SectionB

EMATM0061,TB1 2022

YujieWang

2023-01-08

Contents

Introduction

This is the Section B,which consists of three parts including **B.1**,**B.2**,**B.3**. **B.1** is about the concepts of conditional probability,**B.2** is about the continue random variables,the law of large numbers,the maximum likelihood estimate. **B.3** is about the foundations of statistical estimation-mean,variance,mean squared error and so on.This paper gives the solutions and codes of all answers and explains the reason for the observed relationship.

The solutions of Section B

To finish the study, "tidyverse" and "dplyr" package should be loaded.

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.0      v purrr   1.0.0
## v tibble  3.1.8      v dplyr   1.0.10
## v tidyr   1.2.1      v stringr 1.5.0
## v readr   2.1.3      v forcats 0.5.2
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
```

B.1

In this question, the case considers the lifetime of products from a light bulb factory. The factory has two LED bulb production lines (Line A and Line B) that independently produce light bulbs.The light bulbs, are not always of high quality, and the products from Line A and Line B have different lifetimes.

1. For a light bulb produced by Line A, the probability that its lifetime is equal to or bigger than 2 years is p_A (a number in [0,1]).
2. A light bulb from Line B generally has a shorter lifetime, and the probability of its lifetime being equal to or bigger than 2 years is p_B (which is less than p_A).

The light bulbs produced from Line A and Line B are considered to be the same products by the factory and they are sold randomly to the customers. These light bulbs are sold in the same packages such that, given a light bulb, a customer can not identify which of the two production lines the light bulb is made from. Suppose that if the customers buy a light bulb produced by the factory, then it must come from either Line A or Line B,and the probability that it comes from Line A is p.

A customer bought a light bulb whose lifetime was unfortunately less than 2 years. The customer suspects that this light bulb was made in Line B.

1. Let α denote the probability that this light bulb was made in Line B, given that its lifetime is less than 2 years. Derive a mathematical expression for α in terms of p_A, p_B and p.

$$P(bulb\,made\,in\,Line\,B|the\,lifetime\,is\,less\,than\,2\,years) = \frac{(1-p_B)\cdot(1-p)}{(1-p_A)\cdot p + (1-p_B)\cdot(1-p)}$$

2. Suppose that $p_A = 0.99$, $p_B = 0.5$ and $p = 0.1$,then the numerical value of α is:

$$\alpha = \frac{0.5 \times 0.9}{0.01 \times 0.1 + 0.5 \times 0.9} = 0.9977$$

Next, fix $p_A = 0.99$, $p_B = 0.5$ and $p = 0.1$. The experiment will conduct a simulation study to estimate the conditional probability α with samples.

3. The simulation study should contain 100000 trials. In each of the trials, generate a sample of a light bulb. Each sample is represented by a pair of randomly generated numbers called $(Line, LessThan2Years)$:

   (1) The random number Line represents the product line that the light bulb was made from (with Line=0 representing Line A and Line=1 representing Line B). The probability of Line=0 should be equal to p.
   (2) The random number LessThan2Years should be either 0 or 1, where LessThan2Years = 0 represents that the lifetime of this bulb is not less than 2 years, and LessThan2Years = 1 represents that the lifetime is less than 2 years. The number LessThan2Years should be generated by taking into account the value of Line. If Line = 0, then the probability of LessThan2Years = 0 is equal to p_A. If Line = 1, then the probability of LessThan2Years = 0 is equal to p_B.

4. Based on the samples generated in the simulation study, compute an estimate of the conditional probability α.

   (1) First, select the subset of samples in which LessThan2Years =1.
   (2) Second, within this subset of samples, compute the number of samples in which Line = 1, and divide it by the total number of samples within this subset to get an estimate of the conditional probability α. Display the estimate to at least 5 decimal places.

5. From the simulation study below, it shows that the estimate of the conditional probability α.

```
num_trials<-100000 # set the number of trials
set.seed(0) # set the random seed
p=0.1 # set the proability of "Line=1"
pA=0.99 # set the probability of PA
pB=0.5 # set the probability of PB
Line<-rbinom(num_trials,1,1-p) #generate the samples of product lines
Lessthan2yearsfun<-function(num){ #generate the samples of lifetime
  if(num==0){ # if line=0
    # the proability of LessThan2Years=1 is equal to 1-PA
    return(rbinom(1,1,1-pA))
  }else if(num==1){ # if line=1
    #the probability of LessThanYears=0 is equal to 1-PB
    return(rbinom(1,1,1-pB))
  }
}
Sample_data<-data.frame(trial=1:num_trials,Line)%>%
  # generate the data frame of samples
  mutate(lessthan2years=unlist(map(.x=Line,.f=~Lessthan2yearsfun(.x))))
Sample_lessthan2years<-Sample_data%>%
  filter(lessthan2years==1)%>%
  # Select the subset of samples in which lessthan2years=1
  summarise(num_Line=sum(Line),num_lifetime=sum(lessthan2years))
  #compute the number of samples in which Line=1 and lessthan2years=1
alpha<-(Sample_lessthan2years$num_Line)/(Sample_lessthan2years$num_lifetime)
```

```
  # compute the probability alpha
  #The bulb's lifetime is less than 2 years, and it comes from product line B
print(alpha)
```

```
## [1] 0.9978612
```

B.2

In this question, the case will explore statistical estimation for parameters in continuous random variables.

1. Suppose a product is being sold in a supermarket.
2. The case is interested in knowing how quickly the product returns to the shelf again after it is sold out.
3. Let X be a continuous random variable denoting the length of time between the time point at which it is sold out and the time point at which it is placed on the shelf again. So X should be a non-negative number, and $X = 0$ means that the product gets on the shelf immediately after it is sold out. Here, the case assumes that the probability density function of X is given by

$$p_\lambda(x) = \begin{cases} ae^{-\lambda(x-b)} & \text{if } x \geq b, \\ 0 & \text{if } x < b, \end{cases}$$

where $b > 0$ is a known constant, $\lambda > 0$ is a parameter of the distribution, and a is to be determined by λ and b.

(1) First, determine the value of a: derive a mathematical expression of a in terms of λ and b.

$$\begin{aligned} \int_{-\infty}^{b} 0\mathrm{d}x + \int_{b}^{+\infty} ae^{-\lambda(x-b)} &= 1 \\ [-\frac{a}{\lambda}e^{-\lambda(x-b)}]_b^{+\infty} &= 1 \\ 0 - [-\frac{a}{\lambda}] &= 1 \\ a &= \lambda \end{aligned}$$

(2) Derive a formula for the population mean and stand deviation of the exponential random variable X with parameter λ.

1. **Population mean:**

$$\begin{aligned} \mathrm{E}(X) &= \int_{-\infty}^{+\infty} xp_\lambda(x)\mathrm{d}x \\ &= \int_{b}^{+\infty} x\lambda \cdot e^{-\lambda(x-b)}\mathrm{d}x \\ &= [(-\frac{1}{\lambda} - x)e^{-\lambda(x-b)}]_b^{+\infty} \\ &= [-xe^{-\lambda(x-b)}]_b^{+\infty} + [-\frac{1}{\lambda}e^{-\lambda(x-b)}]_b^{+\infty} \\ &= b + \frac{1}{\lambda} \end{aligned}$$

2. **Standard deviation:**

$$\begin{aligned}\sigma(x) &= \sqrt{\mathrm{Var}(x)} \\ &= \sqrt{\mathrm{E}(X^2) - \mathrm{E}(X)^2}\end{aligned}$$

$$\begin{aligned}\mathrm{E}(X^2) &= \int_b^{+\infty} x^2 p_\lambda(x)\mathrm{d}x \\ &= \int_b^{+\infty} x^2 \cdot \lambda e^{-\lambda(x-b)}\mathrm{d}x \\ &= [-x^2 e^{-\lambda(x-b)}]_b^{+\infty} + 2\int_b^{+\infty} x e^{-\lambda(x-b)}\mathrm{d}x \\ &= b^2 + \frac{2}{\lambda}\int_b^{+\infty} \lambda x e^{-\lambda(x-b)} \\ &= b^2 + \frac{2}{\lambda}\cdot \mathrm{E}(X) \\ &= b^2 + \frac{2}{\lambda}\cdot(\frac{1}{\lambda} + b) \\ &= b^2 + \frac{2}{\lambda^2} + \frac{2b}{\lambda}\end{aligned}$$

$$\begin{aligned}\text{So} \quad \sigma(x) &= \sqrt{b^2 + \frac{2}{\lambda^2} + \frac{2b}{\lambda} - (b + \frac{1}{\lambda})^2} \\ &= \sqrt{\frac{1}{\lambda^2}} \\ &= \frac{1}{\lambda}\end{aligned}$$

(3) Derive a formula for the cumulative distribution function and the quantile function for the exponential random variable X with parameter λ.

1. The **cumulative distribution function** is:

$$F_X(x) = \int_{-\infty}^{x} p_\lambda(x)\mathrm{d}x = [-e^{-\lambda(x-b)}]_b^x = 1 - e^{-\lambda(x-b)}$$

for $x \geq b$, and $F_X(x) = 0$ for $x < b$.

2. The **quantile function** is $F_X^{-1}(p) = \inf\{x \in \mathbb{R} : F_X(x) \leq p\}$, so:

$$F_X^{-1}(p) = \ln(1-p)/(-\lambda) + b$$

(4) Suppose that $X_1, \cdots, X_n$ are independent copies of X with the unknown parameter $\lambda > 0$.The case need to compute the maximum likelihood estimate λ_{MLE} for λ.

1. The **likelihood function** is:

$$\begin{aligned}L(\lambda; X) &= \prod_{i=1}^{n} p_\lambda(X_i) \\ &= \prod_{i=1}^{n} \lambda e^{-\lambda(X_i - b)}\end{aligned}$$

2. The **log-likelihood function** is:

$$\log L(\lambda; X) = n\log\lambda - \lambda\sum_{i=1}^{n}(X_i - b)$$

$$\frac{\partial}{\partial\lambda}\log L(\lambda; X) = \frac{n}{\lambda} - \sum_{i=1}^{n}(X_i - b)$$

Let $Y = X - b$

If $\lambda < 1/(\bar{Y}) := \frac{1}{n}\sum_{i=1}^{n}(Y_i)$, then $\frac{\partial}{\partial\lambda}\log L(\lambda) > 0$.

If $\lambda > 1/(\bar{Y}) := \frac{1}{n}\sum_{i=1}^{n}(Y_i)$, then $\frac{\partial}{\partial\lambda}\log L(\lambda) < 0$.

3. So the maximum likelihood estimate for λ is $\hat{\lambda}_{MLE} = 1/\bar{Y}$, and $Y = X - b$.

4. Now download the "supermarket_data_EMATM0061.csv" file. The .csv file contains synthetic data on the length of time (in seconds) taken by a product to get on the shelf again after being sold out. So the sample is a sequence of time lengths. The case need to model the sequence of time lengths in the sample as independent copies of X with parameter λ and known constant $b = 300$ (seconds).

   (5) Based on the expression shown above, the maximum likelihood estimate of λ_{MLE} of the parameter λ can be computed and displayed below.

   (6) Apply the method of Bootstrap confidence interval to obtain a confidence interval for λ with a confidence level of 95%. To compute the Bootstrap confidence interval, the number of resamples (i.e., subsamples that are generated to compute the bootstrap statistics) should be set to 10000.Based on 10000 bootstrap replicates,the confidence interval is shown below:

```
# the solution of the 5th question
b<-300
supermarket_data<-read.csv("supermarket_data_EMATM0061.csv")
#read the data in .csv file
Supermarket_process<-supermarket_data%>%mutate(Timediff=TimeLength-b)
#compute Y=X-b
lambda_mle<-1/mean(Supermarket_process$Timediff,na.rm=TRUE)
#compute maximum likelihood estimate
print(lambda_mle)
```

```
## [1] 0.01988426
```

```
# the solution of the 6th question
library(boot) #load the library

set.seed(123) #set random seed

#1. define a function which computes the lambda of a column of interest
compute_lambda<-function(df,indicies,col_name){
  sub_sample<-slice(df,indicies)%>% pull(all_of(col_name)) #extract subsample
  return(1/mean(sub_sample,na.rm=TRUE)) #reutrn lambda
}

#2. use the boot function to generate the bootstrap statistics
results<-boot(data=Supermarket_process,statistic=compute_lambda,col_name="Timediff",R=10000)

#3. compute the 95% confidence interval for the lambda
boot.ci(boot.out = results,type="basic",conf=0.95)
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 10000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = results, conf = 0.95, type = "basic")
```

```
##
## Intervals :
## Level      Basic
## 95%   ( 0.0191,  0.0207 )
## Calculations and Intervals on Original Scale
```

5. Next, conduct a simulation study to explore the behaviour of the maximum likelihood estimator:

   (7) Conduct a simulation study to explore the behaviour of the maximum likelihood estimator λ_{MLE} for λ on simulated data $X_1, \cdots, X_n$ (as independent copies of X with parameter λ) according to the following instructions. Take $b = 0.01$ and consider a setting in which $\lambda = 2$ and generate a plot of the mean squared error as a function of the sample size n. The sample size between 100 and 5000 in increments of 10, and 100 trials per sample size should be considered. For each trial of each sample size generate a random sample $X_1, \cdots, X_n$ (as independent copies of X with parameter $\lambda = 2$), then compute the maximum likelihood estimate λ_{MLE} for λ based upon the corresponding sample.

```
set.seed(0)
num_trials_per_sample_size <- 100
min_sample_size <- 100
max_sample_size <- 5000
sample_size_inc <- 10
b <- 0.01
lambda <- 2

# create data frame of all pairs of sample_size and trial
simulation_df<-crossing(trial=seq(num_trials_per_sample_size),
                        sample_size=seq(min_sample_size,
                                        max_sample_size,sample_size_inc)) %>%
  # simulate sequences of random variables
  # create a new column to show the random variables
  # (use rexp() function to generate the random variables)
  mutate(simulation=pmap(.l=list(trial,sample_size),
                         .f=~rexp(.y,rate=lambda))) %>%
  # compute the lambda_mle(maximum likelihood estimator)
  mutate(sample_lambda=map_dbl(.x=simulation,.f=~(1/mean(.x)))) %>%
    group_by(sample_size) %>%
  #compute the mean square error
    summarise(msq_error_lambda=mean((sample_lambda-lambda)^2))
simulation_df
```

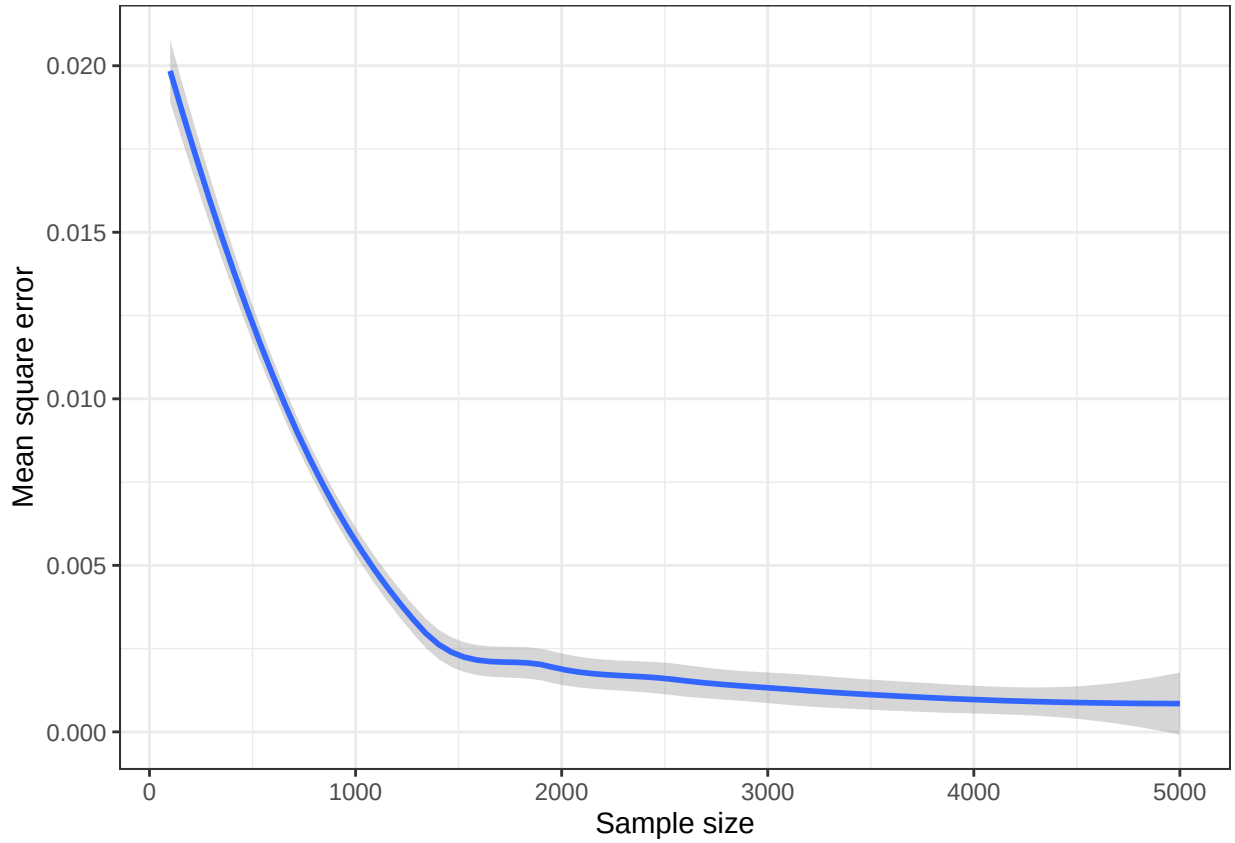
```
## # A tibble: 491 x 2
##    sample_size msq_error_lambda
##          <dbl>            <dbl>
##  1         100           0.0334
##  2         110           0.0586
##  3         120           0.0351
##  4         130           0.0361
##  5         140           0.0346
##  6         150           0.0322
##  7         160           0.0207
##  8         170           0.0247
##  9         180           0.0260
```

```
## 10         190           0.0179
## # ... with 481 more rows
```

The plot of the mean square error of λ_{MLE} as an estimator for λ as a function of the sample size n is displayed in the chart below.It can be seen that the value of mean squared error is descending as the sample size increases.

```
simulation_df%>%
  # create the plot of mean square error
  ggplot(aes(x=sample_size,y=msq_error_lambda)) +
  # use the geom_smooth() function to draw the line
  geom_smooth()+theme_bw()+xlab("Sample size")+ylab("Mean square error")
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



In the code chunk below, it shows another method to compute the maximum likelihood estimate, which uses the self-defined function `bexp()` to generate the random variables and uses the self-defined function `compute_lambda()` to compute the maximum likelihood estimate λ_{MLE}.

```
set.seed(0)
num_trials_per_sample_size <- 100
min_sample_size <- 100
max_sample_size <- 5000
sample_size_inc <- 10
b <- 0.01
```

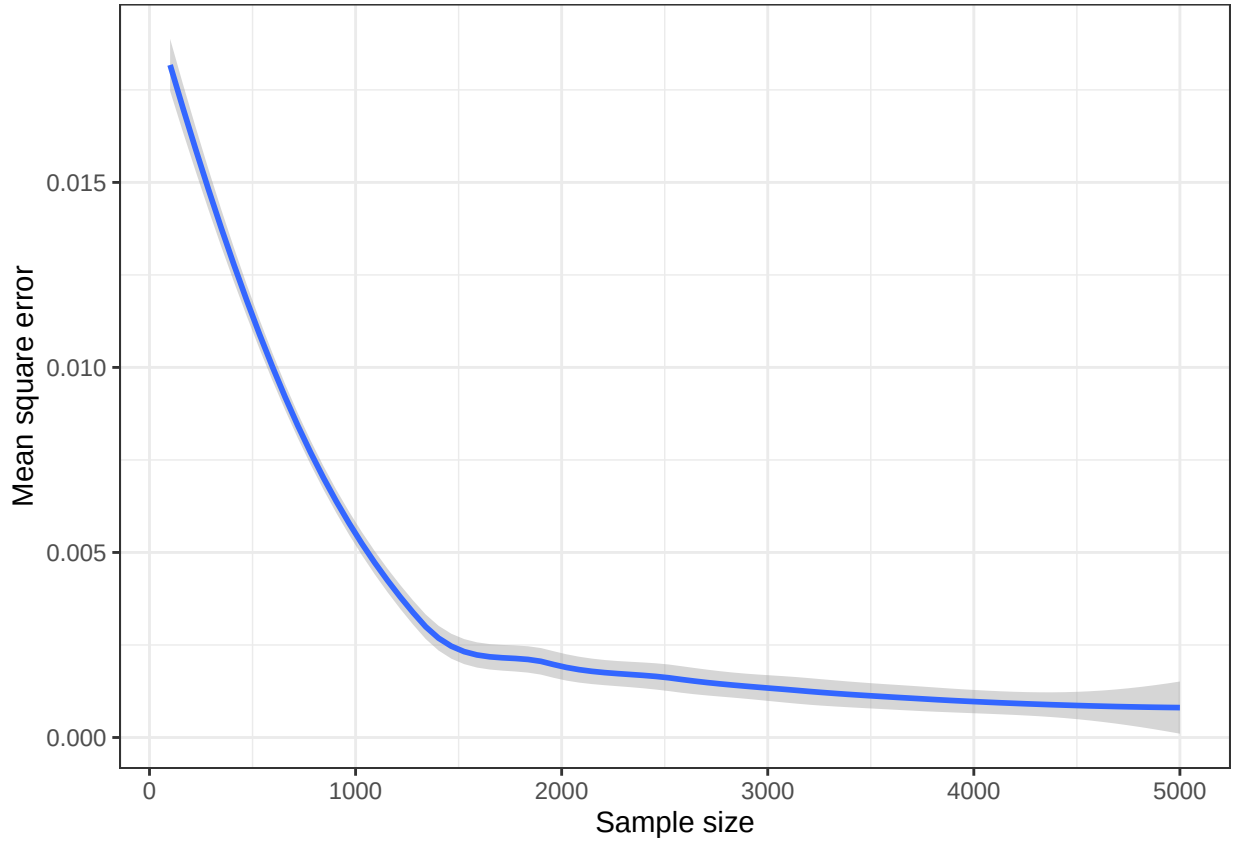
```
lambda <- 2

bexp<-function(n,rate,b){ # the probability density function
  u=runif(n,min=-1,max=0) #
  x<-b-1/rate*log(-u)
  return(x)
}

compute_lambda<-function(l,b){ #the function to compute the lambda_mle
  l1<-unlist(l)
  l2<-l1-b
  lam<-1/mean(l2,na.rm=TRUE) #compute the lambda_mle,remove the missing values
  return(lam)
}

# create data frame of all pairs of sample_size and trial
simulation_lam<-crossing(trial=seq(num_trials_per_sample_size),
                         sample_size=seq(min_sample_size,
                                         max_sample_size,sample_size_inc)) %>%
  # simulate sequences of random variables
  # use the bexp() function created before to generate the random variables
  mutate(simulation=pmap(.l=list(trial,sample_size),
                         .f=~bexp(.y,rate=lambda,b))) %>%
  # compute the lambda_mle(maximum likelihood estimator)
  mutate(sample_lambda=map_dbl(.x=simulation,.f=~compute_lambda(.x,b))) %>%
    group_by(sample_size) %>%
  #compute the mean square error
    summarise(msq_error_lambda=mean((sample_lambda-lambda)^2))
simulation_df
```

```
## # A tibble: 491 x 2
##    sample_size msq_error_lambda
##          <dbl>            <dbl>
##  1         100           0.0334
##  2         110           0.0586
##  3         120           0.0351
##  4         130           0.0361
##  5         140           0.0346
##  6         150           0.0322
##  7         160           0.0207
##  8         170           0.0247
##  9         180           0.0260
## 10         190           0.0179
## # ... with 481 more rows
```

The new plot of the mean square error of λ_{MLE} as an estimator for λ as a function of the sample size n is displayed in the chart below. Actually, the paper just provides an another method to compute the mean square error data,from the chart below,it shows that there is almost no difference between two methods.The value of mean squared error is descending as the sample size is rising.

```
simulation_lam%>%
  ggplot(aes(x=sample_size,y=msq_error_lambda)) +
  geom_smooth()+theme_bw()+xlab("Sample size")+ylab("Mean square error")
```

```
## `geom_smooth()` using method = 'loess' and formula = 'y ~ x'
```



B.3

Consider a bag of a red balls and b blue balls (so the bag has $a+b$ balls in total), where $a \geq 1$ and $b \geq 1$. In this question, we randomly draw two balls from the bag without replacement. That means, we draw the first ball from the bag and, without returning the first ball to the bag, we draw the second one. Each ball has an equal chance of being drawn. Now we record the colour of the two balls drawn from the bag, and let X denote the number of red balls minus the number of blue balls. So X is a discrete random variable. For example, if we draw one red ball and one blue ball, then $X = 0$.

1. Give a formula for the probability mass function $p_X : \mathbb{R} \to [0, 1]$ of X.

   There are three possibilities for the outcome of the experiment, the first is two red balls and no blue balls, the second is two blue balls and no red balls, and the third is one blue ball and one red ball.

   Provided that there is a blue ball and a red ball drawn from the bag, there are still two probabilities.The first probability is that the first ball drawn from the bag is red, and the second is blue. The second probability is that the first ball drawn from the bag is blue and the second is red.

$$p_X(x) = \begin{cases} \dfrac{b(b-1)}{(a+b)(a+b-1)} & \text{if } x = -2, \\ \dfrac{2ab}{(a+b)(a+b-1)} & \text{if } x = 0, \\ \dfrac{a(a-1)}{(a+b)(a+b-1)} & \text{if } x = 2, \\ 0 & \text{otherwise.} \end{cases}$$

2. Use the the probability mass function p_X to obtain an expression of the expectation $\mathrm{E}(X)$ of X (i.e., the population mean) in terms of a and b.

$$\begin{aligned}
\mathrm{E}(X) &= \sum_{x\in\mathbb{R}} x \cdot p_X(x) \\
&= -2 \cdot \frac{b(b-1)}{(a+b)(a+b-1)} + 0 \cdot \frac{2ab}{(a+b)(a+b-1)} + 2 \cdot \frac{a(a-1)}{(a+b)(a+b-1)} + \sum_{x\in\mathbb{R}\setminus 0,-2,2} x \cdot p_X(x) \\
&= \frac{2(a(a-1)-b(b-1))}{(a+b)(a+b-1)} \\
&= \frac{2(a-b)}{(a+b)}
\end{aligned}$$

3. Give an expression of the variance $\mathrm{Var}(X)$ of X in terms of a and b.

$$\begin{aligned}
\mathrm{Var}(x) &= \mathrm{E}[(X-\mathrm{E}(X))^2] \\
&= \sum_{x\in\mathbb{R}} p_X(x) \cdot \left(x - \frac{2(a-b)}{(a+b)}\right)^2 \\
&= \frac{b(b-1)}{(a+b)(a+b-1)} \cdot \left(-2 - \frac{2(a-b)}{(a+b)}\right)^2 + \\
&\frac{2ab}{(a+b)(a+b-1)} \cdot \left(\frac{2(a-b)}{(a+b)}\right)^2 + \\
&\frac{a(a-1)}{(a+b)(a+b-1)} \cdot \left(2 - \frac{2(a-b)}{(a+b)}\right)^2 \\
&= 4 - \frac{8ab}{(a+b)(a+b-1)} - 4\left(\frac{a-b}{a+b}\right)^2
\end{aligned}$$

4. Then, Write a function called `compute_expectation_X` that takes a and b as input and outputs the expectation $\mathrm{E}(X)$. Write a function called `compute_variance_X` that takes a and b as input and outputs the variance $\mathrm{Var}(X)$.

```
compute_expectation_X<-function(a,b){ # compute the expectation of X
  expectation_x<-2*(a-b)/(a+b) # based on the expression,compute the expectation
  return(expectation_x)
}

compute_variance_X<-function(a,b){ #compute the variance of X
  # based on the expression, compute the variance
  variance_x<-4-(8*a*b/((a+b)*(a+b-1)))-4*((a-b)/(a+b))^2
  return(variance_x)
}

print(compute_expectation_X(3,4)) # the expectation of X
```

```
## [1] -0.2857143
```

```
print(compute_variance_X(3,4)) # the variance of X
```

```
## [1] 1.632653
```

Additionally, suppose that $X_1, X_2, \cdots, X_n$ are independent copies of X. So $X_1, X_2, \cdots, X_n$ are i.i.d.random variables having the same distribution as that of X. Let $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ be sample mean.

5. Give an expression of the expectation of the random variable $\bar{X}$ in terms of a, b.

$$\mathrm{E}(\bar{X}) = \mathrm{E}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(X_i) = \frac{2(a-b)}{a+b}$$

6. Give an expression of the variance of the random variable $\bar{X}$ in terms of a, b and n.

$$\mathrm{Var}(\bar{X}) = \mathrm{Var}(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(X_i) = \frac{4}{n}\left(1 - \frac{2ab}{(a+b)(a+b-1)} - \left(\frac{a-b}{a+b}\right)^2\right)$$

7. Create a function called `sample_Xs` which takes as inputs a, b and n and outputs a sample $X_1, X_2, \cdots, X_n$ of independent copies of X.

```
# the 7th question
sample_Xs<-function(a,b,n){
  p1<-(b*(b-1))/((a+b)*(a+b-1)) #compute the probability of "X=-2"
  p2<-(2*a*b)/((a+b)*(a+b-1)) #compute the probability of "X=0"
  p3<-(a*(a-1))/((a+b)*(a+b-1)) #compute the probability of "X=2"
  return(sample(c(-2,0,2),n,replace=TRUE,prob=c(p1,p2,p3)))
  # generate the sample use the function
}

sample_Xs(3,4,10) # compute a sample with the inputs: a=3,b=4,n=10
```

```
##  [1]  0  2  0 -2  2  0 -2  0  0  0
```

8. Let $a = 3$, $b = 5$ and $n = 100000$.Compute the numerical value of $\mathrm{E}(X)$ using the function `compute_expectation_X` and compute the numerical value of $\mathrm{Var}(X)$ using the function `compute_variance_X`. Then use the function `sample_Xs` to generate a sample $X_1, X_2, \cdots, X_n$ of independent copies of X. With the generated sample, compute the sample mean $\bar{X}$ and sample variance.

   - From the outputs, it shows that the difference between the sample mean and the population mean(expectation) is close to 0 (0.00394000000000005).Apart from that, the difference between the sample variance and the population variance is also close to 0 (0.00072316537526524). This is due to the law of large numbers.
   - The weak laws of large numbers: Let $X : \Omega \to \mathbb{R}$ be a random variable with a well-defined expectation $\mu = E(X)$. Let $X_1, ..., X_n : \Omega \to \mathbb{R}$ be a sequence of independent copies of X. Then for all $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\frac{1}{n}\sum X_i - \mu\right| \geq \epsilon\right) = 0$$

     The law of large numbers tells that the sample mean approaches the population mean in some sense when the sample size is large for sequences of independent and identically distributed random variables.

```r
# the 8th question
a<-3
b<-5
n<-100000
Ex<-compute_expectation_X(a,b)  # compute the expectation based on given a,b n
Varx<-compute_variance_X(a,b) # compute the variance
Sample_X<-sample_Xs(a,b,n) #generate a sample based on a,b,n
sample_mean<-mean(Sample_X) # compute the sample mean
sample_variance<-var(Sample_X) # compute the sample variance
print(paste("The expectation is:",Ex))
```

```
## [1] "The expectation is: -0.5"
```

```r
print(paste("The Variance is:",Varx))
```

```
## [1] "The Variance is: 1.60714285714286"
```

```r
print(paste("The sample mean is:",sample_mean))
```

```
## [1] "The sample mean is: -0.49702"
```

```r
print(paste("The sample variance is:",sample_variance))
```

```
## [1] "The sample variance is: 1.60902720987236"
```

```r
diff_mu<-abs(sample_mean-Ex)
diff_var<-abs(sample_variance-Varx)
print(paste(
  "The difference between the population mean and sample mean:",
  diff_mu))
```

```
## [1] "The difference between the population mean and sample mean: 0.00298000000000043"
```

```r
print(paste(
  "The difference between the population variance and sample variance:",
  diff_var))
```

```
## [1] "The difference between the population variance and sample variance: 0.0018843527295016"
```

Moreover, let $\mu := \mathrm{E}(X)$ and $\sigma := \sqrt{\mathrm{Var}(X)/n}$, and let $f_{\mu,\sigma} : \mathbb{R} \to [0, \infty)$ be the probability density function of a Gaussian random variable with distribution $\mathcal{N}(\mu, \sigma^2)$, i.e., the expectation is μ and the variance is σ^2. Next, conduct a simulation study to explore the behaviour of the sample mean $\bar{X}$.

9. Let $a = 3$, $b = 5$ and $n = 900$. Conduct a simulation study with 20000 trials. In each trial, generate a sample $X_1, \cdots, X_n$ of independent copies of X. For each of the 20000 trials, compute the corresponding sample mean X based on $X_1, \cdots, X_n$.

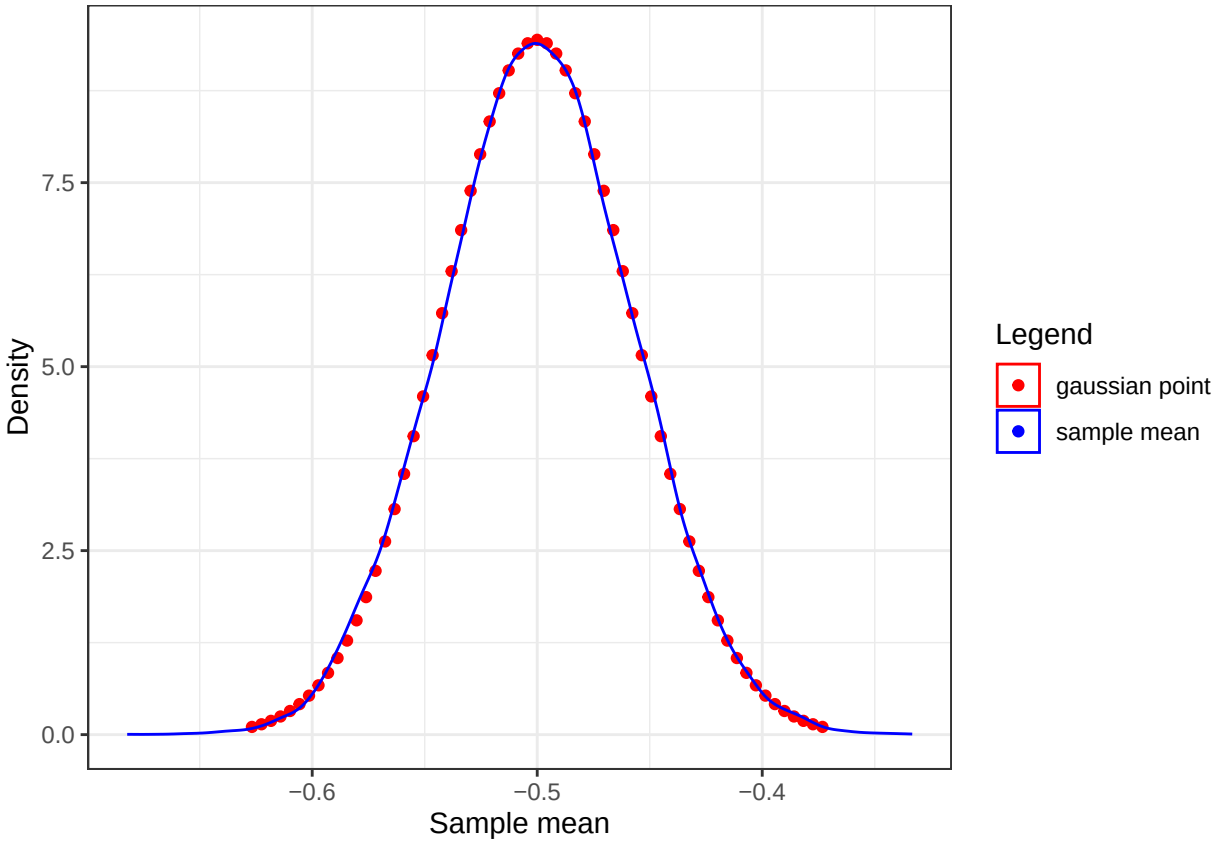
```r
# the 9th question
num_trials<-20000
a<-3
b<-5
n<-900
simulation_data<-data.frame(trials=1:num_trials)%>%
  # generate a column of trials
  mutate(sample_X=map(.x=trials,.f=~sample_Xs(a,b,n)))%>%
  #for each trial, generate a sample
  mutate(sample_mean=map_dbl(.x=sample_X,.f=~mean(.x)))
#for each sample, generate a sample mean
simulation_data%>%head(5)
```

```
##   trials
## 1      1
## 2      2
## 3      3
## 4      4
## 5      5
##
## 1         0, 0, 0, 0, 2, 0, -2, 0, -2, 0, 0, -2, 0, 0, -2, 0, -2, -2, 0, 0, 2, 0, -2, -2, -2, -2, 0,
## 2              -2, 0, 0, 0, 0, 0, 0, 0, -2, -2, 0, 0, 0, 0, 2, 0, 0, -2, 0, 2, 0, -2, -2, 0, 0, 0, -
## 3 0, 0, -2, 2, 2, -2, -2, 0, 2, -2, 0, 0, -2, 0, 0, 0, 0, 0, -2, 0, -2, -2, 0, -2, -2, 0, 0, -2, 0, 0
## 4     0, -2, 2, 2, -2, 0, 0, 0, -2, -2, 0, -2, -2, 2, -2, -2, 0, 2, -2, 0, -2, 0, -2, 0, 0, -2, 0, -
## 5                                   0, 0, -2, -2, 0, 0, 0, -2, 0, -2, 0, -2, 0, -2, 0, -2, 0, 0, 0, -
##   sample_mean
## 1  -0.5355556
## 2  -0.4888889
## 3  -0.5333333
## 4  -0.5177778
## 5  -0.4755556
```

10. Create a scatter plot of the points $\{(x_i, f_{\mu,\sigma}(x_i))\}$ where $\{x_i\}$ are a sequence of numbers between $\mu-3\sigma$ and $\mu+3\sigma$ in increments of 0.1σ . Then append to the scatter plot a curve representing the kernel density of the sample mean $\bar{X}$ within the simulation study (with 20000 trials). Use different colours(red and blue) for the point $\{(x_i, f_{\mu,\sigma}(x_i))\}$ and the curve in the kernel density plot of the sample mean $\bar{X}$. The result is shown below.

```r
a<-3
b<-5
n<-900
mu<-compute_expectation_X(a,b) # compute the E(X)
sigma<-sqrt(compute_variance_X(a,b)/n)
# compute the parameter sigma of Gaussian function
scale_X<-seq(mu-3*sigma,mu+3*sigma,0.1*sigma)
# generate a sequence of numbers

scatter_data<-data.frame(scale_X)%>%
  #generate a data frame including "xi" and corresponding gaussian values
  mutate(f_xi=dnorm(scale_X,mean=mu,sd=sigma))
  # generate the corresponding gaussian values

ggplot()+theme_bw()+ # create a plot
```

```
# create a scatter plot
geom_point(data=scatter_data,aes(x=scale_X,y=f_xi,color="gaussian point"))+
# add a curve representing the kernel density of the sample mean
geom_density(data=simulation_data,aes(x=sample_mean,color="sample mean"))+
#use different colors for the scatter plot and the curve
scale_color_manual(name="Legend",values=c("gaussian point"="red","sample mean"="blue"))+
labs(x="Sample mean",y="Density")
```



11. As can be seen from the chart above,each point $\{(x_i, f_{\mu,\sigma}(x_i))\}$ is close to the curve in the kernel density plot of the sample mean $\bar{X}$. Both of them illustrate a trending of the Gaussian random variables.

    This is due to the central limit theorem:

    Let $X : \Omega \to \mathbb{R}$ be a random variable with expectation $\mu = E(X)$ and variance $\sigma^2 = \mathrm{Var}(x)$. Let $X_1, \cdots, X_n : \Omega \to \mathbb{R}$ be a sequence of independent copies of X. Let $Z \sim \mathcal{N}(0,1)$ be a standard Gaussian random variable. Then for all $x \in \mathbb{R}$,

    $$\lim_{n\to\infty} \mathbb{P}\left\{\sqrt{\frac{n}{\sigma^2}}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right) \le x\right\} = \mathbb{P}(Z \le x)$$

    The distribution of $G_n = \sqrt{\frac{n}{\sigma^2}}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)$ converges to the standard Gaussian distribution $\mathcal{N}(0,1)$.

    The central limit theorem states that given a population with a mean μ and a standard deviation σ and a sufficiently large random sample from the population with replacement, the distribution of sample means will be approximately normally distributed.