

# Report coursework assignment A - 2021

## CS4125 Seminar Research Methodology for Data Science

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

20/04/2021

## Contents

<b>1</b>	<b>Part 1 - Design and set-up of true experiment</b>	<b>1</b>
1.1	The motivation for the planned research . . . . .	1
1.2	The theory underlying the research . . . . .	1
1.3	Research questions . . . . .	1
1.4	The related conceptual model . . . . .	1
1.5	Experimental Design . . . . .	2
1.6	Experimental procedure . . . . .	2
1.7	Measures . . . . .	2
1.8	Participants . . . . .	2
1.9	Suggested statistical analyses . . . . .	2
<b>2</b>	<b>Part 3 - Multilevel model</b>	<b>2</b>
2.1	Visual inspection . . . . .	2
2.2	Frequentist approach . . . . .	5
2.2.1	Multilevel analysis . . . . .	5
2.2.2	Report section for a scientific publication . . . . .	8

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

### 1.1 The motivation for the planned research

(Max 250 words)

### 1.2 The theory underlying the research

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

### 1.3 Research questions

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

### 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)*

## 1.5 Experimental Design

Note that the study should have a true experimental design

## 1.6 Experimental procedure

Describe how the experiment will be executed step by step

## 1.7 Measures

Describe the measure that will be used

## 1.8 Participants

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

## 1.9 Suggested statistical analyses

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

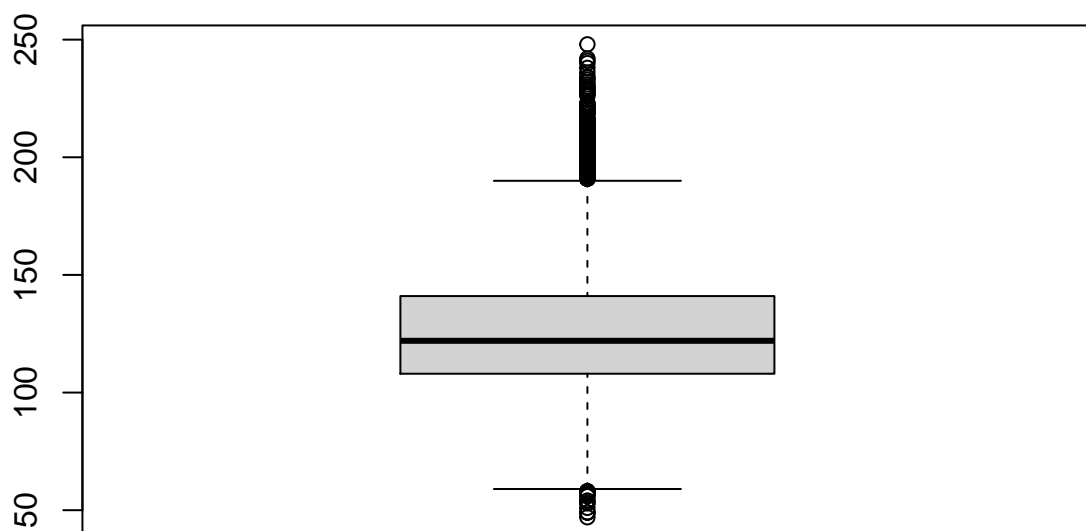
# 2 Part 3 - Multilevel model

## 2.1 Visual inspection

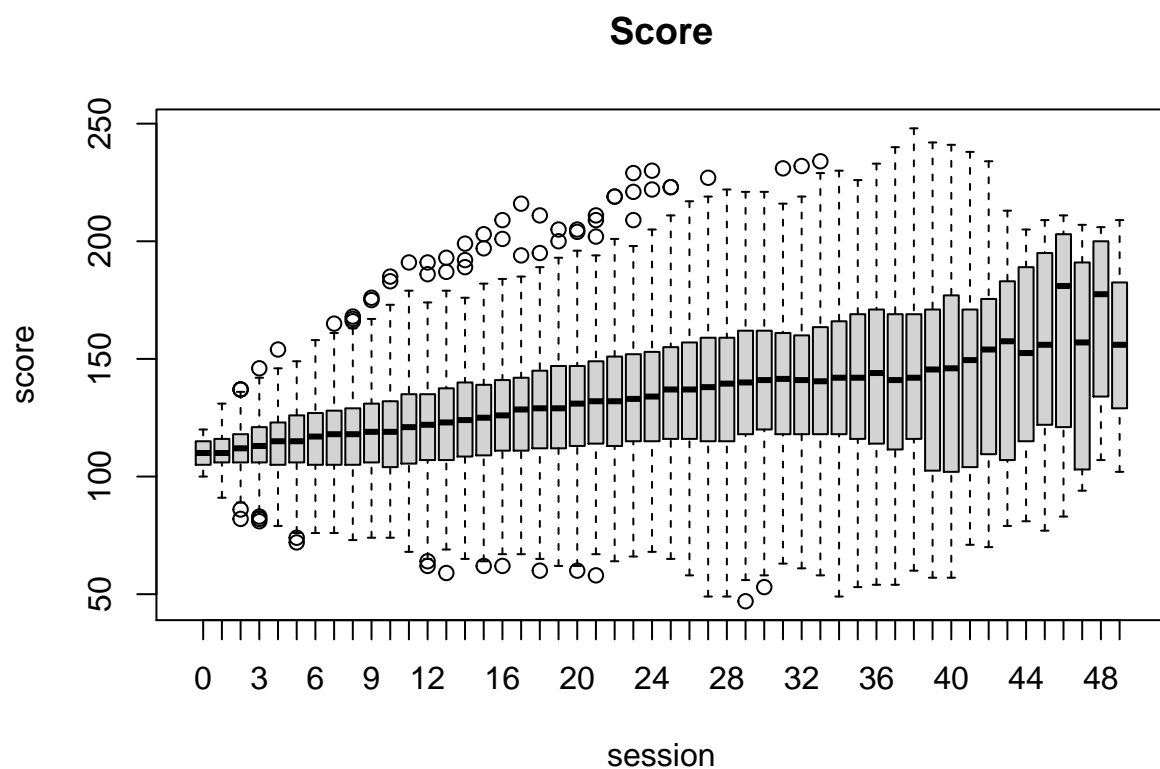
```
library(ggplot2)

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

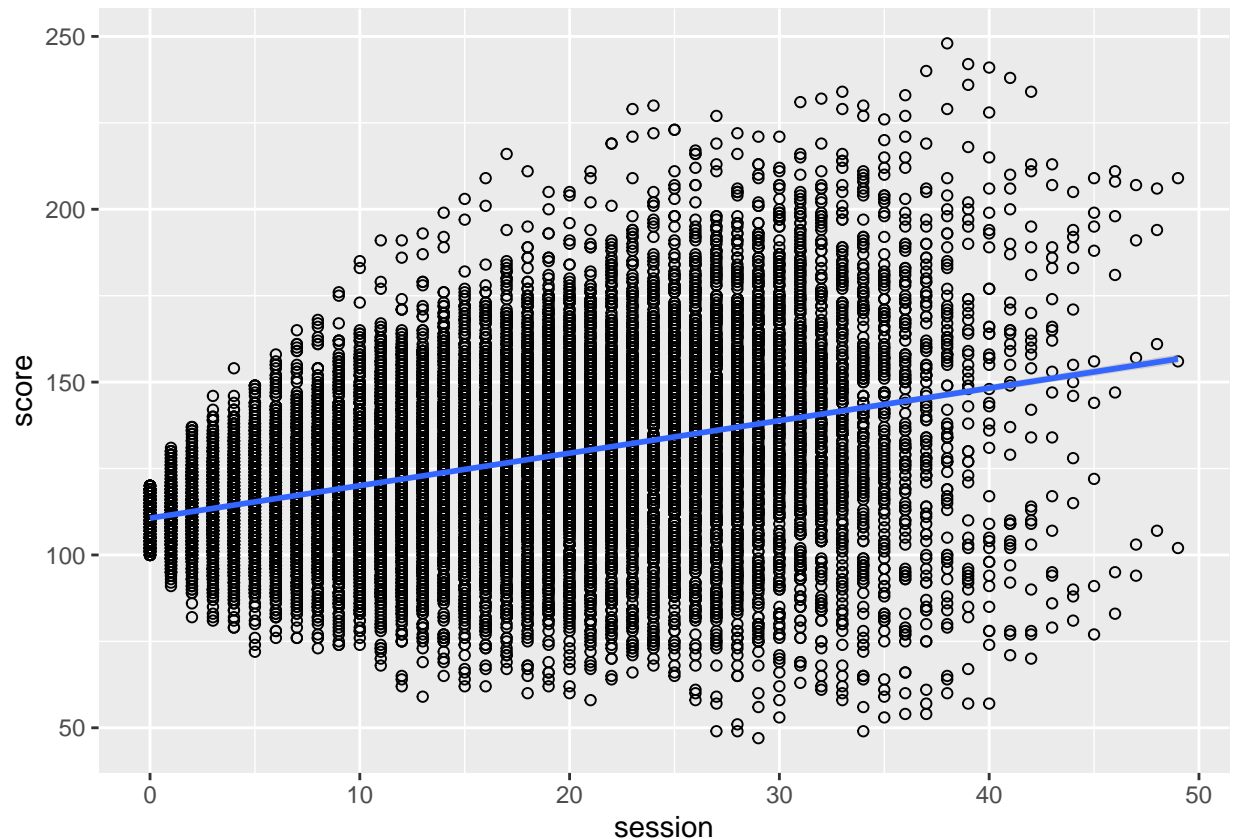
# boxplot score overall distribution (session independent)
boxplot(ds$score)
```



```
# 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")
```



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



## 2.2 Frequentist approach

### 2.2.1 Multilevel analysis

We have conducted a multilevel analysis and calculated the 95% confidence intervals. The results show that the session has impact on the scores: the more sessions, the better the score. It can be observed that this increase in score stagnates around 30 sessions.

There is (no?) significant variance between the participants in their score. TODO

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

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

# 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
anova(model0,model1)

## Analysis of Variance Table
##
## Model 1: score ~ 1
```

```
## Model 2: score ~ sessionF
##   Res.Df      RSS Df Sum of Sq      F      Pr(>F)
## 1  16127 10713477
## 2  16078  9228641 49    1484836 52.793 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

summary(model1)

##
## Call:
## lm(formula = score ~ sessionF, data = ds, na.action = na.exclude)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -92.038 -13.872   0.289  14.176 105.304
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 110.2735     1.0704 103.024 < 2e-16 ***
## sessionF1     0.7166     1.5137   0.473 0.635950
## sessionF2     1.9082     1.5137   1.261 0.207478
## sessionF3     2.9601     1.5137   1.955 0.050543 .
## sessionF4     3.8383     1.5137   2.536 0.011233 *
## sessionF5     5.0040     1.5137   3.306 0.000949 ***
## sessionF6     5.9721     1.5137   3.945 8.01e-05 ***
## sessionF7     7.0160     1.5137   4.635 3.60e-06 ***
## sessionF8     7.6367     1.5137   5.045 4.59e-07 ***
## sessionF9     8.5509     1.5137   5.649 1.64e-08 ***
## sessionF10    9.0973     1.5152   6.004 1.97e-09 ***
## sessionF11   10.3578     1.5152   6.836 8.45e-12 ***
## sessionF12   11.2155     1.5152   7.402 1.41e-13 ***
## sessionF13   12.4380     1.5152   8.209 2.41e-16 ***
## sessionF14   13.5983     1.5152   8.974 < 2e-16 ***
## sessionF15   14.4540     1.5152   9.539 < 2e-16 ***
## sessionF16   15.6284     1.5152  10.314 < 2e-16 ***
## sessionF17   16.6944     1.5160  11.012 < 2e-16 ***
## sessionF18   18.0680     1.5183  11.900 < 2e-16 ***
## sessionF19   19.1757     1.5206  12.610 < 2e-16 ***
## sessionF20   19.8039     1.5214  13.017 < 2e-16 ***
## sessionF21   20.8477     1.5246  13.674 < 2e-16 ***
## sessionF22   21.3461     1.5294  13.957 < 2e-16 ***
## sessionF23   22.3946     1.5360  14.580 < 2e-16 ***
## sessionF24   23.3188     1.5419  15.124 < 2e-16 ***
## sessionF25   25.0569     1.5551  16.113 < 2e-16 ***
## sessionF26   26.1071     1.5740  16.586 < 2e-16 ***
## sessionF27   26.5383     1.5966  16.622 < 2e-16 ***
## sessionF28   27.3690     1.6397  16.691 < 2e-16 ***
## sessionF29   28.7646     1.6805  17.117 < 2e-16 ***
## sessionF30   29.5175     1.7295  17.067 < 2e-16 ***
## sessionF31   30.6567     1.8044  16.990 < 2e-16 ***
## sessionF32   30.0179     1.9082  15.731 < 2e-16 ***
## sessionF33   30.1901     2.0335  14.846 < 2e-16 ***
## sessionF34   30.0795     2.2130  13.592 < 2e-16 ***
## sessionF35   30.8377     2.3877  12.915 < 2e-16 ***
```

```
## sessionF36 31.1742 2.5714 12.123 < 2e-16 ***
## sessionF37 29.2265 2.8247 10.347 < 2e-16 ***
## sessionF38 32.4222 3.0764 10.539 < 2e-16 ***
## sessionF39 31.7720 3.7671 8.434 < 2e-16 ***
## sessionF40 32.7792 4.0312 8.131 4.55e-16 ***
## sessionF41 32.0599 4.5032 7.119 1.13e-12 ***
## sessionF42 34.5526 5.1090 6.763 1.40e-11 ***
## sessionF43 37.3377 5.7475 6.496 8.47e-11 ***
## sessionF44 38.7980 6.4919 5.976 2.33e-09 ***
## sessionF45 43.1710 8.0575 5.358 8.54e-08 ***
## sessionF46 50.1551 9.1184 5.500 3.85e-08 ***
## sessionF47 40.1265 10.7677 3.727 0.000195 ***
## sessionF48 56.7265 12.0268 4.717 2.42e-06 ***
## sessionF49 45.3932 13.8736 3.272 0.001070 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 23.96 on 16078 degrees of freedom
## Multiple R-squared:  0.1386, Adjusted R-squared:  0.136
## F-statistic: 52.79 on 49 and 16078 DF, p-value: < 2.2e-16
```

```
# examine estimators
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: score
##           Df Sum Sq Mean Sq F value    Pr(>F)
## sessionF   49 1484836   30303  52.793 < 2.2e-16 ***
## Residuals 16078 9228641     574
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# gives CI95%
confint(model1)
```

```
##           2.5 %      97.5 %
## (Intercept) 108.175408166 112.371498
## sessionF1   -2.250516726   3.683650
## sessionF2   -1.058899959   4.875267
## sessionF3   -0.007003752   5.927163
## sessionF4    0.871239761   6.805407
## sessionF5    2.036908424   7.971076
## sessionF6    3.004972296   8.939139
## sessionF7    4.048884472   9.983052
## sessionF8    4.669642955  10.603810
## sessionF9    5.583814612  11.517982
## sessionF10   6.127233256  12.067344
## sessionF11   7.387754298  13.327865
## sessionF12   8.245469729  14.185580
## sessionF13   9.467914618  15.408025
## sessionF14  10.628235260  16.568346
## sessionF15  11.483946682  17.424057
## sessionF16  12.658295380  18.598406
## sessionF17  13.722869661  19.665967
## sessionF18  15.091899875  21.044022
```

## sessionF19	16.195112212	22.156356
## sessionF20	16.821787305	22.786093
## sessionF21	17.859365283	23.836028
## sessionF22	18.348321730	24.343857
## sessionF23	19.383949293	25.405297
## sessionF24	20.296539040	26.341104
## sessionF25	22.008713118	28.105135
## sessionF26	23.021851788	29.192263
## sessionF27	23.408790988	29.667775
## sessionF28	24.154986252	30.583054
## sessionF29	25.470619859	32.058497
## sessionF30	26.127442948	32.907644
## sessionF31	27.119815970	34.193572
## sessionF32	26.277524407	33.758178
## sessionF33	26.204148297	34.176029
## sessionF34	25.741803760	34.417172
## sessionF35	26.157470199	35.517846
## sessionF36	26.133865257	36.214467
## sessionF37	23.689820367	34.763273
## sessionF38	26.392055312	38.452343
## sessionF39	24.388086311	39.155917
## sessionF40	24.877536137	40.680821
## sessionF41	23.233118228	40.886642
## sessionF42	24.538428125	44.566840
## sessionF43	26.071853518	48.603463
## sessionF44	26.073075399	51.522876
## sessionF45	27.377482890	58.964500
## sessionF46	32.282111551	68.028125
## sessionF47	19.020590328	61.232503
## sessionF48	33.152698987	80.300395
## sessionF49	18.199443267	72.586984

### 2.2.2 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. TODO