Petter Mostad
Matematisk Statistik
Chalmers

Suggested solution for re-exam in
MSA830: Statistical Analysis and Experimental Design
15 August 2011

NOTE: The notation used in these solutions have been updated February 2012.

1. (a) The mean and variance of the 6 observations for batteries of type X is 4.8167 and 0.5777, respectively. According to the formula for the posterior distribution for the expectation of the distribution of battery lengths for batteries of type X, it is

$$\mathrm{t}(4.8167, 6-1, \log(\sqrt{0.5777/6})) = \mathrm{t}(4.8167, 5, \log(\sqrt{0.09628}))$$

We find from the table of the t distribution that a 95% credibility interval for the standard t distribution with 5 degrees of freedom is

$$[-2.5706, 2.5706]$$

Thus the 95% credibility interval for our distribution is

$$[4.8167 - 2.5706 \cdot \sqrt{0.09628}, 4.8167 + 2.5706 \cdot \sqrt{0.09628}] = [4.02, 5.62].$$

(b) The logged scale λ has distribution

$$\mathrm{ExpGamma}\left(\frac{6-1}{2}, \frac{6-1}{2}0.5777, -2\right) = \mathrm{ExpGamma}\left(\frac{5}{2}, \frac{2.8885}{2}, -2\right).$$

As a χ^2 distribution with 5 degrees of freedom has 95% credibility interval

$$[0.831, 12.833]$$

a 95% credibility interval for e^λ becomes

$$\left[\sqrt{\frac{2.8885}{12.833}}, \sqrt{\frac{2.8885}{0.831}}\right] = [0.4744, 1.8644]$$

Also, a 95%credibility interval for the precision of the distribution becomes

$$[1/1.8644^2, 1/0.4744^2] = [0.2877, 4.4435]$$

(c) For batteries of type Y, we get a mean and variance of 6.0667 and 0.6907, respectively. So we get a pooled variance of

$$s_p^2 = \frac{5 \cdot 0.5777 + 5 \cdot 0.6907}{5+5} = 0.6342.$$

Thus the difference in expected values for the two battery types has posterior distribution

$$\mathrm{t}\left(6.0667 - 4.8167, 6+6-2, \log\sqrt{\left(\frac{1}{6}+\frac{1}{6}\right)0.6342}\right) = \mathrm{t}\left(1.25, 10, \log\sqrt{0.2114}\right)$$

A 95% credibility interval for this distribution becomes

$$[1.25 - 2.2281\sqrt{0.2114}, 1.25 + 2.2281\sqrt{0.2114}] = [0.23, 2.27]$$

It seems like batteries of type Y have longer life lengths.

(d) The test statistic of the relevant hypothesis test is

$$\frac{0.6907}{0.5777} = 1.1956.$$

Comparing this with an F distribution with 5 and 5 degrees of freedom, we see that probability for such a distribution to be above this value is above 0.25. Thus the p-value is above $2 \cdot 0.25 = 0.5$, and so it is between 0.5 and 1. The test supports the decision of using a formula where the precisions of the two distributions are assumed to be equal.

(e) It is possible to perform a non-parametric test on the data: In this case, one could perform a Wilcoxon rank sum test, to test whether the two datasets could come from the same distribution. In principle, permutation tests are also possible, and these would avoid the assumption of normal distributions.

2. Let us use the following notation:

gold	Minable amounts of gold are present
no gold	Minable amounts of gold are not present
other metal present	Traces of the other metal is present
other metal not present	Traces of the other metal is not present

Then we can write

$$\begin{aligned}
&\pi(\text{gold} \mid \text{other metal present}) \\
=\ &\frac{\pi(\text{other metal present} \mid \text{gold})\pi(\text{gold})}{\pi(\text{other metal present})} \\
=\ &\frac{\pi(\text{other metal present} \mid \text{gold})\pi(\text{gold})}{\pi(\text{other metal present} \mid \text{gold})\pi(\text{gold}) + \pi(\text{other metal present} \mid \text{no gold})\pi(\text{no gold})} \\
=\ &\frac{0.1 \cdot 0.04}{0.1 \cdot 0.04 + 0.003 \cdot 0.96} \\
=\ &0.581
\end{aligned}$$

So the posterior probability that there is minable amounts of gold at the location is 58%.

3. (a) First, we compute the relevant sums of squares:

$$SS_{\text{habitat}} = 4 \cdot \left((49.5 - 55.5)^2 + (53 - 55.5)^2 + (44.5 - 55.5)^2 + (75 - 55.5)^2\right) = 2174$$

and

$$SS_{\text{time}} = 8 \cdot \left((47 - 55.5)^2 + (64 - 55.5)^2\right) = 1156$$

The variance of the data can be computed to 234.8, so the total sum of squares becomes

$$SS_{\text{total}} = 15 \cdot 234.8 = 3522$$

Based on this, we construct the following ANOVA table:

	SS	D.f.	M.sq.	F	p
Habitat	2174	3	724.67	41.52	$p < 0.01$
Time	1156	1	1156	66.23	$p < 0.01$
Residuals	192	11	17.4545		
Total	3522	15			

We find that both p-values are less than 0.01, so we conclude that both the habitat and the time of year influences the amount of the chemical. Assumptions are that the observations for each combination of habitat and time of year are from a normal distribution, and that the variances of all these normal distributions are the same. As we are not including interaction, we also assume that the effect of the habitat adds to the effect of the time of year, and that these two factors do not interact.

(b) To compute the sum of squares for the interaction, we need to first compute the 8 averages for the 8 combinations of habitat and time of year. These averages are

$$43, 45, 38, 62, 56, 61, 51, 88$$

The sum of squares representing all effects, including interaction, can now be computed as

$$\begin{aligned} SS_{\text{alleffects}} &= 2 \cdot \Big((43-55.5)^2 + (45-55.5)^2 + (38-55.5)^2 + (62-55.5)^2 \\ &\quad +(56-55.5)^2 + (61-55.5)^2 + (51-55.5)^2 + (88-55.5)^2\Big) = 3444 \end{aligned}$$

The sum of squares for interaction can be computed with

$$SS_{interaction} = SS_{\text{alleffects}} - SS_{habitat} - SS_{time} = 3444 - 2174 - 1156 = 114$$

With this, one can compute that the ANOVA table becomes

	SS	D.f.	M.sq.	F	p
Habitat	2174	3	724.67	74.32	$p < 0.01$
Time	1156	1	1156	118.56	$p < 0.01$
Interaction	114	3	38	3.89	$0.05 < p < 0.1$
Residuals	78	8	9.75		
Total	3522	15			

We conclude that the interaction is not significantly large, and that we can use the conclusions from the previous ANOVA table.

4. (a) The Binomial distribution gives

$$\binom{6}{5} 0.69^5 (1-0.69)^{6-5} = \frac{6!}{5!1!} 0.69^5 0.31 = 6 \cdot 0.1564 \cdot 0.31 = 0.29$$

So the probability that exactly 5 out of the next 6 customers are female is 0.29.

(b) The easiest thing is to first compute the probability that all 6 of the next customers are female. This probability is

$$0.69^6 = 0.11$$

The probability that 4 of fewer of the next 6 customers are female is then

$$1 - 0.29 - 0.11 = 0.60$$

(c) The assumption is that all customers arrive independently. (This is not completely realistic as sometimes people shop in pairs or groups).

5. (a) Lisa could use the following fractional factorial design, where each of the 7 factors have been given names A,B,C,D,E,F,G, respectively:

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

Here, we have used the equations E=ABC, F=BCD, and G=ACD for construction. However, many different fractional factorial designs could be used here.

(b) The name of the particular design above is 2^{7-3}_{IV}.

(c) Other factors that could be controlled should be kept as constant as possible during the testing. To avoid coufounding with uncontrollable factors related with time, it would be nice to randomize the order in which the 16 experimental runs were done.

(d) From the information, we get that the average training time for an enclosed location is $543/8 = 67.875$, and the average training time for an open location is $619/8 = 77.375$. Thus, the expected effect of changing the location from enclosed to open is an increase in training time of $77.375 - 67.875 = 9.5$.

(e) In fact, the formula for the posterior for each of the parameters for the effects show that the parameters governing the size of the interval, $n-k$ and $\frac{SS}{(n-k)n}$, are the same for all the parameters. Thus, in all situations, the two credibility intervals will be equally long.

6. (a) The distribution is a multivariate Normal-Gamma distribution.

(b) The marginal distribution for β_2 is a t-distribution.

(c) In the multiple regression model, the error terms $\epsilon_1, \dots, \epsilon_n$ are assumed to all be a random sample from the same normal distribution with zero expectation. The residuals are approximations of these error terms, so according to the model, they should not show any dependency on other parameters, and their distribution should be approximately normal. Whether this is so can be checked with various plots.