REPORT OF DATA MINING

Lab1

He NI

BI2

20150742

**Part A Discrete series**

1. **Generate a discrete series of 1000 random data (values included between 0 and 10):**

**In order to don’t change the value: I used “set.seed”**

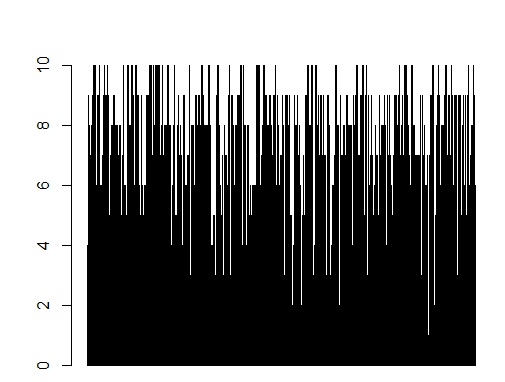
set.seed(250)

A = round(runif(1000,0,10))

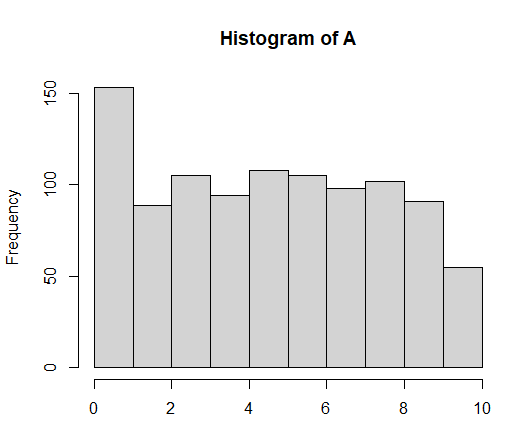


**2. Represent this series in the form of a histogram: to do so, you may use the R functions barplot or hist. See R help to find the right parameters.**

**barplot(A) :**



**hist(A) :**



**3. Determine the mode, the median and the mean of this series without using the predefined R functions.**

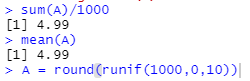
**Mode:**

**names(which.max(table(A)))**



**Mean:**

**sum(A)/1000**



**Median:**

**sum(sort(A)[500:501])/2**



**4. Verify the mean and the median value of your series using the functions mean(·) and median(·). The results should be identical with these of question 3.**

**Mean:** 

**Median:** 

**5. Explain why the mean and median values of this series may be very different.**

***Because outlier values in our data can distort the results and visualizations.***

**6. Determine the range, the variance and the standard deviation:**

• Without using the predefined R functions.

**Range:**

**min(A),max(A)**

**-------------------------------------------------------------------------------**

**Var:**

**ans = 0**

**i = 0**

**for(v in A)**

**{**

**ans = ans + (v -mean(A))^2**

**print(ans)**

**i = i+1**

**}**

**var = sum(ans)/1000**

**print(var)**

**-------------------------------------------------------------------------------**

**Sd:**

**ans = 0**

**i = 0**

**for(v in A){**

 **ans = ans + (v-mean(A))^2**

**print(ans)**

**i = i+1**

**}**

**sd = sqrt(sum(ans)/1000)**

**print(sd)**

• Using the predefined R functions: range(·), var(·) and sd(·).

**Range:**



**Var:**



**Sd:**

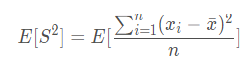


• Comment the results.

For variance

Using the formula: ，The result is ，which is **biased estimate**

Calculate the expected value of the sample variance to estimate the difference between the biased variance **S2** and the unbiased variance **σ2**



By deducing the formula, the final result is ，Bias is,

Then I can get the formula of **unbiased estimates** σ2.

And also using var(A) can get σ2，i.e., **unbiased estimates**，The formula is

I have : S2  and : σ2

So according to then

Then the above formula can also be changed to：

**ans = 0**

**i = 0**

**for(v in A)**

**{**

**ans = ans + (v -mean(A))^2**

**print(ans)**

 **i = i+1**

**}**

**var = sum(ans)/(1000-1)**

**print(var)**

**-------------------------------------------------------------------------------**

Standard deviation ceteris paribus

**set.seed(250)**

**A = round(runif(1000,0,10))**

**ans = 0**

**i = 0**

**for(v in A){**

**ans = ans + (v-mean(A))^2**

**print(ans)**

**i = i+1**

**}**

**sd = sqrt(sum(ans)/(1000-1))**

**print(sd)**

**Part B Grouped discrete series**

1. **Input this series and respresent it as a histogram.**

**• The R function c(v1, · · · , vN ) creates a vector with N values. Use this function to generate the vectors for the marks and number of students having each mark.**

Mark = c(5,8,9,10,11,12,13,14,16)

Number = c(10,12,48,23,24,48,9,7,13)

**• The function plot(data1,data2,type=”h”) is the only one available to generate a histogram from 2 vectors. You can use the command “? plot” to learn more about this function.**

plot(Mark,Number,type="h")

1. **Determine the position and dispersion measures.**

**Measures of Position**

***Mode:***



Mode = Mark[which.max(Number)]

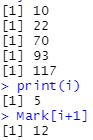
***Mean:***

weighted.mean(Mark,Number)

***Median:***

ans = 0

i = 0

 for(t in Number){

if(ans<sum(Number)/2) {

ans = ans + t

print(ans)

i = i+1

}

}

print(i)

Mark[i+1]

**Measures of Dispersion**

***Range:***



range(Mark)

**-------------------------------------------------------------------------------**

***Variance:***

ans = 0

k = 0

for(c in Mark){

i = 0

k = k+1

 print(k)

print('now is the next in mark')

K = Number[k]

while(i <= K)

{ ans = ans + (c-weighted.mean(Mark,Number))^2

print(ans)

print(i)

i = i+1

}

}

var = sum(ans)/(sum(Number)-1)

print(var)

**-------------------------------------------------------------------------------**

***Standard Deviation:***

ans = 0

k = 0

for(c in Mark){

i = 0

k = k+1

 print(k)

print('now is the next in mark')

K = Number[k]

while(i <= K)

{ ans = ans + (c-weighted.mean(Mark,Number))^2

print(ans)

print(i)

i = i+1

}

}

sd = sqrt(sum(ans)/(sum(Number)-1))

print(sd)

1. **Explain why this series has a bimodal distribution.**

It could be that one group of students is underprepared for the class (perhaps because of a lack of previous classes). The other group may have overprepared.

**Part C Normal distributions**

**The R function rnorm(n,m,sd) generates a sample of n random variables that follow a normal distribution of mean m and standard deviation sd. In this exercise, we propose to generate a sample to simulate the human IQ. Human IQ has a mean value of 100 and a variance of 225.**

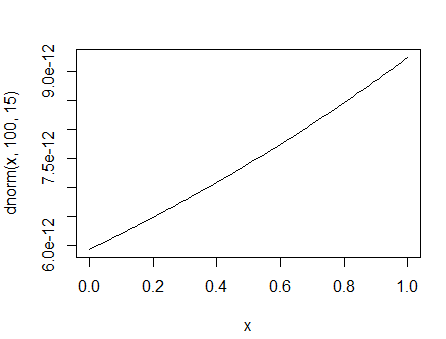
set.seed(250)

x = rnorm(10, mean=100, sd=15)

**1． Use the function “curve(· · ·)” to display the probability density function of this distribution (dnorm(x, µ, σ) for a Gaussian distribution).**

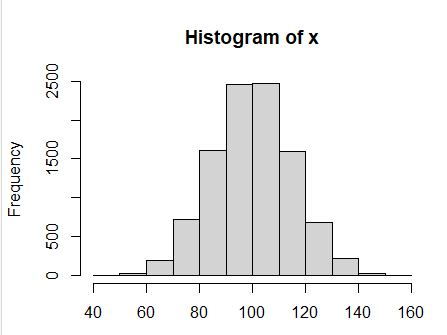
dnorm(x,100,15)

curve(dnorm(x,100,15))



**2． Generate a sample of size 100000 and display its histogram.**

**hist(x)**



**3． Assess the mean value and the standard deviation of your sample. Comment.**

*mean(x)*



*sd(x)*



**4 . Find the percentage of your sample that has an IQ bellow 60.**

v = paste(round(100\*(length(x[x<60])/10000), 2), "%", sep="")

v



**5． Find the percentage of your sample that has an IQ above 130.**

v = paste(round(100\*(length(x[x<60])/10000), 2), "%", sep="")

v



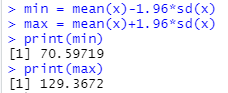
**6． Find the range of values that contains 95 percent of your sample around the mean**

min = mean(x)-1.96\*sd(x)

max = mean(x)+1.96\*sd(x)

print(min)

print(max)



**Part D IQ analysis**

**In this exercise, we want to assess the affect of malnutrition on the human IQ. Knowing that the average IQ is of 100 with a standard deviation of 15, we will modelise the human population with random sample of different sizes and compare them with IQ sample data from people that suffered from malnutrition.**

set.seed(250)

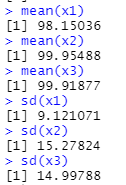
x1 = rnorm(10, mean=100, sd=15)

x2 = rnorm(1000, mean=100, sd=15)

x3 = rnorm(100000, mean=100, sd=15)

**1 .Generate 3 different samples of size 10, 1000 and 100000 with a mean value of 100 and a standard deviation of 15 ( function rnorm(·)).**

**• For each sample, evaluate its mean value and its standard deviation.**

mean(x1)

mean(x2)

mean(x3)

sd(x1)

sd(x2)

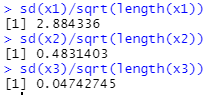
sd(x3)

**• Compare the values you found for the mean and standard deviation with the theoretical values.**

If we have more samples, the SD will be more accurate

**• Calculate the standard error and IC95 of the estimated mean values of each sample.**



sd(x1)/sqrt(length(x1))

sd(x2)/sqrt(length(x2))

sd(x3)/sqrt(length(x3))

SE of x1 is 2.884336

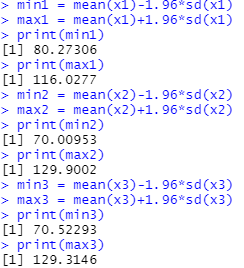
SE of x2 is 0.4831403

SE of x3 is 0.04742745

**-------------------------------------------------------------------------------**

min1 = mean(x1)-1.96\*sd(x1)

max1 = mean(x1)+1.96\*sd(x1)

print(min1)

print(max1)

min2 = mean(x2)-1.96\*sd(x2)

max2 = mean(x2)+1.96\*sd(x2)

print(min2)

print(max2)

min3 = mean(x3)-1.96\*sd(x3)

max3 = mean(x3)+1.96\*sd(x3)

print(min3)

print(max3)

IC95 of x1 is [80.27306, 116.0277]

IC95 of x2 is [70.00953, 129.9002]

IC95 of x3 is [70.52293, 129.3146]

**-------------------------------------------------------------------------------**

**• Comment on your previous results.**

If we have more samples, the IC95 will be more accurate, and the Standard Error will be smaller

**We now want to assess the effect of malnutrition on the IQ. To this end, we will analyze the data from a sample of people that suffered from malnutrition during their childhood.**

**2 .Using the command read.table(file), open the file malnutrition.csv.**

data1=read.table("E:\\addons\\malnutrition.csv")

**3 .Compute the mean and standard deviation of this new sample.**

attach(data1)

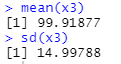
mean(V1)

sd(V1)

**4 .Using the statistical measures at your disposal, what can you conclude on the effect of malnutrition on the IQ when comparing this sample to your previous sample of 100000 elements ?**

**• Compare the mean and standard deviation of both samples.**

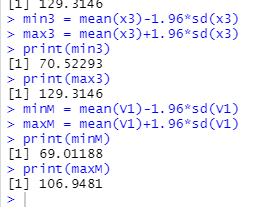
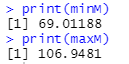




Mean: IQ for malnutrition below 100,000 samples

SD: the smaller the SD, the closer the sample distribution is to the mean, and the variance of IQ for malnutrition is less than the variance of IQ for 100,000 samples.

**• Compute the confidence intervals for both comparisons.**

**• Comment on your results.**

Malnutrition affects IQ, making it lower