Graduate School of Materials Engineering - a.y. 2009-2010
Methods of statistical and numerical analysis (integrated course). Part I
Final test - April 9th 2010

□ Exercise 1 — Points 4

The surface tension of α-bromonaphthalene at room temperature has been repeatedly measured by a tensiometer. The results are listed below (in $\text{mJ} \cdot \text{m}^{-2}$):

42.4	40.6	46.2	44.8	41.0	51.2	48.2	45.0
46.4	47.3	41.7	45.2	46.3	43.5	41.2	

and can be assumed to follow a normal distribution. Check for the eventual presence of an outlier not belonging to the statistical population of the measurements.

□ Exercise 2 — Points 3

The volumetric flow rate Q of water through a cylindrical pipe is given by the Hagen-Poiseuille formula:

$$Q = \frac{\pi \, \Delta P \, r^4}{8\mu L}$$

where:

- $L = (4.20 \pm 0.01)$ m is the length of the pipe;
- $r = (10.0 \pm 0.1)$ cm denotes the radius of the pipe;
- $\Delta P = (80.0 \pm 0.5)$ Pa is the pressure drop across the pipe;
- $\mu = (8.90 \pm 0.02) \cdot 10^{-4}$ Pa · s stands for the dynamic viscosity coefficient of the liquid.

Determine the estimated value of Q and the appropriate absolute error.

□ Exercise 3 — Points 4

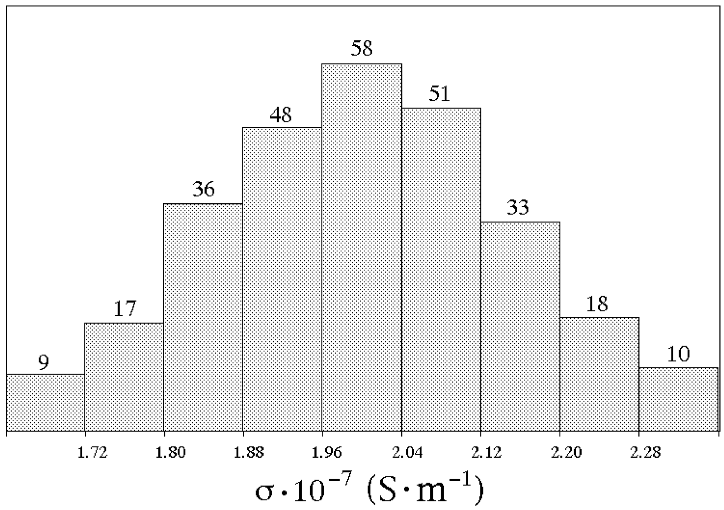
A random sample of 250 resistors, whose nominal electrical resistance is $500\,\Omega$, has been collected. By measuring the actual resistance of each device, one may find that the mean resistance is $495\,\Omega$, with a standard deviation of $14\,\Omega$. Determine:

(a) the confidence interval for the resistance of a resistor at a confidence level of 65%;

(b) the confidence interval for the resistance corresponding to a confidence level of 90%.

☐ Exercise 4 **Points 5**

The data of the electric conductivity of a metallic alloy are believed to obey a normal distribution. The hypothesis is checked by performing 280 independent measurements of the electrical conductivity σ. The sample mean turns out to be $\overline{\sigma} = 2.00 \cdot 10^7\,\mathrm{S}\cdot\mathrm{m}^{-1}$ and the sample estimate of the standard deviation is $s = 0.16 \cdot 10^7\,\mathrm{S}\cdot\mathrm{m}^{-1}$. The experimental data are summarized in the histogram below:



Check if the probability distribution of the sample can be really assumed to be normal, with a significance level (a) of 5% and (b) of 2%.

☐ Exercise 5 **Points 4**

Repeated measurements of the electromotive force v between the poles of a battery provide the values below (in V):

i	v_i	i	v_i	i	v_i
1	1.3770	7	1.3550	13	1.3551
2	1.3724	8	1.3550	14	1.3845
3	1.3570	9	1.3666	15	1.3406
4	1.3047	10	1.3663	16	1.3534
5	1.3585	11	1.3924	17	1.3970
6	1.3766	12	1.3625	18	1.3826

which can be assumed to be normal. Determine, at the confidence level of 90%, (i) the confidence interval of the mean μ and (ii) the CI of the variance σ^2.

☐ Exercise 6 **Points 4**

A sol-gel process provides ceramic samples whose porosity ε and thermal conductivity k vary in an unpredictable way according to a bivariate normal probability distribution. To check if the two quantities may be correlated, 16 measurements of ε and k have been carried out on as many samples. The results are listed in the following table:

i	ε_i	k_i	i	ε_i	k_i
1	1.1	5.34	9	3.3	5.43
2	1.3	6.63	10	3.6	5.08
3	1.7	5.77	11	3.8	4.37
4	2.0	5.98	12	4.0	4.57
5	2.4	5.13	13	4.2	3.66
6	2.7	5.39	14	4.6	3.75
7	2.9	4.49	15	4.8	4.21
8	3.1	5.06	16	4.9	3.84

where the porosity is expressed as a percentage and the thermal conductivity in $W \cdot m^{-1} \cdot K^{-1}$. Check whether the quantities ε and k can be regarded as stochastically independent, at a significance level (a) of 10%, (b) of 5% and (c) of 2%.

☐ Exercise 7 **Points 4**

A physicochemical treatment is applied to 12 polypropylene samples. The Young's modulus E of the material is measured for each sample prior to and after the treatment, providing the results below (in GPa):

sample no.	E before the treatment	E after the treatment
1	1.693	1.931
2	1.820	1.864
3	1.543	1.689
4	1.780	2.010
5	1.735	2.299
6	1.862	2.105
7	1.581	2.087
8	1.807	1.759
9	1.651	1.927
10	1.632	1.627
11	1.647	1.904
12	1.747	2.081

By assuming that the data are normal, check whether the treatment actually affects the Young's modulus of the material, with a significance level of 2%.

□ Exercise 8 **Points 6**

The dynamic viscosity μ (mPa $\cdot$ s) of pure water has been repeatedly measured at different values of the temperature T (°C), providing the following table:

i	T_i	μ_{i1}	μ_{i2}	μ_{i3}	μ_{i4}
1	35	0.7355	0.6913	0.6021	0.7890
2	40	0.5947	0.6190	0.4852	0.7069
3	45	0.6688	0.5211	0.5099	0.6829
4	50	0.5320	0.5709	0.5163	0.5625
5	55	0.4911	0.5524	0.5242	0.4497
6	60	0.4718	0.4667	0.3972	0.4270
7	65	0.4245	0.4128	0.4371	0.3892
8	70	0.4358	0.3192	0.3822	0.3516

The random error on the temperatures T_i is negligible, while the viscosity data μ_i can be regarded as independent normal random variables with the same standard deviation $\sigma = 0.06$. Determine:

(a) the least squares regression straight line of the form

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

where $\overline{T}$ denotes the arithmetic mean of the temperatures;

(b) the goodness of fit of the regression model;

(c) the 90% confidence intervals of the regression parameters α and β;

(d) the 90% confidence region for predictions;

(e) the 90% confidence interval for the value of μ predicted at $T = 52^oC$.

☐ Exercise 9 **Points 4**

A thermal treatment is applied to a metallic alloy in order to increase crystallinity and electrical conductivity. 11 samples of the alloy are thermally treated at a temperature of 1100 K. An analogous treatment of the same duration is applied to further 14 samples of the same material, but at a temperature of 700 K. The electrical conductivity of all the samples is finally measured. The data (in $10^6\,\mathrm{S}\cdot\mathrm{m}^{-1}$) are listed in the table below:

$T = 1100$ K	$T = 700$ K
21.7	15.4
12.5	14.4
12.6	15.9
18.6	13.1
19.4	11.5
15.9	10.0
17.5	16.4
15.4	12.0
22.2	10.5
18.6	15.4
13.6	12.5
	17.5
	18.7
	16.7

and can be assumed to be normal. After having checked whether the relative variances can or cannot be regarded as equal, verify with a significance level of 10% if the temperature of the thermal treatment has a significant effect on the electrical conductivity of the alloy.

Remark The sufficient grade corresponds to 18 points

Solution to Exercise 1

The possible presence of an outlier can be checked by using Chauvenet criterion, since the data are assumed to be normal. We must compute the sample mean $\overline{\gamma}$ and standard deviation s, and find the farthest data from the mean, as illustrated in the table below:

i	γ_i	$\gamma_i - \overline{\gamma}$	$\lvert\gamma_i - \overline{\gamma}\rvert$	$\lvert\gamma_i - \overline{\gamma}\rvert^2$	outlier
1	42.4	−2.3333	2.3333	5.44444444	
2	40.6	−4.1333	4.1333	17.08444444	
3	46.2	1.4667	1.4667	2.15111111	
4	44.8	0.0667	0.0667	0.00444444	
5	41.0	−3.7333	3.7333	13.93777778	
6	51.2	6.4667	6.4667	41.81777778	×
7	48.2	3.4667	3.4667	12.01777778	
8	45.0	0.2667	0.2667	0.07111111	
9	46.4	1.6667	1.6667	2.77777778	
10	47.3	2.5667	2.5667	6.58777778	
11	41.7	−3.0333	3.0333	9.20111111	
12	45.2	0.4667	0.4667	0.21777778	
13	46.3	1.5667	1.5667	2.45444444	
14	43.5	−1.2333	1.2333	1.52111111	
15	41.2	−3.5333	3.5333	12.48444444	

where:

$$\overline{\gamma} = \frac{1}{15}\sum_{i=1}^{15}\gamma_i = 44.7333 \qquad\qquad s = \sqrt{\frac{1}{14}\sum_{i=1}^{15}(\gamma_i - \overline{\gamma})^2} = 3.0210$$

while the outlier is the data $\gamma_6 = 51.2$, whose distance from the mean $\overline{\gamma}$ is maximum. The distance z of the suspect data from the mean, in units of s, can be expressed as

$$z = \frac{\gamma_6 - \overline{\gamma}}{s} = \frac{51.2 - 44.7333}{3.0210} = 2.1405$$

and is greater than the tabulated critical value of Chauvenet test for $n = 15$ data

$$z_{cr,15} = 2.128045234\,.$$

Therefore, the data should be rejected as an outlier not belonging to the statistical population. Alternatively, we can easily calculate the probability that a data has a distance from the mean greater than or equal to $2.1405s$, by using the table of the standard normal cumulative distribution. We have indeed

$$\begin{aligned} P(|\gamma_6 - \overline{\gamma}| \geq 2.1405\,s) &= 1 - P(|\gamma_6 - \overline{\gamma}| < 2.1405\,s) = \\ &= 1 - 2\,P(\overline{\gamma} \leq \gamma_6 < \overline{\gamma} + 2.1405s) = \\ &= 1 - 2\cdot 0.48384 = 0.03232 \end{aligned}$$

due to the value of $P(\overline{\gamma} \leq \gamma_6 < \overline{\gamma} + 2.1405s) = 0.48384$ which can be derived by means of the following linear interpolation scheme:

2.1400	0.48382
2.1405	P
2.1500	0.48422

$$\frac{2.1405 - 2.1400}{2.1500 - 2.1400} = \frac{P - 0.48382}{0.48422 - 0.48382}$$

As a consequence, the expected number of data at such a large distance from the mean, out of 14 datapoints, would be

$$0.03232 \cdot 15 = 0.48475$$

and since the result is smaller than 1/2 we must conclude, as before, that the outlier likely does not belong to the normal population and must be rejected.

Solution to Exercise 2
The volumetric flow rate Q must be calculated according to the formula

$$Q = \frac{\pi \,\Delta P \, r^4}{8\mu L}$$

with

$$L = (4.20 \pm 0.01)\,\mathrm{m} \qquad\qquad r = (10.0 \pm 0.1)\,\mathrm{cm}$$
$$\Delta P = (80.0 \pm 0.5)\,\mathrm{Pa} \qquad \text{and} \qquad \mu = (8.90 \pm 0.02) \cdot 10^{-4}\,\mathrm{Pa \cdot s}\,.$$

The estimate of the rate is reckoned by using the estimated true values of the length, radius, pressure drop and dynamic viscosity, respectively,

$$\overline{L} = 4.20\,\mathrm{m} \qquad \overline{r} = 10.0 \cdot 10^{-2}\,\mathrm{m} \qquad \overline{\Delta P} = 80.0\,\mathrm{Pa} \qquad \overline{\mu} = 8.90 \cdot 10^{-4}\,\mathrm{Pa \cdot s}$$

and writes therefore

$$\overline{Q} = \frac{\pi\,\overline{\Delta P}\,\overline{r}^4}{8\overline{\mu}\overline{L}} = \frac{3.14159 \cdot 80.0 \cdot (10.0 \cdot 10^{-2})^4}{8 \cdot 8.90 \cdot 10^{-4} \cdot 4.20} = 0.840447\,\mathrm{m}^3 \cdot \mathrm{s}^{-1}\,.$$

We can conveniently analyze the error propagation by using the logarithmic differential method. The relationship

$$\ln Q = \ln \pi + \ln \Delta P + 4 \ln r - \ln \mu - \ln L$$

provides indeed the differential

$$\frac{dQ}{Q} = \frac{d\pi}{\pi} + \frac{d\Delta P}{\Delta P} + 4\frac{dr}{r} - \frac{d\mu}{\mu} - \frac{dL}{L}$$

and thus the relative error estimate

$$\frac{\Delta Q}{\overline{Q}} = \frac{\Delta\pi}{\pi} + \frac{\Delta(\Delta P)}{\overline{\Delta P}} + 4\frac{\Delta r}{\overline{r}} + \frac{\Delta\mu}{\overline{\mu}} + \frac{\Delta L}{\overline{L}}$$

where the contribution of π can be taken arbitrarily small by simply considering an adequate number of significant digits — e.g., $\pi = 3.14159$. As a consequence:

$$\frac{\Delta Q}{\overline{Q}} = \frac{0.00001}{3.14159} + \frac{0.5}{80.0} + 4\cdot\frac{0.1}{10.0} + \frac{0.02}{8.90} + \frac{0.01}{4.20} = 0.0509 = 5.09\,\%$$

Whence we derive the absolute error on the volume rate:

$$\Delta Q = \frac{\Delta Q}{\overline{Q}}\cdot\overline{Q} = 0.0509\cdot 0.840447 = 0.042779\,\mathrm{m}^3\cdot\mathrm{s}^{-1}$$

so that the error interval of Q becomes

$$Q = \overline{Q} \pm \Delta Q = (0.840447 \pm 0.042779)\,\mathrm{m}^3\cdot\mathrm{s}^{-1}$$

or, rounding off the absolute error to two significant digits only,

$$Q = (0.840 \pm 0.043)\,\mathrm{m}^3\cdot\mathrm{s}^{-1}\,.$$

Solution to Exercise 3

The sample can be assumed to be large, because of the number of data $n = 250 > 30$. According to the theory of large samples, the statistical population is not required to be normal. The confidence interval for the mean μ, at a confidence level of $1-\alpha$, is expressed by

$$\overline{x} - z_{[1-\frac{\alpha}{2}]}\frac{s}{\sqrt{n}} \le \mu \le \overline{x} + z_{[1-\frac{\alpha}{2}]}\frac{s}{\sqrt{n}}$$

in terms of the sample estimates of the mean and standard deviation:

$$\overline{x} = 495\,\Omega \qquad\qquad s = 14\,\Omega\,,$$

while $z_{[1-\frac{\alpha}{2}]}$ denotes the inverse of the cumulative standard normal distribution at $1-\frac{\alpha}{2}$ (the critical value). The general form of the confidence interval is thus

$$495 - z_{[1-\frac{\alpha}{2}]}\frac{14}{\sqrt{250}} \le \mu \le 495 + z_{[1-\frac{\alpha}{2}]}\frac{14}{\sqrt{250}}$$

or, equivalently,

$$495 - 0.88543775\cdot z_{[1-\frac{\alpha}{2}]} \le \mu \le 495 + 0.88543775\cdot z_{[1-\frac{\alpha}{2}]}\,.$$

(a) Confidence level 65%
The confidence level is $1 - \alpha = 0.65$, so that $\alpha = 0.35$ and

$$1 - \frac{\alpha}{2} = 1 - \frac{0.35}{2} = 1 - 0.175 = 0.825 .$$

The critical value $z_{[1-\frac{\alpha}{2}]} = z_{[0.825]}$ is then calculated by the Excel function NORMINV:

$$\mathrm{NORMINV}(0,825;0;1)$$

which provides

$$z_{[0.825]} = 0.934589291 .$$

Alternatively, we can use the equation

$$\int_0^{z_{[1-\frac{\alpha}{2}]}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \frac{1}{2} - \frac{\alpha}{2} \tag{0.1}$$

and search for the value

$$\frac{1-\alpha}{2} = \frac{0.65}{2} = 0.325$$

among the entries of the standard normal table — which collects the integrals from 0 and $z > 0$ of the standard normal distribution — to obtain the closest approximations

$$\int_0^{0.93} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 0.32381 \qquad \int_0^{0.94} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 0.32639$$

To improve the approximation we can apply the linear interpolation scheme below:

$(1-\alpha)/2$	$z_{[1-\frac{\alpha}{2}]}$
0.32381	0.93
0.325	$z_{[0.825]}$
0.32639	0.94

$$\frac{0.325 - 0.32318}{0.32639 - 0.32381} = \frac{z_{[0.825]} - 0.93}{0.94 - 0.93}$$

which gives the critical value estimate

$$z_{[0.825]} = 0.93461240$$

in good agreement with the "rigorous" value calculated by Excel. We have then the absolute error on the mean

$$0.88543775 \cdot z_{[1-\frac{\alpha}{2}]} = 0.88543775 \cdot z_{[0.825]} = 0.88543775 \cdot 0.93458929 = 0.827521$$

and the confidence interval for the mean μ of the resistance gets:

$$\mu \,=\, (495.0 \pm 0.8)\,\Omega\,.$$

An alternative form is, of course,

$$494.2\,\Omega \,\leq\, \mu \,\leq\, 405.8\,\Omega\,.$$

(b) *Confidence level* 90%
In this case we have $\alpha = 1 - 0.90 = 0.10$ and therefore

$$1 - \frac{\alpha}{2} \,=\, 1 - \frac{0.10}{2} \,=\, 1 - 0.05 \,=\, 0.95\,.$$

The Excel function NORMINV provides the corresponding critical value $z_{[0.95]}$:

$$\mathrm{NORMINV}(0,95;0;1) \qquad \Longrightarrow \qquad z_{[0.95]} \,=\, 1.644853627$$

so that

$$0.88543775 \cdot z_{[1-\frac{\alpha}{2}]} \,=\, 0.88543775 \cdot z_{[0.95]} \,=\, 0.88543775 \cdot 1.644853627 \,=\, 1.456415$$

and the confidence interval for the mean takes the form

$$\mu \,=\, (495.0 \pm 1.5)\,\Omega$$

i.e.

$$493.5\,\Omega \,\leq\, \mu \,\leq\, 496.5\,\Omega\,.$$

As before, a satisfactory approximation can be obtained by means of the equation (0.1):

$$\int_0^{z_{[0.95]}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \,=\, \frac{1}{2} - \frac{0.10}{2} \,=\, 0.5 - 0.05 \,=\, 0.45$$

and looking for the value 0.45 among the entries of the cumulative standard normal distribution. The closest approximations are:

$$\int_0^{1.64} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \,=\, 0.44950 \qquad\qquad \int_0^{1.65} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \,=\, 0.45053$$

which allow us to introduce the following interpolation scheme:

$(1-\alpha)/2$	$z_{[1-\frac{\alpha}{2}]}$
0.44950	1.64
0.45	$z_{[0.95]}$
0.45053	1.65

$$\frac{0.45 - 0.44950}{0.45053 - 0.44950} \,=\, \frac{z_{[0.95]} - 1.64}{1.65 - 1.64}$$

and provide the critical value estimate

$$z_{[0.95]} = 1.64485437$$

in excellent agreement with the "rigorous" value calculated by Excel.
The confidence interval with confidence level of 90% is larger than that at confidence level of 60%, as obviously expected.

Solution to Exercise 4
It is pretty reasonable to suppose that the data of electric conductivity σ belong to a normal population, since the sample histogram appears bell-shaped. To check the conjecture we can apply the $\mathcal{X}^2$ test, since all the empirical frequencies in the histogram are sufficiently high ($f_i \geq 3$). Formally, we must test the null hypothesis

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

versus the alternative

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

The sample data were used to estimate the mean and the standard deviation of the distribution, respectively:

$$\overline{\sigma} = m = 2.00 \cdot 10^7\,\mathrm{S}\cdot\mathrm{m}^{-1} \qquad\qquad s = 0.16\cdot 10^7\,\mathrm{S}\cdot\mathrm{m}^{-1}$$

and the classes of the results (the histogram intervals) are $h = 9$ in all. Due to the $c = 2$ constraints on the mean and the standard deviation, estimated by using the same data of the sample, if H_0 holds true the $\mathcal{X}^2$ of data follows approximately a $\mathcal{X}^2$ distribution with

$$n = h - c - 1 = 9 - 2 - 1 = 6$$

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 best way to carry out the calculation is to *standardize* the normal distribution by means of the transformation

$$z = \frac{\sigma - m}{s}$$

which defines a standard normal random variable z and introduce the following correspondence among the values of σ and those of z:

$\sigma \cdot 10^{-7}$	z
1.72	−1.75
1.80	−1.25
1.88	−0.75
1.96	−0.25
2.04	0.25
2.12	0.75
2.20	1.25
2.28	1.75

The theoretical frequencies can be now derived directly from the cumulative distribution of a standard normal random variable. For simplicity's sake, it si convenient to pose

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

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 available on the statistical table of the standard normal distribution. Denoted with p_i, n_i and f_i the probability, the theoretical frequency and the empirical frequency in the i-th class, respectively, we can fill in the table below:

i	p_i	$n_i = 280 \cdot p_i$	f_i	$f_i - n_i$	$(f_i - n_i)^2/n_i$
1	$\Phi(+\infty) - \Phi(1.75) = 0.04006$	11.216564	9	-2.216564	0.438026802
2	$\Phi(1.75) - \Phi(1.25) = 0.06559$	18.365373	17	-1.365373	0.101508565
3	$\Phi(1.25) - \Phi(0.75) = 0.12098$	33.873722	36	2.126278	0.133468001
4	$\Phi(0.75) - \Phi(0.25) = 0.17466$	48.906570	48	-0.906570	0.016804888
5	$\Phi(0.25) + \Phi(0.25) = 0.19742$	55.275542	58	2.724458	0.134284875
6	$\Phi(0.75) - \Phi(0.25) = 0.17466$	48.906570	51	2.093430	0.089608585
7	$\Phi(1.25) - \Phi(0.75) = 0.12098$	33.873722	33	-0.873722	0.022536354
8	$\Phi(1.75) - \Phi(1.25) = 0.06559$	18.365373	18	-0.365373	0.007268963
9	$\Phi(+\infty) - \Phi(1.75) = 0.04006$	11.216564	10	-1.216564	0.131950193

recalling that $\Phi(z)$ is an even function — i.e. $\Phi(-z) = \Phi(z)\ \forall z \in \mathbb{R}$. The sum of all the entries in the last column provides the $\mathcal{X}^2$ of the sample:

$$\mathcal{X}^2 = \sum_{i=1}^{9} \frac{(f_i - n_i)^2}{n_i} = 1.075457226\,.$$

On the table of the $\mathcal{X}^2$ cumulative probability distribution we read the critical value:

$$\mathcal{X}^2{}_{[1-\alpha](6)} = \mathcal{X}^2{}_{[0.95](6)} = 12.592 \qquad \text{for } \alpha = 0.05\,,$$

while for $\alpha = 0.02$ the critical value ${\mathcal{X}^2}_{[1-\alpha](6)} = {\mathcal{X}^2}_{[0.98](6)}$ is not directly available and must be estimated, for instance by the linear interpolation scheme below:

$\alpha = 0.01$	${\mathcal{X}^2}_{[0.99](6)} = 16.812$
$\alpha = 0.02$	${\mathcal{X}^2}_{[0.98](6)}$
$\alpha = 0.025$	${\mathcal{X}^2}_{[0.975](6)} = 14.449$

$$\frac{0.02-0.01}{0.025-0.01} = \frac{{\mathcal{X}^2}_{[0.98](6)} - 16.812}{14.449 - 16.812}$$

which provides ${\mathcal{X}^2}_{[0.98](6)} = 15.237$. More accurate results can be obtained by the Excel function CHIINV:

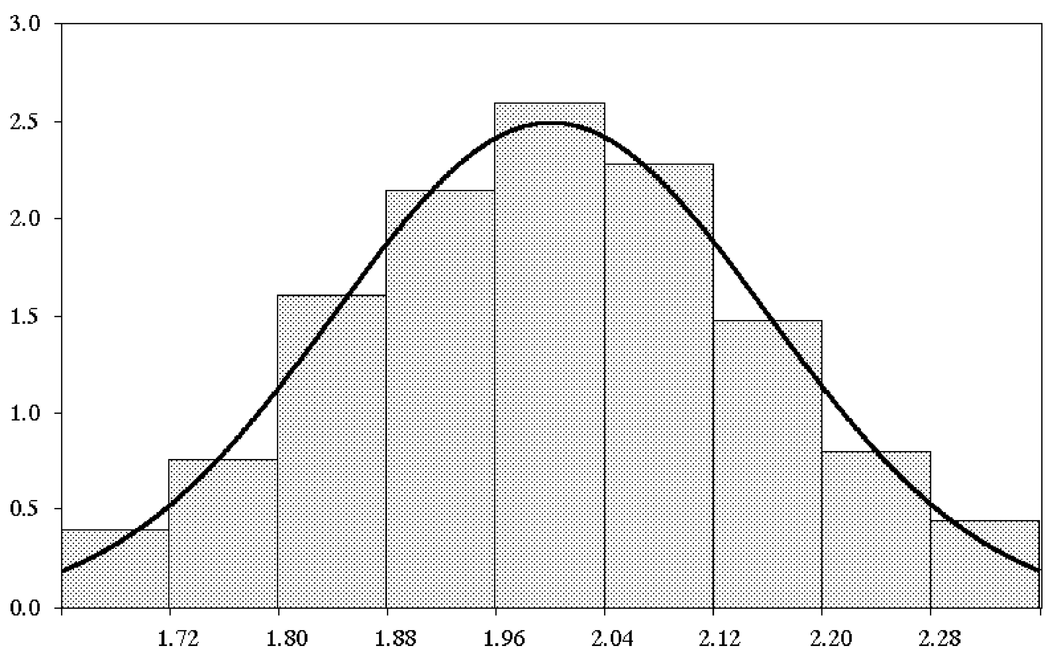
$$\text{CHIINV}(0,05;6) \qquad \Longrightarrow \qquad {\mathcal{X}^2}_{[0.95](6)} = 12.59158724$$

$$\text{CHIINV}(0,02;6) \qquad \Longrightarrow \qquad {\mathcal{X}^2}_{[0.98](6)} = 15.03320775\,.$$

In both cases the $\mathcal{X}^2$ of the sample is smaller than the critical values and we conclude that, with both the significance levels of 5 and 2%, *the null hypothesis cannot be rejected*. The sample data suggest that *the distribution of the electric conductivity of the metallic alloy is presumably normal*. Such a formal conclusion is supported by the good overlap between the theoretical distribution of the data:

$$p(10^{-7}\sigma) = \frac{1}{10^{-7}}\frac{1}{\sqrt{2\pi}\,s}e^{-(10^{-7}\sigma-10^{-7}m)^2/2(10^{-7}s)^2} = \frac{1}{\sqrt{2\pi}\,0.16}e^{-(10^{-7}\sigma-2.00)^2/2\cdot 0.16^2}$$

and the binned distribution obtained from the histogram, as shown in the figure below:



Pay attention that the binned distribution is a piecewise constant function that along the i-th bin takes a constant value $f_i/280/0.08$, where the number of data 280 is introduced to normalize the distribution to 1, while the factor 0.08 is the width of the bin — in this case all the bins have the same width. Thanks to this definition, the area beneath the binned distribution has the meaning of a probability.

Remark. Alternative calculation of the probabilities p_i
The calculation of the theoretical probabilities p_i can also be performed by using the standard normal cumulative probability distribution

$$P(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{\xi^2/2} \, d\xi$$

which is implemented in Excel by the function NORMDIST:

$$P(z) \qquad \leftrightarrow \qquad \mathrm{NORMDIST}(z;\mu;\sigma;\mathrm{TRUE})$$

by posing $\mu = 0$ and $\sigma = 1$. We obtain then the table below:

i	p_i
1	$P(-1.75) - P(-\infty) = 0.040059 - 0.000000 = 0.040059$
2	$P(-1.25) - P(-1.75) = 0.105650 - 0.040059 = 0.065591$
3	$P(-0.75) - P(-1.25) = 0.226627 - 0.105650 = 0.120978$
4	$P(-0.25) - P(-0.75) = 0.401294 - 0.226627 = 0.174666$
5	$P(0.25) - P(-0.25) = 0.598706 - 0.401294 = 0.197413$
6	$P(0.75) - P(0.25) = 0.773373 - 0.598706 = 0.174666$
7	$P(1.25) - P(0.75) = 0.894350 - 0.773373 = 0.120978$
8	$P(1.75) - P(1.25) = 0.959941 - 0.894350 = 0.065591$
9	$P(+\infty) - P(1.75) = 1.000000 - 0.959941 = 0.040059$

whose entries coincide with those derived from the statistical table.

Solution to Exercise 5
The sample consists of $n = 18$ data and cannot be considered large, since $n < 30$. It is then necessary to determine the correct confidence interval for the mean by using the hypothesis of the normal population. Analogously, the sample variance s^2 cannot be regarded as likely equal to the variance σ^2 of the population, as prescribed by the weak law of large numbers

(Kintchine's theorem) for large samples: an appropriate confidence interval is needed also for σ^2. The calculation of the sample mean is straightforward:

$$\overline{v} = \frac{1}{n}\sum_{i=1}^{n} v_i = \frac{1}{18}\sum_{i=1}^{18} v_i = 1.3642889\,.$$

We can then determine the residuals of the data with respect to the mean and the relative squares, as illustrated in the following table:

i	v_i	$(v_i - \overline{v}) \cdot 10^2$	$(v_i - \overline{v})^2 \cdot 10^4$
1	1.3770	1.2711111	1.6157235
2	1.3724	0.8111111	0.6579012
3	1.3570	−0.7288889	0.5312790
4	1.3047	−5.9588889	35.5083568
5	1.3585	−0.5788889	0.3351123
6	1.3766	1.2311111	1.5156346
7	1.3550	−0.9288889	0.8628346
8	1.3550	−0.9288889	0.8628346
9	1.3666	0.2311111	0.0534123
10	1.3663	0.2011111	0.0404457
11	1.3924	2.8111111	7.9023457
12	1.3625	−0.1788889	0.0320012
13	1.3551	−0.9188889	0.8443568
14	1.3845	2.0211111	4.0848901
15	1.3406	−2.3688889	5.6116346
16	1.3534	−1.0888889	1.1856790
17	1.3970	3.2711111	10.7001679
18	1.3826	1.8311111	3.3529679

from which we deduce the sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(v_i - \overline{v})^2 = \frac{1}{17}\sum_{i=1}^{18}(v_i - \overline{v})^2 = 4.4527987 \cdot 10^{-4}$$

and the sample estimate of the standard deviation:

$$s = \sqrt{s^2} = 2.1101656 \cdot 10^{-2}\,.$$

We have thus calculated the basic quantities to determine the confidence intervals of the mean and standard deviation, at the confidence level $1 - \alpha = 0.90$.

(a) Confidence interval for the mean
The CI of the mean, with confidence level $1-\alpha$, takes the form

$$\overline{v} - t_{[1-\frac{\alpha}{2}](n-1)} \frac{s}{\sqrt{n}} \le \mu \le \overline{v} + t_{[1-\frac{\alpha}{2}](n-1)} \frac{s}{\sqrt{n}}$$

and for $\alpha = 0.10$, $n = 18$ has thus the limits

$$\overline{v} - t_{[0.95](17)} \frac{s}{\sqrt{18}} = 1.3642889 - 1.740 \cdot \frac{0.021101656}{\sqrt{18}} = 1.35564$$

$$\overline{v} + t_{[0.95](17)} \frac{s}{\sqrt{18}} = 1.3642889 + 1.740 \cdot \frac{0.021101656}{\sqrt{18}} = 1.37466$$

so that the confidence interval becomes

$$1.35564\,\mathrm{V} \le \mu \le 1.37466\,\mathrm{V}$$

or, equivalently,

$$\mu = (1.36515 \pm 0.00951)\,\mathrm{V}\,.$$

As a matter of fact, such a large number of digits in not meaningful and for all practical purposes an approximation of the form

$$\mu = (1.365 \pm 0.010)\,\mathrm{V}$$

can be regarded as more than satisfactory. *Remember that*, besides looking at the statistical table, the critical value $t_{[0.95](17)}$ can also be calculated by the Excel function TINV, as follows

$$\mathrm{TINV}(0,10;17) \qquad \Longrightarrow \qquad 1.739607\,.$$

(b) Confidence interval for the standard deviation
The CI of the variance is given by

$$\frac{1}{\mathcal{X}^2{}_{[1-\frac{\alpha}{2}](n-1)}}(n-1)s^2 \le \sigma^2 \le \frac{1}{\mathcal{X}^2{}_{[\frac{\alpha}{2}](n-1)}}(n-1)s^2$$

still with $\alpha = 0.10$ and $n = 18$. Therefore the lower and the upper limits are:

$$\frac{1}{\mathcal{X}^2{}_{[0.95](17)}} 17\,s^2 = \frac{1}{27.587}\, 17 \cdot 4.4527987 \cdot 10^{-4} = 2.7439472 \cdot 10^{-4}$$

$$\frac{1}{\mathcal{X}^2{}_{[0.05](17)}} 17\,s^2 = \frac{1}{8.672}\, 20 \cdot 4.4527987 \cdot 10^{-4} = 8.7292055 \cdot 10^{-4}$$

and the CI of the variance becomes

$$2.7439472 \cdot 10^{-4}\,\mathrm{V}^2 \le \sigma^2 \le 8.7292055 \cdot 10^{-4}\,\mathrm{V}^2\,.$$

As before, such a huge number of digits is not meningful and must be adequately shortened:

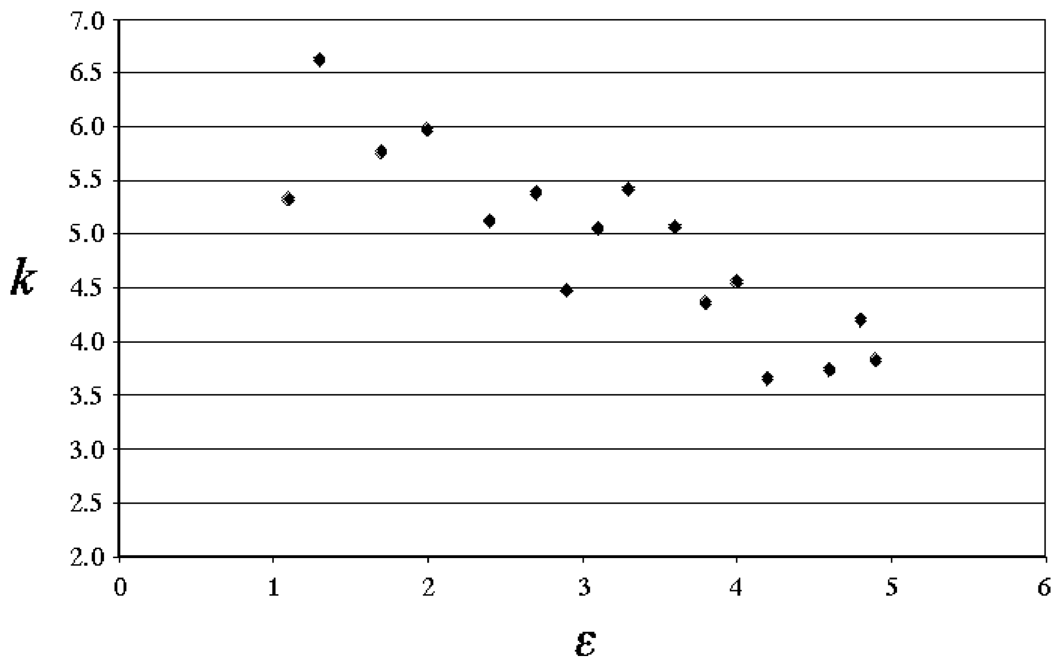
$$2.7 \cdot 10^{-4}\,\mathrm{V} \le \sigma \le 8.7 \cdot 10^{-4}\,\mathrm{V}\,.$$

The critical values of the $\mathcal{X}^2$ variable con be also reckoned by the Excel function CHIINV:

$$\mathrm{CHIINV}(0,05;17) \qquad \Longrightarrow \qquad {\mathcal{X}^2}_{[0.95](17)} = 27.587112$$

$$\mathrm{CHIINV}(0,95;17) \qquad \Longrightarrow \qquad {\mathcal{X}^2}_{[0.05](17)} = 8.671760\,.$$

Solution to Exercise 6
The data plot suggests that the porosity ε and the thermal conductivity k are described by dependent random variables (bivariate normal random variables are stochastically dependent if and only if they are correlated):



since the datapoints appear to be partially aligned along a straight line of negative slope. The sample means $\overline{\varepsilon}$ and $\overline{k}$ of the two quantities are given by:

$$\overline{\varepsilon} = \frac{1}{16}\sum_{i=1}^{16}\varepsilon_i = 3.1500 \qquad\qquad \overline{k} = \frac{1}{16}\sum_{i=1}^{16}k_i = 4.918750$$

and allow us to calculate the sum of products of residuals

$$SS_{\varepsilon k} = \sum_{i=1}^{16}(\varepsilon_i - \overline{\varepsilon})(k_i - \overline{k}) = -13.332000$$

and those of the relative squares:

$$SS_{\varepsilon\varepsilon} = \sum_{i=1}^{16}(\varepsilon_i - \overline{\varepsilon})^2 = 22.240000$$

$$SS_{kk} = \sum_{i=1}^{16}(k_i - \overline{k})^2 = 10.753775\,,$$

as illustrated in the following table:

ε_i	k_i	$\Delta\varepsilon_i$	Δk_i	$\Delta\varepsilon_i^2$	Δk_i^2	$\Delta\varepsilon_i\Delta k_i$
1.1	5.34	−2.050000	0.421250	4.202500	0.177452	−0.863563
1.3	6.63	−1.850000	1.711250	3.422500	2.928377	−3.165813
1.7	5.77	−1.450000	0.851250	2.102500	0.724627	−1.234313
2.0	5.98	−1.150000	1.061250	1.322500	1.126252	−1.220438
2.4	5.13	−0.750000	0.211250	0.562500	0.044627	−0.158438
2.7	5.39	−0.450000	0.471250	0.202500	0.222077	−0.212063
2.9	4.49	−0.250000	−0.428750	0.062500	0.183827	0.107188
3.1	5.06	−0.050000	0.141250	0.002500	0.019952	−0.007062
3.3	5.43	0.150000	0.511250	0.022500	0.261377	0.076688
3.6	5.08	0.450000	0.161250	0.202500	0.026002	0.072563
3.8	4.37	0.650000	−0.548750	0.422500	0.301127	−0.356687
4.0	4.57	0.850000	−0.348750	0.722500	0.121627	−0.296437
4.2	3.66	1.050000	−1.258750	1.102500	1.584452	−1.321688
4.6	3.75	1.450000	−1.168750	2.102500	1.365977	−1.694688
4.8	4.21	1.650000	−0.708750	2.722500	0.502327	−1.169438
4.9	3.84	1.750000	−1.078750	3.062500	1.163702	−1.887813

where, for simplicity's sake, we have posed $\varepsilon_i - \overline{\varepsilon} = \Delta\varepsilon_i$ and $k_i - \overline{k} = \Delta k_i$. The linear correlation coefficient becomes

$$r = \frac{SS_{\varepsilon k}}{\sqrt{SS_{\varepsilon\varepsilon}}\sqrt{SS_{kk}}} = \frac{-13.332000}{\sqrt{22.240000}\sqrt{10.753775}} = -0.862080\,.$$

Since the bivariate random variable (ε, k) may be assumed to be normal, we can check the null hypothesis

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

against the alternative hypothesis

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

by means of the random variable

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

that, if H_0 holds true, follows a Student's distribution with $n-2$ d.o.f. In the present case we have $n = 16$ and obtain:

$$\boxed{t \;=\; \sqrt{16-2}\,\frac{-0.862080}{\sqrt{1-(-0.862080)^2}} \;=\; -6.364998\,.}$$

At a significance level α the critical region of the test takes the form

$$\big\{t \in \mathbb{R} \;:\; |t| \,>\, t_{[1-\frac{\alpha}{2}](n-2)}\big\} = \big\{t \in \mathbb{R} \;:\; |t| \,>\, t_{[1-\frac{\alpha}{2}](14)}\big\}\,.$$

(*a*) *Significance level* $\alpha = 10\%$

In this case the critical value of the test statistic is

$$\boxed{t_{[1-\frac{\alpha}{2}](14)} \;=\; t_{[0.95](14)} \;=\; 1.761}$$

and can be easily calculated by using the table of the Student's t cumulative distribution or the Excel function TINV:

$$\mathrm{TINV}(0,10;14) \qquad\Longrightarrow\qquad t_{[0.95](14)} \;=\; 1.761310115$$

We conclude that the value of the test statistic does belong to the rejection region, so that H_0 *must be rejected*. The random variables ε and k can be considered stochastically dependent (or correlated).

(*b*) *Significance level* $\alpha = 5\%$

For this significance level the critical value of the test statistic becomes

$$\boxed{t_{[1-\frac{\alpha}{2}](14)} \;=\; t_{[0.975](14)} \;=\; 2.145\,.}$$

A more accurate result can be obtained, as before, by using the Excel function TINV:

$$\mathrm{TINV}(0,05;14) \qquad\Longrightarrow\qquad t_{[0.975](14)} \;=\; 2.144786681\,.$$

The value of the test statistic is still outside the acceptance region. Therefore the conclusion is the same as before.

(c) *Significance level* $\alpha = 2\%$
The critical value of the test statistic is now

$$\boxed{t_{[1-\frac{\alpha}{2}](14)} \,=\, t_{[0.99](12)} \,=\, 2.624}$$

and can be more accurately calculated by the TINV function:

$$\mathrm{TINV}(0,02;14) \qquad \Longrightarrow \qquad t_{[0.99](14)} \,=\, 2.624494064\,.$$

Again, since the value -6.364998 of the test statistic does not fall within the acceptance region for H_0, we must conclude that *the random variables* ε *and* k *are probably stochastically dependent*. Due to the negative sign of the correlation coefficient, which is very close to -1, the relation should be inverse.

Solution to Exercise 7
It seems quite natural to apply a paired t-test for the comparison of the means, because Young's modulus E is measured prior to and after the treatment *on each sample*. Therefore, for all the $n = 12$ samples the values relative to the same sample will be coupled:

$$(y_i, z_i) \qquad i \,=\, 1, \ldots, n\,,$$

on having denoted with y_i and z_i the Young's modulus measured before and after the treatment, respectively. If μ_1 and μ_2 denote the mean (true) value of E before and after the treatment, we must check the hypothesis $H_0 : \mu_1 = \mu_2$, that the treatment has no effect on the mean value of E, versus the alternative hypothesis$H_1 : \mu_1 \neq \mu_2$ that the claim is false. The test variable writes

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

which, for H_0 true, follows a Student's distribution with $n - 1 = 11$ d.o.f. The test defines a critical region, with a significance level α, of the form

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

and for $n = 12$, $\alpha = 0.02$ becomes

$$\big\{t \leq -2.718\big\} \cup \big\{t \geq 2.718\big\}$$

since the table of the Student's t cumulative distribution provides the critical value

$$\boxed{t_{[1-\frac{\alpha}{2}](n-1)} \,=\, t_{[0.99](11)} \,=\, 2.718}$$

for which a more accurate estimate is also available

$$t_{[0.99](11)} = 2.718079$$

by using the Excel function TINV(0, 02; 11). In the present case the sample means of the two samples hold:

$$\overline{y} = \frac{1}{12}\sum_{i=1}^{12} y_i = 1.70817 \qquad \overline{z} = \frac{1}{12}\sum_{i=1}^{12} z_i = 1.94025$$

whereas

$$\frac{1}{n-1}\sum_{i=1}^{n}(y_i - z_i - \overline{y} + \overline{z})^2 = \frac{1}{11}\sum_{i=1}^{12}(y_i - z_i - \overline{y} + \overline{z})^2 = 0.03415772$$

and therefore

$$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - z_i - \overline{y} + \overline{z})^2} = \sqrt{0.03415772} = 0.1848181$$

so that the test variable takes the value

$$t = \sqrt{12}\cdot\frac{1.70817 - 1.94025}{0.1848181} = -4.3500\,.$$

The table below shows the detailed calculations:

y_i	z_i	$y_i - z_i$	$y_i - z_i - \overline{y} + \overline{z}$	$(y_i - z_i - \overline{y} + \overline{z})^2$
1.693	1.931	−0.238	−0.00592	0.000035
1.820	1.864	−0.044	0.18808	0.035375
1.543	1.689	−0.146	0.08608	0.007410
1.780	2.010	−0.230	0.00208	0.000004
1.735	2.299	−0.564	−0.33192	0.110169
1.862	2.105	−0.243	−0.01092	0.000119
1.581	2.087	−0.506	−0.27392	0.075030
1.807	1.759	0.048	0.28008	0.078447
1.651	1.927	−0.276	−0.04392	0.001929
1.632	1.627	0.005	0.23708	0.056209
1.647	1.904	−0.257	−0.02492	0.000621
1.747	2.081	−0.334	−0.10192	0.010387

The calculated value of t belongs to the rejection region, because

$$-4.3500 < -2.718\,,$$

and therefore *we can exclude*, with a significance level of 2%, *that the true value of the Young's modulus is the same* before and after the treatment. *The null hypothesis is rejected.*

Solution to Exercise 8

No relevant random errors affect the temperature data, while the values of dynamic viscosity can be regarded as the outcomes of independent normal random variables. The standard theory of linear regression is then applicable, with a further simplification due to the homoskedastic nature of the model — it can be assumed that all the viscosity data share the same variance. The regression straight line must provide the dynamic viscosity μ as a function of the temperature T:

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

where $\overline{T}$ stands for the arithmetic mean of the measured temperatures, while α and β denote the parameters of the regression model. As well known, such a kind of model ensures the stochastic independence of the best-fit estimates a and b to the regression parameters α and β.

Notice that the sample consists in multiple measurements at each temperature: many measurements of viscosity have been performed for each sampled value of T. This circumstance does not constitute a hindrance to the application of the standard linear regression model, provided that all the pairs (T_i, μ_i) with the same T are treated as distinct. According to this criterion the whole number of sample data is thus $n = 32$.

(*a*) *Regression straight line*

Since the standard deviations are equal, the $\mathcal{X}^2$ fitting reduces to the usual least-squares fitting and the best-fit estimates a and b of the parameters can be expressed as

$$a = \overline{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mu_i = 0.522519 \qquad\qquad b = \frac{\displaystyle\sum_{i=1}^{n}(T_i - \overline{T})\,\mu_i}{\displaystyle\sum_{i=1}^{n}(T_i - \overline{T})^2} = -0.008951786$$

with $n = 32$ and

$$\overline{T} = \frac{1}{n}\sum_{i=1}^{n}T_i = 52.5$$

$$\sum_{i=1}^{n}(T_i - \overline{T})\,\mu_i = -37.5975 \qquad\qquad \sum_{i=1}^{n}(T_i - \overline{T})^2 = 4200.0\,.$$

The regression straight line, determined by the least-squares method, takes therefore the following form:

$$\begin{aligned}\mu = a + b\,(T - \overline{T}) &= 0.522519 - 0.008951786\cdot(T - 52.5) = \\ &= 0.9924875 - 0.008951786\cdot T\end{aligned}$$

where the number of digits is left temporarily large before the appropriate confidence region has been determined. The detailed calculations are shown in the table below, which collects all the single terms involved in the computation of the previous summations:

T_i	μ_i	$T_i - \overline{T}$	$(T_i - \overline{T})^2$	$(T_i - \overline{T})\,\mu_i$
35	0.7355	-17.5	306.25	-12.87125
40	0.5947	-12.5	156.25	-7.43375
45	0.6688	-7.5	56.25	-5.01600
50	0.5320	-2.5	6.25	-1.33000
55	0.4911	2.5	6.25	1.22775
60	0.4718	7.5	56.25	3.53850
65	0.4245	12.5	156.25	5.30625
70	0.4358	17.5	306.25	7.62650
35	0.6913	-17.5	306.25	-12.09775
40	0.6190	-12.5	156.25	-7.73750
45	0.5211	-7.5	56.25	-3.90825
50	0.5709	-2.5	6.25	-1.42725
55	0.5524	2.5	6.25	1.38100
60	0.4667	7.5	56.25	3.50025
65	0.4128	12.5	156.25	5.16000
70	0.3192	17.5	306.25	5.58600
35	0.6021	-17.5	306.25	-10.53675
40	0.4852	-12.5	156.25	-6.06500
45	0.5099	-7.5	56.25	-3.82425
50	0.5163	-2.5	6.25	-1.29075
55	0.5242	2.5	6.25	1.31050
60	0.3972	7.5	56.25	2.97900
65	0.4371	12.5	156.25	5.46375
70	0.3822	17.5	306.25	6.68850
35	0.7890	-17.5	306.25	-13.80750
40	0.7069	-12.5	156.25	-8.83625
45	0.6829	-7.5	56.25	-5.12175
50	0.5625	-2.5	6.25	-1.40625
55	0.4497	2.5	6.25	1.12425
60	0.4270	7.5	56.25	3.20250
65	0.3892	12.5	156.25	4.86500
70	0.3516	17.5	306.25	6.15300

(*b*) *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 ρ_{n-2} stands for 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

$$\text{NSSAR} = \sum_{i=1}^{n} \frac{1}{\sigma^2}[a + b\,(T_i - \overline{T}) - \mu_i]^2 = \frac{\text{SSAR}}{\sigma^2}$$

behaves like a $\mathcal{X}^2$ random variable with $n-2$ d.o.f. In order to evaluate the goodness of fit *it is crucial to know the common value of the standard deviation* $\sigma = 0.06$, since we need to determine the NSSAR, and not simply the SSAR. In the present case we have $n = 32$ data and the regression model is based on the two parameters α and β. Consequently, the NSSAR obeys a $\mathcal{X}^2$ distribution with $n-2 = 30$ d.o.f. For the given sample the normalized sum of squares around regression holds

$$\text{NSSAR} = \frac{\text{SSAR}}{\sigma^2} = \frac{0.0980656}{0.06^2} = 27.24045149\,.$$

On the table of the upper critical values of $\mathcal{X}^2$ with $\nu = 30$ d.o.f. we can only find the two far points

Probability $\{\mathcal{X}^2 \geq 20.599\}$	Probability $\{\mathcal{X}^2 \geq 40.256\}$
0.90	0.10

so that the simple linear interpolation scheme:

20.599	0.90
27.2405	Q
40.256	0.10

$$\frac{27.2405 - 20.599}{40.256 - 20.599} = \frac{Q - 0.90}{0.10 - 0.90}$$

does not probably provide an accurate estimate of Q:

$$Q = 0.90 + (0.10 - 0.90)\,\frac{27.2405 - 20.599}{40.256 - 20.599} = 0.6297\,.$$

A more precise value of Q can be obtained by a numerical integration

$$Q = \text{Probability}\{\mathcal{X}^2 \geq 27.24045149\} = \int_{27.24045149}^{+\infty} p_{30}(\mathcal{X}^2)\,d\mathcal{X}^2$$

for instance by using the Excel function CHIDIST:

$$\text{CHIDIST}(27,24045149;30) \qquad \Longrightarrow \qquad \int_{27.24045149}^{+\infty} p_{30}(\mathcal{X}^2)\,d\mathcal{X}^2 \;=\; 0.610618751\,.$$

Alternatively, we may execute the Maple command line

$$1 - \text{stats[statevalf, cdf, chisquare[30]](27.24045149);}$$

to obtain the "exact" value $Q = 0.6106187067$. The goodness of fit of the regression model is thus equal to about 61%: such a percentage would express the probability of (a type I) error if the regression model were rejected.

(*c*) *Confidence intervals for the regression parameters*

The sum of squares around regression has already been determined:

$$\text{SSAR} \;=\; \sum_{i=1}^{n}[a + b(T_i - \overline{T}) - \mu_i]^2 \;=\; 0.0980656\,.$$

At the confidence level $1-\alpha$, the CI of the parameter α and that of the slope β are given by the formulas:

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

$$\beta \;=\; b \pm t_{[1-\frac{\alpha}{2}](n-2)}\sqrt{\left[\sum_{i=1}^{n}(T_i - \overline{T})^2\right]^{-1}\frac{\text{SSAR}}{n-2}}\,.$$

Here we have $\alpha = 0.10$ and $n = 32$, so that the confidence intervals become:

$$\alpha \;=\; a \pm t_{[0.95](30)}\sqrt{\frac{1}{32}\frac{\text{SSAR}}{30}}$$

$$\beta \;=\; b \pm t_{[0.95](30)}\sqrt{\left[\sum_{i=1}^{32}(T_i - \overline{T})^2\right]^{-1}\frac{\text{SSAR}}{30}}$$

where:

$$\begin{aligned} a &= 0.522519 \\ b &= -0.008951786 \\ \text{SSAR} &= 0.0980656 \\ \sum_{i=1}^{32}(T_i - \overline{T})^2 &= 4200.0 \\ t_{[0.95](30)} &= 1.697 \,. \end{aligned}$$

The latter critical value has been read on the table of the Student's t cumulative distribution, but a more accurate result can be obtained by the Excel function TINV:

$$\text{TINV}(0,10;30) \qquad \Longrightarrow \qquad t_{[0.95](30)} = 1.697260851\,.$$

As a satisfactory compromise between the tabulated and the Excel value we may assume $\mathbf{t_{[0.95](30)} = 1.69726}$. By inserting the numerical values and performing the calculations we deduce that:

– the 90%-CI of the parameter α is

$$\alpha = 0.522519 \pm 1.69726 \sqrt{\frac{1}{32}\frac{0.0980656}{30}}$$

i.e.

$$\alpha = 0.522519 \pm 0.017154 = [0.5053645141, 0.5396729859]$$

– the 90%-CI for the slope β holds

$$\beta = -0.008951786 \pm 1.69726 \sqrt{\frac{1}{4200.0}\frac{0.0980656}{30}}$$

or, equivalently,

$$\beta = -0.008951786 \pm 0.001497344468 = [-0.01044913018, -0.007454441246]\,.$$

Leaving out the redundant digits and introducing the physical units, we conclude that

$$\alpha = [0.50536, 0.53967]\,\text{mPa}\cdot\text{s} = (0.52252 \pm 0.01715)\,\text{mPa}\cdot\text{s}$$

whereas

$$\begin{aligned} \beta &= [-0.010449, -0.007454]\,\text{mPa}\cdot\text{s}\cdot\ {}^o\text{C}^{-1} = \\ &= (-0.008952 \pm 0.001497)\,\text{mPa}\cdot\text{s}\cdot\ {}^o\text{C}^{-1}\,. \end{aligned}$$

(d) Confidence region
It has been assumed that the model is homoskedastic. Consequently, the CI at a confidence level $1-\alpha$ for the prediction of $\mu = \mu_0$ at a given $T = T_0$ is expressed by the general formula:

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

where, more specifically, we have:

$$\begin{aligned}
a \;&=\; 0.522519 \\
b \;&=\; -0.008951786 \\
\overline{T} \;&=\; \frac{1}{n}\sum_{i=1}^{n} T_i \;=\; 52.5 \\
t_{[1-\frac{\alpha}{2}](n-2)} \;&=\; t_{[0.95](30)} \;=\; 1.69726 \\
V \;&=\; 1 + \frac{1}{n} + \frac{1}{\displaystyle\sum_{i=1}^{n}(T_i - \overline{T})^2}(T_0 - \overline{T})^2 \;=\; 1 + \frac{1}{32} + \frac{(T_0 - 52.5)^2}{4200.0} \;= \\
&=\; 1.03125 + 0.0002380952 \cdot (T_0 - 52.5)^2 \\
\text{SSAR} \;&=\; \sum_{i=1}^{n}\big[a + b(T_i - \overline{T}) - \mu_i\big]^2 \;=\; 0.0980656 \;.
\end{aligned}$$

The CI for the prediction of the dynamic viscosity μ at $T = T_0$ is then:

$$\begin{aligned}
\mu_0 \;&=\; 0.522519 - 0.008951786 \cdot (T_0 - 52.5) \pm \\
&\pm\; 1.69726 \cdot \sqrt{1.03125 + 0.0002380952 \cdot (T_0 - 52.5)^2}\sqrt{\frac{0.0980656}{30}}
\end{aligned}$$

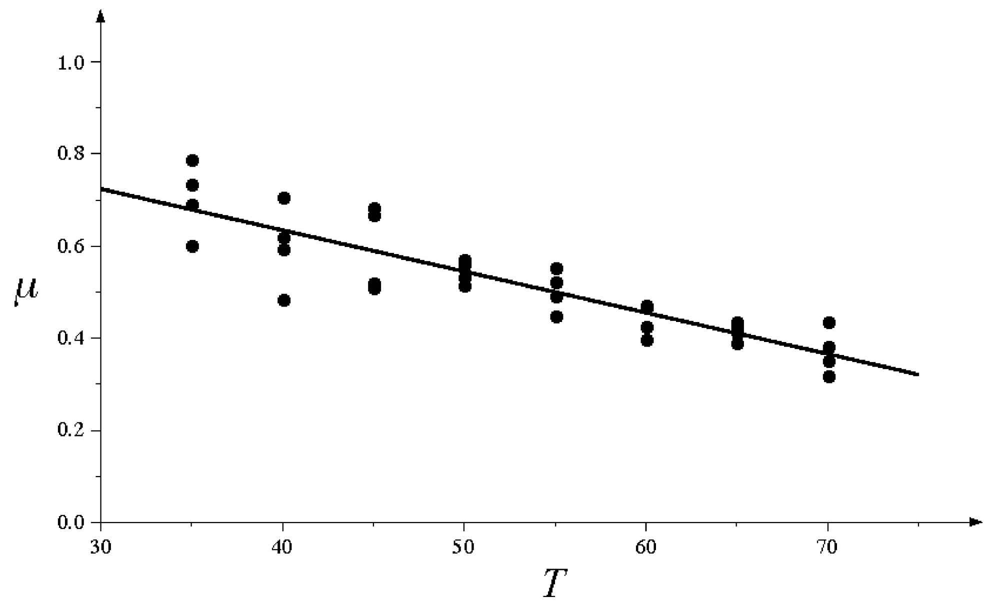
and performing the calculations reduces to:

$$\begin{aligned}
\mu_0 \;&=\; 0.992487500 - 0.0089517857 \cdot T_0 \pm \\
&\pm\; 0.0970390123 \cdot \sqrt{1.03125 + 0.0002380952 \cdot (T_0 - 52.5)^2}
\end{aligned}$$

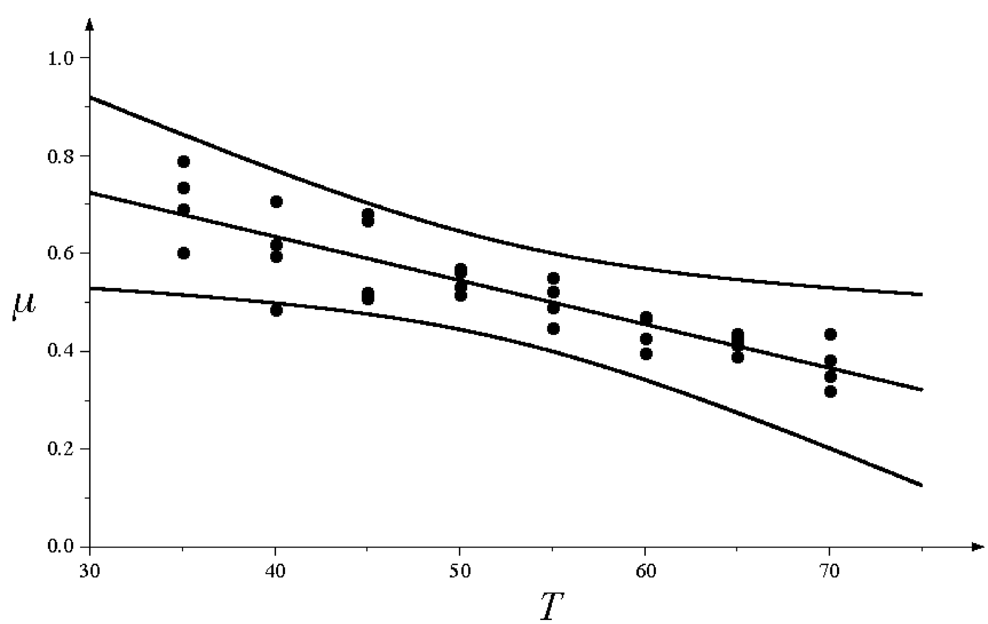
or, more conveniently,

$$\boxed{\begin{aligned}
\mu_0 \;&=\; 0.992487500 - 0.0089517857 \cdot T_0 \pm \\
&\pm\; 0.0970390123 \cdot \sqrt{1.03125 + 2.380952 \cdot \left(\frac{T_0}{100} - 0.525\right)^2}
\end{aligned}}$$

The number of digits in the above formula is certainly excessive, but it costs nothing to carry out the computations by using all the available digits: we must simply remember to round off appropriately the final result, which has a direct physical meaning. To point out the good agreement between the regression model and the data, 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 (by exaggerating the factor V for clarity's sake)



The curves below and over the regression straight line represent the lower and upper limits of the confidence region, respectively. The width of the confidence region, measured parallel to the vertical axis μ, is minimum for $T = \overline{T} = 52.5$ and tends to increase monotonically to the right and to the left of that point. To better stress the effect on the graph, the term $(T_0 - \overline{T})^2$ which appears in the expression of V of the definition has been multiplied by a scale factor 25.

(*e*) *Confidence interval for a prediction*
The CI at a confidence level of 90% for the prediction of γ at $T = 52^oC$ can be obtained by posing $T_0 = 52$ in the previous formula

$$\begin{aligned}\mu_0 \;=&\; 0.992487500 - 0.0089517857 \cdot T_0 \pm \\ &\pm\; 0.0970390123 \cdot \sqrt{1.03125 + 2.380952 \cdot \Big(\frac{T_0}{100} - 0.525\Big)^2}\;.\end{aligned}$$

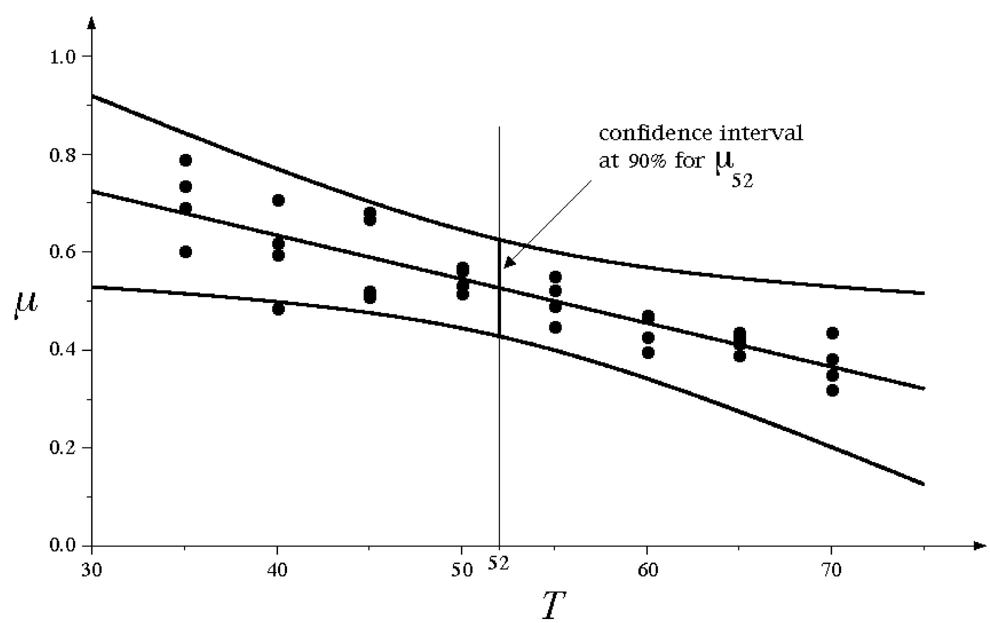
We obtain therefore:

$$\begin{aligned}\mu_0 \;=\; \mu_{52} \;=&\; 0.992487500 - 0.0089517857 \cdot 52 \pm \\ &\pm\; 0.0970390123 \cdot \sqrt{1.03125 + 2.380952 \cdot \Big(\frac{52}{100} - 0.525\Big)^2} \\ =&\; 0.5269946428 \pm 0.0985464268 \;=\; [0.4284482161,\, 0.6255410697]\end{aligned}$$

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

$$\mu_{52} \;=\; [0.43,\, 0.63]\,\mathrm{mPa\cdot s} \;=\; (0.53 \pm 0.10)\,\mathrm{mPa\cdot s}\,.$$

In the following figure the CI at 90% is highlighted as the intersection of the confidence region at 90% with the vertical straight line of equation $T = 52$:



Solution to Exercise 9

Let us denote with μ_1 and μ_2 the mean values of the two samples, that is the "true values" of electrical conductivity for the first and the second treatment, respectively. Let $p = 11$ be the number of data $y_1, \ldots, y_p$ of the first sample and $q = 14$ that of the data $z_1, \ldots, z_q$ of the second sample.

We want to test the hypothesis $H_0 : \mu_1 = \mu_2$ (the two treatments do not modify significantly the electrical conductivity of the material) against the alternative hypothesis $H_1 : \mu_1 \neq \mu_2$ (the two treatments yield materials with a different electrical conductivity). The significance level we choose is 10%.

Checking whether the variances are or are not equal

We can check whether the two normal populations share or do not share the same variance by using the F-test. The test statistic is the ratio of the sample estimates of variances:

$$F = \frac{s_y^2}{s_z^2},$$

with

$$\overline{y} = \frac{1}{p}\sum_{i=1}^{p} y_i = 17.090909 \qquad \overline{z} = \frac{1}{q}\sum_{j=1}^{q} z_j = 14.285714$$

$$s_y^2 = \frac{1}{p-1}\sum_{i=1}^{p}(y_i - \overline{y})^2 = 11.510909 \qquad s_z^2 = \frac{1}{q-1}\sum_{j=1}^{q}(z_j - \overline{z})^2 = 7.345934$$

according to the detailed calculation shown in the table below:

y_i	$y_i - \overline{y}$	$(y_i - \overline{y})^2$	z_i	$z_i - \overline{z}$	$(z_i - \overline{z})^2$
21.7	4.609091	21.243719	15.4	1.114286	1.241633
12.5	−4.590909	21.076446	14.4	0.114286	0.013061
12.6	−4.490909	20.168264	15.9	1.614286	2.605918
18.6	1.509091	2.277355	13.1	−1.185714	1.405918
19.4	2.309091	5.331901	11.5	−2.785714	7.760204
15.9	−1.190909	1.418264	10.0	−4.285714	18.367347
17.5	0.409091	0.167355	16.4	2.114286	4.470204
15.4	−1.690909	2.859174	12.0	−2.285714	5.224490
22.2	5.109091	26.102810	10.5	−3.785714	14.331633
18.6	1.509091	2.277355	15.4	1.114286	1.241633
13.6	−3.490909	12.186446	12.5	−1.785714	3.188776
			17.5	3.214286	10.331633
			18.7	4.414286	19.485918
			16.7	2.414286	5.828776

We obtain therefore:

$$F = \frac{11.510909}{7.345934} = 1.566977\,.$$

The F-test prescribes that the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ that the two populations have the same variance is accepted, at a significance level α, if

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

In the present case we have $p = 11$, $q = 14$ and $\alpha = 0.10$, so that the acceptance condition becomes

$$0.346359 = F_{[0.05](10,13)} < F < F_{[0.95](10,13)} = 2.671024$$

as derived from the Excel function FINV:

$$\text{FINV}(0,95;10;13) \qquad \Longrightarrow \qquad F_{[0.05](10,13)} = 0.346359$$

$$\text{FINV}(0,05;10;13) \qquad \Longrightarrow \qquad F_{[0.95](10,13)} = 2.671024$$

or, with a lesser accuracy, by using the table of the Fisher cumulative distributions — $F_{[0.05](10,13)} = 0.3464$ and $F_{[0.95](10,13)} = 2.6710$. Since the value of the test statistic actually falls within the acceptance region:

$$0.346359 < 1.566977 < 2.671024$$

we conclude that *the variances σ_1^2 and σ_2^2 can be considered as equal.*

T-test for the comparison of the means
By hypothesis the populations can be assumed to be normal. Moreover, we have checked that the variances of the two populations are probably the same. The test variable is then

$$t = \frac{\overline{y} - \overline{z}}{s\sqrt{\dfrac{1}{p} + \dfrac{1}{q}}}$$

where s^2 denotes the pooled variance of the two samples:

$$s^2 = \frac{(p-1)s_y^2 + (q-1)s_z^2}{p+q-2} .$$

When $\mu_1 = \mu_2$ the random variable is known to follow a Student's t distribution with $p + q - 2$ d.o.f. The null hypothesis $H_0 : \mu_1 = \mu_2$ will be rejected if the value of t calculated on the sample belongs to the two-sided critical region

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

In this case we have the pooled variance

$$\frac{10 \cdot 11.510909 + 13 \cdot 7.345934}{11 + 14 - 2} = 9.156792772$$

and the value of the test statistic is

$$t = \frac{\overline{y} - \overline{z}}{s\sqrt{\dfrac{1}{p} + \dfrac{1}{q}}} = \frac{17.090909 - 14.285714}{\sqrt{9.156792772}\sqrt{\dfrac{1}{11} + \dfrac{1}{14}}} = 2.300815$$

whereas

$$\mathrm{TINV}(0,10;23) \qquad \Longrightarrow \qquad t_{[1-\frac{\alpha}{2}](p+q-2)} = t_{[0.95](23)} = 1.713872$$

so that the critical region takes the form

$$\Big\{t < -1.713872\Big\} \cup \Big\{t > 1.713872\Big\}.$$

Clearly, the value of t falls within the upper tail of the critical region and *the null hypothesis must be rejected* at the significance level of 10%. We conclude that *the different temperatures of the thermal treatment probably have an effect* on the electric conductivity of the alloy.