TMA4255 Applied Statistics
Exercise 8

MINITAB/R commands are at the end of each problem.

Problem 1

Use the data given in Table 1. We assume that $X_1, ..., X_n$ and $Y_1, ..., Y_m$ all are independent and normally distributed:

$$E(X_i) = \mu_X \; Var(X_i) = \sigma_X^2, \; i = 1, ..., n$$

$$E(Y_j) = \mu_Y \; Var(Y_i) = \sigma_Y^2, \; j = 1, ..., n$$

Assume that $\sigma_X^2 = \sigma_Y^2$, but unknown.

From A (X_j)	5179	5203	5207	5195	5207	5202	5203	5208	5216	5193
From B (Y_j)	5190	5159	5153	5206	5168	5186	5194	5200		

Table 1: Tensile strength for copper wires

a) Put the data into your statistical software (MINITAB or R) and perform a two sample t-test.
Write down the expressions for the statistics computed.
What is being tested here?
What is the conclusion when the significance level of the test is 1%?

b) By using one-way analysis of variance one can examine if the tensile strength of the copper wires are not equal.
Perform the test.
Explain the statistics in the output and what is being tested here. Why is the p-value for this test the same as in a)?

c) Are there different model assumptions if you perform a two-sample t-test or a one-way ANOVA with one factor with two levels?

Investigate if the model assumptions are met for analysis of the copper data using a one-way ANOVA.

MINITAB:

Stat → Basic Statistics → 2-Sample t
Samples in different columns (C1 C2)
Alternative: Not equal
$\surd$ Assume equal variances

Stat → ANOVA → One-way (Unstacked)
Responses: C1 C2

Graph: three in one

Data → Stack → Columns
Stack the following columns: C1 C2

Stat → ANOVA → Test for equal variances
Responses: C2
Factors: C1

R:

```
dsA <- c(5179,5203,5207,5195,5207,5202,5203,5208,5216,5193)
dsB <- c(5190,5159,5153,5206,5168,5186,5194,5200)
t.test(dsA,dsB,var.equal=TRUE)
copper <- c(dsA,dsB)
coppergrp <- c(rep("A",length(dsA)),rep("B",length(dsB)))
obj <- lm(copper~as.factor(coppergrp))
anova(obj)
plot(obj)
```

Problem 2

We will study the a data set taken from the University of Wisconsin, http://pages.stat.wisc.edu/ jzhu/stat571/lympho.txt, and described on page 11-12 of http://pages.stat.wisc.edu/ jzhu/stat571/chap11.pdf.

In short, an experiment is performed to compare the effect of ve drugs on the lymphocyte counts (in thousands per mm3 of blood) in mice. The experiment started out as a balanced randomized assignment of mice to treatments, with 9 mice in each of the five treatment groups. But, due to the surgical procedure involved not all mice survived the procedure and the resulting sample size for the five treatments vary from 6 to 9 mice. The resulting data set is thus unbalanced (sample size varies between treatments).

a) Is there reason to believe that the treatments differ with respect to lymphocyte count?

Perform a one-way ANOVA, assess the model fit and perform the hypothesis test.

b) Assume that the one-way ANOVA hypothesis test above gave a significant result.

We want to perform 4 hypothesis tests to compare the following pairs of treatments: A vs B, B vs C, C vs D and D vs E, but we want to control the familywise error rate at level 0.05. What method for multiple testing adjustment would you like to choose? Perform the selected comparisons using the chosen method for multiple testing. Comment on your findings.

c) Let us now turn to performing all pairwise comparisons between the treatments. Which method for multiple testing adjustment would you now like to choose? Perform the comparisons and comment on your findings.

Comment: the Tukey method is tailored to the balanced case (sample size equal for each treatment), but will perform conservatively with unequal sample sized. The method is then called Tukey-Kramer.

MINITAB: Read in the lympho.MTW file from the course www-page.

Stat → ANOVA → One-way
Respons=count, Factor=drug
Graphs: Four in one
Comparisons: Tick both Tukey (with 5%) and Fisher LSD (with 1.25%)

Stat → ANOVA → Test for equal variances
Respons=count, Factor=drug

R:

```
ds <- read.table("http://www.math.ntnu.no/~mettela/
TMA4255/Data/lympho.txt",header=T)

# one way ANOVA
obj <- lm(count~drug,data=ds)

#equal variances?
leveneTest(obj)

#normality of errors?
plot(obj)

# effect of treatment?
anova(obj)

# pairwise tests, with no adjustment
pairwise.t.test(ds$count,ds$drug,pool.sd=T,p.adjust.method="none")
# choose the 4 tests in question and use
# the Bonf. ind. level of 0.05/4=0.125

#Tukey needs aov object
aovobj <- aov(count~drug, data=ds)
TukeyHSD(aovobj)
```

Problem 3

One wants to examine the machines' capacities in a factory by recording how many units is produced by each machine in a specified time. The 4 workers $A_1, ..., A_4$ are randomly assigned to operate the 4 machines $M_1, ..., M_4$ and then the following observations are recorded:

Worker/Machine	$M1$	$M2$	$M3$	$M4$
A_1	76	77	81	78
A_2	69	71	72	68
A_3	72	78	80	74
A_4	71	74	75	68

a) Assume first that the workers capabilities do not influence the number of units produced.

We want to test if the capacities of the machines differ. Which model can we use?
Perform the test. What is your conclusion?

b) If the workers' influence on the number of units produced were taken into account, what model should then be used and what test would you use?

c) (Calulate by hand!)
What estimator should be used for the number of produced units by machine M_2? Also find a 90% confidence interval for the expectation when you use the estimate of σ^2 from the model in b).

MINITAB: If the data are put in column C1 in Minitab, A's levels (1,2,3,4) in C2 and M's levels in C3 one may use the following commands:

One-way
Stat $\rightarrow$ ANOVA $\rightarrow$ One-way
Response: C1
Factor: C3

Two-way without interactions
Stat $\rightarrow$ ANOVA $\rightarrow$ Two-way
Response: C1
Row factor: C2
Column factor: C3

R:

```
WMds <- data.frame("units"=c(76,77,81,78,69,71,72,68,72,78,80,74,71,74,75,68),
"workers"=as.factor(rep(1:4,each=4)),"machines"=as.factor(rep(1:4,4)))
# one way anova
anova(lm(units~machines,data=WMds))
# two way anova
obj2w <- lm(units~machines+workers,data=WMds)
anova(obj2w)
```

Problem 4

Poor quality of a thread may cause it to break when weaving material. The quality of a thread is here seen as the ability not to break.

The kind of thread that is used for weaving a particular cloth can be spun of one of three kinds of cotton A_1, A_2, A_3 and of one of four kinds of silk $B_1, ..., B_4$. A cloth was weaved using a certain length of thread for each of the 12 possible combinations of thread A_iB_j and the experiment was repeated once. The number of times Y_{ijk} the thread broke during one length was recorded.

The results were:

Let Y_{ijk} be the number of times the thread broke during weaving in the kth trial, $k = 1, 2$ for the combination A_iB_j. Y_{ijk} are assumed to be independent and normally distributed with the same unknown variance σ^2 and expectation

$$E(Y_{ijk}) = \mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \text{ where } \sum_i \alpha_i = \sum_j \beta_j = \sum_i \gamma_{ij} = \sum_j \gamma_{ij} = 0$$

Cotton/Silk	B_1	B_2	B_3	B_4
A_1	65	76	63	62
	68	69	59	69
A_2	61	69	61	72
	63	62	66	71
A_3	51	57	61	61
	53	54	52	67

a) Explain what the parameters μ, α_i, β_j and γ_{ij} represent and find estimators for the parameters.

b) Set up the ANOVA. How would you test if there is a significant interaction between the cotton and slik?

c) Then test if A or B has significant effect. Comment on the results of the analysis.

Minitab

The response values are put in C1 and the level values for A in C2, and the level values for B in C3. Then use the command: Stat→ ANOVA → Two-way, and select response, row and column factors. Do not tick of "fit additive model" unless you do not want to model the interation between A and B.

Alternatively, data are available in a file `DATA9_1.MTW` from the www-page.

R

Read in the data in three separate vectors, one for the reponse, one for the coding av factor A and one for the coding of factor B. Then make a data frame with these three columns and use `lm` and `anova` to perform a two-way ANOVA with interactions.

```
dsy <- c(65,68,61,63,51,53,76,69,69,62,57,54,63,59,61,66,61,52,62,69,72,71,61,67)
dsA <-c(rep(c(1,1,2,2,3,3),4))
dsB <-rep(1:4,each=6)
ds <- data.frame(dsy,as.factor(dsA),as.factor(dsB))
colnames(ds)<- c("y","A","B")

res <- anova(lm(y~A*B,data=ds))
```