NTNU

Norwegian University of
Science and Technology

Department of Mathematical Sciences

Examination paper for **TMA4245 Statistics**

Academic contact during examination: Sara Martino[a], Torstein Fjeldstad[b]

Phone: [a] 994 03 330, [b] 962 09 710

Examination date: 7 June 2019

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

Permitted examination support material: Support material code C:

– Tabeller og formler i statistikk, Akademika,
– A yellow sheet of paper (A5 with a stamp) with personal handwritten formulas and notes,
– A specific basic calculator

Other information:

All your answers should be justified.

The hand-in material should contain calculations leading to your answer.

There are 10 subtasks which have equal weights in the grading.

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 Electric scooter

A hospital that treats patients having bone fractures wishes to study the connection between the use of an electric scooter and bone fractures. Therefore, they record whether each patient having a bone fracture got the injury by using an electric scooter. In 2018, the hospital treated n patients having bone fractures. Further, assume that the probability that a randomly selected patient with a bone fracture had been exposed to an electric scooter accident is p.

Let X be the number of patients having bone fractures that were involved in an accident by use of an electric scooter. Then X has a binomial distribution with n trials and constant success probability p.

a) Assume, only in this subtask, that $n = 17$ and $p = 0.2$.

Find the probability that exactly 4 of the patients having bone fractures were involved in an accident with an electric scooter.

Find the probability that at least 4 of the patients having bone fractures were involved in an accident with an electric scooter.

Given that at least 4 of the patients having bone fractures were involved in an accident with an electric scooter, find the probability that at least 6 of the patients having bone fractures were involved in an accident with an electric scooter.

Further, assume that the hospital in 2018 treated $n = 215$ patients having bone fractures, and that 54 of those were involved in an accident with an electric scooter.

b) Formulate the central limit theorem.

Suggest a reasonable estimator $\widehat{p}$ for p and, based on this, derive an expression for an approximate 95% confidence interval for p.

Use the values given above to find numerical values for the interval.

Problem 2 Charging an electric car

A housing cooperative having 17 housing units, all with an electric car, offers charging of electric cars to its residents. Assume that the annual power consumptions $X_1, X_2, \ldots, X_{17}$ for each of the housing units are independent and normally distributed with unknown expected value (mean) μ kilowatt hours and unknown standard deviation σ kilowatt hours. The housing cooperative has earlier assumed that the expected power consumption of a randomly selected housing unit is 3000 kilowatt hours. It is suspected that the power consumption is in reality higher, and a consultancy firm was hired to investigate this.

The consultancy firm has collected the power consumptions $x_1, x_2, \ldots, x_{17}$ from 17 housing units that on average consumed $\bar{x} = \frac{1}{17}\sum_{i=1}^{17} x_i = 3200$ kilowatt hours a year with a sample standard deviation of $s = \sqrt{\frac{1}{16}\sum_{i=1}^{17}(x_i - \bar{x})^2} = 300$ kilowatt hours for charging an electric car.

a) Formulate the above question as a hypothesis test.

Perform the hypothesis test that you specified with a significance level of $\alpha = 0.05$. In particular, state the probability distribution of the test statistic that you use.

Can the housing cooperative, based on the result of the hypothesis test, conclude that the power consumption is higher than 3000 kilowatt hours?

Problem 3 Runoff

The annual runoff Y (millimetres per year) is a measure of how large portion of the annual precipitation (millimetres per year) in a specific area, often called a drainage basin, that runs out in connected waterways. The difference between annual precipitation and annual runoff is assumed to have evaporated from the drainage basin.

Assume the following linear relationship between annual runoff Y and annual precipitation x within a drainage basin,

$$Y = \beta_0 + \beta_1 x + \varepsilon, \tag{1}$$

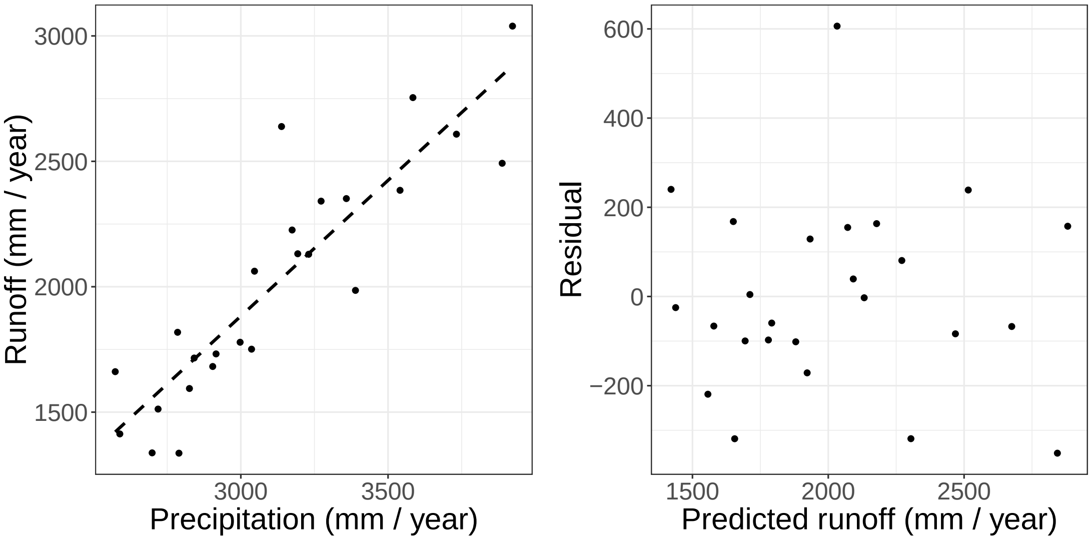
where β_0 and β_1 are unknown constants and ε is normally distributed with expected value (mean) 0 and unknown variance σ^2.

Hydrologists have collected independent observations from the drainage basin of interest over a period of 25 years, that is, a random sample $(x_1, Y_1), (x_2, Y_2), \ldots,$ (x_{25}, Y_{25}) from the model defined in (1). It is given that the following are unbiased estimators of β_1, β_0 and σ^2, respectively:

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^{25}(x_i - \bar{x})Y_i}{\sum_{i=1}^{25}(x_i - \bar{x})^2} \\
\widehat{\beta}_0 &= \bar{Y} - \widehat{\beta}_1 \bar{x} \\
S^2 &= \frac{1}{23}\sum_{i=1}^{25}\left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2 .
\end{aligned} \tag{2}$$

In Figure 1a, the observed values $(x_1, y_1), (x_2, y_2), \ldots, (x_{25}, y_{25})$ are shown together with the fitted regression line $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta} x_i$.

I Figure 1b, the fitted runoff $\widehat{y}_i$ is plotted against the residuals $e_i = y_i - \widehat{y}_i$.



(a) Precipitation x_i against runoff y_i.

(b) Fitted runoff $\widehat{y}_i$ against residuals e_i.

Figure 1: Observations (x_i, y_i) and fitted regression line $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta} x_i$, and fitted runoff $\widehat{y}_i$ against residuals e_i.

a) Briefly explain how the least squares method can be used to find estimators of β_0 and β_1, and illustrate by drawing a figure. You are not required to derive the expressions of the estimators.

Consider the fitted model shown in Figure 1.

Discuss briefly whether it is reasonable to use a linear regression model. In particular, state (briefly) what assumptions must be satisfied for a linear regression model to be used.

Assume that we now are interested in predicting future runoff for a new year Y_0 given annual precipitation $x = x_0$, from the model defined in (1), where (x_0, Y_0) is independent of $(x_1, Y_1), (x_2, Y_2), \dots, (x_{25}, Y_{25})$. It is given that $\widehat{Y}_0 = \widehat{\beta}_0 + \widehat{\beta}_1 x_0$ is a reasonable point estimator for the expected runoff $\mu_{Y|x_0} = \beta_0 + \beta_1 x_0$ when annual precipitation is x_0. It is given that $\widehat{\beta}_0 = -1364$ and $\widehat{\beta}_1 = 1.08$.

You can further in the problem use (without proof) that $\frac{(n-2)S^2}{\sigma^2} \sim \chi^2_{n-2}$, that is, chi-squared distributed with $n-2$ degrees of freedom. You can also use that $\bar{Y}$ and $\widehat{\beta}_1$, and $\widehat{Y}_0$ and $\frac{(n-2)S^2}{\sigma^2}$ are independent random variables.

b) What is the estimated expected runoff for a year in which $x = 2000$ millimetres of precipitation is observed?

Show that

$$\mathrm{E}\left(\widehat{Y}_0 - Y_0\right) = 0$$

and

$$\mathrm{Var}\left(\widehat{Y}_0 - Y_0\right) = \sigma^2\left(1 + \frac{1}{25} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{25}(x_i - \bar{x})^2}\right).$$

Use this to find an expression for a 95% prediction interval for a new observation Y_0 given $x = x_0$.

Problem 4 Weight gold bar

Thomas has inherited a gold bar after his grandparents that he wants to sell. The gold bar is of pure gold and weighs μ grams. Before Thomas sells the gold bar, he wants to make sure that he sells it at a correct price and therefore decides to measure the weight of the gold bar in two ways.

First Thomas uses his own kitchen scales to weigh the gold bar. Assume that the weight measured at his kitchen scales X is a normally distributed random variable with expected value (mean) μ and standard deviation 1 gram. Thereafter he goes to a retailer to have the weight measured professionally. Let Y be the weight measured at the retailer, and assume that Y is normally distributed with expected value μ and standard deviation 0.5 grams. Assume that $\mathrm{Cov}(X, Y) = -0.2$.

To estimate the true weight μ grams of the gold bar, Thomas will compare two different estimators,

$$\widehat{\mu} = Y \qquad \text{and} \qquad \tilde{\mu} = \frac{1}{2}(X + Y).$$

a) Describe briefly what characterizes a good estimator.

Which of the two estimators $\widehat{\mu}$ and $\tilde{\mu}$ should Thomas choose? Justify your answer.

Problem 5 Moment-generating function

Assume that $X_1, X_2, \ldots, X_n$ is a random sample from a gamma distribution with parameters α and β, that is, having probability density function

$$f_X(x;\alpha,\beta) = \begin{cases} \frac{1}{\beta^\alpha \Gamma(\alpha)} x^{\alpha-1} e^{-x/\beta} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

where $\alpha > 0$ is a known parameter and $\beta > 0$ is an unknown parameter. You can use (without proof) that the moment-generating function of a gamma-distributed random variable X is

$$M_X(t) = \frac{1}{(1-\beta t)^\alpha} \qquad \text{for } t < \frac{1}{\beta}.$$

a) Show, by using moment-generating functions, that $Y = X_1 + X_2 + \cdots + X_n$ has a gamma distribution with parameters $n\alpha$ and β, that is, Y has probability density function

$$f_Y(y;n\alpha,\beta) = \begin{cases} \frac{1}{\beta^{n\alpha} \Gamma(n\alpha)} y^{n\alpha-1} e^{-y/\beta} & y > 0 \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Assume that we have an observation y from (3). Derive an expression for the maximum likelihood estimator for β based on this. Remember that α (and n) are known.

Problem 6 Probability

Let X and Y be two independent normally distributed random variables with known expected values (means) $\mu_X = 1$ and $\mu_Y = 0$, respectively, and known standard deviations $\sigma_X = \sigma_Y = 1$.

a) Find $P(2X > 3)$.

Find $P(2X > 3 \mid Y > 0)$.

Explain briefly why $X - Y$ is also normally distributed, and use this fact to find $P(-1 \leq X - Y \leq 1)$.

Assume that we have two independent random samples $X_1, X_2, \ldots, X_{10}$ and $Y_1, Y_2, \ldots, Y_{15}$ from the two distributions specified above.

b) Derive the probability that at most 5 of $X_1, X_2, \ldots, X_{10}$ are less than or equal to x, in terms of the cumulative distribution function of X.

Derive an expression for $P(\max\{X_1, X_2, \ldots, X_{10}, Y_1, Y_2, \ldots, Y_{15}\} \leq z)$ in terms of the cumulative distribution functions of X and Y.

Problem 7 Hypotesis test uniform distribution

Let X_1 and X_2 be two independent uniformly distributed random variables on the interval $[\theta, \theta + 1]$, that is, they have probability density function

$$f(x;\theta) = \begin{cases} 1 & x \in [\theta, \theta + 1] \\ 0 & \text{otherwise} \end{cases}$$

where θ is an unknown constant.

The following hypothesis shall be investigated:

$$\mathrm{H}_0 : \theta = 0 \qquad \text{against} \qquad \mathrm{H}_1 : \theta > 0.$$

To rejection rules are suggested,

$$\begin{aligned} \text{Rejection rule 1 :} \qquad & \text{Reject } \mathrm{H}_0 \text{ if } X_1 > 0.95 \\ \text{Rejection rule 2 :} \qquad & \text{Reject } \mathrm{H}_0 \text{ if } X_1 + X_2 > k, \end{aligned}$$

where k is an unknown critical value that is to be determined.

a) Find the probability of type I error for rejection rule 1.

For rejection rule 1, find an expression for the power of the test as a function of $\theta = \tau$, that is, $P(\text{reject } \mathrm{H}_0 \text{ when } \theta = \tau)$ for $\tau > 0$, and sketch the graph of this function.

We now require that rejection rule 1 and 2 should have identical probability of a type I error. Find k.

Hint: It can be useful to sketch the probability density function of X_1 and the joint probability density function of (X_1, X_2).