Petter Mostad
Mathematical Statistics
Chalmers

MSA830 Statistical analysis and experimental design

Exam 11 March 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. Luigi is wondering whether he is carrying a particular disease gene X. His father has the gene (but not his mother) so based on this, there is a 50% chance that Luigi is carrying the gene. Among those with the disease gene, 84% have type-2 diabetes, whereas 3% of those not carrying the gene has type-2 diabetes. Based on the fact that Luigi does not have type-2 diabetes, what is the probability that he is carrying the disease gene X? (2 points)

2. Alvar wants to compare the amount a pollutant in several Swedish cities. He starts with obtaining the 5 independent measurements 21.9, 22.2, 21.6, 22.3, 21.5 in city A, and the 5 independent measurements 21.7, 22.8, 22.4, 22.2, 23.1, in city B.

   (a) Alvar would like to analyse the data assuming that the observations from each city come from normal distributions. He would like to do a hypothesis test of whether the two normal distributions could have the same precision. Make the computations and find an interval for the p-value. (2 points)

   (b) Alvar decides to choose a model for analysis of his data based on the p-value of the test in (a). Find a 95% credibility interval for the difference between the expectations of the normal distributions. (3 points)

   (c) Alvar goes on to collect more data from three more cities. He now assumes that data from all cities come from distributions with the same precision. In each city, he collects 5 independent observations. In city C, he gets mean 20.66 and variance 0.128. In city D he gets mean 20.26 and variance 0.008. In city E he gets mean 22.98 and variance 0.077. Using this additional information, re-compute the credibility interval for the difference in between the expectations of the normal distributions for city A and B. (2 points)

3. Elisabeth is interviewing 40 randomly selected persons in some town about a current issue. She also asks about how they voted in the last election. Let us assume the results of the last election can be described in the following table

	Party A	Party B	Party C	Party D	Party E
Results	45.1%	29.9%	10.8%	7.2%	7.0%

   (a) What is the probability that exactly 2 of the persons Elisabeth interviews voted for Party E? (1 point)

   (b) What is the probability that exactly 3 of the persons Elisabeth interviews voted for either C, D, or E? (1 point)

(c) What is the approximate probability that 30 or more voted for party A? (2 points)

4. Eric is planning an experiment where he is trying to produce a new deodorant with the best possible smell. He is going to vary the amount of three ingredients, A, B, and C, and the amount of each will be tried at two different levels. He would like to use a full factorial design, and at each combination of levels, he plans two replications. For each single experimental run, he will produce a deodorant stick which will then be smelled by a test panel, which will give a score for the pleasentness of its smell. (The same test panel will smell all the deodorants).

   (a) Make an experimental plan for Eric. Include a list giving the details of all the experimental runs. Include also possible advice about how the experiment should be done. If you recommend to use randomization in any part of the experiment, explain in detail what randomization could mean in this case. (3 points)

   (b) After he has done the experiment, he would like to analyze the data, using a linear model that includes all possible interactions. Write down the design matrix Eric should use. (2 points)

   (c) As part of his analysis, Eric plots the residuals. Among other things, he plots in one column all the residuals when the amount of ingredient A is low, and in a column next to it all the residuals when the amount of ingredient A is high. Is it possible for such a plot to reveal problems with Eric's model? If so, what would the plot look like, if it indicated problems? (1 point)

5. Susan is working at a factory that has 5 machines producing a particular material. Susan has completed an investigation into how the strength of the material depends on the amount of an ingredient X. For each machine, Susan has made 2 experimental runs with low amounts of X and two with high amounts of X. The results are given in the table below. For your convenience, values for the averages for each machine, the averages for each level of X, and the grand average are also given. The variance of the data is 881.1026[1].

	Low X	High X	Average
Machine 1	336 362	354 387	359.75
Machine 2	337 392	420 369	379.5
Machine 3	364 336	361 387	362
Machine 4	273 349	353 350	331.25
Machine 5	357 335	371 338	350.25
Average	344.1	369	356.55

   (a) Make a complete ANOVA table for Susan's data. Do not include interaction. Answer practical questions, such as: Is there a significant difference between the strength when X is high compared to when X is low? Would you recommend high X or low

[1] In the original exam, the variance was erroneously given as 801.3132

X if the goal is to produce a strong material? Is there a significant difference between the strengths of the materials produced by different machines? (4 points)

(b) Assume Susan had analysed the same data disregarding the machine factor. Make a new ANOVA table for this analysis, and answer again the question of whether there is a significant difference beetween the strength when X is high compared to when X is low. (2 points)

6. Consider the following fractional factorial experimental plan:

A	B	C	D	E	F
-	-	-	-	+	+
-	-	-	+	+	-
-	-	+	-	+	-
-	-	+	+	+	+
-	+	-	-	-	-
-	+	-	+	-	+
-	+	+	-	-	+
-	+	+	+	-	-
+	-	-	-	-	-
+	-	-	+	-	+
+	-	+	-	-	+
+	-	+	+	-	-
+	+	-	-	+	+
+	+	-	+	+	-
+	+	+	-	+	-
+	+	+	+	+	+

(a) Write down the name of the plan, in the form $2^{\cdots}$, or, ideally, in the form $2^{\cdots}_{*}$ so that you also specify the resolution. (1 point)

(b) Edith would like to study the effect of the following 6 factors on the egg production from her chicken.

	Low level	High level
Food type	type A	type B
Space per chicken	low	high
Type of compartments	small	large
Temperature in room	low	high
Light in room	low	high
Noise reduction in room	yes	no

In partuclar, she is interested in the possible interaction between space and type of compartments.

She would like to use the experimental design given above. To use it, she needs to assign each of the letters A, B, C, D, E, F to each of the factors listed. Give one example of an assignment of letters to factors that is NOT recommendable to Edith, considering the goals of her study. (1 point)

7. Which of the following statements are true and which are false? Explain for each of them. (2 points)

(a) If you take the sample variance of a large sample from a Gamma distribution, that variance will be approximately equal to the variance of the Gamma distribution.

(b) If you take the average of a large sample from a Gamma distribution, that average will be approximately Gamma distributed.

(c) If X is sampled from a Gamma distribution and Y is sampled from another Gamma distribution, then $X - Y$ will be (approximately) Gamma distributed.

(d) If X is sampled from a Normal distribution and Y is sampled from another Normal distribution, then $X - Y$ will be (approximately) Normally distributed.

8. The plot below shows three observations of bivariate data (x, y). The observations are represented by the points $(1, 1)$, $(1, 2)$, and $(2, 2)$. Three lines are also given, named A, B, and C in the plot. Which one of these represents the linear regression line when x is the predictor and y is the response? Why? (1 point)