Petter Mostad
Matematisk Statistik
Chalmers

Suggested solution for exam in MSA830: Statistical Analysis and Experimental Design, March 2008

1. (a) The probability has a Binomial distribution:

$$P(X = 3) = \binom{8}{3} 0.2^3(1 - 0.2)^{8-3} = \frac{8!}{3!5!} 0.2^3 0.8^5 = 0.1468.$$

   (b) One way to compute is to use that $P(X = 3, Y = 3) = P(X = 3)P(Y = 3|X = 3)$. Given that exactly 3 experiments have outcome X, the probability for each remaining experiment to have outcome Y is $0.5/(1-0.2) = 0.625$ and the probability of outcome Z is $0.3/(1 - 0.2) = 0.375$. So we get

$$\begin{aligned} P(X = 3, Y = 3) &= P(X = 3)P(Y = 3|X = 3) \\ &= 0.1468 \cdot \frac{5!}{3!2!} 0.625^3 0.375^2 \\ &= 0.0504. \end{aligned}$$

   (c) The Binomial distribution for the number of X outcomes has expectation $30 \cdot 0.2 = 6$ and variance $30 \cdot 0.2 \cdot (1 - 0.2) = 4.8$. Thus X is approximately normally distributed with expectation 6 and varianc 4.8. We can compare $\frac{10-6}{\sqrt{4.8}} = 1.826$ with quantiles of a normal distribution, and we get the approximate probability 0.03 from Table A. More accurately, we can compare $\frac{9.5-6}{\sqrt{4.8}} = 1.60$ with quantiles of a normal distribution to get an approximate probability 0.05. (The exact probability using the Binomial distribution is 0.061).

2. (a) One possibiltiy is to use a paired t-test. We then analyze the 5 differences $9 - 4 = 5$, $6-5 = 1$, $7-3 = 4$, $6-7 = -1$, and $5-4 = 1$. We have to assume that these differences come from a normally distributed population, and that they are independent. Our null hypothesis is that this distribution has zero expectation; the alternative hypothesis is that the expectation is nonzero. The mean and variance of the 5 differences is 2 and 6, respectively, and the test statistic becomes

$$\frac{2}{\sqrt{6/5}} = 1.826.$$

   Comparing this with a t-distribution with 4 degrees of freedom, using table B1, we see that the p-value is greater than 0.1, and we cannot reject the null hypothesis that the paints have the same durability.

   (b) He should choose procedure II. The reason is that his data is paired (he has used both paints on each house) and this should be reflected in his computations.

3. (a) The main effect of A can be computed as

$$\frac{-4 - 7 - 5 - 3 + 1 + 4 + 2 + 4}{4} = -2.$$

The signs for computing the interaction effect between B and C can be found by multiplying the signs of the B and C column. The resulting effect calculation is

$$\frac{+4-7-5+3+1-4-2+4}{4} = -1.5.$$

(b) 2^{4-1}_{IV}. A generating relation is $ABCD = I$.

(c) The pooled variance estimate is

$$\frac{3+2+5+3+4+2+3+2}{8} = 3.$$

As 4 experiments are done for each setting of the factors, each effect is computed as an average of $4 \cdot 4 = 16$ experiments minus an average of $4 \cdot 4 = 16$ other experiments. Thus the estimated standard error is

$$\sqrt{3.0\left(\frac{1}{16}+\frac{1}{16}\right)} = 0.61.$$

(d) The mean yield is

$$\frac{4+7+5+3+1+4+2+4}{8} = 3.75.$$

As this estimate is computed as an average of 32 observations, its estimated variance is $3.0/32 = 0.09375$. Thus a 95% confidence interval becomes

$$3.75 \pm t_{3\cdot 8,0.025}\sqrt{0.09375} = 3.75 \pm 2.064 \cdot 0.306 = 3.75 \pm 0.63$$

using table B1.

(e) There is a problem. The plot shows that the observations are autocorrelated, i.e., that each observation is dependent on the previous. This conflicts with the necessary assumption done in the analysis above, that the observations are independent.

(f) If the experiments had been done in a randomized order, the experiments for each factor setting are likely to have been more different, and thus he is likely to have found a larger sample variance.

4. (a) The average of the 9 observations is 12. Subtracting this mean from all the observations, and writing the result as a sum of a group mean and a residual, we get

$$\begin{bmatrix} 2 & -5 & 1 \\ 1 & -3 & 1 \\ 0 & -1 & 4 \end{bmatrix} = \begin{bmatrix} 1 & -3 & 2 \\ 1 & -3 & 2 \\ 1 & -3 & 2 \end{bmatrix} + \begin{bmatrix} 1 & -2 & -1 \\ 0 & 0 & -1 \\ -1 & 2 & 2 \end{bmatrix}.$$

An ANOVA table is given as

Source of variation	Sum of squares	d. f.	Mean square	
Between treatments	$s_T = 42$	$\nu_T = 2$	$m_T = 21$	7.875
Within treatments	$s_R = 16$	$\nu_R = 6$	$m_R = 2.67$	
Total	$s_D = 58$	$\nu_D = 8$		

Comparing 7.875 with an F distribution with 2 and 6 degrees of freedom, using Table D, we get that the p-value is between 0.05 and 0.01.

(b) She should use an unpaired t-test. She should choose a two-sided test, as she has no prior information that type B1 is better or worse than type B2.

(c) She should choose test schedule III. If the efficiency measurements are influenced by the day of the week, this will reduce the value of results using test schedule I. If the efficiency measurements are influenced by the time of day, this will reduce the value of results using test scheduke III. Test schedule III is balanced so that influence of day of week, or time of day, will effect the measurements for each brand equally.

5. (a) The quantity minimized is

$$\sum_{i=1}^{20} \epsilon_i^2.$$

(b) He needs to assume that the cell growth values are a linear function of the temperature plus an error term, and that this error term is independent for all observations, and normally distributed with expectation zero and the same variance for all observations.

(c) There is a problem. When fitting a line to the plot, it is clear that the residuals, i.e., differences between observed values and values predicted by the line, are not independent, which they need to be in order for the confidence interval computations to be correct.

(d) The sum of squares of residuals would be smallest for the last model, and largest for the first model. This is because each model is contained in the following model. For example, when comparing the second and third model, the fitted values for the parameters in the second model could also be used in the third, with $\beta_2 = 0$. Thus the value of $\sum_{i=1}^{20} \epsilon_i^2$ found in the third model when *optimal* parameters are found must be smaller than the value of the sum for optimal parameters in the second model.