

**Doctoral School of Materials Engineering**  
**Methods of statistical and numerical analysis (integrated course). Part I**  
**Final test - June 9th 2009**

**Solution of Exercise 1**

The Young's modulus is expressed by the formula

$$E = 2G(1 + \nu) \quad (1)$$

in terms of the Poisson's ratio  $G$  and the shear modulus  $\nu$ :

$$\nu = \bar{\nu} \pm \Delta\nu = 0.34 \pm 0.01 \quad G = \bar{G} + \Delta G = (44.7 \pm 0.1) \text{ GPa}.$$

Substitution into equation (1) of the estimated values  $\bar{\nu}$  and  $\bar{G}$  provides the estimate of  $E$ :

$$\bar{E} = 2\bar{G}(1 + \bar{\nu}) = 2 \cdot 44.7 \cdot (1 + 0.34) = 119.796 \text{ GPa}.$$

The function is a simple polynomial and the error propagation on it can be easily analyzed by the logarithmic differential method. In fact we have

$$\ln E = \ln 2 + \ln G + \ln(1 + \nu)$$

and whence the differential

$$\frac{dE}{E} = \frac{dG}{G} + \frac{d\nu}{1 + \nu}$$

which provides the needed formula for the upper estimate of the relative error on Young's modulus:

$$\frac{\Delta E}{\bar{E}} = \frac{\Delta G}{\bar{G}} + \frac{\Delta\nu}{1 + \bar{\nu}} = \frac{0.1}{44.7} + \frac{0.01}{1 + 0.34} = 0.009699823.$$

The absolute error of  $E$  is therefore

$$\Delta E = \bar{E} \frac{\Delta E}{\bar{E}} = 119.796 \cdot 0.009699823 = 1.162$$

and the final result takes the form

$$E = (119.8 \pm 1.2) \text{ GPa}.$$

The precision of the estimate is expressed by the percent error

$$100 \cdot \frac{\Delta E}{\bar{E}} = 100 \cdot 0.009699823 = 0.97\%$$

and appears quite satisfactory.

## Solution of Exercise 2

The application of Chauvenet criterion is illustrated in the table below:

$\rho$	$\Delta\rho$	$ABS(\Delta\rho)$	$\Delta\rho^2$	
199,00	9,67	9,67	93,44	Outlier
184,00	-5,33	5,33	28,44	
186,00	-3,33	3,33	11,11	
192,00	2,67	2,67	7,11	
196,00	6,67	6,67	44,44	
191,00	1,67	1,67	2,78	
190,00	0,67	0,67	0,44	
183,00	-6,33	6,33	40,11	
185,00	-4,33	4,33	18,78	
195,00	5,67	5,67	32,11	
189,00	-0,33	0,33	0,11	
182,00	-7,33	7,33	53,78	
mean( $\rho$ ):	189,33	st.dev.( $\rho$ ):	5,4993	
Maximum of $ABS(\Delta\rho)$ :	9,6667	corresponding to the value of $\rho$ :	199,00	Outlier
Distance z of the outlier from the mean in standard deviation units:				1,7578
Probability of a larger distance from the mean (see table):		z	area from 0 to z	residual area
		1,7500	0,45994	0,08012
		1,7600	0,46080	0,07840
Linearly interpolated value:		1,7578	0,46061	0,07878
Mean number of expected events out of 12 measurements:				0,9453

The mean number of outliers is larger than 1/2. Thus, according to Chauvenet criterion, the outlier 199 cannot be rejected as not belonging to the population.

The sample estimates of the mean and standard deviation are given by:

$$\bar{\rho} = \frac{1}{12} \sum_{i=1}^{12} \rho_i = 189.33 \quad s = \sqrt{\frac{1}{11} \sum_{i=1}^{12} (\rho_i - \bar{\rho})^2} = 5.4993.$$

Whence it is apparent that the farthest datapoint from the sample mean  $\bar{\rho}$  is  $\rho_{\text{out}} = 199$ , as shown by the  $ABS(\Delta\rho)$  column in the previous table. Such an outlier may not belong to the statistical population of the normal sample. The distance of the suspect value from the mean, in units of  $s$ , is expressed as

$$z = \frac{\rho_{\text{out}} - \bar{\rho}}{s} = \frac{199.00 - 189.33}{5.4993} = 1.7578.$$

The probability of finding a datapoint at a distance greater than 1.7578 standard deviations from the mean can be calculated from the table of the standard normal cumulative probability distribution:

$$\begin{aligned} P(|\rho_{\text{out}} - \bar{\rho}| \geq 1.7578s) &= 1 - P(|\rho_{\text{out}} - \bar{\rho}| < 1.7578s) = \\ &= 1 - 2 \cdot P(\bar{\rho} \leq \rho_{\text{out}} < \bar{\rho} + 1.7578s) \\ &= 1 - 2 \cdot 0.46061 = 0.07878. \end{aligned}$$

Although the probability  $P = P(\bar{\rho} \leq \rho_{\text{out}} < \bar{\rho} + 1.7578s)$  is not directly readable on the table, it can be estimated with satisfactory accuracy by a linear interpolation scheme:

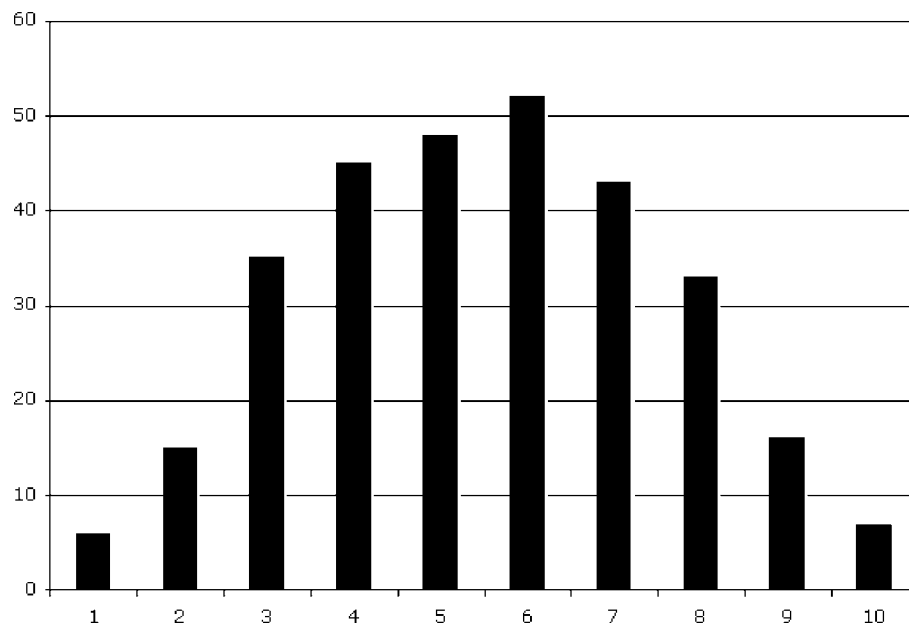
1.7500	0.45994	$\frac{1.7578 - 1.7500}{1.7600 - 1.7500} = \frac{P - 0.45994}{0.46080 - 0.45994}$
1.7578	$P$	
1.7600	0.46080	

which provides  $P = 0.46061$ .

Out of 12 measurements, one presumably expects  $12 \cdot 0.07878 = 0.94536$  outliers at a distance larger than  $1.7578s$  from the mean. *Since*  $0.94536 > 1/2$ , *Chauvenet criterion suggests that*  $\rho_{\text{out}}$  *cannot be rejected as not belonging to the statistical population.*

### Solution of Exercise 3

The sample histogram is bell-shaped, so that it seems rather plausible that the data belong to a normal population:



All the empirical frequencies are sufficiently high ( $f_i \geq 3$ ) to allow the application of the  $\chi^2$  test to check whether the population is normal, i.e. the null hypothesis

$$H_0 : \text{the population is normal, with distribution } N(\mu, \sigma)$$

against the alternative hypothesis

$$H_1 : H_0 \text{ is false .}$$

In the present analysis the sample data are used to estimate the mean and the standard deviation of the distribution:

$$\mu = \overline{m} \quad \sigma = s$$

and the number of result classes (i.e. the histogram intervals) is  $k = 10$ . In the presence of  $c = 2$  constraints on the mean and the standard deviation, if  $H_0$  holds true the  $\mathcal{X}^2$  of data obeys approximately a  $\mathcal{X}^2$  distribution with

$$n = k - c - 1 = 10 - 2 - 1 = 7$$

degrees of freedom. To calculate the  $\mathcal{X}^2$ , the expected frequencies in each class are needed, under the assumption that the normal distribution is correct. The endpoints of the classes differ from the mean  $\mu = \overline{m}$  by half-integer multiples of  $\sigma = s$ , thus the theoretical frequencies can be derived directly from the standard normal cumulative probability distribution. For simplicity's sake, it is convenient to define

$$p(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

and introduce the integral of the standard normal distribution

$$\Phi(z) = \int_0^z \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2} d\xi$$

whose values are tabulated. Denoted with  $n_i$  the frequency in the  $i$ -th class, the expected frequencies are determined as follows:

$$\begin{aligned} n_1 &= 300 \cdot \int_{-\infty}^{-2} p(z) dz = 300 \cdot \int_2^{+\infty} p(z) dz = 300 \cdot \left( \frac{1}{2} - \int_0^2 p(z) dz \right) = \\ &= 300 \cdot \left( \frac{1}{2} - \Phi(2) \right) = 300 \cdot \left( \frac{1}{2} - 0.47725 \right) = 6.825 \\ n_2 &= 300 \cdot \int_{-2}^{-1.5} p(z) dz = 300 \cdot \int_{1.5}^2 p(z) dz = 300 \cdot \left( \Phi(2) - \Phi(1.5) \right) = \\ &= 300 \cdot (0.47725 - 0.43319) = 13.218 \\ n_3 &= 300 \cdot \int_{-1.5}^{-1} p(z) dz = 300 \cdot \int_1^{1.5} p(z) dz = 300 \cdot \left( \Phi(1.5) - \Phi(1) \right) = \\ &= 300 \cdot (0.43319 - 0.34134) = 27.555 \end{aligned}$$

$$\begin{aligned}
n_4 &= 300 \cdot \int_{-1}^{-0.5} p(z) dz = 300 \cdot \int_{0.5}^1 p(z) dz = 300 \cdot (\Phi(1) - \Phi(0.5)) = \\
&= 300 \cdot (0.34134 - 0.19146) = 44.964 \\
n_5 &= 300 \cdot \int_{-0.5}^0 p(z) dz = 300 \cdot \int_0^{-0.5} p(z) dz = 300 \cdot \Phi(0.5) = \\
&= 300 \cdot 0.19146 = 57.438
\end{aligned}$$

whereas, owing to the symmetry of the normal distribution with respect to the mean, the other theoretical frequencies are symmetrically equal to the previous ones:

$$\begin{aligned}
n_6 &= n_5 = 57.438 & n_7 &= n_4 = 44.964 & n_8 &= n_3 = 27.555 \\
n_9 &= n_2 = 13.218 & n_{10} &= n_1 = 6.825.
\end{aligned}$$

A comparison is made then between the empirical frequencies  $f_i$  and the expected ones  $n_i$  for all the classes, as summarized in the table below:

i	class of $G$	empirical frequency	expected frequency
1	$m < \bar{m} - 2.0s$	6	6.825
2	$\bar{m} - 2.0s \leq m < \bar{m} - 1.5s$	15	13.218
3	$\bar{m} - 1.5s \leq m < \bar{m} - 1.0s$	35	27.555
4	$\bar{m} - 1.0s \leq m < \bar{m} - 0.5s$	45	44.964
5	$\bar{m} - 0.5s \leq m < \bar{m}$	48	57.438
6	$\bar{m} \leq m < \bar{m} + 0.5s$	52	57.438
7	$\bar{m} + 0.5s \leq m < \bar{m} + 1.0s$	43	44.964
8	$\bar{m} + 1.0s \leq m < \bar{m} + 1.5s$	33	27.555
9	$\bar{m} + 1.5s \leq m < \bar{m} + 2.0s$	16	13.218
10	$\bar{m} + 2.0s \leq m$	7	6.825

The  $\chi^2$  of the sample is given by:

$$\chi^2 = \sum_{i=1}^{10} \frac{(f_i - n_i)^2}{n_i} = 6.1690.$$

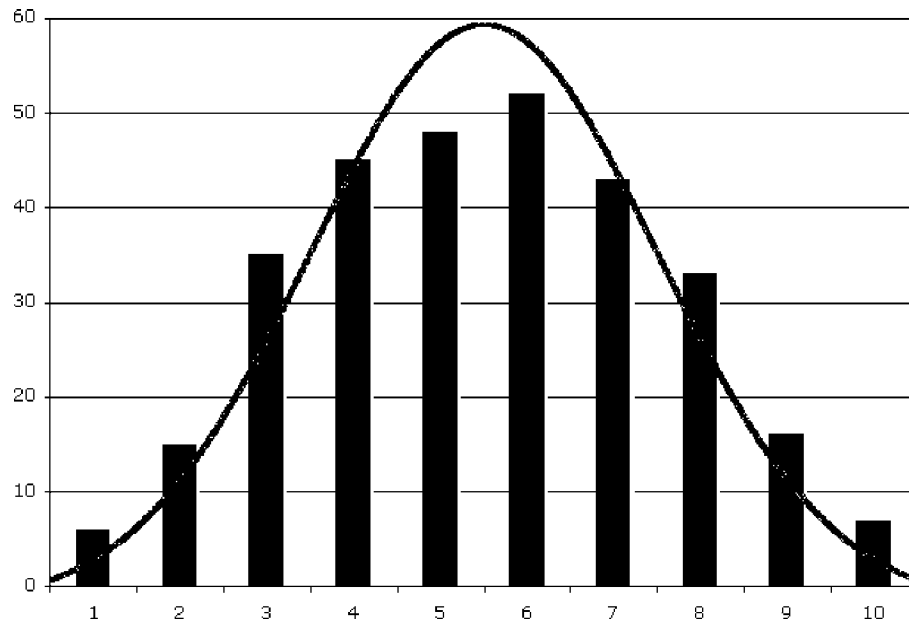
The table of the  $\chi^2$  cumulative distributions provides the critical values:

$$\chi^2_{[1-\alpha](7)} = \chi^2_{[0.90](7)} = 12.017 \quad \text{for } \alpha = 0.10$$

$$\chi^2_{[1-\alpha](7)} = \chi^2_{[0.95](7)} = 14.067 \quad \text{for } \alpha = 0.05.$$

In both cases the  $\chi^2$  of the sample is smaller: as a conclusion, with both the significance levels of 10 and 5% *the null hypothesis cannot be rejected*. The data sample supports the feeling that *the shear modulus of the polymer follows a normal distribution*.

The formal conclusion is also suggested by the good superposition between the theoretical distribution and the histogram, as shown in the figure:



although the histogram appears a little bit “flatter” around the mean. Pay attention to the way the theoretical distribution is calculated (solid curve). In the histogram the class 2 is centred at  $x = 2$  and the class 9 at  $x = 9$ . The centre of the class 2 must correspond to the point

$$x = \frac{\mu - 2.0\sigma + \mu - 1.5\sigma}{2} = \mu - 1.75\sigma$$

and that of the class 9 to the point

$$x = \frac{\mu + 1.5\sigma + \mu + 2.0\sigma}{2} = \mu + 1.75\sigma$$

so that the parameters  $\mu$  and  $\sigma$  of the normal distribution are determined by the linear equations

$$\mu - 1.75\sigma = 2 \qquad \mu + 1.75\sigma = 9$$

which provide

$$\mu = 5.5 \qquad \sigma = 2.0.$$

The normal distribution which must be superimposed to the histogram is then

$$300 \cdot p(x) = 300 \cdot \frac{1}{2\sqrt{2\pi}} e^{-(x-5.5)^2/8}.$$

**Solution of Exercise 4**

As  $n = 500 > 30$  the sample can be regarded as large and it is not necessary to assume a normal distribution of the data.

(a) In this case the confidence level is  $1 - \alpha = 0.67$ , so that  $\alpha = 0.33$  and

$$\frac{\alpha}{2} = \frac{0.33}{2} = 0.165 \quad \implies \quad \frac{1}{2} - \frac{\alpha}{2} = 0.5 - 0.165 = 0.335.$$

The table of the standard normal distribution suggests the following linear interpolation scheme:

$(1 - \alpha)/2$	$z_\alpha$
0.33398	0.97
0.33500	$z_{0.33}$
0.33646	0.98

$$\frac{0.33500 - 0.33398}{0.33646 - 0.33398} = \frac{z_{0.33} - 0.97}{0.98 - 0.97}$$

which provides the critical value

$$z_{0.33} = 0.9741.$$

The confidence interval for the mean  $\mu$  of the bolt length takes then the form

$$\bar{x} - z_{0.33} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{0.33} \frac{s}{\sqrt{n}}$$

with the sample mean and standard deviation given by

$$\bar{x} = 7.45 \quad s = 0.05$$

whereas  $n = 500$ . By inserting all the numbers, the confidence interval becomes

$$7.45 - 0.9741 \cdot \frac{0.05}{\sqrt{500}} \leq \mu \leq 7.45 + 0.9741 \cdot \frac{0.05}{\sqrt{500}}$$

and therefore, after trivial calculations,

$$7.4478 \text{ mm} \leq \mu \leq 7.4522 \text{ mm}.$$

The same confidence interval can be expressed in the equivalent form:

$$\mu = (7.4500 \pm 0.0022) \text{ mm}.$$

(b) For the confidence level  $1 - \alpha = 0.99$  there holds  $\alpha/2 = 0.005$  and therefore

$$\frac{1 - \alpha}{2} = \frac{0.99}{2} = 0.495.$$

The table of the standard normal cumulative probability distribution provides then, to a good degree of accuracy, the estimate

$$z_\alpha = z_{0.01} = 2.58$$

due to the approximate relationship

$$\int_0^{2.58} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 0.49506.$$

The CI of the mean becomes then

$$\bar{x} - z_{0.01} \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{0.01} \frac{s}{\sqrt{n}}$$

and, equivalently,

$$7.45 - 2.58 \cdot \frac{0.05}{\sqrt{500}} \leq \mu \leq 7.45 + 2.58 \cdot \frac{0.05}{\sqrt{500}}$$

and finally

$$7.4442 \text{ mm} \leq \mu \leq 7.4558 \text{ mm}.$$

An alternative expression puts into evidence the absolute error:

$$\mu = (7.4500 \pm 0.0058) \text{ mm}.$$

As expected, the confidence interval with confidence level of 99% is about three times larger than that at confidence level of 67%, a general and well-known feature of the normal distribution. A special comment deserves the number of significant digits of the results. In the problem proposed, both the mean and the standard deviation of the sample were denoted with 2 significant digits after the decimal point. If this were taken as an indication that the measurement procedure adopted is not able to detect lengths smaller than 0.01 mm, then the correct conclusion of the previous procedure would be simply that the estimate of 7.45 mm for the mean length can be regarded as completely certain: the calculation of a CI width smaller than 0.01 mm would be meaningless, because beyond the sensitivity of the experimental procedure used for measurements.

### Solution of Exercise 5

The number of the sample data is  $n = 22$ , and therefore the sample mean is simply the arithmetic mean

$$\bar{\sigma} = \frac{1}{n} \sum_{i=1}^n \sigma_i = 25.7170455,$$

on having denoted with  $\sigma_i$ ,  $i = 1, \dots, 22$ , the conductivity data of the sample.

We can then determine the residuals of the data with respect to the mean and the corresponding squares:

i	$\sigma_i$	$(\sigma_i - \bar{\sigma}) \cdot 10^2$	$(\sigma_i - \bar{\sigma})^2 \cdot 10^4$
1	25.711	-0.6045455	0.3654752
2	25.741	2.3954545	5.7382025
3	25.752	3.4954545	12.2182025
4	25.687	-3.0045455	9.0272934
5	25.713	-0.4045455	0.1636570
6	25.734	1.6954545	2.8745661
7	25.681	-3.6045455	12.9927479
8	25.717	-0.0045455	0.0000207
9	25.739	2.1954545	4.8200207
10	25.753	3.5954545	12.9272934
11	25.671	-4.6045455	21.2018388
12	25.707	-1.0045455	1.0091116
13	25.729	1.1954545	1.4291116
14	25.656	-6.1045455	37.2654752
15	25.720	0.2954545	0.0872934
16	25.695	-2.2045455	4.8600207
17	25.742	2.4954545	6.2272934
18	25.753	3.5954545	12.9272934
19	25.743	2.5954545	6.7363843
20	25.713	-0.4045455	0.1636570
21	25.688	-2.9045455	8.4363843
22	25.730	1.2954545	1.6782025

from which we deduce the sample variance:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (\sigma_i - \bar{\sigma})^2 = 7.7690260 \cdot 10^{-4}$$

and the sample estimate of the standard deviation:

$$s = \sqrt{s^2} = 2.7872973 \cdot 10^{-2}.$$

The sample cannot be regarded as large because the number of data is smaller than 30. As a consequence, it is necessary to compute the correct confidence interval for the mean by using the hypothesis of the normal population. For the same reason, the sample variance  $s^2$  cannot be assumed as practically equal to the variance  $\sigma^2$  of the population, as it would be ensured by the weak law of large numbers (Kintchine's theorem) in the case of a large sample: an appropriate confidence interval is needed also for  $\sigma^2$ .

(a) The CI of the mean, with confidence level  $1 - \alpha$ , is expressed by

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

In the present case we have  $1 - \alpha = 0.90$  and therefore  $\alpha = 0.10$ , while  $n = 22$ . The CI has thus the lower and upper limits

$$\begin{aligned} \bar{\sigma} - t_{[0.95](21)} \frac{s}{\sqrt{22}} &= 25.7170455 - 1.721 \cdot \frac{0.027872973}{\sqrt{22}} = 25.7068183 \\ \bar{\sigma} + t_{[0.95](21)} \frac{s}{\sqrt{22}} &= 25.7170455 + 1.721 \cdot \frac{0.027872973}{\sqrt{22}} = 25.7272726 \end{aligned}$$

so that the confidence interval becomes

$$25.7068183 \cdot 10^6 \text{ S m}^{-1} \leq \mu \leq 25.7272726 \cdot 10^6 \text{ S m}^{-1}$$

or, equivalently,

$$\mu = (25.7170455 \pm 0,0102271) \cdot 10^6 \text{ S m}^{-1}.$$

For all practical purposes an approximation of the form

$$(25.717 \pm 0.010) \cdot 10^6 \text{ S m}^{-1}$$

can be considered more than satisfactory.

(b) The confidence interval of the variance takes the form

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

still with  $\alpha = 0.10$  and  $n = 22$ . Therefore:

$$\begin{aligned} \frac{1}{\chi^2_{[0.95](21)}} 21 s^2 &= \frac{1}{32.671} 21 \cdot 7.7690260 \cdot 10^{-4} = 4.9937 \cdot 10^{-4} \\ \frac{1}{\chi^2_{[0.05](21)}} 21 s^2 &= \frac{1}{11.591} 21 \cdot 7.7690260 \cdot 10^{-4} = 14.0755 \cdot 10^{-4} \end{aligned}$$

and the CI of the variance is expressed as

$$4.9937 \cdot 10^{-4} \cdot 10^{12} (\text{S m}^{-1})^2 \leq \sigma^2 \leq 14.0755 \cdot 10^{-4} \cdot 10^{12} (\text{S m}^{-1})^2.$$

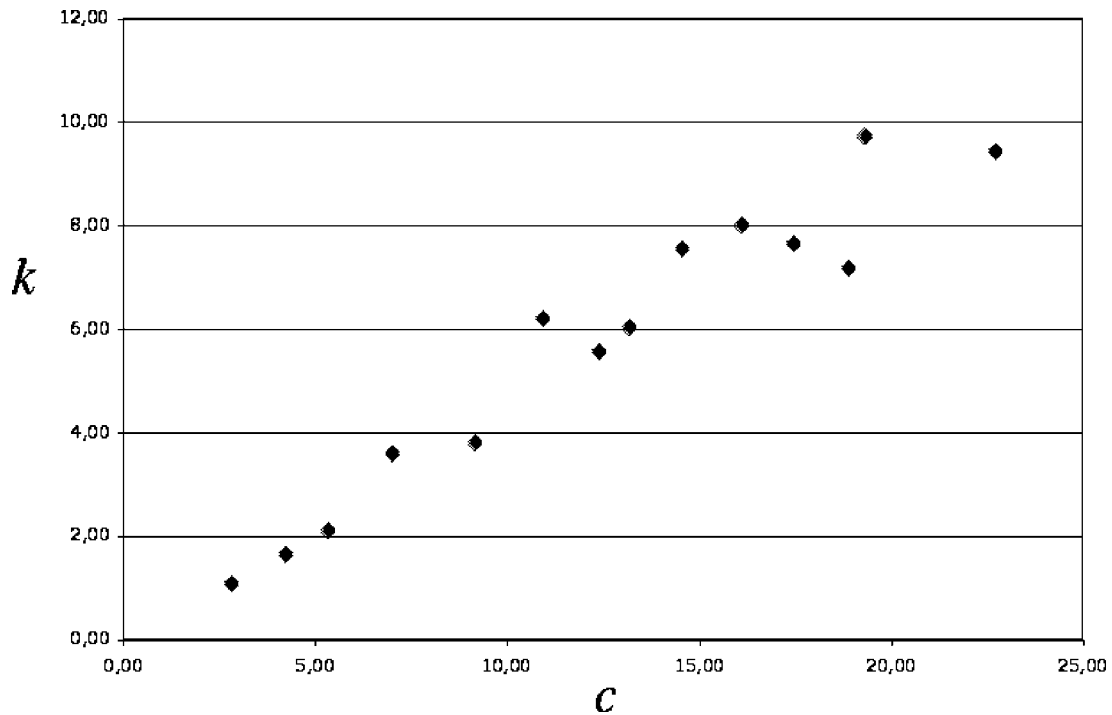
The required CI of the standard deviation is determined by taking the square root of the previous inequality side by side:

$$2.2346614 \cdot 10^4 \text{ S m}^{-1} \leq \sigma \leq 3.7517378 \cdot 10^4 \text{ S m}^{-1},$$

i.e.,  $2.2 \cdot 10^4 \text{ S m}^{-1} \leq \sigma \leq 3.8 \cdot 10^4 \text{ S m}^{-1}$ .

### Solution of Exercise 6

The plot of the data suggests that the quantities  $c$  and  $k$  may be described by dependent random variables (what means that they are correlated, owing to the hypothesis of normal random variables):



Denoted with  $(c_i, k_i)$ ,  $i = 1, 1 \dots, 14$ , each pair of data, the sample means  $\bar{c}$  and  $\bar{k}$  of the crystallinity degree and thermal conductivity are given by:

$$\bar{c} = \frac{1}{14} \sum_{i=1}^{14} c_i = 12.439286 \quad \bar{k} = \frac{1}{14} \sum_{i=1}^{14} k_i = 5.705000$$

and allow us to calculate the sum of products of residuals

$$SS_{ck} = \sum_{i=1}^{14} (c_i - \bar{c})(k_i - \bar{k}) = 218.313450$$

and of the relative squares:

$$SS_{cc} = \sum_{i=1}^{14} (c_i - \bar{c})^2 = 491.026293$$

$$SS_{kk} = \sum_{i=1}^{14} (k_i - \bar{k})^2 = 104.163750,$$

as illustrated in the following table:

c	k	$\Delta c$	$\Delta k$	$\Delta c^2$	$\Delta k^2$	$\Delta c * \Delta k$
16,11	8,02	3,670714	2,315000	13,474143	5,359225	8,497704
9,17	3,82	-3,269286	-1,885000	10,688229	3,553225	6,162604
7,01	3,62	-5,429286	-2,085000	29,477143	4,347225	11,320061
10,93	6,23	-1,509286	0,525000	2,277943	0,275625	-0,792375
14,55	7,57	2,110714	1,865000	4,455115	3,478225	3,936482
18,89	7,20	6,450714	1,495000	41,611715	2,235025	9,643818
13,19	6,05	0,750714	0,345000	0,563572	0,119025	0,258996
2,83	1,10	-9,609286	-4,605000	92,338372	21,206025	44,250761
22,72	9,46	10,280714	3,755000	105,693086	14,100025	38,604082
12,39	5,59	-0,049286	-0,115000	0,002429	0,013225	0,005668
4,23	1,66	-8,209286	-4,045000	67,392372	16,362025	33,206561
5,34	2,12	-7,099286	-3,585000	50,399858	12,852225	25,450939
19,32	9,75	6,880714	4,045000	47,344229	16,362025	27,832489
17,47	7,68	5,030714	1,975000	25,308086	3,900625	9,935661
12,439286	5,705000			491,026293	104,163750	218,313450
c mean	k mean					
				r=	0,965317	
				t=	12,808076	

The linear correlation coefficient becomes

$$r = \frac{SS_{ck}}{\sqrt{SS_{cc}} \sqrt{SS_{kk}}} = \frac{218.313450}{\sqrt{491.026293} \sqrt{104.163750}} = 0.965317.$$

As both  $c$  and  $k$  are assumed normal, we can check the null hypothesis

$$H_0 : c \text{ and } k \text{ are stochastically independent}$$

against the alternative hypothesis

$$H_1 : c \text{ and } k \text{ are stochastically dependent}$$

by means of the Fisher's random variable

$$t = \sqrt{n-2} \frac{r}{\sqrt{1-r^2}}$$

that, whenever  $H_0$  holds true, follows a Student distribution with  $n-2$  d.o.f. In the present case we get:

$$t = \sqrt{14-2} \frac{0.965317}{\sqrt{1-0.965317^2}} = 12.808076.$$

For a significance level  $\alpha = 5\%$  the critical region has the form

$$|t| > t_{[1-\frac{\alpha}{2}](n-2)} = t_{[0.975](12)} = 2.179,$$

while when the requested significance level is  $\alpha = 1\%$  it becomes

$$|t| > t_{[1-\frac{\alpha}{2}](n-2)} = t_{[0.995](12)} = 3.055.$$

*In both cases  $H_0$  must be rejected.* We conclude therefore that a correlation probably exists between the degree of crystallinity  $c$  and the thermal conductivity  $k$ . Moreover, due to the positive sign of the correlation coefficient which is very close to  $+1$ , the relation should be direct: *the thermal conductivity of the material increases with the degree of crystallinity.*

### Solution of Exercise 7

The temperature data are not affected by appreciable random errors, while the corresponding values of dynamic viscosity are regarded as the outcomes of independent normal random variables. It is then possible to apply the standard theory of linear regression, with the further simplification due to the homoscedastic character of the model — we may assume that all the random variables which describe the dynamic viscosity at different temperatures share the same variance. The regression straight line is defined by putting the temperature  $T$  along the abscissa axis and the dynamic viscosity  $\mu$  along the ordinate axis:

$$\mu = \alpha + \beta(T - \bar{T})$$

on having denoted with  $\bar{T}$  the sample mean of the temperatures, while  $\alpha$  and  $\beta$  are the parameters of the regression model. We recall that a model of this form ensures the stochastic independence of the best-fit estimates, say  $a$  and  $b$ , of the regression parameters  $\alpha$  and  $\beta$ .

Notice that the sample consists, as often in common laboratory practice, in multiple measurements at a constant temperature: many measurements of dynamic viscosity have been carried out at each given value of  $T$ . This circumstance does not constitute an obstacle to the application of the standard linear regression model, provided that all the pairs  $(T_i, \mu_i)$  with the same  $T$  are regarded as distinct. According to this criterion the whole number of sample data is thus  $n = 26$ .

#### (i) Regression straight line

Since all the standard deviations are equal, the  $\chi^2$  fitting reduces to the usual least-squares fitting and the best-fit estimates  $a, b$  of the parameters  $\alpha, \beta$  can be written as

$$a = \frac{1}{n} \sum_{i=1}^n \mu_i = 7.483192308 \quad b = \frac{\sum_{i=1}^n (T_i - \bar{T}) \mu_i}{\sum_{i=1}^n (T_i - \bar{T})^2} = -0.1450715928$$

with  $n = 26$  and

$$\bar{T} = \frac{1}{n} \sum_{i=1}^n T_i = 39.61538461$$

$$\sum_{i=1}^n (T_i - \bar{T})\mu_i = -1551.708076 \quad \sum_{i=1}^n (T_i - \bar{T})^2 = 10696.15385.$$

The regression straight line, calculated by the least-squares method, is therefore:

$$\begin{aligned} \mu &= a + b(T - \bar{T}) = 7.483192308 - 0.1450715928 (T - 39.61538461) = \\ &= 13.23025925 - 0.1450715928 T. \end{aligned}$$

(ii) *Confidence intervals for the regression parameters*

By definition, the sum of squares around regression holds:

$$\text{SSAR} = \sum_{i=1}^n [a + b(T_i - \bar{T}) - \mu_i]^2 = 17.30878358.$$

At the significance level  $1 - \alpha \in (0, 1)$ , the CI of the parameter  $\alpha$  and that of the slope  $\beta$  take the form:

$$\begin{aligned} \alpha &= a \pm t_{[1-\frac{\alpha}{2}](n-2)} \sqrt{\frac{1}{n} \frac{\text{SSAR}}{n-2}} \\ \beta &= b \pm t_{[1-\frac{\alpha}{2}](n-2)} \sqrt{\left[ \sum_{i=1}^n (T_i - \bar{T})^2 \right]^{-1} \frac{\text{SSAR}}{n-2}}. \end{aligned}$$

In the present case we have  $n = 26$  and  $1 - \alpha = 0.95$ , i.e.  $\alpha = 0.05$ , so that the confidence intervals become:

$$\begin{aligned} \alpha &= a \pm t_{[0.975](24)} \sqrt{\frac{1}{26} \frac{\text{SSAR}}{24}} \\ \beta &= b \pm t_{[0.975](24)} \sqrt{\left[ \sum_{i=1}^{26} (T_i - \bar{T})^2 \right]^{-1} \frac{\text{SSAR}}{24}} \end{aligned}$$

with:

$$\begin{aligned} a &= 7.483192308 \\ b &= -0.1450715928 \\ \text{SSAR} &= 17.30878358 \\ \sum_{i=1}^{26} (T_i - \bar{T})^2 &= 10696.15385 \\ t_{[0.975](24)} &= 2.064. \end{aligned}$$

Inserting the numerical values and performing the calculations we deduce that:

– the CI at 95% of the parameter  $\alpha$  is

$$7.483192308 \pm 0.3437563046 = [7.139436003, 7.826948613]$$

– the CI at 95% for the slope  $\beta$  holds

$$-0.1450715928 \pm 0.01694819786 = [-0.1620197907, -0.1281233949].$$

It is certainly appropriate to ignore the less significant digits, physically meaningless, and introduce the physical units, to conclude that:

$$\alpha = [7.139, 7.827] \cdot 10^{-4} \text{ Pa s} = (7.483 \pm 0.344) \cdot 10^{-4} \text{ Pa s}$$

while

$$\begin{aligned} \beta &= [-0.1620, -0.1281] \cdot 10^{-4} \text{ Pa s } ^\circ\text{C}^{-1} = \\ &= (-0.1451 \pm 0.0169) \cdot 10^{-4} \text{ Pa s } ^\circ\text{C}^{-1}. \end{aligned}$$

(iii) *Confidence region*

Taking in mind that the model is homoscedastic, the CI at a confidence level  $1 - \alpha$  for the prediction of  $\mu = \mu_0$  at a given  $T = T_0$  can be determined by the general formula:

$$\mathbb{E}(\rho_0) = a + b(T_0 - \bar{T}) \pm t_{[1-\frac{\alpha}{2}](n-2)} \sqrt{V} \sqrt{\frac{\text{SSAR}}{n-2}}$$

where we have:

$$a = 7.483192308$$

$$b = -0.1450715928$$

$$\bar{T} = \frac{1}{n} \sum_{i=1}^n T_i = 39.61538461$$

$$t_{[1-\frac{\alpha}{2}](n-2)} = t_{[0.975](24)} = 2.064$$

$$V = 1 + \frac{1}{n} + \frac{1}{\sum_{i=1}^n (T_i - \bar{T})^2} (T_0 - \bar{T})^2 = 1 + \frac{1}{26} + \frac{(T_0 - 39.61538461)^2}{10696.15385} =$$

$$= 1.038461538 + 0.9349154977 \cdot 10^{-4} (T_0 - 39.61538461)^2$$

$$\text{SSAR} = \sum_{i=1}^n [a + b(T_i - \bar{T}) - \mu_i]^2 = 17.30878358.$$

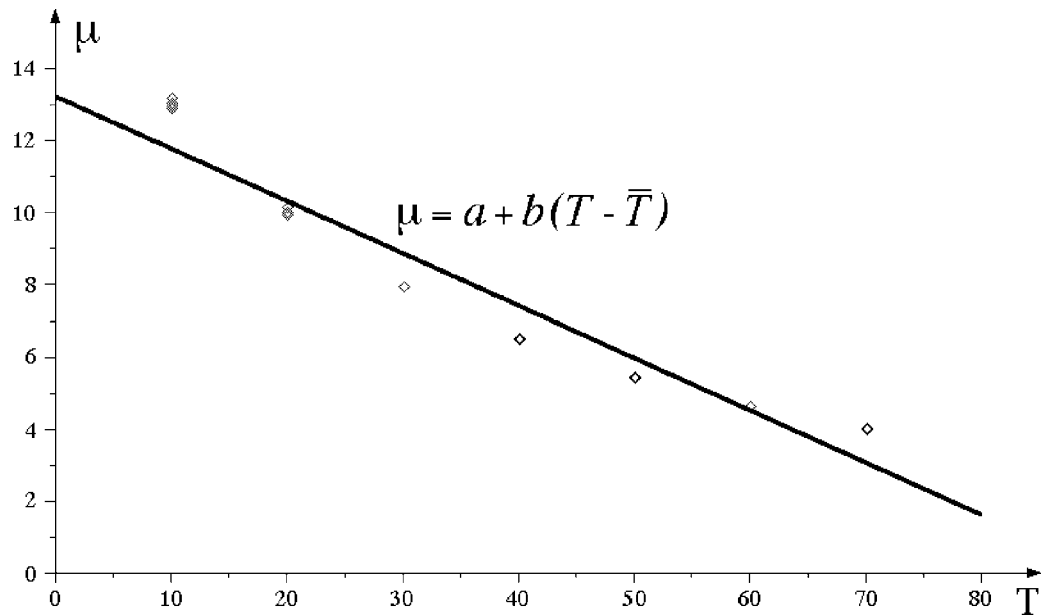
The CI for the prediction of  $\mu$  at  $T = T_0$  becomes then:

$$\begin{aligned} \mu_0 &= 7.483192308 - 0.1450715928 (T_0 - 39.61538461) \pm \\ &\pm 2.064 \cdot \sqrt{1.038461538 + 0.9349154977 \cdot 10^{-4} (T_0 - 39.61538461)^2} \sqrt{\frac{17.30878358}{24}} \end{aligned}$$

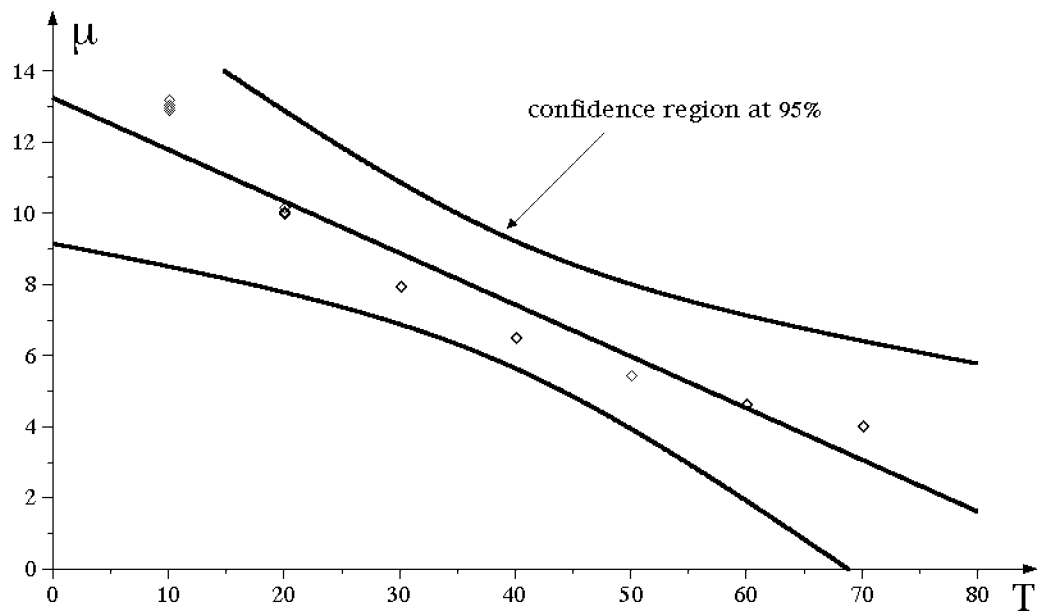
and performing the calculations reduces to:

$$\begin{aligned} \mu_0 &= 13.23025925 - 0.1450715928 T_0 \pm \\ &\pm 1.752820105 \sqrt{1.038461538 + 0.9349154977 \cdot 10^{-4} (T_0 - 39.61538461)^2}. \end{aligned}$$

In the following figure the regression straight line is superimposed to the experimental points:



The confidence region for predictions, at the confidence level of 95%, is shown in the figure below (where the coefficient of the  $(T_0 - T)^2$  term in the factor  $V$  has been enlarged by 30 for clarity's sake)



The regression model is the central straight line, while the upper and the lower curves represent the upper and the lower boundaries of the confidence region, respectively. The

width, measured parallel to the  $\mu$  axis, of the confidence region is minimum for  $T = \bar{T} = 39.61538461$  and tends to increase monotonically on the right and on the left of that point. To better stress the effect on the graph, the term  $(T_0 - \bar{T})^2$  which appears in the expression of  $V$  of the definition, has been enlarged by a scale factor 30.

(iv) *Confidence interval for a prediction*

The CI at a confidence level of 95% for the prediction of  $\mu$  at  $T = 1170$  can be obtained by posing  $T_0 = 25$  in the previous formula

$$\mu_0 = 13.23025925 - 0.1450715928 T_0 \pm 1.752820105 \sqrt{1.038461538 + 0.9349154977 \cdot 10^{-4} (T_0 - 39.61538461)^2} .$$

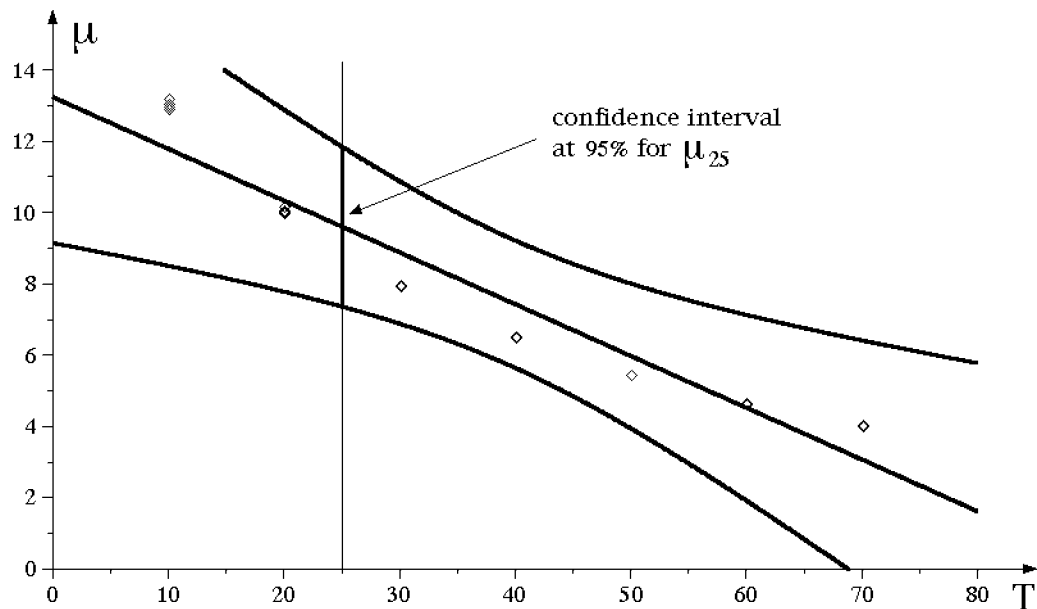
We have therefore:

$$\mu_{25} = \mu_{25} = [7.800165743, 11.40677312] = 9.603469432 \pm 1.803303688$$

i.e., dropping the less significant digits and introducing the unit of measure,

$$\mu_{25} = [7.8, 11.4] \cdot 10^{-4} \text{Pa s} = (9.6 \pm 1.8) \cdot 10^{-4} \text{Pa s} .$$

In the following figure the CI at 95% is simply the intersection of the confidence region at 95% with the vertical straight line of equation  $T = 25$ :



*(v) Goodness of fit*

The goodness of fit  $Q$  of the regression model is defined by the relationship

$$Q = \int_{\text{NSSAR}}^{+\infty} \rho_{n-2}(\mathcal{X}^2) d\mathcal{X}^2$$

where  $\rho_{n-2}$  denotes the  $\mathcal{X}^2$  distribution with  $n - 2$  d.o.f. This is because, if the regression model is correct, the normalized sum of squares around regression

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

constitutes a  $\mathcal{X}^2$  random variable with  $n - 2$  d.o.f. In order to determine the goodness of fit of the regression model *it is essential to know the common value of the standard deviation*  $\sigma = 0.05$ , because we need to compute the NSSAR and not simply the SSAR. In the present case we have  $n = 26$  data and the regression model is based on two parameters,  $\alpha$  and  $\beta$ ; as a consequence, the NSSAR follows a  $\mathcal{X}^2$  distribution with  $n - 2 = 24$  d.o.f. For the given sample the normalized sum of squares around regression holds

$$\text{NSSAR} = \frac{\text{SSAR}}{\sigma^2} = \frac{17.30878358}{0.05^2} = 6923.5143200.$$

On the table of the upper critical values of  $\mathcal{X}^2$  with  $\nu = 24$  d.o.f. we find

$$\text{Probability}\{\mathcal{X}^2 \geq 51.179\} = 0.001$$

so we expect that the goodness of fit  $Q$  be much smaller than 0.001, a value unsatisfactorily small. A precise numerical calculation of  $Q$ , according to the general definition

$$Q = \text{Probabilità}\{\mathcal{X}^2 \geq 6923.5143200\} = \int_{6923.5143200}^{+\infty} p_{24}(\mathcal{X}^2) d\mathcal{X}^2$$

can be carried out for instance by using the Maple 11<sup>TM</sup> command line

$$1 - \text{stats}[\text{statevalf}, \text{cdf}, \text{chisquare}[24]](6923.5143200);$$

which leads to an estimate practically null. If the regression model were rejected, the probability of a type I error would be virtually zero. As a conclusion, the regression model is certainly uncorrect and must be rejected.

It is noticeable that the same conclusion would be qualitatively suggested by the mere trend of the sample data compared with the least squares regression straight line, as shown in the previous figures. The data of repeated measurements appear indeed reproducibly rather far from the regression line, except than the values at  $T = 20$  and  $T = 60$ . The observed

trend suggests that a second order polynomial in  $T$  should be probably more appropriate to model the data.

Another possible interpretation of the results may be the following. If the regression model were correct the NSSAR would likely take a value in the confidence interval

$$[\nu - 3\sqrt{2\nu}, \nu + 3\sqrt{2\nu}] = [24 - 3\sqrt{2 \cdot 24}, 24 + 3\sqrt{2 \cdot 24}] = [3.215, 44.785]$$

of a  $\chi^2$  distribution with  $\nu = n - 2 = 24$  d.o.f. But in the present case *the NSSAR value greatly exceeds the upper limit of the above confidence interval*, what makes the hypothesis of a NSSAR which follows a  $\chi^2$  probability distribution quite unreasonable. The problem could arise from an underestimate of the standard deviation  $\sigma$ , assumed too optimistically small.

### Solution of Exercise 8

Let  $\mu_1$  and  $\mu_2$  be the mean values of the two statistical populations, i.e. the “true values” of electrical conductivity for the alloy after the first and the second treatment, respectively. Let  $p = 10$  be the number of data  $y_1, \dots, y_p$  of the first sample and  $q = 14$  that of the second data sample  $z_1, \dots, z_q$ .

We want to check the hypothesis  $H_0 : \mu_1 = \mu_2$  (the two treatments have essentially the same effect on the electrical conductivity of the material) against the alternative hypothesis  $H_1 : \mu_1 \neq \mu_2$  (the two treatments yield materials with a significantly different electrical conductivity). The significance level we assume is of 5%.

Since the form of the test is different if the variances  $\sigma_1^2$  and  $\sigma_2^2$  of the two populations are equal or not, we firstly check the null hypothesis  $\sigma_1^2 = \sigma_2^2$  versus the alternative  $\sigma_1^2 \neq \sigma_2^2$ .

(i) *Test on the variances*

Denoted with  $s_y^2$  and  $s_z^2$  the sample variances of the two samples, the null hypothesis  $H_0 : \sigma_1^2 = \sigma_2^2$  is accepted at a significance level  $\alpha$  if and only if the random variable

$$F = s_y^2/s_z^2$$

has a value within the acceptance interval

$$F_{[\frac{\alpha}{2}](p-1, q-1)} < F < F_{[1-\frac{\alpha}{2}](p-1, q-1)}.$$

In fact, when  $H_0$  holds true, the random variable  $F = s_y^2/s_z^2$  follows a Fisher distribution with  $(p - 1, q - 1)$  d.o.f. In this case we obtain

$$\begin{aligned} \bar{y} &= \frac{1}{p} \sum_{i=1}^p y_i = 15.517000 & \bar{z} &= \frac{1}{q} \sum_{j=1}^q z_j = 16.361429 \\ s_y^2 &= \frac{1}{p-1} \sum_{i=1}^p (y_i - \bar{y})^2 = 0.448068 \\ s_z^2 &= \frac{1}{q-1} \sum_{j=1}^q (z_j - \bar{z})^2 = 0.916105 \end{aligned}$$

and therefore the test variable holds

$$F = \frac{s_y^2}{s_z^2} = \frac{0.448068}{0.916105} = 0.489101$$

whereas the lower and upper limit of the acceptance region of  $H_0$ , assuming  $\alpha = 5\%$ , are given by

$$F_{[\frac{\alpha}{2}](p-1, q-1)} = F_{[0.025](9, 13)} = 0.2611$$

$$F_{[1-\frac{\alpha}{2}](p-1, q-1)} = F_{[0.975](9, 13)} = 3.3120.$$

Since  $F_{[0.025](9, 13)} < F < F_{[0.975](9, 13)}$ , we must conclude that the two normal populations have presumably the same variance.

(ii) *Test on the mean*

Due to the previous result, the  $t$ -test to compare the means of two normal populations of equal unknown variances is applicable. The test statistics is

$$t = \frac{\bar{y} - \bar{z}}{\sqrt{\frac{1}{p} + \frac{1}{q}} \cdot s}, \quad \text{with} \quad s^2 = \frac{(p-1)s_y^2 + (q-1)s_z^2}{p+q-2},$$

and the rejection region of  $H_0 : \mu_1 = \mu_2$  takes the form

$$\{t \leq -t_{[1-\frac{\alpha}{2}](p+q-2)}\} \cup \{t \geq t_{[1-\frac{\alpha}{2}](p+q-2)}\}.$$

In the present case we have

$$s^2 = \frac{9 \cdot 0.448068 + 13 \cdot 0.916105}{10 + 14 - 2} = 0.724635$$

and consequently

$$t = \frac{15.517000 - 16.361429}{\sqrt{\frac{1}{10} + \frac{1}{14}} \cdot \sqrt{0.724636}} = -2.395861.$$

Moreover, for  $\alpha = 0.05$  there holds

$$t_{[1-\frac{\alpha}{2}](p+q-2)} = t_{[0.975](22)} = 2.074$$

so that  $t$  belongs to the rejection region of  $H_0$ . We can conclude, at the significance level of 5%, that *the value of electrical conductivity of the metallic alloy is significantly affected by the temperature of the thermal treatment.*

The detailed calculations are illustrated in the table below:

$y$	$z$	$\Delta y$	$\Delta y^2$	$\Delta z$	$\Delta z^2$
16,30	15,89	0,783000	0,613089	-0,471429	0,222245
15,00	15,25	-0,517000	0,267289	-1,111429	1,235273
15,80	17,70	0,283000	0,080089	1,338571	1,791773
16,45	17,66	0,933000	0,870489	1,298571	1,686288
14,38	15,44	-1,137000	1,292769	-0,921429	0,849031
15,25	17,10	-0,267000	0,071289	0,738571	0,545488
16,02	17,48	0,503000	0,253009	1,118571	1,251202
14,77	15,95	-0,747000	0,558009	-0,411429	0,169273
15,68	15,05	0,163000	0,026569	-1,311429	1,719845
15,52	16,32	0,003000	0,000009	-0,041429	0,001716
	16,47			0,108571	0,011788
	16,10			-0,261429	0,068345
	15,24			-1,121429	1,257602
	17,41			1,048571	1,099502
mean( $y$ ) =	15,517000		$s_y^2 / s_z^2 =$	0,489101	
mean( $z$ ) =	16,361429		$s^2 =$	0,724636	
$s_y^2 =$	0,448068		$t =$	-2,395860	
$s_z^2 =$	0,916105				

### Solution of Exercise 9

In this case it seems obvious to apply a paired  $t$ -test for the comparison of the means, since the quantity is measured before and after the treatment *on each sample*. Therefore, data relative to the same sample will be paired:

$$(y_i, z_i) \quad i = 1, \dots, n,$$

denoting with  $y_i$  the values measured before the treatment and with  $z_i$  those measured after the treatment, on all the  $n = 10$  samples. We check the hypothesis  $H_0 : \mu_1 = \mu_2$ , that the treatment has no effect on the mean value of the measured quantity, against the alternative hypothesis  $H_1 : \mu_1 \neq \mu_2$  that such an effect actually exists. Whenever  $H_0$  holds true, the test variable

$$t = \sqrt{n} \frac{\bar{y} - \bar{z}}{\sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - z_i - \bar{y} + \bar{z})^2}}$$

follows a Student distribution with  $n - 1 = 9$  d.o.f. The two-sided critical region, with significance level  $\alpha$ , is of the form

$$\{t \leq -t_{[1-\frac{\alpha}{2}](n-1)}\} \cup \{t \geq t_{[1-\frac{\alpha}{2}](n-1)}\}$$

and for  $n = 10$ ,  $\alpha = 0.02$  becomes

$$\{t \leq -2.821\} \cup \{t \geq 2.821\}$$

since

$$t_{[1-\frac{\alpha}{2}](n-1)} = t_{[0.99](9)} = 2.821.$$

A simple calculation provides the means:

$$\bar{y} = \frac{1}{10} \sum_{i=1}^{10} y_i = 13.090 \quad \bar{z} = \frac{1}{10} \sum_{i=1}^{10} z_i = 13.850$$

while the paired sample estimate of the variance holds

$$\frac{1}{n-1} \sum_{i=1}^n (y_i - z_i - \bar{y} + \bar{z})^2 = \frac{1}{9} \cdot 6.524000 = 0.72488889$$

and the corresponding standard deviation is therefore

$$\sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - z_i - \bar{y} + \bar{z})^2} = \sqrt{0.72488889} = 0.85140407.$$

The test variable takes the value

$$t = \sqrt{10} \cdot \frac{13.090 - 13.850}{0.85140407} = -2.8228.$$

Here is the table of the detailed calculations:

$y_i$	$z_i$	$y_i - z_i$	$d_i = y_i - z_i - \text{mean}(y-z)$	$d_i^2$
10,5	12,3	-1,8	-1,0400	1,081600
13,0	13,2	-0,2	0,5600	0,313600
15,3	16,3	-1,0	-0,2400	0,057600
11,2	12,5	-1,3	-0,5400	0,291600
14,6	13,9	0,7	1,4600	2,131600
12,3	14,1	-1,8	-1,0400	1,081600
14,0	15,2	-1,2	-0,4400	0,193600
13,2	12,8	0,4	1,1600	1,345600
14,9	15,5	-0,6	0,1600	0,025600
11,9	12,7	-0,8	-0,0400	0,001600
mean(y-z) =	-0,7600		sum of $d_i^2$ 's =	6,524000
mean(y) =	13,0900		var.(y-z):	0,72488889
mean(z) =	13,8500		st.dev.(y-z):	0,85140407
Student t =	-2,8228			

The calculated value belongs to the lower tail of the rejection region, because

$$t = -2.8228 < -2.821 = -t_{[0.99](9)},$$

and therefore *we can exclude*, with a significance level of 2%, *that the mean value of the quantity is the same before and after the treatment. The null hypothesis must be rejected.*