3 Graphics to Display the Data Distribution

As the Kola Project progressed, chemical analyses for c. 50 chemical elements in four different sample materials collected at about 600 sample sites were received. These needed to be summarised, compared and mapped; now we enter the realms of statistical data analysis. For data collected in space, two types of distribution need to be considered: the spatial distribution of the data in the survey area and the statistical data distribution of the measured values. At the beginning of data analysis the focus is on the statistical distribution, temporarily setting the spatial component aside (see Chapter 5, where the spatial data distribution is studied).

The statistical data distribution can be explained using some small artificial data sets. The data are simply plotted against a straight line as a scale for the data values. Depending on the data, the distribution of the values can look symmetrical (Figure 3.1, upper left, the data are distributed symmetrically around the value 20). The data could fall into two groups, i.e. be bimodal, and show a gap around the value 20 (Figure 3.1, upper right). The data could show a single extreme outlier (Figure 3.1, lower right). The data could show an asymmetrical distribution with a high density of points at the left side and a decreasing density towards the right side (Figure 3.1, lower right). All these different characteristics, or a mixture thereof, define the statistical data distribution. For further statistical treatment of the data it is essential to get a good idea about the data distribution. Many statistical methods are, for example, based on the assumption of a certain model for the data distribution (see Section 4.1, Figure 4.1 for a number of different model distributions).

With more than 600 measurements it will not be sufficient to simply plot the measured values along an axis as in Figure 3.1, because many data values will plot at the same point. Other graphics are needed to visualise the data distribution. These will be introduced on the following pages.

3.1 The one-dimensional scatterplot

Plotting the data along a straight line works fine as long as the data set is small and the data do not plot too close to, or on top of, one another. Figure 3.2, upper diagram, demonstrates the problem using the analytical results of Sc in the Kola C-horizon. Many more data points may be hidden behind one point plotted along the line. To view samples that have the same

Statistical Data Analysis Explained: Applied Environmental Statistics with R. C. Reimann, P. Filzmoser, R. G. Garrett, R. Dutter © 2008 John Wiley & Sons, Ltd. ISBN: 978-0-470-98581-6

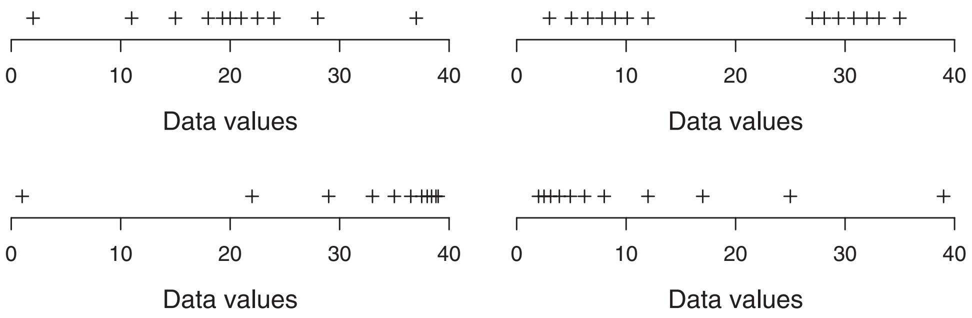

Figure 3.1 Some possible data distributions. Upper left: symmetric; upper right: bimodal with a gap around 20; lower left: left skewed; lower right: right skewed

value it is, however, possible to add a second dimension and add such values as additional symbols against the *y*-axis (Figure 3.2 middle, stacked scatterplot). With many values at the same position the *y*-axis starts to dominate the plot. With some further modifications the one-dimensional scatterplot is obtained (see, e.g., Box *et al.*, 1978); an informative and simple

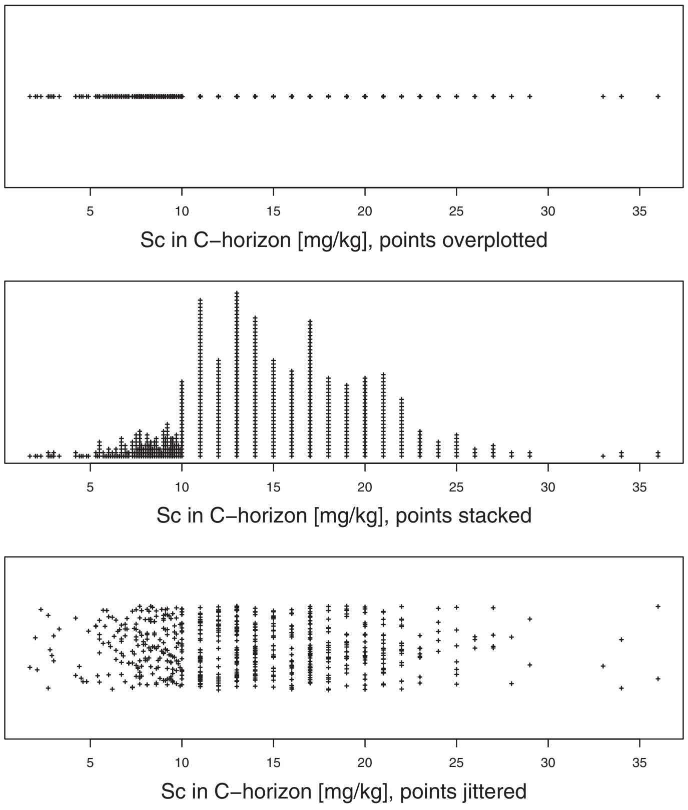

Figure 3.2 Evolution of the one-dimensional scatterplot demonstrated using Sc as measured by instrumental neutron activation analysis (INAA) in the samples of the Kola C-horizon

graphic to study the data distribution. This plot is typically displayed as an elongated rectangle. Each value is plotted at its correct position along the x-axis and at a position selected by chance (according to a random uniform distribution) along the y-axis (Figure 3.2, lower diagram). This simple graphic can provide important insight into structure in the data.

In Figure 3.2 (stacked and one-dimensional scatterplot) a significant feature is apparent that would be important to consider if this variable were to be used in a more formal statistical analysis. The data were reported in 0.1 mg/kg steps up to a value of 10 mg/kg and then rounded to full 1 mg/kg steps – this causes an artifical "discretisation" of all data above 10 mg/kg.

3.2 The histogram

One of the most frequently used diagrams to depict a data distribution is the histogram. It is constructed in the form of side-by-side bars. Within a bar each data value is represented by an equal amount of area. The histogram permits the detection at one glance as to whether a distribution is symmetric (i.e. the same shape on either side of a line drawn through the centre of the histogram) or whether it is skewed (stretched out on one side – right or left skewed). It is also readily apparent whether the data show just one maximum (unimodal) or several humps (multimodal distribution). The parts far away from the main body of data on either side of the histograms are usually called the tails. The length of the tails can be judged. The existence or non-existence of straggling data (points that appear detached from the main body of data) at one or both extremes of the distribution is also visible at one glance. The Kola C-horizon data set provides some good examples.

Figure 3.3 shows four example histograms plotted for the variable Ba (aqua regia extraction) from the Kola C-horizon data set. The x-axis is scaled according to the range of the data, the starting point is usually a "nice looking" value slightly below the minimum value. Intervals along the x-axis are adapted to the number of classes it is required to display, and the number of classes is chosen such that the whole data range of the variable is covered (several rules of thumb exist for the "optimum" interval length or number of classes, one of the easiest is $\sqrt{n}$ for the number of classes, where n is the number of individuals/samples in the data set). The y-axis shows the number of observations in each class or, alternatively, the relative frequency of values in percent.

In theory, when dealing with a very large data set, the length of the intervals could be decreased so much by increasing the number of classes that the typical histogram steps disappear. This results in a plot of a smooth function, the density function. This smooth function is thought to represent the distribution from which the data are sampled by chance. If the data were drawn from a normal distribution the density function would take on the classical bell shape (see Figure 4.1).

One situation that often arises with environmental (geochemical) data is that the distributions are strongly right-skewed. In addition, extreme data outliers occur quite frequently. In such cases a histogram plotted with a "linear" scale may appear as a single bar at the left hand side and a far outlier at the right hand side of the graphic. Such histograms contain practically no information of value about the shape of the distribution (see upper left histogram, Figure 3.3). Statisticians will usually solve such problems via scaling the data differently. In the case at hand it appears advisable to reduce the influence of the high values and to focus on the main body of data that lie in the first bar of the histogram. One solution that meets these requirements is to scale the data logarithmically (lower right histogram, Figure 3.3). This re-scaling is called log-

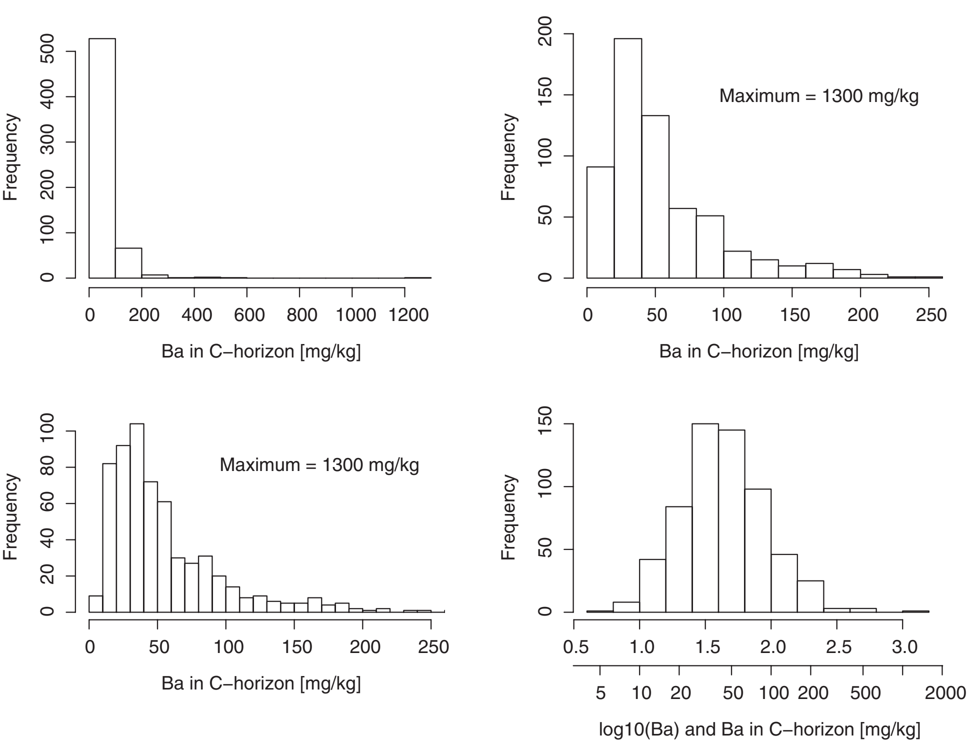

Figure 3.3 Histograms for Ba (aqua-regia extraction) from the C-horizon of the Kola data set. Upper left: original data; upper right: truncated data; lower left: truncated data as upper right but with double the number of classes; and lower right: log-transformed data. The lower right histogram has two scales to demonstrate the logarithmic transformation, the upper scale indicating the logarithms and the lower scale the original untransformed data

transformation of the data; a variety of data transformations are used in statistics to improve the visibility (or more generally, the behaviour) of the data (see Chapter 10). In geochemistry the log-transformation is the transformation that is most frequently used. Note that in the example the log-transformation results in an almost symmetrical distribution of the strongly right-skewed original data and that the range (i.e. maximum value – minimum value) of the data is drastically reduced because the influence of the outliers has been reduced. Plotting a histogram of the log-transformed values has, however, the disadvantage that the direct relation to the original data is missing (upper scale on the x-axis). This can be overcome by using a logarithmic scale for the x-axis (lower scale on the x-axis), this is the procedure adopted by DAS+R.

To permit scaling in the original data units, which are easier to read than a log-scale, it is also possible to plot a histogram for a certain part of the data only, e.g., for a certain data range (upper right histogram, Figure 3.3, range 0 to 250 mg/kg). In that case the outliers are not displayed; and the "real" maximum value should be indicated in the plot so that an unsuspecting reader does not gain an erroneous impression about the complete data distribution from the truncated histogram. Plotting more classes will result in a better resolution of the distribution (lower left) and will often considerably change the histogram's shape. Plotting too many classes will result in ugly gaps in the histogram.

The choice of starting point and class interval will substantially influence the appearance of the resulting histogram. This is shown in the lower left, where the only difference from the histogram in the upper right position is that the number of classes was increased from 13 to 26 (according to the simple $\sqrt{n}$ rule of thumb, 25 would be the "optimum" number of classes for the Kola Project C-horizon data). While the lower right histogram could be taken as an indication that a lognormal distribution is present (i.e. the log-transformed data follow a normal distribution), the more detailed histogram suggests that this may be in fact a multimodal distribution, approaching symmetry if log-transformed. The histogram could clearly be misused to demonstrate a "lognormal" distribution by reducing the number of classes (a statistical test of the distribution – see Chapter 9 – will indicate that even the log-transformed data for Ba do not follow a normal distribution). It may also be possible that the spikes in the distribution only appear because too many classes were chosen for the number of samples – the spikes might then be artefacts of the way the data were reported by the laboratory, e.g., to the nearest 1, 2, 5, or 10 mg/kg. This demonstrates how important the choice of the optimum interval length (number of classes) is when constructing histograms. Modern software packages will automatically use more sophisticated mathematical models than the $\sqrt{n}$ rule of thumb for that purpose (see, e.g., Venables and Ripley, 2002).

It can be concluded that by studying or displaying just one histogram, important aspects of the distribution may be missed. It may thus be necessary to plot a series of histograms for a variable (as above), or the histogram should be augmented with some other graphics displaying the data distribution and showing additional features.

One possibility is to combine the histogram with the one-dimensional scatterplot (Figure 3.4).

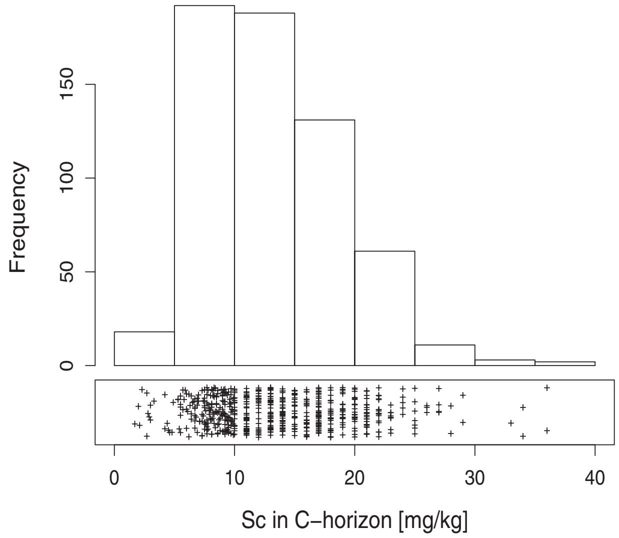

Figure 3.4 One-dimensional scatterplot for Sc (total concentrations as determined by INAA) in the Kola C-horizon data set, combined with the histogram

3.3 The density trace

The density trace (kernel density estimate – Scot, 1992) is on first inspection a “smoothed line, tracing the histogram” (Figure 3.5). It is another approximation of the underlying density function of the data. Each point along the curve is calculated from data within a defined bandwidth using a “weight function” (in most packages; including R, these parameters are chosen by default but can be changed manually). Choice of bandwidth and weight function will crucially determine how the final density trace appears (compare Figure 3.5 density trace upper right with density trace lower left). Although at first glance the density trace is a more “objective” graphic to display the data distribution than the histogram, it can also be manipulated substantially. Density traces are better suited for comparing data distributions than histograms because they can easily be plotted on top of one another (e.g., in different line styles or colours – see Chapter 8). The plots (Figure 3.5) are for the distribution of Ba from the Kola C-horizon data set previously displayed as histograms (Figure 3.3). Compared to the histogram not much information is gained. More or less the same problems are experienced as above when plotting the histograms. Note that for plotting the density trace of the log-transformed data in the lower right display, a log-scale was used to preserve the direct relation to the original data range (Figure 3.5).

A density trace can also be combined with a histogram to give a more realistic impression of the data distribution.

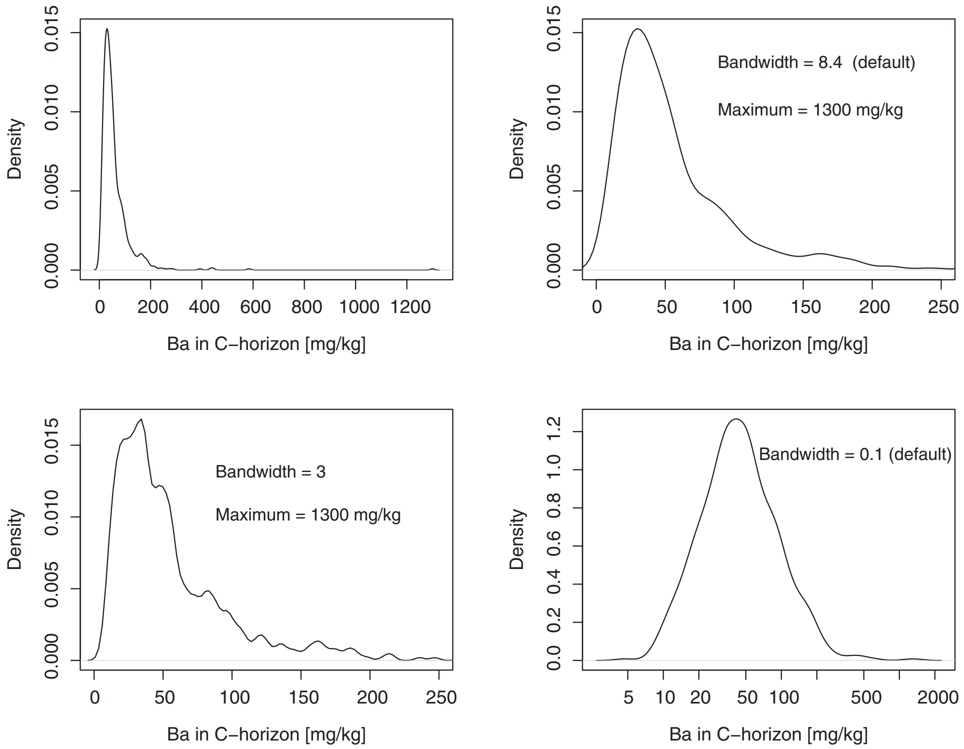

Figure 3.5 Density traces for Ba (aqua regia extraction) from the C-horizon of the Kola data set. Compare to Figure 3.3

3.4 Plots of the distribution function

Plots of the distribution function, e.g., the cumulative probability plot, were originally introduced to geochemists by Tennant and White (1959), Sinclair (1974, 1976) and others. They are one of the most informative graphical displays of geochemical distributions. There exist several different plots of the distribution function that all have their merits.

3.4.1 *Plot of the cumulative distribution function (CDF-plot)*

Histogram and density trace are based on the density function. However, the percentage of samples plotting above or below a certain data value x could be of interest. The percentage equals an area under the curve of the density function. Percentages can also be expressed as probabilities, i.e. the probability that a value smaller than or equal to the chosen data value x_1 (Figure 3.6) appears is p_1. This probability can be taken as a point in a new plot, where the data values along the x-axis are plotted against probabilities along the y-axis (Figure 3.6). By varying the chosen value x, additional probabilities are obtained that can then be drawn as new

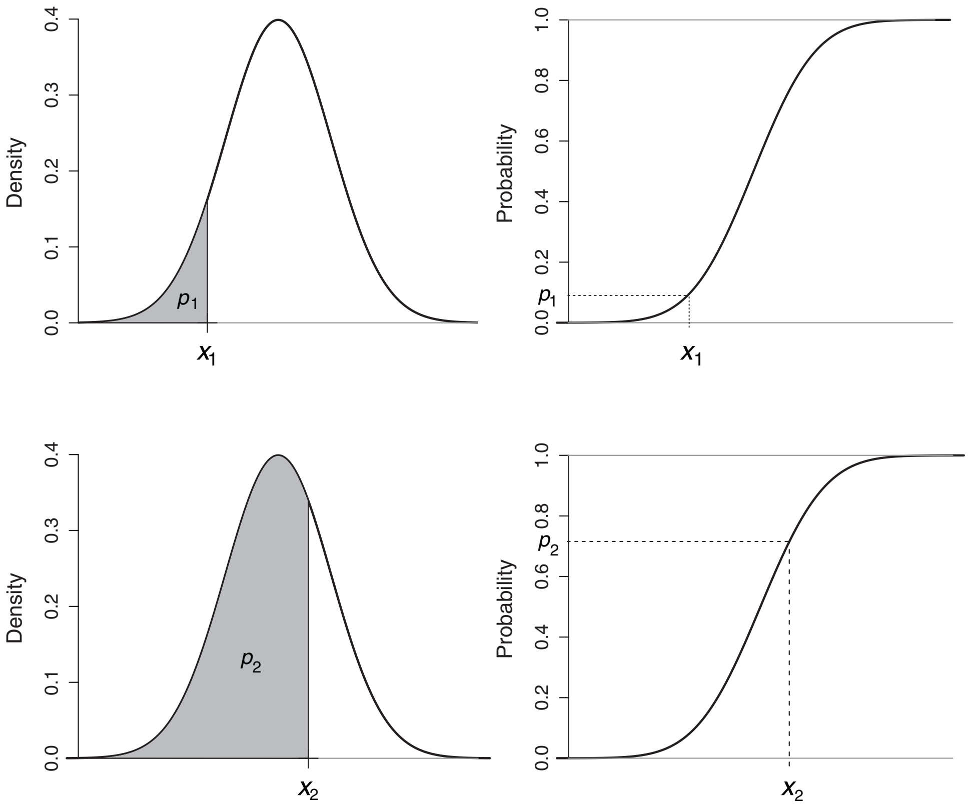

Figure 3.6 Construction of the cumulative distribution function (CDF-plot). The graphic to the left shows the density function and the graphic to the right shows the resulting cumulative distribution function

points in the right hand plot for the corresponding x values (e.g., p_2 for the data value x_2). If this is carried out for the whole x-range, a new plot is generated, the cumulative distribution function or CDF-plot (right hand diagrams).

If the data follow a normal distribution (bell shape in the left hand plot), the distribution function will have a typical symmetrical, sigmoidal S-shape (right hand plot).

3.4.2 Plot of the empirical cumulative distribution function (ECDF-plot)

Just like trying to approach the density function via plotting histograms or density traces, it is possible to approach the distribution function by displaying the empirical cumulative distribution function (the ECDF-plot). The ECDF-plot is a discrete step function that jumps with each data value by $\frac{1}{n}$, where n is the number of data points. As n becomes increasingly large, to infinity, this function will approximate the underlying distribution function.

The ECDF-plot shows the variable values along the x-axis (either scaling according to non-transformed data or following a transformation, e.g., logarithmic). The y-axis shows the probabilities of the empirical cumulative distribution function between 0 and 1 (which could also be expressed as percentages 0–100 percent).

As an example, the gold (Au) results for the Kola Project C-horizon soils are displayed (Figure 3.7). The histogram and density trace demonstrate that the distribution is not normal but extremely right skewed, as a result the typical S-shape mentioned above for the CDF-plot (Figure 3.7, upper right) is not present. Two very high values dominate the plot. The ECDF-plot of the original data is still informative because it graphically displays the distance of these two high values from the main body of the data. However, to provide a more useful visualisation of the main body of data, a log-transformation should be applied. Following log-transformation, the histogram and density trace still show a slight right skew (Figure 3.7, middle left). The resulting ECDF-plot begins to display an S-shape (Figure 3.7, middle right), however, the right skew is still clearly reflected. Instead of using a log-transform, the plotting range of the data could be limited to focus on the main body of data (Figure 3.7, lower left and right). This permits the study of the main body of data in far greater detail than in the upper plot. Displaying the data in this manner requires that the viewer be reminded that the diagrams are displaying only a limited part of the range of the complete data set.

One of the main advantages of the ECDF-plot is that every single data point is visible. A geochemist would now start to search for any unusually high (or low) values and breaks in the distribution. Very high values might, for example, indicate a mineral deposit or anthropogenic contamination source. A break in the distribution could be caused by the presence of different natural factors like geology, weathering and climate, or different contamination sources influencing the data. In a large regional survey, the data may reflect both multiple natural processes and anthropogenic sources. The ECDF-plot can be used with advantage to identify classes for geochemical mapping that have a direct relation to the underlying statistical data distribution via assigning class boundaries to these breaks (see Chapters 5 and 7).

3.4.3 The quantile-quantile plot (QQ-plot)

It is often quite difficult to judge whether the S-shape as displayed in the ECDF-plot indicates a normal or a lognormal (or some other) distribution. When it is necessary to judge the underlying distribution of the empirical data, a different plot is required, and it is advantageous to change

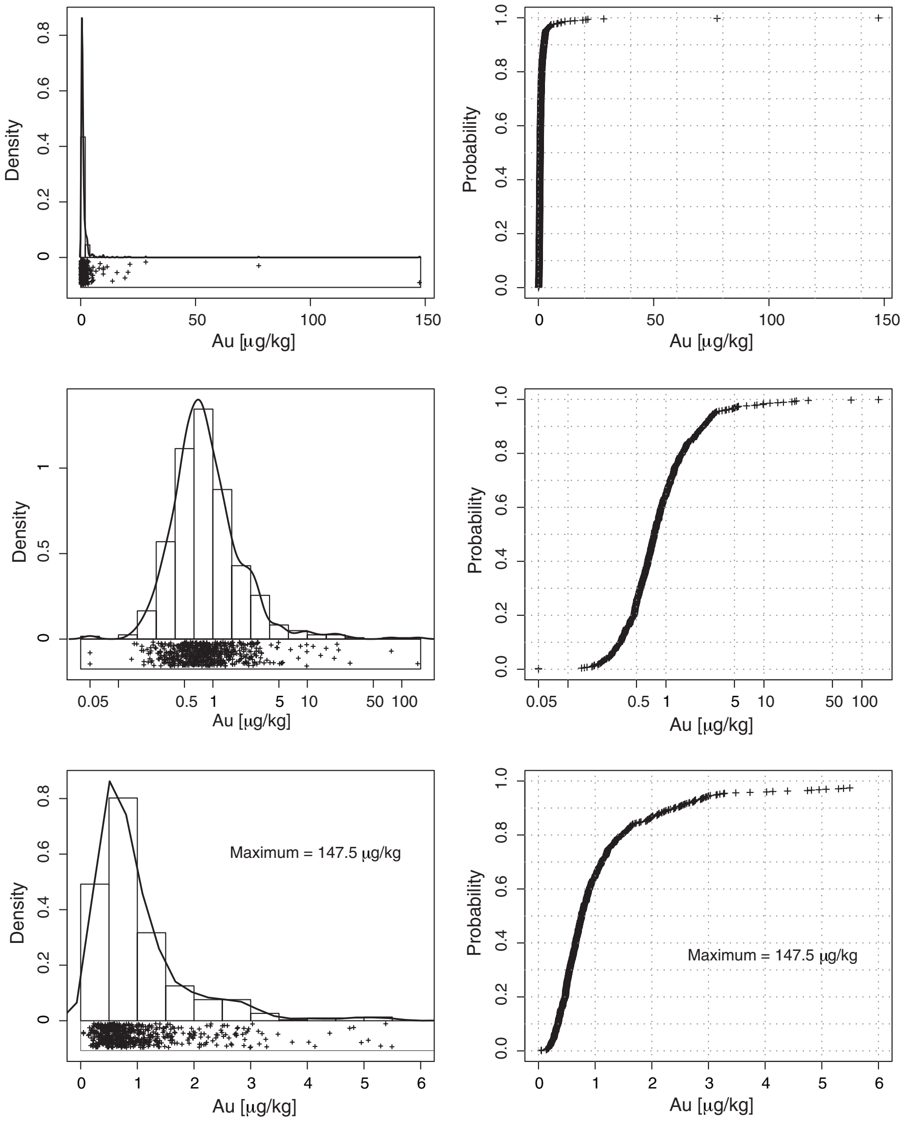

Figure 3.7 Combination of histogram, density trace, and one-dimensional scatterplot (left hand side) and ECDF-plot (right hand side) used to study the distribution of Au in the Kola Project C-horizon soil data set. Upper diagrams: original scale, middle diagrams: log-scale, lower diagrams: truncated plotting range

the y-axis scaling while keeping the x-axis the same. It is easiest to detect changes from an expected distribution if the points fail to follow a straight line. Thus the best approach is to change the y-axis in such a way that the plotted points fall on a straight line if they follow the assumed distribution (normal, lognormal, or any other). To achieve this, the cumulative distribution function must be transformed. A non-linear transformation, the inverse of the expected distribution function of the y-axis, is used for this purpose. The values of the inverse

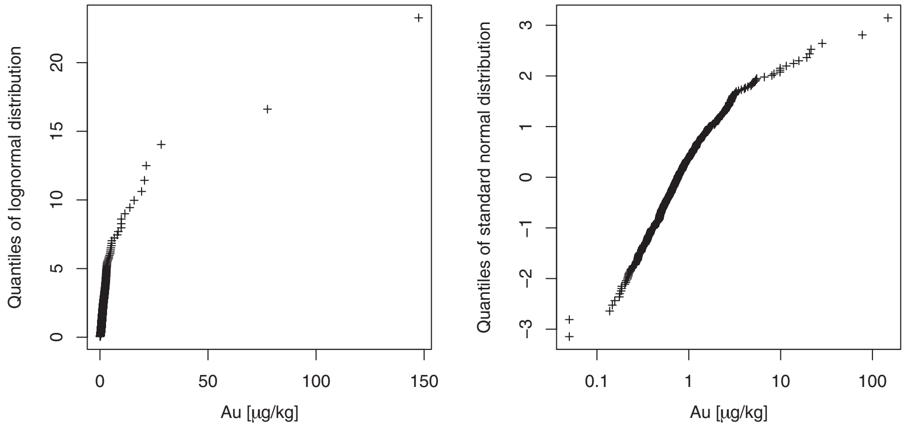

Figure 3.8 QQ-plots for Au in the Kola Project C-horizon soil data set as shown in Figure 3.6. Left hand side: *x*-axis non-transformed data, *y*-axis quantiles of the lognormal distribution; right hand side: *x*-axis log-transformed data, *y*-axis quantiles of the normal distribution

of the expected distribution function are called quantiles. The sorted data values along the *x*-axis can be considered as the quantiles of the empirical data distribution.

Quantiles are the linear measure underlying the non-linear distribution of the probabilities. Quantiles are expressed as positive and negative numbers, they can be compared to standard deviation units (see Chapter 10). This plot is called the quantile-quantile (QQ-) plot because quantiles of the data distribution are plotted against quantiles of the hypothetical normal or lognormal distribution. Figure 3.8 shows the QQ-plot for the Kola C-horizon Au data. In the left hand diagram the original data are plotted along the *x*-axis. In the right hand diagram the log-transformed values are plotted along the *x*-axis to reduce the impact of the two extreme values. By changing the scaling of the *y*-axis, it is possible to check the distribution for log-normality in both diagrams. When plotting the original data, the *y*-axis is scaled according to the quantiles of the lognormal distribution. When plotting the log-transformed data, the *y*-axis is scaled according to the quantiles of the normal distribution. This is possible because the lognormal distribution is simply the logarithm of the normal distribution. For the standard normal distribution 0 corresponds to the MEDIAN and the RANGE $[-1, 1]$ indicates the inner two-thirds of the data distribution. Unfortunately, the construction and interpretation of the QQ-plot for different data distributions requires statistical knowledge about these data distributions and their quantiles that is far beyond the scope of this book. The interested reader can consult Chambers *et al.* (1983).

In general, when checking for other distributions (e.g., Poisson or gamma distributions), the *x*-axis is always scaled according to the original data, and only the *y*-axis is changed according to the quantiles of the expected distribution.

When plotting empirical data distributions in the QQ-plot, it may still be quite difficult to determine whether the data points really follow a straight line. A simple and robust way to plot a straight line into the diagram is to connect first and third quartiles of both axes. In addition to the straight line 95 percent confidence intervals around that line can also be constructed.

The *confidence interval* encloses 95 percent of the data points that could have been drawn from the hypothetical distribution. These limits can be used to support a graphical decision

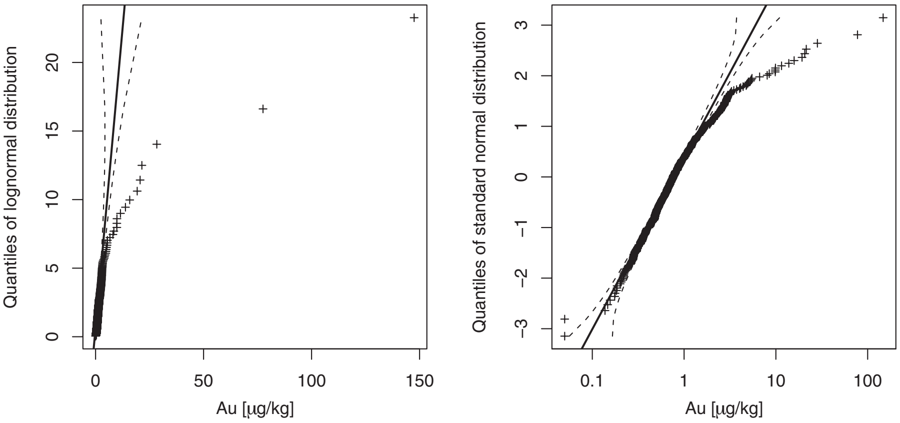

Figure 3.9 QQ-plots as shown in Figure 3.8 with a straight line indicating the hypothetical distribution and the 95 percent confidence intervals as dashed lines

as to whether the empirical data originate from the hypothetical distribution. The width of the confidence interval varies because the "allowed" deviation of data points from the straight line depends on the number of samples. In the example plot (Figure 3.9) the upper tail of the distribution clearly deviates from log-normality.

3.4.4 *The cumulative probability plot (CP-plot)*

The examples above demonstrate that a scale forcing the points in the plot to follow a straight line is useful. Because in the QQ-plot the scaling of the *y*-axis is different from distribution to distribution, it would be much easier if the *y*-axis were expressed in probabilities. In the pre-computer times such a plot was constructed on probability paper that was especially designed for a normal (or lognormal) distribution. This procedure was originally introduced to geochemists by Tennant and White (1959), Sinclair (1974, 1976) and others. The graph in the plot is exactly the same as in the QQ-plot. When using a computer, it is no longer necessary to limit the QQ-plot to a normal (lognormal) distribution. Any other distribution could be introduced for scaling the *y*-axis. If the scale on the *y*-axis is expressed in probabilities rather than in quantiles, the plot is generally named the cumulative probability plot (CP-plot) (Figure 3.10). Note that when checking for normality, the probabilities as expressed along the *y*-axis can never reach zero or one because these values would correspond to quantiles of minus infinity or plus infinity, respectively.

The CP-plot with log-scale for the data (Figure 3.10, right hand side) is especially useful because it allows direct visual estimation of the MEDIAN (50^{th} percentile) or any other value from the *x*-axis or the percentage of samples falling below or above a certain threshold (e.g., a maximum admissible concentration (MAC)) from the *y*-axis. It also allows the assigning of a percentage to any break in the curve; in the example several breaks in the Au data are visible, the first one occurs at about 85 per cent (c. 1.5 μg/kg Au), the next at 95 per cent (c. 3 μg/kg Au) of the data. Just as in the QQ-plot, the straight line to judge whether the empirical data follow the hypothetical distribution can be shown.

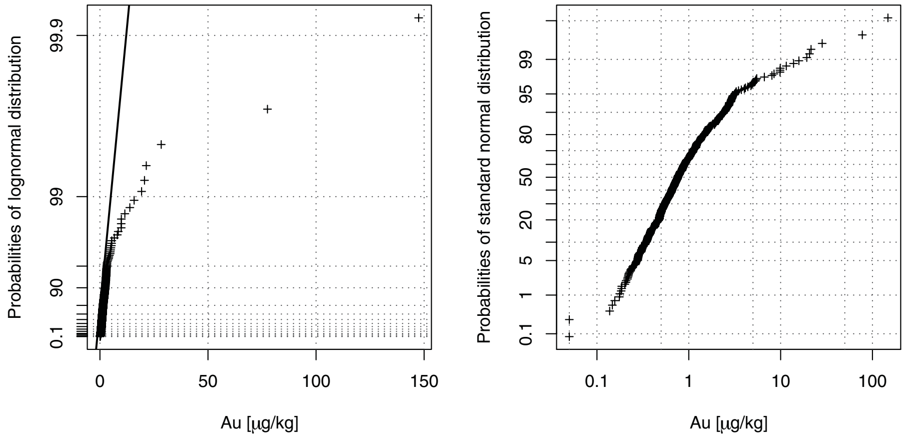

Figure 3.10 CP-plots for Au in the Kola Project C-horizon soil data set as shown in Figures 3.6, 3.7 and 3.8. Left hand side: *x*-axis non-transformed data, *y*-axis probabilities in per cent; right hand side: *x*-axis log-transformed data, *y*-axis probabilities in per cent

3.4.5 The probability-probability plot (PP-plot)

Yet another version of these diagrams is the probability-probability (PP-) plot. Instead of plotting quantiles of the hypothetical distribution against the quantiles of the data distribution at fixed probabilities (QQ-plot), the probability of the hypothetical distribution is plotted against the probability of the empirical data distribution at fixed quantiles (PP-plot). The advantage of the PP-plot is that while the QQ-plot and the CP-plot can be dominated by extreme values, these cannot dominate the PP-plot because of their low probability. The PP-plot will thus focus the attention on the main body of data. In the example plot (Figure 3.11) an additional flexure

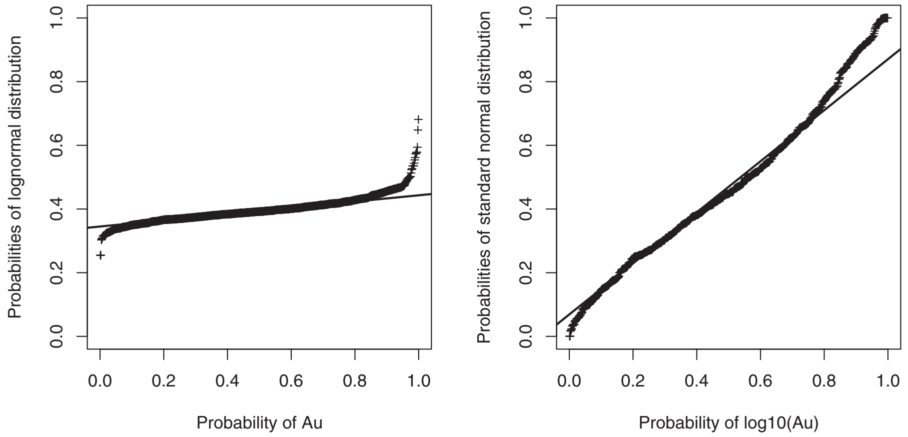

Figure 3.11 PP-plots for Au in the Kola Project C-horizon soil data set as shown in Figures 3.7, 3.8, 3.9, and 3.10. Left hand side *x*-axis: probabilities referring to the non-transformed data, *y*-axis: probabilities for lognormal distribution; right hand side *x*-axis: probabilities referring to the log-transformed data, *y*-axis: probabilities for the normal distribution

at about 20 percent of the Au distribution becomes visible. This is visible as a weak flexure in the ECDF-plot (Figure 3.7) at 0.5 μg/kg Au. The main disadvantage of the PP-plot is that the relation to the original data is completely lost, whereas it is retained in the ECDF- and CP-plot. It is of course possible to use the PP-plot in combination with the CP-plot to identify the data value for 20 percent. A combination of CP-plot and ECDF- or PP-plot may thus be quite powerful in obtaining a more complete picture of the data distribution. Single extreme values are of course hardly visible in the PP-plot though they are in the ECDF-plot. Just as in the QQ- and CP-plot, a straight line can be introduced in the PP-plot to check for agreement with a hypothetical distribution.

3.4.6 Discussion of the distribution function plots

Depending on the empirical data distribution of a given variable, the different versions of these plots all have their merits, especially when the task is to detect fine structures (breaks or flexures) in the data. If only one plot is to be presented, it is advisable to look first at the different possibilities and then select the most informative for the variable under study because this will depend on the actual data distribution of each variable. In applied geochemistry the CP-plot with logarithmic x-axis is probably the most frequently used (Figure 3.10, right). In combination with the less frequently used ECDF-plot and the almost never used PP-plot (Figure 3.11, right), it holds the potential to provide a very realistic picture of the complete data distribution.

As mentioned above, one of the main advantages of these diagrams is that each single data value remains observable. The range covered by the data is clearly visible, and extreme outliers are detectable as single values. It is possible to directly count the number of extreme outliers and observe their distance from the core (main mass) of the data.

When looking at a selection of these plots for As (Figure 3.12), several data quality issues can be directly detected; for example, the presence of discontinuous data values at the lower end of the distribution. Such discontinuous data are often an indication of the method detection limit or too-severe rounding, discretisation, of the measured values reported by the laboratory. The PP-plot corresponding to the log-transformed data shows most clearly how serious an issue this data discretisation due to rounding of the values by the laboratory can become.

Values below the detection limit, set to some fixed value, are visible as a vertical line at the lower end of the plots, and the percentage of values below the detection limit can be visually estimated. The CP-plot with logarithmic scale (middle right figure) displays this best. The detection limit for As was 0.1 mg/kg, about two per cent of all values plot below the detection limit (Figure 3.12). From 0.1 to 1 mg/kg the As values were reported in 0.1 mg/kg steps – obviously a too-harsh discretisation for the data at hand, causing artificial data structures (Figure 3.12). The presence of multiple populations results in slope changes and breaks in the plots (Figure 3.12).

3.5 Boxplots

The boxplot is one of the most informative graphics for displaying a data distribution. It is built around the MEDIAN (see Chapter 4), which divides any data set into two equal halves.

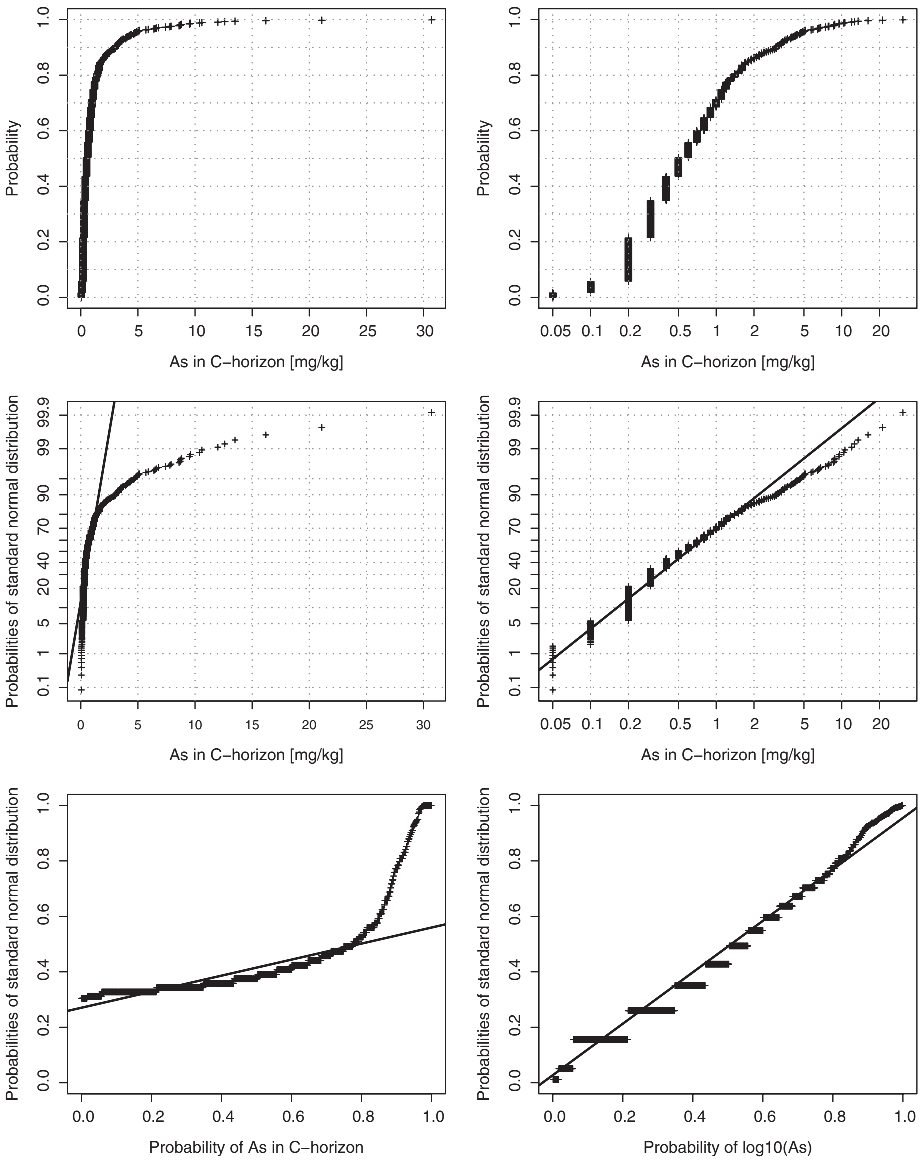

Figure 3.12 Six different ways of plotting distribution functions. Upper row: empirical cumulative distribution function plots (ECDF-plots); middle row: cumulative probability plots (CP-plots); lower row: probability-probability plots (PP-plots). Left half of diagram: data without transformation; right half of diagram: plots for log-transformed data

3.5.1 The Tukey boxplot

Tukey (1977) introduced the boxplot to exploratory data analysis. The construction of the Tukey boxplot is best demonstrated using a simple sample data set, consisting of only nine values:

2.3 2.7 1.7 1.9 2.1 2.8 1.8 2.4 5.9.

The data are sorted to find the MEDIAN:

1.7 1.8 1.9 2.1 **2.3** 2.4 2.7 2.8 5.9.

After finding the MEDIAN (2.3), the two halves (each of the halves includes the MEDIAN) of the data set are used to find the "hinges", the MEDIAN of each remaining half:

1.7 1.8 **1.9** 2.1 **2.3** 2.4 **2.7** 2.8 5.9.

These upper and lower hinges define the central box, which thus contains approximately 50 percent of the data. In the example the "lower hinge" (LH) is 1.9, the "upper hinge" (UH) is 2.7. The "inner fence", a boundary beyond which individuals are considered *extreme values* or potential *outliers*, is defined as the box extended by 1.5 times the length of the box towards the maximum and the minimum. This is defined algebraically, using the upper whisker as an example, as

Upper inner fence (UIF) $= \mathrm{UH}(x) + 1.5 \cdot \mathrm{HW}(x)$.
Upper whisker $= \max(x[x \leq \mathrm{UIF}])$.

where HW (hinge width) is the difference between the hinges (HW = upper hinge–lower hinge), approximately equal to the interquartile range (depending on the sample size), i.e. Q3−Q1 ($75^{th} - 25^{th}$ percentile); and the square brackets, [. . .] indicate the subset of values that meet the specified criterion.

The calculation is simple for the example data:

Hinge width, $\mathrm{HW} = \mathrm{UH} - \mathrm{LH} = 2.7 - 1.9 = 0.8$.
Lower inner fence, $\mathrm{LIF} = \mathrm{LH} - (1.5 \cdot \mathrm{HW}) = 1.9 - (1.5 \cdot 0.8) = 0.7$.
Upper inner fence, $\mathrm{UIF} = \mathrm{UH} + (1.5 \cdot \mathrm{HW}) = 2.7 + (1.5 \cdot 0.8) = 3.9$.

By convention, the upper and lower "whiskers" are then drawn from each end of the box to the furthest observation inside the inner fence. Thus the lower whisker is drawn from the box to a value of 1.7, the lower whisker and minimum value are identical, and the upper whisker is drawn to the value of 2.8. Values beyond the whiskers are marked by a symbol, in the example the upper extreme value of (5.9) is clearly identified as an extreme value or data outlier (see Chapter 7) and is at the same time the maximum value.

Figure 3.13 shows a "classical" Tukey boxplot for Ba in the Kola C-horizon samples. No lower extreme values or outliers are identified. The lower inner fence is lower than the minimum value, and thus the lower whisker terminates at the location of the minimum value. The upper inner fence and termination of the upper whisker fall together in this example, and all values to the right of the upper whisker are identified as upper extreme values or outliers.

In summary, the Tukey boxplot – one of the most powerful EDA graphics – shows in graphical form:

- the "middle" (MEDIAN) of a given data set, identified via the line in the box;
- spread (see Chapter 4) by the length of the box (the hinge width);
- skewness (see Chapter 4) by the symmetry of the box and whisker extents about the median line in the box;
- kurtosis (see Chapter 4) by the length of the whiskers in relation to the width of the box;
- the existence of extreme values (or outliers – see Chapter 7), identified by their own symbol.

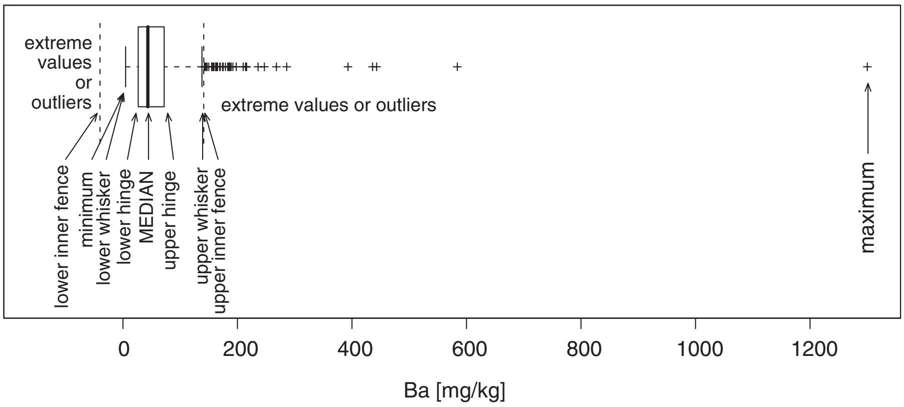

Figure 3.13 Tukey boxplot for Ba in the Kola C-horizon soil samples

Furthermore, because the construction of the boxplot is based on quartiles, it will not be seriously disturbed by up to 25 percent of "wild" data at either end of the distribution. It is not even seriously influenced by widely different data distributions (Hoaglin *et al.*, 2000).

3.5.2 The log-boxplot

It is important to recognise that the calculation of the whiskers in the above formula assumes data symmetry, lack of which is easily recognised by the median line not being close to the middle of the box. The recognition of extreme values and outliers is based on normal theory, and the standard deviation (SD) of the distribution is estimated via the hinge width (HW) or interquartile range (IQR). So for the estimate of SD to be appropriate, there has to be symmetry in the middle 50 percent of the data (see Chapter 4). Thus for strongly right-skewed data distributions, as frequently occur in applied geochemistry, the Tukey boxplot based on untransformed data will tend to seriously underestimate the number of lower extreme values and overestimate the number of upper extreme values.

Figure 3.13 demonstrates that the boxplot detects a high number of upper extreme values for the Ba data and no lower extreme values. The reason for this is due to the right-skewed data distribution, indicated by the MEDIAN falling only one-third of the hinge width (HW) above the lower hinge (LH). This feature was also apparent in the histogram and density trace (Figures 3.3 and 3.5). These figures demonstrate that the Ba distribution approaches symmetry when the data are log-transformed. The Tukey boxplot of the log-transformed data will thus be suitable for providing a realistic estimate of the extreme values at both ends of the data distribution. Figure 3.14 shows the Tukey boxplot for the log-transformed Ba data. As expected, the number of upper extreme values is drastically reduced, and one lower extreme value is now identified (Figure 3.14, upper diagram). It is of course possible to plot a log-scale for the original data to regain the desirable direct relationship (Figure 3.14, middle). Because the boxplot is reliable for symmetric distributions, it is appropriate to calculate the values for the whiskers for the log-transformed distribution and then back-transform the values for the fences to the original data scale (Figure 3.14 lower diagram). Note that the MEDIAN and hinges will not be changed by log- and back-transformation because they are based on order statistics. This version of the boxplot is called the log-boxplot and should be used when the

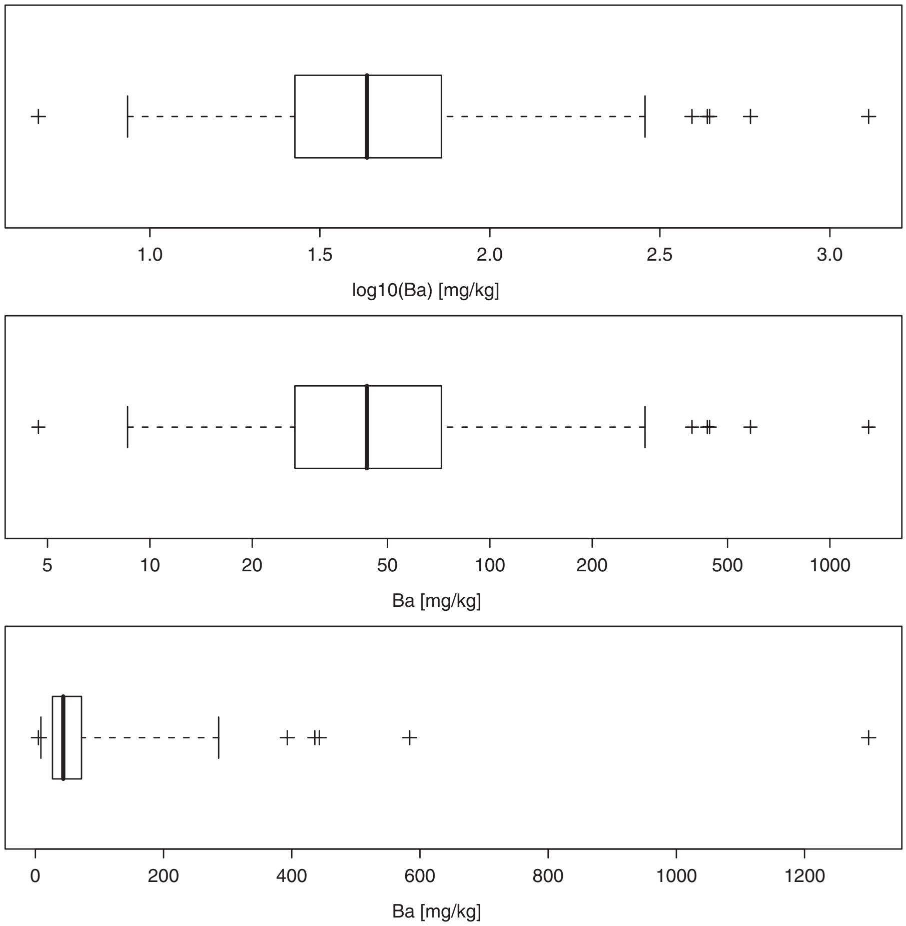

Figure 3.14 Boxplots for Ba. Upper diagram: the boxplot of the log-transformed data. Middle diagram: the same but scaled according to the original data. Lower diagram: log-boxplot of the original data with whiskers calculated according to the symmetrical log-transformed data and back-transformed to the original data scale – compare with the Tukey boxplot in Figure 3.13

original data are strongly skewed and it is still desirable to preserve the data scale. Comparison of Figure 3.13 with Figure 3.14 (lower diagram) demonstrates that the limits of the box are not changed; however, the extent and position of the whiskers and the number of extreme values has changed dramatically.

In conclusion, the Tukey boxplot should not be applied to strongly skewed data distributions without an appropriate data transformation because it will result in a wrong impression about the upper and lower extreme values. Thus for applied geochemical data, the data distribution should always be checked for symmetry before drawing either a Tukey boxplot or the log-boxplot. In the majority of cases the log-boxplot (Figure 3.14, lower plot) will be better suited to identify the number and position of extreme values.

If the log-transformed data still deviate from symmetry, versions of the boxplot exist that can deal with strongly skewed data sets and will still provide useful fences for extreme values at both ends of the data distribution. They are based on a robust measure of skewness (Section 4.4) for the calculation of the fences for the lower and upper whiskers (Vandervieren and Hubert, 2004).

3.5.3 The percentile-based boxplot and the box-and-whisker plot

The data symmetry problems with the Tukey boxplot are probably the reason why some workers prefer to use a modified version, the percentile-based boxplot, where all definitions are based on percentiles (see Chapter 4): MEDIAN and 25^{th} and 75^{th} percentile for the box and 2^{nd} (or 5^{th} to 20^{th}) and 98^{th} (or 80^{th} to 95^{th}) percentile for the whiskers.

However, when using a percentile-based boxplot, one of the major advantages of the Tukey boxplot is lost, i.e. the "automatic" identification of extreme values. The percentile boxplot will always identify a certain percentage of extreme values while it is possible that no extreme values will be identified with the Tukey boxplot. When studying boxplots, it is essential to be aware of the exact conditions used for their construction.

Because the Tukey boxplot, log-boxplot, and percentile-based boxplot all look the same to the unsuspecting reader (compare Figures 3.13, 3.14 and 3.15), it should be good practice to explain in the figure caption which version of the plot was used.

In the box-and-whisker plot (see, e.g., Garrett, 1988) the whiskers are drawn to stated percentiles, e.g., the 5^{th} and 95^{th}, and the minima and maxima plotted as crosses. No attempt is made to identify extreme values. The box-and-whisker plot is simply a graphical summary of the data based on the order statistics (percentiles) with no assumptions concerning the

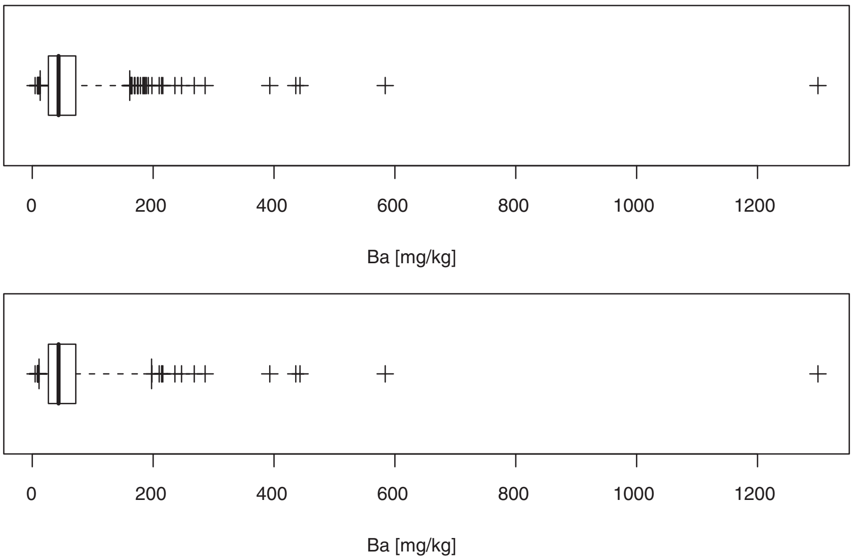

Figure 3.15 Percentile-based boxplot using the 5^{th} and 95^{th} (upper boxplot) and 2^{nd} and 98^{th} percentile (lower boxplot) for drawing the whiskers. Variable Ba, Kola Project C-horizon; compare with Figures 3.13 and 3.14

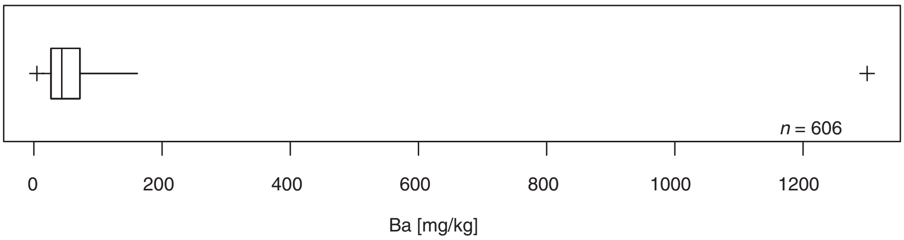

Figure 3.16 Box-and-whisker plot of the variable Ba, Kola Project C-horizon

underlying statistical model. Figure 3.16 is the box-and-whisker plot for Ba in Kola Project C-horizon soils. The resulting plot is the same as the boxplot shown in Figure 3.15, upper boxplot, without identifying all outliers or extreme values.

3.5.4 The notched boxplot

Often the information included in the Tukey boxplot is extended by adding an estimate of the 95 percent confidence bounds on the MEDIAN. This leads to a graphical test of comparability

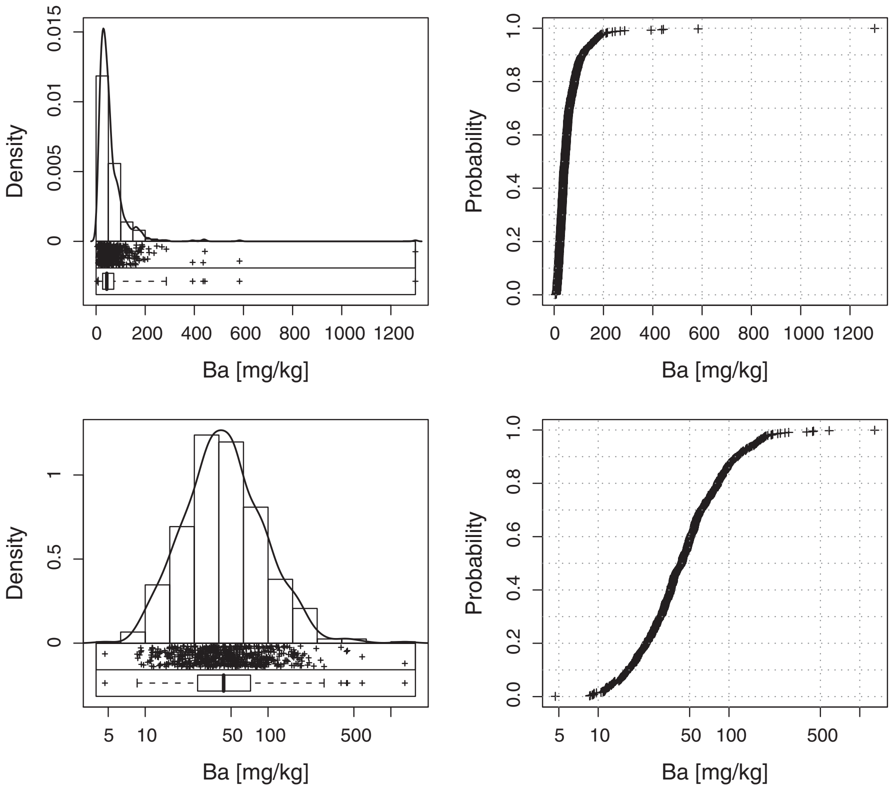

Figure 3.17 Histogram, density trace, one-dimensional scatterplot, and boxplot in just one display, combined with the ECDF-plot. Variable Ba, Kola Project C-horizon. Upper diagrams: original data (with log-boxplot); lower diagrams: log-transformed data

– much like the more formal t-test (see Chapter 9) – of MEDIANS via notches in the boxplot. This use of boxplots is discussed in Section 9.4.1.

3.6 Combination of histogram, density trace, one-dimensional scatterplot, boxplot, and ECDF-plot

Several of the plots named so far – one-dimensional scatterplot, histogram, density trace and boxplot – can advantageously be combined into just one display (Reimann, 1989). Figure 3.17 (left) shows this combined graphic for Ba. The ECDF-plot is another graphic that will reveal interesting properties of the data distribution (Figure 3.17, right). In contrast to the QQ-, CP-, or PP-plot, the ECDF-plot is not based on the assumption of any underlying data distribution. It is thus ideally suited as an exploratory data analysis tool in the first stages of data analysis.

In combination these graphics provide an excellent first impression of the data, as each of them highlights a different aspect of the data distribution (Figure 3.17). Figure 3.17 shows that Ba is strongly right skewed for the original data. It displays an almost symmetrical distribution for the log-transformed data. For the log-transformed data both the one-dimensional scatterplot and ECDF-plot show some minor disturbances at the lower end of the Ba distribution. For further work with most statistical methods requiring symmetrical data, the log-transformed

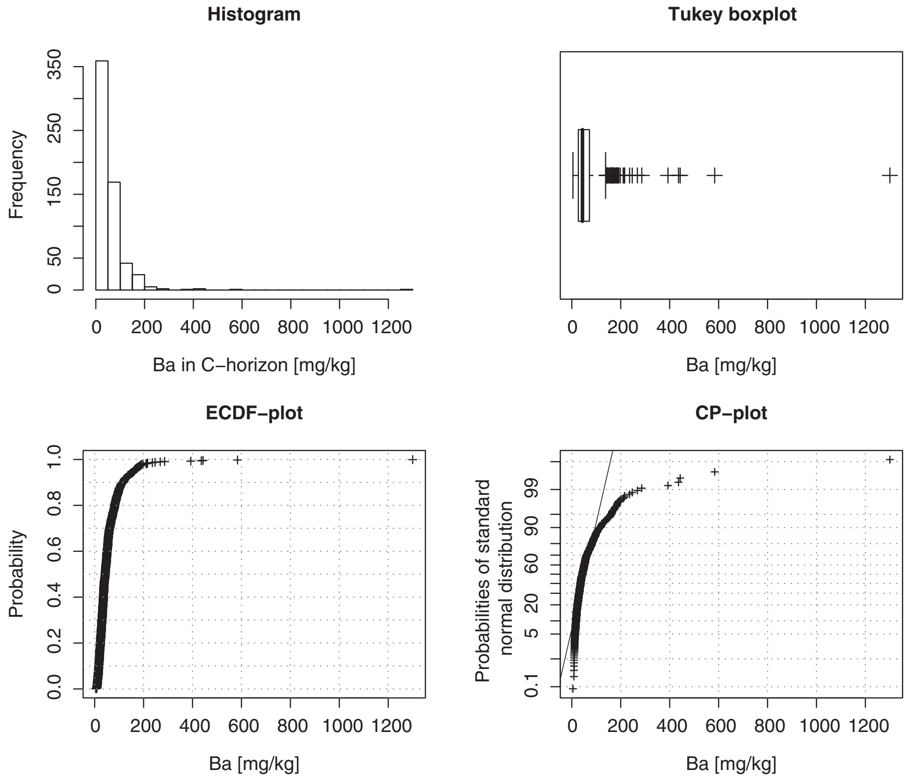

Figure 3.18 Combination of histogram, Tukey boxplot, ECDF-, and CP-plot for the variable Ba, Kola Project C-horizon

values of this variable should obviously be used. One could also conclude that the log-boxplot should be used for further work with this variable for a realistic identification of extreme values whenever the data are kept in the original scale.

3.7 Combination of histogram, boxplot or box-and-whisker plot, ECDF-plot, and CP-plot

A different combination of the plots in this chapter has also proven informative in a single display (Figure 3.18). Using the data for Ba in Kola C-horizon soils, the histogram in the upper left is the same as in Figure 3.17. The upper right display is either a Tukey boxplot or a box-and-whisker plot, the user's choice. The lower left display is an ECDF-plot and the lower right a CP-plot. The ECDF-plot and CP-plot permit easy inspection of the middle and extreme parts of the data distribution, respectively. The choice of Tukey boxplot or box-and-whisker plot depends on whether the user wishes to identify potential outliers or simply have an order statistics based replacement for the histogram. The plotting of the Tukey boxplot above the CP-plot permits easy comparison of the two plots. Figure 3.18 indicates that all plots are dominated by some few extreme values. Thus the whole display should be plotted using logarithmic scaling (Figure 3.19).

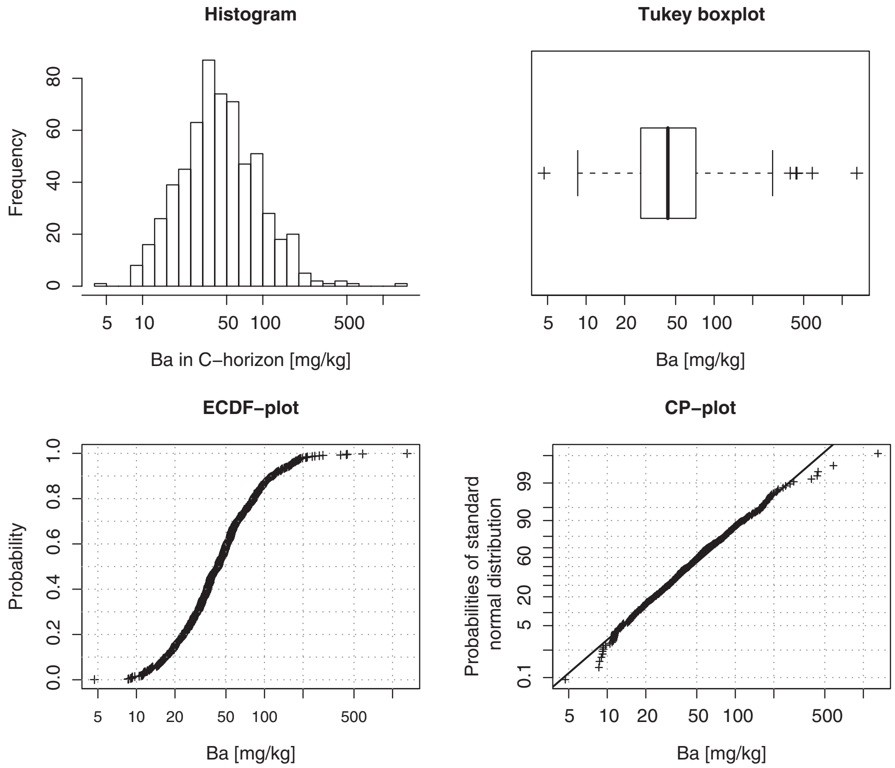

Figure 3.19 The same plot as Figure 3.18 using logarithmic scales

3.8 Summary

When working with a new data file it is highly advisable to study the data distribution in detail before proceeding to all subsequent statistical analyses. Many statistical methods are based on assumptions about the data distribution. A documentation of the data distribution will help to decide which statistical methods are appropriate for the data at hand. A histogram alone is not sufficient to get a good impression of the data distribution. The impression gained from histogram and density trace alone depends strongly on the choice of parameters. A combination of different graphics will often provide greater insight into the data distribution. All variables should thus be documented in a combination of summary plots, e.g., histogram combined with density trace, boxplot and one-dimensional scatterplot and ECDF- or CP-plots.

It is advisable to have copies of a number of distribution graphics for all variables under study at hand for easy reference when more advanced data analysis methods are applied.