Kurskod: TAMS38 - Provkod: TEN1

TAMS38 – Experimental Design and Biostatistics, 6 hp
April 24 2019, 8–12

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 are allowed. No extra notes in the formula collection is allowed.
Score limits: 7-9 points gives grade 3, 9.5-12 gives 4 and 12.5-15 gives 5.
Examinator: Martin Singull, 013-281447
The result will *normally* be published via LADOK within 15 working days.
Clear answers and justifications are required for each problem.

1. One wants to compare six different types of oats. The experiment is designed as a block experiment with blocks = cultivation area and each oat variety has been cultivated within the various blocks. The proportion of protein in the dry weight has then been determined. Results expressed in percentage:

	Cultivation area = block					
Oat variety	1	2	3	4	5	6
1	19.09	20.29	20.31	19.60	18.62	20.10
2	16.28	17.88	16.88	17.57	16.72	17.32
3	16.31	18.17	17.38	17.53	16.34	17.88
4	17.50	18.05	17.59	17.64	17.38	18.04
5	16.25	16.92	15.88	14.78	15.97	16.66
6	21.09	21.37	21.38	20.52	21.09	21.58

Below, the data has been analyzed using Minitab.

   (a) According to which model has the data material been analyzed? Write up the model and specify any constraints on the parameters. (0.5p)

   (b) Rank the oats based on protein content by constructing appropriate confidence intervals with simultaneous confidence level exactly 95%. (2p)

   (c) Estimate the protein content of oat variety 6 grown in area (block) 1. (0.5p)

```
MTB > ANOVA 'Y' = A B;
SUBC>   Means A B.

ANOVA: Y versus A, B

Analysis of Variance for Y

Source  DF       SS      MS
A        5  106.788  21.358
B        5    6.112   1.222
Error   25    4.999   0.200
Total   35  117.898

Means

A    N       Y
1    6  19.668
2    6  17.108
3    6  17.268
4    6  17.700
5    6  16.077
6    6  21.172

B    N       Y
1    6  17.753
2    6  18.780
3    6  18.237
4    6  17.940
5    6  17.687
6    6  18.597
```

2. The purity (y_{ij}) of a chemical product varies between different production batches and of course within a batch. One has randomly selected four production batches from a large number and analyzed three samples from each of them. Results:

Batch			
1	2	3	4
93.88	94.53	95.40	93.16
93.33	94.39	95.88	93.71
93.16	94.16	95.89	93.67

Assume a random effects model $Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$, where $\tau_i \sim N(0, \sigma_\tau)$ and $\varepsilon_{ij} \sim N(0, \sigma)$, $i = 1, \ldots, 4$, $j = 1, \ldots, 3$. The analyze of the data gave $SS_{TREAT} = 10.0625$ and $SS_E = 0.6980$.

   (a) Why should we use a random effects model in this case? (0.5p)

   (b) Test if the batches are homogeneous with an F-test at level 5%. Write out both the null hypothesis H_0 and the alternative H_1. (1.5p)

3. In a cultivation experiment, one has three factors A, B and C and measured the yield Y. Two replicates of a full 2^3 design was perfomred with one observation in the zero point.

x_1	-1	-1	-1	-1	1	1	1	1	0
x_2	-1	-1	1	1	-1	-1	1	1	0
x_3	-1	1	-1	1	-1	1	-1	1	0
Replicate 1	3.70	2.94	4.94	2.31	5.68	3.31	10.57	16.82	5.86
Replicate 2	3.80	2.11	3.92	2.03	5.29	3.09	11.83	9.84	7.24

Since there were large time intervals between the crops, we regard it as a block design. The measurements within a replicate are thus included in the same block. After doing a residual analysis, the observations are transformed according to $\ln Y$.

```
Factorial Fit: lnY versus Block, A, B, C

Estimated Effects and Coefficients for lnY (coded units)

Term      Effect     Coef    SE Coef
Constant           1,5402    0,04053
Block              0,0565    0,03821
A        -0,3792  -0,1896    0,04053
B         0,5313   0,2656    0,04053
C         0,8360   0,4180    0,04053
A*B       0,0948   0,0474    0,04053
A*C       0,1799   0,0899    0,04053
B*C       0,5212   0,2606    0,04053
A*B*C     0,2447   0,1224    0,04053
Ct Pt              0,3337    0,12159

Analysis of Variance for lnY (coded units)

Source              DF   Seq SS   Adj SS   Adj MS
Blocks               1  0,05751  0,05751  0,05751
Main Effects         3  4,50009  4,50009  1,50003
2-Way Interactions   3  1,25190  1,25190  0,41730
3-Way Interactions   1  0,23960  0,23960  0,23960
  Curvature          1  0,19798  0,19798  0,19798
Residual Error       8  0,21026  0,21026  0,02628
Total               17  6,45734
```

(a) Test if the response surface in the studied area is curved. Use significance level 5%. (1.5p)

(b) Given the result in (a), how would you like to continue the analysis? Would you like to take more observations? For which x_1, x_2 and x_3? (1.5p)

4. A producer of plastic film has been commissioned to produce a film as thin as possible. In the chemical process, five factors, A, B, C, D, and E, have been varied on two levels each, i.e., a 2^5 factorial design. For each combination of the five factors, the thickness of the plastic film was measured. The observations are in the table below followed by an analysis using Minitab.

(1)	1.129	d	1.760	e	1.224	de	1.674
a	0.985	ad	1.684	ae	1.092	ade	1.215
b	1.347	bd	1.957	be	1.280	bde	1.275
ab	1.151	abd	1.656	abe	1.381	abde	1.446
c	2.197	cd	2.472	ce	1.859	cde	2.585
ac	1.838	acd	2.147	ace	1.865	acde	2.587
bc	1.744	bcd	2.142	bce	1.867	bcde	2.339
abc	2.101	abcd	2.423	abce	2.005	abcde	2.629

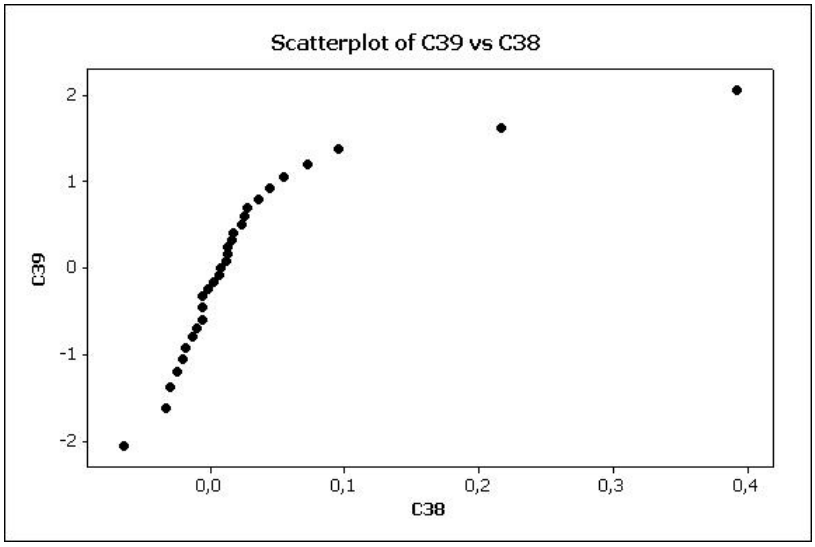
```
MTB > copy c1-c32 m1
MTB > trans m1 m2
MTB > set c33
DATA> 1,129 0,985 1,347 1,151 2,197 1,838 1,744 2,101
DATA> 1,760 1,684 1,957 1,656 2,472 2,147 2,142 2,423
DATA> 1,224 1,092 1,280 1,381 1,859 1,865 1,867 2,005
DATA> 1,674 1,215 1,275 1,446 2,585 2,587 2,339 2,629
DATA> end
MTB > copy c33 m3
MTB > mult m2 m3 m4
MTB > copy m4 c34
MTB > let c34 = c34/32
MTB > set c35
DATA> 1:32
DATA> end
MTB > Sort c34 c35 c36 c37;
SUBC>   By c34.
MTB > print c36 c37

Data Display

Row       C36  C37
  1  -0,06381   24
  2  -0,03219    7
  3  -0,02950   11
  4  -0,02456   27
  5  -0,02019    2
  6  -0,01787   25
  7  -0,01281   17
  8  -0,00950   32
  9  -0,00594   31
 10  -0,00588   19
 11  -0,00588   10
 12  -0,00556   16
 13  -0,00094   26
 14   0,00262   22
 15   0,00744   20
 16   0,00844   12
 17   0,01250   14
 18   0,01281   30
 19   0,01344    3
 20   0,01600   15
 21   0,01763   23
 22   0,02406   13
 23   0,02613   28
```

```
 24   0,02750   18
 25   0,03613    8
 26   0,04456    6
 27   0,05481   21
 28   0,07275    4
 29   0,09538   29
 30   0,21644    9
 31   0,39200    5
 32   1,78300    1

MTB > copy c36 c38;
SUBC> omit 32.
MTB > nscores c38 c39
MTB > plot c39*c38
```



(a) Which four effects appear to have the greatest significance based on the Minitab analysis above? (1p)

Given the result in (a), another Minitab analysis is made.

```
MTB > ANOVA 'Y' = C|D|E;
SUBC>   Means C|D|E;
SUBC>   GFourpack.

ANOVA: Y versus C; D; E

Analysis of Variance for Y

Source  DF       SS       MS       F      P
C        1  4,91725  4,91725  200,01  0,000
D        1  1,49905  1,49905   60,97  0,000
E        1  0,00525  0,00525    0,21  0,648
C*D      1  0,01853  0,01853    0,75  0,394
C*E      1  0,09614  0,09614    3,91  0,060
D*E      1  0,01022  0,01022    0,42  0,525
C*D*E    1  0,29108  0,29108   11,84  0,002
```

```
Error   24  0,59004  0,02458
Total   31  7,42756

S = 0,156796   R-Sq = 92,06%   R-Sq(adj) = 89,74%

Means

C    N       Y
-1  16  1,3910
 1  16  2,1750

D    N       Y
-1  16  1,5666
 1  16  1,9994

E    N       Y
-1  16  1,7958
 1  16  1,7702

C   D   N       Y
-1  -1  8  1,1986
-1   1  8  1,5834
 1  -1  8  1,9345
 1   1  8  2,4155

C   E   N       Y
-1  -1  8  1,4586
-1   1  8  1,3234
 1  -1  8  2,1330
 1   1  8  2,2170

D   E   N       Y
-1  -1  8  1,5615
-1   1  8  1,5716
 1  -1  8  2,0301
 1   1  8  1,9688

C   D   E   N       Y
-1  -1  -1  4  1,1530
-1  -1   1  4  1,2443
-1   1  -1  4  1,7643
-1   1   1  4  1,4025
 1  -1  -1  4  1,9700
 1  -1   1  4  1,8990
 1   1  -1  4  2,2960
 1   1   1  4  2,5350
```

(b) Which model has been used in the second Minitab analysis? Is it possible to find the best combination of the factors C, D and E? Motivate your answer using confidence intervals with simultaneous confidence level 95%. (2p)

(c) Suppose we want to do a 2^{5-2} factorial design, according to $D = AB$ and $E = BC$. Which 8 observations out of the 32 in the table above should be used? (1p)

5. When doing a 2^4 factorial design with the factors A, B, C and D we get the following observations.

	y		y
(1)	24.7	d	43.3
a	45.5	ad	39.7
b	8.6	bd	5.4
ab	9.1	abd	9.7
c	75.5	cd	78.8
ac	86.6	acd	77.8
bc	10.0	bcd	21.1
abc	50.1	abcd	37.8

We strive for as large response values as possible. The data material has been analyzed according to a complete model.

```
MTB > set c17
DATA> 24,7 45,5 8,6 9,1 75,5 86,6 10,0 50,1
DATA> 43,3 39,7 5,4 9,7 78,8 77,8 21,1 37,8
DATA> end
MTB > copy c1-c16 m1
MTB > trans m1 m2
MTB > copy c17 m3
MTB > mult m2 m3 m4
MTB > copy m4 c18
MTB > let c18 = c18/16
MTB > set c19
DATA> 1:16
DATA> end
MTB > Sort C18 C19 c20 c21;
SUBC>   By C18.
MTB > print c20 c21


Data Display

Row       C20  C21
  1  -20,0062    3
  2   -4,9562    7
  3   -3,5063   10
  4   -2,4687   16
  5   -1,0563   13
  6   -0,9313   14
  7   -0,6937   11
  8    0,2187    9
  9    1,0562   12
 10    1,2312   15
 11    2,1438    4
 12    2,8062    6
 13    3,6938    8
 14    5,5563    2
 15   15,7312    5
 16   38,9812    1
```

(a) By looking only at the main effects, estimate the expected value that gives the best response, i.e., the largest expected value. (1p)

(b) The data is analyzed according to a reduced model according to the Minitab analysis below. Use this analysis to derive a 95% confidence interval for the expected value ($E(Y)$) that gives the largest response value. (2p)

```
MTB > ANOVA 'Y' = A B C D;
SUBC>   Means A B C D;
SUBC>   GFourpack.

ANOVA: Y versus A; B; C; D

Analysis of Variance for Y

Source  DF       SS      MS      F      P
A        1    494,0   494,0   4,58  0,056
B        1   6404,0  6404,0  59,37  0,000
C        1   3959,6  3959,6  36,71  0,000
D        1      0,8     0,8   0,01  0,934
Error   11   1186,6   107,9
Total   15  12044,9

S = 10,3862   R-Sq = 90,15%   R-Sq(adj) = 86,57%
```

TAMS38
Some extra formulas

Some nonparametric tests

Tukey-Duckworth's quick test

If $4 \leq n_1 \leq n_2 \leq 30$, and $n_2 \leq \frac{4n_1}{3} + 3$ then we can test the null hypothesis H_0 with the test:

1. Find the smallest and largest observation in each sample, respectively.
2. For the sample with the largest observation, count how many observation that is larger in that sample than in the other sample.
3. For the other sample, count how many observations that is smaller then the smallest observation in the first sample.
4. Let C be the sum of the number of observations in 2. and 3. For $\alpha = 0.05$, 0.01 or 0.001, reject H_0 if $C \geq 7$, 10, or 13, respectively.

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 $T_- \approx N\left(\frac{n(n+1)}{4}, \sqrt{\frac{n(n+1)(2n+1)}{24}}\right)$. 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 treatment 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 treatment effect.