Kurskod: TAMS38 - Provkod: TEN1

TAMS38 – Experimental Design and Biostatistics, 4 p / 6 hp
Examination on 15 August 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 are allowed. No extra notes in the formula collection are allowed.
Grading limits: 7-9 points gives 3, 9.5-12 gives 4 and 12.5-15 gives 5.
Examinator: Martin Singull, 013-281447
The result will be *normally* published via LADOK within 12 working days.
Clear answers and justifications are required for each problem.

1) Rats were given one of four different diets at random, and the response measure was liver weight as a percentage of body weight. The responses were:

Treatment			
1	2	3	4
3.52	3.47	3.54	3.74
3.36	3.73	3.52	3.83
3.57	3.38	3.61	3.87
4.19	3.87		4.08
	3.69		4.31
$s_1 = 0.365$	$s_2 = 0.1995$	$s_3 = 0.0473$	$s_4 = 0.229$

We assume normal distribution of data.
a) State appropriate ANOVA model. Compute the overall mean, estimates of treatment effects and estimate of variance. (0.5p+1.5p)
b) Compute the Analysis of Variance table for these data. What would you conclude about the eventual difference between those four diets? Motivate your answer with test on level $\alpha \approx 5\%$. (1p+1p)

2) Let us now skip assumption about normality of data and use appropriate non-parametric method to analyze measurements from Problem 1).
Answer the question: Can we claim with probability 99% that there is significant difference between four diets investigated in Problem 1)? Motivate your answer with a non-parametric test. (3p)

3) Pipe diameter data. A quality engineer at a company that makes plastic pipes is concerned about the consistency of the pipe diameters. The engineer measures 10 pipes a week for three weeks from each of two machines. Each machine can be operated by operator A or B.

Variable	Mean
Week 1	5.019
Week 2	5.474
Week 3	6.029

↓	C1	C2	C3	C4	C5-T	C6
	Week 1	Week 2	Week 3	Machine	Operator	
1	5.19	5.57	8.73	1	A	
2	5.53	5.11	5.01	2	B	
3	4.78	5.76	7.59	1	A	
4	5.44	5.65	4.73	2	B	
5	4.47	4.99	4.93	1	A	
6	4.78	5.25	5.19	2	A	
7	4.26	7.00	6.77	1	B	
8	5.70	5.20	5.66	2	B	
9	4.40	5.30	6.48	1	A	
10	5.64	4.91	5.20	2	B	
11						

a) Which columns include response variable? Which factors do we have for the collected data? (0.5p)

b) Which model/models can be used to analyze pipe diameter data so that all the factors are used?

i) One-way ANOVA. (YES/NO)

ii) Balanced ANOVA with k factors. (YES/NO) If YES state k for this model.

iii) Generalized ANOVA with k factors. (YES/NO) If YES state k for this model.

iv) Regression model. (YES/NO)

v) 2^k factorial design model. (YES/NO) If YES state k for this model.

(1p)

c) We analyze data with ANOVA model that gives us sample standard deviation $s = 0.89$ and $df_E = 25$. Using this information and given for each week means construct confidence intervals that answer the question if there is significant difference between weeks on simultaneous significance level at most 5%. (1p)

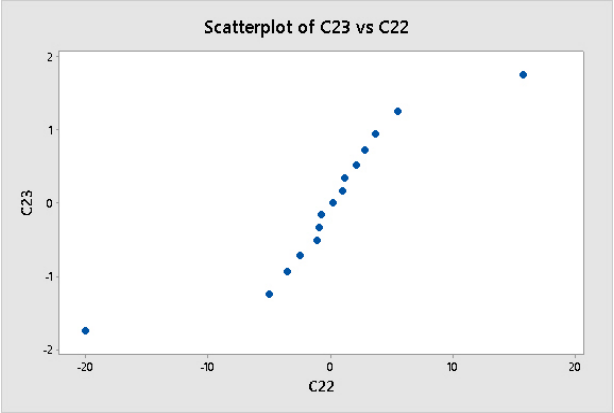
d) Construct confidence t-interval for overall mean μ on $\alpha \leq 1\%$. You can use information about standard deviation and degrees of freedom given in c). (1p)

4) We use 2^4-factorial design with factors A, B, C and D to analyze 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 aim for high response value. The data have been analyzed using matrix calculations for 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
```

```
MTB > copy c20 c22;
SUBC> omit 16.
MTB > nscores c22 c23
MTB > plot c23*c22
```

a) Give two most significant effects. (0.5p)

b) Assume that you want to investigate one more factor E (temperature) but still use only 16 measurements. Which 8 observations would you do in low temperature and which 8 in high temperature?
Blocking rule (of your choice): .....
Measurements in high temperature $E = 1$: .....
Measurements in low temperature $E = -1$: ..... (1p)

c) The same data was analyzed with the reduced ANOVA model. State the model and calculate R^2. (1p)
Minitab output:

```
Analysis of Variance
Source  DF   Adj SS   Adj MS  F-Value  P-Value
  A      1    494.0   493.95     6.22    0.032
  B      1   6404.0  6404.00    80.70    0.000
  C      1   3959.6  3959.56    49.89    0.000
  D      1      0.8     0.77     0.01    0.924
  B*C    1    393.0   393.03     4.95    0.050
Error   10    793.6    79.36
Total   15  12044.9

      S
8.90832

Means
A   N       y        B   N       y         C   N       y        D   N       y
-1  8  33.425        -1  8  58.987         -1  8  23.250        -1  8  38.763
 1  8  44.538         1  8  18.975          1  8  54.712         1  8  39.200

B   C   N       y
-1  -1  4  38.300
-1   1  4  79.675
 1  -1  4   8.200
 1   1  4  29.750
```

d) Using reduced model from c) construct confidence intervals for the choice of best combination of factors A, B and C. Write conclusions. Use $1 - \alpha_{sim} \geq 90\%$. (2p)

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.