CHEMOMETRICS FOR ANALYTICAL CHEMISTRY

Volume 1: PC-Aided Statistical Data Analysis

MILAN MELOUN
Department of Analytical Chemistry, University of Chemical Technology, Pardubice, Czechoslovakia

JIRI MILITKÝ
Department of Textile Materials, College of Mechanical and Textile Design, Liberec, Czechoslovakia

MICHELE FORINA
Faculty of Pharmacy, University of Genoa, Italy

Translation Editor:
DR M. MASSON
Department of Chemistry, University of Aberdeen

ELLIS HORWOOD
NEW YORK LONDON TORONTO SYDNEY TOKYO SINGAPORE

Glossary

FREQUENTLY USED STATISTICAL TERMS AND DIAGNOSTIC TOOLS OF DATA PROCESSING IN CHEMOMETRICS

The following statistical terms and diagnostic tools are frequently used in computer-assisted data treatment in chemometrics. They are explained in this textbook in detail, and will be used throughout the book.

Absolute error. The difference between the true and measured values of the signal.
Abstract factors. The first (significant) eigenvectors of the data matrix, generally without a scaling procedure.
Accuracy. The agreement between an experimentally determined value and the accepted reference value. In chemical work this term is frequently used to express freedom from bias, but in other fields it assumes a broader meaning as a joint index of precision and bias.
Additive systematic error. Arises from imperfect or incorrect zero-point reading on an instrument.
AFA reproduced data matrix. The data matrix reproduced without the noise from the non-significant components or abstract factors.
Analytical signal. Physical quantity measured instrumentally, and characterized by the signal magnitude and the signal position.
Autoscaling. Column centring followed by column standardization.
Bar chart (G13). A graphical representation of a grouped frequency distribution in which all the elements in a given group are represented by the same value.
Barycentre. See Centroid.
Batch of data. A set of values of similar meaning, such as a set of measurements.

Bias. A constant or systematic deviation as opposed to a random error. It appears as a persistent plus or minus deviation of the method average from the accepted reference value.

Bootstrap method. Distribution-free statistical technique based on generation of Bootstrap samples and their analysis.

Box-and-whisker plot (G4). A box from the lower to upper quartile, barred at the median, with 'whisker' to (a) the extremes, (b) the innermost identified values, (c) adjacent values.

Centroid. A row vector with the values of the average of the variables.

Cluster. Group of objects (or variables) with relatively high similarity.

Clustering. A procedure to detect well-separated clusters of objects (or variables).

Coefficient of variation (C.V.) or **Relative standard deviation** is a measure of relative variability; defined as the standard deviation divided by the mean and multiplied by 100 [%].

Column centring. The subtraction of the mean of the variable from each column of a data matrix. The transformed variable has a mean of zero.

Column profiles. The division of the quantities in a data matrix by the column sum.

Column range scaling. Each element of a data matrix has subtracted from it the minimum of the corresponding variable, and is divided by the range of the variable. The transformed variable has a minimum value of zero and a maximum of one.

Column simplicity. The variance of the squares of the elements in a column of a matrix.

Column standardization. The division of each element in a data matrix by the column standard deviation (standard deviation of the variable).

Communality of a variable. The sum of the square of the loadings. It is 1 when the sum is computed over all the eigenvectors in the matrix of loadings.

Components. The eigenvectors of the data matrix, generally after autoscaling or range scaling.

Confidence interval or **confidence limits**. These terms (usually expressed as a percentage; $P\%$ confidence limit), refer to that interval or range of values which will with a probability of $P/100$ include the population parameter.

Confidence level. This term (usually expressed as a percentage; e.g., 95% confidence level), is commonly used in establishing the probability of precision statements, and means that there are, for example, 95 in 100 chances of being correct, and 5 in 100 chances of being wrong, in predicting that the population parameter will fall within the specified limits or range.

Confirmatory data analysis. Stresses evaluating the available evidence.

Correlation coefficient. A measure of the degree of linear dependence between two variables. It is equal to the covariance divided by the square root of the product of the variances of the two variables. It can take values between -1 and $+1$.

Covariance. A measure of the joint dispersion of two variables in the same series of objects. It is the sum of the products of the individual deviations of the first variable from its mean and the deviation of the second variable from its mean, divided by the number of degrees of freedom (generally the number of objects minus one).

D-value. A letter value at depth $(1 + d)/2$, where d is the integer part of the depth of an E value; crudely, a sixteenth.

Data matrix. A rectangular organization of quantities in rows and columns. Each row is a data vector (see Object) and each column is a column vector (see Variable).
Degrees of freedom (D.F.) Generally equal to the number of observations minus the number of parameters calculated from it. For example, variance is the average squared deviation about the mean of n observations. The variance can be calculated only after the mean has been calculated and this uses up one degree of freedom. No other parameters are necessary to estimate the variance. Hence, the estimate of the variance has $(n - 1)$ degrees of freedom.
Depth. Lesser of upward rank and downward rank.
Differential quantile plot (G9). A comparison of the actual distribution with the normal one on the basis of quantile comparison.
Distribution function. A function describing the statistical behaviour of a random variable. On the y-axis are the cumulative probabilities corresponding to the value on the x-axis.
Dot diagram (G2). Univariate projection of the quantile plot onto the x-axis.
Double centring. Column centring followed by row centring, or row centring followed by column centring.
E-value. A letter value at depth $(1 + e)/2$, where e is the integer part of the depth of a quartile; crudely an eighth, sometimes called one.
Eigenvalues. The variances in the space of the eigenvectors.
Eigenvector rotation. An orthogonal rotation (generally around the centroid) in Q-space (or R-space) that produces new variables, by linear combination of the original variables. In the space of the new variables (eigenvectors) the variance–covariance matrix is a diagonal matrix (covariances are zero). The eigenvectors are numbered in order of decreasing variance (eigenvalue).
Ensemble. See universe or population.
Error. Any deviation of an observed value from the true value. When expressed as a percentage of the value measured, it is called a relative error. Types of error include instrumental errors, methodology errors, theoretical errors and data-treatment errors. Another criterion classifies errors as (1) systematic errors, (2) random errors, (3) personal errors, and (4) gross errors.
Error of the first kind. In hypothesis testing, this error is caused by false rejection of the hypothesis when it is true.
Error of the second kind. In hypothesis testing this error is caused by not rejecting the hypothesis when it is false.
Examining for independence of sample elements. Confirmatory data analysis tests for poor stability of measurement equipment, inconstancy of the conditions of measurements, neglect of some measurement factors, and false and non-random choice of measurements in a sample.
Examining for minimum sample size. Confirmatory data analysis searches for minimum samples size.
Examining for homogeneity of sample. Confirmatory data analysis searches for outliers or clusters in sample.
Examining for normality of sample distribution. Confirmatory data analysis tests the normality of the sample distribution.

Exploratory data analysis. This provides the first contact with the data and serves to uncover unexpected departures from familiar models. It emphasizes flexible searching for clues and evidence.

Exponential distribution. This is a distribution of a random variable bounded on one side. There are one-parameter exponential distributions, and two-parameter exponential distributions.

Extreme. The highest or lowest value of a batch. Since it has a depth of 1, it is often labelled "1".

Fisher weights. Weights for classification problems, proportional to the ratio between the intercategory variance and the intracategory variance.

Frequency polygon (G13). A graphical representation of grouped frequency distribution, with points joined by straight lines to form an open polygon.

Frequency ratio plot (G17). Used for distinguishing among various types of discrete distributions.

Global centring. The subtraction from all the elements in a data matrix of the generalized average of the elements.

Global standardization. The division of each element in a data matrix by the overall standard deviation (standard deviation of the elements of the matrix around the generalized average of the matrix).

H-value. A letter value, a quartile, a value with depth $(1 + h)$, where h is the integer part of the depth of the median; crudely, a quartile.

Heterogeneity of a cluster. The sum of the squares of the distance of the objects from the barycentre of the cluster.

Hinge. A letter value corresponding to a quartile.

Histogram (G13). The oldest classical representation of grouped frequency distributions.

Hypothesis (or **statistical hypothesis**). An assumption made concerning a parameter to provide the basis for a statistical test; usually expressed as a null hypothesis H_0 or as an alternative hypothesis H_A. It is a statement about the population distribution of some random variable.

Interval estimation. The interval which contains, with some pre-selected probability, the population parameter (obviously of location, scale or shape).

Interquantile range. The quantile estimate of spread.

Jack-knife method. Distribution-free technique which uses the pseudovalues y_i defined by $y_i = \mathrm{n} \times \hat{\theta} - (n - 1) \times \hat{\theta}_{(i)}$ where $\hat{\theta}_{(i)}$ is the estimate used from all elements except the ith one. For large samples the pseudovalues y are assumed to be approximately normally distributed.

Jittered-dot diagram. A univariate projection of a quantile plot.

Kernel estimation of probability density function (G12). A smooth nonparametric estimate of probability density, based on kernel function.

Kurtosis. A measure of shape characterizing the peakedness of a distribution and its tail length.

Kurtosis plot (G8). This indicates the asymmetry of a distribution.

Laplace Distribution. A symmetric two-tailed exponential distribution with higher kurtosis than normal.

Letter value. One of the values tagged . . . , M, H, (or F) E, D, C, B, A, Z, Y, X, . . . , where the depth of each letter is half the depth of the next letter value, starting from M for median and ending with 1 (the depth of an extreme).

Limiting error of an instrument. The highest possible error which, under given experimental conditions, is not overcome by any other random errors.

Loadings. The scalars in the M-square (M is number of variables) orthogonal rotation matrix used in eigenvector rotation. These are the cosines of the angles between the original variables and the eigenvectors.

Location. The characteristic of a random distribution that measures the position of the distribution on the x-axis.

Log–normal distribution. The logarithmic transformation of the normal distribution. A two-parameter log-normal distribution and a three-parameter log–normal distribution can be distinguished.

M-estimator. This robust estimator represents the maximum likelihood estimate of a population parameter for some special distributions.

M-value. A letter value, the median.

Mahalanobis distance. A statistical distance taking into account the variance of each variable and the correlation coefficients. In the case of a single variable, it is the square of the distance (between two objects, or between an object and the centroid) divided by the variance.

Mean. The arithmetic average of a series of measurements.

Mean error. The limiting error of measurement for the 50% probability level.

Median of sample. For samples with odd numbers of items, the median is the middle item when the items are arranged in order of magnitude. For samples with an even number of items, the median is the arithmetic average of the two middle items when they are arranged in order of magnitude. The median is an estimate of the location and has a depth of $(1 + n)/2$, where there are n values in the batch.

Midsum. The most efficient estimate of location for the rectangular distribution.

Midsum plot (G6). Indicates symmetry of sample distribution.

Mode or **Modus**. The local maximum on the probability density function.

Modified Poisson plot (G19). A plot for examination of the Poisson distribution for data sample.

Modified rankit plot (G16). A plot for examination of the normality of a sample distribution.

Monte-Carlo simulation method for error propagation. The method for calculation of propagation of errors when the probability distribution is replaced by the Monte-Carlo simulation.

Multiplicative systematic error. An error which depends on the value of measured signal.

Normal distribution. The most common symmetrical distribution with probability density in the form of a Gaussian curve. Normality is assumed in classical statistical methods.

Notched box-and-whisker plot (G5). An analogue of the box-and-whisker plot G4, that allows examination the variability of the median.

5-number summary. Values of extremes, hinges, and median; values of 1HMH1.

7-number summary. Values of 1EHMHE1.

9-number summary. Values of 1DEHMHED1.

Object. A sample (or molecule, individual, reaction, process, etc.) described by one or more measured or computed quantities (variables). These quantities constitute a row vector (or data vector).

Observation. A value that was measured experimentally.

Order statistic. When the sample values $x_1, \ldots, x_n$ are sorted in order of ascending magnitude, the order statistics $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$ are obtained. The order statistic $x_{(i)}$ is the ith smallest sample value.

Outlier. A value in a set of observations which is so different from the rest that it is considered to be a member of another set or population.

Parameter of a population. The unknown 'true' value for the population distribution, such as the mean and the variance.

Pivot halfsum. For small samples, a robust estimate of parameter of location.

Pivot range. For small samples, a robust estimate of parameter of spread.

Plot of logarithm of likelihood function (G21). Searches for suitable power data transformation.

Poisson distribution. This discrete distribution relates to the number of events that occur per given interval of time or space when the events occur randomly in time or space at a certain average rate.

Point estimation. A single-value estimate of a population parameter computed from a sample.

Population. Same as universe or ensemble.

Power of statistical test. Represents the probability of making a correct decision when the hypothesis is actually wrong.

Precision. The degree of agreement between repeated measurements of the same property. There are various types of precision such as:

Duplicability: the agreement between duplicate or other multiple determinations performed by the same analyst at essentially the same time.

Repeatability: the precision of a method expressed as the agreement attainable between independent determinations performed by a single analyst using the same apparatus and techniques on more than one day.

Reproducibility: the precision of a method expressed as the agreement attainable between determinations performed in different laboratories.

Principal components (PC). The components with high eigenvalues. These account for the useful information and predictive ability.

Probability interval for random error. The interval in which the random error ε lies with a preselected probability.

Poisson plot (G18). A plot to distinguish whether the actual distribution is of Poisson nature.

Propagation of absolute errors. An approximate expression for calculation of the overall variance when various operations are performed and each is loaded by some absolute error.

Propagation of relative errors. An approximate expression for calculation of relative overall variance when various operations are performed and each is loaded by some relative error.

Q-space. An M-dimensional space (M is number of variables) described by M cartesian co-ordinates. Each point in this space describes an object.

Quantile or **percentile of sample**. The value below which P_i% of the sample lies. The order statistic $x_{(i)}$ is quantile $\tilde{x}_{P_i}$ where P_i is the order probability $P_i = i/(n + 1)$.

Quantile plot (G1). This allows determination of the shape of a sample distribution, which can be symmetrical, or skewed to higher or lower values.

Quantile–box plot (G10). A universal tool for examination of statistical features of data.

Quantile–quantile plot (Q–Q plot, G14). This allows comparison of the sample distribution described by the empirical quantile function Q_E with the selected theoretical quantile function Q_T.

R-space. An I-dimensional space (I is number of objects) described by I cartesian co-ordinates. Each point in this space describes a variable.

Random error. The chance variation encountered in all experimental work despite the closest possible control of all variables. It is characterized by the random occurrence of both positive and negative deviations from the true value.

Range. The absolute value of the algebraic difference between the highest and lowest values in a set of data.

Range scaling. See Column range scaling.

Rank. Ordinal number corresponding to the position of a value when values are sorting into decreasing (rank up) or increasing (rank down) order.

Rank-and-depth technique. This is used for estimation of letter value.

Rankit plot or **normal-probability plot (G15)**. This allows comparison of the actual sample distribution with the normal one, by use of quantile functions.

Real factors. These contrast with the abstract, mathematical, factors. They are found by linear combination of the original variables, and they have physical significance. Real factors can be obtained by orthogonal or oblique rotation in the space of the abstract factors.

Rectangular distribution. A symmetrical distribution of random variable with a lower kurtosis than the normal distribution. It describes e.g. a distribution of errors caused by rounding-off to k decimal places.

Re-expressed statistics. After calculation of statistics for transformed data, these statistics are re-expressed to refer to the original data. As a rough approximation, the re-expressed statistics represent a straight reverse transformation $\bar{x}_R = g^{-1}(y)$. More rigorously, re-expressed statistics can be calculated by the Taylor series of the function $y = g(x)$.

Residual. The difference between an actual data point, and the value calculated from the statistical model (fit).

Response. Variable, or the value of a variable, as it is influenced by (or associated with) certain factors or circumstances.

Relative error. See coefficient of variation.

Relative standard deviation. See coefficient of variation.

Rootogram (G13). The square-root re-expression of a histogram.

Row centring. The subtraction from the rows of a data matrix of the row mean.
Row profiles. The transformation of a data vector by division by the row sum. The row sum of the transformed quantities is 1.
Row simplicity. The variance of the squares of the elements in a row of a matrix.
Row standardization. The division of each element in a data matrix by the row standard deviation (standard deviation of the object).
Sampling. Impartial selection of sample to ensure that it is representative.
Scores. The co-ordinates in the space of the eigenvectors.
Selection nomogram (G20). A diagnostic tool which searches for a suitable data transformation.
Shape parameters. Measures of shape peculiarities of probability density functions.
Similarity. An inverse measure of distance. The similarity between two objects is given by one minus the ratio between their distance and the maximum distance among the objects in the data matrix. The similarity between two variables is defined similarly.
Skewness. Measure of shape characterizing symmetry or asymmetry of the distribution.
Sorting. The process of putting a set of numbers into order.
Spread or **scale** or **variability** or **scatter**. The degree of variability of the measured quantity.
Square row profiles. The division of the quantities in a data matrix by the square root of the row sum of squares. The row sum of squares of the transformed quantities is equal to one.
Statistic or **statistical characteristic**. An estimate of a parameter.
Standard deviation. A measure of the dispersion of a series of results around their mean, expressed as the square root of the variance.
Stem-and-leaf display (G11). A generalized two-digit display, in which the left hand portion of the values displayed is given by a stem value, and the right hand portion makes up a leaf.
Test criterion or **significance criterion**. Hypothesis testing consists of comparing some statistical measures which are called test or significance criteria. These measures are deduced from a data sample and are compared with the values of these criteria that would apply on the assumption that a given hypothesis is correct.
Tolerance interval for random error. The interval that contains $P\%$ of all errors with statistical certainty $(1 - \alpha)$.
Transformation of data. Correct data transformation leads to a symmetrical distribution, stabilizes the variance and makes the distribution closer to normal.
Trimmed mean. A simple, efficient, robust estimate of location defined with the use of order statistics. We may distinguish the symmetric trimmed mean, and for asymmetric and strongly skewed distributions, the asymmetric trimmed mean.
Two-points estimate. The method of calculation of propagation of errors used when the probability distribution is replaced by the two-points distribution with the same mean and variance.
Two-way analysis of variances (ANOVA). Analysis of one response according to two kinds of circumstances.
Universe or **population** or **ensemble**. Statistical reasoning employs the concept of a sample of observations drawn at random from a universe of all possible events. A

universe is generally characterized by a probability function with one or more parameters, such as a mean and a variance. A finite sample can give only estimates of these parameters.

Variable. A measured or computed quantity. The series of values of the same variable measured for several objects constitutes a column vector.

Variance. A measure of the dispersion (or spread) of a random variable around a mean. Given by the sum of the squares of the individual deviations from the mean of the results, divided by the number of results minus one.

Variance about the origin. A measure of the dispersion of a series of results around the origin (the value zero). It is the sum of the squares of the quantities in the series divided by the number of results.

Variance–covariance matrix. An M-square matrix (M is the number of variables). The scalar in the column m' of row m (on the diagonal of the matrix) is the variance of the mth variable. The scalar in the column m of the row m is the covariance of the variables m and m'.

Weights. The multiplying coefficients for the autoscaled (generally) variables, required to give to each variable an appropriate importance. Weights can be obtained from the error in the measurement of the variance, or from the ability of the variable to solve classification or correlation problems.

1

Errors in instrumental measurements

Observations and experiments constitute the basis of all natural science. Observations and experiments that provide numbers—the results of measurements—are of special significance. A correct analysis of these observations leads to a theoretical interpretation of the results and to the final goal of natural science, namely, the establishment of laws that make possible the prediction of the future behaviour of the phenomena. The analysis of observations refers to operations on numbers that are obtained directly from observations. However, to develop a theory on the basis of computation of quantities that are not directly observed but that are derived from the analysis of observations, it is necessary to use various mathematical devices.

Chemistry, physics and other sciences deal with many quantities, some of which are purely physical (time, mass, volume, temperature, electrical parameters). Others are physical-chemical (pH, potential, viscosity), and some are connected with the chemical composition of the system. A description of the chemical composition requires the determination of a number of quantitative characteristics, the concentrations of the components.

A specialized analytical chemistry subsystem is necessary in every branch of chemical research, to obtain data on chemical compositions. This subsystem consists of the following operations:

(1) Sample preparation and treatment, including operations that are mostly carried out outside the laboratory, to ensure that the sample is representative.

(2) Sample preparation for the measurement, including sample decomposition, separation operations, procedures defining the chemical substance, and also the actual instrumental measurement. In instrumental methods, there is an attempt to combine operations necessary for forming a measurable quantity with the actual measuring operation.

(3) The measurement corresponds to monitoring of the analytical signal (or often only the signal or measurement).

The analytical signal usually corresponds to a physical quantity and is measured instrumentally. Two characteristics of the signal can be distinguished; the signal magnitude (e. g., radiation intensity at a given wavelength) and the signal position (e.g., wavelength). In identification analysis, the signal position is decisive, and its determination assumes a certain minimal size; in quantitative analysis, on the other hand, the signal magnitude is decisive.

The magnitude of the signal, S, is in general a function of the concentration of the test component, the analyte, A, M, L, H, ... etc., and also of variable x_i which can be the reagent volumes, temperature, etc.; $S = f(A, M, L, H; x_i)$. The analytical method is characterized by its sensitivity, defined as the derivative of the signal with respect to the concentration of the test component

$$\left[\frac{\delta S}{\delta A}\right]_{M,L,H,x_i}$$

where parameters M, L, H, and variables x_i are kept constant so that the signal can be considered to be a function of a single variable A. The corresponding dependence $S = f(A)$ is then termed the *single calibration function.*

Since all measurements of analytical signals in the chemical laboratory contain errors from a range of origins, the results of calculation with the numbers corresponding to these also contain errors. It is very important to be able to estimate both the errors incurred in making the measurements, and the errors resulting from operations on those measurements. Both the signal measurements and the calculations must be organised in such a way as to minimize the errors in the results.

Among the errors of analytical signals a conspicuous place is occupied by random errors; that is, errors with values that cannot be estimated before the signal measurements. We might also note that they cannot be evaluated even after observations, since the presence of random errors makes it impossible for us to determine the exact value of the analytical signal measured. The analysis of analytical signals containing random errors utilizes the theory of probability, which is also necessary in statistical work. The measured analytical signal is treated according to the following rules.

(1) There is always a limit to the precision of measurements. Even results obtained with the best instrumental precision and experimental care are not precise, but have approximate values.

(2) Only some of physical quantities can be measured directly, i.e., length, mass, and time intervals. Most other quantities are measured indirectly. The relationship between the analytical signal and the analyte concentration is derived from a mathematical relationship and yields a mathematical model. The measured results are approximate values, so the calculated analyte content must also be an approximate value.

(3) The arithmetic mean of repeated measurements corresponds to an estimate of the true value of the measured quantity; however, the true value remains unknown.

(4) Finding empirical model functions is the first step in finding more basic relationships. In addition to the test quantities, the model also contains unknown parameters that are estimated from the experimental data by regression methods.

This course of chemometrics begins with basic principles of statistical data treatment and theory of errors [1], so that the student may become familiar with the system of notation and terminology.

1.1 TYPES OF MEASUREMENT ERROR

The results of laboratory measurements of analytical signals are always approximate. Therefore, all measurements contain errors of various origin. It is customary to classify these errors according to their source in the measurement process into four types.

(1) *Instrumental errors* are caused by the construction of instrument used and are usually known and specified by the instrument manufacturer.

(2) *Methodology errors* are caused by use of an inappropriate method. Examples are inappropriate data acquisition, interference by some external effects, faulty strategy of experimentation, etc.

(3) *Theoretical errors* are caused by the use of a false principle of measurement or inappropriate physical model, etc.

(4) *Data treatment errors* are caused by inappropriate numerical methods of data evaluation or statistical data treatment.

Another way to classify errors is according to their effect on the evaluation of the results. Again, there are four types.

(1) *Systematic errors.* The most important systematic errors are instrumental errors which are a result of incorrect instrumental settings, constant distortions, insufficient chemical purity, and imperfect standardization and calibration.

Additive systematic errors (fixed bias errors) arise from simple instrumental errors such as taking an imperfect or wrong zero point reading. As a result of such an error, all signal measurements are distorted by the same amount, which may be positive or negative.

The multiplicative systematic error (relative bias, an error of sensitivity) depends in some definite way on other quantities, in particular, on the measured signal itself. Systematic instrumental errors must be investigated and eliminated from the results of measurements.

(2) *Random errors.* Experiment has shown that successive measurements of a single fixed quantity, made with the greatest possible care, give different numerical values even after all the known systematic errors are allowed for. This shows that many causes have an effect on the results of measurements—causes for which we cannot make allowance. A whole series of similar random causes may produce deviations from an exact value. In each case, the deviation is slight: otherwise it would be noticed and investigated. However, the total effects of all these causes can yield significant

deviations. The theory of errors usually refers to the theory of random errors. For the construction of such a theory, the nature of random errors suggests the application of probability theory.

(3) Personal errors. The results of measurements depend to some degree on the physical peculiarities of the observer (under otherwise equal conditions). For example, in recording the instant of a phenomenon, one observer may regularly notice a phenomenon somewhat sooner than another. Repeated study of the personal errors of different observations has shown that these errors can be both systematic and random. Some amount of personal error is associated with an observer, and this error should be considered to be systematic and taken into consideration in the analysis of observations. Often observations can be made to determine the personal error, and the results of these observations are analysed in much the same way as for random errors, in order to obtain their average value.

(4) Gross errors. In the analysis of observations, we need to allow for the possibility of blunders or external influences that cause completely inaccurate results. One of the simplest of these will be for an observer to read 20.0 and write down 30.0, for example. The presence of gross errors is detected by the fact that in a succession of comparatively close results only one or only a few values will differ appreciably from the general level of values: that is, these results stand out. If the discrepancy is great enough for us to be sure that it is result of an error, the signal measurement can be disregarded.

Let us suppose that a certain signal quantity has a definite numerical value μ that remains unchanged during the entire process of signal measurement. Let us suppose also that the repeated measurement made of this signal quantity yields the values x_i, $i = 1, \ldots, n$. The difference between the exact and each actual value of signal

$$\Delta_i = x_i - \mu \tag{1.1}$$

is called the *absolute error*. This definition is convenient in that the concept of an absolute error coincides with the concept of a criterion since $x_i = \mu + \Delta_i$; that is, the absolute error is the number that must be added to the exact value μ in order to obtain the approximate number x_i.

When no gross errors are present in a set of n repeated observations, the average value

$$\bar{\Delta} = \frac{1}{n} \sum_{i=1}^{n} \Delta_i \tag{1.2}$$

represents the systematic error found in observations and the difference $\bar{\Delta} - \Delta_i$ is the estimate of random error.

1.2 PRECISION AND ACCURACY OF INSTRUMENTAL MEASUREMENTS

In instrumental metrology [3] the terms accuracy and precision are considered as characteristics of the measuring process. *Accuracy* refers to the typical 'closeness of n measurement results x_i, (or their average $\bar{x}$) to the true value μ' while *precision* refers to the typical 'closeness of n measurement results x_i, together' for the conceptually large population size n of results that might have been, or could be, obtained. When measurements of the same quantity are repeated, the dispersion of the results can be seen by examining the data. A set of data that shows little variation may be said to have greater precision than a set of data showing larger variation.

It is instructive to distinguish between random errors and systematic errors, and between precision and accuracy, in some cases typical of different measurement processes. Figure 1.1 shows four drawings, each of which depicts a distribution of measurements x_i. In case (a) each of the observations x_i is relatively close to the true value μ (accurate) and each other (precise). In case (b) the individual observations do not have good accuracy, but their mean $\bar{x}$ would be reasonably accurate. The precision is relatively poor as indicated by the wide spread of data. In case (c), none of the observations is relatively close to the true value μ. The observations are close together and it can be said that the measurement process is precise but not accurate. Here the distinction between accuracy and precision is quite clear. In case (d), although one of the observations happens to be near the true value, the accuracy exhibited by most of the individual measurements is relatively poor, the accuracy exhibited by the mean of the set is relatively poor, and the precision of the set is relatively poor.

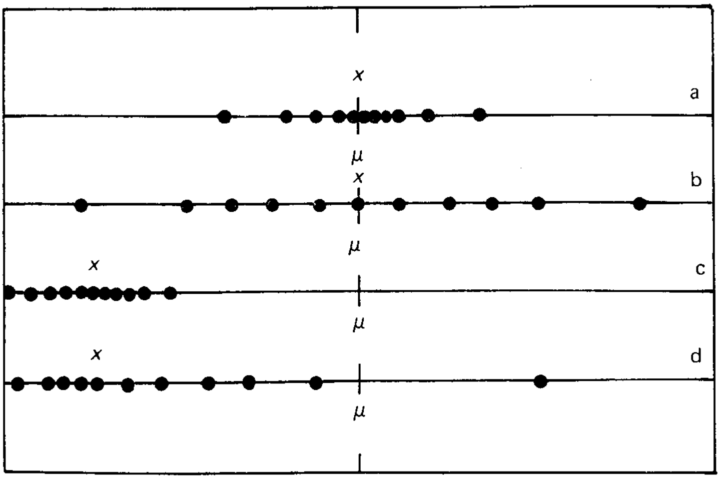


Fig. 1.1—Classification of repeated measurements: (a) accurate and precise, (b) accurate but not precise, (c) not accurate but precise, and (d) not accurate and not precise.

Only in case (a) can the measurement process be called accurate. In case (b) the accuracy can be improved by improving the precision. In case (c) the accuracy of the measurement process can be improved by correcting the systematic error, and in case (d) both factors need to be improved.

1.2.1 Absolute and relative errors

The calibration of an instrument requires that for known values of the input quantity

x_i (e.g., the concentration of hydrogen ions, $[H^+]$, in solution), the corresponding values of the output analytical signal y_i (e.g., the e.m.f. of a glass electrode cell) are measured. Repeated signal measurements are made for each of several values of x_i, so that the dependence $y = f(x)$ can be found. An approximate graphical interpretation shows the uncertainty band in the plot. (Fig. 1.2).

The middle curve in the uncertainty band is called the nominal characteristic y_{nom} (or x_{nom}) and it is usually declared by the manufacturer of instrument. The nominal characteristic y_{nom} (or x_{nom}) differs from the real characteristic y_{real} or x_{real} by the error of the instrument, $\Delta^* = y_{\mathrm{real}} - y_{\mathrm{nom}}$. For any selected y the error due to the instrument is $\Delta = x_{\mathrm{real}} - x_{\mathrm{nom}}$. [We will use errors Δ which are in units of measurement quantities (e.g. pH)].

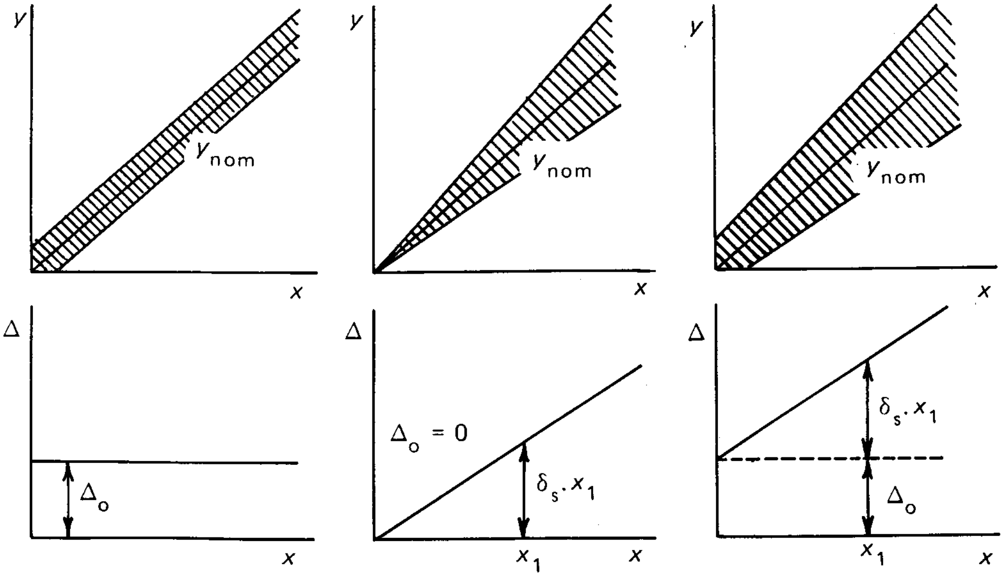


Fig. 1.2--The uncertainty band and three types of instrumental errors: (a) additive, (b) multiplicative, and (c) combined error.

The **absolute error** of signal measurement Δ is not convenient for expressing an instrument's precision because it is given in the specific units of the instrument used. More convenient is the **relative error** defined by

$$\delta = 100 \times \Delta/x \quad [\%] \tag{1.3}$$

or the **reduced relative error** defined by

$$\delta_R = 100 \times \Delta/(x_{\max} - x_{\min}) = 100 \times \Delta/R \quad [\%] \tag{1.4}$$

where R is the range of measurements.

The **limiting error** of an instrument, Δ_0 (absolute) or δ_0 (relative) is, under given experimental conditions, the highest possible error which is not obscured by any other random errors.

The **reduced limiting error** of an instrument, $\delta_{0,\mathrm{R}}$ (relative) for the actual value of measured variable x_i and given experimental conditions is defined by the ratio of the limiting error Δ_0 and the highest value of instrumental range R, $\delta_{0,\mathrm{R}} = \Delta_0/R$. Often, the reduced limiting error is given as a percentage of the instrument range, [see Eq. (1.5)].

From the shape of the uncertainty band, various types of measurement errors can be identified, whence some corrections for their elimination may be suggested.

(a) Absolute measurement errors of an instrument are limited for the whole signal range by the constant limiting error Δ_0 that corresponds to the **additive errors model.** The systematic additive error is a result of incorrect setting of an instrument's zero point. However, nowadays instruments contain an automatic correction for zero and so systematic additive errors rarely appear.

(b) The magnitude of absolute measurement errors grows with the value of input quantity x and for $x = 0$, it is $\Delta = 0$. This is a case of the **multiplicative errors model.** Such errors are called errors of instrument sensitivity. Systematic multiplicative errors are caused by defects of the instrument.

Real instruments have errors that include both types of effects, and are expressed by a nonlinear function $y = f(x)$.

Problem 1.1 *Absolute and relative error of a pH-meter*
A glass electrode for pH measurement has a resistance of 500 MΩ at 25°C, and the input impedence of the pH meter is $2 \times 10^{11}\,\Omega$. Estimate the absolute (Δ_0) and relative (δ_0) error of e.m.f. measurement when the voltage measured is $U = 0.624$ V.
Solution: $U_{\mathrm{corr}} = 0.624(2 \times 10^{11} + 5 \times 10^{8})/(2 \times 10^{11}) = 0.6254$ V; $\Delta_0 = 0.6254 - 0.624 = 0.0016$ V; $\delta_0 = 100 \times 0.0016 / 0.624 = 0.25\%$
Conclusion: The absolute error is 1.6 mV and the relative error 0.25%.

1.2.2 Classification of instrument precision

To express some metrological properties of the instruments the limiting errors of measurement are used. These upper bounds of error will seldom be exceeded, and they also express the class of an instrument precision.

The class of instrument precision is an important precision parameter of an instrument. It describes the highest absolute value of reduced limiting errors found under given experimental conditions for the whole range of the instrument. To express the class of instrument precision for additive, multiplicative and combined errors, the following limiting errors are used:

(1) For the additive errors model, the class of an instrument precision δ_0 is equal to the *limiting reduced relative error* $\delta_{0,\mathrm{R}}$ defined by

$$\delta_{0,\mathrm{R}} = 100 \times \Delta_0/(x_{\max} - x_{\min}) = 100 \times \Delta_0/R \quad [\%] \tag{1.5}$$

where R is the instrument range. The relative error δ decreases hyperbolically with increasing value of x. The sensitivity limit x_c is the input value x for which the limiting absolute error Δ_0 is equal to x_c, i.e. $\Delta_0 = x_c$ or $\delta(x_c) = 100\%$. When the class of instrument precision δ_0 and the range R are known, the sensitivity limit can instead be calculated from

$$x_c = \delta_0 \times R/100 \tag{1.6}$$

To ensure that the relative error of the instrument is sufficiently low, the lower limit of the working interval x_s is defined such that the relative error is kept equal to p (%), usually 4% or 10%. The lower limit of the working interval is defined as

$$x_s = 100\,\Delta_0/p = 100\,x_c/p \tag{1.7}$$

When instrument errors are additive, the range of use is limited to the region of low values of the input quantity x.

(2) In the case of the multiplicative errors model, the class of instrument precision δ_s is expressed by the relative error of sensitivity calculated by

$$\delta_s = 100 \times \Delta_0/x \quad [\%] \tag{1.8}$$

and reaches a constant value in a limited range of instrument scale; this will be declared by the instrument manufacturer.

(3) In the case of the combined errors model, the absolute error Δ may be written as a sum of additive Δ_0 and multiplicative $\delta_s x$ parts by the expression

$$\Delta = \Delta_0 + \delta_s x \tag{1.9}$$

The combined uncertainty band is then formed by addition of the two separate uncertainty bands. The limiting reduced relative error expressed by

$$\delta_{0,\mathrm{R}} = \delta_0 + \delta_s \times x/R \tag{1.10}$$

grows monotonically with increasing x. The growth of $\delta_{0,\mathrm{R}}$ starts later for larger values of the ratio δ_s/δ_0. To express the level of instrument precision in this case two quantities are used:

(a) the reduced relative error δ_0,
(b) the relative error at the upper limit of the measurement scale δ_K expressed by

$$\delta_\mathrm{K} = \delta_0 + \delta_s \tag{1.11}$$

Instruments can be classified according to the level of precision in the series: 6%, 4%, 2.5%, 1.5%, 1.0%, 0.5%, 0.2%, 0.1%, 0.05%, 0.02%, 0.01%, 0.005%, 0.002%, 0.001%, with the error type indicated by δ_s (for a multiplicative error), δ_0 (for an additive error) or $\delta_\mathrm{K}/\delta_0$ (for a combined error), as follows.

(1) For pure multiplicative errors the level of instrument precision is expressed by the relative error of the sensitivity δ_s, and is usually written as the number in a circle, e.g. $\delta_s = 1.5\%$ is written $\textcircled{1.5}$
(2) For pure additive errors the level of instrument precision is expressed by the reduced relative error δ_0 and $R = x_{\mathrm{max}}$ is the upper end of the scale range $x_{\mathrm{max}}$when $x_{\mathrm{min}} = 0$, e.g., $\delta_0 = 1.5\%$ is written 1.5.

(3) When the instrument has a strongly nonlinear scale the level of instrument precision is expressed by a relative error and the associated scale range $R = x_{\max} - x_{\min}$, e.g. $\overset{1.5}{\bigvee}$, 10 means $\delta_0 = 1.5\%$ and $R = 10$.

(4) For combined multiplicative and additive errors the level of instrument precision is expressed by the ratio δ_K/δ_0, e.g., 1.5/1 means $\delta_K = 1.5\%$ and $\delta_0 = 1\%$.

On the basis of the instrument precision it is possible to compute the maximum deviation likely to be caused by instrumental error (Table 1.1).

Table 1.1—Relative and absolute errors of an instrument expressed for the actual measured quantity x and the level of instrument precision p or p_1/p_2

Type of error	Class of precision	Scale range	Relative error δ (%)	Absolute error Δ
Additive	p	$x_{\max}$ $x_{\min} = 0$	$p(x_{\max}/x)$	$p \times x_{\max}/100$
	$\overset{p}{\bigvee}$	$x_{\max} - x_{\min}$	$\dfrac{p(x_{\max} - x_{\min})}{x}$	$\dfrac{p(x_{\max} - x_{\min})}{100}$
Multiplicative	$\textcircled{p}$	$x_{\max}$ $x_{\min} = 0$	p	$p \times x/100$
Combined	p_1/p_2	$x_{\min} = 0$ $x_{\max}$	$p_1 + (p_2 \times x_{\max}/x) - p_2$	$\dfrac{p_1 \times x + p_2(x_{\max} - x)}{100}$

Problem 1.2 *Determination of ammeter precision*

If an ammeter with range $R = 60\,\text{mA}$ gives a mean reading $\bar{x} = 49.6\,\text{mA}$ when the true value of the electric current is 50 mA, what is the level of precision? Use the limiting absolute error and the limiting reduced relative error. Also calculate the sensitivity limit.

Solution: $\Delta_0 = 50.0 - 49.6 = 0.4\,\text{mA}$

$\delta_0 = 100 \times 0.4/60 = 0.67\%$ rounded off to the nearest larger value in the series of allowed values of instrument precision, i.e. 1%; [(Eq. (1.5)]

$$x_c = 0.67 \times (60/100) = 0.402\,\text{mA} \qquad \text{[(Eq. (1.6)]}$$

Conclusion: The class of instrument precision is 1% and the sensitivity limit 0.402 mA.

Problem 1.3 *Limiting and reduced values of errors for ammeter*

The manufacturer claims the following data for the ammeter: 2, $R = 60\,\text{mA}$ meaning that $\delta_{0,R} = 2\%$, $x_{\min} = 0$ and $x_{\max} = 60\,\text{mA}$. Estimate the limiting absolute error Δ_0, the reduced relative error at $x_{\max}$, i.e. $\delta_0(x_{\max})$ and the reduced relative error at $x_{\min}$, i.e. $\delta_0(x_{\min})$ for this instrument.

Solution: $\Delta_0 = 2 \times (60/100) = 1.2\,\text{mA}$;

$$\delta_0(x_{\min}) = \infty \text{ for } x = 0\,\text{mA};$$

$$\delta_0(x_{\max}) = 2\ (60/60) = 2\%\ \text{for}\ x_{\max} = 60\,\text{mA}.$$

Conclusion: The relative error is not useful for expressing the error in a value close to zero. The minimum value for the reduced relative error is equal to the instrument precision, i.e. here 2%.

Problem 1.4 *Determination of level of voltmeter precision*
Determine the precision p for a voltmeter with a range from $x_{\min} = 0$ to $x_{\max} = 40\,\text{mV}$ and for which it is known that the error is a combination of additive and multiplicative errors. It was found that for $x = 10\,\text{mV}$, $\delta_{0,\text{R}}(10) = 2\%$, and for 40 mV, $\delta_{0,\text{R}}(40) = 5.2\%$ so that $\delta_\text{K} = 5.2\%$.
Solution: Since $\delta_\text{K} = 5.2\%$, δ_0 can be calculated from Eqs. (1.10) and (1.11), i.e. $\delta_0 = (\delta_{0,\text{R}}(x) - \delta_\text{K} \times x/R)/(1 - x/R)$ where $R = x_{\max} - x_{\min} = 40$. Therefore $\delta_0 = 0.93\%$. The value δ_K is rounded off to the nearest higher value in the series of instrument precision, i.e. 6%, and then δ_0 to 1%.
Conclusion: The voltmeter precision, expressed by the ratio δ_K/δ_0, is numerically equal to 6/1.

1.2.3 Estimating rounding errors

In practice, limiting relative error is very frequently expressed implicitly by reporting only the number of significant figures that are known with certainty in an approximate number.

Suppose that a positive approximate number contains s definitely known digits. Then, its decimal expansion is of the form

$$a = n_1 \times 10^r + n_2 \times 10^{r-1} + \ldots + n_s \times 10^p$$

where $n_1, n_2, \ldots, n_s$ are digits in the decimal representation with $n_1 \not\equiv 0$. The integers r and p (with $r > p$) and the positive integer s are related by $r - p = s - 1$.

Suppose that the number a is obtained on rounding off. Then, $\varepsilon_a = 10^p/2$, and from the definition of limiting relative error we have $\delta_a = 10^p/2a$. In the expression for δ_a the number a is replaced by its decimal expansion, keeping only the first term:

$$\delta_a < 10^p/(2n_1 \times 10^r) = 10^{1-s} \times 10/2n_1.$$

Here s is the number of significant digits in the approximate number and n_1 is the first digit of the number. We should note that in accordance with the formula obtained, the limiting relative error depends only on the number of known digits and n_1, and not on the position of the decimal point.

To get an approximate value for the relative error, we may take the average value of the first digit; that is, we may set $n_1 = 5$. Then, $\delta_a < 10^{-s}$ here has the value $\delta_a < 10^{-1}$, so the limiting relative error of a number with one definitely known digit is a number of the order of ten percent. For a number with two definitely known digits, it is one percent; for three, it is 0.1 percent, etc. We should note that in many applied problems, a limiting relative error of the order of a tenth of a percent is sufficient. The calculations in such problems can be made with three significant figures.

We have just determined the limiting relative error from the number of digits known. It is easy to solve the inverse problem too, namely, how many digits are required for a given limiting relative error. Suppose we need to find a value for s to make $\delta_a = 10^{-q}$. From formula:

$$10^{-s} \times 10/2n_1 \leqslant 10^{-q}; \quad 10^q \leqslant 10^s \times 2n_1/10; \quad 10^s \geqslant 10^q \times 10/2n_1$$

If we replace n_1 with the average number 5 (n_1 can take any integral value between 1 and 9), we get $10^s \geqslant 10^q$ and $s \geqslant q$. Consequently, on average, the number of known digits must be equal to the absolute value of the power of 10 in the value of δ_a. For example, if we need δ_a to have a value around one percent, the number must have no fewer than two definitely known digits.

To obtain a better estimate than the average, we begin with the decimal expansion of the number a:

$$(n_1 + 1) \times 10^r > a \geqslant n_1 \times 10^r$$

and hence

$$\delta'_a = 10 \times 10^{-s}/[2(n_1 + 1)]$$

if a is an approximate number obtained on rounding off. If we want $\delta_a = 10^{-q}$, we need a value for s such that δ'_a will be less than δ_a; that is

$$10^s \geqslant 10^{q+1}/[2(n_1 + 1)]$$

For the necessary number of known digits, we need to take the smallest integer s that satisfies this inequality. We can get something like the average value if we take $n_1 = 4$. Then, $s \geqslant$ q.

We can make a simple rule: if the first digit does not exceed 3, the number of known digits must exceed by 1 the absolute value of the power of 10 in the given relative error. In other cases, these numbers are equal. The value of 0 for $q = 0$ and $n_1 = 4$ to 9 means that there will be a 100% error if we do not know a single digit in the number for certain but only that the first digit is less than 4. To see this, suppose that the exact value of a one-digit number is 5. If an error of 100% is allowed for the number, the absolute value of the error can attain the value of 5 and the approximate number can have any value from 0 to 10; that is, the first digit will not be known.

1.2.4 Decomposition of measurement error

The precision of instruments is expressed by the absolute error of the instrument Δ_{inst}, which represents the first part of decomposed absolute error of the signal, Δ_{V}. The second part of the signal error is the variability of measured material Δ_{M}, the square of which is proportional to the variance σ^2. When the two parts of the error are uncorrelated, the decomposition of signal error Δ_{V} may be expressed by the equation

$$\Delta_{\text{V}} = \sqrt{\Delta_{\text{inst}}^2 + \Delta_{\text{M}}^2} \qquad (1.12)$$

(1) With the most precise instrument, the smallest error of signal Δ_V will be controlled by the material error Δ_M alone, so that $\Delta_{inst} \ll \Delta_M$. The precision of the signal can be increased by making a greater number of repeated signal measurements *n*.

(2) For an instrument with error $\Delta_{inst} \approx \Delta_M/3$, the signal error Δ_V will be only slightly higher than for a very precise instrument.

(3) For an instrument with $\Delta_{inst} \approx \Delta_M$, the error of the signal will be $\Delta_V \approx 1.4\ \Delta_M$. For *n* repeated signal measurements, the error of signal Δ_V will be decreased by $\sqrt{n}$, and consequently the random part of the instrument error Δ_{inst} will also decrease.

(4) For an instrument with $\Delta_{inst} \gg \Delta_M$, the error of signal measurement Δ_V will be proportional to the instrument error Δ_{inst}, i.e. $\Delta_V \approx \Delta_{inst}$. Repeated signal measurement cannot bring any improvement in the precision of the signal. An improvement of signal measurement is possible only with the use of a more precise instrument.

It may be concluded that a suitable choice of instrument is one with an error Δ_{inst} equal to $\Delta_m/3$ or less.

1.3 MODELS OF SIGNAL MEASUREMENT

Statistical analysis of errors involves the statistical treatment of repeated measurements of the analytical signal *S*. One of the following models of signal measurement is assumed.

(1) The additive model of signal measurement is the simplest one; the *i*th measured observation of the analytical signal is expressed by equation

$$x_i = \mu + \varepsilon_i \tag{1.13}$$

where ε_i represents the *i*th random error. It is obviously assumed that random errors have a mean of zero, constant variance and are not correlated. This model corresponds to the non-random signal μ, measured by an instrument that causes only random errors of measurement, ε_i; and the variability due to the material is equal to zero, $\Delta_M = 0$.

The additive model describes a measurement of a random signal variable ξ by its realizations x_i, with an ideally precise instrument for which $\Delta_{inst} = 0$. Here μ corresponds to the mean value.

A realistic approach involves a random variable ξ with a probability density function f_ξ, measured by an instrument which introduces errors ε with a probability density function f_ε. For an additive model of signal measurement [Eq. (1.13)] the probability density function f_x of the measured signal variable *x*, may be expressed as

$$f_x(x) = \int_{-\infty}^{\infty} f_\xi(x - \varepsilon) \times f_\varepsilon(\varepsilon)\, \mathrm{d}\varepsilon \tag{1.14}$$

It may be concluded that

(a) if f_ξ is a probability density function with normal distribution $N(\mu, \sigma^2)$, and f_ε is a probability density function with normal distribution $N(0, \tau^2)$, the probability density function f_x will then also have a normal distribution, given by $N(\mu, \sigma^2 + \tau^2)$;

(b) by convolution of some unimodal symmetric distributions the bimodal distribution may be formed;

(c) if the variance of an instrument τ^2 is not constant, even for a normal distribution of f_ξ, the distribution of results is rather complicated.

(2) The multiplicative model of signal measurement supposes that the errors have the following effect:

$$x_i = \mu \times \exp(\varepsilon_i) \tag{1.15}$$

where error ε_i has the same properties as in the previous model. The variable $\ln x_i$ has a distribution with mean $\ln \mu$ and variance σ^2. The variance of the measured quantities x_i is given by

$$\sigma^2(x) = x^2 \times \sigma^2 \tag{1.16a}$$

and the corresponding relative error by

$$\delta(x) = \sigma(x)/x = \sigma \tag{1.16b}$$

For the multiplicative model, the relative error is constant. When random errors ε_i have a normal distribution, the measured (signal) variables x_i have a log-normal distribution.

(3) In the model with a systematic error, the measurement includes a systematic error of the instrument. The simplest model of this type is expressed by

$$x_i = \mu + \varepsilon_i + a \tag{1.17}$$

where a is the constant systematic error of an instrument. If measurements are made at only one signal level (i.e., level of μ) it is not possible to determine the systematic error a. If n replicate signal measurements x are made at each of several different signal levels μ_j, $j = 1, \ldots, m$, the resulting model

$$x_{ij} = \mu_j + \varepsilon_{ij} + a \tag{1.18}$$

may be examined by one-way analysis of variance (Chapter 4). This model assumes that

(1) the mean error $\bar{\varepsilon}_j$ of repeated measurements is equal to zero;
(2) the variance of errors is constant;
(3) errors of repeated measurements are not correlated,
(4) errors at different levels μ_j are also not correlated [5].

To analyse a set of measurements it is necessary to know the error distribution. Distribution determination needs a rather large number of signal measurements. We will consider four main types of error distribution; these are illustrated in Fig. 1.3.

(1) The **rectangular distribution** occurs when measurements have errors that are formed by rounding off the numbers.

(2) The **normal distribution** is observed for signal measurements when the errors ε result from a sum of partial errors and the measurements were performed at constant variance.

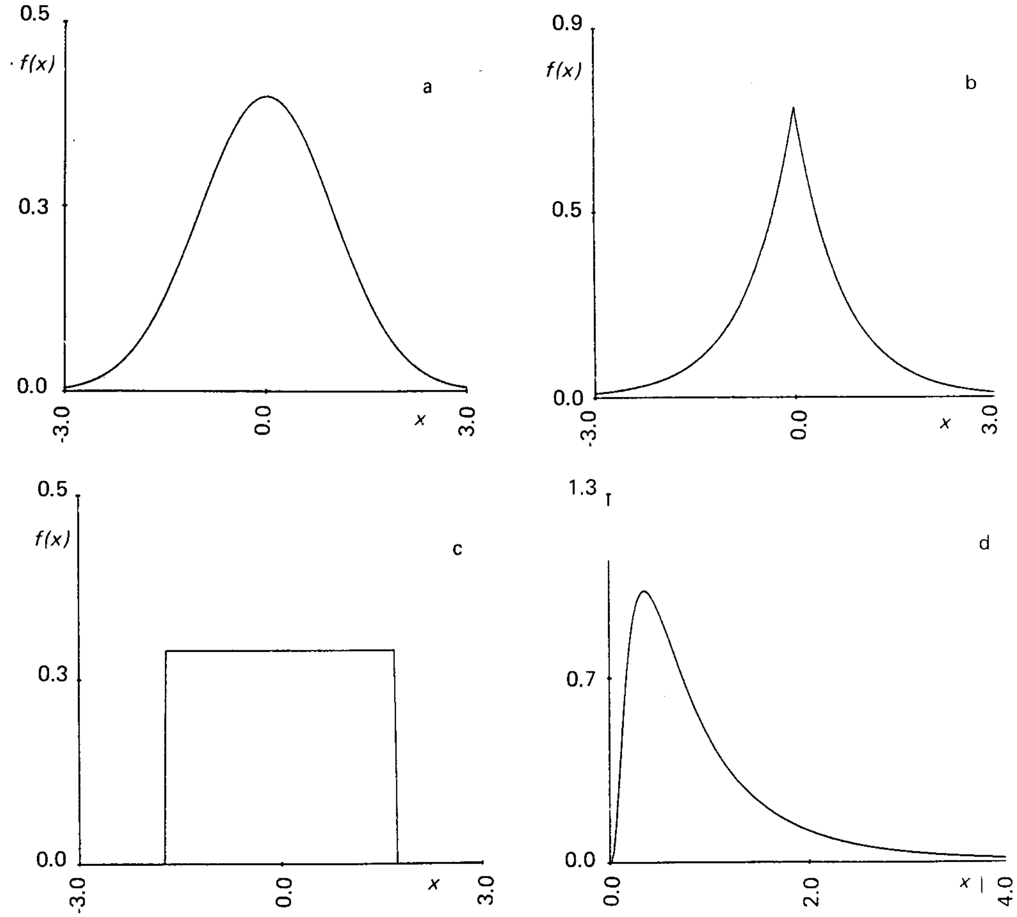


Fig. 1.3 Selected distributions of random errors: (a) normal, (b) Laplace, (c) rectangular, and (d) log-normal.

(3) The **Laplace (two-tailed exponential) distribution** is observed for signal measurements when the variance of measurements oscillates around a mean value.

(4) The **log-normal distribution** is observed for signal measurements when the errors ε result as a product of elemental errors. Measurements must be positive and performed at constant relative error (coefficient of variation).

When the error distribution is bounded by some finite interval (e.g., as for a rectangular distribution) the limiting error of measurement may be calculated as half of the bounding interval. Because there is a need for exact expressions of error ε, various quantile and moment estimates of random errors are used.

1.4 QUANTILE ESTIMATES OF MEASUREMENT ERRORS

One of the measures of the spread of a random variable ε is the interquantile deviation $K_{1-\alpha}$ which defines an interval containing $100(1-\alpha)\%$ of all random errors

$$K_{1-\alpha} = (\tilde{x}_{1-\alpha/2} - \tilde{x}_{\alpha/2})/2 \tag{1.19a}$$

The x_α denote the quantile of the given random variable (here errors). Suppose that the errors ε have a mean of zero. On the base of the $K_{1-\alpha}$, the $100P = 100(1-\alpha)\%$ estimate of the limiting error of measurement is expressed by

$$\Delta_P = K_{1-\alpha} \tag{1.19b}$$

Outside the interval $(-\Delta_P, +\Delta_P)$ there will be $100(1-P)\%$ of all the errors. To calculate the estimate of a limiting error of measurement, σ_{Δ_p}, the following quantities should be chosen.

(1) For $P = 0.5$ (or 50%), the limiting error of measurement is called the mean error $\sigma_{\Delta_{0.5}}$. For a normal distribution it is expressed by

$$\sigma_{\Delta_{0.5}} = 0.68\sigma(x) \tag{1.20}$$

(2) For $P = 0.683$ (or 68.3%), the limiting error of measurement is called the probable error $\sigma_{\Delta_{0.689}}$. For a normal distribution it is calculated by

$$\sigma_{\Delta_{0.689}} = \sigma(x) \tag{1.21}$$

(3) For $P = 0.9$ (or 90%), the limiting error of measurement is given, for various symmetric distributions [4], by

$$\sigma_{\Delta_{0.9}} = 1.65\sigma(x) \tag{1.22}$$

The choice of $P = 0.9$ (or 90%) is convenient in cases where the error distribution is unknown. When some limiting errors of measurement $\Delta_{0.9,i}$, $i = 1, \dots, n$, are known, the total error of a sum of quantile (limiting) errors of measurements $\Delta_{0.9,\mathrm{T}}$ may be calculated as

$$\sigma_{\Delta_{0.9,\mathrm{T}}} = \sum_{i=1}^{n} \sqrt{\sigma^2_{\Delta_{0.9,i}}} \tag{1.23}$$

For other values of P, Eq. (1.23) is not valid.

(4) Generally, when errors have a normal distribution, the limiting error of measurement may be expressed with use of the $100(1+P)/2$ percentage quantile of the standardized normal distribution $u_{(1+P)/2}$ by using the expression

$$\sigma_{\Delta_P} = u_{(1+P)/2} \times \sigma(x) \tag{1.24}$$

whereas for other distributions Δ_P is expressed by

$$\sigma_{\Delta_P} = h \times \sigma(x) \tag{1.25}$$

where the quantile h is a function of P and of the kurtosis g_2. The quantile h may be expressed by the equation [4]

$$h = 1.62\,[3.8(g_2 - 1.6)^{2/3}]^Z \tag{1.26}$$

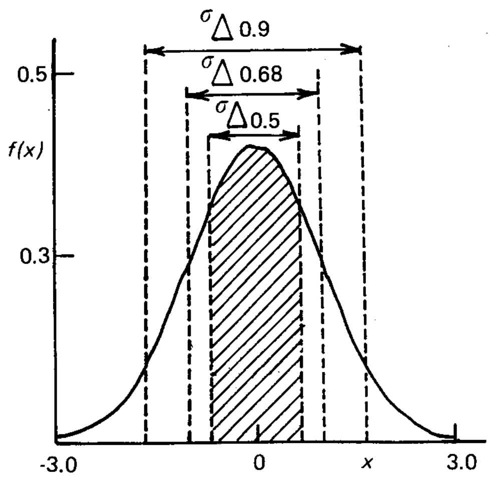


Fig. 1.4—Geometric interpretation of various quantile estimates of instrumental error: the numbers relate to the area under the probability density curve.

where $Z = \log\{\log[1/(1 - P)]\}$, for values P chosen in the interval $0.9 \leq P \leq 0.99$. The kurtosis for the normal distribution is ($g_2 = 3$), for the rectangular distribution it is ($g_2 = 1.8$) and for the Laplace distribution it is ($g_2 = 6$).

(5) For estimation of quantile estimates of errors (Fig. 1.4) the quantiles $\tilde{x}_P$ and $\tilde{x}_{1-P}$ may be applied. The sample values $x_1, x_2, \ldots, x_n$ (where x_i is now the ith error of measurement) are first of all arranged in increasing order of magnitude so that smallest is $x_{(1)}$, the next smallest $x_{(2)}, \ldots$, and the largest $x_{(n)}$. The data now form the set of order statistics. The order statistics divide the x-axis into $(n + 1)$ intervals of the same probability $P_i = 1/(n + 1)$ of occurrence of the random variable x. Therefore, it is not possible from the data set of size n to determine quantiles of any value of statistical certainty $0 \leq P \leq 1$. The minimum size of a data set for a reliable determination of the $100P\%$ quantile (without $x_{(1)}$ and $x_{(n)}$) is calculated from

$$n_{\min} \geq 4/(1 - P) \tag{1.27}$$

For example, for $P = 0.9$ it is necessary to have a minimum of $n_{\min} = 4/(1 - 0.9) = 40$ measurements. If the error distribution is normal, the limiting error of measurement σ_{Δ_P} may be calculated for size $n < 30$ as

$$\sigma_{\Delta_P} = t_\alpha(n - 1) \times s/\sqrt{n} \tag{1.28}$$

where $t_\alpha(n - 1)$ is the $\alpha = (1 + P)/2$ quantile of the Student distribution and s is the sample standard deviation (Chapter 3.1). Quantile errors can not be directly added but some approximate expressions can be used [4].

Problem 1.5 *Quantile estimate of limiting error for two different distributions*
From preliminary experiments, the variance of the signal measurement of the instrument was found to be $\sigma^2 = 0.5$. Find quantiles estimates of the 95% limiting

error $\sigma_{\Delta_{0.95}}$ for measurements having (a) a normal distribution or errors, and (b) a rectangular distribution of errors.
Data: $\sigma^2 = 0.5$
Program: Chemstat: Basic Statistics: One sample analysis
Solution: Equation (1.25) is used for estimation of the limiting quantile error of measurement, $\sigma_{\Delta_{0.95}}$, and Eq. (1.26) for h, with $Z = 1.14287$. For (a) a normal distribution $g_2 = 3$, and therefore $h = 1.936$ and $\sigma_{\Delta_{0.95}} = 0.968$; for (b) a rectangular distribution $g_2 = 1.8$, and therefore $h = 1.669$ and $\sigma_{\Delta_{0.95}} = 0.835$.
Conclusion: The error distribution has a great influence on the quantile estimate of the limiting error of signal measurement: $\sigma_{\Delta_{0.95}} = 0.968$ for a normal distribution, but for a rectangular it is 0.835.

1.5 SUMMATION OF QUANTILE ESTIMATES OF MEASUREMENT ERRORS

The total error of measurement is estimated from known partial errors and their known distributions. Generally, the quantile estimates of limiting error of measurement cannot be added because they contain a quantile that is strongly dependent on the error distribution [4]. The only exception is for $P = 0.9$, for which summation of errors $\sigma_{\Delta_{0.9,i}}$, $i = 1, \ldots, n$, can be done by Eq. (1.23).

We limit ourselves to calculating the quantile estimate of the limiting error $\sigma_{\Delta_p}(x + y)$ of two independent and symmetrically distributed random variables x and y with variance σ^2 (x) and $\sigma^2(y)$, and with kurtosis $g_2(x)$ and $g_2(y)$. To calculate $\sigma_{\Delta_p}(x + y)$ the variance of the sum is calculated first

$$\sigma^2(x + y) = \sigma^2(x) + \sigma^2(y) \tag{1.29}$$

and the kurtosis of the sum [4]

$$g_2(x + y) = g_2(x) \times p^2 + 6p \times (1 - p) + g_2(y) \times (1 - p)^2 \tag{1.30}$$

where p is the relative weight of variance defined by

$$p = \sigma^2(x)/\sigma^2(x + y) \tag{1.31}$$

On the basis of known $\sigma_{\Delta_p}(x)$, $\sigma_{\Delta_p}(y)$ and $g_2(x)$, $g_2(y)$ the resulting $\sigma_{\Delta_p}(x + y)$ is calculated as follows.

(1) The parameter h for both kurtosis is calculated from Eq. (1.26).
(2) The estimates of standard deviations $\sigma(x)$ and $\sigma(y)$ are calculated from Eq. (1.25).
(3) The estimates $\sigma^2(x + y)$ and $g_2(x + y)$ are calculated from Eqs. (1.29) and (1.30), and then the estimates of h from Eq. (1.26).
(4) The quantile estimate of the sum of the two errors $\sigma_{\Delta_p}(x + y)$ is calculated from Eqs. (1.31), (1.26) and (1.25).

When more than two errors are to be added, the procedure for two errors is repeated.

Problem 1.6 *Summation of quantile estimates of voltage errors*
The instrument consists of three blocks. In block A the errors are distributed rectangularly with variance $\sigma^2(\mathrm{A}) = 0.1$, in block B the errors have arcsin distribution with $\sigma^2(\mathrm{B}) = 0.05$, and in block C the errors have a normal distribution with variance $\sigma^2(\mathrm{C}) = 0.1$. Estimate the 95% quantile estimate of the limiting error of measurements, by addition of all voltage errors, and assuming their independence. The kurtosis of the three distributions are: $g_2(\mathrm{A}) = 1.8$, $g_2(\mathrm{B}) + 1.5$ and $g_2(\mathrm{C}) = 3$.
Solution: We calculate $\sigma^2(\mathrm{A} + \mathrm{B}) = \sigma^2(\mathrm{A}) + \sigma^2(\mathrm{B}) = 0.15$ and $g_2(\mathrm{A} + \mathrm{B}) = 1.88 \times 0.66^2 + 6 \times 0.66(1 - 0.66) + 1.5(1 - 0.66)^2 + 2.299$. In the next step the variance $\sigma^2(\mathrm{A} + \mathrm{B} + \mathrm{C})$ and the kurtosis $g_2(\mathrm{A} + \mathrm{B} + \mathrm{C})$ are calculated: $\sigma^2(\mathrm{A} + \mathrm{B}) + \sigma^2(\mathrm{C}) = 0.25$ and $g_2(\mathrm{A} + \mathrm{B} + \mathrm{C}) = g_2(\mathrm{A} + \mathrm{B}) + g_2(\mathrm{C}) = 2.299 \times 0.66^2 + 6 \times 0.66(1 - 0.66) + 3(1 - 0.66)^2 = 2.75$. The next step of calculation estimates h from Eq. (1.26): $h = 1.62(3.8(2.75 - 1.6)^{2/3})^{1.1428}$. The resulting 95% quantile estimate of limiting error of measurement for the three blocks is $\sigma_{\Delta_{0.95}} = 1.907 \times (0.25)^{1/2} = 0.954$. If the differences in the error distributions were neglected an approximately normal distribution of all three block errors were assumed, the answer would be $\sigma_{\Delta_{0.95}} = 1.96 \times (0.25)^{1/2} = 0.98$.
Conclusion: The result of error summation depends on the distribution of the partial errors.

1.6 MOMENT ESTIMATES OF MEASUREMENT ERRORS

For an additive model of measurement, $x_i = \mu + \varepsilon_i$, the measurement of spread is represented by the standard deviation which is closely connected with the variability of the material measured and with all errors of measurement. In the case of a normal distribution of the measured variable x and of the errors of measurement, the magnitude of the standard deviation $\sigma(x)$ is equal to the quadratic average of the signal variance σ^2 and the variance of measurement τ^2, i.e. $\sigma^2(x) = \sigma^2 + \tau^2$.

The standard deviation is a measure of precision and as such gives a measure of the precision of the location estimate. More information about the variability of a measured quantity x can be obtained from the intervals within which, with a given probability, the random errors of measurement exist. Let us limit ourselves to the case of a random error ε with mean equal to zero. When the standard deviation $\sigma(x)$ of errors in measurement is known, Mandel [6] suggests calculation of the probability interval for random error ε. More frequently, only the sample estimate of the standard deviation $s(x)$ is known, and then a tolerance interval for random error ε is used.

1.6.1 Probability interval for random error

Let us assume a known distribution of errors, given by the distribution function $F(x)$ or its inverse i.e. the quantile function $Q(\alpha) = F^{-1}(x)$. When the standard deviation $\sigma(x)$ is also known, it is possible to determine a probability P with which random errors will lie within an interval $\pm k \times \sigma(x)$, where k is a selected number. If random errors have a symmetrical distribution with zero mean, this probability is equal to

$$\begin{aligned} P &= P(-k \times \sigma(x) \le \varepsilon \le k \times \sigma(x)) = F(k \times \sigma(x)) - F(-k \times \sigma(x)) \\ &= 1 - 2F(-k \times \sigma(x)) \end{aligned}$$

where $F(k \times \sigma(x))$ is the value of the distribution function at the point $k \times \sigma(x)$. The interval $(-k \times \sigma(x), k \times \sigma(x))$ may be said, colloquially, to "contain a proportion P of the distribution", in the sense that, in the long run, a proportion P of repeated observations of ε would lie in the interval. If P is large (say 0.95 or 0.99) we are "practically certain" that most realizations of ε will lie within the probability interval $(-k \times \sigma(x), k \times \sigma(x))$.

When $(-k)$ is a standardized α-quantile $Q_0(\alpha)$ of the given distribution and k a standardized $(1-\alpha)$-quantile $Q_0(1-\alpha)$ of the distribution, the quantities $\pm k \times \sigma(x)$ represent non-standardized, original quantiles $Q(\alpha)$ and $Q_0(1-\alpha)$ of error distribution. Because the quantile function is an inverse function to the distribution one, it follows that

$$F[Q(\alpha)] = \alpha \text{ and } F[Q(1-\alpha)] = 1-\alpha$$

and therefore $P = 1 - 2\alpha$. In the interval

$$Q_0(\alpha) \times \sigma(x) \leq \varepsilon \leq Q_0(1-\alpha) \times \sigma(x)$$

lie $100(1-2\alpha)\%$ of all errors.

If a normal distribution of errors ε is assumed and a standardized quantile of the normal distribution is denoted as u_α, the probability interval for random error may be estimated as

$$-\sigma(x) \times \mathrm{u}_\alpha \leq \varepsilon \leq \sigma(x) \times \mathrm{u}_\alpha \qquad (1.32)$$

In this interval the random errors ε lie with probability $(1-2\alpha)$. The actual probability value usually chosen is $\alpha = 0.025$, so that $P_{0.025} = 0.95$ means that 95% of normally spread errors lie in the interval $-1.96\sigma(x) \leq \varepsilon \leq 1.96\sigma(x)$, and only 2.5% of all possible random errors will be larger than $1.96\sigma(x)$ and 2.5% smaller that $-1.96\sigma(x)$. When the distribution of errors is unknown, it is recommended to use the level $P = 0.90$, (*or* $\alpha = 0.05$), for which $u_{0.05} = -1.64$. Some unimodal symmetrical distributions of random errors have 90% of all errors in the probability interval $-1.64\sigma(x) \leq \varepsilon \leqslant 1.64\sigma(x)$. For an asymmetric distribution of errors this interval cannot be used.

Problem 1.7 *Probability interval for the error of Laplace and normal distributions*
Estimate the 99% probability interval of measurement error for the Laplace distribution and also for the normal distribution when $\sigma^2(x) = 1$.
Solution: Within the interval $Q_0(\alpha) \times \sigma(x) \leq \varepsilon \leq Q_0(1-\alpha) \times \sigma(x)$ there will be $100(1-2\alpha)\%$ of all errors. For the 99% probability interval the $\alpha = 0.005$ standardized distribution function of Laplace distribution with zero mean and unity variance has for $s < 0$ the form

$$F(x) = 0.5 \exp(\sqrt{2}s)$$

The corresponding quantile function $Q_0(\alpha)$ for $\alpha = 0.005$ is equal to $Q_0(\alpha) = \ln(2\alpha)/\sqrt{2} = -3.256$. Since the distribution is symmetrical, $Q_0(1-\alpha) = 3.256$ and the 99% probability interval is $-3.256 \leq \varepsilon \leq 3.256$. For comparison, the 99% probability interval for a normal distribution will be calculated. From statistical tables

we take the value $Q_0(\alpha = 0.005) = -2.575$, so the 99% probability interval will be $-2.575 \leq \varepsilon \leq 2.575$.
Conclusion: The Laplace distribution leads to a 99% probability interval significantly broader than the normal one.

1.6.2 Tolerance interval for random error

In many practical cases the standard deviation $\sigma(x)$ is unknown and can only be estimated by $s(x)$. In this case we construct the tolerance interval which will contain $100P\%$ of all errors with statistical certainty $(1 - \alpha)$. When the error mean is zero the tolerance interval for random errors will be

$$-k_{\mathrm{T}} \times s(x) \leq \varepsilon \leq k_{\mathrm{T}} \times s(x) \tag{1.33}$$

where k_{T}, for a normal distribution of error, is given by

$$k_{\mathrm{T}} = u_{(1+P)/2}\sqrt{\frac{n-1}{\chi^2_\alpha(n-1)}} \tag{1.34}$$

where χ^2_α is the quantile of the χ^2-distribution. The tolerance intervals are always broader that the probability ones.

Problem 1.8 *Comparison of probability and tolerance intervals*
An instrument with a declared standard deviation of signal measurement $s(x) = 0.5$ was used to make 15 measurements. Estimate the tolerance interval covering 90% of all measurement errors with a statistical certainty of $1 - \alpha = 0.95$. Assume that the standard deviation $\sigma(x)$ is known and equal to 0.5; then compare the two probability intervals, and say which is broader.
Solution: From Eq. (1.34), $k_{\mathrm{T}} = 1.64\sqrt{(15-1)/6.57} = 2.394$. The tolerance interval is then $-1.197 \leq \varepsilon \leq 1.197$. For $\sigma(x) = 0.5$ and $P = 0.9$ the value $u_{0.05}$ is, from statistical tables, equal to -1.64 and therefore $-0.82 \leq \varepsilon \leq 0.82$.
Conclusion: The tolerance interval is significantly wider than the probability one.

1.7 PROPAGATION OF ERRORS IN EXPERIMENTAL OPERATIONS

Direct results x_j of instrumental measurements in a chemical laboratory are always approximate, mainly because of the limited accuracy of measuring instruments. Results of a chemical analysis y are calculated from several measured quantities $x_1, \ldots, x_m$ by the function $y = G(x_1, \ldots, x_m)$. We will formulate the following steps.

(1) the estimation of the results of the chemical analysis, i.e. the mean value $\bar{y}$ and variance $s^2(y)$;
(2) the estimation of the total error of chemical analysis $s(y)$ from known errors of several measured quantities, $s(x_i)$;
(3) the reverse estimation of limiting errors of measured quantities $s(x_i)$ from the allowed error of the chemical analysis, $s(y)$.

To express the absolute error of the ith variable x_i, the standard deviation $s(x_i)$ is convenient; for the relative error of x_i the relative standard deviation (or coefficient of variation) is used

$$\delta(x_i) = s(x_i)/x_i \tag{1.35}$$

For the first step the estimate of mean $\bar{x}_i$, variance $s^2(x_i)$, skewness $\hat{g}_{1,i}$ and kurtosis $\hat{g}_{2,i}$ calculated by the methods described in Chapter 3 are used.

For the second step the expressions for variance $s^2(y)$ as a function of individual variances $s^2(x_i)$ are used. A simplification can be achieved by the use of relative errors.

For the third step the expressions for variance $s^2(y)$ or coefficient of variation $\delta(y) = s(y)/\bar{y}$ are used, with an assumption that individual measured quantities x_i have the same relative effects.

To solve all three steps the mean $\bar{y}$ and its variance $s^2(y)$ of a function $y = G(x_1, \ldots, x_m)$ must be known. The estimates y and $s^2(y)$ may be obtained by any of the following methods:

(1) Taylor series expansion of the function $y = G(x_1, \ldots, x_m)$;
(2) two-points estimates;
(3) Monte-Carlo simulation.

Whereas the method of Taylor series requires a knowledge of at least the first and second derivative of the function $y = G(x_1, \ldots, x_m)$, the remaining two methods can be computer-assisted, and can give more reliable results.

1.7.1 Method of Taylor series expansion

When a function of random variables is analysed, it should be realized that each non-linear transformation of the random variable distorts its distribution, and therefore changes the dependence of variance on the mean value. In the case when the measured variable x has a constant variance $\sigma^2(x)$, the results of analysis $y = G(x)$ has a non-constant variance

$$\sigma^2(y) = \left(\frac{\mathrm{d}G(x)}{\mathrm{d}x}\right)^2 \sigma^2(x) \tag{1.36}$$

Moreover, the mean $\bar{y}$ cannot be estimated by direct substitution of mean $\bar{x}$ into the function $G(\bar{x})$, i.e.

$$\bar{y} \neq G(\bar{x}) \tag{1.37}$$

To estimate the mean $\bar{y}$, the variance $s^2(y)$, and higher moments, the Taylor series expansion of function $G(\mathbf{x})$ can be used.

Suppose that the function $y = G(x_1, \ldots, x_m)$ is known. Let $G(\mathbf{x})$ be a differentiable function. On writing the Taylor series expansion in the neighbourhood of the vector of means $\bar{\mathbf{x}} = (\bar{x}_1, \ldots, \bar{x}_m)^{\mathrm{T}}$ we obtain

$$\begin{aligned} G(\mathbf{x}) \approx G(\bar{\mathbf{x}}) &+ \sum_{i=1}^{m} \frac{\delta G(\mathbf{x})}{\delta x_i}(x_i - \bar{x}_i) + \frac{1}{2} \sum_{i=1}^{m} \frac{\delta^2 G(\mathbf{x})}{\delta x_i^2}(x_i - \bar{x}_i)^2 \\ &+ \sum_{i=1}^{m-1} \sum_{j>i}^{m} \frac{\delta^2 G(\mathbf{x})}{\delta x_i \delta x_j}(x_i - \bar{x}_i)(x_j - \bar{x}_j) + \ldots \end{aligned} \tag{1.38}$$

where all first and second derivatives are calculated for the vector of mean values $\bar{\mathbf{x}}$. By using a mean value operator $E(.)$ at both sides of Eq. (1.38) the expression for the estimate of mean $\bar{y}$ may be written as

$$\bar{y} \approx G(\bar{x}) + \frac{1}{2}\sum_{i=1}^{m}\frac{\delta^2 G(\mathbf{x})}{\delta x_i^2} \times s^2(x_i) + \sum_{i=1}^{m-1}\sum_{j>i}^{m}\frac{\delta G(x)}{\delta x_i \delta x_j} \times \operatorname{cov}(x_i, x_j) \qquad (1.39)$$

where $\bar{y} = E(y) = E(G(\mathbf{x}))$, $s^2(x_i) = E[(x_i - \bar{x}_i)^2]$ and where $E[x_i - x_i] = 0$. The symbol $\operatorname{cov}(x_i, x_j)$ stands for the covariance which give "a measure of linear dependence" between the two variables x_i and x_j.

Where the variance $s^2(y)$ is determined by an approximation [Eq. (1.38)], higher moments (i.e., the skewness and kurtosis) are neglected. The resulting approximate relation for variance is termed the rule of propagation of absolute errors and expressed by

$$s^2(y) \approx \sum_{i=1}^{m}\left[\frac{\delta G(\mathbf{x})}{\delta x_i}\right]^2 \times s^2(x_i) + 2\sum_{i=1}^{m-1}\sum_{j>i}^{m}\frac{\delta G(\mathbf{x})}{\delta x_i}\frac{\delta^2 G(\mathbf{x})}{\delta x_j}\operatorname{cov}(x_i, x_j)$$

$$+ \sum_{i=1}^{m-1}\sum_{j>i}^{m}\frac{\delta^2 G(\mathbf{x})}{\delta x_i \delta x_j} \times s^2(x_i) \times s^2(x_j) \qquad (1.40)$$

The third term of Eq. (1.40) is usually neglected. When the resulting error $s(y)$ is formed from m sources of errors and each source has its own variance $\sigma^2(x_i)$, the following expression for the error estimate can be used[8]

$$s^2(y) = \sum_{i=1}^{m} s^2(x_i) + 2\sum_{i=1}^{m-1}\sum_{j>i}^{m}\operatorname{cov}(x_i, x_j) \qquad (1.41)$$

where cov (x_i, x_j) again is a measure of the linear dependence between the two variables x_i and x_j. There are two limiting cases of estimation of total error of measurements, $s(y)$:

(1) the sources of errors are quite independent, so that the covariance $\operatorname{cov}(x_i, x_j)$ is equal to zero. The resulting estimate of the error will be proportional only to the quadratic mean of errors $s(x_i)$ coming from m sources,

$$s(y) = \sqrt{\sum_{i=1}^{m} s^2(x_i)} \qquad (1.42)$$

(2) the sources of errors are linearly dependent and the covariance cov (x_i, x_j) is given by

$$\operatorname{cov}(x_i, x_j) = \sqrt{s^2(x_i) \times s^2(x_j)}$$

The resulting estimates of the total error will be proportional to the arithmetic mean of errors $s(x_i)$ coming from all m sources

$$s(y) = \frac{1}{m} \sum_{i=1}^{m} s(x_i) \tag{1.43}$$

For various experimental operations and signal measurements in a chemical laboratory, the function $G(\mathbf{x})$ can be expressed by a power-type relationship

$$y = G(\mathbf{x}) = x_1^{a_1} . \, x_2^{a_2} . \, \ldots . \, x_m^{a_m} = \prod_{i=1}^{m} x_i^{a_i} \tag{1.44}$$

where a_i are known coefficients usually equal to ±1. The estimation of the absolute error $s(y)$ or $s^2(y)$ by Eq. (1.40) is then rather complicated. Simplification results from the logarithmic transformation

$$\ln G(\mathbf{x}) = \sum_{i=1}^{m} a_i \times \ln x_i \tag{1.45}$$

Then

$$\frac{\mathrm{d} \ln G(\mathbf{x})}{\mathrm{d}x} = \frac{1}{G(\mathbf{x})} \times \frac{\mathrm{d}G(\mathbf{x})}{\mathrm{d}x} \tag{1.46}$$

Substitution from Eq. (1.46) into Eq. (1.40) and rearrangement leads to a simplified form for the relative error (variation coefficient)

$$\delta^2(y) \approx \sum_{i=1}^{m} a_i^2 \delta^2(x_i) + 2 \sum_{i=1}^{m-1} \sum_{j>i}^{m} a_i a_j r_{ij} \delta(x_i) \delta(x_j) \tag{1.47}$$

where r_{ij} represents the correlation coefficient expressing the closeness of linear dependence between variables x_i and x_j. Equation (1.47) is called the rule of propagation of relative errors. The quality of the estimates $\bar{y}$, $s^2(y)$ and $\delta^2(y)$ is dependent on the quality of the approximation of the function $G(\mathbf{x})$ by the quadratic function.

Although the estimate $\bar{y}$ is normally sufficiently accurate, some inaccuracy may be found in the estimation $s^2(y)$ [9].

Equation (1.47) may be used for estimation of relative errors $\delta(x_i)$ such that the relative error of chemical results $\delta(y)$ will not be greater than the selected value for H in %, i.e $100\delta(y) \leq H$. In solving this inversion problem, the independence of the measured variables x_i and the principle of the same relative influence

$$|a_1|\delta(x_1) \approx |a_2|\delta(x_2) \approx \ldots . \approx |a_{\mathrm{m}}|\delta(x_m) \approx H/m$$

are assumed. Here a_i, $i = 1, \ldots, m$, are the coefficients of the function $G(\mathbf{x})$ [Eq. (1.44)]. For other types of function $G(\mathbf{x})$ the expression $a_i \approx |\mathrm{d} \ln G(\mathbf{x})/\mathrm{d}x_i|$ is used. For estimating the mean $\bar{y}$, the second derivatives $\mathrm{d}^2G(\mathbf{x})/\mathrm{d}x_i^2$ play an important role.

For the case of a ratio x_1/x_2, an estimate of the mean $\bar{y}$ is controlled only by the variance $\sigma^2(x_2)$ and not by the variance $\sigma^2(x_1)$.

Problem 1.9 *Error in the isotope dilution method by Taylor's formula*
Arsenic was determined by the method of isotope dilution. The initial specific activity was $a_2 = 3.7 \times 10^4 \text{sec}^{-1}$. After addition of the standard, $m_1 = 5 \times 10^{-7}$ g of arsenic, the specific activity was $a_1 = 5.3 \times 10^6 \text{sec}^{-1}$. Estimate the relative error of the arsenic content in the sample when the relative error of weighing is $\delta(m) = 0.03\%$, and relative error of activity measurement $\delta(a_1) = \delta(a_2) = 1\%$.
Program: Chemstat: Basic Statistics: Error propagation.
Solution. The content of arsenic m_x in the samples is calculated by the relation $m_x = m_1(a_1 - a_2)/a_2$. Because this expression is not in the form of Eq. (1.44), Eq. (1.47) cannot be used. The relative error will be estimates by Eq. (1.35). Assuming that the quantities m_1, a_1, and a_2 are not correlated, substitution into Eq. (1.39) gives

$$\bar{m}_x \approx m_1(a_1 - a_2)/a_2 + m_1 a_1 s^2(a_2^3)$$

$$+ 7.112 \times 10^{-5} + 7.162 \times 10^{-9} = 7.113 \times 10^{-5}\text{g}.$$

The variance is expressed by Eq. (1.40), after omitting the third term, as

$$s^2(m_x) = (a_1/a_2 - 1)^2 \times s^2(m) + (m_1/a_2)^2 \times s^2(a_1) - (m_1 a_1/a_2^2)$$

$$s^2(a_2) = (a_1/a_2 - 1)^2 m_1^2 \delta^2(m) + (m_1 a_1/a_2)^2 \times (\delta^2(a_1) + \delta^2(\text{a}_2))$$

$$= 3.2 \times 10^{-18} + 1.0259 \times 10^{-12} = 1.0259 \times 10^{-12}.$$

The relative error is

$$\delta(m_x) = 100\ s(m_x)/m_x = 1.424\%.$$

Conclusion: When an expression for determination of analyte content or concentration is not in the form of Eq. (1.44), Eq. (1.35) should be used.

1.7.2 Method of two-point estimates

Manly's procedure [7] of two-point estimates is based on replacement of the probability distribution of function $G(\mathbf{x})$ by the two-points distribution with the same mean and variance. The estimate of the mean is then expressed by

$$\bar{y} \approx (G(\bar{x} + s(x)) + G(\bar{x} - s(x)))/2 \tag{1.48}$$

and the estimate of variance by

$$s^2(y) \approx (G(\bar{x} + s(x)) - G(\bar{x} - s(x)))^2/4 \tag{1.49}$$

Both simple relations give better results than Taylor's formula for a function of type (1.44).

When the function $G(\mathbf{x})$ is a function of m independent random variables $x_1, \ldots, x_m$, the summation of Eqs. (1.48) and (1.49) can be used

$$\bar{y} \approx \sum_{i=1}^{m} (G(\bar{x}_i + s(x_i)) + G(x_i - s(x_i)))/2m \tag{1.50}$$

and

$$s^2(y) \approx \sum_{i=1}^{m} (G(\bar{x}_i + s(x_i)) - G(\bar{x}_i - s(x_i)))^2/4 \tag{1.51}$$

1.7.3 Monte-Carlo simulation method

The mean $\bar{y}$ and variance $s^2(y)$, as a function of random variables x, may be determined by computer-assisted Monte-Carlo simulation methods. Schwartz [9] showed that this general procedure is well suited for simulation of statistical behaviour of even rather complicated systems. The following steps can be formulated.

(1) Selection of the function $G(\mathbf{x})$: for many chemical problems the function $G(\mathbf{x})$ is usually known. The great advantage of the Monte-Carlo simulation method is that the function $G(\mathbf{x})$ need not necessarily be expressed in explicit form.

(2) Distribution of measured variables: in chemistry it is usually assumed that measured variables are independent and have normal distribution. The Monte-Carlo simulation method requires numerical values of the quantities $\bar{x}_i$, $s(x_i)$, $i = 1, \ldots, m$ only.

When these values are not available, two limiting values of interval $[A, B]$ in which the variables x_i are expected should be supplied. The approximate probability density function is then expressed by the parabolic distribution

$$f(x_i) = 6(x_i - A)(B - x_i)/(B - A)^2 \tag{1.52}$$

for $A \leq x_i \leq B$. The situation is more complicated when some correlation among the input variables exists. Then the simultaneous distribution of all variables x_i, $i = 1, \ldots, m$, should be specified; this will be simple only for the case of the normal distribution.

(3) Generation of random numbers: most computer languages contain a function that will generate pseudo-random numbers from rectangular distribution $R(0,1)$. For two independent random numbers, R_j, R_{j+1}, the Box-Müller transformation is used to generate two independent random numbers N_j, N_{j+1}

$$N_j = \sqrt{(-2 \ln R_j)} \times \sin(2\pi R_{j+1}) \tag{1.53}$$

$$N_{j+1} = \sqrt{(-2 \ln R_j)} \times \cos(2\pi R_{j+1}) \tag{1.54}$$

which have standardized normal distribution. The jth simulated values of the ith variable x_i will be expressed by

$$x^*_{i,j} = N_j s(x_i) + \bar{x}_i \tag{1.55}$$

For the parabolic distribution (1.52), the simulated quantity $x^*_{i,j}$ is the solution of the cubic equation

$$x^{*2}_{i,j}/2 - x^*_{i,j} - x^{*3}_{i,j}/3 + \alpha = R_i \times \beta \tag{1.56}$$

where $\alpha = A^3 - \mathrm{A}^2 + A$ and $\beta = 6/(B - A)^2$.

(4) The choice of the number of simulations: the rules for the determination of the necessary number of simulations are the same as for the determination of sample size (Section 2.7.1). The minimum number of simulations for the requested $100(1 - \alpha)\%$ confidence interval D of the mean is expressed by

$$n_{\min} = [4u_{1-\alpha/2}s^2(y)]/D^2 + 1 \tag{1.57}$$

where $u_{1-\alpha/2}$ is the quantile of standardized normal distribution and $s^2(y)$ is the estimate of variance from the first 50 simulations.

(5) The display of results: this includes a listing of an empirical probability density function of the distribution of simulated data $\{y_j^*\}$, $j = 1, \ldots, n_{\min}$, and then a calculation of the estimates of location and spread, $\bar{y}^*$ and $s(\bar{y}^*)$.

Problem 1.10 *Determination of the error of the measured viscosity*
Calculate the viscosity of glycerol by the Stokes method, from the following experimental data: the radius of the ball $r = 0.0112 \pm 0.0001$ m; density of the ball $d_0 = 1335$ kg m^{-3}, the density of glycerol $d = 1280$ kg m^{-3}, the trajectory $l = 31.23 \pm 0.05$ cm, the time $t = 62.1 \pm 0.2$ sec, and the acceleration due to gravity $g = 9.801$ m.sec^{-1}.

Program: Chemstat: Basic statistics: Error propagation.
Solution: Viscosity η determined by Stokes method is calculated from the expression $\eta = 2gr^2(d_0 - d)t/(91)$. Because this expression is not of type (1.44), the relative error cannot be calculated from a simple relationship. By the two-points method, the following values are calculated: $\bar{\eta} = 0.0299$ Pa.sec, $s(\eta) = 5.422 \times 10^{-4}$ Pa sec and the relative error $\delta(\eta) = 1.82\%$. By the Monte-Carlo simulation method $\bar{\eta} = 0.0299$ Pa sec, $s(\eta) = 5.387 \times 10^{-4}$ Pa sec, and $\hat{g}_1 = 0.038$ and $\hat{g}_2 = 2.77$.
Conclusion: The two methods, the two-point and the Monte-Carlo simulation, give the same results. The viscosity distribution is approximately symmetrical and flatter than the normal one.

1.8 SUMMARY OF DETERMINATION OF MEASUREMENT ERRORS

(1) The relative δ and absolute Δ_0 errors of signal measurement are calculated by using the expressions in Table 1.1. The sensitivity limit x_c (1.6) and the lower limit of working interval x_s (1.7) are also calculated.

(2) The absolute error of an instrument Δ consists of an instrument part Δ_v and a contribution from the variability of the analyte Δ_M, expressed by Eq. (1.12). Measurement of signal may include the additive model of errors (1.13), multiplicative model (1.15) or the model with a systematic error a (1.17). Errors come usually from the rectangular, normal, log-normal or Laplace distributions.

(3) The measurement error may be estimated with the use of the interquantile range (1.19) for a given value of statistical certainty P. For $P = 0.5$ the resulting error $\sigma_{\Delta_{0.5}}$ is termed the mean error (1.20), and for $P = 0.683$, the probable error $\sigma_{\Delta_{0.689}}$. For $P = 0.9$ the limiting quantile error $\sigma_{\Delta_{0.9,i}}$ (1.22) can be added even if the distribution of partial errors $\sigma_{\Delta_{0.9,i}}$ is not known (1.23), or (1.24)–(1.31).

(4) The moment estimate of error with the use of standard deviation enables calculation of the probable error interval (1.32) which contains all random errors with probability $P = 1 - 2\alpha$ or of the tolerance interval of error (1.33)–(1.34) which uses the estimate of standard deviation.

(5) The total error of some analytical quantity (the concentration, the content, etc.) is a result of a law of propagation of all kinds of errors concerning various experimental and instrumental operations. In addition to the classical method of Taylor series expansion (1.40)–(1.47), two computer-assisted methods may be applied, i.e. the two-point method (1.48)–(1.51) and the Monte-Carlo simulation method (1.52)–(1.57).

1.9 ADDITIONAL SOLVED PROBLEMS

Problem 1.11 *Limiting errors of ammeter*

The class of ammeter precision is declared by the ratio δ_K/δ_0 ($= 1.5/0.5$) and its range is $R = 50$ mA. Calculate the limiting absolute error Δ_0 and the limiting relative error δ_0 for an electric current of about 10 mA.

Solution: The limiting relative error (Table 1.1) is

$$\delta_0 = 1.5 + 0.5(50/10 - 1) = 3.5$$

i.e. 3% and the limiting absolute error

$$\Delta_0 = [1.5 \times 10 + 0.5(50 - 10)]/100 = 0.35 \text{ mA}$$

which is rounded off to 0.3 mA.

Conclusion: The value of electric current should be written in the form 10 ± 0.3 mA.

Problem 1.12 *Sensitivity limit of an ammeter*

Calculate the sensitivity limit of the ammeter, x_c, from Problem 1.11 and estimate the working interval x_s in which the relative error does not exceed 4%.

Solution: From Eq. (1.6), the value of x_c is

$$x_c = 2 \times 60/100 = 1.2 \text{ mA}$$

and this value is equal to the limiting absolute error. From Eq. (1.7), the lower limit of working interval is

$$x_s = 100 \times 1.2/4 = 30 \text{ mA}.$$

Conclusion: The relative precision of the ammeter is 4%, and the instrument can be used in the working interval from 30 to 60 mA.

Problem 1.13 *Systematic errors of a pipette*

A 5-ml pipette was calibrated by weighing the volume of water delivered, and 10 values were obtained. Calculate the relative and absolute systematic errors of the pipette.

Data: the volume [ml], $n = 10$, $\alpha = 0.05$: 4.969, 4.945, 5.058, 5.021, 4.945, 5.006, 4.972, 5.022, 5.013, 4.986

Program: Chemstat: Basic Statistics: One sample analysis.

Solution: The mean volume of the pipette $\bar{x}$ is 4.9937 ml with variance $s^2(x) = 0.00134$. The estimate of absolute systematic error, $\hat{a} = \bar{x} - \mu$, is -0.0063 ml and the estimate of the relative systematic error, $\delta(\hat{a}/x)$, is -0.13%. Since $\mu = 5.000$, the hypothetical 'true' value, is fixed, the variance $s^2(a) = s^2(\bar{x}) = s^2(x)/\sqrt{n}$, is equal to 0.0004. If we assume that the systematic error has a normal distribution (Section 3.3.2), then:

(1) The 95% confidence limit of the systematic error,

$$\hat{a} - t_{0.95}(10 - 1) \times s(a) \leq a \leq \hat{a} + t_{0.95}(10 - 1) \times s(a)$$

where the quantile of the Student distribution $t_{0.95}(9) = 2.263$. Therefore $-0.0325 \leq a \leq 0.0199$.

(2) The 95% confidence limit of the systematic error with statistical certainty $(1 - \alpha) = 0.99$ is equal to

$$\hat{a} - k_{\mathrm{T}} \times s(a) \leq a \leq \hat{a} + k_{\mathrm{T}} \times s(a)$$

where k_{T} is calculated from Eq. (1.34),

$$k_{\mathrm{T}} = 1.96(9/2.088)^{1/2} = 4.069$$

Therefore $-0.0534 \leq a \leq 0.0408$.

(3) For the variance of random errors of the water weights, $s^2(x)$, the 95% confidence interval of the variance $s^2(x)$ with statistical certainty $(1 - \alpha) = 0.99$ is given by Eq. (1.33) as

$$-0.1489 \leq \varepsilon \leq 0.1489$$

The limiting quantile error of the pipette

$$\sigma_{\Delta_{0.9}} = 1.65s(x) = 1.65 \times 0.0366 = 0.0604.$$

Conclusion: Since the 95% confidence interval of the systematic error and the tolerance interval of the systematic error cover the value zero, the systematic error of pipette $\hat{a} = -0.0063$ ml and $\delta(\hat{a}/\bar{x}) = -0.13\%$ may be considered not to be significant. The actual volume of the pipette is 4.994 ± 0.060 ml.

Problem 1.14 *Propagation of errors in solution preparation*
Calculate the relative error of the concentration of Fe_2O_3 in a solution which was prepared by mixing $V_1 = 5.0$ ml of the first standard solution of concentration $c_1 = 1.0$ g/l of Fe_2O_3 and $V_2 = 5.0$ ml of the second standard solution of concentration $c_2 = 2.0$ g/l of Fe_2O_3. The relative error of concentration of both standard solutions is the same, $\delta(c_1) = \delta(c_2) = 0.2\%$ and the relative error of pipetting is $\delta(V) = 0.1\%$.
Program: Chemstat: Basic Statistics: Error propagation.
Solution: The concentration of the resulting solution is

$$c = (c_1 V_1 + c_2 V_2)/(V_1 + V_2)$$

and the relative error of this concentration will be, according to Eq. (1.35)

$$\delta(c) = \delta(c_1) \times (c_1^2 + c_2^2)^{1/2}/(c_1 + c_2) = 0.149\%$$

Conclusion: When the relative error of a result is to be calculated, the summation of relative errors cannot be used and it should be considered that these variables are defined by Eq. (1.35).

Problem 1.15 *Not exceeding a declared concentration error*
Prepare $V = 100$ ml of Fe(II) solution of concentration $c_s = 5.0$ g/l. such that the relative error of this concentration will not exceed the value $\delta(c) \leq 0.1\%$. Calculate corresponding relative and absolute errors of weighing $\delta(m)$ and of standard flasks, $\delta(V)$. For the standard flask, the error $\Delta(V)$ is 0.07 ml; calculate the necessary precision of weighing.
Program: Chemstat: Basic Statistics: Error propagation.
Solution: The resulting concentration is calculated by relation

$$c = 1000 \times c_s/V$$

On the basis of the principle of the same relative influences

$$\delta(V) \simeq \delta(m)$$

As

$$\delta(V) = \delta(m) = 0.05\%$$

then

$$\Delta(V) = V \times \delta(V)/100 = 0.05 \text{ ml}$$

and

$$\Delta(m) = c \times V \times \delta(m)/100 = 2.5 \times 10^{-4} \text{ g.}$$

For $\Delta(V) = 0.07$ ml, $\delta(V) = 0.07$ %, then

$$\delta(m) = \delta(c) - \delta(V) = 0.1 - 0.07 = 0.03\%$$

and this corresponds to the absolute error of weighing

$$\Delta(m) = (5/1000) \times 100 \times 0.03/100 = 1.5 \times 10^{-4} \text{ g.}$$

Conclusion: So as not to exceed the required error of concentration of the solution, $\delta(c)$, the sum of partial relative errors of weighing and standard flasks, $\delta(m)$ and $\delta(V)$, must be less than or equal to $\delta(c)$, i.e. $\delta(m) + \delta(V) \leq \delta(c)$.

Problem 1.16 *Propagation of correlated errors in the preparation of solutions*
A mass $m = 0.1$ g of zinc was dissolved in hydrochloric acid and diluted in a standard flask with volume $V = 1000$ ml. The volume $V_1 = 100$ ml of this solution was diluted to volume $V_2 = 1000$ ml. The sample for analysis was prepared by taking $v_3 = 5$ ml and diluting to $V_4 = 25$ ml. Calculate the concentration of the resulting sample and its relative error when the standard deviation of weighing is $s(m) = 0.3$ mg, and for the standard flasks $s(V) = s(V_2) = 0.2$ ml, $s(V_1) = 0.05$ ml, $s(V_3) = 0.005$ ml and $s(V_4) = 0.025$ ml.
Program: Chemstat: Basic Statistics: Error propagation.
Solution: The concentration c is calculated from

$$c = m \times V_1 \times V_3/(V \times V_2 \times V_4)$$

Errors in volumes V_2 and V_4 are strongly correlated with errors in volumes V_1 and V_3. Consider first the ideal case when correlation coefficients

$$r(V_1 V_2) = r(V_3 V_4) = 1.$$

while other variables are uncorrelated. From Eq. (1.47)

$$\delta^2(c) \simeq (s(m)/m)^2 + (s(V)/V)^2 + (s(V_1)/V_1)^2 + (s(V_2)/V_2)^2 + (s(V_3)/V_3)^2$$
$$+ (s(V_4)/V_4)^2 - 2(s(V_1)/V_1)(s(V_2)/V_2) - 2(s(V_3)/V_3)(s(V_4)/V_4)$$

and numerically $\delta(c) = 0.302\%$.

Then, consider that the correlation between V_1 and V_2 and also between V_3 and V_4 is negligible, so that

$$r(V_1 V_2) = r(V_3 V_4) = 0$$

and then

$$\delta(c) = 0.336\%.$$

Calculation of some derivatives in Eq. (1.39) allows the mean concentration $\bar{c}$ to be estimated

$$\bar{c} = m \times V_1 \times V_3/(V \times V_2 \times V_4) + m \times V_1 \times V_3 \times [s^2(V)/(V_3 \times V_2 \times V_4)$$
$$+ s^2(V_2)/(V_2 \times V \times V_4) + s^2(V_4)/(V_4 \times V \times V_2)] - m \times V_3 \times s(V_1)$$
$$\times s(V_2)/(V \times V_2^2 \times V_4) - m \times V_1 \times s(V_3) \times s(V_4)/(V \times V_2 \times V_4^2)$$

where the first term is 2×10^{-6}, the second 2.16×10^{-12} and the third is 2.2×10^{-12}. If the two smaller terms are neglected the mean concentration will be $\bar{c} = 2$ mg/l. *Conclusion*: Correlation between volumes V_1 and V_3 and also between V_2 and V_4 decreases the relative error of the resulting sample concentration.

Problem 1.17 *Propagation of errors in gravimetry*

Iron (III) oxide in iron ore containing about 50% of Fe_2O_3 is determined gravimetrically with the use of an analytical balance with absolute error $s(m) = 0.3$ mg and the sample weight $m = 0.105$ g. Estimate the error of gravimetric determination when the sample weight m and the ash weight m_0 are related by $m_0 \simeq 0.5\ m$.

Program: Chemstat: Basic Statistics: Error propagation.

Solution: The relative mass of Fe_2O_3 in the iron ore is calculated by $\mathrm{R} = 100 \times m_0/m$. Since the sample weight m and the ash weight m_0 are strongly correlated, $r(m_0 m) \neq 0$. From Eq. (1.47) it will be

$$\delta(R) = [\delta^2(m_0) + \delta^2(m) - 2 \times \delta(m_0) \times \delta(m) \times r(m_0 m)]^{1/2}$$

In case of linear dependence,

$$m_0 = k \times m$$

and

$$r(m_0m) = 1$$

and we will have

$$\delta(R) = [(0.3/52.5)^2 + (0.3/105)^2 - 2 \times (0.3/52.5) \times (0.3/105)]^{1/2}$$
$$= 0.286\%$$

When the ash weight is not dependent on the sample weight and

$$r(m_0m) = 0$$

then

$$\delta(R) = 0.639\%$$

For a partial correlation $r(m_0m) = 0.5$, $\delta(R) = 0.49\%$.

The mean ratio $\bar{R}$ and its variance are dependent on a correlation between m and m_0. When

$$s(m_0) \simeq s(m) \simeq 0.3$$

and the measurement is repeated n times, then according to Eq. (1.39)

$$R \simeq 100[m_0/m + s^2(m)/m^3 - (r(m_0m) \times s(m))/m^2]$$

For $0 < r(m_0m) < 1$ the influence of the third term is always negligible and $\bar{R} \simeq 50\%$, and the variance of the ratio is

$$s^2(R) \simeq 10^4[s^2(m_0)/m^2 + m_0^2 \times s^2(m)/m^4 - 2 \times m_0 \times r(m_0 \times m),$$
$$s(m_0) \times s(m)/m^3 + s^2(m_0) \times s^2(m)/m^4]$$

When $r(m_0m) = 1$, then $s^2(R) \simeq 0.103$, and when $r(m_0m) = 0$, then $s^2(R) \simeq 0.102$.

For the case $r(m_0m) = 0$ the relative error $\delta(R) = 0.64\%$ and the same result is also found for $r(m_0m) = 1$.

Conclusion: Positive correlation between sample weight and ash weight decreases the relative error of the method. For sufficiently high sample weights relative to the error of weighing, the estimates $\bar{R}$ and $s^2(R)$ do not depend significantly on the degree of correlation.

Problem 1.18 *Propagation of errors in photometry*

A standard solution of iron (III) containing 0.1 mg of Fe_2O_3 in 1 ml was prepared by dissolving $m = 0.4911$ g of Mohr salt (ammonium iron (II) sulphate) in $V = 1000$ ml water. In the standard flask, with volume $V_2 = 25$ ml, the volume $V_1 = 5$ ml of this solution was diluted by salicylic acid. In the cuvette, length $l = 1.000$ cm, the absorbance $A = 1.000$ was measured by using a photometer with instrumental precision $s_{inst}(A) = 0.007$. If the errors of weighing and of the standard flask are $s(m) = 0.3$ mg, $s(V) = 0.2$ ml, $s(V_1) = 0.005$ ml, $s(V_2) = 0.025$ ml, calculate the relative error of the molar absorptivity.

Program: Chemstat: Basic Statistics: Error propagation.

Solution: Molar absorptivity ε is calculated from:

$$\varepsilon = A/cl = VV_2A/(mV_1l)$$

If there is strong correlation between the volumes V_1 and V_2, i.e.

$$r(V_1 V_2) = 1$$

Eq. (1.47) yields

$$\begin{aligned}\delta^2(\varepsilon) &\simeq (s(m)/m)^2 + (s(V)/V)^2 + (s(V_1)/V_1)^2 + (s(V_2)/V_2)^2 \\ &\quad + (\mathbf{s}_{inst}(A)/A)^2 - 2(s(V_1)/V_1)(s(V_2)/V_2) \\ &= 0.696\%.\end{aligned}$$

If correlation between V_1 and V_2 is ignored, i.e.,

$$r(V_1 V_2) = 0$$

them

$$\delta(\varepsilon) = 0.71\%.$$

Conclusion: The relative error of the molar absorptivity depends mainly on the instrumental error of the spectrophotometer used. Because $\delta(A) = 0.7\%$ the $\delta(\varepsilon)$ is also equal to 0.71%.

Problem 1.19 *Propagation of errors in the solubility of a silver salt*
The solubility product of the silver(I) salt AgX is $K_s = (4.0 \pm 0.4) \times 10^{-8}$. Estimate the error of the calculated solubility of silver(I) ions $[\mathrm{Ag}^+]$ in water.
Program: Chemstat: Basic Statistics: Error propagation.
Solution: The solubility of silver(I) ions is calculated from the expression $[\mathrm{Ag}^+] = (K_s)^{1/2}$. $[\mathrm{Ag}^+]$, $s([\mathrm{Ag}^+])$ and $\delta([\mathrm{Ag}^+])$ can be calculated in various ways:

(1) *The Taylor series*: Eq. (1.39),

$$\begin{aligned}[\mathrm{Ag}^+] &= (K_s)^{1/2} - 0.125K_s^{-3/2}s^2(K_s) = 2 \times 10^{-4} - 2.5 \times 10^{-7} \\ &= 1.9975 \times 10^{-4}\end{aligned}$$

and from Eq. (1.40)

$$s^2([\mathrm{Ag}^+]) = 0.25K_s^{-1}s^2(K_s) = 10^{-10}$$

$$s([\mathrm{Ag}^+]) = 10^{-5} \text{ and } \delta([\mathrm{Ag}^+]) = 5\%.$$

(2) *Method of two-point estimates*: $[\mathrm{Ag}^+] = 1.997 \times 10^{-4}$,

$s([\mathrm{Ag}^+]) = 1.001 \times 10^{-5}$ and $\delta([\mathrm{Ag}^+]) = 5\%$.

(3) *Monte-Carlo simulation method*: $[\mathrm{Ag}^+] = 1.997 \times 10^{-4}$,

$s([\mathrm{Ag}^+]) = 1.019 \times 10^{-5}$, $\delta([\mathrm{Ag}^+]) = 5.1\%$, $\hat{g}_1 = 0.143$ and $\hat{g}_2 = 3$.

Conclusion: All three methods of error determination give the same results. The results are illustrated in Fig. 1.5.

Problems 1.20 *Propagation of errors in solution preparation*
A standard solution of iron(II) ions was prepared by dissolving $m = 0.5458$ g of Mohr salt (the error $s(m) = 0.3$ mg) in $V = 100$ ml (standard flask). Calculate the concen-

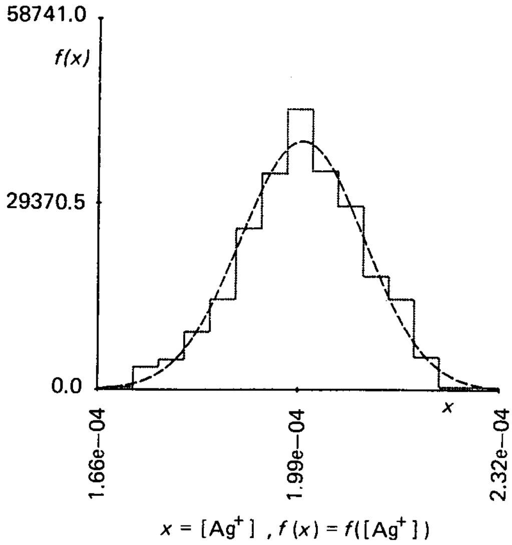


Fig. 1.5—Histogram of the solubility of silver ions $[Ag^+]$ in water.

tration c of Mohr salt in the standard solution and its relative $\delta(c)$ and absolute $s(c)$ errors.

Program: Chemstat: Basic Statistics: Error propagation.

Solution: The concentration of Mohr salt $c = 1000m/V$ (g/l). Three computation methods are compared here:

(1) *The Taylor series*: from Eq. (1.38)

$$c \simeq 100m/V + 1000m\ V^{-3}s(V) = 5.458 \text{ g/l}$$

From Eq. (1.40)

$$s^2(c) = 10^6 V^{-2} s^2(m) + 10^6 V^{-4} m^2 s^2(V) + 10^6 V^{-4}$$
$$+ s^2(V) s^2(m) = 1.377 \times 10^{-5}$$

and $s(c) = 0.0037$ g/l, $\delta(c) = 0.068\%$.

(2) *The method of two-points estimates*: $c = 5.458$ g/l, $s(c) = 0.0037$ g/l and $\delta(c) = 0.068\%$.

(3) *The Monte-Carlo simulation method*: $c = 5.458$ g/l, $s(c) = 0.0036$ g/l, $\delta(c) = 0.0662\%$, $\hat{g}_1 = 0.053$ and $\hat{g}_2 = 3.08$.

Conclusion: All three methods of error estimation lead to the same results.

REFERENCES

[1] J. R. Taylor, *An Introduction to Error Analysis*, University Science Books, Mill Valley, California, 1982.
[2] A. J. Lyon, *Dealing with Data*, Pergamon, Oxford, 1970.
[3] F. Zeleny, *Zakladni vlastnosti mericich pristroju*, SNTL, Praha, 1976.
[4] P. V. Novickij and I. A. Zograf, *Ocenka pogreshnostej rezultatov izmerenij*, Atomizdat, Moskva, 1985.
[5] G. J. Hahn and W. Nelson, *Technometrics*, 1970, **12**, 95.
[6] J. Mandel, *The Statistical Analysis of Experimental Data*, Interscience, New York, 1964.
[7] B. F. J. Manly, *Biom. J.*, 1986, **28**, 949.
[8] J. W. Müller, *Nucl. Instr. Methods*, 1979, **163**, 241.
[9] L. M. Schwartz, *Anal. Chem.*, 1975, **47**, 963.
[10] S. S. Shapiro and A. J. Gross, *Statistical Modeling Techniques*, Dekker, New York, 1981.

2

Exploratory and confirmatory analysis of univariate data

The main aim of exploratory analysis of univariate data is to isolate certain basic statistical features and patterns of data. *Exploratory data analysis (EDA)* often provides the first contact with the data and serves to uncover unexpected departures from familiar models. An important element of the exploratory approach is flexibility in responding to patterns that successive steps of analysis uncover. In brief, exploratory data analysis emphasizes flexible searching for clues and evidence, whereas *confirmatory data analysis (CDA)* stresses evaluation of the available evidence. Four major facets of exploratory data analysis stand out:

(*a*) *Revelation* through visual display meets the analyst's need to look at the behaviour of data, of diagnostic tests, of fits, and of residuals, and thus to highlight the unexpected features as well as the familiar regularities.

(*b*) *Resistance* provides insensitivity to localized misbehaviour in data. Resistant methods are influenced mostly by the main body of the data, and little by outliers. Resistance ensures that a few extraordinary data values do not unduly influence the results of an analysis. We distinguish between resistance and the related notion of robustness. Robustness generally implies insensitivity to departures from assumptions about an underlying model.

(*c*) *Residuals* focus attention on what remains of the data after some analysis, after a fitted model has been subtracted from the data; i.e. residual = measured data − calculated data.

(*d*) *Transformation* with subsequent re-expression of data involves finding a scale (e.g., logarithmic or square root) that can clarify the analysis of the data or simplify the behaviour of the data. A transformation into another scale may help to promote symmetry, constancy of variability, linearity, or additivity of effect, depending on the structure of the data.

2.1 SAMPLING, SORTING AND RANKING

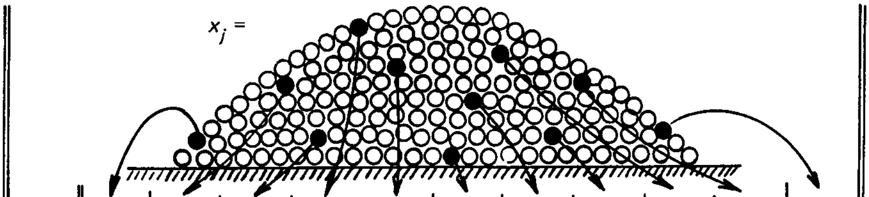
In an ideal case, the known conditions of a chemical experiment fully determine the outcome. However, in practice some factors are usually not fully controlled and others are random in nature. Observations (responses) resulting from experiments are then *random quantities*. The complete collection of all possible outcomes from a chemical experiment, if the experiment is repeated an infinite number of times, is called the *population space*. Observations represent points in this population space. The population is *discrete* when there are a finite number of possible outcomes, and *continuous* when all real values are possible in a certain interval (finite or infinite), or series of intervals. When in an experiment, only one variable is recorded, then the actual observations form a *univariate sample*. If more than one variable is obtained from a single experiment, a *multivariate sample* is obtained; e.g., if two values are obtained, the sample is *bivariate*. The aim of data analysis is to make *inferences* about *population* characteristics on the basis of a *representative random sample* of items from the population. There are several reasons why it is usual to *analyse* a representative sample from the population rather than the whole population:

(*a*) The population, although finite, may be large enough to make all possible inspections too costly, or take too long a time.
(*b*) The experiment may involve a destructive process or consumption of expensive chemicals.
(*c*) The whole population may not be available for analysis.
(*d*) The population may be infinite.

In a common practice the observation could be the result of an experiment under conditions which, for reasons outside the experimenter's control, may vary each time the experiment is repeated. The population in this case represents the set of observations that would be obtained if the experimental were repeated an infinite number of times.

A *sample* is said to be representative if it gives a sufficiently complete view of the population involved. All sample members have the same probability of being selected from the population, equal to $1/n$. If the experimenter has no prior information about the population, the only way to ensure representation is by random sampling or by impartial selection which is given the statistical term *randomization*. From the randomness of samples it then immediately follows that any judgment passed on the population on the basis of a sample is also random as well.

The process of putting a set of numbers into order is known as *sorting*. Because an ordered sample batch makes it easy to pick out the letter values, as well as to detect possible stray values at either end, sorting is important in exploratory data analysis. The sample values $x_1, \ldots, x_n$ can be sorted such that $x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}$. More formally, $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ are called the *order statistics* of the sample $x_1, x_2, \ldots, x_n$, and $x_{(i)}$ is the *i*th *order statistic* (see Fig. 2.1). On the basis of the sorting, we can define the *rank* of an observation in either of two ways: we can count up from the smallest value, or count down from the largest. The first of these yields the observation's *upward rank*



x_j	52	37	90	56	58	45	66	75	36	100	68
j	1	2	3	4	5	6	7	8	9	10	11

$x_{(i)}$	36	37	45	52	56	58	66	68	75	90	100
$R_{P_i} = i$	1	2	3	4	5	6	7	8	9	10	11
K_{P_i}	11	10	9	8	7	6	5	4	3	2	1
H_i	1	2	3	4	5	6	5	4	3	2	1

Fig. 2.1—Sampling, sorting, ranking and depth of a sample. From the population of $N = 59$ values of the melting point of wax (°C), m.p. $= 63.00 + x/100$, the random sample of $n = 11$ values is taken by selection of every fifth value. The order statistic $x_{(i)}$, the upward rank R_{P_i}, the downward rank K_{P_i}, and the depth H_i of the ith statistic, are shown.

$$R_{\mathrm{P}_i} = i$$

that is $x_{(2)}$ has upward rank 2 and, in general, $x_{(i)}$ has upward rank i. Counting down from the largest yields an observation's *downward rank*

$$K_{\mathrm{P}_i} = n + 1 - i$$

$x_{(n-1)}$ has downward rank 2, and generally, $x_{(i)}$ has downward rank $K_{\mathrm{P}_i} = n + 1 - i$. Considering both of these rankings together, we see that for any data value

$$R_{\mathrm{P}_i} + K_{\mathrm{P}_i} = n + 1$$

Sometimes it is useful to think in terms of the original observations. For example, if, through the sorting process, the raw observation x_i becomes the order statistic $x_{(j)}$, then the upward rank of x_i is j.

Often we want to give equal attention to both ends of a sample batch. A convenient way of handling this is to use the two ranks upward and downward, in defining *depth*. The depth of the ith element in a sample is the smaller of its upward rank and its downward rank.

$$H_i = \min (R_{\mathrm{P}_i}, K_{\mathrm{P}_i})$$

The depth of each data value expresses how far it is from the low end or high end of the sample.

2.2 ORDER STATISTICS, QUANTILES AND LETTER VALUES

The method of exploratory data analysis (EDA) examines certain basic features of the statistical properties of the observations (experimental data). Graphical treatment of data is used to identify the type of sample distribution, to analyse and sometimes also to re-express it. EDA is detective work which has a firm probability base and uses quantile descriptive statistics, whereas confirmatory data analysis is judicial or quasi-judicial in character.

The sample values $x_1, \ldots, x_n$ are first of all sorted into ascending order to yield $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$, the order statistics.

The P_ith *sample quantile* (or *percentile*) is defined to be the value of x below or at which $100 \times P_i\%$ of the sample values lie. The P_ith quantile is

$$\tilde{x}_{P_i} = x_{(j)} \qquad \text{where} \quad j = (n+1)P_i$$

Parameter P_i is usually termed the *cumulative* or *rank probability* and is given by

$$P_i = \frac{i}{n+1} \tag{2.1}$$

If index j is not an integer but lies between two integers m and $m + 1$, the P_ith quantile $\tilde{x}_{P_i}$ may be calculated by interpolating between $x_{(m)}$ and $x_{(m+1)}$ according to the formula

$$\tilde{x}_{P_i} = x_{(m)} + (j - m)(x_{(m+1)} - x_{(m)}) \tag{2.2}$$

For $P = 25\%$, 50%, 75% the 25th, 50th, and 75th quantiles (or percentiles) are called the *first* (or *lower*) *quartile*, the *second quartile* (or *median*) and the *third* (or *upper*) *quartile* of the sample.

The method of evaluating P_i depends on the nature of the sample distribution. The order statistics $x_{(i)}$ divide the real x-axis into $(n + 1)$ intervals, and any observation x will have the same probability $1/(n + 1)$ of appearing in any one of them. The cumulative probability is then given by

$$P_i = \frac{i}{n+1}$$

For a normal distribution the expression

$$P_i = (i - 3/8)(n + 1/4)$$

is often used, but EDA uses

$$P_i = (i - 1/3)(n + 1/3) \tag{2.3}$$

The plot of order statistics $x_{(i)}$ against the cumulative probability P_i, when $0 \leq P_i \leq 1$, for $i = 1, \ldots, n$ is called the *quantile function* $Q(P)$. This is, in fact, an inverse function of the sample distribution function. For any value α from the interval $[0, 1]$ the 100αth quantile $\tilde{x}_\alpha$ may be calculated by linear interpolation

$$\tilde{x}_\alpha = x_{(i)} + (n+1)\left[\alpha - \frac{i}{n+1}\right](x_{(i+1)} - x_{(i)}) \tag{2.4}$$

where

$$\frac{i}{n+1} \leq \alpha \leq \frac{i+1}{n+1} \tag{2.5}$$

The variance of sample quantile $\tilde{x}_\alpha$ for a sample size n is given by:

$$D(\tilde{x}_\alpha) = \frac{\alpha(1-\alpha)}{n \times [f(\tilde{x}_\alpha)]^2} \tag{2.6}$$

where $f(\tilde{x}_\alpha)$ is the value of the sample probability density function at point $\tilde{x}_\alpha$. An example of a quantile function is shown in Fig. 2.2

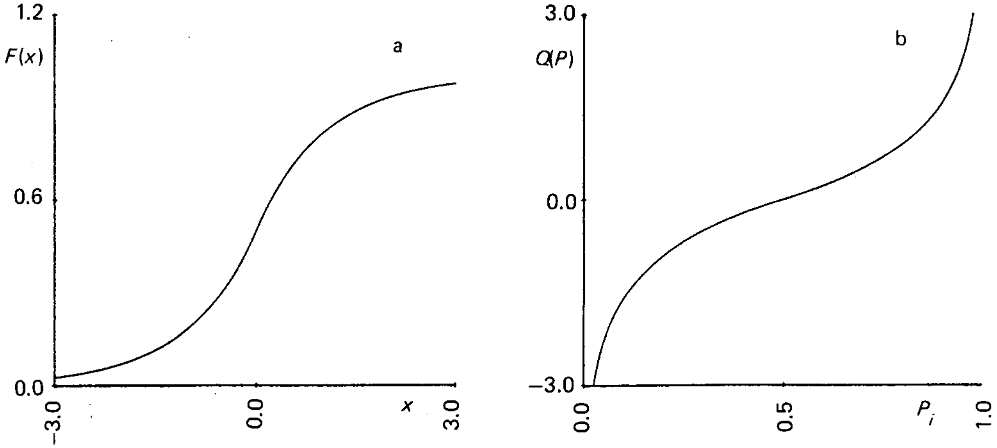


Fig. 2.2—(a) The distribution function $F(x)$, and (b) the quantile function $Q(P)$ for the Laplace distribution with a mean of zero and variance of 2.

Table 2.1—A survey of selected letter values

i	ith quantile	cumulative probability	symbol for letter value	normal quantile u_{P_i}
1	median	$2^{-1} = 0.500$	M	0
2	quartiles	$2^{-2} = 0.250$	F or H	−0.674
3	octiles	$2^{-3} = 0.125$	E	−1.15
4	sedeciles	$2^{-4} = 0.0625$	D	−1.53

Some methods of EDA are based on some selected quantiles Q being calculated for selected cumulative probabilities $P_{(i)} = 2^{-i}$, $i = 1, 2, \ldots$ These quantiles are termed *letter values* (Table 2.1).

The symbol u_{P_i} is used to denote the quantiles of the standard normal distribution $N(0, 1)$, (Section 3.3.2). The median corresponds to $i = 1$, and for each $i > 1$ there is a pair of quantiles, the lower (Q_L) and upper (Q_U) letter values. The lower letter value is calculated for a cumulative probability $P_i = 2^{-i}$ and the upper one for $P_i = 1 - 2^{-i}$.

Letter values are estimated by the *rank-and-depth* method. The *rank* of an observation is defined by counting up from the smallest value (upward rank), or by counting down from the largest (downward rank). The order statistic $x_{(i)}$ has upward rank $R_{P_i} = i$ and downward rank $K_{P_i} = (n + 1 - i)$. The depth H_i of the ith observation is defined as the lower value of the two ranks, $R_{P_i}$and K_{P_i}, i.e., $H_i = \min(R_{P_i}, K_{P_i})$, and the depth of median is given by

$$H_M = \frac{n+1}{2} \tag{2.7}$$

If H_M is an integer (i.e. n is an odd number), the median is equal to $\tilde{x}_{0.5} = M = x_{(H)}$; otherwise it is halfway between, $x_{(n/2)}$ and $x_{(n/2+1)}$. The depth of lower letter values is calculated from

$$H_Q = \frac{1 + \mathrm{int}(H_{Q-1})}{2} \tag{2.8}$$

where Q stands for the letters F, E, D, ... and int(x) means the integer part of a number x. If Q is F, then we would say that Q − 1 = M, etc. When H_Q is an integer, the lower quantile Q_L is $x_{(H_Q)}$ while the upper quantile Q_U is $x_{(n+1-H_Q)}$. When H_Q is not an integer, the following linear interpolation is carried out

$$Q_L = \frac{x_{(\mathrm{int}(H_Q))} + x_{(\mathrm{int}(H_Q))+1)}}{2} \tag{2.9}$$

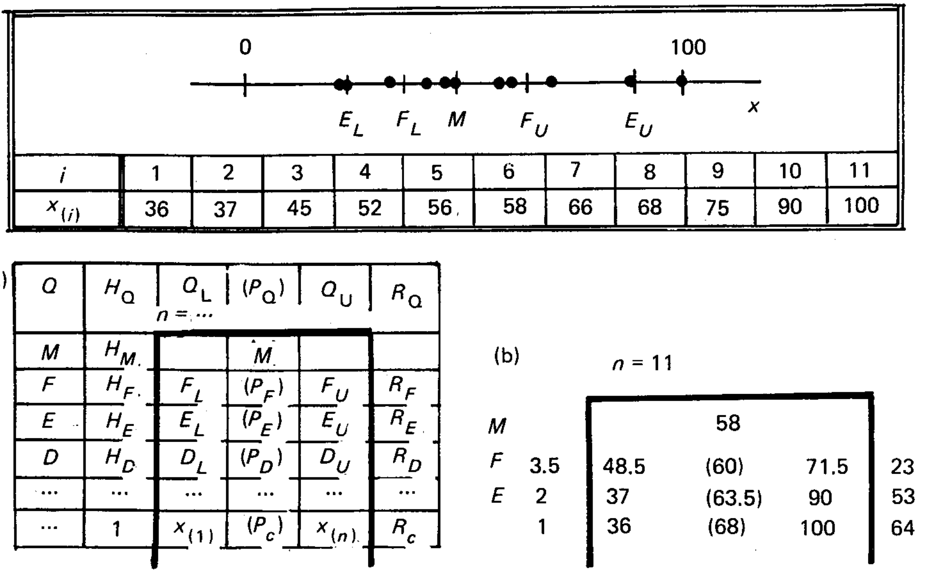
$$Q_U = \frac{x_{(n+1-int(H_Q))} + x_{(n+2-\mathrm{int}(H_Q))}}{2} \tag{2.10}$$

For lower values of H_Q and quantiles near to $x_{(1)}$ and $x_{(n)}$, the procedure based on Eqs. (2.9) and (2.10) is more robust than that based on Eq. (2.4). The number of letter values for a sample depends on the sample size. For a given sample size n, this number, which includes the median, is given by

$$n_Q \simeq 1.44 \ln(n + 1) \tag{2.11}$$

The letters used as tags for the letter values start with M for median and F for fourths (quartiles), E for eighths (octiles), etc. The extremes have no tag other than 1, their depth.

Letter values are used to provide a convenient summary of data, and the 5-number summary (1FMF1) or the 7-number summary (1EFMFE1) provide about the right amount of detail. More information is available in larger batches and we might use a fuller set of seven or more letter values if necessary (Fig. 2.3).



i	1	2	3	4	5	6	7	8	9	10	11
$x_{(i)}$	36	37	45	52	56	58	66	68	75	90	100

(a)

Q	H_Q	Q_L	(P_Q)	Q_U	R_Q
	$n = ...$				
M	H_M		M		
F	H_F	F_L	(P_F)	F_U	R_F
E	H_E	E_L	(P_E)	E_U	R_E
D	H_D	D_L	(P_D)	D_U	R_D
...	...	...	...	...	...
...	1	$x_{(1)}$	(P_c)	$x_{(n)}$	R_c

(b) $n = 11$

M			58		
F	3.5	48.5	(60)	71.5	23
E	2	37	(63.5)	90	53
	1	36	(68)	100	64

Fig. 2.3—Data summarization: (a) General construction of the skeleton of the letter-value display, (b) the letter-value display for a melting-points sample from Fig. 2.1.

Problem 2.1 *Use of the rank-and-depth method*

For the first 9 digits (1, 2, ..., 9) determine letter values and both ranks, with depth. *Solution:* The first row of Table 2.2 shows the order statistics $x_{(i)}$, the second row, the upward rank R, the third row the downward rank K and the fourth row, the depth calculated by Eq. (2.6). From Eq. (2.7), the depth of the median, $H_M = (9 + 1)/2 = 5$ and the median is equal to $M = \tilde{x}_{(H_M)} = 5$. From Eq. (2.8), the depth of both quartiles is $H_F = 3$ and of octiles $H_E = 2$. The letter values corresponding to the quartiles are $F_L = 3$ and $F_U = 7$, and to octiles $E_L = 2$ and $E_U = 8$. The letter values in Table 2.2 are in a square. The corresponding diagram is shown in Fig. 2.4.

Table 2.2—The rank-and-depth method

$x_{(i)}$	1	$\boxed{2}$	$\boxed{3}$	4	$\boxed{5}$	6	$\boxed{7}$	$\boxed{8}$	9
R	1	2	3	4	5	6	7	8	9
K	9	8	7	6	5	4	3	2	1
H_i	1	2	3	4	5	4	3	2	1

Conclusion: The rank-and-depth method allows easy determination of letter values with pencil and paper.

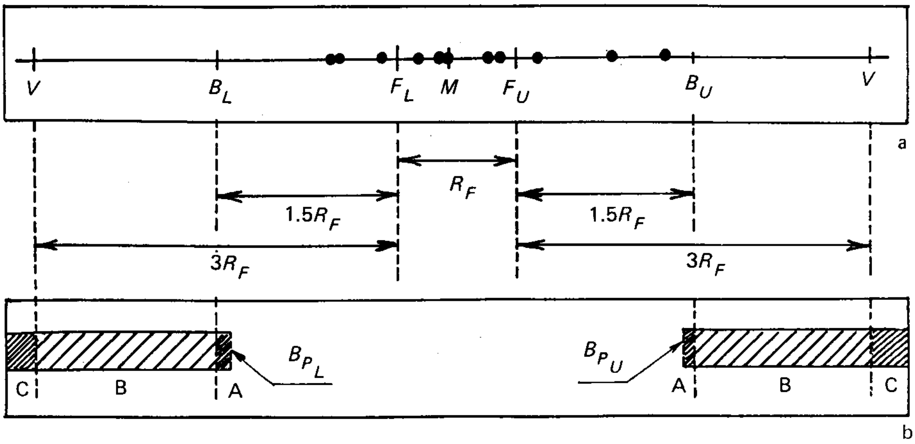


Fig. 2.4—A dot diagram showing the letter values: (a) the dot diagram with median M, F_L (lower) and F_U (upper) quartiles, inner B_L (lower) and B_U (upper) bounds, outer V_L (lower) and V_U (upper) bounds, (b) the area of outliers: A close outliers, B near far outliers, C far outliers.

2.3 PLOTS AND DISPLAYS IN EXPLORATORY DATA ANALYSIS

The basic features and statistical properties of experimental data are described by the symmetry and kurtosis of the sample distribution, the dispersion of the data, and the presence or absence of outliers. The various exploratory diagnostic plots (EDA plots G1 – G21) offer information about these statistical features of the data.

G1: Quantile plot

(*x-axis*: the cumulative (order) probability P_i, *y-axis*: the order statistic $x_{(i)}$).

The quantile plot permits identification of any peculiarities of the shape of the sample distribution, which might be symmetrical or skewed to higher or lower values.

A real sample distribution can readily be compared with the normal one, if the quantile functions for the normal distribution $Q(u_P) = \mu + \sigma u_P$ for $0 \leq P \leq 1$ is plotted on the same graph, with (1) the classical estimators of μ and $\sigma^2(\hat{\mu} = \bar{x}$ and $\sigma^2 = s^2)$ and (2) the robust estimators of μ and σ^2 $(\hat{\mu} = \tilde{x}_{0.5}$ and $\tilde{\sigma}^2 = (R_F/1.349)^2)$.

Problem 2.2 *Generation of samples from five different distributions which frequently appear in chemical data*

To demonstrate the diagnostic investigation of various samples of chemical data, samples from five different common distributions were generated. Each sample of size 50 was taken from an actual population with known population mean μ and population variance σ^2, denoted $X(\mu, \sigma^2)$:

(A) rectangular distribution $R(0.5, 1/12)$ in interval $[0, 1]$;
(B) normal distribution $N(0, 1)$;
(C) Laplace distribution $L(0, 2)$;
(D) exponential distribution $E(1, 1)$;

(E) log – normal distribution $LN(2.718, 47.209)$.

Data: (A) Sample from the $R(0.5, 1/12)$ distribution, $n = 50$

0.531	0.677	0.171	0.065	0.848	0.021	0.380	0.760	0.524	0.283
0.841	0.631	0.645	0.567	0.594	0.141	0.994	0.998	0.211	0.487
0.595	0.751	0.231	0.012	0.487	0.794	0.358	0.823	0.414	0.087
0.147	0.559	0.053	0.217	0.385	0.755	0.853	0.707	0.266	0.878
0.040	0.407	0.839	0.171	0.325	0.295	0.842	0.636	0.172	0.924

(B) Sample from the $N(0, 1)$ distribution, $n = 50$

−1.008	−0.500	0.749	1.723	0.076	0.569	−1.389	0.087
1.112	−0.235	0.519	0.279	−0.758	−0.588	−0.594	−0.885
−0.072	1.980	0.063	0.016	−0.673	−0.993	0.752	0.092
0.236	−2.962	0.109	−1.285	0.634	−0.383	1.134	−0.711
−1.825	2.374	0.500	−1.380	0.046	−0.544	−0.150	−1.129
1.173	1.401	−2.121	0.521	0.280	1.440	−0.415	−0.443
−0.384	0.690						

(C) Sample from the $E(1, 1)$ distribution, $n = 50$

0.757	1.129	0.188	0.067	1.885	0.021	0.478	1.427	0.743	0.333
1.837	0.188	0.998	1.036	0.837	0.902	0.152	5.145	6.170	0.237
0.668	0.903	1.388	0.262	0.012	0.668	2.572	1.580	0.444	1.731
0.535	0.091	0.159	0.819	0.054	0.245	0.487	1.408	1.916	1.228
0.309	2.104	0.040	0.523	1.829	0.188	0.394	0.349	1.846	1.012

(D) Sample from the $L(0, 2)$ distribution, $n = 50$

0.064	0.436	−1.072	−2.036	1.192	−3.162	−0.275	0.734
0.049	−0.569	1.144	−1.070	0.304	0.343	0.144	0.209
−1.269	4.452	5.477	−0.862	−0.026	0.210	0.695	−0.774
−3.723	−0.026	1.879	0.887	−0.333	1.038	−0.188	−1.749
−1.224	0.126	−2.249	−0.835	−0.261	0.715	1.223	0.535
−0.632	1.411	−2.538	−0.206	1.136	−1.070	−0.429	−0.529
1.153	0.319						

(E) Sample from the $LN(2.718, 47.21)$ distribution, $n = 50$

0.191	2.118	0.380	0.264	3.374	2.490	0.509	0.232
3.482	1.746	2.372	4.657	2.507	2.832	0.150	13.673
0.312	0.810	4.080	0.619	1.691	0.088	1.236	0.726
0.157	1.415	1.002	0.035	0.908	15:880	0.047	1.817
0.078	7.606	1.349	0.267	3.649	0.212	0.397	26.475
0.606	0.440	1.849	27.203	0.545	5.690	48.558	4.732
0.006	2.404						

Program: Chemstat: Basic Statistics: Exploratory continuous: QF plot.

Solution: Table 2.3 lists the statistical characteristics (see Chapter 3) of the five samples taken from five different distributions.

Table 2.3—Statistical characteristics of five distributions (upper line) and their estimates from samples of size $n = 50$ (lower line)

	Rectang	Normal	Exponent	Laplace	Lognorm
Median, $\tilde{x}_{0.5}$	0.5 0.51	0 0.03	1 0.75	0 0.02	1 1.38
Mean, $\bar{x}$	0.5 0.49	0 −0.57	1 −1.01	0 −0.025	2.718 4.08
Variance, σ^2	0.0833 0.086	1 1.088	1 1.36	2 2.43	47.209 74.53
Skewness, g_1	0 −0.048	0 −0.14	0 2.68	0 0.80	23.74 3.611
Kurtosis, g_2	1.8 1.75	3 3.37	6 11.506	9 6.10	39.48 16.795

Some of the estimated characteristics, in particular the estimates of skewness and kurtosis, differ from the corresponding population values.

The quantile plot for samples from the *R*-, *N*-, *E*-, and *L*-distributions (Fig. 2.5) shows that the *R*- and *L*-distributions give different tail lengths from the normal distribution, and the *E*-distribution is skewed to higher values.

Conclusion: The quantile plot can distinguish between different distributions because of differences in shape.

G2: Dot diagram

(*x-axis*: x values, *y-axis*: selected level, usually $y = 0$).

The dot diagram is a univariate projection of the quantile plot onto the x-axis. It is a one-dimensional scatter plot of data. The dot diagram indicates local concentrations of data, outliers, and extremes in data. A example is shown in Fig. 2.6.

G3: Jittered-dot diagram

(*x-axis*: x values, *y-axis*: a small interval of random numbers)

The jittered-dot diagram also represents a univariate projection of a quantile plot. The values of the sample points are randomly spread out in the y-direction, so this diagram gives a clearer view of the local concentration of points [2]. An example is shown in Fig. 2.6.

Problem 2.3 *Construction of dot and jittered-dot diagrams*

Construct dot and jittered-dot diagrams for the samples from the (a) rectangular, (b) normal, (c) exponential and (d) Laplace distributions.

Data: as for Problem 2.2

Program: Chemstat: Basic Statistics: Exploratory continuous.

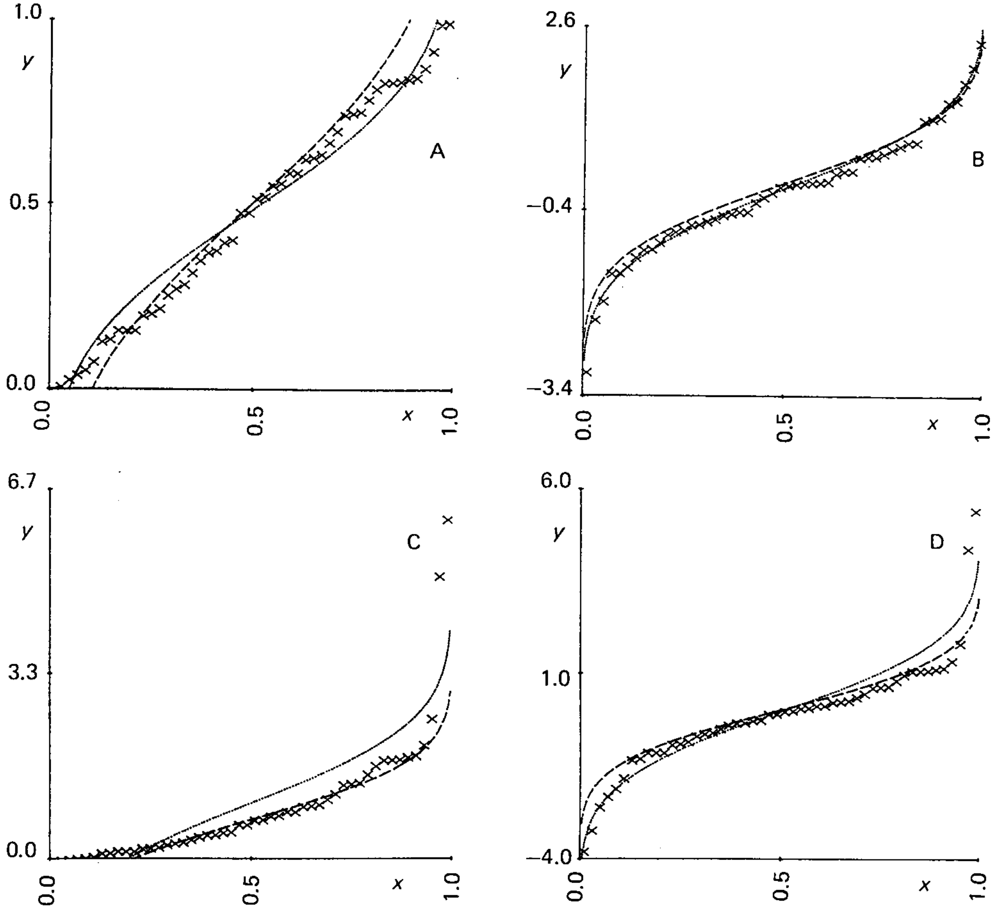


Fig. 2.5—The quantile plots (G1, robust — — — and classical . . .) of samples from four distributions: (A) rectangular, (B) normal, (C) exponential and (D) Laplace.

Solution: You should find that the diagrams indicate the obvious asymmetry in the case of the exponential distribution, and long tails in the case of the Laplace distribution.

Conclusions: The jittered-dot diagram is more informative than the dot diagram.

G4: Box-and-whisker plot

(x-*axis*: x values, y-*axis*: any suitable interval)

The box-and-whisker plot shows the *5-number summary overview* of letter values in the form of median, two quartiles (hinges) and two extremes. This plot permits determination of a robust estimate of the median M, illustrates the spread and skewness of the sample data, shows the symmetry and length of the tails of the distribution, and aids identification of outliers.

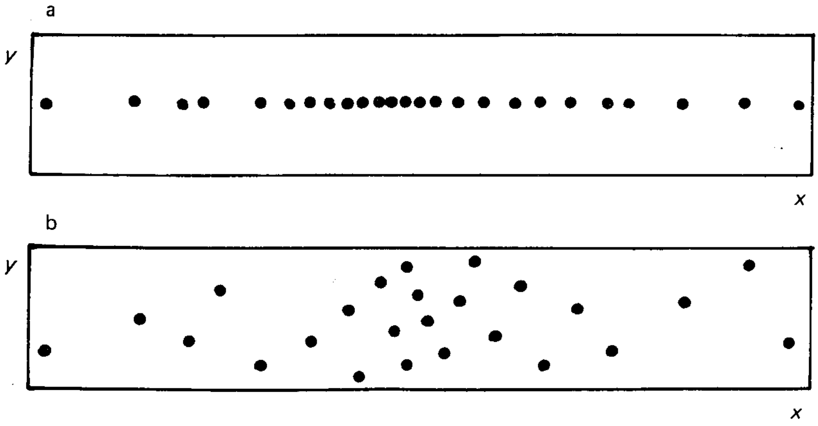


Fig. 2.6—Examples of (a) a dot diagram (G2) and (b) a jittered-dot diagram (G3).

The letter values are shown graphically in this plot. The skeletal box-and-whisker plot has a length from lower quartile F_{L} to upper F_{U} quartile equal to:

$$R_{\mathrm{F}} = F_{\mathrm{U}} - F_{\mathrm{L}} = \tilde{x}_{0.75} - \tilde{x}_{0.25}$$

and the width is proportional to $\sqrt{n}$.

The position of the median is marked by a vertical crossbar inside the box. The classical box-and-whisker plot is then completed by drawing lines (whiskers) out from each quartile to the corresponding extreme values $x_{(1)}$, $x_{(n)}$ at the ends of the order statistics. An example is shown in Fig. 2.7.

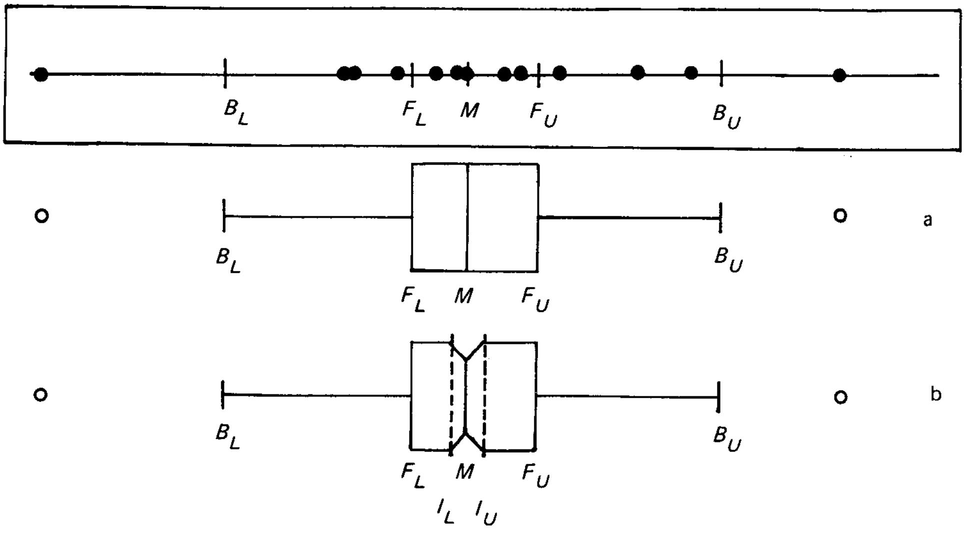


Fig. 2.7—Construction of (a) the box-and-whisker plot (G4) and (b) the notched box-and-whisker plot (G5) from the dot diagram. Empty circles indicate outliers.

This plot is useful in illustrating skewness of a sample. If the distribution has a long tail to the right (*positive skew*) the right-hand section of the box will be longer than the left, and upper extreme point will be further from the median than the lower extreme. The converse will be true if the distribution has *negative skew* with its longer tail to the left.

In the modified box-and-whisker plot, the whiskers are terminated by the "adjacent" values B_{PU} and B_{PL}. These values lie just within the *inner bounds* defined by the cutoffs B_U and B_L, which are given by:

$$B_U = F_U + 1.5\ R_F \tag{3.12a}$$

$$B_L = F_L - 1.5\ R_F \tag{2.12b}$$

For a sample from a normal distribution, $B_U - B_L \simeq 4.2$. The probability that data lie outside this interval is 0.04. Observations outside the inner bounds (smaller than B_L or larger than B_U) are probable *outliers*, and are marked on the G4 plot by circles (Fig. 2.7)

G5: Notched box-and-whisker plot

(*x-axis*: x values, *y-axis*: any suitable interval)

An analogue of the box-and-whisker plot is the notched box-and-whisker plot, which facilitates examination of the variability of the median. The median variability is expressed by notches given by the robust confidence interval $I_L \leq M \leq I_U$, where the lower and upper limits are

$$I_L = M - 1.57\ R_F/\sqrt{n} \tag{2.13a}$$

$$I_U = M + 1.57\ R_F/\sqrt{n} \tag{2.13b}$$

The notches I_L and I_U are placed symmetrically around the median. The properties of the notched box-and-whisker plot are similar to those of the G4 plot.

Problem 2.4 *Construction of box-and-whisker and notched box-and-whisker plots*

Construct a box-and-whisker plot and a notched box-and-whisker plot for the samples from (A) rectangular, (B) normal, (C) exponential and (D) Laplace distributions.

Data: as for Problem 2.2

Program: Chemstat: Basic Statistics: Exploratory continuous.

Solution: The box-and-whisker and the notched box-and-whisker plots in Fig. 2.8 indicate the asymmetry of the exponential distribution (C), and probable outliers in the samples from the normal (B), Laplace (C) and strongly skewed exponential (C) distribution.

Conclusion: The two plots, G4 and G5, can demonstrate asymmetry of sample distributions and outliers in data.

The main statistical features of a sample distribution are examined by comparing the asymmetry and tail lengths with those of the normal (Gaussian) one. The skewness and kurtosis can be characterized at various distances from the median by the following statistical characteristics:

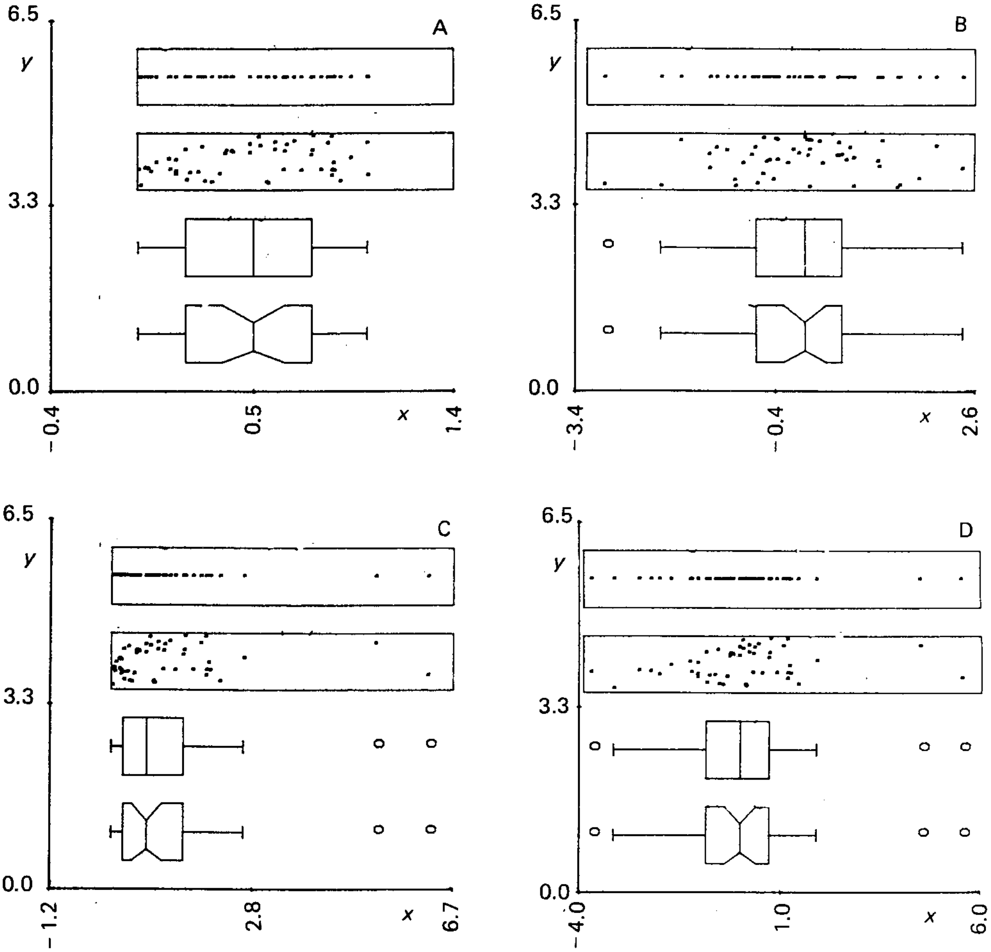


Fig. 2.8—Dot diagrams (G2), jittered-dot diagrams (G3), box-and-whisker plots (G4) and notched box-and-whisker plots (G5) for the samples from the (A) rectangular, (B) normal, (C) exponential and (D) Laplace distributions.

the midsum $Z_Q = (L_L + L_U)/2$

the interquantile range $R_Q = L_U - L_L$

the skewness $S_Q = (M - P_Q)/R_Q$

the pseudosigma $G_Q = R_Q/(-2u_{P_i})$

where u_{P_i} is the quantile of the standardized normal distribution for $P = 2^{-i}$ (Section 3.3.2); and

the length of tails $T_Q = \ln (R_Q/R_F)$

These characteristics are summarized in Table 2.4.

Table 2.4—Characteristics of a distribution shape

Characteristic	Used for	Valid for L
Midsum Z_Q	symmetry (at $P_Q = 0$)	F, E, D, . . .
Interquantile range R_Q	spread	F, E, D, . . .
Skewness S_Q	symmetry (at $S_Q = 0$)	F, E, D, . . .
Pseudosigma G_Q	kurtosis (for Gaussian distribution G_Q = const.)	F, E, D, . . .
Tail lengths T_Q	kurtosis	E, D

For any symmetric distribution, the theoretical length of tails, T_E and T_D, can be computed: for the normal distribution, $T_E = 0.534$ and $T_D = 0.822$, for the rectangular distribution $T_E = 0.405$ and $T_D = 0.559$, and for the Laplace distribution, $T_E = 0.693$ and $T_D = 1.098$.

The skewness S_Q has negative values, for distributions skewed to higher values and positive values for distributions skewed to lower values. For distributions with longer tails than the normal, the values of pseudosigma G_Q increase with the distance from the median. When the values of pseudosigma G_Q decrease with the distance from a median, the sample distribution has shorter tails than the normal.

To examine all statistical features of the sample, various plots of characteristics from Table 2.4 are used. For large samples, the letter values are examined, whereas for small samples the quantile $\tilde{x}_{P_i} = x_{(i)}$ usually for $P_i = (i - 1/3)/(n + 1/3)$, is used.

G6: Midsum plot

[*x-axis*: the order statistic $x_{(i)}$; *y-axis*: the midsum $Z_i = (x_{(n+1-i)} + x_{(i)})/2$]

The midsum plot gives information about the symmetry of a distribution. For a symmetrical distribution, the midsum plot forms a horizontal line $y = M$.

Problem 2.5 *Construction of the midsum plot*

Construct the midsum plot for the samples from the (A) rectangular, (B) normal, (C) exponential, and (D) Laplace distributions.

Data: as for Problem 2.2

Program: Chemstat: Basic Statistics: Exploratory continuous.

Solution: The midsum plot for sample (C) (exponential distribution) in Fig. 2.9 indicates that it deviates from a symmetrical distribution.

G7: Symmetry plot

[*x-axis*: the quantile $u^2_{P_i}/2$ for $P_i = i \quad (n + 1)$; *y-axis*: the midsum $Z_i = (x_{(n+1-i)} + x_{(i)})/2$]

For a symmetrical distribution, the symmetry plot forms the horizontal line $y = M$. When this line has non-zero slope, the slope gives an estimate of skewness [3].

Problem 2.6 *Construction of the symmetry plot*

Construct the symmetry plot for the samples from (A) rectangular, (B) normal, (C) exponential, and (D) Laplace distributions.

Data: as for Problem 2.2

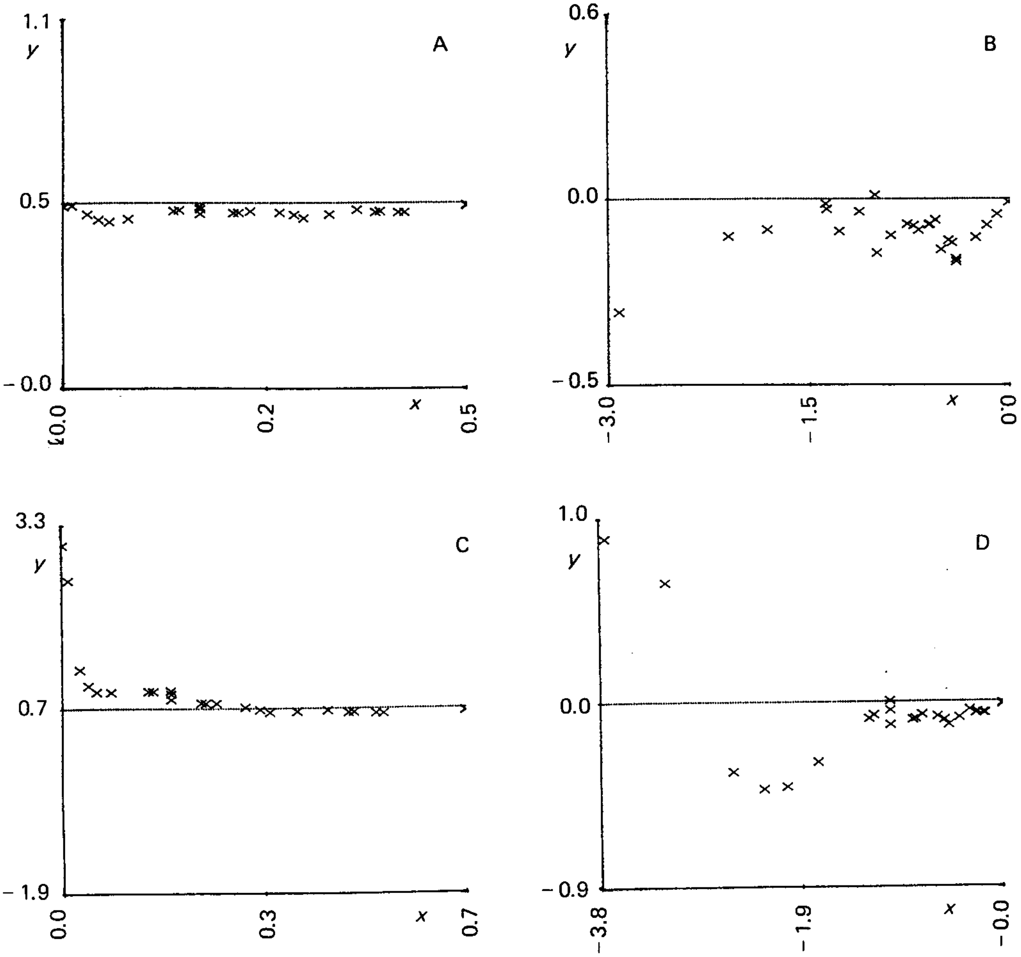


Fig. 2.9—The midsum plot (G6) for samples from the (A) rectangular, (B) normal, (C) exponential and (D) Laplace distributions.

Program: Chemstat: Basic Statistics: Exploratory continuous.
Solution: The symmetry plot is interpreted in the same way as the midsum plot.
Conclusion: From the slope of the line for sample (C) in Fig. 2.10, the skewness can be estimated to be equal to 2.

G8: Kurtosis plot

(*x-axis*: the quantile $u_{P_i}^2/2$ for $P_i = i/(n+1)$;
y-axis: the quantity $\ln [(x_{(n+1-i)} - x_{(i)})/(-2u_{P_i})]$)
The kurtosis plot indicates the peakedness of a distribution. For a normal distribution the kurtosis plot gives a horizontal line. When the line has a non-zero slope, the value of the slope gives an estimate of kurtosis [3].

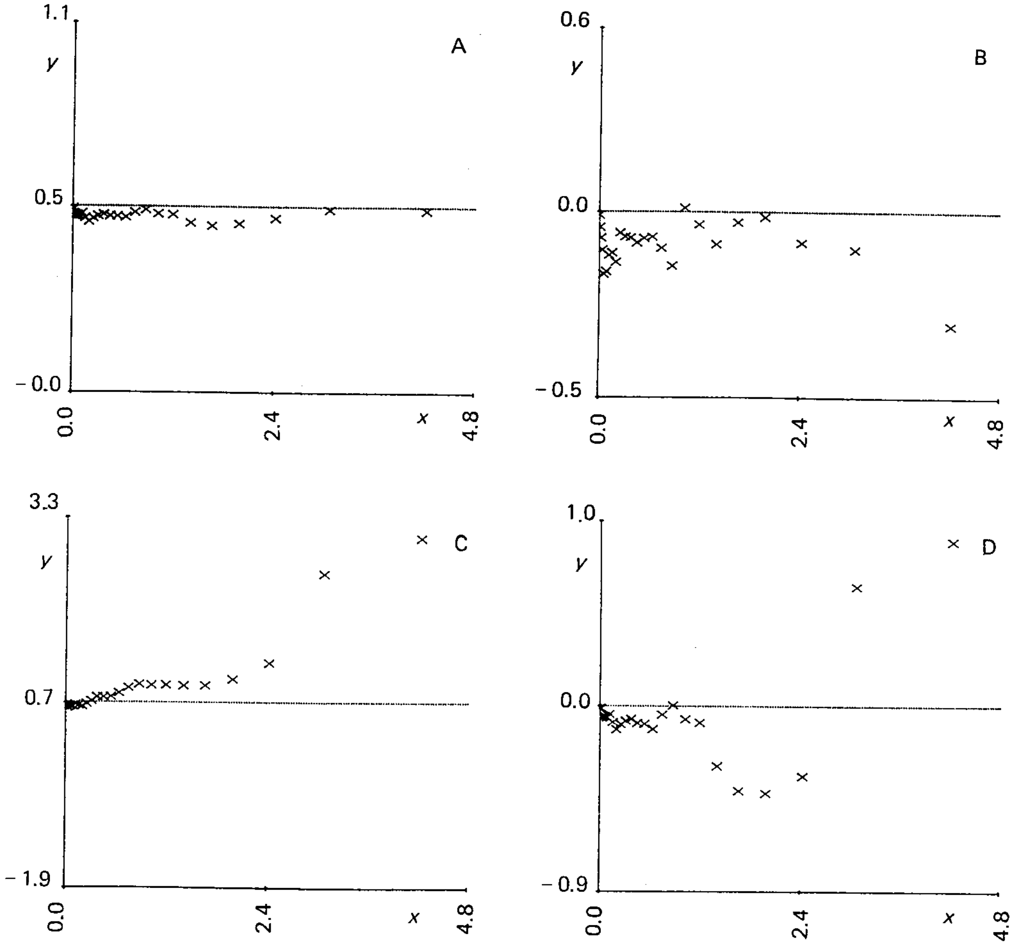


Fig. 2.10—The symmetry plot (G7) for samples from the (A) rectangular, (B) normal, (C) exponential, and (D) Laplace distributions.

Problem 2.7 *Construction of a kurtosis plot*
Construct the kurtosis plot for the samples from the (A) rectangular, (B) normal, (C) exponential, and (D) Laplace distributions.

Data: as for Problem 2.2

Program: Chemstat: Basic Statistics: Exploratory continuous.

Solution: As can be seen in Fig. 2.11, significant systematic deviations from the normal distribution are indicated in the case of a symmetric distribution with short tails [sample (A)] and with long tails [sample (D)].

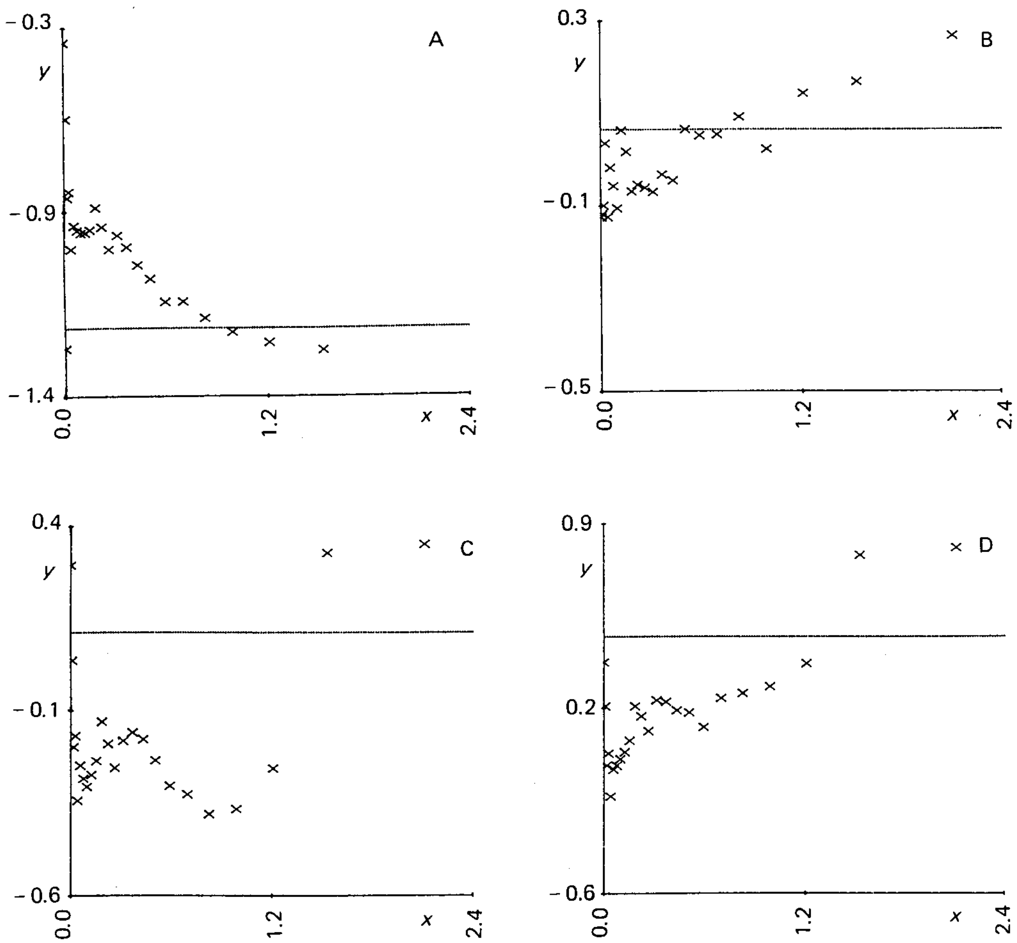


Fig. 2.11—The kurtosis plot (G8) for samples from the (A) rectangular, (B) normal, (C) exponential and (D) Laplace distributions.

G9: The differential quantile plot

(x-*axis*: the quantile u_{P_i}; y-*axis*: the deviation of order statistics $d_{(i)} = x_{(i)} - \tilde{s}u_{P_i}$)
The differential quantile plot compares the sample distribution with the normal one. The statistic $\tilde{s}$ represents the robust estimate of the standard deviation, calculated for example, by the use of the interquantile range. A horizontal line indicates a symmetrical distribution with tails similar to the normal one.

Problem 2.8 *Construction of a differential quantile plot*

Construct the differential quantile plot for the samples from the (A) rectangular, (B) normal, (C) exponential, and (D) Laplace distributions.
Data: as for Problem 2.2

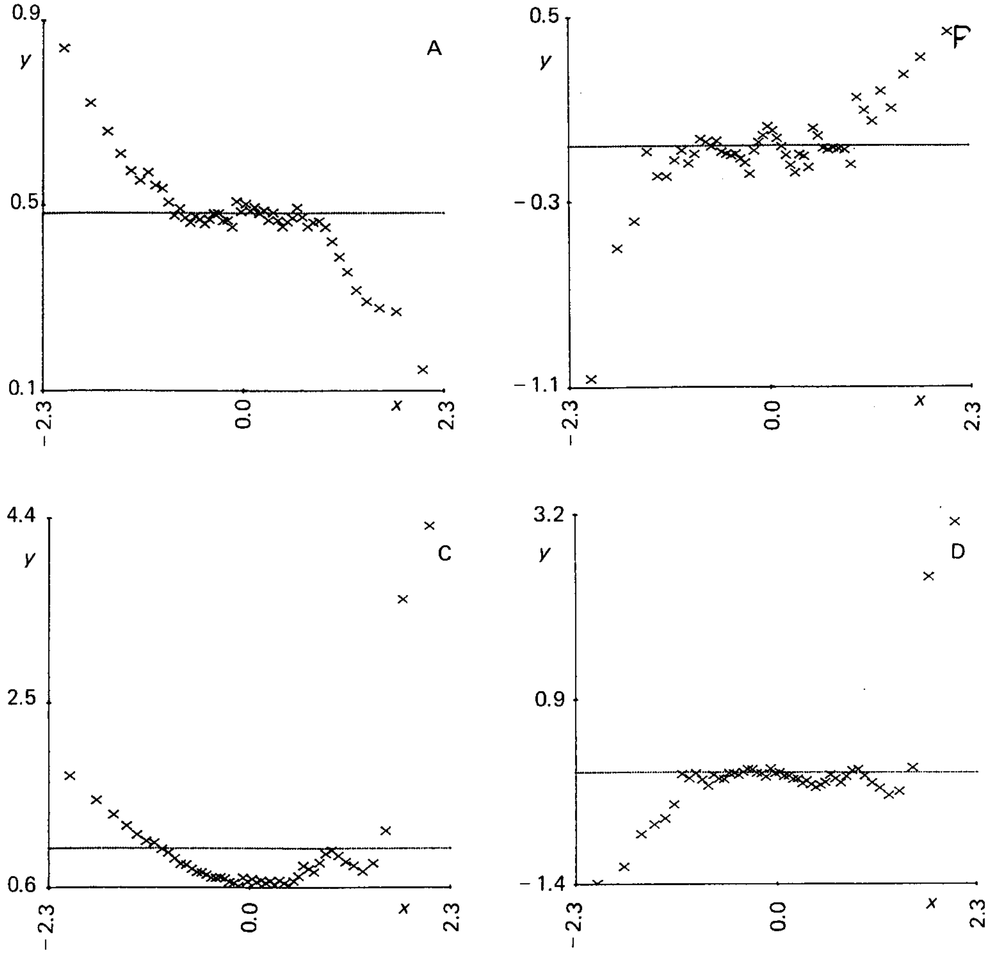


Fig. 2.12—The differential quantile plots (G9) for samples from the (A) rectangular, (B) normal, (C) exponential, and (D) Laplace distributions.

Solution: Figure 2.12 shows the differential quantile plots.
Conclusion: The deviation from normality is smallest for sample (B).

G10: Quantile–box plot

(*x-axis*: the order probability P_i, *y-axis*: the order statistic $x_{(i)}$)
The quantile–box plot (Fig. 2.13) is a simple and universal tool for examining the statistical features of data. The plot is based on the estimate of a sample quantile function formed by connecting points $\{x_{(i)}, P_i\}$ by straight lines. P_i is calculated from

$$P_i = (i - 1/3)/(n + 1/3).$$

For symmetrical distributions, the sample quantile function has a sigmoid shape, whereas for an asymmetrical one, the quantile function is convex or concave increasing. For easier interpretation the following quantile boxes are included on the graph:

(a) *The quartile box F* has on the y-axis two vertices given by quartiles F_L and F_U with corresponding values on the x-axis equal to the cumulative probability values

$$P_2 = 2^{-2} = 0.25 \qquad \text{and} \qquad 1 - 2^{-2} = 0.75.$$

(b) *The octiles box E* has, on the y-axis, the octiles E_L and E_U and on the x-axis the cumulative probabilities

$$P_3 = 2^{-3} = 0.125 \qquad \text{and} \qquad 1 - 2^{-3} = 0.875.$$

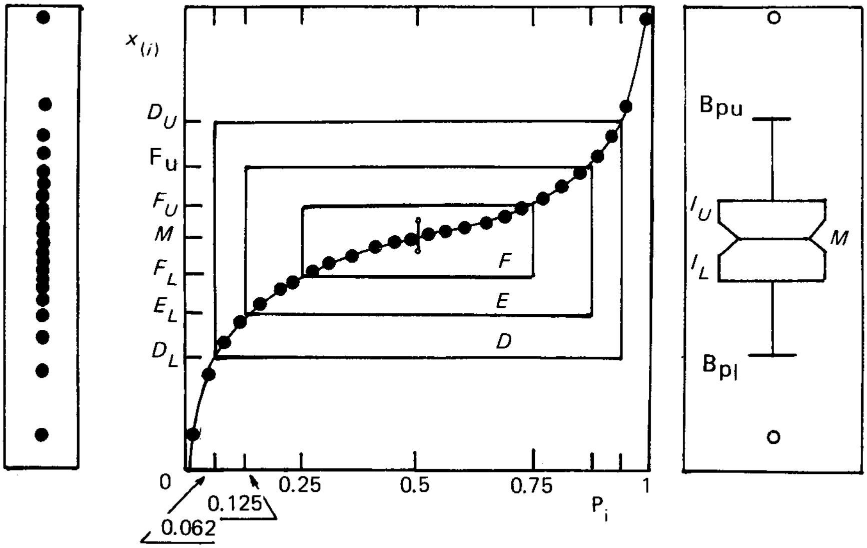


Fig. 2.13—An example of a quantile-box plot (G10). The dot diagram (left) and the notched box-and-whisker plot (right) are given for comparison.

(c) *The sedeciles box D* has on the y-axis the sedeciles D_L and D_U and on the x-axis the cumulative probabilities

$$P_4 = 2^{-4} = 0.0625 \qquad \text{and} \qquad 1 - 2^{-4} = 0.9375.$$

The position of the median M is marked by a horizontal line inside the quartile box. The robust estimate of confidence interval of the median $M \pm 1.57\ R_F/\sqrt{n}$, is drawn as a vertical line at $P = 0.5$. From this plot, and from the estimates of the midsum Z_Q, the interquantile range R_Q, the relative skewness S_Q and the relative lengths of tails T_Q, the following may be stated about the sample distribution:

(1) *A symmetric unimodal sample distribution* contains individual boxes arranged symmetrically inside one another, and the value of relative skewness is close to zero, $S_Q \simeq 0$. When the tail lengths T_Q are approximately equal to their theoretical values for a particular distribution, then the normal distribution, the Laplace (long tails) and the rectangular distribution (short tails) may be distinguished.

(2) *An asymmetric sample distribution.* In the case of a distribution skewed to higher values, there are significantly shorter distances between the lower than between the upper parts of the boxes. The skewness S_Q then has a negative value. For a distribution skewed to lower values, the skewness S_Q is positive.

(3) *Outliers* are indicated by a sudden increase of the quantile function outside the F box; the slope may approach infinity.

(4) *A multimodal sample distribution* is indicated by several parts of the quantile function inside box F reaching zero slope.

The quantile–box plot is one of the most useful diagnostics of exploratory data analysis. The sample values are not transformed and all the original information about the data is available.

Problem 2.9 *Construction of a quantile–box plot*
Construct quantile–box plots for samples from the (A) rectangular, (B) normal, (C) exponential, and (D) Laplace distributions.
Data: as for Problem 2.2
Program: Chemstat: Basic Statistics: One sample analysis or Exploratory continuous.
Solution: Significant differences in the tail lengths of symmetrical distributions (A, B, D) and obvious skewing in the case of the exponential distribution (C) can be observed. For a distribution with long tails (e.g. D), it is difficult to recognize outliers (Fig. 2.14). Table 2.5 lists the estimates of skewness S_Q and tail lengths T_Q in the region of octiles and sedeciles.

Table 2.5—The estimates of sample skewness S_Q and sample kurtosis expressed by the tail lengths T_Q for four samples

	Sample skewness		Tail lengths	
	Octiles E	Sedeciles D	Octiles E	Sedeciles D
A	0.040	0.031	0.318	0.478
B	0.027	0.020	0.633	0.990
C	−0.160	−0.230	0.407	0.736
D	0.076	0.084	0.565	1.007

Conclusion: The samples taken from the normal and Laplace distribution do not differ significantly in the octile and sedecile values.

2.4 EXAMINING A SAMPLE DISTRIBUTION BY EDA

The first step in any data examination is to summarize the information contained in the data. EDA can perform this step in two ways: (a) by use of an appropriate picture

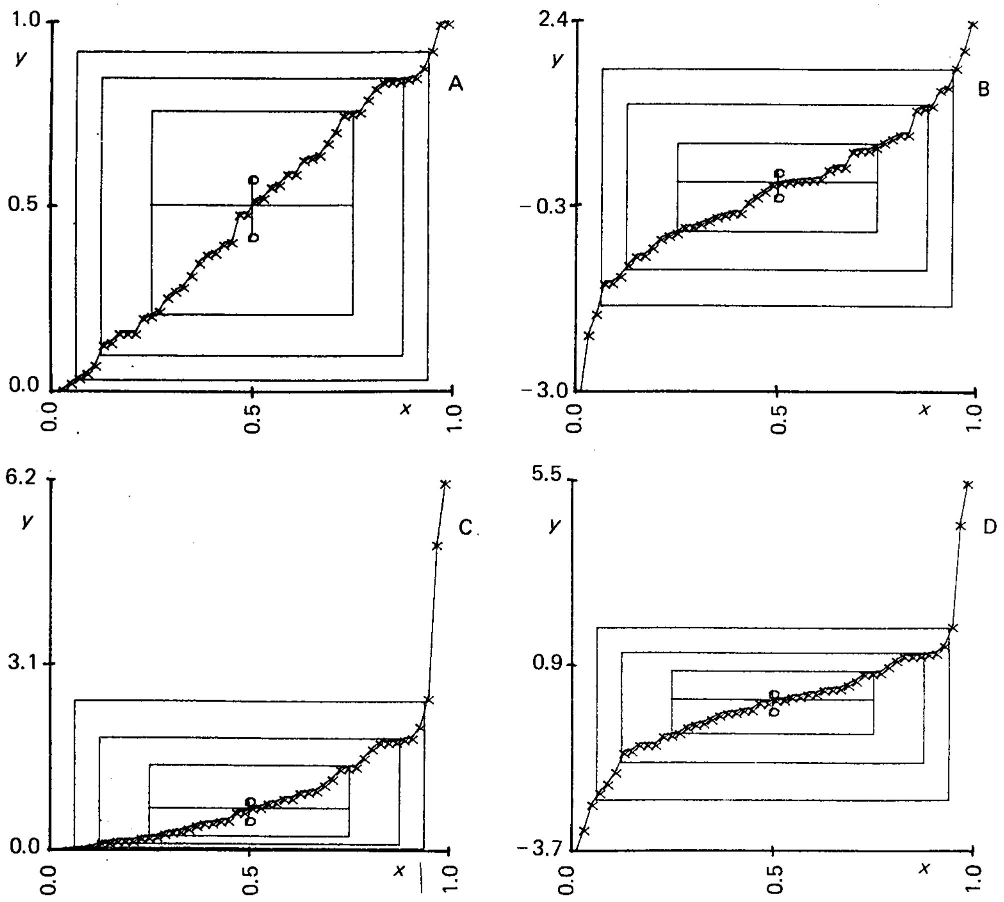


Fig. 2.14—The quantile–box plot (G10) for samples from (A) rectangular, (B) normal, (C) exponential and (D) Laplace distributions.

or display, or (b) by calculation of characteristics from the data which indicate certain basic features.

Some graphical displays can show overall patterns or trends. They can also reveal surprising, unexpected, or amusing features of data that might otherwise go unnoticed. When a large number of observations is available, the estimation of the *probability density function* or other function characterizing the data distribution can help to elucidate the structure of the sample.

In order to elucidate the structure of a large sample we can divide the range covered by the sample into a number of classes, usually of equal length, and then count the number of members f_i of the sample falling into each class. In this way the sample is reduced to a grouped sample characterized by frequencies f_i and mid-point of classes x_i^+, $i = 1, \ldots, k$, $k < n$. Grouping generally leads to a drop in the information content, especially with small and medium sample sizes.

The number of classes taken in forming a grouped sample is to some extent arbitrary. As the class width is reduced, a situation is eventually reached in which there is a frequency f_i of either 1 or 0 in some class. On the other hand if the class width is increased, the observations in the sample will fall into fewer and fewer classes and the picture of the structure of the sample presented will become cruder and less informative. As a compromise between these extremes we use, for example, with a sample of size 100, about 10 classes. Some empirical rules for choosing optimal number of classes are discussed in section G13.

Usually the class widths are chosen to be the same, but this is not essential. If the tails of the distribution contain only a few members of the sample it may be convenient to take wider classes in the tails than in the rest of the range.

G11: Stem-and-leaf display

The stem-and-leaf display shows

(a) the range of values covered by the data;
(b) where the values are concentrated;
(c) how symmetric the sample is;
(d) whether there are gaps where no values were observed; and
(e) whether any values stray markedly from the rest.

A working stem-and-leaf display is constructed such that the numerical values of the observations in each class of the distribution are divided into two parts, (1) the *stem* which consists of all the digits common to the members of the class, and (2) the *leaf*, which consists of the remaining digits. The stems are then written as a column with the smallest at the top, and the leaves are written on the same lines as their stems to give an ordered *stem-and-leaf display.* The leaves in each of the rows may be ordered to give an ordered stem-and-leaf display.

Problems 2.10 *Construction of an ordered stem-and-leaf display*

Construct an ordered stem-and-leaf display for the data sample of the weights of 100 aspirin tablets in 8 classes.

Data: The weights of aspirin tablets [10] grouped into 8 classes, are, to the nearest mg:

	Class	Class boundary	Class mid-value	Tally marks	Freq.
1	0.324–0.325	0.3235–0.3255	0.3245	111	3
2	0.326–0.327	0.3255–0.3275	0.3265	11111 11	7
3	0.328–0.329	0.3275–0.3295	0.3285	11111 11111 11111 111	18
4	0.330–0.331	0.3295–0.3315	0.3305	11111 11111 11111 11111 111	23
5	0.332–0.333	0.3315–0.3335	0.3325	11111 11111 11111 11111 1111	24
6	0.334–0.335	0.3335–0.3355	0.3345	11111 11111 11111	15
7	0.336–0.337	0.3355–0.3375	0.3365	11111 11	7
8	0.338–0.339	0.3375–0.3395	0.3385	111	3
Total					100

Solution: The stem-and-leaf display is as follows:

	Class	
Stem		Leaves
0.32	(4, 5)	545
0.32	(6, 7)	77777 77
0.32	(8, 9)	99998 89988 88998 998
0.33	(0, 1)	00100 00100 01010 10111 100
0.33	(2, 3)	33322 23332 32233 22222 2222
0.33	(4, 5)	54444 55554 45445
0.33	(6, 7)	76776 77
0.33	(8, 9)	988

G12: Kernel estimation of probability density

[*x-axis*: the variable x, *y-axis*: the probability density $f(x)$]
Let x be a continuous random variable. The statistical properties of x may be determined by specifying the *probability density function* of x (also termed the frequency function), $f(x)$ say. A computer may be used to estimate the kernel of the sample probability density function $\hat{f}(x)$ for small and medium samples from:

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left[\frac{x - x_i}{h}\right] \tag{2.14}$$

In this equation h is bandwidth, which controls the smoothness of $\hat{f}(x)$, and $K(x)$ is the kernel function, which is symmetric around zero, and also has the properties of a frequency function. The actual choice of shape for the kernel function is not important, so here we consider a bi-quadratic kernel estimate

$$K(x) = \begin{cases} 0.9375\ (1 - x^2)^2 & \text{for } -1 \leq x \leq 1 \\ 0 & \text{for } x \text{ outside } [-1;\ 1] \end{cases} \tag{2.15}$$

The quality of the kernel estimate $\hat{f}(x)$ is controlled mainly by the selection of parameter h. If h is too small, the estimate is too rough; if it is too large, the shape of $\hat{f}(x)$ is flattened too much. For samples taken from a normal distribution, the optimal bandwidth h can be calculated from an expression suggested by Scott and Sheater [4]

$$h_{\text{opt}} = 2.34\ \sigma n^{-0.2} \tag{2.16}$$

Lejenne, Dodge and Koelin[5] recommend the following procedure for construction of the kernel estimate of the probability density function.

(1) From Eq. (2.14), calculate an initial guess for the probability density function $\hat{f}(x)^{(0)}$ with the bandwidth

$$h^{(0)} = 0.75 \times (n/100)^{-0.2} \times [x_{(i+\text{int}(n/2))} - x_{(i)}],$$

then calculate the kernel function $K(x)$ from Eq. (2.15).

(2) Find the final estimate of the probability density function with the kernel function (2.15) and non-constant bandwidth from

$$\hat{f}(x)^{(k)} = \frac{1}{n} \sum_{i=1}^{n} K\left[\frac{x - x_i}{h_i}\right] \tag{2.17}$$

Here the local bandwidth h_i is calculated from

$$h_i = h^{(0)} \times [\hat{f}(x_i)^{(0)}/\max \hat{f}(x_i)^{(0)}]^{-\alpha} \tag{2.18}$$

Parameter α is defined in the interval [0, 1] and controls the smoothness of $\hat{f}(x)$. Higher values of α lead to a smoothed estimate $\hat{f}(x)$. The parameter α is usually chosen to be equal to 1/3. For complex sample distributions, it is useful to construct $\hat{f}(x)$ with various values of α and select the one corresponding to maximal visual smoothness.

Problem 2.11 *Construction of the kernel estimate of the probability density function*

Construct the kernel estimate of the probability density function for samples from the (A) rectangular, (B) normal, (C) exponential, and (D) Laplace distributions.
Data: as for Problem 2.2
Solution: The kernel estimates for the four sample probability density functions are constructed with the use of Eqs. (2.15), (2.17), (2.18), for $\alpha = 0.7$ (Fig. 2.15).
Conclusion: The kernel estimates of the probability density function are in a good agreement.

G13: Histogram, frequency polygon, bar chart and rootogram
[*x-axis*: the variable x; *y-axis*: the probability density function $\hat{f}(x)$]
The histogram is one of the oldest classical representations of grouped frequency distributions. The vertical axis represents roughly the class frequency, and the class mid-values x_i, $i = 1, \ldots, k$, are plotted on the horizontal axis. With the class mid-value x_i as the centre of its base, a vertical bar of width equal to the class width and height equal to an empirical relative frequency f_i, is drawn for each of the classes.

If the class widths Δx_i are not all equal, a histogram constructed by the above method will give a distorted picture of the distribution—it will overemphasize the contributions of the classes with the larger widths. In this situation the correct histogram is constructed with $f_i/\Delta x_i$ along the vertical axis instead of f_i. When the Δx_i are all the same ($= \Delta x$, say) the shape of the histogram will be the same whether f_i or $f_i/\Delta x$ is plotted against x_i.

In an *ungrouped* data sample, the class boundaries x_j^*, $j = 1, \ldots, L + 1$, and the number of classes L should be defined. Then the jth class has two boundaries,

$$x_j^* \leq x \leq x_{j+1}^*,$$

and their difference represents the class width,

$$\Delta x_i = x_{j+1}^* - x_j^*.$$

The quality of a histogram will depend on the width of the classes used. For approximately symmetric distributions, a suitable number of classes $L$is given by

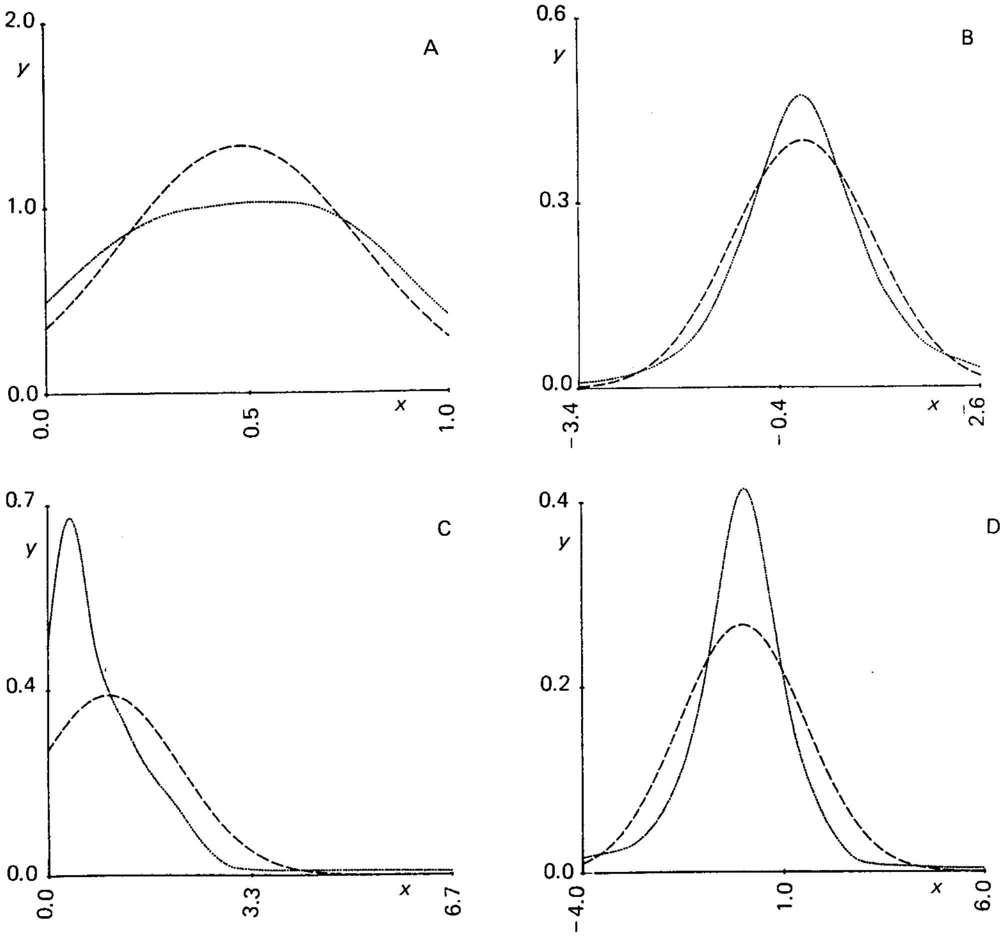


Fig. 2.15—The kernel estimate of the probability density function (...) and the Gaussian (— — —) function for samples from the distributions: (A) rectangular, (B) normal, (C) exponential, and (D) Laplace.

$$L = \text{int}(2\sqrt{n})$$

where int(x) is the integer part of a number x. For a large range of sample sizes

$$L = \text{int}[2.46 \times (n - 1)^{0.4}]$$

may also be used. For samples from the normal distribution, the optimal class width is

$$\Delta x_{\text{opt}} = 3.49\ s/n^{1/3}$$

where s is the standard deviation. A robust estimate of class width for approximately normal data is

$$\Delta x_{\text{rob}} = 2(F_{\text{U}} - F_{\text{L}})/n^{1/3}$$

where F_U and F_L are the upper and lower values of sample quantiles.

For more complicated shapes of sample distribution, the number of classes should be increased, or some special technique of classes with a non-constant length can be used.

If the class boundaries for all classes, x_j^*, are known, the histogram is calculated from

$$\hat{f}(x) = \frac{1}{n(x_{j+1}^* - x_j^*)} C(x_j^*, x_{j+1}^*) \qquad \text{for} \quad x_j^* \le x \le x_{j+1}^*$$

where $C(x_j^*, x_{j+1}^*)$ is a function equal to the number of sample observations in the interval

$$x_j^* \le x \le x_{j+1}^*.$$

An alternative method for graphical representation of a grouped frequency distribution is the *frequency polygon.* The class frequency values are joined by straight lines to form an open polygon which is referred to as the frequency polygon. If the class widths are not all equal, the construction is based on the points $(x_i, f_i/\Delta x_i)$, as in the case of the histogram.

A *bar chart* is used for the graphical representation of a sample distribution in which all the elements in a given class have the same value. Here the class values are plotted along the x-axis and a vertical line (or bar) of height equal to the class frequency is drawn at the class value.

The square-root re-expression of a histogram is the *rootogram.* The class widths have not changed; so we *keep* the same bar widths as in the histogram, but we now use $\sqrt{f_i}/\Delta x_i$ as the height of the bar for class i. A suitable re-expression can make data more regular and easier to look at.

Examples of these graph types are given in Fig. 2.16.

Problem 2.12 *Construction of a histogram*

Construct a histogram with a constant class width for the samples from the (A) rectangular, (B) normal, (C) exponential, and (D) Laplace distributions.

Data: as for Problem 2.2

Program: Chemstat: Basic Statistics: Exploratory continuous.

Solution: The histograms in Fig. 2.17 indicate quite obviously the type of distribution the sample was taken from.

G14: Quantile–quantile plot (Q–Q plot)

[*x-axis*: the quantile $Q_S(P_i)$; *y-axis*: the order statistic $x_{(i)}$]

Given a random sample, we often need to find whether the data can be regarded as a sample from a population with a given theoretical distribution. To look at the closeness of the sample distribution to a given theoretical one, the quantile – quantile plot (Q–Q plot) is used [6]. The Q–Q plot allows comparison of the sample distribution being described by the empirical $Q_E(P_i)$ quantile function with the given theoretical one, with the theoretical $Q_T(P_i)$ quantile function. The empirical Q_E function is approximated by the sample order statistic $x_{(i)}$. If there is close agreement between the sample and theoretical distributions, it must be true that

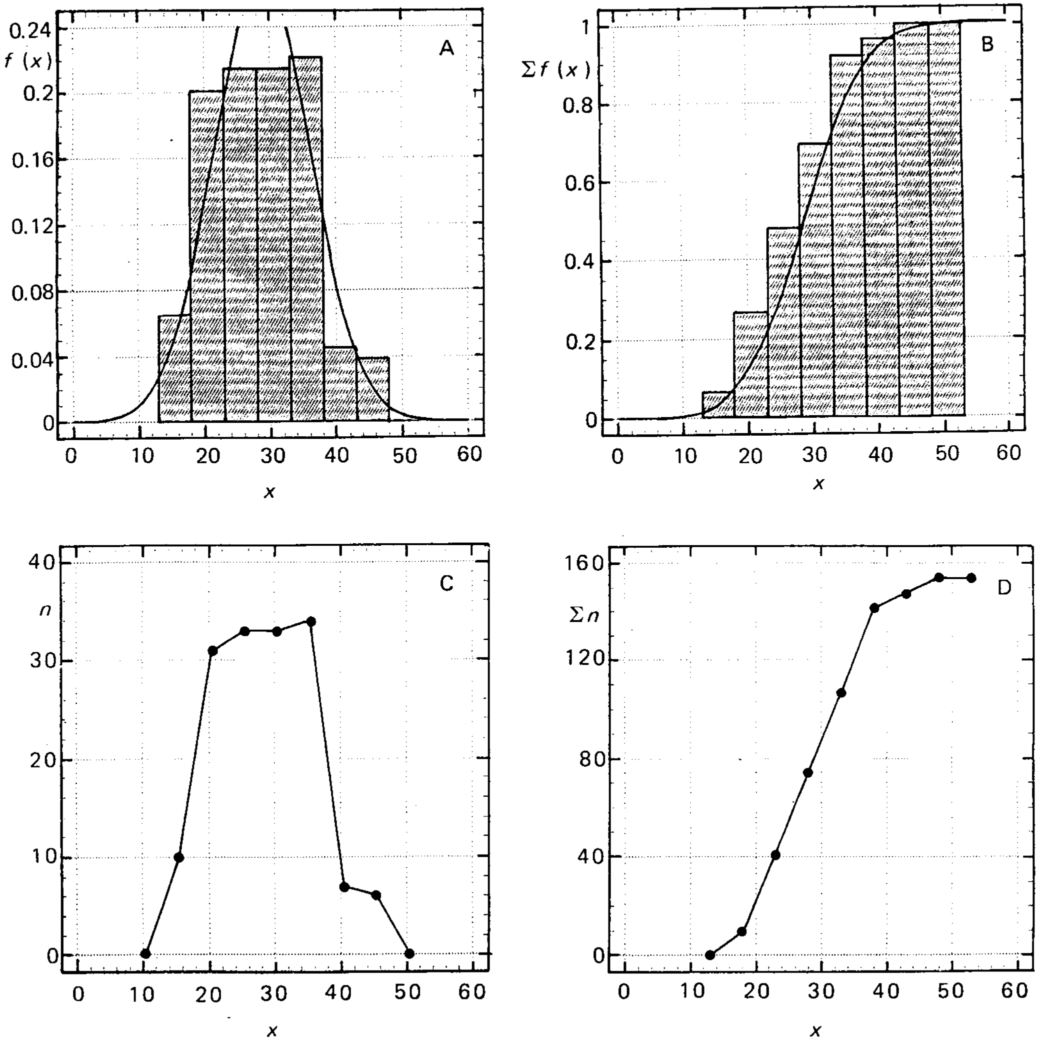


Fig. 2.16—(A) Histogram and probability density function, (B) cumulative histogram, (C) frequency polygon, and (D) cumulative frequency polygon.

$$x_{(i)} \simeq Q_{\mathrm{T}}(P_i) \tag{2.19}$$

where P_i is the cumulative probability chosen as

$$P_i = (i - 1/3)(n + 1/3).$$

When the empirical sample distribution is the same as the theoretical one, the resulting Q–Q plot is represented by a straight line [see Eq. (2.19)].

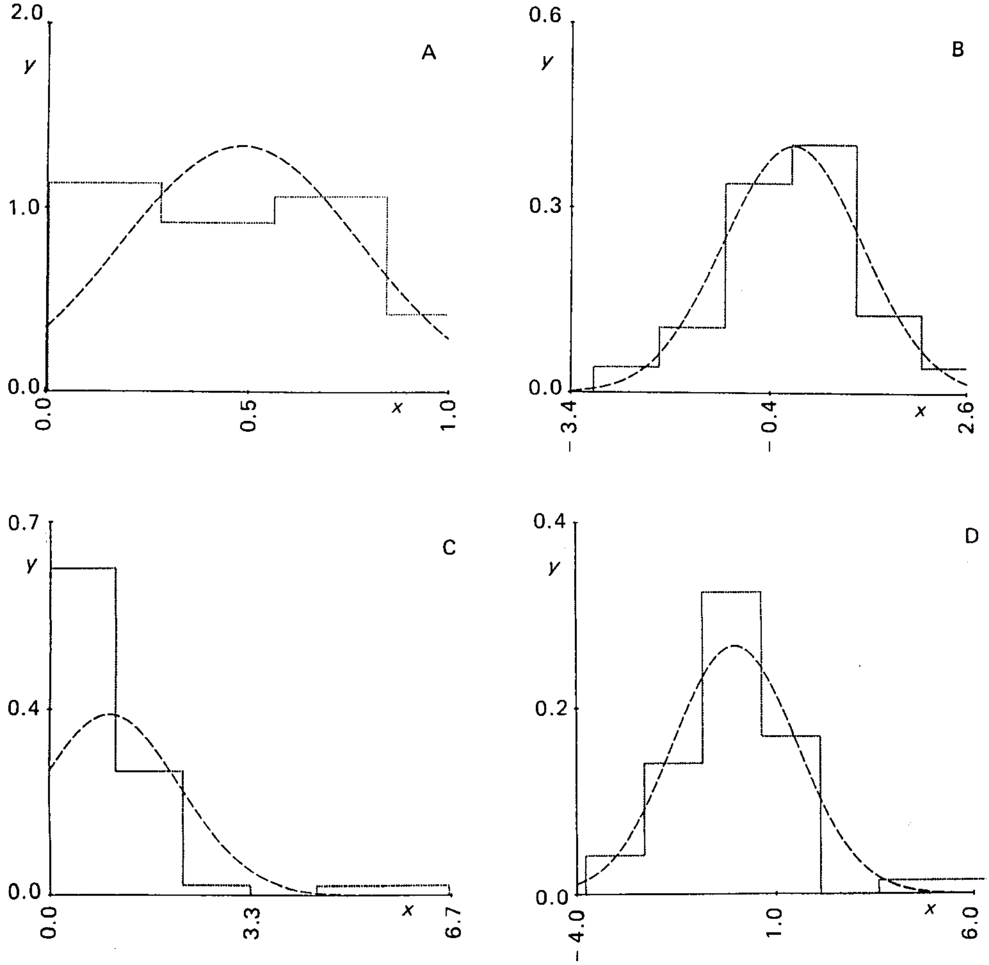


Fig. 2.17—The histograms (G13) for samples from the (A) rectangular, (B) normal, (C) exponential, and (D) Laplace distributions.

To construct this plot, the parameters of location and spread of the theoretical distribution (or their estimates) must be known. For many theoretical distributions, the standardized variable S may be used

$$S = (x - Q)/R \tag{2.20}$$

where Q stands for a parameter of location or threshold and R for a parameter of spread. The standardized (theoretical) quantile function $Q_S(P_i)$ then contains only shape parameters (their magnitude may be varied).

When there is agreement between the empirical sample and the theoretical distribution, the Q–Q plot is a straight line

$$x_{(i)} = Q + RQ_S(P_i) \tag{2.21}$$

For selected theoretical distributions the x and y co-ordinates of the Q–Q graph are given in Table 2.6.

Table 2.6—Standardized frequency $f_T(s)$ and distribution $F_T(s)$ functions, and corresponding co-ordinates (x, y) of the Q–Q plot

Distribution	$F_T(s)$	$f_T(s)$	y	x
Rectangular	s	1	$x_{(i)}$	P_i
Exponential	$1 - \exp(-s)$	$\exp(-s)$	$x_{(i)}$	$-\ln(1 - P_i)$
Normal	$\Phi(s)$	$(2\pi)^{-1/2} \exp(0.5s^2)$	$x_{(i)}$	$\phi^{-1}(P_i)$
Laplace				
$x \leq Q$	$0.5 \exp(s)$	$0.5 \exp(s)$	$x_{(i)}$	$\ln(2P_i)$ for $P_i \leq 0.5$
$x > Q$	$0.5[2 - \exp(-s)]$	$0.5 \exp(-s)$	$x_{(i)}$	$-\ln(2(1 - P_i))$ for $P_i > 0.5$
Log-normal	$\Phi[\ln(s)]$	$(2\pi)^{-1\cdot 2}\exp(-0.5 \ln s^2)$	$x_{(i)}$	$\exp[\Phi^{-1}(P_i)]$

In Table 2.6 the normal distribution function $\Phi(s)$ is defined as

$$\Phi(s) = \frac{1}{\sqrt{2\pi}} \int_{-x}^{s} \exp(-0.5u^2)\, \mathrm{d}u$$

To calculate the inverse function $\Phi^{-1}(P_i)$, the following simple approximate expression may be used

$$\Phi^{-1}(P_i) = -9.4 \ln[1/P_i - 1]/[\mathrm{abs}(\ln(1/P_i - 1))] + 14.$$

Problem 2.13 *Construction of the Q–Q plot*
Construct the quantile–quantile plot for investigation of agreement between the distribution of sample (B) in Problem 2.2 and the theoretical (A) rectangular, (B) normal, (C) exponential, and (D) Laplace distributions.
Data: as for Problem 2.2
Program: Chemstat: Basic Statistics: Distribution checking
Solution: Agreement between the empirical distribution and the theoretical distributions is assessed by the goodness of the fit to a straight line, which can also be estimated by the correlation coefficient r_{xy}. When the sample distribution is compared with the normal distribution (B), the correlation coefficient is 0.993, for the rectangular (A) distribution, 0.963, for the exponential distribution, 0.908, and for the Laplace one, 0.994.
Conclusion: From the Q–Q graphs shown in Fig. 2.18 is clear that the sample distribution has slightly longer tails than the theoretical normal one. It is then difficult

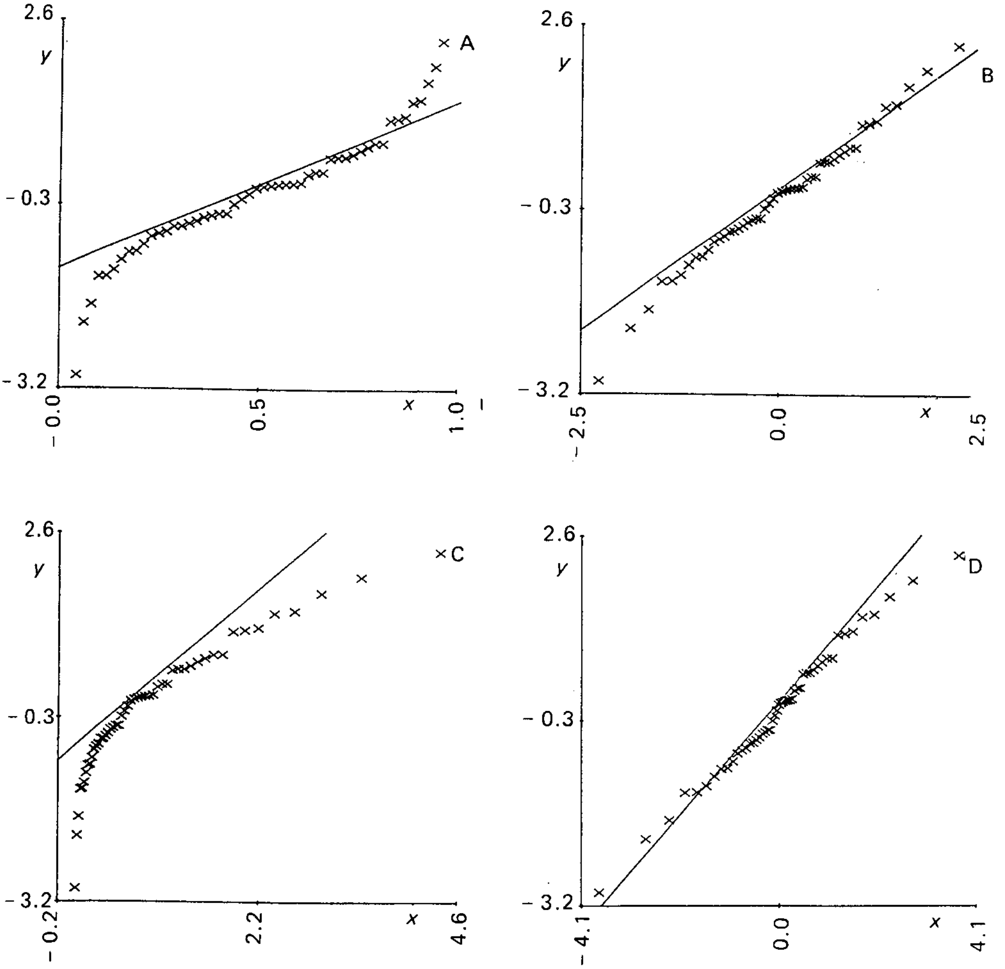


Fig. 2.18—The Q–Q plot (G14) for comparison of the empirical distribution of sample (B) with the theoretical (A) rectangular, (B) normal, (C) exponential, and (D) Laplace distributions.

to decide if the sample is taken from the normal (correct) or Laplace (incorrect) distribution.

G15: Rankit plot

(*x-axis*: the standardized normal quantile u_{P_i}; *y-axis*: the order statistic $x_{(i)}$)
When it is desired to test whether a given random sample can be regarded as a sample from a normal (Gaussian) distribution, the resulting Q–Q plot is called the rankit plot or the normal probability plot. This plot enables classification of a sample distribution according to its skewness, kurtosis and tail length. A convex or concave shape indicates a skewed sample distribution. A sigmoidal shape indicates that the tail lengths of the sample distribution differ from those of the normal one.

Problem 2.14 *Construction of a rankit plot*

Construct the rankit plot for samples from (A) rectangular, (B) normal, (C) exponential, and (D) Laplace distributions with the theoretical one.

Data: as for Problem 2.2

Program: Chemstat: Basic Statistics: Exploratory continuous: Q–Q plot.

Solution: The graphs in Fig. 2.19 indicate short tails (A), or long tails (D), and skewing of the distribution to higher values in case (C). Also, sample (B) has rather longer tails than an ideal normal distribution.

G16: Conditioned rankit plot

(*x-axis*: the function $\phi^{-1}[(u_{(i-1)} + U_{(i+1)})/2]$; *y-axis*: the order statistic $x_{(i)}$)

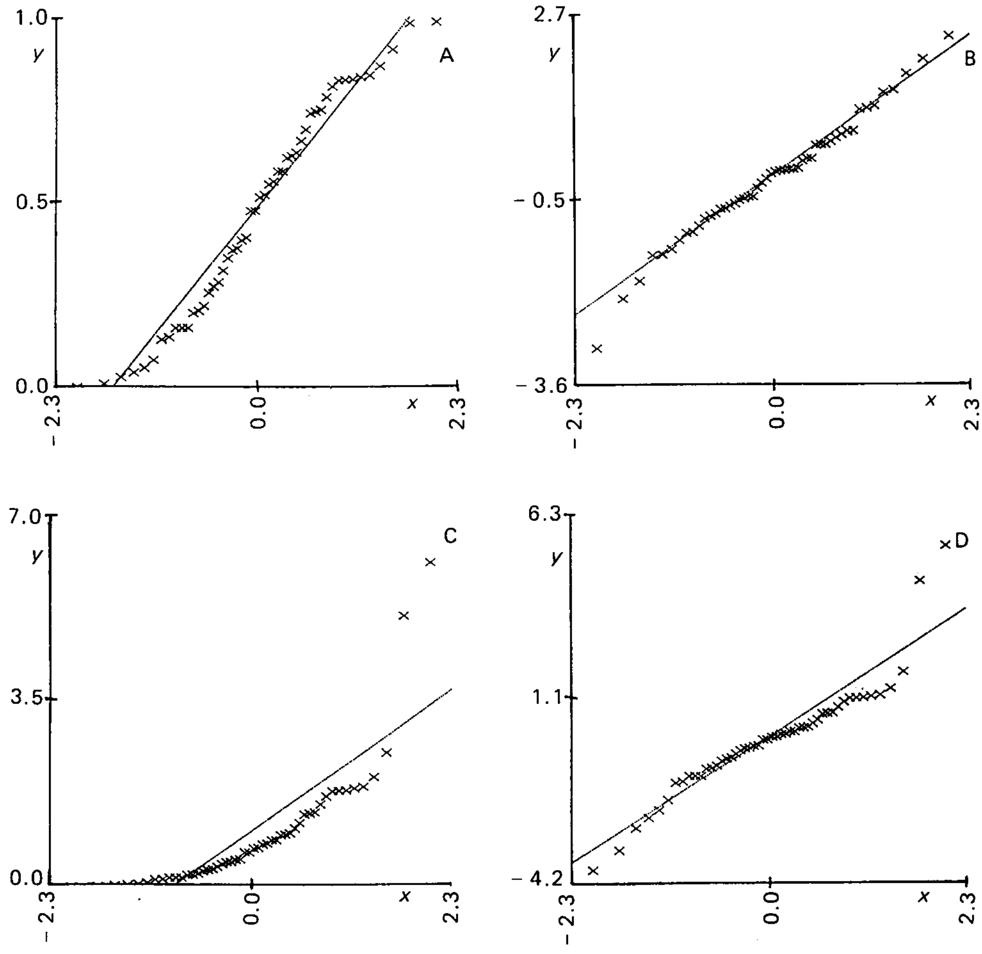


Fig. 2.19—The rankit plot (G15) for samples from (A) rectangular, (B) normal, (C) exponential and (D) Laplace distributions.

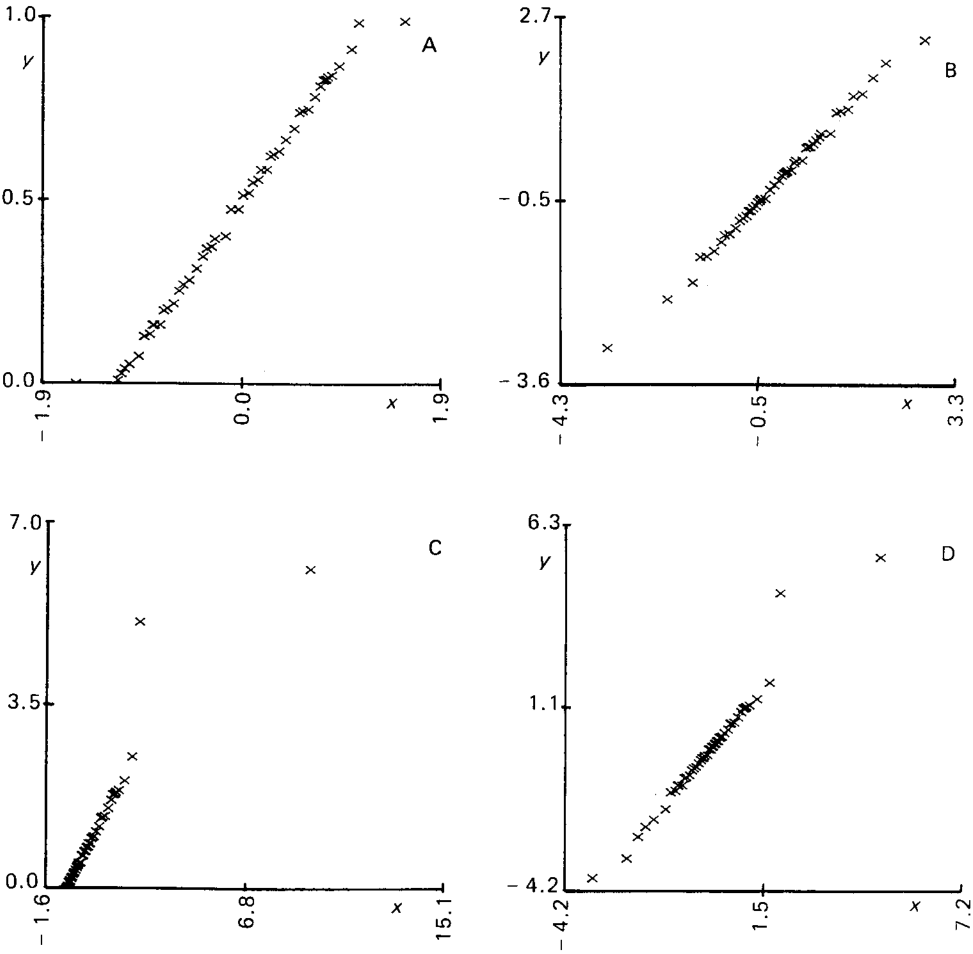


Fig. 2.20—The conditioned rankit plot (G16) for samples from the (A) rectangular, (B) normal, (C) exponential, and (D) Laplace distributions.

Kafander and Spiegelman [7] have recommended the conditioned rankit plot for the examination of the normality of a sample distribution. The symbol $\Phi^{-1}(U)$ denotes the standardized quantile function of the standardized normal distribution where, for $U = P_i$, it corresponds to the normal quantile u_{P_i}. The order statistic $U_{(i)}$ corresponds to the random variable U_i defined by

$$U_i = \Phi[(x_i - \hat{\mu}_R)/\hat{\sigma}_R^2] \qquad (2.22)$$

where the symbol $\Phi(x)$ stands for the distribution function of the standardized normal distribution. The robust estimate of location $\hat{\mu}_R = M$ is equal to the median and the robust estimate of the standard deviation is

$$\hat{\sigma}_R = 0.75(\tilde{x}_{0.75} - \tilde{x}_{0.25}).$$

For complete definition, $U_{(0)} = 0$ and $U_{(n+1)} = 1$ are also required. Approximate linearity of the conditioned rankit plot indicates normality of the sample distribution.

Problem 2.15 *Construction of a conditioned rankit plot*
Construct the conditioned rankit plot for the samples from (A) rectangular, (B) normal, (C) exponential, and (D) Laplace distributions.
Data: as for Problem 2.2
Program: Chemstat: Basic Statistics: Exploratory continuous: Q–Q conditioned
Solution: The conditioned rankit plot shown in Fig. 2.20 shows smaller local variability than the rankit plot (Fig. 2.19). Sample (B) is proved here to be from the normal distribution.

G17: Frequency-ratio plot

(*x-axis*: the variable x; *y-axis*: the function $xp(x)/[p(x-1)]$)
To distinguish between various types of discrete distributions the frequency ratio plot is used. This plot is based on the expression

$$\frac{x\ p(x)}{p(x)-1} = C_0 + C_1 x \tag{2.23}$$

where the discrete variable $x = 1, 2, \ldots, k$, and the symbol $p(x)$ stands for the probability frequency function. Equation (2.23) is valid for many discrete distributions. By comparing estimated values of the slope C_1 and intercept C_0 of a straight line on the frequency ratio plot with theoretical values from Table 2.7, the actual type of discrete distribution may be identified.

Table 2.7—The slope C_1 and intercept C_0 of the straight line in the frequency ratio plot

Distribution	Probability function $p(x)$	Slope C_1	Intercept C_0
Poisson	$\exp(-\lambda)\lambda^x/x!$	0	λ
Binomial†	$\binom{n}{x} p^x(1-p)^{n-x}$	$-p/(1-p)$	$\dfrac{p(n+1)}{(1-p)}$
Negative binomial†	$\binom{n+x-1}{x} p^n(1-p)^x$	$1-p$	$(n-1)(1-p)$
Geometric†	$p(1-p)^{x-1}$	$1-p$	0

†where n, p are parameters of the distribution

Problem 2.16 *Construction of a frequency ratio plot*
Four hundred values from the Poisson distribution with $\lambda = 2$ were generated. The sample estimate $\hat{\lambda}$ is 2.053. Construct a frequency ratio plot and indicate the actual distribution.
Data: A random number generator was used to generate 400 numbers for the Poisson distribution ($\lambda = 2$).

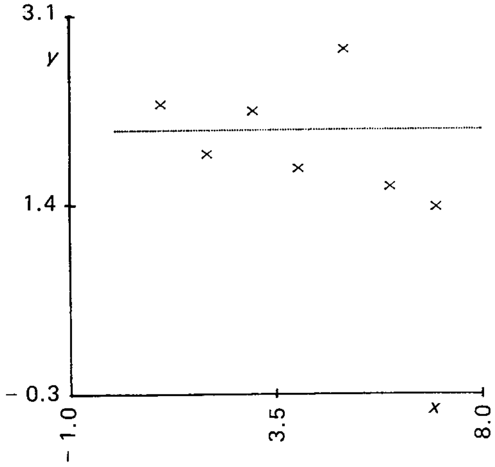


Fig. 2.21—The frequency ratio plot (G17) for a sample from the Poisson distribution.

Program: Chemstat: Basic Statistics: Exploratory discrete.
Solution: The frequency ratio plot is shown in Fig. 2.21.
Conclusion: This plot (Fig. 2.21) shows great scatter about the line $y = \hat{\lambda}$ but the sample distribution can be approximated by the Poisson distribution.

G18: Poisson plot

(x-*axis*: the variable x; y-*axis*: the function $\ln(x!\, n_x/n)$)
The Poisson plot is based on the validity of the equation

$$\ln(x!\, n_x/n) = -\lambda + x \ln \lambda \tag{2.24}$$

where the absolute frequency n_x represents the number of sample values reaching the magnitude x, and n is the sample size. If the actual distribution is of Poisson nature, the Poisson plot is a straight line with slope ln λ and intercept λ. When an estimate of $\hat{\lambda}$ is known, the "theoretical" straight line $y = -\hat{\lambda} + x \ln \hat{\lambda}$ may be drawn.

Problem 2.17 *Construction of a Poisson plot*
Construct the Poisson plot for the data used in Problem 2.16.
Program: Chemstat: Basic Statistics: Exploratory discrete.
Solution: The Poisson plot shows that the data points are in good agreement with the "theoretical" straight line

$$y = -2.053 + x(\ln 2.053)$$

and therefore the sample comes from the Poisson distribution.
Conclusion: the Poisson plot is shown in Fig. 2.22.

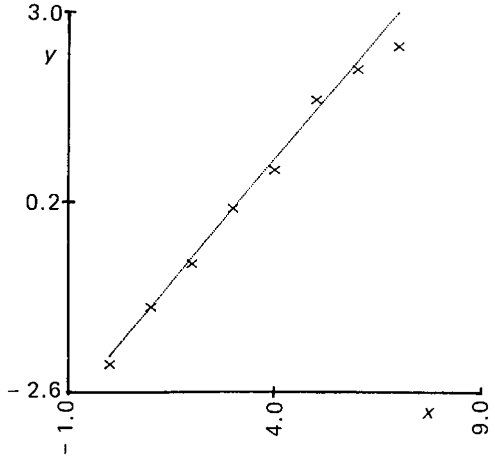


Fig. 2.22—The Poisson plot (G18).

G19: Modified Poisson plot

[*x-axis*: the variable x; *y-axis*: the function $\ln(x!\,n_x/n) + (\lambda_0 - x \ln \lambda_0)$]

To examine the suitability of the value selected for parameter λ_0 in the Poisson distribution, the modified Poisson plot can be used. When the estimate λ_0 is reasonably suitable, the sample points lie on the horizontal line $y = 0$.

Problem 2.18 *Construction of a modified Poisson plot*

Construct the modified Poisson plot for the data sample used in Problem 2.15.

Solution: The estimate $\lambda_0 = 2$ is used to construct the modified Poisson plot. Since all the points are randomly spread around the horizontal straight line $y = 0$, the sample comes from a Poisson distribution with the mean equal to 2.

Conclusion: The plot is shown in Fig. 2.23.

2.5 DATA TRANSFORMATION

When exploratory data analysis proves that the sample distribution strongly differs from the normal one, we are faced with the problem of how to analyse the data. Raw data may require re-expression to produce an informative display, effective summary, or a straightforward analysis. We may need to change not only the units in which the data are stated, but also the basic scale of the measurement. To change the shape of a data distribution, we must do more than change the origin and/or unit of measurement. Changes of origin and scale mean linear transformations, and they leave shape alone. Nonlinear transformations such as the logarithm and square root are necessary to change shapes.

Data must be examined so as to find the *proper transformation* which leads to symmetric distribution of data, stabilizes the variance, or makes the distribution closer

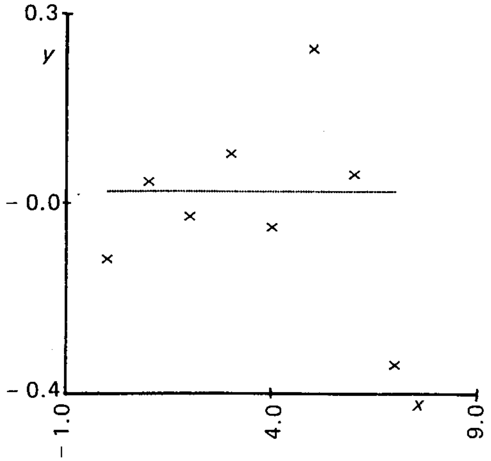


Fig. 2.23—The modified Poisson plot (G19).

to normal. Such transformation of original data x to a new variable $y = g(x)$ is based on an assumption that the data represent a nonlinear transformation of the normally distributed variable y, according to $x = g^{-1}(y)$.

Transformation for variance stabilization involves finding a transformation $y = g(x)$ in which the variance $\sigma^2(y)$ is constant. If the variance of the original variable x is a function of type $\sigma^2(x) = f_1(x)$, the variance $\sigma^2(y)$ may be expressed by

$$\sigma^2(y) \simeq \left(\frac{\mathrm{d}g(x)}{\mathrm{d}x}\right)^2 f_1(x) = C \tag{2.25}$$

where C is a constant. The chosen transformation $g(x)$ is the solution of the differential equation

$$g(x) \simeq C \int \frac{\mathrm{d}x}{\sqrt{f_1(x)}} \tag{2.26}$$

In some instrumental methods of analytical and physical chemistry, the relative standard deviation $\delta(x)$ of the measured variable is constant. This means that the variance $\sigma^2(x)$ is described by a function $\sigma^2(x) = f_1(x) = \delta^2(x)x^2 = \text{const} \times x^2$. The substitution into Eq. (2.26) will be $g(x) = \ln x$, so that one form of transformation of original data is the logarithmic transformation. This transformation leads to the use of a geometric mean.

When the dependence $\sigma^2(x) = f_1(x)$ is of power nature, the optimal transformation will also be a power transformation. Since for a normal distribution the mean is not dependent on the variance, a transformation that stabilizes the variance makes the distribution closer to normal.

Transformation for symmetry is carried out by a simple power transformation

$$y = g(x) = \begin{array}{ll} x^{\lambda} & \text{for parameter } \lambda > 0 \\ \ln\, x & \text{for parameter } \lambda = 0 \\ -x^{-\lambda} & \text{for parameter } \lambda < 0 \end{array} \qquad (2.27)$$

which does not retain the scale, is not always continuous, and is suitable only for positive data x. Optimal estimates of parameter $\hat{\lambda}$ are sought by minimizing the absolute values of particular characteristics of asymmetry. In addition to the classical estimate of skewness $\tilde{g}_1(y)$, [Eq. (3.29)] the robust estimate $\tilde{g}_{1,R}(y)$ is used:

$$\tilde{g}_{1,R}(y) = \frac{(\tilde{y}_{0.75} - \tilde{y}_{0.50}) - (\tilde{y}_{0.50} - \tilde{y}_{0.25})}{\tilde{y}_{0.75} - \tilde{y}_{0.25}} \qquad (2.28)$$

The relative distance between the arithmetic mean and the median may also be utilized:

$$\tilde{g}_P(y) = \frac{\bar{y} - \tilde{y}_{0.50}}{\left[\sum_{i=1}^{n} (y_i - \bar{y})^2/(n-1)\right]^{1/2}} \qquad (2.28)$$

because for symmetrical distributions this is equal to zero.

The estimate of parameter $\hat{\lambda}$ may be found also from a rankit plot, because for an optimal value of $\hat{\lambda}$ the transformed quantiles $y_{(i)}$ will lie on the straight line.

G20: Hines–Hines selection graph

(x-*axis*: the ratio $\tilde{x}_{0.5}/\tilde{x}_{1-P_i}$, y-*axis*: the ratio $\tilde{x}_{P_i}/\tilde{x}_{0.5}$)

An excellent diagnostic tool enabling estimation of parameter λ is represented by the Hines–Hines selection graph [8]. This is based on an assumption of symmetry of individual quantiles around a median

$$(\tilde{x}_{P_i}/\tilde{x}_{0.5})^{\lambda} + (\tilde{x}_{0.5}/\tilde{x}_{1-P_i})^{-\lambda} = 2 \qquad (2.30)$$

where, for the cumulative probability $P_i = 2^{-i}$, the letter values F, E ($i = 2,3$) are usually chosen.

To compare the empirical dependence of the experimental points with the ideal one, patterns for various values of parameter λ are drawn in a selection graph. These patterns λ represent a solution of the equation $y^{\lambda} + x^{-\lambda} = 2$ in the range $0 \leq x \leq 1$ and $0 \leq y \leqslant 1$:

(1) for $\lambda = 0$ the solution is a straight line $y = x$;
(2) for $\lambda \leq 0$ the solution takes the form $y = (2 - x^{-\lambda})^{1/\lambda}$;
(3) for $\lambda > 0$ the solution takes the form $x = (2 - y^{\lambda})^{-1/\lambda}$.

The estimate $\hat{\lambda}$ is guessed from a selection graph, according to the location of experimental points near to the various theoretical patterns.

Problem 2.19 *Estimation of λ from a Hines–Hines selection graph*

Construct the Hines–Hines selection graph for the sample (E) taken from the log normal distribution (Problem 2.2) and find the optimal power transformation with the use of the estimate of parameter λ.

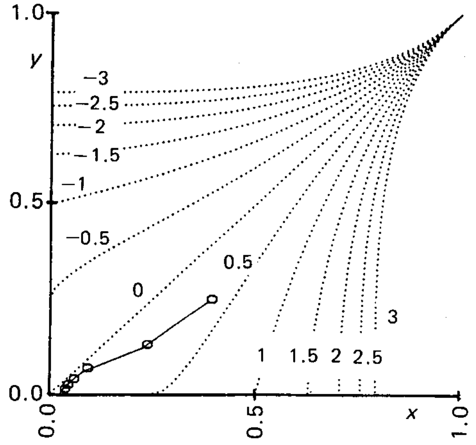


Fig. 2.24—Determination of λ from a Hines–Hines selection graph (G20).

Data: sample (E) from Problem 2.2
Program: Chemstat: Basic Statistics: Power transformation: Hines–Hines plot
Solution: The selection graph shown in Fig. 2.24 suggests that the best estimate for parameter λ is $\lambda = 0$, because the experimental points oscillate around the curve $\lambda = 0$. The value $\lambda = 0$ corresponds to a logarithmic transformation.
Conclusion: The simulated data from a log–normal distribution have shown that a selection graph was able to find a suitable value of the transformation parameter $\lambda = 0$ that would lead to normality.

In many cases sample distributions can be transformed to approximate normality by use of the family of *Box–Cox transformations* defined as

$$y = g(x) = \begin{matrix} (x^{\lambda} - 1)/\lambda & \text{for } \lambda \neq 0 \\ \ln x & \text{for } \lambda = 0 \end{matrix} \tag{2.31}$$

where x is a positive variable and λ is real number. Box–Cox transformation has following properties:

(1) The curves of transformation $g(x)$ are monotonic and continuous with respect to parameter λ, because

$$\lim_{\lambda \to 0} (x^{\lambda} - 1)/\lambda = \ln x \tag{2.32}$$

(2) All transformation curves share one point $[y = 0, x = 1]$ for all values of λ. The curves nearly coincide at points close to [0,1); that is, they share a common tangent line at that point.
(3) The power transformations with exponent -2, $-3/2$, -1, $-1/2$, 0, 1/2, 1, 3/2, 2 have equal spacing between curves in the family of Box–Cox transformation graphs.

The Box–Cox transformation defined by Eq. (2.32) can be applied only for positive data. To extend this transformation, the x values are replaced by $(x - x_0)$ values, which are always positive. Here x_0 is the threshold value $x_0 < x_{(1)}$.

G21: Plot of logarithm of likelihood function
(x-*axis*: the parameter λ; y-*axis*: the logarithm of the likelihood function ln L)
To estimate parameter λ in the Box–Cox transformation, Eq. (2.31), the method of maximum likelihood may be used, because for $\lambda = \hat{\lambda}$ a distribution of the transformed variable y is considered to be normal, $N[\mu_y, \sigma^2(y)]$. The logarithm of the maximum likelihood function may be written as

$$\ln L(\lambda) = -\frac{n}{2} \ln s^2(y) + (\lambda - 1) \sum_{i=1}^{n} \ln x_i \tag{2.33}$$

where $s^2(y)$ is the sample variance of the transformed data. y. The function ln $L = f(\lambda)$ is expressed graphically for a suitable interval, for example, $-3 \leq \lambda \leq 3$. The maximum on this curve represents the maximum likelihood estimate $\hat{\lambda}$.

The asymptotic 100(1 − α)% confidence interval of parameter λ is expressed by

$$2 \ln L(\hat{\lambda}) - \ln L(\lambda)] \leq \chi^2_{1-\alpha}(1) \tag{2.34}$$

where $\chi^2_{1-\alpha}(1)$ is the quantile of the χ^2 distribution with 1 degree of freedom. This interval contains all values λ for which it is true that:

$$\ln L(\lambda) \geq \ln L(\hat{\lambda}) - 0.5\, \chi^2_{1-\alpha}(1) \tag{2.35}$$

This Box–Cox transformation is less suitable for wide confidence intervals. When the value $\lambda = 1$ is also covered by this confidence interval, the transformation is not efficient.

Problem 2.20 *Construction of the plot of the logarithm of maximum likelihood*
Construct the plot of the logarithm of maximum likelihood for sample (E) from the log normal distribution.
Data: sample (E) from Problem 2.2
Program: Chemstat: Basic Statistics: Power transformations.
Solution: The plot of logarithm of maximum likelihood for sample (E) is shown in Fig. 2.25. The optimal estimate is $\hat{\lambda} = 0$, for which the logarithm of maximum likelihood reaches the value ln $L = -31.4$. The 95% confidence level is graphed too. Figure 2.26 shows the quantile box plot and rankit plot for both the original and transformed data. With respect to skewness, the optimal estimate of μ is $\hat{\lambda} = 0$ but with respect to the robust skewness $\hat{g}_R(y)$, it is $\hat{\lambda} = 0.133$. The optimal value of x is that for which $\hat{g}^1$ (or $\hat{g}_R$) reaches a value near zero.

2.6 RE-EXPRESSION OF STATISTICS FOR TRANSFORMED DATA

After an appropriate transformation of the original data $\{x\}$ has been found, so that the transformed data gives an approximately normal symmetrical distribution with constant variance, the statistical measures of location and spread for the transformed data $\{y\}$ are calculated. These include the sample arithmetic mean $\bar{y}$, the sample

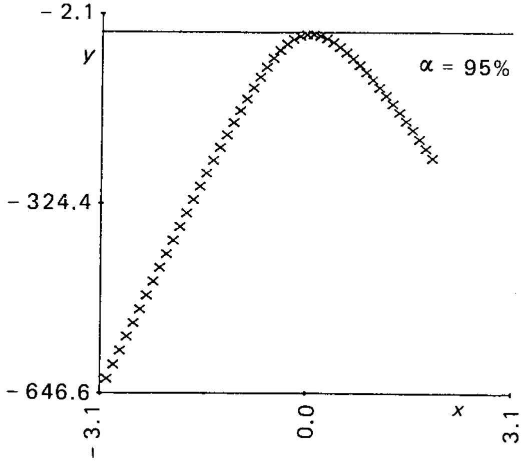


Fig. 2.25—The plot of the logarithm of maximum likelihood (G21).

variance $s^2(y)$, and the confidence interval of the mean $\bar{y} \pm t_{1-\alpha/2}(n-1)s(y)/\sqrt{n}$. These estimates must then be recalculated for the original data $\{x\}$. Two different approaches to re-expression of the statistics for transformed data exist.

(1) *Rough re-expressions* represent a single reverse transformation $\bar{x}_R = g^{-1}(y)$. This re-expression for a simple power transformation leads to the general mean

$$\bar{x}_R = \bar{x}_\lambda = \left[\frac{\sum_{i=1}^{n} x_i^\lambda}{n}\right]^{1/\lambda} \tag{2.36}$$

where for $\lambda = 0$, ln x is used instead of x^λ and e^x instead of $x^{1/\lambda}$. The re-expressed mean $\bar{x}_R = \bar{x}_{-1}$ stands for the *harmonic mean*, $\bar{x}_R = \bar{x}_0$ for the *geometric mean*, $\bar{x}_R = \bar{x}_1$ for the *arithmetic mean* and $\bar{x}_R = \bar{x}_2$ for the *quadratic mean*.

(2) *More correct re-expressions* are based on the Taylor series expansion of the function $y = g(x)$ in the neighbourhood of the value $\bar{y}$. The re-expressed mean $\bar{x}_R$ is then given by

$$\bar{x}_R \approx g^{-1}\left(\bar{y} - \frac{1}{2}\frac{\mathrm{d}^2 g(x)}{\mathrm{d}\ x^2}\left(\frac{\mathrm{d}\ g(x)}{\mathrm{d}\ x}\right)^{-2} s^2(y)\right) \tag{2.37}$$

For the variance

$$s^2(x_R) \approx \left(\frac{\mathrm{d}\ g(x)}{\mathrm{d}\ x}\right)^{-2} s^2(y) \tag{2.38}$$

where individual derivatives are calculated at the point $x = \bar{x}_R$. The $100(1-\alpha)\%$ confidence interval of the re-expressed mean for the original data may be defined as

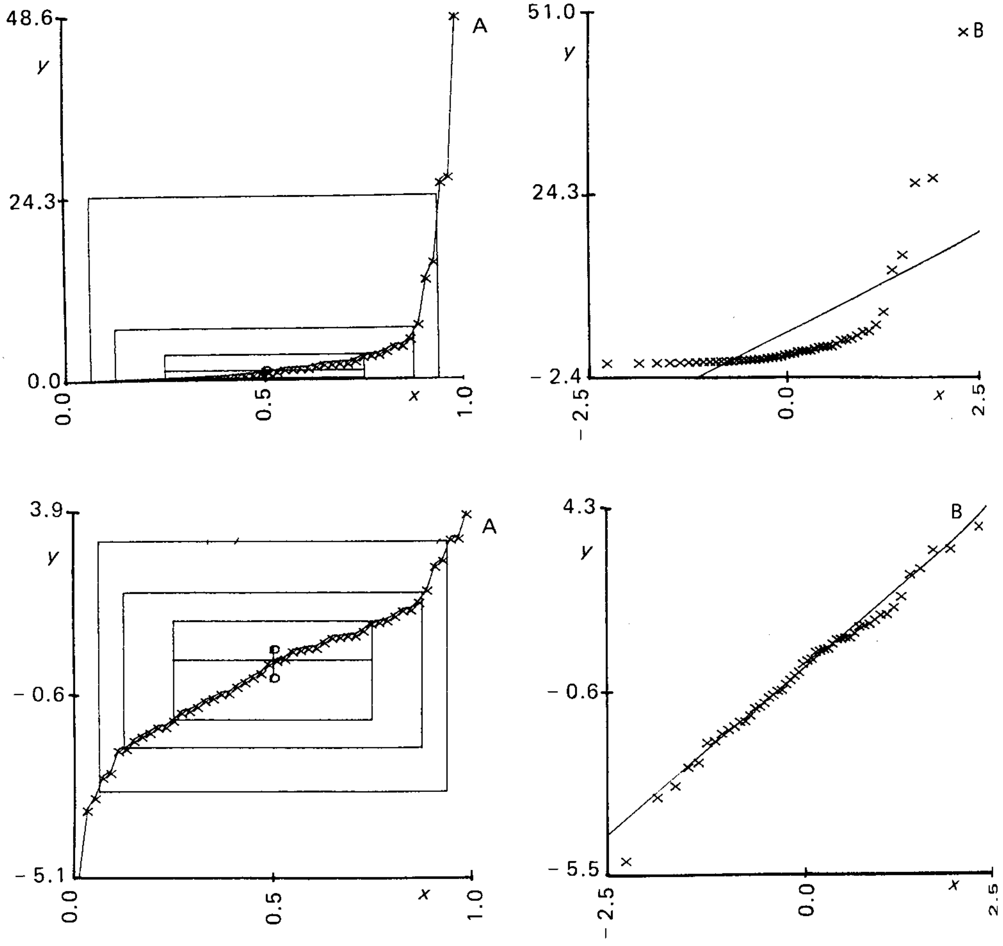


Fig. 2.26—The EDA graphical examination of a sample from the log-normal distribution: (A) the quantile–box plot (G10), (B) the rankit plot (G15). Upper graphs are for the original data, and lower for logarithmically transformed data.

$$\bar{x}_R - I_L \leq \mu \leq \bar{x}_R + I_U \tag{2.39}$$

where

$$I_L = g^{-1}[\bar{y} + G - t_{1-\alpha/2}(n-1)s(y)/\sqrt{n}] \tag{2.39a}$$

$$I_U = g^{-1}[\bar{y} + G + t_{1-\alpha/2}(n-1)s(y)/\sqrt{n}] \tag{2.39b}$$

$$G = -\frac{1}{2}\frac{\mathrm{d}^2 g(x)}{\mathrm{d}\ x^2}\left(\frac{\mathrm{d}\ g(x)}{\mathrm{d}\ x}\right)^{-2} s^2(y) \tag{2.39c}$$

On the basis of the (known) actual transformation function $y = g(x)$ and the estimates $\bar{y}$, $s^2(y)$, it is easy to calculate re-expressed estimates $\bar{x}_R$ and $s^2(x_R)$:

(1) For a logarithmic transformation ($\lambda = 0$) and $g(x) = \ln x$, the re-expressed mean and variance will be given by Eq. (2.37), i.e.

$$\bar{x}_R \approx \exp[\bar{y} + 0.5\, s^2(y)] \tag{2.40}$$

and

$$s^2(x_R) \approx \bar{x}_R^2 s^2(y) \tag{2.41}$$

(2) For $\lambda \neq 0$ and the Box–Cox transformation, Eq. (2.31), the re-expressed mean $\bar{x}_R$ will be represented by one of the two roots of the quadratic equation

$$\bar{x}_{R,1,2} = [0.5(1 + \lambda\bar{y}) \pm 0.5\{1 + 2\lambda(\bar{y} + s^2(y)) + \lambda^2(\bar{y}^2 - 2s^2(y))\}^{1/2}]^{1/\lambda} \tag{2.42}$$

which is close to the median $\tilde{x}_{0.5} = g^{-1}(\tilde{y}_{0.5})$. If $\bar{x}_R$ is known, the corresponding variance may be calculated from

$$s^2(x) = \bar{x}_R^{(-2\lambda+2)} s^2(y) \tag{2.43}$$

Problem 2.21 *Re-expressed statistics for logarithmic data*
Make a reverse transformation of the statistics estimated for the sample (E) from the log–normal distribution, and compare the rough and correct approaches to re-expressed estimates.
Data: sample (B) from Problem 2.2
Program: Chemstat: Basic Statistics: Power transformations.
Solution: Logarithmic transformation of the sample (E) leads to estimates $\bar{y} = 0.41$ and $s^2(y) = 3.226$. The correct approach of Eq. (2.40) gives for the re-expressed mean

$$\bar{x}_R = \exp(0.041 + 0.5 \times 3.226) = 5.23.$$

The rough approach gives

$$\bar{x}_R = \exp(0.041) = 1.05$$

which is significantly smaller. From Eq. (2.41) the re-expressed variance is

$$s^2(\bar{x}_R) = 5.23^2 \times 3.226 = 88.24.$$

2.7 CONFIRMATORY ANALYSIS OF ASSUMPTIONS ABOUT DATA

Statistical treatment of experimental data supposes that the data are independent random variables from the same distribution, which may be normal in nature, and that the sample size is sufficient for precise estimates of location and spread to be obtained.

When some of these assumptions about data are not fulfilled, the data analysis is rather complicated. These assumptions are examined in *confirmatory data analysis (CDA)*.

2.7.1 Examination for minimum sample size

The sample size has an influence on the precision of estimates; e.g., the variance of the parameter estimate is a function of $1/n$. The sample size n controls the size of confidence intervals; i.e., for larger values of n, the confidence interval is smaller. For very small sample sizes it may happen that the class width and hypothesis tests are affected more by the sample size n than by the variability of data. The procedure for finding the sample size that is sufficient is as follows:

(1) From n_1 starting values, the sample variance $s_0^2(x)$ is calculated. The minimum size $n_{\min}$ of a sample taken from a normal distribution is calculated in such a way that for a given probability $(1-\alpha)$ and value of d, the confidence interval will be

$$\mu - d \le \bar{x} \le \mu + d,$$

and $n_{\min}$ is then given by

$$n_{\min} = s_0^2(x)[t_{1-\alpha/2}(n_1 - 1)/d]^2 \tag{2.44}$$

where $t_{1-\alpha/2}(n_1 - 1)$ is the quantile of the Student distribution with $(n_1 - 1)$ degrees of freedom.

(2) The minimum size $n_{\min}$ of a sample from the normal distribution may be chosen such that the relative error of the standard deviation $\delta(s)$ has a particular value

$$n_{\min} = 1 + [\hat{g}_2(x) - 1]/[4\delta^2(s)] \tag{2.45}$$

where $\hat{g}_2(x)$ is the estimate of the kurtosis of the sample distribution given by Eq. (3.30). The value of $\delta(s)$ usually chosen is 10%, i.e. 0.1. The minimum size $n_{\min}$ is several tens, so typical sample sizes used in chemical laboratories $n = 5, 10, \ldots$ are too small from the statistical point of view.

Problem 2.22 *Minimum sample size for samples from various distributions*
Determine the minimum sample size $n_{\min}$ for samples from the (A) rectangular, (B) normal, (C) exponential, (D) Laplace, and (E) log-normal distribution.
Data: as for Problem 2.2
Program: Chemstat: Basic Statistics: Assumptions testing.
Solution: For a relative error of the standard deviation of $\delta(s) = 10\%$, $n_{\min}$ can be calculated from Eq. (2.45). The results are given in Table 2.8.

Table 2.8—Minimal sample size $n_{\min}$ for five distributions, calculated for $\delta(s) = 10\%$

Distribution	Skewness	Minimal size
Rectangular	1.8	21
Normal	3.0	51
Exponential	6.0	126
Laplace	9.0	176
Log-normal	15.0	351

Conclusion: The sample size ($n = 50$) chosen for Problem 2.2 is suitable for the rectangular and normal distributions only. For the other three distributions it is too small, and does not ensure the requested value of $\delta(s) = 10\%$.

2.7.2 Examination for independence of sample elements

The basic assumption of good measurement is that the individual measurements (observations) in the sample set are independent. Interdependence of measurements may be caused by

(1) instability of the measurement equipment, for example, a shift in readings with time;
(2) variable conditions of measurements, which could suddenly change;
(3) neglect of factor(s) which have a great influence on measurement, for example, the sample volume, temperature, purity of chemicals, etc.
(4) false and non-random (stratified) choice of values in a sample.

When all the experimental factors change over time, a time dependence in the observations may be indicated if the observations are arranged in order of time. When there is a sudden change in observations, a heterogeneous sample is formed. In both the above cases, a higher value for the variance is found than for a homogeneous sample.

Any time dependence or dependence on the order of observations is tested for by examining the significance of the autocorrelation coefficient ρ_α according to

$$t_{\mathrm{n}} = T_1\sqrt{(n+1)}/\sqrt{(1-T_1)} \tag{2.46}$$

where

$$T_1 = (1 - T/2)\sqrt{[(n^2-1)/(n^2-4)]} \tag{2.46a}$$

and T is the von Neumann ratio defined by

$$T = \frac{\sum_{i=1}^{n-1} (x_{(i+1)} - x_{(i)})^2}{\sum_{i=1}^{n} (x_{(i)} - \bar{x})^2} \tag{2.46b}$$

When the null hypothesis H_0: $\rho_{\mathrm{a}} = 0$ is valid, the test criterion t_n has the Student distribution with $(n + 1)$ degrees of freedom. The alternative hypothesis H_{A} is $\rho_\alpha \neq 0$. When $|t_n| > t_{1-\alpha/2}(n_1 + 1)$, the null hypothesis about the independence of sample observations is rejected at the significance level α.

There are other nonparametric tests and tests for autocorrelation of higher order which are applicated individually or simultaneously. To find any interdependence of data, the whole measurement process and data collection should be examined.

Problem 2.23 *Test for independence of sample elements*
The sample (A) used in Problem 2.2 was generated from $R_{i+1} = (\pi + R_i)^8 - \mathrm{int}\,(\pi + R_i)^8$, $i = 1, \ldots, 50$, with $R_0 = 0$. Test the data for independence.
Program: Chemstat: Basic Statistics: Assumptions testing.
Solution: From Eq. (2.46b) the von Neumann ratio $T = 2.149$ and the test criterion $|t_n| = 0.534$. Since the quantile of the Student t-test $t_{0.975}(51) = 1.96$ is significantly larger, there is no evidence for autocorrelation of the elements of the data.

2.7.3 Testing for normality of sample distribution

Normality of a sample distribution is the basic assumption of most statistical data treatment, because many statistical tests require normality. When the type of deviation from normality of the sample is known before statistical inference, the *directional* tests are used; when the type of deviation from normality is unknown, the *omnibus tests* are used.

Generally, statistical tests are less sensitive to deviations from normality than diagnostic graphs. Moreover, deviation from normality can be caused by the presence of outliers. When the normality of a sample distribution is not proved, the data should be analysed with great care. To test normality of a sample distribution, the rankit plot is one of the most useful tools, but other useful tests are available.

(1) Test for combined sample skewness and kurtosis

The test criterion is defined as

$$C_1 = \frac{\hat{g}_1^2(x)}{D(\hat{g}_1(x))} + \frac{[\hat{g}_2(x) - 3]^2}{D(\hat{g}_2(x))} \tag{2.47}$$

where $\hat{g}_1(x)$ is the sample skewness and $D(\hat{g}_1(x))$ is its variance, $\hat{g}_2(x)$ is the sample kurtosis and $D(\hat{g}_2(x))$ is its variance (calculated from Eqs. (3.19a)–(3.20a)). For a normal distribution, the test criterion C_1 has approximately the χ^2 distribution, so that when $C_1 > \chi^2_{1-\alpha}(2)$, the null hypothesis about normality of sample distribution is rejected.

(2) Anderson – Darling test

This test is based on the empirical distribution function $F_E(x)$. The null hypothesis, H_0: $F_E(x) = F_T(x)$ is tested *vs.* H_A: $F_E(x) \neq F_T(x)$ where $F_T(x)$ is the distribution function of the fully specified distribution. The test criterion is defined as

$$AD = n - \left[\sum_{i=1}^{n} (2i - 1)[\ln Z_i + \ln (1 - Z_{n-i+1})]\right]/n \tag{2.48}$$

where Z_i is the standardized variable

$$Z_i = F_T(x_{(i)}).$$

To test for normality, the null hypothesis is formulated as H_0: $F_E = N(\bar{x}; s^2)$ and the variable

$$Z_i = \Phi[(x_{(i)} - \bar{x})/s]$$

represents the quantities of the normal distribution. When $AD > D_{1-\alpha}$, the null hypothesis about normality is rejected. The quantile $D_{0.95}$ may, for large samples, be approximated by

$$D_{0.95} = 1.0348\ (1 - 1.013/n - 0.93/n^2) \tag{2.49}$$

Problem 2.24 *Examination of the normality of five samples*

Apply the normality tests to samples from the (A) rectangular, (B) normal, (C) exponential, (D) Laplace, and (E) log–normal distributions.

Data: as for Problem 2.2.
Program: Chemstat: Basic Statistics: Assumptions testing.
Solution: Two different tests were applied to the five samples, and the results are given in Table 2.9.

Table 2.9—Normality tests made at the significance level $\alpha = 0.05$; H_0: $F_E = N(\bar{x}; s^2)$ *vs.* H_A: $F_E \neq N(\bar{x}; s^2)$

Sample	C_1 test: H_0 is	*AD* test: H_0 is
(a)	accepted	accepted
(b)	accepted	accepted
(c)	rejected	rejected
(d)	rejected	rejected
(e)	rejected	rejected

Conclusion: Neither test can distinguish between the rectangular and normal distributions. The other distributions are correctly indicated as not being normal.

2.7.4 Testing for homogeneity of sample

Sample heterogeneity becomes evident when a sample contains outliers or when the sample can be logically divided into several sub-samples, each of which can be analysed separately. Testing the difference between sub-sample averages can indicate whether the separation into sub-samples can be taken as significant or not. We limit ourselves here to the situation when outliers exist in a data batch. Outliers significantly differ from all other values and can be readily identified by EDA plots. Outliers cause distortion of the estimates $\bar{x}$ and s^2 and may impair the subsequent statistical testing.

There are many different techniques for identifying outliers, when a normal distribution of data can be assumed. One of the simplest and most efficient methods seems to be *Hoaglin's modification of inner bounds* B_L^* and B_U^* (Fig. 2.7)

$$B_L^* = \tilde{x}_{0.25} - K(\tilde{x}_{0.75} - \tilde{x}_{0.25}) \tag{2.50a}$$

and

$$B_U^* = \tilde{x}_{0.75} + K(\tilde{x}_{0.75} - \tilde{x}_{0.25}) \tag{2.50b}$$

where the value of parameter K is selected such that the probability $P(n, K)$ that no observation from a sample of size n will lie outside the modified inner bounds $[B_L^*, B_U^*]$ is sufficiently high, for example, $P(n, K) = 0.95$. For $P(n, K) = 0.95$ and $8 \leq n \leqslant 100$, Hoaglin [9] uses the following equation for calculation of K:

$$K \approx 2.25 - 3.6/n \tag{2.51}$$

All elements lying outside the modified inner bounds $[B_L^*, B_U^*]$ are considered to be outliers.

Problem 2.25 *Identification of outliers in a sample*
Find outliers in samples from the (A) rectangular, (B) normal, (C) exponential, (D) Laplace and (E) log–normal distributions with the simplifying assumption that each sample comes from a normal distribution.

Data: as for Problem 2.2.
Program: Chemstat: Basic Statistics: Assumptions testing.
Solution: According to Hoaglin's modification of inner bounds B_L^* and B_U^*, n_{out} values are excluded from each sample as outliers (Table 2.19)

Table 2.10—Excluding outliers by an external hinges technique

Sample	B_L^*	B_U^*	n_{out}
A	−0.956	1.928	0
B	−3.38	3.27	0
C	−2.289	3.941	2
D	−4.21	4.09	2
E	−6.358	10.04	5

Note: the simplifying assumption of normality gives misleading results for samples with skewed distribution, or samples of a distribution with long tails. The apparent outliers should not therefore be excluded from these samples.

2.8 SUMMARY OF THE PROCEDURE FOR EDA AND CDA OF UNIVARIATE DATA

The extent of exploratory (EDA) and confirmatory (CDA) data analysis of univariate data is best chosen according to experience from previous data analyses. Here, we consider two common situations:

(a) the treatment of routine data, and
(b) the treatment of new data when no preliminary information is available.

(a) The analysis of routine data

With routine data, some knowledge of the sample distribution is assumed–it is usually normal, and the data elements are homogeneous and independent. Tests for examining all assumptions about data should include

(i) a test for minimal sample size;
(ii) a test for independence of sample elements;
(iii) a test for normality;
(iv) a test for homogeneity of sample.

Graphical EDA techniques such as the rankit plot (G15) and quantile–box plot (G10) are often used.

When no preliminary information about the data is available, the full range of EDA plots should be followed by determination and construction of the sample distribution. When no suitable distribution has been found, a power transformation of the data is recommended.

To summarize a batch of experimental data, the quantile–box plot (G10) is *always* used.

(b) The analysis of new data

There are several cases that require different strategies for the EDA and CDA procedures.

Case I. No independence of sample elements

When the sample elements are not proved to be independent, a danger of systematically biased and over evaluated estimates for a positive value of ρ_α Eq. (2.46) arises. Therefore, a new logical analysis of the experimental equipment and data measurement procedures is necessary: after an improvement in the experimental strategy, the new data should be examined again.

Case II. The sample distribution is not normal

The actual sample distribution is not normal in nature, or outliers are present in the data. When the distribution is not normal, the deviation can be in the lengths of tail or in skewing. When tails differ in length, robust estimates (Section 3.3) may be used, or a power transformation chosen. For skewed distributions, a power transformation should be always used. When a power transformation is successful and the optimal value λ is found, the estimates of the parameters of location and spread can be calculated and re-expressed in the measure of the original variables. If the power transformation is not successful, exploratory data analysis can be used to find a suitable approximate theoretical distribution. The estimates of location and scale can then be found as appropriate.

When the actual distribution is strongly skewed, with skewness $\hat{g}_1$, the modified random variable t_c is used,

$$t_c = \left[(\bar{x} - \mu) + \frac{\hat{g}_1}{6\sigma^2 n} + \frac{\hat{g}_1}{3\sigma^4} (\bar{x} - \mu)^2 \right] \frac{\sqrt{n}}{s} \tag{2.52}$$

where t_c has the Student distribution with $n - 1$ degrees of freedom. In practical calculations the variance σ^2 is replaced by its unbiased estimate s^2 and the skewness $\hat{g}_1$ by its unbiased estimate

$$\hat{g}_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} (x_i - \bar{x})^3 \tag{2.53}$$

For construction of the confidence intervals according to Eq. (2.32)

$$H_L \leq \mu \leq H_U,$$

the quadratic equation for μ should be solved. The limits H_L and H_U will then be

$$H_L = \bar{x} + \frac{1 - \sqrt{d_1}}{2C_2} \tag{2.54}$$

$$H_U = \bar{x} + \frac{1 - \sqrt{d_2}}{2C_2} \tag{2.55}$$

where

$$C_2 = \hat{g}_1/3s^4 \tag{2.56a}$$

$$d_1 = 1 - 4C_2(C_1 - C) \tag{2.56b}$$

$$d_2 = 1 - 4C_2(C_1 + C) \tag{2.56c}$$

$$C_1 = \hat{g}_1/(6s^2n) \tag{2.56d}$$

$$C = t_{1-\alpha/2}(n-1)s/\sqrt{n} \tag{2.56e}$$

The confidence interval of the mean, $H_L \leq \mu \leq H_U$, can also be used for statistical inference about this parameter of location.

Case III. Sample not homogeneous

It should first be considered whether the distribution is skewed or not, because some points would appear to be outliers for a symmetrical (normal) distribution, but would be accepted in a skewed distribution. When some points may be extremes or outliers there are two alternatives. (1) Exclude the outliers from the data batch. For a small sample size, this may lead to loss of valuable information. (2) Apply robust methods. In both cases the experimenter should be consulted about the suspect points from the physical point of view, in order to consider the possibility of gross errors.

Case IV. The sample size is not sufficient

The best solution is to carry out new experimental measurements. As a general rule, when the variance of the data is small, a relatively smaller sample size will be required for any given precision of estimate. When no extra experiments can be carried out, the technique for small sample sizes should be applied (Section 3.3.4). This is convenient for routine data analysis, but for new data, exploratory data analysis should be used first, so that any statistical peculiarities of the sample are determined.

2.9 ADDITIONAL SOLVED PROBLEMS

Problem 2.26 *Use of EDA in the determination of phosphorus in blood*

A random sample of fifty milk cows was taken from a herd of 2900 cows, and the blood of each was analysed for phosphorus content, in mmoles per litre. Apply EDA to determine the sample distribution, and to decide whether the measures of location and spread should be computed from classical moment or robust quantile estimators. *Data*: the phosphorus content [mmol/1]; $n = 50$.

1.74, 2,05, 2.35, 2.21, 1.50,
2.17, 2.72, 1.91, 1.95, 2.20, 2.29, 1.84,
2.21, 1.40, 2.51, 2.19, 1.51, 2.09, 2.28,
1.43, 2.50, 2.44, 2.08, 2.27, 1.79, 2.16,
2.03, 2.08, 2.12, 1.96, 1.96, 1.79, 2.76,
1.83, 2.10, 2.17, 1.55, 2.14, 1.92, 2.28,
2.43, 2.17, 2.21, 1.82, 2.28, 2.17, 2.40,
1.87, 1.98, 2.56.

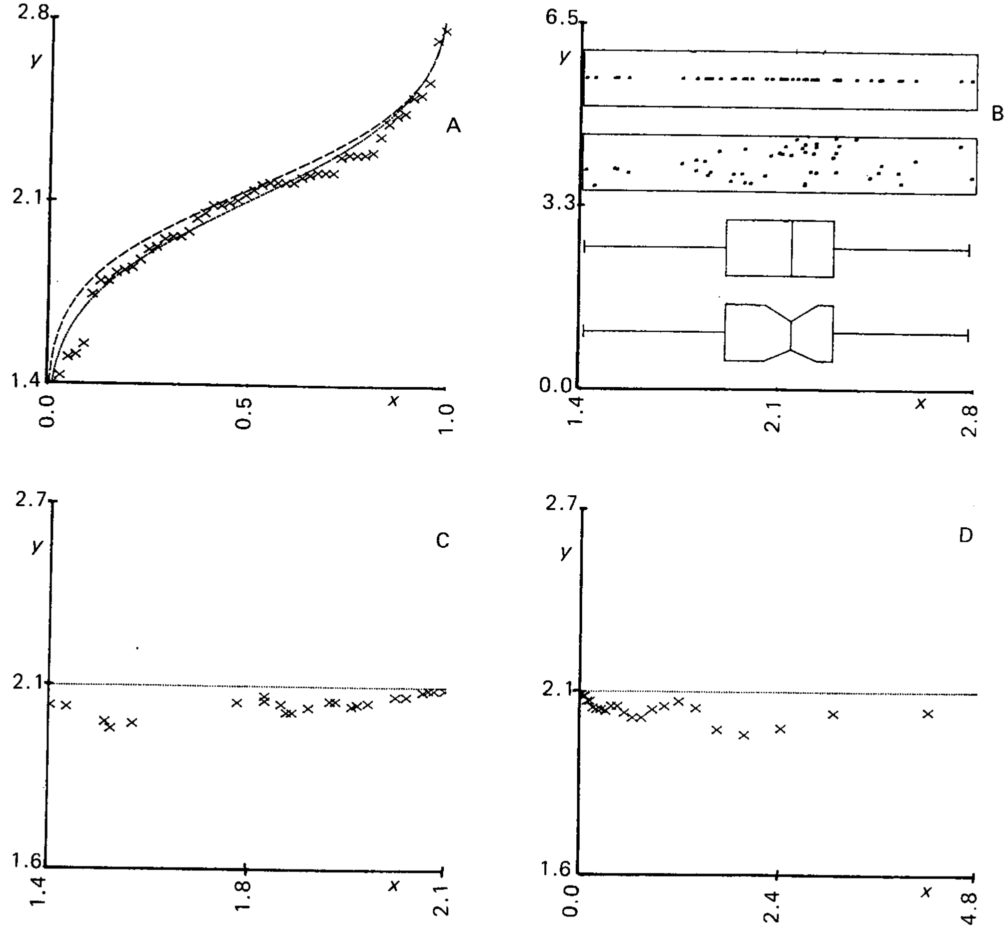


Fig. 2.27—The EDA diagnostics for Problem 2.26: (A) the quantile plot G1; (B) dot diagram G2, and jittered dot diagram G3; the box-and-whisker plot G4, and the notched box-and-whisker plot G5; (C) the midsum plot G6; (D) the symmetry plot G7.

Program: Chemstat: Basic Statistics: Exploratory continuous.

Solution: The diagnostic graphs of EDA are used as follows. The quantile plot G1 shows small deviations from the normal distribution, especially at low values. The dot diagrams G2–G3, and the box-and-whisker plots G4–G5 (Fig. 2.27) indicate that five low measurements and two high measurements differ from the rest of the sample. The distribution is skewed to lower values, and both skewness and kurtosis differ from the expected values (the diagnostics G6, G7 and G8 in Figs. 2.27 and 2.28).

The non-parametric kernel estimate of probability density function G12 in Fig. 2.28 and the histogram G13 indicate that the distribution is skewed to lower values in comparison with the normal distribution. Therefore the mode $\tilde{x}_{\text{mod}}$ is also shifted from the arithmetic mean $\bar{x}$.

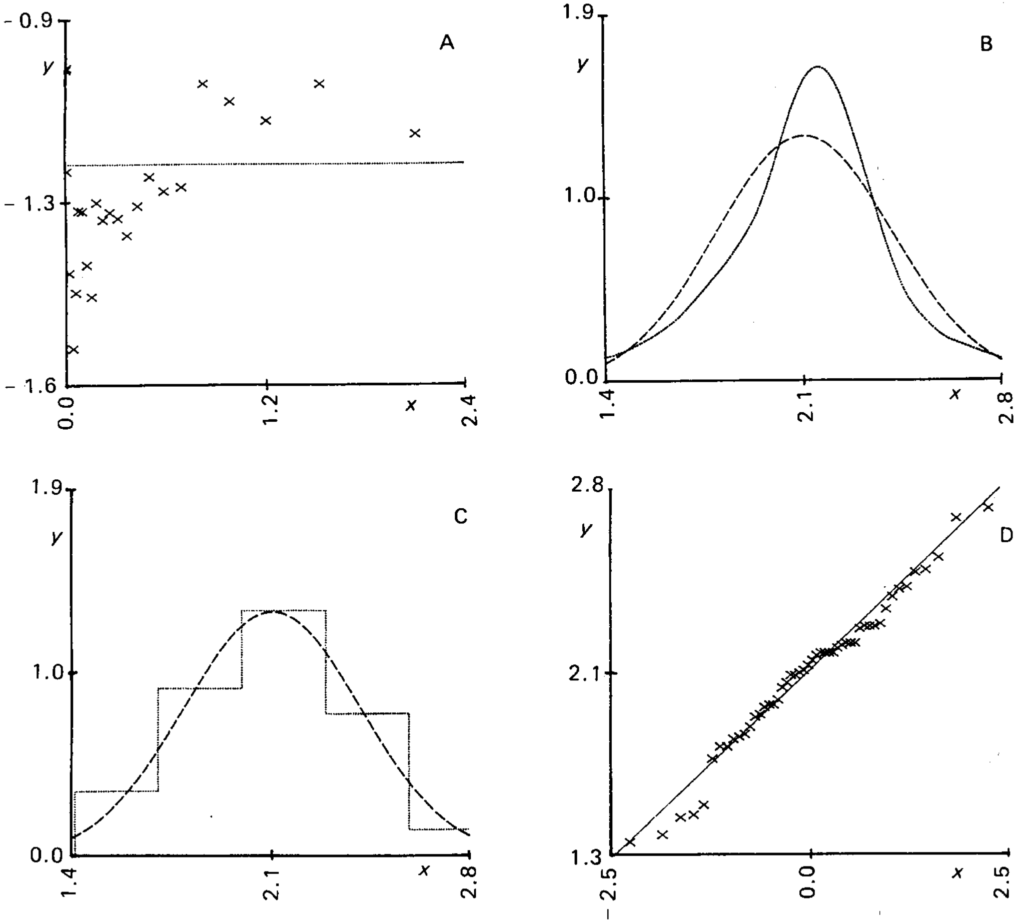


Fig. 2.28—The EDA plots for Problem 2.26: (A) the kurtosis plot G8; (B) the kernel estimate of probability density function G12; (C) the histogram G13; (D) the rankit plot G15.

Both the rankit plot G15 (Fig. 2.28) and the modified rankit plot G16 (Fig. 2.29) prove that there is a significant separation of five lowest and two highest values of the sample. The quantile-box plot G10 (Fig. 2.29) and the numerical values of quantile measures of location and scale in Table 2.11 show (1) there is asymmetric skewing, which is largest in the quartile range, and (2) the five lowest and two highest values are outliers.

To improve the distribution the power transformation was used. A Hines–Hines selection graph was analysed and a convenient transformation was found to be in the range 1.5–2. With $\lambda = 1.6$, the distribution is near to normality (Table 2.11). The plot of the logarithm of the likelihood function G21 gave $\hat{\lambda} = 1.5$. The 95% confidence interval of λ is so broad that it also covers the value $\lambda = 1$; thus from a statistical point of view this transformation is not significant.

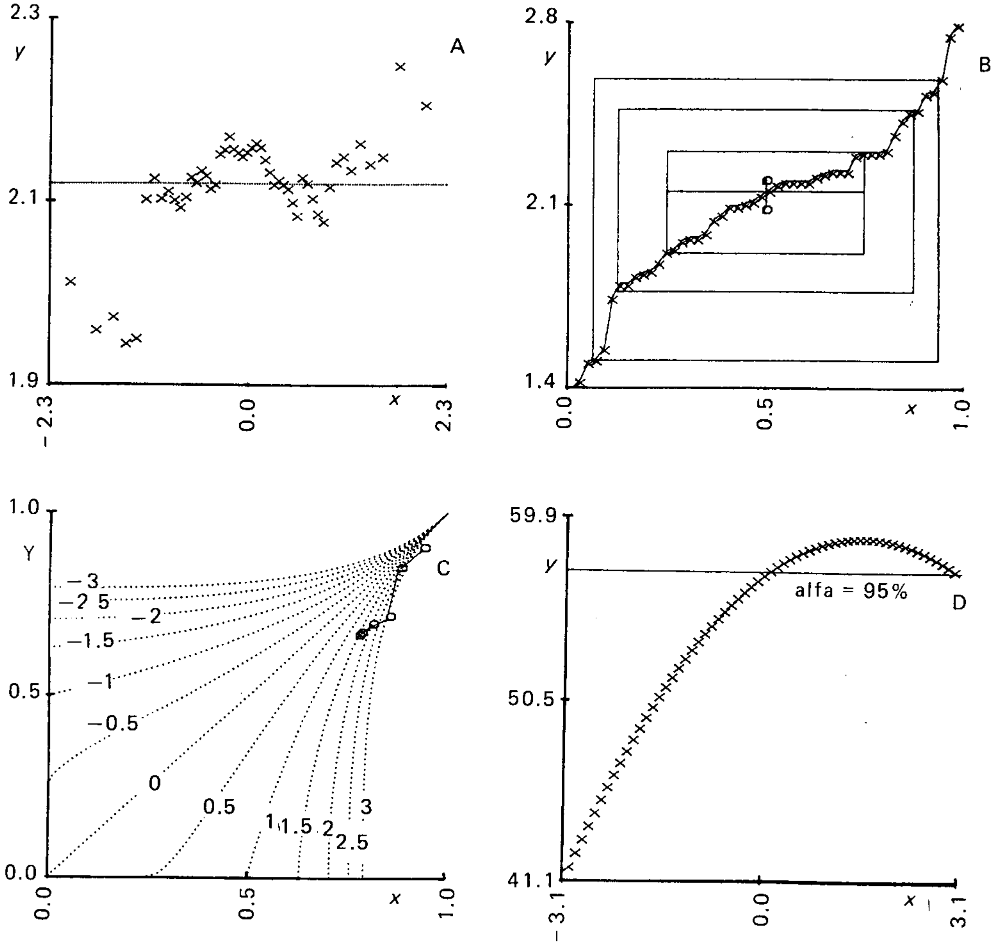


Fig. 2.29—The EDA diagnostics for Problem 2.26: (A) the modified rankit plot, G16, (B) the quantile–box plot, G10, (C) the Hines–Hines selection graph, G20, and (D) the plot of the logarithm of the likelihood function, G21.

Table 2.11(a)—The quantile measures of location

Quantile	*P*	Lower quantile	Upper quantile	Range
Median	0.5	2.130	2.130	–
Quartile	0.25	1.900	2.280	0.380
Octile	0.125	1.759	2.436	0.678
Sedecile	0.0625	1.502	2.551	1.049

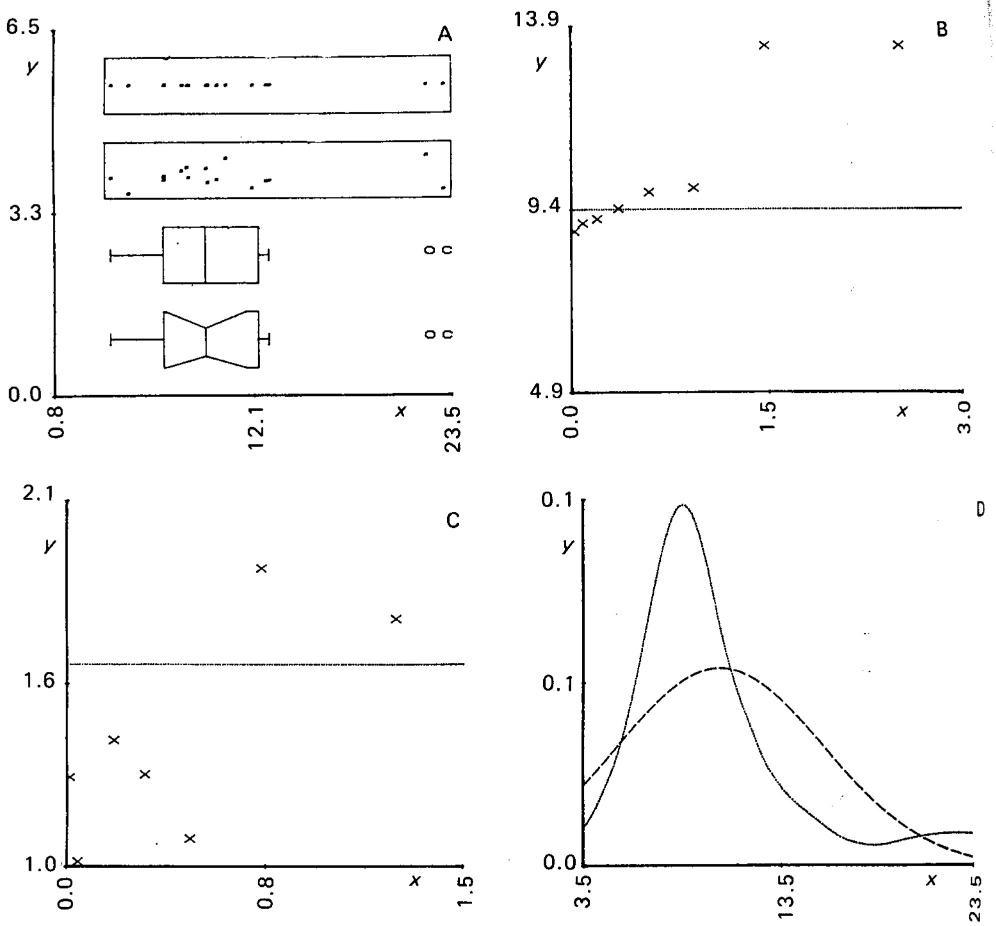


Fig. 2.30—The EDA diagnostics for Problem 2.27: (A) the dot diagrams G2 – G3 and the box-and-whisker plots G4 – G5, (B) the symmetry plot G7, (C) the kurtosis plot G8, and (D) the kernel estimation of the probability density function G12.

Table 2.11(b)—The quantile measures of spread and shape (the values for the transformed data are given in brackets).

Quantile	*P*	Midsum	Skewness	Tails length of	
				the sample distribution	the normal distribution
Quartile	0.25	2.090 (1.354)	0.105 (0.094)	0.000 (0.000)	0.000 (0.000)
Octile	0.125	2.098 (1.378)	0.048 (0.0280	−1.025 (−0.037)	0.578 (0.517)
Sedecile	0.0625	2.026 (1.305)	0.099 (0.068)	0.125 (0.725)	0.801 (0.801)

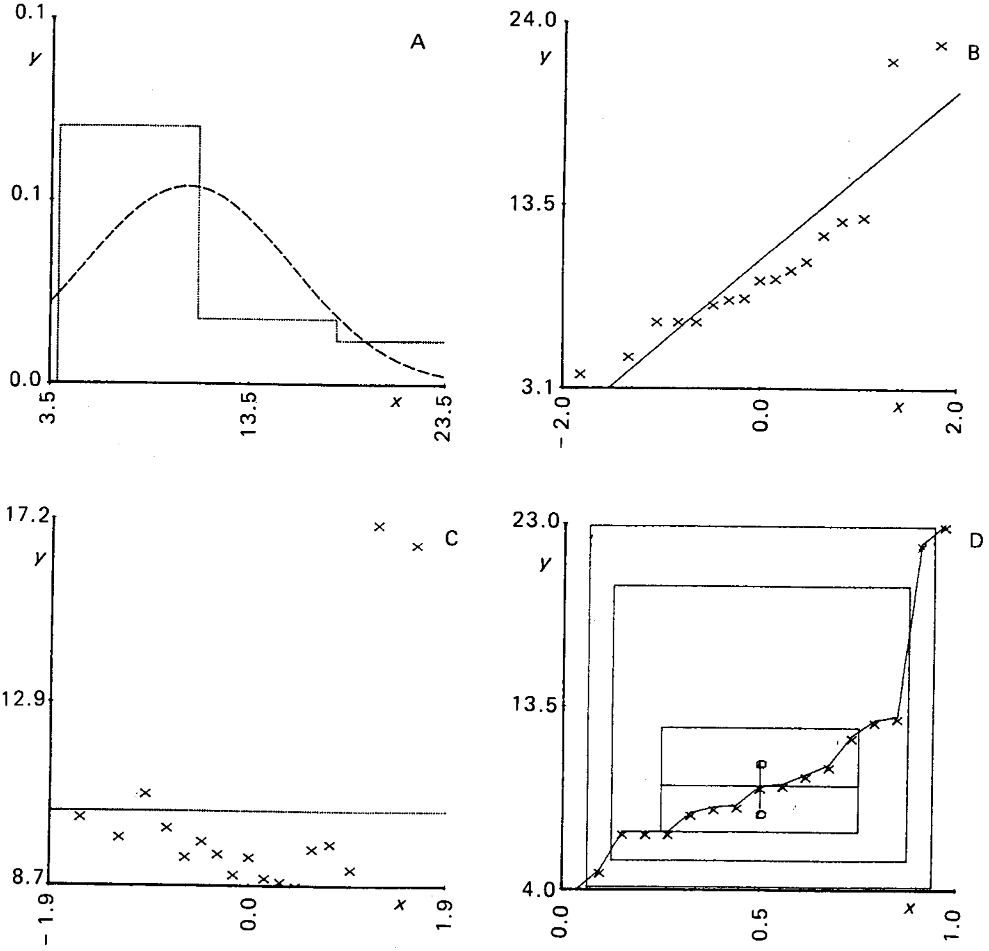


Fig. 2.31—The EDA diagnostics for Problem 2.27: (A) the histogram, G13. (B) the rankit plot, G15, (C) the modified rankit plot, G16 and (D) the quantile-box plot, G10.

Conclusion: The sample batch deviates from normality, and this has significant influence on the measures of location and spread. The robust quantile estimators are more suitable for these data.

Problem 2.27 *EDA in determination of trace copper in kaolin*
Trace copper was determined in a standard sample of kaolin, and the values were arranged in increasing order. Examine the type of sample distribution and decide what type of measures of location and spread should be used.
Data: copper concentration [ppm]; $n = 17$.

4, 5, 7, 7, 7, 8, 8.3, 8.4, 9.4, 9.5, 10, 10.5, 12, 12.8, 13, 22, 23.

Program: Chemstat: Basic Statistics: Exploratory continuous, power transformations, assumptions testing.

Solution: On examination of the EDA diagnostics, the following were noted. The dot diagrams G2−G3 and the box-and-whisker plots G4−G5 (Fig. 2.30) indicate two outliers, but these could be accepted if the distribution is skewed.

The symmetry plot G7 and the kurtosis plot G8, the non-parametric kernel estimation of the probability density function G12, (Fig. 2.30) and the histogram G13 (Fig. 2.31) indicate that the distribution is skewed towards higher values.

The rankit plot G15, with a convex increasing shape, confirms that the distribution is skewed to higher values. The modified rankit plot G16 (Fig. 2.31) indicates that, if the two highest and two lowest points are omitted, the distribution would appear to be normal. The box plot with quantiles G10 and corresponding quantile measures confirm that the distribution is skewed because of the two highest points.

The second part of EDA concerns the search for a suitable symmetric transformation of the data. The selection graph G20 (Fig. 2.32) shows that the optimal power reaches a value above −0.5 in the range near zero which corresponds to a logarithmic transformation.

From the plot of the logarithm of the likelihood function for the Box–Cox transformation the maximum of the curve is at $\lambda = -0.2$. The corresponding 95% confidence interval does not contain the value $\lambda = 1$, so this transformation is statistically significant. The quantile–box plot G10 together with the rankit plot G15 (Fig. 2.32) show that there is a significant improvement in the distribution symmetry with the transformation $\hat{\lambda} = -0.2$.

The measures of location, spread and shape for the original data have the values $\bar{x} = 10.406$, $s^2(x) = 26.834$, $\hat{g}^1 = 1.399$, $\hat{g}^2(x) = 4.272$. After a logarithmic transformation ($\lambda = 0$) the values are 2.243, 0.203, 0.304 and 3.070, and after a power transformation ($\lambda = -0.2$) they are 1.795, 0.081, 0.041 and 3.052.

By rough re-expression [Eq. (2.36)] $\bar{x}_R = \exp(\bar{x}^*) = 9.337$. The corresponding confidence limits are $I_L = 7.742$ and $I_U = 11.878$ (Eq. (2.39a,b)). Quantile $t_{0.975}(16) = 2.12$. By the approximate re-expression [Eq. (2.37)] $\bar{x}_R = 10.42$ with $I_L = 8.272$ and $I_U = 13.147$ [Eq. (2.42)].

In the comparison of the sample distribution with the theoretical exponential one, the correlation coefficient r_{xy} of the Q−Q plot G14 is found to be 0.967, while for the log–normal one, r_{xy} is 0.961.

Conclusion: The assumption of a log–normal distribution is acceptable. Because of the small sample size it is difficult to be certain whether there are outliers in the sample, or if the sample distribution is of skewed log–normal or of skewed exponential nature.

Problem 2.28 *Investigation of number of micro-organisms*

In biomedical laboratories, the counting of micro-organisms in individual fields of a square net under the microscope is common. Micro-organisms were counted in $n = 118$ rectangular fields. The number of fields n_x containing x (=0, 1, 2, ..., 6)

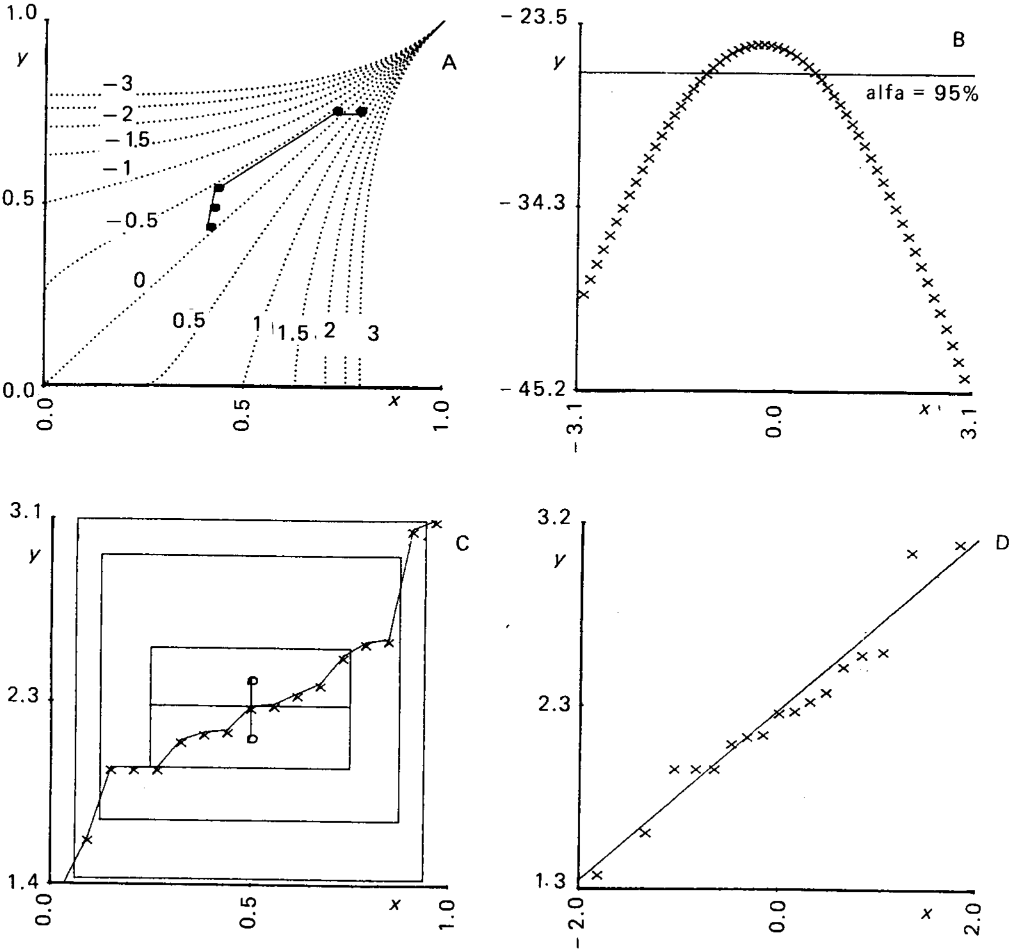


Fig. 2.32—The transformation of data: (A) the selection graph G20, (B) the plot of the logarithm of the likelihood function G21, (C) the quantile–box plot G10 for the transformed data, (D) the rankit plot G15 for the transformed data.

micro-organisms make up the sample. It is assumed that the numbers of micro-organisms in the individual fields of square net have the Poisson distribution with $\lambda_0 = 2.960$. Examine this assumption and estimate parameter λ and its confidence interval.

Data:

x	0	1	2	3	4	5	6
n_x	5	19	26	26	21	13	8

Program: Chemstat: Basic statistics: Exploratory discrete.

Solution: The maximum likelihood estimate of parameter λ is calculated from Eq. (3.34) to be $\hat{\lambda} = 2.932$. The calculated numbers of fields n_x are estimated from

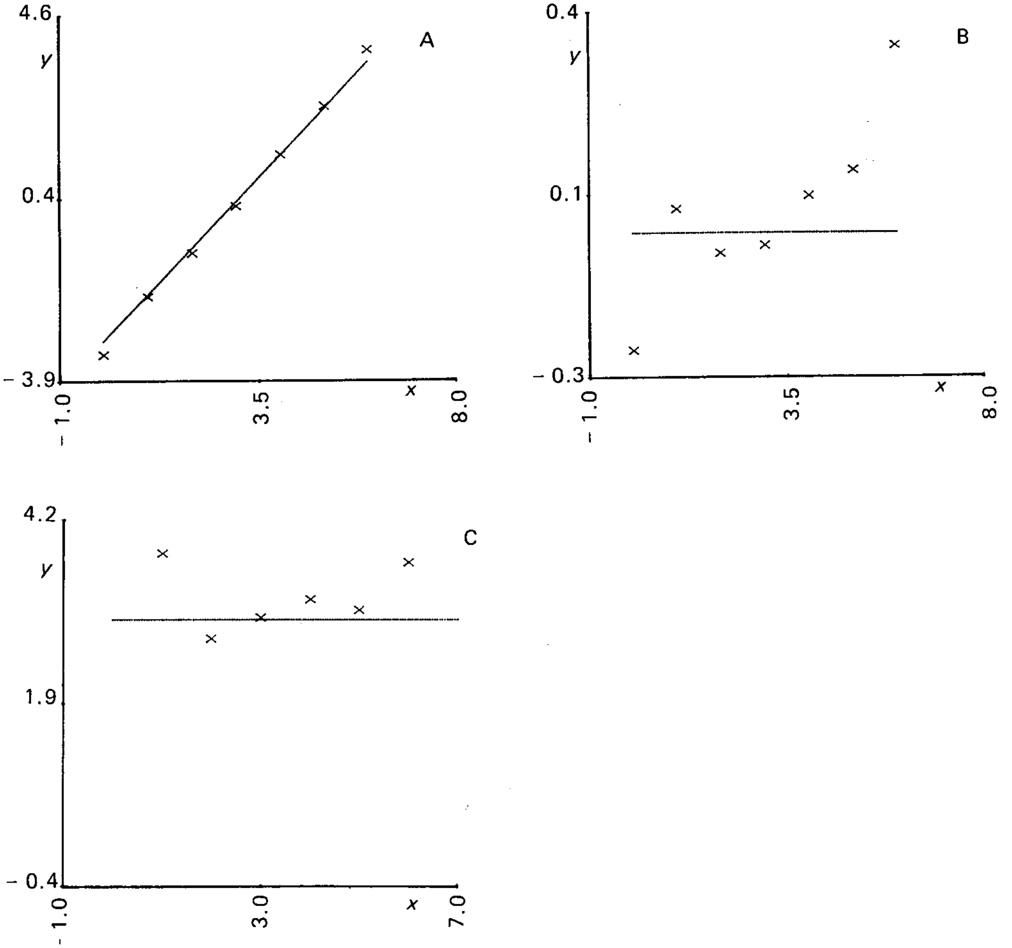


Fig. 2.33—The EDA diagnostics for a discrete distribution: (A) the Poisson plot, G18, (B) the modified Poisson plot, G19, and (C) the frequency ratio plot, G17.

$$\hat{n}_x = p(x, \lambda)n$$

and are given in Table 2.12. The function $p(x, \lambda)$ is found from Eq. (3.31). The lower (L) and upper (U) limits of $\ln(n_x)$ are calculated from

$$L = \ln(n_x) - 1.96\ [(1 - \hat{p})/(n_x - (0.47 + 0.25\hat{p})\ \sqrt{n_x})]^{1/2}$$

$$U = \ln(n_x) + 1.96\ [(1 - \hat{p})/(n_x - (0.47 + 0.25\hat{p})\ \sqrt{n_x})]^{1/2}$$

The Poisson plot G18 shows significant linearity and hence confirms the hypothesis about the Poisson distribution of the numbers of micro-organisms. Significant deviations occur only at $x = 0$ and $x = 6$. Confirmation of the assumed value for parameter λ_0 leads to the modified Poisson plot G19 (Fig. 2.33), which also indicates two possible outliers, at $x = 0$ and at $x = 6$. The frequency ratio plot G17 shows that (a) when one outlier (for $x = 0$) is excluded the non-zero straight line with a non-zero intercept suggests a binomial distribution; and (b) when two outliers ($x = 0$ and $x = 6$) are excluded, the non-zero trend of the straight line is not significant, so the distribution can be of Poisson nature.

Table 2.12—The quantiles of the Poisson distribution

x	n_x	$\hat{n}_x$	$\ln(n_x)$	L	U
0	5	6.3	1.60	0.48	2.43
1	19	18.4	2.94	2.46	3.34
2	26	27.0	3.26	2.86	3.58
3	26	26.4	3.26	2.86	3.58
4	21	19.4	3.04	2.59	3.41
5	13	11.4	2.56	1.95	3.05
6	8	5.5	2.07	1.24	2.72

REFERENCES

[1] J. W. Tukey, *Exploratory Data Analysis*. Addison Wesley, Reading, Mass., 1977.

[2] J. Chambers, W. Cleveland, W. Kleiner and P. Tukey, *Graphical Methods for Data Analysis*, Duxbury Press, Boston, 1983.

[3] D. C. Hoaglin, F. Mosteler and J. W. Tukey, *Exploring Data Tables, Trends and Shapes*, Wiley, New York, 1985.

[4] D. W. Scott and S. J. Sheater, *Commun. Statist.*, 1985, **14**, 1353.

[5] M. Lejenne, Y. Dodge and E. Koelin: *Proc. Conf. COMSTAT'82, Toulouse*, p. 173, Vol. III.

[6] D. C. Hoaglin, F. Mosteler and J. W. Tukey, eds., *Understanding Robust and Exploratory Data Analysis*, Wiley, New York, 1983.

[7] K. Kafander and C. H. Spiegelman, *Comput. Stat. Data Anal.*, 1986, **4**, 167.

[8] W. G. S. Hines and R. J. H. Hines, *Am. Statist.*, 1987, **41**, 21.

[9] D. C. Hoaglin, B. Iglewicz and J. W. Tukey, *J. Am. Statist. Assoc.*, 1986, **81**, 991.

[10] K. Stoodley, *Applied and Computational Statistics*, Ellis Horwood, Chichester, 1984.

3

Statistical analysis of univariate data

After the exploratory data analysis, the next step is statistical analysis. With small samples the statistical characteristics are estimated directly, but with large samples the data are divided into classes and the statistical characteristics of each class are estimated. For univariate data, a single property or quantitative parameter is examined.

Univariate samples come from a population with an unknown probability distribution. A univariate population (or ensemble) is considered to be a set in which only one property is studied, and one quantity with frequency N is measured. The population is characterized both by measures of the *location*, i.e. the level at which the quantity values vary, by the degree of the *dispersion* (or *spread, scatter, variability*) of the quantity of interest, and by the *shape* of the distribution.

In chemical practice, the large population of all measured quantities is rarely available. Therefore, statistical analysis examines a *representative random sample* (or *sample*) of n measurements. A representative random sample has the properties:

(1) All sample elements $\{x_i\}$, $i = 1, \ldots, n$, are random quantities from the *same distribution*; that is the sample is homogeneous.

(2) All sample elements x_i are selected *independently*. The choice of one element does not affect the value of any other element in sample.

The sample is characterized by information about the *mean value* of the sample elements and their *variability* around this mean. In addition, there may be interest in the shape of the sample distribution. Statistical characteristics of location, spread and shape are called *the sample characteristics.* From these sample characteristics, the measures for the population are derived.

In statistical analysis it is assumed that the sample distribution is the same as the population distribution. For continuous random quantities the sample distribution is described by the probability density function $f(x, \theta)$, and for discrete random quantities by the probability density function $p(x, \theta)$. The probability density depends on vector $\boldsymbol{\theta}$, which contains the parameters of location, scale and shape. The purpose of an analysis is the estimation of these parameters. Since these estimates are also random quantities, their distribution or at least their characteristics should be estimated.

The main task of statistical analysis is to collect information about a population, so the sample estimates are used to find confidence intervals of parameters. With a given probability, the confidence interval of a population parameter will include the true value of this "unknown" parameter. Statistical testing of hypotheses about "unknown" parameters of the population is also carried out.

A main purpose of chemometrics experimentation is to draw inferences about a population from samples of the population. We can identify three different types of inferences, namely:

(1) parameter (point) estimation;
(2) interval estimation,
(3) hypothesis testing.

If we want to make the best estimate of one or more parameters of a probability distribution, the problem is said to be parameter estimation. By parameters we usually mean measures of location, scale and the shape of probability distribution. Estimation of a single value for a parameter is called *point estimation.*

Interval estimation is concerned with estimation of the interval that will include the population parameter with a specified probability. An interval estimate is more informative than a point estimate.

Interval estimation is closely related to *hypothesis testing.* In hypothesis testing, one or more propositions are selected about parameters of population probability distribution. Hypotheses are stated, a criterion of some sort is formulated, and a decision is reached.

"Good" estimates should, if possible, be: (1) unbiased, (2) consistent, (3) efficient, and (4) sufficient.

(1) *Unbiased.* An estimate $\hat{\theta}$ of a parameter θ is said to be unbiased if its expected value, $E(\theta)$, is equal to the population value θ.
(2) *Consistent.* An estimator is said to be consistent if the estimate tends to approach the population value more and more closely as the sample size is increased; that is, if $E[(\hat{\theta} - \theta)^2]$ approaches zero as the sample size n approaches infinity.
(3) *Efficient.* The estimate $\hat{\theta}$ is efficient when its variance around the population value θ is the smallest of all the possible estimates. If two point estimates of a single parameter θ are calculated from the same sample size n, the one with the smaller variance has the higher efficiency.
(4) *Sufficient.* If $\hat{\theta}$ is a sufficient estimate of population parameter θ, then it contains all sample information. An estimate of θ is denoted as *best unbiased* if it is unbiased, efficient and sufficient simultaneously.

We now turn to methods for estimation of parameters.

3.1 POINT ESTIMATES FOR PARAMETERS OF LOCATION, SPREAD AND SHAPE

3.1.1 Maximum likelihood method

There are many varied methods of point estimation. Regression uses the least-squares method, but for univariate samples the simple method of moments is often used. A well-known and desirable estimation procedure is that of maximum likelihood, which leads asymptotically to estimates that are efficient but not necessarily unbiased. A desirable feature of the maximum likelihood method is that, under certain conditions, the estimated parameters are normally distributed for large samples.

Suppose that $p(x; \theta)$ is a probability density function of known form for the *discrete* random variable x. This function contains one or more unknown parameters $\theta_1, \ldots, \theta_m$. One way to estimate the parameters $\theta_1, \ldots, \theta_m$, is to maximize the *likelihood function* $L(\theta, x)$. The estimators $\hat{\theta}_1, \ldots, \hat{\theta}_m$ are known as maximum likelihood estimators.

The likelihood function for one discrete observation x_1 is just the probability density at the point x_1

$$L(\theta_1, \theta_2, \ldots\ \theta_m; x_1) = p(x_1; \theta_1, \theta_2, \ldots\ \theta_m) \tag{3.1a}$$

where the lower case xs and the number subscripts refer to the value of the observation that is inserted into the probability function. The likelihood function based on several discrete observations is the product of the individual functions if the discrete observations are independent

$$L(\theta_1, \theta_2, \ldots, \theta_m; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p(x_i; \theta)$$

$$= p(x_1; \theta) \times p(x_2; \theta) \ldots. \times p(x_n; \theta) \tag{3.1b}$$

For the continuous case, the likelihood function based on one observation x_1 is equal to the probability density

$$L(\theta_1, \theta_2, \ldots, \theta_m; x_1) = f(x_1; \theta_1, \theta_2, \ldots, \theta_m) \tag{3.2a}$$

The likelihood function for several independent observations is a product of densities

$$L(\theta_1, \theta_2, \ldots, \theta_m; x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} f(x_i; \theta)$$

$$= f(x_1; \theta) \times f(x_2; \theta). \ldots. f(x_n; \theta) \tag{3.2b}$$

The estimates of θ_i that are chosen are the ones that give the maximum value of L for the given data $(x_1, \ldots, x_n)$. However, it is more convenient to work with $\ln L$ which is not affected by the position of the extreme.

$$\ln L = \ln p(x_1; \theta) + \ln p(x_2; \theta) + \ldots + \ln p(x_n; \theta)$$

$$= \sum_{i=1}^{n} \ln p(x_i; \theta) \tag{3.3a}$$

or

$$\ln L = \ln f(x_1;\, \theta) + \ln f(x_2;\, \theta) + \ldots + \ln f(x_n;\, \theta)$$

$$= \sum_{i=1}^{n} \ln f(x_i;\, \theta) \tag{3.3b}$$

The value of ln L can be maximized with respect to the vector $\boldsymbol{\theta}$ by equating to zero the partial derivatives of ln L with respect to each of the parameters:

$$\frac{\delta \ln L}{\delta\, \theta_1} = \frac{\sum_{i=1}^{n} \ln f(x_i;\, \theta)}{\delta\, \theta_1} = 0 \tag{3.4a}$$

$$\frac{\delta \ln L}{\delta\, \theta_2} = \frac{\sum_{i=1}^{n} \ln f(x_i;\, \theta)}{\delta\, \theta_2} = 0 \tag{3.4b}$$

$$\ldots \qquad \ldots$$

$$\ldots \qquad \ldots$$

(and analogously for $p(x_i;\, \theta)$ for a discrete random variable). Solution of Eqs. (3.4a), (3.4b), etc. yields the maximum likelihood estimates $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m$. It can be shown that as n approaches infinity the maximum likelihood estimates have the desired asymptotic properties; that is (1) they are best unbiased; and (2) values of $[\sqrt{n}(\hat{\theta}_i - \theta_i)]$ have the normal distribution $N(0,\, D(\theta))$. When one parameter θ is estimated, the variance of the most probable estimate can be expressed by

$$D(\hat{\theta}) = -1/E\left[\frac{\mathrm{d}^2 \ln L}{\mathrm{d}\theta^2}\right] \tag{3.5a}$$

where $D()$ is an operator of the variance and $E()$ is an operator of the mean value.

To estimate a parameter vector $\boldsymbol{\theta}$ for characterization of variability, the covariance matrix $\mathbf{C} = \mathbf{A}^{-1}$ containing on its diagonal the variances $D(\hat{\theta}_j)$, is calculated. The elements A_{ij} of matrix $\mathbf{A}$ are given by

$$A_{ij} = -E\left[\frac{\mathrm{d}^2 \ln L}{\mathrm{d}\theta_j\, \mathrm{d}\theta_i}\right] \tag{3.5b}$$

The distribution of the estimates $\hat{\theta}$ is asymptotically normal $N(\boldsymbol{\theta},\, \mathbf{A}^{-1})$. For real samples of finite size n, the maximum likelihood estimates lose some of their asymptotic properties. They are biased and non-effective.

3.1.2 Sample characteristics

The maximum likelihood estimates of location $\hat{\theta}_1$ and of dispersion $\hat{\theta}_2$ based on data from a normal distribution are the *sample arithmetic mean* $\bar{x}$, ($\hat{\theta}_1 = \bar{x}$), and the *sample variance* s^2, ($\hat{\theta}_2 = s^2$).

The sample arithmetic mean $\bar{x}$ and sample variance s^2 can be used for data sampled from all other distributions. If the sample comes from a symmetric population distribution with the mean μ, variance σ^2 and kurtosis g_2, it can be proved that

$$E(\bar{x}) = \mu \tag{3.6a}$$

$$D(\bar{x}) = \sigma^2/n \tag{3.6b}$$

and

$$E(s^2) = \sigma^2 \tag{3.7a}$$

$$D(s^2) = \frac{\sigma^4}{n}\left[g_2 - \frac{n-3}{n-1}\right] \tag{3.7b}$$

In addition to the sample arithmetic mean and the sample variance, other parameters of location and dispersion can be used:

The *sample mode* (or just *mode*) $\hat{x}_M$ is the most frequently found element value in the sample. The *sample quantiles* are descriptive statistics from exploratory data analysis and are sometimes used to supplement the information obtained from the mean and the variance. The sample values $x_1, \ldots, x_n$ are first of all arranged in order of ascending magnitude $x_{(1)} \leq x_{(2)} \leq \ldots x_{(n)}$. The quantities $x_{(i)}$ are called the *order statistics.* The pth quantile (or percentile) is defined to be the value of x below which p% of the sample values lie. The pth quantile separates the order statistics into two parts so that each contains the required percentage of the sample elements, p% and $(100 - p)$%.

The *sample median* $\tilde{x}_{0.5}$ is the quantile that separates order statistics into two parts: 50% of the elements lie below $\tilde{x}_{0.5}$ and 50% of the elements lie above $\tilde{x}_{0.5}$. The sample median for an odd sample size has the form

$$\tilde{x}_{0.5} = x_{(k)}$$

where $k = (n + 1)/2$. For an even sample size, it is

$$\tilde{x}_{0.5} = (x_{(k)} + x_{(k+1)})/2$$

where $k = n/2$. The mean, mode and median are compared in Fig. 3.1.

The 25th, 50th and 75th percentiles may be called the *first* (or *lower*) *quartile, median* (or *second quartile*) and *third* (or *upper*) *quartile* of the sample. The median represents the maximum likelihood estimate of location for the Laplace distribution. For this distribution the variance of the median is expressed by

$$D_L(\tilde{x}_{0.5}) = \sigma^2/2n \tag{3.8}$$

For the normal distribution, however, the sample median is not efficient (Table 3.1).

For the rectangular distribution, the efficient estimate of location is the *midsum* $\hat{x}_P$ *defined by*

$$\hat{x}_P = (x_{(1)} + x_{(n)})/2 \tag{3.9}$$

where $x_{(1)}$ is the smallest and $x_{(n)}$ the largest element of the ordered sample. The variance of the midsum estimate for the rectangular distribution is defined by

$$D_{\mathrm{R}}(\hat{x}_P) = \frac{6\sigma^2}{(n-1)(n-2)} \tag{3.10}$$

Index R denote the rectangular distribution. The variance of x_P for the normal distribution is much higher.

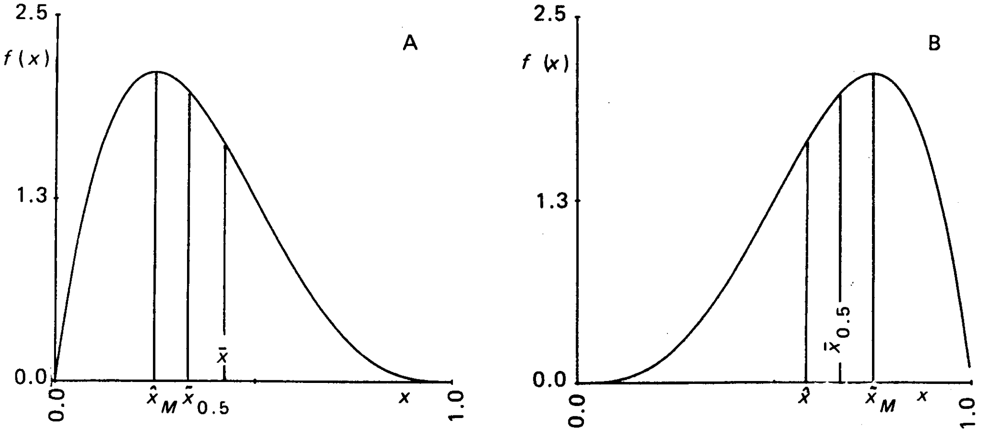


Fig. 3.1—Comparison of three measures of location: mean $\hat{x}$, mode $\hat{x}_{\mathrm{M}}$ and median $\tilde{x}_{0.5}$ for (A) negatively and (B) positively skewed distributions.

Often the condition of constant variance of all sample elements is not maintained. If each x_i has a normal distribution with variance σ_i the statistical weights are calculated as $w_i = 1/\sigma_i^2$. Instead of the sample mean $\bar{x}$, the weighted sample mean $\bar{x}_{\mathrm{w}}$ is computed from:

$$\bar{x}_{\mathrm{w}} = \frac{\sum_{i=1}^{n} x_i w_i}{\sum_{i=1}^{n} w_i} = \frac{\sum_{i=1}^{n} x_i/\sigma_i^2}{\sum_{i=1}^{n} 1/\sigma_i^2} \tag{3.11}$$

The variance of the weighted mean is

$$D(\bar{x}_{\mathrm{w}}) = 1\Big/\left[\sum_{i=1}^{n} 1/\sigma_i^2\right] \tag{3.12}$$

If the relative error has a constant value, $\delta = \sigma/x_i =$ constant, then

$$\sigma_i^2 = \bar{x}_i^2 \times \delta^2.$$

Then $w_i = 1/x_i^2$ and the sample mean is calculated from

$$\bar{x}_{\mathrm{w}} = \frac{\sum_{i=1}^{n} 1/x_i}{\sum_{i=1}^{n} 1/x_i^2} \tag{3.13}$$

with variance

$$D(\bar{x}_{\mathrm{w}}) = \delta^2 \Big/ \left[\sum_{i=1}^{n} 1/x_i\right] \tag{3.14}$$

The dispersion parameters describe the degree of dispersion, (scale, spread, variability or scatter) of the population elements. *The range* is one of the measures of scale which represents the difference between the largest and the smallest value of sample. The *interquartile range* is the quantile estimate of population standard deviation σ defined by

$$R = 0.7413\ (\tilde{x}_{0.75} - \tilde{x}_{0.25}) \tag{3.15}$$

where $\tilde{x}_{0.75}$ is the upper and $\tilde{x}_{0.25}$ the lower quartile. Table 3.1 surveys the sample estimates of location and dispersion, with their variances, efficiency and distribution. Sample estimates are for sample size n, and the sample comes from a population with normal distribution $N(\mu, \sigma^2)$.

Table 3.1—Estimates of location and dispersion for sample of size n from a population with normal distribution $N(\mu, \sigma)$

Parameter	Estimate	Variance estimate	Efficiency	Estimate distribution
Mean μ	$\bar{x}$	σ^2/n	1	$N(\mu, \sigma^2)$
	$\tilde{x}_{0.5}$	$\sigma^2\pi/(2n)$	0.63	$N(\mu, \sigma^2)$
	$\hat{x}_P$	$\sigma^2\pi^2/(24 \ln(n))$	$24 \ln(n)/(\pi^2 n)$	$N(\mu, \sigma^2)$
Variance σ^2	$\hat{\sigma}^2$	$2\ \sigma^4/n$	1*	$N(\sigma^2, D(\sigma^2))$
	s^2	$2\sigma^4/(n-1)$	1	
Standard deviation σ	$\hat{\sigma}$	$\sigma^2/(2n)$	~ 1*	$N(\sigma, D(\sigma))$
	s	$\sigma^2/[2(n-1)]$	1	
	R	$\sim 1.36\ \sigma^2/n$	~ 0.368	
	d	$\sigma^2/[(\pi-2)n]$	~ 0.876	

*biased estimate

Another measure of dispersion is the *mean deviation d* defined by

$$d = \sqrt{\frac{\pi}{2}} \left[\frac{1}{n} \sum_{i=1}^{n} |x_{\mathrm{i}} - \mu|\right] \tag{3.16}$$

where the factor $\sqrt{\pi/2}$ ensures that for normal distribution the value of d approaches that of the standard deviation σ.

The widely used the *coefficient of variation* δ (CV) also known as the *relative standard deviation* s_{rel} (RSD) is given by $100\sigma/\mu$ and may be estimated by

$$\hat{\delta} = s/\bar{x} \tag{3.17}$$

The variance of $\hat{\delta}$ is approximately equal to

$$D(\hat{\delta}) = \hat{\sigma}^2\left[\frac{n + \hat{\delta}^2(2n + 1)}{2n(n - 1)}\right] \tag{3.18}$$

The error $\hat{\delta}$, units %, is called a *relative error.* Relative errors are frequently used in the comparison of the precision of results with different units or magnitudes, and are again important in calculations of error propagation.

To characterize the shape of a distribution, skewness and kurtosis are used. *Skewness* g_1 is a measure characterizing symmetry, which is equal to zero for a symmetrical distribution. Positive values of g_1 indicate smaller scattering of lower values of elements x_I than of the larger values and negative values of g_1 indicate the opposite case. The *moment estimate of skewness* is defined by

$$\hat{g}_1 = \frac{\sqrt{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^3\right]^{3/2}} \tag{3.19}$$

Its asymptotic variance is

$$D(\hat{g}_1) \approx \frac{6\ (n - 2)}{(n + 1)(n + 3)} \tag{3.19a}$$

The effect of skewness on the shape of the probability density function is shown in Fig. 3.2.

Kurtosis characterizes the shape of the distribution near a modal value, and provides a picture of the shape of the distribution peak. For higher values of kurtosis than 3, the distribution has a sharper peak than the normal distribution, while a flat shape is indicated for values of kurtosis lower than 3 (see Fig. 3.3). The *moment estimate of kurtosis* is defined by

$$\hat{g}_2 = \frac{n\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{2}} \tag{3.20}$$

Its asymptotic variance has the form

$$D(\hat{g}_2) \approx \frac{24n\ (n - 2)(n - 3)}{(n + 1)^2(n + 3)(n + 5)} \tag{3.20a}$$

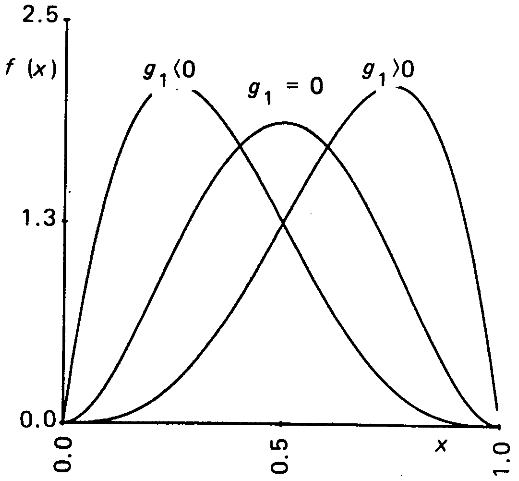


Fig. 3.2—The probability density function for various degrees of skewness: $g_1 < 0$, $g_1 = 0$, $g_1 > 0$.

When a point estimate of any parameter is determined, the variance of the parameter must also be calculated. To achieve the same "precision" of estimates when less effective estimates are used, a greater number of measurements n should be used. To achieve the same parameter precision for data of normal distribution, for example, the calculation of median $\tilde{x}_{0.5}$ needs 1.6 times more measurements that would the arithmetic mean $\bar{x}$.

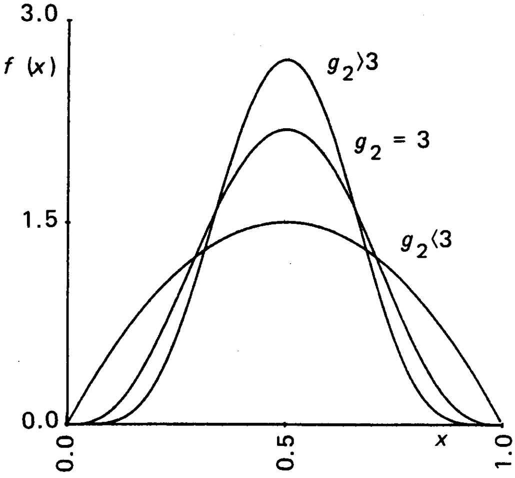


Fig. 3.3—The probability density function for various values of the kurtosis: $g_2 > 3$, $g_2 = 3$, $g_2 < 3$.

Problem 3.1 *Mode, midsum, skewness sand kurtosis of samples from five different distributions*
Calculate the mode, midsum, skewness and kurtosis of a random sample taken from the rectangular, normal, exponential, Laplace and log–normal distribution.
Data: from Problem 2.2
Program: Chemstat: Basic Statistics: One sample analysis.
Solution: For a sample size $n = 50$ taken from five distributions, the mode $\hat{x}_M$, the halfsum $\hat{x}_P$, the skewness $\hat{g}_1$ and the curtosis $\hat{g}_2$ are estimated. The rectangular distribution has no mode, for the exponential, normal and Laplace distributions, the mode is equal to zero, and for the log–normal distribution to 0.135. The midsum for the rectangular distribution is near to the mean value 0.5. Skewness and kurtosis were discussed in Problem 2.2.

Table 3.2—Parameter estimates of location ($\hat{x}_M$, $\hat{x}_P$) and shape ($\hat{g}_1$, $\hat{g}_2$) for a sample of $n = 50$ taken from five distributions

Distribution	Mode $\hat{x}_M$	Midsum $\hat{x}_P$	Skewness $\hat{g}_1$	Kurtosis $\hat{g}_2$
Normal $N(0; 1)$	0.0818	−0.294	−0.137	3.369
Rectangular $R(0.5; 0.083)$	0.841	0.505	−0.0052	1.752
Exponential $E(1; 1)$	0.213	3.09	2.68	11.5
Laplace $L(0; 2)$	0.135	0.877	0.801	6.099
Log-normal $LN(2.71; 47.21)$	0.222	2.428	3.61	16.795

Conclusion: Different estimators can lead to very different values.

For samples from a population with a normal distribution, the random variable

$$t = \frac{\bar{x} - \mu}{s} \sqrt{n} \tag{3.21}$$

has the Student distribution with $(n - 1)$ degrees of freedom (Fig. 3.4). Also, the random variable

$$\chi^2 = \frac{(n - 1)s^2}{\sigma^2} \tag{3.22}$$

has the χ^2 distribution with $(n - 1)$ degrees of freedom. The random variable t and χ^2 are mutually independent. For sufficiently large samples ($n \geq 40$) from the normal distribution, for some estimate $\hat{\theta}$ of parameter θ, the random variable

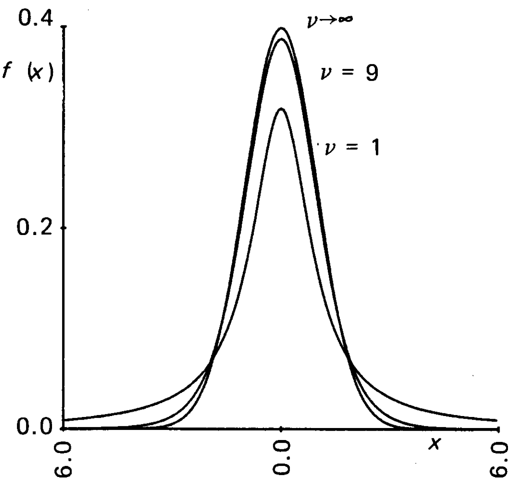


Fig. 3.4—The Student distribution for the degrees of freedom $\nu = 1$, $\nu = 9$, and $\nu \to \infty$, in comparison with the normal distribution.

$$U = \frac{\hat{\theta} - \theta}{D(\hat{\theta})} \tag{3.23}$$

has an approximately standard normal distribution $N(0,\ 1)$. Equation (3.23) is asymptotically valid for any estimate $\hat{\theta}_i$ with variance $D(\hat{\theta}_i)$ determined by the maximum likelihood method, and for any theoretical distribution $f(x,\ \theta)$.

Instead of maximum likelihood estimates, another statistical characteristic called the *likelihood ratio*

$$L_1 = -2\ [\ln L(\bar{\theta}) - \ln L(\hat{\theta})] \tag{3.24}$$

is often used. The *likelihood ratio* L_1 has the χ^2 (1) distribution with 1 degree of freedom. In Eq. (3.24) $\bar{\theta}$ stands for the maximum likelihood estimate $\hat{\theta}$ of parameters θ for which $\hat{\theta}_i = \theta_i$.

It is clear that the distribution of estimators is connected with sample distributions like the Student and χ^2 ones. The Student and χ^2-distributions are both among basic sample distributions which depend only on degrees of freedom, ν. For various values of the degree of freedom ν, the quantiles of the Student distribution and χ^2-distribution may be found in statistical tables.

3.2 INTERVAL ESTIMATES FOR PARAMETERS OF LOCATION AND SPREAD

In the previous section, we described ways of obtaining point estimates of parameters of location, spread and shape. Better than these point estimates are *confidence intervals*. The confidence interval is calculated from the sample estimators. It includes the value of the population parameter within the interval limits, termed *confidence limits*, for a specified degree of assurance, called the *confidence coefficient*. Here, the confidence limits are random variables dependent on the sample.

The parameter of location is then described not by one value ($\bar{x}$) but by two numerical values L_1 and L_2. It is expected that the confidence interval (L_1, L_2) will include the unknown population parameter θ with a preselected probability $(1 - \alpha)$. The degree of trust associated with the confidence statement is called the confidence coefficient; it expresses the degree of certainty or reliability $(1 - \alpha)$ about the unknown population parameter θ.

$$P(L_1 < \theta < L_2) = 1 - \alpha \tag{3.25}$$

where α is called the significance level; the value chosen for α is usually 0.05 or 0.01.

It is useful to know that

(1) the confidence interval is small if the variance of estimate $D(\theta)$ is small,
(2) a large sample size n gives a small confidence interval $\langle L_1, L_2\rangle$, and
(3) higher degrees of certainty $(1 - \alpha)$ give broader confidence intervals $\langle L_1, L_2\rangle$.

Confidence interval $\langle L_1, L_2\rangle$ is referred to as a two-tailed interval, but one-tailed intervals are also used in the chemical laboratory. One-tailed confidence intervals can be

(1) the left-side or lower-tail interval $\langle L_2; \infty)$, or
(2) the right-side or upper-tail interval $(-\infty, L_1\rangle$.

3.2.1 Derivation of the confidence interval

Finding the confidence interval $L_{1,2}$ requires knowledge of the distribution of the parameter in question. Let us find the confidence interval of the population mean of the normal distribution $N(\mu, \sigma^2)$. Let $\bar{x}$ be the mean of a sample of n observations on a normally distributed random variable x with unknown mean μ and known variance σ^2. Then the $100(1 - \alpha)\%$ confidence interval $L_{1,2}$ for μ may be found from

$$\bar{x} - u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + u_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} \tag{3.26}$$

where $u_{1-\alpha/2}$ is the $100(1 - \alpha)\%$ quantile of the standardized normal distribution, (e.g., for $u_{0.975} = 2$, $L_{1,2} = \bar{x} \pm 2\sigma/\sqrt{n}$).

In cases where the sample size n is not large enough and the variance σ^2 is not known, the confidence limit for μ may be found from Eq. (3.26), but using quantiles for the Student t-distribution instead of from the normal one. The $100(1 - \alpha)\%$ confidence limits $L_{1,2}$ are then given by

$$\bar{x} - t_{1-\alpha/2}(\nu)s/\sqrt{n} \leq \mu \leq \bar{x} + t_{1-\alpha/2}(\nu)s/\sqrt{n} \tag{3.27}$$

where $\nu = n - 1$ is the number of degrees of freedom and $t_{1-\alpha/2}(\nu)$ is the $100(1 - \alpha/2)\%$ quantile of the Student distribution. For large sample sizes $(n > 30)$ instead of $t_{1-\alpha/2}(\nu)$ the quantile $u_{1-\alpha/2}$ can be used.

According to Eq. (3.23), the $100(1 - \alpha)\%$ asymptotic confidence interval of any parameter θ may be expressed by

$$\hat{\theta} - u_{1-\alpha/2}\sqrt{D(\hat{\theta})} \leqslant \theta \leqslant \hat{\theta} + u_{1-\alpha/2}\sqrt{D(\hat{\theta})} \tag{3.28}$$

The $100(1-\alpha)\%$ two-tailed confidence interval of the variance σ^2 is given by

$$\frac{\nu s^2}{\chi^2_{1-\alpha/2}(\nu)} \leqslant \sigma^2 \leqslant \frac{\nu s^2}{\chi^2_{\alpha/2}(\nu)} \tag{3.29}$$

where $\chi^2_{1-\alpha/2}(\nu)$ is the upper and $\chi^2_{\alpha/2}(\nu)$ the lower quantile of the χ^2-distribution, and $\nu = n-1$ is the number of degrees of freedom.

Construction of a confidence interval depends on the population distribution from which the sample comes. For example, the variance of the median may be calculated from

$$D(\tilde{x}_{0.5}) = 1/(4nf^2(\text{med}))$$

where $f(\text{med})$ is the value of the probability density function at the position of the median. For the Laplace distribution, $f(\text{med}) = 1/(\sigma\sqrt{2})$ and therefore $D(\tilde{x}_{0.5}) = \sigma^2/2n$, and the confidence interval of the median is given by

$$\tilde{x}_{0.5} - u_{1-\alpha/2} \times 0.707\ s/\sqrt{n} \leqslant \text{med} \leqslant \tilde{x}_{0.5} + u_{1-\alpha/2} \times 0.707\ s/\sqrt{n} \tag{3.30}$$

The Equation (3.30) is valid only if the sample size n is big enough for the median of the Laplace distribution to have approximately normal distribution.

Problem 3.2 *Analysis of five samples with a false assumption of normality*
Make an analysis of samples of size $n = 50$ from rectangular, normal, exponential, Laplace and log–normal distributions, with a false assumption of normality.
Data: from Problem 2.2
Program: Chemstat: Basic Statistics: One sample analysis.
Solution: The assumption of normality is valid only for the sample from the $N(0, 1)$ distribution. Table 3.3 lists statistical characteristics $\bar{x}$, s^2 and limits L_1 and L_2 of the 95% confidence interval of the mean.

Table 3.3—Statistical analysis of samples from five distributions, with a false assumption of sample normality; $n = 50$

Population distribution $X(\mu;\ \sigma^2)$	$\bar{x}$	s	The limits of the 95% confidence interval L_1	L_2
Normal $N(0;\ 1)$	−0.0574	1.088	−0.354	0.239
Rectangular $R(0.5;\ 0.083)$	0.488	0.0865	0.404	0.571
Exponential $E(1;\ 1)$	1.0059	1.362	0.674	1.338
Laplace $L(0;\ 2)$	−0.0246	2.431	−0.468	0.419
Log-normal $LN(2.71;\ 47.21)$	4.077	74.57	1.622	6.532

Conclusion: Although the mean in four sample distributions was estimated with a false assumption of normality, the confidence interval always covers the true mean value. For the Laplace and log-normal distributions the confidence interval is rather broader.

3.3 POINT AND INTERVAL ESTIMATORS FOR SELECTED DISTRIBUTIONS

Chemists would like to replace a large volume of experimental data with a few easily grasped numbers. Under favorable circumstances in the EDA, the experimental data are associated with a known function, a probability density function, which corresponds reasonably with the data.

We shall describe some of the most useful probability density functions that the chemist may meet in the laboratory. Most samples have normal distribution, but there are some tasks when the random quantity is constrained on one side, i.e. it must be in some interval. Then the normality assumption is not warranted. In this section the point and interval estimates for one discrete and five continuous distributions are described. These distributions cover all types of data commonly found in chemical practice. Some details about these distributions may be found in the textbook by Johnson and Kotz [3].

3.3.1 The Poisson distribution

This discrete distribution relates to the number of events that occur in a given interval of time or space when the events occur randomly (in time or space) at a certain average rate. Some examples of random variables for which the Poisson distribution is assumed to apply are: the number of particles emitted from a radioactive source in a given time, the number of typing errors per page of manuscript, the number of calls received at a telephone exchange in a given time period, the number of goals scored by a particular team in a football match. The sample space for the random variable consists of the integers (0, 1, 2, ...).

Suppose a discrete random variable x has a range of possible integer values 0, 1, 2, ... which has a Poisson distribution with the probability function

$$p(x, \lambda) = \frac{\lambda^x \exp(-\lambda)}{x!} \tag{3.31}$$

where λ is a positive parameter (Fig. 3.5).

For the Poisson distribution it can be shown that $\mu = E(x) = \lambda$ and $\sigma^2 = D(x) = \lambda$. That is, for a Poisson distribution the mean and the variance are equal. For a set of $k + 1$ elements, $x = 0, 1, 2, \ldots, k$, the number n_x of observations which have a magnitude x is estimated. From Eqs. (3.31) and (3.1), for n_x replicated values of x, the likelihood function is

$$L(\lambda) = \prod_{x=1}^{k} \lambda^{n_x x} \times \frac{\exp(-n_x \lambda)}{n_x x!} \tag{3.32}$$

After taking logarithms and differentiating

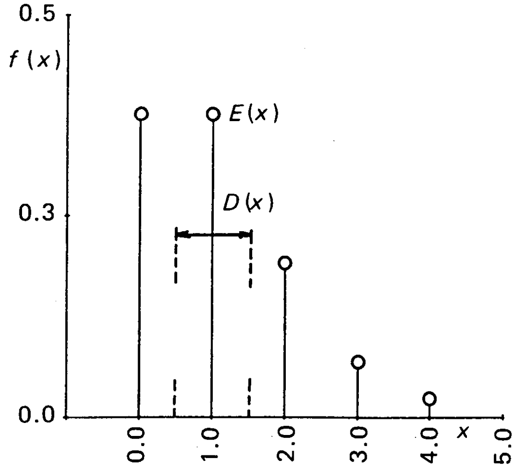


Fig. 3.5—The probability density function for the Poisson distribution with $\lambda = 1$.

$$\frac{\mathrm{d}\ \ln\ L(\lambda)}{\mathrm{d}\ \lambda} = \sum_{x=1}^{k} \left(\frac{n_x x}{\lambda} - n_x \right) = 0 \tag{3.33}$$

The estimate $\hat{\lambda}$ can then be calculated

$$\hat{\lambda} = \frac{\sum_{x=1}^{k} n_x x}{\sum_{x=1}^{k} n_x} = \frac{\sum_{x=1}^{k} n_x x}{n} \tag{3.34}$$

where n is the total sample size. The parameter estimate $\hat{\lambda}$ corresponds to the arithmetic mean. To calculate the variance of λ, Eq. (3.33) must be differentiated again

$$\frac{\mathrm{d}^2\ \ln\ L(\lambda)}{\mathrm{d}\ \lambda^2} = -\frac{1}{\lambda^2} \sum_{x=1}^{k} x n_x$$

and since

$$E\left(\sum_{x=1}^{k} x n_x \right) \simeq n\lambda$$

use of Eq. (3.4) leads to

$$D(\hat{\lambda}) = \lambda / n \tag{3.35}$$

Construction of the confidence interval of parameter λ for a large sample ($n > 30$) is based on an assumption that the random variable $\sqrt{n}(\hat{\lambda} - \lambda)/\sqrt{\lambda}$ has a standardized normal distribution. Although the square of a random variable with normal distribution has the $\chi^2(1)$ distribution, the confidence limits λ_1 and λ_2 of $100(1 - \alpha)\%$ confidence interval of parameter λ may be estimated by solving a quadratic equation

$$\lambda^2 - \left[2\hat{\lambda} + \frac{\chi^2_{1-\alpha}(1)}{n}\right]\lambda + \hat{\lambda}^2 = 0 \tag{3.36}$$

The asymptotic confidence interval of parameter λ based on Eq. (3.23) will be

$$\hat{\lambda} - u_{1-\alpha/2}\sqrt{\hat{\lambda}}/\sqrt{n} \leq \lambda \leq \hat{\lambda} + u_{1-\alpha/2}\sqrt{\hat{\lambda}}/\sqrt{n} \tag{3.36a}$$

It is convenient to use the $100(1-\alpha)\%$ confidence interval of parameter λ, calculated from

$$\frac{\chi^2_{\alpha/2}(2\hat{\lambda}n)}{2n} \leq \lambda \leq \frac{\chi^2_{1-\alpha/2}(2\hat{\lambda}n + 2))}{2n} \tag{3.36b}$$

For large samples ($n > 100$) and for large values of λ (e.g., $\lambda > 10$) the simple expression (3.36a) is recommended.

Problem 3.3 *Confidence interval of cosmic ray "particles"*
A laboratory counter was set up to measure cosmic ray "particles". For the purpose of this example, the number of particles arriving in 0.1-sec intervals was counted. From 200 time intervals the mean of the measurements $\hat{\lambda} = 10.5$ was calculated. With the assumption that the data are described by the Poisson distribution, calculate the 95% confidence interval of the number or particles in 0.1-sec intervals.
Data: the numbers of particles k and frequency of detected ray particles n_k in sample size $n = 200$ are as follows:

k:	0	1	2	3	4	5	6	7	8	9	10	11	12	13
n_k:	0	0	0	2	1	11	12	12	20	22	17	30	20	20

k:	14	15	16	17	18	19	20	21	22	23
n_k:	10	18	6	4	3	1	0	0	1	0

Program: Chemstat: Basic Statistics: Exploratory discrete.
Solution: The 95% confidence interval of parameter λ will be calculated from Eqs. (3.36), (3.36a) and (3.36b).

(1) Equation (3.36): with $n = 200$, $\chi^2_{0.95}(1) = 3.842$ and $\hat{\lambda} = 10.5$ we obtain the quadratic equation

$$\lambda^2 - 21.0192\lambda + 110.95 = 0$$

which has two roots, $\lambda_1 = 10.06$ and $\lambda_2 = 10.96$. The 95% confidence interval of λ will be

$$10.06 \leq \lambda \leq 10.96.$$

(2) Equation (3.36a): the 95% confidence interval will be

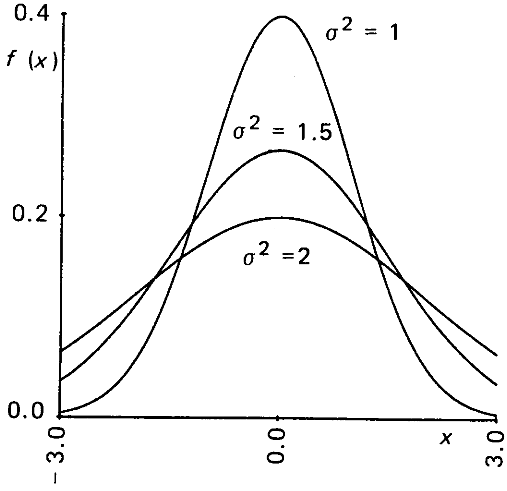


Fig. 3.6—The probability density function of the normal distribution for $\sigma^2 = 1, 1.5, 2$.

$$10.05 \leq \lambda \leq 10.95.$$

(3) Equation (3.36b): since the values of $\chi^2_{0.025}(4200)$ and $\chi^2_{0.975}$ (4600) are not available in statistical tables, we use an approximate expression due to Wilson—Wilferty that

$$\chi^2(v) \approx v(1 - (2/9v + u_{P_i}\sqrt{2/(9v)})^3$$

where u_{P_i} is the 100P% quantile of the standardized normal distribution:

For $v = 4200$, $u_{0.025} = -1.96$, we can calculate $\chi^2_{0.025}$ (4200) = 4022.26 and for $v = 4202$, $u_{0.975} = 1.96$, $\chi^2_{0.975}$ (4202) = 4383.57. The 95% confidence interval will be

$$10.06 \leq \lambda \leq 10.96.$$

Conclusion: All three equations (3.36), (3.36a) and (3.36b) yield essentially the same confidence interval for parameter λ, $10.06 \leq \lambda \leq 10.96$.

3.3.2 The normal distribution

The most important and widely used distribution in chemical practice is the normal or Gaussian distribution. Many continuous random variables encountered in practice follow, at least to a good approximation, this distribution. These include variations in measurement processes; random experimental errors occurring in experiments in physical sciences such as chemistry and physics. In the life sciences, (biology, agriculture, medicine) many directly measured experimental variables do not follow a normal distribution, but transformations can often be made to improve normality (e.g. taking logarithms).

The probability density function of a normally distributed continuous random variable x defined in an infinite interval has a rather complicated mathematical form, namely

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right] \tag{3.37}$$

The mean of variable x is $\mu = E(x)$ and the variance of variable x is $\sigma^2 = D(x)$. The graph of $f(x)$ *vs.* x forms a bell-shaped curve symmetrical about the mean ordinate $x = \mu$ (Fig. 3.6).

Suppose we have a sample $\{x_i\}$, $i = 1, \ldots, n$, with elements that are independent and come from the same normal distribution. From the logarithm of the likelihood function

$$\ln L = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{3.38}$$

we can calculate the estimate of *the sample mean*

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{3.39}$$

The second derivative of ln L with respect to parameter μ, together with Eq. (3.4), yield the variance of this sample mean

$$D(\hat{\mu}) = \sigma^2/n \tag{3.40}$$

Analogously the estimate of *the sample variance* has the form

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \tag{3.41}$$

The variance of this estimate is

$$D(\hat{\sigma}^2) = 2\sigma^4/n \tag{3.42}$$

In practice, parameter μ is unknown and is replaced by its sample estimate, $\hat{\mu} = \bar{x}$. Then the variance $\hat{\sigma}^2$ defined by Eq. (3.41) is a biased estimate since $E(\hat{\sigma}^2) = K\sigma^2$, where $K = (n-1)/n$. For an unbiased estimate of variance, we calculate instead the sample variance

$$s^2 = \frac{n}{n-1}\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \tag{3.43}$$

The confidence interval of the mean is calculated from Eqs. (3.26) and (3.27a) and the confidence interval of variance from Eq. (3.29).

If we know that a random variable x has a normal distribution and the values of μ and σ^2 are given (we write $x \approx N(\mu;\ \sigma^2)$) then we can calculate a probability $Pr(a < x < b)$ as the area bounded by the curve $y = f(x)$, the x-axis and the ordinates $x = a$ and $x = b$ for any a and b. To evaluate this area we proceed as follows. The standardized normal random variable u is defined as

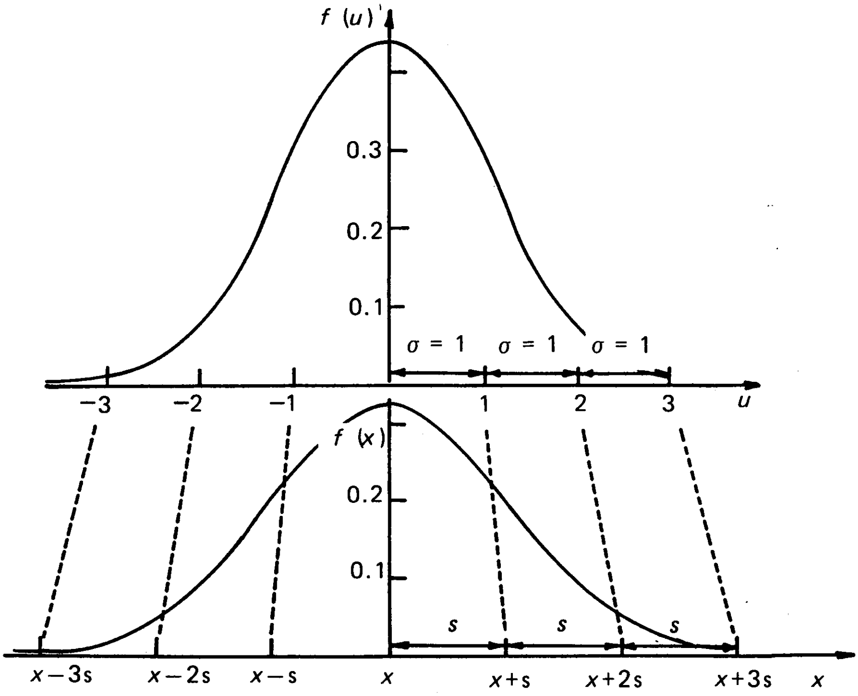


Fig. 3.7—The probability density function for the original variable $x \approx N(\mu, \sigma^2)$, and for the standardized normal variable $u \approx N(1, 0)$.

$$u = \frac{x - \mu}{\sigma}$$

Then $u \approx N(0; 1)$, i.e. u is normally distributed with mean zero and variance 1 (Fig. 3.7).

Tables of the area under the standard normal distribution (either from $-\infty$ to u or from 0 to u) may be found in standard statistical tables.

Problem 3.4 *Estimation of the mass of aspirin tablets*

A sample of $n = 156$ aspirin tablets were weighed (to the nearest mg). The declared weight of one tablet is $\mu = 330$ mg. The mean calculated from all 156 tablets was $x = 330.43$ mg and the variance $s^2 = 2.32$. Subsample A contained $n_A = 32$ tablets, with mean $x = 330.6$ mg and $s^2 = 2.135$. Subsample B contained $n_B = 10$ tablets and had mean $x_B = 330.7$ mg and $s^2 = 2.05$. Estimate the confidence interval of the mean μ and of the variance σ^2, on the assumption that the population has a normal distribution.

Data: (1) The complete sample of aspirin tablets, $n = 156$.

328.99	329.75	331.62	333.08	330.61	331.35	328.42
330.63	332.17	330.15	331.28	330.92	329.36	329.62
329.61	329.17	330.39	333.47	330.59	330.52	329.49
329.01	331.63	330.64	330.85	326.06	329.92	330.66
328.57	331.45	331.54	332.20	329.43	327.76	334.06
331.25	328.43	330.57	329.68	330.27	328.81	332.26
332.60	327.32	331.28	330.92	332.66	329.88	329.84
329.92	329.32	333.37	330.28	330.78	333.19	330.84
330.70	329.73	328.87	331.71	329.76	329.82	330.59
328.57	332.20	328.03	330.28	331.02	330.58	333.35
329.86	331.22	329.99	330.34	331.85	332.88	331.99
330.02	328.14	330.03	330.10	330.03	330.47	330.62
331.78	329.33	330.16	329.46	331.89	330.65	329.35
331.84	330.31	331.31	328.06	332.59	327.57	329.10
331.61	331.69	329.47	332.09	330.45	329.41	331.78
330.50	330.23	329.89	331.53	331.49	330.52	329.59
334.53	329.04	330.88	330.08	330.11	331.38	331.85
328.51	328.56	332.26	330.98	330.91	330.18	325.47
330.99	330.54	329.74	332.55	329.70	328.99	330.63
330.69	331.00	329.29	328.02	330.16	333.56	331.72
325.47	330.72	331.93	329.23	327.87	331.83	330.58
330.94	331.51	330.00	331.21	331.23	330.57	329.59
327.88	328.86					

(2) Subsample A, $n_A = 32$.

328.99	329.75	331.62	333.08	333.61	331.35	328.42
330.63	332.17	330.15	331.28	330.92	329.36	329.62
329.61	329.17	330.39	333.47	330.59	330.52	329.49
329.01	331.63	330.64	330.85	326.06	329.92	330.66
328.57	331.45	331.54	332.20			

(3) Subsample B, $n_B = 10$.

328.99	329.75	331.62	333.08	330.61	331.35	328.42
330.63	332.17	330.15				

Program: Chemstat: Basic Statistics: One sample analysis.
Solution: The 95% confidence interval of the mean is calculated from Eqs. (3.26) and (3.27b).

From the original sample, size $n = 156$: $330.19 \leq \mu \leq 330.67$

From the subsample, size $n_A = 32$: $329.96 \leq \mu \leq 331.09$

From the subsample, size $n_B = 10$: $329.65 \leq \mu \leq 331.70$.

The 95% confidence interval of the variance σ^2 is calculated from Eq. (3.29).

From the elements of the original sample: $1.879 \leq \sigma^2 \leq 2.937$

From the subsample, size $n_A = 32$: $1.372 \leq \sigma^2 \leq 3.43$

From the subsample, size $n_B = 10$: $0.970 \leq \sigma^2 \leq 6.83$

Conclusion: Small samples from a population with normal distribution may lead to inaccurate results. The hypothesis that $\mu = 330$ and $\sigma^2 = 1$ is accepted here at level $\alpha = 0.05$, for sample sizes bigger than $n = 100$.

3.3.3 The Laplace distribution

When random elements are measured under condition of non-constant variance, the Laplace (two-tailed exponential) distribution often occurs. The Laplace probability density function $f(x)$ of random variable x in the interval $(-\infty, \infty)$ is described by

$$f(x) = 0.5\Phi^{-1} \exp\left[-\frac{|x-\theta|}{\Phi}\right] \tag{3.44}$$

The mean of the Laplace distribution $E(x) = \theta$, the variance $D(x) = 2\Phi^2$, and the skewness and kurtosis are $g_1 = 0$ and $g_2 = 6$. The Laplace distribution has a more peaked shape than the normal distribution, with longer tails. For example, for the Laplace distribution the 1% quantile is equal to $E(x) - 2.72\sqrt{D(x)}$, but for the normal distribution it is $E(x) - 2.33\sqrt{D(x)}$. The Laplace distribution is taken as a natural "robust" alternative for the normal one.

From Eq. (3.2), the logarithm of the maximum likelihood function is

$$\ln L = -n \ln(2\Phi) - \Phi^{-1} \sum_{i=1}^{n} |x_i - \theta| \tag{3.45}$$

For the known parameter Φ the maximum likelihood estimate of θ, say $\hat{\theta}$, minimizes the sum

$$\sum_{i=1}^{n} |x_i - \theta|$$

and is equal to the sample median $\hat{\theta} = \tilde{x}_{0.5}$. Differentiation of the Eq. (3.45) with respect to Φ and equating to zero yields the maximum likelihood estimate as

$$\hat{\Phi} = \frac{1}{n} \sum_{i=1}^{n} |x_i - \hat{\theta}| \tag{3.46}$$

Although the median $\tilde{x}_{0.5}$ is unbiased, it does not have minimal variance for small samples. The confidence interval of parameter θ is derived from an asymptotic formula Eq. (3.23) which leads to Eq. (3.30). Since the variance of $\hat{\Phi}$ estimate is

$$D(\hat{\Phi}) = \Phi^2/n \tag{3.47}$$

the resulting confidence interval of parameter Φ is calculated from Eq. (3.23).

When the mean value Φ is known, the $100(1-\alpha)\%$ confidence interval of Φ is

$$\frac{2n\hat{\Phi}}{\chi^2_{1-\alpha/2}(2n)} \leq \Phi \leq \frac{2n\hat{\Phi}}{\chi^2_{\alpha/2}(2n)} \tag{3.48}$$

Problem 3.5 *Confidence interval of parameters of the Laplace distribution*
From a random sample of size $n = 50$ from the Laplace distribution $L(0, 2)$ the following estimates were calculated: $\hat{\theta} = \tilde{x}_{0.5} = 0.0119$ and $\hat{\Phi} = 1.0596$. Calculate the 95% confidence intervals for these parameters.
Data: as for Problem 2.2
Program: Chemstat: Basic Statistics: One sample analysis.
Solution: The variance estimate is $s^2 = 2\hat{\Phi}^2 = 2.246$. Substitution of this value into Eq. (3.30) yields the 95% confidence interval of parameter $\hat{\theta}$ as $-0.282 \leq \hat{\theta} \leq 0.306$.

With the use of Eq. (3.47), Eq. (3.23) can be rewritten as

$$\hat{\Phi} - u_{1-\alpha/2} \times \hat{\Phi}/\sqrt{n} \leq \Phi \leq \hat{\Phi} + u_{1-\alpha/2} \times \hat{\Phi}/\sqrt{n}$$

The 95% confidence interval of parameter can be calculated as $0.765 \leq \Phi \leq 1.353$. With the assumption that $\hat{\theta} = \tilde{x}_{0.5} = 0$, the estimate $\hat{\Phi}$ according to Eq. (3.46), is 1.0596. Then, from Eq. (3.48), the 95% confidence interval of parameter Φ is $0.818 \leq \Phi \leq 1.428$.
Conclusion: The confidence intervals of parameter Φ calculated from Eqs. (3.23) and (3.48) do not significantly differ because the mean θ is equal to zero.

3.3.4 The rectangular distribution

The simplest type of distribution is the rectangular distribution for a random variable constrained on both sides.

$$a - h \leq x \leq a + h$$

When $a = 0$ and $h = 0.5 \times 10^{-k}$, the rectangular distribution describes the distribution of errors that appeared in the rounding-off to k decimal places.

The probability density function for a rectangular distribution is

$$f(x) = 1/2\ h \tag{3.49}$$

where $a - h \leq x \leq a + h$.

The mean of the rectangular distribution $E(x) = a$, the variance $D(x) = h^2/3$, and the skewness and kurtosis are $g_1 = 0$ and $g_2 = 1.8$. The logarithm of the maximum likelihood function is

$$\ln L = -n \ln (2h) \tag{3.50}$$

for

$$a - h \leq \min (x_1, \ldots, x_n) \leq \max (x_1, \ldots, x_n) \leq a + h$$

Equation (3.50) reaches a maximum for a minimum value of h. It is evident that

$$\min (x_1, \ldots, x_n) = x_{(1)} \text{ and } \max (x_1, \ldots, x_n) = x_{(n)}$$

The maximum likelihood estimate of parameter h is given by

$$\hat{h} = 0.5\ (x_{(n)} - x_{(1)}) \tag{3.51}$$

and the maximum likelihood estimate of parameter a is

$$\hat{a} = 0.5\,(x_{(n)} + x_{(1)}) \tag{3.52}$$

The estimate $\hat{a}$ is identical to the midsum $\tilde{x}_{\text{p}}$ defined by Eq. (3.9).

The estimate $\hat{h}$ is biased. The unbiased estimate $\hat{h}_0$ is calculated by correction of the previous estimate of $\hat{h}$:

$$\hat{h}_0 = \hat{h} \times (n + 1)/(n - 1)$$

The variances of estimates $\hat{h}$ and $\hat{a}$ are calculated from

$$D(\hat{h}) = \frac{2\ h^2}{(n - 1)(n + 2)} \tag{3.53}$$

and

$$D(\hat{a}) = \frac{2\ h^2}{(n + 1)(n + 2)} \tag{3.54}$$

The variance estimates $D(\hat{h})$ and $D(\hat{a})$ are not correlated but not independent.

The confidence intervals for these parameters is calculated for large samples by Eq. (3.25).

Problem 3.6 *Examination of copper content*

For one month the concentration of copper (II) ions (μg/1) in the cooling water from an electric power station was measured. The sample of size $n = 90$ measurements contained the smallest value $x_{(1)} = 6$ μg/l, and largest $x_{(90)} = 30\mu$g/l. Exploratory data analysis indicated that the sample comes from a rectangular distribution. Calculate the parameters of location and dispersion, and the corresponding 95% confidence intervals.

Data: the concentration of Cu^{2+} in μg/l.:

18.744	22.241	10.107	7.566	26.358	6.000	15.117
24.240	18.578	12.794	26.177	10.118	21.149	21.482
19.605	20.259	9.374	29.860	29.950	11.066	17.691
20.274	24.013	11.533	6.290	17.696	28.167	25.056
14.600	25.751	15.941	8.088	9.528	19.419	7.266
11.207	15.247	24.127	26.467	22.971	12.378	27.074
6.948	15.771	26.146	10.116	13.811	13.072	26.211
21.274	8.838	28.514	29.339	27.463	10.702	11.517
26.881	17.015	15.607	26.432	25.141	21.155	16.466
17.813	9.247	15.693	28.386	28.468	9.946	6.109
25.531	27.227	28.519	22.850	10.568	7.973	19.874
13.189	12.783	23.244	28.047	20.710	30.000	29.166
18.310	14.841	24.431	19.203	21.527	11.599	

Program: Chemstat: Basic Statistics: One variable analysis.

Solution: From Eq. (3.52), the parameter $\hat{a} = 18$ and for $h = 12$, the variance of $\hat{a}$ from Eq. (3.54) is $D(\hat{a}) = 0.077$. The parameter $\hat{a}$ represents the estimate of the mean. The 95% confidence interval of parameter a from Eq. (3.28) is

$$17.45 \leq a \leq 18.55$$

Substitution in Eq. (3.51) of $\hat{h} = 12$ gives the unbiased estimate of h as $\hat{h}_0 = 12.27$. The variance estimate $\hat{\sigma}^2$ is $\hat{\sigma}^2 = 50.18$. From Eq. (3.53), $D(\hat{h}) = 0.0368$ and the 95% confidence interval of parameter h is

$$11.62 \leqslant h \leqslant 12.38$$

Conclusion: The point estimates of location and dispersion are $\hat{a} = 18$ μg/l and $h = 12.27$ μg/l. The 95% confidence intervals of the parameters, a and h, are

$$17.45 \leq a \leq 18.55 \; [\mu\text{g/l}]$$

and

$$11.62 \leq h \leq 12.38 \; [\mu\text{g/l}].$$

3.3.5 The exponential distribution

The exponential distribution is constrained on one side (upper part) and concerns the time elapsed between consecutive events in a Poisson process. Examples could be listed corresponding to those given for the Poisson distribution. Often, the lifetime of a component in a piece of apparatus is assumed to have such a distribution. Another example of an exponentially distributed random variable could be the distance travelled between successive collisions in a low pressure gas.

3.3.5.1 The one-parameter exponential distribution

This distribution describes the behaviour of a continuous random variable for which the sample space is the positive half of the real line, $x \geq 0$. We say that this random variable x has the one-parameter exponential distribution of probability density function $f(x)$ described by

$$f(x) = \theta^{-1} \exp(-x/\theta) \tag{3.55}$$

The mean of this distribution is defined by $E(x) = \theta$, the variance $D(x) = \theta^2$, the skewness $g_1 = 2$ and kurtosis $g_2 = 9$. The median is $\tilde{x}_{0.5} = \theta \times \ln 2$. The logarithm of the likelihood function is

$$\ln L = -n \ln \theta - \sum_{i=1}^{n} x_i/\theta \tag{3.56}$$

From Eq. (3.3), the maximum likelihood estimate is

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{3.57}$$

and from Eq. (3.4) its variance is

$$D(\hat{\theta}) = \theta^2/n \tag{3.58}$$

Construction of the confidence interval is based on the fact that the random variable $2\,\hat{\theta}n/\theta$ has the $\chi^2(2n)$ distribution. The 95% confidence interval of parameter θ is

$$\frac{2\,n\hat{\theta}}{\chi^2_{1-\alpha/2}(2n)} \leq \theta \leq \frac{2\,n\hat{\theta}}{\chi^2_{\alpha/2}(2n)} \tag{3.59}$$

For large samples the confidence interval defined by Eq. (3.28) can be also used.

Problem 3.7 *Decomposition kinetics of DTBP*
Di-*tert*-butyl peroxide (DTBP) decomposes at 154.6°C in the gas phase by a first-order process with rate constant $k = 3.46 \times 10^{-4}\text{sec}^{-1}$. Calculate the time (called the half-life) when 50% of the molecules have decomposed.
Solution: The number of DTBP molecules remaining at time t is given by $N_t = N_0 \exp(-kt)$, where n_0 is the number of molecules present at $t = 0$. The probability that one of the original molecules will survive for this time $t \leq T \leq t + \mathrm{d}t$, is

$$P(t \leq T \leq t + \mathrm{d}t) = -\frac{\mathrm{d}N_t}{N_0} = k \exp(-kt)\, \mathrm{d}t \tag{3.60}$$

since the probability density function (3.55) of the survival time is $f(t) = k \exp(-kt)$, where in Eq. (3.55) $\theta = 1/k$, the average survival time of the DTBP molecules is $E(x) = 1/k = 10^4\text{sec}/3.46 = 2.89 \times 10^3\text{sec}$.
Conclusion: 50% of the DTBP molecules disappeared in 2.00×10^3sec, the time that corresponds to the median $\tilde{x}_{0.5}$ (usually called the half-time), the time satisfying $f(t) = 0.5$, that is $t_{0.5} = (\ln 2)/k$. Thus, fewer than half of the molecules survive for that average time.

3.3.5.2 The two-parameter exponential distribution

This distribution describes the statistical behaviour of a constrained random variable which can reach only values $x \geq \mu$. The probability density function is defined by

$$f(x) = \theta^{-1} \exp\left[\frac{\mu - x}{\theta}\right] \tag{3.61}$$

The mean of this distribution is $E(x) = \mu + \theta$. The variance, skewness and kurtosis are the same as for the one-parameter exponential distribution. The logarithm of the likelihood function is

$$\ln L = n \ln \theta - \sum_{i=1}^{n} (x_i - \mu)/\theta \tag{3.62}$$

and the maximum likelihood estimate $\hat{\mu}$ of Eq. (3.62) is then

$$\hat{\mu} = \min(x_1, \ldots, x_n) = x_{(1)} \tag{3.63}$$

For the maximum likelihood estimate of $\hat{\theta}$ on the basis of Eq. (3.3)

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \hat{\mu}) \approx (\bar{x} - x_{(1)}) \tag{3.64}$$

The estimate $\hat{\mu}$ has expectation

$$E(\hat{\mu}) = \mu + \theta/n \tag{3.65a}$$

and variance

$$D(\hat{\mu}) = \theta^2/n^2 \tag{3.65b}$$

For an estimate $\hat{\theta}$ from Eq. (3.64)

$$E(\hat{\theta}) = \theta(1 - 1/n) \tag{3.66a}$$

and

$$D(\hat{\theta}) = \theta^2(1/n + 1/n^2 + 2/n^3) \tag{3.66b}$$

The maximum likelihood estimates $\hat{\mu}$ and $\hat{\theta}$ are clearly biased. Unbiased estimates $\hat{\mu}_0$ and $\hat{\theta}_0$ take the form

$$\hat{\mu}_0 = \frac{nx_{(1)} - \bar{x}}{n - 1} \tag{3.67a}$$

with variance

$$D(\hat{\mu}_0) = \frac{\theta^2}{n(n - 1)} \tag{3.67b}$$

and

$$\hat{\theta}_0 = \frac{n(\bar{x} - x_{(1)})}{n - 1} \tag{3.68a}$$

with variance

$$D(\hat{\theta}_0) = \frac{\theta^2}{n - 1} \tag{3.68b}$$

However, the estimates $\hat{\mu}_0$ and $\hat{\theta}_0$ are correlated, with correlation coefficient of $(-1/\sqrt{n})$.

The estimate of a confidence interval is based on the fact that the random variables $2n(x_{(1)} - \mu)/\theta$ and $2(n - 1)\hat{\theta}_0/\theta$ are independent and have the χ^2-distribution with 2 and $2(n - 1)$ degrees of freedom. Therefore, the $100(1 - \alpha)\%$ confidence interval of parameter θ is calculated from

$$\frac{2(n - 1)\hat{\theta}_0}{\chi^2_{1-\alpha/2}(2n - 2)} \leq \theta \leq \frac{2(n - 1)\hat{\theta}_0}{\chi^2_{\alpha/2}(2n - 2)} \tag{3.69}$$

Since the ratio $n(x_{(1)} - \mu)/\hat{\theta}_0$ has Fisher–Snedecor F distribution with 2 and $2n - 2$ degrees of freedom, the lower limit μ_1 of the $100(1 - \alpha)\%$ confidence interval of parameter μ is given by

$$\mu_1 = x_{(1)} - \hat{\theta}_0\, F_{1-\alpha}(2, 2n - 1)/n \tag{3.70}$$

The upper limit μ_2 is equal to the smallest sample element, $x_{(1)}$. For determination of quantiles of the F-distribution, the following approximate expression may be used

$$F_P(2, 2n - 2) = (n - 1)[(1 - P)^{-1/(n-1)} - 1] \tag{3.71}$$

Problem 3.8 *Examination of data for biologically cleaned flowing sink-water*
For 4 months, BSK_5 values in g/m^3 were recorded for biologically cleaned outflowing sink-water. In the sample of size $n = 125$, the smallest value was $x_{(1)} = 9$ g/m^3 and the arithmetic mean was $\bar{x} = 27$ g/m^3. Exploratory data analysis proved that the distribution was exponential. Estimate values for the parameters of location and dispersion with their 95% confidence intervals.

Data:

20.948	26.818	11.962	10.065	38.752	9.338	16.541
31.521	20.717	14.252	37.989	11.970	24.741	25.346
22.203	23.229	11.391	90.195	106.359	12.741	19.536
23.253	30.910	13.135	9.192	19.543	49.589	33.930
16.002	36.321	17.439	10.436	11.509	21.924	9.855
12.859	16.678	31.214	39.235	28.378	13.875	42.205
9.636	17.249	37.861	11.969	15.213	14.508	38.130
24.965	10.986	52.896	65.683	44.458	12.441	13.122
41.200	18.693	17.068	39.077	34.205	24.752	18.040
19.694	11.294	17.163	51.594	52.413	11.834	9.072
35.525	43.054	52.956	28.108	12.332	10.354	22.617
14.618	14.241	29.003	48.586	23.978	100.200	62.008
20.351	16.250	32.053	21.604	25.430	13.192	20.004
43.407	39.788	61.421	10.680	40.342	10.595	26.927
21.804	11.014	18.395	46.739	31.923	20.167	27.753
13.546	46.334	22.184	24.920	11.151	37.217	19.819
9.267	14.745	51.224	16.619	35.590	51.926	14.082
17.200	12.859	40.870	48.021	11.344	9.000	

Program: Chemstat: Basic Statistics: One variable analysis.
Solution: From Eq. (3.63), $\hat{\mu} = 9$, and then from Eq. (3.64), $\hat{\theta} = 18$. Equation (3.67) gives the unbiased estimate $\hat{\mu}_0 = 8.854$ and $D(\hat{\mu}_0) = 0.0212$. Analogously, from Eq. (3.68) the unbiased estimate $\hat{\theta}_0 = 18.15$ and $D(\hat{\theta}_0) = 2.655$. The 95% confidence interval of parameter θ is $15.31 \leq \theta \leq 21.64$. Because for $P = 0.95$, $F_{0.95}(2{,}248) = 3.03$, the lower limit of the 95% confidence interval of parameter μ [Eq. (3.70)] is $\mu_1 = 8.56$.
Conclusion: The confidence interval of parameter θ is broad compared with that of parameter μ.

3.3.6 The log–normal distribution

This distribution is one of the commonest distributions related to the normal one. For many types of physical measurements and other types of chemical data, the measured values are either positive only (pressure, volume, concentration, weight, absorbance, etc.) or have a defined origin (e.g., absolute zero for temperature). When the measured values are far from an origin, the normal distribution is found to be appropriate, but when measured values are near an origin the approximation by the normal distribution is not convenient, and the log–normal or other distribution should be used. This distribution may be found in the analysis of samples containing low and very low concentrations, i.e. in *trace analysis*. The log–normal distribution is also applicable to the distribution of powder particles in the atmosphere or the size distribution of powder pigments. A total error that is a product of partial small errors belongs to a log–normal distribution.

The log–normal distribution is derived from the logarithmic transformation of the normal distribution. The random variable x of the log–normal distribution is related to the random variable u of the standardized normal distribution by

$$u = [\ln (x - \theta) - \mu]/\sigma \tag{3.72}$$

where μ, σ and θ are parameters. The probability density function of the log–normal distribution is defined by

$$f(x) = \frac{1}{(x - \theta)\sigma\sqrt{2\pi}} \exp\left[- \frac{\ln (x - \theta) - \mu)^2}{2\ \sigma^2} \right] \tag{3.73}$$

3.3.6.1 The two-parameter log–normal distribution

This distribution concerns the positive random variable defined in the range $0 \leq x \leq \infty$. The probability density function is defined by Eq. (3.73) when $\theta = 0$. The random variable x has a two-parameter log–normal distribution if the random variable $\ln x$ has a normal distribution $N(\mu, \sigma^2)$. The mean and variance of the random variable x are calculated from

$$E(x) = \exp(\mu + 0.5\sigma^2) \tag{3.74}$$

and

$$D(x) = \exp(2\mu)\omega(\omega - 1) \tag{3.75}$$

where $\omega = \exp \sigma^2$. The skewness g_1 and kurtosis g_2 depend only on the variable ω according to

$$g_1 = \sqrt{\omega - 1}(\omega + 2) \tag{3.76a}$$

and

$$g_2 = \omega^4 + 2\omega^3 + 3\omega^2 - 3 \tag{3.76b}$$

The coefficient of variation δ defined by Eq. (3.17) is a function of parameter ω only for the log–normal distribution.

$$\delta = \sqrt{\omega - 1} \tag{3.77}$$

The mode $\hat{x}_{\mathrm{M}}$ and median $\tilde{x}_{0.5}$ are given by

$$\tilde{x}_{\mathrm{M}} = \exp(\mu - \sigma^2) \tag{3.78a}$$

and

$$\tilde{x}_{0.5} = \exp(\mu) \tag{3.78b}$$

The logarithm of the likelihood function of the two-parameter log–normal distribution is

$$\ln L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \sum_{i=1}^{n}\ln x_i - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\ln x_i - \mu)^2 \tag{3.79}$$

From Eq. (3.3), the maximum likelihood estimate of parameter μ is

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}\ln x_i \tag{3.80}$$

and of parameter σ^2

$$\hat{\sigma}^2 = \frac{1}{2}\sum_{i=1}^{n}(\ln x_i - \hat{\mu})^2 \tag{3.81a}$$

which is biased. The unbiased estimate $\hat{\sigma}_0^2$ is calculated from

$$\hat{\sigma}_0^2 = \hat{\sigma}^2\, n/(n-1) \tag{3.81b}$$

The variance of parameter $\hat{\mu}$ is calculated from Eq. (3.40) and of variance $\hat{\sigma}^2$ from Eq. (3.42). The confidence interval of parameters μ and σ^2 is found in the same as for a normal distribution (Section 3.3.2).

There are some cases when the investigation of data in logarithmic transformation is not convenient, and parameters of location and spread with their confidence intervals should be calculated in the original data scale. The $100(1-\alpha)\%$ confidence interval of median $\tilde{x}_{0.5}$ is then calculated from

$$\exp(\hat{\mu} - t_{1-\alpha/2}(n-1)\hat{\sigma}/\sqrt{n}) \leq \tilde{x}_{0.5} \leq \exp(\hat{\mu} + t_{1-\alpha/2}(n-1)\hat{\sigma}/\sqrt{n}) \tag{3.82}$$

The confidence interval for the coefficient of the variation, the skewness or kurtosis may be calculated analogously as they are functions only of parameter σ^2. The $100(1-\alpha)\%$ two-tailed confidence interval of the coefficient of variation is estimated by

$$\left[\exp\frac{(n-1)\,\hat{\sigma}^2}{\chi^2_{1-\alpha/2}(n-1)} - 1\right]^{1/2} \leq \delta \leq \left[\exp\frac{(n-1)\,\hat{\sigma}^2}{\chi^2_{\alpha/2}(n-1)} - 1\right]^{1/2} \tag{3.83}$$

The point estimate of the mean $M = E(x)$ and the corresponding estimate of the variance $V = D(x)$ for the original data scale is given by

$$\hat{M} = \exp(\hat{\mu})\, g(0.5\hat{\sigma}^2) \tag{3.84}$$

and

$$\hat{V} = \exp(2\hat{\mu})\left[g(2\hat{\sigma}^2) - g\left\{\frac{(n-2)\,\hat{\sigma}^2}{n-1}\right\}\right] \tag{3.85}$$

In both expressions the function $g(t)$ is found from an infinite series

$$g(t) = 1 + \frac{n-1}{n}t + \sum_{j=2}^{\infty}\frac{(n-1)^{2j-1}}{n^j(n+1)(n+3)\ldots(n+2j-1)} \times \frac{t^j}{j!} \tag{3.86}$$

For large samples with $n > 50$, or for sufficiently small values of the variance σ^2, the following approximation [3] may be used

$$g(t) \approx \exp(t)\left[1 - \frac{t(t+1)}{n} + \frac{t^2(3t^2 + 22t + 21)}{6n^2}\right] \tag{3.87}$$

The variances of the $\hat{M}$ and $\hat{V}$estimates are calculated from the expression [3]

$$D(\hat{M}) = \sigma^2 \exp(2\mu)\ \omega(1 + 0.5\sigma^2)/n \tag{3.88}$$

and

$$D(\hat{V}) = 2\sigma^2 \exp(\mu)\ \omega^2[2(\omega - 1)^2 + \sigma^2(2\omega - 1)^2] \tag{3.89}$$

For large samples, the approximate confidence intervals of $\hat{M}$ and $\hat{V}$ can be calculated from Eqs. (3.23) and (3.28). The confidence interval of the mean M may be calculated with the use of an estimate $\hat{\tau}$ and its variance $D(\hat{\tau})$

$$\hat{\tau} = \hat{\mu} + n\hat{\sigma}^2/(2(n-1)) \tag{3.90a}$$

and

$$D(\hat{\tau}) = \sigma^2/(n-1) + n^2\sigma^4/(4(n-1)^3) \tag{3.90b}$$

By using the estimate $\hat{\tau}$ and $D(\hat{\tau})$ the *100(1 − α)% confidence interval of the mean* M may be expressed as

$$\exp[\hat{\tau} - \mu_{1-\alpha/2}\sqrt{D(\hat{\tau})}] \le M \le \exp[\hat{\tau} + \mu_{1-\alpha/2}\sqrt{D(\hat{\tau})}] \tag{3.91}$$

In the case of the log–normal distribution, the data should be analysed in logarithmic transformation, or the estimates $\hat{M}$ and $\hat{V}$should be used, except in cases when $\hat{\sigma}^2 \ll 1$.

Problem 3.9 *Examination of trace concentrations of copper in kaolin*
A set of $n = 32$ samples of kaolin was used to determine the trace concentration of copper in raw kaolin. Exploratory data analysis indicated that the sample came from a two-parameter log–normal distribution. By analysis of logarithms of the measured quantities, the arithmetic mean $\hat{\mu} = 23$ and the sample variance $\sigma^2 = 0.0004$ were estimated. Calculate point and interval estimates of parameters of location and of dispersion by analysing the original measured quantities.
Data

9.467	9.785	9.806	9.863	9.867	9.889	9.915
9.933	9.951	9.950	9.969	9.994	10.025	10.059
10.078	10.088	10.093	10.091	10.094	10.097	10.124
10.134	10.185	10.195	10.209	10.222	10.234	10.235
10.317	10.446	10.313	10.502			

Progam: Chemstat: Basic Statistics: One variable analysis.
Solution: From Eq. (3.27b) the 95% confidence interval of parameter μ is calculated to be $2.297 \le \mu \le 2.307$. The confidence limits are the arguments of the exponential in Eq. (3.82). The 95% confidence interval of the median $\tilde{x}_{0.5}$ of the original data is $10.012 \le \tilde{x}_{0.5} \le 10.167$.

The 95% confidence interval of the coefficient of variation δ (Eq. (3.83)) is $0.0160 \le \delta \le 0.0266$. Equation (3.87) gives the values $g(0.5\hat{\sigma}^2) = 1.0002$, Introduction of the maximum likelihood estimates $\hat{\mu}(\theta)$ and $\hat{\sigma}^2(\theta)$ into the logarithm of the likelihood function yields a modified maximum likelihood function:

$$\ln L_1 = -n[\hat{\mu}(\theta) + 0.5 \ln \hat{\sigma}(\theta)] \tag{3.92a}$$

This function depends on just a single parameter, θ. A graph of the function $\ln L_1 = f(\theta)$ is shown in Fig. 3.8. The first step is a search over the interval in which the function $\ln L_1$ is unimodal. The second stage is to search iteratively in this interval.

Problem 3.10 *Examination of trace concentration data on antimony in a copper ore* The set of $n = 40$ samples of copper ore was analysed to determine the trace concentration of antimony (in ppm). Exploratory analysis indicated that the antimony concentration is described by a log–normal distribution. Estimate the point parameters of location and dispersion.

Data:

145.4	151.6	83.7	109.6	124.1	124.4	128.1	139.7
140.3	145.4	156.0	160.3	160.6	163.1	165.6	172.5
182.0	187.7	193.1	199.6	201.9	203.1	204.2	205.0
205.4	206.7	217.0	220.6	240.6	242.4	247.2	253.8
259.6	265.8	266.1	308.8	311.7	401.3	449.3	536.3

Program: Chemstat: Probability models.
Solution: On maximizing Eq. (3.92a), with the use of Eqs. (3.80) and (3.81) with x_i replaced by $(x_i - \theta)$, the maximum likelihood estimates $\hat{\mu} = 5.065$, $\hat{\sigma} = 0.478$ and $\theta = 36.5$ are calculated. The maximum of the $\ln L$ function reaches the value $\ln L = -258.5$. Figure 3.9 shows a graph of the log–normal distribution with its simple nonparametric estimate.

By taking the minimum value $\theta = 0$, the estimates $\hat{\mu}_1 = 5.29$, $\hat{\sigma}_1 = 0.384$ and $\ln L = -258.8$ are calculated.
Conclusion: The difference between the estimates results from the choice $\theta = 0$; the estimates calculated from original data are not influenced. For the three-parameter model the median is,

$$\tilde{x}_{0.5} = \hat{\theta} + \exp(\mu) = 194.88$$

whereas for the two-parameter model

$$\tilde{x}_{0.5} = \exp(\hat{\mu}_1) = 198.34.$$

3.4 ROBUST ESTIMATES FOR PARAMETERS OF LOCATION AND SPREAD

The sample mean $\bar{x}$ and the sample variance s^2 are efficient estimates of the parameters of location and spread only for data from the normal distribution. If the sample is not normally distributed, or if some outliers are present, the efficiency of both $\bar{x}$ and s^2 decreases. We shall introduce many statistical techniques based on normal distribution of the original observations, and these still remain approximately correct

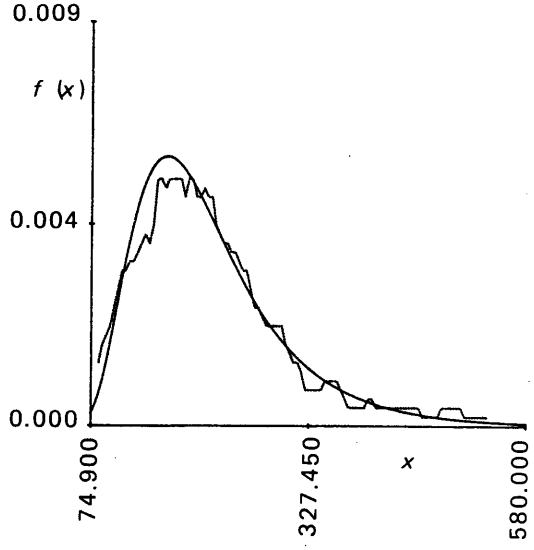


Fig. 3.9—The probability density function for the sample log–normal distribution and its nonparametric estimate distribution.

for reasonable departures from normality. In this regard they are said to be *robust* to *non-normality*. Robustness can relate to the separate effects of deviations from normality, independence, equal variance, and randomness.

3.4.1 The median

The median $\tilde{x}_{0.5}$ is the oldest robust estimate of the parameter of location. Despite other robust statistics, the median has a precise interpretation for a symmetrical and also for a non-symmetrical distribution: the median $\tilde{x}_{0.5}$ is the second quartile or 50%-percentile, which divides the probability distribution area into two equal areas. Thus, the probability of x being less than $\tilde{x}_{0.5}$ is 1/2 and equal to the probability of x being greater than $\tilde{x}_{0.5}$. When the sample size n is odd, the sample median $x_{0.5}$ is defined to be the middle value in order of size; when n is even it lies between the two middle values, and usually we take the average of these two as its value.

For an unknown distribution and when some outliers are present, the nonparametric estimate of the standard deviation of a median is calculated from

$$s_M = \frac{x_{(n-k+1)} - x_{(k)}}{2\ u_{\alpha/2}} \tag{3.93}$$

where

$$k = \frac{n+1}{2} - |u_{\alpha/2}|\ \sqrt{n/4} \tag{3.94}$$

Best results are obtained with $\alpha = 0.05$, for which $|u_{0.025}| = 1.96$.

For small samples, the Marritz–Jarret estimate of the standard deviation of median can be used.

$$s_{\mathrm{M}}^{*} = \left[\sum_{i=1}^{n} w_i x_{(i)}^2 - \left[\sum_{i=1}^{n} w_i x_{(i)}\right]^2\right]^{1/2} \tag{3.95}$$

where

$$w_i = J\left[\frac{i-0.5}{n}\right]\left[\sum_{i=1}^{n} J\left[\frac{i-0.5}{n}\right]\right]^{-1} \tag{3.95a}$$

and function $J(x)$ is defined by

$$J(x) = \frac{n!}{(m!)^2} \times x^m(1-x)^m \tag{3.95b}$$

where $m = \mathrm{int}((n-1)/2)$ and $\mathrm{int}(x)$ means the integer part of a number x.

The random variable t_{M} defined by

$$t_{\mathrm{M}} = \frac{\tilde{x}_{0.5} - Me}{s_{\mathrm{M}}} \approx \frac{\tilde{x}_{0.5} - Me}{s_{\mathrm{M}}^{*}} \tag{3.96}$$

has approximately the Student distribution with $(n-1)$ degrees of freedom. The symbol Me means the median of the population from which the sample comes.

Problem 3.11 *Robust sample estimates of five distributions*
Apply robust analysis to five samples of size $n = 50$, from the normal, rectangular, exponential, Laplace and log–normal distribution, with the use of the median measures.
Data: $n = 50$, data from Problem 2.2
Program: Chemstat: Basic Statistics: One sample analysis.
Solution: Results of the robust analysis of the five sample are shown in Table 3.4. These are the median $\tilde{x}_{0.5}$, the square of interquantile deviation R^2 (Eq. (3.15)), the standard deviations of median s_{M}^{*} and the limits of the 95% confidence interval L_1 and L_2 (Eq. (3.96)).

Comparison of the results of Table 3.4 with those of Table 3.3 shows that

(a) for the normal and rectangular distribution, the median characteristics give nearly the same results as the mean characteristics;
(b) for the exponential distribution the confidence interval is nearly the same, but the median is too robust, compared with the mean.
(c) for the Laplace distribution the median characteristics are better than the mean ones;
(d) for the log–normal distribution the median characteristics are not objective enough and the 95% confidence interval does not contain the mean value.

Table 3.4—The median characteristics of five samples with various population distributions, $n = 50$

Population distribution $X(\mu, \sigma^2)$	$\tilde{x}_{0.5}$	R^2	s_M	s_M^*	L_1	L_2
Normal $N(0; 1)$	0.0309	1.334	0.032	0.423	−0.329	0.391
Rectangular $R(0.5; 0.083)$	0.506	0.242	0.0061	0.279	0.349	0.663
Exponential $E(1; 1)$	0.705	0.923	0.024	0.423	0.395	1.02
Laplace $L(0; 2)$	0.0115	1.964	0.032	0.423	−0.347	0.371
Log–normal $LN(2.71; 47.21)$	1.293	4.109	0.190	0.660	0.415	2.169

Conclusion: The median and other robust characteristics are not suitable for asymmetrical distributions, because robust techniques "cut" good values and make the population look symmetric. Robust techniques cannot be adopted for general use.

3.4.2 The trimmed mean

One of the simplest and most efficient robust estimates of location is the *trimmed mean* $\bar{x}(\varkappa)$ defined with the use of the order statistics $x_{(i)}$ as

$$\bar{x}(\varkappa) = \frac{1}{n - 2M} \sum_{i=M+1}^{n-M} x_{(i)} \tag{3.97}$$

where $M = \text{int}\,(\varkappa \times n/100)$. The parameter $\varkappa$ determines the percentage of order statistics $x_{(i)}$ that are to be trimmed off at each (low and high) tail. The usual value of $\varkappa$ is 10%, and this results in the 10% trimmed mean, $\bar{x}(10)$. When there are many outliers, $\bar{x}(25)$ is preferred.

The trimmed mean is used with the *winsorized sum of squared differences*

$$S_w(\varkappa) = \sum_{i=M+2}^{n-M-1} (x_{(i)} - \bar{x}_w(\varkappa))^2 + (M+1)[(x_{(M+1)} - \bar{x}_w(\varkappa))^2 + (x_{(n-M)} - x_w(\varkappa))^2] \tag{3.98}$$

where $\bar{x}_w(\varkappa)$ is the *winsorized mean* defined by

$$\bar{x}_w(\varkappa) = \frac{1}{n}\left[(M+1)(x_{(M+1)} + x_{(n-M)}) + \sum_{i=M+2}^{n-M-1} x_{(i)}\right] \tag{3.99}$$

Tukey and McLaughlin [2] recommend for statistical testing about parameter of location μ the test statistics

$$t_{\mathrm{R}}(\varkappa) = \frac{(\bar{x}(\varkappa) - \mu)\sqrt{h(h-1)}}{S_{\mathrm{w}}(\varkappa)} \tag{3.100}$$

where $h = n - 2M$. The test statistic $t_{\mathrm{R}}(\varkappa)$ has approximately the Student distribution with $(n - 1)$ degrees of freedom. For small samples from $n = 5$, the recommended value for constant M is 1, i.e. $\varkappa = 20\%$. It may be concluded from Eq. (3.100) that the winsorized variance is equal to $s_{\mathrm{w}}^2(\varkappa) = S_{\mathrm{w}}(\varkappa)/(h-1)$.

For asymmetric and strongly skewed distributions the *asymmetric trimmed mean* $\bar{x}(\varkappa_1, \varkappa_2)$ is defined by

$$\bar{x}(\varkappa_1, \varkappa_2) = \frac{\sum_{i=n_1}^{n_2} x_{(i)}}{n_2 - n_1 + 1} \tag{3.101}$$

where $n_1 = \mathrm{int}(\varkappa_1 \times n/100)$ and $n_2 = n - \mathrm{int}(\varkappa_2 \times n/100)$. When $\varkappa_1$ and $\varkappa_2$ are chosen so that resulting trimmed sample has a symmetrical distribution, the *variance of asymmetrically trimmed mean* may be calculated by

$$\begin{aligned} s_{\mathrm{w}}^2(\varkappa_1, \varkappa_2) = \frac{1}{h(h+1)} \Bigg[& n_1(x_{(n_1)} - \bar{x}(\varkappa_1, \varkappa_2))^2 \\ & + \sum_{i=n_1+1}^{n_2-1} (x_{(i)} - \bar{x}(\varkappa_1, \varkappa_2))^2 + (n - n_2 + 1)(x_{(n_2)} - \bar{x}(\varkappa_1, \varkappa_2))^2 \\ & - ((n_1 - 1)(x_{(n_1)} - \bar{x}(\varkappa_1, \varkappa_2)) + (n - n_2)(x_{(n_2)} - \bar{x}(\varkappa_1, \varkappa_2))^2 \Bigg] \end{aligned} \tag{3.102}$$

where $h = n_2 - n_1 + 1$. If the resulting trimmed sample is already symmetrical, the test criterion

$$t_{\mathrm{R}}(\varkappa_1, \varkappa_2) = \frac{\bar{x}(\varkappa_1, \varkappa_2) - \mu}{s_{\mathrm{w}}(\varkappa_1, \varkappa_2)} \tag{3.103}$$

has approximately the Student distribution with $\nu = [(n_2 - n_1 + 1) - 2]$ degrees of freedom.

Various selector criteria are available for choosing the magnitude for trimming off data. These serve as the estimates of the length of tails or of skewness of the sample distribution. The Q_1-criterion for estimation of the *relative tail length of the sample distribution* is defined by

$$Q_1 = \frac{\bar{U}(0.05) - \bar{L}(0.05)}{\bar{U}(0.5) - \bar{L}(0.5)} \tag{3.104}$$

and also for estimation of the *relative skewness of the sample distribution* by

$$Q_2 = \frac{\bar{U}(0.05) - \bar{M}(0.5)}{\bar{M}(0.05) - \bar{L}(0.05)} \tag{3.105}$$

where $\bar{U}(\beta)$ is the average of the $n\beta$ largest ordered values, $\bar{L}(\beta)$ is the average of the $n\beta$ smallest ordered values, and $\bar{M}\ (\beta)$ is the average of ordered values from the middle part of the distribution. When $n\beta$ is not an integer, it is found as a ratio of neighbouring ordered values by linear interpolation.

According to the values of Q_1 and Q_2, the parameter $\varkappa$ of the trimmed mean $\bar{x}(\varkappa)$ is chosen as follows.

(a) When $Q_2 \approx 0$ and the sample distribution is symmetric, $\varkappa$ is selected according to sample size (Table 3.5).

(b) When the sample distribution is symmetric and 90% efficiency is desired, parameter $\varkappa$ is chosen such that $\varkappa = 15\%$ for $Q_1 < 2.9$, $\varkappa = 25\%$ for $2.9 \leq Q_1 \leq 3.5$ and $\varkappa = 35\%$ for $Q_1 \geq 3.5$.

Table 3.5—The choice of best trimming parameter, κ in %. Q_1^* is calculated from Eq. (3.104), with difference $[\bar{U}\ (0.2) - \bar{L}(0.2)]$ in the numerator

n	$\kappa(\%)$	Q_1^*
< 10	6.35	for all Q^*
	12.5	$Q_1^* < 1.84$
$10-20$	25	$Q_1^* > 1.84$
	9.375	$Q_1^* < 1.81$
$20-30$	18.75	$1.81 < Q_1^* < 1.87$
	28.125	$Q_1^* > 1.87$

(c) When $Q_2 \neq 0$, the asymmetrically trimmed average may be used, and parameters $\varkappa_1$ and $\varkappa_2$ are selected as follows:

$$\text{for } Q_2 \leq 1.4 \text{ and } Q_1 > 2.68,\ \varkappa_1 = 25\% \text{ and } \varkappa_2 = 25\%,$$

$$\text{for } Q_2 > 1.4 \text{ and } Q_1 < 1.98,\ \varkappa_1 = 0\% \text{ and } \varkappa_2 = 50\%,$$

$$\text{for } Q_2 > 1.4 \text{ and } Q_1 > 2.68,\ \varkappa_1 = 25\% \text{ and } \varkappa_2 = 25\%.$$

More detailed information is available [1].

Problem 3.12 *Trimmed mean and winsorized variance of samples from five distributions* Apply robust analysis to five samples of size $n = 50$ from the normal, rectangular, exponential, Laplace and log-normal distributions, with use of the trimmed mean and winsorized variance.

Data: from Problem 2.2
Program: Chemstat: Basic Statistics: One sample analysis.
Solution: Table 3.6 lists the trimmed mean $\bar{x}(\varkappa)$ and the winsorized variance $s_{\mathrm{w}}^2(\varkappa)$ for two values of $\varkappa$, $\varkappa = 0.05$ and $\varkappa = 0.10$.

Table 3.6—The trimmed mean $\bar{x}(\varkappa)$ and the winsorized variance $s_{\mathrm{w}}^2(\varkappa)$ for (*a*) $\varkappa = 0.05$, and (*b*) $\varkappa = 0.10$, and the limits L_1 and L_2 of the confidence interval of the mean $\bar{x}_{\mathrm{w}}$

(*a*)

Population distribution	$\varkappa = 0.05$			
$X(\mu, \sigma^2)$	$\bar{x}(\varkappa)$	$s_{\mathrm{w}}^2(\varkappa)$	L_1	L_2
Normal $N(0; 1)$	−0.046	0.825	−0.325	0.232
Rectangular (0.5; 0.083)	0.486	0.088	0.395	0.578
Exponential $E(1; 1)$	0.836	0.519	0.615	1.06
Laplace $L(0; 2)$	−0.088	1.271	−0.434	0.259
Log-normal LN (2.71; 47.21)	2.551	22.65	1.087	4.02

(b)

Population distribution	$\varkappa = 0.10$			
$X(\mu, \sigma^2)$	$\bar{x}(\varkappa)$	$s_{\mathrm{w}}^2(\varkappa)$	L_1	L_2
Normal $N(0; 1)$	−0.053	0.787	−0.340	0.234
Rectangular (0.5; 0.083)	0.489	0.092	0.391	0.587
Exponential $E(1; 1)$	0.805	0.512	0.573	1.036
Laplace $L(0; 2)$	−0.049	1.119	−0.392	0.293
Log-normal LN (2.71; 47.21)	1.796	7.30	0.921	2.671

Table 3.7—The selector statistics Q_1, Q_1^* and Q_2 for five samples from various distributions with $n = 50$

Population distribution $X(\mu, \sigma^2)$	Q_1	Q_1^*	Q_2
Normal $N(0; 1)$	2.856	1.795	0.919
Rectangular $R(0.5; 0.083)$	1.922	1.571	0.054
Exponential $E(1; 1)$	3.576	1.784	6.175
Laplace $L(0; 2)$	3.66	1.885	1.358
Log–normal LN (2.71; 47.21)	5.02	2.064	25.182

When these results are compared with those of Table 3.3, it is evident that (1) for the symmetric distributions N, R and L, the trimming leads to shortening of the confidence interval,

(2) for the asymmetric distribution E and LN the trimming for $\varkappa = 0.10$ causes negative results, and for the log–normal distribution the 95% confidence interval does not contain a true value. A small amount of trimming $\varkappa = 0.05$ always leads to confidence intervals which are narrow and contain the true value μ. Table 3.7 lists some selector statistics Q_1 and Q_2 which indicate the differences in skewness and kurtosis of the samples. Use of selector statistics in the symmetric distributions does not cause significant improvement.

Conclusion: Trimmed means in connection with selector statistics enable determination of robust estimates even for asymmetric distributions.

3.4.3 The robust *M*-estimates

The robust M-estimates represent the maximum likelihood estimates of parameters for some special distributions. Maximization of the likelihood function according to the parameter μ_M leads here to a minimization of the function

$$\sum_{i=1}^{n} \left[\frac{x_i - \mu_M}{\sigma} \right] = \min \qquad (3.106)$$

The shape of the function $\rho(u)$ determines the property of the estimate. Among M-estimates are the arithmetic mean and median.

The *M-estimate of location parameter* μ_M is generally defined by

$$\hat{\mu}_M = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \qquad (3.107)$$

where $w_i = W[(x_i - \mu_M)/\sigma]$ and $W(u) = \mathrm{d}\,\rho(u)/dt$. For a robust estimate the function $W(u)$ must be bounded. The bi-quadratic function $W(u)$ of the following type is recommended

$$W(u)\begin{cases} \left[1 - \left[\dfrac{u}{4.69}\right]^2\right]^2 & \text{for } |u| < 4.69 \\ 0 & \text{for } |u| \geqslant 4.69 \end{cases} \tag{3.108}$$

where the numerical constant 4.69 means that for normally distributed data the asymptotic efficiency of estimate $\hat{\mu}_M$ is equal to 0.95. Since the standard deviation σ is usually unknown it is replaced by a suitable robust estimate. Du Mond and Lenth [5] recommend for the *M-estimate of standard deviation* the expression

$$s_M = - \left[\frac{\sum\limits_{i=1}^{n} V_i (x_i - \hat{\mu}_M)^2}{\sum\limits_{i=1}^{n} V_i}\right]^{1/2} \tag{3.109}$$

where

$$V_i = W\left[\left[\Delta\left[\frac{x_i - \hat{\mu}_M}{s_M}\right]\right]^{1/2}\right] \tag{3.110}$$

The weight function $W()$ is defined by Eq. (3.108) and $\Delta(u)$ is a deviation function, for which:

$$\Delta(u) = \begin{cases} u^2 - \ln\,(u^2) - 1 & \text{for } u \neq 0 \\ \infty & \text{for } u = 0 \end{cases} \tag{3.111}$$

Du Mond and Lenth [5] ***procedure for $\hat{\mu}_M$ and $\hat{s}_M$ estimation***

(1) For initial guesses $\mu_M^{(0)}$ and $s_M^{(0)}$ the median $\tilde{x}_{0.5}$ and the interquantile range $\hat{\sigma}_R = 0.75\,(\tilde{x}_{0.75} - \tilde{x}_{0.25})$ are computed.

(2) For initial guess $\hat{\mu}_M^{(0)}$ and $s_M^{(0)}$ the weights w_i and V_i are calculated from Eqs. (3.107) and (3.109), then the refined estimates $\hat{\mu}_M^{(1)}$ and $s_M^{(1)}$ are obtained.

(3) In the second iteration new values of weights w_i and V_i are calculated and hence new refined estimates $\hat{\mu}_M^{(2)}$ and $s_M^{(2)}$.

(4) Iteration refinement terminates when the estimates from two iterations do not differ significantly.

Because the robust M-estimate $\hat{\mu}_M$ represents the weighted arithmetic mean, its variance is expressed by:

$$D(\hat{\mu}_M) = s_M^2 / \sum_{i=1}^{n} w_i \tag{3.112}$$

In constructing confidence intervals and statistical testing the random variable

$$t_M = \frac{(\mu_M - \hat{\mu}_M)\left[\sum_{i=1}^{n} w_i\right]^{1/2}}{s_M} \tag{3.113}$$

which has approximately the Student t-distribution with $\nu = n - 1$ degrees of freedom can be used.

Two weight functions are compared in Fig. 3.10.

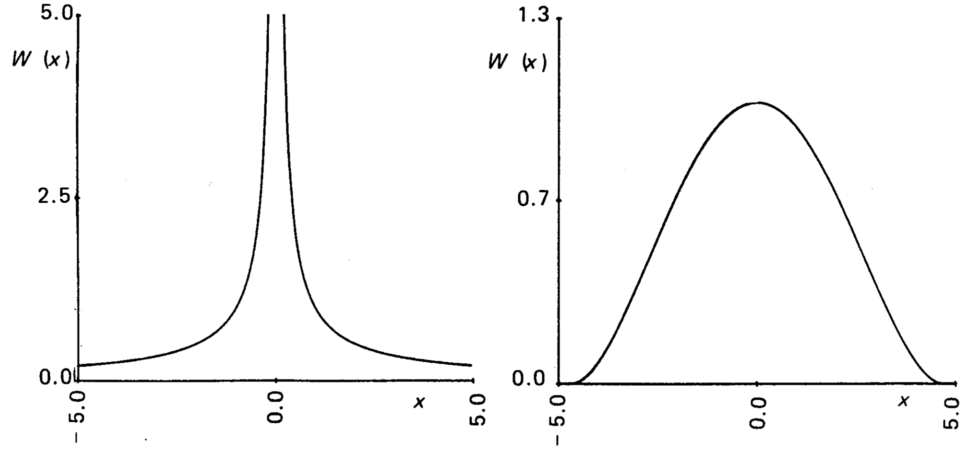


Fig. 3.10—Comparison of two weight functions of M-estimates: (a) for the median, and (b) for the bi-quadratic function.

Problem 3.13 *Robust bi-quadratic sample estimates from five distributions*
Apply robust analysis to five samples of size $n = 50$ from normal, rectangular, exponential, Laplace and log–normal distributions with the use of bi-quadratic estimates. %
Data: from Problem 2.2

Program: Chemstat: Basic Statistics: One sample analysis.

Solution: Robust estimates $\hat{\mu}_M$, variances $D(\hat{\mu})$ and the limits of the 95% confidence interval of the mean are listed in Table 3.8. For the symmetric distributions N, R and L the robust analysis gives accurate estimates quite near to the true values, and the confidence interval is narrow.

Worse results were achieved with the asymmetric skewed distributions: for the exponential and log–normal distributions the 95% confidence interval does not contain theoretical value μ.

Table 3.8—Robust analysis of samples from five distributions with the use of bi-quadratic estimates

Population distribution $X(\mu, \sigma^2)$	$\hat{\mu}_M$	$D(\hat{\mu}_M)$	L_1	L_2
Normal $N(0; 1)$	−0.0458	1.039	−0.349	0.257
Rectangular $R(0.5; 0.0833)$	0.488	0.089	0.399	0.577
Exponential $E(1; 1)$	0.762	0.442	0.561	0.964
Laplace $L(0; 2)$	−0.124	1.464	−0.490	0.242
Log–normal $LN(2.71; 47.21)$	1.375	2.378	0.893	1.858

Conclusion: The robust M-estimates of this type are not suitable for analysis of skewed distributions.

Problem 3.14 *Comparison of various estimates of location and spread for data with outliers*

The advantage of robust estimates is their lower sensitivity to outliers in data. A sample of size $n = 50$ was generated in which 47 observations came from the normal distribution $N(0, 1)$ with $\mu = 0$ and $\sigma^2 = 1$, and the remaining three from the normal distribution $N(A, 1)$ with the population mean $\mu = A$ and variance $\sigma^2 = 1$. For three different values of A, 100, 50 and 10, three samples were synthesized. Calculate the sample mean $\bar{x}$, variance s^2, median $\tilde{x}_{0.5}$, squared interquantile range R^2, trimmed mean $\bar{x}(10)$, winsorized variance $s_w^2(10)$, bi-quadratic estimate of a mean $\hat{\mu}_M$ with its corresponding variance, and comment on each estimate.

Data:

(1) $n = 50$, 47 elements from $N(0, 1)$ and 3 elements from $N(100, 1)$:

−1.008 −0.500 0.749 1.723 0.076 0.569 −1.389
0.087 1.112 −0.235 0.519 0.279 −0.758 −0.588
−0.594 −0.885 −0.072 1.980 0.063 0.016 −0.673
−0.993 0.752 0.092 0.236 −2.962 −0.383 0.109
−1.285 0.634 0.690 1.134 −0.711 −1.825 2.374
0.500 −1.380 0.046 −0.544 −0.150 −1.129 1.173
1.401 −2.121 0.521 0.280 1.440 99.585 99.557
99.616

(2) $n = 50$, 47 elements from $N(0, 1)$ and 3 elements from $N(50, 1)$: replace last three points with 49.585 49.557 49.616

(3) $n = 50$, 47 elements from $N(0, 1)$ and 3 elements from $N(10, 1)$: replace last three points with 9.585 9.557 9.616

Program: Chemstat: Basic Statistics: One sample analysis.

Solution: The parameters of location and spread are listed in Table 3.9. Whereas $\bar{x}$ and s^2 are quite far from the true values, all the robust methods leads to estimates that are not influenced by the outliers.

Table 3.9—Parameter estimates of location and spread for three samples containing outliers.

Magnitude of outlier	$\bar{x}$	s^2	$\tilde{x}_{0.5}$	R^2	$\bar{x}(10)$	$s_w^2(10)$	$\hat{\mu}_M$	s_M^2
100	5.94	572.2	0.082	1.56	0.092	1.14	−0.020	1.152
50	2.94	142.8	0.082	1.56	0.092	1.14	−0.018	1.136
10	0.54	6.4	0.082	1.56	0.092	1.18	−0.017	1.120

Conclusion: The main advantage of robust estimates is the low sensitivity to outliers. Robust statistics are convenient for statistical data analysis when the data are not homogeneous.

3.4.4 The analysis of small samples

The analysis of small samples is not reliable and results are usually rather uncertain. Small samples are used in cases when experiment repetition is expensive or scarcely possible.

For $n = 2$, statistical analysis is very difficult. If observations are close enough, the arithmetic mean is calculated. If observations do not agree, it is not possible to say which is the outlier. The $100(1 - \alpha)\%$ confidence interval of the mean μ may be calculated by an approximation

$$\frac{x_1 + x_2}{2} - T_\alpha \times \frac{|x_1 - x_2|}{2} \leq \mu \leq \frac{(x_1 + x_2)}{2} + T_\alpha \times \frac{|x_1 - x_2|}{2}$$

The critical value of T_α depends on the distribution of the data population that the two values come from. For the normal distribution it is $T_\alpha = \cot(\pi\alpha/2)$, and for $\alpha = 0.05$, T_α is 12.71. For the rectangular distribution $T_\alpha = 1/\alpha - 1$, i.e. $T_{0.05} = 19$ [6].

For $n = 3$ it is also difficult to use statistical analysis. The calculation of the arithmetic mean $\bar{x}$ from two near observations is better than the use of the median from all three values. The $100(1 - \alpha)\%$ confidence interval of the mean μ is then calculated by an approximation

$$\bar{x} - s \times T'_\alpha/\sqrt{3} \leq \mu \leqslant \bar{x} + s \times T'_\alpha/\sqrt{3}.$$

For the normal distribution, $T'_\alpha \approx 1/\sqrt{\alpha} - 3\sqrt{\alpha}/4 + \ldots$, and when $\alpha = 0.05$, T'_α is 4.30. For the rectangular distribution $T'_{0.05} = 5.74$ [6].

For $4 \leq n \leq 20$ a procedure based on order statistics was introduced by Horn [7]. This is based on the depths which correspond to the sample quartiles (the letter F), cf. Section 2.2. The pivot depth is expressed by $H_L = \mathrm{int}[(n + 1)/2]/2$ or

$H_L = \text{int}[(n+1)/2+1]/2$ according to which of the H_L is an integer. The lower pivot is $x_L = x_{(H)}$ and the upper one is $x_U = x_{(n+1-H)}$. The estimate of the parameter of location is then expressed by the *pivot halfsum*

$$P_L = 0.5(x_L + x_U) \tag{3.114}$$

and the estimate of the parameter of spread is expressed by the *pivot range*

$$R_L = x_U - x_L \tag{3.115}$$

The random variable

$$T_L = \frac{P_L}{R_L} = \frac{x_L + x_U}{2(x_U - x_L)} \tag{3.116}$$

has approximately a symmetric distribution and its quantiles are given in Table 3.10.

Table 3.10—The quantile $t_{L,1-\alpha}(n)$ of the T_L-distribution

$1-\alpha$ n	0.9	0.95	0.975	0.99	0.995
4	0.477	0.555	0.738	1.040	1.331
5	0.869	1.370	2.094	3.715	5.805
6	0.531	0.759	1.035	1.505	1.968
7	0.451	0.550	0.720	0.978	1.211
8	0.393	0.469	0.564	0.741	0.890
9	0.484	0.688	0.915	1.265	1.575
10	0.400	0.523	0.668	0.878	1.051
11	0.363	0.452	0.545	0.714	0.859
12	0.344	0.423	0.483	0.593	0.697
13	0.389	0.497	0.608	0.792	0.945
14	0.348	0.437	0.525	0.661	0.776
15	0.318	0.399	0.466	0.586	0.685
16	0.299	0.374	0.435	0.507	0.591
17	0.331	0.421	0.502	0.637	0.774
18	0.300	0.380	0.451	0.555	0.650
19	0.288	0.361	0.423	0.502	0.575
20	0.266	0.337	0.397	0.464	0.519

The 95% confidence interval of the mean is expressed by pivot statistics as

$$P_L - R_L \times t_{L,0.975}(n) \le \mu \le P_L + R_L \times t_{L,0.975}(n) \tag{3.117}$$

and analogously hypothesis testing also may be carried out. For small samples $(4 \le n \le 20)$, the pivot statistics lead to more reliable results than the application of Student's t-test or robust t-tests.

Problem 3.15 *Analysis of a small sample ($n = 5$) taken from five distributions*
Make an analysis of a small sample ($n = 5$) taken from normal, rectangular, exponential, Laplace and log–normal distributions. Try to use Horn's procedure [7].

Data: $n = 5$.

$N(0, 1)$	-1.008	-0.500	0.749	1.723	0.076
$R(0.5, 0.083)$	0.531	0.677	0.171	0.065	0.848
$E(1, 1)$	0.757	1.129	0.188	0.067	1.885
$L(0, 2)$	0.064	0.436	-1.072	-2.036	1.192
$LN(2.71, 47.21)$	0.191	2.118	0.380	0.264	3.374

Program: Chemstat: Basic Statistics: One sample analysis.
Solution: The classical and pivot approaches are compared. In the first half of Table 3.11 the classical statistical measures ($\bar{x}$, s^2 and limits L_1, L_2 of the 95% confidence interval of the mean) were calculated, and the second half contains the pivot statistics (P_L, R_L and limits L_1, L_2 of the 95% confidence interval of the mean).
Conclusion: For a small sample size, the pivot approach seems to lead to more reliable parameter estimates.

Table 3.11—Horn's procedure for analysis of a small sample ($n = 5$), in comparison with the classical statistical approach

Population distribution $X(\mu, \delta)^2$	Classical approach				Pivot approach			
	$\bar{x}$	s^2	L_1	L_2	P_L	R_L	L_1	L_2
Normal $N(0; 1)$	0.208	1.146	-1.120	1.536	0.125	1.249	-0.084	0.333
Rectangular $R(0.5; 0.083)$	0.458	0.110	0.046	0.081	0.424	0.506	-1.332	2.180
Exponential $E(1; 1)$	0.805	0.550	-0.115	1.725	0.658	0.941	-0.806	2.123
Laplace $L(0; 2)$	-0.283	1.628	-1.866	1.299	-0.318	1.508	-0.760	0.124
Log-normal $LN(2.71; 47.21)$	1.265	2.029	-0.502	3.033	1.190	1.855	-0.154	2.536

3.4.5 The nonparametric estimates of variance

A survey of various nonparametric methods for estimating variance was made by Efron [8]. Here only two techniques, the Bootstrap method and Jackknife method, will be demonstrated. Both methods enable a calculation of confidence interval of variance and both are computer-assisted.

The Bootstrap method is based on the estimate $\hat{\theta}$ of parameter θ which is a known function of n independent random observations $x_1, \ldots, x_n$[i.e., $\hat{\theta} = f(x_1, \ldots, x_n)$]. The sample arises from an unknown distribution F_B. The variance of estimate $\hat{\theta}$ depends on the unknown distribution function F_B. The Bootstrap method substitutes the F_B distribution by a discrete distribution with probability $p_i = 1/n$ in points x_i, $i = 1, \ldots, n$.

The Bootstrap procedure:

(1) From the original sample (X_i), $i = 1, \ldots, n$, random sampling with replacement is used to create the B Bootstrap samples of size n. The ith element of the original sample may be present in the ith Bootstrap sample several times while another element may not be present at all. A selection of Bootstrap sample elements is made on the basis of random index i

$$i = \text{int}[\text{rnd}(0) \times n + 1] \tag{3.118}$$

where rnd(0) is the pseudorandom number from a generator of rectangularly distributed pseudorandom numbers between 0 and 1. When rnd(0) = 1, then $i = n$.

(2) Calculation of parameter estimates $\hat{\theta} = g(x_i)$, $i = 1, \ldots, B$, where x_i stands for the ith Bootstrap sample of size n.

(3) Calculation of the estimate of variance

$$\hat{\sigma}_B^2 = \frac{\sum_{i=1}^{B} (\hat{\theta}_i - \bar{\theta}_B)^2}{B - 1} \tag{3.119}$$

where

$$\bar{\theta} = \sum_{i=1}^{B} \hat{\theta}_i / B$$

represents the estimate of the mean. Usually the number of Bootstrap samples B is set between 200 and 1000.

(4) For construction of the confidence interval of the mean parameter and statistical tests the criterion

$$t_B = \frac{(\bar{\theta} - \mu)\sqrt{B}}{\hat{\sigma}_B} \tag{3.120}$$

is used; for large B values this has a standardized normal distribution.

The Jack-knife method is based on the use of "pseudovalues", y_i defined by

$$y_i = n \times \hat{\theta} - (n - 1) \times \hat{\theta}_{(i)} \tag{3.121}$$

where $\hat{\theta}_{(i)}$ is an estimate computed from all sample elements except the ith one. The mean $\bar{\theta}_J$ is calculated from

$$\bar{\theta}_J = \frac{1}{n} \times \sum_{i=1}^{n} y_i = n\hat{\theta} - \frac{n-1}{n} \sum_{i=1}^{n} \hat{\theta}_{(i)} \tag{3.122a}$$

With large or medium sample sizes, the pseudovalues y are taken to be approximately normally distributed. The random variable

$$t_J = \frac{\bar{\theta}_j - \theta}{\hat{\sigma}_J} \tag{3.122b}$$

has the standardized normal distribution. The variance of the Jack-knife estimate σ_J^2 is calculated from

$$\hat{\sigma}_J^2 = \frac{1}{n(n-1)} - \sum_{i=1}^{n} (y_i - \bar{\theta}_J)^2 \tag{3.122c}$$

Both nonparametric estimates $\hat{\sigma}_B^2$ and $\hat{\sigma}_J^2$ enable determination of a variance of various types of estimators of parameter θ.

Problem 3.16 *Confidence interval of variance by Bootstrap and Jack-knife methods* Calculate the 95% confidence interval of the sample variance s^2 for a sample of $n = 30$ taken from the Laplace distribution, with the use of the Bootstrap and Jack-knife methods.

Data: $n = 30$.

0.064	0.436	−1.072	−2.036	1.192	−3.162	−0.275
0.734	0.049	−0.569	1.144	−1.070	0.304	0.343
0.144	0.209	−1.269	4.452	5.477	−0.862	−0.026
0.210	0.695	−0.774	−3.723	−0.026	1.879	0.887
−0.333	1.038					

Solution: (*a*) *The Bootstrap method*: 400 simulated Bootstrap samples were selected. The average variance $\bar{s}_B^2 = 3.12$ and Bootstrap variance $\hat{\sigma}_B^2 = 1.53$. From Eq. (3.120) the limits of the 95% confidence interval of the variance are calculated as

$$L_{1,2} = \bar{s}_B^2 \pm 1.95 \times \hat{\sigma}_B^2 = 3.12 \pm 2.41.$$

A histogram of the values generated is shown in Fig. 3.11.

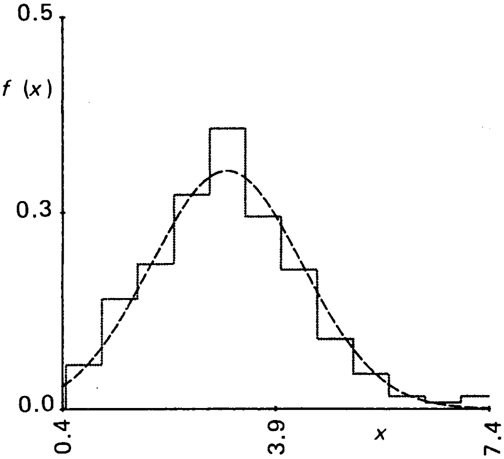


Fig. 3.11—The histogram of the 400 Bootstrap sample values simulated from an original sample size of $n = 30$ from the Laplace distribution.

(*b*) *The Jack-knife method*: the average value of variance $\bar{s}_J^2$ is 3.2 with variance $\hat{\sigma}_J^2 = 1.289$. From Eq. (3.122) the limits of the 95% confidence interval

$$L_{1,2} = \bar{s}_J^2 \pm 1.95 \times \hat{\sigma}_J^2 = 3.2 \pm 2.51.$$

Conclusion: The values for the confidence interval of the sample variance estimated by the Bootstrap and Jack-knife methods are in reasonable agreement.

3.5 STATISTICAL HYPOTHESIS TESTING

In many applications of statistics we are interested in making inferences about *population* (or ensemble) characteristics on the basis of observations made on a *random sample* of items from the population. The characteristics of interest may often be expressed in terms of population *parameters*, such as the population mean μ, or variance σ^2, or the proportion p of the population which has a certain characteristic. In other situations we may wish to make inferences about the difference between two (or more) populations, such as the difference between two population means μ_1 and μ_2.

A *statistical hypothesis* is a statement about the population distribution of some random variable. Hypothesis testing consists of comparing some statistical measures called ***test criteria*** (or statistics) deduced from a data sample with the values of these criteria taken on the assumption that a given hypothesis is correct. In hypothesis testing, one examines a ***null hypothesis*** H_0 against one or more ***alternative hypotheses*** $H_1, H_2, \ldots, H_A$ which are stated explicitly or implicitly. To reach a decision about the hypothesis, we select a value of α, which is termed the *significance level* for the test. Significance level α is usually arbitrarily selected to be fairly small, for example, α might be 0.05 or 0.01. The significance level α is related to the confidence coefficient $1 - \alpha$.

Since the alternative hypothesis is the hypothesis which is accepted when a null hypothesis is rejected, the procedure of hypothesis testing seems to be, in fact, a process of rejection of alternative hypotheses.

For hypothesis testing *the test statistic* (or *the test criterion*) is set up. When this statistic falls into the *the range of acceptance*, the null hypothesis is not rejected. When this statistic falls into *the region of rejection* (or *the critical range*) the null hypothesis is rejected. The probability of the test statistic falling in the region of rejection is equal to *the significance level.* It is expressed in %, e.g 5% or 1%.

A region of rejection can be set up as two-tailed (or two-sided) leading to a two-tailed test, or as a one-tailed test. The two-tailed test is used when test statistic can take values with either positive or negative sign. The significance level α is then split into two equal parts of magnitude $\alpha/2$.

3.5.1 Procedure for hypothesis testing

The procedure of statistical hypothesis testing is as follows:

(1) The null hypothesis H_0 and an alternative one are formulated.
(2) The significance level α is selected.
(3) The test statistic is chosen.

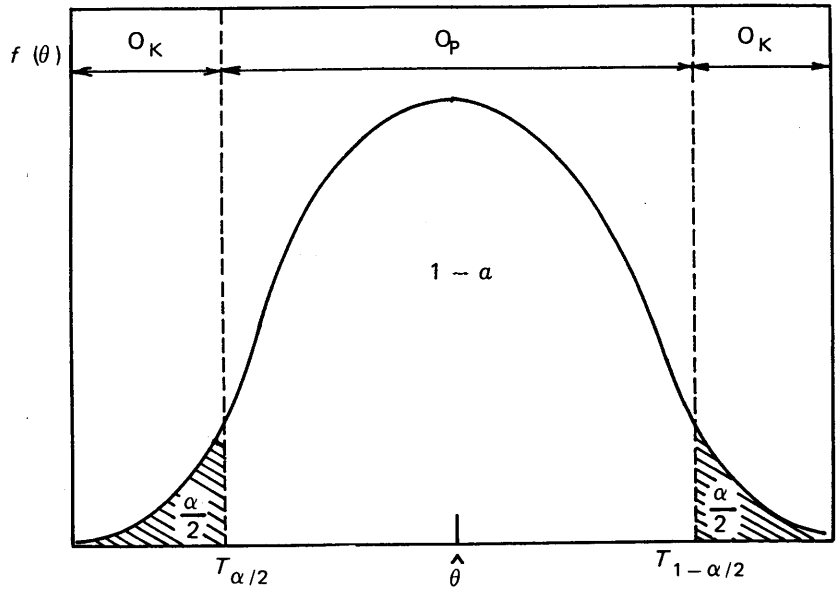


Fig. 3.12—Regions of rejection (O_K) and acceptance (O_P) for a symmetrical (two-tailed) hypothesis test.

(4) The region of rejection of the test statistic, on the basis of its probability distribution and the significance level, is determined.

(5) For the sample, the test statistic is calculated, and the limits of the region of rejection (Fig. 3.12).

(6) (*a*) The null hypothesis is rejected and the alternative one accepted when the value of the test statistic falls into the region of rejection;

(*b*) the null hypothesis is not rejected when the value of the test statistic does not fall into the region of rejection.

In judgement, it is necessary to remember that

(*i*) *rejection* of a null hypothesis H_0 does not necessarily mean that the tested null hypothesis is not *valid.* The rejection of a null hypothesis H_0 means only that we do not trust its validity because of the statistical test performed. It is understood in the following work that if the H_0 hypothesis is not valid then the alternative one H_A is valid;

(*ii*) *no rejection* of a null hypothesis H_0 does not imply its *acceptance.* When we do not reject a null hypothesis H_0 it means only that hypothesis testing did not provide sufficient reason for rejection of the hypothesis. If the H_0 hypothesis is not rejected, it can usually be assumed that either the H_0 (or some other hypothesis close to H_0) is valid.

We illustrate the procedure of hypothesis testing with an example in which we want to test the parameter θ. The sample size is large enough to allow us to use Eq.

(3.23). The null hypothesis H_0: $\theta = K$ where K is a known number, is tested against the alternative H_A: $\theta \neq K$. The test statistic

$$u_s = \frac{|\theta - K|}{\sqrt{D(\theta)}}$$

will have a normal distribution if the null hypothesis H_0 is valid. Testing the H_0 hypothesis can lead to the following results:

(1) the test statistic falls into a region of acceptance of the null hypothesis i.e.

$$u_{\alpha/2} < u_s < u_{1-\alpha/2}$$

and therefore H_0 is not rejected. If H_0 is valid the probability that u_s will fall out of range O_P is equal to the significance level α. The magnitude of α determines the magnitude of ***the error of the first kind***, i.e. wrong rejection of correct hypothesis H_0.

(2) The test statistic falls into a range of rejection O_K i.e. into the interval

$$u_s < u_{\alpha/2} \quad \text{or} \quad u_s > u_{1-\alpha/2}, \text{ respectively.}$$

The null hypothesis H_0 is then rejected in favour of the alternative H_A. The probability that u_s falls into the region of acceptance O_P even if H_0 is wrong represents the magnitude of ***the error of the second kind***, β. This error results when a wrong hypothesis H_A is accepted. The two types of error are illustrated in Fig. 3.13.

Both errors should be minimized. The probability that we will not make an error of the first kind is $(1 - \alpha)$. This is the probability of making the correct decision about the test hypothesis. The second correct decision is made with probability $(1 - \beta)$. This is the probability of not making an error of the second kind. We call *the power of the test* to discriminate $S = 1 - \beta$. It represents the hope of making a correct decision when the hypothesis is actually wrong.

The power of the test is affected by the sample size n: with bigger sample sizes, more information is available, and therefore the wrong hypothesis will be rejected with more confidence in favour of the alternative one. For $n \to \infty$, $S \to 1$.

Attention should be paid to the choice of significance level α in hypothesis testing:

(1) When the null hypothesis H_0 is not rejected at the significance level $\alpha = 0.05$, the difference between the theoretical value K and the estimated parameter $\hat{\theta}$ is not significant.

(2) When the null hypothesis H_0 is rejected also at the significance level $\alpha = 0.01$, the difference between the theoretical value K and the estimated parameter $\hat{\theta}$ is statistically significant.

(3) When the null hypothesis H_0 is rejected at the significance level $\alpha = 0.05$ but not at $\alpha = 0.01$, the sample size did not give sufficient information for a correct decision.

3.5.2 Hypothesis tests for the parameters of one population

Tests on the parameters of a single population enable the chemist to tell if the population mean (or variance) of a new product or variable (1) is different from, (2)

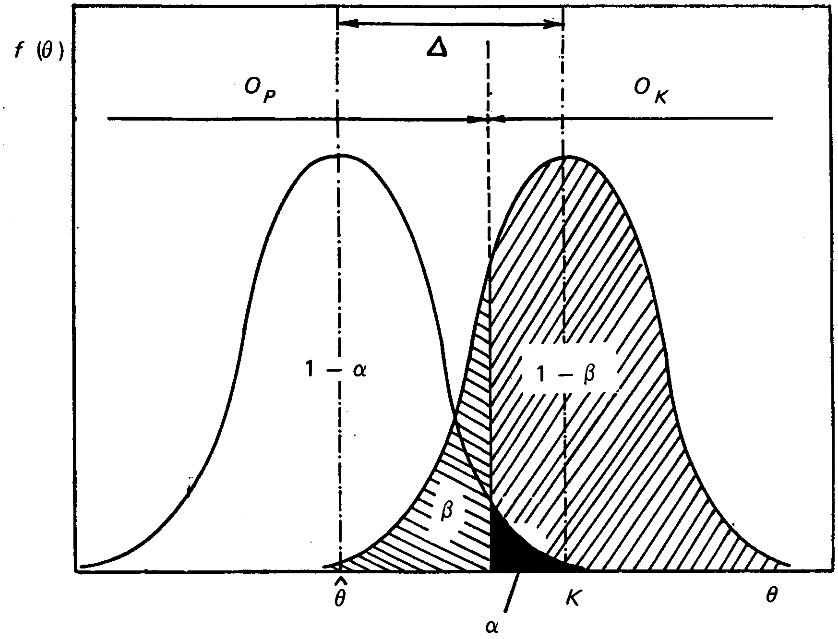


Fig. 3.13—Relationship between an error of the first kind and an error of the second kind.

exceeds, or (3) is less than the population mean (or variance) of a standard product or variable.

The hypothesis selected assumes that we know the value of the standard population parameters, μ_0, (or σ_0^2, respectively) from past experience or otherwise. As indicated in the third column of Table 3.12, a decision can be reached as follows:

(1) If the inequality proves to be true, i.e. if the calculated difference exceeds the right-hand side of the inequality, the hypothesis *is accepted.*

(2) If the inequality does not prove to be true, i.e. if the calculated difference does not exceed the right-hand side of the inequality, then the hypothesis *is rejected,* and there is little likelihood that the hypothesis is correct.

For a hypothesis test of a normal distribution based on a random sample x_i, $i = 1, \ldots, n$, with mean $\bar{x}$ and variance s^2, the random variables u (for large samples) and t (for small samples, $n < 30$) are used as test statistics:

$$u = \frac{\bar{x} - \mu_0}{\sigma} \times \sqrt{n} \approx N(0; 1) \qquad \text{and} \qquad t = \frac{\bar{x} - \mu_0}{s} \sqrt{n}$$

where $\bar{x}$ is the sample mean, s is the sample standard deviation and n is the sample size. The null hypothesis H_0: $\mu = \mu_0$ is followed by the alternative hypotheses (1) H_A: $\mu > \mu_0$, (2) H_A: $\mu < \mu_0$, (3) H_A: $\mu \neq \mu_0$, (see Table 3.12).

To test the hypothesis H_0: $\sigma^2 = \sigma_0^2$ the test statistic

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

is used. The tests assume that the observations are taken randomly from a normal random variable.

(1) *Test the hypothesis* H_0: $\mu = \mu_0$ *vs.* H_A: $\mu > \mu_0$. In this case, large positive values of t (or u) are evidence in favour of H_A. Values of t around zero are evidence in favour of H_0. Large negative values of t(or u), while they are unlikely if H_0 is true, are even more unlikely if H_A is true.

Hypothesis testing may be carried out also by calculating the probability $\alpha = P(t_{1-\alpha} > t)$ as an *one-tailed* (upper tail) *test*:

(*a*) If $\alpha > 0.05$ we say that *t (or u) is not significant* and there is *no evidence* for rejecting H_0 in favour of H_A.

(*b*) If α lies in the region $0.05 > \alpha > 0.01$ we say that *t (or u) is significant at the 5% level* and there is *some evidence* for rejecting H_0 in favour of H_A.

(*c*) If α lies in the region $0.01 > \alpha > 0.001$ we say that *t (or u) is significant at the 1% level* and it is generally interpreted as *strong evidence* for rejecting H_0 in favour of H_A.

(*d*) If $0.001 > \alpha$ we say that *t (or u) is significant at the 0.1% level* and there is *almost conclusive evidence* for rejecting H_0 in favour of H_A.

In the formulation of statistical hypothesis testing given above, the results of the tests are presented as evidence at various levels (none, some, strong, almost conclusive) in favour of the alternative hypothesis H_A.

An alternative formulation of statistical hypothesis testing is as follows: before the experiment takes place we decide on a fixed value of α (usually, but not necessarily, one of the values 0.05, 0.01, 0.001). This will determine a critical region in the tails of the distribution, such that if t falls within the critical region H_A is accepted, otherwise H_0 is accepted. It will be noted that even if H_0 is true there is a probability α that $t_\alpha(v)$ will fall into the critical region. Thus α may be interpreted as the probability of accepting H_A when H_0 is true; that is $\alpha = P$ (accepting H_A/H_0). The probability of incorrectly accepting H_A when H_0 is true could obviously be reduced by decreasing α.

(2) *Test of the hypothesis* H_0: $\mu = \mu_0$ *vs.* H_A: $\mu < \mu_0$. This time large negative values of t (or u) are evidence in favour of H_A. We calculate $\alpha = P(t_{1-\alpha} < t)$ and the values of α obtained are interpreted as in the previous case. This test is a *one-tailed* (lower tail) *test.*

(3) *Test of the hypothesis* H_0: $\mu = \mu_0$ *vs.* H_A: $\mu \neq \mu_0$. As in (1) and (2) we calculate $t = (\bar{x} - \mu_0)\sqrt{n}/s$ (or $u = (\bar{x} - \mu_0)\sqrt{n}/\sigma$). This time both large positive and large negative values of t (or u) are evidence in favour of H_A. We calculate

$$\begin{aligned}\alpha &= P(|t_{1-\alpha}| - |t|)\\ &= P(t_{1-\alpha} < |t|) + P(t_{1-\alpha} > |t|)\\ &= 2P(t_{1-\alpha} > |t|) \text{ from symmetry.}\end{aligned}$$

Again the values of α obtained are interpreted as in (1). This third type of test is referred to as a *two-tailed test.*

The critical regions for the three types of tests are shown in Fig. 3.14.

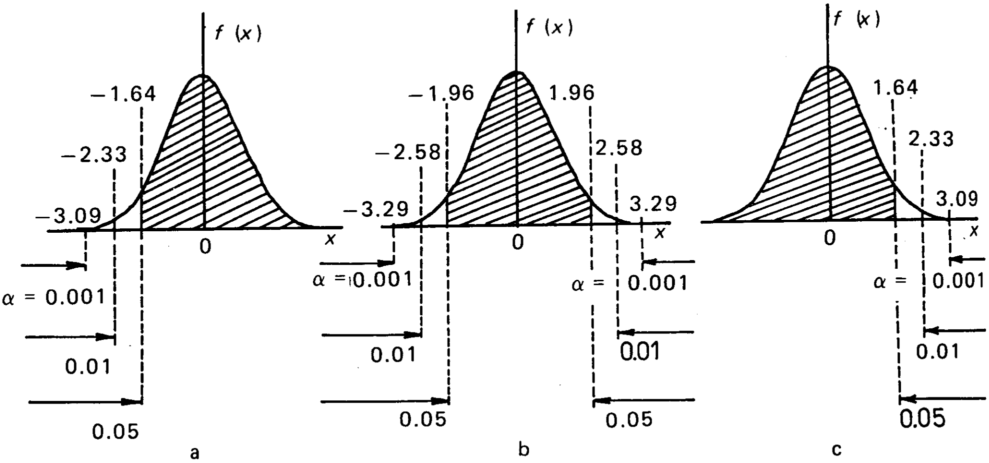


Fig. 3.14—Critical regions for (a) a one-tailed test (lower tail), (b) a two-tailed test, (c) a one-tailed test (upper tail).

To investigate whether t (or u) is significant, and if so, at what level, it is not necessary in practice to calculate the value of α. We simply compare the value of t (or u) with the critical values $t_{1-\alpha}$ (or $u_{1-\alpha}$) obtained form critical statistical tables.

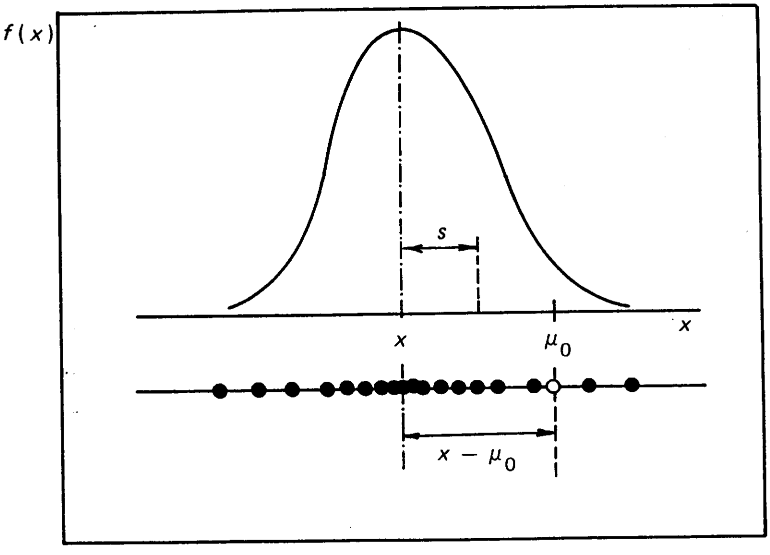


Fig. 3.15—The test of accuracy of a result μ, H_0: $\mu = \mu_0$ *vs.* H_A: $\mu \neq \mu_0$.

Table 3.12—Tests for comparing (a) the mean μ of a new product with a standard μ_0, and (b) these two products with regard to their variability: (a) H_0: $\mu = \mu_0$, σ unknown, s from sample used $t_{1-\alpha}(n-1)$ or $t_{1-\alpha/2}(n-1)$ is the quantile of the Student distribution. Instead of t, the u quantile may be used (see also Fig. 3.15).

H_A Hypothesis	Region of rejection	t-test
$\mu \neq \mu_0$	$\left\lvert \frac{(\bar{x}-\mu_0)}{s}\sqrt{n} \right\rvert \geqslant t_{1-\alpha/2}(n-1)$	Two-tailed
$\mu > \mu_0$	$\frac{(\bar{x}-\mu_0)}{s}\sqrt{n} \geqslant t_{1-\alpha}(n-1)$	One-tailed
$\mu < \mu_0$	$\frac{(\bar{x}-\mu_0)}{s}\sqrt{n} < t_{1-\alpha}(n-1)$	One-tailed

(b) H_0: $\sigma^2 = \sigma_0^2$, $\chi^2_\alpha(n-1)$, $\chi^2_{1-\alpha}(n-1)$, $\chi^2_{\alpha/2}(n-1)$ and $\chi^2_{1-\alpha/2}(n-1)$ are the quantiles of the Pearson χ^2 distribution.

H_A Hypothesis	Region of rejection	χ^2-test
$\sigma \neq \sigma_0^2$	$\chi^2_{\alpha/2}(n-1) \leq \frac{s^2(n-1)}{\sigma_0^2} \leq \chi^2_{1-\alpha/2}(n-1)$	Two-tailed
$\sigma^2 > \sigma_0^2$	$\frac{s^2(n-1)}{\sigma_0^2} \geq \chi^2_{1-\alpha}(n-1)$	One-tailed
$\sigma^2 < \sigma_0^2$	$\frac{s^2(n-1)}{\sigma_0^2} < \chi^2_\alpha(n-1)$	One-tailed

Problem 3.17 *Test of the sample mean from the log–normal distribution*
A random sample is taken from the log–normally distributed population of copper in kaolin (Problem 3.9). Test whether the sample mean is equal to the expected value $M = 10$, i.e. H_0: $M = 10$ *vs.* H_A: $M \neq 10$.
Data: from Problem 3.9.
Solution: Since $n = 32 > 30$, we use the test statistic for large samples $u = |\hat{M} - 10|/\sqrt{D(\hat{M})}$, which has the standardized normal distribution. Calculation of $u = |9.976 - 10|/\sqrt{0.02} = 1.2$ leads to a lower value than the $u_{1-\alpha/2}$ quantile ($u_{1-0.05/2} = 1.96$ for $\alpha = 0.05$), $1.2 < 1.96$ and hence the hypothesis H_0 is accepted.
Conclusion: the sample mean is equal to the expected value of 10.

3.5.3 Hypothesis tests for the parameters of two populations

Comparison of two samples $\{x_i\}$, $i = 1, \ldots, n_1$ and $\{y_i\}$, $i = 1, \ldots, n_2$, is a frequent problem in the instrumental laboratory, for

(1) comparing results of different instrumental methods or laboratories,
(2) examining the need to separate heterogeneous samples into homogeneous classes, and
(3) classifying the difference between various materials or various instruments.

Sometimes the problem of two populations may be tested as the problem of one population. If the elements x_i represent some response before a treatment of a material, and elements y_i the same response after the treatment, the difference between the responses of a pair of values, $d_i = x_i - y_i$, then gives a measure of the effectiveness of the treatment. Let $\bar{d}$ and s_{d} be the mean and standard deviation of these differences. If the differences d are independently normally distributed (or nearly so) with mean zero and (unknown) variance σ^2 it can be shown that the statistic

$$t = d \times \sqrt{n}/s_{\mathrm{d}}$$

has Student's t-distribution with $v = n - 1$ degrees of freedom. Here n is the number of matched pairs in the experiment (Fig. 3.16). It is assumed here that each set of measurements may be regarded as a sample from a normal population. This statistic may be used to test the hypothesis H_0: $\mu_{\mathrm{D}} = \mu_x - \mu_y = 0$, (i.e. there is no treatment effect) *vs.* H_A: $\mu_{\mathrm{D}} \neq 0$. From the hypothesis H_A: $\mu_{\mathrm{x}} \neq \mu_{\mathrm{y}}$ we see that a two-tailed test is required.

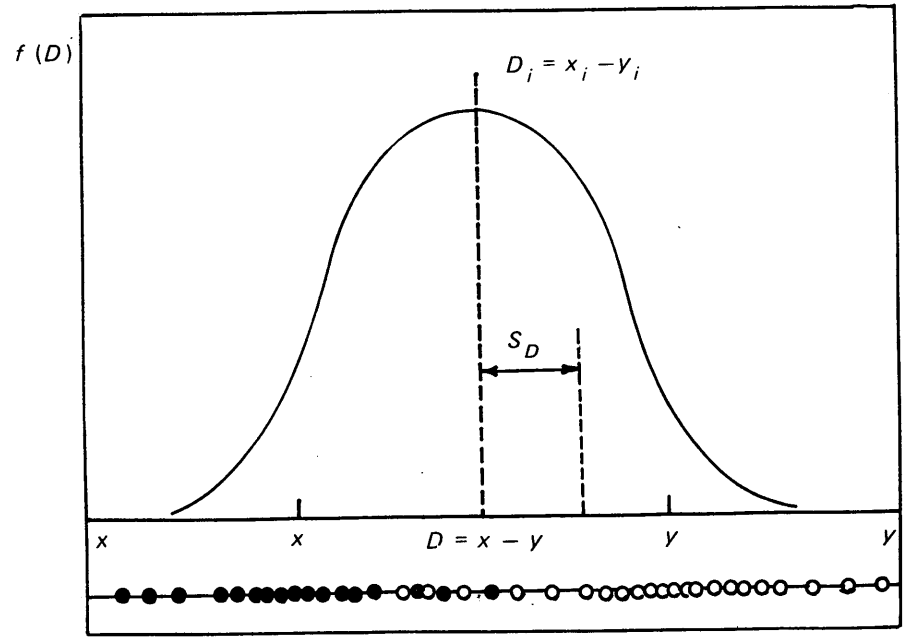


Fig. 3.16—The pair test.

Before statistical testing, the methods of exploratory data analysis should be applied to both samples $\{x_i\}$ and $\{y_i\}$, $i = 1, \ldots, n$. For each sample the box-and-whisker plot G4 and notched box-and-whisker plot G5 are examined. The assumption that both samples have the same distribution is examined by the empirical quantile–quantile plot G14 (Q–Q plot) with the ordered quantities $y_{(i)}$ on the y-axis and $x_{(i)}$ values on the x-axis. When both samples have the same distribution, the points $(x_{(i)}, y_{(i)})$ should lie on a line $y = x$ with slope equal to 1. When the empirical Q–Q plot has the equation $y = kx + q$, mean $\bar{y} = k\bar{x} + q$ and variance $s_y^2 = k^2 s_x^2$. A nonlinear pattern in the Q–Q plot indicates that the distributions of the samples differ significantly.

For different sample sizes, $n_1 > n_2$, the empirical Q–Q plot is drawn for the ordered elements of the smaller sample i.e. for $y_{(i)}$ here. The values of the quantiles of the larger sample $\tilde{x}_j$ are computed according to

$$\tilde{x}_j = (1 - z)x_{(k)} + zx_{(k+1)} \tag{3.123}$$

where $k = \text{int}(v_j)$ is the integer part of real number v_j,

$$v_j = (j - 0.5)\, n_1/n_2 + 0.5$$

and $z = v_j - k$. An empirical Q–Q plot is constructed from n_2 pairs of points $(y_{(j)}, \tilde{x}_j)$.

After exploratory data analysis, the classical tests of significance of difference in the parameters of location and spread is applied. Classical tests assume that

(1) the two samples (x_i), $i = 1, \ldots, n_1$, and (y_i), $i = 1, \ldots, n_2$ are independent,
(2) the two populations distributions are normal, $x_i \simeq N(\mu_x, \sigma_x^2)$ and $y_i \simeq N(\mu_y, \sigma_y^2)$.

3.5.3.1 Comparisons of population means

Consider a random sample of size n_1, with mean $\bar{x}$ and variance s_x^2, from a population P_1 with unknown mean μ_x and unknown variance σ_x^2, and an independent random

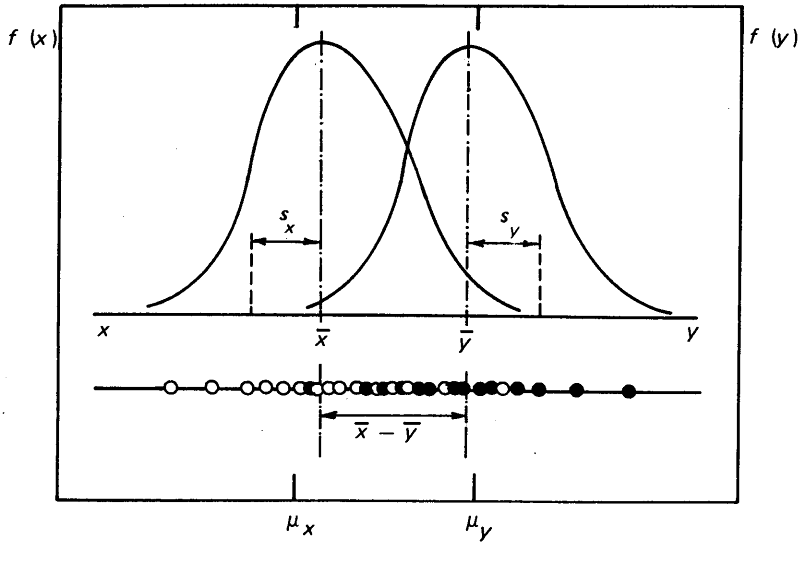


Fig. 3.17—The test for comparison of mean values.

sample of size n_2 with mean $\bar{y}$ and variance s_y^2, from population P_2 with unknown mean μ_y and unknown variance σ_y^2. The hypothesis H_0: $\mu_x = \mu_y$ is tested against the alternative H_A: $\mu_x \neq \mu_y$ (Fig. 3.17).

To proceed further we distinguish two cases:

(1) If we know that $\sigma_x^2 = \sigma_y^2$, the test statistic

$$t_1 = \frac{|\bar{x} - \bar{y}|}{\sqrt{(n_1 - 1)s_x^2 + (n_2 - 1)s_y}} \times \left[\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}\right]^{1/2} \tag{3.124}$$

has Student's t-distribution with $v = n_1 + n_2 - 2$ degrees of freedom and significance level α. When $t_1 > t_{1-\alpha/2}(v)$, the null hypothesis H_0 is rejected.

(2) If we know that $\sigma_x^2 \neq \sigma_y^2$, the test statistic

$$t_2 = |\bar{x} - \bar{y}| \, (s_x^2/n_1 + s_y^2/n_2)^{-1/2} \tag{3.125}$$

has Student's t-distribution with v degrees of freedom expressed by

$$v = \frac{s_x^2/n_1 + s_y^2/n_2}{\dfrac{s_x^4}{n^2(n_1 - 1)} + \dfrac{s_y^4}{n_2^2(n - 1)}} \tag{3.126}$$

When $t_2 > t_{1-\alpha/2}(v)$, the null hypothesis H_0 is rejected at the significance level α.

In some problems of comparison of means, the variances σ_x^2 and σ_y^2 are unknown. Posten, Yeh and Owen [9] found that for $n_1 = n_2 > 8$ the test statistic t_1 can be applied even for $\sigma_x^2 \neq \sigma_y^2$.

For different sample sizes $n_1 \neq n_2$ and variance ratio $\sigma_x^2/\sigma_y^2 \approx 1$ the test statistic t_1 can be used. If $n_2 > n_1$ and also n_2 is large enough, the test statistic t_1 can be used provided that

$$0.82 \leq \frac{\dfrac{n_2}{n_1} \times \dfrac{s_x^2}{s_y^2} + 1}{\dfrac{s_x^2}{s_y^2} \times \dfrac{n_2}{n_1}} \leq 1.17 \tag{3.127}$$

where $s_x^2 > s_y^2$. Test statistic t_1 is non-robust when the variance of data is not constant. The test criterion t_1 is not robust against heteroscedasticity i.e. against a case of data being measured with different precision. For such case the test criterion t_2 is preferable because it is more robust. However, the number of degrees of freedom v calculated by Eq. (3.116) are less than $n_1 + n_2 - 2$, so that the power of test t_2 is lower than that of t_1, and also the probability of a Type II error β increases.

When both samples are not from a normal distribution, the modified test criterion t_3 is used

$$t_3 = \frac{|\bar{x} - \bar{y}| + C + D(\bar{x} - \bar{y})^2}{\sqrt{s_x^2/n_1 + s_y^2/n_2}} \tag{3.128}$$

where

$$C = \frac{1}{6}\left[\frac{\hat{g}_{1x}}{n_1^2} \times \frac{s_x^3}{\sqrt{n_1}} - \frac{\hat{g}_{1y}}{n_2^2} \times \frac{s_y^3}{\sqrt{n_2}}\right]\left[\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}\right]^{-1} \tag{3.129}$$

$$D = \frac{1}{3}\left[\frac{\hat{g}_{1x}}{n_1^2} \times \frac{s_x^3}{\sqrt{n_1}} - \frac{\hat{g}_{1y}}{n_2^2} \times \frac{s_y^3}{\sqrt{n_2}}\right]\left[\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}\right]^{-2} \tag{3.130}$$

Here $\hat{g}_{1x}$ and $\hat{g}_{2x}$ are sample skewness. In order to use the quantiles of Student's t-distribution for a declared significance level α, another statistic t'_3 should be used

$$t'_3 = t_2 + B_x - B_y \tag{3.131}$$

where

$$B_x = \left[\frac{\hat{g}_{1x} s_x^3}{6\ n_1^2\sqrt{n_1}\left[\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}\right]} + \frac{\hat{g}_{1x} s_x^2(\bar{x} - \bar{y})^2}{3\ n_2^2\sqrt{n_2}\left[\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}\right]}\right] \times \left[\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}\right]^{-1/2} \tag{3.132}$$

Quantity B_y is calculated analogously with $\hat{g}_{1y}$, σ_y^2 and n_2. The test criterion t'_3 for H_0 has the Student t-distribution with $v = n_1 + n_2 - 2$ degrees of freedom. This statistic is robust for distorted samples distributions, also for heteroscedasticity in data and different sample variances $\sigma_x^2 \neq \sigma_y^2$.

The Brown and Forsythe test concerns comparison of population means on the basis of small samples taken from normal populations. The null hypothesis is H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$. Suppose that there are k samples, each of size n_i from a normal distribution with sample mean $\bar{x}_i$ and variance s_i^2, $i = 1, \ldots, k$. The test statistic

$$F = \frac{\sum_{i=1}^{k} n_1(\bar{x}_i - \bar{\bar{x}})^2}{\sum_{i=1}^{k}\left[1 - \frac{n_i}{\bar{n}}\right] s_i^2} \tag{3.133}$$

where

$$\bar{\bar{x}} = \frac{1}{\bar{n}} \sum_{i=1}^{k} n_i \bar{x}_i \tag{3.134}$$

and

$$\bar{n} = \sum_{i=1}^{k} n_i \tag{3.135}$$

has in H_0 the Fisher–Snedecor distribution with $(k - 1)$ and v degrees of freedom. When the sample value F is greater than the quantile $F_{1-\alpha/2}(k - 1, v)$ at the 5% level, H_0 is rejected. The number of degrees of freedom is here given by

$$v = 1/\sum_{i=1}^{n} \frac{o_i}{n_i - 1} \tag{3.136}$$

where

$$o_i = \left[[1 - n_i/\bar{n}]s_i^2\right] \Big/ \left[\sum_{i=1}^{n} [1 - n_i/\bar{n}]s_i^2\right] \tag{3.137}$$

Data: (*A*) The sample from $N(0, 1)$:

−1.008	−0.500	0.749	1.723	0.076	0.569	−1.389
0.087	1.112	−0.235	0.519	0.279	−0.758	−0.588
−0.594	−0.885	−0.072	1.980	0.063	0.016	−0.673
−0.993	0.752	0.092	0.236	−2.962	−0.383	0.109
−1.285	0.634	0.690	1.134	−0.711	−1.825	2.374
0.500	−1.380	0.046	−0.544	−0.150	−1.129	1.173
1.401	−2.121	0.521	0.280	1.440	−0.415	−0.443
−0.384						

(*B*) The sample from $N(3, 1)$:

1.992	2.500	3.749	4.723	3.076	3.569	1.611
3.087	4.122	2.765	3.519	3.279	2.242	2.412
2.406	2.115	2.928	4.980	3.063	3.016	2.327
2.007	3.752	3.092	3.236	0.038	2.617	3.109
1.715	3.634	3.690	4.134	2.289	1.175	5.374
3.500	1.620	3.046	2.456	2.850	1.871	4.173
4.401	0.809	3.521	3.280	4.440	2.585	2.557
2.616						

(*C*) The sample from $L(0, 2)$:

0.064	0.436	−1.072	−2.036	1.192	−3.162	−0.275
0.734	0.049	−0.569	1.144	−1.070	0.304	0.343
0.144	0.209	−1.269	4.452	5.477	−0.862	−0.026
0.210	0.695	−0.774	−3.723	−0.026	1.879	0.887
−0.333	1.038	−0.188	−1.749	−1.224	0.126	−2.249
−0.835	−0.261	0.715	1.223	0.535	−0.632	1.411
−2.538	−0.206	1.136	−1.070	−0.429	−0.529	1.153
0.319						

(*D*) The sample from $L(2, 2)$:

2.064	2.436	0.928	−0.036	3.192	−1.162	1.725
2.734	2.049	1.431	3.144	0.930	2.304	2.343
2.144	2.209	0.731	6.452	7.477	1.138	1.974
2.210	2.695	1.226	−1.723	1.974	3.879	2.887
1.667	3.038	1.812	0.251	0.776	2.126	−0.249
1.165	1.739	2.715	3.223	2.535	1.368	3.411
−0.538	1.794	3.136	0.930	1.571	1.471	3.153
2.319						

Program: Chemstat: Basic Statistics: Two sample testing.
Solution: Examining the variance of the two samples by the Fisher–Snedecor test proved (*cf.* Problem 3.19) that $\sigma_1^2 = \sigma_2^2$ and therefore statistics t_1, t_3 and t_4 can be used. Table 3.14 shows the test statistics with the quantiles of the Student t-test for significance level $\alpha = 0.05$.

Table 3.13—Testing of the sample means of two populations (a) N (0, 1) and N (0, 3), and (b) L (0, 2) and L (2, 2) at significance level $\alpha = 0.05$.

Test statistic	Sample from N(0, 1) and sample from N(3, 1)	Sample from L(0, 2) and sample from L(2, 2)
t_1	15.65	4.42
quantile	1.985	1.985
t_3	14.72	4.7
quantile	1.984	1.985
t_4	16.11	5.22
quantile	1.988	1.988

Conclusion: All the test statistics indicate a significant difference between the means.

3.5.3.2 Comparisons of the variances of two populations

The comparison of variances is particularly important for analysis of responses obtained from experimental design. A random sample of size n_1 is taken from a normal population with unknown mean μ_x and variance σ_x^2 and an independent

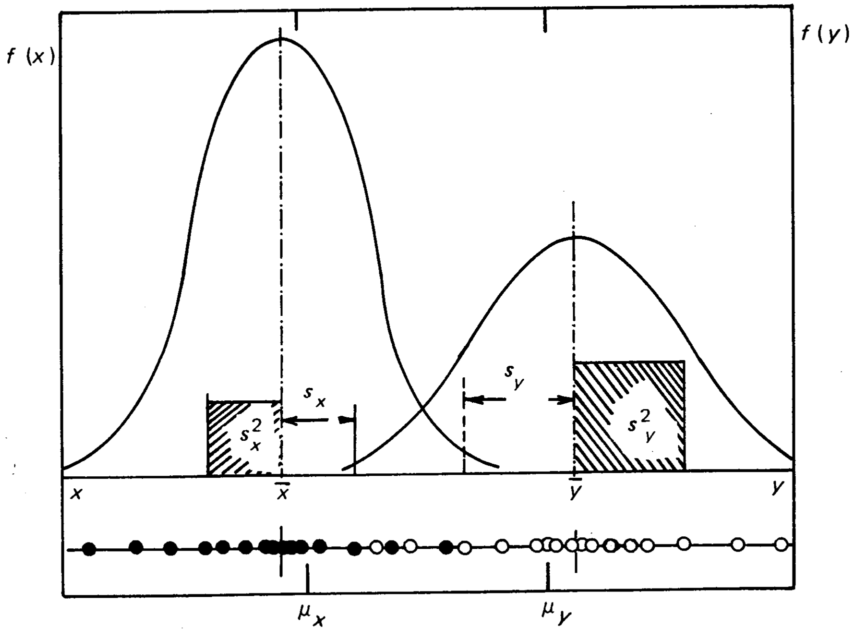


Fig. 3.18—The Fisher–Snedecor test for identity of two variances.

random sample of size n_2 from a second normal population with unknown mean μ_y and variance σ_y^2. We take $\hat{\sigma}_x^2 = s_x^2$ and $\hat{\sigma}_y^2 = s_y^2$ where s_x^2 and s_y^2 are the sample variances.

For normal and independent populations, the hypothesis H_0: $\sigma_x^2 = \sigma_y^2$ against H_A: $\sigma_x^2 \neq \sigma_y^2$ can be tested by using the statistics

$$F = \max\left(\frac{s_x^2}{s_y^2}, \frac{s_y^2}{s_x^2}\right) \tag{3.145}$$

When $s_x^2 > s_y^2$, the F statistic has the Fisher–Snedecor F-distribution with $v_1 = n_x - 1$ and $v_2 = n_y - 1$ degrees of freedom. In the case when $s_y^2 > s_x^2$ the order of degrees of freedom must be changed. The test is illustrated in Fig. 3.18.

Fisher–Snedecor test is sensitive to the assumption of normality: if the kurtosis of the samples differs from the normal distribution, for the quantile $F_{1-\alpha/2}(v_1, v_2)$ the numbers of degrees of freedom v_1 and v_2 should be calculated from

$$v_1 = (n_1 - 1)/(1 + \hat{g}_{2c}/2) \tag{3.146}$$

$$v_2 = (n_2 - 1)/(1 + \hat{g}_{2c}/2) \tag{3.147}$$

where

$$\hat{g}_{2c} = \frac{2(n_1 + n_2)\left[\sum_{i=1}^{n_1} (x_i - \bar{x})^4 + \sum_{i=1}^{n_2} (y_i - \bar{y})^4\right]}{\left[\sum_{i=1}^{n_1} (x_i - \bar{x})^2 + \sum_{i=1}^{n_2} (y_i - \bar{y})^2\right]^2} - 3 \tag{3.148}$$

When more than two sample variances are to be tested, H_0: $\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_k^2$, the *Jack-knife test* is recommended. The Jack-knife statistic is calculated by

$$F_j = \frac{[n_1(\bar{z}_1 - \bar{z})^2 + n_2(\bar{z}_2 - \bar{z})^2](n_1 + n_2 - 2)}{\sum_{i=1}^{n_1} (z_{1i} - \bar{z}_1)^2 + \sum_{i=1}^{n_2} (z_{2i} - \bar{z}_2)^2} \tag{3.149}$$

where

$$\bar{z} = \frac{n_1\bar{z}_1 + n_2\bar{z}_2}{n_1 + n_2} \tag{3.150}$$

$$z_j = \frac{\sum_{i=1}^{n_j} z_{ji}}{n_j}, \quad j = 1, 2 \tag{3.151}$$

$$z_{1i} = n_1 \times \ln s_x^2 - (n_1 - 1) \times \ln s_{1(i)} \tag{3.152}$$

$$s_{1(i)}^2 = \frac{1}{n_1 - 2} \sum_{j \neq i}^{n_1} (x_j - \bar{x}_{(i)})^2 \tag{3.153}$$

In Eq. (3.153) the sample mean $\bar{x}_{(i)}$ is calculated from a sample with the ith element omitted

$$\bar{x}_{(i)} = \frac{1}{n_1 - 1} \sum_{j=i}^{n_1} x_j \tag{3.154}$$

Analogously, for the sample y_i, $i = 1, \ldots, n_2$ the z_{2i}, $s_y^2, \ldots$ etc. are calculated. The Jack-knife statistic has the Fisher–Snedecor distribution with $\nu_1 = 2$ and $\nu_2 = n_1 + n_2 - 2$ degrees of freedom. When the sample value F is greater than the quantile $F_{1-\alpha/2}(\nu_1, \nu_2)$, the H_0 is rejected at the significance level α.

Problem 3.19 *The variances of two samples*
For the two pairs of samples from Problem 3.18, test whether both samples in the pair have the same variance. Do the test at significance level $\alpha = 0.05$.
Data: As for Problem 3.18
Program: Chemstat: Basic Statistics: Two sample testing.
Solution: The variances of two samples are tested with the use of two statistical tests, the Fisher–Snedecor test (F) and Jack-knife test (F_J). Table 3.15 compare the results of the tests.

Table 3.14—Test statistics and quantiles for the 5% level, for data from Problem 3.18

Test statistics	Sample of $N(0, 1)$ with sample of $N(3, 1)$	Sample of $L(0, 2)$ with samples of $L(2, 2)$
F	1.006	1.003
quantile	1.762	1.762
F	0.0003	2.21×10^{-5}
quantile	3.831	3.831

Conclusion: Both tests prove the validity of the H_0 hypothesis, the variances are the same. Generally, F-tests are more sensitive to the classical assumptions of an actual distribution than the T-tests.

3.6 SUMMARY OF PROCEDURES FOR UNIVARIATE DATA ANALYSIS

Univariate data analysis involves the following steps:

1. *Confirmatory data analysis*, examining assumptions about data
   1.1 Examining the independence of sample elements.
   1.2 Examining the normality of the sample distribution.
   1.3 Examining the outliers by the use of modified external hinges $[V_L^*, V_U^*]$.
   1.4 Examining the minimum sample size n_{min}

In addition to these four tasks, other diagnostics of exploratory data analysis may be used to verify the actual sample distribution.

2. *Determination of point and interval estimates of parameters.*

The choice of the type of statistical characteristics depends on the distribution of the population the sample comes from. With an assumption of normality, independence and homogeneity of sample, the moment characteristics are calculated. If outliers are present in the data, robust characteristics are calculated.

2.1 *The moment characteristics of location and spread* include the arithmetic mean $\bar{x}$, Eq. (3.39), with its variance $D(\bar{x})$, Eq. (3.40), and the confidence interval, Eq. (3.27), in which the true value μ exists; the estimate of sample variance s^2, Eq. (3.43), and its confidence interval, Eq. (3.29), the mean absolute deviation $\bar{d}$, Eq. (3.16), the variance coefficient δ, Eq. (3.17), with its variance $D(\delta)$, Eq. (3.18); the weighted arithmetic mean $\bar{x}_w$ Eq. (3.11), with its variance $D(\bar{x}_w)$, Eq. (3.12).

2.2 *The characteristics of shape* give information about the distribution shape; they include the skewness $\hat{g}_1$, Eq. (3.19), with its variance $D(\hat{g}_1)$, Eq. (3.19); the kurtosis $\hat{g}_2$. Eq. (3.20), with its variance $d\,(\hat{g}_2)$, Eq. (3.20a); and also the selector characteristics Q_1, Eq. (3.104), and Q_2, Eq. (3.105).

2.3 *The quantile and robust characteristics of location and spread* are less sensitive to outliers than the moment characteristics. These characteristics include the median $\tilde{x}_{0.5}$ (Section 3.4.1) with its nonparametric estimate of variance s_M^2, Eq. (3.93), or s_M^*, Eq. (3.95), and the confidence interval of the median; the mode (Section 3.1); the upper $\tilde{x}_{0.75}$ and the lower $\tilde{x}_{0.25}$ quartile, which are useful for the calculation of the interquantile range R, Eq. (3.15). The simplest and the most effective robust estimate of location is the trimmed mean $\bar{x}(\varkappa)$, Eq. (3.97), with its winsorized variance $s_w(\varkappa)$, Eq. (3.98), and also the asymmetric trimmed mean $\bar{x}(\varkappa_1, \varkappa_2)$, Eq. (3.101), with its variance $s_w^2(\varkappa_1, \varkappa_2)$, Eq. (3.103), which is suitable for asymmetric skewed distributions. To select the extent of trimming, some selector criteria are used: the relative tail length Q_1, Eq. (3.104), or the estimate of relative skewness Q_2, Eq. (3.105). When a sample contains outliers the robust M-estimates with bi-quadratic function are used. This leads to the value of location $\hat{\mu}_M$ (3.107) with its variance $D(\hat{\mu}_M)$, Eq. (3.112), and the confidence interval calculated with use of the random variable t_M, Eq. (3.113). For small samples, the pivot halfsum P_L, Eq. (3.114), and the pivot range R_L, Eq. (3.115), are used. The random variable T_L, Eq. (3.116), is used for calculation of the confidence interval of the mean μ.

2.4 *The nonparametric estimates of variance* permit calculation of an estimate of the variance of any parameter of distribution θ and also construction of the corresponding confidence interval. The estimate of location $\hat{\theta}_B$ and variance $\hat{\sigma}_B^2$, Eq. (3.119), by the Bootstrap method or $\hat{\theta}_J$, Eq. (3.122a) with its variance $\hat{\sigma}_J^2$, Eq. (3.122c), by the Jack-knife method, are used.

2.5 *The maximum likelihood estimates* for distributions other than normal are calculated as follows:

(1) For the *Poisson distribution* $\hat{\lambda}$, Eq. (3.34), $D(\hat{\lambda})$, Eq. (3.35), and the confidence interval of parameter λ, Eq. (3.36a,b);

(2) For the *Laplace distribution* $\hat{\phi}$, Eq. (3.46), $D(\hat{\phi})$, Eq. (3.47), and the confidence interval of parameter ϕ, Eq. (3.48).

(3) For the *rectangular distribution* ln L, Eq. (3.50), $\hat{h}$, Eq. (3.51), $\hat{a}$, Eq. (3.52), with their variances $D(\hat{h})$, Eq. (3.53), $D(\hat{a})$, Eq. (3.54), and the confidence interval, Eq. (3.28).

(4) For the *exponential distribution (one-parameter)* $\hat{\theta}$, Eq. (3.57), $D(\hat{\theta})$ Eq. (3.58) and the confidence interval for θ, Eq. (3.59).

(5) For the *exponential distribution (two-parameters)* $\hat{\theta}$, Eqs. (3.64, 3.65a, 3.66a), $D(\hat{\theta})$, Eqs. (3.65b, 3.66b) and unbiased estimates, Eqs. (3.67, 3.68), the confidence interval of parameter θ, Eq. (3.69) and Eq. (3.70).

(6) For *log–normal distribution (two-parameters)* $E(x)$, Eq. (3.74), $D(x)$, Eq. (3.75), $\hat{g}_1$, Eq. (3.76a), $\hat{g}_2$, Eq. (3.76b), δ, Eq. (3.77), $\hat{x}_M$, Eq. (3.78a), $\tilde{x}_{0.5}$, Eq. (3.78b), the confidence interval $\tilde{x}_{0.5}$, Eq. (3.82) and δ, Eq. (3.83). The estimate of the mean value is $\hat{M}$, Eq. (3.84), with variance $\hat{V}$, Eq. (3.85), with variances $D(\hat{M})$, Eq. (3.88) and $D(\hat{V})$, Eq. (3.89). The confidence interval of parameter M is calculated from Eq. (3.91).

(7) For *the log–normal distribution (three-parameters)* $\hat{\mu}(\theta)$, Eq. (3.80), and $\sigma^2(\theta)$, Eq. (3.81).

3. *Statistical hypothesis testing*

The simple test of the parameters of population on the basis of one sample uses the $100(1-\alpha)\%$ confidence interval of parameter θ. If the given value θ_0 lies in this interval, the null hypothesis H_0: $\theta = \theta_0$ is accepted, otherwise the alternative one, H_A: $\theta \neq \theta_0$.

For testing hypotheses about two populations on the basis of two samples, the first step is the test of homogeneity of variances of the two samples by the Fisher–Snedecor F-test, Eq. (3.145). However, this test is rather sensitive to any deviation of the distribution from normality; the Jack-knife test F_J, Eq. (3.149) or some robust test of location t_4, Eq. (3.138) or t_5, Eq. (3.139) are more suitable.

The classical Student t-test t_1, Eq. (3.124), or t_2, Eq. (3.125), is robust enough when the actual distribution deviates from normality, but the two sample sizes are the same. When both samples deviate in skewness from the normal distribution, the test characteristic t_3, Eq. (3.128), is more convenient.

3.7 ADDITIONAL SOLVED PROBLEMS

Problem 3.20 *Probability calculation for data from the normal distribution*
For normally distributed data $N(\mu, \sigma^2)$ evaluate

(a) the relative number of sample elements which lie in the interval $\langle \mu, \mu + \sigma \rangle$;
(b) the relative number of sample element which lie above the limit $\mu + 2\sigma$;
(c) the limiting element $x_{(m)}$ above which just 1% of all sample elements lie.

Solution: (*a*) The probability P of existence of a random element in the interval $\langle x_1, x_2 \rangle$ is equal to the difference between values of distribution function, $P = F(x_2) - F(x\ )$. The relative number of sample elements in the given interval as a percentage is equal to $P \times 100\%$. The relative number of sample elements in the interval $\langle \mu, \mu + \sigma \rangle$ can be computed by the use of the tabulated distribution function for normalized variable

$$u = (x - \mu)/\sigma$$

For this case

$$u_1 = (\mu - \mu)/\sigma = 0,$$

$$u_2 = (\mu + \sigma - \mu)/\sigma = 1$$

and

$$P = F(1) - F(0) = 0.8413 - 0.5 = 0.3413.$$

Therefore in the interval $\langle \mu, \mu + \sigma \rangle$ there are 34.13% of all sample elements.

For the interval $\langle \mu - k\sigma, \mu + k\sigma \rangle$, we have

$$P = F(\mu + k\sigma) - F(\mu - k\sigma) = 2F(k) - 1.$$

For $k = 1$, $F(1) = 0.8413$ and $P = 0.6826$. Therefore, in the interval $\langle \mu - \sigma, \mu + \sigma \rangle$ there are 68.26% of all sample elements.

(*b*) Determination of the probability P for which the sample elements reach higher values than the limiting element x_{M}. The expression

$$P = 1 - F(x_{\mathrm{M}})$$

is valid. Since the normalized variables are used, we need to determine the corresponding u_{M}. For

$$x_{\mathrm{M}} = \mu + 2\sigma,$$

$$u_{\mathrm{M}} = (\mu + 2\sigma - \mu)/\sigma = 2$$

and

$$P = 1 - F(2) = 0.02275.$$

Above the value $x_{\mathrm{M}} = \mu + 2\sigma$, there are 2.275% of all elements.

(*c*) For limiting element x_{M} above which there are $100 \times P\%$ elements the relation

$$P = 1 - u_{\mathrm{M}}$$

is valid, where

$$u_{\mathrm{M}} = (x_{\mathrm{M}} - \mu)/\sigma$$

and

$$x_{\mathrm{M}} = \sigma F^{-1}(1 - P) + \mu.$$

The symbol $F^{-1}()$ means the quantile function which is inverse to the distribution function: it is available in statistical tables. For $P = 0.01$ it is

$$F(1 - 0.01) = 2.33$$

and therefore, above the limiting element

$$x_{\mathrm{M}} = 2.33\sigma + \mu$$

there are only 1% of all samples.

Conclusion: For this sort of problem, the normalizing transformation is very useful.

Problem 3.21 *Calibration of a pipette, with large sample size*
A pipette of volume 10 ml was calibrated by weighing the water delivered, and 32 measurements were obtained. Determine the point and interval estimates of the real volume of the pipette.
Data: $n = 32$, V[ml]:

9.9820, 9.9656, 9.9940, 9.9877, 9.9865, 9.9755, 9.9820,
9.9794, 9.9184, 9.9848, 9.9914, 9.9905, 9.9726, 9.9661,
9.9857, 9.9889, 9.9832, 9.9923, 9.9877, 9.9779, 9.9936,
9.9666, 9.9903, 9.9666, 9.9713, 9.9762, 9.9840, 9.9723,
9.999, 9.9887, 9.9921. 9.9889

Program: Chemstat: Basic Statistics: Assumptions testing, one sample analysis, exploratory continuous.
Solution: The point estimates of location, spread and shape are:

$$\bar{x} = 9.9810 \text{ ml}, s^2 = 2.153 \times 10^{-4}, \hat{g}_1 = -2.44, \hat{g}_2 = 11.17.$$

Examination of the data leads to the following conclusions:

(1) data are independent of the time, from $t_n = 0.19$;
(2) in the tests of sample skewness and kurtosis, $C_1 = 136.8$ reaches a higher value than the quantile $\chi^2_{0.95}(2) = 5.99$, so the sample distribution is not normal;
(3) because the value $x_1 = 9.9134$ ml lies outside the modified external hinges $V_L^* = 9.94$ ml and $V_U^* = 10.02$ ml, $x_{(1)}$ is an outlier and is excluded from the sample. The estimates of location and shape of the new sample are:

$$\bar{x} = 9.9827 \text{ ml}, s^2 = 8.88 \times 10^{-5}, \hat{g}_1 = -0.42, \hat{g}_2 = 2.22.$$

After exclusion of two outliers, $x_{(1)} = 9.9134$ and $x_{(2)} = 9.9656$ ml, the resulting sample is described by

$$\bar{x} = 9.9833 \text{ ml}, s^2 = 8.14 \times 10^{-5}, \hat{g}_1 = -0.44, \hat{g}_2 = 2.27.$$

Classical and robust statistics of the original sample ($n = 32$) are given in Table 3.16. The classical arithmetic mean is also calculated for the sample after elimination of one ($n = 31$) or two ($n = 30$) outliers.

The robust characteristics, $\tilde{x}_{0.5}$, $\bar{x}(10)$ and $\hat{\mu}_M$ are near the value $\bar{x}$ for the reduced sample without the two outliers. Also, the robust confidence interval reaches the same values as the classical confidence interval for the sample without outliers.

The diagnostics G2–G5 of exploratory data analysis exhibit two lowest values which can be understood as the outliers in the sample (Fig. 3.19). The quantile plot G1 shows that the assumption of normality is not fulfilled either when using classical $\bar{x}$ and s^2 or robust (median) characteristics in the G1 plot. The sample distribution is rather skewed to lower values.

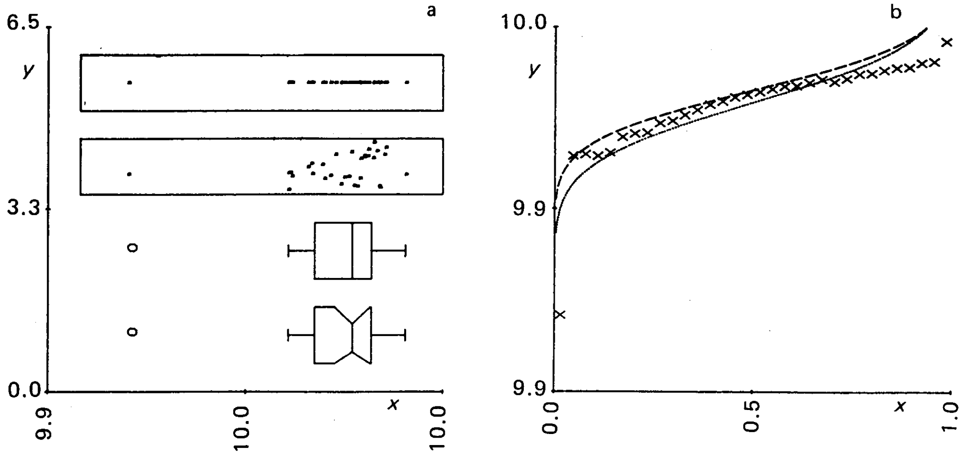


Fig. 3.19—(*a*) Exploratory data analysis by G2–G5 plots and (*b*) G1 plot.

Table 3.15—Point and interval estimates of location

Parameter	Estimate $\hat{\mu}$ [ml]	Estimate $\hat{\sigma}$ [ml]	95% confidence interval L_1	L_2
mean $\bar{x}$ ($n = 32$)	9.9810	6.73×10^{-6}	9.975	9.986
mean $\bar{x}$ ($n = 31$)	9.9827	2.86×10^{-6}	9.979	9.986
mean $\bar{x}$ ($n = 30$)	9.9833	2.71×10^{-6}	9.980	9.990
median $\tilde{x}_{0.5}$	9.9844	6.62×10^{-6}	9.979	9.987
trimmed mean $\bar{x}$ (10)	9.9826	3.55×10^{-6}	9.979	9.987
"biweight" mean $\hat{\mu}_M$	9.9831	3.15×10^{-5}	9.980	9.987

Conclusion: Assumption of normality is not correct because of the presence of outliers. The use of robust estimates is equivalent to excluding outliers from sample. Excluding outliers decreases the relative error of the pipette volume from 0.026% for the original data ($n = 32$) to 0.016% for reduced data ($n = 30$).

Problem 3.22 *Calibration of a pipette, with small sample size*
The pipette of a volume 25 ml was calibrated by weighing the water delivered, and 7 measurements were obtained. Determine the point and interval estimates of the real volume of the pipette.
Data: $n = 7$, V[ml]:
24.96439, 24.97758, 24.96809, 24.97409, 24.96880, 24.94759, 24.97119.
Program: Chemstat: Basic statistics: Assumptions testing, one sample analysis, exploratory continuous.

Solution: The point estimates of location, spread and shape are

$$\bar{x} = 24.9670 \text{ ml}, s^2 = 9.447 \times 10^{-5}, \hat{g}_1 = -1.25, \hat{g}_2 = 3.64.$$

Examination of the data leads to the following conclusions:

(1) the test criterion $t_n = 0.61$ shows that the sample elements are independent;
(2) the test criterion $C_1 = 2.64$ is lower than the quantile $\chi^2_{0.95}(2) = 5.99$, so the sample has a normal distribution;
(3) outside the modified interval hinges $V^*_L = 24.96$ ml and $V^*_U = 24.98$ ml is an outlier value, 24.94759 (Fig. 3.20).

The 95% confidence interval of the mean is

$$24.958 \leq \mu \leq 24.976$$

When the robust "biweight" estimates are calculated, nearly the same values are obtained

$$24.960 \leq \mu \leq 24.977.$$

There are bigger differences between the classical and robust estimates of the values of the variance of the mean,

$$s^2(\bar{x}) = 1.349 \times 10^{-5} \quad \text{and} \quad D(\hat{\mu}_M) = 1.307 \times 10^{-5}.$$

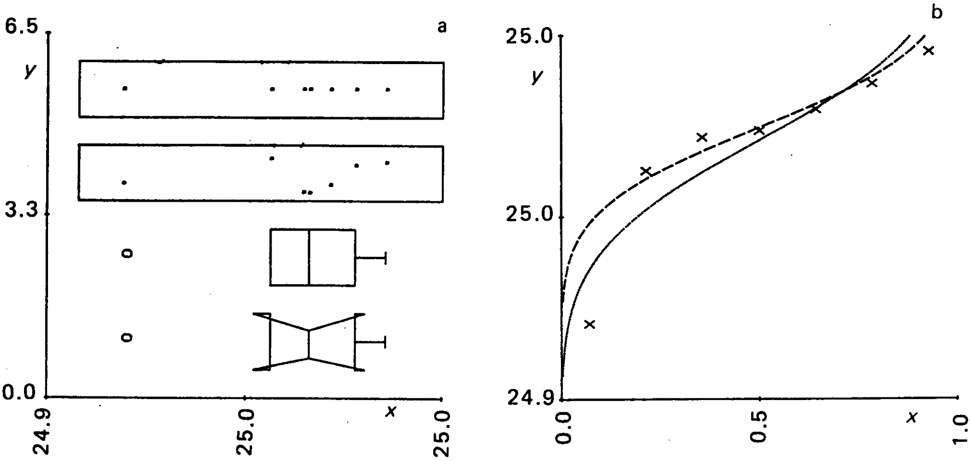


Fig. 3.20—(*a*) Exploratory data analysis by G2–G5 plots and G1 (*b*) plot.

Because of the small sample size the technique of Section 3.4.4 is used. The pivot depth $H_L = 2$, the lower pivot is $x_{(2)} = 24.96439$ and the upper pivot is $x_{(6)} = 24.97409$. From the pivot halfsum $P_L = 24.9692$ ml (3.114), the pivot range $R_L = 0.0097$ ml (3.115) and from Table 3.10 the quantile $t_{L,0.975}(7) = 0.72$, the 95% confidence interval of the mean can be calculated,

$$24.9622 \leq \mu \leq 24.9762.$$

Conclusion; Because the sample is small, the pivot technique is suitable. Then, the outlier has only a small influence on the estimate of the mean. Since the EDA indicates presence of one outlier, the robust methods are preferred.

Problem 3.23 *Examination of* $[Na^+]$, $[Ca^{2+}]$, $[Mg^{2+}]$ *and urea in blood*
A random sample of 32 dairy cows, taken from a herd of 280, was tested for the sodium(I), calcium(II), magnesium(II) and urea content. Find whether the contents of elements correspond to the relevant norms, i.e. $134.5 \le [Na^+] \le 150$ mM, $2.24 \le [Ca^{2+}] \le 3.0$ mM, $0.77 \le [Mg^{2+}] \le 1.07$ mM, and $2.5 \le$ urea ≤ 5.1 mM at the significance level $\alpha = 0.05$.
Data: (*a*) Sample $[Na^+]$, mM.

138.5, 139.8, 145.7, 134.9, 138.2, 131.9, 142.9, 141.7,
132.2, 135.5, 134.7, 132.3, 137.2, 136.7, 133.6, 136.9,
139.8, 140.8, 136.6, 140.4, 132.7, 136.4, 139.0, 142.3,
142.0, 132.1, 139.0, 141.8, 135.8, 139.0, 133.4, 138.8.

(*b*) Sample $[Ca^{2+}]$, mM.

2.50, 2.56, 2.69, 2.47, 2.49, 2.56, 2.47, 2.54, 2.51,
2.39, 2.43, 2.48, 2.41, 2.42, 2.35, 2.37, 2.50, 2.46,
2.35; 2.45, 2.52, 2.42, 2.45, 2.46, 2.49, 2.38, 2.46,
2.51, 2.37, 2.37, 2.49, 2.41.

(*c*) Sample $[Mg^{2+}]$, mM.

0.715, 0.791, 0.810, 0.754, 0.647, 0.767, 0.775, 0.970,
0.878, 0.814, 0.600, 0.684, 0.718, 0.646, 0.786, 0.982,
0.776, 0.716, 0.810, 0.807, 0.841, 0.783, 0.788, 0.665,
0.869, 0.759, 0.834, 0.710, 0.982, 0.719, 0.749, 0.707,

(*d*) Sample urea, mM.

5.76, 4.03, 3.66, 4.85, 5.54, 5.33, 4.46, 6.36, 5.64,
3.14, 4.41, 4.99, 4.29, 4.65, 5.57, 6.40, 5.46, 5.08,
3.96, 3.99, 5.45, 4.99, 7.74, 5.34, 5.11, 4.48, 3.74,
5.29, 5.86, 7.66, 5.69, 7.16.

Program: Chemstat: Basic Statistics: One sample analysis.
Solution: The 95% confidence intervals of the expected values of the arithmetic mean $E(\bar{x})$, of the median $E(\tilde{x}_{0.5})$, and the 40% trimmed mean $E(\bar{x}(0.4))$ are as follows:

(a) *Sample* Na^+: $136.27 \le E(\bar{x}) \le 138.89$, $135.62 \le E(\tilde{x}_{0.5}) \le 139.78$ and $135.74 \le E(\bar{x}(0.4)) \le 139.71$.
(b) *Sample* Ca^{2+}: $2.434 \le E(\bar{x}) \le 2.487$, $2.429 \le E(\tilde{x}_{0.5}) \le 2.491$ and $2.438 \le E(\bar{x}(0.4)) \le 2.486$.
(c) *Sample* Mg^{2+}: $0.743 \le E(\bar{x}) \le 0.810$, $0.741 \le E(\tilde{x}_{0.5}) \le 0.810$ and $0.746 \le E(\bar{x}(0.4)) \le 0.802$.
(d) *Sample* *urea*: $4.796 \le E(\bar{x}) \le 5.584$, $4.765 \le E(\tilde{x}_{0.5}) \le 5.635$ and $4.816 \le E(\bar{x}(0.4)) \le 5.568$.

Conclusion: Robust and classical characteristics lead to the same conclusion that the estimates of content $[Na^+]$, $[Ca^{2+}]$, $[Mg^{2+}]$ and urea correspond to the relevant norms.

Problem 3.24 *Determination of nicotine by GC*
The nicotine content in blood can be determined by gas chromatography down to concentrations of 1 ng/ml. Test the accuracy and precision of the determination of an artificial sample containing 10 ng/ml and the sample with 50 ng/ml. Does the reliability of the determination depend on the concentration?
Data: Sample (A): $\mu = 10$ ng/ml, $n = 12$, $\alpha = 0.05$:
8.40, 9.59, 9.38, 9.10, 10.78, 11.41, 9.94, 10.08, 12.11, 9.10, 9.59, 10.36.
Sample (B): $\mu = 50$ ng/ml, $n = 10$, $\alpha = 0.05$:
47.5, 48.4, 48.8, 48.4, 46.8, 46.2, 48.6, 50.6, 45.5, 46.1.
Program: Chemstat: Basic Statistics: One sample analysis.

Since the samples contain small numbers of measurements, we use the pivot technique.

For sample (A), $H_L = 3$, $P_L = 9.94$, $R_L = 1.68$, and the 95% confidence interval of the mean is $9.13 \le \mu \le 10.75$. This interval contains the value 10, so the gas chromatography technique can be used for determination of a concentration of 10 ng/ml of nicotine in blood.

For sample (B), $H_L = 3$, $P_L = 47.4$, $R_L = 2.4$ and the 95% confidence interval of the mean is $45.79 \le \mu \le 49.00$. This does not contain the true value 50, so the gas chromatography technique for this concentration of nicotine in blood has a systematic negative error.
Conclusion: Sample (A) is determined accurately, but sample (B) is determined with a systematic error.

Problem 3.25 *Comparison of two methods for determination of P_2O_5 in fertilizer*
The concentration of P_2O_5 in artificial fertilizer was determined by two different methods, with use of (A) citrate, and (B) sulphuric acid. Do the results of the two methods come from one common population?

Data: Content of P_2O_5 [%]:

A:	16.3,	15.5,	16.7,	16.0,	13.7,	11.0,	12.5,
	13.4,	14.4,	14.7,	16.9,	15.7,	13.5,	14.0.
B:	16.5,	15.9,	16.6,	15.8,	13.3,	11.2,	12.4,
	13.6,	14.9,	14.6,	16.8,	16.2,	13.8,	14.3.

Program: Chemstat: Basic Statistics: One sample analysis exploratory continuous.
Solution: To test the distributions of the two samples, exploratory data analysis is used. Figure 3.21 shows on the Q−Q plot G14 the linear dependence (the straight line has slope 0.995, intercept −0.041 and $r_{xy} = 0.9961$). Because the slope of the Q−Q line is near to one and the intercept to zero, the two samples come from the same population.
Conclusion: The two methods give the same results.

Problem 3.26 *Determination of glucose in blood*
The concentration of glucose in blood was determined for 20 persons. Estimate a convenient parameter of location and test whether it lies in the interval of medical norm $\langle 2.95, 3.90 \rangle$.

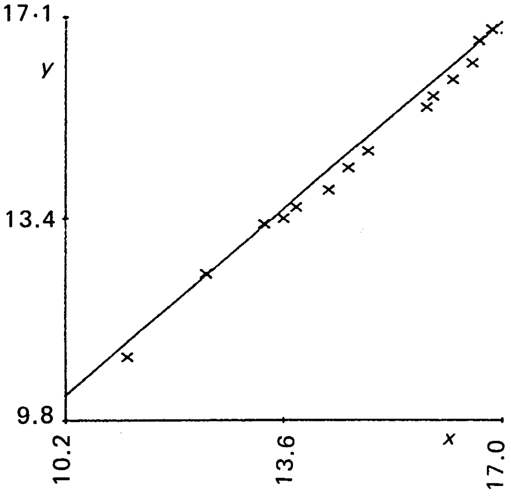


Fig. 3.21—The empirical quantile–quantile plot G14.

Data: the concentration of glucose, mM, $n = 20$:

1.53, 2.94, 3.38, 3.34, 3.45, 3.51, 3.73, 3.33,
3.62, 3.45, 4.24, 2.85, 3.53, 3.97, 3.59, 3.55,
3.68, 3.98, 3.57, 5.53.

Program: Chemstat: Basic Statistics: One sample analysis.

Solution: Because the sample of 200 people may include unhealthy persons (i.e. outliers in a sample), in addition to the arithmetic mean $\bar{x}$ and its variance $D(\bar{x})$, the robust median $\tilde{x}_{0.5}$ and its variance $D(\tilde{x}_{0.5})$ are also used:

$$\bar{x} = 3.539 \text{ m}M,\ D(\bar{x}) = 0.0260,\ \tilde{x}_{0.5} = 3.540 \text{ m}M,\ D(\tilde{x}_{0.5}) = 0.0112.$$

The presence of two outliers has little influence on the parameters of location but significantly changes the variance of the median in comparison to the variance of mean. Therefore, the 95% confidence intervals of the mean and of the median differ significantly,

$$3.20 \le E(\bar{x}) \le 3.88,\ 3.44 \le E(\tilde{x}_{0.5}) \le 3.63.$$

In both cases the mean estimate lies in the requested interval.

Conclusion: The average concentration of glucose in the sample corresponds to the medical norm $\langle 2.95; 3.90 \rangle$.

Problem 3.27 *Examination of the purity of a commercial chemical*

Ammonium phosphate is declared by the manufacturer to contain at least 97% of pure ammonium phosphate. Test whether the chemical really reaches the declared purity.

Data: the content of ammonium phosphate, %, $n = 18$.

99.7, 97.2, 97.9, 97.8, 98.2, 97.4, 97.3, 98.0,
97.9, 98.0, 98.1, 98.4, 98.7, 97.3, 97.9, 96.7, 97.0, 98.1.

Program: Chemstat: Basic Statistics: One sample analysis.

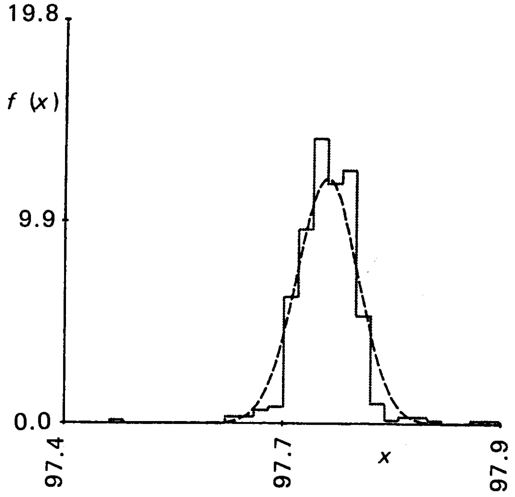


Fig. 3.22—Histogram of the content of ammonium phosphate in a commercial reagent.

Solution: If the highest and lowest values are excluded from the data, calculation of the arithmetic mean corresponds to the 5% trimmed mean $\bar{x}(0.05)$. To calculate the mean value and its variance, the Bootstrap method with 400 simulations was used: $\bar{x}_B = 97.7\%$, $\sigma_B^2 = 0.1102$, $B = 400$ and the 95% confidence interval is $97.1\% \leq E(\bar{x}_B) \leq 98.35\%$. The histogram in Fig. 3.22 shows that the distribution has longer tails and seems to be bimodal.
Conclusion: Because the 95% confidence interval does not cover the value 97%, the ammonium phosphate could not have exactly 97% parity. However, it is obvious that the purity is better than the specified value.

Problem 3.28 *Upper limit of the confidence interval of the average content of fluorine in fertilizer*
Fluorine is an undesirable impurity in phosphate fertilizers. Determine the upper limit of the 95% confidence interval of the average content of fluorine, on the basis of 20 samples of phosphate fertilizer.
Data: The fluorine content, %, $n = 20$.

0.16, 0.16, 0.15, 0.13, 0.18, 0.19, 0.13, 0.19,
0.18, 0.14, 0.29, 0.14, 0.12, 0.10, 0.16, 0.13,
0.16, 0.16, 0.13, 0.14.

Program: Chemstat: Basic Statistics: One sample analysis.
Solution: Because of heterogeneity of the fertilizer, outliers should be expected. The confidence intervals of the mean $E(\bar{x})$ and median $E(\tilde{x}_{0.5})$, and also the robust "biweight" estimate $E(\hat{\mu}_M)$ were evaluated:

$$0.139 \leq E(\bar{x}) \leq 0.174, 0.134 \leq E(\tilde{x}_{0.5}) \leq 0.176$$

and

$$0.138 \leq E(\hat{\mu}_M) \leq 0.163.$$

Conclusion: Because of one outlier (0.29) the confidence interval of the arithmetic mean is too broad and the arithmetic mean is biased to higher values. The upper limit of the confidence interval of average fluorine content is better expressed by the robust estimate, i.e. 0.16%.

Problem 3.29 *Comparison of concentrations of folic acid in two samples*
Folic acid can be determined spectrophotometrically from its reaction with 1,2-naphthoquinone-4-sulphonic acid. Ten determinations were made for each of two tablets with a declared content of 5 mg. Test whether the concentration of folic acid in the two tablets is the same.
Data: folic acid found, mg, $n = 10$
Tablet A: 5.45, 5.15, 7.71, 5.55, 4.75, 5.32, 5.53, 5.09, 5.70, 4.42.
Tablet B: 4.98, 4.84, 4.77, 4.91 4.84, 4.98 4.91, 5.21, 4.67, 5.21.
Program: Chemstat: Basic Statistics: Two sample testing.
Solution: The moment characteristics of tablet (A) and tablet (B) (in the brackets) are:

$$\bar{x} = 5.467\ (4.932)\ \text{mg},\ s^2 = 0.775\ (0.030),\ \hat{g}_1 = 1.665\ (0.432),$$
$$\hat{g}_2 = 2.51\ (-0.63).$$

The determination from tablet (A) contains one outlier and the corresponding distribution is skewed to higher values. Table 3.16 shows the results of tests of the variance for both tablets, H_0: $\sigma_A^2 = \sigma_B^2$ *vs.* H_A: $\sigma_A^2 \neq \sigma_B^2$.

Table 3.16—Results of three tests of variances, H_0: $\sigma_A^2 = \sigma_B^2$ *vs.* H_A: $\sigma_A^2 \neq \sigma_B^2$, $\alpha = 0.05$

F-test used	Test criterion	Quantile	Conclusion
Fisher–Snedecor	25.63	4.026	Reject H_0
With correction of degrees of freedom	25.63	647.79	Accept H_0
Jack-knife test	6.452	4.560	Reject H_0

The conclusion of the *F*-test with correction of degrees of freedom is influenced by the skewed distribution but not by outliers.

The test for whether the means for the two tablets are in agreement, H_0: $\mu_A = \mu_B$ *vs.* H_A: $\mu_A \neq \mu_B$, is shown in Table 3.17.

The classical t-test suggests that there is the same content of folic acid in both tablets, but the robust test indicates that the means do differ significantly.
Conclusion: When one outlier (tablet (A), 7.71) is not excluded from data, both classical tests, the Fisher–Snedecor F-test and the Student t-test give incorrect results. The robust test gives true conclusions because the influence of outliers in data is eliminated.

Table 3.17—Comparison of the means: H_0: $\mu_A = \mu_B$ *vs.* H_A: $\mu_A \neq \mu_B$, $\alpha = 0.05$

Test used	Test criterion	Quantile	Conclusion
Student t-test for $\sigma_A^2 = \sigma_B^2$	1.886	2.100	Accept H_0
Student t-test for $\sigma_A^2 \neq \sigma_B^2$	1.886	2.201	Accept H_0
With allowance for skewness, t_3'	2.557	2.201	Reject H_0
Robust t-test for $\sigma_A^2 = \sigma_B^2$	2.585	2.121	Reject H_0
Robust t-test for $\sigma_A^2 \neq \sigma_B^2$	2.508	2.201	Reject H_0

Problem 3.30 *Pair test for validation of new method*
For analytical determination of dinitrocresol in herbicides a polarographic method (P) is used, but titration (T) is, however, faster and cheaper. Both methods were used and compared. Test whether the titration method gives the same results as the polarographic one.
Data: the content, %, $n = 8$
P: 18.60, 27.60, 27.50, 25.00, 24.50, 26.80, 29.50, 26.50
T: 18.58, 27.37, 27.70, 24.64, 24.10, 26.33, 29.33, 26.63
Program: Chemstat: Basic Statistics: Two sample testing.
Solution: The statistical characteristics of location, spread, and shape for methods P and T and for differences between pairs of values

$$d_i = P_i - T_i$$

are given in Table 3.18.

Table 3.18—Statistical characteristics of samples P and T, and their differences

Sample characteristic	Sample P	Sample T	Difference d
$\bar{x}$	25.75	25.59	0.165
s^2	10.79	10.79	0.191
$\hat{g}_1$	−1.302	−1.201	–
$\hat{g}_2$	3.985	3.709	–

All three variants of the F-test indicate agreement between the two sample variances at a significance level of $\alpha = 0.05$. All variants of the two-sample t-test indicate agreement between the two sample means. For the null hypothesis of the pair test, H_0: $d = 0$ *vs.* H_A: $d \neq 0$, the test criterion has the value $t_p = 2.552$, while the quantile of the Student t-test $t_{0.975}(7) = 2.364$. Therefore, at the significance level $\alpha = 0.05$ the difference between the pair values is significant.

Conclusion: The variability caused by the different levels of dinitrocresol content overlaps with the variability of the two methods, P and T, and therefore the simple t-tests show agreement between the sample means. Use of the differences between the polarographic and titration measurements allows the variability from the different levels of dinitrocresol to be eliminated. It can then be shown that the two methods differ in results.

Problem 3.31 *Agreement between two analytical methods*

The iodine number of soya bean oil was determined by the Hanuš method (H) and by the Wijss method (W). Test whether the two methods yield the same results.

Data: $n = 8$

H: 139.90, 139.80, 138.90, 136.40, 139.40, 140.90, 139.20, 139.40.

W: 139.40, 139.90, 140.20, 140.30, 140.60, 140.90, 140.10, 140.30.

Program: Chemstat: Basic Statistics: Two sample testing.

Solution: The statistical characteristics of location, spread, and shape for sample H and sample W (in brackets) are:

$$\bar{x} = 139.24\ (140.21),\ s^2 = 1.677\ (0.201),\ \hat{g}_1 = -1.25\ (-0.31),$$

$$\hat{g}_2 = 1.22\ (-0.21).$$

The differences in variances and skewness prove that there is an outlier with low value in sample H. The test for equality of sample variances is reported in Table 3.19.

Table 3.19—Tests of sample variances, H_0: $\sigma^2_H = \sigma^2_W$ *vs.* H_A: $\sigma^2_H \neq \sigma^2_W$, $\alpha = 0.05$

Test used	Test criterion	Quantile	Conclusion
Fisher–Snedecor	8.33	4.995	Reject H_0
With correction for degrees of freedom	8.33	39.00	Accept H_0
Jack-knife test	2.83	4.857	Accept H_0

To examine the agreement between the sample means, the null hypothesis H_0: $\mu_H = \mu_W$ is tested against the alternative H_A: $\mu_A \neq \mu_W$ at $\alpha = 0.05$; the results are shown in Table 3.20.

Conclusion: At the significance level $\alpha = 0.05$ the classical test leads to the opposite conclusion from the robust one. The robust test confirms that the differences between the methods are not negligible, although the variances differ only insignificantly. Tests modified for non-zero skewness give the same conclusion.

Table 3.20—Results of testing for equality of means H_0: $\mu_H = \mu_W$ *vs.* H_A: $\mu_H \neq \mu_W$, $\alpha = 0.05$

Test used	Test criterion	Quantile	Conclusion
Student t-test for $\sigma_H^2 = \sigma_W^2$	2.012	2.145	Accept H_0
Student t-test for $\sigma_H^2 \neq \sigma_W^2$	2.012	2.228	Accept H_0
Test modified for skewness, t'_3	2.573	2.228	Reject H_0
Robust t-test for $\sigma_H^2 = \sigma_W^2$	3.533	2.179	Reject H_0
Robust t-test for $\sigma_H^2 \neq \sigma_W^2$	3.394	2.228	Reject H_0

Problem 3.32 *Difference between gravimetric and titrimetric determinations of P_2O_5 in bone*

For the determination of P_2O_5 in calcinated bone meal, gravimetric (G) and titration (T) techniques were used. Find out whether the methods are significantly different.

Data: amount of P_2O_5, mg, $n = 15$

Sample G: 40.24, 40.30, 40.15, 40.20, 40.50, 40.40, 40.12,
40.12, 39.88, 40.23, 40.24, 40.12, 40.17, 40.11,
40.26.

Sample T: 39.90, 40.22, 39.85, 39.93, 39.70, 40.12, 40.20,
39.62, 40.01, 39.77, 39.79, 39.98, 40.26, 39.77,
40.01.

Program: Chemstat: Basic Statistics: Two sample testing.

Solution: Statistical characteristics of location, spread, and shape for method (G) and method (T) (in brackets) are:

$$\bar{x} = 39.94\ (40.203)\ \text{mg},\ s^2 = 0.039\ (0.020),\ \hat{g}_1 = 0.146\ (-0.027),$$

$$\hat{g}_2 = -1.05\ (0.90).$$

The results of the test of equality of variance are given in Table 3.21.

Table 3.21—Results of three tests of variances, H_0: $\sigma_G^2 = \sigma_T^2$ *vs.* H_A: $\sigma_G^2 \neq \sigma_T^2$, $\alpha = 0.05$

F-test used	Test criterion	Quantile $\alpha/2 = 0.025$	Conclusion
Fisher–Snedecor	1.932	2.989	Accept H_0
With correction of degrees of freedom	1.932	2.673	Accept H_0
Jack-knife test	0.743	4.221	Accept H_0

Results of the tests for equality of the means of the two methods are given in Table 3.22.

Table 3.22—Results of the tests for equality of means, H_0: $\mu_G = \mu_T$ *vs.* H_A: $\mu_G \neq \mu_T$, $\alpha = 0.05$

Test used	Test criterion	Quantile $\alpha/2 = 0.025$	Conclusion
Student *t*-test for $\sigma_G^2 = \sigma_T^2$	4.164	2.049	Reject H_0
Student *t*-test for $\sigma_G^2 \neq \sigma_T^2$	4.164	2.052	Reject H_0
Modification for skewness, t_3'	4.036	2.052	Reject H_0
Robust *t*-test for $\sigma_G^2 = \sigma_T^2$	4.295	2.056	Reject H_0
Robust *t*-test for $\sigma_G^2 \neq \sigma_T^2$	4.215	2.086	Reject H_0

Conclusion: The two methods should not be considered to give the same results, regardless of the equality or non-equality of the variances.

3.8 REFERENCES

[1] D. Hensgaard, *Commun. Statist*, 1979, **B8**, 359.
[2] J. W. Tukey and McLaughlin, *Sankya*, 1963, **125**, 331.
[3] N. L. Johnson and S. Kotz, Distributions in Statistics: *Continuous Univariate Distributions*, Houghton Mifflin, Boston, 1970.
[4] R. V. Hogg, *J. Am. Stat. Assoc.*, 1974, **69**, 909.
[5] Ch. Du Mond and R. V. Lenth, *Technometrics*, 1987, **29**, 211.
[6] N. M. Blackman and R. E. Machol, *IEEE Trans. Inform. Theory*, 1987, **IT-33**, 373.
[7] J. Horn, *J. Am. Stat. Assoc.*, 1983, **78**, 930.
[8] B. Efron, *Can. J. Statist.*, 1981, **9**, 139.
[9] H. O. Posten, H. C. Yeh and D. B. Oven, *Commun. Statist.* 1982, **A11**, 109.
[10] N. A. C. Cressie and H. J. Whitford, *Biom. J.*, 1986, **28**, 131.
[11] K. Yuen and W. J. Dixon, *Biometrika*, 1973, **60**, 369.
[12] D. B. Owen, *Handbook of Statistical Tables*, Addison Wesley, Reading, 1963.

4

Analysis of variance (ANOVA)

The results of observations vary because of changes in the basic factors (both qualitative and quantitative) that control the conditions of the chemical experiment, and also in accidental factors. It is the objective of **analysis of variance (ANOVA)** to investigate the effect of the various factors on the variability of data and to determine which part of the variation in a population is due to systematic reasons (called factors) and which is due to random effects. ANOVA has been defined as a statistical technique for analysing measurements that depend on several kinds of effects operating simultaneously, in order to decide which kind of effects are important and to estimate the effects.

The profusion of instrumental techniques in analytical chemistry is such that often more than two possible techniques have to be compared. The techniques to be examined may be subject to systematic errors. The choice of a technique is called a **controlled factor**. Moreover, the results of the analytical determinations are subject to **random errors**. The analysis of variance compares both causes of error, with the purpose of deciding whether or not the controlled factor has a significant effect.

In a chemical laboratory the analysis of variance often serves

(*a*) to distinguish between sources of variability between laboratories, between samples and between replicates,

(*b*) to investigate the influence of human factors instrument factors, methodology, concentration or time on the results of chemical analysis.

A survey of ANOVA techniques may be found in the literature [1–5]. We limit ourselves to techniques suitable for evaluation of chemical data.

4.1 OBJECTIVES OF ANALYSIS OF VARIANCE

Let us consider an example from the chemical laboratory. It is desired to examine the influence of different methods of sample homogenization on the result of a chemical analysis. With the use of three different homogenizers Z_1, Z_2 and Z_3, three different samples were prepared and analysed. The observed values are y_{ij}, $i = 1, 2, 3$ and $j = 1, 2, 3$, where y_{ij} denotes the observation for the ith homogenizer and jth sample. The method of homogenization of sample is called a *qualitative factor*. There are also *quantitative factors*, such as, for example, the mean particle size of the homogenized sample or various physico-chemical parameters.

The individual factor Z exists on different levels Z_1, Z_2, Z_3, which are called *treatments*. The treatments are the main sources of variability and may be also be of qualitative or quantitative nature.

The model of the response in one-factor ANOVA can be written

$$y_{ij} = \mu_i + \varepsilon_{ij} \tag{4.1}$$

where y_{ij} represents the jth observation ($j = 1,2, \ldots, n_i$) for the ith treatment ($i = 1,2 \ldots, k$), μ_i is the true response (mean) at a factor level Z_i, and ε_{ij} is the random error present in the jth observation for the ith treatment. The mean μ_i may be divided into two parts

$$\mu_i = \mu + \alpha_i \tag{4.2}$$

where μ represents a general overall mean and α_i represents the effect of the ith treatment Z_i. The total number of observations is

$$n = n_1 + n_2 + \ldots + n_k.$$

In our example we have $k = 3$ and $n = 9$. We will now test the null hypothesis that there are no differences caused by the method of homogenization, H_0: $\mu_1 = \mu_2 = \mu_3$ which corresponds to the null hypothesis H_0: $\alpha_1 = \alpha_2 = \alpha_3 = 0$. If we investigate the differences between just three methods of homogenization, we have a **fixed-effect model**. If the levels (Z_1, Z_2, Z_3) are random samples from the population of all possible methods of homogenization, we have a **random-effect model**.

The types of effects proposed lead to distinctive model assumptions and associated statistical analysis.

The analysis of variance can be applied in several distinct forms, according to the structure of the process being investigated. The selection of a particular form usually constitutes a major difficulty in the practical application of the analysis of variance.

The choice between a fixed- or random-effect model depends on the *purpose* of the analysis. If we suppose that three homogenizing machines make three different levels of particle size, then instead of considering homogenizing machine to be the factor, we use the mean particle size.

(*a*) We speak about a *fixed-effects model* when three homogenizing machines correspond to three different milling finenesses, and we examine whether these homogenizing machines affect significantly the results of the chemical analysis.

(*b*) We speak about a *random-effects model* when we test whether the mean particle size has an influence on the results of the chemical analysis. From a population of different particle sizes we randomly select three.

The criteria for choosing between the fixed- and random-effects model are:

(*a*) Factors with fixed effects are usually a type of chemical treatment, a type of instrument, an analytical method, a type of raw material, etc.

(*b*) Factors with random effects are laboratories, days, people, animals, etc.

One-way ANOVA deals with the influence of a single factor on a single response variable. When that one factor has fixed effects, one-way ANOVA (*"fixed-effects one-way ANOVA"*) involves comparison of several (two or more) population means. Each population corresponds to one treatment (factor level). Often, the influence of more factors is examined and then we speak about a multi-way ANOVA. In this book, we concentrate on one- and two-way ANOVA only.

An example of two factors could be the milling of samples using three grinding mills by two chemists. The second factor is here "chemist" with levels L_1 and L_2. The result of chemical analysis y_{ijk}, $i = 1, 2, 3; j = 1, 2;$ and $k = 1, 2, 3$ means the result for the ith way of milling mode by the jth chemist on the kth sample. The observation is replicated for a given combination of both factors $Z_i L_j$, and the corresponding ANOVA model is expressed by

$$y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \tag{4.3}$$

where μ_{ij} is the true, theoretical mean chemical analysis for the combination of factors $Z_i L_1$ and ε_{ijk} is a random error. The mean μ_{ij} can be written as a sum of effects α_i and β_j of the factors Z_i and L_j, an overall mean μ, and the interaction effect τ_{ij} due to

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \tau_{ij} \tag{4.4}$$

The term τ_{ij} represents the effect of **interaction** of levels Z_i and L_j. It is used when the variability of y_{ijk} cannot be explained by the additive influence of factors. When the effects of both factors are fixed or random, we speak about *models with fixed* or *random effects*. It may happen that, e.g., effects of factor L are random while factor Z has fixed effects. Such models are called *model with mixed effects*. For data treatment it is important that at all factor combinations the same number of replicate measurements is performed.

A combination of levels of individual factors (e.g. $Z_i L_j$) is called a **cell**. When the number of replicate measurement is the same for each, the experiment is termed a **balanced experiment** (or *a balanced plan* of experiments); for different numbers it is called an **unbalanced experiment**.

Treatment of data is easier for balanced experiments. Treatment of data from unbalanced experiments is more complicated and sometimes the ANOVA assumptions (e.g. about normality) are not fulfilled, and this will cause distortion in the ANOVA results.

4.2 ONE-WAY ANOVA

Suppose that some factor A, which we postulate as having some effect on a response variable y, has k levels. We set up an experiment in which n_1 observations are made of the response y at level A_1, n_2 observations at level A_2, and so on, with n_i observations at a given level A_i. The levels A_i are called treatments, and there are k treatments in the experimental design. The total number of observations $n = \Sigma_{i=1}^{k} n_i$. At each level A_i there are n_i observations y_{ij}, $j = 1, \ldots, n_i$. The layout of a one-way ANOVA experiment with different number of replicates for each treatment is shown in Table 4.1.

Table 4.1-Data layout for one-way ANOVA with an unequal number of replications on each treatment

	Factor A levels (treatments)						Overall
i	1	2	...	i	...	k	mean
	A_1	A_2	...	A_i	...	A_k	
Replicate j							
1	y_{11}	y_{21}	...	y_{i1}	...	y_{k1}	
2	y_{12}	y_{22}	...	y_{i2}	...	y_{k2}	
.	.	.	...	.	...	.	
.	.	.	...	.	...	.	
.	.	.	...	.	...	.	
n_1	y_{1n_1}	y_{2n_2}	...	y_{in_i}	...	y_{kn_k}	
Mean	$\hat{\mu}_1$	$\hat{\mu}_2$	...	$\hat{\mu}_i$	...	$\hat{\mu}_k$	$\hat{\mu}$
Sample size	n_1	n_2	...	n_i	...	n_k	n

Let $\hat{\mu}_i$ denote the mean of the ith (partial) sample, that is, the mean of observations at the ith treatment or level

$$\hat{\mu}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \tag{4.5}$$

The overall (or grand mean) $\hat{\mu}$ of the all samples can be defined as

$$\hat{\mu} = \frac{1}{k} \sum_{i=1}^{k} \hat{\mu}_i = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} \tag{4.6}$$

Equations (4.5) *and* (4.6) *determine the estimates of parameters* μ_i in Eq. (4.1) or μ in Eq. (4.2). To determine an estimate of effect α_i the following expression is used:

$$\hat{\alpha}_i = \hat{\mu}_i - \hat{\mu} \tag{4.7}$$

To avoid identifiability problems the parameters α_i are constrained by

$$\sum_{i=1}^{k} n_i \alpha_i = 0 \tag{4.8}$$

When the sample size is the same for each treatment (balanced experiment), this condition simplifies to

$$\sum_{i=1}^{k} \alpha_i = 0 \tag{4.9}$$

The estimator $\hat{\alpha}_i$ corresponds to this constraint. The procedure adopted now depends on whether fixed-effects or random-effects ANOVA is considered.

4.2.1 Fixed-effects one-way ANOVA

4.2.1.1 Assumptions

In a *fixed-effects ANOVA model*, all the factor levels being considered are fixed, i.e. the levels of each factor are the only levels of interest. The "effects" referred to in such a model represent measures of the influence (i.e., the effect) that different levels of the factor have on the observed variable. Such measures are often expressed in the form of differences between *a* at given level and an overall mean. The *i*th level corresponds to the *i*th population from which the sample of size n_i can be selected. The effect of the *i*th population is often measured by the amount that the *i*th population mean differs from an overall mean.

The assumptions needed for fixed-effects one-way ANOVA may be stated simply as follows:

(*a*) Random samples of observations (chemical results, measurements, observations, etc.) are selected from each of *k* fixed populations or groups. For the *i*th level we have sample $y_{ij}, j = 1, \ldots, n_i$.

(*b*) The model of ANOVA defined by Eq. (4.1) is valid.

(*c*) The observations are normally distributed with constant variance σ^2 in the whole population, $y_{ij} = N(\mu_i, \sigma^2)$.

(*d*) Random errors ε_{ij} are mutually independent random variables normally distributed with a mean equal to zero and variance σ^2, $\varepsilon_{ij} = N(0, \sigma^2)$.

In general, classical ANOVA analysis can be applied if none of the assumptions is very badly violated. This is true for more complex ANOVA situations as well as for fixed-effects one-way ANOVA. The term generally used to refer to this property of broad applicability is called *robustness*. We say that a procedure is robust with respect to moderate departures from the basic assumptions.

We must nevertheless be careful to avoid using robustness as an automatic justification for blindly applying the ANOVA model. Certain facts should be borne in mind when the use of ANOVA in a given situation is considered. For example, the normality assumption does not have to be exactly satisfied as long as we are dealing with relatively large samples (e.g., 20 or more observations from each population), although the consequences of large deviations from normality are more severe for random effects than for fixed effects. The assumption of variance homogeneity can also be mildly violated without serious risk, provided that the numbers of observations selected from each population are more or less the same, although, again, the consequences are more severe for random effects.

Violation of the assumption of independence of the observations, however, can lead to very serious errors in both the fixed- and random-effects cases. In general, great care should be taken to ensure that the observations are independent. This concern arises primarily in studies where repeated observations are recorded on the same experimental subjects, since very often the level of response of a subject on one occasion has a decided effect on subsequent responses.

What, then, should be done when one or more of these assumptions are in serious question? One possibility is for the data to be transformed (e.g., by means of a log, square root, or other transformation) so that they more closely satisfy the assumptions.

4.2.1.2 Methodology

The null hypothesis that there is no treatment effect, i.e., the hypothesis of equal population means H_0: $\mu_1 = \mu_2 = \ldots = \mu_k = \mu$ is usually tested first.

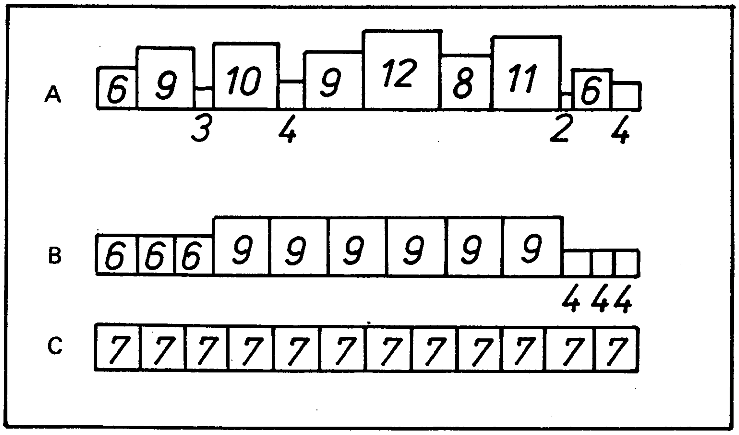


Fig. 4.1—Partitioning the sum of squared deviations into members A, B and C, where A are data, B are treatment means and C is the overall mean: S_C = A − C, S_A = B − C, S_R = A − B.

We begin the analysis by partitioning the sum of squared deviations from the overall $\hat{\mu}$ defined by

$$S_C = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu})^2 \tag{4.10}$$

into two components, one attributable to the identifiable source of variation (factor A), and denoted by S_A; the other representing the variation due to uncontrolled factors and random errors associated with the response measurement, and denoted by S_R. Rearrangement of Eq. (4.10) leads to

$$S_C = \sum_{i=1}^{k} \sum_{j=1}^{n_i} [(y_{ij} - \hat{\mu}_i) + (\hat{\mu}_i - \hat{\mu})]^2 = S_A + S_R \tag{4.11}$$

where

$$S_A = \sum_{i=1}^{k} n_i(\hat{\mu}_i - \hat{\mu})^2 \tag{4.12}$$

represents the variability *between* individual treatments of a given factor A, and

$$S_R = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \hat{\mu}_i)^2 \tag{4.13}$$

represents the variability *within* all treatments. A graphical interpretation of the partitioning of S_C is shown in Fig. 4.1. Results of an ANOVA procedure are usually presented in a so-called ANOVA table (Table 4.2)

Table 4.2-ANOVA table for one-way fixed-effects model

Source of variation	Degrees of freedom	Mean square	Expected mean square
Between treatments S_A	$k - 1$	$S_A/(k - 1)$	$\sigma^2 + \dfrac{\Sigma_{i=1}^{k} n_i \alpha_i^2}{k - 1}$
Residual (within treatments) S_R	$n - k$	$S_R/(n - k)$	σ^2
Totals	$n - 1$	—	—

The last column in Table 4.2 shows the expected mean square. An unbiased estimate of the variance of errors σ^2 is the mean square of residuals defined by

$$\hat{\sigma}^2 = S_R/(n - k) \tag{4.14}$$

The null hypothesis that the treatments are equal, i.e. insignificance of effect; H_0: $\alpha_i = 0$, $i = 1, \ldots, k$, and the alternative hypothesis is H_A: $\alpha_i \neq 0$, $i = 1, \ldots, k$. The test is based on the fact that S_A/σ^2 has the χ^2 distribution with $(k - 1)$ degrees of freedom, and quantity S_R/σ^2 has an independent χ^2-distribution with $(n - k)$ degrees of freedom. Their ratio has the Fisher–Snedecor F-distribution with $(k - 1)$ and $(n - k)$ degrees of freedom. The test statistic is calculated by

$$F_e = \frac{S_A(n - k)}{S_R(k - 1)} \tag{4.15}$$

When the null hypothesis H_0 is valid, the statistic F_e has the F-distribution with $(k - 1)$ and $(n - k)$ degrees of freedom. When F_e is greater than the quantile $F_{1-\alpha}(k - 1, n - k)$, the null hypothesis is rejected, and the effect of factor A is taken as significant.

If $F_e \leq F_{1-\alpha}(k-1,\ n-k)$, the effect of factor A should be considered to be insignificant. Then the total variance σ^2 is related only to the uncontrolled (random) factor and may serve as an estimate of the replication variance.

In interpreting the results of an ANOVA, it is important to bear in mind that a very low value of the variance ratio S_A/S_R may be related to the fact that some important uncontrolled factor was not randomized in the course of the experiment. This may lead to an increased variance within the treatments while leaving the variance between the treatments unchanged, resulting in a reduced variance ratio. In these circumstances, the experimental results will not obey the model defined by Eq. (4.1). If, on the other hand, the inequality $F_e > F_{1-\alpha}(k-1,\ n-k)$ holds, the difference between the two variances is significant, and so is the effect of factor A.

Problem 4.1 *Test of quality of silver nitrate made by various companies*
The bottles containing silver nitrate were manufactured by five different companies. Random samples of silver nitrate taken from the bottles were used for determination of chlorine in organic samples. From each of the five bottles different numbers of random samples are taken to prepare stock solutions of $AgNO_3$:$n_1 = n_3 = 6$; $n_2 = n_5 = 3, n_4 = 4$. Test whether the quality of $AgNO_3$ coming from various chemical companies differs.
Data: The percentages of chlorine in a single organic compound determined by using five stock solutions of silver nitrate are listed in Table 4.3

Table 4.3—Determination of Cl by various stock solutions of $AgNO_3$

	Source of $AgNO_3$				
Replicate	V_1	V_2	V_3	V_4	V_5
1	4.40	4.90	5.55	4.45	5.15
2	4.40	4.95	5.10	5.45	6.25
3	5.20	5.40	5.50	4.65	6.14
4	5.45	—	5.98	4.40	—
5	5.80	—	5.60	—	—
6	5.60	—	5.56	—	—

Program: Chemstat: ANOVA-1: One-way.
Solution: The estimates of the individual means $\hat{\mu}_i$, the overall mean $\hat{\mu}$ and effects $\hat{\alpha}_i$ are:

$$\hat{\mu} = 5.2715;\ \hat{\mu}_1 = 5.1417;\ \hat{\mu}_2 = 5.083;\ \hat{\mu}_3 = 5.548;\ \hat{\mu}_4 = 4.738;$$

$$\hat{\mu}_5 = 5.8467;\ \hat{\alpha}_1 = -0.1298;\ \hat{\alpha}_2 = -0.1885;\ \hat{\alpha}_3 = 0.276;$$

$$\hat{\alpha}_4 = -0.534;\ \hat{\alpha}_5 = 0.575.$$

The sums of squared deviations and variance components are summarized in Table 4.4.

Table 4.4—One-way ANOVA table for quality of silver nitrate

Source of variation	Degrees of freedom	Mean square	F_e
Between companies $S_A = 2.7999$	4	0.6999	3.1
Residual $S_R = 3.8322$	17	0.2254	—
Totals $S_C = 6.632$	21	—	—

For significance level $\alpha = 0.05$ the quantile $F_{0.95}$ (4,17) = 2.96.
Conclusion: Since the experimental value F_e is greater than quantile $F_{0.95}$ (4,17), the null hypothesis H_0: $\alpha_i = 0, i = 1, \ldots, 5$ is rejected, and the quality of $AgNO_3$ coming from different chemical manufacturers significantly differs.

4.2.1.3 Multiple-comparison procedure

Whenever an ANOVA F-test for simultaneous comparison of several population means is found to be statistically significant, it is of interest to determine which *specific* differences there are among the population means. For example, if four means are being compared (fixed-effects case) and the null hypothesis H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$ is rejected, it is usually desirable to determine which subgroups of means are different by considering some more specific hypothesis such as H_{01}: $\mu_1 = \mu_2$; H_{02}: $\mu_2 = \mu_3$; or even H_{04}: $(\mu_1 + \mu_2)/2 = (\mu_3 + \mu_4)/2$, which compares the average effect of populations 1 and 2 with the average effect of populations 3 and 4. Such specific comparisons may have been of interest to the investigator *before* the data were collected, or may arise in completely exploratory studies only *after* the data have been examined. In either event, a seemingly reasonable first approach to making inferences about differences among the population means would be to make several t-tests, and to focus on all the tests found to be significant. The justification for all the $Z = {}_kC_2 = k(k-1)/2$ tests being applied simultaneously comes from the Bonferroni inequality. Special quantiles of the t distribution for level $\alpha/2 \times Z$ must then be used. Testing of differences between population means in ANOVA is called the multiple comparisons technique. In Scheffe's multiple comparison procedure the null hypothesis H_0: $\mu_i = \mu_j$ is rejected for all pairs of (i, j) populations for which

$$|\hat{\mu}_i - \hat{\mu}_j| \geq \sqrt{(k-1) \times \sigma^2 \times F_{1-\alpha}(k-1,\ n-\mathrm{k}) \times [1/n_i + 1/n_j]} \tag{4.16}$$

where n is the total number of observations, k is the number of means considered, n_i and n_j are the sizes of the samples selected from the ith and jth populations (treatments), respectively, $\hat{\sigma}^2$ is the estimate of variance calculated by Eq. (4.14). Equation (4.16) is used for all pairs of indices (i, j). In some cases, only selected linear constraints can be proved.

In general, a *linear contrast* is defined to be any linear function of the population means, say

$$L = \sum_{i=1}^{k} c_i \mu_i \tag{4.17}$$

with known coefficients c_i such that

$$\sum_{i=1}^{k} c_i = 0 \text{ and } \sum_{i=1}^{k} c_i^2 > 0 \tag{4.18a,b}$$

The contrast estimate is defined by

$$\hat{L} = \sum_{i=1}^{k} c_i \hat{\mu}_i \tag{4.19}$$

When all the observations come from a normal distribution $N(\mu_i, \sigma^2)$, the associated null hypothesis H_0: $L = 0$ may be tested against the alternative H_A: $L \neq 0$ by using the test criterion

$$F_L = \frac{\hat{L}^2}{\left[\hat{\sigma}^2 \sum_{i=1}^{k} c_i^2/n_i\right]} \tag{4.20}$$

If the null hypothesis H_0 is valid, the test criterion F_L has a Fisher–Snedecor distribution with 1 and $(n - k)$ degrees of freedom. The null hypothesis is rejected when F_L reaches a value higher than quantile $F_{1-\alpha}(1,n - k)$.

Problem 4.2 *Differences in quality of $AgNO_3$ from two suppliers*

For the data from Problem 5.1 test whether the difference between μ_4 and μ_2 is statistically significant. Is the silver nitrate from supplier No. 4 of better quality than that from supplier No. 2?

Data: Problem 4.1

Program: Chemstat: ANOVA – 1: One-way.

Solution: To test for a difference between μ_4 and μ_2, the linear contrast with coefficients $c_1 = c_3 = c_5 = 0$; $c_2 = 1$, $c_4 = -1$ may be calculated. The estimate of contrast $\hat{L} = \hat{\mu}_2 - \hat{\mu}_4 = 0.345$. From Eq. (4.20), $F_L = 0.119/(0.2254(1/4 + 1/3)) = 0.905$. Because the quantile $F_{0.95}$ (1,17) = 4.451 is greater than F_L, the difference between μ_2 and μ_4 is not statistically significant.

Conclusion: The difference between the quality of silver nitrate from suppliers 4 and 2 is not statistically significant.

4.2.1.4 Regression model

The procedure of analysis of variance is applicable only when the observations are independent, the errors ε_{ij} have the normal distribution $N(0, \sigma^2)$, with constant variance σ^2. Before use of the ANOVA procedure, all the assumptions should be examined. For this, it is advantageous to convert the ANOVA model into a linear regression model and apply regression diagnostics (from Chapter 6, Vol. 2).

Most ANOVA procedures can also be considered in a regression analysis setting; this can be done by defining appropriate dummy variables in a regression model. The ANOVA model, Eq. (4.1), may be expressed as the linear regression model

$$y_{ij} = \mu_1 w_1 + \mu_2 w_2 + \ldots + \mu_k w_k + \varepsilon_{ij} \tag{4.21}$$

where the w_i are dummy variables which take the following values:

$$w_i = \begin{cases} 1 \text{ for effect } i \\ 0 \text{ otherwise} \end{cases}$$

The means $\mu_1, \mu_2, \ldots, \mu_k$ are understood as the regression parameters. If all the assumptions about errors are valid, the parameters estimates $\hat{\mu}_i$ can be calculated by the least-squares method—i.e. by minimizing

$$U(\mu) = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \sum_{i=1}^{k} w_i \mu_i)^2 \tag{4.22}$$

Analytical minimizing of $U(\mu)$ leads to the system of equations:

$$\frac{\delta U(\mu)}{\delta \mu_i} = 0,\ i = 1, \ldots, k \tag{4.23}$$

Since $w_j = 1$ only for $i = j$ the solution of Eq. (4.23) has the simple form

$$\hat{\mu}_i = \frac{\sum_{j=1}^{n_i} y_{ij}}{n_i} \tag{4.24}$$

Chapter 6 (Vol. 2) explains the important role played by the diagonal elements H_{ii} of the projection matrix $\mathbf{H}$ in analysis of residuals and leverage points,

$$\mathbf{H} = \begin{bmatrix} (1/n_i)\mathbf{J}_1\mathbf{J}_1^{\mathrm{T}} & 0 & 0 \\ 0 & (1/n_2)\mathbf{J}_2\mathbf{J}_2^{\mathrm{T}} & 0 \\ 0 & 0 & (1/n_k)\mathbf{J}_k\mathbf{J}_k^{\mathrm{T}} \end{bmatrix}$$

where $\mathbf{J}_i$ is a column unit vector of size $(n_i \times 1)$. Matrix $\mathbf{H}$ consists of blocks of size $(n_i \times n_i)$ with values $1/n_i$.

For the same number of observations (replicates) of each treatment (balanced experiments) all the diagonal elements H_{ii} have the same magnitude. It means that estimates $\hat{\mu}_i$ have constant variance.

For different numbers of observation of each treatment, the variances

$$D(\hat{\mu}_i) = \sigma^2/n_i \tag{4.26}$$

are not constant. Similarly, the variances of the residuals

$$D(\hat{e}_{ij}) = \sigma^2(1 - 1/n_i) \tag{4.27}$$

also are not constant. Residuals $\hat{e}_{ij}$ in ANOVA models are expressed as

$$\hat{e}_{ij} = y_{ij} - \hat{\mu}_i \tag{4.28}$$

For very different numbers of replicates of the treatments (unbalanced experiments) the residuals will have nonconstant variance even in cases when the errors have constant variance.

From the theory of regression models (Chapter 6, Vol. 2) it is evident that for extreme points, the diagonal elements of the projection matrix become larger than $2k/n$. This means that for small sample sizes there is a danger that levels with a particularly small number of observations will have a strong influence on the results of the statistical analysis.

Problem 4.3 *Investigation of the influence of individual suppliers of silver nitrate on ANOVA result*
Examine the influence of the individual suppliers of silver nitrate in Problem 4.1 especially for small sample sizes, and test whether any supplier can be taken as an extreme.
Data: As for Problem 4.1
Solution: For samples from the first and third supplier the diagonal elements of the projection matrix $1/n_i = 1/6 = 0.16$, from the second and the fifth, $1/n_i = 1/3 = 0.33$ and from the fourth $1/n_i = 1/4 = 0.25$. The critical value is $2 \times 5/22 = 0.4545$.
Conclusion: Since all diagonal elements of projection matrix H_{ij}, $i = 1, \ldots, 5$ have values under the critical limit, all samples can be considered as not to be leverages.

4.2.1.5 Checking for data normality

To check the data normality, the rankit plot (Chapter 2) may be used. Examination of standardized residuals is also helpful (see Chapter 6, Vol. 2)

$$\hat{e}_{si} = \frac{\hat{e}_{ij}}{\hat{\sigma}\sqrt{1 - 1/n_i}}$$

where $\hat{e}_{ij}$ are residuals, $\hat{\sigma}$ is the estimate of the standard deviation and n_i is the number of observations for a treatment. The standardized residuals, in a classical analysis of variance, exhibit approximately a normal distribution with zero mean and unit variance $\hat{e}_{si} \approx N(0,\sigma^2)$. If the errors are normally distributed, $\varepsilon_{ij} \approx N(0, \sigma^2)$, the rankit plot of the standardized residuals is linear, with zero intercept and unit slope.

Problem 4.4 *Check of data normality*
Check whether the data from Problem 4.1 have a normal distribution, with the use of a rankit plot G12 for standardized residuals $\hat{e}_{si}$.
Data: Problem 4.1
Program: Chemstat: ANOVA-1: One-way.
Solution: The rankit plot is shown in Fig. 4.2
Conclusion: The rankit plot proves that the data have an approximately normal distribution.

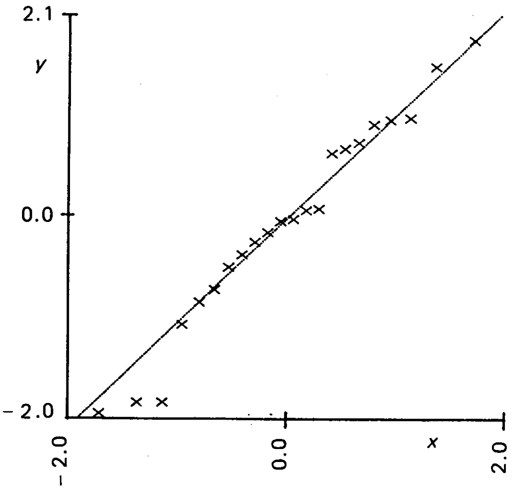


Fig. 4.2—The rankit plot G12 for standardized residuals from Problem 4.4.

When data do not belong to the normal distribution, some transformation (logarithm, square-root, or other functions) can often be applied. In many practical cases, data are skewed to higher values. Then the logarithmic transformation

$$y^* = \ln (2y + C) \tag{4.30}$$

is suitable. The value C is selected such that

(1) the distribution of residuals is symmetrical with kurtosis near to 3,
(2) a rankit plot of standardized residuals is linear.

In ANOVA models the outliers play an important role (as in regression). Outliers may be detected by Jack-knife residuals $\hat{e}_{Jij}$ which are defined by

$$\hat{e}_{Jij} = \hat{e}_{Sij}\sqrt{\frac{n-k-1}{n-k-\hat{e}_S^2}}$$

For normally distributed ^{ij}data, these residuals have approximately the Student distribution with $(n - k - 1)$ degrees of freedom. Roughly, if $\hat{e}_{Jij}^2 > 10$, the given value y_{ij} is taken as an outlier. Other diagnostics for finding outliers, described in Chapter 6, may be applied also.

Problem 4.5 *Detection of outliers in data*
Test whether the value 5.15 in the 1st sample from the fifth supplier in Problem 4.1 is an outlier.
Data: As for Problem 4.1
Solution: The residual is

$$\hat{e}_{51} = y_{51} - \hat{\mu}_5 = 5.15 - 5.8467 = -0.6967$$

and corresponding standardized residual is

$$\hat{e}_{S51} = -0.6967/\sqrt{0.2254(1 - 1/3)} = -1.7973.$$

The Jack-knife residual is

$$\hat{e}_{J51} = -1.7973\sqrt{(22 - 6)/(22 - 5 - 1.797^2)} = -1.932.$$

Conclusion: Because $\hat{e}^2_{J51} = 3.74$ is smaller than 10, the value y_{51} is not an outlier.

4.2.1.6 Checking for homoscedasticity

The assumption of homoscedasticity (i.e. constant variance) may be tested by use of the same diagnostics as for the linear regression model. For non-constant numbers of observations on treatments, the heteroscedasticity of classical residuals (Eq. (4.27) should be considered. For a sufficient number of observations on a treatment, in addition to the mean $\hat{\mu}_i$, a treatment variance s_i^2 can also be estimated. A test of homoscedasticity may be carried out on the basis of a plot of s_i *vs.* $\hat{\mu}_i$. If a random pattern of points results, the homoscedasticity in different treatments is accepted. If s_i is related to $\hat{\mu}_i$ by a monotonic function $s = f_2(\hat{\mu})$, the data might be transformed to stabilize the variance. A suitable transformation can be determined from

$$g(y) = \int_y \frac{d\mu}{f_2(\mu)} \tag{4.32}$$

Problem 4.6 *Determination of the transformation for stabilizing a variance*
It was found that a plot of s_i *vs.* μ_i is linear. Determine a convenient transformation for stabilizing the variance.
Solution: By using Eq. (4.32) for $f_2(\mu) \approx a \times \mu$, the differential equation will be

$$g(y) = \int_y \frac{d\mu}{\mu}$$

with solution $g(y) = a \times \ln y$. Thus, the optimum transformation will be a logarithmic one, $y^*_{ij} = \ln(y_{ij})$.
Conclusion: When the dependence of s_i on $\hat{\mu}_i$ is monotonic, a convenient transformation can easily be found. Such a transformation improves the normality of the data too.

4.2.2. The random-effects model for one-way ANOVA

A *random-effect factor* is a factor which has levels that may be regarded as a sample from some large population of levels, whereas a fixed-effect factor is one which has levels that are the only levels of interest. The distinction is important in any ANOVA, since different tests of significance are required for different configurations of random- and fixed-effect factors. For now, it is perhaps useful to give some examples of random and fixed effects factors:

(a) "Subjects" is usually considered to be a random-effects factor, since we ordinarily wish to infer from the subjects used, to a population of potential subjects.

(b) "Observers" is a random-effects factor often considered in the examination of the effect of different observers on the response variable of interest.

(c) "Days", "weeks", and so on are usually considered as random factors in investigation of the effect of time on a response variable observed for different time periods. We usually use many levels for such temporal factors to represent a large number of time periods.

(d) "Sex" is always a fixed-effects factor, since its two levels include all possible levels of interest.

(e) "Location" (e.g., cities, plants, laboratories) may be fixed- or random-effects factors, depending on whether only specific sites are of interest or whether a larger geographical universe is to be considered.

(f) "Treatments", "drugs", "chemicals", "tests" and so on, are usually considered as fixed factors, but they may be considered random if their levels are representative of a much large group of possible levels.

In the random-effects model, the effect α_i is considered as a random variable. For the random-effects model, it is assumed that

(a) α_i are mutually independent random variables with normal distribution $N(0, \sigma_A^2)$;

(b) ε_{ij} are mutually independent random variables having normal distribution $N(0, \sigma_e^2)$;

(c) random variables α_i are independent of random variables ε_{ij}. The purpose of ANOVA is to test the differences of theoretical effects $\alpha_i - \alpha_j$ or the variances σ_A^2 and σ_e^2. In some cases, the overall mean μ or means $\mu_i = \mu + \alpha_i$ for individual treatments are estimated.

For estimates of variances α_A^2 and σ_e^2 a residual sum of squares S_R and a sum of squares S_A explained by effects are used. Both quantities are calculated as for the fixed-effects model (Table 4.2).

The variances σ_e^2 and σ_A^2 are calculated from

$$\hat{\sigma}_e^2 = \frac{S_R}{n - k} \tag{4.33}$$

and

$$\hat{\sigma}_A^2 = \frac{n(k-1)[S_A/(k-1) - S_R/(n-k)]}{n^2 - \sum_{i=1}^{k} n_i^2} \tag{4.34}$$

The estimate $\hat{\sigma}_e^2$ must be non-negative, i.e. $\hat{\sigma}_A^2 = \max(0, \hat{\sigma}_A^2)$. Estimates $\hat{\mu}$, $\hat{\sigma}_e^2$ and $\hat{\sigma}_A^2$ have some useful statistical properties. Provided that the assumptions about errors ε_{ij} and effects α_i are valid, they have from all possible unbiased estimates the minimum variance. The variances $\hat{\sigma}_e^2$ and $\Sigma\hat{\sigma}_A^2$ may be estimated with the use of a maximum likelihood or some other method[5].

The requirement of a zero mean for all α_i is similar in philosophy to the requirement that $\Sigma_{i=1}^k \alpha_i = 0$ for the fixed-effects model. When the ANOVA random model applies, we assume that the average effect of treatments α_i is 0 over the entire population of all treatments α. That is, we assume that $\mu_i = 0$. Because we have required the treatment effects to average out to 0 over the entire population of possible effects, there is only one way to assess whether there are any significant treatment effects at all, and this involves consideration of σ_A^2. If there is no variability (i.e., $\sigma_A^2 = 0$), all treatment effects must be 0. If there is variability (i.e., $\sigma_A^2 > 0$), there are some non-zero effects in the population of treatment effects. Thus, our null hypothesis of no treatment effects should be stated as H_0: $\sigma_A^2 = 0$. This hypothesis is therefore analogous to the null hypothesis used in the fixed-effects case, although it happens to be stated in terms of a population variance rather than in terms of population means. The F-test criterion is stated for the random-effects model in exactly the same way as that used for the fixed-effects model,

$$F_e = \frac{S_A}{S_R} \times \frac{(n-k)}{(k-1)} \tag{4.35}$$

If the null hypothesis is valid, this F_e statistic has the Fisher–Snedecor distribution with $(k-1)$ and $(n-k)$ degrees of freedom.

When the means μ_i are estimated, the procedure is the same as in the fixed-effects model [5]. For an estimate of an overall mean μ in balanced experiments, the arithmetic mean $\hat{\mu}$ is used. The variance of $\hat{\mu}$ can be estimated from the equation

$$\hat{\sigma}_{\hat{\mu}}^2 = \frac{S_A}{(k-1)k \times n^*} \tag{4.36}$$

where n^* is number of observations, which should be the same for all treatments. For a significance test of the overall mean μ, the test statistic $t = \hat{\mu}/\hat{\sigma}_{\hat{\mu}}$ may be used. This statistic has, for a valid null hypothesis H_0: $\mu = 0$, the Student distribution with $(k-1)$ degrees of freedom.

For unbalanced experiments, in addition to the arithmetic mean

$$\hat{\mu}_N = \frac{1}{k} \sum_{i=1}^{k} \hat{\mu}_i$$

the weighted arithmetic mean

$$\hat{\mu}_W = \frac{1}{n} \sum_{i=1}^{k} n_i \times \hat{\mu}_i \tag{4.37}$$

is also computed. The corresponding variances of $\hat{\mu}_N$ and $\hat{\mu}_W$ are

$$D(\hat{\mu}_W) = \frac{1}{n^2} \sum_{i=1}^{k} n_i^2 \left[\frac{\sigma_e^2}{n_i} + \sigma_A^2 \right] \tag{4.38}$$

and

$$D(\hat{\mu}_N) = \frac{1}{k^2} \sum_{i=1}^{k} \left[\frac{\sigma_e^2}{n_i} + \sigma_A^2 \right] \tag{4.39}$$

In the numerical calculation of σ_A^2 and σ_e^2 their estimates are substituted, and from both of these, the mean value taken is the one for which the variance has the lower value.

Problem 4.7 *Evaluation of quality of $AgNO_3$*
Use the data of Problem 4.1, but suppose that instead of 5 bottles of silver nitrate from known suppliers, five bottles randomly selected from stores were used. Test the quality of $AgNO_3$ available in the stores.
Data: As for Problem 4.1
Program: Chemstat: ANOVA-1: One-way.
Solution: The residual variance $\sigma_e^2 = 0.2254$ from Problem 4.1 is substituted into Eq. (4.34)

$$\sigma_A^2 = \frac{22 \times 4[0.6999 - 0.2254]}{22^2 - 106} = 0.1104$$

Since the test statistic $F_e = 3.10$ computed from Eq. (4.35) is higher than the quantile of the Fisher–Snedecor distribution at a significance level $\alpha = 0.05$, $F_{0.95}(4, 17) = 2.9$, the null hypothesis $H_0{:}\sigma_A^2 = 0$ is rejected.
Conclusion: The variability of the quality of silver nitrate in bottles in the stores is significant. The fixed-effects model thus gives a different answer from the random-effects model in interpretation of results.

In random-effects models, the assumption of normality may be violated for variables e_{ij} and also for α_i. Normality may be checked by the rankit plots for residuals. Rankit plots can also be drawn for means $\hat{\mu}_i$ or $\hat{\alpha}_i$, although the results are not absolutely correct[5].

Like the fixed-effects models, the random-effects models can have the normality of data improved by use of a suitable transformation, but variance estimation can be a problem in the transformed scale. When the assumption of normality is violated, the Jack-knife technique (Chapter 3) can be used for estimation of the variance σ_A^2 and testing its significance. For detection of outliers and heteroscedasticity, the same technique as for the fixed-effects model is adopted.

4.3 TWO-WAY ANOVA

In the previous section we explained the simplest kind of ANOVA problem, that involving a single factor. We now focus on the two-factor case, which is generally referred to as two-way ANOVA. This extension is by no means trivial. We shall describe how the two-factor situation may be classified according to the pattern of the data.

4.3.1 Two-way data patterns

Several different types of data patterns for two-way ANOVA are illustrated in Table 4.5. Each of these tables describes a two-factor study with three levels of factor B (the "column" factor) and four levels of factor A (the "row" factor). The combination of level A_i and B_j is called a *cell*. The *y*s in each table correspond to individual observations on a single dependent variable y. The number of ys in a given cell is denoted by n_{ij} for the ith level of factor A and the jth level of factor B. The marginal total for the ith row is denoted by $n_{i.}$ and for the jth column by $n_{.j}$. The total number of observations is denoted by n.

Table 4.5—Some two-way data patterns for a 4 × 3 table

Factor B

Factor A	y	y	y
	y	y	y
	y	y	y
	y	y	y

(a) Single observation per cell ($n_{ij} = 1$) (balanced case)

Factor B

Factor A	$yyyy$	$yyyy$	$yyyy$
	$yyyy$	$yyyy$	$yyyy$
	$yyyy$	$yyyy$	$yyyy$
	$yyyy$	$yyyy$	$yyyy$

(b) Equal number of replicates per cell ($n_{ij} = 4$) (balanced case)

Factor B

Factor$A$	$yyyy$	yy	yyy	$n_{1.} = 9$
	$yyyy$	yy	yyy	$n_{2.} = 9$
	$yyyy$	yy	yyy	$n_{3.} = 9$
	$yyyy$	yy	yyy	$n_{4.} = 9$

$n_{.1} = 16$ $n_{.2} = 8$ $n_{.3} = 12$

(c) Equal replications by column, proportional replications by row ($n_{ij} = n_{.j}/4$) (unbalanced case)

Factor B

Factor A	$yyyy$	yy	yyy	$n_{1.} = 9$
	$yyyy$	yy	yyy	$n_{2.} = 18$
	$yyyy$	yy	yyy	
	$yyyy$	yy	yyy	$n_{3.} = 27$
	$yyyy$	yy	yyy	
	$yyyy$	yy	yyy	
	$yyyy$	yy	yyy	$n_{4.} = 18$
	$yyyy$	yy	yyy	

$n_{.1} = 32$ $n_{.2} = 16$ $n_{.3} = 24$ $n = 72$

(d) Proportional row and column replications ($n_{ij} = n_{i.}n_{.j}/n$) (unbalanced case)

Factor B

Factor A	yy	yyy	$yyyyyy$	$n_{1.} = 11$
	yyy	$yyyy$	yy	$n_{2.} = 9$
	y	yyy	$yyyy$	$n_{3.} = 8$
	$yyyyy$	yy	y	$n_{4.} = 8$

$n_{.1} = 11$ $n_{.2} = 12$ $n_{.3} = 13$ $n_{.3} = 36$

(e) Nonsystematic replications (unbalanced case)

The simplest two-factor pattern (Table 4.5a), arises when there is a single observation in each cell (i.e., $n_{ij} = 1$ for all i and j). A second type of pattern (Table 4.5b) occurs when there are equal numbers of observations in each cell. Here, $n_{ij} = 4$ for all i and j. The common property of the last three patterns is that all cells do not have the same number of observations. Unequal cell replications often arise in observational studies in which the levels of certain effects are determined after, rather than before, the data are collected.

For the pattern in Table 4.5c, cells in the same column have the same number of observations, whereas cells in the same row are in the ratio 4:2:3. For this table each of the four cell frequencies in the jth column is equal to the same fraction of the corresponding total column frequency (i.e., $n_{ij} = n_j/4$ in this case). Note, for example, that $n_1/4 = 16/4$, which is the number of observations in any cell in column 1.

For Table 4.5d the cells in a given column are in the ratio 1:2:3:2, whereas the cells in a given row are in the ratio 4:2:3. This pattern results because n_{ij} is determined as $n_{ij} = n_{i.} \times n_{.j}/n$, which means that any cell frequency can be obtained by multiplying the corresponding row and column marginal frequencies together and then dividing by the total number of observations. Thus, for cell (1,2) in Table 4.5d, we have $n_{1.} \times n_{.2}/n = 9(16)/72 = 2$, which equals n_{12}. Similarly, for cell (4, 3), $n_{4.} \times n_{.3}/n = 18(24)/72 = 6$, which equals n_{43}.

There is no mathematical rule for describing the pattern of cell frequencies in Table 4.5e, and so we say that such a pattern is nonsystematic.

4.3.2 Formulation of various two-way ANOVA models

We shall consider the case in which we must set up an experiment to study the effects of two factors A and B on a response variable y. Factor A has N levels $\alpha_1, \alpha_2, \ldots, \alpha_N$, whereas factor B has M levels $\beta_1, \beta_2 \ldots, \beta_M$. For each combination of levels $(\alpha_i \beta_j)$, we measure the response y_{ij} by carrying out n_{ij} observations. The total number of observations is $n = \Sigma_{i=1}^{N} \Sigma_{j=1}^{M} n_{ij}$. The model of the response to each treatment may be written

$$y_{ijk} = \mu + \alpha_\mathrm{i} + \beta_j + \tau_{ij} + \varepsilon_{ijkq} \tag{4.40}$$

where μ represents an overall mean or a common effect, α_i represents the row effect on the ith level of factor A ($i = 1, 2, \ldots, N$), β_j represents the column effects on the jth level of factor B ($j = 1, 2, \ldots, M$) and τ_{ij} represents the effect due to the interaction of two factors, A and B.

The interaction term τ_{ij} is the deviation of the mean of observations in the (ij)th set from the sum of the first three terms in the model defined by Eq. (4.40), and ε_{ijk} ($k = 1, 2, \ldots, n_{ij}$) represents the error term.

The simplest model of interaction of rows with columns is the *Tukey model of interaction* defined by

$$\tau_{ij} = C \times \alpha_i \times \beta_j \tag{4.41}$$

where C is a constant. More complicated models of interaction are the *row-linear interaction model* expressed by

$$\tau_{ij} = \gamma_i \times \beta_j \times C_\mathrm{R} \tag{4.42}$$

or the *column-linear interaction model* expressed by

$$\tau_{ij} = \delta_j \times \alpha_i \times C_\mathrm{K} \tag{4.43}$$

The extended model is the *additive-multiplicative interaction model* expressed by

$$\tau_{ij} = \delta_j \times \gamma_i \times C_{\mathrm{W}} \tag{4.44}$$

These expressions contain, in addition to the column and row constants δ_j and γ_i, also the general constants C_{R}, C_{K} and C_{W}.

Interactions of second and higher orders also exist, and these can express rather complicated structures in data. Analysis of such non-linear interaction models may be found in the literature[10].

We limit ourselves to the simplest (Tukey) model of interaction Eq. (4.41). Since this model contains only one parameter C, it is called the model with *one degree of freedom for non-additivity*. This model expresses approximately the interaction effects in quadratic models for which:

$$\mu_{ij} \approx (\mu + \alpha_i + \beta_j)^2$$

After rearrangement of this equation, the interaction term is of the type $2\alpha_i\beta_j$. Use of the Tukey model of interaction is suitable for cases when each cell contains just one observation.

4.3.3 Fixed-effects two-way ANOVA

This type of two-way screening is the most frequently used. It allows evaluation of the effect of two factors on the results of a chemical analysis. Two-way ANOVA problems can be separated into three groups:

(1) models with a single observation per cell
(2) balanced models
(3) unbalanced models.

For each model, a different computational procedure is required.

4.3.3.1 Models with a single observation per cell

In these models the each cell contains only one observation, and the model is described by Eq. (4.40). The errors ε_{ij} are assumed to be independent identically distributed random variables with zero mean and constant variance. In testing, it is assumed that the error distribution is normal. In the ANOVA model, there are the following constraints

$$\sum_{i=1}^{N} \alpha_i = 0; \ \sum_{j=1}^{M} \beta_j = 0; \ \sum_{i=1}^{N} \tau_{ij} = 0; \ \sum_{j=1}^{M} \tau_{ij} = 0 \tag{4.45}$$

For pure additive effect of individual factors, $\tau_{ij} = 0$ for all $i = 1, \ldots, N$ and $j = 1, \ldots, M$. The estimates of parameters μ, α_i, and β_j are in this case calculated from

$$\hat{\mu} = \frac{1}{N \times M} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} \tag{4.46}$$

$$\hat{\alpha}_i = \frac{1}{M} \sum_{j=1}^{M} y_{ij} - \hat{\mu} \tag{4.47}$$

$$\hat{\beta}_j = \frac{1}{N}\sum_{i=1}^{N} y_{ij} - \hat{\mu} \tag{4.48}$$

For residuals

$$\hat{\mathrm{e}}_{ij} = y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_{\mathrm{j}} \tag{4.49}$$

When Eq. (4.40) is considered as the special linear regression model, the diagonal elements H_{ii} of a projection matrix **H** have the same value [7]

$$H_{ii} = \frac{N + M - 1}{N \times M} \tag{4.50}$$

Off-diagonal elements are not zero, so an outlier in one cell affects the estimates of parameters for all cells.

To determine the interaction we use the fact that

$$\tau_{ij} = E(y_{ij}) - \mu - \alpha_i - \beta_j \tag{4.51a}$$

Then, the estimate of interaction is given approximately by

$$\hat{\tau}_{ij} \approx \hat{\mathrm{e}}_{ij} \tag{4.51}$$

By using Eq. (4.51), the *Tukey model* of interaction [Eq. (4.41)] may be identified. If the plot of $\hat{\mathrm{e}}_{ij}$ *vs.* $\hat{\alpha}_i\hat{\beta}_j$ is linear, the Tukey model of interaction is accepted. The parameter C is calculated from the slope of this straight line

$$C = \left[\sum_{i=1}^{N}\sum_{j=1}^{M} \hat{e}_{ij} \times \hat{\alpha}_i \times \hat{\beta}_j\right] \Bigg/ \left[\sum_{i=1}^{N}\sum_{j=1}^{M} \hat{\alpha}_i^2 \times \hat{\beta}_j^2\right] \tag{4.52}$$

The slightly modified plot $\hat{e}_{ij}$ *vs.* $\hat{\alpha}_i \times \hat{\beta}_j/\hat{\mu}$ is called *the non-additivity graph.* If this plot exhibits a non-random trend, interactions probably exist.

The sums of squares of ANOVA model for a Tukey model of interaction are given in Table 4.5. The quantity S_{T} is the sum of squared deviations corresponding to the Tukey interaction [3].

$$S_{\mathrm{T}} = \left[\sum_{i=1}^{N}\sum_{j=1}^{M} y_{\mathrm{ij}} \times \hat{\alpha}_i \times \hat{\beta}_j\right]^2 \Bigg/ \left[\sum_{i=1}^{N}\sum_{j=1}^{M} \hat{\alpha}_i^2 \times \hat{\beta}_j^2\right] \tag{4.53}$$

and the symbol S_{AB} means a residual sum of squares for the case without interaction

$$S_{\mathrm{AB}} = \sum_{i=1}^{N}\sum_{j=1}^{M} (y_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j)^2 \tag{4.54}$$

The corresponding mean square M_{AB}

$$M_{\mathrm{AB}} = \frac{S_{\mathrm{AB}}}{(N-1)\ (M-1)} \tag{4.54a}$$

represents an unbiased estimate of the variance σ^2.

Table 4.6—Two-way ANOVA model with interaction of Tukey type

Sum of squares for	Degrees of freedom	Mean square	Test criterion F
Factor A $S_A = M \sum_{i=1}^{N} \hat{\alpha}_i^2$	$N - 1$	$M_A = S_A/(N - 1)$	$F_A = M_A/M_{AB}$
Factor B $S_B = N \sum_{j=1}^{M} \beta_j^2$	$M - 1$	$M_B = S_B/(M - 1)$	$F_B = M_B/M_{AB}$
Interaction (Tukey) S_T = [Eq.(4.53)]	1	$M_T = S_T$	$F_T = M_T/M_E$
Residuals $S_R = S_{AB} - S_T$	$N \times M - N - M$	$M_E = S_R/(N \times M - N - M)$	—
Totals $S_C = \sum_{i=1}^{N} \sum_{j=1}^{M} (\hat{\mu} - y_{ij})^2$	$N \times N - 1$	—	—

Statistical tests based on the Fisher–Snedecor F-criterion may be performed. The null hypothesis H_0: "Tukey interaction is not significant" is tested by the F_T criterion from Table 4.6. If the null hypothesis H_0 is valid, F_T has the Fisher–Snedecor F-distribution with 1 and $(N \times M - N - M)$ degrees of freedom. When this hypothesis is not rejected, a test of the null hypothesis H_0: $\alpha_i = 0$, $i = 1, \ldots, N$ (the effects of rows or factor A are not significant) using the statistic F_A, or a test of the null hypothesis H_0: $\beta_j = 0$, $j = 1, \ldots, M$ (the effects of columns or factor B are not significant) using the statistic F_B may be made. If the null hypothesis H_0 is valid, the F_A statistic has the Fisher–Snedecor F-distribution with $(N - 1)$ and $(N - 1)(M - 1)$ degrees of freedom. F_B has the same distribution with $(M - 1)$ and $(N - 1)(M - 1)$ degrees of freedom.

If F_T is higher than corresponding quantile of the Fisher–Snedecor distribution, the effect of Tukey interaction is significant.

Problem 4.8 *Determination of water content in solvents, in various laboratories*
In three samples of solvent A_1, A_2 and A_3, the content of water was determined in four laboratories B_1, B_2, B_3 and B_4. Test whether there are significant differences between the water contents of the three samples and in the results coming from the four laboratories.
Data: $N = 3$, $M = 4$

Table 4.7—Water content, %, in solvent found by various laboratories

Sample	Laboratory			
	B_1	B_2	B_3	B_4
A_1	1.35	1.13	1.06	0.98
A_2	1.40	1.23	1.26	1.22
A_3	1.49	1.46	1.40	1.35

Program: Chemstat: ANOVA-2P: Two-way, One Observation Per Cell.
Solution: It is found that $\hat{\mu} = 1.277$; $\hat{\alpha}_1 = -0.147$; $\hat{\alpha}_2 = 0$; $\hat{\alpha}_3 = 0.1475$; $\hat{\beta}_1 = 0.1358$; $\hat{\beta}_2 = 0.0042$; $\hat{\beta}_3 = -0.0375$ and $\hat{\beta}_4 = -0.0942$. The slope estimate from Eq. (4.52) is $C = -3.532$. Figure 4.3 shows a non-additivity graph with a slightly significant trend.

The sum of square deviations corresponding to Tukey interaction is $S_T = 0.0156$. The sum of square is $S_{AB} = 0.02215$ and $M_{AB} = 0.003692$. The results are summarized in the ANOVA Table 4.8.

Table 4.8—ANOVA table of water content in different solvents determined by different laboratories

Sum of squares for	Degrees of freedom	Mean square	Test criterion F
Samples A $S_A = 0.174$	2	0.087	19.64
Laboratories B $S_B = 0.0862$	3	0.0287	6.49
Interaction (Tukey) $S_T = 0.0156$	1	0.0156	3.522
Residuals $S_R = 0.0222$	5	0.0044	—
Totals $S_C = 0.2824$	11	0.0257	—

The quantiles of the Fisher–Snedecor distribution are $F_{0.95}$ (1,5) = 6.61; $F_{0.95}$ (2,5) = 5.79 and $F_{0.95}$ (3,5) = 5.41.
Conclusion: The interaction effect is not significant and the additive model of ANOVA can be used: sample effect and laboratory effect are significant. There are non-random difference between laboratory results and samples of solvent.

In some cases it is convenient to apply the power transformation

$$y_{ij}^* = \begin{cases} \dfrac{(y_{ij} + Q)^\lambda - 1}{\lambda} & \text{for } \lambda \neq 0 \\ \ln\,(y_{ij} + Q) & \text{for } \lambda = 0 \end{cases} \tag{4.55}$$

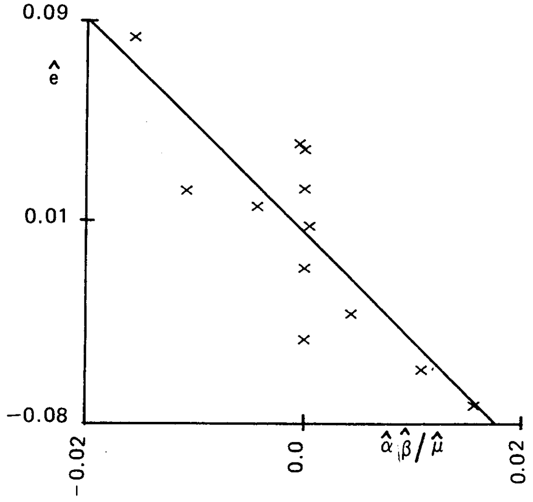


Fig. 4.3—Non-additivity graph for water content in different solvents.

where Q is a constant selected to make $(y_{ij} + Q) > 0$ and λ is a parameter of transformation. Details about the power transformation are given in Chapter 2. In ANOVA, the power transformation can be used for eliminating non-additivity. The parameter C is estimated from Eq. (4.52) (it is equal to the slope of the regression straight line divided by $\hat{\mu}$ in a non-additivity graph). The value of λ may then be evaluated from the equation

$$\lambda = 1 - \hat{\mu} \times C \tag{4.56}$$

In practice, λ is rounded to the nearest number from the following series: ..., − 1.0; − 0.5; − 0.3; 0; 0.3; 0.5; 1; 1.5; 2; .... For the estimate of variance $\hat{\sigma}^2(\lambda)$ the following expression is used [11]

$$\hat{\sigma}^2(\lambda) = \frac{\hat{\mu}^2 \times N \times M \times S_{AB}}{(N-1)(M-1)\sum_{i=1}^{N}\sum_{j=1}^{M}\hat{\alpha}_i^2\hat{\beta}_j^2} \tag{4.57}$$

For a first guess, transformations in the range $\lambda \pm \hat{\sigma}(\lambda)$ are acceptable.

Table 4.6 can be also used for ANOVA without interaction when $S_T = 0$. Since each observation y_{ij} affects parameters in all cells and a projection matrix is not diagonal, the classical analysis of residuals by Eq. (4.49) can lead to wrong conclusions. The non-random trend on a rankit plot indicates an interaction or non-normality. Detection of outliers cannot be performed by residuals $\hat{e}_{ij}$.

For identification of outliers, so-called "proper tetrads" $T_{ij:eg}$ are recommended [8]. These are given by

$$T_{ij:eg} = y_{ij} - y_{ej} - y_{ig} + y_{eg} \qquad i \neq e, j \neq g \tag{4.58}$$

Instead of residuals, medians Q_{ij} of all tetrads are calculated which contain observations y_{ij}. Medians Q_{ij} are plotted in rankit graphs.

Another technique for identification of outliers is a calculation of robust estimates of parameters α_i and β_j, and consequently also robust residuals. A suitable robust technique here is the *median-polish procedure* [9] which is a nonparametric technique corresponding to a regression, and which minimizes the absolute deviations. The median polish is very simple iterative technique, involving the successive subtraction of row and column medians from the data values. The procedure will remove row and column effects to create a new polished table containing only the residuals. In the resulting table of residuals each row and each column has zero median. Also, centring of row and column effects is carried out, so that their medians are equal to zero.

Problem 4.9 *Median-polish procedure for data of water content in various solvents*
To demonstrate the median-polish procedure, calculate one iteration for the data of water content in various solvents from Problem 4.6.
Data: As for Problem 4.6
Solution:

			Data		Median
	1.35	1.13	1.06	0.98	1.095
	1.40	1.23	1.26	1.22	1.245
	1.49	1.46	1.40	1.35	1.430

	after subtraction of row medians				Median
	0.255	0.035	– 0.035	– 1.115	1.095
	0.155	– 0.015	0.015	– 0.025	1.245
	0.060	0.030	– 0.030	– 0.080	1.430
Median	0.155	0.030	– 0.030	– 0.080	1.245

	after subtraction of column medians				Median
	0.100	0.005	– 0.005	– 0.035	– 0.150
	0.000	– 0.045	0.045	0.055	0.000
	–0.095	0.000	0.000	0.000	0.185
Median	0.155	0.030	– 0.030	– 0.080	1.245

Although in the table of residuals, all the column and row medians are not zero, it is possible to make a first guess of parameter estimates:

$$\hat{\mu} = 1.245;\ \hat{\alpha}_1 = -0.150;\ \hat{\alpha}_2 = 0.0;\ \hat{\alpha}_3 = 0.185;\ \hat{\beta}_1 = 0.155;$$

$\hat{\beta}_2 = 0.030;\ \hat{\beta}_3 = -0.030;\ \hat{\beta}_4 = -0.080.$

Conclusion: Comparison of estimates of parameters made from one iteration of the median-polish procedure shows that they are quite close to the estimates found by the classical procedure in Problem 4.6. In practice, usually three or five iterations of the median-polish procedure are used.

In the case of more complicated interactions, a matrix of residuals $\hat{e}_{ij}$ is formed, $\mathbf{E}(N \times M)$. Then, the matrix $\mathbf{E}^T\mathbf{E}$ is decomposed into eigenvalues and eigenvectors. From these quantities the parameters δ_j and γ_i are estimated. More details may be found in [10].

4.3.3.2 Balanced models

These models contain $n_{ij} = n^*$ observations in each cell. The ANOVA model is expressed by Eq. (4.3) or (4.4). An estimate of parameter μ_{ij} is represented by the arithmetic mean

$$\hat{\mu}_{ij} = \frac{1}{n^*} \sum_{k=1}^{n} y_{ijk} \tag{4.59}$$

For estimation of other parameters, the following expressions are used

$$\hat{\mu} = \frac{1}{N \times M} \sum_{i=1}^{N} \sum_{j=1}^{M} \hat{\mu}_{ij} \tag{4.60}$$

$$\hat{\alpha}_i = \frac{1}{M} \sum_{j=1}^{M} \hat{\mu}_{ij} - \hat{\mu} \tag{4.61}$$

$$\hat{\beta}_j = \frac{1}{N} \sum_{j=1}^{N} \hat{\mu}_{ij} - \hat{\mu} \tag{4.62}$$

Residuals are given by

$$\hat{e}_{ijk} = y_{ijk} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j \tag{4.63}$$

In this case as in Eq. (4.51a), an estimate of interaction is defined

$$\hat{\tau}_{ij} = \hat{\mu}_{ij} - \hat{\mu} - \hat{\alpha}_i - \hat{\beta}_j \tag{4.64}$$

Notice that this expression differs from Eq. (4.51) by the fact that instead of the variable y_{ij} the mean $\hat{\mu}_{ij}$ is used. For a test of non-additivity of Tukey interaction, the plot $\hat{\tau}_{ij}$ *vs.* $\hat{\alpha}_i \times \hat{\beta}_j$ can be adopted. A random pattern in this graph indicates additive effects of the two factors. The sums of squares for this type of ANOVA model are given in Table 4.9.

Table 4.9—Two-way ANOVA table for a balanced model

Sum of squares for	Degrees of freedom	Mean square	Test F-criterion
Factor A $S_A = n^* \times M \sum_{i=1}^{N} \hat{\alpha}_i^2$	$N - 1$	$M_A = \frac{S_A}{N-1}$	$F_A = \frac{M_A}{M_R}$
Factor B $S_B = n^* \times N \sum_{j=1}^{M} \hat{\beta}_j^2$	$M - 1$	$M_B + \frac{S_B}{M-1}$	$F_B = \frac{M_B}{M_R}$
Interaction AB $S_{AB} = n^* \sum_{i=1}^{N} \sum_{j=1}^{M} \hat{\tau}_{ij}^2$	$(N-1)(M-1)$	$M_{AB} = \frac{S_{AB}}{(N-1)(M-1)}$	$F_{AB} = \frac{M_{AB}}{M_R}$
Residuals $S_R = \sum_i^N \sum_j^M \sum_k^{n^*} (y_{ijk} - \hat{\mu}_{ij})^2$	$M \times N(n^* - 1)$	$M_R = \frac{S_R}{M \times N(n^* - 1)}$	—
Totals $S_C = \sum_i^N \sum_j^M \sum_k^{n^*} (y_{ijk} - \hat{\mu})^2$	$M \times N \times n^* — 1$	—	—

The corresponding expected values of mean squares are

$$E(M_A) = \sigma^2 + \frac{n^* \times M \sum_{i=1}^{N} \alpha_i^2}{(N-1) \times \sigma^2} = \sigma^2 + n^* \times M\sigma_A^2$$

$$E(M_B) = \sigma^2 + \frac{n^* \times M \sum_{j=1}^{M} \beta_j^2}{(M-1) \times \sigma^2} = \sigma^2 + n^* \times N\sigma_B^2$$

$$E(M_{AB}) = \sigma^2 + \frac{n^* \sum_{i=1}^{N} \sum_{j=1}^{M} \tau_{ij}^2}{(N-1)(M-1)\sigma^2} = \sigma^2 + n^* \times \sigma_{AB}^2$$

The expected value $E(M_R) = \sigma^2$ shows that the variance M_R is an unbiased estimate σ^2 of an error variance. Variances σ_A^2, σ_B^2 and σ_{AB}^2 correspond to the effects of rows, columns and interaction. These expressions are used also in calculation of estimates of variance of factors and interaction. Then instead of mean values $E(.)$, the mean squares, and instead of variance σ^2, the residual variance $\hat{\sigma}^2$ (cf. Problem 4.10) are substituted. It is important to note that the mean squares are not estimates of the corresponding variances.

Statistical criteria F_{AB}, F_B and F_A defined in the ANOVA Table 4.9 are used to test whether the interaction effects, column effects and row effects are significant. To test a null hypothesis H_0: $\tau_{ij} = 0$, $i = 1, \ldots, N$ and $j = 1, \ldots, M$, the test criterion

F_{AB} is used; if H_0 is valid, this has the Fisher–Snedecor distribution with $(N - 1)$ $(M - 1)$ and $M \times N(n^* - 1)$ degrees of freedom. To test the significance of row effects (factor A), the null hypothesis H_0 is $\alpha_i = 0$, $i = 1, \ldots, N$. When H_0 is valid, the test criterion F_A has the Fisher–Snedecor distribution with $(N - 1)$ and $M \times N(n^* - 1)$ degrees of freedom. For column effects (factor B), the null hypothesis H_0 is $\beta_j = 0$, $j = 1, \ldots, M$. When H_0 is valid, the test criterion F_B has the Fisher–Snedecor distribution with $(M - 1)$ and $M \times N(n^* - 1)$ degrees of freedom. The unbiased estimator of variance is here M_R.

For balanced models, the diagonal elements H_{ii} of a projection matrix $\mathbf{H}$ are constant.

It may be concluded that the two models, single-observation-per-cell (I) and balanced (II) differ only in replacement of quantities.

An advantage of all balanced models is mutual orthogonality of the individual terms of the ANOVA model, so that the individual partial sums of squares in Table 4.9 and 4.6 may be added. This may be exploited for simultaneous testing of several hypotheses.

Problem 4.10 *Derivation of a test criterion for examining the independence of factor B*
Let us suppose that Table 4.9 is available and we wish to test whether the results depend on a factor B. That is, we test the validity of the two hypotheses, $H_{01}\colon \tau_{ij} = 0$ and $H_{02}\colon \beta_j = 0$, simultaneously.
Solution: Since partial sums of squares may be added, we can calculate the sum of squares resulting from factor B simply as

$$S_{PB} = S_B + S_{AB}.$$

The corresponding number of degrees of freedom $N(M - 1)$ is the sum of degrees of freedom $(M - 1)$ and $(N - 1)(M - 1)$, so that the mean square is

$$M_{PB} = S_{PB}/[N(M - 1)].$$

The test criterion

$$F_{PB} = M_{PB}/M_R$$

has the Fisher–Snedecor distribution with $N(M - 1)$ and $M \times N(n^* - 1)$ degrees of freedom, if the partial hypotheses H_{01} and H_{02} are valid.
Conclusion: It is possible to test another simultaneous hypotheses, similarly. That is, is the variability of y due to interaction only, which corresponds to $H_{01}\colon \sigma_i = 0$ and $H_{02}\colon \beta_j = 0$.

With the use of the procedure shown in Problem 4.9 the validity of various ANOVA submodels may be tested, from the simplest

$$y_{ijk} = \mu + \varepsilon_{ijk}$$

over all partial models (containing only some of the parameters α, β and τ) to the total analysis expressed by Eqs. (4.3) and (4.4). Summation of sums of squares is recommended also in cases when the influence of some factors or interactions is proved to be not significant. Then, a corresponding sum of squares is added to the residual one, and corresponding terms are omitted.

Problem 4.11 *Finding the residual sum of squares for a two-way layout with insignificant interaction*

Suppose that Table 4.9 is available, and that the Fisher–Snedecor test has proved that the interaction effect is not significant and the null hypothesis $H_0\colon \tau_{ij} = 0$ has been accepted. Calculated the corrected residual sum of squares without forming a new ANOVA table.

Solution: The sum of squares S_R is combined with the sum of squares S_{AB}, and the corresponding number of degrees of freedom is found. The corrected residual sum of squares is

$$S_R^* = S_R + S_{AB}$$

and the corrected mean square is

$$M_R^* = \frac{S_R^*}{M \times N(n^* - 1) + (N - 1)(M - 1)}$$

In the Fisher–Snedecor F-test, M_R^* is used in the denominator of the F-criterion, and the corresponding degrees of freedom are corrected.

Conclusion: The orthogonality of parameters permits use of the various partial ANOVA models without having to calculate new estimates and an ANOVA table.

Problem 4.12 *Precision of chromatographic determination of diethyleneglycol in ethyleneglycol*

Three laboratory technicians A_1, A_2 and A_3 can work on three chromatographs B_1, B_2 and B_3 and determine diethylene glycol (DEG) in ethylene glycol. Each technician made two determinations with each chromatograph. Besides ANOVA, estimate how much of the variance corresponds to the instrument (the instrumental error) and how much to the technician (the error of the experimenter).

Data: $N = 3$, $M = 3$, $n^* = 2$. Data for percentage of DEG found in ethylene glycol are given in Table 4.10

Table 4.10—Concentration of DEG [%] measured twice by three technicians A_1, A_2 and A_3 on three instruments B_1, B_2 and B_3

Technician	Instrument		
	B_1	B_2	B_3
	0.110	0.101	0.108
A_1	0.116	0.102	0.109
	0.112	0.115	0.111
A_2	0.111	0.106	0.109
	0.114	0.107	0.113
A_3	0.112	0.109	0.110

Program: Chemstat: ANOVA–2B: Two-way balanced.
Solution: The ANOVA table is shown in Table 4.11. For the significance level $\alpha = 0.05$ the quantiles are $F_{0.95}$ (2,9) = 4.26 and $F_{0.95}$ (4,9) = 3.63. From F criteria in Table 4.11, it is evident that the effects of factor A and interaction AB are separately not significant. Let us test whether technicians have an influence on the determination of DEG. Two null hypotheses are formulated, H_{01}: $\alpha_i = 0$ and H_{02}: $\tau_{ij} = 0$. The corresponding mean square is

$$S_{\mathrm{PA}} = (S_{\mathrm{A}} + S_{\mathrm{AB}})/(2 + 4) = 1.63 \times 10^{-5}$$

and the test criterion

$$F_{\mathrm{PA}} = 1.63 \times 10^{-5}/7.83 \times 10^{-6} = 2.09.$$

Since the corresponding quantile $F_{0.95}$ (6, 9) = 3.373 is higher than the criterion F_{PA}, the technicians have no significant influence on the determination of DEG and the ANOVA model is formulated by the equation

$$y_{ijk} = \mu + \beta_j + \varepsilon_{ijk}$$

Let us test the significance of factor B with H_0: $\beta_j = 0$. In addition, to avoid construction of a new ANOVA table, the sums $S_{\mathrm{A}} + S_{\mathrm{AB}}$ will be added to the residual sum of squares S_{R}. The mean square will then be

$$M_{\mathrm{R}}^* = \frac{S_{\mathrm{R}} + S_{\mathrm{A}} + S_{\mathrm{AB}}}{2 + 4 + 9} = \frac{1.688 \times 10^{-5}}{15} = 1.12 \times 10^{-5}$$

Table 4.11—ANOVA table for DEG determination

Sum of squares	Degrees of freedom	Mean square	Test criterion F
Technicians (A) $S_{\mathrm{A}} = 3.81 \times 10^{-5}$	2	1.906×10^{-5}	2.433
Instruments (B) $S_{\mathrm{B}} = 1.027 \times 10^{-4}$	2	5.139×10^{-5}	6.560
Interaction (AB) $S_{\mathrm{AB}} = 6.02 \times 10^{-5}$	4	1.506×10^{-5}	1.922
Residual $S_{\mathrm{R}} = 7.05 \times 10^{-5}$	9	7.833×10^{-6}	—
Totals $S_{\mathrm{C}} = 2.716 \times 10^{-4}$	17	—	—

The test criterion F_B is

$$F_B = M_B/M_B^* = 4.58$$

and the corresponding quantile $F_{0.95}(3, 15) = 3.68$ is lower. Thus, the null hypothesis H_0 is rejected, and the influence of the instrument on the determination of DEG has been shown to be significant at significance level $\alpha = 0.05$.

The expected value of a mean square is given by

$$E(M_B) = \sigma^2 + 6\sigma_B^2,$$

and the estimate of σ^2 is $\hat{\sigma}^2$. The estimate of instrumental error may be calculated from

$$\hat{\sigma}_B^2 = (M_B - \hat{\sigma}^2)/6 = 6.68 \times 10^{-6}.$$

Conclusion: It has been shown that the precision of determination of DEG is affected only by the instrument used. The variability caused by technicians and other random effects is $\hat{\sigma}^2 = 1.12 \times 10^{-5}$, and the variability caused by instruments (instrumental error) is $\hat{\sigma}_B^2 = 6.68 \times 10^{-6}$.

4.3.3.3 Unbalanced models

For unbalanced models, there are n_{ij} observations in the (i,j)th cell. When an experiment is poorly organized, so that differences in n_{ij} values are in tens, the ANOVA is rather complicated. Parameters of Eqs. (4.3) and (4.4) are not orthogonal, and the partitioning of the sum of squares is ambiguous. The analysis of variance is performed with the use of programs for linear regression, with models (4.3) and (4.4) considered as special regression models with dummy variables taking only the values 0 or 1.

For practical calculations in chemometrics, an approximate partitioning of overall sum of squares is used. This begins with a calculation of means:

$$\hat{\mu}_{ij} = \frac{1}{n_k} \sum_{k=1}^{n_k} y_{ijk} \tag{4.65}$$

for the cells. From these values the residual sum of squares is estimated

$$S_R = \sum_{i=1}^{N} \sum_{j=1}^{M} \sum_{k=1}^{n_k} (y_{ijk} - \hat{\mu}_{ij})^2 \tag{4.66}$$

For calculation of other components of the partitioned overall sum of squares, the estimates of means $\hat{\mu}_{ij}$ are used, with the assumption that they have been estimated from an equivalent number of observations n_e^* defined by

$$n_e^* = \left[\frac{1}{N \times M} \sum_{i=1}^{N} \sum_{j=1}^{M} (1/n_{ij})\right]^{-1} \tag{4.67}$$

Analysis of variance is done by the same technique as for balanced models (Table 4.8), and individual sums are defined by

$$S_A = n_e^* \times M \sum_{i=1}^{N} (\hat{\mu}_i - \hat{\mu})^2 \tag{4.68}$$

with $(N-1)$ degrees of freedom

$$S_{\mathrm{B}} = n_{\mathrm{e}}^{*} \times N \sum_{j=1}^{M} (\hat{\mu}_j - \hat{\mu})^2 \tag{4.69}$$

with $(M-1)$ degrees of freedom, and

$$S_{\mathrm{AB}} = n_{\mathrm{e}}^{*} \sum_{i=1}^{N} \sum_{j=1}^{M} (\hat{\mu}_{ij} - \hat{\mu}_i - \hat{\mu}_j + \hat{\mu})^2 \tag{4.70}$$

with $(N-1)(M-1)$ degrees of freedom. In these expressions, the following notation is used

$$\hat{\mu}_i = \frac{1}{M} \sum_{j=1}^{M} \hat{\mu}_{ij}; \quad \hat{\mu}_j = \frac{1}{N} \sum_{i=1}^{N} \hat{\mu}_{ij}; \quad \hat{\mu} = \frac{1}{N \times M} \sum_{i=1}^{N} \sum_{j=1}^{M} \hat{\mu}_{ij}$$

The sum $S_{\mathrm{A}} + S_{\mathrm{B}} + S_{\mathrm{AB}} + S_{\mathrm{R}}$ is not here exactly equal to S_{C}, but differences are relatively small. Tests for row, column and interaction effects are made as previously for balanced models.

When there are several replicates in individual cells, for each cell the sample variance s_{ij}^2 may be estimated, and a plot s_{ij}^2 *vs.* $\hat{\mu}_{ij}$ used to check for any variance on the mean (heteroscedasticity).

4.3.4 Mixed effects two-way ANOVA

Suppose that some factor A corresponds to fixed effects and another factor B to random effects. The factor B is usually considered as the noisy factor, and it is not usually tested. The ANOVA model is defined by Eqs. (4.3) and (4.4). For fixed effects the constraints are $\alpha_i = 0, i = 1, \ldots, N$. For random variables β_j and ε_{ijk} it is assumed that

(a) β_j are mutually independent random variables with the normal distribution $N(0, \sigma_{\mathrm{B}}^2)$;

(b) $\varepsilon_{\mathrm{ijk}}$ are mutually independent random variables with the normal distribution $N(0, \sigma_{\mathrm{e}}^2)$;

(c) the random variables β_j are independent of the random variables $\varepsilon_{\mathrm{ijk}}$.

Here is an example of the elucidation of an interaction. Some authors consider that there are random mutually independent variables with the normal distribution $\mathrm{N}(0,\sigma_{\mathrm{AB}}^2)$. We will consider that τ_{ij} are independent of β_j and identically distributed variables, so that their variance is equal to [5]

$$D(\tau_{ij}) = (1 - 1/N)\sigma_{\mathrm{AB}}^2 \tag{4.71}$$

For equal cell numbers $n_{ij} = n^*$, to analyse a mixed-effects model the ANOVA Table 4.8 may be used. The null hypotheses concerning nonsignificance of interaction, H_0: $\sigma_{\mathrm{AB}}^2 = 0$, and nonsignificance of factor B, H_0: $\sigma_{\mathrm{B}}^2 = 0$, are tested with the use of test criteria F_{AB} and F_{B} from Table 4.8. For a test of significance of fixed effects of factor A, another test criterion is used [5]

$$F_{\mathrm{PA}} = M_{\mathrm{A}}/\mathrm{M}_{\mathrm{AB}} \tag{4.72}$$

If the null hypothesis H_0: $\alpha_i = 0$ is valid, the test statistic F_{PA} has the Fisher–Snedecor distribution with $(N - 1)$ and $(N - 1)(M - 1)$ degrees of freedom. If there is no interaction, the test criterion F_{A} from Table 4.8 may be used. This procedure can also be modified for cases where the effects of factor A are random and the effects of factor B fixed.

Problem 4.13 *Influence of the instrument of determination of DEG*
Consider the same task as in Problem 4.12, with the difference that three technicians were chosen randomly from a team of laboratory staff. Factor A now has random effects and factor B fixed. By doing ANOVA for this mixed-factors model, examine the influence of the type of chromatograph.
Solution: In Problem 4.12 it was estimated that $M_{\mathrm{B}} = 5.139 \times 10^{-5}$ and $M_{\mathrm{AB}} = 1.506 \times 10^{-5}$. The test criterion F_{PA} (4.72) is now written as

$$F_{\mathrm{PB}} = M_{\mathrm{B}}/M_{\mathrm{AB}} = 3.412$$

The corresponding quantile $F_{0.95}\,(2,4) = 6.944$ is higher than this F_{PB} value. Therefore, the null hypothesis H_0: $\beta_j = 0$ is rejected and the type of chromatograph does have a significant effect on the determination of DEG.
Conclusion: The ANOVA result demonstrates that a small change in the assumptions about an experiment (here the random selection of technicians) leads to a change in the results of ANOVA.

4.3.5 Random-effects two-way ANOVA

In this case, the effects of both factors are random. The ANOVA model can be defined by Eqs. (4.3) and (4.4). The random components of the model, α_i, β_j, τ_{ij} and ε_{ijk} are characterized by following properties:

(a) α_i are mutually independent random variables with the normal distribution $N(0,\sigma_{\mathrm{A}}^2)$;
(b) β_j are mutually independent random variables with the normal distribution $N(0,\sigma_{\mathrm{B}}^2)$;
(c) τ_{ij} are mutually independent random variables with the normal distribution $N(0,\sigma_{\mathrm{AB}}^2)$;
(d) ε_{ijk} are mutually independent random variables with the normal distribution $N(0,\sigma_{\mathrm{e}}^2)$;
(e) all these random variables are independent of one another. For interaction τ_{ij} the estimates $\hat{\tau}_{ij}$ are dependent on $\hat{\alpha}_i$ and β_j. For the whole population it is assumed that the interactions are independent for this model [5].

The primary goal of ANOVA here is to find the estimates of the variance components σ_{A}^2, σ_{B}^2, σ_{AB}^2 and σ_{e}^2 and test them. We will limit the discussion to balanced models with n^* replication per cell. To find the estimates of the variance components, the statistics from Table 4.8 can be adopted. The following expressions are used for the estimates of the partial variances

$$\hat{\sigma}_{\mathrm{A}}^2 = \frac{M_{\mathrm{A}} - M_{\mathrm{AB}}}{n^* \times M} \tag{4.73}$$

$$\hat{\sigma}_{\mathrm{B}}^2 = \frac{M_{\mathrm{B}} - M_{\mathrm{AB}}}{n^* \times M} \tag{4.74}$$

$$\hat{\sigma}_{\mathrm{AB}}^2 = \frac{M_{\mathrm{AB}} - M_{\mathrm{R}}}{n^*} \tag{4.75}$$

and $\sigma_{\mathrm{e}}^2 = M_{\mathrm{R}}$

The mean squares M_{A}, M_{B} and M_{AB} are defined in Table 4.8.

For testing the significance of the individual variance components, the F-tests are used as follows:

(a) For a *test of the null hypothesis* H_0: $\sigma_{\mathrm{AB}}^2 = 0$, the test criterion F is

$$F_{\mathrm{EAB}} = \frac{M_{\mathrm{AB}}}{M_{\mathrm{R}}} \tag{4.77}$$

If the null hypothesis H_0 is valid, the statistic F_{EAB} has a Fisher–Snedecor distribution with $(N-1)(M-1)$ and $N \times M(n^*-1)$ degrees of freedom.

(b) For a test of the null hypothesis H_0: $\sigma_{\mathrm{A}}^2 = 0$ the test criterion F is

$$F_{\mathrm{EA}} = \frac{M_{\mathrm{A}}}{M_{\mathrm{AB}}} \tag{4.78}$$

If the null hypothesis H_0 is valid the statistic F_{EA} has a Fisher–Snedecor distribution with $(N-1)$ and $(N-1)(M-1)$ degrees of freedom.

(c) For a test of the null hypothesis H_0: $\sigma_{\mathrm{B}}^2 = 0$ the test criterion F is

$$F_{\mathrm{EB}} = \frac{M_{\mathrm{B}}}{M_{\mathrm{AB}}} \tag{4.79}$$

If the null hypothesis is valid the statistic F_{EB} has a Fisher–Snedecor distribution with $(M-1)$ and $(N-1)(M-1)$ degrees of freedom.

When the estimates of parameters are interesting, the estimates $\hat{\mu}_{ij}$ and $\hat{\mu}$ defined for fixed-factors models can be applied. The estimate of variance of parameter $\hat{\mu}$ is defined by[5]

$$\hat{\sigma}_{\hat{\mu}}^2 = \frac{M_{\mathrm{A}}}{N \times M \times n^*} \tag{4.80}$$

Some alternative procedures for estimating the variance components are discussed in the literature[3–5].

Problem 4.14 *Factors affecting the chromatographic determination of DEG*

Consider the data of Problem 4.10, with the difference that both technicians and instruments were chosen randomly. Moreover, in the laboratory there are more than three technicians and instruments. This represents a random-effects model, and we want to find the effect of the variances of the individual sources on the variability of the final results.

Solution: A test of the null hypothesis H_0: $\sigma_B^2 = 0$ yields the same result as in Problem 4.13. To test the null hypothesis H_0: $\sigma_A^2 = 0$ the criterion is $F_{PA} = 1.264$. The corresponding quantile $F_{0.95}$ (2,4) = 6.944 is higher, so that the variance σ_A^2 cannot be considered significantly different from zero. For the null hypothesis H_0: $\sigma_{AB}^2 = 0$ in Problem 4.10 it was found that $F_{AB} = F_{EAB} = 1.93$, which is lower than the corresponding quantile $F_{0.95}(4,9) = 3.63$, so the null hypothesis H_0 cannot be rejected. The estimate of variance corresponding to instruments is

$$\sigma_B^2 = (M_B - M_{AB})/6 = 6 \times 10^{-6}.$$

Conclusion: The variance of any partial source of variance, σ_A^2, σ_B^2 and σ_{AB}^2, cannot be considered to be significantly different from zero, so the result of determination of DEG is loaded by random effects only.

4.4 SUMMARY

The first step of any ANOVA procedure is to recognize, on the basis of the data layout, whether the model is a fixed-, mixed- or random-effects one. For all three models the hypotheses to be tested must be specified, and the parameters to be estimated formulated. It is useful to know whether interaction is likely. The general procedure of ANOVA involves the following steps:

(1) Estimate the parameters of the ANOVA model.
(2) Test the significance of model and construct submodels or models with fixed-effects.
(3) Express the variance components for the random-effects model and test their significance.
(4) Test the assumptions of normality, homogeneity of variance and outliers. Residuals other than the classical ones may be used (Chapter 6 in Vol. 2).
(5) Interpret the ANOVA results with reference to data and assumptions.

4.5 ADDITIONAL SOLVED PROBLEMS

Problem 4.15 *Effect of type of penicillin on Bacillus subtilis*

An effect of four types of penicillin on growth of *Bacillus subtilis* was examined. Factor A (the type of penicillin) has 4 levels and at each level 5 replicated experiments was carried out. Test whether the effects of all the penicillins is the same.

Data: $N = 4$, $n_1 = n_2 = n_3 = n_4 = 5$

Replication	Type of penicillin			
	A_1	A_2	A_3	A_4
1	10.6	7.3	8.2	7.5
2	8.5	9.1	7.7	6.6
3	9.8	8.4	8.0	5.1
4	8.3	8.8	7.2	7.1
5	8.1	7.6	6.4	6.7

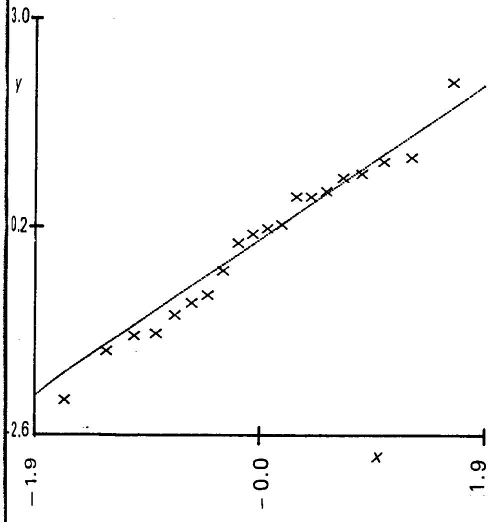


Fig. 4.4—Rankit plot for Jack-knife residuals

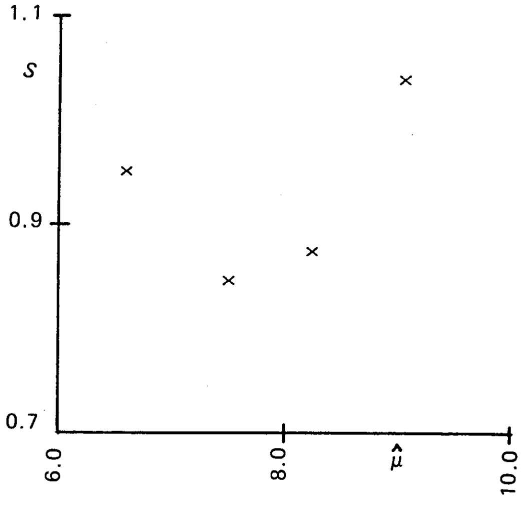


Fig. 4.5—Heteroscedasticity graph.

Program: Chemstat: ANOVA-1: One-way

Solution: The estimates of the ANOVA model parameters were found to be

$$\hat{\mu} = 7.85;\ \hat{\mu}_1 = 9.06;\ \hat{\mu}_2 = 8.24;\ \hat{\mu}_3 = 7.50 \text{ and } \hat{\mu}_4 = 6.60.$$

and the estimates of effects

$$\hat{\alpha}_1 = 1.21;\ \hat{\alpha}_2 = 0.39;\ \hat{\alpha}_3 = -0.35;\ \hat{\alpha}_4 = -1.25.$$

The diagonal elements of the projection matrix are

$$H_{11} = H_{22} = H_{33} = H_{44} = 0.2.$$

Since the quantile $F_{0.95}$ (3,16) = 3.24 is lower than $F = 7.04$, the null hypothesis H_0: $\alpha_i = 0$ is rejected and the effect of penicillin type is significant. Figure 4.4 shows a rankit plot for Jack-knife residuals and Fig. 4.5 a plot of s_i *vs.* $\hat{\mu}_i$

Table 4.12—ANOVA table for the effect of penicillin on *Bacillus subtilis*

Sum of squares for	Degrees of freedom	Mean square	Test criterion F
Type of penicillin (A) $S_a = 16.506$	3	5.502	7.04
Residual $S_R = 12.504$	16	0.782	—
Totals $S_c = 29.01$	19	—	—

In the analysis of Jack-knife residuals, the point $y_{11} = 10.6$ is indicated as the extreme ($\hat{e}_{J1} = 2.16$). The Scheffe method of multiple comparison causes the null hypothesis H_0: $\mu_1 = \mu_4$ to be rejected because $|\hat{\mu}_1 - \hat{\mu}_4| = 2.46$ is higher than the right-hand side of Eq. (4.16) which is equal to 1.741. For the small sample size, the systematic heteroscedasticity cannot be accepted.
Conclusion: The type of penicillin has a significant effect on growth of *Bacillus subtilis.*

Problems 4.16 *Effect of purity and mineralization on determination of organically bound nitrogen*
The effect of purity of organic compound (factor A) and conditions of Kjeldahl digestion (factor B) on organic nitrogen found, were studied. Five bottles of a single organic compound from five different producers, and five methods of digestion were used, and the nitrogen content was determined. Examine the significance of the variances of the two parameters.
Data: $N = M = 5$; $n = 1$

Bottle			Digestion		
	B_1	B_2	B_3	B_4	B_5
1	127	162	155	124	169
2	166	156	140	95	147
3	136	123	125	88	166
4	182	136	115	97	157
5	133	127	117	98	169

Program: Chemstat: ANOVA-2P: One observation per cell
Solution: The ANOVA model is formulated as the fixed-effects model with a single observation per cell. We will suppose that no interaction between bottle factors exists.

Table 4.13—ANOVA table for a two-way fixed-effects model with a single observation per cell

Sum of squares	Degrees of freedom	Mean square	Test criterion F
Factor A $S_A = 1382.8$	4	345.7	1.184
Factor B $S_B = 10\,700.8$	4	2675.2	9.165
Residual $S_R = 4378.4$	16	291.9	—
Totals $S_C = 16462$	24	—	—

Because the relevant quantile, $F_{0.95}$ (4,16) = 3.01, the null hypothesis H_0: $\alpha_i = 0$ is accepted but the null hypothesis H_0: $\beta_j = 0$ is rejected. Figure 4.6 shows the non-additivity plot.

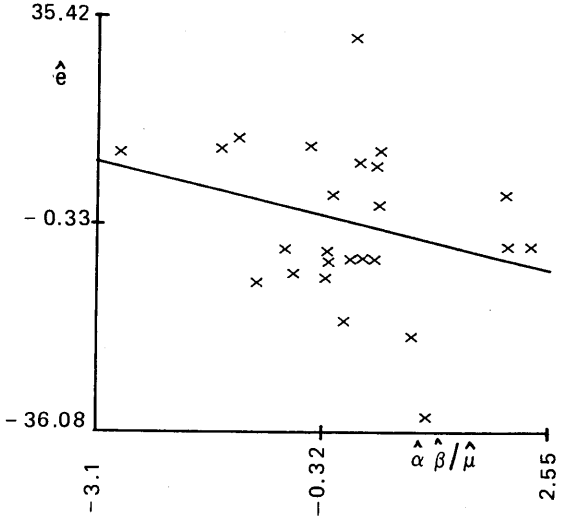


Fig. 4.6—The non-additivity plot.

Conclusion: The assumption of additivity of factors is approximately fulfilled. The determination of organically bound nitrogen is affected significantly by the digestion but not by the source of the reagents.

Problem 4.17 *Effect of two factors on yield of chemical reaction*
Test whether factor A and factor B affect the yield of a chemical reaction. Two levels of factor A were randomly chosen, A_1 and A_2, and three randomly chosen levels of factor B, i.e., B_1, B_2 and B_3.

Factor A		Factor B	
	B_1	B_2	B_3
	21.3	22.3	23.8
A_1	20.9	21.6	23.7
	20.4	21.0	22.6
	12.7	12.0	14.5
A_2	14.9	14.2	16.7
	12.9	12.1	14.5

For each combination of factors, three replicates of the chemical reaction were made, and the yield in [%] was determined.
Data: $N = 2$, $M = 3$, $n = 3$
Program: Chemstat: ANOVA–2B Balanced experiments
Solution: The two-way random-effects ANOVA model is assumed.

To test the null hypothesis H_0: $\sigma^2_{AB} = 0$, the test criterion is given by Eq. (4.77), i.e.

$$F_{EAB} = 0.85/0.95 = 0.894.$$

The corresponding quantile $F_{0.95}$ (2,12) = 3.89 is higher, so the null hypothesis H_0 *cannot be* rejected.

Table 4.14—ANOVA table

Sum of squares	Degrees of freedom	Mean square	Test criterion F
Factor A $S_A = 296.87$	1	296.87	312.13
Factor B $S_B = 17.78$	2	8.89	9.35
Interaction AB $S_{AB} = 1.69$	2	0.84	0.89
Residual $S_A = 11.41$	12	0.95	—
Totals $S_C = 327.75$	17	—	—

To test the null hypothesis H_0: $\sigma^2_A = 0$, the test criterion is given by Eq. (4.78), i.e.

$$F_{EA} = 296.87/0.85 = 349.26$$

so the null hypothesis H_0 must be rejected for $\alpha = 0.05$. To test the null hypothesis H_0: $\sigma^2_B = 0$ the test criterion is (Eq. 4.79):

$$F_{EB} = 8.89/0.85 = 10.16$$

and the corresponding quantile $F_{0.95}$ (2,2) = 19.00, so the null hypothesis H_0 cannot be rejected at $\alpha = 0.05$.
Conclusion: The yield of the chemical reaction is affected only by factor A. From the ANOVA table it is evident that assumption of a fixed-effects model would make factor B also significant. Therefore, the properties and quality of the selected factors must be exactly specified.

REFERENCES

[1] S. R. Searle, *Biometrics*, 1971, **27**, 1.
[2] M. S. Bartlett and D. G. Kendall, *J. Roy. Stat. Soc.*, 1946, **B8**, 128.
[3] H. Scheffe, *The Analysis of Variance*, Wiley, New York, 1959.
[4] S. R. Searle, *Linear Models*, Wiley, New York, 1971.
[5] P. G. Miller, *Beyond ANOVA, Basic of Applied Statistics*, Wiley, New York, 1986.
[6] T. P. Speed, *Ann. Statist.* 1987, **15**, 885.
[7] J. D. Emerson, D. C. Hoaglin and P. I. Kempthorne, *J. Am. Statist. Assoc.* 1984, **79**, 329.
[8] D. Bradu and D. M. Hawkins, *Technometrics*, 1982, **24**, 103.
[9] P. Bloomfield and W. Steiger, *Least Absolute Deviations: Theory, Applications and Algorithms.* Birkhauser, Boston, 1983.
[10] K. R. Gabriel, *J. R. Stat. Soc.*, 1972, **B40**, 186.
[11] N. A. C. Cressie, *Biometrics*, 1978, **34**, 505.