# Applied Data Analysis, Übung 5

Henry Haustein

## Vorbereitung

```r
install.packages("dplyr")
library(dplyr)
install.packages("readxl")
library(readxl)
data = read_excel("data.xlsx", na = "NA")

data$gender = factor(data$gender)
levels(data$gender) = c("male", "female", "diverse")
data$employment = factor(data$employment)
levels(data$employment) = c("student", "employed", "unemployed")
data$education = factor(data$education)
levels(data$education) = c("no degree", "secondary", "intermediate
    ", "high school", "academic")
data$play_frequency = factor(data$play_frequency)
levels(data$play_frequency) = c("never", "every few months", "
    every few weeks", "1-2 days a week", "3-5 days a week", "daily
    ")
data$treatment = factor(data$treatment)
levels(data$treatment) = c("control", "lootbox in task reward", "
    lootbox picture", "badge")
data$age = sapply(data$age, function(year) {2016-year})
data$rt6 = as.numeric(data$rt6)
data$rt7 = as.numeric(data$rt7)
data$rt8 = as.numeric(data$rt8)
data$rt9 = as.numeric(data$rt9)
data$rt10 = as.numeric(data$rt10)
data$rt11 = as.numeric(data$rt11)
data$rt12 = as.numeric(data$rt12)
data$rt13 = as.numeric(data$rt13)
data$rt14 = as.numeric(data$rt14)
```

## Task 1

```r
aov_tc_treatment = aov(tasks_completed ~ treatment, data = data)
summary(aov_tc_treatment)

```

```r
pairwise.t.test(data$tasks_completed, data$treatment, p.adjust.method = "none")
pairwise.t.test(data$tasks_completed, data$treatment, p.adjust.method = "bonferroni")
pairwise.t.test(data$tasks_completed, data$treatment, p.adjust.method = "holm")
```

Ergebnis von ANOVA ist

| | Df | Sum Sq | Mean Sq | F value | Pr(>F) |
|---|---|---|---|---|---|
| treatment | 3 | 1585.61 | 528.54 | 34.24 | $< 2 \cdot 10^{-16}$ |
| Residuals | 433 | 6683.72 | 15.44 | | |

Da ANOVA auf

- $H_0 : \mu_1 = \cdots = \mu_g$
- $\exists i, j : \mu_i \neq \mu_j$

testet, sehen wir am p-value, dass die Treatments einen unterschiedlichen Mittelwert haben. Die paarweisen t-Tests haben folgendes Ergebnis (ohne p-value-adjustment, Bonferroni-Korrektur[1], Bonferroni-Holm Korrektur[2]):

| | control | lootbox in task reward | lootbox picture |
|---|---|---|---|
| lootbox in task reward | $< 2 \cdot 10^{-16}$, $< 2 \cdot 10^{-16}$, $< 2 \cdot 10^{-16}$ | | |
| lootbox picture | 0.05967, 0.35800, 0.05967 | $7.3 \cdot 10^{-14}$, $4.4 \cdot 10^{-13}$, $3.6 \cdot 10^{-13}$ | |
| badge | $2.8 \cdot 10^{-8}$, $1.7 \cdot 10^{-7}$, $1.1 \cdot 10^{-7}$ | $1.9 \cdot 10^{-5}$, 0.00011, $5.6 \cdot 10^{-5}$ | 0.00013, 0.00079, 0.00026 |

Wir sehen, dass bei fast allen Gruppen sich die Mittelwerte unterscheiden, bis auf $\mu_{\text{lootbox pic}} = \mu_{\text{control}}$.

# Task 2

```r
install.packages("car")
library(car)
data %>% group_by(treatment) %>% summarise(completeTasks_var = var(tasks_completed))
leveneTest(data$tasks_completed, data$treatment)
oneway.test(tasks_completed ~ treatment, data = data, var.equal = FALSE)

qqnorm(data$tasks_completed)
qqline(data$tasks_completed)
kruskal.test(tasks_completed ~ treatment, data = data)
```

Die Varianzen unterscheiden sich:

[1] $p^* = \frac{\alpha}{n}$ mit $n$ Anzahl der Tests
[2] $p^* = \frac{\alpha}{m+1-k}$ mit $p_1, ..., p_m$ geordneten p-values und für die $k$-te These

| | treatment | completeTasks_var |
|---|---|---|
| 1 | control | 12.43 |
| 2 | lootbox in task reward | 12.29 |
| 3 | lootbox picture | 15.56 |
| 4 | badge | 19.82 |

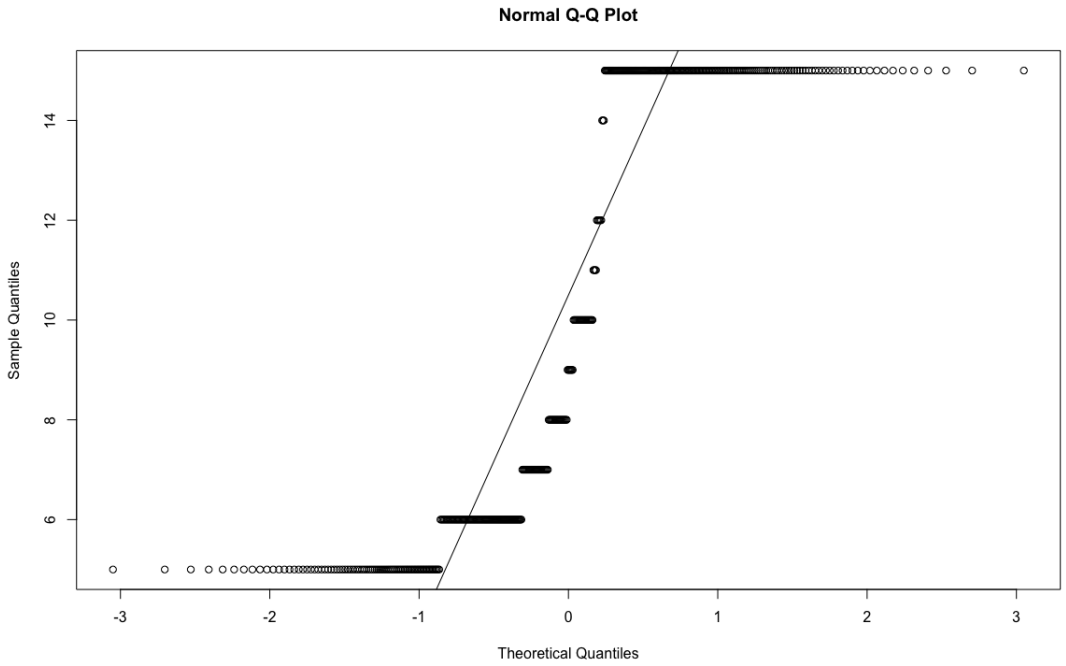
Der Levene Test testet

- $H_0 : \sigma_1^2 = \cdots = \sigma_g^2$
- $H_1 : \exists i, j : \sigma_i^2 \neq \sigma_j^2$

und hat folgendes Ergebnis

| | Df | F value | Pr(>F) |
|---|---|---|---|
| group | 3 | 14.14 | $< 2 \cdot 10^{-16}$ |
| | 433 | | |

Ja, die Varianzen unterscheiden sich. Der Oneway-ANOVA-Test hat einen p-value von $< 2 \cdot 10^{-16}$, also gibt es auch hier Unterschiede zwischen den Gruppen.




Offensichtlich kann man hier auch nicht von Normalverteilung ausgehen. Der Kruskal-Wallis-Test testet, ob die Gruppen aus der gleichen Population ($H_0$) kommen. Das scheint hier nicht der Fall zu sein, der p-value ist $1.51 \cdot 10^{-15}$.