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: August 2014

Examination time (from–to):

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: 9

Number pages enclosed: 0

Checked by:

Date Signature

Problem 1 Reiki treatment of Fibromyalgia

Fibromyalgia is a chronic pain condition of unknown cause, affecting $2-4\%$ of the population. Patients with fibromyalgia often use complementary and alternative medicine, such as Reiki. Reiki is a form of energy medicine, where the healer channels "universal life energy" through light touch.

A study was conducted to explore the usefulness of Reiki as management of pain. Hundred patients with fibromyalgia were recruited. To serve as placebo treatment the study included an actor giving the Reiki treatments, in addition to a Reiki master.

Pain was measured using a visual analogue scale (VAS) (0 = no pain, 10 = worst pain ever).

(Source: Nassim Assefi, M.D., Andy Bogart, M.S., Jack Goldberg, Ph.D., and Dedra Buchwald, M.D. (2008). Reiki for the Treatment of Fibromyalgia: A Randomized Controlled Trial. Journal of Alternative and Complementary Medicine, 14(9): 1115-1122.)

a) The 100 patients were randomized into 4 groups, each of size $n = 25$, receiving treatment either with direct touch or no touch (distant) from either a Reiki master (referred to as Reiki treatment) or an actor (referred to as placebo treatment). The VAS score at enrollment (e.g. before treatment) was measured for each of the $n = 25$ patients in each of the 4 groups. Let group A denote patients receiving treatment from a Reiki master with direct touch, group B denote patients receiving treatment from a Reiki master with no touch (distant), group C denote patients receiving treatment from a actor with direct touch and group D denote patients receiving treatment from a actor with no touch (distant).

The descriptive measures for the different groups in VAS score at enrollment (e.g. before treatment) are given in Table 1.

Groups of pain	Sample size	Mean	Standard deviation
Group A	25	6.3	2.2
Group B	25	6.4	2.6
Group C	25	6.8	2.1
Group D	25	6.1	2.4
Total	100	6.4	

Table 1: Descriptive measures of groups A, B, C and D in the Reiki treatment dataset.

To be able to compare VAS scores after treatment across groups, the VAS score should be equal across groups at enrollment (e.g. before treatment). From this study can we conclude that the enrollment score (e.g. before treatment score) differs between groups, A, B, C and D?

Write down the null hypothesis and the alternative hypothesis, perform one hypothesis test based on the descriptive measures above. Use significance level $\alpha = 0.05$.
Specify the assumptions you make and the conclusion of the test.

b) The researchers were interested in comparing the direct touch treatment, either by a Reiki master or an actor.

Let X_i denote the VAS score after treatment by the Reiki master with direct touch, $i = 1, ..., 25$, and Y_j denote the VAS score after treatment by the actor with direct touch, $j = 1, ..., 25$.

Assume that X_i and Y_j are normally distributed, $X_i \sim N(\mu_X, \sigma^2)$, $i = 1, ..., 25$ and $Y_j \sim N(\mu_Y, \sigma^2)$, $j = 1, ..., 25$, respectively.

Descriptive measures for this dataset are $\bar{d} = \bar{x} - \bar{y} = -0.4$ and $s_p^2 = \frac{1}{48}(\sum_{i=1}^{25}(x_i - \bar{x})^2 + \sum_{i=1}^{25}(y_i - \bar{y})^2) = 1.23^2$, where $\bar{x} = \frac{1}{25}\sum_{i=1}^{25} x_i$ and $\bar{y} = \frac{1}{25}\sum_{i=1}^{25} y_i$.

Based on these data, do we have reason to believe that Reiki treatment has an larger effect on pain, measured as VAS score, than placebo treatment given by an actor? 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.

Problem 2 Cheddar cheese Taste

As cheddar cheese matures, various chemical processes take place that determine the taste of the final product. The concentration of several chemicals in this chemical process is related to the taste.

In an observational study of cheddar cheese taste, a sample of $n = 30$ mature cheddar cheeses were analyzed for their chemical composition and were subjected to taste tests. Overall taste scores were obtained by combining the scores from several tasters. The following variables were measured for each cheddar cheese.

- y: Subjective taste test score, average taste score ranged from 0.7 to 57.2.
- x_1: Natural log of concentration of acetic acid in the cheese.
- x_2: Natural log of concentration of hydrogen sulfide in the cheese.
- x_3: Concentration of lactic acid in the cheese.

First, three separate simple regressions were used to study the relationship between the taste and each of the variables,

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

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

The MINITAB outputs from statistical analyses are found in Figure 1, 2 and 3.

```
Predictor     Coef  SE Coef       T      P
Constant    -61,50    24,85   -2,48  0,020
X1          15,648    4,496    3,48  0,002

S = 13,8212   R-Sq = 30,2%   R-Sq(adj) = 27,7%
```

Figure 1: Printout from statistical analysis of the cheddar cheese data for model in Equation (1).

```
Predictor     Coef  SE Coef       T      P
Constant    -9,787    5,958  -1,64  0,112
X2          5,7761   0,9458   6,11  0,000

S = 10,8334   R-Sq = 57,1%   R-Sq(adj) = 55,6%
```

Figure 2: Printout from statistical analysis of the cheddar cheese data for the model in Equation (2).

```
Predictor     Coef  SE Coef       T      P
Constant    -29,86    10,58  -2,82  0,009
X3          37,720    7,186   5,25  0,000

S = 11,7450   R-Sq = 49,6%   R-Sq(adj) = 47,8%
```

Figure 3: Printout from statistical analysis of the cheddar cheese data for the model in Equation (3).

a) From Figure 1 we see that the acetic acid variable, X_1, is significant at a 5% significance level. What can you conclude from the given p-values in Figure 1, 2 and 3 about the three chemicals influence on taste? Comment on the values of R^2 in Figure 1, 2 and 3. Justify your answer.

Find a 90% confidence interval for β_1.

What is the predicted taste score for an acetic acid value of $x_1^0 = 7$?

b) Find an appropriate estimate for σ, and calculate a 90% confidence interval for σ in the regression model in Equation (1) (Hint: use that SSE$/\sigma^2$ is chi-square distributed).

How can we use this confidence interval to test the null hypothesis $H_0 : \sigma = 1$? Write down the alternative hypothesis used, give the conclusion of this test and the significance level.

Further, regression of taste on all three chemicals were performed.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \epsilon_i \tag{4}$$

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

The MINITAB output from statistical analysis of this three-variable model is found in Figure 4. A correlation matrix and a pairwise scatter plot are found in Figure 5 and Figure 6, respectively. Residual plots are found in Figure 7.

```
The regression equation is
y = - 28,9 + 0,33 X1 + 3,91 X2 + 19,7 X3

Predictor    Coef  SE Coef      T      P
Constant   -28,88    19,74  -1,46  0,155
X1          0,328    4,460   0,07  0,942
X2          3,912    1,248   3,13  0,004
X3         19,671    8,629   2,28  0,031

S = 10,1307   R-Sq = 65,2%   R-Sq(adj) = 61,2%

Analysis of Variance

Source          DF      SS      MS      F      P
Regression       3  4994,5  1664,8  16,22  0,000
Residual Error  26  2668,4   102,6
Total           29  7662,9
```

Figure 4: Printout from statistical analysis of the cheddar cheese data for the model in Equation (4).

```
Correlations: X1; X2; X3

       X1     X2
X2  0,618

X3  0,604  0,645
```

Figure 5: Pearson correlation between variable X_1, X_2 and X_3 in the cheddar cheese data set.

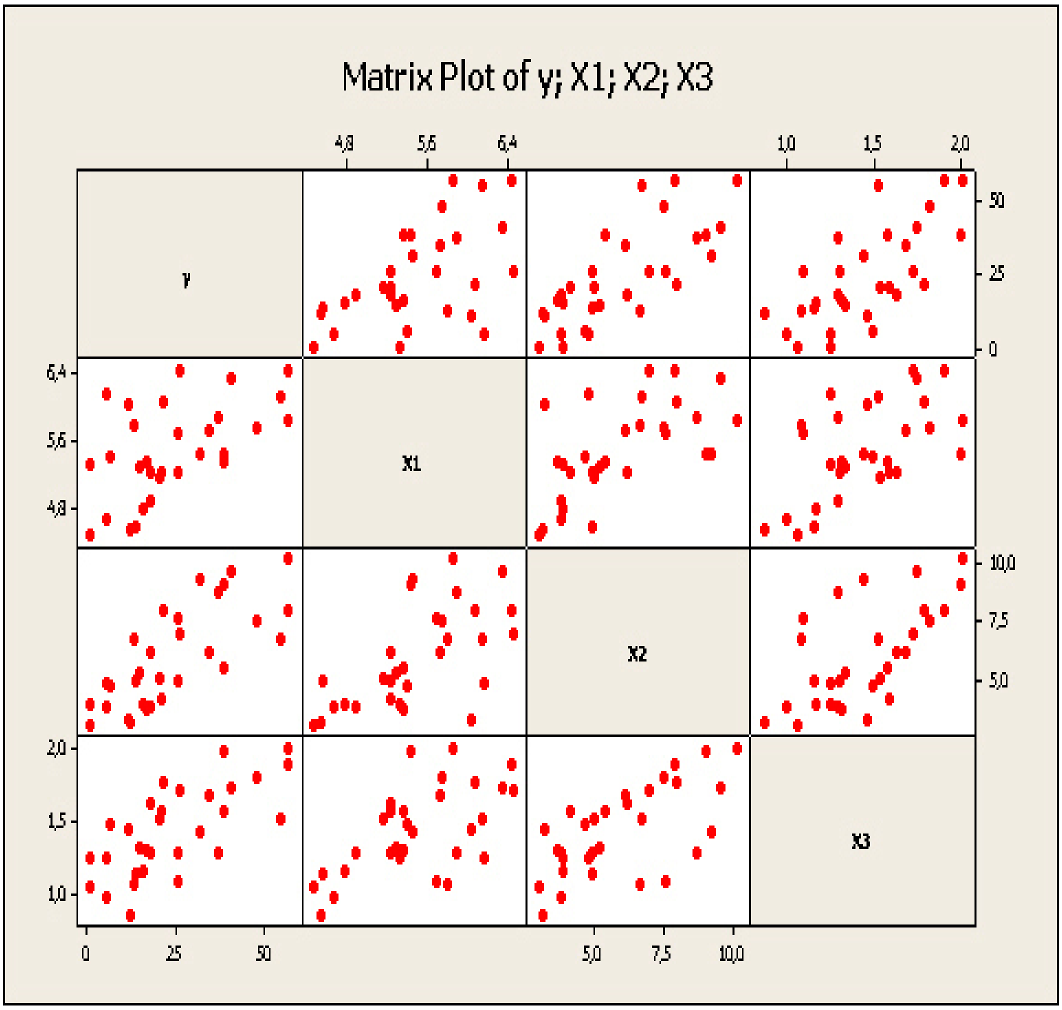

Figure 6: Pairwise scatterplot of the variables in the cheddar cheese data set.

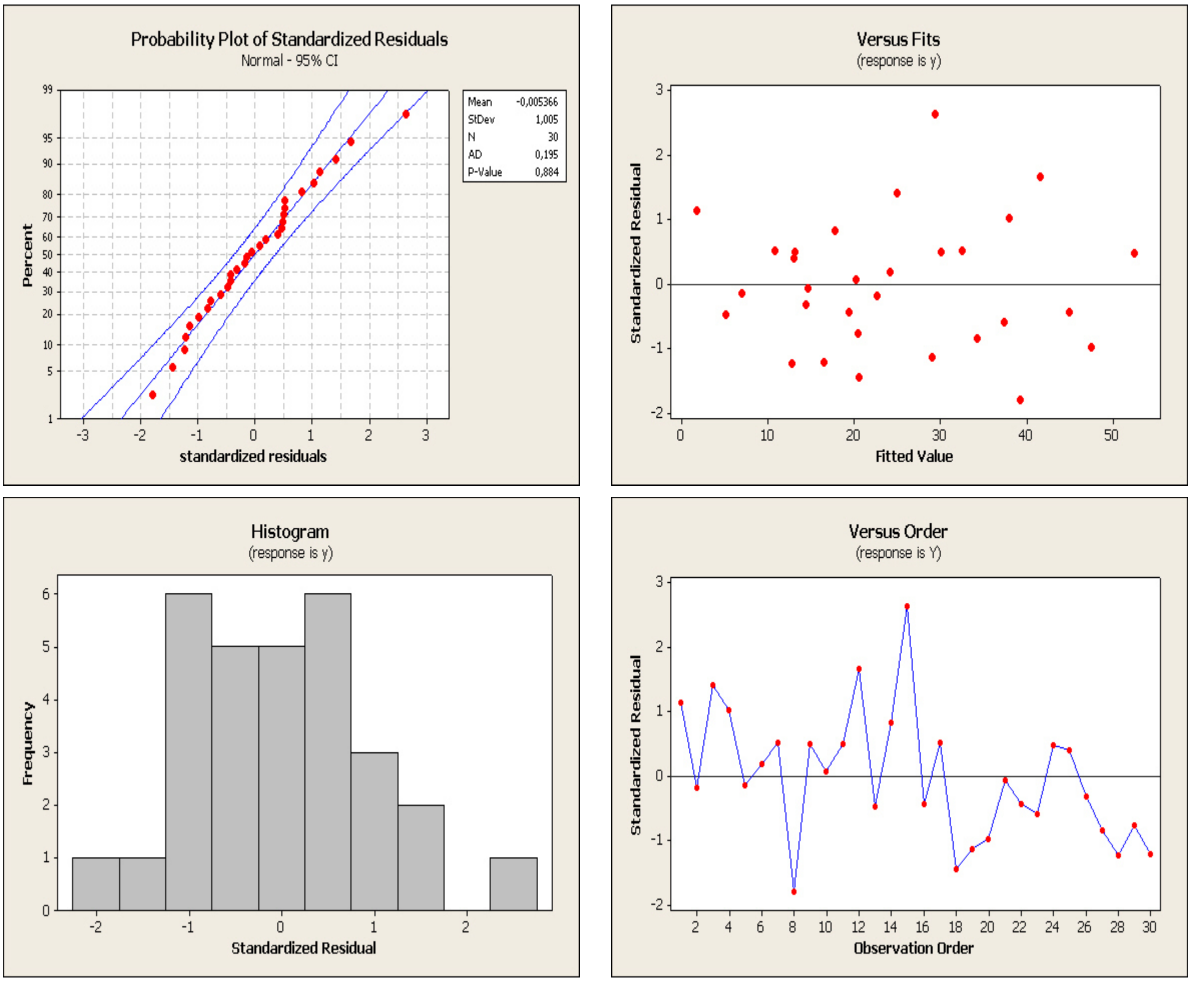

Figure 7: Residual plots (normal plot based on standardized residuals in the upper left panel, standardized residual versus fitted values in the upper right panel, histogram based on standardized residuals in lower left panel and standardized residual versus observation order in the lower right panel) for the regression model in Equation (4) for the cheddar cheese data set.

c) Explain the term multicollinearity. Could this be a problem in the regression model in Equation (4)? Comment on the correlation matrix in Figure 5 and pairwise scatter plot in Figure 6.
From the three-variable regression models in Figure 4 we see that acetic acid, X_1 is not significant, but in Figure 1 (Equation (1)) X_1 was significant. What may be the reason behind this? Justify your answer.

This is an observational study. Would it be possible to design an experiment (design of experiment) to investigate the problem under study? Elaborate.

Problem 3 Toy plastic bricks

A factory produces interlocking toy plastic bricks. To be able to assemble the bricks, they must have no defects. The factory takes samples to ensure that the bricks are of good quality. Every day a sample of $n = 250$ random bricks is taken and each brick classified as either OK or defect. The results after two weeks are given below. The process is assumed to be in control during this two weeks of sampling.

Day i	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Number of defects	3	3	1	3	4	6	4	6	4	4	5	2	3	1

a) Make the appropriate control chart to control the probability of defects, p. Use all the data in the table above to calculate control limits (use 3σ limits). Can we assume that the number of defects in each sample is approximately normally distributed?

b) The factory is interested in detecting that the process is out-of-control when an increase in the probability of defects has occurred. How many observations, n, in each rational subgroup is needed to detect a change from $p = 0.2$ to $p_1 = 0.21$?

Problem 4 Obesity and alcohol intake

In a study exploring the connection between obesity and alcohol intake in Australia, a random sample of $n_{\text{Low}} = 165$ persons with low measure of obesity, $n_{\text{Average}} = 161$ persons with average measure of obesity and $n_{\text{High}} = 165$ persons with high measure of obesity were selected and their alcohol intake in drinks per day, was registered. The alcohol intake was classified into four groups and the following cross-tabulation was observed.

	Alcohol Intake				
Obesity	0	1-2	3-5	6+	Total
Low	45	45	41	34	165
Average	39	32	46	44	161
High	33	37	47	48	165
Total	117	114	134	126	491

a) Based on these data can we conclude that Low, Average and High obesity populations differ with respect to their alcohol intake? 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. It is given that the χ^2-test statistics equals 6.952. You need to show the calculation of only one of the 12 terms in the sum. What is the conclusion from the test?