

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)

Definition of the confidence interval for the mean of a normal population by using a small sample.

Topic (2)

Describe the general ideas of hypothesis testing. Discuss also, in particular, a specific type of test among the following ones:

- χ^2 -test for adapting a probability distribution to a sample;
- one-sided and two-sided t -test for the mean of a normal population;
- one-sided and two-sided F -test for the variance of a normal population;
- unpaired t -test for comparing the mean values of two normal populations;
- nonparametric sign test for the median of a random variable.

Part II - Open book exercises (1 hour)**Exercise 1**

The repeated measurement of a mass m led to the following table of data (Kg):

i	m_i	i	m_i
1	2.950	8	2.969
2	3.105	9	2.933
3	2.789	10	2.901
4	2.973	11	2.865
5	2.807	12	2.995
6	2.890	13	2.953
7	3.065	14	2.921

Assuming that the population is normal, compute the confidence interval of the mean and that of the variance, both at the same confidence level 95%.

Exercise 2

The temperature data below (in K)

347 , 345 , 346 , 348 , 338 ,
344 , 359 , 345 , 344 , 343

can be assumed to be extracted from a normal population. By applying Chauvenet criterion, check if the sample may contain an outlier.

Exercise 3

Let us consider the following table of (x, y) data:

x_i	y_{i1}	y_{i2}	y_{i3}	y_{i4}	y_{i5}
2.00	1.56	1.62	×	1.54	1.60
2.12	1.52	1.48	1.54	1.51	1.55
2.35	1.47	1.38	1.45	1.38	1.42
2.58	1.35	1.39	×	1.37	1.41
3.05	×	1.12	1.07	×	1.11

All the y data are assumed to be independent normal variables with the same standard deviation, while the random error on the x values is negligible. 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 must be calculated with the least squares method. In particular, determine:

- (i) the 95%-confidence interval of the intercept m ;
- (ii) the 95%-confidence interval of the slope q ;
- (iii) the 95%-confidence interval for the prediction of y at a given value $x = 2.55$ of the independent variable

Answer to Exercise 1

Number of data: $n = 14$

$$\text{Sample mean: } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 2.936857$$

$$\text{Sample variance: } s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = 75.755 \cdot 10^{-4}$$

$$\text{Estimated standard deviation: } s = \sqrt{s^2} = 0.087037$$

□ The CI of the mean is

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

and for $\alpha = 0.05$, $n = 14$ has therefore the limits

$$\bar{x} - t_{[0.975](13)} \frac{s}{\sqrt{14}} = 2.936857 - 2.160 \cdot \frac{0.087037}{\sqrt{14}} = 2.8866$$

$$\bar{x} + t_{[0.975](13)} \frac{s}{\sqrt{14}} = 2.936857 + 2.160 \cdot \frac{0.087037}{\sqrt{14}} = 2.9871$$

so that the CI writes

$$2.8866 \leq \mu \leq 2.9871$$

or, equivalently,

$$[2.9369 \pm 0.0502]$$

□ 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

$$\frac{1}{\chi^2_{[0.975](13)}} 13 s^2 = \frac{1}{24.736} 13 \cdot 75.755 \cdot 10^{-4} = 39.813 \cdot 10^{-4}$$

$$\frac{1}{\chi^2_{[0.025](13)}} 13 s^2 = \frac{1}{5.009} 13 \cdot 75.755 \cdot 10^{-4} = 196.61 \cdot 10^{-4}$$

and the CI becomes

$$39.813 \cdot 10^{-4} \leq \sigma^2 \leq 196.61 \cdot 10^{-4}$$

Answer to Exercise 2

The sample mean and standard deviation are given by

$$\bar{x} = \frac{1}{10} \sum_{i=1}^{10} x_i = 345.9 \quad s = \sqrt{\frac{1}{9} \sum_{i=1}^{10} (x_i - \bar{x})^2} = 5.3427$$

The suspect value is 359, the farthest from the mean \bar{x} . The distance of the suspect value from the mean, in units of s , holds

$$\frac{x_{\text{sus}} - \bar{x}}{s} = \frac{359 - 345.9}{5.3427} = 2.45$$

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.**

Answer to 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:

$$m = \frac{1}{n} \sum_{i=1}^n y_i = 1.420952 \quad q = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = -0.432164$$

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

$$\text{SSAR} = \sum_{i=1}^n [m + q(x_i - \bar{x}) - y_i]^2 = 0.029966564$$

while $\alpha = 0.05$. 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:

$$\mu = m \pm t_{[0.975](19)} \sqrt{\frac{1}{21} \frac{\text{SSAR}}{19}}$$

$$\kappa = q \pm t_{[0.975](19)} \sqrt{\left[\sum_{i=1}^{21} (x_i - \bar{x})^2 \right]^{-1} \frac{\text{SSAR}}{19}}$$

with:

$$m = 1.420952$$

$$q = -0.432164$$

$$\text{SSAR} = 0.029966564$$

$$\sum_{i=1}^{21} (x_i - \bar{x})^2 = 2.425580951$$

$$t_{[0.975](19)} = 2.093$$

As a conclusion:

(i) the 95%-confidence interval for the intercept μ is

$$1.4209 \pm 0.0181 = [1.4028, 1.4391]$$

(ii) the 95%-confidence interval for the slope κ holds

$$-0.4322 \pm 0.0534 = [-0.4855, -0.3788]$$

(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 $\bar{x} = 2.372381$, so that:

$$y_0 = 1.420952 - 0.432164(x_0 - 2.372381) \pm t_{[0.975](19)} \cdot \sqrt{1 + \frac{1}{21} + \frac{1}{2.425580951} (x_0 - 2.372381)^2} \sqrt{\frac{0.029966564}{19}}$$

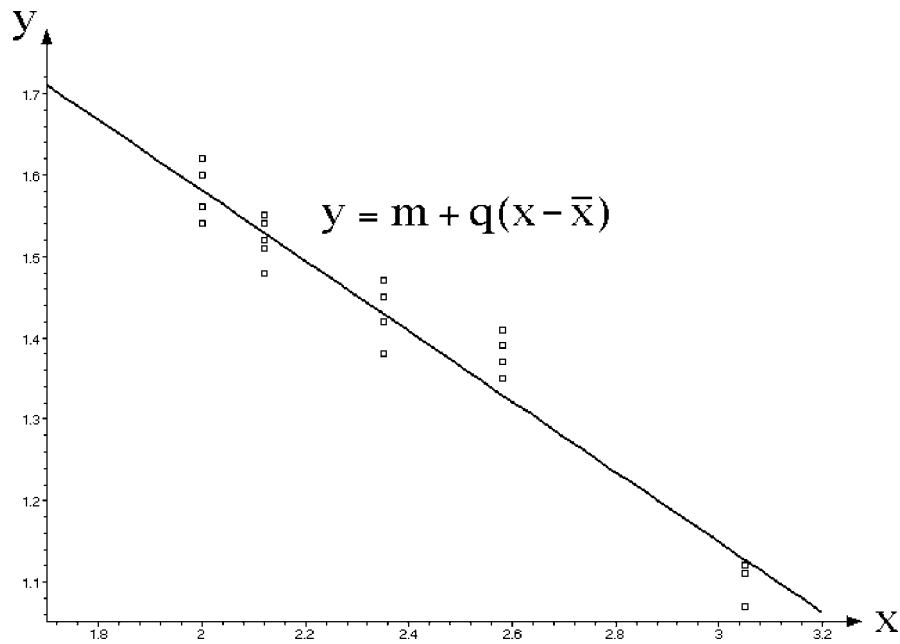
and since $t_{[0.975](19)} = 2.093$ the confidence interval for the prediction of y at $x = x_0$ reduces to:

$$y_0 = 1.420952 - 0.432164(x_0 - 2.372381) \pm 0.08312103 \sqrt{1.04761905 + 0.412272367(x_0 - 2.372381)^2}$$

The confidence interval of y at $x = x_0 = 2.55$ writes therefore:

$$y_0 = [1.34419 \pm 0.08560] = [1.25859, 1.42979]$$

In the following picture the regression line is superimposed to the experimental data (dots):



The confidence region for predictions (at a confidence level of 95%) is evidenced in the figure below (factor V exaggerated)

