NTNU
Norwegian University of
Science and Technology

Department of Mathematical Sciences

Examination paper for **TMA4255 Applied Statistics**

Academic contact during examination: Jo Eidsvik

Phone: 901 27 472

Examination date: May 12, 2021

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

Permitted examination support material:

- Home exam.

Other information:
Note that all answers must be justified. All ten subproblems are equally weighted.

Language: English

Number of pages: 6

Number of pages enclosed: 0

Checked by:

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

Date Signature

Problem 1

In the production of caffeine-free coffee brands, it is important to achieve sufficiently low levels of caffeine in the final product. Some use a specified mean level of caffeine in these products of 3.5 mg per 100 g. A company on this market claims to deliver a new brand of caffeine-free coffee that has a lower level. We will conduct hypothesis tests to check if the brand has the specified caffeine level or a lower one.

To get all answers you should use the following measurements of the caffeine level in the new brand: There are 11 independent random samples (tilfeldig utvalg av uavhengige data) in mg / 100 g : 2.8, 3.2, 4.1, 3.2, 3.3, 3.4, 2.6, 3.3, 3.4, 3.7, 4.4.

a)

Formulate hypotheses to test if the mean value of caffeine in the new brand is the same or lower than the specified level.

What are the assumptions going into the T-test for testing this hypothesis?

Use the T-test to conduct the hypothesis test for this data. Use $\alpha = 0.05$ significance level. Find the associated p-value of the test.

Alternative tests here include the non-parametric (ikke-parametrisk) sign-test and Wilcoxon sign-rank test.

b)

Formulate the sign-test for this situation. What are the assumptions going into this hypothesis test? Find the exact p-value of the sign-test.

Formulate the Wilcoxon sign-rank test for this situation. What are the assumptions going into this hypothesis test? Use a normal approximation for the test statistic to find a p-value of this test.

Problem 2

Celebrities Henry and Maggie just moved to a new country. They are speculating whether there are less false news about their lifestyle in the new country compared with the old one. Their data analytics team is asked to study this, and they decide to sample random independent data (tilfeldig utvalg av uavhengige data) in both countries. Results show that there are $X_1 = 3$ false news articles out of $n_1 = 10$

relevant news pieces in the new country. The corresponding numbers for the old country are $X_2 = 6$ of $n_2 = 11$ pieces.

a)

Phrase the question as a hypothesis test.

Use Gaussian approximation results for binomial data to conduct an approximate level $\alpha = 0.05$ hypothesis test.

b)

Define the power (teststyrke) of a hypothesis test. Discuss how the power tends to depend on the data size and the specified parameter in the alternative hypothesis.

Assume the true probability of false news in the new country is 0.2 lower than that in the old country. Compute the approximate power of the hypothesis test from a).

Assume again that the true difference in false news probabilities is 0.2. How many different news piece samples would be required to achieve a power of 0.5?

Problem 3

A student climbing group are at a location with $k = 4$ climbing routes. They are discussing if all routes take the same time, and decide to formulate this question as a hypothesis test and use an analysis of variance (ANOVA) setup. Data is acquired by timing $n = 11$ random persons climbing the routes.

Assume a model $Y_{ij} = \mu_i + \epsilon_{ij}$ for time data on climbing routes $i = 1, \ldots, k$ and with $j = 1, \ldots, n$ samples in each group. Here, μ_i, $i = 1, \ldots, k$ are fixed but unknown values, while ϵ_{ij} are Gaussian zero-mean error terms with fixed but unknown variance σ^2.

The results (seconds) of the data collection are $\bar{y}_{1\cdot} = 128$, $\bar{y}_{2\cdot} = 136$, $\bar{y}_{3\cdot} = 157$, $\bar{y}_{4\cdot} = 141$, with notation $\bar{y}_{i\cdot} = \frac{1}{n}\sum_{j=1}^{n} y_{ij}$. Moreover, we have $\sum_{j=1}^{n}(y_{1j} - \bar{y}_{1\cdot})^2 = 4900$, $\sum_{j=1}^{n}(y_{2j} - \bar{y}_{2\cdot})^2 = 5100$, $\sum_{j=1}^{n}(y_{3j} - \bar{y}_{3\cdot})^2 = 4600$ and $\sum_{j=1}^{n}(y_{4j} - \bar{y}_{4\cdot})^2 = 5000$.

a)

Formulate hypotheses to test if the mean values of climbing times are the same for all routes.

Conduct the hypothesis test using $\alpha = 0.05$ significance level.

b) Some suggest that the first two routes could take about equal time, and the same for the last two, but not all four.

Phrase this statement as a hypothesis test on a linear contrast.

Conduct the hypothesis test at $\alpha = 0.05$ significance level.

Problem 4

Rock physicists study models for the acoustic velocity of rock types as a function of different constituents. In this study the main focus is on porosity in relatively homogeneous rocks and its relation to velocity. The goal is to fit a model to 14 data samples of both porosity (x, in fraction) and velocity (y, in m/s).

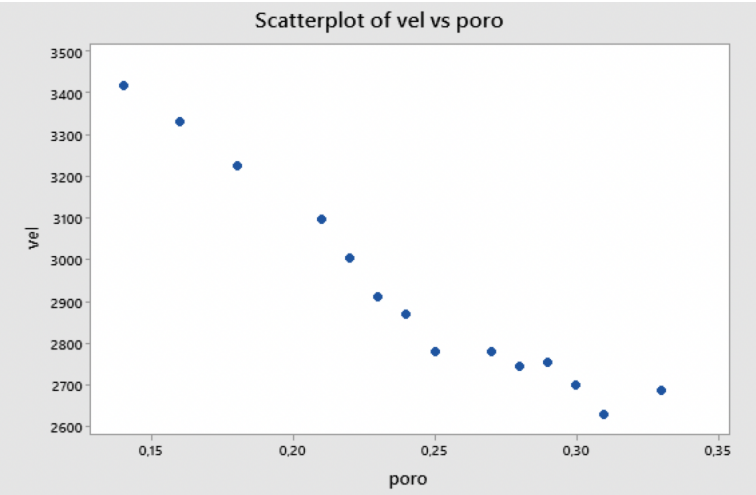

Figure 1: Plot of velocity (second axis) and porosity (first axis) data in rocks.

A suggested quadratic model for the velocity is

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i, \quad i = 1, \ldots, 14, \tag{1}$$

where β_0, β_1 and β_2 are unknown parameters, while $\epsilon_i \sim N(0, \sigma^2)$, $i = 1, \ldots, 14$, are independent (uavhengige) error terms.

a)

Describe and sketch how the method of least squares (minste kvadratsums metode) can be used to fit the parameters of this model. You do not need to finish any formal mathematical derivations here.

Figure 1 shows a cross plot of (x_i, y_i), $i = 1, \ldots, 14$. Do you think the assumptions of the regression model in equation (1) hold?

Figure 2 shows a MINITAB print-out for the model in equation (1).

Regression Equation

vel = 4793 - 11704 poro + 15743 poro2

Coefficients

Term	Coef	SE Coef	95% CI	T-Value
Constant	4793	200	(4352; 5233)	23,95
poro	-11704	1756	?	-6,66
poro2	15743	?	(7581; 23906)	?

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)
42,3578	97,66%	97,24%	40568,7	95,20%

Figure 2: Results of fitting a quadratic model to the velocity and porosity data.

b)

Construct a 95 % confidence interval for β_1.

Find the estimated standard error of $\hat{\beta}_2$.

Find the T-value for $\hat{\beta}_2$.

Some claim that the rocks with low porosity (less than $c = 0.25$) were exposed to high temperatures at some geological time. This could have lead to cementation effects that likely changed the porosity and velocity relation for these rocks,

compared with those of higher porosity. An alternative piecewise linear model is suggested:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 (x_i - c) I(x_i > c) + \epsilon_i, \quad i = 1, \ldots, 14, \tag{2}$$

where the indicator function $I(x_i > c) = 1$ if the porosity x_i is larger than c and 0 otherwise.

Figure 3 shows a print-out from MINITAB with the analysis of this model.

Regression Equation

vel = 4260,0 - 5790 poro + 3720 indporo

Coefficients

Term	Coef	SE Coef	95% CI	T-Value
Constant	4260,0	61,6	(4124,5; 4395,5)	69,20
poro	-5790	291	(-6429; -5150)	-19,93
indporo	3720	608	(2382; 5059)	6,12

Model Summary

S	R-sq	R-sq(adj)	PRESS	R-sq(pred)	
32,7884	98,60%	98,35%	22823,4	97,30%	1

Figure 3: Results of fitting a piecewise linear model to the velocity and porosity data.

c)

How are S and R^2 defined?

What is the definition of PRESS in the print-outs?

Based on the results in Figure 2 and 3, which one of these models would you prefer for the velocity and porosity relation?

Figure 4 shows MINITAB print-out results of a forward selection model fitting using up to three covariates (forklaringsvariable); x_i, x_i^2 and $(x_i - c)I(x_i > c)$.

Forward Selection of Terms

Candidate terms: poro; poro2; indporo

	-----Step 1----		-----Step 2----		-----Step 3----	
	Coef	P	Coef	P	Coef	P
Constant	3971,3		4260,0		3732	
poro	-4299	0,000	-5790	0,000	-151	0,969
indporo			3720	0,000	6670	0,009
poro2					-14500	0,169
S		65,8718		32,7884		31,1406
R-sq		93,84%		98,60%		98,85%
R-sq(adj)		93,33%		98,35%		98,51%

Figure 4: Results of sequential model fitting.

d)

What does it mean to conduct a sequential forward selection in this situation? What happens in Step 1, 2 and 3?

Why do the coefficient for the 'poro' covariate change in Step 1, 2, and 3?

Which model would be preferred based on these results? Would you try yet other types of models based on the results?