Petter Mostad
Matematisk Statistik
Chalmers

Suggested solution for exam in
MSG830: Statistical Analysis and Experimental Design
30 May 2013

1. (a) Dataset A: Negative correlation. Dataset B: Negative correlation. Dataset C: Positive correlation. Dataset D: Positive correlation.

   (b) Dataset A: Simple linear regression is not appropriate, as the data cleary vary around a line that is not straight. In this case, a more complex regression model might be appropriate: Polynomial regression or some picewise-linear regression. Dataset B: Linear regression is appropriate. Dataset C: Linear regression is appropriate. Dataset D: Simple linear regression is not appropriate: Even though the observations vary around a straight line, the variability around the line is clearly larger for larger X than for smaller, violating an assumption of the simple linear regression model. A more complex model where the variance around the line also depends on X might be used.

2. Type I and Type II errors are concepts related to hypothesis testing: If the hypothesis is true, while the hypothesis test results in a rejection of the hypothesis, we have a type I error. If the hypothesis is false, while the hypothesis test does not result in a rejection of the hypothesis, we have a thype II error.

3. The number of selections of shirts can be computed with the Binomial coefficient:

$$\binom{12}{4} = \frac{12 \cdot 11 \cdot 10 \cdot 9}{1 \cdot 2 \cdot 3 \cdot 4} = 495$$

4. Principal Component Analysis is primarily a way to visualize high-dimensional data in 2 or 3 dimensions. It can also be used to determine how much variability in the data is lost if one considers only this 2 or 3 dimensional representation. More formally, the first principal component is the axis along which the variation of the data is largest. After this has been fixed, the second principal component is the orthogonal axis along which the remaining variation of the data is largest, and so on.

5. (a) Both A and B show clear interaction between the two factors. If there is no interaction, the line connecting the averages for the four observations at the three temperatures with low pressures should be roughly parallel to the line connecting the averages for the four observations at the three temperatures with high pressure. In this case, these lines are clearly not parallel, so there seems to be interaction.

   (b) The most natural would be two-way ANOVA.

(c) The assumption would be that the observations made with a particular combination of the two factors are independently normally distributed, where the normal distribution has the same variance for all combinations of the two factors.

(d) The variability does play a role: The lines mentioned in (a) need to be roughly parallel for there to be no interaction. How far away from exactly parallel do they need to be in order to indicate interaction? This is determined by the variability at each combination of factors: The smaller this variability is, the smaller the deviation from parallel needs to be in order to indicate interaction.

6. (a) There are two major problem with Randi's experiment: First, her two factors "bike type" and "clothes" are *confounded*: In other words, if any of them have an effect on her respons variable time, it is impossible from her experiment to tell which one has the effect, and how much of the total effect is due to each factor. Secondly, both her factors "bike type" and "clothes" are confounded with time: It seems natural that she gets better and better times each day, as her condition improves. When all observations with the new bike type and with more comfortable clothes are made at the end of her training period, she cannot tell whether improvements in time are due to her improved general condition, or to the changes in the factors.

(b) As Randi is investigating two factors with two levels each, she has 4 different settings of these variables. As her experiment goes over 16 days, she should try out each combination of the factor levels for 4 days. Moreover, to separate their effect from any time effect, she should randomize the order in which she does these 16 experimental runs.

7. Bayes formula gives

$$\begin{aligned}
&\Pr(\text{X on crime scene} \mid \text{glass type}) \\
&= \frac{\Pr(\text{glass type} \mid \text{X on crime scene})\Pr(\text{X on crime scene})}{\Pr(\text{glass type})} \\
&= \frac{\Pr(\text{glass type} \mid \text{X on crime scene})\Pr(\text{X on crime scene})}{\Pr(\text{glass type} \mid \text{X on cs})\Pr(\text{X on cs}) + \Pr(\text{glass type} \mid \text{X not on cs})\Pr(\text{X not on cs})} \\
&= \frac{0.8 \cdot 0.5}{0.8 \cdot 0.5 + 0.002 \cdot 0.5} = 0.9975
\end{aligned}$$

so the probability that X was on the crime scene is now 99.75%.

8. (a) It is natural to assume in this context that the observations for each soil are normally distributed: We assume that the values for soil A are normally distributed with expectation μ_A and distribution variance σ_A^2, while the values for soil B are normally distributed with expectation μ_B and distribution variance σ_B^2. We may assume that $\sigma_A^2 = \sigma_B^2$, or we may not. In this case, both choices are reasonable: That the observed sample variances 688.0 and 902.3 are fairly similar means that it is reasonable to assume that the distribution variances σ_A^2 and σ_B^2 are equal. A general argument against this assumption is that it seems better to minimize the assumptions you make.

If one assumes that $\sigma_A^2 = \sigma_B^2$, we can compute

$$s_p^2 = \frac{(7-1)\cdot 688.0 + (7-1)\cdot 902.3}{7-1+7-1} = 795.15$$

and the 95% confidence interval can be computed as

$$\begin{aligned}
&\left[203.4 - 155.6 - t_{14-2,0.025}\sqrt{795.15}\sqrt{\frac{1}{7}+\frac{1}{7}},\right.\\
&\left.203.4 - 155.6 + t_{14-2,0.025}\sqrt{795.15}\sqrt{\frac{1}{7}+\frac{1}{7}}\right]\\
=\ & [203.4 - 155.6 - 2.1788\cdot 15.07268, 203.4 - 155.6 + 2.1788\cdot 15.07268]\\
=\ & [15.0, 80.6]
\end{aligned}$$

If one assumes that σ_A^2 and σ_B^2 are estimated separately from data, we compute

$$\nu = \frac{\left(\frac{s_A^2}{7}+\frac{s_B^2}{7}\right)^2}{\frac{(s_A^2/7)^2}{7-1}+\frac{(s_B^2/7)^2}{7-1}} = \frac{\left(\frac{688}{7}+\frac{902.3}{7}\right)^2}{\frac{(688/7)^2}{7-1}+\frac{(902.3/7)^2}{7-1}} = 11.786$$

and the 95% confidence interval can be computed as

$$\begin{aligned}
&\left[203.4 - 155.6 - t_{12,0.025}\sqrt{\frac{688}{7}+\frac{902.3}{7}},\right.\\
&\left.203.4 - 155.6 + t_{12,0.025}\sqrt{\frac{688}{7}+\frac{902.3}{7}}\right]\\
=\ & [203.4 - 155.6 - 2.1788\cdot 15.07268, 203.4 - 155.6 + 2.1788\cdot 15.07268]\\
=\ & [15.0, 80.6]
\end{aligned}$$

(b) It is possible to use a non-parametric test, specifically the Mann-Whitney test. In it, the order of the observations from the different soils in the total ordering of all observations are used to test whether it is reasonable that the observations come from the same distribution.

9. (a) The expected counts become $40\cdot 0.34 = 13.6$, $40\cdot 0.22 = 8.8$, $40\cdot 0.18 = 7.2$, $40\cdot 0.17 = 6.8$, and $40\cdot 0.09 = 3.6$ for types A, B, C, D, and E, respectively, so the test statistic becomes

$$\chi^2 = \frac{(9-13.6)^2}{13.6} + \frac{(8-8.8)^2}{8.8} + \frac{(9-7.2)^2}{7.2} + \frac{(6-6.8)^2}{6.8} + \frac{(8-3.6)^2}{3.6} = 7.550505$$

Comparing this with a χ^2 distribution with 4 degrees of freedom, we find that the p-value is above 0.05. The conclusion is that Harri has not found in his study a significant difference between the prevalences of gut bacterial systems for people in Japan and people in Sweden.

(b) The added number would increase the probability of rejecting the null hypothesis of the test: The power of the test would increase with increased number of persons studied.

10. The complete ANOVA table becomes

	Sum of Squares	Deg. Freedom	Mean Squ.	F	p
Plane design	45.77	3	15.25667	3.7242	$0.01 < p < 0.025$
Paper type	16.51	2	8.255	2.0151	$0.1 < p < 0.25$
Residuals	221.22	54	4.096667		
Total	283.50	59			

From this, we can conclude that the plane design seems to have a significant influence on the flight distance, whereas we cannot conclude that the paper type has such a significant influence.