METHODS QUALIFYING EXAM

January 2007

INSTRUCTIONS:

1. DO NOT put your NAME on the exam. Place the NUMBER assigned to you on the UPPER LEFT HAND CORNER of EACH PAGE of your exam.
2. Please start your answer to EACH QUESTION on a SEPARATE sheet of paper.
3. Answer all the questions.
4. Be sure to attempt all parts of every question. It may be possible to answer a later part of a question without having solved the earlier parts.
5. Be sure to hand in all of your exam. No additional material will be accepted once the exam has ended and you have left the exam room.

Problem I.

A researcher was interested in studying the effect of the distance of an object from the eye on the eye focus time. There are t different distances of interest. The researcher has r subjects available for the experiment. Because there may be differences among subjects, she decides to conduct the experiment in a randomized block design, that is, let y_{ij} denote the focus time from the jth subject using the ith distance. The following model was used for this data: For $i = 1, \ldots, t$ and $j = 1, \ldots, r$

$$y_{ij} = \mu + \tau_i + b_j + e_{ij}$$

where $\mu, \tau_1, \ldots, \tau_t$ are unknown parameters; $b_1, \ldots, b_r$ are iid $N(0, \sigma_b^2)$ random variables; $e_{ij} (i = 1, \ldots, t; \quad j = 1, \ldots, r)$ are iid $N(0, \sigma_e^2)$ random variables; all b_j and e_{ij} are jointly independent; and σ_b^2 and σ_e^2 are unknown positive variances.

a) In the following AOV table, fill in the formulas for the missing elements in the df and SS columns:

Source of Var.	D.F.	Sum of Squares
Subject	_____	_____
Distance	_____	_____
Error	_____	_____
Total	n-1	$\sum_{i=1}^{t} \sum_{j=1}^{r} (y_{ij} - \bar{y}_{..})^2$

b) What is the expected MS for the Block term?

c) Using your expression from part b), provide an estimator of σ_b^2.

d) Provide an expression for the test statistic for testing $H_o : \tau_1 = \cdots = \tau_t$.

e) The researcher in fact was also interested in the impact of Age of the subject on the eye focus time. The age variable was divided into m groups. Let y_{ijk} denote the focus time from the jth subject from the kth age group using the ith distance. The following model was used for this data: For $i = 1, \ldots, t$; $j = 1, \ldots, r$; and $k = 1, \ldots, m$;

$$y_{ijk} = \mu_{ik} + b_j + e_{ijk}$$

where $\mu_{ik} (i = 1, \ldots, t; \quad k = 1, \ldots, m)$ are unknown parameters; $b_1, \ldots, b_r$ are iid $N(0, \sigma_b^2)$ random variables; $e_{ijk} (i = 1, \ldots, t; \quad j = 1, \ldots, r)$ are iid $N(0, \sigma_e^2)$ random variables; all b_j and e_{ijk} are jointly independent; and σ_b^2 and σ_e^2 are unknown positive variances. Under the above model, answer the following questions:

1) Provide an expression for the correlation between the eye focus times for a subject in Age Group 1 viewing the object at Distances 2 and 3.

2) The researcher was very interested in determining if there was an interaction between Age and Distance. In terms of the model parameters given above, state the null hypothesis for the test of interaction between Age and Distance.

PROBLEM II.

The American Red Cross does a detailed analysis every month on 100 randomly selected blood samples from donors in Texas. The samples produced a measurement,

X_i = Serum lead level (in micrograms per deciliter) for sample i.

From past experience, the distribution of X_i is known to be lognormal. Thus, a statistician defined the transformed variable $Y_i = \ln(X_i)$. For purposes of our analysis, we will assume that

$$Y_i = \mu + e_i \tag{1}$$

where μ is a fixed mean parameter for blood donors in Texas and $e_i,\ i = 1, \ldots, 100$ are independent and identically distributed $N(0, \sigma_e^2)$ random variables.

a) Using the above model, provide an expression for a 95% confidence interval for μ.

b) Now consider a new observation Y_{n+1} that also satisfies model (1). Provide an expression for a 95% *prediction interval* for Y_{n+1}, based on the old data $Y_i,\ i = 1, \ldots, 100$.

c) Give a customary interpretation of your prediction interval in b), paying special attention to: i) what is fixed; ii) what is random; and iii) to what probability does the term "95%" refer?

d) Explain how the interpretation of your prediction interval from c) differs from the customary interpretation of the confidence interval in a).

e) Recall that our transformed observations $Y_i = \ln(X_i)$ were recorded on a log-transformed scale. Using the results from the preceding steps to provide an expression for a 95% prediction interval for a new observation X_{n+1} on the *original* (untransformed) scale.

f) Explain why the procedure you used in e) to obtain the 95% prediction interval for X_{n+1} would not yield an **exact** 95% confidence interval for $E(X_i)$.

g) Describe an alternative method for finding a 95% confidence interval for $E(X_i)$.

h) Suppose the Red Cross wants to compare the mean serum lead levels for the previous 12 months.

   1) How would you adjust the individual monthly C.I.s in order to obtain a simultaneous coverage probability of 95% for the 12 monthly C.I.s?
   2) What would be the impact on the average widths of the monthly C.I.s?
   3) Describe how the simultaneous monthly C.I.s could be used to determine which pairs of months had different means.

i) The prediction interval in part b) is based on the assumption that the Y_i values are normally distributed. List *two* specific methods that you could use to check this assumption based on the transformed observations. For *each* of these two methods, give explicit rules (omitting numerical critical values) you would follow to decide whether the observations are consistent with the normal-distribution assumption.

Problem III

A scientist wants to study the effect of jazz music on intelligence in rats. She uses 4 female rats from a single birth episode. Two of the rats are raised in a quiet laboratory and the other two rats are raised under the same conditions, except that they listen to Thelonious Monk on the piano for 6 hours each day. At maturity, each rat runs through a maze, the times (in minutes) for the rats to run the maze are $y_1 = 5$ and $y_2 = 7$ (quiet), and $y_3 = 1$ and $y_4 = 9$ (Monk piano). The scientist proposed the following model for analyzing the data:

$$\mathbf{Y} = \mathrm{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where the elements in $\mathbf{Y}$ are the times, $\boldsymbol{\varepsilon}$ is a random vector with $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ and $Var(\boldsymbol{\varepsilon}) = \sigma^2 \mathrm{I}$, $\boldsymbol{\beta} = (\beta_0, \beta_1)^{\mathrm{T}}$ and

$$\mathrm{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

Give numerical answers to the following questions whenever possible.

a) Explain in one sentence each interpretations of the parameters β_0 and β_1.

b) Determine the least squares estimate of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$.

c) Determine $Var(\hat{\boldsymbol{\beta}})$.

d) Determine the "hat" matrix, $\mathbf{H}$.

e) Give an unbiased estimate of $\beta_0 + \beta_1$.

f) What additional assumptions are needed so the estimate in e) is the best linear unbiased estimate? Why?

g) Suppose there is a fifth female rat from the same litter and suppose that she was raised in a quiet lab. Predict her time, $\hat{y}_5$, to run the maze on her first try at maturity.

h) Estimate σ^2 unbiasedly using the original four observations.

i) Suppose you had only the 2 rats in the quiet group, how would you then estimate σ^2 using these two observations?

j) Based on the answers to h) and i) what assumption(s) may be violated?

k) Estimate the variance of the estimate in e) based on the original four observations.

l) Estimate $Var(\hat{y}_5 - y_5)$ where $\hat{y}_5$ is defined in g).

m) After correcting for the mean, what proportion of the total variability in y is explained by the model?

PROBLEM IV.

Below is a message from Nicki to the SPSS discussion board asking for help. Following her message are suggestions from an anonymous respondent, Stephen.

After reading Nicki's questions and Stephen's suggestions, answer the following questions a), b), and c). Give reasons for all of your answers. Be very critical of Stephen's suggestions. Do they make sense? Would you recommend doing what he suggests?

a) For each of Stephen's suggestions i), ii), and iii); do you agree or disagree with Stephen's point? Give a brief discussion of why you agree or disagree.

b) Explain to Nicki what it means to have a significant interaction term between the fixed factor and the continuous covariate.

   1) give the model in terms she can understand;
   2) sketch a graph to explain what a significant interaction means;
   3) explain to her in terms of the model what a significant interaction means.

c) Explain what is being tested by the pairwise comparisons. Answer in terms of the model.

Nicki's Questions:

I ran an ANCOVA model which yielded a significant interaction between a fixed factor (having four levels) and a continuous covariate. I am interested in investigating the interaction further. I have the additional problem that the pairwise comparisons were not significant. So I am wondering if there are any other statistical methods to investigate an interaction between a continuous covariate and a factor.

Stephen's Suggestions:

i) One of the core assumptions of ANCOVA is that the relationship between the covariate and the dependent variable is the same for both groups. When you have a significant group by covariate interaction, then this assumption is not met.

ii) As a result, I would not use ANCOVA to analyze these data. I would use a regression framework instead, with the following predictors: 1) the continuous variable; 2) dummy codes for the categorical; 3) a cross-product, or interaction term, between the continuous variable and each of the dummy codes.

iii) You can use the regression coefficients to compute values of the dependent variable for each group at one standard deviation above and below the mean on the continuous variable. These points will allow you to plot the slope and intercept of the continuous variable for each group.