Stat 500 **MIDTERM EXAM** **Fall 2020**

Due Date: Wednesday, October 7, at 2 pm. You may submit your paper any time before the due date and time to Canvas. **Late papers will not be accepted.**

Instructions: You must do your own work on this exam. You may not discuss this exam with anyone except the course instructor or the course TA. You may use SAS to perform calculations or construct graphs except for questions that require hand calculation. You should **not** submit pages of unedited computer output, but you may include specific tables or graphs in your report. Write concise answers that clearly describe the steps in your analysis and your conclusions. The full score is 100 points.

Problem 1 (10 points):

A researcher wants to conduct a randomized experiment to examine the effect of an expensive new drug compared with a control drug on running speed of mice. The population standard deviation for running speed of mice is 2.5 mph. The researcher obtained the new expensive drug only enough for 6 mice. The researcher wants to obtain a 95% confidence interval with width of at most 5mph for the difference between means of new drug and control group. How many mice should be used in the control group? ***Use hand calculation for this problem***.

Problem 2 (10 points):

A state highway department wants to study the wear characteristics of four different paints (paint A, B, C, and D) for marking lines on highways. Here is how the study is designed. Ten locations will be selected, each a mile (1600 meters) long. The locations were selected reflecting variations in traffic densities and weather conditions throughout the state. Paints A and B will be used at five randomly selected locations. At each of those five locations, two adjacent 200-meter road segments will be randomly picked. Paint A will be used to paint lines in the middle and along both sides of one road segment, and paint B will be used to paint lines in the middle and along both sides of the other road segment. Paints A and B will be randomly assigned to the two road segments at each of those five locations. Paints C and D will be applied in a similar manner at the other five locations. One year after exposure to weather and traffic, the road segment at each location will be evaluated and assigned a score to reflect wear with respect to the visibility of the paint. Higher scores will indicate less wear.

a. Examine the proposed study and comment on how well it incorporates the principle of designs (randomization, replication, and blocking). Be brief but specific.

b. If you think the design of this study could be improved without substantially increasing the cost, describe the changes you would make. Otherwise, simply state that no improvements are necessary.

Problem 3 (35 points):

In a study of crop losses due to air pollution at a farm in North Carolina, plots of green beans were grown in open-top field chambers and continuously exposed to various conditions of air pollution. Four plots were randomly assigned to each of four treatment groups. Treatment A is air without pollution; treatment B is air with sulfur dioxide pollution; treatment C is air with nitrous oxides pollution, and treatment D is air with both sulfur dioxide and nitrous oxides pollution. After a month, the plants were harvested and the yield (kg) of bean pods was recorded for each of 16 plots. The sample statistics of the observed yield for each treatment group is shown in Table 1.

Table 1: sample statistics of the observed bean pods yield for each treatment group

Treatment	Sample Mean	Sample Variance
A	31	2.3
B	29	2.4
C	25	1.8
D	27	2.3

a. (*10 points*) Complete the ANOVA table including columns as shown below.

Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**

b. (*2 points*) State the null and alternative hypothesis corresponding to the *F*-test in the ANOVA table in part (a). Define any parameters or symbols you use.

c. (*3 points*) Perform the *F*-test and state your conclusion while controlling the type I error rate at 5%.

d. (*5 points*) State the model assumptions for the ANOVA *F*-test.

e. (*10 points*) Construct a contrast to test whether air pollution results in crop loss. Give the expression of the contrast, using μ_A, μ_B, μ_C, and μ_D to denote the population mean yield for treatments A, B, C, and D, respectively. Test whether this contrast is zero or not, and state your conclusion about whether air pollution results in crop loss or not.

f. (*5 points*) For an experiment-wise error rate of 5%, compute the value of Tukey's HSD for this study. Use your answer to determine which pair of means are significantly different.

Problem 4 (15 points):

Current nitrogen fertilization recommendations for wheat include applications of fertilizer at specified stages of plant growth. An experiment was conducted to evaluate the effect of two fertilization timing schedules (Schedule 1 and Schedule 2) on plant nitrogen content measured by the stem tissue nitrate amount. The experiment was carried out in an irrigated field divided into four blocks such that plots within each block were in the same part of the water gradient. Two plots within each block were randomly assigned to the two schedules with the same fertilizer amount (at a low level). The observed nitrate nitrogen content (ppm $\times 10^{-2}$) from each plot is shown in Table 2.

Table 2: Observed nitrate nitrogen content (ppm $\times 10^{-2}$)

Block	Schedule 1	Schedule 2
1	34.98	37.18
2	41.22	45.85
3	36.94	40.23
4	39.97	39.20

a. (*3 points*) What design is this experiment?
b. (*3 points*) What are the experimental units?
c. (*9 points*) Use a non-parametric test to detect any difference between fertilization timing schedules. Specify the following parts in your answer: (i) the name of the non-parametric test that you choose to apply, (ii) show your formula to calculate the **exact** *p*-value and calculate the numerical answer of the exact *p*-value, (iii) your conclusion in the context of the study.

Problem 5 (30 points)::

In a study to investigate the effects of group therapy on survival times of breast cancer patients, 56 breast cancer patients were randomly assigned to either a control group or a treatment group. Patients in the treatment group received weekly 90 minute sessions of group therapy that involved discussion and support for coping with the patients' quality of life during treatment, but patients were not led to believe that group therapy would affect the progression of their disease. Data on months of survival for each of 56 patients can be found in the file **survuval.csv**. Suppose that you are the statistician working on this dataset to answer the research question of whether group therapy treatment is effective or not in increasing survival time. Perform an appropriate analysis with SAS to answer the research question. For this question, you will be graded on the quality of your investigation of the data, the justification of your method, the interpretation of your results and conclusions, and how well your analysis addresses the objective of the study. Provide the following parts in your report of analysis.

a. (*4 points*) Identify whether this study is an observational study or an experiment. Justify your answer.
b. (*10 points*) A brief description of your examination of the dataset. Comment on how it leads you to the final analysis reported in part (c). Provide evidence (e.g., in the form of graphs, tests, diagnostic methods) to justify the appropriateness of your method for statistical analysis. Excluding graphs and tables, this should not exceed one page.
c. (*10 points*) A detailed report of your final analysis. Describe the null and alternative hypothesis. List any assumptions you make for your statistical method. Report a test statistic and the corresponding *p*-value. Excluding graphs and tables, this should not exceed one page.
d. (*6 points*) A one paragraph summary of your findings and conclusions with respect to the research question. This paragraph should not contain any formulas, tables, or graphs. Write the summary in a manner that would be understood by medical graduate students who have taken one introductory undergraduate course in statistics.