Petter Mostad
Mathematical Statistics
Chalmers

MSA830 Statistical analysis and experimental design

Exam 16 January 2013, 8:30 - 13:30
Examiner: Petter Mostad, phone 0707163235,
visits the exam at 9.30 and at 12.30.
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. Anna is working for an organization that helps students at a school with math questions. Individual students at the school make appointments with her and meet her when she visits the school. At the school, 57% of the students are female.

   (a) Anna would like to calculate the probability that out of the students at the next 10 appointments, exactly 8 will be female. Describe a set of assumptions that would enable her to compute such a probability. (1 point)

   (b) Making the assumptions from (a), compute the probability that out of the students at Annas next 10 appointments, exactly 8 will be female. (1 point)

   (c) Making the same assumptions, what is the approximate probability that during all of Annas 132 appointments at the school, 89 or more of the students contacting here are female, while 43 are male? (1 point)

   (d) In fact, Anna had 132 appointments at the school, with 89 female and 43 male students. What conclusions could be reasonable from this, based on what you have done in parts (a), (b), and (c) above? (1 point)

2. Anton would like to investigate whether and how a large number of factors might influence the quality of the new measurement process he is developing as part of his PhD project. He decides to test each factor at two plausible levels, and do a number of experiments with each factor at one of two possible levels. For simplicity he calls the factors A, B, C, D, ... and the levels "+" and "-".

   (a) If he is prepared to do 8 experimental runs, what is the maximal number of factors he can vary and still get some information about the main effect of each? (1 point)

   (b) Set up a fractional factorial experimental plan describing the levels of factors for 8 experimental runs, where the number of factors is the maximal you found in (a). (1 point)

   (c) A possible experimental plan in this situation would be to decide on a "base case" level for all factors, make one experimental run with this base case, and then make 7 more experimental runs, where in each only one of the factors has a different level compared to the base case. Describe in your own words how learning from the data from such an experiment would be different from learning from the experimental plan of (b). (1 point)

   (d) If Anton is prepared to do 16 experimental runs, what is the maximal number of factors he can vary and still get some information about the main effect of each? (1 point)

3. Jatta is studying the differences in the percentage of a particular element between three rock types, X, Y, and Z. For each rock type, she has 6 independently selected samples, and thus 6 measurements of the percentage of the element. The data is listed in the table below:

	Rocktype X	Rocktype Y	Rocktype Z
	16.44	19.60	18.16
	15.65	19.75	20.92
	16.06	16.35	20.60
	17.29	14.87	19.24
	17.29	14.69	19.81
	15.10	19.36	18.52
Mean	16.305	17.437	19.542
Sample variance	0.77995	5.8087	1.2271

   (a) Assuming that the data for rock type X comes from a normal distribution, find a 95% credibility interval for the expected value of this distribution. (1 point)

   (b) Assuming that the data for rock type X comes from a normal distribution, find a 95% credibility interval for the standard deviation of this distribution. (2 points)

   (c) Assuming that the data for both rock types X and Y come from normal distributions, find a 95% credibility interval for the difference between the expected values of these distributions. (2 points)

   (d) Assuming that the data for both roch types X and Y come from normal distributions with the same distribution standard deviations, find a 95% credbility interval for the difference between the expected values of these distributions. (2 points)

   (e) Assuming that the data from all rock types X, Y, and Z come from normal distributions with the same distribution standard deviations, find a 95% credibility interval for the difference between the expected values for the distributions for rock types X and Y. (2 points)

   (f) Is there a way to say something about the differences in the percentages between the rocks of type X and Y without making assumptions about normal distributions? Explain (you do not have to compute). (1 point)

4. Miro is analyzing data for a large on-line movie company. Customers are categorized into three types, A, B, and C, based on what kind of movies they prefer. Based on previous data, Miro estimates that a brand new customer is of type A, B, or C with probabilities 10%, 37%, and 53%, respectively. He is then using the customer's scoring of movies to update the information about what type the customer is. For example, for the new movie "M", Miro estimates from previous data that 59% of customers of type A give it top scores, 20% of customers of type B give it top scores, and 3% of customers of type C give it top scores. Also, 6% of customers of type A give the movie "M" bottom scores, 44% of customers of type B give it bottom scores, and 18% of customers of type C give it bottom scores.

   (a) Given that a brand new customer gives "M" a top score, what is the probability that the new customer is of type A? (2 points)

   (b) Given that the new customer instead gives "M" the bottom score, what is the probability that the new customer is of type A? (2 points)

5. Lurleen is a materials scientist, and is trying to optimize the strength of a new material she has invented. She is considering two factors, A and B, and how they influence the strength. In an experiment, she tries out three levels (x, y, z) of factor A, and three levels (r, s, t) of factor B. For each combination of these levels, she does three independent experimental runs, producing a total of 27 strength measurements. Averages for different combinations of the levels are given in the table below. The variance of the strength measurements was 54.0114.

	x	y	z	Average
r	40.67	50.00	52.00	47.56
s	39.00	53.67	53.00	48.56
t	35.33	48.00	45.67	43.00
Average	38.33	50.56	50.22	46.37

(a) Set up two ANOVA tables for this situation: One that includes the effects of factors A and B, but not interaction effects, and one that also includes interaction. The tables should include intervals for the relevant p-values. (4 points)

(b) Use the two ANOVA tables to draw conclusions, and describe these conclusions. Should interaction be included in the model? Do each of the factors influence the material strength? What kind of combinations for the factors would you recommend for the maximum strength? (2 points)

(c) In order to draw conclusions in the way above, some assumptions about the data need to be made. Describe these assumptions, and describe how one can check in this situation whether the assumptions are reasonable for the data. (1 point)

(d) It turns out that the experiment was performed in the following way: Lurleen and her two friends Selma and Patty decided to divide the work between them so that Lurleen, Selma, and Patty did experiments with factor A at level x, y, and z, respectively. They also decided that they should all do experiments with factor B at level r the first day, at level t the second day, and at level s the third day. Are there any problems with this experimental procedure? If you think so, describe the problem in detail, why it is a problem, and what might have been done instead. (1 point)