Petter Mostad
Matematisk Statistik
Chalmers

Suggested solution for exam in
MSA830: Statistical Analysis and Experimental Design
16 March 2012

1. (a) One can use the Binomial distribution to compute this probability:

$$\binom{10}{5} \cdot 0.25^5 \cdot (1-0.25)^{10-5} = \frac{10 \cdot 9 \cdot 8 \cdot 7 \cdot 6}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} \cdot 0.25^5 \cdot 0.75^5 = 0.0583992$$

The probability is about 6%.

(b) Betty will buy a total of $52 \cdot 4 = 208$ tickets. We need to compute the probability that a variable with distribution Binomial(208, 0.25) has value 75 or more. We approximate this distribution with a normal distribution with expectation $208 \cdot 0.25 = 52$ and variance $208 \cdot 0.25 \cdot (1-0.25) = 39$: Comparing

$$\frac{74.5 - 52}{\sqrt{39}} = 3.602883$$

with the table for a normal distribution, we find the probability 0.00016, so this is the approximate answer.

(c) For each person, the probability of exactly one win is

$$\binom{4}{1} \cdot 0.25^1 \cdot (1-0.25)^{4-1} = \frac{4}{1} \cdot 0.25 \cdot 0.75^3 = 0.421875.$$

As it is independent whether or not each of the family members win, the probability that each of the four familiy members has won exactly one time is

$$0.421875^4 = 0.031676$$

The probability is about 3%.

2. Using the notation

A	:	Virus A is responsible
B	:	Virus B is responsible
C	:	Virus C is responsible
symptom	:	The cows show the particular symptom

we can write

$$\begin{aligned}
\pi(C \mid \text{symptom}) &= \frac{\pi(\text{symptom} \mid C)\pi(C)}{\pi(\text{symptom})} \\
&= \frac{\pi(\text{symptom} \mid C)\pi(C)}{\pi(\text{symptom} \mid A)\pi(A) + \pi(\text{symptom} \mid B)\pi(B) + \pi(\text{symptom} \mid C)\pi(C)} \\
&= \frac{0.9 \cdot 0.1}{0.1 \cdot 0.7 + 0.1 \cdot 0.2 + 0.9 \cdot 0.1} \\
&= 0.5
\end{aligned}$$

so the probability that variant C is responsible for the outbreak is 50%.

3. (a) For design A, we compute the average 75.5 and the sample variance 46.3, while for design B, we compute the average 89.16667 and the sample variance 211.3667. To make the hypothesis test mentioned, we compute

$$\frac{211.3667}{46.3} = 4.565$$

and compare it with an F distribution with 5 and 5 degrees of freedom. Consulting the relevant table, we find that the probability for a variable with this distribution to be above 4.565 is between 0.05 and 0.1, so the p-value for the test is between 0.1 and 0.2. This means that we do *not* reject the null hypothesis, that the two normal distributions have the same standard deviations.

(b) One possibility is the following: One assumes the data for design A and B come from two normal distributions. One may either assume that these two distributions have the same, unknown scale, or, one may assume that they have different scales. The last possibility is an equally valid choice, even if the p-value in (a) is above 0.05, as there could be other reasons to believe that the scales of the distributions were different. Finally, one needs to assume flat priors for these parameters.

(c) If one assumes that the two normal distribution have the same standard deviations, we would compute the pooled variance

$$s_p^2 = \frac{5 \cdot 46.3 + 5 \cdot 211.3667}{5 + 5} = 128.8333$$

and get that the difference mentioned in the question has distribution

$$\mathrm{t}\left(89.16667 - 75.5, 6 + 6 - 2, \log\sqrt{s_p^2(1/6 + 1/6)}\right) = \mathrm{t}\left(13.66667, 10, \log(6.5532)\right)$$

A 95% credibility interval is thus given as

$$[13.66667 - 2.2281 \cdot 6.5532, 13.66667 + 2.2281 \cdot 6.5532] \approx [-0.9, 28.3]$$

and a 90% credibility interval is given as

$$[13.66667 - 1.8125 \cdot 6.5532, 13.66667 + 1.8125 \cdot 6.5532] \approx [1.8, 25.6]$$

We may instead not assume that the two normal distributions have the same standard deviation. As we have variance $s_A^2 = 46.3$ and $n = 6$ observations for A, and variance $s_B^2 = 211.3667$ and $m = 6$ observations for B, we compute the degrees of freedom parameter ν as

$$\nu = \frac{\left(\frac{s_A^2}{n} + \frac{s_B^2}{m}\right)^2}{\frac{(s_A^2/n)^2}{n-1} + \frac{(s_B^2/m)^2}{m-1}} = \frac{\left(\frac{46.3}{6} + \frac{211.3667}{6}\right)^2}{\frac{(46.3/6)^2}{5} + \frac{(211.3667/6)^2}{5}} = 7.084709$$

so the difference mentioned in the question has (approximate) distribution

$$\mathrm{t}\left(89.16667 - 75.5, 7, \log\sqrt{46.3/6 + 211.3667/6}\right) = \mathrm{t}\left(13.66667, 7, \log(6.5532)\right)$$

A 95% credibility interval is thus given as

$$[13.66667 - 2.3646 \cdot 6.5532, 13.66667 + 2.3646 \cdot 6.5532] \approx [-1.8, 29.2]$$

and a 90% credibility interval is given as

$$[13.66667 - 1.8946 \cdot 6.5532, 13.66667 + 1.8946 \cdot 6.5532] \approx [1.3, 26.1]$$

(d) Irrespective of which model Billy chooses in (b), the main conclusion is the same: As the 95% credibility interval contains zero, he can *not* say that he has found a significant difference between the expected weight each design can take before it collapses. However, as the 90% credibility interval does not contain zero, he can still claim that he has fairly strong evidence that design B is stronger than design A.

4. (a) A possible experimental plan is given by

Band length	Band width	Elasticity	Stick material	Opening
-	-	-	+	-
-	-	+	+	+
-	+	-	-	+
-	+	+	-	-
+	-	-	-	+
+	-	+	-	-
+	+	-	+	-
+	+	+	+	+

Note that the column for “Stick material” has been produced by multiplying the columns for “Band length” and “Band width”, while the column for “Opening” has been produced by multiplying the columns for “Band length”, “Band width” and “Elasticity”. For the experimental design to fulfill Bart’s requirements, the important thing is that none of the columns are equal to the product of the “Band width” and “Elasticity” columns, and none of the columns are equal to the product of the “Band length” and “Elasticity” columns.

(b) The design matrix not taking into account interaction would become

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

The design when taking into account the two types of interactions that interest Bart would become

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

5. (a) We start with computing various sums of squares. We get

$$\begin{aligned} & SS_{\text{soap}} \\ = \; & 16 \cdot \left((31.625 - 35.0625)^2 + (39.1875 - 35.0625)^2 + (34.375 - 35.0625)^2 \right) \\ = \; & 468.875 \end{aligned}$$

and

$$\begin{aligned} & SS_{\text{design}} \\ = \; & 12 \cdot \left((35 - 35.0625)^2 + (32.3333 - 35.0625)^2 + (41.4167 - 35.0625)^2 + (31.5 - 35.0625)^2 \right) \\ = \; & 726.2364 \end{aligned}$$

and for the total sums of squares

$$SS_{\text{Total}} = 47 \cdot 163.5918 = 7688.815$$

We can now find $SS_{\text{Residuals}}$ by subtraction, and we get the ANOVA table

	SS	D.f.	M.sq.	F	p
Soap	468.875	2	234.4375	1.516296	$0.1 < p < 0.25$
Design	726.2364	3	242.0788	1.565718	$0.1 < p < 0.25$
Residuals	6493.704	42	154.612		
Total	7688.815	47			

As both the p-values are above 0.05, one would conclude that Alex has not found any significant effect of the soap factor or the design factor. This does not mean that there is no such effect, it just means that he has not found clear evidence of such an effect in his experiment, assuming a linear model without interaction for the analysis of the data.

(b) To compute the sum of squares for the interaction, we first compute the sum of squares for all the factors including interaction:

$$\begin{aligned} SS_{\text{All}} &= SS_{\text{Soap+Design+Interaction}} \\ &= 4 \cdot \left((29 - 35.0625)^2 + (40.75 - 35.0625)^2 + (35.25 - 35.0625)^2 \right. \\ &\quad + (30.5 - 35.0625)^2 + (29.75 - 35.0625)^2 + (36.75 - 35.0625)^2 \\ &\quad + (45 - 35.0625)^2 + (54.5 - 35.0625)^2 + (24.75 - 35.0625)^2 \\ &\quad \left. + (22 - 35.0625)^2 + (31.75 - 35.0625)^2 + (40.75 - 35.0625)^2 \right) \\ &= 3671.562 \end{aligned}$$

We then get

$$SS_{\text{Interaction}} = SS_{\text{All}} - SS_{\text{Soap}} - SS_{\text{Design}} = 3671.562 - 468.875 - 726.2364 = 2476.451$$

and

$$SS_{\text{Residuals}} = SS_{\text{Total}} - SS_{\text{All}} = 7688.815 - 3671.562 = 4017.253$$

and the ANOVA table including interaction becomes

	SS	D.f.	M.sq.	F	p
Soap	468.875	2	234.4375		
Design	726.2364	3	242.0788		
Interaction	2476.451	6	412.7418	3.6987	$p < 0.01$
Residuals	4017.253	36	111.5904		
Total	7688.815	47			

Thus the interaction is clearly significant, and Alex should continue his analyses with a model containing interaction.

(c) When Alex did his 48 experimental runs, there may have been other factors than the two factors he mentioned that influenced the outcome, for example changes in the air movements around the apparatus, changes in the temperature or humidity, changes in the amount of residual soap present in the machine, etc. Many such factors would be correlated with time. For example, Alex might gain more routine over time in how he set up each experiment. When Alex did his 48 experiments in the same order as the results are listed in the table, the influence of such factors would become confounded with the influence of in particular one of his factors (the soap if he did his experiments columnwise in relation to the table, the design if he did his experiments rowwise in relation to the table). Thus he should not really trust his results for one of his factors.

6. (a) Computing

$$\begin{aligned} \sum_{i=1}^{5} x_i &= 0.3 \\ \sum_{i=1}^{5} y_i &= 16.6 \\ \sum_{i=1}^{5} x_i y_i &= 0.85 \\ \sum_{i=1}^{5} x_i^2 &= 0.022 \end{aligned}$$

we get the estimate for the slope to be

$$\widehat{\beta_2} = \frac{5\sum_{i=1}^{5} x_i y_i - \sum_{i=1}^{5} x_i \sum_{i=1}^{5} y_i}{5\sum_{i=1}^{5} x_i^2 - \left(\sum_{i=1}^{5} x_i\right)^2} = \frac{5 \cdot 0.85 - 0.3 \cdot 16.6}{5 \cdot 0.022 - 0.3^2} = -36.5$$

and the intercept

$$\widehat{\beta_1} = \overline{y} - \widehat{\beta_2}\overline{x} = \frac{16.6}{5} + 36.5 \cdot \frac{0.3}{5} = 5.51$$

(b) The simple linear regression model assumes that expected production Y of the compound is given as $\beta_1 + \beta_2 X$, where X is the concentration of the chemical X. It is unlikely that the expected production exactly follows a linear function of X, as the way X influences Y is probably highly non-linear. As one example, such a simple model would predict for large X that either Y becomes negagive, which is impossible, is unchanged as a function of X, or increases without bounds, which is also extremely unlikely.

(c) The best prediction when $X = 0.2$ is

$$5.51 - 36.5 \cdot 0.2 = -1.79$$

Cleary, it is unrealistic to predict that the production becomes negative.

(d) The 5 residuals become

$$\begin{aligned} 4.3 - (5.51 - 36.5 \cdot 0.02) &= -0.48 \\ 4.6 - (5.51 - 36.5 \cdot 0.04) &= 0.55 \\ 3.7 - (5.51 - 36.5 \cdot 0.06) &= 0.38 \\ 2.1 - (5.51 - 36.5 \cdot 0.08) &= -0.49 \\ 1.9 - (5.51 - 36.5 \cdot 0.1) &= 0.04 \end{aligned}$$

so the sum of the squares of the residuals becomes 0.919.

(e) The λ parameter for the linear model has posterior distribution

$$\begin{aligned} \lambda &\sim \text{ExpGamma}\left(\frac{n-k}{2}\frac{1}{2}SS, -2\right) \\ &= \text{ExpGamma}\left(\frac{5-2}{2}, \frac{1}{2} \cdot 0.919, -2\right) = \text{ExpGamma}(3/2, 0.4595, -2) \end{aligned}$$

Thus a 95% credibility interval for the standard deviation e^{λ} becomes

$$\left[\sqrt{\frac{0.919}{\chi^2_{0.025,3}}}, \sqrt{\frac{0.919}{\chi^2_{0.975,3}}}\right] = \left[\sqrt{\frac{0.919}{9.348}}, \sqrt{\frac{0.919}{0.216}}\right] = [0.31, 2.06]$$