

**Graduate School of Materials Engineering - a.y. 2010-2011**  
**Methods of statistical and numerical analysis (integrated course). Part I**  
**Final test - April 1st 2011**

□ **Exercise 1** **Points 4**

A random sample of 250 bolts produced by an automatic machine is collected. The weight of the bolts has a sample mean of 15.50 g, with a standard deviation of 0.45 g. Determine:

- (a) the confidence interval for the weight of a bolt, at a confidence level of 80%;
- (b) the same as before, but at a confidence level of 95%.

□ **Exercise 2** **Points 4**

Repeated measurements of the electrical conductivity  $\kappa$  of a HCl solution in water gave the results listed below (in  $\Omega^{-1} \cdot \text{cm}^{-1}$ ):

0.6237	0.6342	0.5702	0.6102	0.6223	0.5575
0.5329	0.5584	0.6212	0.5633	0.7020	0.6438

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

□ **Exercise 3** **Points 3**

The equivalent resistance of two resistors in parallel is expressed by the formula

$$R = \frac{R_1 R_2}{R_1 + R_2}$$

with resistances  $R_1 = (200 \pm 1) \Omega$  and  $R_2 = (100 \pm 1) \Omega$ . Determine the estimated value of  $R$  along with the appropriate absolute error.

□ **Exercise 4****Points 4**

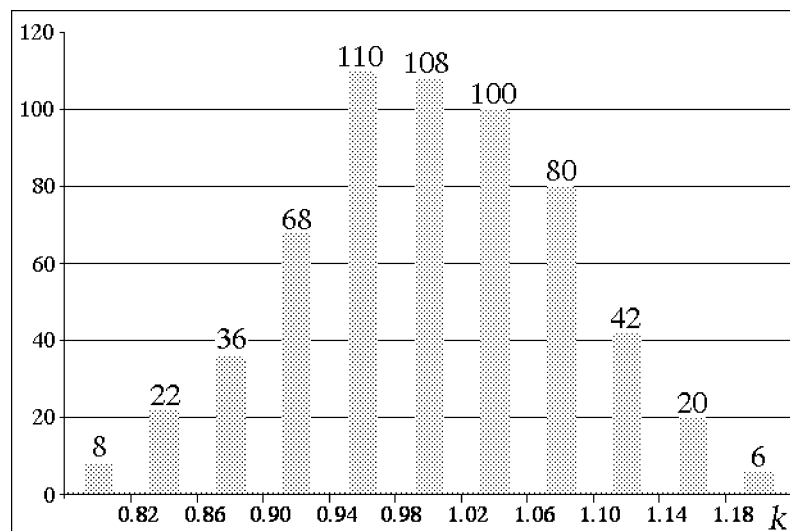
The electromotive force  $v$  between the poles of a battery has been repeatedly measured, yielding the values below (in V):

$i$	$v_i$	$i$	$v_i$	$i$	$v_i$
1	1.3845	7	1.3826	13	1.3666
2	1.3570	8	1.3970	14	1.3406
3	1.3766	9	1.3924	15	1.3625
4	1.3551	10	1.3550	16	1.3770
5	1.3534	11	1.3663	17	1.3347
6	1.3724	12	1.3550	18	1.3585

which can be assumed to be normally distributed. Determine the confidence interval of the mean  $\mu$  and that of the standard deviation  $\sigma$ , both at the confidence level of 90%.

□ **Exercise 5****Points 5**

The data of the thermal conductivity  $k$  of a semiconductor are believed to follow a normal distribution. To check the conjecture 600 measurements of thermal conductivity have been carried out and the relative sample mean  $\bar{k} = 1.00$  and standard deviation  $s = 0.08$  calculated (in arbitrary units). The histogram below summarizes the data:



Check the hypothesis with a significance level ( $\alpha$ ) of 5% and ( $\beta$ ) of 1%.

**□ Exercise 6****Points 4**

A thermal treatment is applied to 12 samples of the same metallic alloy. The electrical resistivity  $\rho$  of the material is measured for each sample prior to and after the treatment, providing the results below (in  $10^{-8} \Omega \cdot \text{m}$ ):

sample no.	$\rho$ before the treatment	$\rho$ after the treatment
1	3.681	3.990
2	3.747	4.081
3	3.666	3.595
4	3.651	3.927
5	3.647	3.904
6	3.862	4.105
7	3.820	3.864
8	3.807	3.759
9	3.632	3.627
10	3.693	3.931
11	3.748	3.697
12	3.735	4.099

Assume that the data are normal and check, with a significance level of 2%, whether the treatment really affects the electrical resistivity of the material.

**□ Exercise 7****Points 4**

A sintering process yields metallic alloy samples whose roughness  $Ra$  and kinematic friction coefficient  $\mu$  vary in an unpredictable way according to a bivariate normal probability distribution. To check if the two quantities may be correlated, 16 measurements of  $Ra$  and  $\mu$  have been carried out on as many samples. The results are listed in the table below:

$i$	$Ra_i$	$\mu_i$	$i$	$Ra_i$	$\mu_i$
1	1.1	2.65	9	3.3	2.58
2	1.3	1.36	10	3.6	2.93
3	1.7	2.24	11	3.8	3.64
4	2.0	1.99	12	4.0	3.40
5	2.4	2.88	13	4.2	4.33
6	2.7	2.59	14	4.6	4.27
7	2.9	3.54	15	4.8	3.80
8	3.1	3.02	16	5.0	4.05

where the roughness is expressed in  $10^{-6}\text{m}$ . Apply Pearson's linear correlation coefficient to check whether the quantities  $Ra$  and  $\mu$  can be regarded as stochastically independent, at a significance level (a) of 10%, (b) of 5% and (c) of 1%.

**□ Exercise 8****Points 4**

A thermal treatment is applied to a material in order to increase toughness. 11 samples of the material are thermally treated at a temperature of 900 K. An analogous treatment of the same duration is applied to further 15 samples of the same material, but at a temperature of 1300 K. The toughness of all the samples is finally measured. The results are given below (in undimensionalized units):

$T = 800 \text{ K}$	$T = 1200 \text{ K}$
2.99	2.80
2.86	2.31
2.94	2.15
3.22	2.05
2.59	2.00
3.16	2.54
2.25	2.44
2.26	2.59
2.75	2.64
2.54	2.20
2.86	2.54
	2.67
	2.25
	2.85
	2.77

Assuming that the data are normal, and after having checked whether the relative variances may or may not be regarded as equal, verify with a significance level of 2% if the temperature of the thermal treatment has a significant effect on the toughness of the material.

**□ Exercise 9****Points 6**

The table below shows the results of some experimental measurements concerning the surface tension  $\gamma$  ( $\text{mJ} \cdot \text{m}^{-2}$ ) of glycerol as a function of the temperature  $T$  ( $^{\circ}\text{C}$ ):

$i$	$T_i$	$\gamma_{i1}$	$\gamma_{i2}$	$\gamma_{i3}$	$\gamma_{i4}$	$\gamma_{i5}$
1	20	63.39	64.35	63.90	64.09	63.75
2	30	62.49	63.48	63.30	63.49	63.15
3	40	62.19	62.88	62.70	62.89	62.55
4	50	61.96	62.29	62.11	62.30	61.45
5	60	61.36	61.69	61.51	60.90	60.79
6	70	60.25	61.09	60.91	61.10	60.76
7	80	59.82	60.49	60.31	60.50	60.16

While the random error on the temperatures  $T_i$  is negligible, the surface tension data  $\gamma_i$  can be assumed to be independent normal random variables with the same standard deviation  $\sigma = 0.27$ . Determine:

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

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

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

- (b) the 95% confidence intervals of the regression parameters  $\mu$  and  $\kappa$ ;  
 (c) the 95% confidence region for predictions;  
 (d) the 95% confidence interval for the value of  $\gamma$  predicted at  $T = 45^{\circ}\text{C}$ .  
 (e) the goodness of fit of the regression model;

**Remark** The sufficient grade corresponds to 18 points

**Solution to Exercise 1**

For a number  $n = 250 > 30$  of data the sample can be regarded as large, and it is not necessary to assume that the statistical population is normal. At a confidence level  $1 - \alpha$  the confidence interval for the mean  $\mu$  is given by

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

and depends on the sample estimates of the mean and standard deviation:

$$\bar{x} = 15.50 \text{ g} \quad s = 0.45 \text{ g},$$

along with the critical value  $z_{[1-\frac{\alpha}{2}]}$ , which denotes the inverse of the cumulative standard normal distribution at  $1 - \frac{\alpha}{2}$ . The general form of the confidence interval becomes therefore

$$15.50 - z_{[1-\frac{\alpha}{2}]} \frac{0.45}{\sqrt{250}} \leq \mu \leq 15.50 + z_{[1-\frac{\alpha}{2}]} \frac{0.45}{\sqrt{250}}$$

i.e., equivalently,

$$15.50 - 0.0284605 \cdot z_{[1-\frac{\alpha}{2}]} \leq \mu \leq 15.50 + 0.0284605 \cdot z_{[1-\frac{\alpha}{2}]}.$$

We must specialize the latter formula to the prescribed confidence levels.

(a) *Confidence level 80%*

The confidence level is  $1 - \alpha = 0.80$ , so that  $\alpha = 0.20$  and

$$1 - \frac{\alpha}{2} = 1 - \frac{0.20}{2} = 1 - 0.10 = 0.90.$$

The critical value  $z_{[1-\frac{\alpha}{2}]} = z_{[0.90]}$  is then calculated by using the Excel function NORMINV:

$$\text{NORMINV}(0, 90; 0; 1)$$

which provides

$$z_{[0.90]} = 1.281551566.$$

As an alternative approach, 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 look for the probability value

$$\frac{1}{2} - \frac{\alpha}{2} = 0.5 - \frac{0.20}{2} = 0.5 - 0.10 = 0.40$$

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:

$$\int_0^{1.28} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 0.39973 \qquad \int_0^{1.29} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 0.40147.$$

We can improve the approximation by adopting the linear interpolation scheme below:

$(1 - \alpha)/2$	$z_{[1-\frac{\alpha}{2}]}$
0.39973	1.28
0.40	$z_{[0.90]}$
0.40147	1.29

$$\frac{0.40 - 0.39973}{0.40147 - 0.39973} = \frac{z_{[0.90]} - 1.28}{1.29 - 1.28}$$

which gives the critical value estimate

$$z_{[0.90]} = 1.28 + (1.29 - 1.28) \cdot \frac{0.40 - 0.39973}{0.40147 - 0.39973} = 1.281552$$

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

$$0.0284605 \cdot z_{[1-\frac{\alpha}{2}]} = 0.0284605 \cdot z_{[0.90]} = 0.0284605 \cdot 1.281552 = 0.036474$$

and the confidence interval for the mean  $\mu$  of the bolt weight becomes:

$$\mu = (15.50 \pm 0.04) \text{ g},$$

or, equivalently,

$$15.46 \text{ g} \leq \mu \leq 15.54 \text{ g}.$$

(b) *Confidence level 95%*

In this case we have  $\alpha = 1 - 0.95 = 0.05$  and thus

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

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

$$\text{NORMINV}(0, 975; 0; 1) \implies z_{[0.975]} = 1.959963985$$

so that

$$0.0284605 \cdot z_{[1-\frac{\alpha}{2}]} = 0.0284605 \cdot z_{[0.975]} = 0.0284605 \cdot 1.959964 = 0.055782$$

and the confidence interval for the mean takes the form

$$\mu = (15.50 \pm 0.06) \text{ g}$$

i.e.

$$15.44 \text{ g} \leq \mu \leq 15.56 \text{ g}.$$

As in the previous case, a satisfactory approximation can be obtained by 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 approximation 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 expected.

### Solution to Exercise 2

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  $\bar{\kappa}$  and the sample standard deviation  $s$ , and determine then the farthest data from the mean, as illustrated in the table below:

$i$	$\kappa_i$	$\kappa_i - \bar{\kappa}$	$ \kappa_i - \bar{\kappa} $	$ \kappa_i - \bar{\kappa} ^2$	outlier
1	0.6237	0.0204	0.0204	0.00041582	
2	0.6342	0.0309	0.0309	0.00095430	
3	0.5702	-0.0331	0.0331	0.00109616	
4	0.6102	0.0069	0.0069	0.00004750	
5	0.6223	0.0190	0.0190	0.00036068	
6	0.5575	-0.0458	0.0458	0.00209840	
7	0.5329	-0.0704	0.0704	0.00495733	
8	0.5584	-0.0449	0.0449	0.00201676	
9	0.6212	0.0179	0.0179	0.00032011	
10	0.5633	-0.0400	0.0400	0.00160067	
11	0.7020	0.0987	0.0987	0.00974005	×
12	0.6438	0.0405	0.0405	0.00163958	

where:

$$\bar{\kappa} = \frac{1}{12} \sum_{i=1}^{12} \kappa_i = 0.6033 \quad s = \sqrt{\frac{1}{11} \sum_{i=1}^{12} (\kappa_i - \bar{\kappa})^2} = 0.0479$$

while the outlier is the data  $\kappa_{11} = 0.7020$ , whose distance from the mean  $\bar{\kappa}$  turns out to be maximum. The distance  $z$  of the suspect data from the mean, in units of  $s$ , can be expressed as

$$z = \frac{\kappa_{11} - \bar{\kappa}}{s} = \frac{0.7020 - 0.6033}{0.0479} = 2.0600$$

and is (slightly) greater than the tabulated critical value of Chauvenet test for  $n = 12$  data

$$z_{cr,12} = 2.036834132.$$

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  $2.0600s$ , by using the table of the standard normal cumulative distribution. We have indeed

$$\begin{aligned} P(|\kappa_{11} - \bar{\kappa}| \geq 2.0600s) &= 1 - P(|\kappa_{11} - \bar{\kappa}| < 2.0600s) = \\ &= 1 - 2P(\bar{\kappa} \leq \kappa_{11} < \bar{\kappa} + 2.0600s) = \\ &= 1 - 2 \cdot 0.48030 = 0.03940 \end{aligned}$$

due to the value of  $P(\bar{\kappa} \leq \kappa_{11} < \bar{\kappa} + 2.0600s) = 0.48030$  which can be read directly on the table of the standard normal distribution. As a consequence, the expected number of data at such a large distance from the mean, out of 12 datapoints, would be

$$0.03940 \cdot 12 = 0.47260$$

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

### Solution to Exercise 3

The equivalent resistance is given by the formula

$$R = \frac{R_1 R_2}{R_1 + R_2}$$

with the partial resistances

$$R_1 = (200 \pm 1) \Omega \quad R_2 = (100 \pm 1) \Omega.$$

The estimate of  $R$  is determined by using the estimated true values of  $R_1$  and  $R_2$ :

$$\bar{R}_1 = 200 \Omega \quad \bar{R}_2 = 100 \Omega$$

and writes therefore

$$\bar{R} = \frac{200 \cdot 100}{200 + 100} = 66.667 \Omega.$$

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

$$\ln R = \ln R_1 + \ln R_2 - \ln(R_1 + R_2)$$

we deduce indeed the first order partial derivatives:

$$\frac{\partial \ln R}{\partial R_1} = \frac{1}{R_1} - \frac{1}{R_1 + R_2} = \frac{R_2}{R_1(R_1 + R_2)}$$

$$\frac{\partial \ln R}{\partial R_2} = \frac{1}{R_2} - \frac{1}{R_1 + R_2} = \frac{R_1}{R_2(R_1 + R_2)}$$

and therefore the differential

$$\frac{dR}{R} = \frac{R_2}{R_1(R_1 + R_2)} dR_1 + \frac{R_1}{R_2(R_1 + R_2)} dR_2$$

which provides the relative error estimate:

$$\frac{\Delta R}{\bar{R}} = \frac{\bar{R}_2}{\bar{R}_1 + \bar{R}_2} \frac{\Delta R_1}{\bar{R}_1} + \frac{\bar{R}_1}{\bar{R}_1 + \bar{R}_2} \frac{\Delta R_2}{\bar{R}_2}.$$

As a consequence, we obtain the relative error

$$\frac{\Delta R}{\bar{R}} = \frac{100}{200 + 100} \frac{1}{200} + \frac{200}{200 + 100} \frac{1}{100} = 0.008333333 = 0.8333\%$$

and the absolute error on  $R$

$$\Delta R = \frac{\Delta R}{\bar{R}} \bar{R} = 0.008333 \cdot 66.667 = 0.555 \Omega,$$

so that the error interval for  $R$  takes the form

$$R = \bar{R} \pm \Delta R = (66.667 \pm 0.555) \Omega$$

or, rounding off the absolute error to one significant digit only,

$$R = \bar{R} \pm \Delta R = (66.7 \pm 0.6) \Omega.$$

**Solution to Exercise 4**

The sample, consisting of  $n = 18$  data, cannot be considered as large because the number of data is smaller than 30. We need calculate therefore the correct confidence interval for the mean by using the hypothesis of the normal statistical population. Analogously, the sample variance  $s^2$  cannot be regarded as approximately equal to the variance  $\sigma^2$  of the population, as prescribed by the weak law of large numbers (Kintchine's theorem) for large samples: an appropriate confidence interval must be reckoned also for  $\sigma^2$ , and the goal can be achieved thanks to the normal distribution of the data. The calculation of the sample mean provides:

$$\bar{v} = \frac{1}{n} \sum_{i=1}^n v_i = 1.365956.$$

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

$i$	$v_i$	$(v_i - \bar{v}) \cdot 10^2$	$(v_i - \bar{v})^2 \cdot 10^4$
1	1.3845	1.8544444	3.4389642
2	1.3570	-0.8955556	0.8020198
3	1.3766	1.0644444	1.1330420
4	1.3551	-1.0855556	1.1784309
5	1.3534	-1.2555556	1.5764198
6	1.3724	0.6444444	0.4153086
7	1.3826	1.6644444	2.7703753
8	1.3970	3.1044444	9.6375753
9	1.3924	2.6444444	6.9930864
10	1.3550	-1.0955556	1.2002420
11	1.3663	0.0344444	0.0011864
12	1.3550	-1.0955556	1.2002420
13	1.3666	0.0644444	0.0041531
14	1.3406	-2.5355556	6.4290420
15	1.3625	-0.3455556	0.1194086
16	1.3770	1.1044444	1.2197975
17	1.3347	-3.1255556	9.7690975
18	1.3585	-0.7455556	0.5558531

from which we deduce the sample estimate of the variance:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (v_i - \bar{v})^2 = 2.8496614 \cdot 10^{-4}$$

and that of the standard deviation:

$$s = \sqrt{s^2} = 1.688094 \cdot 10^{-2}.$$

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

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

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

$$\begin{aligned} \bar{v} - t_{[0.95](17)} \frac{s}{\sqrt{18}} &= 1.365956 - 1.740 \cdot \frac{0.01688094}{\sqrt{18}} = 1.35903 \\ \bar{v} + t_{[0.95](17)} \frac{s}{\sqrt{18}} &= 1.365956 + 1.740 \cdot \frac{0.01688094}{\sqrt{18}} = 1.37288 \end{aligned}$$

so that the 90%-confidence interval becomes

$$1.35903 \text{ V} \leq \mu \leq 1.37288 \text{ V}$$

or, equivalently,

$$\mu = (1.36596 \pm 0.00693) \text{ V}.$$

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

$$\mu = (1.366 \pm 0.007) \text{ V}$$

can be regarded as more than satisfactory. We recall 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

$$\text{TINV}(0, 10; 17) \quad \Rightarrow \quad 1.739607$$

(b) *Confidence interval for the standard deviation*

For a normal sample the CI of the variance is given by

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

With  $\alpha = 0.10$  and  $n = 18$  the lower and the upper limits are:

$$\begin{aligned} \frac{1}{\chi^2_{[0.95](17)}} 17 s^2 &= \frac{1}{27.587} \cdot 17 \cdot 2.8496614 \cdot 10^{-4} = 1.7560534 \cdot 10^{-4} \\ \frac{1}{\chi^2_{[0.05](17)}} 17 s^2 &= \frac{1}{8.672} \cdot 17 \cdot 2.8496614 \cdot 10^{-4} = 5.5862828 \cdot 10^{-4} \end{aligned}$$

and the CI of the variance becomes

$$1.7560534 \cdot 10^{-4} V^2 \leq \sigma^2 \leq 5.5862828 \cdot 10^{-4} V^2.$$

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

$$1.3251616 \cdot 10^{-2} V \leq \sigma \leq 2.3635318 \cdot 10^{-2} V.$$

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

$$1.32 \cdot 10^{-2} V \leq \sigma \leq 2.36 \cdot 10^{-2} V.$$

Pay attention that the critical values of the  $\chi^2$  variable can be also computed by the Excel function CHIINV:

$$\text{CHIINV}(0, 05; 17) \quad \Longrightarrow \quad \chi^2_{[0.95](17)} = 27.587112$$

$$\text{CHIINV}(0, 95; 17) \quad \Longrightarrow \quad \chi^2_{[0.05](17)} = 8.671760.$$

### Solution to Exercise 5

The sample histogram is bell-shaped, thus the hypothesis of the normal population appears pretty reasonable. Formally, we want to check the null hypothesis

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

versus the alternative hypothesis that  $H_0$  is false. We can apply the  $\chi^2$  test, since all the empirical frequencies in the histogram are sufficiently high ( $f_i \geq 3$  in each bin). In the present case the sample data were used to estimate the mean and the standard deviation of the distribution:

$$\mu = \bar{k} = 1.00 \quad \sigma = s = 0.08$$

and the the histogram intervals of the results (the classes) are  $h = 11$  in all. Due to the  $c = 2$  constraints on the mean and the standard deviation, if  $H_0$  holds true the  $\chi^2$  of data follows approximately a  $\chi^2$  distribution with

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

degrees of freedom. To calculate the  $\chi^2$  the expected frequencies in each class are needed, by assuming 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 correspon-

dence among the values of  $k$  and those of  $z$ :

$k$	0.82	0.86	0.90	0.94	0.98	1.02	1.06	1.10	1.14	1.18
$z$	-2.25	-1.75	-1.25	-0.75	-0.25	0.25	0.75	1.25	1.75	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 is convenient to denote with  $p(z)$  the standard normal distribution and introduce the integral

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

whose values are readable 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 compile the table below:

$i$	$p_i$	$n_i = 600 \cdot p_i$	$f_i$	$f_i - n_i$	$(f_i - n_i)^2/n_i$
1	$\Phi(+\infty) - \Phi(2.25) = 0.01222$	7.334684	8	0.665316	0.060349695
2	$\Phi(2.25) - \Phi(1.75) = 0.02784$	16.700811	22	5.299189	1.681439895
3	$\Phi(1.75) - \Phi(1.25) = 0.06559$	39.354370	36	-3.354370	0.285909764
4	$\Phi(1.25) - \Phi(0.75) = 0.12098$	72.586547	68	-4.586547	0.289811491
5	$\Phi(0.75) - \Phi(0.25) = 0.17466$	104.799793	110	5.200207	0.258036303
6	$\Phi(0.25) + \Phi(0.25) = 0.19742$	118.447591	108	-10.447591	0.921522786
7	$\Phi(0.75) - \Phi(0.25) = 0.17466$	104.799793	100	-4.799793	0.219828816
8	$\Phi(1.25) - \Phi(0.75) = 0.12098$	72.586547	80	7.413453	0.757155205
9	$\Phi(1.75) - \Phi(1.25) = 0.06559$	39.354370	42	2.645630	0.177854649
10	$\Phi(2.25) - \Phi(1.75) = 0.02784$	16.700811	20	3.299189	0.651743888
11	$\Phi(+\infty) - \Phi(2.25) = 0.01222$	7.334684	6	-1.334684	0.242870776

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  $\chi^2$  of the sample:

$$\chi^2 = \sum_{i=1}^{11} \frac{(f_i - n_i)^2}{n_i} = 5.546523266.$$

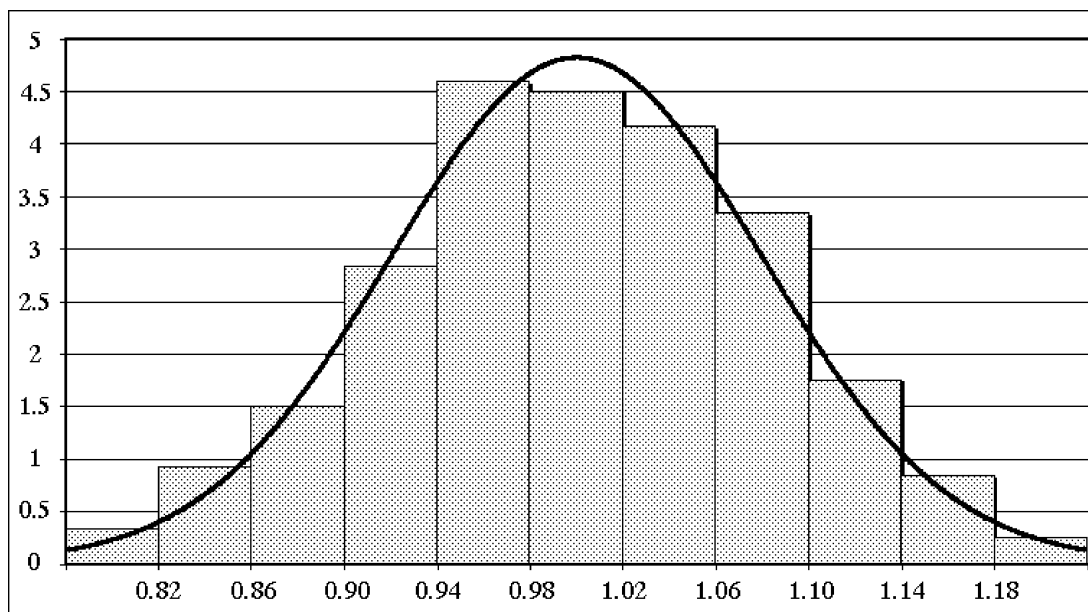
On the table of the  $\chi^2$  cumulative probability distribution we find the critical values:

$$\begin{aligned}\chi^2_{[1-\alpha](8)} &= \chi^2_{[0.95](8)} = 15.507 && \text{for } \alpha = 0.05 \\ \chi^2_{[1-\alpha](8)} &= \chi^2_{[0.99](8)} = 20.090 && \text{for } \alpha = 0.01.\end{aligned}$$

In both cases the  $\chi^2$  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 thermal conductivity of the semiconductor is probably normal*. The formal conclusion is supported by the excellent overlap between the theoretical distribution of the data:

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

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



The binned distribution is a piecewise constant function that along the  $i$ -th bin takes a constant value  $f_i/600/0.04$ , where the number of data — 600 — is introduced to normalize the distribution to 1, while the factor 0.04 is the width of the bin (the same for all the bins, in this case). Owing 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:

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

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 found in the statistical table.

### Solution to Exercise 6

Electrical resistivity  $\rho$  is measured *on each sample* before and after the thermal treatment, thus it seems quite reasonable to apply a paired  $t$ -test for the comparison of the means. Therefore, the values relative to the same sample will be coupled:

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

on having denoted with  $y_i$  and  $z_i$  the resistivities measured before and after the treatment, respectively. If  $\mu_1$  and  $\mu_2$  stand for the mean (true) value of the electrical resistivity prior to 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  $\rho$ , versus the alternative hypothesis  $H_1 : \mu_1 \neq \mu_2$  that the claim is false. For a paired  $t$ -test the test variable writes:

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

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

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

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

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

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

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

As usual, a more accurate estimate can be derived

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

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

$$\bar{y} = \frac{1}{12} \sum_{i=1}^{12} y_i = 3.72408 \qquad \bar{z} = \frac{1}{12} \sum_{i=1}^{12} z_i = 3.88158$$

whereas

$$\frac{1}{n-1} \sum_{i=1}^n (y_i - z_i - \bar{y} + \bar{z})^2 = \frac{1}{11} \sum_{i=1}^{12} (y_i - z_i - \bar{y} + \bar{z})^2 = 0.028334818$$

and therefore

$$\sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - z_i - \bar{y} + \bar{z})^2} = \sqrt{0.028334818} = 0.1683295$$

so that the test variable takes the value

$$t = \sqrt{12} \cdot \frac{3.72408 - 3.88158}{0.1683295} = -3.2412.$$

The table below accounts for the detailed calculations:

$y_i$	$z_i$	$y_i - z_i$	$y_i - z_i - \bar{y} + \bar{z}$	$(y_i - z_i - \bar{y} + \bar{z})^2$
3.681	3.990	-0.309	-0,15150	0,022952
3.747	4.081	-0.334	-0,17650	0,031152
3.666	3.595	0.071	0,22850	0,052212
3.651	3.927	-0.276	-0,11850	0,014042
3.647	3.904	-0.257	-0,09950	0,009900
3.862	4.105	-0.243	-0,08550	0,007310
3.820	3.864	-0.044	0,11350	0,012882
3.807	3.759	0.048	0,20550	0,042230
3.632	3.627	0.005	0,16250	0,026406
3.693	3.931	-0.238	-0,08050	0,006480
3.748	3.697	0.051	0,20850	0,043472
3.735	4.099	-0.364	-0,20650	0,042642

The calculated value of  $t$  belongs to the lower tail of the rejection region, since

$$-3.2412 < -2.718,$$

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 7

We can assume that the roughness  $Ra$  and the friction coefficient  $\mu$  are two normal random variables. The sample means  $\bar{Ra}$  and  $\bar{\mu}$  of the two quantities are given by:

$$\bar{Ra} = \frac{1}{16} \sum_{i=1}^{16} Ra_i = 3.156250 \quad \bar{\mu} = \frac{1}{16} \sum_{i=1}^{16} \mu_i = 3.079375$$

and allow us to calculate the sum of products of residuals

$$SS_{Ra\mu} = \sum_{i=1}^{16} (Ra_i - \bar{Ra})(\mu_i - \bar{\mu}) = 13.307563$$

along with those of the relative squares:

$$SS_{RaRa} = \sum_{i=1}^{16} (Ra_i - \bar{Ra})^2 = 22.599375$$

$$SS_{\mu\mu} = \sum_{i=1}^{16} (\mu_i - \bar{\mu})^2 = 10.658694,$$

as illustrated in the following table:

$Ra_i$	$\mu_i$	$\Delta Ra_i$	$\Delta \mu_i$	$\Delta Ra_i^2$	$\Delta \mu_i^2$	$\Delta Ra_i \Delta \mu_i$
1.1	2.65	-2.056250	-0.429375	4.228164	0.184363	0.882902
1.3	1.36	-1.856250	-1.719375	3.445664	2.956250	3.191590
1.7	2.24	-1.456250	-0.839375	2.120664	0.704550	1.222340
2.0	1.99	-1.156250	-1.089375	1.336914	1.186738	1.259590
2.4	2.88	-0.756250	-0.199375	0.571914	0.039750	0.150777
2.7	2.59	-0.456250	-0.489375	0.208164	0.239488	0.223277
2.9	3.54	-0.256250	0.460625	0.065664	0.212175	-0.118035
3.1	3.02	-0.056250	-0.059375	0.003164	0.003525	0.003340
3.3	2.58	0.143750	-0.499375	0.020664	0.249375	-0.071785
3.6	2.93	0.443750	-0.149375	0.196914	0.022313	-0.066285
3.8	3.64	0.643750	0.560625	0.414414	0.314300	0.360902
4.0	3.40	0.843750	0.320625	0.711914	0.102800	0.270527
4.2	4.33	1.043750	1.250625	1.089414	1.564063	1.305340
4.6	4.27	1.443750	1.190625	2.084414	1.417588	1.718965
4.8	3.80	1.643750	0.720625	2.701914	0.519300	1.184527
5.0	4.05	1.843750	0.970625	3.399414	0.942113	1.789590

where, for simplicity's sake, we have posed  $Ra_i - \overline{Ra} = \Delta Ra_i$  and  $\mu_i - \overline{\mu} = \Delta \mu_i$ . The linear correlation coefficient becomes then

$$r = \frac{SS_{Ra\mu}}{\sqrt{SS_{RaRa}} \sqrt{SS_{\mu\mu}}} = \frac{13.307563}{\sqrt{22.599375} \sqrt{10.658694}} = 0.857429277.$$

Since the bivariate random variable  $(Ra, \mu)$  may be assumed to be normal, we may check the null hypothesis

$$H_0 : Ra \text{ and } \mu \text{ are stochastically independent}$$

in contrast to the alternative hypothesis

$$H_1 : Ra \text{ and } \mu \text{ are stochastically dependent}$$

by means of the test variable

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

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

$$t = \sqrt{16-2} \frac{0.857429277}{\sqrt{1-0.857429277^2}} = 6.2343475.$$

At a significance level  $\alpha$  the critical region of the test takes the form

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

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

In this case the critical value of the test statistic is

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

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

$$\text{TINV}(0, 10; 14) \quad \Longrightarrow \quad t_{[0.95]}(12) = 1.761310.$$

We conclude that the value of the test statistic belongs to the upper tail of the rejection region, so that  $H_0$  *must be rejected*. The random variables  $Ra$  and  $\mu$  can be considered stochastically dependent/correlated.

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

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

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

As in the previous case, a more accurate result can be obtained by using the function TINV:

$$\text{TINV}(0, 05; 14) \quad \Longrightarrow \quad t_{[0.975]}(14) = 2.144787.$$

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

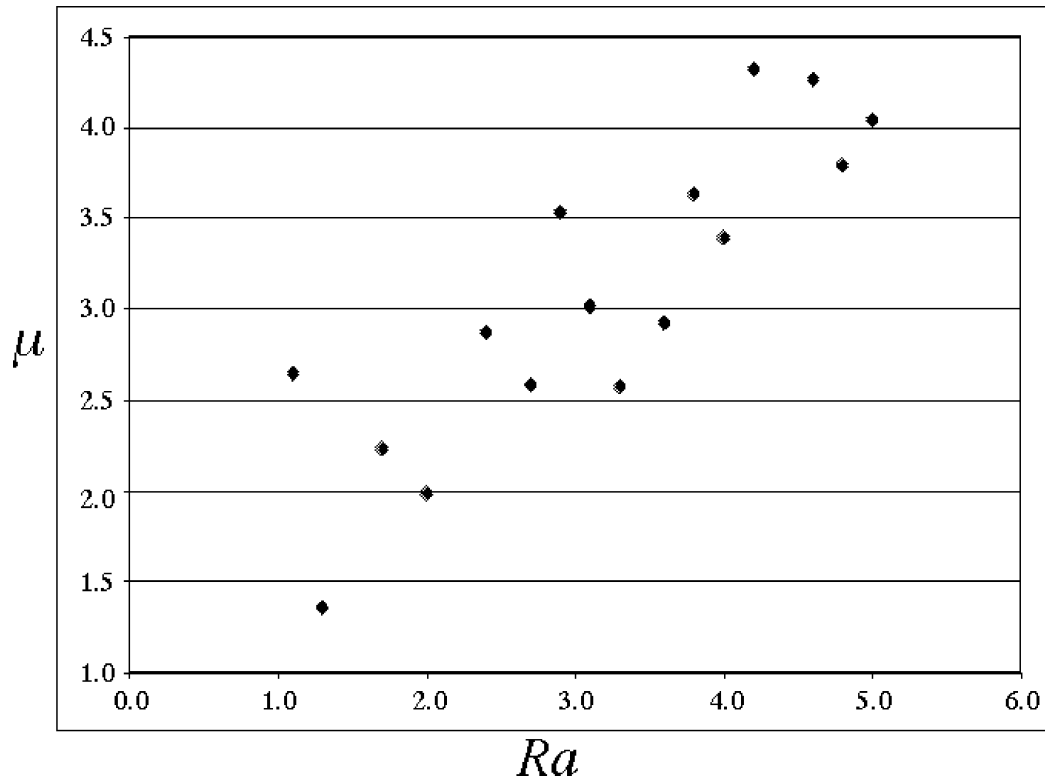
$$t_{[1-\frac{\alpha}{2}]}(14) = t_{[0.995]}(14) = 2.977$$

and can be more accurately calculated by the TINV function:

$$\text{TINV}(0, 01; 14) \quad \Longrightarrow \quad t_{[0.995]}(14) = 2.976843.$$

Since the value 6.2343475 of the test statistic again does not fall within the acceptance region for  $H_0$ , we must conclude that *the random variables  $Ra$  and  $\mu$  are perhaps stochastically dependent*. Due to the positive sign of the correlation coefficient, which is rather close to +1, the relation should be direct.

It is worthy of note that our formal conclusion is also supported by the trend of the data plot:



because the datapoints appear to be partially aligned along a straight line of positive slope.

### Solution to Exercise 8

Let us denote with  $\mu_1$  and  $\mu_2$  the mean values of the two samples, that is the “true values” of toughness for the first and the second thermal treatment, respectively. Let  $p = 11$  be the number of data  $y_1, \dots, y_p$  of the first sample and  $q = 15$  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 toughness of the material) against the alternative hypothesis  $H_1 : \mu_1 \neq \mu_2$  (the two treatments yield materials with a different toughness). The significance level we choose is 2%.

#### *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:

$$\bar{y} = \frac{1}{p} \sum_{i=1}^p y_i = 2.765455 \qquad \bar{z} = \frac{1}{q} \sum_{j=1}^q z_j = 2.453333$$

$$s_y^2 = \frac{1}{p-1} \sum_{i=1}^p (y_i - \bar{y})^2 = 0.106407 \qquad s_z^2 = \frac{1}{q-1} \sum_{j=1}^q (z_j - \bar{z})^2 = 0.077124$$

according to the detailed calculation shown in the table below:

$y_i$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$z_i$	$z_i - \bar{z}$	$(z_i - \bar{z})^2$
2.99	0.224545	0.050421	2.80	0.346667	0.120178
2.86	0.094545	0.008939	2.31	-0.143333	0.020544
2.94	0.174545	0.030466	2.15	-0.303333	0.092011
3.22	0.454545	0.206612	2.05	-0.403333	0.162678
2.59	-0.175455	0.030784	2.00	-0.453333	0.205511
3.16	0.394545	0.155666	2.54	0.086667	0.007511
2.25	-0.515455	0.265693	2.44	-0.013333	0.000178
2.26	-0.505455	0.255484	2.59	0.136667	0.018678
2.75	-0.015455	0.000239	2.64	0.186667	0.034844
2.54	-0.225455	0.050830	2.20	-0.253333	0.064178
2.86	0.094545	0.008939	2.54	0.086667	0.007511
			2.67	0.216667	0.046944
			2.25	-0.203333	0.041344
			2.85	0.396667	0.157344
			2.77	0.316667	0.100278

We obtain therefore:

$$F = \frac{0.106407}{0.077124} = 1.379694.$$

According to the  $F$ -test, 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  $\alpha$ , 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  $p = 11$ ,  $q = 15$  and  $\alpha = 0.02$ , so that the acceptance condition becomes

$$0.217352 = F_{[0.01](10,14)} < F < F_{[0.99](10,14)} = 3.939396$$

as derived from the Excel function FINV:

$$\text{FINV}(0, 99; 10; 14) \implies F_{[0.01](10,14)} = 0.217352$$

$$\text{FINV}(0, 01; 10; 14) \implies F_{[0.99](10,14)} = 3.939396$$

or, to a lesser approximation, by using the table of the Fisher cumulative distributions —  $F_{[0.01](10,14)} = 0.2174$  and  $F_{[0.99](10,14)} = 3.9394$ . Since the value of the test statistic actually falls within the acceptance region:

$$0.217352 < 1.379694 < 3.939396$$

we conclude that *the variances  $\sigma_1^2$  and  $\sigma_2^2$  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

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

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

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

In this case we have the pooled variance

$$\frac{(11-1) \cdot 0.106407 + (15-1) \cdot 0.077124}{11+15-2} = 0.089325253$$

and the value of the test statistic is therefore

$$t = \frac{\bar{y} - \bar{z}}{s \sqrt{\frac{1}{p} + \frac{1}{q}}} = \frac{2.765455 - 2.453333}{\sqrt{0.089325253} \sqrt{\frac{1}{11} + \frac{1}{15}}} = 2.630822$$

whereas

$$\text{TINV}(0, 02; 24) \quad \Longrightarrow \quad t_{[1-\frac{\alpha}{2}](p+q-2)} = t_{[0.99](24)} = 2.492159$$

so that the critical region takes the form

$$\left\{ t < -2.492159 \right\} \cup \left\{ t > 2.492159 \right\}.$$

Clearly, the value of  $t$  belongs to the upper tail of the critical region and *the null hypothesis must be rejected* at the significance level of 2%. We conclude that *the different temperatures of the thermal treatment probably have an effect* on the toughness of the material.

*Remark. T-test in the case of unequal variances*

If the variances  $\sigma_1^2$  and  $\sigma_2^2$  were different, the test variable would be

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

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 the present case we have

$$\bar{y} = 2.765455 \quad \bar{z} = 2.453333 \quad s_y^2 = 0.106407 \quad s_z^2 = 0.077124$$

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

$$n = \frac{\left(\frac{0.106407}{11} + \frac{0.077124}{15}\right)^2}{\frac{1}{10}\left(\frac{0.106407}{11}\right)^2 + \frac{1}{14}\left(\frac{0.077124}{15}\right)^2} = 19.5170625,$$

while the test statistic assumes the value

$$t = \frac{2.765455 - 2.453333}{\sqrt{\frac{1}{11} \cdot 0.106407 + \frac{1}{15} \cdot 0.077124}} = 2.564324.$$

The rejection region writes

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

with  $\alpha = 0.02$  and  $n = 19.5170625$ , and therefore we need the  $t$ -value

$$t_{[1-\frac{\alpha}{2}]}(n) = t_{[0.99]}(19.5170625)$$

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

of d.o.f. is automatically truncated to the integer 19. However, we can read on the table the critical values at  $\alpha = 0.02$  for  $n = 19$  and  $n = 20$  d.o.f.

$$t_{[0.99](19)} = 2.539 \qquad t_{[0.99](20)} = 2.528$$

and apply a linear interpolation scheme:

19	2.539
19.5170625	$t_{[0.99](19.5170625)}$
20	2.528

$$\frac{19.5170625 - 19}{20 - 19} = \frac{t_{[0.99](19.5170625)} - 2.539}{2.528 - 2.539}$$

which provides the relationship

$$t_{[0.99](19.5170625)} = 2.539 + (2.528 - 2.539) \frac{19.5170625 - 19}{20 - 19}$$

and finally the approximate critical value

$$t_{[0.99](19.5170625)} = 2.533 .$$

The critical region of  $H_0$  becomes then

$$\{t \leq -2.533\} \cup \{t \geq 2.533\}$$

and *contains the value*  $t = 2.564324$  *of the test statistics*. Therefore, the null hypothesis  $H_0 : \mu_1 = \mu_2$  must be refused, as before.

### Solution to Exercise 9

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 tension data share the same variance. That's why the regression straight line is written by expressing the surface tension  $\gamma$  as a function of the temperature  $T$ :

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

where  $\bar{T}$  stands for the arithmetic mean of the measured temperatures, while  $\mu$  and  $\kappa$  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  $\mu$  and  $\kappa$ .

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  $(T_i, \gamma_i)$  with the same  $T$  are treated as distinct. According to this criterion the whole number of sample data is thus  $n = 35$ .

(a) *Regression straight line*

Since the standard deviations are equal, the  $\chi^2$  fitting reduces to the usual least-squares fitting and the best-fit estimates  $m$  and  $q$  of the parameters can be easily calculated by using the table below:

$T_i$	$\gamma_i$	$T_i - \bar{T}$	$(T_i - \bar{T})^2$	$(T_i - \bar{T}) \gamma_i$
20	63.39	-30.0	900.0	-1901.700
30	62.49	-20.0	400.0	-1249.800
40	62.19	-10.0	100.0	-621.900
50	61.96	0.0	0.0	0.000
60	61.36	10.0	100.0	613.600
70	60.25	20.0	400.0	1205.000
80	59.82	30.0	900.0	1794.600
20	64.35	-30.0	900.0	-1930.500
30	63.48	-20.0	400.0	-1269.600
40	62.88	-10.0	100.0	-628.800
50	62.29	0.0	0.0	0.000
60	61.69	10.0	100.0	616.900
70	61.09	20.0	400.0	1221.800
80	60.49	30.0	900.0	1814.700
20	63.90	-30.0	900.0	-1917.000
30	63.30	-20.0	400.0	-1266.000
40	62.70	-10.0	100.0	-627.000
50	62.11	0.0	0.0	0.000
60	61.51	10.0	100.0	615.100
70	60.91	20.0	400.0	1218.200
80	60.31	30.0	900.0	1809.300
20	64.09	-30.0	900.0	-1922.700
30	63.49	-20.0	400.0	-1269.800
40	62.89	-10.0	100.0	-628.900
50	62.30	0.0	0.0	0.000
60	60.90	10.0	100.0	609.000
70	61.10	20.0	400.0	1222.000
80	60.50	30.0	900.0	1815.000
20	63.75	-30.0	900.0	-1912.500
30	63.15	-20.0	400.0	-1263.000
40	62.55	-10.0	100.0	-625.500
50	61.45	0.0	0.0	0.000
60	60.79	10.0	100.0	607.900
70	60.76	20.0	400.0	1215.200
80	60.16	30.0	900.0	1804.800

which provides the estimates:

$$m = \bar{\gamma} = \frac{1}{n} \sum_{i=1}^n \gamma_i = \frac{2170.35}{35} = 62.01$$

$$q = \frac{\sum_{i=1}^n (T_i - \bar{T}) \gamma_i}{\sum_{i=1}^n (T_i - \bar{T})^2} = \frac{-851.6000}{14000.0} = -0.060828571$$

with  $n = 35$  and

$$\bar{T} = \frac{1}{n} \sum_{i=1}^n T_i = \frac{1750.00}{35} = 50.0.$$

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

$$\begin{aligned} \gamma &= m + q(T - \bar{T}) = 62.01 - 0.060828571 \cdot (T - 50.0) = \\ &= 65.05142857 - 0.060828571 \cdot T \end{aligned}$$

where the number of digits is left temporarily large before the appropriate confidence region has been determined.

(b) *Confidence intervals for the regression parameters*

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

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

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

Here we have  $\alpha = 0.05$  and  $n = 35$ , so that the confidence intervals become:

$$\mu = m \pm t_{[0.975](33)} \sqrt{\frac{1}{35} \frac{\text{SSAR}}{33}}$$

$$\kappa = q \pm t_{[0.975](33)} \sqrt{\left[ \sum_{i=1}^{35} (T_i - \bar{T})^2 \right]^{-1} \frac{\text{SSAR}}{33}}$$

where the sum of squares around regression holds:

$$\text{SSAR} = \sum_{i=1}^n [m + q(T_i - \bar{T}) - \gamma_i]^2 = 3.608988579,$$

while:

$$\begin{aligned} m &= 62.01 \\ q &= -0.060828571 \\ \sum_{i=1}^{35} (T_i - \bar{T})^2 &= 14000.0 \\ t_{[0.975](33)} &= 2.035 . \end{aligned}$$

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:

$$\text{TINV}(0, 05; 33) \quad \Longrightarrow \quad t_{[0.975](33)} = 2.034515287 .$$

As a satisfactory compromise between the tabulated and the Excel value we may assume  $t_{[0.975](33)} = \mathbf{2.03452}$ . By inserting the numerical values and performing the calculations we deduce that:

- the 95%-CI of the parameter  $\mu$  is

$$\mu = 62.01 \pm 2.03452 \sqrt{\frac{1}{35} \frac{3.608988579}{33}}$$

i.e.

$$\mu = 62.01 \pm 0.1137270372 = [61.89627296, 62.12372704]$$

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

$$\kappa = -0.060828571 \pm 2.03452 \sqrt{\frac{1}{14000.0} \frac{3.608988579}{33}}$$

or, equivalently,

$$\kappa = -0.060828571 \pm 0.005686351860 = [-0.06651492329, -0.05514221957] .$$

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

$$\mu = [61.8963, 62.1237] \text{ mJ} \cdot \text{m}^{-2} = (62.0100 \pm 0.1137) \text{ mJ} \cdot \text{m}^{-2}$$

whereas

$$\begin{aligned} \kappa &= [-0.0665149, -0.0551422] \text{ mJ} \cdot \text{m}^{-2} \cdot \text{ }^\circ\text{C}^{-1} = \\ &= (-0.0608286 \pm 0.0056864) \text{ mJ} \cdot \text{m}^{-2} \cdot \text{ }^\circ\text{C}^{-1} . \end{aligned}$$

*(c) Confidence region for predictions*

Since the model is assumed to be homoskedastic, 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:

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

where, more specifically, we have:

$$m = 62.01$$

$$q = -0.060828571$$

$$\bar{T} = \frac{1}{n} \sum_{i=1}^n T_i = 50.0$$

$$t_{[1-\frac{\alpha}{2}](n-2)} = t_{[0.975](33)} = 2.03452$$

$$V = 1 + \frac{1}{n} + \frac{1}{\sum_{i=1}^n (T_i - \bar{T})^2} (T_0 - \bar{T})^2 = 1 + \frac{1}{35} + \frac{(T_0 - 50.0)^2}{14000.0} =$$

$$= 1.028571429 + 0.00007142857143 \cdot (T_0 - 50.0)^2$$

$$\text{SSAR} = \sum_{i=1}^n [m + q(T_i - \bar{T}) - \gamma_i]^2 = 3.608988579 .$$

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

$$\begin{aligned} \gamma_0 &= 62.01 - 0.060828571 \cdot (T_0 - 50.0) \pm \\ &\pm 2.03452 \cdot \sqrt{1.028571429 + 0.00007142857143 \cdot (T_0 - 50.0)^2} \sqrt{\frac{3.608988579}{33}} \end{aligned}$$

and performing the calculations reduces to:

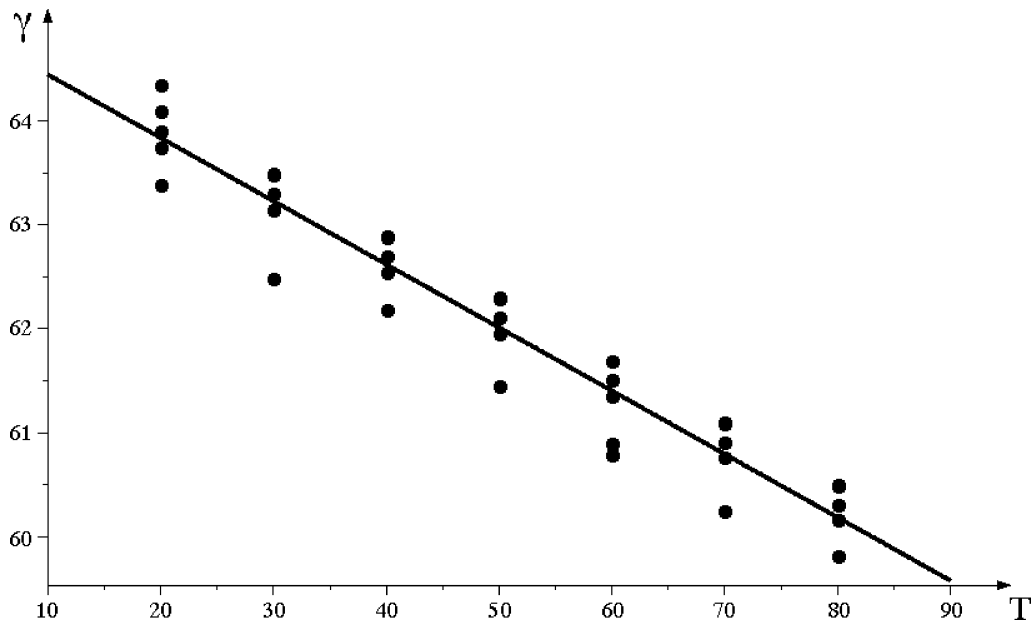
$$\begin{aligned} \gamma_0 &= 65.05142857 - 0.060828571 \cdot T_0 \pm \\ &\pm 0.6728182258 \cdot \sqrt{1.028571429 + 0.00007142857143 \cdot (T_0 - 50.0)^2} \end{aligned}$$

or, more conveniently,

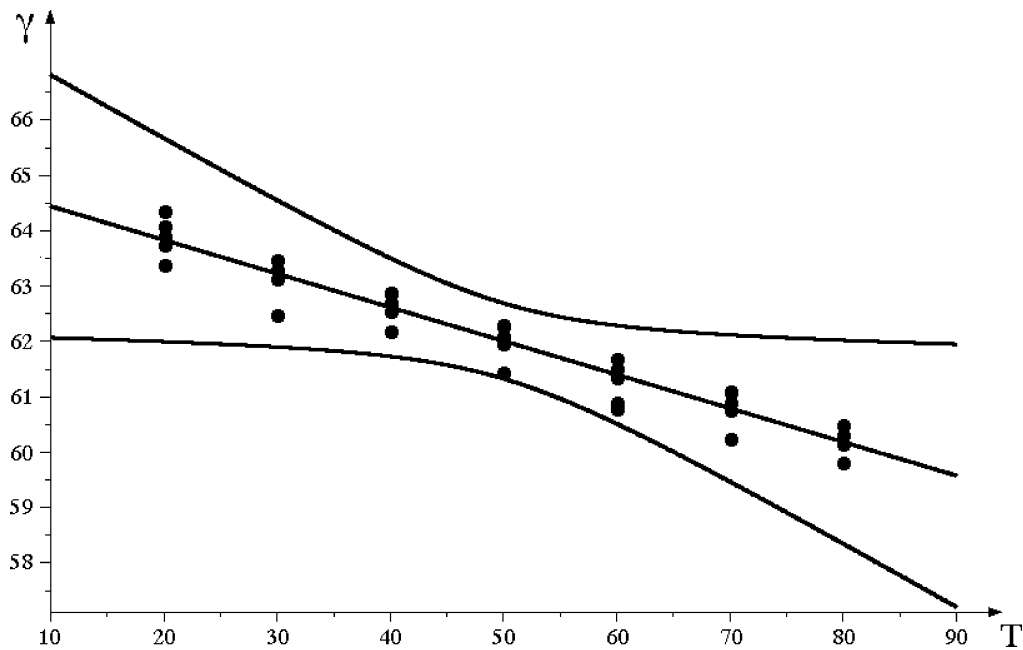
$$\begin{aligned} \gamma_0 &= 65.05142857 - 0.060828571 \cdot T_0 \pm \\ &\pm 0.6728182258 \cdot \sqrt{1.028571429 + 0.7142857143 \cdot \left(\frac{T_0}{100} - 0.50\right)^2} \end{aligned}$$

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  $\gamma$  axis, is minimum for  $T = \bar{T} = 50.0$  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  $(T_0 - \bar{T})^2$  which appears in the expression of  $V$  of the definition has been multiplied by a scale factor 100.

(d) *Confidence interval for a prediction*

The CI at a confidence level of 95% for the prediction of  $\gamma$  at  $T = 45^\circ C$  can be obtained by posing  $T_0 = 45$  in the previous formula

$$\begin{aligned} \gamma_0 &= 65.05142857 - 0.060828571 \cdot T_0 \pm \\ &\pm 0.6728182258 \cdot \sqrt{1.028571429 + 0.7142857143 \cdot \left(\frac{T_0}{100} - 0.50\right)^2}. \end{aligned}$$

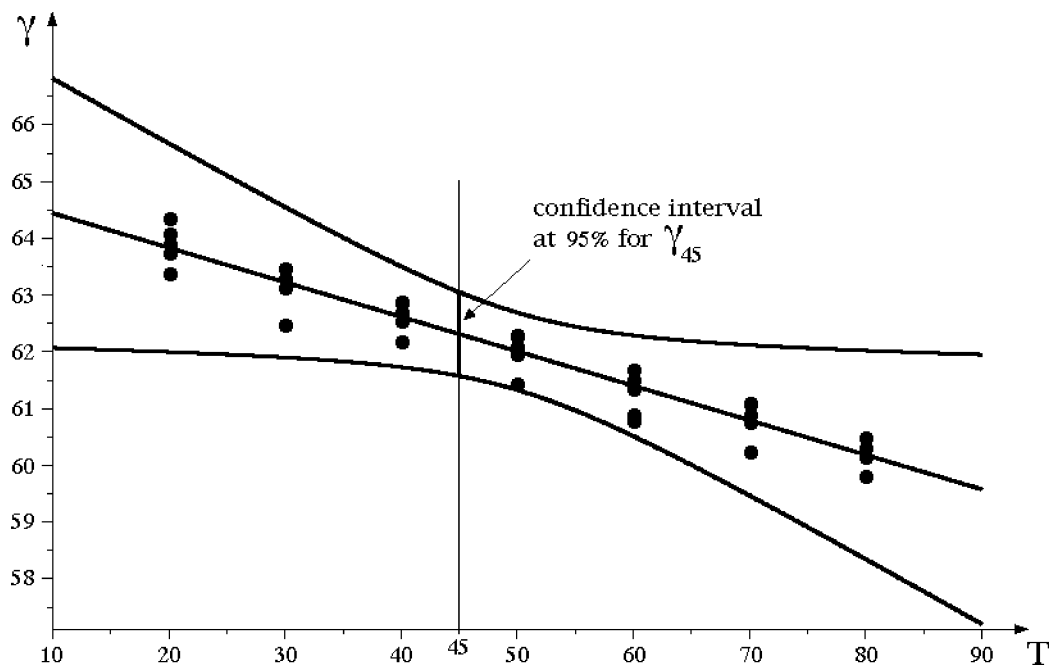
We obtain therefore:

$$\begin{aligned} \gamma_0 &= \gamma_{45} = 65.05142857 - 0.060828571 \cdot 45 \pm \\ &\pm 0.6728182258 \cdot \sqrt{1.028571429 + 0.7142857143 \cdot \left(\frac{45}{100} - 0.50\right)^2} = \\ &= 62.31414286 \pm 0.68295430 = [61.63118857, 62.99709715] \end{aligned}$$

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

$$\gamma_{45} = [61.63, 62.99] \text{ mJ} \cdot \text{m}^{-2} = (62.31 \pm 0.68) \text{ mJ} \cdot \text{m}^{-2}.$$

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  $T = 45$ :



(e) *Goodness of fit*

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

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

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

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

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*  $\sigma = 0.32$ , since we need to determine the NSSAR, and not simply the SSAR. In the present case we have  $n = 35$  data and the regression model is based on the two parameters  $\mu$  and  $\kappa$ . Consequently, the NSSAR obeys a  $\mathcal{X}^2$  distribution with  $n - 2 = 33$  d.o.f. For the given sample the normalized sum of squares around regression holds

$$\text{NSSAR} = \frac{\text{SSAR}}{\sigma^2} = \frac{3.6089886}{0.27^2} = 49.50601607.$$

The goodness of fit can then be calculated by a numerical integration

$$Q = \text{Probability}\{\mathcal{X}^2 \geq 49.50601607\} = \int_{49.50601607}^{+\infty} p_{33}(\mathcal{X}^2) d\mathcal{X}^2$$

for instance by using the Excel function CHIDIST:

$$\text{CHIDIST}(49, 50601607; 33) \quad \Rightarrow \quad \int_{49.50601607}^{+\infty} p_{33}(\mathcal{X}^2) d\mathcal{X}^2 = 0.032443826.$$

Alternatively, we may execute the Maple command line

$$1 - \text{stats}[\text{statevalf}, \text{cdf}, \text{chisquare}[33]](49.50601607);$$

to obtain the “exact” value  $Q = 0.0324438259$ .

A satisfactory approximation can be obtained, however, by using the table of the upper critical values of  $\mathcal{X}^2$  with  $\nu = 33$  d.o.f., which provides

Probability $\{\mathcal{X}^2 \geq 47.400\}$	Probability $\{\mathcal{X}^2 \geq 50.725\}$
0.05	0.025

so that a simple linear interpolation scheme:

47.400	0.05
49.5060	$Q$
50.725	0.025

$$\frac{49.506 - 47.400}{50.725 - 47.400} = \frac{Q - 0.05}{0.025 - 0.05}$$

leads to the reasonable estimate:

$$Q = 0.05 + (0.025 - 0.05) \frac{49.506 - 47.400}{50.725 - 47.400} = 0.03416.$$

The goodness of fit of the regression model is thus equal to about 3.2%: such a percentage, if the regression model were rejected, would express the probability of a type I error.