

### ***Instructions:***

1. Please use R markdown to produce your document as a PDF or Word document.
2. Provide all of your R commands with an R markdown code chunk.
3. If you produced a Word document, then please convert your Word document to a PDF format.
4. Submit your answers as a PDF in brightspace.

## **Questions**

1. Consider the data in the file **P12-50-Startups.csv**. It is a description of 50 startup companies.
  - (a) Import the data and display a few rows.
  - (b) There is a categorical variable in the dataframe. Display its levels.
  - (c) Fit a linear model to describe the profit as a function of the other variables, and (i) Give the estimated model for a company in New York; (ii) Give the estimated model for a company in California.
2. Melanoma is a type of skin cancer which forms from melanocytes. Consider the data in the file **SkinCancer.csv**. It contains the latitude of the largest city in each state or province that was used as an estimate of geographical center of population. The mortality for the male population is the number of deaths per year per 100,000 individuals.
  - (a) Fit a linear model to describe the mortality rate for the male population against the latitude of the state or province.
  - (b) Produce a scatter plot of melanoma mortality rates for the male population against the latitude of the state or province and overlay the estimated line from (a) onto the plot.
  - (c) Give a 95% prediction interval of the mortality rate for the male population for a city with a latitude of 40.
3. Consider the data in the file **Preparation.csv**. It is data from a study comparing the effect of two preparations of a virus on tobacco plants. For each plant, half of a leaf is inoculated with preparation 1 and the other half is inoculated with preparation 2. The number of lesions are measured (they are columns 2 and 3 in the dataframe).

- (a) Construct a variable called `diff` which is the difference between the number of lesions under preparation 1 and preparation 2. Produce a normal qq-plot for the difference. Does it appear reasonable to assume that the difference is normally distributed?
- (b) Produce a paired data plot to display the number of lesions under each preparation for each plant.
- (c) Give the sample size `n`. That is, give the number of plants involved in the study.
- (d) Give the mean, and the standard deviation for the number of lesions under each preparation.
- (e) Conduct a paired t-test to compare the number of lesions under each preparation. Give a conclusion within the context of the problem.

4. Consider the data in the file `sales.csv`.

- (a) Import the data and display a few rows.
- (b) Coerse the variable `design` to be a factor, and display its levels.
- (c). Fit a linear model to describe the sales according to design, and apply Levene's test on this fitted model. Levene's test is used to assess what? Give the conclusion of Levene's test.
- (d) Produce a qq-plot for the studentized residuals of the fitted linear model. Is it reasonable to assume that the populations are normal?
- (e) Assume that it is reasonable to assume that the populations are normal with equal variance. Conduct an ANOVA to determine whether of not the mean sales differ according to design.
- (f) Assume that it is reasonable to assume that the populations are normal with equal variance. Use the Tukey procedure to compare the mean sales according to design differ pairwise. What are your conclusions?
- (g) Give a comparative boxplot (with an overlay of a jitter plot), to visualize the distribution of the sales according to design. Surperimpose onto the plot letters from the Tukey procedure from (f).