Report coursework assignment A - 2021

CS4125 Seminar Research Methodology for Data Science

Nikki Bouman (4597648), Anuj Singh (), Gwennan Smitskamp ()

20/04/2021

Contents

1 Introduction

We will be using the following packages:

```
library(ggplot2) # plotting
library(AICcmodavg) # aictab
library(pander) #for rendering output
library(rethinking) # for stan
```

```
## Loading required package: rstan

## Loading required package: StanHeaders

## rstan (Version 2.21.2, GitRev: 2e1f913d3ca3)

## For execution on a local, multicore CPU with excess RAM we recommend calling
## options(mc.cores = parallel::detectCores()).
```

```
## To avoid recompilation of unchanged Stan programs, we recommend calling
## rstan_options(auto_write = TRUE)

## Do not specify '-march=native' in 'LOCAL_CPPFLAGS' or a Makevars file

## Loading required package: parallel

## Loading required package: dagitty

## rethinking (Version 2.01)

##
## Attaching package: 'rethinking'

## The following object is masked from 'package:AICcmodavg':
##
##     DIC

## The following object is masked from 'package:stats':
##
##     rstudent
```

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

2.1 The motivation for the planned research

(Max 250 words)

2.2 The theory underlying the research

(Max 250 words) Preferable based on theories reported in literature

2.3 Research questions

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

2.4 The related conceptual model

This model should include: *Independent variable(s)* Dependent variable *Mediating variable (at least 1)* Moderating variable (at least 1)

2.5 Experimental Design

Note that the study should have a true experimental design

2.6 Experimental procedure

Describe how the experiment will be executed step by step

2.7 Measures

Describe the measure that will be used

2.8 Participants

Describe which participants will recruit in the study and how they will be recruited

2.9 Suggested statistical analyses

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

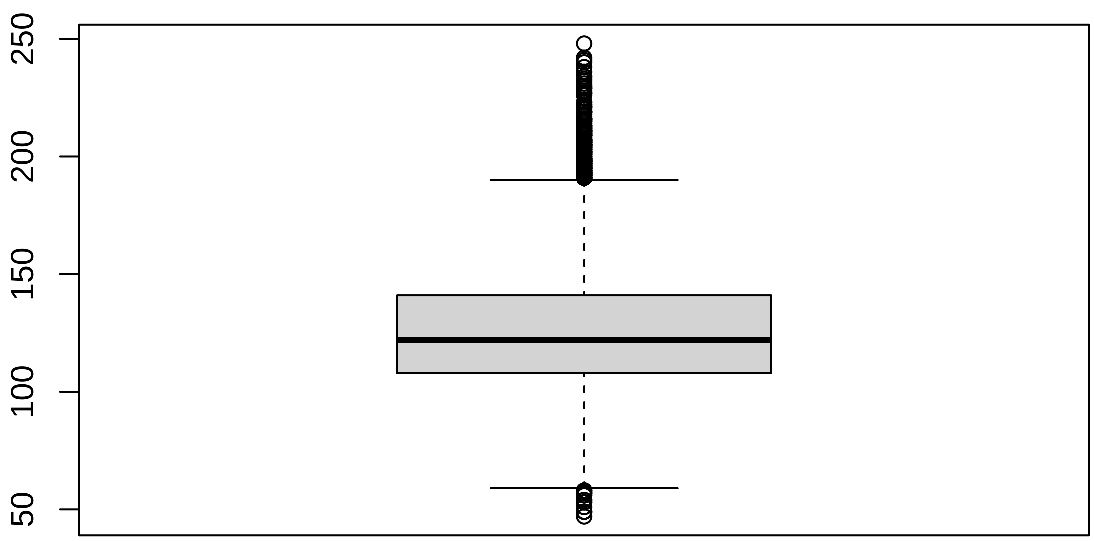
3 Part 3 - Multilevel model

3.1 Visual inspection

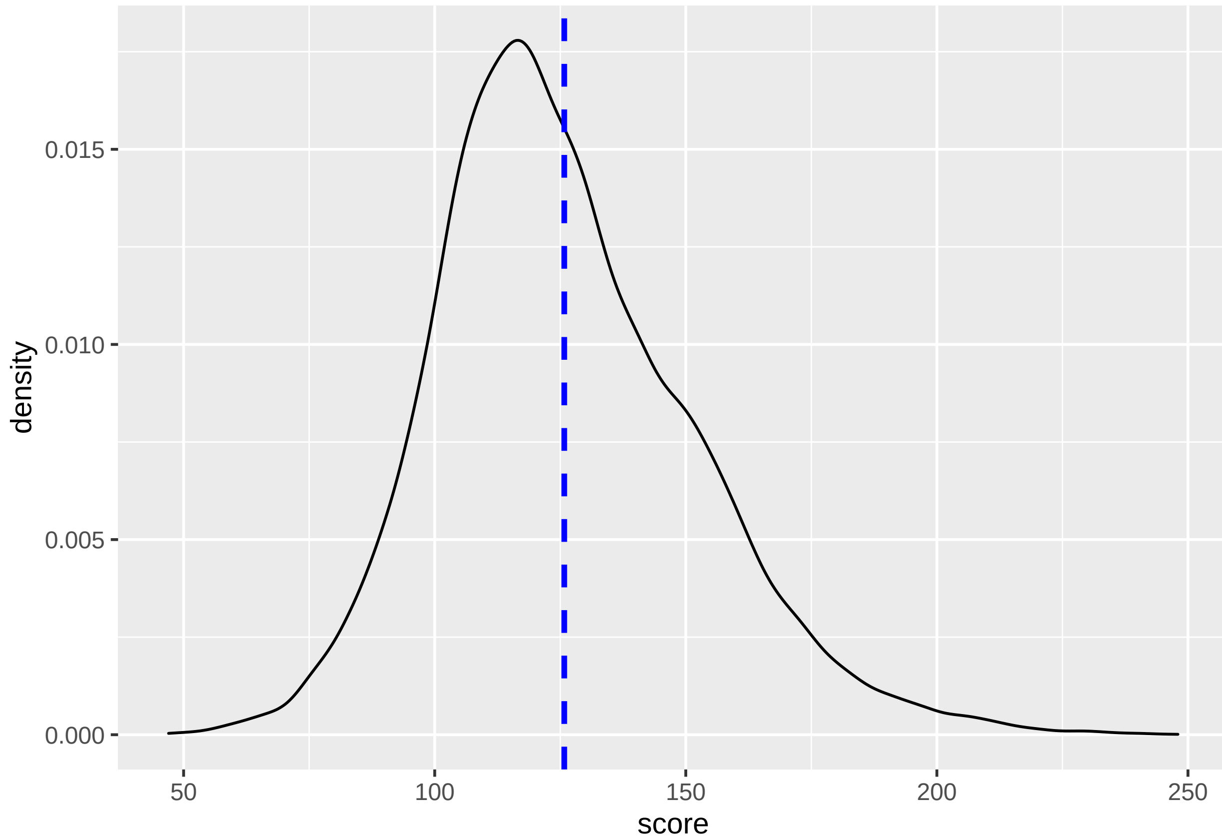
The boxplot and density plot show the distribution of the score. We can see that the mean score is 122 points. The minimum is set at 59, with outliers until 46, while the maximum is set at 190, with outliers until 248.

```r
# Get data
filepath <- ("set0.csv")
ds <- read.csv(file=filepath, header=TRUE)
ds <- data.frame(ds)

# boxplot score overall distribution (session independent)
boxplot(ds$score)
```



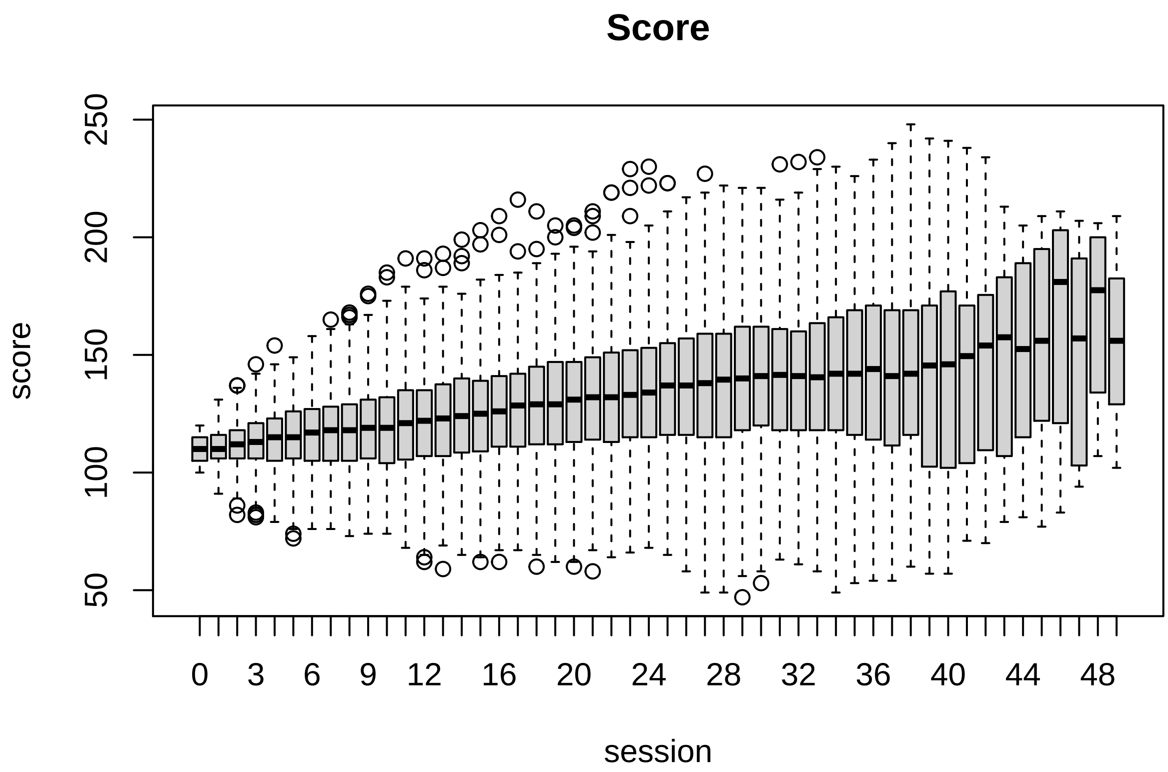

```r
# density score overall distribution (with mean line)
p <- ggplot(ds, aes(x=score)) + geom_density()
p + geom_vline(aes(xintercept=mean(score)), color="blue", linetype="dashed",size=1)
```



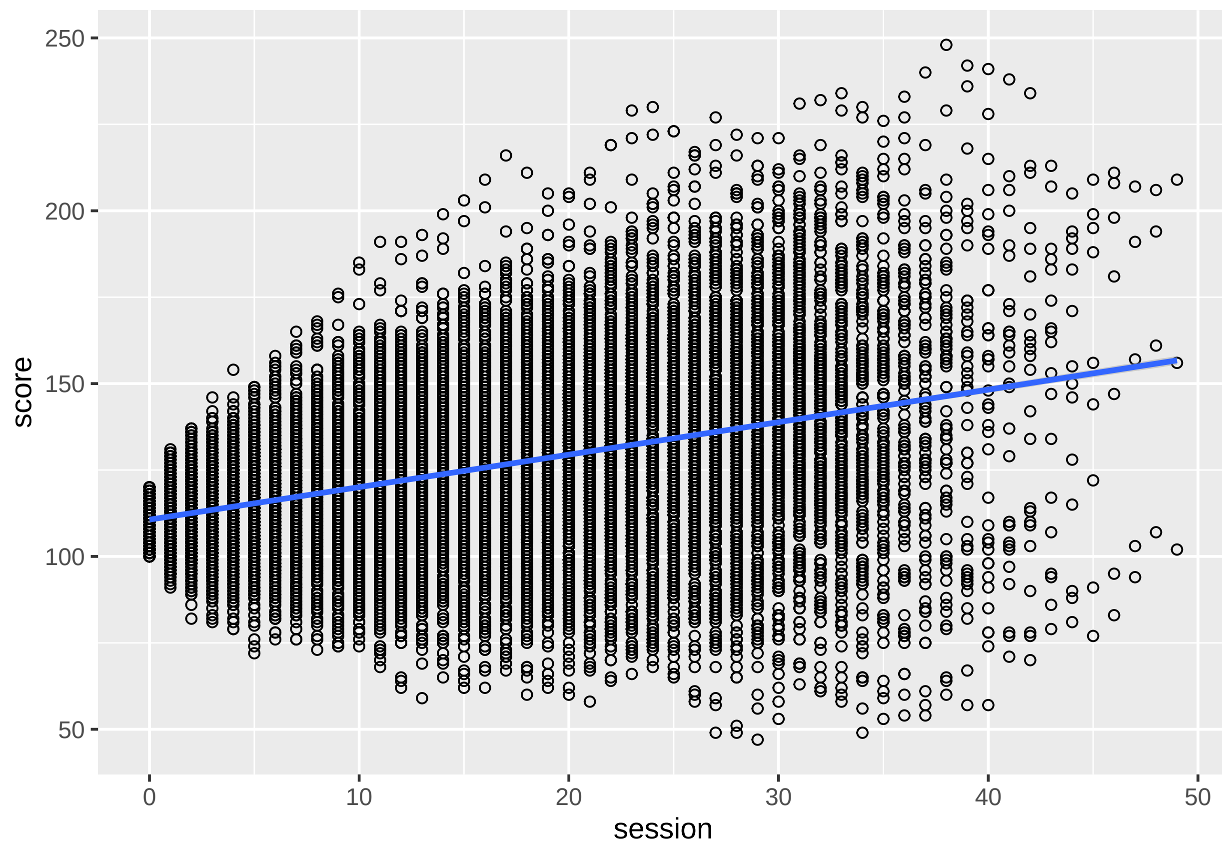
The relationship between the score and the session can be observed with the next two figures. The regression line (blue) in the scatterplot clearly shows how the score rises with the amount of sessions. This can also be observed in the box plot when looking at the mean (black line) for every box.

```
# set labels
ds$sessionF <- factor(ds$session, levels=c(0:49), labels=c(0:49))

# boxplot score per session
boxplot(score~sessionF, data=ds, main="Score", xlab="session", ylab="score")
```




```r
# ggplot score per session
hp <- ggplot(ds, aes(x=session, y=score)) + geom_point(shape=1) + geom_smooth(formula = y ~ x,method=lm
hp
```



3.2 Frequentist approach

3.2.1 Multilevel analysis

We have conducted a multilevel analysis. We have an intercept only model (model0) which we compare to a model that includes a predictor parameter for the session (model1). By comparing these two models, we will know whether there is a difference in the score over the sessions.

```
# create models as given in slides lecture 4
model0 <- lm(formula=score~1, data=ds, na.action=na.exclude)
model1 <- lm(formula=score~sessionF, data=ds, na.action=na.exclude)

# analysis, see if predictor improves fitting
pander(anova(model0,model1))
```

Table 1: Analysis of Variance Table

Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
16127	10713477	NA	NA	NA	NA
16078	9228641	49	1484836	52.79	0

```
pander(anova(model1))
```

Table 2: Analysis of Variance Table

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
sessionF	49	1484836	30303	52.79	0
Residuals	16078	9228641	574	NA	NA

From this analysis we can see there is a significant variation between the sessions. We take a further look at the summary results.

```
pander(summary(model1))
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	110.3	1.07	103	0
sessionF1	0.7166	1.514	0.4734	0.636
sessionF2	1.908	1.514	1.261	0.2075
sessionF3	2.96	1.514	1.955	0.05054
sessionF4	3.838	1.514	2.536	0.01123
sessionF5	5.004	1.514	3.306	0.0009494
sessionF6	5.972	1.514	3.945	8.005e-05
sessionF7	7.016	1.514	4.635	3.599e-06
sessionF8	7.637	1.514	5.045	4.585e-07
sessionF9	8.551	1.514	5.649	1.642e-08
sessionF10	9.097	1.515	6.004	1.969e-09
sessionF11	10.36	1.515	6.836	8.453e-12
sessionF12	11.22	1.515	7.402	1.41e-13
sessionF13	12.44	1.515	8.209	2.407e-16
sessionF14	13.6	1.515	8.974	3.161e-19
sessionF15	14.45	1.515	9.539	1.644e-21
sessionF16	15.63	1.515	10.31	7.277e-25
sessionF17	16.69	1.516	11.01	4.212e-28
sessionF18	18.07	1.518	11.9	1.62e-32
sessionF19	19.18	1.521	12.61	2.751e-36
sessionF20	19.8	1.521	13.02	1.539e-38
sessionF21	20.85	1.525	13.67	2.489e-42
sessionF22	21.35	1.529	13.96	5.131e-44
sessionF23	22.39	1.536	14.58	7.593e-48
sessionF24	23.32	1.542	15.12	2.555e-51
sessionF25	25.06	1.555	16.11	5.923e-58
sessionF26	26.11	1.574	16.59	2.818e-61
sessionF27	26.54	1.597	16.62	1.578e-61
sessionF28	27.37	1.64	16.69	5.045e-62
sessionF29	28.76	1.68	17.12	4.195e-65
sessionF30	29.52	1.73	17.07	9.765e-65
sessionF31	30.66	1.804	16.99	3.552e-64
sessionF32	30.02	1.908	15.73	2.406e-55
sessionF33	30.19	2.034	14.85	1.568e-49
sessionF34	30.08	2.213	13.59	7.575e-42
sessionF35	30.84	2.388	12.92	5.712e-38
sessionF36	31.17	2.571	12.12	1.116e-33
sessionF37	29.23	2.825	10.35	5.185e-25
sessionF38	32.42	3.076	10.54	6.942e-26
sessionF39	31.77	3.767	8.434	3.618e-17
sessionF40	32.78	4.031	8.131	4.553e-16

	Estimate	Std. Error	t value	Pr(>|t|)
sessionF41	32.06	4.503	7.119	1.13e-12
sessionF42	34.55	5.109	6.763	1.397e-11
sessionF43	37.34	5.748	6.496	8.474e-11
sessionF44	38.8	6.492	5.976	2.33e-09
sessionF45	43.17	8.057	5.358	8.536e-08
sessionF46	50.16	9.118	5.5	3.846e-08
sessionF47	40.13	10.77	3.727	0.0001948
sessionF48	56.73	12.03	4.717	2.417e-06
sessionF49	45.39	13.87	3.272	0.00107

Table 4: Fitting linear model: score ~ sessionF

Observations	Residual Std. Error	R^2	Adjusted R^2
16128	23.96	0.1386	0.136

The summary compares the first session (intercept) with the other sessions. Looking at the estimates, we can see that compared to the first session, the scores are higher every later session. Next, we will take a look at the Akaike Information Criterion (AIC) to compare the models on the goodness-of-fit concering the out-of-sample deviance.

```
#AIC
models <- list(model0, model1)
model.names <- c("model0", "model1")
pander(aictab(cand.set = models, modnames=model.names), caption="Model selection based on AICc.")
```

Table 5: Model selection based on AICc.

	Modnames	K	AICc	Delta_AICc	ModelLik	AICcWt	LL	Cum.Wt
2	model1	51	148277	0	1	1	-74087	1
1	model0	2	150584	2308	0	0	-75290	1

Here we can see that model 1 has the best goodness-of-fit as it has the smallest AICc value. Lastly, we will obtain a 95% confidence interval of the estimates we have obtained earlier.

```
# gives CI95%
pander(confint(model1), caption="95% confidence interval of the estimates.")
```

Table 6: 95% confidence interval of the estimates.

	2.5 %	97.5 %
(Intercept)	108.2	112.4
sessionF1	-2.251	3.684
sessionF2	-1.059	4.875
sessionF3	-0.007004	5.927
sessionF4	0.8712	6.805
sessionF5	2.037	7.971
sessionF6	3.005	8.939
sessionF7	4.049	9.983
sessionF8	4.67	10.6

	2.5 %	97.5 %
sessionF9	5.584	11.52
sessionF10	6.127	12.07
sessionF11	7.388	13.33
sessionF12	8.245	14.19
sessionF13	9.468	15.41
sessionF14	10.63	16.57
sessionF15	11.48	17.42
sessionF16	12.66	18.6
sessionF17	13.72	19.67
sessionF18	15.09	21.04
sessionF19	16.2	22.16
sessionF20	16.82	22.79
sessionF21	17.86	23.84
sessionF22	18.35	24.34
sessionF23	19.38	25.41
sessionF24	20.3	26.34
sessionF25	22.01	28.11
sessionF26	23.02	29.19
sessionF27	23.41	29.67
sessionF28	24.15	30.58
sessionF29	25.47	32.06
sessionF30	26.13	32.91
sessionF31	27.12	34.19
sessionF32	26.28	33.76
sessionF33	26.2	34.18
sessionF34	25.74	34.42
sessionF35	26.16	35.52
sessionF36	26.13	36.21
sessionF37	23.69	34.76
sessionF38	26.39	38.45
sessionF39	24.39	39.16
sessionF40	24.88	40.68
sessionF41	23.23	40.89
sessionF42	24.54	44.57
sessionF43	26.07	48.6
sessionF44	26.07	51.52
sessionF45	27.38	58.96
sessionF46	32.28	68.03
sessionF47	19.02	61.23
sessionF48	33.15	80.3
sessionF49	18.2	72.59

Here we can again see an increase in score related to the sessions. From this we can conclude that the session has a positive effect on people's score. Also, it seems there is a significant variance between the participants in their score.

3.2.2 Report section for a scientific publication

A Linear Model analysis was conducted to test the difference between sessions on the score. The results found a significant effect ($F(49,16078) = 52.793$, $p < .001$) for the sessions on the score. We can conclude that over sessions the score per participant generally increases.

3.3 Bayesian approach

3.3.1 Model description

For model 2, the model with session as a factor, we take as prior a normal distribution of N(125,30). This comes from the mean of the score, 125, and a bit more than the standard deviation, which is around 27. Our sigma is set at a uniform distribution of U(0.001,30).

3.3.2 Model comparison

We will create and compare the three described models. From the results we can see that m1 has the best fit since it has the smallest WAIC value

```
ds <- ds[!(ds$Subject>99),] # select first 100 subjects
ds$Subject <- ds$Subject +1 # increase subject number with 1 to overcome Stan zero index problem

mean(ds$score) # check mean
```

```
## [1] 125.5142
```

```
sd(ds$score) # check standard deviation
```

```
## [1] 27.402
```

```
ds$sessionF <- factor(ds$session, levels=c(0:49), labels=c(0:49))
ds$subjF <- factor(ds$Subject, levels=c(1:100), labels=c(1:100))

da <- subset(ds, select=c(score, sessionF))
da1 <- subset(ds, select=c(score, subjF))

# create model with fixed intercept (i)
m0 <- map2stan(
  alist(
    score ~ dnorm(mu, sigma),
    mu <- a,
    a ~dnorm(125,30), # mean and sd from what we found above
    sigma ~dunif(0.001,30)
  ), data = da, iter = 10000, chains = 4, cores = 4
)
```

```
## Computing WAIC
```

```
# create model extended with an adaptive prior for subject id (ii)
m1 <- map2stan(
  alist(
    score ~ dnorm(mu, sigma),
    mu <- a[subjF],
    a[subjF] ~ dnorm(125,sigma_subj),
    sigma_subj ~ dcauchy(0,10),
    sigma ~dunif(0.001,30)
  ), data = da1, iter = 10000, chains = 4, cores = 4
)
```

```
## Computing WAIC
```

```
# create model with session as a factor (iii)
m2 <- map2stan(
  alist(
    score ~ dnorm(mu, sigma),
```

```
    mu <- a[sessionF],
    a[sessionF] ~ dnorm(125,30),
    sigma ~dunif(0.001,30)
  ), data = da, iter = 10000, chains = 4, cores = 4
)
```

```
## Computing WAIC
```

```
pander(compare(m0,m1,m2,func=WAIC))
```

	WAIC	SE	dWAIC	dSE	pWAIC	weight
m1	28322	96.41	0	NA	93.31	1
m2	30578	98.57	2255	120.9	67.33	0
m0	31039	98.23	2716	104.4	2.485	0

3.3.3 Estimates examination

From the previous comparison we could see that m1 is the best fit model. We will further examine this model with 95% credible intervals of the parameters of this model.

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

	mean	sd	2.5%	97.5%	n_eff	Rhat4
a[1]	118.3	5.452	107.6	128.9	46863	0.9998
a[2]	105.4	3.037	99.6	111.4	49275	0.9998
a[3]	169.4	3.169	163.2	175.6	47051	0.9999
a[4]	139.1	2.977	133.3	144.9	46109	0.9998
a[5]	119	3.053	113	125	49912	0.9999
a[6]	114.3	3.199	108.1	120.5	48861	0.9999
a[7]	120.4	3.152	114.3	126.5	46010	0.9998
a[8]	134.4	3.104	128.2	140.4	50871	0.9999
a[9]	122.4	4.137	114.4	130.5	49844	0.9999
a[10]	145.6	3.207	139.4	151.8	47889	0.9998
a[11]	138.2	2.668	132.9	143.4	46579	0.9999
a[12]	105.3	3.157	99.04	111.5	52917	0.9998
a[13]	134.5	2.917	128.8	140.2	48173	0.9999
a[14]	135.2	2.704	129.9	140.5	45895	0.9998
a[15]	99.3	2.85	93.75	104.9	45343	0.9999
a[16]	151.5	2.988	145.6	157.3	45624	0.9999
a[17]	127.4	2.817	121.9	132.9	48267	0.9999
a[18]	82.56	3.322	76.08	89.06	50601	0.9999
a[19]	123.8	3.321	117.3	130.3	48748	0.9998
a[20]	122.5	3.32	116	129	47095	0.9998
a[21]	111.7	2.957	105.9	117.5	46187	0.9999
a[22]	135.8	3.901	128.2	143.5	49686	0.9999
a[23]	118.4	2.93	112.6	124.1	46719	0.9999
a[24]	118.7	3.675	111.5	125.9	49665	0.9998
a[25]	109.4	3.29	103	115.9	45479	0.9998
a[26]	117.4	2.722	112	122.6	48505	0.9999
a[27]	127.2	2.853	121.6	132.8	47420	0.9999
a[28]	132.8	3.158	126.7	138.9	51363	0.9998
a[29]	94.88	3.098	88.78	100.9	44499	0.9998
a[30]	148.7	3.157	142.5	155	45543	0.9999

	mean	sd	2.5%	97.5%	n_eff	Rhat4
a[31]	148.4	3.445	141.7	155.1	46831	0.9998
a[32]	81.77	3.086	75.71	87.78	45759	0.9999
a[33]	159.2	2.858	153.6	164.8	45197	0.9998
a[34]	95.96	3.225	89.68	102.2	48347	0.9999
a[35]	129.9	3.389	123.2	136.6	52065	0.9998
a[36]	92.8	2.725	87.52	98.14	46483	0.9999
a[37]	120.2	3.138	114	126.3	47614	0.9999
a[38]	145.3	2.808	139.8	150.7	48202	0.9999
a[39]	112.4	3.23	106.1	118.8	50442	0.9998
a[40]	114.1	3.361	107.6	120.8	45820	0.9999
a[41]	100.4	3.296	93.97	106.9	53785	0.9999
a[42]	139.1	3.255	132.7	145.4	45588	0.9999
a[43]	122.3	3.207	116	128.6	49526	0.9998
a[44]	95.53	3.233	89.18	101.8	45333	0.9999
a[45]	108	3	102.1	113.8	47190	0.9998
a[46]	126.4	2.816	120.9	131.9	49853	0.9999
a[47]	134.1	3.023	128.3	139.9	49155	0.9998
a[48]	114.3	3.059	108.3	120.3	48282	0.9998
a[49]	163.7	2.927	157.9	169.4	44999	0.9999
a[50]	144.3	2.935	138.5	150.1	47865	0.9999
a[51]	126.2	3.11	120.1	132.3	47868	0.9999
a[52]	116.9	2.811	111.4	122.4	48539	0.9998
a[53]	142	3.21	135.7	148.4	51044	0.9999
a[54]	132.9	3.091	126.8	139	45872	0.9999
a[55]	111.4	3.459	104.7	118.2	48772	0.9999
a[56]	95.98	3.562	89.09	103	45293	0.9999
a[57]	135.5	3.12	129.3	141.6	42763	1
a[58]	139.1	3.26	132.8	145.6	47416	0.9999
a[59]	100	3.705	92.79	107.3	48328	0.9999
a[60]	146.3	2.901	140.5	152	51934	0.9999
a[61]	107	2.924	101.4	112.8	48097	1
a[62]	197.8	2.758	192.3	203.2	46559	0.9999
a[63]	90.87	3.189	84.53	97.1	47514	1
a[64]	105.3	3.311	98.9	111.9	45335	0.9998
a[65]	97.18	2.915	91.5	102.9	46678	0.9999
a[66]	104.6	2.844	98.96	110.1	42626	0.9999
a[67]	142.1	3.321	135.6	148.6	50650	0.9999
a[68]	109.3	2.944	103.5	115.1	50291	0.9999
a[69]	138.3	3.173	132.1	144.5	46584	0.9999
a[70]	158.6	3.009	152.7	164.4	47839	0.9998
a[71]	97.19	3.15	91.06	103.4	47457	0.9999
a[72]	108.5	3.031	102.5	114.4	46872	1
a[73]	130.9	3.343	124.4	137.4	48174	0.9998
a[74]	138.3	2.977	132.5	144.2	49628	0.9999
a[75]	107.1	2.669	101.8	112.2	48569	0.9999
a[76]	137.8	3.003	131.9	143.8	50039	0.9999
a[77]	154.6	2.829	149.1	160.2	49248	0.9999
a[78]	119.4	3.058	113.5	125.5	46443	0.9999
a[79]	109	3.247	102.6	115.4	47524	0.9999
a[80]	121.4	2.908	115.7	127.1	49514	0.9999
a[81]	131.8	3.042	125.9	137.8	48432	0.9999
a[82]	115.2	2.979	109.4	121.1	50496	0.9999

	mean	sd	2.5%	97.5%	n_eff	Rhat4
a[83]	121	3.302	114.5	127.5	44464	0.9999
a[84]	91.57	3.049	85.62	97.55	49403	0.9999
a[85]	117.1	3.123	111	123.2	51796	0.9999
a[86]	145.4	2.726	140	150.7	52371	0.9998
a[87]	132.4	3.191	126.2	138.8	47567	0.9999
a[88]	146.2	3.403	139.5	152.8	46597	0.9998
a[89]	123.7	2.818	118.1	129.2	47402	0.9999
a[90]	144.1	3.182	138	150.3	44826	0.9999
a[91]	124.2	3.074	118.1	130.3	47071	0.9999
a[92]	140.9	3.207	134.6	147.1	47419	0.9999
a[93]	128.4	3.017	122.4	134.3	48548	0.9999
a[94]	128	2.905	122.2	133.7	47732	0.9999
a[95]	116.6	3.654	109.4	123.6	52927	0.9998
a[96]	126.1	3.329	119.5	132.6	49726	0.9999
a[97]	154.2	3.35	147.7	160.8	44233	0.9998
a[98]	149.2	2.846	143.6	154.8	42799	0.9999
a[99]	125.2	2.927	119.5	130.9	49088	0.9999
a[100]	139.6	2.874	134	145.2	47560	0.9999
sigma_subj	20.4	1.473	17.75	23.52	41017	0.9999
sigma	17.86	0.2228	17.44	18.31	43378	0.9998

Looking at the mean we can see that in general the scores are increasing over the sessions. Moreover, the standard deviating is increasing as well. This could be due to the fact that not every subject has completed the same amount of sessions.