Petter Mostad
Matematisk Statistik
Chalmers

Suggested solution for exam in MSA830: Statistical Analysis and Experimental Design 11 March 2011 (revised notation 3 October 2011)

1. Let us write, for short,

$$\begin{aligned} \text{Has X} &: \text{Luigi is carrying the disease gene X} \\ \text{No X} &: \text{Luigi is not carrying the disease gene X} \\ \text{Has diab} &: \text{Luigi has type-2 diabetes} \\ \text{No diab} &: \text{Luigi does not have type-2 diabetes} \end{aligned}$$

Then we get

$$\begin{aligned} \pi(\text{Has X} \mid \text{No diab}) &= \frac{\pi(\text{No diab} \mid \text{Has X})\pi(\text{Has X})}{\pi(\text{No diab})} \\ &= \frac{\pi(\text{No diab} \mid \text{Has X})\pi(\text{Has X})}{\pi(\text{No diab} \mid \text{Has X})\pi(\text{Has X}) + \pi(\text{No diab} \mid \text{No X})\pi(\text{No X})} \\ &= \frac{(1-0.84)\cdot 0.5}{(1-0.84)\cdot 0.5 + (1-0.03)\cdot 0.5} \\ &= 0.1416 \end{aligned}$$

So there is roughly a 14% chance that Luigi is carrying the disease gene.

2. (a) First, we compute that the mean and sample variances are $m_A = 21.9$ and $s_A^2 = 0.125$ for city A and $m_B = 22.44$ and $s_B^2 = 0.293$ for city B. In the hypothesis test where the null hypothesis is that the data come from two normal distributions with the same standard deviationss, and the alternative is that the standard deviationss are different, the test statistic is

$$F = \frac{s_B^2}{s_A^2} = \frac{0.293}{0.125} = 2.344$$

which should be compared to an F distribution with 4 and 4 degrees of freedom. According to the tables for the F distribution, the probability to that distribution to be above 2.344 is between 0.25 and 0.1. Thus the p-value is twice this, i.e., it is between 0.2 and 0.5. This means that the null hypothesis can not be rejected.

(b) If Alvar uses the p-value above to make a decision about which model to use, he should choose a model where the standard deviationss of the normal distributions are assumed equal, as this model was not rejected in the hypothesis test. For this model, we first compute the pooled variance

$$s_p^2 = \frac{(5-1)\cdot s_A^2 + (5-1)\cdot s_B^2}{5+5-2} = 0.209$$

We then get for the posterior of the difference between the expectations $\mu_A - \mu_B$:

$$\mu_A - \mu_B \sim \mathrm{t}\left(21.9 - 22.44, 5+5-2, \log\left(\sqrt{(1/5+1/5)\cdot 0.209}\right)\right) = \mathrm{t}(-0.54, 8, \log(0.2891))$$

As a 95% credibility interval for the standard t distribution with 8 degrees of freedom is $[-2.306, 2.306]$, a 95% credibility interval for $\mu_A - \mu_B$ then becomes

$$[-0.54 - 2.306 \cdot 0.2891, -0.54 + 2.306 \cdot 0.2891] = [-1.207, 0.127]$$

(c) The credibility interval not only depends on the average values for cities A and B, but also on the amount of variability in each city. If we assume that the standard deviations of the normal distributions of all the 5 cities is the same, then the sample variances of all the cities give information about the standard deviationss in the distributions for cities A and B. In fact, we can use that the posterior for $\mu_A - \mu_B$ is

$$\mu_A - \mu_B \sim \mathrm{t}\left(m_A - m_B, n-k, \log\sqrt{\frac{SS}{n-k}\left(\frac{1}{n_A} + \frac{1}{n_A}\right)}\right),$$

where n is the total number of observations in all 5 cities, i.e., $n = 25$, k is the number of groups, i.e., $k = 5$, n_A and n_B are the numbers of observations in cities A and B respectively, so that $n_A = n_B = 5$, and SS is the sum of squares of the residuals, when we have 5 normal distributions with the same precisions and 5 observations from each group. As the residuals correspond to the differences used to compute the sample variances for each group, we get that

$$SS = 4\cdot s_A^2 + 4\cdot s_B^2 + 4\cdot s_C^2 + 4\cdot s_D^2 + 4\cdot s_E^2 = 4(0.125+0.293+0.128+0.008+0.077) = 2.524$$

Thus we get for the posterior that

$$\mu_A - \mu_B \sim \mathrm{t}\left(-0.54, 25-5, \log\sqrt{\frac{2.524}{25-5}\left(\frac{1}{5} + \frac{1}{5}\right)}\right) = \mathrm{t}\left(-0.54, 20, \log(0.2247)\right)$$

As a 95% credibility interval for a standard t distribution with 20 degrees of freedom is $[-2.086, 2.086]$, the 95% credibility interval for the difference $\mu_A - \mu_B$ now becomes

$$[-0.54 - 2.086 \cdot 0.2247, -0.54 + 2.086 \cdot 0.2247] = [-1.009, -0.071]$$

3. (a) The probability is given by the Binomial distribution:

$$\binom{40}{2}(0.07)^2(1-0.07)^{40-2} = \frac{40 \cdot 39}{1 \cdot 2} 0.07^2 \cdot 0.93^{38} = 0.2425$$

So the probability is approximately 24%.

(b) The probability that a randomly chosen person voted for C, D, or E, is

$$0.108 + 0.072 + 0.07 = 0.25$$

Thus the probability asked for is again given by the Binomial distribution:

$$\binom{40}{3}(0.25)^3(1-0.25)^{40-3} = \frac{40 \cdot 39 \cdot 38}{1 \cdot 2 \cdot 3} 0.25^3 \cdot 0.75^{37} = 0.00368$$

So the probability is approximately 0.4%.

(c) We would like to compute the probability for observing 30 or more in a Binomial distribution with parameters $n = 40$ and $p = 0.451$. We approximate this probability by using a normal distribution. The expectation and variance of the Binomial distribution is $np = 40 \cdot 0.451 = 18.04$ and $np(1-p) = 18.04 \cdot (1-0.451) = 9.90396$, so we use a normal distribution with the same expectation and variance. The probability is approximated by the probability for a variable with the standard normal distribution to be above

$$\frac{29.5 - 18.04}{\sqrt{9.90396}} = 3.6415$$

From the table for the standard normal distribution, we get that this probability is approximately 0.00014, so the probability we seek is about 0.01%.

4. (a) The list of all experimental runs would be

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

The list indicates how the factors should be set in each of the 16 experimental runs. However, they should *not* be performed in the order listed above, but rather in a randomized order.

More precisely, the smelling panel should be give 16 samples in a randomized order. The deodorants should be as similar as possible except for the smell. The persons in the panel should not be given information about which deodorants are replicates, and they should not get information about the settings of the factors for the different deodorants.

If there is a suspicion that there could be a time effect on Eric's manufacturing of the deodorants, that manufacturing should also be done in a randomized order.

(b) Assuming that Eric lists his results in an order corresponding to the lines of the experimental plan, the design matrix should be:

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

(c) The residuals should be (approximately) independent and come from a normal distribution with a precision that is constant and does not depend on any of the predictors. So one thing Eric could check in the plot is whether the variance of the residuals seemed to be roughly the same when the level of A is low and the level of A is high. If one group of residuals are more spread out than the other, it could indicate a problem with the model. (Eric could also attempt to see if the residuals looked normally distributed in each group, but this would be generally more difficult to check with such a limited amount of data.) He could also be able to spot outliers in the plot.

5. (a) We can compute

$$SS_{\text{Machine}} = 4 \cdot \Big((359.75 - 356.55)^2 + (379.5 - 356.55)^2 + (362 - 356.55)^2 \\ +(331.25 - 356.55)^2 + (350.25 - 356.55)^2\Big) = 4985.70$$

and

$$SS_{\text{X}} = 10 \cdot \Big((344.1 - 356.55)^2 + (369 - 356.55)^2\Big) = 3100.05$$

The total sum of squares can be computed from the given variance as

$$SS_{\text{Total}} = 19 \cdot s^2 = 19 \cdot 881.1026 = 16740.95$$

The sum of squares of residuals can then be computed by subtraction:

$$SS_{\text{Residuals}} = SS_{\text{Total}} - SS_{\text{Machine}} - SS_{\text{X}} = 16740.95 - 4985.70 - 3100.05 = 8655.20$$

We can now set up the ANOVA table as follows:

	SS	D.f.	M.sq.	F	p
Machine	4985.70	4	1246.425	2.02	$0.1 < p < 0.25$
X	3100.05	1	3100.05	5.01	$0.025 < p < 0.05$
Residuals	8655.20	14	618.2286		
Total	16740.95	19			

As the p-value for X is below 0.05, one would say there is a significant effect of X. Looking at the averages, we see that high X produces the highest response values, so one should recommend high levels of X. The effect of the machines is not significant, as the p-value is not below 0.05.

(b) The ANOVA table for the situation when the machines are disregarded would have the same SS_{X} and SS_{Total} as above, but no line for "Machine". Thus the $SS_{\text{Residuals}}$ in this new table will be the sum of the $SS_{\text{Residuals}}$ and SS_{Machine} from the old table. Similarly for degrees of freedom, and we get the ANOVA table

	SS	D.f.	M.sq.	F	p
X	3100.05	1	3100.05	4.0907	$0.05 < p < 0.1$
Residuals	13640.90	18	757.8278		
Total	16740.95	19			

Note that the effect of X is no longer significant, as the p-value is now above 0.05.

6. (a) As the design has 6 columns but is based on a full design for 4 factors, it is a 2^{6-2} design. To find the resolution, note that column E is equal to the product of columns A and B. This means that $ABE = I$. Thus the resolution is at most 3. If the resolution were 2, there would exist two columns whose product would be equal to the identity column of only pluses. For this to happen, the two columns would have to be identical. As there are no identical columns in the design, the resolution is exactly equal to 3, and the design can be named a 2^{6-2}_{III} design.

   (b) The only specific goal that is mentioned is that Edith would like to investigate the interaction between space and type of compartments. Thus it is important that this interaction is not confounded with any other factors. This would be the case if the product of the columns corresponding to space and type of compartments was equal to some other column. As already mentioned, we see that $AB = E$, so an example of assignments of factors to letters that would *not* be recommendable would be one where A and B are assigned to space and type of compartments.

7. (a) This is true, and was used as an intuitive definition of the variance of a probability distribuion in our course.

   (b) This is false: The average of a large samle from a Gamma distribution will be approximately normally distributed, according to the central limit theorem.

   (c) This is false: One way to see this is that values that are Gamma distributed are always greater than or equal to zero: The difference between two values that are Gamma distributed can easily become negative.

   (d) This is true, in fact, it is exactly true and not only approximately true. It is an important property of the normal distribution.

8. The linear regression line can be constructed by minimizing the sum of the squares of the residuals (leading to the "least squares" estimates for the parameters). When x is the predictor and y is the response, the residuals are the *vertical* distances beteween the data points and the regression line. One can see from the figure that for line A, these residuals are 1, 0, and something larger than 1, so the sum of squares is larger than 2. For line C, the residuals are 0.5, -0.5, and 0, so the sum of the squares of these is 0.5. It is easy to see that the sum of the squares of the residuals for line B must be larger than this, so C is the regression line.