Petter Mostad
Mathematical Statistics
Chalmers

MSG830 Statistisk analys och experimentplanering

Exam 30 May 2013, 8:30 - 12:30
Examiner: Petter Mostad, phone 0707163235,
visits the exam at 9:30 and at 11:30
Allowed to use during the exam: Pocket calculator
Number of points on the exam: 30. To pass the exam, at least 12 points are needed

1. In Figure 1 are scatterplots of four datasets. In each dataset, bivariate data with variables X and Y has been observed.

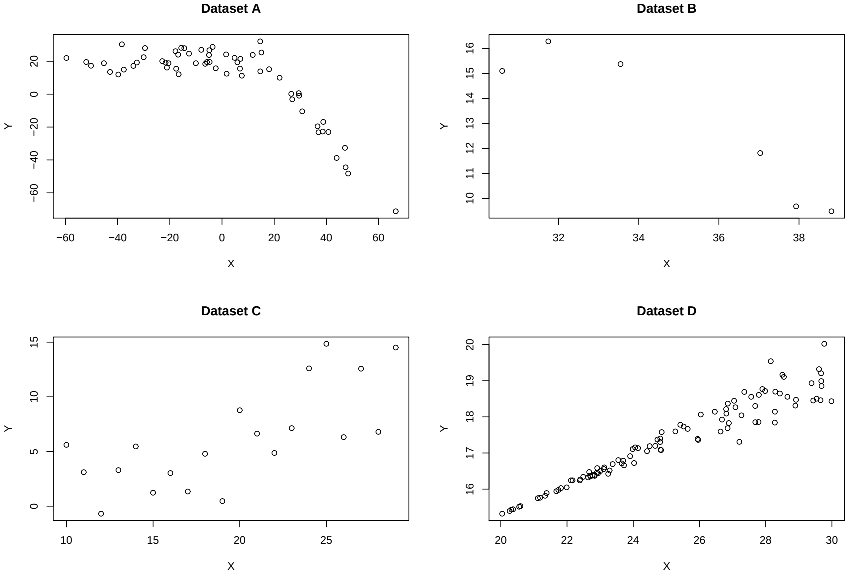

Figure 1: The datasets for question 1.

   (a) From looking at each of the plots for the datasets A, B, C, and D, do you believe the correlation is positive or negative in each case? (2 points)

   (b) For each dataset A, B, C, and D, assume X is the predictor variable and Y the response variable. For each dataset, answer whether you think a linear regression model would be appropriate for the data. If you think it is not appropriate, explain why, and if you have an alternative model, mention it. (2 points)

2. Explain the concepts of Type I error and Type II error. (2 points)

3. Yusuf is packing for a holiday. He has 12 shirts and would like to bring 4 of these. Compute the number of different selections of shirts he can make. (2 points)

4. Explain what Principal Component Analysis is, and how it is used (2 points).

5. Anna works with various chemical processes at a factory. In case A, she has investigated how the yield of a process varies when she tries the temperature at 3 different levels, and the pressure at 2 different levels. Each combination is tried out four times. A plot of the results is given in Figure 2, using x for observations with the low pressure and o for observations with the high pressure. Similar results for another process is shown as case B.

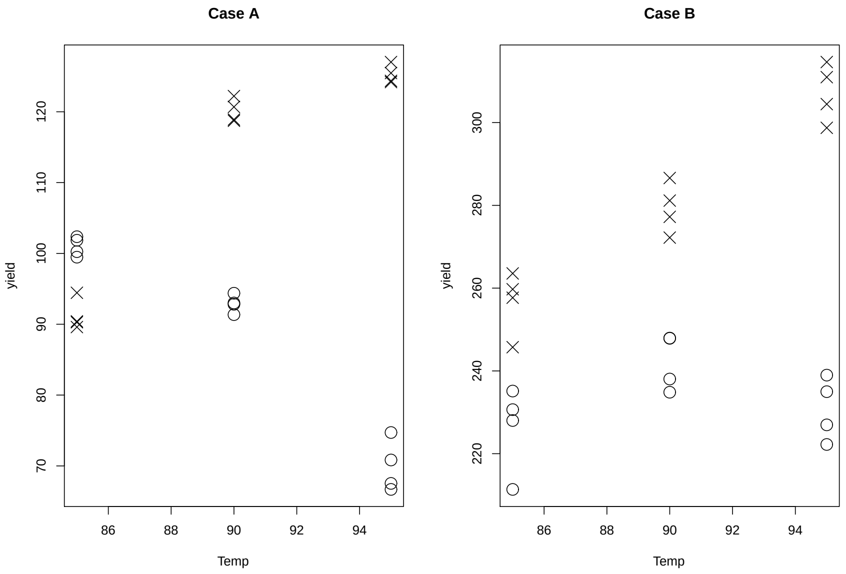

Figure 2: The datasets for question X.

   (a) Does either A or B, or both, show clear interaction between the two factors? If so, why? (1 point)

   (b) What kind of analysis method(s) can one use for such data? (1 point)

   (c) What kind of assumptions does the analysis method you mentioned in (b) require you to make? (1 point)

   (d) Does the variability of the yield at each combination of the factors play a role in determining whether or not there is an interaction between the factors? If so, how? (1 point)

6. Randi is trying to minimize the time she takes to ride her bike to work each day. For eight consecutive days, she rides as fast as she can, and records the times. Then, she figures that she might be able to improve her times by changing to a different bike with more gears,

and also by wearing more comfortable gym clothes. She makes these changes, and after recording her times the next eight days, the times are indeed improving.

(a) Is the experiment Randi has done appropriate to learn about the effect of changing the bike or the effect of changing her clothes? Explain about any problems there might be. Give the name for what happens with the two variables in this experimental design. (2 points)

(b) Assuming that Randi would indeed want to know the effect of her variables, propose a better experimental plan for her. (1 point)

7. Sonya is a forensic investigator and she is trying to find out if person X was at a crime scene or not. Based on initial evidence, she estimates that there is a 50% chance that X was at the crime scene (and thus 50% chance that he was not). Sonya has now discovered some traces of glass on X's clothes that exactly match the glass type of a broken window at the crime scene. She estimates that if X were at the crime scene, the probability that he would have this glass trace on his clothes is 80%, while if X were not at the crime scene, and the glass trace comes from some other source, the probability of the observation is only only 0.2%, as the glass type is fairly rare. Given this information, what is the probability that X was at the crime scene? (3 points)

8. Two types of soil are compared for growing potato plants. Seven potatoes are planted in soil A and seven in soil B. After harvesting, the weight of potatoes produced by each of the 14 plants is measured, with the results given below:

Soil A	176, 156, 114, 161, 144, 196, 142
Soil B	199, 166, 231, 171, 201, 250, 206

The mean and sample variance of the 7 observations for soil A is 155.6 and 688.0, respectively. The mean and sample variance of the 7 observations for soil B is 203.4 and 902.3, respectively.

(a) Compute a 95% confidence interval for the added weight of potatoes when switching from soil A to soil B: In addition to the result, describe in detail which assumptions you make in order to arrive at the result. If there are some choices regarding the assumptions, mention why you choose as you do. (3 points)

(b) If there are reasons why one should not make the assumption that the weight of new potatoes produced by a potato plant is normally distributed, is there some way one can compare how well potatoes grow in the two soils, without making distributional assumptions? Explain; you do not need to make computations. (1 point)

9. Harri is studying the bacterial genomics in people's guts, and he has found that a person's gut bacterial system can be classified into 5 different types: A, B, C, D, or E. He has also found the prevalences of 34%, 22% 18% 17% and 9% of each of these types, respectively, in the Swedish population. In a study in Japan, he investigated 40 people and found the following gut bacterial systems in these people:

A	B	C	D	E
9	8	9	6	8

(a) Do a goodness-of-fit test of whether Japanese people tend to have the same prevalences of gut bacterial systems as Swedes. State your conclusions. (2 points)

(b) Assume Harri had investigated 400 Japanese instead of 40 Japanese. Would the added number of investigated persons increase or decrease the probability that Harri would reject the hypothesis that Japanese people tend to have the same prevalences of gut bacterial systems as Swedes? Explain. (1 point)

10. In a statistics class, each student does a project investigating paper plane designs. Each student chooses four different folding desings, A, B, C, and D, and three different paper types, X, Y, and Z. For each combination of design and paper type, the student makes 5 paper airplanes. He or she then files each plane 10 times and records the average flight distance for each plane, thus recording a total of 60 average flight distances. Astrid has started to fill out an ANOVA table below, based on her data:

	Sum of Squares	Deg. Freedom	Mean squ.	F	p
Plane design	45.77				
Paper type	16.51				
Residuals					
Total	283.50				

Fill in the missing values in the table. Then, formulate the conclusions that Astrid should draw from her experiment. (3 points)