Assignment 3

Tim But, Mehmet Bedirhan Gursoy, Vincentas Ryliskis, group 035

30 November 2023

Setting a Seed

In order to ensure that the results in this report are reproducible, seed (123) was used when generating random numbers.

```
set.seed(123)
```

Task 1

a)

We have r = 0.9469 based on n = 5 pairs.

H_0: p = 0
H_a: p $\neq$ 0

significance level = 0.05

We calculate r based on the formula below

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

We calculate the t-statistic in the context of correlation coefficient of p testing. The formula is shown below.

$$t_\rho = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$$t_\rho = 5.10$$

We calculte our critical values through a t-distrubution with an α of 0.05 and n = 5. Because it is a two tailed test the area in one tail is equal to 0.025.

$$-t_{4,0.025} = -2.776$$

$$t_{4,0.025} = 2.776$$

Our critical values are between -2.776 and 2.776. Since $t_\rho = 5.10 > 2.776$, we reject H_0.

There is sufficient evidence to reject the claim that there is no linear correlation opening bids suggested by the auctioneer in the auction the winning bids.

b)

The fitted regression equiation is given with the formula below. We assume the opening bid to be equal to x and the winning bid to be $\hat{y}$.

$$\hat{y} = b_1 x + b_0$$

We can calculate b_1 and b_0 through the formula: To calculate b_1 we need to find the values of s_y and s_x. We have already found these values in the question above so we can insert them to the formula and find 0.429.

$$b_1 = r\frac{s_y}{s_x}$$

$$b_1 = 0.429$$

We will use the value of b_1 to find the the value of b_0. Moreover, we will use the mean values of y and x calculated from the question above.

$$b_0 = \bar{y} - b_1\bar{x}$$

$$b_0 = -4.56$$

Finally, when we replace the variables in the formula with our numbers we get an equation shown below.

$$\hat{y} = 0.429x - 4.56$$

c)

Again, we are assuming x values to be the opening bid in this question. When we replace the x value, which is 300 in this case, to the equation calculated from the question above we receive the following answer:

$$\hat{y} = 0.429 * 300 - 4.56 = 124.14$$

Now, when $x = 1500$:

$$\hat{y} = 0.429 * 1500 - 4.56 = 638.94$$

The actual winning bid of the opening bid with a value of 300 is 125. We received an answer of 124.14. The actual winning bid of the opening bid with a value of 1500 is 650. We received an answer of 638.94. These answers are not exact howeever, we calculated our fitted regression equation on the best fit line meaning that we will get an estimate value for each item since they are not one to one correlated. This is the reason why the numbers are not exactly the same but close to the actual value.

Task 2

a)

The probability of any given fiber to break, given the table below, can be achieved by multiplying each number of breaks by it's frequency, and then dividing by the total number of fibers produced (5*280). This gives us the probability that any given fiber broke during testing.

Breaks	0	1	2	3	4	5
Frequency	157	69	35	17	1	1

$$p = \frac{0*157 + 1*69 + 2*35 + 3*17 + 4*1 + 5*1}{5*(157+69+35+17+1+1)} = \frac{199}{5*280} = 0.14214285714$$

Therefore, $p \approx 0.142$.

b)

At the moment, we are working with the following values:
Number of trials $(n) = 1400$
$p = 0.142$

In order to test whether the prediction regarding the distribution holds, we need to construct a table with expected values for the distribution assuming the same n and p. In order to do this, we need to calculate the probability that over 5 trials, a fiber will break 0, 1, 2, or more than 3 times. We can do this by using the Binomial distribution formula:

$P(x) = \frac{n!}{(n-x)!x!}p^x(1-p)^{n-x} = \frac{5!}{(5-x)!x!}0.142^x(0.858)^{5-x}$

When calculating the values, we assume that n = 5, p is as in part a, and x is a number between 0 and 5, based on the value we are calculating with P(x>=3) = P(3) + P(4) + P(5). This gives us the table below:

	0	1	2	3+
Probability	0.465	0.385	0.128	0.023

Using this table, we can then multiply the probabilities by the expected 280 trials, in order to get the expected number of breaks. Below are the tables for both the expected number of breaks (rounded to the nearest values in order to have a sum of 280), as well as the observed frequency from part a):

	0	1	2	3+
Expected Frequency	130	108	36	6

	0	1	2	3+
Observed Frequency	157	69	35	19

Now, we can use Pearson's chi-squared test in order to determine whether the distributions match:

$X^2 = \sum_{i=1}^{k=3} \frac{(O_i - E_i)^2}{E_i}$

Calculating using this formula, we get that $X^2 = 47.88547$. We can then use R in order to calculate the critical values for $X^2_{279,0.05}$, where we assume $\alpha = 0.05$, and use n = 280.

This gives us $X^2_{279,0.05} = 239.4547$, which is larger than X^2. This supports the notion that the distribution is binomial

Task 3

a)

b)

c)

d)

Task 4

a)

For this question, we need to perform a Chi-square test, in order to establish that the proportional performance of Andy's friends is independent of which friend is playing. As such, we have H_0 that all of Andy's friends are equally strong opponents, and H_1 that their skill levels are not even. The code below compares the resulting p value with $\alpha = 0.05$, and will print it's findings after the code block.

```
alpha = 0.05
                    #  W    L   D   T
results = matrix(c(179, 47, 57,# 283
                   96, 17, 36,# 149
                   52, 13, 18,# 83
                   39, 15, 15,# 69
                   84, 37, 39# 160
                  ), nrow=5, byrow=TRUE)
#           Total:  450 129 165 744
colnames(results) <- c("Won", "Lost", "Draw")
rownames(results) <- c("Bob", "Cecilia", "David", "Emma", "Freddy")

part_a_chi_test = chisq.test(results)
```

```r
if (part_a_chi_test$p.value < alpha) {
  cat("The difference in Andy's game results is statistically significant",
"at the alpha level of 0.05, allowing us to reject H0. This suggests",
"that Andy's friends are not equally strong opponents.\n")
} else {
  cat("There is no statistically significant difference in Andy's game",
"results at the alpha level of 0.05, so H0 is not rejected. This suggests that",
"there insufficient evidence to say Andy's friends are not equally strong",
"opponents.\n")
}
```

There is no statistically significant difference in Andy's game results at the alpha level of 0.05, so H0 is not rejected. This suggests that there insufficient evidence to say Andy's friends are not equally strong opponents.

b)

In order to investigate which entries contributed the most to the test statistics in a), we have to look at the individual contributions to the X^2 statistics of every element of the table. We can get these values through the code below (which will print a table containing these values).

```r
library(knitr)
kable(part_a_chi_test$res)
```

	Won	Lost	Draw
Bob	0.5985281	-0.2953004	-0.7273305
Cecilia	0.6192884	-1.7381577	0.5141664
David	0.2538191	-0.3667077	-0.0949238
Emma	-0.4231883	0.8778299	-0.0773089
Freddy	-1.2985352	1.7577286	0.5902681

In this table, the values furthest away from zero have the most impact on the P value, and as such, we should be looking at those. Specifically, most notable are Cecilia's "Lost" value, as well as Freddy's values for "Lost" and "Won", since they have the highest absolute values. Andy lost and won less games than expected against Cecilia and Freddy (respectively), and lost more games against Freddy than would be expected.

c)

The code below prints the expected amount of games that Andy would win against Freddy, assuming the null hypothesis holds.

```r
cat("Andy is expected to win",part_a_chi_test$exp["Freddy", "Won"],
    "games vs Freddy")
```

Andy is expected to win 96.77419 games vs Freddy

As we can see, this is more than the 84 that Andy won in actuality, implying that Freddy might be a better player than the others. However, this does not imply that Freddy is an outlier, as that would normally require a contribution to the X^2 statistics of greater than 2 (the current value is -1.29).