NTNU – Trondheim
Norwegian University of
Science and Technology

Department of Mathematical Sciences

Examination paper for **TMA4255 Applied statistics**

Academic contact during examination: Anna Marie Holand

Phone: 951 38 038

Examination date: 30 May 2014

Examination time (from–to): 09:00-13:00

Permitted examination support material: All printed and handwritten material. Special calculator.

Other information:

- In outputs from MINITAB comma is used as decimal separator.
- Significance level 5% should be used unless a different level is specified.
- All answers need to be justified.

Language: English

Number of pages: 6

Number pages enclosed: 0

Checked by:

Date Signature

Problem 1 Darwins corn plants

Darwin (1876) studied the growth of pairs of corn plants, where one plant was produced by cross-fertilization and the other produced by self-fertilization. His goal was to demonstrate the greater fitness (e.g. survival and growth) of cross-fertilized plants compared to self-fertilized plants.

Fifteen pairs of plants were grown together under identical conditions. The data recorded are the height (in inches) of the plants in each pair.

For pair i let X_{1i} denote the height of the plant from the seedling produced by cross-fertilization and X_{2i} denote the height of the plant from seedling produced by self-fertilization, $i = 1, ..., 15$. Further, let $D_i = X_{1i} - X_{2i}$. Data from the experiment is presented below.

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
x_{1i}	188	96	168	176	153	172	177	163	146	173	186	168	177	184	96
x_{2i}	130	163	160	160	147	149	149	122	132	144	130	144	102	124	144

Descriptive measures for this dataset are $\bar{x}_1 = \frac{1}{15}\sum_{i=1}^{15} x_{1i} = 161.53$,
$s_{x1} = \sqrt{\frac{1}{14}\sum_{i=1}^{15}(x_{1i} - \bar{x}_1)^2} = 28.94$, $\bar{x}_2 = \frac{1}{15}\sum_{i=1}^{15} x_{2i} = 140$,
$s_{x2} = \sqrt{\frac{1}{14}\sum_{i=1}^{15}(x_{1i} - \bar{x}_2)^2} = 16.64$, $\bar{d} = \bar{x}_1 - \bar{x}_2 = 21.53$,
$s_d = \sqrt{\frac{1}{14}\sum_{i=1}^{15}(d_i - \bar{d})^2} = 38.29$.

(Source: Darwin, C. (1876). The Effect of Cross- and Self-fertilization in the Vegetable Kingdom, 2nd Ed. London: John Murray.)

a) Assume that X_{1i} and X_{2i} are normally distributed, $X_{1i} \sim N(\mu_1, \sigma^2)$ and $X_{2i} \sim N(\mu_2, \sigma^2)$, $i = 1, ..., 15$.

Based on this experiment, could Darwin conclude that cross-fertilized plants are taller than self-fertilized plants? Write down the null hypothesis and the alternative hypothesis, choose a test statistics and perform a hypothesis test. Use significance level $\alpha = 0.05$.
Specify the assumptions you make.

b) Assume that X_1 and X_2 are not normally distributed. Perform a sign test to test if cross-fertilized plants are taller than self-fertilized plants.
Comment on your findings.

Problem 2 House sparrows

House sparrow (Passer domesticus) populations along the coast of the central and northern Norway have been studied by scientists at NTNU for decades.

The male house sparrows display a black plumage patch on their chest (badge). The badge is variable in size and earlier studies indicate that the size of the badge is a signal of social status or fitness. In a study the following variables were measured

- y, visible badge size measured in mm.
- x_1, the length of the tarsus (foot) measured in mm.
- x_2, the bill depth measured in mm.
- x_3, the total badge size measured in mm.

Scientist believe that the visible badge size, y, is dependent on the length of their tarsus, x_1, (which indicates general body size), the bill depth, x_2, (that tells us something about the ability to eat food of certain type) and the total badge size, x_3, (which tell us something about the potential for visible size of the badge).

These variables were measured for $n = 901$ males house sparrows. We analyze the data using a multiple regression model:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3_i} + \epsilon_i, \tag{1}$$

where ϵ_i is i.i.d. $N(0, \sigma^2)$ for $i = 1, ..., n$.

The MINITAB output from a statistical analysis is found in Figure 1 and Figure 2 and plots of standardized residuals are found in Figure 3.

a) Perform a t-test to test the null hypothesis H_0: $\beta_1 = 0$ against the alternative hypothesis H_1: $\beta_1 \neq 0$. Use a 5% level of significance.

What is the percentage of variability explained by the regression?

Predictor	Coef	SE Coef	T	P
Constant	4,966	2,020	2,46	0,014
x1	0,11584	0,06609	?	?
x2	-0,6629	0,1905	-3,48	0,001
x3	0,71509	0,04441	16,10	0,000

S = 1,54791   R-Sq = 24,1%   R-Sq(adj) = 23,8%

Figure 1: Printout from statistical analysis of the house sparrow data set.

Analysis of Variance

Source	DF	SS	MS	F	P
Regression	3	682,54	227,51	?	0,000
Residual Error	897	?	?		
Total	900	?			

Source	DF	Seq SS
x1	1	14,11
x2	1	47,08
x3	1	621,36

Figure 2: Printout from statistical analysis of the house sparrow data set.

b) What is an appropriate estimate for σ^2?

In the printout from performing an analysis of variance on the fitted multiple linear model in Figure 2 four numerical values are substituted by question marks. Calculate numerical values for each of these, and explain what each of the numbers means.

c) Use the estimated regression model in Figure 1 to calculate a point prediction for visible badge size, $\hat{y}_0$, for observed tarsus length $x_1^0 = 20$, bill depth $x_2^0 = 8$ and total badge size $x_3^0 = 19$. It is given that the estimated standard deviation of $\hat{y}_0$ is 0.097.

Calculate a 95% confidence interval for the expected value of $\hat{y}_0$ and a 95% prediction interval for y_0.

What is the difference in interpretation of the two intervals?

d) Which model assumptions were made in the linear regression?

Based on the plots presented in Figure 3, do you think that these assumptions are satisfied? Justify your answer.

How do you think this effects the results given in the outputs?

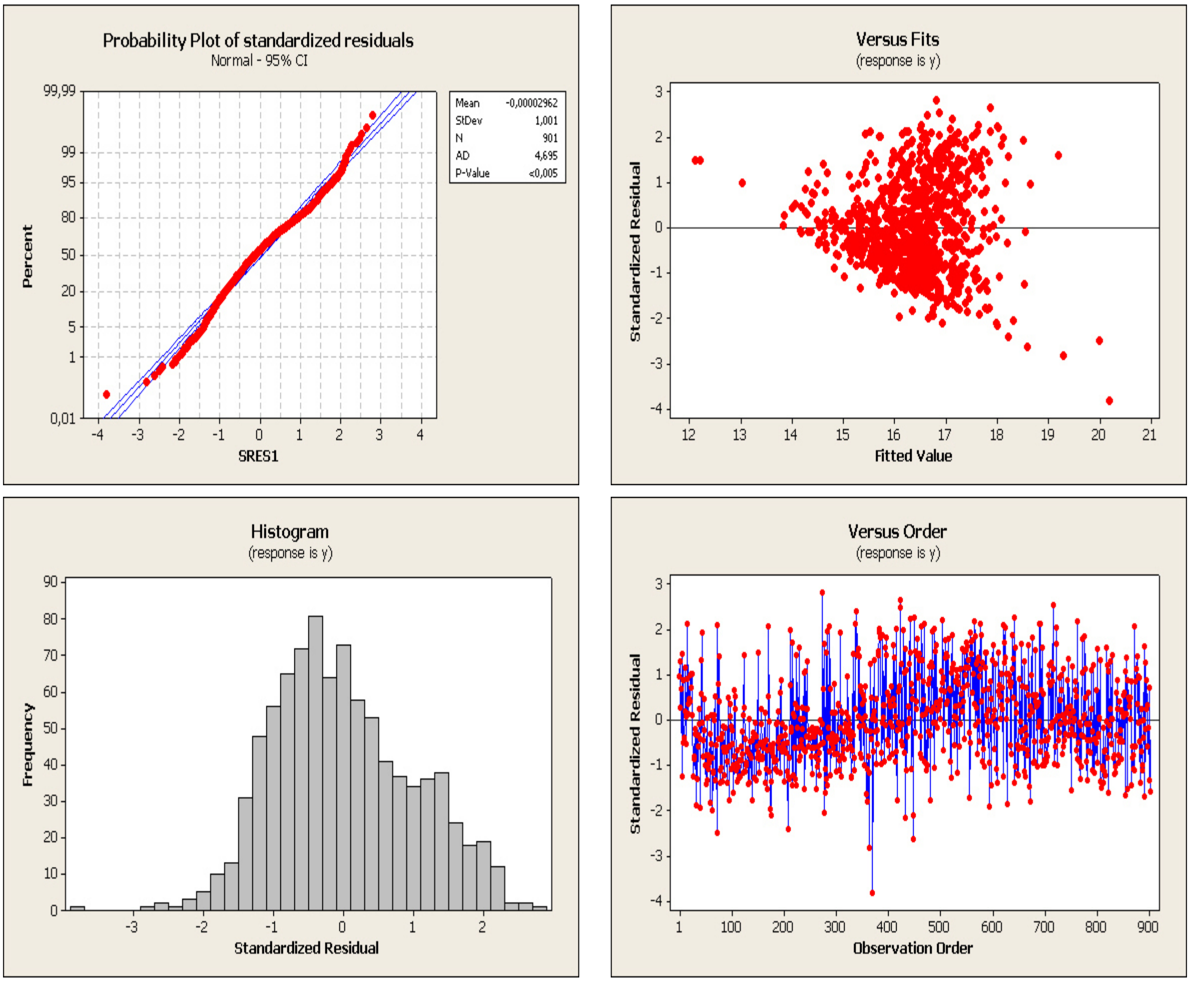


Figure 3: Residual plots (normal plot based on standardized residuals in the upper left panel, standardized residual versus fitted values in the upper right panel, histogram based on standardized residuals in lower left panel and standardized residual versus observation order in the lower right panel) for the regression model in Equation (1) for the house sparrow data set.

Problem 3 Forging of piston rings

A factory produces piston rings for an automotive engine by a forging process. The factory is interested in controlling the inside diameter of the piston rings.

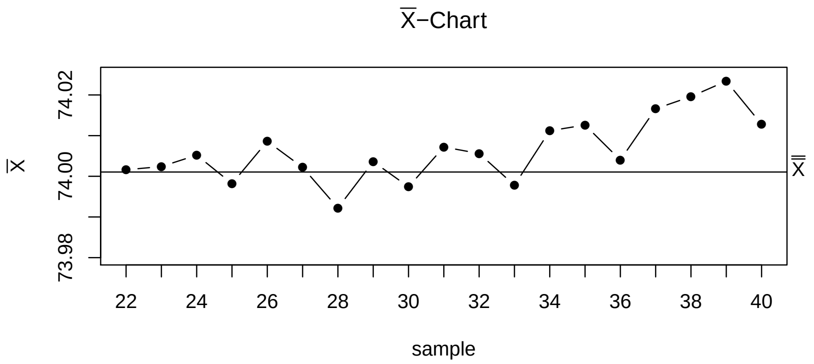
21 samples, each of size 5 were taken. The process is considered to be in control when the samples were taken.

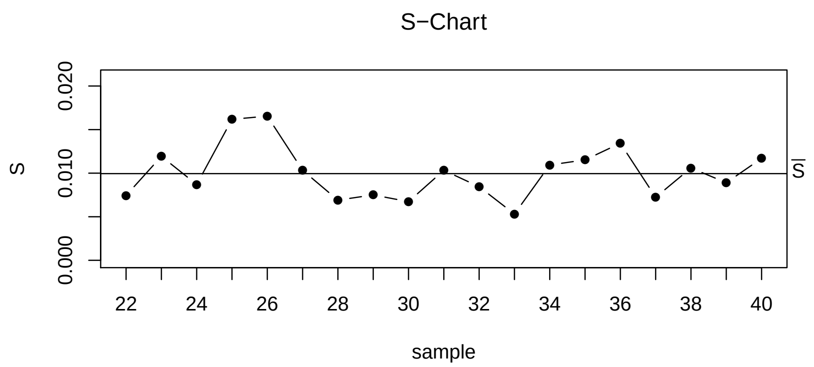
Let X_{ij} be the measure of diameter for piston ring j, on sample i, where $j = 1, 2, 3, 4, 5$ and $i = 1, 2, ..., 21$. Further, $\bar{X}_i = \frac{1}{5}\sum_{j=1}^{5} X_{ij}$, $S_i = \sqrt{\frac{1}{4}\sum_{j=1}^{5}(X_{ij} - \bar{X}_i)^2}$, $\bar{\bar{X}} = \frac{1}{21}\sum_{i=1}^{21} \bar{X}_i$, and $\bar{S} = \frac{1}{21}\sum_{i=1}^{21} S_i$.

Based on these 21 samples, we find $\bar{\bar{x}} = 74.001$ and $\bar{s} = 0.00995$.

a) Construct a S-chart and a $\bar{X}$-S-chart (with 3σ limits).

New samples were taken from the process, shown in the figure below. Are the process in control? Justify your answer.





Problem 4 Good and bad husbands and wives

A study of the independence of the temperament of husbands and wives was conducted. 111 married couples were randomly selected and a relative of the couple crossclassified the husband and wife into either having good or bad temperament.

	Good wife	Bad wife
Good husband	24	27
Bad husband	34	26

a) Is it any reason to believe that the temperament (good/bad) of the husband is dependent on the temperament (good/bad) of the wife? Write down the null hypothesis and the alternative hypothesis and perform a hypothesis test on the basis of the table above. Use a 5% level of significance.