~~~
---
title: "Report Template coursework assignment A - 2019"
subtitle: "CS4125 Seminar Research Methodology for Data Science"
author: "Alexander Bieniek, Wesley Quispel, David Nyrnberg"
date: "04/02/2019"
output:
   pdf_document:
      fig_caption: true
      number_sections: true
---

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


\tableofcontents



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

## Motivation
  Students, researchers, and most people beyond these buckets believe that caffeine improves their productivity. Many
even insist on having their morning coffee to do their work each day. We design an experiment to empirically evaluate
the effects of caffeine as well as other intake trends on one's mental acuity.
  Perhaps caffeine truly improves one's performance in their studies or work. However, we may also investigate the
existence of a placebo effect on one's work ethic. We also investigate more complex effects of caffeine with its
relationship on mental acuity, such as the influence of the caffeine-induced "crash", and caffeine tolerance, and
possible relationships between caffeine effectiveness and amount of sleep.

## Theory



## Research Questions
Put succinctly, here are some of the questions that we investigate in this experiment:
- Does caffeine intake change one's mental acuity?
- Does the "caffeine crash", or the perceived exhaustion after wearing off of caffeine, change one's mental acuity?
- Does there exist a placebo affect in caffeine intake and one's mental acuity?
- Does caffeine tolerance change the effectiveness of a single dose of caffeine?
- Does one's recent amount of sleep change the effectiveness of caffeine at changing one's mental acuity?

## Participants
  For convenience and consistency, the experiment will use students from TUDelft as participants. A sample from this
population would likely generalize to broader student populations for the effects of caffeine intake. We can also find
various levels of caffeine intake habits and regular sleep amounts. Lastly, because the students belong to the same
university, we can expect that, with a lower variance in mental acuity, a smaller sample size could suffice for testing
of statistical significance.
  One thing to consider is that, because TUDelft is a linguistically diverse university, we can make no assumptions
about the language backgrounds of the students. As such, it is important to test subjects with means independent of
reading ability, domain experience, etc.
  Administering of caffeine shall be transparent and consentual. Caffeine will be administered in commercially available
forms and otherwise ordinary forms, with the possibility of caffeine-free doses.

## Conceptual Model
**Dependent Variable: reading test score**
Mental acuity is difficult to measure, so we will approximate it by administering a brief intelligence examination. In
particular, the subject will be given an IQ test, which can be used as an unbiased approximation of the mentual acuity
of a subject.

**Independent Variables:**
* Amount of Administered Caffeine (in mg)
* Duration Since Caffeine Dose

**Mediating Variables:**
* Caffeine Tolerance (as approximated by amount of caffeine per day)
* Average Sleep Amount (in Hours)

**Moderating Variables:**
* Percieved Inake of Caffeine  (Placebo or No Placebo)

## Experimental Design
The experiment explores the change in mental accuity as described by various factors around caffeine intake. There are
many variables which many complicate and influence the change in mental accuity, such as caffeine tolerance and duration
since dose, and these variables will be tested in experiments where other variables are controlled.

## Experimental Procedure
For the colection of data from test subjects:
1. Randomly sample a test subject from a place of study, and preemptively, randomly assign treatment to the subject
 * treatment would be in the form of a dose of caffeine, a placebo dose of caffeine, or no dose of caffeine
 * We would need to find people who have not had any caffeine yet on the given day
~~~

~~~
2. Collect some preliminary information about the subject, including estimated sleep amount and estimated caffeine
intake per day
3. administer a waiting period for the subject to go about other activities and allow the caffeine to kick in and/or
wear off
 * Duration of waiting period will be influenced by academic literature about caffeine
4. After the waiting period, administer the IQ test
 * The test should be realistically impossible to complete in the given amount of time to avoid statistically
meaningless score distributions


## Experiments and Suggested Statistical Analyses
In this section, we describe the types of independent variables that we will observe and what statistical tests would
allow for hypothesis testing:
1. Effect of Caffeine Intake on Mental Accuity -> two-sample t-statistic
2. Placebo Effect of Caffeine Intake -> two-sample t-statistic
2. Effectiveness of Caffeine Intake Over Time -> Regression Analysis
3. Influence of Caffeine Tolerance on Caffeine Effectiveness -> Regression Analysis
3. Influence of Sleep on Caffeine Effectiveness -> Multiple Regression Analysis

#Part 2 - Generalized linear models

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

## Collecting Preliminary Data

Set up libraries...
```{r, echo=FALSE, message=FALSE, warning=FALSE, include = FALSE}
if (FALSE) {
  install.packages("base64enc")
  install.packages("twitteR", dependencies = TRUE)
  install.packages("RCurl", dependencies = T)
  install.packages("bitops", dependencies = T)
  install.packages("plyr", dependencies = T)
  install.packages('stringr', dependencies = T)
  install.packages("NLP", dependencies = T)
  install.packages("tm", dependencies = T)
  install.packages("wordcloud", dependencies=T)
  install.packages("RColorBrewer", dependencies=TRUE)
  install.packages("reshape", dependencies=T)
}

library(base64enc)
library(twitteR)
library(bitops)
library(RCurl)
library(plyr)
library(stringr)
library(NLP)
library(tm)
library(RColorBrewer)
library(wordcloud)
library(reshape)
library(car)
```

Collect positive and negative words and related functions, set up twitter oauth...
```{r, echo=FALSE, message=FALSE, warning=FALSE, include = FALSE}
# load local textfiles listing key positive and negative words
#taken from https://github.com/mjhea0/twitter-sentiment-analysis
positive_words = scan('./data/positive-words.txt', what = 'character', comment.char=';') #read the positive words
negative_words = scan('./data/negative-words.txt', what = 'character', comment.char=';') #read the negative words

# set up twitter sesssion
source("twitter_keys.R") # imports consumer_key, consumer_secret, access_key, and access_secret
source("sentiment3.R")
setup_twitter_oauth(consumer_key, consumer_secret,access_token, access_secret)
```

Collect tweets about celebrities...
```{r, echo=FALSE, message=FALSE, warning=FALSE, include = FALSE}
if (TRUE) {
  tweets_KW = searchTwitter("@KanyeWest", n=1e3, lang="en", resultType="recent")
  tweets_D  = searchTwitter("@Drake", n=1e3, lang="en", resultType="recent")
  tweets_AG = searchTwitter("@ArianaGrande", n=1e3, lang="en", resultType="recent")
  tweets_CB = searchTwitter("@IAmCardiB", n=1e3, lang="en", resultType="recent")
  tweets_LP = searchTwitter("@LilPump", n=1e3, lang="en", resultType="recent")
  tweets_NM = searchTwitter("@NickiMinaj", n=1e3, lang="en", resultType="recent")
}
```
~~~

~~~
Parse tweets, collect sentiment scores from them...
```{r, echo=FALSE, message=FALSE, warning=FALSE, include = FALSE}
tweets_KW.parsed= laply(tweets_KW, function(t)t$getText())
tweets_D.parsed  = laply(tweets_D,  function(t)t$getText())
tweets_AG.parsed = laply(tweets_AG, function(t)t$getText())
tweets_CB.parsed = laply(tweets_CB, function(t)t$getText())
tweets_LP.parsed = laply(tweets_LP, function(t)t$getText())
tweets_NM.parsed = laply(tweets_NM, function(t)t$getText())

tweets_KW.analysis = score.sentiment(tweets_KW.parsed, positive_words, negative_words)
tweets_D.analysis  = score.sentiment(tweets_D.parsed,  positive_words, negative_words)
tweets_AG.analysis = score.sentiment(tweets_AG.parsed, positive_words, negative_words)
tweets_CB.analysis = score.sentiment(tweets_CB.parsed, positive_words, negative_words)
tweets_LP.analysis = score.sentiment(tweets_LP.parsed, positive_words, negative_words)
tweets_NM.analysis = score.sentiment(tweets_NM.parsed, positive_words, negative_words)
```

Aggregate tweet scores, put into data frames...
```{r, echo=FALSE, message=FALSE, warning=FALSE, include = FALSE}
sem = data.frame(KW = tweets_KW.analysis$score,
                 D  =  tweets_D.analysis$score,
                 AG = tweets_AG.analysis$score,
                 CB = tweets_CB.analysis$score,
                 LP = tweets_LP.analysis$score,
                 NM = tweets_NM.analysis$score)
sem_frame = melt(sem, measured=c(tweets_KW.analysis$score,
                                 tweets_D.analysis$score,
                                 tweets_AG.analysis$score,
                                 tweets_CB.analysis$score,
                                 tweets_LP.analysis$score,
                                 tweets_NM.analysis$score))
names(sem_frame) <- c("candidate", "score")
sem_frame$candidate <-factor(sem_frame$candidate, labels=c("KanyeWest",
         "Drake",
         "ArianaGrande",
         "IAmCardiB",
         "LilPump",
         "NickiMinaj"))
```

## Tweet Sentiment Analysis
Is there a difference in the sentiment of the tweets related to the different celebrities? We inspect this question
using context independent sentiment analysis of tweets about them. The data was collected as follows:
 1. We used the twitter api to collect the 1000 most recent tweets which contain the twitter tag of the celebrity of
interest
 2. Punctuation and links were removed from the tweets
 3. Upon these "parsed" tweets, we counted the number of "positive" and "negative" words present in the tweet, as
provided in lists for the assignment, and the tweets were given a sentiment score as a difference of the positive and
negative word counts

 Tweets about the following American music artists were collected:
 - Kanye West
 - Drake
 - Ariana Grande
 - Cardi B
 - Lil Pump
 - Nicki Minaj

### Statistical Considerations:
Making no assumptions, we would expect that all tweets about our celebrities have similar sentiment scores. That is, the
sentiment scores found in tweets about celebrities have come from the same, broad distribution of sentiment scores in
tweets about all celebrities. The alternative hypothesis would be that tweet sentiment scores come from different
distributions when we sample tweets about the various celebrities.

## Assessing Homogeneity of Variance
```{r}
leveneTest(score ~ candidate, sem_frame)
```

We use the Levene Test to assess homogeneity of variance. At an alpha value of .05, our p-value is less than our alpha
value, so we do not believe that the tweet sentiment distributions from the various celebrities have homogeneous
variances.

## Graphical Examination of Means and Variations of Tweet Sentiment by Celebrity
```{r}
boxplot(score ~ candidate, sem_frame)
par(mfrow=c(2,3))
for (name in unique(sem_frame$candidate)) {
  hist(sem_frame[sem_frame$candidate == name,]$score, breaks = seq(-8, 8, 1),
       main=paste("Tweet Sentiments for", name), xlab="sentiment score", ylim=c(0, 750))
~~~

~~~
}
```
Above, we have produced visualizations of the sampled distributions of tweet sentiments about various American celebrity
artists. They have different medians, levels of spread, and numbers of outliers. In particular:
 - All celebrities were found to have a median neutral sentiment score for their sampled tweets except for Ariana
Grande, who has a median +1 tweet sentiment
 - Most celebrities have an interquartile range of 1, while Drake and Cardi B have an interquartile range of 2
 - Most celebrities have outliers of similar scores, but:
   - there is at least one tweet about Nicki Minaj with a sentiment score of +6, which isat least 2 sentiment points
higher than all other tweets
   - there is one or more tweets about Drake which have a sentiment score of -5, which is at least 1 point lower than
all other tweets
 - Distributions tend to center around neutral tweets, but some celebrities recieve far less neutral tweets than others:
   - Lil Pump appears to reach ~550 neutral tweets
   - Offset only reaches ~350 neutral tweets

## Tweet Knowledge as Used to Describe Tweet Sentiment
```{r}

```

This part of the assignment prompt feels unclear and will be saved for a later time.

## Assessing Model Quality Using Tweet Knowledge as a Predictor

This part of the assignment prompt feels unclear and will be saved for a later time.

## Findings

Through statistical significance tests and inspection of visualizations, we believe that the sentiment score
distribution from tweets pertaining to various American celebrity artists do not come from the same distribution. The
sampled distributions of sentiment scores about artists have different variances, median sentiment scores, and numbers
of outliers.

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

```{r}
#setwd("/Users/wesleyquispel/Documents/Universiteit/2018-2019/p3-seminar")
# apple , note use / instead of \, which used by windows


library(twitteR)
library(RCurl)
library(bitops)
library(plyr)
library(stringr)
library(NLP)
library(tm)
library(RColorBrewer)
library(wordcloud)
library(reshape)

library(Rcmdr)
library(foreign)
library(sm)
library(car)

library(tidyr)  # for wide to long format transformation of the data
library(ggplot2)
library(QuantPsyc) #include lm.beta()
library(gmodels)
library(pander) #for rendering output
library(ez) #for ezANOVA
library(nlme)
```

```{r}
# We use version 1
data<-read.csv("./data/webvisit1.csv", header = TRUE)
cat("Mean=", mean(data$pages), "\n", "SD=", sd(data$pages), "\n", "median=", median(data$pages), "\n")
cat("Max=", max(data$pages), "\n", "Min=", min(data$pages))
```

### Conceptual model
Make a conceptual model underlying this research question

### Visual inspection
Graphically examine the variation in page visits for different factors levels (e.g. histogram, density plot etc.)
~~~

```{r}
hist(data$pages, xlab="Number of Pages", main="Histogram of distribution page visits")

#d <-density(data$pages) #density plot
#plot(d)
```
```{r}
data$versionCat<-factor(data$version, levels = c(0:1), labels = c("Old","New"))
data$portalCat<-factor(data$portal, levels = c(0:1), labels = c("Consumers","Companies"))

sm.density.compare(data$pages, data$portal, xlab = "Number of Pages")
title(main="Number of pages per portal")
legend('topright', legend=levels(data$portalCat), col=c('red', 'green'), lty=1:2, cex=0.8,             title="Portal",
text.font=4, bg='lightblue')

sm.density.compare(data$pages, data$version, xlab = "Number of Pages")
title(main="Number of pages per version")
legend('topright', legend=levels(data$versionCat), col=c('red', 'green'), lty=1:2, cex=0.8,
       title="Version", text.font=4, bg='lightblue')
```

```{r}
scatterplot(pages ~ portalCat, data = data,
            main="Scatterplot with extra features")

scatterplot(pages ~ user, data = data,
            main="Scatterplot with extra features")

scatterplot(pages ~ versionCat, data = data,
            main="Scatterplot with extra features")
```

### Normality check
Statistically test if variable page visits deviates from normal distribution

```{r}
shapiro.test(data$pages)
```

### 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.

```{r}
#include your code and output in the document
CrossTable(data$portal, data$pages, prop.r=FALSE, prop.c = FALSE,
           prop.t = FALSE,
           prop.chisq=FALSE, format = "SPSS",
           fisher = FALSE, chisq = TRUE,
           expected = FALSE, sresid = FALSE)
```
```{r}
#include your code and output in the document
CrossTable(data$version, data$pages, prop.r=FALSE, prop.c = FALSE,
           prop.t = FALSE,
           prop.chisq=FALSE, format = "SPSS",
           fisher = FALSE, chisq = TRUE,
           expected = FALSE, sresid = FALSE)
```

Linear model with
```{r}
model0 <- lm(data$pages ~ 1, data=data)
model1 <- lm(data$pages ~ data$portal, data=data)
model2 <- lm(data$pages ~ data$version, data=data)
anova(model0, model1, test = "F")
anova(model0, model2, test = "F")
cat("Summary: \n ----------------------------------------------------- \n")
summary(model1)
summary(model2)
cat("Anova: \n ----------------------------------------------------- \n")
Anova(model1)
Anova(model2)
```
```{r}
model0 <- lm(data$pages ~ 1 , data = data, na.action = na.exclude)
model1 <- lm(data$pages ~ portal , data = data, na.action = na.exclude)
model2 <- lm(data$pages ~ version , data = data, na.action = na.exclude)
```

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

```{r}
data$unstandardizedResiduals1 <- resid(model1)
hp <- ggplot(data, aes(x= unstandardizedResiduals1)) + geom_histogram() + labs(title="distribution residuals")
hp + facet_grid(.~portalCat)
```
```{r}
data$unstandardizedResiduals2 <- resid(model2)
hp <- ggplot(data, aes(x= unstandardizedResiduals2)) + geom_histogram() + labs(title="distribution residuals")
hp + facet_grid(.~versionCat)
```

###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 in a figure


```{r}
data$simple <- interaction(data$portal, data$version)
contractSimple <- c(1,-1,0,0)
contractComplex <- c(0,0,1,-1)
SimpleEff <- cbind(contractSimple,contractComplex)
contrasts(data$simple) <- SimpleEff
simpleEffectModel <- aov(pages ~ simple , data = data, na.action = na.exclude)
summary.lm(simpleEffectModel)
```



### 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.

**Independent variables.** Version used (old, new), Portal used (consumers, companies). Interaction by any combination
thereof ((old:consumer), (old:companies), (new:consumers), (new:companies))

**Dependent variables.** Number of pages visited by user.

**Null hypothesis.** There is no observed difference in the number of pages visited based on either the versions, the
portals, or a combination thereof used. The observed difference in the sample is based on a sampling error and there is
no observed difference in the entire population.

**Alternative hypothesis.** The observed difference in the sample is a real effect plus some change variation.


# Part 3 - Multilevel model

```{r}
ml_data<-read.csv("./data/set1.csv", header = TRUE)
```

## Visual inspection
Use graphics to inspect the distribution of the score, and relationship between session and score


```{r}
#include your code and output in the document
hist(ml_data$score, xlab="Score", main="Distribution of score")
hist(ml_data$session, xlab="Session", main="Distribution of sessions")
# scatterplot(ml_data$session, ml_data$score, xlab="Sessions", ylab="Score", main="Relation between score and sessions")
# pairs(~Subject+session+score,data=ml_data,main="Scatterplot Matrix")
#
#
# model0 <- lm(ml_data$score ~ ml_data$Subject , data = ml_data, na.action = na.exclude)
# model1 <- lm(ml_data$score ~ ml_data$session , data = ml_data, na.action = na.exclude)
# summary(model0)
# summary(model1)

range_vec <- c(0:9)
new_data <- ml_data[ml_data$Subject %in% range_vec, ]

scatter <-ggplot(new_data, aes(x=session, y=score, color=Subject))
scatter +geom_point() + geom_smooth(method="lm", se= F) + scale_color_gradientn(colours = rainbow(100))
~~~

~~~
scatter <-ggplot(ml_data, aes(x=session, y=score, color=Subject))
scatter +geom_point() + geom_smooth(method="lm", se= F) + scale_color_gradientn(colours = rainbow(100))

# scatter <- qplot(x = session, y = score, color = Subject, data =  ml_data, geom = "point")
# scatter + scale_fill_gradient(heat.colors(unique(ml_data$Subject)))
```

## 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


```{r}
#include your code and output in the document

randomInterceptOnly <- lme(score ~ 1, data = ml_data,
                           random = ~1|Subject, method = "ML")
summary(randomInterceptOnly)


intervals(randomInterceptOnly, 0.95)
```
```{r}
randomInterceptSession <- lme(score ~ session,
                              data = ml_data, random = ~1|Subject, method = "ML")
summary(randomInterceptSession)

anova(randomInterceptOnly,randomInterceptSession)
```
```{r}
```

## 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.
~~~