

# Statistical Methods Fall 2021

## Assignment 2: Probability, Normality, CLT and Law of Large Numbers

**Deadline:** see Canvas

### *Topics of this assignment*

The exercises below concern topics that were covered in Lectures 2, 3 and 4: probability, including the Law of Total Probability and Bayes' Theorem, random variables, model distributions, the Central Limit Theorem (CLT), and the Law of Large Numbers (see the respective sections in Chapters 3, 4 and 5 of the book and the handouts of Lectures 2, 3, 4). Before making the assignment, study these topics.

### *How to solve and present the solutions to the exercises?*

See the first page of [Assignment 1](#) and the [example file](#) about the required format.

### Theoretical exercises

**Exercise 2.1** Take a look at Exercise 1.3 from the set of additional exercises ([link](#)) and consider the same setting here.

- What is the probability that a random person, which does the test, gets a positive result? Explain the difference between this probability and the probability you are asked to calculate in the Exercise 1.3 from the set of additional exercises.
- Solve Exercise 1.3 from the additional exercises ([link](#)).
- Are the two events that *a person has cancer* and that *the test is positive* dependent? Does the fact that a test result was positive increase the risk of having cancer, when compared to the probability that a random individual from the population has cancer?

*Hand in: Answers with your calculations, explanations, and motivations.*

*Also, mention which formulas you are using and state the formulas you are using.*

*Warning: it is possible that you receive 0 points for certain parts of the exercise if you don't give a motivation/explanation!*

**Exercise 2.2** Suppose you walk to a bus stop to catch a bus; busses of the line you take arrive there every 15 minutes. You don't look at your watch and arrive at the bus stop totally at random. For simplicity, we assume that all arriving times (you/the busses) are rounded down to whole minutes. When you and the bus arrive simultaneously, you will be able to catch it.

- Describe the probability space that models the above-described waiting time experiment, i.e. sample space and the related probabilities.

Suppose now that you still start walking towards the bus stop at a completely random time; but that you *do* look at your watch at the beginning of your journey. Knowing your walking pace and the distance to the bus stop, you know by how many minutes you will miss the previous bus. If you would miss the bus by 4 or less minutes (i.e. hypothetically, the waiting time would be 11 minutes or more), you decide to hurry up; in this case, you will still be able to catch it. Otherwise, if you would miss the bus by 5 or more minutes (i.e. waiting times 10 minutes or less), you simply go on with your normal walking pace.

- Let the random variable  $X$  model your waiting time at the bus station in the just-described situation. Calculate the probability that you will need to wait for at least 5 minutes.
- Calculate the expectation of  $X$ .
- Calculate the variance of  $X$ .

- e) Suppose you walk to the bus stop in the previously described way year 160 times per year. Denote by  $X_1, \dots, X_{160}$  your waiting times on all these days; assume they are independent of each other. Describe the (approximate) distribution of your average waiting time  $\bar{X}_{160} = \frac{1}{n} \sum_{i=1}^{160} X_i$  across the whole year.

*Note: only if you are not able to derive the expectation in c) you may continue in d) and e) with the **wrong** value  $E(X) = 5$ , and if you are not able to derive the variance in d), you may continue in e) with the **wrong** value  $\text{var}(X) = 4$ .*

*Make formal calculations in all parts and do not apply mere reasoning!*

*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  $\mathbf{x}$  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 relevant 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)`, histograms with `hist(x)`, and boxplots with `boxplot(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 arguments 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  $a + bx$  on top of an existing plot. Otherwise specify `type="l"` in the parameters of the function `plot()`.
- To concatenate text and numbers (useful for titles of plots) use the *R*-function `paste()`.
- Use the command `set.seed(...)` to make your results based on the generated samples reproducible.

### Exercise 2.3

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

- one sample of size 115 from the chisquared distribution with degrees of freedom 2;
- one sample of size 105 from the  $t$ -distribution with 4 degrees of freedom.

Evaluate the usefulness of the normal distribution as a model distribution for both samples based on the QQ plots. Comment briefly on each plot and each peculiarity.

*Hint: in a normal QQ plot, data are compared to a theoretical normal distribution.*

*Hand in: the 2 plots concisely presented using the command `par(mfrow=c(1,2))`, and your answers.*

- b) Generate the following samples and make for each of them a histogram and a boxplot.

- one sample of size 115 from the chisquared distribution with degrees of freedom 2;
- one sample of size 105 from the  $t$ -distribution with 4 degrees of freedom.

Relate the peculiarities visible in the histograms to what you see in the corresponding boxplots, and describe your findings. In particular, address the heaviness of the tails, symmetry, and outliers.

*Hand in: the 4 plots concisely presented using the command `par(mfrow=c(2,2))`, and your answers.*

- c) 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.

Also, for each dataset, point out the peculiarities of each sample by comparing the histogram, boxplot, and QQ-plot with each other. Indicate whether you detect (some/all) peculiarities in some/all of these diagnostic plots.

- (i) **titanic3.csv**: Data about many passengers of the Titanic; we are going to analyse the passengers’ ages.

More information on <http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/titanic.html>

*Hints: use `titanic <- read.csv("titanic3.csv")` to read the dataset.*

*Then use `titanic$age` to obtain the passengers’ ages.*

- (ii) **diabetes.csv**: Data from a study about understanding the prevalence of obesity, diabetes, and other cardiovascular risk factors; we are going to analyse the individuals’ total cholesterol values.

More information on <http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/diabetes.html>

*Hints: use `diabetes <- read.csv("diabetes.csv")` to read the dataset.*

*Then use `diabetes$chol` to obtain the individuals’ total cholesterol values.*

- (iii) **vlbw.csv**: Data about newborn babies with very low birth weight; we are going to analyse the babies’ birth weights.

More information on <http://biostat.mc.vanderbilt.edu/wiki/pub/Main/DataSets/vlbw.html>

*Hints: use `vlbw <- read.csv("vlbw.csv")` to read the dataset.*

*Then use `vlbw$bwt` to obtain the babies’ birth weights.*

*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. Also, the explanation on whether you identified some/all peculiarities in some/all plots for each dataset. Use the function `par(mfrow=c(1,3))` to print the three plots next to each other. Adjust the size of the figure so that the ratio becomes approximately 1:3, and each plot is more or less square.*

**Exercise 2.4** Study the R-function `diffdice` from the file `function02.txt`. Load it by using the command `source("function02.txt")`.

- a) Consider two dice and the random variable ‘the absolute difference of two die rolls’. Illustrate the Law of Large Numbers for this random variable by considering ‘the mean of the absolute difference of two die rolls’ in  $n$  trials for different values of  $n$  and making a plot similar to the one on slide 24 of the Lecture 3 handout.

*Hint:  $n$  trials means that both of two dice are rolled  $n$  times, and for each of the trials the absolute difference is calculated.*

*You may use (without proof) that the theoretical expected value of the absolute difference is about 1.9444.*

- b) Use the function `diffdice` to find an approximate value of expectation of the random variable ‘the absolute difference of two die rolls’ and the probability of the event ‘the absolute difference of two die rolls is 3’.

- c) Use the function `diffdice` to graphically illustrate the Central Limit Theorem for the random variable ‘the mean absolute difference of two dice rolls after  $n$  trials’, by making 4 plots similar to the 4 plots on slide 15 of the Lecture 4 handout.

*You may use (without proof) that the standard deviation of the **absolute difference of two** dice rolls is approximately 1.4326.*

- 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).*

# Assignment 1

## Statistical Methods 2021

Project Group 3:

Caspar M. S. Grevelhörster (2707848)

Zülfiye Tülin Manisa (2664758)

Sidharth R. Singh (2708756)

Vrije Universiteit Amsterdam,

De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands

### Exercise 1

a) The probability of someone having a positive test is the probability of someone getting a positive test and having cancer in addition to the probability that they have a positive test and do not have cancer: It is given that  $P(\text{Cancer}) = 0.004$ ,  $P(\text{NoCancer}) = 0.996$ ,  $P(\text{PositiveTest}|\text{Cancer}) = 0.95$ , and  $P(\text{PositiveTest}|\text{NoCancer}) = 0.05$ . So  $P(\text{PositiveTest} \cap \text{Cancer}) = P(\text{Cancer}) * P(\text{PositiveTest}|\text{Cancer}) = 0.004 * 0.95 = 0.0038$  and  $P(\text{PositiveTest} \cap \text{NoCancer}) = P(\text{NoCancer}) * P(\text{PositiveTest}|\text{NoCancer}) = 0.996 * 0.05 = 0.0498$ . Finally,  $P(\text{PositiveTest}) = P(\text{PositiveTest} \cap \text{Cancer}) + P(\text{PositiveTest} \cap \text{NoCancer}) = 0.0038 + 0.0498 = 0.0536$ . Exercise 1.3 asks to calculate the probability of someone having cancer when being tested positive, so:  $P(\text{Cancer}|\text{PositiveTest})$  which is clearly different from the probability of someone just being tested positive given above.

b) We can calculate if someone has cancer, given that they have a positive test with Bayes Theorem.  $P(\text{Cancer}|\text{PositiveTest}) = \frac{P(\text{PositiveTest}|\text{Cancer}) * P(\text{Cancer})}{P(\text{PositiveTest})}$ . All expressions to the right side of the equation are either given or were calculated in a):  $P(\text{Cancer}|\text{PositiveTest}) = \frac{0.95 * 0.004}{0.0536} \approx 0.0709$ .

c) The two events that a person has cancer and that that person is tested positive are dependent, because the test result heavily depends on whether a person has cancer or not. We also see that they are dependent because  $P(\text{Cancer}|\text{PositiveTest}) \approx 0.0709$  and  $P(\text{Cancer}) = 0.004$ . If the events were independent  $P(\text{Cancer}|\text{PositiveTest})$  would be equal to  $P(\text{Cancer})$ , but clearly it is not. This also means that a positive test increases the risk of having cancer, because the probability of someone having cancer given a positive test  $P(\text{Cancer}|\text{PositiveTest}) \approx 0.0709$  is larger than a random person having cancer  $P(\text{Cancer}) = 0.004$ .

### Exercise 2

a) It is possible to arrive at the bus stop at every (full) minute, meaning that it is possible to wait from 0 to at most 14 minutes for the bus. Thus, the sample space for the waiting time experiment is  $\Omega = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14\}$ . Each outcome is equally likely, so:  $P(\omega) = \frac{1}{15}$  for all  $\omega$  in  $\Omega$ .

b) Let  $A = \{\text{waiting 5 or more minutes}\} = \{5, 6, 7, 8, 9, 10\}$ . There are 6 outcomes in event  $A$  (it is not possible to wait for 11 or more minutes in this scenario), which are all equally likely to occur. Since at the moment of looking at the watch there can be 15 hypothetical waiting times, each outcome in  $A$  has the probability of  $\frac{1}{15}$ . Thus,  $P(A) = 6 * \frac{1}{15} = \frac{2}{5}$ . Thus the probability for waiting 5 or more minutes in this scenario is 0.4.

c) We catch the bus when we wait for 0 minutes. Let  $A = \{\text{catch the bus}\} = \{0\}$ . We miss the bus when we wait for 1 up to 10 minutes, because in this scenario waiting for 11 to 14 minutes is not possible (since 5 minutes were already spent walking). Let  $B = \{\text{miss the bus}\} = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ . Each of the 10 outcomes in event  $B$  is equally likely to occur with a probability of  $\frac{1}{15}$  (see b)). Thus,  $P(B) = 10 * \frac{1}{15} = \frac{2}{3}$ . Since the events  $A$  and  $B$  are disjoint,  $P(A \cup B) = P(A) + P(B)$ . Furthermore, we can either catch the bus or miss the bus. Thus,  $P(A \cup B) = 1$ . Thus,  $P(A) + P(B) = 1$ . If we solve this equation for  $P(A)$  we get:  $P(A) = 1 - P(B) = 1 - \frac{2}{3} = \frac{1}{3}$ . This means that catching the bus and thus waiting for 0 minutes has a probability of  $\frac{1}{3}$ . With this information it is possible to calculate the expected value using  $E(X) = \sum_{i=1}^k x_i * P(X = x_i) = 0 * \frac{1}{3} + 1 * \frac{1}{15} + 2 * \frac{1}{15} + 3 * \frac{1}{15} + 4 * \frac{1}{15} + 5 * \frac{1}{15} + 6 * \frac{1}{15} + 7 * \frac{1}{15} + 8 * \frac{1}{15} + 9 * \frac{1}{15} + 10 * \frac{1}{15} = \frac{11}{3} \approx 3.67$ .

d) We can calculate the variance of  $X$  with the mean  $\mu = \frac{11}{3}$  for  $X$  that we calculated in c):  $Var(X) = \sum_{i=1}^k (x_i - \mu)^2 * P(X = x_i) = \sum_{i=1}^k (x_i - \frac{11}{3})^2 * P(X = x_i) \approx 12.22$ .

e) According to the Central Limit Theorem, the distribution will be approximately a normal distribution with mean  $\mu = 12.22$  and standard deviation  $\sigma \approx 0.276$  described by  $\bar{X}_{160} \sim N(\frac{11}{3}, \frac{12.22}{160})$ .

### Exercise 3

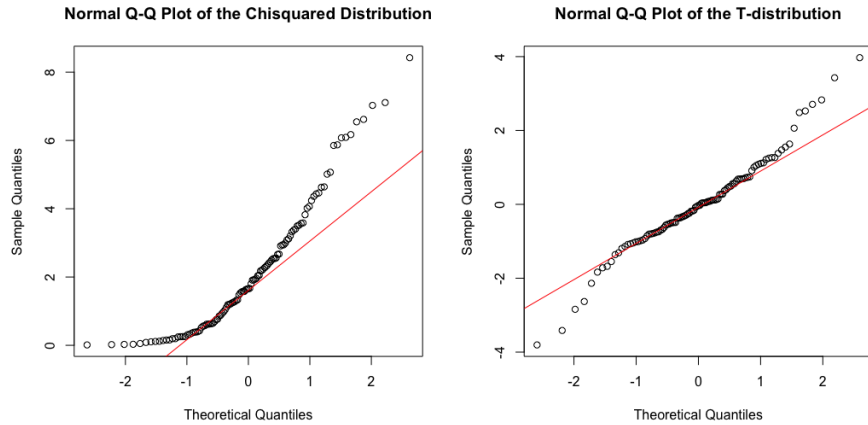
a)

i) On the left side of figure 1 the chisquared distribution is plotted, we can see that this QQ-plot is right

skewed and deviates from the red line. From this we can conclude that the normal distribution is not useful to model the distribution of this sample.

ii) On the right side of the figure the t-distribution is plotted, one both ends the tails are heavy and clearly deviate from the red line. Thus, we can conclude that the normal distribution is not useful to model the distribution of this sample.

Fig. 1: The Normal Q-Q Plots for chisquared distribution (left) and t-distribution(right).

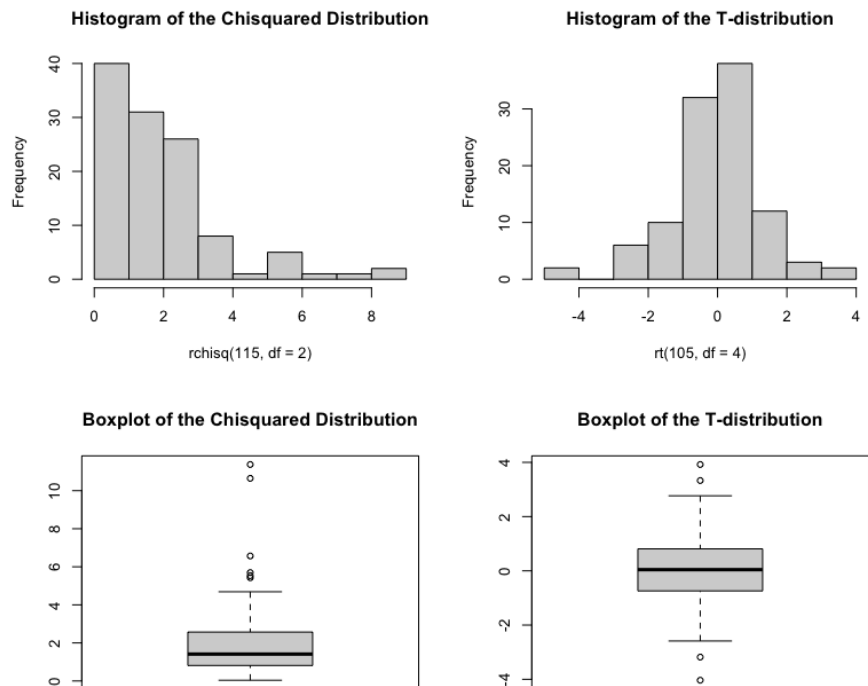


b)

i) The histogram of the chisquared distribution has a long tail to the right (right skewed). This is also seen in the boxplot right underneath the histogram. The median is not centered, there are long whiskers on the upper (positively skewed) and lastly outliers on the upper side.

ii) The histogram of the t-distribution, although, not perfect shows a bell shape. The boxplot looks nearly symmetrical and the median is roughly in the middle of the interquartile range box. Additionally, the whiskers on both sides are of approximately the same size, with the same number of outliers beyond them.

Fig. 2: The Histogram and Boxplot for chisquared distribution (left) and t-distribution(right).

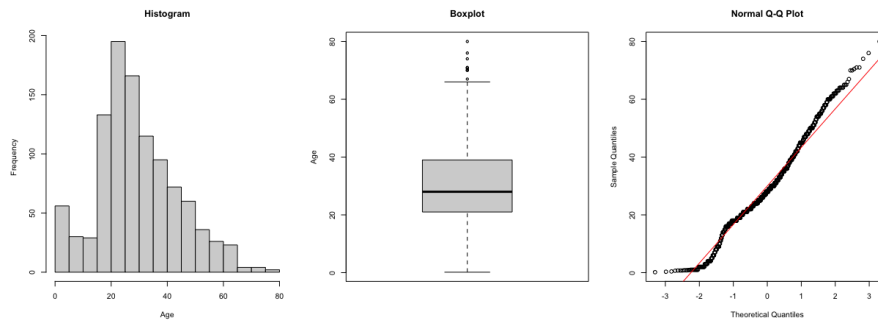


c)

i) Obviously not from a normal distribution:

The histogram has a tail to the right (right-skewed), this is again visible in the boxplot where the upper whiskers are longer (positively skewed). The median is not centered and some outliers on the upper side are present. The QQ-plot shows a relatively straight line with a light tail.

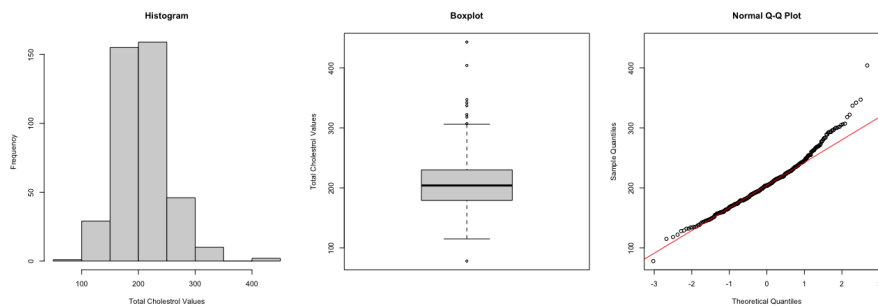
Fig. 3: Histogram, Boxplot and Q-Q Plot of the Titanic Dataset



**ii) Normality cannot be excluded:**

The histogram is thin tailed and shows a slight bell shape, although, it is too peaked in the center. The boxplot shows a nearly centered median, the upper side has a slightly longer whisker and more outliers than the bottom. In the QQ-plot the outliers can again be seen on the upper side. It does, however, show a relatively straight line.

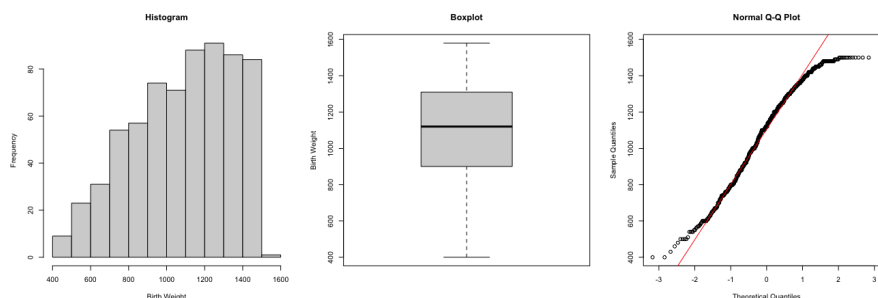
Fig. 4: Histogram, Boxplot and Q-Q Plot of the Diabetes Dataset



**iii) Obviously not from a normal distribution**

The histogram has a tail to the left (left skewed), this is again seen in the boxplot, where the bottom whisker is almost twice as long as the upper whisker (negatively skewed). Additionally, we can see that the median is only slightly higher than the center. In the QQ-plot a heavy tail at the upper end can be seen, again confirming it is left skewed.

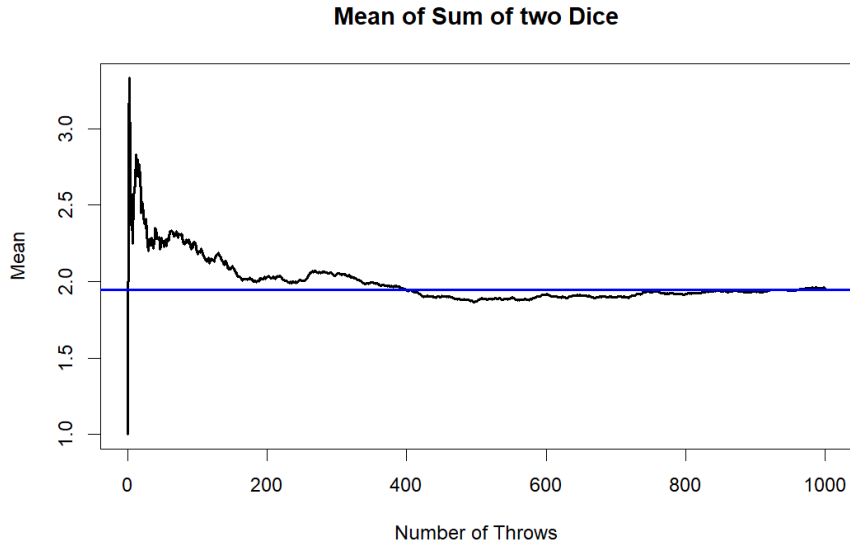
Fig. 5: Histogram, Boxplot and Q-Q Plot of the Very Low Birth Weight (VLBW) Dataset



## Exercise 4

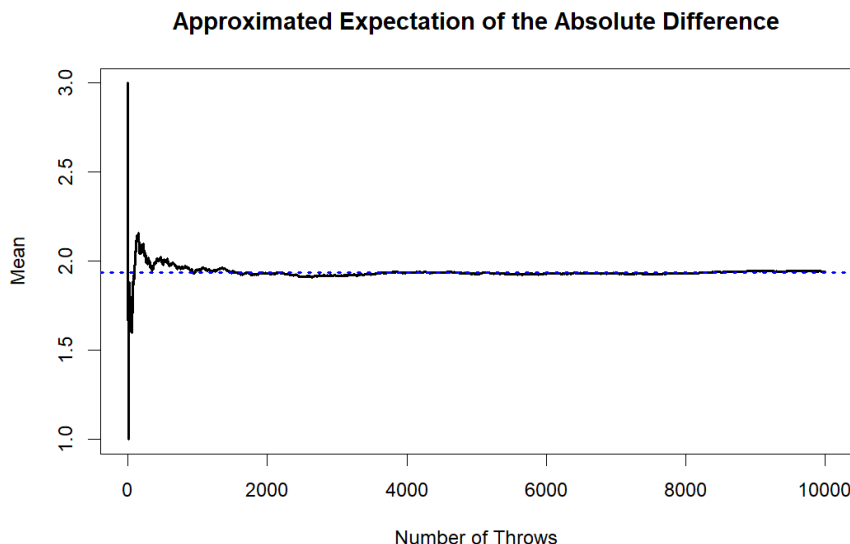
a) The following diagram was generated using R, the code can be found in the appendix. The blue line represents the theoretically expected value of the absolute difference of 1.9444. The mean of the dice throws are depicted by the black line and illustrate the Law of Large Numbers since the line converges towards the blue line the higher the number of throws.

Fig. 6: Illustration of the Law of Large Numbers



b) In order to find an approximate value of expectation of the random variable "the absolute difference of two die rolls", a linear regression method was used to fit a linear model to the values of the mean. The linear model can be seen as a blue dotted line in the following diagram. The y-value of this model was found to be 1.936.

Fig. 7: Illustration of the Law of Large Numbers

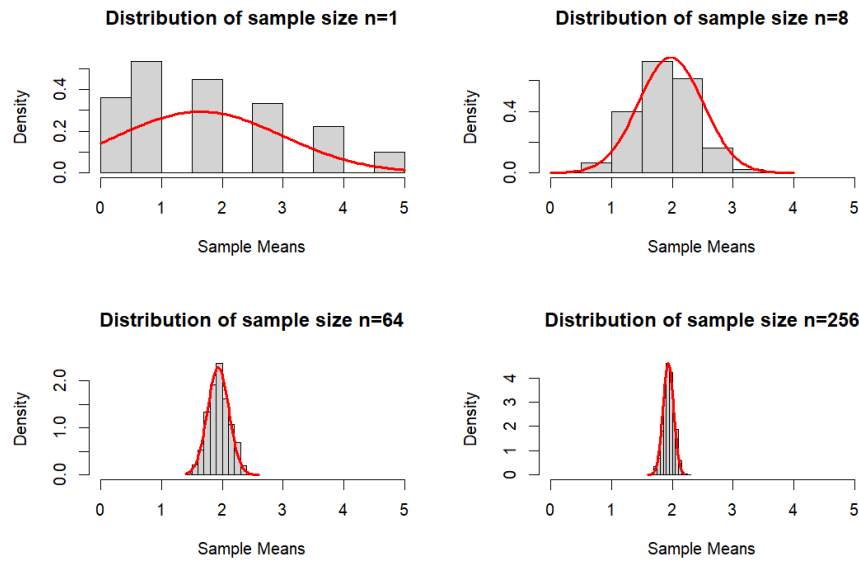


In contrast to part a), the sample space was increased to 10000 throws to give a more accurate approximation. The probability of the event "the absolute difference of two die rolls is 3" was computed in R and found to be 0.1618 for this graph. The utilized code as well as the compiled numerical results can be found in the appendix.

c) The following graph depicts the central limit theorem for the random variable "the mean absolute difference of two dice rolls after n trials". The graphs were plotted using R, the utilized code can be found in the appendix.

d) The four plots illustrate the central limit theorem. The latter describes that with a sufficiently large sample size, the distribution of the sample means will approximately be normally distributed. With increasing

Fig. 8: Illustration of the Central Limit Theorem



sample size, the graphs are also more accurately taking form of a normal distribution. Thus, they depict the central limit theorem.



# 1 Appendix

## 1.1 Code of exercise 3 a) i)

```
set.seed(12)
par(mfrow = c(1, 2))
qqnorm(rchisq(115, df = 2), main = "Normal Q-Q Plot of the Chisquared Distribution")
qqline(rchisq(115, df = 2), col = "red")
qqnorm(rt(105, df = 4), main = "Normal Q-Q Plot of the T-distribution")
qqline(rt(105, df = 4), col = "red")
```

## 1.2 Code of exercise 3 b)

```
set.seed(3)
par(mfrow = c(2, 2))
hist(rchisq(115, df = 2), main = "Histogram of the Chisquared Distribution")
hist(rt(105, df = 4), main = "Histogram of the T-distribution")
boxplot(rchisq(115, df = 2), main = "Boxplot of the Chisquared Distribution")
boxplot(rt(105, df = 4), main = "Boxplot of the T-distribution")
```

## 1.3 Code of exercise 3 c) i)

```
par(mfrow = c(1, 3))
titanic = read.csv("titanic3.csv")
hist(titanic$age, main = "Histogram", xlab = "Age")
boxplot(titanic$age, main = "Boxplot", ylab = "Age")
qqnorm(titanic$age, main = "Normal Q-Q Plot")
qqline(titanic$age, col = "red")
```

## 1.4 Code of exercise 3 c) ii)

```
diabetes <- read.csv("diabetes.csv")
hist(diabetes$chol, main = "Histogram", xlab = "Total Cholestrol Values")
boxplot(diabetes$chol, main = "Boxplot", ylab = "Total Cholestrol Values")
qqnorm(diabetes$chol)
qqline(diabetes$chol, col = "red")
```

## 1.5 Code of exercise 3 c) iii)

```
vlbw <- read.csv("vlbw.csv")
hist(vlbw$bwt, main = "Histogram", xlab = "Birth Weight")
boxplot(vlbw$bwt, main = "Boxplot", ylab = "Birth Weight")
qqnorm(vlbw$bwt)
qqline(vlbw$bwt, col = "red")
```

## Code of exercise 4 a)

```

set.seed(2)
diffdice = function(n = 0) {
  replicate(n, abs(diff(sample(6, 2, replace = T))))
}
differences = diffdice(1000)
means = c()
for (i in 1:length(differences)) {
  means[i] = mean(differences[1:i])
}
plot(means, type = "l", lwd = 2, main = "Mean of Sum of two Dice",
     ylab = "Mean", xlab = "Number of Throws")
abline(h = 1.9444, col = "blue", lwd = 2)

```

#### Code of exercise 4 b)

```

set.seed(6)
nr_throws = 10000
diffdice = function(n = 10) {
  replicate(n, abs(diff(sample(6, 2, replace = T))))
}

differences = diffdice(nr_throws)
data = table(differences)
p_of_3s = data[names(data) == 3]/nr_throws

means = c()
for (i in 1:length(differences)) {
  means[i] = mean(differences[1:i])
}

plot(means, type = "l", lwd = 2, main = "Approximated Expectation of the Absolute Difference",
     ylab = "Mean", xlab = "Number of Throws")

approximated_expectation = lm(means ~ 1)
abline(approximated_expectation, col = "blue", lwd = 2,
      lty = "dotted")
approximated_expectation

##
## Call:
## lm(formula = means ~ 1)
##
## Coefficients:
## (Intercept)
##      1.936

p_of_3s

##      3
## 0.1618

```

#### Code of exercise 4 c)

```

set.seed(3)
par(mfrow = c(2, 2))
nr_samples = 1000
diffdice = function(n = 10) {
  replicate(n, abs(diff(sample(6, 2, replace = T))))
}

```

```

}

samples_1 = c()
samples_8 = c()
samples_64 = c()
samples_256 = c()

for (i in 1:nr_samples) {
  samples_1[i] = mean(diffrdice(1))
}
for (i in 1:nr_samples) {
  samples_8[i] = mean(diffrdice(8))
}
for (i in 1:nr_samples) {
  samples_64[i] = mean(diffrdice(64))
}
for (i in 1:nr_samples) {
  samples_256[i] = mean(diffrdice(256))
}

rnorm_1 <- rnorm(100, mean = mean(samples_1), sd = sd(samples_1))
hist(samples_1, xlim = c(0, 5), main = "Distribution of sample size n=1",
  ylab = "Density", xlab = "Sample Means", probability = TRUE)
curve(dnorm(x, mean = mean(rnorm_1), sd = sd(rnorm_1)),
  add = TRUE, col = "red", lwd = 2)

rnorm_8 <- rnorm(100, mean = mean(samples_8), sd = sd(samples_8))
hist(samples_8, xlim = c(0, 5), main = "Distribution of sample size n=8",
  ylab = "Density", xlab = "Sample Means", probability = TRUE)
curve(dnorm(x, mean = mean(rnorm_8), sd = sd(rnorm_8)),
  add = TRUE, col = "red", lwd = 2, xlim = c(0, 4))

rnorm_64 <- rnorm(100, mean = mean(samples_64), sd = sd(samples_64))
hist(samples_64, xlim = c(0, 5), main = "Distribution of sample size n=64",
  ylab = "Density", xlab = "Sample Means", probability = TRUE)
curve(dnorm(x, mean = mean(rnorm_64), sd = sd(rnorm_64)),
  add = TRUE, col = "red", lwd = 2, xlim = c(1.4,
    2.6))

rnorm_256 <- rnorm(100, mean = mean(samples_256), sd = sd(samples_256))
hist(samples_256, xlim = c(0, 5), main = "Distribution of sample size n=256",
  ylab = "Density", xlab = "Sample Means", probability = TRUE)
curve(dnorm(x, mean = mean(rnorm_256), sd = sd(rnorm_256)),
  add = TRUE, col = "red", lwd = 2, xlim = c(1.6,
    2.2))

```