EXAMINATION: Experimental design (MSA250/TMS031)
Wednesday, May 2, 2017, 8:30 - 12:30
Lecturers on call: Kerstin Wiklander and Torbjörn Lundh, tel 772 5355 and tel 772 3503.
Tools: A pocket calculator with emtied memory. At the examination, sheets with statistical distributions and tables will be handed out.

Give explanations to the notation you use and motivation to your conclusions.

1. (2 p) Explain in general terms what the concept "Yates Adjustment" means and give a simple example.

   Short solution indication:
   An adjusetment for going between a discrete distribution (such as the binomial, or a randomized reference) and a continuious (such as the normal). See for example pp. 53 and 106.

2. (2 p) Describe what the term "Lack of fit" means and how one compute the lack of fit sum of squares S_L.

   Short solution indication:

   $$S_L = S_R - S_E,$$

   where S_R is the residual sum of squares and S_E the "pure" error, i.e. the sum of the square of errors from genuine replicates. See p. 369 for more details.

3. (3 p) Several factors were varied in a 2^4 full factorial experiment in a welding experiment. Those were A: Stroke distance, B: Type of material, C: Welding current and D: Electrode force. Other relevant factors were held constant during this experiment. The response variable was Y: Weld strength. The resulting estimates of the effects were: $l_A = -37.3, l_B = -19.3, l_C = 234.7, l_D = 100.5, l_{AB} = -350, l_{AC} = -5.2, l_{AD} = 2.7, l_{BC} = -69.6, l_{BD} = 3.8, l_{CD} = 4.7$ and the mean value $\bar{y} = 331.0$. The following effects were significant in the test performed on significance level 5%: C, D, AB and BC.

   Use this information to express a general formula for the predicted model $\hat{y}$ in regression form and without omitting any main effect.

   What settings do you recommend in order to maximize the weld strength? And which predicted value do you get for that choice?

   Short solution:
   In a model with main effects and the significant two-factor interactions (and without index for the y-variable):
   $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_3 + \hat{\beta}_4 x_4 + \widehat{\beta_{12}} x_1 x_2 + \widehat{\beta_{23}} x_2 x_3 =$
   $= l_M + \frac{l_A}{2} x_1 + \frac{l_B}{2} x_2 + \frac{l_C}{2} x_3 + \frac{l_D}{2} x_3 + \frac{l_{AB}}{2} x_1 x_2 + \frac{l_{BC}}{2} x_2 x_3$, with $x_i = \pm 1$.

Since the interactions have such large effects, we can not find max from the signs of the main effects only. Whether we use all second order interactions (model 1) or only those that were significant (model 2), the optimal setting is high level for factors A, C and D and low level for factor B. The predicted value of this setting in model 1 is 698.6 and 699.4 in model 2.

4. (6 p) Suppose you want to evaluate the sell efficiency of the two used car sales men mr A and mr B by looking at the number of sold cars for January through April.

	Jan	Feb	Mar	Apr
mr A	30	29	30	29
mr B	32	31	31	30

Estimate the probability associated with the two sided significance test of the hypotheses that they are equally efficient sales men. Motivate your choice of method.

Short solution indication:

For all the displayed months, mr B has won the internal competition, so one might assume that the probability that they should be equally good is rather low. To investigate, we can make the bold assumption that the selling situation was equally good all four months and then construct the randomized reference distribution where all the selling number were distributed randomly four to mr A and the remaing four to mr B. This gives us a list of $\binom{8}{4} = 70$ combinations where no combination gives a greater mean difference (1.5) than the actual outcome, but 3 give the same. Using Yates adjustment, we have than that the one sided p-value can be estimated as $\frac{3/2}{70} \approx 0.021$. The two sided could then be estimated as 0.04. Alternatively, we can use a t-test and there obtain a two sided p-value of 0.024, with $s_A^2 = \frac{1}{3}$, $s_B^2 = \frac{2}{3}$, estimated pooled variance of $\frac{1}{2}$ and $t_0 = 3$.

5. (6 p) You have been appointed by some biologists to plan a test on significance level 5%. The objective with the investigation is to detect an biologically relevant deviation (Δ) between the true expected value and that according to a null hypothesis. In the application, this deviation has the value of at least two. We know that the theoretical variance is $\sigma^2 = 3$.

The formula for the sample size with the normal distribution with known variance is:

$$n = \frac{(z_{\alpha/2} + z_\beta)^2 \sigma^2}{\Delta^2}$$

(a) If you want a power of at least $P(Z \leq z_\beta) = 1 - \beta = 0.90$, which sample size should you choose?

(b) Suppose now that it was decided to have the sample size $n = 7$. What power will it be in this case?

Short solution:
(a) We get $n = 7.87$, that is, choose $n = 8$.
(b) With $n = 7$ we get $z_\beta = 1.095$. Then $P(Z \leq 1.095) = 0.86$ according tho the table. Thus, the power is 86%.

6. (3 p) In a 2^{7-3} fractional factorial design, the generators E=ABC, F=ABD and G=ACD were selected.

   Give the total alias pattern for two of the main effects and one of the two-factor interactions (free choice). State also the resolution for this design.

   Short solution:
   $I_1 = ABCE, I_2 = ABDF, I_3 = ACDG, I_4 = I_1 I_2 = CDEF, I_5 = BDEG, I_6 = BCFG, I_7 = AEFG$

   Derive the alias for e.g. A by taking $AI_1, AI_2, \ldots, AI_7$. This gives that l_A estimates $A + BCE + BDF + CDG + ACDEF + ABDEG + ABCFG + EFG$. The same principle for other main factors and interactions.
   The resolution is IV.

7. (3 p) A batch contains products, which are different with respect to the factors A and B. Suppose a random sample of four products resulted in the set-up below. These were used to measure some response Y.

A	B	y
$-$	$-$	y_1
$+$	$-$	y_2
$-$	$+$	y_3
$+$	$-$	y_4

   (a) Express the full model in regression form.
   (b) What is the expected value of the estimator of the coefficient connected to the main effect from factor A?

   Short solution:
   Use this table

M	A	B	AB	y
$+$	$-$	$-$	$+$	Y_1
$+$	$+$	$-$	$-$	Y_2
$+$	$-$	$+$	$-$	Y_3
$+$	$+$	$-$	$-$	Y_4

   from which you can express the model in regression form:
   $Y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_{12} x_{1i} x_{2i} + \epsilon_i$. This gives for the first variable: $Y_1 = \beta_0 - \beta_1 - \beta_2 + \beta_{12} + \epsilon_1$
   The estimator of β_1 is $\hat{\beta}_1 = \frac{l_A}{2} = \frac{-Y_1+Y_2-Y_3+Y_4}{4}$. Then $E[\hat{\beta}_1] = E[-Y_1 + Y_2 - Y_3 + Y_4]/4 = \beta_1 - \beta_2/2 - \beta_{12}/2$. This design is not orthogonal and the estimator is affected

by other parameters ($\hat{\beta}_1$ is not an unbiased estimator of β_1, nor are the other estimators unbiased).

8. (5 p) By using matrix notation we describe a model by

$$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta},$$

where $\boldsymbol{\eta}$ is the vector of expected values for the response, $\boldsymbol{X}$ is the matrix of regressors (independent variabels) and $\boldsymbol{\beta}$ is the vector of parameters.

(a) What is the definition of the so called normal equations?

(b) Give an expression for the normal equations using the vectors $\mathbf{y}, \hat{\mathbf{y}}$ of observed and estimated values.

(c) Give the one line expression to obtain the parameter vector $\boldsymbol{b}$ for the model.

Short solution indication:

(a) Let $\boldsymbol{S(\beta)}$ be the sum of the squares of the discrepancies between the data values and the calculated model values. We want to find the global minimum for $\boldsymbol{S(\beta)}$, hence we are interested when the set of partial derivatives of $\boldsymbol{S(\beta)}$ with respect to the elements in $\boldsymbol{\beta}$ is zero. Those equations are the *normal equations*.

(b) The normal equations can be written as

$$\boldsymbol{X^t(y - \hat{y}) = 0.}$$

(c)

$$\boldsymbol{b} = \left(\boldsymbol{X^t X}\right)^{\boldsymbol{-1}} \boldsymbol{X^t y.}$$