NTNU

Kunnskap for en bedre verden

Institutt for matematiske fag

Eksamensoppgave i **ST1201 Statistiske metoder**

Faglig kontakt under eksamen: Bo Lindqvist

Tlf: 975 89 418

Eksamensdato: 19. desember 2020

Eksamenstid (fra–til): 09:00 – 13:00

Hjelpemiddelkode/Tillatte hjelpemidler: Hjelpemiddelkode C:

– Tabeller og formler i statistikk, Tapir forlag,
– K.Rottman: Matematisk formelsamling,
– Ett gult ark (A4 med stempel) med egne håndskrevne formler og notater,
– Bestemt, enkel kalkulator

Annen informasjon:

Alle svar må begrunnes.

Du må ha med nok mellomregninger til at tenkemåten din klart fremgår.

Oppgaven består av 10 delpunkter som har lik vekt ved sensur.

Målform/språk: bokmål

Antall sider: 7

Antall sider vedlegg: 0

Kontrollert av:

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

Dato Sign

Merk! Studenter finner sensur i Studentweb. Har du spørsmål om din sensur må du kontakte instituttet ditt. Eksamenskontoret vil ikke kunne svare på slike spørsmål.

Oppgave 1 Distribution of chocolate pieces in a Christmas bag

A chocolate factory offers various favorite packages where customers find familiar and cherished pieces of chocolate in the same bag. According to the factory is the combination in the Christmas bags 26% Christmas balls, 14% chocolate mice, 24% Christmas figures, 20% party candies and 16% marzipan pigs. Five students have bought several Christmas bags and found the following distribution for 300 chocolate pieces:

Sjokoladebit	Antall
Christmas balls	100
Chocolate mice	47
Christmas figures	62
Partry candies	53
marzipan pigs	38

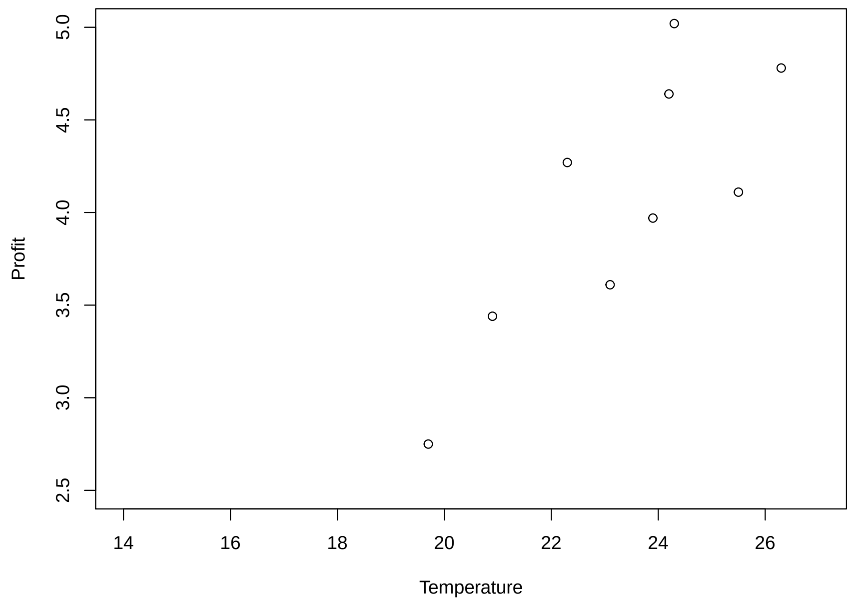
a) Can the students come to the conclusion that the distribution of chocolate pieces is different than the chocolate factory says? Use $\alpha = 0.05$ level of significance

Oppgave 2 Icecream profit over summer

The owners of a icecream shop in Germany investigated their profit (in Euro) over a warm period of 11 days in July 2020. They considered the corresponding days maximum temperature (in °C) as an explanatory variable and modelled the relationship using a linear regression model:

$$Y_i = \mu + \beta x_i + \epsilon_i; \qquad i = 1, \ldots, 11.$$

Unfortunately, the profit was only recorded for 10 of the 11 days. The observed profit (Y) and maximum temperatures (X) are shown in the following graph, while the exact values can be found in the table:



i	1	2	3	4	5	6	7	8	9	10	11
Y_i	4.27	3.44	5.19		2.75	3.61	5.02	4.64	3.97	4.11	4.78
x_i	22.3	20.9	25.6	15.6	19.7	23.1	24.3	24.2	23.9	25.5	26.3

a) Compute the least squares estimator $\hat{\mu}$ and $\hat{\beta}$ by only considering the complete data. You can use that

$$\sum_{i\neq 4} x_i = 235.8, \quad \sum_{i\neq 4} x_i^2 = 5600.44, \quad \sum_{i\neq 4} y_i = 41.78, \quad \sum_{i\neq 4} y_i^2 = 179.8286,$$
$$\sum_{i\neq 4} x_i y_i = 997.223, \quad \sum_{i\neq 4} (x_i - \sum_{j\neq 4} x_j/10)^2 = 40.276.$$

b) Provide an estimate $\hat{Y}_4$ for $E(Y_4 \mid x_4 = 15.6)$ and construct the corresponding 95%-confidence intervall for for $E(Y_4 \mid x_4 = 15.6)$.

c) Assume that the observed value for Y_4 would be equal to the estimated value, that is $Y_4 = \hat{Y}_4$. Do the estimates for the parameters μ and β change, if the regression was now performed with all 11 observations?

(Do not do the computation! An answer which is only a computation will not be considered.)

Oppgave 3 Development times of a queen bee

It is assumed that the developent time of a queen bee depends on the temperature time in the bee hive. Three experiments have been performed where queens are reared at different temperatures in a hive: 1) low temperature of 31 degrees Celsius; 2) medium temperature of 32.5 degrees Celsius; 3) high temperature of 34 degrees Celsius. The development times are summarised in the following table, where n_j denotes the number of bees reared under temperature $j = 1, 2, 3$, $\bar{Y}_{.j}$ the average development time in days and S_j the sample standard deviation for temperature group j.

j	n_j	$\bar{Y}_{.j}$	S_j
low temperature hive	14	14.5	0.45
medium temperature hive	14	15.1	0.48
high temperature hive	16	15.5	0.42

a) Construct the complete ANOVA table and test the null hypothesis that the expected queen development time in days for queen eggs reared in a low, a medium, and a high temperature hive is the same. Let 0.05 be the level of significance.

b) Construct a constrast to test the null hypothesis that the average expected development time under low and high temperature is the same as the expected development time under the medium temperature. Let 0.05 be the level of significance.

Oppgave 4 Physical or digital teaching?

A course in economics was taught to two groups of students. One group was given normal classroom teaching, while the other group was given some form of digital teaching. There were 16 students in each group. In order to divide the students into groups, the students were first arranged in 16 subgroups of size 2 so that the two students within each subgroup were as equal as possible in terms of abilities and prior knowledge. Then a coin was tossed to decide which type of education they should receive. At the end of the course, all students received the same exam and each candidate received a score from 0 to 100. The aim was to find out if there was a basis for saying that these two teaching methods gave different results.

Let X_i for $i = 1, 2, \ldots, 16$ be the score in subgroup number i for those who had classroom teaching and let Y_i for $i = 1, 2, \ldots, 16$ be the score in subgroup i for those who had digital teaching. Let D_i be $X_i - Y_i$. The observed values d_i for D_i were:

12	8	-3	13	7	2	6	21
7	2	11	-3	-14	-2	17	-4

a) Use a paired t-test to examine whether there is a basis for concluding that the two teaching methods give different results.

Which assumptions do you need to make in order to use this test?

Formulate the null hypothesis H_0 and the alternative hypothesis H_1.

Set up the test statistic and perform the test when the significance level is set to 5 %. What is the conclusion of the test?

You can use that $\sum_{i=1}^{16} d_i = 80$, $\sum_{i=1}^{16} d_i^2 = 1604$.

b) Questions were asked regarding the assumptions for the paired t-test, and therefore a request was made to use the sign test for paired data.

This test is based on the test statistic V which is the number of the $n = 16$ pairs of (X_i, Y_i) where $X_i > Y_i$.

Let $p = P(X_i > Y_i)$ and specify the assumptions that lead to V being binomially distributed with $n = 16$ and probability p.

Explain why the null hypothesis for our test of the difference between the two teaching methods can now be written $H_0 : p = 1/2$. What is the alternative hypothesis?

Then derive V from the data and use the tables to find the p-value for the test you want to use based on V.

What is the conclusion of using this test when you select a significance level of 5%?

c) An alternative non-parametric test is the Wilcoxon's signed rank test, also called Wilcoxon's one-sample test.

Explain what this test is about and which assumptions it is based on. Compare with the assumptions made for the sign test in the previous point.

Formulate a null hypothesis and an alternative hypothesis.

Which test statistic is used in this test? Explain how to calculate it and perform the calculation with the given data.

Calculate the p-value for the test using the approximation to the normal distribution.

What is the conclusion of using this test with a significance level of 5%? Compare with the conclusions above.

d) It is well known that if the conditions for using the t-test are met, this will give the strongest test of the three tests considered in this problem.

Here, we are interested in comparing the probabilities for making a Type II error for the sign test and the test based on that the D_is are normally distributed.

First briefly explain what is meant by Type I error and Type II error in hypothesis testing.

First consider the sign test in point b). Let $n = 16$ and show that a two-sided test of $H_0 : p = 1/2$ with significance level 5% will reject H_0 if

$$V \leq 3 \text{ or } V \geq 13$$

where V is defined in point (b).

Show that the probability of a Type I error is 0.022.

The test in point a) is based on that $D_1, \ldots, D_{16}$ are independent and normally distributed, $N(\mu_D, \sigma_D^2)$. At this point we assume that $\sigma_D^2 = 10^2$ is known.

Show that a two-sided test for $H_0 : \mu_D = 0$ with a probability of a Type I error equal to 0.022 rejects H_0 if

$$\bar{D} \leq -5.275 \text{ or } \bar{D} \geq 5.275$$

Assume that in reality $\mu_D = 4$. Find the probabilities of making a Type II error for both tests that are considered in this point. Leave a comment at the end.