Design and Analysis of Experiments/Försöksplanering

(KBT120, KKR031)

Thursday 25/10 2013 8:30-13:30 V

Bengt Andersson will be available at ext 3026 and will visit the examination room ca 11:00.

The examination results will be available for review 25/11 12:45 – 13:15 KRT sem. room.

Time for examination = 5 h

Examination aids:

Textbook (Douglas C. Montgomery: Design and Analysis of Experiments) with notes. No calculation examples (in book or on paper) are allowed as aid.

All type of calculators, with emptied memory, are allowed.

Standard Math. Tables, TEFYMA table, Beta Mathematics Handbook or Handbook of Chemistry and Physics and Language dictionaries are accepted.

NOTE: Problem 6 is available in two versions. The first example Multivariate Statistics is for students taking the course KBT 120 (7.5 credits) (Biotechnology registered 2011 and later, master students and all other students) **and the second problem for students taking the course KKR031 (6 credits)** (Biotechnology students registered before 2011)

Use significance lever α=0.05 if nothing else is specified

Problem 1 (12credits)

The effect of five factors (A,B,C,D,E) has beed studied in 8 experiments with D=AB and E=AC resulting in l_A=20, l_B=22,5, l_C=-3, l_D=11, l_E=-1,5, l_{BC}=-5, l_{ABC}=-1,5 and the average M=60,75.

The physics in the process makes it reasonable that only the interactions AD and BD have an effect. Suggest a suitable experimental design for additional experiments that makes it possible to evaluate the interactions AD and BD without confounding with the main effects and each other. Calculate the main effects, AD and BD. Evaluate if they are significant and determine their aliases. All other factors are assumed to be random.

The experimental results are obtained from the table below. The letters indicate which factors that are high in the experiment.

(1)	= 62	d	= 69	e	= 56	de	= 44
a	= 53	ad	= 61	ae	= 63	ade	= 45
b	= 63	bd	= 94	be	= 59	bde	= 88
ab	= 61	abd	= 93	abe	= 65	abde	= 77
c	= 53	cd	= 46	ce	= 69	cde	= 49
ac	= 56	acd	= 60	ace	= 55	acde	= 42
bc	= 54	bcd	= 95	bce	= 67	bcde	= 81
abc	= 61	abcd	= 98	abce	= 65	abcde	= 82

Problem 2 (8 credits)

In linear regression the parameters are estimated from

$$\mathbf{b=(X^TX)^{-1}X^Ty}$$

and the variance

$$\mathrm{V(b)= (X^TX)^{-1}\sigma^2}$$

What is required from the model, the independent variable **X** and the independent variable **y** in the model y=Xβ for the equations above to give true expected value i.e. b is a correct estimation of true β and **V(b)** is a correct estimation of the true variance-covariance matrix? How do you test that these conditions are fulfilled?

How can you determine the reason for large confidence region for estimated parameters? Suggest an evaluation method and additional experiments that can help to decide if the reason is bad experiments, a poor model or a poor experimental design.

Problem 3 (10 credits)

The hydrolysis of cellulose was studied at four different pH using four different acids. The table below shows the fraction of undesired byproduct in ppm that may inhibit fermentation. Unfortunately one experiment failed. Make a statistical analysis and calculate which combination of acid and pH that statistically is expected to produce the lowest fraction of byproduct.

	pH 1	**pH 2**	**pH 3**	**pH 4**
HCl	20	26	21	25
HNO_3	20	26	23	27
H_2SO_4	16	13	-	16
H_3PO_4	20	15	17	20

$y_{..} = 305$ ~~$y^2_{..} = 6471$~~ $\sum y_{ij}^2 = 6471$

Problem 4 (12 credits)

Matlab's subrutine cordexch(4,15,'quadratic') has suggested the first 15 experiments below. Five additional experiments in the central point has also been performed. The model $\hat{y} = b_0 + \sum_{i=1}^{4} b_i x_i + \sum_{i=1}^{4}\sum_{j \geq i}^{4} b_{ij} x_i x_j$ has been fitted to the data.

	x_1	x_2	x_3	x_4	y
	1.00	1.00	-1.00	-1.00	-34.43
	-1.00	0	-1.00	-1.00	6.33
	1.00	-1.00	1.00	1.00	12.13
	-1.00	1.00	1.00	1.00	-17.71
	1.00	-1.00	-1.00	1.00	-27.15
6	-1.00	-1.00	1.00	-1.00	11.19
7	-1.00	1.00	-1.00	0	6.19
8	1.00	0	1.00	-1.00	-8.04
9	1.00	0	0	0.36	2.29
10	0	0	1.00	0	9.17
11	1.00	1.00	1.00	1.00	9.81
12	-1.00	-1.00	-1.00	1.00	46.73
13	1.00	-1.00	-1.00	-1.00	-36.59
14	0	1.00	-1.00	1.00	-10.82
15	-0.26	1.00	0	-1.00	-21.85
	0	0	0	0	10.11
	0	0	0	0	11.07
	0	0	0	0	10.06
	0	0	0	0	9.90
	0	0	0	0	9.17

$H=X(X^TX)^{-1}X^T$

Diag(H)= [1.00 0 .99 1.00 0.99 0.96 0.97 0.98 0.96 0.94 0.93 0.99 0.97 1.00 ....
0.97 0.49 0.17 0.17 0.17 0.17 0.17]

Diag($(X^TX)^{-1}$)=[0.17 0.13 0.11 0.14 0.11 0.14 0.13 0.14 0.12 0.15 0.14 0.58 0.38 0.51 0.47]

b^T= [10.17 -10.09 -12.07 3.12 8.36 12.09 13.21 -1.43 -1.65 0.51 1.58 0.93 1.90 -4.42 -12.33]

$\bar{y} = 0,75$

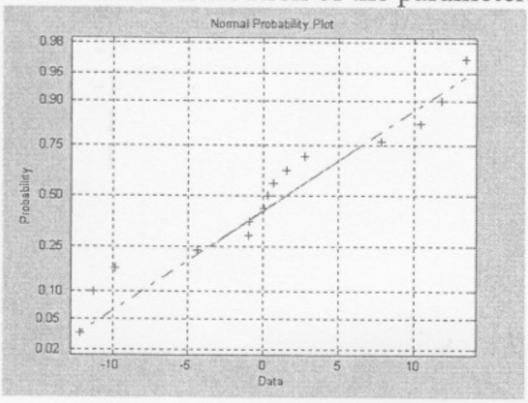
$$SST = \sum (y_i - \bar{y})^2 = 5269.25$$

$$SSE = \sum (y_i - \hat{y}_i)^2 = 2.21$$

Correlation matrix

b_0	b_1	b_2	b_3	b_4	b_{12}	b_{13}	b_{14}	b_{23}	b_{24}	b_{34}	b_{11}	b_{22}	b_{33}	b_{44}
1.00	-0.04	-0.20	-0.06	0.10	0.09	-0.01	-0.09	-0.04	0.08	0.14	-0.19	-0.07	-0.24	-0.08
-0.04	1.00	0.05	-0.09	-0.18	0.04	0.20	0.03	0.07	0.18	-0.18	-0.11	0.21	0.09	-0.20
-0.20	0.05	1.00	0.05	-0.08	-0.13	-0.23	0.29	0.25	-0.26	-0.38	0.24	-0.20	0.13	-0.04
-0.06	-0.09	0.05	1.00	-0.16	-0.08	-0.04	-0.10	0.09	-0.29	-0.11	0.01	0.19	0.04	-0.13
0.10	-0.18	-0.08	-0.16	1.00	0.11	-0.29	-0.24	-0.48	0.04	0.21	0.16	-0.47	-0.18	0.39
0.09	0.04	-0.13	-0.08	0.11	1.00	-0.09	0.01	0.05	-0.04	0.23	-0.03	0.10	0.08	-0.18
-0.01	0.20	-0.23	-0.04	-0.29	-0.09	1.00	-0.16	0.09	0.31	-0.02	-0.17	0.41	-0.17	-0.06
-0.09	0.03	0.29	-0.10	-0.24	0.01	-0.16	1.00	0.34	-0.11	-0.12	-0.03	-0.00	0.22	-0.15
-0.04	0.07	0.25	0.09	-0.48	0.05	0.09	0.34	1.00	-0.19	-0.19	-0.16	0.27	0.29	-0.37
0.08	0.18	-0.26	-0.29	0.04	-0.04	0.31	-0.11	-0.19	1.00	-0.02	0.05	0.10	-0.31	0.11
0.14	-0.18	-0.38	-0.11	0.21	0.23	-0.02	-0.12	-0.19	-0.02	1.00	-0.23	-0.05	0.02	0.10
-0.19	-0.11	0.24	0.01	0.16	-0.03	-0.17	-0.03	-0.16	0.05	-0.23	1.00	-0.18	-0.43	-0.08
-0.07	0.21	-0.20	0.19	-0.47	0.10	0.41	-0.00	0.27	0.10	-0.05	-0.18	1.00	-0.11	-0.55
-0.24	0.09	0.13	0.04	-0.18	0.08	-0.17	0.22	0.29	-0.31	0.02	-0.43	-0.11	1.00	-0.38
-0.08	-0.20	-0.04	-0.13	0.39	-0.18	-0.06	-0.15	-0.37	0.11	0.10	-0.08	-0.55	-0.38	1.00

The normal distribution of the parameters



Discuss:
Is the experimental design acceptable?
What experiments have the largest effect on the parameters in the model?
Is the model significant?
Is there a lack of fit?
What parameters are significant?
How can you estimate how large fraction of the total sum of squares that is determined by each parameter?

Problem 5 (10 credits)

In pulp production the quality has been determined as function of boiling time (x_1), temperature (x_2) and the influence from cutting the wood (z1) in the summer (-1) or winter (+1). The quality can be described by the polynomial

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \gamma_1 z_1 + \delta_{11} x_1 z_1 + \delta_{21} x_2 z_1 + \varepsilon$$

där

$$\beta_0 = 55 \quad \beta_1 = 5 \quad \beta_2 = 3 \quad \beta_{12} = -5 \quad \gamma_1 = 2 \quad \delta_{11} = 1 \quad \delta_{21} = -2$$

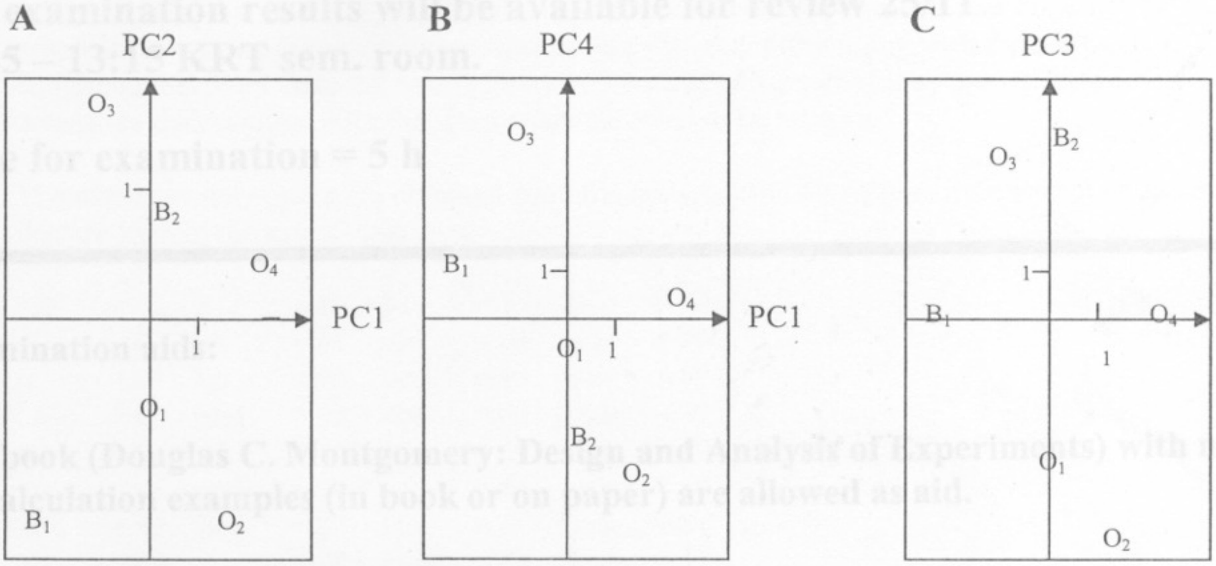
Estimate the best running conditions and calculate what quality the can be guaranteed if 95% of the production should reach that quality. The variation in repeated identical experiments (the same wood) is estimated as $s^2=1$ and the variation in wood is $s_z^2=0.25$. The model is based on many experiments i.e. all parameters and variances are known exactly (degrees of freedom $=\infty$).

Do the calculations for the following conditions

x_1	x_2
-1	-1
1	-1
-1	1
1	1

Problem 6 (8 credits)
(For master student and KB students registered 2011 or later taking the 7.5 credit course)

1. The diagrams below A-C show score plots of B_1 the dependent variables B_1 and B_2 and the independent variables $O_1 - O_4$ with three different principal components. You want to know how the dependent variables B_1 and B_2 correlates with the independent variables $O_1 - O_4$. Which diagram(s) gives most information? Motivate your answer!
2. What can you deduce from the diagrams? How are the dependent and independent variables correlated? Is the correlation strong or weak, positive or negative? Motivate your conclusions,



Problem 6 (8 credits)
(For biotechnology students registered before 2011 taking the 6 credits course)

The quality assessment of raw material that was used in a production unit gave the results below. A random selection of raw material from three different batches from each supplier were separated into two random samples and each sample was analyzed twice.

	Supplier 1						Supplier 2						Supplier 3					
batch	1		2		3		1		2		3		1		2		3	
Sample	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2	1	2
Analysis	23	24	25	23	24	25	21	23	26	25	23	22	22	24	26	23	21	23
	24	25	25	23	22	23	22	22	24	25	24	21	21	26	23	25	25	22

Batch average 24.0 24.0 23.5 22.0 25.0 22.5 23.25 24.25 22.75

Total $\sum y_i^2 = 19913$ $\overline{y} = 23.47$ $\sum (y_i - \overline{y})^2 = 78.97$ The mean sum of squares for error in Analysis is MS_E=1.47

Deside if you should continue with all three suppliers or is any of them has significantly lower variation in quality seen as the variation between batches and within batches.