STAT 510 **Final Exam** **Spring 2014**

Instructions: The 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 paper. Please do not write your answers on the pages with the questions. For questions that require extensive numerical calculations that cannot be done easily 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.

1. Consider an experiment with two factors: A (with levels A_1 and A_2) and B (with levels B_1, B_2, and B_3). Suppose an unbalanced, completely randomized design was used to assign experimental units to treatments. For $i = 1, 2$ and $j = 1, 2, 3$, let n_{ij} be the number of experimental units treated with level A_i of factor A and level B_j of factor B. Let y_{ijk} denote the response of the kth experimental unit treated with level A_i of factor A and level B_j of factor B for $i = 1, 2$, $j = 1, 2, 3$, and $k = 1, \ldots, n_{ij}$. Assume that

$$y_{ijk} = \mu_{ij} + e_{ijk}, \tag{1}$$

where each μ_{ij} term is an unknown mean parameter and all e_{ijk} terms are independent and identically distributed as $N(0, \sigma^2)$ for some unknown variance parameter σ^2. The following tables contain the sample averages, sample variances, and sample sizes for each treatment group. (Whole numbers are presented to make calculation easier.)

Sample Averages ($\bar{y}_{ij\cdot}$ values)

	B_1	B_2	B_3
A_1	11	7	2
A_2	10	9	5

Sample Variances $\left(\frac{1}{n_{ij}-1}\sum_{k=1}^{n_{ij}}(y_{ijk} - \bar{y}_{ij\cdot})^2 \text{ values}\right)$

	B_1	B_2	B_3
A_1	3	1	2
A_2	2	3	1

Sample Sizes (n_{ij} values)

	B_1	B_2	B_3
A_1	5	3	2
A_2	3	5	5

(a) Find an LSMEAN for each level of factor A.

(b) Provide the numerical value of the estimator you would use to estimate σ^2.

(c) Find a 95% confidence interval for the simple effect of factor A when factor B is fixed at level B_3. (Use the notation $t_{df,q}$ to represent the qth quantile of a t distribution with df degrees of freedom. Specify both df and q in your answer.)

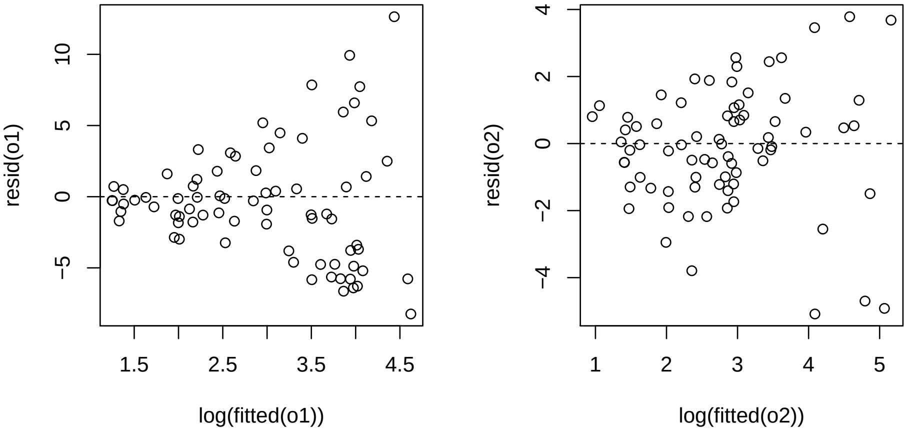
(d) Compute the test statistic you would use to determine if there is a significant factor A main effect.

(e) Now suppose each experimental unit is a pot of soil containing a seedling. Factor A is a soil chemical treatment factor with A_1 representing one chemical and A_2 representing another. Factor B is the amount of chemical applied to the soil with B_1, B_2, and B_3 representing 0, 10, and 20 units of chemical, respectively. The response variable is the weight of the seedling 28 days after emergence from the soil. Based on this new information, would you recommend any changes to the model provided in equation (1) above? Explain why or why not.

2. A total of 75 intersections were selected for use in an experiment to compare the safety of three different traffic control systems (labeled 1, 2, and 3). A completely randomized design was used to assign 25 intersections to each of the systems. The amount of traffic passing through each intersection during the study period was recorded in an approximately continuous quantitative variable x. Low values of x indicate relatively little traffic while high values of x indicate a high traffic volume. The number of traffic accidents at each intersection during the study period was recorded in a non-negative-integer-valued response variable y. Some potentially relevant `R` code and output is as follows:

```
> #s=factor containing the traffic system information.
> #s=1 <=> system 1, s=2 <=> system 2, s=3 <=> system 3
> summary(s)
 1  2  3
25 25 25
>
> #x=amount of traffic
> summary(x)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.10   34.05   55.90   54.89   78.55   98.50
>
> #y=number of crashes
> summary(y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    1.0     6.0    16.0    28.8    31.0   225.0
>
> o1=glm(y˜x,family=poisson(link="log"))
> summary(o1)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.2392  -3.1111  -0.7144   1.3178  12.6448

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.865029   0.088445    9.78   <2e-16 ***
x           0.038152   0.001159   32.91   <2e-16 ***

    Null deviance: 2620.8  on 74  degrees of freedom
Residual deviance: 1204.1  on 73  degrees of freedom
AIC: 1548.9
```

```
> o2=glm(y~x+s+x:s,family=poisson(link="log"))
> summary(o2)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.0841  -1.2134  -0.0385   0.8344   3.7845

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.438769   0.260385   1.685 0.091973 .
x           0.030749   0.003367   9.132  < 2e-16 ***
s2          0.643431   0.296200   2.172 0.029834 *
s3          0.595104   0.288490   2.063 0.039130 *
x:s2        0.001201   0.003973   0.302 0.762443
x:s3        0.013336   0.003719   3.586 0.000336 ***

    Null deviance: 2620.83  on 74  degrees of freedom
Residual deviance:  239.27  on 69  degrees of freedom
AIC: 592.03
```



(a) Provide the value of the test statistic you would use to test the null hypothesis that says, “For any given amount of traffic x, the mean number of crashes is the same for all three traffic control systems.”

(b) State the approximate null distribution of the test statistic provided in part (a).

(c) Provide an estimate of the mean number of accidents as a function of the amount of traffic x for intersections using traffic control system 3.

3. An experiment was conducted to study the impact of two testing methods (online vs. on paper) and two teaching styles (traditional lecture vs. active learning) on test scores of introductory statistics students. Within each of ten universities, two instructors participated in the study. In five universities randomly selected from the ten universities, both instructors taught using a traditional lecture style. In the other five universities, both instructors taught using an active learning style. Each of the 20 instructors taught two sections of 30 students each. For each instructor, students in one section (randomly selected from the two sections) took a comprehensive final exam online while students in the other section took the same final on paper. Suppose a normally distributed comprehensive final exam score is available for each of the 1200 students who participated in the study.

   (a) Does this experiment involve blocking? If so, describe the blocks.

   (b) What are the experimental units in this experiment?

   (c) Provide the *Source* and *Degrees of Freedom* columns of the ANOVA table that corresponds to the model you would fit to the exam score data.

   (d) Name the term in the *Source* column whose mean square would be the denominator of the F-statistic for testing for the main effect of testing method.

   (e) Name the term in the *Source* column whose mean square would be the denominator of the F-statistic for testing for the main effect of teaching style.

4. Scientists were interested in studying the immune response of pigs to a bacterial infection. A total of 16 pigs were randomly assigned to either the bacterial infection or a mock infection. The design was balanced, so 8 pigs received each infection type (bacterial or mock). Blood samples were taken from each pig immediately before infection (time 0) and at 1, 2, and 3 hours after infection (times 1, 2, and 3, respectively). The concentration of a particular protein involved in immune response was measured in each blood sample. Let y_{ijk} be the protein concentration in the blood sample collected at time k from the jth pig that received treatment i, where $k \in \{0, 1, 2, 3\}$, $j \in \{1, \dots, 8\}$, and $i = 1$ for bacterial infection and $i = 2$ for mock infection. For all i, j, let $\boldsymbol{y}_{ij} = [y_{ij0}, y_{ij1}, y_{ij2}, y_{ij3}]'$. We can define a general linear model for these data as follows:

   > Suppose all $\boldsymbol{y}_{ij}$ vectors are independent of each other and that $\boldsymbol{y}_{ij} \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}_1 = [\mu_{10}, \mu_{11}, \mu_{12}, \mu_{13}]'$ and $\boldsymbol{\mu}_2 = [\mu_{20}, \mu_{21}, \mu_{22}, \mu_{23}]'$ are unknown vectors in $\mathbb{R}^4$ and $\boldsymbol{\Sigma}$ is an unknown variance matrix.

   Four special cases of the general linear model were fit to the data. A description of the four models and the value of the maximized log likelihood for each model are provided in the following table.

Model	Mean	Variance	Maximized Log Likelihood
A	$\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^4$	$\boldsymbol{\Sigma}$ has compound symmetric structure	-189.9
B	$\boldsymbol{\mu}_1, \boldsymbol{\mu}_2 \in \mathbb{R}^4$	$\boldsymbol{\Sigma}$ has AR(1) structure	-192.4
C	$\mu_{ik} = \beta_0 + k\beta_i \; \forall \, i, k$	$\boldsymbol{\Sigma}$ has compound symmetric structure	-191.7
D	$\mu_{ik} = \beta_0 + k\beta_i \; \forall \, i, k$	$\boldsymbol{\Sigma}$ has AR(1) structure	-194.5

   Note that in models C and D, β_0, β_1, and β_2 are each unknown parameters in $\mathbb{R}$.

   (a) For each pair of models that can be compared with a likelihood ratio test, write down which model is the null model, which model is the alternative model, the value of the likelihood ratio statistic, and the degrees of freedom associated with the likelihood ratio statistic. In other words, create a table with the following headings and one row for each possible likelihood ratio test.

Null Model	Alternative Model	Likelihood Ratio Test Statistic	Degrees of Freedom

(b) Based on the information provided, which of the four models do you think is most appropriate for the data? Briefly explain your choice.

(c) Suppose Model B was fit to the data using the R code

```
o=gls(y~inf*tm, correlation = corAR1(form=~1|pig),method="ML")
```

where y is the response vector, inf is a factor with levels 1 and 2 for bacterial infection and mock infection, respectively, tm is a factor with levels 0, 1, 2, and 3 corresponding to the four times, and pig is a factor with one level for each of the 16 pigs. Use the following code and output to find the MLEs of $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$.

```
> coef(o)
(Intercept)         inf2          tm1          tm2          tm3
         16           -4            8           14           25

   inf2:tm1     inf2:tm2     inf2:tm3
         -8          -12          -19
```

5. Suppose researchers were interested in testing for the presence of a certain type of bacteria on Iowa hog farms. On each of 50 randomly selected hog farms in Iowa, 20 fecal samples (one from each of 20 randomly selected hogs on the farm) were tested for presence of the bacteria. Let y_i denote the proportion of positive samples on the ith farm. Let π_i denote the probability that a fecal sample from a randomly selected hog on the ith farm will test positive for the bacteria. For all $i = 1, \ldots, 50$, suppose

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \theta + e_i, \text{ where } e_i \sim N(0, \sigma^2),$$

and

$$y_i|\pi_i \sim \text{binomial}(20, \pi_i)/20.$$

Furthermore, suppose all the e_i terms are independent and that the conditional distributions $y_i|\pi_i$ are independent across all $i = 1, \ldots, 50$. Suppose the MLEs of θ and σ^2 obtained by fitting this model to the data are $\hat{\theta} = -0.9$ and $\hat{\sigma}^2 = 0.05$, respectively. Suppose the standard error for $\hat{\theta}$ is 0.1.

(a) Provide an estimate of π_i for a farm where $e_i = 0$.

(b) Suppose a new farm is randomly selected from hog farms in Iowa. Let π_{51} denote the probability that a fecal sample from a randomly selected hog on this new farm will test positive for the bacteria. Assume

$$\log\left(\frac{\pi_{51}}{1-\pi_{51}}\right) = \theta + e_{51}, \text{ where } e_{51} \sim N(0, \sigma^2),$$

and e_{51} is independent of $e_1, \ldots, e_{50}$. Find an interval that will contain π_{51} with approximate coverage probability 0.95.