```
---
title: "Report Template coursework assignment A - 2019"
author: "Alexander Bieniek, Wesley Quispel, David Nyrnberg"
date: "04/02/2019"
output:
  pdf_document:
    fig_caption: yes
    number_sections: yes
  word_document: default
subtitle: CS4125 Seminar Research Methodology for Data Science
bibliography: references.bib
---

```{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
  Academics have performed a host of analyses on the effects of caffeine on cognitive performance. For example, a study by Nehlig at UDS found that caffeine changes memory performance in nuanced ways, and it likely does not change the aggregation of long-term memory [@Neglig2010]. The study concluded that "caffeine cannot be considered a 'pure' cognitive enhancer", although it may indirectly influence, and possibly enhance, one's cognitive performance. [@Neglig2010]. Another study by Pasman et al. found that, when taking cognition tests, scores of subjects did not improve, but the tests were completed "approximately 10% faster" [@Pasman2017]. The findings of these studies suggest that, while caffeine may not directly improve cognitive performance, indirect factors may still lead users to enjoy increased efficiency when doing their work.

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)
Set up libraries...
```{r, echo=FALSE}
if (F) {
  install.packages("base64enc", dependencies=T)
  install.packages("twitteR", dependencies=T)
  install.packages("plyr", dependencies=T)
  install.packages("stringr", dependencies=T)
  install.packages("container", dependencies=T)
}

library(container)
library(base64enc)
library(twitteR)
library(plyr)
library(stringr)
library(car)
```

Collecting Preliminary Data
Collect positive and negative words and related functions, set up twitter oauth...
```{r echo=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
source("sentiment3.R")

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

Collect tweets about celebrities, calculate sentiment scores...
```{r, echo=FALSE}
tweetsFilename = 'data/tweets.csv'
n_tweets = 1e3
celebHandles = Dict$new(c("Kanye West" = "@KanyeWest",
                          "Drake" = "@Drake",
                          "Ariana Grande" = "@ArianaGrande",
                          "Cardi B" = "@IAmCardiB",
                          "Lil Pump" = "@LilPump",
                          "Nicki Minaj" = "@NickiMinaj")
                        )
if (file.exists(tweetsFilename)) {
  tweets = read.csv(tweetsFilename, header=T)
```

```
} else {
  tweets = NULL
    for (name in celebHandles$keys()) {
    handle = celebHandles[name]
    retTweets = searchTwitter(handle, n=n_tweets, lang="en", resultType="recent")
    out = data.frame("text"=laply(retTweets, function(t)t$getText()))
    out$name = name
    out$handle = handle
    tweets = rbind(tweets, out)
  }
  write.csv(tweets, file='data/tweets.csv', row.names=T)
}

tweets$score = score.sentiment(tweets$text, positive_words, negative_words)[[1]]
```

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(tweets$score, group=tweets$name)
```

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 reject the assumption 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 ~ name, data=tweets)
par(mfrow=c(2,3))
for (name in unique(tweets$name)) {
  hist(tweets[tweets$name == name,]$score, breaks = -8:8 + .5,
       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 neutral sentiment score for the mode for their sampled tweet scores for Ariana Grande and Lil Pump, who have a mode tweet sentiment of +1
 - Most celebrities have an interquartile range of 1, while Drake's tweet scores have an IQR of 2, and Kanye's tweet scores have an IQR of 0
 - The boxplots show that a lot of celebrities have many outliers in their tweet score distributions, as defined by the boxplot function
 - Distributions tend to center around neutral tweets, but some celebrities recieve far less neutral tweets than others:
   - Kanye West appeared to recieve about ~650 neutral tweets
   ~ Ariana Grande recieved only about ~250 neutral tweets, and most of her tweets were at a score of +1

Tweet Knowledge as Used to Describe Tweet Sentiment
```{r}
tweets_lm0 <- lm(score ~ 1, data = tweets,
na.action = na.exclude)
tweets_lm1 <- lm(score ~ name, data = tweets,
na.action = na.exclude)
anova(tweets_lm0, tweets_lm1)
```

Using the ANOVA function in R, we find a p-value which is less than 0.05. With that, we reject the hypothesis that the

sentiment score of a tweet cannot be predicted by knowing which celebrity the tweet pertains to.

Assessing Model Quality Using Tweet Knowledge as a Predictor
```{r}
pairwise.t.test(tweets$score, tweets$name, paired=FALSE, p.adjust.method="bonferroni")
```

The Bonferroni Correction analysis shows that the tweet distributions for most pairs of celebrities differ from eachother. However, for some pairs, it appears that we fail to reject that the tweets come from different distributions. In this example, assuming an alpha value of 0.05, we see that we fail to reject the given null hypothesis for the following pairs of celebrities:
  - Drake and Kanye West
  - Nicki Minaj and Cardi B
  - Nicki Minaj and Kanye West

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 and median sentiment scores. However, some pairs of celebrities appear to come from similar distributions.

Question 2 - Website visits (between groups - Two factors)
Set up libraries, load data...
```{r, echo=FALSE}
if (F) {
  install.packages("gmodels", dependencies=T)
}

library(gmodels)

visits = read.csv("data/webvisit1.csv", header = TRUE)
```

Conceptual model
We are tasked with analyzing the results of an A-B study of a webserver as administered in two different versions to two different groups. Also, notice that we are using the webvisit dataset 1. Through this analysis, we investigate linear modeling between groups of two different factors:

Independent Variables:
* Version of webserver (0 or 1)
* Type of User (0=consumer, 1=company)
* All combinations of the previously listed independent variables

Dependent Variables:
* Number of pages visited

In this analysis, we inspect whether the independent variables, individually and/or in combination, effected the numper of pages visited:

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.

Visual inspection
```{r}
xlim = c(0, max(visits$pages))
ylim = c(0, 250)
breaks=max(visits$pages)

# histogram of all page visits
hist(visits$pages, xlab="Number of Visits", main="Histogram of Page Visit Counts", xlim=xlim)

# histogram of page visits by different explanator variables:
# by design
par(mfrow=c(2,1))
for (v in unique(visits$version)) {
  hist(visits[which(visits$version == v),]$pages,
       xlab="Number of Visits", main=paste("Visits for version =", v),
       xlim=xlim, ylim=ylim, breaks=seq(0,breaks,1))
}

# by user type
par(mfrow=c(2,1))
for (p in unique(visits$portal)) {
```

```
  hist(visits[which(visits$portal == p),]$pages,
       xlab="Number of Visits", main=paste("Visits for portal =", p),
       xlim=xlim, ylim=ylim, breaks=seq(0,breaks,1))
}

# by all combinations:
par(mfrow=c(2,2))
for (v in unique(visits$version)) {
  for (p in unique(visits$portal)) {
    hist(visits[which(visits$version == v & visits$portal == p),]$pages,
         xlab="Number of Visits",
         main=paste("Visits for version =", v, ", portal =", p),
         xlim=xlim, ylim=ylim, breaks=seq(0,breaks,1))
  }
}
```
Upon visual inspection, it appears that the portal type doesn't change the distsribution of page visit counts. However, in both cases, it appears that the version of the website causes the mean page visit count to shift to the right, and the page visit count distributions no longer seem as right skewed.

Normality check
Statistically test if variable page visits deviates from normal distribution
```{r}
shapiro.test(visits$pages)
```
A simple Shapiro-Wilk normality test reveals, with a p-value of 2.2e-16, it is unlikely that the true distribution of page visit counts across all scenarios can be described by a linear model.

Model analysis
We construct linear models for the number of page visits as described by some factors:

* model0 - page visit count without predicting variables
* model1 - page visit count as linearly described by website version
* model2 - page visit count as linearly described by user type
* model3 - page visit count as linearly described by website version and user type, independently
* model3 - page visit count as linearly described by website version and user type, with interaction effects

```{r}
pages_model0 = lm(pages~1, data=visits,
                  na.action=na.exclude)
pages_model1 = lm(pages~version, data=visits,
                  na.action=na.exclude)
pages_model2 = lm(pages~portal, data=visits,
                  na.action=na.exclude)
pages_model3 = lm(pages~version+portal, data=visits,
                  na.action=na.exclude)
pages_model4 = lm(pages~version+portal+version:portal,
                  data=visits, na.action=na.exclude)
```

```{r}
anova(pages_model0, pages_model1)
```
Model 1 attempts to predict the page visit count using page version, and an ANOVA test finds, with p-value = 2.2e-16, that page visit count is not independent of the page version.

```{r}
anova(pages_model0, pages_model2)
```
Model 2 attempts to predict the page visit count using user type, and an ANOVA test finds, with p-value = 1.734-15, that page visit count is not independent of the user type.

```{r}
anova(pages_model3, pages_model4)
```
Because model1 and model2, which use page version and user type as predictors, can be used to explain trends in page visits, we would expect that a model using both of these predictors would work as well. This is how we constructed model3. Now we use an ANOVA test to see whether including interaction effects would further help predict the number of page visits, as we construct in model4. Through this ANOVA test with a p-value of 1.264e-12, we see that interaction effects can be further used to explain the number of page visits seen.

###Simple effect analysis
```{r}
visits$interaction = interaction(visits$portal, visits$version)
allPortalsVersion0 = c(1,-1,0,0)
allPortalsVersion1 = c(0,0,1,-1)
SimpleEff = cbind(allPortalsVersion0, allPortalsVersion1)
contrasts(visits$interaction) = SimpleEff
simpleEffectModel = aov(pages~interaction , data=visits, na.action=na.exclude)
summary.lm(simpleEffectModel)
```

```
Our analysis shows that, indeed, there is an interaction effect, but only in some cases:

* For version 0, the type of user doesn't change the page visit count. The test specifically finds a p-value of 0.317, so we don't have reason to believe that there is a statistically significant difference in page visit counts for the consumer vs business users when using this version.
* For version 1, the type of user indeed changes the page visit count. The test specifically finds a p-value of 2e-16, so we have reason to believe that there is a statistically significant difference in page visit counts for the consumer vs business users when using this version.

### Report section for a scientific publication
We analyzed the number of page visits for a given website. In particular, we inspected the effects of the type of user and the version of the website presented to the type of user. In general, we find that the number of page visits is indeed dependent on the type of user visiting the page and the version of the page. We also found that some conditions of the experiment seem to have interaction effects. In particular, we find that for users that are using portal version 0, there is a statistically significant difference in the trends of page visit counts for the business users and the consumer users.

##Question 3 - Linear regression analysis

Set up libraries, load data...
```{r, echo=FALSE}
if (F) {
  install.packages("ggpubr", dependencies=T)
  install.packages("ggExtra", dependencies=T)
  install.packages("ppcor", dependencies=T)
  install.packages("mctest", dependencies=T)
}

library(ggpubr)
library(ggExtra)
library(car)
library(mctest)
library(ppcor)

airfare <- read.csv(file="data/airfare.csv", header=T)
```

Conceptual model
For a self-guided linear regression analysis, we investigate a dataset which records airfares from ities to cities. We would like to see if the chosen independent variables can be used to predict the price of a ticket from one city to nother. For the analysis:

Dependent Variable: Average Fare
The average price of the ticket to get from City1 to City2

Independent Variables:
*
* Distance - the distance between City1 and City2
* Average Weekly Passengers - the average number of passengers that fly from City1 to City2 per week
* Lead Share - the Percentage of the flights from City1 to City2 which are served through the leading airline of the route

Visual inspection
Graphical analysis of the distribution of the dependent variable, e.g. histogram, density plot
```{r, echo=FALSE}
hist(airfare$averageFare, main="Airfare Average Prices by Route", xlab="Average Price")
```
The price of airfares appears normally distributed, centered somewhere around 175 units. The prices are right skewed, although this is expected because the price has a lower bound of 0.

Scatter plot
```{r}
# Basic Scatterplot Matrix
pairs(~averageFare+distance+leadShare+weeklyPassengers,data=airfare,
   main="Simple Scatterplot Matrix")
```
We produce a scatterplot matrix for our independent and dependent variables. Some of the scatterplots which stand out:

* Average fare seems to increase with distance, which is quite intuitive.
* There doesn't appear to be a strong relationship between the percentage of flights owned by the leading airline and the average fare price.
* It's very hard to see a relationship, visually, between the amount of weekly passengers for a route and the price of the flight. However, the routes which have extremely high weekly passenger counts seem to have lower prices. This trend is visually supported by very few data points, though.
```

Linear regression
Conduct a multiple linear regression (including confidence intervals, and beta-values)
```{r}
fare_model0 = lm(averageFare ~ 1, data=airfare, na.action=na.exclude)
confint(fare_model0)
coef(fare_model0)
```

```{r}
fare_model1 = lm(averageFare ~ distance, data=airfare, na.action=na.exclude)
confint(fare_model1)
coef(fare_model1)
anova(fare_model0, fare_model1)
```
Our ANOVA test finds that the average price of the fare is dependent on the distance of the flight. The test reports a p-value of 2.2e-16, so indeed we reject that the average fare price is independent of the distance between the cities. We also report the confidence interval and the weights of the independent variables above.

```{r}
fare_model2 = lm(averageFare ~ distance + weeklyPassengers,
                 data=airfare, na.action=na.exclude)
confint(fare_model2)
coef(fare_model2)
anova(fare_model1, fare_model2)
```
Our ANOVA test finds that the average price of the fare is dependent on the number of weekly passengers of the route. The test reports a p-value .004049, so indeed we reject that the average fare price is independent of the number of weekly passengers. We also report the confidence interval and the weights of the independent variables above.

```{r}
fare_model3 = lm(averageFare ~ distance + weeklyPassengers + leadShare,
                 data=airfare, na.action=na.exclude)
confint(fare_model3)
coef(fare_model3)
anova(fare_model2, fare_model3)
```
Our ANOVA test finds that the average price of the fare is dependent on the percentage of flights which are provided by the lead airline. The test reports a p-value .001181, so indeed we reject that the average fare price is independent of the percentage of flights which are controlled by the leading airline of the route. We also report the confidence interval and the weights of the independent variables above.

Examine assumption
```{r}
X = airfare[c('distance', 'weeklyPassengers', 'leadShare')]
Y = airfare['averageFare']
imcdiag(x=X, y=Y)
pcor(X, method='pearson')
```
The output of the partial correlation coefficients analysis shows that, for all combinations, the testing of independence of all pairs of independent variables produces p-values close to zero. That is, all pairs of independent variables are found to have some amount of correlation.

Impact analysis of individual cases
Examine effect of single cases on the predicted values (e.g. DFBeta, Cook's distance)
```{r}
plot(cooks.distance(fare_model3))
```
The plotting of cook's distance shows that all DFBeta are far less than one. Of the 1000 airfare data points, only a handful of points have values which seem to deviate from the rest, but we do not believe that these values would not reveal any sort of influence of single cases on predicted values.

Report section for a scientific publication
We performed a multiple regression on the average airfare of a route from one city to another. We found that the average airfare is dependent on the distance between cities, the amount of flights on the route owned by a dominant airline, and the average number of weekly passengers of the given route. We also found that each of these independent variables exhibit high degrees of correlation.

Question 4 - Logistic regression analysis
Set up libraries, collect data...
```{r, echo=FALSE}
library(gmodels)

shf <- read.csv("data/logisticDataStatureHandFoot.csv")
```

Conceptual model

In this logistic regression analysis, we consider some size measurements of subjects and look for a relationship between these measurements and the sex of the subject. Of course, we assume that the lengths of the hands and feet of our subjects as well as their sexes are independent of those observations in other subjects.

Dichotomous Dependent Variable: sex
Note: the experiment collected and recorded this variable as a "gender". We shall call this variable "sex" because we believe that this is what the experimenters were actually observing.

Independent Variables:
* Hand Length (in mm)
* Foot Length (in mm)

The Null Hypothesis would suggest that the independent variables, each individually and combinations of them, do not have statistically significant trends such that we can reliably predict the sex of a new subject given their feature size measurements. The Alternative Hypothesis, then, would be to say that we can reliably attempt to predict the sex of a subject given their feature size measurements because we find different sexes to exhibit different trends in these measurements.

Visualization of Data
```{r, echo=FALSE}
pairs(~handLen+footLen,data=shf,
   main="Simple Scatterplot Matrix")

p1Q4 <- ggscatter(shf, x = "handLen", y = "footLen",
               palette = "jco",
               size = 2, alpha = 0.6)
ggMarginal(p1Q4, type = "density")
plot(p1Q4)
```

Logistic Regression
```{r, echo=FALSE}
shf$sex[shf$gender ==1] <- 0
shf$sex[shf$gender ==2] <- 1

model0 <- glm(sex ~ 1, data = shf, family = binomial())
model1 <- glm(sex ~ handLen, data = shf, family = binomial())
model2 <- glm(sex ~ handLen + footLen, data = shf, family = binomial())

anova(model0, model1, model2, test="Chisq")
#pander(anova(model0,model1,model2,test = "Chisq" ),
#       caption = "Model comparison of binominal variable of sex")
```
We use of ANOVA for the comparison of our models:

* From random classification of subjects, we find that adding in Hand Lenth as a predictor improves our linear model. Statistically, the chance that Hand Length as an indicator improves our model without it being a truly good indicator is 2.2e-16: we reject that Hand Length is independent of the sex of a subject.
* Similarly, we find that additionally adding Foot Length as a predictor improves our model. Statistically, the chance that Foot Length as an indicator improves our previous model, with just Hand Length as an indicator, without it being a truly good indicator is 3.954e-8: We reject that Foot Length is independent of the sex of the subject.

Visualization of Results
```{r, echo=FALSE}
sexProbs = predict.glm(model2, shf, type="response")
sexPreds = sapply(sexProbs, function(x) as.integer(x > .5))
paste("training accuracy:", sum(diag(as.matrix(table(shf$sex, sexPreds))))/length(sexPreds))
```

Report section for a scientific publication
We find that, when predicting the sex of an individual, Hand Length and Foot Length can each be used to attempt to predict the sex of the individual. Using these measurements as indicators are both statistically significantly better than attempting to predict the sex of a person by random chance. Our model considering both of these factors achieves a training error of ~89.7%.

Part 3 - Multilevel model
Collecting data and setting up libraries...
```{r}
library(ggplot2)
library(hexbin)
library(lattice)
library(nlme)

learningData<-read.csv("./data/set1.csv", header = TRUE)
```

Visual inspection
```{r}
hist(learningData$score, xlab="Score", main="Distribution of score")
plot(hexbin(learningData$score ~ learningData$session, xbins=50, xlab="session number", ylab="score"))
xyplot(score~session | Subject, data=learningData[learningData$Subject %in% seq(1,191,10),])
```
Through a visual inspection:

* scores alone seem normally distributed
* When plotting a scatterplot of scores against session number, it appears that the score of a testee still centers around ~100 as session number increases, although one might see that the center increases slightly with session number. Also, the spread of scores increases as a function of session number.
* We sample some ~20 subjects from our data and plot their scores as a function of the session after which they were tested. We see quite inconsistent trends, where some subjects score similarly before and after sessions, some subjects have increasing scores over time, and some subjects even have decreasing scores over time.

Multilevel analysis with scentific findings
Is there significant variance between the participants in their score?
```{r}
randomInterceptOnly <- lme(score ~ 1, data = learningData,
                           random = ~1|Subject, method = "ML")
summary(randomInterceptOnly)
intervals(randomInterceptOnly, 0.95)
```
We find that there is very high variance between the scores of each subject on a given session. With a p-value of 0, we find approximately no chance that, if the distributions of scores as a function of lesson number came from the same distribution, we would find a collection of this sort of data. That is, We reject the null hypothesis that the scores of a subject as a function of the number of lessons he had recieved are not from the same distribution. Subjects' scores respond differently to recieving lessons.

Does session have an impact on people score?
```{r}
randomInterceptSession <- lme(score ~ session,
                              data = learningData, random = ~1|Subject, method = "ML")
summary(randomInterceptSession)
anova(randomInterceptOnly,randomInterceptSession)
```
We perform an analysis to observe the linear regressions of scores on the number of sessions from a subject across all subjects. We create linear models with and without use of session as an explanatory variable. We then compare these linear models, and with a p-value of <.001, we find that it cannot be the case that scores of a subject are independent of the number of sessions they attend. That is, constructing a linear model for the score of a subject is changed when we consider the number of sessions they attend.