BIEN/CENG 2310
MODELING FOR CHEMICAL AND BIOLOGICAL ENGINEERING

HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY, FALL 2022

HOMEWORK #7 (DUE NOV. 29, 2022)

1. In the last homework, we created a statistical model to simulate polling error in a two-candidate election. This time, we will try to reach the same conclusion from a theoretical angle.

   (a) Suppose in the whole population, a fraction p of the people plan to vote for Candidate A, and the rest for Candidate B. We define a random variable M to take on the value of 1 if a randomly selected person plans to vote for A, and 0 if he/she plans to vote for B. Express the mean and standard deviation of M in terms of p.

   (b) We randomly select n people, and we ask who they will vote for. This forms our sample. From their answers, we determine the fraction of votes for A for this sample, x. Estimate the standard deviation of x using the Central Limit Theorem.

   (c) Using the Central Limit Theorem, find a general expression for 95% confidence interval of x in terms of p and n, such that p will be within this interval with 95% probability. This is often reported as the "margin of error" of a poll.

   (d) Suppose in an election, 53% of people voted for candidate A. Estimate the probability that a pre-election poll of sample size 400 would have predicted the election results incorrectly; that is, the polling data showed that B was preferred over A.

DELIVERABLES:

For all parts, submit your type-written answers or scanned hand-written answers on Canvas.

2. Recall the rocket problem we studied in Module 6. The altitude of the rocket as a function of time, $y(t)$ can be modeled by the second-order ODE:

$$\frac{d^2y}{dt^2} = -g - D\left(\frac{dy}{dt}\right)^2 \operatorname{sgn}\left(\frac{dy}{dt}\right); \quad y(t=0) = 0 \quad ; \quad \left.\frac{dy}{dt}\right|_{t=0} = v_0$$

where D is the drag coefficient per mass of the rocket, v_0 is the initial speed, and g is the acceleration due to gravity at the Earth's surface, $g = 9.8\ \mathrm{ms}^{-2}$. For simplicity, since our rocket is not flying very high, we will ignore the altitude dependence of the acceleration due to gravity.

Our unknown parameters are D and v_0. (Although we can adjust the initial propulsion force on the rocket, the actual velocity is hard to control exactly, so we decided to learn it from the data.) We install an altimeter on the rocket to measure the altitude once every second during the flight. The data is in the table `hw7_q2_data` in the provided file `hw7_data.mat`. The column `t` contains the time points (in seconds), and `y1` contains the measured altitudes (in meters) at those time points.

(a) Use `nlinfit`, `lsqcurvefit`, or `fitnlm` to obtain the least-square estimates for D and v_0, and also report their 95% confidence intervals.

(b) Out of curiosity, you decided to treat the acceleration due to gravity, g, as a parameter to be fitted as well. What least-square estimates do you get for D, v_0, and g? How does it compare to what you got in Part (a)? What lesson do you learn from this exercise about choosing parameters in nonlinear regression?

(c) To get more accurate estimates, you decide to repeat the experiment, but with greater initial propulsion force you used before. Let the initial velocity in your second attempt to be v_0'. You expect that the drag coefficient D should be the same both times. The altitude data of the second attempt is in the vector `y2` in the same table `hw7_q2_data` (the time points are the same). Perform a least-square fit on these two sets of data with `lsqcurvefit` to find D, v_0, and v_0'. How does your estimate of D and its confidence interval compare to what you had in Part (a)?

DELIVERABLES:

This question will be done in class as a demo. No need to submit anything.

3. Have you ever wondered about the $(n-1)$ in the denominator of the formula for the sample variance? In this problem, we will use simulation to verify that the sample variance:

$$s^2 = \frac{\sum_i (x_i - \overline{x})^2}{n-1}$$

is an unbiased estimator of the population variance, σ^2. We will also check that the statistic:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}$$

follows the Chi-square distribution, and learn how to test the equality of variances of two normally distributed distributions by the F-test.

(a) Write a MATLAB function that takes in the population mean, μ, and population standard deviation, σ, the sample size n, and the number of sampling experiments, K. To simulate one experiment, draw a sample of n values from this population and calculate the sample variance by the formula above. Assume that the population is normally distributed.

Your function definition should be:

```
function s2bar = verifySampleVar(mu, sigma, n, K)
```

The function should repeat the sampling K times, return the mean of the sample variances of the K samples, $\overline{s^2}$, and plot a histogram of χ^2 fitted to the Gamma distribution.

What does the mean of the sample variance approach? Try different inputs of μ, σ and n. How are the fitted parameters a and b of the Gamma distribution relate to the inputs? *(Hint: The Chi-square distribution is a special case of the Gamma distribution. You can look it up in the MATLAB help or Wikipedia.)*

(b) Suppose we are interested in testing if the variances of two normally distributed populations are the same or not. The appropriate statistical test to use is called the F-test. Use MATLAB to generate two samples of sizes n_1 and n_2 from the same population, and perform the F-test to test the equality of the two sample variances (s_1^2 and s_2^2). Repeat the experiment K times. Verify that it will incorrectly reject the null hypothesis (i.e., committing a Type I error) with a probability of α (the significance level). Your function definition should be:

```
function FPR = verifyFTest(mu, sigma, n1, n2, K, alpha)
```

Your function should return the false positive rate (the fraction of Type I errors among the K experiments) of the F-test, and plot a histogram of the F-statistic, s_1^2/s_2^2. Overlay a suitable F-distribution with the right parameters. (Do not fit the histogram; instead, look up what parameters we should use from the MATLAB help or Wikipedia.) Mark the critical region with vertical dotted lines. You are NOT allowed to use the MATLAB function `vartest2`, but you may use it to check your answers.

DELIVERABLES:

Part (a) will be done in class as a demo. Submit your MATLAB program `verifyFTest_<LastName>_<FirstName>.m` for Part (b) and an example plot it produces for any input parameter setting of your choosing (use a large K for a stable distribution).

4. You recently gave an exam to a class of 100 students as the instructor of a HKUST course. After the exam, you discovered that half of the class got hold of a past exam paper, which was unfortunately quite similar to the present one. You want to find out if having the past exam paper gives an unfair advantage to some students.

(a) Using MATLAB, perform a two-tailed, two-sample, equal-variance *t*-test to compare the mean exam scores of the two groups of students, those with the past exam paper, and those without. The function should return 1 if the null hypothesis (that the two groups have the same mean exam scores) is rejected, 0 otherwise. Use a significance level of 0.05. It should also output a p-value. You are NOT allowed to use the functions `ttest` or `ttest2` function of MATLAB. The function definition should be:

`function [] = examScores_<LastName>_<FirstName>(dataTable)`

where `dataTable` is a table with 3 columns: `haspaper` (a Boolean indicating whether the student has the past paper), `studytimes` (the amount of time she spent studying – see Part (b) below), and `scores` (the exam score she got). You can find the data in the table `hw7_q4_data` in the file `hw7_data.mat`.

(b) Upon inspection of the data, you realize that another variable you need to consider is how long the students studied for the exam. If they did not study much, the past exam paper would not have helped anyway. To test your idea, perform a linear regression on each group of students using the MATLAB function `polyfit` in the same MATLAB function from Part (a). Overlay the two fitted lines (one for each group) and the data points in the same plot, using different colors for the two groups. Compute the coefficient of determination, R^2, for both fitted lines, and put them in the title.

For each fitted line, compute the 95% confidence bounds of the regressed mean exam scores, using the provided `linear_muCI.m` function. Show them in your plot as dotted curves on either side of the fitted lines.

(c) Based on the results of the *t*-test in Part (a) and the regression in Part (b), do you believe the access to the past exam paper has helped the students achieve a higher score in the present exam? Explain how you reach this conclusion.

DELIVERABLES:

Submit your MATLAB program `examScores_<LastName>_<FirstName>.m`, the required plot of the regression for Part (b), and type-written or scanned hand-written answer for Part (c).

5. In Module 5, we learned that the "rate law" of a chemical reaction is often empirically determined, from which we can guess the reaction mechanism. Suppose we run the reaction $A \rightarrow B$ in a solvent (water) in a continuous flow reactor, with no A or B in the reactor initially, and A introduced at a fixed concentration $[A]_{in}$ in the feed stream. The inlet and outlet flow rates are equal (i.e., $F = F_{in} = F_{out}$) such that the volume of the reacting mixture, V, does not change. We measure the concentration of B in the outlet stream as a function of time, and fit the data points to the prediction of two possible models:

Reaction mechanism 1. A spontaneously converts to B:

$$r_B = k_1[A]$$

Reaction mechanism 2. Two molecules of A activate by collision, then convert to B:

$$r_B = k_2[A]^2$$

where r_B is the rate of generation of B per unit volume, k_1, k_2, are parameters to be determined by regression, and $[A]$ is the concentration of A in the reactor.

(a) In our first experiment, we set the inlet concentration of A, $[A]_{in}$ to be 1, and the ratio F/V (the reciprocal of the residence time) to be 1. The data is provided in the table `hw7_q5_data` in the file `hw7_data.mat`. The column `t` containing the time points of measurements, and `B1` containing the concentration of B in the outlet stream at those time points. Using MATLAB, fit the two models to the data points and provide the coefficient of determination, R^2, for each of the two fits. Overlay the best-fit curves and the data points in one plot. Which model appears to better explain the data? Why?

(b) Not satisfied with your answer in Part (a), you decide to do one more experiment with the same setup, but increasing $[A]_{in}$ to 1.2. The data for this new experiment is in the column `B2` in the same `hw7_q5_data` table (the time points are the same). Using `lsqcurvefit`, fit the data of <u>*both experiments simultaneously*</u> to find the best parameters for each model. This time, to avoid making the plot too difficult to read, provide separate plots (including the data points of both experiments and the fitted curves) for each of the two models. Which model appears to better explain the data? Why? Compared to Part (a), are you now more or less confident in your decision?

DELIVERABLES:

Submit any MATLAB programs you used for this analysis, including functions for solving the ODEs, and functions for performing the nonlinear regression. Also submit a typewritten or scanned handwritten write-up to show how you set up the ODEs, and answer the questions. You may insert the required plots into this write-up, or submit them as separate images.