BASKIN SCHOOL OF ENGINEERING
Department of Applied Mathematics and Statistics

First Year Exam, Take Home Question (Statistics)

Due by 2PM, Wednesday, June 20, 2012

Instructions:

Please work individually on this problem. Do not share with anyone any information or comments about your findings or the models and methods you use. You are required to write a report using a word processing software (i.e., LaTeX or Microsoft Word). You are required to email your report as ONE PDF file to Hongyun Wang at `hongwang@ams.ucsc.edu`

by 2PM, Wednesday, June 20, 2012.

Please take care to organize and present the material in the best possible way. Be informative but concise. You should include a summary of your work at the beginning of the report, include and annotate all relevant figures and tables in the body of the report, write your conclusions in a separate section and list the references (if any) that you consider appropriate. Provide information about the method and the software you use to fit the models. Your report should consist of no more than 8 letter-size pages (typeset with 11pt or larger font and margins on all four sides of at least 1 inch), including all figures, tables, and appendices; answers longer than 8 pages will lose credit for excess length.

General Guide and Requirements:

Start your paper with an abstract that contains a short description of the problem and the main findings. Then, the first part of the body of the paper will correspond to an introduction with a description of the problem. The methods and the analysis will follow. The paper will finish with concluding remarks and references. Tables and figures, if any, need to be part of the text. Do not append them to the end of the paper. You have a maximum of ten pages, including figures and tables.

Your answers should be carefully in Latex. You are suggested to use the template from http://classes.soe.ucsc.edu/ams207/Spring11/

You are allowed to consult any material you wish, but you should not communicate with any other individual (student or faculty). All work must be your own.

Questions:

Analyzing a designed experiment: Table 1 displays the results of a randomized blocks experiment on penicillin production.

Block	Treatment			
	A	B	C	D
1	89	88	97	94
2	84	77	92	79
3	81	87	87	85
4	87	92	89	84
5	79	81	80	88

Table 1: Yields of penicilin produced by four maufacturing processes, each applied in five different conditions. Four runs were made within each block, with the treatments assigned to the runs at random.

a. Fit a normal linear regression model to the randomized block data above by Bayesian approach. Let the model be $Y = X^T\beta + e$, with $e \sim N(0, \sigma^2 I)$, where I denotes the identity matrix. Use prior $p(\beta, \sigma^2) \propto 1/\sigma^2$.

b. Summarize posterior inference for the (superpopulation) average penicillin yields, averaging over the block conditions, under each the four treatments. Under this measure, what is the probability that each of the treatments is best? given a 95% posterior interval for the difference in yield between the best and the worst treatments.

c. Set up a hierarchical extension of the model, in wich you have indicators for all five blocks and all four treatments, and the blodk and treatment indicators are two sets of random effects. Explain why the means for the block and treatment indicator groups should be fixed at 0. Write the joint distribution of all model parameters (including the hierarchical parameters).

d. Compute the posterior mode of the three variance components of your model in c. using EM algorithm. Construct a normal approximation about the mode and use this to obtain posterior inferences for all parameters and answer the questions in b.

e. Check the fit of your model to the data. Discuss the relevance of the randomized block design to your check; how would the posterior predictive simulations change if you were told that the treatments had been assigned by complete randomization?

f. Obtain draws from the actual posterior distribution using the Gibbs sampler, you may use your results from d. to obtain starting points, or use some other proper start points. Run multiple sequences and monitor the convergence of the simulations by computing $\sqrt{\hat{R}}$ for all parameters in the model.

g. Discuss how your inferences in b. d. and f. differ.