Petter Mostad
Matematisk Statistik
Chalmers

Suggested solution for exam in MSG830: Statistical Analysis and Experimental Design 20 August 2013

1. (a) The probability is 9/20 = 0.45.

   (b) The conditional probability becomes $\frac{7/20}{(4+7)/20} = \frac{7}{4+7} = 0.6364$.

2. We have that

$$\begin{aligned} \Pr(\text{A painter}) &= \frac{150}{150+400} = 0.2727273 \\ \Pr(\text{pigment} \mid \text{A painter}) &= 0.5 \\ \Pr(\text{pigment} \mid \text{B painter}) &= \frac{1}{20} = 0.05 \end{aligned}$$

   Thus Bayes formula gives

$$\begin{aligned} &\Pr(\text{A painter} \mid \text{pigment}) \\ =\ & \frac{\Pr(\text{pigment} \mid \text{A painter})\Pr(\text{A painter})}{\Pr(\text{pigment} \mid \text{A painter})\Pr(\text{A painter}) + \Pr(\text{pigment} \mid \text{B painter})\Pr(\text{B painter})} \\ =\ & \frac{0.5 \cdot 0.2727273}{0.5 \cdot 0.2727273 + 0.05 \cdot (1 - 0.2727273)} \\ =\ & 0.789 \end{aligned}$$

   So the probability is about 80 percent.

3. The expected counts for the 20 new olives would be

$$\begin{aligned} 20 \cdot 13/40 &= 6.5 \\ 20 \cdot 5/40 &= 2.5 \\ 20 \cdot 11/40 &= 5.5 \\ 20 \cdot 11/40 &= 5.5 \end{aligned}$$

   The goodness-of-fit test statistic becomes

$$\chi^2 = \frac{(4-6.5)^2}{6.5} + \frac{(4-2.5)^2}{2.5} + \frac{(9-5.5)^2}{5.5} + \frac{(3-5.5)^2}{5.5} = 5.2252$$

   and this should be compared with a χ^2 distribution with 3 degrees of freedom. We find from the table that the p-value is above 0.05. So one would say that the tastes of the olives from the new tree has not been found to differ significantly from those of the general orchard.

4. (a) To reduce the effect of the amount of gold varying between creeks, Bart should use a paired test. The increase in found gold in the 6 creeks when moving from the old to the new pan was $-4, 6, 18, 13, 28, 0$. The average of these numbers is 10.167, so Bart finds on average 10.167 more gold with the new pan. The sample standard deviation of the differences is 11.9066, and a 95% confidence interval becomes

$$\begin{aligned} & \left[10.167 - t_{5,0.025}\frac{11.9066}{\sqrt{6}}, 10.167 + t_{5,0.025}\frac{11.9066}{\sqrt{6}}\right] \\ = & \left[10.167 - 2.5706\frac{11.9066}{\sqrt{6}}, 10.167 + 2.5706\frac{11.9066}{\sqrt{6}}\right] \\ = & [-2.33, 22.66] \end{aligned}$$

(b) To use the interval over as a confidence interval one has to assume that the observed differences are independent of each other, and from a normal distribution. From the values of the differences, there are no particular reasons to doubt that they are from a normal distribution. However, from the context it would be reasonable to expect that some creeks have no gold, no matter what pan is used. This makes it doubtful that the assumption can be trusted.

(c) The null hypothesis would be that the differences are from a normal distribution with expectation zero. The reasonable alternative hypothesis would be that the new pans might be either better or worse, i.e., that the differences come from a normal distribution with an expectation different from zero. The hypothesis would conclude that that one can reject the null hypothesis precisely when the 95% confidence interval does not contain zero. As the confidence interval does contain zero, the hypothesis test would not conclude that we can reject the null hypothesis.

(d) It is possible to use a non-parametric test where the comparison is made only based on the size-ordering of the observations. The test to use here would be a Wilcoxon paired-sample test.

5. (a) A possible 8-experiment fractional factorial experimental plan would be

Temperature	Enzyme	Mixerspeed	Light
Normal	X	Normal	No
Normal	X	High	Yes
Normal	Y	Normal	Yes
Normal	Y	High	No
High	X	Normal	Yes
High	X	High	No
High	Y	Normal	No
High	Y	High	Yes

(b) The major problem with Andrea's plan is that it is quite unbalanced. Only one experiment is run with high mixerspeed, and only one with light irradiation. Thus it will be very difficult to make conclusions about any effect of these factors. In the fractional factorial plan, the different factors, and combinations of factors, are all investigated equally often. Thus conclusions from such data would be more reliable. The fractional factorial experimental plan is to be recommended.

6. A dendrogram is a way to illustrate the results of a hierarchical clustering. It shows a tree-like structure, where any vertical line crossing the figure will indicate a particular way to cluster the clustered items. As the line moves downwards, the number of clusters increases stepwise. An illustration is Figure 1.

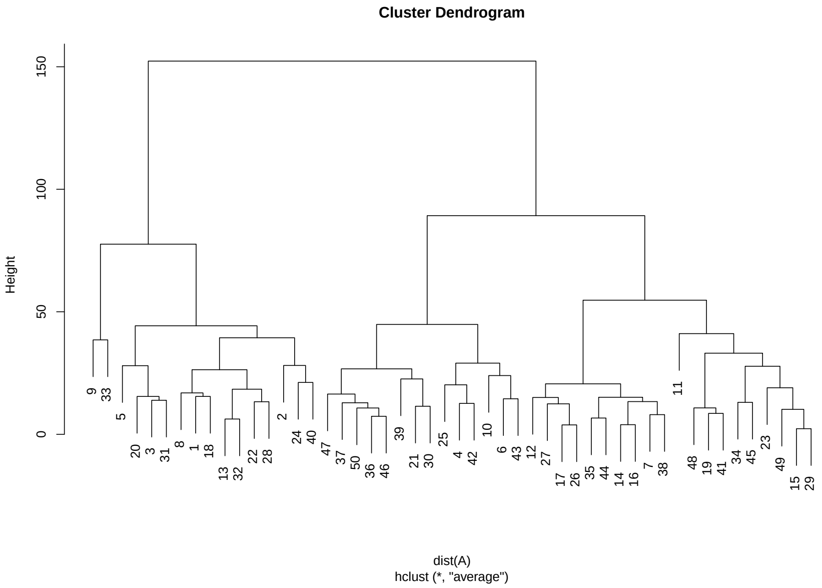

Figure 1: An example of a dendrogram

7. (a) From the data we get that

$$\begin{aligned} \sum x_i &= 4 \\ \sum y_i &= 5 \\ \sum x_i y_i &= 11 \\ \sum x_i^2 &= 10 \end{aligned}$$

which gives the simple linear regression parameters

$$\widehat{\beta_1} = \frac{3 \cdot 11 - 4 \cdot 5}{3 \cdot 10 - 4^2} = 0.9285714$$

and

$$\widehat{\beta_0} = \frac{5}{3} - 0.9285714 \cdot \frac{4}{3} = 0.4285715$$

(b) The assumptions are that the value of each observation y_i for each given x_i is equal to $\beta_0 + \beta_1 x_i + \epsilon_i$, where all ϵ_i are independent and from a normal distribution with expectation zero.

8. In a permutation test, the null hypothesis is that the data comes from some type of distribution with some symmetry properties, so that certain permutations of the data would have been equally likely to have been observed as the actual data. By computing the test statistic on not only the actual data but also on these permutations of it, one obtains a sample from the reference distribution of the test statistic under the null hypothesis, with which the actual value of the test statistic can be compared.

   For example, assume 10 children try out shoe sole material by using one shoe with old-type material and one with new material (where it is randomly selected which is which) for some period of time. If the difference between the new material wear and the old material wear are designated $d_1, \dots, d_{10}$, any addition of + or − signs to these numbers should yield a sequence of numbers that is equally likely under the null hypothesis that the shoe materials get the same amount of wear. Using the average of the numbers as the test statistic, one investigates in the permutation test how large $(d_1 + \cdots + d_{10})/10$ is compared with the numbers $(\pm d_1 \pm \cdots \pm d_{10})/10$ for all possible choices of signs.

9. The ANOVA table becomes

	Sum of squares	Deg. freedom	Mean squ.	F	p
Temperature	212.9	3	70.9667	3.8861	$0.01 < p < 0.025$
Pressure	235.8	4	58.95	3.2281	$0.01 < p < 0.025$
Residuals	949.6	52	18.2615		
Total	1398.3	59			

10. Roughly, the data is ordered and divided into 4 groups according to size, with approximately equal number of members in each group. From this, the median and the 25th and 75th percentiles of the data are computed. The three values are marked with vertical lines, and a box is drawn extending from the 25th to the 75th percentile. Then, lines are drawn from this box down to the smallest value and up to the larges value in the dataset. If some data values are so large or so small that they might be considered outliers, they are marked with circles, and the line from the box does not extend to them.