QRM II Graded Assignment (11), Period 1 2025

Material by Sjoerd van Alten and Klervie Toczé

Fill in your group number and names here

25-09-2025

# Introduction

This assignment is to be completed in groups of 3-4. Further, all students in your group need to be assigned to the same R tutorial group (Friday’s tutorial). You can sign yourself up for a group on Canvas. Please do so You can use the Discussion Board in Canvas if you do not have a group yet or if your group is incomplete.

The assignment has 5 parts, and each part corresponds to the course material of that week (with the exclusion of week 6, for which there is no R programming material).

You are supposed to hand in these assignments on Canvas at the following dates:

* **Deadline 1** *Thursday September 25th, at 23:59pm*: you are supposed to hand in weeks 1, 2, and 3 of this assignment. This will determine 18% of your overall course grade
* **Deadline 2** *Thursday October 9th, at 23:59pm*: you are supposed to hand in weeks 4, and 5 of this assignment. This will determine 12% of your overall course grade

The R tutorials (each Friday) will consist of two halves. During the first half, you will discuss the tutorial exercises. These can be downloaded separately from Canvas. During the second half, you can work on this graded assignment within your own group. The purpose is that you find out how to work with R for doing statistical analyses by yourself. The tutorial exercises are meant to teach you basic commands to get you started, but to answer the problem sets in this assignment, you might need to research your own solutions, and use functions and commands not described in the tutorial exercises. Learning how to solve your own research problems is integral part of learning R. When you and your group get stuck on how to approach an exercise, the hierarchy in finding your way is as follows:

* use the concepts from the tutorial exercises;
* use the cheat sheets available on Canvas;
* use Google, YouTube, StackOverflow, or another website;
* ask the teacher.

The use of generative AI is permitted and may result in a grade of 0. See the AI protocol in the course manual for details.

To answer the assignment, you can simply fill out this R markdown document. There are designated places which you can fill with R code. There are also designated spaces for you to answer each question. Often, the structure of an answer will be as follows. First, you type the R code in the designated box. This will show how you analyzed the data to get the answer to the question. Below the box for the R code, you will then summarize your answer to the question, i.e. what are the conclusions that you draw from the data analysis?

When handing in, you are supposed to submit this .Rmd file, and a knitted version of this document. You can knit this document to pdf, word, or html. Knitting to pdf requires you to have a .tex distribution installed on your computer. Knitting to Word requires you to have Word installed.

The exercises are designed such that you should be able to finish the majority of them during the tutorial each week. If you are not able to finish them fully during that time, you are expected to work on it in your own time using the computers on campus or your own device. It is best to meet as a group in-person when working together. If you want to work remotely, github is a good platform to guarantee smooth collaboration. Alternatively, you can email this .Rmd file back and forth to one another as a group, but this is not recommended as it is more cumbersome.

We encourage you to keep your code blocks, printing statements, and final answers, as short as possible. In any case, there is a page limit of 6 pages per week, which encompasses the total length of this document which consists of the questions, your coding lines, and your answers. When your answers to questions of the respective week exceed this page limit, they will not be graded, resulting in zero points.

Each week consists of 1, 2, or 3 subquestions. The total amount of points you can earn per week is 20 points.

# Week 1

1. Find the dataset “movies4.tsv” on Canvas. Describe your data set: How many observations does it have. How many variables are there? How many subjects? What consists of a subject?

## Warning: package 'readr' was built under R version 4.5.1

## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──  
## ✔ dplyr 1.1.4 ✔ readr 2.1.5  
## ✔ forcats 1.0.0 ✔ stringr 1.5.1  
## ✔ ggplot2 3.5.2 ✔ tibble 3.2.1  
## ✔ lubridate 1.9.4 ✔ tidyr 1.3.1  
## ✔ purrr 1.0.4   
## ── Conflicts ────────────────────────────────────────── tidyverse\_conflicts() ──  
## ✖ dplyr::filter() masks stats::filter()  
## ✖ dplyr::lag() masks stats::lag()  
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors  
## Rows: 808 Columns: 19  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: "\t"  
## chr (8): keywords, original\_language, title, genre, first\_actor, first\_act...  
## dbl (10): index, budget, popularity, revenue, runtime, vote\_average, vote\_c...  
## date (1): release\_date  
##   
## 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.

## # A tibble: 808 × 2  
## index n  
## <dbl> <int>  
## 1 1 1  
## 2 7 1  
## 3 16 1  
## 4 20 1  
## 5 26 1  
## 6 29 1  
## 7 33 1  
## 8 37 1  
## 9 42 1  
## 10 46 1  
## # ℹ 798 more rows

## # A tibble: 808 × 19  
## index budget keywords original\_language title popularity release\_date revenue  
## <dbl> <dbl> <chr> <chr> <chr> <dbl> <date> <dbl>  
## 1 2256 0 <NA> en The … 0.286 2012-08-29 0   
## 2 2473 1.5 e7 barcelo… en Vick… 32.8 2008-08-15 9.64e7  
## 3 1514 3.20e7 gunslin… en The … 16.5 1995-02-09 1.86e7  
## 4 1512 3.20e7 robbery… en A Hi… 34.6 2005-09-23 6.07e7  
## 5 395 8.5 e7 holiday… en The … 45.0 2006-12-08 1.94e8  
## 6 3156 0 1970s h… en Red … 5.50 2011-08-04 0   
## 7 4753 6 e4 haunted… en Hayr… 0.412 2012-10-13 0   
## 8 4590 0 kidnapp… en Show… 0.231 2004-09-23 0   
## 9 1665 0 venice … en The … 12.5 2004-09-03 0   
## 10 2714 1.4 e7 new yor… en Marg… 4.90 2011-09-30 4.65e4  
## # ℹ 798 more rows  
## # ℹ 11 more variables: runtime <dbl>, vote\_average <dbl>, vote\_count <dbl>,  
## # genre <chr>, release\_year <dbl>, release\_month <dbl>, release\_day <dbl>,  
## # first\_actor <chr>, first\_actor\_gender <chr>, director\_first\_name <chr>,  
## # director\_gender <chr>

**Your Answer:**

Write your response here.

1. Which of the following types of variables are present in your data set? (i) nominal; (ii) ordinal; (iii); interval; (iv) ratio. If present, name one example of such a variable present in your data set.

#Nominal variables: keywords, original\_language, title, genre, first\_actor, first\_actor\_gender, director\_first\_name, director\_gender  
  
#Ordinal variables: NA  
  
#Interval variables: index, release\_date, release\_year, release\_month, release\_day  
  
#Ratio variables: budget, popularity, revenue, runtime, vote\_average, vote\_count

**Your Answer:** Nominal variables: keywords, original\_language, title, genre, first\_actor, first\_actor\_gender, director\_first\_name, director\_gender

Ordinal variables: NA

Interval variables: index, release\_date, release\_year, release\_month, release\_day

Ratio variables: budget, popularity, revenue, runtime, vote\_average, vote\_count

1. A movie studio wants to know which types of movies give maximal profit. Perform the following steps to provide the movie studio with an analysis which corresponds to their request:
2. Create the variable profits as the revenue of a movie minus its budget. Report its mean, median, maximum, and minimum.
3. Which movie has the highest profits in your data set and how much are these profits. Which movie has the lowest and how much are its profits? If multiple movies have the exact same highest or lowest profits, give only one example.
4. Create a boxplot of the variable profits. Make sure it has an appropriate title, and appropriate titles and labels for the x- and y-axis. Give Q1, Q2, Q3, and Q4. What does this tell you about the nature of making money in the movies industry?
5. Add a new variable to your data set the log of profits. When creating this variable, what happens to movies for which profits is zero or negative? What then happens when you calculate the mean of log of profits?
6. For movies that have a profit of zero or less, replace log of profits with “NA”. What is now the mean of log of profits? Create a boxplot for log of profits, again with an appropriate title, x- and y-axis labels. How does it compare to the boxplot you made under c.)?
7. Create a scatterplot of with the runtime of movies on the x-axis and the average vote of movies on the y-axis. What do you conclude from the scatterplot? Are movies with a longer runtime considered worse or better by the audience, or does the audience not have a preference? Why do you think this is the case?

For each step, you should provide first all the code you used to answer the question and then formulate an answer using full sentences.

*Step a*

## Rows: 808 Columns: 19  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: "\t"  
## chr (8): keywords, original\_language, title, genre, first\_actor, first\_act...  
## dbl (10): index, budget, popularity, revenue, runtime, vote\_average, vote\_c...  
## date (1): release\_date  
##   
## 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.

## # A tibble: 1 × 4  
## mean\_profit median\_profit max\_profit min\_profit  
## <dbl> <dbl> <dbl> <dbl>  
## 1 52885669. 0 1299557910 -150000000

**Your Answer:**

mean\_profit = 52885669 median\_profit = 0 max\_profit = 1299557910 min\_profit = -1.5e+08

*Step b*

## Rows: 808 Columns: 19  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: "\t"  
## chr (8): keywords, original\_language, title, genre, first\_actor, first\_act...  
## dbl (10): index, budget, popularity, revenue, runtime, vote\_average, vote\_c...  
## date (1): release\_date  
##   
## 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.

## # A tibble: 1 × 2  
## title profits  
## <chr> <dbl>  
## 1 The Avengers 1299557910

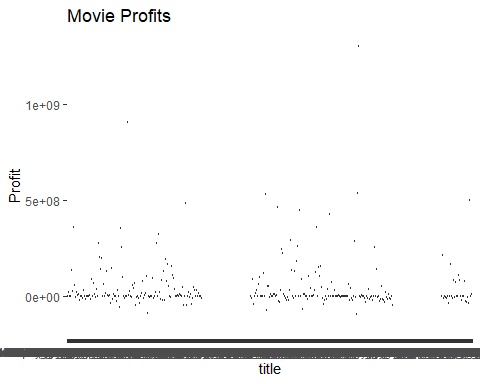
## # A tibble: 1 × 2  
## title profits  
## <chr> <dbl>  
## 1 The Wolfman -150000000

**Your Answer:**

Highest profits: The Avengers, profits=1299557910 Lowest profits: The wolfman, profits=-1.5e+08 (-$150 million)

*Step c*

library (ggplot2)  
  
boxplot\_profits<-ggplot(movies4, aes(x = title, y = profits)) +  
 geom\_boxplot() +  
 labs(title="Movie Profits",  
 y="Profit")  
boxplot\_profits



quartiles = quantile(movies4$profits)  
quartiles

## 0% 25% 50% 75% 100%   
## -150000000 -1849854 0 49711134 1299557910

**Your Answer:**

Q1: -1849854 Q2: 0 Q3: 49711134 Q4: 12999557910

This shows that making money in the movies industry is risky, as the bottom 50% is in loss or break even, while the top 25% is in massive profits.

*Step d*

**Your Answer:**

Where profit is 0, the log is -inf.

Where profit is negative, the log is NaN

The mean results in NaN, and the mean with na.rm=T results in -inf

*Step e*

movies4=read\_tsv("movies4.tsv")

## Rows: 808 Columns: 19  
## ── Column specification ────────────────────────────────────────────────────────  
## Delimiter: "\t"  
## chr (8): keywords, original\_language, title, genre, first\_actor, first\_act...  
## dbl (10): index, budget, popularity, revenue, runtime, vote\_average, vote\_c...  
## date (1): release\_date  
##   
## 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.

movies4=movies4%>%  
 mutate(profits=revenue-budget)%>%  
 mutate(log\_profits = ifelse(profits <= 0, NA, log(profits)))

## Warning: There was 1 warning in `mutate()`.  
## ℹ In argument: `log\_profits = ifelse(profits <= 0, NA, log(profits))`.  
## Caused by warning in `log()`:  
## ! NaNs produced

movies4

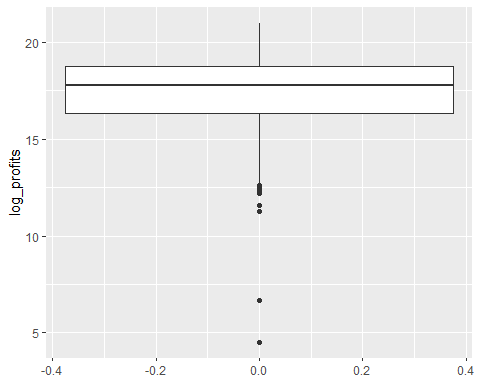
## # A tibble: 808 × 21  
## index budget keywords original\_language title popularity release\_date revenue  
## <dbl> <dbl> <chr> <chr> <chr> <dbl> <date> <dbl>  
## 1 2256 0 <NA> en The … 0.286 2012-08-29 0   
## 2 2473 1.5 e7 barcelo… en Vick… 32.8 2008-08-15 9.64e7  
## 3 1514 3.20e7 gunslin… en The … 16.5 1995-02-09 1.86e7  
## 4 1512 3.20e7 robbery… en A Hi… 34.6 2005-09-23 6.07e7  
## 5 395 8.5 e7 holiday… en The … 45.0 2006-12-08 1.94e8  
## 6 3156 0 1970s h… en Red … 5.50 2011-08-04 0   
## 7 4753 6 e4 haunted… en Hayr… 0.412 2012-10-13 0   
## 8 4590 0 kidnapp… en Show… 0.231 2004-09-23 0   
## 9 1665 0 venice … en The … 12.5 2004-09-03 0   
## 10 2714 1.4 e7 new yor… en Marg… 4.90 2011-09-30 4.65e4  
## # ℹ 798 more rows  
## # ℹ 13 more variables: runtime <dbl>, vote\_average <dbl>, vote\_count <dbl>,  
## # genre <chr>, release\_year <dbl>, release\_month <dbl>, release\_day <dbl>,  
## # first\_actor <chr>, first\_actor\_gender <chr>, director\_first\_name <chr>,  
## # director\_gender <chr>, profits <dbl>, log\_profits <dbl>

movies4%>%  
 summarise(mean(log\_profits, na.rm=T))

## # A tibble: 1 × 1  
## `mean(log\_profits, na.rm = T)`  
## <dbl>  
## 1 17.3

ggplot(movies4, aes(y=log\_profits)) +  
 geom\_boxplot()

## Warning: Removed 407 rows containing non-finite outside the scale range  
## (`stat\_boxplot()`).



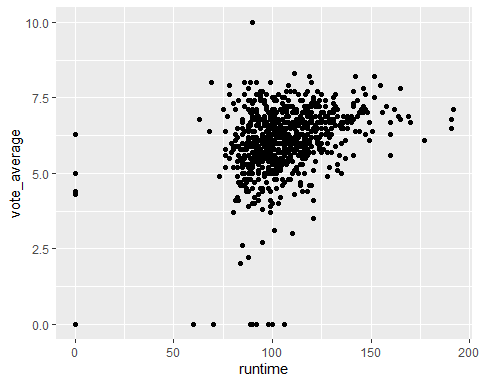
**Your Answer:**

Mean of log of profits = 17.33131

This boxplot seems to have less extreme outliers.

*Step f*

ggplot(movies4, aes(x=runtime, y=vote\_average)) +  
 geom\_point()



**Your Answer:**

There seems to be a slight correlation between runtime and higher average votes, as the points at the relatively shorter runtime of 100 are more dense around the 5.0 to 7.0 ratings, while the point at the relatively longer runtime of 125 to 150 minutes seem to be more dense around the 6.5 to 7.5 range.

# Week 2

1 Is your dataset movies4.tsv the full population, or is it a sample of a larger population? If the latter, how would you describe the full population?

The dataset is a sample of a larger population. It does not represent the full population of movies. The full population would consist of all movies released worldwide, covering a wide variety of genres, budgets, release years, and popularity levels.

2

1. For which actor in your data set do you observe the most movies?
2. What is the average revenue of the movie in which this actor plays and does the revenue lie above or below the revenue of an average movie according to your data set?
3. How trustworthy do you consider your conclusion to answer 2b? Use the term “law of large numbers” in your explanation.

*step a*

#WRITE YOUR CODE HERE  
library(tidyverse)  
  
y<- movies4 %>%  
 group\_by(first\_actor)%>%  
 count(first\_actor)%>%  
 arrange(desc(n))  
  
head(y,2)

## # A tibble: 2 × 2  
## # Groups: first\_actor [2]  
## first\_actor n  
## <chr> <int>  
## 1 Bruce Willis 9  
## 2 Denzel Washington 8

**Your Answer:** Bruce Willis is the actor who plays in the most movies in the data set. We pulled the top 2 rows to confirm that there is not any other actor with the same amount of movies.

*step b*

#WRITE YOUR CODE HERE  
z<- movies4%>%  
 group\_by(first\_actor)%>%  
 summarise(mean\_r = mean(revenue, na.rm = TRUE))  
z = slice(z, -(1:64))  
head(z,1)

## # A tibble: 1 × 2  
## first\_actor mean\_r  
## <chr> <dbl>  
## 1 Bruce Willis 255652124.

mean\_revenue = mean(movies4$revenue)  
mean\_revenue

## [1] 82727885

The avarage revenue of Bruce Willis is 255652124 this lies above the average revenue of 82727855

*step c*

mean(y$n)

## [1] 1.453237

According to the law of large numbers, a higher number of observations leads to a more accurate estimate. Since Bruce Willis has a relatively high number of observations (9 movies, compared to an average of 1.45 movies for other actors), our conclusion from 2b can be considered quite trustworthy.

3 For this question, you will assume that your data set is the full population.

1. Recode profits such that it is expressed in millions. What is the variance of the variable profits (in millions) in your data set?
2. Create a new data set, called movies\_sample. Make sure that it is a random sample of your data set of 25 movies. What is the variance of profits in this random sample? How does it compare to the variance of profits in 2a?
3. In a for loop, create 100 different samples of 25 movies, as in b, and estimate the variance within each sample. Save the variance of each sample in a vector called sample\_vars. So the first position of the vector would have the variance of the first sample, the second position the variance of the second sample, etc. Print the start of this vector.
4. Summarize and make a histogram of sample\_vars. What is the mean, standard deviation and shape of its distribution?
5. In your opinion, is a sample of 25 movies sufficient to get a reliable estimate of the population variance of profits, using the sample variance? Explain?

*step a*

library(tidyverse)  
  
movies4 <- movies4 %>%  
 mutate(profit\_in\_millions = profits / 1000000)  
  
variance <- var(movies4$profit\_in\_millions)  
  
variance

## [1] 20270.56

**Your Answer:**

The variance of the profit in millions in the data set is 20270.56 million.

*step b*

set.seed(56564)  
movies\_sample <- movies4[sample(nrow(movies4), 25), ]  
  
variance\_sample <- var(movies\_sample$profit\_in\_millions)  
  
variance\_sample

## [1] 29823.65

variance

## [1] 20270.56

**Your Answer:**

The variance of the sample is significantly higher than in the original data set. This because the random sample is less precise.

*step c*

set.seed(293)  
sample\_vars<- c()  
for(i in 1:100){  
 mov\_samp<- slice\_sample(movies4, n=25)  
 sample\_vars[i]<- var(mov\_samp$profit\_in\_millions)  
   
}  
print(sample\_vars[1:9])

## [1] 37056.903 9338.942 2122.148 69705.140 11612.124 24541.473 49198.948  
## [8] 7123.823 3566.988

min(sample\_vars)

## [1] 1598.791

max(sample\_vars)

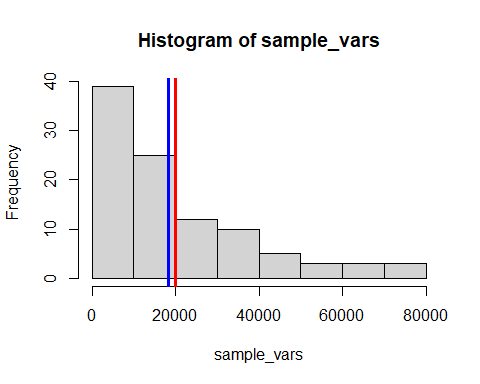
## [1] 73529.14

**Your Answer:**

In the different variances of the samples we took you can see a very big difference, the lowest value in the vector is 1598.8 million and the highest variance is 73529.1 million.

*step d*

hist(sample\_vars)  
abline(v=mean(sample\_vars),  
 col = "red",  
 lwd = 3)  
abline(v=sd(sample\_vars),  
 col="blue",  
 lwd = 3)



cat("mean = ", mean(sample\_vars), "\n")

## mean = 19923.88

cat("standard deviation =",sd(sample\_vars))

## standard deviation = 18251.93

mean(sample\_vars)

## [1] 19923.88

sd(sample\_vars)

## [1] 18251.93

**Your Answer:**

In the histogram you can see that the variances of the samples are not normal distributed, and that there are more low values than higher ones. This principles is called right-skewed, meaning that most of the variances are low and a few variances have a way higher variance. The mean of the variances is: 19923.88 million, and the standard deviation is 18251.93 million

*step e*

**Your Answer:**

Since the variances vary a lot, this shows that estimates depend greatly on which movies are chosen. so a sample size of 25 is not enough to get reliable estimates

**Your answer here**

# Week 3

For the next part of the assignment, assume that the movies in your data frame are a random sample of a larger population of movies.

1

1. Create a new data set that only includes movies that are of the genre “Thriller”. For these thriller movies, give a 99 percent confidence interval for the variable *runtime*. Interpret the result.
2. Now, assume that the variance of *runtime* amongst thriller movies in your data is exactly the same as the variance of *runtime* in the population. Under this assumption, give a 99 percent confidence interval for the variable *runtime* among thriller movies. Interpret the result. Is you confidence interval wider or less wide than the one you found under question 1a? Why is that the case?

*step a*

library(tidyverse)  
library(readr)  
movies4

## # A tibble: 808 × 22  
## index budget keywords original\_language title popularity release\_date revenue  
## <dbl> <dbl> <chr> <chr> <chr> <dbl> <date> <dbl>  
## 1 2256 0 <NA> en The … 0.286 2012-08-29 0   
## 2 2473 1.5 e7 barcelo… en Vick… 32.8 2008-08-15 9.64e7  
## 3 1514 3.20e7 gunslin… en The … 16.5 1995-02-09 1.86e7  
## 4 1512 3.20e7 robbery… en A Hi… 34.6 2005-09-23 6.07e7  
## 5 395 8.5 e7 holiday… en The … 45.0 2006-12-08 1.94e8  
## 6 3156 0 1970s h… en Red … 5.50 2011-08-04 0   
## 7 4753 6 e4 haunted… en Hayr… 0.412 2012-10-13 0   
## 8 4590 0 kidnapp… en Show… 0.231 2004-09-23 0   
## 9 1665 0 venice … en The … 12.5 2004-09-03 0   
## 10 2714 1.4 e7 new yor… en Marg… 4.90 2011-09-30 4.65e4  
## # ℹ 798 more rows  
## # ℹ 14 more variables: runtime <dbl>, vote\_average <dbl>, vote\_count <dbl>,  
## # genre <chr>, release\_year <dbl>, release\_month <dbl>, release\_day <dbl>,  
## # first\_actor <chr>, first\_actor\_gender <chr>, director\_first\_name <chr>,  
## # director\_gender <chr>, profits <dbl>, log\_profits <dbl>,  
## # profit\_in\_millions <dbl>

movies\_thriller <- movies4 %>%  
 filter(genre=="Thriller")%>%  
 summarise(  
 n=n(),  
 mean\_runtime=mean(runtime, na.rm=T),  
 sd\_runtime=sd(runtime, na.rm=T),  
 se=sd\_runtime/sqrt(n()),  
 t\_critical=qt(0.995, df=n-1),  
 ci\_lower=mean\_runtime-t\_critical\*se,  
 ci\_upper=mean\_runtime+t\_critical\*se)%>%  
 print()

## # A tibble: 1 × 7  
## n mean\_runtime sd\_runtime se t\_critical ci\_lower ci\_upper  
## <int> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 116 105. 19.2 1.78 2.62 101. 110.

**Your Answer:**

According to this 99% confidence interval, the runtime of the movies in the Thriller genre is reasonably consistent, as the confidence interval is pretty narrow, going from 100.55 to 109.90.

*step b*

movies4 <- movies4%>%  
 mutate(runtime = ifelse(runtime == 0, NA, runtime))  
  
movies\_thrillerb <- movies4 %>%  
 filter(genre == "Thriller")  
  
mean\_runtime\_thriller = mean(movies\_thrillerb$runtime, na.rm = T)  
n = sum(!is.na(movies\_thrillerb$runtime))  
sd\_thriller = sd(movies\_thrillerb$runtime, na.rm = T)  
se\_thriller = sd\_thriller/ sqrt(n)  
z\_value = qnorm(0.995)  
  
lowerbound = mean\_runtime\_thriller-z\_value \* se\_thriller  
upperbound = mean\_runtime\_thriller+z\_value \* se\_thriller  
print(c(lowerbound,upperbound))

## [1] 102.1602 110.1180

data.frame(  
 Methode = c("a","b"),  
 Lower = c(movies\_thriller$ci\_lower, lowerbound),  
 Upper = c(movies\_thriller$ci\_upper, upperbound)  
)

## Methode Lower Upper  
## 1 a 100.5517 109.8965  
## 2 b 102.1602 110.1180

interval\_a = movies\_thriller$ci\_upper - movies\_thriller$ci\_lower  
interval\_b = upperbound-lowerbound  
  
interval\_a > interval\_b

## [1] TRUE

**Your Answer:**

When runtime is 0 in the dataset this is actually a NA, because a movie can not have 0 run time. When taking this into account we come to the conclusion that the interval of a [100.55,109.90] is bigger than the interval of b [102.16,110.12].

2

1. Using an appropriate five-step procedure, set up a test for the null hypothesis that the variance of runtime equals . Clearly state your null hypothesis, alternative hypothesis your test statistic, your critical value, and your conclusion.
2. For the validity of your test in 2a, what assumption about the distribution of revenue needs to hold? Make an appropriate plot to test this assumption. What do you conclude?

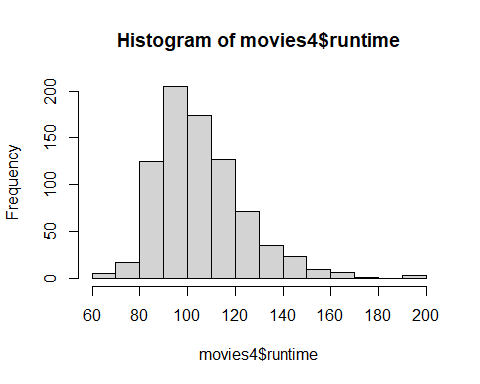
*step a*

sample\_var = var(movies4$runtime, na.rm = T)  
n <- sum(!is.na(movies4$runtime))   
H0 = 500  
chi\_square = ((n-1)\*sample\_var)/H0  
a = 0.05  
critl = qchisq(a/2, df=n-1)  
crith = qchisq(1-a/2, df = n-1)  
if (chi\_square<critl|chi\_square>crith){  
 conc = "reject H0"  
}else{  
 conc = "accept H0"  
}  
conc

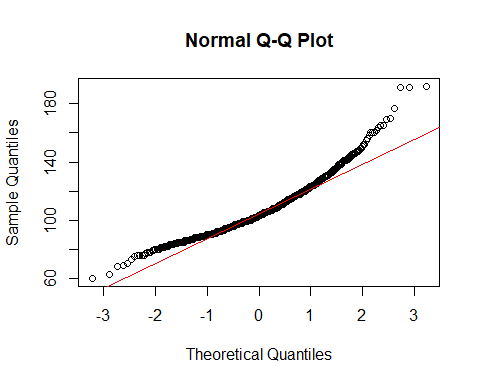
## [1] "reject H0"

**Your Answer:** step 1: cleary state H0 and HA step 2: computation of the sample variance and length step 3: identify the test statistics step 4: create critical values step 5: conclude H0, the variance of the runtime is 500 HA, th variance of the runtime in not 500 test statistics are: H0 =500 and chi\_square = ((n-1)*sample\_var)/H0 Conclusion: reject H0, since the variance is significantly different from 500* step b\*

hist(movies4$runtime)



qqnorm(movies4$runtime)  
qqline(movies4$runtime, col = "red")



**Your Answer:** For the chi square test to be valid, the runtime must be normally distributed.

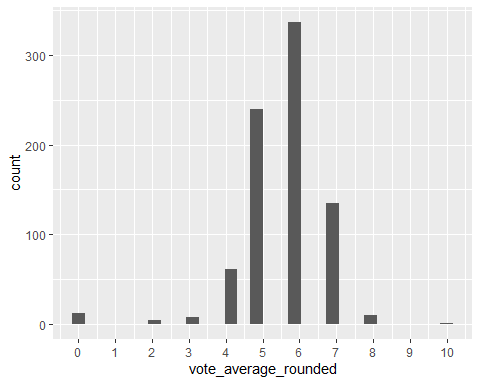
Based on the Plot, the runtime is not normally distributed, as there is a slight left- and right tail deviation and an overall S-shape. This means that the chi square test is not valid.

1. There is an argument going on in the movie studio. *Bob* claims that they should make higher-quality movies, as this will bring in more profits. *Chantal* disagrees. She tells Bob that mediocre movies bring in the most profits. You are asked to advise on who is right.
2. Create a new variable called vote\_average\_rounded. Make sure this variable is the same as vote\_average, but without any decimals (i.e., a 6.3 becomes a 6, a 8.7 an 8, etc.). Display a histogram of vote\_average\_rounded.
3. Create a scatter plot with vote\_average\_rounded on the x axis and the mean of profits within each category of vote\_average\_rounded on the y-axis. Make sure it has an appropriate title, and appropriate titles and labels for the x- and y-axis. At which rating of movies are profits the highest?
4. Recreate the scatter plot with year on the x axis and mean\_profits on the y-axis, but now add bars around each point, indicating the 95% confidence interval.
5. Write an advice to settle the argument between Bob and Chantal.

*step a*

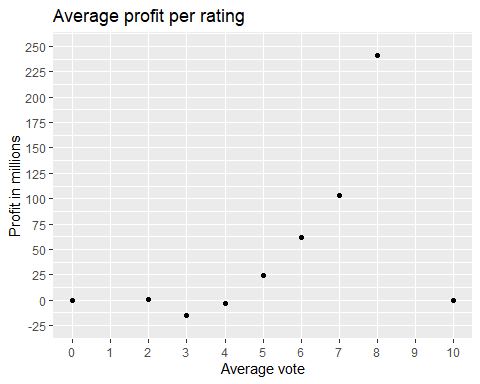
#WRITE YOUR CODE HERE  
movies4 <- movies4%>%  
 mutate(vote\_average\_rounded = floor(vote\_average))  
  
ggplot(movies4, aes(x= vote\_average\_rounded)) +  
 scale\_x\_continuous(breaks = seq(0,10, by = 1))+  
 geom\_histogram()

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



*step b*

movies\_mean\_per\_vote <- movies4 %>%  
 group\_by(vote\_average\_rounded) %>%  
 summarise(Mean\_profit\_per\_vote = mean(profit\_in\_millions))  
  
ggplot(movies\_mean\_per\_vote, aes(x = vote\_average\_rounded, y= Mean\_profit\_per\_vote))+  
 scale\_x\_continuous(breaks = seq(0,10, by = 1), minor\_breaks = F)+  
 scale\_y\_continuous(breaks = seq(-50, 250, 25), limits = c(-25,250)) +  
 labs(x = "Average vote",  
 y = "Profit in millions",  
 title = "Average profit per rating")+  
 geom\_point()



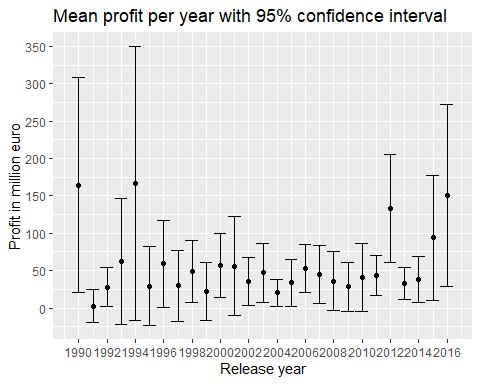
movies\_mean\_per\_vote[8,2]

## # A tibble: 1 × 1  
## Mean\_profit\_per\_vote  
## <dbl>  
## 1 241.

**Your Answer:**

Movies rated with an 8 earn the highest profit, on average the profit is 241.18 million euro. *step c*

mean\_profit\_per\_year <- movies4%>%  
 group\_by(release\_year)%>%  
 summarise(mean\_profit\_per\_year = mean(profit\_in\_millions),  
 n = n(),  
 sd = sd(profit\_in\_millions),  
 se = sd / sqrt(n),  
 t\_value = qt(0.975, df = n-1),  
 lowerbound = mean\_profit\_per\_year - t\_value \* se,  
 upperbound = mean\_profit\_per\_year + t\_value \* se,  
 .groups = "drop")  
  
ggplot(mean\_profit\_per\_year, aes(x = release\_year,y = mean\_profit\_per\_year))+  
 geom\_point()+  
 scale\_x\_continuous(breaks = seq(1990,2016, by = 2))+  
 scale\_y\_continuous(breaks = seq(-50,350, by = 50))+  
 labs(x = "Release year",  
 y = "Profit in million euro",  
 title = "Mean profit per year with 95% confidence interval")+  
 geom\_errorbar(aes(ymin = lowerbound,  
 ymax = upperbound))



**Your Answer:**

The scatterplot shows the mean and the corresponding confidence interval per year.We can conclude out of the plot that some years have a bigger interval than others.

*step d*

**Your Answer:** Bob and Chantal argued about which movies made the biggest profit. Bob thought that high-quality movies made a bigger profit, while Chantal argued that mediocre movies led to higher profits. Chantal says that movies with a rating of around 5 or 6 will lead to the highest profit, and Bob said that movies with higher quality, a rating of around 7-8 or higher, will lead to higher profits. Upon examining the data, the majority of movies have a rating of around 6. However, upon analyzing the mean profit of movies with a rating of 6, we can conclude that it is significantly lower than that of movies with a rating of 8. Movies with a rating of 6 have around 60 million profit, while movies with a rating of 8 have a profit of 240 million. So to give the movie studio an advise is to focus on higher quality movies and listen to Bob.

# Week 4

1. There is another argument going on in the movie studio. *Bob* claims that production budgets are getting out of hand, and that the studio should focus on making cheaper movies. *Chantal* disagrees. She tells Bob that ``Every dollar we spend on movie production is more than offset by the increase in movie profits’’.
2. Set up a regression model to test Chantal’s claim, and estimate it. That is, estimate:

* Print a summary of your estimated model.

1. What is the estimated value of and how do you interpet it?
2. Test for the null hypothesis that . Report the p-value and state your conclusion.
3. Next, estimate the model

* When creating the variables Log Profits and Log Budget, make sure that movies with a Revenue or Budget of zero are assigned the value “NA”. Print a summary of your estimated model

1. What is the estimated value of and how do you interpet it?
2. Which model has better fit? The level-level model or the log-log model? Explain.
3. Who do you think is correct? Bob or Chantal? What would you advise the movie studio to do?

*step a*

#WRITE YOUR CODE HERE

**Your Answer:**

Write your formulated response here.

*step b*

#WRITE YOUR CODE HERE

**Your Answer:**

Write your formulated response here.

*step c*

#WRITE YOUR CODE HERE

**Your Answer:**

Write your formulated response here.

*step d*

#WRITE YOUR CODE HERE

**Your Answer:**

Write your formulated response here.

*step e*

#WRITE YOUR CODE HERE

**Your Answer:**

Write your formulated response here.

*step f*

#WRITE YOUR CODE HERE

**Your Answer:**

Write your formulated response here.

*step g*

#WRITE YOUR CODE HERE

**Your Answer:**

Write your formulated response here.

2

1. Make a plot with a 95% confidence interval with the mean log of budget on the y-axis, and whether the first actor of the movie is male or female on the x-axis. What do you conclude?
2. Estimate the following simple OLS model: $log(budget)\_i=\beta\_0+\beta\_1 \text(FirstActorMale)\_i + \varepsilon\_i.$ Is the estimated coefficient for significantly different from zero? How do you interpret its estimate, and how does this relate to your conclusion in 2a?
3. Now, have a close look at your data frame. Can you find any instances of male first actors who are wrongly labeled as being female, or vice versa? What would such mislabelling mean for the coefficient you estimated under 2b?

*step a*

#WRITE YOUR CODE HERE

**Your Answer:**

Write your formulated response here.

*step b*

#WRITE YOUR CODE HERE

**Your Answer:**

Write your formulated response here.

*step c*

#WRITE YOUR CODE HERE

**Your Answer:**

Write your formulated response here.

# Week 5

1. Create a plot of the mean profits by month of release. Do you see any indication that month of release matters to the profits of the movie?
2. Estimate an OLS model which has as dependent variable the log of profits of a movie, and as independent variable the log of budget, a dummy for whether the movie was released in english or not, and a linear term for the month of release. Show a summary of the resulting model and interpret each coefficient.
3. Test for the hypothesis that the coefficient that belongs to month of release is zero.
4. Based on your plot in a.) do you consider the choice that month of release enters the model linearly under b.) reasonable? Estimate a specification that allows for a more flexible curve. In this new specification, test for the null hypothesis that month of release does not impact profits. This might require testing multiple terms at once.
5. One executive at the studio wants to time the release of the movie to a specific month of the year such that they can maximize revenue. Based on your model under d.), What would you advise the movie studio regarding the timing of the release of the movie?

The movie studio that you work at is releasing a new movie in 2026. It will be an English-spoken Thriller movie with a budget of 40,000,0000.

1. Estimate a model that is able to predict the revenue of this movie. Give its predicted revenue and include a 99% prediction interval.

*step a*

#WRITE YOUR CODE HERE

**Your Answer:**

Write your formulated response here.

*step b*

#WRITE YOUR CODE HERE

**Your Answer:**

Write your formulated response here.

*step c*

#WRITE YOUR CODE HERE

**Your Answer:**

Write your formulated response here.

*step d*

#WRITE YOUR CODE HERE

**Your Answer:**

Write your formulated response here.

*step e*

#WRITE YOUR CODE HERE

**Your Answer:**

Write your formulated response here.

*step f*

#WRITE YOUR CODE HERE