Petter Mostad
Matematisk Statistik
Chalmers

Suggested solution for exam in
MSA830: Statistical Analysis and Experimental Design
21 August 2012

1. (a) With the additional assumption that the cases are independent, the number of cases where the putative father is found to be the true father is Binomially distributed. The probability becomes

$$\binom{10}{2} 0.43^2 \cdot (1 - 0.43)^8 = 0.09271463$$

   (b) We use a normal approximation with expectation $0.43 \cdot 100 = 43$ and variance $0.43 \cdot (1 - 0.43) \cdot 100 = 24.51$:

$$\frac{49.5 - 43}{\sqrt{24.51}} = 1.31293$$

   compared to the standard normal distribution gives 0.095, so the probability is approximately 9.5%.

   (c) The probability assuming the rate 43% becomes

$$\binom{10}{9} 0.43^9 (1 - 0.43)^1 = 0.002864778$$

   while assuming the rate 65% it becomes

$$\binom{10}{9} 0.65^9 (1 - 0.65)^1 = 0.07249169$$

2. (a) The prior probability that the patient has varant A is $0.02 \cdot 0.8 = 0.016$. Given that the test is positive, Bayes formula shows that

$$\frac{0.9 \cdot 0.016}{0.9 \cdot 0.016 + 0.1 \cdot (1 - 0.016)} = 0.1276596$$

   is the probability that the patient has variant A.

   (b) In this case, Bayes formula shows that

$$\frac{0.1 \cdot 0.016}{0.1 \cdot 0.016 + 0.9 \cdot (1 - 0.016)} = 0.001803427$$

   (c) One way to think is the following: From the information in the question, we know that a proportion $0.02 \cdot 0.8 = 0.016$ of the population has variant A, while a proportion $0.02 \cdot (1 - 0.8) = 0.004$ has one of the variants B, C, or D. So if a patient either does not have X or has variants B, C, or D, the probability for variants B, C, or, D is $0.004/(1 - 0.016) = 0.004065$. From (b) we know that the probability that this particular patient does not have X or has one of the variants B, C, or D is

$$1 - 0.001803427 = 0.9981966$$

   So the probability that the patient has variants B, C, or D is $0.9981966 \cdot 0.004065 = 0.004057669$, or very slightly above 0.4%.

3. (a) The mean, variance, and standard deviation of the data is 178.67, 89.8667 and 9.4798, respectively. The distribution for the expeted length becomes

$$\mathrm{t}(178.67, 5, \log(9.4798/\sqrt{6})) = \mathrm{t}(178, 5, \log(3.87))$$

The credibility interval becomes

$$[178.67 - t_{5,0.025} \cdot 3.87, 178.67 + t_{5,0.025} \cdot 3.87] = [168.7, 188.6]$$

using that $t_{5,0.025} = 2.5706$.

(b) The distribution of the logged standard deviation becomes

$$\mathrm{ExpGamma}\left(\frac{5}{2}, \frac{5}{2} \cdot 89.8667, -2\right) = \mathrm{ExpGamma}\left(\frac{5}{2}, \frac{449.3335}{2}, -2\right)$$

and a 95% credibility interval for the standard deviation becomes

$$\left[\sqrt{\frac{449.3335}{\chi^2_{0.025,5}}}, \sqrt{\frac{449.3335}{\chi^2_{0.975,5}}}\right] = \left[\sqrt{\frac{449.3335}{12.833}}, \sqrt{\frac{449.3335}{0.831}}\right] = [5.91, 23.25]$$

(c) We can compute that

$$(192-181)^2+(187-181)^2+(173-181)^2+(173-181)^2+(180-181)^2+(167-181)^2 = 482$$

and so the distribution of the logged standard deviation now becomes

$$\mathrm{ExpGamma}\left(\frac{6}{2}, \frac{482}{2}, -2\right)$$

and a 95% credibility interval becomes

$$\left[\sqrt{\frac{482}{\chi^2_{0.025,6}}}, \sqrt{\frac{482}{\chi^2_{0.975,6}}}\right] = \left[\sqrt{\frac{482}{14.449}}, \sqrt{\frac{482}{1.237}}\right] = [5.77, 19.74]$$

(d) In all these questions we need to assume that the lengths of male students at the university has a normal distribution, and that the observations are an independent random sample from the group of students.

4. (a) We get

$$SS_{\text{Temperature}} = 36 \cdot ((32.69 - 31.01)^2 + (29.33 - 31.01)^2) = 201.2128$$

and

$$SS_{\text{Lighting}} = 24 \cdot ((27.46 - 31.01)^2 + (29.37 - 31.01)^2 + (36.21 - 31.01)^2) = 1015.97$$

so the ANOVA table becomes

	SS	D.f.	M.sq.	F	p
Temperature	203.2128	1	203.2128	4.41	$0.025 < p < 0.05$
Lighting	1015.97	2	507.985	11.023	$p < 0.01$
Residuals	3133.803	68	46.085		
Total	4352.986	71			

We get that both the temperature and the lighting has a significant effect on the sales.

(b) For the smell we get

$$\begin{aligned} SS_{\text{smell}} &= 18 \cdot ((28.39 - 31.01)^2 + (29.61 - 31.01)^2 \\ &\quad +(29 - 31.01)^2 + (37.06 - 31.01)^2) \\ &= 889.49 \end{aligned}$$

so the ANOVA table now becomes

	SS	D.f.	M.sq.	F	p
Temperature	203.2128	1	203.2128	5.89	$0.01 < p < 0.025$
Lighting	1015.97	2	507.985	14.71	$p < 0.01$
Smell	889.49	3	296.497	8.59	$p < 0.01$
Residuals	2244.313	65	34.528		
Total	4352.986	71			

(c) For the total effect of temperature and light we get

$$\begin{aligned} SS_{\text{temperature + light}} &= 12 \cdot ((27.92 - 31.01)^2 + (31.42 - 31.01)^2 + (38.75 - 31.01)^2 \\ &\quad +(27 - 31.01)^2 + (27.33 - 31.01)^2 + (33.67 - 31.01)^2) \\ &= 1275.863 \end{aligned}$$

and by subtraction we then get the sum of squares for the interaction:

$$SS_{\text{temp : light}} = 1275.863 - 203.2128 - 1015.97 = 56.6802$$

The ANOVA table now becomes

	SS	D.f.	M.sq.	F	p
Temperature	203.2128	1	203.2128		
Lighting	1015.97	2	507.985		
Smell	889.49	3	296.497		
Temp:Light	56.6802	2	28.34	0.84	$p > 0.025$
Residuals	2186.633	65	33.66		
Total	4352.986	71			

and we conclude that the interaction is NOT significant, and should not be included in the model.

(d) According to the model, the residuals shoul be (approximately) independent, and normally distributed. The top two plots show no indications that the residuals depend in any way on the light levels or smell types. The bottom left plot shows that the residuals appear to be normally distributed. However, the bottom right plot indicates that residuals from experiments performed late are larger than the other residuals. Thus there is a time effect. Apparently, sales increase with time independently of the levels of the factors, and one should probably include this effect in the model.

5. (a) Except for the different catalysts, there will be in each experiment a number of factors, or conditions, that Alizadeh can control, and which are likely to influence the yield. Generally, she should try to keep these as constant as possible, to increase the possibility that she can get useful results from her experiments. There will also be

a number of factors, or conditions, that is likely to influence the result, and which are not under the control of Alizadeh. The best general way to deal with such factors is to use randomization. Thus Alizadeh should perform the 12 experiments in a randomized order.

(b) The test statistic of the hypothesis test is

$$\frac{12925.2}{2549.5} = 5.0697$$

which should be compared with an F distribution with 5 and 5 degrees of freedom. Such a comparison shows that the the probability that such a distribution is above 5.0697 is slightly below 0.5, thus the p-value, which is twice this, is slightly below 0.1. According to this test, one can then not reject the null hypothesis that the variances of the underlying normal distributions are the same.

(c) The pooled variance becomes

$$s_p^2 = \frac{5 \cdot 2549.5 + 5 \cdot 12925.2}{5+5} = 7737.35$$

the difference in expected yields has distribution

$$\mathrm{t}\left(1033 - 1017.5, 6 + 6 - 2, \log\left(\sqrt{7737.35}\sqrt{\frac{1}{6} + \frac{1}{6}}\right)\right) = \mathrm{t}\left(15.5, 10, \log(50.785)\right)$$

This gives the 90% credibility interval

$$\begin{aligned} & [15.5 - t_{10,0.05} \cdot 50.785, 15.5 + t_{10,0.05} \cdot 50.785] \\ = \; & [15.5 - 1.8125 \cdot 50.785, 15.5 - 1.8125 \cdot 50.785] \\ = \; & [-76, 107] \end{aligned}$$

(d) In this case, it is important that the effect of the decreasign yields is balanced as well as possible against the effect of the change in catalyst. Thus, instead of doing the 12 experiments in a randomized order, she might do them in pairs, with one catalyst of each type in each pair, ordering the pairs so that in three of the pairs, the old catalyst is used first and in the remaining pairs, the new catalyst is used first.