Petter Mostad
Matematisk Statistik
Chalmers

Suggested solution for exam in
MSA830: Statistical Analysis and Experimental Design
16 January 2013

1. (a) If anna assumes that the next 10 students will be a random sample from the students at the school, she can compute the probability. This means that at each appointment, each student at the school has the same probability of coming, and that there is independence between which student comes at each appointment.

   (b) The probability can be found with the Binomial formula:

$$p = \binom{10}{8} 0.57^8 (1-0.57)^{10-8} = 0.09271463$$

   (c) We can use a normal approximation. One should then use a normal distribution with expectation $132 \cdot 0.57 = 75.24$ and variance $132 \cdot 0.57 \cdot (1-0.57) = 32.3532$. To find the approximate probability one should compare

$$\frac{88.5 - 75.24}{\sqrt{32.3532}} = 2.331229$$

   with the standard normal distribution. From the table we find the approximate probability 0.0099.

   (d) In fact, it seems unlikely that the assumptions from (a) are valid, i.e., the students making appointments with Anna are not a random sample from all the students. A possible alternative model would be that female students would tend to visit Anna more often than male students.

2. (a) He can vary maximally 7 factors. Estimating the effect of each factor will demand one degree of freedom, and one degree of freedom is needed to estimate the grand average of his measurements.

   (b) The experimental plan can look like

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

   where $D = AB$, $E = BC$, $F = AC$, and $G = ABC$.

(c) It would still be possible to get some estimate of the main effect of each of seven factors, so the number of factors one could learn about would be the same. However, for each factor, you would have 7 measurements where it was at its "base" setting and only one where it was at its non-base setting. This would lead to a less accurate measurement of the factor than with the experimental plan from (b).

(d) In this case, Anton could get information about $16 - 1 = 15$ factors.

3. (a) The expected value would have the distribution

$$\mathrm{t}\left(16.305, 5, \log\left(\sqrt{0.77995/6}\right)\right) = \mathrm{t}\left(16.305, 5, \log(0.3605436)\right)$$

so a 95% credibility interval would be

$$[16.305 - 2.5706 \cdot 0.3605436, 16.305 + 2.5706 \cdot 0.3605436] \approx [15.38, 17.23].$$

(b) The logged standard deviation would have the distribution

$$\mathrm{ExpGamma}\left(\frac{5}{2}, \frac{5}{2}0.77995, -2\right)$$

so a 95% credibility interval for the standard deviation would be

$$\left[\sqrt{\frac{5 \cdot 0.77995}{12.833}}, \sqrt{\frac{5 \cdot 0.77995}{0.831}}\right] \approx [0.55, 2.17].$$

(c) If we don't assume that the two normal distributions have the same distribution standard deviations, the difference between their expectations has the (approximate) distribution

$$\begin{aligned} & \mathrm{t}\left(17.437 - 16.305, \frac{\left(\frac{0.77995}{6} + \frac{5.8087}{6}\right)^2}{\frac{(0.77995/6)^2}{5} + \frac{(5.8087/6)^2}{5}}, \log\left(\sqrt{\frac{0.77995}{6} + \frac{5.8087}{6}}\right)\right) \\ = \; & [1.131667, 6.319, \log(1.047907)] \end{aligned}$$

so a 95% credibility interval becomes

$$[1.131667 - 2.45 \cdot 1.047907, 1.131667 + 2.45 \cdot 1.047907] \approx [-1.4, 3.7]$$

(d) If we assume that the two normal distributions have the same distribution standard deviations, we get the pooled variance

$$s_p^2 = \frac{5 \cdot 0.77995 + 5 \cdot 5.8087}{10} = 3.294325$$

and the difference between the expectations of the distributions has the (approximate) distribution

$$\begin{aligned} & \mathrm{t}\left(17.437 - 16.305, 10, \log\left(\sqrt{\frac{3.294325}{6} + \frac{3.294325}{6}}\right)\right) \\ = \; & [1.131667, 10, \log(1.047907)] \end{aligned}$$

so a 95% credibility interval becomes

$$[1.131667 - 2.2281 \cdot 1.047907, 1.131667 + 2.2281 \cdot 1.047907] \approx [-1.2, 3.4]$$

(e) In this case, we first get the sum of the squares of the residuals for all the rock types:

$$SS = 5 \cdot 0.77995 + 5 \cdot 5.8087 + 5 \cdot 1.2271 = 39.07875$$

so that the difference between the expectations of the distributions has the distribution

$$\begin{aligned} & \mathrm{t}\left(17.437 - 16.305, 18-3, \log\left(\sqrt{\frac{39.07875}{18-3}\left(\frac{1}{6}+\frac{1}{6}\right)}\right)\right) \\ = \quad & \mathrm{t}\left(1.131667, 15, \log\left(0.9318888\right)\right) \end{aligned}$$

and a 95% credibility interval becomes

$$[1.131667 - 2.1314 \cdot 0.9318888, 1.131667 + 2.1314 \cdot 0.9318888] \approx [-0.8, 3.1]$$

(f) One can make a non-parametric test of the hypothesis that the two sets of measurements come from the same population. Such a test would be the Mann-Whitney U test, also called the Wilcoxon rank sum test.

4. (a) We use Bayes formula and find

$$\begin{aligned} \pi(A \mid \text{topscore}) &= \frac{\pi(\text{topscore} \mid A)\pi(A)}{\pi(\text{topscore})} \\ &= \frac{\pi(\text{topscore} \mid A)\pi(A)}{\pi(\text{topscore} \mid A)\pi(A) + \pi(\text{topscore} \mid B)\pi(B) + \pi(\text{topscore} \mid C)\pi(C)} \\ &= \frac{0.59 \cdot 0.1}{0.59 \cdot 0.1 + 0.2 \cdot 0.37 + 0.03 \cdot 0.53} \\ &= 0.396 \end{aligned}$$

(b) We use Bayes formula and find

$$\begin{aligned} \pi(A \mid \text{bottomscore}) &= \frac{\pi(\text{bottomscore} \mid A)\pi(A)}{\pi(\text{bottomscore})} \\ &= \frac{\pi(\text{bottomscore} \mid A)\pi(A)}{\pi(\text{bottomscore} \mid A)\pi(A) + \pi(\text{bottomscore} \mid B)\pi(B) + \pi(\text{bottomscore} \mid C)\pi(C)} \\ &= \frac{0.06 \cdot 0.1}{0.06 \cdot 0.1 + 0.44 \cdot 0.37 + 0.18 \cdot 0.53} \\ &= 0.023 \end{aligned}$$

5. (a) The sums of squares for the factors A and B become

$$\begin{aligned} SS_A &= 9 \cdot ((38.33 - 46.37)^2 + (50.56 - 46.37)^2 + (50.22 - 46.37)^2) = 873.1818 \\ SS_B &= 9 \cdot ((47.56 - 46.37)^2 + (48.56 - 46.37)^2 + (43.00 - 46.37)^2) = 158.1219 \end{aligned}$$

and the total sum of squares is $SS_{\text{Total}} = 26 \cdot 54.0114 = 1404.296$, so the ANOVA table without interaction becomes

	SS	D.f.	M.sq.	F	p
A	873.1818	2	436.5909	25.7	$p < 0.01$
B	158.1219	2	79.06095	4.66	$0.01 < p < 0.025$
Residuals	372.9923	22	16.95420		
Total	1404.296	26			

To find the ANOVA table with interaction, we first compute that

$$\begin{aligned} SS_A + SS_B + SS_{\text{Interaction}} & \\ &= 3 \cdot ((40.67-46.37)^2 + (50-46.37)^2 + (52-46.37)^2 \\ &\quad +(39-46.37)^2 + (53.67-46.37)^2 + (53-46.37)^2 \\ &\quad +(35.33-46.37)^2 + (48-46.37)^2 + (45.67-46.37)^2) \\ &= 1061.868 \end{aligned}$$

and from this we get that $SS_{\text{Interaction}} = 1061.868 - 873.1818 - 158.1219 = 30.5643$. The ANOVA table with interaction then becomes

	SS	D.f.	M.sq.	F	p
A	873.1818	2	436.5909	22.95	
B	158.1219	2	79.06095	4.16	
Interaction	30.5643	4	7.641075	0.40	$p > 0.25$
Residuals	342.428	18	19.02378		
Total	1404.296	26			

(b) Interaction should not be included in the model as the interaction seems very insignificant. Both the two factors A and B seem to influence the material strength, as the factors have p-values below 0.05 in the ANOVA table. From the table of averages, it seems that the combination $A = y$ and $B = s$ gives the highest strength.

(c) The assumptions are those of a linear model: That the actual observed values are those predicted by the linear model plus error terms, where the error terms are independently sampled from one common normal distribution with zero expectation. This can be checked most easily by studying the residuals of the model: The data values minus the values predicted by the fitted linear model. One can plot these residuals in various ways to detect ways in which the assumptions do not hold.

(d) There are two obvious problems: First, that a possible effect of the person doing the experiment, the “person effect”, is confounded with the effect of factor A. Secondly, a possible time effect is confounded with the effect of factor B. So the two significant effects found above could be due to a person effect and a time effect, respectively.