- As a general rule, normality and independence of the data is required in Statistical Process Control and the multivariate extensions are not the exception.
- In a multivariate control chart with the use of rational subgroups according to the central limits theorem certain grade of normality is achieved. But in alternatives called charts for individuals, this rule is not satisfied. The same occurs in capability indices that rarely are computed using subgroups.
- Many authors have proposed nonparametric alternatives to deal with the departures of normality and techniques based on PCA as the studied in Sections 2.10 and 3.6 which are robust to the lack of normality.
- However, nowadays it results quite unproblematic to test multivariate normality and randomness. In this chapter we introduce a wide range of tools to fulfill these requirements.

The first section of this chapter will examine two graphical techniques: histogram and Q-Q plot that facilitate the assumption of normality. Histogram is a graphical technique that allows a visual summary of the data. It provides information about the center, the spread, the skewness, and the existence of outliers. (NIST / SEMATECH e-Handbook of Statistical Methods).

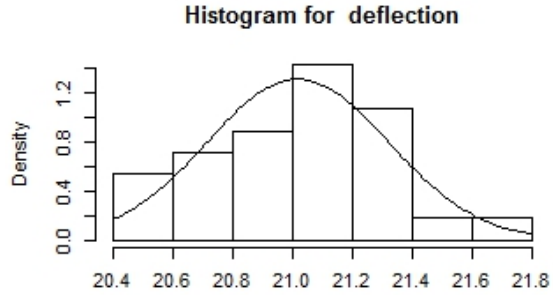
A visual inspection of a histogram permits to establish an initial hypothesis of the distribution; in this case a bell-shaped is desired. Although histograms are basically used in univariate scenarios, univariate normality per se does not imply multivariate normality; if a departure from normality is founded in individual variables, this has a negative effect in the multinormality.

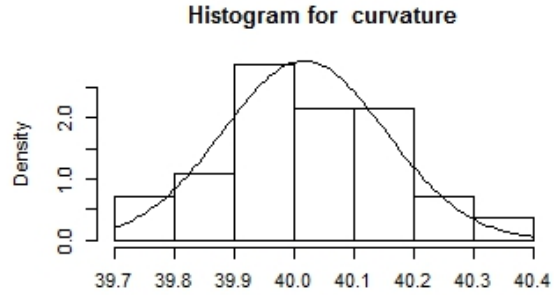
In this example we will illustrate the use of histogram in a multivariate context. For that, return to the bimetal dataset introduced in Sect. 2.6. To put multiple figures in one graph device the parameter mfrow can be used by specifying `mfrow = c(n,m)` being n the number of figures by row and m by columns. As for each quality characteristic a histogram is desireda simple loop is used.

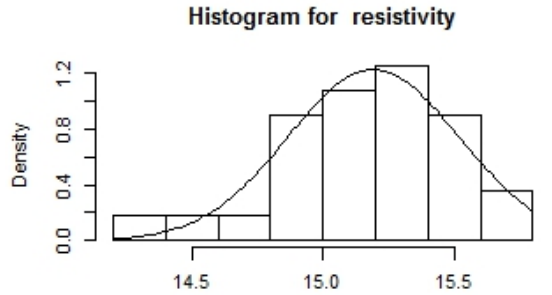
```
> par(mfrow = c(3,2))
> for( i in 1 : ncol(bimetal1) ){
> x <- bimetal1[,i]
> mean<-mean(bimetal1[,i])
> sd<-sd(bimetal1[,i])
> hist(x, prob = TRUE, main = paste( "Histogram for ", coln
> Finally, adding the normal curve
> points(curve(dnorm(x, mean = mean, sd = sd), add = TRUE)
```

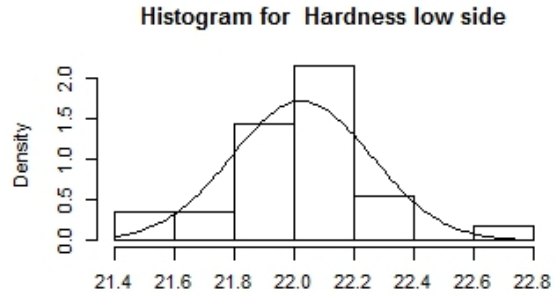
From this chart we can appreciate that most of the classes are located in the center, no significant skewness is revealed, no long tails are presented, and no considerable outliers are detected. The form of the classes does not differ drastically to the normal shape. Finally, there is no visual evidence to reject the univariate normality hypothesis.

This visual inspection can be complemented with the quantile-quantile plot, or simply Q-Q plot. The Q-Q plot is a graphical tool for comparing a two dataset or a dataset with a theoretical distribution. The most common use is to plot the quantiles against a Histogram for Deflection Density


Histogram for deflection
Density
0.0
0.4
0.8
1.2
20.4
20.6
20.8
21.0
21.2
21.4
21.6
21.8


Histogram for curvature
Density
0.0
1.0
2.0
39.7
39.8
39.9
40.0
40.1
40.2
40.3
40.4


Histogram for resistivity
Density
0.0
0.4
0.8
1.2
14.5
15.0
15.5


Histogram for Hardness low side
Density
0.0
0.5
1.0
1.5
2.0
21.4
21.6
21.8
22.0
22.2
22.4
22.6
22.8

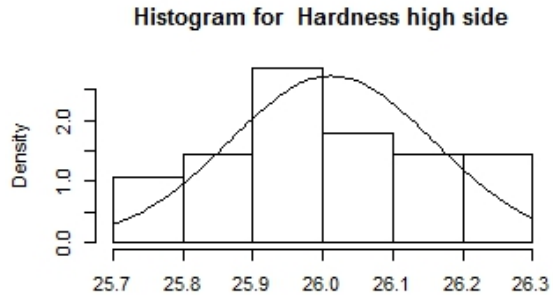

Histogram for Hardness high side
Density
0.0
1.0
2.0
25.7
25.8
25.9
26.0
26.1
26.2
26.3

Figure:

Fig. 4.1 Histogram of the individual variables in the bimetal1 dataset
reference line from a normal distribution. When the points fall approximately over the line there is evidence that both come from an identical distribution. The performing of a Q-Q plot in R is done through the qqnorm function.

Example 4.2

To construct a Q-Q for each variable from bimetal1 dataset:

```
> par(mfrow = c(3,2))
> for( i in 1 : ncol(bimetal1) ){
> qqnorm(bimetal1[,i], main = paste( "Q-Q plot for ", colna
> qqline(bimetal1[,i])
> }
```

From these graphs it appears that each variable is normally distributed since no departure from diagonal line is presented (Fig. 4.2).

Although in a p-variate data the marginal normality does not imply joint normality, deviation from normality frequently affects the marginal distributions. 2 1 0 1 2 20.4 21.2 QQ plot for Deflection Theoretical Quantiles Sample Quantiles 2 1 0 1 2 39.8 40.1

QQ plot for Curvature Theoretical Quantiles Sample Quantiles 2 1 0 1 2 14.4 15.2 QQ plot for Resistivity Theoretical Quantiles Sample Quantiles 2 1 0 1 2 21.6 22.2 QQ plot for Hardness low exp side Theoretical Quantiles Sample Quantiles 2 1 0 1 2 25.8 26.1 QQ plot for Hardness high exp side Theoretical Quantiles Sample Quantiles

Fig. 4.2 The Q-Q plot of the individual variables in bimetal1 dataset 4.1 Tools of Support to MSQC 89
There are many well-known univariate normality tests like: w2, Anderson-Darling, Kolmorov-Smirnov, DAgostino, Jarque-Bera, and Shapiro-Wilks tests, etc. In this section, we present an approach to the last three previouslymentioned tests.

The DAgostino(1970) test is based on the power transformation of the sample kurtosis and skewness. It consists of three tests: for skewness, kurtosis, and an omnibus (see DAgostino et al. 1990) for an excellent exposition of the method.

The skewness test is used to test

H_0

H_1

The kurtosis test is based on the following hypothesis

H_0

H_1

The Z(b2) statistics has approximately a normal distribution

In order to integrate both tests, DAgostino and Pearson (1973) proposed an omnibus test with the following statistics Conversely, the Kurtosis Test detects a positive grade of peakness in a low expansion side variable since the kurtosis coefficient was 4.16 although not significant at alpha = 0.05 (see p-value: 0.09) On the other hand the omnibus test does not found departures from normality. According to this test, there is no evidence for rejecting the normality assumption.

The Jarque and Bera (1980) Test is an elegant and powerful goodness of fit test, likewise based on kurtosis and skewness. It is defined as:

$$JB = \frac{m}{6}\left[S^2 + \frac{1}{4}(K-3)^2\right]$$

Figure:

where m is the sample size and S and K the skewness and kurtosis respectively. The JB statistics follows a χ^2 distribution with two degrees of freedom. For more details see Jarque and Bera (1980), Jarque and Bera (1987), or Jarque (2010). Jarque (2010) offers the significance points table although statistical software usually computes the p-values as: p-value=1 - pchisq(STATISTIC, df = 2) or p-value = 1 w2 JB,2

At least three R packages include this test. They are: tseries, moments, and lawstat. In this context we use the first one: ¿ library(”tseries”) Example 4.4 Using the `jarque.bera.test` function for each quality characteristics from the bimetal1 dataset:

Jarque-Bera Test data: bimetal1[, 1] X-squared = 0.22, df = 2, p-value = 0.90 Jarque-Bera Test data: bimetal1[, 3] X-squared = 1.87, df = 2, p-value = 0.39 Jarque-Bera Test data: bimetal1 [, 5] X-squared = 0.57, df = 2, p-value = 0.75 Jarque Bera Test data: bimetal1[, 2] X-squared = 1.74, df = 2, p-value = 0.42 Jarque Bera Test data: bimetal1[, 4] X-squared = 0.96, df = 2, p-value = 0.62

Notice that according to the p-values the normality assumption cannot be rejected at alpha level = 0.05 or 0.10. 4.1 Tools of Support to MSQC 93

The Shapiro and Wilk (1965) Test has become one of the most popular tests due to its high performance. The null hypothesis H0 is the sample that proceeds from a normal distribution and possesses the statistics
R includes the built-in function `shapiro.test()` to compute this test.

The example below illustrates its use over the bimetal1 dataset. Using this function individually for every quality characteristic

```
Shapiro-Wilk normality test
data: deflection
W = 0.98, p-value = 0.86
Shapiro-Wilk normality test
data: resistivity
W = 0.97, p-value = 0.46
Shapiro-Wilk normality test
data: high expansion side
W = 0.98, p-value = 0.78
Shapiro-Wilk normality test
data: curvature
W = 0.98, p-value = 0.89
Shapiro-Wilk normality test
data: low expansion side
W = 0.97, p-value = 0.46
```

Figure:

On the other hand, Thode (2010) offers an excellent presentation of the most powerful test and suggests a test based on moments like Shapiro-Wilks, Anderson- Darling, and Jarque Bera.

Though the literature reflects that the proposals to test multivariate normality exceed the 50 methods (see e.g.: (Mecklin and Mundfrom 2004)) these tools are rarely applied in MSPC publications. This is due to the fact that as a general rule these methods lack of simplicity and the software availability is limited. Three of the most powerful tests are introduced in this section.

Mardia Test

The Mardia (1970) test is a generalization of the univariate skewness and kurtosis test and becomes one of the most popular ones on assessment of multivariate normality. The multivariate skewness and kurtosis are given by:

$$b_{1,p} = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} g_{jk}^3$$

$$b_{2,p} = \frac{1}{n} \sum_{i=1}^{n} g_{jj}^2$$

Figure:

Mardia Test

Mardia (1970, 1974) provides the percentiles for b1,p and b2,p for many values of p (quality characteristics) and many numbers of samples (m). Mardia also proposed for b1,p an approximation to the χ^2 distribution as follows: The Mardia test is available from QRMlib and dprep R packages.

Mardia Test

Example 4.6
Then, to illustrate the Mardia Test return to the bimetal1 dataset. Using the QRMlib package:

```
> MardiaTest(bimetal1)
# The R returns
$skewness
[1] 6.982112
$p.value
[1] 0.585327
$kurtosis
[1] 33.77373
$p.value
[1] 0.3490892
```

Regarding the p-value for skewness and kurtosis, there is no evidence of departures from normality.

Henze and Zirkler

Henze and Zirkler (1990) proposed a multivariate normality test based on the empirical characteristic function. A wide number of simulation studies point out the high performance of this test. See e.g.: (Thode 2002) The statistics is given by:

Henze and Zirkler

The following example shows the application of the test using also the bimetal1 data:

```
> HZ.test(bimetal1)
p-value HZ statistic
[1] 0.61 0.77
```

According to the results obtained, p-value = 0.77, which is a high value; there is no evidence to reject the assumption of multivariate normality.

Royston Test

Another powerful test was proposed by Royston (1983) which is a multivariate extension of the Shapiro andWilks normality test (see Royston 1982, 1983, 1992, 1995). The statistic recommended by Royston is [Royston TS]
There are two ways to compute Zj according to the number of observations:
(4.41) Wj is the statistics of the univariate Shapiro-Wilks test. (See the previous section.)

Departures from Normality

- Practically, it is common to get variables with non-normal distribution and one alternative is to transform the data.
- The transformation of the data is the application of a mathematical function to the original dataset.
- In a multivariate context this solution could be addressed to a marginal or multivariate approach.
- In this section two marginal solutions and one multivariate are introduced.

- There are many simple transformations used in practice: $\sqrt{x}$, log(x), arcsin($\sqrt{x}$), etc (see, e.g., (Juran and Godfrey 1998) Sect. 4.4)
- Another is the well-known Box-Cox Transformation (BCT) that is probably the most used one for practitioners and professionals of quality control.
- Finally, another type of transformation (although not so well known) is the Johnsons system of distributions recognized as the ***Johnson Transformation (JT)***.

The family of Box-Cox is a power transformation suggested by Box and Cox (1964). It is given by: where xi is the original dataset, l (lambda) is the power and yi the new observations.

- One alternative, in order to find the optimal value of l, is using the value that maximizes the logarithm of the likelihood function.
- The BCT is widely used to improve the normality in some practical situations and a lot of statistical packages provide this application.
- An advantage is the easy algorithm to transform the data while a disadvantage is that it does not allow negative data values, though it can be solved by adding a constant to the original dataset.
- There are many functions in `R` that perform the BCT transformation but we will use the `powerTransform` included in `car` package.

The Z family of distributions was presented in Johnson (1949) and is composed by three distributions named ***Unbounded (SU)***, ***Lognormal (SL)***, and ***Bounded (SB)*** which allow to transform into a normal distribution through selecting one of the three of them. The transformations are:

Coming back to the bimetal1 dataset, the marginal independence could be assessed.

```
> par(mfrow = c(3,2))
> for( i in 1 : ncol(bimetal1) ){
> par(mar = c(4.1,4.5,1,1))
> acf(bimetal1[,i],lag = 7,las = 1)}
```

Notice that when lag = 0 the correlation is 1. This can be proved easily in formula x. There is no evidence of relation between adjacent observations; that is, there is marginal randomness. This tool can be complemented with the use of another such as: Box-Pierce, Ljung-Box or Runs Test (Fig. 4.5).

When time dependence is detected the problem should be addressed in two different ways: by using a specific control chart such as the proposal by Apley and Tsung (2002) and Kalagonda and Kulkarni (2004) or by modifying the data removing the autocorrelation effects. About the latter point a possible solution is to decompose it in multivariate autoregressive model and analyze the resultant residuals which should present independency and MVN (Mason and Young 2001).