Kurskod: TAMS38 - Provkod: TEN1

TAMS38 – Experimental Design and Biostatistics, 6 hp
January 17 2019, 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 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 12 working days.
Clear answers and justifications are required for each problem.

1. The tensile property of a metal wire depends on different treatments. Three different treatments $A_1 - A_3$ were tested and compared to a non-treated raw wire A_4. The tensile property for 11 different wires were

	y_{ij}			
A_1	9.07	8.75		
A_2	9.76	9.88		
A_3	9.28	9.15	9.21	
A_4	8.47	7.38	7.78	8.72

   with $SS_A = 4.6098$ and $SS_E = 1.2083$.

   (a) Given normal distribution, write out the one-factor model and test the hypothesis of equal means for the four wires at level 1%. (1.5p)

   (b) Assume we can NOT use normal distribution, use a non-parametric method to test the hypothesis of equal medians for the four wires at level 1%. (1.5p)

2. Burning cigarettes produce appreciable quantities of carbon monoxide. When cigarette smoke is inhaled, carbon monoxide is combined with carboxyhemoglobin. In a study (*Carbon Monoxide and Exercise in Chronic Bronchitis and Emphysema*, Brit. Med. J., Vol. 283, pp. 877-880, 1981) researchers wanted to determine whether an appreciable concentration of carboxyhemoglobin reduces the exercise tolerance of patients suffering from chronic bronchitis and emphysema. Seven such person was randomly selected and asked to walk for 12 minutes breathing one of four gas mixtures: air, oxygen, air plus CO, oxygen plus CO. The walked distances (in meters) are given in the table below.

Subjets	Gas mixture Air	Oxygen	Air+CO	Oxygen+CO	$\bar{y}_{i\cdot}$	$s_{i\cdot}$
1	835	874	750	854	828.25	54.5
2	787	827	755	829	799.50	35.4
3	724	738	698	726	721.50	16.84
4	336	378	210	279	300.75	72.8
5	252	315	168	336	267.75	75.5
6	560	672	558	642	608.00	57.9
7	336	341	260	336	318.25	38.9
$\bar{y}_{\cdot j}$	547.1	592.1	485.6	571.7		
$s_{\cdot j}$	240.9	240.8	265	248.8		$\bar{y}_{\cdot\cdot} = 549.1$

(a) Give an appropriate model and choose one table below (i, ii, iii or iv) to use in an analysis. (0.5p)

i - Two-way ANOVA: y versus subj, gas

Source	d.f	Sum Sq.
subj	7	1471772
gas	4	44827
Error	18	16316
Total	29	1532915

ii - Two-way ANOVA: y versus subj, gas

Source	d.f	Sum Sq.
subj	6	1471772
gas	3	44827
Error	18	16316
Total	27	1532915

iii - Two-way ANOVA: y versus subj, gas

Source	d.f	Sum Sq.
subj	6	1471772
gas	3	44827
Error	20	16316
Total	27	1532915

iv - Two-way ANOVA: y versus subj, gas

Source	d.f	Sum Sq.
subj	7	1471772
gas	4	44827
Error	16	16316
Total	27	1532915

(b) Use an appropriate test to test at level 5% the null hypothesis: the distance does not depend on the gas mixture. Give both H_0 and H_1. (1p)

(c) Which gas mixture will give the shortest distance? Answer the question with appropriate intervals with simultaneous significance level at most 5%. (1.5p)

3. A canning factory uses many different machines to fill their cans. They suspected that there is great variation in how much each machine fills. They select four machines randomly to test this. Result:

	Machine 1	2	3	4
y_{ij}	1.24	1.20	1.19	1.18
	1.22	1.20	1.20	1.18
	1.22	1.21	1.19	1.19
	1.23	1.22	1.20	1.18
	1.23	1.20	1.21	1.20
$\bar{y}_{i\cdot}$	1.228	1.206	1.198	1.186

Model: $y_{ij} = \mu + \tau_i + \varepsilon_{ij}$, where $\tau_i \sim N(0, \sigma_\tau)$ and $\varepsilon_{ij} \sim N(0, \sigma)$, $i = 1, \ldots, 4$ and $j = 1, \ldots, 5$, independent.

```
Analysis of Variance for y

Source  DF       SS
M        3   0.0047
Error   16   0.0012
Total   19   0.0059
```

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

(b) Give an interval for μ with confidence level 95%. Conclusion? (1.5p)

4. In nitride etching, C_2F_6 is used as the acting gas. It is possible to vary the flow of the gas, the effect on the cathode, the pressure in the etching chamber and the distance between the cathode and the anode. We will use a 2^4 design to analyze the process. The factors are varied at the following levels.

	Factors			
Level	Distance (A) (cm)	Pressure (B) (mTorr)	C_2F_6 flow (C) (SCCM)	Effect (D) (w)
−	0.80	450	125	275
+	1.20	550	200	325

Many different response variables would be interesting to analyze, but in our experiment we measure the etching rate of silicon nitride. High speed is good and we get

(1)	550	d	1037
a	669	ad	749
b	604	bd	1052
ab	650	abd	868
c	633	cd	1075
ac	642	acd	860
bc	601	bcd	1063
abc	635	abcd	729

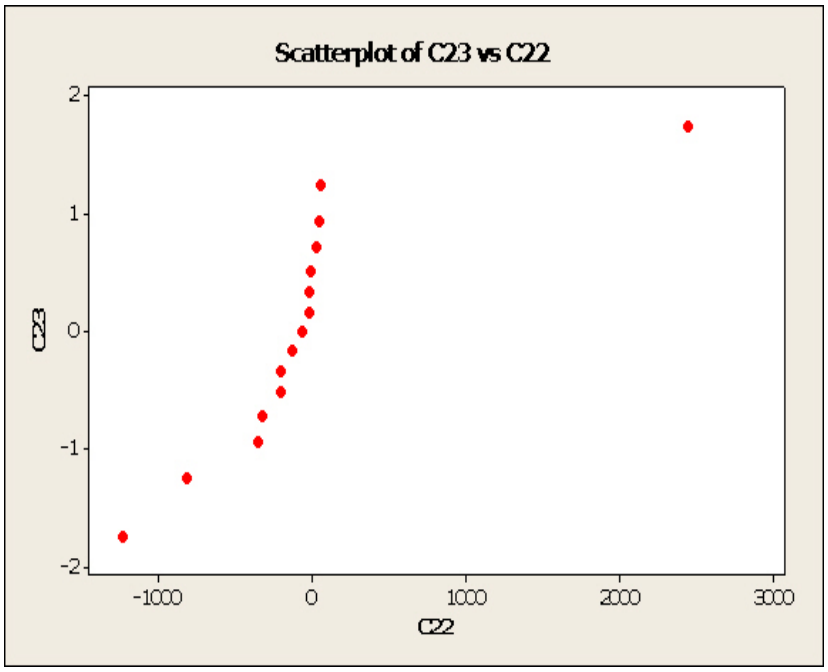
with the following Minitab analysis where the 2^4 design matrix is in c1-c16 and the response variable in c17.

```
MTB > copy c1-c16 m1
MTB > trans m1 m2
MTB > copy c17 m3
MTB > mult m2 m3 m4
MTB > copy m4 c18
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  -1229   10
  2   -813    2
  3   -351    7
  4   -321   16
  5   -203   15
  6   -199    6
  7   -125    8
  8    -63    4
  9    -17   13
 10    -13    3
 11     -5   11
 12     33   12
 13     45   14
 14     59    5
 15   2449    9
 16  12417    1

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



(a) What effects appear to be of the greatest importance based on the Minitab analysis above? (1p)

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

```
ANOVA: y versus A, D

Analysis of Variance for y

Source  DF      SS      MS       F      P
A        1   41311   41311   23.77  0.000
D        1  374850  374850  215.66  0.000
A*D      1   94403   94403   54.31  0.000
Error   12   20858    1738
Total   15  531421

S = 41.6911   R-Sq = 96.08%   R-Sq(adj) = 95.09%

Means
A   N       y
-1  8  826.88
 1  8  725.25

D   N       y
-1  8  623.00
 1  8  929.13
```

```
A   D   N       y
-1  -1  4   597.0
-1   1  4  1056.8
 1  -1  4   649.0
 1   1  4   801.5
```

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

5. In one study, two different methods would be compared A_1 and A_2 to determine the sulfur content of carbon. In total, 14 equivalent samples were distributed to 7 randomly selected laboratories and each laboratory $(B_1, ..., B_7)$ analyzed on sample with A_1 and one with A_2. Result:

	B_1	B_2	B_3	B_4	B_5	B_6	B_7
A_1	0.106	0.124	0.113	0.108	0.096	0.116	0.150
A_2	0.111	0.125	0.115	0.116	0.105	0.119	0.162

Since the laboratories were chosen at random, a random effects model should be used. Method no i and laboratory no j gives the observation y_{ij} from the random variable

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij},$$

where the τ_i:s are fix, $\sum \tau_i = 0$, $\beta_j \sim N(0, \sigma_\beta)$, $\varepsilon_{ij} \sim N(0, \sigma)$ with all the random variables β_j and ε_{ij} independently distributed.

(a) Estimate $\tau_2 - \tau_1$ and show that the estimator is unbiased. (1p)

(b) The usual sum of squares for an additive model are given by

$$SS_A = 0.0001143, \quad SS_B = 0.003856, \quad SS_E = 0.00004971.$$

Estimate σ_β^2 and show that the estimator also is unbiased. (2p)

Hint: $\dfrac{SS_E}{\sigma^2} \sim \chi^2(6)$

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.