Petter Mostad
Matematisk Statistik
Chalmers

Suggested solution for exam in
MSA830: Statistical Analysis and Experimental Design
June 2010

1. (a) The mean μ of the distribution is distributed as[1]

$$\mu \sim \mathrm{t}\left(\overline{x}, 4, s_x^2/5\right)$$

where $\overline{x}$ and s_x^2 are the mean and variance of the observations for species X, respectively. We get

$$\overline{x} = 48.8$$

and

$$s_x^2 = 39.7$$

so

$$\mu \sim \mathrm{t}(48.8, 4, 7.94).$$

A 95% credibility interval for this distribution is given by

$$[48.8 - \sqrt{7.94} \cdot 2.776, 48.8 + \sqrt{7.94} \cdot 2.776] = [40.98, 56.62],$$

where 2.776 is found in the table for the t distribution: A t distribution with 4 degrees of freedom has 2.5% probability of being above 2.776.

(b) The precision τ is distributed as

$$\tau \sim \mathrm{Gamma}\left(\frac{4}{2}, \frac{4s_x^2}{2}\right) = \mathrm{Gamma}\,(2, 79.4)\,.$$

A 95% credibility interval for a Chi-square distribution with 4 degrees of freedom is $[0.48, 11.14]$. Thus, a 95% credibility interval for the distribution above is given by

$$[0.48/(2 \cdot 79.4), 11.14/(2 \cdot 79.4)] = [0.0030, 0.0702].$$

(c) The difference in means has distribution[2]

$$\mathrm{t}\left(\overline{x} - \overline{y}, 5 + 6 - 2, \left(\frac{1}{5} + \frac{1}{6}\right) s_p^2\right)$$

where $\overline{x} = 48.8$, $\overline{y} = 39.17$, and $s_p^2 = \frac{4s_x^2 + 5s_y^2}{4+5} = \frac{4 \cdot 39.7 + 5 \cdot 41.367}{9} = 40.63$, so the distribution becomes

$$\mathrm{t}\,(9.63, 9, 14.898)$$

A 95% credibility interval for this distribution becomes

$$[9.63 - \sqrt{14.898} \cdot 2.262, 9.63 + \sqrt{14.898} \cdot 2.262] = [0.90, 18.36].$$

[1] Using the notation of spring 2010, the last parameter in the t distribution would be the inverse of the one used below

[2] Using the notation of spring 2010, the last parameter in the t distribution would be the inverse of that used below

2. (a) She would need to experiment for $2^6 = 64$ days.

   (b) A possible experimental plan would be

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

   where the columns for E and F has been derived from the multiplications E=ABC and F=BCD, respectively.

   (c) As the results are influenced by a number of factors Tanya cannot influence, and as the sales may have a trend over the summer, it is important that Tanya randomize the order in which she does the experiments above. Also, for the factors that she can control, but which are not listed among the factors A-F, she should try to keep them as constant as possible over the experimental days. For example, she should not change her prices in the middle of the experimental period.

   (d) In the design above, and using the generating relations E=ABC and F=BCD, a little experimentation will show that the following two-way interactions are confounded with each others:

$$\begin{aligned} AB &= CE \\ AC &= BE \\ AD &= EF \\ AE &= BC = DF \\ AF &= DE \\ BD &= CF \\ BF &= CD \end{aligned}$$

3. (a) There are $4^6 = 4096$ possible sequences of length 6.

   (b) The three nucleotides of type T can be in any of the 6 positions. The number of ways to place them in this sequence is given by the binomial coefficient:

$$\binom{6}{3} = \frac{6!}{3!3!} = \frac{4 \cdot 5 \cdot 6}{1 \cdot 2 \cdot 3} = 20.$$

For each of these ways to place the nucleotides of type T, the remaining 3 places can be filled with either of the remaining three nucleotides. Each time, there are $3^3 = 27$ ways to do this. Thus the total number of seqences of length 6 containing exactly 3 nucleotides of type T is $20 \cdot 27 = 540$.

(c) The probability is given by the binomial distribution, where the probability of "success" (i.e., a T) is 0.15. Thus the sought probability is

$$\binom{10}{3} 0.15^3 0.85^{10-3} = \frac{8 \cdot 9 \cdot 10}{1 \cdot 2 \cdot 3} 0.15^3 0.85^7 = 0.1298$$

(d) One way to compute this is to say that first, the sequence needs to have 3 T's, then (conditionally on this) the remaining 7 places need to have 4 of type A, and then, the remaining 3 places must be filled with G. The probability for filling places with T is 0.15, while the probability of filling places with A, given that only A, C, and G are possible, is $0.15/(0.15 + 0.35 + 0.35) = 0.1765$. The probability of filling places with G, given that only G and C are possible, is $0.35/(0.35 + 0.35) = 0.5$. Thus, the sought probability is

$$\begin{aligned} & 0.1298 \cdot \binom{7}{4} 0.1765^4 (1 - 0.1765)^3 \cdot \binom{3}{3} 0.5^3 (1 - 0.5)^0 \\ = \; & 0.1298 \cdot 0.0190 \cdot 0.125 = 0.00003083 \end{aligned}$$

4. Write
A: Dan has observed species A.
B: Dan has observed species B.
behaviour: The birds have a certain behaviour.
Then we have $\Pr(A) = 0.75$, $\Pr(B) = 0.25$, $\Pr(\text{behaviour} \mid A) = 0.1$, and $\Pr(\text{behaviour} \mid B) = 0.9$. We can now use Bayes formula to write

$$\begin{aligned} \Pr(A \mid \text{behaviour}) &= \frac{\Pr(\text{behaviour} \mid A) \Pr(A)}{\Pr(\text{behaviour} \mid A) \Pr(A) + \Pr(\text{behaviour} \mid B) \Pr(B)} \\ &= \frac{0.1 \cdot 0.75}{0.1 \cdot 0.75 + 0.9 \cdot 0.25} \\ &= 0.25 \end{aligned}$$

So the probability that Dan has observed species A is 25%.

5. (a)

	Sum of squ.	Deg. freed.	Mean squ.	F value	p value
Shoes	7.8104	2	3.9052	5.5487	0.05<p<0.1
Clothes	3.1599	2	1.5800	2.2450	0.1<p<0.25
Residuals	2.8151	4	0.7038		
Total	13.7854	8			

(b) As none of the p-values are below 0.05, one would traditionally conclude that Roza does not have enough evidence to support a conclusion that her shoes, or her clothes, influence her running times.

(c) The averages can be computed as
Clothes X: 32.8
Clothes Y: 32.3

Clothes Z: 33.2
Shoe A: 31.37
Shoe B: 33.57
Shoe C: 33.37
Total: 32.77
Thus the sum of squares for shoes can be computed as

$$3 \cdot (31.37 - 32.77)^2 + 3 \cdot (33.57 - 32.77)^2 + 3 \cdot (33.37 - 32.77)^2 = 8.88$$

The sum o0f squares for clothes can be computed as

$$3 \cdot (32.8 - 32.77)^2 + 3 \cdot (32.2 - 32.77)^2 + 3 \cdot (33.2 - 32.77)^2 = 1.2201$$

The total sum of squares can be computed as

$$\begin{aligned} & (30.0 - 32.77)^2 + (31.1 - 32.77)^2 + (33.0 - 32.77)^2 \\ + \; & (33.5 - 32.77)^2 + (33.5 - 32.77)^2 + (33.7 - 32.77)^2 \\ + \; & (34.9 - 32.77)^2 + (32.3 - 32.77)^2 + (32.9 - 32.77)^2 \\ = \; & 17.22 \end{aligned}$$

Thus the first numbers in the ANOVA table become

	Sum of squ.	Deg. freed.	Mean squ.	F value	p value
Shoes	8.88				
Clothes	1.2201				
Residuals					
Total	17.22				

6. (a) The difference between the means is approximately distributed as[3]

$$\mathrm{t}\left(\overline{z} - \overline{y}, \nu, \frac{s_y^2}{6} + \frac{s_z^2}{6}\right),$$

where $\overline{y}$ and $\overline{z}$ are the averages for Type A and Type B of clothes, respectively, where s_y^2 and s_z^2 are the variances, respectively, and where ν is the degrees of freedom, given below. We have

$$\begin{aligned} \overline{y} &= 31.67 \\ \overline{z} &= 37.5 \\ s_y^2 &= 40.67 \\ s_z^2 &= 58.3 \end{aligned}$$

and the degrees of freedom is given by

$$\nu = \frac{\left(\frac{s_y^2}{6} + \frac{s_z^2}{6}\right)^2}{\frac{(s_y^2/6)^2}{5} + \frac{(s_z^2/6)^2}{5}} = 9.692$$

[3]Using the notation of spring 2010, the last parameter of the t distribution would be the inverse of the one used below

Thus the distribution for the difference in the means is approximately

$$\text{t}(5.833, 9.692, 16.495)$$

A 95% credibility interval for this distribution is given by

$$[5.883 - \sqrt{16.495} \cdot 2.228, 5.883 + \sqrt{16.495} \cdot 2.228] = [-3.17, 14.93]$$

where $[-2.228, 2.228]$ is a 95% credibility interval for the standard t distribution with 10 degrees of freedom.

(b) With the new information, Lurleen should do a paired t-test. This means that she should look at the *differences* between the observations for each person, and find the distribution for the mean of the distribution for such differences. The differences (Type B - Type A) are 2, 2, 3, 10, 6, and 12. The mean and variance of these numbers is 5.83 and 18.57, respectively. Assuming that the differences follow a normal distribution, the mean of this distribution is distributed as[4]

$$\text{t}(5.83, 5, 18.57/6) = \text{t}(5.83, 5, 3.095)$$

As a 95% credibility interval for a standard t distribution with 5 degrees of freedom is $[-2.571, 2.571]$, so a 95% credibility interval for our distribution is

$$[5.83 - \sqrt{3.095} \cdot 2.571, 5.83 + \sqrt{3.095} \cdot 2.571] = [1.31, 10.35]$$

[4]Using the notation of spring 2010, the last parameter of the t distribution would be the inverse of the one used below