EXAMINATION: Experimental design (MSA250/TMS031)
Wednesday, March 15th, 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. (5 p) Given a randomized experiment of two brands of baking yeast, A and B where the respond is cup-cake hight in mm.

A	A	B	A	B	B
48	43	52	50	49	52

Construct a randomization test and give the p-value. Compare this with the result from a t-test at 5% significance level for the two cases:

(a) if the new yeast brand B is better, i.e. give rise to higher cakes
(b) if one of the yeast brands is better than the other.

For the last task, using the t-distribution, compute the 95% confidence interval of the expected cup-cake-hight-difference between B and A.

Short answer:

$$\binom{6}{3} = 20$$

There is one more extreeme configuration and one different that give an equal difference. This gives the one-sided significance levels 0.1 for the randomization and 0.08 for the t-test. ($s_A^2 = 13, s_B^2 = 3$ and the estimate of the pooled variance is 8. $t_0 = \frac{51-47}{\sqrt{8}\sqrt{1/3+1/3}} \approx 1.73$.) For the second task, we will instead have 0.2 and 0.16. The confidence interval is $(-2.4, 10.4)$

2. (8 p) Below we have the set-up and data from a factorial experiment aiming to investigate the time for a fix amount of liquid to be absorbed in a material. Four measurements were taken for each combination of the levels. The factors and their levels are:

Factors:	Levels: (coded by $-$)	(coded by $+$)
C: Type of outer material (cover)	type C1	type C2
F: Type of filling	type F1	type F2
T: Thickness	thin	thick

The set-up and data:

Cover	Filling	Thickness	mean	sample variance
low	low	low	41.25	2.92
high	low	low	37.00	3.33
low	high	low	43.25	2.92
high	high	low	45.25	4.25
low	low	high	41.00	3.33
high	low	high	36.50	4.33
low	high	high	43.00	3.33
high	high	high	43.00	4.67

(a) Estimate the main effects. Pick one of the interactions and estimate that effect.

(b) Test the main effects on significance level $\alpha = 0.05$. Formulate also for one of the factors the two hypotheses for the test. State the assumptions for your test.

Short answer:

(a) Estimates of the effects:
$l_C = -1.6875, l_F = 4.6875, l_T = -0.8125$
$l_{CF} = 2.6875, l_{CT} = -0.5625, l_{FT} = -0.4375, (l_{CFT} = -0.4375)$

(b) Estimate of σ^2 is $s_p^2 = 3.635$. The estimate of the standard deviation of an effect estimate is $s_{\text{effect}} = \sqrt{4\frac{s_p^2}{2^3 * 4}} = 0.674$ (See the lecture notes in connection to Section 5.7 in the book.)

Test of effect from factor C: $H_0 : \tau_C = 0, H_1 : \tau_C \neq 0$, where τ_C is the effect from factor C. Test statistic for factor C: $\frac{l_C - 0}{s_{\text{effect}}} = -2.50$. It is t-distributed with $2^3 \times (n - 1) = 24$ degrees of freedom. Reject H_0 if the test statistic is $\geq t_{24,0.025} = 2.064$ or $\leq t_{24,0.975} - 2.064$. Thus, $H_0 : \tau_C = 0$ can be rejected and a significant effect (negative) has been found. For the other cases: significant effect from F but not from T. (The interaction CF is also significant.) The assumptions for doing t-test are that the $Y's$ are independent and following the normal distribution with constant variance.

3. (1 p) There is an example in the text book on genuine replicates with n=2 in a 2^3 full factorial experiment. Show why you in this case, with sample size two, can calculate the pooled sample variance s_p^2 (estimator of σ^2) using the formula $\sum_{i=1}^{2^k} \frac{d_i^2/2}{2^k}$, where d_i is the difference $y_{i1} - y_{i2}$ for level combination i.

Short answer:

In general: $s_p^2 = \sum_{i=1}^{2^k} \frac{s_i^2}{2^k}$ with constant n_i. Take for example $i = 1$. Then $s_1^2 = \sum_{j=1}^{2} \frac{(y_{1j} - \bar{y}_1)^2}{2-1} = [y_{11} - (y_{11} + y_{12})/2]^2 + (y_{12} - (y_{11} + y_{12})/2]^2 = (y_{11}/2 - y_{12}/2)^2 + (y_{12}/2 - y_{11}/2)^2 = (d_1/2)^2 + (-d_1/2)^2 = d_1^2/2$.

4. (2 p) In a 2^{7-2}, you are given four suggestions of the fractional design:
No. 1: F=ABCD and G=ACE
No. 2: F=ABCD and G=ABDE
No. 3: F=ABCD and G=ABCE
No. 4: F=ABCD and G=ABD

Are all equally good? Which one would you choose and why?

Short answer:

No. 1: $I_3 : BDEFG$
No. 2: $I_3 : CEFG$
No. 3: $I_3 : DEFG$
No. 4: $I_3 : CFG$

Together with the two generating relations in the four cases, we check the length of the "shortest word". This gives the resolutions IV for No. 1, 2 and 3 while No. 4 has resolution III. Choose any of the suggestions 1,2 and 3.

5. (4 p) An experiment was performed to investigate the influence from temperature and from pH on growth of fish held in different salt water tanks for a time. There were totally 30 fishes with equal size in the beginning and from the same species. The temperatures were varied between $6°, 12°$ and 18°C and for the pH between 7.7 and 8.1, with equal sample sizes for each combination. The response variable was the increase using a growth measure.

   (a) Part of the result is given in the Anova table:

Source of Variation	SS	df	MS	F
Temperature	0.027			
pH	0.0008			
Interaction	xxx	xxx	xxx	xxx
Within groups (error)	0.004	24		
Total	0.0077	29		

   Fill in the Anova table (except for the Interaction row). Draw conclusions about the main effects on significance level 5%.

   A comment: Note that there is a missprint somewhere in this table. This can be seen since the SS for Total is impossible for the other sums of squares. The correct SS for Temperature is 0.0027. However, this does not change any conslusions or marking of the answers.

   (b) The mean values for the combinations are:

pH/Temp	6	12	18
7.7	2.071	2.092	2.070
8.1	2.084	2.101	2.080

Construct an interaction plot. Do you, from that illustration only, think there could be a significant interation between temperature and pH?

Short answer:

(a) The Anova table (from the figures given in the question):

Source of Variation	SS	df	MS	F
Temperature	0.027	2	0.014	78.4
pH	0,0008	1	0.0008	4.86
Interaction	xxx	xxx	xxx	xxx
Within groups (error)	0.004	24	0.00017	
Total	0.0077	29		

A comment: With the correct value of SS Temp, the answer would be:

Source of Variation	SS	df	MS	F
Temperature	0.0027	2	0.0014	7.84
pH	0,0008	1	0.0008	4.86
Interaction	xxx	xxx	xxx	xxx
Within groups (error)	0.004	24	0.00017	
Total	0.0077	29		

We can conclude an effect from temperature since the test statistic (F) has exceeded the percentile (rejection limit) $F_{2,24,0.05} = 3.40$. The same conclusion for pH since that test statistic (F) has exceeded the percentile $F_{1,24,0.05} = 4.26$.

A comment: The percentiles and the conclusions are not affected by the wrong value in the question.

(b) The interaction plot (with Temperature representing the x-axis):

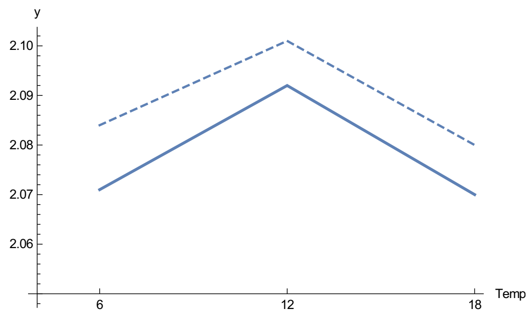

Figure 1: Solid line for Low pH. Dashed line for Normal pH.

The lines looks parallel (they do not need to be straight lines), so there are no indications on an active interaction effect. (The value of the F-statistic was F=0.06 with a p-value of 0.94.)

6. (5 p)

You want to see how the time for pre-fermenting the dough will effect the final cupcakes of your new miracle strain of yeast. You get the following data from your baking skilled friend where t is time in minutes and h height in mm.

t [minutes]	10	10	15	20	20	25	25	25	30	35
h [mm]	73	78	85	90	91	87	86	91	75	65

(a) Find the best (in least square meaning) second order model for how the height might be estimated from the pre-fermenting time of the dough. First, formally in matrix form, $\hat{\mathbf{h}} = \mathbf{X}\mathbf{b}$. That is, give the expression for the matrix $\mathbf{A}$ where $\mathbf{b} = \mathbf{A}\mathbf{h}$.

(b) Find the numerical values for the vector $\mathbf{b}$ and give the fitted model $\hat{h} = \ldots$, when the matrix $\mathbf{A}$ numerically can be approximated by

$$\begin{pmatrix} 1.2152 & 1.2152 & 0.1369 & -0.4359 & -0.4359 & -0.5032 & -0.5032 & -0.5032 & -0.0649 & 0.8790 \\ -0.0949 & -0.0949 & 0.0102 & 0.0609 & 0.0609 & 0.0573 & 0.0573 & 0.0573 & -0.0008 & -0.1133 \\ 0.0018 & 0.0018 & -0.0005 & -0.0015 & -0.0015 & -0.0012 & -0.0012 & -0.0012 & 0.0003 & 0.0032 \end{pmatrix}.$$

(c) What would be an optimal pre-fermentation time (if you want to have as high cup-cakes as possible)?

<u>Short answer</u>: The quadratic model can be written as

$$h = \beta_0 + \beta_1 t + \beta_2 t^2 + \varepsilon.$$

Using vector form with $\mathbf{b} = (\beta_0, \beta_1, \beta_2)^T$, the so called normal equations can be written as $\mathbf{X^T}(\mathbf{h} - \hat{\mathbf{h}}) = \mathbf{0}$. With some algebraic manipulation, this will lead to $\mathbf{b} = (\mathbf{X^T X})^{-1}\mathbf{X^T h}$. Hence our sought after matrix $\mathbf{A} = (\mathbf{X^T X})^{-1}\mathbf{X^T}$. Using the numerical given value of the matrix $\mathbf{A}$, one can obtain via matrix multiplication $\mathbf{b} = (\mathbf{35.66}, \mathbf{5.26}, \mathbf{0.128})^{\mathbf{T}}$. Hence, the fitted model will then be

$$\hat{y} = 35.7 + 5.3t - 0.13t^2.$$

Take the derivative to find the estimate of the maximal hight at $t = \frac{5.3}{2 \cdot 0.13} \approx 20.55 \approx 21$ minutes.

7. (5 p)

Suppose you have a 2^2 facorial design (first order) without center point with IID errors.

(a) Give an expression of the model.

(b) Suppose $\{y_i\}_{i=1}^4$ are observations at the four design points. Give the least–square estimations of the coefficient in your model using these four observed values and give the estimated $\hat{y}$.

(c) State and explain the meaning of the *information function* of this design.

Short answer: The first order model can be written as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon.$$

Using the observed values, $\{y_i\}_{i=1}^4$, at the four design points $(\pm 1, \pm 1)$, we get via the normal equations, i.e. least square, that an estimates of the model parameters is

$$b_0 = \frac{1}{4}(y_1 + y_2 + y_3 + y_4)$$

$$b_1 = \frac{1}{4}(-y_1 + y_2 - y_3 + y_4)$$
$$b_2 = \frac{1}{4}(-y_1 - y_2 + y_3 + y_4).$$

The estimated coefficients are distributed independetly, hence the standardized variane of the estimate would be

$$\frac{V(\hat{y})}{\sigma^2} = \frac{1}{4}(1 + x_1^2 + x_2^2) = \frac{1}{4}(1 + r^2) = \frac{1}{I(x_1, x_2)},$$

where $I(x_1, x_2)$ is the design information function measuring the information the design harwest at a point (x_1, x_2) with distance r from the design center.

See page 448 in the book for more information on the *information function*, or your lecture notes.