Exam in MSA830 Statistical Analysis and Experimental design

December 5th 2007, 8:00 – 12:00, at Väg och Vatten
Jour: Petter Mostad (phone 0707163235), who will be available for questions about the formulations of the exam questions at 9:00 and 11:00.
Allowed during the exam: An optional calculator, and one single page of your own notes.
For a grade G you need 12 points and for a grade VG you will need 24 points, out of a total of 30 points.

1. Lisa is investigating the growth of a certain plant under two different growing conditions, A and B. She has made 5 experiments with growing conditions A and 7 with growing conditions B, and has obtained the following output results:

   A: 2, 11, 7, 12, 13
   B: 21, 25, 3, 4, 24, 19, 23

   Some computations that might be useful:
   $(2+11+7+12+13)/5=9$
   $(21+25+3+4+24+19+23)/7=17$
   $(2-9)^2+(11-9)^2+(7-9)^2+(12-9)^2+(13-9)^2=82$
   $(21-17)^2+(25-17)^2+(3-17)^2+(4-17)^2+(24-17)^2+(19-17)^2+(23-17)^2=534$

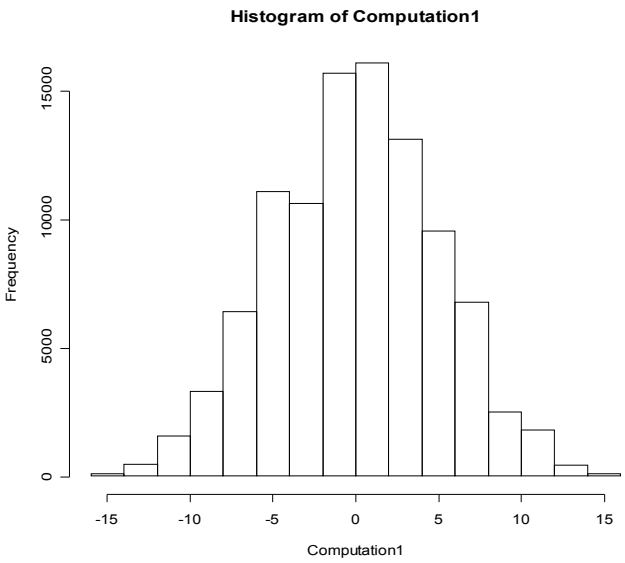
   a) Lisa would like to test whether the population mean output for growing condition B is higher than the population mean output for growing condition A, against the null hypothesis that the population means are equal. What kind of assumptions does she need to make in order to do a t-test? (2 points)

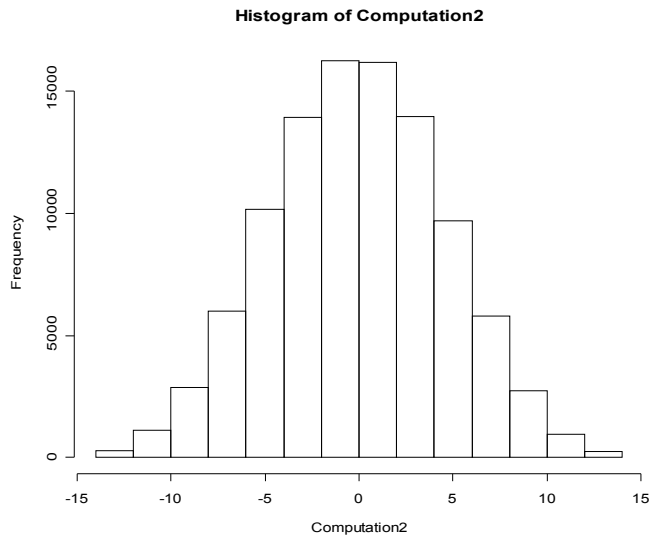
   b) Make the assumptions above, and assume also equal variance for the two conditions. Perform a t-test. Find an approximate p-value under these assumptions. What would you conclude from the test? (3 points)

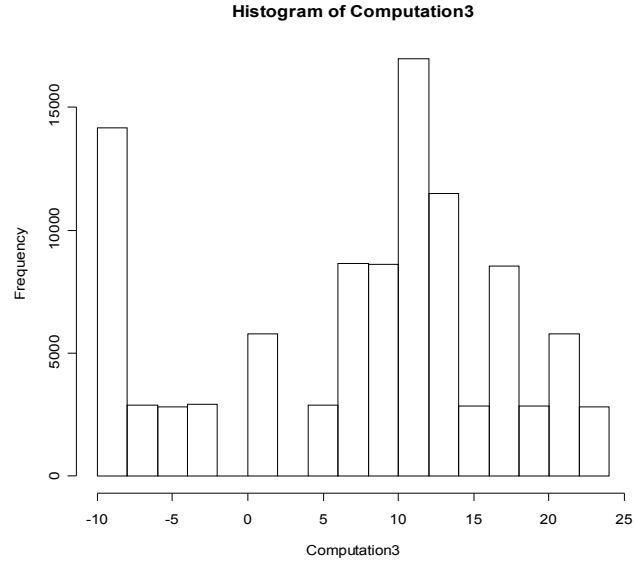
   c) Lisa would like to re-analyse the data using a randomization test. She was not quite sure what to compute, so she made the three different computations below. Which one is the correct one to use in a randomization test, and why? (2 points)

      1. In the first computation, she does the following 100000 times: She selects randomly 7 data values from the 12 data values above, takes their mean, and subtracts the mean of the remaining 5 data values. The results are shown in the first histogram and in the table below.
      2. In the second computation, she does the following 100000 times: She multiplies each of the 12 data values with either +1 or -1, chosen randomly with equal probability, and then takes the mean. The results are shown in the second histogram and in the table below.
      3. In the third computation, she does the following 100000 times: She randomly selects one of the data values for B, and subtracts a randomly selected data value from A. The results are shown in the third histogram and in the table below.

   d) What is the result of the correct randomization test? What is the approximate p-value? (2 points)


Histogram of Computation1
Frequency
0
5000
10000
15000
-15
-10
-5
0
5
10
15
Computation1


Histogram of Computation2
Frequency
0
5000
10000
15000
-15
-10
-5
0
5
10
15
Computation2


Histogram of Computation3
Frequency
0
5000
10000
15000
-10
-5
0
5
10
15
20
25
Computation3

Threshold	***Frequency of simulations below threshold for computation 1***	***Frequency of simulations below threshold for computation 2***	***Frequency of simulations below threshold for computation3***
-12	0.00602	0.00249	0
-11	0.01219	0.00596	0
-10	0.02241	0.01158	0
-9	0.03476	0.02227	0.02917
-8	0.05631	0.03787	0.08486
-7	0.08431	0.06293	0.14184
-6	0.12101	0.09467	0.17065
-5	0.17425	0.13977	0.17065
-4	0.21219	0.19163	0.17065
-3	0.27366	0.25651	0.19888
-2	0.33696	0.32835	0.22807
-1	0.4125	0.40766	0.22807
0	0.4927	0.49051	0.22807
1	0.57368	0.57289	0.22807
2	0.65193	0.65273	0.25686
3	0.72043	0.72706	0.28547
4	0.78367	0.79524	0.28547
5	0.8364	0.85029	0.28547
6	0.8802	0.8963	0.28547
7	0.91972	0.93069	0.31434
8	0.94032	0.95727	0.34297
9	0.96425	0.97423	0.40086
10	0.97559	0.98701	0.43026
11	0.98721	0.99308	0.48713
12	0.99343	0.99739	0.54301

2. Johan is selling hot-dogs from a mobile hot-dog stand, and he wants to investigate at which one of two positions, P1 and P2, he sells the most hot-dogs. He suspects that the number of hot-dogs sold is also influenced by whether he has a large advertising board on top of his stand, and by whether he plays music[1] or not at his stand.

   a) Johan has the possibility to try out different combinations of these sales techniques over 8 weeks one summer. Give a detailed plan for how he might perform his experiment. (2 points)

   b) Explain how Johan could compute the main effect from the position, the main effect from the music, and the interaction effects between the two effects. (2 points)

   c) Johan realizes that the sale might be influenced by whether he or his partner Susan is manning the stand. Propose an experimental plan where the main effects of the three factors mentioned above can be estimated independently of the effect of the salesperson. (2 points)

   d) Over the summer there might be some changes in the weather, which Johan suspects might influence his results. Mention one or two ways in which he could attempt to avoid that weather changes influence his results too much. (1 point)

3. Sonia wants to investigate whether 4 different types of winter tyres have an effect on the stopping distance for cars under winter conditions. She believes that the stopping distance for cars will also be influenced by the driver, and so she performs an experiment where 4 different drivers each test each of the 4 different types of tyres. The results are analyzed, and below is part of an ANOVA table with results:

Source of variation	***Sum of squares***	***Degrees of freedom***	***Mean square***
Tyres	15.3	3	5.1
Drivers	12.6	3	4.2
Residuals	9.9	9	1.1[2]
Total	37.8	15	

   a) Find an approximate p-value for the null hypothesis that the type of tyres has no effect on the stopping distance. What can you conclude? (2 points)

   b) Find an approximate p-value for the null hypothesis that the driver has no effect on the stopping distance. What can you conclude? (2 points)

   c) Sonia could also have analyzed the data by looking at it as 4 repeated experiments for each type of tyre, disregarding the information about the drivers. Look at the table above, and find out what an ANOVA table would look like in this case. What would you now get as an approximate p-value for the null hypothesis that the tyres have no effect on the stopping distance? (2 points)

1 In the original exam, there was a misprint here: It was written ”musing” instead of ”music”.
2 In the original exam, there was a misprint here: Instead of the correct 1.1 (=9.9/9), it was written 0.9.

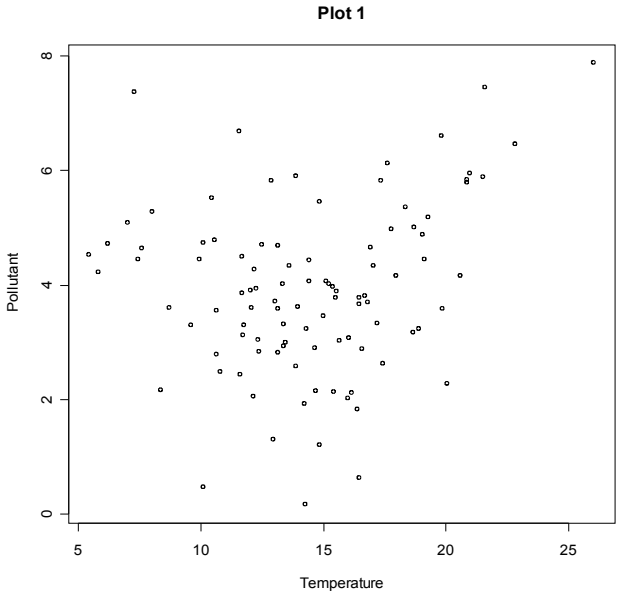
4. Reza is studying how the concentration y of a certain pollutant at a certain place in Göteborg depends on the temperature x_1 and the humidity x_2. He beileves the concentration can be modelled as $b_0 + b_1x_1 + b_2x_2$ plus some normally distributed random error. Over a period of 100 days, he makes 100 measurements of x_1, x_2, and y. He then analyzes the results using a regression program. The program outputs, among other things, -0.058 as an estimate for b_1, with standard error 0.035, and 0.130 as an estimate for b_2, with standard error 0.018.

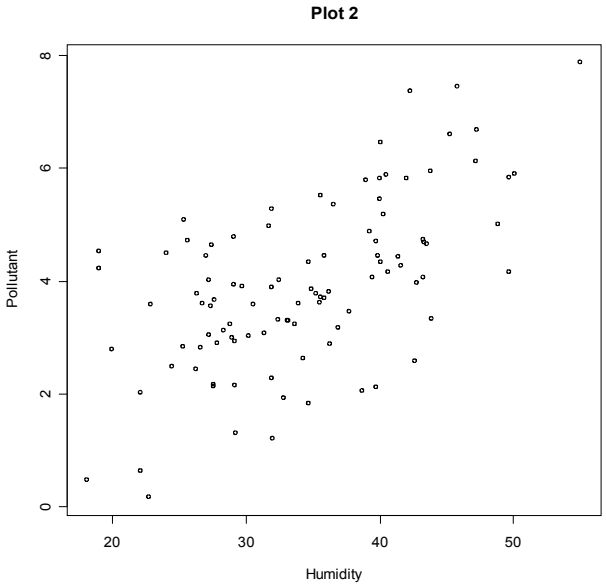
   a) Make a 95% confidence interval and a 99% confidence interval for each of b_1 and b_2. (3 points)

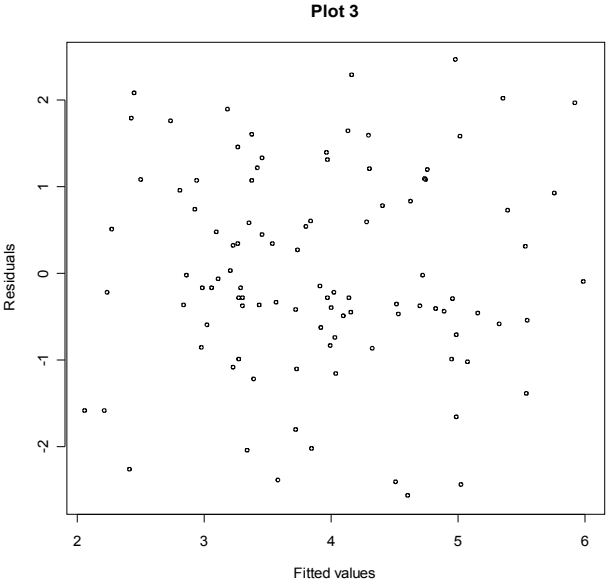
   b) What is meant by a confidence *region* for b_1 *and* b_2? What would be the shape of such a region? (1 point)

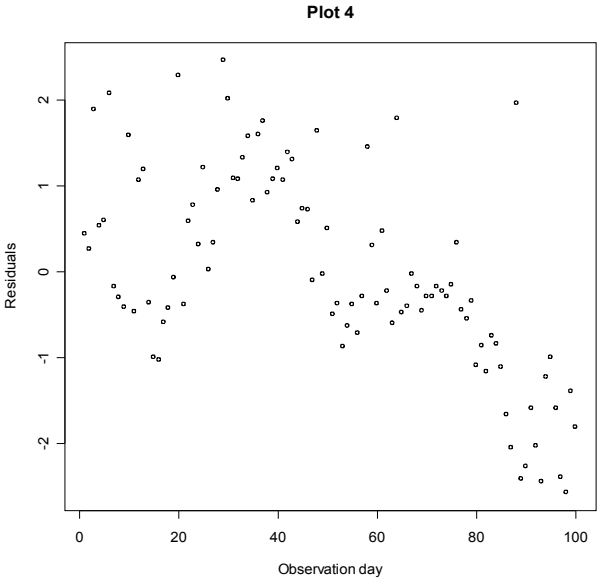
   c) Reza gives you the numbers above, and assures you that all the assumptions of a linear regression are fulfilled. What can you then conclude from his study, about the relationship between the pollutant and temperature and humidity? (1 points)

   d) Reza then gives you the following plots based on his data:
   1) A plot with the temperature on the x axis and the pollutant on the y axis
   2) A plot with the humidity on the x anxis and the pollutant on the y axis
   3) A plot of the predicted values from the regression on the x axis, and the residuals (actual observed values minus predicted values) on the y axis
   4) A plot with the day number on the x axis and the residual on the y axis.
   Do the plots show that there are some problems with the assumptions of linear regression, and if so, what are these problems? (3 points)


Plot 1
Pollutant
Temperature
0
2
4
6
8
5
10
15
20
25


Plot 2
Pollutant
Humidity
0
2
4
6
8
20
30
40
50


Plot 3
Residuals
Fitted values
-2
-1
0
1
2
2
3
4
5
6


Plot 4
Residuals
Observation day
-2
-1
0
1
2
0
20
40
60
80
100