Name: _____ Fall, 2019

**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. (23 points) A recent study in Yellowstone National Park aimed to compare the proportion of black and brown bear cubs surviving to at least one year in age. Of the 181 black bear cubs studied, 150 survived to a year, while out of the 42 brown bear cubs studied, 32 survived.

   (a) (10 points) Create a 95% confidence interval for the difference in the proportion of each species of bear cub surviving to one year.

   (b) (3 points) For inference for two proportions, we must assume that the sample sizes are "large enough." Briefly explain the rationale behind one common check of this assumption, $n_1\widehat{p}_1 \geq 5$, $n_1(1-\widehat{p}_1) \geq 5$, $n_2\widehat{p}_2 \geq 5$, and $n_2(1-\widehat{p}_2) \geq 5$.

   (c) (10 points) Suppose that the main goal of the study was actually to test at the 0.05 level whether the one-year survival rate for black bear cubs is different than the one-year survival rate for brown bear cubs. Use the confidence interval to draw a conclusion for the test of interest. Make sure to do each of the following:

      i. write the hypotheses of interest symbolically, defining symbols as is necessary.

      ii. briefly explain what the conclusion is and how the confidence interval helped you reach that conclusion.

      iii. interpret the conclusion in terms of the problem.

2. (30 points) Four local amateur golfers participated in a long drive contest where each golfer hit four drives. The drive lengths (in yards) and the sample means and sample standard deviations are given in the table below. You are interested in comparing the mean drive lengths for the four golfers.

| Golfer | Drive Lengths (Yards) | | | | $\bar{x}$ | $s$ |
|---|---|---|---|---|---|---|
| A | 248 | 125 | 193 | 229 | 199 | 54 |
| B | 271 | 251 | 247 | 225 | 249 | 19 |
| C | 163 | 149 | 177 | 189 | 170 | 17 |
| D | 253 | 252 | 284 | 301 | 273 | 24 |

(a) (10 points) Using level 0.05, test for evidence that the mean drive length differs from one golfer to another.

(b) (10 points) Apply Tukey's method at level 0.05 to compare the mean drive lengths for the four golfers. Present your conclusions either in words or with an appropriate diagram.

(c) (10 points) List all the assumptions required for the inference procedures that you applied in parts (a) and (b). For each assumption that you list, provide one way to check the assumption. Please be specific. It is not necessary that you actually do the assumption-checking.

3. (20 points) In the 1980-1981 and 1981-1982 NBA basketball seasons, Larry Bird shot 338 pairs of free throws from the foul line (from *Wardrop 1995, The American Statistician*). The observed distribution of the number of free throws (out of two) that he made each time at the foul line is given by the following table:

| Number of Free Throws Made | 0 | 1 | 2 |
|---|---|---|---|
| Number of Occasions | 5 | 82 | 251 |

   (a) (15 points) If all of the free throws Larry Bird took over his career were independent with the same probability of success, we would expect the number of free throws made out of two to follow a binomial distribution with success probability 0.89 (Larry Bird made 89% of his career free throws). Conduct a 0.05 level goodness-of-fit test to decide whether the binomial distribution with $\pi = 0.89$ is reasonable for these two seasons.

   (b) (5 points) What assumptions are required for the test in part a? Comment on whether you have any concerns about the validity of these assumptions.

4. (25 points) A no-intercept simple linear regression model (or a regression model through the origin), that is $y_i = \beta x_i + \epsilon_i$, $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$, $i = 1, 2, ..., n$, is often appropriate in analyzing data from chemical and other manufacturing processes.

   (a) (10 points) Show that the least-squares estimator for the slope $\beta$ is $\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$.

   (b) (7 points) Derive the mean and variance of $\hat{\beta}$.

   (c) (3 points) Since two points determine a straight line, a student proposes another estimator for $\beta$ that uses only the first observation in the dataset, i.e., $(x_1, y_1)$, and the origin, i.e., $(0, 0)$, as $\tilde{\beta} = \frac{y_1 - 0}{x_1 - 0} = \frac{y_1}{x_1}$. Is $\tilde{\beta}$ an unbiased estimator for $\beta$? Show your work.

   (d) (5 points) Which estimator is better, $\hat{\beta}$ or $\tilde{\beta}$? Explain.

5. (37 points) From a random sample of all domestic flights departing from Newark Liberty International Airport (EWR) and John F. Kennedy International Airport (JFK) in the year 2013, the following summary statistics were observed for the departure delay in (minutes):

| | Mean | Standard Deviation | n |
|---|---|---|---|
| EWR | 18.5 | 41.7 | 178 |
| JFK | 20.3 | 44.6 | 153 |

(a) (10 points) Conduct an appropriate statistical test to determine if there is evidence that the variances in departure delay time from each airport are unequal, at $\alpha = 0.05$. Note: since the provided distribution table does not actually contain the proper degrees of freedom based on the data provided, please use the closest that you can.

(b) (8 points) Since the desired degrees of freedom for the test in part (a) are not in the table, the Type I and Type II Error rates will be different than the ones resulting from the decisions based on the proper degrees of freedom. Answer each of the following:

i. Using the provided distribution table, the Type I Error rate will be:

lower than the same as higher than

that of the decisions made with the proper degrees of freedom (choose one, and explain).

ii. Using the provided distribution table, the Type II Error rate will be:

lower than the same as higher than

that of the decisions made with the proper degrees of freedom (choose one, and explain).

(c) (10 points) Based on your answer to part (a), conduct an appropriate statistical test to determine if there is evidence that the mean departure delays from each airport are different, at $\alpha = 0.05$. If a required degrees of freedom is not exactly available, please use the closest that you can.

(d) (6 points) Based on the information given, do either of the tests performed in parts (a) or (c) require an assumption that departure delays follow a normal distribution? Why or why not?

(e) (3 points) List any other required assumptions for each of the tests in parts (a) and (c), and comment briefly on whether they appear to be reasonable here.

6. (30 points) A student ran an experiment to study how the average time required for her to blow up a balloon varied with the color of the balloon (Red, White, or Blue) and the brand (A, B, or C). She blew up three balloons for each combination of a color and a brand, recording the time required (in seconds). The treatment sample means are given in the table below.

| | Brand | | |
|---|---|---|---|
| Color | A | B | C |
| Red | 22 | 20 | 26 |
| White | 20 | 17 | 23 |
| Blue | 25 | 20 | 18 |

   (a) (10 points) Create appropriate interaction plots, and comment on both the presence/absence of interaction and the type of interaction.

   (b) (10 points) Write down a complete set of orthogonal interaction contrasts.

   (c) (10 points) Suppose that the SSE for an effects model with interaction is 72. Using level 0.05, test for an interaction involving colors White and Blue and brands B and C.

7. (35 points) Researchers from the Netherlands were interested in whether smoking by the mother in pregnancy is related to higher childhood blood pressure in their offspring. In a recent study, they recruited 200 mothers and their newborns (one newborn per mother). They considered the following regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \epsilon,\ \epsilon \overset{iid}{\sim} N(0, \sigma^2),$$

to relate the infant's systolic blood pressure in mm Hg ($y$) to the infant's age in weeks ($x_1$) and weight in kg ($x_2$), and the mother's smoking status in pregnancy (no smoking, passive exposure to smoking, or smoking). Mothers who did not smoke in pregnancy but were exposed to smoke by others were considered as having "passive exposure to smoking". Two indicator variables were defined to account for mother's smoking status:

$$x_3 = \begin{cases} 1, & \text{passive exposure to smoking,} \\ 0, & \text{otherwise;} \end{cases} \qquad x_4 = \begin{cases} 1, & \text{smoking,} \\ 0, & \text{otherwise.} \end{cases}$$

The R output of the analysis for this full model is presented in Table 1. Table 2 presents the ANOVA tables for the full model, the reduced model 1 (only including the independent variables $x_1$ and $x_2$), and the reduced model 2 (only including the independent variables $x_3$ and $x_4$). Figure 1 shows the residual plots for the full model.

(a) (5 points) Interpret the estimated coefficient associated with the variable age ($x_1$) in the full model.

(b) (5 points) Calculate and interpret the coefficient of determination, $R^2$, for the full model.

(c) (10 points) Use appropriate sums of squares or mean squares in Table 2 to test whether the infant's systolic blood pressure is associated with the mother's smoking status given that the age and weight of the infant are in the model at the significance level of 0.05.

(d) (9 points) List all model assumptions. Based on the residual plots for the full model in Figure 1, comment on whether the model assumptions appear to be satisfied. Provide a way to address each assumption violation if there are any.

(e) (6 points) Later on, one researcher suspected that the association between the infant's systolic blood pressure and the age might be different across different smoking statuses. Propose a model that allows the slope of age to differ among the three smoking statuses, and write down the null and alternative hypotheses that would be used to test this researcher's idea.

Table 1: R output of regression analysis for the full model.

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4, data = sdata)

Residuals:
   Min       1Q     Median     3Q      Max
-13.0408  -2.6090  -0.2574   1.6563  20.9125

Coefficients:
               Estimate   Std.Error  t value    Pr(>|t|)
(Intercept)     72.0164     2.5112    28.678     < 2e-16 ***
   x1            1.0006     0.2393     4.181    4.38e-05 ***
   x2            4.0882     0.5007     8.166    3.92e-14 ***
   x3            0.2803     0.9379     0.299       0.765
   x4            5.6025     0.8522     6.574    4.36e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 4.286 on 195 degrees of freedom
F-statistic: 30.08 on 4 and 195 DF,  p-value: < 2.2e-16
```

Table 2: ANOVA tables for the full model (including $x_1$, $x_2$, $x_3$, and $x_4$), the reduced model 1 (including $x_1$ and $x_2$), and the reduced model 2 (including $x_3$ and $x_4$).

| Models | | Degrees of Freedom | Sum of Squares | Mean Squares |
|---|---|---|---|---|
| Full Model | Regression | 4 | 2210.1 | 552.53 |
| | Error | 195 | 3581.4 | 18.37 |
| Reduced Model 1 | Regression | 2 | 1403.5 | 701.75 |
| | Error | 197 | 4388.0 | 22.27 |
| Reduced Model 2 | Regression | 2 | 699.3 | 349.65 |
| | Error | 197 | 5092.2 | 25.85 |

Figure 1: Residual plots for the full model.