Name: _____ Spring, 2021

Applied Statistics Comprehensive Examination

- Calculators are permitted on this examination.
- When you compute a confidence interval, always give an interpretation of the interval in the context of the problem.
- When you perform a hypothesis test, always write down the null and alternative hypotheses, and write the conclusion in the context of the problem.
- There are 200 points on this examination.
- You must give complete explanations to receive full credit.
- Please put your answers and explanations on the separate sheets provided. Please use only the front side of these sheets.

1. (30 points) The following 16 data points are the "seeds" of the "Sweet Sixteen" of the 2021 Men's College Basketball tournament:

   1, 1, 1, 2, 2, 3, 4, 5, 5, 6, 7, 8, 11, 11, 12, 15.

   Below are several summary statistics for these 16 data points:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	St. Dev.
1.000	2.000	5.000	5.875	8.750	15.000	4.425

   The tournament starts with 64 teams* split into four separate regions, and in each region the teams are assigned a "seed" 1 through 16, with smaller-valued seeds indicating that the team is stronger. Thus, among the 64 teams, there are four of each seed (that is, there are four teams assigned a 1 seed, four teams asssigned a 2 seed, etc.)

   The tournament is a single-elimination tournament, meaning that when teams play each other, the winning team advances and the losing team is eliminated. The first round of games reduces the field from 64 to 32 teams, and the second round reduces the field to 16 teams. These 16 teams are the "Sweet Sixteen," with the seeds for this year's tournament shown above.

   Due to various impacts of the COVID-19 pandemic, it has been speculated that the results from the first couple rounds of the tournament may be particularly unusual in various ways. This question will explore that possibility based on the data above.

   * *This is not including the initial play-in round which we will ignore for the sake of this question.*

   (a) (10 points) From historical data, the average value of the seeds in the Sweet Sixteen prior to this year is 4.41. Perform a parametric test to determine whether this year's Sweet Sixteen seeds have a higher average seed than 4.41. In doing so, state your null and alternative hypotheses, define any symbols used, and report your conclusions at $\alpha = 0.05$, including a description in context.

   (b) (5 points) State the conditions required for the test you performed in part (a), and comment on whether they are satisfied or can be reasonably assumed based on the description of the data above. If they are not reasonable, comment on the impact that this might have on the validity of the test that you performed.

   (c) (10 points) We might also like to investigate if the values of the seeds in the Sweet Sixteen are more varied than we would expect. Based again on historical data, suppose that the true standard deviation of the Sweet Sixteen seeds under typical circumstances is $\sigma = 2.84$. Perform a parametric test using the data above to determine whether there is significant evidence to conclude that the seeds of this year's Sweet Sixteen have a standard deviation higher than $\sigma = 2.84$. In doing so, state your null and alternative hypotheses, define any symbols used, and report your conclusions at $\alpha = 0.05$, including a description in context.

   (d) (5 points) State the conditions required for the test you performed in the previous part, and comment on whether they are reasonable based on the description of the data. If they are not reasonable, comment on the impact that this might have on the validity of the test that you performed.

2. (25 Points) An experiment was conducted to determine the effect of three different pesticides and two different fertilizers on the yield of fruit from a citrus tree. Twelve trees were randomly selected from an orchard. Each of the six pesticide by fertilizer combinations were then randomly assigned to two of the trees. The yield of fruit, in bushels per tree, was obtained for each tree after the test period. The mean yield for each pesticide by fertilizer combination is given in the following table:

	Fertilizer	
Pesticide	1	2
1	44	48
2	39	59
3	42	52

A partially completed ANOVA table is given below:

Source	df	SS	MS
Pesticide			9.33
Fertilizer		385.33	
Interaction		130.67	
Error			
Total		592.67	

Suppose the following effects model was fit to the data:

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk},$$

where μ is the overall mean, α_i is the effect of the ith pesticide (i=1, 2, 3) , β_j is the effect of the jth fertilizer (j = 1, 2), γ_{ij} is the interaction effect of the ith pesticide and jth fertilizer, and ϵ_{ijk} is the random error term.

(a) (5 Points) Create interaction plots and discuss the type of interaction found in the data.

(b) (10 Points) Fill out the partially completed ANOVA table and test the significance of the interaction. How should the statistician proceed with the analysis from this point if the goal of the study is to determine the pesticide and fertilizer combination that maximize yield? Do not actually perform the additional analyses, just discuss next steps.

(c) (10 Points) Is $\gamma_{11} - \gamma_{12} - \gamma_{21} - \gamma_{22}$ estimable? Explain.

3. (27 points) A researcher is studying three different methods for teaching English as a second language, which we will call Method 1, Method 2, and Method 3. Sixty participants were randomly assigned to the three methods (20 for each), and each participant was given an exam to measure English proficiency after completing the program. Let y be the score on the exam. The researcher uses the following model to compare the effectiveness of the three methods:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_3 + \beta_5 x_2 x_3 + \epsilon,$$

where

$$x_1 = \begin{cases} 1, & \text{Method 2,} \\ 0, & \text{otherwise,} \end{cases} \qquad x_2 = \begin{cases} 1, & \text{Method 3,} \\ 0, & \text{otherwise,} \end{cases}$$

and x_3 is a baseline score that indicates each participant's English proficiency prior to the beginning of the study. The baseline score (x_3) is 0 if a participant had not studied English before the program.

(a) (10 points) Write out three separate models, one for each method, relating the score after three months in the program, y, to the baseline score, x_3. Make sure to provide the distributional assumptions for the error term ϵ.

(b) (7 points) Interpret β_3 and β_4 in the context of this study.

(c) (10 points) The researcher believes that, for those who have not studied English before, there is no difference among the three methods in terms of the mean score after completing the program. Describe how to test the researcher's theory. Make sure to state the hypotheses, the general form of the test statistic, the underlying probability distribution (including degrees of freedom), and the decision rule.

4. (50 points) A retrospective study of women in Washington state from March to June 2020 appeared to indicate that pregnant women were more likely to be infected with COVID-19 than non-pregnant women of reproductive age. Suppose that the following is a random sample of the data represented in this study.

	Pregnant Women	Non-Pregnant Women
Tested Positive for COVID-19	95	125
Tested Negative for COVID-19	55	125

In the following, you will use two different strategies for comparing positivity rates for pregnant and non-pregnant women.

(a) (10 points) Use a z-test to decide if the proportions of pregnant women and non-pregnant women testing positive for COVID-19 are different at the $\alpha = 0.05$ level. State the null and alternative hypotheses, define any symbols used, and report your conclusions including a description in context.

(b) (5 points) State the required conditions for the validity of the test in part (a). Identify any concerns you may have regarding the conditions in this situation.

(c) (5 points) Discuss how an increase in sample size may affect the results of your test in part (a).

(d) (10 points) Now use a χ^2 test to decide if the proportions of pregnant women and non-pregnant women testing positive for COVID-19 are different at the $\alpha = 0.05$ level. State the null and alternative hypotheses, define any symbols used, and report your conclusions including a description in context.

(e) (5 points) State the required conditions for the validity of the test in part (d). Identify any concerns you may have regarding the conditions in this situation.

(f) (5 points) Discuss how an increase in sample size may affect the results of the test in part (d), including the number of degrees of freedom.

(g) (10 points) Suppose that we wanted to use a single-proportion z-test to determine at the $\alpha = 0.05$ level if the proportion of pregnant women in the population of interest who test positive for COVID-19 is greater than 0.5. What is the power for this test if the sample size is 150 and the true proportion who test positive is 0.55?

5. (35 Points) The following table gives coagulation times (seconds) for samples of blood drawn from 12 animals receiving three different diets. The diets were randomly allocated to the animals. The data are:

	Diet A	Diet B	Diet C
	63	62	68
	67	60	66
	71	63	71
	64	59	67
Sum:	265	244	272
Mean:	66.25	61	68
SD:	3.59	1.83	2.16

(a) (5 Points) Write an appropriate model and describe each component. Be sure to state the assumptions of the model.

(b) (15 Points) Fill out an analysis of variance table and test to determine at the 0.05 level if there is a significant difference between the diets in terms of average time to blood coagulation. You may use the fact that the total sum of squares equals 168.92. Whether or not a significant difference exists, use Tukey's Honest Significant Difference (HSD) multiple comparisons procedure to compare the diets in an attempt to determine which diet(s) minimize the average time to blood coagulation.

(c) (5 Points) Construct two orthogonal contrasts for main effects and explain why they are orthogonal.

(d) (10 Points) Estimate the difference between the mean coagulation time for Diet B and the average of the mean coagulation times for Diets A and C with a 95% confidence interval.

6. (33 points) The following model relates the carbonation level of a soft drink beverage (y) to the temperature of the product (x_1) and filler operating pressure (x_2).

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,$$

Table 1 presents the R output of the regression analysis for this model, and Figure 1 shows the residual plots.

(a) (7 points) Construct and interpret a 95% confidence interval for β_1 in the context of this study.

(b) (10 points) List all model assumptions. Use the residual plots in Figure 1 to assess the assumptions and comment on how to address the violations if there are any.

(c) (8 points) Describe how to calculate PRESS residuals and explain the purpose of them.

(d) (8 points) Explain what extrapolation is in the context of this problem, what issues it may cause, and how you would detect it.

Table 1: R output of the regression analysis.

```
Call:
lm(formula = y ~ x1 + x2, data = carb)

Coefficients:
             Estimate   Std. Error  t value   Pr(>|t|)
(Intercept) -134.9957    14.7954     -9.124    9.77e-10 ***
x1             1.3006     0.5316      2.447    0.0212 *
x2             4.5804     0.1855     24.694   < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.9422 on 27 degrees of freedom
Multiple R-squared:  0.9713,Adjusted R-squared:  0.9692
F-statistic: 457.3 on 2 and 27 DF,  p-value: < 2.2e-16
```

Figure 1: Residual plots