Lab 3 - Group 16

Eric Börjesson (ericbor@student.chalmers.se)
Ludvig Ekman (eludvig@student.chalmers.se)
Wenli Zhang (wenliz@student.chalmers.se) Tim Grube (gusgruti@student.gu.se)

11.12.2020

Design an experiment

1. Description of context and problem

The Covid pandemic has changed the way of working for many companies and their employees. With social distancing and restrictions being implemented by governments and companies, employees are now working from home more than ever before. Even before the pandemic, experiments have been conducted in order to understand the effects of having employees working from home rather in the office. Bloom et al. (2014) published a study where they randomly assigned employees to either work in the office or at home. The experiment indicated that working from home led to an increase in work performance, and the positive effect of working from home has also been discussed in other studies (Hill, Ferris, and Märtinson 2003).

The concept of technical debt within the software engineering field is recognized as something that is unavoidable and it is therefore necessary for companies to track, monitor and take action upon the technical debt (Lim, Taksande, and Seaman 2012). This experiment is intended to combine these two research topics and investigate if the work location influences the technical debt of software artefacts.

2. General objectives of the experiment

The objective of this experiment is to compare the quality of software artifacts produced by teams working in different locations. The aim is to investigate if the work place influences the quality of the code.

3. Design of the experiment

The experiment will consist of workplace being a factor and the experience of the team being a blocking variable. The factor workplace will be composed of three levels: everyone working from home (Home), a mixture of the team working from home and from the office (Mix), everyone working in the office (Office).

The experience of the team is believed to effect the outcome of the experiment, but this influence and variation is not of main interest in our study and thus as a blocking variable. The blocking variable is measured as the average time each team member has worked as a software engineer and it has two levels: less experienced (LE) and more experienced (ME). The half of the teams with lower experience are placed in the group less experienced (LE), and the half with more experience are placed in the group more experienced (ME).

The experiment will have two controlling variables. The first is to ensure that the teams have the same workload and the second is to control the work hours and breaks, and thus ensuring consistency in work time.

The experiment will be a block-design experiment with the following 6 experimental groups:

- G1: Workplace = Office, Experience = LE
- G2: Workplace = Office, Experience = ME

- G3: Workplace = Home, Experience = LE
- G4: Workplace = Home, Experience = ME
- G5: Workplace = Mix, Experience = LE
- G6: Workplace = Mix, Experience = ME

The population consists of software engineering teams from the company ABC, where the teams are working within various business areas of the company. We aim at achieving a balanced design, so it is therefore necessary to first stratify the population based on the experience level of the team. After this, an equal amount of less and more teams can be randomly selected and the three treatments can be evenly distributed within each.

4. Power analysis

We have used the *pwr* package in R to conduct a power analysis. The package is based on the concepts described by Cohen (Cohen 1988). According to Cohen, the five parameters power, significance criterion, degrees of freedom for numerator, degrees of freedom for denominator, and effect size, are so tightly related so that if four of them are fixed the fifth can be given. We have used the *pwr::pwr.f2.test* method to get the estimated effect size, as shown below.

Base on the experiment background we have 6 groups, so the value of u(degrees of freedom for the numerator) is 5(6-1). Compared to the control(LE, Office), we assume that the factor of workplace would explain 20% influence on the outcome(Technical Debt), so effect size is 0.25(f2=0.2/1-0.2) with a confidence of 95% (sig.level=0.05) and has the power of 90% (power=0.09) to know the more efficient workplace and run experiments.

```
pwr::pwr.f2.test(u=5, f2=0.2/0.8, power=0.9)
```

```
##
##      Multiple regression power calculation
##
##               u = 5
##               v = 65.64045
##              f2 = 0.25
##       sig.level = 0.05
##           power = 0.9
```

From the result, we can deduce that the sample size should be at least 72 teams(66+5+1) and 72 teams will be divided into 6 groups to be able to have a statistically significant model, so each group should have 12 teams to conduct the experiment.

5. Independent variables

The experiment has one independent variable, namely the workplace of the team. It is a categorical variable with three categories: home, mix and office.

6. Dependent variables

The dependent variable of the experimental design is the amount of technical debt. Technical debt is measured in time and is thus measured on a continuous scale.

7. Specific experiment hypotheses

The design includes one null and one alternative hypothesis:

- h0: The work location has **no** effect on the technical debt.
- h1: The work location has **an** effect on the technical debt.

8. Data collection procedure

SonarQube is a tool used for continuous inspection of code quality and includes several metrics such as technical debt. It is the tool to be used to measure technical debt of the software artifacts produced by the teams.

9. Plan for analysis of data to answer hypotheses

As we have only one factor of the workplace and there are same numbers of observations in each group we will do a balanced one-way ANOVA to determine whether there is a significant difference between the groups.

10. Possible threats to validity and replication considerations

Validity. We have focused on the four main types of validity threats discussed by Wohlin et al. (2012):

1. **Conclusion validity**. The conclusion validity is strongly related to analyzing the outcome and determining if there is a significant effect of the factor. This could be determined by using the different tools provided by R and supporting packages and functions. However, we must analyze the outcome together with the assumptions (such as chosen power, effect size and significance level) to make appropriate conclusions.
2. **Construct validity**. This aspect is related to the generalization of the results to the theory behind it and is connected to how well the experiment measures what is intended. One threat related to this is the fact that the behavior of the subjects participating in the experiment might be affected just by the fact that they are part of the study. It is a valid threat in our case since it is quite hard to apply the treatment in this study without them being aware of it.
3. **Internal validity**. If we observe a significant effect of a treatment on the outcome, we might still not be sure that the observed factor was the actual cause. There might exist other factors that we have not included or controlled that effect the outcome. One example could be that employees working from home do not need to commute to work and therefore are less stressed at the beginning of their work day.
4. ***External validity**. This is described in the replication section below, but summarizing it can be hard to repeat the experiment since there are many specific conditions that can not be measured or controlled in our experiment.

Replication. Replication is an extremely important aspect since a new knowledge is generally not accepted until it has been repeated and verified by external agents (Juristo and Moreno 2013). According to Juristo and Moreno (2013), there are two types of replication: External replication where the experiment is repeated by independent researchers and internal validation which is run within the experiment itself. A possible threat to both internal and external validations is the problem of similarity between repeated experiments. Replication of an experiment aims at repeating the experiment under specific preconditions. In our experiment, there are multiple elements and conditions, such as the team members, the tasks and experience levels which are very specific and thereby hard to replicate.

Data generation and analysis

Data generation

First, a population size of 2000 is generated.

```
N <- 2000
```

Then we designed the model. “Less experienced” and “Office” is our reference group. We took the value 5 as our expected mean for technical debt that exists at the end of the experiment.

As the main effect we set that working from home decreases the amount of technical debt, whereas a mixed team increases the technical debt. The experience has a slightly negative effect.

```
model <- function(N, X_experience, X_workplace){

  ref <- 5 #LE, Office (reference group)

# Main effects
 experience_ME <- -0.1
 workplace_Home <- -0.3
 workplace_Mix <- 0.7

# Input
 x_experience_ME <- X_experience[1]
 x_workplace_Home <-  X_workplace[1]
 x_workplace_Mix <- X_workplace[2]

 response_std <- 1.0

#Linear model that controls the response variable
 techdebt <- ref +
   experience_ME*x_experience_ME +
   workplace_Home*x_workplace_Home +
   workplace_Mix*x_workplace_Mix

# Generate normal distribution
 techdebt_out<- rnorm(N, mean=techdebt, sd = response_std)

 return(techdebt_out)
}
```

As described in the experiment design, we have 6 different combinations of experience and workplace, so we have 6 groups.

```
set.seed(8123)

experience_LE <- c(0)
experience_ME <- c(1)

workplace_Office <- c(0,0)
workplace_Home <- c(1,0)
workplace_Mix <- c(0,1)

g1 <- data.frame(
Experience = rep('LE', N),
Workplace = rep('Office', N),
techdebt=model(
  N=N,
  X_experience = experience_LE,
  X_workplace = workplace_Office
  )
)

g2 <- data.frame(
Experience = rep('ME', N),
Workplace = rep('Office', N),
techdebt=model(
```

```
  N=N,
  X_experience = experience_ME,
  X_workplace = workplace_Office
  )
)

g3 <- data.frame(
Experience = rep('LE', N),
Workplace = rep('Home', N),
techdebt=model(
  N=N,
  X_experience = experience_LE,
  X_workplace = workplace_Home
  )
)

g4 <- data.frame(
Experience = rep('ME', N),
Workplace = rep('Home', N),
techdebt=model(
  N=N,
  X_experience = experience_ME,
  X_workplace = workplace_Home
  )
)

g5 <- data.frame(
Experience = rep('LE', N),
Workplace = rep('Mix', N),
techdebt=model(
  N=N,
  X_experience = experience_LE,
  X_workplace = workplace_Mix
  )
)

g6 <- data.frame(
Experience = rep('ME', N),
Workplace = rep('Mix', N),
techdebt=model(
  N=N,
  X_experience = experience_ME,
  X_workplace = workplace_Mix
  )
)
```

For the first assignment, we take the calculated sample size of 72 which is divided into 6 groups with 12 teams each - so a balanced design:

```
n <- 12
all_groups <- list(g1, g2, g3, g4, g5, g6)

#Creating an empty data frame from an existing one (with the same column)
d<-g1[0,]
```

```
for(g in all_groups) {
 d<-rbind(d, dplyr::sample_n(g, size=n)) #Appending rows at the end for every group
}

d$Experience <-as.factor(d$Experience)
d$Workplace <- as.factor(d$Workplace)

#Set correct levels
levels(d$Experience) <- c("LE", "ME")
levels(d$Workplace) <- c("Office", "Home", "Mix")
```

Then we created a linear model. The fitted model looks like this:

y = 5 + 0.2503 * ExperienceME + 1.1963 * WorkplaceHome + 0.2264 * WorkplaceMix

We expected the coefficient for WorkplaceHome to be negative. However, it is strongly positive and even higher than for WorkplaceMix. The model obviously has a problem here.

```
m1 <- lm(techdebt ~ Experience + Workplace, data=d)
summary(m1)
```

```
##
## Call:
## lm(formula = techdebt ~ Experience + Workplace, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.41591 -0.49031 -0.02615  0.64116  2.42732
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.3074     0.2286  18.844  < 2e-16 ***
## ExperienceME    0.2503     0.2286   1.095    0.277
## WorkplaceHome   1.1963     0.2800   4.273 6.14e-05 ***
## WorkplaceMix    0.2264     0.2800   0.809    0.421
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9698 on 68 degrees of freedom
## Multiple R-squared:  0.2428, Adjusted R-squared:  0.2094
## F-statistic:  7.27 on 3 and 68 DF,  p-value: 0.0002667
```

This can also be seen, when checking whether the true coefficients are within the confidence interval. This is the case for ExperienceME and WorkplaceMix, but not for WorkplaceHome (as already mentioned, the WorkplaceHome coefficient is strongly positive instead of negative).

```
confint(m1, level = 0.95)
```

```
##                    2.5 %    97.5 %
## (Intercept)    3.8513057 4.7635643
## ExperienceME  -0.2058080 0.7064505
## WorkplaceHome  0.6376561 1.7549401
## WorkplaceMix  -0.3322192 0.7850648
```

Now the same is done for half the sample size:

```r
n <- 6
all_groups <- list(g1, g2, g3, g4, g5, g6)

#Creating an empty data frame from an existing one (with the same column)
d<-g1[0,]
for(g in all_groups) {
 d<-rbind(d, dplyr::sample_n(g, size=n)) #Appending rows at the end for every group
}

d$Experience <-as.factor(d$Experience)
d$Workplace <- as.factor(d$Workplace)

#Set correct levels
levels(d$Experience) <- c("LE", "ME")
levels(d$Workplace) <- c("Office", "Home", "Mix")
```

This time, the fitted model looks like this:

y = 5 + 0.5381 * ExperienceME + 1.6355 * WorkplaceHome + 0.4346 * WorkplaceMix

The problems from before still exist.

```r
m2 <- lm(techdebt ~ Experience + Workplace, data=d)
summary(m2)
```

```
##
## Call:
## lm(formula = techdebt ~ Experience + Workplace, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7290 -0.5223  0.1544  0.4847  1.7432
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.2497     0.3156  13.464 9.97e-15 ***
## ExperienceME    0.5381     0.3156   1.705 0.097956 .
## WorkplaceHome   1.6355     0.3866   4.231 0.000182 ***
## WorkplaceMix    0.4346     0.3866   1.124 0.269242
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9469 on 32 degrees of freedom
## Multiple R-squared:  0.4087, Adjusted R-squared:  0.3532
## F-statistic: 7.372 on 3 and 32 DF,  p-value: 0.000687
```

As before, when checking the true coefficients, the coefficients for ExperienceME and WorkplaceMix are in the confidence interval, WorkplaceHome is not.

```r
confint(m2, level = 0.95)
```

```
##                   2.5 %   97.5 %
## (Intercept)    3.6067807 4.892668
## ExperienceME  -0.1048900 1.180997
## WorkplaceHome  0.8481034 2.422987
## WorkplaceMix  -0.3528015 1.222082
```

The main difference between the two differently powered experiments is the different explanatory power. This is for example shown by an increased standard error for all coefficients.

Conclusion

The Anova table shows a F value of 10.305 for the workplace. The critical F value is 2.3933. This means that there is a significant effect of the workplace.

```
Anova(m1)
```

```
## Anova Table (Type II tests)
##
## Response: techdebt
##             Sum Sq Df F value    Pr(>F)
## Experience  1.128  1  1.1992 0.2773354
## Workplace  19.384  2 10.3054 0.0001232 ***
## Residuals  63.954 68
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The Tukey HSD post-hoc supports this. It shows that there are significant differences between Home and Office as well as between Home and Mix. There is not a big difference between Office and Mix. Summarized, if the whole team works from home, a lower amount of technical debt is usually achieved.

```
TukeyHSD(aov(techdebt ~ Workplace + Experience, data=d))
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = techdebt ~ Workplace + Experience, data = d)
##
## $Workplace
##                    diff        lwr        upr     p adj
## Home-Office  1.6355453  0.6855695  2.5855212 0.0005220
## Mix-Office   0.4346404 -0.5153355  1.3846163 0.5064195
## Mix-Home    -1.2009049 -2.1508808 -0.2509291 0.0107299
##
## $Experience
##             diff      lwr      upr     p adj
## ME-LE 0.5380537 -0.10489 1.180997 0.0979562
```

Pre-registration

Our pre-registering experiment can provide a guideline for researchers and students about how to design a standardized experiment for a problem in software engineering field. We design an experiment intending to combine the workplace and technical debt topics and then investigate if the work location influences the technical debt of software artefacts. So if they want to dig into those topics they can refer to our design and then conduct the experiment.

All members in our group had been added in the experiment design and we all approve the registration of our experiment in the *Open Science Framework platform*. Here is the link: https://osf.io/qd942.

Contributions

All four of us attended the lab. During the lab, we started discussing the design of the experiment. We started by creating a draft registration in *osf.io* which helped guiding us through the first steps of the design.

Eric continued with writing and completing the report regarding the experiment design. Ludvig completed the online pre-registration. Wenli, Ludvig and Tim were responsible for simulating the data and analyzing it. At the end, everybody read through the whole document, so that everybody knows about all parts of the lab.

References

Bloom, Nicholas, James Liang, John Roberts, and Zhichun Jenny Ying. 2014. " Does Working from Home Work? Evidence from a Chinese Experiment *." *The Quarterly Journal of Economics* 130 (1): 165–218. https://doi.org/10.1093/qje/qju032.

Cohen, Jacob. 1988. *Statistical Power Analysis for the Behavioral Sciences*. Academic press.

Hill, E Jeffrey, Maria Ferris, and Vjollca Märtinson. 2003. "Does It Matter Where You Work? A Comparison of How Three Work Venues (Traditional Office, Virtual Office, and Home Office) Influence Aspects of Work and Personal/Family Life." *Journal of Vocational Behavior* 63 (2): 220–41.

Juristo, Natalia, and Ana M Moreno. 2013. *Basics of Software Engineering Experimentation*. Springer Science & Business Media.

Lim, E., N. Taksande, and C. Seaman. 2012. "A Balancing Act: What Software Practitioners Have to Say About Technical Debt." *IEEE Software* 29 (6): 22–27. https://doi.org/10.1109/MS.2012.130.

Wohlin, Claes, Per Runeson, Martin Höst, Magnus C Ohlsson, Björn Regnell, and Anders Wesslén. 2012. *Experimentation in Software Engineering*. Springer Science & Business Media.