Kurskod: TAMS38 - Provkod: TEN1

TAMS38 – Experimental Design and Biostatistics, 4 p / 6 hp
Examination on 19 April 2017, 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 be *normally* published via LADOK within 12 working days.
Clear answers and justifications are required for each problem.

1) A manufacturer suspects that the batches of raw material furnished by his supplier differ significantly in calcium content. There are a large number of batches currently in the warehouse. Five of these are randomly selected for study. A chemist makes five determinations on each batch and obtains the following data:

	Batch 1	Batch 2	Batch 3	Batch 4	Batch 5
	23.46	23.59	23.51	23.28	23.29
	23.48	23.46	23.64	23.40	23.46
	23.56	23.42	23.46	23.37	23.37
	23.39	23.49	23.52	23.46	23.32
	23.40	23.50	23.49	23.39	23.38
$\bar{y}_{i\cdot} =$	23.458	23.492	23.524	23.380	23.364
$s_i =$	0.0687	0.0630	0.0688	0.0652	0.0650

$$\bar{y}_{\cdot\cdot} = 23.444$$

a) Write the model and estimate its parameters. (1.5p)
b) Is there significant variation in calcium content from batch to batch? Use $\alpha = 5\%$. Remember to state your H_0 and H_1. (1.5p)
c) Test if one can assume that Batch 2 and Batch 3 have same variance. Use $\alpha = 5\%$. Remember to state your H_0 and H_1. (1p)

2) An industrial engineer is investigating the effect of four assembly methods (A, B, C, D) on the assembly time for a color television component. Four operators are selected for the study. Furthermore, the engineer knows that time for the earlier and later assembly can be different regarding assembly method procedure. To account for this, the engineer uses the Latin Square design shown below.

Order of Assembly	Operator			
	1	2	3	4
1	C=10	D=14	A=7	B=8
2	B=7	C=18	D=11	A=8
3	A=5	B=10	C=11	D=9
4	D=10	A=10	B=12	C=14

a) State the Latin Square model used to analyze data. (0.5p)

```
General Linear Model: time versus Method, Operator, Order

Analysis of Variance
Source      DF  Adj SS
  Method     3   72.50
  Operator   3   51.50
  Order      3   18.50
Error        6   10.50
Total       15  153.00

Method    Mean          Operator   Mean          Order     Mean
  1      7.500            1       8.000           1      9.750
  2      9.250            2      13.000           2     11.000
  3     13.250            3      10.250           3      8.750
  4     11.000            4       9.750           4     11.500
```

b) Is there a significant difference between assembly methods regarding assembly time? Use $\alpha = 5\%$. Remember to state your H_0 and H_1. (1p)
c) Use confidence intervals to determinate which operator and method gives the shortest assembling time with $\alpha_{sim} \approx 10\%$. (1.5p)

3) Male residents of some city in aged between 40 and 59 were chosen to investigate probability of developing coronary heart disease. During a 6-year follow-up period 1329 males were classified according to several factors, including blood pressure and whether they developed coronary heart disease. The results and Minitab analysis using Logistic Regression ($logit(p) = \beta_0 + \beta_1 * pressure$, where p is probability that subject develops heart disease) for these two variables are given below.

Blood pressure	Heart disease		Total (n)
	YES	NO	
<117	153	3	156
117-126	235	17	252
127-136	272	12	284
137-146	255	16	271
147-156	127	12	139
157-166	77	8	85
167-186	83	16	99
>186	35	8	43
Total	1237	92	1329

```
MODEL 1
Binary Logistic Regression: Yes versus pressure
Response Information
                              Event
Variable  Value      Count    Name
Yes       Event       1237    Event
          Non-event     92
n         Total       1329

Model Summary
Coefficients
Term         Coef  SE Coef
Constant    3.692    0.270
pressure  -0.2694   0.0550

Regression Equation
P(Event)  =  exp(Y')/(1 + exp(Y'))

Y' = 3.692 -0.2694 pressure

Goodness-of-Fit Tests
Test             DF  Chi-Square
Deviance          6        6.39
Pearson           6        6.92
Hosmer-Lemeshow   4        6.75
```

a) Can we claim with probability 95% that logistic model 1 is equally good to the maximal model? Motivate your answer with appropriate test. (1p)
b) Construct 95% confidence interval for β_1 for results obtained for Model 1. (1p)
c) One decided to collect the data also for females and add variable 'Sex' into logistic model 2. The results for new data set are given below. Is the new variable 'Sex' significant on $\alpha = 5\%$? Motivate answer using appropriate test. State H_0 and H_1. (1.5p)

```
MODEL 2
Binary Logistic Regression: Yes versus pressure, Sex

Coefficients
Term           Coef  SE Coef
Constant      2.169    0.156
pressure    -0.1055   0.0342
Sex           0.831    0.134

Goodness-of-Fit Tests
Test             DF  Chi-Square  P-Value
Deviance         13       56.12    0.000
Pearson          13       68.67    0.000
Hosmer-Lemeshow   6       14.87    0.021
```

4) In the experiment one studied three factors on the growth of a certain type of sulfur bacteria. The cultivation of bacteria occurred during the continuous light and at temperature of $25° - 30°C$. The following factors have been used:

	Low level	High level
A ($Na_2S \cdot 9H_2O$)	0.5	1.5
B ($Na_2S_2O_3$)	1	5
C ($NaHCO_3$)	1	5

where concentrations are given in g/l. The following observations were obtained

(1)	15.5	c	14.5
a	7.0	ac	14.0
b	17.0	bc	48.0
ab	14.0	abc	63.0

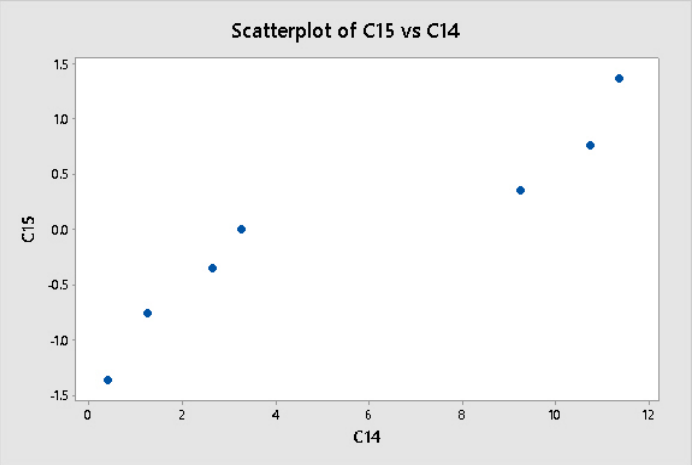
and analyzed using 2^3 factorial design

```
MTB > copy c1-c8 m1
MTB > trans m1 m2
MTB > copy c9 m3
MTB > mult m2 m3 m4
MTB > copy m4 c10
MTB > set c11
DATA> 1:8
DATA> end
MTB > let c12 = c10/8
MTB > Sort c11 c12 c13 c14;
SUBC>   By c12.
MTB > print c13-c14

Data Display

Row  C13     C14
  1    2   0,375
  2    8   1,250
  3    4   2,625
  4    6   3,250
  5    7   9,250
  6    5  10,750
  7    3  11,375
  8    1  24,125

MTB > copy c14 c14;
SUBC> omit 8.
MTB > nscores c14 c15
MTB > plot c15*c14
```



a) Determine three most significant effects. (0.5p)
b) Is it possible to use ANOVA analysis directly instead of given above procedure based on formula $\hat{\xi} = \frac{1}{8}F'y$? Motivate your answer with one sentence. (0.5p)
c) By only looking at main effects estimate the expected value that gives the best response, i.e, the highest expected value. (1p)
d) Using results of ANOVA analysis (given below) and a t-interval, determine level for factor A that give highest response with $\alpha = 1\%$. (1p)
e) Using results of ANOVA analysis (given below) and Tukey-interval method determine level combination for factors B and C that give highest response with $\alpha = 5\%$. (1p)
f) What is simultaneous confidence level for the two sets of intervals you obtained in d) and e)? (0.5p)

ANOVA: Y versus A, B, C

Analysis of Variance for Y

Source	DF	SS
A	1	1.13
B	1	1035.13
C	1	924.50
B*C	1	684.50
Error	3	152.13
Total	7	2797.38

Means

A	N	Y
-1	4	23.750
1	4	24.500

B	N	Y
-1	4	12.750
1	4	35.500

C	N	Y
-1	4	13.375
1	4	34.875

B	C	N	Y
-1	-1	2	11.250
-1	1	2	14.250
1	-1	2	15.500
1	1	2	55.500