STAT 50, Group Project 1, Spring 2025

Instructions. This project assignment consists of two problems. The relevant Excel and Matlab data files are in the Files section (Quizzes-Projects folder) on Canvas. The project is due by Friday, April 11th. You can either submit a hand-written or printed copy of your document in-person or upload an electronic copy (preferably as a single pdf file) on Canvas but don't upload an Excel file or the original data sets via Canvas Assignments. You are allowed to collaborate in pairs or groups of size three and submit the assignment jointly. In that case, one group member can upload the combined document with the names of the collaborators on the first page (or in the comments section on Canvas). You may also choose to submit or upload your document individually by stating names of your collaborators, if any.

1. (2 pts) This problem is about using randomly generated binomial sample data to estimate a population proportion (or binomial success probability) p. Let $X \sim Bin(40, p)$ where p is unknown and can be estimated using sample proportion, $\hat{p} = X/n$. For example, if $x = 28$ successes occur in a random sample of $n = 40$ trials, then $\hat{p} = 28/40 = 0.7$ is the corresponding (point) estimate of p.

   Check the Files section —> Quizzes-Projects folder on Canvas for the Excel file Data_Project1.xlsx. The worksheet "binom" includes 200 random numbers from $Bin(40, p)$ population. The sample successes are in column A and the sample proportions are in column B of the worksheet.

   A histogram and a vertical boxplot of the sample proportions are also given there but feel free to use other software (e.g. R, Matlab etc.) to construct better histograms/boxplots.

   Briefly describe the center (e.g. using the mean and/or median), the variability (e.g. using the standard deviation, range and/or IQR) and shape of the $\hat{p}$ values from both summary measures and the visual evidence. Moreover, find an estimate of the standard error for $\hat{p}$ values.

2. This problem concerns applications to dissolved oxygen concentration levels in a water science scenario. It is essential to have a sufficient amount of dissolved oxygen in rivers for healty and sustainable aquatic life (for fish and other organisms). In particular, if this level drops below 5 mg/L especially in warm water, aquatic life is endangered.

   Consider a random sample of 130 measurements of oxygen concentration in a few locations of a river in California. The corresponding data set is in worksheet Oxygen of the same Excel file. It is also in the Matlab file Proj1_Pr2.mat if you want to work with Matlab.

   (a) (2 pts) Obtain a histogram, a boxplot and a normality plot (QQ-plot) of the data. Describe the shape of the data distribution. Does it appear to come from an approximate normal population? Why or why not? You may also consider applying *kstest* function of Matlab (to a standardized version of the sample data).

   (b) (1 pt) Estimate both the mean and standard deviation of oxygen concentration in the river. Moreover, determine the standard error of the mean estimate. Based on these measures, can you reasonably conclude that the average oxygen concentration in the river is above 5 mg/L? Briefly explain.