Teo Rebull - 2721042
Jacquiline Rose Roney - 2761539
Selma Kocabiyik - 2759524

Assignment 1 - Group 84

Exercise 1

a)
The probability that a random person, which does the test, gets a positive result can be shown like P(Positive). Since the accuracy of the test is 95% correct and so 5% considered as wrong test results. The following table demonstrates the probabilities.

	Positive Tested	Negative Tested
Has Cancer	95% (true positive)	5% (false negative)
Has No Cancer	5% (false positive)	95% (true negative)

To find out $P(Positive)$, we used the Total Law of Probability. Which give us:
(Considering the probabilities as independent events, since it stated as that in context).

$$Probability\ of\ True\ Positive\colon\ P(Cancer) \times P(Positive \mid Cancer) = 0.004 \times 0.95 = 0.0038$$
$$Probability\ of\ False\ Positive\colon\ P(No\ Cancer) \times P(Positive \mid No\ Cancer) = 0.996 \times 0.05 = 0.0498$$
$$Total\ Probability\ of\ a\ Positive\ Test\ Result\colon 0.0038 + 0.0498 = 0.0536$$
rounded : 0.054

In this question we calculated the probability that a random person gets a positive result from the cancer diagnostic test. However, in exercise 1.3 asked the probability that a person has cancer, given that they have received a positive test result. Both questions should have different approaches to achieve the wanted probability result.

b)
Since they asked the probability of having cancer given a positive test result , Bayes' Theorem would be used.

$$Bayes'\ Theorem$$
$$P(Cancer \mid Positive) = \frac{P(Positive \mid Cancer) \times P(Cancer)}{P(Positive)}$$

$$Total\ Law\ of\ Probability$$
$$P(Positive) = P(Positive \mid Cancer) \times P(Cancer) + P(Positive \mid No\ Cancer) \times P(No\ Cancer)$$

$P(Positive \mid Cancer) = 95\%\ or\ 0.95$ Since the accuracy of the results are 95% correct.
$P(Cancer) = 4\%\ or\ 0.004$ Since 4% of the population has cancer.
$P(Positive) = As\ we\ calculated\ in\ exercise\ 1a,\ 0.0536$

Filling the values in the theorem:

$$P(\,Cancer \mid Positive) = \frac{0.95 \times 0.004}{0.0536} = 0.070895522\ or\ 0.071\ .$$

c)

The events “person has cancer” and “positive test” are dependent, since the occurrence of one affects the probability of the other. So it is true to say that if a person has cancer (Event A), it significantly increases the likelihood of a positive test result (Event B). Moreover, if the result of the test is positive for a person, the probability of having cancer would increase. This dependency can be seen in the calculations using conditional probabilities and Bayes' Theorem.
The received risk of having cancer rises significantly because of the positive test result. It moves from the general prevalence rate (0.4% from the exercise) to a much higher probability (approximately 7.09% as calculated in Exercise 1.3). The probability of a positive test result is higher if the person has cancer. This illustrates how new information provided by test findings allows us to update and improve our assessment of the health of a person.

Exercise 2

a) A sample space is a set of all possible events that cannot be broken down further. In this case, the sample space consists of all the possible timings the individual can wait for the bus. As per the information provided, a bus arrives every 15 minutes. Hence the sample space consists of minutes 0-14. Perhaps the sample space S is:

$$S = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14\}$$

As the question does not state any restriction on the time an individual can arrive, the probability that the individual can arrive at a certain time t can be described using the probability P(t). t being a value of minute from the sample space S, the probability that the individual reaches the stop at a time t can be as follows:

$$P(t) = \frac{1}{15}$$

b) Now, the question states that if the previous bus will be missed by 4 or less minutes, the waiting time is more than 11 minutes and then the individual hurries and catches the bus (so waiting time will be 0 in this case). Whereas, if the previous bus is missed by more than 5 minutes, that is the waiting time is 10 minutes or less, then the individual continues at their normal pace. A random variable X models the waiting time at the bus stop. So the probability an individual will need to wait at least 5 minutes at the bus stop or P(X ≥ 5) is the sum of probability of x=5, x=6, x=7, x=8, x=9 and x=10. So:

$$P(5 \leqslant X) = \frac{1}{15} + \frac{1}{15} + \frac{1}{15} + \frac{1}{15} + \frac{1}{15} + \frac{1}{15}$$

$$P(5 \leqslant X) = \frac{6}{15}$$

$$P(5 \leqslant X) = \frac{2}{5}$$

$$P(5 \leqslant X) = 0.4$$

c) X is a random variable demonstrating waiting time, from the sample space, an individual waits at the bus stop from 0 to 10 minutes. So, the random variable X waiting at the bus stop consists of time 0-10, so the expectation of X is as follows:

$$E(X) = \sum [t * P(t)]$$

$$E(X) = (\frac{1}{15} * 0) + (\frac{1}{15} * 1) + (\frac{1}{15} * 2) \ldots . (\frac{1}{15} * 10)$$

$$E(X) = \frac{11}{3}$$

$$E(X) = 3.67$$

As it can be seen above using the expectation formula when substituting t with values from 1 to 10. The expectation is 3.67minutes.

d) The variance is as follows:

$$Var[X] = \sum [t^2 * P(X)] - E(X)^2$$

$$Var[X] = \sum_{0}^{10} [t^{^2} * \frac{1}{15}] - 3.67^2$$

$$Var[X] = \frac{1}{15} * (0^2 + 1^2 \ldots . 10^{10}) - 3.67^2$$

$$Var[X] = 12.2$$

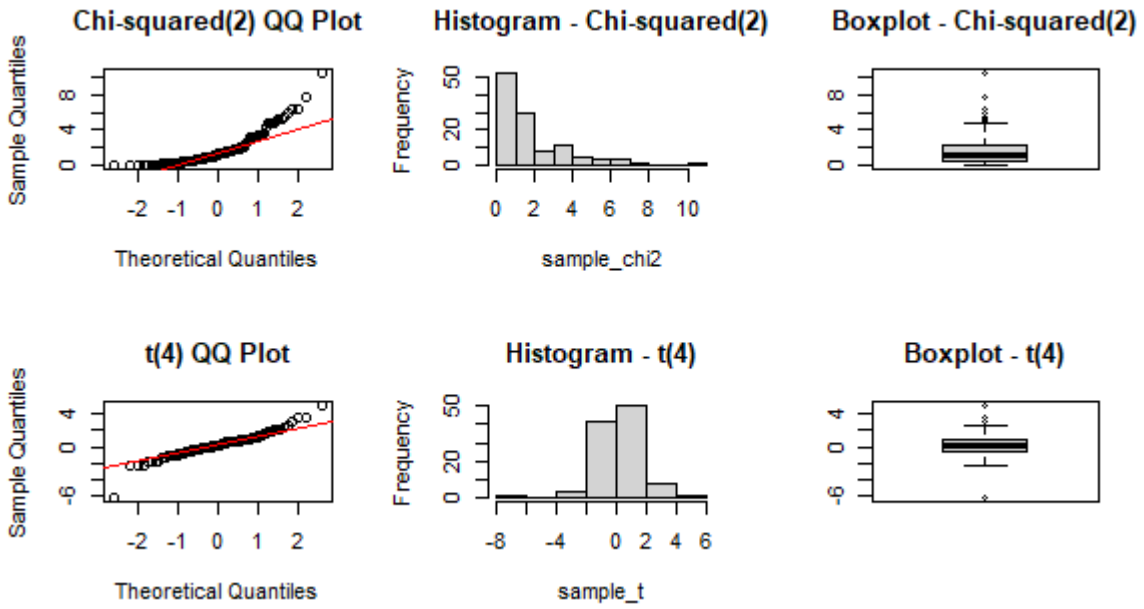
Similarly to the mean, the variance of the X (waiting time at the bus stop) can be found by substituting t with values from 1-10 into the variance formula. Which leads to variance being 12.2 rounded to one decimal place.

e) According to the central limit theorem a large number of independent randomly distributed variables will approach a normal distribution regardless of the distribution of the sample. In this case the sample of the current distribution is not a normal distribution. However, assuming that the waiting times are independent of each other 160 times a year the distribution of your average waiting time across the whole year can be estimated to be normal on the basis of the central limit theorem.

Exercise 3

a)
chi-squared:
qqnorm(sample_chi2, main = "Chi-squared(2) QQ Plot")
qqline(sample_chi2, col = "red")
hist(sample_chi2, main = "Histogram - Chi-squared(2)")
boxplot(sample_chi2, main = "Boxplot - Chi-squared(2)")

t-distribution:
qqnorm(sample_t, main = "t(4) QQ Plot")
qqline(sample_t, col = "red")
hist(sample_t, main = "Histogram - t(4)")
boxplot(sample_t, main = "Boxplot - t(4)")



It is very visible the difference between the first set of plots and the second one.

On the one hand, we have the first set. This is clearly not in normal distribution because even though some of the points in the middle of the QQ-Plot are on the line, its tails are very heavy which means that there are big outliers that deform the whole normal distribution

structure. If we move onto the histogram, there is a right-skewed design which means that most of the data is on the left side. Finally, in the boxplot, we find that the median is very low compared to the whole range of random variables.

In the second set(t-distribution) there is normal distribution because in the QQ-plot the data is aligned with the diagonal and the tails are quite light despite having an outlier here and there. In the histogram it's also visible that although not perfect, it has a solid normal distribution with most of the data around the mean. Finally, in the boxplot, the median is also very centered which normally is an indicator of having a well distributed sample.

As a final comment, the reason why I say that the tails are light is because there is nearly no point that doesn't follow the tendency of normal distribution. However, since these points in the sample are very far apart from the rest, it means that they have a meaningful impact individually to the data.

b)
(i)
Arbitrary outcome is smaller than 3: 0.5
Arbitrary outcome is bigger than -0.5: 0.6914625 (rounded: 0.691)
Arbitrary outcome is between -1 and 2: 0.2426384 (rounded: 0.243)

Relevant R-code

```
p_smaller_than_3 <- pnorm(3, mean = 0, sd = 1)
p_bigger_than_neg_05 <- 1 - pnorm(-0.5, mean = 0, sd = 1)
p_between_neg_1_and_2 <- pnorm(2, mean = 0, sd = 1) - pnorm(-1, mean = 0, sd = 1)
```

(ii)
Arbitrary outcome is smaller than 3: 0.5
Arbitrary outcome is bigger than -0.5: 0.6914625 (rounded: 0.691)
Arbitrary outcome is between -1 and 2: 0.2857874 (rounded: 0.286)
95% of the outcomes are smaller: 6.289707 (rounded: 6.290)

Relevant R-code

```
p_smaller_than_3 <- pnorm(3, mean = 3, sd = 2)
p_bigger_than_neg_05 <- 1 - pnorm(-0.5, mean = 3, sd = 2)
p_between_neg_1_and_2 <- pnorm(2, mean = 3, sd = 2) - pnorm(-1, mean = 3, sd = 2)
value_95_percentile = qnorm(0.95, mean=3, sd=2)
```

Z-score is 1.645 for the 95th percentile from the Z-table(from book).
So to transform Z-score into the value of the distribution:
$X = \sigma \cdot Z + \mu$
$X = 2 \cdot 1.645 + 3 = 6.29$
Since the value from R (which is 6.289707) is more precise, it's then true to say that R calculates (with the function 'qnorm') the result based on the continuous normal distribution rather than a discrete set of tabulated values (such as the Z-score table from the book).

(iii)

Sample mean: -0.9132614 (rounded: -0.9)
Sample standard deviation: 5.114544 (rounded: 5.1)

Relevant R-code

```
standard_normal_sample = rnorm(1000, mean = 0, sd = 1)
transformed_sample = standard_normal_sample * 5 - 1
transformed_sample_mean = mean(transformed_sample)
transformed_sample_sd = sd(transformed_sample)
print(transformed_sample_mean)
print(transformed_sample_sd)
```

As can be seen from the results, mean -0.9 and standard deviation 5.1 are close to the theoretical values, which are -1 and 5.

(iv)
The size of the sample affects the accuracy of the probability estimations. The probabilities estimated from a larger sample, in our case demonstrated by 100,000 observations, are closer to the theoretical probabilities. For instance, the result of P(Z < 3) for 100,000 observations was approximately 0.99861, which is really close to 0.9987(theoretical probability). On the other hand, 100 observations give 1, a clear deviation likely due to sampling variability. Here we can also see the law of large numbers, which states that if sample size grows, the average of the whole population gets closer to its mean.

```
sample_100 <- rnorm(100, mean = 0, sd = 1)
sample_100000 <- rnorm(100000, mean = 0, sd = 1)
calculate_probabilities <- function(sample) {
  p_smaller_than_3 <- mean(sample < 3)
  p_bigger_than_neg_05 <- mean(sample > -0.5)
  p_between_neg_1_and_2 <- mean(sample > -1 & sample < 2)
  return(list(P_smaller_than_3 = p_smaller_than_3,
              P_bigger_than_neg_0.5 = p_bigger_than_neg_05,
              P_between_neg_1_and_neg_2 = p_between_neg_1_and_2))
}
P_100 <- calculate_probabilities(sample_100)
P_100000 <- calculate_probabilities(sample_100000)
```

Results

```
> print(P_100)
$P_smaller_than_3 [1] 1
$P_bigger_than_neg_0.5 [1] 0.65
$P_between_neg_1_and_neg_2 [1] 0.83

> print(P_100000)
$P_smaller_than_3 [1] 0.99861
$P_bigger_than_neg_0.5 [1] 0.68968
$P_between_neg_1_and_neg_2 [1] 0.81741
```

c)

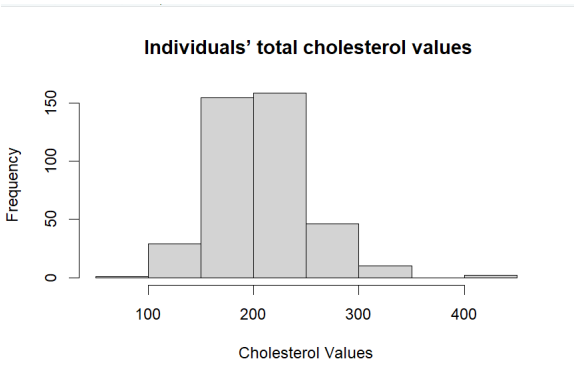


Chart 1.1: Histogram Diabetes Chart

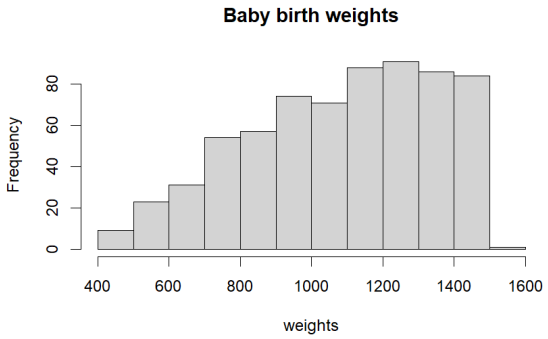


Chart2: Histogram Newborn weights

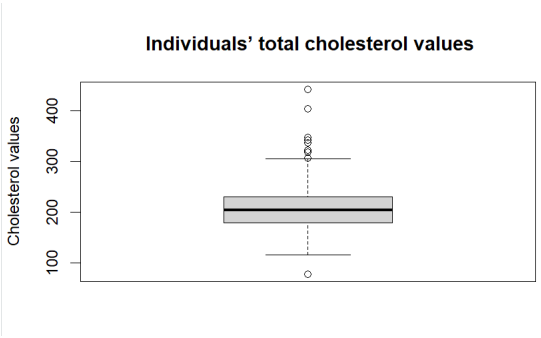


Chart 1.2 Boxplot of Diabetes data

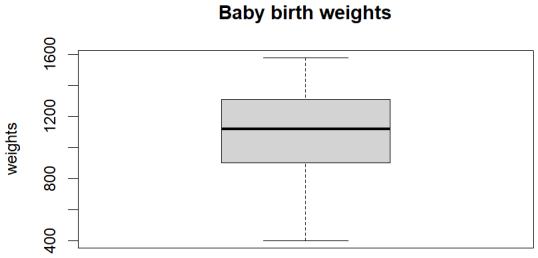


Chart 2.2 Boxplot of Newborn weights

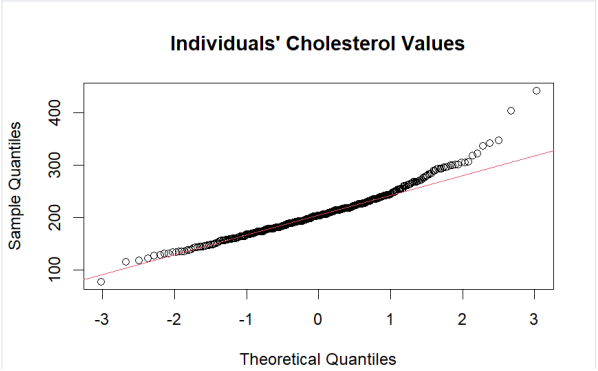


Chart.1.3 QQ plots of Diabetes data

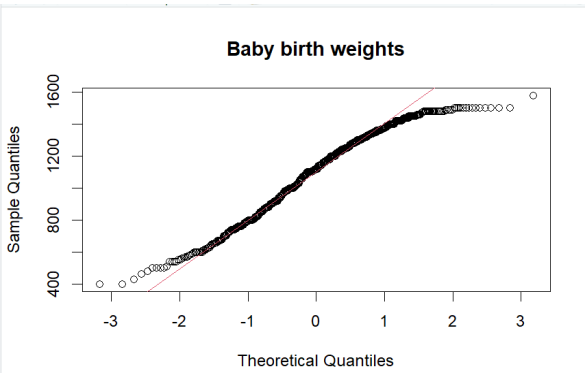


Chart 2.3 QQ plots for Newborn weights

When comparing chart 1.1 and 2.1 it can be noticed that chart 1.1 is more symmetrical than 2.1 as 2.1 is more right skewed. Chart 1.1 also has a smaller dispersion compared to chart 2.1. If the shape of the histogram is compared to the bell curve of a normal distribution, chart 1.1 has more of a bell curve shape compared to 2.1. However it could be that with more data 2.1 can be seen as a bell curve as it could very likely be that only the left half of the bell curve is visible in chart 2.1. This could be because 2.1 is a sample extracted from a larger population that follows a normal distribution. However, based on the current shape of the chart 2.1 it does not follow the bell curve shape of a normal distribution strictly.

Comparing chart 1.2 and 2.2 is insightful to understand the spread of the data. It must be noticed that chart 1.2 has a lot more peculiar data points compared to 2.2 this is due to the greater presence of outliers in chart 1.2 compared to 2.2 whose data is more within the certain range. From the box plot itself it is visible that chart 2.2's data is more spread within

the range of values on the y axis compared to chart 1.2 whose majority of the data is concentrated within 200-250.

The QQ plots of both datasets visualized in chart 1.3 and 2.3 are very similar, in both of them, the data points seem to deviate from the red linear line at the beginning and end as visible. However, the middle portion a of the data seemed to align with the line of distribution in both cases. Perhaps, on a more critical note, chart 1.3 seemed to be more aligned to the liner line than chart 2.3.

In conclusion, it can be derived that for **dataset 1 about diabetes "Normality cannot be excluded"** for sure. The histogram's resemblance to a bell curve, the spread of the data demonstrated in the box plot and the alignment of the data with the line in the QQ plot all support this as explained. However, the situation of dataset 2 with the newborn weights is more complex, as explained the histogram partially follows the bell curve shape of a normal distribution indicating that the dataset could be a sample of a larger population that shows a clear normal distribution. However, the provided chart the spread of data demonstrated through the box plot and QQ plots alone does not support the standards for a normal distribution as it is very right skewed to be normally distributed. Perhaps, the resemblance and qualitative analysis is not strong enough to solidify that the data might be part of a normal distribution. Hence, for the sake of choosing from the options provided and based on what is shown in the charts **dataset 2 about newborn weights is "Obviously not from a normal distribution".** However, with the provision of more data it could be normal distributed.

Relevant code:

- Chart1.1:

```
> diabetes <- read.csv("diabetes.csv")
> hist(diabetes$chol,main=" Individuals' total cholesterol values",xlab="Cholesterol
Values",ylab="Frequency")
```

- Chart1.2

```
> diabetes <- read.csv("diabetes.csv")
> boxplot(diabetes$chol, main=" Individuals' total cholesterol values", ylab="Cholesterol values")
```

- Chart1.3

```
> diabetes <- read.csv("diabetes.csv")
> qqnorm(diabetes$chol, main = "Individuals' Cholesterol Values")
> qqline(diabetes$chol, col = 2)
```

- Chart2.1

```
> vlbw <- read.csv("vlbw.csv")
> hist(vlbw$bwt,main="Baby birth weights",xlab="weights",ylab="Frequency")
```

- Chart2.2

```
> vlbw <- read.csv("vlbw.csv")
> boxplot(vlbw$bwt, main="Baby birth weights", ylab="weights")
```

- Chart2.3

```
> vlbw <- read.csv("vlbw.csv")
```

```
> qqnorm(vlbw$bwt, main="Baby birth weights")
> qqline(vlbw$bwt, col = 2)
```

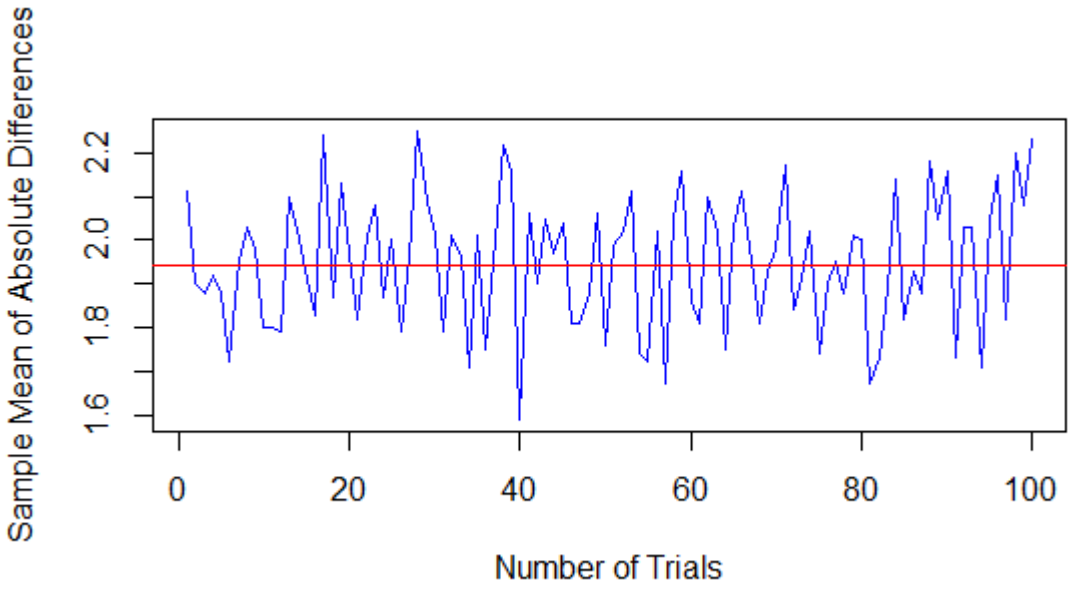
Exercise 4

a)
To make the plot:

plot(1:n_trials, trial_means, type = "l", col = "blue", xlab = "Number of Trials", ylab = "Sample Mean of Absolute Differences")

To draw the red dashed line:

abline(h = expected_value, col = "red")



b)

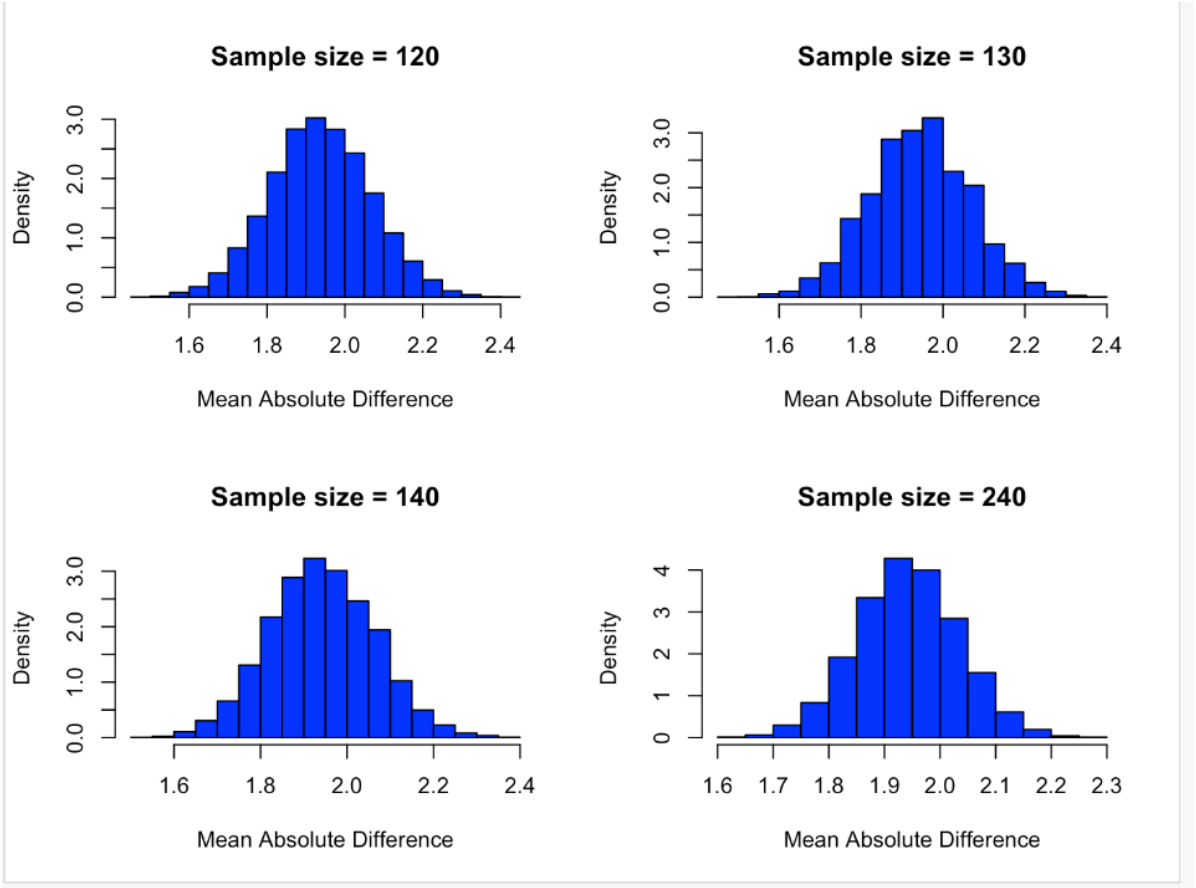
results <- diffdice(num_simulations)
approximate_probability_of_event <- mean(results == 3)

Results:

cat("Approximate Expectation:", approximate_expectation, "\n")
cat("Approximate Probability of Absolute Difference = 3:",
approximate_probability_of_event, "\n")

```
> cat("Approximate Expectation:", approximate_expectation, "\n")
Approximate Expectation: 1.9421
> cat("Approximate Probability of Absolute Difference = 3:", approximate_probabil
ity_of_event, "\n")
Approximate Probability of Absolute Difference = 3: 0.1675
```

c)



R-code for the histograms

```
par(mfrow = c(2, 2))
sample_sizes <- list(120, 130, 140, 240)
for (n in sample_sizes) {
  means <- replicate(10000, mean(diffdice(n)))
  hist(means,
       main = paste("Sample size =", n),
       xlab = "Mean Absolute Difference",
       breaks = 20,
       col = "blue",
       freq = FALSE)#bc in the slides density is used
}
```

d)

The resulting histograms of the mean of absolute differences get more accurate and start to approximate the bell curve of a typical normal distribution as we increase the number of trials

inside each replicate (n) and increase the size of the samples. This is explained by the Central Limit Theorem, which states that regardless of the initial distribution of the population, as sample size increases, the means of the sample distribution will resemble a standard normal distribution. So as the sample size increases, each of our four plots illustrates this impact, becoming more symmetric and narrowly peaked around the mean, which is what the CLT indicates.