STA2002 Probability and Statistics II
Assignment 2
Due date: 16 October 2020 5pm TC414 Assignment dropbox
Please submit an electronic copy of your answers on Blackboard too

Please answer all questions. Please work on the assignment by yourself only. We follow a strict rule on academic honesty.

1. (Dinosaur and star) Download the datasets "D.csv" and "S.csv" on Blackboard. They both have 2 columns and 142 data points. For the dataset "D.csv", we assume the usual linear regression model with
$$Y_i^D = \alpha + \beta(x_i^D - \overline{x^D}) + \epsilon_i^D, \quad \epsilon_i^D \overset{i.i.d.}{\sim} N(0, \sigma^2), \quad i = 1, \ldots, n,$$
and σ^2 is unknown. Similarly, for the dataset "S.csv", we assume the usual linear regression model with
$$Y_i^S = a + b(x_i^S - \overline{x^S}) + \epsilon_i^S, \quad \epsilon_i^S \overset{i.i.d.}{\sim} N(0, \gamma^2), \quad i = 1, \ldots, n,$$
and γ^2 is unknown. Use your favourite computing language to answer this question, and attach your code at the end. **Do not use any existing regression packages or functions in your choice of language for this question.**

   (a) Report the sample mean and sample variance for both x and y in the two datasets, that is, fill in the following table. Express all your answers in this question to 2 decimal places. There is no need to round up or round down.

	x^D	x^S	y^D	y^S
Sample mean				
Sample variance				

   (b) Report the maximum likelihood estimate of α, β, a, b respectively.

   (c) Construct 95% confidence intervals for α, β, a, b respectively. Use $z_{0.025} = 1.96$ for the confidence intervals instead of the t-distribution quantile.

   (d) Based solely upon the statistics you computed in part (a), (b) and (c), how do the two datasets compare?

   (e) Construct scatterplots for each of the two datasets. Surprise :)

   (f) What's the moral of the story? (That is, what does this example suggest about what should be done when analyzing data?)

2. Suppose you flip a coin 314 times, and heads appears 159 times.

   (a) Construct an approximate 90% confidence interval for the probability that the coin comes up heads.

   (b) Approximately how many samples would you need to obtain an approximate 90% confidence interval with width 0.02, while keeping exactly 50.64% of flips appearing heads?

   (c) You give the coin to a friend, who also flips the coin 314 times, and obtains an approximate confidence interval [0.390, 0.500] instead. What confidence level did (s)he use?

3. Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu_X, \sigma_X^2)$, $Y_1, \ldots, Y_m \overset{i.i.d.}{\sim} N(\mu_Y, \sigma_Y^2)$ and they are independent. Suppose that both σ_X^2 and σ_Y^2 are known.

(a) Construct a two-sided $100(1-\alpha)\%$ confidence interval for the difference $\mu_X - \mu_Y$.

(b) We want to obtain a 90% confidence interval for the difference between true average cable strengths made by Company X and by Company Y. Suppose cable strength is normally distributed for both types of cables with $\sigma_X = 50$ and $\sigma_Y = 30$. If we can make $n+m = 6000$ observations, how many of these should be on Company X cable if we want to minimize the width of the interval?

4. Michael owns a bakery. The number of chocolate chips that he adds to his cookies is normally distributed with some mean μ and known variance $\sigma^2 = 121$. A customer buys a dozen of these cookies, and obtains the sample

$$31,\ 41,\ 59,\ 26,\ 53,\ 59\,47,\ 43,\ 23,\ 34,\ 42,\ 44$$

for which the sample mean is $\bar{x} = 41.83$ and the sample variance is $s^2 \approx 11.8^2$.

(a) Compute a 90% confidence interval for μ.

(b) Compute 95% and 99% confidence intervals for μ.

(c) Repeat part (a), assuming that the customer has no idea what σ^2 is.

5. Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$. In this question, we will construct confidence interval for σ^2. Let $C \sim \chi^2(r)$, and as usual for any $\alpha \in [0,1]$ we denote $\chi^2_\alpha(r)$ to be

$$P(C > \chi^2_\alpha(r)) = \alpha.$$

(a) Suppose that μ is known. By considering the distribution of

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2,$$

prove that

$$\left[\frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi^2_{\alpha/2}(n)}, \frac{\sum_{i=1}^n (X_i - \mu)^2}{\chi^2_{1-\alpha/2}(n)} \right]$$

is a $100(1-\alpha)\%$ confidence interval for σ^2.

(b) Suppose that μ is unknown. By considering the distribution of

$$\sum_{i=1}^{n} \left(\frac{X_i - \overline{X}}{\sigma} \right)^2 = \frac{(n-1)S^2}{\sigma^2},$$

prove that

$$\left[\frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\chi^2_{\alpha/2}(n-1)}, \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\chi^2_{1-\alpha/2}(n-1)} \right]$$

is a $100(1-\alpha)\%$ confidence interval for σ^2.

(c) In the same setting as part (b), that is, suppose that μ is unknown. Construct a $100(1-\alpha)\%$ confidence interval for σ.

6. (Weighted least squares) In lecture, we derive the least squares regression line. One popular alternative of least squares regression is called the weighted least squares. In weighted least squares, we instead minimize

$$\min_{\alpha,\beta} \sum_{i=1}^{n} w_i (y_i - \alpha - \beta x_i)^2,$$

where w_i for $i = 1, \ldots, n$ are some known weights and $(x_1, y_1), \ldots, (x_n, y_n)$ are the observations. Let α^* and β^* be the minimizers of α and β respectively. Prove that they can be written as

$$\beta^* = \frac{\sum_{i=1}^n w_i \sum_{i=1}^n w_i x_i y_i - \sum_{i=1}^n w_i x_i \sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i \sum_{i=1}^n w_i x_i^2 - \left(\sum_{i=1}^n w_i x_i\right)^2},$$

$$\alpha^* = \frac{\sum_{i=1}^n w_i y_i}{\sum_{i=1}^n w_i} - \beta^* \frac{\sum_{i=1}^n w_i x_i}{\sum_{i=1}^n w_i}.$$

7. (Traffic volume) Michael lives in Minneapolis. He is interested in the traffic volume of Minneapolis. Download the dataset "traffic.csv" on Blackboard. There are 9 columns and 48,204 data points. We will only use two columns for this question: "weather_main" and "traffic_volume". "weather_main" is a short description of the weather, and "traffic_volume" is the hourly traffic volume. Similar to Assignment 1, use your favourite computing language and attach your code at the end of your answer.

   Note: the dataset in this question is a real-world dataset from UC Irvine Machine Learning depository. The link is here: CLICK.

   (a) Michael is interested in doing a comparison between the hourly traffic volume when the weather is "Clear" verus the hourly traffic volume when the weather is "Rain". He has the feeling that there is less traffic when the weather is "Rain", and would like to derive some statistical insights from the data to support his idea. What is the sample mean of the traffic volume when the weather is "Clear", say $\overline{x}$? What is the sample mean of the traffic volume when the weather is "Rain", say $\overline{y}$?

   (b) Assume the hourly traffic volume when the weather is "Rain" are $Y_1, \ldots, Y_m \overset{i.i.d.}{\sim} N(\mu_Y, \sigma^2)$, and the hourly traffic volume when the weather is "Clear" are $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu_X, \sigma^2)$, and X_i and Y_i are independent. These assumptions are not realistic but still provide us a model to work on. Compute the 95% two-sample pooled t-interval for the difference $\mu_X - \mu_Y$. As n and m are large in this dataset, please use $z_{0.025}$ to replace $t_{0.025}(n + m - 2)$ when you compute the interval.

   (c) Assume the hourly traffic volume when the weather is "Rain" are $Y_1, \ldots, Y_m \overset{i.i.d.}{\sim} N(\mu_Y, \sigma_Y^2)$, and the hourly traffic volume when the weather is "Clear" are $X_1, \ldots, X_n \overset{i.i.d.}{\sim} N(\mu_X, \sigma_X^2)$, and X_i and Y_i are independent and $\sigma_X^2 \neq \sigma_Y^2$. Compute the 95% Welch's t-interval for the difference $\mu_X - \mu_Y$. As n and m are large in this dataset, please use $z_{0.025}$ to replace the t distribution quantiles when you compute the interval.

   (d) At 95% confidence, are there really less cars in Minneapolis when the weather is "Rain"?