) + (b1 + b3)*hp
#
# where the b's and the model output above have the following correspondences:
# b0 = "(Intercept)", b1 = "hp", b2 = "am", b3 = "hp:am"

# We interpret the coefficients b1, b2, b3 each in turn:
#
# b3    describes the difference of the hp-slopes for the two transmission types.
#       p-value is almost 1 and we conclude that it is plausible that b3 = 0
#       -> the slopes do not differ from each other (the horsepowers do not affect mpg
#       differently for the two types of transmission).
#
# b1    describes the slope of the group with am == 0 (but also of the group with am ==
#       1, since we now believe that b3 = 0). p-value is below 0.05 so the slope differs
#       significantly from zero -> we conclude that a unit increase in horsepowers
#       lowers mpg by -0.06.
#
# b2    describes the difference of the intercept terms for the two transmission types.
#       p-value >= 0.05 and we conclude that it is plausible that b2 = 0.
#       -> the lines do not differ in their vertical position (that is, even though the
#       points of am == 1 are higher in the plot, we established that there is
#       not enough evidence to show that this effect is not caused by randomness).
```

---

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**.*
