Petter Mostad
Matematisk Statistik
Chalmers

Suggested solution for exam in
MSA830: Statistical Analysis and Experimental Design
26 October 2012

1. (a) We first compute the probability that all bottles are properly cleaned. This probability is $(1-0.04)^{20} = 0.442$. We can also use a Binomial distribution with probability 0.04, obtaining

$$\binom{20}{0} 0.04^0 \cdot (1-0.04)^{20-0} = 0.442$$

The probability that one or more bottle is not properly cleaned is then $1 - 0.442 = 0.558$.

(b) We can use a Normal approximation. This approximation has expectation $50 \cdot 0.558 = 27.9$ and variance $50 \cdot 0.558 \cdot (1 - 0.558) = 12.3318$. So we compare the number

$$\frac{19.5 - 27.9}{\sqrt{12.3318}} = -2.392$$

According to the normal probability table, the probability to be above 2.39 is 0.00842, so the probability to be above -2.391992 is $1 - 0.00842 = 0.99158$, so there is approximately a 99% probability that 20 or more cases will be problem cases.

2. (a) One way to compute is

$$\begin{aligned}\pi(\text{C shown by host}) &= \pi(\text{C shown by host} \mid \text{prize in A})\pi(\text{prize in A}) \\ &\quad + \pi(\text{C shown by host} \mid \text{prize in B})\pi(\text{prize in B}) \\ &\quad + \pi(\text{C shown by host} \mid \text{prize in C})\pi(\text{prize in C}) \\ &= 0.5 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} = 0.5\end{aligned}$$

The probability that the host will show the contents of box C is 0.5. Alternatively, one can note that the problem is completely symmetric with respect to box B and C, and as the host must choose one of them, the probability for each is 0.5.

(b)

$$\begin{aligned}\pi(\text{prize in A} \mid \text{C shown by host}) &= \frac{\pi(\text{C shown by host} \mid \text{prize in A})\pi(\text{prize in A})}{\pi(\text{C shown by host})} \\ &= \frac{0.5 \cdot \frac{1}{3}}{0.5} = \frac{1}{3}\end{aligned}$$

The probability that the prize is in box A, given that the host has chosen to show box C, is $\frac{1}{3}$.

(c) As the probability that the prize is in box A is $\frac{1}{3}$, and the prize is not in box C, the probability that it is in box B must be $1 - \frac{1}{3} = \frac{2}{3}$. It would be advantageous for the contestant to change his choice from box A to box B.

3. (a) Note that the data is paired, so to get the most information out of the data, we should analyze the 5 differences between the types of feed. Computing the results for Billy's feed minus the traditional feed, we get

$$21, -6, 4, 36, -20.$$

It seems fairly reasonable to assume that this data comes from a normal distribution. A credibility interval for the expectation of this distribution shows what the data tells us about the difference in expected weight measurements for the different types of feed.

(b) The mean and sample variance of the 5 numbers above are 7 and 486, respectively. The expectation of the normal distribution from which the numbers come has distribution

$$\mathrm{t}(7, 4, \log(\sqrt{486/5})) = \mathrm{t}(7, 4, \log(9.86))$$

A 95% credibility interval becomes

$$[7 - t_{0.025,4}9.86, 7 + t_{0.025,4}9.86] = [7 - 2.7764 \cdot 9.86, 7 + 2.7764 \cdot 9.86] = [-20.38, 34.38].$$

(c) Using the t distribution found above, we must compare

$$\frac{7 - 0}{9.86} = 0.7099$$

with the table for the standard t distribution with 4 degrees of freedom. The table shows that the probability for such a distribution to be above 0.7099 above 0.25, so the probability asked for is above 0.5.

4. Note: There was a printing mistake in the exam question: The true grand average of the data is 46.25, not 45.25. We first show the results as they appear using 45.25 as the grand average, then the results as they appear using the 46.25 grand average.

(a) The effect sum of squares become

$$SS_{\text{colors}} = 4 \cdot (43.25 - 45.25)^2 + 4 \cdot (42.25 - 45.25)^2 + 4 \cdot (50 - 45.25)^2 + 4 \cdot (49.5 - 45.25)^2 = 214.5$$

and

$$SS_{\text{print}} = 4 \cdot (43.25 - 45.25)^2 + 4 \cdot (42.25 - 45.25)^2 + 4 \cdot (50 - 45.25)^2 + 4 \cdot (49.5 - 45.25)^2 = 67.5$$

while the total sum of squares becomes $SS_{\text{Total}} = 15 \cdot 23.8 = 357$. Thus the ANOVA table becomes

	SS	D.f.	M.sq.	F	p
Colors	214.5	3	71.5	8.58	$p < 0.01$
Print	67.5	3	22.5	2.7	$0.1 < p < 0.25$
Residuals	75	9	8.33333		
Total	357	15			

The table shows that colors have a significant influence on the sales, while the print does not.

(b) Dropping Colors, the ANOVA table becomes

	SS	D.f.	M.sq.	F	p
Print	67.5	3	22.5	0.9326	$p > 0.25$
Residuals	289.5	12	24.125		
Total	357	15			

The conclusion is still that the print type does not have a significant effect on the sales.

(c) In this situation, we assume that the sales for all the four prints are normally distributed with the same distribution standard deviations. In this model, we found above that the sum of squares of residuals $SS_{\text{Residuals}}$ is 289.5. The difference between the expected sales between print A and print D has distribution

$$\text{t}\left(48-47, 16-4, \log\sqrt{\frac{289.5}{16-4}\left(\frac{1}{4}+\frac{1}{4}\right)}\right) = \text{t}\left(1, 12, \log(3.4731)\right)$$

and a 95% credibility interval becomes

$$[1-t_{0.025,12}3.4731, 1+t_{0.25,12}3.4731] = [1-2.1788\cdot 3.4731, 1+2.1788\cdot 3.4731] = [-6.57, 8.57]$$

(d) The assumptions are that of a linear model without interaction, i.e., that the sales for each combination of color and print is normally distributed with an expectation that is a linear combination effects for color and print, and with a distribution standard deviation that is the same for all combinations of color and print. The best way to study whether these results are reasonable is to study the residuals for the model: They should be approximately normally distributed, and independent. Plotting the residuals against each of the predictors, and against time, are some ways that departures from this independence can be detected.

We now continue with the results as they become when the true grand average 46.25 is used:

(a) The effect sum of squares become

$$SS_{\text{colors}} = 4\cdot(43.25-46.25)^2+4\cdot(42.25-46.25)^2+4\cdot(50-46.25)^2+4\cdot(49.5-46.25)^2 = 198.5$$

and

$$SS_{\text{print}} = 4\cdot(43.25-46.25)^2+4\cdot(42.25-46.25)^2+4\cdot(50-46.25)^2+4\cdot(49.5-46.25)^2 = 51.5$$

while the total sum of squares becomes $SS_{\text{Total}} = 15\cdot 23.8 = 357$. Thus the ANOVA table becomes

	SS	D.f.	M.sq.	F	p
Colors	198.5	3	66.1667	5.57	$0.01 < p < 0.025$
Print	51.5	3	17.1667	1.44	$p > 0.25$
Residuals	107	9	11.8889		
Total	357	15			

The table shows that colors have a significant influence on the sales, while the print does not.

(b) Dropping Colors, the ANOVA table becomes

	SS	D.f.	M.sq.	F	p
Print	51.5	3	17.1667	0.6743	$p > 0.25$
Residuals	305.5	12	25.4583		
Total	357	15			

The conclusion is still that the print type does not have a significant effect on the sales.

(c) In this situation, we assume that the sales for all the four prints are normally distributed with the same distribution standard deviations. In this model, we found above that the sum of squares of residuals $SS_{\text{Residuals}}$ is 305.5. The difference between the expected sales between print A and print D has distribution

$$\mathrm{t}\left(48-47, 16-4, \log\sqrt{\frac{305.5}{16-4}\left(\frac{1}{4}+\frac{1}{4}\right)}\right) = \mathrm{t}\left(1, 12, \log(3.5678)\right)$$

and a 95% credibility interval becomes

$$[1-t_{0.025,12}3.5678,\ 1+t_{0.25,12}3.5678] = [1-2.1788\cdot 3.5678,\ 1+2.1788\cdot 3.5678] = [-6.77, 8.77]$$

(d) The assumptions are that of a linear model without interaction, i.e., that the sales for each combination of color and print is normally distributed with an expectation that is a linear combination effects for color and print, and with a distribution standard deviation that is the same for all combinations of color and print. The best way to study whether these assumptions are reasonable is to study the residuals for the model: They should be approximately normally distributed, and independent. Plotting the residuals against each of the predictors, and against time, are some ways that departures from this independence can be detected.

5. (a) A possible fractional factorial plan for Sally to follow is

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

This plan has been generated by starting with a full factorial design for A, B, C, and D, and defining $E = ABC$ and $F = CD$. With this design, Sally will for example not get information about the interaction between C and D independently from the information about F. On the other hand, she whould be able to for example get information about the interaction between A and B independently from other main effects.

(b) Apart from the factors she is investigating, there will be a number of other factors which could influence the life-lengths of her flowers, and which she can influence. These factors should generally be kept as constant as possible over her experiments. An alternative might be to use blocking for some such factors. Sally should also randomize the order in which she runs her 16 experiments, to avoid that time dependent factors are confounded with the factors she is investigating.

6. (a) The least squares estimates are

$$\begin{aligned}\widehat{\beta_2} &= \frac{n\sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n\sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \\ &= \frac{8 \cdot 2911.7 - 49 \cdot 458.5}{8 \cdot 325 - 49^2} \\ &= 4.156281\end{aligned}$$

and

$$\widehat{\beta_1} = \frac{458.5}{8} - 4.156281 \cdot \frac{49}{8} = 31.85528$$

(b) For example, the residuals for the two first observations listed in the data table are

$$\begin{aligned}r_1 &= 41.3 - (31.85528 + 4.156281 \cdot 3) = -3.0241 \\ r_2 &= 50.0 - (31.85528 + 4.156281 \cdot 5) = -2.6367\end{aligned}$$

The remaining residuals are

$$\begin{aligned}r_3 &= 56.5 - (31.85528 + 4.156281 \cdot 5) = 3.9633 \\ r_4 &= 56.7 - (31.85528 + 4.156281 \cdot 6) = -0.0929 \\ r_5 &= 59.1 - (31.85528 + 4.156281 \cdot 6) = 2.3070 \\ r_6 &= 62.6 - (31.85528 + 4.156281 \cdot 7) = 1.6508 \\ r_7 &= 68.0 - (31.85528 + 4.156281 \cdot 8) = 2.8945 \\ r_8 &= 64.2 - (31.85528 + 4.156281 \cdot 9) = -5.0618\end{aligned}$$

(c) The model is a linear model, were we assume that

$$y_i = \beta_1 + \beta_2 x_i + \epsilon_i$$

where the ϵ_i are assumed to be independent and normally distributed with expectation zero and the same scale. The residuals estimate the ϵ_i, so one may use these to theck if they are approximately normally distributed and independent.

(d) The logged standard deviation for the distribution of the ϵ_i has distribution

$$\text{ExpGamma}\left(\frac{8-2}{2}, \frac{1}{2} \cdot 73.8612, -2\right) = \text{ExpGamma}\,(3, 36.9306, -2)$$

and a 95% credibility interval for the standard deviation then becomes

$$\left[\sqrt{\frac{2 \cdot 36.9306}{14.449}}, \sqrt{\frac{2 \cdot 36.9306}{1.237}}\right] = [2.26, 7.73]$$

(e) The length of the credibility interval where he adds 12 units of his chemical will be larger than the length of the credibility interval where he adds 6 uints of his chemical. The reason is that the credibility intervals reflect both the variability around the regression line $y = \beta_1 + \beta_2 x$ and the uncertainty of the line itself at the value x. The first type of variability will be the same for all observations, whereas, because of the uncertainty in the slope β_2, the uncertainty of the line will be larger the further x is from the mean of the observed x values.