Exam in MSA830 Statistical Analysis and Experimental design

February 20th 2008, 8:30 – 12:30

Jour: Petter Mostad (phone 0707163235), 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. Johan has heard that cows produce more milk when they get to listen to classical music. He sets up an experiment to test this, where he from his 12 cows randomly selects 6 which get to listen to classical music, and 6 that do not. The resulting milk production over a week is then measured for each cow:

Cows listening to classical music	113	118	126	121	111	119
Cows with no music	110	116	110	121	117	110

   a) Johan would like to make a hypothesis test. Assuming that the observations in each group come from a normal distribution, choose a test, and compute a p-value for your test. Do you have to make additional assumptions in order to use your test? How would you conclude? (3 points)

   b) Based on the computations above, find a 90% confidence interval for the difference between the expected milk production with and without music. (1 point)

   c) Johan would like to study the question further, so he takes the music away from the cows that had been listening, and instead lets the other six cows listen to the music for a week. This gives him 12 new measurements. How can he analyze his new and old data together? Describe with formulas how to do the computations. Why is he likely to get a better result this time? (2 points)

   d) Is there a test Johan could use that does not assume that the measurements in each group come from a normal distribution? Give a short description of how to make computations with such a test. (2 points)

   e) Assume Johan wanted to find out whether the musical style mattered. He selects 5 different musical styles, and performs a new experiment lasting 5 weeks, where each cow gets to listen to one week of each type of music, whereafter its milk production for that week is measured. What kind of analysis method could he use on the resulting data? (It is enough to describe the method in one sentence.) (1 point)

2. Sara is investigating the rate of triplet births in her town. The intial information she has is that there were 12 in 2006 and 17 in 2007.

   a) How would you estimate the rate of triplet births per year, given this information? (1 point)

   b) Sara later gets some more information about the issue, and establishes the rate 14.2 triplet births per year in her town. With this rate, what is the probability of observing exactly 3 triplet births in the coming year? (1 point)

   c) What is the approximate probability of observing 20 or more triplet births in the coming year? (1 point)

3. Li is comparing the effect of three different types of running shoes on his running time. He has made, in a randomized order, 5 test runs with each type of running shoes. He would like to test his null hypothesis that the running shoes do not influence his running times.

   a) He has started to fill out the ANOVA table below for his data analysis. Making the usual NIID assumptions, complete the table and compute a p-value for Li's null hypothesis. (1 point)

	Sum of squares	Degrees of freedom	Mean sum of squares	Test statistic
Treatment	213			
Residuals				
Total	432			

   b) Is it possible to make a randomization test in this context? Describe shortly what the algorithm would do. What would be the advantages of such a test compared to the ANOVA test above? (2 points).

4. A group of students have performed a two-level factorial experiment where they investigate how the factors temperature, amount of water, and amount of light influence the growth of a certain type of plants. For each combination of factor levels, they have 3 repetitions of the experiment, and their initial results are given in the table below:

Temperature	Water	Light	Sample mean	Sample variance
-	-	-	4	5.5
-	-	+	8	1.3
-	+	-	6	4.0
-	+	+	17	1.3
+	-	-	3	0.2
+	-	+	9	1.5
+	+	-	6	4.0
+	+	+	11	10.2

   a) Compute the main effect of the Water factor and the interaction effect between Temperature and Light. (2 points)

   b) Making the usual NIID assumptions, find the 95% confidence interval for the two effects computed in a. (2 points)

   c) How can the dependence of the sample means on the factor settings be illustrated? Make a quick drawing. (1 point)

   d) Imagine that all the 3 observations for the last row of the table were lost or destroyed, for some reason. Would it still be possible to analyze the data? Mention the name of an analysis method that could be used. (1 point)

   e) What exactly were the assumptions made in b[1]? Mention three things we can do with the data

1 In the original exam, there was a printing error here; it was written a instead of b.

to check whether these assumptions are reasonable or not. (1 points)

5. Carmen is planning an experiment investigating the yield of a production process. She has 6 factors, T, C, X, P, B, Y, and she has two levels for each factor,written ”+” and ”-”, that she wants to try out. She would like to do 16 experimental runs in her initial experiment, where each experiment combines different settings of the 6 factors.

   a) Propose an experimental design for her, by writing down a table with 6 columns, each headed by the name of a factor, and with 16 rows, each consisting of ”+” and ”-” signs , indicating the levels of the factors for this experimental run. The experimental design should be balanced, so that each main effect can be estimated independently, and no main effect should be confounded with a two-way interaction effect. (2 points)

   b) Write down the generating relations of your design, and its name, on the form 2^{*}_{*}. (1 point)

6. Anna is doing clinical medical research, and wants to compare the effects of medication A with those of medication B. A will be adiministered to n patiends, and B to n different patients, in a randomized trial. For each patient, a response y, on a continuous scale, will be measured. Based on earlier studies, Anna believes that the standard deviation of the responses in both cases will be 14. The minimal difference in effect that it is clinically relevant to detect is 5, and Anna wants to detect this effect with 90% probability, if it is there, using a test with a 5% significance.

   a) Approximately how many patients does Anna need in each group of patients? (1 point)

   b) If the standard deviation of the effects had been 7 instead of 14, how would that change the results found in a? (1 point)

   c) If she decides that the analysis should be done with a test with 10% significance instead of one with 5% significance, would she need to increase or decrease the sample size to achieve the same power to reject the null hypothesis? (1 point)

7. Nils is working at a factory, and he is experimenting with the levels of two different additives, A and B, in the process he is monitoring, in order to optimize the yield of the process. He has performed 10 experiments with different combinations of the additives, measuring the yield each time. The results are given in the table below. The yields have sample mean 35.6 and sample standard deviation 3.8. Nils analyzes the data using regression.

Additive A	Additive B	Yield
10	3	31.3
10	5	35.4
20	3	34.4
20	5	37.4
30	5	38.3
30	7	42.1
25	7	39.9
15	2	33.3
15	1	29.9
15	5	33.8

a) After he has done the regression, Nils tells you that the sum of the squares of the residuals in the regression is 152.6. How can you tell that he must have done an error? (1 point)

b) The sample correlation between the values of the two additives is 0.59. If Nils had performed the experiment so that this correlation was closer to zero, would he have been likely to get smaller or larger confidence intervals? Explain your answer. (1 point)