NTNU

Norwegian University of
Science and Technology

Department of Mathematical Sciences

Examination paper for **ST0103 Statistics with applications**

Academic contact during examination: Jarle Tufto

Phone: 99 70 55 19

Examination date: 3 December 2016

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

Permitted examination support material: C: Specific simple calculator. Tabeller og formler i statistikk (Tapir akademisk forlag). One yellow A4 sheet with your own handwritten notes.

Other information:
Some formulas for use in regression analysis are given in enclosures.

You should demonstrate how you arrive at your answers (e.g. by including intermediate answers or by referring to theory or examples from the reading list).

In the grading, each of the ten points counts equally.

Language: English

Number of pages: 5

Number of pages enclosed: 1

Checked by:

Informasjon om trykking av eksamensoppgave	
Originalen er:	
1-sidig □	**2-sidig** ☒
sort/hvit ☒	**farger** □
skal ha flervalgskjema □	

Date Signature

Problem 1

At a botanical institute one investigated the occurrence of a certain type of fungus in soil from moor. At a certain moor area one took cylinders of soil with an earth auger, washed the roots, and then determined the proportion of root cells with fungi. The proportion was recorded in percent.

Let X be the result of such a sample. From long experience it is considered as known that X is normally distributed with expectation 42.0 and standard deviation 4.0. Results from different samples are assumed to be stochastically independent.

a) What is the probability that a measurement of X is less than 35?

Find the probability that a measurement is between 35 and 45.

Draw the two probabilities as areas in a simple sketch of the probability density of X.

b) Assume in this point that four samples were taken.

What is the probability that the average of the four samples is below 35?

What is the probability that at least three of the four samples are between 35 and 45?

Problem 2

Let the situation be as in Problem 1. In another part of the moor area the soil mainly consists of bleached sand. One wanted to determine whether the proportion of fungi in this type of soil is different from that in the moor soil where the samples described in Problem 1 were taken.

There were taken 20 samples in the new area in the same way as described at the beginning of Problem 1. This gave the measurements $X_1, X_2, \ldots, X_{20}$ given below. The measurements are assumed to be realizations of independent and identically normally distributed variables with unknown expectation μ and the same standard deviation as in the first samples, $\sigma = 4.0$. *(The analysis below should therefore be made with known σ.)*

Measurements:

37.99	46.49	36.92	48.10	42.70	49.91	39.10	46.43	39.52	40.40
47.11	41.32	42.61	47.99	44.19	47.40	49.42	45.52	44.84	41.45

You may use that $\sum_{i=1}^{20} X_i = 879.41$.

a) Find a point estimate for μ based on the measurements, and calculate its standard error, i.e. the standard deviation of the estimator.

Find also a 95% confidence interval for μ.

b) Explain briefly why the botanists' aim involves testing

$$H_0 : \mu = 42.0 \text{ mot } H_1 : \mu \neq 42.0.$$

Perform the testing with the given data and write down the conclusion when the significance level is set to 0.05. Also calculate the corresponding p-value. What is its interpretation?

c) What is meant by type I error and type II error in hypothesis testing?

What is the probability of type I error in the test of the previous point?

What is the probability of type II error with this test if μ in reality is 45.0?

Problem 3

As part of a larger geological investigation one has studied cores of sand aiming to describe the connection between the effective diffusion and the degree of cementation.

The effective diffusion describes how easily gases diffuse through the core of sand, and is represented by a measurement d. The degree of cementation of a core of sand is a number between 0 and 1 and is denoted by c.

One wants to describe d as a function of c and expects a relationship

$$d = \theta(1 - c)^{\beta}, \tag{1}$$

where θ and β are unknown parameters.

To estimate θ and β one does an experiment where one starts by measuring the effective diffusion D for a core of sand with $c = 0$. Then cementation is increased gradually by adding copper sulfate, and one measures the corresponding values of c and D. This gives a total of 10 corresponding values of D and c, $(D_1, c_1), (D_2, c_2), \ldots, (D_{10}, c_{10})$.

a) Show, by taking the natural logarithm of each side of the equality in (1), that we can describe the expected relationship (1) by

$$y = \alpha + \beta x$$

where $y = \ln d$, $\alpha = \ln \theta$ and $x = \ln(1 - c)$.

Based on this one decides to analyze the data with a simple linear regression model

$$Y_i = \alpha + \beta x_i + e_i, \quad i = 1, 2, \ldots, 10 \tag{2}$$

where $Y_i = \ln D_i$ and $x_i = \ln(1 - c_i)$.

Which assumptions are generally made for use of the model (2)?
Discuss briefly to what extent they may be assumed to be satisfied in the given situation.

The following table gives the values of c_i and the associated responses $D_i = d_i$, as well as the transformed x_i and y_i used in the regression analysis.

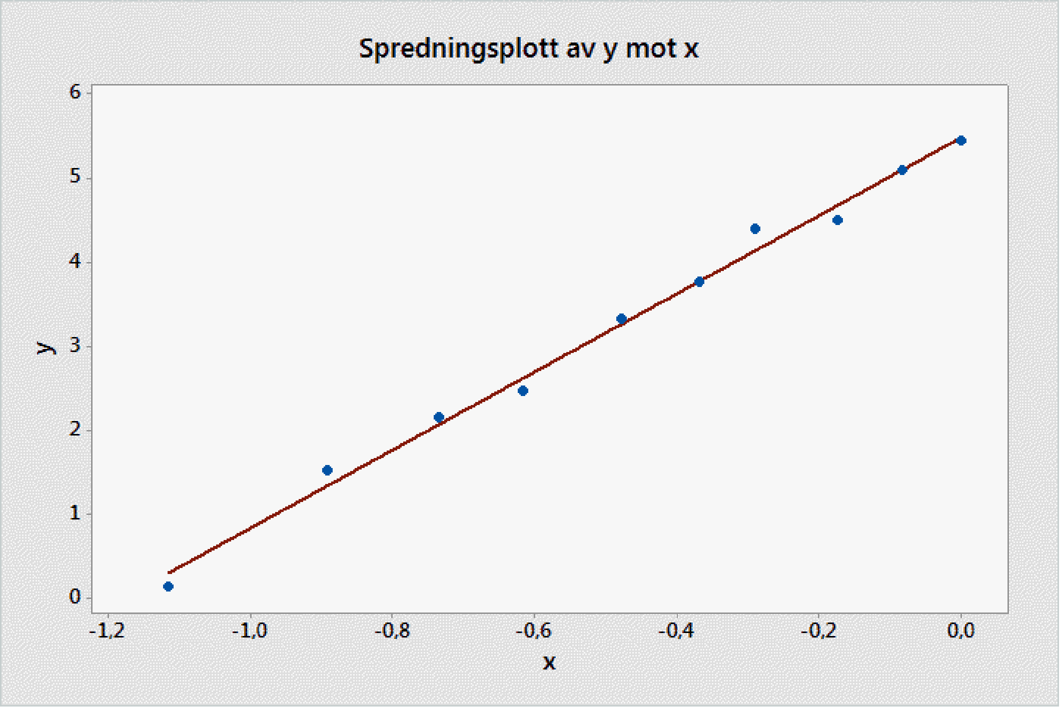
i	c_i	d_i	$x_i = \ln(1 - c_i)$	$y_i = \ln d_i$
1	0,000	234.102	0.000	5.456
2	0.080	165.961	-0.083	5.112
3	0.160	90.990	-0.174	4.511
4	0.250	82.496	-0.288	4.413
5	0.307	43.413	-0.367	3.771
6	0.380	28.156	-0.478	3.338
7	0.460	11.903	-0.616	2.477
8	0.520	8.747	-0.734	2.169
9	0.590	4.589	-0.892	1.524
10	0.672	1.165	-1.115	0.153

It is given that $\sum x_i = -4.747$, $\sum x_i^2 = 3.440$, $\sum y_i = 32.924$, $\sum y_i^2 = 134.273$, $\sum x_i y_i = -10.108$.

You can also use that $SS_E = 0.1898$ and $\sum (x_i - \bar{x})^2 = 1.187$.

In the calculations, you can use formulas that are given in the enclosures and in the tables.

A scatter plot of the calculated (x_i, y_i) with a drawn regression line is for your information given on the next page.



b) Calculate the point estimates $\hat{\alpha}$ and $\hat{\beta}$ for α and β based on the least squares method using the given results.

Also find a point estimate of the parameter θ from equation (1).

Calculate the point estimate S for σ and use it to determine the standard error of the estimate $\hat{\beta}$. Also find a 95% confidence interval for β.

What proportion of the variation in the responses is explained by the regression model? Give a comment in light of the scatter plot.

Problem 4

A biologist wants to investigate the incidence of a rare insect species, in the following called species A, in a particular area. A so-called Malaise trap (a large tent-like structure) is set up and it is assumed that insects of species A are caught in the trap as a Poisson process with intensity (rate) λ pr. hour. Let X be the number of insects of species A captured during one week (= 168 hours). The biologist judges before the investigation that λ is 0.01. This value of λ is used in points a) and b).

a) Explain that X is Poisson distributed with expectation $\mu = 1.68$.

What is the probability that the biologist will find no insects of species A in the trap after one week?

What is the probability that there are at least three insects of species A in the trap?

Assume now that the trap was kept for two weeks. Given that at least three insects of species A were caught during the two weeks, what is the probability that no one was caught in the first week?

b) Let T be the time, measured in hours, until the first insect of species A gets caught in the trap. *(Disregard in this point the fact that the trap is mounted only a limited time.)*

What is the distribution of T?

What is the probability that no insects of species A are captured during the first 48 hours, i.e. $T > 48$?

Given that no insects of species A are captured during the first 48 hours, what is the probability that still no one are captured during the next 48 hours? Comment on the result.

The biologist is no longer sure of his previous estimate of λ and would like to estimate λ. He lets the trap stand for five weeks and finds Y captured insects of species A.

c) Which is the probability distribution of Y?

Write down an unbiased estimator $\hat{\lambda}$ of λ based on Y.

What is the estimate if it is observed that $Y = 15$? What is the standard error (i.e. estimated standard deviation) for this estimate?

Supplement to "Noen resultater fra regresjonsanalysen"
in *Tabeller og formler i statistikk, Tapir akademisk forlag.*

The formulas are based on the sums:

$$\sum_{i=1}^{n} x_i,\ \ \sum_{i=1}^{n} x_i^2,\ \ \sum_{i=1}^{n} y_i,\ \ \sum_{i=1}^{n} y_i^2,\ \ \sum_{i=1}^{n} x_i y_i$$

The least squares method then gives:

$$\begin{aligned}\hat{\beta} &= \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}\\ \hat{\alpha} &= \frac{(\sum_{i=1}^{n} y_i) - \hat{\beta}(\sum_{i=1}^{n} x_i)}{n}\end{aligned}$$

From the text book:

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SS_T} = \underbrace{\sum_{i=1}^{n}(\hat{\alpha} + \hat{\beta}x_i - \bar{y})^2}_{SS_R} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{\alpha} - \hat{\beta}x_i)^2}_{SS_E}$$

Here are:

$$\begin{aligned}SS_T &= (\sum_{i=1}^{n} y_i^2) - \frac{1}{n}(\sum_{i=1}^{n} y_i)^2\\ SS_E &= (\sum_{i=1}^{n} y_i^2) - \hat{\alpha}(\sum_{i=1}^{n} y_i) - \hat{\beta}(\sum_{i=1}^{n} x_i y_i)\end{aligned}$$

Unbiased estimator of σ^2:

$$S^2 = \frac{SS_E}{n-2}$$

Statistical properties of the estimators:

$$E(\hat{\alpha}) = \alpha, \qquad E(\hat{\beta}) = \beta$$

$$Var(\hat{\alpha}) = \frac{\sigma^2 \sum_{i=1}^{n} x_i^2}{n \sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad Var(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$