Norges teknisk–
naturvitenskapelige universitet
Mathematical Sciences

Contact during exam:
Jo Eidsvik
Tlf. 90127472

EXAM IN TMA4255

Experimental design and applied statistical methods

Monday 6. august 2007

Time: 09:00–13:00

Assisting material:
All printed and handwritten material. All calculators.

Sensur: 27. august 2007

Problem 1

To compare three different types of investment funds (called A, B, C), 2000 kroner was invested in 18 funds, 6 of each type, in a period of five years. For each investment the pay-off (in kroner) was registrered and is given in the table:

A	B	C
13288	15738	14790
12782	14249	13827
12812	12369	13680
11713	12822	13150
11201	12117	12669
12233	12605	14267

We assume a model as follows

$$Y_{ij} = \mu_i + \epsilon_{ij}; \quad i = 1, 2, 3; \; j = 1, 2, \ldots, 6$$

where the ϵ_{ij}s are independent and $N(0, \sigma^2)$.

A modified printout from MINITAB is presented below. Some numbers are not given and replaced by *.

One-way ANOVA: A; B; C

```
Source  DF        SS      MS     F      P
Factor   2   6134625       *     *  0,085
Error   15  15725266       *
Total   17  21859890

S = *   R-Sq = 28,06%   R-Sq(adj) = 18,47%


                             Individual 95% CIs For Mean Based on
                             Pooled StDev
Level  N   Mean  StDev  -------+---------+---------+---------+--
A      6  12338    777  (----------*----------)
B      6  13317   1401            (----------*-----------)
C      6  13731    760                (-----------*----------)
                        -------+---------+---------+---------+--
                           12000     12800     13600     14400

Pooled StDev = *
```

a) Specify the null hypothesis that is tested.

Compute the test statistic F. Find the critical value and draw a conclusion on the test using a significance level of 5%.

Explain how you can also read this conclusion directly from the MINITAB printout?

b) Let S_A, S_B, S_C be estimators for σ, each using data from their fund only. Show how the estimator S that is used in (a) can be computed from these three estimators.

Compute S based on S_A, S_B, S_C in the lower half of the MINITAB printout. Compute also S directly from the upper part of the MINITAB printout. (Answer: $S = 1024$).

Assume next that one wishes to test whether the expectation for fund A is equal to the mean expectation for the two other funds, i.e. to test

$$\tilde{H}_0 : \mu_1 = \frac{\mu_2 + \mu_3}{2} \text{ against } \tilde{H}_1 : \mu_1 \neq \frac{\mu_2 + \mu_3}{2}$$

Define

$$U = \bar{Y}_{1\cdot} - \frac{\bar{Y}_{2\cdot} + \bar{Y}_{3\cdot}}{2}$$

c) Show that under hypothesis $\tilde{H}_0$ one has $U \sim N(0, \sigma^2/4)$.
Use this result to explain why

$$T = \frac{2U}{S}$$

is a natural test statistic to test $\tilde{H}_0$ against $\tilde{H}_1$.
What is the probability distribution of T under $\tilde{H}_0$?
What is the conclusion of this test given the data?

Problem 2

Genetic variability is assumed to play a central role in the survival of species. According to this theory a population with many 'different' individuals has a better chance of adapting to new conditions. The table shows results of a study designed to test the hypothesis of genetic variability experimentally.

Two populations of banana flies, where population A is more heterogeneous, while population B is more homogeneous, was put in designed cans with minimum food and space. Every 100 days the number of individuals in each population was counted.

Day number (x)	Population A (yA)	Population B (yB)
0	100	100
100	250	203
200	304	214
300	403	295
400	446	330
500	482	324

The underlying models are assumed to be as

$$y_i^A = \beta_0^A + \beta_1^A x_i + \epsilon_i^A, \quad i = 1, \ldots, 6 \qquad (1)$$

$$y_i^B = \beta_0^B + \beta_1^B x_i + \epsilon_i^B, \quad i = 1, \ldots, 6 \qquad (2)$$

where all ϵ terms are independent, with $\epsilon_i^A \sim N(0, \sigma_A^2)$, $\epsilon_i^B \sim N(0, \sigma_B^2)$.

Below we present the MINITAB printout from each model separately.

Regression Analysis: yA versus x

```
The regression equation is
yA = 145 + 0,742 x

Predictor      Coef  SE Coef     T      P
Constant     145,33    26,87  5,41  0,006
x           0,74200  0,08874  8,36  0,001

S = 37,1219   R-Sq = 94,6%   R-Sq(adj) = 93,2%
```

Regression Analysis: yB versus x

```
The regression equation is
yB = 131 + 0,452 x

Predictor      Coef  SE Coef     T      P
Constant     131,33    22,77  5,77  0,004
x           0,45200  0,07522  6,01  0,004

S = 31,4648   R-Sq = 90,0%   R-Sq(adj) = 87,5%
```

In Figure 1 the data are plotted with the fitted regression lines.

a) Does the data imply that the population sizes increase with time? Formulate this as hypothesis tests concerning the parameters in the models.

Use the MINITAB printout to test the hypotheses. What are the p-values?

b) Let S_A og S_B be estimators for σ_A og σ_B.

Explain why

$$F = \frac{S_B^2}{S_A^2} \cdot \frac{\sigma_A^2}{\sigma_B^2}$$

is Fisher distributed with 4 and 4 degrees of freedom.

Use this result to create a 95% confidence interval for $\frac{\sigma_A}{\sigma_B}$.

What is your conclusion to a 5%-level hypothesis test on $\sigma_A = \sigma_B$, when the alternative is that they are unequal?

(*Note:* You can find computed estimates S_A og S_B in the MINITAB printouts.)

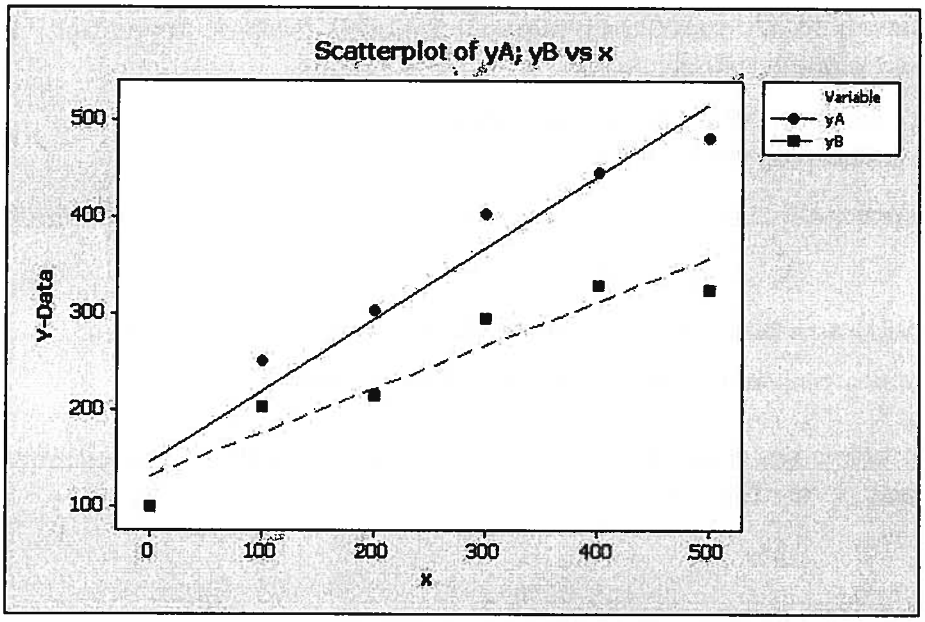

Figure 1: Survival of banana flies

Based on the conclusion from the previous task we will next assume that $\sigma_A^2 = \sigma_B^2 = \sigma^2$.

The β_1 parameters in (1)-(2) denote the development of the populations and can be used as indicators for survival among the two populations. We will now test

$$H_0:\ \beta_1^A = \beta_1^B \text{ against } H_1:\ \beta_1^A > \beta_1^B$$

c) Let $\hat{\beta}_1^A$ and $\hat{\beta}_1^B$ be least squares estimators for β_1^A og β_1^B under the two regression models in (1)-(2).
What is the expectation and variance of the difference $\hat{\beta}_1^A - \hat{\beta}_1^B$?
(You can use that $\text{Var}(\hat{\beta}_1^A) = \text{Var}(\hat{\beta}_1^B) = \sigma^2 / \sum_{i=1}^{6}(x_i - \bar{x})^2$ without proving this).
Explain heuristically why

$$T = \frac{\hat{\beta}_1^A - \hat{\beta}_1^B}{S_{AB}} \tag{3}$$

is a natural test statistic for testing H_0 against H_1, where S_{AB} is the estimated standard deviation for $\hat{\beta}_1^A - \hat{\beta}_1^B$.

One can show that under H_0 T is t distributed with 8 degrees of freedom (you do not have to prove this).

d) For the given data we get $T = 2.49$. Used this to carry out the hypothesis test of H_0 against H_1. What is the conclusion?

The testing procedure above is based on separate regression models for each population. An alternative model is to have a unified model

$$y_i = \begin{cases} \beta_0 + \beta_1^A x_i + \epsilon_i \text{when } i \text{ is observed in population A} \\ \beta_0 + \beta_1^B x_i + \epsilon_i \text{when } i \text{ is observed in population B} \end{cases} \quad (4)$$

where i is an index running over all 12 observations, and where the constant term β_0 is common for the two populations.

e) Phrase the model in (4) as a multiple regression problem on matrix-vector form.

Why does this seem as a reasonable model for the current situation?

In the MINITAB printout below the main results of model (4) are presented. The estimated correlation betweeen $\hat{\beta}_1^A$ and $\hat{\beta}_1^B$ for this model is 0.52.

Parameter	Estimate	Std. Error	T-value
β_0	138,33	16,77	8,25
β_1^A	0,76109	0,06358	11,97
β_1^B	0,43291	0,06358	6,81

The T statistics given in (3) can be used to test H_0 in this case too, but with new values for the difference and its standard deviation. In this case T under model (4) and under H_0 is t distributed with 9 degrees of freedom (you do not have to prove this). The observed T is 5.268.

f) Provide a heuristic argument for getting 9 degrees of freedom in this model, rather than 8 which we had above.

What is the conludion when testing H_0 based on model (4)?

Why is it natural to get a smaller p-value in this case?

Problem 3

The table shows the number of survivors at the *Titanic* split in groups indicating classes. Classes are matroses, first to third class. The numbers in parantheses show the expected values based on a chi-square test for this data. The chi-square test statistic is $X = 187.79$.

	Matros	1. class	2. class	3. class	Total
Survived	212	202	118	178	710
	(285.48)	(104.84)	(91.94)	(227.74)	
Dead	673	123	167	528	1491
	(599.52)	(220.16)	(193.06)	(478.26)	
Total	885	325	285	706	2201

We want to test whether the chances of survival is the same no matter which class you had on the ship. Express this as a null hypothesis.

Carry out the test and draw the conclusion.

What is the most surpring elements according to the null hypothesis?