Name: _____ Fall, 2017

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.

1. (20 points) A large company plans to offer employees a choice of three different health insurance plans. Human resources has projected that 50% will choose Plan A, 30% Plan B, and 20% Plan C. To test these projections, the company has sampled 120 employees and asked which of the three plans they would prefer. Among the 120 sampled employees, 59 indicated a preference for Plan A, 34 for Plan B, and 27 for Plan C. Using a significance level of 0.05, test the projections made by human resources.

2. (30 points) A manufacturing plant has had difficulty maintaining good production rates in a catalyst plant. An experiment was run to investigate the effect of four reagents (A, B, C, D) and three catalysts (X, Y, Z) on production rates. Experimentation was costly and time-consuming. Thus, only one observation was obtained at each combination of reagent and catalyst. The observations were obtained in random order. The sums of squares for reagent, catalyst, and total were 120, 48, and 252, respectively.

   (a) (10 points) Write a complete mathematical model that would be appropriate for this experiment.

   (b) (10 points) Complete the ANOVA table and make appropriate inferences at a 5% level of significance. Be sure to state what hypotheses are being tested.

   (c) (10 points) State what plots would be of interest in this situation. For each plot, state why the plot would be of interest.

3. (25 points) In order to maintain a certain level of proficiency, a bowling league desires to only admit bowlers who score at least 200 in more than 1/3 of the games they bowl. Thus, for an individual wishing to join the league, the goal is to test $H_0 : \pi \leq \frac{1}{3}$ vs. $H_a : \pi > \frac{1}{3}$ where π is the true proportion of games that a bowler has a score of at least 200. In practice, the league allows bowlers to join the league if they score at least 200 in at least 4 out of 6 "tryout" games.

   (a) (10 points) What is the α-level for this test?

   (b) (10 points) If a particular bowler scores at least 200 in 50% of the games they bowl, what is the power of the league's test?

   (c) (5 points) If the number of practice games stays at 6 and H_0 and H_a remain the same, is it possible to reduce the Type I error rate to less than 0.05 while keeping the power of the test constant? Briefly explain.

4. (30 points) Suppose a 2×2 factorial experiment is run in a completely randomized fashion, with one observation per factor combination. Assume an effects model with α_i for factor A effects, β_j for factor B effects, and γ_{ij} for the interaction effects.

   (a) (10 points) Write the design matrix X.

   (b) (10 points) Determine if $\alpha_1 - \alpha_2$ is estimable and justify your answer.

   (c) (10 points) Write an interaction contrast and show why it is an interaction contrast.

5. (25 points) A controversy has arisen in the mathematics department at a large university over the proportion of freshman who had AP statistics in high school. The department chair insists that at least 70% of freshman had AP statistics in high school, but other department members suspect that the proportion is lower. To resolve this issue, the department surveys 55 freshmen, finding that 32 had AP statistics in high school.

   (a) (10 points) Using a significance level of 0.05, test for evidence that the "other department members" are right. Give the p-value.

   (b) (10 points) Find a 95% confidence interval for the proportion of freshmen who had AP statistics in high school.

   (c) (5 points) The department now wishes to do a larger study. If they want the new 95% confidence interval for the proportion of interest to have margin of error no more than 0.04, how large should their sample be?

6. (10 points) A set of data are modeled using the following regression equation:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1$$

   The statistician is considering adding a new variable x_2 to the model.

   (a) (5 points) How does the value of $\hat{\beta}_1$ (estimated coefficient for x_1) change if x_1 and x_2 are orthogonal? Explain your answer.

   (b) (5 points) How does the value of $\hat{\beta}_1$ change if x_1 and x_2 have a correlation coefficient (r) close to 1? Explain your answer.

7. (20 points) In a 1991 journal article, the following question was posed: Are male college students more easily bored than their female counterparts? To answer the question, the authors developed a Boredom Proneness Scale, where higher scores indicate more boredom, and then administered tests to randomly chosen male and female college students. They obtained the following results:

Gender	Sample Size	Sample Mean	Sample SD
Male	97	10.40	4.83
Female	148	9.26	4.68

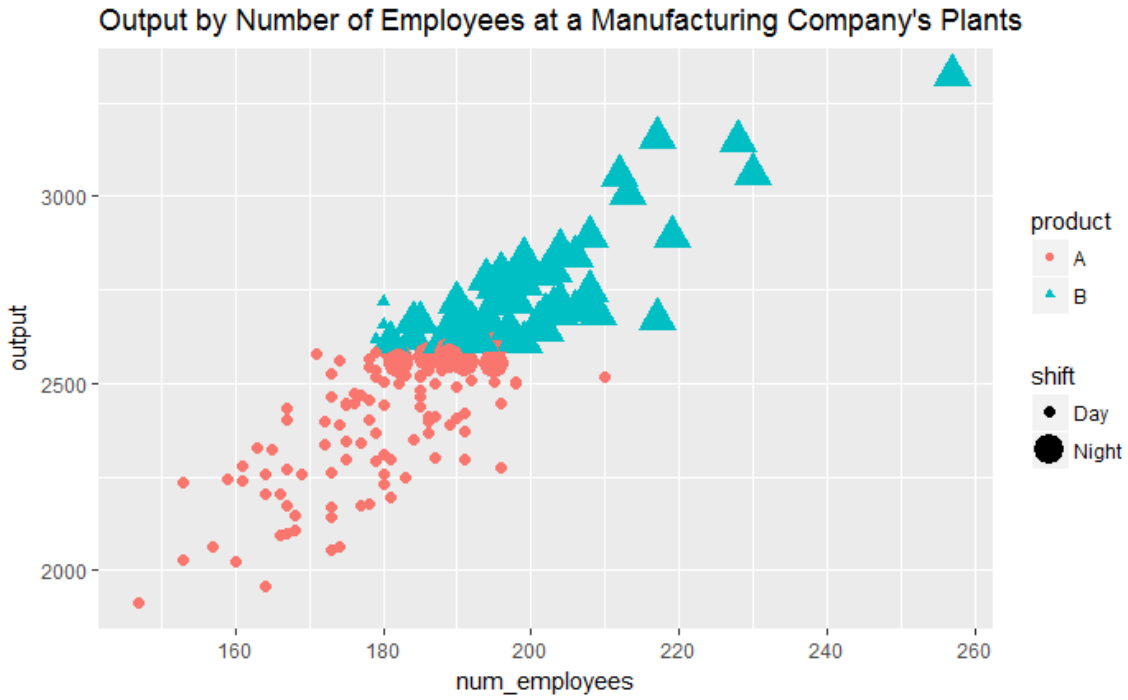
(a) (15 points) Test the hypotheses of interest at the 0.05 level.

(b) (5 points) What assumptions are needed for conducting the test in part a?

8. (40 points) A manufacturing company was interested in determining factors associated with the level of output produced at each plant (*output*). At each plant, they measured the following:

num_employees	The number of employees who worked at the plant
plant_age	The age of the plant (in years)
productB	Product type was either A or B. The variable *productB* is an indicator variable for product B.
shiftNight	Shift was either "day" or "night". The variable *shiftNight* is the indicator variable for the night shift.
productB:shiftNight	This is an interaction term between *productB* and *shiftNight*.

The following is a plot of the data:



A multiple regression model produced the following results:

```
Coefficients:
                  Estimate Std. Error t value  Pr(>|t|)
(Intercept)       760.4685   150.0178   5.069  < 0.0001 ***
num_employees       9.1089     0.7985  11.407  < 0.0001 ***
plant_age          -3.5482     4.4258  -0.802    0.4240
productB          280.8238    68.3550   4.108  < 0.0001 ***
shiftNight        100.4470    39.2607   2.558    0.0115 *
productB:shiftNight -197.6038    79.3426  -2.491    0.0138 *
```

a) (10 points) Write down the model to predict level of output produced (*output*) **for product B only** based on the above table.

b) (10 points) Explain what is being tested by the *productB:shiftNight* interaction term. Conduct the appropriate hypothesis test at the 5% level of significance for this term.

c) (5 points) In a backwards stepwise regression model using the partial-F (or partial P-value) method, what would the next step be, based on the output above? Explain your answer.

d) (15 points) Consider the two plots below. For <u>each</u> one, answer the following questions:

i) What model assumptions are being tested in this plot?

ii) Based on this plot, comment on whether these model assumptions are met.