Name	SNR	ANR
Fernando Catala	2048042	u270800
Francesco Lauria	2041074	u819639
Konstantinos Soileidis	2037001	u226393

Group 12

Assignment 1

Exercise 1

1. X~hypergeometric (N, K, n) where:

    N = 1734 ⇒ total n° of shifts
    K = 26 ⇒ total n° of incidents
    n = 203 ⇒ Lucia's n° of shifts
    k = 7 ⇒ Lucia's n° of incidents

    R code: *(see annex)*

    Result:

    P(X >= 7) = 0.02547351

The probability that 7 or more incidents happen in 203 shifts, considering the other total numbers, is 2.5%, which is to say 1/40. This probability of course, is higher than 1/342 million. Therefore, this analysis proves that there is actually a higher probability that incidents may have happened regardless of Lucia's wrongdoing, than in the previous analysis. Anyway, this figure is still very small to clearly exonerate Lucia.

2.

    R code: *(see annex)*

    Result:

    P(X >= 7) = 0.0334 ***(with 10,000 iterations)***

The probability of having 7 or more incidents during 203 shifts under the model of an *innocent nurse* is 3,3%, which is to say 1/30. Using a Poisson distribution resulted in a larger probability than when using a hypergeometric distribution. Therefore, in line with the conclusion of the previous exercise, it is more likely that Lucia's high records of accidents may have happened regardless of a wrongdoing.

3.

R code: *(see annex)*

Result:

P(X >= 7) = 0.1367 ***(with 10,000 iterations)***

The probability of having 7 or more incidents during 203 shifts is 13.7%, which is to say 1/7.
This was achieved by generating a nurse-specific incidence rate with an exponential distribution. Plugging this rate in the simulation with a Poisson distribution resulted in a larger probability than in the previous exercise. Therefore, this method provided a larger probability of Lucia being innocent.

4.

With each subsequent model the assumption of interchangeability of nurses has been relaxed, which resulted each time in a higher probability that Lucia's oddly high number of accidents may have occurred regardless of any deliberate wrongdoing.

Both the original model and the hypergeometric model assume that both nurse shifts and incidents are independent and uniform. However, in reality this assumptions are incorrect. Regarding nurses, they are not interchangeable (= uniform) since they are human beings with different professional skills and personalities. Also, they may have different shift schemes over a week and over the year. About incidents, they are also not uniform since hospitals may have different caring policies and standards.

When relaxing the assumption of interchangeability, which means admitting heterogeneity in the likeliness of experience an incident, it finally appeared that Lucia's odd situation was not unique, but just slightly unusual.

Exercise 2

1. a.

R code: *(see annex)*

Result:

SE Student = 0.3452053
SE Welch = 0.4582576

1. b.

Pseudocode:

```
Input:
    - S = integer defining the number of independent data sets to generate
        - S = 10,000
    - $n_1$ : sample size of the dataset for the 1st group
        - $n_1 = 5$
```

- n_2: sample size of the dataset for the 2nd group
    - $n_2 = 100$
- μ_1, μ_2, σ_1^2, σ_2^2: values for the population parameters of the data generating distribution, which is a normal distribution for both groups
    - $\mu_1 = 0$
    - $\mu_2 = 1$
    - $\sigma_1^2 = 2$
    - $\sigma_2^2 = 1$

Output:

- S estimates of the statistics of interest, namely SE Student and SE Welch

Body:

```
- Initialize output Out_SE_t and Out_SE_w both as a vector of length S
- For s in 1:S
    - Generate data (out1) for group 1 with sample size $n_1$ from $N_1(\mu_1, \sigma_1^2)$
    - Generate data (out2) for group 2 with sample size $n_2$ from $N_2(\mu_2, \sigma_2^2)$
    - Obtain the SE_student and SE_welch from out1 and out2
    - Store the values of SE_student and SE_welch:
        - Out_SE_t [s] = SE_student
        - Out_SE_w [s] = SE_welch
- Return Out_SE_t, Out_SE_w as the output
```

1. c.

Conditions for MC simulation:

According to the task, these parameters are fixed:

- $\sigma_1^2 = 2$
- $n_2 = 100$
- $\mu_1 = 0$
- $\mu_2 = 1$

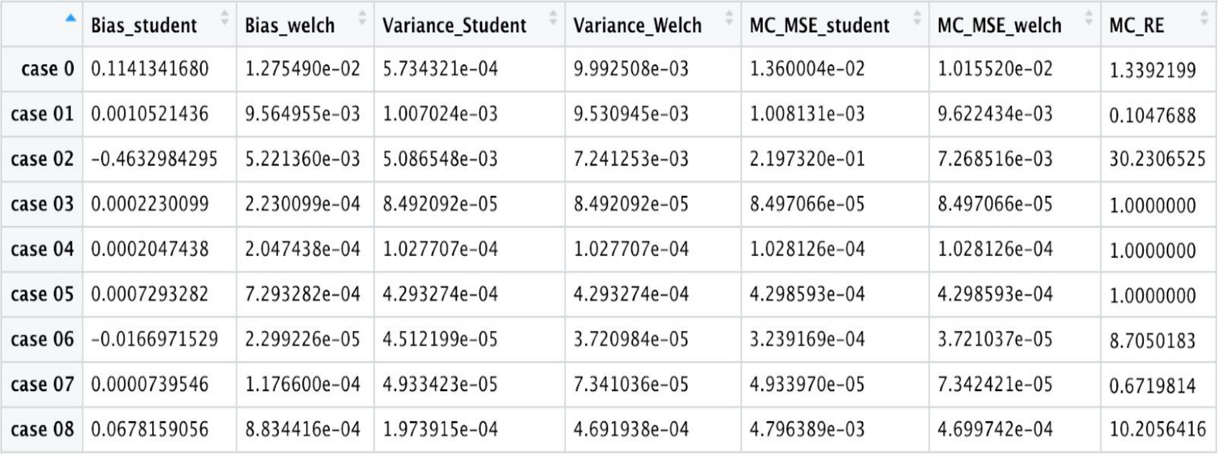
And then we change σ_2^2 and n_1 as following:

# Condition	σ_2^2	n_1
0	1	10
1	2	10
2	10	10
3	1	100
4	2	100

5	10	100
6	1	200
7	2	200
8	10	200

R code: *(see annex)*

Results:



	Bias_student	Bias_welch	Variance_Student	Variance_Welch	MC_MSE_student	MC_MSE_welch	MC_RE
case 0	0.1141341680	1.275490e-02	5.734321e-04	9.992508e-03	1.360004e-02	1.015520e-02	1.3392199
case 01	0.0010521436	9.564955e-03	1.007024e-03	9.530945e-03	1.008131e-03	9.622434e-03	0.1047688
case 02	-0.4632984295	5.221360e-03	5.086548e-03	7.241253e-03	2.197320e-01	7.268516e-03	30.2306525
case 03	0.0002230099	2.230099e-04	8.492092e-05	8.492092e-05	8.497066e-05	8.497066e-05	1.0000000
case 04	0.0002047438	2.047438e-04	1.027707e-04	1.027707e-04	1.028126e-04	1.028126e-04	1.0000000
case 05	0.0007293282	7.293282e-04	4.293274e-04	4.293274e-04	4.298593e-04	4.298593e-04	1.0000000
case 06	-0.0166971529	2.299226e-05	4.512199e-05	3.720984e-05	3.239169e-04	3.721037e-05	8.7050183
case 07	0.0000739546	1.176600e-04	4.933423e-05	7.341036e-05	4.933970e-05	7.342421e-05	0.6719814
case 08	0.0678159056	8.834416e-04	1.973915e-04	4.691938e-04	4.796389e-03	4.699742e-04	10.2056416

1. d.
   The Student's t-test has the following assumptions:
   - Populations are normally distributed (with unknown population mean and variance)
     $X_1 \sim \mathrm{N}(\mu_1, \sigma^2)$ and $X_2 \sim \mathrm{N}(\mu_2, \sigma^2)$
   - Variances of the two populations are the same
   - Samples are drawn independently

   The Welch's t-test is mostly the same as a Student's t-test except that it does not assume equal variances.

   Under the specified conditions, the relative performance of the SE of the difference in means estimator should be better in the Welch's t-test than in Student's t-test, when variances are not equal. However, when variances are equal, Student's t-test estimator should perform better. *Thorough explanation is below in point 2.*

2.

   Which estimator is biased, and under what conditions of sample size and variance?
   - SE_student is more biased under conditions 0, 2, 6, 8 while SE_welch is more biased under conditions 1, 7. Conditions 0, 2, 6, 8 have different variances, therefore they are

violating the assumption of equal variances for Student's t-test, and that's why SE_student is more biased than SE_welch. On the other hand, conditions 1, 7 have equal variances, which is why SE_student is less biased than SE_welch. Finally, conditions 3, 4, 5 have all the same sample sizes, which results in the estimators having same bias.

What is the influence of the sample size and variance σ_2^2 on the variance of the two estimators?

- SE_welch has larger variance under conditions 0, 1, 2, 7, 8 which all have different sample sizes. On the other hand, variances are equal under conditions 3, 4, 5 which all have the same sample size.

Which of the two methods (Student's t versus Welch) is the more efficient one, and under what circumstances?

- MSE & RE: under conditions 0, 2, 6, 8, RE > 1, which means SE_welch is more efficient than SE_student, which also means it has a smaller MSE. This holds true when both sample sizes and variances are different. On the other hand, when variances are equal, SE_student has a lower MSE and therefore RE < 1. This is because equal variances are an assumption of SE_student. Also, when sample sizes are equal, MSEs are equal.

We also compared our results with the suggested literature, in order to further elaborate our conclusions[1]. Some interesting coincidences are:

- Student's t-test is considered stronger if the assumptions are met: data coming from two normal distributions and the variances of the distributions are the same (cases 1 and 7)
- Student's is more biased than Welch when the difference between variances and sample size is large among the groups (evident cases are 2 and 8)
- Welch t-test is assumed as of equal power by literature, so it is suggested that it should be used by default, in order to avoid assumptions that might not be true in the end, regarding the population (equal variances and populations normal distributed). According to our experiment, Welch t-test performs better than Student t-test on 4 out of 6 conditions by excluding the 3 conditions where the two estimators behave the same because of equal sample size.

Annex

[1] International review of social psychology, 30(1), 92–101. DOI: http://doi.org/10.5334/irsp.82.

```
#Exercise 1.1.

x = 6
k = 26
m = 1734
n = 203

phyper (x, k, m-k, n, lower.tail = FALSE)
```

```
## [1] 0.02547351
```

```
#Exercise 1.2.

incidents<-function(S,n){
  # S is the number of experiments
  # n is the number of the sample size
  i<-0
  out_vec<-rep(i,S)
  lambda<-26/1734
  for(i in 1:S){
    incidents<-rpois(n,lambda)
    number_incidents<-sum(incidents)
    if(number_incidents>=7){
      out_vec[i]=1
    }
  }
  solution<-(sum(out_vec)/S)
  return(solution)
}

set.seed(123)
incidents(10000,203)
```

```
## [1] 0.0334
```

```
#Exercise 1.3.

incidents2<-function(S,n){
  # S is the number of experiments
  # n is the number of the sample size
  i<-0
  out_vec<-rep(i,S)
  lambda<- 26/1734
  for(i in 1:S){
    rateexp=rexp(1,1/lambda)
    incidents<-rpois(n,rateexp)
    number_incidents<-sum(incidents)
    if(number_incidents>=7){
      out_vec[i]=1
    }
  }
  solution<-(sum(out_vec)/S)
  return(solution)
}

set.seed(123)
incidents2(10000,203)

## [1] 0.1367

#Exercise 2.1.a

n1<-10
n2<-100
sigma_squared1<-2
sigma_squared2<-1

sp_num<-((n1-1)*sigma_squared1)+((n2-1)*sigma_squared2)
sp_den<-n1+n2-2
SP<-sqrt(sp_num/sp_den)
SE_student<-SP*(sqrt((1/n1)+(1/n2)))

SE_welch<-sqrt((sigma_squared1/n1)+(sigma_squared2/n2))

#Exercise 2.1.c

set.seed(123)
##inputs
sigma_squared1<-2; n1<-c(10,100,200);
sigma_squared2<-c(1,2,10);n2<-100;mu1=0;mu2=1;S=10000

#final matrix to put the properties values of all the 9 conditions
all_results<-data.frame(row.names =
c("Bias_student","Bias_welch","Variance_student","Variance_welch","MSE_stu
dent","MSE_welch","RE_student/welch"))


MonteCarlo=function(S,n1,n2,sigma_squared1,sigma_squared2,mu1,mu2){

  #matrix for calculations
  result<-matrix(nrow=S,ncol=2)
  colnames(result)<-c("SE_student","SE_welch")

  for (s in 1:S){
    #computing new variances
    out1<-rnorm(n1,mu1,sqrt(sigma_squared1))
    out2<-rnorm(n2,mu2,sqrt(sigma_squared2))
    out1_var<-var(out1)
    out2_var<-var(out2)

    ##SE_student
    sp_num<-((n1-1)*out1_var)+((n2-1)*out2_var)
```

```r
    sp_den<-n1+n2-2
    SP<-sqrt(sp_num/sp_den)
    SE_student<-SP*(sqrt((1/n1)+(1/n2)))
    SE_student<-SP*(sqrt((1/n1)+(1/n2)))

    ##SE_Welch
    SE_welch<-sqrt((out1_var/n1)+(out2_var/n2))

    ##Storing results
    result[s,]<-c(SE_student,SE_welch)
  }

  return(result)
}


Evaluation=function(result,n1,n2,sigma_squared1,sigma_squared2){
  #True value parameter
  teta<-sqrt((sigma_squared1/n1)+(sigma_squared2/n2))

  #Bias
  MC_mean<-apply(result,2,mean)
  Bias_student<-teta-MC_mean[1]
  Bias_welch<-teta-MC_mean[2]

  #Variance
  MC_var<-apply(result,2,var)
  Variance_Student<-MC_var[1]
  Variance_Welch<-MC_var[2]

  #MSE
  MC_MSE_student<-(Bias_student^2)+MC_var[1]
  MC_MSE_welch<-(Bias_welch^2)+MC_var[2]

  #Relative efficiency
  MC_RE<-MC_MSE_student/MC_MSE_welch

  #Storing Values

resultdf<-data.frame(Bias_student,Bias_welch,Variance_Student,Variance_Wel
ch,MC_MSE_student,MC_MSE_welch,MC_RE,row.names = "case 0" )
  return(resultdf)
}



#Create the final table thanks to the functions above
for (i in 1:3){
  for (j in 1:3){

MC_result<-MonteCarlo(S,n1[i],n2,sigma_squared1,sigma_squared2[j],mu1,mu2)
```

```
appendThat<-Evaluation(MC_result,n1[i],n2,sigma_squared1,sigma_squared2[j]
)
    all_results<-rbind.data.frame(all_results,appendThat)
  }
}
```