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

Assignment 3 - Group 84

Exercise 1

a) Based on the data alone, it cannot be concretely said if there is a linear correlation on not. However, if the data is closely observed it can be seen that as the opening bid values increase, the final bids also increase, suggesting that there could be a positive linear correlation. Calculating the r value or the correlation coefficient, which can be done using the program R as follows:

```
> opening_bid <- c(1500, 500, 500, 400, 300)
> winning_bid <- c(650, 175, 125, 275, 125)
> correlation_coefficient <- cor(opening_bid, winning_bid)
> print(correlation_coefficient)
0.9468915
```

As the correlation coefficient; r = 0.947 is very close to 1 a very strong positive correlation can be suggested.

b) The intercept and slope of the linear regression line can be calculated with the formula:

$$b_1 = r\frac{s_y}{s_x} \qquad \text{and} \qquad b_0 = \overline{y} - b_1\overline{x};$$

where b1 is the slope and b0 is the y-intercept.

- to calculate b1 the r value is already calculated while sy and sx the standard deviation of y and x needs to be calculated. using R:
```
> standard_deviation_opening_bid <- sd(opening_bid)
> print(standard_deviation_opening_bid)
487.8524
> standard_deviation_winning_bid<- sd(winning_bid)
> print(standard_deviation_winning_bid)
221.0769
```
- so now the values can be substitute to calculate b1:

$$b_1 = 0.947 \times \frac{221.0769}{487.8524}$$

$$b_1 = 0.429$$

- Now the values of b0 can be also computed which requires the means of x and y which can again be calculated using R:
```
> mean_opening_bid <- mean(opening_bid )
> print(mean_opening_bid)
640
```

```
> mean_winning_bid <- mean(winning_bid)
> print(mean_winning_bid)
270
```

- So, b0 can be now calculated as follows:

$$b_0 = 270 - 0.429 \times 640$$

$$b_0 = -4.56$$

So, the intercept is -4.56 while the slope is 0.429 as demonstrated.

c) Use the regression line from b) to predict winning bid if the opening bid were 1000. Derive the residual values for Opening bids equal to 300 and 1500. Comment

The regression line $y = -4.56 + 0.429x$ derived previously as can be used to calculate the residual values for opening bids equal 300 and 1500 as follows:

Residual for 300:
= 125 - (-4.56 + 0.429(300))
= 0.86

Residual for 1500:
= 650 - (-4.56 + 0.429(1500))
= 11.06

In both these cases, it can be seen that the value is positive, suggesting that the model underestimates the actual predictions. The residual value for 300 is quite small while the residual value for 1500 is much more significant suggesting that the model deviates from true values as the value gets higher.

Exercise 2

a)
Since the number of breaks of a given fiber should be binomially distributed within five trials, we can use the given formula in the assignment sheet, which is X ~ Bin(n, p), EX=np. And so If X ~ Bin(n, p) (with unknown p) is observed, the success probability p is estimated by ˆp = X/n. Hereby, X is the number of successes(which means in this problem “breaking of a fiber”) in all the trials. In this problem X is the total number of breaks. To find out the X, we have to find the total broken fibers for each break(0, 1, 2, 3, 4, 5) in 5 trials. The frequency row gives us the amount of fibers that are broken. Breaks on the other hand show us how many times in 5 trials the given amount of fibers was broken. For instance if we look at the second column, it means that 1 time in 5 trials 69 fibers were broken. So X is the product of the total number of breaks with the corresponding frequencies, since we have 6 different amounts of breaks in 5 trials for 280 fibers, we are calculating the total breaks. Hereby, n is the total number of trials.
So since each of the 280(157+69+35+17+1+1) fibers is tested 5 times, total trials will be 280 x 5 = 1400.

$$\hat{p} = \frac{X}{n} = \frac{0\,x\,157 + 1\,x\,69 + 2\,x\,35 + 3\,x\,17 + 4\,x\,1 + 5\,x\,1}{5(157 + 69 + 35 + 17 + 1 + 1)} = 0.142142857 \approx 0.142$$

Thus our estimated probability p is 0.142 under the binomial distribution.

b)

To test the observed distribution with the binomial distribution using a chi-squared test we first have to determine the hypotheses.

Null hypothesis: The frequency distribution of fiber breaks follows a binomial distribution with a constant probability p, which makes all fibers to have the same strength.
Alternative hypothesis: The frequency distribution of fiber breaks does not follow a binomial distribution with a constant probability p, which means fibers can have different strength.

Pooling the last three cells:

Breaks	0	1	2	3-5
Frequency	157	69	35	19

$$\mathbb{X}^2 = \sum \left(\frac{(O-E)^2}{E} \right)$$

$$Observed\ values: 157,\ 69,\ 35,\ 19 (= 17 + 1 + 1)$$

$$Expected\ Frequency (using\ binomial\ distribution): E = P(X = k) = \left(\frac{n}{k} \right) p^k (1-p)^{n-k}$$

$$binomial\ coefficient: \left(\frac{n}{k} \right) = \frac{n!}{k!(n-k)!}$$

$$estimated\ probability: p = 0.142$$

$$n = 5\ trials$$

$$k = succes(s)\ out\ of\ n\ trials.\ (for\ breaks\ 0, 1, 2)$$

$$For\ breaks\ 3-5:$$

$$P(X \geq 3) = 1 - \left[P(X = 0) + P(X = 1) + P(X = 2) \right]$$

We are looking here for a goodness-of-fit so df is in this case k-1=4-1=3.
If we fill all the values into R;

```
> breaks_non_pooled <- c(0, 1, 2)
> frequencies <- c(157, 69, 35, 17, 1, 1)
```

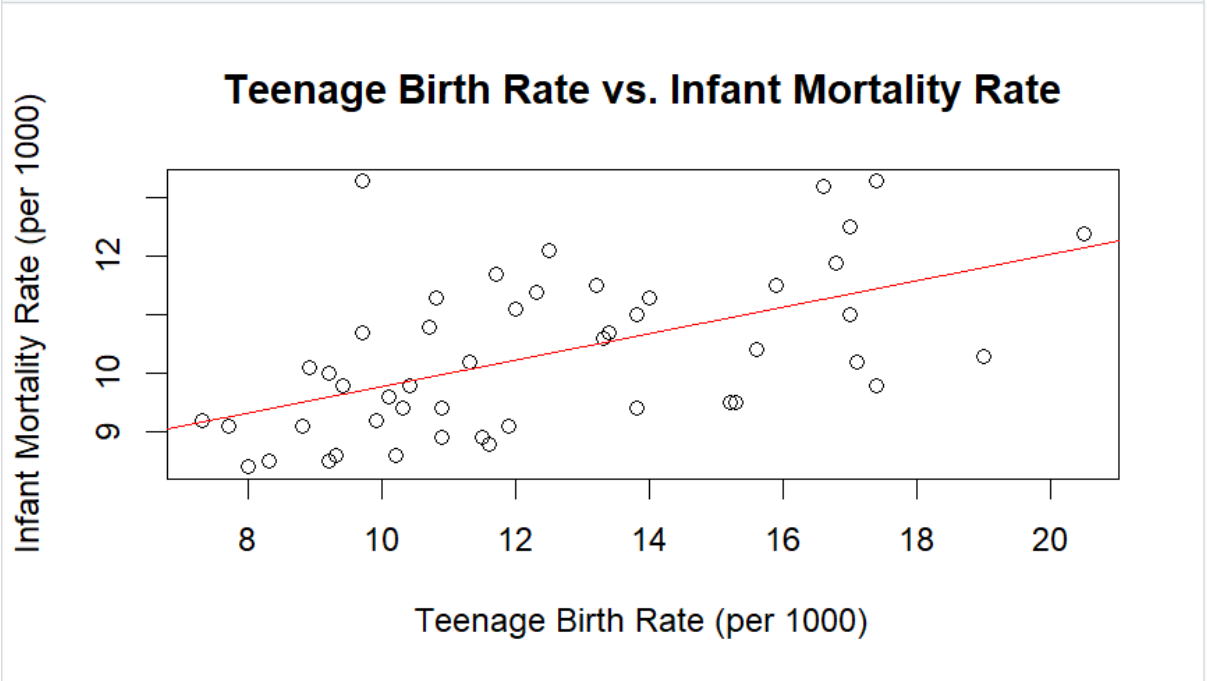
```
> trials_n <- 5
> estimated_p <- 0.1421429
> total_trials <- 280
>
> E_non_pooled <- dbinom(breaks_non_pooled, trials_n,
estimated_p) * total_trials
> E_pooled <- (1 - sum(dbinom(breaks_non_pooled, trials_n,
estimated_p))) *  total_trials
>
> observed_freq<- c(frequencies[1:3], sum(frequencies[4:6]))
> expected_freq <- c(E_non_pooled, E_pooled)
>
> chi_squared_test <- chisq.test(observed_freq, p =
expected_freq / total_trials) #here we divide E by total
trials to make it probability, since in R chisq.test requires
probabilities in place of
> chi_squared_test

     Chi-squared test for given probabilities

data:  observed_freq
X-squared = 44.149, df = 3, p-value = 1.403e-09
```

From table 4 of the book with the df=3 and significance level 0.05, the critical value is 7.815. To conclude the result; the p-value is 1.403e-09 which is less than significance level 0.05 and also the critical value is 7.815 which is less than our chi square test result(X-squared > X-squared k-1, alpha). Thus the null hypothesis can be rejected and therefore, the frequency distribution of fiber breaks does not follow a binomial distribution with a constant probability p, which means fibers can have different strength.

Exercise 3

a)

```
plot(mortality$teen, mortality$mort, main = "Teenage Birth Rate vs. Infant Mortality Rate",
     xlab = "Teenage Birth Rate (per 1000)", ylab = "Infant Mortality Rate (per 1000)")
linear_model <- lm(mort ~ teen, data = mortality)
abline(linear_model, col = "red")
summary(linear_model)
```



Given the scatterplot above, it can be concluded that as the teenage birth rate increases there is a positive tendency with the infant mortality rate increasing as well. In other words, when the mother of the child is older, it is more likely that the child will live longer than if the mother is younger.

b)

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.6429 -1.0348 -0.0184  0.6831  3.5902

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.52640    0.64930  11.592 3.03e-15 ***
teen         0.22509    0.05052   4.456 5.32e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.14 on 46 degrees of freedom
Multiple R-squared:  0.3015,   Adjusted R-squared:  0.2863
F-statistic: 19.85 on 1 and 46 DF,  p-value: 5.317e-05
```

The p-value = 5.32e-05 here shows the relationship between independent and dependent variables. Since null hypothesis stands for no relationship between teen and mort so like Betha = 0 which also stands for slope coefficient. For alternative H; there is a relationship, meaning slope coefficient is not zero. Thus since the found p-value is much less than the 0.01(significance level) we reject the null hypothesis. Which also means that it is almost not possible to randomly observe 0.22509 as slope coefficient .
Dependent variable (mort): this is what we are trying to predict/estimate.
Independent variable (teen): with using this we are trying to explain the dependent variable.
Intercept, 7.52640 : constant term for this case.
Slope, 0.22509 : coefficient of the independent variable
Filling these values into the formula of simple linear regression gives as estimated mort, which also gives the wanted regression equation.

```
linear_model <- lm(mort ~ teen, data = mortality)
summary(linear_model)
conf_interval <- confint(linear_model, level = 0.98)
conf_interval
residual_std_error <- summary(linear_model)$sigma
error_variance <- residual_std_error^2
```

```
                1 %      99 %
(Intercept) 5.961478 9.0913266
teen        0.103335 0.3468416
```

These are the outputs we get for a 98% confidence interval. These are the lower and upper bounds for each coefficient. In this context, this means that we are 98% sure that the true value of the intercept and the true value of the slope coefficient falls within their respective bounds.

Now, going back to the results from part a, we had that the positive estimate for the slope coefficient was 0.22509 which suggests that there is a positive relationship between teenage birth rate (**teen**) and child mortality rate (**mort**). Moreover, the fact that the CI for the slope coefficient doesn't include 0 supports the idea that there is a significant relationship.

estimate error variance -> `Residual standard error: 1.14`

determination of the coefficient R-squared → `Multiple R-squared:  0.3015`

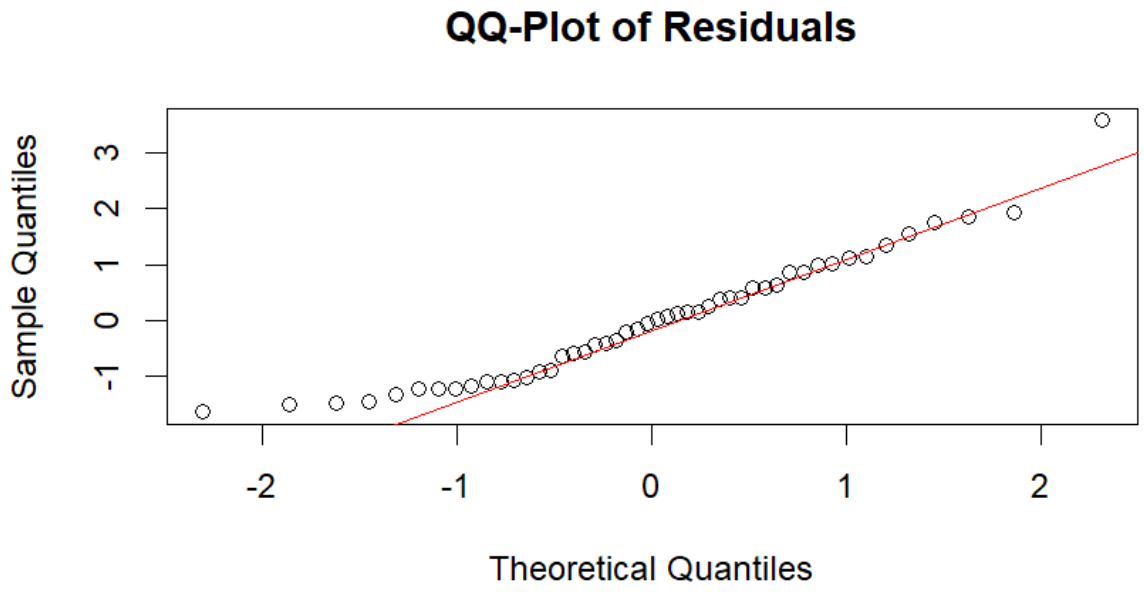
c)
The formula to get the answers based on the known data is the following:

Predicted Mortality Rate = 7.52640(**Intercept**) + 0.22509(**slope**) x 10(**teen**)

We get 9.77729 which is a prediction that for every 1000 births, the rate would be around 9.78.

d)

```
residuals <- residuals(linear_model)
plot(mortality$teen, residuals, main = "Scatterplot of Predictor against Residuals",
     xlab = "Teenage Birth Rate (per 1000)", ylab = "Residuals")
abline(h = 0, col = "red")
qqnorm(residuals, main = "QQ-Plot of Residuals")
qqline(residuals, col = "red")
```



As we can see in the plot above, it looks like we are talking about a normally distributed plot. However, there are a few deviations on the left-tail. They are above the red line which means that it is a heavy tail. Since this happens, the residuals might not be normally distributed. However, even though there are a few extreme values to the left side, the rest is quite well fitted altogether which might suggest that the rest are just some outliers and might not have a huge impact overall.

If that's the case, we can safely say that the results obtained from part b are reliable.
Also, something to note here is that linear regression models are generally robust, meaning they can handle some degree of deviation from the assumptions and especially if the sample size is sufficiently large, which is our case.

Exercise 4

a)

What we want to investigate is whether Andy's friends are equally strong, which also means that his friends(which in this case have different proportions) have the same amount of (same proportion of) strength in the game(some characteristics)?

Null Hypothesis: Andy's friends are equally strong opponents.
Alternative hypothesis: Andy's friends are not equally strong opponents.

Since we are dealing with categorical data and so contingency tables, here we used the Chi-Squared test to test for homogeneity.

From R:

```
> data_matrix <- matrix(c(179, 47, 57, 283,
+                         96, 17, 36, 149,
+                         52, 13, 18, 83,
+                         39, 15, 15, 69,
+                         84, 37, 39, 160,
+                         450, 129, 165, 744),
+                       nrow = 6, ncol = 4, byrow = TRUE)
> rownames(data_matrix) <- c("Bob", "Cecilia", "David",
"Emma", "Freddy", "Total")
> colnames(data_matrix) <- c("Won", "Lost", "Draw", "Total")
> print(data_matrix)
        Won Lost Draw Total
Bob     179   47   57   283
Cecilia  96   17   36   149
David    52   13   18    83
Emma     39   15   15    69
Freddy   84   37   39   160
Total   450  129  165   744
> #  chi-squared test with excluding totals, since R doesn't
need total values for chi-test.
> data_matrix_without_total <- data_matrix[1:5, 1:3]
> chi_test_result <- chisq.test(data_matrix_without_total)
> print(chi_test_result)

     Pearson's Chi-squared test

data:  data_matrix_without_total
X-squared = 10.931, df = 8, p-value = 0.2056
```

As the results we got chi-squared as 10.931 and p-value as 0.2056. We can approach our conclusion in two ways.

1. **P-value:** p-value from the test is greater than significance level. Which is 0.2056 > 0.05. So in this case we don't have enough evidence to reject the null hypothesis and to conclude that Andy's friends are not equally strong opponents.
2. **Chi-squared critical value:** To find the critical value in R:

```
> alpha = 0.05
> df <- chi_test$parameter
> critical_value <- qchisq(1 - alpha, df)
> print(critical_value)
```

```
[1] 15.50731
```

Since this critical value(15.50731) is greater than our chi-squared value(10.931) (chi-squared test is right tailed), we cannot reject the null hypothesis.

b)

Our null hypothesis stated that Andy's friends are equally strong when it comes to the game. Hereby we calculated the contribution of each cell with the squared residuals from our chi-squared test. A residual is the difference between the observed and expected values for a cell. Larger residuals demonstrate more contribution within the cell to the result of chi-square test. With the table it can be seen that how much each game's result differed from what would be really expected if the null hypothesis(that if all friends were equally strong) were true and how they would impact the result from Chi-Squared statistics.

```
> contributions <- chi_test_result$residuals^2 #The
squared residuals are to prevent negative values and get
larger inconsistencies.
> print(contributions)
                Won       Lost        Draw
Bob     0.35823588 0.08720234 0.529009720
Cecilia 0.38351808 3.02119217 0.264367041
David   0.06442415 0.13447451 0.009010529
Emma    0.17908836 0.77058531 0.005976667
Freddy  1.68619355 3.08960990 0.348416422
```

If we look under the 'Lost', the highest contributions which are shown with squared residuals, are 3.02119217 against Cecelia and 3.08960990 against Freddy. If all friends were equally strong there wouldn't be this huge contribution difference between the other friends' contribution values. This also can be monitored in 'Won', since Freddy's won contribution (1.68619355) is unexpectedly higher, which indicates deviation from expected values. So 3.02119217, 3.08960990 and 1.68619355 would be the most contributing to the chi-test statistics under the null hypothesis. However, even though there are significant differences seen from the contributions, it is still not sufficient to conclude that there is indeed a difference between Andy's friends' strengths.

c)

If the null hypothesis were true, this would mean that Andy's all friends has the same strength which also would mean that Andy would have the same probability of winning against each of his friends. So we can assume p as probability or proportion of Andy winning a game with his total winnings divided by the amount of games played. To find the expected wins(E), hereby we took n as 160.
From the R:

```
> total_wins <- data_matrix[nrow(data_matrix), 1]
> total_games <- data_matrix[nrow(data_matrix), ncol(data_matrix)]
> p_probability_winning <- total_wins / total_games
> n_games_against_freddy <- 160
> E_expected_wins_freddy <- p_probability_winning *
n_games_against_freddy
> print(E_expected_wins_freddy)
```

```
[1] 96.77419
```

So around 97 games Andy expected to win against Freddy under the assumption of equal strength. Thus, if we compare this result with the result from the table (84 wins against Freddy), under the assumption of the null hypothesis Andy would win more games (since 97 > 84).