STAT 510 **Final Exam** **Spring 2017**

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 paper. 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 figure out 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.

1. Suppose μ_1 and μ_2 are real-valued parameters, and suppose σ_u^2 and σ_e^2 are positive variance components. Suppose $u_1, u_2, u_3 \overset{iid}{\sim} N(0, \sigma_u^2)$ independent of $e_1, e_2, e_3, e_4 \overset{iid}{\sim} N(0, \sigma_e^2)$. Suppose

$$\begin{aligned} y_1 &= \mu_1 + u_1 + e_1, \\ y_2 &= \mu_1 + u_1 + e_2, \\ y_3 &= \mu_1 + u_2 + e_3, \text{ and} \\ y_4 &= \mu_2 + u_3 + e_4. \end{aligned}$$

   (a) Suppose $\sigma_u^2 = 3$ and $\sigma_e^2 = 2$. Find values a_1, a_2, a_3, and a_4 so that $a_1y_1 + a_2y_2 + a_3y_3 + a_4y_4$ is the BLUE of $\mu_1 - \mu_2$.

   (b) Now suppose σ_u^2 and σ_e^2 are unknown. Provide a set of error contrasts that could be used to find the REML estimators of σ_u^2 and σ_e^2. (You do NOT need to find REML estimators; just give error contrasts.)

2. A total of 20 patients suffering from foot pain associated with the disease diabetes participated in an experiment to evaluate two oral medications (OM_1 and OM_2) and two foot treatments (FT_1 and FT_2) intended to relieve foot pain. The two oral medications were randomly assigned to patients using a balanced, completely randomized design, with 10 patients per oral medication. Within each oral medication treatment group, 5 patients were randomly selected to receive foot treatment FT_1 on their left foot and foot treatment FT_2 on their right foot. The other 5 patients in each oral medication treatment group received FT_2 on their left foot and foot treatment FT_1 on their right foot. After six weeks on their assigned oral medication and foot treatment regime, the pain experienced by each patient was measured separately for left and right feet to obtain a dataset with a total of 40 pain measurements. Consider the following `R` code and output.

```
> #y = vector of pain measurements
> #OM = factor specifying oral medication (levels 1 and 2)
> #FT = factor specifying foot treatment  (levels 1 and 2)
> #foot = factor specifying foot (levels L and R for left foot
> #                               and right foot, respectively)
> #subject = factor specifying the subject (levels 1 to 20)
> #d = data.frame containing the dataset
>
```

```
> head(d)
  OM subject foot FT   y
1  1       1    L  1 7.8
2  1       1    R  2 4.2
3  1       2    L  1 5.7
4  1       2    R  2 3.1
5  1       3    L  1 6.1
6  1       3    R  2 2.2
> library(nlme)
> o = lme(y ~ foot * OM * FT, random = ~ 1 | subject, data = d)
> o
Linear mixed-effects model fit by REML
  Data: d
  Log-restricted-likelihood: -44.95243
  Fixed: y ~ foot * OM * FT
  (Intercept)         footR           OM2           FT2
         6.44         -3.08         -1.94          1.22

  footR:OM2     footR:FT2       OM2:FT2
       0.52         -1.32          0.66

  footR:OM2:FT2
          -1.26

Random effects:
 Formula: ~1 | subject
        (Intercept)  Residual
StdDev:     1.13231 0.3941764

Number of Observations: 40
Number of Groups: 20
```

For the model fit to the data in the R code, the best linear unbiased estimate of any cell or marginal mean is equal to the corresponding response data average.

(a) What value of BIC would R report for the lme output contained in the object o.

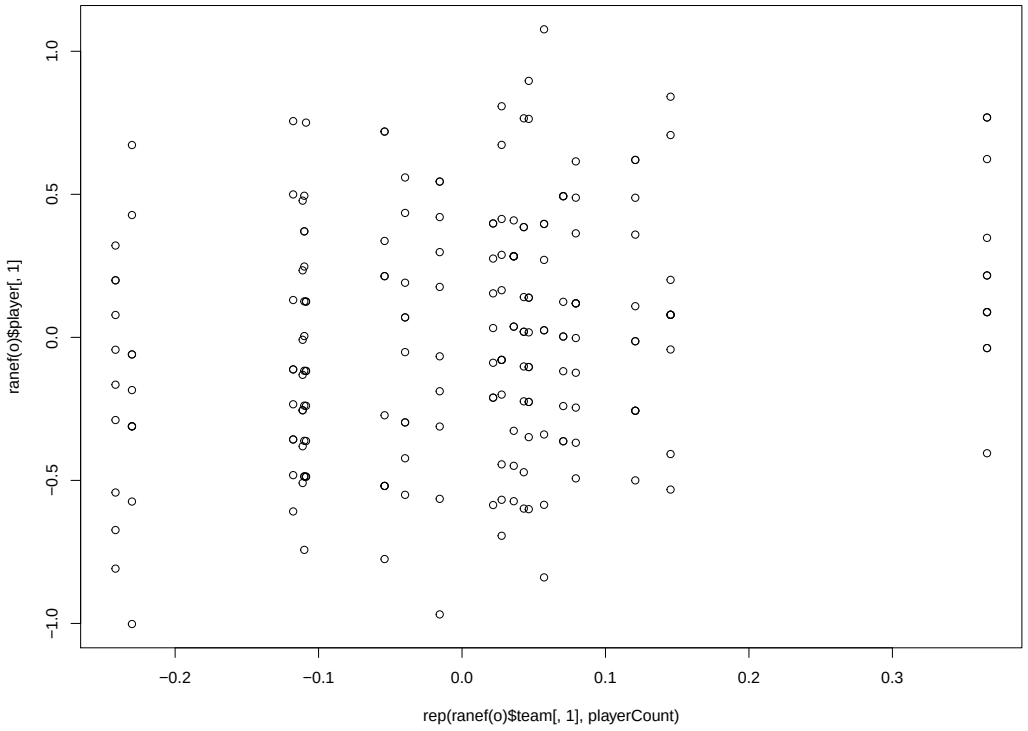
(b) Give the standard error for the estimate 0.66 that is labeled OM2:FT2 in the R output.

3. A free throw is an unguarded attempt to throw a basketball through a hoop from behind a line 15 feet from the backboard that supports the hoop. A free throw attempt is successful if the ball passes through the hoop and unsuccessful otherwise. An experiment was conducted to compare the effectiveness of two methods for training 5th grade boys to shoot (i.e., attempt) free throws. A total of 20 teams were involved in the experiment. Each team consisted of a coach and 8 to 12 5th grade boys. Coaches for 10 of the teams, randomly selected from the 20, agreed to teach their players to shoot free throws using method 1. Coaches for the other 10 teams agreed to teach their players to shoot free throws using method 2. At the end of the basketball season, each 5th grade boy attempted 20 free throws, and the number of successful attempts was recorded.

Some R code and output related to the analysis of the free throw data are provided below. When answering parts (a) and (b) after the R code and output, assume the model fit to the data with the R code is an appropriate model for the data.

```
> #y is a vector.  Element i of the vector is the number of successful
> #free throw attempts out of 20 attempts for player i.
> length(y)
[1] 204
> summary(y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    7.00   10.00   10.09   13.00   19.00
> #team is a factor with 20 levels, one for each team.
> length(team)
[1] 204
> table(team)
team
 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
12 12 10 11 12  9 10 10 10  9 11 10  8 10 11 10 10  9 10 10
> playerCount = as.vector(table(team))
> playerCount
 [1] 12 12 10 11 12  9 10 10 10  9 11 10  8 10 11 10 10  9 10 10
> #player is a factor with one level for each 5th grade boy.
> player = factor(1:204)
> #method is factor with two levels.
> table(method, team)
      team
method  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
     1 12 12 10 11 12  9 10 10 10  9  0  0  0  0  0  0  0  0  0  0
     2  0  0  0  0  0  0  0  0  0  0 11 10  8 10 11 10 10  9 10 10
> library(lme4)
>
> o = glmer(cbind(y, 20 - y) ~ method + (1 | team) + (1 | player),
+           family = binomial(link = "logit"))
>
> summary(o)
Generalized linear mixed model fit by maximum likelihood
 Family: binomial  ( logit )
Formula: cbind(y, 20 - y) ~ method + (1 | team) + (1 | player)

     AIC      BIC   logLik deviance df.resid
  1091.8   1105.1   -541.9   1083.8      200

Scaled residuals:
     Min       1Q   Median       3Q      Max
-1.77970 -0.46539  0.00035  0.42116  1.87455
```

```
Random effects:
 Groups Name        Variance Std.Dev.
 player (Intercept) 0.3076   0.5546
 team   (Intercept) 0.0413   0.2032
Number of obs: 204, groups:  player, 204; team, 20

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.1026     0.0958   1.070    0.284
method2      -0.1779     0.1364  -1.304    0.192

Correlation of Fixed Effects:
        (Intr)
method2 -0.702
> plot(rep(ranef(o)$team[, 1], playerCount), ranef(o)$player[, 1])
```




```
> ranef(o)$team[, 1]
 [1] -0.10994442  0.02760587 -0.24155857  0.12077402  0.36596393 -0.10875557
 [7] -0.05398734  0.03605508 -0.11770779  0.05718532  0.07931555 -0.23008532
[13] -0.11110321  0.07054687  0.04651782  0.14536446  0.04306034  0.02167510
[19] -0.03969921 -0.01555833
> summary(ranef(o)$player[, 1])
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-1.0020000 -0.3111000  0.0002420 -0.0001582  0.2833000  1.0770000
```

(a) Is one of the methods for training 5th grade boys to shoot free throws better than the other? Conduct one test to address this question. Provide a test statistic, a p-value, and a brief conclusion.

(b) Predict the success probability for the worst free throw shooter among all the 5th grade boys who participated in this experiment.

4. Researchers were interested in determining which of two sets of written instructions were best for informing people how to correctly complete a task. The task involves examining an image on a computer screen and using a computer mouse to define boxes around certain features that appear in the image. The researchers have many thousands of images that must be processed in this way, and they need to employ many people to do the work. Because the task is nontrivial, a good set of written instructions is very important for obtaining accurate processing of all images.

   From a large collection of potential employees, the researchers selected 8 people for participation in a pilot experiment to compare the two instruction sets. Instruction set 1 was given to 4 participants randomly selected from the 8, and instruction set 2 was given to the other 4 participants. The researchers selected 8 images randomly from their collection of images and presented a different subset of 2 of these 8 images to each of the 8 participants in the pilot experiment. Each participant was asked to process their assigned images. Numerical scores, representing the quality of image processing, are provided in the following `R data.frame d`.

```
> d
   InstructionSet Participant Image Score
1               1           1     1  43.4
2               1           1     2  64.6
3               1           2     2  57.1
4               1           2     3  48.3
5               1           3     3  45.0
6               1           3     4  18.6
7               1           4     4  34.7
8               1           4     5  44.7
9               2           5     5  48.3
10              2           5     6  43.2
11              2           6     6  45.4
12              2           6     7  67.8
13              2           7     7  73.9
14              2           7     8  79.9
15              2           8     8  73.1
16              2           8     1  46.3
```

   These data can be modeled using a linear mixed-effects model of the form $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}$, where $\boldsymbol{y}$ is the vector `d$Score` in the `R data.frame d`.

   (a) Consider the specific linear mixed-effects model you would fit to these data and provide the matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ that correspond to your model.

   (b) State the distributional assumptions you would make for $\boldsymbol{u}$ and $\boldsymbol{e}$ in your linear mixed-effects model.

5. An experiment was conducted to study the immune response of pigs to injection with a pathogen. A total of 60 pigs were used in the experiment. A set of 30 pigs, randomly selected from the 60, were injected with the pathogen. The other 30 pigs were injected with a harmless saline solution. The 30 pigs that received the pathogen injection were randomly assigned to 5 sampling times (1, 2, 3, 4, and 5 hours after injection) with 6 pigs per sampling time. Likewise, the 30 pigs that received the saline injection were randomly assigned to the same 5 sampling times (1, 2, 3, 4, and 5 hours after injection) with 6 pigs per sampling time. A blood sample was drawn from each pig at the assigned sampling time. Let y_{ijk} be the measurement of a response variable of interest for the blood sample of the kth pig sampled at time j in injection group i, where $i = 1$ for pathogen injection and $i = 2$ for saline injection, $j = 1, \ldots, 5$, and $k = 1, \ldots, 6$. The researchers are considering two models for the data:

$$\textbf{Model 1} \quad y_{ijk} = \mu + \alpha_i + \beta_i x_j + \gamma x_j^2 + \epsilon_{ijk},$$

where $x_j = j$ for $j = 1, \ldots, 5$ and μ, α_1, α_2, β_1, β_2, and γ are unknown parameters, and

$$\textbf{Model 2} \quad y_{ijk} = \mu_{ij} + \varepsilon_{ijk},$$

where μ_{ij} $(i = 1, 2; j = 1, \ldots, 5)$ are unknown parameters. In both models, the error terms are assumed to be independent and identically distributed normal random variables with mean 0. In the `R` code below `y` is the response vector, `Injection` is a factor with two levels (1=pathogene injection, 2=saline injection), and `x` is a quantitative vector that contains the time (1, 2, 3, 4, or 5) for each observation.

```
> o = lm(y ~ Injection + x +I(x^2) + I(x^3) + I(x^4) +
+        Injection:x +Injection:I(x^2) + Injection:I(x^3) + Injection:I(x^4),
+        data = d)
>
> anova(o)
Analysis of Variance Table

Response: y
                  Sum Sq
Injection          756.1
x                   50.7
I(x^2)             100.6
I(x^3)               3.0
I(x^4)              17.4
Injection:x         24.3
Injection:I(x^2)   180.2
Injection:I(x^3)     4.4
Injection:I(x^4)     3.3
Residuals          222.8
```

(a) Using the output provided, construct the test statistic you would use to test whether Model 1 fits adequately relative to Model 2.

(b) Now suppose the 60 blood samples actually come from only 12 pigs, 6 treated with pathogen injection and 6 with saline injection. Suppose each pig provided one blood sample at each of the five time points (1, 2, 3, 4, and 5 hours after injection). Let y_{ijk} be the measurement of a response

variable of interest for the time j blood sample of the kth pig in injection group i, where $i = 1$ for pathogen injection and $i = 2$ for saline injection, $j = 1, \ldots, 5$, and $k = 1, \ldots, 6$. Consider

$$\textbf{Model 3} \qquad y_{ijk} = \mu_{ij} + e_{ijk},$$

where where μ_{ij} $(i = 1, 2; j = 1, \ldots, 5)$ are unknown parameters and the vector of errors

$$\boldsymbol{e} = (e_{111}, e_{121}, \ldots, e_{151}, e_{112}, e_{122}, \ldots, e_{152}, \ldots, e_{216}, e_{226}, \ldots, e_{256})'$$

is multivariate normal with mean $\mathbf{0}$ and has block diagonal variance with one block corresponding to each pig, and each block of the form $\sigma^2 \boldsymbol{W}$, where $\boldsymbol{W}$ has an $AR(1)$ structure with correlation parameter ρ. Give a simplified expression for the variance of the best linear unbiased estimator of $\mu_{11} - \mu_{15}$ in terms of Model 3 parameters.