Department:	**Chemistry, biotechnology and food sciences**	
Examination in:	**STAT 310**	**Design of experiments and analysis of variance**
	Course code	*Course name*
Time for exam:	**Tuesday Dec 17, 2013**	**14:00-17:30 (3,5 hours)**
	Day and date	*As from – to and duration of examination (hours)*
Person responsible for the course:	**Thore Egeland, 41479582**	
	Name and telephone	

C3: all types of calculators, all other aids

The exam includes:	5
	Number of pages incl. attachment

Each sub-question will be given the same score in the evaluation of the exam. You may answer in English or Norwegian (or "Scandinavian"). There are three exercises. Use significance level 0.05 throughout.

The complete data and excerpts of R output appear in the appendix. Different parts of the data are used for Exercise 1 and 2. The purpose of a study was to investigate the impact of cadmium exposure on vital capacity (a measure of lung capacity). The variables are:

- **group:** A factor variable with two levels:
  - -1 Exposed to cadmium for more than 10 years.
  - 1 Not exposed (control group).
- **age:** A continuous variable giving the age of the individual when data was recorded.
- **factory:** Data was collected from randomly selected factories denoted 'f1', 'f2', 'f3' and 'f4'.
- **vital.capacity:** A continuous variable used as response.

Exercise 1

a) The first analysis was done by a representative of the workers. He considered only one (fixed) effect, **group,** and the model

$$y_{ij} = \mu + \tau_i + e_{ij},$$
$$i = 1,2;\ j = 1,2,\ldots,12,$$
$$y_{ij} = \text{vital.capacity},$$
$$\tau_1 + \tau_2 = 0,$$
$$e_{ij} \sim N(0,\sigma^2). \text{ The } e_{ij} \text{ are assumed independent.}$$

Table 1 gives output based on the model. Fill in the missing numbers denoted (1), (2) and (3).

b) Formulate hypotheses and test if there is a significant difference between the exposed workers and the control group. Formulate a conclusion.

c) Comment on the assumptions of the model in 1a). Refer to Figure 1.

d) A representative of the industry claimed that the difference in lung capacity (**vital.capacity**) can be explained by age differences. His analysis was based on the regression model

$$y_i = \beta_0 + \beta_1 * \text{age}_i + e_i,$$
$$y_i = \text{vital.capacity},$$
$$i = 1,2,\ldots,24,$$
$$e_i \sim N(0,\sigma^2). \text{ The } e_i \text{ are assumed independent.}$$

The output is in Table 2. Formulate a conclusion regarding whether there is a significant effect of age based on this model. What is the estimate of the intercept β_0 and the slope β_1? Draw the regression line.

e) A statistician suggested that an analysis of covariance should be performed to study the effect of **group**, **age** and interaction between **group** and **age** on **vital.capacity.** Formulate the model. Include the assumptions in your description of the model. Test if there is *some* **group** effect based on the output from Table 2 and Table 3; in other words, this corresponds to testing whether the models giving rise to Table 2 and Table 3 differ significantly.

Exercise 2

In this exercise, we only study the impact of the variables **factory** and **group** on the response **vital.capacity.** Four factories were chosen randomly from the cadmium industry. From each factory, three workers exposed more than 10 years were randomly sampled. Furthermore, to obtain a control group, four factories not involved in cadmium production were randomly chosen and three workers were randomly selected from each factory.

a) Explain why it is reasonable to consider **factory** to be a random effect and **group** a fixed effect. Explain why this is a nested design. Draw a figure.

b) Consider the model

$$y_{ijk} = \mu + \tau_i + \beta_{j(i)} + e_{ijk},$$
$$i = 1, 2;\ j = 1, 2, 3, 4;\ k = 1, 2, 3,$$
$$\tau_1 + \tau_2 = 0,$$
$$\beta_{j(i)} \sim N(0, \sigma_\beta^2),\ e_{ijk} \sim N(0, \sigma^2),$$
$$\beta_{j(i)}, e_{ijk} \text{ are all independent.}$$

Some output from the model is given in Table 4. Formulate hypotheses and test if there is a significant difference between the exposed workers and the control group. Formulate a conclusion.

c) Estimate σ^2 and find a 95% confidence interval for σ^2. Interpret the confidence interval. Estimate the correlation coefficient for **vital.capacity** for workers in the same factory.

Exercise 3

a) It was argued that differences in **vital.capacity** can be explained by different smoking habits among exposed workers and those in the control group. We would like to design a new study where **vital.capacity** and **smoking.status** (‘Yes’ or ‘No’) are the only variables recorded. **vital.capacity** is assumed to be normally distributed with known standard deviation 1.1. We require the study to be balanced. Explain what this means. Explain what we need to further specify to calculate the required sample size. (You are not asked to do any calculations *here*.)

b) Assume we would like to find a difference of 0.5 in **vital.capacity** between non-smokers and smokers with probability 0.8. We use significance level 0.05. Determine the required total sample size.

Appendix

vitcap.tmp

	age	factory	vital.capacity	group
1	39	f1	3.7	-1
2	40	f2	5.3	-1
3	41	f3	4.6	-1
4	41	f4	5.5	-1
5	45	f1	2.2	-1
6	49	f2	5.1	-1
7	52	f3	4.3	-1
8	47	f4	4.0	-1
9	61	f1	2.0	-1
10	65	f2	3.0	-1
11	58	f3	2.7	-1
12	59	f4	3.7	-1
13	27	f1	3.7	1
14	25	f2	5.8	1
15	24	f3	5.3	1
16	32	f4	4.3	1
17	23	f1	4.5	1
18	25	f2	4.6	1
19	32	f3	4.6	1
20	18	f4	5.7	1
21	19	f1	4.4	1
22	26	f2	5.2	1
23	33	f3	5.5	1
24	27	f4	5.9	1

```
Table 1

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.4000     0.1986  22.159   <2e-16
group(-1)   -0.5583     0.1986  -2.812   0.0102

          Df  Sum Sq Mean Sq F value  Pr(>F)
group     (1)  7.4817  7.4817  (2)    0.01016
Residuals 22      (3)  0.9463
```

```
Table 2
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.34326    0.51175  12.395 2.13e-11
age         -0.05136    0.01271  -4.042 0.000544

          Df Sum Sq Mean Sq F value    Pr(>F)
age        1  12.06 12.0602  16.338 0.0005444
Residuals 22  16.24  0.7382
```

```
Table 3

            Df  Sum Sq Mean Sq F value Pr(>F)
age          1  4.8510  4.8510  6.3560 0.0203 *
group        1  0.2725  0.2725  0.3570 0.5569
age:group    1  0.7031  0.7031  0.9212 0.3486
Residuals 20 15.2643  0.7632
```

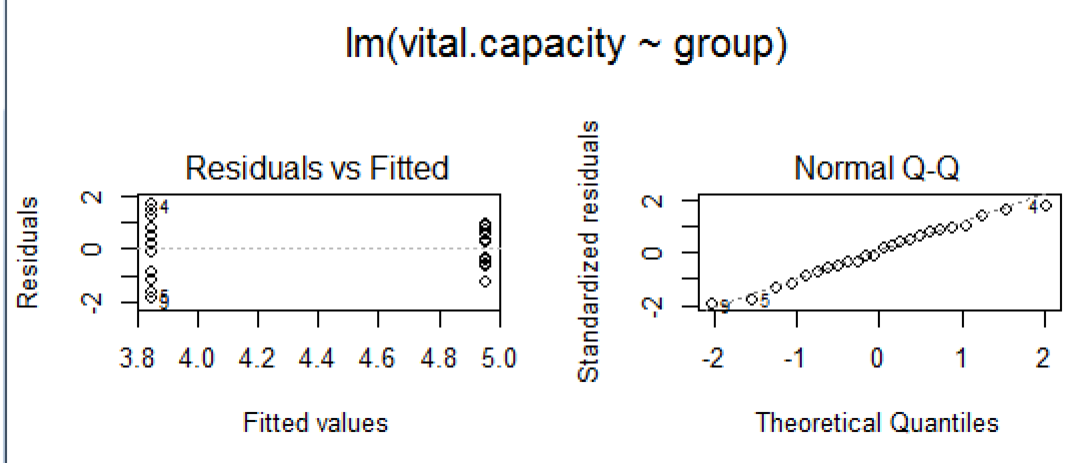
```
Table 4

Response: vital.capacity
               Sum Sq Df Mean Sq
group            7.48  1    7.48
group:factory    8.83  6    1.47
Residuals       11.99 16    0.75

                Err.term(s) Err.df
1 group                 (2)      6
2 group:factory         (3)     16
3 Residuals              -       -
(VC = variance component)

              Expected mean squares
group         (3) + 3 (2) + 12 Q[1]
group:factory (3) + 3 (2)
Residuals     (3)
```

Figure 1



Person responsible for the course: Thore Egeland

External examiner: Torfinn Torp