Petter Mostad
Matematisk Statistik
Chalmers

Suggested solution for exam in
MSA830: Statistical Analysis and Experimental Design
12 January 2011 (updated notation)

1. (a) The observed values for route A has mean 72.17 and variance 19.77. The observed values for route B has mean 59.33 and variance 177.47. As there are 6 observations for each route, the pooled variance becomes

$$s_p^2 = \frac{5 \cdot 19.77 + 5 \cdot 177.47}{5+5} = 98.62$$

Under the assumptions mentioned in the question, the difference between the expected values for the two routes is distributed as

$$\mathrm{t}\left(72.17 - 59.33, 6+6-2, \log\left(\sqrt{(1/6+1/6)s_p^2}\right)\right) = \mathrm{t}(12.84, 10, \log(5.733))$$

A 95% credibility interval for this distribution is given by

$$[12.84 - 2.228 \cdot 5.773, 12.84 + 2.228 \cdot 5.773] = [0.07, 25.61]$$

(b) As the 6 observations for route A have variance 19.77, the distribution for the standard deviation of the population of times using route A, under the assumptions given, is

$$\mathrm{ExpGamma}\left(\frac{5}{2}, \frac{5 \cdot 19.77}{2}, -2\right) = \mathrm{ExpGamma}\left(\frac{5}{2}, \frac{98.85}{2}, -2\right)$$

A 90% credibility interval for the standard deviation in question is then

$$\left[\sqrt{\frac{98.85}{11.07}}, \sqrt{\frac{98.85}{1.145}}\right] = [2.988, 9.291]$$

As the variance is the square of the standard deviation, we get that a credibility interval for it is

$$[2.988^2, 9.291^2] \approx [8.9, 86.3]$$

(c) With the variance of the data for the A and B routes computed to 19.77 and 177.47, respectively, we get the test statistic

$$\frac{177.47}{19.77} = 8.98$$

which should be compared with an F statistic with 5 and 5 degrees of freedom. From the relevant tables, we find that such a distribution has a probability between 0.025 and 0.01 of being above 8.98. Thus the p-value for the test becomes twice this, i.e., the p-value is between 0.5 and 0.2. As the p-value is smaller than 0.05, we should *reject* the null hypothesis that the two population variances are equal, and in fact we should use a computation where they are not assumed to be equal.

(d) Under these assumptions, the distribution for the expected time difference between the two routes is given by

$$\mathrm{t}\left(72 - 17 - 59.33, \nu, \log\left(\sqrt{19.77/6 + 177.47/6}\right)\right) = \mathrm{t}(12.84, \nu, \log(5.733))$$

where

$$\nu = \frac{\left(\frac{19.77}{6} + \frac{177.47}{6}\right)^2}{\frac{(19.77/6)^2}{5} + \frac{(177.47/6)^2}{5}} = 6.1$$

A 95% credibility interval for this distribution is given by

$$[12.84 - 2.447 \cdot 5.733, 12.84 + 2.447 \cdot 5.733] = [-1.14, 26.81]$$

(e) An important condition for the computations above to hold is that both data sets are *random samples* from the distribution of possible travelling times. In other words, each observation should be independent of all other observations. If for example Alex did all observations for route A during one week, and then all observations for route B during the next week, it would be unreasonable to assume independence. For example, special road work, or a special holiday, might influence the traffic in the whole city one of the weeks but not the other. If on the other hand Alex had chosen 12 days, spread over time throughout the whole year, and for each of these days he randomly chose which of the two routes to follow, it would make the assumption of independence more reasonable, and would strengthen the belief in the scientific reproducibility of his results.

2. (a) The sum of squares for location:

$$SS_{\text{Location}} = 6\left((23.17 - 26)^2 + (24.17 - 26)^2 + (30.67 - 26)^2\right) = 199$$

The sum of squares for sex:

$$SS_{\text{Sex}} = 9\left((27 - 26)^2 + (25 - 26)^2\right) = 18$$

As the variance is 28.59, the total sum of squares is

$$SS_{\text{Total}} = 17 \cdot 28.59 = 486.03$$

To compute the two remaining sums of squares, one can either compute the sum of suqares of the residuals from teh data and the cell averages, and then compute the sum of squares for the interaction by subtraction. One gets

$$SS_{\text{Residuals}} = (24 - 24)^2 + (23 - 24)^2 + \cdots + (26 - 32)^2 + (31 - 32)^2 = 198.67$$

and then

$$SS_{\text{Interaction}} = SS_{\text{Total}} - SS_{\text{Residuals}} - SS_{\text{Sex}} - SS_{\text{Location}} = 70.33$$

Another way is to first compute the sum of the squares of all effects including interaction:

$$SS_{\text{Location+Sex+Interaction}} = 3\left((24 - 26)^2 + (27.67 - 26)^2 + (29.33 - 26)^2 + (22.33 - 26)^2 + (20.67 - 26)^2 + (32 - 26)^2\right) = 287.33$$

then

$$SS_{\text{Interaction}} = 287.33 - 199 - 18 = 70.33$$

and then

$$SS_{\text{Residuals}} = SS_{\text{Total}} - SS_{\text{Interaction}} - SS_{\text{Sex}} - SS_{\text{Location}} = 198.67$$

The ANOVA table becomes

	SS	D.f.	M.sq.	F	p
Location	199	2	99.5	6.01	$0.01 < p < 0.025$
Sex	18	1	18	1.09	$p > 0.25$
Interaction	70.33	2	35.16	2.12	$0.1 < p < 0.25$
Residuals	198.67	12	16.56		
Total	486	17			

(b) As the p-value for the location is smaller than 0.025, it may be considered clear that wombats from different locations perform differently at the task. However, one cannot say that there is a clear effect of either the sex of the interaction.

(c) This ANOVA table can be obtained from the previous one by adding together the sums of squares and the degree of freedom form the Interaction and the Residuals of the previous table:

	SS	D.f.	M.sq.	F	p
Location	199	2	99.5	5.18	$0.01 < p < 0.025$
Sex	18	1	18	0.94	$p > 0.25$
Residuals	269	14	19.21		
Total	486	17			

3. Let us use the notation

A=M The animal is male.
A=F The animal is female.
S=L The size is “large” (above the limit)
S=S The size is “small” (below the limit)

Then we get that

$$\begin{aligned}
\pi(A = M \mid S = L) &= \frac{\pi(S = L \mid A = M)\pi(A = M)}{\pi(S = L)} \\
&= \frac{\pi(S = L \mid A = M)\pi(A = M)}{\pi(S = L \mid A = M)\pi(A = M) + \pi(S = L \mid A = F)\pi(A = F)} \\
&= \frac{0.4 \cdot 0.5}{0.4 \cdot 0.5 + 0.1 \cdot 0.5} \\
&= \frac{0.2}{0.25} = 0.8
\end{aligned}$$

The probability is 80% that the animal is male.

4. (a) The expected number is $54/20 = 2.7$ (including all of 1990 and all of 2009 in the counting).

(b) The probability can be computed with the Poisson distribution: The probability of two or fewer comets is

$$\begin{aligned}\pi(2)+\pi(1)+\pi(0) &= e^{-2.7}\frac{2.7^2}{2!}+e^{-2.7}\frac{2.7}{1}!+e^{-2.7}\frac{2.7^0}{0!}\\ &= e^{-2.7}\left(\frac{7.29}{2}+2.7+1\right)\\ &= 0.0672\cdot 7.345 = 0.49\end{aligned}$$

The probability is about 49%.

5. (a) After the completion of the experiment, an argument that is has been proved that B is better than A would be based on the x values in the B group being on average better than the x values in the A group. The argument would be stronger if the experiment is organized so that the only explanation is that B works better than A, i.e., that other explanations for the difference can be ruled out. For example, if the treatment for different patients is selected based on the patients medical history, an explanation for any difference could be in this history. If the treatment is selected based on time, an explanation for any difference could be that other factors influencing the patient got better (or worse) over time. So the best way to select the treatment would be to use randomization, i.e., to select based on some completely separate "random" process, but in such a way that (roughly) half of the patients get each treatment. If patients are informed about what medication they get, a placebo effect could explain an observed difference in the result: People might have a higher belief in the efficacy of the new drug, and this might effect their values. Also, if Lars knew which drug he administered, he could have been more enthusiastic in regards to the patients who got B, in a way that could have influenced the outcome.

(b) If there are n patients in the study, with $n/2$ getting each of the treatments, then, according to the assumptions of the study, the difference in the expected effects would have a distribution

$$\mathrm{t}\left(\overline{b}-\overline{a},\frac{n}{2}+\frac{n}{2}-2,\log\left(\sqrt{\left(\frac{1}{n/2}+\frac{1}{n/2}\right)s_p^2}\right)\right)=\mathrm{t}\left(\overline{b}-\overline{a},n-2,\log(2s_p/\sqrt{n})\right)$$

where $\overline{a}$ and $\overline{b}$ are the average observations for treatments A and B, respectively. Thus the length of the 95% credibility interval will be

$$2t_{0.025,n-2}2s_p/\sqrt{n}\approx 2\cdot 1.96\cdot 2s_p/\sqrt{n}\approx 7.84\cdot 2.1/\sqrt{n}=16.464/\sqrt{n}$$

because, when the number of degrees of freedom is large, the standard t distribution is very similar to the standard normal distribution. To get this interval smaller than 0.3, one would need

$$16.464/\sqrt{n}<0.3$$

which results in $n > (16.464/0.3)^2 = 3011$. With 100 patients per doctor, the company would need to enroll at least 31 doctors.

6. (a) She could use the design matrix

$$\begin{bmatrix} 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & 1 & 1 & -1 \\ 1 & 1 & 1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

(b) She could use the design matrix

$$\begin{bmatrix} 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & -1 & 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 & 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & -1 & 1 & 1 & -1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & -1 & -1 & -1 & 1 & 1 \\ 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & -1 & 1 & -1 & 1 & -1 & -1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & -1 & 1 & -1 & -1 & -1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

7. (a) There is a problem: if the model were appropriate for the data, the ϵ for each observation should be independent, so that the residuals should be approximately independent. Yet it is clear from the figure that the residuals depend on x_2: Small and large values of x_2 are related to small residuals, while medium values of x_2 are related to high residuals. (To obtain a more appropriate model, Lurleen might try a quadratic term in her regression).

(b) Intuitively, if the residuals did not sum to zero, one could get a better fitted value for the parameter β_0 by adjusting it: Adjusting this fitted value would change all the residuals with the same amount, and this could be done so that the sum of the squares of the residuals decreased. As the fitted values should have been found such that

the sum of the squares of the residuals was minimized, this is a contradition, so the residuals must sum to zero.

Mathematically, let $\widehat{\epsilon_1}, \widehat{\epsilon_2}, \ldots, \widehat{\epsilon_{30}}$ be the residuals, so that, for $i = 1, \ldots, 30$,

$$y_i = \widehat{\beta_1} + \widehat{\beta_2} x_{1i} + \widehat{\beta_3} x_{2i} + \widehat{\beta_4} x_{3i} + \widehat{\epsilon_i}$$

where $\widehat{\beta_1}, \widehat{\beta_2}, \widehat{\beta_3}, \widehat{\beta_4}$ are the fitted values, chosen so that

$$R = \widehat{\epsilon_1}^2 + \widehat{\epsilon_2}^2 + \cdots + \widehat{\epsilon_{30}}^2$$

is minimized. Let

$$S = \widehat{\epsilon_1} + \widehat{\epsilon_2} + \cdots + \widehat{\epsilon_{30}}.$$

Then we can write

$$y_i = \widehat{\beta_1} + \frac{S}{30} + \widehat{\beta_2} x_{1i} + \widehat{\beta_3} x_{2i} + \widehat{\beta_4} x_{3i} + \widehat{\epsilon_i} - \frac{S}{30}$$

and the sum of the squares of these residuals becomes

$$\sum_{i=1}^{30} \left(\widehat{\epsilon_i} - \frac{S}{30} \right)^2 = \sum_{i=1}^{30} \widehat{\epsilon_i}^2 - 2 \sum_{i=1}^{30} \widehat{\epsilon_i} \frac{S}{30} + 30 \left(\frac{S}{30} \right)^2 = R - \frac{S^2}{30}$$

If S is not zero, then the sum of the squares of these new residuals would be smaller than the sum of the squares of the old residuals, contradicting that the parameters were chosen to minimize the sum of the squares of the residuals.