Mappeeksamen IDR4000

Student Studentson

2024-09-10

Table of contents

Introduksjon til mappetemplat og Lærinsgrefleksjon

Emnet IDR4000, kvantitativ metode og statistikk bruker mappeeksamen som evalueringsform. Dette er et "templat" for skriving og innlevering av eksamen. Lag en egen versjon av templatet ved å laste ned det (clone/fork) fra GitHub og bruk denne som utgangspunkt. Ta vekk alle instruksjoner før du leverer den siste versjonen av din mappe.

Mappeeksamen består av følgende deler:

- Refleksjoner på læring i emnet
- Rapport: "Deskriptiv statistikk, reliabilitet og validitet og verktøy for reproduserbar vitenskap".
- Laborasjonsrapport fra molekylærlabb
- Arbeidskrav i vitenskapsteori
- Rapport: "Statistisk inferens, statistiske modeller og statistisk styrke"
- Rapport: "Studiedesign"
- Rapport: "Analyse av eksperimenter med repeterte målinger"

I templatet organiseres hver del som et kapittel og instruksjoner for hvert del finner du i templatet.

Instruksjoner for læringsrefleksjon

Mappeeksamen skal starte med en læringsrefleksjon. Her forventes du reflektere over hvordan de forskjellige delene i mappeeksamen (og andre aktiviteter i emnet) bidrar til oppnåelse av emnets læringsutbytte. Du finner læringsutbytte her. Emnet har flere, delvis overlappende mål. En form for refleksjon kan være å ta utgangspunkt i læringsutbytte og beskrive aktiviteter som bidratt til at du oppnådd målene. Refleksjonen bør være et levende dokument i løpet av emnet.

Emnet har Bestått/Ikke bestått som karakterskala men læringsrefleksjonen skal inneholde en egen vurdering av læring. Her kan du bruke de generelle kvalitative beskrivelsene i tabellen under. Her er det ikke tilstrekkelig å sette en symbol eller betegnelse. Det forventes at refleksjonen inneholder motivering til vurderingen og hva som eventuelt savnes for et høyere

nivå. Bruk enten læringsutbytte eller aktiviteter for å vise på eksempel hvor læringsstrategier eller innsats har bidratt/ikke bidratt til måloppnåelse.

Symbol	Betegnelse	Generell, ikke fagspesifikk beskrivelse av vurderingskriterier
A	fremragende	Fremragende prestasjon som klart utmerker seg. Kandidaten viser svært god vurderingsevne og stor grad av selvstendighet.
B	meget god	Meget god prestasjon. Kandidaten viser meget god vurderingsevne og selvstendighet.
C	god	Jevnt god prestasjon som er tilfredsstillende på de fleste områder. Kandidaten viser god vurderingsevne og selvstendighet på de viktigste områdene.
D	nokså god	En akseptabel prestasjon med noen vesentlige mangler. Kandidaten viser en viss grad av vurderingsevne og selvstendighet.
E	tilstrekkelig	Prestasjonen tilfredsstiller minimumskravene, men heller ikke mer. Kandidaten viser liten vurderingsevne og selvstendighet.
F	ikke bestått	Prestasjon som ikke tilfredsstiller de faglige minimumskravene. Kandidaten viser både manglende vurderingsevne og selvstendighet.

For en godkjent mappeeksamen kreves et fullstendig refleksjonsnotat.

1 Assignment 1: Reliability and tools for reproducible data science

The purpose of this assignment is to present estimates of reliability of measures collected in the physiology lab. A second purpose is to use tools for reproducible data science. The report that you are expected to hand in therefore has some strict requirements in its format (see assignment description). The assignment is a group assignment and at least three students are expected to contribute to each report.

1.1 Elements of the report

Importantly, the report should contain:

- At least one table (created from your data)
- At least one figure (created from your data), and
- data presented in the text.
- The report should use a bibliography file to manage references.

2 Assignment 2: Regression models, predicting from data

2.1 Part 1: Lactate thresholds

There are several suggestions on how to best capture the physiological "essence" of the lactate threshold test (See Tanner and Gore 2012, chap. 6). A simple, and very common way to analyze the relationship between exercise intensity and blood lactate is to determine exercise intensity at fixed blood lactate values. This can be done by fitting a regression model that captures the relationship and then "inverse predict" the exercise intensity value. An example of such inverse prediction can be found in the lecture notes.

Your report should use data from the reliability project in the lab. Calculate at least two lactate thresholds (e.g. exercise intensity at 2 and 4 mmol L-1) and compare the reliability (typical error as a percentage of the mean) between the two thresholds. If you want to complicate things further you may want to implement other lactate threshold concepts (described in Tanner and Gore 2012; Newell et al. 2007).

2.2 Part 2: Predicting sizes of DNA fragments

In the molecular laboratory you have been tasked to extract and analyze DNA. In this process we have to determine the size of resulting PCR (polymerase chain reaction) amplified DNA fragments. A tutorial using Image J and R can be found here. In your report you should show how you arrived to your predicted sizes by including the code chunk in your report.

2.3 Part 3: Intepreting a regression table

Using the `hypertrophy` data set, state a question that concerns a linear relationship between two variables in the data set. These variables might be related to muscle size and strength, or two molecular markers or any other variables you are interested in. Include a regression table from your analysis in the report and interpret its components in plain language (e.g. for an unit difference in the independent variable the dependent variable differs by y units). The interpretation should also include a description and explanation of the standard error, the

t-value and the p-value. Valuable guidance on how to interpret the table may be found in for example and in (Frigessi and Aalen 2018), (Campbell, Walters, and Machin 2020) and (Spiegelhalter 2019, chap. 5).

Special attention should be made concerning the p-value. How do you define and interpret the p-value in your regression table. *What does it mean?*.

2.4 How to hand in the report

The report is a group assignment, it is not to be included in the portfolio (mappeeksamen). However, it is required in order to pass the course (arbeidskrav).

Create a new project on github and collaborate with your group there. The repository with all data and coded needed to create the report, and the report itself (in `html`, `docx` or `pdf` format) should be reported on canvas as a link to the repository. **Each member of the group hand in the link in canvas**. The repository should be the same for all group members.

3 Assignment 3: Drawing inference from statistical models, and statistical power

This assignment is set up as a statistical laboratory, we will perform simulations and your assignment is to interpret and explain the results. Create a report based on the code used in the lab and make sure you answer the specified questions (1-8). You can be as creative as you want and explore the results further.

The report should be handed in on canvas as a link to github repository containing a reproducible `.Rmd` (or `qmd`) file.

3.1 Setting up a simulation

In this assignment we will simulate a population of possible values, from this population we will draw random samples, calculate statistics and interpret them. The population of values can be regarded as the possible differences between two treatments in a cross-over study where participants have performed both treatments. The values in the population are calculate as $Treatment - Control$.

We will simulate a population of one million numbers with a mean of 1.5 and a standard deviation of 3. We will make two different set of studies, one set with a sample size of 8 and one set with a sample size of 40. In order to be sure you replicate your results, include and run `set.seed()` before simulations in your final script.

We will use the `lm` function to estimate the average value of the population. We do this in an "intercept-only" model. This model can be written as

$$Y_i = \beta_0 + \epsilon_i$$

where β_0 is the intercept and can be interpreted as the average value of Y, our dependent variable. ϵ is the error term, each observation (i) deviates from the intercept to some degree. If the intercept term is positive or negative we can interpret it as a difference between the two treatments (described above). This model is equivalent to a one-sample t-test. Let's get started!

In the code chunk below, we will simulate the population of differences between treatments. We will then draw two random samples corresponding sample sizes of 8 and 40 and save these data in data frames with the dependent variable named `y`. We fit the very simple model `y ~ 1` as a linear model and save the model object as `m1` and `m2`.

```
library(tidyverse)

set.seed(1)
population <- rnorm(1000000, mean = 1.5, sd = 3)


samp1 <- data.frame(y = sample(population, 8, replace = FALSE))

samp2 <- data.frame(y = sample(population, 40, replace = FALSE))


m1 <- lm(y ~ 1, data = samp1)
m2 <- lm(y ~ 1, data = samp2)

summary(m1)
```

```
Call:
lm(formula = y ~ 1, data = samp1)

Residuals:
    Min      1Q  Median      3Q     Max
-6.5322 -1.2523 -0.0883  1.3540  4.8692

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.840      1.251    1.47    0.185

Residual standard error: 3.539 on 7 degrees of freedom
```

The results from a simple model can be calculated by hand. The `Estimate` corresponds to the average of all values in the sample, from the smaller sample, `samp1` we can do `mean(samp1$y)`. This average should correspond to `coef(m1)` which should be 1.84. The variation of the data is most often described with the standard deviation (SD). The SD of `y` in the smaller sample is `sd(samp1$y)` (corresponding to 3.539). However, the regression table (`summary(m1)`) show you the standard error (SE). This statistic is an attempt to estimate the variation in a hypothetical distribution of means. The standard error is (in this simple case) $SE_y = \frac{SD_y}{\sqrt{n}}$.

Calculating by hand using the data in `samp1` we would do `sd(samp1$y)/sqrt(8)`. Amazingly this corresponds to 1.251!

By using the estimate 1.84 and the corresponding SE (1.251) we can calculate the t-value as the ratio $\frac{Estimate}{SE}$. The t-value may in turn be used to determine the are under the curve of a t-distribution. The t-value from the above calculation is 1.4702611. Using our single $n = 8$ study, we estimate that values of t, as extreme or even more extreme as our observed value both above and below 0, would occur in 18.5% of studies if the null-hypothesis was true. This corresponds to a p-value of 0.185. The figure below shows a graphical representation of a t-value distribution under the assumption that the null-hypothesis is true.

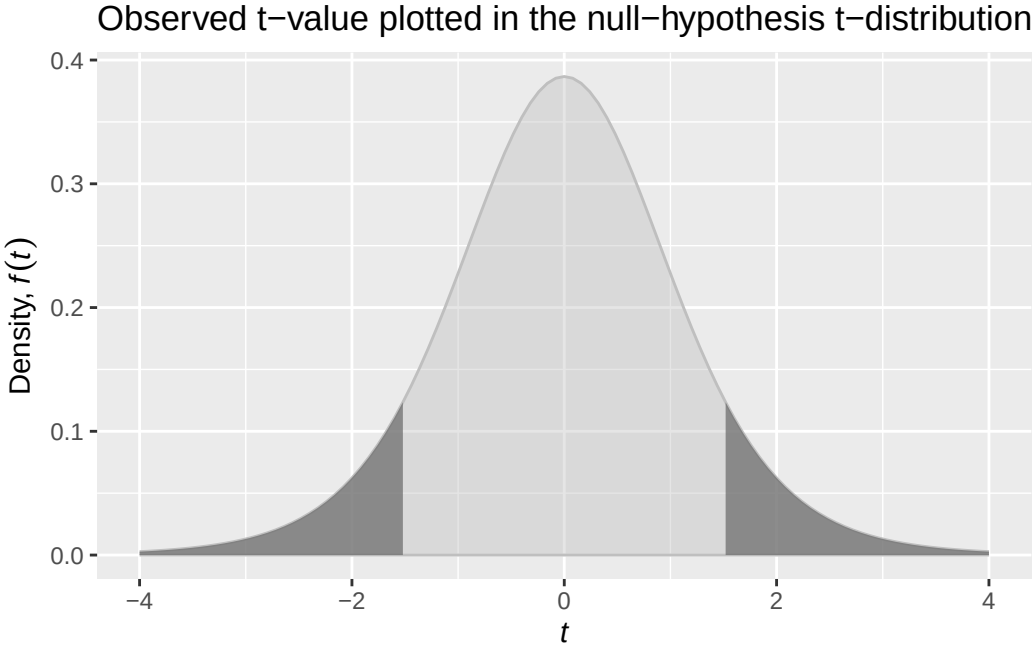


Figure 3.1: A t-distribution estimated from model `m1` with the shaded area corresponding to the observed p-value.

In light of what you know now about the process of conducting a study with a random sample, use your own words and…

1. Explain the estimate, SE, t-value, and p-value from the regression models that we created previously (`m1` and `m2`).
2. Discuss what contributes to the different results in the two studies (`m1` and `m2`).
3. Why do we use the shaded area in the lower and upper tail of the t-distribution (See Figure @ref(fig:t-dist-fig)).

3.2 Many studies

Below we will perform 1000 studies and save the results from each study. This will make it possible for us to get an actual sampling distribution. Copy the code to your own document to run the experiment.

```
# Create data frames to store the model estimates
results_8 <- data.frame(estimate = rep(NA, 1000),
                        se = rep(NA, 1000),
                        pval = rep(NA, 1000),
                        n = 8)

results_40 <- data.frame(estimate = rep(NA, 1000),
                        se = rep(NA, 1000),
                        pval = rep(NA, 1000),
                        n = 40)

# A for loop used to sample 1000 studies, each iteration (i) will draw a new sample
# from the population.

for(i in 1:1000) {

  # Draw a sample
  samp1 <- data.frame(y = sample(population, 8, replace = FALSE))
  samp2 <- data.frame(y = sample(population, 40, replace = FALSE))

  # Model the data
  m1 <- lm(y ~ 1, data = samp1)
  m2 <- lm(y ~ 1, data = samp2)

  # Extract values from the models
  results_8[i, 1] <- coef(summary(m1))[1, 1]
  results_8[i, 2] <- coef(summary(m1))[1, 2]
  results_8[i, 3] <- coef(summary(m1))[1, 4]

  results_40[i, 1] <- coef(summary(m2))[1, 1]
  results_40[i, 2] <- coef(summary(m2))[1, 2]
  results_40[i, 3] <- coef(summary(m2))[1, 4]


}
```

```r
# Save the results in a combined data frame

results <- bind_rows(results_8, results_40)
```

Using the `results` data frame…

4. Calculate the standard deviation of the `estimate` variable, and the average of the `se` variable for each of the study sample sizes (8 and 40). Explain why these numbers are very similar. How can you define the Standard Error (SE) in light of these calculations?
5. Create a histogram (see example code below) of the p-values from each study sample-size. How do you interpret these histograms, what do they tell you about the effect of sample size on statistical power?
6. Calculate the number of studies from each sample size that declare a statistical significant effect (specify a threshold for α, your significance level).
7. Using the `pwr` package, calculate the power of a one-sample t-test, with a effect size of `1.5/3`, your specified significance level and sample sizes 8 and 40. Explain the results in the light of your simulations.

```r
# Example code for copy and paste

# A two facets histogram can be created with ggplot2
results %>%
  ggplot(aes(pval)) +
  geom_histogram() +
  facet_wrap(~ n)


# Count the proportion of tests below a certain p-value for each
results %>%
  filter(pval < 0.05) %>%
  group_by(n) %>%
  summarise(sig_results = n()/1000)

# Using the pwr package
library(pwr)

pwr.t.test(n = 40, sig.level = 0.05, d = 1.5/3, type = "one.sample")
```

3.3 Many studies without population effect

We will now simulate a population without differences between treatment and control. The code below is very similar to the one we use above, except that we use an average effect of 0 in the population.

```
population <- rnorm(1000000, mean = 0, sd = 3)


# Create data frames to store the model estimates
results_8 <- data.frame(estimate = rep(NA, 1000),
                        se = rep(NA, 1000),
                        pval = rep(NA, 1000),
                        n = 8)

results_40 <- data.frame(estimate = rep(NA, 1000),
                        se = rep(NA, 1000),
                        pval = rep(NA, 1000),
                        n = 40)

# A for loop used to sample 1000 studies, each iteration (i) will draw a new sample
# from the population.

for(i in 1:1000) {

  # Draw a sample
  samp1 <- data.frame(y = sample(population, 8, replace = FALSE))
  samp2 <- data.frame(y = sample(population, 40, replace = FALSE))

  # Model the data
  m1 <- lm(y ~ 1, data = samp1)
  m2 <- lm(y ~ 1, data = samp2)

  # Extract values from the models
  results_8[i, 1] <- coef(summary(m1))[1, 1]
  results_8[i, 2] <- coef(summary(m1))[1, 2]
  results_8[i, 3] <- coef(summary(m1))[1, 4]

  results_40[i, 1] <- coef(summary(m2))[1, 1]
  results_40[i, 2] <- coef(summary(m2))[1, 2]
  results_40[i, 3] <- coef(summary(m2))[1, 4]
```

```
}


# Save the results in a combined data frame

results_null <- bind_rows(results_8, results_40)
```

Using the new data frame with results from studies of a population with an average effect of zero, create new histograms.

8. With a significance level of 5%, how many studies would give you a "false positive" result if you did many repeated studies?

4 Assignment 4: Study designs

4.1 Overview

Choose an area of interest (e.g. protein supplementation for muscle hypertrophy or the effect of block periodization on VO2max). Find at least five *original research studies*[1] in your selected area and describe strength and weakness of these studies (see below). The report should focus on the design of the studies and selection of statistical tests to answer study aims. Conclude your report with a recommendation, how should future studies in your area be designed to best answer similar questions?

The report should be handed in on canvas as a link to a github folder containing a reproducible report. This is an individual assignment.

4.2 Details

When **analyzing** your studies you can use the QALMRI method[2].

Note that the report **should not** contain your QALMRI table but should instead be focused on describing differences and similarities in all studies together (see also below)!

4.2.1 The questions may help you analyze your studies

1. What was the broader problem the authors are trying to resolve in the study?
2. What are the specific questions the authors are trying to answer?

   The first point should be similar in all your studies, this could be e.g. the effect of age on physical functioning, effect of certain training protocols on VO2max, etc. The second point is potentially different between your studies.

[1] Avoid using review articles or meta-analyses

[2] See Teaching undergraduate students to read empirical articles: An evaluation and revision of the QALMRI method, this advice was also heavily influenced by this website

4.2.2 Alternative explanations

3. Is the specific question framed as an hypothesis or a question?
4. If the authors have formulated a hypothesis, what alternative explanations can you think of that could potentially explain the data that the authors hypothesize?

4.2.3 Logic

5. What is the logic of the hypothesis or the question. Try to create a "line of logic" between the introduction and the question/hypothesis.

4.2.4 Methods

6. Describe the study design. Use Hulley, (2013), Chapters 7-13[3] in your analysis.
7. Describe the sample and if the study defines the population.
8. Describe the method of recruiting participants to the study. Did the authors justify their sample size (i.e. did they do a power calculation)?
9. Describe how the study was conducted (what tests was performed when etc.)
10. Describe the variables in the study, what variables relate to the question/hypothesis?
11. What methods did the authors use to make claims (what statistical tests were used)

4.2.5 Results

12. What were the main results of the study, did the authors answer their question/address their hypothesis?

4.2.6 Inference

13. What could the authors conclude from the study?
14. What did the authors conclude about the study population?

[3]Hulley, S. B. (2013). Designing clinical research. Philadelphia, Wolters Kluwer/Lippincott Williams & Wilkins.

4.3 Performing your literature review and writing the report

Your report should not contain a detailed summary of all studies for all these questions, instead you should summarize your results. Highlight differences and similarities between studies. *As the main point of this review is study designs and statistical analyses, this should be your main focus.* When doing your literature review, it is however a good idea to structure it in a table with the above mentioned headings:

- Question
- Alternative
- Logic
- Methods
- Results
- Inference

5 Assignment 5: Analyzing repeated measures experiments

Below you will find a basic outline of the report and example code that we worked on today in class. Further below is the description of the assignment.

5.1 Introduction

5.2 Methods

5.2.1 Participants and study overview

5.2.2 Muscle strength and hypertrophy

5.2.3 Data analysis and statistics

5.3 Results

The average difference in lean mass changes between sets were 122.8, 95% CI: [8.6, 237], p = 0.036.

```
## Time points in strength data set

strengthvolume %>%
  distinct(exercise)
```

```
# A tibble: 6 x 1
  exercise
  <chr>
1 legpress
2 legext
3 isok.60
4 isok.120
```

```
5 isok.240
6 isom
```

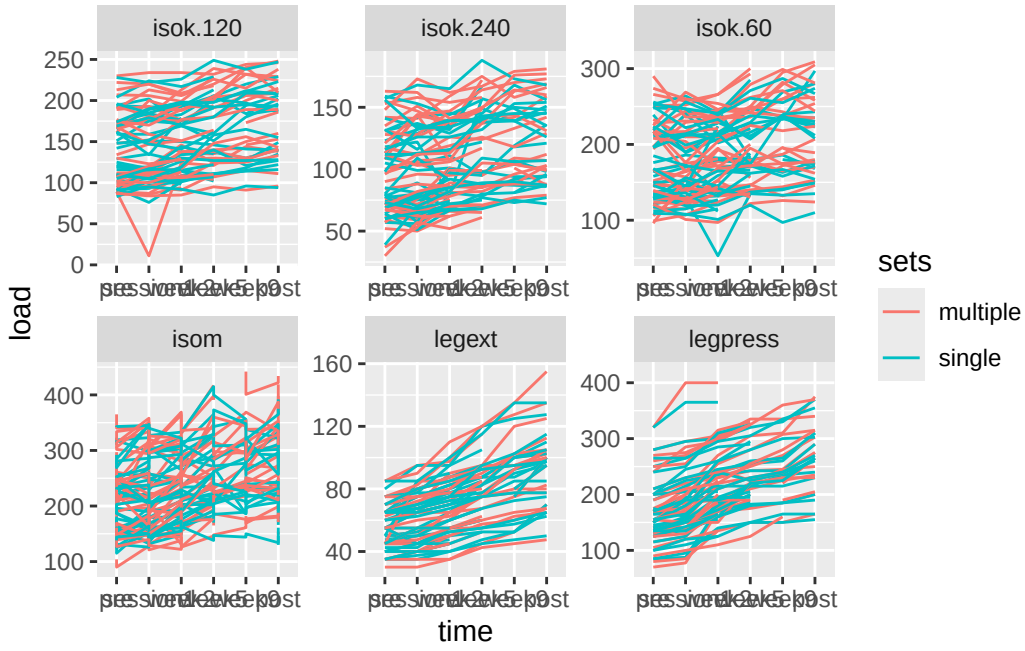
```
## Exploratory plot of strength data

str <- strengthvolume %>%
  filter(include == "incl") %>%
  mutate(time = factor(time, levels = c("pre", "session1",
                                        "week2", "week5",
                                        "week9", "post"))) %>%
  print()
```

```
# A tibble: 2,856 x 8
   participant sex   include time     sets     leg   exercise  load
   <chr>       <chr> <chr>   <fct>    <chr>    <chr> <chr>    <dbl>
 1 FP13        male  incl    pre      single   R     legpress   115
 2 FP13        male  incl    pre      multiple L     legpress   115
 3 FP13        male  incl    pre      single   R     legext      55
 4 FP13        male  incl    pre      multiple L     legext      55
 5 FP13        male  incl    session1 single   R     legpress   125
 6 FP13        male  incl    session1 multiple L     legpress   125
 7 FP13        male  incl    session1 single   R     legext      55
 8 FP13        male  incl    session1 multiple L     legext      55
 9 FP13        male  incl    week2    single   R     legpress   185
10 FP13        male  incl    week2    multiple L     legpress   175
# i 2,846 more rows
```

```
str %>%
  ggplot(aes(time,
             load,
             group = paste(participant, sets),
             color = sets)) +
  geom_line() +
  facet_wrap(~ exercise, scales = "free")
```

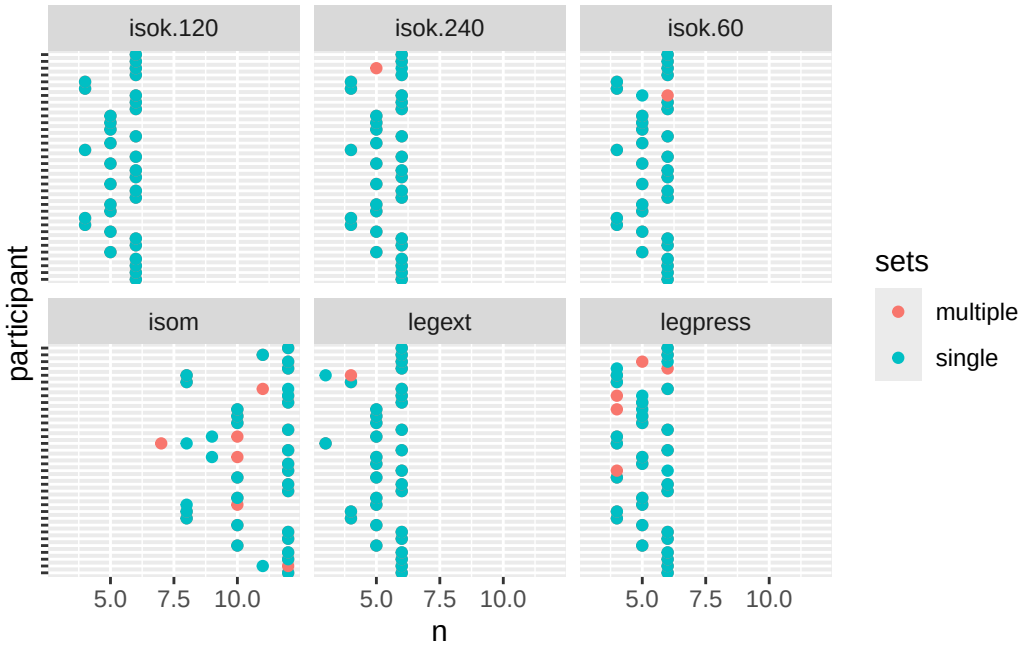
```
Warning: Removed 5 rows containing missing values or values outside the scale range
(`geom_line()`).
```




```
## How many measurements per participant

str %>%
  filter(!is.na(load)) %>%
  group_by(participant, exercise, sets) %>%
  summarise(n = n() ) %>%
  ggplot(aes(n, participant, color = sets)) +
  geom_point() +
  facet_wrap(~ exercise) +
  theme(axis.text.y = element_blank())
```

```
`summarise()` has grouped output by 'participant', 'exercise'. You can override
using the `.groups` argument.
```




```
## Use pre and post data
# Combine pre data prior to data analysis
# per exercise, leg, participant, and sets

str %>%
  mutate(time = if_else(time %in% c("pre", "session1"), "pre", time)) %>%

  filter(time %in% c("pre", "post")) %>%

  summarise(load = max(load, na.rm = TRUE),
            .by = c(participant,
                    sex,
                    time,
                    sets,
                    exercise,
                    leg)) %>%

  print()
```

```
Warning: There were 7 warnings in `summarise()`.
The first warning was:
i In argument: `load = max(load, na.rm = TRUE)`.
```

```
i In group 62: `participant = "FP6"`, `sex = "female"`, `time = "post"`, `sets
  = "multiple"`, `exercise = "legpress"`, `leg = "L"`.
Caused by warning in `max()`:
! no non-missing arguments to max; returning -Inf
i Run `dplyr::last_dplyr_warnings()` to see the 6 remaining warnings.
```

```
# A tibble: 816 x 7
   participant sex    time  sets     exercise leg    load
   <chr>       <chr>  <chr> <chr>    <chr>    <chr> <dbl>
 1 FP13        male   pre   single   legpress R     125
 2 FP13        male   pre   multiple legpress L     125
 3 FP13        male   pre   single   legext   R      55
 4 FP13        male   pre   multiple legext   L      55
 5 FP13        male   post  single   legpress R     230
 6 FP13        male   post  multiple legpress L     235
 7 FP13        male   post  single   legext   R      97.5
 8 FP13        male   post  multiple legext   L     100
 9 FP16        female pre   single   legpress R      95
10 FP16        female pre   multiple legpress L      85
# i 806 more rows
```

5.4 Discussion

5.5 Conclusion

In this assignment you will analyze and report on trial investigating the effect of resistance training volume on lean mass and muscle strength. The data are part of the `exscidata` package and can be accessed as `data("strengthvolume")` and `data("dxadata")`. Read the instructions carefully!

5.6 Organizing the report

Your report should consist of the sections Introduction, Methods, Results and Discussion. Each part of the report should be written as a reproducible document and a link or reference to the repository containing the source document(s) should be included in the report. Below follows detailed descriptions and requirements for each section.

5.6.1 Introduction

This section should consist of a description of the field, resistance-training volume and muscle strength and mass. Use at least five to ten references to introduce your audience and explain why you are doing the analysis/study. A tip is to use the QALMRI method, introduced in Assignment 4 to structure the reading of background information. It is up to you how you motivate the study and how you phrase the purpose of the study. It could be a hypothesis based on previous studies, it could also be question to fill a knowledge gap that you have identified in your literature review.

Structure the introduction in paragraphs. A first paragraph could contain a general introduction to the field, why is it of interest to investigate resistance-training? A second paragraph could specifically describe the specific field of resistance-training volume, why is important to know more about how we are likely to respond to different training volumes. The second paragraph should incorporate definitions important for your report, e.g., training volume, muscle mass and strength. Try to incorporate these definition as a fluid part of the text.

A third (or last) paragraph of the introduction should contain a statement regarding the purpose of the study. The purpose could be descriptive, hypothesis-driven or guided by a question. Although it could be considered a bit backward, you should explore the data sets before you select your question/hypothesis/purpose for it to be possible to answer.

5.6.2 Methods

The method should give a thorough overview of the study and specific details regarding data collection. You can read about the details of this specific study in (Hammarström et al. 2020). Use your own words to describe the study based on this description. A nice way to structure the methods section is to include subheadings:

- Participants and study overview: Describe the participants and give an overview of all tests/measurements. Participants should be described in the first table of the report (Table 1). The overview of the tests/measurements should be done without double presentation as details should be presented in subsequent sections.
- Specific descriptions (e.g. strength tests): Describe in detail how tests/measurements that you mentioned in the overview where conducted.
- Data analysis and statistics: Describe how you treated the data prior to statistical tests or procedures and what tests/procedures were used to draw inference (or more generally, to answer your purpose). Describe how you present data (e.g. descriptive data with mean (SD), inference with confidence intervals etc.).

5.6.3 Results

Describe the results of your analysis. This description should make use of table and figures as well as a text that guides and structures the content to the reader. Think about it this way, the text should describe when and how to read the figures and tables. This means that all aspects of the results should be covered in the text. The figures and tables should also be "self explanatory", this means that you have to include descriptive figure captions and descriptions of tables (see below for tips).

As the main purpose of the analysis should concern the effect of training volume on muscle mass and strength, it is natural that the comparison of training outcomes between volume conditions is the main analysis in the results. You may also have questions regarding the relationship between muscle strength and mass gains, if there are differences between men and women etc. Selection of statistical/analysis techniques should reflect the study question/purpose.

5.6.4 Discussion

Structure the discussion with a first paragraph describing the main results of the analysis, this could be the answer to your question or a statement regarding the study hypothesis. In the following paragraphs discuss all results that you have presented in the light of previous studies. It is your job to give the reader plausible interpretations and explanations of your results. This is how single scientific results are incorporated in our collective understanding. These interpretations can later be challenged, however if you give the reader good arguments and clear descriptions, your insights will be valuable to collective reasoning even if they turn out to be wrong in light of new data.

End the discussion with a summary or conclusion. Some journals request that you state your conclusions under a specific heading in the end of the report/article.

5.7 Organizing the data analysis

5.7.1 Data preparation

The data is already structured in the `exscidata` package. To access the data, use the following code:

```
library(exscidata)
data("dxadata"); data("strengthvolume")
```

To get and overview of the variables in each data set use `?strengthvolume` and `?dxadata`. In the `dxadata` the variables of interest are organized in a more convenient way using the code below:

```
library(tidyverse)

dxadata %>%
  select(participant:include, lean.left_leg, lean.right_leg) %>%
  pivot_longer(names_to = "leg",
               values_to = "lean.mass",
               cols = lean.left_leg:lean.right_leg) %>%
  mutate(leg = if_else(leg == "lean.left_leg", "L", "R"),
         sets = if_else(multiple == leg, "multiple", "single")) %>%
  select(participant, time, sex, include, sets, leg, lean.mass) %>%
  print()
```

```
# A tibble: 160 x 7
   participant time  sex    include sets     leg   lean.mass
   <chr>       <chr> <chr>  <chr>   <chr>    <chr>     <dbl>
 1 FP28        pre   female incl    multiple L          7059
 2 FP28        pre   female incl    single   R          7104
 3 FP40        pre   female incl    single   L          7190
 4 FP40        pre   female incl    multiple R          7506
 5 FP21        pre   male   incl    single   L         10281
 6 FP21        pre   male   incl    multiple R         10200
 7 FP34        pre   female incl    single   L          6014
 8 FP34        pre   female incl    multiple R          6009
 9 FP23        pre   male   incl    single   L          8242
10 FP23        pre   male   incl    multiple R          8685
# i 150 more rows
```

6 Philosophy of science

See instructions on canvas.

7 Molecular Laboratory report

References

Campbell, Michael J., Stephen John Walters, and David Machin. 2020. *Medical Statistics: A Textbook for the Health Sciences*. Fifth edition. Hoboken, NJ: Wiley-Blackwell.

Frigessi, Arnoldo, and Odd O Aalen. 2018. *Statistiske Metoder i Medisin Og Helsefag*. Oslo: Gyldendal akademisk.

Hammarström, Daniel, Sjur Øfsteng, Lise Koll, Marita Hanestadhaugen, Ivana Hollan, William Apró, Jon Elling Whist, Eva Blomstrand, Bent R. Rønnestad, and Stian Ellefsen. 2020. "Benefits of Higher Resistance-Training Volume Are Related to Ribosome Biogenesis." Journal Article. *The Journal of Physiology* 598 (3): 543–65. https://doi.org/10.1113/JP278455.

Newell, J., D. Higgins, N. Madden, J. Cruickshank, J. Einbeck, K. McMillan, and R. McDonald. 2007. "Software for Calculating Blood Lactate Endurance Markers." Journal Article. *Journal of Sports Sciences* 25 (12): 1403–9. https://doi.org/10.1080/02640410601128922.

Spiegelhalter, D. J. 2019. *The Art of Statistics : How to Learn from Data*. Book. First US edition. New York: Basic Books.

Tanner, R. K., and C. J. Gore. 2012. *Physiological Tests for Elite Athletes 2nd Edition*. Book. Human Kinetics. https://books.google.no/books?id=0OPIiMks58MC.