Suggested solution to

exam in MSA830: Statistical Analysis and Experimental design

February 20th 2008, 8:30 – 12:30

1. a) The data is not paired, so an unpaired t-test is appropriate. The mean and variance for the cows with music is 118 and 29.6 respectively, while for the cows without music, it is 114 and 22. We can choose between a t-test where variances are assumed equal, and one where they are not. As 22 is close to 29.6, the data does not indicate that the population variances in the two groups are different. For a test where the variances are assumed equal, we get a pooled variance estimate of $\frac{5\cdot 29.6+5\cdot 22}{5+5}=25.8$ and a test statistic $\frac{118-114}{\sqrt{25.8(1/6+1/6)}}=1.36$ which should be compared with a t distribution with 10 degrees of freedom. Looking at the table for the t-distribution, and choosing a 2-sided t-test, we find that the p-value is slightly above 0.2, so we conclude that the music does not make a statistically significant difference. If we instead choose a t-test where variances are not assumed equal, the test statistic becomes $\frac{118-114}{\sqrt{29.6/6+22/6}}=1.36$ which should be compared with a t distribution with $\frac{(29.6/6+22/6)^2}{(29.6/6)^2/5+(22/6)^2/5}=9.79$ degrees of freedom. Again, choosing a two-sided test, the table for the t-distribution shows that the p-value is somewhere between 0.2 and 0.5, and we get the same conclusion.

   The additional assumption we must make in order to use the test is that the observations in each group are independent.

   b) If we continue with the assumption that the popluation variances are equal, the confidence interval becomes $118-114\pm 1.812\sqrt{25.8(1/6+1/6)}=4\pm 5.31$, where 1.812 is found in the table for the t-distribution with 10 degrees of freedom. Similarly if we do not assume equal variance, the 90% confidence interval becomes $118-114\pm 1.812\sqrt{29.6/6+22/6}=4\pm 5.31$ as before, as the degrees of freedom is still close to 10.

   c) He can now use a paired t-test, analyzing all the 12 differences. If the differences are $d_1,d_2,\ldots,d_{12}$, with average $\overline{d}$ and sample variance s_d^2 then the test statistic $\frac{\overline{d}}{\sqrt{s_d^2/12}}$ should be compared with a t-distribution with 11 degrees of freedom, and the 90% confidence interval should have the form $\overline{d}\pm t_{0.05,11}\, s_d/\sqrt{12}$. The results are likely to be better, as the individual variation between cows in milk production will not influence the differences used in the inference.

   d) Johan could use a randomization test. Applied to his initial data, it could be performed as follows: Under the null hypothesis that the music does not make a difference, all possible possible milking values are exchangeable. Johan could find all possible selections of 6 values out of the twelwe, and subtract from their mean the mean of the remaining values. This would make a reference population with which his actual difference of means 118-114=4, could be compared.

   e) He could use a two-way ANOVA analysis, blocking for the cows and testing whether the musical style made a difference.

2. a) A good estimate is the average of the observations: (12+17)/2 = 14.5

   b) It is reasonable to use the Poisson distribution for the number of triplet births every year. The formula for Poisson probabilities gives $P(3)=\frac{e^{-14.2}14.2^3}{3!}=0.000325$

   c) If X has a Poisson distribution with rate 14.2, its distribution is reasonably well approximated by a normal distribution with expectation 14.2 and variance 14.2 (the same expectation and variance as the Poisson distribution). Thus $\frac{X-14.2}{\sqrt{14.2}}$ has an approximate standard normal distribution. Comparing $\frac{20-14.2}{\sqrt{14.2}}=1.54$ with a standard normal distribution, we get that the probability of observing 20 or more triplet births is approximately 6%. (Using Yates adjustment, we get $\frac{19.5-14.2}{\sqrt{14.2}}=1.41$ giving a p-value of around 8%.)

3. a) The full ANOVA table becomes

	Sum of squares	Degrees of freedom	Mean sum of squares	Test statistic
Treatment	213	2	106.5	5.84
Residuals	219	12	18.25	
Total	432	14		

   We compare 5.84 with the F distribution with 2 and 12 degrees of freedom. From the tables, we see that such a distribution has a 1% probability of values above 6.93, and a 5% probability of values above 3.89. Thus the p-value is somewhere between 0.01 and 0.05, and it is reasonable that Li rejects his null hypothesis.

   b) One way to do a randomization test would be to keep the same test statistic as the one computed as 5.84 in the ANOVA table above, but to compare 5.84 with the results of applying the same computation on a large set of random selections of 3 groups of 5 numbers from the same total data set. (As the question was unclear, some credit has also been given for discussing randomized experimental plans in this context).

4. a) The effect of water is $(-4-8+6+17-3-9+6+11)/4=4$. The interaction effect between temperature and light is $(4-8+6-17-3+9-6+11)/4=-1$.

   b) The pooled variance estimate is $(5.5+1.3+4.0+1.3+0.2+1.5+4.0+10.2)/8=3.5$. The estimate for the standard error of an effect then becomes $\sqrt{3.5(1/12+1/12)}=0.76$ and, finding 2.120 in the table for a t-distribution with 16 degrees of freedom (as the pooled variance estimate has 16 degrees of freedom) we get 95% confidence intervals $4\pm 2.12\cdot 0.76=4\pm 1.61$ and $-1\pm 2.12\cdot 0.76=-1\pm 1.61$.

   c) For example, one may make a drawing of a cube, with the 8 measurements placed appropriately at the 8 corners.

d) One may still analyse the data using linear regression.

e) The assumptions are that the observations at each combination of settings of the factors are independently normally distributed with expectation zero and a variance that is equal for all settings of the factors. To see if it is reasonable to assume that the observations come from normal distributions, one could plot the residuals. (Actually, as the residuals are computed based on the average observation and not on the true expectation, they have a scaled t-distribution with 2 degrees of freedom under the NIID hypothesis. Thus all the residuals could be investigated in a probability plot against the t-distribution with 2 degrees of freedom). Secondly, the sample variances could be compared to see if it is reasonable to assume that the population variances are equal. (Specifically, under the NIID hypothesis, the sample variances have a scaled χ^2 distribution with 2 degrees of freedom, and they could be studied in a probability plot against such a distribution). Finally, the assumption of independence should be investigated, by plotting the residuals against other variables such as the predicted values, time, etc.

5. a) One possible solution is

T	***C***	***X***	***P***	***B***	***Y***
-	-	-	-	-	-
-	-	-	+	-	+
-	-	+	-	+	+
-	-	+	+	+	-
-	+	-	-	+	+
-	+	-	+	+	-
-	+	+	-	-	-
-	+	+	+	-	+
+	-	-	-	+	-
+	-	-	+	+	+
+	-	+	-	-	+
+	-	+	+	-	-
+	+	-	-	-	+
+	+	-	+	-	-
+	+	+	-	+	-
+	+	+	+	+	+

b) The generating relations used above are B=TCX and Y=CXP. It can be described as a 2_{IV}^{6-2} design.

6. a) Using the formula for sample size for such experiments, and finding k=10.5 in the enclosed table, we get $n=2\left(\frac{\sigma}{d}\right)^2 k=2\cdot\left(\frac{14}{5}\right)^2\cdot 10.5=164.64$, so Anna should have at least 165 patients in each group.

b) We see from the formula that the number of patients in each group could be divided by 4. Thus, Anna should use approximately 41 patients in each group in this case.

c) We can see from the table for k that k would decrease, so fewer patients would be needed. One can also argue that with 10% significance instead of 5%, it will be easier to reject the null hypothesis, so fewer patients are needed in order to reject it equally often.

7. a) When the sample mean is 3.8, it means that the sum of squares of each observation minus the average observation is $9 \cdot 3.8^2 = 130$. The sum of squares of residuals is always less than this number, so Nils must have done something wrong in his computations.

   b) He would be likely to get smaller confidence intervals. It is possible to see this from the formula for the variance of regression estimators, but one can also argue that if two predictor variables have a high correlation, it will be more difficult to estimate the effect due to each of them separately. Indeed, if the two predictiors are perfectly aligned so that they have a correlation 1, then any linear effect of one of them will clearly be indistinguishable from a linear effect of the other.