Petter Mostad
Matematisk Statistik
Chalmers

Suggested solution for exam in
MSA830: Statistical Analysis and Experimental Design
8 June 2012

1. Let L, M, and H designate that the stone has low, medium, or high levels of the element, respectively. Let I and O designate that the stone comes from the area from which export is illegal, respectively that it comes from some other area. Then we are given the prior $\pi(I) = 0.05$, and the conditional probabilities $\pi(L \mid I) = 0.03$, $\pi(M \mid I) = 0.15$ and $\pi(H \mid I) = 0.82$, and also $\pi(L \mid O) = 0.93$, $\pi(M \mid O) = 0.06$, and $\pi(H \mid O) = 0.01$. Using Bayes formula, we get

$$\begin{aligned}\pi(I \mid H) &= \frac{\pi(H \mid I)\pi(I)}{\pi(H)} \\ &= \frac{\pi(H \mid I)\pi(I)}{\pi(H \mid I)\pi(I) + \pi(H \mid O)\pi(O)} \\ &= \frac{0.82 \cdot 0.05}{0.82 \cdot 0.05 + 0.01 \cdot (1 - 0.05)} \\ &= 0.812\end{aligned}$$

   so there is a 81.2% probability that the stone is from the area from which exports are illegal.

2. (a) We can model the number of incidences per hour as Poisson distributed with rate λ, where λ can be estimated as the average number of incidences per hour observed: $\lambda = 13/100 = 0.13$. The probability of observing no new incidences during the additional hour of observations is then

$$e^{-0.13}\frac{0.13^0}{0!} = e^{-0.13} = 0.8781$$

   so the probability is 87.8% that no new incidences will be observed.

   (b) Using the same model, we get that the probability of 2 or more incidences is

$$1 - e^{-0.13}\frac{0.13^0}{0!} - e^{-0.13}\frac{0.13^1}{1!} = 1 - 0.8780954 - 0.1141524 = 0.0077522$$

   so the probability is 0.8% for 2 or more observations.

   (c) A total of 500 additional hours of observations is planned. We would like to find the probability that $100 - 13 = 87$ or more observations are made during these hours. The number of observations made will be Poisson distributed, and the expected number of observations will be $5 \cdot 13 = 65$. Thus the actual number of observations is approximately normally distributed with expectation 65 and variancd 65. The probability that a variable with such a distribution is 87 or above can be approximated by comparing

$$\frac{86.5 - 65}{\sqrt{65}} = 2.666747$$

   to a standard normal distribution. Using the correct table, we get the probability 0.00379, so the probability is approximately 0.4%.

3. A possible experimental plan is shown below:

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

Here, the settings for the E and F factors are generated using E=ABC and F=BCD, respectively. In Carolines experiments there will be a number of other factors which can be controlled by her and which may influence her results. There are two ways to deal with such factors: Either, she can try to keep them as constant as possible. This keeps the total variability in the results as small as possible, but it also limits the validity of the results to the one value the factors are fixed at. The alternative is blocking, i.e., to balance the settings of the factors against the settings of the factors A,B,C,D,E,F. There will also be a number of factors that are not under the control of Caroline. To make sure the variability of these factors do not influence the results unevenly, it is important to use randomization. For example, Caroline should randomized the order in which she does the experients, to the extent that this is practical.

4. (a) The sample mean and sample variance for species A are $m_A = 463$ and $s_A^2 = 7927.5$, respectively. Thus the distribution for the expectation μ_A of measurements of species A is

$$\mu_A \sim \text{t}(m_A, 5-1, \log(\sqrt{s_A^2/5})) = \text{t}(463, 4, \log(\sqrt{7927.5/5})) = \text{t}(463, 4, \log(39.81834))$$

as there are 5 observations for species A. A 95% credibility interval is thus

$$\begin{aligned} & [463 - t_{4,0.025} \cdot 39.81834, 463 + t_{4,0.025} \cdot 39.81834] \\ = \; & [463 - 2.7764 \cdot 39.81834, 463 + 2.7764 \cdot 39.81834] \\ = \; & [352, 374] \end{aligned}$$

(b) The logged standard deviation λ of measurements for birds of species A has distribution

$$\lambda \sim \text{ExpGamma}\left(\frac{5-1}{2}, \frac{5-1}{2}7927.5, -2\right) = \text{ExpGamma}(2, 15855, -2)$$

A 95% credibility interval for the standard deviation e^{λ} then becomes

$$\left[\sqrt{\frac{2 \cdot 15855}{\chi^2_{0.025,2\cdot 2}}}, \sqrt{\frac{2 \cdot 15855}{\chi^2_{0.975,2\cdot 2}}}\right] = \left[\sqrt{\frac{31710}{11.143}}, \sqrt{\frac{31710}{0.484}}\right] = [53, 256]$$

(c) If μ_B is the expectation of measurements for species B and $s_B^2 = 37420.67$ is the sample variance for species B, we get that, approximately,

$$\mu_B - \mu_A \sim \mathrm{t}\left(689.3333 - 463, \nu, \log \sqrt{\frac{s_A^2}{5} + \frac{s_B^2}{6}}\right)$$

where the degrees of freedom ν can be computed with

$$\nu = \frac{\left(\frac{s_A^2}{5} + \frac{s_B^2}{6}\right)^2}{\frac{(s_A^2/5)^2}{5-1} + \frac{(s_B^2/6)^2}{6-1}} = 7.277$$

so that we get

$$\mu_B - \mu_A \sim \mathrm{t}(226.3333, 7.277, \log(88.44364))$$

A 95% credibility interval then becomes

$$\begin{aligned} & [226.3333 - t_{7,0.025} \cdot 88.44364, 226.3333 + t_{7,0.975} \cdot 88.44364] \\ = \; & [226.3333 - 2.3646 \cdot 88.44364, 226.3333 + 2.3646 \cdot 88.44364] \\ = \; & [17, 435] \end{aligned}$$

(d) As we now have three groups of observations that are assumed to come from distributions with equal standard deviations, we must use the theory for linear models. The sum of squares for such a model will become

$$SS = 4 \cdot 7927.5 + 5 \cdot 37420.67 + 3 \cdot 57578.25 = 391548.1$$

The distribution for μ_A now becomes, as we have a total of $5+6+4 = 15$ observations,

$$\mu_A \sim \mathrm{t}\left(463, 15 - 3, \log \sqrt{\frac{391548.1}{(15-3)5}}\right) = \mathrm{t}(463, 12, \log(80.78243))$$

and a 95% credibility inteval becomes

$$\begin{aligned} & [463 - t_{12,0.025} \cdot 80.78243, 463 + t_{12,0.975} \cdot 80.78243] \\ = \; & [463 - 2.1788 \cdot 80.78243, 463 + 2.1788 \cdot 80.78243] \\ = \; & [287, 639] \end{aligned}$$

For the logged standard deviation λ for all the three normal distributions, we get

$$\lambda \sim \mathrm{ExpGamma}\left(\frac{12-3}{2}, \frac{1}{2}391548.1, -2\right) = \mathrm{ExpGamma}(6, 195774, -2)$$

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

$$\left[\sqrt{\frac{2 \cdot 195774}{\chi^2_{0.025,2\cdot 6}}}, \sqrt{\frac{2 \cdot 195774}{\chi^2_{0.975,2\cdot 6}}}\right] = \left[\sqrt{\frac{391548}{23.337}}, \sqrt{\frac{391548}{4.404}}\right] = [130, 298]$$

(e) David can make non-parametric hypothesis tests comparing the different groups of observations. For example, the Mann-Whitney U test can be used to make pairwise comparisons between species. (But note that it will be difficult to get good results this way with so few observations).

5. (a) The sum of squares for seed type can be computed as

$$SS_{\text{Seed}} = 10 \cdot ((57.1 - 58.8)^2 + (59 - 58.8)^2 + (61.6 - 58.8)^2 + (57.5 - 58.8)^2) = 124.6$$

For the total sum of squares we get from the variance that

$$SS_{\text{Total}} = 39 \cdot 28.47179 = 1110.4$$

This gives the ANOVA table

	SS	D.f.	M.sq.	F	p
Seed	124.6	3	41.5333	1.5172	$0.1 < p < 0.25$
Residuals	985.8	36	27.375		
Total	1110.4	39			

As the p-value is above 0.05, one would say that we cannot determine from our data whether the grass seed type has an influence on the grass output.

(b) The sum of squares for fertilizer can be computed as

$$SS_{\text{Fertilizer}} = 20 \cdot ((57 - 58.8)^2 + (60.6 - 58.8)^2) = 129.6$$

This gives the ANOVA table

	SS	D.f.	M.sq.	F	p
Fertlilizer	129.6	1	129.6	5.0212	$0.025 < p < 0.05$
Residuals	980.8	38	25.81053		
Total	1110.4	39			

(c) The sum of squares of residuals computed in question (b), $SS_{\text{Residuals}} = 980.8$ is equal to the sum over both fertilizers of the sum of squares of differences of observed values and the average for that fertilizer. So in fact, if s_p^2 denotes the pooled variance for the two fertilizers, we get that

$$s_p^2 = SS_{\text{Residuals}}/(19 + 19) = 908.8/38 = 25.81053$$

i.e., the mean square for the residuals. So the difference between expected responses for the fertilizers has distribution

$$\text{t}(60.6 - 57, 20 + 20 - 2, \log \sqrt{s_p^2(1/20 + 1/20)}) = \text{t}(3.6, 38, \log(1.606566))$$

So a 90% credibility interval is

$$\begin{aligned} & [3.6 - t_{38,0.05} \cdot 1.606566, 3.6 + t_{38,0.05} \cdot 1.6066566] \\ = \; & [3.6 - 1.685 \cdot 1.606566, 3.6 + 1.685 \cdot 1.6066566] \\ = \; & [0.89, 6.31] \end{aligned}$$

(d) We first find the sum of squares corresponding to both factors and the interaction:

$$\begin{aligned} & SS_{\text{Seed+Fertilizer+Interaction}} \\ = \; & 5 \cdot ((55.3 - 58.8)^2 + (57.2 - 58.8)^2 + (59.8 - 58.8)^2 + (55.7 - 58.8)^2 \\ & +(58.9 - 58.8)^2 + (60.8 - 58.8)^2 + (63.4 - 58.8)^2 + (59.3 - 58.8)^2) \\ = \; & 254.2 \end{aligned}$$

This means that

$$\begin{aligned} SS_{\text{Interaction}} &= SS_{\text{Seed+Fertilizer+Interaction}} - SS_{\text{Seed}} - SS_{\text{Fertilizer}} \\ &= 254.2 - 124.6 - 129.6 = 0 \end{aligned}$$

and we get the ANOVA table including interaction

	SS	D.f.	M.sq.	F	p
Seed	124.6	3	41.5333	1.552	
Fertlilizer	129.6	1	129.6	4.843	
Interaction	0	3	0	0	$p < 0.01$
Residuals	856.2	32	26.75625		
Total	1110.4	39			

Based on this computation, one should not include interaction in the model.

(e) An ANOVA table without interaction becomes

	SS	D.f.	M.sq.	F	p
Seed	124.6	3	41.5333	1.69781	$0.1 < p < 0.25$
Fertlilizer	129.6	1	129.6	5.29783	$0.025 < p < 0.05$
Residuals	856.2	35	24.46286		
Total	1110.4	39			

According to this table there is a significant influence of the fertilizer, but we can not conclude whether there is a significant influence of the type of grass seed.

(f) We need to assume that the data follow a linear model, i.e., that they come from normal distributions with the same standard deviations and with expectations given by the linear combination of the effects of the parameters. To investigate whether this is a reasonable assumption for these data, one may study the residuals. The residuals should be approximately normally distributed, and they should be independent of all factors, and independent of all other possible predictors, such as time. This can be investigated using various types of plots.