TMA4245 Statistics
Exam May 21 2013

Corrected June 10 2013

Norges teknisk-naturvitenskapelige universitet
Institutt for matematiske fag

Oppgave 1

Let X and Y be two independent normally distributed random variables. Assume that X has expected value equal to 0 and standard deviation equal to 1, while Y has expected value equal to 1 and standard deviation equal to 2.

Make a sketch of the probability densities for X and Y in one common plot. Determine the probabilities $\mathrm{P}(X \leq 1.2)$, $\mathrm{P}(Y > 2)$ and $\mathrm{P}(X + Y \leq 2)$.

Oppgave 2

Let A and B be two events in a sample space S, where $\mathrm{P}(A) = 0.2$, $\mathrm{P}(B) = 0.5$ and $\mathrm{P}(A \cup B) = 0.6$.

Are the events A and B disjoint? Are the events A and B independent?

Oppgave 3

The lifetime T (measured in number of days (24 h)) of a new type of valve that is being considered for use on oil platforms in the North Sea, is to be checked. It is well known that the lifetime is influenced by for example the temperature and pressure where the valve is used, and also by the chemical composition of the oil that passes through the valves. Assume that the effect of these factors are measured as a stress factor z. It is also assumed that the cumulative distribution function for the lifetime T is known by experience to be

$$F(t) = \mathrm{P}(T \leq t) = \begin{cases} 1 - e^{-\frac{zt^2}{\theta}} & \text{for } t > 0, \\ 0 & \text{otherwise,} \end{cases}$$

where θ is an unknown parameter. The parameter θ is a characteristic feature for a particular type of valve, while z describes the environment where the valve is used.

a) When $z = 2.0$ and $\theta = 2 \cdot 10^6$, determine the probabilities

$$\mathrm{P}(T > 1000) \quad \text{and} \quad \mathrm{P}(T > 2000 | T > 1000).$$

Let T_1, T_2 and T_3 be the lifetimes of three valves that all operate under conditions where $z = 2.0$, and assume that T_1, T_2 and T_3 are independent random variables. When $\theta = 2 \cdot 10^6$, find the probability that at least two of the three lifetimes are greater than 1000 days.

b) Show that the probability density of T is given by

$$f(t) = \begin{cases} \frac{2zt}{\theta} e^{-\frac{zt^2}{\theta}} & \text{for } t > 0, \\ 0 & \text{otherwise.} \end{cases}$$

Make a sketch of $f(t)$ for $t \in [0, 3000]$ when $z = 2.0$ and $\theta = 2 \cdot 10^6$, and mark on the sketch the area that equals the probability $\mathrm{P}(T > 1000)$.

Let a random variable V be defined by

$$V = \frac{2zT^2}{\theta}.$$

c) Apply the transformation formula to show that V is χ^2 distributed with 2 degrees of freedom.

Use this result to show that $\mathrm{E}\left(T^2\right) = \frac{\theta}{z}$ and $\mathrm{Var}\left(T^2\right) = \left(\frac{\theta}{z}\right)^2$.

To check the quality of this type of valve, a random sample of $n = 10$ valves has been tested. Let $z_1, z_2, \ldots, z_n$ denote the stress factors under which these valves operate, and let $T_1, T_2, \ldots, T_n$ denote the corresponding lifetimes. It is assumed that $T_1, T_2, \ldots, T_n$ are independent random variables. The observed lifetimes are given in the following table:

Valve i	1	2	3	4	5	6	7	8	9	10
z_i	1.0	3.4	1.9	2.4	1.2	4.0	3.2	2.2	1.4	3.2
t_i	1297.2	834.2	1265.8	331.7	1937.8	727.6	869.6	746.7	1965.3	280.9

It has been calculated that $\sum_{i=1}^{n} z_i t_i^2 = 23\ 287\ 125$.

d) Derive the maximum likelihood estimator for θ, and show that it can be written on the form

$$\widehat{\theta} = \frac{1}{n} \sum_{i=1}^{n} z_i T_i^2.$$

Is $\widehat{\theta}$ unbiased? Find $\mathrm{Var}(\widehat{\theta})$.

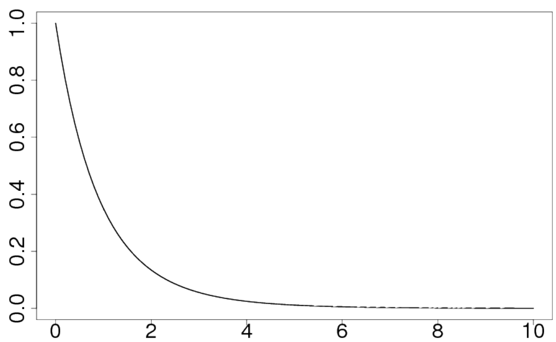
e) Justify that

$$U = \sum_{i=1}^{n} \frac{2 z_i T_i^2}{\theta} \sim \chi^2_{2n}.$$

Use this to derive a $(1 - \alpha) \cdot 100\%$ confidence interval for θ. What is the confidence interval when the data are as given above and $\alpha = 0.05$?

On the basis of the observed data, one would also like to find a prediction interval for the lifetime, T_0, of a new valve which is to operate under conditions with a stress factor equal to z_0. To do this, one may start with the random variable

$$Y = n \cdot \frac{\frac{2T_0^2 z_0}{\theta}}{\sum_{i=1}^{n} \frac{2 z_i T_i^2}{\theta}},$$



$P(Y > y_\alpha) = \alpha$

α	y_α
0.975	0.025
0.950	0.051
0.900	0.106
0.100	2.59
0.050	3.49
0.025	4.46

Figur 1: The probability density of Y and some quantiles of this distribution.

where the numerator and the denominator in the fraction are independent random variables. The numerator and the denominator are both χ^2 distributed, and they have 2 and $2n$ degrees of freedom, respectively. When $n = 10$, it can be shown that the probability density of Y becomes as shown in Figure 1. Note that some quantiles of this distribution are given to the right in the same figure.

f) Derive a 90% prediction interval for T_0. What is the prediction interval when the data are as given above and $z_0 = 3.0$?

Oppgave 4

As part of a physical fitness test of male soldiers, a random sample of $n = 42$ male soldiers were selected to go through a series of physical tests. In this problem we are going to study the results of two of these tests, namely the number of push-ups the soldiers were able to do during 2 minutes and the time (measured in seconds) they needed to run 3 kilometers. A simple linear regression model is assumed for these data, where the number of push-ups is the regressor (independent variable) and the running time the response variable (dependent variable). This gives the model

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,$$

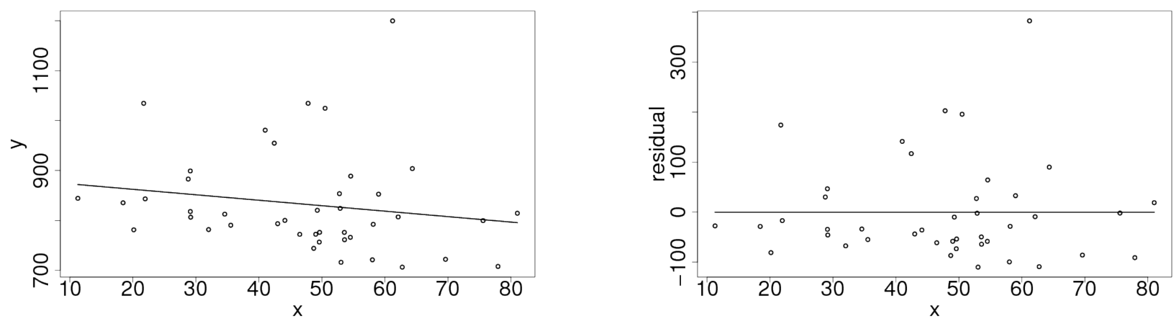
where Y_i and x_i denotes the running time and number of push-ups, respectively, and $\varepsilon_1, \ldots, \varepsilon_n$ are assumed to be independent and normally distributed with mean value 0 and variance σ^2. The values of all three parameters β_0, β_1 and σ^2 are unknown. To estimate these, the following standard estimators are applied.

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})Y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \widehat{\beta}_0 = \bar{Y} - \widehat{\beta}_1 \bar{x}$$

and

$$S^2 = \frac{1}{n-2} \sum_{i=1}^n \left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i \right)^2.$$

Figure 2(a) shows the $n = 42$ observations, together with the estimated regression line. From the observations it is obtained that $\sum_{i=1}^{n}(x_i - \bar{x})y_i = -12\ 163.6$, $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 11\ 113.9$ and $\sum_{i=1}^{n}\left(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2 = 414\ 563.1$.



(a) (b)

Figur 2: (a) Observed data and estimated regression line, with the number of push-ups on the x-axis and the running time on the y-axis. (b) The corresponding residual plot with the number of push-ups on the x-axis and estimated residuals $\widehat{\varepsilon} = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i$ on the y-axis.

For the remaining part of this problem you may (without proof) use that

$$\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\frac{S^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}}$$

is t-distributed with $n - 2$ degrees of freedom.

It is required to investigate whether the observed data justifies the claim that the expected running time decreases with the number of push-ups.

a) Formulate this as a hypothesis test and construct a test for this purpose with significance level 5%.

What is the conclusion of the hypothesis test when the data are given as above.

b) The model we have been using in this problem assumes that the residuals, ε_i, are normally distributed with mean value zero and with the same standard deviation σ. Based on the plot in Figure 2(a), and the plot of the estimated residuals $\widehat{\varepsilon}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i$ in Figure 2(b), would you agree that this assumption is satisfied for the observed data? Justify your answer.

Fasit

1. 0.8849,0.3085,0.6736

3. **a)** 0.368,0.050,0.306 **d)** $\mathrm{E}[\widehat{\theta}] = \theta$, $\mathrm{Var}[\widehat{\theta}] = \theta^2/n$ **e)** $[1\ 363\ 016, 4\ 856\ 037]$ **f)** $[198.97, 1\ 645.92]$

4. **a**) Do not reject H_0