

Doctoral School of Materials Engineering
Methods of statistical and numerical analysis (integrated course). Part I
Final test - April 2nd 2009

Solution of Exercise 1

The electrical conductivity is given by the formula

$$\sigma = LI/abV$$

where:

$$L = (2.8 \pm 0.1) \cdot 10^{-3} \text{ m}$$

$$I = (5.120 \pm 0.002) \cdot 10^{-3} \text{ A}$$

$$a = (1.00 \pm 0.01) \cdot 10^{-6} \text{ m}$$

$$b = (4.00 \pm 0.01) \cdot 10^{-6} \text{ m}$$

$$V = (101 \pm 1) \cdot 10^{-3} \text{ V.}$$

The estimate of the conductivity is obtained by reckoning the relationship by the estimated true values \bar{L} , \bar{I} , \bar{a} , \bar{b} , \bar{V} of all the factors:

$$\bar{\sigma} = \frac{\bar{L}\bar{I}}{\bar{a}\bar{b}\bar{V}} = \frac{2.8 \cdot 10^{-3} \cdot 5.120 \cdot 10^{-3}}{1.00 \cdot 10^{-6} \cdot 4.00 \cdot 10^{-6} \cdot 101 \cdot 10^{-3}} = 0.0354851 \cdot 10^9 = 3.54851 \cdot 10^7.$$

The result is expressed in $\Omega^{-1} \cdot \text{m}^{-1}$. Since the function which defines σ is a simple polynomial, it is convenient to apply the logarithmic differential method to analyze the propagation of the *relative* error on $\bar{\sigma}$. We have indeed:

$$\ln \sigma = \ln L + \ln I - \ln a - \ln b - \ln V$$

and therefore, by computing the partial derivatives and replacing the absolute errors ΔL , ΔI , Δa , Δb , ΔV :

$$\frac{\Delta \sigma}{\bar{\sigma}} = \frac{\Delta L}{\bar{L}} + \frac{\Delta I}{\bar{I}} + \frac{\Delta a}{\bar{a}} + \frac{\Delta b}{\bar{b}} + \frac{\Delta V}{\bar{V}} = \frac{0.1}{2.8} + \frac{0.002}{5.120} + \frac{0.01}{1.00} + \frac{0.01}{4.00} + \frac{1}{101} = 0.058505901.$$

The greatest absolute error on the electrical conductivity holds then

$$\Delta \sigma = \frac{\Delta \sigma}{\bar{\sigma}} \cdot \bar{\sigma} = 0.058505901 \cdot 3.54851 \cdot 10^7 = 0.207608774 \cdot 10^7$$

in such a way that the error interval of σ becomes

$$\sigma = (\bar{\sigma} \pm \Delta \sigma) = (3.54851 \pm 0.207608774) \cdot 10^7 = (3.5485 \pm 0.2076) \cdot 10^7 \Omega^{-1} \cdot \text{m}^{-1}.$$

Since repeated measurements are lacking, the precision of the measurement of σ can be estimated by means of the percentage error:

$$\text{err}\% = \frac{\Delta \sigma}{\bar{\sigma}} \cdot 100 = 0.058505901 \cdot 100 = 5.85\%$$

which seems rather small.

Solution of Exercise 2

The computations needed to apply Chauvenet criterion are illustrated in the table below:

x	Δx	ABS(Δx)	Δx^2	
786,00	-3,58	3,58	12,84	
790,00	0,42	0,42	0,17	
792,00	2,42	2,42	5,84	
796,00	6,42	6,42	41,17	
784,00	-5,58	5,58	31,17	
791,00	1,42	1,42	2,01	
802,00	12,42	12,42	154,17	Outlier
789,00	-0,58	0,58	0,34	
783,00	-6,58	6,58	43,34	
785,00	-4,58	4,58	21,01	
795,00	5,42	5,42	29,34	
782,00	-7,58	7,58	57,51	
mean(x):	789,58	st.dev.(x):	6,0221	
Maximum of ABS(Δx):	12,4167	corresponding to the value of x:		Outlier 802,00
Distance z of the outlier from the mean in standard deviation units:				2,0619
Probability of a larger distance from the mean (see table):		z	area from 0 to z	residual area
		2,0600	0,48030	0,03940
		2,0700	0,48077	0,03846
Linearly interpolated value:		2,0619	0,48039	0,03922
Mean number of expected events out of 12 measurements:		0,4707		

The mean number of outliers is smaller than 1/2. Thus, according to Chauvenet criterion, the outlier 802 must be rejected as not belonging to the population.

The sample estimates of the mean and standard deviation are immediate:

$$\bar{x} = \frac{1}{12} \sum_{i=1}^{12} x_i = 789.58 \quad s = \sqrt{\frac{1}{11} \sum_{i=1}^{12} (x_i - \bar{x})^2} = 6.0221.$$

The datapoint farthest from the sample mean \bar{x} is clearly $x_{\text{sus}} = 802$, as shown by the ABS(Δx) column in the previous table. It is this outlier that could not belong to the statistical population of the normal data. The distance of the suspect value from the mean, in units of s , is given by

$$z = \frac{x_{\text{sus}} - \bar{x}}{s} = \frac{802 - 789.58}{6.0221} = 2.0619.$$

The probability that a datapoint is placed at a distance greater than 2.0619 standard deviations from the mean can be calculated from the table of the cumulative distribution

of the standard normal random variable:

$$\begin{aligned} P(|x_{\text{sus}} - \bar{x}| \geq 2.0619s) &= 1 - P(|x_{\text{sus}} - \bar{x}| < 2.0619s) = \\ &= 1 - 2 \cdot P(\bar{x} \leq x_{\text{sus}} < \bar{x} + 2.0619s) \\ &= 1 - 2 \cdot 0.48039 = 0.03922. \end{aligned}$$

Notice that the probability $P = P(\bar{x} \leq x_{\text{sus}} < \bar{x} + 2.0619s)$ is not directly available on the table, but can be estimated with sufficient accuracy by a linear interpolation scheme:

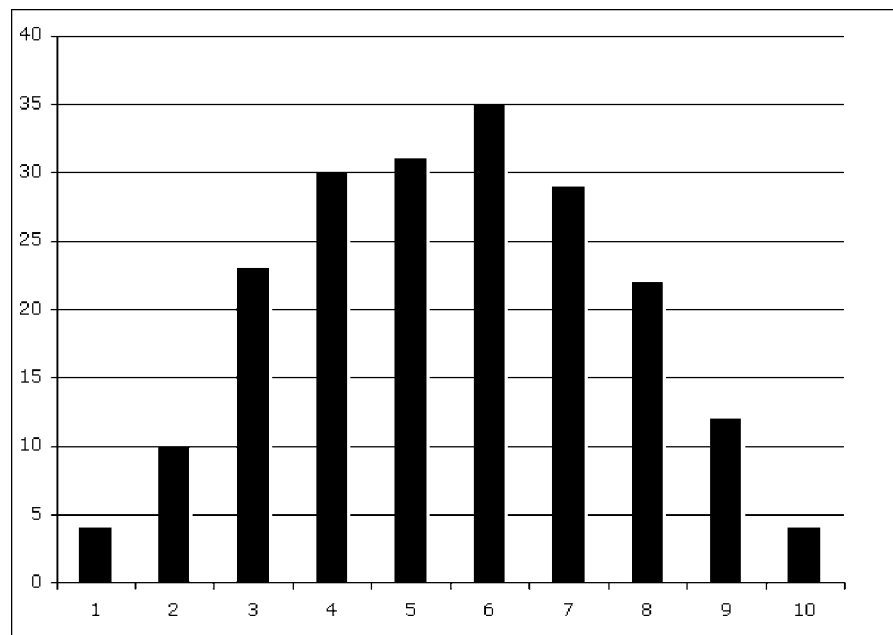
2.0600	0.48030	
2.0619	P	
2.0700	0.48077	$\frac{2.0619 - 2.0600}{2.0700 - 2.0600} = \frac{P - 0.48030}{0.48077 - 0.48030}$

which provides $P = 0.48039$.

Out of 12 measurements, typically we expect $12 \cdot 0.03922 = 0.4707$ outliers at a distance larger than $2.0619s$ from the mean. *Since $0.4707 < 1/2$, Chauvenet criterion suggests that x_{sus} must be rejected as not belonging to the statistical population.*

Solution of Exercise 3

The sample histogram shows a bell-shaped trend, so that it is rather reasonable to suppose that the data belong to a normal population:



All the empirical frequencies are relatively high ($f_i \geq 3$), what allows us to apply the χ^2 test to check whether the population is normal, i.e. the null hypothesis

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

versus the alternative hypothesis

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

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

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

and the classes of the results (the histogram intervals) are $k = 10$ in all. In the presence of $c = 2$ constraints on the mean and the standard deviation, if H_0 holds true the \mathcal{X}^2 of data follows approximately a \mathcal{X}^2 distribution with

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

degrees of freedom. To calculate the \mathcal{X}^2 , we firstly need the expected frequencies in each class, under the assumption that the normal distribution is correct. The endpoints of the classes differ from the mean $\mu = \bar{m}$ by half-integer multiples of $\sigma = s$, so that the theoretical frequencies can be derived directly from the cumulative distribution of a standard normal random variable. For simplicity's sake, it is 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 n_i the frequency in the i -th class, the expected frequencies are the following:

$$\begin{aligned} n_1 &= 200 \cdot \int_{-\infty}^{-2} p(z) dz = 200 \cdot \int_2^{+\infty} p(z) dz = 200 \cdot \left(\frac{1}{2} - \int_0^2 p(z) dz \right) = \\ &= 200 \cdot \left(\frac{1}{2} - \Phi(2) \right) = 200 \cdot \left(\frac{1}{2} - 0.47725 \right) = 4.550 \\ n_2 &= 200 \cdot \int_{-2}^{-1.5} p(z) dz = 200 \cdot \int_{1.5}^2 p(z) dz = 200 \cdot \left(\Phi(2) - \Phi(1.5) \right) = \\ &= 200 \cdot (0.47725 - 0.43319) = 8.812 \\ n_3 &= 200 \cdot \int_{-1.5}^{-1} p(z) dz = 200 \cdot \int_1^{1.5} p(z) dz = 200 \cdot \left(\Phi(1.5) - \Phi(1) \right) = \\ &= 200 \cdot (0.43319 - 0.34134) = 18.370 \end{aligned}$$

$$\begin{aligned}
n_4 &= 200 \cdot \int_{-1}^{-0.5} p(z) dz = 200 \cdot \int_{0.5}^1 p(z) dz = 200 \cdot (\Phi(1) - \Phi(0.5)) = \\
&= 200 \cdot (0.34134 - 0.19146) = 29.976 \\
n_5 &= 200 \cdot \int_{-0.5}^0 p(z) dz = 200 \cdot \int_0^{-0.5} p(z) dz = 200 \cdot \Phi(0.5) = \\
&= 200 \cdot 0.19146 = 38.292
\end{aligned}$$

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

$$\begin{aligned}
n_6 &= n_5 = 38.292 & n_7 &= n_4 = 29.976 & n_8 &= n_3 = 18.370 \\
n_9 &= n_2 = 8.812 & n_{10} &= n_1 = 4.550.
\end{aligned}$$

We can then compare the empirical frequencies f_i and the expected ones n_i for all the classes, as summarized in the table below:

i	class of E	empirical frequency	expected frequency
1	$m < \bar{m} - 2.0s$	4	4.550
2	$\bar{m} - 2.0s \leq m < \bar{m} - 1.5s$	10	8.812
3	$\bar{m} - 1.5s \leq m < \bar{m} - 1.0s$	23	18.370
4	$\bar{m} - 1.0s \leq m < \bar{m} - 0.5s$	30	29.976
5	$\bar{m} - 0.5s \leq m < \bar{m}$	31	38.292
6	$\bar{m} \leq m < \bar{m} + 0.5s$	35	38.292
7	$\bar{m} + 0.5s \leq m < \bar{m} + 1.0s$	29	29.976
8	$\bar{m} + 1.0s \leq m < \bar{m} + 1.5s$	22	18.370
9	$\bar{m} + 1.5s \leq m < \bar{m} + 2.0s$	12	8.812
10	$\bar{m} + 2.0s \leq m$	4	4.550

The χ^2 of the sample is given by:

$$\chi^2 = \sum_{i=1}^{10} \frac{(f_i - n_i)^2}{n_i} = 5.0342.$$

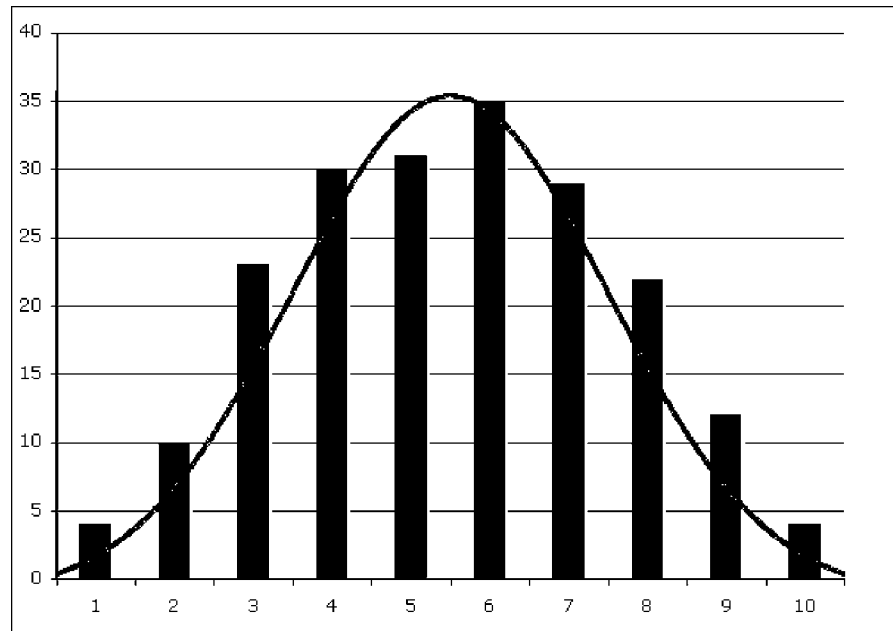
The table of the χ^2 cumulative distributions provides the critical values:

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

$$\chi^2_{[1-\alpha](7)} = \chi^2_{[0.99](7)} = 18.475 \quad \text{for } \alpha = 0.01.$$

In both cases the χ^2 of the sample is smaller: 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 Young modulus of the polymer is normal*.

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



Pay attention to the way the theoretical distribution is calculated (solid curve). In the histogram the class 2 is centred at $x = 2$ and the class 9 at $x = 9$. The centre of the class 2 must correspond to the point

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

and that of the class 9 to the point

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

so that the parameters μ and σ of the normal distribution are determined by the linear equations

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

which provide

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

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

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

Solution of Exercise 4

Since $n = 400 > 30$, we can apply the theory of large samples and it is not required that the statistical population is normal.

(a) The confidence level is $1 - \alpha = 0.67$, so that $\alpha = 0.33$ and

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

From the table of the standard normal distribution we derive the following linear interpolation scheme:

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

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

which provides the critical values

$$z_{0.33} = 0.9741.$$

The confidence interval for the mean μ of the weight of the bolts turns out to be

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

with the sample mean and standard deviation given by

$$\bar{x} = 7.47 \quad s = 0.15$$

while $n = 400$. By inserting all the numbers, we obtain therefore

$$7.47 - 0.9741 \cdot \frac{0.15}{\sqrt{400}} \leq \mu \leq 7.47 + 0.9741 \cdot \frac{0.15}{\sqrt{400}}$$

and finally

$$7.4627 \text{ g} \leq \mu \leq 7.4773 \text{ g}.$$

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

$$\mu = (7.4700 \pm 0.0073) \text{ g}.$$

(b) In this case the confidence level holds $1 - \alpha = 0.99$, so that $\alpha/2 = 0.005$ and

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

From the table of the standard normal distribution we get then, to a good approximation,

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

since

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

The CI of the mean becomes

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

i.e.

$$7.47 - 2.58 \cdot \frac{0.15}{\sqrt{400}} \leq \mu \leq 7.47 + 2.58 \cdot \frac{0.15}{\sqrt{400}}$$

and finally

$$7.4507 \text{ g} \leq \mu \leq 7.4894 \text{ g}.$$

An alternative expression puts into evidence the absolute error:

$$\mu = (7.470 \pm 0.0194) \text{ g}.$$

The confidence interval with confidence level of 99% is obviously larger than that at confidence level of 67%.

Solution of Exercise 5

The number of the sample data is $n = 22$, and therefore the sample mean turns out to be

$$\bar{V} = \frac{1}{n} \sum_{i=1}^n V_i = 1.5715045.$$

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

i	V_i	$(V_i - \bar{V}) \cdot 10^2$	$(V_i - \bar{V})^2 \cdot 10^4$
1	1.5710	-0.0504545	0.0025457
2	1.5742	0.2695455	0.0726548
3	1.5751	0.3595455	0.1292729
4	1.5686	-0.2904545	0.0843638
5	1.5712	-0.0304545	0.0009275
6	1.5734	0.1895455	0.0359275
7	1.5682	-0.3304545	0.1092002
8	1.5718	0.0295455	0.0008729
9	1.5737	0.2195455	0.0482002
10	1.5757	0.4195455	0.1760184
11	1.5668	-0.4704545	0.2213275
12	1.5703	-0.1204545	0.0145093
13	1.5729	0.1395455	0.0194729
14	1.5668	-0.4704545	0.2213275
15	1.5725	0.0995455	0.0099093
16	1.5696	-0.1904545	0.0362729
17	1.5742	0.2695455	0.0726548
18	1.5711	-0.0404545	0.0016366
19	1.5729	0.1395455	0.0194729
20	1.5711	-0.0404545	0.0016366
21	1.5689	-0.2604545	0.0678366
22	1.5731	0.1595455	0.0254548

from which we deduce the sample variance:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (V_i - \bar{V})^2 = 0.0653093 \cdot 10^{-4}$$

and the sample estimate of the standard deviation:

$$s = \sqrt{s^2} = 0.2555569 \cdot 10^{-2}.$$

The sample 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 .

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

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

and for $\alpha = 0.10$, $n = 22$ has therefore the limits

$$\begin{aligned} \bar{V} - t_{[0.95](21)} \frac{s}{\sqrt{22}} &= 1.5715045 - 1.721 \cdot \frac{0.002555569}{\sqrt{22}} = 1.5705669 \\ \bar{V} + t_{[0.95](21)} \frac{s}{\sqrt{22}} &= 1.5715045 + 1.721 \cdot \frac{0.002555569}{\sqrt{22}} = 1.5724422 \end{aligned}$$

so that the confidence interval becomes

$$1.5705669 \text{ V} \leq \mu \leq 1.5724422 \text{ V}$$

or, equivalently,

$$\mu = (1.5715045 \pm 0.0009377) \text{ V}.$$

It is understood that, for all practical purposes, an approximation of the type

$$(1.57150 \pm 0.00094) \text{ V}$$

can be considered more than satisfactory.

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

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

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

$$\begin{aligned} \frac{1}{\chi^2_{[0.95](21)}} 21 s^2 &= \frac{1}{32.671} 21 \cdot 0.0653093 \cdot 10^{-4} = 4.19790 \cdot 10^{-6} \\ \frac{1}{\chi^2_{[0.05](21)}} 21 s^2 &= \frac{1}{11.591} 21 \cdot 0.0653093 \cdot 10^{-4} = 11.83242 \cdot 10^{-6} \end{aligned}$$

and the CI of the variance is expressed as

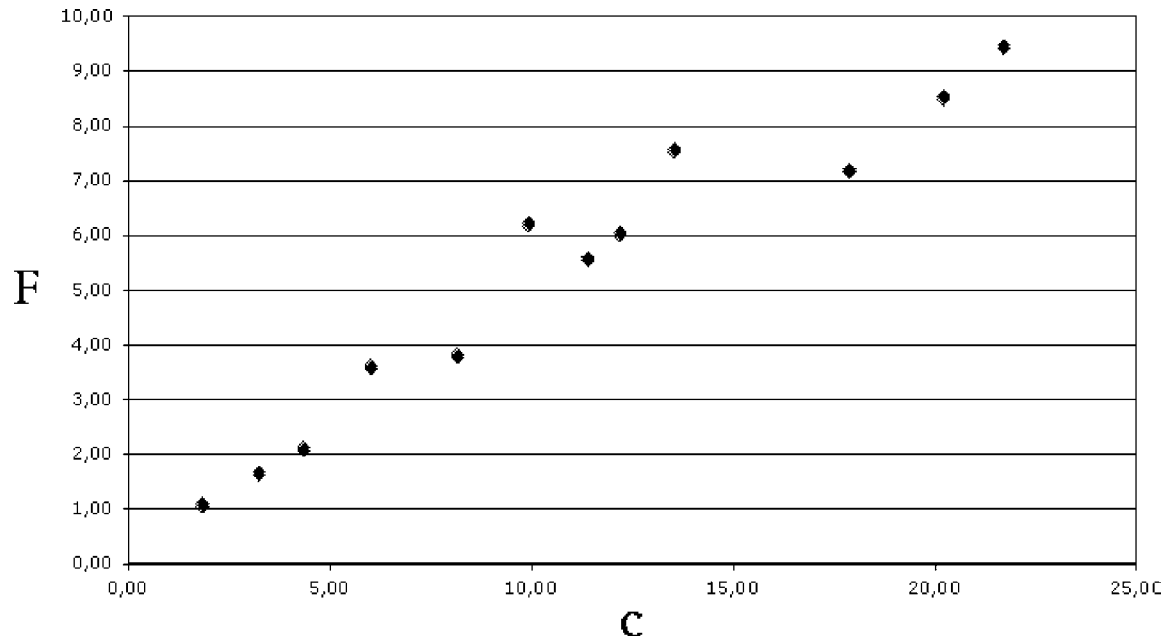
$$4.19790 \cdot 10^{-6} \text{ V}^2 \leq \sigma^2 \leq 11.83242 \cdot 10^{-6} \text{ V}^2.$$

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

$$2.048877 \cdot 10^{-3} \text{ V} \leq \sigma \leq 3.439828 \cdot 10^{-3} \text{ V}.$$

Solution of Exercise 6

The plot of the data suggests that the quantities c and F are described by dependent random variables (i.e., correlated, owing to the hypothesis of normal random variables):



The sample means \bar{c} and \bar{F} of the two quantities are given by:

$$\bar{c} = \frac{1}{12} \sum_{i=1}^{12} c_i = 10.873333 \quad \bar{F} = \frac{1}{12} \sum_{i=1}^{12} F_i = 5.245833$$

and allow us to calculate the sum of products of residuals

$$SS_{cF} = \sum_{i=1}^{12} (c_i - \bar{c})(F_i - \bar{F}) = 193.762967$$

and of the relative squares:

$$SS_{cc} = \sum_{i=1}^{12} (c_i - \bar{c})^2 = 478.332867$$

$$SS_{FF} = \sum_{i=1}^{12} (F_i - \bar{F})^2 = 83.992492,$$

as illustrated in the following table:

c	F	Dc	DF	Dc ²	DF ²	Dc * DF
20,23	8,53	9,356667	3,284167	87,547211	10,785751	30,728853
8,17	3,82	-2,703333	-1,425833	7,308011	2,033001	3,854503
6,01	3,62	-4,863333	-1,625833	23,652011	2,643334	7,906969
9,93	6,23	-0,943333	0,984167	0,889878	0,968584	-0,928397
13,55	7,57	2,676667	2,324167	7,164544	5,401751	6,221019
17,89	7,20	7,016667	1,954167	49,233611	3,818767	13,711736
12,19	6,05	1,316667	0,804167	1,733611	0,646684	1,058819
1,83	1,10	-9,043333	-4,145833	81,781878	17,187934	37,492153
21,72	9,46	10,846667	4,214167	117,650178	17,759201	45,709661
11,39	5,59	0,516667	0,344167	0,266944	0,118451	0,177819
3,23	1,66	-7,643333	-3,585833	58,420544	12,858201	27,407719
4,34	2,12	-6,533333	-3,125833	42,684444	9,770834	20,422111
10,873333	5,245833			478,332867	83,992492	193,762967
c mean	F mean					
				r=	0,966686	
				t=	11,942756	

The linear correlation coefficient becomes

$$r = \frac{SS_{cF}}{\sqrt{SS_{cc}}\sqrt{SS_{FF}}} = \frac{193.762967}{\sqrt{478.332867}\sqrt{83.992492}} = 0.966686.$$

As both c and F may be assumed normal, we can check the null hypothesis

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

against the alternative hypothesis

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

by using the random variable

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

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

$$t = \sqrt{12-2} \frac{0.966686}{\sqrt{1-0.966686^2}} = 11.942756.$$

The critical region takes the form

$$|t| > t_{[1-\frac{\alpha}{2}]}(n-2) = t_{[0.975]}(10) = 2.228$$

for a significance level $\alpha = 5\%$, while it becomes

$$|t| > t_{[1-\frac{\alpha}{2}]}(n-2) = t_{[0.995]}(10) = 3.169$$

when the requested significance level is $\alpha = 1\%$. *In both cases H_0 must be rejected!* We conclude therefore that a correlation presumably exists between the degree of crystallinity c and the resistance to traction F ; due to the positive sign of the correlation coefficient, which is very close to $+1$, the relation must be direct.

Solution of Exercise 7

The temperature data are not affected by appreciable random errors, whereas the relative values of resistivity are the outcomes of independent normal random variables. It is then possible to apply the standard theory of linear regression, with the further simplification due to the homoscedastic character of the model — we may assume that all the random variables which describe the resistivity at different temperatures share the same variance. That is way the regression straight line is calculated by putting the temperature T along the abscissa axis and the resistivity ρ along the ordinate axis:

$$\rho = \mu + \kappa(T - \bar{T})$$

where \bar{T} denotes the arithmetic mean of the measured temperatures, while μ and κ are the parameters of the regression model. We recall that a model of this form ensures the stochastic independence of the best-fit estimates m and q of the regression parameters μ and κ .

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

(i) Regression straight line

Since the standard deviations are equal, the χ^2 fitting reduces to the usual least-squares fitting and the best-fit estimates m and q of the parameters can be written as

$$m = \frac{1}{n} \sum_{i=1}^n \rho_i = 2.815384615 \quad q = \frac{\sum_{i=1}^n (T_i - \bar{T}) \rho_i}{\sum_{i=1}^n (T_i - \bar{T})^2} = 0.002273762022$$

with $n = 26$ and $\bar{T} = 956.1538461$. The regression straight line, calculated by the least-squares method, is therefore:

$$\begin{aligned} \rho &= m + q(T - \bar{T}) = 2.815384615 + 0.002273762022(T - 956.1538461) = \\ &= 0.641318312 + 0.002273762022 T . \end{aligned}$$

(ii) *Confidence intervals for the regression parameters*

By definition, the sum of squares around regression holds:

$$\text{SSAR} = \sum_{i=1}^n [m + q(T_i - \bar{T}) - \rho_i]^2 = 0.2809376284.$$

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

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

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

In the present case we have $\alpha = 0.05$, $n = 26$ and the confidence intervals become then:

$$\mu = m \pm t_{[0.975](24)} \sqrt{\frac{1}{26} \frac{\text{SSAR}}{24}}$$

$$\kappa = q \pm t_{[0.975](24)} \sqrt{\left[\sum_{i=1}^{26} (T_i - \bar{T})^2 \right]^{-1} \frac{\text{SSAR}}{24}}$$

where:

$$m = 2.815384615$$

$$q = 0.002273762022$$

$$\text{SSAR} = 0.2809376284$$

$$\sum_{i=1}^{26} (T_i - \bar{T})^2 = 3.034415385 \cdot 10^6$$

$$t_{[0.975](24)} = 2.064.$$

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

– the CI at 95% of the parameter μ is

$$2.815384615 \pm 0.04379478051 = [2.771589834, 2.859179396]$$

– the CI at 95% for the slope κ holds

$$0.002273762022 \pm 0.0001281951256 = [0.002145566898, 0.002401957150].$$

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

$$\mu = [2.7716, 2.8592] 10^{-7} \Omega \cdot \text{m} = (2.8154 \pm 0.0438) 10^{-7} \Omega \cdot \text{m}$$

while

$$\begin{aligned} \kappa &= [2.145567, 2.401957] 10^{-3} 10^{-7} \Omega \cdot \text{m} \cdot \text{K}^{-1} = \\ &= (2.273762 \pm 0.128195) 10^{-10} \Omega \cdot \text{m} \cdot \text{K}^{-1} . \end{aligned}$$

(iii) *Confidence region*

The model being homoscedastic, the CI at a confidence level $1 - \alpha$ for the prediction of $\rho = \rho_0$ at a given $T = T_0$ can be calculated by the general formula:

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

where, more specifically, we have:

$$m = 2.815384615$$

$$q = 0.002273762022$$

$$\bar{T} = \frac{1}{n} \sum_{i=1}^n T_i = 956.1538461$$

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

$$V = 1 + \frac{1}{n} + \frac{1}{\sum_{i=1}^n (T_i - \bar{T})^2} (T_0 - \bar{T})^2 = 1 + \frac{1}{26} + \frac{(T_0 - 956.1538461)^2}{3.034415385 \cdot 10^6} =$$

$$= 1.038461538 + 3.295527715 \cdot 10^{-7} (T_0 - 956.1538461)^2$$

$$\text{SSAR} = \sum_{i=1}^n [m + q(T_i - \bar{T}) - \rho_i]^2 = 0.2809376284 .$$

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

$$\rho_0 = 2.815384615 + 0.002273762022 (T_0 - 956.1538461) \pm$$

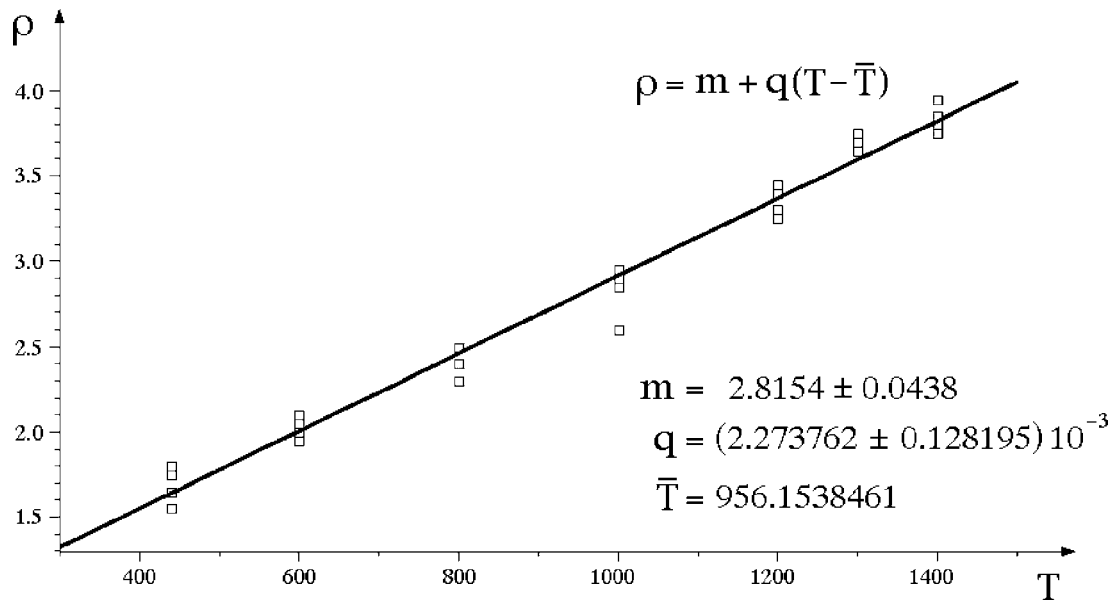
$$\pm 2.064 \cdot \sqrt{1.038461538 + 3.295527715 \cdot 10^{-7} (T_0 - 956.1538461)^2} \sqrt{\frac{0.2809376284}{24}}$$

and performing the calculations reduces to:

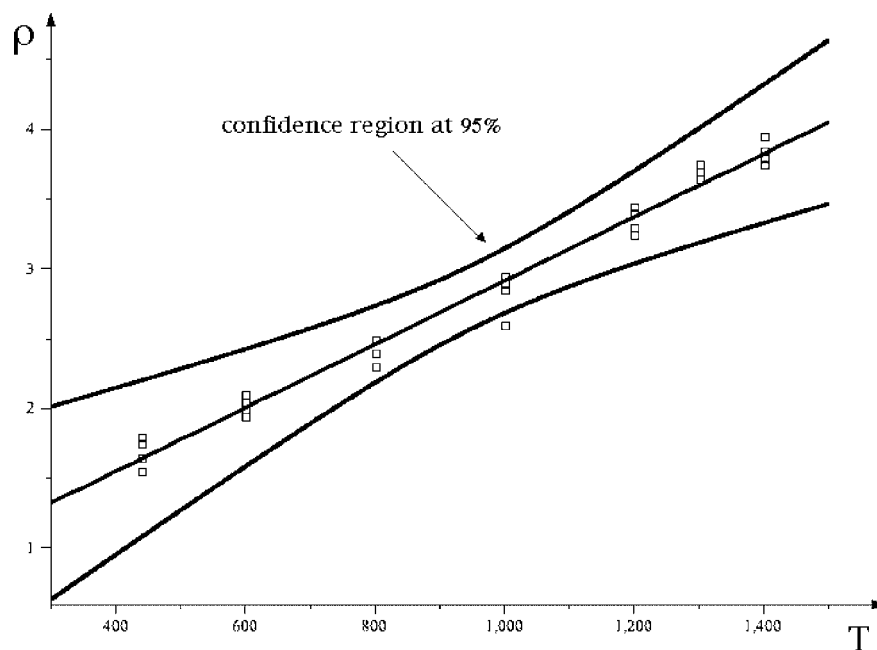
$$\rho_0 = 0.641318312 + 0.002273762022 T_0 \pm$$

$$\pm 0.2233104403 \sqrt{1.038461538 + 3.295527715 \cdot 10^{-7} (T_0 - 956.1538461)^2}$$

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 regression straight line of the model is marked in red, while in blue and in black are represented the upper and lower boundary of the confidence region, respectively. The width, measured parallel to the ρ axis, of the confidence region is minimum for $T = \bar{T} =$

956.15 and tends to increase monotonically on the right and on the left of that point. To better stress the effect on the graph, the term $(T_0 - \bar{T})^2$ which appears in the expression of V of the definition, has been enlarged by a scale factor 60.

(iv) *Confidence interval for a prediction*

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

$$\rho_0 = 0.641318312 + 0.002273762022 T_0 \pm 0.2233104403 \sqrt{1.038461538 + 3.295527715 \cdot 10^{-7} (T_0 - 956.1538461)^2} .$$

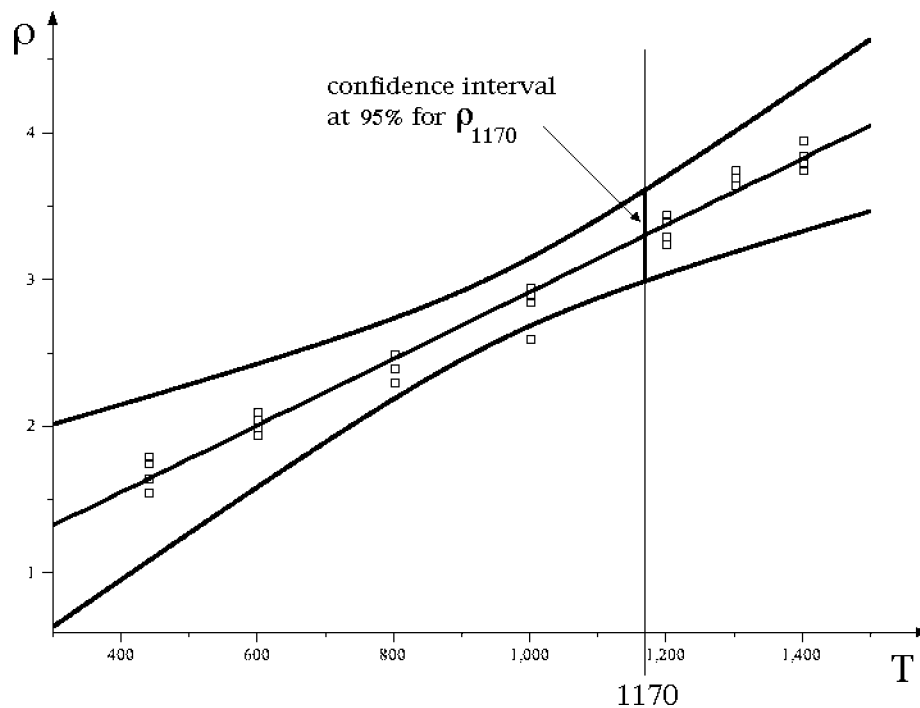
We have therefore:

$$\rho_0 = \rho_{1170} = [3.072410226, 3.530829532] = 3.301619879 \pm 0.229209653$$

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

$$\rho_{1170} = [3.07, 3.53] 10^{-7} \Omega \cdot \text{m} = (3.30 \pm 0.23) 10^{-7} \Omega \cdot \text{m} .$$

In the following figure the CI at 95% is simply the intersection of the confidence region at 95% with the vertical straight line of equation $T = 1170$:



(v) *Goodness of fit*

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

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

where ρ_{n-2} stands for the χ^2 distribution with $n-2$ d.o.f. This is because, if the regression model is correct, the normalized sum of squares around regression

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

constitutes a χ^2 random variable with $n-2$ d.o.f. To evaluate the goodness of fit *it is crucial to know the common value of the standard deviation* $\sigma = 0.09$, since we need to determine the NSSAR, and not simply the SSAR. In the present case we have $n = 26$ data and the regression model is based on two parameters, μ and κ ; as a consequence, the NSSAR obeys a χ^2 distribution with $n-2 = 24$ d.o.f. For the given sample the normalized sum of squares around regression holds

$$\text{NSSAR} = \frac{\text{SSAR}}{\sigma^2} = \frac{0.2809376284}{0.09^2} = 34.68365782.$$

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

Probability $\{\chi^2 \geq 33.196\}$	Probability $\{\chi^2 \geq 36.415\}$
0.10	0.05

so that a simple linear interpolation scheme:

33.196	0.10	$\frac{34.6836 - 33.196}{36.415 - 33.196} = \frac{Q - 0.10}{0.05 - 0.10}$
34.6836	Q	
36.415	0.05	

provides the required estimate of Q :

$$Q = 0.10 + (0.05 - 0.10) \frac{34.6836 - 33.196}{36.415 - 33.196} = 0.07689344517.$$

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

$$Q = \text{Probabilità}\{\chi^2 \geq 34.68365782\} = \int_{34.68365782}^{+\infty} p_{24}(\chi^2) d\chi^2$$

for instance by using the Maple command line

$$1 - \text{stats}[\text{statevalf}, \text{cdf}, \text{chisquare}[24]](34.68365782);$$

which leads to the “exact” value $Q = 0.0732313109$. The goodness of fit of the regression model is thus equal to about 7.3%, a value that, if the regression model were rejected, would express the probability of a type I error.

Solution of Exercise 8

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, \dots, y_p of the first sample and $q = 14$ that of the data z_1, \dots, 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 of 5%.

Since, by hypothesis, the populations can be assumed to be normal but the variances are not necessarily equal, the test variable is

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

and for H_0 true follows *approximately* a Student distribution with a number of d.o.f. equal to

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

In this case we get

$$\bar{y} = \frac{1}{p} \sum_{i=1}^p y_i = 5.572727 \quad \bar{z} = \frac{1}{q} \sum_{j=1}^q z_j = 6.362143$$

$$s_y^2 = \frac{1}{p-1} \sum_{i=1}^p (y_i - \bar{y})^2 = 0.437422$$

$$s_z^2 = \frac{1}{q-1} \sum_{j=1}^q (z_j - \bar{z})^2 = 0.916049$$

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

$$n = \frac{\left(\frac{0.437422}{11} + \frac{0.916049}{14}\right)^2}{\frac{1}{10}\left(\frac{0.437422}{11}\right)^2 + \frac{1}{13}\left(\frac{0.916049}{14}\right)^2} = 22.7022297 .$$

The rejection region writes

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

with $\alpha = 0.05$ and $n = 22.7022297$, and therefore we need the t -value

$$t_{[1-\frac{\alpha}{2}]}(n) = t_{[0.975]}(22.7022297)$$

which obviously is not tabulated, as the number of d.o.f. is not an integer. From the table we get, however,

$$t_{[0.975]}(22) = 2.074 \quad t_{[0.975]}(23) = 2.069$$

and thus we can apply a linear interpolation scheme

22	2.074
22.7022297	$t_{[0.975]}(22.7022297)$
23	2.069

$$\frac{22.7022297 - 22}{23 - 22} = \frac{t_{[0.975]}(22.7022297) - 2.074}{2.069 - 2.074}$$

which provides the relationship

$$t_{[0.975]}(22.7022297) = 2.074 + (2.069 - 2.074) \frac{22.7022297 - 22}{23 - 22}$$

and finally the required critical value

$$t_{[0.975]}(22.7022297) = 2.0705 .$$

The critical region of H_0 becomes then

$$\{t \leq -2.0705\} \cup \{t \geq 2.0705\} .$$

On the other hand, the test variable takes the value

$$t = \frac{5.572727 - 6.362143}{\sqrt{\frac{1}{11} \cdot 0.437422 + \frac{1}{14} \cdot 0.916049}} = -2.4339$$

which *falls within the rejection region of H_0* . The values of electrical conductivity which can be obtained by the two temperatures of treatment are *significantly different*, at the significance level of 5%.

The detailed calculations are illustrated in the table below:

	y _i	z _i	Dy	Dy ²	Dz	Dz ²
	5,52	6,33	-0,052727	0,002780	-0,032143	0,001033
	6,30	5,89	0,727273	0,528926	-0,472143	0,222919
	5,00	5,25	-0,572727	0,328017	-1,112143	1,236862
	5,80	7,70	0,227273	0,051653	1,337857	1,789862
	6,45	7,66	0,877273	0,769607	1,297857	1,684433
	4,38	5,44	-1,192727	1,422598	-0,922143	0,850347
	5,25	7,10	-0,322727	0,104153	0,737857	0,544433
	6,02	7,48	0,447273	0,200053	1,117857	1,249605
	4,77	5,95	-0,802727	0,644371	-0,412143	0,169862
	5,68	5,05	0,107273	0,011507	-1,312143	1,721719
	6,13	6,47	0,557273	0,310553	0,107857	0,011633
		6,10			-0,262143	0,068719
		5,24			-1,122143	1,259205
		7,41			1,047857	1,098005
mean of y =	5,572727					
mean of z =	6,362143					
var. of y =	0,437422					
var. of z =	0,916049					
num d.o.f. =	0,0110666					
den d.o.f. =	0,0004875					
d.o.f. =	22,7022297					
critical t =	2,0705					
sample t =	-2,4339					

Remark

In the previous calculations we have assumed that the variances of the two statistical populations are unequal. In principle, we could test such an assumption by using an F-test. The test statistics is

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

and the null hypothesis $H_0 : \sigma_1^2 = \sigma_2^2$ 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 this case the value of F for our samples is

$$F = \frac{0.437422}{0.916049} = 0.477509$$

whereas $p = 11$, $q = 14$ and $\alpha = 0.05$. Therefore, the acceptance region becomes

$$0.279081 = F_{[0.025](10,13)} < F < F_{[0.975](10,13)} = 3.249668$$

since

$$F_{[0.025](10,13)} = 0.279081 \quad F_{[0.975](10,13)} = 3.249668.$$

The above critical values are not available on the table of the Fisher cumulative distributions and should be calculated by appropriate numerical tools. In particular, the following Maple commands are applicable:

$$\text{stateval}[icdf, fratio[10, 13]](0.025)$$

$$\text{stateval}[icdf, fratio[10, 13]](0.975)$$

which provide $F_{[0.025](10,13)}$ and $F_{[0.975](10,13)}$, respectively. An alternative way to carry out the calculation makes use of the Excel function FINV, which gives the inverse of the Fisher cumulative distribution. *Pay attention, however, that the Excel definition of the Fisher cumulative distribution with (n_1, n_2) d.o.f. takes the form:*

$$P_{n_1, n_2}^{\text{Excel}}(F) = \int_F^{+\infty} p_{n_1, n_2}(F) dF$$

and differs from the usual one:

$$P_{n_1, n_2}(F) = \int_0^F p_{n_1, n_2}(F) dF$$

so that $P_{n_1, n_2}^{\text{Excel}}(F) = 1 - P_{n_1, n_2}(F)$. As a consequence, the critical values $F_{[0.025](10,13)}$ and $F_{[0.975](10,13)}$ that we need are obtained by digiting in an Excel worksheet cell the commands

$$= \text{FINV}(0, 975; 10; 13) \quad = \text{FINV}(0, 025; 10; 13)$$

respectively: *the p -probability values 0.025 and 0.975 are interchanged.*

However the calculation is made, since the value of the F statistics falls within the acceptance region:

$$0.279081 < 0.477509 < 3.249668$$

we should conclude that *the variances σ_1^2 and σ_2^2 are equal* at the significance level $\alpha = 5\%$. As a conclusion, *a rigorous unpaired t -test with equal variances could be applied.*

Solution of Exercise 9

In this case it seems natural to apply a paired t -test for the comparison of the means, because the quantity is measured prior to and after the treatment *on each sample*. Therefore, values relative to the same sample will be coupled:

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

denoting with y_i the values measured before the treatment and with z_i those measured after the treatment, on all the $n = 8$ samples. We check the hypothesis $H_0 : \mu_1 = \mu_2$,

that the treatment has no effect on the mean value of the measured quantity, versus the alternative hypothesis $H_1 : \mu_1 \neq \mu_2$ that the claim is not true. The test variable is

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

which, for H_0 true, follows a Student distribution with $n - 1 = 7$ d.o.f. The critical region, with significance level α , is of the form

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

and for $n = 8$, $\alpha = 0.02$ becomes

$$\{t \leq -2.998\} \cup \{t \geq 2.998\}$$

since

$$t_{[1-\frac{\alpha}{2}](n-1)} = t_{[0.99](7)} = 2.998.$$

In the present case, we obtain:

$$\bar{y} = \frac{1}{8} \sum_{i=1}^8 y_i = 0.8438 \quad \bar{z} = \frac{1}{8} \sum_{i=1}^8 z_i = 1.0188$$

while

$$\frac{1}{n-1} \sum_{i=1}^n (y_i - z_i - \bar{y} + \bar{z})^2 = \frac{1}{7} \cdot 0.219600 = 0.03137143$$

and therefore

$$\sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - z_i - \bar{y} + \bar{z})^2} = \sqrt{0.03137143} = 0.1771198$$

so that the test variable takes the value

$$t = \sqrt{8} \cdot \frac{0.8438 - 1.0188}{0.1771198} = -2.7946.$$

Here is the table of the detailed calculations:

y_i	z_i	$y_i - z_i$	$d_i = y_i - z_i - \text{mean}(y.z)$	d_i^2
1,02	1,16	-0,14	0,0350	0,001225
0,95	1,04	-0,09	0,0850	0,007225
0,73	0,85	-0,12	0,0550	0,003025
0,96	1,10	-0,14	0,0350	0,001225
0,79	1,09	-0,30	-0,1250	0,015625
1,01	0,93	0,08	0,2550	0,065025
0,42	0,95	-0,53	-0,3550	0,126025
0,87	1,03	-0,16	0,0150	0,000225
mean(y-z):	-0,1750		st.dev.(y-z):	0,177120
Student t =	-2,7946			
mean(y) =	0,8438			
mean(z) =	1,0188			
		var.(y-z):	0,03137143	
		st.dev.(y-z):	0,1771198	

The calculated value does not belong to the rejection region, because

$$-2.998 < -2.7946 < 2.998,$$

and therefore *we cannot exclude*, with a significance level of 2%, *that the mean value of the quantity is the same* before and after the treatment. *The null hypothesis cannot be rejected.*