Report Template coursework assignment A - 2021

CS4125 Seminar Research Methodology for Data Science

Thomas Bos (4543408), Daniël van Gelder (4551028), Jessie van Schijndel (5407397)

20/04/2021

Contents

1 Part 1 - Design and set-up of true experiment **2**

1.1 The motivation for the planned research . . . 2
1.2 The theory underlying the research . . . 2
1.3 Research questions . . . 3
1.4 The related conceptual model . . . 3
1.4.1 Independent Variable (IV) . . . 3
1.4.2 Dependent Variable (DV) . . . 3
1.4.3 Mediating Variable . . . 3
1.4.4 Moderating Variable . . . 3
1.5 Experimental Design . . . 3
1.6 Experimental procedure . . . 4
1.7 Measures . . . 4
1.8 Participants . . . 4
1.9 Suggested statistical analyses . . . 4

2 Part 2 - Generalized linear models **5**

2.1 Question 1 Twitter sentiment analysis (Between groups - single factor) . . . 5
2.1.1 Conceptual model . . . 5
2.1.2 Collecting tweets, and data preparation . . . 5
2.1.3 Homogeneity of variance analysis . . . 5
2.1.4 Visual inspection Mean and distribution sentiments . . . 5
2.1.5 Frequentist approach . . . 7
2.1.6 Bayesian Approach . . . 8
2.2 Question 2 - Website visits (between groups - Two factors) . . . 13
2.2.1 Conceptual model . . . 16
2.2.2 Visual inspection . . . 16

2.2.3 Normality check 19
2.2.4 Frequentist Approach 21
2.2.5 Bayesian Approach 24

3 Part 3 - Multilevel model **27**
3.1 Visual inspection 27
3.2 Frequentist approach 32
3.2.1 Multilevel analysis 32
3.2.2 Report section for a scientific publication 34
3.3 Bayesian approach 34
3.3.1 Model description 34
3.3.2 Model comparison 35
3.3.3 Estimates examination 37

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

1.1 The motivation for the planned research

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.

1.2 The theory underlying the research

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.

1.3 Research questions

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.

1.4 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:

1.4.1 Independent Variable (IV)

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

1.4.2 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.

1.4.3 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.

1.4.4 Moderating Variable

There are several factors which may a moderating effect on the relationship between the IV and the DV which are difficult 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

1.5 Experimental Design

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 trail 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.

1.6 Experimental procedure

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.

1.7 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 to tests for each particpant: the score of the post-test divided by the score of the pre-test.

1.8 Participants

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.

1.9 Suggested statistical analyses

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 ass 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.

2 Part 2 - Generalized linear models

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

2.1.1 Conceptual model

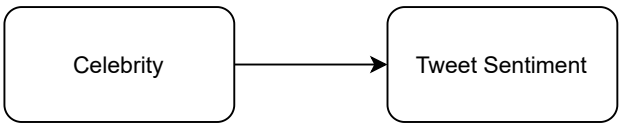


Figure 1: The conceptual model. Different attributes of a celebrity are shown which may influence the sentiment of tweets related to a certain celebrity.

2.1.2 Collecting tweets, and data preparation

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.

2.1.3 Homogeneity of variance analysis

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.

```
library(car)
library(pander)
```

```
boxplot(semFrame$score ~ semFrame$Candidate)
```

```
pander(leveneTest(semFrame$score, semFrame$Candidate, center = median))
```

Table 1: Levene's Test for Homogeneity of Variance (center = median)

	Df	F value	Pr(>F)
group	2	31.39	6.663e-14
	897	NA	NA

2.1.4 Visual inspection Mean and distribution sentiments

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.

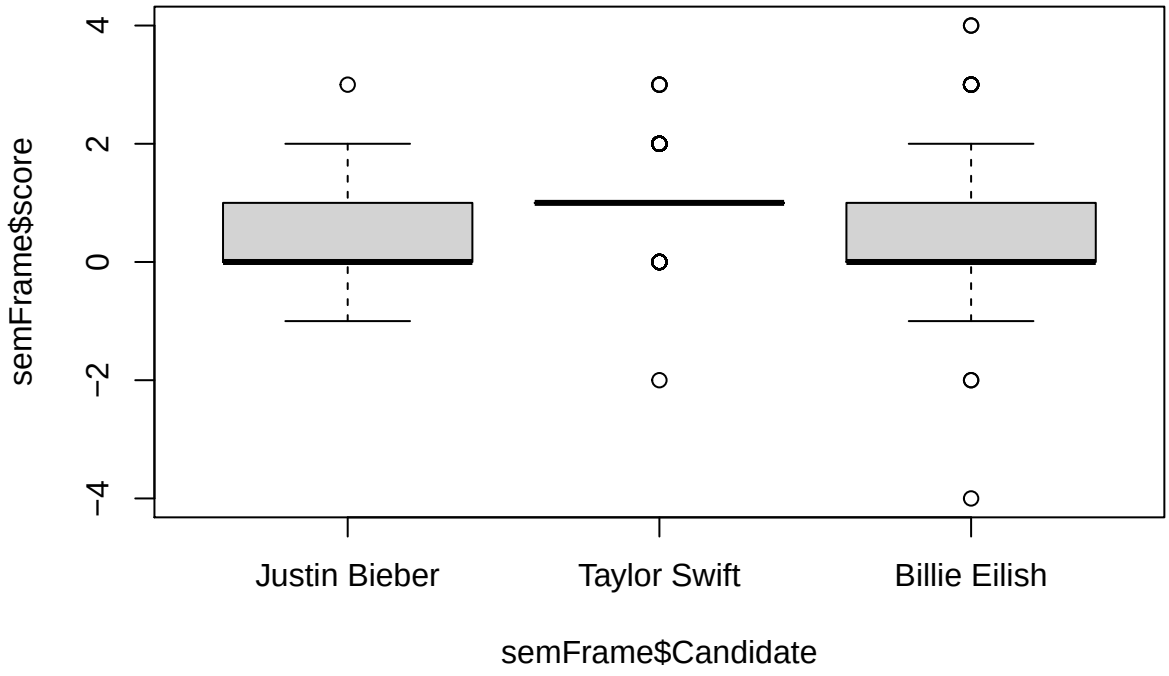


Figure 2: Boxplot of the sentiment values of the tweets for each celebrity.

```
library(ggplot2)

p <- semFrame %>% ggplot( aes(x=score)) + geom_histogram( color="#e9ecef", alpha=0.6, position = 'identi
```

```
plot(p)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

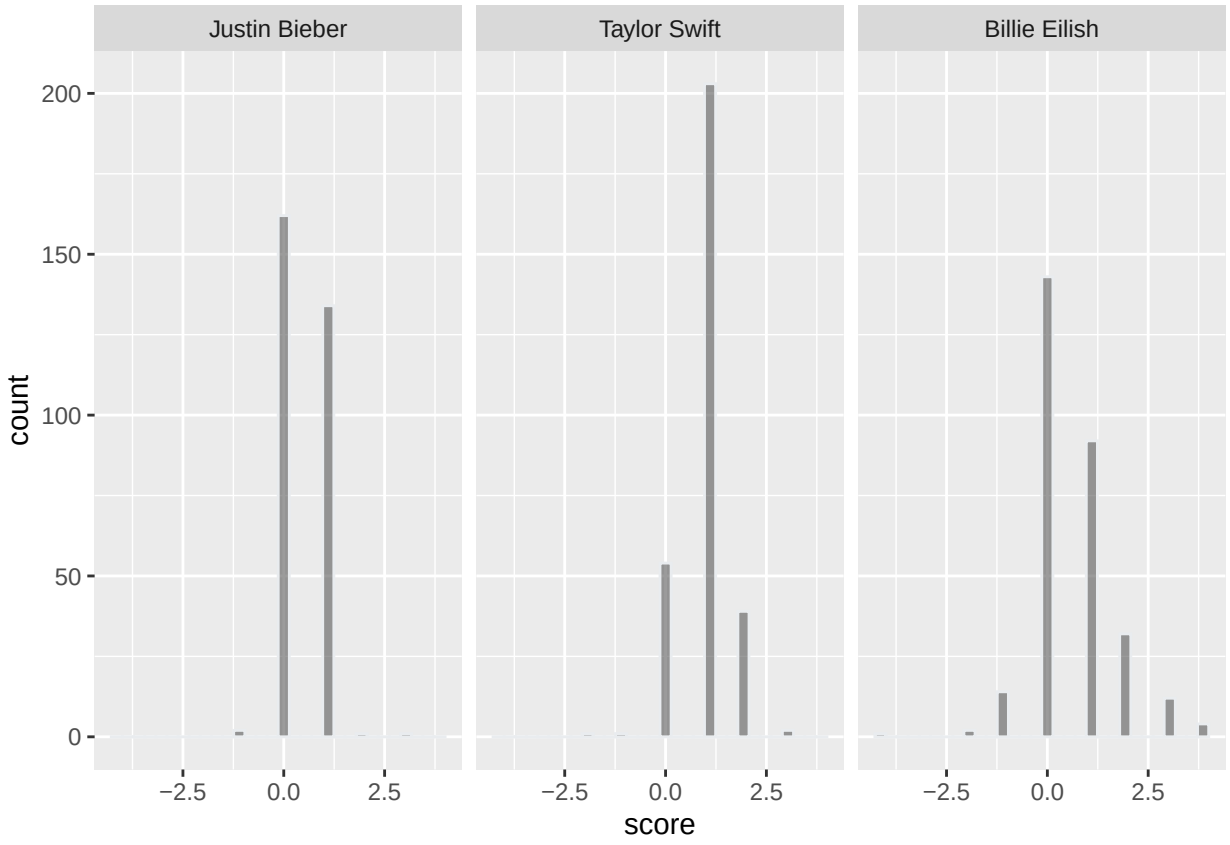


Figure 3: Bar plot for the tweet sentiment values for each celebrity.

2.1.5 Frequentist approach

2.1.5.1 Linear model 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.

```
#include your code and output in the document
library(pander)
library(multcomp)

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

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

```
pander(res.aov)
```

Table 2: Analysis of Variance Model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Candidate	2	37.35	18.67	32.57	2.208e-14
Residuals	897	514.3	0.5733	NA	NA

2.1.5.2 Post Hoc analysis 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.

```
library(stats)
```

```
TukeyHSD(res.aov)
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = score ~ Candidate, data = semFrame, na.action = na.exclude)
##
## $Candidate
##                                   diff         lwr        upr     p adj
## Taylor Swift-Justin Bieber   0.4900000  0.34486375  0.6351362 0.0000000
## Billie Eilish-Justin Bieber  0.1633333  0.01819709  0.3084696 0.0228031
## Billie Eilish-Taylor Swift  -0.3266667 -0.47180291 -0.1815304 0.0000005
```

2.1.5.3 Report section for a scientific publication 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).

2.1.6 Bayesian Approach

2.1.6.1 Model description 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 μ, and we allow σ 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,

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

```r
semFrame <- subset(semFrame, select = c(score, CandidateF))
x <- seq(-5, 5, length=100)
y <- dnorm(x, mean=0, sd=2)
```

```r
hist(semFrame$score)
```

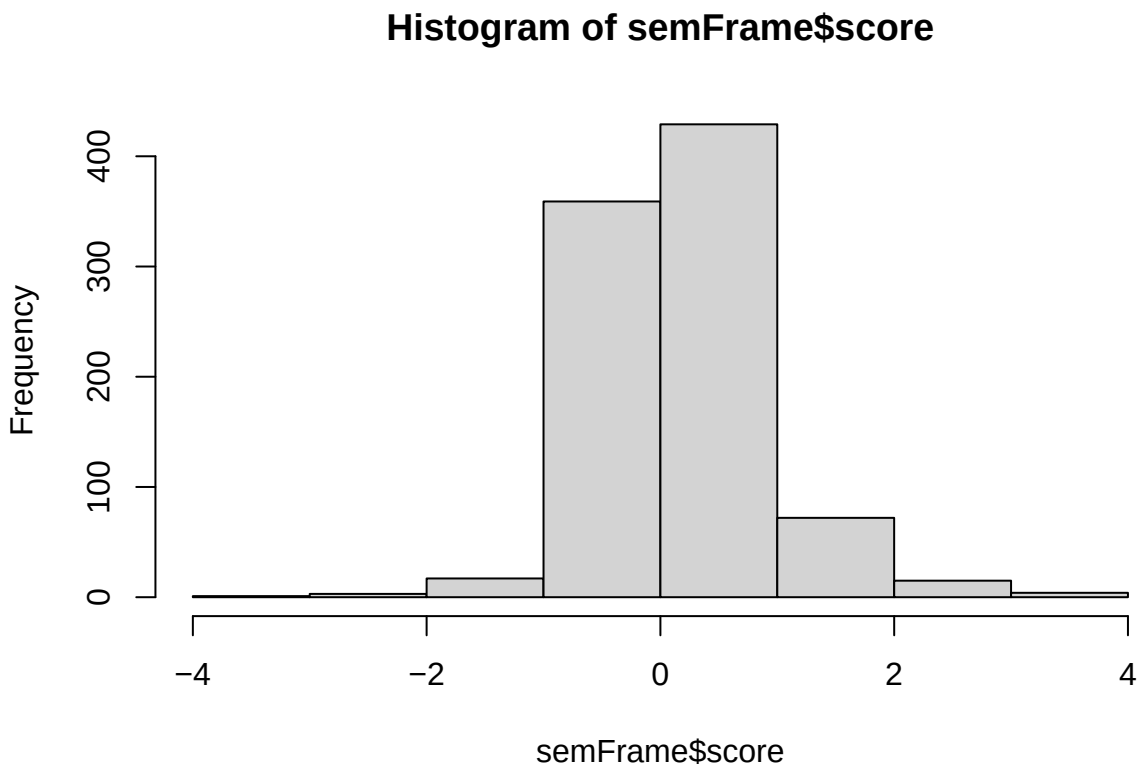


Figure 4: Sentiment values of all tweets.

```r
plot(x, y, type="l", lwd=1)
```

2.1.6.2 Model comparison 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
# ```{r STAN CHUNK}
#include your code and output in the document
library("rstan")
library("rethinking")

m0 <-ulam(
```

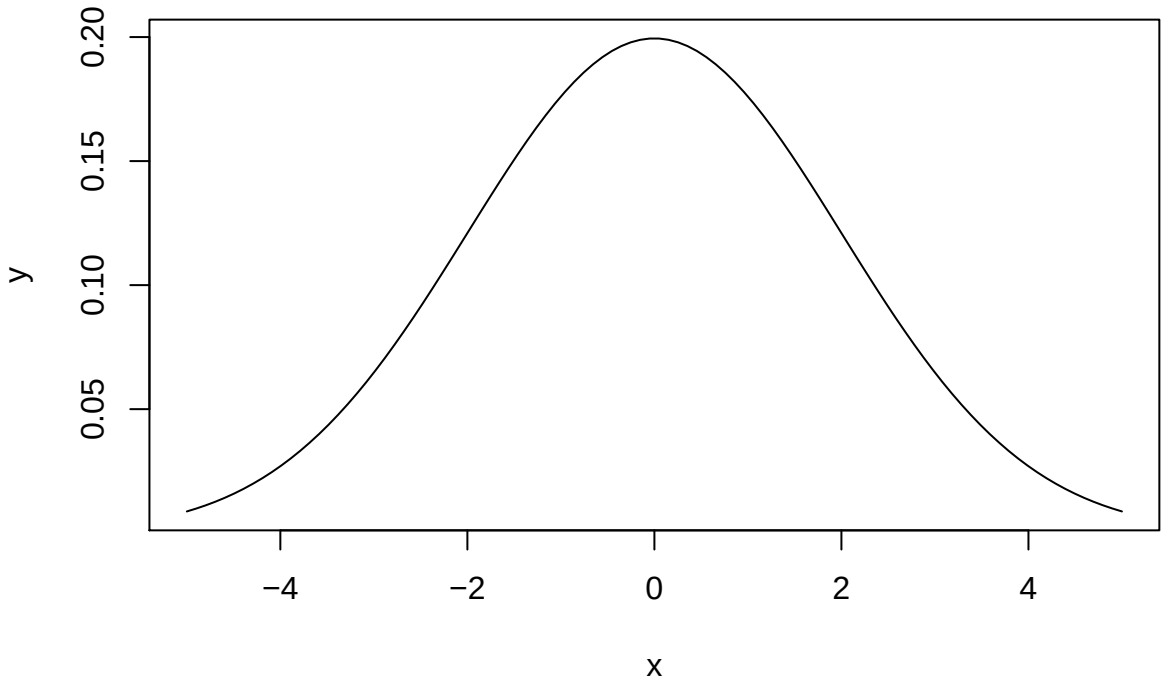


Figure 5: Normal distribution with a mean of 0 and a standard deviation of 2.

```
  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
)
```

```
plot(precis(m1, depth=2, prob = .95))
```

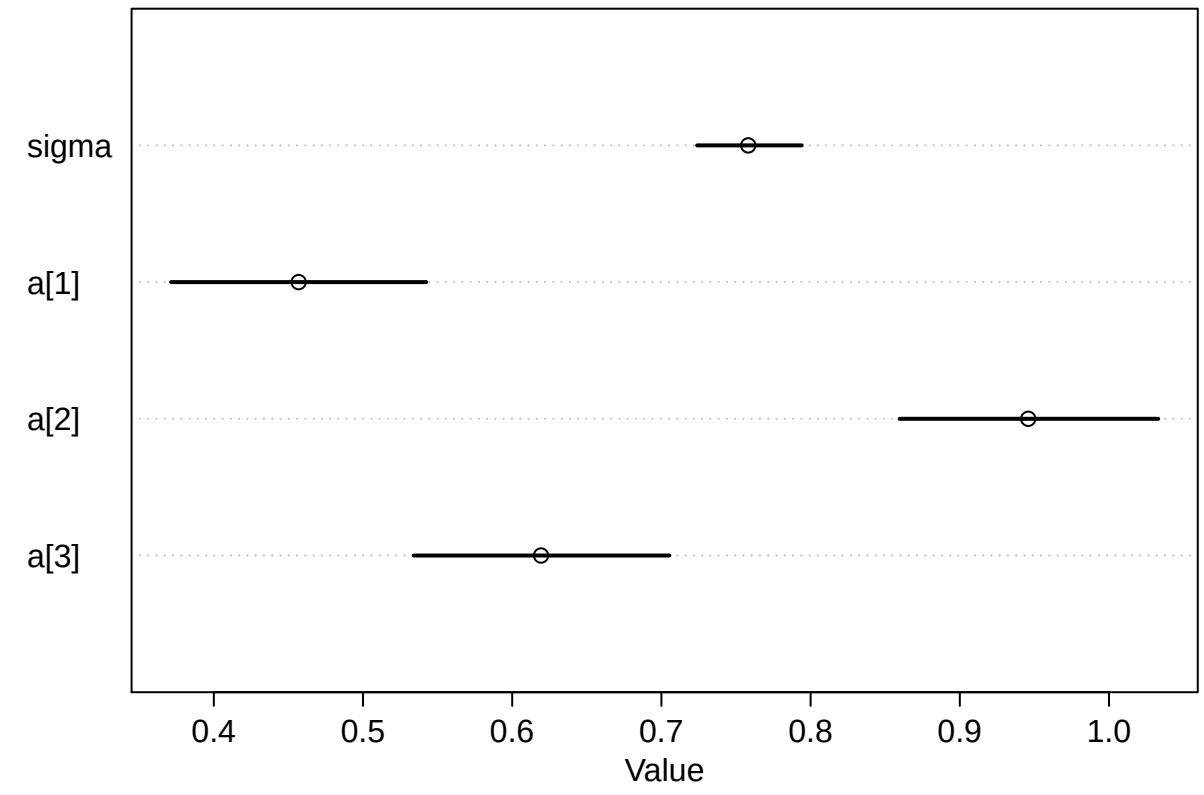


Figure 6: Confidence intervals for the parameters of the model with celebrity added as a predictor.

```
plot(compare(m0, m1, func=WAIC))
```

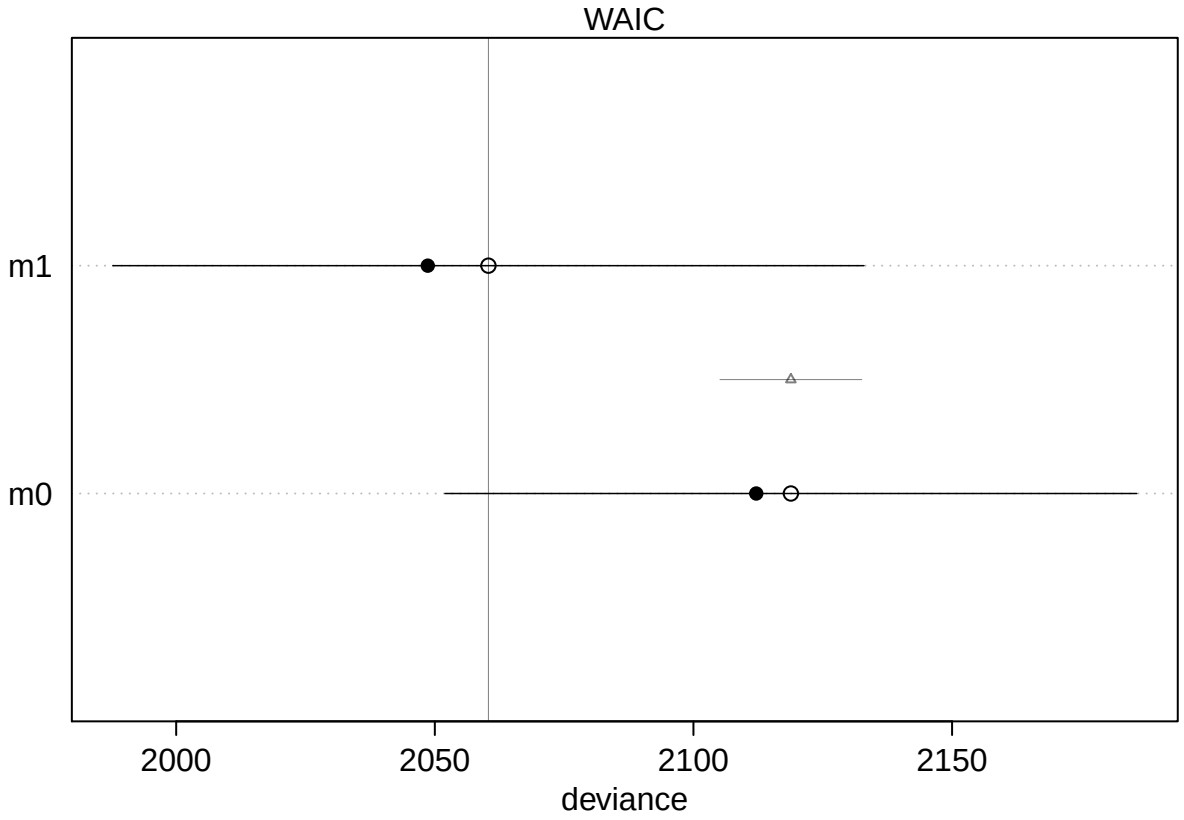


Figure 7: Comparison of the WAIC score for each model.

2.1.6.3 Comparison celebrity pair 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.

```
# 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"))
```

```
plot(bayes.t.test(jbSub$score, tsSub$score))
```

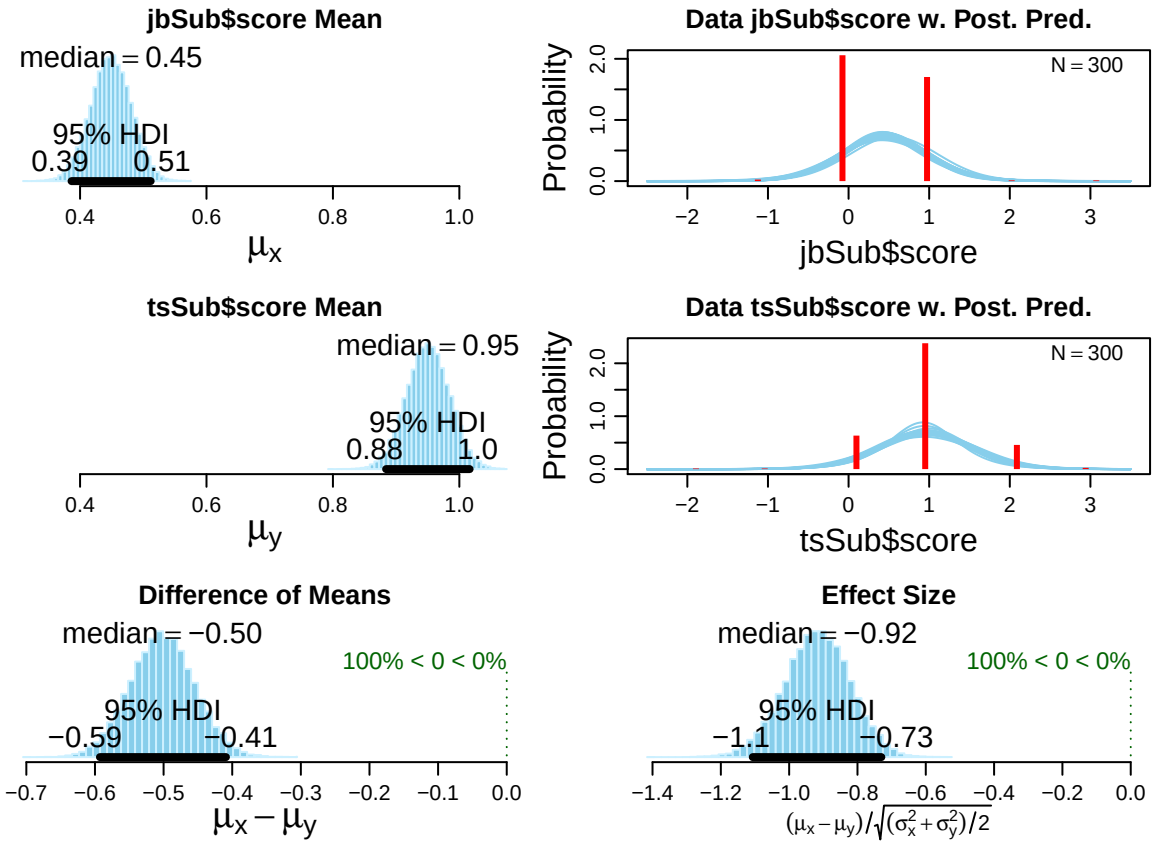


Figure 8: Results of performing the bayes t-test on the pair Justin Bieber - Taylor Swift.

```
plot(bayes.t.test(jbSub$score, beSub$score))
```

```
plot(bayes.t.test(tsSub$score, beSub$score))
```

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

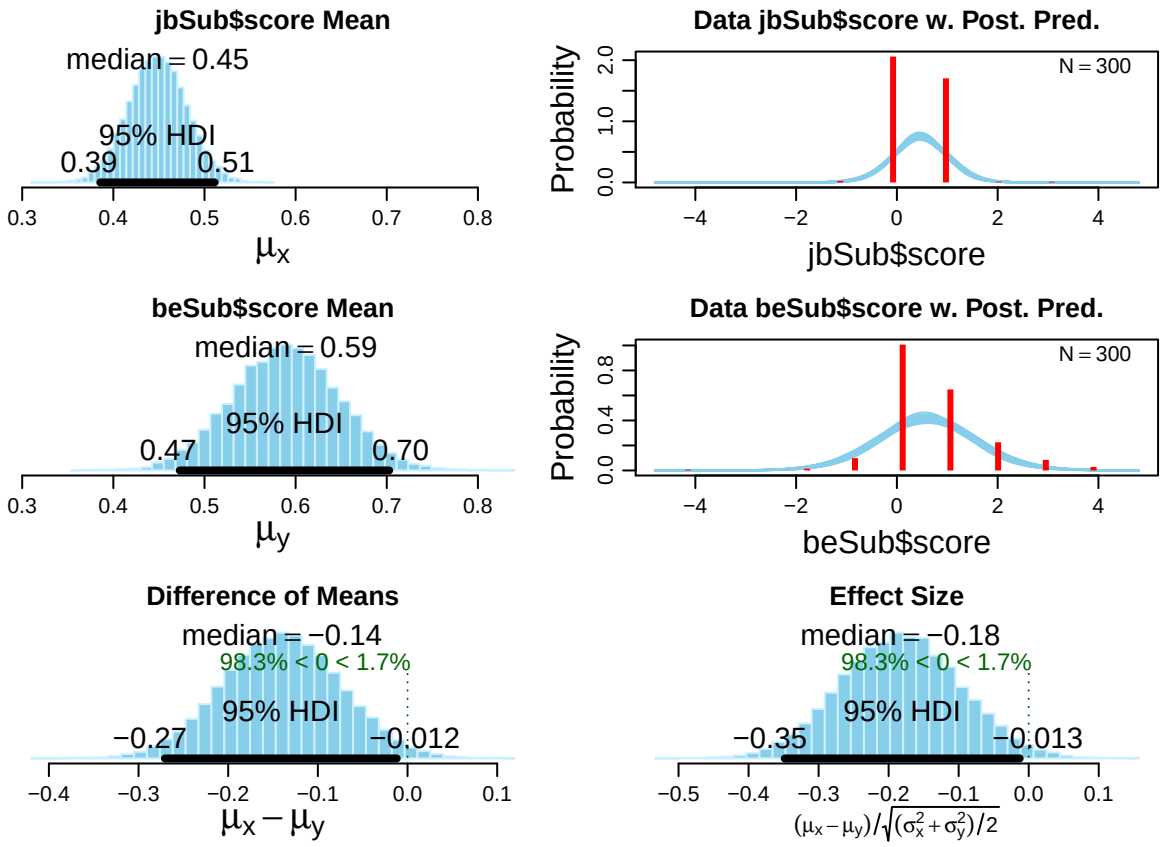


Figure 9: Results of performing the bayes t-test on the pair Justin Bieber - Billie Eilish.

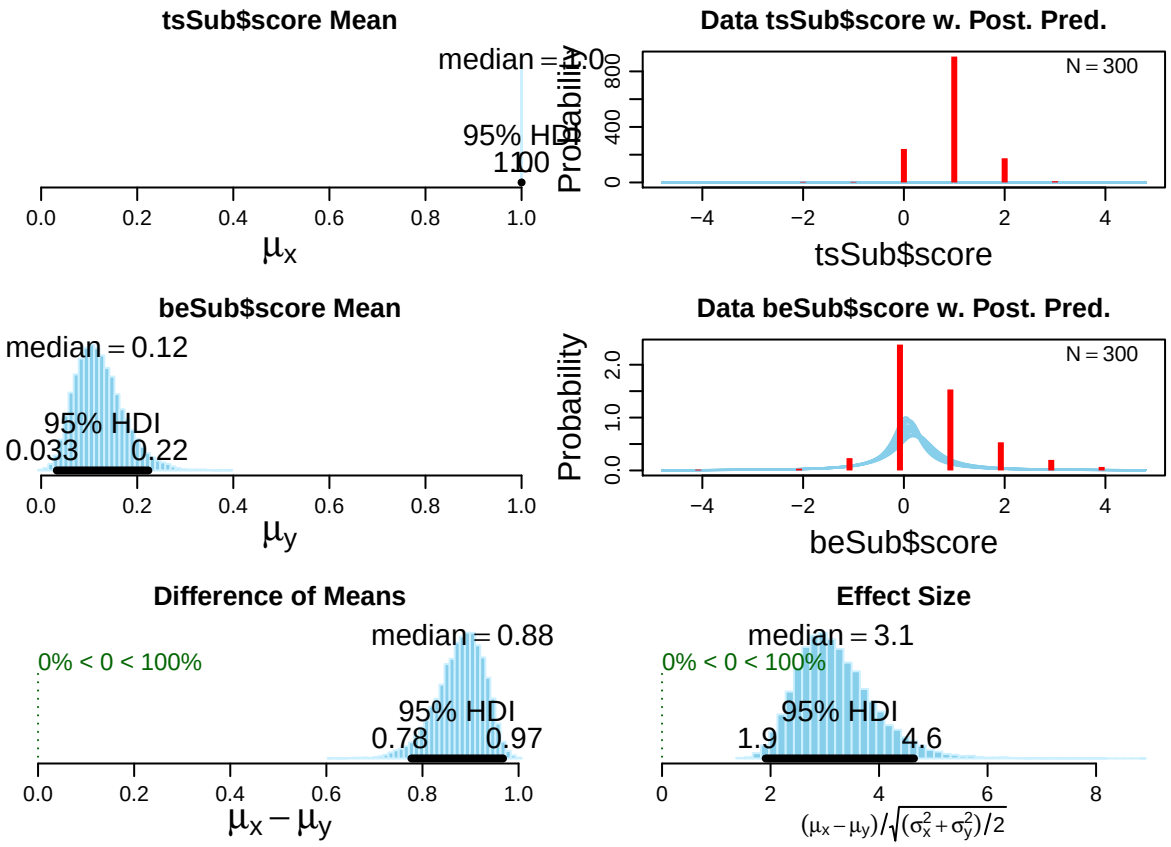


Figure 10: Results of performing the bayes t-test on the pair Taylor Swift - Billie Eilish.

2.2.1 Conceptual model

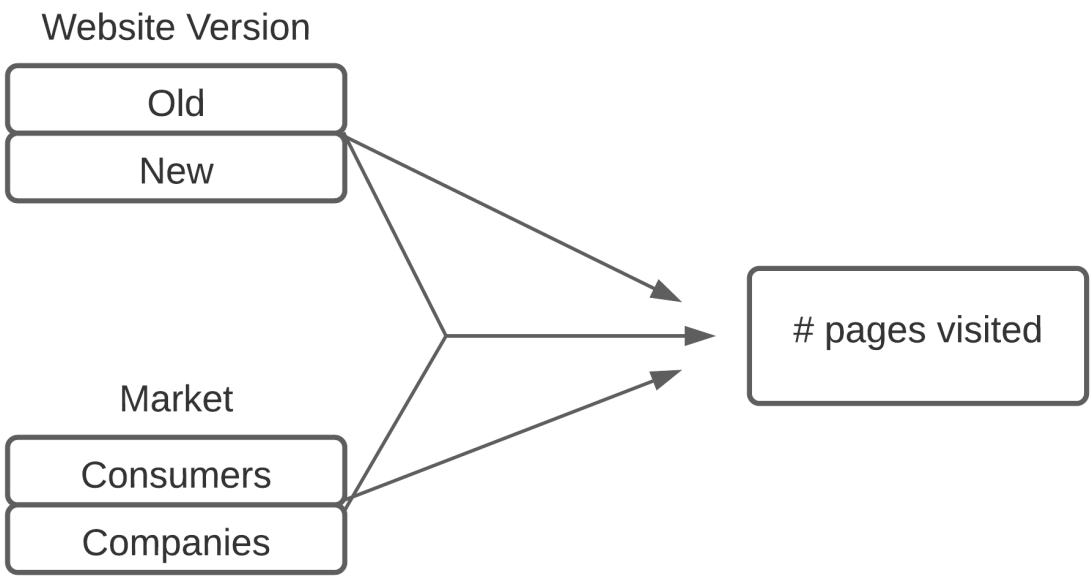


Figure 11: Conceptual Model underlying this research question

2.2.2 Visual inspection

The first step in our inspection is to observe the distribution of the dependent variable overall regardless of the factors.

```
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"))
```

```
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:

```
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:

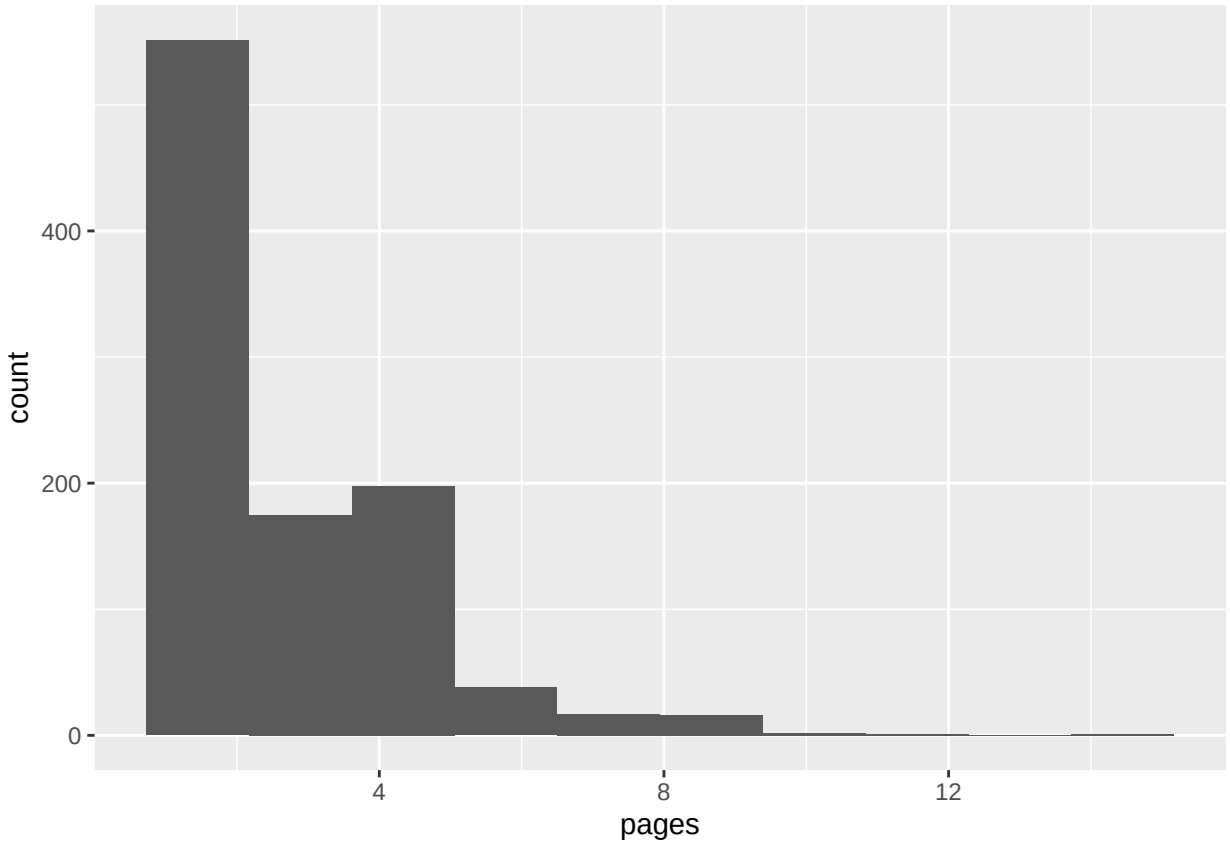


Figure 12: Histogram showing page visits on the website over all factors.

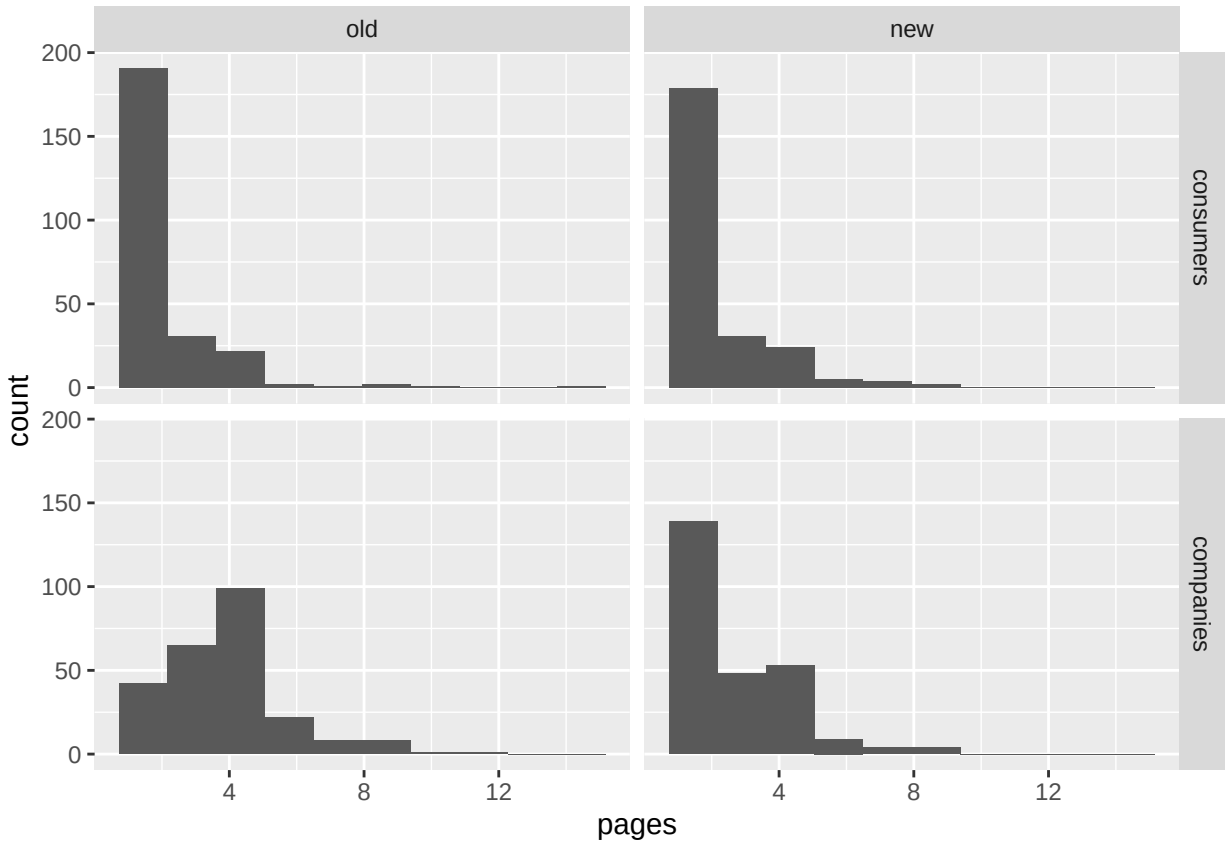
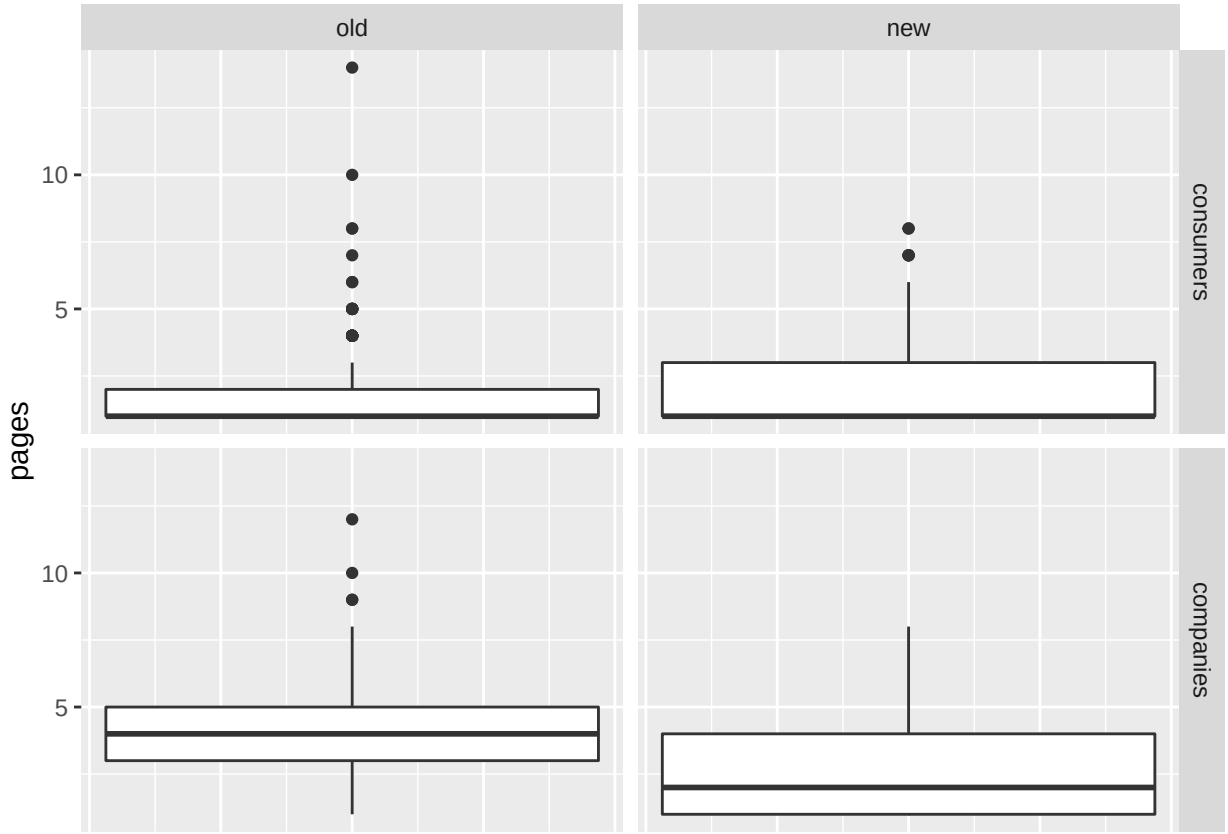


Figure 13: Histograms of page visits on the website per independent variable.

```
p <- web_visit %>% ggplot( aes(y=pages)) + geom_boxplot() + facet_grid(portal ~ version, scales = "free
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.

2.2.3 Normality check

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 μ parameter the average of the data and as σ the variance of the data.

```
m<-mean(web_visit$pages)
std<-sqrt(var(web_visit$pages))
```

```
p <- web_visit %>% ggplot( aes(x=pages)) + geom_histogram(aes(y =..density..), bins=15) + stat_function

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.

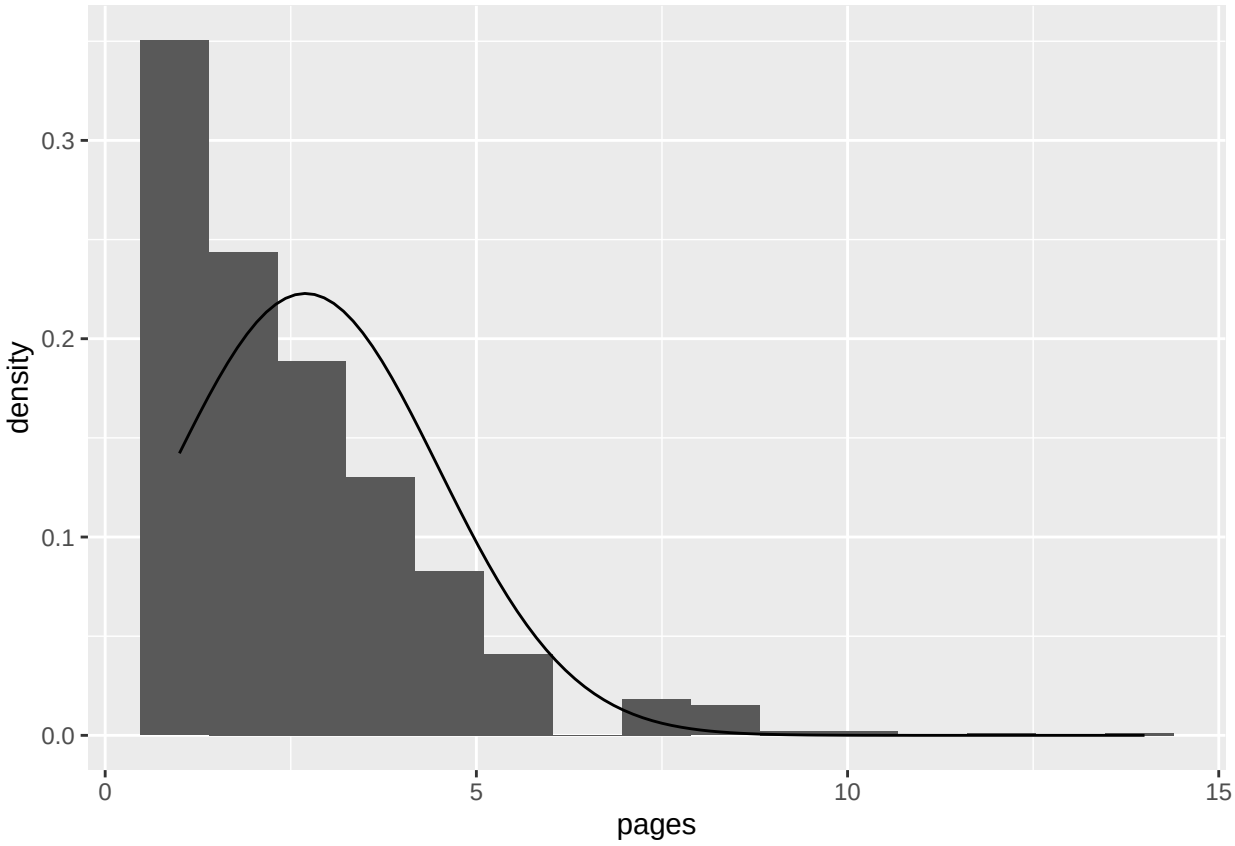


Figure 14: Normal distribution fitted to data of page visits

2.2.4 Frequentist Approach

2.2.4.1 Model analysis First we perform Levene's Test to determine the homogeneity of variance.

```
library(car)
library(pander)

pander(leveneTest(web_visit$pages, interaction(web_visit$version , web_visit$portal)))
```

Table 3: Levene's Test for Homogeneity of Variance (center = median) 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.

	Df	F value	Pr(>F)
group	3	2.562	0.05358
	995	NA	NA

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.

```
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)
```

```
pander(anova(model0,model1), caption = "Website version as main effect on page visits")
```

Table 4: Website version as main effect on page visits

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
998	3199	NA	NA	NA	NA
997	3107	1	92.2	29.59	6.731e-08

```
pander(anova(model0,model2), caption = "Portal type as main effect on page visits")
```

Table 5: Portal type as main effect on page visits

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
998	3199	NA	NA	NA	NA
997	2751	1	448.2	162.4	1.409e-34

```
pander(anova(model3,model4),caption = "Interation effect on top of two main effects")
```

Table 6: Interation effect on top of two main effects

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
996	2652	NA	NA	NA	NA
995	2534	1	117.8	46.25	1.793e-11

```
pander(anova(model4),caption = "Effect of Website version, Portal type and interaction effect on page v
```

Table 7: Effect of Website version, Portal type and interaction effect on page visits

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
version	1	92.2	92.2	36.2	2.495e-09
portal	1	455.3	455.3	178.8	1.283e-37
version:portal	1	117.8	117.8	46.25	1.793e-11
Residuals	995	2534	2.547	NA	NA

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:

```
library(AICcmodavg)
models <-list(model0, model1, model2, model3, model4)
model.names <-c("model0","model1","model2","model3","model4")
```

```
pander(aictab(cand.set = models, modnames=model.names))
```

	Modnames	K	AICc	Delta_AICc	ModelLik	AICcWt	LL	Cum.Wt
5	model4	5	3775	0	1	1	-1882	1
4	model3	4	3818	43.37	3.824e-10	3.824e-10	-1905	1
3	model2	3	3853	78.07	1.117e-17	1.117e-17	-1923	1
2	model1	3	3975	199.6	4.451e-44	4.451e-44	-1984	1
1	model0	2	4002	226.8	5.517e-50	5.517e-50	-1999	1

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.

2.2.4.2 Simple effect analysis 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.

```
bar <- ggplot(web_visit, aes(portal , pages, fill = version))
bar + stat_summary(fun.y = mean, geom = "bar", position="dodge")
```

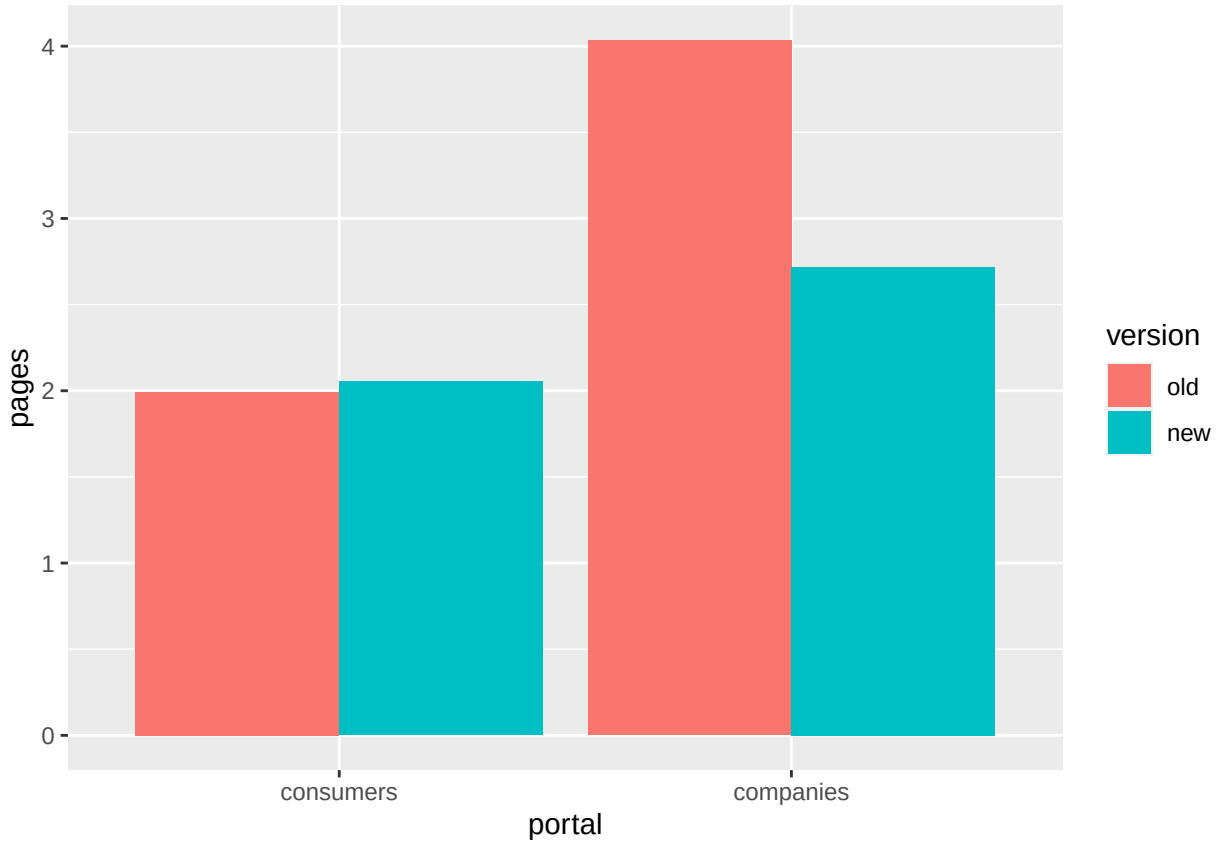


Figure 15: Bar plot showing the mean page visits per independent variable.

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.

```
web_visit$simple <- interaction(web_visit$version, web_visit$portal) #merge two factors levels(Lec7g$si
levels(web_visit$simple)
```

```
## [1] "old.consumers" "new.consumers" "old.companies" "new.companies"
```

```
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
```

```
simpleEffectModel <-lm(pages ~ simple , data = web_visit, na.action = na.exclude)
pander(summary.lm(simpleEffectModel))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.699	0.0505	53.45	9.614e-295
simplecontrastConsumers	-0.03051	0.07166	-0.4258	0.6703
simplecontrastCcompanies	0.6563	0.07117	9.222	1.695e-19
simple	1.354	0.101	13.4	8.88e-38

Table 10: Fitting linear model: pages ~ simple

Observations	Residual Std. Error	R^2	Adjusted R^2
999	1.596	0.2079	0.2056

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.

2.2.4.3 Report section for a scientific publication 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.

2.2.5 Bayesian Approach

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

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 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.

```
p <- web_visit %>% ggplot( aes(x=pages)) + geom_histogram(aes(y =..density..), bins=15) + stat_function
plot(p)
```

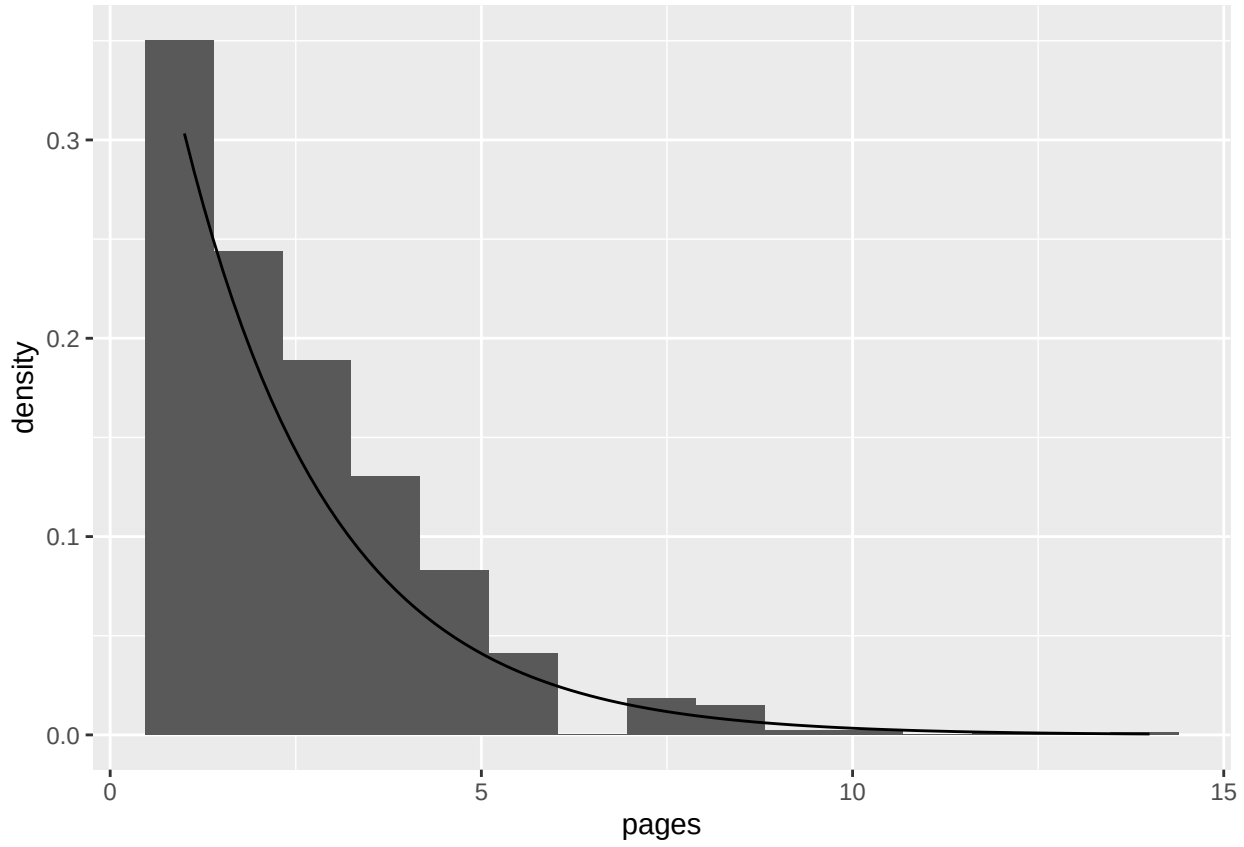


Figure 16: Gamma distribution (1, 2) fitted to data of page visits

2.2.5.2 Model comparison 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.

```
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.

```
plot(compare(m0,m1,m2,m3,m4, func=WAIC))
```

```
pander(compare(m0,m1,m2,m3,m4))
```

	WAIC	SE	dWAIC	dSE	pWAIC	weight
m4	3786	84.14	0	NA	6.737	1
m3	3821	82.16	34.53	8.273	6.26	3.175e-08
m2	3855	82.21	69.04	15.65	5.229	1.019e-15

	WAIC	SE	dWAIC	dSE	pWAIC	weight
m1	3976	69.38	189.7	28.76	4.266	6.343e-42
m0	4003	69.95	216.9	30.94	3.388	7.91e-48

```
pander(precis(m4, prob= .95))
```

	mean	sd	2.5%	97.5%	n_eff	Rhat4
a	0.1399	0.1282	0.004442	0.4752	1464	1.002
b	0.4526	0.1288	0.1825	0.6952	999.9	1.008
c	2.411	0.1266	2.136	2.645	1107	1.004
d	-0.7672	0.09913	-0.9581	-0.5666	802.4	1.01
sigma	1.606	0.03684	1.536	1.68	1873	1.001

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.

3 Part 3 - Multilevel model

3.1 Visual inspection

For this exericse, we use `set1.csv`. Below we inspect the distribution of the score through a histogram of the 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.

```
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.

```
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)
```

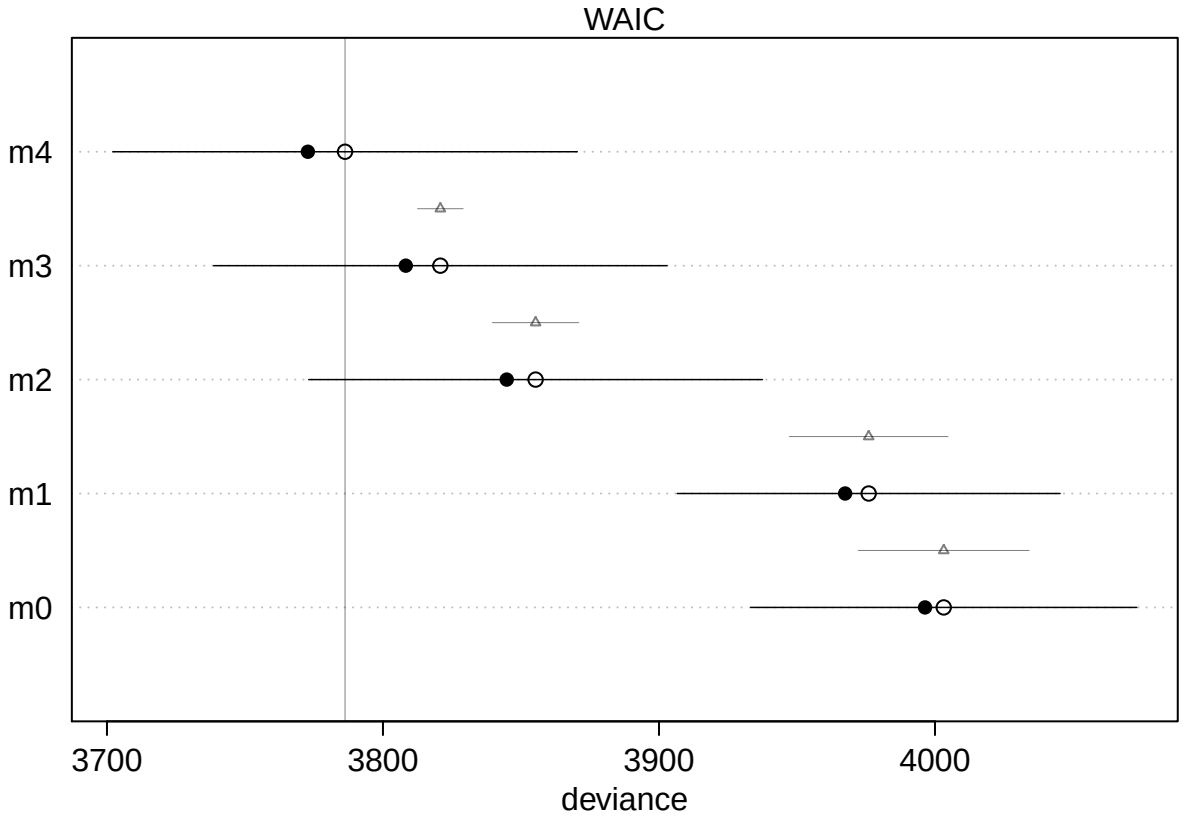


Figure 17: WAIC comparison between Bayesian models capturing effects of different variables.

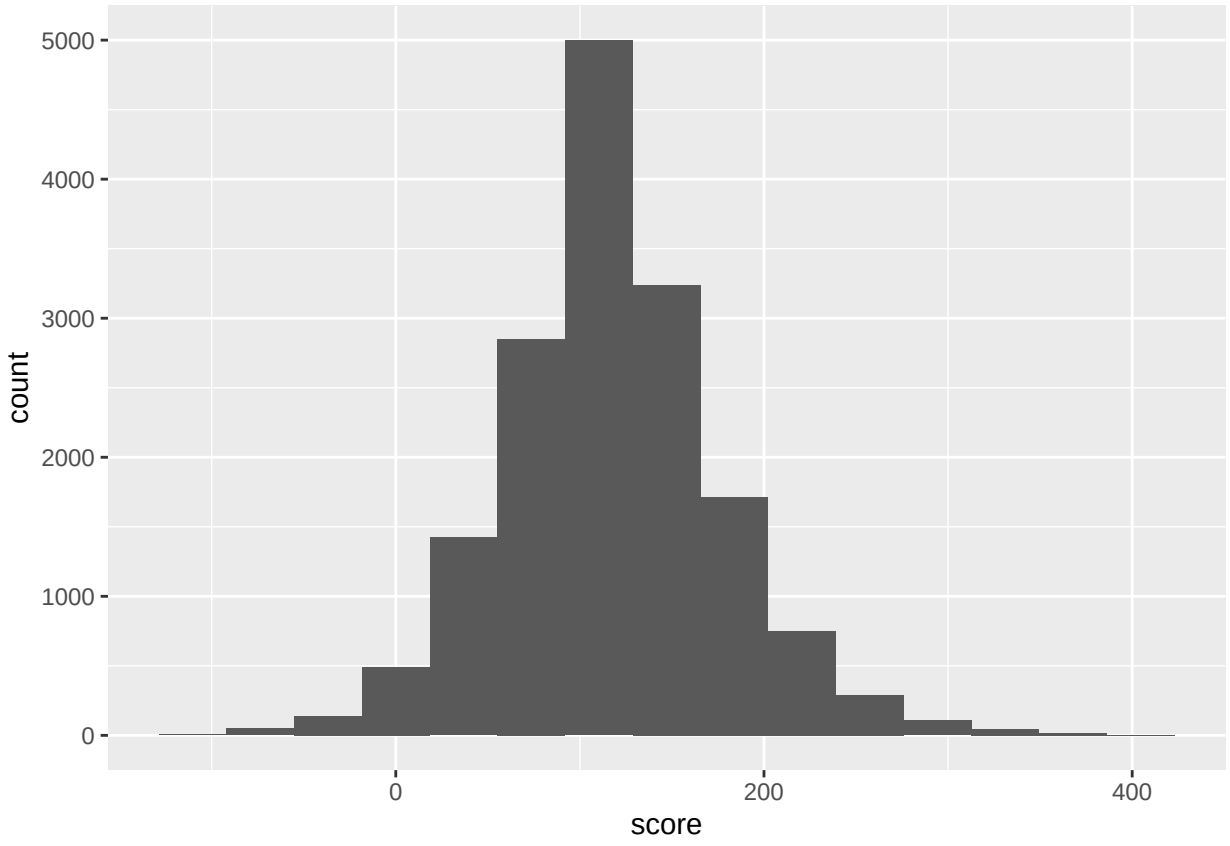


Figure 18: Histogram of score

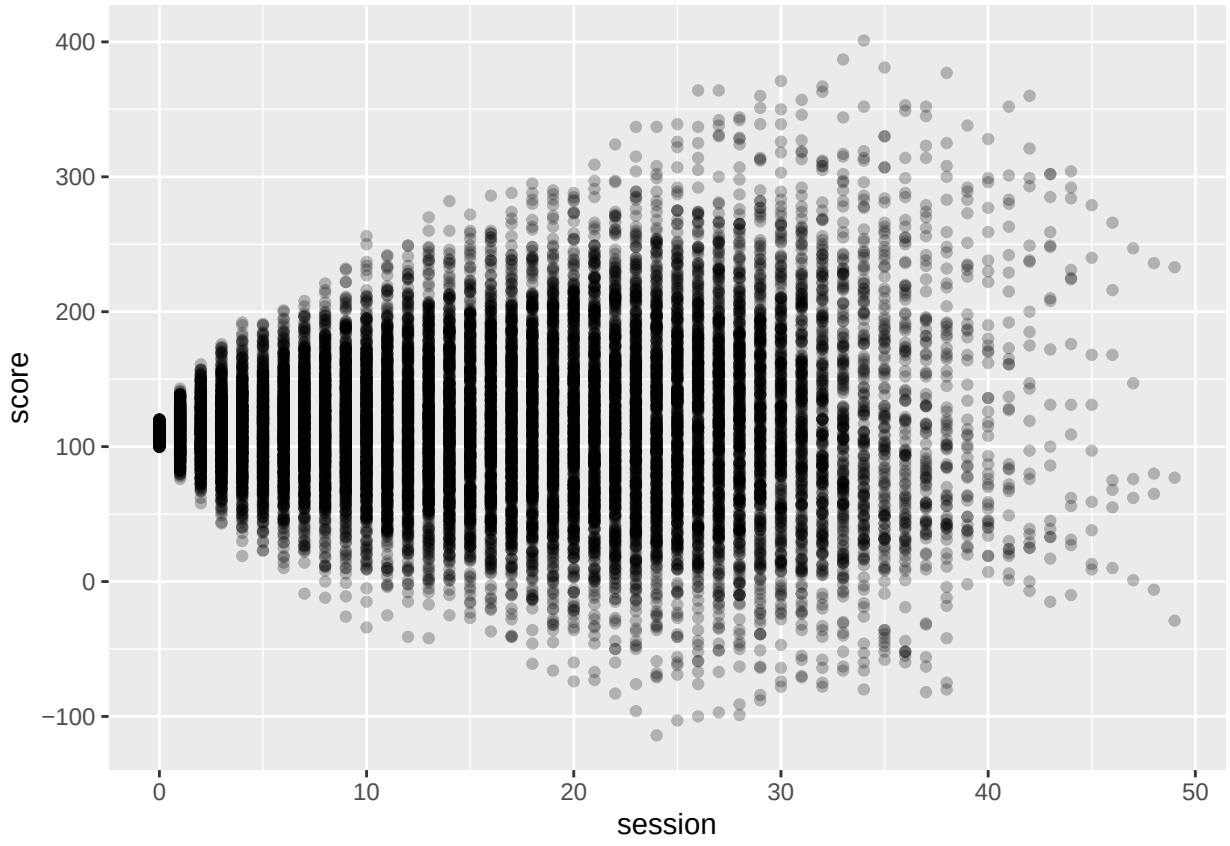


Figure 19: Scatterplot relation between score and session

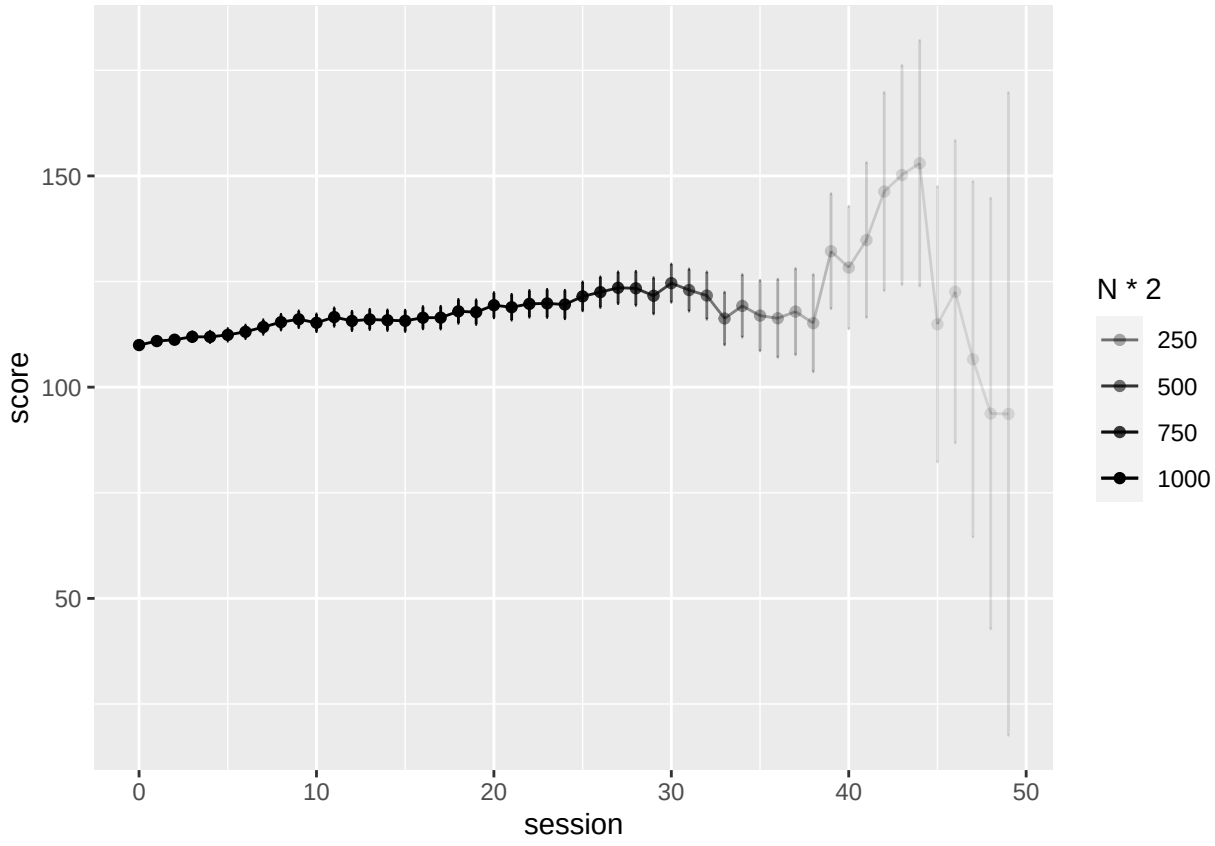


Figure 20: Mean score and standard error over sessions.

3.2 Frequentist approach

3.2.1 Multilevel analysis

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 (`~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.

```
library(nlme)

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

```
## Linear mixed-effects model fit by maximum likelihood
##   Data: exp_data
##        AIC    BIC    logLik
##   162710.9 162734 -81352.45
##
## Random effects:
##  Formula: ~1 | Subject
##         (Intercept) Residual
## StdDev:    46.52747 35.25763
##
## Fixed effects:  score ~ 1
##                Value Std.Error    DF  t-value p-value
## (Intercept) 116.8139  2.097897 15627 55.68142       0
##
## Standardized Within-Group Residuals:
##         Min          Q1         Med          Q3         Max
## -4.22644590 -0.61530909  0.01016836  0.62959973  4.10477262
##
## Number of Observations: 16128
## Number of Groups: 501
```

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.

```
intervals(model0, 0.95)
```

```
## Approximate 95% confidence intervals
##
##  Fixed effects:
##                lower     est.    upper
## (Intercept) 112.7019 116.8139 120.9259
## attr(,"label")
## [1] "Fixed effects:"
##
##  Random Effects:
```

```
##   Level: Subject
##                     lower     est.    upper
## sd((Intercept)) 43.68633 46.52747 49.55338
##
##  Within-group standard error:
##    lower     est.    upper
## 34.86891 35.25763 35.65067
```

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.

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

```
## Linear mixed-effects model fit by maximum likelihood
##   Data: exp_data
##        AIC      BIC    logLik
##   162545.2 162575.9 -81268.58
##
## Random effects:
##  Formula: ~1 | Subject
##         (Intercept) Residual
## StdDev:     46.5146 35.06933
##
## Fixed effects:  score ~ session
##                Value Std.Error    DF  t-value p-value
## (Intercept) 111.0676  2.143371 15626 51.81911       0
## session       0.3682  0.028356 15626 12.98493       0
##  Correlation:
##         (Intr)
## session -0.206
##
## Standardized Within-Group Residuals:
##          Min           Q1          Med           Q3          Max
## -4.120041920 -0.613554431  0.009847298  0.627208531  3.952634923
##
## Number of Observations: 16128
## Number of Groups: 501
```

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.

```
intervals(model1, 0.95)
```

```
## Approximate 95% confidence intervals
##
##  Fixed effects:
##                    lower        est.       upper
## (Intercept) 106.8665678 111.0675622 115.2685566
## session       0.3126229   0.3682005   0.4237781
## attr(,"label")
## [1] "Fixed effects:"
```

```
##
##  Random Effects:
##   Level: Subject
##                     lower    est.    upper
## sd((Intercept)) 43.67471 46.5146 49.53914
##
##  Within-group standard error:
##    lower     est.    upper
## 34.68269 35.06933 35.46028
```

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

Table 13: Model comparison. (continued below)

	call	Model	df	AIC	BIC
model0	lme.formula(fixed = score ~ 1, data = exp_data, random = ~1 | Subject, method = "ML")	1	3	162711	162734
model1	lme.formula(fixed = score ~ session, data = exp_data, random = ~1 | Subject, method = "ML")	2	4	162545	162576

	logLik	Test	L.Ratio	p-value
model0	-81352		NA	NA
model1	-81269	1 vs 2	167.7	2.317e-38

3.2.2 Report section for a scientific publication

To investigate whether session has an impact on people's scores and wheter 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), $p = 0$.

3.3 Bayesian approach

3.3.1 Model description

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

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

For the prior of the standard deviations σ_{Subject} and σ 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_{Subject} and b we chose to use a normal distribution as prior.

3.3.2 Model comparison

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

```
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=TRU
)
```

```
## Running /Library/Frameworks/R.framework/Resources/bin/R CMD SHLIB foo.c
## clang -mmacosx-version-min=10.13 -I"/Library/Frameworks/R.framework/Resources/include" -DNDEBUG   -I
## In file included from <built-in>:1:
## In file included from /Library/Frameworks/R.framework/Versions/4.0/Resources/library/StanHeaders/incl
## In file included from /Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/inclu
## In file included from /Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/inclu
## /Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/include/Eigen/src/Core/util/
## namespace Eigen {
## ^
## /Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/include/Eigen/src/Core/util/
## namespace Eigen {
##                 ^
##                 ;
## In file included from <built-in>:1:
## In file included from /Library/Frameworks/R.framework/Versions/4.0/Resources/library/StanHeaders/incl
## In file included from /Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/inclu
## /Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/include/Eigen/Core:96:10: fa
## #include <complex>
##          ^~~~~~~~~
## 3 errors generated.
## make: *** [foo.o] Error 1
```

```
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=TRU
)
```

```
## Running /Library/Frameworks/R.framework/Resources/bin/R CMD SHLIB foo.c
## clang -mmacosx-version-min=10.13 -I"/Library/Frameworks/R.framework/Resources/include" -DNDEBUG   -I
## In file included from <built-in>:1:
## In file included from /Library/Frameworks/R.framework/Versions/4.0/Resources/library/StanHeaders/incl
## In file included from /Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/inclu
## In file included from /Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/inclu
## /Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/include/Eigen/src/Core/util/
## namespace Eigen {
## ^
## /Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/include/Eigen/src/Core/util/
## namespace Eigen {
##                 ^
##                 ;
## In file included from <built-in>:1:
## In file included from /Library/Frameworks/R.framework/Versions/4.0/Resources/library/StanHeaders/incl
## In file included from /Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/inclu
## /Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/include/Eigen/Core:96:10: fa
## #include <complex>
##          ^~~~~~~~~
## 3 errors generated.
## make: *** [foo.o] Error 1
```

```
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=TRU
)
```

```
## Running /Library/Frameworks/R.framework/Resources/bin/R CMD SHLIB foo.c
## clang -mmacosx-version-min=10.13 -I"/Library/Frameworks/R.framework/Resources/include" -DNDEBUG   -I
## In file included from <built-in>:1:
## In file included from /Library/Frameworks/R.framework/Versions/4.0/Resources/library/StanHeaders/incl
## In file included from /Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/inclu
## In file included from /Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/inclu
## /Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/include/Eigen/src/Core/util/
## namespace Eigen {
## ^
## /Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/include/Eigen/src/Core/util/
## namespace Eigen {
##                 ^
##                 ;
## In file included from <built-in>:1:
## In file included from /Library/Frameworks/R.framework/Versions/4.0/Resources/library/StanHeaders/incl
## In file included from /Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/inclu
## /Library/Frameworks/R.framework/Versions/4.0/Resources/library/RcppEigen/include/Eigen/Core:96:10: fa
## #include <complex>
##          ^~~~~~~~~
```

```
## 3 errors generated.
## make: *** [foo.o] Error 1
```

```
compare(m0,m1,m2)
```

```
##        WAIC       SE      dWAIC       dSE     pWAIC      weight
## m2 33055.53 87.73112    0.00000        NA 95.503738 0.998956252
## m1 33069.25 87.26925   13.72779  9.700142 94.055538 0.001043748
## m0 36527.12 98.01394 3471.59595 99.446576  2.510276 0.000000000
```

```
plot(compare(m0,m1,m2))
```

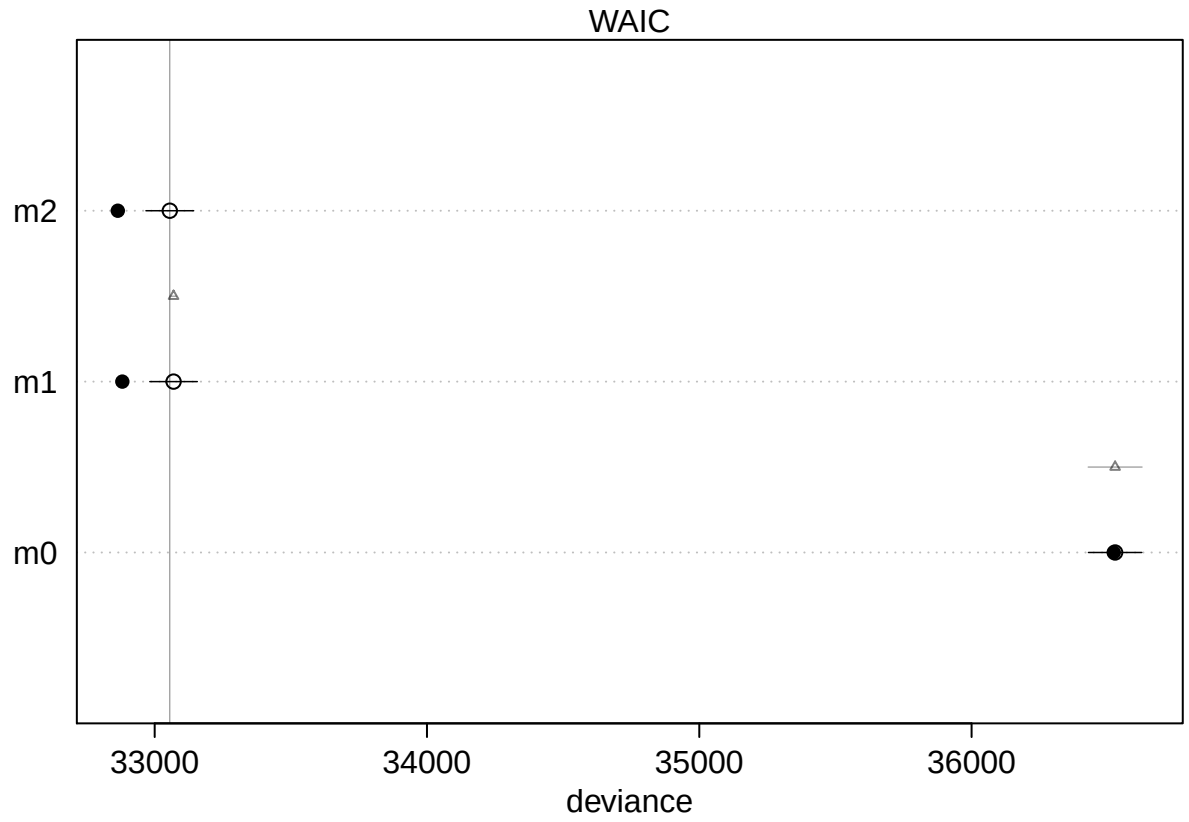


Figure 21: Comparing three Bayesian models of increasing complexity.

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 ($pWAIC$) increases for both these models.

3.3.3 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 σ_{Subject} and a is relatively high compared to that for parameters b and σ. Since the value for b is close to zero, it seems that session does not impact score.

```
plot(m2)
```

```
## 100 vector or matrix parameters hidden. Use depth=2 to show them.
```

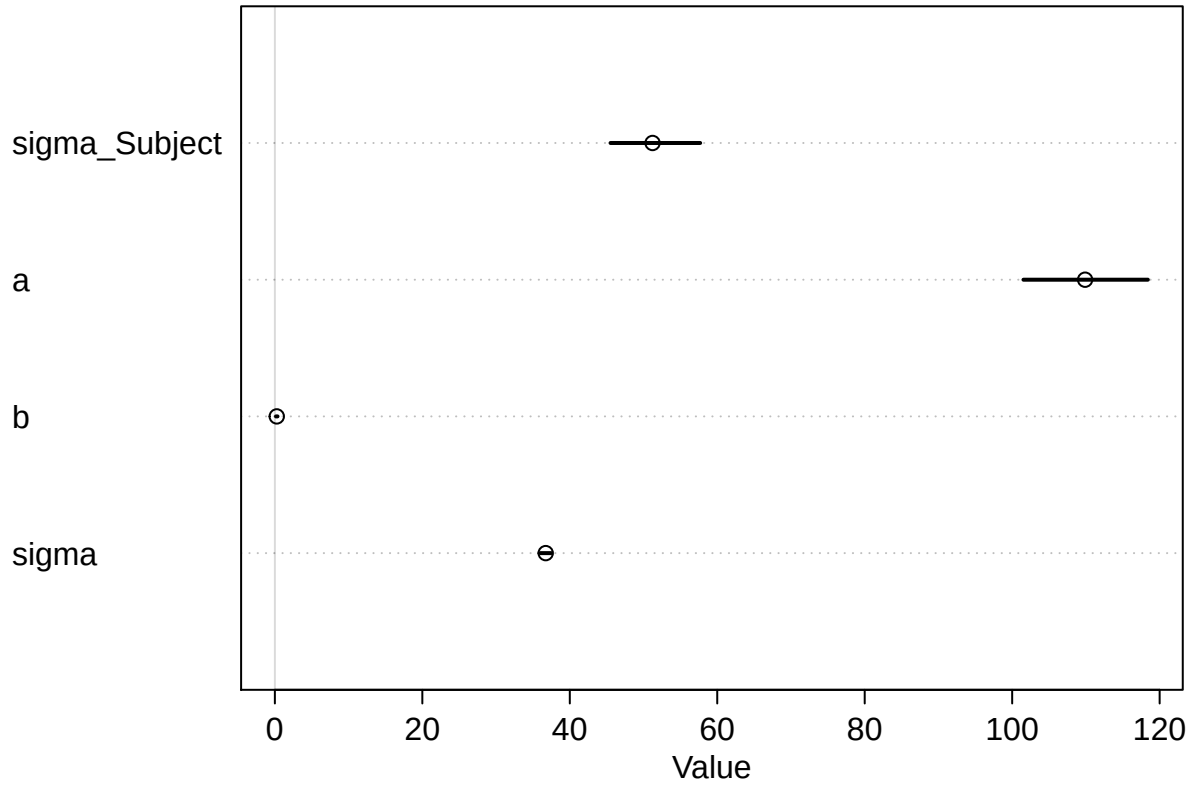


Figure 22: Parameters of m2.