**MATH 3070: Final Project**

Your final project activity involves either the analysis of your own data or the completion of one of the suggested problems (see below), and should result in a short report (no longer than 5 pages) prepared for the activity of your choice: You may select data available at our e-Statistics web site, or find data someplace else. Your report must contain three basic components:

**Introduction.** It must include:

* *Statement of study objective:* What is the topic of your analysis? What is/are the study question/questions you planned to investigate? *One of the study questions, presumably the main question, must be answered by hypothesis test.*
* *Background:* Describe why you are interested in this particular study. What kind of background information do you have for the study?
* *Method for data collection:* How did you or the authors of article gather data (via experiment, survey, etc.)?

**Analysis.** This section should exhibit your understanding of data analysis, including graphics such as histogram, boxplots, or scatter plots used in your analysis. Discuss other calculations required for the completion of the study, and include:

* *Methods of statistical inference:* Describe what type of statistical inference, and state the hypotheses (null and alternative) you performed. It must be related to one of the following topics:
  1. inference on mean (lecture note #1 and #3);
  2. simple linear regression (lecture note #4 and #7);
  3. comparison of two groups, paired or unpaired (lecture note #3 and #5);
  4. analysis of variance (lecture note #6).
* *Results of data analysis:* The results include graphical and numerical summaries of the data, and the hypothesis test with respect to the study objective.

**Conclusion.** In this section you will summarize your findings. You may start with the highlight of the result of analysis you have written in the previous section, and explain what you can support or fail to support regarding the result. If you see the weakness in your conclusion, make suggestions for future study. Tell us what you have learned from this whole activity.

## Generative AI Policy. Generative AI can produce statistical calculations as well as R code, and create a statistical report using them. However, the AI providers are concerned with fundamental accuracy as well as legal issues, and have chosen to obscure their results on purpose. Consequently, it will randomly alter the output of R code and make sure that the statistical report remains fictitious. A possible remediation is to run the R code generated by the Generative AI tools in your machine, and revise their results accordingly.

## When your report is assisted by Generative AI tools, you must acknowledge them in your report, publish the R code, and take full responsibility. To indicate the use of Generative AI resources, a student should include the following statement: "I acknowledges the utilization of [Generative AI Tool Name], a language model developed by [Generative AI Tool Provider], in the preparation of this assignment, and I am fully responsible for the integrity of the content generated by it."

**Data and study outside suggested problems.** If your final project is not one of the suggested problems (see below), you must share the data set of your final project with the instructor by uploading it to iLearn.

**Suggested problems.** You may choose to complete one of the following problems as your final project activity.

1. It is expensive and cumbersome to determine the body fat in humans as it involves immersion of the person in water. It would therefore be very helpful if a regression model with some or all of these predictor variables could provide reliable predictions of the amount of body fat, since the measurements needed for the predictor variables are easy to obtain. The dataset contains the variables for body fat, thigh circumference, and mid-arm circumference. The data set is available at e-Statistics. Go under the title “Textbook Data Sets” → “Body fat in women.”

*Objective of study.* The researcher is interested in the model that best predicts body fat from one of the two variables, thigh or midarm circumference, that is easy to measure.

1. Researchers into the genetic disease sickle cell anemia are interested in how red blood cells adhere to endothelial cells, which form the innermost lining of blood vessels. A set of 14 blood samples are obtained, and each sample is split in half. One half of the blood sample is profuse over an endothelial monolayer of type A and the other half of the blood sample is profuse over an endothelial monolayer of type B. The two types differ in respect to the stimulation conditions of the endothelial cells. The data represent the number of adherent red blood cells per mm2.

The data set is available at e-Statistics. Go under the title “Textbook Data Sets” → “Red blood cell adhesion.”

*Study Question: Is there any evidence that the different stimulation conditions affect the adhesion of red blood cells?*

1. Reproduce the result obtained in the article “Are women really more talkative than men?” regarding “Word Counts by Male and Female” under “Data in Statistical Studies” → “Worksheet Data Sets.” Data set is available at e-Statistics. Then in your own analysis and conclusion, explain how you have chosen a test procedure and reached your own conclusion.
2. Apply the supplementary topic “Model Validation,” and revise your statistical report assignment No.2 on tumor size study by completing the following study questions. Assume again that we are interested in prediction of the tumor weight from the radioactivity.

Go under the title “Textbook Data Sets” → “Tumor size.”

*Study Questions:*

* 1. Make a residual plot for the model. Is it reasonable to assume variance homogeneity?
  2. Try a transformation of the tumor weight variable so that a linear regression model on the transformed data fits the data in a satisfactory way (in e-Statistics: “Descriptive Statistics” → “Transformed Data”)
  3. A patient shows his radioactivity value of 8, and his tumor weighs about 0.7. It that unusual?

1. Report your findings regarding “Realtor Study” under “Data in Statistical Studies” → “Worksheet Data Sets.” A realtor studied the relationship between yearly income of home purchasers and sale price of the house. Present a scatter plot. Find the fitted linear equation, and calculate the correlation coefficient. In the investigation, you should also answer the following study questions:
   1. Which variable, Income or Price, should be the explanatory variable?
   2. Are there any influential points? If so, which ones are they?
   3. Removing influential points as potential outliers, recalculate the linear regression.
2. Report your findings regarding “Biomass in soil” under the title “Textbook Data Sets”. In ecological research it is often of interest to quantify the biomass in the upper layers of soil. One commonly used measure of biomass is the weight of live animals, FW (fresh weight), in the respective layer. Some practical difficulties exist, however, when the FW is to be measured in the field. Instead, animals collected from the soil are dried and the dry matter, DM, is used to determine the FW. Ten individuals of the species *Isotoma notabilis* were collected and their dry matter and fresh weights were recorded. In the investigation, answer the following study questions:
   1. Which variable should be the explanatory variable?
   2. Specify the linear regression model for FW as a function of DM and estimate the parameters.
   3. How can you predict the fresh weight if the weight of dry matter of an animal is known, say it is 1.8 or 2.6 for example?
3. Report your findings regarding a linear regression associated with “Stomach experiment” under the title “Textbook Data Sets”. As explained in lecture note #4 (i) find the standard error for the slope parameter in the linear regression. (ii) Calculate a 95% confidence interval for the slope parameter. Then answer the following study questions:
   1. Is there evidence of an association between the hormone concentration and the time it takes for the stomach to become empty? What is your evidence?
   2. A physiological theory suggests that the time until the stomach becomes empty increases by 1.2 when the hormone concentration increases by 0.1. Do the current data contradict or support this theory? Why or why not?
   3. Consider two persons’ hormone concentrations, 0.20 and 0.30, respectively. For each person, calculate a 95% confidence interval for the expected time it takes for stomach to become empty.
4. Report your findings regarding “Sucrase Activity Study” under “Data in Statistical Studies” → “Worksheet Data Sets.” For the measurement of the sucrase activity, the pellet method is more time-consuming, but provides a more accurate measure than the homogenate method. The researchers are interested in whether they can estimate the pellet reading based on a particular homogenate reading.
   1. State the null hypothesis corresponding to each p-value you presented.
   2. Describe all your findings regarding the null hypotheses.
   3. Conclude the adequacy of linear model, and discuss evidence of relationship between the two methods