

Family name **First name**

Ph. D. in Materials Engineering
Methods of statistical and numerical analysis
Final test on statistical methods

Part I - Discussion (1 hour)

Topic (1)

Basic definitions concerning random variables and probability distributions. Remarkable examples of discrete and continuous random variables.

Topic (2)

Linear regression of experimental data by the chi-square fitting: the basic statistical model of linear regression according to the chi-square fitting method.

Exercise 1

The below table of data has been obtained by the repeated measurement of a temperature T (K):

n	T_n	n	T_n
1	206.5	8	192.1
2	199.5	9	180.7
3	210.5	10	193.3
4	178.9	11	190.1
5	186.5	12	197.3
6	195.0	13	196.9
7	189.0	14	195.3

The statistical population is assumed to be normal. Compute the confidence interval of the mean at the confidence level of 90%. Reckon also the confidence interval of the standard deviation at the confidence level 95%.

Exercise 2

A normal sample of mass data (in Kg) is summarized in the table below:

24.5 , 24.7 , 24.8 , 24.6 , 25.9 ,
24.4 , 23.8 , 24.4 , 24.5 , 24.3

Test, by using Chauvenet criterion, whether the sample may contain an outlier.

Exercise 3

Repeated measurements of two quantities x and y have led to the following table of data (various y values corresponding to the same value of x mean that the measurement of y has been carried out many times for the same x):

x_i	y_{i1}	y_{i2}	y_{i3}	y_{i4}	y_{i5}
4.00	3.12	×	3.24	3.20	3.08
4.24	3.04	2.96	3.08	3.02	3.10
4.70	2.94	2.76	2.90	2.76	2.84
5.16	2.82	2.78	2.70	2.74	×
7.10	×	×	2.14	2.24	2.22

The random error on the x values is negligible, whereas all the y data are assumed to be independent normal variables with the same standard deviation. Model the data by a best-fit straight line of the form

$$y = m + q(x - \bar{x}),$$

where \bar{x} stands for the sample mean of the independent variables x_i 's and the model parameters m and q are calculated by the least squares method. In particular, determine:

- (i) the 90%-confidence interval of the intercept m ;
- (ii) the 90%-confidence interval of the slope q ;
- (iii) the 90%-confidence interval for the prediction of y at a given value $x = 5.10$ of the independent variable

Solution of Exercise 1

Number of data: $n = 14$

$$\text{Sample mean: } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 193.685714$$

$$\text{Sample variance: } s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = 75.7551648$$

Estimated standard deviation: $s = \sqrt{s^2} = 8.70374429$

□ The CI of the mean is

$$\bar{x} - t_{[1-\frac{\alpha}{2}](n-1)} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{[1-\frac{\alpha}{2}](n-1)} \frac{s}{\sqrt{n}}$$

In the present case the confidence level for the confidence interval of the mean is requested to be $1 - \alpha = 90\%$, so that $\alpha = 0.10$; the number of datapoints is $n = 14$. Therefore the confidence interval has the limits

$$\begin{aligned} \bar{x} - t_{[0.95](13)} \frac{s}{\sqrt{14}} &= 193.685714 - 1.771 \cdot \frac{8.70374429}{\sqrt{14}} = \\ &= 189.56606 \end{aligned}$$

$$\begin{aligned} \bar{x} + t_{[0.95](13)} \frac{s}{\sqrt{14}} &= 193.685714 + 1.771 \cdot \frac{8.70374429}{\sqrt{14}} = \\ &= 197.80537 \end{aligned}$$

and writes

$$189.56606 \leq \mu \leq 197.80537$$

or, equivalently,

$$[193.68571 \pm 4.119653]$$

□ The CI of the variance takes the form

$$\frac{1}{\chi^2_{[1-\frac{\alpha}{2}](n-1)}}(n-1)s^2 \leq \sigma^2 \leq \frac{1}{\chi^2_{[\frac{\alpha}{2}](n-1)}}(n-1)s^2$$

with $\alpha = 0.05$ and $n = 14$. Thus

$$\frac{1}{\chi^2_{[0.975](13)}} 13 s^2 = \frac{1}{24.736} 13 \cdot 75.7551648 = 39.8131122$$

$$\frac{1}{\chi^2_{[0.025](13)}} 13 s^2 = \frac{1}{5.009} 13 \cdot 75.7551648 = 196.609531$$

and the CI of the variance becomes

$$39.8131122 \leq \sigma^2 \leq 196.609531$$

Taking the square root we obtain the CI of the standard deviation

$$6.30976324 \leq \sigma \leq 14.0217521$$

which can also be put into the equivalent form

$$[10.166 \pm 3.856]$$

Solution of Exercise 2

The sample mean and standard deviation are given by

$$\bar{x} = \frac{1}{10} \sum_{i=1}^{10} x_i = 24.59 \quad s = \sqrt{\frac{1}{9} \sum_{i=1}^{10} (x_i - \bar{x})^2} = 0.53427$$

The farthest value from the mean \bar{x} is 25.9, the main candidate to be an outlier of the sample. The distance of the suspect value from the mean, in units of s , holds

$$\frac{x_{\text{sus}} - \bar{x}}{s} = \frac{25.9 - 24.59}{0.53427} = 2.45$$

The probability that a measurement falls at a distance larger than 2.45 standard deviations from the mean can be calculated from the Table of the cumulative distribution function of the standard normal distribution:

$$\begin{aligned} P(|x_{\text{sus}} - \bar{x}| \geq 2.45s) &= 1 - P(|x_{\text{sus}} - \bar{x}| < 2.45s) = \\ &= 1 - 2 \cdot P(\bar{x} \leq x_{\text{sus}} < \bar{x} + 2.45s) = \\ &= 1 - 2 \cdot 0.49286 = 0.01428 \end{aligned}$$

Out of 10 measurements we typically expect $10 \cdot 0.01428 = 0.1428$ “bad” results, at a distance larger than $2.45s$ from the mean. **Since $0.1428 < 1/2$, Chauvenet criterion suggests that x_{sus} should be rejected as an outlier.**

Solution of Exercise 3

Since the standard deviations are assumed to be the same, the chi-square fitting reduces to the usual least squares fitting and the best-fit estimates of the parameters m and q can be written in the form:

$$m = \frac{1}{n} \sum_{i=1}^n y_i = 2.841905 \quad q = \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = -0.299860$$

with $n = 21$. The sum of squares around regression holds then:

$$\text{SSAR} = \sum_{i=1}^n [m + q(x_i - \bar{x}) - y_i]^2 = 0.09716684$$

while $\alpha = 0.10$. At a confidence level $1 - \alpha \in (0, 1)$ the confidence interval of the intercept μ and that of the slope κ are given by:

$$\mu = m \pm t_{[1-\frac{\alpha}{2}](n-2)} \sqrt{\frac{1}{n} \frac{\text{SSAR}}{n-2}}$$

$$\kappa = q \pm t_{[1-\frac{\alpha}{2}](n-2)} \sqrt{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^{-1} \frac{\text{SSAR}}{n-2}}$$

In the present case the confidence intervals become therefore:

$$\mu = m \pm t_{[0.95](19)} \sqrt{\frac{1}{21} \frac{\text{SSAR}}{19}}$$

$$\kappa = q \pm t_{[0.95](19)} \sqrt{\left[\sum_{i=1}^{21} (x_i - \bar{x})^2 \right]^{-1} \frac{\text{SSAR}}{19}}$$

with:

$$m = 2.841905$$

$$q = -0.299860$$

$$\text{SSAR} = 0.09716684$$

$$\sum_{i=1}^{21} (x_i - \bar{x})^2 = 20.40518096$$

$$t_{[0.95](19)} = 1.729$$

As a consequence:

(i) the 90%-confidence interval for the intercept μ is

$$2.8419 \pm 0.0269 = [2.8150, 2.8688]$$

(ii) the 90%-confidence interval for the slope κ holds

$$-0.2999 \pm 0.0274 = [-0.3273, -0.2725]$$

(iii) For a homoscedastic model, the $(1 - \alpha)$ -confidence interval for the prediction of $y = y_0$ at a given abscissa $x = x_0$ is expressed by the formula:

$$\mathbb{E}(y_0) = m + q(x_0 - \bar{x}) \pm t_{[1-\frac{\alpha}{2}](n-2)} \sqrt{V} \sqrt{\frac{\text{SSAR}}{n-2}}$$

where:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

$$\text{SSAR} = \sum_{i=1}^n [-y_i + m + q(x_i - \bar{x})]^2$$

$$V = 1 + \frac{1}{n} + \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} (x_0 - \bar{x})^2.$$

In the present case we have $\bar{x} = 4.887619$, so that:

$$y_0 = 2.841905 - 0.299860(x_0 - 4.887619) \pm t_{[0.95]}(19) \cdot \sqrt{1 + \frac{1}{21} + \frac{1}{20.40518096}(x_0 - 4.887619)^2} \sqrt{\frac{0.09716684}{19}}$$

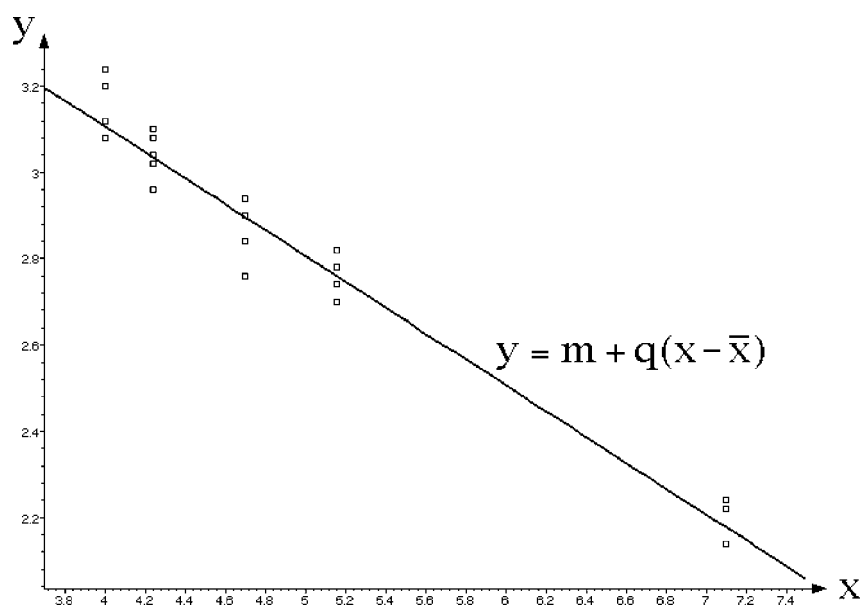
and since $t_{[0.95]}(19) = 1.729$ the confidence interval for the prediction of y at $x = x_0$ reduces to:

$$y_0 = 2.841905 - 0.299860(x_0 - 4.887619) \pm 0.12364519 \sqrt{1.04761905 + 0.049007162(x_0 - 4.887619)^2}$$

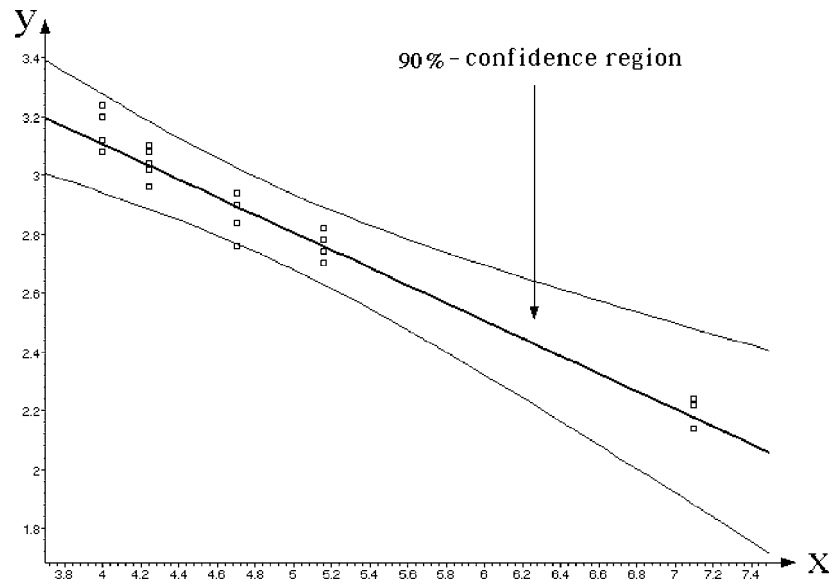
The confidence interval of y at $x = x_0 = 5.10$ is therefore:

$$y_0 = [2.77822 \pm 0.12669] = [2.65153, 2.90491]$$

In the following picture the regression line is superimposed to the experimental data (dots). The number of displayed dots is only 20, because two values of y at $x = 4.70$ are equal (the common value is 2.76).



The confidence region for predictions (at a confidence level of 90%) is evidenced in the figure below



where the factor V is magnified for clarity.