Homework 1

Denis Ostroushko

Question 1

a)

Outcomes for 20th subject are shown in Table 1

Table 1: Outcomes for 20th subject in the study

Y20
0.101
0.074
0.059
0.019
0.082

b)

In the long format of the data, where we have 107 unique participants, and 5 observations per Id, the full length of responses $\mathbf{Y}$ is 535

c)

We are given that for a random variable that generate observed values for the 20th participant, $Var(Y_{20\ j}) = 0.05$, so variance is constant for each observation $j = 1, 2, 3, 4, 5$.

Covariance of i^{th} and j^{th} observations is also constant at 0.2.

Covariance of i^{th} observation with itself is variance, an is 0.05.

Table 2 displays resulting variance-covariance matrix.

Table 2: Variance-covariance matrix for 20th participant

0.05	0.02	0.02	0.02	0.02
0.02	0.05	0.02	0.02	0.02
0.02	0.02	0.05	0.02	0.02
0.02	0.02	0.02	0.05	0.02
0.02	0.02	0.02	0.02	0.05

We can now convert this matrix into the correlation matrix.

$\rho_{ij} = \frac{Cov(Y_{ij}, Y_{jk}}{\sqrt{Var(Y_{ij}) * Var(Y_{jk})}}$

Since all variances are equal to 0.05, the calculation is pretty straightforward.

Resulting correlation matrix is shown in Table 3

Table 3: Correlation matrix for 20th individual

1.0	0.4	0.4	0.4	0.4
0.4	1.0	0.4	0.4	0.4
0.4	0.4	1.0	0.4	0.4
0.4	0.4	0.4	1.0	0.4
0.4	0.4	0.4	0.4	1.0

Each observation on the diagonal is equal to 1, which makes the results more credible

d)

A variance-covariance matrix for each individual is 5×5 in size, and we have 107 individuals, therefore the size of full matrix is $(5 \times 107) \times (5 \times 107)$, which means that the dimension of variance covariance matrix Σ is 535×535.

e)

We consider trial number, age, and sex of a participant as predictors. Values provided in the dataset are given in the Table 4.

Table 4: Covariates for 20th participant

Trial Number	Age	Sex
1	31	0
2	31	0
3	31	0
4	31	0
5	31	0

However, we also need a column of 1s in order to estimate $\hat{\beta}_0$. A full matrix $\mathbf{X}_{20}$ for the 20th participant is given in Table 5

Table 5: Full matrix for model estimation

1	1	31	0
1	2	31	0
1	3	31	0
1	4	31	0
1	5	31	0

f)

Including intercept column, the full dimension of $\mathbf{X}$ is 535×4. Without intercept, the size of a matrix with three predictors is 535×3.

Question 2

a)

Figure 1 shows the distribution of scores in each trial for all participants. While each trial has extreme values, ranging from 0.5 to 0.6, consistency of these values across the five trials suggests that these values are not necessarily outliers. However, reaction times in trial number two close to 0.8 should be considered outliers.

Using boxplots we assessed the shape of distribution and how it looks over time. It appears that for a hard task like showing reaction to quick events, short term results do not seem to vary much.

Figure 1: Distribution of Reaction Measurements in Each Trial

b)

Figure 2 shows that the average trend *seems* to be downward, but it is rather weak. There is not much visual evidence that there are any individual trajectories that can be considered outliers. Individuals tend to have higher measurements in a given trial, however, it appears to be 'random' noise or variations.

c)

Figure 3 shows that, on average, adults have lower response time in each trial, however, by trial number 4 and 5, the difference seems to be smaller.

Question 3

a)

Figure 4 shows that sample correlation between time points, or trials, seems to be pretty equal, with some random sampling variation. There are a few values that are a little more different than most values in the sample, however, this visual pattern suggests that the correlation structure is Exchangeable.

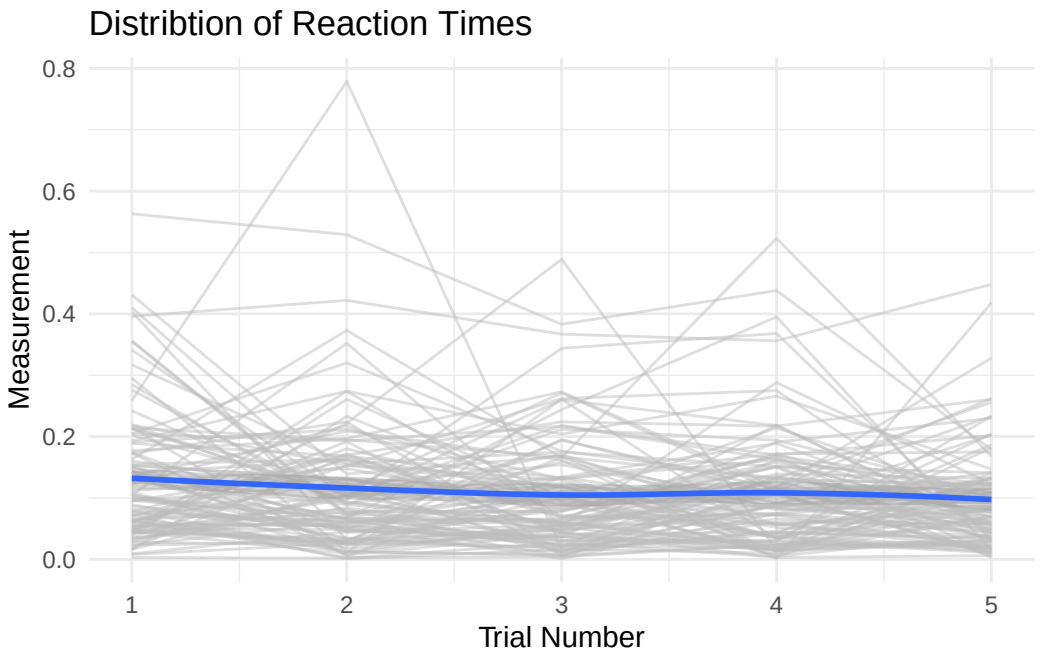

Figure 2: Individual Profiles of Measurements for each participant

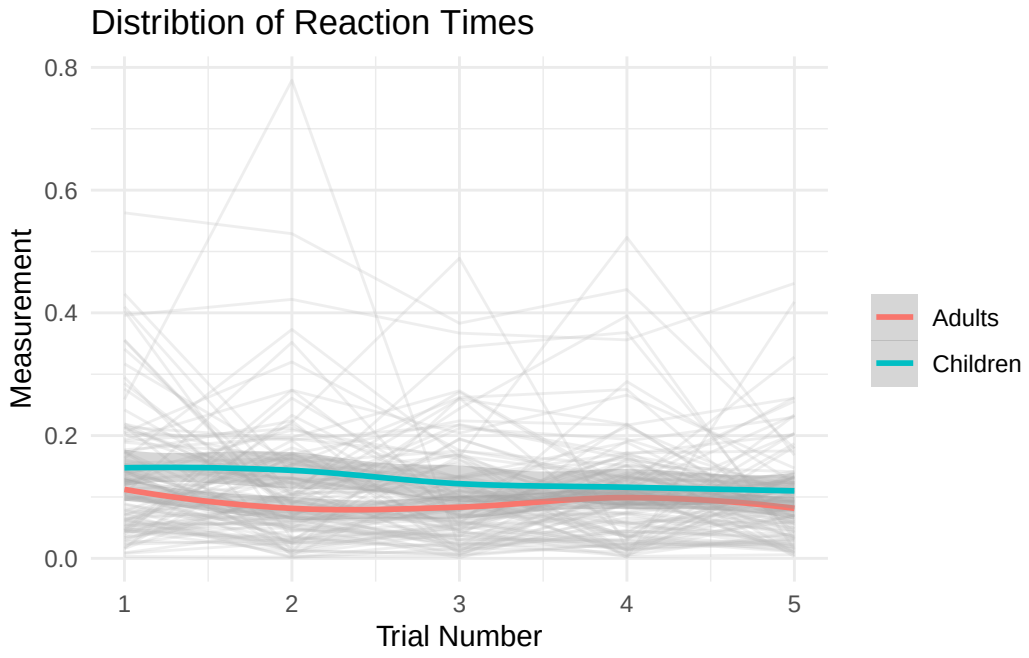

Figure 3: Individual Profiles of Measurements for each participant

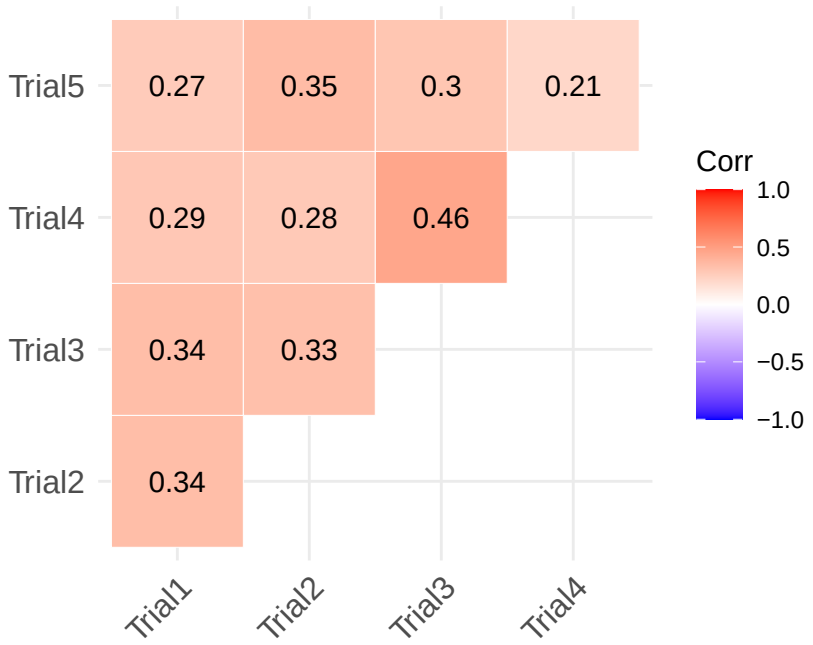

Figure 4: **?(caption)**

b)

Of displayed values in the sample correlation matrix on Figure 4, the average value is 0.32. From a set of 10 correlation values, we obtain a 95% confidence interval for the mean value, which is (0.28, 0.36).

As most values fall within this range, there is little evidence that the values are auto-regressive, or M-independent.

Question 4

a)

If there is a learning effect, we want to test if the average response time is lower in the fifth trial compared to the first trial. Therefore, we want to use a one sided t-test for the difference in means.

Table 6: Average response time in each trial

Trial	Average Response	Estiamte SE
Trial1	0.13	0.01
Trial2	0.12	0.01

- Null Hypothesis: $Mean_{Trial\ 5} = Mean_{Trial\ 1}$
- Alternative Hypothesis: $Mean_{Trial\ 5} < Mean_{Trial\ 1}$
- Difference in means: 0.03, on average, the measurements in Trial 5 decreased
- Test statistic and p-value: -2.66, 0.0042
- P-value is less than 0.05, therefore, we can conclude that the average response time decreased for trial number 5

b)

i)

To detect any differences between adults and children in each trial we use a two-sided t-test.

Table 7 provides estimates for adults in children in each trial. Since we are performing multiple comparisons here, I performed a Benjamini-Hochberg adjustments of p-values to control for false discovery rate. It appears that the only difference appears for trial number 2. This corresponds to Figure 3, which shows that the largest gap in the average trends for two age categories also took place at trial number 2

Table 7: Two-sided T-test comparisons between children and adults in each trial

Trial	Children Average	Adults Average	Est. Differnce	Estimate Std. Error	P-value	BH-Adjusted P-value
1	0.15	0.11	0.11	0.02	0.07	0.09
2	0.14	0.08	0.08	0.02	0.00	**0.01**
3	0.12	0.08	0.08	0.02	0.02	0.05
4	0.12	0.10	0.10	0.02	0.34	0.34
5	0.11	0.08	0.08	0.01	0.06	0.09

[a] N Children in each trial: 60
[a] N Adults in each trial: 47
[a] Used Benjamini-Hochberg procedure for multiple comparison adjustment

ii)

A data summary measure is the average response for each participant. In this section, we perform a two-sample t-test comparing averages for adults vs averages for children. Table 8 shows the average of average values from all 5 trials.

Table 8: Average response time in each trial

Trial	Average Response	Estiamte SE
Adults	0.09	0.01
Children	0.13	0.01

- Null Hypothesis: $Mean_{Adults} = Mean_{Children}$
- Alternative Hypothesis: $Mean_{Adults} \neq Mean_{Children}$
- Difference in means: 0.04, on average, adults scored lower
- Test statistic and p-value: -3.09, 0.0026
- P-value is less than 0.05, therefore, we can conclude that on average, the average for adults over 5 trials was lower than that of children.

iii)

Part A advantages

It is a simple and interpretable way. We can also see the difference from trial to trial, which is beneficial. A more sound statistical approach would be to take the difference from trial 1 to trial 5, and then compare the distribution of differences to zero.

Part A disadvantages

However, last point from *advantages* would also be not optimal in this setting. Ignoring trial change for each subject, a cluster variant covariate, results in conservative inference.

Part B advantages

Analyzing the difference in average of averages allows us to analyze independent observations, so we do not have to deal with covariance structures in this approach.

Part B disadvantages

Means come from a sampling distribution, and means, as a statistic, have inherent variance attributed to them. Therefore, taking variance of a sample of means, who also come from other random samples, is not as straightforward. Analytical solution may be complicated, because

we can not treat them as observed constants. Bootstrap approach may be able to approximate variance estimates that more closely resemble reality.

Appendix

```
path_main_folder = substr(getwd(), 1, nchar(getwd()) - nchar("HW1"))

source(paste0(path_main_folder, "Master Packages.R"))
```

```
timetrial <- read_csv('timetrial.csv')


timetrial_long <-
  timetrial %>%
  pivot_longer(
    cols = c("Trial1", "Trial2", "Trial3", "Trial4", "Trial5"),
    names_to = "Trial",
    values_to = "Measure"
  ) %>%
  mutate(
    Trial_num = as.numeric(substr(Trial, nchar("Trial")+1,  nchar("Trial") + 2)
                                    )
  )
```

```
# answer 1a  appendix

timetrial_long %>%
  filter(Id == '20') %>%
  select(Measure) %>%
  kable(booktabs = T,
        align = 'c',
      col.names = c("Y20")) %>%
  kable_styling(full_width = F, latex_options = c("HOLD_position"))
```

```
# 1c egenrate vcov matrix appendix

vcov_20 <- matrix(c(rep(0.02, 25)), nrow = 5, ncol = 5)

for(i in 1:5){
  vcov_20[i,i] = 0.05
}
```

```r
vcov_20 %>%
  kable(booktabs = T
        ) %>%
  kable_styling(latex_options = "HOLD_position")
```

```r
# correlation matrix 1c  appendix

(vcov_20 /( sqrt(0.05) *sqrt( 0.05))) %>%
  kable(booktabs = T
        ) %>%
  kable_styling(latex_options = "HOLD_position")
```

```r
# predictor matrix 1e appendix

timetrial_long %>%
  filter(Id == 20)  %>%
  select(Trial_num, Age, Sex) ->

  x_20

x_20 %>%
  kable(
    col.names = c("Trial Number", "Age", "Sex"),
    booktabs = T
  ) %>%
  kable_styling(latex_options = "HOLD_position", full_width = F)
```

```r
# predictors appendix

x_20 %>%
  cbind(
    data.frame(Intercept = rep(1, 5)),
    .
  )%>%
  kable(
    col.names = c("","  ", "", ""),
    booktabs = T
  ) %>%
  kable_styling(latex_options = "HOLD_position", full_width = F)
```

```r
# 2a boxplot for outliers  appendix

ggplot(
  data = timetrial_long,
  aes(x = Trial_num, y = Measure, group = Trial)
) +
  theme_minimal() +
  geom_boxplot(fill = NA) +
  geom_jitter(alpha = 0.5, width = .25) +

  labs(
    title = "Distribtion of Reaction Times",
    x = "Trial Number",
    y = "Measurement"
  )
```

```r
# 2b spahetti  appendix

ggplot(
  data = timetrial_long,
  aes(x = Trial_num, y = Measure, group = Id)
) +
  theme_minimal() +
  geom_line(alpha = 0.5, color = "grey") +
  stat_smooth(aes(group = 1)) +

  labs(
    title = "Distribtion of Reaction Times",
    x = "Trial Number",
    y = "Measurement"
  )
```

```r
# 2c spahetti by age group  appendix

timetrial_long <-
  timetrial_long %>%
  mutate(
    age_cat = ifelse(Age < 18, "Children", "Adults")
  )

ggplot(
```

```
  data = timetrial_long,
  aes(x = Trial_num, y = Measure, group = Id)
) +
  theme_minimal() +
  geom_line(alpha = 0.25, color = "grey") +
  stat_smooth(aes(group = age_cat, color = age_cat), se = T) +

  labs(
    title = "Distribtion of Reaction Times",
    x = "Trial Number",
    y = "Measurement",
    color = ""
  )
```

```
# correlation matrix viz 3a  appendix

ggcorrplot(
  timetrial %>% select(Trial1, Trial2, Trial3, Trial4, Trial5) %>% cor(),
  type = "upper",
  outline.color = "white",
  lab = TRUE
)
```

```
# 3b correlation  appendix

timetrial %>% select(Trial1, Trial2, Trial3, Trial4, Trial5) %>% cor() %>% matrix(., nrow
res <- res[res != 1]
```

```
# 4a t test appendix

timetrial_long %>%
  group_by(Trial) %>%
  summarize(mean(Measure),
            sd(Measure)/sqrt(n())
            ) %>%
  filter(Trial %in% c("Trial1", "Trial2") ) %>%
  kable(booktabs = T,
        digits = 2,
        align = 'c',
        col.names = c("Trial", "Average Response", "Estiamte SE")) %>%
```

```
  kable_styling(latex_options = 'HOLD_position')
```

```
# 4a t test appendix

t.test(
  x = timetrial %>% select(Trial5) %>% unlist() ,
  y = timetrial %>% select(Trial1) %>% unlist() ,
  alternative = "less"
) -> ttest_res
```

```
# 4b part 1 appendix

res <-
  data.frame(
    trial = c(1:5),
    kids = rep(NA, 5),
    kids_n = rep(NA, 5),
    adult = rep(NA, 5),
    adults_n = rep(NA, 5),
    diffrence = rep(NA, 5),
    sd = rep(NA, 5),
    p_val = rep(NA, 5)
  )

for(i in 1:5){

  X = timetrial_long %>% filter(Trial_num == i & age_cat == "Adults") %>% select(Measure)
  Y = timetrial_long %>% filter(Trial_num == i & age_cat == "Children") %>% select(Measure

  res$adult[i] = mean(X)
  res$kids[i] = mean(Y)

  res$adults_n[i] = length(X)
  res$kids_n[i] = length(Y)

  t.test(
    x = X,
    y = Y,
    alternative= 'two.sided'
  ) -> iter_t
```

```
  res$diffrence[i] <- iter_t$estimate
  res$sd[i] <- iter_t$stderr
  res$p_val[i] <- iter_t$p.value

  }

res <-
  res %>%
  mutate(
    bh_adj_p_val = p.adjust(p_val, method="BH")
  )

kids_n <- res$kids_n %>% unique()
adults_n <- res$adults_n %>% unique()

res %>%
  select(-kids_n, -adults_n) %>%
  kable(
    align = 'c',
    digits = 2,
    booktabs = T,
    col.names = c("Trial", "Children Average", "Adults Average", "Est. Differnce", "Estima
                  "BH-Adjusted P-value") ) %>%
  add_footnote(paste0("N Children in each trial: ", kids_n)) %>%
  add_footnote(paste0("N Adults in each trial: ", adults_n)) %>%
  add_footnote("Used Benjamini-Hochberg procedure for multiple comparison adjustment") %>%

  column_spec(c(1,3,6), width = '1.5cm') %>%
  column_spec(c(2,4,5,7), width = '2cm') %>%
  column_spec(7, bold = ifelse(res$bh_adj_p_val <= 0.05, T, F) ) %>%

  kable_styling(
    latex_options = c("hover", "condensed", "HOLD_position")
  )
```

```
# 4b part 2 appendix

timetrial_long %>%
  group_by(Id, age_cat) %>%
  summarise(mean_response = mean(Measure)) %>%
  ungroup() -> mean_responses
```

```r
mean_responses %>%
  group_by(age_cat) %>%
  summarise(mean(mean_response),
            sd(mean_response)/sqrt(n())
            ) %>%
  kable(booktabs = T,
        digits = 2,
        align = 'c',
        col.names = c("Trial", "Average Response", "Estiamte SE")) %>%
  kable_styling(latex_options = 'HOLD_position')

t.test(
  x = mean_responses %>% filter(age_cat == "Adults") %>% select(mean_response) %>% unlist(
  y = mean_responses %>% filter(age_cat == "Children") %>% select(mean_response) %>% unlis
  alternative = 'two.sided'
) -> ttest_res
```