**Part I**

Exercise 1:

Downloading the data from the course site and reading it into R under the name "soil".

**Command:**



a. Running a linear regression of lead against zinc concentrations.

**Command:**

**A close-up of a logo

Description automatically generated**

**Output:**

**A screenshot of a computer

Description automatically generated**

b. Plotting the lead and zinc data and adding regression line.

**Command:**

**A screen shot of a white sheet

Description automatically generated**

**Output:**

**A graph of a graph with dots and lines

Description automatically generated with medium confidence**

c.Plotting the residuals of the regression from (a) and adding horizontal line.

**Command:**

**A screenshot of a computer screen

Description automatically generated**

**Output:**

**A graph with black dots and red line

Description automatically generated**

**USING R:**

d. Writing theequation of the linear regression line.

summary(lead\_zinc\_model) [output provided in a.]

🡪 Looking at the output,

Intercept estimate = 16.582928

Zinc slope estimate = 0.291335

The equation is:

**lead = 16.582928 + 0.291335 \* zinc**

e. Lead concentration at new data point of zinc concentration 1000 ppm.

🡪 Plugging the new data point into the regression equation,

**16.582928 + 0.291335 \* 1000**

A close-up of numbers

Description automatically generated

🡪 Expected lead concentration at this new data point (1000 ppm) is 3**07.9179 ppm.**

f. Based on 2 data points at A and B, looking at how much higher the lead concentration is expected to be in A compared to B.

lead\_A = 16.582928 + 0.291335 \* zinc\_A

lead\_B = 16.582928 + 0.291335 \* zinc\_B

lead\_A = 16.582928 + 0.291335 \* (zinc\_B + 100)

lead\_B = 16.582928 + 0.291335 \* zinc\_B

lead\_A – lead\_B = 16.582928 + 0.291335 \* (zinc\_B + 100) – (16.582928 + 0.291335 \* zinc\_B)

lead\_A – lead\_B = 0.291335 \* (zinc\_B + 100) - 0.291335 \* zinc\_B

lead\_A – lead\_B = 0.291335 \* (100)

lead\_A – lead\_B = 29.1335

🡪 We expect the lead concentration at site A to be **29.1335 ppm higher** than the concentration at site B.

g. Reporting and interpreting the R-Squared value.

summary(lead\_zinc\_model)

A close up of numbers

Description automatically generated



🡪 R2 value is **0.912**. This means that approximately 91.2% of the variance in lead concentration can be explained by the zinc concentration of the soil.

h. Commenting on whether the three main assumptions for linear regression are met for this data and listing any concerns.

a

Exercise 2:

Downloading the data from the course site and reading it into R under the name "ice". Further, converting the Date column into class “date”.

**Command:**

A close-up of a website

Description automatically generated

a. Producing a summary of a linear model of sea ice extent against time.

**Command:**

****

**Output:**

**A screenshot of a computer code

Description automatically generated**

b. Plotting the data and adding the regression line and checking for trends.

**Command:**

A computer code with green and black text

Description automatically generated

**Output:**

A graph with numbers and dots

Description automatically generated

🡪 Observing the data ,

**Linearity-** The regression line is almost flat, indicating no strong linear relationship in the data provided.

**Symmetry-** Cannot be assessed without a residual plot; the graph doesn't provide information on the distribution of residuals.

**Equal Variance-**  The spread of data points does not obviously change across the range of dates, but some clusters suggest variability, so this condition is uncertain without further analysis.

Therefore, there doesn’t seem to be a clearly defined trend in the given data.

c. Plotting the residuals of the model over time and adding a horizontal line. What assumption(s) about the linear model should we be concerned about?

**Command:**

**A screenshot of a computer code

Description automatically generated**

**Output:**

**A graph of a graph with circles and lines

Description automatically generated with medium confidence**

🡪 Assumptions that could be of concern are:

**Equal Variance-** While there is no clear pattern suggesting increasing or decreasing variance in the residuals, the clustering of residuals could be a sign of equal variance, which would be an assumption of concern.

**No outliers**- Significant outliers can affect the regression line. In this plot, there don't appear to be extreme values, but further analysis would help identify any influential outliers.

**Linearity of Relationship:** The assumption is that the relationship between the predictors and the response is linear. In the plot, there's no strong indication of non-linearity, but this could be more thoroughly checked.

Exercise 3:

a. Calculating the chance Adam will double his money in the first round of the game and the chance Adam will lose his money in the first round of the game.

When you roll 2 dice, there are 36 unique combinations.

- How many ways are there to roll two dice that add up to 7?

(1,6), (2,5), (3,4), (4,3), (5,2), (6,1): 6 combinations

- Add up to 11?

(5,6), (6,5): 2 combinations

🡪 In total, **8/36 = 2/9 probability** that Adam doubles his money in the first round.

p

b. Approximating the results in (a) by simulation.

**Command:**

A computer code with text

Description automatically generated with medium confidence

**Output:**

A close-up of a computer code

Description automatically generated

Visualizing the distribution of these outcomes.

**Command:**

A screenshot of a computer

Description automatically generated

**Output:**

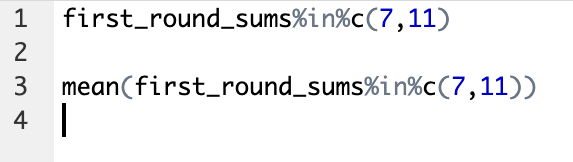
**A graph of s outcomes

Description automatically generated**

c. Calculating the percentage of time Adam doubled his money and the percentage of time Adam lost his money.

- Doubled his money:

**Command:**



**Output:**

A screenshot of a computer generated image

Description automatically generated

A screenshot of a computer

Description automatically generated

🡪 Therefore the percentage of time Adam doubled his money is approx 0.2188 = **21.88%**

- Lost his money:

**Command:**

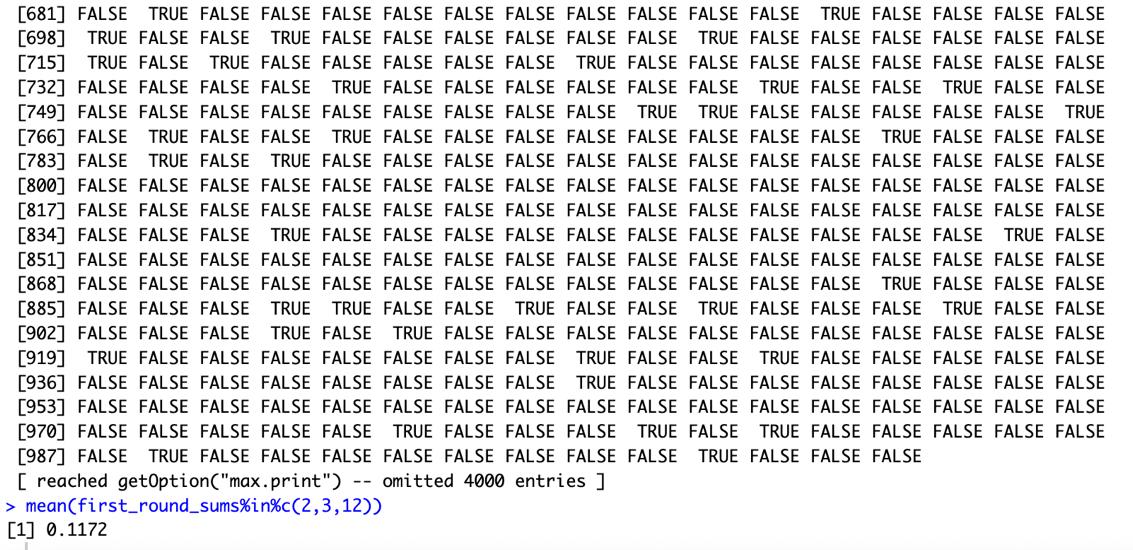
A close-up of a number

Description automatically generated

**Output:**

**A screenshot of a computer generated image

Description automatically generated**

****

🡪 Therefore the percentage of time Adam lost his money is approx 0.1172 = **11.72%**

d. Explaining why the two events of Adam winning money and losing money are independent, disjoint, or both.

🡪 Are winning and losing money independent?

**Intuitively:** If two events are independent, then information about the first event will not impact your belief in whether or not the second event will occur. If we know that Adam doubled his money, we instantly know that he could not have lost any money. Thus, the events are not independent.

**Mathematically:** If two events are independent then P(A and B) = P(A)\*(B).

P(winning and losing money) = 0, P(winning) \* P(losing) = 2/9 \* 1/9 = 2/81.

Thus, the events are not independent.

🡪 Are winning and losing money disjoint?

**Intuitively:** Two events are disjoint if they cannot occur at the same time. Can Adam lose and win money at the same time? No, because the dice cannot sum up to two numbers at the same time. Thus, the events are disjoint.

**Mathematically:** Two events are disjoint if P(A and B) = 0. We know that it is impossible for Adam to win and lose money at the same time, so P(winning and losing) = 0.

Thus, the events are disjoint.

e. Mathematically verifying if those events are independent.

Mathematically, we can check independence using the formula: P(A∩B) = P(A)⋅P(B)

P(Winning and Losing) = 0

P(Winning)⋅P(Losing) = 2/9.1/9 = 2/81

Since, P(Winning and Losing) ≠ P(Winning)⋅P(Losing), the events are not independent.

Additionally, since P(Winning and Losing) = 0, the events are disjoint.

So, both the intuition and the mathematical verification agree that winning and losing money are not independent events, and they are disjoint.

**Part II**

Exercise 1

Possible grades in a history course are A, B, C, or lower than C; probability that a randomly selected student will get an A in the course is 0.32, probability that a student will get a B in the course is 0.21, probability that a student will get a C in the course is 0.23.

a. Probability that a student will get an A OR a B:

Given that a student cannot receive an A and B at the same time, these events are mutually exclusive / disjoint.

P(A or B) = P(A) + P(B)

0.32 + 0.21 = 0.53

Thus, the probability of a student getting an A or B is **53%.**

b. Probability that a student will get an A OR a B OR a C:

Given that a student cannot receive an A or B or C at the same time, these events are mutually exclusive / disjoint.

P(A or B or C) = P(A) + P(B) + P(C)

0.32 + 0.21 + 0.23 = 0.76

Thus, the probability of a student getting an A or B or C is **76%.**

c. Probability that a student will get a grade lower than a C:

P(<C) = 1 - (P(A) + P(B) + P(C))

1 - (0.32 + 0.21 + 0.23) = 0.24

Thus, the probability that a student will get a grade lower than a C knowing the probability of those getting an A or B or C (76%) is **24%.**

Exercise 2

De Mere’s Dice Problem

a. Find P(E) if E is the event of getting at least one six in four rolls of a single die:

P(E) = 1 - (⅚)4

= 1 - (625/1296)

= 671/1296

= 0.5177

The probability of getting at least one six in four rolls of a single die (P(E)) is approximately **51.77%.**

b. Find P(F) if F is the event of getting at least one double six in 24 throws:

P(F) = 1 - (35/36)24

= 1 - (1.1419131e+37/2.2452258e+37)

= 1 - 0.5086

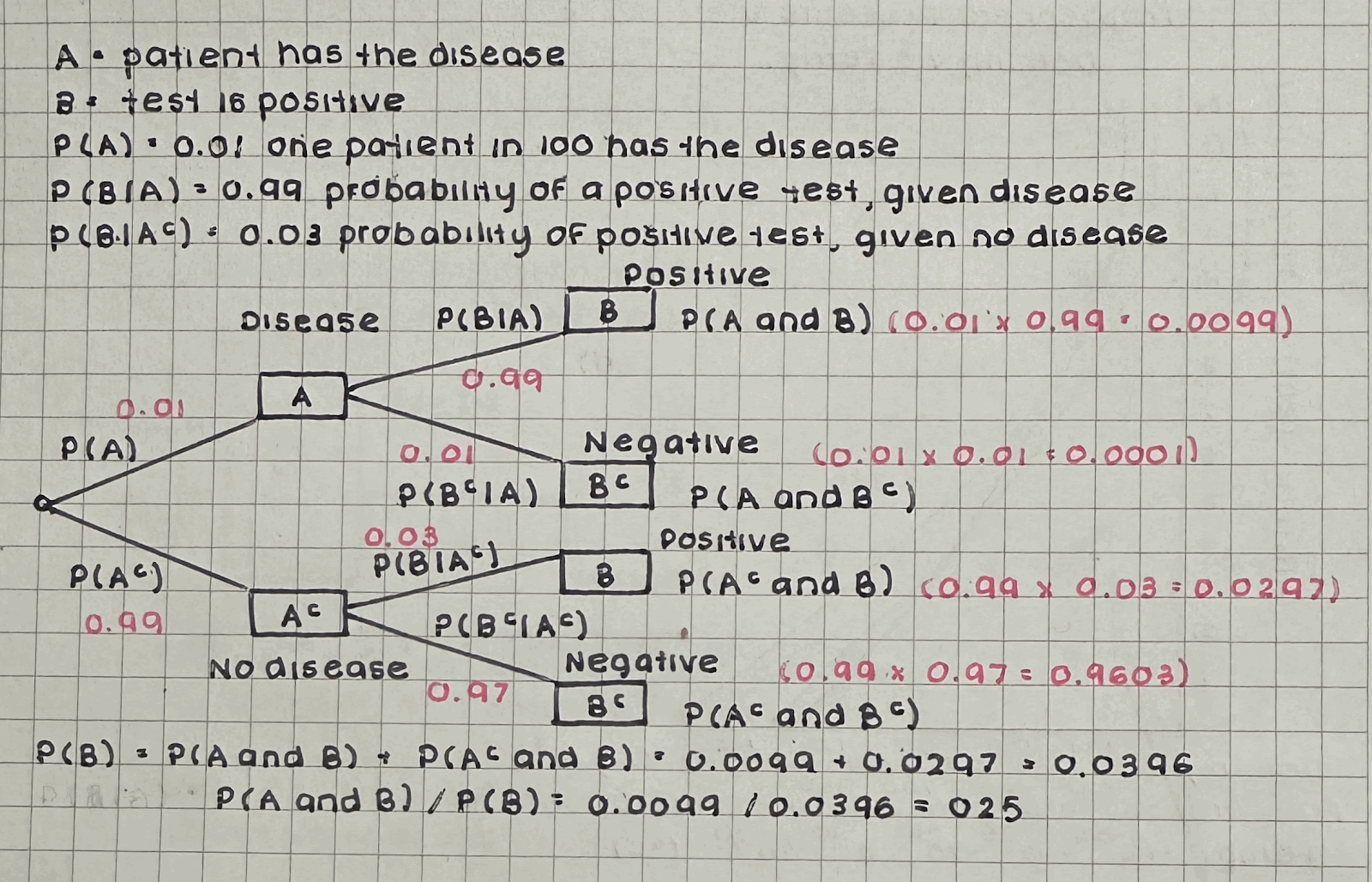
= 0.4914

The probability of getting at least one double six in 24 rolls (P(F)) is approximately **49.14%.**

Exercise 3

Given that the test comes back positive 99% of the time for people who have the disease, comes back negative 97% of the time for people who do not have the disease, and the disease affects 1 in 100 people in the country.

Probability that the patient actually has the disease if the test result for the patient came back positive:

 🡪 Given that the test result came back positive, the probability that the patient actually has the disease is 0.25.

Exercise 4

Given that a fair coin is flipped 100 times and the recorded results are 58 heads and 42 tails.

a. Theoretical probability and empirical probability of getting heads:

Theoretical probability of getting heads = ½ = **0.5**

Empirical probability of getting heads = 58/100 = **0.58**

b. Theoretical probability and the empirical probability of getting tails:

Theoretical probability of getting tails = ½ = **0.5**

Empirical probability of getting tails = 42/100 = **0.42**

c. Empirical probability observed if the coin was flipped 1000 times and proportion of times getting heads is recorded:

As the number of trials increases, the empirical probability of getting heads gets closer to its theoretical probability (½). If we try to use empirical probability from 100 heads and run an experiment, then 0.58 x 1000 = 580. But since theoretical probability is 0.5, then 0.5 x 1000 = 500 times. The upper bound is 580, but since the theoretical value should be 500, we would expect a number between 500 and 580 but closer to 500 due to the law of large numbers. Thus, we can expect an approximate empirical probability of **0.5.**

d. Real-life situation where empirical probabilities would be useful:

Empirical probabilities could be useful in real life for **sports forecasting and coaching**, such as knowing how players will perform for offense or defence based on past trials by testing each of their skills and setting them up according to what they excel at.

Exercise 5

Given the table displaying the outcomes of rolling a fair six-sided die in each of the three experiments conducted :

a. Empirical probability of rolling a 4 for 20 trials:

2/20 = **0.1**

b. Empirical probability of rolling a 4 for 100 trials:

20/100 = **0.2**

c. Empirical probability of rolling a 4 for 1000 trials:

166/1000 = **0.166**

d. Theoretical probability of rolling a 4 with a fair six-sided die:

P(4) = ⅙ = **0.167**

e. Comparing the empirical probabilities to the theoretical probability, and explaining what they show:

- Theoretical probability vs empirical probability of rolling a 4 for 20 trials = |0.1 - 0.167|

= **0.067**

- Theoretical probability vs empirical probability of rolling a 4 for 100 trials = |0.2 - 0.167|

= **0.033**

- Theoretical probability vs empirical probability of rolling a 4 for 1000 trials = |0.166 - 0.167| = **0.001**

🡪 Comparing the empirical probabilities to the theoretical probability of rolling a 4 with a fair six-sided die, we can observe that as we conduct more trials, the empirical probability value gets closer to its theoretical probability.