

```
---
title: "Report Template coursework assignment A – 2021"
author: "Thomas Bos (4543408), Daniël van Gelder (4551028), Jessie van
Schijndel (5407397)"
date: "20/04/2021"
output:
  pdf_document:
    fig_caption: yes
    number_sections: yes
  word_document: default
subtitle: CS4125 Seminar Research Methodology for Data Science
---
```

```
```${r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```
```

```
\tableofcontents
```

```
# Part 1 – Design and set-up of true experiment
```

```
## The motivation for the planned research
```

```
<!-- (Max 250 words) -->
```

The coronavirus pandemic has had a great impact on many aspects of society. University education, in particular, has changed significantly. As education in many countries has shifted from physical lectures to online teleconferencing lectures, concerns have been raised with regards to the effectiveness of this method of education. While the technological developments surrounding teleconferencing have enabled an almost seamless transition from offline to online education, it may be that the lack of a physically present lecturer affects the comprehensibility of the lecture material for students. With this research, we aim to address whether the students' understanding of the lecture material is affected by a different learning setting (i.e., from home watching an online lecture). The results may reveal whether online education is a way to move forward out of the pandemic. Moreover, if the results indicate no significant change in student understanding of material it may open up the way for new form of education, where students could enroll into "digital universities" without needing to be present at any time.

```
## The theory underlying the research
```

```
<!-- (Max 250 words) Preferable based on theories reported in
literature-->
```

```
<!-- I am not really covering how the experiments should be done, this
may be important to add? Currently ~210 words.-->
```

Figlio et al. (2013) presented, according to them, the first experimental evidence on the effects of live versus online instruction. In this research, participants took an entire microeconomics course either only attending live lectures or online lectures. Exam performance was then compared between both groups and all students which did not volunteer to participate in the experiment but did still follow the course. Result showed that there is a modest difference in exam scores in favour of the students only attending live lectures, although the authors state that the experiments had many limitations and that further research is necessary. In contrary, a more recent survey by Nguyen (2015), which summarizes results of multiple studies, has found that 92% makes online education to be at least as effective, if not better, than live education. However, it is also important to recognize other issues that may arise when switching teaching modalities, which becomes clear when such a shift is forced due to, for example, the onset of COVID-19. In a very recent study by Finnegan (2021), results showed that while results are marginally worse after the shift to online teaching, student experience has deteriorated when their learning environment is suddenly changed, especially with students with poor online access.

## ## Research questions

<!--

The research question that will be examined in the experiment (or alternatively the hypothesis that will be tested in the experiment)

-->

Our research question is the following: "How is students' understanding of lecture material affected by attending the lecture live rather than online?". We describe our null hypothesis and alternative hypothesis in the section on suggested statistical analyses.

## ## The related conceptual model

This model should include:

- \*Independent variable(s)
- \*Dependent variable
- \*Mediating variable (at least 1)
- \*Moderating variable (at least 1)

The following sections describe the conceptual model for each type of variable:

### ### Independent Variable (IV)

The IV of this research is whether the participant (student) attends the lecture physically or from home through online teleconferencing.

### ### Dependent Variable (DV)

The DV of this research is the relative score increase on the test that students make. Before the experiment the participants make a small test regarding the lecture material for which the score is expected to be low as the participants are expected to have no prior knowledge regarding the material. Then after the lecture the students make the same test regarding the lecture material. The relative increase (or unlikely decrease) of score will be the DV.

### ### Mediating Variable

As the students perform the test in a different setting (from home or on campus) depending on the IV. The change in setting is expected to have a mediating effect on the relationship between the IV and DV.

### ### Moderating Variable

There are several factors which may have a moderating effect on the relationship between the IV and the DV which are difficult to control on the experiment. These mostly have to do with the environment in which the lecture is attended. The following list describes the specific variables which are believed to have this moderating effect:

- (online lecture) video/audio quality
- (online lecture) device that is used to attend lecture (e.g. laptop, tablet, smartphone)
- (both physical and online lecture) presence of noise and/or distraction in environment of watching lecture

### ## Experimental Design

<!-- Note that the study should have a true experimental design -->

<!--

- Multiple lectures? (Different group sizes, different topics, etc.)
- How to place the online group? (isolated in cubicles, Drebbelweg Tentamenzaal, etc.)

-->

In order to determine the difference between live and online lectures on students with respect to acquired knowledge the experimental design Pre-test Post-test randomized controlled trial was chosen. This means the participants can be tested before and after the lecture so that the difference in test results, the dependent variable, can be used as an indicator of knowledge gained from said lectures. For the lecture itself, the participants will be divided randomly over live and online groups such that the live group will attend a lecture face-to-face with a lecturer, and the online group will attend the lecture via an online platform such as Zoom. In order to minimize the influence of moderating variables such as video/audio quality and distractions, the online group will watch the lecture in a quiet, moderated environment on identical systems specifically set up for the experiment.

### ## Experimental procedure

<!--

Describe how the experiment will be executed step by step

-->

First, we ask all students in the class who have agreed to participate in our experiment to perform a pre-test a day before the lecture. The pre-test will consist of questions composed by the teacher giving the lecture. The questions should reflect the main learning goals of the lecture. Ideally, this pre-test is done in a controlled setting on campus. If this is not possible due to governmental restrictions, the pre-test is performed online. All students perform the pre-test at the same time. After the pre-test, students are assigned to either the live lecture condition or the online lecture condition. To reduce unexplained variability, we will opt for a randomized block design. We will divide similar participants into blocks based on their pre-test scores. Then, we randomly assign participants from each block to the live condition or the online condition. Students in both conditions will follow the same lecture at the same time. A day after the lecture, the students perform a post-test. Just like the pre-test, the post-test will consist of questions composed by the teacher giving the lecture and should reflect the main learning goals of the lecture. However, the questions from the pre-test should not be repeated. Again, this post-test is ideally done in a controlled setting on campus, but may have to be performed online.

### ## Measures

In the experiment, both participant groups will take a pre-test and a post-test. This test aims to evaluate the participants' comprehension of the lecture material. The pre-test is meant to serve as a baseline measurement to rule out any pre-existing knowledge of the participants. Both tests will be identical and will be in the form of a multiple choice exam of ten questions to be taken in a short time span (10 minutes). The score of the test is defined as the proportion of correct answers. The measure of the experiment is the ratio between these two tests for each participant: the score of the post-test divided by the score of the pre-test.

### ## Participants

<!-- Describe which participants will recruit in the study and how they will be recruited -->

<!-- What to add? -->

Participants should be students and could be recruited by asking for volunteers across a university campus. A small compensation could be offered in return as a sign of appreciation.

### ## Suggested statistical analyses

<!--

Describe the statistical test you suggest to carry out on the collected data

-->

First, we determine our null hypothesis  $H_0$  and alternative hypothesis  $H_1$ . Our null hypothesis states that there is no difference in student understanding of the lecture material between the two different conditions. Our alternative hypothesis states that there is a difference in student understanding. We create two linear models to predict student understanding of lecture material. First, we create a model which has only an intercept. This model does not use the information about which condition a participant was in. This model will be referred to as  $m_0$ . Second, we create a model which does include this information as a predictor. This model will be referred to as  $m_1$ . Then, we compare the fits of the two models to the data. We determine whether  $m_1$  fits significantly better than  $m_0$  through an ANOVA F-test. If this is not the case, we cannot reject our null hypothesis. We may also inspect the significance of the parameters of  $m_1$ . If the effect of the condition parameter is not significant, we cannot reject our null hypothesis.

## # Part 2 - Generalized linear models

### ## Question 1 Twitter sentiment analysis (Between groups - single factor)

#### ### Conceptual model

```
<!-- Make a conceptual model for the following research question: Is
there a difference in the sentiment of the tweets related to the
different celebrities? -->
```

```
![The conceptual model. Different attributes of a celebrity are shown
which may influence the sentiment of tweets related to a certain
celebrity.](ConceptualModel2.1.pdf)
```

#### ### Collecting tweets, and data preparation

```
<!-- Include the annotated R script (excluding your personal Keys and
Access Tokens information), but put echo=FALSE, so code is not included
in the output pdf file. -->
```

```
<!-- TODO: Dit willen we misschien dubbel checken voordat we hem
inleveren -->
```

Note from the author: as this data is updated with each run, analyses may not reflect accompanying graphs in certain sections as they were performed on older data.

```
`r, echo=FALSE, message=FALSE, warning=FALSE, include = FALSE}
```

```
#during writing you could add "eval = FALSE", knitr will than not run
this code chunk (take some time do)
```

```
source("working_dir.R")
```

```
setwd(wd)
```

```
# apple , note use / instead of \, which used by windows
```

```
#install.packages("twitter", dependencies = TRUE)
```

```
library(twitterR)
```

```
#install.packages("RCurl", dependencies = T)
```

```
library(RCurl)
```

```
#install.packages("bitops", dependencies = T)
```

```
library(bitops)
```

```
#install.packages("plyr", dependencies = T)
```

```
library(plyr)
```

```
#install.packages('stringr', dependencies = T)
```

```
library(stringr)
```

```

#install.packages("NLP", dependencies = T)
library(NLP)
#install.packages("tm", dependencies = T)
library(tm)
#install.packages("wordcloud", dependencies=T)
#install.packages("RColorBrewer", dependencies=TRUE)
library(RColorBrewer)
library(wordcloud)
#install.packages("reshape", dependencies=T)
library(reshape)

##### functions

clearTweets <- function(tweets, excl) {

  tweets.text <- sapply(tweets, function(t)t$getText()) #get text out of
tweets

  tweets.text = gsub('[:cntrl:]', '', tweets.text)
  tweets.text = gsub('\\d+', '', tweets.text)
  tweets.text <- str_replace_all(tweets.text,"^[[:graph:]]", " ") #remove
graphic

  corpus <- Corpus(VectorSource(tweets.text))

  corpus_clean <- tm_map(corpus, removePunctuation)
  corpus_clean <- tm_map(corpus_clean, content_transformer(tolower))
  corpus_clean <- tm_map(corpus_clean, removeWords, stopwords("english"))
  corpus_clean <- tm_map(corpus_clean, removeNumbers)
  corpus_clean <- tm_map(corpus_clean, stripWhitespace)
  corpus_clean <- tm_map(corpus_clean, removeWords,
c(excl,"http","https","httpst"))

  return(corpus_clean)
}

## capture all the output to a file.

##### Collect from Twitter

# for creating a twitter app (apps.twitter.com) see youtube https://
youtu.be/1T4Kosc_ers
#consumer_key <- 'your key'
#consumer_scret <- 'your secret'
#access_token <- 'your access token'
#access_scret <- 'your access scret'

source("your_twitter.R") #this file will set my personal variables for my
twitter app, adjust the name of this file. use the provide template
your_twitter.R

setup_twitter_oauth(consumer_key,consumer_scret,
access_token,access_scret) #connect to twitter app

```

```

## Note that it will take the computer some to collect the tweets

tweets_J <- searchTwitter("#justinbieber", n=300, lang="en",
resultType="recent") #300 recent tweets about Justin Bieber, in English
tweets_T <- searchTwitter("#taylorswift", n=300, lang="en",
resultType="recent") #300 recent tweets about Taylor Swift, in English
tweets_B <- searchTwitter("#billieeilish", n=300, lang="en",
resultType="recent") #300 recent tweets about Billie Eilish, in English

##### Sentiment analysis

tweets_J.text <- laply(tweets_J, function(t)t$text()) #get text out of
tweets
tweets_T.text <- laply(tweets_T, function(t)t$text()) #get text out of
tweets
tweets_B.text <- laply(tweets_B, function(t)t$text()) #get text out of
tweets

#taken from https://github.com/mjhea0/twitter-sentiment-analysis
pos <- scan('positive-words.txt', what = 'character', comment.char=';')
#read the positive words
neg <- scan('negative-words.txt', what = 'character', comment.char=';')
#read the negative words

source("sentiment3.R") #load algorithm
# see sentiment3.R form more information about sentiment analysis. It
assigns a intereger score
# by substracitng the number of occurrence of negative words from that of
positive words

analysis_J <- score.sentiment(tweets_J.text, pos, neg)
analysis_T <- score.sentiment(tweets_T.text, pos, neg)
analysis_B <- score.sentiment(tweets_B.text, pos, neg)

sem<-data.frame(analysis_J$score, analysis_T$score, analysis_B$score)

semFrame <-melt(sem, measured=c(analysis_J.score,analysis_T.score,
analysis_B.score ))
names(semFrame) <- c("Candidate", "score")
semFrame$Candidate <-factor(semFrame$Candidate, labels=c("Justin Bieber",
"Taylor Swift", "Billie Eilish"))
``

### Homogeneity of variance analysis
<!-- Analyze the homogeneity of variance of sentiments of the tweets of
the different celebrities, and provide interpretation -->
From the boxplot containing the distribution of tweet sentiments for all
three celebrities we can conclude that there is a visible difference in
sentiment variance. Performing Levene's test verifies this. In the
results of that test we can see that the effect is significant
($p<0.05$). This indicates that there is variance inequality between all
three groups of tweets.

``{r, message=FALSE, warning=FALSE, echo = T, results = 'hide'}

library(car)

```

```

library(pander)
```
```{r, fig.cap="Boxplot of the sentiment values of the tweets for each
celebrity."}
boxplot(semFrame$score ~ semFrame$Candidate)
```
```{r, fig.cap="Results of performing Levene's test on the three groups
of tweets about the three celebrities."}
pander(leveneTest(semFrame$score, semFrame$Candidate, center = median))
```

### Visual inspection Mean and distribution sentiments
<!-- Graphically examine the mean and distribution sentiments of tweets
for each celebrity, and provide interpretation -->
Analysing the bar graph of sentiment counts for all tweets for each
celebrity, we can see that in general tweets about Justin Bieber are most
neutral, with most tweets having a sentiment of either 0 or 0.5, while
tweets about Taylor Swift vary greatly, having sentiment values between
-1 and 4. Tweets about Billie Eilish are the most negative and also have
the most tweets with value 0.

```{r, message=FALSE, warning=FALSE, echo = T, results = 'hide',
cache=TRUE}
library(ggplot2)

p <- semFrame %>% ggplot( aes(x=score)) + geom_histogram(
color="#e9ecef", alpha=0.6, position = 'identity') + facet_grid(. ~
Candidate)
```
```{r, fig.cap="Bar plot for the tweet sentiment values for each
celebrity."}
plot(p)
```

### Frequentist approach

#### Linear model
<!-- Use a linear model to analyze whether the knowledge to which
celebrity a tweet relates has a significant impact on explaining the
sentiments of the tweets. Provide interpretation of results -->
The results of performing an F-test on a null model and a model where the
celebrity is added as a independent variable show that there is a
significant improvement with the latter model with respect to the quality
of the fit. This means that knowing the celebrity gives information about
the distribution of tweet sentiments in the data set.

```{r, message=FALSE, warning=FALSE, echo = T, results = 'hide',
cache=TRUE}
#include your code and output in the document
library(pander)
library(multcomp)

semFrame$CandidateF <-factor(semFrame$Candidate, levels =c("Justin
Bieber", "Taylor Swift", "Billie Eilish"), labels =c("Justin Bieber",
"Taylor Swift", "Billie Eilish"))

res.aov <- aov(score ~ Candidate, data=semFrame, na.action=na.exclude)
```

```



```
``` {r fig.cap="Effect pre-post measuring condition on the tweet
sentiments of the three celebrities."}
pander(res.aov)
```
```

#### #### Post Hoc analysis

```
<!-- If a model that includes the celebrity is better in explaining the
sentiments of tweets than a model without such predictor, conduct a post-
hoc analysis with e.g. Bonferroni correction, to examine which of
celebrity tweets differ from the other celebrity tweets. Provide
interpretation of the results -->
```

Performing Post Hoc analysis allows us to determine whether there is a significant difference in tweet sentiments between all three celebrities, and how the distribution changes between them. The results, obtained by performing Tukey's Honest Significant Difference test, show that there is a significant difference in tweet sentiments for all three celebrities. Tweets about Taylor Swifts are the most positive, while those about Billie Eilish's are most negative, confirming our findings in the visual inspection.

```
```{r, message=FALSE, warning=FALSE, echo = T, results = 'hide',
cache=TRUE}
library(stats)
```
```

```
```{r, fig.cap="Results of performing Tukey's HSD."}
TukeyHSD(res.aov)
```
```

#### #### Report section for a scientific publication

```
<!-- Write a small section for a scientific publication, in which you
report the results of the analyses, and explain the conclusions that can
be drawn. -->
```

In order to determine whether sentiment of tweets depend on the celebrity they are about, a null model and a model with celebrity added as a independent variable predictor have been constructed. Results showed that the fit was significantly better ( $F(2, 897) = 44.39$ ,  $p. < 0.01$ ) with the added predictor. We can therefore conclude that the Twitter user base tweets differently depending on the three celebrities in the data set. Further analysis using Post Hoc analysis using Tukey's Honest Significant Difference test showed that there is a significant difference in tweet sentiment distribution in all three celebrity pairs, Taylor Swift – Justin Bieber ( $p. < 0.01$ ), and Billie Eilish – Justin Bieber ( $p. < 0.01$ ), and Taylor Swift – Billie Eilish ( $p. < 0.01$ ).

#### ### Bayesian Approach

#### #### Model description

```
<!-- Describe the mathematical model fitted on the most extensive model.
(hint, look at the mark down file of the lectures to see example on
formulate mathematical models in markdown). Justify the priors. -->
```

The most complex model tested uses the celebrity as an independent variable to predict the score. As we can expect the tweet sentiments to be normally distributed around 0 with almost all data between -5 and 5 due to the tweet word limit, a normally distributed prior with  $\mu=0$  and  $\sigma=2$  is used for  $\mu$ , and we allow  $\sigma$  to take a value between 0.0001 and 5. When adding celebrity as an independent variable predictor,  $a$  becomes a vector of parameters for each celebrity, giving us the following model where  $N$  and  $U$  represent a normal and uniform distribution respectively,

$$\mu_{\text{celebrity}} = N(a_{\text{celebrity}}, \sigma) \quad a \sim N(0, 2) \quad \sigma \sim U(0.0001, 5)$$

```
```{r, message=FALSE, warning=FALSE, echo = T, results = 'hide',
cache=TRUE}
semFrame <- subset(semFrame, select = c(score, CandidateF))
x <- seq(-5, 5, length=100)
y <- dnorm(x, mean=0, sd=2)
```
```

```
```{r, fig.cap="Sentiment values of all tweets."}
hist(semFrame$score)
```
```

```
```{r, fig.cap="Normal distribution with a mean of 0 and a standard
deviation of 2."}
plot(x, y, type="l", lwd=1)
```
```

#### #### Model comparison

```
<!-- Conduct model analysis and provide brief interpretation of the
results -->
Analysing the parameters  $a_1$ ,  $a_2$ , and  $a_3$ , we can conclude that the
means of the distributions of tweet sentiments vary significantly between
each of the three celebrities. Furthermore, comparing the null model with
the model using celebrity as a independent variable predictor, we can see
that the latter model has a lower WAIC and thus a better fit, indicating
that using celebrity as a predictor improves the fit of the model.
```

```
```{r, message=FALSE, warning=FALSE, echo = T, results = 'hide',
cache=TRUE}
# ```{r STAN CHUNK}
#include your code and output in the document
library("rstan")
library("rethinking")
```

```
m0 <-ulam(
  alist(
    score ~ dnorm(mu, sigma),
    mu <- a,
    a ~ dnorm(0, 2),
    sigma ~ dunif(0.0001, 5)),
  data = semFrame, iter = 10000, chains = 4, cores = 4,
  control=list(adapt_delta=.99), log_lik=TRUE
)
```

```
m1 <-ulam(
```

```

alist(
  score ~ dnorm(mu, sigma),
  mu <- a[CandidateF] ,
  sigma ~ dunif(0.0001, 5),
  a[CandidateF] ~ dnorm(0, 2)),
data = semFrame ,iter = 10000, chains = 4, cores = 4,
control=list(adapt_delta=.99), log_lik=TRUE
)

...

```{r fig.cap="Confidence intervals for the parameters of the model with
celebrity added as a predictor."}
plot(precis(m1, depth=2, prob = .95))
```

```{r fig.cap="Comparison of the WAIC score for each model."}
plot(compare(m0, m1, func=WAIC))
```

#### Comparison celebrity pair

<!-- Compare sentiments of celebrity pairs and provide a brief
interpretation (e.g. CIs) -->
Using `_bayes.t.test_` we can analyse the difference in tweet sentiment
distributions for each celebrity pair. From the plots we can deduce that
indeed there is a significant ( $p < 0.01$ ) probability that the true
distributions of the tweet sentiments for the three celebrities are
different from one another.

```{r, message=FALSE, warning=FALSE, echo = T, results = 'hide'}
# Might have to do 'brew install jags' on Mac to make this work.
library(rjags)
devtools::install_github("rasmusab/bayesian_first_aid")
library(BayesianFirstAid)
jbSub <- subset(semFrame, (CandidateF == "Justin Bieber"))
tsSub <- subset(semFrame, (CandidateF == "Taylor Swift"))
beSub <- subset(semFrame, (CandidateF == "Billie Eilish"))
```

``` {r, fig.cap="Results of performing the bayes t-test on the pair
Justin Bieber – Taylor Swift."}
plot(bayes.t.test(jbSub$score, tsSub$score))
```

``` {r, fig.cap="Results of performing the bayes t-test on the pair
Justin Bieber – Billie Eilish."}
plot(bayes.t.test(jbSub$score, beSub$score))
```

``` {r, fig.cap="Results of performing the bayes t-test on the pair
Taylor Swift – Billie Eilish."}
plot(bayes.t.test(tsSub$score, beSub$score))
```

## Question 2 – Website visits (between groups – Two factors)

### Conceptual model

```

![Conceptual Model underlying this research question]  
(ConceptualModel\_2.2.png)

### ### Visual inspection

<!--Graphically examine the variation in page visits for different factors levels (e.g. histogram, density plot etc.)-->  
The first step in our inspection is to observe the distribution of the dependent variable overall regardless of the factors.

```
```{r, message=FALSE, warning=FALSE}
library(ggplot2)
library(magrittr)
library(dplyr)
web_visit <- read.csv("webvisit0.csv")
web_visit$version <- factor(x=web_visit$version, labels=c("old", "new"))
web_visit$portal <- factor(x=web_visit$portal, labels=c("consumers",
"companies"))
```
```

```
```{r, warning=FALSE, message=FALSE, fig.cap="Histogram showing page
visits on the website over all factors."}
p <- web_visit %>% ggplot(aes(x=pages)) + geom_histogram(bins=10)
plot(p)
```
```

To determine whether either of the IV's have an effect on the distribution of the page visits we observe the distributions under different instantiations of the IV's in the following figure:

```
```{r, warning=FALSE, message=FALSE, fig.cap="Histograms of page visits
on the website per independent variable."}
p <- web_visit %>% ggplot( aes(x=pages)) + geom_histogram(bins=10) +
facet_grid(portal ~ version)
plot(p)
```
```

To further assist our understanding of the distribution we will also draw boxplots of the same data to observe differences in the distribution:

```
```{r, warning=FALSE, message=FALSE, fig.cap="Boxplots of page visits on
the website per independent variable"}
p <- web_visit %>% ggplot( aes(y=pages)) + geom_boxplot() +
facet_grid(portal ~ version, scales = "free_x", space = "free_x") +
theme( axis.title.x=element_blank(), axis.text.x=element_blank(),
axis.ticks.x=element_blank())
plot(p)
```
```

Based on the figures it seems that the IV's do affect the distribution of the DV. While the distributions between the old and new website for consumers do not show clearly different distributions, in the other cases there seems to be a significant difference. In particular, the distributions over the old and new website for companies shows a clearly different distribution in the histograms as well as in the boxplots.

### ### Normality check

<!--Visually inspect if variable page visits deviates from a Gaussian distribution, and discuss implication for general linear model analysis.-->

Again, we take the histogram of the distribution of all data and try to fit a normal distribution to the data. Here, we have taken a normal distribution with as  $\mu$  parameter the average of the data and as  $\sigma$  the variance of the data.

```
```{r, message=FALSE, warning=FALSE}
m<-mean(web_visit$pages)
std<-sqrt(var(web_visit$pages))
```

```{r, message=FALSE, warning=FALSE, fig.cap="Normal distribution fitted
to data of page visits"}
p <- web_visit %>% ggplot( aes(x=pages)) + geom_histogram(aes(y
=..density..), bins=15) + stat_function(fun = dnorm, args = list(mean =
mean(web_visit$pages), sd = sd(web_visit$pages)))

plot(p)
```
```

The figure shows that the distribution of the data does not fit to a normal distribution. This does not have to be an issue for the general linear model analysis as the assumption over the distribution of the data only applies to the errors. The assumption is that the errors are normally distributed.

### ### Frequentist Approach

#### #### Model analysis

<!--Conduct a model analysis, to examine the added values of adding 2 factors and interaction between the factors in the model to predict page visits, and include brief interpretation of the results.-->  
First we perform Levene's Test to determine the homogeneity of variance.

```
```{r}
library(car)
library(pander)

pander(leveneTest(web_visit$pages, interaction(web_visit$version ,
web_visit$portal)), caption = "Results of Levene's Test")
```
```

The p-value of Levene's Test in this case is larger than 0.05 and this thus indicates that the variance is homogeneous. Our assumptions on the data which allows for linear model fitting still hold.

The following tables show the results of fittings linear models and comparing them using ANOVA.

The first table shows whether adding the website version as a predictor for the page visits has a significant effect. The results show indeed that there is a significant effect. The second table indicates the same, but in this case for the portal of the website (i.e. market), this effect is also significant. The third table indicates whether there is an interaction effect between the two factors, which is also significantly demonstrated. The final table shows the combined effect of the two factors as well as the interaction.

```
```{r, warning=FALSE, message=FALSE}
```

```

model0 <- lm(pages ~ 1 , data = web_visit, na.action = na.exclude)
model1 <- lm(pages ~ version , data = web_visit, na.action = na.exclude)
model2 <- lm(pages ~ portal , data = web_visit, na.action = na.exclude)
model3 <- lm(pages ~ version + portal , data = web_visit, na.action =
na.exclude)
model4 <- lm(pages ~ version + portal + version:portal , data =
web_visit, na.action = na.exclude)
```


```

```{r}
pander(anova(model0,model1), caption = "Website version as main effect on
page visits")
pander(anova(model0,model2), caption = "Portal type as main effect on
page visits")
pander(anova(model3,model4),caption = "Interaction effect on top of two
main effects")
pander(anova(model4),caption = "Effect of Website version, Portal type
and interaction effect on page visits")
```

```


```

The results justify a further investigation into the effects of the factors. However, we will first also evaluate the goodness of fit of the final statistical model through a Akaike Information Criterion (AIC) comparison:

```

```{r, warning=FALSE, message=FALSE}
library(AICcmodavg)
models <-list(model0, model1, model2, model3, model4)
model.names <-c("model0","model1","model2","model3","model4")
```



```

```{r}
pander(aictab(cand.set = models, modnames=model.names), caption = "AIC
Comparison for the 5 models")
```

```


```

The analysis shows that the final model had the best goodness of fit and, in fact, captures all predictive power that could be found in the full set of models.

#### #### Simple effect analysis

<!--If the analysis shows a significant two-way interaction effect, conduct a Simple Effect analysis to explore this interaction effect in more detail.It helps first to look at the means of different conditions in a figure. Provide brief interpretation of the results.-->  
As we have previously found a two-way interaction effect between the version and portal factors of the experiment, we will conduct a Simple Effect analysis to explore this interaction effect.

The following bar plot shows how the page visits vary over the factors.

```

```{r, warning=FALSE, message=FALSE, fig.cap="Bar plot showing the mean
page visits per independent variable."}
bar <- ggplot(web_visit, aes(portal , pages, fill = version))
bar + stat_summary(fun.y = mean, geom = "bar", position="dodge")
```

```

The plot clearly shows that there is a greater difference between the old and new versions of the website for the company portal. The Simple effect analysis will be conducted through a linear model fitted to contrasts on the consumer and company variables.

```
```{r, warning=FALSE, message=FALSE}
web_visit$simple <- interaction(web_visit$version, web_visit$portal)
#merge two factors levels(Lec7g$simple) #to see the level in the new
factor
levels(web_visit$simple)

contrastConsumers <-c(1,-1,0,0) #Only the consumer portal data
contrastCompanies <-c(0,0,1,-1) #Only the company portal data

SimpleEff <- cbind(contrastConsumers,contrastCompanies)
contrasts(web_visit$simple) <- SimpleEff
```

```{r, warning=FALSE, message=FALSE}
simpleEffectModel <-lm(pages ~ simple , data = web_visit, na.action =
na.exclude)
pander(summary.lm(simpleEffectModel), caption="Summary of Simple Effect
linear model")
```
```

The Simple Effect analysis demonstrates no significant difference between page visits for users in the consumer portal using new and old versions of the website. However, it demonstrates that there is a significant difference for users in the company portal.

```
#### Report section for a scientific publication
<!-- Write a small section for a scientific publication, in which you
report the results of the analyses, and explain the conclusions that can
be drawn. -->
A linear model was fitted on the amount of page visits for users of a
website, taking as independent variables the website version (new or old)
and the user portal (consumer or company). The analysis found a
significant main effect ( $F(1,995)=36.2, \ p.<0.01$ ) for the website
version as well as for the portal ( $F(995,1)=178.8, \ p.<0.01$ ). Moreover,
the analysis found a significant two-way interaction effect
( $F(1,995)=46.2, \ p.<0.01$ ). A Simple Effect analysis was conducted to
further examine the two-way interaction between the independent
variables. It revealed a significant ( $t=9.22, \ p. 0.01$ ) difference for
website version in the company portal of the website but no significant
effect ( $t=13.4, \ p=0.67$ ) for the website version in the consumer
portal.
```

### Bayesian Approach

```
#### Model description
<!-- Describe the mathematical model fitted on the most extensive model.
(hint, look at the mark down file of the lectures to see example on
formulate mathematical models in markdown). Justify the priors. -->
```

Our most extensive Bayesian model captures each individual factor as well as the interaction of the factors and is defined as follows:

```


$$\text{PageVisits} \sim N(\mu, \sigma)$$


$$\mu = a + b \cdot \text{Version} + c \cdot \text{Portal} + d \cdot \text{Version} \cdot \text{Portal}$$


$$a \sim \text{Gamma}(1, 2)$$


$$b \sim N(0, 1)$$


$$c \sim N(0, 1)$$


$$d \sim N(0, 1)$$


$$\sigma \sim U(0.1, 2)$$


```

Here the functions  $\text{Gamma}$ (\$),  $N$ (\$\$) and  $U$ (\$\$) denote the probability functions of the Gamma, Normal and Uniform distributions respectively.

The Gamma distribution was chosen as a prior as previous plots of the page visits shows the distribution to follow a Gamma distribution. By means of visual inspection the parameters shape  $k=1$  and scale  $\theta=2$  were chosen. See the figure below for a fit of the distribution to the data with these parameters.

```

```{r, warning=FALSE, message=FALSE, fig.cap="Gamma distribution (1, 2)
fitted to data of page visits"}
p <- web_visit %>% ggplot( aes(x=pages)) + geom_histogram(aes(y
=..density..), bins=15) + stat_function(fun = dgamma, args = list(shape =
1, scale=2))
plot(p)
```

```

#### #### Model comparison

<!-- Conduct model analysis and provide brief interpretation of the results -->

Besides the previously defined model, we defined simpler models which capture the effects of the factors on their own and their joint as well as interaction effect.

```

```{r, warning=FALSE, message=FALSE, results = 'hide', echo=T,
cache=TRUE}
web_visit <- subset(web_visit, select = c(pages,version, portal ))
web_visit$versionN <- as.numeric(web_visit$version)
web_visit$portalN <- as.numeric(web_visit$portal)

m0 <-map2stan( alist(
  pages ~ dnorm(mu, sigma), mu <- a ,
  a ~ dgamma(shape=1, scale=2),
  sigma ~ dunif(0.1, 2)
), data = web_visit, iter = 10000, chains = 4, cores = 4
)

m1 <-map2stan( alist(
  pages ~ dnorm(mu, sigma), mu <- a + b*versionN,
  a ~ dgamma(shape=1, scale=2),
  b ~ dnorm(0, 1),
  sigma ~ dunif(0.1, 2) ), data = web_visit, iter = 10000, chains = 4,
cores = 4
)

m2 <-map2stan( alist(

```



```

    pages ~ dnorm(mu, sigma), mu <- a + c*portalN ,
    a ~ dgamma(shape=1, scale=2),
    c ~ dnorm(0, 1),
    sigma ~ dunif(0.1, 2)),
    data = web_visit, iter = 10000, chains = 4, cores = 4
)

m3 <-map2stan( alist(
  pages ~ dnorm(mu, sigma), mu <- a + b*versionN + c*portalN ,
  a ~ dgamma(shape=1, scale=2),
  b ~ dnorm(0, 1),
  c ~ dnorm(0, 1), sigma ~ dunif(0.1, 2)),
  data = web_visit, iter = 10000, chains = 4, cores = 4
)

m4 <-map2stan( alist(
  pages ~ dnorm(mu, sigma),
  mu <- a + b*versionN + c*portalN + d*versionN*portalN,
  a ~ dgamma(shape=1, scale=2),
  c(b,c,d) ~ dnorm(0, 1),
  sigma ~ dunif(0.1, 2)),
  data = web_visit, iter = 10000, chains = 4, cores = 4
)
...

```

The figure below demonstrates the differences in WAIC between varying models. Here, `m0` is a model with just the prior, `m1` is a model capturing the effect of the website version, `m2` captures the effect of the portal, `m3` captures the effects of both the version and the portal and `m4` represents the model described earlier.

```

```{r, warning=FALSE, message=FALSE, fig.cap="WAIC comparison between
Bayesian models capturing effects of different variables."}
plot(compare(m0,m1,m2,m3,m4, func=WAIC))
```

```

```

```{r, warning=FALSE, message=FALSE}
pander(compare(m0,m1,m2,m3,m4), caption = "Details of the fitted Bayesian
models")
pander(precis(m4, prob= .95), caption = "Details of the fitted extensive
model described at the beginning of the section.")
```

```

The WAIC analysis above indicates that the extensive model described at the beginning of the section has the best goodness of fit according to the WAIC score. The table below the WAIC comparisons shows the 95% credible intervals for all the model parameters. As the null is not present in any of the variables' 95% credible intervals, we deem the results significant.

### # Part 3 – Multilevel model

#### ## Visual inspection

```

<!-- Use graphics to inspect the distribution of the score, and
relationship between session and score -->
For this exercise, we use `set1.csv`. Below we inspect the distribution
of the score through a histogram of the score.

```

```

```{r, fig.cap="Histogram of score"}
exp_data = read.csv("set1.csv")
p <- exp_data %>% ggplot( aes(x=score)) + geom_histogram(bins=15) +
theme()
plot(p)
```

```

We also inspect the relationship between session and score through a scatterplot.

```

```{r, fig.cap="Scatterplot relation between score and session"}
p <- exp_data %>% ggplot( aes(x=session, y=score)) + geom_point(alpha=0.
25) + theme()
plot(p)
```

```

Lastly we inspect the same relationship through a plot showing the mean score and standard error over sessions.

```

```{r, fig.cap="Mean score and standard error over sessions."}
library(Rmisc)

tgc = summarySE(exp_data, measurevar="score", groupvars=c("session"))
p <- tgc %>% ggplot(aes(x=session, y=score, alpha=N*2)) +
  geom_errorbar(aes(ymin=score-se, ymax=score+se), width=.1) +
  geom_line() +
  geom_point()

plot(p)
```

```

## Frequentist approach

### Multilevel analysis

<!--

Conduct multilevel analysis and calculate 95% confidence intervals, determine:

\* If session has an impact on people score

\* If there is significant variance between the participants in their score

-->

Below we conduct a multilevel analysis to determine if session has an impact on people's score and to determine if there is a significant variance between the participants in their score.

The first model, \$m\_0\$, includes a fixed intercept ( $\sim 1$ ) and a random intercept, indicated by `random = 1|Subject`. This model will have a general intercept (the fixed effect intercept) and an intercept for each of the subjects. In the output we see the standard deviation between the random-effect terms, which is 46.53 for the intercept per clinic, and 35.26 for the residuals. The fixed intercept value is 116.81.

```

```{r}
library(nlme)

```

```
model0 <- lme(score ~ 1 , random = ~1|Subject, data = exp_data,
method="ML")
summary(model0)
```

```

We also inspect the 95% confidence interval. We see that 0 is not included in the intervals, therefore the found effects seem to be significant. We conclude that there is significant variance between participants in their score.

```
```{r}
intervals(model0, 0.95)
```

```

The second model \$m\_1\$ includes the variable `session`. The summary function shows an estimated fixed effect for session on the score of 0.37. With a p-value of 0.0 this fixed effect is significant.

```
```{r}
model1 <- lme(score ~ session , random = ~1|Subject, data = exp_data,
method="ML")
summary(model1)
```

```

Inspecting the 95% confidence intervals, we can again confirm that the effects are significant, since the intervals do not include 0. We conclude that session has an impact on people's score.

```
```{r}
intervals(model1, 0.95)
```

```

```
```{r}
pander(anova(model0, model1), caption = "Model comparison.")
```

```

### Report section for a scientific publication

<!--

Write a small section for a scientific publication, in which you report the results of the analyses, and explain the conclusions that can be drawn.

-->

To investigate whether session has an impact on people's scores and whether there is a significant variance between the participants in their score, multilevel analysis was performed. A model with a random intercept only and a model with a random intercept and fixed effects were created. There was a significant relationship between session and score,  $M = 116.8139$  (95% CI 112.70, 120.93),  $p = 0$ . The relationship between score and session showed significant variance in intercepts across subjects,  $SD = 46.51$  (95% CI 43.67, 49.54),  $p = 0$ .

## Bayesian approach

### Model description

<!--

Describe the mathematical model fitted on the most extensive model. (hint, look at the mark down file of the lectures to see example on formulate mathematical models in markdown). Justify the priors.

-->

In most extensive mathematical model, the score is drawn from a normal distribution with mean  $\mu$  and standard deviation  $\sigma$ . Parameter  $\mu$  is created through a linear equation involving parameters  $a$  (an intercept),  $a_{\text{Subject}}$  (a subject-specific intercept) and  $b$  multiplied by  $\text{session}$ .

```
 $\text{Score} \sim N(\mu, \sigma)$   
 $\mu = a + a_{\text{Subject}}[\text{Subject}] + b * \text{session}$   
 $a_{\text{Subject}}[\text{Subject}] \sim N(0, \sigma_{\text{Subject}})$   
 $\sigma_{\text{Subject}} \sim \text{Cauchy}(0, 100)T(0, \infty)$   
 $a \sim N(0, 100)$   
 $b \sim N(0, 5)$   
 $\sigma \sim \text{Cauchy}(0, 50)T(0, \infty)$ 
```

For the prior of the standard deviations  $\sigma_{\text{Subject}}$  and  $\sigma$  we use a half Cauchy distribution. We use a Cauchy distribution with a large scale value, since we do not have much prior knowledge or belief about the true parameters. With this large scale, the half Cauchy distribution has, unlike the normal distribution, fat tails. This makes the distribution over values away from the mean more uniform than in a normal distribution. For parameters  $a$ ,  $a_{\text{Subject}}$  and  $b$  we chose to use a normal distribution as prior.

### ### Model comparison

We select the first 100 participants from the data set. Then, we create three models with increasing complexity.

```
``{r, message=FALSE, warning=FALSE, cache=TRUE}  
exp_data_1 = data.frame(exp_data)  
exp_data_1 = exp_data_1[exp_data_1$Subject < 100,]  
exp_data_1$Subject <- exp_data_1$Subject + 1  
  
m0 <-ulam( alist(  
  score ~ dnorm(mu, sigma),  
  mu <- a ,  
  a ~ dnorm(0, 100),  
  sigma ~ dcauchy(0, 50)  
) , data = exp_data_1, iter = 10000, chains = 4, cores = 4,  
control=list(adapt_delta=.99), log_lik=TRUE  
)  
  
m1 <-ulam( alist(  
  score ~ dnorm(mu, sigma),  
  mu <- a + a_Subject[Subject] ,  
  
  a_Subject[Subject] ~ dnorm(0, sigma_Subject),  
  sigma_Subject ~ dcauchy(0, 100),  
  
  a ~ dnorm(0, 100),  
  sigma ~ dcauchy(0, 50)  
) , data = exp_data_1, iter = 10000, chains = 4, cores = 4,  
control=list(adapt_delta=.99), log_lik=TRUE  
)  
  
m2 <-ulam( alist(  
  score ~ dnorm(mu, sigma),
```

```

mu <- a + a_Subject[Subject] + b * session,

a_Subject[Subject] ~ dnorm(0, sigma_Subject),
sigma_Subject ~ dcauchy(0, 100),

a ~ dnorm(0, 100),
b ~ dnorm(0, 5),
sigma ~ dcauchy(0, 50)
), data = exp_data_1, iter = 10000, chains = 4, cores = 4,
control=list(adapt_delta=.99), log_lik=TRUE
)

compare(m0,m1,m2)
```


```

```{r fig.cap="Comparing three Bayesian models of increasing
complexity."}
plot(compare(m0,m1,m2))
```

```


```

From the plot we can see that models  $m_1$  and  $m_2$  outperform model  $m_0$ . However, there is not much of a difference between  $m_1$  and  $m_2$ . The effective number of parameters ( $spWAIC$ ) increases for both these models.

### Estimates examination

Below we examine the estimate of parameters of the model with best fit, which is  $m_2$ . We see that the uncertainty for parameters  $\sigma_{\text{Subject}}$  and  $a$  is relatively high compared to that for parameters  $b$  and  $\sigma$ . Since the value for  $b$  is close to zero, it seems that session does not impact score.

```

```{r, fig.cap="Parameters of m2."}
plot(m2)
```

```