STAT 510 **Final Exam** **Spring 2019**

Instructions: This is a closed-notes, closed-book exam. No calculator or electronic device of any kind may be used. Use nothing but a pen or pencil. Please write your name and answers on blank answer sheets. Please do NOT write your answers on the pages with the questions. For questions that require extensive numerical calculations that you should not be expected to do without a calculator, simply set up the calculation and leave it at that. For example, $(3.45 - 1.67)/\sqrt{2.34}$ would be an acceptable answer. On the other hand, some quantities that are very difficult to compute one way may be relatively easy to compute another way. Part of this exam tests your ability to discover the easiest way to compute things based on the information provided and the relationships between various quantities. If you find yourself trying to do exceedingly complex or tedious calculations, there is probably a better way to solve the problem.

Provided below are 0.975 quantiles from t distributions with degrees of freedom 20 through 40. One or more or these quantiles may be needed for answering some questions on this exam.

```
> rbind(20:40, round(qt(0.975, 20:40), 3))

 20.000 21.00 22.000 23.000 24.000 25.00 26.000 27.000 28.000 29.000 30.000 31.00
  2.086  2.08  2.074  2.069  2.064  2.06  2.056  2.052  2.048  2.045  2.042  2.04

 32.000 33.000 34.000 35.00 36.000 37.000 38.000 39.000 40.000
  2.037  2.035  2.032  2.03  2.028  2.026  2.024  2.023  2.021
```

1. Suppose $y_1, y_2, \ldots, y_{100}$ are non-negative integers with average $\bar{y}_. = 62$ and sample variance $s^2 = \sum_{i=1}^{100}(y_i - \bar{y}_.)^2/(100-1) = 93$. Do these non-negative integers seem like they might be independent and identically distributed observations from some Poisson distribution? Compute Pearson's Chi-Square Statistic and explain how to use its value to answer this question.

2. A total of 50 Iowa corn fields, each containing a waterway passing through the field, were used in a study. Within each field, a strip of land (30 meters wide and 100 meters long) immediately adjacent to the waterway was converted to prairie. A strip of land the same size as the prairie strip but on the opposite side of the waterway was kept in corn production and treated as a control strip. A coin was flipped to determine which side of the waterway would contain the prairie strip and which side would contain the control strip. Within each strip (whether prairie or control), four butterfly traps were positioned along the length of the strip at 20, 40, 60, and 80 meters from the end of the strip closest to an access road. Following a capture period, the number of butterflies in each trap was recorded. The researchers are primarily interested in learning whether there tend to be more butterflies in control strips or in prairie strips.

   Let $i = 1, \ldots, 50$ index fields. Let $j = 1, 2$ index strip types, where $j = 1$ corresponds to control and $j = 2$ corresponds to prairie. Let $k = 1, 2, 3, 4$ index traps, with $k = k^*$ corresponding to the trap set at $20k^*$ meters from the end of the strip nearest the access road ($k^* = 1, 2, 3, 4$). Let y_{ijk} be the number of butterflies for field i, strip type j, and trap k. Provide a thorough description of the model you would propose to fit to the y_{ijk} data. Use the same level of detail that was provided in our course notes whenever models for data were described. There is not one right answer to this question, but some answers are better than others. The most appropriate model would depend on the actual data, which are not available to you. Define the model you think is most appropriate based on the information available to you.

3. The Enviratron facility at Iowa State University can be used to study effects of environmental conditions on plant growth. The facility consists of eight growth chambers and a robotic rover that is programmed to visit each chamber at specified times to collect plant measurements. An experiment was conducted to study the effects of temperature, humidity, and carbon dioxide (CO_2) level on plant growth. Ten pots were placed in each growth chamber. Two genotypes were randomly assigned to the pots, with one plant per pot and five plants of each genotype in each growth chamber. Eight combinations of temperature (low vs. high), humidity (low vs. high), and CO_2 level (low vs. high) were assigned to the eight growth chambers. Total leaf area was measured for each plant when plants were one week old and again when the same plants were two weeks old. This process was carried out in January with one set of 80 plants to produce 160 measurements (8 chambers $\times$ 10 plants per chamber $\times$ 2 measurements per plant). The process was repeated again in February, March, and April, each time with a new set of 80 plants. The entire experiment yielded 640 measurements of total leaf area from 320 plants. Let $y_{mthcgap}$ be the total leaf area for month m ($m = 1$ for January, $m = 2$ for February, $m = 3$ for March, $m = 4$ for April), temperature t ($t = 1$ for low, $t = 2$ for high), humidity h ($h = 1$ for low, $h = 2$ for high), CO_2 level c ($c = 1$ for low, $c = 2$ for high), genotype g ($g = 1, 2$), age a ($a = 1$ for one week old and $a = 2$ for two weeks old), and plant p ($p = 1, \ldots, 5$). Suppose

$$y_{mthcgap} = \mu_{thcga} + u_m + v_{mthc} + w_{mthcgp} + e_{mthcgap},$$

$$u_m \sim N(0, \sigma_u^2), \;\; v_{mthc} \sim N(0, \sigma_v^2), \;\; w_{mthcgp} \sim N(0, \sigma_w^2), \;\; e_{mthcgap} \sim N(0, \sigma_e^2),$$

where μ_{thcga} is an unknown mean parameter, σ_u^2, σ_v^2, σ_w^2, and σ_e^2 are unknown, positive variance components, and all random effects and random errors are assumed to be independent. Fitting this assumed model to the 640 total leaf area measurements produced REML estimates

$$\hat{\sigma}_u^2 = 0.3, \quad \hat{\sigma}_v^2 = 0.1, \quad \hat{\sigma}_w^2 = 0.4, \;\text{ and }\; \hat{\sigma}_e^2 = 0.2.$$

Data averages of the form $\bar{y}_{.thcga.}$ are provided in the central portion of Table 1. Other averages are provided in the last row and column of Table 1.

Table 1. Data averages of the form $\bar{y}_{.thcga.}$ (center), $\bar{y}_{.thc...}$ (last column), and $\bar{y}_{....ga.}$ (last row).

t	h	c	$g=1, a=1$	$g=1, a=2$	$g=2, a=1$	$g=2, a=2$	$\bar{y}_{.thc...}$
1	1	1	4.3	12.5	3.5	14.1	8.6
1	1	2	5.2	13.5	4.8	14.9	9.6
1	2	1	2.9	11.0	3.2	12.9	7.5
1	2	2	3.5	12.7	4.3	13.9	8.6
2	1	1	5.9	14.1	6.7	16.5	10.8
2	1	2	7.3	15.6	6.6	16.9	11.6
2	2	1	8.2	16.2	9.4	19.4	13.3
2	2	2	9.1	18.0	9.5	19.4	14.0
		$\bar{y}_{....ga.}$	5.8	14.2	6.0	16.0	

(a) Compute a t-statistic for testing the null hypothesis of no three-way interaction between temperature, humidity, and CO_2 level.

(b) State the degrees of freedom associated with the t statistic in part (a).

(c) Determine the degrees of freedom for the F-test used to test for genotype main effects.

(d) Compute the standard error for the BLUE of $\mu_{11111} - \mu_{11112}$.

(e) Find the degrees of freedom associated with the standard error in part (d).

4. Consider another Enviratron experiment intended to identify an ideal temperature for maximum plant growth. Temperatures 22, 24, 26, and 28 degrees Celsius were randomly assigned to the eight Enviratron growth chambers, with two chambers per temperature. Ten pots, each containing one plant, were placed in each growth chamber. After four weeks, the total leaf area for each plant was measured, and the average total leaf area for the ten plants in each chamber was recorded. This entire process was carried out five times (once each month for five months) to obtain a total of 40 data points, where each data point is the average total leaf area of ten plants after four weeks in an Enviratron growth chamber set at a specific temperature. Let y_{mtc} be the average total leaf area of the ten plants for month m ($m = 1, \ldots, 5$), temperature t ($t = 1$ for 22, $t = 2$ for 24, $t = 3$ for 26, $t = 4$ for 28), and chamber c ($c = 1, 2$).

   Use the R code and output on page 4 to answer the following questions.

   (a) Consider two models:

$$\text{Model 1: } y_{mtc} = \alpha_{0m} + \alpha_1 x_t + \alpha_2 x_t^2 + \varepsilon_{mtc}, \quad \varepsilon_{mtc} \overset{iid}{\sim} N(0, \sigma_\varepsilon^2),$$

$$\text{Model 2: } y_{mtc} = \mu + \beta_m + \tau_t + \epsilon_{mtc}, \quad \epsilon_{mtc} \overset{iid}{\sim} N(0, \sigma_\epsilon^2),$$

   where $\alpha_{01}, \ldots, \alpha_{05}$, α_1, α_2, μ, $\beta_1, \ldots, \beta_5$, and $\tau_1, \ldots, \tau_4$ are unknown parameters in $\mathbb{R}$,
   $x_1 = 22, x_2 = 24, x_3 = 26, x_4 = 28$,
   and σ_ε^2 and σ_ϵ^2 are unknown, positive variance parameters. Does Model 1 fit the data adequately compared to Model 2? Compute one F-statistic that could be used to answer this question.

   (b) Under the assumptions of Model 2, compute an F statistic for testing

$$H_0 : \tau_1 = \tau_2 = \tau_3 = \tau_4.$$

   (c) Now assume

$$\text{Model 3: } y_{mtc} = \gamma_0 + \gamma_1 x_t + \gamma_2 x_t^2 + e_{mtc}, \quad e_{mtc} \overset{iid}{\sim} N(0, \sigma_e^2),$$

   where $\gamma_0, \gamma_1, \gamma_2$ are unknown parameters in $\mathbb{R}$ and σ_e^2 is an unknown, positive variance parameter. Compute a 95% confidence interval for γ_1.

   (d) Assuming Model 3 holds, estimate the temperature at which the expected value of total leaf area is largest.

```
> #y is average of total leaf area for 10 plants
> d
   month temp    y
1      1   22 48.3
2      1   22 47.0
3      1   24 48.0
4      1   24 49.0
5      1   26 48.3
6      1   26 48.2
7      1   28 47.5
8      1   28 47.7
9      2   22 48.0
10     2   22 47.6
.
.<28 LINES DELETED>
.
39     5   28 47.0
40     5   28 47.8
> is.factor(d$month)
> [1] TRUE
> is.numeric(d$temp)
> [1] TRUE
> o = lm(y ~ month + temp + I(temp^2) + I(temp^3), data = d)
> anova(o)
Analysis of Variance Table

Response: y
            Sum Sq
month          1.1
temp           0.2
I(temp^2)      7.9
I(temp^3)      0.1
Residuals     10.2
>
> oQuad = lm(y ~ temp + I(temp^2), data = d)
> coef(oQuad)
(Intercept)        temp   I(temp^2)
      -20.3         5.5       -0.11
> vcov(oQuad)
            (Intercept)     temp   I(temp^2)
(Intercept)     185.982  -14.959       0.298
temp            -14.959    1.200      -0.020
I(temp^2)         0.298   -0.020       0.001
```