Petter Mostad
Matematisk Statistik
Chalmers

Suggested solution for exam in
MSA830: Statistical Analysis and Experimental Design
March 2010

1. We have the two variables

G: G=1 means X has the gene, G=0 the opposite
H: H=1 means X has hypertension, H=0 the opposite

and we know that

$$\begin{aligned} \Pr(G=1) &= 0.12 \\ \Pr(H=1 \mid G=0) &= 0.18 \\ \Pr(H=1 \mid G=1) &= 0.89. \end{aligned}$$

Bayes theorem then gives that

$$\begin{aligned} \Pr(G=1 \mid H=1) &= \frac{\Pr(H=1 \mid G=1)\Pr(G=1)}{\Pr(H=1)} \\ &= \frac{\Pr(H=1 \mid G=1)\Pr(G=1)}{\Pr(H=1 \mid G=1)\Pr(G=1) + \Pr(H=1 \mid G=0)\Pr(G=0)} \\ &= \frac{0.89 \cdot 0.12}{0.89 \cdot 0.12 + 0.18 \cdot (1-0.12)} \\ &= 0.403. \end{aligned}$$

So the probability is 40.3% that X has the gene G.

2. (a) The increases in scores from the old to the new control are

$$65, 148, 39, -157, 132, -34$$

The sample mean of these numbers is 32.17, and the sample variance is 12934.17. Assuming that these differences are normally distributed, and using the standard prior for the mean and precision of this normal distribution, we get that the posterior for the mean of the normal distribution is the t distribution

$$\mathrm{t}(32.17, 6-1, \log\sqrt{\frac{12934.17}{6}}) = \mathrm{t}(32.17, 5, \log\sqrt{2155.659})$$

The 95% credibility interval thus becomes

$$[32.17 - \sqrt{2155.659}t_{0.025,5}, 32.17 + \sqrt{2155.659}t_{0.025,5}] = [-87.20, 151.54]$$

where $t_{0.025,5} = 2.571$ is the value such that a standard t-distributed variable with 5 degrees of freedom is above it with probability 0.025. The value is found in the table for the t distribution.

(b) The analysis in (a) can be described as using a linear model with one parameter (in addition to the parameter for the unknown precision):

$$\beta_1 : \text{Expected increase in points switching from old to new}$$

To take the learning effect into account, Morten could extend this model with an additional parameter

$$\beta_2 : \text{Expected increase in points from first to second hour}$$

and find the posterior for β_1 in this model, still using the differences as his data. For example, if friends 1,2, and 5 used the old control first, Morten would use the linear model

$$\begin{aligned}
778 - 713 &= \beta_1 + \beta_2 + \epsilon_1 \\
569 - 421 &= \beta_1 + \beta_2 + \epsilon_2 \\
655 - 616 &= \beta_1 - \beta_2 + \epsilon_3 \\
845 - 1001 &= \beta_1 - \beta_2 + \epsilon_4 \\
700 - 568 &= \beta_1 + \beta_2 + \epsilon_5 \\
695 - 729 &= \beta_1 - \beta_2 + \epsilon_6
\end{aligned}$$

3. (a) According to Emelie's assumptions, we can model the number of counted earthquakes as Poisson distributed with rate 14. For a Poisson distriubution with rate 14, the probability of observing 5 is given by

$$\frac{14^5 e^{-14}}{5!} = 0.003727$$

(b) To answer this question, it is natural to use a normal approximation. A Poisson distribution with rate 14 is approximated by a Normal distribution with expectation 14 and variance 14. The probability of observing 20 or more earthquakes can thus be approximated by the probability for a standard normal distribution to be above

$$\frac{19.5 - 14}{\sqrt{14}} = 1.47$$

and according to the table for the standard normald distribution, this probability is $1 - 0.92922 = 0.07$. Using 19 or 20 instead of 19.5 also give acceptable approximations.

4. (a) The table becomes

	Sum of squ.	Deg. freed.	Mean squ.	F value	p value
Thickness	11.8	1	11.8	1.26	$p > 0.25$
Fold design	116.6	2	58.3	6.25	$0.025 < p < 0.05$
Interaction	11.2	2	5.6	0.60	$p > 0.25$
Residuals	56.0	6	9.33		
Total	195.6	11			

(b) The folding design seems to make a significant difference for the flight length. However, the thickness of the paper does not seem to make a difference, and there seems to be no interaction between the two factors. From the values in the table, it is impossible to tell which of the three designs gives the longest-flying paper airplane.

(c) First, let's compute the sum of squares for the thickness. The average measurements are 11.633 and 14.767 for the thick and thin planes, respectively, while the grand mean is 13.2. Thus the sum of squares for the thickness becomes

$$6 \cdot (11.633 - 13.2)^2 + 6 \cdot (14.767 - 13.2)^2 = 29.45$$

Secondly, the average measurements for the three designs are 11.775, 13.1, and 14.725 for design A, B, and C, respectively. Thus the sum of squares for the design becomes

$$4 \cdot (11.775 - 13.2)^2 + 4 \cdot (13.1 - 13.2)^2 + 4 \cdot (14.725 - 13.2)^2 = 17.46$$

As the analysis includes interaction, we can find the sum of squares for the residuals as follows: The average values in each of the 6 cells in the data table, i.e., the fitted values in a model including interaction, are

10.5	13.05
13.5	12.7
10.9	18.55

Thus the measurements minus the fitted values are

2.5	-0.25
-2.5	0.25
2.5	0.1
-2.5	-0.1
0.5	1.35
-0.5	-1.35

and the sum of the squares of these numbers are 29.29. Finally, as always, the total sum of squares can be found by computing the sample variance, which is 10.22, and multiplying it with 11 (one minus the number of measurements), which gives 112.42. The sum of squares for interaction can be found by subtraction, and the first part of Hans' ANOVA table becomes

	Sum of squ.	Deg. freed.	Mean squ.	F value	p value
Thickness	29.45				
Fold design	17.46				
Interaction	36.23				
Residuals	29.29				
Total	112.42				

5. An important aspect is that Dmitrii should try to make sure that if his data indicate a difference in the "goodness" of the two recipes, this difference cannot be explained by other factors. For example

- If Dmitrii wanted results valid for the recipes themselves, and not for his interpretation of them, he should ideally get 10 of his friends to each bake a pizza, with 5 following the old and 5 following the new recipe. However, more realistically, Dmitrii would bake all the pizzas himself. With this setup, his results would be valid for his interpretation of the recipes.
- It is important that differences between the pizzas are connected to the recipes, and to nothing else. For example, Dmitrii should avoid making five pizzas of one type first, and then the other type, as then either the first type would be cold, or re-heated, or his friends would eat one type first, when they were hungry, and then the other type. If he can bake two pizzas in his oven at the same time, he could make 5 runs with one of each type of pizza in the oven simultaneously. To the extent that he cannot do blocking in this way, he should consider randomization.
- In the tasting, each person should ideally get pieces of both pizzas simultaneously, so as to be able to make a direct comparison.
- Dmitrii should try to avoid that his friends influence each other in their grading of the pizzas. If there is such an influence, the results become dependent, and difficult to analyze.
- Dmitrii should ask his friends to give each pizza type a score from 0 to 10. Although it woudl be easier for his friends simply to indicate which pizza they like best, a score could also given some information about how much better one is compared to the other, or if they are basically the same (and if the person likes kebab-pizza or not).
- Dmitrii should not inform his friends which is the old recipe and which is the new, as his friends then might indicate that the new one is the best, just to be a good friend.

6. (a) Using "+" and "-" for the two settings of his parameters, a possible experimental plan could look like

A	B	C	D
-	-	-	-
-	-	+	+
-	+	-	+
-	+	+	-
+	-	-	+
+	-	+	-
+	+	-	-
+	+	+	+

where the settings for factor D has been constructed as the product ABC.

(b) The confounded interaction effects in the experimental plan above are

$$\begin{aligned} AB &= CD \\ AC &= BD \\ AD &= BC \end{aligned}$$

7. (a) For machine X, the sample mean is 75.4 and the sample variance 35.3. For machine Y, the sample mean is 85.8, and the sample variance is 112.2. Thus the pooled variance

is

$$s_p^2 = \frac{4 \cdot 35.3 + 4 \cdot 112.2}{4 + 4} = 73.75.$$

If we assume a common precision for both normal distributions, and if we use a standard prior, then the posterior for the difference between the expectations of the two distributions is the t distribution

$$\mathrm{t}\left(85.8 - 75.4, 5 + 5 - 2, \log\sqrt{\left(\frac{1}{5} + \frac{1}{5}\right)73.75}\right) = \mathrm{t}(10.4, 8, \log\sqrt{29.5})$$

Thus the expectation is 10.4, and a 95% credibility interval is given by

$$[10.4 - \sqrt{29.5}2.306, 10.4 + \sqrt{29.5}2.306] = [-2.12, 22.92]$$

where $2.306 = t_{0.025,8}$ is the value such that a t distribution with 8 degrees of freedom has probability 0.025 of being above this value.

(b) The posterior for the logged scale λ is

$$\mathrm{ExpGamma}\left(\frac{5+5-2}{2}, \frac{5+5-2}{2}s_p^2, -2\right) = \mathrm{ExpGamma}(4, 295, -2)$$

Thus a 95% credibility interval for the standard deviation e^λ is given by

$$\left[\sqrt{\frac{2 \cdot 295}{17.53}}, \sqrt{\frac{2 \cdot 295}{2.18}}\right] = [5.8014, 16.4512]$$

where $2.18 = \chi^2_{0.975,8}$ and $17.53 = \chi^2_{0.025,8}$ are the values such that a χ^2 distribution with 8 degrees of freedom has probabilities 0.975 and 0.025 of being above these values, respectively. Thus a 95% credibility interval for the *variance* is given by

$$[5.8014^2, 16.4512^2] = [33.66, 270.64].$$

(c) The test statistic becomes

$$\frac{s_Y^2}{s_X^2} = \frac{112.2}{35.3} = 3.178$$

and this should be compared with an F distribution with 4 and 4 degrees of freedom. Comparing with the relevant tables, we see that the probability for such a distribution to be above 3.178 is between 0.1 and 0.25, so the p-value is between 0.2 and 0.5. Thus, one should not reject the null hypothesis in this case, and it is not necessary for Emma to re-do her calculations.

8. The two plots on the left indicate that there are problems with his current model. The normal probability plot on the top left has a slight S-shape, indicating that the residuals are not normally distributed. The plot of the residuals against the predictor x_2 indicate that the residuals depend on this predictor, which they should not according to the linear model. What seems to be wrong is that the residuals make a jump at roughly $x_2 = 55$. This indicates that the output changes abruptly in size when x_2 passes 55. This phenomenon might also be the reason why the residuals do not seem to be normally distributed. A way to do the analysis might then be to include a predictor that is 0 when $x_2 < 55$ and 1 when $x_1 \geq 55$.