Contact person during exam:
Mette Langaas (98 84 76 49)

EXAM
KLMED8005 Medical statistics II

June 2, 2010
09:00–13:00

ECTS credits: 7.5

Supporting materials:

- All written and printed material permitted.
- Calculator.

Examination results: June 24, 2010.
Examination results are announced on `http://studweb.ntnu.no/`.

ENGLISH

Problem 1 Smoking and deaths

This problem is based on data presented in Vollset, Tverdal, Gjessing (2006), «Smoking and Deaths between 40 and 70 Years of Age in Women and Men», in *Annals of Internal Medicine.*

In a cohort study women and men from three Norwegian counties were followed over a time period of approximately 25 years, with the aim to investigate the possible association between smoking and death.

At the start of the study the participants were 35–49 years of age (start of study in 1974–1978). Mortality was followed-up through the year 2000. The participants that died during the follow-up period were between 40 and 70 years of age.

Smoking habits were reported at the start of the study and in surveys 5 and 10 years later. Smoking habits were divided into three groups:

- Never smokers.
- Former smokers.
- Continuing smokers.

We look into data from women, presented in the table below.

Smoking habit	Survivied	Died	Total
Never smokers	11000	823	11823
Former smokers	3931	368	4299
Continuing smokers	7241	1142	8383
Total	22172	2333	24505

a) Use a statistical test to investigate possible association between smoking and death. Justify your choice of statistical test.
Use significance level 5 %. What is the conclusion from the test?

b) We now disregard the group of «Former smokers».

We will look at the odds ratio defined as the odds for death among continuing smokers divided by the odds for death among never smokers.

Find a point estimate and a 95 % confidence interval for this odds ratio.
Interpret these results.

c) In the cohort study the mortality of the participants were followed-up through the year 2000. For all participants the age at death (within the year 2000) or age at the end of the study was recorded.

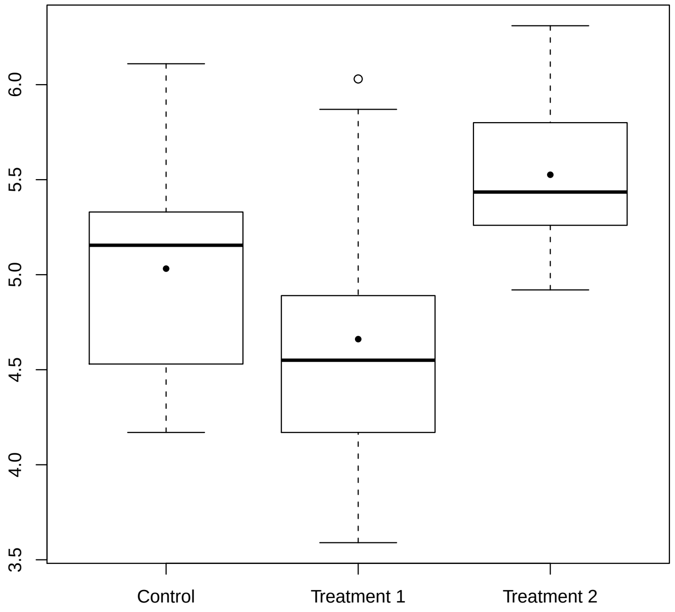
Information about age at the debut of smoking, physical activity, duration of education, county and marital status, were collected at the start of the study.

How can this information be used in an analysis of the data?
Which statistical methods can be used?

Problem 2 Yield

We will look at data from an experiment to compare yields (as measured by weight in grams of dried plants) obtained under three conditions: one control condition («Control») and two different treatment conditions («Treatment 1» and «Treatment 2»). The data set consists of 10 observations for each condition, that is, 30 observations in total.

In the graph below the you find a boxplot of the distribution of yield for the three conditions. The solid dots represent the mean yield for each condition.



A one-way analysis of variance model was fitted to the data, and the results are presented in the table below.

Source of variation	Df	SS	MS	F value
Between treatments	2	3.77	1.88	4.85
Error (within treatments)	27	10.49	0.39	
Total	29	14.26	0.49	

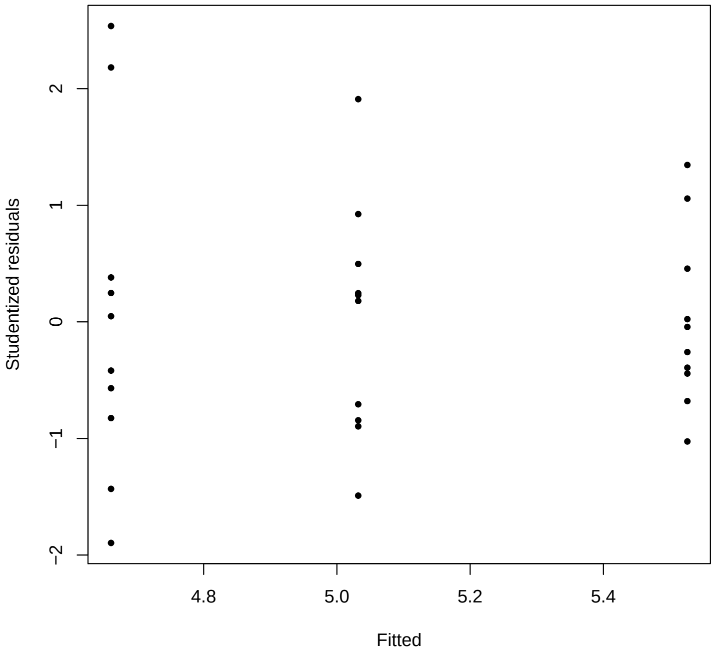
a) Write down the model that this analysis of variance is based on.
Give a brief explanation of the goal of the analysis of variance.

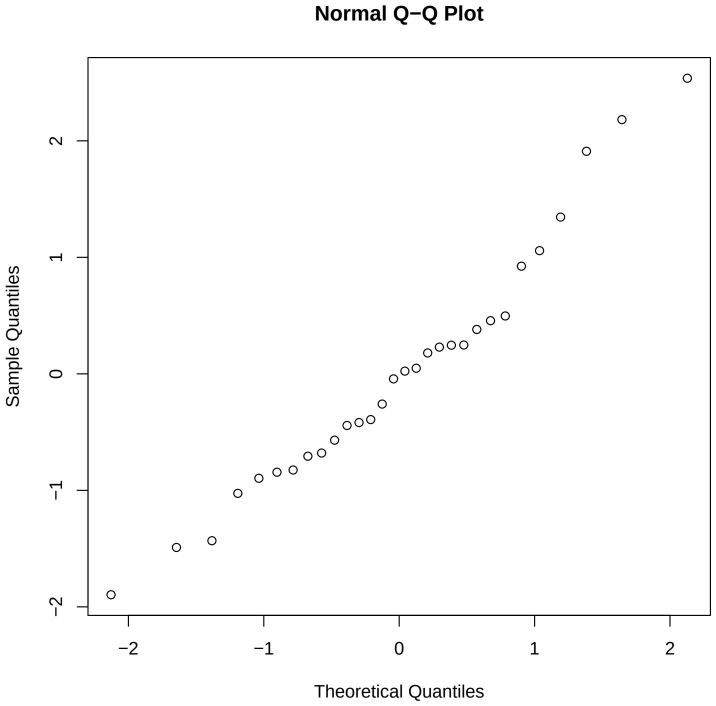
On the next page you find two graphs. In the graph on top of the next page studentized residuals are plotted versus predicted yield, and the graph on the top of next page shows a normal quantile-quantile plot of studentized residuals. Both graphs are based on the fitted one-way analysis of variance model for yield. What can you conclude from examining these graphs?

b) Write down the null hypothesis tested in a one-way analysis of variance.
Using a 5 % significance level, what can you conclude from this test when you use the numercial results presented in the analysis of variance table on the top of this page.

How would you go about to compare the yield produced by each of the three condiionts? You only need to give a brief explaination of how this is done, since you do not have enough information to perform the analyses numerically.

What is the relationship between your suggested method and a two-sample t-test?


Studentized residuals
2
1
0
−1
−2
4.8
5.0
5.2
5.4
Fitted


Normal Q−Q Plot
Sample Quantiles
2
1
0
−1
−2
−2
−1
0
1
2
Theoretical Quantiles

Problem 3 Coronary heart disease

We will analyse a data set from an American heart study. The data set consists of measurement from 1615 men in the age group 31–65 years.

Our aim is to look at association between coronary heart disease (0 = no, 1 = yes) and

- age (in years),
- systolic blood pressure (mmHg),
- smoking (0 = non-smoker, 1 = smoker), and
- serum cholesterol (mg/dl).

Age, systolic blood pressure, smoking status and serum cholesterol were registered and measured at the start of the study. The follow-up period for this data set is eight years, and we look at the development of coronary heart disease during the follow-up period.

We have fitted a logistic regression to the data. Age, systolic blood pressure and serum cholesterol are all continuous variables. For the categorical variable smoking our analyses are relative to the category that is coded with the smallest number (e.g. 0). Results from the logistic regression are presented in the table below. One of the numbers in the table is on purpose deleted and replaced with a question mark (?).

Variable	Regression coefficient B	Standard deviation for B	p-value	$\exp B$
Constant	-10.015	0.941	<0.0001	-
Age	0.058	0.012	<0.0001	1.060
Systolic blood pressure	0.014	0.004	0.0007	1.014
Smoking	0.611	0.251	?	1.843
Serum cholesterol	0.010	0.002	<0.0001	1.010

a) In the table presenting the results from the logistic regression (on the previous page) one of the columns is given the name «p-value». What is the interpretation of this p-value?

Consider the variable «smoking», where we have not reported a p-value.
Write down the null hypothesis and the alternative hypothesis and perform the hypothesis test. Use significance level 5 %.
What is the conclusion from the test?

b) Look at the table with results from the logistic regression (on the previous page). In the column named «$\exp B$» for the variable «systolic blood pressure» the value reported is 1.014. What is this an estimate of?

Find a 95 % confidence interval for the odds ratio to develop coronary heart disease when the systolic blood pressure increases with 5 mmHg.

c) How will you interpret the results from the logistic regression presented in the table on the previous page?

Estimate the probability of developing coronary heart disease during the follow-up period for a man that at the start of the study is 45 years of age, has a systolic blood pressure of 131 mmHg, smokes, and has a serum cholesterol level of 227 mg/dl.