# Chapter 3

# Customer Data

As underlined in the previous chapter, customer load values play an important role in every distribution network study. They are needed for any load flow study for example. Load flow is a fundamental tool for the analysis of distribution systems and is used in the operational as well as planning stages. Certain applications in distribution automation also require repeated load flow solutions. Thus, it is very important to solve the load flow problem as efficiently as possible. To carry out load flow studies related to distribution networks, the values of power consumption have to be known at each node. The values of power consumption at each node also influence different distribution network factors such as the size of distribution transformers and conductors, peak load demand period and capacitor banks.

In distribution networks, typically, there are very few available real-time measurements of load values. This means that many values of nodal power consumption are not recorded. To provide data for the studies mentioned previously, it is necessary to estimate what the load looks like on these different parts of the circuit. Each study requires different types of input consumption values. For load flow studies, the values of nodal hourly power consumption are needed. For the sizing studies, the values of nodal maximum power consumption are important. To estimate the missing values at some nodes, data recorded at other nodes are utilized. In this chapter, a large collection of real data is described first. It is seen how the database is processed in order to use it more efficiently in the later intervals calculations. In a second part, the data will be analyzed and statistical conclusions will be inferred.

## 3.1-Background Information

In general, the only information commonly available regarding loads at location other than distribution substations and major equipment installations, consists of customer kWhr consumption values.

In fact, annual load data are recorded for every single demand point monitored on a circuit. A demand point can account for one or several single customer load demands. In the present case, each point corresponds to one single customer. At that point, one value per hour is

recorded, yielding 24*365 = 8760 data per demand point and per year. Data is gathered each month by service personnel using electronic readers and are stored on a mainframe computer.

In order to carry out data analysis, it is required to separate customers into different *classes* that are statistically homogeneous. That is to say, it is necessary to separate customers who have electric heating (or cooling) from those who have not electric heating, and from commercial loads. This can easily done thanks to the wide availability of electronic demand recorders.

A real database belonging to the three different classes mentioned above has been provided to us. *These data are supposed to be representative from the customer's population*. As it will be used in later calculations, we need to correctly process it. Next, a data analysis will allow us to make statistical inferences. From the results, it will be explained why the classical parametric methods can not be utilized to calculate confidence intervals for the power consumption values.

## 3.2-Database Processing

There does not exist specific format to record networks' data. In the situation in which we are concerned with, the database can be described as follows. The power consumption customer real-time measurements are contained in three different files: '*H*eat *R*esidential' (*HR*), '*N*on-*H*eat *R*esidential' (*NHR*) and '*C*ommercial *L*oads' (*CL*). These files consist respectively of 139, 158 and 180 customers. The routine *PROCESS 1* presented in Appendix A, is used to transform each of these three files in several individual files. A part of the results are shown in Appendix B.

To calculate a confidence interval for nodal hourly power consumption, nodal hourly power consumption values recorded during the previous years are used. This calculation will be carried out for a given node and a given hour of the year. Unfortunately, these values are not available. A new data processing will allow us to create artificially these consumption values. The program *PROCESS 2* in Appendix C, wrote in FORTRAN, uses the individual data files created previously to infer one hundred nodal hourly power consumption values. The input parameters for this calculation are: the number of the *H*eat *R*esidential class customers (N1), the number of the *N*on-*H*eat *R*esidential class customers (N2), the number of the *C*ommercial *L*oad class customers (N3) and the hour of the year. A subroutine will calculate randomly the consumption values. It will be described more carefully in the next chapter.

To estimate a confidence interval for maximum power consumption per customer, values of nodal maximum consumption power per customer are required by the calculations. These estimations will be carried out for each month of the year. Thus, it would be very interesting to process data in order to gather these maximum values of power consumption, for a month and a

class given, inside the same file. The routine *PROCESS 3* in Appendix D, wrote in FORTRAN language, was implemented to obtain this result.

## 3.3-Data Analysis

Any confidence interval is estimated for a given network's node. For this node, the load is known. It consists of N1 customers of the *H*eat *R*esidential class, N2 customers of the *N*on-*H*eat *R*esidential class, and N3 of the *C*ommercial *L*oad class. In practice, N1, N2 and N3 can take all possible integer values.

Two statistical analyses are presented in the following pages. One analysis deals with the nodal hourly power consumption values, given the hour and the node. The other one deals with the values of nodal maximum power consumption per customer, given the month and the node. As it will be described in Chapters 5 and 6, the results are used in the calculation of confidence intervals for nodal hourly power consumption and for nodal maximum power consumption per customer.

S-PLUS, a powerful language designed for data analysis and graphics, will allow us to explore visually the data and to infer statistical conclusions **[22, 23]**. Several S-PLUS functions were designed to infer numerical descriptive measures for a sample such as estimates of location and spread. Other functions are used to create a graphical representation of the dataset. The two most common numerical descriptive measures are measures of central tendency and measures of variability. This means that we seek to estimate the center of the probability distribution of the population and its spread about that center. The three measures of central tendency considered in the next parts will be the *mean*, the *median* and the *mode*. The measures of dispersion will be the *standard deviation* and the *median-absolute-deviation from the median* called *MAD*. Let us recall that the mean of a distribution is defined to be the center of gravity of the distribution. The median is defined to be the center of probability. The mode is defined to be the value that occurs with the highest frequency. The standard deviation is the positive square-root of the variance. Finally, the MAD is a measure of dispersion about the median.

### 3.3.1-Nodal Hourly Power Consumption Analysis

Once the nodal hourly power consumption values were inferred by using *PROCESS 2*, the distribution of these values can be analyzed. Three different examples are going to be presented next. The characteristics of each distribution are given by the composition of the load and the hour of consumption that is selected.

*Example 1*

In the following example, the demand point consists of one hundred customers: fifty customers from the *H*eat *R*esidential class, forty from the *N*on-*H*eat *R*esidential class and ten from the *C*ommercial *L*oad class. The chosen hour of the year is the third hour of January 1. One hundred nodal hourly consumption values were calculated as explained in a previous paragraph. These one hundred values form the basic set of measurements which will be utilized to calculate the confidence interval given in Example 1, Chapter 5. The distribution of this sample is shown in Figure 3.1.

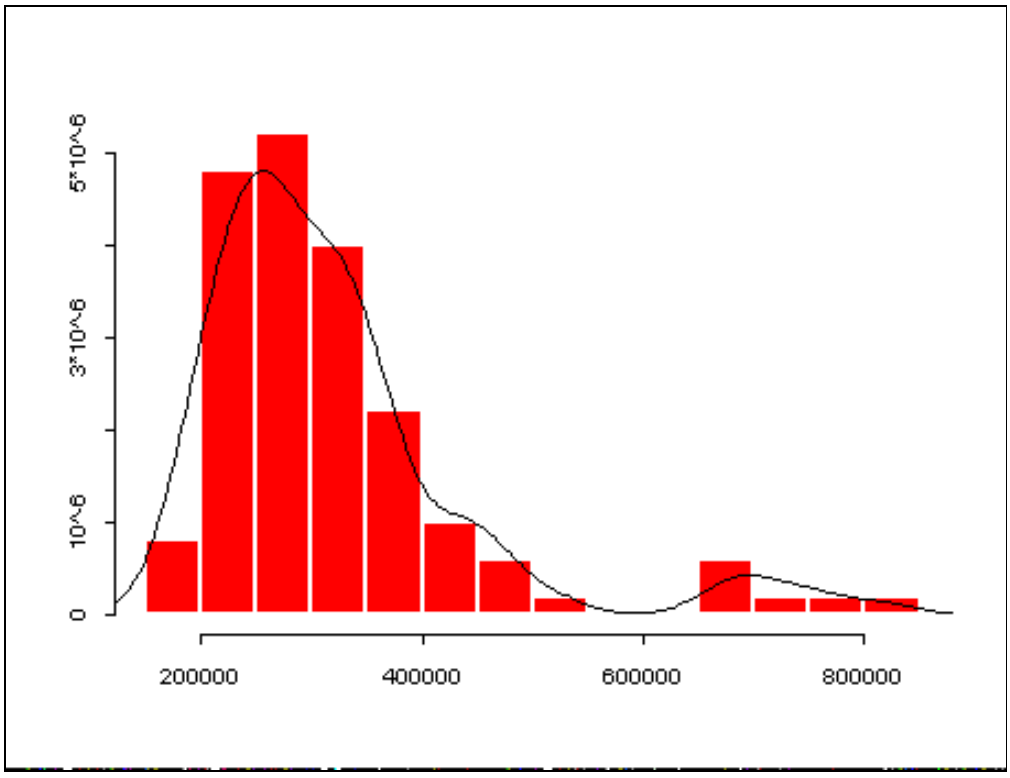


**Fig 3.1** Histogram and Smoothed Density Estimate
of the Distribution of the Nodal Hourly Power Consumption for Example 1

The values of the power consumption range from 174,630 kW to 814,583 kW for this first example. Their location and spread, estimated by means of S-PLUS, are as follows

mean = 351,832

median = 269,281
mode = 245,364
standard deviation = 126,433
MAD = 81,085

*Example 2*

In this second example, the node consists of one hundred and fifty customers: seventy five customers from the *H*eat *R*esidential class, fifty from the *N*on-*H*eat *R*esidential class and twenty five from the *C*ommercial *L*oad class. The hour of the year chosen is the forth hour of May 1. As in the previous example, one hundred nodal hourly consumption values were calculated. These values form the basic set of measurements that is utilized to calculate the confidence interval in Example 2, Chapter 5.

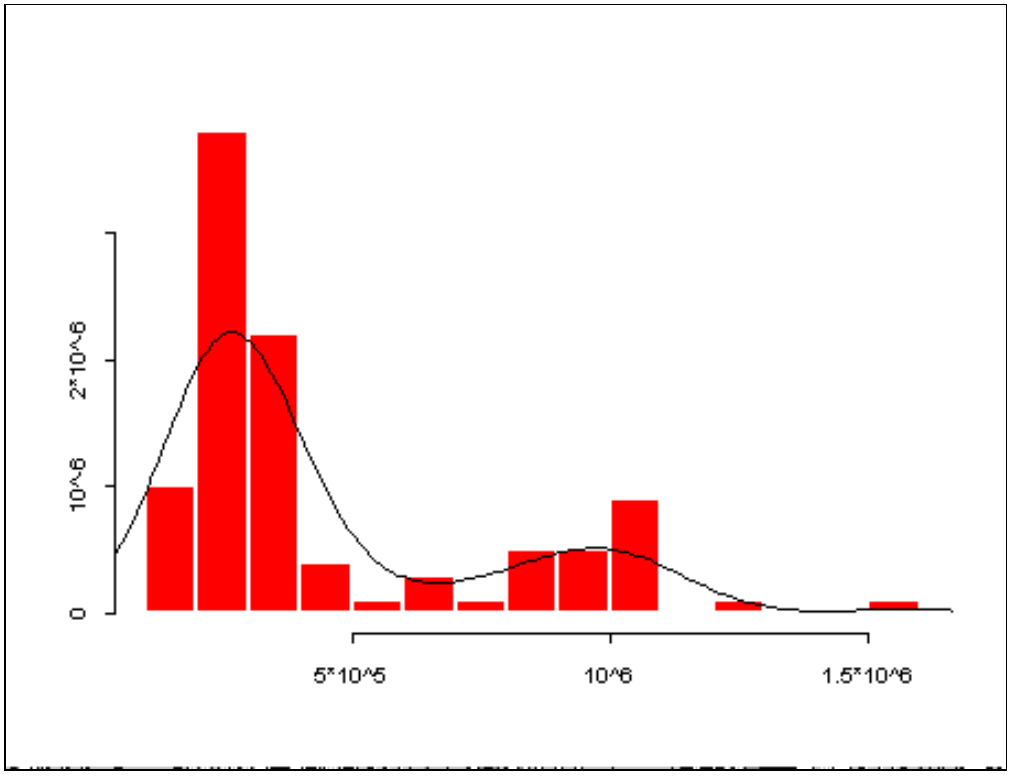


**Fig 3.2** Histogram and Smoothed Density Estimate
of the Distribution of the Nodal Hourly Power Consumption for Example 2

The values of the power consumption range from 140,091 kW to 1,572,379 kW. Their location and spread are as follows

mean = 447,016
median = 306,115
mode = 254,599
standard deviation = 316,679
MAD = 122,685

*Example 3*

For this last example, the composition of the node is a little bit different. It includes sixty customers: that consists of thirty customers from the *H*eat *R*esidential class, thirty from the *N*on-*H*eat *R*esidential class and no customer from the *C*ommercial *L*oad class. The hour of the year chosen is the $7^{th}$ hour of January 25. This composition for the node was chosen in order to show the impact of the customers consumption values from the *CL* class on the distribution.

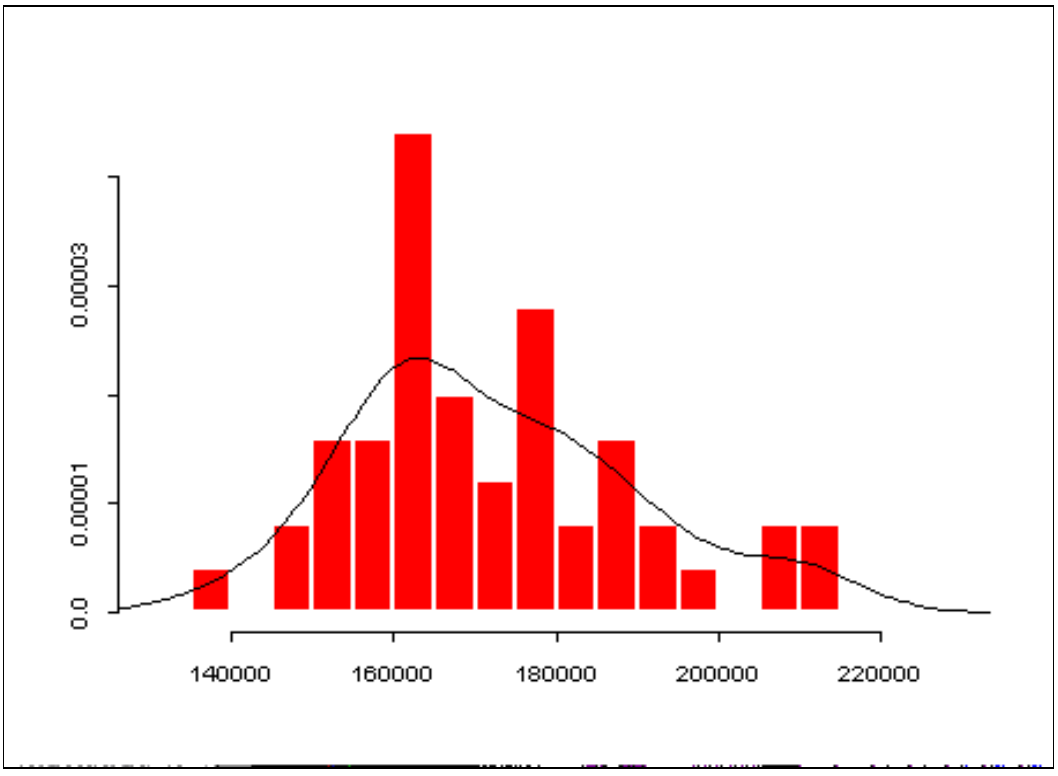


**Fig 3.3** Histogram and Smoothed Density Estimate
of the Distribution of the Nodal Hourly Power Consumption for Example 3

The values of the power consumption range from 138,091 kW to 210,750 kW. Their location and spread, estimated by means of S-PLUS, are as follows

mean = 174,310
median = 168,176
mode = 166,250
standard deviation = 17,323
MAD = 15,958

*Discussions*

The histograms shown in the figures above display the relative frequency of the data points interval. Generally speaking, the histogram gives a qualitative appreciation about the distribution of the data set. To quantify the distribution asymmetry, numerical measures are needed. If the values of the mode, the median and the mean are close to each other, the distribution is then likely symmetric. Otherwise, it is asymmetric.

From the histograms shown in Figures 3.1, 3.2, 3.3, we infer that the distributions of the samples in the 3 examples strongly depart from Gaussianity. In the two first examples, the distributions are very asymmetric with a long tail. This is confirmed by the large differences between the numerical values of the mode, the median and the mean. The measures of dispersion show that the values of consumption inside the sample are largely spread around the center of the distribution. In the third example, the measures of central tendency are not the same but they are much closer than in the two other examples. The shape of the distribution, however, remains asymmetric.

The difference between the Examples 1,2 and the Example 3 lies in the composition of the load. In the third case, no customer from the *CL* class was considered. In fact, the consumption values for these customers often are much more important than for the customers from the two other classes. This explains why in the two first examples the distribution is skewed to the left with a long tail toward large values. Furthermore, the values of consumption for the customers from the *HR* and *NHR* classes are quite comparable when the same period of the year is considered. This explains why in the third example, the distribution is more grouped around a central position. This distribution is, however, far from a Gaussian distribution.

To sum up, the distributions of the nodal hourly power consumption may strongly depart from a Gaussian distribution. Moreover, these distributions are different from one case to another one. These differences depend on the composition of the load and the hour of the year considered. Therefore, classical statistical methods can not be used to calculate confidence intervals in this case since they assume a Gaussian distribution for the load distributions.

### 3.3.2-Nodal Maximum Power Consumption per Customer Analysis

For a given demand point, the load is known and consists of N = N1 + N2 + N3 customers. Given the month, the basic sample of values of nodal maximum power consumption per customer for the node has to be created. The values of nodal maximum power consumption per customer gathered by *PROCESS 3* are used to build this sample. More precisely, a program randomly select without replacement, N values of individual maximum power consumption. This routine is included inside the program of confidence intervals estimation presented in Appendix G. Two examples of the basic sample's distribution are presented in this section. The characteristics of each distribution are given by the composition of the load and the month that is picked.

*Example 1*

In the following example, the nodal load consists of two hundred and eighty customers (N1 = 120, N2 = 120, N3 = 40). The sample was built for the month of April. The distribution of the consumption values is displayed in Figure 3.4.

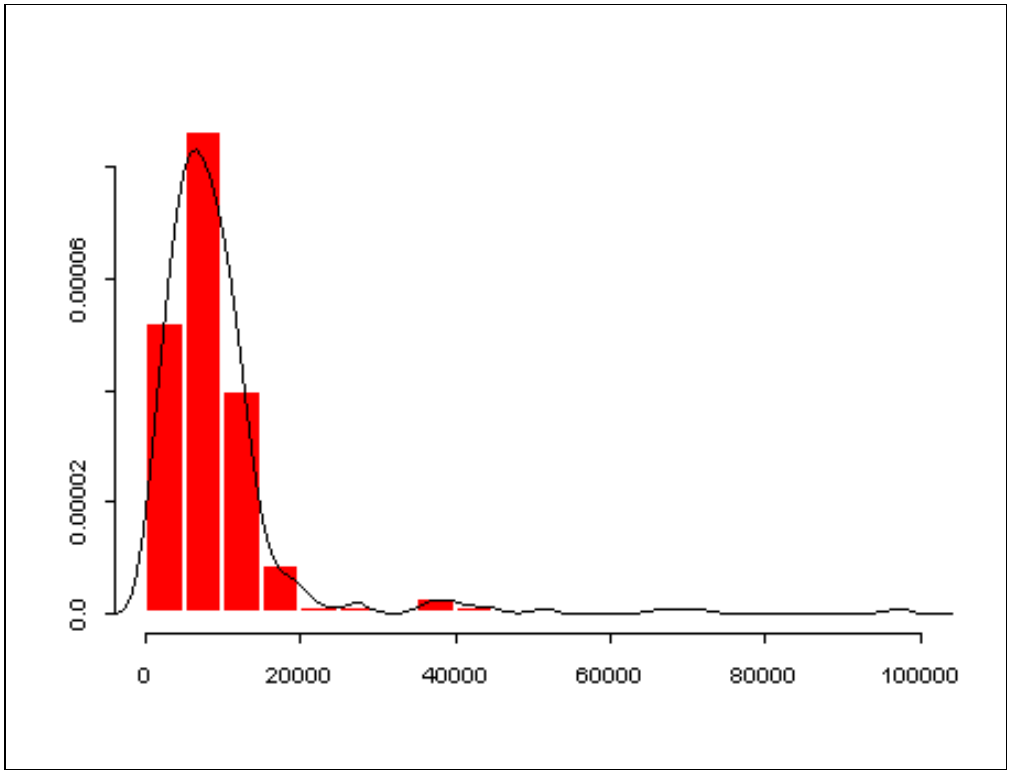


**Fig 3.4** Histogram and Smoothed Density Estimate of the Distribution of the Nodal Maximum Power Consumption per Customer for Example 1

The values of the power consumption range from 144 kW to 202,032 kW. The values of some location and spread estimates are

mean = 11,660
median = 7,584
mode = 5,866
standard deviation = 21,472
MAD = 4,438

Note that the MAD is about one fifth of the standard deviation. This is due to the skewness of the distribution.

*Example 2*

In this example, the nodal load consists of two hundred customers (*HR* = 120, *NHR* = 80, *CL* = 0). The sample was built for the month of April. This set of measurements will be used to calculate the confidence interval in Example 1, Chapter 6.

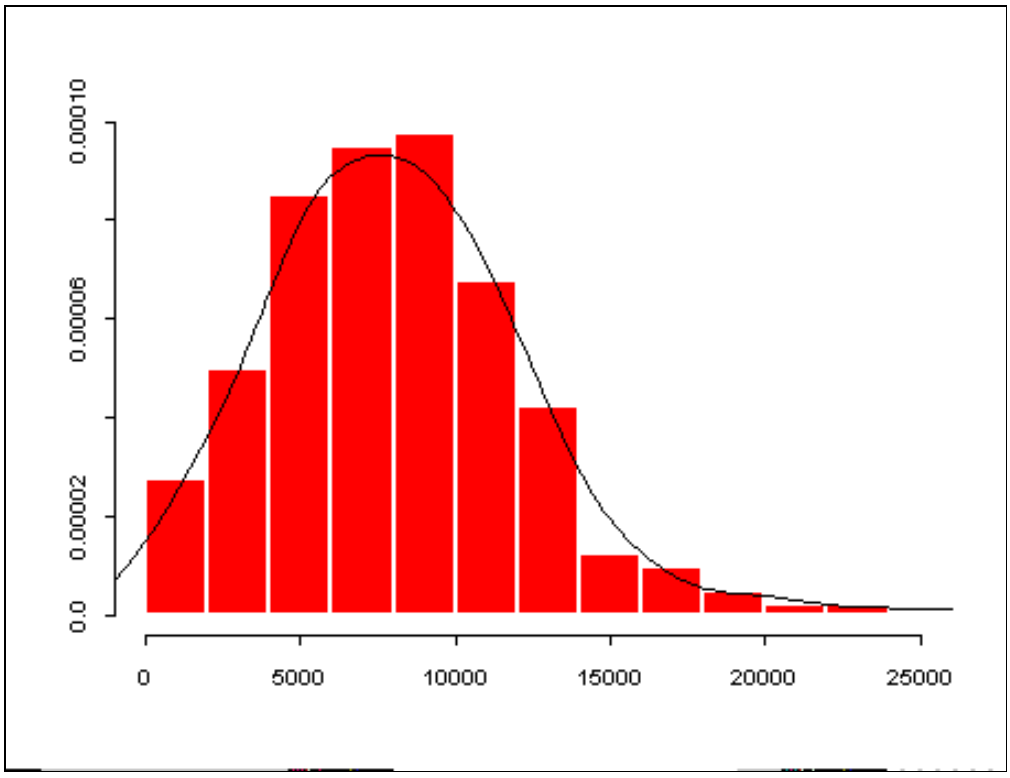


**Fig 3.5** Histogram and Smoothed Density Estimate of the Distribution of the Nodal Maximum Power Consumption per Customer for Example 2

The values of the power consumption range from 144 kW to 27,468 kW. Their location and spread, estimated by means of S-PLUS, are as follows

mean = 8,369
median = 7,920
mode = 7,850
standard deviation = 4,283
MAD = 3,812

Note that the estimates of the mean, median, and mode are very close to each other. Similarly, the difference between the standard deviation and the MAD is small. This is due to the fact that the density function is not very skewed, due to the absence of customers from the *CL* class. It still departures from Gaussianity.

## 3.4-Conclusion

In the previous paragraphs, analyses were carried out on two types of distributions: nodal hourly power consumption distribution and nodal maximum power consumption per customer distribution. These distributions are needed to calculate confidence intervals for the nodal hourly power consumption and the nodal maximum power consumption per customer. These studies revealed that the distributions of power consumption are very asymmetric with values spread over a large interval when the load includes customers from the three classes. If the load only consists of customers from the *H*eat *R*esidential and *N*on-*H*eat *R*esidential classes, the distribution is more symmetric. It departures, however, from a Gaussian distribution.

The shape of the distribution is also very related to the composition of the load and may vary thoroughly from one case to another one. From these observations, it is obvious to infer that it is not possible to use classical methods based on Gaussian assumptions to calculate confidence intervals. Another method has to be considered to estimate these intervals. The advent of fast computers have open new avenues for statistics. They have prompted the development of computationally intensive methods that are based on fewer assumptions than classical methods, and are robust against small departures from these assumptions. Idealized model assumptions can now be replaced by more or less model free analyses. One of these new methods is the bootstrap method. It can be used in two different modes: nonparametric and parametric. The nonparametric bootstrap method is going to be presented in the next chapter. The parametric bootstrap method will be discussed in Chapter 6.