UNIVERSITY *of* LIMERICK

OLLSCOIL LUIMNIGH

FACULTY OF SCIENCE AND ENGINEERING

DEPARTMENT OF MATHEMATICS & STATISTICS

END OF SEMESTER ASSESSMENT PAPER

MODULE CODE: MA4605

MODULE TITLE: Chemometrics

LECTURER: Krzysztof Podgorski

EXTERNAL EXAMINER:
Prof. A. Bowman

SEMESTER: Autumn 2009

DURATION OF EXAM: 2.5 hour

GRADING SCHEME: 100%

INSTRUCTIONS TO CANDIDATES

The exam is made of five composite problems worth 20 points each for the maximal total of 100pts. Solutions should be written in the provided answer sheets. Additional computations can be made on the backsides of provided pages. One page of formulas can be brought for the examination (two sides of the page can be used). It can not contain any solutions to the assignment or sample exams. Calculators are allowed.

Part 1 – Question sheet

Problem 1 (20pts)

Below two data sets are described and some of their characteristics are given. The goal is to obtain confidence intervals for the mean value in each of these two cases and interpret the results by answering some questions.

Data Set I. Seven measurements of the pH of a buffer solution gave the following results:

```
5.12 5.20 5.15 5.17 5.16 5.19 5.15
```

These data have been read into an R-variable called `pf` and the following computation using R have been performed: `mean(pf)` and `sd(pf)`. The obtained values have been approximated and these approximations are among the following numerical values: 0.14, 26.5, 0.03, 5.16.

- Please, identify the approximated values of the two characteristics of the data and name them.
- Name the values obtained in the following R code:

```
qt(0.95,6)
[1] 1.943180
qt(0.99,6)
[1] 3.142668
```

- What a distribution is associated with these values?
- Based on the above values evaluate confidence intervals for the mean of the pH and provide with their significance levels.
- Suppose that there is a report claiming that the mean value of the pH is 5.13. Based on your findings from the data would you question this value? Why?

Data Set II. The concentration of lead in the bloodstream was measured for a sample of 50 children from a large school near a busy main road. The sample mean was 10.12 ng/ml and the standard deviation was 0.64 ng ml/l. Additionally, there are following computations made in R

```
qt(0.95,49)
[1] 1.676551
qt(0.99,49)
[1] 2.404892
```

- Compute confidence intervals for the mean lead concentration based on the above values and provide with their confidence levels.
- Name the following values that can be obtained from R: `qnorm(0.95)` and `qnorm(0.99)`. What a distribution do they relate to?
- Could you compute the confidence intervals for the mean lead concentrations if the values from the preceding question were known to you? Justify your answer.
- Could you use these values for computing confidence intervals for Data Set I? Justify.
- Suppose that there are available measurements for 200 instead of just 50 children. What would have happened to the lengths of confidence intervals for the mean lead concentration? Compute approximate confidence interval lengths for this bigger sample size.

Part 1 – Answer sheet

Problem 1 (20pts)

(2pts) The name of the first characteristics is _____ and its approximate value is _____.
The name of the second characteristics is _____ and its approximate value is _____.

(2pts) The calculated values are _____

(1pts) The name of distribution on which these values are based on is _____

(3pts) Confidence intervals: _____ and _____ with the corresponding confidence levels: _____ and _____.

(2pts) Would you question the value 5.13 as the mean pH value? Circle your answer: **Yes/No**.
Explain why:

_____

(3pts) Confidence intervals: _____ and _____ with the corresponding confidence levels: _____ and _____.

(2pts) These are _____ and they relate to _____ distribution.

(2pts) Could you compute the confidence intervals for the mean lead concentration if the values from the preceding question were known to you? Circle your answer: **Yes/No**. Explain why:

_____

(1pts) Could you compute the confidence intervals based on these values for Data Set I? Circle your answer: **Yes/No**. Explain why:

_____

(2pts) What would have happened to the lengths of confidence intervals for the mean lead concentration?

_____

Compute approximate confidence interval lengths for this bigger sample size

_____

Part 2 – Question sheet

Problem 1 (12pts)

Six analysts each made seven determinations of the paracetamol content of the same batch of tablets. The results are shown below

Analyst	Paracetamol content						
A	84.32	84.51	84.63	84.61	84.64	84.51	84.62
B	84.24	84.25	84.41	84.13	84.00	84.30	84.02
C	84.29	84.40	84.68	84.28	84.40	84.36	84.63
D	84.14	84.22	84.02	84.48	84.27	84.33	84.22
E	84.50	83.88	84.49	83.91	84.11	84.06	83.99
F	84.70	84.17	84.11	84.36	84.61	83.81	84.15

The following table has been produced as a result of analysis of these data:

```
Analysis of Variance Table

Response: xx
          Df  Sum Sq Mean Sq F value   Pr(>F)
An         5 0.86108 0.17222  4.2362 0.003941 **
Residuals 36 1.46351 0.04065
---
```

Based on this analysis answer the following questions.

(1pts) Describe what is the purpose of this procedure.

(2pts) Find the total sum of squares SS_T and determine the number of degrees-of-freedom associated with it.

(2pts) Determine the within-sample and between-sample estimators of variance.

(2pts) How are these estimators used to decide if there is a significant difference between analysts?

(2pts) What is the name of the distribution used in the above procedure? Specify its parameters. In what other important statistical procedure is this distribution also used?

(1pts) What is the conclusion following from the above analysis?

(2pts) The following row means (analysts' means) have been computed 84.54857 84.19286 84.43429 84.24000 84.13429 84.27286. Based on these values identify which analysts differ from others in their determinations using the least significant difference method (the 97.5% quantile of t-distriubution with 36 degrees of freedom is equal to `qt(0.975,36)=2.028094`).

Problem 2 (8pts)

It has been observed that measurements of concentration of a certain chemical have a standard deviation of 0.045. It is believed that the true concentration is either 2.0 or 2.06. Propose a statistical procedure based on a sample of measurements which would allow to decide between these two values and such that the chances of making any kind of error in the final claim are at most 5%. How big a sample size is needed for your procedure?

Part 2 – Answer sheet

Problem 1

(1pts) Describe what is the purpose of this procedure

(2pts) The total sum of squares SS_T is [ ] and its number of degrees of freedom is [ ].

(2pts) The within-sample estimator of variance is [ ].

The between-sample estimators of variance is [ ].

(2pts) How are these estimators used to decide if there is a significant difference between analysts?

(2pts) The name of distribution: [ ]; Parameters: [ ].

Used also in: [ ].

(1pts) What is the conclusion following from the above analysis?

(2pts) The least significant difference [ ]

Analysts that significantly differ from A: [ ] from B: [ ]

from C: [ ] from D: [ ] from E: [ ] from F: [ ].

Problem 2

(2pts) Name of the procedure: [ ].

(4pts) Detailed description:

(2pts) Sample size: [ ].

Part 3 – Question sheet

Problem 1 (10pts)

Standard aqueous solutions of fluorescein represented by concentrations 0,2,4,6,8,10,12, pg/ml, are examined in a fluorescence spectrometer and yield the following fluorescence intensities measurements 2.1,5.0,9.0,12.6,17.3,21.0,24.7. The following analysis in R has been performed:

```
Int=c(2.1,5.0,9.0,12.6,17.3,21.0,24.7)
Conc=c(0,2,4,6,8,10,12)

summary(lm(Int~Conc))

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.5179 0.2949 5.146 0.00363 **
Conc 1.9304 0.0409 47.197 8.07e-08 ***
---
Residual standard error: 0.4328 on 5 degrees of freedom
Multiple R-squared: 0.9978,     Adjusted R-squared: 0.9973
F-statistic:  2228 on 1 and 5 DF,  p-value: 8.066e-08
```

Based on this analysis answer the following questions.

(2pts) Is there a strong evidence of linear relation in the data? Explain your answer.

(2pts) Write down explicitly the linear fit to the data.

(2pts) Evaluate the residuals for the above data and plot them against the concentration values.

(2pts) Suppose that a measurement was made to determine some unknown concentration and resulted in the intensity 13.00. To what a concentration does it correspond?

(4pts) The following results have been obtained in R

```
a=coef(lm(Int~Conc))[1]
b=coef(lm(Int~Conc))[2]
sres=0.4328
 bary=mean(Int)
 barx=mean(Conc)
 y_0=13.0
 x_0=(y_0-a)/b
 ssx=sum((Conc-barx)^2)
 Sx_0=(sres/b)*sqrt(1+1/7+(y_0-bary)^2/(b^2*ssx)) #Standard dev

 tn_2=qt(0.975,5)
```

The obtained values of `Sx_0` and `tn_2` are approximately 0.24 and 2.57, respectively. Explain what these values represent. Utilize the values to assess accuracy of the concentration determination found in the previous problem.

(8pts) Suppose that there were two additional measurements based on the same sample of solution that have resulted in the intensities: 12.9, 13.4. How this would affect the answer to the previous two questions? Perform all required computations and compare the obtained results with the one when only one measurement of intensity was available.

Part 3 – Answer sheet

Problem 1

(2pts) Is there a strong evidence of a linear relation in the data? Circle your answer: **Yes/No**. Explain why:

(2pts) The linear fit to the data is:

(2pts) The residuals vs. concentration values plot (mark approximate values of residuals on the plot):

(2pts) The concentration corresponding to the intensity 13.0 is:

(4pts) The values `Sx_0` and `tn_2` which approximately are 0.24 and 2.57, respectively, represent: and , respectively.

The accuracy of determination is expressed by:

(8pts) Having three measurements instead of one for a given concentration requires:

The computations needed in this case follow:

From these computations we conclude that adding more measurements resulted in:

Part 4 – Question sheet

Problem 1 (10pts)

The fluorescence of each of a series of acidic solutions of quinine with concentrations 0,10,20,30,40,50 was determined five times. The mean values and standard deviations of these determinations have been obtained as follows:

```
Means:  4.0  21.2  44.6  61.8  78.0 105.2
StDev:  0.71 0.84 0.89 1.64 2.24 3.03
```

The following two analyses have been performed on the data

```
lm(formula = Means ~ Concentrations)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.9238     2.1648   1.351    0.248
Conc          1.9817     0.0715  27.715 1.01e-05 ***
---
Residual standard error: 2.991 on 4 degrees of freedom
Multiple R-squared: 0.9948,Adjusted R-squared: 0.9935
F-statistic: 768.1 on 1 and 4 DF,  p-value: 1.008e-05
```

```
weights=SdInt^(-2)/mean(SdInt^(-2))
lm(formula = MInt ~ Conc, weights = weights)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.48066    1.15736   3.007   0.0397 *
Conc         1.96315    0.06765  29.018  8.4e-06 ***
---
Residual standard error: 2.034 on 4 degrees of freedom
Multiple R-squared: 0.9953,Adjusted R-squared: 0.9941
F-statistic:   842 on 1 and 4 DF,  p-value: 8.396e-06
```

Answer the following questions.

(3pts) What kind of analyses have been performed above? Write down the fits found by each of the two analyses.

(4pts) Describe differences between the two methods. When one is preferable over the other?

(3pts) Find the concentrations that follow from each of these two fits for the observed intensity of 45.

Problem 2 (10pts)

In an experiment to determine hydrolysable tannins in plants by absorption spectroscopy the following results were obtained:

```
Absorbance (Abs)           0.084 0.183 0.326 0.464 0.643
Concentration (Conc), mg/ml 0.123 0.288 0.562 0.921 1.420
```

The following two analyses have been performed on the data

```
Conc2=Conc^2

lm(formula = Abs ~ Conc + Conc2)
Residuals:
       1        2        3        4        5
-0.004572  0.003127  0.008089 -0.009119  0.002474

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.01651    0.01186   1.392  0.29841
Conc         0.59973    0.03953  15.172  0.00432 **
Conc2       -0.11288    0.02483  -4.546  0.04514 *
---
Residual standard error: 0.009628 on 2 degrees of freedom
Multiple R-squared: 0.9991,Adjusted R-squared: 0.9981
F-statistic:  1065 on 2 and 2 DF,  p-value: 0.0009384
```

```
lm(formula = Abs ~ Conc)

Residuals:
       1        2        3        4        5
-0.026572  0.002299  0.028842  0.014259 -0.018828

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.05829    0.02060    2.83 0.066209 .
Conc         0.42502    0.02544   16.71 0.000467 ***
---
Residual standard error: 0.02646 on 3 degrees of freedom
Multiple R-squared: 0.9894,Adjusted R-squared: 0.9858
F-statistic: 279.1 on 1 and 3 DF,  p-value: 0.0004669
```

Answer the following questions.

(3pts) What kind of analyses have been performed above? Write down the fits found by each of them.

(4pts) Describe differences between the two fits. Examine the residuals and R^2 coefficients. Explain their role in assessing quality of the fits. Do you see any evidence that one fit is preferable over the other? Explain why.

(3pts) Find the concentrations that follows from each of these two fits for the observed intensity of 0.5.

Part 4 – Answer sheet

Problem 1

(3pts) The analysis presented on the left hand side corresponds to: _____

and the resulting fit to the data is _____.

The analysis presented on the right hand side corresponds to: _____

and the resulting fit to the data is _____.

(4pts) The two methods differ in the following aspects:

_____

The analysis presented on the right hand side is preferable over the other one if

_____

(3pts) The concentration corresponding to the intensity 45.0 is:

for the left hand side fit: _____

for the right hand side fit: _____

Problem 2

(3pts) The analysis presented on the left hand side corresponds to: _____ and

the resulting fit to the data is _____.

The analysis presented on the right hand side corresponds to: _____

and the resulting fit to the data is _____.

(4pts) The two methods differ in the following aspects:

_____

After examining the residuals for each fit we may say that:

_____

After examining the R^2 and adjusted R^2 for each fit we may say that:

_____

(3pts) The concentration corresponding to the intensity 0.5 is:

for the left hand side fit: _____

for the right hand side fit: _____

Part 5 – Question sheet

Problem 1 (20pts)

A new microwave-assisted extraction method for the recovery of 2-chlorophenol from soil samples is to be evaluated. There is a question of its performance for five different soils with the main focus on detecting if there is any difference depending on a type of soil. The lab can accomodate only 10 extractions per day and available funding allows for 30 extractions total. Design an experiment for performing the study and address the following issues.

(1pts) Name the type of design that you propose.

(2pts) Identify factors and their levels.

(2pts) With which of the factors would you associate the term *blocks* and which of the factors is a *controlled factor*. Explain the difference.

(3pts) Explain randomization that is needed in your design. Is it a complete randomization or there is a restriction on it? Explain the difference between the two.

(4pts) Provide with a complete description how the conditions of experiments will be set and how the data will be colected.

The following data have been collected and read into *R* variable `Percentages` and the levels of factors in each experiment have been coded to `Factor1` and `Factor2` variables as follows:

```
Percentages
67 68 69 70 82 81 76 78 66 65 76 76 78 76 73 74 75 77 70 70 69 70 87 85 69 68 71 72 80 82
Factor1
"1""1""2""2""3""3""1""1""2""2""3""3""1""1""2""2""3""3""1""1""2""2""3""3""1""1""2""2""3""3"
Factor2
"1""1""1""1""1""1""2""2""2""2""2""2""3""3""3""3""3""3""4""4""4""4""4""4""5""5""5""5""5""5"
```

The following analysis has been performed on the data:

```
Analysis=lm(Percentages~Factor1+Factor2)

Analysis of Variance Table

Response: Percentages
          Df Sum Sq Mean Sq F value    Pr(>F)
Factor1    2 580.20  290.10 18.4760 1.641e-05 ***
Factor2    4  38.67    9.67  0.6157    0.6557
Residuals 23 361.13   15.70
```

(2pts) Do you think that these data have been collected accordingly to your proposed design?

(4pts) Write conclusions to the presented analysis.

(2pts) Have been interactions between factors accounted in the above analysis? Do you think that it is important to account for them? Explain.

Part 5 – Answer sheet

Problem 1

(1pts) Name of the design [ ].

(2pts) Factors [ ], factors' levels [ ]

(2pts) The term *blocks* can be associated with: [ ]. A controlled factor is: [ ] and this is because

[ ]

(3pts) Randomization needed in the design

[ ]

Is it a complete randomization? Circle your answer: **Yes/No**. If your answer is 'No' explain the restriction on randomization:

[ ]

Explain the difference between a complete randomization and a restricted one.

[ ]

(4pts) The conditions of experiments are set and the data are collected in the following way:

[ ]

(2pts) The data have been collected accordingly to your proposed design: Circle your answer: **Yes/No**

(4pts) The conclusion to the presented analysis:

[ ]

(2pts) Have been interactions between factors accounted in the above analysis? Circle your answer: **Yes/No**.

Do you think that it is important to account for them? Circle your answer: **Yes/No**

This is because:

[ ]