

**Methods of statistical and numerical analysis  
(integrated course). Part I**

**Stefano Siboni**

**Working examples of statistical methods**

**Example 1. Confidence intervals**

Let us consider the following table of data, concerning some repeated measurements of a quantity (in arbitrary units). The sample is assumed to be normal.

$i$	$x_i$	$i$	$x_i$
1	1.0851	13	1.0768
2	834	14	842
3	782	15	811
4	818	16	829
5	810	17	803
6	837	18	811
7	857	19	789
8	768	20	831
9	842	21	829
10	786	22	825
11	812	23	796
12	784	24	841

We want to determine the confidence interval (CI) of the mean and that of the variance, both with confidence level  $1 - \alpha = 0.95$

**Solution**

Number of data:  $n = 24$

$$\text{Sample mean: } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 1.08148$$

$$\text{Sample variance: } s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = 6.6745 \cdot 10^{-6}$$

$$\text{Estimated standard deviation: } s = \sqrt{s^2} = 0.002584$$

□ The CI of the mean is

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

and for  $\alpha = 0.05$ ,  $n = 24$  has therefore the limits

$$\bar{x} - t_{[0.975](23)} \frac{s}{\sqrt{24}} = 1.08148 - 2.069 \cdot \frac{0.002584}{\sqrt{24}} = 1.08039$$

$$\bar{x} + t_{[0.975](23)} \frac{s}{\sqrt{24}} = 1.08148 + 2.069 \cdot \frac{0.002584}{\sqrt{24}} = 1.08257$$

so that the CI writes

$$1.08039 \leq \mu \leq 1.08257$$

or, equivalently,

$$[1.08148 \pm 0.00109]$$

□ The CI of the variance takes the form

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

with  $\alpha = 0.05$  and  $n = 24$ . Thus

$$\frac{1}{\chi^2_{[0.975](23)}} 23 s^2 = \frac{1}{38.076} 23 \cdot 6.6745 \cdot 10^{-6} = 4.03177 \cdot 10^{-6}$$

$$\frac{1}{\chi^2_{[0.025](23)}} 23 s^2 = \frac{1}{11.689} 23 \cdot 6.6745 \cdot 10^{-6} = 13.1332 \cdot 10^{-6}$$

and the CI becomes

$$4.03177 \cdot 10^{-6} \leq \sigma^2 \leq 13.1332 \cdot 10^{-6}$$

**Example 2. CI and hypothesis testing on the mean**

Let us consider the data of the following table:

$i$	$x_i$	$i$	$x_i$
1	449	16	398
2	391	17	472
3	432	18	449
4	459	19	435
5	389	20	386
6	435	21	388
7	430	22	414
8	416	23	376
9	420	24	463
10	381	25	344
11	417	26	353
12	407	27	400
13	447	28	438
14	391	29	437
15	480	30	373

By assuming a normal population, we want to test the null hypothesis that the mean  $\mu$  is  $\mu_0 = 425$ . against the alternative hypothesis that  $\mu \neq 425$ , with significance level  $\alpha = 0.02$ .

**Solution**

Number of data:  $n = 30$

$$\text{Sample mean: } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = 415.66$$

$$\text{Sample variance: } s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 = 1196.44$$

$$\text{Estimated standard deviation: } s = \sqrt{s^2} = 34.59$$

The test variable is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

and the two-sided rejection region writes

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

with  $\alpha = 0.02$ ,  $n = 30$ .

We calculate the value of the test variable

$$t = \frac{415.66 - 425}{34.59/\sqrt{30}} = -1.47896$$

and the critical point

$$t_{[1-\frac{\alpha}{2}](n-1)} = t_{[0.99](29)} = 2.462$$

so that the rejection region becomes

$$\{t < -2.462\} \cup \{t > 2.462\}$$

and **does not contain** the value of  $t$ .

**The hypothesis  $\mu = 425$  cannot be rejected** on the basis of the given sample and with significance level 0.02.

**Remark - Confidence interval for  $\mu$** 

The CI of the mean has the lower and upper limits:

$$\bar{x} - t_{[0.99](29)} \frac{s}{\sqrt{30}} = 415.66 - 2.462 \cdot \frac{34.59}{\sqrt{30}} = 400.11$$

$$\bar{x} + t_{[0.99](29)} \frac{s}{\sqrt{30}} = 415.66 + 2.462 \cdot \frac{34.59}{\sqrt{30}} = 431.21$$

and reduces then to

$$400.11 \leq \mu \leq 431.21$$

The suggested mean  $\mu_0 = 425$  **actually belongs** to the confidence interval of  $\mu$ .

**Generally speaking:**

The test variable takes a value in the rejection region of  $H_0 : \mu = \mu_0$  if and only if the tested value  $\mu_0$  of the mean belongs to the confidence interval of  $\mu$  — the significance level of the test,  $\alpha$ , and the confidence level of the CI,  $1 - \alpha$ , are complementary to 1.

### Example 3. Chauvenet criterion

Repeated measurements of a length  $x$  have provided the following results (expressed in mm):

46 , 48 , 44 , 38 , 45 , 47 , 58 , 44 , 45 , 43

and can be assumed to be extracted from a normal population. By applying Chauvenet criterion, we want to check whether the anomalous result (outlier)  $x_{\text{sus}} = 58$  must be rejected.

#### Solution

The sample mean and standard deviation are given by

$$\bar{x} = \frac{1}{10} \sum_{i=1}^{10} x_i = 45.8 \quad s = \sqrt{\frac{1}{9} \sum_{i=1}^{10} (x_i - \bar{x})^2} = 5.1$$

The distance of the suspect value from the mean, in units of  $s$ , holds

$$\frac{x_{\text{sus}} - \bar{x}}{s} = \frac{58 - 45.8}{5.1} = 2.4$$

The probability that a measurement falls at a distance larger than 2.4 standard deviations from the mean can be calculated from the Table of the cumulative distribution function of the standard normal distribution:

$$\begin{aligned} P(|x_{\text{sus}} - \bar{x}| \geq 2.4s) &= 1 - P(|x_{\text{sus}} - \bar{x}| < 2.4s) = \\ &= 1 - 2 \cdot P(\bar{x} \leq x_{\text{sus}} < \bar{x} + 2.4s) = \\ &= 1 - 2 \cdot 0.49180 = 0.0164 \end{aligned}$$

Out of 10 measurements we typically expect  $10 \cdot 0.0164 = 0.164$  “bad” results, at a distance larger than  $2.4s$  from the mean. **Since  $0.164 < 1/2$  Chauvenet criterion suggests that the outlier  $x_{\text{sus}}$  should be rejected.**

**Example 4. Hypothesis testing on the mean**

Four measurements of a quantity  $x$  have yielded the following results:

$$1269 \quad , \quad 1271 \quad , \quad 1263 \quad , \quad 1265$$

In the hypothesis of a normal population, we want to test at a 5%-significance level whether the previous data are in agreement with a hypothetical true value  $\mu_0 = 1260$  of  $x$ .

**Solution**

The estimated mean and standard deviation can be easily calculated:

$$\bar{x} = \frac{1}{4} \sum_{i=1}^4 x_i = 1267 \quad s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} = 3.65$$

so that the test variable holds:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{1267 - 1260}{3.65/\sqrt{4}} = 3.83$$

Whenever the null hypothesis  $\mu = \mu_0$  is correct, the latter obeys a Student's  $t$  distribution with  $n-1$  d.o.f.; at a significance level  $\alpha = 0.05$ , the null hypothesis should be accepted if

$$|t| \leq t_{[1-\frac{\alpha}{2}](n-1)} = t_{[0.975](3)} = 3.182$$

Since  $t = 3.83 > 3.182 = t_{[0.975](3)}$ , we conclude that the hypothesis  $\mu = \mu_0$  should be rejected. **The experimental data do not support the conjecture that the true value of the quantity  $x$  is 1260.**

### Example 5. Unpaired $t$ -test for the comparison of two means (same variance)

Two processes are characterized by the values listed in the table below:

Process 1	Process 2
9	14
4	9
10	13
7	12
9	13
10	8
	10

We want to test, with a 5%-significance level, the hypothesis that the mean values  $\mu_1$  and  $\mu_2$  of the processes are the same, by assuming normal populations and equal variances ( $\sigma_1^2 = \sigma_2^2$ ).

#### Solution

The two samples are composed by  $p = 6$  and  $q = 7$  elements, respectively. The mean and the variance of the first sample can be written as:

$$\bar{x} = \frac{1}{p} \sum_{i=1}^p x_i = 8.17 \quad s_x^2 = \frac{1}{p-1} \sum_{i=1}^p (x_i - \bar{x})^2 = 5.37$$

while those of the second sample hold:

$$\bar{y} = \frac{1}{q} \sum_{j=1}^q y_j = 11.29 \quad s_y^2 = \frac{1}{q-1} \sum_{j=1}^q (y_j - \bar{y})^2 = 5.24$$

The pooled variance of the two samples is then calculated in the form:

$$s^2 = \frac{(p-1)s_x^2 + (q-1)s_y^2}{p+q-2} = \frac{5 \cdot 5.37 + 6 \cdot 5.24}{11} = 5.299$$

As a test variable for testing the hypothesis  $H_0 : \mu_1 = \mu_2$ , we use the quotient:

$$t = \frac{\bar{x} - \bar{y}}{s \sqrt{\frac{1}{p} + \frac{1}{q}}} = \frac{8.17 - 11.29}{\sqrt{5.299 \left( \frac{1}{6} + \frac{1}{7} \right)}} = -2.436$$

If the null hypothesis holds true, such a random variable is a Student's  $t$  with  $p+q-2$  d.o.f. so that the two-sided acceptance region must be taken as:

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

In the present case we have  $\alpha = 0.05$ ,  $p = 6$ ,  $q = 7$  and the acceptance region becomes

$$|t| \leq t_{[0.975]}(11) = 2.201$$

Since  $|t| = 2.436 > 2.201 = t_{[0.975]}(11)$ , our samples suggest that the null hypothesis  $\mu_1 = \mu_2$  must be rejected: **we conclude that the difference between the means of the two processes appears significant.**

**Example 6. Regression straight line**

The following  $(x, y)$  data have been obtained by an experiment:

$x_i$	$y_{i1}$	$y_{i2}$	$y_{i3}$	$y_{i4}$	$y_{i5}$
1.00	2.11	2.12	2.07	×	×
0.50	2.34	2.39	2.39	2.37	×
0.33	2.47	2.45	2.38	2.38	2.42
0.14	2.51	2.48	2.54	2.52	2.55
0.04	2.62	2.63	2.61	2.55	2.57

All the data are assumed to be independent normal variables with the same standard deviation. We want to model the data by a best-fit straight line of the form

$$y = m + q(x - \bar{x})$$

where  $\bar{x}$  denotes the sample mean of the independent variables  $x_i$ 's and the model parameters  $m$  and  $q$  must be calculated with the least-square method. In particular, we want to determine:

- (i) the 95%-confidence interval of the intercept  $m$ ;
- (ii) the 95%-confidence interval of the slope  $q$ ;
- (iii) the 95%-confidence interval for the prediction of  $y$  at a given value  $x = 0.45$  of the independent variable

## Solution

Since the standard deviations are assumed to be the same, the chi-square fitting reduces to the usual least square fitting and the best-fit estimates of the parameters  $m$  and  $q$  can be written in the form:

$$m = \frac{1}{n} \sum_{i=1}^n y_i = 2.4305 \quad q = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = -0.4995$$

with  $n = 22$ . The sum of squares around regression holds then:

$$\text{SSAR} = \sum_{i=1}^n [m + q(x_i - \bar{x}) - y_i]^2 = 0.02223246$$

while  $\alpha = 0.05$ . At a confidence level  $1 - \alpha \in (0, 1)$  the confidence interval of the intercept  $\mu$  and that of the slope  $\kappa$  are given by:

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

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

In the present case the confidence intervals become therefore:

$$\mu = m \pm t_{[0.975](20)} \sqrt{\frac{1}{22} \frac{\text{SSAR}}{20}}$$

$$\kappa = q \pm t_{[0.975](20)} \sqrt{\left[ \sum_{i=1}^{22} (x_i - \bar{x})^2 \right]^{-1} \frac{\text{SSAR}}{20}}$$

with:

$$m = 2.4305$$

$$q = -0.4995$$

$$\text{SSAR} = 0.02223246$$

$$\sum_{i=1}^{22} (x_i - \bar{x})^2 = 2.05947727$$

$$t_{[0.975](20)} = 2.086$$

As a conclusion:

(i) the 95%-confidence interval for the intercept  $\mu$  is

$$2.4305 \pm 0.0148 = [2.415, 2.445]$$

(ii) the 95%-confidence interval for the slope  $\kappa$  holds

$$-0.4995 \pm 0.0485 = [-0.548, -0.451]$$

(iii) Since the model is assumed to be homoscedastic, the confidence interval (at a confidence level  $1 - \alpha$ ) for the prediction of  $y = y_0$  at a given abscissa  $x = x_0$  is expressed by the general formula:

$$\mathbb{E}(y_0) = m + q(x_0 - \bar{x}) \pm t_{[1-\frac{\alpha}{2}](n-2)} \sqrt{V} \sqrt{\frac{\text{SSAR}}{n-2}}$$

where:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

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

$$V = 1 + \frac{1}{n} + \frac{1}{n \sum_{i=1}^n (x_i - \bar{x})^2} (x_0 - \bar{x})^2.$$

In the present case we have  $\bar{x} = 0.34318$ , so that:

$$y_0 = 2.4305 - 0.4995(x_0 - 0.34318) \pm t_{[0.975](20)} \cdot \sqrt{1 + \frac{1}{22} + \frac{1}{2.05947727}(x_0 - 0.34318)^2} \sqrt{\frac{0.02223246}{20}}$$

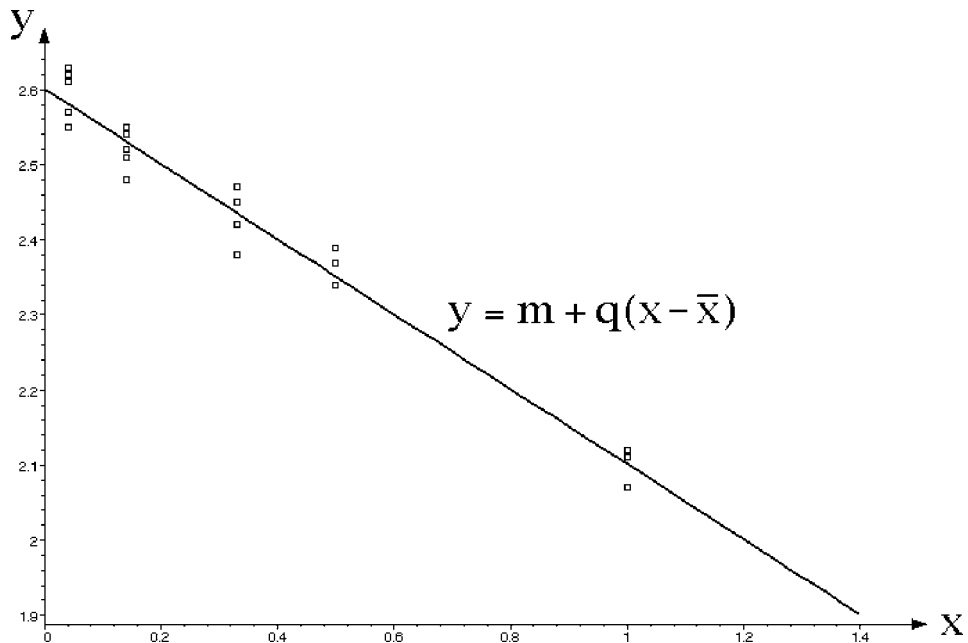
and the confidence interval for the prediction of  $y$  at  $x = x_0$  reduces to:

$$y_0 = 2.4305 - 0.4995(x_0 - 0.34318) \pm 0.06955 \sqrt{1.04545 + 0.48556(x_0 - 0.34318)^2}$$

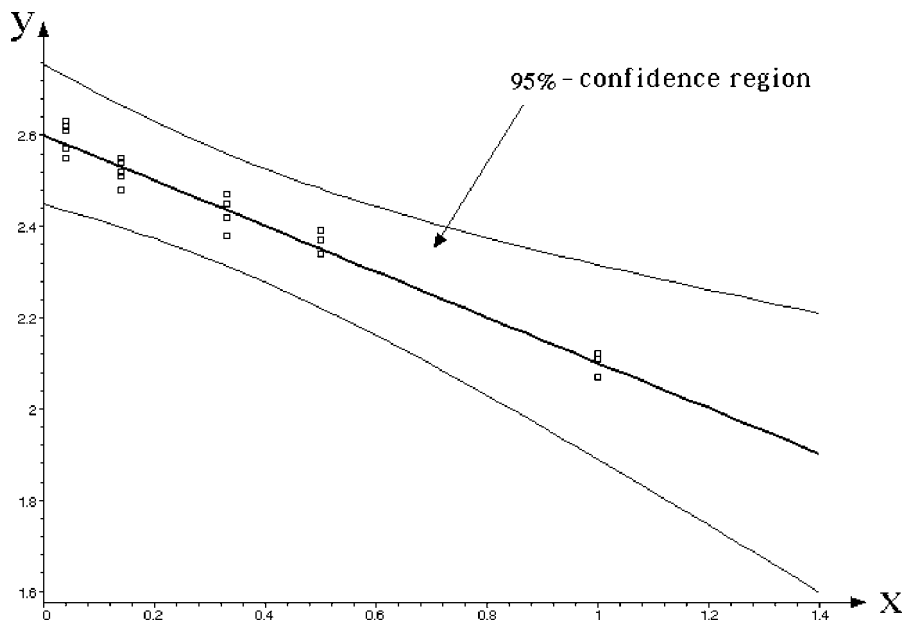
For  $x_0 = 0.45$  we get:

$$y_0 = [2.377 \pm 0.075] = [2.302, 2.452]$$

In the following picture the regression line is superimposed to the experimental data (dots):



The confidence region for predictions (at a confidence level of 95%) is evidenced in the figure below (factor  $V$  exaggerated)



**Example 7. Regression straight line**

Experimental measurements of a quantity  $y$  versus another quantity  $x$  are collected in the table below

$k$	$x_k$	$y_{k1}$	$y_{k2}$	$y_{k3}$	$y_{k4}$
1	0.6	0.55	0.80	0.75	×
2	1.0	1.10	0.95	1.05	×
3	1.5	1.40	1.30	1.50	×
4	2.0	1.60	1.90	1.85	1.95
5	2.5	2.25	2.30	2.40	2.45
6	2.9	2.65	2.70	2.75	×
7	3.0	2.65	2.70	2.80	2.85

The random error on the abscissas  $x_k$ 's is negligible, while the data  $y_k$ 's are independent random variables with the same variance  $\sigma^2$  (homoscedastic system).

Determine:

- (i) the regression straight-line by the least-square method, in the form

$$y = \mu + \kappa(x - \bar{x});$$

- (ii) the 95%-confidence interval for the intercept  $\mu$  and that for the slope  $\kappa$ ;
- (iii) the 95%-confidence region for predictions;
- (iv) the 95%-confidence interval for the predicted value of  $y$  at  $x = 2.3$ ;
- (v) the goodness of fit of the regression model if  $\sigma = 0.08$ .

**Solution**

- (i) Since the standard deviations are equal, the chi-square fitting reduces to the usual least square fitting and the best-fit estimates of parameters  $m$  and  $q$  can be written as:

$$m = \frac{1}{n} \sum_{i=1}^n y_i = 1.8833 \quad q = \frac{\sum_{i=1}^n (x_i - \bar{x}) y_i}{\sum_{i=1}^n (x_i - \bar{x})^2} = 0.87046$$

with  $n = 24$  and  $\bar{x} = 2.000$ . The regression line, as calculated with the least square method, writes therefore:

$$\begin{aligned} y &= m + q(x - \bar{x}) = 1.8833 + 0.87046(x - 2.000) = \\ &= 0.1424 + 0.87046 x \end{aligned}$$

- (ii) The sum of squares around regression holds:

$$\text{SSAR} = \sum_{i=1}^n [m + q(x_i - \bar{x}) - y_i]^2 = 0.22704376$$

At a confidence level  $1 - \alpha \in (0, 1)$  the confidence interval of the intercept  $\mu$  and that of the slope  $\kappa$  take the form:

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

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

In the present case we have  $\alpha = 0.05$ ,  $n = 24$  and the confidence intervals become therefore:

$$\mu = m \pm t_{[0.975](22)} \sqrt{\frac{1}{24} \frac{\text{SSAR}}{22}}$$

$$\kappa = q \pm t_{[0.975](22)} \sqrt{\left[ \sum_{i=1}^{24} (x_i - \bar{x})^2 \right]^{-1} \frac{\text{SSAR}}{22}}$$

with:

$$m = 1.8833$$

$$q = 0.87046$$

$$\text{SSAR} = 0.22704376$$

$$\sum_{i=1}^{24} (x_i - \bar{x})^2 = 17.060000$$

$$t_{[0.975](22)} = 2.074$$

As a conclusion:

- the 95%-confidence interval for the intercept  $\mu$  is

$$1.8833 \pm 0.0430 = [1.840, 1.926]$$

- the 95%-confidence interval for the slope  $\kappa$  holds

$$0.87046 \pm 0.05101 = [0.819, 0.921]$$

(iii) Since the model is assumed to be homoscedastic, the confidence interval (at a confidence level  $1 - \alpha$ ) for the prediction of  $y = y_0$  at a given abscissa  $x = x_0$  can be calculated by the general formula:

$$\mathbb{E}(y_0) = m + q(x_0 - \bar{x}) \pm t_{[1-\frac{\alpha}{2}](n-2)} \sqrt{V} \sqrt{\frac{\text{SSAR}}{n-2}}$$

where:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

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

$$V = 1 + \frac{1}{n} + \frac{1}{n \sum_{i=1}^n (x_i - \bar{x})^2} (x_0 - \bar{x})^2.$$

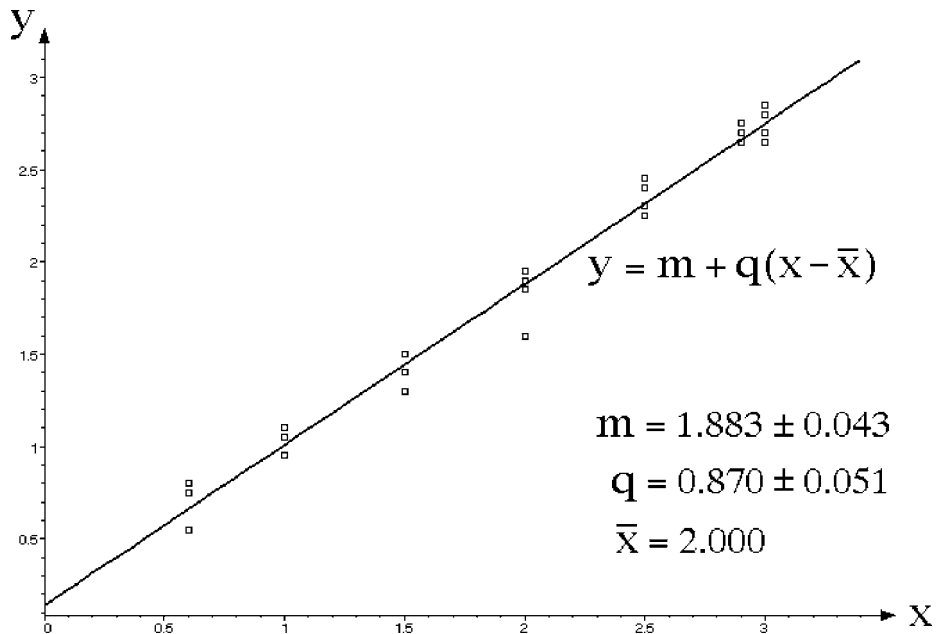
In the present case we have  $\bar{x} = 2.000$ , so that:

$$y_0 = 1.8833 + 0.87046(x_0 - 2.000) \pm t_{[0.975](22)} \cdot \sqrt{1 + \frac{1}{24} + \frac{1}{17.060000} (x_0 - 2.000)^2} \sqrt{\frac{0.22704376}{22}}$$

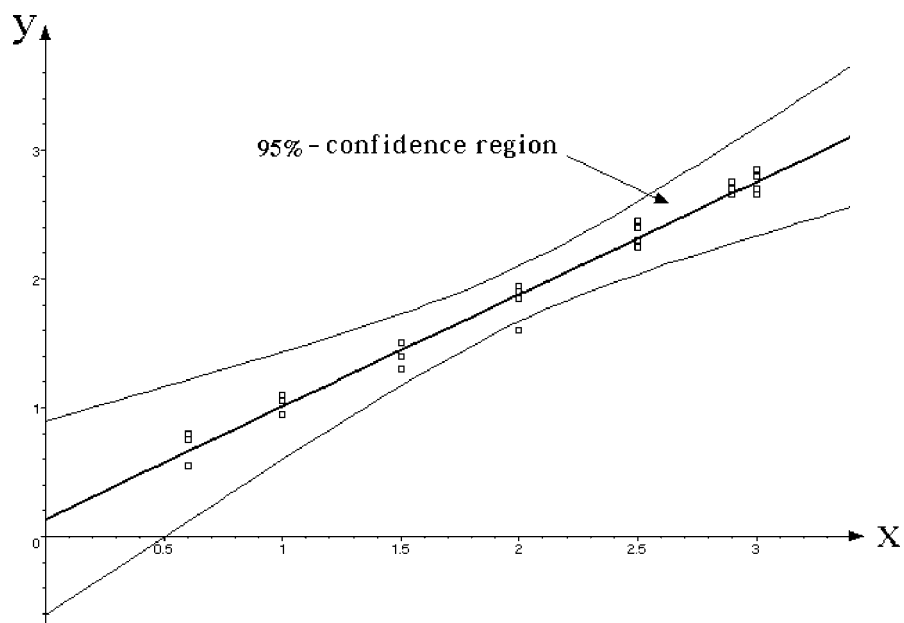
and the confidence interval for the prediction of  $y$  at  $x = x_0$  reduces to:

$$y_0 = 1.8833 + 0.87046(x_0 - 2.000) \pm 0.21069 \sqrt{1.04167 + 0.058617(x_0 - 2.000)^2}$$

In the following picture the regression line is superimposed to the experimental data (dots):



The confidence region for predictions (at a confidence level of 95%) is evidenced in the figure below (factor  $V$  exaggerated)



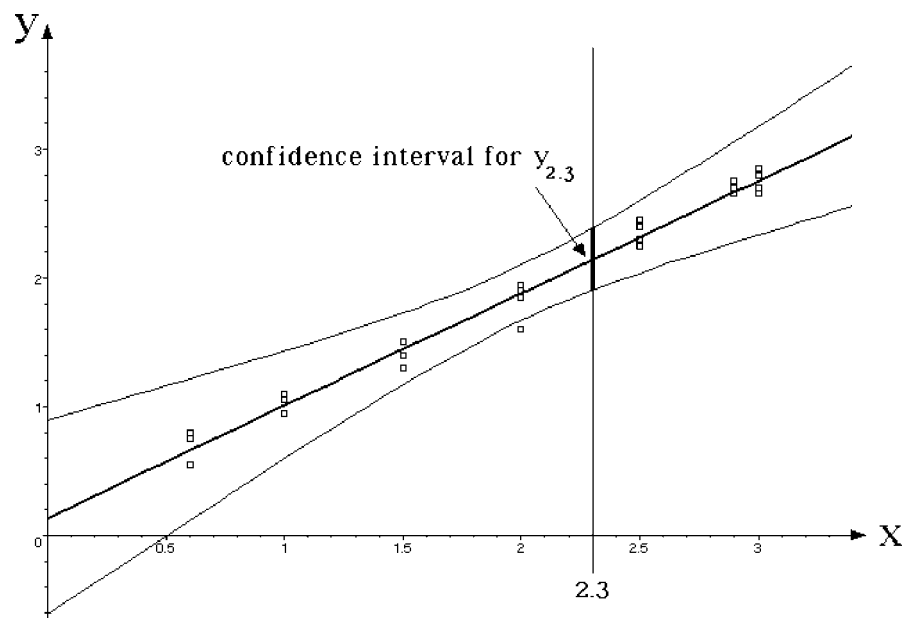
(iv) the 95%-confidence interval for the prediction of  $y$  at  $x = 2.3$  can be obtained by replacing  $x_0 = 2.3$  in the previous formula

$$y_0 = 1.8833 + 0.87046(x_0 - 2.000) \pm 0.21069 \sqrt{1.04167 + 0.058617(x_0 - 2.000)^2}$$

We have therefore:

$$y_{2.3} = [1.903, 2.385] = 2.144 \pm 0.241$$

In the following graph the confidence interval is the intersection of the 95%-confidence region with the vertical line  $x = 2.3$ :



(v) the goodness of fit  $Q$  of the regression model is defined by the relationship

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

where  $\rho_{n-p}$  denotes the  $\mathcal{X}^2$  distribution with  $n - p$  d.o.f.

This is because, if the regression model holds true, the normalized sum of squares around regression

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

is a  $\mathcal{X}^2$  random variable with  $n - p$  d.o.f.

In order to make calculations **it is crucial to know the common value of the standard deviation  $\sigma = 0.08$** , since NSSAR and not simply SSAR is needed.

In the case under scrutiny we have  $n = 24$  data and the regression model is based on two parameters,  $\mu$  and  $\kappa$ ; as a consequence,  $p = 2$  and the NSSAR follows a  $\mathcal{X}^2$  distribution with  $n - p = 22$  d.o.f.

For the given sample the normalized sum of squares holds

$$\text{NSSAR} = \frac{\text{SSAR}}{\sigma^2} = \frac{0.22704376}{0.08^2} = 35.47558$$

From the table of upper critical values of  $\chi^2$  with  $\nu = 22$  d.o.f. we find

Probability $\{\chi^2 \geq 33.924\}$	Probability $\{\chi^2 \geq 36.781\}$
0.05	0.025

so that a simple linear interpolation scheme:

33.924	0.05
35.476	$Q$
36.781	0.025

$$\frac{35.476 - 33.924}{36.781 - 33.924} = \frac{Q - 0.05}{0.025 - 0.05}$$

provides the required estimate of  $Q$ :

$$Q = 0.05 + (0.025 - 0.05) \frac{35.476 - 33.924}{36.781 - 33.924} = 0.0364$$

A more accurate  $Q$  follows *via* a numerical integration

$$Q = \text{Probability}\{\chi^2 \geq 35.47558\} = \int_{35.47558}^{+\infty} p_{22}(\chi^2) d\chi^2$$

for instance by the Maple V command line

$$1 - \text{stats}[\text{statevalf}, \text{cdf}, \text{chisquare}[22]](35.475587);$$

which leads to  $Q = 0.0345$ .

If the regression model were rejected,  $Q$  would express the probability of a type I error — about 3.5% in the present case.

**Example 8. Linear correlation coefficient (Pearson)**

Let us consider the following table of data, concerning some repeated measurements of two quantities  $x$  and  $y$  (in arbitrary units), in the same experimental conditions. The sample is assumed to be normal.

$i$	$x_i$	$y_i$
1	1.8	1.1
2	3.2	1.6
3	4.3	1.1
4	5.9	3.5
5	8.1	3.8
6	9.9	6.2
7	11.4	5.6
8	12.1	4.0
9	13.5	7.5
10	17.9	7.1

We want to use Pearson's correlation coefficient to check if the quantities  $x$  and  $y$  can be regarded as stochastically independent or not, with a significance level  $\alpha = 5\%$  and  $\alpha = 1\%$ .

**Solution**

The sample means of the two quantities are given by

$$\bar{x} = \frac{1}{10} \sum_{i=1}^{10} x_i = 8.8100 \quad \bar{y} = \frac{1}{10} \sum_{i=1}^{10} y_i = 4.1500$$

and allow us to compute the following sums of residuals

$$SS_{xy} = \sum_{i=1}^{10} (x_i - \bar{x})(y_i - \bar{y}) = 87.7220$$

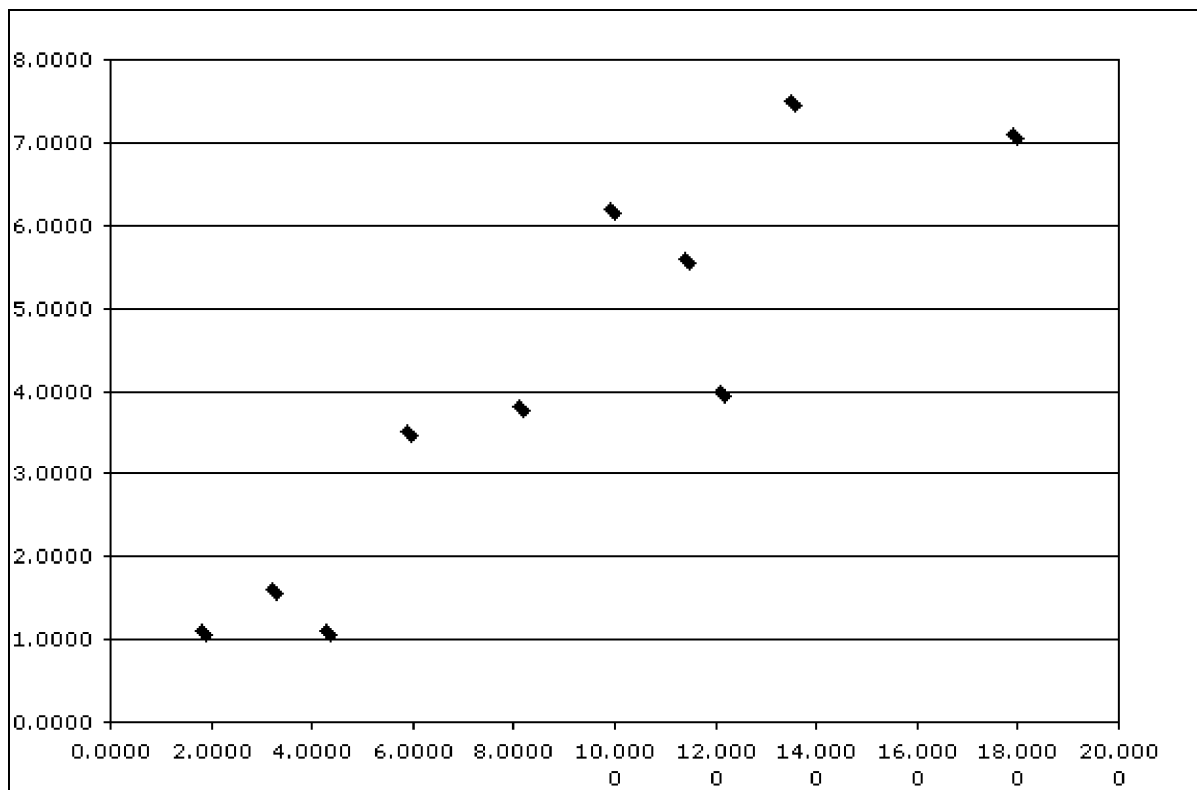
$$SS_{xx} = \sum_{i=1}^{10} (x_i - \bar{x})^2 = 202.6994$$

$$SS_{yy} = \sum_{i=1}^{10} (y_i - \bar{y})^2 = 51.9050$$

so that the linear correlation coefficient becomes

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}} \sqrt{SS_{yy}}} = \frac{87.7220}{\sqrt{202.6994} \sqrt{51.9050}} = 0.8552$$

The plot of data suggests that  $x$  and  $y$  are dependent:



Since  $x$  and  $y$  can be assumed normal, we can check the null hypothesis

$$H_0 : x \text{ and } y \text{ are stochastically independent}$$

against the alternative hypothesis

$$H_1 : x \text{ and } y \text{ are stochastically dependent}$$

by using the test variable

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

which follows a Student's  $t$  distribution with  $n-2$  d.o.f. whenever  $H_0$  holds true.

In the present case we have

$$t = \sqrt{10-2} \frac{0.8552}{\sqrt{1-0.8552^2}} = 4.6673$$

The rejection region takes the form

$$|t| > t_{[1-\frac{\alpha}{2}](n-2)} = t_{[0.975](8)} = 2.306$$

for a significance level  $\alpha = 5\%$ , and the form

$$|t| > t_{[1-\frac{\alpha}{2}](n-2)} = t_{[0.995](8)} = 3.355$$

if the significance level is assumed to be  $\alpha = 1\%$ .

**In both cases  $H_0$  must be rejected!**

**Example 9.  $F$ -test on the variances of two independent normal populations**

In order to enhance a chemical reaction we must choose between two different catalysts, say,  $A$  and  $B$ . To check if the variance of the amount of product is the same or not, for the two catalysts, we carry out  $p = 10$  experiments with  $A$  and  $q = 12$  experiments with  $B$  and obtain the following sample estimates:

$$s_A^2 = 0.1405 \qquad s_B^2 = 0.2820$$

At a significance level of 5%, can we reject the hypothesis that  $\sigma_A^2 = \sigma_B^2$ ? The two populations are assumed independent and normal.

**Solution**

We can use the test variable

$$F = s_A^2/s_B^2$$

which follows a Fisher distribution with  $(p - 1, q - 1) = (9, 11)$  d.o.f. if the null hypothesis  $H_0 : \sigma_A^2 = \sigma_B^2$  is satisfied.

We have

$$F = 0.1405/0.2820 = 0.49822695$$

The rejection region is

$$\{F \leq F_{[\frac{\alpha}{2}]}(p-1, q-1)\} \cup \{F \geq F_{[1-\frac{\alpha}{2}]}(p-1, q-1)\}$$

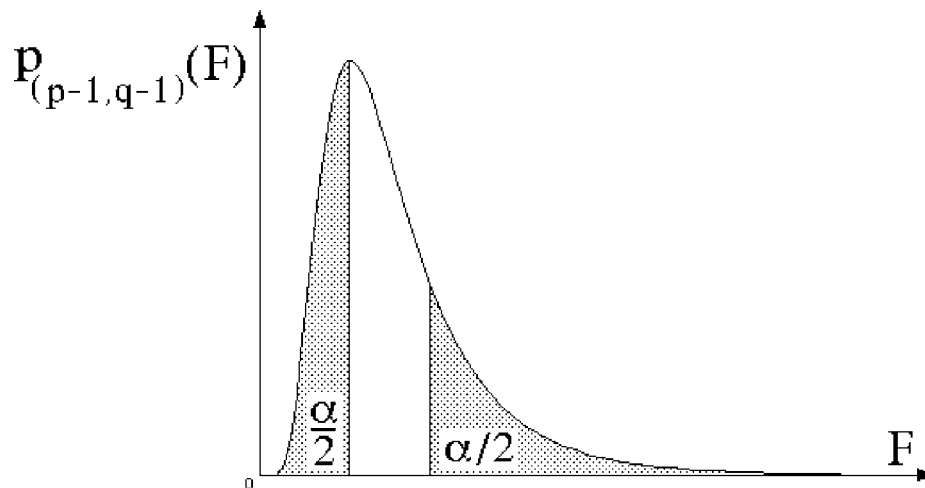
i.e. (since  $p = 10$ ,  $q = 12$  and  $\alpha = 0.05$ )

$$\{F \leq F_{[0.025]}(9, 11)\} \cup \{F \geq F_{[0.975]}(9, 11)\}$$

where

$$F_{[0.025](9,11)} = 0.2556188$$

$$F_{[0.975](9,11)} = 3.587898$$



The  $F$ -values are not available on the table, but can be easily calculated by numerical tools. For instance, by the following Maple commands

$$\text{statevalf}[\text{icdf}, \text{fratio}[9, 11]](0.025)$$

$$\text{statevalf}[\text{icdf}, \text{fratio}[9, 11]](0.975)$$

In the present case

$$0.49822695 \notin \{F \leq 0.2556188\} \cup \{F \geq 3.587898\}$$

thus we must conclude that our samples cannot exclude that the variances of the two populations are actually the same.

**$H_0$  cannot be rejected!**

**Example 10. Comparison of the means of two normal populations with unknown, but presumably equal variances (unpaired  $t$ -test)**

At a medical research center, a group of 22 volunteers is exposed to various types of influenza viruses and kept under medical control. A prescribed dose of vitamin C is supplied to a random sample of  $p = 10$  volunteers. A placebo, indistinguishable from the drug, is supplied instead to the other  $q = 12$  volunteers. For each volunteer, the duration (in days) of the illness is recorded. The result is the list below

vitamin C	placebo
5.5	6.5
6.0	6.0
7.0	8.5
6.0	7.0
7.5	6.5
6.0	8.0
7.5	7.5
5.5	6.5
7.0	7.5
6.5	6.0
	8.5
	7.0

The populations may be assumed independent, normal and presumably with the same (unknown) variance. We want to test, with a 5%-significance level, the hypothesis that the mean duration of the illness is the same for volunteers treated with vitamin C ( $\mu_C$ ) and for those treated with placebo ( $\mu_p$ ), against the alternative hypothesis that  $\mu_C < \mu_p$ .

## Solution

The two samples consist in  $p = 10$  and  $q = 12$  elements, respectively. Since the alternative hypothesis is one-sided, we can apply a one-sided version of the unpaired  $t$ -test for the means of two independent normal populations with the same unknown variance to check

$$H_0 : \mu_C = \mu_p \quad \text{against} \quad H_1 : \mu_C < \mu_p$$

Denoted with  $x_i$  the illness durations for the vitamin-treated sample and with  $y_j$  the placebo-treated ones, the mean and the variance of the first sample can be written as:

$$\bar{x} = \frac{1}{p} \sum_{i=1}^p x_i = 6.4500 \quad s_x^2 = \frac{1}{p-1} \sum_{i=1}^p (x_i - \bar{x})^2 = 0.5806$$

while those of the second sample hold:

$$\bar{y} = \frac{1}{q} \sum_{j=1}^q y_j = 7.1250 \quad s_y^2 = \frac{1}{q-1} \sum_{j=1}^q (y_j - \bar{y})^2 = 0.7784$$

The pooled variance of the two samples is then reckoned as:

$$\begin{aligned} s^2 &= \frac{(p-1)s_x^2 + (q-1)s_y^2}{p+q-2} = \\ &= \frac{9 \cdot 0.5806 + 11 \cdot 0.7784}{20} = 0.6894 \end{aligned}$$

For testing the hypothesis  $H_0 : \mu_C = \mu_p$ , we use the variable:

$$t = \frac{\bar{x} - \bar{y}}{s \sqrt{\frac{1}{p} + \frac{1}{q}}} = \frac{6.4500 - 7.1250}{\sqrt{0.6894 \left( \frac{1}{10} + \frac{1}{12} \right)}} = -1.8987$$

Whenever the null hypothesis holds true, such a random variable is a Student's  $t$  with  $p + q - 2$  d.o.f. so that the one-sided rejection region must be defined as:

$$t \leq t_{[\alpha](p+q-2)} = t_{[0.05](20)} = -t_{[0.95](20)} = -1.725$$

since  $t_{[\alpha](p+q-2)} = -t_{[1-\alpha](p+q-2)}$  and  $p = 10$ ,  $q = 12$ ,  $\alpha = 0.05$ .

In the present case we have

$$t = -1.8987 < -1.725 = -t_{[0.95](20)}$$

Therefore our samples suggest that the null hypothesis  $\mu_C = \mu_p$  must be rejected: **we conclude that the difference between the means of the two processes appears significant** and that presumably the mean duration  $\mu_C$  of the illness under vitamin C treatment is smaller than that ( $\mu_p$ ) recorded after placebo supply.

### **Example 11. Test on the probability parameter of a Bernoulli population**

A manufacturer of integrated circuits claims that no more than 2% of the pieces he produces are faulty. A company of electronic products buys a large stock of such circuits. A sample of 300 pieces is tested, and 10 of them are found to be faulty. Is it possible to deny manufacturer's claim?

#### **Solution**

As usual, although not too realistically, we can assume that all the pieces are manufactured independently and that a constant probability  $\pi$  of producing a faulty piece is defined. Then the

whole number of faulty circuits out of a sample of  $n$  pieces follows a Bernoulli distribution with parameters  $\pi$  and  $n$ :

$$p_{n,\pi}(x) = \frac{n!}{x!(n-x)!} \pi^x (1-\pi)^{n-x}, \quad x = 0, 1, \dots, n.$$

We must check the null hypothesis

$$H_0 : \pi = \pi_0 = 2/100 = 0.02$$

against the alternative hypothesis

$$H_1 : \pi > \pi_0 = 2/100 = 0.02$$

We adopt a one-sided test and reject  $H_0$  whenever the number  $x$  of pieces is sufficiently large. The rejection region takes then the form

$$\{x = 0, 1, \dots, n, \quad x \geq x_{cr}\}$$

where the positive integer  $x_{cr}$  is defined in terms of the the significance level  $\alpha$  of the test by the equation

$$\text{probability}(x \geq x_{cr}) = \sum_{x=x_{cr}}^n \frac{n!}{x!(n-x)!} \pi_0^x (1-\pi_0)^{n-x} = \alpha$$

with  $n = 300$  and  $\pi_0 = 0.02$ .  $H_0$  is rejected if  $10 \geq x_{cr}$ .

Alternatively, assuming  $H_0$  correct, we can simply compute the probability that the number of faulty pieces is  $\geq 10$ :

$$\text{probability}(x \geq 10)$$

and reject  $H_0$  if the latter probability is  $\leq \alpha$  — which means that  $10 \geq x_{cr}$ .

In the present case we obtain

$$\begin{aligned} \text{probability}(x \geq 10) &= 1 - \text{probability}(x < 10) = \\ &= 1 - \sum_{x=0}^9 \binom{300}{x} 0.02^x (1 - 0.02)^{300-x} \sim 0.0818 = 8.18\% \end{aligned}$$

where it is convenient to reckon the probability of  $(x < 10)$  instead of  $(x \geq 10)$  — the evaluation is much faster because only 10 terms must be summed up instead of 291.

But for both standard significance levels  $\alpha = 5\%$  and  $\alpha = 1\%$  of the test we have

$$8.18\% > \alpha$$

and conclude that the null hypothesis cannot be rejected. **The information is insufficient to disprove manufacturer's claim that the probability of a faulty circuit is  $\pi_0 = 2\%$ .**

**Example 12.  $\chi^2$ -test for a normal distribution**

An anthropologist is interested in the heights of the natives of a certain island. He suspects that the heights of male adults should be normally distributed and measures the heights of a sample of 200 men. By using these data, he computes the sample mean  $\bar{x}$  and standard deviation  $s$ , and applies the results as an estimate for the mean  $\mu$  and the standard deviation  $\sigma$  of the expected normal distribution. Then he chooses eight equal intervals where he groups the results of his records (empirical frequencies). He deduces the table below:

i	height interval	empirical frequency	expected frequency
1	$x < \bar{x} - 1.5s$	14	13.362
2	$\bar{x} - 1.5s \leq x < \bar{x} - s$	29	18.370
3	$\bar{x} - s \leq x < \bar{x} - 0.5s$	30	29.976
4	$\bar{x} - 0.5s \leq x < \bar{x}$	27	38.292
5	$\bar{x} \leq x < \bar{x} + 0.5s$	28	38.292
6	$\bar{x} + 0.5s \leq x < \bar{x} + s$	31	29.976
7	$\bar{x} + s \leq x < \bar{x} + 1.5s$	28	18.370
8	$\bar{x} + 1.5s \leq x$	13	13.362

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

**Solution**

The empirical frequencies  $f_i$  are rather large ( $f_i \gg 5$ ), so that the  $\chi^2$ -test is applicable. The expected frequencies  $n_i$  are calculated multiplying by the whole number of individuals — 200 — the integrals of the standard normal distribution over each interval. For intervals  $i = 5, 6, 7, 8$  the table of integrals be-

tween 0 and  $z$  of the standard normal provides indeed:

$$P(\bar{x} \leq x < \bar{x} + 0.5s) = \int_0^{0.5} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 0.19146$$

$$\begin{aligned} P(\bar{x} + 0.5s \leq x < \bar{x} + s) &= \int_{0.5}^1 \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \\ &= \int_0^1 \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz - \int_0^{0.5} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \\ &= 0.34134 - 0.19146 = 0.14674 \end{aligned}$$

$$\begin{aligned} P(\bar{x} + s \leq x < \bar{x} + 1.5s) &= \int_1^{1.5} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \\ &= \int_0^{1.5} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz - \int_0^1 \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \\ &= 0.43319 - 0.34134 = 0.09185 \end{aligned}$$

$$\begin{aligned} P(\bar{x} + 1.5s \leq x) &= \int_{1.5}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \\ &= \int_0^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz - \int_0^{1.5} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \\ &= 0.5 - 0.43319 = 0.06681 \end{aligned}$$

while the probabilities of the intervals  $i = 1, 2, 3, 4$  are symmetrically equal to the previous ones (due to the symmetry of the standard normal distribution with respect to the origin).

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

$$\chi^2 = \sum_{i=1}^8 \frac{(f_i - n_i)^2}{n_i} = 17.370884$$

If the proposed probability distribution is correct, the test variable obeys a  $\chi^2$  distribution with

$$8 - 1 - 2 = 5$$

d.o.f. since  $k = 8$  are the bins introduced and  $c = 2$  the parameters of the distribution ( $\mu$  and  $\sigma$ ) which are estimated by using the same data of the sample.

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

$$\chi^2_{[1-\alpha](5)} = \chi^2_{[0.95](5)} = 11.070 \quad \text{for } \alpha = 0.05$$

$$\chi^2_{[1-\alpha](5)} = \chi^2_{[0.99](5)} = 13.277 \quad \text{for } \alpha = 0.01$$

In both cases the  $\chi^2$  calculated on the sample is greater: we conclude that, with both levels of significance of 5 and 1%, **the null hypothesis must be rejected**. The data of the sample suggest that **the height distribution is not normal** for the population of the natives.

**Example 13.  $\chi^2$ -test for a discrete distribution**

A pair of dice is rolled 360 times and for each trial the score is recorded. The possible scores are, of course, 2, 3, 4, ..., 11, 12 and each of them has been obtained the number of times shown in the following table

Score	Number of results	Score	Number of results
2	6	8	44
3	14	9	49
4	23	10	39
5	35	11	27
6	57	12	16
7	50		

Check the hypothesis that the dice are not fixed with a significance level ( $\alpha$ ) of 5% and ( $\beta$ ) of 1%.

**Solution**

If the dice are not fixed, the probability of the various scores can be calculated *a priori*, with no use of the sample data:

Score	Probability	Score	Probability
2	1/36	8	5/36
3	2/36	9	4/36
4	3/36	10	3/36
5	4/36	11	2/36
6	5/36	12	1/36
7	6/36		

keeping in mind that each face has a probability 1/6 to be drawn.

The expected frequencies  $n_i$  of the various scores  $i = 2, \dots, 12$  are then those listed in the third column of the following table:

Score	Empirical frequency	Expected frequency
2	6	10
3	14	20
4	23	30
5	35	40
6	57	50
7	50	60
8	44	50
9	49	40
10	39	30
11	27	20
12	16	10

The empirical frequencies  $f_i$  are all large enough ( $> 5$ ) to allow the use of a  $\chi^2$ -test to check the hypothesis **without collecting and pooling the scores with a too low empirical frequency** (a score  $\Leftrightarrow$  an interval).

The  $\chi^2$  of data is written as:

$$\chi^2 = \sum_{i=2}^{12} \frac{(f_i - n_i)^2}{n_i} = 19.800$$

If the proposed probabilities are correct (fair dice), the test variable will follow a  $\chi^2$  distribution with

$$11 - 1 = 10$$

d.o.f., attributable to the  $n = 11$  calculated frequencies and to the unique constraint provided by the total number of recorded

results.

From the table of the  $\chi^2$  cumulative distributions we deduce the following critical values:

$$\chi^2_{[1-\alpha](10)} = \chi^2_{[0.95](10)} = 18.307 \quad \text{for } \alpha = 0.05$$

$$\chi^2_{[1-\alpha](10)} = \chi^2_{[0.99](10)} = 23.209 \quad \text{for } \alpha = 0.01$$

We find then

$$\chi^2 = 19.800 > 18.307 = \chi^2_{[0.95](10)}$$

so that **with a significance level  $\alpha = 5\%$  the hypothesis that the two dice are fair must be rejected.**

In contrast, we clearly have

$$\chi^2 = 19.800 < 23.209 = \chi^2_{[0.99](10)}$$

and therefore **the same hypothesis cannot be disproved with a significance level  $\alpha = 1\%$ .**

Notice that the probability of obtaining a value of  $\chi^2$  greater than or equato to that calculated by the sample is equal to

$$\int_{19.800}^{+\infty} p_{10}(\chi^2) d\chi^2 = 0.033 = 3.3\%$$

where  $p_{10}(\chi^2)$  is the probability density of a  $\chi^2$  random variable with 10 d.o.f. The latter value is not tabulated, but it can be easily derived by linear interpolation from the tabulated critical values:

$$\chi^2_{[0.95](10)} = 18.307 \quad \chi^2_{[0.975](10)} = 20.483 .$$

**Example 14.  $\chi^2$ -test for a Poisson distribution**

The commercial description of a radionuclide states that it yields a mean of two decays per minute. To check this statement we count the number of decays yielded during 40 separate time intervals, all one minute long. The results are listed in the table below:

Number of decays per minute	Empirical frequency
0	11
1	12
2	11
3	4
4	2
5 or more	0

We want to verify, with a significance level of 5%, if the number of decays per minute follows:

- (a) a Poisson distribution with mean  $\mu = 2$ ;
- (b) a Poisson distribution with a mean estimated by means of the sample itself.

**Solution**

The empirical frequencies of the observations with at least 3 decays per minute — 4, 2 and 0, in the table — are all too small to allow the direct application of the  $\chi^2$ -test.

It is then convenient to collect the observations with at least 3 decays into a unique bin, to which we will attribute a total empirical frequency of  $4 + 2 + 0 = 6$ .

(a) If the correct distribution is a Poisson distribution of mean  $\mu = 2$ , the probabilities of 0, 1, 2 and at least 3 decays per minute are given by:

$$P(0) = e^{-2} = 0.135335 \quad P(1) = \frac{2}{1!} e^{-2} = 0.270671$$

$$P(2) = \frac{2^2}{2!} e^{-2} = 0.270671$$

$$\begin{aligned} P(\geq 3) &= 1 - e^{-2} \sum_{i=0}^2 \frac{2^i}{i!} = \\ &= 1 - 0.135335 - 0.270671 - 0.270671 = 0.323324 \end{aligned}$$

and therefore the relative expected frequencies  $n_i$  out of 40 trials are those listed in the third column of the following table

Number of decays per minute	Empirical frequency	Expected frequency
0	11	5.413411
1	12	10.826823
2	11	10.826823
3 o più	6	12.932943

The  $\chi^2$  of the data is then

$$\chi^2 = \sum_{i=0}^3 \frac{(f_i - n_i)^2}{n_i} = 9.611732$$

For  $4 - 1 = 3$  d.o.f. and significance level  $\alpha = 0.05$ , the critical  $\chi^2$  turns out to be

$$\chi^2_{[0.95](3)} = 7.815$$

and leads to **reject the hypothesis**.

(b) The mean number of decays per minute is estimated by the sample mean

$$\bar{x} = \frac{0 \cdot 11 + 1 \cdot 12 + 2 \cdot 11 + 3 \cdot 4 + 4 \cdot 2}{40} = \frac{54}{40} = 1.3500$$

The probabilities of 0, 1, 2 and at least 3 decays per minute are thus

$$P(0) = e^{-1.35} = 0.259240$$

$$P(1) = \frac{1.35}{1!} e^{-1.35} = 0.349974$$

$$P(2) = \frac{1.35^2}{2!} e^{-1.35} = 0.236233$$

$$\begin{aligned} P(\geq 3) &= 1 - e^{-1.35} \sum_{i=0}^2 \frac{1.35^i}{i!} = \\ &= 1 - 0.259240 - 0.349974 - 0.236233 = 0.154553 \end{aligned}$$

and the expected frequencies become

Number of decays per minute	Empirical frequency	Expected frequency
0	11	10.369610
1	12	13.998974
2	11	9.449308
3 or more	6	6.182108

The calculation of  $\chi^2$  provides then

$$\chi^2 = \sum_{i=0}^3 \frac{(f_i - n_i)^2}{n_i} = 0.583608$$

On the other hand, the only parameter of the Poisson distribution has been estimated by using the sample, so that the number of d.o.f. is

$$4 - 1 - 1 = 2$$

and for  $\alpha = 0.05$  leads to the critical  $\chi^2$  value

$$\chi^2_{[0.95](2)} = 5.991 .$$

The calculated  $\chi^2$  is significantly smaller than the critical one.

On the basis of the sample we conclude, with a significance level of 5%, that **we can accept the hypothesis that the number of decays per minute of the radionuclide is described by a Poisson distribution of mean  $\mu = 1.35$ :**

$$P(i) = \frac{1.35^i}{i!} e^{-1.35} \quad i = 0, 1, 2, \dots$$

Thus we find, which sounds very reasonable, that an accurate estimate of the parameter  $\mu$  is crucial in determining the result of the test.

**Example 15. Paired  $t$ -test on two means**

A semiconductors company has recently introduced a workplace safety programme. In the table below the week's averages of the man-hours lost due to accident are listed for 10 plants of similar characteristics. The averages are calculated over a month before and a month after the introduction of the new programme.

Plant	Before	After	Difference
1	30.5	23.0	-7.5
2	18.5	21.0	+2.5
3	24.5	22.0	-2.5
4	32.0	28.5	-3.5
5	16.0	14.5	-1.5
6	15.0	15.5	+0.5
7	23.5	24.5	+1.0
8	25.5	21.0	-4.5
9	28.0	23.5	-4.5
10	18.0	16.5	-1.5

We want to determine, with a significance level of 5%, whether the workplace safety programme has been effective.

**Solution**

Denoted with  $\mu_p$  and  $\mu_d$  the week's averages of the man-hours lost due to accidents, we must test the null hypothesis

$$H_0 : \mu_d - \mu_p = 0$$

against the alternative hypothesis

$$H_1 : \mu_d - \mu_p < 0;$$

this allows us to establish if the data may support the idea that the man-hours lost owing to accidents have really diminished because of the workplace safety programme.

Since **at each** plant the week's averages of the man-hours lost have been estimated **before** and **after** the adoption of the new workplace safety programme, it seems natural to carry out the comparison of the averages by a **paired  $t$ -test**. Denoted with

$$(y_1, z_1), \quad (y_2, z_2), \quad \dots, \quad (y_n, z_n)$$

the different pairs of data (after,before), the test statistics is given by the ratio

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

where  $d_i = y_i - z_i$  for  $i = 1, \dots, n$  and  $n = 10$ . If  $H_0$  holds true, the latter is a Student random variable with  $n - 1 = 9$  d.o.f. The test is one-sided and the critical (rejection) region of  $H_0$  consists in the interval

$$\{t \leq t_{[\alpha](n-1)}\} = \{t \leq -t_{[1-\alpha](n-1)}\}$$

which for  $n = 10$  and  $\alpha = 0.05$  becomes

$$\{t \leq -t_{[0.95](9)}\} = \{t \leq -1.833\}.$$

In this case, the mean of the differences turns out to be

$$\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i = -2.1500$$

whereas the sample estimate of the variance is

$$s_d^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - z_i - \bar{y} + \bar{z})^2 = 9.002778$$

so that the statistic  $t$  of the test takes the value

$$t = \frac{-2.1500}{\sqrt{9.002778}/\sqrt{10}} = -2.265949 .$$

Since, however,

$$t = -2.265949 < -1.833 = -t_{[0.95]}(9)$$

we deduce that the value of the statistics falls within the critical region of  $H_0$ .

We conclude therefore, at a significance level of 5%, that **the alternative hypothesis must be rejected** and that the workplace safety programme must be regarded as effective.

**Example 16.  $\chi^2$ -test on the variance of a normal population**

A new device has been installed that must control the quantity of ribbon wrapped on a bobbin. The device can be considered efficient if the standard deviation of the quantity of ribbon wrapped does not exceed 0.15 cm. If a sample of 20 pieces provides a sample variance of  $s^2 = 0.034 \text{ cm}^2$  can we conclude that the device is not efficient?

**Solution**

Let us denote with  $\sigma^2$  the variance of the normal population and with  $\sigma_0^2 = 0.15^2 = 0.0225$  the value of the variance that we want to test. As a null hypothesis we assume

$$H_0 : \sigma^2 = \sigma_0^2$$

against the alternative hypothesis

$$H_1 : \sigma^2 > \sigma_0^2$$

which thus leads to a two-sided test. The rejection region of  $H_0$  is a unique interval of the form

$$\{\chi^2 \geq \chi^2_{[1-\alpha](n-1)}\}$$

in term of the test statistics

$$\chi^2 = \frac{(n-1)s^2}{\sigma_0^2},$$

which, if the null hypothesis holds true, follows a  $\chi^2$  distribution with  $n - 1$  d.o.f.

In the present case the sample contains  $n = 20$  data and the significance level  $\alpha$  is not specified.

We can then reckon the value taken by the test statistics

$$\chi^2 = \frac{(20 - 1) \cdot 0.034}{0.15^2} = 28.7111$$

and determine the probability that, for  $H_0$  true, the statistics has a value equal to or larger than that observed:

$$\text{P-value} = \int_{28.7111}^{+\infty} p_{19}(\chi^2) d\chi^2$$

where  $p_{19}(\chi^2)$  denotes the probability density of a  $\chi^2$  random variable with 19 d.o.f. Such a value is not tabulated, but it can be easily calculated by applying a linear interpolation scheme to the tabulated critical values just smaller and just larger than 28.7111:

$$\chi^2_{[0.90](19)} = 27.204 \quad \chi^2_{[0.95](19)} = 30.144$$

We have therefore

$$\frac{28.7111 - 27.204}{30.144 - 27.204} = \frac{(1 - \text{P-value}) - 0.90}{0.95 - 0.90}$$

and whence

$$\text{P-value} = 0.074369$$

As a conclusion, **if the significance level of the test is smaller than 7.437% the hypothesis  $H_0$  must be accepted**, and we conclude that the device can be regarded as efficient. In contrast, if  $\alpha \geq 7.437\%$ , the calculated  $\chi^2$  falls within the critical region and  $H_0$  must be rejected.

**Example 17.  $z$ -test on the mean of a normal population**

An electrochemical company has developed a battery for pocket calculators which is believed to last significantly longer than the normal batteries. It is known that the duration of a common battery for pocket calculator obeys a normal distribution with mean 1003 h and standard deviation 62.5 h. Preliminary investigations suggest that also the duration of the new batteries follows a normal distribution with the same standard deviation. A sample of 15 (independent) durations of the new devices is considered and a sample mean of  $\bar{x} = 1056$  h is calculated. Test, with a significance level of 1%, whether the new batteries offer **really different** duration performances.

**Solution**

Let us denote with  $\mu_0 = 1003$  h the mean duration of a common battery and with  $\mu$  that of a battery of new kind. Since we are interested in checking only if the new type of battery has a duration **different** from that of a common battery, we will have to test the null hypothesis

$$H_0 : \mu = \mu_0$$

against the **two-sided** alternative hypothesis

$$H_1 : \mu \neq \mu_0$$

with a significance level of  $\alpha = 0.01$ , by using a sample of  $n = 15$  data.

The population of the new batteries is normal and with a known standard deviation. It is then convenient to use as a test variable the statistic

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

which for  $H_0$  true follows a standard normal distribution. For the present sample the statistic takes the value

$$z = \frac{1056 - 1003}{62.5/\sqrt{15}} = 3.284290$$

and the probability of getting a value of  $z$  farther than this from zero (the so-called P-value) can be calculated in a straightforward way

$$\begin{aligned} \text{P-value} &= \int_{-\infty}^{-3.284290} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz + \int_{+3.284290}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \\ &= 1 - 2 \int_0^{3.284290} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 1 - 2 \cdot 0.49948 = 0.001040 \end{aligned}$$

by using the tabulated value:

$$\int_0^{3.284290} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 0.49948 .$$

As the P-value is smaller than the significance level of the test:

$$\text{P-value} = 0.001040 < 0.01 = \alpha ,$$

we conclude that **the value of  $z$  belongs to the critical region** — more precisely, to the upper tail of the two-sided critical region — and that therefore the null hypothesis  $H_0$  must be rejected. At the significance level of 1%, **it is justified to claim that the mean duration of the new batteries is different from that of the normal ones.**

**Example 18. One-sided  $z$ -test on the mean**

In the previous example, the analysis on the batteries of new type is repeated by increasing the number of data to  $n = 20$  and a sample mean  $\bar{x} = 1050$  h is found. By assuming as before that the duration measurements of the new batteries have a normal distribution with  $\sigma = 62.5$  h, we want to test with a significance level of 5% whether the duration of the new batteries is **larger** than that of the standard batteries.

**Solution**

Now we want to check, with a certain significance level, if the new batteries have really a mean duration larger than that of the normal batteries, so that we shall test the same null hypothesis as before

$$H_0 : \mu = \mu_0 = 1003 \text{ h}$$

against the new **one-sided** alternative hypothesis

$$H_1 : \mu > \mu_0$$

with a significance level of  $\alpha = 0.05$ , by using a sample of  $n = 20$  data.

The population is however normal, as also is however known the relative standard deviation  $\sigma = 62.5$  h; thus the statistic of the test we must consider does not change

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

and is always a standard normal for  $H_0$  true. By inserting the new values of  $n$  and  $\bar{x}$  we obtain, however,

$$z = \frac{1050 - 1003}{62.5/\sqrt{20}} = 3.363046$$

and the corresponding P-value now turns out to be

$$\begin{aligned} \text{P-value} &= \int_{3.363046}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = \\ &= \frac{1}{2} - \int_0^{3.363046} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz = 0.5 - 0.49960 = 0.00040 \end{aligned}$$

Since

$$\text{P-value} = 0.00040 < 0.05 = \alpha,$$

we conclude that  $H_0$  must be rejected in favor of  $H_1$ . **The mean duration of the new batteries seems significantly larger than that of the common batteries.**

The same result can be achieved, of course, by computing the one-sided critical region of  $H_0$ :

$$\{z \geq z_\alpha\}$$

with the critical value  $z_\alpha$  determined by the equation

$$\alpha = \int_{z_\alpha}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz \iff \frac{1}{2} - \alpha = \int_0^{z_\alpha} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$

It is enough to search in the table the critical value corresponding to  $0.5 - \alpha = 0.45$  to obtain

$$z_{0.05} = 1.655$$

and, being  $z = 3.363046 > 1.655 = z_{0.05}$ , conclude that  $H_0$  must be rejected.

**Example 19. Linear correlation coefficient**

Owing to the poor epidermic transpiration, birds usually remove the excess of heat of their body by panting. In order to check if there is a relationship between the body temperature and the respiratory frequency, we consider a random sample of 15 ravens, in different environmental conditions, and for each individual we measure the body temperature  $T$  (in  $^{\circ}C$ ) and the respiratory frequency  $f$  (in breathes per minute). The sample is assumed to be normal. The experimental results are summarized in the table below.

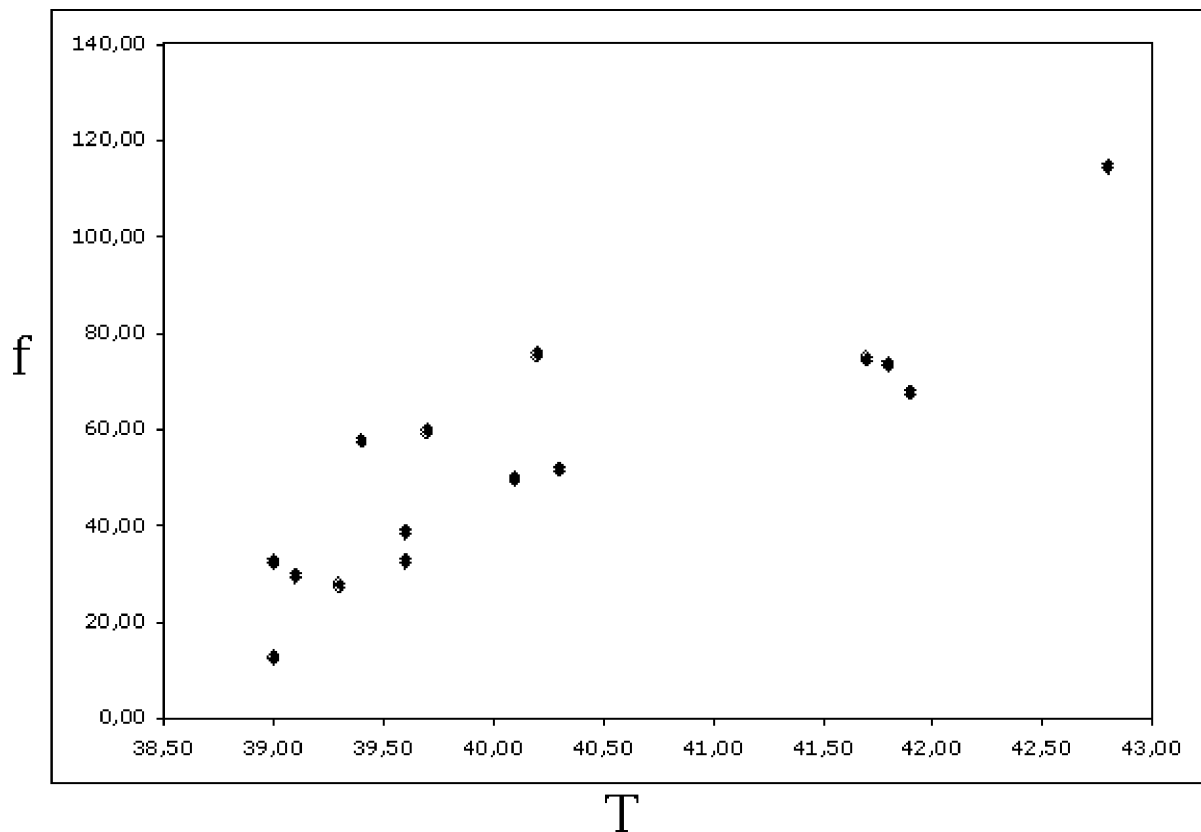
individual $i$	$T_i$	$f_i$
1	39.6	33
2	40.1	50
3	41.7	75
4	39.0	13
5	41.9	68
6	42.8	115
7	40.3	52
8	39.0	33
9	39.7	60
10	39.3	28
11	41.8	74
12	39.6	39
13	40.2	76
14	39.1	30
15	39.4	58

Test the hypothesis that the variables  $T$  and  $f$  are independent against the alternative hypothesis that a linear correlations ex-

ists, with a significance level (a) of 5% and (b) of 1%.

## Solution

Since the raven sample is completely random, both temperature  $T$  and respiratory frequency  $f$  are random variables and the problem of their stochastic independence can be tackled by analysing Pearson's linear correlation coefficient  $r$ . Notice that the variables are assumed to be normal, so that stochastic independence is equivalent to lack of correlation. The graph of the data suggests a linear relationship between  $T$  and  $f$ :



To put the analysis on a quantitative ground, we proceed to the calculation of  $r$ . We compute the sample means of the two

quantities:

$$\bar{T} = \frac{1}{15} \sum_{i=1}^{15} T_i = 40.2333 \quad \bar{f} = \frac{1}{15} \sum_{i=1}^{15} f_i = 53.6000$$

and derive then the sums of residual products/squares

$$SS_{Tf} = \sum_{i=1}^{15} (T_i - \bar{T})(f_i - \bar{f}) = 385.9000$$

$$SS_{TT} = \sum_{i=1}^{15} (T_i - \bar{T})^2 = 20.9733$$

$$SS_{ff} = \sum_{i=1}^{15} (f_i - \bar{f})^2 = 9351.6000$$

so that the linear correlation coefficient becomes

$$r = \frac{SS_{Tf}}{\sqrt{SS_{TT}}\sqrt{SS_{ff}}} = \frac{385.9000}{\sqrt{20.9733}\sqrt{9351.6000}} = 0.8714$$

As  $T$  and  $f$  are normal, we can test the null hypothesis

$$H_0 : T \text{ and } f \text{ are stochastically independent}$$

against the alternative hypothesis

$$H_1 : T \text{ and } f \text{ are stochastically dependent}$$

by using the random variable

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

which, for true  $H_0$ , follows a Student distribution with  $n-2$  d.o.f. In the present case we have

$$t = \sqrt{15-2} \frac{0.8714}{\sqrt{1-0.8714^2}} = 6.4033$$

The rejection region takes then the form

$$|t| > t_{[1-\frac{\alpha}{2}](n-2)} = t_{[0.975](13)} = 2.160$$

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

$$|t| > t_{[1-\frac{\alpha}{2}](n-2)} = t_{[0.995](13)} = 3.012$$

whenever the requested significance level is  $\alpha = 1\%$ .

**In both cases  $H_0$  must be rejected!**

Therefore we conclude that a stochastic dependence between the body temperature and the respiratory frequency of birds is plausible.

In the hypothesis of normal variables, such a dependence is equivalent to a nonvanishing correlation between the same random variables.

**Example 20.  $F$ -test to check the goodness of a linear regression fit with repeated measurements**

At some fixed values  $x_i$  of a physical quantity  $x$  we have carried out a number of repeated measurements of another physical quantity  $y$ . The data are summarized in the following tables:

$x_i$	$y_{i1}$	$y_{i2}$	$y_{i3}$	$y_{i4}$	$y_{i5}$
1.0	2.3	2.5	3.2	4.2	4.4
2.0	6.1	7.0	7.7	8.4	6.5
3.0	10.0	11.5	12.8	12.0	12.9
4.0	14.2	15.3	16.8	14.9	15.5
5.0	20.0	21.0	19.3	19.9	20.5
6.0	21.8	24.8	24.6	24.9	23.1
7.0	28.3	27.1	27.8	29.1	30.2
8.0	32.0	28.0	32.7	30.3	31.0

$x_i$	$y_{i6}$	$y_{i7}$	$y_{i8}$	$y_{i9}$	$y_{i10}$
1.0	4.9	4.1	5.0	5.5	6.0
2.0	8.5	7.5	8.8	7.4	9.9
3.0	11.1	11.3	12.5	13.2	14.0
4.0	16.8	17.0	15.0	16.1	17.9
5.0	20.9	19.1	22.5	19.2	23.8
6.0	23.5	23.2	24.1	23.0	26.1
7.0	27.5	28.7	28.1	=	=
8.0	31.5	34.0	=	=	=

Assuming that the system is homoskedastic and that the data  $y$  are normally distributed, determine the least-squares straight line of the sample and decide whether the regression model is acceptable at a significance level of 5%.

## Solution

In the standard case of  $\mathcal{X}^2$  linear fitting to a straight line with non-repeated measurements, for a regression model of the form

$$y = \mu + \kappa(x - \bar{x})$$

the merit function would be

$$\mathcal{L}(\mu, \kappa) = \sum_{i=1}^n \frac{1}{\sigma_i^2} \left[ \mu + \kappa(x_i - \bar{x}) - y_i \right]^2$$

with

$$\bar{x} = \sum_{i=1}^n \frac{1}{\sigma_i^2} x_i \cdot \left( \sum_{i=1}^n \frac{1}{\sigma_i^2} \right)^{-1}$$

and the corresponding normal equations would provide the best-fit estimates

$$m = \frac{\sum_{i=1}^n \frac{1}{\sigma_i^2} y_i}{\sum_{i=1}^n \frac{1}{\sigma_i^2}} \quad q = \frac{\sum_{i=1}^n \frac{1}{\sigma_i^2} (x_i - \bar{x}) y_i}{\sum_{i=1}^n \frac{1}{\sigma_i^2} (x_i - \bar{x})^2}$$

for the model parameters  $\mu$  and  $\kappa$ , respectively.

In the present case of repeated measurements of  $y$  for each sampled value of  $x$ , the values of the quantity  $y$  collected at a given  $x_i$  are denoted with

$$y_{ik} , \quad k = 1, 2, \dots, n_i$$

for  $i = 1, \dots, n$ . The whole number of  $y$  data is thus

$$N = \sum_{i=1}^n n_i .$$

The merit function for the  $\mathcal{X}^2$  fitting is now

$$\mathcal{L}^*(\mu, \kappa) = \sum_{i=1}^n \frac{1}{\sigma_i^2/n_i} \left[ \mu + \kappa(x_i - \bar{x}) - \bar{y}_i \right]^2$$

where  $\bar{y}_i$  stands for the arithmetic mean of the repeated measurements of  $y$  at  $x = x_i$ :

$$\bar{y}_i = \frac{1}{n_i} \sum_{k=1}^{n_i} y_{ik},$$

whereas  $\bar{x}$  is given by:

$$\bar{x} = \sum_{i=1}^n \frac{n_i}{\sigma_i^2} x_i \cdot \left( \sum_{i=1}^n \frac{n_i}{\sigma_i^2} \right)^{-1}.$$

By comparing  $\mathcal{L}(\mu, \kappa)$  and  $\mathcal{L}^*(\mu, \kappa)$  it is apparent that the new normal equations and the corresponding solutions can be obtained from the previous ones by means of the substitutions:

$$\begin{aligned} y_i &\longrightarrow \bar{y}_i \\ \sigma_i^2 &\longrightarrow \sigma_i^2/n_i \end{aligned}$$

for all  $i = 1, \dots, n$ . We have therefore the new relationships:

$$m = \frac{\sum_{i=1}^n \frac{n_i}{\sigma_i^2} \bar{y}_i}{\sum_{i=1}^n \frac{n_i}{\sigma_i^2}} \qquad q = \frac{\sum_{i=1}^n \frac{n_i}{\sigma_i^2} (x_i - \bar{x}) \bar{y}_i}{\sum_{i=1}^n \frac{n_i}{\sigma_i^2} (x_i - \bar{x})^2}$$

which in the homoskedastic case reduce to

$$m = \frac{\sum_{i=1}^n n_i \bar{y}_i}{\sum_{i=1}^n n_i} = \frac{1}{N} \sum_{i=1}^n n_i \bar{y}_i \quad q = \frac{\sum_{i=1}^n n_i (x_i - \bar{x}) \bar{y}_i}{\sum_{i=1}^n n_i (x_i - \bar{x})^2}$$

while  $\bar{x}$  becomes the simple arithmetic mean of all the  $x$  data with their multiplicities

$$\bar{x} = \sum_{i=1}^n n_i x_i \cdot \left( \sum_{i=1}^n n_i \right)^{-1} = \frac{1}{N} \sum_{i=1}^n n_i x_i.$$

The sample considered here consists of  $n = 8$  values of  $x$  and

$$N = \sum_{i=1}^8 n_i = 10 + 10 + 10 + 10 + 10 + 10 + 8 + 7 = 75$$

values of  $y$ , as easily checked on the experimental tables. The means of the repeated measurements at a given  $x_i$  can be calculated immediately, and the results are listed in the table below

$x_i$	$n_i$	$\bar{y}_i$
1.0	10	4.2100
2.0	10	7.7800
3.0	10	12.1300
4.0	10	15.9500
5.0	10	20.6200
6.0	10	23.9100
7.0	8	28.3500
8.0	7	31.3571

The mean of  $x_i$ 's is simply

$$\bar{x} = \frac{1}{75} \sum_{i=1}^n n_i x_i = \frac{1}{75} \cdot 322 = 4.2933.$$

As for the least-squares estimate of  $\mu$ , we get

$$m = \frac{1}{N} \sum_{i=1}^n n_i \bar{y}_i = \frac{1}{75} \cdot 1292.3000 = 17.2307$$

while the estimate of  $\kappa$  becomes

$$q = \frac{\sum_{i=1}^n n_i (x_i - \bar{x}) \bar{y}_i}{\sum_{i=1}^n n_i (x_i - \bar{x})^2} = \frac{1460.525333}{367.5466667} = 3.9737$$

due to the partial terms listed in the table below

$n_i$	$x_i - \bar{x}$	$\bar{y}_i$	$n_i(x_i - \bar{x})^2$
10	-3.2933	4.2100	108.460444
10	-2.2933	7.7800	52.593778
10	-1.2933	12.1300	16.727111
10	-0.2933	15.9500	0.860444
10	0.7067	20.6200	4.993778
10	1.7067	23.9100	29.127111
8	2.7067	28.3500	58.608356
7	3.7067	31.3571	96.175644

As a consequence, the least-squares straight line takes the form

$$y = m + q(x - \bar{x})$$

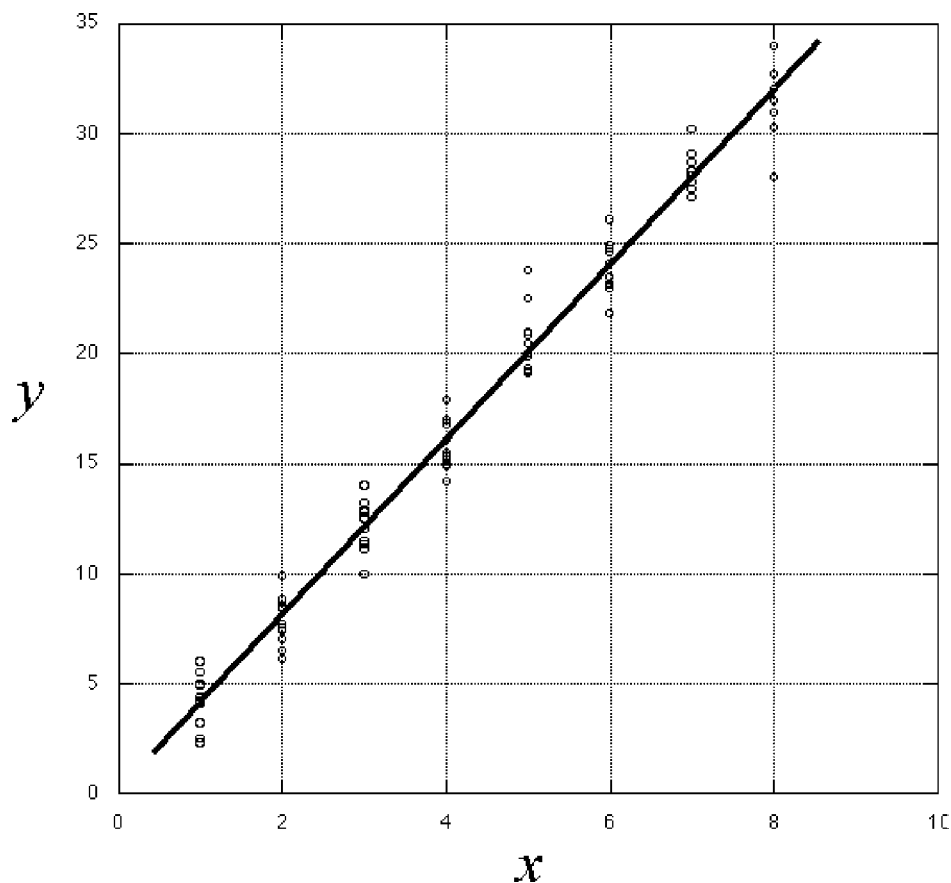
i.e.

$$y = 17.2307 + 3.9737(x - 4.2933)$$

or, finally,

$$y = 0.17041 + 3.9737x .$$

In the following graph the least-squares straight line is superimposed to the experimental datapoints:



showing, at a first glance, a satisfactory agreement.

To check the quality of the linear regression model we can use the  $F$  test with repeated measurements in the homoskedastic version.

We need to compare the **model dependent** sum of squares (or sum of squares around regression)

$$SS_{\text{m.d.}} = \sum_{i=1}^n n_i [m + q(x_i - \bar{x}) - \bar{y}_i]^2$$

with the **model independent** sum of squares

$$SS_{\text{m.i.}} = \sum_{i=1}^n \sum_{k=1}^{n_i} (y_{ik} - \bar{y}_i)^2.$$

The first is easily calculated by using the data listed below

$n_i$	$m + q(x_i - \bar{x})$	$\bar{y}_i$	$n_i [m + q(x_i - \bar{x}) - \bar{y}_i]^2$
10	4.143901908	4.2100	0.043689577
10	8.117615904	7.7800	1.139844984
10	12.0913299	12.1300	0.014953767
10	16.06504389	15.9500	0.132350977
10	20.03875789	20.6200	3.378423903
10	24.01247189	23.9100	0.105004873
8	27.98618588	28.3500	1.058885705
7	31.95989988	31.3571	2.543212172

The model independent sum of squares is derived from the following partial squares

$(y_{i1} - \bar{y}_i)^2$	$(y_{i2} - \bar{y}_i)^2$	$(y_{i3} - \bar{y}_i)^2$	$(y_{i4} - \bar{y}_i)^2$	$(y_{i5} - \bar{y}_i)^2$
3.64810	2.92410	1.02010	0.00010	0.03610
2.82240	0.60840	0.00640	0.38440	1.63840
4.53690	0.39690	0.44890	0.01690	0.59290
3.06250	0.42250	0.72250	1.10250	0.20250
0.38440	0.14440	1.74240	0.51840	0.01440
4.45210	0.79210	0.47610	0.98010	0.65610
0.00250	1.56250	0.30250	0.56250	3.42250
0.41327	11.27041	1.80327	1.11755	0.12755

$(y_{i6} - \bar{y}_i)^2$	$(y_{i7} - \bar{y}_i)^2$	$(y_{i8} - \bar{y}_i)^2$	$(y_{i9} - \bar{y}_i)^2$	$(y_{i10} - \bar{y}_i)^2$
0.47610	0.01210	0.62410	1.66410	3.20410
0.51840	0.07840	1.04040	0.14440	4.49440
1.06090	0.68890	0.13690	1.14490	3.49690
0.72250	1.10250	0.90250	0.02250	3.80250
0.07840	2.31040	3.53440	2.01640	10.11240
0.16810	0.50410	0.03610	0.82810	4.79610
0.72250	0.12250	0.06250	=	=
0.02041	6.98469	=	=	=

The results are

$$SS_{m.d.} = 8.4163659 \quad SS_{m.i.} = 112.97314$$

and provide the value of the test statistics

$$F = \frac{N - n}{n - p} \frac{SS_{m.d.}}{SS_{m.i.}} = \frac{75 - 8}{8 - 2} \cdot \frac{8.4163659}{112.97314} = 0.8319035$$

If the regression model is correct the test statistics obeys a Fisher distribution with  $(n - p, N - n) = (6, 67)$  d.o.f.

The critical region at a significance level  $\alpha$  is then of the form

$$\{F > F_{[1-\alpha],(n-p,N-n)}\}$$

and in the present case of  $\alpha = 0.05$  becomes

$$\{F > F_{[0.95],(6,67)}\}.$$

From the table of the Fisher cumulative distribution — the table corresponding to an upper tail probability  $\alpha = 0.05$ , at the intersection between the column  $\nu_1 = 6$  and the row  $\nu_2 = 67$  — we obtain the critical value

$$F_{[0.95],(6,67)} = 2.237$$

and since

$$F = 0.8319035 < 2.237$$

we must conclude that the regression model **cannot be rejected** at the significance level of 5%.

Thus the result of the test confirms the rough, first impression suggested by the graphical representation of the sample data and of the regression straight line.

Once the regression model has been statistically validated, we can complete it by computing the **confidence intervals** for the regression parameters  $\mu$  and  $\kappa$ .

As in the case of single measurements, the definition of appropriate CIs for the parameters is made possible by the circumstance that if the model is correct the random variable

$$\text{NSS}_{\text{m.d.}} = \sum_{i=1}^n \frac{1}{\sigma_i^2/n_i} [m + q(x_i - \bar{x}) - \bar{y}_i]^2$$

follows a  $\mathcal{X}^2$  distribution with  $n - p$  d.o.f. and is stochastically independent on the independent normal random variables  $m$  and  $q$ , whose variances are given by

$$\sigma_m^2 = \left( \sum_{i=1}^n \frac{1}{\sigma_i^2/n_i} \right)^{-1} \quad \sigma_q^2 = \left( \sum_{i=1}^n \frac{1}{\sigma_i^2/n_i} (x_i - \bar{x})^2 \right)^{-1}$$

as derived from the corresponding formulas for the single measurement regression by replacing  $\sigma_i^2$  with  $\sigma_i^2/n_i$ .

The CI interval of  $\mu$ , at the confidence level  $1 - \xi$ , is thus

$$m \pm t_{[1-\frac{\xi}{2}](n-2)} \sqrt{\left( \sum_{i=1}^n \frac{1}{\sigma_i^2/n_i} \right)^{-1} \frac{\text{NSS}_{\text{m.d.}}}{n-2}}$$

whereas for the CI of  $\kappa$  there holds

$$q \pm t_{[1-\frac{\xi}{2}](n-2)} \sqrt{\left( \sum_{i=1}^n \frac{1}{\sigma_i^2/n_i} (x_i - \bar{x})^2 \right)^{-1} \frac{\text{NSS}_{\text{m.d.}}}{n-2}}.$$

As usual, in the homoskedastic case the common variance  $\sigma^2$  disappears in all formulas

$$m \pm t_{[1-\frac{\xi}{2}](n-2)} \sqrt{\frac{1}{N} \frac{SS_{m.d.}}{n-2}}$$

$$q \pm t_{[1-\frac{\xi}{2}](n-2)} \sqrt{\left(\sum_{i=1}^n n_i (x_i - \bar{x})^2\right)^{-1} \frac{SS_{m.d.}}{n-2}}$$

with the simple sum of squares around regression

$$SS_{m.d.} = \sum_{i=1}^n n_i [m + q(x_i - \bar{x}) - \bar{y}_i]^2$$

instead of the normalized one.

In the present case the CI of  $\mu$  becomes

$$17.2307 \pm t_{[1-\frac{\xi}{2}](6)} \sqrt{\frac{1}{75} \frac{8.4163659}{6}}$$

and that of  $\kappa$  holds

$$3.9737 \pm t_{[1-\frac{\xi}{2}](6)} \sqrt{(367.5466667)^{-1} \frac{8.4163659}{6}}.$$

Performing the calculations, we finally obtain:

$$\mu \in 17.2307 \pm 0.13676 \cdot t_{[1-\frac{\xi}{2}](6)}$$

and

$$\kappa \in 3.9737 \pm 0.061778 \cdot t_{[1-\frac{\xi}{2}](6)}$$

In both intervals the confidence level  $1 - \xi$  is left indetermined and can be chosen at will (e.g., 67%, 90% or 95%).

**Example 21. Incremental  $F$ -test on a linear regression**

At some given values of a quantity  $x$  we have measured the corresponding value of another quantity  $y$  which is supposed to depend on  $x$ . The measurements have provided the following sample

$x_i$	$y_i$
0.5	1.49
1.0	1.37
1.5	2.45
2.0	2.16
2.5	3.53
3.0	3.78
3.5	5.04
4.0	5.59
4.5	7.11
5.0	7.83

The data  $y_i$  are normal with a common variance  $\sigma^2$ . Determine, with a significance level of 5%, which of the following regression models

$$y = \mu + \kappa(x - \bar{x})$$

$$y = \mu + \kappa(x - \bar{x}) + \lambda(x - \bar{x}^2)$$

is more satisfactory to describe the sample, on having denoted with  $\bar{x}$  the arithmetic mean of the values  $x_i$ .

**Solution**

We have to compare two polynomial models, of the first and of the second order. The comparison is made by computing the NSSAR of the two models and determining if the decrease

of the sum of squares is large enough to detect a significant difference.

*First order polynomial*

The  $\mathcal{X}^2$  merit function is

$$\mathcal{L}_1(\mu, \kappa) = \sum_{i=1}^n \frac{1}{\sigma^2} [\mu + \kappa(x_i - \bar{x}) - y_i]^2$$

and its minimum is defined by the solution of the normal equations

$$\begin{cases} \sum_{i=1}^n [\mu + \kappa(x_i - \bar{x}) - y_i] = 0 \\ \sum_{i=1}^n [\mu + \kappa(x_i - \bar{x}) - y_i](x_i - \bar{x}) = 0 \end{cases}$$

which can be simplified as

$$\begin{cases} n\mu = \sum_{i=1}^n y_i \\ \kappa \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n y_i(x_i - \bar{x}) \end{cases}$$

and provide the best fit estimates of the parameters

$$\mu = \frac{1}{n} \sum_{i=1}^n y_i = m$$

$$\kappa = \sum_{i=1}^n y_i(x_i - \bar{x}) \left( \sum_{i=1}^n (x_i - \bar{x})^2 \right)^{-1} = q.$$

In the present case we have

$$m = \frac{1}{10}40.3500 = 4.03500$$

$$q = \frac{30.45750000}{20.62500000} = 1.476727273$$

and the regression model becomes

$$y = 4.03500 + 1.476727273 (x - 2.75)$$

or, equivalently,

$$y = -0.026000001 + 1.476727273 x .$$

The normalized sum of squares around regression is then

$$\text{NSSAR}_1 = \mathcal{L}(m, q) = \frac{1}{\sigma^2} \sum_{i=1}^n [m + q(x_i - \bar{x}) - y_i]^2 = \frac{2.235429}{\sigma^2}$$

*Second order polynomial*

For this polynomial model the  $\mathcal{X}^2$  objective function to minimize is given by

$$\mathcal{L}_2(\mu, \kappa, \lambda) = \sum_{i=1}^n \frac{1}{\sigma^2} [\mu + \kappa(x_i - \bar{x}) + \lambda(x_i - \bar{x})^2 - y_i]^2$$

and leads to the normal equations

$$\left\{ \begin{array}{l} \sum_{i=1}^n [\mu + \kappa(x_i - \bar{x}) + \lambda(x - \bar{x})^2 - y_i] = 0 \\ \sum_{i=1}^n [\mu + \kappa(x_i - \bar{x}) + \lambda(x - \bar{x})^2 - y_i] (x - \bar{x}) = 0 \\ \sum_{i=1}^n [\mu + \kappa(x_i - \bar{x}) + \lambda(x - \bar{x})^2 - y_i] (x - \bar{x})^2 = 0 \end{array} \right.$$

which can be partially simplified and rewritten into the equivalent form

$$\begin{cases} n\mu + \lambda \sum_{i=1}^n (x - \bar{x})^2 = \sum_{i=1}^n y_i \\ \kappa \sum_{i=1}^n (x_i - \bar{x})^2 + \lambda \sum_{i=1}^n (x - \bar{x})^3 = \sum_{i=1}^n y_i (x - \bar{x}) \\ \mu \sum_{i=1}^n (x - \bar{x})^2 + \kappa \sum_{i=1}^n (x_i - \bar{x})^3 + \lambda \sum_{i=1}^n (x - \bar{x})^4 = \sum_{i=1}^n y_i (x - \bar{x})^2 \end{cases}$$

and determine the unique minimum of  $\mathcal{L}_2$  — the best fit estimates  $m, q, l$  of the model parameters  $\mu, \kappa, \lambda$ , respectively.

In the present case the matrix representation of the normal equations is

$$\begin{pmatrix} 10. & 0. & 20.6250 \\ 0. & 20.6250 & 0. \\ 20.6250 & 0. & 75.5390625 \end{pmatrix} \begin{pmatrix} m \\ q \\ l \end{pmatrix} = \begin{pmatrix} 40.3500 \\ 30.457500 \\ 90.221875 \end{pmatrix}$$

where the inverse of the representative matrix holds

$$\begin{pmatrix} 0.22890625 & 0. & -0.0625 \\ 0. & 0.04848484849 & 0. \\ -0.0625 & 0. & 0.03030303031 \end{pmatrix} \quad (1)$$

and provides the solution

$$\begin{pmatrix} m \\ q \\ l \end{pmatrix} = \begin{pmatrix} 3.597500000 \\ 1.476727273 \\ 0.212121213 \end{pmatrix}. \quad (2)$$

Notice that, according to the standard theory of linear regression, (1) constitutes the covariance matrix of the normal RVs  $m, q, l$ , which therefore are not stochastically independent. From (2) we finally deduce the requested regression model

$$y = 3.59750 + 1.476727273(x - 2.75) + 0.212121213(x - 2.75)^2$$

that after some algebraic manipulation becomes

$$y = 1.140666672 + 0.3100606015x + 0.2121212130x^2.$$

The normalized sum of squares around regression holds then

$$\begin{aligned} \text{NSSAR}_2 &= \frac{1}{\sigma^2} \sum_{i=1}^n [m + q(x_i - \bar{x}) + l(x_i - \bar{x})^2 - y_i]^2 = \\ &= \frac{1}{\sigma^2} 0.750580606. \end{aligned}$$

### *Choice of the regression model*

We test the null hypothesis

$$H_0 : y = \mu + \kappa(x - \bar{x}) \text{ is correct}$$

against the alternative hypothesis

$$H_1 : y = \mu + \kappa(x - \bar{x}) + \lambda(x - \bar{x})^2 \text{ holds true}$$

by using the test variable

$$F = \frac{\text{NSSAR}_1 - \text{NSSAR}_2}{\text{NSSAR}_2 / (10 - 2 - 1)} \quad (3)$$

which for a homoskedastic system is actually independent on  $\sigma^2$ , the latter thus being unnecessary to perform the test.

If  $H_0$  holds true,  $F$  is known to be a Fisher random variable with  $(1, n - p - 1) = (1, 10 - 2 - 1) = (1, 7)$  d.o.f.

The null hypothesis is rejected if  $F$  turns out to be sufficiently large, because in this case the decrease of  $\text{NSSAR}_2$  with respect to  $\text{NSSAR}_1$  seems to justify the introduction of the further regression term  $\lambda(x - \bar{x})^2$ .

The rejection region takes then the form

$$\{F > F_{[1-\alpha]}(1, n-p-1)\}$$

with

$$F_{[1-\alpha]}(1, n-p-1) = F_{[0.95]}(1, 7) = 5.591.$$

For the given sample the test variable holds

$$F = \frac{2.235429091 - 0.750580606}{0.750580606/7} = 13.84786565$$

and since  $F > F_{[0.95]}(1, 7)$  we conclude that  $H_0$  must be rejected at the significance level of 5%.

**The second order polynomial is statistically more appropriate to describe the sample data.**

The detailed data are listed in the following table.

### Incremental F-test on a linear regression

#### Table of data

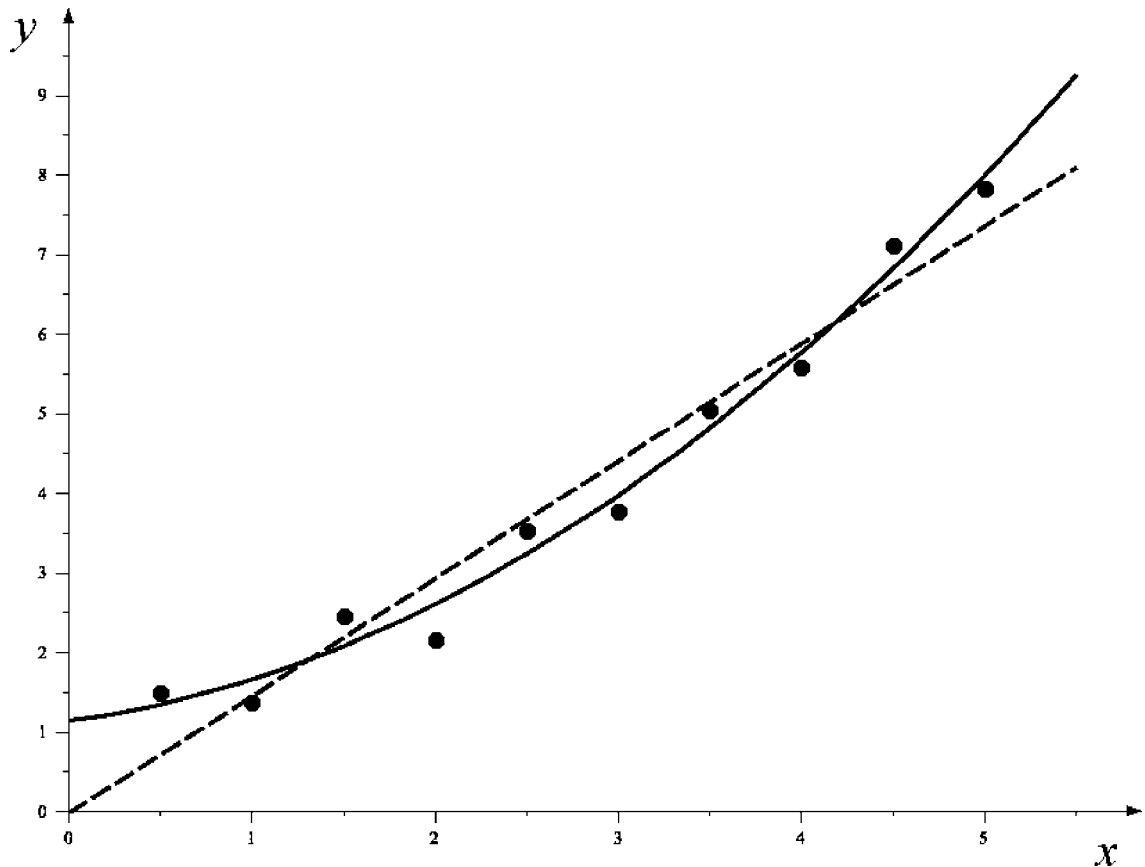
x(i)	y(i)	x(i)-xs	[x(i)-xs]^2	[x(i)-xs]^3	[x(i)-xs]^4	y(i)*[x(i)-xs]	y(i)*[x(i)-xs]^2
0,5	1,49	-2,2500	5,06250000	-11,39062500	25,62890625	-3,35250000	7,54312500
1,0	1,37	-1,7500	3,06250000	-5,35937500	9,37890625	-2,39750000	4,19562500
1,5	2,45	-1,2500	1,56250000	-1,95312500	2,44140625	-3,06250000	3,82812500
2,0	2,16	-0,7500	0,56250000	-0,42187500	0,31640625	-1,62000000	1,21500000
2,5	3,53	-0,2500	0,06250000	-0,01562500	0,00390625	-0,88250000	0,22062500
3,0	3,78	0,2500	0,06250000	0,01562500	0,00390625	0,94500000	0,23625000
3,5	5,04	0,7500	0,56250000	0,42187500	0,31640625	3,78000000	2,83500000
4,0	5,59	1,2500	1,56250000	1,95312500	2,44140625	6,98750000	8,73437500
4,5	7,11	1,7500	3,06250000	5,35937500	9,37890625	12,44250000	21,77437500
5,0	7,83	2,2500	5,06250000	11,39062500	25,62890625	17,61750000	39,63937500
Sums:	40,3500	0,0000	20,62500000	0,00000000	75,53906250	30,45750000	90,22187500

$$xs = 2,7500$$

2nd order polyn. prediction	2nd order polyn. squares	1st order polyn. prediction	1st order polyn. squares
1,348727276	0,019957983	0,712363636	0,604718315
1,662848487	0,085760236	1,450727272	0,006516892
2,083030304	0,134666758	2,189090909	0,068073554
2,609272727	0,201845983	2,927454545	0,588986479
3,241575757	0,083188544	3,665818182	0,018446578
3,979939394	0,039975761	4,404181818	0,389602942
4,824363637	0,046499041	5,142545455	0,01051557
5,774848486	0,034168963	5,880909091	0,084628099
6,831393942	0,077621336	6,619272728	0,240813256
7,994000005	0,026896001	7,357636364	0,223127405
Sum of squares:	0,750580606	Sum of squares:	2,235429091

F = 13,84786565 If the correct model is the first order polynomial, F is a Fisher RV with  $(1, n-p-1) = (1, 10-2-1) = (1, 7)$  d.o.f.

As a further illustration, in the figure below we plot the sample data (solid circles), the first order polynomial (dashed line) and the second order polynomial (solid line).



It is rather apparent that the parabolic curve (second order polynomial) passes significantly closer to the datapoints than the straight line (first order polynomial).

**Example 22. Logarithmic differential**

A conducting line on an IC chip is  $L = (2.8 \pm 0.1)$  mm long and has a rectangular cross section  $a \times b$ , where  $a = (1.00 \pm 0.01)$   $\mu\text{m}$  and  $b = (4.00 \pm 0.01)$   $\mu\text{m}$ . A current of intensity  $I = (5.120 \pm 0.002)$  mA produces a voltage drop of  $V = (101 \pm 1)$  mV across the line. Determine the conductivity

$$\sigma = \frac{LI}{abV}$$

and the corresponding absolute error. Estimate the precision of the result.

**Solution**

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  $\sigma$  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  $\Delta L$ ,  $\Delta I$ ,  $\Delta a$ ,  $\Delta b$ ,  $\Delta V$ :

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

The greatest absolute error on  $\bar{\sigma}$  is 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  $\sigma$  becomes

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

Since repeated measurements are lacking, the precision of the measurement of  $\sigma$  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.

### Example 23. Chauvenet criterion

By assuming that the temperatures below (in K) follow a normal distribution

786 , 790 , 792 , 796 , 784 , 791 ,  
 802 , 789 , 783 , 785 , 795 , 782 ,

check for the possible presence of an outlier not belonging to the statistical population of the sample.

### Solution

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

x	$\Delta x$	$ABS(\Delta x)$	$\Delta x^2$	
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 $ABS(\Delta x)$ : 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
Linearly interpolated value:				2,0700      0,48077      0,03846
				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  $\text{ABS}(\Delta x)$  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  $x_{\text{SUS}}$  must be rejected as not belonging to the statistical population.*

### Example 24. $\chi^2$ -test for a normal distribution

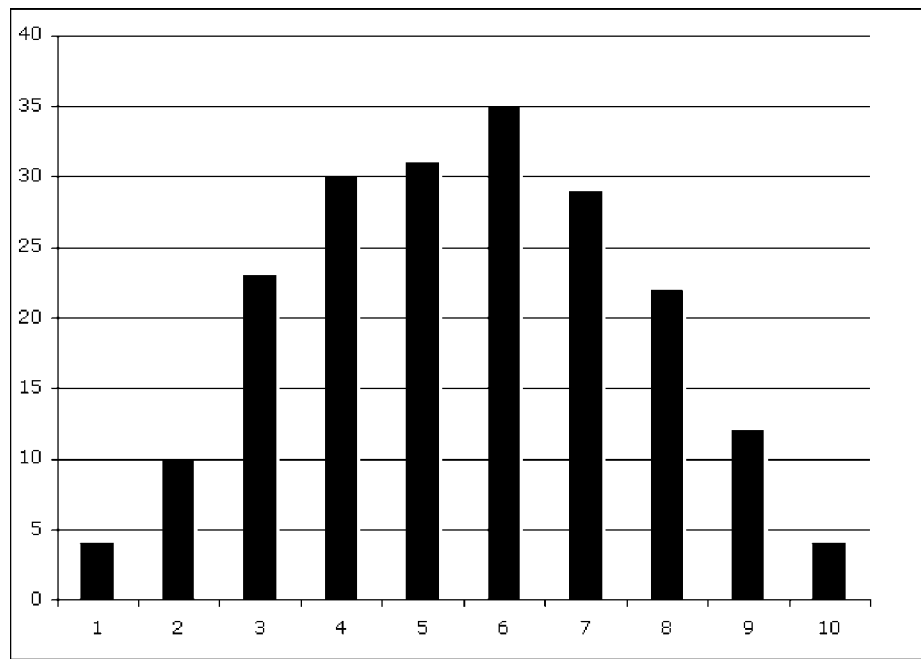
We want to check whether the Young modulus  $E$  of a polymer follows a normal distribution. To this aim we carry out 200 measurements of Young modulus and compute the relative sample mean  $\bar{m}$  and standard deviation  $s$ , which can be assumed as estimates of the mean and standard deviation of the whole population, respectively. The results are binned according to the following frequency table:

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

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

## Solution

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  $\chi^2$  test to check whether the population is normal, i.e. the null hypothesis

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

versus the alternative hypothesis

$H_1$  :  $H_0$  is false .

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

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

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:

$$\begin{aligned}n_1 &= 200 \cdot \int_{-\infty}^{-2} p(z) dz = 200 \cdot \int_2^{+\infty} p(z) dz = \\&= 200 \cdot \left( \frac{1}{2} - \int_0^2 p(z) dz \right) = 200 \cdot \left( \frac{1}{2} - \Phi(2) \right) = \\&= 200 \cdot \left( \frac{1}{2} - 0.47725 \right) = 4.550 \\n_2 &= 200 \cdot \int_{-2}^{-1.5} p(z) dz = 200 \cdot \int_{1.5}^2 p(z) dz = \\&= 200 \cdot \left( \Phi(2) - \Phi(1.5) \right) = \\&= 200 \cdot (0.47725 - 0.43319) = 8.812 \\n_3 &= 200 \cdot \int_{-1.5}^{-1} p(z) dz = 200 \cdot \int_1^{1.5} p(z) dz = \\&= 200 \cdot \left( \Phi(1.5) - \Phi(1) \right) = \\&= 200 \cdot (0.43319 - 0.34134) = 18.370 \\n_4 &= 200 \cdot \int_{-1}^{-0.5} p(z) dz = 200 \cdot \int_{0.5}^1 p(z) dz = \\&= 200 \cdot \left( \Phi(1) - \Phi(0.5) \right) = \\&= 200 \cdot (0.34134 - 0.19146) = 29.976\end{aligned}$$

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

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:

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

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

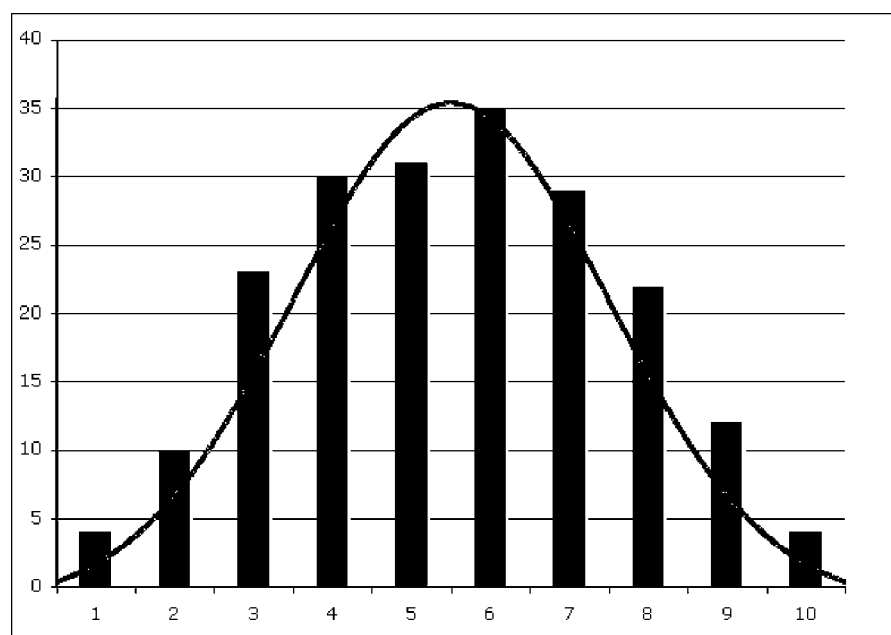
$$\chi^2 = \sum_{i=1}^{10} \frac{(f_i - n_i)^2}{n_i} = 5.0342.$$

The table of the  $\chi^2$  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  $\chi^2$  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  $\mu$  and  $\sigma$  of the normal distribution are determined by the linear equations

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

which provide

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

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

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

### **Example 25. CI for the mean of a large sample**

A random sample of 400 bolts produced by an automatic machine has a mean weight of 7.47 g, with a standard deviation of 0.15 g. We want to determine, for the weight of the bolts:

- (a) the confidence interval at a confidence level of 67%;
- (b) the confidence interval at a confidence level of 99%.

**Solution**

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

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

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

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

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

which provides the critical values

$$z_{0.33} = 0.9741.$$

The confidence interval for the mean  $\mu$  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%.

### Example 26. CI for the mean and standard deviation of a normal population

Repeated measurements of the electromotive force between the poles of an electric battery have yielded the following results (in V):

$i$	$V_i$	$i$	$V_i$
1	1.5710	12	1.5703
2	1.5742	13	1.5729
3	1.5751	14	1.5668
4	1.5686	15	1.5725
5	1.5712	16	1.5696
6	1.5734	17	1.5742
7	1.5682	18	1.5711
8	1.5718	19	1.5729
9	1.5737	20	1.5711
10	1.5757	21	1.5689
11	1.5668	22	1.5731

By assuming that the population is normal, calculate the confidence interval of the mean and that of the standard deviation, both at the confidence level of 90%.

### Solution

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

i	$V_i$	$(V_i - \bar{V}) \cdot 10^2$	$(V_i - \bar{V})^2 \cdot 10^4$
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  $\sigma^2$  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  $\sigma^2$ .

(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:

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

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

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:

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

□ **Example 27. Pearson's linear correlation coefficient**

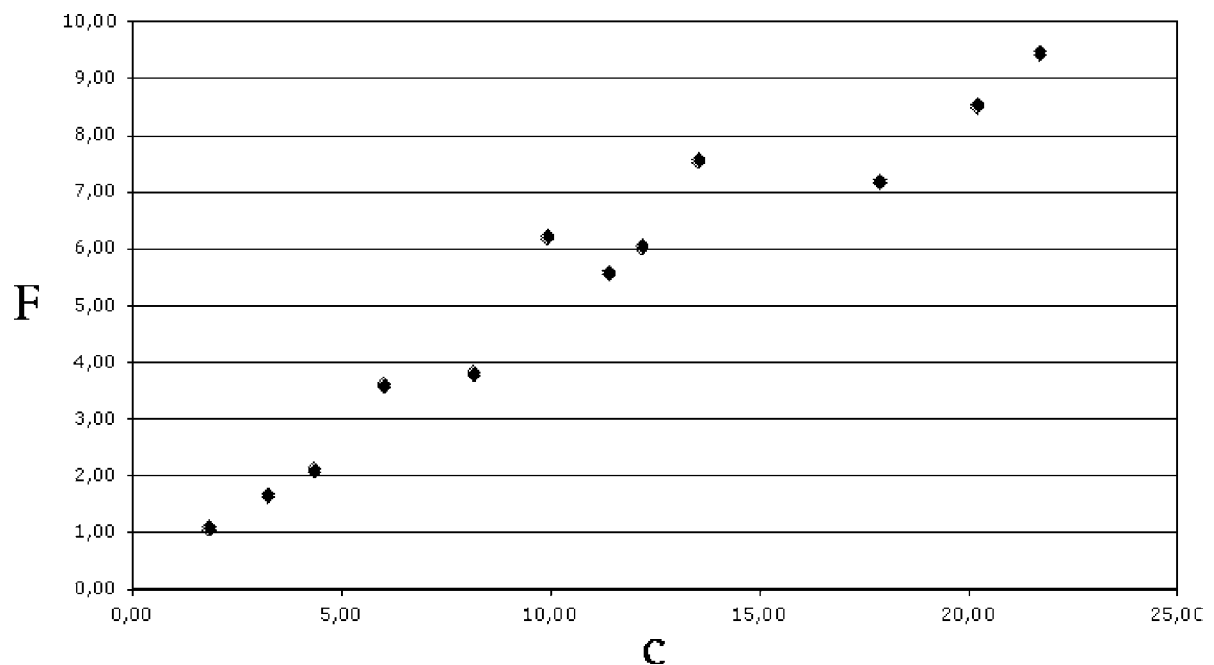
A polymerization process, not yet well standardized, yields polymer samples whose degree of crystallinity  $c$  and resistance to traction  $F$  vary at random according to a joint probability distribution which can be assumed normal. We wonder whether the two quantities may be correlated. To test the conjecture we carry out 12 measurements on the same number of samples, whose results are summarized in the following table (in arbitrary units):

$i$	$c_i$	$F_i$	$i$	$c_i$	$F_i$
1	20.23	8.53	7	12.19	6.05
2	8.17	3.82	8	1.83	1.10
3	6.01	3.62	9	21.72	9.46
4	9.93	6.23	10	11.39	5.59
5	13.55	7.57	11	3.23	1.66
6	17.89	7.20	12	4.34	2.12

Apply Pearson's linear correlation coefficient to check whether the quantities  $c$  and  $F$  can be regarded as stochastically independent, at a significance level of 5% and 1%. Comment on the physical meaning of the result.

### Solution

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:

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

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.

### Example 28. Regression straight line

The table below collects some experimental measurements of the electrical resistivity  $\rho$  (in  $10^{-7} \Omega \cdot \text{m}$ ) of a metallic alloy as a function of the absolute temperature  $T$  (in K):

$k$	$T_k$	$\rho_{k1}$	$\rho_{k2}$	$\rho_{k3}$	$\rho_{k4}$
1	440	1.55	1.80	1.75	1.65
2	600	2.10	1.95	2.05	2.00
3	800	2.40	2.30	2.50	×
4	1000	2.60	2.90	2.85	2.95
5	1200	3.25	3.30	3.40	3.45
6	1300	3.65	3.70	3.75	×
7	1400	3.75	3.85	3.80	3.95

The random error on the temperatures  $T_k$  is negligible, whereas the resistivity data  $\rho_k$  are independent normal random variables with the same standard deviation  $\sigma$  (i.e., the system is homoscedastic).

Find:

- (i) the least squares regression straight line of the form

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

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

- (ii) the 95% confidence intervals of the parameter  $\mu$  and of the slope  $\kappa$ ;  
 (iii) the 95% confidence region for predictions;

- (iv) the 95% confidence interval for the value of  $\rho$  as predicted at  $T = 1170$  K;
- (v) the goodness of fit of the regression model if  $\sigma = 0.09$  is the common standard deviation of all the data  $\rho$ .

### Solution

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  $\rho$  along the ordinate axis:

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

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

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  $(T_i, \rho_i)$  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  $\mathcal{X}^2$  fitting reduces to the usual least-squares fitting and the best-fit estimates  $m$  and  $q$  of the parameters can be written as

$$m = \frac{1}{n} \sum_{i=1}^n \rho_i = 2.815384615$$

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

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  $\mu$  and that of the slope  $\kappa$  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  $\mu$  is

$$2.815384615 \pm 0.043794781 = [2.771589834, 2.859179396]$$

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

$$\begin{aligned} & 0.002273762022 \pm 0.0001281951256 = \\ & = [0.002145566898, 0.002401957150]. \end{aligned}$$

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:

$$m = 2.815384615$$

$$q = 0.002273762022$$

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

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

$$\begin{aligned}
V &= 1 + \frac{1}{n} + \frac{1}{\sum_{i=1}^n (T_i - \bar{T})^2} (T_0 - \bar{T})^2 = \\
&= 1 + \frac{1}{26} + \frac{(T_0 - 956.1538461)^2}{3.034415385 \cdot 10^6} = \\
&= 1.0384615 + 3.2955277 \cdot 10^{-7} (T_0 - 956.1538461)^2
\end{aligned}$$

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

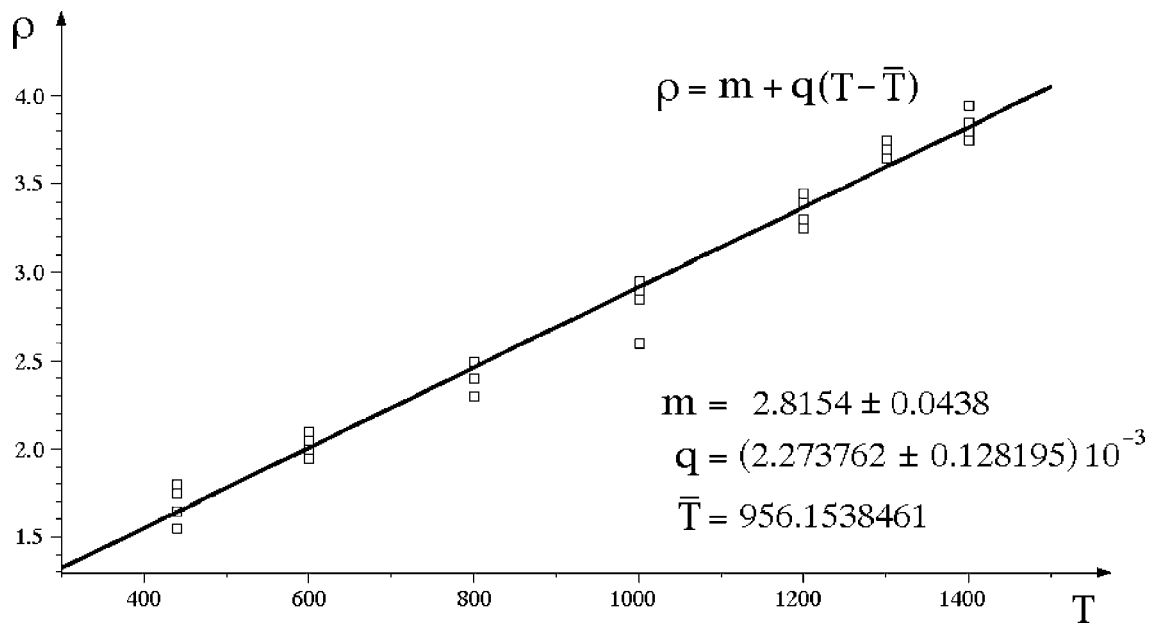
The CI for the prediction of  $\rho$  at  $T = T_0$  becomes then:

$$\begin{aligned}
\rho_0 &= 2.815384615 + 0.002273762022 (T_0 - 956.1538461) \pm \\
&\pm 2.064 \cdot \sqrt{1.0384615 + 3.2955277 \cdot 10^{-7} (T_0 - 956.1538461)^2} \cdot \\
&\quad \cdot \sqrt{\frac{0.2809376284}{24}}
\end{aligned}$$

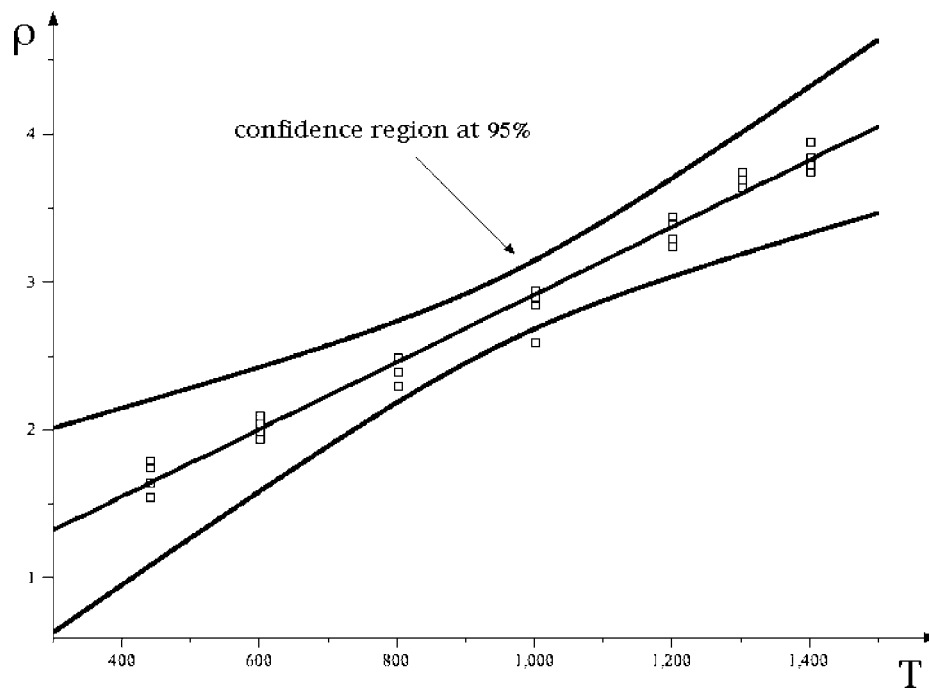
and performing the calculations reduces to:

$$\begin{aligned}
\rho_0 &= 0.641318312 + 0.002273762022 T_0 \pm \\
&\pm 0.22331044 \cdot \\
&\quad \cdot \sqrt{1.0384615 + 3.2955277 \cdot 10^{-7} (T_0 - 956.1538461)^2}
\end{aligned}$$

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



The confidence region for predictions, at the confidence level of 95%, is shown in the figure below (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  $\rho$  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  $\rho$  at  $T = 1170$  can be obtained by posing  $T_0 = 1170$  in the previous formula

$$\begin{aligned} \rho_0 &= 0.641318312 + 0.002273762022 T_0 \pm \\ &\pm 0.2233104403 \cdot \\ &\cdot \sqrt{1.0384615 + 3.2955277 \cdot 10^{-7} (T_0 - 956.1538461)^2} . \end{aligned}$$

We have therefore:

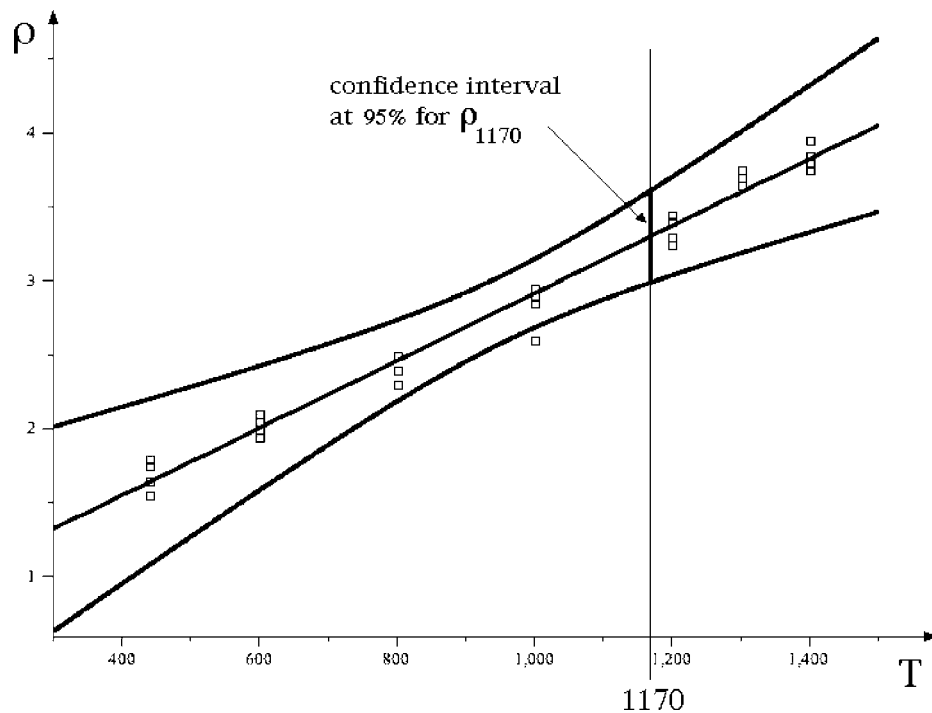
$$\begin{aligned} \rho_0 = \rho_{1170} &= [3.072410226, 3.530829532] = \\ &= 3.301619879 \pm 0.229209653 \end{aligned}$$

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  $\rho_{n-2}$  stands for the  $\chi^2$  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  $\chi^2$  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,  $\mu$  and  $\kappa$ ; as a consequence, the NSSAR obeys a  $\chi^2$  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  $\chi^2$  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
34.6836	$Q$
36.415	0.05

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

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

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

for instance by using the Maple command line

```
1 - stats[statevalf, cdf, 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.

### **Example 29. Unpaired $t$ -test on the means of two normal populations**

11 samples of a conductive polymer are subjected to a thermal treatment at the temperature of 320 K. Further 14 samples of the same material are subjected to a treatment of the same duration, but at a temperature of 370 K. The electrical conductivity of all the samples is then measured, obtaining the results in the table (data in  $10^{-1}\text{S} \cdot \text{m}^{-1}$ ):

$T = 320 \text{ K}$	$T = 370 \text{ K}$
5.52	6.33
6.30	5.89
5.00	5.25
5.80	7.70
6.45	7.66
4.38	5.44
5.25	7.10
6.02	7.48
4.77	5.95
5.68	5.05
6.13	6.47
	6.10
	5.24
	7.41

Assuming that the populations are normal, and after having checked whether the relative variances can be regarded as equal or not, determine with a significance level of 5% if the temperature of the thermal treatment has a significant effect on the electrical conductivity of the material.

### Solution

Let us denote with  $\mu_1$  and  $\mu_2$  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 <sub>i</sub>	z <sub>i</sub>	Dy	Dy <sup>2</sup>	Dz	Dz <sup>2</sup>
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  $\alpha$ , if

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

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

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

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:

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

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

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

$$= FINV(0, 975; 10; 13)$$

$$= FINV(0, 025; 10; 13)$$

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  $\sigma_1^2$  and  $\sigma_2^2$  are equal at the significance level  $\alpha = 5\%$ . As a conclusion, a rigorous unpaired  $t$ -test with equal variances could be applied.*

### **Example 30. Paired $t$ -test on the means of two normal populations**

A set of 8 samples is subjected to a physico-chemical treatment which is supposed to affect significantly the thermal conductivity of the material constituting the samples. The thermal conductivity is measured for each sample before and after the treatment, providing the following results (in  $\text{W} \cdot \text{m}^{-1} \cdot \text{K}^{-1}$ ):

before the treatment	after the treatment
1.02	1.16
0.95	1.04
0.73	0.85
0.96	1.10
0.79	1.09
1.01	0.93
0.42	0.95
0.87	1.03

We want to check, with a significance level of 2%, the hypothesis that the means  $\mu_1$  and  $\mu_2$  of the measured quantity are the same before and after the treatment, by assuming that the populations are normal.

### Solution

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  $\alpha$ , is of the form

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

and for  $n = 8$ ,  $\alpha = 0.02$  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.*

### Example 31. Logarithmic differential method

The Young's modulus  $E$  of a material is related to the Poisson's ratio  $\nu$  and the shear modulus  $G$  of the same material by means of the general relationship

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

Let the Poisson's ratio and the shear modulus of a tin-copper alloy be given by:

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

Determine the Young's modulus of the material and the corresponding absolute error. Estimate the precision of the result.

### Solution

The Young's modulus is expressed by the formula

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

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

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

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

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

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

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

and whence the differential

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

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

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

The absolute error of  $E$  is therefore

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

and the final result takes the form

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

The precision of the estimate is expressed by the percent error

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

and appears quite satisfactory.

### **Example 32. Chauvenet's criterion**

The following density data of an expanded polystyrene foam are assumed to obey a normal distribution (data in  $\text{kg m}^{-3}$ )

$$\begin{array}{cccccc} 199, & 184, & 186, & 192, & 196, & 191, \\ 190, & 183, & 185, & 195, & 189, & 182. \end{array}$$

Check for the possible presence of a result not belonging to the statistical population of the sample.

## Solution

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

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

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

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

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

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

the suspect value from the mean, in units of  $s$ , is expressed as

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

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

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

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

1.7500	0.45994
1.7578	$P$
1.7600	0.46080

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

which provides  $P = 0.46061$ .

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

**Example 33.  $\chi^2$ -test for a normal distribution**

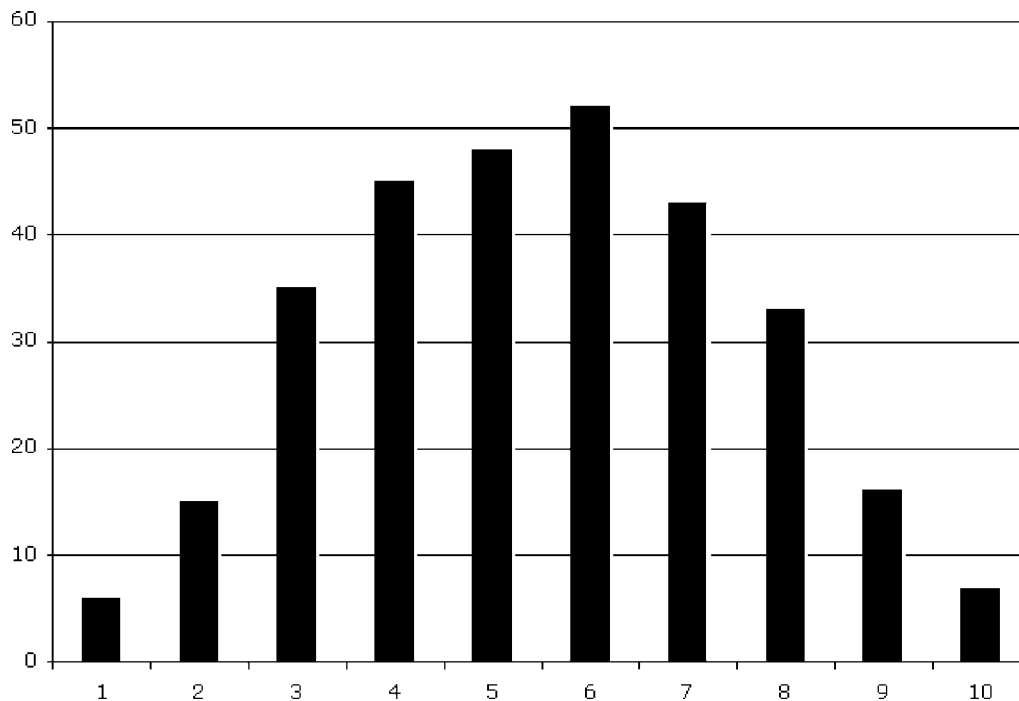
In order to check whether the shear modulus  $G$  of a polymer follows a normal distribution, we carry out 300 measurements of shear modulus and compute the relative sample mean  $\bar{m}$  and standard deviation  $s$ . Due to the large size of the sample,  $\bar{m}$  and  $s$  can be assumed as good estimates of the mean and standard deviation of the whole population, respectively. The binned results are shown in the following frequency table:

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

Test the hypothesis of the normal distribution with a significance level: (a) of 10%; (b) of 5%.

**Solution**

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



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

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

against the alternative hypothesis

$H_1$  :  $H_0$  is false .

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

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

and the number of result classes (i.e. the histogram intervals) is  $k = 10$ . In the presence of  $c = 2$  constraints on the mean and

the standard deviation, if  $H_0$  holds true the  $\mathcal{X}^2$  of data obeys approximately a  $\mathcal{X}^2$  distribution with

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

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

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

and introduce the integral of the standard normal distribution

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

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

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

$$\begin{aligned}
 n_2 &= 300 \cdot \int_{-2}^{-1.5} p(z) dz = 300 \cdot \int_{1.5}^2 p(z) dz = \\
 &= 300 \cdot \left( \Phi(2) - \Phi(1.5) \right) = \\
 &= 300 \cdot (0.47725 - 0.43319) = 13.218
 \end{aligned}$$

$$\begin{aligned}
 n_3 &= 300 \cdot \int_{-1.5}^{-1} p(z) dz = 300 \cdot \int_1^{1.5} p(z) dz = \\
 &= 300 \cdot \left( \Phi(1.5) - \Phi(1) \right) = \\
 &= 300 \cdot (0.43319 - 0.34134) = 27.555
 \end{aligned}$$

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

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

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

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

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

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

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

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

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

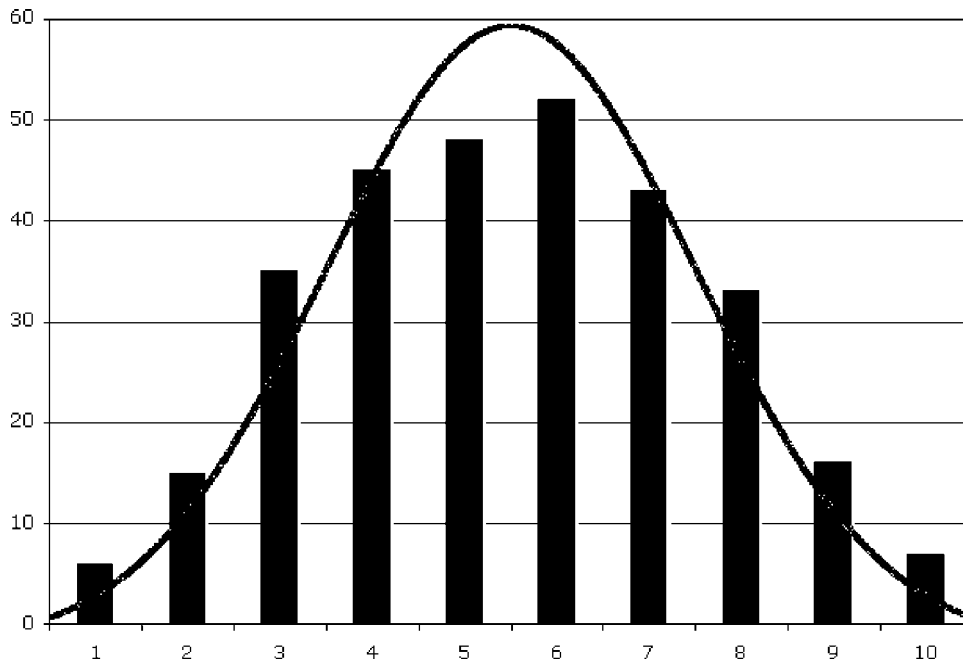
$$\chi^2_{[1-\alpha](7)} = \chi^2_{[0.90](7)} = 12.017 \quad \text{for } \alpha = 0.10$$

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

In both cases the  $\chi^2$  of the sample is smaller: as a conclusion, with both the significance levels of 10 and 5% *the null*

*hypothesis cannot be rejected.* The data sample supports the feeling that *the shear modulus of the polymer follows a normal distribution.*

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



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

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

and that of the class 9 to the point

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

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

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

which provide

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

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

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

### **Example 34. CI for the mean of a large sample**

A random sample of 500 screws produced by an automatic machine has a mean length of 7.45 mm, with a standard deviation of 0.05 mm. Determine the confidence interval for the length of the screws:

- (a) at a confidence level of 67%;
- (b) at a confidence level of 99%.

### **Solution**

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

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

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

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

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

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

which provides the critical value

$$z_{0.33} = 0.9741 .$$

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

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

with the sample mean and standard deviation given by

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

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

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

and therefore, after trivial calculations,

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

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

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

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

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

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

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

due to the approximate relationship

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

The CI of the mean becomes then

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

and, equivalently,

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

and finally

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

An alternative expression puts into evidence the absolute error:

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

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

### Example 35. Regression straight line

The table below collects some experimental measurements of the dynamic viscosity  $\mu$  (in  $10^{-4}$  Pa s) of pure water as a function of the temperature  $T$  (in  $^{\circ}\text{C}$ ):

$k$	$T_k$	$\mu_{k1}$	$\mu_{k2}$	$\mu_{k3}$	$\mu_{k4}$
1	10	13.08	13.00	12.92	13.20
2	20	10.03	9.992	10.06	10.18
3	30	7.978	7.970	7.989	×
4	40	6.531	6.539	6.520	6.524
5	50	5.471	5.454	5.465	5.482
6	60	4.668	4.652	4.679	×
7	70	4.044	4.032	4.048	4.055

The temperatures  $T_k$  are affected by no significant error, while the viscosity data  $\mu_k$  can be described as independent normal random variables with the same standard deviation  $\sigma$ .

Determine:

- (i) the least squares regression straight line of the form

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

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

- (ii) the 95% confidence intervals of the regression parameters;  
(iii) the 95% confidence region for predictions;  
(iv) the 95% confidence interval for the value of  $\mu$  predicted at  $T = 25$  °C;  
(v) the goodness of fit of the regression model if  $\sigma = 0.05$  is the common standard deviation of all the data  $\mu$ .

### Solution

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

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

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

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

(i) *Regression straight line*

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

$$a = \frac{1}{n} \sum_{i=1}^n \mu_i = 7.483192308$$

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

with  $n = 26$  and

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

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

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

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

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

(ii) *Confidence intervals for the regression parameters*

By definition, the sum of squares around regression holds:

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

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

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

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

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

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

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

with:

$$a = 7.483192308$$

$$b = -0.1450715928$$

$$\text{SSAR} = 17.30878358$$

$$\sum_{i=1}^{26} (T_i - \bar{T})^2 = 10696.15385$$

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

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

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

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

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

$$\begin{aligned} & - 0.1450715928 \pm 0.01694819786 = \\ & = [-0.1620197907, -0.1281233949] . \end{aligned}$$

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

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

while

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

(iii) *Confidence region*

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

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

where we have:

$$a = 7.483192308$$

$$b = -0.1450715928$$

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

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

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

$$= 1 + \frac{1}{26} + \frac{(T_0 - 39.61538461)^2}{10696.15385} =$$

$$= 1.038461538 +$$

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

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

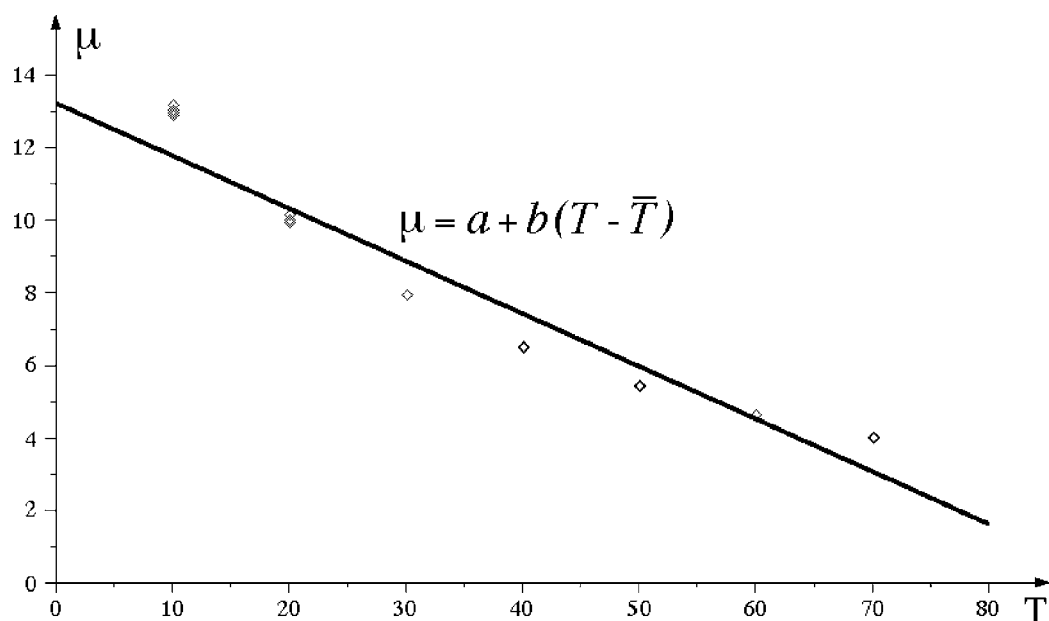
The CI for the prediction of  $\mu$  at  $T = T_0$  becomes then:

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

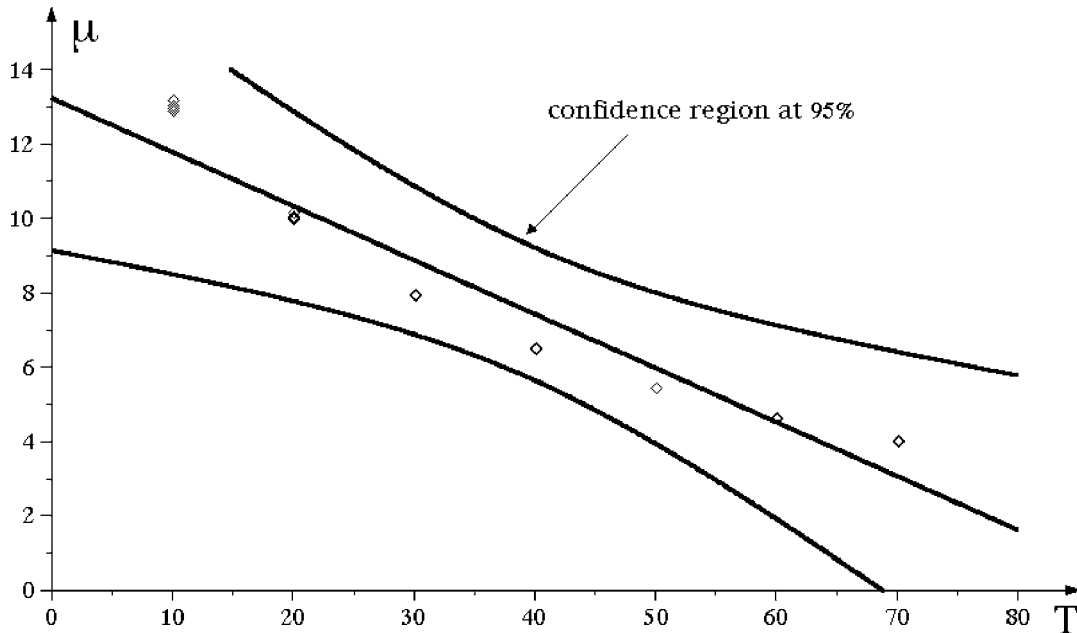
and performing the calculations reduces to:

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

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



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



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

*(iv) Confidence interval for a prediction*

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

formula

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

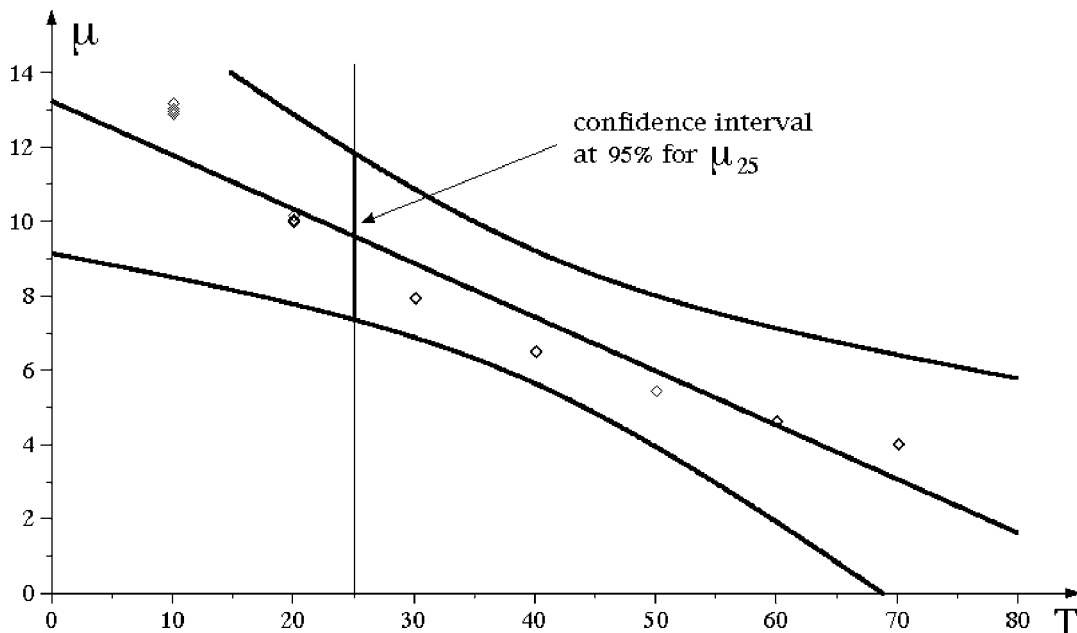
We have therefore:

$$\begin{aligned} \mu_0 = \mu_{25} &= [7.800165743, 11.40677312] = \\ &= 9.603469432 \pm 1.803303688 \end{aligned}$$

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

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

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



*(v) Goodness of fit*

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

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

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

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

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

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

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

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

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

$$\begin{aligned} Q &= \text{Probability}\{\chi^2 \geq 6923.5143200\} = \\ &= \int_{6923.5143200}^{+\infty} p_{24}(\chi^2) d\chi^2 \end{aligned}$$

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

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

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

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

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

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

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

### Example 36. CI for the mean and standard deviation of a normal population

The electrical conductivity of a metallic alloy has been repeatedly measured, providing the data table below (in  $10^6 \text{ S m}^{-1}$ ):

$i$	$\sigma_i$	$i$	$\sigma_i$
1	25.711	12	25.707
2	25.741	13	25.729
3	25.752	14	25.656
4	25.687	15	25.720
5	25.713	16	25.695
6	25.734	17	25.742
7	25.681	18	25.753
8	25.717	19	25.743
9	25.739	20	25.713
10	25.753	21	25.688
11	25.671	22	25.730

By assuming a normal population, compute the confidence interval of the mean and that of the standard deviation, both at the confidence level of 90%.

### Solution

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

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

on having denoted with  $\sigma_i$ ,  $i = 1, \dots, 22$ , the conductivity data of the sample. We can then determine the residuals of the data with respect to the mean and the corresponding squares:

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

i	$\sigma_i$	$(\sigma_i - \bar{\sigma}) \cdot 10^2$	$(\sigma_i - \bar{\sigma})^2 \cdot 10^4$
19	25.743	2.5954545	6.7363843
20	25.713	-0.4045455	0.1636570
21	25.688	-2.9045455	8.4363843
22	25.730	1.2954545	1.6782025

from which we deduce the sample variance:

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

and the sample estimate of the standard deviation:

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

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

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

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

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

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

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

so that the confidence interval becomes

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

or, equivalently,

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

For all practical purposes an approximation of the form

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

can be considered more than satisfactory.

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

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

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

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

$$\frac{1}{\chi^2_{[0.05](21)}} 21 s^2 = \frac{1}{11.591} 21 \cdot 7.7690260 \cdot 10^{-4} =$$

$$= 14.0755 \cdot 10^{-4}$$

and the CI of the variance is expressed as

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

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

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

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

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

### Example 37. Pearson's linear correlation coefficient

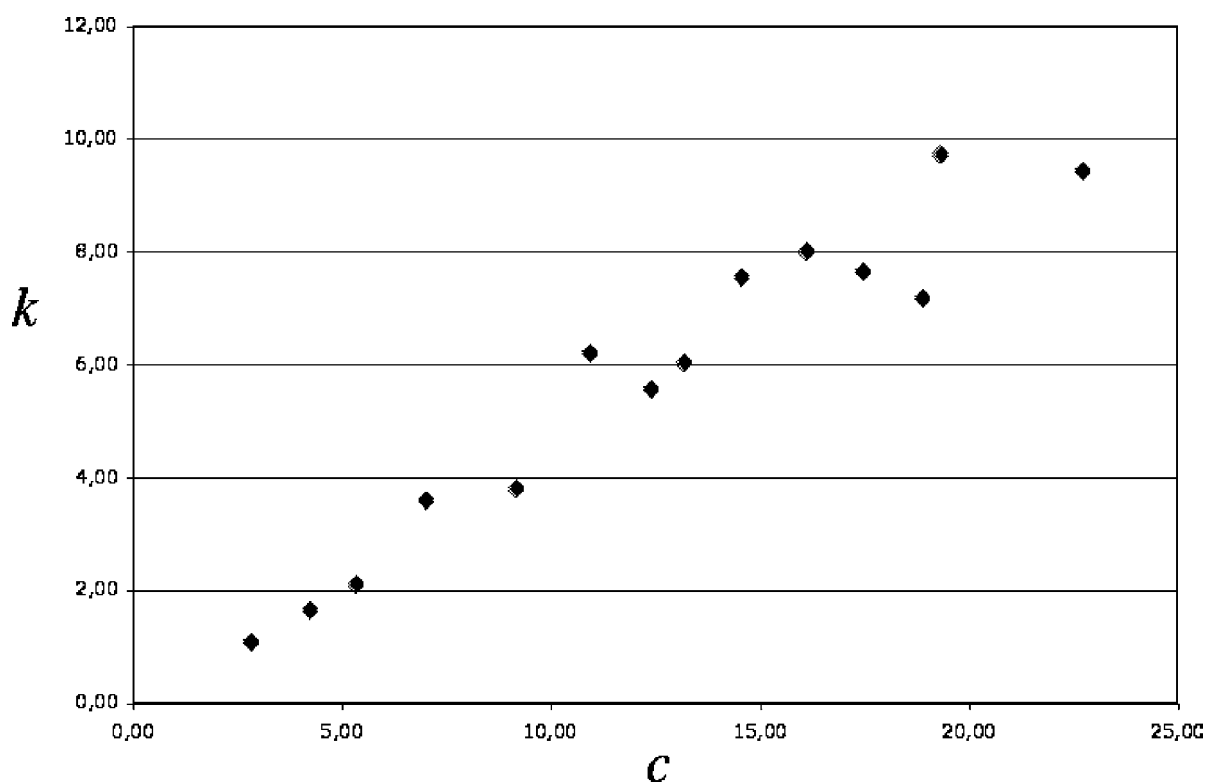
A chemical process, not yet well standardized, provides product samples whose degree of crystallinity  $c$  and thermal conductivity  $k$  vary at random according to a normal joint probability distribution. We guess that the two quantities may be correlated. To test the conjecture, 14 measurements are carried out on the same number of samples and the results are summarized in the table below (in arbitrary units):

$i$	$c_i$	$k_i$	$i$	$c_i$	$k_i$
1	16.11	8.02	8	2.83	1.10
2	9.17	3.82	9	22.72	9.46
3	7.01	3.62	10	12.39	5.59
4	10.93	6.23	11	4.23	1.66
5	14.55	7.57	12	5.34	2.12
6	18.89	7.20	13	19.32	9.75
7	13.19	6.05	14	17.47	7.68

Applying Pearson's linear correlation coefficient, check whether  $c$  and  $k$  can be regarded as stochastically independent at a significance level ( $\alpha$ ) of 5% and ( $\beta$ ) of 1%. Comment on the physical meaning of the result.

### Solution

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



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

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

and allow us to calculate the sum of products of residuals

$$SS_{ck} = \sum_{i=1}^{12} (c_i - \bar{c})(k_i - \bar{k}) = 218.313450$$

and of the relative squares:

$$SS_{cc} = \sum_{i=1}^{12} (c_i - \bar{c})^2 = 491.026293$$

$$SS_{kk} = \sum_{i=1}^{12} (k_i - \bar{k})^2 = 104.163750,$$

as illustrated in the following table:

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

The linear correlation coefficient becomes

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

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

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

against the alternative hypothesis

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

by means of the Fisher's random variable

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

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

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

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

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

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

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

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

the relation should be direct: *the thermal conductivity of the material increases with the degree of crystallinity.*

**Example 38. Unpaired  $t$ -test on the means of two normal populations**

10 samples of a metallic alloy are thermally treated at the temperature of 800 K. Other 14 samples of the same material are subjected to a treatment of the same duration, but at a temperature of 1200 K. The electrical conductivity of all the samples is then measured. The table below lists the results (data in  $10^6 \text{ S m}^{-1}$ ):

$T = 800 \text{ K}$	$T = 1200 \text{ K}$
16.30	15.89
15.00	15.25
15.80	17.70
16.45	17.66
14.38	15.44
15.25	17.10
16.02	17.48
14.77	15.95
15.68	15.05
15.52	16.32
	16.47
	16.10
	15.24
	17.41

Assuming that the populations are normal, check whether the relative variances can be regarded as equal or not. Determine

then, with a significance level of 5%, if the temperature of the thermal treatment significantly affects the electrical conductivity of the material.

### Solution

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

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

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

#### (i) Test on the variances

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

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

has a value within the acceptance interval

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

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

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

and therefore the test variable holds

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

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

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

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

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

*(ii) Test on the mean*

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

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

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

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

In the present case we have

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

and consequently

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

Moreover, for  $\alpha = 0.05$  there holds

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

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

The detailed calculations are illustrated in the table below:

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

### Example 39. Paired $t$ -test on the means of two normal populations

We guess that a thermal treatment may affect the degree of crystallinity of a polymer. To check our conjecture, the degree of crystallinity of 10 samples is measured prior to and after the treatment. The data are listed below:

before treatment	after treatment
10.5	12.3
13.0	13.2
15.3	16.3
11.2	12.5
14.6	13.9
12.3	14.1
14.0	15.2

before treatment	after treatment
13.2	12.8
14.9	15.5
11.9	12.7

We want to check, with a significance level of 2%, the hypothesis that the treatment actually affects the degree of cristallinity, by assuming that the populations are normal.

### Solution

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

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

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

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

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

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

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

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

since

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

A simple calculation provides the means:

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

while the paired sample estimate of the variance holds

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

and the corresponding standard deviation is therefore

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

The test variable takes the value

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

Here is the table of the detailed calculations:

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

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

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

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

#### **Example 40. Multiple linear regression by chi-square fitting**

The following table summarizes the results of many repeated measurements a physical quantity  $f$  carried out for different values of two independent quantities  $x$  and  $y$  (arbitrary units). Each result is a mean of many independent measurements, whose distribution can be assumed normal with a common

standard deviation  $\sigma = 0.030$ . In contrast, the variables  $x$  and  $y$  are affected by negligible random errors.

$i$	$x_i$	$y_i$	$f_i$
1	1.575	3.307	2.248
2	3.884	0.420	2.086
3	0.382	3.569	2.072
4	0.021	3.015	1.489
5	3.182	3.925	2.800
6	3.377	3.347	2.566
7	1.679	0.957	2.351
8	0.509	0.330	1.719
9	2.428	3.175	1.605
10	0.454	1.165	1.863
11	3.426	0.174	2.315
12	0.151	0.416	1.632
13	3.752	3.915	3.647
14	1.422	1.243	2.388
15	0.663	1.795	2.145
16	2.774	3.425	1.814
17	0.881	0.460	1.912
18	0.736	3.601	2.428
19	2.704	3.005	1.635
20	3.531	1.652	1.414
21	2.010	2.277	2.009

The reader is asked to determine:

(i) a linear regression model of the form

$$\alpha_1 + \alpha_2 xy + \alpha_3 \sin(x\sqrt{y}),$$

where  $\alpha_1, \alpha_2, \alpha_3$  are appropriate constants to be fitted by the least-squares method;

- (ii) the CI, at a confidence level of 99%, of all the model parameters;
- (iii) the 99%-confidence region for predictions;
- (iv) the 99%-confidence interval for the value of  $f$  predicted at  $(x, y) = (3.100, 2.200)$ ;
- (v) the goodness of fit of the regression model, specifying if the value can be reliably used to accept or reject the model.

### Answer

(i) The **linear regression model** is based on the functions

$$\phi_1(z) = 1 \quad \phi_2(z) = xy \quad \phi_3(z) = \sin(x\sqrt{y})$$

where the independent variable  $z = (x, y)$  is a vector of  $\mathbb{R}^2$ . Thus we have to solve a multiple linear regression problem by using the measured values  $f_i$  of  $f$  at the sampled points

$$z_i = (x_i, y_i) \quad i = 1, \dots, 21.$$

Moreover, the system can be regarded as homoscedastic since the variance  $\sigma^2 = 0.030^2$  is assumed independent on the sampled point. Consequently, the chi-square method simply reduces to a least-squares fitting. The normal equations take the matrix form

$$F \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix}$$

and provide the estimates  $a_1, a_2, a_3$  of the regression parameters. The symmetric matrix  $F$  takes the form

$$F = \begin{pmatrix} F_{11} & F_{12} & F_{13} \\ F_{12} & F_{22} & F_{23} \\ F_{13} & F_{23} & F_{33} \end{pmatrix}$$

and its entries can be readily calculated, for instance by an Excel worksheet:

$$F_{11} = \sum_{i=1}^{21} \phi_1(z_i)\phi_1(z_i) \frac{1}{\sigma^2} = \sum_{i=1}^{21} 1 \frac{1}{\sigma^2} = 21 \frac{1}{\sigma^2}$$

$$F_{12} = \sum_{i=1}^{21} \phi_1(z_i)\phi_2(z_i) \frac{1}{\sigma^2} = \sum_{i=1}^{21} x_i y_i \frac{1}{\sigma^2} = 91.468929 \frac{1}{\sigma^2}$$

$$\begin{aligned} F_{13} &= \sum_{i=1}^{21} \phi_1(z_i)\phi_3(z_i) \frac{1}{\sigma^2} = \sum_{i=1}^{21} \sin(x_i \sqrt{y_i}) \frac{1}{\sigma^2} = \\ &= 4.832307611 \frac{1}{\sigma^2} \end{aligned}$$

$$\begin{aligned} F_{22} &= \sum_{i=1}^{21} \phi_2(z_i)\phi_2(z_i) \frac{1}{\sigma^2} = \sum_{i=1}^{21} x_i^2 y_i^2 \frac{1}{\sigma^2} = \\ &= 816.829787221 \frac{1}{\sigma^2} \end{aligned}$$

$$\begin{aligned} F_{23} &= \sum_{i=1}^{21} \phi_2(z_i)\phi_3(z_i) \frac{1}{\sigma^2} = \sum_{i=1}^{21} x_i y_i \sin(x_i \sqrt{y_i}) \frac{1}{\sigma^2} = \\ &= -5.448276423 \frac{1}{\sigma^2} \end{aligned}$$

$$\begin{aligned}
 F_{33} &= \sum_{i=1}^{21} \phi_3(z_i) \phi_3(z_i) \frac{1}{\sigma^2} = \sum_{i=1}^{21} \sin^2(x_i \sqrt{y_i}) \frac{1}{\sigma^2} = \\
 &= 10.541634112 \frac{1}{\sigma^2}
 \end{aligned}$$

We obtain therefore

$$F = \begin{pmatrix} 21 & 91.468929 & 4.832307611 \\ 91.468929 & 816.829787221 & -5.448276423 \\ 4.832307611 & -5.448276423 & 10.541634112 \end{pmatrix} \frac{1}{\sigma^2}$$

and by matrix inversion deduce the covariance matrix  $F^{-1}$  of the parameter estimates

$$F^{-1} = \begin{pmatrix} 0.125980119 & -0.014542628 & -0.065265683 \\ -0.014542628 & 0.002907222 & 0.008168923 \\ -0.065265683 & 0.008168923 & 0.129001860 \end{pmatrix} \sigma^2$$

The known terms of the normal equations are finally given by

$$c_1 = \sum_{i=1}^{21} \phi_1(z_i) f_i \frac{1}{\sigma^2} = \sum_{i=1}^{21} f_i \frac{1}{\sigma^2} = 44.138 \frac{1}{\sigma^2}$$

$$c_2 = \sum_{i=1}^{21} \phi_2(z_i) f_i \frac{1}{\sigma^2} = \sum_{i=1}^{21} x_i y_i f_i \frac{1}{\sigma^2} = 216.428423589 \frac{1}{\sigma^2}$$

$$\begin{aligned}
 c_3 &= \sum_{i=1}^{21} \phi_3(z_i) f_i \frac{1}{\sigma^2} = \sum_{i=1}^{21} \sin(x_i \sqrt{y_i}) f_i \frac{1}{\sigma^2} = \\
 &= 14.097932518 \frac{1}{\sigma^2}
 \end{aligned}$$

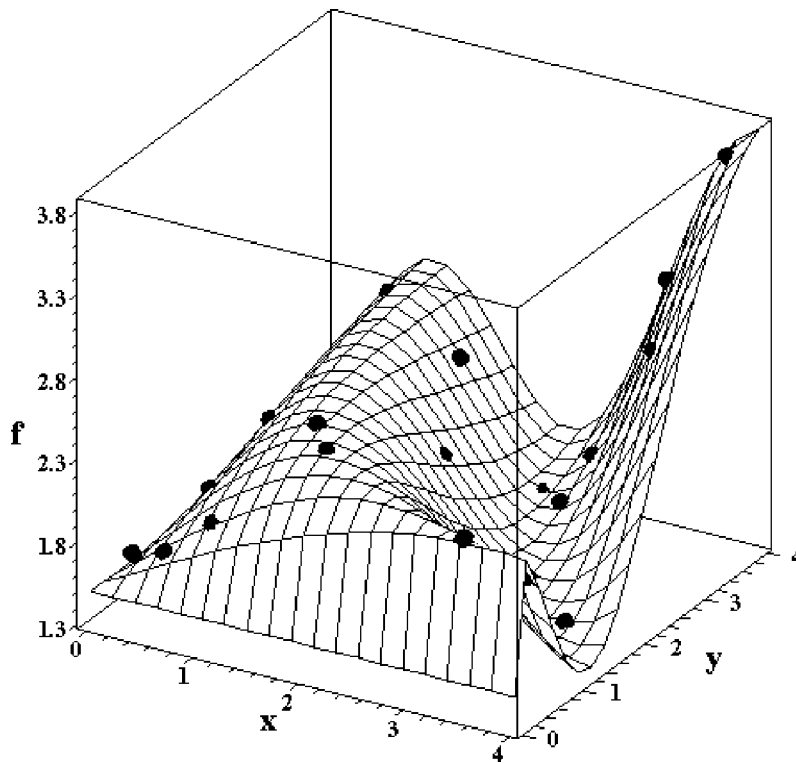
so that the best-fit estimates of the regression parameters hold

$$\begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} = F^{-1} \begin{pmatrix} c_1 \\ c_2 \\ c_3 \end{pmatrix} = \begin{pmatrix} 1.492961166 \\ 0.102487785 \\ 0.705950009 \end{pmatrix}$$

and the requested model becomes

$$f(x, y) = 1.492961166 + 0.102487785xy + 0.705950009 \sin(x\sqrt{y})$$

A graphical illustration of the result is shown in the figure below



where the graph of the model is the 2-surface immersed in the space  $\mathbb{R}^3$ , while the dots represent the experimental data.

Notice that same model can be obtained by the below sequence of Maple 11 commands:

*with(Statistics);*

to load the Maple package “Statistics”,

$$X := Matrix([[x_1, y_1], [x_2, y_2], \dots [x_{21}, y_{21}]], datatype = float);$$

to define the sampled set of independent variables  $(x, y)$ ,

$$Y := Array([f_1, f_2, \dots f_{21}], datatype = float);$$

to introduce the the corresponding values of the dependent variable  $f$ ,

$$V := Array([1, 1, 1, \dots 1, 1]);$$

in order to assign all the data the same statistical weight 1 in the application of the least-squares procedure, and finally

$$LinearFit([1, x \cdot y, \sin(x \cdot \sqrt{y})], X, Y, [x, y], weights = V);$$

to compute and display the regression model.

(ii) The **confidence intervals of the parameters**  $\alpha_1$ ,  $\alpha_2$ ,  $\alpha_3$  with a confidence level of  $1 - \eta = 0.99$  are determined by the general formula

$$\alpha_j = a_j \pm t_{[1-\frac{\eta}{2}](n-p)} \sqrt{(F^{-1})_{jj} \frac{NSSAR}{n-p}} \quad j = 1, 2, 3,$$

by posing:

$$\eta = 0.01 \quad n - p = 21 - 3 = 18$$

and inserting the numerical estimates

$$t_{[1-\frac{\alpha}{2}](n-p)} = t_{[0.995](18)} = 2.878440473$$

$$\begin{aligned} \text{NSSAR} &= \sum_{i=1}^{21} \left[ -f_i + a_1 + a_2 x_i y_i + a_3 \sin(x_i \sqrt{y_i}) \right]^2 \frac{1}{\sigma^2} = \\ &= 0.0155447625051 \frac{1}{\sigma^2} \end{aligned}$$

$$(F^{-1})_{11} = 0.125980119 \sigma^2$$

$$(F^{-1})_{22} = 0.002907222 \sigma^2$$

$$(F^{-1})_{33} = 0.129001860 \sigma^2$$

along with the estimates  $a_1$ ,  $a_2$ ,  $a_3$  of the parameters

$$a_1 = 1.492961166 \quad a_2 = 0.102487785 \quad a_3 = 0.705950009$$

Whence we obtain, for  $j = 1, 2, 3$ ,

$$\alpha_j = a_j \pm 2.878440473 \sqrt{(F^{-1})_{jj} \frac{0.0155447625051}{18} \frac{1}{\sigma^2}},$$

and consequently

$$\alpha_1 = 1.492961166 \pm 0.03002368898$$

$$\alpha_2 = 0.102487785 \pm 0.00456091641$$

$$\alpha_3 = 0.705950009 \pm 0.03038162728$$

A more meaningful approximation to the CIs may simply be

$$\alpha_1 = 1.492961 \pm 0.030024$$

$$\alpha_2 = 0.102488 \pm 0.004561$$

$$\alpha_3 = 0.705950 \pm 0.030382$$

so that the percent errors on the best-fit parameter estimates are expressed as

$$\begin{aligned}\frac{\Delta\alpha_1}{\alpha_1} &= \frac{0.030024}{1.492961} = 2.01\% \\ \frac{\Delta\alpha_2}{\alpha_2} &= \frac{0.004561}{0.102488} = 4.45\% \\ \frac{\Delta\alpha_3}{\alpha_3} &= \frac{0.030382}{0.705950} = 4.30\%\end{aligned}$$

Notice that the percent uncertainty on the parameters  $\alpha_2$  and  $\alpha_3$  is more than twice that on the parameter  $\alpha_1$ .

(iii) The **CI for the prediction**  $f_0$  of  $f$  at an arbitrary point  $z_0 = (x_0, y_0)$  of the independent variables is defined by the relationship

$$\mathbb{E}(f_0) = f(x_0, y_0) \pm t_{[1-\frac{\alpha}{2}], (n-p)} \sqrt{\text{var}(f_0) \frac{\text{NSSAR}}{n-p}}$$

where the variance of the prediction  $f_0$  can be calculated by the formula

$$\text{var}(f_0) = \sum_{j,k=1}^3 \phi_j(z_0)\phi_k(z_0)(F^{-1})_{jk} + \sigma^2$$

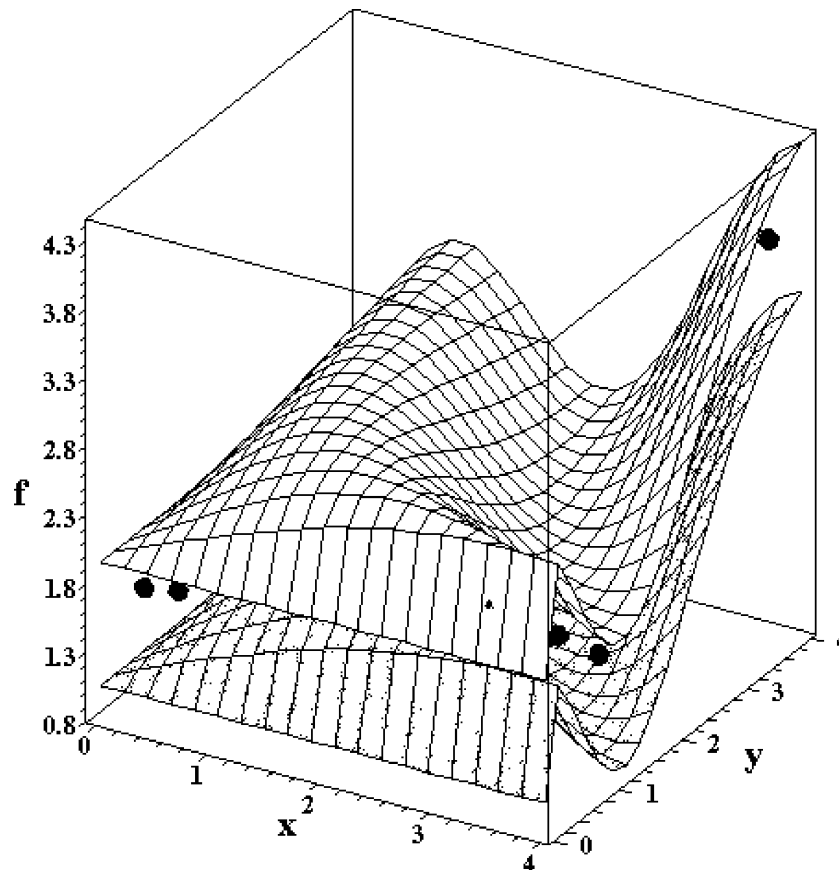
and in the homoscedastic case reduces to the form

$$\text{var}(f_0) = \left[ \sum_{j,k=1}^3 \phi_j(z_0)\phi_k(z_0) \frac{1}{\sigma^2} (F^{-1})_{jk} + 1 \right] \sigma^2,$$

the coefficients  $(F^{-1})_{jk}/\sigma^2$  being independent on  $\sigma$ . As a consequence, the absolute error on the prediction becomes

$$\begin{aligned}
& t_{[1-\frac{\eta}{2}](n-p)} \sqrt{\text{var}(f_0) \frac{\text{NSSAR}}{n-p}} = \\
& = t_{[1-\frac{\eta}{2}](n-p)} \left[ \sum_{j,k=1}^3 \phi_j(z_0) \phi_k(z_0) \frac{1}{\sigma^2} (F^{-1})_{jk} + 1 \right]^{1/2} \cdot \\
& \quad \cdot \sqrt{\frac{\text{NSSAR} \sigma^2}{n-p}} = \\
& = 2.878440473 \left[ \sum_{j,k=1}^3 \phi_j(z_0) \phi_k(z_0) \frac{1}{\sigma^2} (F^{-1})_{jk} + 1 \right]^{1/2} \cdot \\
& \quad \cdot \sqrt{\frac{0.0155447625051}{18}} = \\
& = \left[ 89.79736249 - 2.081128958 xy \right. \\
& \quad - 9.339873013 \sin(x\sqrt{y}) + 0.2080195854 x^2 y^2 \\
& \quad + 1.169017211 xy \sin(x\sqrt{y}) \\
& \quad \left. - 9.230432762 \cos^2(x\sqrt{y}) \right]^{1/2} \cdot 10^{-2}
\end{aligned}$$

and is independent on the variance  $\sigma^2$ , as expected. In the figure below the lower and the upper limits of the 99%-CI for predictions at an arbitrary point  $(x, y)$  are represented as a green and a yellow surface, while the dots denote the experimental points (as before).



The 3D-domain between the lower and the upper surface defines the confidence region of the regression model, at the assigned confidence level  $1 - \eta = 99\%$ . Notice that the two surfaces tend to divaricate at points  $(x, y)$  far from the sampled domain, which means that the CI width increases and so does the absolute error on the prediction: *we find again the same phenomenon already described for the regression straight-line* — for better clarity, in the picture the CI half-width of the predicted  $f(x, y)$  has been enlarged by a factor 5 at any point  $(x, y)$ .

(iv) The **99%-CI for the prediction** of the quantity  $f$  at  $(x, y) = (3.100, 2.200)$  can be calculated immediately by the

previous formula, which leads to

$$f(3.100, 2.200) = 1.4906 \pm 0.0930$$

or, equivalently, to

$$1.3976 \leq f(3.100, 2.200) \leq 1.5836$$

This corresponds to a relative error

$$|\Delta f(x, y)/f(x, y)| = 0.0930/1.4906 = 6.2\%$$

(v) The **goodness of fit of the model** can be determined by means of the NSSAR:

$$\begin{aligned} \text{NSSAR} &= 0.0155447625051 \frac{1}{\sigma^2} = \\ &= 0.0155447625051 \frac{1}{0.03^2} = 17.27195834 \end{aligned}$$

from which we deduce the very satisfactory value

$$\begin{aligned} Q &= \int_{\text{NSSAR}}^{+\infty} p_{n-p}(\mathcal{X}^2) d\mathcal{X}^2 = \\ &= \int_{17.27195834}^{+\infty} p_{18}(\mathcal{X}^2) d\mathcal{X}^2 = 0.5044877015 \end{aligned}$$

where  $p_{18}(\mathcal{X}^2)$  denotes the  $\mathcal{X}^2$  probability distribution with 18 d.o.f. The integral is easily calculated by the Maple 11 command

$$Q := 1 - \text{statevalf}[\text{cdf}, \text{chisquare}[18]](17.27195834);$$

The value of  $Q$  can be considered adequate since the NSSAR is between the lower bound

$$n - p - 3\sqrt{2(n - p)} = 18 - 3\sqrt{36} = 0$$

and the upper bound

$$n - p + 3\sqrt{2(n - p)} = 18 + 3\sqrt{36} = 36$$

which are typically expected for a true  $\chi^2$  random variable with  $n - p = 18$  d.o.f. — remember that a  $\chi^2$  distribution with  $\nu$  d.o.f. behaves approximately as a normal RV of mean  $\nu$  and standard deviation  $\sqrt{2\nu}$ .

As a conclusion, we can be pretty sure that the proposed linear regression model is satisfactory from a statistical point of view.

## Index of examples

1. Confidence intervals .....	1
2. CI and hypothesis testing on the mean .....	3
3. Chauvenet's criterion .....	6
4. Hypothesis testing on the mean .....	7
5. Unpaired $t$ -test for the comparison of two means (same variance) .....	8
6. Regression straight line.....	10
7. Regression straight line.....	15
8. Linear correlation coefficient (Pearson) .....	23
9. $F$ -test on the variances of two independent normal populations .....	26
10. Comparison of the means of two normal populations with unknown, but presumably equal variances (unpaired $t$ -test).....	28
11. Test on the probability parameter of a Bernoulli population .....	30
12. $\chi^2$ -test for a normal distribution .....	33
13. $\chi^2$ -test for a discrete distribution .....	36
14. $\chi^2$ -test for a Poisson distribution .....	39
15. Paired $t$ -test on two means .....	43
16. $\chi^2$ -test on the variance of a normal population .....	46
17. $z$ -test on the mean of a normal population .....	48
18. One-sided $z$ -test on the man .....	50
19. Linear correlation coefficient.....	52
20. $F$ -test to check the goodness of a linear regression fit with repeated measurements .....	56
21. Incremental $F$ -test on a linear regression .....	67
22. Logarithmic differential .....	75
23. Chauvenet's criterion .....	77
24. $\chi^2$ -test for a normal distribution .....	79

25. CI for the mean of a large sample .....	85
26. CI for the mean and standard deviation of a normal population .....	88
27. Pearson's linear correlation coefficient .....	92
28. Regression straight line .....	96
29. Unpaired <i>t</i> -test on the means of two normal populations	106
30. Paired <i>t</i> -test on the means of two normal populations ..	113
31. Logarithmic differential method .....	116
32. Chauvenet's criterion .....	118
33. $\chi^2$ -test for a normal distribution .....	121
34. CI for the mean of a large sample .....	127
35. Regression straight line .....	130
36. CI for the mean and standard deviation of a normal population .....	141
37. Pearson's linear correlation coefficient .....	145
38. Unpaired <i>t</i> -test on the means of two normal populations	149
39. Paired <i>t</i> -test on the means of two normal populations ..	153
40. Multiple linear regression by chi-square fitting .....	156