STATS 10 Assignment 3

Please submit both parts of the assignment in one single PDF file. You can use any PDF editor software to merge the two parts into one file. Please make sure that the questions are in the correct order and clearly labeled, and that the answers are legible and easy to read.
To submit your assignment, upload the PDF file under the designated assignment page on the course website before the deadline specified. Email or hard copy submissions are not accepted.

Part I

Include both the R commands and their corresponding outputs, results, or answers for all exercise questions in Part I.

Exercise 1

We will be working with some soil mining data and are interested in looking at some of the relationships between metal concentrations (in ppm). Download the data 'soil_complete.txt' from the course website and read it into R. When you read in the data, name your object "soil".

a. Run a linear regression of lead against zinc concentrations (treat lead as the response variable). Use the summary function just like in the example above and paste the output into your report.

b. Plot the lead and zinc data, then use the abline() function to overlay the regression line onto the data.

c. In a separate plot, plot the residuals of the regression from (a), and again use the abline() function to overlay a horizontal line.

Parts d-h can be answered by hand, using a calculator, or any R functions of your choice.

d. Based on the output from (a), what is the equation of the linear regression line?

e. Imagine we have a new data point. We find out that the zinc concentration at this point is 1,000 ppm. What would we expect the lead concentration at this point to be?

f. Imagine two locations (A and B) for which we only observe zinc concentrations. Location A contains 100ppm higher concentration of zinc than location B. How much higher would we expect the lead concentration to be in location A compared to location B?

g. Report the R-squared value and explain in words what it means in context.

h. Comment on whether you believe the three main assumptions (linearity, symmetry, equal variance) for linear regression are met for this data. List any concerns you have.

Exercise 2

Our next data set is what is known as a time series, or data in time. It contains the measurements via satellite imagery of sea ice extent in millions of square kilometers for each month from 1988 to 2011. Please download the "sea_ice" data from the course website and read it into R. If you have your working directory properly set, you can use the line below:

```
ice <- read.csv("sea_ice.csv", header = TRUE)
```

Note that currently R does not know what class the Date column is. We need to convert the Date column into class "date" using the following line:

```
ice$Date <- as.Date(ice$Date, "%m/%d/%Y")
```

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

b. Plot the data and overlay the regression line. Does there seem to be a trend in this data?

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

Exercise 3

One of Adam's favorite casino games is called "Craps". In the first round of this game, two fair 6-sided dice are rolled. If the sum of the two dice equal 7 or 11, Adam doubles his money! If a 2, 3, or 12 are rolled, Adam loses all the money he bets ☹.

a. Based on your lecture notes, what is the chance Adam will double his money in the first round of the game? What is the chance Adam will lose his money in the first round of the game?

b. Let's now approximate the results in (a) by simulation. First, set the seed to 123. Then, create an object that contains 5,000 sample first round Craps outcomes (simulate the sum of 2 dice, 5,000 times). Use the appropriate function to visualize the distribution of these outcomes (*hint: are the outcomes discrete or continuous?*).

c. Imagine these sample results happened in real life for Adam. Using R functions of your choice, calculate the percentage of time Adam doubled his money. Calculate the percentage of time Adam lost his money.

d. Adam winning money and Adam losing money can both be considered events. Are these two events independent, disjoint, or both? Explain why.

e. Quickly mathematically verify by calculator if those events are independent using part (a) and what you learned in lecture. Show work.

Part II

You may choose to type or write your answers electronically or scan your handwritten solutions. Please ensure that you show all steps and explanations to receive full credit, unless otherwise instructed.

Exercise 1

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

a. What is the probability that a student will get an A OR a B?

b. What is the probability that a student will get an A OR a B OR a C?

c. What is the probability that a student will get a grade lower than a C?

Exercise 2

De Mere's Dice Problem

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

b. Let F be the event of getting at least one double six in 24 throws. Find P(F).

Exercise 3

A patient is displaying some symptoms and received a disease screening test. The test comes back positive 99% of the time for people who have the disease, and comes back negative 97% of the time for people who do not have the disease. The doctor knows that the disease affects 1 in 100 people in the country. Suppose the test result for the patient came back positive, what is the probability that the patient actually has the disease?

Exercise 4
Suppose you flip a fair coin 100 times and record the results. You get 58 heads and 42 tails.

a. Find both the theoretical probability and the empirical probability of getting heads.

b. Find both the theoretical probability and the empirical probability of getting tails.

c. If you were to flip the coin 1000 times and record the proportion of times that you get heads, what empirical probability would you expect to observe? Why?

d. Give an example of a real-life situation where empirical probabilities would be useful.

Exercise 5
Three experiments, each comprising a different number of trials, were conducted. The table below displays the outcomes of rolling a fair six-sided die in each of these experiments. Answer the questions about empirical probabilities using the table. Compare the empirical probabilities to the theoretical probability, and explain what they show.

Outcome on Die	20 Trials	100 Trials	1000 Trials
1	3	20	169
2	4	20	166
3	4	14	167
4	2	20	166
5	4	13	166
6	3	13	166

a. What is the empirical probability of rolling a 4 for 20 trials?

b. What is the empirical probability of rolling a 4 for 100 trials?

c. What is the empirical probability of rolling a 4 for 1000 trials?

d. What is the theoretical probability of rolling a 4 with a fair six-sided die?

e. Compare the empirical probabilities to the theoretical probability, and explain what they show.

Please provide your answers using decimal numbers and round to three digits if needed.