Intermediate Quantitative Methods

Lucas Lemmann

2023-10-24

Contents

About

What is this book about? What for?

0.1 How to use these exercises?

- Besides the 14 lectures, the course will be organized around 12 non-graded exercises:
    - 5 labs
    - 7 do-it-yourself (DIYS)
- The labs' solutions will be discussed in detail between TAs and students in the corresponding sessions, while DIYS will not. In both cases, we will publish the solutions the week after the exercise is due.
- We encourage you to prepare for the lab sessions in advance as well as to attend them to discuss any doubts they might have related to the labs material.
- To prevent redundant communications (i.e., emails with the same information), share your questions regarding the exercises in the forum. Labs will emphasize the most voted questions.
- While we encourage and foster a collaborative learning process, we expect you to work individually first.
    - I.e., try to address the task on your own first, identify what is limiting you, try to solve it on your own (not for too long), and, if you cannot find a solution, reach out your classmates. Once you find your solution, consider discussing the solution with your classmates.

0.2 Schedule

Week	Dates	Exercise type
1	19-25/02	DIYS 1

Week	Dates	Exercise type
2	26/02-03/03	Lab 1
3	04/03-10/03	Lab 1
4	11/03-17/03	DIYS 2
5	18/03-24/03	Lab 2
6	25/03-31/03	DIYS 3
Spring Break	28/03-07/04	None?
7	08/04-14/04	Lab 3
8	15/04-21/04	DIYS 4
9	22/04-28/04	Lab 4
10	29/04-05/05	DIYS 5
11	06/05-12/05	Lab 5
12	13/05-19/05	Lab 5
13	20/05-26/05	DIYS 6
14	27/05-02/06	DIYS 7

Chapter 1

Week 1: DIYS 1

1.1 Aim:

To refresh your R skills by performing some basic analyses (i.e., descriptive, exploratory, and hypothesis testing ones).

1.2 First Part: Descriptive Analysis

1. Download the files `f.txt` and `m.txt`. They contain information on the number of steps in a day and the body mass index (BMI) for female and male individuals respectively. Open them and explore the first 5 observations for each file.

Adjust using the links from GitHub

```
# Your code goes here
```

For the exercise before publishing the solution

```
# open data
female <- read.table("~/Documents/0_IPZ/2023_2/Leemann-QuantMethods/QuantitativeMethods/Quantitat

# explore data
head(female, 3)
```

```
##   ID steps  bmi
## 1  3 15000 17.0
## 2  4 14861 17.2
## 3  5 14861 17.2
```

```
# open data
male <- read.table("~/Documents/0_IPZ/2023_2/Leemann-QuantMethods/QuantitativeMethods/(

# explore data
head(male, 3)
```

```
##   ID steps  bmi
## 1  1 15000 16.9
## 2  2 15000 16.9
## 3  6 14861 16.8
```

Some key functions in dplyr can be categorized as dealing with columns (e.g., `select`, `mutate`), rows (e.g., `filter`, `distinct`, `arrange`), or groups (e.g., `group_by`, `summarise`, and `count`). Let's use some of them!

2. Select only the columns 'steps' and 'bmi'. Do it only for the first three observations of the data on females.

```
# It's necessary to restate it in each r code section so the book can be rendered.
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
head(female, 3) %>%
  select(steps, bmi)
```

```
##   steps  bmi
## 1 15000 17.0
## 2 14861 17.2
## 3 14861 17.2
```

3. Select all columns except 'ID'. Do not use `steps` nor `bmi`. Do it only for the first three observations of the data on females. Is the resulting table the same as the previous point? If not, check your answer.

```
library(dplyr)
head(female, 3) %>%
  select(-ID)
```

```
##   steps  bmi
## 1 15000 17.0
## 2 14861 17.2
## 3 14861 17.2
```

Note: to check the documentation of `select`, use `?select` on the console.

4. Use `mutate` to create a new column in the dataframe `female` called `StepsTimesBmi` formed as the product of `steps` and `bmi`. Show the first three observations for the new variable.

```
library(dplyr)
female <-female %>%
  mutate(StepsTimesBmi= steps * bmi)

head(female$StepsTimesBmi,3)
```

```
## [1] 255000.0 255609.2 255609.2
```

5. Get rid of the column `StepsTimesBmi`. Use `subset`.

```
female <- subset(female, select= - StepsTimesBmi)
```

6. Use filter to find the share of female individuals with a `bmi` higher than 20 and lower than 21.

```
library(dplyr)
f20_21 <- female %>%
  filter(bmi>20, bmi<21)

cat("The share of female individuals with a `bmi` higher than 20 and lower than 21 is:", nrow(f20
```

```
## The share of female individuals with a `bmi` higher than 20 and lower than 21 is: 2.28013 %
```

7. Use filter to find the share of female individuals with a `bmi` higher than 20 and lower than 21 while at the same time having less than 14000 `steps`.

```
library(dplyr)
fBMI20_21_Step14000 <- female %>%
  filter(bmi>20, bmi<21, steps<14000)

cat("The share of female individuals with a `bmi` higher than 20 and lower than 21 whil
```

```
## The share of female individuals with a `bmi` higher than 20 and lower than 21 while
```

8. Use filter to find the share of male individuals with `ID` number lower than 5 **and** higher than 860. Notice that you can use either `&` between conditions or simply a comma. Could any data set generate a different answer? Why?

```
library(dplyr)
m_5_860 <- male %>%
  filter(ID<5 & ID>860)

cat("The share of male individuals with `ID` number lower than 5 AND higher than 860 i
```

```
## The share of male individuals with `ID` number lower than 5 AND higher than 860 is:
```

9. Use filter to find the share of male individuals with `ID` number lower than 5 **or** higher than 860. Use `|` between conditions. Could any data set generate a different answer? Why?

```
library(dplyr)
m_5_or_860 <- male %>%
  filter(ID<5 | ID>860)

cat("The share of male individuals with `ID` number lower than 5 OR higher than 860 is
```

```
## The share of male individuals with `ID` number lower than 5 OR higher than 860 is: 4
```

10. Use `distinct` to identify the share of male IDs that are unique.

```
unique_m_IDs <- male %>%
  distinct(ID)

cat("The share of male IDs that are unique is:", nrow(unique_m_IDs)*100/nrow(male), "%
## The share of male IDs that are unique is: 100 %
```

11. Use `arrange` to find the three highest and lowest BMI values for males. Use `slice_head`.

```
# Max
top_3_m <- male %>%
  arrange(desc(bmi)) %>%
  slice_head(n = 3)
print(top_3_m)
```

```
##    ID steps bmi
## 1 786  7894  32
## 2 847  7593  32
## 3 863  7431  32
```

```
# Min
bottom_3_m <- male %>%
  arrange(bmi) %>%
  slice_head(n = 3)
print(bottom_3_m)
```

```
##     ID steps  bmi
## 1 1170  6366 15.7
## 2  614  9097 15.8
## 3  615  9097 15.8
```

```
9. group_by summarise count
```

2. Are there repeated ids within each data set?

```
# get package
# install.packages("dplyr")
library(dplyr)


# Check for repeated IDs in the female data set. How many are there?
repeated_ids_female <- female %>%
  group_by(ID) %>%
  filter(n() > 1)

cat("Number of repeated IDs in the female data set:", nrow(repeated_ids_female), "\n")
```

```
## Number of repeated IDs in the female data set: 0
```

```r
# Check for repeated IDs in the male data set. How many are there?
repeated_ids_male <- male %>%
  group_by(ID) %>%
  filter(n() > 1)

cat("Number of repeated IDs in the male data set:", nrow(repeated_ids_male), "\n")
```

```
## Number of repeated IDs in the male data set: 0
```

1.3 Second Part: Exploratory VS. Hypothesis-Testing Analysis

Please read the whole instruction before solving the exercise.

Each student will be randomly allocated to either doing the task 1 or 2 (a list containing those numbers will published). Both tasks are based on the same data sets used in the first part.

Notes:

- The details of the data origin will be published with the solution.
- Students allocated to each group are encouraged to do the task for the other group ***only*** after finishing their own task.

1.3.1 Task 1:

- What do you conclude from the combined data set (i.e., the one formed using both the one for males and the one for females)?
- What questions did you ask yourself?
  - Why did you ask those questions? Is there an intuition behind them?
    - If so, what was your intuition?
    - If not, how did you proceed?
- Hint: consider visualizing how variables interact.

1.3.2 Task 2:

- Is the average number of steps for males and females statistically different?
- How do BMI and daily steps statistically relate to each other?

- Does that relationship depend on whether individuals are of one sex or another? If so, how?
  - Is there an statistically significant negative correlation between the number of steps and the BMI for females?
  - Is there an statistically significant positive correlation between the number of steps and the BMI for males?

1.3.3 Preliminary steps: do this before doing the task that you were assigned to

1. For each data set, create a new variable called `sex`. Assign any value to each case, but make sure they are different.

```r
female$sex <- 'F'
male$sex <- 'M'
```

2. Create one data frame with all the IDs present in **both** data sets. How many cases are there? Use `dplyr`'s join methods.

```r
library(dplyr)

in_both <- inner_join(female, male, by="ID")

cat("The number of cases where an ID is in both data sets is:", nrow(in_both), "\n")
```

```
## The number of cases where an ID is in both data sets is: 0
```

3. Now that you know that there are no repeated individuals across the data sets, consider whether a join method is the appropriate way of unifying both data sets. Try first with `full_join` and then with `bind_rows`. Which one should you use? Why? Finally, how many individuals does the new dataframe have?

```r
library(dplyr)

all <- full_join(female, male, by="ID", copy=FALSE)
cat("The new dataframe has ", nrow(all), "individuals\n")
```

```
## The new dataframe has  1786 individuals
```

```
# Assuming that `sex` was created for each dataframe
all <- bind_rows(female, male, .id = NULL)
cat("The new dataframe has ", nrow(all), "individuals\n")
```

```
## The new dataframe has  1786 individuals
```

```
# Without assuming that `sex` was created for each dataframe
female <- read.table("~/Documents/0_IPZ/2023_2/Leemann-QuantMethods/QuantitativeMethods
male <- read.table("~/Documents/0_IPZ/2023_2/Leemann-QuantMethods/QuantitativeMethods/

all <- bind_rows(female, male, .id = 'sex')
cat("The new dataframe has ", nrow(all), "individuals\n")
```

```
## The new dataframe has  1786 individuals
```

```
# Which assigns a number 1 for the first binded dataframe, and 2 for the second one. H
all$sex <- ifelse(all$sex == 1, 'F', ifelse(all$sex == 2, 'M', all$sex))
```

4. What’s the share per sex in the unified dataframe from the previous point?

Consider using the packages `dplyr`, “

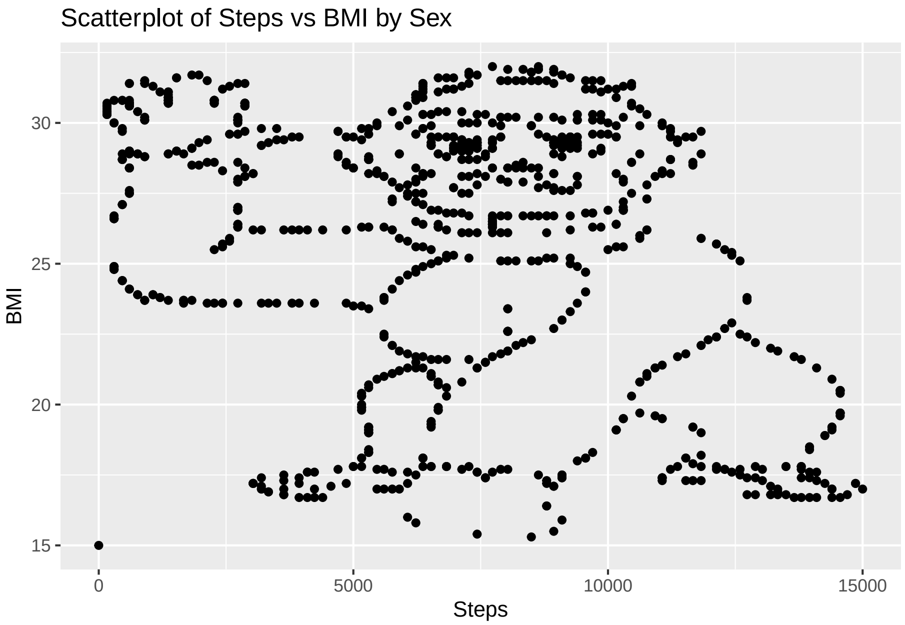
- 1st weeks, dplier: to check> to statistical analysis
  - Doing basic code to make analysis (which is fine enough), but in dplier you could do it like this.
  - Make descriptive statistics using an interesting

looking for something unknown in the dark, grope, feel blindly and make conjectures on what things are and how they are related. - Two groups: random selection: description similar? The smaller the group, the likelier that a random selection is not balanced? What about attrition?

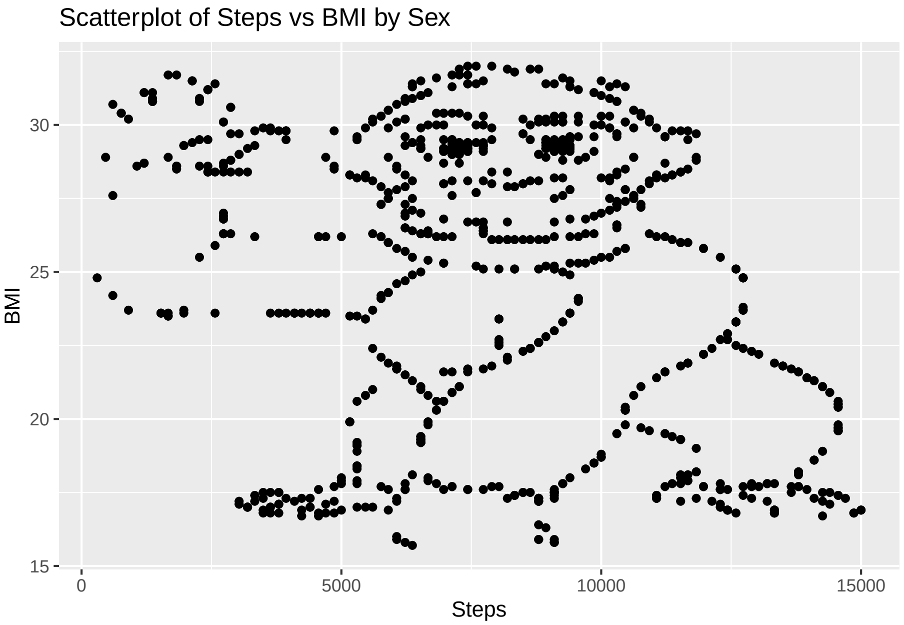
Looking!=seeing: Different beliefs (non- and knowledge ones), different preferences, different attention focus -> different attention investment and emphasis Value of diverse academic community while keeping a minimal set of shared assessment rules: objectivity as continuum of increasing inter-subjective agreement

1.4 Graph

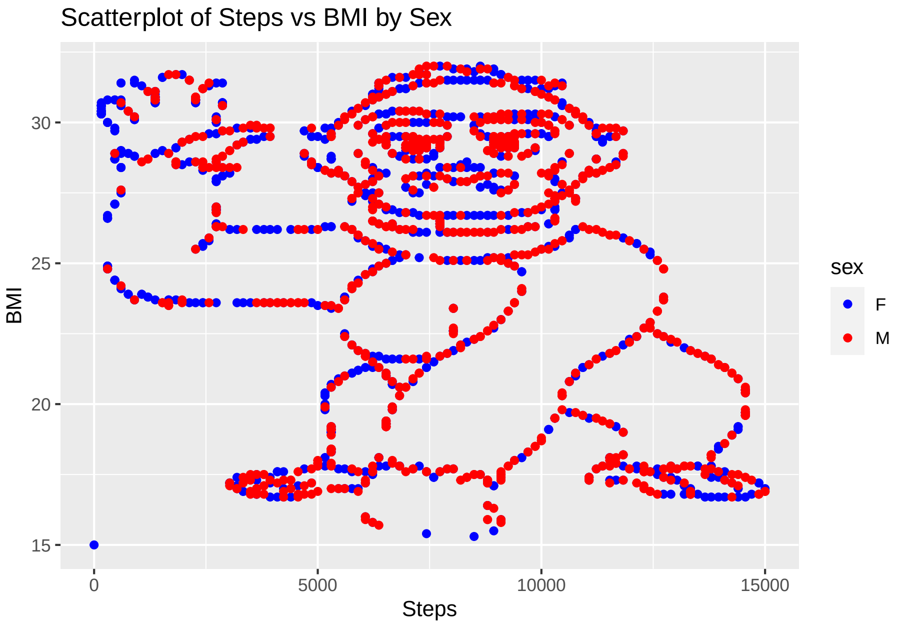
```r
# Install and load the ggplot2 package if you haven't already
#install.packages("ggplot2")
library(ggplot2)

# Assuming you have a consolidated dataset named 'combined_data'

# Create a scatterplot of steps vs bmi

ggplot(female, aes(x = steps, y = bmi)) +
  geom_point() +
  labs(x = "Steps", y = "BMI") +
  ggtitle("Scatterplot of Steps vs BMI by Sex")
```



```r
ggplot(male, aes(x = steps, y = bmi)) +
  geom_point() +
  labs(x = "Steps", y = "BMI") +
  ggtitle("Scatterplot of Steps vs BMI by Sex")
```



```r
ggplot(all, aes(x = steps, y = bmi, color = sex)) +
  geom_point() +
  labs(x = "Steps", y = "BMI") +
  ggtitle("Scatterplot of Steps vs BMI by Sex") +
  scale_color_manual(values = c("F" = "blue", "M" = "red"))  # Optional: Define color
```



1.5 Solution

Will be made available.

Chapter 2

Week 2

2.1 Exercise

- 2nd: simulated dataset and increase the variance: how does that affects the standard error

2.2 Solution

- Data taken from here.
- Original selective attention, here.
- Suicide awareness campaign, here.

Chapter 3

Week 3

3.1 Exercise

3.2 Solution

Chapter 4

Week 4

4.1 Exercise

4.2 Solution

Chapter 5

Week 5

5.1 Exercise

5.2 Solution

Chapter 6

Week 6

6.1 Exercise

6.2 Solution

Chapter 7

Week 7

7.1 Exercise

7.2 Solution

Chapter 8

Week 8

8.1 Exercise

8.2 Solution

Chapter 9

Week 9

9.1 Exercise

9.2 Solution

Chapter 10

Week 10

10.1 Exercise

10.2 Solution

Chapter 11

Week 11

11.1 Exercise

11.2 Solution

Chapter 12

Week 12

12.1 Exercise

12.2 Solution

Chapter 13

Week 13

13.1 Exercise

13.2 Solution

Chapter 14

Week 14

14.1 Exercise

14.2 Solution