Problem Set Week 1

Helge Moes

15-01-2023

Table of Contents

## Directions for the student

* Put all R code in code chunks and verbal answers outside code chunks.
* If you get a piece of R code to work, set the code chunk option eval=FALSE to ensure the document can still be knitted.
* Use tidyverse functions whenever possible.
* Comment your code to communicate your intentions.
* Ensure that the R Markdown document knits without problems into a PDF or Word document.
* Submit the R Markdown document on Canvas (under Assignments) before the deadline.

#Load all libraries in this code chunk.  
library(tidyverse)

## ── Attaching packages ─────────────────────────────────────── tidyverse 1.3.2 ──  
## ✔ ggplot2 3.4.0 ✔ purrr 1.0.1   
## ✔ tibble 3.1.8 ✔ dplyr 1.0.10  
## ✔ tidyr 1.2.1 ✔ stringr 1.5.0   
## ✔ readr 2.1.3 ✔ forcats 0.5.2   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()

library(dplyr)  
library(readr)  
library(ggplot2)

## Data

**sleepers.csv**. The effect of two drugs (variable *group*) on the increase in hours of sleep compared to a control condition (variable *extra*) in 67 patients (variable *ID*). The patient’s age group (*age*, 1 = 18-25, 2 = 26-35, 3 = 36-50, 4 = 50+) is provided as well.

## Questions

1. Import the data in R with readr:: and count the missing values on each variable using R code.

#Add your R code for answering this question here.  
#Read file  
sleepers <- readr::read\_csv("sleepers.csv")

## Rows: 134 Columns: 4  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (1): group  
## dbl (3): ID, age, extra  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

sleepers#table of data

## # A tibble: 134 × 4  
## ID group age extra  
## <dbl> <chr> <dbl> <dbl>  
## 1 1 Drug1 1 4.41   
## 2 2 Drug1 4 1.81   
## 3 3 Drug1 1 4.51   
## 4 4 Drug1 3 -0.269  
## 5 5 Drug1 1 4.32   
## 6 6 Drug1 1 2.38   
## 7 7 Drug1 4 1.03   
## 8 8 Drug1 4 1.68   
## 9 9 Drug1 3 -1.05   
## 10 10 Drug1 3 1.31   
## # … with 124 more rows

#Check: count missing values  
sum(is.na(sleepers$group)) #0

## [1] 0

sum(is.na(sleepers$age)) #3

## [1] 3

sum(is.na(sleepers$extra)) #0

## [1] 0

#Count missing values  
sleepers %>%   
 summarise\_all(funs(sum(is.na(.))))

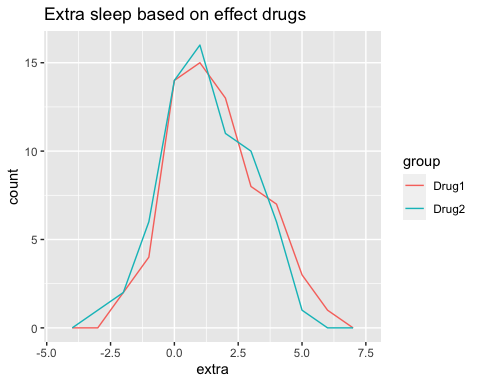
## Warning: `funs()` was deprecated in dplyr 0.8.0.  
## ℹ Please use a list of either functions or lambdas:  
##   
## # Simple named list: list(mean = mean, median = median)  
##   
## # Auto named with `tibble::lst()`: tibble::lst(mean, median)  
##   
## # Using lambdas list(~ mean(., trim = .2), ~ median(., na.rm = TRUE))

## # A tibble: 1 × 4  
## ID group age extra  
## <int> <int> <int> <int>  
## 1 0 0 3 0

| Grading | Max points | Awarded |  |
| --- | --- | --- | --- |
| Ex. 1 | 1 |  |  |

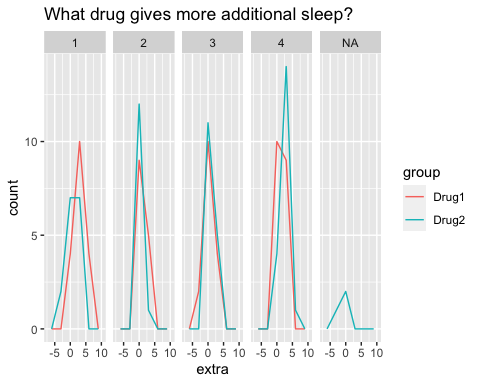
1. Use ggplot2:: to create one figure containing a frequency polygon of sleep increase for each drug. Use different colors for the two polygons. Which drug seems to increase sleep more?

#Add your R code for answering this question here.  
  
ggplot2::ggplot(data = sleepers, aes(x = extra, fill = group, color = group)) + geom\_freqpoly(binwidth = 1) + ggtitle("Extra sleep based on effect drugs") + scale\_fill\_discrete(name = "Drug", labels = c("Drug 1", "Drug 2"))

 Grading | Max points | Awarded | ——- | ———- | —— | ——- Ex. 2 | 1 | |

1. For each age group, create a figure as in Question 2 using facetting. Again, use ggplot2::. Which drug works better (gives more additional sleep) for which age group?

#Add your R code for answering this question here.  
  
ggplot2::ggplot(data = sleepers, aes(x = extra, fill = group, color = group)) + geom\_freqpoly(binwidth = 3) + ggtitle("What drug gives more additional sleep?") + scale\_fill\_discrete(name = "Drug", labels = c("Drug 1", "Drug 2")) + facet\_grid(~ age)



#Answer: Based on the figures with a binwidth of 3, Drug 2 portrays better results (and therefore addtional sleep) for age groups 2, 3 and 4 than Drug 1. Yet, Drug 1 is more succesful for age group 1. Furthermore, Drug 2 also portrays missing values in figure 'NA'. However, this shows what is lost during the conduct of this research

### Answer:

Based on the figures with a binwidth of 3, Drug 2 portrays better results (and therefore addtional sleep) for age groups 2 (26-35), 3 (36-50) and 4 (50+) than Drug 1. Yet, Drug 1 is more succesful for age group 1 (18-25). Furthermore, Drug 2 also portrays missing values in figure ‘NA’. However, this shows what is lost during the conduct of this research

| Grading | Max points | Awarded |  |
| --- | --- | --- | --- |
| Ex. 3 | 2 |  |  |

1. Use dplyr:: to create a table (tibble/data frame) containing the number of observations, the minimum, maximum, and average sleep increase for each drug.

#Add your R code for answering this question here.  
  
sleepers %>%  
 group\_by(group) %>%  
 summarise(n = n(), min = min(extra), max = max(extra), mean = mean(extra))

## # A tibble: 2 × 5  
## group n min max mean  
## <chr> <int> <dbl> <dbl> <dbl>  
## 1 Drug1 67 -2.07 5.98 1.61  
## 2 Drug2 67 -2.72 4.98 1.32

| Grading | Max points | Awarded |  |
| --- | --- | --- | --- |
| Ex. 4 | 1 |  |  |

1. Calculate the difference in extra sleep for each patient between drug 2 and drug 1, again using dplyr::. Show the first 10 rows of the results.

Tip: Sort the data and, in the end, only retain one case for each patient.

#Add your R code for answering this question here.  
  
sleepers\_difference <- sleepers %>%  
 group\_by(ID) %>%  
 summarize(extra\_difference = sum(extra[group == "Drug2"] -- extra[group == "Drug2"])) %>%  
 select(ID, extra\_difference)  
  
head(sleepers\_difference, 10)

## # A tibble: 10 × 2  
## ID extra\_difference  
## <dbl> <dbl>  
## 1 1 -0.353   
## 2 2 8.53   
## 3 3 0.0637  
## 4 4 4.01   
## 5 5 4.38   
## 6 6 3.58   
## 7 7 9.95   
## 8 8 2.59   
## 9 9 4.07   
## 10 10 1.84

| Grading | Max points | Awarded |  |
| --- | --- | --- | --- |
| Ex. 5 | 2 |  |  |

1. Create a plot to explore the covariation (association) between two continuous variables in *your project’s data set* that are of interest to you. Use comments in the R code to justify the choices that you made to create this plot.

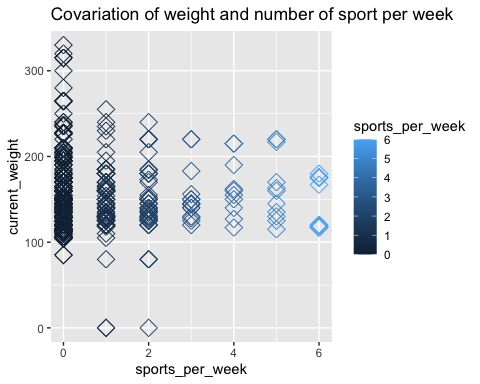
Note: You may use grouped summaries (aggregation) to create continuous variables.

#Add your R code for answering this question here.  
  
#The dataset Health.csv is used in order to research the covariation between the frequency of sport per week and weight of patients. To address the covariation of these two symptoms, a geom\_point has been implemented, since it allows the user to use x and y variables in order to portray covariates in the data.  
  
Health <- readr::read\_csv("Health.csv")

## Rows: 385 Columns: 10  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: ","  
## chr (2): healthy\_diet, current\_smoking  
## dbl (8): user\_id, current\_weight, current\_height, salads\_per\_week, veggies\_f...  
##   
## ℹ Use `spec()` to retrieve the full column specification for this data.  
## ℹ Specify the column types or set `show\_col\_types = FALSE` to quiet this message.

ggplot2::ggplot(data = Health, aes(x = sports\_per\_week, y = current\_weight, color = sports\_per\_week)) + geom\_point(size = 4, shape = 5) + ggtitle("Covariation of weight and number of sport per week") + scale\_fill\_discrete(name = "sports\_per\_week", labels = c("sports\_per\_week", "current\_weight"))

## Warning: Removed 4 rows containing missing values (`geom\_point()`).



### Answer:

Based on the figure, the frequency of patients that do not sport per week weigh more than those that sport more frequently. Therefore, the more one sports, the less they weigh. The number of people that sport more a week decreases as so does the weight level. The gradiant color of black and blue emphasizes the number of times a patient sports a week. The missing values are set at 0, since the patient was reluctant to share information of their weight in the research. Furthermore, 4 rows with missing values were removed of the data.

| Grading | Max points | Awarded |  |
| --- | --- | --- | --- |
| Ex. 6 | 2 |  |  |
| Flawless knitting | 1 |  |  |
| **Total** | 10 |  |  |