Assignment 2

Hongyu He (2632195) & Bruno Hoevelaken (2645065)

Group CS 6

& Theoratical exercises

Exercise 2.1 (Additional exercises 1.3)

- ***"+ / -"** : positive / negative results;*
- ***"disease / !disease"**: have / don't have cancer*

a)

$$\begin{aligned} P(+) &= P(+ \mid disease) \cdot P(disease) + P(+ \mid !disease) \cdot P(!disease) \\ &= 0.95 \cdot 0.004 + (1 - 0.95) \cdot (1 - 0.004) \\ &= 0.054 \end{aligned}$$

The probability `P(+)` calculated here is called the **belief** in statistics, representing the possibility of a certain event happening, whilst the probability `P(disease | +)` calculated in the additional exercises is called **conditional probability** illustrating the possibility of a certain event happening based on a **revised model** tunned by new information (an event has already happened).

b)

$$\begin{aligned} P(disease \mid +) &= P(disease) \cdot \frac{P(+ \mid disease)}{P(+)} \\ &= 0.95 \cdot \frac{0.004}{0.054} = \frac{19}{268} \\ &= 0.071 \end{aligned}$$

c)

(1) Yes, they are dependant because

$$\begin{aligned} P(disease \cap +) &= P(disease) \cdot P(+ \mid disease) \\ &= 0.004 \cdot 0.95 \end{aligned}$$

is no the same as

$$P(disease) \cdot P(+) = 0.004 \cdot 0.054$$

(2) Yes, a test result was positive increase the risk of having cancer because

$$P(disease \mid +) = 0.071 > P(disease) = 0.004$$

Exercise 2.2

a)

- *Let* $\boldsymbol{X}$ *denote the random variable 'outcome of two 8-sided die-rolls';*
- *Let* $\boldsymbol{x_i}$ *(where i = 1, 2, . . . , 64) denote the possible outcomes*

Because multiple rolls do not influence each other, we have

1)

$x_1 = (1,1);\ x_2 = (1,2);\ x_3 = (1,3);\ x_4 = (1,4);\ x_5 = (1,5);\ x_6 = (1,6);\ x_7 = (1,7);\ x_8 = (1,8);$

$x_9 = (2,1);\ x_{10} = (2,2);\ x_{11} = (2,3);\ x_{12} = (2,4);\ x_{13} = (2,5);\ x_{14} = (2,6);\ x_{15} = (2,7);\ x_{16} = (2,8);$

...

$x_{57} = (8,1);\ x_{58} = (8,2);\ x_{59} = (8,3);\ x_{60} = (8,4);\ x_{61} = (8,5);\ x_{62} = (8,6);\ x_{63} = (8,7);\ x_{64} = (8,8);$

2)

$P(X = x_1) = P(X = x_2) = \ ... \ = P(X = x_{64}) = \frac{1}{64}$

b)

- *Let* $\boldsymbol{Y}$ *denote the random variable 'number of 1's in two 8-sided die rolls';*
- *Let* $\boldsymbol{y_j}$*(where j = 1, 2, 3) denote the possible outcomes*

j	y_j	$P(Y=y_j)$
1	0	49/64
2	1	7/32
3	2	1/64

c) The expectation of the random variable Y:

$E(Y) = 0 * \frac{49}{64} + 1 * \frac{7}{32} + 2 * \frac{1}{64} = \frac{1}{4}$

d)

Proof:

$Var(Y) = 0^2 * \frac{49}{64} + 1^2 * \frac{7}{32} + 2^2 * \frac{1}{64} - E(Y)^2 = \frac{7}{32}$

e)

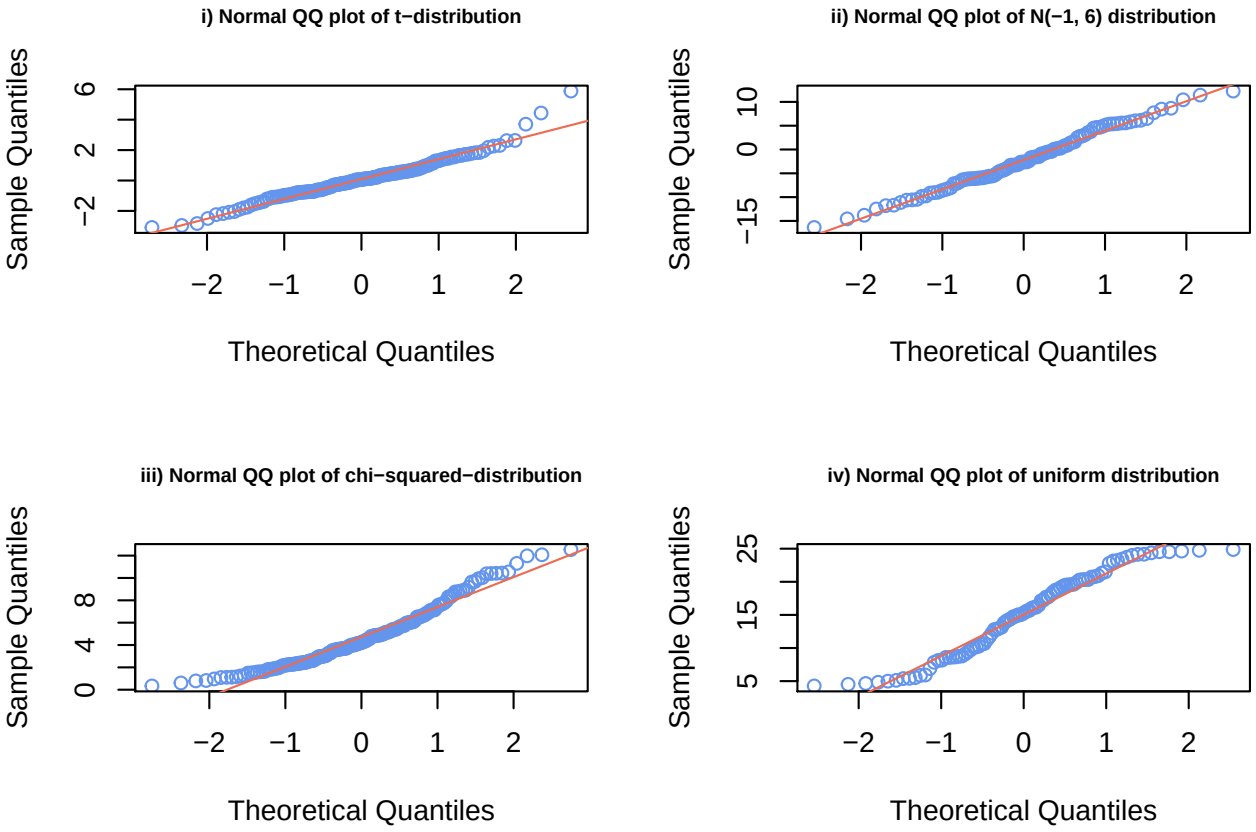
Since the sample size is much greater than 30, according to the **central limit theorem**, the random variables X_i approximately have a normal distribution with an expectation μ of 1/4 and a standard deviation of $\frac{\sigma}{\sqrt{n}}$, which is

$$N(\frac{1}{4}, \frac{7/32}{n})$$

& R-exercises

Exercise 2.3

a)



Observations:

i) The left part of the plot is below the QQ-line and the right part is above the QQ-line. Therefore, we conclude that both the left tail and right tail of the *t-distribution* is **heavier** than that of the *standard normal distribution*;

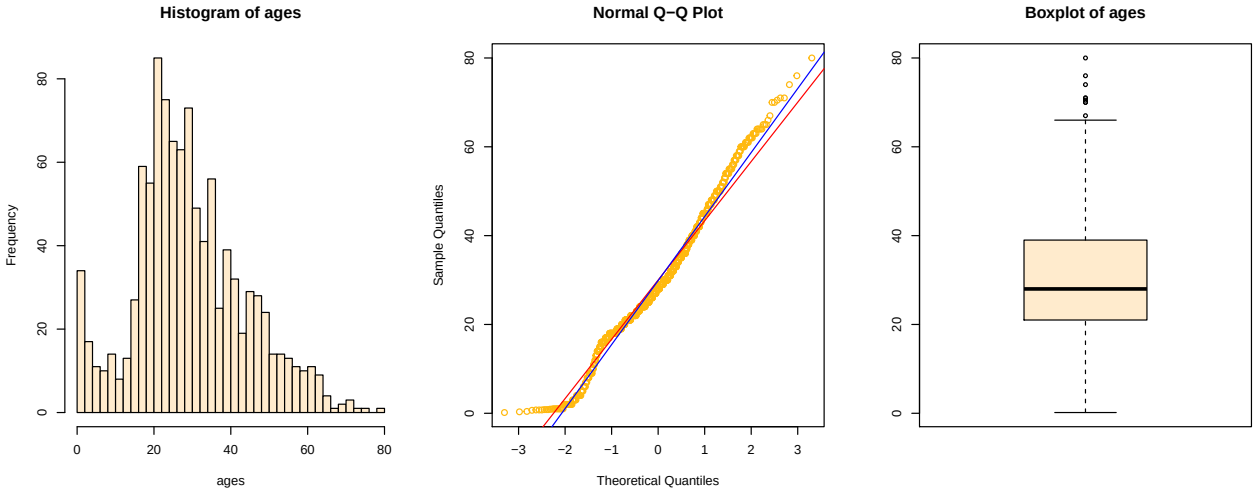
ii) As the top-right figure demonstrates, because N(-1, 6) and *the standard normal distribution* are in the same **location-scale family**, the QQ-line of the N(-1, 6) distribution **approximately corresponds to** the QQ-line of *the standard distribution* very well;

iii) Both the left part and the right part of the plot are above the QQ-line. Therefore, we conclude that the left tail of *the chi-squared-distribution* is **lighter** than that of *the standard normal distribution*, whilst the right tail of the chi-squared-distribution is **heavier** than that of the standard normal distribution;

iv) The left part of the plot is above the QQ-line and the right part is below the QQ-line. Therefore, we conclude that the left tail of *the uniform distribution* is **heavier** than that of *the standard normal distribution*, whilst the right tail of *the uniform distribution* is **lighter** than that of *the standard normal distribution*.

b)

(i) ***titanic3.csv*** **dataset:**



- —— *QQ-line*
- —— $y = \mu x + \sigma^2$ *family*

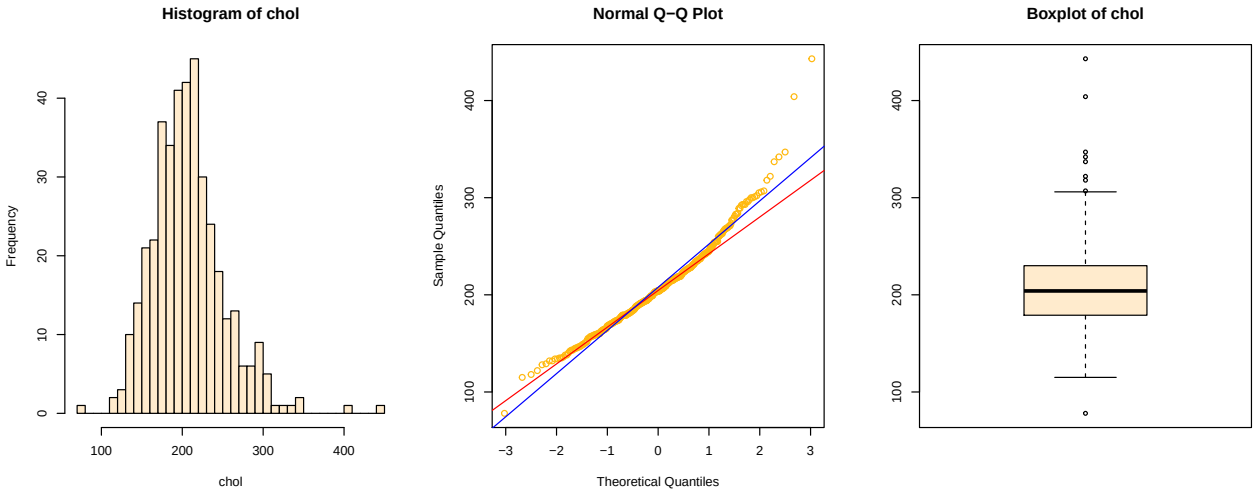
(1) According to the plots above, we conclude that the answer for the first question is **Normality cannot be excluded** for the following reasons:

a. the histogram demonstrates a **roughly bell shape**;
b. as the normal Q-Q plot shows, although the right tail is heavier and right tail is lighter, the major quantiles **approximately correspond to the Q-Q line** and the $\mathbf{y = \mu x + \sigma^2}$ family as well;
c. as we can see from the boxplot, excluding outliers, the major body of the date, ranging from the 1st quartile to the 3rd quartile, lies in the middle of the range, as well as the median value. Thus, the distribution is **approximately symmetric.**

(2) **Peculiarities:**

a. as the histogram shows, the left part of the distribution is not only heavier but also has **an abnormal decreasing trend**;
b. the location of this distribution is dragged left (**left-skewed**) because of too many children.

(ii) ***diabetes.csv*** **dataset:**



- —— *QQ-line*
- —— $y = \mu x + \sigma^2$ *family*

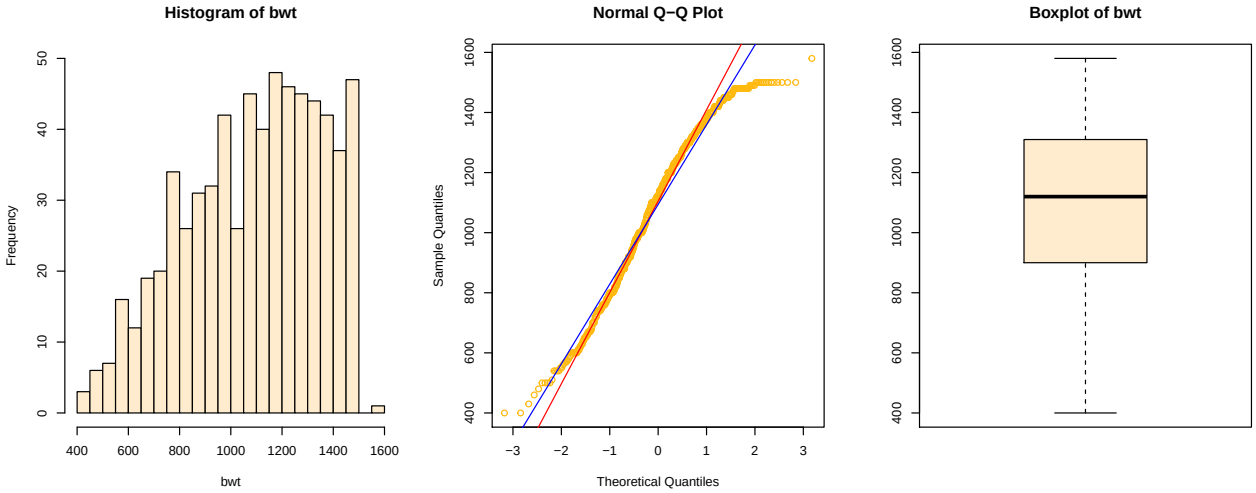
(1) According to the plots above, we conclude that the answer for the first question is **Normality cannot be excluded** for the following reasons: red, violets are blue

a. the histogram demonstrates a **roughly bell shape**;
b. as the normal Q-Q plot shows, although the right tail is lighter and right tail is heavier, the major quantiles are roughly correspond to **the Q-Q line** and the $\mathbf{y = \mu x + \sigma^2}$ family as well;
c. as we can see from the boxplot, excluding outliers, the major body of the date, ranging from the 1st quartile to the 3rd quartile, lies in the middle of the range, as well as the median value. Thus, the distribution is **approximately symmetric.**

(2) **Peculiarities:**

As we can see from the histogram, there are some "holes" which represents the totally absences of some certain cholesterol values.

(iii) *vlbw.csv* dataset:



- *QQ-line*
- $y = \mu x + \sigma^2$ *family*

(1) According to the plots above, we conclude that the answer for the first question is **Obviously not from a normal distribution"** for the following reasons:

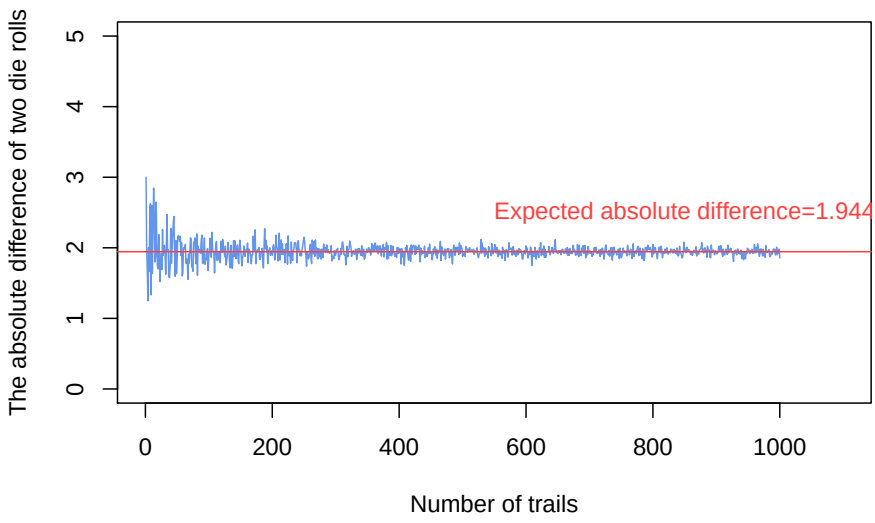
a. the shape of the histogram is by no means a **bell shape**;
b. as the normal Q-Q plot shows, although the major quantiles adhere to the Q-Q line and y = μx + σ^2 family, the left tail and, especially, the right tail **horribly deviate** from two lines;
c. as we can see from the boxplot, the distribution is quite **left-skewed** and therefore, it is not **symmetric** whatsoever.

(2) **Peculiarities:**

Compared to the previous two datasets, this distribution demonstrates a more **centralized pattern** and there are **no outliers**.

Exercise 2.4

a)



As the diagram illustrates, when the number of trails becomes larger, the absolute difference of two di

b)

(1)

#	outcome	frequency
1	0	168
2	1	260
3	2	219
4	3	190
5	4	112
6	5	51

```
source("function02.txt")
num_trails = 1000
outcomes = diffdice(num_trails)

E = 0
for (x_i in 1:6) {
  # calculate the probability of P(X=xi);
  P_xi = length(outcomes[outcomes==x_i]) / num_trails
  # calculate the approximate expectation which is the weighted average of x_i;
  E = E + x_i * P_xi
  E
}
```

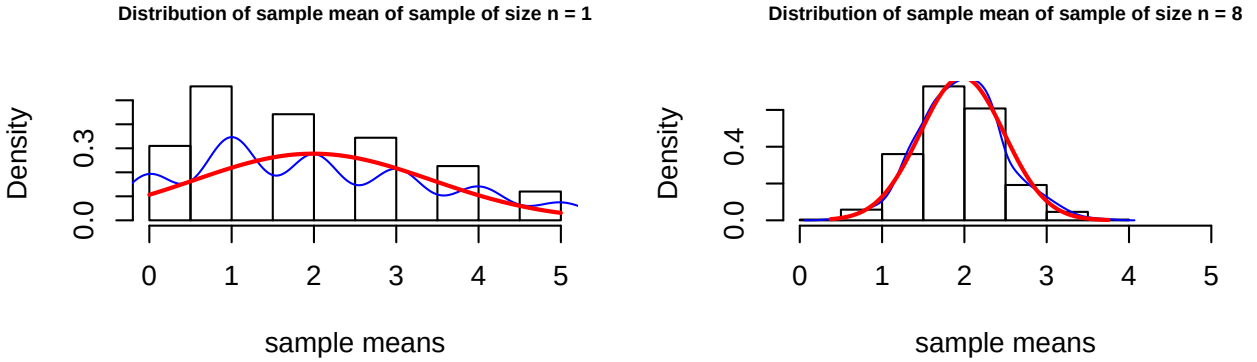
Using the function `diffdice` and above code, we found an approximate value of expectation `E` of the random variable 'the absolute difference of two die rolls' : **1.971**

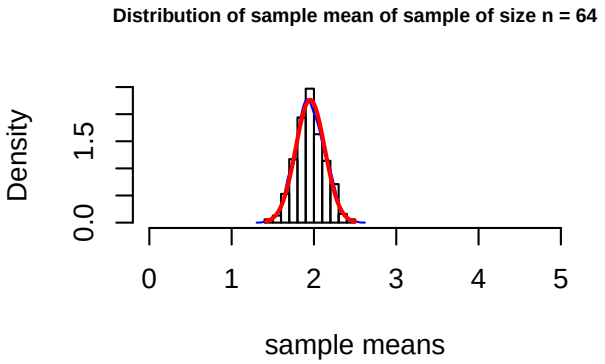
(2)

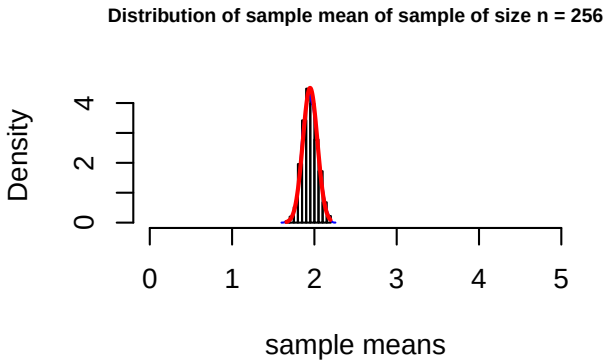
```r
P_3 = length(outcomes[outcomes==3]) / num_trails
```

Using the above code, we found that the probability of the event 'the absolute difference of two die rolls is 3' `P_3` is **0.19**

c)







- *Normal distribution*
- *Empirical distribution*

d)

The blue lines in the above figures represent empirical distributions and the red lines are the normal distribution. As these plots demonstrate, those two lines gradually merge together as the sample size become larger. The shape of the histogram also increasingly adheres to the shape of the normal distribution. In other words, the distribution of the sample means steadily move toward a normal distribution as the sample size goes larger.

Appendix

```r
# 2.3 a
par(mfrow=c(2,2))
# i)
a = rt(150, 4)
```

```r
qqnorm(a, main = "i) Normal QQ plot of t-distribution", cex.main=0.7, col = "cornflowerblue")
abline(mean(a), sd(a), col = "coral2")
# ii)
b = rnorm(100, -1, 6)
qqnorm(b, main = "ii) Normal QQ plot of N(-1, 6) distribution", cex.main=0.7, col = "cornflowerblue")
abline(mean(b), sd(b), col = "coral2")
# iii)
c = rchisq(170, 5)
qqnorm(c, main = "iii) Normal QQ plot of chi-squared-distribution", cex.main=0.7, col = "cornflowerblue
abline(mean(c), sd(c), col = "coral2")
# iv)
d = runif(90, 4, 25)
qqnorm(d, main = "iv) Normal QQ plot of uniform distribution", cex.main=0.7, col = "cornflowerblue")
abline(mean(d), sd(d), col = "coral2")

# 2.3 b
par(mfrow=c(1,3))
# (i)
titanic = read.csv("titanic3.csv")
ages = titanic$age
sd_titanic=sd(ages, na.rm=TRUE)
mean_titanic=summary(ages)[4]

# Add legends later
hist(ages, breaks = 50, col = "blanchedalmond")
qqnorm(ages, col = "darkgoldenrod1")
qqline(ages, col = "red")
abline(as.numeric(mean_titanic), sd_titanic, col = "blue")

boxplot(ages, col = "blanchedalmond", main = "Boxplot of ages")

# (ii)
par(mfrow=c(1,3))
diabetes = read.csv("diabetes.csv")
chol = diabetes$chol
sd_chol=sd(chol, na.rm=TRUE)
mean_chol=summary(chol)[4]

hist(chol, breaks =50, col = "blanchedalmond")
qqnorm(chol, col = "darkgoldenrod1")
qqline(chol, col = "red")
abline(as.numeric(mean_chol), sd_chol, col = "blue")

boxplot(chol, col = "blanchedalmond", main = "Boxplot of chol")

# (iii)
par(mfrow=c(1,3))
vlbw = read.csv("vlbw.csv")
bwt = vlbw$bwt
sd_bwt=sd(bwt, na.rm=TRUE)
mean_bwt=summary(bwt)[4]

hist(bwt, breaks = 30, ylim = c(0, 50), col = "blanchedalmond")
```

```r
qqnorm(bwt, col = "darkgoldenrod1")
qqline(bwt, col = "red")
abline(as.numeric(mean_bwt), sd_bwt, col = "blue")

boxplot(bwt, col = "blanchedalmond", main = "Boxplot of bwt")

# 2.4
# (a)
dice = (1: 6)
results = c()

for (j in 1:1000) {
  mean_diff = 0
  for(i in 1:j) {
    trail = sample(6, 2, replace = TRUE)
    diff = abs(trail[1]-trail[2])
    mean_diff = mean_diff + diff
  }
  mean_diff = mean_diff/j
  results[j] <- mean_diff
}

plot(results, type="l", ylim=c(0, 5), xlim=c(0,1100), xlab = "Number of trails", ylab = "The absolute di
abline(1.9444, 0, col = "brown1")

text(850,2.5, paste0("Expected absolute difference=", 1.944), col = "brown1")

# (d)
source("function02.txt")
num_trails = 1000
outcomes = diffdice(num_trails)

samples_of_1 = numeric(1000)
samples_of_8 = numeric(1000)
samples_of_64 = numeric(1000)
samples_of_256 = numeric(1000)

for (i in 1:1000) {
  samples_of_1[i] <- mean(diffdice(1))
  samples_of_8[i] <- mean(diffdice(8))
  samples_of_64[i] <- mean(diffdice(64))
  samples_of_256[i] <- mean(diffdice(256))
}

par(mfrow=c(2,2))

hist(samples_of_1, prob = T, xlim = c(0,5), main = "Distribution of sample mean of sample of size n = 1
lines(density(samples_of_1), col = "blue")
x <- seq(min(samples_of_1), max(samples_of_1), by=.001)
y <- dnorm(x, mean=mean(samples_of_1), sd=sd(samples_of_1))
lines(x, y, type="l", lwd=2, col = "red")
```

```r
hist(samples_of_8, prob = T, xlim = c(0,5), main = "Distribution of sample mean of sample of size n = 8
lines(density(samples_of_8), col = "blue")
x <- seq(min(samples_of_8), max(samples_of_8), by=.001)
y <- dnorm(x, mean=mean(samples_of_8), sd=sd(samples_of_8))
lines(x, y, type="l", lwd=2, col = "red")

hist(samples_of_64, prob = T, xlim = c(0,5), main = "Distribution of sample mean of sample of size n = 
lines(density(samples_of_64), col = "blue")
x <- seq(min(samples_of_64), max(samples_of_64), by=.001)
y <- dnorm(x, mean=mean(samples_of_64), sd=sd(samples_of_64))
lines(x, y, type="l", lwd=2, col = "red")

hist(samples_of_256, prob = T, xlim = c(0,5), main = "Distribution of sample mean of sample of size n =
lines(density(samples_of_256), col = "blue")
x <- seq(min(samples_of_256), max(samples_of_256), by=.001)
y <- dnorm(x, mean=mean(samples_of_256), sd=sd(samples_of_256))
lines(x, y, type="l", lwd=2, col = "red")
```