Comparing means and fitting distributions

Introduction

I will be comparing the weights of mice and rats before and after taking a nutritional supplement over a duration of 6 months. To do this I will be generating two artificial datasets: one containing the weights of mice before and after receiving the treatment, and the other for the weights of rats before and after receiving the treatment. Will be used to evaluate the impact of the treatment on the weight of mice and rats. For evaluation purposes, I will be performing several tests to identify whether the data fits a normal distribution or not and provide and analysis of the results: Shapiro Wilk test for normality, paired t-test and non-parametric tests, etc.

Table of Contents

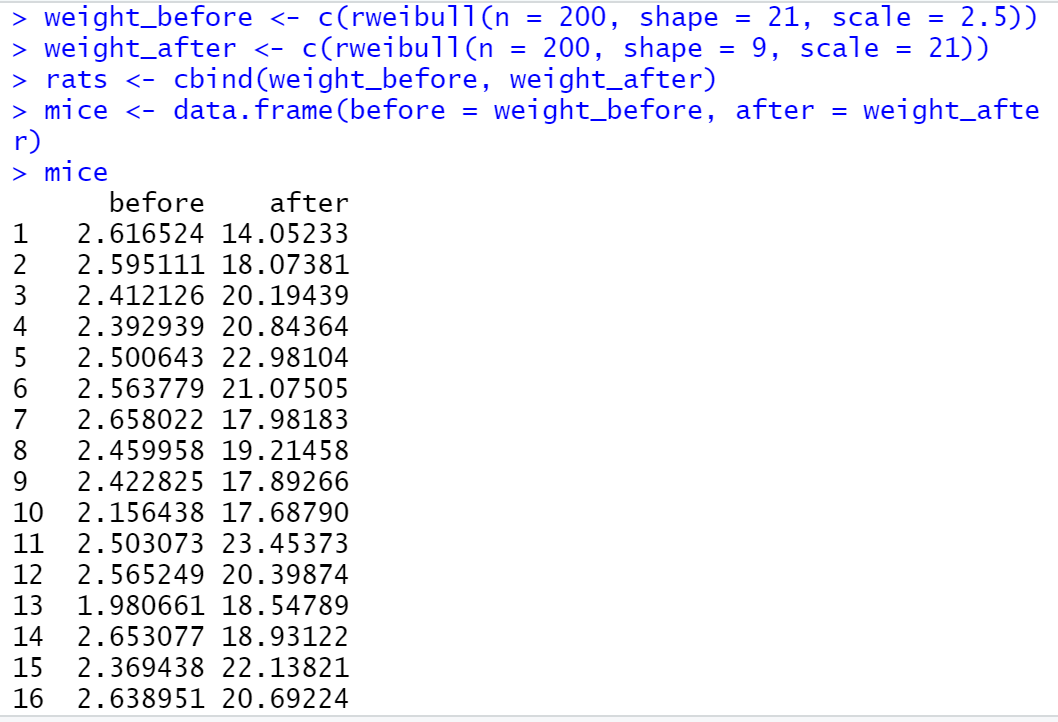
[**Task 1: Data generation** 3](#_Toc67770877)

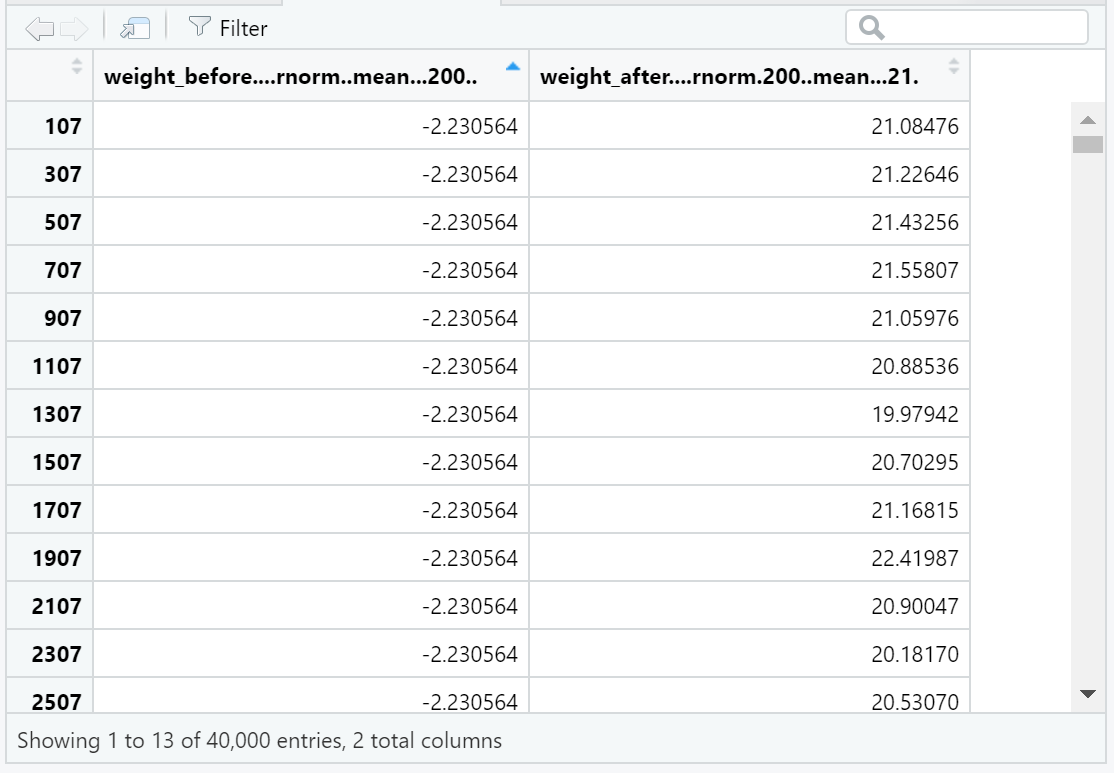
[Bibliography 4](#_Toc67770878)

# **Task 1: Data generation**

In this section of the report, I will be generating two artificial datasets, one for mice and the other for rats, that will be used to measure the suitability of a trial run for the development of a new treatment for weight loss. For evaluation purposes I will be calculating the mean and variance, creating several plots,; and fitting the data with a variety of distributions. and discussing my findings.

The generated mice and rats’ datasets use the rnorm (random distribution) and Weibull (Weibull distribution) base packages respectively to generate 200 values of a numeric type in a data frame for analysis. These datasets can be seen below, along with a summary:





The reason for using a data frame is the data is stored in a way that is convenient and easy to interpret. This data structures stores data in a structure like a table. The random distribution is a type of discrete and random probability distribution that contains random variables of a numeric set that are mapped to a sample space (). In the discrete case, the probability mass function ( is used to analyse the data. This works by assigning probabilities to each outcome. Whereas, the probability density function () and cumulative distribution function () can be used to analyse the data in the continuous case. By integrating the probability density function over an interval.

There are many variations of the Weibull distribution, each with different properties. some of these variations can be considered continuous, and others discrete. This distribution is used to assess the reliability of products, analyse models of data from real life scenarios, and model failures over a specified period. The formula for this distribution depends on the use case. For this report, I will be using the following formulas, as I will be retrieving the mean, variance, and kurtosis of my dataset:

Mean: or mean () in R.

Variance: or var () in R.

|  |  |  |  |
| --- | --- | --- | --- |
| gamma_1 | = |  | (9) |
| gamma_2 | = | |  |  |  |  | | --- | --- | --- | --- | | gamma_1 | = |  | (9) | | gamma_2 | = | (f(alpha))/([Gamma(1+2alpha^(-1))-Gamma^2(1+alpha^(-1))]^2), |  | |  |

Kurtosis(2Gamma^3(1+alpha^(-1))-3Gamma(1+alpha^(-1))Gamma(1+2alpha^(-1)))/([Gamma(1+2alpha^(-1))-Gamma^2(1+alpha^(-1))]^(3/2))+(Gamma(1+3alpha^(-1)))/([Gamma(1+2alpha^(-1))-Gamma^2(1+alpha^(-1))]^(3/2))rtosis: (2Gamma^3(1+alpha^(-1))-3Gamma(1+alpha^(-1))Gamma(1+2alpha^(-1)))/([Gamma(1+2alpha^(-1))-Gamma^2(1+alpha^(-1))]^(3/2))+(Gamma(1+3alpha^(-1)))/([Gamma(1+2alpha^(-1))-Gamma^2(1+alpha^(-1))]^(3/2))

From the statistics obtained by running the summary using the mice dataset, I can conclude that the data comes from a normal distribution, as the values do not differ significantly. The values obtained can be seen below:

weight\_before

Min. :15.95

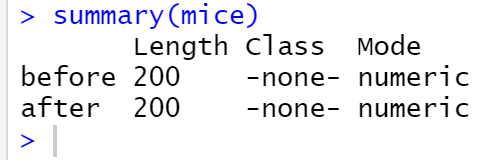
1st Qu.:19.12

Median :20.12

Mean :20.07

3rd Qu.:21.16

Max. :24.55



|  |  |
| --- | --- |
|  | Weight\_before |
| Min. | :15.95 |
| 1st Quartile | 19.12 |
| Median | 20.12 |
| Mean | 20.07 |
| 3rd Quartile | 21.16 |
| Max | 24.55 |

As you can see the maximum difference between the values obtained is 153.125. This value is insignificant, as it is positive and not negative. Another conclusion that can be made from these statistics is when plotting the points on a graph, the graph will be normal; as the data points are close together.

Chart, line chart

Description automatically generated

From this graph I can conclude that the values conform to a normal distribution with a few outliers, as the points are close to the centre. An alternative inference is that the graph is light skewed because that the curvature in the graph are outliers. Making it light skewed. This graph may not show an accurate representation as the dataset is relatively small (n = 200). The line of best fit suggests there is a positive correlation between the size of the rats.

Below is a boxplot for the generated rats’ dataset.

Chart

Description automatically generated

From the image above, I can see that my dataset contains many outliers, as the upper and lower quartile merge with the interquartile range (middle box). Although this is the case, the diagram illustrates a normal distribution with a positively skewed dataset, as the lower quartile is larger than the upper quartile. This suggests there is a variation with respect to the weights.

Task 2: Appropriateness for Hypothesis testing

In this section I will be proposing explanations I have made based on limited evidence. These observations can be used for further investigation. To do this I will be comparing the averages of my two datasets with the null hypothesis and alternative hypothesis. The methods I will be using for these tests are T-test statistics, degrees of freedom, p-values, confidence intervals and sample estimates. I will be gathering this data from graphs.

The null hypotheses and alternative hypotheses I will be using to carry out these tests can be found in the table below:

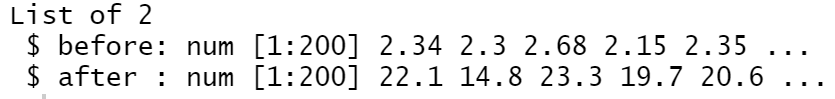
|  |  |  |  |
| --- | --- | --- | --- |
| Null hypotheses | Corresponding values | Alternative hypotheses | Corresponding values |
| H0: m = 0 | Equal to 0 | Ha: m ≠ 0 | Not equal to 0 |
| H0: m ≤ 0 | Less than or equal to 0 | Ha: m < 0 | Less than |
| H­0: m ≥ 0 | Greater than or equal to 0 | Ha: m > 0 | Greater than |

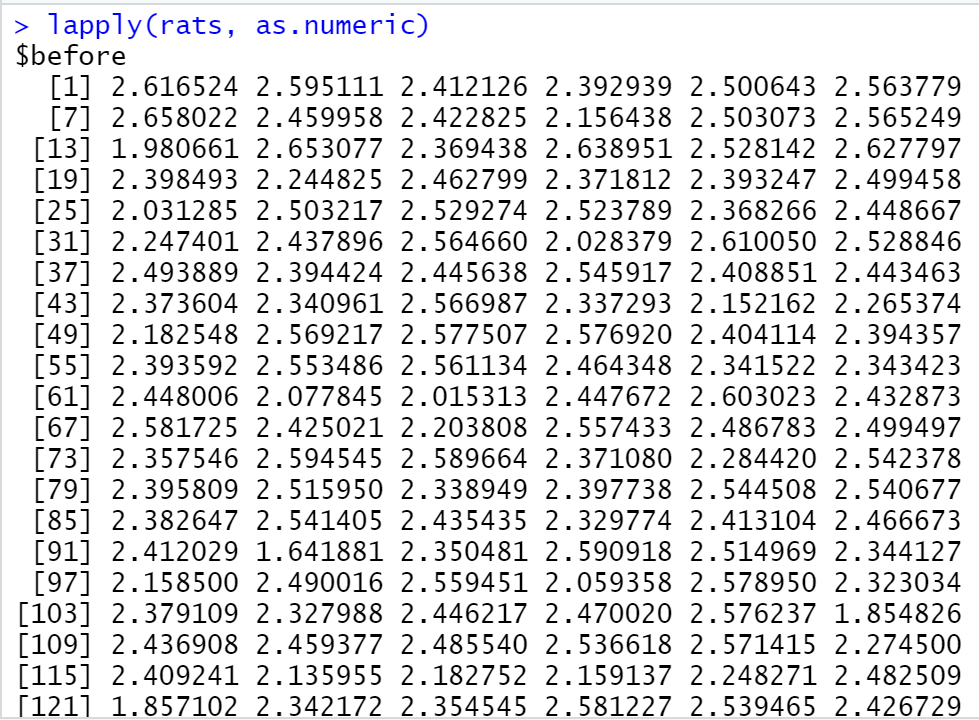
Firstly, I will be carrying out the Shapiro Wilk test on both datasets. I will be doing this to determine if they meet the null hypothesis. If this is false, my alternative hypothesis will be true. The Shapiro Wilk test is a test for normality. I will be using this test to identify if the datasets I am using came from a normal distribution. I will be performing this test using the R function: shapiro.test(x). This function takes the argument x as a parameter. X is a numeric vector containing data values.

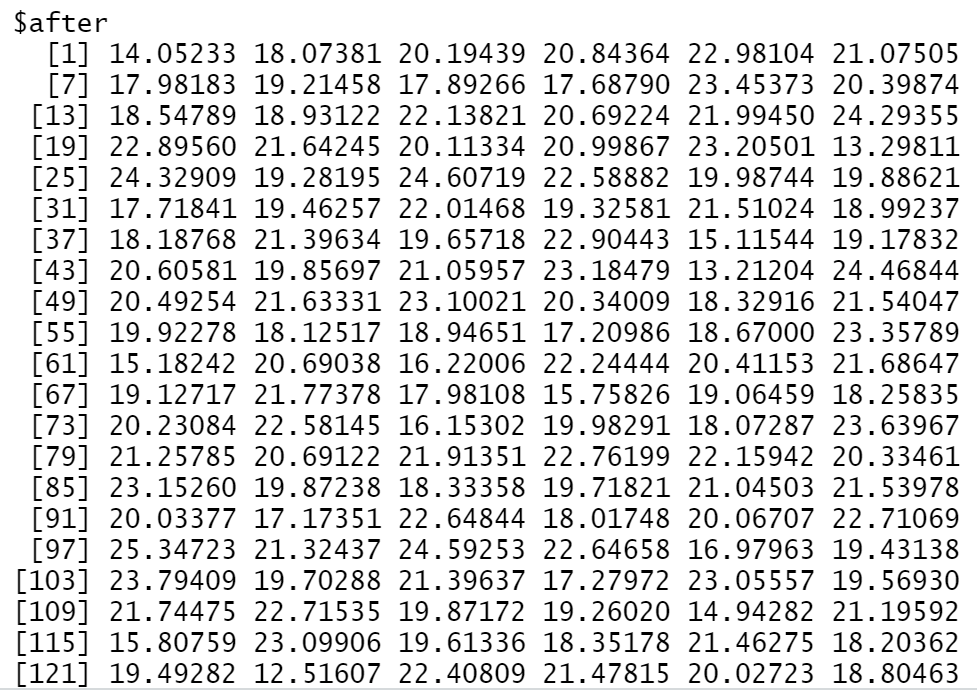
Before I perform the Shapiro Wilk test, I will be converting my datasets to numeric values, as this function requires numeric values, and not dataframes. I will do this using the sapply/lapply function. These functions are wrappers that apply a specified function over to datasets. I will be using the function in the form of:

Lapply(mice, as.numeric)

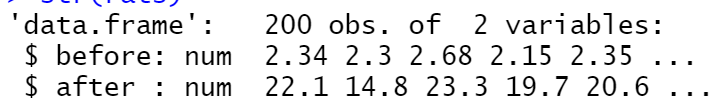
This converts the mice dataset to numeric values, as seen below:

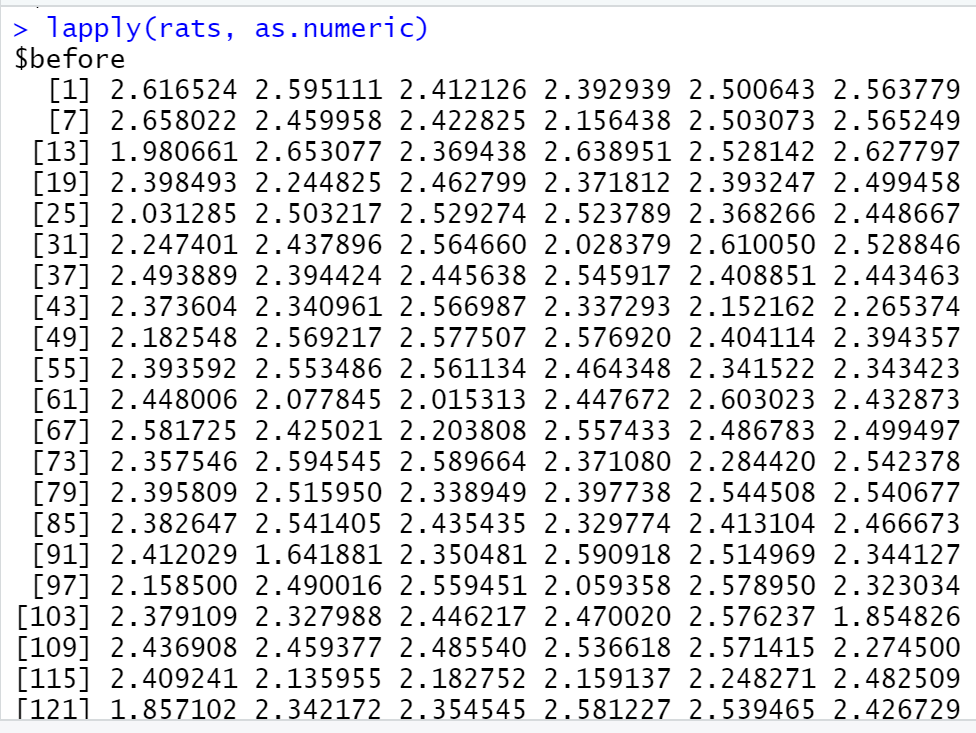






I will now do the same for the rats’ dataset. The result can be seen below:





Chart, line chart

Description automatically generated

Chart

Description automatically generated

## Comparing means and fitting distributions

* **Exercise 5 Data import/export in R**

**(Please make sure you first go through and complete the tutorial sheet “data frames”)**

**Q1**. Import the dataset from Internet and named it as *dat\_csv*. Dataset: <https://stats.idre.ucla.edu/stat/data/hsbraw.csv>.

Option 1 copy and paste R commands with comments as solution

# create a dataset called dat\_csv by loading a .csv dataset from Internet

> dat\_csv<-read.csv("https://stats.idre.ucla.edu/stat/data/hsbraw.csv")

Option 2 insert a screenshot of R code as solution





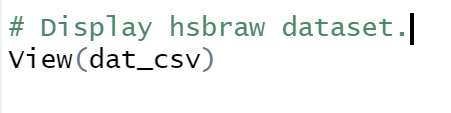
**Q2.1.** Either use the RStudio Environment pane or the View() function to view the imported dataset dat\_csv.

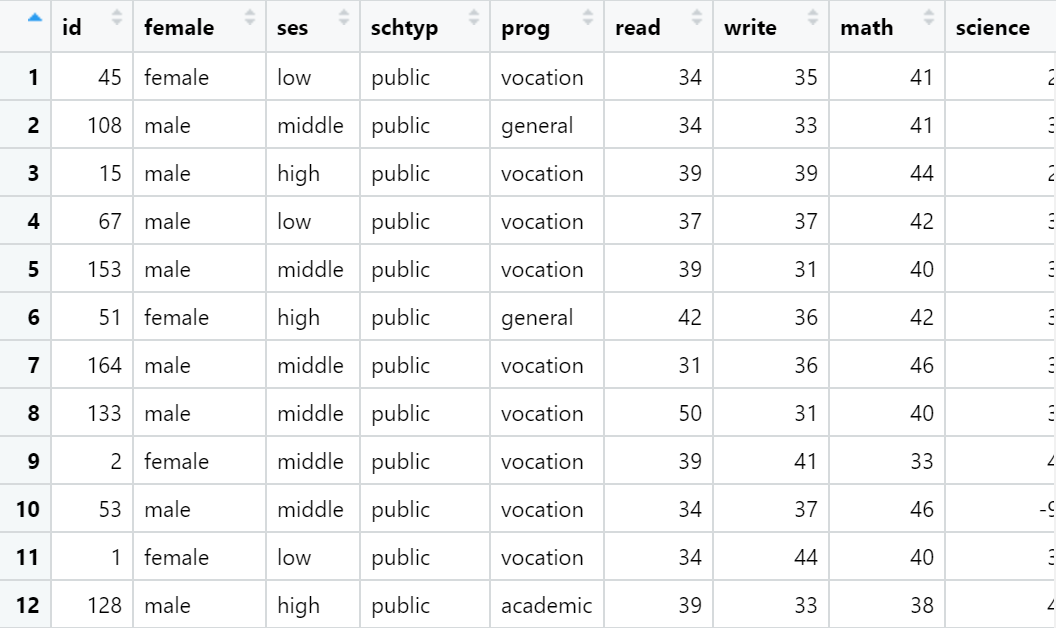
* How many variables in total and how many observations in total?

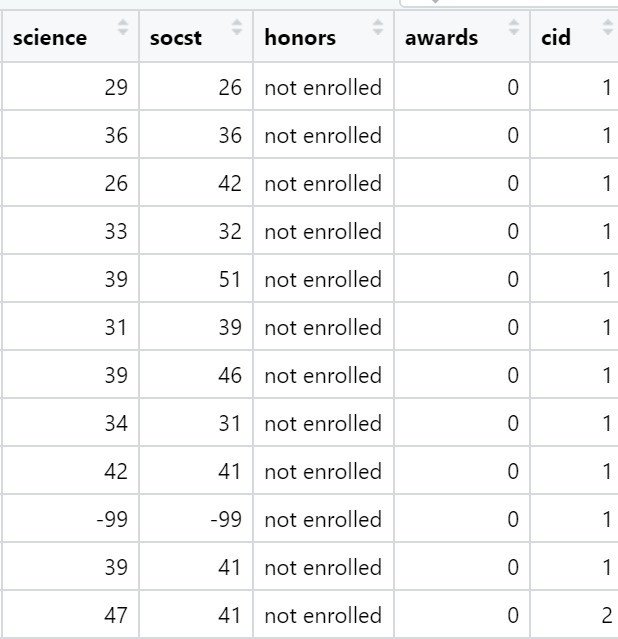
There are 13 variable and 200 observations in the hsbraw dataset.

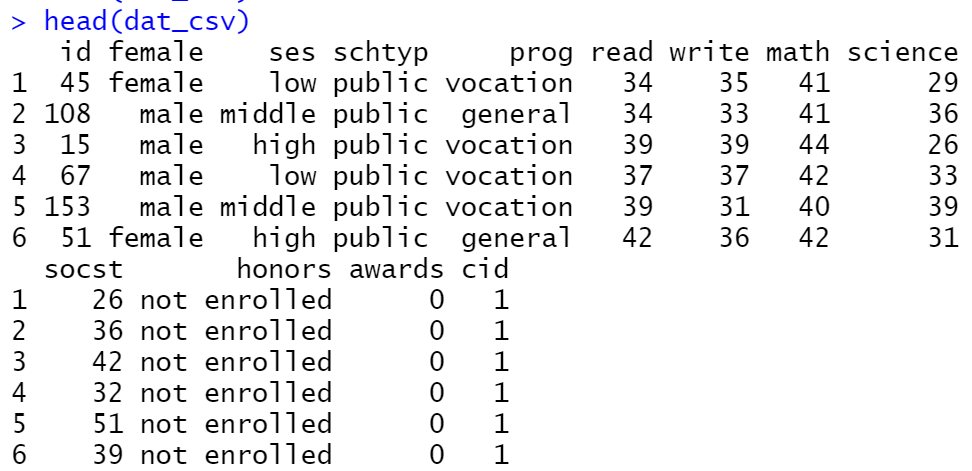
* Use help(function\_name) or ?function\_name to check details of functions head() and tail(),and then choose the right functions to view the first 3 rows and the last 3 rows, respectively.

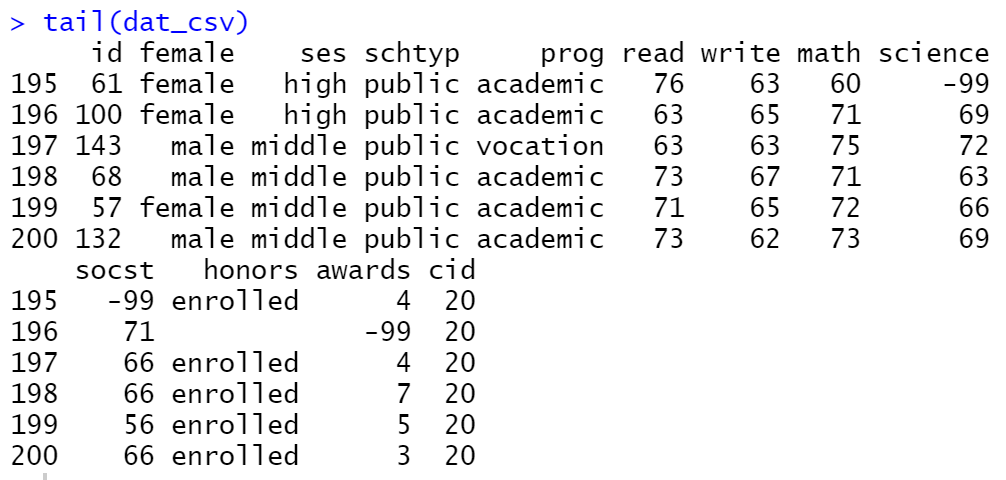
Copy and paste the R codes (with comments) used to perform the task here. Copy and paste what is printed in your R console after executing the codes. Instead, take a screenshot and make sure it is readable.











**Q2.2.** Search for the right function to examine the structure of dat\_csv.

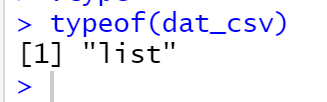
* What data structure is it (e.g., a numeric vector, a matrix, a list, a data frame, an array…)?

The hsbraw dataset is stored as a list.

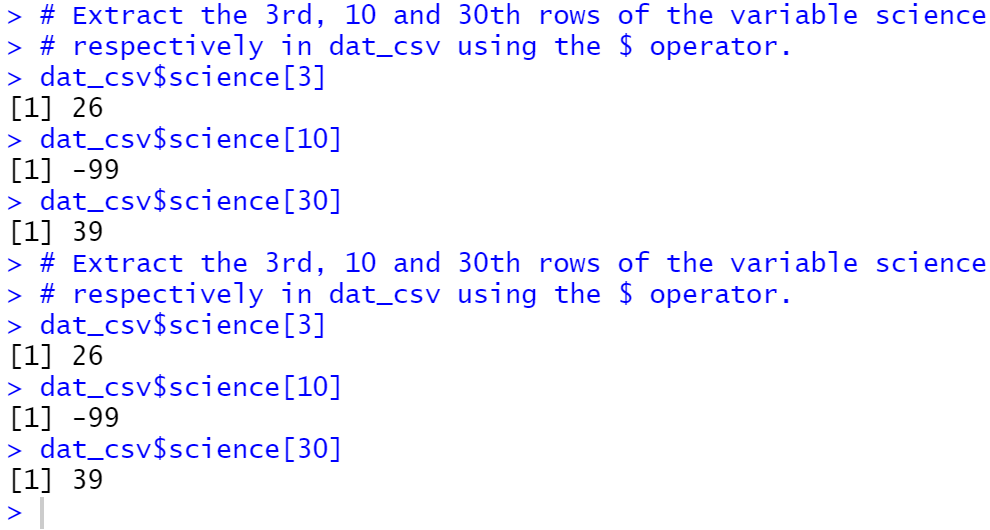
* How many variables of the dataset are factors? Provide details of the level information for each factor variable.

Copy and paste the R codes (with comments) used to perform the task. Copy and paste what is printed in your R console after executing the codes.

Instead, take a screenshot and make sure it is readable.



**Q3.** Extract the 3rd, 10th, and 30th rows of the variable science in dat\_csv using the $ operator. Copy and paste the R codes (with comments) used to perform the task here. Copy and paste what is printed in your R console after executing the codes. Instead, use screenshots and make sure it is readable.



**Q4.** Find theright function to view the summary information of dat\_csv.

* How many females? And how many males?

There are 78 male and 128 females.

* How many public schools?

There are 13 public shoos.

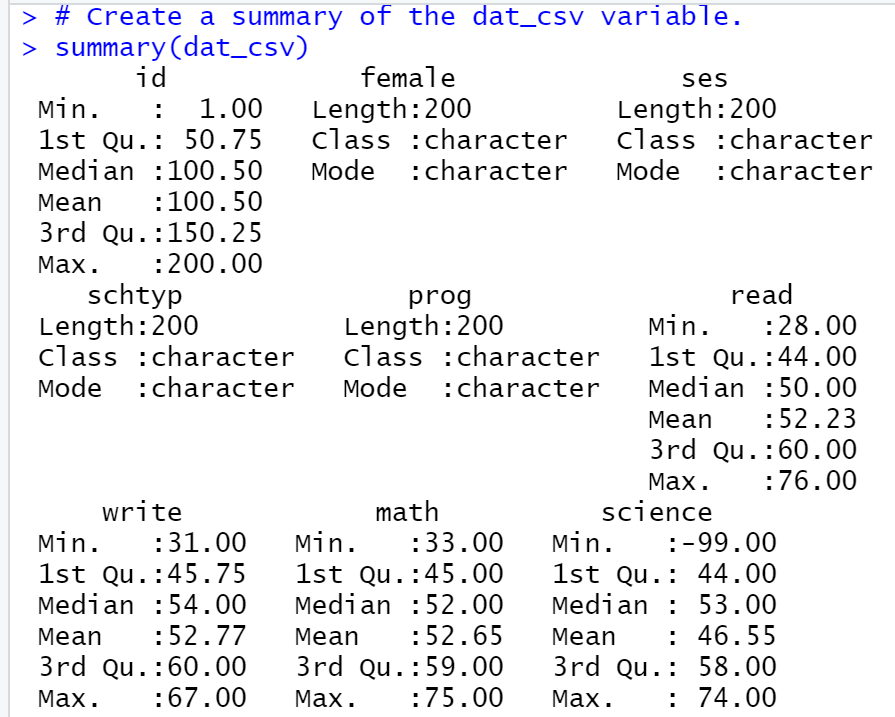
* What are the min/max/median of variable math?

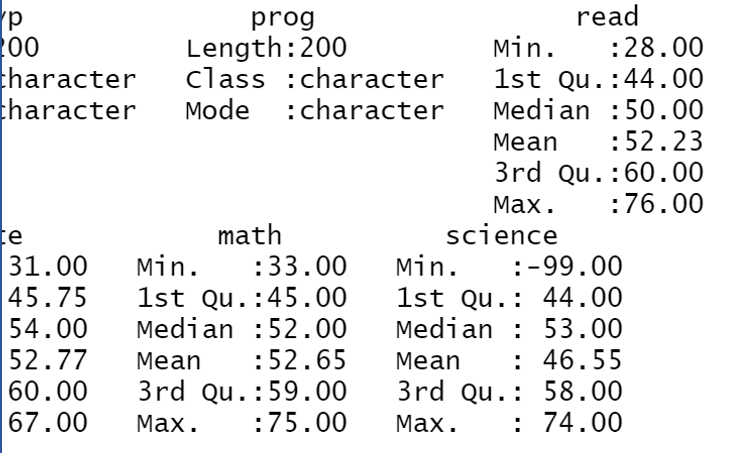
Minimum quantile: 33

75 (max) quantile: 59

Median: 52

Copy and paste your R commands and what is printed in your R console after executing the codes. Instead, use screenshots.





**Q5.1.** Create a new data set called test that is all rows of the 3 column variables math, read, and write from dat\_csv.

Copy and paste your R commands and what is printed in your R console after executing the codes. Instead, use screenshots. Don’t forget to add comments to make your codes understandable.

**Q5.2.** Add a new variable to the dataset test and name the new variable as test\_mean which is the mean of the variables math, read, and write.

Copy and paste your codes and what is printed in your R console after executing the codes. Instead, use screenshots. Don’t forget to add comments to make your codes understandable.

**Q2.1.** Either use the RStudio Environment pane or the View() function to view the imported dataset dat\_csv.

* How many variables in total and how many observations in total?
* Use help(function\_name) or ?function\_name to check details of functions head() and tail(),and then choose the right functions to view the first 3 rows and the last 3 rows, respectively.

Copy and paste the R codes (with comments) used to perform the task here. Copy and paste what is printed in your R console after executing the codes. Instead, take a screenshot and make sure it is readable.

**Q2.2.** Search for the right function to examine the structure of dat\_csv.

* What data structure is it (e.g., a numeric vector, a matrix, a list, a data frame, an array…)?
* How many variables of the dataset are factors? Provide details of the level information for each factor variable.

Copy and paste the R codes (with comments) used to perform the task. Copy and paste what is printed in your R console after executing the codes.

Instead, take a screenshot and make sure it is readable.

**Q3.** Extract the 3rd, 10th, and 30th rows of the variable science in dat\_csv using the $ operator. Copy and paste the R codes (with comments) used to perform the task here. Copy and paste what is printed in your R console after executing the codes. Instead, use screenshots and make sure it is readable.

**Q4.** Find theright function to view the summary information of dat\_csv.

* How many females? And how many males?
* How many public schools?
* What are the min/max/median of variable math?

Copy and paste your R commands and what is printed in your R console after executing the codes. Instead, use screenshots.

**Q5.1.** Create a new data set called test that is all rows of the 3 column variables math, read, and write from dat\_csv.

Copy and paste your R commands and what is printed in your R console after executing the codes. Instead, use screenshots. Don’t forget to add comments to make your codes understandable.

**Q5.2.** Add a new variable to the dataset test and name the new variable as test\_mean which is the mean of the variables math, read, and write.

Copy and paste your codes and what is printed in your R console after executing the codes. Instead, use screenshots. Don’t forget to add comments to make your codes understandable.

# Bibliography

Kassamba, A.K. (2020) Available athttps://CRAN.R-project.org/package=ggpubr:

Wickham, H.W., Francois, R.F., Henry, L.H., Muller, K.M. (2021) ‘*dlpyr: A Grammar of Data*

*ManipulationI’* (1.0.3) [R DPLYR] Available at: https://cran.r-project.org/web/packages/dplyr/index.html

Appendix

R code for coursework

# Install relevant packages.

install.packages("dplyr")

install.packages("DT")

install.packages("ggplot2")

install.packages("ggpubr")

install.packages("fitdistrplus")

install.packages("tidyverse")

# Load relevant packages.

library(dplyr)

library(ggplot2)

library(ggpubr)

library(DT)

library(ggplot2)

library(fitdistrplus)

library(tidyverse)

# Task 1: Data generation

# Generate dataset consisting of the weights of mice before

# and after treatment

# with the parameters: mean = 20 and variance = 2.

weight\_before < c(rnorm(n = 200, mean = 20, sd = sqrt(2)))

weight\_after <- c(rnorm(n = 200, mean = 20, sd = sqrt(2.5)))

#weight\_before <- rnorm(n = 200, mean = 20, sd = sqrt(2))

#weight\_after <- rnorm(n = 200, mean = 21, sd = sqrt(2.5))

#weight\_before

#weight\_after

# combine the data frames for mice.

mice <- cbind(weight\_before, weight\_after)

mice

# Generate a dataset containing the weights of rats before and after treatment

# with mean = 21 and variance = 2.5

rats <- data.frame(before = weight\_before, after = weight\_after)

rats <- data.frame(weight\_before = c(rweibull(n = 200,shape = 10, scale = 20)))

weight\_before <- c(rweibull(n = 200, shape = 21, scale = 2.5))

weight\_after <- c(rweibull(n = 200, shape = 9, scale = 21))

rats <- cbind(weight\_before, weight\_after)

mice <- data.frame(before = weight\_before, after = weight\_after)

mice

mice$weight\_before

rats <- data.frame(before = weight\_before, after = weight\_after)

mice$weight\_after

rats

# Convert mice dataset from numeric to a factor variable.

mice$weight\_before < as.factor(mice$weight\_before)

plot(mice)

#rats <- ggplot(data=mice, aes(x=weight\_after)) +

# geom\_qqplot(aes(x = weight\_before, y = weight\_after))

class(mice)

class(rats)

dim(mice)

dim(rats)

# Generate a qq plot and box plot with data from the mice dataset.

ggplot(mice, aes(sample = weight\_before)) +

stat\_qq() + stat\_qq\_line()

ggplot(mice, aes(x = weight\_before, y = weight\_after)) +

stat\_boxplot()

# Generate a box plot with data from the rats dataset.

ggplot(rats, aes(x = weight\_before, y = weight\_after)) +

stat\_boxplot()

#geom\_qq(

# geom = "density")

str(mice)

str(rats)

summary(mice)

summary(rats)

head(mice)

write.table(mice)

gather("weight\_before", "weight\_after")

#unite("weight\_before", "weight\_after")

View(mice)

#ggpubr::ggqqplot(mice$weight\_before)

mice <- data.frame(before = weight\_before, after = weight\_after)

mice

mice$weight\_before

mice$weight\_after

mice <- lapply(mice, as.numeric)

str(mice)

lapply(rats, as.numeric)

str(rats)

# Perform shapiro wilk test on rat and mice dataset.

shapiro.test(rats)

summary(rats)

summary(mice)