Due date: Friday Week 11, by 5pm

You must submit your assignment electronically and as a single file via the LMS page for this subject. Your solutions must include your workings. **Note that this assignment is worth** 10% **of your overall mark for this subject.** Advice for preparing an electronic version of your assignment is provided on **this LMS page (click here)** under the heading: 'Some ideas for submitting your assignments online'. Do not submit your file as a ZIP file.

In submitting your work, you are consenting that it may be copied and transmitted by the University for the detection of plagiarism. Please start with the following statement of originality, which must be included near the top of your submitted assignment:

"This is my own work. I have not copied any of it from anyone else."

Please round your answers to three decimal places if rounding is necessary. Alternatively, if you choose to express your answers as fractions, please ensure that the fraction is reduced to its simplest form. E.g. 10/20 should be expressed as 1/2.

1. Suppose you would like to compare the efficacy of two weight loss programs (programs X and Y). To control for the confounding potential of genetics, you collect data from twelve pairs of identical twins of similar weight. For each pair of twins, one twin is randomly assigned to program X, with the other twin then assigned to program Y. The amount of weight (in pounds) lost over a three-month period by each individual was recorded. Your aim is to determine if there is a difference in the efficacy of the two programs, using the data presented below.

	Weight lost (pounds)											
Program/Pairs	1	2	3	4	5	6	7	8	9	10	11	12
X	10.97	8.53	7.61	9.19	11.22	12.12	5.80	8.62	8.41	7.91	7.32	10.48
Y	11.82	10.05	8.54	10.45	12.62	14.26	6.83	10.21	9.52	8.72	9.25	12.53
X-Y	-0.85	-1.52	-0.93	-1.26	-1.40	-2.14	-1.03	-1.59	-1.11	-0.81	-1.93	-2.05

   (a) State your null and alternate hypotheses. *Hint: You are assessing the differences between X and Y.*

   (b) Calculate the appropriate test statistic by hand. Assume the differences follow a normal distribution.

   (c) Calculate the 95% confidence interval for the difference score.

   (d) State your conclusion based on the 95% confidence interval and the p-value.

   (e) This question should be carried out in R. You must include your R code and output in your answers. Carry out the appropriate test in R to verify your answer to (d). Use the R function **shapiro.test()** to confirm that the differences in weight losses between programs follow a normal distribution.

2. Suppose that a hotel chain conducts a comparison study of customer satisfaction at two of its hotels, using data collected over a one-month period. At Hotel W, 386 out of 402 customers gave positive feedback, while at Hotel Z, 544 out of 581 customers gave positive feedback. The remaining customers at both hotels gave negative feedback. Using this information, answer the following questions.

   (a) This question should be carried out in R. You must include your R code and output in your answers. Calculate the proportions p_w and p_z of satisfied customers at both hotels, and then carry out a hypothesis test comparing these two proportions, using the R function `prop.test`. From the R output, find and report the following:

      i. The estimates to p_w and p_z.

      ii. The approximate 90% confidence interval for $p_w - p_z$.

      iii. The p-value for the test comparing p_w and p_z.

   (b) Using the p-value you reported above, can you reject that the proportions are equal at the level of significance $\alpha = 0.1$? Explain.

   (c) Does your confidence interval suggest that one hotel performs better than the other with respect to the proportion of satsfied customers? If so, which hotel performs better and why? Otherwise, clearly explain why this is not the case.

   (d) Provide a simple statement that summarises the findings you have reported above and which could be understood by a non-statistician.

3. This question should be carried out in R. You must include your R code, output and plots in your answers.

In R, the sample data set **Loblolly** contains information on the variables *height* (in feet) and *age* (in years) of Loblolly pine trees. For the purposes of this question, ignore the variable *Seed*. Suppose that US scientists are concerned with the growth rate of Loblolly pine trees in Texas, and would like to know, using this sample data set, if the relationship between the height and age of these trees can typically be modelled using a linear model.

Note that the data set Loblolly is already loaded in R - you can consider it as a 'pre-defined' data.frame object. E.g. to access it, simply type Loblolly in the console. You do not need to download the data set from an external source.

(a) Using the R command *scatter.smooth()*, create a scatter plot with a smooth line, to visualize any linear relationship between the dependent (response) variable and independent (predictor) variable in the **Loblolly** data set.

(b) Produce individual boxplots for the variables *height* and *age* and check for outliers.

Hint: To produce 2 or more plots side-by-side, using the command par(mfrow(nrows, ncols)), replacing nrows and ncols with the desired numbers of rows and columns respectively.

(c) Execute the following R commands to obtain least squares estimates and associated output for the **Loblolly** data.

```
loblolly_lm <- lm(height ~ age, data = Loblolly)
summary(loblolly_lm)
```

From the R output, find and report the following:

i. What are the intercept and age coefficient estimates? Interpret these values.

ii. Does the R output suggest that your regression model fits the data well? Explain.

iii. Create a Residual versus Fitted plot and a Normal Q-Q plot of the standardised residuals for your regression model. Do the Residuals versus Fitted plot and/or the Normal Q-Q plot of the standardised residuals suggest that there are any linear regression model violations about which we need to be concerned? Justify your answer with references to both plots.

(d) Let β_1 denote the true coefficient for the *age* explanatory variable and consider the hypotheses

$$H_0 : \beta_1 = 0 \text{ versus } H_1 : \beta_1 \neq 0.$$

Do you reject the null hypothesis at the level of significance $\alpha = 0.05$? Explain.

(e) Repeat (d), but this time for the intercept coefficient. You may denote this coefficient as β_0.

(f) Construct 95% confidence intervals for β_1 and β_0. *Hint: Inspect the output from part c, and use the R command qt() to find the appropriate percentile of the t-distribution with the appropriate degrees of freedom*

(g) Using your *loblolly_lm* regression model, what is the estimated height for a 9 year old Loblolly pine tree?

(h) Provide a 95% confidence interval and 95% prediction interval for the height of the tree considered above in (g). Provide a justification as to why these intervals are different?

Hint: Discuss the context in which these intervals would be used.

(i) Based on your analysis, is there statistical evidence to suggest that as the Loblolly pine trees age, their height increases? Explain.