Petter Mostad
Matematisk Statistik
Chalmers

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

1. (a) Statement 1 is true: The upper bound of the box for the A values is below the lower bound for the box for the B values. This means that 75% of the A values are below 75% of the B values, which is what statement 1 says. Statement 2 is false: If the variance was 148.3, the standard deviation, which is the square root of this number, would be above 10. However the total spread between the smallest and the largerst values in A is less than 10, so their standard deviation cannot be larger than 10. Statement 3 is also false: The values in B are generally larger, and so that condition is better for the plants.

   (b) The box plots give no strong indication that the values in each group are not normally distributed, so we use a t-test (although using a nonparametric Wilcoxon rank-sum test would be a good alternative). However, the box plot indicates that the spread, or variance, in the two groups are different, so we use the t-test that does not assume equal variances. We then compute

$$
\begin{aligned}
\overline{a} &= \sum_{i=1}^{20} a_i \\
\overline{b} &= \sum_{i=1}^{20} b_i \\
s_a^2 &= \frac{1}{19}\sum_{i=1}^{20}(a_i - \overline{a})^2 \\
s_b^2 &= \frac{1}{19}\sum_{i=1}^{20}(b_i - \overline{b})^2 \\
t &= \frac{\overline{b} - \overline{a}}{\sqrt{\frac{s_a^2}{20} + \frac{s_b^2}{20}}} \\
df &= \frac{\left(\frac{s_a^2}{20} + \frac{s_b^2}{20}\right)^2}{\frac{(s_a^2/20)^2}{19} + \frac{(s_b^2/20)^2}{19}},
\end{aligned}
$$

   (it is not required to remember the formula for df) and then compare the computed t value with a t-distribution with approximately df degrees of freedom. This would give us the probability of getting numbers larger than the computed t under the null hypothesis, and we multiply this probability by 2 to get the p-value, as we should use a two-sided test.

2. (a) The central limit effect is the effect that a sum of many independent random variables from the same distribution thends to be approximately normally distributed. The approximation improves as the number of variables in the sum increases. The variables can be from any distribution as long as it has an expectation and a variance.

(b) A normal probability plot is used to investigate whether a set of n values can reasonably be assumed to be a sample from a normal distribution. It is constructed by first sorting the n values, denoting them $x_1, \ldots, x_n$, and then constructing $y_1, \ldots, y_n$ so that these represent approximately the middle of the n quantiles in the standard normal distribution. The points $(x_1, y_1), \ldots, (x_n, y_n)$ are then plotted. If the points stay approximately on a straight line, the original data can be assumed to be a normal sample. This can be useful for example when checking whether the residuals in linear modelling are approximately normally distributed.

3. (a) For an efficient experimental plan, the columns in the table should be orthogonal. One way to achieve that is to complete the first three columns so that they consitute a 2^3 experimental plan on standard form. The column for W seems to be generated in the first six lines as the product of all the previous columns, so we generate it in the same way for the last to lines as well. This gives us the following lines:

$$\begin{array}{cccc} + & + & - & - \\ + & + & + & + \end{array}$$

This is an 2^{4-1}_{IV} experimental plan. As $W = XYU$, we have that for example $WX = YU$, so the interaction effect between W and X is confounded with the interaction effect between Y and U.

(b) The main effect of U is

$$\frac{21.3 + 10.1 + 14.3 + 19.9}{4} - \frac{14.2 + 4.7 + 6.0 + 5.2}{4} = 8.875$$

The interaction effect of X and Y is

$$\frac{14.2 + 21.3 + 5.2 + 19.9}{4} - \frac{4.7 + 10.1 + 6.0 + 14.3}{4} = 6.375.$$

(c) The pooled variance estimate is

$$s^2_{pooled} = \frac{2.5 + 2.0 + 0.7 + 1.7 + 0.3 + 1.2 + 0.7 + 1.3}{8} = 1.3,$$

with $8 \cdot (7 - 1) = 48$ degrees of freedom. The standard error of any of the estimates is

$$SE = \sqrt{s^2_{pooled}\left(\frac{1}{4 \cdot 7} + \frac{1}{4 \cdot 7}\right)} = \sqrt{1.3 \cdot \frac{1}{14}} = 0.305$$

and so the 95% confidence interval for the effect of U becomes

$$8.875 \pm t_{0.025,48} \cdot 0.305 \approx 8.875 \pm 2.01 \cdot 0.305 = 8.875 \pm 0.613,$$

where $t_{0.025,48} \approx 2.01$ is estimated from table II.

(d) We can extend the table into a full 2^4 factorial design, so that the full table looks like

X	Y	U	W
-	-	-	-
-	-	+	+
-	+	-	+
-	+	+	-
+	-	-	+
+	-	+	-
+	+	-	-
+	+	+	+
-	-	-	+
-	-	+	-
-	+	-	-
-	+	+	+
+	-	-	-
+	-	+	+
+	+	-	+
+	+	+	-

(The order of the lines can be different; the main point is that each 16 possible combination of + and - for the 4 factors occur in exactly one line).

4. (a) To make a computation, we need to assume that the defectiveness of a component is independent of the defectiveness of the other components produced that day. Whith this assumption, the number of defective components is Binomially distributed with parameters $n = 10$ and $p = 0.2$, and the probability of zero defects is given by

$$P(0) = \binom{10}{0} p^0(1-p)^{10} = \frac{10!}{0!10!}0.8^{10} = 0.8^{10} = 0.107.$$

The probability is about 10.7%.

(b) We could sum the Binomial probabilities for 6,7,8,9, and 10, and this gives the probability 0.00637. For a rough approximation, we can use a normal approximation. The Binomial distribution has an expectation of $0.2 \cdot 10 = 2$ and a variance of $0.2 \cdot (1 - 0.2) \cdot 10 = 1.6$. Thus we can approximately compare $(6 - 2)/\sqrt{1.6} = 3.16$ with a standard normal distribution. From the corresponding table, we get that the probability of being above this number is $1 - 0.99921 = 0.00079$. Thus this is the approximate probability (other formulas can give more accurate approximations).

(c) In the new situation, we may use a Poisson distribution with rate 2, as the event of a failure is now very rare. The probability of exactly two failures can now be computed as

$$P(2) = \frac{2^2 e^{-2}}{2!} = 0.271$$

The probability is approximately 27.1%.

5. (a) The table becomes

	Sums of squ.	D. freedom	Mean squ.	F	p
Tyre	41	2	20.5	23.81	<0.01
Biker	87	3	2.9		
Interaction	13	6	2.167		
Error	93	108	0.8611		
Total	234	119			

as the sums of squares and degrees of freedom for the Biker and Interaction rows are now added to the Error row. Jack should conclude that the type of tyre influences the braking distance.

(b) There are many factors that could influence the braking distance, which are not measured in this experiment, and which could vary with time. Examples of such factors are changing weather and how well the bikers do the braking. If the order of the experiments were not randomized, the influence of such factors could be confounded with the influence of factors that are changed in in a time sequence in the experiment.

(c) The table now becomes

	Sums of squ.	D. freedom	Mean squ.	F	p
Tyre	41	2	20.5	12.43	<0.01
Error	193	117	1.6496		
Total	234	119			

The p-value would increase, but it would still be below 0.01, and Jack could make the same conclusion.

6. (a) The means for the weight gains for feed A and feed B are

$$\frac{43 + 20 + 35 + 39}{4} = 34.25$$

and

$$\frac{40 + 41 + 34 + 49}{4} = 41,$$

respectively, and an estimate for the average added weight gain is then $41 - 34.25 = 6.75$. The sample standard deviation for feed A and feed B is

$$s_A^2 = \frac{1}{3}\left(8.75^2 + (-14.25)^2 + 0.75^2 + 4.75^2\right) = 100.92$$

and

$$s_B^2 = \frac{1}{3}\left(1^2 + 0^2 + (-7)^2 + 8^2\right) = 38,$$

respectively. If we assume that the population variance in the two groups is the same, we get a pooled variance estimate

$$s_p^2 = \frac{s_A^2 + s_B^2}{2} = 69.46,$$

and a 90% confidence interval can be written

$$6.75 \pm \sqrt{s_p^2\left(\frac{1}{4} + \frac{1}{4}\right)}t_{0.05,6} = 6.75 \pm 5.89 \cdot 1.943 = 6.75 \pm 11.44$$

(b) The changes are that we can use 8.0 instead of the pooled variance estimate, and that we can use the normal distribution instead of the t-distribution. The confidence interval would become

$$6.75 \pm \sqrt{8.0^2\left(\frac{1}{4}+\frac{1}{4}\right)}z_{0.05} = 6.75 \pm 5.66 \cdot 1.65 = 6.75 \pm 9.34$$

(c) The length of the confidence interval is given by

$$2 \cdot \sqrt{8.0^2\left(\frac{1}{n}+\frac{1}{n}\right)}z_{0.05},$$

where n is the number of pigs in each of the groups given feed A or feed B. If we want to find n so that

$$2 \cdot \sqrt{8.0^2\left(\frac{1}{n}+\frac{1}{n}\right)}z_{0.05} = \frac{1}{3} \cdot 2 \cdot \sqrt{8.0^2\left(\frac{1}{4}+\frac{1}{4}\right)}z_{0.05}$$

we get that

$$\begin{aligned}
\sqrt{8.0^2\left(\frac{1}{n}+\frac{1}{n}\right)} &= \frac{1}{3}\sqrt{8.0^2\left(\frac{1}{4}+\frac{1}{4}\right)} \\
8.0^2 \cdot \frac{2}{n} &= \frac{1}{9} \cdot 8.0^2 \cdot \frac{2}{4} \\
\frac{2}{n} &= \frac{2}{36} \\
n &= 36
\end{aligned}$$

In other words, in addition to the 4 pigs in each group she already has data for, she needs data for additional 32 pigs in each group.