STAT 505/506 MS Comprehensive Exam–Take Home

Released 9am 23 Aug 2016
Due by NOON 25 Aug 2016

Instructions: Make sure to read each question carefully and ask for clarification as needed. You are allowed to use any resources from Stat 505/506, the internet and resources from any other course. **If you do use non-Stat 505/506 resources they must be referenced, including internet resources, but you may not receive help from other people except me**. Please provide your numbered answers separately (you do not need to include the questions) and email me (`laura.hildreth@montana.edu`) your answers by noon on 25 August 2016. **By turning in this exam, you acknowledge that you have completed this exam in accordance with the Student Conduct Code for Academic Honesty found online at** `http://www.montana.edu/policy/student_conduct/academicmisconduct`. Failure to comply with this code will result in an automatic score of 0, failure of the comprehensive exam, and you will be reported for academic dishonesty.

The questions below are based on the article *Guppies show behavioural but not cognitive sex differences in a novel object recognition test* by Lucon-Xiccato and Dadda (2016). The pdf of this article is found at:
`http://journals.plos.org/plosone/article/asset?id=10.1371%2Fjournal.pone.0156589.PDF`
and the data can be downloaded at:
`http://journals.plos.org/plosone/article?id=info%3Adoi%2F10.1371%2Fjournal.pone.0156589`
Please note that you may need to do some minor cleaning of the data due to how the data were recorded by the authors. For this exam we will look at the response variable `preference for the novel object` (I'd recommend renaming this variable to `preference`).

1. Identify which variables/factors the authors are treating as fixed effects and which are treated as random effects. Explain if these modeling choices are reasonable/appropriate given the study. [10pts]

2. Recall that we can write a linear mixed model using the notation $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\epsilon}$ where $\boldsymbol{u} \sim N(0, \boldsymbol{G})$ and $\boldsymbol{\epsilon} \sim N(0, \boldsymbol{R})$. For the matrices $\boldsymbol{G}$ and $\boldsymbol{R}$,

   (a) Clearly explain what they represent *in context of this analysis*. [5pts]

   (b) Provide/sketch out the hypothesized structure of these matrices that the authors used in their analysis. **Remember to include dimensions!** [10pts]

   (c) Using the matrices you provided in 2b, explain what assumptions the authors are making regarding the guppies and the responses of preference. [6pts]

3. On page 5 the authors note that they transformed the response by using the arcsine square root in order to meet normality assumptions. An alternative to transforming the response is to choose a model that makes alternative distributional assumptions regarding the response variable including the use of logistic and Poisson models. Explain why neither of these two models would be appropriate to use in this analysis (regardless of whether the data were transformed or not). [7pts]

4. Let's turn to the linear mixed model the authors describe on page 5.

(a) Write out the model used by the authors making sure to carefully define all elements of the model. [5pts]

(b) Fit the model proposed by the authors using either SAS or R (no need to do both). No output is necessary for this question as in Problem 7 you will need to provide your SAS or R code and because you will use this model in the next problem. Note: if you are using R, the function `asin` calculates the arcsine for you (hopefully you remember how to do the square root and keep in mind that you square root THEN take the arcsine!). In SAS, you need to add a new variable to your data set so try modifying the following code to do so:

```
data guppies;
set guppies;
asinsqPreference = arsin(sqrt(Preference));
run;
```

Note: You will NOT get the same values as the authors. SAS and R, different versions of R, and different packages in R (such as nlme and lme4) use different algorithms to obtain parameter estimates. Also, R and SAS may compute the df differently and use different sums of squares to calculate the F-statistic. What we are concerned with is if you used the correct model, if you can read/interpret your results correctly, and if you chose a reasonable approach to obtain results similar to those in the paper. [4pts]

(c) Using the appropriate diagnostics, evaluate if the transformation did indeed help meet normality assumptions. In your evaluation, clearly state what diagnostic(s) you are evaluating to assess this assumption and also clearly explain if, based on the diagnostic(s), the assumption is or is not reasonably met. [8pts]

5. On the last paragraph on page 5 the authors report the test statistic $F_{1,24} = 0.051$ associated with a hypothesis test for the effect of sex.

   (a) Explain what the authors are testing with this test statistic and provide the hypotheses being tested. [7pts]

   (b) The authors conclude that "Sex did not significantly influence the preference for the novel object." Explain whether or not you agree with their conclusion. [8pts]

6. Using Figure 2a, the authors examine the relationship between block of time and preference for males and females. To further explore this relationship, one analysis done by the authors was to compare males' preference in the first period (minutes 1 to 6) to females' preference in the second period (minutes 7 to 12). Briefly explain or outline the analysis done by the authors to compare sex effects over time. Do you think that this was a reasonable statistical approach? If so, explain why and if not, explain why not by providing a more appropriate statistical approach (you do NOT need to conduct this analysis–just describe what you would do). [15pts]

7. Provide a *Scope of Inference* for this problem. That is, clearly explain to what population the results of this analysis can be inferred and whether causal inferences can be made supporting your claims using information from the article. [10pts]

8. Provide your SAS or R code for your analysis AND also provide the first 10 rows of the data set so that we can see the data you are analyzing. [5pts]