Kurskod: TAMS38 - Provkod: TEN1

TAMS38 – Experimental Design and Biostatistics, 4 p / 6 hp
Examination on 10 January 2017, 14–18.

The collection of the formulas in mathematical statistics prepared by Department of Mathematics LiU and calculator with empty memory are allowed on the exam. Dictionary English-other language is allowed. No extra notes in the formula collection is allowed.
Score limits: 7-9 points gives 3, 9.5-12 gives 4 and 12.5-15 gives 5.
Examinator: Jolanta Pielaszkiewicz
The result will be *normally* published via LADOK within 12 working days.
Clear answers and justifications are required for each task.

1) In semiconductor manufacturing wet chemical etching is often used to remove silicon from the backs of wafers prior to metalization. The each rate is important characteristic of this process. Two different etching solutions have been evaluated. We assume that measurements for those two solutions come from $N(\mu_1, \sigma)$ and $N(\mu_2, \sigma)$ distribution, respectively. We know that $\sigma = 1.2$.
a) How many measurements should we do for each of solutions so that the length of interval $I_{\mu_1-\mu_2}$ is at most 3 and $\alpha = 0.05$? (1p)
b) In the next step the following data are measured:

Solution 1		Solution 2	
9.9	10.6	10.2	10.6
9.4	10.3	10.0	10.2
10.0	9.3	10.7	10.4
10.3	9.8	10.5	10.3
$s_1 = 0.45$		$s_2 = 0.233$	

Do the data indicate that both solutions have the same mean etch rate? Use $\alpha = 0.05$ and assumption about equal variance. Estimate σ with s, as low values of s_1 and s_2 indicate that assumption $\sigma = 1.2$ may we not relevant. (1p)

2) A chemist wishes to test the effect of four chemical agents on the strength of a particular type of cloth. Because there might be variability from one bolt to another, the chemist decides to use a randomized block design with the bolts of cloth considered as blocks. Analyze the eventual difference between chemical agents on $\alpha = 10\%$.

	Bolt				
Chemical	1	2	3	4	5
1	72	68	74	71	67
2	73	67	75	72	70
3	75	68	78	73	68
4	73	71	75	75	69

a) Using suitable parametric test. Remember to identify factor1 and factor2. (1p)

```
Analysis of Variance for strength
Source     DF        SS
factor1     4   153.700
factor2     3    15.400
Error      12    21.100
Total      19   190.200

Means
factor1  N  strength            factor2   N  strength
   1     4    73.250               1      5    70.400
   2     4    68.500               2      5    71.400
   3     4    75.500               3      5    72.400
   4     4    72.750               4      5    72.600
   5     4    68.500
```

b) Using suitable non-parametric test. (1.5p)
c) Construct confidence interval for the difference between position parameters of Chemical 1 and 4. Do not assume normal distribution on $\alpha = 14\%$. (1p)
You can use Minitab results below if you find them suitable for the task:

```
Sorted pairwise averages chemical1-chemical4
   -4.0   -3.5   -3.0   -3.0   -2.5   -2.5   -2.5   -2.0   -2.0
   -2.0   -1.5   -1.5   -1.0   -1.0   -1.0

Sorted pairwise differences chemical1-chemical4
   -8    -8    -7    -7    -6    -5    -4    -4    -4    -3    -3    -3
   -2    -2    -1    -1    -1    -1     0     1     1     2     3     3     5
```

If needed the Wilcoxons sign rang test tabel is given. The other table for Wilcoxons sum rang test is in the table and formula collection.

TABELL FÖR WILCOXONS TECKENRANGTEST

APPENDIX 4 Statistical Tables

TABLE F Cumulative probabilities $P(W \leq c)$, where W has the signed rank distribution for sample size n

n: / c	2	3	4	5	6	7	8
0	.250	.125	.062	.031	.016	.008	.004
1	.500	.250	.125	.062	.031	.016	.008
2		.375	.188	.094	.047	.023	.012
3		.625	.312	.156	.078	.039	.020
4			.438	.219	.109	.055	.027
5			.562	.312	.156	.078	.039
6				.406	.219	.109	.055
7				.500	.281	.148	.074
8					.344	.188	.098
9					.422	.234	.125
10					.500	.289	.156
11						.344	.191
12						.406	.230
13						.469	.273
14						.531	.320
15							.371
16							.422
17							.473
18							.527

d) Both of the factors in the model are **not** dummy variables. Propose the dummy variables that can be used instead of multilevel variable *Chemical*. (0.5p)

3) The data shown in the table below were collected in an experiment to optimaze crystal growth as a function of three variables x_1, x_2 and x_3. Large values of y (yield in grams) are desirable.
a) Check if a second-order model surface is suitable for modeling data on $\alpha = 5\%$? Remember to write H_0 and H_1. One can assume normal distribution. (2.5p)
b) Draw conclusion accordingly to the order of the obtained surface how you should proceed with analysis if your aim is to find conditions assuring maximum growth. (0.5p)

x_1	x_2	x_3	y
-1	-1	-1	66
-1	-1	1	70
-1	1	-1	78
-1	1	1	64
1	-1	-1	80
1	-1	1	70
1	1	-1	100
1	1	1	75
0	0	0	113
0	0	0	100
0	0	0	118
0	0	0	88
0	0	0	100

4) Heat treating is often used to carbonize metal parts, such a gears. The thickness of the carbonized layer is a critical output variable from this process, and it is usually measured by performing a carbon analysis on the gear pitch (the top of the gear tooth). Six factors were studied in a 2^{6-2} design:
A = furnace temperature; B = cycle time;
C = carbon concentration; D = duration of the carbonizing cycle;
E = carbon concentration of the diffuse cycle; F = duration of the diffuse cycle.
Measurements are given below.

(1)	74	df	119
ae	190	adef	188
bef	133	bde	135
abf	127	abd	170
cef	115	cde	126
acf	101	acd	175
bc	54	bcdf	126
abce	144	abcdef	193

a) Which of below mentioned blocking rules have been used in the experiment.

i) E=ABC and F=ABCD

ii) E=ABC and F=BCD

iii) E=ABCD and F=BCD

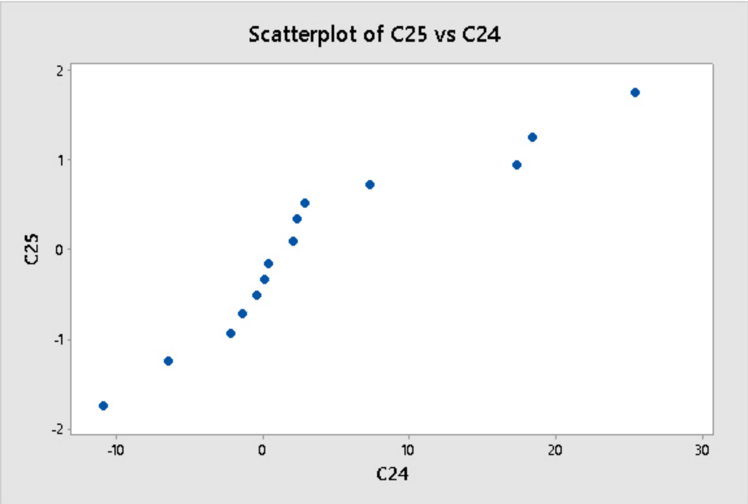
Motivate your answer. Calculate all generators of the design. (1p)
b) What are the four most significant effects? Give the corresponding parameter estimates and all their aliases. (1p)

Minitab analysis:

```
MTB > copy c1-c16 m1
MTB > copy c17 m2
MTB > trans m1 m3
MTB > mult m3 m2 m4
MTB > copy m4 c19
MTB > let c20=c19/16
MTB > set c18
DATA> 1:16
DATA> end
MTB > Sort c20 c18 c21 c22;
SUBC> By c20.
MTB > print c21 c22

Data Display
Row       C21   C22
  1   -10.875    16
  2    -6.375     5
  3    -2.125     4
  4    -1.375     6
  5    -0.375     3
  6     0.125    12
  7     0.375     7
  8     2.125    10
  9     2.125    15
 10     2.375    11
 11     2.875    14
 12     7.375    13
 13    17.375     8
 14    18.375     9
 15    25.375     2
 16   135.625     1

 MTB > copy c21 c24;
 SUBC> omit 16.
 MTB > nscores c24 c25
 MTB > plot c25*c24
```



c) Below we give ANOVA table.

```
Analysis of Variance
Source   DF    Adj SS    Adj MS  F-Value  P-Value
  A       1   10302.3   10302.3    23.04    0.002
  B       1       2.2       2.2     0.01    0.945
  C       1     650.3     650.3     1.45    0.267
  D       1    5402.3    5402.3    12.08    0.010
  A*B     1      72.3      72.3     0.16    0.700
  A*C     1      30.2      30.2     0.07    0.802
  B*C     1       2.3       2.3     0.01    0.945
  A*B*C   1    4830.2    4830.2    10.80    0.013
Error     7    3129.8     447.1
```

```
Total     15  24421.8

Model Summary
      S    R-sq  R-sq(adj)  R-sq(pred)
21.1449  87.18%     72.54%      33.05%

Means
A   N        y        B   N        y       C   N        y       D   N        y
-1  8  110.25        -1  8  136.00      -1  8  142.00      -1  8  117.25
 1  8  161.00         1  8  135.25       1  8  129.25       1  8  154.00

A   B   N       y        A   C   N       y       B   C   N       y
-1  -1  4  108.50       -1  -1  4  115.25       -1  -1  4  142.75
-1   1  4  112.00       -1   1  4  105.25       -1   1  4  129.25
 1  -1  4  163.50        1  -1  4  168.75        1  -1  4  141.25
 1   1  4  158.50        1   1  4  153.25        1   1  4  129.25

A   B   C   N       y
-1  -1  -1  2   96.50
-1  -1   1  2  120.50
-1   1  -1  2  134.00
-1   1   1  2   90.00
 1  -1  -1  2  189.00
 1  -1   1  2  138.00
 1   1  -1  2  148.50
 1   1   1  2  168.50
```

State a model used for analyzing data in ANOVA analysis above and using it find the best level combination of the given factors on $\alpha \approx 10\%$. Assume that high value of response variable is preferable. (2p)

d) Now instead of ANOVA we use the regression model (for the same data). The Minitab results are given below. State a model used for analysis and write obtained regression line equation. Check significance of main effect of D as explanatory variable with test or confidence interval on $\alpha = 5\%$. (2p)

```
Coefficients
Term         Coef  SE Coef
Constant   135.63     5.29
A           25.38     5.29
B           -0.37     5.29
C           -6.38     5.29
D           18.38     5.29
A*B         -2.13     5.29
A*C         -1.37     5.29
B*C          0.38     5.29
A*B*C       17.38     5.29
```

TAMS38 (TAMS12)
Some extra formulas for non-parametric tests

The Wilcoxon signed rank test

Let r_i be the rank for the observations $|y_i| \neq 0, i = 1, \ldots, n$. Let $T_+ = \sum_{\{y_i>0\}} r_i$ and $T_- = \sum_{\{y_i<0\}} r_i$. When H_0 is true and $n > 15$ us that T_+ and $T_- \approx N\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right)$. For $n \leq 15$ us table for Wilcoxons signed rank distribution.

For confidence interval us the $N = n(n+1)/2$ ordered pairwise means A_i and $P(A_{(k)} < \mu < A_{(N-k+1)}) = 1 - 2P(W_S \leq k-1)$, where W_S is Wilcoxons signed rank distributed.

The Wilcoxon-Mann-Whitney test

Let d_{ij} be the differences $d_{ij} = x_i - y_j$, $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$ and $d_{(k)}$, $k = 1, \ldots, n_1 n_2$, the ordered differences. The confidence interval for the difference in mean/median is given by

$$I = \left(d_{(c+1)}, d_{(n_1 n_2 - c)}\right),$$

where $c = T_l - \frac{n_1(n_1+1)}{2}$ and T_l is from the Wilcoxon table for the rank sum test.

The Kruskal-Wallis test

Assume a treatments. Let r_{ij} be the rank for the observation y_{ij}. Test statistic

$$T = \begin{cases} \dfrac{12S_a}{N(N+1)} - 3(N+1), & \text{if no ties,} \\ \\ \dfrac{(N-1)(S_a - C)}{S_r - C}, & \text{if ties,} \end{cases}$$

where $s_i = \sum_{j=1}^{n_i} r_{ij}$, $S_a = \sum_{i=1}^{a} \frac{s_i^2}{n_i}$, $S_r = \sum_{i=1}^{a} \sum_{j=1}^{n_i} r_{ij}^2$, $C = \frac{1}{4} N(N+1)^2$ and $N = \sum_{i=1}^{a} n_i$. For *small* values of $n_1, \ldots, n_a$ ($a \leq 3$ and $n_i \leq 5$) use table and for *large* values of $n_1, \ldots, n_a$ use that $T \approx \chi^2(a-1)$ when there is no treatmet effect.

The Friedman test

Assume t treatments and b blocks. Let r_{ij} be the rank of y_{ij} within each block, i.e., for each j, $r_{ij} = 1, \ldots, t$. Test statistic for the treatments is given by

$$T = \begin{cases} \dfrac{12S_t}{t(t+1)} - 3b(t+1), & \text{if no ties,} \\ \\ \dfrac{b(t-1)(S_t - C)}{S_r - C}, & \text{if ties,} \end{cases}$$

where $s_i = \sum_{j=1}^{b} r_{ij}$, $S_t = \frac{1}{b} \sum_{i=1}^{t} s_i^2$, $S_r = \sum_{i=1}^{t} \sum_{j=1}^{b} r_{ij}^2$ and $C = \frac{1}{4} bt(t+1)^2$. For *small* values of b and t ($t = 3$, $b \leq 15$ and $t = 4$, $b \leq 8$) use table and for *large* values of b and t use that $T \approx \chi^2(t-1)$ when there is no treatmet effect.