Petter Mostad
Mathematical Statistics
Chalmers

MSA830 Statistical analysis and experimental design

Exam 12 January 2011, 8:30 - 13:30
Examiner: Petter Mostad, phone 0707163235,
visits the exam at 9.30 and at 11.30.
Allowed to use during the exam: Pocket calculator, books, copies, and notes
Number of points on the exam: 30. To pass the exam, at least 12 points are needed

1. Alex has two different ways to drive to work: using route A or using route B. The time he uses on either route varies each day he tries it, but he is interested in the expected time difference between the two routes. The observations he has collected are:

Route A	67, 66, 76, 75, 74, 75
Route B	42, 64, 71, 73, 44, 62

   (a) Find a 95% credibility interval for the expected time difference between the two routes, assuming the observations are random samples from two normally distributed populations with equal variances. Which route would you recommend Alex to follow? (2 points)

   (b) Assume now that the variances of the two populations may be different. Find 90% credibility intervals for the variance of the population of times using route A. (2 points)

   (c) Make a hypothesis test of whether the variances in the two populations of observations are the same. (2 points).

   (d) Find a 95% credibility interval for the expected time difference between the two routes, assuming the observations are random samples from two normally distributed populations which do not necessarily have equal variances. (2 points)

   (e) The scientific reproducibility of the conclusions found above depend on how Alex has performed his testing. Give an example on how he might have organized or performed his testing that would cast doubt on the scientific reproducibility of the conclusions. Give another example on a way he might have organized or performed his testing that would strengthen your belief on the scientific reproducibility of the conclusions. (2 points).

2. Lisbeth works with animal psychology, and is studying the ability of wombats to perform a particular task. She has selected 6 wombats from each of three different locations, for a total of 18 wombats. From each location, she has selected 3 males and 3 females. The resulting scores are given in the table below. The average in each group of 3 wombats is given in each cell[1]. Averages over each location is given at the bottom line, while averages for females and males are given in the right column. The grand average is 26, while the variance of the data is 28.59.

[1] In the original exam, there was an error: The average for Female and location A was given as 23

	Location A	Location B	Location C	Average
Female	24 23 (av. 24) 25	32 24 (av. 27.67) 27	33 30 (av 29.33) 25	27
Male	24 21 (av. 22.33) 22	25 21 (av. 20.67) 16	39 26 (av. 32) 31	25
Average	23.17	24.17	30.67	26

(a) Make a complete ANOVA table for Lisbeth's experiment. Includ interaction between the factors. For the p-values: Find an interval containing each p-value. (4 points).

(b) Explain the results of the analysis using words that can be understood by somebody who knows nothing about ANOVA tables or statistics. (1 point)

(c) Re-write the entire table so that the analysis does not include interaction. (2 points)

3. Sara is observing at a distance an animal of a particular species, and she is trying to determine if the animal is male or female. Initially, she believes there is a 50% chance of each possibility, but then she finds that the size of the animal is above a certain limit: 40% of all males are larger than this limit, while only 10% of all females. Given this information, what is the probability that the anmial is male? (2 points)

4. Hugo is researching amateur observations of a particular type of comets. In Sweden, in the time period 1990-2009, there were a total of 54 observations of this type of comet.

   (a) Assuming all observations are independent, and that the rate of observations is constant, what is the expected number of observations during 2011? (1 point)

   (b) What is the probability that two or fewer comets will be observed during 2011? (1 point)

5. Lars is working for a medical company that wants to test whether whether a new medication B works better than the standard medication A for treating a particular disease. The effect of the medication is measured with a number x at a follow-up visit after a few weeks, for each patient. Over the period of the experiment, Lars can expect to have about 100 patients with the particular disease, who might be enrolled in the study.

   (a) Describe how to organize the experiment in such a way that the results are as scientifically reproducible as possible. Discuss such things as how a medication should be selected for each patient, and what kind of information the patients, and Lars, should be given. Give an argument why your recommendations are correct. (2 points)

   (b) The effect of medication A, as measured with the variable x above, has been found to be approximately normally distributed, with expectation 3.4 and standard deviation 2.1. The medical company would like to find a 95% credibility interval of length at most 0.3 for the difference in expected effect between medication A and medication B. If each doctor can expect around 100 patients over the period of the experiment who could be enrolled in the study, how many doctors should the company at least include in the study? Assume that the standard deviation of the effect of medication B will be as for medication A. (2 points)

6. Jeanette is setting up an experiment where she has three factors, A, B, and C, which can each have two possible levels, and for each possible combination of levels, she can measure an outcome X. She uses a full factorial design with two replications (i.e., trying out each combination of levels twice).

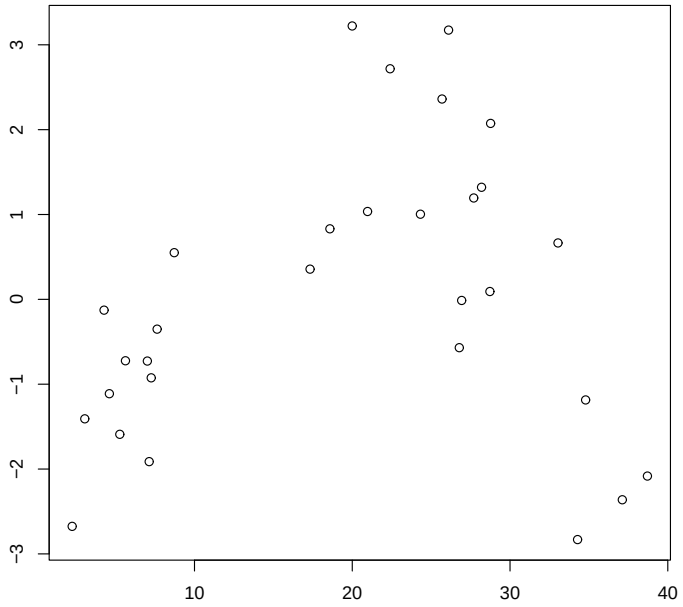
   (a) Jeanette first wants to analyze the data with a linear model disregarding interactions between the factors. Write down the design matrix she should use.(1 point)

   (b) Jeanette then wants a analysis of the data that includes all possible types of interactions between the factors. Write down the design matrix she shouls use. (1 point)

7. Lurleen is studying how a certain output y depends on three different predictors x_1, x_2, and x_3. For each of 30 experiments, Lurleen has measured the values of y, x_1, x_2, and x_3, and the resulting data has been analyzed using multiple regression, i.e., a linear model of the form

$$y = \beta_1 + \beta_2 x_1 + \beta_3 x_2 + \beta_4 x_3 + \epsilon$$

   where ϵ is assumed to be normally distributed with expectation zero. Plotting the residuals on the y-axis and the corresponding values of the predictor x_2 on the x-axis, Lurleen gets the plot below.



   (a) Does the plot indicate that the model is appropriate for the data, or that there is a problem? Explain. (1 point)

   (b) The residuals plotted above sum to zero. This is always the case, no matter what the data is. Give an argument why this must be so. (2 points)