

Error analysis and data modeling

Part 1

Andrzej Lasia
Université de Sherbrooke

version 2.3 revised and extended
October 2022

Table of Contents

1	Means and errors.....	5
1.1	Introduction	5
1.2	Significant digits.....	5
1.3	Measures of errors	6
1.4	Type of errors	7
1.5	Distribution of errors	7
1.6	Integration of the Gauss curve.....	13
1.7	Standard deviation of the population and sample standard deviation	19
1.8	Standard deviation of the true value and of the mean	19
1.9	Confidence intervals	20
1.10	Confidence intervals when σ is known.....	20
1.11	Confidence intervals when σ is not known	21
1.12	Two-tailed and one-tailed tests.....	22
1.13	Pooling data	26
1.14	Weighted mean	26
2	Propagation of errors.....	30
2.1	Standard deviation of the calculated value	30
2.2	Maximal error	31
3	Linear regression	38
3.1	Introduction	38
3.2	Determination of the parameters and standard deviations of linear regression.....	39
3.3	Properties of the least-squares method	42
3.4	Standard deviation of the calculated values \hat{y}_i	43
3.5	Standard deviations of the experimental y_i	44
3.6	Correlation and determination coefficients	45
3.7	Linear regression for $y = b_1 x$	48
3.8	Error of x_c value calculated from regression	51
3.9	Calibration	54
3.10	Sensitivity	54
3.10.1	Detection limit and dynamic range	55
3.10.2	Selectivity.....	57
3.11	The method of standard additions	59

3.12	Matrix description of the least-squares method.....	63
3.13	Polynomial regression	66
3.14	Multiple linear regression.....	67
3.15	Weighted least squares regression.....	67
3.16	Linear regression with errors in y and x	70
3.17	Variances and covariances in error propagation.....	71
3.18	Intersection of two straight lines	72
3.19	Numerical problems related to the regression analysis	76
3.20	Nonlinear regression.....	79
3.21	Dealing with nonlinear regression with errors in y and x.....	83
4	Statistical tests on average(s).....	85
4.1	Introduction	85
4.2	Test χ^2	86
4.3	Test for outliers, Dixon's Q-test.....	90
4.4	Test for outliers, Grubbs' G test	93
4.5	<i>p</i> -level test.....	97
4.6	Test <i>u</i>	97
4.7	Test <i>t</i> , comparison with the standard	99
4.8	Comparison of two means	103
4.8.1	Test of equality of two means when the variances are the same.....	104
4.8.2	Test of equality of two means when the variances are different.....	104
4.9	Paired <i>t</i> -test for comparing individual differences of two samples	106
4.10	Test <i>F</i> for the comparison of variances	108
5	Test of regression parameters	114
5.1	Rejection of the point in regression, outliers.....	114
5.1.1	Simple <i>t</i> -test.....	115
5.1.2	Internally studentized residuals.....	116
5.1.3	Jack-knifed or externally studentized residuals	116
5.1.4	Cook's distance	117
5.2	Statistical importance of the regression parameters	118
5.2.1	<i>t</i> -test of the importance of regression parameters	119
5.2.2	<i>F</i> -test of the importance of regression parameters	119
5.3	ANOVA.....	121
5.4	Tests in multiple regression.....	139

5.5	Akaike information criterion	144
5.5.1	General equation	144
5.5.2	Unit weights	145
5.5.3	Proportional weights	145
5.5.4	General weights problem	146
5.5.5	Corrected <i>AIC</i>	147
5.5.6	Akaike weights.....	147
6	Interpolation	157
6.1	Polynomial interpolation	157
6.2	Splines	159
7	Smoothing.....	169
7.1	Simple data reduction	170
7.2	Simple digital filters	171
7.2.1	Moving central average square filter.....	172
7.2.2	Exponential filter.....	173
7.2.3	Symmetrical triangular filter	174
7.2.4	Bi-exponential filter	175
7.2.5	Adjacent-averaging filter.....	176
7.3	Savitzky-Golay filter	178
7.4	Polynomial approximation	186
7.5	FFT smoothing	190
7.6	Smoothing splines.....	197
7.7	Cross-validation.....	198
7.8	B-splines	199
7.9	LOESS/LOWESS.....	202
7.10	Digital differentiation and integration	205
7.10.1	Digital differentiation.....	205
7.10.2	Numerical integration.....	206
7.11	Conclusion.....	209
8	Excel functions.....	210
9	References.....	210

1 Means and errors

1.1 Introduction

In the research in analytical, physical, and other domains of chemistry the researchers are confronted with modeling of the experimental data, error determination, and statistical tests to obtain the model parameters. This is a very important part of the experimental research but often poorly understood.

The purpose of this text is to present these topics together with the exercises. Most of the exercises might be performed using easily available Excel, Origin, and SigmaPlot. Unfortunately, Excel is not certified for the statistical analysis.¹⁻³ From its version 2013 it was improved and most of the earlier errors eliminated. It can be used in typical cases, however, certified programs as Minitab should be used for the certified statistical analysis.

Although excellent books on statistics and data analysis were published,⁴⁻¹⁹ however, they are often too general or too advanced for chemists, and often do not use the new corrected Excel functions. The purpose of these notes is to present a comprehensive survey of methods used in error and data analysis and modeling of the experimental data. All these issues are illustrated by Examples (files in Excel or Origin) which can be inspected. Besides, there are Examples (with solutions) which should be solved to better understand these methods. This book is divided in two parts, first mainly for simple modeling, error analysis, and statistical tests, second on more advanced data reduction, modeling, interpolation, smoothing and numerical integration and differentiation. The second book: Data analysis and modeling, Part 2, Chemometrics, treats the analysis of larger amount of experimental data and multivariate analysis.

1.2 Significant digits

All the experimental measurements are obtained with certain error. Number of significant digits must correspond to the precision of the measurements. In the physicochemical or analytical measurements **standard deviation and confidence limits determine the precision of the results**. If the numbers are too precise, they should be rounded. Of course, in mathematics there are precise numbers which can be determined with any desired precision, e.g. π or e (base of the natural logarithms).

The general rule in calculations is **all the calculations are carried out with the maximal precision and they are rounded only at the end**, otherwise the errors might be introduced at each operation. Below, there are examples of rounding off the numbers.

Rounding

37.56 \rightarrow 37.6

37.54 \rightarrow 37.5

37.65 \rightarrow ?

In this case, to avoid accumulation of errors when the last number is 5 and the number before it is even one should round to the smaller value and if it is odd to the larger:

37.65 \rightarrow 37.6

37.35 \rightarrow 37.4

Other examples:

23.4	3 significant digits
12.40	4 significant digits
0.002	1 significant digits
0.0023	2 significant digits
0.00270	3 significant digits

How many significant digits is in:

240	?
2.4×10^2	2 significant digits
2.40×10^2	3 significant digits

Multiplication/division

One should keep number of significant digits corresponding to the least precise number:

$$\begin{aligned}
 7.643 \times 15.3 &= 116.9379 && 4 \text{ significant digits} \times 3 \text{ significant digits} = 3 \text{ significant digits} \\
 &= 117 \\
 7.8933 \times 15 &= 118.3995 && 2 \text{ significant digits} = 1.2 \times 10^2 \\
 68.233^2 &= 4655.7423 && 5 \text{ significant digits} = 4655.7
 \end{aligned}$$

Addition/subtraction

One should keep precision corresponding to the least precise number

$$386.0 + 67.241 = 453.241 \approx 453.2$$

$$\begin{array}{r}
 386 \\
 67.241 \\
 1.32 \\
 + 64.5 \\
 \hline
 516.421 \approx 516
 \end{array}$$

1.3 Measures of errors

The purpose of the statistics is to make conclusions about the experimental data. There are two principal measures of errors:

1) Accuracy

a. **absolute error**, that is the difference between the measured, x_i , and the true value, μ :

$$x_i - \mu$$

b. **relative error**, $(x_i - \mu)/\mu$, it is often expressed in %

2) **Precision** characterizes reproducibility of the data when one repeats the same measurements several times in the same way. The measures are:

a. **standard deviation**, σ

b. **variance**, σ^2

Differences between accuracy and precision are displayed in Fig. 1.1.

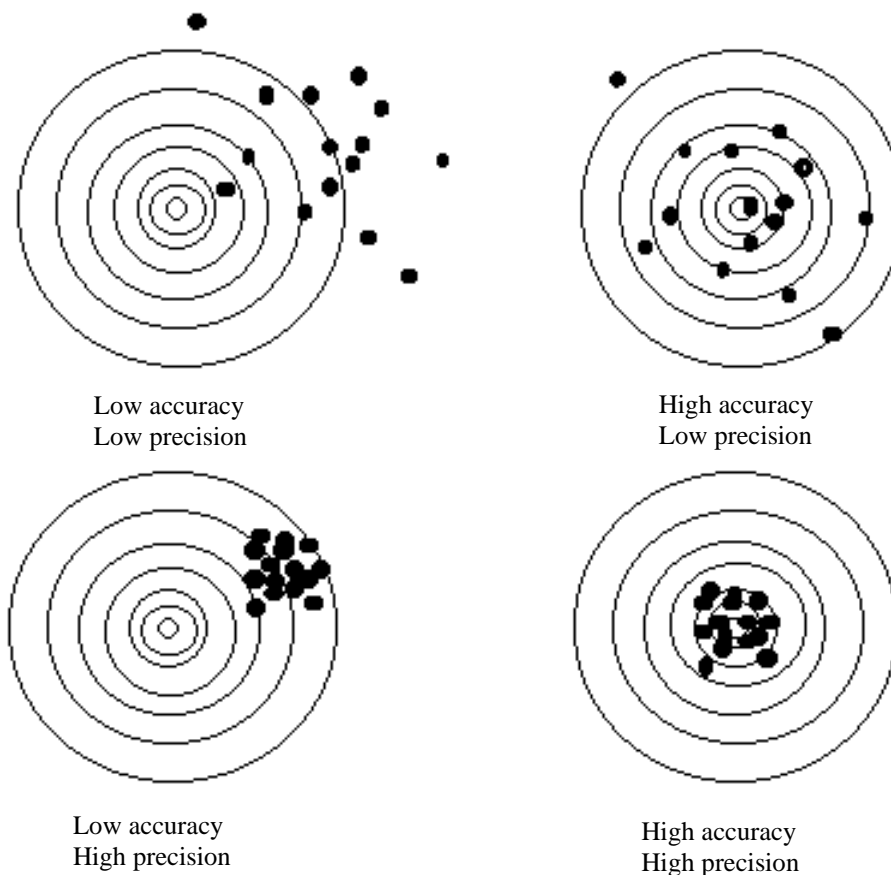


Fig. 1.1. Illustration of the precision and accuracy.

1.4 Type of errors

There several types of errors:

- 1) **systematic**, they have an origin, which can be determined and corrected
 - a. **instrumental errors**, for example weak battery, resistive electric contacts, etc.
 - b. **errors of the method**, they are caused by the non-ideal behavior of the reactants, e.g. complexing reaction too slow, contamination, instability or decomposition of the reagents, chemical interferences
 - c. **errors caused by the experimenter**, e.g., reading from the incorrect scale, apparatus not correctly adjusted, etc.
- 2) **random errors**, indeterminable, might be positive or negative, caused by then random fluctuations, noise, they do not have one cause. **Only the random errors can be studied using statistical methods.**

1.5 Distribution of errors

Population is the complete ensemble of data. It might be finite, e.g., number of habitants in the city, number of cells in a sample. In typical physical measurements it is an infinite number of results that could be obtained in an experiment. Of course, one cannot determine the whole

population and acquirement of a very large amount of data is impractical. In physical measurements one works with samples.

Sample is a limited number of repeated measurements which allows to conclude about the parameters of the population.

When one repeats the measurements several times in the same way it is possible to plot frequency of obtaining given results versus the obtained value. This is illustrated in Fig. 1.2 where number of measured values of the 10 ml pipet volume in the small intervals of 0.01 ml is plotted versus the volume found for the repetition of 30, 100, and 1000 times. When the number of measurements N goes to infinity, $N \rightarrow \infty$, the obtained distribution curve approaches the **Gauss or normal distribution curve**.

This normal distribution curve describes probability density function, P_G , that is the probability that the measured value is between x and $x + dx$. The normal (Gauss) distribution is fully described by three parameters: **the value x , true value μ , and standard deviation of the population σ** :

$$P_G(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (1.1)$$

An example of such continues curve for $P_G(x, 10.00, 0.01)$ that is $\mu = 10.00$ and $\sigma = 0.01$ is shown in Fig. 1.3.

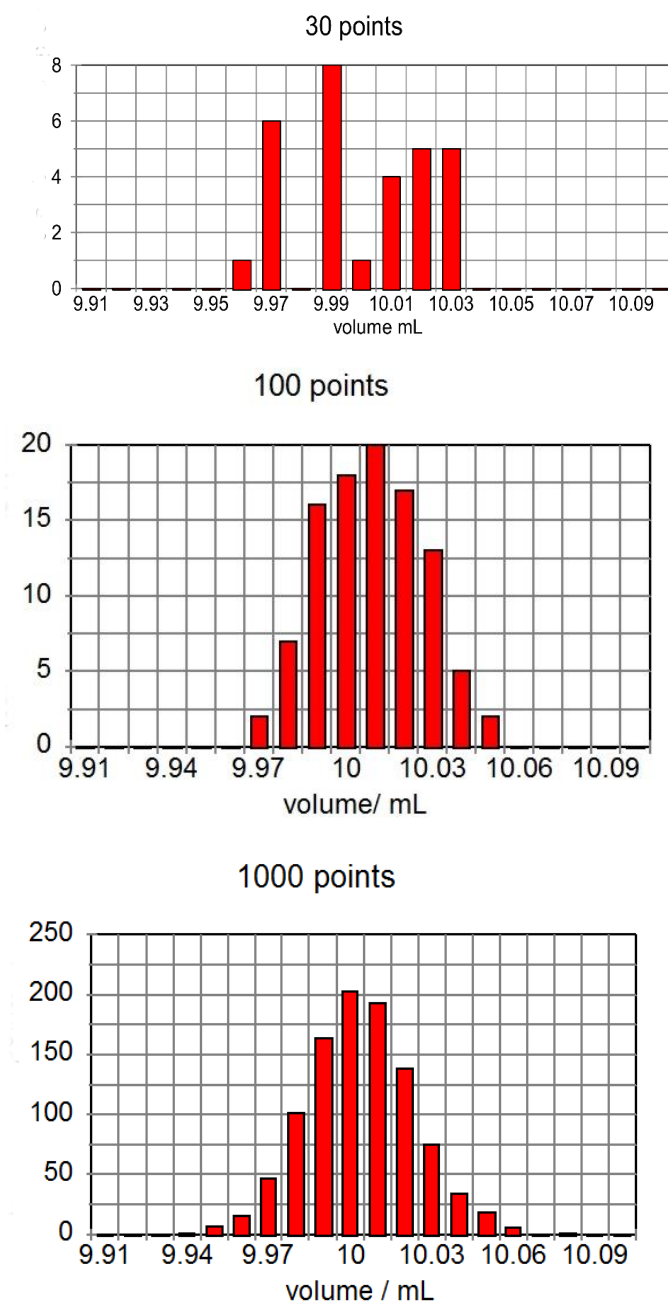


Fig. 1.2. Distribution of the measured values $P_G(x, 10.00, 0.01)$ of the volume of 10 ml pipet with standard deviation 0.01 for 30, 100, and 1000 measurements.

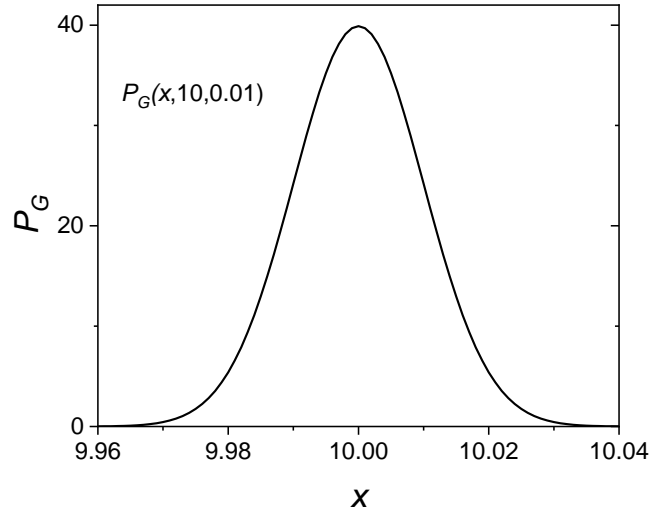


Fig. 1.3. Normal distribution curve for $\mu = 10$ and $\sigma = 0.01$.

Very often normalized distribution is presented for the **reduced parameter** z :

$$z = \frac{(x - \mu)}{\sigma} \quad (1.2)$$

Using this substitution, it is possible to transform $P_G(x, \mu, \sigma)$ into $P_G(z, 0, 1)$ with $\mu = 0$ and $\sigma = 1$.

The Gauss distribution function, Eq. (1.1), becomes:

$$P_G(z, 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad (1.3)$$

A plot of this function is shown in Fig. 1.4.

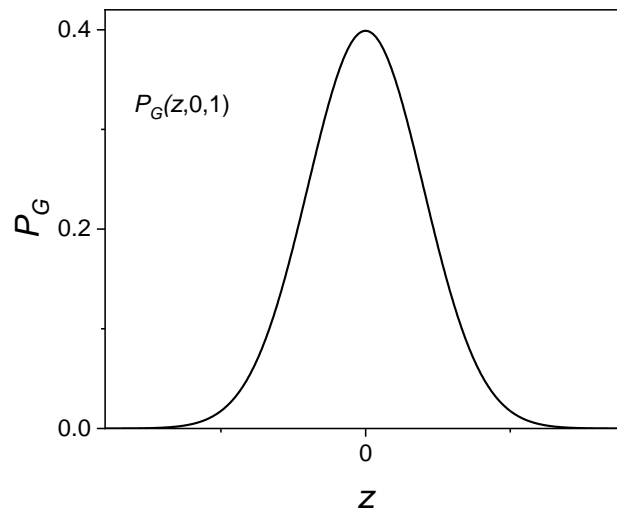


Fig. 1.4. Normalized Gaussian probability function, $P_G(z, 0, 1)$.

In Excel, $P_G(x, \mu, \sigma)$, is calculated using $\text{NORM.DIST}(x, \mu, \sigma, \text{FALSE})$ and the normalized Gauss probability function, $P_G(z, 0, 1)$, as $\text{NORM.DIST}(z, 0, 1, \text{FALSE})$ or $\text{NORM.S.DIST}(z, \text{FALSE})$. The logical values FALSE or TRUE are only used to choose the formula (two different formulas are used in NORM.DIST) to calculate the value, see also Eq. (1.9) ; with FALSE probability P_G is calculated while with TRUE its integral.

The true value is approximated by the **mean**:

$$\mu = \lim_{N \rightarrow \infty} \left(\frac{x_1 + x_2 + \dots + x_N}{N} \right) = \frac{\sum_{i=1}^N x_i}{N} = \bar{x} \quad (1.4)$$

It might be calculated using Excel function $\text{AVERAGE}(\text{cell1}:\text{cellN})$. It can be noticed that the sum of deviations from the mean is zero:

$$\sum_{i=1}^N (x_i - \bar{x}) = 0 \quad (1.5)$$

The **standard deviation of the population** is:

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}} \quad (1.6)$$

and the **variance**, σ^2 :

$$\sigma^2 = \frac{\sum_{i=1}^N (x_i - \mu)^2}{N} \quad (1.7)$$

The **variation coefficient**, CV , is simply the **relative standard deviation** expressed in %:

$$CV = \frac{\sigma}{\mu} 100\% \quad (1.8)$$

In Excel standard deviation of the population it can be calculated using function $\text{STDEV.P}(\text{cell1}:\text{cellN})$.

An example of the distribution of the deviations from the mean is presented in Fig. 1.5.

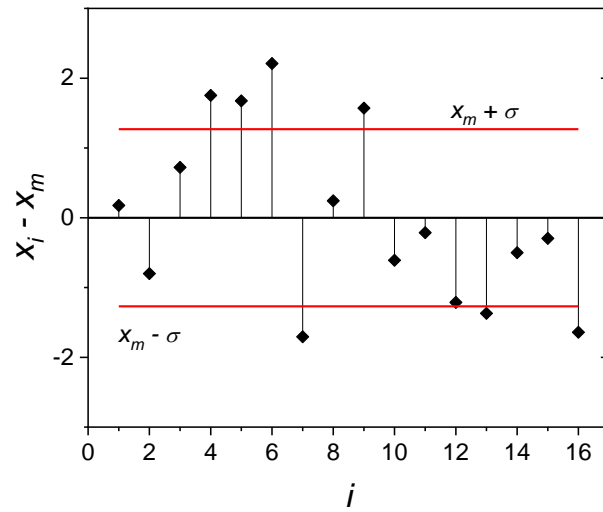


Fig. 1.5. Distribution of the deviations from the mean of 16 points distributed normally. The sum of the deviations is zero.

Gaussian curves obtained for different σ and μ are presented in Fig. 1.6 - 1.8.

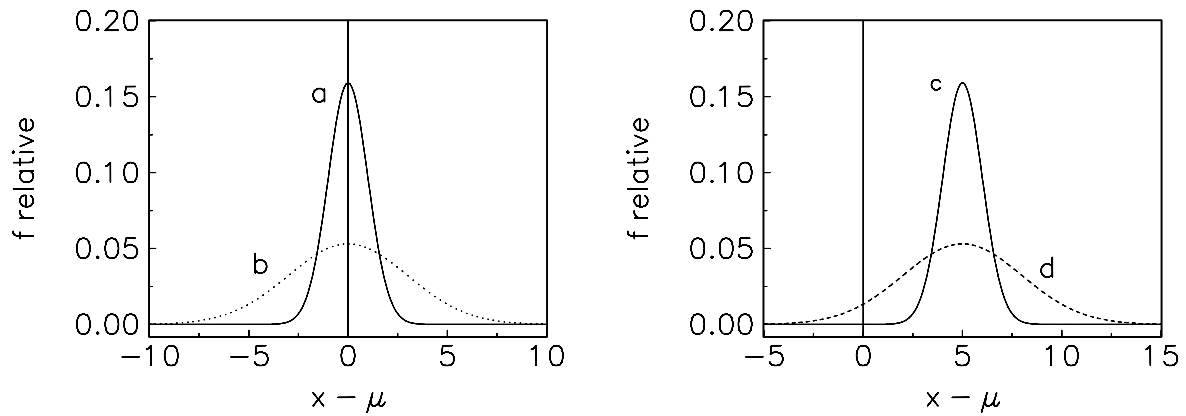


Fig. 1.6. Gaussian curves for: a) good precision and accuracy, b) bad precision and good accuracy, c) good precision and bad accuracy, d) bad precision and accuracy.

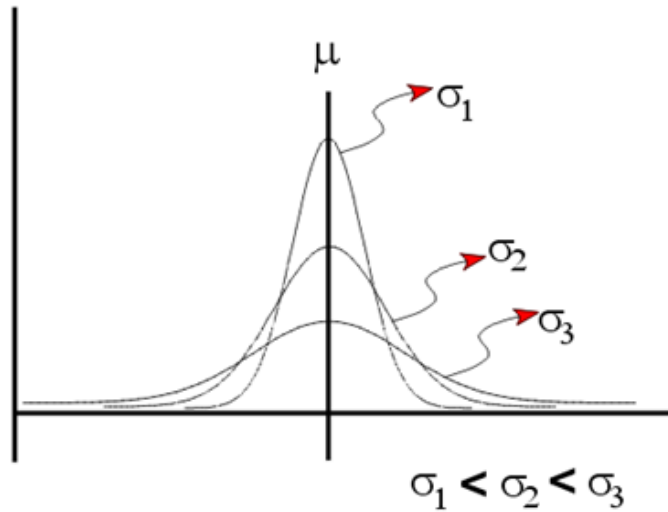


Fig. 1.7. Gaussian curves for different values of σ and the same μ .

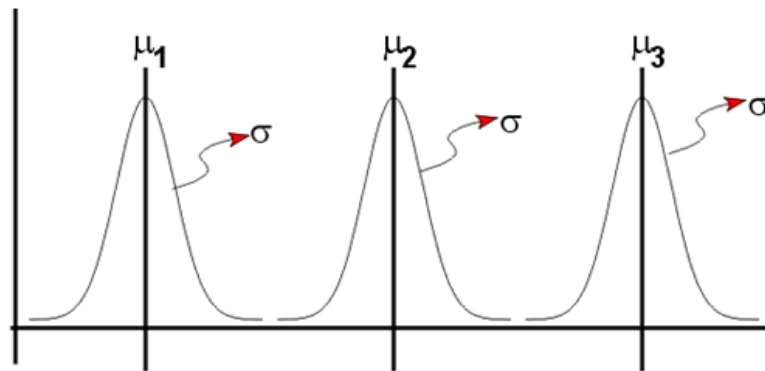


Fig. 1.8. Gaussian curves for different μ and the same σ ,

1.6 Integration of the Gauss curve

The area of the Gaussian curve between $\mu - \sigma$ and $\mu + \sigma$ or for the normalized curve between -1 and 1 is 68.3% of the total surface area. This is illustrated in Fig. 1.9.

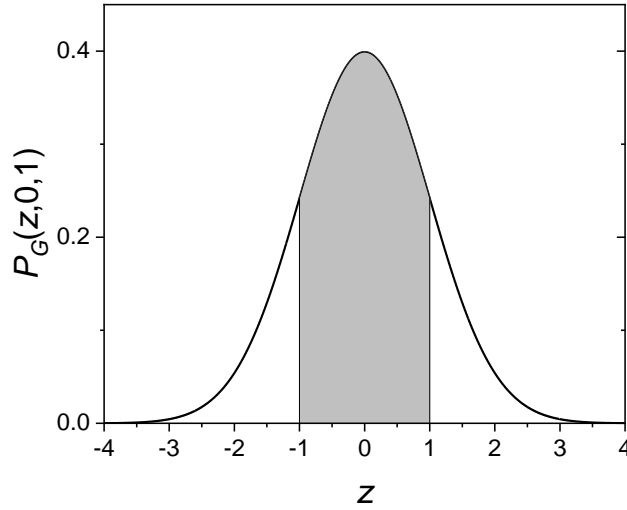


Fig. 1.9. Surface area under the Gauss curve. The area between -1 and +1 is 68.3%.

For other ranges see Table 1.1 below.

Table 1.1. Surface area under the Gauss curve.

Non normalized	Normalized	Area
$\mu \pm \sigma$	-1... +1	68.3%
$\mu \pm 2\sigma$	-2... +2	95.5%
$\mu \pm 3\sigma$	-3... +3	99.7%
$\mu \pm 1.96\sigma$	-1.96...+1.96	95%

The integral under the Gauss curve is calculated using NORM.DIST function with the logical (cumulative) value TRUE:

$$\text{NORM.DIST}(x, \mu, \sigma, \text{TRUE}) = \int_{-\infty}^x P_G(x, \mu, \sigma) dx = \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \quad (1.9)$$

or for the normalized curve:

$$\text{NORM.S.DIST}(z, \text{TRUE}) = \int_{-\infty}^z P_G(z, 0, 1) dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{z^2}{2}} dz \quad (1.10)$$

If one needs to determine the area under -1 and +1 the following formula should be used:

$$\begin{aligned} \int_{-1}^1 P_G(z, 0, 1) dz &= \frac{1}{\sqrt{2\pi}} \int_{-1}^1 e^{-\frac{z^2}{2}} dz \\ &= \text{NORM.S.DIST}(1, \text{TRUE}) - \text{NORM.S.DIST}(-1, \text{TRUE}) \\ &= \text{NORM.DIST}(1, 0, 1, \text{TRUE}) - \text{NORM.DIST}(-1, 0, 1, \text{TRUE}) \end{aligned} \quad (1.11)$$

These integrals used in Excel are illustrated in Fig. 1.10.

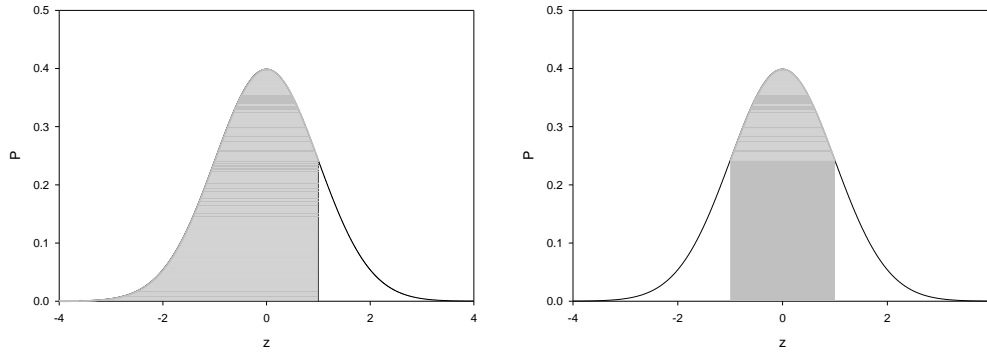


Fig. 1.10. The grey area under Gauss curve represents the integral form $-\infty$ to $z = 1$ and the integral form $z = -1$ to $z = 1$.

It should be stressed that integration of the non-normalized and normalized Gaussian curves in the same range gives the same results. For example, using data form Fig. 1.3 and 1.4 that is $\mu = 10$ and $\sigma = 0.01$, integrations from $\mu - \sigma$ to $\mu + \sigma$ gives:

$$\begin{aligned} \int_{x-\sigma}^{x+\sigma} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx &= \int_{9.99}^{10.01} \frac{1}{0.01\sqrt{2\pi}} e^{-\frac{(x-10)^2}{2 \times 0.01^2}} dx \\ &= \text{NORM.DIST}(10.01, 10, 0.01, \text{TRUE}) - \text{NORM.DIST}(9.99, 10, 0.01, \text{TRUE}, \text{TRUE}) \quad (1.12) \\ &= 0.682689 \end{aligned}$$

and using equivalent reduced parameter z

$$\begin{aligned} \frac{1}{\sqrt{2\pi}} \int_{-1}^1 e^{-\frac{z^2}{2}} dz &= \text{NORM.S.DIST}(1, \text{TRUE}) - \text{NORM.S.DIST}(-1, \text{TRUE}) \quad (1.13) \\ &= 0.682689 \end{aligned}$$

To illustrate application of the normal distribution few examples will be shown below.

Example 1.1.

For the normal distribution with $\mu = 25$ and $\sigma = 5$ find:

a) probability $P(x \geq 20)$

We need to calculate integral of the Gauss curve form 20 to ∞ , Fig. 1.11:

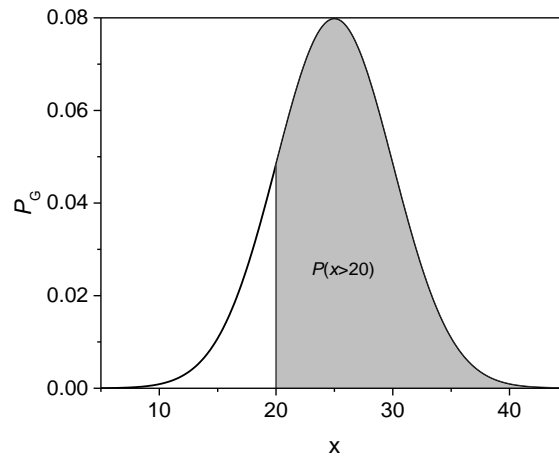


Fig. 1.11. Distribution of $P_G(x, 25, 5)$ and the calculation of the probability $P(x \geq 20)$.

$$\begin{aligned}
 P(x \geq 20) &= \int_{20}^{\infty} P_G(x, 25, 5) dx = 1 - \int_{-\infty}^{20} P_G(x, 25, 5) dx \\
 &= 1 - \text{NORM.DIST}(20, 25, 5, \text{TRUE}) \\
 &= 1 - 0.15866 = 0.84134
 \end{aligned} \tag{1.14}$$

b) $P(x < 40)$

Fig. 1.12. Distribution of $P_G(x, 25, 5)$ and the calculation of the probability $P(x < 40)$.

$$\begin{aligned}
 P(x < 40) &= \int_{-\infty}^{40} P_G(x, 25, 5) dx \\
 &= \text{NORM.DIST}(40, 25, 5, \text{TRUE}) = 0.99865
 \end{aligned} \tag{1.15}$$

c) $P(21 \leq x \leq 30)$

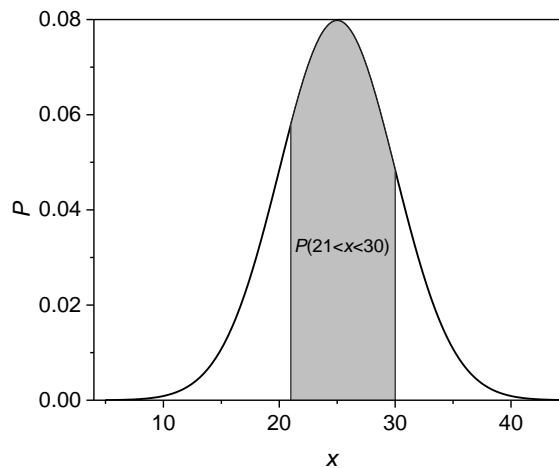


Fig. 1.13. Distribution of $P_G(x, 25, 5)$ and the calculation of the probability $P(21 \leq x \leq 30)$.

$$\begin{aligned}
P(21 \leq x \leq 30) &= \int_{-\infty}^{30} P_G(x, 25, 5) dx - \int_{-\infty}^{21} P_G(x, 25, 5) dx \\
&= \text{NORM.DIST}(30, 25, 5, \text{TRUE}) - \text{NORM.DIST}(21, 25, 5, \text{TRUE}) = 0.62949
\end{aligned} \tag{1.16}$$

See calculations in *Examples1.xlsx*, sheet *Ex. 1.1*.

Example 1.2.

The average of notes of one course measured during several years is $\mu = 65$ and $\sigma = 15$. Calculate:

- 1) % of students with the average $x \geq 85$
- 2) % of students with the average $x \leq 50$
- 3) % of students with the average $60 \leq x \leq 85$
- 4) what values of notes have lower 20% of students?

Re. 1.

$$\begin{aligned}
P(x \geq 85) &= \int_{85}^{\infty} P(x, 65, 15) dx = 1 - \int_{-\infty}^{85} P(x, 65, 15) dx \\
&= 1 - \text{NORM.DIST}(85, 65, 15, \text{TRUE}) = 1 - 0.9088 = 0.0912 = 9.12\%
\end{aligned} \tag{1.17}$$

Re. 2.

$$\begin{aligned}
P(x \leq 50) &= \int_{-\infty}^{50} P(x, 65, 15) dx \\
&= \text{NORM.DIST}(50, 65, 15, \text{TRUE}) = 0.159 = 15.9\%
\end{aligned} \tag{1.18}$$

Re. 3.

$$\begin{aligned}
P(60 \leq x \leq 85) &= \int_{-\infty}^{85} P_G(x, 65, 15) dx - \int_{-\infty}^{60} P_G(x, 65, 15) dx \\
&= \text{NORM.DIST}(85, 65, 15, \text{TRUE}) - \text{NORM.DIST}(60, 65, 15, \text{TRUE}) = 0.5393 = 53.9\%
\end{aligned} \tag{1.19}$$

Re. 4.

In this case one should find the value of x for which:

$$\int_{-\infty}^x P_G(x, 65, 15) dx = 0.20 \tag{1.20}$$

This is illustrated in Fig. 1.14 and for the normalized parameters in Fig. 1.15.

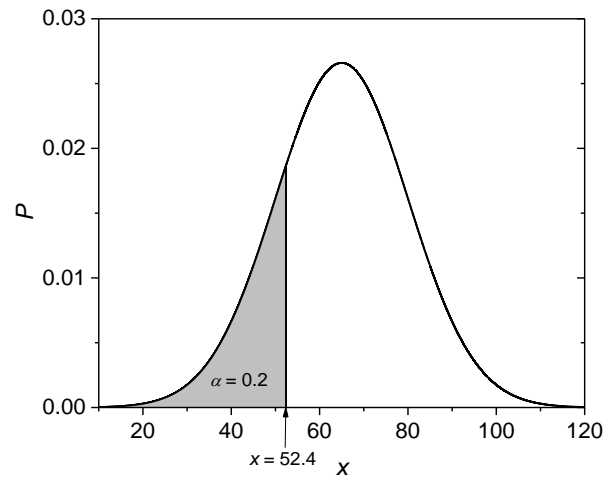


Fig. 1.14. Illustration of Example 1.2-4.

The shaded area is 0.2 of the total area under the curve.

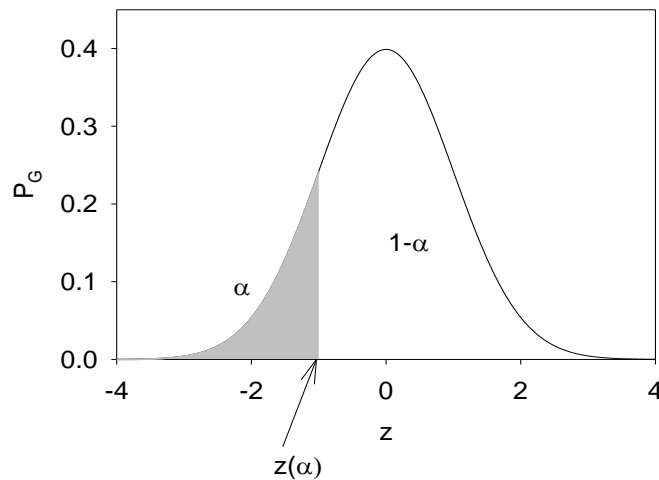


Fig. 1.15. Illustration of Example 1.2-4.

The shaded area is 0.2 of the total area. Graph is shown for the normalized distribution which can be obtained by normalization of the parameters, Eq. (1.2) and $\alpha = 0.2$.

This can be calculated using Excel function $\text{NORM.INV}(\alpha, \mu, \sigma)$:

$$\begin{aligned} x(\alpha) &= \text{NORM.INV}(\alpha, \mu, \sigma) \\ x(0.2) &= \text{NORM.INV}(0.2, 65, 15) = 52.4 \end{aligned} \quad (1.21)$$

There are 20% of students have results lower than $x = 52.4$.

There is a similar function for the normalized distribution:

$$z(\alpha) = \text{NORM.S.INV}(0.2) = -0.8416 \quad (1.22)$$

Using Eq. (1.2) one gets $x(0.2) = \mu + z \sigma = 65 - 0.8416 \times 15 = 52.4$. See calculations in *Examples1.xlsx*, sheet *Ex. 1.2*.

1.7 Standard deviation of the population and sample standard deviation

The main purpose of the statistics is to find **mean, standard deviation, and confidence intervals**. There are two methods used depending on the amount of data available.

- a) When large amount of data points is available, in practice when its number $N \geq 30$, one can estimate the true μ and **standard deviation of the population**, σ :

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}} \quad \text{where} \quad \mu = \lim_{x \rightarrow \infty} \frac{\sum_{i=1}^N x_i}{N} \quad (1.23)$$

- b) In general, $\bar{x} \neq \mu$ but $\bar{x} \rightarrow \mu$ when $N \rightarrow \infty$. When $N < 30$ one cannot determine σ . In such a case one can determine **sample standard deviation** that is standard deviation of a small sample of data points, s :

$$s = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N-1}} \quad \text{where} \quad \bar{x} = \frac{\sum_{i=1}^N x_i}{N} \quad (1.24)$$

The difference between σ and s is the term in the denominator indicating number of degrees of freedom. There are N points and the number of degrees of freedom should be N , but these points were used to determine \bar{x} , therefore number of degrees of freedom is $N-1$.

The average is calculated in Excel using AVERAGE, standard deviation of the population using STDEV.P and sample standard deviation using STDEV.S.

1.8 Standard deviation of the true value and of the mean

Standard deviations of mean are calculated using the following formula (which will be developed in the section on the error propagation in Example 2.4.):

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} \quad s_{\bar{x}} = \frac{s_x}{\sqrt{N}} \quad (1.25)$$

Example 1.3

Calculate mean, standard deviation and standard deviation of the mean of the data (100 points) in *Examples1.xlsx*, sheet *Ex 1.3*.

Calculation of these parameters is shown in *Ex. 1.3*. The results are displayed in in Fig. 1.16, see details in *Example 1.3* in *Examples1.xlsx*, sheet *Ex 1.3*. It can be noticed, that the average and standard deviations approach μ and σ while the standard deviation of the mean, $s_{\bar{x}}$, always decreases with increase of the number of points. However, $s_{\bar{x}}$, initially decreases rapidly and then very slowly.

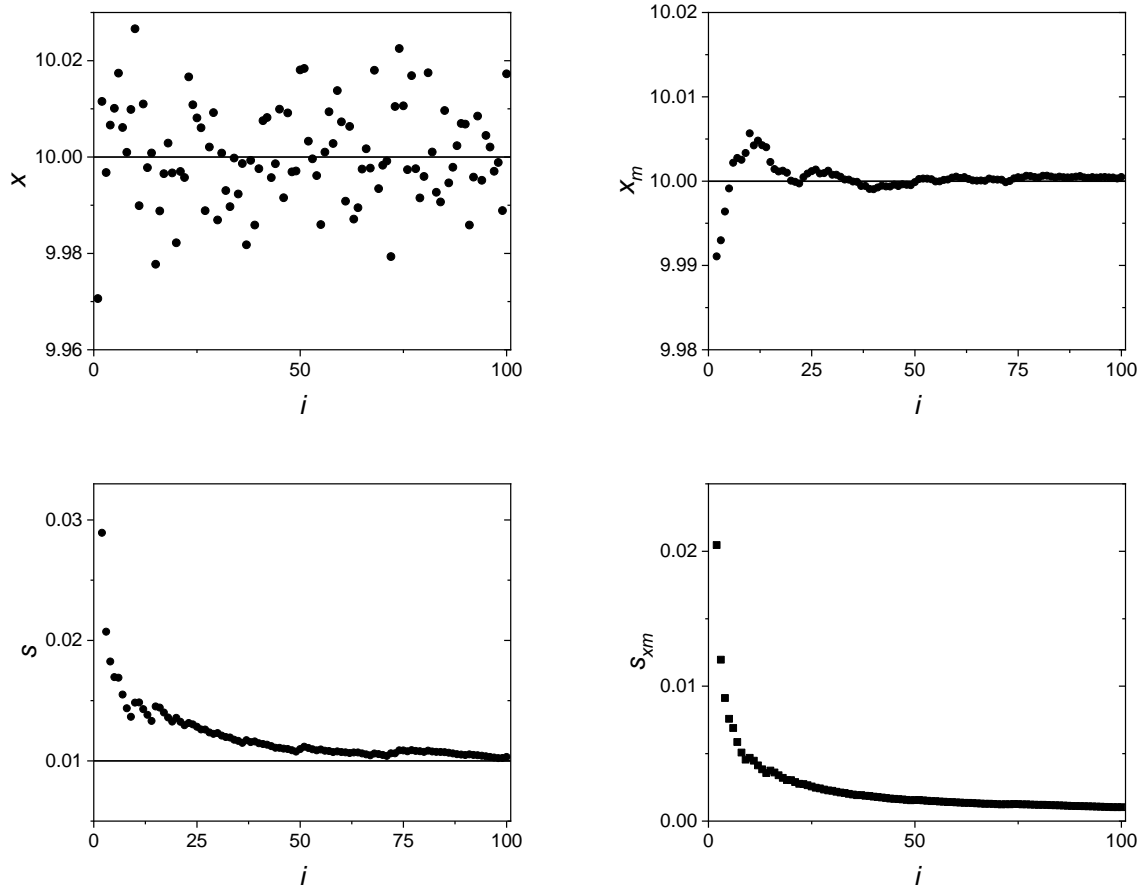


Fig. 1.16. Dependence of the random values x , average, x_m , standard deviation, s , and standard deviation of the average, $s_{\bar{x}} (=s_{xm})$, on number of points.

1.9 Confidence intervals

Using the statistical methods, it is possible to estimate **confidence intervals** around \bar{x} where the real value might be situated. One can determine an interval: $a \geq \bar{x} \geq b$ with certain probability. More precisely, one can say with the probability $1-\alpha$ that the real value is within some interval. The value of α is called **confidence level** but often this term is used for $(1-\alpha) \times 100\%$. There are two methods used depending on the fact if the standard deviation of the population is known.

1.10 Confidence intervals when σ is known

If one makes more than 30 measures, $s \rightarrow \sigma$. In such a case one can use normal (Gaussian) distribution to evaluate the confidence intervals:

$$IC(\mu) = \bar{x} \pm z(\alpha/2) \frac{\sigma}{\sqrt{N}} = \bar{x} \pm z(\alpha/2) \sigma_{\bar{x}} \quad (1.26)$$

This means that with the confidence of $1 - \alpha$ the real value is between the limits shown in Eq. (1.26) that is with the confidence α that it is outside. For example, for the confidence level of 95%,

that is for $\alpha = 0.05$, the value of $z(\alpha/2) = z(0.025) = -1.96$ and $z(1-\alpha/2) = z(0.975) = 1.96$ that is $|z(0.025)| = z(0.975)$ (two-tailed test, see Section 1.12). These zones are illustrated in Fig. 1.17.

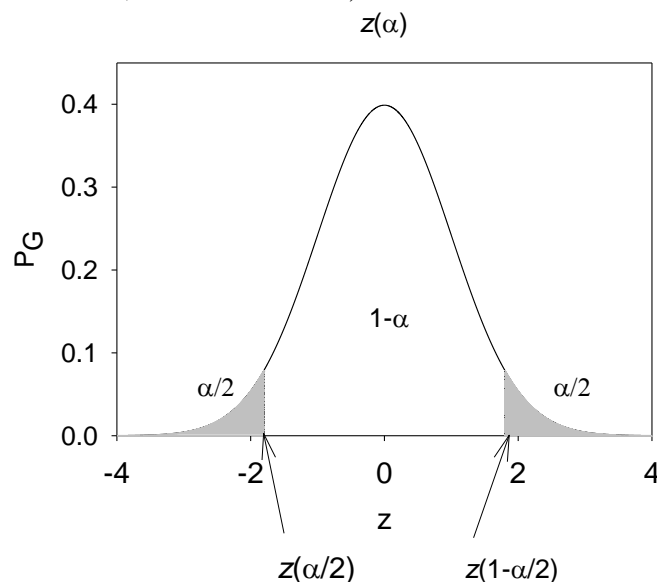


Fig. 1.17. Illustration of the zone $1-\alpha$ and two tails $\alpha/2$ each. The grey area constitutes α part under the Gauss curve.

Example 1.4.

Copper was determined by the atomic spectroscopy. For three measurements the average was $\bar{x} = 2.30$ ppm and the standard deviation determined earlier pooling large number of data was $\sigma = 0.20$ ppm. Calculate the confidence intervals for the probability 95% and 99%.

The value of $z(\alpha/2)$ are: $z(0.025) = -1.96$ and $z(0.005) = -2.58$, respectively and those $z(1-\alpha/2)$ are: $z(0.975) = 1.96$ and $z(0.995) = 2.58$. These values were calculated using NORM.S.INV(p). Therefore, the confidence intervals are:

$$95\% (\alpha = 0.05, \alpha/2 = 0.025) \text{ CI}(\mu) = \bar{x} \pm z \sigma_{\bar{x}} = 2.30 \pm 1.96 \times 0.20 / \sqrt{3} = 2.30 \pm 0.23$$

$$99\% (\alpha = 0.01, \alpha/2 = 0.005) \text{ CI}(\mu) = \bar{x} \pm z \sigma_{\bar{x}} = 2.30 \pm 2.58 \times 0.20 / \sqrt{3} = 2.30 \pm 0.30.$$

There is a probability of 5% that the real value is outside intervals 2.30 ± 0.23 and probability of 1% that the real value is outside the intervals 2.30 ± 0.30 . See calculations in *Examples1.xlsx*, sheet *Ex. 1.4*.

1.11 Confidence intervals when σ is not known

If σ is unknowns one must use s (sample standard deviation) to estimate the confidence interval. However, in this case one cannot use normal distribution. Because less data is known the result will depend on the number of points used and one has to use the **Student distribution function**. It gives larger confidence intervals than normal distribution. Only when $N \rightarrow \infty$, Student distribution becomes normal. The values of the Student distribution function are presented in Fig. 1.18, and Example 1.5.

Example 1.5.

Simulate Student distribution functions and their integrals for $df = 4$ and 9 degrees of liberty. Compare with normal distribution function.

The results are shown in *Examples1.xlsx*, sheet *Ex. 1.5*.

Student distribution depends on the number of degrees of freedom $df = N - 1$. In the calculation of the confidence intervals instead of $z(\alpha/2)$ one should use $t(\alpha'', df)$:

$$IC(\mu) = \bar{x} \pm t(\alpha'', df) \frac{s}{\sqrt{N}} = \bar{x} \pm t(\alpha'', df) s_{\bar{x}} \quad (1.27)$$

where $df = N - 1$ and symbol α'' indicates so called two-tailed test (see below). The values of the Student distribution function might be calculated using Excel function: T.DIST(t, df, FALSE). The values of $t(\alpha'', df)$ are calculated in Excel using T.INV.2T(α, df); this is so called two-tailed Student distribution (symbolized by $''$) which means that the surface area outside the central $1-\alpha$ is α and there are two tails, $\alpha/2$ each, similarly to Fig. 1.17.

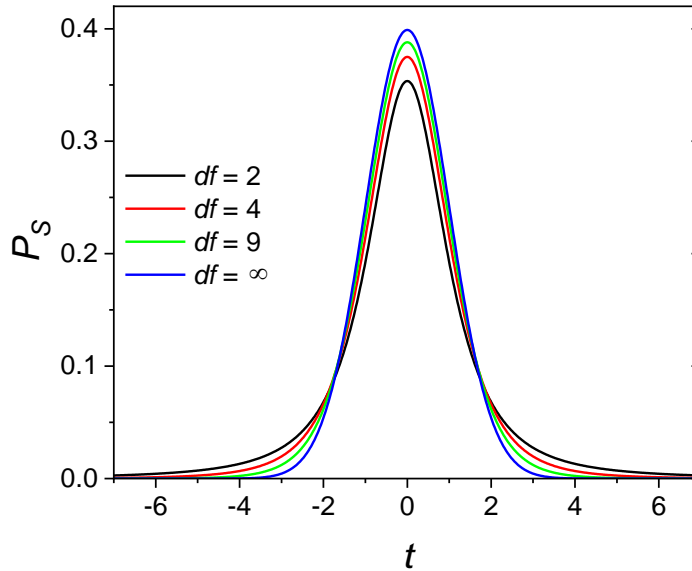


Fig. 1.18. Student probability, P_s , distribution for the number of degrees of freedom $df = 2, 4, 9$, and ∞ , when it becomes normal.

For example, using probability of 95%, $|z(0.05/2)| = z(1-0.05/2) = 1.96$ and for three measurements $t(0.05'', 2) = 4.30$, which means that the confidence intervals are over two times larger than for the normal distribution for $df = 2$.

1.12 Two-tailed and one-tailed tests

In the above tests we were determining the confidence intervals for the normal: $\bar{x} \pm z(\alpha/2) \sigma_{\bar{x}}$ and for the Student $\bar{x} \pm t(\alpha'', df) s_{\bar{x}}$ distributions for which the probability of finding our values was 0.95 or 95%. This means that the probability of finding results outside this interval is 0.05 or

5%. This case is illustrated in Fig. 1.19, left, for the Student distribution function. Confidence interval values, Eq. (1.27), x_{CI} are:

$$\frac{x_{CI} - \bar{x}}{s_{\bar{x}}} = \pm t(\alpha'', df)$$

$$-t(\alpha'', df) \leq \frac{x_{CI} - \bar{x}}{s_{\bar{x}}} \leq +t(\alpha'', df)$$
(1.28)

This case is called **two-tailed** t -test shown in Fig. 1.19, left. There are two tails containing each 0.025 of the total surface area, therefore, the total surface area of two tails is 0.05 and the part inside the confidence intervals is 0.95. Index '' indicates that this is a **two-tailed test**. It might be calculated using Excel function: T.INV.2T(α, df) or in this particular case T.INV.2T(0.05,4).

Another case is so called **one-tailed** test. It corresponds to the condition $x_{CI} \leq \bar{x} + t(\alpha', df)s_{\bar{x}}$ illustrated in Fig. 1.19, right. Here we are interesting only in the data below the confidence interval:

$$\frac{x_{CI} - \bar{x}}{s_{\bar{x}}} \leq +t(\alpha', df)$$
(1.29)

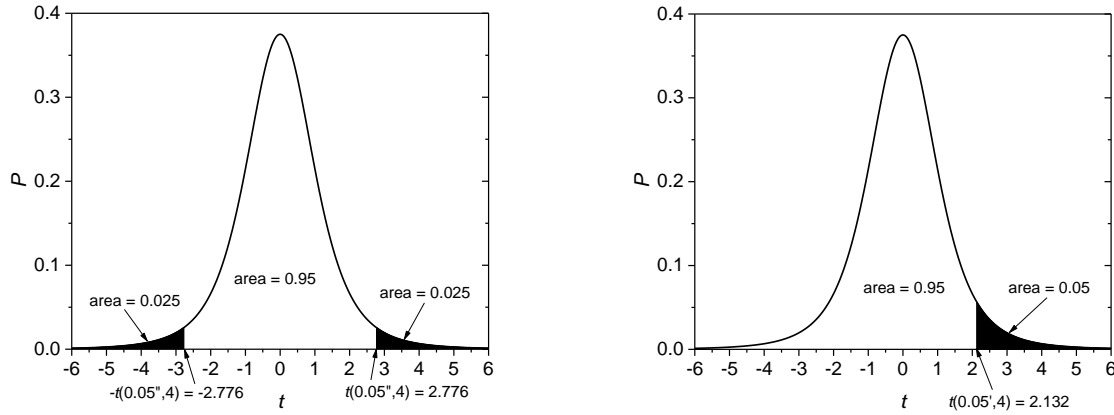


Fig. 1.19. Student t density of probability distribution function for 4 degrees of freedom showing two-tailed (left) and one-tailed (right) tests for the confidence level $\alpha = 0.05$. The black areas of two tails (left) have the probability of 0.05 and the central part of 0.95. The corresponding value of $t(0.05'', 4)$ is calculated using Excel function T.INV.2T(0.05,4). The black area of the one-tailed distribution graph is 0.05 and that of the rest is 0.95, calculated as $t(0.05', 4)$ i.e. T.INV(0.95,4) = |T.INV(0.05,4)|.

The probability that

$$\frac{x_{CI} - \bar{x}}{s_{\bar{x}}} \geq +t(\alpha', df)$$
(1.30)

is α . The sign ' means that the **one-tailed distribution** is used. The corresponding value of $t(\alpha', df)$ is calculated using Excel function T.INV(α, df) and in the case of $t(0.05', 4)$ as T.INV(0.95,4). These calculations are shown in Example 1.5 for $df = 4$.

Example 1.6.

The analysis of alcohol in blood gave the following results: 0.084%, 0.089% and 0.079%. Calculate the confidence intervals for the confidence limit of 95%.

- using only these data
- assuming that σ is known, $\sigma = 0.006\%$.

Re. a)

Student statistics is used, σ is unknown

$$\bar{x} = (0.084 + 0.089 + 0.079)/3 = 0.084$$

Sample standard deviation:

$$s = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N-1}} = 0.005$$

$$df = N - 1 = 3 - 1 = 2$$

$$t(0.05, 2) = 4.30$$

$$CI = 0.084 \pm 4.30 * 0.005 / \sqrt{3} = 0.084 \pm 0.012.$$

The results should be presented in the following form:

$$\bar{x} = 0.084, s = 0.005, s_{\bar{x}} = 0.003 \text{ for } N = 3 \text{ or } df = 2$$

With the probability of 95% the true value is between:

$$0.072 \leq \bar{x} \leq 0.096 \text{ or } \bar{x} = 0.084 \pm 0.012.$$

It should be stressed that \pm is **reserved for confidence intervals** and must not be used for the standard deviations!

Re. b)

σ is known, normal distribution is used

$$|z(0.05/2)| = z(1-0.05/2) = 1.96$$

$$IC = 0.084 \pm 1.96 * 0.006 / \sqrt{3} = 0.084 \pm 0.007$$

$$\bar{x} = 0.084, \sigma = 0.006, \sigma_{\bar{x}} = 0.003 \text{ for } N = 3 \text{ or } df = 2$$

With the probability of 95% the true value is between:

$$0.077 \leq \bar{x} \leq 0.091 \text{ or } \bar{x} = 0.084 \pm 0.007$$

It should be noticed that using known value of σ the CI are smaller. See calculations in *Examples1.xlsx*, sheet *Ex. 1.6*.

Example 1.7.

Measurements of the iron concentration in the sample carried out using atomic absorption spectroscopy gave the following results: 3.2, 2.9, 3.0, 3.3, 3.1 ppm. Calculate the mean standard deviation, standard deviation of the mean, and confidence intervals for the confidence level of 95% and 99%.

- 95%

The analysis might be carried out as in Example 1.6b but it is easier to use Descriptive Statistics in Data Analysis in Excel (for the first use it must be installed: File, Options, Add-Ins, Excel Add-

Ins, Analysis ToolPack; the necessary files are already on the disk). Below there is the obtained result. Explanations were added on the right.

Column1

Mean	3.1	\bar{x}
Standard Error	0.071	$s_{\bar{x}}$
Median	3.1	value in the middle
Mode	#N/A	
Standard Deviation	0.16	s
Sample Variance	0.025	s^2
Kurtosis	-1.2	
Skewness	8.7E-15	
Range	0.4	$x_{\max} - x_{\min}$
Minimum	2.9	x_{\min}
Maximum	3.3	x_{\max}
Sum	15.5	$\sum_{i=1}^N x_i$
Count	5	N
Confidence Level(95.0%)	0.196324316	$t(\alpha'', df) s_{\bar{x}}$

To present the results they should be rounded:

$$\bar{x} = 3.10, s_x = 0.16, s_{\bar{x}} = 0.07, N = 5 \text{ (} df = 4 \text{)}$$

With the probability of 95% the true value is between:

$$2.90 \leq \bar{x} \leq 3.30$$

(that is $3.10 - 0.20$ and $3.1 + 0.20$) or $x = 3.10 \pm 0.20$

b) 99%

Column1

Mean	3.1
Standard Error	0.071
Median	3.1
Mode	#N/A
Standard Deviation	0.16
Sample Variance	0.025
Kurtosis	-1.2
Skewness	8.7E-15
Range	0.4
Minimum	2.9
Maximum	3.3
Sum	15.5
Count	5
Confidence Level(99.0%)	0.33

The only difference in comparison with the earlier example is the value of “Confidence Level(99%)” which is in fact the value which should be subtracted and added to the mean to obtain CI.

$$\bar{x} = 3.10, s_x = 0.16, s_{\bar{x}} = 0.07, N = 5 (df = 4)$$

With the probability of 99% the true value is between:

$$2.77 \leq \bar{x} \leq 3.43 \text{ or } x = 3.10 \pm 0.33.$$

See calculations in *Examples1.xlsx*, sheet *Ex. 1.7*.

1.13 Pooling data

Very often many measurements are carried out in a similar way, but the number of repetitions is small. Nevertheless, all these results might be applied to determine better estimation of the sample standard deviation or even standard deviation of the population. The standard deviation of the pooled data is calculated using:

$$s_{\text{pooled}} = \sqrt{\frac{\sum (x_i - \bar{x})^2}{N - n}} \quad (1.31)$$

where n is the number of samples and $N = \sum N_i$ is the total number of measurements.

Example 1.8.

Let us suppose that the measurements were carried out on 8 samples, each sample was different:

Sample #	N_i	\bar{x}	$\sum (x_i - \bar{x})^2$
1	5	1.673	0.029
2	4	1.015	0.0115
3	5	3.24	0.0242
4	6	2.018	0.0611
5	4	0.57	0.0114
6	5	2.482	0.0658
7	4	1.13	0.0175
8	7	1.27	0.0319
$N = 40$			$\sum = 0.2524$

The value $df = N - n = 40 - 8 = 32$ is the number of degrees of freedom (total number of points minus number of means). In the above case $s_{\text{pooled}} = 0.089$. Because number of degrees of freedom is larger than 30 one can consider $s_{\text{pooled}} = \sigma$. See calculations in *Examples1.xlsx*, sheet *Ex. 1.8*.

1.14 Weighted mean

When the different measures are characterized by different precision one should take it into account and use the weighted mean. Let us assume that for each measurement its standard deviation is known:

$$x_1 \quad s_{x_1}$$

$$x_2 \quad s_{x_2}$$

$$x_3 \quad s_{x_3}$$

...

$$x_N \quad s_{x_N}$$

The statistical weights, w_i , of each measurement is inversely proportional to the variance: $w_i = 1 / s_{x_i}^2$. The weighted mean is calculated as:

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i} \quad (1.32)$$

and its standard deviation is:

$$s_{\bar{x}}^2 = \frac{1}{\sum w_i} = \frac{1}{\sum \frac{1}{s_{x_i}^2}} \quad (1.33)$$

This will be illustrated in the following example.

Example 1.9.

In the case of the radioactive decay the standard deviation is equal to the square root of number of impulses measured, x_i (Poisson distribution):

$$s_{x_i} = \sqrt{x_i} \quad (1.34)$$

The following results were obtained during data acquisition in different times:

Time of data acquisition t_i / min	Number of impulses measured x_i
5	10255
20	41200
2	4084
10	20650

Calculate the sample specific activity i.e. impulses per minute.

The specific activity $r_i = x_i / t_i$ and its standard deviation: $s_{r_i} = \frac{s_{x_i}}{t_i} = \frac{\sqrt{x_i}}{t_i}$. The following results are obtained:

Activity $r_i = x_i / t_i$	Standard deviation $s_i = \sqrt{x_i} / t_i$	w_i	$w_i r_i$	$w_i (r_i - \bar{r})^2$
2051	20.25	0.002438	5	0.161742
2060	10.15	0.009709	20	0.007092
2042	31.95	0.0009794	2	0.287916
2065	14.37	0.004843	10	0.165991
sum		0.01797	37	0.62274

Using Eqns. (1.32) and (1.33) the following results are obtained:

$\bar{r} = 2059.1$ pulses/min, $s_{\bar{x}} = 7.5$ pulses/min.

These results are in *Examples1.xlsx*, sheet *Ex. 1.9*.

In certain cases standard deviation of the mean seems to be too small with respect to the standard deviations of the samples.^{7,20} This can arrive when the distribution of the measured data is not normal. This hypothesis can be determined using χ^2 (chi-square) test, defined as:

$$\chi^2 = \sum_{i=1}^N \frac{(x_i - \bar{x})^2}{s_{x_i}^2} = \sum_{i=1}^N w_i (x_i - \bar{x})^2 \quad (1.35)$$

This experimental function should follow $\chi^2(\alpha, k)$ statistics. It can be determined in Excel using CHISQ.INV.RT(α, k). In the above example is: $\chi_{\text{exp}}^2 = 0.6227$. The value of $\chi^2(0.05, 3) = 7.81$.

Because the experimental value of $\chi_{\text{exp}}^2 < \chi^2(0.05, 3)$ we can accept the obtained result (see later in this test in the verification of the statistical hypotheses). χ^2 distribution function will be shown in Section 4.10. See calculations in *Examples1.xlsx*, sheet *Ex. 1.9*.

Let us consider another example.

Example 1.10.

Five measurements were carried out and the following results were obtained:

x_i	s_{x_i}
1.4	0.2
0.9	0.15
3.0	0.3
1.8	0.2
2.5	0.25

Calculate the mean and the standard deviation.

The results of calculations are presented below.

Table 1.2. Calculations of the weighted mean.

x_i	s_{x_i}	w_i	$w_i x_i$	$w_i (x_i - \bar{x})^2$
1.4	0.2	25	35	0.907195
0.9	0.15	44.44444	40	21.19029
3.0	0.3	11.11111	33.33333	22.07454
1.8	0.2	25	45	1.097323
2.5	0.25	16	40	13.23523
sum=		121.5556	193.3333	$58.50457 = \chi_{\text{exp}}^2$

$$\bar{x} = 1.59$$

$$\chi^2(0.05, 4) = 9.488$$

$$s_{\bar{x}} = 0.09$$

It can be noticed that the experimental value of $\chi_{\text{exp}}^2 = 58.50457$ is much greater than $\chi^2(0.05, 4) = 9.488$. This might indicate that the standard deviations of the means are too small and the error distribution is not normal. In such a case it was suggested^{7,20} that the standard deviation should be multiplied by a factor $\sqrt{\chi_{\text{exp}}^2 / (N - 1)}$:

$$s_{\bar{x}, \text{corr}} = s_{\bar{x}} \sqrt{\frac{\chi_{\text{exp}}^2}{N - 1}} \quad (1.36)$$

In the case above:

$$s_{\bar{x}, \text{corr}} = 0.091 \sqrt{\frac{58.505}{4}} = 0.35 \quad (1.37)$$

See calculations in *Examples1.xlsx*, sheet *Ex. 1.10*.

2 Propagation of errors

In physico-chemical measurements it is necessary to calculate function of several (n) parameters p_i , $z(p_1, p_2, p_3, \dots, p_n)$ where each parameter is determined with its standard deviation. To answer the question what is the standard deviation of z calculated using parameters p_i one should use the error propagation method.

2.1 Standard deviation of the calculated value

Let us assume a function z of n parameters:

$$z = f(p_1, p_2, \dots, p_n) \quad (2.1)$$

Its total derivative is:

$$dz = \left(\frac{\partial f}{\partial p_1} \right)_{p_2, \dots, p_n} dp_1 + \left(\frac{\partial f}{\partial p_2} \right)_{p_1, p_3, \dots, p_n} dp_2 + \left(\frac{\partial f}{\partial p_n} \right)_{p_1, \dots, p_{n-1}} dp_n \quad (2.2)$$

Let us assume that the deviations dx_i are very small, that is $|\Delta p_i| \ll |p_i|$ where:

$$dp_i = p_i - \mu_i \quad (2.3)$$

and

$$dz = z(p_1 + \Delta p_1, p_2 + \Delta p_2, \dots, p_n + \Delta p_n) - z(p_1, p_2, \dots, p_n) \quad (2.4)$$

The square of dz is:

$$\begin{aligned} dz^2 &= \left(\frac{\partial f}{\partial p_1} \right)^2 dp_1^2 + \left(\frac{\partial f}{\partial p_2} \right)^2 dp_2^2 + \dots \\ &+ 2 \left(\frac{\partial f}{\partial p_1} \right) \left(\frac{\partial f}{\partial p_2} \right) dp_1 dp_2 + 2 \left(\frac{\partial f}{\partial p_1} \right) \left(\frac{\partial f}{\partial p_3} \right) dp_1 dp_3 + \dots \end{aligned} \quad (2.5)$$

The values dp_i^2 are always positive but the terms $dp_i dp_j$ might be positive or negative and in the summation they might cancel, therefore:

$$\sum_{i=1}^N (dz_i)^2 \approx \left(\frac{\partial f}{\partial x_1} \right)^2 \sum_{i=1}^N (dp_{1,i})^2 + \left(\frac{\partial f}{\partial x_2} \right)^2 \sum_{i=1}^N (dp_{2,i})^2 + \dots \quad (2.6)$$

and the sum of squares leads to standard deviation:

$$\frac{\sum dz^2}{N} = \frac{\sum (p_i - \mu_i)^2}{N} = \sigma_z^2 \quad (2.7)$$

This equation might be written for standard deviations of a population or of a sample (assuming that the parameters p_i are independent):

$$\begin{aligned} \sigma_z^2 &= \left(\frac{\partial f}{\partial p_1} \right)^2 \sigma_{p_1}^2 + \left(\frac{\partial f}{\partial p_2} \right)^2 \sigma_{p_2}^2 + \dots \\ s_z^2 &= \left(\frac{\partial f}{\partial p_1} \right)^2 s_{p_1}^2 + \left(\frac{\partial f}{\partial p_2} \right)^2 s_{p_2}^2 + \dots \end{aligned} \quad (2.8)$$

Eq. (2.8) allows us to calculate the standard deviation of z when the standard deviations of the parameters are known.

To determine the confidence interval of z it is necessary to estimate the effective number of degrees of freedom, ν_{eff} , for z , and the value of t . Let us suppose that each parameter p_i is determined N_i times with the number of degrees of freedom $df_i = N_i - 1$. Of course, the number of measurements might be different for each p_i . Then, the effective number of degrees of freedom for z is calculated using the following equation:^{21,22}

$$\frac{\left(\sum_{i=1}^n \left(\frac{\partial z}{\partial p_i} \right)^2 s_{p_i}^2 \right)^2}{df_{eff}} = \frac{s_z^4}{df_{eff}} = \sum_{i=1}^n \frac{\left(\frac{\partial z}{\partial p_i} \right)^4 s_{p_i}^4}{df_i} \quad (2.9)$$

$$df_{eff} = \frac{s_z^4}{\sum_{i=1}^n \frac{\left(\frac{\partial z}{\partial p_i} \right)^4 s_{p_i}^4}{df_i}}$$

and the confidence interval is calculated as $\pm t(\alpha, df_{eff}) s_z$. It should be noticed that $df_{eff} \leq \sum df_i$.

More detailed definition of the error propagation will be given later in Eqs. (3.134)-(3.135).

2.2 Maximal error

When standard deviation is not known because the error analysis was not performed it is possible to estimate the maximal error which could be found when there is no cancelation of the terms. It can be found, assuming $|\Delta p_i| \ll |p_i|$, Eq. (2.10):

$$dz = \left(\frac{\partial f}{\partial p_1} \right)_{p_2, \dots, p_n} dp_1 + \left(\frac{\partial f}{\partial p_2} \right)_{p_1, p_3, \dots, p_n} dp_2 + \dots + \left(\frac{\partial f}{\partial p_n} \right)_{p_1, \dots, p_{n-1}} dp_n \quad (2.10)$$

becomes:

$$|dz| = \left| \left(\frac{\partial f}{\partial p_1} \right)_{p_2, \dots, p_n} \Delta p_1 \right| + \left| \left(\frac{\partial f}{\partial p_2} \right)_{p_1, p_3, \dots, p_n} \Delta p_2 \right| + \dots + \left| \left(\frac{\partial f}{\partial p_n} \right)_{p_1, \dots, p_{n-1}} \Delta p_n \right| \quad (2.11)$$

or:

$$\Delta z = \sum_{i=1}^n \left| f'_{p_i}(p_1, p_2, \dots, p_n) \Delta x_i \right| \quad (2.12)$$

where Δz represents the maximal error without the compensation of the random deviations. This is the worst-case scenario but it allows for a quick estimation of possible errors. This will be illustrated in Example 2.1.

Example 2.1.

Calculate the standard deviation and confidence interval of the volume of the box having dimensions a , b , and c using the following data:

$a = 5.5$ $b = 3.6$ $c = 1.9$ $s_a = s_b = s_c = 0.1$ (all in mm). The measurements of each parameter were repeated 4 times ($df_i = 4 - 1 = 3$).

$$V = abc = 37.62$$

$$s_V^2 = (f'_a \cdot s_a)^2 + (f'_b \cdot s_b)^2 + (f'_c \cdot s_c)^2 = (b \cdot c \cdot s_a)^2 + (a \cdot c \cdot s_b)^2 + (a \cdot b \cdot s_c)^2 = 5.48$$

$$s_V = 2.3 \quad (2.13)$$

$$df_{eff} = \frac{s_V^4}{\frac{\left(\frac{\partial V}{\partial a}\right)^4 s_a^4}{3} + \frac{\left(\frac{\partial V}{\partial b}\right)^4 s_b^4}{3} + \frac{\left(\frac{\partial V}{\partial c}\right)^4 s_c^4}{3}} = 5.37 \approx 5 \quad (2.14)$$

and $t(0.05, 5) = 2.57$ and confidence interval is $s_V t(0.05, 5) = 6.0$.

The answer is: $V = 37.6$, $s_V = 2.3$, and $CI = 6.0$. It is important to note that the number of significant digits in the standard deviation is one or two and the calculated function must be rounded accordingly. See calculations in *Examples2.xlsx*, sheet *Ex. 2.1-2.2*.

Example 2.2.

Calculate the maximal error in the above example assuming: $\Delta a = \Delta b = \Delta c = 0.1$.

$$\Delta V = |f'_a \Delta a| + |f'_b \Delta b| + |f'_c \Delta c| = 3.7 \quad (2.15)$$

The answer $V = 37.6$, $\Delta V = 3.7$ (or $V = 38$, $\Delta V = 4$). See calculations in *Examples2.xlsx*, sheet *Ex. 2.1-2.2*.

Example 2.3.

Calculate the standard deviation of the sum:

$$y = a + b$$

$$s_y^2 = \left(\frac{\partial y}{\partial a}\right)^2 s_a^2 + \left(\frac{\partial y}{\partial b}\right)^2 s_b^2 \quad (2.16)$$

$$s_y^2 = s_a^2 + s_b^2$$

The variance of sum equals sum of variances.

Example 2.4.

Calculate standard deviation of the mean assuming that the standard deviation of each measurement is the same, s_x .

$$\begin{aligned}
\mu &= \frac{\sum_{i=1}^N x_i}{N} = \frac{1}{N} (x_1 + x_2 + \dots + x_N) \\
s_\mu^2 &= \sum_{i=1}^N \left(\frac{\partial \mu}{\partial x_i} \right)^2 s_x^2 = \left(\frac{1}{N} \right)^2 s_x^2 + \left(\frac{1}{N} \right)^2 s_x^2 + \dots + \left(\frac{1}{N} \right)^2 s_x^2 \\
&= \frac{1}{N^2} s_x^2 + \frac{1}{N^2} s_x^2 + \dots + \frac{1}{N^2} s_x^2 = \frac{N}{N^2} s_x^2 = \frac{s_x^2}{N}
\end{aligned} \tag{2.17}$$

$$s_\mu = \frac{s_x}{\sqrt{N}}$$

Standard deviation of the mean is the standard deviation of a single measurement divided by the square root of the number of points, see Eq. **(1.25)**.

Example 2.5.

Calculate the standard deviation of the product: $y = ab$:

$$\begin{aligned}
s_y^2 &= b^2 s_a^2 + a^2 s_b^2 \quad | : y^2 = a^2 b^2 \\
\frac{s_y^2}{y^2} &= \frac{s_a^2}{a^2} + \frac{s_b^2}{b^2}
\end{aligned} \tag{2.18}$$

$$s_y = y \sqrt{\left(\frac{s_a}{a} \right)^2 + \left(\frac{s_b}{b} \right)^2}$$

In the multiplication (or division) the relative variances are added.

Example 2.6

$$\begin{aligned}
 y &= e^a \\
 s_y^2 &= (e^a)^2 s_a^2 \\
 \frac{s_y}{y} &= s_a
 \end{aligned} \tag{2.19}$$

Example 2.7.

$$\begin{aligned}
 y &= a^n \text{ where } n \text{ is a constant} \\
 s_y^2 &= (na^{n-1} s_a)^2 \\
 \frac{s_y}{y} &= \frac{na^{n-1}}{a^n} s_a = \frac{n}{a} s_a
 \end{aligned} \tag{2.20}$$

Example 2.8.

Calculate volume, standard deviation, and confidence interval of a cylinder characterized by the diameter D and the height h : $D = 7.76$ mm, $s_D = 0.02$ mm; $h = 62.33$ mm, $s_h = 0.02$ mm. The individual measurements of D and h were repeated six times.

$$\begin{aligned}
 V &= \frac{\pi D^2 h}{4} = 2947.88 \text{ mm}^3 \\
 s_V &= \sqrt{\left(\frac{\partial V}{\partial D}\right)^2 s_D^2 + \left(\frac{\partial V}{\partial h}\right)^2 s_h^2} \\
 \frac{\partial V}{\partial D} &= \frac{\pi D h}{2}, \quad \frac{\partial V}{\partial h} = \frac{\pi D^2}{4} \\
 s_V &= \sqrt{(759.8)^2 (0.02)^2 + (47.29)^2 (0.02)^2} = \\
 &= \sqrt{230.9 + 0.849} = 15.2 \text{ mm}^3
 \end{aligned} \tag{2.21}$$

The value of df_{eff} calculated using Eq. (2.9) is $5.04 \approx 5$, $t(0.05, 5) = 2.57$ and $df_{eff} t(0.05, 5) = 39$. The final results are: $V = 2948 \text{ mm}^3$, $s_V = 15 \text{ mm}^3$, $CI = 39$. It is evident that the largest contribution to the standard deviation comes from the measurements of the diameter. See calculations in *Examples2.xlsx*, sheet *Ex. 2.8*.

Example 2.9.

Few more examples.

a)

$$\begin{aligned}
y &= \frac{ab}{c} \\
s_y^2 &= \left(\frac{b}{c}\right)^2 s_a^2 + \left(\frac{a}{c}\right)^2 s_b^2 + \left(-\frac{ab}{c^2}\right)^2 s_c^2 \quad \Big| : y^2 \\
s_y &= y \sqrt{\frac{s_a^2}{a^2} + \frac{s_b^2}{b^2} + \frac{s_c^2}{c^2}}
\end{aligned} \tag{2.22}$$

b)

$$\begin{aligned}
y &= \frac{a-b}{c+d} \\
s_y^2 &= \left(\frac{1}{c+d}\right)^2 s_a^2 + \left(-\frac{1}{c+d}\right)^2 s_b^2 + \left(-\frac{a-b}{(c+d)^2}\right)^2 (s_c^2 + s_d^2) \\
s_y^2 &= \frac{s_a^2 + s_b^2}{(c+d)^2} + \frac{(a-b)^2}{(c+d)^4} (s_c^2 + s_d^2)
\end{aligned} \tag{2.23}$$

c)

$$\begin{aligned}
y &= \frac{a-b}{a+b} \\
s_y^2 &= \left[\frac{2b}{(a+b)^2}\right]^2 s_a^2 + \left[\frac{2A}{(a+b)^2}\right]^2 s_b^2
\end{aligned} \tag{2.24}$$

d)

$$\begin{aligned}
y &= \left(\frac{a}{b+c}\right)^{1/3} \\
s_y^2 &= \left[\frac{1}{3} \left(\frac{a}{b+c}\right)^{-2/3} \frac{1}{b+c}\right]^2 s_a^2 + \left[-\frac{1}{3} \left(\frac{a}{b+c}\right)^{-2/3} \frac{a}{(b+c)^2}\right]^2 (s_b^2 + s_c^2)
\end{aligned} \tag{2.25}$$

e)

$$\begin{aligned}
y &= 10^a \\
\ln y &= a \ln 10; \quad y = e^{a \ln 10} \\
s_y &= e^{a \ln 10} \ln(10) s_a = y \ln(10) s_a
\end{aligned} \tag{2.26}$$

f)

$$\begin{aligned}
y &= \log_{10} a \\
a &= 10^y; \quad \ln a = y \ln 10 \\
y &= \frac{\ln a}{\ln 10} \\
s_y &= \frac{1}{a \ln 10} s_a
\end{aligned} \tag{2.27}$$

g)

$$\begin{aligned}
y &= ab^c \\
s_y^2 &= (b^c)^2 s_a^2 + (acb^{c-1})^2 s_b^2 + (ab^c \ln b)^2 s_c^2
\end{aligned} \tag{2.28}$$

Example 2.10.

pH of the solution is 2.10. What is the standard deviation of the H^+ concentration when:
 $s_{pH} = 0.01$

$$\begin{aligned}
a_{H^+} &= 10^{-pH} = 7.94 \times 10^{-3} \\
s_{a_{H^+}} &= 10^{-pH} \ln 10 s_{pH} = 1.8 \times 10^{-4} \\
s_{a_{H^+}} &= 0.18 \times 10^{-3}
\end{aligned} \tag{2.29}$$

 $s_{pH} = 0.02$

$$s_{a_{H^+}} = 0.37 \times 10^{-3} \tag{2.30}$$

See calculations in *Examples2.xlsx*, sheet *Ex. 2.10*.

Example 2.11.

From the Nernst equation $E = E^0 + p \log a_x$ determine the standard deviation of a_x .

$$\begin{aligned}
a_x &= 10^{(E-E^0)/p} \\
\frac{\partial a_x}{\partial E} &= 10^{(E-E^0)/p} (\ln 10) / p \\
\frac{\partial a_x}{\partial E^0} &= -10^{(E-E^0)/p} (\ln 10) / p \\
\frac{\partial a_x}{\partial p} &= -10^{(E-E^0)/p} [(E-E^0) \ln 10] / p^2 \\
s_{a_x}^2 &= \left[\frac{10^{(E-E^0)/p} (\ln 10)}{p} \right]^2 (s_E^2 + s_{E^0}^2) + \left[\frac{10^{(E-E^0)/p} [(E-E^0) \ln 10]}{p^2} \right]^2 s_p^2
\end{aligned} \tag{2.31}$$

Example 2.12.

To determine the activity using ion selective electrodes two measurements were performed, first in the standard solution containing $a_1 = 10^{-3}$ where $E_1 = 237.1$ mV and the second in the unknown solution where $E_2 = 250.7$ mV. The slope $p = 25.9$ mV, $s_p = 0.4$ mV, the standard deviations of the potentials are $s_E = 1.0$ mV, and that of the standard $s_{a_1} = 3 \times 10^{-5}$. What is the value, standard deviation, and confidence interval of the unknown solution assuming that the measurements of a_1 were repeated 6 times, E_1 4 times, E_2 4 times, and p 5 times?

$$\begin{aligned}
 E_1 &= E^0 + p \ln a_1 & E_2 &= E^0 + p \ln a_2 \\
 a_2 &= a_1 e^{\frac{E_2 - E_1}{p}} = 1.691 \times 10^{-3} \\
 \frac{\partial a_2}{\partial a_1} &= e^{\frac{E_2 - E_1}{p}} & \frac{\partial a_2}{\partial E_1} &= a_1 e^{\frac{E_2 - E_1}{p}} \left(-\frac{1}{p} \right) \\
 \frac{\partial a_2}{\partial E_2} &= a_1 e^{\frac{E_2 - E_1}{p}} \left(\frac{1}{p} \right) & \frac{\partial a_2}{\partial p} &= a_1 e^{\frac{E_2 - E_1}{p}} \frac{(E_2 - E_1)}{p^2} \\
 s_{a_2} &= 3.7 \times 10^{-4} \\
 a_2 &= 1.69 \times 10^{-3} & s_{a_2} &= 0.37 \times 10^{-3}
 \end{aligned} \tag{2.32}$$

The effective number of degrees of freedom for a_2 , $df_{eff} = 4.719 \approx 5$, and $CI = 0.95 \times 10^{-3}$. See calculations in *Examples2.xlsx*, sheet *Ex. 2.12*.

3 Linear regression

3.1 Introduction

Very often there is a correlation between data y and x , see for example Fig. 3.1.

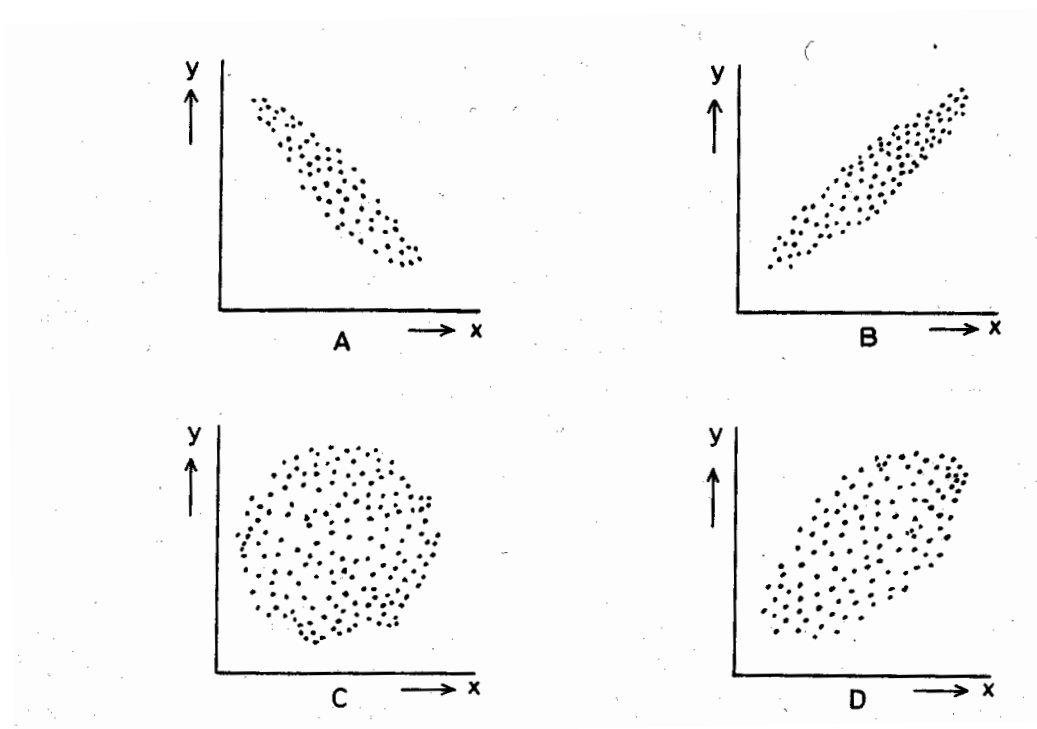


Fig. 3.1. Examples of correlation between x and y : a) and b) negative and positive correlations with low dispersion, c) no correlation, d) correlation with large dispersion.

If there is a correlation between data one can use different models to describe those using equations. Let us first consider linear correlation between data.

In analytical and physical chemistry very often linear relations exist. In many other cases, the nonlinear equation might be linearized, e.g.:

$$E = E^0 + p \ln a$$

$$y = b_0 + b_1 x \quad (3.1)$$

where $y = E$, $b_0 = E^0$, $b_1 = p$, $x = \ln a$

or

$$k = A \exp(E / RT)$$

$$\ln k = \ln A + \frac{E}{RT}$$

$$y = b_0 + b_1 x \quad (3.2)$$

$$y = \ln k, \quad b_0 = \ln A, \quad b_1 = \frac{E}{R}, \quad x = \frac{1}{T}$$

It should be added that linear regression means that regression is linear versus parameters, that is polynomial regression is linear and $y = \exp(b x)$ nonlinear, because y is a nonlinear function of parameter b .

3.2 Determination of the parameters and standard deviations of linear regression

Let us suppose that N values y_i are measured as function of x_i :

$$\begin{array}{ll} x_1 & y_1 \\ x_2 & y_2 \\ x_3 & y_3 \\ \dots & \\ x_N & y_N \end{array} \quad (3.3)$$

Let us also assume that only the measured values y_i are determined with certain error while the values x_i are known without error that is their precision is much larger than that of y_i . Then we can postulate a linear regression between the data:

$$\hat{y} = b_0 + b_1 x \quad (3.4)$$

where \hat{y} are the values calculated using above equation and parameters b_0 and b_1 are calculated for N pairs x_i and y_i . Of course, in reality, because of the experimental errors, for each experimental point one can write the relation:

$$y_i = b_0 + b_1 x_i + \varepsilon_i \quad (3.5)$$

where ε_i is:

$$\varepsilon_i = y_i - \hat{y}_i \quad (3.6)$$

the difference between the experimental and calculated value for x_i . The least squares method minimizes the sum of squares S^2 :

$$S^2 = \sum_i^N \varepsilon_i^2 = \sum_{i=1}^N (y_i - b_0 - b_1 x_i)^2 = \sum_{i=1}^N (y_i - \hat{y}_i)^2 = \min \quad (3.7)$$

This means that the parameters b_0 and b_1 minimize the sum of squares. At the minimum one can write that the derivative of the function is zero:

$$\left(\frac{\partial S^2}{\partial b_0} \right)_{b_1} = 0 \quad \left(\frac{\partial S^2}{\partial b_1} \right)_{b_0} = 0 \quad (3.8)$$

This is illustrated in Fig. 3.2. Differentiation, Eq. (3.8), of S^2 , Eq. (3.7), gives:

$$\begin{aligned} \sum 2(y_i - b_0 - b_1 x_i) &= 0 \\ \sum 2(y_i - b_0 - b_1 x_i) x_i &= 0 \end{aligned} \quad (3.9)$$

where the summation goes from $i = 1$ to N . These two equations might be rearranged into:

$$\begin{aligned} N b_0 + \left(\sum x_i \right) b_1 &= \sum y_i \\ \left(\sum x_i \right) b_0 + \left(\sum x_i^2 \right) b_1 &= \sum x_i y_i \end{aligned} \quad (3.10)$$

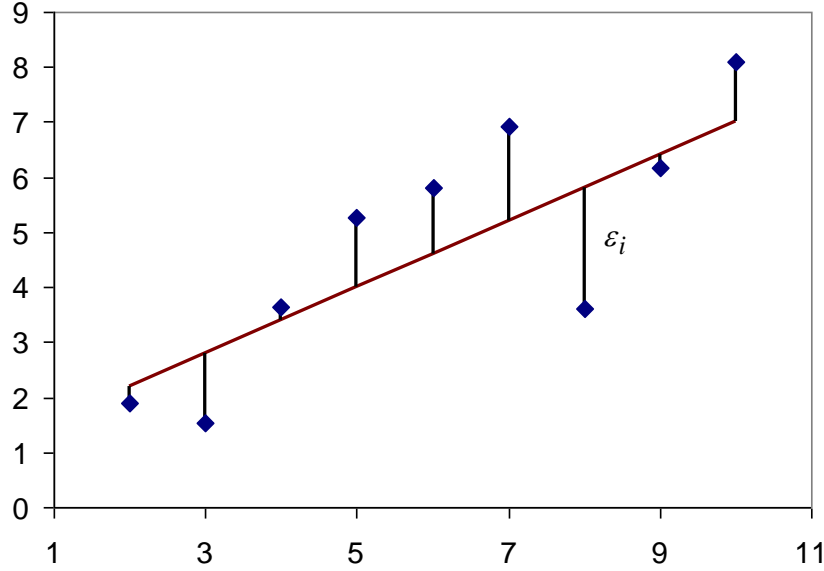


Fig. 3.2. The experimental points and a straight line which minimizes the sum of squares of deviations $\sum \varepsilon_i^2$.

Eq. (3.10) presents a system of two equations with two unknowns. It can be written in a matrix form as:

$$\begin{bmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} \quad (3.11)$$

with solution:

$$b_0 = \frac{\sum x^2 \sum y - \sum x \sum xy}{d} \quad (3.12)$$

and

$$b_1 = \frac{N \sum xy - \sum x \sum y}{d} = \frac{S_{xy}}{S_{xx}} \quad (3.13)$$

where

$$d = N \sum x^2 - (\sum x)^2 = NS_{xx} \quad (3.14)$$

Matric method of solution of Eq. (3.11) will be presented in Section 3.12. One can give another form of the solutions introducing new parameters:

$$\begin{aligned}
S_{xx} &= \sum (x_i - \bar{x})^2 = \sum \left(x_i^2 - \frac{2x_i \sum x_i}{N} + \frac{(\sum x_i)^2}{N^2} \right) \\
&= \left(\sum x_i^2 - \frac{2(\sum x_i)^2}{N} + \frac{N(\sum x_i)^2}{N^2} \right) \\
&= \sum x_i^2 + \frac{-2(\sum x_i)^2 + (\sum x_i)^2}{N} = \sum x_i^2 - \frac{(\sum x_i)^2}{N} \\
&= \frac{d}{N} = \sum x_i^2 - N \bar{x}^2
\end{aligned} \tag{3.15}$$

$$S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{N} \tag{3.16}$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{\sum x_i \sum y_i}{N} \tag{3.17}$$

and

$$b_1 = \frac{S_{xy}}{S_{xx}} \tag{3.18}$$

It should be added that to avoid numerical errors S_{ij} should be calculated using formulas with mean values instead of sums of x_i and y_i . Standard deviation of each value y_i , s_y (also called residual standard deviation, s_r) is calculated from the sum of deviations $y_i - \hat{y}_i$, similarly to the standard deviation of the arithmetic mean, but with the number of degrees of freedom is now $df = N - 2$, because 2 parameters: b_0 and b_1 must be calculated first to obtain the standard deviation:

$$s_y = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{N - 2}} = \sqrt{\frac{\sum (y_i - b_0 - b_1 x_i)^2}{N - 2}} \tag{3.19}$$

The standard deviation of b_1 is calculated from Eq. (3.18) using Eq. (3.17):

$$s_{b_1}^2 = \sum_i \left(\frac{\partial b_1}{\partial y_i} \right)^2 s_y^2 = s_y^2 \sum_i \left(\frac{\partial b_1}{\partial y_i} \right)^2 \tag{3.20}$$

$$\frac{\partial b_1}{\partial y_i} = \frac{Nx_i - \sum x_i}{d} \tag{3.21}$$

$$\left(\frac{\partial b_1}{\partial y_i} \right)^2 = \frac{N^2 x_i^2 - 2Nx_i \sum x_k + (\sum x_k)^2}{d^2} \tag{3.22}$$

and finally:

$$s_{b_1}^2 = \frac{N^2 \sum x_i^2 - 2N \left(\sum x_i \right)^2 + N \left(\sum x_i \right)^2}{d^2} s_y^2 = N \frac{N \sum x_i^2 - \left(\sum x_i \right)^2}{\left[N \sum x_i^2 - \left(\sum x_i \right)^2 \right]^2} s_y^2 = \frac{N}{d} s_y^2 \quad (3.23)$$

and

$$s_{b_1} = \sqrt{\frac{N}{d}} s_y = \frac{s_y}{\sqrt{S_{xx}}} = \frac{s_y}{\sqrt{\sum (x_i - \bar{x})^2}} \quad (3.24)$$

To calculate the standard deviations of the origin b_0 one should use condition that regression passes through point \bar{x} , \bar{y} (see below)

$$\bar{y} = b_0 + b_1 \bar{x} \quad (3.25)$$

or

$$b_0 = \bar{y} - b_1 \bar{x} \quad (3.26)$$

and applying the law of error propagation:

$$\begin{aligned} s_{b_0}^2 &= \left(\frac{\partial b_0}{\partial \bar{y}} \right)^2 s_{\bar{y}}^2 + \left(\frac{\partial b_0}{\partial b_1} \right)^2 s_{b_1}^2 = 1 \cdot \frac{s_y^2}{N} + (-\bar{x})^2 \frac{N s_y^2}{d} = \frac{\sum x_i^2}{d} s_y^2 \\ s_{b_0} &= \sqrt{\frac{\sum x_i^2}{d}} s_y = \sqrt{\frac{\sum x_i^2}{N} \frac{1}{\sum (x_i - \bar{x})^2}} s_y = \sqrt{\frac{\sum x_i^2}{N S_{xx}}} s_y \end{aligned} \quad (3.27)$$

3.3 Properties of the least-squares method

Let us determine the sum of deviations, ε_i :

$$\begin{aligned} \sum_{i=1}^N \varepsilon_i &= \sum_{i=1}^N (y_i - \hat{y}_i) = \sum (y_i - b_0 - b_1 x_i) \\ &= \sum \left[\frac{y_i - \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{d}}{-\frac{(N \sum x_i y_i - \sum x_i \sum y_i) x_i}{d}} \right] \end{aligned} \quad (3.28)$$

$$\begin{aligned}
&= \frac{\sum y_i d - N \sum x_i^2 \sum y_i + N \sum x_i \sum x_i y_i}{d} \\
&\quad + \frac{-N \sum x_i \sum x_i y_i + \sum y_i (\sum x_i)^2}{N} \\
&= \frac{N \sum x_i^2 \sum y_i - \sum y_i (\sum x_i)^2}{d} \\
&\quad + \frac{-N \sum x_i^2 \sum y_i + \sum y_i (\sum x_i)^2}{d} = 0
\end{aligned} \tag{3.29}$$

or

$$\sum y_i = \sum \hat{y}_i \tag{3.30}$$

$$\bar{y}_i = \bar{\hat{y}}_i \tag{3.31}$$

which means that the averages of the experimental and calculated values is the same and the straight line passes through this point: \bar{x}, \bar{y} .

3.4 Standard deviation of the calculated values \hat{y}_i

The calculated values are given by:

$$\hat{y}_i = b_0 + b_1 x_i \tag{3.32}$$

and might be rearranged into:

$$\hat{y}_i = \bar{y} + b_1 (x_i - \bar{x}) \tag{3.33}$$

The standard deviation is calculated using error propagation method:

$$s_{\hat{y}_i}^2 = s_{\bar{y}}^2 + s_{b_1}^2 (x_i - \bar{x})^2 \tag{3.34}$$

but

$$s_{\bar{y}}^2 = \frac{s_y^2}{N} \quad s_y = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{N - 2}} \quad s_{b_1}^2 = \frac{s_y^2}{S_{xx}} \quad S_{xx} = \sum (x_i - \bar{x})^2 \tag{3.35}$$

and

$$s_{\hat{y}_i} = \sqrt{\frac{1}{N} + \frac{(x_i - \bar{x})^2}{S_{xx}}} s_y \tag{3.36}$$

This equation indicate that the smallest standard deviation is at $x_i = \bar{x}$, $s_{\hat{y}_i} = s_y / \sqrt{N}$, and it increases going further in both directions. The confidence intervals for these parameters are calculated by multiplying the standard deviations by $t_{0.05, N-2}$ (assuming 95% probability):

$$\begin{aligned}
&b_0 \pm t_{0.05, N-2} s_{b_0} \\
&b_1 \pm t_{0.05, N-2} s_{b_1} \\
&\hat{y}_i \pm t_{0.05, N-2} s_{\hat{y}_i}
\end{aligned} \tag{3.37}$$

This is illustrated in Fig. 3.3.

3.5 Standard deviations of the experimental y_i

In a similar way it is possible to calculate the standard deviations and confidence intervals of y_i . Deviation, Δ , of the experimental point y_i from that predicted by regression, \hat{y}_i is:

$$\Delta = \varepsilon_i = y_i - \hat{y}_i \quad (3.38)$$

$$s_{\Delta}^2 = s_y^2 + s_{\hat{y}_i}^2 = s_y^2 \left(1 + \frac{1}{N} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right) \quad (3.39)$$

and

$$s_{\Delta} = s_y \sqrt{1 + \frac{1}{N} + \frac{(x_i - \bar{x})^2}{S_{xx}}} \quad (3.40)$$

It is evident that the standard deviation of the experimental y_i is larger than that of \hat{y}_i calculated and it also depends on the distance from \bar{x} . Comparison of the confidence intervals $y_i \pm s_{\Delta} t(\alpha, k)$ is presented in Fig. 3.4 $y_i \pm s_{\Delta} t(\alpha, k)$. Certain graphical programs (Origin, SigmaPlot) make these plots automatically. They might be also calculated manually in Excel using formulas presented above.

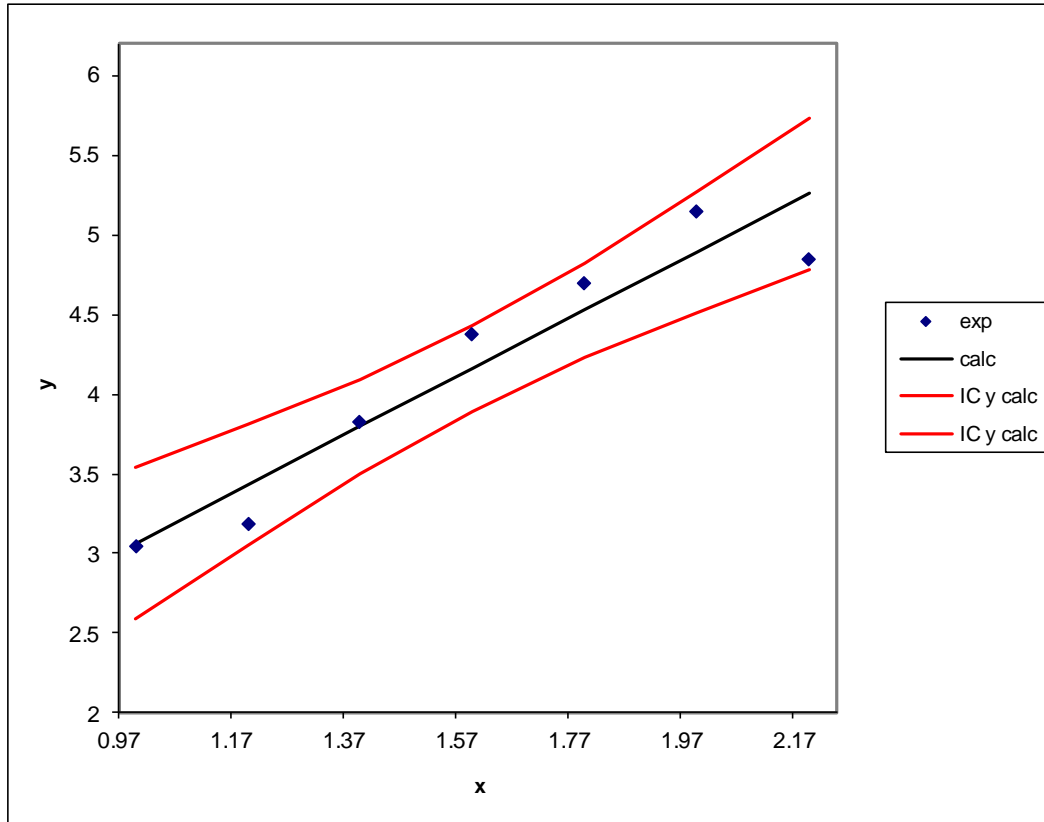


Fig. 3.3. Plot of the experimental points (symbols), calculated straight line (black line) and confidence intervals of \hat{y}_i ($\hat{y}_i \pm t_{0.05, N-2} s_{\hat{y}_i}$, red line).

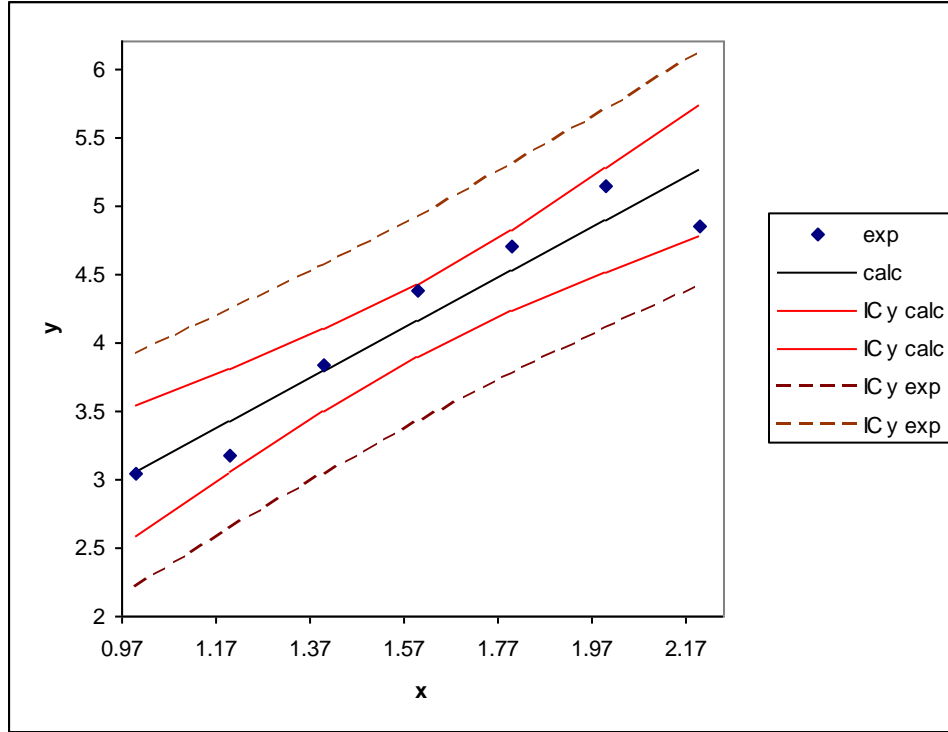


Fig. 3.4. Plot of the experimental points (symbols) calculated straight line (black line), and confidence intervals of \hat{y}_i (red line), and that of y_i (dashed line) for $\alpha = 0.05$.

3.6 Correlation and determination coefficients

Correlation coefficient, r , describes how good is the regression. It takes values between $-1 \leq r \leq 1$. It is defined as:

$$r = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}} \quad (3.41)$$

When it is 0 there is no correlation and when it is $|1|$ the correlation is ideal, and all the experimental points lay exactly on the line. Qualitatively, one can use terms:

- $|r| > 0.95$ good correlation
- $|r| > 0.99$ very good correlation.

The statistical meaning has the **determination coefficient** r^2 (or R^2). It represents the ratio of sum of squares explained by regression, SS_{reg} to the total sum of squares, SS_{tot} . Eq. (3.41) might be written in another form:

$$r^2 = \frac{SS_{\text{reg}} \text{ (explained by regression)}}{SS_{\text{tot}} \text{ (total sum of squares)}} = \frac{\sum_{i=1}^N (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^N (y_i - \bar{y})^2} \quad (3.42)$$

Its meaning is how much of the total variation if y_i can be explained by regression. It is sometimes expressed in %.

To better understand how to perform regression an example is shown below. Excel allows to determine regression and its errors using Regression in Data Analysis (in Data).

Example 3.1.

Determine regression parameters for the following data:

x	y
-5	-7.4
-3	-4.3
-1	-0.4
1	3.3
3	6.7
5	10.2
7	12.4
9	16.4

Use of Regression in Excel is displayed below.

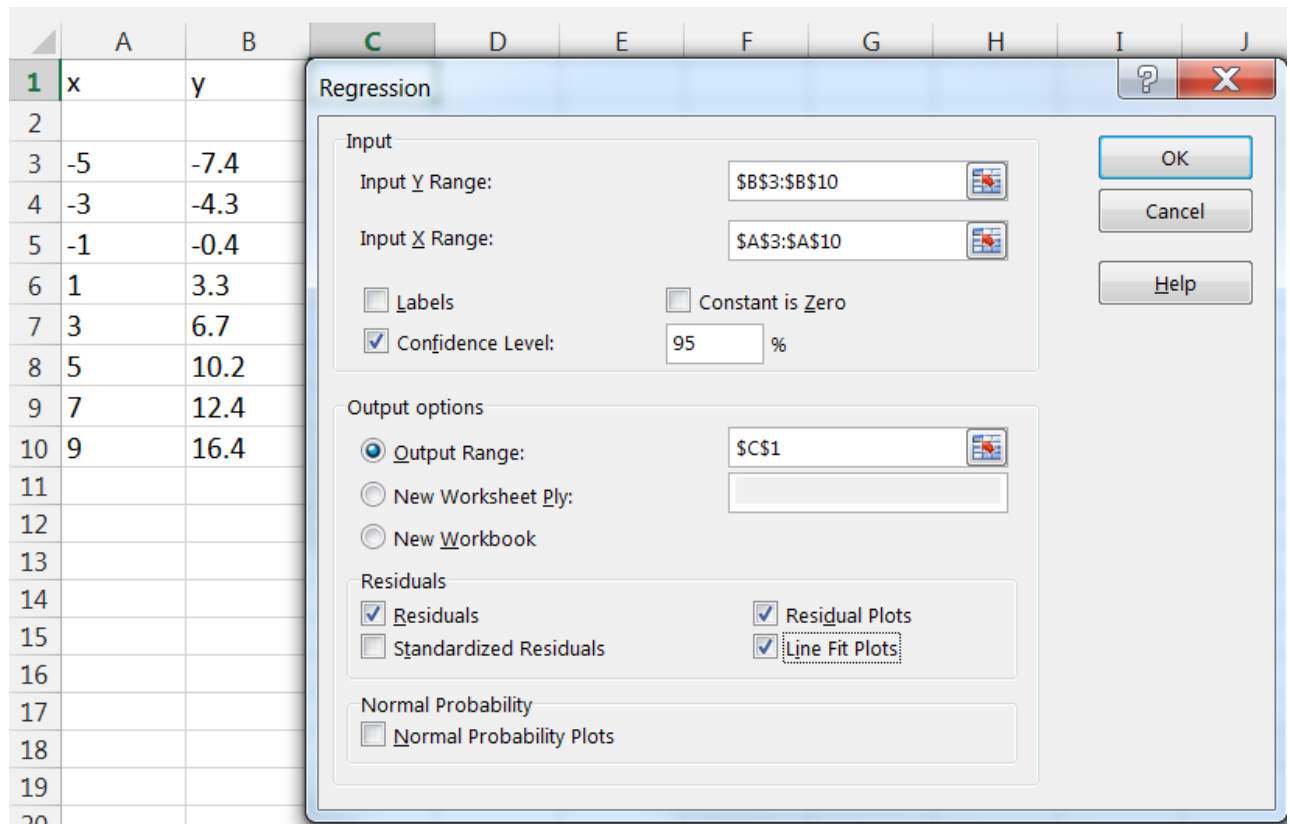


Fig. 3.5. Use of the Regression in Excel.

The obtained results are displayed in Table 3.1.

First, Regression Statistics presents determination coefficient: R Square (r^2) = 0.99743, and that there are 8 data points, N , Observations = 8.

The Analysis of Variances ANOVA will be presented later.

Table 3.1. Results of the regression analysis in Excel for Example 3.1.

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.998714
R Square	0.99743
Adjusted R Square	0.997002
Standard Error	0.456109
Observations	8

<i>ANOVA</i>					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	484.5005357	484.5005	2328.93	5.30764E-09
Residual	6	1.248214286	0.208036		
Total	7	485.74875			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	1.22	0.18	6.911539	0.000454	0.79	1.65
X Variable 1	1.698	0.035	48.25898	5.31E-09	1.612	1.784

RESIDUAL OUTPUT

<i>Observation</i>	<i>Predicted Y</i>	<i>Residuals</i>
1	-7.275	-0.125
2	-3.87857	-0.421428571
3	-0.48214	0.082142857
4	2.914286	0.385714286
5	6.310714	0.389285714
6	9.707143	0.492857143
7	13.10357	-0.703571429
8	16.5	-0.1

Test F of the significance of the parameter b_1 is $2328.9 \gg F(0.05, 1, 6) = 5.99$ (parameter b_1 is very important) and the probability, p , of the hypothesis $H_0: b_1 = 0$ is shown as *Significance F* and

calculated using Excel function F.DIST.RT(F,1,df₂) in this case F.DIST.RT(F,1,6), see Excel file. It is $p = 5.31 \times 10^{-9}$ and is much lower than 0.05 and this hypothesis must be rejected.

Next, there are regression results:

Intercept (b_0) = 1.22,

Standard deviation: Standard Error $s_{b_0} = 0.18$,

Confidence Intervals for the assumed probability of 95%: $0.79 \leq b_0 \leq 1.65$

X Variable 1 (b_1) = 1.698

Standard deviation: Standard Error $s_{b_1} = 0.035$

Confidence Intervals for the assumed probability of 95%: $1.612 \leq b_1 \leq 1.784$

These values were rounded using Excel function $\rightarrow .00$ to keep the precision according to the standard deviations (not more than 2 significant digits in standard deviation).

The probability, p , that $H_0 (b_i = 0)$ is true is called in Excel “*P-value*” and it is 0.00045 for b_0 , and 5.3×10^{-9} for b_1 , both much lower than 0.05. It is calculated using Excel function T.DIST.2T(t , df), here $df = N - 2 = 6$.

Finally there are calculated values, \hat{y}_i as “*Predicted Y*” and “*Residuals*”: $y_i - \hat{y}_i$. There are also automatic plots created for the regression and residuals.

One can write the summary results as follows:

Model: $y = b_0 + b_1 x$

$b_0 = 1.22$, $s_{b_0} = 0.18$, CI for 95% : $0.79 \leq b_0 \leq 1.65$ or $b_0 = 1.22 \pm 0.43$

$b_1 = 1.698$, $s_{b_1} = 0.035$, CI for 95% : $1.612 \leq b_1 \leq 1.784$ or $b_1 = 1.698 \pm 0.086$

$r^2 = 0.99743$

See calculations in Examples3.xlsx, sheet Ex. 3.1.

3.7 Linear regression for $y = b_1 x$

In certain cases, the free term b_0 is insignificant and a simpler linear model might be postulated:

$$\begin{aligned}\hat{y} &= b_1 x \\ y_i &= b_1 x_i + \varepsilon_i\end{aligned}\tag{3.43}$$

To find the best value of b_1 one should minimize the sum of squares:

$$\begin{aligned}S^2 &= \sum_{i=1}^N (y_i - b_1 x_i)^2 = \sum_{i=1}^N (y_i^2 - 2b_1 x_i y_i + b_1^2 x_i^2) = \\ &\sum y_i^2 - 2b_1 \sum x_i y_i + b_1^2 \sum x_i^2\end{aligned}\tag{3.44}$$

that is the derivative of sum of squares by the coefficient should be equal zero:

$$\frac{dS^2}{db_1} = -2 \sum x_i y_i + 2b_1 \sum x_i^2 = 0\tag{3.45}$$

from which:

$$b_1 = \frac{\sum x_i y_i}{\sum x_i^2} \quad (3.46)$$

The standard deviation of the slope is obtained as:

$$s_{b_1}^2 = \sum \left(\frac{db_1}{dy_i} \right)^2 s_y^2 \quad (3.47)$$

$$\frac{db_1}{dy_i} = \frac{x_i}{\sum x_i^2} \quad \left(\frac{db_1}{dy_i} \right)^2 = \frac{x_i^2}{\left(\sum x_i^2 \right)^2} \quad (3.48)$$

and

$$s_{b_1}^2 = \frac{\sum x_i^2}{\left(\sum x_i^2 \right)^2} s_y^2 = \frac{s_y^2}{\sum x_i^2} \quad (3.49)$$

$$s_{b_1} = \frac{s_y}{\sqrt{\sum x_i^2}}$$

where

$$s_y^2 = \frac{\sum (y_i - \hat{y}_i)^2}{N-1} = \frac{\sum (y_i - b_1 x_i)^2}{N-1} \quad (3.50)$$

The standard deviation of \hat{y}_i is calculated from the condition that the line passes through \bar{x}, \bar{y}

$$\begin{aligned} \hat{y}_i - \bar{y} &= b_1 (x_i - \bar{x}) \\ \hat{y}_i &= \bar{y} + b_1 (x_i - \bar{x}) \end{aligned} \quad (3.51)$$

and using the error propagation technique:

$$\begin{aligned} s_{\hat{y}_i}^2 &= s_y^2 + s_{b_1}^2 (x_i - \bar{x})^2 \\ s_{\hat{y}_i}^2 &= \frac{s_y^2}{N} + \frac{(x_i - \bar{x})^2}{\sum x_i^2} s_y^2 \end{aligned} \quad (3.52)$$

$$s_{\hat{y}_i} = \sqrt{\frac{1}{N} + \frac{(x_i - \bar{x})^2}{\sum x_i^2}} s_y$$

Similarly as before the standard deviation of y_i is:

$$s_{y_i} = \sqrt{1 + \frac{1}{N} + \frac{(x_i - \bar{x})^2}{\sum x_i^2}} s_y \quad (3.53)$$

Example 3.2.

Determine the linear regression parameters assuming the following model: $\hat{y} = b_1x$ for the data below:

x	y
0.2	0.57
0.4	1.24
0.6	1.78
0.8	2.43
1.0	3.12
1.2	3.57

Using Excel Regression analysis and selection option: Constant is Zero, the following results are displayed in Table 3.2.

Table 3.2. Results of the regression analysis in Excel for Example 3.2.

SUMMARY OUTPUT

<i>Regression Statistics</i>						
Multiple R	0.99976					
R Square	0.999521					
Adjusted R Square	0.799521					
Standard Error	0.056592					
Observations	6					

ANOVA						
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
Regression	1	33.39909	33.39909	10428.62	1.71E-09	correct
Residual	5	0.016013	0.003203		5.51E-08	incorrect
Total	6	33.4151				

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A
X Variable 1	3.029	0.030	102.1206	1.71E-09	2.953	3.105

RESIDUAL OUTPUT

<i>Observation</i>	<i>Predicted Y</i>	<i>Residuals</i>
1	0.605824	-0.03582
2	1.211648	0.028352
3	1.817473	-0.03747
4	2.423297	0.006703

5	3.029121	0.090879
6	3.634945	-0.06495

The results might be presented as:

Model : $y = b_1 x$

$b_1 = 3.029$, $s_{b_1} = 0.030$, CI (95%) : $2.953 \leq b_1 \leq 3.105$ or $b_1 - 3.029 \pm 0.076$

$r^2 = 0.9995$

It should be noticed that in older Excel versions incorrect table of ANOVA and R^2 were presented. It was corrected in Excel 2013. In Excel 2013 the value of p (*Significance F*) is calculated with incorrect number of degrees of freedom, F.DIST.RT(F,1,4) instead of F.DIST.RT(F,1,5). See the Excel file for details. Because there is only one parameter in regression equation p values “*Significance F*” and “*P-value*” for the significance of the slope are now identical. $p = 1.71 \times 10^{-9}$. See calculations in Examples3.xlsx, sheet Ex. 3.2.

3.8 Error of x_c value calculated from regression

In chemical analysis often the calibration curve is determined (with its parameters) and then from the measured signal of unknown, y_c , the corresponding unknown concentration, x_c is determined.¹⁷ It is important to know what is the standard deviation of x_c determined from the working curve. Two cases will be considered:

a) The working curve is described by equation: $\hat{y} = b_0 + b_1 x$

For the unknown concentration one can write the following equation:

$$\begin{aligned} y_c &= \bar{y} + b_1 (x_c - \bar{x}) \\ x_c &= \bar{x} + \frac{y_c - \bar{y}}{b_1} \end{aligned} \quad (3.54)$$

The standard deviation of x_c is:

$$s_{x_c}^2 = \left(\frac{\partial x}{\partial y_c} \right)^2 s_y^2 + \left(\frac{\partial x}{\partial \bar{y}} \right)^2 s_{\bar{y}}^2 + \left(\frac{\partial x}{\partial b_1} \right)^2 s_{b_1}^2 \quad (3.55)$$

but:

$$\begin{aligned} s_{\bar{y}}^2 &= \frac{s_y^2}{N} & s_{b_1}^2 &= \frac{s_y^2}{S_{xx}} \\ \frac{\partial x}{\partial y_c} &= \frac{1}{b_1} \\ \frac{\partial x}{\partial \bar{y}} &= -\frac{1}{b_1} \\ \frac{\partial x}{\partial b_1} &= -\frac{(y_c - \bar{y})}{b_1^2} \end{aligned} \quad (3.56)$$

and:

$$s_{x_c} = \frac{s_y}{b_1} \sqrt{1 + \frac{1}{N} + \frac{(y_c - \bar{y})^2}{b_1^2 S_{xx}}} \quad (3.57)$$

When the unknown y_c is measured m times the standard deviation is smaller:

$$s_{x_c} = \frac{s_y}{b_1} \sqrt{\frac{1}{m} + \frac{1}{N} + \frac{(y_c - \bar{y})^2}{b_1^2 S_{xx}}} \quad (3.58)$$

where the parameters: s_y , b_1 , S_{xx} and N are related to the calibration curve and were determined prior to the determination of the unknown.

b) The working curve is described as: $\hat{y} = b_1 x$

In this case the concentration is calculated using Eq. (3.54) and its standard deviation is:

$$s_{x_c} = \frac{s_y}{b_1} \sqrt{\frac{1}{m} + \frac{1}{N} + \frac{(y_c - \bar{y})^2}{b_1^2 \sum x_i^2}} \quad (3.59)$$

Application of this method is shown in Example 3.3.

Example 3.3.

In analytical chemistry a calibration curve was obtained from the following measurements of the analytical signal, y , versus concentration, x :

x	y
0.100	0.34
0.200	0.80
0.300	1.20
0.400	1.77
0.500	2.14
0.600	2.42
0.700	2.90
0.800	3.36
0.900	3.74

Next the analytical signal of the unknown sample was measured three times and the mean value was $y_c = 1.15$. Determine the unknown concentration, x_c , and its standard deviation.

First, calibration line was calculated using equation $y = b_0 + b_1 x$. The results are presented below:

Table 3.3. Fit of the experimental working curve (above) to the equation $y = b_0 + b_1x$:

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.998677
R Square	0.997357
Adjusted R Square	0.996979
Standard Error	0.063629
Observations	9

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	10.69348	10.69348	2641.246	2.77E-10
Residual	7	0.028341	0.004049		
Total	8	10.72182			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	-0.03639	0.046225	-0.78721	0.456983	-0.14569	0.072917
X Variable 1	4.221667	0.082145	51.39305	2.77E-10	4.027425	4.415908

It is clear that the t -test shows that the parameter b_0 is not important, $|t_{\text{exp}}| = 0.787 < t(0.05, 7) = 2.364$, see t -test for the importance of the parameters in Section 5.2.1. A new regression using equation $y = b_1x$ must be recalculated. It is presented below.

Table 3.4. Fit of the experimental working curve (above) to the equation $y = b_1x$:

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.999688
R Square	0.999376
Adjusted R Square	0.874376
Standard Error	0.062098
Observations	9

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	49.42085	49.42085	12816	1.11E-12

Residual	8	0.030849	0.003856
Total	9	49.4517	

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	0	#N/A	#N/A	#N/A	#N/A	#N/A
X Variable 1	4.164211	0.036784	113.2078	4.14E-14	4.079387	4.249034

Using this regression, the following results were obtained:

$$s_y = 0.062098182$$

$$x_c = 0.278$$

$$s_{x_c} = 0.011$$

$$CI(x_c) = t(0.05, 8) \times s_{x_c} = 0.025$$

At the confidence level of 95% the unknown concentration is: $x_c = 0.278 \pm 0.025$. See the calculations in Examples3.xlsx, sheet Ex. 3.3.

3.9 Calibration

In analytical chemistry, in order to determine the unknown concentration, a relation between the analytical signal and the concentration must be determined. For example, using spectroscopy, a linear relation between the absorbance, emission, or fluorescence and concentration is determined by measuring analytical signal for certain number of standards. The best straight line is then determined using the least-squares method described above. In the case of intrinsic nonlinear correlation other models (usually polynomial) are used. Let us look at the case when the linear relation exists. The concentrations must cover all the concentration range studied. Extrapolations outside this range should not be used because we do not know anything about the behavior of the analytical signal outside the studied zone.

3.10 Sensitivity

First, let us define the sensitivity. There are two types of sensitivity:

- Calibration sensitivity, m :** this is simply the slope of the calibration curve, $y = b_0 + b_1x$ which is often rewritten as $S = mC + S_{bl}$ where S (or y) is the analytical signal, C (or x) is the concentration and S_{bl} (or b_0) is the signal of the blank in the absence of the analyte. The larger the slope the larger is the calibration sensitivity.
- Analytical sensitivity, γ :** this is the ratio of the slope of the calibration curve, i.e. calibration sensitivity, to the standard deviation, s_C of the analytical signal at a given concentration:

$$\gamma = \frac{m}{s_C} \quad (3.60)$$

It can be noticed that m has units and is independent of concentration while γ is dimensionless but concentration dependent. It cannot be determined for the zero concentration because s_C does not exist in such a case.

3.10.1 Detection limit and dynamic range

One of the important factors in calibration is the **detection limit**. Qualitatively, detection limit is the smallest concentration which is distinguished from the blank. The smallest signal distinguished from the noise, S_m , is defined as:

$$S_m = \bar{S}_{bl} + 3s_{bl} \quad (3.61)$$

where \bar{S}_{bl} is the mean signal of the blank and s_{bl} is its standard deviation. This means that the signal three times larger than the standard deviation is the minimal signal distinguished. This is based on the accepted convention. For normal distribution, 99.4% of the results is found between $\pm 3s_{bl}$. This is illustrated in Fig. 3.6.

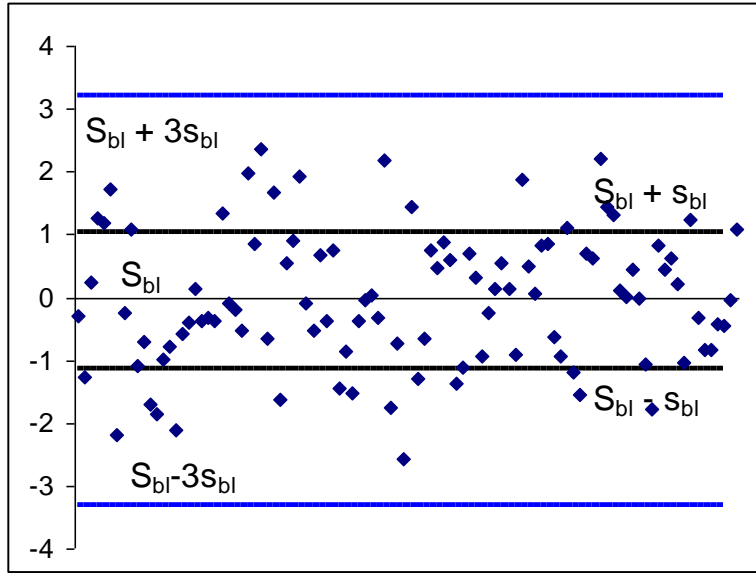


Fig. 3.6. Normal distribution of the results around \bar{S}_{bl} ; the black lines show $\bar{S}_{bl} \pm s_{bl}$ and blue lines $\bar{S}_{bl} \pm 3s_{bl}$.

The minimal concentration distinguished from the noise is calculated from the calibration curve:

$$C_m = \frac{S_m - \bar{S}_{bl}}{m} = \frac{3s_{bl}}{m} \quad (3.62)$$

To better understand the meaning of the limit of detection let us look at Fig. 3.7.

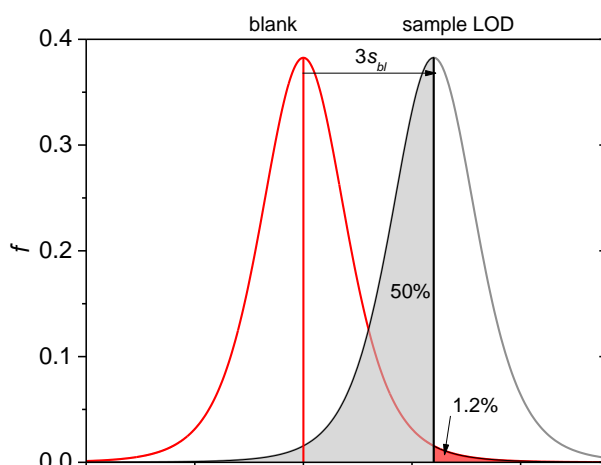


Fig. 3.7. Distribution of measurements of blank and of the sample at the detection limit (assuming Student distribution of samples for $df = 6$). Only 1.2% of blank measurements exceed LOD while 50% of measurements at LOD is below this limit (and 50% above).

It can be noticed that only 1.2% of measurements of the blank exceeds the detection limit while 50% of the measurements of a sample containing analyte at the detection limit is below the detection limit and 50% above. For the sample at the detection limit there is 50% chance of concluding that the analyte is absent because the signal is below the detection limit. The distribution curves were calculated assuming Student's distribution for $df = 6$; these curves are broader than those for the normal distribution.

The minimal concentration, which can be quantitatively determined, is called limit of quantitation (LOQ), C_{LOQ} , is larger than C_m . It is assumed that C_{LOQ} must be 10 times larger than the noise of the blank:

$$C_{LOQ} = \frac{10s_{bl}}{m} \quad (3.63)$$

It should be mentioned that this limit is a bit arbitrary and other definitions also exist.²³

When linear working curve is used in the analysis at higher concentrations one can observe nonlinearity and this is the end of the limit of linearity, LOL. This is illustrated in Fig. 3.8.

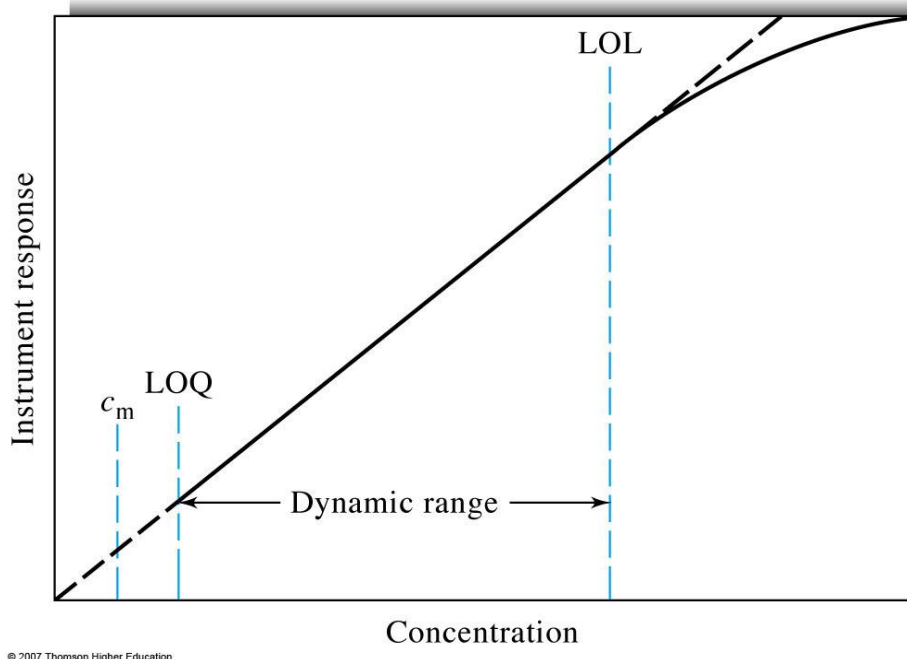


Fig. 3.8. Limit of detection limit, C_m , limit of quantitation, LOQ, and limit of linearity, LOL. The distance between LOQ and LOL determines useful dynamic range where analysis can be performed.²⁴

3.10.2 Selectivity

In analysis we expect that the measured signal is related to the measured analyte only. However, there might be interfering substances which influence the measured signal. Selectivity (or specificity) is the capability to distinguish of the analyte from other interfering species. Let us assume that besides the determined substance A there are interfering substances B and C. The measured signal S in such a case is described as:

$$S = m_A C_A + m_B C_B + m_C C_C + S_{bl} \quad (3.64)$$

Introducing selectivity coefficients for A with respect to B, $k_{A,B}$, and C, $k_{A,C}$ are

$$k_{A,B} = \frac{m_B}{m_A} \quad k_{A,C} = \frac{m_C}{m_A} \quad (3.65)$$

and Eq. (3.64) might be rearranged to

$$S = m_A (C_A + k_{A,B} C_B + k_{A,C} C_C) \quad (3.66)$$

The method is selective if the coefficients of selectivity are small.

Example 3.4.

The calibration data were determined and they are presented below.

x	y
0.2	0.61
0.4	0.98
0.6	1.43

0.8	1.85
1.0	2.32
1.2	2.69

- a) Determine the calibration curve.
- b) The signal of the blank was measured 15 times and the following results were obtained: $\bar{s}_{bl} = 0.079$, $s_{bl} = 0.055$. Determine the limit of detection (LOD) and limit of quantitation (LOQ).
- c) Besides, measurements for $x = 0.4$ and 1.0 were repeated 15 times and the following results were obtained:
 $x = 0.4$, $y = 0.98$, $s_y = 0.032$
 $x = 1.0$, $y = 2.36$, $s_y = 0.062$.
Determine the calibration and analytical sensitivity at these concentrations.
- d) Unknown concentration was measured 3 times and the average signal is $y_C = 2.28$. Calculate the unknown concentration and its standard deviation and confidence limits at 95%.
- All the calculations are shown in Example3.xlsx, sheet Ex. 3.4.
The results of the regression analysis are shown below.

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.999834
R Square	0.999668
Adjusted R Square	0.999585
Standard Error	0.016662
Observations	6

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	3.339773	3.339773	12030.06	4.14E-08
Residual	4	0.00111	0.000278		
Total	5	3.340883			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	0.132667	0.015511	8.552856	0.001026	0.0896	0.175733
X Variable 1	2.184286	0.019915	109.6816	4.14E-08	2.128993	2.239578
	<i>t(0.05,4)=</i>	<i>2.776</i>				

- a) The determination coefficient is large, 0.9997. The regression equation is:
 $y = (0.132 \pm 0.043) + (2.184 \pm 0.055) x$. Both parameters are statistically important based on the t -test and p -level tests.
- b) Concentration LOD is

$$C_{LOD} = \frac{3 \times 0.055}{2.184} = 0.075$$

and the limit of quantitation:

$$C_{\text{LOQ}} = \frac{10 \times 0.055}{2.184} = 0.25$$

- c) Calibration sensitivity is the slope of the regression curve, $m = 2.184$; it is independent of concentration.

The analytical sensitivity is:

$$x = 0.4, \gamma = \frac{2.184}{0.032} = 68$$

$$x = 1.0, \gamma = \frac{2.184}{0.042} = 52$$

- d) Using regression equation the unknown concentration is calculated as:

$$x_C = \frac{2.28 - 0.133}{2.184} = 0.983$$

The standard deviation of the concentration is calculated using Eq. (3.58):

$$s_{x_C} = \frac{0.01666}{2.1843} \sqrt{\frac{1}{3} + \frac{1}{6} + \frac{(2.28 - 1.662)^2}{2.1842^2 \times 0.70}} = 0.006$$

and the confidence intervals:

$$CI = 0.006 \times 2.776 = 0.02$$

The unknown concentration is: 0.983 ± 0.017 (or 0.98 ± 0.02).

For details see *Examples3.xlsx*, Sheet *Ex3.4*.

Example 3.5

Determine the error of the analytical signal of 3×10^{-3} M K^+ in the presence of 2×10^{-2} M Na^+ if selectivity coefficient is $k_{\text{K}^+, \text{Na}^+} = 0.052$.

The analytical signal is:

$$S = m_{\text{K}^+} (C_{\text{K}^+} + k_{\text{K}^+, \text{Na}^+} C_{\text{Na}^+}) + 0$$

or

$$\frac{S}{m_{\text{K}^+}} = 3 \cdot 10^{-3} + 0.052 \cdot 2 \cdot 10^{-2} = 4.04 \cdot 10^{-3}$$

and the relative error of the signal is:

$$E_{\text{rel}} = \frac{4.04 \cdot 10^{-3} - 3 \cdot 10^{-3}}{3 \cdot 10^{-3}} \times 100\% = 35\%$$

The presence of the interfering ion Na^+ causes error of 35%.

3.11 The method of standard additions

The method of standard additions is often used in analysis. To the analyzed sample one or several aliquots of standard solution are added and the analytical signal is measured. The advantage of this method over standard calibration technique is that the unknown solution matrix components (which can influence/interfere with the analytical signal) are the same. The disadvantage is that

this is an extrapolation technique and the assumption is made that beyond the concentration range studied the equations are the same.

For the simple addition of one sample and assuming simple relation between analytical signal, y , and concentration, x : $y = b_1x$ one makes two measurements, one for the unknown:

$$y_u = b_1x_u \quad (3.67)$$

and after that the addition of the standard:

$$y_1 = b_1(x_u + x_s) \quad (3.68)$$

where y_u is the signal of the unknown, y_1 is the signal of the mixture of unknown and standard, x_u is the concentration of the unknown, and x_s is the concentration of the standard. Division of these equations to eliminate the slope gives the concentration of unknown (assuming that dilution is neglected):

$$x_u = \frac{x_s y_u}{y_1 - y_u} \quad (3.69)$$

However, when addition of the standard volume V_s caused dilution, Eq. (3.68) must be replaced by:

$$y_1 = b_1 \left(x_u \frac{V}{V + V_s} + x_s \frac{V_s}{V + V_s} \right) \quad (3.70)$$

where V is the initial volume of the unknown. This leads to the concentration of unknown:

$$x_u = \frac{y_u x_s \frac{V_s}{V + V_s}}{y_1 - y_u \frac{V}{V + V_s}} \quad (3.71)$$

Much better method is the method of repeated additions of the standard solution where the obtained analytical signal is plotted versus concentration (or volume) added to the solution of unknown. Example of such plot is displayed in Fig. 3.9.

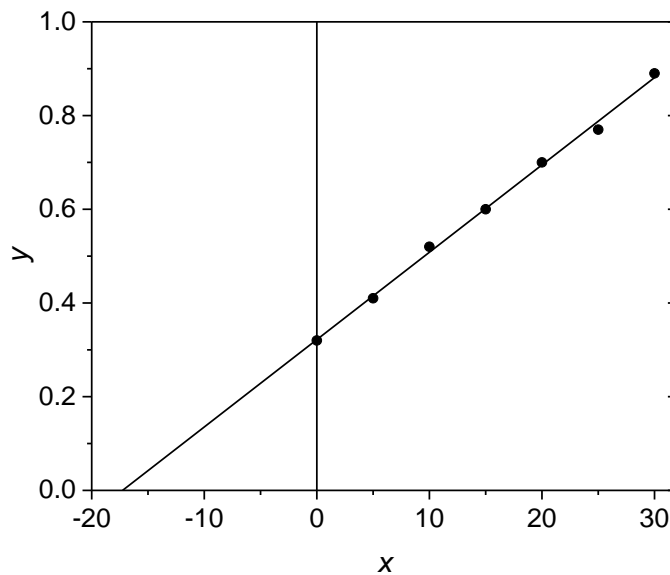


Fig. 3.9. Illustration of the method of standard additions. For $x = 0$ the signal corresponds to the unknown sample. The value of x extrapolated to $y = 0$ corresponds to the concentration of the unknown sample.

In order to take into account dilution one can plot $y(V+V_s)$ versus the volume of added standard V_s :

$$\begin{aligned} y(V + V_s) &= b_1(x_u V + x_s V_s) \\ Y &= B_0 + B_1 V_s \end{aligned} \quad (3.72)$$

where

$$B_0 = b_1 x_u V \quad \text{and} \quad B_1 = b_1 x_s \quad (3.73)$$

Extrapolation of the straight line to $Y = 0$ gives the standard volume V_s' and the concentration of the unknown is:

$$\begin{aligned} B_0 + B_1 V_s' &= 0 \\ V_s' &= -\frac{B_0}{B_1} \\ x_u &= \frac{B_0 x_s}{B_1 V} = -\frac{x_s V_s'}{V} \end{aligned} \quad (3.74)$$

from which unknown concentration might be determined. Notice, that V_s' is negative and, of course, x_u positive. In order to find the standard deviation of the found x_u , Eq. (3.72) might be rearranged to:

$$\begin{aligned} Y - \bar{Y} &= B_1(V_s - \bar{V}_s) \\ \text{for } Y &= 0 \\ -\bar{Y} &= B_1(V_s' - \bar{V}_s) \\ V_s' &= \bar{V}_s - \frac{\bar{Y}}{B_1} \end{aligned} \quad (3.75)$$

The error propagation equation might be applied to V_s' keeping in mind that V_s is the independent variable determined without error:

$$s_{V_s'}^2 = \frac{1}{B_1^2} \frac{s_{\bar{Y}}^2}{N} + \frac{\bar{Y}^2}{B_1^4} s_{B_1}^2 \quad (3.76)$$

the standard deviation of the slope was found earlier, Eq. (3.24), which leads to:

$$s_{V_s'}^2 = \frac{s_{\bar{Y}}^2}{B_1^2} \left(\frac{1}{N} + \frac{\bar{Y}^2}{B_1^2 \sum_{i=1}^N (V_s - \bar{V}_s)^2} \right) \quad (3.77)$$

and the standard deviation of the unknown concentration is:

$$s_{x_u} = \frac{x_s}{V} s_{V_s'} \quad (3.78)$$

Calculations of the standard deviation are shown in the following Example 3.6.

Example 3.6.

Determine concentration and its standard deviation for the method of standard additions using the following data: $V = 100$ ml, standard concentration: $x_s = 20$ ppm, and the measured atomic emission signal, y , and different volumes of the standard, V_s , added:

V_s	V_s+V	y	Y	$(V_s-\bar{V}_s)^2$
0	100	2.09	209.00	225
5	105	2.91	305.55	100
10	110	3.82	420.20	25
15	115	4.42	508.30	0
20	120	5.12	614.40	25
25	125	5.54	692.50	100
30	130	6.30	819.00	225
$\bar{V}_s =$	15	$\bar{Y} =$	509.85	700
				$= S_{xx}$

To linearize the signal it was multiplied by $V + V_s$: $Y = y(V + V_s)$, The plot of Y vs. V_s is shown in Fig. 3.10.

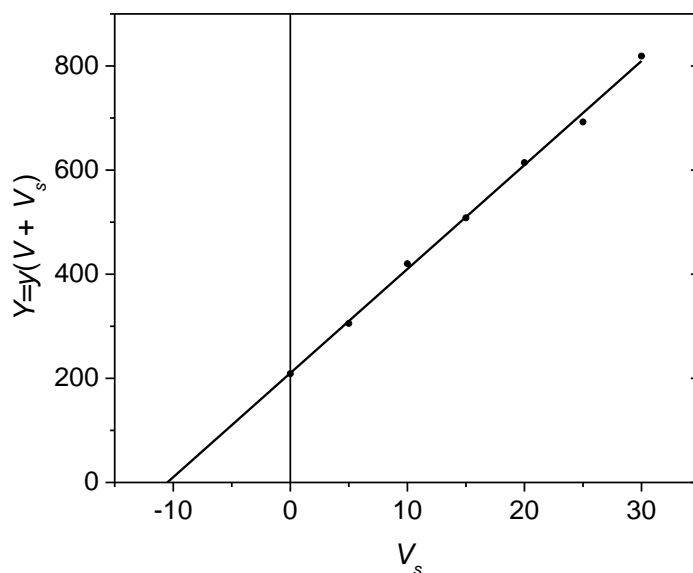


Fig. 3.10. Data analysis for the standard additions method, Example 3.6.

Regression analysis was carried out in Excel and it is in *Examples3.xlsx*, sheet *Ex. 3.5*. The results are shown below:

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.999046
R Square	0.998094
Adjusted R Square	0.997712
Standard Error	10.33495
Observations	7

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	279620.1	279620.1	2617.891	5.39E-08
Residual	5	534.0561	106.8112		
Total	6	280154.2			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	210.0536	7.042082	29.82833	7.94E-07	191.9513	228.1558
X Variable 1	19.98643	0.390624	51.16533	5.39E-08	18.9823	20.99056

Using Eq. (3.74), volume $V_s' = -10.51$ ml. Then, the unknown concentration is:

$$x_u = -\frac{x_s V_s'}{V} = 2.102 \quad (3.79)$$

The standard deviation of V_s' using Eq. (3.77) is:

$$s_{V_s'}^2 = \frac{s_Y^2}{B_1^2} \left(\frac{1}{N} + \frac{\bar{Y}^2}{B_1^2 \sum_{i=1}^N (V_s - \bar{V}_s)^2} \right) = \frac{106.81}{19.986^2} \left(\frac{1}{7} + \frac{509.85^2}{19.986^2 \times 700} \right) = 0.2868 \quad (3.80)$$

$$s_{V_s'} = 0.54$$

and standard deviation of the sample concentration, Eq. (3.78), $s_{x_u} = 0.11$. The confidence intervals for the unknown, using $t(0.05, 5) = 2.5706$, are:

$$x_u = 2.10 \pm 0.28 \text{ ppm} \quad (3.81)$$

See *Examples3.xlsx*, sheet *Ex. 3.5* for details.

3.12 Matrix description of the least-squares method

Matrix description of the regression simplifies the problem and allows for the generalization of the linear regression. It should be added that the term **linear regression** denotes systems where

the unknown parameters b_i are linear functions of y . In this sense polynomial approximation is linear:

$$\hat{y} = b_0 + b_1 x + b_2 x^2 + b_3 x^3 \quad (3.82)$$

and exponential function of b_i is nonlinear:

$$\hat{y} = b_1 \exp(b_2 x) \quad (3.83)$$

To describe a simple linear regression, i.e. fit to a straight line ($y = b_0 + b_1 x$) let us consider the following matrices: \mathbf{X} , \mathbf{Y} , \mathbf{b} , and $\boldsymbol{\varepsilon}$:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \dots & \\ 1 & x_N \end{bmatrix} \quad \mathbf{Y} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \dots \\ y_N \end{bmatrix} \quad \mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \dots \\ \varepsilon_N \end{bmatrix} \quad (3.84)$$

The linear model postulated might be written in matrix form as, see Eq. (3.5):

$$\mathbf{Y} = \mathbf{X}\mathbf{b} + \boldsymbol{\varepsilon} \quad (3.85)$$

Using the definition of transposed matrices:

$$\mathbf{X}' = \begin{bmatrix} 1 & 1 & 1 & \dots & 1 \\ x_1 & x_2 & x_3 & \dots & x_N \end{bmatrix} \quad (3.86)$$

$$\mathbf{Y}' = [y_1 \quad y_2 \quad y_3 \quad \dots \quad y_N]$$

one can write:

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix} \quad \mathbf{X}'\mathbf{Y} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} \quad (3.87)$$

and

$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y} \quad (3.88)$$

This equation is identical with Eq. (3.11). To determine vector \mathbf{b} both sides of this matrix equation must be multiplied by the inverse matrix:

$$(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}) \quad (3.89)$$

which gives the solution:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{Y} \quad (3.90)$$

The latter might be also obtained using matrix description of the data. The sum of squares, S^2 , Eq. (3.7), is given by:

$$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = (\mathbf{Y} - \mathbf{X}\mathbf{b})'(\mathbf{Y} - \mathbf{X}\mathbf{b}) \quad (3.91)$$

and taking into account that:

$$\frac{\partial}{\partial \mathbf{x}} [(\mathbf{a} + \mathbf{B}\mathbf{x})'(\mathbf{c} + \mathbf{D}\mathbf{x})] = \mathbf{B}'(\mathbf{c} + \mathbf{D}\mathbf{x}) + \mathbf{D}'(\mathbf{a} + \mathbf{B}\mathbf{x}) \quad (3.92)$$

Eq. (3.91) becomes

$$\frac{\partial}{\partial \mathbf{b}} [(\mathbf{Y} - \mathbf{X}\mathbf{b})'(\mathbf{Y} - \mathbf{X}\mathbf{b})] = -\mathbf{X}'(\mathbf{Y} - \mathbf{X}\mathbf{b}) - \mathbf{X}'(\mathbf{Y} - \mathbf{X}\mathbf{b}) = -2\mathbf{X}'(\mathbf{Y} - \mathbf{X}\mathbf{b}) \quad (3.93)$$

and the vector \mathbf{b} is found at the minimum

$$\frac{\partial(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon})}{\partial \mathbf{b}} = -2\mathbf{X}'(\mathbf{Y} - \mathbf{Xb}) = \mathbf{0} \quad (3.94)$$

$$\mathbf{X}'\mathbf{Xb} = \mathbf{X}'\mathbf{Y}$$

where $\mathbf{0}$ is the vector of zeros, which is identical with Eq. (3.88). The calculated values of \hat{y}_i are obtained as:

$$\hat{\mathbf{Y}} = \mathbf{Xb} \quad (3.95)$$

This equation might be rearranged into another form using Eq. (3.90) for \mathbf{b} :

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{HY} \quad (3.96)$$

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

where \mathbf{H} is called hat matrix as it changes experimental y_i into calculated \hat{y}_i from regression (it puts a hat on y_i). Its diagonal elements are called **leverage** which describe the influence of each response value on the fitted value for that same observation.

The matrix method might also be used to determine standard deviations. Matrix of variances and covariances of the regression parameters, $\mathbf{C_b}$ is:

$$\mathbf{C_b} = (\mathbf{X}'\mathbf{X})^{-1} s_y^2 = s_y^2 \begin{bmatrix} \frac{\sum x_i^2}{NS_{xx}} & -\frac{\sum x_i}{NS_{xx}} \\ -\frac{\sum x_i}{NS_{xx}} & \frac{1}{S_{xx}} \end{bmatrix} = s_y^2 \begin{bmatrix} \frac{\sum x_i^2}{NS_{xx}} & -\frac{\bar{x}}{S_{xx}} \\ -\frac{\bar{x}}{S_{xx}} & \frac{1}{S_{xx}} \end{bmatrix} \quad (3.97)$$

where

$$s_y^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} / (N - 2) \quad (3.98)$$

Eq. (3.97) can be written in another format:

$$\mathbf{C_b} = \begin{bmatrix} \text{var}(b_0) & \text{cov}(b_0, b_1) \\ \text{cov}(b_1, b_0) & \text{var}(b_1) \end{bmatrix} \quad (3.99)$$

where var denotes variance and cov covariance. It is evident that covariances of b_0 and b_1 are not equal zero and $\text{cov}(b_0, b_1) = \text{cov}(b_1, b_0)$.

The diagonal elements of this matrix are variances (standard deviation squared) of the regression parameters and the non-diagonal elements are covariances.

The calculated values of \hat{y}_k for one value of x_k may be obtained introducing vector:

$$\mathbf{X}_k' = [1 \quad x_k] \quad (3.100)$$

which gives:

$$\hat{y}_i = \mathbf{X}_k' \mathbf{b} = \mathbf{b}' \mathbf{X}_k \quad (3.101)$$

and its variance:

$$C_{\hat{y}_k} = \mathbf{X}_k' \mathbf{C_b} \mathbf{X}_k = \mathbf{X}_k' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_k s_y^2 \quad (3.102)$$

The variance of y_k is:

$$C_{y_k} = \left[(1 + \mathbf{X}_k' (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}_k) \right] s_y^2 \quad (3.103)$$

3.13 Polynomial regression

Polynomial regression is a form of linear regression (linear versus regression parameters) in the form:

$$y = b_0 + b_1x + b_2x^2 + b_3x^3 + \dots \quad (3.104)$$

Polynomial regression might be written simply using matrix notation. Let us assume polynomial model:

$$y_i = b_0 + b_1x_i + b_2x_i^2 + \varepsilon_i \quad (3.105)$$

For such model matrix \mathbf{X} , called Jacobian, must be modified:

$$\mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \\ \dots & & \\ 1 & x_N & x_N^2 \end{bmatrix} \quad (3.106)$$

but the regression equation, Eq. (3.90) is exactly the same.

Matrix \mathbf{X} might be easily constructed for other models keeping in mind the meaning of its parameters. Its elements are the derivatives of the regression equations, e.g. Eq. (3.105), by the parameters b_i :

$$x_{ij} = \frac{\partial y_i}{\partial b_j} \quad \mathbf{X} = \begin{bmatrix} \frac{\partial y_1}{\partial b_0} & \frac{\partial y_1}{\partial b_1} & \frac{\partial y_1}{\partial b_2} \\ \frac{\partial y_2}{\partial b_0} & \frac{\partial y_2}{\partial b_1} & \frac{\partial y_2}{\partial b_2} \\ \frac{\partial y_3}{\partial b_0} & \frac{\partial y_3}{\partial b_1} & \frac{\partial y_3}{\partial b_2} \\ \dots & \dots & \dots \\ \frac{\partial y_N}{\partial b_0} & \frac{\partial y_N}{\partial b_1} & \frac{\partial y_N}{\partial b_2} \end{bmatrix} \quad (3.107)$$

For Eq. (3.105) it is simply matrix Eq. (3.106). Another example is presented below.

Example 3.7.

Write matrix \mathbf{X} for the model: $y_i = b_1x_i + b_3x_i^3 + \varepsilon_i$

Using Eq. (3.107) one obtains:

$$\mathbf{X} = \begin{bmatrix} x_1 & x_1^3 \\ x_2 & x_2^3 \\ x_3 & x_3^3 \\ \dots & \dots \\ x_N & x_N^3 \end{bmatrix} \quad (3.108)$$

3.14 Multiple linear regression

Multiple linear regression is an extension of a simple (univariate) linear regression ($y = b_0 + b_1x$) to more than one independent parameter (multivariable regression) in the form:

$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + \dots + b_px_p \quad (3.109)$$

An example of such problem is the study of the amounts of various chemicals included in the cement mixture on the amount of heat evolved in the curing of cement, where parameters x_i are the amounts of the components. This problem also appears in factorial design, multivariable calibration, etc. For one point Eq. (3.109) might be written as:

$$y_i = b_0 + b_1x_{1,i} + b_2x_{2,i} + b_3x_{3,i} + \dots + b_{p,i}x_{p,i} \quad (3.110)$$

The problem is easily solved writing Jacobian matrix \mathbf{X} , Eq. (3.107), for Eq. (3.110):

$$\mathbf{X} = \begin{bmatrix} 1 & x_{1,1} & x_{2,1} & \dots & x_{p,1} \\ 1 & x_{1,2} & x_{2,2} & \dots & x_{p,2} \\ 1 & x_{1,3} & x_{2,3} & \dots & x_{p,3} \\ 1 & \dots & \dots & \dots & \dots \\ 1 & x_{1,N} & x_{2,N} & \dots & x_{p,N} \end{bmatrix} \quad (3.111)$$

with the solution in the form of Eq. (3.90). An example of multiple regression will be shown in Section 5. There is also multivariate regression where several different y values are measured at different sets of $x_{i,j}$. This type of regression is used in chemometrics, but it is not accessible in Excel. This method will be presented in Part 2, Data analysis and modeling, Chemometrics.

3.15 Weighted least squares regression

It is possible that the experimental data used for the regression analysis are determined with different precision, that each value of y_i has different standard deviation, s_{y_i} . In such a case one should use weighted least-squares method. Let us define the diagonal matrix \mathbf{G}_y of the statistical weights, $w_i = 1/s_{y_i}^2$:

$$\mathbf{G}_y = \begin{bmatrix} 1/s_{y_1}^2 & 0 & 0 & 0 & 0 \\ 0 & 1/s_{y_2}^2 & 0 & 0 & 0 \\ 0 & 0 & 1/s_{y_3}^2 & 0 & 0 \\ 0 & 0 & 0 & \dots & 0 \\ 0 & 0 & 0 & 0 & 1/s_{y_N}^2 \end{bmatrix} \quad (3.112)$$

The problem described earlier by Eq. (3.88) must be modified by including the statistical weights:

$$\mathbf{X}'\mathbf{G}_y\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{G}_y\mathbf{Y} \quad (3.113)$$

which has solution:

$$\mathbf{b} = (\mathbf{X}'\mathbf{G}_y\mathbf{X})^{-1} \mathbf{X}'\mathbf{G}_y\mathbf{Y} \quad (3.114)$$

and

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b} = \mathbf{X}(\mathbf{X}'\mathbf{G}_y\mathbf{X})^{-1} \mathbf{X}'\mathbf{G}_y\mathbf{Y} = \mathbf{H}\mathbf{Y} \quad (3.115)$$

$$\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{G}_y\mathbf{X})^{-1} \mathbf{X}'\mathbf{G}_y$$

The variance/covariance matrix is in this case:

$$\mathbf{C}_b = (\mathbf{X}'\mathbf{G}_y\mathbf{X})^{-1} s_y^2 \quad (3.116)$$

with

$$s_y^2 = \frac{\boldsymbol{\varepsilon}'\mathbf{G}_y\boldsymbol{\varepsilon}}{N-2} \quad (3.117)$$

The matrix of the variances of \hat{y}_k is:

$$\mathbf{C}_{\hat{y}_k} = \mathbf{X}'_k \mathbf{C}_b \mathbf{X}_k = \mathbf{X}'_k (\mathbf{X}'\mathbf{G}_y\mathbf{X})^{-1} \mathbf{X}_k s_y^2 \quad (3.118)$$

and that of y_i :

$$C_{y_k} = \left[(1 + \mathbf{X}'_k (\mathbf{X}'\mathbf{G}_y\mathbf{X})^{-1} \mathbf{X}_k) \right] s_y^2 \quad (3.119)$$

Equations for weighted regression might be also developed as for linear regression replacing sum of squares in Eq. (3.7) by weighted sum of squares:

$$S^2 = \sum_i \varepsilon_i^2 = \sum_{i=1}^N \frac{(y_i - b_0 - b_1 x_i)^2}{s_{y_i}^2} = \sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{s_{y_i}^2} = \sum_{i=1}^N w_i (y_i - \hat{y}_i)^2 = \min \quad (3.120)$$

To determine regression parameters derivatives in Eq. (3.8) must be calculated:

$$\begin{aligned} \frac{\partial S^2}{\partial b_0} &= \sum 2(y_i - b_0 - b_1 x_i)(-w_i) = 0 \\ \frac{\partial S^2}{\partial b_1} &= \sum 2(y_i - b_0 - b_1 x_i)(-x_i w_i) = 0 \end{aligned} \quad (3.121)$$

leading to the system of two liner equations, analog of Eq. (3.11):

$$\begin{bmatrix} \sum w_i & \sum w_i x_i \\ \sum w_i x_i & \sum w_i x_i^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} = \begin{bmatrix} \sum w_i y_i \\ \sum w_i x_i y_i \end{bmatrix} \quad (3.122)$$

Solution of Eqs. (3.122) gives the values of parameters b_0 and b_1 , compare with Eqs. (3.12)-(3.13) :

$$\begin{aligned} b_0 &= \frac{\sum w_i x_i^2 \sum w_i y_i - \sum w_i x_i \sum w_i x_i y_i}{d} \\ b_1 &= \frac{\sum w_i \sum w_i x_i y_i - \sum w_i x_i \sum w_i y_i}{d} \end{aligned} \quad (3.123)$$

where d is now defined as:

$$d = \sum w_i \sum w_i x_i^2 - (\sum w_i x_i)^2 \quad (3.124)$$

Similarly, the standard deviations of the regression parameters are determined, as in Eqs. (3.23) and (3.27):

$$\begin{aligned} s_{b_0}^2 &= \frac{\sum w_i x_i^2}{d} s_y^2 \\ s_{b_1}^2 &= \frac{\sum w_i}{d} s_y^2 \end{aligned} \quad (3.125)$$

where s_y^2 is:

$$s_y^2 = \frac{\sum w_i (y_i - \hat{y}_i)^2}{N - 2} \quad (3.126)$$

These equations allow for easy calculation of the weighted regression in Excel. They also follow from the matrix notation above.

Example 3.8.

Find coefficients of the weighted regression using the following data.

x_i	y_i	s_i	w_i
0	1.9	0.4	6.25
1	2.3	0.5	4
2	3.5	0.7	2.040816327
3	4.5	0.9	1.234567901
4	5.2	1	1
5	6	1.2	0.694444444
6	5.5	1.1	0.826446281

This problem might be solved using Origin, using program *polfit.exe*, or in Excel. The results are shown below:

Parameter	Value	Standard Error	t -Value	95% LCL	95% UCL
Intercept	1.84859	0.17677	10.45743	1.39419	2.303
Slope	0.74423	0.07619	9.76859	0.54839	0.94008

Using t -test, the values for both parameters are larger than the critical value of $t(0.05, 5) = 2.447$, therefore both parameters are important. The fit and confidence intervals at the confidence level of 95% are displayed in Fig. 3.11.

Calculations in Excel using directly Eqs. (3.123)-(3.126) are also presented. See details in *Examples3.xlsx*, sheet *Ex. 3.8* and Origin *Ex3-8.opj*.

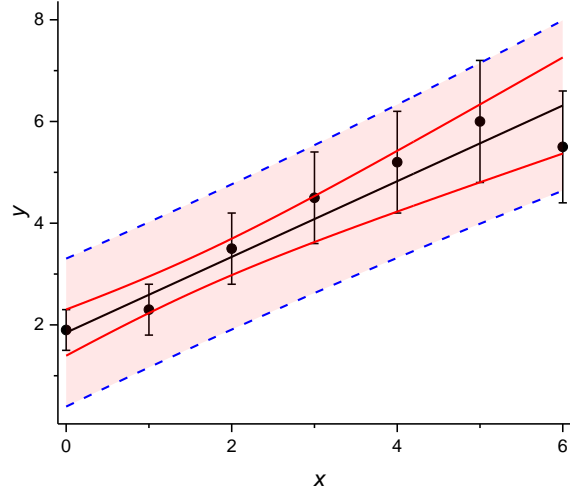


Fig. 3.11. Plot of the experimental points with their standard deviations, prediction line (black), confidence interval for the calculated values (red line) and for the experimental points (dashed blue) for weighed regression in Example 3.8 calculated in Origin (Ex3-8.opj).

3.16 Linear regression with errors in y and x

In all the developments until now we have assumed that the independent variable x is determined precisely that is its error is negligible. This variable is usually time, volume, temperature, etc., which might be determined precisely. However, there are cases where the independent variable is measured with certain error which cannot be neglected.^{25,26}

In such cases in estimation of the total variance, $s_{eff,i}^2$, we have to take into account two contributions: 1) that of y_i , $s_{y_i}^2$ and 2) influence of the variance of x on y , $(dy/dx)^2 s_{x_i}^2$ which gives the total effective variance, $s_{eff,i}^2$:

$$s_{eff,i}^2 = s_{y_i}^2 + (dy/dx)^2 s_{x_i}^2 \quad (3.127)$$

For the linear regression:

$$dy/dx = b_1 \quad (3.128)$$

and Eq. (3.127) becomes:

$$s_{eff,i}^2 = s_{y_i}^2 + b_1^2 s_{x_i}^2 \quad (3.129)$$

Taking into account that $w_i = 1/s_{eff,i}^2$ the problem is formally similar to the ordinary weighted regression Eq. (3.120) with solution in Eqs. (3.123)-(3.126), however, w_i depends on b_1 . This problem might be solved iteratively by minimization of the weighted sum of squares, Eq.

(3.117) or (3.120), $S^2 = \sum_{i=1}^N w_i (y_i - \hat{y}_i)^2$, assuming that in each iteration w_i is constant, calculating the new parameters b_0 and b_1 , \hat{y}_i , and new values of weights and repeating this procedure until reaching the minimum, i.e. parameters which do not change anymore and the sum of squares reaches minimum.²⁵

Example 3.9

Determine parameters and their standard deviations of the linear regression with error in x and y using data below.

x	y	s_y	s_x
0	1.9	0.4	0.1
1	2.3	0.5	0.1
2	3.5	0.7	0.2
3	4.5	0.9	0.3
4	5.2	1.0	0.4
5	6.0	1.2	0.5
6	5.5	1.1	0.6

This problem might be easily solved in Excel by minimizing the weighted sum of squares using Solver and, after finding the best values of regression parameters, calculating the standard deviations of the regression parameters using Eq. (3.125).²⁶

An example is presented in *Examples3.xlsx*, sheet *Ex. 3.9* Values of the parameters were calculated using Solver and Eq. (3.123).

3.17 Variances and covariances in error propagation

In development of the equations concerning error propagation, it was assumed that all the parameters were completely independent that is their covariances were zero. We can recall that in the calculations of the error propagation in linear regression equations were rearranged in order to have only one regression parameter. For example Eq. (3.32) $\hat{y}_i = b_0 + b_1 x_i$ was rearranged into Eq. (3.33) $\hat{y}_i = \bar{y} + b_1 (x_i - \bar{x})$. This was done to avoid problem with covariances. The error propagation for Eq. (3.33) should be written as:

$$s_{\hat{y}_i}^2 = \left(\frac{\partial \hat{y}_i}{\partial \bar{y}} \right)^2 \text{var}(\bar{y}) + \left(\frac{\partial \hat{y}_i}{\partial b_1} \right)^2 \text{var}(b_1) + 2 \left(\frac{\partial \hat{y}_i}{\partial \bar{y}} \right) \left(\frac{\partial \hat{y}_i}{\partial b_1} \right) \text{cov}(\bar{y}, b_1) \quad (3.130)$$

but as $\text{cov}(\bar{y}, b_1) = 0$, Eq. (3.34) is obtained. However, for Eq. (3.32) one obtains:

$$s_{\hat{y}_i}^2 = \left(\frac{\partial \hat{y}_i}{\partial b_0} \right)^2 \text{var}(b_0) + \left(\frac{\partial \hat{y}_i}{\partial b_1} \right)^2 \text{var}(b_1) + 2 \left(\frac{\partial \hat{y}_i}{\partial b_0} \right) \left(\frac{\partial \hat{y}_i}{\partial b_1} \right) \text{cov}(b_0, b_1) \quad (3.131)$$

It is obvious from Eqs. (3.97) and (3.99) that $\text{cov}(b_0, b_1) \neq 0$; it is equal, see Eqs. (3.97) and (3.99)

$$\text{cov}(b_0, b_1) = -\frac{\sum x_i}{NS_{xx}} s_y^2 = -\frac{\bar{x}}{\sum (x_i - \bar{x})^2} s_y^2 = -\frac{\bar{x}}{\sum x_i^2 - \frac{(\sum x_i)^2}{N}} s_y^2 \quad (3.132)$$

The variance of the calculated value of \hat{y}_i , Eq. (3.32), might be correctly evaluated using Eq. (3.131) using $\text{var}(b_0) = s_{b_0}^2$ and $\text{var}(b_1) = s_{b_1}^2$ from Eqs, (3.27) and (3.24):

$$\begin{aligned} s_{\hat{y}_i}^2 &= \frac{\sum x_i^2}{N} + x_i^2 - 2\bar{x} x_i \over \sum (x_i - \bar{x})^2} s_y^2 = \frac{\sum x_i^2}{N} + x_i^2 - 2\bar{x} x_i + \bar{x}^2 - \bar{x}^2 \over \sum (x_i - \bar{x})^2} s_y^2 \\ &= \frac{\sum x_i^2}{N} - \frac{(\sum x_i)^2}{N^2} + (x_i - \bar{x})^2 \over \sum (x_i - \bar{x})^2} s_y^2 = \frac{\frac{1}{N} \left(\sum x_i^2 - \frac{(\sum x_i)^2}{N} \right) + (x_i - \bar{x})^2}{\sum x_i^2 - \frac{(\sum x_i)^2}{N}} s_y^2 \\ &= \left[\frac{1}{N} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right] s_y^2 \end{aligned} \quad (3.133)$$

which is exactly Eq. (3.36). It should be kept in mind that a general form of Eq. (2.8) for function z of parameters p_i , $z = f(p_1, p_2, \dots, p_p)$ is:

$$\begin{aligned} s_z^2 &= \left(\frac{\partial f}{\partial p_1} \right)^2 \text{var}(p_1) + \left(\frac{\partial f}{\partial p_2} \right)^2 \text{var}(p_2) + \dots + \left(\frac{\partial f}{\partial p_n} \right)^2 \text{var}(p_n) \\ &+ 2 \left(\frac{\partial f}{\partial p_1} \right) \left(\frac{\partial f}{\partial p_2} \right) \text{cov}(p_1, p_2) + \dots + 2 \left(\frac{\partial f}{\partial p_1} \right) \left(\frac{\partial f}{\partial p_k} \right) \text{cov}(p_1, p_k) \\ &+ 2 \left(\frac{\partial f}{\partial p_2} \right) \left(\frac{\partial f}{\partial p_3} \right) \text{cov}(p_2, p_3) + \dots \end{aligned} \quad (3.134)$$

or in another form:

$$s_z^2 = \sum_{i=1}^n \left(\frac{\partial f}{\partial p_i} \right)^2 \text{var}(p_i) + 2 \sum_{i=1}^{n-1} \sum_{j=i+1}^n \left(\frac{\partial f}{\partial p_i} \right) \left(\frac{\partial f}{\partial p_j} \right) \text{cov}(p_i, p_j) \quad (3.135)$$

where summation runs over variances of all p parameters and all the covariances of each two parameters. The term “2” appears because $\text{cov}(x_i, x_j) = \text{cov}(x_j, x_i)$, therefore, the covariances appear twice for each two parameters.

3.18 Intersection of two straight lines

In analytical chemistry often one has to determine a parameter from the intersection of two straight lines.^{15,27} For example, the end point in conductometric or spectrophotometric titrations is

determined in that way. It is simple to determine the intersection from equations of straight lines, but the problem is with the standard deviation of the obtained concentration/quantity.

Assuming that there are two straight lines: $y = b_{0,1} + b_{1,1} x$ determined using N_1 points and $y = b_{0,2} + b_{1,2} x$ determined using N_2 points where the second index in regression parameters indicates the equation number. At the intersection both y and x for two lines are the same. From this condition one gets that the x -value (concentration) at the intersection, X , is:

$$X = -\frac{b_{0,2} - b_{0,1}}{b_{1,2} - b_{1,1}} = -\frac{\Delta b_0}{\Delta b_1} \quad (3.136)$$

First, Eq. (3.136) may be rearranged into:

$$\Delta b_0 + X \Delta b_1 = 0 \quad (3.137)$$

and to obtain the confidence limits for X the Fieller's theorem should be applied to Eq. (3.137)^{28,29} which gives:

$$\Delta b_0 + X \Delta b_1 = \pm t(\alpha, N_1 + N_2 - 4) s_\varepsilon \quad (3.138)$$

where variance s_ε^2 (left hand side of Eq. (3.138)) is:

$$s_\varepsilon^2 = s_{\Delta b_0}^2 + X^2 s_{\Delta b_1}^2 + 2X \text{cov}(\Delta b_0, \Delta b_1) \quad (3.139)$$

and taking square of Eq. (3.138) and substituting Eq. (3.139) the following relation is obtained

$$\Delta b_0^2 + 2X \Delta b_0 \Delta b_1 + X^2 \Delta b_1^2 = t^2 \left[s_{\Delta b_0}^2 + 2X \text{cov}(\Delta b_0, \Delta b_1) + X^2 s_{\Delta b_1}^2 \right] \quad (3.140)$$

It can be noticed that variances/covariances on the right correspond to the terms on left.

Pooling variances, see Eq. (1.31) of the two lines and assuming that variances of the two regression lines are similar, $s_{y_1}^2 \approx s_{y_2}^2$ gives the pooled variance, s_p^2 :

$$s_p^2 = \frac{S_1^2 + S_2^2}{N_1 + N_2 - 4} = \frac{(N_1 - 2)s_{y_1}^2 + (N_2 - 2)s_{y_2}^2}{N_1 + N_2 - 4} \quad (3.141)$$

Eq. (3.140) gives the following second order equation:

$$X^2 (\Delta b_1^2 - t^2 s_{\Delta b_1}^2) + 2X (\Delta b_0 \Delta b_1 - t^2 s_{\Delta b_0 \Delta b_1}) + (\Delta b_0^2 - t^2 s_{\Delta b_0}^2) = 0 \quad (3.142)$$

and the roots X_i are the lower and higher confidence intervals for X .

The variances and covariances are calculated from the error propagation equation. Variance of Δb_1 is calculated using Eq. (3.24):

$$s_{\Delta b_1}^2 = s_{b_{1,1}}^2 + s_{b_{1,2}}^2 = s_p^2 \left[\frac{1}{\sum_{i=1}^{N_1} (x_{i,1} - \bar{x}_1)^2} + \frac{1}{\sum_{i=1}^{N_2} (x_{i,2} - \bar{x}_2)^2} \right] \quad (3.143)$$

covariance $s_{\Delta b_0, \Delta b_1}$ using Eq. (3.132)

$$\text{cov}(\Delta b_0, \Delta b_1) = s_{\Delta b_0, \Delta b_1} = -s_p^2 \left[\frac{\bar{x}_1}{\sum_{i=1}^{N_1} (x_{i,1} - \bar{x}_1)^2} + \frac{\bar{x}_2}{\sum_{i=1}^{N_2} (x_{i,2} - \bar{x}_2)^2} \right] \quad (3.144)$$

and variance of Δb_1 using Eq. (3.27)

$$s_{\Delta b_0}^2 = s_p^2 \left[\frac{1}{N_1} + \frac{1}{N_2} + \frac{\bar{x}_1^2}{\sum_{i=1}^{N_1} (x_{i,1} - \bar{x}_1)^2} + \frac{\bar{x}_2^2}{\sum_{i=1}^{N_2} (x_{i,2} - \bar{x}_2)^2} \right] \quad (3.145)$$

The two roots of Eq. (3.142) describe the confidence intervals of the parameter X . Application of this method is illustrated in Example 3.10.

Example 3.10.

In the conductometric titration the following results were obtained,

x_A	y_A	x_B	y_B
4	0.7254	18	0.3929
6	0.6683	19	0.401
8	0.6089	20	0.4104
10	0.5522	22	0.4297
12	0.4964	24	0.4505
14	0.4406	26	0.4715
15	0.4152	28	0.4915
16	0.3961	30	0.513
17	0.3894	32	0.5349
		33	0.5449
		34	0.5554

where y is the analytical signal (e.g. conductivity or absorbance) and x is the volume, and the indices A and B correspond to the branches before and after the final point of titration. The titration curve is displayed in Fig. 3.12.

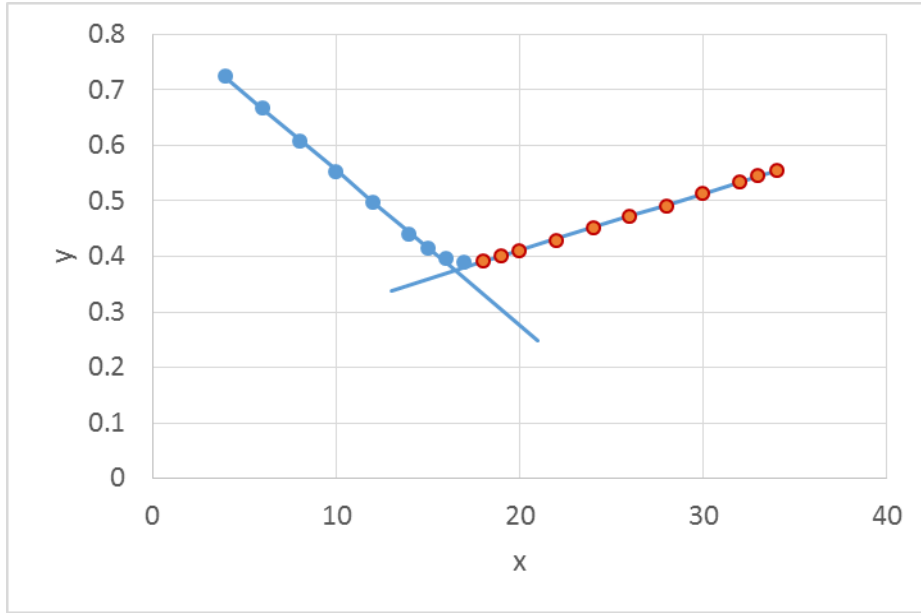


Fig. 3.12. Visualization of the data in Example 3.10 with the regression lines and the intersection in the final point.

Because of the curvature of the experimental relation points for $x = 4$ to 16 and 18 to 34 were selected to regression analysis. The parameters of the linear regressions of these two lines are:

A)

$$\begin{array}{ll} b_0 & 0.833525397 \\ b_1 & -0.027824743 \end{array}$$

B)

$$\begin{array}{ll} b_0 & 0.205452 \\ b_1 & 0.010265 \end{array}$$

Using Eq. (3.136) gives $X = 16.489$. The value of $t(0.05, 8+11-4) = 2.13145$, and using Eqns. (3.141)-(3.144) the following values are obtained:

$$s_p^2 = 8.44022 \times 10^{-6}$$

$$s_{\Delta b_0}^2 = 2.582 \times 10^{-5}, \quad s_{\Delta b_1}^2 = 8.80166 \times 10^{-8}, \quad s_{\Delta b_0, \Delta b_1} = -1.31911 \times 10^{-6} \text{ and the following}$$

equation for X :

$$0.00145041X^2 - 0.0478341X + 0.394359 = 0 \quad (3.146)$$

with two roots: $X_1 = 16.3499$ and $X_2 = 16.6297$ which are the lower and the higher confidence intervals. Therefore, with the probability of 95% the final result is: $X = 16.49 \pm 0.14$. See calculations in *Examples3.xlsx*, sheet *Ex. 3.8*.

3.19 Numerical problems related to the regression analysis

Polynomial regression analysis involves matrix inversion, Eq. (3.96), which is sensitive to numerical errors. To understand this problem let us consider example of the U.S. Census data. The statisticians tried to fit the U.S. population to a second order equation.³⁰ The data were collected every 10 years and are presented below in Example 3.11.

Example 3.11.

Fit the following census data to second order polynomial.

year (x)	population (y)
1900	75994575
1910	91972266
1920	105710620
1930	122775046
1940	131699275
1950	150697361
1960	179323175
1970	203235298

The fit to $y = b_0 + b_1x + b_2x^2$ was carried out using single (~7 significant digits) and double (~14 significant digits) precision on the IBM main frame computer. However, both methods gave two completely different sets of parameters which even had different signs.³⁰ This problem is related to the precision of the calculations and might be shown using **singular value decomposition**.^{30,31,32}

In general, matrix A can be factorized into three matrices:

$$A = \mathbf{u} \mathbf{w} \mathbf{v}' \quad (3.147)$$

where \mathbf{u} and \mathbf{v} are the orthonormal matrices (for which $\mathbf{u}' \mathbf{u} = \mathbf{v}' \mathbf{v} = \mathbf{I}$, where \mathbf{I} is the unit matrix) and \mathbf{w} is the diagonal matrix which contains ordered singular values $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_N \geq 0$.

The inverse of matrix A equals:

$$A^{-1} = \mathbf{v} \mathbf{w}^{-1} \mathbf{u}' \quad (3.148)$$

It should be noticed that for the orthonormal matrices \mathbf{u} and \mathbf{v} inversion is equivalent to transposition and for the diagonal matrix it is inversion of the diagonal values. It can also be added that the determinant of a diagonal matrix is simply a product of its singular values, $\det(\mathbf{w}) =$

$\sigma_1 \times \sigma_2 \times \dots \times \sigma_N = \prod_{i=1}^N \sigma_i$. Inversion of matrices involves division by the determinant. If its value

is very large (inverse very small) numerical problems arise and the problem is called ill-conditioned.

Let us look at the results obtained for our problem. \mathbf{X} matrix is:

$$X := \begin{bmatrix} 1 & 1900 & 3610000 \\ 1 & 1910 & 3648100 \\ 1 & 1920 & 3686400 \\ 1 & 1930 & 3724900 \\ 1 & 1940 & 3763600 \\ 1 & 1950 & 3802500 \\ 1 & 1960 & 3841600 \\ 1 & 1970 & 3880900 \end{bmatrix} \quad (3.149)$$

and $\mathbf{X}'\mathbf{X}$:

$$X'X := \begin{bmatrix} 8 & 15480 & 29958000 \\ 15480 & 29958000 & 57984984000 \\ 29958000 & 57984984000 & 112248125160000 \end{bmatrix} \quad (3.150)$$

Inversion of this matrix may introduce numerical inaccuracies. Let us look at the singular value decomposition of this matrix producing three matrices:

$$u := \begin{bmatrix} -2.66890836280178 \cdot 10^{-7} & -0.00103368019100357 & -0.999999465752453 \\ -0.000516578572539778 & -0.999999332325698 & 0.00103368019095306 \\ -0.999999866573245 & 0.000516578572438714 & -2.67086379283904 \cdot 10^{-7} \end{bmatrix} \quad (3.151)$$

$$w := \begin{bmatrix} 1.12248155113812 \cdot 10^{14} & 0 & 0 \\ 0 & 4195.74438234611 & 0 \\ 0 & 0 & 1.23470202728823 \cdot 10^{-7} \end{bmatrix} \quad (3.152)$$

$$v := \begin{bmatrix} -2.66890836371118 \cdot 10^{-7} & -0.00103368087498222 & -0.999999465751746 \\ -0.000516578572539820 & -0.999999332324991 & 0.00103368087493163 \\ -0.999999866573245 & 0.000516578572438535 & -2.67086732685257 \cdot 10^{-7} \end{bmatrix} \quad (3.153)$$

In the matrix w the ratio of the largest to the smallest singular values is:

$$\frac{\sigma_1}{\sigma_3} = 9.0911 \times 10^{20} \quad (3.154)$$

which is called **matrix condition number**. It is very large and during inversion the smallest singular value, $\sigma_3 = 1.2347 \times 10^{-7}$ becomes the largest! However, this value is calculated with the largest error because of the computer internal precision. The above results were obtained using Maple version 13, but the results obtained using Mathematica version 9 are a little different and $\sigma_3 = 1.19856 \times 10^{-7}$.

It should be noticed that all these problems are related to values of x and not to y . This problem might be easily avoided by scaling the values of x , for example introducing new values: $(x_i - \bar{x})/10$ where $\bar{x} = 1935$. Then, the matrix \mathbf{X} is:

$$X := \begin{bmatrix} 1 & -3.5 & 12.25 \\ 1 & -2.5 & 6.25 \\ 1 & -1.5 & 2.25 \\ 1 & -.5 & .25 \\ 1 & .5 & .25 \\ 1 & 1.5 & 2.25 \\ 1 & 2.5 & 6.25 \\ 1 & 3.5 & 12.25 \end{bmatrix} \quad (3.155)$$

and $\mathbf{X}'\mathbf{X}$

$$XTX := \begin{bmatrix} 8 & 0. & 42.00 \\ 0. & 42.00 & 0. \\ 42.00 & 0. & 388.5000 \end{bmatrix} \quad (3.156)$$

For this matrix the singular value decomposition gives \mathbf{w} :

$$w := \begin{bmatrix} 393.080856129105 & 0 & 0 \\ 0 & 42. & 0 \\ 0 & 0 & 3.41914387089504 \end{bmatrix} \quad (3.157)$$

and the condition number

$$\frac{\sigma_1}{\sigma_3} = 114.96 \quad (3.158)$$

is much smaller than the earlier value, Eq. (3.154), where the calculations might be carried out with large precision. The new version of Excel Regression must involve scaling because the results obtained using unscaled and scaled x values are essentially the same (large difference were obtained with older versions). See calculations in *Examples3.xlsx*, sheet *Ex. 3.11* and in Mathematica *SVD.nb* and Maple *svd.mw* files.

The professional programs use scaling of x values, e.g. between -2 and 2 for internal computing. Singular value decomposition is currently used in all problems demanding matrix

inversion. In ill conditioned problems truncated SVD is used^{31,32} where the smallest value(s) in \mathbf{w} are neglected and replaced by zero in \mathbf{w}^{-1} . In polynomial approximations orthogonal polynomials are usually used where addition of a higher order polynomial term does not affect the terms with lower orders.

3.20 Nonlinear regression

Nonlinear regression is linear in the parameters. Let us consider two functions:⁸

$$y = \exp(b_1 + b_2 x^2) \quad (3.159)$$

$$y = \frac{b_1}{b_1 - b_2} \left(e^{-b_2 x} - e^{-b_1 x} \right) \quad (3.160)$$

The first equation (3.159) might be linearized by taking logarithm:

$$\ln y = b_1 + b_2 x^2 \quad (3.161)$$

and it is called intrinsically linear. However, the second cannot be linearized and it is called intrinsically nonlinear.

Let us suppose that the postulated nonlinear model is presented in the form:

$$y = f(\mathbf{x}, \mathbf{b}) \quad (3.162)$$

where $\mathbf{x} = (x_1, x_2, \dots, x_N)$ is the independent variable and $\mathbf{b} = (b_1, b_2, \dots, b_p)$ are the unknown parameters. In the least-squares method the sum of squares is:

$$S^2 = \sum_i^N \varepsilon_i^2 = \sum_{i=1}^N [y_i - y(x_i, \mathbf{b})]^2 = \sum_{i=1}^N (y_i - \hat{y}_i)^2 = \min \quad (3.163)$$

see also Eq. (3.7). The minimum of the sum of squares occurs when its gradient versus parameters is zero:

$$\frac{\partial S^2}{\partial b_j} = -2 \sum_{i=1}^N [y_i - y(x_i, \mathbf{b})] \frac{\partial y(x_i, \mathbf{b})}{\partial b_j} = 0 \quad j = 1, \dots, p \quad (3.164)$$

The nonlinear method proceeds iteratively. First, the initial guess of p parameters b_i is chosen. These parameters are improved iteratively and new set of parameters b^{k+1} of $k+1$ iteration replaces values obtained in iteration k , that is parameter b_j^k is replaced by a new value:

$$b_j^{k+1} = b_j^k + \Delta b_j \quad \text{or} \quad \Delta b_j = b_j^{k+1} - b_j^k \quad (3.165)$$

The nonlinear model (3.162) is linearized using Taylor expansion in b and keeping only the first order terms:

$$f(x_i, \mathbf{b}^{(k+1)}) = f(x_i, \mathbf{b}^{(k)}) + \sum_{j=1}^p \frac{\partial f(x_i, \mathbf{b}^{(k)})}{\partial b_j} \Delta b_j \quad (3.166)$$

Introducing Jacobian \mathbf{J} , Eq. (3.107), Eq. (3.166) might be rewritten as:

$$\varepsilon_i = f\left(x_i, \mathbf{b}^{(k+1)}\right) - f\left(x_i, \mathbf{b}^{(k)}\right) - \sum_{j=1}^p J_{i,j} \Delta b_j$$

or

(3.167)

$$\varepsilon_i = \Delta y_i - \sum_{j=1}^p J_{i,j} \Delta b_j$$

where

$$\Delta y_i = y_i - \hat{y}_i(x_i, \mathbf{b}) \quad (1.168)$$

and matrix \mathbf{J} is the Jacobian, Eq. (3.107), written for parameters b_1 to b_p is:

$$\mathbf{J} = \begin{bmatrix} \frac{\partial y_1}{\partial b_1} & \frac{\partial y_1}{\partial b_2} & \frac{\partial y_1}{\partial b_3} & \dots & \frac{\partial y_1}{\partial b_p} \\ \frac{\partial y_2}{\partial b_1} & \frac{\partial y_2}{\partial b_2} & \frac{\partial y_2}{\partial b_3} & \dots & \frac{\partial y_2}{\partial b_p} \\ \frac{\partial y_3}{\partial b_1} & \frac{\partial y_3}{\partial b_2} & \frac{\partial y_3}{\partial b_3} & \dots & \frac{\partial y_3}{\partial b_p} \\ \dots & \dots & \dots & \dots & \dots \\ \frac{\partial y_N}{\partial b_1} & \frac{\partial y_N}{\partial b_2} & \frac{\partial y_N}{\partial b_3} & \dots & \frac{\partial y_N}{\partial b_p} \end{bmatrix} \quad (1.169)$$

Substitution into Eqs. (3.164) gives:

$$\sum_{i=1}^N J_{i,j} \left(\Delta y_i - \sum_{j=1}^p J_{i,j} \Delta b_j \right) = 0$$

or

(3.170)

$$\sum_{i=1}^N \sum_{m=1}^p J_{i,j} J_{i,m} \Delta b_m = \sum_{i=1}^N J_{i,j} \Delta y_i \quad \text{for } j = 1, \dots, p$$

which can be written in a matrix form:

$$\mathbf{J}' \mathbf{J} \Delta \mathbf{b} = \mathbf{J}' \Delta \mathbf{y} \quad (3.171)$$

This equation is an analog of Eq. (3.94) but written for Δ and the calculations are repeated until the differences are negligible. In the nonlinear problem the derivatives are usually calculated numerically using very small increments and iterations are continued until relative changes of parameters $\Delta b_j / b_j$ are very small. It is important to notice that if the initial guess of parameters is far from the system parameters the nonlinear least-squares method may diverge without producing results (singular matrix message). There are numerous programs which contain nonlinear regression (Origin, SigmaPlot) and source programs (e.g. in Netlib).^{5,833}

Example 3.12.

There are 1024 points (see Excel file *Examples3.xlsx*, sheet *Ex. 3.12*) and the possible model describing them is the Gauss function:

$$y = b_1 \exp \left\{ -[b_2(x - b_3)]^2 \right\} \quad (3.172)$$

Using Origin nonlinear fit with user defined function the plot of the experimental points and calculated function is presented in Fig. 3.13.

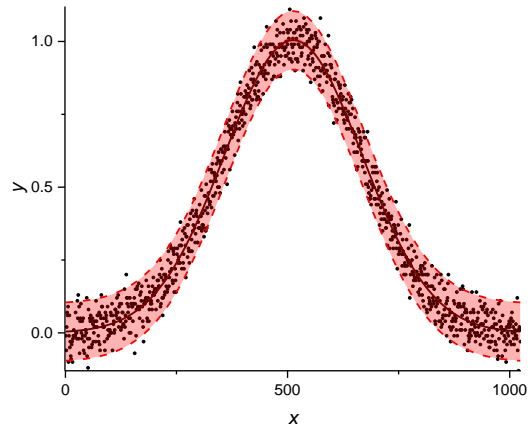


Fig. 3.13. Experimental point, prediction using Gaussian model (black line) and confidence intervals for y_i at the probability of 95%. Statistically, 5% of experimental points might be outside this confidence interval.

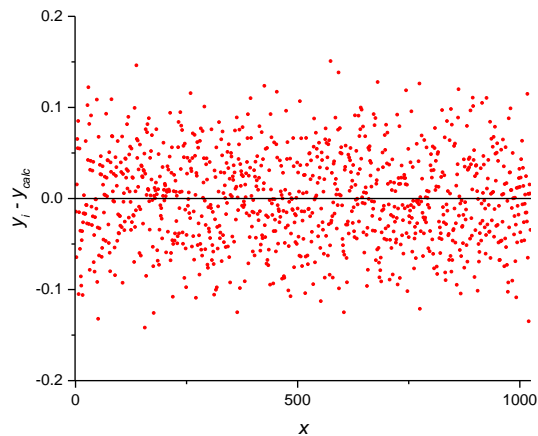


Fig. 3.14. Distribution of deviations of the experimental points from the calculated values. These deviations are randomly distributed.

The fit results are:

No	b	s_b	t_{exp}
1	1.0044	0.0038	266.41
2	0.004550	0.000020	229.51
3	512.71	0.67	762.53

All the parameters are statistically important and t_{exp} are very large. For $N = 1024$, $t(0.05, 1024-3) = 1.962$ (because there are many points this value is close to that for normal distribution, $z(0.975) = 1.960$) one can write that with the probability of 95% the following values of the parameters were found:

$$b_1 = 1.0044 \pm 0.0075$$

$$b_2 = 0.004550 \pm 0.000039$$

$$b_3 = 512.7 \pm 1.3$$

These calculations are shown in Origin *Ex3-120.opj* file, see also calculations in *Examples3.xlsx*, sheet *Ex. 3.12*.

Example 3.13.

Approximate the experimental points displayed in the Excel file *Examples3.xlsx*, sheet *Ex.3.13*. They are displayed in Fig. 3.15.

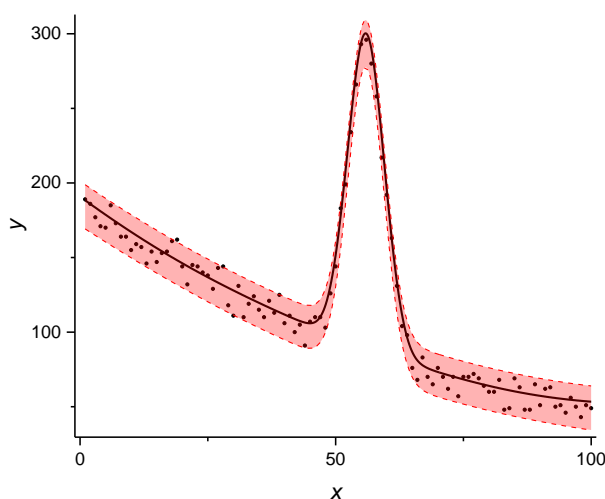


Fig. 3.15. Experimental points, approximation – black line, and confidence intervals for the experimental values for the confidence interval of 95% using model in Example 3.13.

The model proposed here is the Gaussian peak and the baseline approximated by a second order polynomial:

$$y = b_1 \exp \left[-0.5 \left(\frac{x - b_2}{b_3} \right)^2 \right] + b_4 + b_5 x + b_6 x^2 \quad (3.173)$$

The results obtained in Origin are:

b	s_b	t_{exp}
206.5	3.6182	57.0859
55.94	0.06946	805.3392
3.56	0.07631	46.58896
186.3	2.29131	81.31064

-2.30	0.11604	-19.7983
0.0093	0.00111	8.34281

All the parameters found are important and t_{exp} values are larger than $t(0.05, 94) = 1.985$.
With the probability of 95% the parameters found are:

$$\begin{aligned} b_1 &= 206.5 \pm 7.2 \\ b_2 &= 55.94 \pm 0.14 \\ b_3 &= 3.56 \pm 0.15 \\ b_4 &= 186.3 \pm 4.5 \\ b_5 &= -2.30 \pm 0.23 \\ b_6 &= 0.0093 \pm 0.0022 \end{aligned}$$

The calculations are shown in *Origin Ex3-13.opj* file, see also calculations in *Examples3.xlsx*, sheet *Ex. 3.13*. Such an analysis is used in chemistry and physics to determine the parameters of Gaussian peaks.

3.21 Dealing with nonlinear regression with errors in y and x

Linear regression with errors in x and y presented in Chapter 3.16 may be easily extended to nonlinear regression.²⁶ Its application will be illustrated using approximation of the van Deemter equation for gaseous chromatography using data from Moody.^{34,35} This problem may be easily solved using Excel Solver.^{26,35} It should be added that van Deemter equation is linear versus the unknown parameters:

$$y = b_1 x + b_2 / x + b_3 \quad (3.174)$$

where y is the plateau height, x is the flow rate and b_1 , b_2 , and b_3 are the unknown parameters which must be determined using weighted regression. The total effective variance $s_{\text{eff},i}^2$ is a sum of the variance of y_i and the influence of variance of x_i on the variance of y , see Eq. (3.127), $s_{\text{eff},i}^2 = s_{y_i}^2 + (dy/dx)^2 s_{x_i}^2$. The derivative dy/dx calculated from Eq. (3.174) is:

$$\frac{dy}{dx} = b_1 - \frac{b_2}{x^2} \quad (3.175)$$

and the weights $w_i = 1/s_{\text{eff},i}^2$. Starting from the first guess of the parameters b_1 , b_2 , and b_3 values of \hat{y}_i , w_i and the weighted sum of squares S^2 were calculated, and the procedure repeated until the minimum of the weighted sum of squares was found. Details are presented in *Example 3.14*.

Example 3.14

Data below describe dependence of the plateau height, y , vs. the flow rate, x . Approximate this data by van Deemter Eq. (3.174).

x	y
3.4	9.59
7.1	5.29
16.1	3.63

20.0	3.42
23.1	3.46
34.4	3.06
40.0	3.25
44.7	3.31
65.9	3.50
78.9	3.86
96.8	4.24
115.4	4.62
120.0	4.67

Assume standard deviations for x and y as proportional, 3% for x and 2% for y , that is $s_{x_i} = 0.03x_i$ and $s_{y_i} = 0.02y_i$. Starting from the first guess of the parameters b_1 , b_2 , and b_3 values of \hat{y}_i , w_i and the weighted sum of squares S^2 were calculated, and the procedure repeated until the minimum of the weighted sum of squares was found. This procedure was carried out in Excel using Solver. After finding the optimal values of the regression parameters it is possible to determine their standard deviations from the variance/covariance matrix, Eq. (3.116). First, the value of the variance/covariance matrix $(\mathbf{X}'\mathbf{GX})^{-1}$, Eq. (3.116), must be calculated from x_i , y_i , and weights w_i (weights calculated in each iteration from new parameters and \hat{y}_i). This can be carried out in Excel³⁶ or in any other program (Maple, Mathematica, Matlab). Example of the calculations is shown in *Examples3.xlsx*, sheet *Ex. 3.14*. The obtained results are:

$$\begin{array}{ll} b_1 = 0.0237 & s_{b1} = 0.0012 \\ b_2 = 26.2 & s_{b2} = 1.1 \\ b_3 = 1.628 & s_{b3} = 0.087 \end{array}$$

4 Statistical tests on average(s)

4.1 Introduction

The main purpose of the statistical data analysis is to estimate the parameters and carry out statistical tests on them. A **hypothesis** is a statement about experimental data. For example, let us assume that the average concentration of the sample is μ_0 . We want to prove, by repeating the determination N times if our average \bar{x} is statistically equal to μ_0 or there is a significant difference between these two values (bias).

We have to pose two hypotheses H_0 and H_1 about the data. These hypotheses are exclusives i.e. only one of them is true.

Null hypothesis, H_0 , states that the parameters compared (mean, variance) are the same or a value is zero, i.e. in the above example $\bar{x} - \mu_0 = 0$.

Alternative hypothesis, H_1 , states that the parameters compared are different.

Statistical test gives us a rule permitting to state which of two hypotheses is correct. This process is carried out with certain probability. One usually choses a **level of confidence (significance level), α** , in advance. This significance level shows a probability of rejecting the null hypothesis when it is true. Typically, these significance levels are 0.05 (or the probability of 95%) or 0.01 (probability 99%). Moreover, it is usually possible to calculate the **probability, p** , of H_0 being true. **If p is smaller than α one should reject the hypothesis H_0** . For example if $p < 0.05$ the probability of finding data consistent with H_0 is less than 5% and this hypothesis should be rejected at such confidence level that is there is less than one chance in 20 of finding values in accordance with H_0 . The smaller is p the less probable is hypothesis H_0 . In such a case we can state that the null hypothesis is rejected at the 95% level of confidence.

Carrying out such a procedure we can arrive at wrong conclusions (our conclusions are based on probability). There are two possible errors:

Type I error occurs when we reject hypothesis H_0 when it is true (that is we accept H_1 when it is false). It is called false positive.

Type II error occurs when hypothesis H_0 is accepted while it is false (that is we reject H_1 when it is true). It is also called false negative.

Types of decisions and errors are shown in Table 4.1.

Table 4.1. Type of decisions in hypotheses testing and possible error.

Decision		Situation	
		H_0 true	H_1 true
	Do not reject H_0	Good decision Probability $1 - \alpha$	Type II error Risk of error $\beta = ???$ False negative
	Reject H_0	Type I error Risk of error α False positive	Good decision

To better understand these errors let us look at the example of HIV testing. Two hypotheses are formulated:

H_0 - the patient does not have HIV

H_1 - the patient has HIV

A false positive is to reject H_0 and to accept H_1 when H_0 is true that is to say that patient has HIV while he has not.

A false negative is to accept H_0 and reject H_1 when H_1 is true that is to say the patient does not have HIV while it has.

Examples of the hypotheses testing for the mean(s) will be presented below.

4.2 Test χ^2

This test is commonly used to compare observed data with data we would expect to obtain according to a specific hypothesis. In other words it can be used to compare the observed sample distribution with the expected probable distribution. In our case we will apply this test to check if the statistical weights for the average are normally distributed.³⁷ The hypotheses tested are:

H_0 : errors are normally distributed

H_1 : errors are not distributed normally

To check this hypothesis one should compare the experimental (calculated) value of χ^2_{exp} :

$$\chi^2_{\text{exp}} = \sum_{i=1}^N \frac{(x_i - \bar{x})^2}{s_{x_i}^2} \quad (4.1)$$

with the theoretical (critical) value of this distribution $\chi^2_{\text{cr}}(\alpha, N-1)$. If $\chi^2_{\text{exp}} < \chi^2(\alpha, N-1)$ one should keep hypothesis H_0 at the confidence level of 0.05 i.e. probability of 95%.

Example 4.1.

Simulate χ^2 distribution function and its integral for $df = 2, 4$, and 8.

The probability distribution function χ^2 and $P_{\chi^2}(\chi^2, df)$ are shown in Fig. 4.1, and in Excel file *Examples4.xlsx*, sheet *Ex. 4.1* and *Origin Fig4-1.opj*.

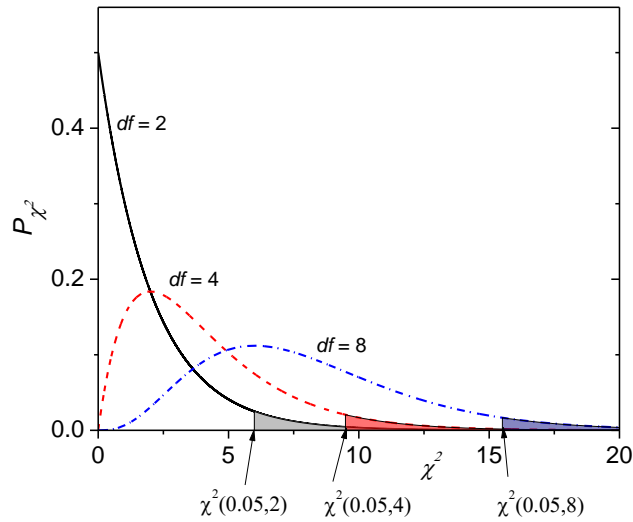


Fig. 4.1. χ^2 -probability distribution functions for 2, 4, and 8 degrees of freedom. The shaded area corresponds to the surface area of $\alpha = 0.05$ and the critical values of $\chi^2_{\text{cr}}(\alpha, df)$ are indicated with arrows.

This test will be applied to the determination of the weighted mean. The values of $P_{\chi^2}(\chi^2, df)$ are calculated in Excel using function: CHISQ.DIST(χ^2 , df , FALSE) and $\chi^2(\alpha, df)$ are calculated as: CHISQ.INV.RT(α , df).

Example 4.2.

Calculate the weighted mean and its standard deviation using the following data:

x_i	σ_i
2.10	0.20
2.30	0.20
2.15	0.22
1.95	0.32
4.00	0.20
3.90	0.21

Details of calculations are shown in *Examples4.xlsx*, sheet *Ex. 4.2*. The calculation are displayed below.

x_i	σ_i	w_i	$w_i x_i$	$w_i (x_i - x_m)^2$
2.1	0.20	25	52.5	13.14352
2.3	0.20	25	57.5	6.892718
2.15	0.22	20.66116	44.42149	9.415964
1.95	0.32	9.765625	19.04297	7.478169
4	0.20	25	100	34.51094
3.9	0.21	22.67574	88.43537	26.20076
sum		128.1025	361.8998	97.64207

$$\chi^2(0.05, 5) = 11.0705$$

$$\bar{x}_m = 2.825$$

$$s_{\bar{x}_m} = 0.088$$

One should notice that the experimental value of $\chi_{\text{exp}} = 97.64$ while $\chi_{\text{cr}}(0.05, 5) = 11.07$. Because the experimental value is much larger than the critical value at 95% confidence level, distribution of errors is not Gaussian. The results displayed in Fig. 4.2.

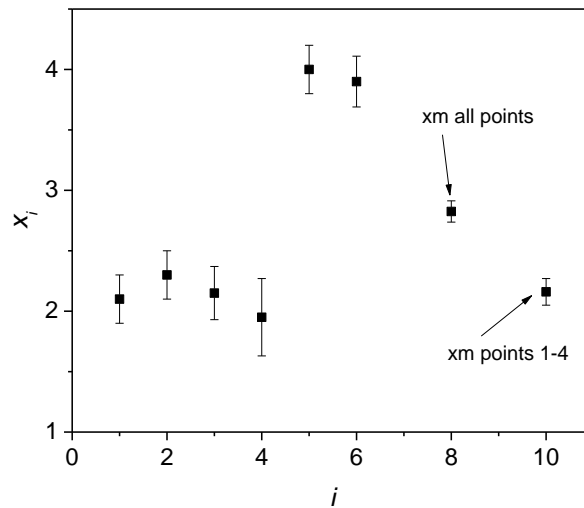


Fig. 4.2. Plot of the data in Example 4.2. The means and their standard deviations were calculated from all six points and from the first four points.

The average of all points: $\bar{x} = 2.825$ but the standard deviation of the mean, $s_{\bar{x}} = 0.088$, looks too small in comparison with the spread of the experimental points. We can also notice that the last two points have values much larger than the first four points. The calculations were repeated using only the first four points. The results obtained are shown below.

x_i	σ_i	w_i	$w_i x_i$	$w_i (x_i - \bar{x}_m)^2$
2.1	0.2	25	52.5	0.080655
2.3	0.2	25	57.5	0.512658
2.15	0.22	20.66116	44.42149	0.000955
1.95	0.32	9.765625	19.04297	0.417638
		80.42678	173.4645	1.011906
	chi2(0.05,3)=	7.814728		
	$\bar{x} =$	2.16		
	$s_{\bar{x}} =$	0.11		

In this case the experimental value $\chi_{\text{exp}} = 1.012$ is much smaller than $\chi_{\text{cr}}(0.05, 3) = 7.81$ and the distribution of errors is correct. The obtained result is: $\bar{x} = 2.16$ and $s_{\bar{x}} = 0.11$.

Example 4.3.

Calculate the weighted mean and its standard deviation using the following data:

x	σ
2.50	0.20
2.30	0.19
3.20	0.30
3.00	0.25
3.10	0.20

Carrying out calculations one can get the following results:

x_i	σ_i	w_i	$w_i x_i$	$w_i (x_i - \bar{x}_m)^2$	
2.50	0.20	25	62.5	1.449489	
2.30	0.19	27.70083	63.71191	5.382143	
3.20	0.30	11.11111	35.55556	2.343048	
3.00	0.25	16	48	1.075042	
3.10	0.20	25	77.5	3.225805	
	sum	104.8119	287.2675	13.47553	$= \chi^2_{\text{exp}}$
$\bar{x} =$	2.741				
$s_{\bar{x}} =$	0.098		$t(0.05, 4) =$	2.776445	
$\chi^2(0.05, 4) =$	9.487729				
corrected					
$s_{\bar{x}} =$	0.18				
CI =	0.50				

The following results are obtained: $\bar{x} = 2.741$, $s_{\bar{x}} = 0.098$, $\chi^2_{\text{exp}} = 13.48$ and $\chi^2_{\text{cr}}(0.05, 4) = 9.49$. In this case $\chi^2_{\text{exp}} > \chi^2_{\text{cr}}(0.05, 5)$, distribution of errors is not normal and the standard deviation is too small, see Fig. 4.3.

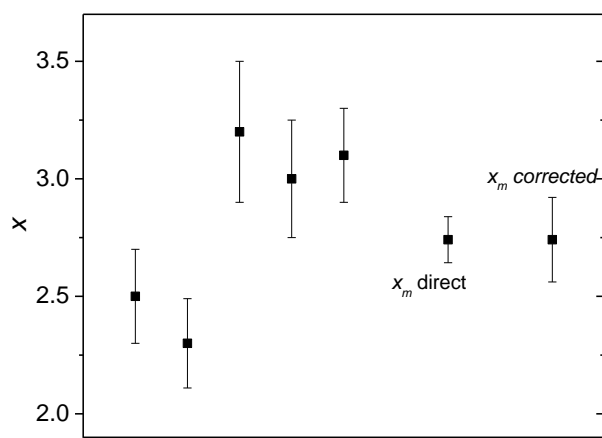


Fig. 4.3. Plot of the data and results in Example 4.3. The directly calculated standard deviation is too small and after correction for χ_{exp}^2 it is larger.

After correction for χ_{exp}^2 , Eq. (1.36), the standard deviation of the mean is $s_{\bar{x}} = 0.18$. The confidence intervals are:

$$\bar{x} = 2.74 \pm 0.50 \text{ using } t(0.05, 4) = 2.776.$$

See calculations in *Examples4.xlsx*, sheet *Ex. 4.3*.

4.3 Test for outliers, Dixon's Q-test

Outlier is a data point taken from a sample, assumed to be normally distributed, which lies beyond the mean at a stated probability. According to a statistical test it does not belong to the distribution of the rest of the data. However, points should not be rejected lightly. Good Manufacturing Practices (GMP) forbids such practices and rejecting data might lead to the process in court (although this is mainly in health sciences and testing production pharmaceutical). One should try to understand why such an outlier appeared. However, in the physicochemical sciences usually the experiments can be repeated to better understand the data distribution.

The oldest test for outliers is **Dixon's Q – test**.³⁸ It answers the question: should we keep this point in the calculation of the mean and standard deviation. It is used to quickly check for outliers. Two hypotheses should be posed:

H_0 this point should be kept

H_1 this point should be rejected

To use this test the points should be sorted from the smallest, x_1 , to the largest, x_N , and then the experimental value of Q_{exp} calculated:

$$Q_{\text{exp}} = \frac{\text{gap}}{\text{range}} = \frac{|x_1 - x_2|}{|x_N - x_1|} \text{ or } \frac{|x_N - x_{N-1}|}{|x_N - x_1|} \quad (4.2)$$

where gap is the absolute of difference between the suspected outlier and the closest number to it and gap is the difference between the maximal and minimal points of all data studied.

When $Q_{\text{exp}} > Q(P, N)$, where P is the probability in %, the point should be rejected. Eq. (4.2) may be used for number of points from 3 to 7. The values of $Q(P, N)$ are displayed in Table 4.2.

For larger N the formula should be modified calculating the gap and range omitting the extreme points and using different table values³⁹ but presently Grubbs test is recommended for such tests (see below). Caution should be used when working with small number of points. For example if three points are analyzed and two numbers are identical, e.g. 17.0, 17.0, 17.1 or 17.00, 17.00, 17.01, Q_{exp} test will always give the value of 1 forcing rejection although the spread of the results might be small. In such a case this test fails and more points should be acquired.

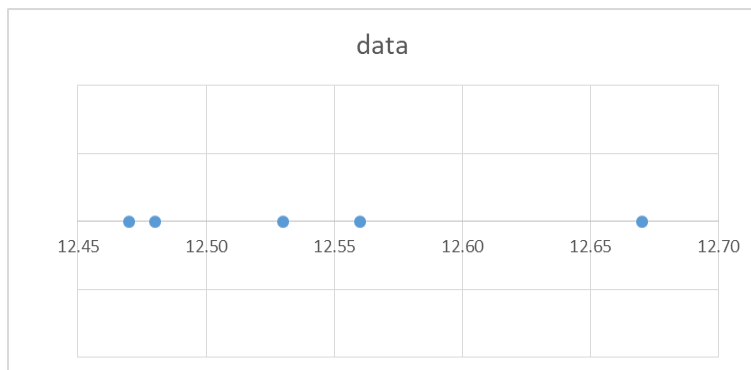
Table 4.2. Theoretical values of the Dixon's $Q(P, N)$ test for different levels of confidence, P : 90%, 95%, and 99%.

Number of observations N	Q (90%)	Q (95%)	Q (99%)
3	0.941	0.970	0.994
4	0.765	0.829	0.926
5	0.642	0.710	0.821
6	0.560	0.625	0.740
7	0.507	0.568	0.680
8	0.468	0.526	0.634
9	0.437	0.493	0.598
10	0.412	0.466	0.568

Example 4.4.

Is the greatest value in the following series an outlier?

The data are set from the smallest to the largest and the gap and range calculated:



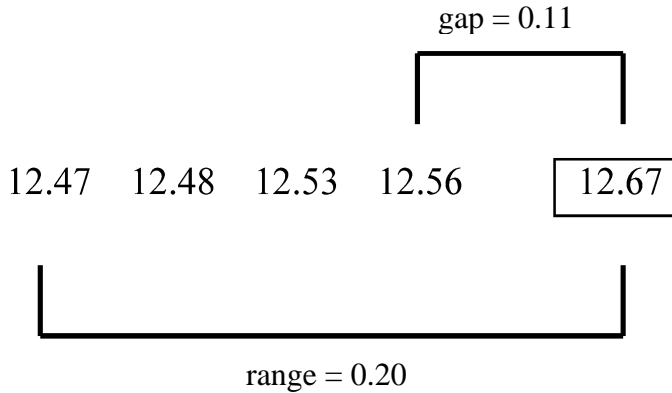


Fig. 4.4. Use of the Q – test.

The value of Q_{exp} is:

$$Q_{\text{exp}} = \frac{0.11}{0.50} = 0.55 \quad (4.3)$$

The value from the Table 4.2 is $Q_{\text{cr}}(95\%, 5) = 0.71$. Because $Q_{\text{exp}} < Q_{\text{cr}}(95\%, 5)$ this point should not be rejected.

Example 4.5.

Should the extreme values in the series be rejected?

5.00 5.10 5.10 5.15 5.20 5.30 6.20

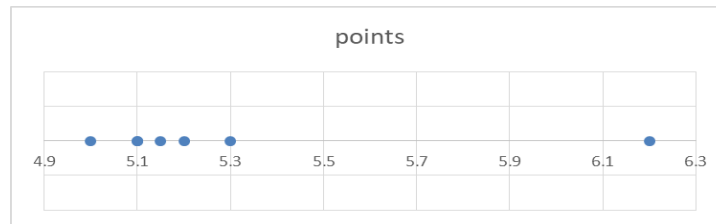


Fig. 4.5. Distribution of points in Example 4.5.

Calculations for the first and seventh point give the following results:

$$Q_1 = \frac{|5.00 - 5.10|}{|6.20 - 5.00|} = 0.083 \quad Q_7 = \frac{6.20 - 5.30}{6.20 - 5.00} = 0.75 \quad (4.4)$$

The value from the Table is: $Q(95\%, 7) = 0.57$ and $Q(99\%, 7) = 0.68$. The comparison shows that:

$Q_1 < Q_{\text{cr}}(95\%, 7)$ and $Q_7 > Q_{\text{cr}}(95\%, 7)$ which means that the first point should be kept and the seventh (6.20) rejected. The same answer is obtained for the probability of 99%. This test is included in the Origin program.

4.4 Test for outliers, Grubbs' G test

Grubbs' G' test⁴⁰ is recommended by ISO⁴¹ and ASTM⁴² and should be used instead of Q test although many analytical handbooks still recommend Q test. It is defined for deletion of one point at a time as:

$$G'_{\text{exp}} = \frac{x_{\text{max}} - \bar{x}}{s} \quad \text{or} \quad G'_{\text{exp}} = -\frac{x_{\text{min}} - \bar{x}}{s} \quad (4.5)$$

where s is sample standard deviation of the whole data set, including the suspected outlier and x_{max} and x_{min} are the suspected outliers. Critical values of one-sided test $G'(\alpha, N)$ are calculated using the following formula:

$$G'(\alpha, N) = \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2[(\alpha/N)', N-2]}{N-2+t^2[(\alpha/N)', N-2]}} \quad (4.6)$$

where one sided t -test was used. These values are easily calculated in Excel using T.INV(α/N , $N-2$) function, see the Excel file.

Similarly, the two sided t -test is also used to determine the largest absolute deviation from the mean as:

$$G'_{\text{exp}} = \frac{|x_{\text{suspected}} - \bar{x}|}{s} \quad (4.7)$$

where $x_{\text{suspected}}$ is x_{max} or x_{min} . To calculate the critical values of $G'(\alpha, N)$ the two sided values of T.INV.2T(α/N , $N-2$) should be computed using similar formula:

$$G'(\alpha, N) = \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2[(\alpha/N)", N-2]}{N-2+t^2[(\alpha/N)", N-2]}} \quad (4.8)$$

These values are shown in Table 4.3. If $G'_{\text{exp}} > G'_{\text{cr}}(\alpha, N)$ this point is an outlier and can be rejected. It should be noticed that although Grubbs⁴⁰, ASTM⁴², and Harris⁴³ recommend one-sided test, Hibbert and Gooding¹⁸, Thompson and Lowthian,⁴⁴ Brereton,⁴⁵ Ellison et al.⁴⁶ recommend using two-sided test. The two-sided test is also included in Origin.

Table 4.3. Critical values of the Grubbs' test for outliers, $G(\alpha, N)$.

one-sided				two sided			
$\alpha =$	0.1	0.05	0.01	$\alpha =$	0.1	0.05	0.01
N				N			
3	1.1484	1.1531	1.1546	3	1.1531	1.1543	1.1547
4	1.4250	1.4625	1.4925	4	1.4625	1.4813	1.4963
5	1.6016	1.6714	1.7489	5	1.6714	1.7150	1.7637
6	1.7289	1.8221	1.9442	6	1.8221	1.8871	1.9728
7	1.8280	1.9381	2.0973	7	1.9381	2.0200	2.1391
8	1.9089	2.0317	2.2208	8	2.0317	2.1266	2.2744
9	1.9773	2.1096	2.3231	9	2.1096	2.2150	2.3868
10	2.0362	2.1761	2.4097	10	2.1761	2.2900	2.4821
12	2.1341	2.2850	2.5494	12	2.2850	2.4116	2.6357

14	2.2132	2.3717	2.6585	14	2.3717	2.5073	2.7554
16	2.2793	2.4433	2.7470	16	2.4433	2.5857	2.8521
20	2.3853	2.5566	2.8838	20	2.5566	2.7082	3.0008
30	2.5651	2.7451	3.1029	30	2.7451	2.9085	3.2361
40	2.6840	2.8675	3.2395	40	2.8675	3.0361	3.3807
50	2.7719	2.9570	3.3366	50	2.9570	3.1282	3.4825

These values were calculated in Excel file *Examples4.xlsx*, sheet *Grubbs Table*.

The problem described earlier discussing limited number of measurements when two points are identical and the third different also exists for Grubb's test (e.g. 17.0, 17.0, 17.1 or 17.0, 17.0, 18.0 give the same G –test values) and more data should be acquired or the standard deviation analyzed more closely before possible rejection.

The above one or two sided G' test was for the rejection of one extremal point. There are two more tests for rejection of two points, G'' for rejection of two extremal points, x_1 and x_N and G''' for rejection of two points on the same side, that is x_1 and x_2 or x_{N-1} and x_N .

Test G'' is defined as:

$$G'' = \frac{x_N - x_1}{s} \quad (4.9)$$

where the critical values are calculated using the following equation: ⁴⁷

$$Q''(\alpha, N) = \sqrt{\frac{2(N-1)[t'(\alpha', k)]^2}{k + [t'(\alpha', k)]^2}} \quad (4.10)$$

where

$$\alpha' = \frac{\alpha}{N(N-1)} \quad \text{and} \quad k = N - 2$$

Similarly as for G' , when $G''_{\text{exp}} > G''_{\text{cr}}(\alpha, N)$ the two extreme pints may be rejected. Experimental values of test G''' are calculated using standard deviation of all points, s , and standard deviation excluding these two extremal points:

$$G''_{\text{low}} = \frac{(N-3)s_{\text{excluding 2 lowest}}^2}{(N-1)s^2} \quad \text{or} \quad G''_{\text{high}} = \frac{(N-3)s_{\text{excluding 2 highest}}^2}{(N-1)s^2} \quad (4.11)$$

In this case when $G'''_{\text{exp}} < G'''_{\text{cr}}(\alpha, N)$ the pair of points may be rejected because G''' becomes smaller as the suspected outliers become more extreme. The critical values of G'' and G''' are presented in Table 4.4.

Table 4.4. Critical values of the Grubbs' test for outliers, $G''(\alpha, N)$ and $G'''(\alpha, N)$.

$\alpha =$ N	$G''(\alpha, N)$		$G'''(\alpha, N)$	
	0.05	0.01	0.05	0.01
3	1.999	2.000	—	—
4	2.429	2.445	0.0002	0.0000
5	2.755	2.803	0.0090	0.0018
6	3.012	3.095	0.0349	0.0116
7	3.222	3.338	0.0708	0.0308
8	3.399	3.543	0.1101	0.0563
9	3.552	3.720	0.1492	0.0851
10	3.685	3.875	0.1864	0.1150
11	3.803	4.012	0.2213	0.1448
12	3.909	4.134	0.2537	0.1738
13	4.005	4.244	0.2836	0.2016
14	4.093	4.344	0.3112	0.2280
15	4.173	4.435	0.3367	0.2530
16	4.247	4.519	0.3603	0.2767
17	4.316	4.597	0.3822	0.2990
18	4.380	4.669	0.4025	0.3200
19	4.440	4.737	0.4214	0.3398
20	4.496	4.800	0.4391	0.3585
21	4.549	4.859	0.4556	0.3761
22	4.599	4.914	0.4711	0.3927
23	4.646	4.967	0.4857	0.4085
24	4.691	5.017	0.4994	0.4234
25	4.734	5.064	0.5123	0.4376
26	4.775	5.109	0.5245	0.4510
27	4.814	5.151	0.5360	0.4638
28	4.851	5.192	0.5470	0.4759
29	4.886	5.231	0.5574	0.4875
30	4.921	5.268	0.5672	0.4985
40	5.201	5.571	0.6445	0.5862
50	5.407	5.790	0.6966	0.6462
60	5.568	5.960	0.7343	0.6901
70	5.700	6.098	0.7630	0.7236
80	5.811	6.213	0.7856	0.7501
90	5.906	6.311	0.8040	0.7717
100	5.990	6.397	0.8192	0.7896

Values of G'' calculated in Excel file *Examples 4* sheet *Grubbs Table*, and values of G''' are from refs. 40 and 46.

In any case, one should be very cautious while rejecting a point from a small group of points, e.g. three points. Although the values are presented for number of points starting from 3, some authors recommend using it starting from 4 points. It is recommended⁴⁶ that when one outlier is found, one should not proceed with the two-point tests until the origin of this outlier was studied.

Example 4.6.

Use two-tailed Grubbs' test for the data in Example 4.4.

$$\bar{x} = 12.542, s = 0.0804$$

$$G'_{\text{exp}} = \frac{|12.67 - 12.542|}{0.0804} = 1.591 \quad (4.12)$$

$$G'(\alpha, N)_{\text{exp}} = 1.591 < G_{\text{cr}}(0.05, 5) = 1.715 \quad (4.13)$$

Therefore there are no reasons to reject this point. See calculations in *Examples4.xlsx*, sheet *Ex. 4.6*.

Example 4.7.

Use two-tailed Grubbs' test for the data in Example 4.5.

$$\bar{x} = 5.2929, s = 0.4107$$

$$G_{\text{exp}} = \frac{|5.2929 - 6.2|}{0.4107} = 2.209 \quad (4.14)$$

$$G'_{\text{exp}} = 2.209 > G'_{\text{cr}}(0.05, 7) = 2.020 \quad (4.15)$$

and $G_{\text{exp}} > G'_{\text{cr}}$, therefore this point ($x = 6.20$) should be rejected. This test is also included in the Origin program. See calculations in *Examples4.xlsx*, sheet *Ex. 4.7*.

Example 4.8.

Use the tests G' , G'' , and G''' to check for one or two outliers using the following data:
20.6 21.7 23.0 23.0 24.3 28.0 36.5

First calculate $\bar{x} = 25.3$ and $s = 5.4675$ (see Excel file). Then,

$$G'_{\text{exp}} = \frac{36.5 - 25.3}{5.4675} = 2.05 \quad (4.16)$$

$$G'_{\text{cr}}(0.05, 7) = 2.02$$

and because $G'_{\text{exp}} > G'_{\text{cr}}(0.05, 7)$ point 36.5 should be rejected as an outlier.

For the rejection of two extreme points (20.6 and 36.5), Eq. (4.9):

$$G''_{\text{exp}} = \frac{36.5 - 20.6}{5.4675} = 2.91 \quad (4.17)$$

$$G''_{\text{cr}}(0.05, 7) = 3.222$$

Because $G''_{\text{exp}} < G''_{\text{cr}}(0.05, 7)$ there is no reason to reject two extreme points x_1 and x_7 .

To calculate G''' first the standard deviation of the first 5 points (without 28.0 and 36.5) should be calculated. $s_{\text{excluding 2 highests}} = 1.41315$. Then, G''' is, Eq. (4.11):

$$G_{\text{exp}}''' = \frac{(7-3) \times 1.41315^2}{(7-1) \times 5.4675^2} = 0.0445 \quad (4.18)$$

$$G'''(0.05, 7) = 0.1101$$

In this case $G_{\text{exp}}''' < G_{\text{cr}}'''(0.05, 7)$ and two largest points might be rejected.

4.5 p -level test

In statistical analysis a specific level of significance at which the test might be rejected is calculated.³⁷ This probability is called p -level or p -probability. It should be compared with the confidence level, α . If $p < \alpha$ the hypothesis H_0 should be rejected and when $p > \alpha$ it should be kept. This p values are included in professional software, e.g. Minitab and also calculated automatically by Excel in several tests. It will be shown below how these values are calculated when discussing specific tests.

4.6 Test u

Test u is used to compare the experimental mean value, \bar{x} , with the true value, μ , when the standard deviation of the population, σ_x , is known. This might happen during massive production allowing determination of the standard deviation of the population by pooling all the data.

There are two hypotheses:

$$H_0: \mu_x = \mu \text{ (or } \mu_x - \mu = 0)$$

$$H_1: \mu_x \neq \mu \text{ (or } \mu_x - \mu \neq 0)$$

Let us use two-tailed test. In this case one should compare value of u

$$u = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\sigma_x} \sqrt{N} \quad (4.19)$$

with $|z(\alpha)|$ of the normal distribution. If $u < |z_{\text{cr}}(\alpha/2)|$ there are no reasons to reject H_0 and the experimental mean is statistically equal to the true value at the level of confidence α . p -value of this test is:

$$p = 2[1 - \text{NORM.S.DIST}(|u|, \text{TRUE})] \quad (4.20)$$

If $p < \alpha$ H_0 should be rejected. Function NORM.S.DIST($|u|$, TRUE) calculates the integral of the Gaussian probability function:

$$\int_{-\infty}^{|u|} P_G(z, 0, 1) dz = \int_{-\infty}^{|u|} \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \quad (1.21)$$

and p is the two-sided probability, Fig. 4.6.

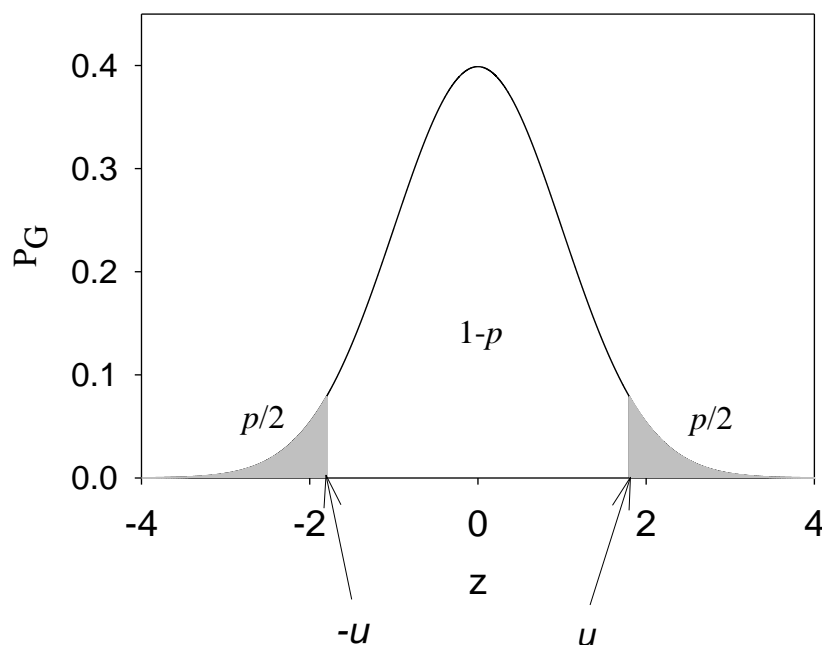


Fig. 4.6. Calculation of the probability p by the integration of the Gauss probability function.

Example 4.9.

The manufacturer produces volumetric flasks of $V_0 = 100.0$ ml. To check if the production goes well a volume of a sample flask was measured 5 times and the following results were obtained: 99.89, 100.42, 100.11, 99.96, 100.33 ml. The standard deviation of the population (determined previously on a large sample of data) is $\sigma_x = 0.2$ ml. Can one say that the volume of the flask is 100.0 ml at the confidence level of 0.05? Use the two-tailed test.

One should formulate two hypotheses:

$$H_0: \bar{V} = 100.0 \text{ ml}$$

$$H_1: \bar{V} \neq 100.0 \text{ ml}$$

$$\bar{V} = 100.142 \text{ ml}$$

$$u_{\text{exp}} = \frac{100.142 - 100.00}{0.2} \sqrt{5} = 1.59 \quad (4.22)$$

$$z_{\text{cr}}(0.975) = |z(0.025)| = 1.96$$

In this case the total probability is 95% or $\alpha = 0.05$ on both sides of the Gaussian curve. Because it is the two-tailed test the values of $|z(0.025)| = z(0.975)$ should be used. In this Example $u = 1.59 < z_{\text{cr}}(0.975) = 1.96$ and one can say that the volume is (statistically) equal to 100.0 ml. There are 5 chances in 100 (one in 20) that the volume is not 100.0 ml. Besides $p = 0.112 > \alpha = 0.05$ and H_0 should not be rejected. See *Examples4.xlsx*, sheet *Ex. 4.9*.

4.7 Test t , comparison with the standard

This test is used to compare the mean value, \bar{x} , with the true value, μ , when the standard deviation of the population, σ_x , is unknown. In this case one can estimate only the standard deviation of the sample, s_x (not population). The test below is two-tailed. The null and alternative hypotheses are:

$$H_0: \bar{x} = \mu$$

$$H_1: \bar{x} \neq \mu$$

and one should compare value of t_{exp} defined as:

$$t_{\text{exp}} = \frac{|\bar{x} - \mu|}{s_{\bar{x}}} = \frac{|\bar{x} - \mu|}{s_x} \sqrt{N} \quad (4.23)$$

with the value of $t_{\text{cr}}(\alpha'', df) = \text{T.INV.2T}(\alpha, df)$ where $df = N - 1$. When $t_{\text{exp}} < t_{\text{cr}}(\alpha'', df)$ there are no reasons to reject the hypothesis. p -value for this test is calculated in Excel as:

$$p = \text{T.DIST.2T}(t_{\text{exp}}, df) \quad (4.24)$$

Example 4.10.

The true value is $\mu = 0.123$. Experimental measurements gave the following results: 0.112, 0.118, 0.115, and 0.119. Can one say that the mean of the experimental measurements is statistically equal to the true value at the confidence levels of 0.05 and 0.01? Use the two-tailed test.

$$\bar{x} = 0.116, s_{\bar{x}} = 0.00158 \text{ (using Descriptive Statistics)}$$

$$t_{\text{exp}} = \frac{|0.116 - 0.123|}{0.00158} = 4.43 \quad (4.25)$$

From Excel, $t_{\text{cr}}(0.05'', 3) = 3.18$ and $t_{\text{cr}}(0.01'', 3) = 5.84$. That is at the confidence level of 95% ($\alpha = 0.05''$) $t_{\text{exp}} = 4.43 > t_{\text{cr}}(0.05'', 3) = 3.18$ and the hypothesis H_0 should be rejected which means that the mean value is different from the true value at the probability of 95%.

However, at the confidence level of 99% ($\alpha = 0.01''$) $t_{\text{exp}} = 4.43 < t_{\text{cr}}(0.01'', 3) = 5.84$ and the hypothesis H_0 cannot be rejected which means that if we assume that there is only one chance in 100 that the mean is not equal to the real value we have to accept our experimental \bar{x} . The same conclusions are obtained using p -values. The calculated $p = 0.0214$ is lower than $\alpha = 0.05$ (reject H_0) but larger than $\alpha = 0.01$ (keep H_0). The calculations are in the Excel sheet *Ex4.10* in the file *Examples4.xlsx*.

Example 4.11.

Taking data from Example 1.6 and assuming that $\mu = 0.08$ answer the following questions:

- using two-tailed t -test is the mean value different from μ ?
- using right one-tailed t -test is the mean value larger than μ ?
- using left one-tailed t -test is the mean value smaller than μ ?

Repeat all these tests using test u and assuming that the standard deviation of the population is known $\sigma_x = 0.005$.

All the calculations are shown in *Examples 4*, sheet *Ex. 4.11*.

t-tests

Re. a).

We have to test the following hypotheses:

$$H_0 \quad \bar{x} = \mu$$

$$H_1 \quad \bar{x} \neq \mu$$

The experimental t_{exp} value is:

$$t_{\text{exp}} = \frac{|\bar{x} - \mu|}{s_{\bar{x}}} = \frac{|0.0840 - 0.080|}{0.002887} = 1.386 \quad (4.26)$$

the critical value t_{cr} :

$$t_{\text{cr}}(\alpha'', df) = t_{\text{cr}}(0.05'', 2) = \text{T.INV.2T}(0.05, 2) = 4.303 \quad (4.27)$$

and the probability p :

$$p = \text{T.DIST.2T}(t_{\text{exp}}, df) = \text{T.DIST.2T}(1.386, 2) = 0.300$$

Student P_S distribution function for $df = 2$ is displayed in Fig. 4.7. The surface of the shaded area outside $\pm t_{\text{cr}}$ is $\alpha = 0.05$ (0.025 on each side). The experimental value of $t_{\text{exp}} = 1.386$ and the surface area outside $\pm t_{\text{exp}}$ is $p = 0.300$. In this case $t_{\text{exp}} < t_{\text{cr}}$ and $p > \alpha$ therefore H_0 should not be rejected and the mean is not different from the true value at the confidence level of 0.05 (or 95%).

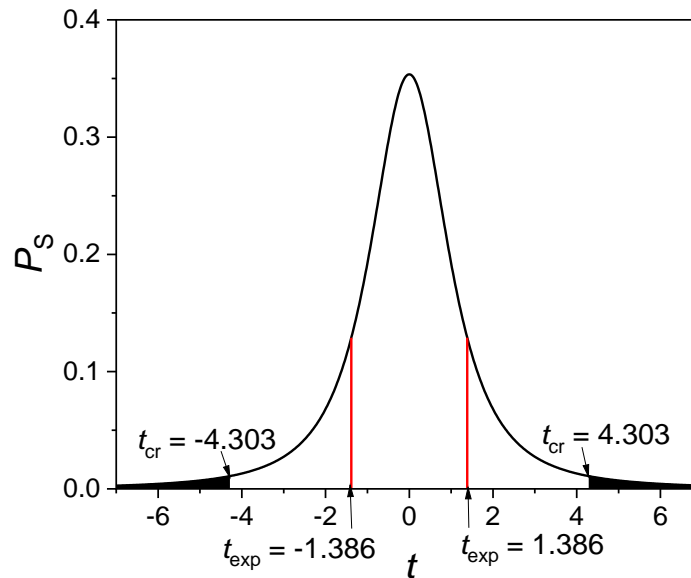


Fig. 4.7. Student distribution function for $df = 2$ (continues black line), the black surface area outside $\pm t_{\text{cr}}$ is $\alpha = 0.05$ and the surface area outside $\pm t_{\text{exp}}$ is $p = 0.300$.

Re. b)

We have to test the following hypotheses:

$$H_0 \quad \bar{x} = \mu$$

$$H_1 \quad \bar{x} > \mu$$

The experimental t_{exp} value is:

$$t_{\text{exp}} = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{0.0840 - 0.080}{0.002887} = 1.386 \quad (4.28)$$

the critical value $t_{\text{cr}}(\alpha', df)$ for the right one-tailed test is:

$$t_{\text{cr}}(\alpha', df) = t_{\text{cr}}(0.05', 2) = \text{T.INV}(0.95, 2) = 2.920 \quad (4.29)$$

and p -value is:

$$p = \text{T.DIST.RT}(t_{\text{exp}}, df) = \text{T.DIST.RT}(1.386, 2) = 0.150 \quad (4.30)$$

The Student probability distribution function is displayed in Fig. 4.8.

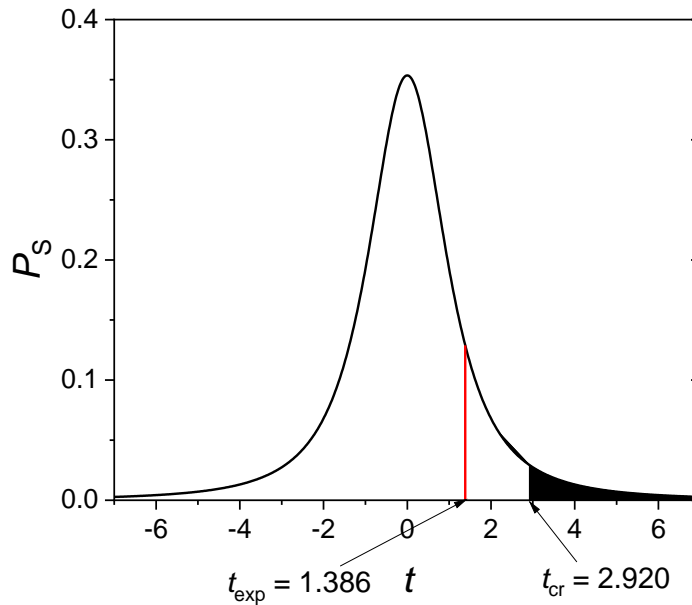


Fig. 4.8. Student distribution function for $df = 2$ (continuous black line), the black surface area right to t_{cr} is $\alpha = 0.05$ and the surface area right to t_{exp} is $p = 0.15$.

The surface of the shaded area right of t_{cr} is $\alpha = 0.05$. The experimental value of $t_{\text{exp}} = 1.386$ and the surface area right to t_{exp} is $p = 0.150$. In this case $t_{\text{exp}} < t_{\text{cr}}$ and $p > \alpha$ therefore H_0 should not be rejected and the mean is not larger than the true value at the confidence level of 0.05 (or 95%).

Re. c)

We have to test the following hypotheses:

$$H_0 \quad \bar{x} = \mu$$

$$H_1 \quad \bar{x} < \mu$$

Of course because numerically $\bar{x} > \mu$ we already know it cannot be smaller. This test is shown only to demonstrate how to carry out the calculations.

The experimental t_{exp} value is:

$$t_{\text{exp}} = \frac{\bar{x} - \mu}{s_{\bar{x}}} = \frac{0.0840 - 0.080}{0.002887} = 1.386 \quad (4.31)$$

the critical value $t_{\text{cr}}(\alpha', df)$ for left one-tailed test is:

$$t_{cr}(\alpha', df) = t_{cr}(0.05', 2) = T.INV(0.05, 2) = -2.920 \quad (4.32)$$

and p -value is:

$$\begin{aligned} p &= T.DIST(t_{exp}, df, TRUE) = 1 - T.DIST.RT(t_{exp}, df) \\ &= T.DIST.(1.386, 2, TRUE) = 0.850 \end{aligned} \quad (4.33)$$

The Student probability distribution function is displayed in Fig. 4.8.

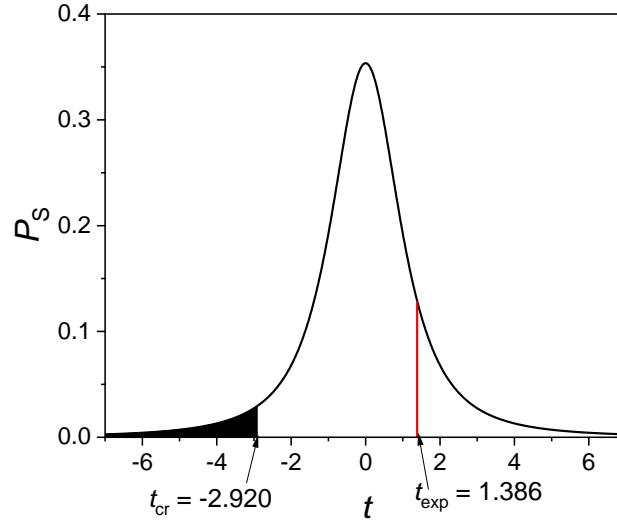


Fig. 4.9. Student distribution function for $df = 2$ (continues black line), the black surface area left of t_{cr} is $\alpha = 0.05$ and the surface left of t_{exp} is $p = 0.85$.

The surface of the shaded area left of t_{cr} is $\alpha = 0.05$. The experimental value of $t_{exp} = 1.386$ and the surface area left to t_{exp} is $p = 0.850$. In this case $t_{exp} > t_{cr}$ (left-tailed t -test) and $p > \alpha$ therefore H_0 should not be rejected and the mean is not smaller than the true value at the confidence level of 0.05 (or 95%).

In general, one can state that if t_{exp} is outside the shaded area defined by t_{ct} hypothesis H_0 cannot be rejected.

u-tests

Re. a)

For two-tailed test we have to examine the following hypotheses:

$$H_0 \quad \bar{x} = \mu$$

$$H_1 \quad \bar{x} \neq \mu$$

The standard deviation of the mean is:

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}} = \frac{0.05}{\sqrt{3}} = 0.002887 \quad (4.34)$$

The experimental u_{exp} value is:

$$u_{exp} = \frac{|\bar{x} - \mu|}{\sigma_{\bar{x}}} = \frac{|0.0840 - 0.080|}{0.002887} = 1.386 \quad (4.35)$$

the critical value z_{cr} :

$$z_{cr}(\alpha'') = z_{cr}(0.05'') = |NORM.S.INV(0.05/2)| = NORM.S.INV(1-0.05/2) = 1.960 \quad (4.36)$$

and the probability p :

$$p = 2(1 - \text{NORM.S.DIST}(z_{\text{exp}}, \text{TRUE})) = 2(1 - \text{NORM.S.DIST}(1.386, \text{TRUE})) = 0.166 \quad (4.37)$$

In the above case $u_{\text{exp}} < z_{\text{cr}}$ and $p > \alpha$ therefore hypothesis H_0 cannot be rejected, and one can say that at the confidence level of 0.05 (95%) the experimental mean is statistically equal to the true value.

Re. b)

For right one-tailed test we have to examine the following hypotheses:

$$H_0 \quad \bar{x} = \mu$$

$$H_1 \quad \bar{x} > \mu$$

The experimental u_{exp} value is:

$$u_{\text{exp}} = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{0.0840 - 0.080}{0.002887} = 1.386 \quad (4.38)$$

the critical value z_{cr} :

$$z_{\text{cr}}(\alpha, df) = z_{\text{cr}}(0.05, 2) = \text{NORM.S.INV}(1 - 0.05) = 1.645 \quad (4.39)$$

and the probability p :

$$p = 1 - \text{NORM.S.DIST}(z_{\text{exp}}, \text{TRUE}) = 1 - \text{NORM.S.DIST}(1.386, \text{TRUE}) = 0.0829 \quad (4.40)$$

In the above case $u_{\text{exp}} < z_{\text{cr}}$ and $p > \alpha$ therefore hypothesis H_0 cannot be rejected, and one can say that at the confidence level of 0.05 (95%) the experimental mean is not larger than the true value; it is statistically equal to the true value.

Re. c)

For left one-tailed test we have to examine the following hypotheses:

$$H_0 \quad \bar{x} = \mu$$

$$H_1 \quad \bar{x} < \mu$$

As described earlier, this test does not make any sense as $\bar{x} > 0.08$ but it is presented for illustration purposes.

The experimental u_{exp} value is:

$$u_{\text{exp}} = \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{0.0840 - 0.080}{0.002887} = 1.386 \quad (4.41)$$

the critical value z_{cr} :

$$z_{\text{cr}}(\alpha, df) = z_{\text{cr}}(0.05) = \text{NORM.S.INV}(0.05) = -1.645 \quad (4.42)$$

and the probability p :

$$p = \text{NORM.S.DIST}(z_{\text{exp}}, \text{TRUE}) = \text{NORM.S.DIST}(1.386, \text{TRUE}) = 0.917 \quad (4.43)$$

In this case $u_{\text{exp}} > z_{\text{cr}}$ and $p \gg \alpha$ therefore hypothesis H_0 cannot be rejected, and one can say that at the confidence level of 0.05 (95%) the experimental mean is not smaller than the true value; it is statistically equal to the true value.

The plots in this test are similar to those for t -test above.

4.8 Comparison of two means

Very often in analytical or physical chemistry we ask ourselves a question: are these two averages statistically equal? This question may arise when two different analytical or

physicochemical methods are compared or when comparing results of different analysts or labs. To answer this question, one should use a *t*-test. There are two tests used in two cases:

- a) when the variances of two sets (methods) are statistically the same
- b) when the variances of two sets (methods) are statistically different.

In these cases we have to decide between the following hypotheses:

$$H_0: \bar{x}_1 = \bar{x}_2$$

$$H_1: \bar{x}_1 \neq \bar{x}_2$$

4.8.1 Test of equality of two means when the variances are the same

When the variances or standard deviations of two sets of data are the same the value of t_{exp} is calculated as:

$$t_{\text{exp}} = \frac{\bar{x}_1 - \bar{x}_2}{s \sqrt{\frac{1}{N_1} + \frac{1}{N_2}}} \quad (4.44)$$

where N_1 and N_2 are the number of points and \bar{x}_1 and \bar{x}_2 are the means of the two sets of data. The average variance, s^2 , is:

$$s^2 = \frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2} \quad (4.45)$$

with the number of degrees of freedom, df :

$$df = N_1 + N_2 - 2 \quad (4.46)$$

The value of t_{exp} must be compared with $t_{\text{cr}}(\alpha, N_1 + N_2 - 2)$. If $t_{\text{exp}} < t_{\text{cr}}(\alpha, N_1 + N_2 - 2)$ one should keep hypothesis H_0 . *p*-test is calculated using Eq. (4.24).

This comparison might be also performed using t-Test: Two-Sample Assuming Equal Variances in Data Analysis in Excel which automatically calculated all the necessary values.

4.8.2 Test of equality of two means when the variances are different

In this case t_{exp} is defined as:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2}}} \quad (4.47)$$

with the number of degrees of freedom calculated as:

$$df = \frac{\left(\frac{s_1^2}{N_1} + \frac{s_2^2}{N_2} \right)^2}{\left(\frac{s_1^2}{N_1} \right)^2 \frac{1}{N_1 - 1} + \left(\frac{s_2^2}{N_2} \right)^2 \frac{1}{N_2 - 1}} \quad (4.48)$$

rounded to the whole number. As above one should compare t_{exp} with $t_{\text{cr}}(\alpha, df)$.

Example 4.12.

Moisture in two samples was determined by two different methods with different variances (see test F below for the determination of the equality of variances). Are there systematic differences at the confidence level of 95% between these methods? Use the following data and assume unequal variances.

x_{1i}	x_{2i}
6.4	6.5
6.2	6.7
6.2	6.5
6.5	6.1
6.3	6
6.4	6.8
6.4	6.2

The values of t_{exp} and k might be calculated manually but there is a program in Excel in Data Analysis, t -Test: Two-Sample Assuming Unequal Variances which can calculate Eqns. (4.47)-(4.48) automatically. Using this test the following screen is produced and the corresponding values must be filled.

t-Test: Two-Sample Assuming Unequal Variances

Input

Variable 1 Range:

Variable 2 Range:

Hypothesized Mean Difference:

☐ Labels

Alpha:

Output options

☒ Output Range:

☐ New Worksheet Ply:

☐ New Workbook

OK Cancel Help

Fig. 4.10. Screen for the t -test for comparison of two samples assuming unequal variances, confidence level 0.05 (see the Excel file *Examples4.xlsx* sheet *Ex. 4.12*).

Using this program the following results are obtained:

Table 4.5. *t*-test for comparison of two samples assuming unequal variances for the above data.

t-Test: Two-Sample Assuming Unequal Variances

	Variable 1	Variable 2
Mean	6.342857	6.4
Variance	0.012857	0.093333
Observations	7	7
Hypothesized Mean Difference	0	
df	8	
t Stat	-0.46395	
P(T<=t) one-tail	0.327524	
t Critical one-tail	1.859548	
P(T<=t) two-tail	0.655049	
t Critical two-tail	2.306004	

In the above table “t Stat” is our t_{exp} . It should be taken as the absolute value, $t_{\text{exp}} = 0.464$, and the mean number of degrees of freedom df is 8. The value of $t_{\text{cr}}(0.05, 8) = 2.306$ (calculated automatically, see the last line, using two-tailed test) and $t_{\text{exp}} < t_{\text{cr}}(0.05, 8)$ and one can say that two methods give the same result. Besides, the probability that H_0 is true, “P(T<=t) two-tail”, $p = 0.655$, which is much larger than $\alpha = 0.05$ which confirms that H_0 cannot be rejected. The value of p is calculated using Eq. (4.24). See calculations in *Examples4.xlsx*, sheet *Ex. 4.12*.

4.9 Paired *t*-test for comparing individual differences of two samples

Let us suppose that two different methods are used to make single measurements of several different samples. The measurements are carried only once for each sample. The question is if the two methods give the same results i.e. the differences between them are statistically unimportant. In such a case the differences between two measurements $d_i = a_i - b_i$ represents our variable with the average \bar{d} and the standard deviation s_d :

$$\bar{d} = \sum_{i=1}^N \frac{(a_i - b_i)}{N}$$

$$s_d = \sqrt{\frac{\sum_{i=1}^N (d_i - \bar{d})^2}{N - 1}} \quad (4.49)$$

where N is the number of measurements. The hypothesis studied are:

H_0 : differences between two series are negligible

H_1 : differences between two series are not negligible

The *t*-test for the differences is defined as:

$$t_{\text{exp}} = \frac{|\bar{d}|}{\frac{s_d}{\sqrt{N}}} = \frac{|\bar{d}|}{s_{\bar{d}}} \quad (4.50)$$

This value should be compared with $t(\alpha, N-2)$. If $t_{\text{exp}} < t(\alpha, N-2)$.
Let us look at the example.

Example 4.13.

Two methods a and b were used to measure concentration of the analyte. Eight samples were analyzed using these methods. The results are shown below:

a	b
1.79	2.01
1.78	2.51
6.14	5.94
5.82	7.23
1.73	1.41
5.37	4.95
6.41	6.59
2.37	2.50

Although parameters in Eq. (4.49) are easily calculable there is a program in Excel Data Analysis: “t-Test: Paired Two Sample for Means” which calculates all the necessary results.

t-Test: Paired Two Sample for Means

Input

Variable 1 Range: \$A\$3:\$A\$10

Variable 2 Range: \$B\$3:\$B\$10

Hypothesized Mean Difference: 0

☐ Labels

Alpha: 0.05

Output options

☒ Output Range: \$D\$2

☐ New Worksheet Ply:

☐ New Workbook

OK Cancel Help

The obtained results are:

t-Test: Paired Two Sample for Means

	<i>Variable 1</i>	<i>Variable 2</i>
Mean	3.92625	4.1425
Variance	4.736541	5.254707
Observations	8	8
Pearson Correlation	0.964646	
Hypothesized Mean Difference	0	
df	7	
t Stat	-1.01075	
P(T<=t) one-tail	0.172896	
t Critical one-tail	1.894579	
P(T<=t) two-tail	0.345791	
t Critical two-tail	2.364624	

The calculated value $|t_{\text{exp}} (t \text{ Stat})| = 1.01$ and the $t_{\text{cr}}(0.05'', 7) = 2.36$. In this case $|t_{\text{exp}}| < t_{\text{cr}}(0.05'', 7)$ therefore with the confidence of 95% one can say that these two method give the same results. The same result is obtained using two-tail p -test, $p = 0.346$ therefore H_0 should not be rejected. See the calculations in *Examples4.xlsx*, sheet *Ex. 4.13*.

4.10 Test F for the comparison of variances

Test F proposed by Fisher and Snedecor is used to determine the equality of variances. The null and alternative hypotheses studied are:

$$H_0: s_1^2 = s_2^2$$

$$H_1: s_1^2 \neq s_2^2$$

Example 4.14.

Simulate the probability distribution function $P_F(f, df_1, df_2)$ versus parameter f (where df_1 and df_2 are the numbers of degrees of freedom) for the degrees of freedom: 1, 9 and 5, 9 and its integrals.

The simulations are in Excel file *Examples4.xlsx*, sheet *Ex. 4.14* and the plots of probability distribution function $P_F(f, df_1, df_2)$ versus f are displayed in Fig. 4.11 for the degrees of freedom: 1, 9 and 5, 9.

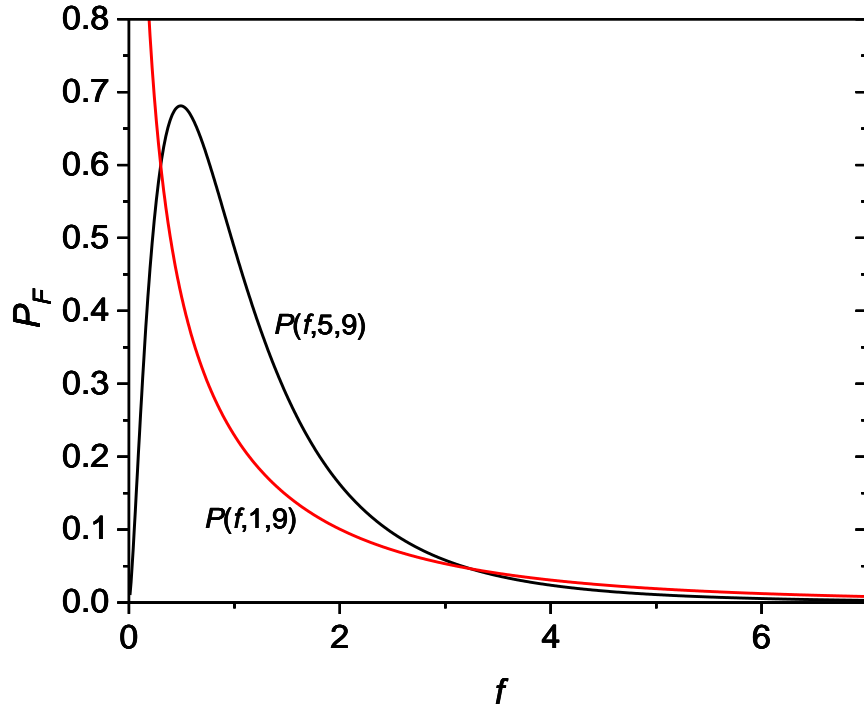


Fig. 4.11. Probability F -distribution for two different sets of degrees of freedom: 1,9 and 5,9.

The integral of the F -distribution probability function is chosen to give the confidence level α :

$$\alpha = \int_{F(\alpha, df_1, df_2)}^{\infty} P_F(f, df_1, df_2) df \quad (4.51)$$

This integral is illustrated in Fig. 4.12. The values of the probability $P_F(f, df_1, df_2)$ may be obtained using Excel function $F.DIST(f, df_1, df_2, FALSE)$ and the values of the critical values using: $F(\alpha, df_1, df_2) = F.INV.RT(\alpha, df_1, df_2)$. The integral:

$$\int_0^F P_F(f, df_1, df_2) df \quad (4.52)$$

is calculated using $F.DIST(F, df_1, df_2, TRUE)$. To better understand these functions an example for $\alpha = 0.05$ and the number of degrees of freedom $df_1 = 5$ and $df_2 = 9$ is presented.

The values in Fig. 4.12 were calculated using:

$P = F.DIST(f, 5, 9, FALSE)$ and for $f = 0.1$, $P = F.DIST(0.1, 5, 9, FALSE) = 0.232008$,

$F(0.05, 5, 9) = F.INV.RT(0.05, 5, 9) = 3.48166$,

$\alpha = F.DIST.RT(F, 5, 9) = F.DIST.RT(3.48166, 5, 9) = 0.05$,

$1 - \alpha = F.DIST(3.48166, 5, 9, TRUE) = 0.95$.

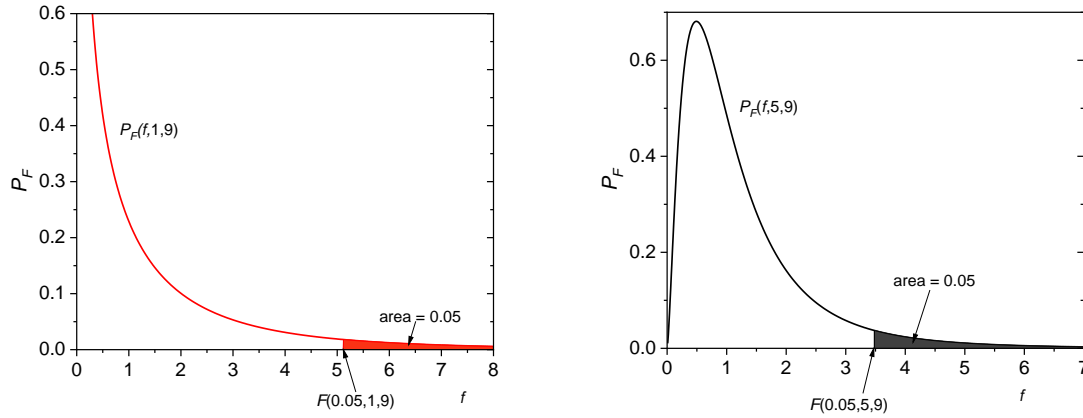


Fig. 4.12. The integral under the F -distribution function, corresponds to the confidence level, in this case 0.05 and the value corresponding to the beginning of integration is $F_{cr}(\alpha, df_1, df_2)$.

To test it function F_{exp} is determined:

$$F_{exp} = \frac{s_1^2}{s_2^2} \geq 1 \text{ for } s_1^2 > s_2^2 \quad (4.53)$$

Function F_{exp} must always be larger than 1 and if $s_1 < s_2$ the numerator and denominator in Eq. (4.53) must be exchanged. F_{exp} must be compared with the calculated critical value $F_{cr}(\alpha, df_1, df_2)$ where df_1 and df_2 are the numbers of degrees of freedom of the numerator and denominator, respectively. In the case of the comparison of means they are simply $df_i = N_i - 1$. If $F_{exp} < F_{cr}(\alpha, df_1, df_2)$ there are no reasons to reject hypothesis H_0 and the variances are equal at the confidence level α . It should be mentioned that $F(\alpha, df_1, df_2) \neq F(\alpha, df_2, df_1)$.

The value of probability, p , that H_0 is true is calculated as:

$$p = \int_F^{\infty} P_F(f, df_1, df_2) df = 1 - \int_0^F P_F(f, df_1, df_2) df \quad (4.54)$$

It can be evaluated as $F.DIST.RT(F, df_1, df_2) = 1 - F.DIST(F, df_1, df_2, TRUE)$. When $p < \alpha$, hypothesis H_0 should be rejected because it is very little probable.

Example 4.15.

Let us verify if the two sets of data in

Example 4.12 have the same variances. Function F_{exp} might be calculated manually or using F-test Two-Samples for Variances from Data Analysis. Taking the first data column to numerator and the second to denominator the following results are obtained using F-Test Two-Sample for Variances in Data Analysis.

Table 4.6. Results of the F-test for the data in Example 4.15 taking the first data column to numerator and the second to denominator.

F-Test Two-Sample for Variances

	Variable 1	Variable 2
Mean	6.342857	6.4
Variance	0.012857	0.093333
Observations	7	7
df	6	6
F	0.137755	
P(F<=f) one-tail	0.014682	
F Critical one-tail	0.233434	

The above results indicate that the experimental F value, $F_{\text{exp}} = 0.138$ is lower than one. As F_{exp} should always be greater than or equal to one, $F_{\text{exp}} \geq 1$, the columns should be inversed, i.e. the second to numerator and the first to denominator. The following results are obtained:

Table 4.7. Results of the F-test for the data in Example 4.15 taking the first data column to denominator and the second to numerator.

F-Test Two-Sample for Variances

	Variable 1	Variable 2			
Mean	6.4	6.342857			
Variance	0.093333	0.012857	0.093333	0.012857	=VAR.S(A3:A9)
Observations	7	7			
df	6	6			
F	7.259259				
P(F<=f) one-tail	0.014682		0.014682	=F.DIST.RT(D19,6,6)	
F Critical one-tail	4.283866		4.283866	=F.INV.RT(0.05,6,6)	

In this case $F_{\text{exp}} = 7.259$. This program also calculates the critical one-tailed value of the test for the assumed value of α , in this case for $\alpha = 0.05$, which is $F_{\text{cr}}(0.05, 6, 6) = 4.284$. In this example $F_{\text{exp}} > F_{\text{cr}}(0.05, 6, 6)$ and the hypothesis H_0 must be rejected and the two variances are different.

In the table there is another value called: “P(F<=f) one tail” which is the parameter p discussed earlier (probability of H_0). In the above case $p = 0.01468 < 0.05$ which confirms that H_0 should be rejected.

See calculations in *Examples4.xlsx*, sheet *Ex. 4.15*.

Example 4.16.

Two analytical methods were used to determine titanium in a sample. Determine if these two methods give the same results and have the same precision at the confidence level of 95%. The obtained data are displayed below.

$x_{1,i}$	$x_{2,i}$
1.00	1.12
1.22	1.05
1.29	1.19
1.11	1.06
1.10	1.10
1.24	1.24
1.16	1.127

First, the test F of the equality of variances should be performed. The following hypotheses are tested:

$$H_0: s_1^2 = s_2^2$$

$$H_1: s_1^2 \neq s_2^2$$

The test F in Excel gives the following results:

F-Test Two-Sample for Variances

	<i>Variable 1</i>	<i>Variable 2</i>
Mean	1.16	1.126714
Variance	0.009767	0.004656
Observations	7	7
df	6	6
F	2.097845	
P(F<=f) one-tail	0.194529	
F Critical one-tail	4.283866	

The value of $F_{\text{exp}} = 2.098 < F_{\text{cr}}(0.05, 6, 6) = 4.284$ therefore there are no reasons to reject H_0 . Besides, $p = 0.194 > \alpha = 0.05$ which confirms that the variances of two experiments are statistically the same and the two methods have the same precision.

Next, let us perform t test for the equality of means assuming equal variances.

t-Test: Two-Sample Assuming Equal Variances

	<i>Variable 1</i>	<i>Variable 2</i>
Mean	1.186667	1.127833
Variance	0.005747	0.005576
Observations	6	6
Pooled Variance	0.005661	
Hypothesized Mean Difference	0	
df	10	
t Stat	1.354321	
P(T<=t) one-tail	0.102722	
t Critical one-tail	1.812461	
P(T<=t) two-tail	0.205444	
t Critical two-tail	2.228139	

It is evident that $t_{\text{exp}} = 1.354 < t_{\text{cr}}(0.05'', 10) = 2.228$ and H_0 cannot be rejected. The p test gives $p = 0.205 > \alpha = 0.05$ and confirms the two methods have statistically the same means. The two methods give the same results and have the same means and variances. See calculations in *Examples4.xlsx*, sheet *Ex. 4.16*.

5 Test of regression parameters

Below, different tests used in regression analysis are presented.

5.1 Rejection of the point in regression, outliers

Sometimes it happens that one (or more) point(s) lie further from the regression line than the others. In such a case one poses a question: should I keep this point in the calculation of the regression parameters? The points which lie far from the predicted are called **outliers**. Operation of removing such points must be applied with great care to avoid removing important information.^{48,49,50} See also remarks in Section 4.3.

Let us look at the example below.

Example 5.1.

Compare the following data and look if the points exceed confidence limits for the experimental values.

x	y
0.0	-0.9
0.5	-0.6
1.0	0.2
1.5	0.4
2.0	2.0
2.5	1.5
3.0	1.8
3.5	2.5
4.0	2.9
4.5	3.7
5.0	4.0

Performing regression analysis with 95% and 99% confidence intervals are displayed in Fig. 5.1 (see also Origin file *outliers band.opj*). The prediction line of experimental y_i corresponds to testing with Eq. (3.40). Assuming confidence intervals of 95% one point lie outside this interval while for confidence intervals of 99% all the experimental points are inside the interval. The results are in Excel file *Examples 5*, sheet *Ex. 5.1-5.2*.

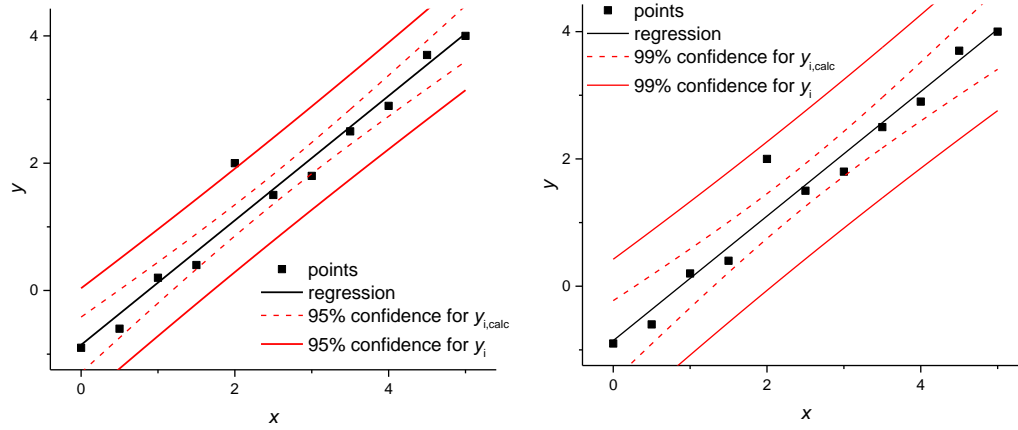


Fig. 5.1. Experimental (points), calculated regression (black line), confidence intervals of \hat{y}_i (dashed lines) and of y_i (continuous red lines) for $\alpha = 0.05$ and 0.01 , that is 95% and 99% confidence intervals. Calculated using Origin.

5.1.1 Simple t -test

As it can be seen from Fig. 5.1 one point lies outside 95% confidence level for experimental points, which means that there is one chance in 20 that it lies outside the 95% prediction band. However, it is inside 99% confidence level.

Deviations of the experimental points from the calculated regression line are plotted in Fig. 5.2.

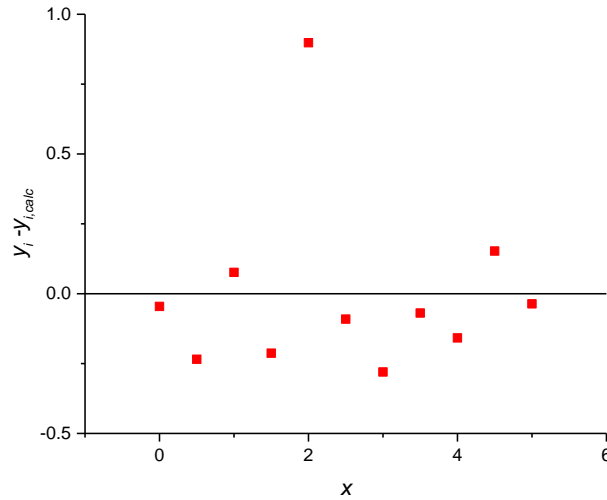


Fig. 5.2. Plot of the deviations of the experimental points from the regression line.

Intuitively, one point seems to be an outlier. Comparison of the point versus confidence limit corresponds to the t -test for Δ , Eqs. (3.38) and (3.40):

$$t_{\text{exp}} = \frac{\Delta}{s_{\Delta}} \quad (5.1)$$

This value should be compared with $t_{cr}(\alpha, df)$ or, in Example 5.1, $t_{cr}(0.05, 9) = 2.262$. The experimental value of $t_{exp} = 2.500$ which is larger than $t(0.05, 9)$ at 95% confidence level. However, in this test the standard deviation is calculated with the possible outlier(s) which increases its value and lowers t_{exp} . This can be corrected using standardized residuals.^{48,50}

5.1.2 Internally studentized residuals

To detect outliers it was proposed to use the standardized or internally studentized residuals:

$$s_i = \frac{y_i - \hat{y}_i}{\sqrt{1 - h_{ii}} s_y} \quad (5.2)$$

where h_{ii} is the diagonal element of the hat matrix, **(3.96)**, s_y is the regression standard deviation, and \hat{y}_i is the calculated value using full regression for N points. Diagonal elements of the hat matrix can be easily found as:

$$h_{ii} = \frac{(x_i - \bar{x})^2}{\sum_{i=1}^N (x_i - \bar{x})^2} + \frac{1}{N} \quad (5.3)$$

Standardized residuals are useful in detecting of outliers. Specific calculations for the regression are presented in Example 5.1. This value for a suspected outlier is $s_i = 2.764$. Usually, values of $s_i > 2$ indicate possibility of an outlier.

5.1.3 Jack-knifed or externally studentized residuals

The problem with standardized residuals is that they also depend on the estimated standard deviation which is affected by outliers. To avoid this problem one can use standard deviation calculated omitting the suspected outlier.

The Student ***t*** test used here is:

$$t_i = \frac{y_i - \hat{y}_{i(i)}}{s_{y_i - \hat{y}_{i(i)}}} \quad (5.4)$$

where $\hat{y}_{i(i)}$ is the value of point i calculated using regression without the suspected outlier (i) that is using $N - 1$ points and $s_{y_i - \hat{y}_{i(i)}}$ its standard deviation. The difference $y_i - \hat{y}_{i(i)}$ may be calculated without performing new regression analysis without point i :

$$y_i - \hat{y}_{i(i)} = \frac{y_i - \hat{y}_i}{1 - h_{ii}} \quad (5.5)$$

where \hat{y}_i is calculated from regression using N points. The value calculated using Eq. (5.4) is called Jack-knifed residual, externally studentized residual, or studentized deleted residual:

$$t_i = \frac{y_i - \hat{y}_i}{\sqrt{1 - h_{ii}} s_{y_{i(i)}}} \quad (5.6)$$

where $s_{y_{i(i)}}$ is the standard deviation calculated using $N - 1$ points i.e. without the suspected outlier (i). This equation might be transformed to a simpler form:

$$t_i = s_i \sqrt{\frac{N-n-1}{N-n-s_i^2}} \quad (5.7)$$

where s_i is the standardized deviation calculated using N points, Eq. (5.2), and n is the number of parameters in the regression ($n = 2$ for the linear regression). It could be compared with $t_{cr}(\alpha'', N-n-1) = t_{cr}(0.05'', 8) = 2.31$. However as we do not know in advance which point has the largest $|t_i|$ we should carry out N tests and the critical value should be^{48,50} $t_{cr}(\alpha/N'', N-n-1)$ where the confidence level was divided by the number of points and the number of degrees of freedom is calculated using $N-1$ experimental points and n parameters. It of course gives much larger values than $t(\alpha'', N-n-1)$. In our example experimental value $t_i = 6.70$, $t_{cr}(0.05, 8) = 2.305$ and $t_{cr}(0.05/11'', 8) = 3.90$. In this case the experimental value $t_i > t_{cr}(0.05/11'', 8)$ and this point looks like an outlier. Removing this outlier from the regression reduces the standard deviation, s_y , from 0.34 to 0.14.

This means that the point (2, 2) is an outlier and the regression should be carried out without it. Removal of the outlier gives the following results at the confidence level of 95% ($\alpha = 0.05$):

$$b_0 = -0.99 \pm 0.19$$

$$b_1 = 0.996 \pm 0.062$$

This test is also calculated in Origin.

5.1.4 Cook's distance

There is a statistical method proposed by Cook^{48,50-52} which helps to decide if the point should be rejected. It is based on the comparison of the regression with and without the suspected outlier and inspection how much the estimation would change after such operation. It is a normalized measure of the influence of point i on all predicted values. The **Cook's distance** for point i , D_i , is defined as:

$$D_i = \sum_{j=1}^N \frac{(\hat{y}_j - \hat{y}_{j(i)})^2}{n s_y^2} \quad (5.8)$$

where \hat{y}_j is the predicted (calculated) value using full regression containing N points and $\hat{y}_{j(i)}$ is the predicted value calculated from regression excluding point i (using $N-1$ points). D_i can be calculated using a simpler relation:

$$D_i = s_i^2 \frac{h_{ii}}{(1-h_{ii})n} = \frac{(y_i - \hat{y}_i)^2}{n s_y^2} \frac{h_{ii}}{(1-h_{ii})^2} \quad (5.9)$$

where n is the number of parameters used in regression ($n = 2$ for linear regression) and s_i is calculated using Eq. (5.2).

Cook's test is not strictly speaking a statistical test. The threshold for data rejection is not well fixed and different statisticians proposed different criteria. Data point can be rejected if $D_i > 1$ ⁵³, $D_i > 0.7$ ⁵⁴, $D_i > 4/N$ ⁵⁵, or $D_i > 4/(N-p-1)$.^{56,57} The Cook's test will be illustrated in the following example.

Example 5.2.

Apply the standardized residual and Cook's test to the data in *Examples 5*, sheet *Ex. 5.1-5.2*.

First, linear regression should be used for all 11 points. The obtained results at 95% are:

$b_0 = -0.85 \pm 0.44$; $b_1 = 0.98 \pm 0.15$, $s_y = 0.34$.

Using Excel the following parameters were calculated:

x	y	residuals	h_{ii}	s_i	t_i	D_i
0.0	-0.9	-0.04545	0.318182	0.160706	0.151733	0.006026
0.5	-0.6	-0.23455	0.236364	0.783561	0.765314	0.095019
1.0	0.2	0.076364	0.172727	0.245105	0.231862	0.006272
1.5	0.4	-0.21273	0.127273	0.664773	0.642732	0.032224
2.0	2.0	0.898182	0.1	2.763963	6.70235	0.424416
2.5	1.5	-0.09091	0.090909	0.278351	0.263569	0.003874
3.0	1.8	-0.28	0.1	0.86164	0.848096	0.041246
3.5	2.5	-0.06909	0.127273	0.215909	0.20409	0.003399
4.0	2.9	-0.15818	0.172727	0.507717	0.485686	0.026911
4.5	3.7	0.152727	0.236364	0.510226	0.488158	0.040289
5.0	4.0	-0.03636	0.318182	0.128565	0.121324	0.003857
$D_{\text{theor}} = 4/(11-2-1) =$		0.5				
$D_{\text{theor}} = 4/11 =$		0.363636				

For the suspected point No 5, (2.0,2.0) the value of $s_i = 2.76$, Eq. (5.2), it is >2 which suggests that it is an outlier. The Jack-knifed value of $t_i = 6.70$, Eq. (5.6), is much larger than $t(0.05/11, 8) = 3.90$ and this point looks like an outlier.

The Cook's distance for this point, is 0.424 that is lower than one. However other threshold values are 0.5 or 0.364, depending on criterion used. Based on Cook's distance in social and bio/medical sciences this point probably would not be rejected. In physical sciences where we are much stricter (better correlations expected) this point should be rejected using Jack knifed test.

The regression should be repeated without this outlier and the obtained results at the confidence level of 95% are: $b_0 = -0.99 \pm 0.19$, $b_1 = 0.996 \pm 0.062$, $s_y = 0.14$. It can be noticed that the confidence intervals of the regression parameters and the standard deviation are largely reduced after elimination of the suspected outlier.

The Cook's test is more often used in social and biomedical sciences. In sciences it is usually possible to repeat the experiment to see if the outliers are always the same. To check for outliers one can use Jack-knifed or studentized residuals t_i -test (see above) which is a modified t -test.

5.2 Statistical importance of the regression parameters

In sciences the regression parameters have a physical meaning therefore it is important to decide if the obtained parameters are statistically important i.e. if they should be kept or rejected.^{8,18} For a linear regression model:

$$y = b_0 + b_1 x \quad (5.10)$$

two simpler models are possible:

$$\begin{aligned} 1) \quad y &= b_0 = \bar{y} \\ 2) \quad y &= b_1 x \end{aligned} \quad (5.11)$$

The general rule in statistics is that in the analysis of experimental data the **number of adjustable parameters should be kept at minimum**. This is so called Occam's Razor (adapted for statistics): Select the simplest model that describes the data sufficiently well.

In the first case above the slope, b_1 , is negligible and the data can be described by the arithmetic mean and in the second case the origin is small i.e. does not have statistical meaning, therefore only slope should be used.

There are two tests which can be used to check importance of parameters: t and F . They will be discussed below.

5.2.1 t -test of the importance of regression parameters

Fist, let us look into importance of the slope, b_1 . One should write two hypotheses:

$$\begin{aligned} H_0: & \quad b_1 = 0 & \quad y = b_0 = \bar{y} \\ H_1: & \quad b_1 \neq 0. & \quad y = b_0 + b_1 x \end{aligned}$$

Student t -test is defined as:

$$t_{\text{exp}} = \frac{b_1}{s_{b_1}} \quad (5.12)$$

This value should be compared with $t_{\text{cr}}(\alpha'', N-2)$. If $t_{\text{exp}} > t_{\text{cr}}(\alpha'', N-2)$ i.e. the value of the parameter is much larger than its standard deviation H_0 should be rejected and the slope is important.

Similarly, one should proceed in the determination of the importance of b_0 . The hypotheses tested are:

$$\begin{aligned} H_0: & \quad b_0 = 0 & \quad y = b_1 x \\ H_1: & \quad b_0 \neq 0. & \quad y = b_0 + b_1 x \end{aligned}$$

Student t -test is:

$$t_{\text{exp}} = \frac{b_0}{s_{b_0}} \quad (5.13)$$

and the hypothesis H_0 should be rejected if $t_{\text{exp}} > t_{\text{cr}}(\alpha'', N-2)$.

5.2.2 F -test of the importance of regression parameters

F -test is defined as a ratio of variances. When a parameter is important, then starting from a simpler model and adding this parameter should significantly decrease the residual sum of squares. This is the base of the sequential F -test. Let us start with studying of the importance of the slope, b_1 . The hypotheses studied are:

$$\begin{aligned} H_0: & \quad b_1 = 0 & \quad y = b_0 = \bar{y} \\ H_1: & \quad b_1 \neq 0. & \quad y = b_0 + b_1 x \end{aligned}$$

Let us determine the sum of squares for both models:

S_1^2 (sum of squares corrected for mean) for the model: $y = b_0 = \bar{y}$ is $S_1^2 = \sum (y_i - b_0)^2 = \sum (y_i - \bar{y})^2$ which has $N-1$ degrees of freedom.

S_2^2 (sum of squares about regression) for the model: $y = b_0 + b_1x$ is $S_2^2 = \sum (y_i - \hat{y}_i)^2$ which has $N-2$ degrees of freedom.

Then, we can construct sequential F -test function which describes decrease of the sum of squares due to addition of one parameter:

$$F_{\text{exp}} = \frac{s_1^2}{s_2^2} = \frac{\frac{S_1^2 - S_2^2}{1}}{\frac{S_2^2}{N-2}} = \frac{S_1^2 - S_2^2}{s_y^2} \quad (5.14)$$

where the difference in s_1 has one degree of freedom as inly one parameter, b_1 , is added to the regression: $(N-1) - (N-2) = 1$

$$s_1^2 = \frac{S_1^2 - S_2^2}{(N-1) - (N-2)} = \frac{S_1^2 - S_2^2}{1} = S_1^2 - S_2^2 \quad (5.15)$$

and

$$s_2^2 = \frac{S_2^2}{N-2} \quad (5.16)$$

For the parameter b_1 to be important, decrease of the sum of squares $S_1^2 - S_2^2$ must be statistically significant in comparison with $s_y^2 = S_2^2 / N - 2$. F_{exp} must be compared with $F(\alpha, 1, N-2)$.

If $F_{\text{exp}} > F_{\text{ct}}(\alpha, 1, N-2)$ hypothesis H_0 must be rejected which means that parameter b_1 is important (large improvement after adding it to the regression). It should be value stressed that the test t is formally identical with the test F because $F(\alpha, 1, df_2) = t^2(\alpha', df_2)$. The values of F_{exp} are calculated in the table of analysis of variances ANOVA.

In a similar way one can determine importance of the parameter b_0 (intercept). The hypotheses are:

$$\begin{array}{lll} H_0: & b_0 = 0 & y = b_1x \\ H_1: & b_0 \neq 0. & y = b_0 + b_1x \end{array}$$

and test F is calculated using:

S_1^2 (sum of squares corrected for mean) for the model: $y = b_1x$, $S_1^2 = \sum (y_i - b_1x_i)^2$ with $N-1$ degrees of freedom,

S_2^2 (sum of squares about regression) for the model: $y = b_0 + b_1x$, $S_2^2 = \sum (y_i - \hat{y}_i)^2 = \sum (y_i - b_0 - b_1x_i)^2$ which has $N-2$ degrees of freedom, and the subsequent formulas are identical.

In certain more complex cases (nonlinear regression) two (or more) parameters must be added in one step. Assuming that the first model contains p parameters and the second one $p+k$ parameters the test- F is defined as:

$$F_{\text{exp}} = \frac{s_1^2}{s_2^2} = \frac{\frac{S_1^2 - S_2^2}{(N-p) - (N-p-k)}}{\frac{S_2^2}{N-p-k}} = \frac{\frac{S_1^2 - S_2^2}{k}}{s_y^2} \quad (5.17)$$

It should be stressed that the number of added parameters **should be kept to a strict minimum**, typically one.

Before showing examples the table of Analysis of Variances, ANOVA, which is calculated automatically by the regression programs, will be discussed in detail.

5.3 ANOVA

Deviations of the experimental y_i from the mean value can be decomposed into two parts:

$$(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}) \quad (5.18)$$

that is:

Total difference corrected for mean = residual difference + difference explained by regression (5.19)

The following equation may be also written for sum of squares:

$$\sum_{i=1}^N (y_i - \bar{y})^2 = \sum_{i=1}^N (y_i - \hat{y}_i)^2 + \sum_{i=1}^N (\hat{y}_i - \bar{y})^2 \quad (5.20)$$

where the total sum of squares corrected for mean equals sum of squares about regression (residual sum of squares) plus sum of squares due to regression. Number of degrees of freedom of these terms are: $(N-1) = (N-2) + 1$. These sums of squares and mean sums of squares are displayed in the table of ANOVA.

Table 5.1. Table of analysis of variances, ANOVA, for linear regression.

Source of variation	Degrees of freedom	Sum of Squares	Mean square	Test F
Due to regression	1	$\sum(\hat{y}_i - \bar{y})^2$	$MS_R = \sum(\hat{y}_i - \bar{y})^2 / 1$	$F = \frac{MS_R}{s_y^2}$
About Regression (residual)	$N - 2$	$SS = \sum(y_i - \hat{y}_i)^2$	$s_y^2 = \frac{SS}{N - 2}$	
Total, corrected for mean, \bar{y}	$N - 1$	$\sum(y_i - \bar{y})^2$		

Test- F in the table of ANOVA is the test of **importance of the slope**, b_1 . It is given here as:

$$F = \frac{\sum(\hat{y}_i - \bar{y})^2}{\frac{\sum(y_i - \hat{y}_i)^2}{N - 2}} \quad (5.21)$$

which is equivalent to Eq. (5.14) because:

$$\sum_{i=1}^N (y_i - \bar{y})^2 = \sum_{i=1}^N (y_i - \hat{y}_i)^2 + \sum_{i=1}^N (\hat{y}_i - \bar{y})^2$$

that is

$$\sum_{i=1}^N (y_i - \bar{y})^2 - \sum_{i=1}^N (y_i - \hat{y}_i)^2 = \sum_{i=1}^N (\hat{y}_i - \bar{y})^2 \quad (5.22)$$

It can be added that there is a simple relation between F and r^2 :

$$F = \frac{r^2}{1 - r^2} (N - 2) \quad (5.23)$$

The following hypothesis will be tested:

$$\begin{aligned} H_0: & \quad b_1 = 0 & \quad y = b_0 = \bar{y} \\ H_1: & \quad b_1 \neq 0. & \quad y = b_0 + b_1 x \end{aligned}$$

which should be compared with $F(\alpha, 1, N - 2)$. ANOVA analysis will be explained in the following example.

It should be added that for the model with the origin equal to zero, that is $y = b_1 x$ ANOVA is calculated differently.^{1,2,58} This is because regression is forced through origin (0,0) and it is generally inconsistent with the best fit. In this case $\sum(y_i - \hat{y}_i) \neq 0$ and $\bar{y}_i = \hat{\bar{y}}_i$. This leads to a different table of ANOVA (see below).

Table 5.2. Table of ANOVA for the regression through the origin: $y = b_1x$.

Source of variation	Degrees of freedom	Sum of Squares	Mean square	Test F
Due to regression	1	$\sum (\hat{y}_i)^2$	$MS_R = \sum (\hat{y}_i)^2 / 1$	$F = \frac{MS_R}{s_y^2}$
About Regression (residual)	$N - 1$	$SS = \sum (y_i - \hat{y}_i)^2$	$s_y^2 = \frac{SS}{N - 1}$	
Total, corrected for mean, \bar{y}	N	$\sum (y_i)^2$		

In earlier versions of Excel this table was calculated incorrectly. The coefficient of determination for this model is calculated as:

$$r^2 = \frac{\sum \hat{y}_i^2}{\sum y_i^2} \quad (5.24)$$

and it can be absurdly large even when the correlation between x and y is weak. Because it is meaningless it should not be used in this case.²

Example 5.3.

Carry out regression analysis for the following data.

x	y
-2.0	-0.93
-1.8	-0.93
-1.6	-0.68
-1.4	-0.47
-1.2	-0.38
-1.0	-0.23
-0.8	-0.52
-0.6	-0.22
-0.4	0.01
-0.2	-0.11
0.0	0.03

Using regression analysis in Excel gives the results shown in Fig. 5.3 and Table 5.3.

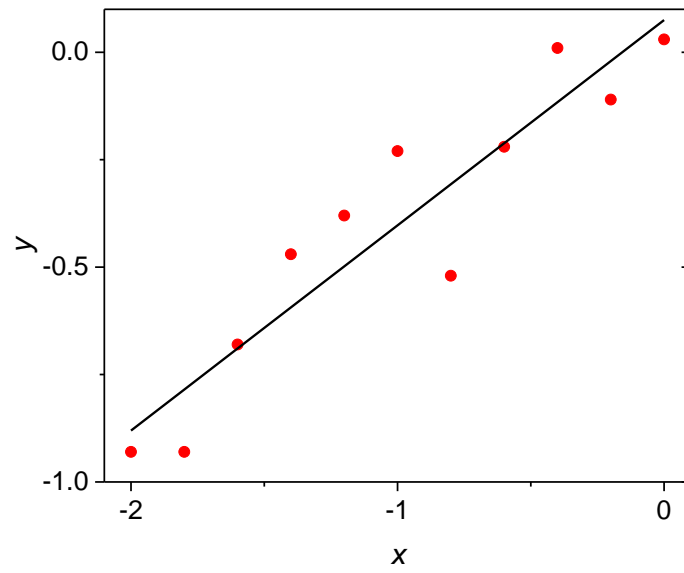


Fig. 5.3. Plot of the experimental points and predicted regression line for the data in Example 5.3.

Table 5.3. Excel output for the regression analysis in Example 5.3 assuming model: $y = b_0 + b_1x$.

SUMMARY
OUTPUT
y=bo+b1 x

Regression Statistics	
Multiple R	0.931131854
R Square	0.86700653
Adjusted R Square	0.852229478
Standard Error	0.130824503
Observations	11

ANOVA	F(0.05,1,9)=	5.11735503
	F.DIST.RT(58.6725,1,9)	3.12689E-5

	df	SS	MS	F	Significance F
Regression	1	1.004182727	1.004182727	58.6724957	3.12689E-5
Residual	9	0.154035455	0.017115051		
Total	10	1.158218182			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	0.075	0.073794972	1.016329406	0.33601429	-0.091935824	0.241935824
X Variable 1	0.477727273	0.062368135	7.659797367	3.1269E-05	0.33664075	0.618813795

These results should be compared with those in Table 5.1. In the Table of ANOVA SS means sum of squares and MS (mean square) is the mean sum of squares that is SS divided by the number of degrees of freedom, df . In that table there is a value of test for the importance of parameter b_1 , F_{exp} , that is the ratio of MS in Regression and MS Residual: $1.004183 / 0.017115 = 58.67$. This value is much larger than $F(0.05, 1, 9) = 5.117$ that is the slope is important at the confidence level of 95%. This is also confirmed by the test p of Significance of F equal to $p = 3.1269 \times 10^{-5} \ll \alpha = 0.05$. This value is the probability that the slope is zero i.e. of hypothesis H_0 being true (not significant). This probability is very small. It is calculated using Excel function $F.DIST.RT(F, df_1, df_2)$ here equal to $F.DIST.RT(58.67, 1, 9)$. The value of s_y^2 is the MS Residual equal to 0.017115 and $s_y = 0.13$. Besides, r^2 called R Square is the ratio of SS in Regression to SS Total, $r^2 = 1.004183 / 1.158218 = 0.867$.

The plots of deviations in Eq. (5.18) are presented in Fig. 5.4 -Fig. 5.6.

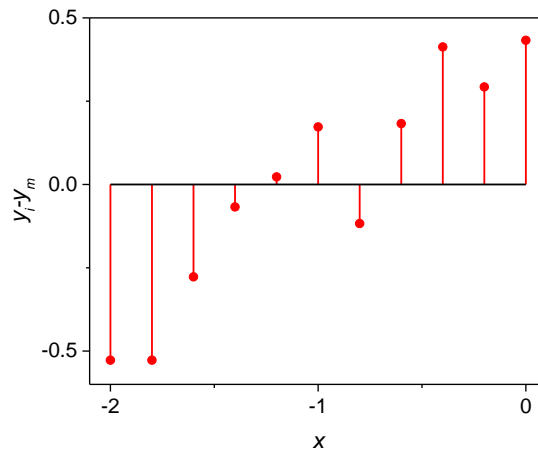


Fig. 5.4. Plot of the total difference corrected for mean, $y_i - \bar{y}$.

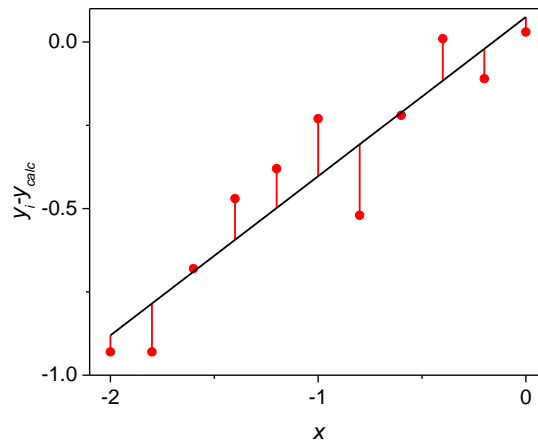


Fig. 5.5. Plot of the residual difference unexplained by regression, $y_i - \hat{y}_i$.

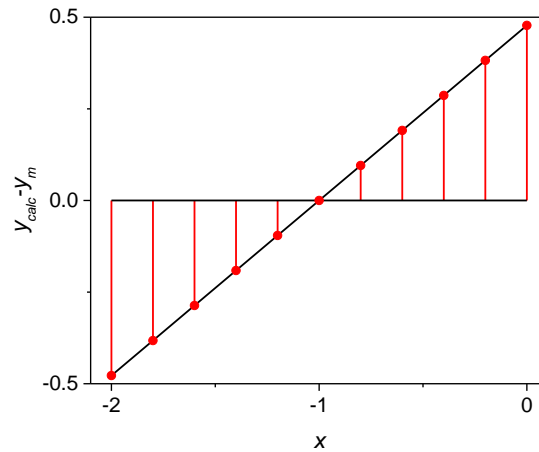


Fig. 5.6. Plot of the differences explained by regression corrected for mean, $\hat{y}_i - \bar{y}$.

Let us now check importance of regression parameters using t -test. These values are displayed as t Stat. They are:

for intercept b_0 : $t_{\text{exp}} = 1.016, p = 0.3360$

for slope b_1 : $t_{\text{exp}} = 7.660, p = 3.1269 \times 10^{-5}$

while the value of $t(0.05'', 9) = \text{T.INV.2T}(0.05, 9) = 2.262$. It is evident that for b_0 : $t_{\text{exp}} < t_{\text{cr}}(0.05'', 9)$ and this parameter is not important in regression while for b_1 : $t_{\text{exp}} > t_{\text{cr}}(0.05'', 9)$ and the slope is important. This is confirmed by the values of the probabilities for b_0 which is $p = 0.3360$, much larger than the assumed here value $\alpha = 0.05$ which confirms that the hypothesis $b_0 = 0$ cannot be rejected. This means, that the regression must be repeated using simpler model $y = b_1 x$.

Table 5.4. Regression results assuming simpler model $y = b_1x$.

SUMMARY OUTPUT

Regression Statistics

Multiple R

R Square

Adjusted R Square

Standard Error

Observations

0.970381124

0.941639525

0.841639525

0.131039699

11

ANOVA

F_{exp}(b₀)=

t²(b₀)=

1.032925461

1.032925461

F(0.05,1,10)=

sy=

5.117355029

0.13

df

SS

MS

F

Significance F

Regression

1

2.770585974

2.770585974

161.348845

4.73643E-07

Residual

10

0.171714026

0.017171403

Total

11

2.9423

F.DIST.RT(161.348845,1,10)=

1.70818E-07

Coefficients

Standard Error

t Stat

P-value

Lower 95%

Upper 95%

Intercept

0

#N/A

#N/A

#N/A

#N/A

#N/A

X Variable 1

0.424

0.033

12.70231652

1.7082E-07

0.350

0.499

Now, the test F for the importance of the parameter b_0 can be carried out. Using Eq. (5.17) where S_1^2 corresponds to the simpler model and S_2^2 to the full model:

$$F_{\text{exp}} = \frac{0.171714026 - 0.154035455}{0.017115051} = 1.03292546 \quad (5.25)$$

This value should be compared with $F(0.05, 1, 9) = 5.117355$. Because $F_{\text{exp}} < F_{\text{cr}}(0.05, 1, 9)$ term b_0 is not statistically important and cannot be determined from the experimental data.

It should be noticed that the value of p : **Significance F is calculated incorrectly** in Excel. This value should be $\text{F.DIST.RT}(F, 1, 10) = \text{F.DIST.RT}(161.3488, 1, 10) = 1.70818 \times 10^{-7}$ and not $4.736431291 \times 10^{-7}$ as it is in Excel ($F = H49 = 161.3488$). The latter value was calculated with wrong number of degrees of freedom (9 instead of 10) as $\text{F.DIST.RT}(F, 1, 9)$. With this correction the significance value for t -test and F -test are the same (1.7082×10^{-7})!

The **final model** describing the data is:

$y = b_1 x$ with:

$$r^2 = 0.9416, b_1 = 0.424, s_{b_1} = 0.033, s_y^2 = 0.017171, s_y = 0.13,$$

and the confidence interval for the regression parameter assuming 95% confidence is:

$$0.350 \leq b_1 \leq 0.499 \text{ or } b_1 = 0.424 \pm 0.074.$$

The confidence intervals are directly displayed as *Lower 95%* and *Upper 95%* while standard deviation of b_1 (X Variable 1) is called *Standard Error*.

As it has been seen **t and F tests give the same answer**. It is because:

$$t^2(\alpha, df) = F(\alpha, 1, df) \quad (5.26)$$

In fact, $t_{\text{exp}} = 1.0163294056275$, and $t_{\text{exp}}^2 = 1.03292546$ which is the value F_{exp} calculated in Eq. (5.25). Besides, $t(0.05, 9) = 2.262157$, $t^2(0.05, 9) = 5.117355 = F(0.05, 1, 9)$.

All the calculations are in *Examples4.xlsx*, sheet *Ex. 5.3*.

Example 5.4.

Calculate regression coefficients for the following data:

x	y
-2.0	0.27
-1.8	0.19
-1.6	0.36
-1.4	0.49
-1.2	0.50
-1.0	0.57
-0.8	0.20
-0.6	0.42
-0.4	0.57
-0.2	0.37
0.0	0.43

The results of the regression analysis in Excel are given below.

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.383636
R Square	0.147176
Adjusted R Square	0.052418
Standard Error	0.130825
Observations	11

<i>ANOVA</i>					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	0.026583	0.026583	1.553178	0.244129
Residual	9	0.154035	0.017115		
Total	10	0.180618			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	0.475	0.073795	6.436753	0.00012	0.308064	0.641936
X Variable 1	0.077727	0.062368	1.246266	0.244129	-0.06336	0.218814
		$t_{cr}(0.05,9)=$	2.262157			

The plot of the experimental and regression data (from Origin) is displayed in Fig. 5.7.

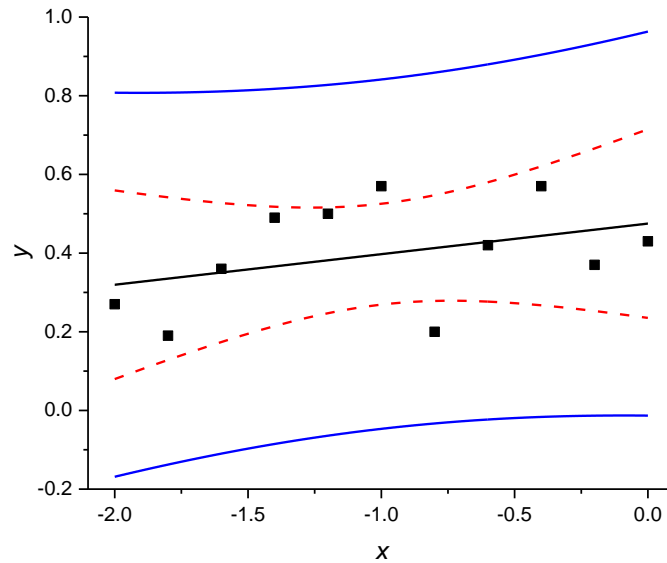


Fig. 5.7. Plot of the experimental points, calculated regression (black line), confidence intervals for the calculated data (dashed red lines), and the confidence intervals for the experimental points (continuous blue lines) for data in Example 5.4.

First of all one can notice that the **determination coefficient** $r^2 = 0.142$, which is a very low meaning that only 14.2% of the total variation of y can be explained by the regression.

Then, the **F -test for the importance of slope**, $F_{\text{exp}} = 1.553$ is much lower than the value $F_{\text{cr}}(0.05, 1, 9) = 5.117$. The **Significance F** equals $p = 0.224$, much larger than the assumed value of the significance level of 0.05. This suggests that the hypothesis $H_0: b_1 = 0$ cannot be rejected and the slope is not statistically important.

Finally, **test t** reveals $t_{\text{exp}} = 1.246$, much lower than the value $t_{\text{cr}}(0.05, 9) = 2.262$.

All these tests indicate that the parameter b_1 is not statistically significant and there is no statistically important relation between y and x . In this case the data should be described by an arithmetic mean. The results obtained using Descriptive Statistics in Excel are shown below.

<i>Column1</i>	
Mean	0.397
Standard Error	0.040
Median	0.42
Mode	0.57
Standard Deviation	0.13
Sample Variance	0.018062
Kurtosis	-1.02408
Skewness	-0.32632
Range	0.38
Minimum	0.19

Maximum	0.57
Sum	4.37
Count	11
Confidence Level(95.0%)	0.090

The results may be presented as:

$$\bar{y} = \mathbf{0.397}, s_y = \mathbf{0.13}, s_{\bar{y}} = \mathbf{0.04} \text{ (11 points, } df = 10\text{)}$$

Confidence intervals (95%)

$$0.397 - 0.090 \leq \bar{y} \leq 0.397 + 0.090$$

that is:

$$\mathbf{0.307 \leq \bar{y} \leq 0.488 \text{ or } \bar{y} = 0.397 \pm 0.090}$$

Example 5.5.

Analyze the data below and find the model describing them.

x	y
0	0.4
1	0.7
2	1.3
3	1.7
4	2.1
5	2.5
6	2.4
7	2.9
8	3.3
9	3.3
10	3.6
11	3.7
12	3.9
13	4.1
14	4.3
15	4.3

Usually, the first step is to visualize the data. They are presented in Fig. 5.8.

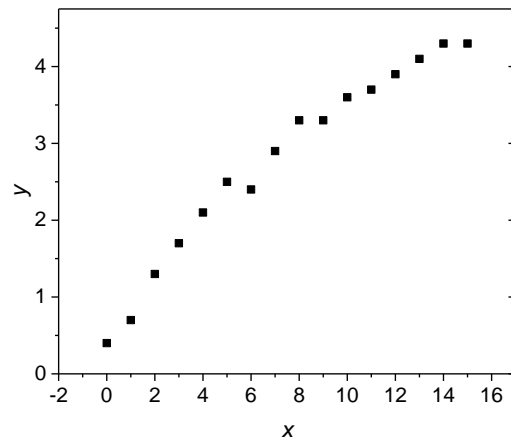


Fig. 5.8. Plot of data in Example 5.5.

The plot displays relation between y and x . One should start with the simplest model in this case a linear regression. The results of the linear regression in Excel are presented in Table 5.5.

Table 5.5. Results of the linear fit to the data in Example 5.5.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.978149
R Square	0.956775
Adjusted R Square	0.953688
Standard Error	0.270645
Observations	16

ANOVA					
	df	SS	MS	F	Significance F
Regression	1	22.69889	22.69889	309.8869	6.02E-11
Residual	14	1.025485	0.073249		
Total	15	23.72438			

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	0.843382	0.129215	6.526982	1.34E-05	0.566244	1.12052
X Variable 1	0.258382	0.014678	17.6036	6.02E-11	0.226902	0.289863

The plot of the regression results together with the confidence intervals is shown in Fig. 5.9.

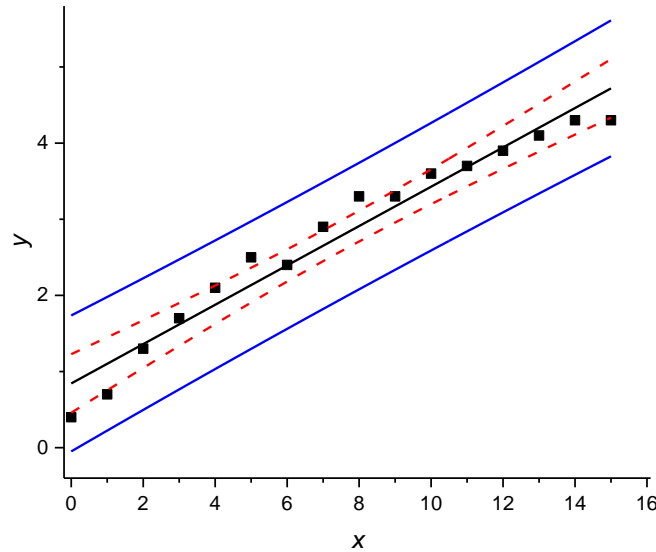


Fig. 5.9. Plot of the experimental points, regression line (black), confidence intervals for \hat{y}_i calculated (red dashed lines), and that for y_i experimental (continuous blue lines) assuming linear model: $y = b_0 + b_1x$.

Inspection of the regression results reveals $r^2 = 0.9568$ (good correlation).

F -test of the regression $F_{\text{exp}} = 309.9$ shows that the hypothesis $y = b_0$ i.e. $b_1 = 0$ must be rejected, $F_{\text{cr}}(0.05, 1, 14) = 4.600$ and the *Significance of F*, $p = 6.02 \times 10^{-11}$ is very low, much lower than 0.05.

t -tests of the significance of the parameters b_0 and b_1 are 6.53 and 17.60, respectively, much larger than $t_{\text{cr}}(0.05, 14) = 2.145$ therefore both parameters are statistically important.

However, inspection of the residuals reveal that they are not randomly distributed and show a parabolic dependence, Fig. 5.10. This might suggest that our model is not adequate. In fact, the experimental data could be approximated by a second order equation: $y = b_0 + b_1x + b_2x^2$. Let us use this model to analyze the experimental data. Another column of x^2 must be added to data in the Excel file:

x	x^2	y
0	0	0.4
1	1	0.7
2	4	1.3
3	9	1.7
4	16	2.1
5	25	2.5
6	36	2.4
7	49	2.9
8	64	3.3
9	81	3.3

10	100	3.6
11	121	3.7
12	144	3.9
13	169	4.1
14	196	4.3
15	225	4.3

In the Regression in Excel two first column (x and x^2) must be chosen in “Input X Range”. The results obtained are displayed below in Table 5.6.

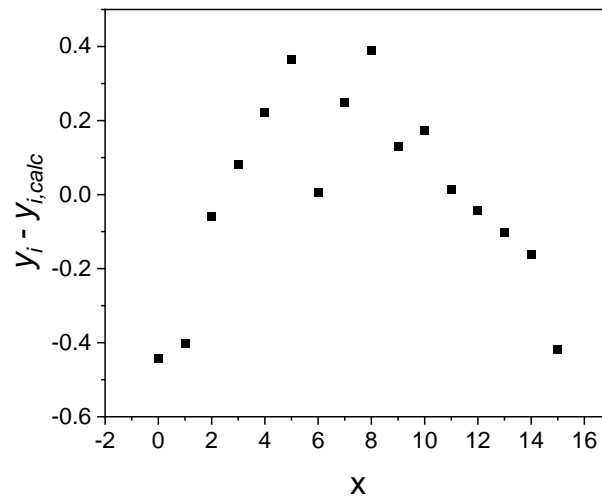


Fig. 5.10. Plot of residuals $y_i - \hat{y}_i$ for linear regression model.

Table 5.6. Results of the parabolic regression approximation of the data in Example 5.5.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.996305
R Square	0.992624
Adjusted R Square	0.99149
Standard Error	0.116017
Observations	16

ANOVA					
	df	SS	MS	F	Significance F
Regression	2	23.5494	11.7747	874.7951	1.38E-14
Residual	13	0.174979	0.01346		
Total	15	23.72438			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	0.416299	0.077167	5.394798	0.000122	0.24959	0.583008
X Variable 1	0.441418	0.02387	18.49245	1.02E-10	0.38985	0.492986
X Variable 2	-0.0122	0.001535	-7.94908	2.4E-06	-0.01552	-0.00889

Then plot for parabolic model is displayed in Fig. 5.11 and the residual plot in Fig. 5.12.

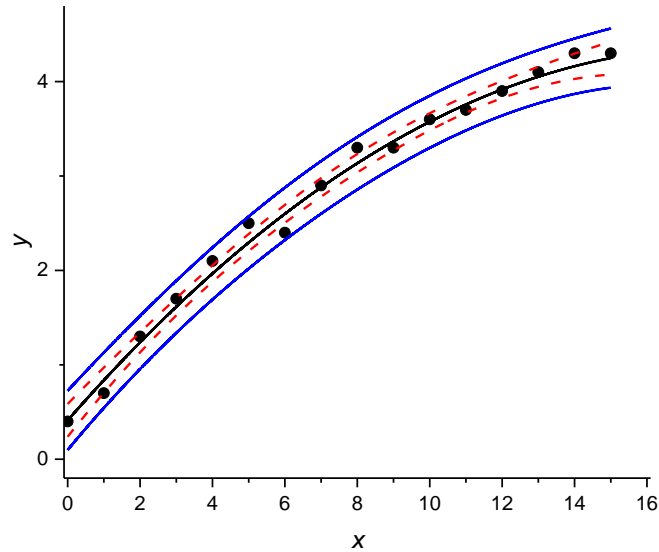


Fig. 5.11. Plot of the parabolic fit to the data in Example 5.5, experimental points, regression line (black), confidence intervals for \hat{y}_i calculated (red dashed lines), confidence interval for y_i experimental (continuous blue lines) assuming parabolic model: $y = b_0 + b_1x + b_2x^2$.

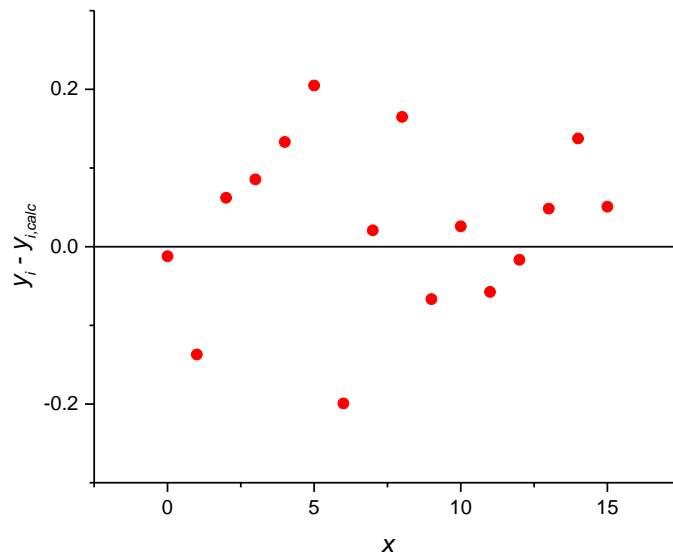


Fig. 5.12. Plot of residuals for the parabolic model fit to data in Example 5.5.

Visual inspection suggests that parabolic plot is better and the residuals are distributed more randomly, Fig. 5.12. The determination coefficient is $r^2 = 0.9926$, larger than for the linear plot (0.9568).

The F -test value is large $F_{\text{exp}} = 874.8$ and *Significance of F* , $p = 1.38 \times 10^{-14}$ very small therefore model $y = b_0$ is very little probable.

t -tests for the parameters b_0 , b_1 , and b_2 , called Intercept, X Variable 1, and X Variable 2, respectively are:

$$t_{b_0} = \frac{b_0}{s_{b_0}} = 5.395$$

$$t_{b_1} = \frac{b_1}{s_{b_1}} = 18.49$$

$$t_{b_2} = \left| \frac{b_2}{s_{b_2}} \right| = 7.949$$

All these values are larger than the value of $t_{\text{cr}}(0.05, 13) = 2.160$ ($df - N - 3 = 13$), therefore all the parameters are statistically important.

It is also possible to assess the importance of the parameter b_2 using F test for addition of the parameter b_2 , Eq. (5.14), where S_1^2 corresponds to the residual sum of squares for the linear approximation and S_2^2 residual sum of squares for the parabolic approximation. Taking the values of sums of squares from ANOVA one can obtain:

$$F_{\text{exp}}(b_2) = \frac{1.025485 - 0.174979}{0.01346} = 952.8 \quad (5.27)$$

which should be compared with $F(0.05, 1, 13) = 4.667$. These results indicate that the parameter b_2 is highly significant. The regression results with probability 95% might be presented as:

$$r^2 = 0.9926$$

$$b_0 = 0.416, s_{b_0} = 0.077, \quad 0.25 \leq b_0 \leq 0.58 \quad \text{or} \quad b_0 = 0.42 \pm 0.17$$

$$b_1 = 0.441, s_{b_1} = 0.0240, \quad 390 \leq b_1 \leq 0.493 \quad \text{or} \quad b_1 = 0.441 \pm 0.052$$

$$b_2 = -0.0122, s_{b_2} = 0.0015, \quad -0.01552 \leq b_2 \leq -0.00899 \quad \text{or} \quad b_2 = -0.0122 \pm 0.0033.$$

All the calculations are in Excel file *Examples5.xlsx*, sheet *Ex. 5.5 linear* and *Ex. 5.5 parabolic*.

5.4 Tests in multiple regression

Let us start first with the example of multiple regression.

Example 5.6.

Find equation describing the following data:

x_1	x_2	y
35.3	20	10.98
29.7	20	11.13
30.8	23	12.51
58.8	20	8.40
61.4	21	9.27
71.3	22	8.73
74.4	11	6.36
76.7	23	8.50
70.7	21	7.82
57.5	20	9.14
46.4	20	8.24
28.9	21	12.19
28.1	21	11.88
39.1	19	9.57
46.8	23	10.94
48.5	20	9.58
59.3	22	10.09
70.0	22	8.11
70.0	11	6.83
74.5	23	8.88
72.1	20	7.68
58.1	21	8.47
44.6	20	8.86
33.4	20	10.36
28.6	22	11.08

First, we can check two simpler regressions:

$$y = b_{0,1} + b_{1,1}x_1 \quad (5.28)$$

$$y = b_{0,2} + b_{1,2}x_2 \quad (5.29)$$

where the regression parameters in both equations are different. This can be simply done and the results are shown below for Eq. (5.28), variable x_1 :

SUMMARY OUTPUT

 $y=f(x_1)$

<i>Regression Statistics</i>	
Multiple R	0.845244
R Square	0.714438
Adjusted R Square	0.702022
Standard Error	0.890125
Observations	25

ANOVA		F(0.05,2,23) 3.422132			
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	45.5924	45.592402	57.54279	1.05495E-07
Residual	23	18.2234	0.7923217		
Total	24	63.8158			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	13.62299	0.581463	23.428795	1.5E-17	12.42014039	14.82584
X Variable 1	-0.07983	0.010524	-7.585697	1.05E-07	-0.101598379	-0.05806
	t(0.05,23)=	2.068658				

and for Eq. (5.29), variable x_2 :

SUMMARY OUTPUT

 $y=f(x_2)$

<i>Regression Statistics</i>	
Multiple R	0.536122
R Square	0.287427
Adjusted R Square	0.256446
Standard Error	1.406095
Observations	25

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	18.3424	18.3424	9.277409	0.005736
Residual	23	45.4734	1.977104		
Total	24	63.8158			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	3.560549	1.945473	1.830171	0.080215	-0.46397	7.585067
X Variable 1	0.289696	0.095111	3.045884	0.005736	0.092945	0.486448

In first case, Eq. (5.28), the t -test indicates that parameters $b_{0,1}$ and $b_{1,1}$ are important and $t_{\text{exp}} > t_{\text{cr}}(0.05, 23) = 2.07$. One can also notice that in the second case, Eq. (5.29) the parameters $b_{0,2}$ is not important while parameter $b_{1,2}$ is significant. However, both correlation coefficients are low, for the first equation it is $r^2 = 0.714$ and for the second very low, 0.287 which means that simple equations can explain 71.4% and 28.7% of the total variation of $y = f(x_i)$.

We can now postulate a multiple linear regression:

$$y = b_0 + b_1x_1 + b_2x_2 \quad (5.30)$$

The results using Excel are:

SUMMARY OUTPUT

y=f(x1,x2)

Regression Statistics	
Multiple R	0.921476
R Square	0.849117
Adjusted R Square	0.835401
Standard Error	0.661565
Observations	25

First, we can notice that the determination coefficient increased to 0.849. However, determination (and correlation) coefficient increases always with the increase of the number of parameters. To compare correlations which have different number of variables one can use **adjusted correlation coefficient**:

$$r_a^2 = 1 - \frac{N-1}{N-df-1} (1-r^2) \quad (5.31)$$

where df is the number of degrees of freedom in regression, that is number of parameters, n , minus one, $df = n - 1$. This value is displayed in ANOVA and in our case it is 1 for the simple linear regression ($n = 2$) and 2 for multiple regression ($n = 3$), see sheet *Ex.5.6* in *Examples5.xlsx*. In our

case the adjusted correlation coefficient is 0.702 for Eq. (5.28), 0.256 for Eq. (5.29) and it increases to 0.835 for the multiple regression, Eq. (5.30).

Next one can notice that the test t_{exp} for all three parameters: 8.28, 9.05, and 4.43, is larger than the critical value for $t_{\text{cr}}(0.05, 22) = 2.074$. This is confirmed by p -level tests; these values for all the parameters are much lower than the value 0.05 which confirms that all three parameters are important.

One can also perform sequential test F for the importance of adding variable x_2 (parameter b_2) to Eq. (5.28). There are two hypotheses:

$$H_0 \ y = b_{0,1} + b_{0,1} x_1$$

$$H_1 \ y = b_0 + b_1 x_1 + b_2 x_2$$

then the sequential F test is:

$$F_{\text{exp}} = \frac{18.223 - 9.629}{0.4377} = 19.64 \quad (5.32)$$

which is larger than $F_{\text{cr}}(0.05, 1, 22) = 4.30$. Similarly, test of adding variable x_1 to Eq. (5.29) for testing hypotheses:

$$H_0 \ y = b_{0,2} + b_{0,2} x_2$$

$$H_1 \ y = b_0 + b_1 x_1 + b_2 x_2$$

$$F_{\text{exp}} = \frac{45.473 - 9.629}{0.4377} = 81.90 \quad (5.33)$$

This confirms importance of the multiple regression, Eq. (5.30).

In the case of more parameters the analysis demands to verify more correlations.

5.5 Akaike information criterion

Akaike information criterion, AIC , is used to compare statistical models in order to choose the best one.⁵⁹⁻⁶² It is based on the information theory. It deals with the risks of over and underfitting and finding the optimal model. However, it tells nothing about the absolute quality of a model, it gives only the quality relative to other models. But when all the models give poor fits, AIC will not give any warning.

AIC contain two components. First, characterizes goodness of fit and the second gives penalty for the number of adjustable parameters in the model:

$$AIC = -2 \ln(\hat{L}) + 2df \quad (5.34)$$

where \hat{L} is the likelihood function which characterises goodness of fit and contains the residual sum of squares of the model and df is the number of estimated parameters. The preferred model is the one with the smallest AIC value. If errors are normally distributed the model has $df = p + 1$ degrees of freedom where p is the number of parameters determined and one degree is for σ^2 .

5.5.1 General equation

Maximum likelihood function is defined as:

$$\hat{L} = \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(y_i - \hat{y}_i)^2}{2\sigma_i^2}} \quad (5.35)$$

where N is the number of points, σ_i is the standard deviation of point i , y_i is the experimental value of the approximated function and \hat{y}_i is the value calculated using the studied model. For the optimal values of the parameters this function reaches maximum. Logarithm of \hat{L} is:

$$\ln(\hat{L}) = -\frac{1}{2} \sum_{i=1}^N \ln(2\pi) - \frac{1}{2} \sum_{i=1}^N \ln \sigma_i^2 - \frac{1}{2} \sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{\sigma_i^2} \quad (5.36)$$

and

$$-2 \ln(\hat{L}) = N \ln(2\pi) + \sum_{i=1}^N \ln(\sigma_i^2) + \sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{\sigma_i^2} \quad (5.37)$$

AIC is then:

$$AIC = N \ln(2\pi) + \sum_{i=1}^N \ln(\sigma_i^2) + \sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{\sigma_i^2} + 2(p+1) \quad (5.38)$$

5.5.2 Unit weights

When unit weights are used all the standard deviations are the same:

$$\sigma_i^2 = \sigma^2 = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y}_i)^2 = \frac{RSS}{N} \quad (5.39)$$

$$RSS = \sum_{i=1}^N (y_i - \hat{y}_i)^2$$

where RSS is the residual sum of squares. Then

$$\sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{\sigma_i^2} = \frac{1}{\sigma^2} \frac{\sum_{i=1}^N (y_i - \hat{y}_i)^2}{N} = \frac{RSS}{\sigma^2} = N \quad (5.40)$$

$$\sum_{i=1}^N \ln(\sigma_i^2) = N \ln(\sigma^2) = N \ln\left(\frac{RSS}{N}\right) = -N \ln N + N \ln(RSS) \quad (5.41)$$

and AIC for unit weights is:

$$AIC = N \ln(2\pi) + N \ln(\sigma^2) + N + 2(p+1) \quad (5.42)$$

or

$$AIC = N \ln(2\pi) - N \ln N + N \ln(RSS) + N + 2(p+1) \quad (5.43)$$

5.5.3 Proportional weights

Let us consider another case of proportional weights, i.e. standard deviation proportional to the estimated value of \hat{y}_i ,

$$\sigma_i = \alpha \hat{y}_i \quad (5.44)$$

Eq. (5.37) may be written as:

$$\begin{aligned}
-2\ln(\hat{L}) &= N\ln(2\pi) + \sum_{i=1}^N \ln(\alpha^2 \hat{y}_i^2) + \sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{\alpha^2 \hat{y}_i^2} \\
&= N\ln(2\pi) + N\ln(\alpha^2) + \sum_{i=1}^N \ln(\hat{y}_i^2) + \frac{1}{\alpha^2} \sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i^2}
\end{aligned} \tag{5.45}$$

To determine optimal value of α^2 let us calculate the derivative

$$\frac{\partial[-2\ln(\hat{L})]}{\partial\alpha^2} = \frac{N}{\alpha^2} - \frac{1}{(\alpha^2)^2} \sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i^2} = 0 \tag{5.46}$$

$$\alpha^2 = \frac{1}{N} \sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i^2} \tag{5.47}$$

Substitution gives:

$$-2\ln(\hat{L}) = N\ln(2\pi) - N\ln N + N + \sum_{i=1}^N \ln \hat{y}_i^2 + N \ln \sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i^2} \tag{5.48}$$

and AIC is

$$AIC = N\ln(2\pi) - N\ln N + N + \sum_{i=1}^N \ln \hat{y}_i^2 + N \ln \sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{\hat{y}_i^2} + 2(p+1) \tag{5.49}$$

5.5.4 General weights problem

Let us consider a case of weighted regression where the standard deviation of each y_i is s_i . In such a case

$$\sigma_i = \alpha s_i \tag{5.50}$$

where parameter α corrects for the quality of fit. Substitution in Eq. (5.37) gives:

$$\begin{aligned}
-2\ln(\hat{L}) &= N\ln(2\pi) + \sum_{i=1}^N \ln(\sigma_i^2) + \sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{\sigma_i^2} \\
&= N\ln(2\pi) + \sum_{i=1}^N \ln(\alpha^2 s_i^2) + \sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{\alpha^2 s_i^2} \\
&= N\ln(2\pi) + N\ln(\alpha^2) \sum_{i=1}^N \ln(s_i^2) + \frac{1}{\alpha^2} \sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{s_i^2}
\end{aligned} \tag{5.51}$$

To determine the value of α^2 one should find the derivative of $-2\ln(\hat{L})$ and its minimum:

$$\frac{\partial[-2\ln(\hat{L})]}{\partial\alpha^2} = \frac{N}{\alpha^2} - \frac{1}{(\alpha^2)^2} \sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{s_i^2} = 0 \tag{5.52}$$

which gives:

$$\alpha^2 = \frac{1}{N} \sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{s_i^2} \quad (5.53)$$

Substitution into Eq. (5.51) gives:

$$-2\ln(\hat{L}) = N \ln(2\pi) - N \ln N + N + \sum_{i=1}^N \ln s_i^2 + N \ln \sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{s_i^2} \quad (5.54)$$

and AIC is:

$$AIC = N \ln(2\pi) - N \ln N + N + \sum_{i=1}^N \ln s_i^2 + N \ln \sum_{i=1}^N \frac{(y_i - \hat{y}_i)^2}{s_i^2} + 2(p+1) \quad (5.55)$$

5.5.5 Corrected AIC

When the number of points in the sample is small there is a danger that AIC as defined in Eq. (5.34) will select models with too many parameters (overfit). In such cases a corrected AIC , AIC_c , should be used:

$$AIC_c = AIC + \frac{2(p+1)(p+2)}{N-p-2} \quad (5.56)$$

$$AIC_c = -2\ln(\hat{L}) + \frac{2(p+1)N}{N-p-2}$$

which adds penalty when N is small; when $N \rightarrow \infty$ the second term decreases to zero. When comparing models, the one with the smallest AIC (AIC_c) should be selected.

5.5.6 Akaike weights

To help with such selection Akaike weight were introduced. First, when comparing few different models $\Delta(AIC)$ should be calculated:

$$\Delta_i(AIC) = AIC_i - AIC_{\min} \quad (5.57)$$

where AIC_{\min} is the smallest value for all the models. Next, the relative weights of models are calculated:

$$w_i(AIC) = \frac{\exp\left[-\frac{1}{2}\Delta_i(AIC)\right]}{\sum_{i=1}^n \exp\left[-\frac{1}{2}\Delta_i(AIC)\right]} \quad (5.58)$$

The model with the largest weight is the most probable. It should be noticed that there is no clear criterion (statistical test) which would say if the difference between models is significant.

All these values might be calculated in Excel, but they are already included in free statistical software R. The calculations of the AIC_c and weight is correct in package `AICcmodavg` and incorrect in `qpcR` (wrong number of degrees of freedom). One can calculate these parameters in `AICcmodavg` (rather complicated method for weights) but otherwise, after calculation of AIC_c , one can change the package and calculate weights using `qpcR`. Below, three examples are presented, the details of calculations are in the Excel file.

Example 5.7

Compare which model better describes the data (data file '10'): $y = b_1 x$ or $y = b_0 + b_1 x$. Use classical statistics and AIC criterion. The results are in Examples5.xls, sheet Ex. 5.7.

x	y
-2.0	-0.838
-1.8	-0.838
-1.6	-0.588
-1.4	-0.378
-1.2	-0.288
-1.0	-0.138
-0.8	-0.428
-0.6	-0.128
-0.4	0.102
-0.2	-0.018
0.0	0.122

The results of fits are displayed in Fig. 5.13. Fit of the experimental data (points) to the models: $y = b_1 x$ (black line) and $y = b_0 + b_1 x$ (red line) in Example 5.7..

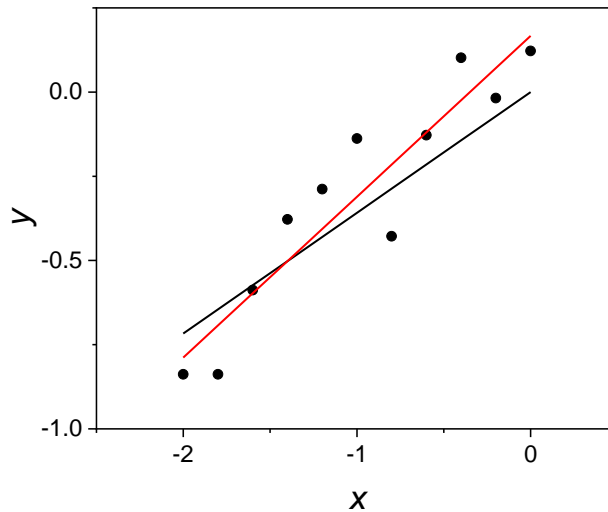


Fig. 5.13. Fit of the experimental data (points) to the models: $y = b_1 x$ (black line) and $y = b_0 + b_1 x$ (red line) in Example 5.7.

The classical analysis may be performed using the sequential F-test, Eq. (5.14), for adding a new parameter:

$$F_{\text{exp}} = \frac{\text{RSS}_1 - \text{RSS}_2}{s_{y,2}^2} = \frac{0.241686597 - 0.1540355}{0.017115} = 5.121 \quad (5.59)$$

which should be compared with $F_{\text{cr}}(0.05, 1, 9) = 5.117$. Because F_{exp} and F_{cr} are practically the same one cannot say that full linear model is better than that without b_0 and, with the probability of 95%, one should reject full linear model and accept the simpler one $y = b_1 x$.

However, using the *AIC* criterion the following results are obtained:

model	$y=b_1 x$	$y=b_0+b_1 x$
<i>AIC</i>	-6.781	-9.736
<i>AICc</i>	-5.281	-6.308
<i>AICc</i> rel. weights	0.598	1

AIC criterion indicates that the linear model is better (*AICc* is lower) and its weight is larger (1 vs. 0.598). The ratio of *AIC* weights is 1.67 times larger for the full linear model than for the simpler without b_0 . However, it does not say which model should be kept. Classical F-test suggests that at $\alpha = 0.05$ the simpler model should be kept.

The detailed calculations are presented in Excel *Examples5.xlsx*, sheet *Ex. 5.7* and compared with the results of the calculations in R. The R program is 'b0'.

Example 5.8

Another example will be for comparing linear and parabolic regressions using classical and *AIC* criterions. Data are in file '5'. Results are in *Examples5.xlsx* sheet *Ex 5.8*.

x	y
0	0.40
1	0.71
2	1.34
3	1.78
4	2.24
5	2.72
6	2.72
7	3.34
8	3.87
9	4.02
10	4.49
11	4.77
12	5.18
13	5.60
14	6.04
15	6.30

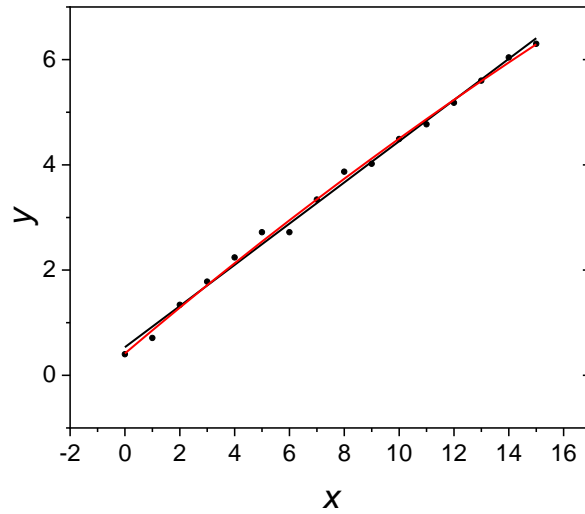


Fig. 5.14. Fit of the experimental data (points) to the linear (black line) and parabolic (red line) models in Example 5.8.

Regression in Excel shows that

$$F_{\text{exp}} = \frac{\text{RSS}_{\text{lin}} - \text{RSS}_{\text{parab}}}{s_{y,\text{parab}}^2} = \frac{0.237604706 - 0.17467}{0.013436} = 4.68 \quad (5.60)$$

the value of $F_{\text{cr}}(0.05, 1, 13) = 4.67$ which is practically the same as F_{exp} .

The results of AIC criterion are shown below:

model	linear	parabolic
AIC	-15.950	-18.873
$AICc$	-13.950	-15.237
$AICc$ rel. weights	0.52546	1

These results indicate that parabolic model shows lower value of $AICc$ and its weight is 1.9 time larger. It looks better and displays lower s^2 , but it is not statistically more important at the confidence level of 95%.

Example 5.9

This example shows calculation of the AIC parameters for weighted regression. Compare which model: linear or parabolic is better? Use classical statistics and AIC criterion. Data are in file *dataw* in E5-9 and the results in *Examples5.xlsx* sheet *Ex 5.9*.

x	y	s_i
0	1.9	0.4
1	2.3	0.5
2	3.5	0.7
3	4.5	0.9
4	5.2	1.0
5	6.0	1.2
6	5.5	1.1

In this case the regression parameters cannot be simply calculated in Excel but one can also use program *polfit.exe* (in exercises to Error analysis and data modeling, Part 1), Origin or R.

Comparison of the results obtained using *polfit* and in R for weighted regression is shown in sheet Ex. 5.9.

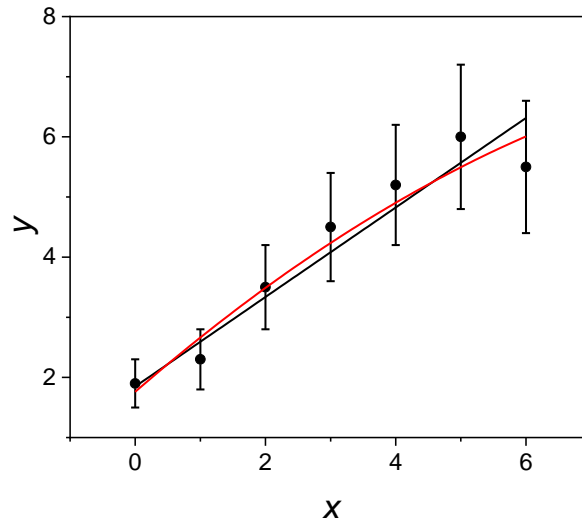


Fig. 5.15. Weighted fit of the experimental data (points) to the linear (black line) and parabolic (red line) models in Example 5.9.

Use of F-test for addition of the parabolic term b_2 is:

$$F_{\text{exp}} = \frac{\text{RSS}_{\text{lin}} - \text{RSS}_{\text{parab}}}{s_{y,\text{parab}}^2} = \frac{1.4465 - 1.2146}{0.30364} = 0.76373 \quad (5.61)$$

F_{exp} is much smaller than the critical value of $F(0.05, 1, 4) = 7.709$ which indicated that addition of a new term is statistically unimportant. This is of course confirmed by the t -test for b_2 which is $t_{\text{exp}} = 0.874$ while $t(0.05, 4) = 2.776$ as these two tests are identical ($t_{\text{exp}}^2 = F_{\text{exp}}$ because $t(\alpha, k)^2 = F(\alpha, 1, k)$). These tests confirm that parabolic model is not justified and linear model should be kept.

The *AIC* criterion gives the following results:

model	Linear	Parabolic
<i>AIC</i>	11.24	12.02
<i>AICc</i>	19.24	32.02
<i>AICc</i> relative	0	12.78
Rel. weights	1	0.001681

The relative weight of the parabolic model is very small (595 time less important than the linear one) which suggests that linear weighted model should be kept.

Data, program in R and the results are also in the folder *E5-9*.

Example 5.10.

Let us consider more complex example of model selection. For the data file *parab-comp*:

<i>x</i>	<i>y</i>
0	0.4
1	2.0
2	2.9
3	5.7
4	7.0
5	13.5
6	15.1
7	20.5
8	18.7
9	37.1
10	44.9
11	70.0
12	63.7
13	50.5
14	115.5
15	113.6

find the best model describing it between the following:

- 1 $y = b_0 + b_1 x$
- 2 $y = b_0 + b_1 x + b_2 x^2$
- 3 $y = b_0 + b_2 x^2$
- 4 $y = b_2 x^2$
- 5 $y = b_1 x + b_2 x^2$

Use classical statistics and AIC criterion.

The results in Examples5.xlsx sheet Ex3.10 obtained using R program *parab-comp-model* using *t*-test show that in models 2, 3, and 5 there are statistically unimportant parameters for which $t_{\text{exp}} < t_{\text{cr}}(0.05, df)$:

Model					
1	t(0.05,14)=	2.144787			
glm(formula = y ~ x)					
Deviance Residuals:					
Min	1Q	Median	3Q	Max	
-25.300	-10.103	-4.534	9.503	32.522	
Coefficients:					
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-17.5191	7.8606	-2.229	0.0427	*
x	7.1784	0.8929	8.039	1.29E-06	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
(Dispersion parameter for gaussian family taken to be 271.0775)					
Null deviance: 21315.0 on 15 degrees of freedom					
Residual deviance: 3795.1 on 14 degrees of freedom					
AIC: 138.91					

Model					
2	t(0.05,13)=	2.160369			
glm(formula = y ~ x + I(x^2))					
Deviance Residuals:					
Min	1Q	Median	3Q	Max	
-30.791	-1.211	0.861	2.142	19.710	
Coefficients:					
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	3.8332	7.5368	0.509	0.61956	
x	-1.9726	2.3314	-0.846	0.41279	
I(x^2)	0.6101	0.1499	4.069	0.00133	**

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
(Dispersion parameter for gaussian family taken to be 128.399)					
Null deviance: 21315.0 on 15 degrees of freedom					
Residual deviance: 1669.2 on 13 degrees of freedom					
AIC: 127.77					

Model					
3	t(0.05,14)=	2.144787			
glm(formula = y ~ I(x^2))					
Deviance Residuals:					
Min	1Q	Median	3Q	Max	
-30.443	-2.038	1.276	2.839	21.389	
Coefficients:					
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-1.47768	4.12944	-0.358	0.726	
I(x^2)	0.4877	0.03912	12.468	5.72E-09	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
(Dispersion parameter for gaussian family taken to be 125.7934)					
Null deviance: 21315.0 on 15 degrees of freedom					
Residual deviance: 1761.1 on 14 degrees of freedom					
AIC: 126.62					

Model					
4	t(0.05,15)=	2.13145			
glm(formula = y ~ 0 + I(x^2))					
Deviance Residuals:					
Min	1Q	Median	3Q	Max	
-30.1840	-2.8549	-0.1194	1.5331	21.9257	
Coefficients:					
	Estimate	Std. Error	t value	Pr(> t)	
I(x^2)	0.47742	0.02578	18.52	9.57E-12	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
(Dispersion parameter for gaussian family taken to be 118.4811)					
Null deviance: 42419.8 on 16 degrees of freedom					
Residual deviance: 1777.2 on 15 degrees of freedom					
AIC: 124.77					

Model					
5	t(0.05,14)=	2.144787			
glm(formula = y ~ 0 + x + I(x^2))					
Deviance Residuals:					
Min	1Q	Median	3Q	Max	
-30.8223	-0.1586	1.4942	3.2036	20.1245	
Coefficients:					
	Estimate	Std. Error	t value	Pr(> t)	
x	-0.9851	1.2559	-0.784	0.445882	
I(x^2)	0.557	0.1047	5.318	0.000109	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
(Dispersion parameter for gaussian family taken to be 121.6)					
Null deviance: 42419.8 on 16 degrees of freedom					
Residual deviance: 1702.4 on 14 degrees of freedom					
AIC: 126.08					

The two possible models are 1 and 4. One can use F -test for variances, Eq. (4.53), to decide if there is a statistical difference between them:

$$F_{\text{exp}} = \frac{s_1^2}{s_4^2} = \frac{271.077}{118.48} = 2.288 \quad (5.62)$$

The critical value $F_{\text{cr}}(0.05, 14, 15) = 2.424$ is larger than the experimental and one cannot say that one model is better than the other. Models 1 and 4 are probable at the probability 95%. The results of fitting models 1 and 4 to the experimental data are displayed in Fig. 5.16.

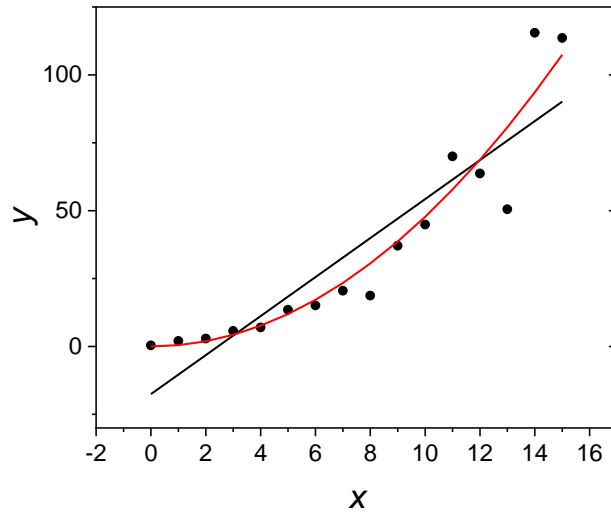


Fig. 5.16. Fit of the experimental data (points) to the linear model 1 (black line) and parabolic model 4 (red line) in Example 5.10.

Now, let us look at the results of AICc test. The results are shown below.

Model	1	2	3	4	5
AICc	140.908	131.402	128.624	125.693	128.081
rel. AICc	15.2155	5.7099	2.93125	0	2.38878
rel. weights	0.0005	0.05756	0.23093	1	0.30289

These results indicate that the most probable model is number 4, i.e. $y = b_2 x^2$ and the next is number 5, $y = b_1 x + b_2 x^2$. However, this test cannot tell us if it is much (statistically) better than others.

6 Interpolation

Interpolation is a method of obtaining new data points within the range of known discrete data points. Sometimes we know the values of the function in some points but would like to know the intermediate values between them and the cost of obtaining more data values is high. This might be necessary for plotting smooth functions, its differentiations or integration. The interpolating function **must pass by all the known values** and interpolate in between. There are few methods of interpolation discussed below. Of course, such interpolation introduces errors because it tries to interpolate the unknown function by some simple models.

6.1 Polynomial interpolation

Given a set of $N+1$ data points: $(x_0, y_0), \dots, (x_i, y_i), \dots, (x_N, y_N)$ there is a polynomial of the degree N which passes exactly through these points. To obtain such a polynomial it recommended to use Lagrange interpolating polynomial:

$$L_N(x) = y_0 l_0(x) + y_1 l_1(x) + \dots + y_N l_N(x) \quad (6.1)$$

where $l_i(x)$ are the polynomials of degree N defined as:

$$l_i(x) = \frac{(x-x_0)(x-x_1)\dots(x-x_{i-1})(x-x_{i+1})\dots(x-x_N)}{(x_i-x_0)(x_i-x_1)\dots(x_i-x_{i-1})(x_i-x_{i+1})\dots(x_i-x_N)} \quad (6.2)$$

where in each $l_i(x)$ the point i was omitted. These polynomials have a property:

$$l_i(x_j) = \delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases} \quad (6.3)$$

therefore

$$L_N(x_i) = y_i l_i(x_i) = y_i \quad (6.4)$$

It passes exactly through all the data points and it interpolates the values between these points.

Example 6.1

Let us interpolate data *ex* shown in Table 6.1 and Fig. 6.1 using Lagrange interpolation.

Table 6.1. Example data (7 points) for the Lagrange interpolation.

x	y
0	0.0
1	0.9
2	1.0
3	0.1
4	-0.8
5	-1.0
6	0.2

This interpolation can be carried out using program *polfit.exe* with 6th degree polynomial. The polynomial passes exactly by all the points and interpolates smoothly between the points. The results are in file *ex2* ($x, y_{\text{calculated}}$). Yet, as we do not know the functional dependence it is difficult

to judge the quality of interpolation. The results are displayed in Excel file *Examples6.xlsx*, sheet *Ex. 6.1* and data file *ex2* in folder *E6-1*.

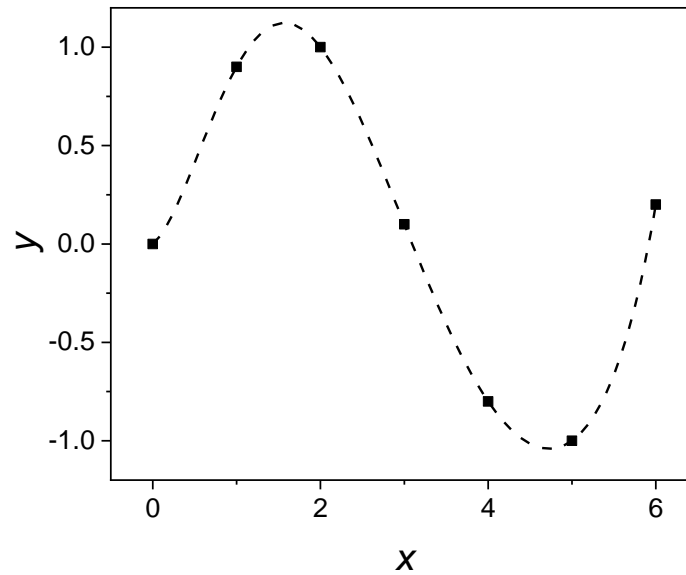


Fig. 6.1. Example of the interpolation of 7 points by the Lagrange polynomial of 6th degree.

However, in some cases such an approximation is not correct and Lagrange polynomial oscillates above and below the true function. Sometimes, such deviations are very large.

Example 6.2.

Let us consider so called Runge's phenomenon i.e. approximation of the Lorentzian-type function of the form:

$$y = \frac{1}{1 + 25x^2} \quad (6.5)$$

by the polynomial. Data set (11 points) in Table 6.2 (data file *runge*) were fitted to the Lagrange polynomial of 10th degree and the results are displayed in Fig. 6.2.

Table 6.2. Example of 11 data points obtained using Eq. (6.5).

x	y
-1.0	0.03846154
-0.8	0.05882353
-0.6	0.10000000
-0.4	0.20000000
-0.2	0.50000000
0.0	1.00000000
0.2	0.50000000
0.4	0.20000000
0.6	0.10000000
0.8	0.05882353
1.0	0.03846154

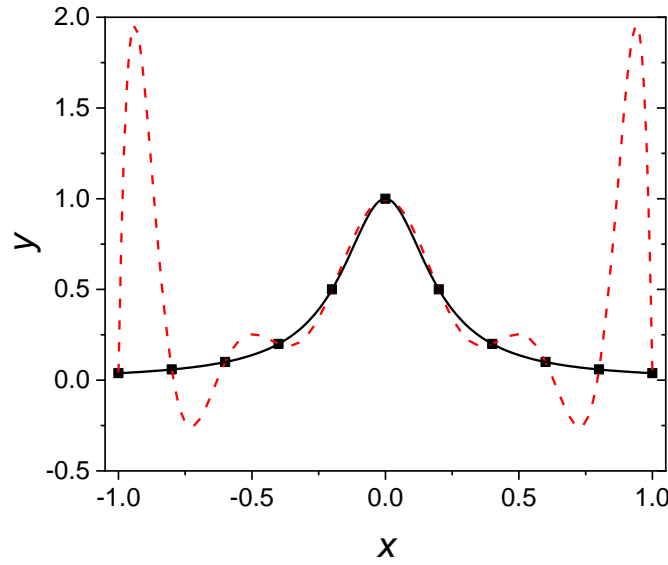


Fig. 6.2. Example of the interpolation of 11 data points in Table 6.2 using Lagrange polynomial of 10th degree; points – symbol, continuous line – function in Eq. (6.5), dashed line – Lagrange approximating polynomial.

It is evident that although the polynomial interpolates exactly all the points it oscillates between these points and such an interpolation is incorrect. To avoid these problems interpolation by the piecewise polynomials of low degree called splines are used.

6.2 Splines

Splines are formed by joining polynomials together at fixed points. The most popular are cubic splines that is polynomials of third degree:

$$p_i(x) = a_i x^3 + b_i x^2 + c_i x + d_i \quad (6.6)$$

Cubic spline interpolation of function $S(x)$ is defined by a series of cubic polynomials $p_i(x)$:

$$S(x) = \begin{cases} p_1(x) & x_0 \leq x \leq x_1 \\ p_2(x) & x_1 \leq x \leq x_2 \\ \dots & \dots \\ p_i(x) & x_{i-1} \leq x \leq x_i \\ \dots & \dots \\ p_N(x) & x_{N-1} \leq x \leq x_N \end{cases} \quad (6.7)$$

This indicated that between each two points a different cubic polynomial $p_i(x)$ is used. To obtain smooth continuous function $S(x)$ two consecutive polynomials must join at point i and their first and second derivatives should be the same:

$$\begin{aligned}
p_i(x_i) &= p_{i+1}(x_i) = y_i \\
p'_i(x_i) &= p'_{i+1}(x_i) \\
p''_i(x_i) &= p''_{i+1}(x_i)
\end{aligned} \tag{6.8}$$

The coefficients in Eq. (6.6) for all the cubic polynomials might be obtained following the procedure below.

The second derivative of Eq. (6.6) is:

$$p''_i(x) = 6a_i x + 2b_i \tag{6.9}$$

Let us call the second derivative in point i : M_i . The coefficients a_i and b_i can be found from the second derivatives in i and $i-1$:

$$\begin{aligned}
M_i &= y''(x_i) = 6a_i x_i + 2b_i \\
M_{i-1} &= y''(x_{i-1}) = 6a_i x_{i-1} + 2b_i \\
h_i &= x_i - x_{i-1} \\
a_i &= \frac{M_i - M_{i-1}}{6h_i} \\
b_i &= \frac{M_i x_{i-1} - M_{i-1} x_i}{2h_i}
\end{aligned} \tag{6.10}$$

Then $p''_i(x)$ can be expressed as:

$$\begin{aligned}
p''_i(x) &= \frac{(M_i - M_{i-1})x}{h_i} + \frac{-M_i x_{i-1} + M_{i-1} x_i}{h_i} \\
&= \frac{x_i - x}{h_i} M_{i-1} + \frac{x - x_{i-1}}{h_i} M_i
\end{aligned} \tag{6.11}$$

To find other coefficients c_i and d_i let us integrate $p''_i(x)$

$$p'_i(x) = -\frac{M_{i-1}(x_i - x)^2}{2h_i} + \frac{M_i(x - x_{i-1})^2}{2h_i} + c_i \tag{6.12}$$

and

$$p_i(x) = \int p'_i(x) dx = \frac{M_{i-1}(x_i - x)^3}{6h_i} + \frac{M_i(x - x_{i-1})^3}{6h_i} + c_i x + d_i \tag{6.13}$$

Now, we can use conditions $p_i(x_{i-1}) = y_{i-1}$ and $p_i(x_i) = y_i$ which give which give:

$$c_i = \frac{y_i - y_{i-1}}{h_i} - \frac{h_i(M_i - M_{i-1})}{6} \tag{6.14}$$

and

$$d_i = \frac{x_i y_{i-1} - x_{i-1} y_i}{h_i} - \frac{h_i(x_i M_{i-1} - x_{i-1} M_i)}{6} \tag{6.15}$$

Using the polynomial coefficients one can write the cubic spline i :

$$\begin{aligned}
p_i(x) &= \frac{M_{i-1}(x_i - x)^3}{6h_i} + \frac{M_i(x - x_{i-1})^3}{6h_i} + \left[\frac{y_i - y_{i-1}}{h_i} - \frac{h_i}{6}(M_i - M_{i-1}) \right] x \\
&+ \frac{x_i y_{i-1} - x_{i-1} y_i}{h_i} - \frac{h_i(x_i M_{i-1} - x_{i-1} M_i)}{6} = \\
&= \frac{M_{i-1}(x_i - x)^3}{6h_i} + \frac{M_i(x - x_{i-1})^3}{6h_i} + \left(\frac{y_{i-1}}{h_i} - \frac{M_{i-1}h_i}{6} \right) (x_i - x) \\
&+ \left(\frac{y_i}{h_i} - \frac{M_i h_i}{6} \right) (x - x_{i-1})
\end{aligned} \tag{6.16}$$

The only unknown parameters in Eq. (6.16) are the second derivatives M_i at $i=1, \dots, N-1$. To determine M_i let us calculate the derivatives $p'_i(x)$:

$$\begin{aligned}
p'_i(x) &= -M_{i-1} \frac{(x_i - x)^2}{2h_i} + M_i \frac{(x - x_{i-1})^2}{2h_i} - \frac{1}{h_i} \left(y_{i-1} - \frac{M_{i-1}h_i^2}{6} \right) + \frac{1}{h_i} \left(y_i - \frac{M_i h_i^2}{6} \right) \\
&= -M_{i-1} \frac{(x_i - x)^2}{2h_i} + M_i \frac{(x - x_{i-1})^2}{2h_i} + \frac{y_i - y_{i-1}}{h_i} - \frac{h_i(M_i - M_{i-1})}{6}
\end{aligned} \tag{6.17}$$

and evaluate them at x_{i-1} and x_i :

$$\begin{aligned}
p'_i(x_i) &= \frac{h_i}{3} M_i + \frac{h_i}{6} M_{i-1} + \frac{y_i - y_{i-1}}{h_i} \\
p'_i(x_{i-1}) &= -\frac{h_i}{3} M_{i-1} + \frac{h_i}{6} M_i + \frac{y_i - y_{i-1}}{h_i}
\end{aligned} \tag{6.18}$$

but because $p'_i(x_i) = p'_{i+1}(x_i)$ the above expressions lead to:

$$\frac{h_i}{3} M_i + \frac{y_i - y_{i-1}}{h_i} + \frac{h_i}{6} M_{i-1} = -\frac{h_{i+1}}{3} M_i + \frac{y_{i+1} - y_i}{h_{i+1}} - \frac{h_{i+1}}{6} M_{i+1} \tag{6.19}$$

Multiplication by $6/(h_{i+1} + h_i) = 6/(x_{i+1} - x_{i-1})$ and rearranging produces:

$$\begin{aligned}
\frac{h_i}{h_{i+1} + h_i} M_{i-1} + 2M_i + \frac{h_{i+1}}{h_{i+1} + h_i} M_{i+1} &= \frac{6}{h_{i+1} + h_i} \left(\frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i} \right) \\
&= dd_i
\end{aligned} \tag{6.20}$$

where dd_i is known

$$dd_i = \frac{6}{h_{i+1} + h_i} \left(\frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i} \right) = \frac{6}{x_{i+1} - x_{i-1}} \left(\frac{y_{i+1} - y_i}{x_{i+1} - x_i} - \frac{y_i - y_{i-1}}{x_i - x_{i-1}} \right) \tag{6.21}$$

Eq. (6.20) forms a tridiagonal matrix which permits to determine the second derivatives M_i . There are $N-1$ equations with $N+1$ unknowns. To determine all the unknowns, we have to get additional equations for x_0 and x_N . There are few conditions which assume additional information about boundary conditions:

- setting second derivatives to zero $M_0 = M_N = 0$ (so called natural boundary condition)
- using known values of $p'_1(x_0)$ and $p'_{N-1}(x_N)$ or setting them equal to zero (so called clamped boundary condition).

c) assuming $M_1 - M_0 = 0$ and $M_{N-1} - M_N = 0$

For the natural boundary condition (a) $M_0 = M_N = 0$ the following tridiagonal matrix is obtained:

$$\begin{bmatrix} 2 & \lambda_1 & & & \\ \mu_2 & 2 & \lambda_2 & & \\ & & \dots & & \\ & & & \dots & \\ & & & & \mu_{N-2} & 2 & \lambda_{N-2} \\ & & & & \mu_{N-1} & 2 \end{bmatrix} \begin{bmatrix} M_1 \\ M_2 \\ \dots \\ \dots \\ M_{N-2} \\ M_{N-1} \end{bmatrix} = \begin{bmatrix} dd_1 \\ dd_2 \\ \dots \\ \dots \\ dd_{N-2} \\ dd_{N-1} \end{bmatrix} \quad (6.22)$$

where

$$\mu_i = \frac{h_i}{h_{i+1} + h_i} \quad \lambda_i = \frac{h_{i+1}}{h_{i+1} + h_i} = 1 - \mu_i \quad (6.23)$$

All M_i values are obtained by solving Eq. (6.22) and the splines are calculated using Eq. (6.16). Let us look at an example of the application of cubic splines in interpolation.

Example 6.3.

Use cubic splines to interpolate data in Table 6.2.

The program which uses cubic splines is *spl.exe* with ITYPE=1. It reads the data file, *runge*, containing data from Table 6.2, then asks for the number of points and the name of the output data file containing interpolation and its first derivative, The results for interpolating 200 points are included in *runge_int*. The results are also shown in Fig. 6.3. They are included in Excel file *Examples6.xlsx* sheet *Ex. 6.3* and folder *E6-3*.

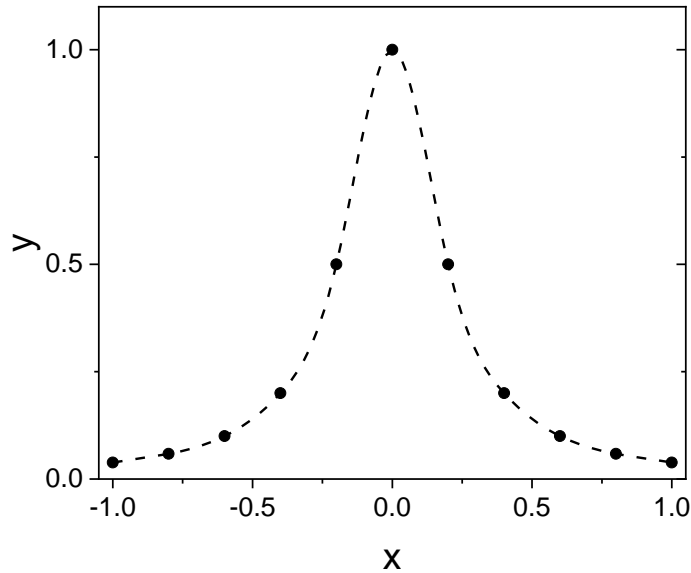


Fig. 6.3. Plot of the data in Table 6.2 and their interpolation by cubic splines.

The first derivative of the cubic splines is shown in Fig. 6.4.

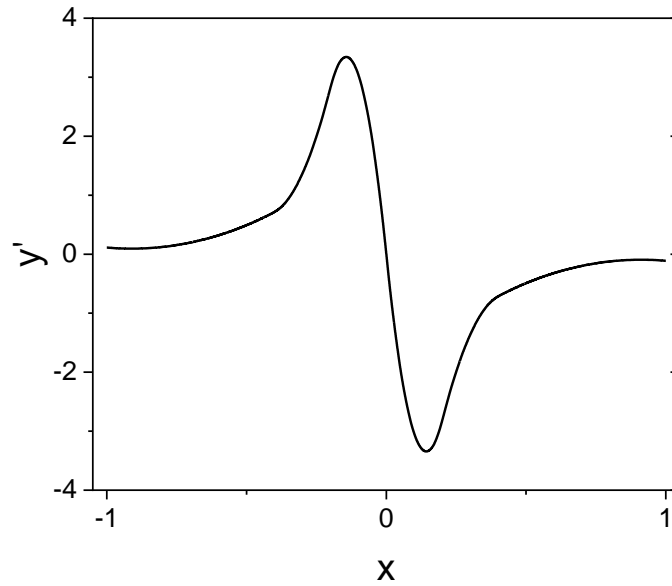


Fig. 6.4. First derivative of the cubic spline in Fig. 6.3.

Finally, the difference between the cubic spline interpolation and the exact function calculated using Eq. (6.5) is displayed in Fig. 6.5.

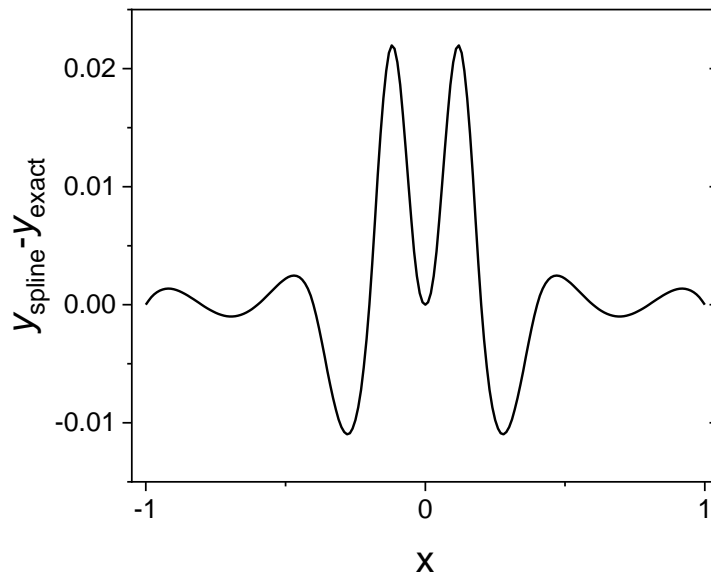


Fig. 6.5. Plot of the differences between cubic spline interpolation and the function calculated using Eq. (6.5).

Although the spline function passes through all the points in Table 6.2, the differences appear in between these points. This is not surprising as there are few points in the fast changing zone for x between -0.5 and 0.5 and the program tries to fit cubic function to the Lorentzian type relation, Eq. (6.5). Of course, with the increasing number of points the interpolation becomes better.

In some cases, the cubic spline interpolation may lead to overshoot or wiggles of the approximating functions. Two methods were proposed to avoid these problems, one by Akima⁶⁴ and another to preserve convex and concave regions.^{65,66} An example below illustrates applications of these methods.

Example 6.4.

Use spline, Akima, and concave interpolations of the data in Table 6.3.

Table 6.3. Data for interpolation in Example 6.4.

x	y
0.00	0.00
0.10	0.90
0.20	0.95
0.30	0.90
0.40	0.10
0.50	0.05
0.60	0.05
0.80	0.20
1.00	1.00

These data are in data file *concave*. The interpolations were carried out using program *spl.exe* with ITYPE=1 for cubic spline, ITYPE=2 for Akima method, and ITYPE=3 for concave producing 100 interpolation points. These results containing interpolated functions and its first derivative are in data files *concave_1*, *concave_2*, and *concave_3*, respectively. The corresponding plots are shown in Fig. 6.6. The results are also included in Excel file *Examples6.xlsx* sheet *Ex. 6.4*.

In this case spline interpolation produces wiggles visible for $x = 0.26, 0.44$, and 0.53 . Akima and concave algorithms remove these problems but in different ways. One should check use of these programs and decide which method is the best.

Finally, besides cubic splines one can use interpolating B-splines (explained in Section 7.6) of different order. Such interpolations are shown in Example 6.5.

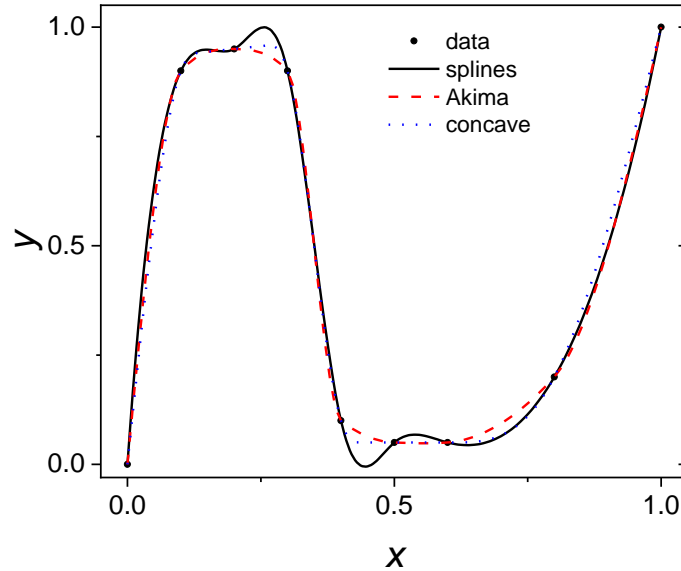


Fig. 6.6. Interpolation of data (points) in Table 6.3 using cubic splines, Akima, and concave algorithms.

Example 6.5.

Use B-splines of the order 2 to 5 to interpolate data in file *runge* (11 points), Table 6.2 and Fig. 6.3, and determine error of such a procedure. Program *bsint.exe* was used to generate files containing 200 points, r_2 , r_3 , r_4 , and r_5 (for orders 2 to 5). The errors of such interpolation i.e. difference between interpolated and calculated, Eq. (6.5), values $y_{interpolated} - y_{calculated}$ are displayed in Fig. 6.7. Of course, interpolation of the exact values with piecewise polynomials must introduce some errors. In this case one can notice that the smallest errors are observed for the second order B-spline and this error increases with the degree of the B-spline. It can also be noticed that using interpolating B-spline of the third order is equivalent to the interpolation using cubic splines. The results are included in Excel file *Examples6.xlsx* sheet *Ex. 6.5*.

However, such propagation of errors is not a rule. Let us look into Example 6.6 where the same Eq. (6.5) was used to generate 21 data points (instead of 11).

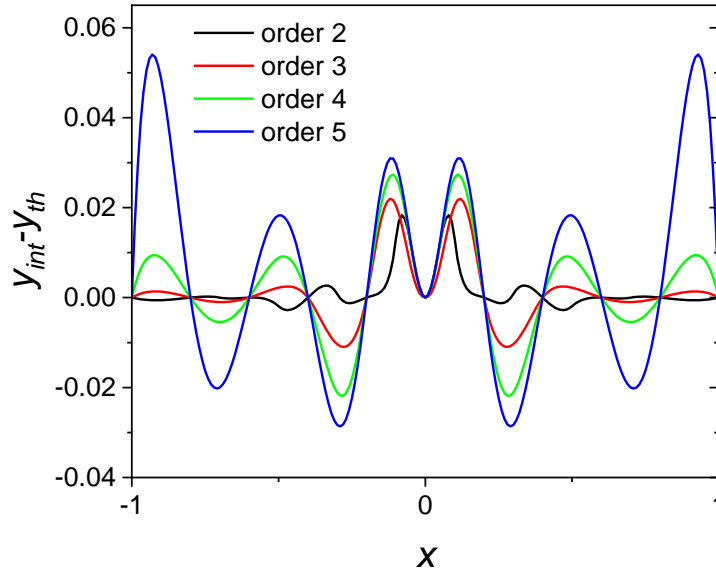


Fig. 6.7. Comparison of the values interpolated using interpolating B-splines of the second to fifth order with the values calculated using Eq. (6.5) (used to generate 11 points data file in Table 6.2).

Example 6.6.

Use interpolating B-splines of the order 2 to 5 to interpolate data in file *rungea* (21 points), Table 6.4 generated using Eq. (6.5) and displayed in Fig. 6.8, and determine error of such a procedure. Program *bsint.exe* was used to generate files containing 200 points, *ra_2*, *ra_3*, *ra_4*, and *ra_5* (for orders 2 to 5). The errors of such interpolation i.e. difference between interpolated and calculated, Eq. (6.5), values $y_{interpolated} - y_{calculated}$ are displayed in Fig. 6.9.

It can be noticed that increasing the interpolating B-spline order from 2 to 5 leads to decrease of the interpolation errors that is the best interpolation is obtained for the order 5. This result is different from that obtained in Example 6.5. However, in Example 6.6 there are more points used as a base for interpolation. The results are in Excel file *Examples6.xlsx* sheet *Ex. 6.6*. This means that the interpolation must be used with prudence. Interpolation with the splines cannot exactly reproduce the unknown data behavior.

Table 6.4. Example of 21 data points obtained using Eq. (6.5).

-1.0	0.038462
-0.9	0.047059
-0.8	0.058824
-0.7	0.075472
-0.6	0.100000
-0.5	0.137931
-0.4	0.200000
-0.3	0.307692
-0.2	0.500000
-0.1	0.800000
0.0	1.000000
0.1	0.800000
0.2	0.500000
0.3	0.307692
0.4	0.200000
0.5	0.137931
0.6	0.100000
0.7	0.075472
0.8	0.058824
0.9	0.047059
1.0	0.038462

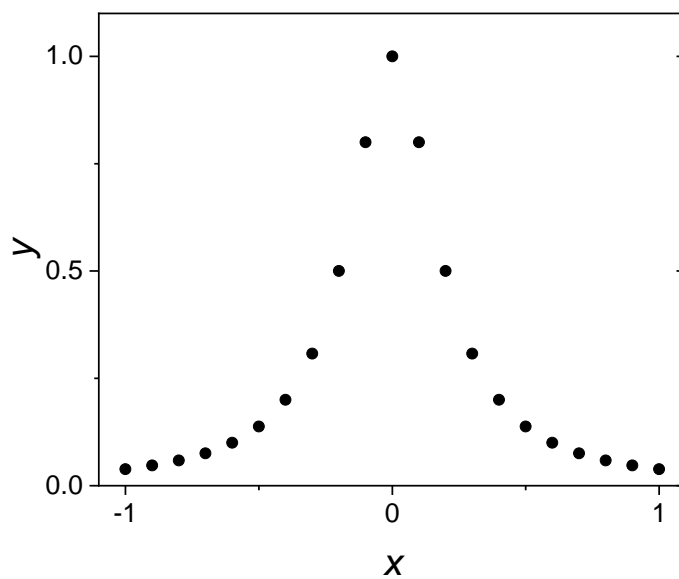


Fig. 6.8. Plot of the data in Table 6.4.

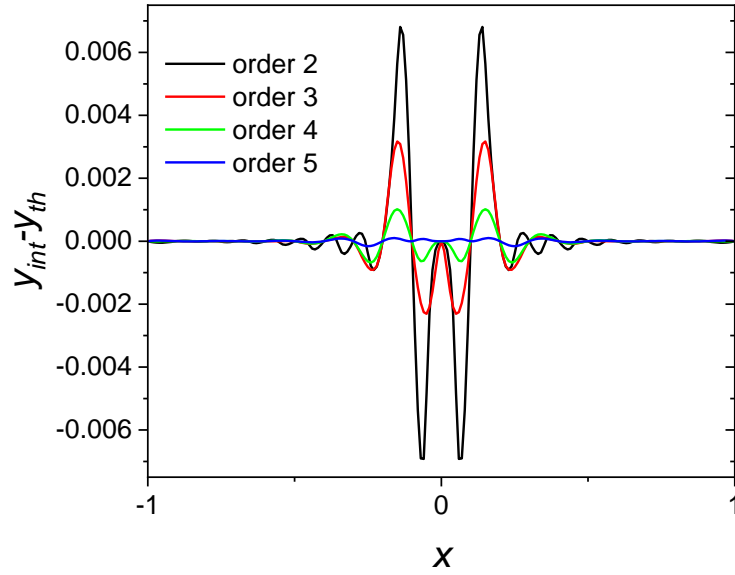


Fig. 6.9. Comparison of the values interpolated using interpolating B-splines of the second to fifth order with the values calculated using Eq. (6.5) (used to generate 21 points data file in Table 6.4, Fig. 6.8).

7 Smoothing

Data obtained experimentally are acquired with some experimental noise. In such a case we do not want to interpolate the all the noisy data points but to eliminate noise. Let us look into an example in Fig. 7.1.

It is not very probable that the approximating line should pass through all the experimental points. In such a case some smoothing must be introduced. There are many different methods which can be used, and one should choose the most appropriate. Of course, the best way is to increase number of experimental points but this is not always possible.

It should be stressed that smoothing is a semi-quantitative method and the amount of smoothing depends on the operator. In the case when we “know” what the data behavior should be, e.g. linear behavior, Gaussian or Lorentzian peaks on a linear or parabolic base line, exponential behavior, etc., we should use fitting the experimental data to the appropriate mathematical model (see e.g. Example 3.12.-Example 3.13.). But in general, we do not know what the data behavior should be and we use smoothing as a graphical technique to guide us through the noisy data. Smoothing is also necessary when determining derivative of the noisy data which increases noise, while integration averages the noise. It is obvious that too much smoothing distorts the tendency in data while too little smoothing leaves the noise. The choice is always a little ambiguous.

Few popular methods will be presented below with the corresponding programs, some are incorporated in the popular plotting software (Origin, SigmaPlot, etc.).

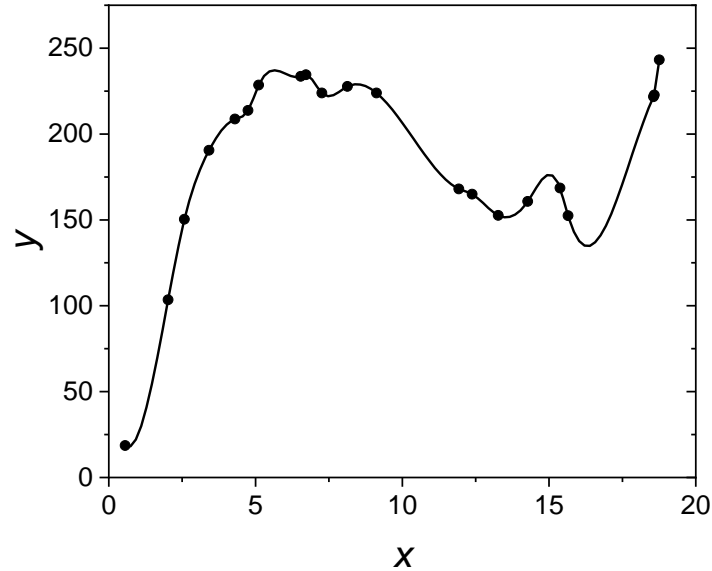


Fig. 7.1. An example of the interpolation of the noisy data by cubic splines; the approximating line passes through all the experimental points (which contain some noise).

7.1 Simple data reduction

To acquire data an analog to digital converter at high speed (at equal intervals) might be used and produces large amount of data. To reduce their amount the data acquisition program usually averages periodically blocks of data. For example, if the experiment lasts 1 min and the data acquisition rate chosen is 1000 Hz (one point every 1 ms) an average might be calculated every 100 points i.e. every 100 ms. This reduces 60,000 data points to 600 with important noise reduction. Such method works well when the signal is changing relatively slowly, the experiment is long, and a lot of data points are acquired, otherwise a deformation might be observed. This procedure also allows reduction of the number of points acquired and consequently stored. An example of such averaging illustrated in Example 7.1.

Example 7.1.

Carry out averaging of the noisy data file *data6*, acquired by the data acquisition system (A/D analog to digital converter) and containing 60,000 points, every 100 points. The average should appear at the end of each averaging window.

The plot of the data is shown in Fig. 7.2a. The averages of 100 points were calculated in Excel and the number of points was reduced to 600 improving the signal to noise ratio. Plot of the averages is displayed in Fig. 7.2b, in data file *data6_averaged* and in Excel file *Examples7.xlsx* sheet *Ex. 7.1*. The data files are in the folder *E7-1*. An important noise reduction was obtained.

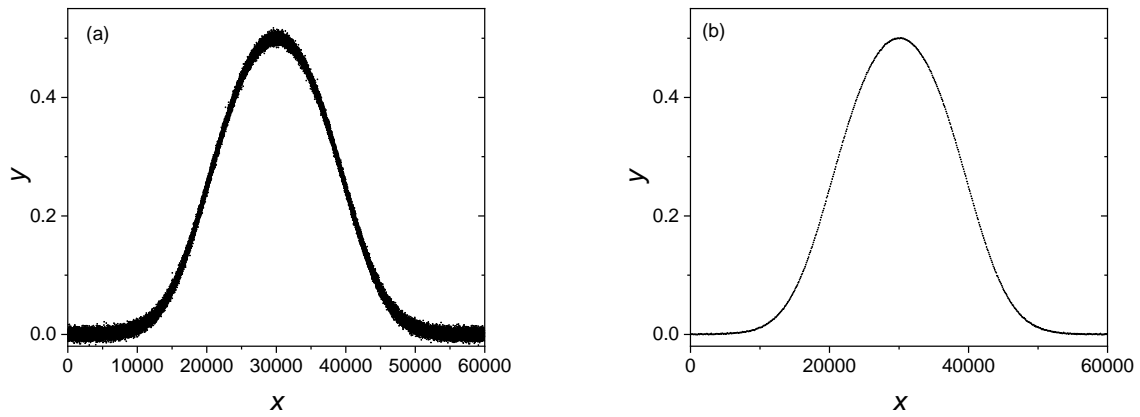


Fig. 7.2. (a) 60,000 raw data points acquired by the A/D converter; (b) data averaged every 100 points reducing number of points to 600 and reducing noise.

7.2 Simple digital filters

Simple digital filters are based on a weighted average of the data. Below, data smoothing techniques of already acquired data using simple digital filters are presented.

The most popular digital filters are displayed in Fig. 7.3.

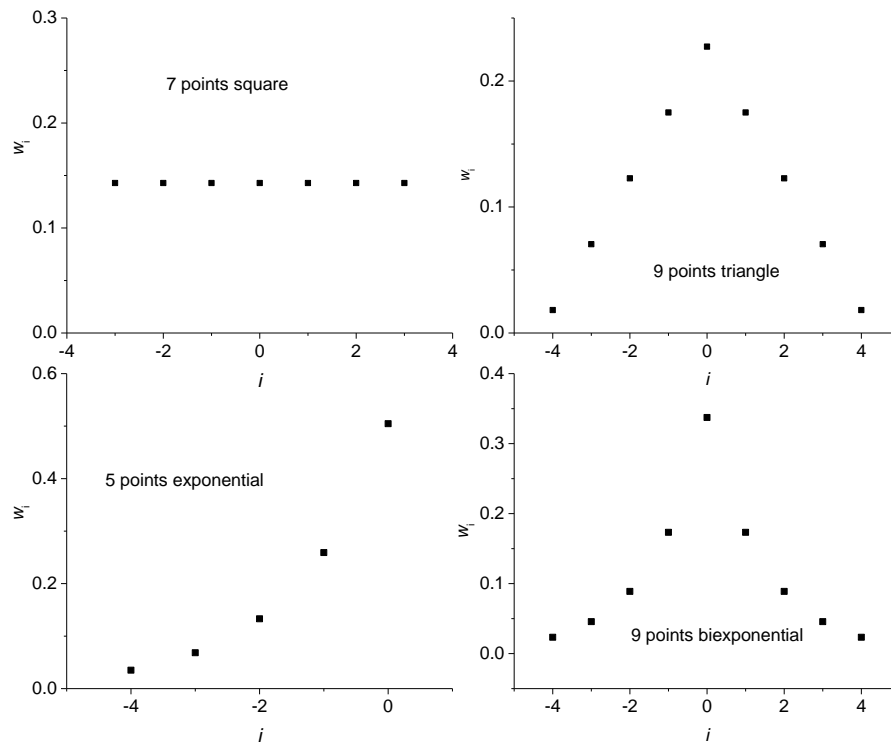


Fig. 7.3. Examples of simple digital filters: simple square 5 points filter (central average of 5 points), 9 points triangle filter, 5 points exponential, 9 points bi-exponential.

7.2.1 Moving central average square filter

Square filter is a simple central average. It has an advantage that it does not cause any horizontal shift although it can attenuate quickly changing signal. This filter calculates an average value in a central point is described by equation:

$$\hat{y}_i = \frac{\sum_{j=i-m}^{i+m} y_j}{2m+1} \quad (7.1)$$

where number of averaged points $np = 2m+1$ is the length of the filter and the data are acquired with the constant rate. In this case each average is calculated after acquiring np points.

For example, for averaging 5 points Eq. (7.1) is:

$$\hat{y}_0 = \sum_{j=-2}^{j=2} \frac{y_j}{5} = \frac{1}{5}(y_{-2} + y_{-1} + y_0 + y_1 + y_2) \quad (7.2)$$

This filter is analog to that in Section 7.1, the difference is that the former method provides the average at last point np of the interval while Eq. (7.1) determines the central average at the central point $m+1$ of the interval np . An example of Excel calculations is shown in Example 7.2.

Example 7.2.

Calculate the central averages 21 points of data in file *data3* containing 401 points. These averages were calculated in the file Excel file *Examples7.xlsx* sheet *Ex. 7.2* which explains how the calculations were carried out. The smoothed data are in file *data3_cent*. Data files are also in the folder *E7-2*, see Fig. 7.4.

Such a procedure works well when the number of points is large but it attenuates the peak when too wide filter or too few data points are used.

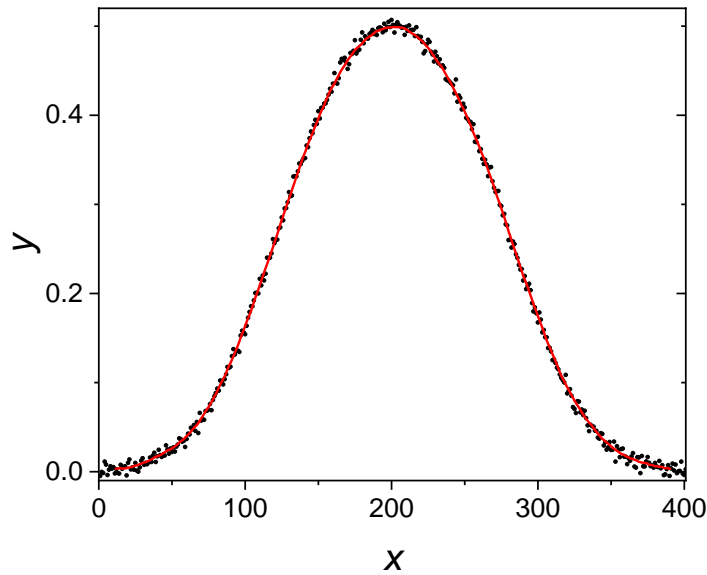


Fig. 7.4. Plot of the noisy 401 data points (black symbols) in file *data3* and 21 point central averages (red line).

In general, weighted average filters are used. They are applied after data acquisition process and are described by the general equation:

$$\hat{y}_i = \sum_{j=-m}^{j=m} \frac{w_j y_j}{w} \quad i = 1, 2, \dots, N$$

$$w = \sum_{i=1}^{2m+1} w_j$$
(7.3)

where w_j is the smoothing weighting parameter and w is the normalization factor.

7.2.2 Exponential filter

The exponential filter corresponds to the analog data averaging using R-C filter, however, it can distort the fast changing signal. An example is shown in Example 7.3.

Example 7.3.

Use 5 points exponential filter to smooth data in file *data2* containing 51 noisy points. They are displayed in Fig. 7.5 and are in folder *E7-3*.

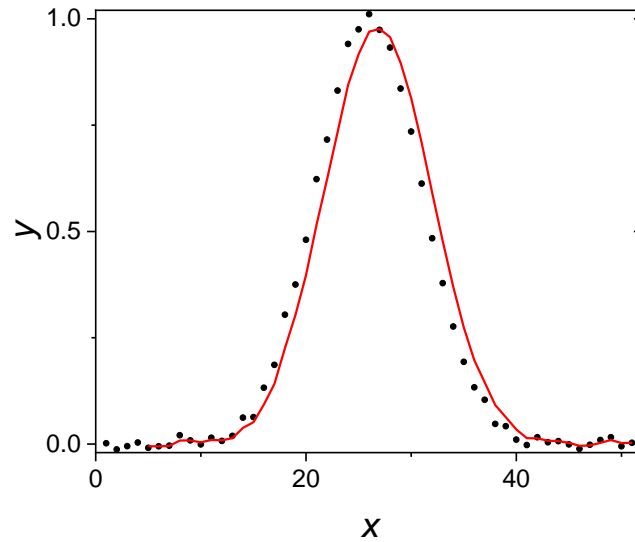


Fig. 7.5. Plot of the data in *data2* (symbols) and the exponential smoothing using 5 points exponential filter (red line).

The filter used contained weights $w_i = \exp(-i/1.5)$ and the sum of weights was $w = 1.981933$. It is displayed in Fig. 7.6.

The smoothing was carried out in Excel file *Examples7.xlsx* sheet *Ex. 7.3* and the smoothed data are in *data2_exp*. It is clear that applying this filter to a very small data file (51 points) causes attenuation of the peak and its shift to the higher values. However, when number of points is very large it works very well. It can be used online because it uses only information on the previous points.

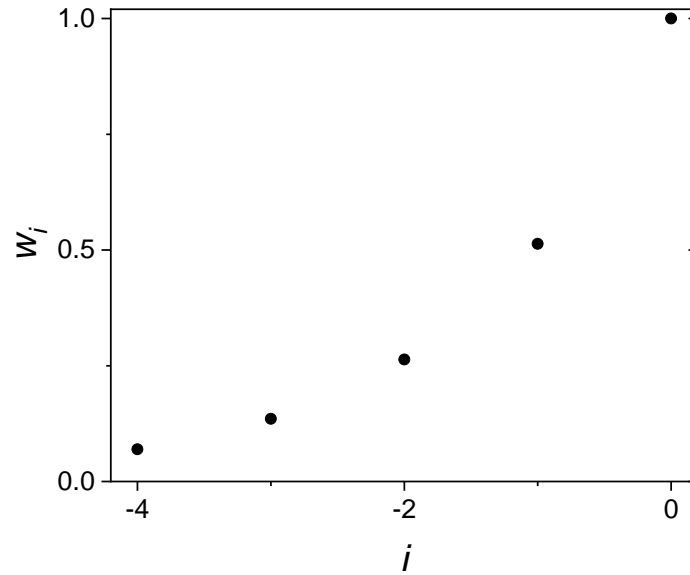


Fig. 7.6. Example of the 5 points exponential filter used for data smoothing.

7.2.3 Symmetrical triangular filter

An example of the symmetrical triangular or bi-triangular filter is shown in Fig. 7.3. Let us look at its application in Example 7.4, the files are in folder *E7-4*.

Example 7.4.

Apply 11 points symmetrical triangular filter to data file *data2*. The filter weights were calculated as $w_i = 1 - |i/6|$ and they are shown in Fig. 7.7.

The results of the application of this bi-triangular filter to data smoothing is shown in Fig. 7.8 and in file *data2_bitri*. The details of calculations are in the Excel file *Examples7sheet Ex. 7.4*.

It is evident that the calculated curve is smooth but the peak is attenuated although its position is unchanged. Of course, this is related to the fact that there are too few experimental points, 51, for the filter size, 11. But using less points in the filter would leave more unfiltered noise.

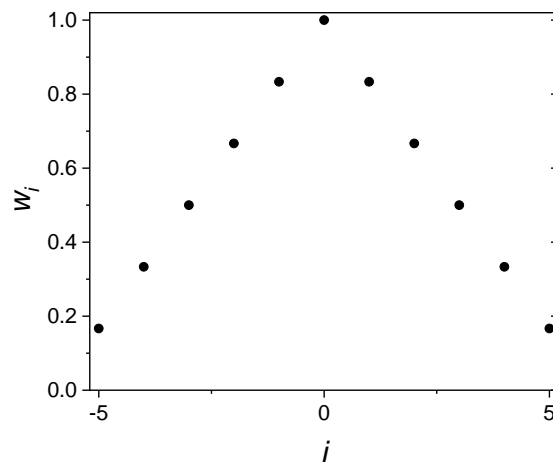


Fig. 7.7. Symmetrical triangular 11 points filter weights calculated as: $w_i = 1 - |i/6|$.

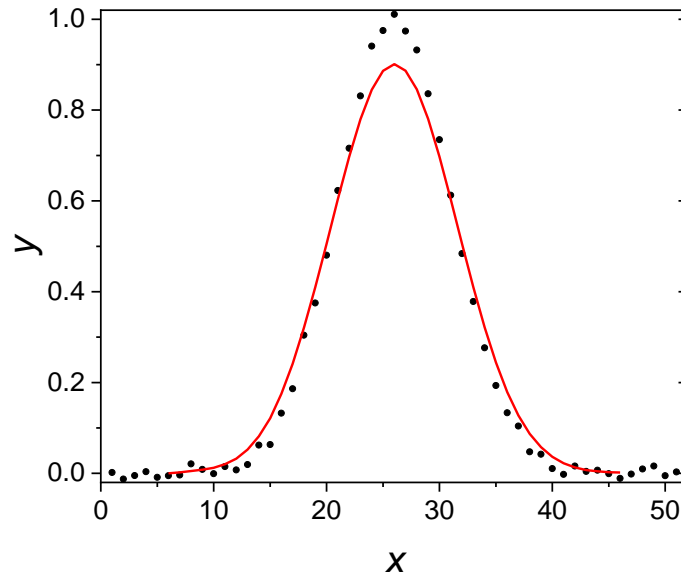


Fig. 7.8. Plot of the noisy data in *data2* (black symbols) and the result of smoothing using 11 points bi-triangular filter with weights $w_i = 1 - |i/6|$ (red line).

7.2.4 Bi-exponential filter

Bi-exponential digital filter is more often used as it puts greater emphasis on the central point. An example of application of such a filter is shown below.

Example 7.5.

Let us assume filtering using 15 points bi-exponential filter, centered at $m = 8$ ($i = 0$), shown in Fig. 7.9. It was defined as in Eq. (7.3) with:

$$w_i = \exp(-|i/5|)$$

$$w = \sum_{i=-7}^{i=7} w_i = 7.8057 \quad (7.4)$$

This filter is displayed in Fig. 7.9. Raw data in *data3* (see folder *E7-5*) containing 401 points are displayed in Fig. 7.10a and the calculations are shown in *Examples7.xlsx*, sheet *Ex. 7.5*. The results of smoothing are shown in Fig. 7.10b where the majority of noise was removed.

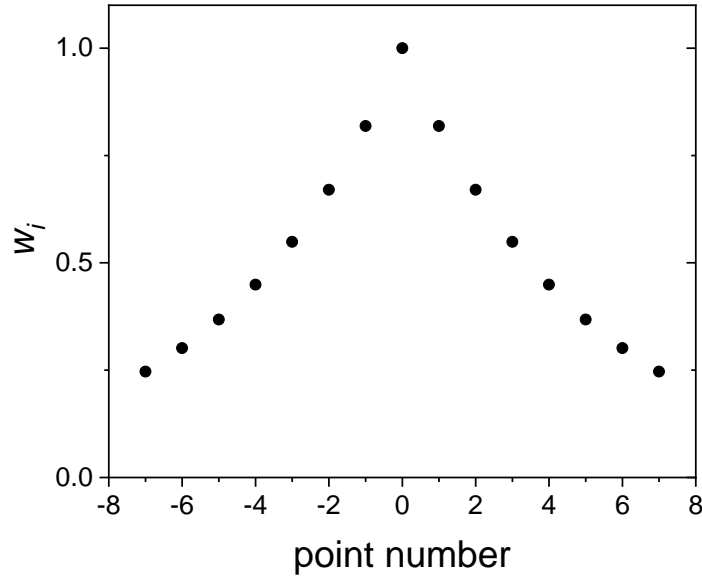


Fig. 7.9. Example of the bi-exponential digital filter, for 15 points, Eq. (7.4).

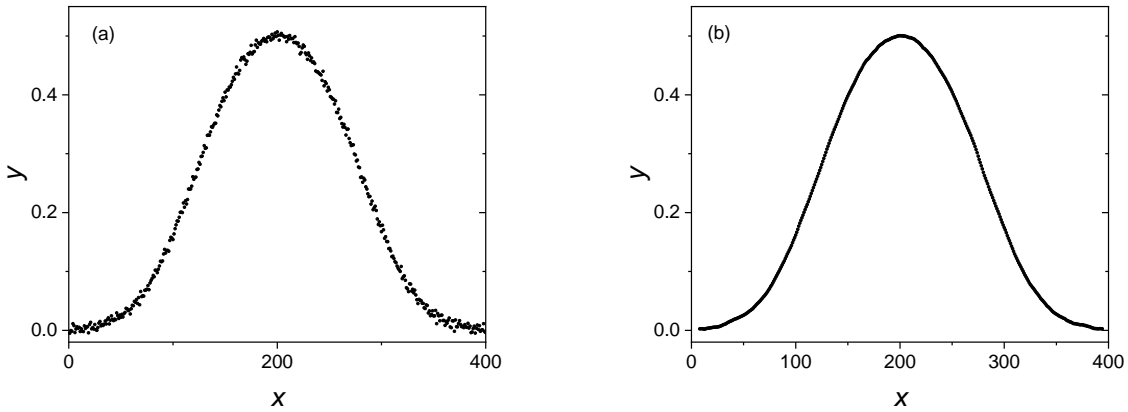


Fig. 7.10. (a) Raw noisy data and (b) smoothed by application of the bi-exponential filter using 15 points to data file *data3*, Eq. (7.4).

Manual application of this filter was used in Excel file but computer programs are usually used to carry out such calculations. Of course, one could try different number of points and weight parameters w_i to get the best results.

The above presented filters are very simple and work when many data points are acquired and large averaging can be used.

7.2.5 Adjacent-averaging filter

Finally, adjacent-averaging filter is a moving weighted average filter where the weights of point j corresponding to uniformly distributed np data points: $j = 1, 2 \dots np$ are defined as:

$$w_j = 1 - \left[\frac{j-i}{(np+1)/2} \right]^2 \quad (7.5)$$

where i is the central point where smoothed value is calculated and np is the total number of points in the filter. An example of a plot of these weights for $np = 11$ (central value $i = 6$) is displayed in Fig. 7.11. The smoothed values are calculated using Eq. (7.3).

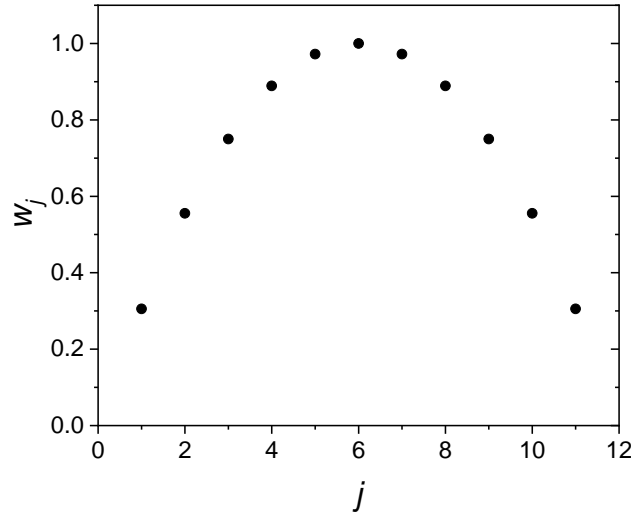


Fig. 7.11. Plot of the weights in adjacent-averaging for 11 points ($i = 6$), Eq. (7.5).

Application of this filter to data smoothing is illustrated in Example 7.6.

Example 7.6.

Use adjacent-smoothing with 11 points window to data *data3* containing 401 points, Fig. 7.12; see also folder *E7-6*.

These calculations were carried out in Excel and the results are in *Examples7.xlsx*, sheet *Ex. 7.6*. The weights were calculated using Eq. (7.5) and the values of the smoothed function using Eq. (7.3). The results are displayed in Fig. 7.12, red line; they are also in file *data3_adj*. Good smoothing was obtain using 11 points window.

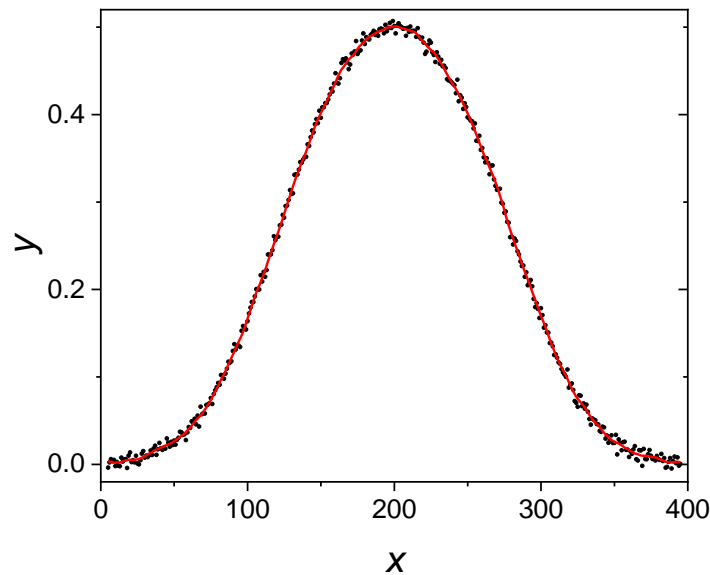


Fig. 7.12. Plot of the *data3* data, points, and 11 points adjacent-averaging filter, red line.

7.3 Savitzky-Golay filter

Savitzky and Golay⁶⁷ have proposed moving second order polynomial to smooth the experimental data. It consists of applying the least-squares method to smooth a subset of data, then it is moved to next points. A parabola passes exactly through 3 points therefore more points must be used to smooth the data. In practice odd number of points: 5, 7, 9... are used. Below, an example of approximation of 5 points by a second order polynomial will be shown but in a similar way other formulas might be obtained. It should be stressed that the points are acquired in equal distances therefore x are not important and only y_i values will be considered.

The second order polynomial is $\hat{y} = b_2x^2 + b_1x + b_0$ and the parameters b_i must found. The data points used locally are: x_{-2}, y_{-2} ; x_{-1}, y_{-1} ; x_0, y_0 ; x_1, y_1 ; x_2, y_2 . Using general method of the least squares we have to create matrix \mathbf{X} for 5 points from -2 to 2, Eqns. (3.106)-(3.107). Matrix \mathbf{X} contains the derivatives $\partial y_i / \partial b_j$. The central point for which the approximation is calculated corresponds to $x = 0$:

$$\mathbf{X} = \begin{bmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix} \quad (7.6)$$

The matrix of the values of y_i is:

$$\mathbf{Y} = \begin{bmatrix} y_{-2} \\ y_{-1} \\ y_0 \\ y_1 \\ y_2 \end{bmatrix} \quad (7.7)$$

The solution of the problem is $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ where

$$\mathbf{b} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} -\frac{3}{35}y_{-2} + \frac{12}{35}y_{-1} + \frac{17}{35}y_0 + \frac{12}{35}y_1 - \frac{3}{35}y_2 \\ -\frac{2}{10}y_{-2} - \frac{1}{10}y_{-1} + \frac{1}{10}y_1 + \frac{2}{10}y_2 \\ \frac{2}{14}y_{-2} - \frac{1}{14}y_{-1} - \frac{2}{14}y_0 - \frac{1}{14}y_1 + \frac{2}{14}y_2 \end{bmatrix} \quad (7.9)$$

The smoothed values of the function and its derivatives calculated at $x = 0$ from the approximating polynomial are:

$$\begin{aligned}
\hat{y}(x=0) &= b_0 = \frac{1}{35}(-3y_{-2} + 12y_{-1} + 17y_0 + 12y_1 - 3y_2) \\
\hat{y}'(x=0) &= b_1 = \frac{1}{10}(-2y_{-2} - y_{-1} + y_1 + 2y_2) \\
y''(x=0) &= 2b_2 = \frac{1}{7}(2y_{-2} - y_{-1} - 2y_0 - y_1 + 2y_2)
\end{aligned} \tag{7.10}$$

This smoothing, in practice, corresponds to multiplication of y_i by some coefficients (weights). Savitzky and Golay have published such coefficients for the second and fourth order smoothing for 5 to 25 points. Coefficients for the parabolic smoothing are shown for the function, its first and second derivative are shown in Table 7.1-7.3.

Table 7.1. Savitzky-Golay coefficients used to calculate smoothed values of \hat{y}_i using 5 to 25 points.

POINTS	25	23	21	19	17	15	13	11	9	7	5
-12	-253										
-11	-138	-42									
-10	-33	-21	-171								
-9	62	-2	-76	-136							
-8	147	15	9	-51	-21						
-7	222	30	84	24	-6	-78					
-6	287	43	149	89	7	-13	-11				
-5	342	54	204	144	18	42	0	-36			
-4	387	63	249	189	27	87	9	9	-21		
-3	422	70	284	224	34	122	16	44	14	-2	
-2	447	75	309	249	39	147	21	69	39	3	-3
-1	462	78	324	264	42	162	24	84	54	6	12
0	467	79	329	269	43	167	25	89	59	7	17
1	462	78	324	264	42	162	24	84	54	6	12
2	447	75	309	249	39	147	21	69	39	3	-3
3	422	70	284	224	34	122	16	44	14	-2	
4	387	63	249	189	27	87	9	9	-21		
5	342	54	204	144	18	42	0	-36			
6	287	43	149	89	7	-13	-11				
7	222	30	84	24	-6	-78					
8	147	15	9	-51	-21						
9	62	-2	-76	-136							
10	-33	-21	-171								
11	-138	-42									
12	-253										
NORM	5175	805	3059	2261	323	1105	143	429	231	21	35

Table 7.2. Savitzky-Golay coefficients to calculate the first derivative \hat{y}'_j .

POINTS	25	23	21	19	17	15	13	11	9	7	5
-12	-12										
-11	-11	-11									
-10	-10	-10	-10								
-9	-9	-9	-9	-9							
-8	-8	-8	-8	-8	-8						
-7	-7	-7	-7	-7	-7	-7					
-6	-6	-6	-6	-6	-6	-6	-6				
-5	-5	-5	-5	-5	-5	-5	-5	-5			
-4	-4	-4	-4	-4	-4	-4	-4	-4	-4		
-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	-3	
-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2	-2
-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1
0	0	0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3	3	
4	4	4	4	4	4	4	4	4	4		
5	5	5	5	5	5	5	5	5			
6	6	6	6	6	6	6	6				
7	7	7	7	7	7	7					
8	8	8	8	8	8						
9	9	9	9	9							
10	10	10	10								
11	11	11									
12	12										
NORM	1300	1012	770	570	408	280	182	110	60	28	10

Table 7.3. Savitzky-Golay coefficients to calculate the second derivative \hat{y}_j'' .

POINTS	25	23	21	19	17	15	13	11	9	7	5
-12	92										
-11	69	77									
-10	48	56	190								
-9	29	37	133	51							
-8	12	20	82	34	40						
-7	-3	5	37	19	25	91					
-6	-16	-8	-2	6	12	52	22				
-5	-27	-19	-35	-5	1	19	11	15			
-4	-36	-28	-62	-14	-8	-8	2	6	28		
-3	-43	-35	-83	-21	-15	-29	-5	-1	7	5	
-2	-48	-40	-98	-26	-20	-44	-10	-6	-8	0	2
-1	-51	-43	-107	-29	-23	-53	-13	-9	-17	-3	-1
0	-52	-44	-110	-30	-24	-56	-14	-10	-20	-4	-2
1	-51	-43	-107	-29	-23	-53	-13	-9	-17	-3	-1
2	-48	-40	-98	-26	-20	-44	-10	-6	-8	0	2
3	-43	-35	-83	-21	-15	-29	-5	-1	7	5	
4	-36	-28	-62	-14	-8	-8	2	6	28		
5	-27	-19	-35	-5	1	19	11	15			
6	-16	-8	-2	6	12	52	22				
7	-3	5	37	19	25	91					
8	12	20	82	34	40						
9	29	37	133	51							
10	48	56	190								
11	69	77									
12	92										
NORM	26910	17710	33649	6783	3876	6188	1001	429	462	42	7

In simple cases smoothing can be performed manually in Excel. This is illustrated in Example 7.7. Example 7.7.

Perform Savitzky-Golay smoothing of the data in file *sg1* (folder *E7-7*), containing 101 points, using 15 points formula and the values from Table 7.1.

The values of y_i were multiplied by the values from Table 7.1 for 15 point filter (-78, -13, 42, 87, 122, 147, 162, 167, 162, 147, 122, 87, 42, -13, -78), added, and divided by the norm 1105 to obtain the smoothed value for point number 8. The details of calculations are in Excel file *Examples7.xlsx*, sheet *Ex. 7.7*. The raw (*sg1*) and smoothed data (*sg1_15*) are compared in Fig. 7.13.

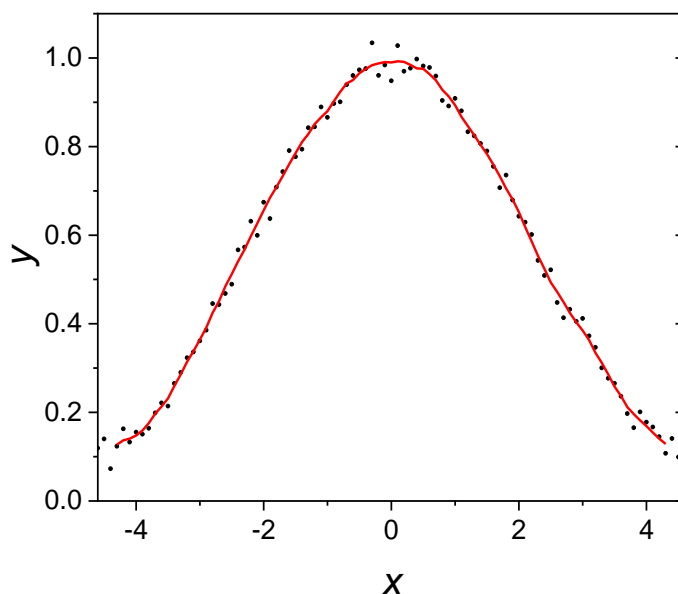


Fig. 7.13. Comparison of the raw and smoothed data (101 points) using Savitzky-Golay 15 point filter.

While the results are smoothed they still contain some noise and may be larger filter should be applied.

Although Savitzky-Golay smoothing can be carried out manually it is easier to use a computer program, especially for larger number of points. For such smoothing program *sg.exe* is provided. An example is shown in Example 7.8.

Example 7.8.

Program *sg.exe* can accommodate up to 40000 points and smooth locally up to 99 points. It should be pointed out that using very large windows may cause distortions and decrease of peaks. Program *sgd.exe* may also be used to calculate the first derivative.

An example of Savitzky-Golay smoothing is presented in application to data in *data3a* containing 401 data points, see Fig. 7.14. To choose the number of points and the degree of the polynomial the first derivative was inspected. It is important to choose the smallest number of points in the filter and the lowest order of polynomial. Using program *sgd.exe* to calculate dy/dx it was found that 49 filter points and the second order of the polynomial is sufficient to obtain relatively smooth derivative (derivative is much more sensitive to the noise than the smoothed function). Then, the smoothed function was calculated using program *sg.exe*. The raw data are in the file *data3a*. The results are displayed in Fig. 7.14 and in Excel file *Examples7.xlsx* sheet *Ex. 7.8*. They are also in files *data3a_sg49* and *data3a_sgd49* in folder *E7-8*.

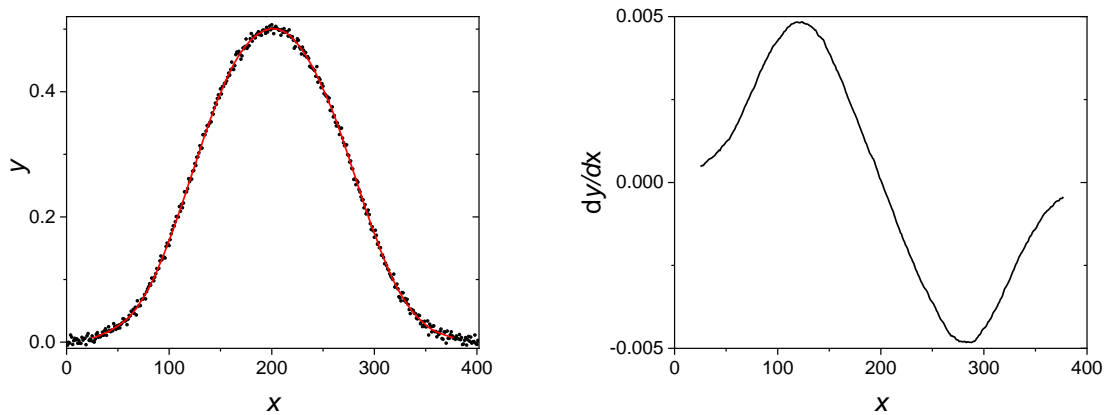


Fig. 7.14. Application of the Savitzky- Golay filter to smooth data in Fig. 7.10a; (a) raw data and smoothed function, (b) first derivative, obtained using 49 points filter and second order polynomial.

It should be noticed that because Savitzky- Golay filter calculated values in the center of the window, m points on two extreme sides of the data are not smoothed. In the example presented $np = 49$ then $m = 24$ points on both sides of the data file are not smoothed.

Limits and applicability of the simple central average and Savitzky-Golay filters are presented in Fig. 7.15. The noisy data (top) contain several peaks of decreasing width. Application of the simple central average for $2 \times 16 + 1 = 33$ points effectively filters noise of the large peak but narrow peaks are deformed and lose their amplitude. On the other hand, application of the Savitzky-Golay filter to 33 points and using fourth order polynomial smoothes the sharp peaks but leaves noise at the large peak and flat area.

The use of a wider window of 65 points and different orders of smoothing polynomials is displayed in Fig. 7.16. Good smoothing is obtained for the flat initial part and a large peak when using second order polynomial. However, sharp peaks practically disappear. Increase of the polynomial order to 4 or 6 leave larger noise at the initial part and the wide peak but better reproduces sharp peaks. These examples show that it is impossible to smooth both the flat part and wide and narrow peaks.

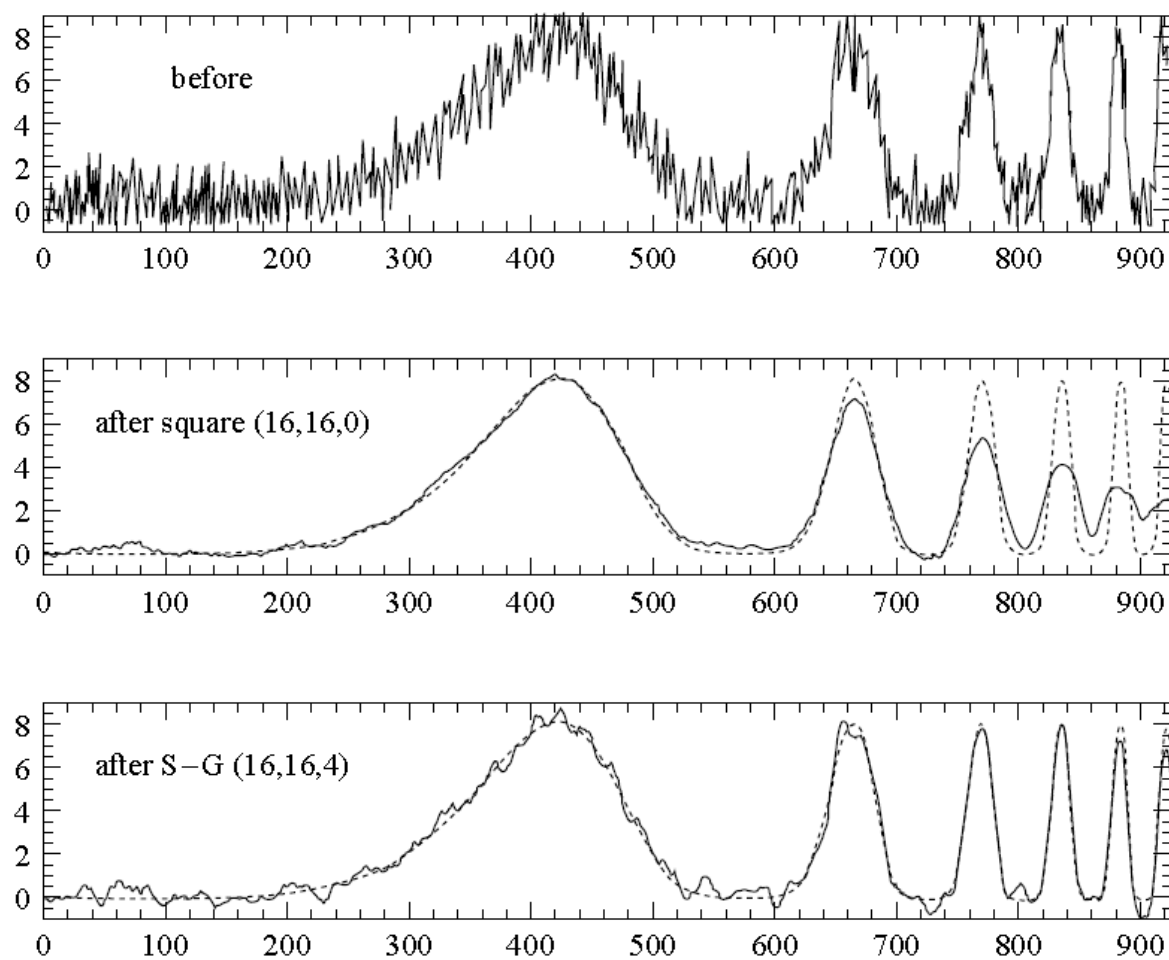


Fig. 7.15. Smoothing of the noisy data (top) using square central average (middle) and Savitzky-Golay filter of the second degree for 33 points.⁶⁸

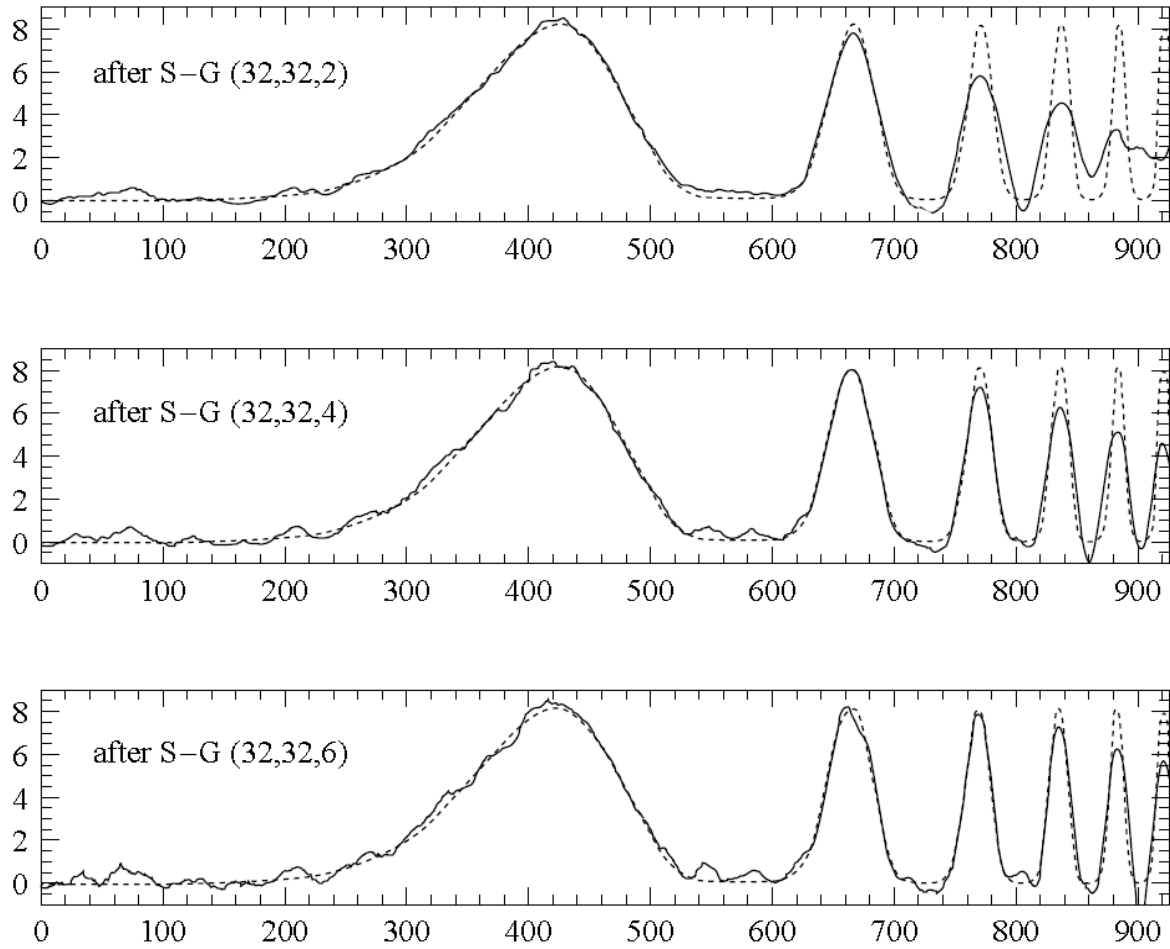


Fig. 7.16. Application of a wider 65 points Savitzky-Golay filter to the data set in Fig. 7.15 (top) and polynomial orders of the second, fourth, and sixth order.⁶⁸

All above smoothing techniques demand that data points are acquired at equal intervals that is at constant sample spacing Δx . Bellow, other methods for varying sample spacing will be presented.

7.4 Polynomial approximation

As it was mentioned in the section 6.1 polynomial of high degree can interpolate exactly the data points but it usually oscillates between. Therefore, one should use the lowest polynomial degree which can smooth well the raw data without oscillations. It must also be added that not all functions can be well approximated by polynomials.

In general, least squares approximation involves matrix inversion and with the increase if the polynomial order such an operation becomes instable (determinant too small). To avoid these problems the modern programs use orthogonal polynomials. Orthogonal polynomials are defined as:

$$\int_a^b p_i(x) p_j(x) dx = 0 \quad \text{for } i \neq j \quad (7.11)$$

and

$$\begin{aligned} p_0 &= 1 \\ p_1(x) &= x - a_1 \\ p_j(x) &= (x - a_j) p_{j-1}(x) - b_j p_{j-2}(x) \end{aligned} \quad (7.12)$$

where parameters a_j and b_j are calculated from the orthogonality condition, Eq. (7.11). The whole data set is approximated by:

$$\hat{y}(x) = \alpha_0 + \alpha_1 p_1(x) + \alpha_2 p_2(x) + \dots + \alpha_n p_n(x) \quad (7.13)$$

where n is the highest order of the polynomial. The parameters α_i are determined by the least-squares method. Usually, x values are scaled between -2 and 2 to avoid subtraction of very large values and problems found in U.S. Census data, where two completely different sets of coefficients were obtained using single and double precision data, see Section 3.19.³⁰ The advantage of orthogonal polynomials is that addition of the next order of polynomial does not change the earlier polynomials and problems of matrix inversion are avoided. The basis of the classical orthogonal polynomials are Hermite, Laguerre, Jacobi, Chebyshev, or Legendre polynomials.

For example, a series of Hermite orthogonal polynomials are:

$$\begin{aligned} p_0(x) &= 1 \\ p_1(x) &= 2x \\ p_2(x) &= 4x^2 - 2 \\ p_3(x) &= 8x^3 - 12x \\ p_4(x) &= 16x^4 - 48x^2 + 12 \\ &\dots \end{aligned} \quad (7.14)$$

Let us look at the application of the approximation by orthogonal polynomials to smoothing of the data file *d3*.

Example 7.9.

Use orthogonal polynomials of different orders to approximate data in file *d3*. Generate 200 points of approximating function and compare the results.

The lowest polynomial order which can be used here is the third. Increasing order from 3 to 10 changes a little the approximation, see Fig. 7.17. In *polfit.exe*, first option for equal statistical weights should be chosen because *d3* contains only two columns x and y (unit weights, $w_i = 1$, are assumed). Two files were produced for the 3rd and 10th orders, first containing approximations at the original x_i points and the approximation parameters recalculated into classical polynomials (*d3_3*, *d3_10*) and the second containing function generated at 200 equally spaced x values (*d3_3_200*, *d3_10_200* for graphing). The results are in *Examples7.xlsx*, sheet *Ex. 7.9* and in folder *E7-9*.

However, when the polynomial order is too high, although the polynomial passes closer to the experimental noisy points, it starts to oscillate for between the points. This is illustrated in **Fig. 7.18** where 13th polynomial order was used. The results are in the files *d3_13* and *d3_13_200*. It

is evident that this order is too large and the polynomial starts produce large deviations between the points.

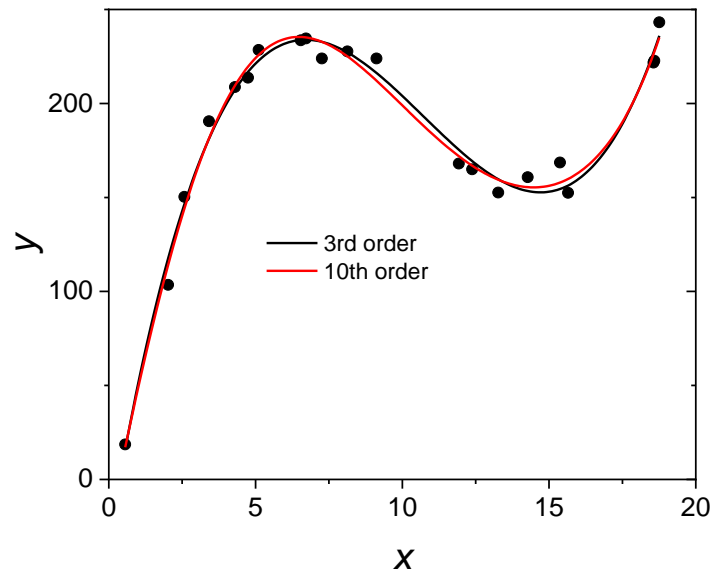


Fig. 7.17. Results of the approximation of the data file *d3* using orthogonal polynomials of the order 3 and 10.

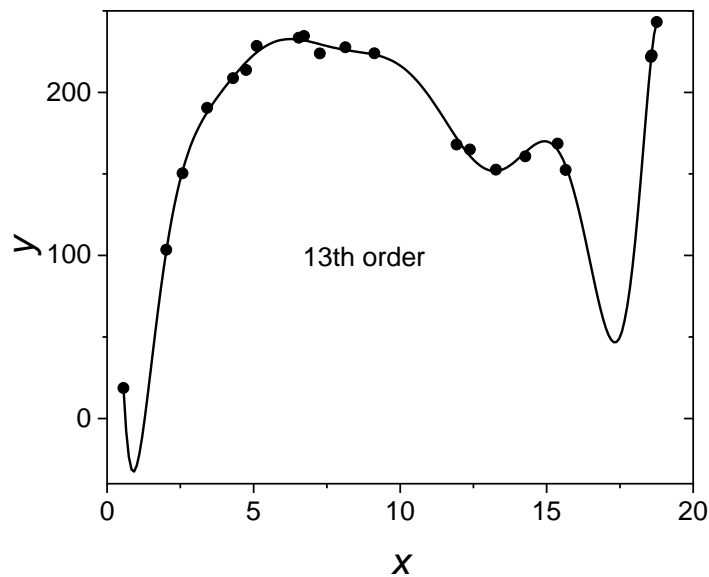


Fig. 7.18.
Results of the

approximation of the data file *d3* using orthogonal polynomials of 13th order.

As it was mentioned earlier in the Section 6.1 on interpolation, polynomials cannot reproduce some shapes, e.g. Runge's Eq. (6.5). The results for approximation of such a function are presented in Example 7.10.

Example 7.10.

Use orthogonal polynomials to approximate data in file *rungea*, **Table 6.4**. The approximations were performed using *polfit.exe* for two polynomial degrees 8 and 13 and 200 points of smoothed function were generated. The results are in the files *r_8*, *r_8_200* and *r_13*, *r_13_200*. The results are also in Excel file *Examples7.xlsx*, sheet *Ex. 7.10*. The plots are displayed in Fig. 7.19.

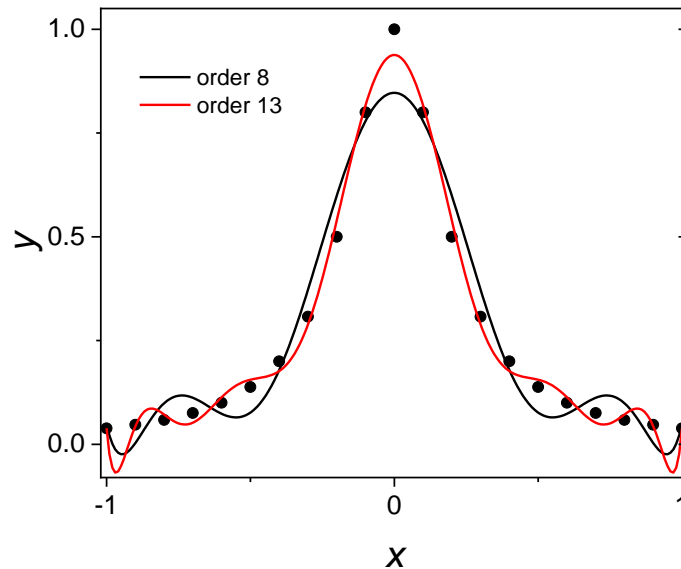


Fig. 7.19. Results of use of the polynomial approximation to the data file *rungea* in **Table 6.4**; points – raw data, black line – approximation by the polynomial of 8th order, red line – approximation by 13th order.

It is clear that the 8th order is not sufficient to approximate the data because the peak of the curve is attenuated and 13th order reproduces better the peak but it increases oscillations between points. In this case polynomial approximation cannot be used for such a function.

However, the approximation is improved when many more experimental points are used. This is illustrated in Example 7.11.

Example 7.11.

Use polynomial approximation for data file *data3* containing 401 noisy data points. Approximations were carried out for few polynomial orders and order 13 was chosen (see data file *data_13_401* and Excel file *Examples7* sheet *Ex. 7.11* and folder *E7-11*). The results of such approximation are displayed in Fig. 7.20. In this case the approximating line is quite smooth because number of points is large.

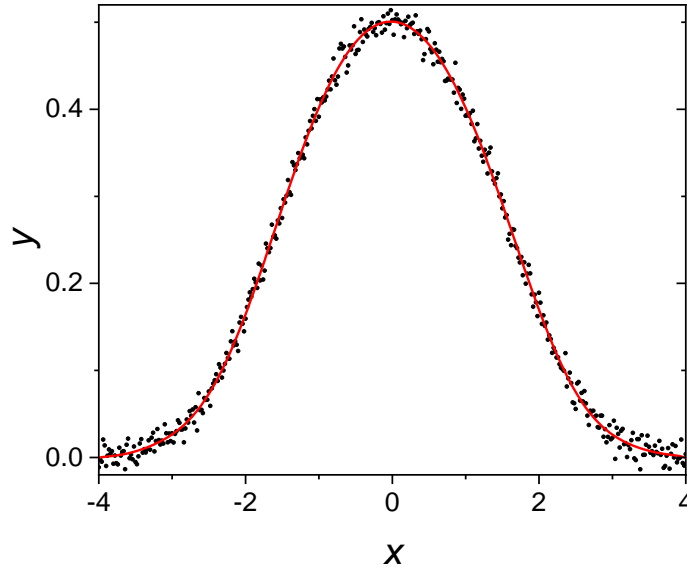


Fig. 7.20. Results of the polynomial smoothing of 401 points in data file *data* by the polynomial of 13th order.

Polynomial smoothing using orthogonal polynomials is a powerful technique which works well for very dense data files and might not work for certain functions especially when number of data points is low.

7.5 FFT smoothing

In section 6.1 it was pointed out that each series of discrete points, x_i , can be described exactly by a polynomial. Such a series of points can also be exactly described by the Fourier series. Fourier transform⁶⁹⁻⁷¹ is a linear integral transform which converts function of time, $f(t)$, into the complex function of the parameter called (angular) frequency, ω , $F(\omega)$:

$$F(\omega) = \int_{-\infty}^{\infty} f(t)e^{-j\omega t} dt \quad (7.15)$$

Although historically parameter t is called time it might be any x value.

In practice a finite time interval, 0 to T , is considered (in which data are acquired) and it is assumed that the same function is repeated between T and $2T$, $2T$ and $3T$, and so on. Then, Eq. (7.15) becomes:

$$F(\omega) = \int_0^T f(t)e^{-j\omega t} dt = \int_0^T f(t)\cos(\omega t)dt + j \int_0^T f(t)\sin(\omega t)dt \quad (7.16)$$

where Euler's formula was used:

$$e^{-j\omega t} = \cos(\omega t) - j\sin(\omega t) \quad (7.17)$$

In data analysis we are interested in the discrete Fourier transform, DFT, applied to the series uniformly distanced N points numbered from 0 to $N-1$:

$$f(0), f(\Delta t), f(2\Delta t), \dots, f(i\Delta t), \dots, f((N-1)\Delta t) \quad (7.18)$$

Because all points are uniformly distributed the value of Δt is not important and Eq. (7.18) might be written as:

$$f(0), f(1), f(2), \dots, f(i), \dots, f(N-1) \quad (7.19)$$

Application of Eq. (7.16) to the data in Eq. (7.19) leads to the formula for the DFT:

$$F(k) = \frac{1}{N} \sum_{i=0}^{N-1} f(i) \exp(-j\omega_k t_i) = \frac{1}{N} \sum_{i=0}^{N-1} f(i) \exp\left(-\frac{j2\pi ki}{N}\right) \quad (7.20)$$

where

$$\omega_k t_i = 2\pi \nu_k t_i = \frac{2\pi k}{T} i \Delta t = \frac{2\pi ki \Delta t}{N \Delta t} = \frac{2\pi ki}{N} \quad (7.21)$$

$\omega_k = 2\pi \nu_k$, $\nu_k = k / T$ is the frequency. Eq. (7.20) transforms points $f(i)$ into complex function $F(\omega)$:

$$\begin{array}{ccc} f(0) & & F(0) \\ f(1) & & F(1) \\ f(2) & DFT & F(2) \\ \dots & \Leftrightarrow & \dots \\ f(i) & & F(k) \\ \dots & & \dots \\ f(N-1) & & F(N-1) \end{array} \quad (7.22)$$

The fundamental frequency ($k = 1$) is:

$$\nu_1 = \frac{1}{T} = \frac{1}{N \Delta t} \quad (7.23)$$

and the harmonic frequencies are: $2\nu_1, 3\nu_1, 4\nu_1 \dots$. Therefore, each frequency is

$$\nu_k = \frac{k}{N \Delta t} \quad (7.24)$$

Although the DFT produces N points in the frequency space the new information is included only for frequencies from 0 to $k = N/2$ that is $\nu_{N/2} = 1/(2\Delta t)$ and further the same values of F are repeated (with the same sign for the real and negative sign for the imaginary part). This largest frequency, $\nu_{N/2}$, is called **Nyquist frequency**.

The FT programs accept one column of data only and they are internally numbered from 0 to $N - 1$. The first value of $F(0)$ is real and it is simply the average of all the data points.

From Fourier transformed values the original data might be calculated using inverse Fourier transform:

$$f(i) = \sum_{k=0}^{N-1} F(k) \exp\left(\frac{j2\pi ki}{N}\right) \quad (7.25)$$

This equation might be also written using real values of the modulus $|F|$ and phase angle φ :

$$f(t_i) = \sum_{k=1}^{N-1} |F_k| \cos(\omega_k t_i + \phi_k)$$

where

$$|F_k| = \sqrt{(F_k')^2 + (F_k'')^2}, \quad \phi_k = \text{atan}\left(\frac{F_k''}{F_k'}\right) \quad (7.26)$$

$$F_k' = \text{Re}(F_k), \quad F_k'' = \text{Im}(F_k)$$

and indices ' and '' denote real and imaginary functions, respectively.

It can be noticed that the presence of discontinuities (producing so called leakage⁶⁹) or the noise introduces high frequencies but the main shape of the function might be usually approximated with a few low frequencies. This leads to the smoothing based on DFT. It should be added that so called fast Fourier transform, FFT, avoids repetition of many calculations and the calculations are performed more quickly. However, FFT demands that the number of points is $N = 2^n$ where n is an integer number.

The idea of FFT smoothing⁷² is to take noisy experimental data, make FT, cut out high frequencies responsible for noise, and carry out inverse transform. Then, compare the original noisy and the smoothed data. The highest frequency to be cut must be found experimentally. Two main types of filters in Fourier space are square and parabolic.⁷² The square filter cuts data abruptly but the parabolic filter does this more smoothly. These filters for 12 points are shown in Fig. 7.21. Parabolic filter was calculated using:

$$\begin{aligned} w_i &= 1 - \left(\frac{i}{12}\right)^2 & \text{for } i = 0 \dots 12 \\ w_i &= 0 & \text{for } i > 12 \end{aligned} \quad (7.27)$$

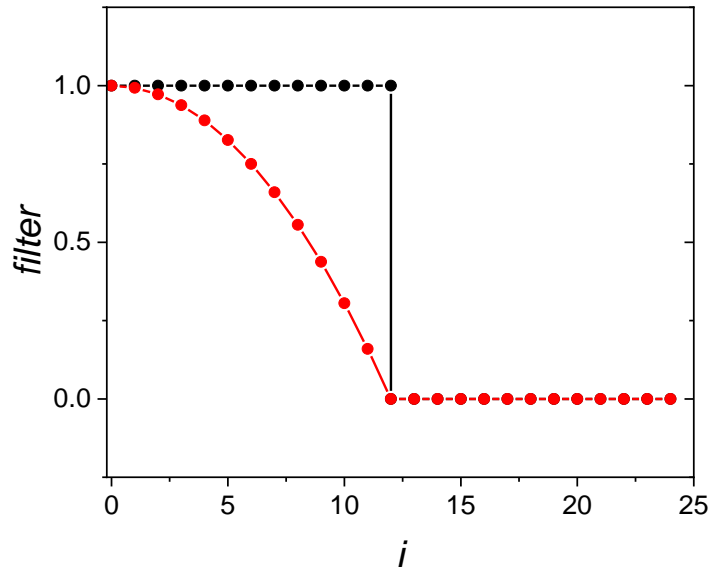


Fig. 7.21. Comparison of the square (black) and parabolic (red) 12 points filters used in FT smoothing.

This filter is multiplied by the FT values $F(k)$ which cuts-off high frequencies. This procedure is illustrated in Example 7.12. The calculations are carried out using DFT/FFT program *fftsm.exe*.

Example 7.12.

Use FT smoothing of the data in file *data1a* (containing one column of 401 points) using square and parabolic filters. These data are displayed in Fig. 7.22. (Because initial and final data values are the same, data rotation to avoid discontinuities is not necessary, see Example 7.13).

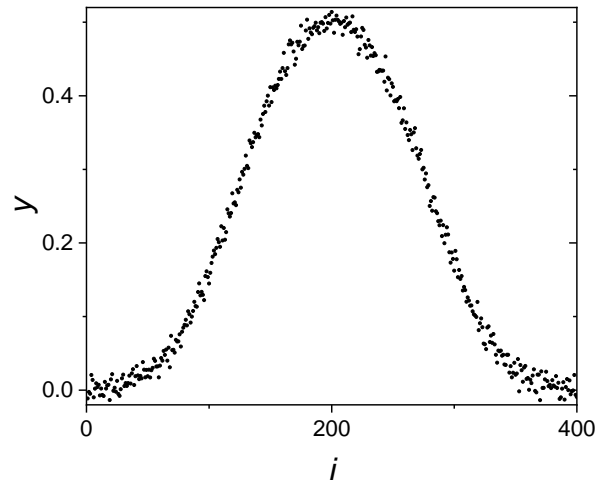


Fig. 7.22. Raw data to be smoothed, in file *data1a*.

The Fourier transform of these points is displayed in Fig. 7.23.

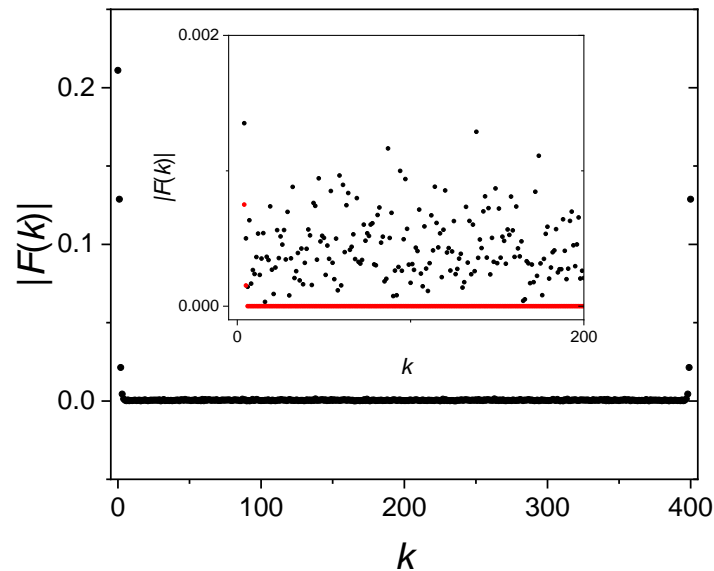


Fig. 7.23. Magnitude (modulus) of the Fourier transform of the data in Fig. 7.22, black points, for $k = 0$ to 400; inset – zoom for $k \geq 4$, data filtered using parabolic filter for 6 points – red points.

The Fourier transform contains few larger values at low frequencies followed by the uniformly distributed noisy data at higher frequencies. These data are responsible for the observed noise.

Using parabolic filter only 6 first points of the Fourier transform were conserved, they are attenuated by multiplication by the filter, Eq. (7.27), but with the number of points 6 instead of 12. This operation puts zero value for all the frequency points for $k > 6$, Fig. 7.23, red points.

Next, inverse Fourier transform was carried out to obtain the results in Fig. 7.24. It can be noticed that only few frequencies are sufficient to reproduce the shape of data. However, using 6 points parabolic filter deformed slightly the obtained curve and the peak values lay below the experimental points and are attenuated. Increasing filter to 20 points causes some oscillation at the foot of the peak, see the inset. It seems that using 12 point parabolic filter presents a smooth line without attenuation. The results are in data files *dp6*, *dp12*, and *dp20*.

However, when using square filter less points could be used in smoothing. The results of using 5 and 12 point square filter are shown in Fig. 7.25. It seems that good smoothing is obtained using 5 point filter while using 12 points displays some small oscillation at the foot of the peak. The results are in the data files *ds5*, *ds12*, and *ds20* and also in Excel file *Examples7.xlsx* sheet *Ex.7.12*.

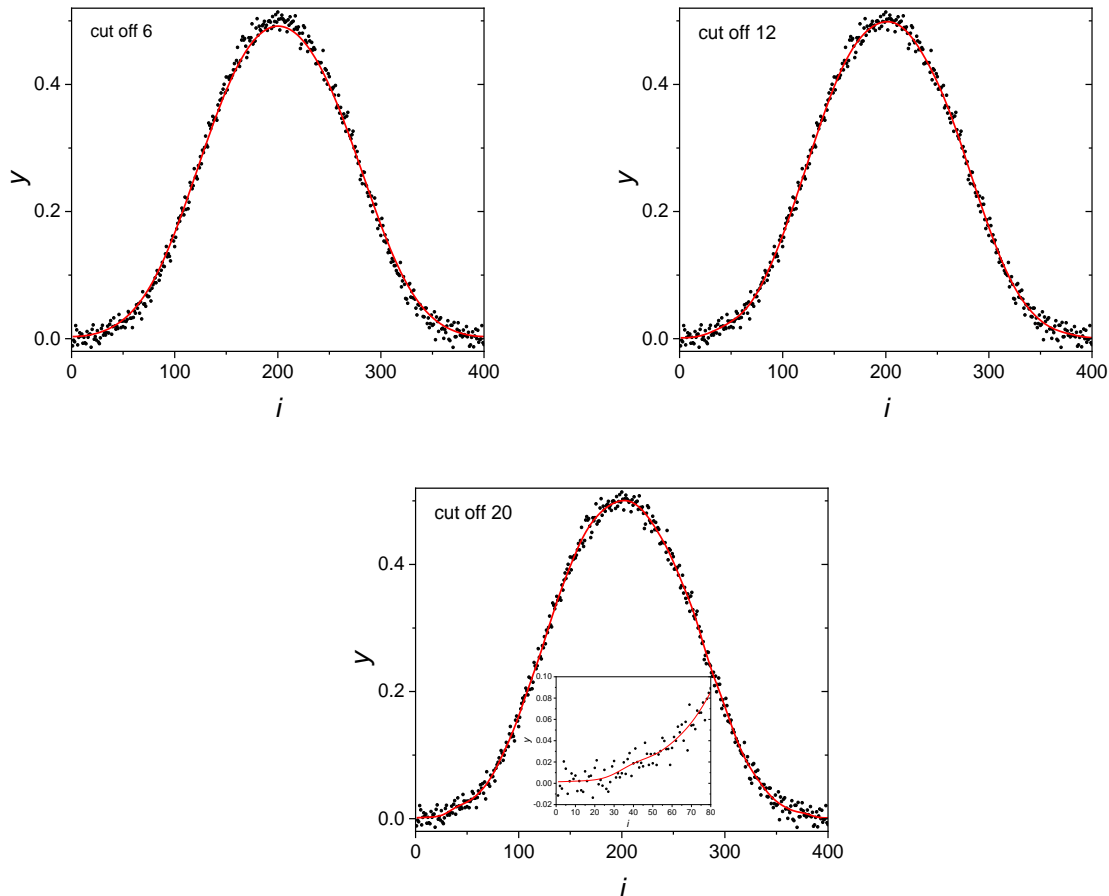


Fig. 7.24. FT smoothing of the data in Fig. 7.22 using parabolic filter for 6, 12, and 20 points.

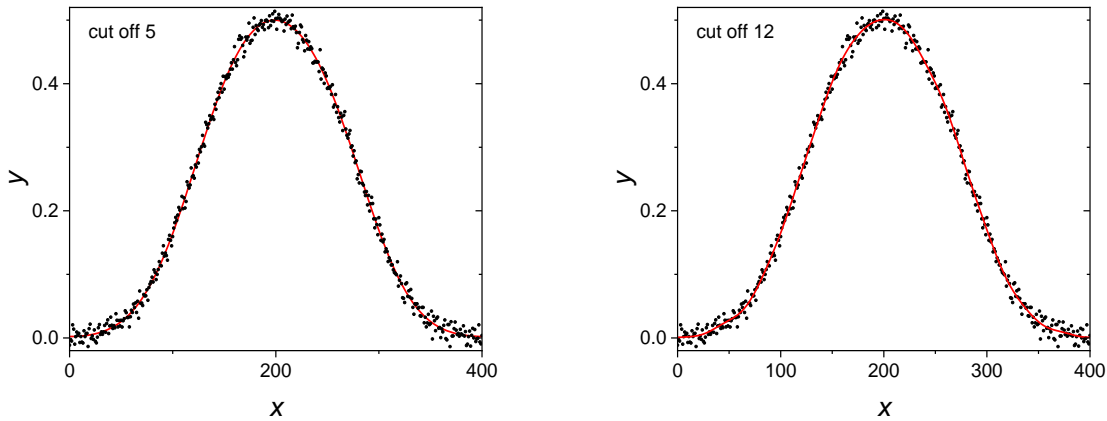


Fig. 7.25. FT smoothing of the data in Fig. 7.22 using square filter for 5 and 12 points.

Of course FT smoothing does not work well when only few data points are used. Next example will be shown for smoothing of the S-shape curve.

Example 7.13.

Carry out FT smoothing of the data in the data file *ss*, containing 501 points, Fig. 7.26.

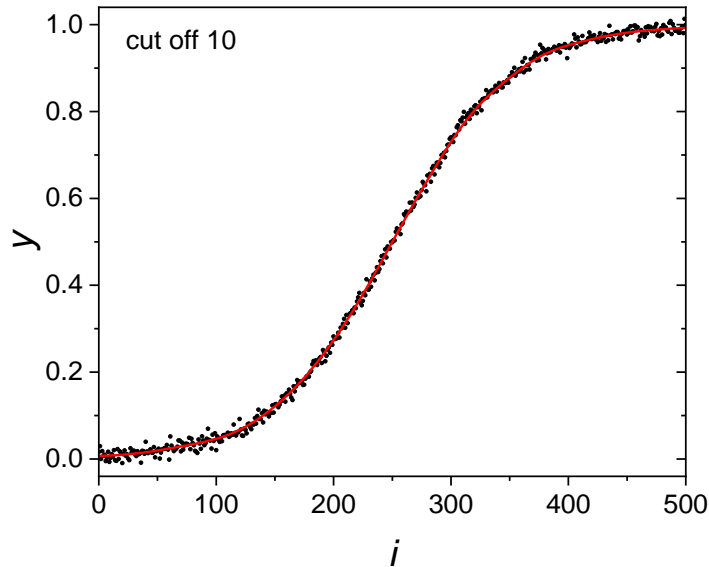


Fig. 7.26. Plot of raw data *ss* (symbols) and the smoothed red line obtained by FT smoothing with 10 points parabolic filter (after rotation of the data).

It is obvious that the first and the last points are very different which creates discontinuity of the FT (and introduces high frequencies which are important for reproduction of the initial shape), therefore, rotation of the spectrum must be first performed (to get the initial and the final data the same). The program used is *fftsm.exe*. The rotated data are displayed in Fig. 7.29.

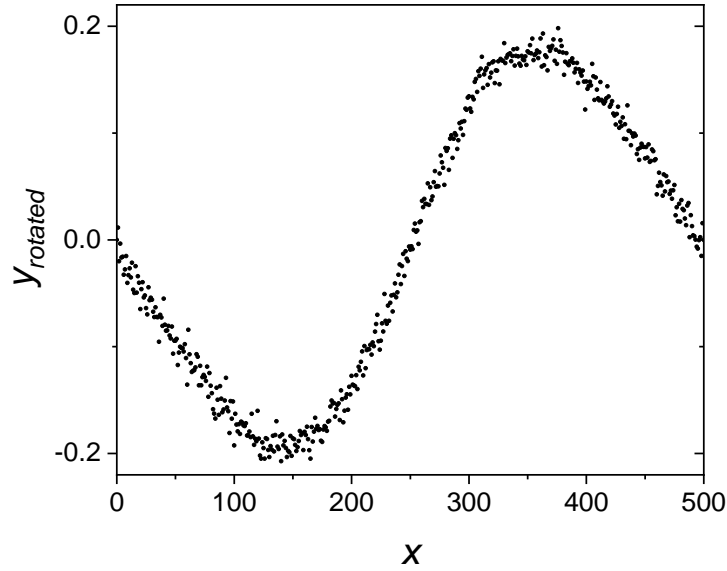


Fig. 7.27. Rotated spectrum of raw data from Fig. 7.26.

Now, FT can be applied to the rotated spectrum using parabolic filter. It was found that choosing 10 points filter is optimal. The FT spectrum (modulus) is shown in Fig. 7.28.

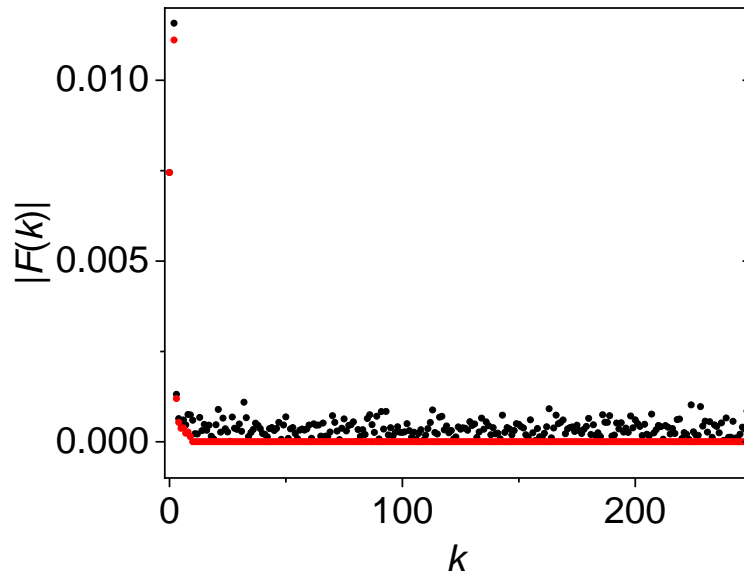


Fig. 7.28. Fourier transform of the data in Fig. 7.27 (black symbols) and filtered spectrum using 10 points parabolic filter (red symbols).

The results of FT smoothing are shown in Fig. 7.26. Only 10 frequencies are sufficient to reproduce the shape of the original data. The results are in file *ss_10* and in Excel file *Examples7.xlsx* sheet *Ex. 7.13* and folder *E7-13*.

FT is a very powerful technique, but it is applicable to the uniformly distributed data without discontinuities. It works better for larger data files.

7.6 Smoothing splines

In Section 6.2 approximating splines which pass a third order polynomial through all the points were presented. There are also smoothing splines can be used to approximate the data, but the approximating line does not necessarily pass by all the points. The smoothing spline minimizes the second derivative $S''(x)$ of the spline described in Eq. (6.7):

$$\int_{x_1}^{x_N} S''(x)^2 dx \quad (7.28)$$

subject to the constraint:

$$\sum_{i=1}^N \left| \frac{S(x_i) - y_i}{w_i} \right|^2 \leq \sigma \quad (7.29)$$

where σ is the assumed smoothing parameter corresponding to the standard deviation and w_i is the weight of the point. If we know the weights of the individual points we can use them, however, usually unit weights are assumed. The use of smoothing splines is illustrated in the examples below.

Example 7.14.

Take data file *d3* from Example 7.9, they are non-uniformly distributed with noise. Let us try with some standard deviations σ . First, let us try $\sigma = 2$, the results (function and its first derivative) were generated for 200 points and are in the file *d3_2*. This file contains 4 columns: *x*, *y* smoothed, dy/dx , and d^2y/dx^2 . These results are displayed in Fig. 7.29.

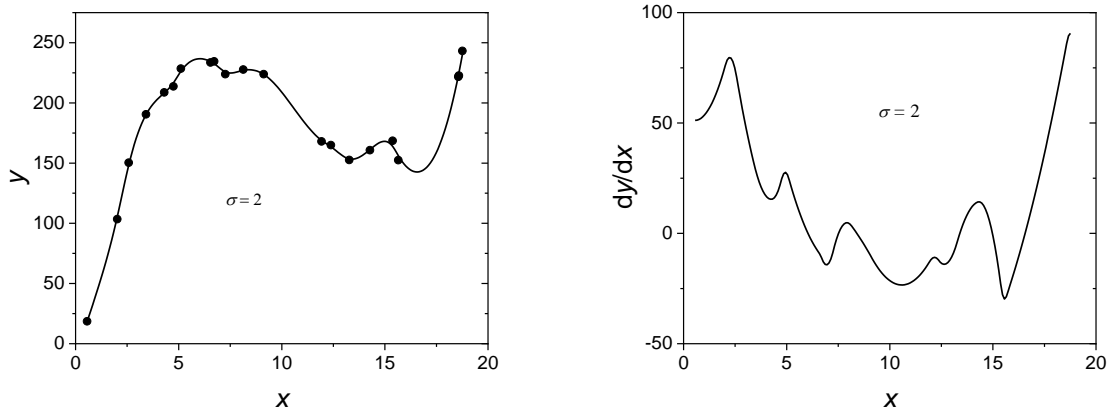


Fig. 7.29. Application of smoothing splines to data, *d3*, represented as symbols assuming smoothing parameter $\sigma = 2$; left function and right its first derivative.

It is obvious that the assumed smoothing parameter is too small, and the approximating spline tries to follow too closely the experimental points therefore little smoothing is done. Moreover, the first derivative oscillates which confirms the presence of the noise.

Next, larger value $\sigma = 15$ was chosen. The results are in Fig. 7.30. Now, although the first derivative is smooth, the approximating line is passing below or above some groups of points indicating that there is too much smoothing.

Finally, after several trials the optimal value of $\sigma = 7$ was chosen, see Fig. 7.31. In this case the first derivative looks smooth and the smoothing line is passing close to the experimental points. But as it was mentioned this smoothing contains some subjectivity. These results are also displayed in *Examples7.xlsx* sheet *Ex. 7.14* and folder *E7-14*.

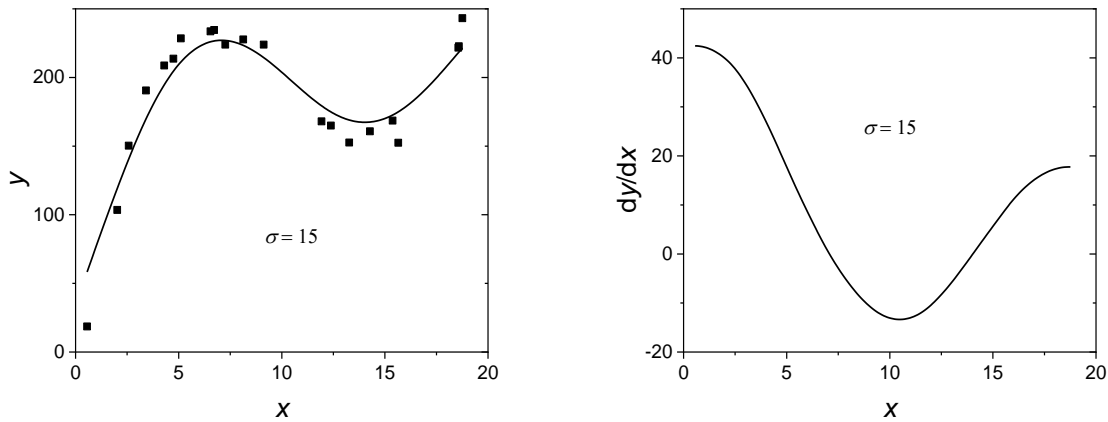


Fig. 7.30. Application of smoothing splines to data represented as symbols assuming smoothing parameter $\sigma = 15$; left function and right its first derivative.

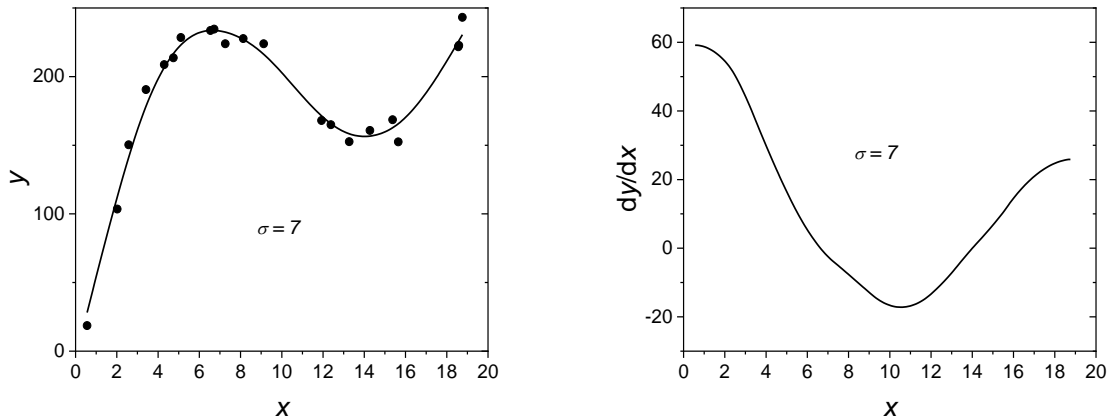


Fig. 7.31. Application of smoothing splines to data represented as symbols assuming smoothing parameter $\sigma = 7$; left function and right its first derivative.

7.7 Cross-validation

Finally, when there is large amount of noisy data one can use method which tries to automatically find the optimal value of the smoothing factor by cross-validation. This method uses cubic smoothing splines and tests the model's ability to predict new data that was not used in its estimation.

An example is shown below.

Example 7.15.

Let us apply spline smoothing with the automatically found optimal value of the smoothing parameter to 1024 data points. Noisy data are file *data1* and the program *smsplcv.exe*. The results displaying the smoothed line and its derivative are displayed in Fig. 7.32. It is evident that the calculated line is smooth without any distortions and its first derivative is also smooth. These results are also shown in Excel file *Exercises7.xlsx* sheet *Ex. 7.15* and folder *E7-15*.

It should be added that applying this program to the data in Example 7.14 (file *d3*) does not produce good smoothing because the number of points is too small.

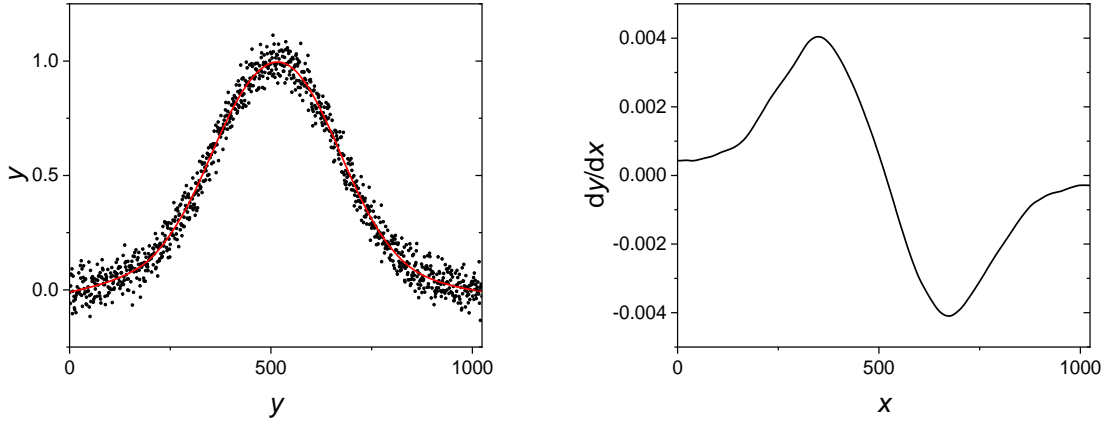


Fig. 7.32. Application smoothing splines with the computer optimization of the smoothing parameter by cross-validation; left: experimental data (black points) and the computed smoothed line (red); right: computed first derivative of the smoothed function.

7.8 B-splines

B-splines or basic splines are the type of spline functions that demand minimal operator support necessary for the calculations.⁷³ Any spline function of a given degree can be expressed as a linear combination of B-splines of the same degree. B-splines are used in curve fitting and numerical differentiation of the experimental data.

B-splines, $B_{i,k}(x)$, of the order k and interval x_i to x_{i+1} , are defined recursively. The B-spline of the order 0 is defined as:

$$B_{i,0}(x) = \begin{cases} 1, & \text{for } x_i \leq x \leq x_{i+1} \\ 0, & \text{otherwise} \end{cases} \quad (7.30)$$

Higher order B-splines are obtained by recurrence:

$$B_{i,k}(x) = \frac{x - x_i}{x_{i+k} - x_i} B_{i,k-1}(x) + \frac{x_{i+k+1} - x}{x_{i+k+1} - x_{i+1}} B_{i+1,k-1}(x) \quad (7.31)$$

These equations indicate that $B_{i,0}(x)$ is a step function of 1 defined between x_i and x_{i+1} , $B_{i,1}(x)$ is piecewise linear function going from 0 to 1 and back to 0, defined between x_i and x_{i+2} , $B_{i,2}(x)$ is a piecewise quadratic function defined between x_i and x_{i+3} , etc.

Assuming $x_i = i$ the first three B-splines are:

$$B_{1,0}(x) = \begin{cases} 1, & \text{for } 1 \leq x \leq 2 \\ 0, & \text{otherwise} \end{cases}$$

$$B_{2,0}(x) = \begin{cases} 1, & \text{for } 2 \leq x \leq 3 \\ 0, & \text{otherwise} \end{cases} \quad (7.32)$$

$$B_{3,0}(x) = \begin{cases} 1, & \text{for } 3 \leq x \leq 4 \\ 0, & \text{otherwise} \end{cases}$$

$$B_{1,1}(x) = \begin{cases} x-1, & \text{for } 1 \leq x \leq 2 \\ 3-x, & \text{for } 2 \leq x \leq 3 \\ 0, & \text{otherwise} \end{cases} \quad (7.33)$$

$$B_{2,1}(x) = \begin{cases} x-2, & \text{for } 3 \leq x \leq 3 \\ 4-x, & \text{for } 3 \leq x \leq 4 \\ 0, & \text{otherwise} \end{cases}$$

$$B_{1,2}(x) = \begin{cases} \frac{(x-1)^2}{2}, & \text{for } 1 \leq x \leq 2 \\ \frac{(x-1)(3-x)}{2} + \frac{(4-x)(x-2)}{2}, & \text{for } 2 \leq x \leq 3 \\ \frac{(4-x)^2}{2}, & \text{for } 3 \leq x \leq 4 \\ 0, & \text{otherwise} \end{cases} \quad (7.34)$$

and so on for higher order splines. The first three B-splines are displayed in Fig. 7.33.

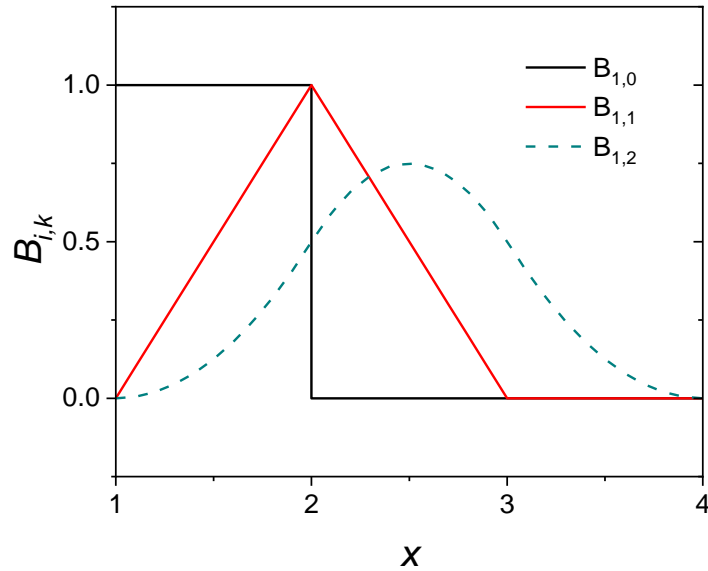


Fig. 7.33. Plot of the first three B-splines calculated assuming $x_i = i$.

Finally, the approximating B-spline function of the order n is the sum of individual B-splines:

$$S_n(x) = \sum_{j=1}^n \alpha_j B_{j,n}(x) \quad (7.35)$$

where α_j are coefficients which should be found by minimization of the sum of squares (least-squares method):

$$S^2 = \sum_{i=1}^N [y(x_i) - S_n(x_i)]^2 \quad (7.36)$$

The approximated function values might be calculated for any x producing a smoothed function. The approximation might be carried out for different spline orders, usually between 2 and 4. Often a weighted least-squares are used:

$$S^2 = \sum_{i=1}^N w_i [y(x_i) - S_n(x_i)]^2 \quad (7.37)$$

where w_i must be supplied by the operator.

Application of B-splines to smoothing data is shown in Example 7.16.

Example 7.16.

Apply B-spline smoothing to the data in Example 7.9. They are included in data file *d3*.

The program used is *bssm.exe* and the second order (parabolic) and third order (cubic) approximations were used producing 200 calculated points. The obtained results are in files *d3_2* for parabolic and *d3_3* for cubic smoothing. These results are displayed in Fig. 7.34.

It is evident that parabolic B-splines produce smoother curve while their derivative consists of segments of straight line while cubic B-splines follows closer the data and display some overshoot (for $x \sim 17.5$) while the first derivative consists of smooth second order parabolas.

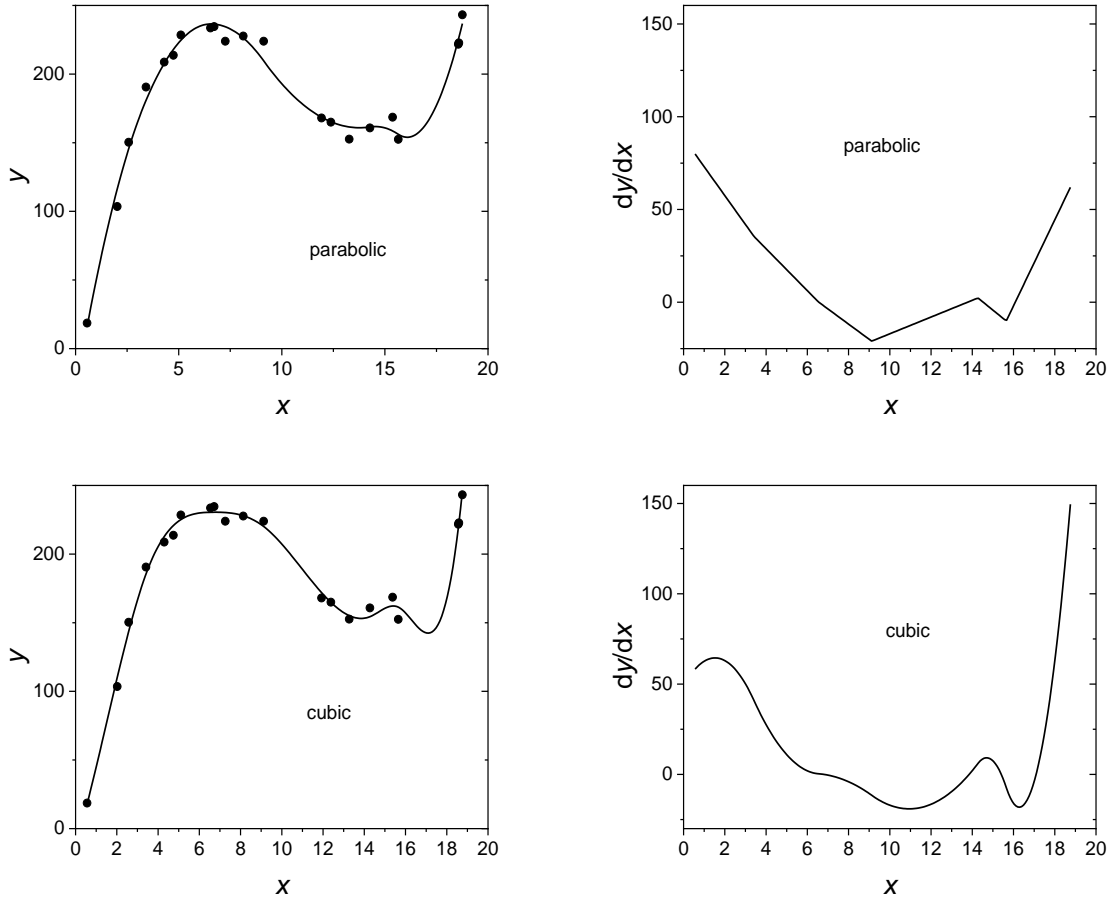


Fig. 7.34. Application of B-splines to smooth the experimental data using parabolic and cubic splines; approximations and the first derivatives are shown.

7.9 LOESS/LOWESS

Sometimes one has to deal with fairly large and densely sampled data to find a trend in such plots. These problems appear in biology, climatology, social sciences, etc., but more rarely found in physical sciences. Very often in such cases the classical procedures do not work well. LOESS/LOWESS techniques are new non-parametric methods which combine local regression models: LOESS (LOcal regrESSion) and LOWESS (LOcally Weighted regrESSion or locally weighted scatterplot smoothing).^{74,75} In these methods local polynomials usually of the first (or second order) are used to fit a fraction of the data using a weighted moving average. In each case a window Δx is chosen to calculate the smoothed value at a chosen x_0 .

The regression weights, w_i , for each point inside a given subset (window) are defined as:

$$w_i = \left(1 - \left| \frac{x_0 - x_i}{d(x)} \right|^3 \right)^3 \quad (7.38)$$

where x_0 is the point in which the function is smoothed, x_i are the points in the defined window, and $d(x)$ is the distance from x_i to the most distant x value in the window. It is clear, that the data

close to the point x_0 are the most important in regression and those which are further are less important. The plot of w as a function of the parameter $u = (x_0 - x_i)/d(x)$ is shown in Fig. 7.35.

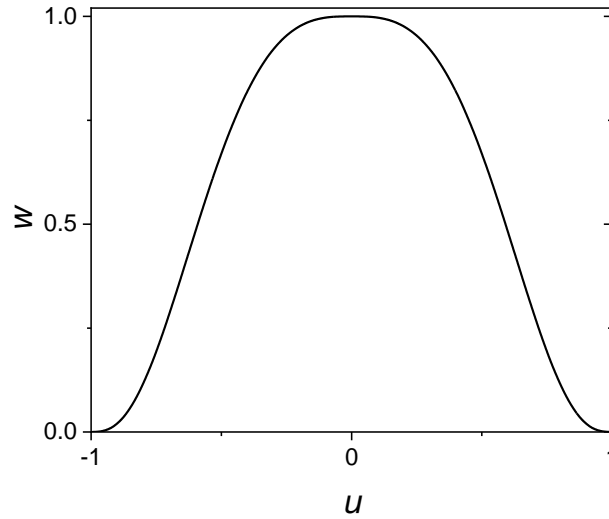


Fig. 7.35. Plot of the weights in LOWESS method versus parameter $u = (x_0 - x_i)/d(x)$, Eq. (7.38).

It is obvious that the highest weights are around the central point: $u \approx 0$, i.e. $x \approx x_0$ and they decrease rapidly with the distance from x_0 . After calculating weights in the window, the weighted linear (or parabolic) regression is performed and the value of smoothed function at x_0 is calculated. This procedure does not produce any function describing the data. An example of such calculations is shown in Example 7.17.

Example 7.17.

Use LOWESS to smooth data in file *d3* (see Example 7.16). The manual calculation is shown in Excel file *Examples7*, sheet *Ex. 7.17*. Data file contains 21 points and 7 point window was chosen for linear smoothing. For each data subset first the distance $|x_0 - x_i|$ is calculated then the distance is scaled by division by $|x_0 - x_{max}|$, $u = (x_0 - x_i)/d(x)$, where x_{max} is the most distant point in the window (it might be situated before or after x_0). Finally, the weights of all the points in the window are calculated as $(1 - u^3)^3$. Then a weighted linear regression is applied to each window. The program *polfit.exe* has an option 2 to read x , y , σ_i therefore standard deviations were calculated as $\sigma_i = 1/\sqrt{w_i}$ and are included in the last column. Such data files are shown in the columns P to Q and are in the files *t1* to *t8*. In *polfit.exe* polynomial order of 1 was chosen and the results are in files *t1r* to *t8r*. The intercept and slope of the straight line were copied and used to calculate the value at x_0 . Following these steps allows to better understand how the LOWESS works but are not practical for large data files. The program *lowess.exe* was supplied to calculate automatically the smoothed values. Instead of asking for the number of points in the window it asks for the fraction of all data. In our example number of points is 7 and the fraction is $7/21 = 0.3333333$.

The program allows also to use either classical least squares regression (option 0) or the iterative robust regression (option 2) which is less sensitive to outliers (which can affect the results). In this case classical linear least-squares method was used. The results are included in file

$d3res$, displayed in Fig. 7.36, and are included in Excel file *Examples7.xlsx*, sheet *Ex. 1.17*. The obtained line is smooth and passes close to the original data.

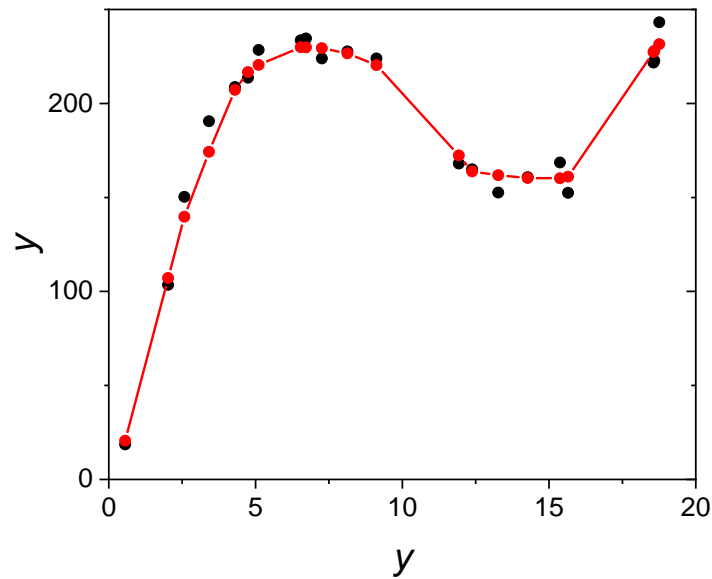


Fig. 7.36. Plot of the raw (black) and the smoothed (red) data using LOWESS for 7 points window with linear approximation.

Another example of a very noisy data with possible outliers are considered in Example 7.18.

Example 7.18.

Use LOWESS to find the trend in the raw data (file *dat*), see Fig. 7.37.

Two methods were used: non-robust linear least-squares and robust linear regression methods. It can be noticed that at larger values of x there are several outliers. LOWESS was used for the data fraction 0.6 and these two methods. The non-robust method shows increase of the smoothed data at higher values of x while the robust method shows a decreasing line in that zone, Fig. 7.37. The results of calculations are included in file *dat_nr_0_6* for non-robust (option 0) and *dat_rob_0_6* for robust regression (option 2). These results are also included in Excel file *Examples7.xlsx*, sheet *Ex. 7.18* and folder *E7-18*.

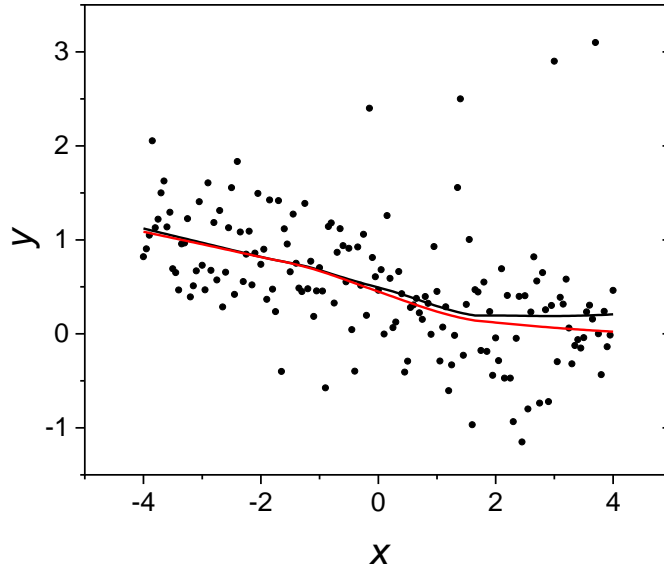


Fig. 7.37. Plot of the raw data *dat* (black symbols) and smoothing lines using LOWESS and non-robust (black) and robust (red) regression.

LOWESS is included in popular graphing programs like Origin. However, use of the supplied program gives more control on the smoothing procedures.

7.10 Digital differentiation and integration

7.10.1 Digital differentiation

The derivative of the function $f(x)$ in the point x_0 is defined as the limit for $h \rightarrow 0$:

$$f'(x_0) = \lim_{h \rightarrow 0} \frac{f(x_0 + h) - f(x_0)}{h} + \left[-\frac{h}{2} f''(\xi) \right] \quad (1.39)$$

that is the error is proportional to the function increment h and $x_0 \leq \xi \leq x_0 + h$. This is so called forward differencing or one-sided differencing.

A better description might be a central difference

$$f'(x_0) = \lim_{h \rightarrow 0} \frac{f(x_0 + h) - f(x_0 - h)}{2h} + \left[-\frac{h^2}{6} f'''(\xi) \right] \quad (1.40)$$

where the error is proportional to h^2 . There are also other formulas, e.g. four point formula:

$$f'(x) = \frac{-f(x+2h) + 8f(x+h) - 8f(x-h) + f(x-2h)}{12h} + \left[O(h^4) \right] \quad (1.41)$$

With the error proportional to h^4 . As the value of h is small because $h > h^2 > h^4$ and the error of estimation of the derivative decreases.

Such formulas are used if function $f(x)$ might be calculated for any value of x . However, in experimental sciences we acquire limited number of experimental points which contain some noise. Differentiation of the noisy data leads to oscillations of the derivative. Therefore, a

smoothing must be used before calculation the derivative. Of course, too much smoothing distorts the function while too little leaves some noise (oscillations).

In earlier chapters it was shown that smoothing methods using Savitzky-Golay filter, polynomial approximation, spline and B-spline smoothing furnish the first derivative. They can be easily used for such a purpose. However, the extent of smoothing influences strongly the obtained results and minimal smoothing threshold must be found by the experimentalist.

7.10.2 Numerical integration

Numerical integration is an easier operation than differentiation because integration reduces the random errors. This is because the integral of the random noise approaches zero as the number of points increases. Let us consider few simple integration formulas called quadratures.

The simplest method is the method of rectangles. Starting with the first point the surface area is calculated assuming that it is equal to that of a rectangle: $y_1h, y_2h, \dots, y_{N-1}h$ and the total surface area is the sum of the areas of these rectangles:

$$\int_{x_1}^{x_N} f(x)dx = y_1h + y_2h + \dots + y_{N-1}h = h \sum_{i=1}^{N-1} y_i \quad (1.42)$$

This is a very simple method which can be used during data acquisition of many points because it is a simple sum of the values. An example of such an operation is illustrated in Example 7.19 where function $y = x^3$ is integrated for x between 0.5 and 3 with the choice of only 6 points. First of all, it is clear that the sum of rectangles underestimates the integral. This is observed always on the increasing function while it is overestimating the integral for the decreasing part.

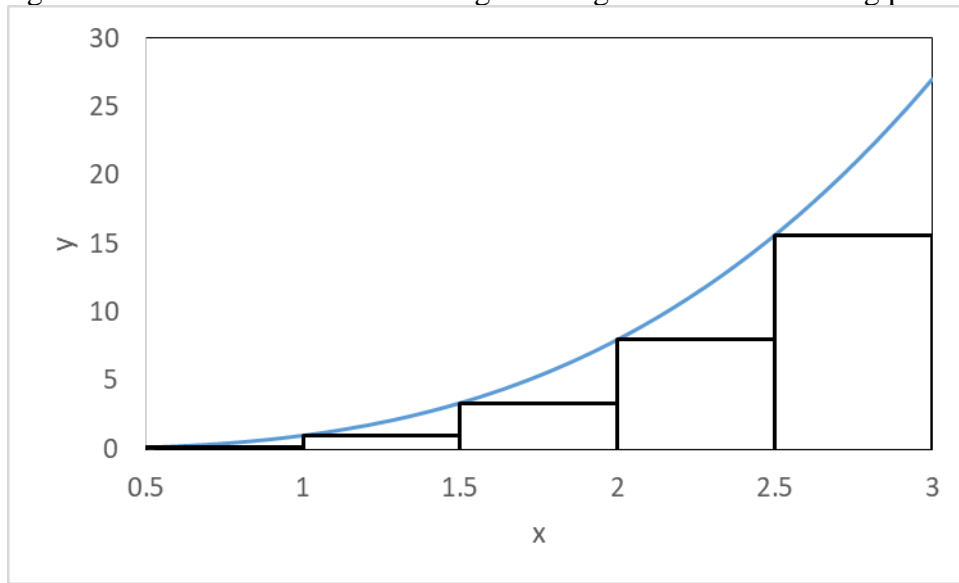


Fig. 7.38. Integration of the function $y = x^3$ defined at 6 points, using method of rectangles.

In this case the numerical integral is 14.0625 while the analytical integral is 20.2344 which produces relative error of -30.5%. The calculations are illustrated in Example 7.19. However, decreasing the step h to 0.02 produces error of -1.32% and further decrease to 0.005 produces error of -0.33%. When $h \rightarrow 0$ the error also $\rightarrow 0$.

Of course, method of rectangle is a very crude method and a very simple improvement is the trapezoidal rule. In this case two adjacent points are connected by a straight line and a trapezium

is constructed. Its surface area $A = \Delta x(y_1 + y_2)/2$ approximates the integral. The total surface area is obtained by summation:

$$\int_{x_1}^{x_N} f(x)dx = \left[\frac{y_1 + y_2}{2} + \frac{y_2 + y_3}{2} + \dots + \frac{y_{N-1} + y_N}{2} \right] h = \left[\frac{y_1}{2} + \sum_{i=2}^{N-1} y_i + \frac{y_N}{2} \right] h \quad (1.43)$$

This method is illustrated in Fig. 7.39.

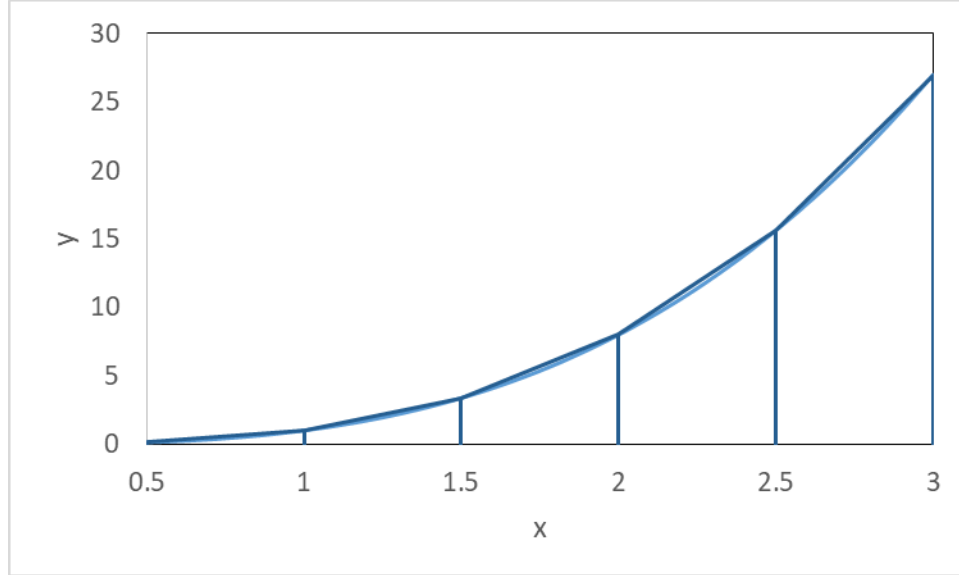


Fig. 7.39. Integration of the function $y = x^3$ defined at 6 points, using trapezoidal rule.

Integration in Fig. 7.39 produces value of 20.781 which gives the error of 2.7%, a very important decrease from -30.5% for method of rectangles. Decreasing the step h to 0.02 gives error of 0.0043% and further decrease to 0.005 produces error of $2.7 \times 10^{-4}\%$. It is evident that the error of numerical integration decreases very quickly with decrease of the integration step.

One more improvement is obtained using interpolation by piecewise polynomials. Let us consider interpolation using second order polynomial. Such a polynomial passes exactly by three points. The surface area under such parabola is easily found as:

$$\int_{x_1}^{x_3} f(x)dx = \left(\frac{y_1 + 4y_2 + y_3}{3} \right) h \quad (1.44)$$

and for $N = 3 + 2i$ points it is:

$$\int_{x_1}^{x_N} f(x)dx = \left(\frac{y_1 + 4y_2 + 2y_3 + 4y_4 + 2y_5 + 4y_6 + \dots + y_N}{3} \right) h \quad (1.45)$$

which can also be written as:

$$\int_{x_1}^{x_N} f(x)dx = \frac{h}{3} \left(f(x_1) + 4 \sum_{j \text{ even}} f(x_j) + 2 \sum_{j \text{ odd}} f(x_j) + f(x_N) \right) \quad (1.46)$$

This is so called Simpson's rule.

In the data above, Example 7.19, there were 6 points and in accordance with the condition above the number of points for $i = 2$ should be $N = 3 + 2 \times 2 = 7$ points therefore the distance

between points was recalculated to $(3 - 0.5)/6 = 0.4166667$. Using this new set of points and using Simpson's rule the integration error is zero, because the integrated function is cubic and the error of this procedure is proportional to the fourth derivative $(h^4/180(x_N - x_1)f^{(4)}(\xi))$, which here is zero. There are other modifications as Simpson's rule (3/8 rule), Romberg rule, Gauss quadrature, etc., which yield greater precision.⁷⁶ Increasing number of points increases precision of the procedures.

Example 7.19.

Integrate data using method of rectangles, trapezoidal and Simpson's rule. Use 6 data points in Table 7.4 calculated using formula: $y = x^3$. The results are shown in *Examples7.xlsx*, sheet *Ex. 7.20*.

Table 7.4. Data calculated using $y = x^3$.

x	y
0.500	0.125
1.000	1.000
1.500	3.375
2.000	8.000
2.500	15.625
3.000	27.000

The analytical integral is:

$$Int = \int_{0.5}^3 x^3 dx = \frac{x^4}{4} \Big|_{0.5}^3 = 20.23438 \quad (1.47)$$

Applying method of rectangles, Eq. (1.42), the estimated integral is 14.0625, which shows error of -30.5%. This integration is shown in Fig. 7.38. Of course, the surface area of rectangles underestimates integral of the function. However, decreasing the step h from 0.5 to 0.02 or 0.005 decreases the error to -1.33% and -0.33%, respectively. Even such a primitive method might lead to reasonable results.

Using the method of trapezes, Fig. 7.39, reduces the error for the steps 0.5, 0.02 and 0.005 to 2.70%, 0.0043% and 0.00027%, respectively.

Finally using the Simpson's rule gives the exact answer because of the nature of the approximation by the parabola which after integration gives the cubic function. Of course, in other cases only reduction of error will be observed.

Example 7.20.

Integrate numerically exponential function: $\int_{0.5}^{3.5} \exp(x/2) dx$ from 0.5 to 3.5 step 0.5, using methods of rectangles, trapezes, and Simpson's rule.

The value of the integral calculated using Maple is 8.941154519. Calculations are shown in *Examples7.xlsx*, sheet *Ex. 7.20* give the following results:

Method	value of integral	err %
rectangles	7.87003028	-11.979708
trapezes	8.98767459	0.520292
Simpson's	8.94134712	0.002154

It is obvious that the Simpson's rule gives quite good results even for a large step of $h = 0.5$ although the shape of the exponential function is quite different from the parabolic function.

However, in above described methods, the values of function must be available at desired intervals. When fewer points are available, and they contain noise, approximations must be carried out and integration of the approximating function carried out. One of the useful methods is spline interpolation with subsequent integration of splines. The data file used might be the raw noisy data or smoothed by one of the smoothing methods. Below, an example of spline integration is presented.

Example 7.21.

Use spline integration of the raw data file from Example 7.15 (file *raw*) and of the smoothed data by splines with cross-validation (file *smoo*).

The program used for spline integration is *splint.exe* in folder *E7-21*. The results are included in files *rawint* and *smooint*. The total integral in the x range $[1, 1024]$ is 390.751 for the raw and 390.742 for the smoothed function. The difference between these two integrals is only 0.0023% which confirms that integration reduces the noise. Of course, if the integral should be determined from 1 to any x the integration of the noisy data shows initially oscillations which decrease with the increase in x and the integration of smoothed data is preferred.

7.11 Conclusion

The procedures presented above are used to smooth the experimental noisy data or to find general trend in the noisy data. As it was mentioned above smoothing is a little subjective procedure. It is mainly used in plotting the data, but it is indispensable when a derivative of the noisy data must be estimated. The methods using simple digital filters, Savitzky-Golay or FT demand uniformly distributed data with constant Δx and the other: polynomial approximation, smoothing cubic splines or B-splines and LOWESS accept non-uniformly distributed data. The user should experiment with different techniques to find the best method.

Determination of the derivative of the function is easy when it can be calculated for any value of x and value of the step h is small. In the case of noisy data these data must be smoothed first to reduce oscillations of the derivative.

Numerical integration can also be carried out easily when the function might be calculated for any value of x . Integration of the noisy data reduces the random noise (average of the random noise is zero), however, when there are fewer points, initial smoothing should be carried out.

8 Excel functions

Normal distribution:

$$P_G(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{NORM.DIST}(x, \mu, \sigma, \text{FALSE})$$

Normalized normal distribution

$$P_G(z, 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad \text{NORM.S.DIST}(z, \text{FALSE})$$

Standard deviation of the population

$$\sigma = \sqrt{\frac{\sum_{i=1}^N (x_i - \mu)^2}{N}} \quad \text{STDEV.P}(\text{cell1:cellN})$$

Arithmetic mean

$$\bar{x} = \frac{\sum_{i=1}^N x_i}{N} \quad \text{AVERAGE}(\text{cell1:cellN})$$

Integral of the normal distribution

$$\int_{-\infty}^x P_G(x, \mu, \sigma) dx = \int_{-\infty}^x \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx = \text{NORM.DIST}(x, \mu, \sigma, \text{TRUE})$$

or for the normalized normal deviation:

$$\int_{-\infty}^z P_G(z, 0, 1) dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-\frac{z^2}{2}} dz = \text{NORM.S.DIST}(z, \text{TRUE})$$

Value of x for which $\int_{-\infty}^x P_G(x, \mu, \sigma) dx = \alpha$

$$x(\alpha) = \text{NORM.INV}(\alpha, \mu, \sigma)$$

and for the normalized distribution

$$z(\alpha) = \text{NORM.S.INV}(\alpha)$$

Sample standard deviation

$$s = \sqrt{\frac{\sum_{i=1}^N (x_i - \bar{x})^2}{N - 1}}$$

STDEV.S(cell1:cellN)

Variance
of the population
of the sample

VAR.P(cell1:cellN)

VAR.S(cell1:cellN)

Statistics of the mean are obtained from DATA, Data Analysis, Descriptive Statistics

Student t distribution function

T.DIST(t, df, FALSE)

t -value of the two-tailed test
 $t(\alpha'', df)$

T.INV.2T(α, df)

t -value of the one-tailed test
 $t(\alpha', df)$

T.INV(α, df)

Regression is calculated using DATA, Data Analysis, Regression

Slope in the linear regression

LINEST($y1:yN, x1:xN$)

Intercept in the linear regression

INTERCEPT($y1:yN, x1:xN$)

Correlation coefficient in the linear regression

CORREL($y1:yN, x1:xN$)

F test, probability function
 $P(f, df_1, df_2)$

F.DIST($f, df_1, df_2, \text{FALSE}$)

F test, critical value of F
 $F(\alpha, k_1, k_2)$

F.INV.RT(α, df_1, df_2)

9 References

- ¹ H. Pottel, Statistical flaws in Excel,
www.pucrs.br/famat/viali/tic_literatura/artigos/planilhas/pottel.pdf.
- ² J.G. Eisenhauer, Teaching Statistics, 25 (2003) 76.
- ³ B.D. McCullough, David A. Heiser, Comput. Stat. Data Anal., 52 (2008) 4570.
- ⁴ International Encyclopedia of Statistical Science, M. Lovric (Ed.), Springer, Heidelberg, 2011.
- ⁵ P.R. Bevington, D.K. Robinson, Data reduction and error analysis for physical sciences, McGraw-Hill, Boston, 2003.
- ⁶ P.D. Lark, B.R. Craven, R.C.L. Bosworth, The Handling of Chemical Data, Pergamon Press, Oxford, 1968.
- ⁷ S. Brandt, Data Analysis, Statistical and Computational Methods for Scientists and Engineers, Springer, 2014.
- ⁸ N.R. Draper, H. Smith, Applied regression analysis, Wiley, New York, 1998.
- ⁹ S.L.R. Ellison, V.J. Barwick, T.J. Duguid Farrant, Practical Statistics for the Analytical Scientist A Bench Guide, RSC Publishing, LGC Limited, 2009.
- ¹⁰ F.L. Ramsey, D.W. Schafer, The Statistical Sleuth A Course in Methods of Data Analysis, Brooks/Cole, Cengage Learning, 2013.
- ¹¹ Z.B. Alfassi, Z. Boger, Y. Ronen, Statistical Treatment of Analytical Data, Blackwell Science, 2005.
- ¹² G. Baillargeon, Probabilité, statistique et techniques de régression, Les Éditions SMG, 1989.
- ¹³ N. Gilbert, J.G. Savard, Statistiques, Édition Études Vivantes, 1992.
- ¹⁴ D. Livingstone, Data analysis for chemists, Oxford University Press, Oxford, 1995.
- ¹⁵ J.N. Miller, J.C. Miller, Statistics and Chemometrics for Analytical Chemistry, Prentice Hall, 2000.
- ¹⁶ R.G. Brereton, Chemometrics. Data analysis for the laboratory and chemical plant, Wiley, Chichester, 2003.
- ¹⁷ P.C. Meier, R.E. Zünd, Statistical Methods in Analytical Chemistry, Wiley-Interscience, New York, 2000.
- ¹⁸ D.B. Hibbert, J.J. Gooding, Data Analysis for Chemistry: An Introductory Guide for Students and Laboratory Scientists, Oxford University Press, 2005.
- ¹⁹ Comprehensive Chemometrics: Chemical and Biochemical Data Analysis, R. Tauler, B. Walczak, S.D. Brown, Edts., Elsevier, 2009.
- ²⁰ A.H. Rosenfeld, H.A. Barbero-Galtieri, W.J. Podolski, L.R. Price, P. Söding, Ch.G. Wohl, M. Roos, W.J. Willis, Rev. Mod. Phys. 39 (1967) 1.
- ²¹ Evaluation of measurement data - Guide to the expression of uncertainty in measurement, JCGM 100:2008,
https://ncc.nesdis.noaa.gov/documents/documentation/JCGM_100_2008_E.pdf.
- ²² A. Williams, Accred. Qual. Assur. 4 (1999) 14.
- ²³ E. Bernal in Advances in Chromatography, X. Guo, Ed., InTech, 2014, p. 57.
- ²⁴ D.A. Skoog, F.J. Holler, T.A. Nieman, Principles of Instrumental Analysis, Saunders, Philadelphia, 1998.
- ²⁵ J.A. Irving, T.I. Quickenden, J. Chem. Educ., 60 (1983) 711.
- ²⁶ J. Tellinghuisen, J. Chem. Educ., 95 (2018) 970.

-
- 27 P.D. Lark, B.R. Craven, R.C.L. Bosworth, *The Handling of Chemical Data*, Pergamon Press, Oxford, 1968
- 28 K.N. Carter, Jr., D.M. Scott, J.K. Salmon, G.S. Zarcone, *Anal. Chem.*, 63 (1991) 1270.
- 29 L.M. Schwartz, R.I. Gelb, *Anal. Chem.*, 56 (1984) 1487.
- 30 G.E. Forsythe, M.A. Malcolm, C.B. Moler, *Computer Methods for Mathematical Computations*, Prentice Hall, Englewood Cliffs, N.J., 1977.
- 31 P.C. Hansen, S. Christiansen, *J. Comp. Appl. Math.*, 12-13 (1985) 341.
- 32 P.C. Hansen, *BIT* 27 (1987) 534.
- 33 W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, *Numerical Recipes in Fortran 77*, Cambridge University Press, New York, 1992.
- 34 H.W. Moody, *J. Chem. Educ.*, 59 (1982) 291.
- 35 D.C. Harris, *J. Chem. Educ.*, 75 (1998) 75.
- 36 R. de Levie, *Advanced Excel for scientific data analysis*, 3rd Edition, Atlantic Academic, 2012.
- 37 M. Otto, *Chemometrics*, Wiley-VCH, Weinheim, 2007.
- 38 R.B. Dean, W.J. Dixon, *Anal. Chem.*, 23 (1951) 636.
- 39 <http://www.statisticshowto.com/dixons-q-test/>
- 40 F.E. Grubbs, G. Beck, *Technometrics*, 14 (1972) 847.
- 41 Norm ISO 16269-4-2010, Statistical interpretation of data - Part 4: Detection and treatment of outliers.
- 42 ASTM E178, Standard Practice for Dealing With Outlying Observations.
- 43 Daniel C. Harris, *Quantitative Chemical Analysis*, Eighth Edition, W. H. Freeman and Company, New York, 2010.
- 44 Michael Thompson, Philip J Lowthian, *Notes on Statistics and Data Quality for Analytical Chemists*, Imperial College Press, London, 2011.
- 45 R.G. Brereton, *Applied Chemometrics for Scientists*, Wiley, Chichester, 2007.
- 46 S.L.R. Ellison, V.J. Barwick, T.J. Duguid Farrant, *Practical Statistics for the Analytical Scientists. A Bench Guide*, RCPublishing, 2nd Edition, 2009.
- 47 H.A. David, H.O. Hartley, E.S. Pearson, *Biometrika*, 41 (1954) 482.
- 48 G. Rodriguez, *Lecture Notes*, 2007, Princeton University, Chapter 2.9 Regression diagnostics, <http://data.princeton.edu/wws509/notes/c2s9.html>.
- 49 P.J. Rousseeuw, A.M. Leroy, *Robust Regression and Outlier Detection*, Wiley, New York, 1987.
- 50 M.H. Kutner, C.J. Nachtsheim, J. Neter, W. Li, *Applied linear statistical models*, McGraw-Hill, 2005.
- 51 R.D. Cook, *Technometrics*, 19 (1977) 15.
- 52 J.O. Rawlings, G. Sasthy Pantula, D.A. Dickey, *Applied Regression Analysis: A Research Tool*, 2nd Ed., Springer, New York, 1998.
- 53 H. Chatterjee, A.S. Hadi, B. Price, *Regression Analysis by Example*, 3rd edition, Wiley, New York, (2000).
- 54 B. McDonald, *Res. Lett. Inf. Math. Sci.*, 3 (2002) 127.
- 55 Spider Financial, <http://www.spiderfinancial.com/support/documentation/numxl/users-guide/factor-analysis/regression-analysis-mlr/influential-data-analysis>.
- 56 T.E. Smith, http://www.seas.upenn.edu/%7Eese302/extra_mtls/REGRESSION_OUTLIERS.pdf.

-
- 57 A.J. Schwab,
http://www.utexas.edu/courses/schwab/sw388r7/Tutorials/IllustrationofRegressionAnalysis_doc_html/052_Identifying_Influential_Cases_Cook_s_Distance.html.
- 58 H.A. Gordon, Errors in Computer Packages. Least Squares Regression Through the Origin, Statistician, 30 (1981) 23.
- 59 A.M. Awad, Properties of the Akaike information criterion, Microelectron. Reliab., 36 (1996) 457.
- 60 K.P. Burnham, D.R. Anderson, Model Selection and Multimodel Inference, A Practical Information-Theoretic Approach, Second Edition, Springer, 2002.
- 61 S. Konishi, G. Kitagawa, Information Criteria and Statistical Modeling, Springer, 2008.
- 62 E.J. Wagenmakers, S. Farrell, AIC model selection using Akaike weights, Psychonomic Bull. Rev., 11 (2004) 192.
- 63 M. Ingdal, R. Johnsen, D.A. Harrington, The Akaike information criterion in weighted regression of immittance data, Electrochim. Acta, 317 (2019) 648.
- 64 H. Akima, J. ACM, 17 (1970) 589.
- 65 C.A. Micchelli, T.J. Rivlin, S. Winograd, *Numerische Mathematik*, 26 (1970) 279.
- 66 L.D. Irvine, S.P. Marin, P. W. Smith, *Constructive Approximation*, 2 (1986) 129.
- 67 A. Savitzky, M.J.E. Golay, Anal. Chem., 36 (1964) 1627.
- 68 W.H. Press, S.A. Teukolsky, W.T. Vetterling, B.P. Flannery, *Numerical Recipes in Fortran 77, The art of scientific computing*, vol. 1, Cambridge University Press, 1997.
- 69 E.O. Brigham, The Fast Fourier Transform, Prentice-Hall, 1974.
- 70 Ken'iti Kido, Digital Fourier Analysis: Fundamentals, Springer 2015; Digital Fourier Analysis: Advanced Techniques, Springer, 2015.
- 71 A. Lasia, Electrochemical impedance spectroscopy and its applications, Springer, 2014.
- 72 J.W. Hayes, D.E. Glover, D.E. Smith, M.W. Overton, Anal. Chem., 45 (1973) 277.
- 73 Carl de Boor, *A Practical Guide to Splines*, Springer-Verlag, 1978.
- 74 W.S. Cleveland, J. Am. Stat. Assoc., 74 (368) (1979) 829.
- 75 W.S. Cleveland, S.J. Devlin, J. Am Stat. Assoc., 83 (403) (1988) 596.
- 76 A. K. Kaw, E.E. Kalu, D. Nguyen, Numerical Methods and Applications, 2008,
http://nm.mathforcollege.com/topics/textbook_index.html.