

Stat 402 A - Spring 2009 - Final Exam, (Takehome)

Due: Friday, May 8, by 5 pm.

You can turn in the exam during office hours, to my mailbox in 3010 Agronomy (available until 4pm), or under my office door (3405 Agronomy).

There are three questions, each with multiple parts.

I grade anonymously. Please **do not put your name on the front** of your answers. **Instead, put your name on the back of the last page** of your written/typed answers and also the last page of your SAS output (if any). Make sure all pages are securely attached.

Reminders: **You are not allowed to ask classmates or friends for help.** Please ask me, not a classmate, for help with SAS (or other computer program), even if you are working next to that classmate in the same computer room. Please sign the appropriate line on the last page when you are finished working on the exam.

You are allowed to use your book(s) and your notes. You are always welcome to e-mail me with questions or requests for clarification. I will not provide SAS code, but I am very happy to answer questions about syntax, output, or errors. If you do not understand the context for each problem, ask for clarification.

Some questions should be easy, some are harder, and a few require you to integrate ideas, which many folks find hard. You don't have to get a perfect score to do well on the exam. I adjust my letter grades to account for the difficulty of the exams. Your answers may be written or typed. You may submit just your answers. But, it is easier to determine partial credit if you include your SAS code and output.

Arrange your answers in the following order:

Answers and SAS **code** to question 1, ditto to question 2, ditto to question 3,
your name on the back of the last page of question 3 answers,
signature page

SAS output for Q 1, ditto for Q 2, ditto for Q 3 (if included)
your name again on the back of the last page of SAS output

You **DO NOT** need to copy your answers. All I'm asking is that you put the SAS output after the main part of your answers.

Reminder: e-mail, call (4-6828 or 4-2142), or come to office hours for help or clarification. Don't spend more than 30 minutes trying to fix a SAS issue. If you haven't found the problem in 30 minutes, get help **from me**.

Office hours during the exam:

over the weekend: e-mail to pdixon@iastate.edu

Monday: 3-5 pm, 3405 Agronomy

Tuesday: 9:15-11 am, 3405 Agronomy

Wednesday: 2-4 pm, 3405 Agronomy

Thursday: 2-3 pm, 3405 Agronomy

Friday: tentatively none because of an NSF site visit.

1. 40 pts. The data in viral.txt are motivated by studies of the effectiveness of vaccines against the PRRS virus in pigs. In this study, pigs were randomly assigned to receive one of three vaccine treatments (TRT): a placebo (A), a vaccine currently on the market (B) and a new vaccine (C). There are five pigs per vaccine and a total of 15 pigs in the study. A month after receiving the vaccine, pigs are challenged by deliberately exposing them to a known amount of the PRRS virus. The amount of virus in a pig is measured daily for 14 days after challenge. Hence, there are 14 measurements per pig and a total of 210 observations in the data set. The response (ITCID50) is the log transformed amount of virus in each pig.

- (a) 5 pts. Are these data an example of a repeated experiment, an example of repeated measurements, neither, or both? Explain your choice.
- (b) 5 pts. Consider only the five pigs in treatment C. Plot their responses over time. (The plot is easier to interpret if you put data from all five pigs on the same plot and mark each point with the pig id). What pattern(s) in this plot indicate that observations from a specific pig are correlated over time?
Note: You don't have to explain why the specific correlation model identified in part 1d is appropriate. You only have to explain why observations are not independent.
- (c) 10 pts. Consider a split-plot-in-time approach to these data. Outline the skeleton ANOVA table, indicating sources of variability, d.f., and whether a term should be considered fixed or random. Indicate the appropriate error term to test each fixed effect.

(d) 5 pts. We considered various correlation models for data like these. The fit statistics for various models are summarized here:

Model	d.f.	AIC	AICc	BIC
Independence	1	-14.9	-14.9	-11.8
cs	2	-81.2	-81.1	-79.8
ar(1)	2	-150.4	-150.4	-149.0
ar(1)+RE	3	-149.2	-149.0	-147.0
arh(1)	15	-142.3	-139.1	-131.6
ante(1)	27	-129.5	-118.7	-110.4
un	27	-129.5	-118.7	-110.4

Which model is the most appropriate model for the correlation? Justify your choice.

Notes: 1) Remember that -10 is **LARGER** than -15.

2) These fit statistics are calculated for one possible fixed effects model. Other choices of fixed effects model will give you different fit statistics.

- (e) 5 pts. Are the differences between drugs the same at all times? Explain why or why not. Your explanation should include a test statistic and p-value.
- (f) 5 pts. The investigators want to compare mean responses at each time. Previous inspection of the data suggests that it is reasonable to use a single error variance pooled over all treatments and times. At what times is there evidence of a difference among drugs?
Note: You don't have to tell me which drugs differ from each other at time 1, at time 2, ... Just test whether the three drugs have the same mean at time 1, at time 2, ...
- (g) 5 pts. After reviewing the conclusions from the previous analyses, the investigators decide to focus on a simpler analysis. The average of the 14 daily responses from a specific pig is a measure of the overall impact of the virus on that pig. Compute the mean for each pig, then use those summary statistics to answer:
Is there evidence of a difference among drugs? Explain why or why not.
Consider the three pairwise differences among drugs. Which drugs are different from each other? I suggest a Tukey adjustment for multiple comparisons.

2. 30 pts. The data in `lentil.txt` come from a multi-location study of the effect of FERTilizer, WEED control, and WEEVIL control on yield of lentils in Syria. 8 treatments, all combinations of high/low FERT, yes/no WEED control and yes/no WEEVIL control) were used in the study (i.e. the treatment design is a 2x2x2 complete factorial). These treatments were randomly assigned to 16 plots on a farm in a CRD. There were 2 replicates per treatment on a farm. The entire study was repeated at 8 farms throughout one province of Syria. There are 16 observations per farm and a total of 128 observations.

- (a) 5 pts. For a preliminary analysis, ignore the factorial treatment structure. There are 8 treatments in this study. Provide the degrees of freedom for the following ANOVA table, assuming intermediate sense inference.

Source	d.f.
Location	
Treatment	
Loc*Trt	
Error	

- (b) 5 pts. Explain what the error MeanSquare “means”. In other words, describe what sort of variability is being summarized by the Error MS. A biological interpretation is very much preferred to a statistical definition.
- (c) 5 pts. If you are interested in treatment differences on new farms in the region, which terms in the above ANOVA table would be considered fixed and which should be considered random? Explain briefly.
- (d) 5 pts. If you are interested in broad sense inference about the treatments, which terms in the above ANOVA table should be considered fixed and which should be considered random?
- (e) 5 pts. The 8 treatments have a 2x2x2 factorial structure, so it is natural to divide the 7 df for treatments into main effects, 2way interactions and the 3 way interactions. In lecture, I discussed two approaches for handling the location*treatment interaction. You could divide the interaction into components for each treatment effect (e.g. location*weed, location*weed*fert, ...), or you could pool all the components into a single location*treatment effect (as in the above ANOVA table). The effect of weed control depends on the weed intensity and the effect of weevil control depends on the weevil abundance. You know that some locations have almost no weeds while others have substantial numbers. Similarly, some locations have almost no weevils while others have substantial numbers. The effect of fertilizer is expected to be similar in all locations. Knowing this, will it be more appropriate to divide the location*treatment interaction into components (e.g. location*weed, ...) or to pool into a single location*treatment interaction? Briefly explain.
- (f) 5 pts. You decide that your questions require intermediate sense inference. Use an appropriate pooling of location interactions (i.e. your choice from part 2e) to test:
 is there an effect of weed control?
 does the effect of weed control depend on the amount of fertilizer?
 does the interaction of weed control and weevil control depend on the amount of fertilizer?
 For each question, report the appropriate test statistic and a p-value.

3. 30 pts. The data in spi.txt are motivated by a recent ISU study on the possibility that eating soy protein isolate (SPI) can reduce bone loss in older women. SPI has two components that may influence bone loss: a particular group of chemicals called isoflavones and the soy proteins. (I'm simplifying a lot of details here, so my apologies to soy experts). Previous studies suggested there was an effect of SPI, but the contributions of protein and isoflavones were not clear. In this study, the investigators compared 3 diets: a control with no soy (control), a diet with soy proteins and isoflavones (spi+) and a diet with soy proteins but no isoflavones (spi-). These diets were randomly assigned to perimenopausal women. The bone mineral content (BMC) was measured on each woman (BMC0), they ate their assigned diet for 24 weeks, then the BMC was measured again (BMC24). A total of 32 women completed the study, 10 on the control diet, 12 on the spi+ diet, and 10 on the spi- diet. The data file includes subject id, BMC0, BMC24, and diet.
- (a) 5 pts. The investigators decide to use ANCOVA to compare treatment means after adjusting for baseline BMC values. Write the skeleton ANOVA table, indicating sources of variability and d.f., for this analysis. Which effects should be considered random and which should be considered fixed?
 - (b) 5 pts. Is there evidence of any difference among treatments when treatments are compared at the same baseline value? Report an appropriate test statistic and p-value.
 - (c) 5 pts. Using the ANCOVA model, estimate the mean BMC24 for each treatment for women who had an initial BMC0 of 60.
 - (d) 5 pts. i) Estimate the effect (on BMC24) of adding soy protein without isoflavones to the diet (i.e. comparing spi- to the control diet). Report your estimate and its precision. ii) Estimate the additional effect of adding isoflavones to a soy protein diet. In other words, if you were consuming a soy diet without isoflavones, what additional effect would you expect if you added isoflavones to your diet? Again, report your estimate and its precision.
 - (e) 5 pts. The treatments in this study could be viewed as a two factorial: 2 levels of soy protein (none, some) x 2 levels of isoflavone (none, some). Is it possible to estimate the interaction between protein and isoflavone in this study? Explain why or why not.
 - (f) 5 pts. After examining the results from the ANCOVA, the investigators ask their statistician to analyze the differences, i.e. use $BMC24 - BMC0$ as the response variable. They find that their results (e.g. estimates of treatment effects) are similar to those from the ANCOVA. Is this a surprise for these data? (It certainly does not always happen). Explain why, for these data, you would (or would not) expect ANCOVA to be similar to an analysis of post-pre differences.

That's all. It's been a delight working with you all semester. I appreciate all your hard work. Good luck on your other finals and have a great summer!

Please check the appropriate box, sign, and include at the end of your answers.

_____ I completed this exam without assistance from friends or classmates.

_____ I received the following assistance from friends or classmates (please describe below).

Signed: _____