Assignment Data Analysis

Kevin Saner, Phillip Gachnang, Raphael Denz

18. 3. 2021

Starting situation

The case is adapted from Sharpe, N.R., Ali, A., Potter, M.E. (2001): A Casebook for Business Statistics, Wiley, New York, p 13-20. The data represents a survey where entrepreneurs were asked about the satisfaction of their product development process. Because developing a new product is often challenging, the participants were also asked if the used the help of an incubator program to jump-start their business. The survey should further give answers on satisfaction with the incubator programs, and also give insights on what kind of companies used those programs.

To-Do

Carry out the following tasks. Comment on each of the results in a few words! Work in groups of 3 members. Establish a report. Write the full names of all members of the group on the top of the first page of the report. Upload the report as a pdf to moodle: Only one person per group must upload the pdf.

Prerequisites

```
library(DescTools) # for CramerV() function
library(haven) # to read SPSS files
library(labelled) # to deal with variable and value labels
library(knitr) # to get the kable() function for nice tables
library(gplots) # to get the plotmeans() function
library(ggplot2) # for grouped boxplot
```

1 Data Preparation

1.1 Read the data

```
rm(list = ls()) # clear workspace
dataset = read_sav("npd.sav")
```

1.2 Check the dimension and the variables

Check the dimensions of the data set. What are the names. Is the data labelled?

```
a<- dim(dataset)
a

## [1] 247  57
```

The above output shows that the dataset contains 247 observations and 57 variables.

```
names(dataset)

##  [1] "SurveyID"  "A1"        "A2"        "A3"        "A4"        "A5"
##  [7] "A6"        "A7"        "A8"        "A9"        "A10"       "A11"
## [13] "B1"        "B2"        "B3"        "B4"        "B5"        "B6"
## [19] "B7"        "B8"        "B9"        "B10"       "B11"       "B12"
## [25] "B13"       "C1"        "C2"        "C3"        "C4"        "C5"
## [31] "C6"        "C7"        "C8"        "C9"        "D1"        "D2"
## [37] "D3AVC"     "D3BLoan"   "D3CSBIR"   "D3DIPO"    "D3EPP"
"D3FGrant"
## [43] "D4"        "D5"        "D6"        "D7"        "D8"
"D9IncYes1"
## [49] "D9BNP"     "D9CU"      "D9DFP"     "D9EGov"    "COMPANY"   "STATE"
## [55] "COMPLETE"  "DELAY"     "GROUP"
```

The above command delivers the names of the columns in the data set. It can be seen that the names of the columns match the questions of the questionnaire. The data set contains no labels. Some of the data in the data set is stored as free text such as the product category (A1). Most of data, however, is stored as numerical data where it only becomes clear when looking at the questionnaire whether the data refers to qualitative or quantitative data. The below example marks an exception:

```
# example of different encoding styles
head(dataset$A2)

## [1] 1 0 1 1 1 1

head(dataset$B1)

## [1] "N" "Y" "Y" "Y" "N" "N"
```

Although, both questions A2 and B1 refer to a “Yes”- or “No”-question, the way the answers are stored look different. This due to a inconsistent encoding style.

1.3 Response

Calculate the unit response rate.

```
# store gross sample size
gross_sample_size <- 592
# keep only rows that indicated completeness
dataset <- subset(dataset,dataset$COMPLETE!="")
100 /gross_sample_size * nrow(dataset)

## [1] 41.55405
```

To calculate the unit response rate only data is considered that is complete. Incomplete data is identified using the column “Complete”. This way a unit reponse rate of 41.55% is calculated, meaning that less than half the companies that were asked completed the questionnaire.

1.4 Measurement Levels and Missingness

Check the measurement levels and missigness of sections C & D.

1.4a Measurement Levels

The measurement levels of the variables in section C & D are follows:

- C1 - C4 are ratios
- C 5 -7 are ordinally scaled variable
- C8, C9 & D1 are ratios
- D2 is a nominal variable
- D3 is a nominal variable that uses one-hot encoding
- D4 - D8 are ratios
- D9IncYes1 is a nominal variable
- D9* is dependent on D9IncYes1 and also uses one-hot encoding

To make the data better readable it is useful to label the ordinal and nominal variables. Variables C5 - C7 follow a 7-level Likert-Scale, therefore the labels will be:

1 = Strongly Disagree
2 = Disagree
3 = Somewhat Disagree
4 = Neither Agree or Disagree
5 = Somewhat Agree
6 = Agree
7 = Strongly Agree

```
# transform the variables to factor and add labels
to_7_level_likert <- function(x){
x <- factor(x,
                     levels = c(1,2,3,4,5,6,7),
                     labels = c("Strongly Disagree",
                                "Disagree",
                                "Somewhat Disagree",
                                "Neither Agree or Disagree",
                                "Somewhat Agree",
                                "Agree",
                                "Strongly Agree"),
                     ordered = TRUE)
}
dataset$C5 <- to_7_level_likert(dataset$C5)
dataset$C6 <- to_7_level_likert(dataset$C6)
dataset$C7 <- to_7_level_likert(dataset$C7)
# example of the newly introduced factors
levels(dataset$C5)
```

```
## [1] "Strongly Disagree"         "Disagree"
## [3] "Somewhat Disagree"         "Neither Agree or Disagree"
## [5] "Somewhat Agree"            "Agree"
## [7] "Strongly Agree"
```

The example shows the levels of the transformed variable C5.

D2 and D9IncYes1 are nominal variables with two levels each. The variables are also transformed to factors and labels are added.

```
dataset$D2 <- factor(dataset$D2,
                     levels = c(1,2),
                     labels = c("Privately held","Publicly Traded"),
                     ordered = FALSE)
dataset$D9IncYes1 <- factor(dataset$D9IncYes1,
                            levels = c(0,1),
                            labels = c("No","Yes"),
                            ordered = FALSE)
summary(dataset$D2)

##  Privately held Publicly Traded            NA's
##             230               6              10

summary(dataset$D9IncYes1)

##   No  Yes NA's
##  159   68   19
```

The output shows the count of occurrences per label.

1.4b Missingness

Check data for missing values in sections C and D.

```
#create a subset that only contains sections C and D
dataset_c_d <-
dataset[,grep("C1",colnames(dataset)):grep("D9EGov",colnames(dataset))]
out <- table(sapply(dataset_c_d, is.na))
out

##
## FALSE  TRUE
##  4086  2556
```

There are 2556 missing values in the data set. The values are distributed as follows among the variables:

```
data.frame(colSums(is.na(dataset_c_d)))

##             colSums.is.na.dataset_c_d..
## C1                                   26
## C2                                   41
## C3                                   45
## C4                                   76
## C5                                   17
## C6                                   16
## C7                                 22
## C8                                 21
## C9                                 30
## D1                                 13
## D2                                 10
## D3AVC                             212
## D3BLoan                           144
## D3CSBIR                           173
## D3DIPO                            240
## D3EPP                             165
## D3FGrant                          221
## D4                                 13
## D5                                 21
## D6                                 74
## D7                                 44
## D8                                 19
## D9IncYes1                          19
## D9BNP                             212
## D9CU                              225
## D9DFP                             235
## D9EGov                            222
```

1.5 Missingness of D9IncYes1

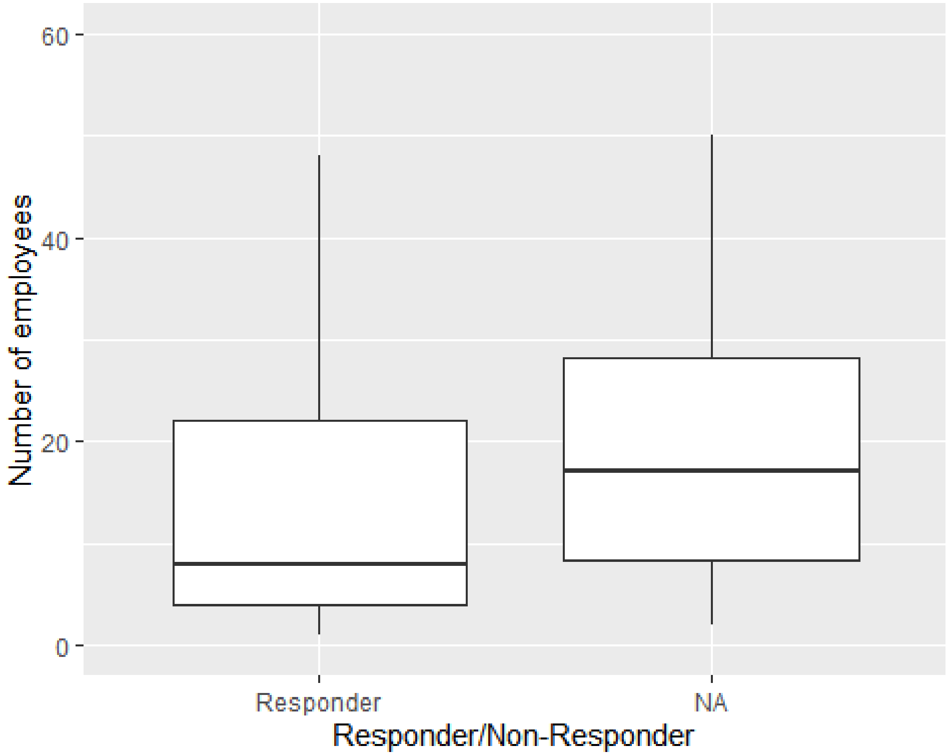
Create a contingency table between is.na(D4) and is.na(D9IncYes1)

```
# contingency table
table(is.na(dataset$D4),is.na(dataset$D9IncYes1))

## 
##         FALSE TRUE
##   FALSE   223   10
##   TRUE      4    9
```

The result of the contigency tables shows that, 223 have answered both questions, 9 have answered neither D4 nor D9IncYes1, 10 have answered D4 but not D9IncYes1 and 4 have answered D9IncYes1 but not D4. To sum, there are only 23 companies that did not answer one of the questions, which relates to roughly 10% of responders.

Create a grouped boxplot to compare responders and non-responders.

```
#make a subset with only the variables of interest
dataset_D4_D9IncYes1 <- dataset[,c("D4","D9IncYes1")]
#keep only rows where the company size is meaningful
dataset_D4_D9IncYes1<- subset(dataset_D4_D9IncYes1,!is.na(dataset$D4))
#if D9IncYes1 is !NA, it counts as reponder
dataset_D4_D9IncYes1$D9IncYes1 <- factor(dataset_D4_D9IncYes1$D9IncYes1,
                            levels = c("Yes","No"),
                            labels = c("Responder","Responder"),
                            ordered = FALSE)
# outliers are hidden in the boxplot
ggplot(dataset_D4_D9IncYes1, aes(x=D9IncYes1,y=D4)) +
geom_boxplot(outlier.colour = NA) +  coord_cartesian(ylim = c(0, 60)) +
xlab("Responder/Non-Responder") + ylab("Number of employees")
```



The median of the NA-group is close to 20, whereas the median of the responders is under 10, which indicates a size difference of the companies when looking at the employees. Additionally, a wilcoxon test is done to test this hypothesis:

```
# test for a statistically significant difference between responder and non-
responder
wilcox.test(dataset$D4[which(is.na(dataset$D9IncYes1))],dataset$D4[which(!is.
na(dataset$D9IncYes1))])

## 
##  Wilcoxon rank sum test with continuity correction
## 
## data:  dataset$D4[which(is.na(dataset$D9IncYes1))] and
dataset$D4[which(!is.na(dataset$D9IncYes1))]
## W = 1382.5, p-value = 0.1997
## alternative hypothesis: true location shift is not equal to 0
```

The Wicoxon-Test shows we can reject the null hypothesis, which means that companies that did not respond to D9IncYes1 have more employees.

2 Analysis

2.1 Proportion of Programme

Calculate the proportion and 95% confidence interval for incubator participants.

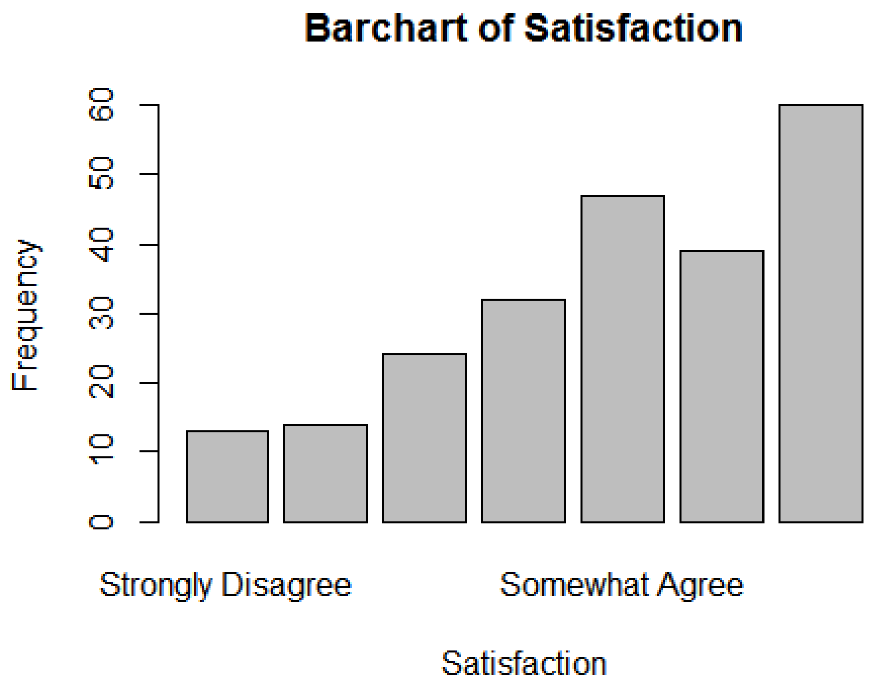
```r
#only non-NA values are used to calculate the total
D9IncAll <- length(subset(dataset$D9IncYes1,!is.na(dataset$D9IncYes1)))
D9IncYes <- length(subset(dataset$D9IncYes1,dataset$D9IncYes1=="Yes"))

D9IncAll

## [1] 227

D9IncYes

## [1] 68

prop.test(D9IncYes,D9IncAll)

## 
##  1-sample proportions test with continuity correction
## 
## data:  D9IncYes out of D9IncAll, null probability 0.5
## X-squared = 35.683, df = 1, p-value = 2.322e-09
## alternative hypothesis: true p is not equal to 0.5
## 95 percent confidence interval:
##  0.2416672 0.3643670
## sample estimates:
##         p 
## 0.2995595
```

The output shows that there are 227 companies that responded to question D9IncYes1, of which 68 answered “yes - they did participate in an incubator programme”. This results in a proportion of approximately 30% with a 95% confidence interval between 24,2% and 36,4%. In other words, we are 95% confident that the true proportion of survey participants that participated in an incubator programme lies between 24,2% and 36,4%.

2.2 Satisfaction with product development process

Analyse the satisfaction with the product development process (C5) using a barchart.

```r
barplot(table(dataset$C5), main="Barchart of Satisfaction",
xlab="Satisfaction", ylab="Frequency")
```



The barchart shows in this case a histogram, the frequency distribution of product development process satisfaction of all participants. The plot depicts the comparison of the satisfaction ratio from Strongly Disagree to Strongly Agree in 7 levels. The frequency distribution leans towards Strongly Agree. The highest frequency with 60 participants are Strongly Agree while the lowest frequency with only 13 participants Strongly Disagree with the satisfaction of their product development process. We can conclude here, that most are at least satisfied by their product development process.

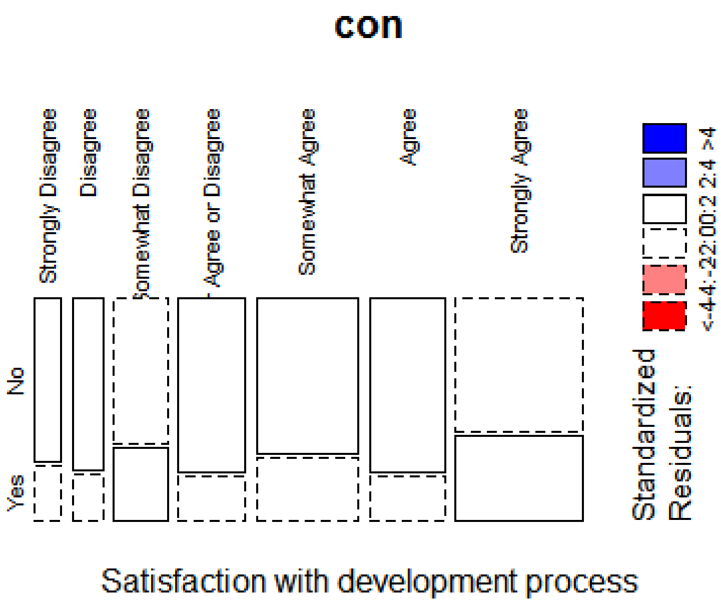
2.3 Association of incubator participation and satisfaction with the product development process

Does the satisfaction with the product development process differ between participants and non-participants in incubator programs? Create a corresponding contingency table, a mosaicplot and test for independence.

```
con = table(dataset$C5, dataset$D9IncYes1)
con

##
##                               No Yes
##   Strongly Disagree            9   3
##   Disagree                    11   3
##   Somewhat Disagree           16   8
##   Neither Agree or Disagree 24   6
##   Somewhat Agree              33  13
##   Agree                       27   7
##   Strongly Agree              35  22
```

The contigency table shows the distribution of the answers to the following question: “We are satisfied with our product development process.”

```
mosaicplot(con, shade = TRUE, las=3, xlab="Satisfaction with development 
process")
```



From the mosaic plot, and the contigency table no obvious differences in satisfaction with the product development process between incubator participant and non-particpants compared to the general satisfaction shown in 2.2 can be seen.

```
chiresult <- chisq.test(dataset$C5, dataset$D9IncYes1)
chiresult

## 
##  Pearson's Chi-squared test
## 
## data:  dataset$C5 and dataset$D9IncYes1
## X-squared = 5.6426, df = 6, p-value = 0.4644

cV <- CramerV(con)
cV

## [1] 0.1612541
```

As p-value of the chi-squered test of 0.4643891 is greater than the 0.05 significance level, we do not reject the null hypothesis that the satisfaction of the product development process is independent on the incubation program. According to the documentation, a Cramer’s V in the range of [0, 0.3] is considered as weak and the result of a value of 0.1612541 is therefore considered as a weak. To conclude the two results, even if there is an dependency of the satisfaction of the product development process with the incubation program, this dependency seems to be weak.

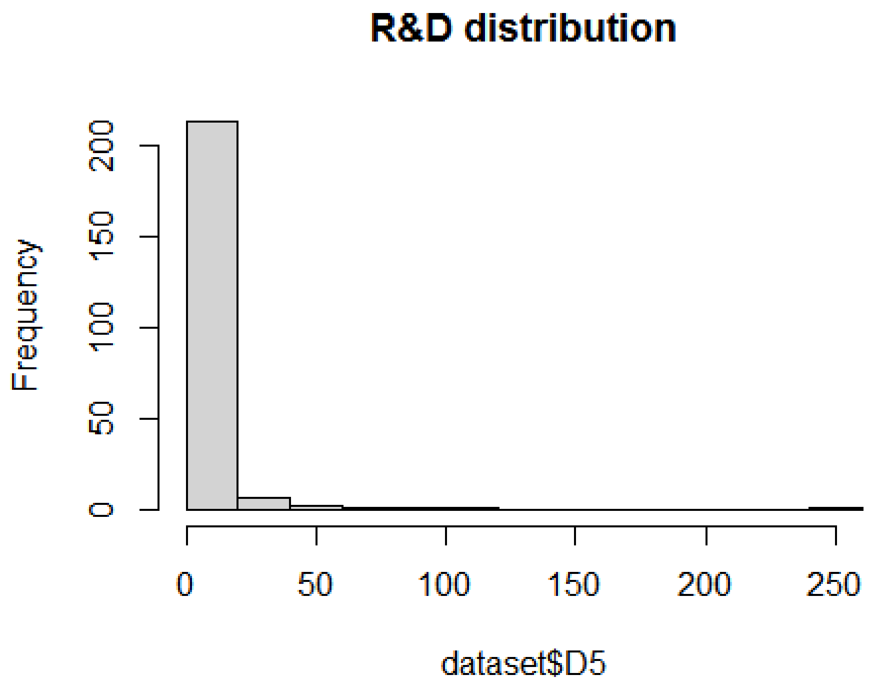
2.4 Number of R+D personnel

2.4a Graphical distribution of R&D personnel.

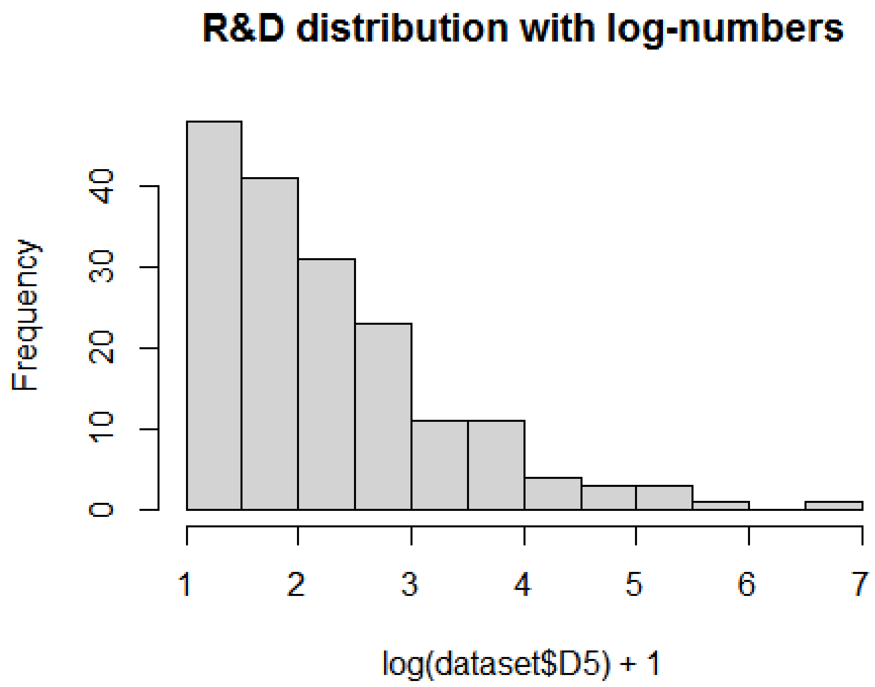
Quantile-Quantile Plot

Create a Histogram to compare the R&D personnel of incubator and non-participants.

```
hist(dataset$D5, main="R&D distribution")
```



```
hist(log(dataset$D5)+1, main="R&D distribution with log-numbers")
```

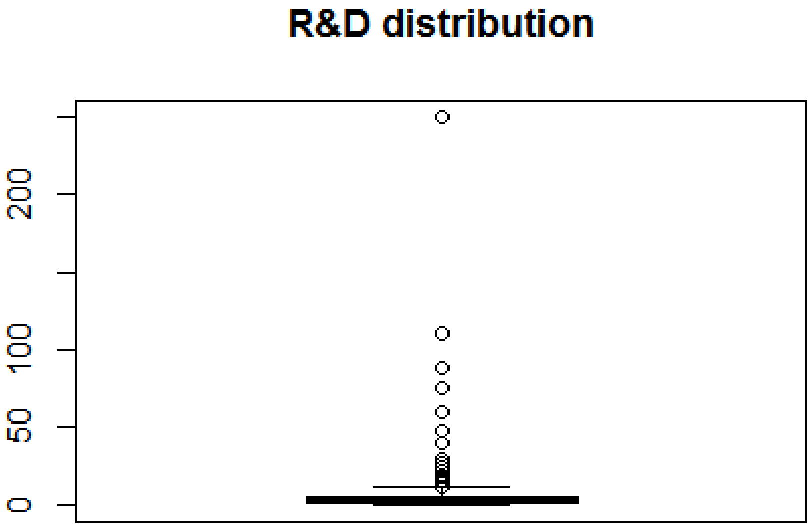


To analyze the distribution of R&D personnel, a Histogram is used. As visible, most companies have very few R&D personnel compared to the companies with the most personnel. Most have around 1 to 3 employees, with a heavy focus on 1. Only very few companies have more employees, however, this goes up to 250 R&D employees. The right-skew of the data is to extreme to receive a normal distribution even when applying the log-transformation.

Boxplot

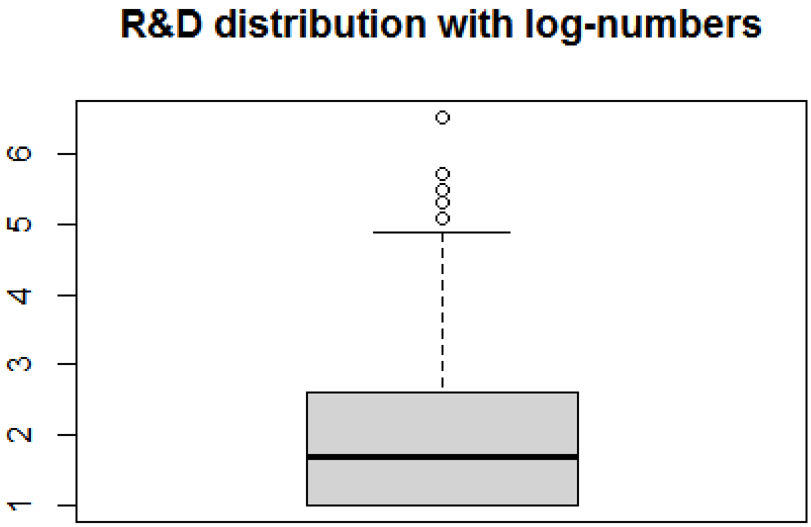
Create a Boxplot to compare the R&D personnel of incubator and non-participants.

```
boxplot(dataset$D5, main="R&D distribution")
```



Since the data might not be normally distributed it is diffcult to read the boxplot. Therefore a log-transformed boxplot is created.

```
boxplot(log(dataset$D5)+1, main="R&D distribution with log-numbers")
```



In the non-log based boxplot, the distribution is extremely shifted and makes if very difficult to read the plot. However, we are easily able to identify that the majority of companies only have very few R&D employees and only few have a lot of R&D personnel.

2.4b Bowley Skeweness

Calculate the Bowley Skeweness to see if the curve is skewed.

```
## Quartile calculation for Bowley Skeweness
Q1 <- quantile(dataset$D5, prob=c(.25), na.rm=TRUE)
Q2 <- quantile(dataset$D5, prob=c(.5), na.rm=TRUE)
Q3 <- quantile(dataset$D5, prob=c(.75), na.rm=TRUE)

bowley <- ((Q3 - Q2) - (Q2 - Q1))/(Q3 - Q1)
bowley

## 75%
## 0.5
```

The Bowley Skeweness tells us, if the curve is symmetrical, positively skewed or negatively skewed. The following numbers can be used for orientation:

- Skewness = 0 means that the curve is symmetrical.
- Skewness > 0 means the curve is positively skewed.
- Skewness < 0 means the curve is negatively skewed.

We are calculating the Bowley Skeweness with the formula [(Q3 - Q2) - (Q2 - Q1)] / (Q3 - Q1). The result of the skeweness is 0.5 and as it is above 0, we conclude that the curve is positively skewed. The calculation of the Bowley Skeweness underlines that we are not dealing with normally distributed data but with a positive or right-skewed sample.

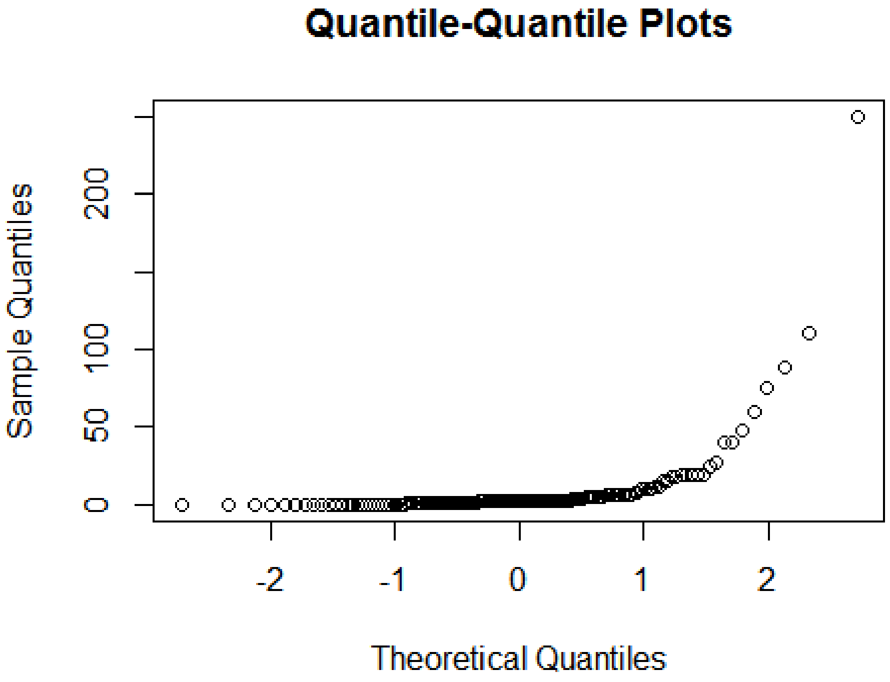
2.5 R+D personnel vs participation

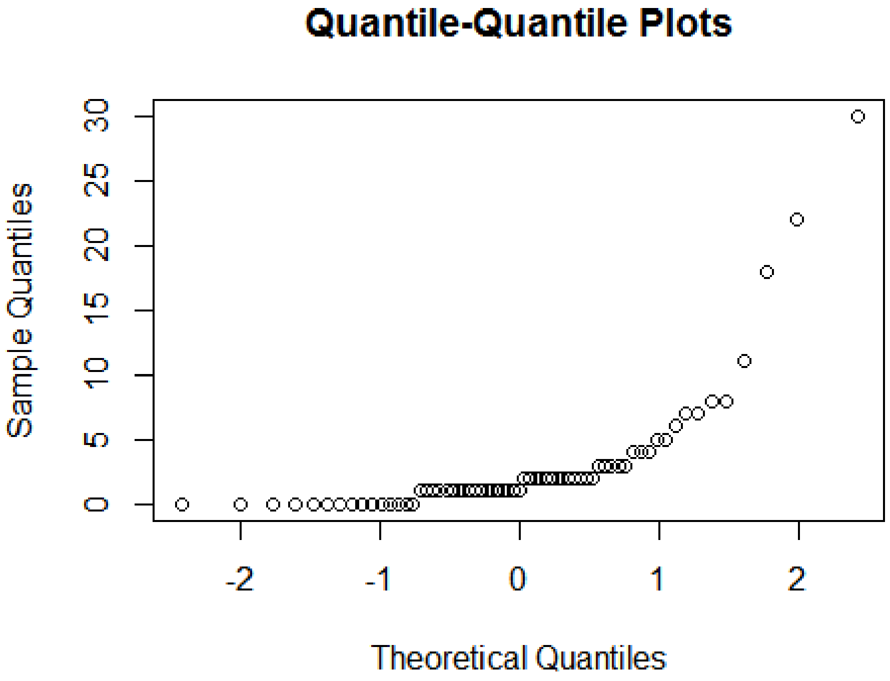
2.5a Graphical distribution of R&D personnel between incubators and non-participants.

Quantile-Quantile Plot

Create a Normal plot to compare the R&D personnel of incubator and non-participants.

```
### QQPlot without NA's
test = aggregate(dataset$D5, by=list(dataset$D9IncYes1), FUN="qqnorm",
na.action = na.omit, main="Quantile-Quantile Plots")
```



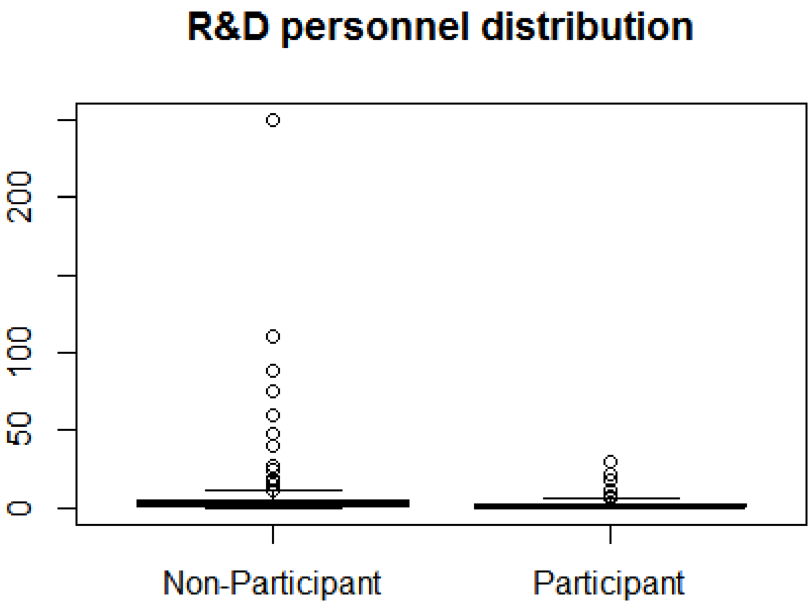


The U-shape of the Q-Q-plots indicate that we are not dealing with a normal distribution, neither in incubator participants nor in non-participants.

Boxplot

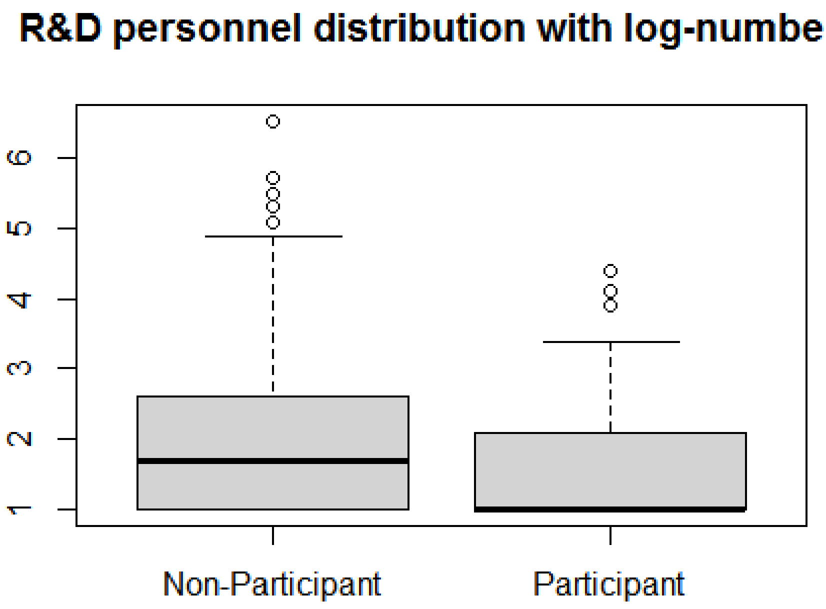
Create a grouped boxplot to compare the R&D personnel of incubator and non-participants.

```
# Boxplot outliers are not drawn
boxplot(test$x[1,]$y, test$x[2,]$y, names = c("Non-Participant",
"Participant"), main="R&D personnel distribution")
```



Since the data is not normally distributed it is diffcult to read the boxplot. Therefore a log-transformed boxplot is created.

```
boxplot(log(test$x[1,]$y)+1, log(test$x[2,]$y)+1, names = c("Non-
Participant", "Participant"), main="R&D personnel distribution with log-
numbers")
```



From the log-transformed boxplot it can be seen that number of R&D staff seems to be slightly higher in non-participants of the incubator programs.

2.5b Hypothesis tests

Shapiro-Test

Calculate the Shapiro-Test to know if the T-Test or the Wilcoxon-Test is applicable.

```
## Shapiro-Test
# P-value above 0.05 or 5% means it is normally distributed, the Null
hypothesis is not rejected and the t-test can be used. Else the Wilcoxon-test
has to be used

shapiro <- shapiro.test(dataset$D5[which(dataset$D9IncYes1==c("No","Yes"))])
shapiro

##
##  Shapiro-Wilk normality test
##
## data:  dataset$D5[which(dataset$D9IncYes1 == c("No", "Yes"))]
## W = 0.48234, p-value < 2.2e-16
```

With p-value < 2.2e-16, the Shapiro test is below 0.05 or 5%, which means the null hypothesis has to be rejected and the t-test is not seen as an applicable test, since it is not applicable on data that is not normally distributed. Instead we recommend to use the Wilcoxon-Test, which is able to handle such situations.

T-Test

Calculate the T-Test

```
## T-Test
# The Null hypothesis shows at higher than 0.05 or 5%  that both populations
are significant differently
ttest <- t.test(test$x[1,]$y, test$x[2,]$y) # H0 auch rejected
ttest

##
##  Welch Two Sample t-test
##
## data:  test$x[1, ]$y and test$x[2, ]$y
## t = 2.5316, df = 175.88, p-value = 0.01223
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.190456 9.610570
## sample estimates:
## mean of x mean of y
##  8.446667  3.046154
```

When we apply the two sample t-test despite knowing, we should not use it (see section “T-Test”), we receive a p-value of p-value = p-value = 0.01223. It lies below 0.05 or 5%, this indicates that the null hypothesis should be rejected and the two populations are significantly different.

Wilcoxon-Test

Calculate the Wilcoxon-Test

```
## Wilcoxon-Test
# The Wilcoxon-Test can be used when there is no normal distribution. The
Null hypothesis shows at higher than 0.05 or 5%  that both populations are
significant differently
wilcox.test(test$x[1,]$y, test$x[2,]$y, alternative="two-sided")

##
##  Wilcoxon rank sum test with continuity correction
##
## data:  test$x[1, ]$y and test$x[2, ]$y
## W = 5775.5, p-value = 0.02956
## alternative hypothesis: true location shift is not equal to 0
```

The Wilcoxon-Test helps to analyse data which is not normally distributed. Applied we receive the p-value = 0.02956, which means that we reject the null hypothesis, as the values lies below 0.05 or 5%. This means that amount of R&D staff for non-participants of the incubator program is significantly higher than for the participants. Even though, the data is not normally distributed the t-test still gives the correct result.

THE END!

Thank you, Beat & Fabian, for your good explanations.