Petter Mostad
Matematisk Statistik
Chalmers

Suggested solution for exam in
MSG830: Statistical Analysis and Experimental Design
15 March 2013

1. (a) (i)

   (b) (iii) The means can not be read from the plot.

   (c) (iii) The sums can not be read from the plot.

   (d) (ii) In fact, the smallest value in A is larger than the smallest value in B.

   (e) (iii) The variances can not be read from the plot. (In fact, one can determine from teh plot which interval each variance is in, but these intervals overlap.)

   (f) (i)

2. The null hypothesis is that people in Gothenburg will select the funniest joke with the same frequencies as the rest of Sweden. The test statistic becomes

$$\chi^2 = \frac{(5 - 20 \cdot 0.43)^2}{20 \cdot 0.43} + \frac{(6 - 20 \cdot 0.12)^2}{20 \cdot 0.12} + \frac{(9 - 20 \cdot 0.45)^2}{20 \cdot 0.45} = 6.906977$$

   It should be compared with a χ^2 distribution with 2 degrees of freedom. The cutoff value for such a distribution is 5.991. Thus the p-value is less than 0.05, and Andrew has shown that people from Gothenburg have a significantly different humor than people generally in Sweden.

3. (a) The probability can be found with the Binomial distribution:

$$\binom{11}{3} 0.1^3 \cdot 0.9^{11-3} = \frac{11 \cdot 10 \cdot 9}{1 \cdot 2 \cdot 3} 0.001 \cdot 0.4304672 = 0.07102709$$

   So the probability that he will win on exactly 3 of his 11 tickets is about 7%.

   (b) We can use a normal approximation. The Binomial distribution in question has expectation $120 \cdot 0.1 = 12$ and variance $120 \cdot 0.1 \cdot (1 - 0.1) = 10.8$. Thus we need to compare

$$\frac{19.5 - 12}{\sqrt{10.8}} = 2.282177$$

   with the normal distribution. We find from the table the probability 0.0113, so the probability to win on 20 or more tickets is approximately 1%.

4. Using Bayes formula, we can write

$$\begin{aligned} P(A \mid \text{coppertrace}) &= \frac{P(\text{coppertrace} \mid A)P(A)}{P(\text{coppertrace})} \\ &= \frac{P(\text{coppertrace} \mid A)P(A)}{P(\text{coppertrace} \mid A)P(A) + P(\text{coppertrace} \mid B)P(B)} \\ &= \frac{0.08 \cdot 0.3}{0.08 \cdot 0.3 + 0.006 \cdot 0.7} = 0.8510638 \end{aligned}$$

   So the probability that the rock sample is from location A is about 85%.

5. There are the following major problems with Sara's experimental design:

   - Two of the factors she is investigating, light and temperature in her restaurant, are confounded: Whenever one is changed, she also changes the other, so that if an effect of the change is found, it would be impossible to tell whether it is due to the temperature or the light change.
   - The three factors price, light, and temperature are all confounded with the effect of the weekdays: Sales in a lunch restaurant can be expected to be different at different weekdays. For Sara, this effect would be impossible to separate from the effects of the three factors. In particular, whether or not possible good sales on fridays were due to low prices or that people like to eat out on Fridays would not be known.
   - The fourth factor, adverticing, is confounded with the longer term time effect in a very unfortunate way. Sara should expect a long term effect of her restaurant getting gradually better known. Even more importantly, she should expect that a Swedish lunch restaurant has better sales in the months April and May than in the months June and July. Bot these effects would be confounded with any effect the advertising might have.

   What would be a better experimental plan would of course depend on what is practical. It would be very difficult to separate any effect of advertising from a general time effect. To measure the effect of advertising, Sara could instead possibly try to ask a random selection of her customers about how they had found her restaurant.

   The effects of the other three factors could more easily be studied with experimentation. First of all, all 8 possible combinations of the 2 levels of each of these 3 factors should occur in her experimental plan. However, Sara might for economical reasons limit the number of days when she drastically reduces her prices. To decide which day she should use which combination of the factor levels, the best approach might be to use randomization. A possible alternative could be to use blocking in relation to the weekdays, i.e., to do the randomization conditional on each weekday having (approximately) the same number of days with the various factor level combinations.

6. A possible fractional factorial experimental plan where 5 two-level factors are investigated in 8 experimental runs is given below:

A_1	A_2	A_3	A_4	A_5
-	-	-	+	-
-	-	+	+	+
-	+	-	-	+
-	+	+	-	-
+	-	-	-	+
+	-	+	-	-
+	+	-	+	-
+	+	+	+	+

Here, the A_4 column has been created by multiplying the A_1 and A_2 columns, while the A_5 column has been created by multiplying columns A_1, A_2 and A_3.

7. A type 1 error happens when the hypothesis is true, but it is rejected in a hypothesis test. A type 2 error happens when the hypothesis is false, but it is not rejected in a hypothesis test.

8. (At least) two possible solutions can be given here. First, one can use the assumption that that the two sets of observations are random samples from normal distributions with equal but unknown variances. Then we first compute the pooled variance as

$$s_p^2 = \frac{(n-1)s_x^2 + (m-1)s_y^2}{n-1+m-1} = \frac{15 \cdot 4.2 + 10 \cdot 3.8}{15+10} = 4.04$$

We then get the 95% confidence interval

$$\begin{aligned}
& \left[\overline{y} - \overline{x} - t_{n+m-2,0.025} s_p \sqrt{\frac{1}{n} + \frac{1}{m}}, \overline{y} - \overline{x} - t_{n+m-2,0.025} s_p \sqrt{\frac{1}{n} + \frac{1}{m}}\right] \\
= & \left[15.7 - 14.3 - t_{25,0.025} s_p \sqrt{\frac{1}{16} + \frac{1}{11}}, 15.7 - 14.3 + t_{25,0.025} s_p \sqrt{\frac{1}{16} + \frac{1}{11}}\right] \\
= & [1.4 - 2.0595 \cdot 2.009975 \cdot 0.3916747, 1.4 + 2.0595 \cdot 2.009975 \cdot 0.3916747] \\
= & [-0.22, 3.02]
\end{aligned}$$

Secondly, one can use the assumption that the two sets of observations are random samples from normal distributions with unknown variances which are not necessarily equal. We then start with computing the degrees of freedom with

$$\nu = \frac{\left(\frac{s_x^2}{n} + \frac{s_y^2}{m}\right)^2}{\frac{(s_x^2/n)^2}{n-1} + \frac{(s_y^2/m)^2}{m-1}} = \frac{\left(\frac{4.2}{16} + \frac{3.8}{11}\right)^2}{\frac{(4.2/16)^2}{15} + \frac{(3.8/11)^2}{10}} = 22.36 \approx 22$$

The approximate 95% confidence interval now becomes

$$\begin{aligned}
&\left[\bar{y}-\bar{x}-t_{\nu,0.025}\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}, \bar{y}-\bar{x}-t_{\nu,0.025}\sqrt{\frac{s_x^2}{n}+\frac{s_y^2}{m}}\right] \\
= &\left[15.7-14.3-t_{22,0.025}\sqrt{\frac{4.2}{16}+\frac{3.8}{11}}, 15.7-14.3+t_{22,0.025}s_p\sqrt{\frac{4.2}{16}+\frac{3.8}{11}}\right] \\
= &\ [1.4-2.074\cdot 0.7797144, 1.4+2.074\cdot 0.7797144] \\
= &\ [-0.22, 3.02]
\end{aligned}$$

9. (a) A: The correlation must be zero, as the points are perfectly symmetric around the y-axis. Thus the mean of the x-values is zero, and any point contributing positively to the correlation is balanced with a point contributing negatively to the correlation.

   B: One cannot tell from the picture exactly what the correlation is, although it must be negative.

   C: The correlation must be exactly 1, as the points lie on a perfectly straight line with a positive slope.

   D: As the 4 points are symmetrically placed, one can again argue by symmetry that the correlation cannot be positive or negative, so it must be zero.

   (b) A: The Spearman rank correlation must be zero: This is again by the symmetry of the picture.

   B: The Spearman rank correlation must be -1. This is because the ranking of the x-values is exactly in the opposite order as the ranking of the y-values.

   C: The Spearman rank correlation must be 1: This is because the ranking of the x-values is exactly in the same order as the ranking of the y-values.

   D: Because of symmetry, the Spearman rank correlation can neither be positive or negative, so it must be exactly zero.

10. It is plot X: For example, this is because in Figure 3, A and B, together with E and C, are clearly the two pairs of points that are closest together. Only in plot X do these two pairs appear as clusters at the bottom of the clustering.

11. (a) It is plot B that represents a linear regression: The linear regression line should be so that the sum of the squares of the *vertical* distances from the line to the data points is minimized. This is clearly true for the line in plot B, while it is clearly not true for the line in plot A.

    (b) For Plot B, the residuals are 0, -1, 1, and 0.