Exam in MSA830 Statistical Analysis and Experimental design

March 25th 2008, 8:30 – 13:30
Jour: Malin Östensson, who will be available for questions about
the formulations of the exam questions at 9:30 and 11:30.
Allowed during the exam: An optional calculator, and one single page of your own notes.
Number of points on the exam: 30. To pass the exam, at least 12 points are needed.

1. Astrid is repeatedly performing an experiment with three possible outcomes: X, Y, or Z. Previous work in her scientific area has established that the probability for observing X is 0.2, the probability for observing Y is 0.5, and the probability for observing Z is 0.3.

   a) If Astrid performs 8 experiments, what is the probability that 3 of them have outcome X? (2 points)

   b) If Astrid performs 8 experiments, what is the probability that 3 of them have outcome X, 3 of them have outcome Y, and the rest outcome Z? (1 point)

   c) If Astrid performs 30 experiments, what is the approximate probability that 10 or more of the outcomes will be X? (2 points)

2. Fredrik is comparing the durability of two types of house paint: Type A and Type B. In an experiment, he has selected 5 houses. For each house, he randomly selects two of the walls and paints them with A, while the two other walls are painted with B. After one year, he measures the durability of the paints, with the results in the table below, where each line corresponds to one house:

Type A	***Type B***
4	9
5	6
3	7
7	6
4	5

   a) Make a hypothesis test of whether the two paint types have equal durability. Use a significance level of 5%. What assumptions does your test depend on? (4 points)

   b) Fredrik would like to analyse the data using a randomization test. He is uncertain about which one of three procedures to choose, to do his calculations:

   I) Compute the average of the Type B responses, and subtract the average of the Type A responses. Then, randomly select 5 values from the table, compute their average, and subtract the average of the remaining 5 values. Do this many times by simulation on a computer. Finally, compare the original difference with this simulated set of differences; if it is larger than all or most of them, the hypothesis that there is no difference in durability between the two types can be rejected.

   II) For each line, compute the Type B response minus the Type A response. Compute the average

of these 5 differences. Then, take these 5 differences and multiply each with +1 or -1; there are 32 different ways to do this. For each of these ways, compute the average of the result. Compare the first computed average with these 32 averages: If it is larger than all or most of them, we can reject the hypothesis that there is no difference in durability between the two types.

III) Select randomly between 1 and 10 numbers from the table, find their average, and subtract the average of the remaining values. Do this a large number of times on a computer. If the average of the computed numbers is larger than zero, reject the hypothesis that the two types have the same durability.

Which procedure would you recommend him to use? (1 point)

3. Eric has performed a factorial experiment investigating the dependence of the yield of some process on 4 different factors, A, B, C, and D, where each can have a high and a low value, denoted + and – in the table below. He has performed 32 experiments in total, with 4 experiments for each row in the table below. For each such set of 4 experiments, he has computed their average yield, and the sample variance:

A	***B***	***C***	***D***	***Average yield***	***Sample var.***
-	-	-	-	4.0	3.0
-	-	+	+	7.0	2.0
-	+	-	+	5.0	5.0
-	+	+	-	3.0	3.0
+	-	-	+	1.0	4.0
+	-	+	-	4.0	2.0
+	+	-	-	2.0	3.0
+	+	+	+	4.0	2.0

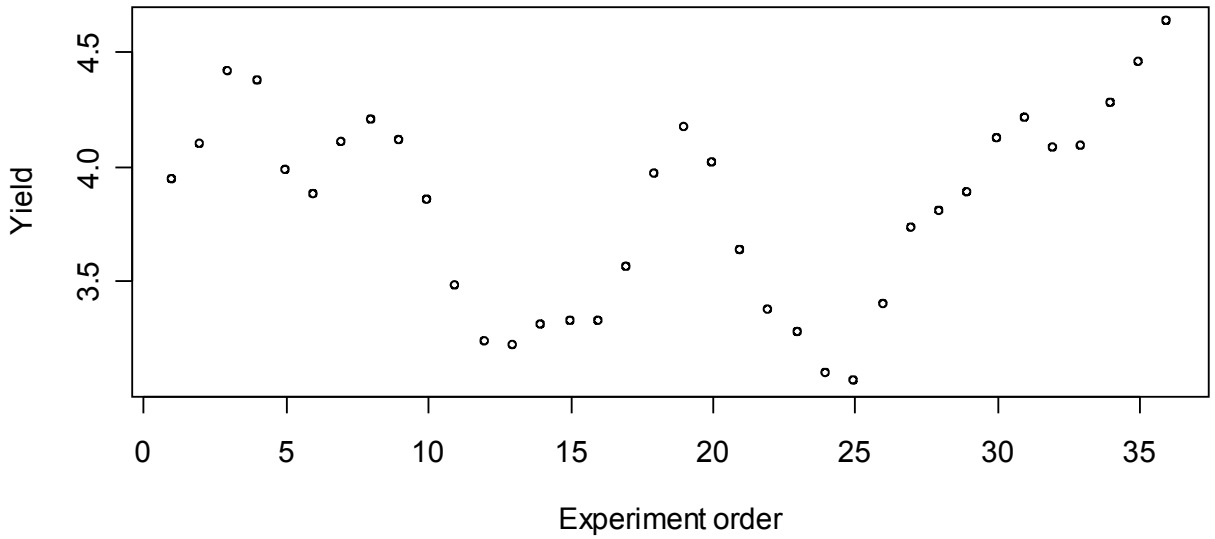
a) From the table, compute the main effect of A, and the interaction effect between B and C. (2 points)

b) Write down the name of the design Eric has used, on the form 2^{*}_{*}. Write down a generating relation for the design. (2 points)

c) Compute the estimated standard error for any of the effects. (2 points)

d) Compute the mean yield of a process of the type above, disregarding any effect of the factors, and find a 95% confidence interval for this mean. (2 points)

e) After he has finished the experiment above, Eric continues with a number of experiments using the settings of the last line in the table above. His yields are plotted in the diagram below against the order in which he performed the experiments. Eric also tells you that he has performed the original experiments sequentially, starting with the 4 in the first line, going on with the 4 in the next line, etc. Does this mean there is a problem with his analysis? Explain. (1 point)



f) If he had done the experiments in a randomized order, is he likely to have found smaller or larger sample variances? (1 point)

4. Mary is trying out new air filtering machines at the hospital where she works. She has to choose between three brands: B1, B2, and B3. For each brand, she does three independent tests, and records the efficiency. The results for B1 are 14.0, 13.0, and 12.0. For B2, they are 7.0, 9.0, and 11.0. Finally, for B3, they are 13.0, 13.0, and 16.0.

   a) Construct an ANOVA table which can be used to test whether there is a difference in the efficiency between the brands. Compute the F statistic, and find a p-value for testing the null hypothesis that there is now difference in the efficiency between the brands. (4 points)

   b) It turns out that brand B3 will be too expensive, so Mary only has to compare brands B1 and B2. If she would like to do a comparison using a t-test, what kind of t-test should she choose? (You only need to give the name). Should she choose a one-sided or two-sided t-test? (1 point)

   c) When she was planning the tests, Mary could choose between the following test schedules:

   I) Monday 8:00-12:00: B1, Monday 12:00-16:00: B1, Monday 16:00-20:00: B1.
   Tuesday 8:00-12:00: B2, Tuesday 12:00-16:00: B2, Tuesday 16:00-20:00: B2.
   Wednesday 8:00-12:00: B3, Wednesday 12:00-16:00: B3, Wednesday 16:00-20:00: B3.

   II) Monday 8:00-12:00: B1, Monday 12:00-16:00: B2, Monday 16:00-20:00: B3.
   Tuesday 8:00-12:00: B1, Tuesday 12:00-16:00: B2, Tuesday 16:00-20:00: B3.
   Wednesday 8:00-12:00: B1, Wednesday 12:00-16:00: B2, Wednesday 16:00-20:00: B3.

   III) Monday 8:00-12:00: B1, Monday 12:00-16:00: B2, Monday 16:00-20:00: B3.
   Tuesday 8:00-12:00: B3, Tuesday 12:00-16:00: B1, Tuesday 16:00-20:00: B2.
   Wednesday 8:00-12:00: B2, Wednesday 12:00-16:00: B3, Wednesday 16:00-20:00: B1.

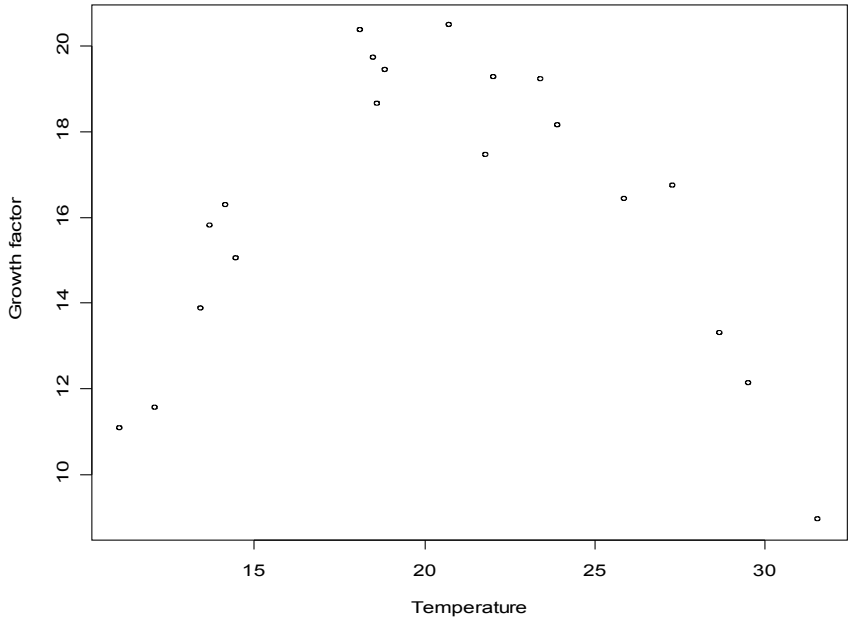
   Which one would you advice her to choose, and why? (1 point)

5. Jamal is studying how the growth of a certain cell line depends on temperature. He chooses 20 values $x_1,x_2,...,x_{20}$ for the temperature and observes the 20 resulting values $y_1,y_2,...,y_{20}$ for cell growth (over 24 hours). He would like to fit the following model to the data, using least squares:
$y_i=\beta_0+\beta_1 x_i+\epsilon_i$ for $i=1,\dots,20$

a) Using the symbols from the model above, write down the quantity that is minimized when the least squares solution is found. (1 point)

b) Using least squares, Jamal finds the solution $\hat{\beta}_0=16.95$ and $\hat{\beta}_1=-0.04$. He would like to compute confidence intervals for these two estimates. What assumptions does Jamal have to make in order to be able to compute confidence intervals? (1 point)

c) He finally plots his data, getting the plot below. Is there a problem with the confidence intervals he computed in b? Explain. (1 point)



d) Three possible models that Jamal could use to model his data are given below:
$y_i=\beta_1 x_i+\epsilon_i$
$y_i=\beta_0+\beta_1 x_i+\epsilon_i$
$y_i=\beta_0+\beta_1 x_i+\beta_2 x_i^2+\epsilon_i$
where ϵ_i is normally distributed with expectation zero and variance σ^2 . If he estimated the parameters of each model using least squares, for which model would the resulting sum of squares of residuals be smallest? For which model would it be largest? Explain. (1 point)