Lucas Faijdherbe(2594812) & Ruben van der Ham(2592271) – CS40

**Assignment 2**

**Exercise 2.1**

Exercise 1.6 from the set of additional exercises.

Hand in: Answers with your calculations and explanations.

**Exercise 1.6**

”Where people turn to for news is different for various age groups.” Suppose that

a study conducted on this issue was based on 300 respondents who were between the ages of 46

and 60 and 300 respondents who were over age 60. Of the 300 respondents who were between the

ages of 46 and 60, 82 got their news primary from newspapers. Of the 300 respondents who were

over age 60, 120 got their news primarily from newspapers.

a) Given that a respondent is over age 60, what then is the probability that he or she gets news

primarily from newspapers?

120/300 = 0.4

b) Given that a respondent gets news primarily from newspapers, what is the probability that

he or she is over age 60?

P(he or she is over age of 60|respondent gets new primarily from newspapers)

A = respondent gets new primarily from newspapers

B = he or she is over age of 60

P(A) = 202/600 = 0.3366

P(B) = 300/600 = 0.5

P(B|A) = P(A && B)/P(A)

P(B|A) = (120/600)/0.3366 = 0.59

c) Explain the difference in the results in parts a and b.

Because the probabilities are conditional they are calculated relatively to the other probability. For example if we have 2 types of people: people who own a Ferrari, and people who are rich. Almost all Ferrari owners are rich, but not that many rich people do own a Ferrari. Since they are relative to each other the outcome is different. The same holds for a and b, the order of the conditional probability is different, the outcome aswell.

d) Are the two events, whether the respondent is over age 60 and whether he or she gets news

primarily from newspapers, independent?

If the two events are independent the following statement should hold:

P(A&&B) = P(A) \* P(B)

0.2 = 0.3366\*0.5

0.2 = 0,1683

The above statement is NOT true, therefore the 2 events are dependent.

**Exercise 2.2**

Consider the experiment of tossing a biased coin for which the probability of heads coming up equals 0.7. We assume that multiple coin tosses do not influence each other.

a) Describe the probability distribution, i.e. the possible values and their probabilities, of the random variable ‘number of heads in one coin toss’.

If we have only one coin toss, we have a sample space of Ω{H, T}, and X = number of heads in one coin toss.

X(T) = 0,

X(H) = 1

P(X = 0) = P({T}) = 0.3

P(X = 1) = P({H}) = 0.7

|  |  |  |
| --- | --- | --- |
| X | P(X = x) | Numerical value of P(X = x) |
| 0 | 7/10 | 0.7 |
| 1 | 3/10 | 0.3 |

To do: Implement a graph.

b) Describe the probability distribution, i.e. the possible values and their probabilities, of the random variable ‘number of heads in two coin tosses’.

If we have two coin tosses, we have a sample space of Ω = {HH, HT, TH, TT}, and X = number of heads in two coin tosses.

X(TT) = 0

X(HT) = 1, X(TH) = 1

X(HH) = 2

P(X = 0) = P({TT}) = 0.3 \* 0.3 = 0.09

P(X = 1) = P({HT, TH}) = (0.3 \* 0.7) + (0.3 \* 0.7) = 0.21 + 0.21 = 0.42

P(X = 2) = P({HH}) = 0.7 \* 0.7 = 0.49

0.49+0.42+0.09 = 1

|  |  |  |
| --- | --- | --- |
| x | P(X = x) | Numerical value of P(X = x) |
| 0 | 9/100 | 0.09 |
| 1 | 21/50 (42/100) | 0.42 |
| 2 | 49/100 | 0.49 |

Input graph

c) What is the expected number of heads in two coin tosses?

|  |  |  |  |
| --- | --- | --- | --- |
| x | P(X=x) | Numerical value of P(X = x) | x \* P(X = x) |
| 0 | 9/100 | 0.09 | 0 |
| 1 | 21/50 | 0.42 | 0.42 |
| 2 | 49/100 | 0.49 | 0.98 |

µ = E(X) = 0 \* P(X = 0) + 1 \* P(X = 1) + 2 \* P(X = 2)

µ = E(X) = 0 + 0.42 + 0.98

µ = E(X) = 1.4

The expected number of heads in two coin tosses is 1.4

d) Show analytically that the standard deviation of the variable ‘number of heads in one coin toss’ is approximately equal to 0.46.

First, we have to calculate the expected value of numbers of heads in one coin toss. This is:

µ = E(X) = 0 \* P(X = 0) + 1 \* P(X = 1)

µ = E(X) = 0 + 0.7 (see 2.2a for table + graph)

µ = E(X) = 0.7

Variance of X: 0^2 \* P(X = 0) + 1^2 \* P(X = 1)

0 \* 0 + 1 \* 0.7 = 0.7

0.7 - µ2 = 0.7 – 0,49 = 0,21

So variance of X = 0,21

Standard deviation of X is sqrt(0.21) = 0.458 -> 0.46

e) Which distribution does the random variable ‘the mean number of heads per coin toss after n tosses’ have for large values of n (also give the expectation and standard deviation of the distribution)?

Hand in: Answers with your calculations and explanations.

R-exercises

Hints concerning R:

• Recall that a simple random sample of size n from a set of values x can be drawn in R using the function sample(x,n). By default, the sample is drawn without replacement; by setting the additional parameter replace to TRUE, the sample is drawn with replacement. This function can be used to simulate a die.

• A sample from a certain distribution can be obtained in R with the function rdist(n,par) where dist stands for the name of the distribution, n for the sample size, and par for the re- levant parameters: x=rnorm(50,5,1), x=rexp(25,1), x=runif(30,-1,1), x=rt(10,df=5), x=rchisq(25,df=8). For example, the function rnorm(n,mean,sd) generates a sample of

size n from the normal distribution with expectation mean and standard deviation sd. The parameters of the other distributions are documented in the help-function.

• A normal QQ plot can be obtained with qqnorm(x).

• The command dnorm(u) computes the value of the probability density function of the standard normal distribution in u. For non-standard normal distributions adjust the argu-ments of the function.

• The command lines(x,y) joins the corresponding points in the vectors x and y with line segments. This is useful to draw a curve on top of an existing plot. Similarly, abline(a,b) draws the line ax + b on top of an existing plot. Otherwise specify type="l" in the parameters of the function plot().

• If you want to concatenate text and numbers (which could be useful for instance for titles of plots) you could use the R-function paste().

**Exercise 2.3**

a) Generate the following samples and make for each of the four samples a normal QQ plot:

(i) one sample of size 40 from the t-distribution with 3 degrees of freedom;

(ii) one sample of size 35 from the N (2, 1) distribution;

(iii) one sample of size 45 from the chi-squared-distribution with two degrees of freedom;

(iv) one sample of size 55 from the uniform distribution on the interval [−1, 1].

What can you say about the shapes of these model distributions based on the QQ plots? Comment briefly on each plot.

Hand in: Present the 4 plots concisely using the command par(mfrow=c(2,2)).

b) Answer for each of the data sets below the following question: “Is it reasonable to assume that the data come from a normal distribution?” In each case choose from the two answers: “Obviously not from a normal distribution” or “Normality cannot be excluded”. Base your answer on histograms, boxplots and normal QQ-plots.

(i) klm.txt (delivery time in days of products by Boeing to KLM)

(ii) iqdata2.txt (IQ data)

(iii) dell.txt (trading volumes Dell shares)

(iv) logdell.txt (log trading volume Dell shares)

Hand in: Present for each data set: a suitable histogram, boxplot and QQ-plot, your answer to the question, and a short motivation of this answer. Use the function par(mfrow=c(1,3)) to print the three plots next to each other. Adjust the size of the

**Exercise 2.4**

Study the R-function maxdice from the file function2.txt. Load it by using the

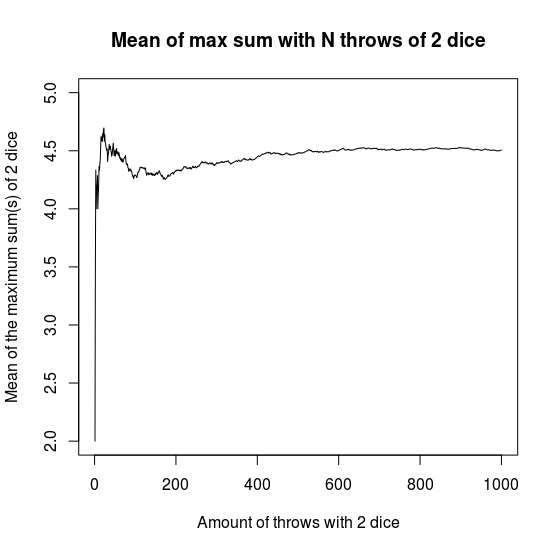
command source("function2.txt").

a) Consider two dice and the random variable ‘the maximum on two dice’ (see Lecture 3). Illustrate the Law of Large Numbers for this random variable by considering ‘the mean of the maximum on two dice’ in n double rolls for different values of n and making a plot similar to the one on slide 24 of the Lecture 3 handout.

The law of large numbers says that for large sample sizes the mean of the random variable values of the all the samples will tend to the expectation of the random variable ,E(X).

Random variable X is “the maximum on 2 dice”.

If we plot the mean of the current outcome of X+ all previous outcomes of X (if any) against the amount of throws of the dice, we obtain the following graph:



As seen in the graph, the value grows fast when N is small. However, when N grows the law of large numbers can be seen clearly. The mean grows until it stabilizes to a mean of approximately 4.5. If we calculate the expected value of X for the same dataset we will get the following:

> table

randomVar probability expectation

[1,] 1 0.027 0.027

[2,] 2 0.083 0.166

[3,] 3 0.127 0.381

[4,] 4 0.192 0.768

[5,] 5 0.262 1.310

[6,] 6 0.309 1.854

> sum(expectation)

[1] 4.506

As described before “ The mean grows until it stabilizes to a mean of approximately 4.5”. The expectation of 4.51 is accurately close to where the mean tends to grow to.

b) Use the function maxdice to find an approximate value of expectation of the random variable ‘the maximum on 5 dice’ and the probability of the event ‘the maximum on 2 dice is a 3’.

To get an approximate value of expectation of the maximum of 5 dice we run the following commands:

*source("function2.txt")*

*maximaOf5Dice = maxdice(1000000,5)*

*probability = c()*

*randomVar = c()*

*uniques = unique(maximaOf5Dice,incomparables = FALSE)*

*for(i in sort(uniques)){*

*probability = c(probability,length(which(maximaOf5Dice == i))/length(maximaOf5Dice))*

*randomVar = c(randomVar,i)*

*}*

*expectation = c(randomVar\*probability)*

To find the expectation of the variable we run:

> sum(expectation)

[1] 5.432212

The expectation of the maximum of 5 dice is approximately 5.43 with a dataset of 1000000 throws.

The probability of the event “the maximum of 2 dice is a 3 is obtained by the following sets of commands:

First we use the same commands as for the previous question, except for the maximaOf2Dice function. This will be:

maximaOf2Dice = maxdice(1000000,2)

Then:

> cbind(randomVar,probability)

randomVar probability

[1,] 1 0.027468

[2,] 2 0.083475

[3,] 3 0.138614

[4,] 4 0.194948

[5,] 5 0.250006

[6,] 6 0.305489

> sum(probability)

[1] 1

> probability[[3]]

[1] 0.138614

The probability of the event “the maximum of 2 dice is a 3” is around 0.139.

c) Use the function maxdice to illustrate the Central Limit Theorem for the random variable ‘the mean maximum value of two dice rolls after n double rolls’ for the present context of two dice rolls graphically, by making 4 plots similar to the 4 plots on slide 15 of the Lecture 4 handout.

d) Explain briefly why the 4 plots of part c) illustrate the Central Limit Theorem in the present context. Hand in: Properly described plots (part a and c), answers with motivation (parts b and d).