Graduate School of Materials Engineering - a.y. 2009-2010
Methods of statistical and numerical analysis (integrated course). Part I
Final test - March 12th 2010

☐ Exercise 1 **Points 4**

Repeated measurements of the electrical conductivity $\varkappa$ of a H_2SO_4 solution in water have led to the results listed below (in $\Omega^{-1} \cdot \mathrm{cm}^{-1}$):

0.4329	0.4584	0.5212	0.5060	0.4633	0.6045	0.5438
0.5237	0.5342	0.4702	0.5102	0.5223	0.4575	0.4681

which are assumed to follow a normal distribution. There is the suspect that the experimental device underwent a possible malfunction during the measurements. Check for the eventual presence of an outlier not belonging to the statistical population of the measurements.

☐ Exercise 2 **Points 3**

The resonance frequency of a LC-circuit is given by the formula

$$f = \frac{1}{2\pi} \frac{1}{\sqrt{LC}}$$

with inductance $L = (15.0 \pm 0.3) \cdot 10^{-4}$ H and capacitance $(20.5 \pm 0.1) \cdot 10^{-6}$ F. Determine the estimated value of f along with the appropriate absolute error. Would it be possible to obtain a precision within 1% by improving the measurement of *only* a quantity C or L? If yes, how?

☐ Exercise 3 **Points 4**

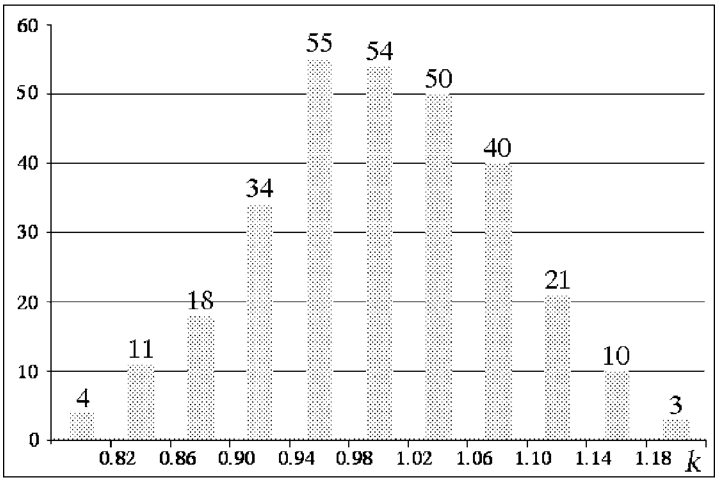
A random sample of 300 screws produced by an automatic machine is collected. The mean weight of the screws turns out to be 18.45 g, with a standard deviation of 0.60 g. Calculate:

(*a*) the confidence interval for the weight of a screw at a confidence level of 70%;

(*b*) the confidence interval for the weight of a screw at a confidence level of 95%.

☐ Exercise 4 Points 5

The data of the thermal conductivity k of a ceramic are believed to follow a normal distribution. To check the conjecture 300 measurements of thermal conductivity have been carried out and the relative sample mean $\overline{k} = 1.00$ and standard deviation $s = 0.08$ calculated (in arbitrary units). The histogram below summarizes the data:



Check the hypothesis of the normal distribution with a significance level (a) of 5% and (b) of 1%.

☐ Exercise 5 Points 4

The electromotive force v between the poles of a photovoltaic cell has been repeatedly measured, yielding the values below (in V):

i	v_i	i	v_i	i	v_i
1	1.4585	8	1.4924	15	1.4970
2	1.4570	9	1.4666	16	1.4406
3	1.4766	10	1.4826	17	1.4625
4	1.4551	11	1.4550	18	1.4770
5	1.4534	12	1.4663	19	1.5047
6	1.4724	13	1.4550	20	1.4845
7	1.4887	14	1.4719	21	1.4834

which can be assumed to be normal. Determine the confidence interval of the mean μ and that of the standard deviation σ, both at the confidence level of 95%.

☐ Exercise 6 **Points 4**

A sintering process yields ceramic samples whose porosity ε and toughness K_{Ic} vary in an unpredictable way according to a bivariate normal probability distribution. To check if the two quantities may be correlated, 14 measurements of ε and K_{Ic} have been performed on as many samples. The results are listed in the following table:

i	ε_i	$K_{Ic\,i}$	i	ε_i	$K_{Ic\,i}$
1	1.1	4.35	8	3.3	4.42
2	1.3	5.64	9	3.6	4.07
3	1.7	4.76	10	3.8	3.36
4	2.0	5.01	11	4.0	3.60
5	2.4	4.12	12	4.2	2.67
6	2.7	4.41	13	4.6	2.73
7	2.9	3.46	14	4.8	3.20

where the porosity is expressed as a percentage and the toughness is in $\mathrm{MPa} \cdot \mathrm{m}^{1/2}$. Apply Pearson's linear correlation coefficient to check whether the quantities ε and K_{Ic} can be regarded as stochastically independent, at a significance level (*a*) of 10%, (*b*) of 5% and (*c*) of 1%.

☐ Exercise 7 **Points 4**

A physicochemical treatment is carried out on 10 samples of the same material. The volume density ρ of the material is measured for each sample prior to and after the treatment, providing the results below (in $10^3\,\mathrm{Kg} \cdot \mathrm{m}^{-3}$):

sample no.	ρ before the treatment	ρ after the treatment
1	2.651	2.927
2	2.647	2.904
3	2.735	3.299
4	2.862	3.105
5	2.820	2.864
6	2.807	2.759
7	2.632	2.627
8	2.581	3.087
9	2.747	3.081
10	2.693	2.931

Check, with a significance level of 1% and assuming that the data are normal, whether the treatment actually affects the density of the material.

☐ Exercise 8 **Points 6**

The table below collects the results of some experimental measurements concerning the surface tension γ ($\mathrm{mJ \cdot m^{-2}}$) of pure water as a function of the temperature T (^oC):

i	T_i	γ_{i1}	γ_{i2}	γ_{i3}	γ_{i4}	γ_{i5}
1	10	74.10	73.76	73.76	74.46	73.27
2	20	72.93	72.72	72.08	73.68	71.67
3	30	70.29	70.52	71.63	70.92	71.08
4	40	70.56	69.01	68.81	69.64	70.18
5	50	68.38	67.74	67.33	66.92	68.26
6	60	65.53	66.34	66.66	65.77	65.16
7	70	64.35	64.65	64.19	65.06	64.64
8	80	62.89	61.71	62.17	63.42	62.76

While the random error on the temperatures T_i is negligible, the surface tension data γ_i are independent normal random variables with the same standard deviation $\sigma = 0.50$ (homoskedastic system). Determine:

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

$$\gamma = \mu + \varkappa(T - \overline{T}),$$

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

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

(c) the 95% confidence intervals of the regression parameters μ and $\varkappa$;

(d) the 95% confidence region for predictions;

(e) the 95% confidence interval for the value of γ predicted at $T = 55^oC$.

☐ Exercise 9 Points 4

A thermal treatment is applied to a metallic alloy in order to increase crystallinity and reduce electrical resistivity. 10 samples of the alloy are thermally treated at a temperature of 800 K. An analogous treatment of the same duration is applied to further 14 samples of the same material, but at a temperature of 1200 K. The electrical resistivity of all the samples is finally measured. Look at the table below for the results (data in $\mu\Omega \cdot \mathrm{cm}$):

$T = 800$ K	$T = 1200$ K
2.22	1.05
1.59	1.00
2.16	1.54
1.25	1.44
1.26	1.59
1.86	1.31
1.94	1.15
1.75	1.64
1.54	1.20
1.86	1.54
	1.25
	1.67
	1.85
	1.87

Assuming that the data are normal, and after having checked whether the relative variances can or cannot be regarded as equal, verify with a significance level of 5% if the temperature of the thermal treatment has a significant effect on the electrical resistivity of the alloy.

Remark The sufficient grade corresponds to 18 points

Solution to Exercise 1

Since the data are assumed to be normal, to check the possible presence of an outlier we can apply Chauvenet criterion. We must compute the sample mean $\overline{\varkappa}$ and standard deviation s, and determine the farthest data from the mean, as illustrated in the table below:

i	$\varkappa_i$	$\varkappa_i - \overline{\varkappa}$	$\lvert\varkappa_i - \overline{\varkappa}\rvert$	$\lvert\varkappa_i - \overline{\varkappa}\rvert^2$	outlier
1	0.4329	−0.0683	0.0683	0.00466001	
2	0.4584	−0.0428	0.0428	0.00182878	
3	0.5212	0.0200	0.0200	0.00040143	
4	0.5060	0.0048	0.0048	0.00002338	
5	0.4633	−0.0379	0.0379	0.00143370	
6	0.6045	0.1033	0.1033	0.01067827	×
7	0.5438	0.0426	0.0426	0.00181780	
8	0.5237	0.0225	0.0225	0.00050786	
9	0.5342	0.0330	0.0330	0.00109136	
10	0.4702	−0.0310	0.0310	0.00095879	
11	0.5102	0.0090	0.0090	0.00008164	
12	0.5223	0.0211	0.0211	0.00044672	
13	0.4575	−0.0437	0.0437	0.00190657	
14	0.4681	−0.0331	0.0331	0.00109325	

where:

$$\overline{\varkappa} = \frac{1}{14}\sum_{i=1}^{14}\varkappa_i = 0.5012 \qquad\qquad s = \sqrt{\frac{1}{13}\sum_{i=1}^{14}(\varkappa_i - \overline{\varkappa})^2} = 0.0455$$

while the outlier is the data $\varkappa_6 = 0.6045$, whose distance from the mean $\overline{\varkappa}$ is maximum. The distance z of the suspect data from the mean, in units of s, can be expressed as

$$z = \frac{\varkappa_i - \overline{\varkappa}}{s} = \frac{0.6045 - 0.5012}{0.0455} = 2.2704$$

and is greater than the tabulated critical value of Chauvenet test for $n = 14$ data

$$z_{cr,14} = 2.100165493\,.$$

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 $2.2704s$, by using the table of the standard normal cumulative distribution. We have indeed

$$\begin{aligned} P(|\varkappa_6 - \overline{\varkappa}| \geq 2.2704\,s) &= 1 - P(|\varkappa_6 - \overline{\varkappa}| < 2.2704\,s) = \\ &= 1 - 2\,P(\overline{\varkappa} \leq \varkappa_6 < \overline{\varkappa} + 2.2704s) = \\ &= 1 - 2\cdot 0.48841 = 0.02318 \end{aligned}$$

due to the value of $P(\overline{\varkappa} \leq \varkappa_6 < \overline{\varkappa} + 2.2706s) = 0.48842$ which can be derived by means of the following linear interpolation scheme:

2.2700	0.48840
2.2704	P
2.2800	0.48870

$$\frac{2.2704 - 2.2700}{2.2800 - 2.2700} = \frac{P - 0.48840}{0.48870 - 0.48840}$$

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

$$0.02318 \cdot 14 = 0.32450$$

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

We must compute the resonance frequency according to the formula

$$f = \frac{1}{2\pi} \frac{1}{\sqrt{LC}}$$

with

$$L = (15.0 \pm 0.3) \cdot 10^{-4}\,\mathrm{H} \qquad \text{and} \qquad C = (20.5 \pm 0.1) \cdot 10^{-6}\,\mathrm{F}\,.$$

The estimate of the frequency is reckoned by using the estimated true values of inductance and capacitance, respectively,

$$\overline{L} = 15.0 \cdot 10^{-4}\,\mathrm{H} \qquad\qquad \overline{C} = 20.5 \cdot 10^{-6}\,\mathrm{F}$$

and writes therefore

$$\overline{f} = \frac{1}{2 \cdot 3.14159} \cdot \frac{1}{\sqrt{15.0 \cdot 10^{-4} \cdot 20.5 \cdot 10^{-6}}} = 907.6\,\mathrm{Hz}$$

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

$$\ln f = -\ln 2 - \ln \pi - \frac{1}{2}\ln L - \frac{1}{2}\ln C$$

provides indeed the differential

$$\frac{df}{f} = -\frac{d\pi}{\pi} - \frac{1}{2}\frac{dL}{L} - \frac{1}{2}\frac{dC}{C}$$

and therefore the relative error estimate

$$\frac{\Delta f}{\overline{f}} = \frac{\Delta\pi}{\pi} + \frac{1}{2}\frac{\Delta L}{\overline{L}} + \frac{1}{2}\frac{\Delta C}{\overline{C}}$$

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:

$$\frac{\Delta f}{\overline{f}} = \frac{0.00001}{3.14159} + \frac{1}{2}\frac{0.3}{15.0} + \frac{1}{2}\frac{0.1}{20.5} = 0.000003 + 0.01 + 0.002439 = 0.0124 = 1.24\%$$

Whence we derive the absolute error on the frequency:

$$\Delta f = \frac{\Delta f}{\overline{f}} \cdot \overline{f} = 0.0124 \cdot 907.6 = 11.29\,\mathrm{Hz}$$

so that the error interval of f becomes

$$f = \overline{f} \pm \Delta f = (907.6 \pm 11.29)\,\mathrm{Hz}$$

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

$$f = (908 \pm 11)\,\mathrm{Hz}\,.$$

The precision of the result, expressed by the relative error $\Delta f/\overline{f}$, is essentially the sum of the relative errors

$$\frac{1}{2}\frac{\Delta L}{\overline{L}} = 1\,\% \qquad\qquad \frac{1}{2}\frac{\Delta C}{\overline{C}} = 0.2439\,\%$$

and can be reduced to within $1\,\%$ by improving the precision of the inductance:

$$\frac{1}{2}\frac{\Delta L}{\overline{L}} = 0.01 - 0.002439 = 0.007561 \simeq 0.75\,\%\,.$$

Thus a reduction of the absolute error on the inductance would be needed:

$$\Delta L = 0.3 \cdot 10^{-4}\,\mathrm{H} \quad \longrightarrow \quad \Delta L = 0.3 \cdot 10^{-4} \cdot 0.75\,\% = 0.22 \cdot 10^{-4}\,\mathrm{H}\,.$$

Solution to Exercise 3

The sample can be regarded as large, because the number of data is $n = 300 > 30$. According to the theory of large samples it is not required that the statistical population is 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:

$$\overline{x} = 18.45\,\mathrm{g} \qquad\qquad s = 0.60\,\mathrm{g}\,,$$

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

$$18.45 - z_{[1-\frac{\alpha}{2}]}\frac{0.60}{\sqrt{300}} \;\leq\; \mu \;\leq\; 18.45 + z_{[1-\frac{\alpha}{2}]}\frac{0.60}{\sqrt{300}}$$

or, equivalently,

$$18.45 - 0.034641\cdot z_{[1-\frac{\alpha}{2}]} \;\leq\; \mu \;\leq\; 18.45 + 0.034641\cdot z_{[1-\frac{\alpha}{2}]}\,.$$

(a) *Confidence level* 70%
The confidence level is $1-\alpha = 0.7$, so that $\alpha = 0.3$ and

$$1-\frac{\alpha}{2} \;=\; 1-\frac{0.3}{2} \;=\; 1-0.15 \;=\; 0.85\,.$$

The critical value $z_{[1-\frac{\alpha}{2}]} = z_{[0.85]}$ is then calculated by using the Excel function NORMINV:

NORMINV(0, 85; 0; 1)

which provides

$$z_{[0.85]} \;=\; 1.036433389\,.$$

Alternatively, we can use the equation

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

and search for the value

$$\frac{1}{2}-\frac{\alpha}{2} \;=\; 0.5-\frac{0.3}{2} \;=\; 0.5-0.15 \;=\; 0.35$$

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

$$\int_0^{1.03} \frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz \;=\; 0.34849 \qquad\qquad \int_0^{1.04} \frac{1}{\sqrt{2\pi}}e^{-z^2/2}dz \;=\; 0.35083$$

To improve the approximation we can adopt the following linear interpolation scheme:

$(1-\alpha)/2$	$z_{[1-\frac{\alpha}{2}]}$
0.34849	1.03
0.35	$z_{[0.85]}$
0.35083	1.04

$$\frac{0.35-0.34849}{0.35083-0.34849} \;=\; \frac{z_{[0.85]}-1.03}{1.04-1.03}$$

which gives the critical value estimate

$$z_{[0.85]} = 1.036452991$$

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

$$0.034641 \cdot z_{[1-\frac{\alpha}{2}]} = 0.034641 \cdot z_{[0.85]} = 0.034641 \cdot 1.036433 = 0.035903$$

and the confidence interval for the mean μ of the weight of the screws becomes:

$$\mu = (18.45 \pm 0.04)\,\mathrm{g}\,.$$

An alternative form is, of course,

$$18.41\,\mathrm{g} \leq \mu \leq 18.49\,\mathrm{g}\,.$$

(*b*) *Confidence level* 95%
In this case we have $\alpha = 1 - 0.95 = 0.05$ and therefore

$$1 - \frac{\alpha}{2} = 1 - \frac{0.05}{2} = 1 - 0.025 = 0.975\,.$$

The Excel function NORMINV provides the corresponding critical value $z_{[0.975]}$:

$$\mathrm{NORMINV}(0,975;0;1) \qquad \Longrightarrow \qquad z_{[0.975]} = 1.959963985$$

so that

$$0.034641 \cdot z_{[1-\frac{\alpha}{2}]} = 0.034641 \cdot z_{[0.975]} = 0.034641 \cdot 1.959964 = 0.067895$$

and the confidence interval for the mean takes the form

$$\mu = (18.45 \pm 0.07)\,\mathrm{g}$$

i.e.

$$18.38\,\mathrm{g} \leq \mu \leq 18.52\,\mathrm{g}\,.$$

As before, a satisfactory approximation can be obtained by means of the equation

$$\int_0^{z_{[0.975]}} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \frac{1}{2} - \frac{0.05}{2} = 0.5 - 0.025 = 0.475$$

and looking for the value 0.475 among the entries of the cumulative standard normal distribution. The closest approcimation is:

$$\int_0^{1.96} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 0.47500$$

so that $z_{[0.975]} = 1.96$, a value which differs from the true one only by less than 10^{-4}. The confidence interval with confidence level of 95% is larger than that at confidence level of 70%, as obviously expected.

Solution to Exercise 4

The sample histogram is bell-shaped, so that it is pretty reasonable to suppose that the data belong to a normal population. 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 test the null hypothesis

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

versus the alternative hypothesis

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

In this case the sample data were used to estimate the mean and the standard deviation of the distribution:

$$\mu = \overline{k} = 1.00 \qquad\qquad \sigma = s = 0.08$$

and the classes of the results (the histogram intervals) are $h = 11$ 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

$$n = h - c - 1 = 11 - 2 - 1 = 8$$

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

$$z = \frac{k - \mu}{\sigma}$$

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

k	z
0.82	−2.25
0.86	−1.75
0.90	−1.25
0.94	−0.75
0.98	−0.25
1.02	0.25
1.06	0.75
1.10	1.25
1.14	1.75
1.18	2.25

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	$n_i = 300 \cdot p_i$	f_i	$f_i - n_i$	$(f_i - n_i)^2/n_i$
1	$\Phi(+\infty) - \Phi(2.25) = 0.01222$	3.667342	4	0.332658	0.030174848
2	$\Phi(2.25) - \Phi(1.75) = 0.02784$	8.350405	11	2.649595	0.840719947
3	$\Phi(1.75) - \Phi(1.25) = 0.06559$	19.677185	18	-1.677185	0.142954882
4	$\Phi(1.25) - \Phi(0.75) = 0.12098$	36.293274	34	-2.293274	0.144905745
5	$\Phi(0.75) - \Phi(0.25) = 0.17466$	52.399897	55	2.600103	0.129018151
6	$\Phi(0.25) + \Phi(0.25) = 0.19742$	59.223795	54	-5.223795	0.460761393
7	$\Phi(0.75) - \Phi(0.25) = 0.17466$	52.399897	50	-2.399897	0.109914408
8	$\Phi(1.25) - \Phi(0.75) = 0.12098$	36.293274	40	3.706726	0.378577602
9	$\Phi(1.75) - \Phi(1.25) = 0.06559$	19.677185	21	1.322815	0.088927324
10	$\Phi(2.25) - \Phi(1.75) = 0.02784$	8.350405	10	1.649595	0.325871944
11	$\Phi(+\infty) - \Phi(2.25) = 0.01222$	3.667342	3	-0.667342	0.121435388

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:

$$\mathcal{X}^2 = \sum_{i=1}^{11} \frac{(f_i - n_i)^2}{n_i} = 2.773261633\,.$$

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

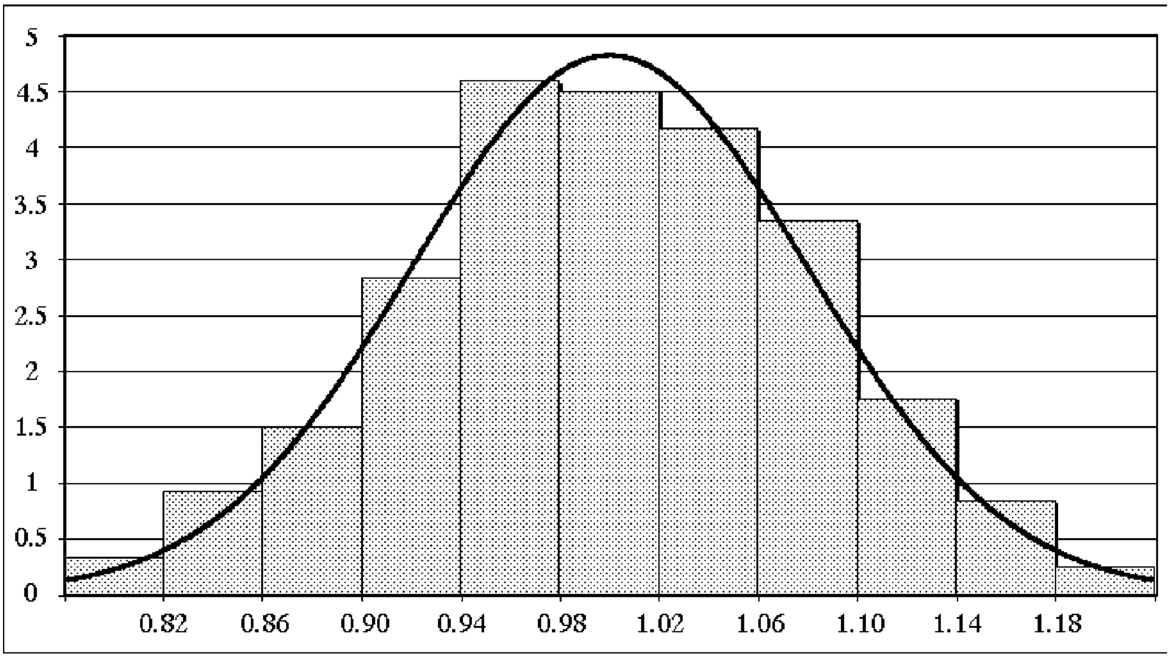
$$\mathcal{X}^2{}_{[1-\alpha](8)} = \mathcal{X}^2{}_{[0.95](8)} = 15.507 \qquad \text{for } \alpha = 0.05$$

$$\mathcal{X}^2{}_{[1-\alpha](8)} = \mathcal{X}^2{}_{[0.99](8)} = 20.090 \qquad \text{for } \alpha = 0.01\,.$$

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 1%, *the null hypothesis cannot be rejected.* The sample data suggest that *the distribution of the conductivity of the ceramic is presumably normal.* The formal conclusion is supported by the good overlap between the theoretical distribution of the data:

$$p(k) = \frac{1}{\sqrt{2\pi}\,s} e^{-(k-\overline{k})^2/2s^2} = \frac{1}{\sqrt{2\pi}\,0.08} e^{-(k-1.00)^2/2\cdot 0.08^2}$$

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 $f_i/300/0.04$, where the number of data 300 is introduced to normalize the distribution to 1, while the factor 0.04 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:

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

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

i	p_i
1	$P(-2.25) - P(-\infty) = 0.012224 - 0.000000 = 0.012224$
2	$P(-1.75) - P(-2.25) = 0.040059 - 0.012224 = 0.027835$
3	$P(-1.25) - P(-1.75) = 0.105650 - 0.040059 = 0.065591$
4	$P(-0.75) - P(-1.25) = 0.226627 - 0.105650 = 0.120978$
5	$P(-0.25) - P(-0.75) = 0.401294 - 0.226627 = 0.174666$
6	$P(0.25) - P(-0.25) = 0.598706 - 0.401294 = 0.197413$
7	$P(0.75) - P(0.25) = 0.773373 - 0.598706 = 0.174666$
8	$P(1.25) - P(0.75) = 0.894350 - 0.773373 = 0.120978$
9	$P(1.75) - P(1.25) = 0.959941 - 0.894350 = 0.065591$
10	$P(2.25) - P(1.75) = 0.987776 - 0.959941 = 0.027835$
11	$P(+\infty) - P(2.25) = 1.000000 - 0.987776 = 0.012224$

whose results coincide with those derived from the statistical table.

Solution to Exercise 5

The sample consists of $n = 21$ data and cannot be considered large, since the number of data is smaller than 30. It is then necessary to reckon 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 regarded as essentially 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:

$$\overline{v} = \frac{1}{n}\sum_{i=1}^{n} v_i = 1.471486\,.$$

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	1.4585	−1.2985714	1.6862878
2	1.4570	−1.4485714	2.0983592
3	1.4766	0.5114286	0.2615592
4	1.4551	−1.6385714	2.6849163
5	1.4534	−1.8085714	3.2709306
6	1.4724	0.0914286	0.0083592
7	1, 4887	1.7214286	2.9633163
8	1.4924	2.0914286	4.3740735
9	1.4666	−0.4885714	0.2387020
10	1.4826	1.1114286	1.2352735
11	1.4550	−1.6485714	2.7177878
12	1.4663	−0.5185714	0.2689163
13	1.4550	−1.6485714	2.7177878
14	1.4719	0.0414286	0.0017163
15	1.4970	2.5514286	6.5097878
16	1.4406	−3.0885714	9, 5392735
17	1.4625	−0.8985714	0.8074306
18	1.4770	0.5514286	0.3040735
19	1.5047	3.3214286	11.0318878
20	1.4845	1.3014286	1.6937163
21	1.4834	1.1914286	1.4195020

from which we deduce the sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(v_i - \overline{v})^2 = 2.79168 \cdot 10^{-4}$$

and the sample estimate of the standard deviation:

$$s = \sqrt{s^2} = 1.67083 \cdot 10^{-2}\,.$$

We have thus calculated the basic quantities to determine the confidence intervals of the mean and standard deviation, at the confidence level $1 - \alpha = 0.95$.

(*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 $\alpha = 0.05$, $n = 21$ has thus the limits

$$\begin{aligned} \overline{v} - t_{[0.975](20)}\frac{s}{\sqrt{21}} &= 1.471486 - 2.086 \cdot \frac{0.0167083}{\sqrt{21}} = 1.46388 \\ \overline{v} + t_{[0.975](20)}\frac{s}{\sqrt{21}} &= 1.471486 + 2.086 \cdot \frac{0.0167083}{\sqrt{21}} = 1.47909 \end{aligned}$$

so that the confidence interval becomes

$$1.46388\,\mathrm{V} \le \mu \le 1.47909\,\mathrm{V}$$

or, equivalently,

$$\mu = (1.47149 \pm 0.00760)\,\mathrm{V}\,.$$

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

$$\mu = (1.472 \pm 0.008)\,\mathrm{V}$$

can be regarded as more than satisfactory. *Remember that*, besides looking at the statistical table, the critical value $t_{[0.975](20)}$ can also be calculated by the Excel function TINV, as follows

$$\mathrm{TINV}(0,05;20) \qquad \Longrightarrow \qquad 2.085963$$

(*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 $\alpha = 0.05$ and $n = 21$. Therefore the lower and the upper limits are:

$$\frac{1}{\mathcal{X}^2{}_{[0.975](20)}}\,20\,s^2 = \frac{1}{34.170}\,20 \cdot 2.79168 \cdot 10^{-4} = 1.6339964 \cdot 10^{-4}$$

$$\frac{1}{\mathcal{X}^2{}_{[0.025](20)}}\,20\,s^2 = \frac{1}{9.591}\,20 \cdot 2.79168 \cdot 10^{-4} = 5.8214636 \cdot 10^{-4}$$

and the CI of the variance becomes

$$1.6339964 \cdot 10^{-4}\,\mathrm{V}^2 \le \sigma^2 \le 5.8214636 \cdot 10^{-4}\,\mathrm{V}^2\,.$$

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

$$1.2782787 \cdot 10^{-2}\,\mathrm{V} \le \sigma \le 2.4127709 \cdot 10^{-2}\,\mathrm{V}\,.$$

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

$$1.28 \cdot 10^{-2}\,\mathrm{V} \le \sigma \le 2.41 \cdot 10^{-2}\,\mathrm{V}\,.$$

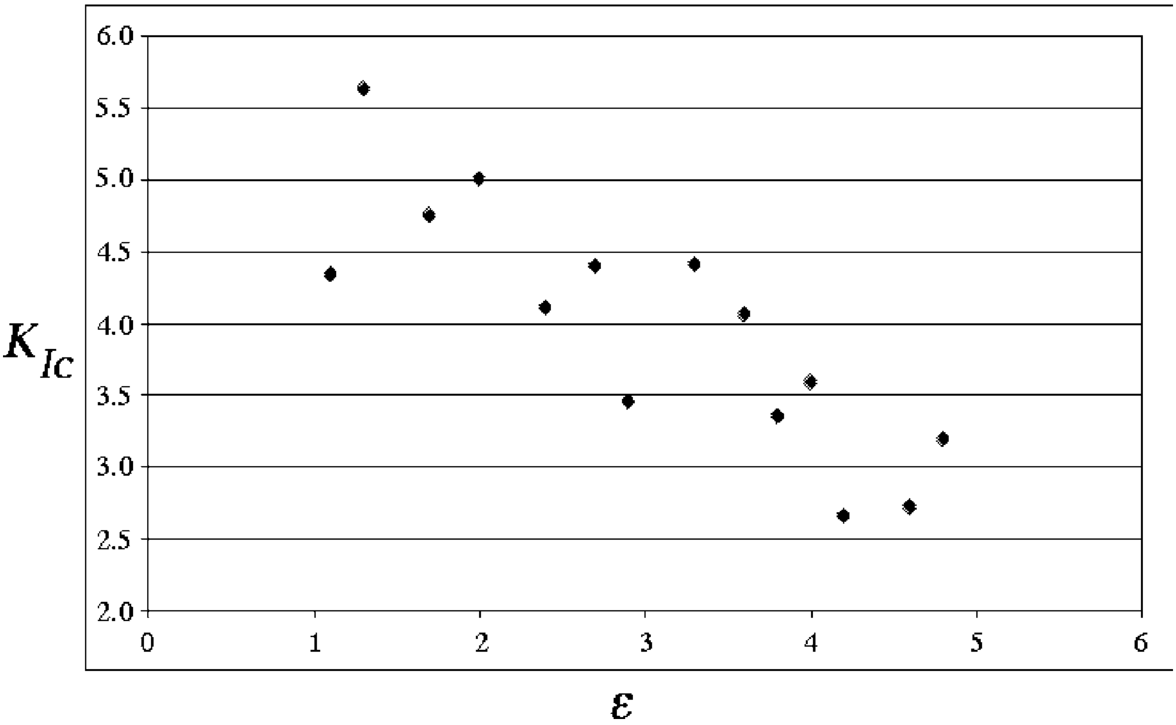
The critical values of the $\mathcal{X}^2$ variable con be also reckoned by the Excel function CHIINV:

$$\mathrm{CHIINV}(0,025;20) \qquad \Longrightarrow \qquad {\mathcal{X}^2}_{[0.975](20)} = 34.169607$$

$$\mathrm{CHIINV}(0,975;20) \qquad \Longrightarrow \qquad {\mathcal{X}^2}_{[0.025](20)} = 9.590778\,.$$

Solution to Exercise 6

The data plot suggests that the porosity ε and the toughness K_{Ic} are described by dependent random variables (remember that 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 $\overline{K}_{Ic}$ of the two quantities are given by:

$$\overline{\varepsilon} = \frac{1}{14}\sum_{i=1}^{14}\varepsilon_i = 3.028571 \qquad\qquad \overline{K}_{Ic} = \frac{1}{14}\sum_{i=1}^{14}K_{Ic\,i} = 3.985714$$

and allow us to calculate the sum of products of residuals

$$SS_{\varepsilon K} = \sum_{i=1}^{14}(\varepsilon_i - \overline{\varepsilon})(K_{Ic\,i} - \overline{K}_{Ic}) = -11.398286$$

and those of the relative squares:

$$SS_{\varepsilon\varepsilon} = \sum_{i=1}^{14}(\varepsilon_i - \overline{\varepsilon})^2 = 18.968571$$

$$SS_{KK} = \sum_{i=1}^{14}(K_{Ic\,i} - \overline{K}_{Ic})^2 = 9.653743\,,$$

as illustrated in the following table:

ε_i	$K_{Ic\,i}$	$\Delta\varepsilon_i$	$\Delta K_{Ic\,i}$	$\Delta\varepsilon_i^2$	$\Delta K_{Ic\,i}^2$	$\Delta\varepsilon_i \Delta K_{Ic\,i}$
1.1	4.35	−1.928571	0.364286	3.719388	0.132704	−0.702551
1.3	5.64	−1.728571	1.654286	2.987959	2.736661	−2.859551
1.7	4.76	−1.328571	0.774286	1.765102	0.599518	−1.028694
2.0	5.01	−1.028571	1.024286	1.057959	1.049161	−1.053551
2.4	4.12	−0.628571	0.134286	0.395102	0.018033	−0.084408
2.7	4.41	−0.328571	0.424286	0.107959	0.180018	−0.139408
2.9	3.46	−0.128571	−0.525714	0.016531	0.276376	0.067592
3.3	4.42	0.271429	0.434286	0.073673	0.188604	0.117878
3.6	4.07	0.571429	0.084286	0.326531	0.007104	0.048163
3.8	3.36	0.771429	−0.625714	0.595102	0.391518	−0.482694
4.0	3.60	0.971429	−0.385714	0.943673	0.148776	−0.374694
4.2	2.67	1.171429	−1.315714	1.372245	1.731104	−1.541265
4.6	2.73	1.571429	−1.255714	2.469388	1.576818	−1.973265
4.8	3.20	1.771429	−0.785714	3.137959	0.617347	−1.391837

where, for simplicity's sake, we have posed $\varepsilon_i - \overline{\varepsilon} = \Delta\varepsilon_i$ and $K_{Ic\,i} - \overline{K}_{Ic} = \Delta K_{Ic\,i}$. The linear correlation coefficient becomes

$$r = \frac{SS_{\varepsilon K}}{\sqrt{SS_{\varepsilon\varepsilon}}\,\sqrt{SS_{KK}}} = \frac{-11.398286}{\sqrt{18.968571}\,\sqrt{9.653743}} = -0.842315\,.$$

Since the bivariate random variable (ε, K_{Ic}) may be assumed to be normal, we can check the null hypothesis

$$H_0 \;:\; \varepsilon \text{ and } K_{Ic} \text{ are stochastically independent}$$

against the alternative hypothesis

$$H_1 \;:\; \varepsilon \text{ and } K_{Ic} \text{ are stochastically dependent}$$

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 $n = 14$ and get:

$$\boxed{t = \sqrt{14-2}\frac{-0.842315}{\sqrt{1-(-0.842315)^2}} = -5.413616\,.}$$

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

$$\left\{t \in \mathbb{R} \,:\, |t| > t_{[1-\frac{\alpha}{2}](n-2)}\right\} = \left\{t \in \mathbb{R} \,:\, |t| > t_{[1-\frac{\alpha}{2}](12)}\right\}.$$

(a) *Significance level* $\alpha = 10\%$
In this case the critical value of the test statistic is

$$\boxed{t_{[1-\frac{\alpha}{2}](12)} = t_{[0.95](12)} = 1.782}$$

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

$$\text{TINV}(0,10;12) \qquad \Longrightarrow \qquad t_{[0.95](12)} = 1.782287548$$

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_{Ic} can be considered stochastically dependent/correlated.

(b) *Significance level* $\alpha = 5\%$
For this significance level the critical value of the test statistic becomes

$$\boxed{t_{[1-\frac{\alpha}{2}](12)} = t_{[0.975](12)} = 2.179\,.}$$

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

$$\text{TINV}(0,05;12) \qquad \Longrightarrow \qquad t_{[0.975](12)} = 2.178812827\,.$$

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

(*c*) *Significance level* $\alpha = 1\%$
The critical value of the test statistic is now

$$\boxed{t_{[1-\frac{\alpha}{2}](12)} = t_{[0.995](12)} = 3.055}$$

and can be more accurately calculated by the TINV function:

$$\mathrm{TINV}(0,01;12) \qquad \Longrightarrow \qquad t_{[0.995](12)} = 3.054539586\,.$$

Nihil novi sub solem: since the value -5.413616 of the test statistic again does not fall within the acceptance region for H_0, we must conclude that *the random variables* ε *and* K_{Ic} *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 the density ρ is measured prior to and after the treatment *on each sample.* Therefore, for all the $n = 10$ 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 densities measured before and after the treatment, respectively. If μ_1 and μ_2 stand for the mean (true) value of the density 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 ρ, 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 $n - 1 = 9$ 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 $n = 10$, $\alpha = 0.01$ becomes

$$\big\{t \leq -3.250\big\} \cup \big\{t \geq 3.250\big\}$$

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

$$\boxed{t_{[1-\frac{\alpha}{2}](n-1)} = t_{[0.995](9)} = 3.250}$$

for which a more accurate estimate is available

$$t_{[0.995](9)} \;=\; 3.249836$$

by using the Excel function TINV(0, 01; 9). In the present case the sample means of the two samples hold:

$$\overline{y} \;=\; \frac{1}{10}\sum_{i=1}^{10} y_i \;=\; 2.7175 \qquad\qquad \overline{z} \;=\; \frac{1}{10}\sum_{i=1}^{10} z_i \;=\; 2.9584$$

whereas

$$\frac{1}{n-1}\sum_{i=1}^{n}(y_i - z_i - \overline{y} + \overline{z})^2 \;=\; \frac{1}{9}\cdot 0.36754290 \;=\; 0.04083810$$

and therefore

$$\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - z_i - \overline{y} + \overline{z})^2} \;=\; \sqrt{0.04083810} \;=\; 0.2020844$$

so that the test variable assumes the value of

$$t \;=\; \sqrt{10}\cdot\frac{2.7175 - 2.9584}{0.2020844} \;=\; -3.7697\,.$$

The table below shows the detailed calculations:

y_i	z_i	$y_i - z_i$	$y_i - z_i - \overline{y} + \overline{z}$	$(y_i - z_i - \overline{y} + \overline{z})^2$
2.651	2.927	−0.276	−0.03510	0.001232
2.647	2.904	−0.257	−0.01610	0.000259
2.735	3.299	−0.564	−0.32310	0.104394
2.862	3.105	−0.243	−0.00210	0.000004
2.820	2.864	−0.044	0.19690	0.038770
2.807	2.759	0.048	0.28890	0.083463
2.632	2.627	0.005	0.24590	0.060467
2.581	3.087	−0.506	−0.26510	0.070278
2.747	3.081	−0.334	−0.09310	0.008668
2.693	2.931	−0.238	0.00290	0.000008

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

$$-3.7697 \;<\; -3.250\,,$$

and therefore *we can exclude*, with a significance level of 1%, *that the true value of the density 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 surface tension are 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 surface tensione data share the same variance. That's why the regression straight line is written by expressing the surface tension γ as a function of the temperature T:

$$\gamma = \mu + \varkappa(T - \overline{T})$$

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

Notice that the sample consists in multiple measurements at each temperature: many measurements of surface tension 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, γ_i) with the same T are treated as distinct. According to this criterion the whole number of sample data is thus $n = 40$.

(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 m and q of the parameters can be expressed as

$$m = \overline{\gamma} = \frac{1}{n}\sum_{i=1}^{n}\gamma_i = 68.475 \qquad q = \frac{\displaystyle\sum_{i=1}^{n}(T_i - \overline{T})\,\gamma_i}{\displaystyle\sum_{i=1}^{n}(T_i - \overline{T})^2} = -0.161966667$$

with $n = 40$ and

$$\overline{T} = \frac{1}{n}\sum_{i=1}^{n}T_i = 45.0\,.$$

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

$$\begin{aligned}\gamma = m + q\,(T - \overline{T}) &= 68.475 - 0.161966667\cdot(T - 45.0) = \\ &= 75.7635 - 0.161966667\,\cdot\,T\end{aligned}$$

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	γ_i	$T_i - \overline{T}$	$(T_i - \overline{T})^2$	$(T_i - \overline{T})\,\gamma_i$
10	74.10	-35.0	1225.0	-2593.500
20	72.93	-25.0	625.0	-1823.250
30	70.29	-15.0	225.0	-1054.350
40	70.56	-5.0	25.0	-352.800
50	68.38	5.0	25.0	341.900
60	65.53	15.0	225.0	982.950
70	64.35	25.0	625.0	1608.750
80	62.89	35.0	1225.0	2201.150
10	73.76	-35.0	1225.0	-2581.600
20	72.72	-25.0	625.0	-1818.000
30	70.52	-15.0	225.0	-1057.800
40	69.01	-5.0	25.0	-345.050
50	67.74	5.0	25.0	338.700
60	66.34	15.0	225.0	995.100
70	64.65	25.0	625.0	1616.250
80	61.71	35.0	1225.0	2159.850
10	73.76	-35.0	1225.0	-2581.600
20	72.08	-25.0	625.0	-1802.000
30	71.63	-15.0	225.0	-1074.450
40	68.81	-5.0	25.0	-344.050
50	67.33	5.0	25.0	336.650
60	66.66	15.0	225.0	999.900
70	64.19	25.0	625.0	1604.750
80	62.17	35.0	1225.0	2175.950
10	74.46	-35.0	1225.0	-2606.100
20	73.68	-25.0	625.0	-1842.000
30	70.92	-15.0	225.0	-1063.800
40	69.64	-5.0	25.0	-348.200
50	66.92	5.0	25.0	334.600
60	65.77	15.0	225.0	986.550
70	65.06	25.0	625.0	1626.500
80	63.42	35.0	1225.0	2219.700
10	73.27	-35.0	1225.0	-2564.450
20	71.67	-25.0	625.0	-1791.750
30	71.08	-15.0	225.0	-1066.200
40	70.18	-5.0	25.0	-350.900
50	68.26	5.0	25.0	341.300
60	65.16	15.0	225.0	977.400
70	64.64	25.0	625.0	1616.000
80	62.76	35.0	1225.0	2196.600

(*b*) *Goodness of fit*
The goodness of fit Q of the regression model is defined by the relationship

$$Q = \int\limits_{\mathrm{NSSAR}}^{+\infty} \rho_{n-2}(\mathcal{X}^2)\, d\mathcal{X}^2$$

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

$$\mathrm{NSSAR} = \sum_{i=1}^{n} \frac{1}{\sigma^2}[m + q\,(T_i - \overline{T}) - \gamma_i]^2 = \frac{\mathrm{SSAR}}{\sigma^2}$$

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* $\sigma = 0.50$, since we need to determine the NSSAR, and not simply the SSAR. In the present case we have $n = 40$ data and the regression model is based on the two parameters μ and $\varkappa$. Consequently, the NSSAR obeys a $\mathcal{X}^2$ distribution with $n-2 = 38$ d.o.f. For the given sample the normalized sum of squares around regression holds

$$\mathrm{NSSAR} = \frac{\mathrm{SSAR}}{\sigma^2} = \frac{13.2433767}{0.50^2} = 52.97350667\,.$$

On the table of the upper critical values of $\mathcal{X}^2$ with $\nu = 38$ d.o.f. we find

Probability $\{\mathcal{X}^2 \geq 49.513\}$	Probability $\{\mathcal{X}^2 \geq 53.384\}$
0.10	0.05

so that a simple linear interpolation scheme:

49.513	0.10
52.9735	Q
53.384	0.05

$$\frac{52.9735 - 49.513}{53.384 - 49.513} = \frac{Q - 0.10}{0.05 - 0.10}$$

may provide a reasonable estimate of Q:

$$Q = 0.10 + (0.05 - 0.10)\,\frac{52.9735 - 49.513}{53.384 - 49.513} = 0.0553\,.$$

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

$$Q \,=\, \text{Probability}\{\mathcal{X}^2 \geq 52.97350667\} \,=\, \int\limits_{52.97350667}^{+\infty} p_{38}(\mathcal{X}^2)\, d\mathcal{X}^2$$

for instance by using the Excel function CHIDIST:

$$\text{CHIDIST}(52,97350667;38) \qquad \Longrightarrow \qquad \int\limits_{52.97350667}^{+\infty} p_{38}(\mathcal{X}^2)\, d\mathcal{X}^2 \,=\, 0.054005458\,.$$

Alternatively, we may execute the Maple command line

$$1 - \text{stats}[\text{statevalf}, \text{cdf}, \text{chisquare}[38]](52.97350667);$$

to obtain the “exact” value $Q \,=\, 0.0540054578$. The goodness of fit of the regression model is thus equal to about 5.4%: such a percentage, if the regression model were rejected, would express the probability of a type I error.

(*c*) *Confidence intervals for the regression parameters*
The sum of squares around regression has already been determined:

$$\text{SSAR} \,=\, \sum_{i=1}^{n} [m + q(T_i - \overline{T}) - \gamma_i]^2 \,=\, 13.2433767\,.$$

At the significance level $1-\alpha$, the CI of the parameter μ and that of the slope $\varkappa$ are given by the formulas:

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

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

Here we have $\alpha = 0.05$ and $n = 40$, so that the confidence intervals become:

$$\mu \,=\, m \pm t_{[0.975](38)} \sqrt{\frac{1}{40}\frac{\text{SSAR}}{38}}$$

$$\varkappa \,=\, q \pm t_{[0.975](38)} \sqrt{\left[\sum_{i=1}^{40}(T_i - \overline{T})^2\right]^{-1} \frac{\text{SSAR}}{38}}$$

where:

$$\begin{aligned} m &= 68.475 \\ q &= -0.16196667 \\ \mathrm{SSAR} &= 13.2433767 \\ \sum_{i=1}^{40}(T_i - \overline{T})^2 &= 21000.0 \\ t_{[0.975](38)} &= 2.024 \,. \end{aligned}$$

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:

$$\mathrm{TINV}(0,05;38) \qquad \Longrightarrow \qquad t_{[0.975](38)} = 2.024394147 \,.$$

As a satisfactory compromise between the tabulated and the Excel value we may assume $\mathbf{t}_{[\mathbf{0.975}](\mathbf{38})} = \mathbf{2.02439}$. By inserting the numerical values and performing the calculations we deduce that:

- the 95%-CI of the parameter μ is

  $$\mu = 68.475 \pm 2.02439\sqrt{\frac{1}{40}\frac{13.2433767}{38}}$$

  i.e.

  $$\mu = 68.475 \pm 0.1889608158 = [68.28603918, 68.66396082]$$

- the 95%-CI for the slope $\varkappa$ holds

  $$\varkappa = -0.16196667 \pm 2.02439\sqrt{\frac{1}{21000.0}\frac{13.2433767}{38}}$$

  or, equivalently,

  $$\varkappa = -0.16196667 \pm 0.008246926108 = [-0.1702135928, -0.1537197406] \,.$$

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

$$\mu = [68.2860, 68.6639]\,\mathrm{mJ}\cdot\mathrm{m}^{-2} = (68.4750 \pm 0.1889)\,\mathrm{mJ}\cdot\mathrm{m}^{-2}$$

while

$$\begin{aligned} \varkappa &= [-0.170214, -0.153719]\,\mathrm{mJ}\cdot\mathrm{m}^{-2}\cdot\ {}^o\mathrm{C}^{-1} = \\ &= (-0.161967 \pm 0.008246)\,\mathrm{mJ}\cdot\mathrm{m}^{-2}\cdot\ {}^o\mathrm{C}^{-1} \,. \end{aligned}$$

(*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 $\gamma = \gamma_0$ at a given $T = T_0$ is expressed by the general formula:

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

where, more specifically, we have:

$$\begin{aligned}
m \;&=\; 68.475 \\
q \;&=\; -0.16196667 \\
\overline{T} \;&=\; \frac{1}{n}\sum_{i=1}^{n} T_i \;=\; 45.0 \\
t_{[1-\frac{\alpha}{2}](n-2)} \;&=\; t_{[0.975](38)} \;=\; 2.02439 \\
V \;&=\; 1 + \frac{1}{n} + \frac{1}{\displaystyle\sum_{i=1}^{n}(T_i - \overline{T})^2}(T_0 - \overline{T})^2 \;=\; 1 + \frac{1}{40} + \frac{(T_0 - 45.0)^2}{21000.0} \;= \\
&=\; 1.025 + 0.00004761904762 \cdot (T_0 - 45.0)^2 \\
\mathrm{SSAR} \;&=\; \sum_{i=1}^{n}\big[m + q(T_i - \overline{T}) - \gamma_i\big]^2 \;=\; 13.2433767\;.
\end{aligned}$$

The CI for the prediction of the surface tension γ at $T = T_0$ is then:

$$\begin{aligned}
\gamma_0 \;&=\; 68.475 - 0.161966667 \cdot (T_0 - 45.0) \pm \\
&\pm\; 2.02439 \cdot \sqrt{1.025 + 0.00004761904762 \cdot (T_0 - 45.0)^2}\sqrt{\frac{13.2433767}{38}}
\end{aligned}$$

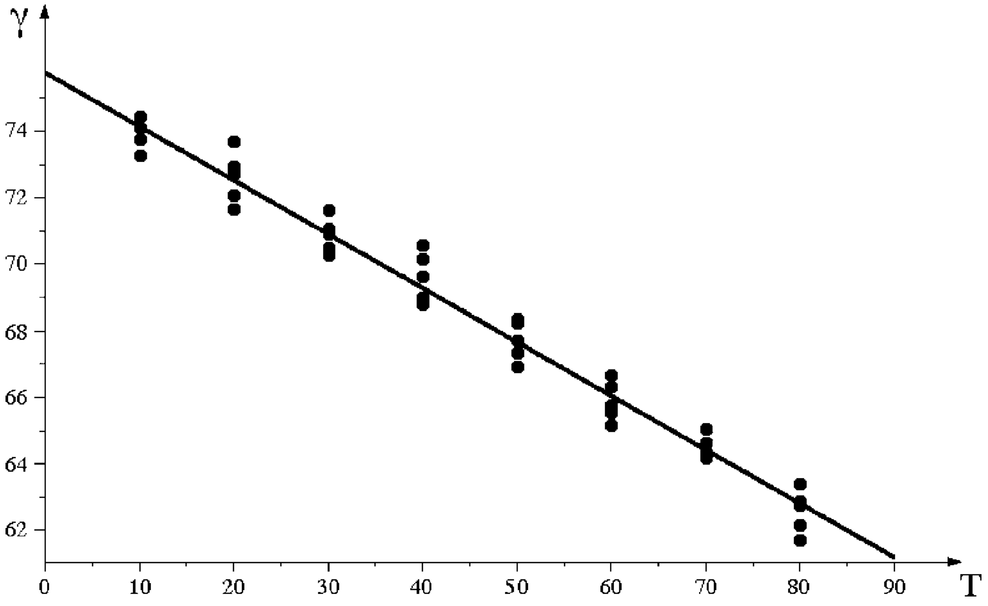
and performing the calculations reduces to:

$$\begin{aligned}
\gamma_0 \;&=\; 75.76350000 - 0.1619666667 \cdot T_0 \pm \\
&\pm\; 1.195093133 \cdot \sqrt{1.025 + 0.00004761904762 \cdot (T_0 - 45.0)^2}
\end{aligned}$$

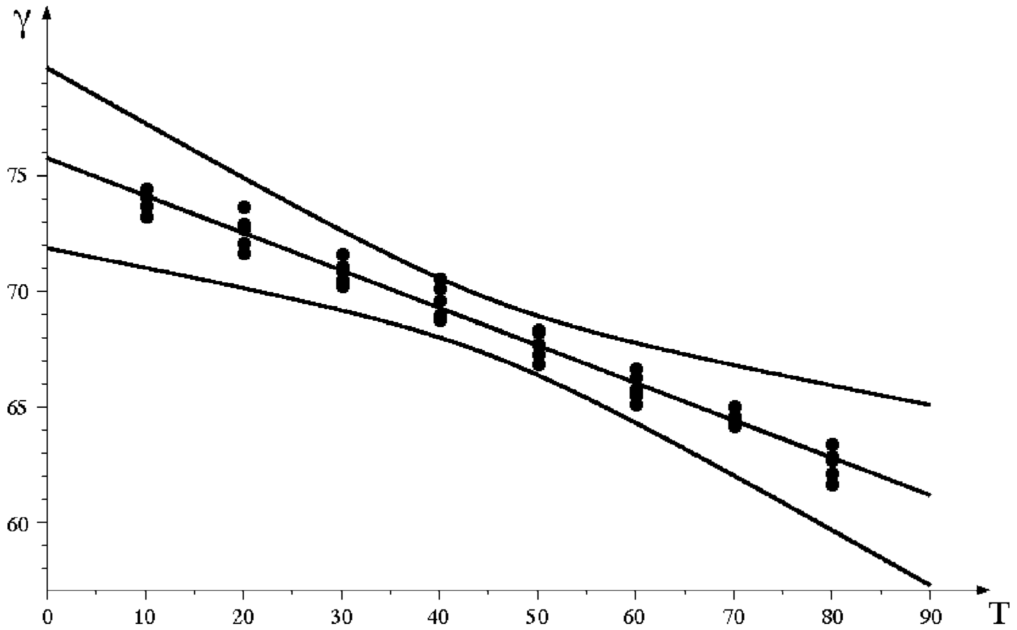
or, more conveniently,

$$\boxed{\begin{aligned}
\gamma_0 \;&=\; 75.76350000 - 0.1619666667 \cdot T_0 \pm \\
&\pm\; 1.195093133 \cdot \sqrt{1.025 + 0.4761904762 \cdot \Big(\frac{T_0}{100} - 0.45\Big)^2}
\end{aligned}}$$

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 γ axis, is minimum for $T = \overline{T} = 45.0$ 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 100.

(*e*) *Confidence interval for a prediction*

The CI at a confidence level of 95% for the prediction of γ at $T = 55^oC$ can be obtained by posing $T_0 = 55$ in the previous formula

$$\begin{aligned}\gamma_0 \,&=\, 75.76350000 - 0.1619666667 \cdot T_0 \pm \\ &\pm\, 1.195093133 \cdot \sqrt{1.025 + 0.4761904762 \cdot \Big(\frac{T_0}{100} - 0.45\Big)^2} \;.\end{aligned}$$

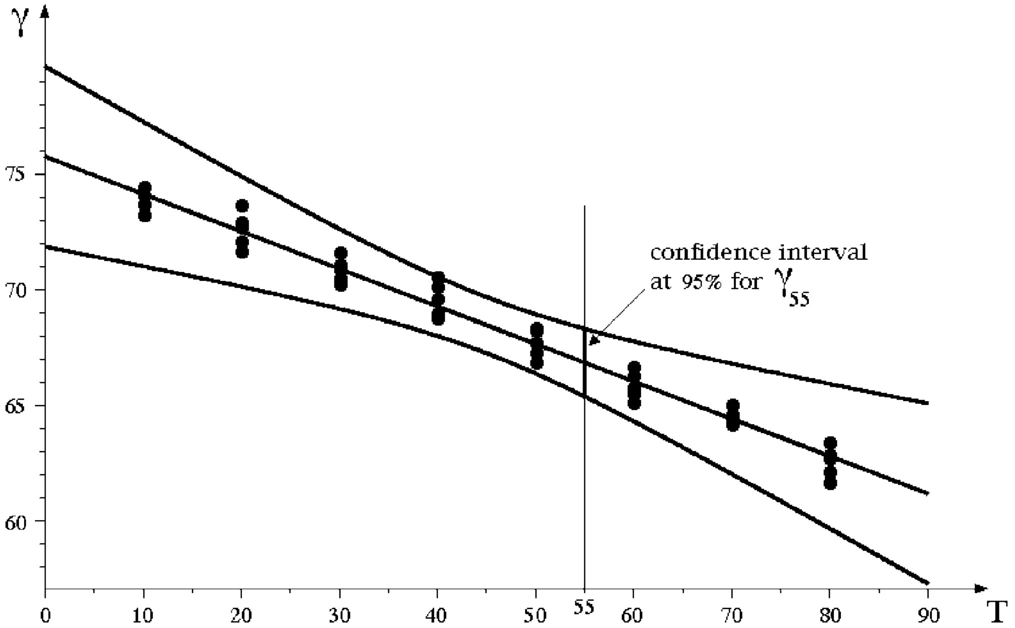
We obtain therefore:

$$\begin{aligned}\gamma_0 \,=\, \gamma_{55} \,&=\, 75.76350000 - 0.1619666667 \cdot 55 \pm \\ &\pm\, 1.195093133 \cdot \sqrt{1.025 + 0.4761904762 \cdot \Big(\frac{55}{100} - 0.45\Big)^2} \,= \\ &=\, 66.85533333 \pm 1.21274687 \,=\, [65.64258646,\, 68.06808020]\end{aligned}$$

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

$$\gamma_{55} \,=\, [65.7,\, 68.1]\, \mathrm{mJ} \cdot \mathrm{m}^{-2} \,=\, (66.9 \pm 1.2)\, \mathrm{mJ} \cdot \mathrm{m}^{-2} \,.$$

In the following figure the CI at 95% is highlighted as the intersection of the confidence region at 95% with the vertical straight line of equation $T = 55$:



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 resistivity for the first and the second treatment, respectively. Let $p = 10$ 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 resistivity of the material) against the alternative hypothesis $H_1 : \mu_1 \neq \mu_2$ (the two treatments yield materials with a different electrical resistivity). The significance level we choose is 5%.

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

$$\overline{y} = \frac{1}{p}\sum_{i=1}^{p} y_i = 1.7430 \qquad \overline{z} = \frac{1}{q}\sum_{j=1}^{q} z_j = 1.435714$$

$$s_y^2 = \frac{1}{p-1}\sum_{i=1}^{p}(y_i - \overline{y})^2 = 0.112068 \qquad s_z^2 = \frac{1}{q-1}\sum_{j=1}^{q}(z_j - \overline{z})^2 = 0.079119$$

according to the detailed calculation shown in the table below:

y_i	z_i	$y_i - \overline{y}$	$(y_i - \overline{y})^2$	$z_i - \overline{z}$	$(z_i - \overline{z})^2$
2.22	1.05	0.4770	0.227529	−0.385714	0.148776
1.59	1.00	−0.1530	0.023409	−0.435714	0.189847
2.16	1.54	0.4170	0.173889	0.104286	0.010876
1.25	1.44	−0.4930	0.243049	0.004286	0.000018
1.26	1.59	−0.4830	0.233289	0.154286	0.023804
1.86	1.31	0.1170	0.013689	−0.125714	0.015804
1.94	1.15	0.1970	0.038809	−0.285714	0.081633
1.75	1.64	0.0070	0.000049	0.204286	0.041733
1.54	1.20	−0.2030	0.041209	−0.235714	0.055561
1.86	1.54	0.1170	0.013689	0.104286	0.010876
	1.25			−0.185714	0.034490
	1.67			0.234286	0.054890
	1.85			0.414286	0.171633
	1.87			0.434286	0.188604

We obtain therefore:

$$F = \frac{0.112068}{0.079119} = 1.4165 .$$

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 $p = 10$, $q = 14$ and $\alpha = 0.05$, so that the acceptance condition becomes

$$0.261056 = F_{[0.025](9,13)} < F < F_{[0.975](9,13)} = 3.312032$$

as derived from the Excel function FINV:

$$\begin{aligned} &\mathrm{FINV}(0,975;9;13) &\Longrightarrow\quad & F_{[0.025](9,13)} = 0.261056 \\ &\mathrm{FINV}(0,025;9;13) &\Longrightarrow\quad & F_{[0.975](9,13)} = 3.312032 \end{aligned}$$

or, with a lesser accuracy, by using the table of the Fisher cumulative distributions — $F_{[0.025](9,13)} = 0.2611$ and $F_{[0.975](9,13)} = 3.3120$. Since the value of the test statistic actually falls within the acceptance region:

$$0.261056 < 1.4165 < 3.312032$$

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

$$\frac{0.112068 \cdot 9 + 0.079119 \cdot 13}{10 + 14 - 2} = 0.092597857$$

and the value of the test statistic is

$$t = \frac{\overline{y} - \overline{z}}{s\sqrt{\frac{1}{p} + \frac{1}{q}}} = \frac{1.7430 - 1.435714}{\sqrt{0.092597857}\sqrt{\frac{1}{10} + \frac{1}{14}}} = 2.438935$$

whereas

$$\mathrm{TINV}(0,05;22) \qquad \Longrightarrow \qquad t_{[1-\frac{\alpha}{2}](p+q-2)} = t_{[0.975](22)} = 2.073873$$

so that the critical region takes the form

$$\Big\{t < -2.073873\Big\} \cup \Big\{t > 2.073873\Big\}.$$

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 5%. We conclude that *the different temperatures of the thermal treatment probably have an effect* on the electric resistivity of the alloy.

Remark. T-test in the case of unequal variances
If the variances σ_1^2 and σ_2^2 were different, the test variable would be

$$t = \frac{\overline{y} - \overline{z}}{\sqrt{\frac{1}{p}s_y^2 + \frac{1}{q}s_z^2}}$$

and for H_0 true it would follow *approximately* a Student's distribution with a number of d.o.f. given by

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

In this case we have

$$\overline{y} = 1.7430 \qquad \overline{z} = 1.435714 \qquad s_y^2 = 0.112068 \qquad s_z^2 = 0.079119$$

so that the number of d.o.f. of the test statistic, if H_0 holds true, turns out ot be

$$n = \frac{\left(\frac{0.112068}{10} + \frac{0.079119}{14}\right)^2}{\frac{1}{9}\left(\frac{0.112068}{10}\right)^2 + \frac{1}{13}\left(\frac{0.079119}{14}\right)^2} = 17.3169953\ ,$$

while the test statistic assumes the value

$$t = \frac{1.7430 - 1.435714}{\sqrt{\frac{1}{10}\cdot 0.112068 + \frac{1}{14}\cdot 0.079119}} = 2.366671\,.$$

The rejection region writes

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

with $\alpha = 0.05$ and $n = 17.3169953$, and therefore we need the t-value

$$t_{[1-\frac{\alpha}{2}](n)} = t_{[0.975](17.3169953)}$$

which obviously is not tabulated, as the number of d.o.f. is not an integer. The critical value can be calculated by using the Excel function TINV:

$$\mathrm{TINV}(0,05;17,3169953) \qquad \Longrightarrow \qquad t_{[0.975](17.3169953)} = 2.109816$$

which works also for a noninteger number of d.o.f. Alternatively, we can read on the table the critical values at $\alpha = 0.05$ for $n = 17$ and $n = 18$ d.o.f.

$$t_{[0.975](17)} = 2.110 \qquad\qquad t_{[0.975](18)} = 2.101$$

and apply a linear interpolation scheme:

17	2.110
17.3169953	$t_{[0.975](17.3169953)}$
18	2.101

$$\frac{17.3169953 - 17}{18 - 17} = \frac{t_{[0.975](17.3169953)} - 2.110}{2.101 - 2.110}$$

which provides the relationship

$$t_{[0.975](17.3169953)} = 2.110 + (2.101 - 2.110)\,\frac{17.3169953 - 17}{18 - 17}$$

and finally the approximate critical value

$$t_{[0.975](17.3169953)} = 2.107\ .$$

The critical region of H_0 becomes then

$$\left\{t \leq -2.109816\right\} \cup \left\{t \geq 2.109816\right\}$$

and *contains the value* $t = 2.366671$ *of the test statistics.* Therefore, the null hypothesis $H_0\ :\ \mu_1 = \mu_2$ must be refused, as before.