Norwegian University of Science and Technology
Department of Mathematical Sciences

Contact during the exam:
Mette Langaas (988 47 649)

ENGLISH

EXAM IN TMA4255 APPLIED STATISTICS

Wednesday, August 8, 2012
Time: 9:00–13:00

Number of credits: 7.5.

Permitted aids: All printed and handwritten material. Special calculator.

Grading finished: August 30, 2012.

Exam results are annonunced at `http://studweb.ntnu.no/`.

Note that:

- Significance level 5% should be used unless a different level is specified.
- All answers need to be justified.

Problem 1 Use of cannabis among youths in Norway

Norwegian Institute for Alcohol and Drug Research is responsible for maintaining a statistics database, RusStat, on alcohol, drugs and tobacco.

The data for use of cannabis among youths are shown below for the years 2004–2008. The row "Total" gives the total number of youths surveyed for each year.

	2004	2005	2006	2007	2008	All years
Used	222	261	206	357	281	1327
Not used	1432	1482	1365	2891	2394	9564
Total	1654	1743	1571	3248	2675	10891

a) We would like to investigate if there has been a change in the frequency of use of cannabis over the years 2004–2008. Write down the null- and alternative hypothesis and perform one hypothesis test using the information given in the table above. To easen the computational burden you may use that the χ^2-test statistic equals 27.7 and only show the calculation of one of the 10 terms in the sum.
What is the conclusion from the test?

Problem 2 Process development

We will look at a designed experiment to develop an etching process. There are three design factors (A, B and C). The three design factors were run at two levels each.

- A: The gap between the electrodes. Low: 0.80 cm. High: 1.20 cm.
- B: The flow of gas. Low: 125. High: 200.
- C: The power applied to the cathode. Low: 275 W. High: 325 W.

The response variable is the etch rate for the process (Å/m). A high value for the etch rate is desired. A 2^3 factorial design was run, and the results are presented in Table 1.

Run	A	B	C	Response
1	−1	−1	−1	550
2	1	−1	−1	669
3	−1	1	−1	633
4	1	1	−1	642
5	−1	−1	1	1037
6	1	−1	1	749
7	−1	1	1	1075
8	1	1	1	729

Table 1: The etch experiment with one replicate.

From these results we have the following design of experiment (DOE) effect estimates.

A	B	C	AB	AC	BC	ABC
−126.5	*	274.0	−42.0	−190.5	−9.5	13.0

a) Fill in the missing effect estimate for factor B. Construct a main effects plot for factor B, and explain with words how the estimated main effect of factor B is interpreted.

In addition to this first replicate of the experiment a second replicate was made, such that each factor combination was run twice, that is $n = 16$. The result from fitting a multiple linear regression model to these data, with intercept, A, B, C, AB, AC, BC and ABC as covariates, is given in the printout below. We call this the *full* regression model. Note that the estimates ("Estimate") are the estimated regression coefficients and not the effect estimates.

	Estimate	Std. Error	t value	p-value
Intercept	776.062	11.865	65.406	3.32e-12
A	−50.812	11.865	−4.282	0.002679
B	3.688	11.865	*	0.763911
C	153.062	11.865	12.900	1.23e-06
AB	−12.437	11.865	−1.048	0.325168
AC	−76.812	11.865	−6.474	0.000193
BC	−1.062	11.865	−0.090	0.930849
ABC	2.813	11.865	0.237	0.818586
Residual standard error: 47.46 on 8 degrees of freedom				

b) What does the "Std. Error" column give, and why is the "Std. Error" the same for all covariates?

Now turn to factor B, in the second row of the printout table. What is the relationship between an estimated coefficient for B and the estimated effect for B? Calulate the missing number for the t-statistic. What are the hypotheses underlying the p-value and what is the conclusion of that test?

What are the significant covariates in the model?

c) We now assume that we use a regression model with intercept and covariates A, C and AC, and get the following printout for this new regression model. Let us call this the *reduced* regression model.

	Estimate	Std. Error	t value	p-value
Intercept	776.06	10.42	74.458	<2e-16
A	−50.81	10.42	−4.875	0.000382
C	153.06	10.42	14.685	4.95e-09
AC	−76.81	10.42	−7.370	8.62e-06

Residual standard error: 41.69 on 12 degrees of freedom

Why are the estimated coefficients for A, C and AC in the reduced model the same as in the full model? Why have the estimated standard deviations changed from the full to the reduced model?

What would you suggest are the optimal choices of level (low or high) for each of these two factors when the aim is to use the fitted regression model to arrive at the combination with the highest estimated etch rate?

Calculate a 95% prediction interval for the etch rate based on your chosen levels for A and C.

d) We now assume that in a pilot study with three factors only runs 1, 4, 6 and 7 of Table 1 were performed.

What type of experiment is this?
What is the generator and the defining relation for the experiment?
Write down the alias structure of the experiment.
What is the resolution of the experiment?

Problem 3 Gene activity

We will look at the results from a biological study where the aim was to compare the gene activity of three genes (B_1, B_2, and B_3) in three different cell populations (A_1, A_2 and A_3).

The gene activity was measured using the polymerase chain reaction method. Measurements are continuous and will be analysed on the logarithmic scale (base 2). In total $n = 45$ measurements were made.

a) A two-way analysis of variance model (ANOVA) with interaction was fitted to the data, and gave the following results.

Source	Df	Sum Sq	Mean Sq	F value	p-value
A	2	3.240	*	5.935	*
B	2	253.161	126.581	463.768	<2e-16
AB	4	*	30.645	112.277	<2e-16
Error	36	9.826	0.273		
Total	*	388.807	8.837		

Four of the entries in the result table are each replaced with an asterisk. Explain the column headings and calculate numerical values for these four missing entries. Is the interaction term AB significant? How would you now proceed for further analyses?

For the rest of the problem we will only look at cell population A_1.

b) The biologist would like to compare the gene activity for gene B_1 and B_2 for cell population A_1. There were $n_1 = 3$ observations for gene B_1 in cell population A_1 and the average was $\bar{x} = -29.1$ and the empirical standard deviation $s_1 = 0.34$. Further, there were $n_2 = 9$ observations for gene B_2 in cell population A_1 and the average was $\bar{y} = -33.7$ and the empirical standard deviation $s_2 = 0.42$.

Perform a test to investigate if the gene activity for the two genes differ in this cell population. List the assumptions you need to make to perform the test.

c) Let μ_1 be the expected gene activity on the logaritmic scale (base 2) for gene B_1 in cell population A_1, and let μ_2 be the expected gene activity on the logaritmic scale (base 2) for gene B_2 in cell population A_1. The biologist is interested in the expected *fold change* between the activity of the two genes on the original scale, that is

$$\gamma = \frac{2^{\mu_1}}{2^{\mu_2}}.$$

Based on two independent random samples of size n_1 and n_2 from the two groups suggest an estimator, $\hat{\gamma}$, for γ.

Use approximate methods to find the expected value and variance of this estimator, that is, $\mathrm{E}(\hat{\gamma})$ and $\mathrm{Var}(\hat{\gamma})$. Use the sample sizes, averages and empirical standard deviations given in b) above to calulate $\hat{\gamma}$ numerically and give estimated numerical values for $\mathrm{E}(\hat{\gamma})$ and $\mathrm{Var}(\hat{\gamma})$.

Hint: You may use that $2^x = \exp(x \ln 2)$, where ln is the natural logarithm.