4

Exploratory Data Analysis

4.1 Introduction

This book is about the statistical analysis of financial markets data such as equity prices, foreign exchange rates, and interest rates. These quantities vary randomly thereby causing financial risk as well as the opportunity for profit. Figures 4.1, 4.2, and 4.3 show, respectively, time series plots of daily log returns on the S&P 500 index, daily changes in the Deutsch Mark (DM) to U.S. dollar exchange rate, and changes in the monthly risk-free return, which is 1/12th the annual risk-free interest rate. A *time series* is a sequence of observations

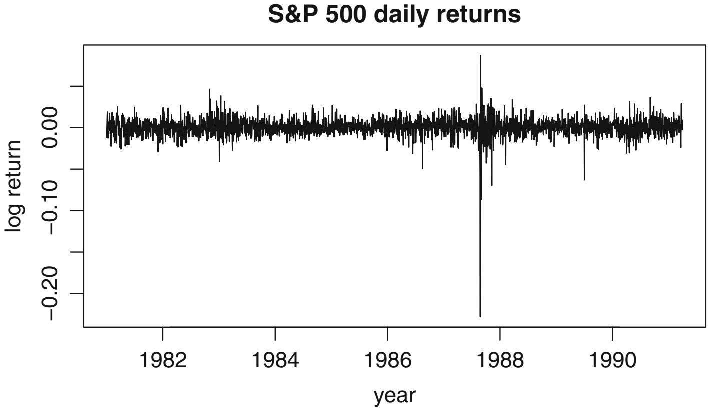

Fig. 4.1. *Daily log returns on the S&P* 500 *index from January* 1981 *to April* 1991*. This data set is the variable* `r500` *in the* `SP500` *series in the* `Ecdat` *package in* `R`*. Notice the extreme volatility in October* 1987*.*

D. Ruppert, D.S. Matteson, *Statistics and Data Analysis for Financial Engineering*, Springer Texts in Statistics,
DOI 10.1007/978-1-4939-2614-5_4

of some quantity or quantities, e.g., equity prices, taken over time, and a *time series plot* is a plot of a time series in chronological order. Figure 4.1 was produced by the following code:

```
data(SP500, package = "Ecdat")
SPreturn = SP500$r500
n = length(SPreturn)
year_SP = 1981 + (1:n) * (1991.25 - 1981) / n
plot(year_SP, SPreturn, main = "S&P 500 daily returns",
     xlab = "year", type = "l", ylab = "log return")
```

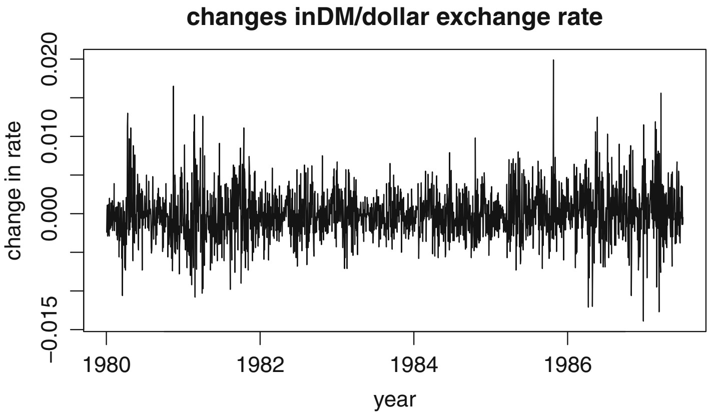

Fig. 4.2. *Daily changes in the DM/dollar exchange rate, January* 2, 1980*, to May* 21, 1987*. The data come from the* `Garch` *series in the* `Ecdat` *package in* `R`*. The DM/dollar exchange rate is the variable* `dm`*.*

Despite the large random fluctuations in all three time series, we can see that each series appears *stationary*, meaning that the nature of its random variation is constant over time. In particular, the series fluctuate about means that are constant, or nearly so. We also see *volatility* clustering, because there are periods of higher, and of lower, variation within each series. Volatility clustering does *not* indicate a lack of stationarity but rather can be viewed as a type of dependence in the conditional variance of each series. This point will be discussed in detail in Chap. 14.

Each of these time series will be modeled as a sequence $Y_1, Y_2, \dots$ of random variables, each with a CDF that we will call F.[1] F will vary between series but, because of stationarity, is assumed to be constant within each series. F is also called the marginal distribution function. By the *marginal distribution* of a stationary time series, we mean the distribution of Y_t given no knowledge

[1] See Appendix A.2.1 for definitions of CDF, PDF, and other terms in probability theory.

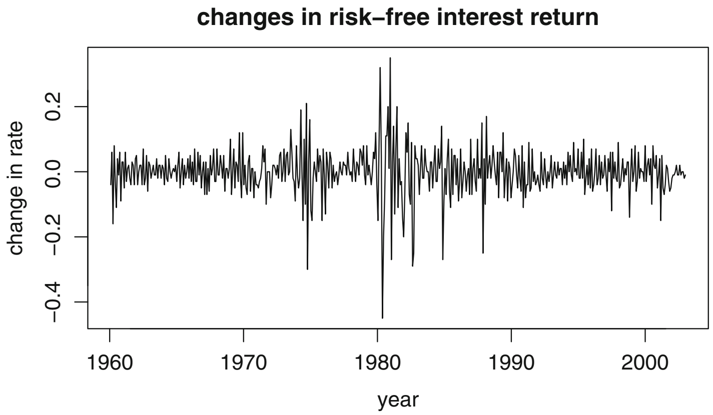

Fig. 4.3. *Monthly changes in the risk-free rate, January* 1960 *to December* 2002*. The rates are the variable* `rf` *in the* `Capm` *series in the* `Ecdat` *package in* `R`*.*

of the other observations, that is, no knowledge of Y_s for any $s \neq t$. Thus, when modeling a marginal distribution, we disregard dependencies in the time series. For this reason, a marginal distribution is also called an *unconditional distribution*. Dependencies such as autocorrelation and volatility clustering will be discussed in later chapters.

In this chapter, we explore various methods for modeling and estimating marginal distributions, in particular, graphical methods such as histograms, density estimates, sample quantiles, and probability plots.

4.2 Histograms and Kernel Density Estimation

Assume that the marginal CDF F has a probability density function f. The histogram is a simple and well-known estimator of probability density functions. Panel (a) of Fig. 4.4 is a histogram of the S&P 500 log returns using 30 cells (or bins). There are some outliers in this series, especially a return near −0.23 that occurred on Black Monday, October 19, 1987. Note that a return of this size means that the market lost 23 % of its value in a single day. The outliers are difficult, or perhaps impossible, to see in the histogram, except that they have caused the x-axis to expand. The reason that the outliers are difficult to see is the large sample size. When the sample size is in the thousands, a cell with a small frequency is essentially invisible. Panel (b) of Fig. 4.4 zooms in on the high-probability region. Note that only a few of the 30 cells are in this area.

The histogram is a fairly crude density estimator. A typical histogram looks more like a big city skyline than a density function and its appearance is sensitive to the number and locations of its cells—see Fig. 4.4, where panels (b), (c), and (d) differ only in the number of cells. A much better estimator is

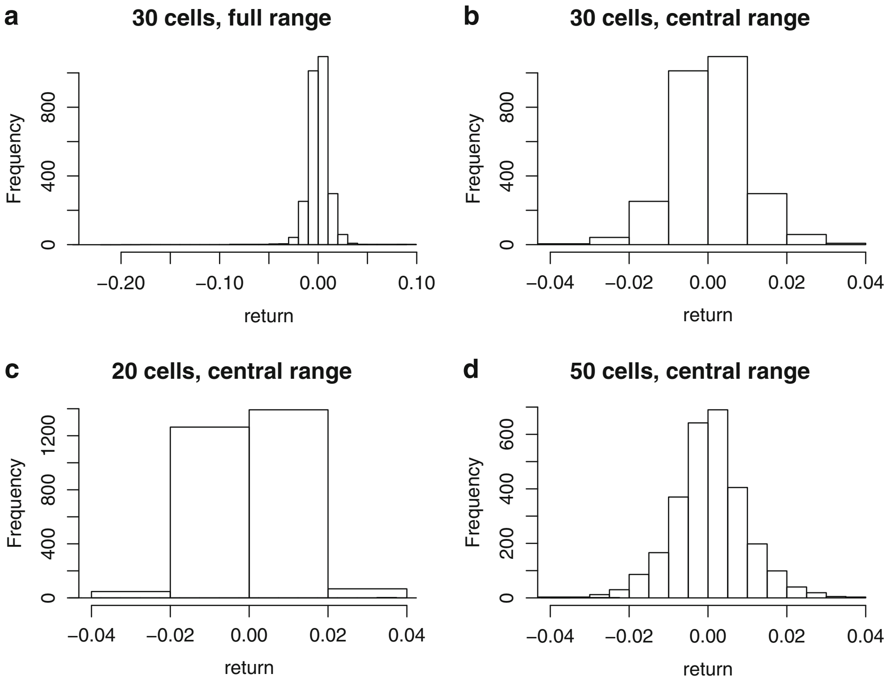

Fig. 4.4. *Histograms of the daily log returns on the S&P* 500 *index from January* 1981 *to April* 1991*. This data set is the* `SP500` *series in the* `Ecdat` *package in* `R`*.*

the *kernel density estimator* (KDE). The estimator takes its name from the so-called kernel function, denoted here by K, which is a probability density function that is symmetric about 0. The standard[2] normal density function is a common choice for K and will be used here. The kernel density estimator based on $Y_1, \ldots, Y_n$ is

$$\widehat{f}(y) = \frac{1}{nb} \sum_{i=1}^{n} K\left(\frac{y - Y_i}{b}\right) \tag{4.1}$$

where b, which is called the bandwidth, determines the resolution of the estimator.

Figure 4.5 illustrates the construction of kernel density estimates using a small simulated data set of six observations from a standard normal distribution. The small sample size is needed for visual clarity but, of course, does not lead to an accurate estimate of the underlying normal density. The six data points are shown at the bottom of the figure as short vertical lines called a “rug.” The bandwidth in the top plot is 0.4, and so each of the six dashed lines is 1/6 times a normal density with standard deviation equal to 0.4 and

[2] “Standard” means having expectation 0 and variance 1.

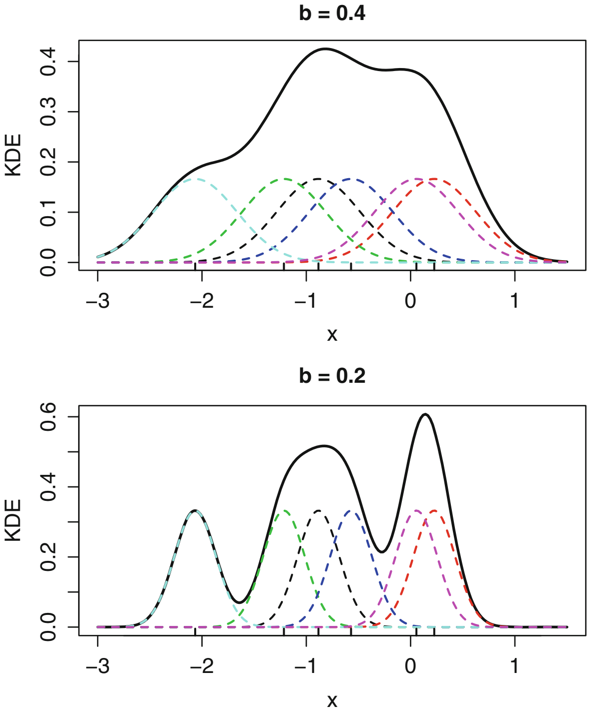

Fig. 4.5. *Illustration of kernel density estimates using a sample of size* 6 *and two bandwidths. The six dashed curves are the kernels centered at the data points, which are indicated by vertical lines at the bottom. The solid curve is the kernel density estimate created by adding together the six kernels. Although the same data are used in the top and bottom panels, the density estimates are different because of the different bandwidths.*

centered at one of the data points. The solid curve is the superposition, that is, the sum as in Eq. (4.1), of the six dashed curves and estimates the density of the data.

A small value of b allows the density estimator to detect fine features in the true density, but it also permits a high degree of random variation. This can be seen in the plot in the bottom of Fig. 4.5 where the bandwidth is only half as large as in the plot on the top. Conversely, a large value of b dampens random variation but obscures fine detail in the true density. Stated differently, a small value of b causes the kernel density estimator to have high variance and low bias, and a large value of b results in low variance and high bias.

Choosing b requires one to make a tradeoff between bias and variance. Appropriate values of b depend on both the sample size n and the true density and, of course, the latter is unknown, though it can be estimated. Roughly speaking, nonsmooth or “wiggly” densities require a smaller bandwidth.

Fortunately, a large amount of research has been devoted to automatic selection of b, which, in effect, estimates the roughness of the true density. As a result of this research, modern statistical software can select the bandwidth automatically. However, automatic bandwidth selectors are not foolproof and density estimates should be checked visually and, if necessary, adjusted as described below.

The solid curve in Fig. 4.6 has the default bandwidth from the `density()` function in `R`. The dashed and dotted curves have the default bandwidth multiplied by 1/3 and 3, respectively. The tuning parameter `adjust` in `R` is the multiplier of the default bandwidth, so that `adjust` is 1, 1/3, and 3 in the three curves. The solid curve with `adjust` equal to 1 appears to have a proper amount of smoothness. The dashed curve corresponding to `adjust` $= 1/3$ is wiggly, indicating too much random variability; such a curve is called undersmoothed and overfit. The dotted curve is very smooth but underestimates the peak near 0, a sign of bias. Such a curve is called oversmoothed or underfit. Here *overfit* means that the density estimate adheres too closely to the data and so is unduly influenced by random variation. Conversely, *underfit* means that the density estimate does not adhere closely enough to the data and misses features in the true density. Stated differently, over- and underfitting means a poor bias–variance tradeoff with an overfitted curve having too much variance and an underfitted curve having too much bias.

Automatic bandwidth selectors are very useful, but there is nothing magical about them, and often one will use an automatic selector as a starting

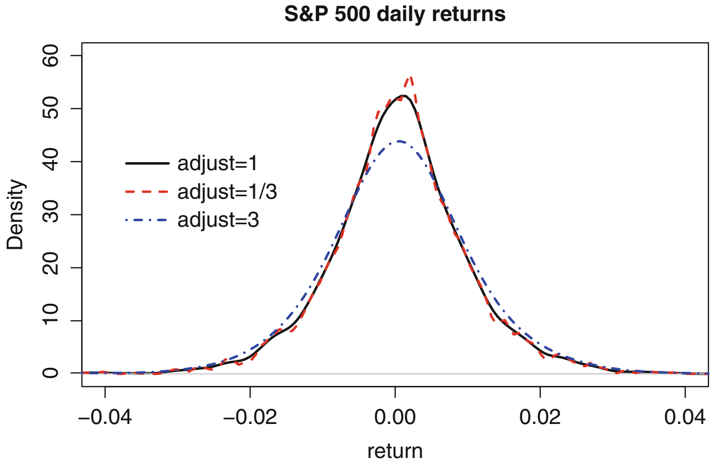

Fig. 4.6. *Kernel density estimates of the daily log returns on the S&P* 500 *index using three bandwidths. Each bandwidth is the default bandwidth times* `adjust` *and* `adjust` *is* 1/3, 1, *and* 3. *This data set is the* `SP500` *series in the* `Ecdat` *package in* `R`. *The KDE is plotted only for a limited range of returns to show detail in the middle of the distribution.*

point and then "fine-tune" the bandwidth; this is the point of the `adjust` parameter. Generally, `adjust` will be much closer to 1 than the values, 1/3 and 3, used above. The reason for using 1/3 and 3 in Fig. 4.6 was to emphasize the effects of under- and oversmoothing.

Often a kernel density estimate is used to suggest a parametric statistical model. The density estimates in Fig. 4.6 are bell-shaped, suggesting that a normal distribution might be a suitable model. To further investigate the suitability of the normal model, Fig. 4.7 compares the kernel density estimate with `adjust` = 1 with normal densities. In panel (a), the normal density has mean and standard deviation equal to the sample mean and standard deviation of the returns. We see that the kernel estimate and the normal density are somewhat dissimilar. The reason is that the outlying returns inflate the sample standard deviation and cause the fitted normal density to be too dispersed in the middle of the data. Panel (b) shows a normal density that is much closer to the kernel estimator. This normal density uses robust estimators which are less sensitive to outliers—the mean is estimated by the sample median and the MAD estimator is used for the standard deviation. The MAD estimator is the median absolute deviation from the median but scaled so that it estimates the standard deviation of a normal population.[3] The sample

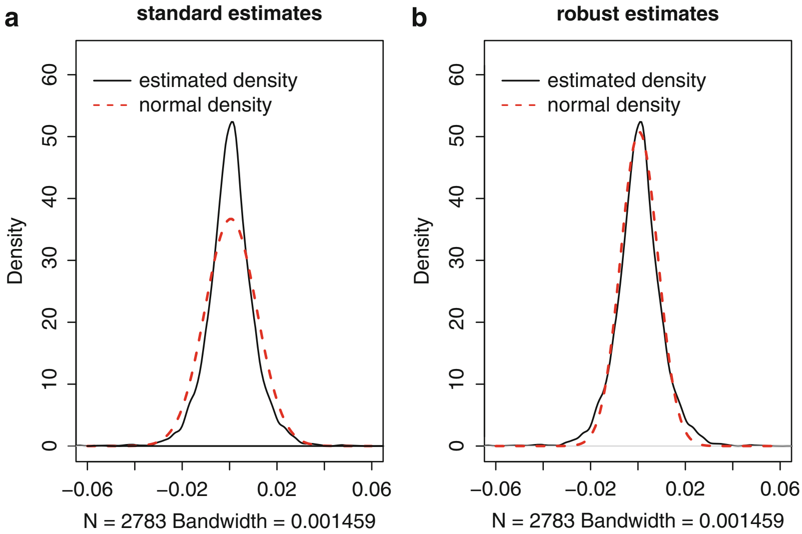

Fig. 4.7. *Kernel density estimates (solid) of the daily log returns on the S&P* 500 *index compared with normal densities (dashed). (***a***) The normal density uses the sample mean and standard deviation. (***b***) The normal density uses the sample median and MAD estimate of standard deviation. This data set is the* `SP500` *series in the* `Ecdat` *package in* `R`*.*

[3] See Sect. 5.16 for more discussion of robust estimation and the precise definition of MAD.

standard deviation is 0.011, but the MAD is smaller, 0.0079; these values were computed using the R functions sd() and mad(). Even the normal density in panel (b) shows some deviation from the kernel estimator, and, as we will soon see, the t-distribution provides a better model for the return distribution than does the normal distribution. The need for robust estimators is itself a sign of nonnormality.

We have just seen a problem with using a KDE to suggest a good model for the distribution of the data in a sample—the parameters in the model must be estimated properly. Normal probability plots and, more generally, quantile–quantile plots, which will be discussed in Sects. 4.3.2 and 4.3.4, are better methods for comparing a sample with a theoretical distribution.

Though simple to compute, the KDE has some problems. In particular, it is often too bumpy in the tails. An improvement to the KDE is discussed in Sect. 4.8.

4.3 Order Statistics, the Sample CDF, and Sample Quantiles

Suppose that $Y_1, \ldots, Y_n$ is a random sample from a probability distribution with CDF F. In this section we estimate F and its quantiles. The *sample* or *empirical CDF* $F_n(y)$ is defined to be the proportion of the sample that is less than or equal to y. For example, if 10 out of 40 ($= n$) elements of a sample are 3 or less, then $F_n(3) = 0.25$. More generally,

$$F_n(y) = \frac{\sum_{i=1}^{n} I\{Y_i \leq y\}}{n}, \tag{4.2}$$

where $I\{\cdot\}$ is the indicator function so that $I\{Y_i \leq y\}$ is 1 if $Y_i \leq y$ and is 0 otherwise. Therefore, the sum in the numerator in (4.2) counts the number of Y_i that are less than or equal to y. Figure 4.8 shows F_n for a sample of size 150 from an $N(0,1)$ distribution. The true CDF (Φ) is shown as well. The sample CDF differs from the true CDF because of random variation. The sample CDF is also called the empirical distribution function, or EDF.

The function ecdf() computes a sample CDF. The code to produce Fig. 4.8 is:

```
set.seed("991155")
edf_norm = ecdf(rnorm(150))
pdf("normalcdfplot.pdf", width = 6, height = 5)  ##  Figure 4.8
par(mfrow = c(1, 1))
plot(edf_norm, verticals = TRUE, do.p = FALSE, main = "EDF and CDF")
tt = seq(from = -3, to = 3, by = 0.01)
lines(tt, pnorm(tt), lty = 2, lwd = 2, col = "red")
legend(1.5, 0.2, c("EDF", "CDF"), lty = c(1, 2),
   lwd = c(1.5, 2), col = c("black", "red"))
graphics.off()
```

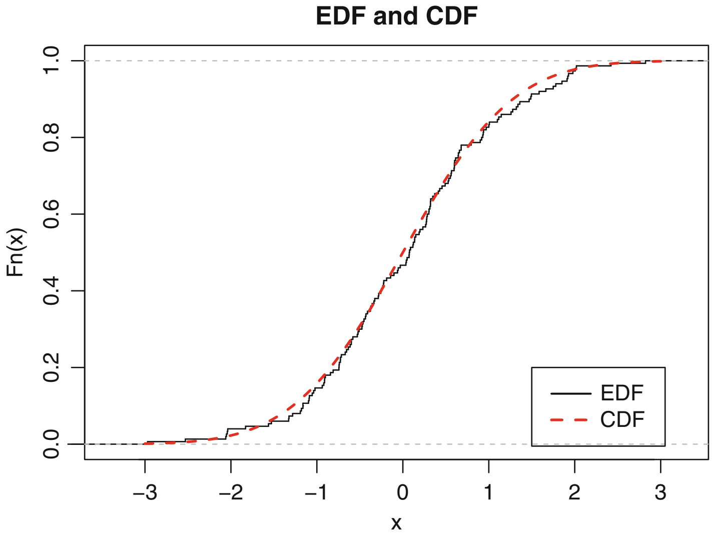

Fig. 4.8. *The EDF* F_n *(solid) and the true CDF (dashed) for a simulated random sample from an* $N(0,1)$ *population. The sample size is* 150.

The *order statistics* $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ are the values $Y_1, \ldots, Y_n$ ordered from smallest to largest. The subscripts of the order statistics are in parentheses to distinguish them from the unordered sample. For example, Y_1 is simply the first observation in the original sample while $Y_{(1)}$ is the smallest observation in that sample. The *sample quantiles* are defined in slightly different ways by different authors, but roughly the q-sample quantile, $0 < q < 1$, is $Y_{(k)}$, where k is qn rounded to an integer. Some authors round up, others round to the nearest integer, and still others interpolate. The function `quantile()` in `R` has nine different types of sample quantiles, the three used by SAS$^{\text{TM}}$, S-PLUS$^{\text{TM}}$, and SPSS$^{\text{TM}}$and Minitab$^{\text{TM}}$, plus six others. With the large sample sizes typical of financial markets data, the different choices lead to nearly identical estimates, but for small samples they can be somewhat different.

The qth quantile is also called the $100q$th *percentile*. Certain quantiles have special names. The 0.5 sample quantile is the 50th percentile and is usually called the *median*. The 0.25 and 0.75 sample quantiles are called the first and third *quartiles*, and the median is also called the second quartile. The 0.2, 0.4, 0.6, and 0.8 quantiles are the *quintiles* since they divide the data into five equal-size subsets, and the 0.1, 0.2, …, 0.9 quantiles are the *deciles*.[4]

[4] Somewhat confusingly, the bottom 10 % of the data is also called the first decile and similarly for the second 10 %, and so forth. Thus, the first decile could refer to the 10th percentile of the data or to all of the data at or below this percentile. In like fashion, the bottom 20 % of the sample is called the first quintile and the second to fifth quantiles are defined analogously.

4.3.1 The Central Limit Theorem for Sample Quantiles

Many estimators have an approximate normal distribution if the sample size is sufficiently large. This is true of sample quantiles by the following central limit theorem.

Result 4.1 *Let $Y_1, \ldots, Y_n$ be an i.i.d. sample with a CDF F. Suppose that F has a density f that is continuous and positive at $F^{-1}(q)$, $0 < q < 1$. Then for large n, the qth sample quantile is approximately normally distributed with mean equal to the population quantile $F^{-1}(q)$ and variance equal to*

$$\frac{q(1-q)}{n\left[f\{F^{-1}(q)\}\right]^2}. \tag{4.3}$$

This result is not immediately applicable, for example, for constructing a confidence interval for a population quantile, because $\left[f\{F^{-1}(q)\}\right]^2$ is unknown. However, f can be estimated by kernel density estimation (Sect. 4.2) and $F^{-1}(q)$ can be estimated by the qth sample quantile. Alternatively, a confidence interval can be constructed by resampling. Resampling is introduced in Chap. 6.

4.3.2 Normal Probability Plots

Many statistical models assume that a random sample comes from a normal distribution. *Normal probability* plots are used to check this assumption, and, if the normality assumption seems false, to investigate how the distribution of the data differs from a normal distribution. If the normality assumption is true, then the qth sample quantile will be approximately equal to $\mu + \sigma\,\Phi^{-1}(q)$, which is the population quantile. Therefore, except for sampling variation, a plot of the sample quantiles versus Φ^{-1} will be linear. One version of the normal probability plot is a plot of $Y_{(i)}$ versus $\Phi^{-1}\{(i-1/2)/n\}$. These are the $(i-1/2)/n$ sample and population quantiles, respectively. The subtraction of $1/2$ from i in the numerator is used to avoid $\Phi^{-1}(1) = +\infty$ when $i = n$.

Systematic deviation of the plot from a straight line is evidence of nonnormality. There are other versions of the normal plot, e.g., a plot of the order statistics versus their expectations under normality, but for large samples these will all be similar, except perhaps in the extreme tails.

Statistical software differs about whether the data are on the x-axis (horizontal axis) and the theoretical quantiles on the y-axis (vertical axis) or vice versa. The `qqnorm()` function in `R` allows the data to be on either axis depending on the choice of the parameter `datax`. When interpreting a normal plot with a nonlinear pattern, it is essential to know which axis contains the data. In this book, the data will always be plotted on the x-axis and the theoretical quantiles on the y-axis, so in `R`, `datax = TRUE` was used to construct the plots rather than the default, which is `datax = FALSE`.

If the pattern in a normal plot is nonlinear, then to interpret the pattern one checks where the plot is convex and where it is concave. A convex curve is one such that as one moves from left to right, the slope of the tangent line increases; see Fig. 4.9a. Conversely, if the slope decreases as one moves from left to right, then the curve is concave; see Fig. 4.9b. A convex-concave curve is convex on the left and concave on the right and, similarly, a concave-convex curve is concave on the left and convex on the right; see Fig. 4.9c and d.

A convex, concave, convex-concave, or concave-convex normal plot indicates, respectively, left skewness, right skewness, heavy tails (compared to the normal distribution), or light tails (compared to the normal distribution)—these interpretations require that the sample quantiles are on the horizontal axis and need to be changed if the sample quantiles are plotted on the vertical axis. *Tails* of a distribution are the regions far from the center. Reasonable definitions of the "tails" would be that the left tail is the region from $-\infty$ to $\mu - 2\sigma$ and the right tail is the region from $\mu + 2\sigma$ to $+\infty$, though the choices of $\mu - 2\sigma$ and $\mu + 2\sigma$ are somewhat arbitrary. Here μ and σ are the mean and standard deviation, though they might be replaced by the median and MAD estimator, which are less sensitive to tail weight.

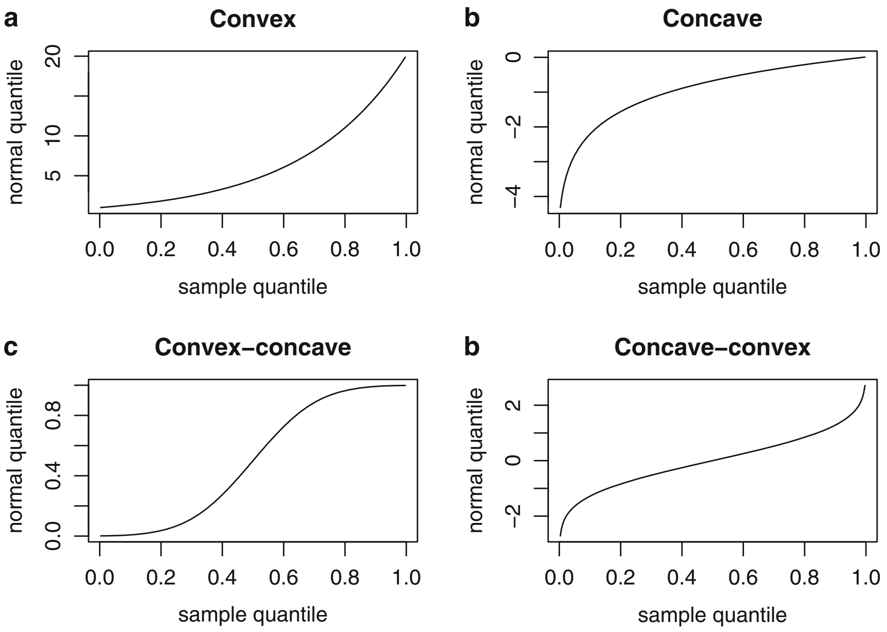

Fig. 4.9. *As one moves from (***a***) to (***d***), the curves are convex, concave, convex-concave, and concave-convex. Normal plots with these patterns indicate left skewness, right skewness, heavier tails than a normal distribution, and lighter tails than a normal distribution, respectively, assuming that the data are on the x-axis and the normal quantiles on the y-axis, as will always be the case in this textbook.*

Figure 4.10 contains normal plots of samples of size 20, 150, and 1000 from a normal distribution. To show the typical amount of random variation in normal plots, two independent samples are shown for each sample size. The plots are only close to linear because of random variation. Even for normally distributed data, some deviation from linearity is to be expected, especially for smaller sample sizes. With larger sample sizes, the only deviations from linearity are in the extreme left and right tails, where the plots are more variable.

Often, a reference line is added to the normal plot to help the viewer determine whether the plot is reasonably linear. One choice for the reference line goes through the pair of first quartiles and the pair of third quartiles; this is what R's `qqline()` function uses. Other possibilities would be a least-squares fit to all of the quantiles or, to avoid the influence of outliers, some subset of the quantiles, e.g., all between the 0.1 and 0.9-quantiles.

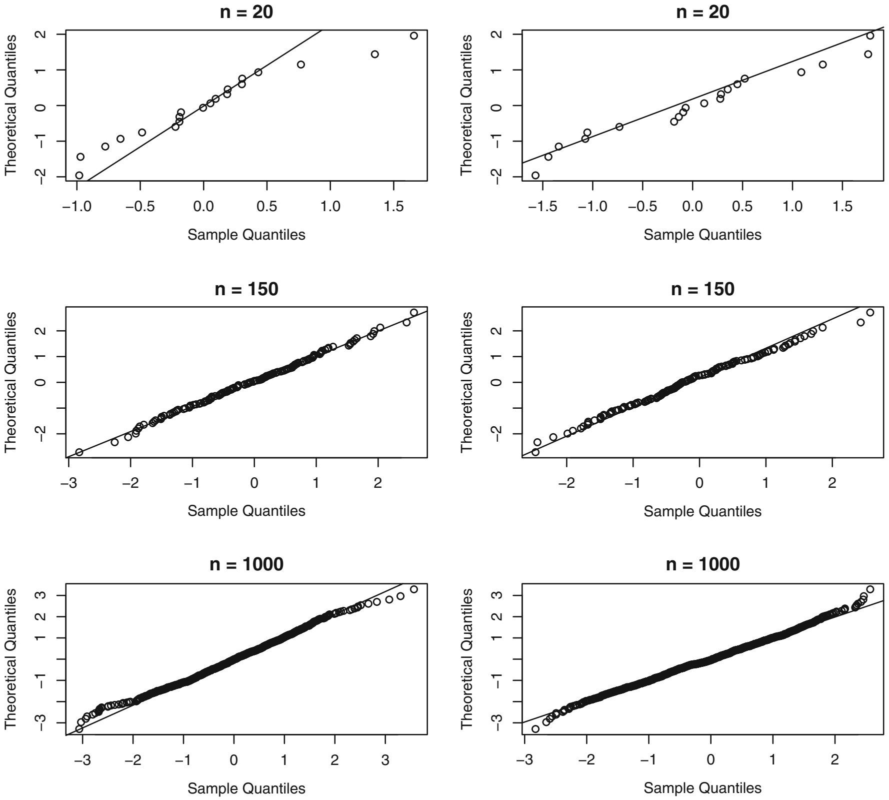

Fig. 4.10. *Normal probability plots of random samples of size* 20, 150*, and* 1000 *from an* $N(0,1)$ *population. The plots were produced by the* R *function* `qqnorm()`*. The reference lines pass through the first and third quartiles and were produced by* R*'s* `qqline()` *function.*

Figure 4.11 contains normal probability plots of samples of size 150 from lognormal $(0, \sigma^2)$ distributions,[5] with the log-standard deviation $\sigma = 1$, $1/2$, and $1/5$. The concave shapes in Fig. 4.11 indicate right skewness. The skewness when $\sigma = 1$ is quite strong, and when $\sigma = 1/2$, the skewness is still very noticeable. With σ reduced to $1/5$, the right skewness is much less pronounced and might not be discernable with smaller sample sizes.

Figure 4.12 contains normal plots of samples of size 150 from t-distributions with 4, 10, and 30 degrees of freedom. The first two distributions have heavy tails or, stated differently, are outlier-prone, meaning that the extreme observations on both the left and right sides are significantly more extreme than would be expected for a normal distribution. One can see that the tails are heavier in the sample with 4 degrees of freedom compared to the sample with

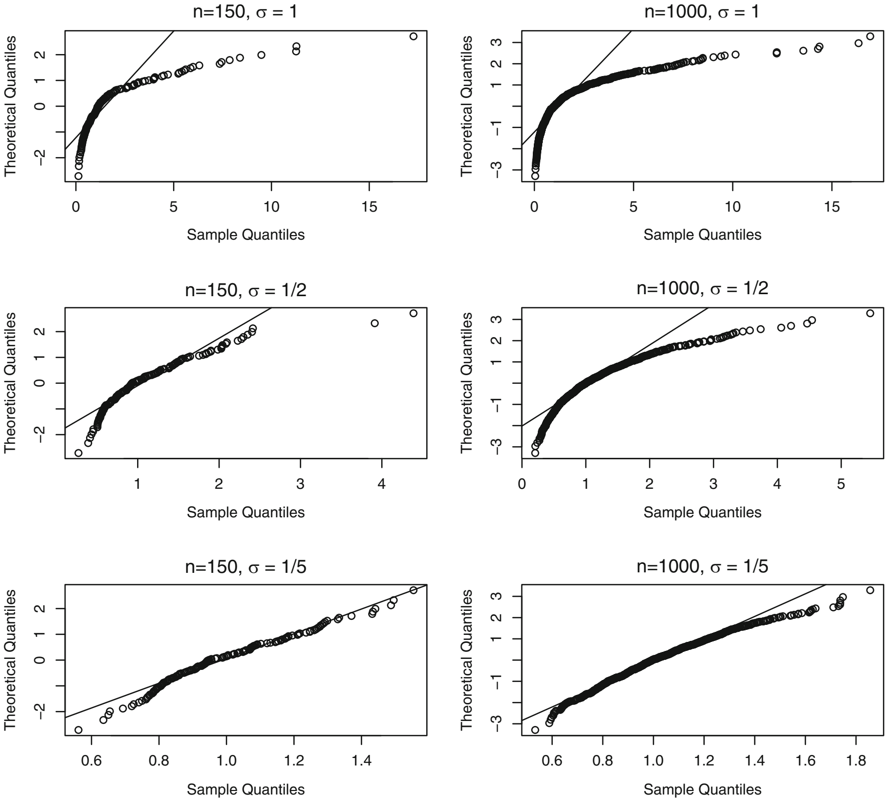

Fig. 4.11. *Normal probability plots of random samples of sizes* 150 *and* 1000 *from lognormal populations with* $\mu = 0$ *and* $\sigma = 1$, $1/2$, *or* $1/5$. *The reference lines pass through the first and third quartiles.*

[5] See Appendix A.9.4 for an introduction to the lognormal distribution and the definition of the log-standard deviation.

10 degrees of freedom, and the tails of the t-distribution with 30 degrees of freedom are not much different from the tails of a normal distribution. It is a general property of the t-distribution that the tails become heavier as the degrees of freedom parameter decreases and the distribution approaches the normal distribution as the degrees of freedom approaches infinity. Any t-distribution is symmetric,[6] so none of the samples is skewed. Heavy-tailed distributions with little or no skewness are common in finance and, as we will see, the t-distribution is a reasonable model for stock returns and other financial markets data.

Sometimes, a normal plot will not have any of the patterns discussed here but instead will have more complex behavior. An example is shown in Fig. 4.13, which uses a simulated sample from a trimodal density. The alternation of the QQ plot between concavity and convexity indicates complex behavior which should be investigated by a KDE. Here, the KDE reveals the trimodality. Multimodality is somewhat rare in practice and often indicates a mixture of several distinct groups of data.

It is often rather difficult to decide whether a normal plot is close enough to linear to conclude that the data are normally distributed, especially when the sample size is small. For example, even though the plots in Fig. 4.10 are close to linear, there is some nonlinearity. Is this nonlinearity due to nonnormality or just due to random variation? If one did not know that the data were simulated from a normal distribution, then it would be difficult to tell, unless one were very experienced with normal plots. In such situations, a test of normality is very helpful. These tests are discussed in Sect. 4.4.

4.3.3 Half-Normal Plots

The half-normal plot is a variation of the normal plot used for detecting outlying data rather than checking for a normal distribution. For example, suppose one has data $Y_1, \ldots, Y_n$ and wants to see whether any of the absolute deviations $|Y_1 - \overline{Y}|, \ldots, |Y_n - \overline{Y}|$ from the mean are unusual. In a half-normal plot, these deviation are plotted against the quantiles of $|Z|$, where Z is $N(0,1)$ distributed. More precisely, a half-normal plot is a scatterplot of the order statistics of the absolute values of the data against $\Phi^{-1}\{(n+i)/(2n+1)\}$, $i = 1, \ldots, n$, where n is the sample size. The function `halfnorm()` in `R`'s `faraway` package creates a half-normal plot and labels the `nlab` most outlying observations, where `nlab` is an argument of this function with a default value of 2.

Example 4.1. DM/dollar exchange rate—Half-normal plot

Figure 4.14 is a half-normal plot of changes in the DM/dollar exchange rate. The plot shows that case #1447 is the most outlying, with case #217

[6] However, t-distributions have been generalized in at least two different ways to the so-called skewed-t-distributions, which need not be symmetric. See Sect. 5.7.

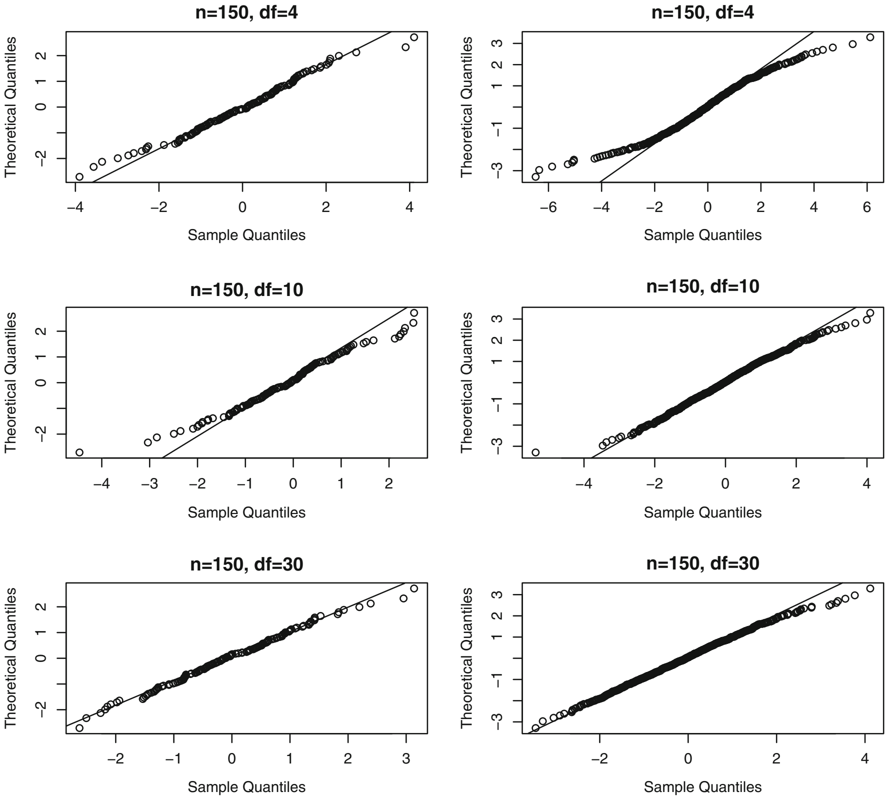

Fig. 4.12. *Normal probability plot of a random sample of size* 150 *and* 1000 *from a t-distribution with* 4*,* 10*, and* 30 *degrees of freedom. The reference lines pass through the first and third quartiles.*

the next most outlying. Only the two most outlying cases are labeled because the default value of `nlab` was used. The code to produce this figure is below.

```
data(Garch, package = "Ecdat")
diffdm = diff(dm)  #  Deutsch mar
pdf("dm_halfnormal.pdf" ,width = 7, height = 6)  # Figure 4.14
halfnorm(abs(diffdm), main = "changes in DM/dollar exchange rate",
   ylab = "Sorted data")
graphics.off()
```

□

Another application of half-normal plotting can be found in Sect. 10.1.3.

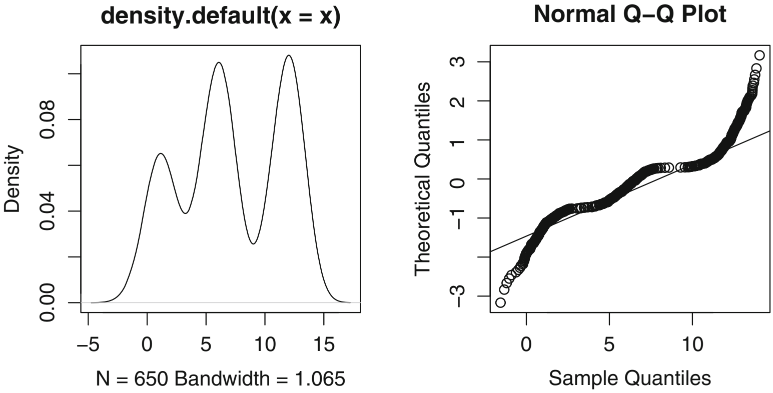

Fig. 4.13. *Kernel density estimate (left) and normal plot (right) of a simulated sample from a trimodal density. The reference lines pass through the first and third quartiles. Because of the three modes, the normal plot changes convexity three times, concave to convex to concave to convex, going from left to right.*

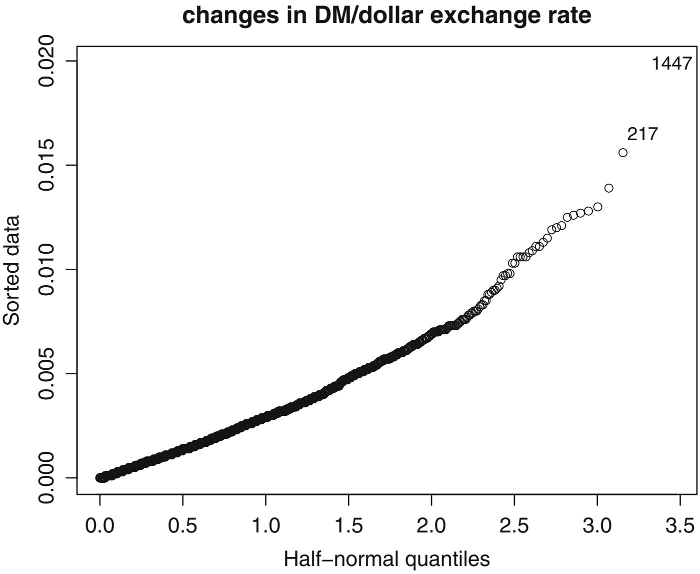

Fig. 4.14. *Half-normal plot of changes in DM/dollar exchange rate.*

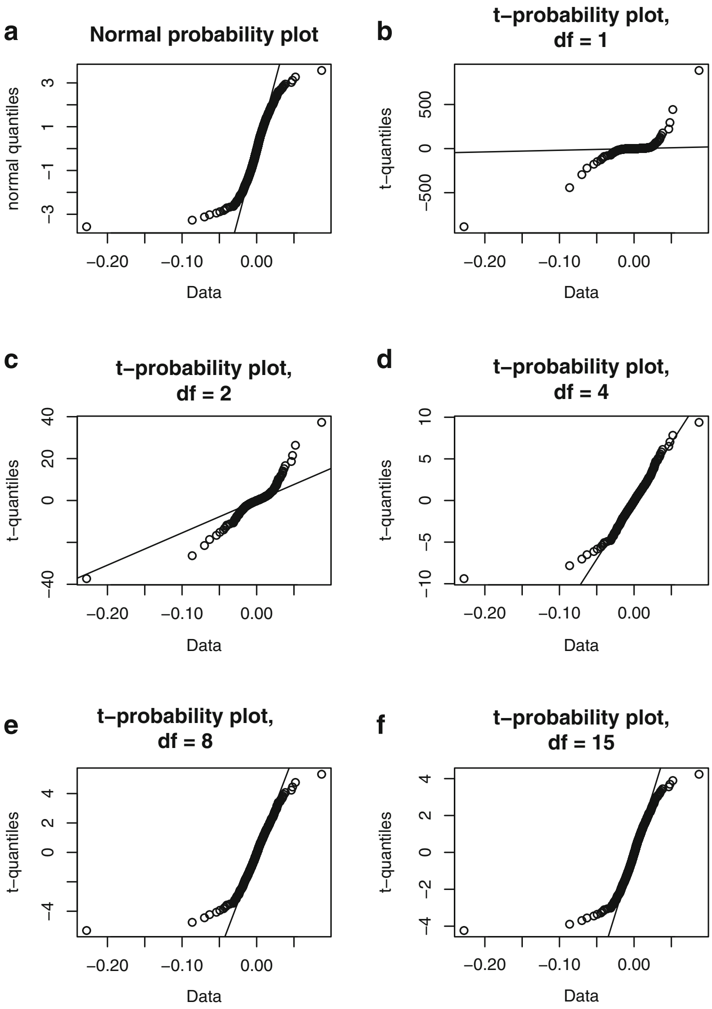

Fig. 4.15. *Normal and t probability plots of the daily returns on the S&P* 500 *index from January* 1981 *to April* 1991. *This data set is the* `SP500` *series in the* `Ecdat` *package in* `R`. *The reference lines pass through the first and third quartiles.*

4.3.4 Quantile–Quantile Plots

Normal probability plots are special cases of *quantile-quantile plots*, also known as QQ plots. A *QQ plot* is a plot of the quantiles of one sample or distribution against the quantiles of a second sample or distribution.

For example, suppose that we wish to model a sample using the $t_\nu(\mu, \sigma^2)$ distribution defined in Sect. 5.5.2. The parameter ν is called the "degrees of freedom," or simply "df." Suppose, initially, that we have a hypothesized value of ν, say $\nu = 6$ to be concrete. Then we plot the sample quantiles against the quantiles of the $t_6(0,1)$ distribution. If the data are from a $t_6(\mu,\sigma^2)$ distribution, then, apart from random variation, the plot will be linear with intercept and slope depending on μ and σ.

Figure 4.15 contains a normal plot of the S&P 500 log returns in panel (a) and t-plots with 1, 2, 4, 8, and 15 df in panels (b) through (f). None of the plots looks exactly linear, but the t-plot with 4 df is rather straight through the bulk of the data. There are approximately nine returns in the left tail and four in the right tail that deviate from a line through the remaining data, but these are small numbers compared to the sample size of 2783. Nonetheless, it is worthwhile to keep in mind that the historical data have more extreme outliers than a t-distribution. The t-model with 4 df and mean and standard deviation estimated by maximum likelihood[7] implies that a daily log return of -0.228, the return on Black Monday, or less has probability 3.2×10^{-6}. This means approximately 3 such returns every 1,000,000 days or 40,000 years, assuming 250 trading days per year. Thus, the t-model implies that Black Monday was extremely unlikely, and anyone using that model should be mindful that it did happen.

There are two reasons why the t-model does not give a credible probability of a negative return as extreme as on Black Monday. First, the t-model is symmetric, but the return distribution appears to have some skewness in the extreme left tail, which makes extreme negative returns more likely than under the t-model. Second, the t-model assumes constant conditional volatility, but volatility was unusually high in October 1987. GARCH models (Chap. 14) can accommodate this type of volatility clustering and provide more realistic estimates of the probability of an extreme event such as Black Monday.

Quantile–quantile plots are useful not only for comparing a sample with a theoretical model, as above, but also for comparing two samples. If the two samples have the same sizes, then one need only plot their order statistics against each other. Otherwise, one computes the same sets of sample quantiles for each and plots them. This is done automatically with the `R` command `qqplot()`.

The interpretation of convex, concave, convex-concave, and concave-convex QQ plots is similar to that with QQ plots of theoretical quantiles versus sample quantiles. A concave plot implies that the sample on the x-axis is more right-skewed, or less left-skewed, than the sample on the y-axis. A convex plot implies that the sample on the x-axis is less right-skewed, or more left-skewed, than the sample on the y-axis. A convex-concave (concave-convex) plot implies that the sample on the x-axis is more (less) heavy-tailed than the sample on the y-axis. As before, a straight line, e.g., through the first and third quartiles, is often added for reference.

Figure 4.16 contains sample QQ plots for all three pairs of the three time series, S&P 500 returns, changes in the DM/dollar rate, and changes in the risk-free return, used as examples in this chapter. One sees that the S&P 500

[7] See Sect. 5.14.

returns have more extreme outliers than the other two series. The changes in DM/dollar and risk-free returns have somewhat similar shapes, but the changes in the risk-free rate have slightly more extreme outliers in the left tail. To avoid any possible confusion, it should be mentioned that the plots in Fig. 4.16 only compare the marginal distributions of the three time series. They tell us nothing about dependencies between the series and, in fact, the three series were observed on different time intervals so correlations between these time series cannot be estimated from these data.

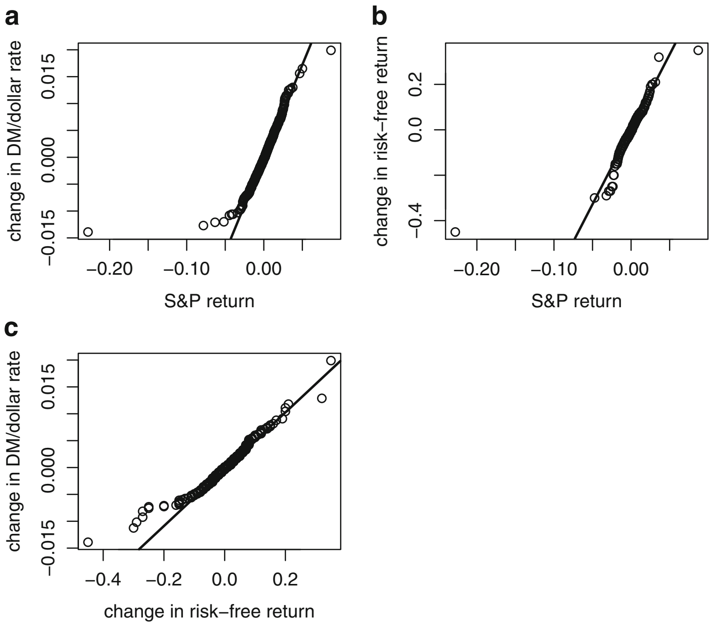

Fig. 4.16. *Sample QQ plots. The straight lines pass through the first and third sample quantiles. (**a**) Change in DM/dollar rate versus S&P return. (**b**) Change in risk-free rate versus S&P return. (**c**) Change in DM/dollar rate versus change in risk-free rate.*

The code for panel (a) of Fig. 4.16 is below. The code for the other panels is similar and so is omitted.

```
qqplot(SPreturn, diffdm, xlab = "S&P return",
   ylab = "change in DM/dollar rate", main = "(a)")
xx = quantile(SPreturn, c(0.25, 0.75))
yy = quantile(diffdm, c(0.25, 0.75))
slope = (yy[2] - yy[1]) / (xx[2] - xx[1])
inter = yy[1] - slope*xx[1]
abline(inter, slope, lwd = 2 )
```

4.4 Tests of Normality

When viewing a normal probability plot, it is often difficult to judge whether any deviation from linearity is systematic or instead merely due to sampling variation, so a statistical test of normality is useful. The null hypothesis is that the sample comes from a normal distribution and the alternative is that the sample is from a nonnormal distribution.

The Shapiro–Wilk test of these hypotheses uses something similar to a normal plot. Specifically, the Shapiro–Wilk test is based on the association between sample order statistics $Y_{(i)}$ and the expected normal order statistics which, for large samples, are close to $\Phi^{-1}\{i/(n+1)\}$, the quantiles of the standard normal distribution. The vector of expected order statistics is multiplied by the inverse of its covariance matrix. Then the correlation between this product and the sample order statistics is used as the test statistic. Correlation and covariance matrices will be discussed in greater detail in Chap. 7. For now, only a few facts will be mentioned. The *covariance* between two random variables X and Y is

$$\mathrm{Cov}(X,Y) = \sigma_{XY} = E\Big[\{X - E(X)\}\{Y - E(Y)\}\Big],$$

and the *Pearson correlation coefficient* between X and Y is

$$\mathrm{Corr}(X,Y) = \rho_{XY} = \sigma_{XY}/\sigma_X\,\sigma_Y. \tag{4.4}$$

A correlation equal to 1 indicates a perfect positive linear relationship, where $Y = \beta_0 + \beta_1 X$ with $\beta_1 > 0$. Under normality, the correlation between sample order statistics and the expected normal order statistics should be close to 1 and the null hypothesis of normality is rejected for small values of the correlation coefficient. In `R`, the Shapiro–Wilk test can be implemented using the `shapiro.test()` function.

The Jarque–Bera test uses the sample skewness and kurtosis coefficients and is discussed in Sect 5.4 where skewness and kurtosis are introduced. Other tests of normality in common use are the Anderson–Darling, Cramér–von Mises, and Kolmogorov–Smirnov tests. These tests compare the sample CDF to the normal CDF with mean equal to $\overline{Y}$ and variance equal to s_Y^2. The Kolmogorov–Smirnov test statistic is the maximum absolute difference between these two functions, while the Anderson–Darling and Cramér–von Mises tests are based on a weighted integral of the squared difference. The p-values of the Shapiro–Wilk, Anderson–Darling, Cramér–von Mises, and Kolmogorov–Smirnov tests are routinely part of the output of statistical software. A small p-value is interpreted as evidence that the sample is not from a normal distribution.

A recent comparison of eight tests of normality (Yap and Sim 2011) found that the Shapiro-Wilk test was as powerful as its competitors for both short- and long-tailed symmetric alternatives and was the most powerful test for asymmetric alternatives. The tests in this study were: Shapiro-Wilk, Kolmogorov-Smirnov, Lilliefors, Cramér-vo Mises, Anderson-Darling, D'Agostino-Pearson, Jarque-Bera, and chi-squared.

For the S&P 500 returns, the Shapiro–Wilk test rejects the null hypothesis of normality with a p-value less than 2.2×10^{-16}. The Shapiro–Wilk also strongly rejects normality for the changes in DM/dollar rate and for the changes in risk-free return. With large sample sizes, e.g., 2783, 1866, and 515, for the S&P 500 returns, changes in DM/dollar rate, and changes in risk-free return, respectively, it is quite likely that normality will be rejected, since any real population will deviate to some extent from normality and any deviation, no matter how small, will be detected with a large enough sample. When the sample size is large, it is important to look at normal plots to see whether the deviation from normality is of practical importance. For financial time series, the deviation from normality in the tails is often large enough to be important.[8]

4.5 Boxplots

The boxplot is a useful graphical tool for comparing several samples. The appearance of a boxplot depends somewhat on the specific software used. In this section, we will describe boxplots produced by the `R` function `boxplot()`. The three boxplots in Fig. 4.17 were created by `boxplot()` with default choice of tuning parameters. The "box" in the middle of each plot extends from the first to the third quartile and thus gives the range of the middle half of the data, often called the *interquartile range*, or IQR. The line in the middle of the box is at the median. The "whiskers" are the vertical dashed lines extending from the top and bottom of each box. The whiskers extend to the smallest and largest data points whose distance from the bottom or top of the box is at most 1.5 times the IQR.[9] The ends of the whiskers are indicated by horizontal lines. All observations beyond the whiskers are plotted with an "o". The most obvious differences among the three boxplots in Fig. 4.17 are differences in scale, with the monthly risk-free return changes being the most variable and the daily DM/dollar changes being the least variable. It is not surprising that the changes in the risk-free return are most variable, since these are changes over months, not days as with the other series.

These scale differences obscure differences in shape. To remedy this problem, in Fig. 4.18 the three series have been standardized by subtracting the median and then dividing by the MAD. Now, differences in shape are clearer. One can see that the S&P 500 returns have heavier tails because the "o"s are farther from the whiskers. The return of the S&P 500 on Black Monday

[8] See Chap. 19 for a discussion on how tail weight can greatly affect risk measures such as VaR and expected shortfall.

[9] The factor 1.5 is the default value of the `range` parameter and can be changed.

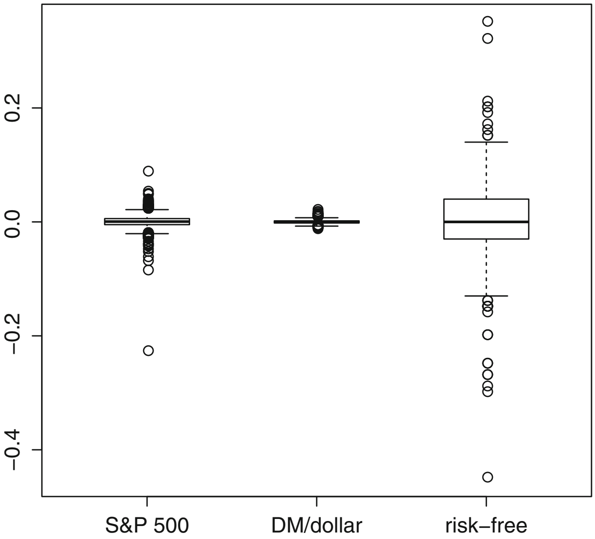

Fig. 4.17. *Boxplots of the S&P* 500 *daily log returns, daily changes in the DM/dollar exchange rate, and monthly changes in the risk-free returns.*

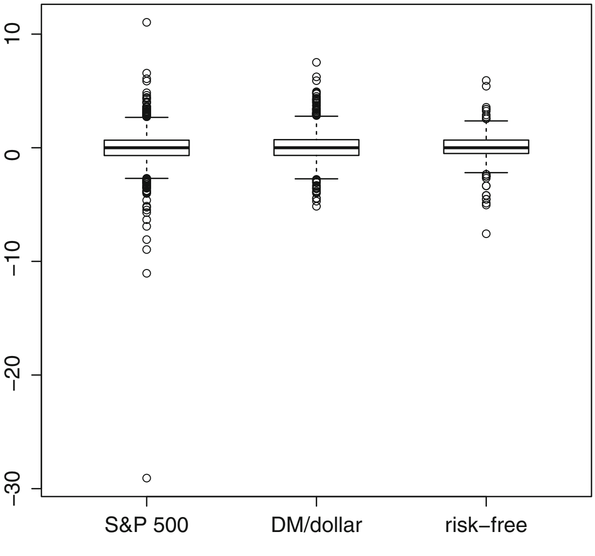

Fig. 4.18. *Boxplots of the standardized S&P* 500 *daily log returns, daily changes in the DM/dollar exchange rate, and monthly changes in the risk-free returns.*

is quite detached from the remaining data. Of course, one should be aware of differences in scale, so it is worthwhile to look at boxplots of the variables both without and with standardization.

When comparing several samples, boxplots and QQ plots provide different views of the data. It is best to use both. However, if there are N samples, then the number of QQ plots is $N(N-1)/2$ or $N(N-1)$ if, by interchanging axes, one includes two plots for each pair of samples. This number can get out of hand quickly, so, for large values of N, one might use boxplots augmented with a few selected QQ plots.

4.6 Data Transformation

There are a number of reasons why data analysts often work not with the original variables, but rather with transformations of the variables such as logs, square roots, or other power transformations. Many statistical methods work best when the data are normally distributed or at least symmetrically distributed and have a constant variance, and the transformed data will often exhibit less skewness and a more constant variance compared to the original variables, especially if the transformation is selected to induce these features.

A transformation is called *variance stabilizing* if it removes a dependence between the conditional variance and the conditional mean of a variable. For example, if Y is Poisson distributed with a conditional mean depending on X, then its conditional variance is equal to the conditional mean. A transformation h would be variance-stabilizing for Y if the conditional variance of $h(Y)$ did not depend on the conditional mean of $h(Y)$.

The logarithm transformation is probably the most widely used transformation in data analysis, though the square root is a close second. The log stabilizes the variance of a variable whose conditional standard deviation is proportional to its conditional mean. This is illustrated in Fig. 4.19, which plots monthly changes in the risk-free return (top row) and changes in the log of the return (bottom row) against the lagged risk-free return (left column) or year (right column). Notice that the changes in the return are more variable when the lagged return is higher. This behavior is called nonconstant conditional variance or conditional heteroskedasticity. We see in the bottom row that the changes in the log return have relatively constant variability, at least compared to changes in the return.

The log transformation is sometimes embedded into the power transformation family by using the so-called Box–Cox power transformation

$$y^{(\alpha)} = \begin{cases} \frac{y^{\alpha}-1}{\alpha}, & \alpha \neq 0 \\ \log(y), & \alpha = 0. \end{cases} \tag{4.5}$$

In (4.5), the subtraction of 1 from y^{α} and the division by α are not essential, but they make the transformation continuous in α at 0 since

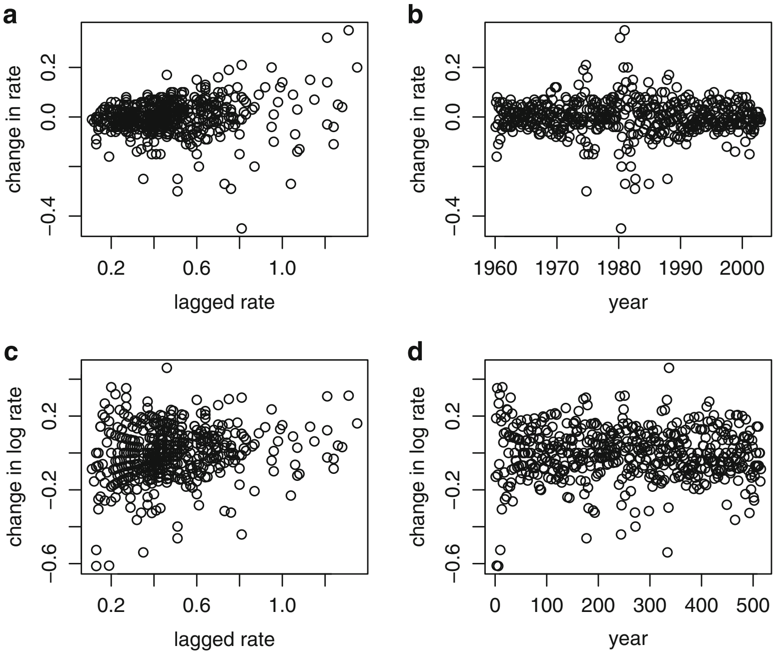

Fig. 4.19. *Changes in risk-free rate (top) and changes in the logarithm of the risk-free rate (bottom) plotted against time and against lagged rate. The risk-free returns are the variable* `rf` *of the* `Capm` *data set in* `R`*'s* `Ecdat` *package. (**a**) Change in risk-free rate versus change in lagged rate. (**b**) Change in rate versus year. (**c**) Change in log(rate) versus lagged rate. (**d**) Change in log(rate) versus year.*

$$\lim_{\alpha \to 0} \frac{y^\alpha - 1}{\alpha} = \log(y).$$

Note that division by α ensures that the transformation is increasing even when $\alpha < 0$. This is convenient though not essential. For the purposes of inducing symmetry and a constant variance, y^α and $y^{(\alpha)}$ work equally well and can be used interchangeably, especially if, when $\alpha < 0$, y^α replaced by $-y^\alpha$ to ensure that the transformation is monotonically increasing for all values of α. The use of a monotonically decreasing, rather than increasing, transformation is inconvenient since decreasing transformations reverse ordering and, for example, transform the pth quantile to the $(1-p)$th quantile.

It is commonly the case that the response is right-skewed and the conditional response variance is an increasing function of the conditional response mean. In such cases, a concave transformation, e.g., a Box–Cox transformation with $\alpha < 1$, will remove skewness and stabilize the variance. If a Box–Cox transformation with $\alpha < 1$ is used, then the smaller the value of α, the greater the effect of the transformation. One can go too far—if the transformed response is *left*-skewed or has a conditional variance that is decreasing as a function of the conditional mean, then α has been chosen too small. Instances of this type of overtransformation are given in Examples 4.2, 4.4, and 13.2.

Typically, the value of α that is best for symmetrizing the data is not the same value of α that is best for stabilizing the variance. Then, a compromise is needed so that the transformation is somewhat too weak for one purpose and somewhat too strong for the other. Often, however, the compromise is not severe, and near symmetry and homoskedasticity can both be achieved.

Example 4.2. Gas flows in pipelines

In this example, we will use a data set of daily flows of natural gas in three pipelines. These data are part of a larger data set used in an investigation of the relationships between flows in the pipelines and prices. Figure 4.20 contains histograms of the daily flows. Notice that all three distributions are left-skewed. For left-skewed data, a Box–Cox transformation should use $\alpha > 1$.

Figure 4.21 shows KDEs of the flows in pipeline 1 after a Box–Cox transformation using $\alpha = 1, 2, 3, 4, 5, 6$. One sees that α between 3 and 4 removes most

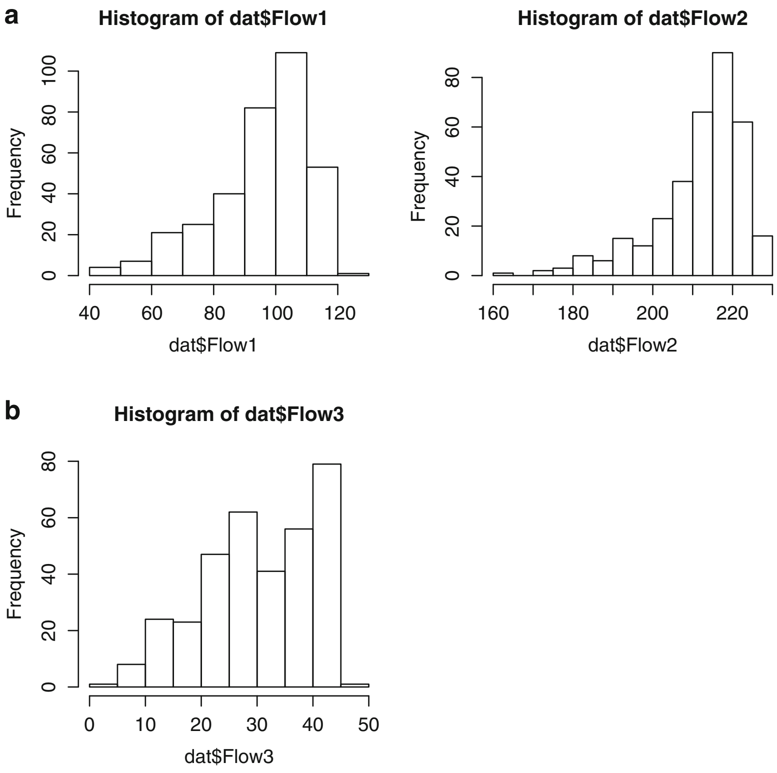

Fig. 4.20. *Histograms of daily flows in three pipelines.*

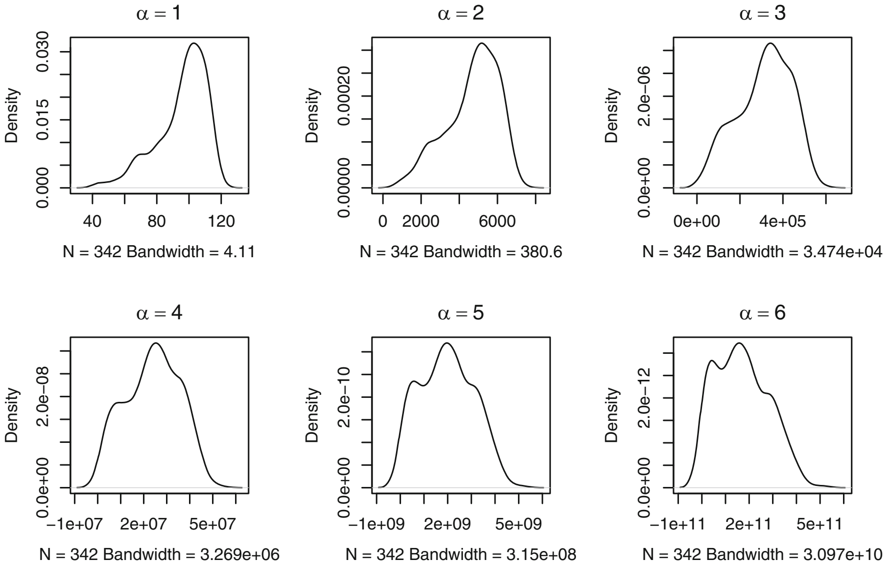

Fig. 4.21. *Kernel density estimates for gas flows in pipeline* 1 *with Box–Cox transformations.*

of the left-skewness and $\alpha = 5$ or greater overtransforms to right-skewness. Later, in Example 5.7, we will illustrate an automatic method for selecting α and find that $\alpha = 3.5$ is chosen. □

Example 4.3. t-Tests and transformations

This example shows the deleterious effect of skewness and nonconstant variance on hypothesis testing and how a proper data transformation can remedy this problem. The boxplots on the panel (a) in Fig. 4.22 are of independent samples of size 15 from lognormal(1,4) (left) and lognormal(3,4) distributions. Panel (b) shows boxplots of the log-transformed data.

Suppose one wants to test the null hypothesis that the two populations have the same means against a two-sided alternative. The transformed data satisfy the assumptions of the t-test that the two populations are normally distributed with the same variance, but of course the original data do not meet these assumptions. Two-sided independent-samples t-tests have p-values of 0.105 and 0.00467 using the original data and the log-transformed data, respectively. These two p-values lead to rather different conclusions, for the first test that the means are not significantly different at the usual $\alpha = 0.05$, and not quite significant even at $\alpha = 0.1$, and for the second test that the difference is highly significant. The first test reaches an incorrect conclusion because its assumptions are not met. □

The previous example illustrates some general principles to keep in mind. All statistical estimators and tests make certain assumptions about the distribution of the data. One should check these assumptions, and graphical methods are often the most convenient way to diagnose problems. If the assumptions are not met, then one needs to know how sensitive the estimator or test is to violations of the assumptions. If the estimator or test is likely to be seriously degraded by violations of the assumptions, which is called *nonrobustness*, then there are two recourses. The first is to find a new estimator or test that is suitable for the data. The second is to transform the data so that the transformed data satisfy the assumptions of the original test or estimator.

4.7 The Geometry of Transformations

Response transformations induce normality of a distribution and stabilize variances because they can stretch apart data in one region and push observations together in other regions. Figure 4.23 illustrates this behavior. On the horizontal axis is a sample of data from a right-skewed lognormal distribution. The transformation $h(y)$ is the logarithm. The transformed data are plotted on the vertical axis. The dashed lines show the transformation of y to $h(y)$ as one moves from a y-value on the x-axis upward to the curve and then to $h(y)$ on the y-axis. Notice the near symmetry of the transformed data. This symmetry is achieved because the log transformation stretches apart data with small values and shrinks together data with large values. This can be seen by observing the derivative of the log function. The derivative of $\log(y)$ is $1/y$, which is a decreasing function of y. The derivative is, of course, the slope of the tangent line and the tangent lines at $y = 1$ and $y = 5$ are plotted to show the decrease in the derivative as y increases.

Consider an arbitrary increasing transformation, $h(y)$. If x and x' are two nearby data points that are transformed to $h(x)$ and $h(x')$, respectively, then

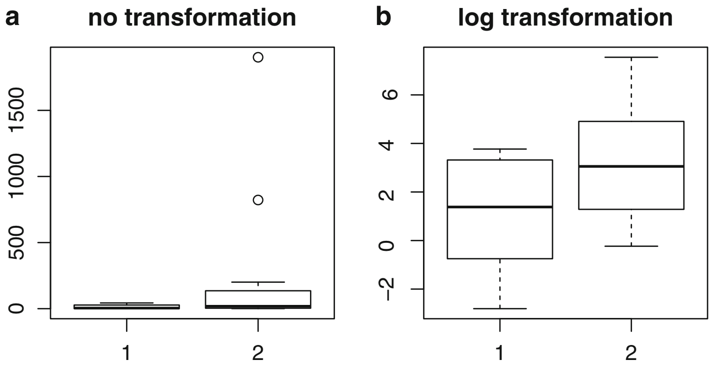

Fig. 4.22. *Boxplots of samples from two lognormal distributions without (***a***) and with (***b***) log transformation.*

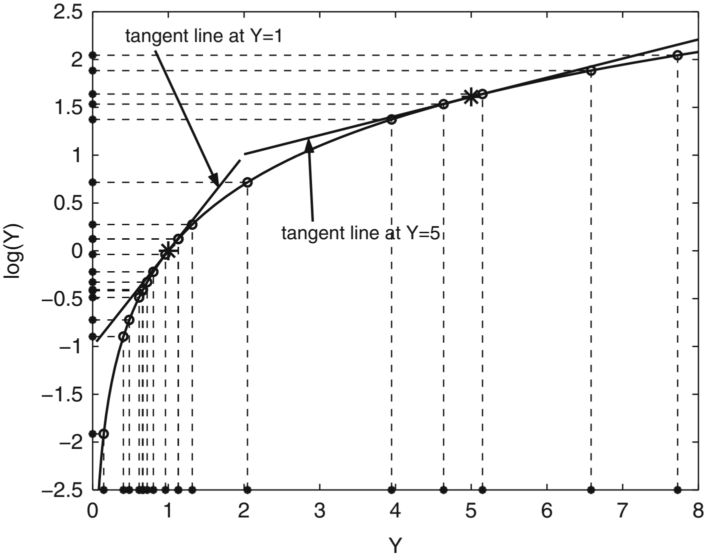

Fig. 4.23. *A symmetrizing transformation. The skewed lognormal data on the horizontal axis are transformed to symmetry by the log transformation.*

the distance between transformed values is $|h(x) - h(x')| \approx h^{(1)}(x)|x - x'|$. Therefore, $h(x)$ and $h(x')$ are stretched apart where $h^{(1)}$ is large and pushed together where $h^{(1)}$ is small. A function h is called concave if $h^{(1)}(y)$ is a decreasing function of y. As can be seen in Fig. 4.23, concave transformations can remove right skewness.

Concave transformations can also stabilize the variance when the untransformed data are such that small observations are less variable than large observations. This is illustrated in Fig. 4.24. There are two groups of responses, one with a mean of 1 and a relatively small variance and another with a mean of 5 and a relatively large variance. If the expected value of the response Y_i, conditional on $\boldsymbol{X}_i$, followed a regression model $m(\boldsymbol{X}_i; \boldsymbol{\beta})$, then two groups like these would occur if there were two possible values of $\boldsymbol{X}_i$, one with a small value of $m(\boldsymbol{X}_i; \boldsymbol{\beta})$ and the other with a large value. Because of the concavity of the transformation h, the variance of the group with a mean of 5 is reduced by transformation. After the transformation, the groups have nearly the same variance, as can be seen by observing the scatter of the two groups on the y-axis.

The strength of a transformation can be measured by how much its derivative changes over some interval, say a to b. More precisely, for $a < b$, the strength of an increasing transformation h is the derivative ratio $h'(b)/h'(a)$. If the transformation is concave, then the derivative ratio is less than 1 and the smaller the ratio the stronger the concavity. Conversely, if the transformation is convex, then the derivative ratio is greater than 1 and the larger the ratio,

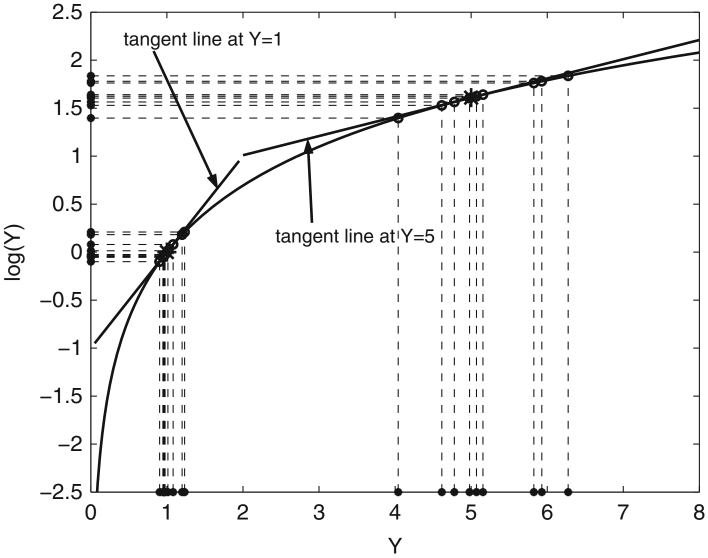

Fig. 4.24. *A variance-stabilizing transformation.*

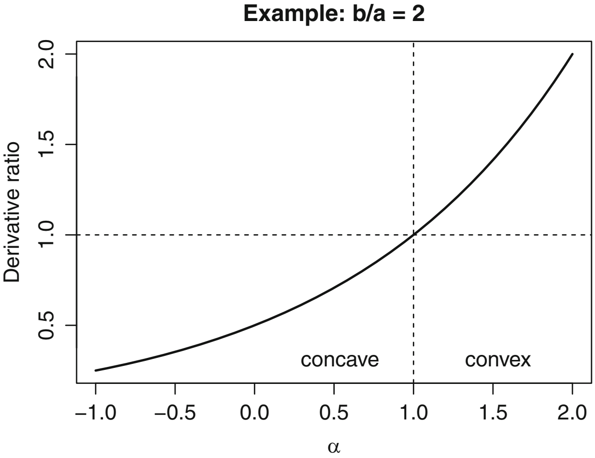

Fig. 4.25. *Derivative ratio for Box–Cox transformations.*

the greater the convexity. For a Box–Cox transformation, the derivative ratio is $(b/a)^{\alpha-1}$ and so depends on a and b only through the ratio b/a. Figure 4.25 shows the derivative ratio of Box–Cox transformations when $b/a = 2$. One can see that the Box–Cox transformation is concave when $\alpha < 1$, with the concavity becoming stronger as α decreases. Similarly, the transformation is convex for $\alpha > 1$, with increasing convexity as α increases.

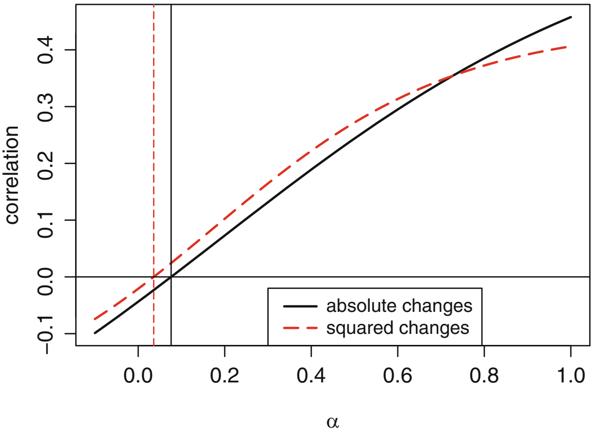

Fig. 4.26. *Correlations between the lagged risk-free returns and absolute (solid) and squared (dashed) changes in the Box–Cox transformed returns. A zero correlation indicates a constant conditional variance. Zero correlations are achieved with the transformation parameter α equal to* 0.036 *and* 0.076 *for the absolute and squared changes, respectively, as indicated by the vertical lines. If $\alpha \approx 0$, then the data are conditionally homoskedastic, or at least nearly so.*

Example 4.4. Risk-free returns—Strength of the Box–Cox transformation for variance stabilization

In this example, we return to the changes in the risk-free interest rates. In Fig. 4.19, it was seen that there is noticeable conditional heteroskedasticity in the changes in the untransformed rate but little or no heteroskedasticity in the changes in the logarithms of the rate. We will see that for a Box–Cox transformation intermediate in strength between the identity transformation ($\alpha = 1$) and the log transformation ($\alpha = 0$), some but not all of the heteroskedasticity is removed, and that a transformation with $\alpha < 0$ is too strong for this application so that a new type of heteroskedasticity is induced.

The strength of a Box–Cox transformation for this example is illustrated in Fig. 4.26. In that figure, the correlations between the lagged risk-free interest returns, r_{t-1}, and absolute and squared changes, $|r_t^{(\alpha)} - r_{t-1}^{(\alpha)}|$ and $\{r_t^{(\alpha)} - r_{t-1}^{(\alpha)}\}^2$, in the transformed rate are plotted against α. The two correlations are similar, especially when they are near zero. Any deviations of the correlations from zero indicate conditional heteroskedasticity where the standard deviation of the change in the transformed rate depends on the previous value of the rate. We see that the correlations decrease as α decreases from 1 so that the concavity of the transformation increases. The correlations are equal to zero when α is very close to 0, that is, the log transformation. If α is much below 0, then the transformation is too strong and the overtransformation induces a negative correlation, which indicates that the conditional standard deviation is a decreasing function of the lagged rate. □

4.8 Transformation Kernel Density Estimation

The kernel density estimator (KDE) discussed in Sect. 4.2 is popular because of its simplicity and because it is available on most software platforms. However, the KDE has some drawbacks. One disadvantage of the KDE is that it undersmooths densities with long tails. For example, the solid curve in Fig. 4.27 is a KDE of annual earnings in 1988–1989 for 1109 individuals. The data are in the `Earnings` data set in `R`'s `Ecdat` package. The long right tail of the density estimate exhibits bumps, which seem due solely to random variation in the data, not to bumps in the true density. The problem is that there is no single bandwidth that works well both in the center of the data and in the right tail. The automatic bandwidth selector chose a bandwidth that is a compromise, undersmoothing in the tails and perhaps oversmoothing in the center. The latter problem can cause the height of the density at the mode(s) to be underestimated.

A better density estimate can be obtained by the *transformation kernel density estimator* (TKDE). The idea is to transform the data so that the density of the transformed data is easier to estimate by the KDE. For the earnings data, the square roots of the earnings are closer to being symmetric and have a shorter right tail than the original data; see Fig. 4.28, which compares histograms of the original data and the data transformed by the square root. The KDE should work well for the square roots of the earnings.

Of course, we are interested in the density of the earnings, not the density of their square roots. However, it is easy to convert an estimate of the latter to

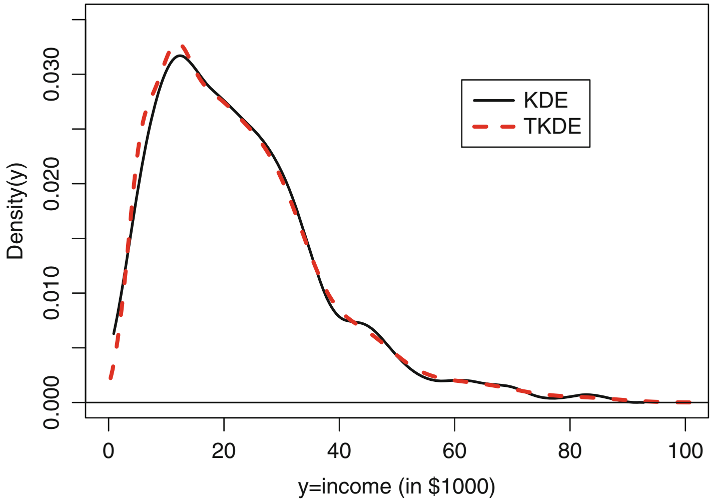

Fig. 4.27. *Kernel density and transformation kernel density estimates of annual earnings in* 1988–1989 *expressed in thousands of* 1982 *dollars. These data are the same as in Fig.* 4.28.

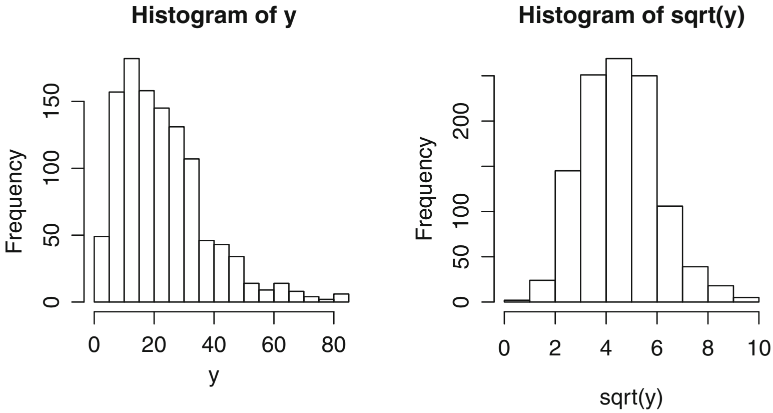

Fig. 4.28. *Histograms of earnings (y) and the square roots of earnings. The data are from the* `Earnings` *data set in* `R`*'s* `Ecdat` *package and use only age group* `g1`*.*

one of the former. To do that, one uses the change-of-variables formula (A.4). For convenience, we repeat the result here—if $X = g(Y)$, where g is monotonic and f_X and f_Y are the densities of X and Y, respectively, then

$$f_Y(y) = f_X\{g(y)\}\,|g'(y)|. \tag{4.6}$$

For example, if $x = g(y) = \sqrt{y}$, then $g'(y) = y^{-1/2}/2$ and

$$f_Y(y) = \{f_X(\sqrt{y})y^{-1/2}\}/2.$$

Putting $y = g^{-1}(x)$ into Eq. (4.6), we obtain

$$f_Y\{g^{-1}(x)\} = f_X(x)\,|g'\{g^{-1}(x)\}|. \tag{4.7}$$

Equation (4.7) suggests a convenient method for computing the TKDE:

1. start with data $Y_1, \ldots, Y_n$;
2. transform the data to $X_1 = g(Y_1), \ldots, X_n = g(Y_n)$;
3. let $\widehat{f}_X$ be the usual KDE calculated on a grid $x_1, \ldots, x_m$ using $X_1, \ldots, X_n$;
4. plot the pairs $\left[g^{-1}(x_j),\ \widehat{f}_X(x_j)\,\left|g'\{g^{-1}(x_j)\}\right|\right]$, $j = 1, \ldots, m$.

The red dashed curve in Fig. 4.27 is a plot of the TKDE of the earnings data using the square-root transformation. Notice the smoother right tail, the faster decrease to 0 at the left boundary, and the somewhat sharper peak at the mode compared to the KDE (solid curve).

When using a TKDE, it is important to choose a good transformation. For positive, right-skewed variables such as the earnings data, a concave transformation is needed. A power transformation, y^α, for some $\alpha < 1$ is a common choice. Although there are automatic methods for choosing α (see Sect. 4.9), trial-and-error is often good enough.

4.9 Bibliographic Notes

Exploratory data analysis was popularized by Tukey (1977). Hoaglin, Mosteller, and Tukey (1983,1985) are collections of early articles on exploratory data analysis, data transformations, and robust estimation. Kleiber and Zeileis (2008) is an introduction to econometric modeling with `R` and covers exploratory data analysis as well as material in latter chapters of this book including regression and time series analysis. The `R` package `AER` accompanies Kleiber and Zeileis's book.

The central limit theorem for sample quantiles is stated precisely and proved in textbooks on asymptotic theory such as Serfling (1980); Lehmann (1999), and van der Vaart (1998).

Silverman (1986) is an early book on nonparametric density estimation and is still well worth reading. Scott (1992) covers both univariate and multivariate density estimation. Wand and Jones (1995) has an excellent treatment of kernel density estimation as well as nonparametric regression, which we cover in Chap. 21. Wand and Jones cover more recent developments such as transformation kernel density estimation. An alternative to the TKDE is variable-bandwidth KDE; see Sect. 2.10 of Wand and Jones (1995) as well as Abramson (1982) and Jones (1990).

Atkinson (1985) and Carroll and Ruppert (1988) are good sources of information about data transformations.

Wand, Marron, and Ruppert (1991) is an introduction to the TKDE and discusses methods for automatic selection of the transformation to minimize the expected squared error of the estimator. Applications of TKDE to losses can be found in Bolance, Guillén, and Nielsen (2003).

4.10 `R` Lab

4.10.1 European Stock Indices

This lab uses four European stock indices in `R`'s `EuStockMarkets` database. Run the following code to access the database, learn its mode and class, and plot the four time series. The `plot()` function will produce a plot tailored to the class of the object on which it is acting. Here four time series plots are produced because the class of `EuStockMarkets` is `mts`, multivariate time series.

```
data(EuStockMarkets)
mode(EuStockMarkets)
class(EuStockMarkets)
plot(EuStockMarkets)
```

If you right-click on the plot, a menu for printing or saving will open. There are alternative methods for printing graphs. For example,

```
pdf("EuStocks.pdf", width = 6, height = 5)
plot(EuStockMarkets)
graphics.off()
```

will send a pdf file to the working directory and the `width` and `height` parameters allow one to control the size and aspect ratio of the plot.

Problem 1 *Write a brief description of the time series plots of the four indices. Do the series look stationary? Do the fluctuations in the series seem to be of constant size? If not, describe how the volatility fluctuates.*

Next, run the following `R` code to compute and plot the log returns on the indices.

```
logR = diff(log(EuStockMarkets))
plot(logR)
```

Problem 2 *Write a brief description of the time series plots of the four series of log returns. Do the series look stationary? Do the fluctuations in the series seem to be of constant size? If not, describe how the volatility fluctuates.*

In `R`, data can be stored as a data frame, which does not assume that the data are in time order and would be appropriate, for example, with cross-sectional data. To appreciate how `plot()` works on a data frame rather than on a multivariate time series, run the following code. You will be plotting the same data as before, but they will be plotted in a different way.

```
plot(as.data.frame(logR))
```

Run the code that follows to create normal plots of the four indices and to test each for normality using the Shapiro–Wilk test. You should understand what each line of code does.

```
par(mfrow=c(2, 2))
for(i in colnames(logR))
{
  qqnorm(logR[ ,i], datax = T, main = i)
  qqline(logR[ ,i], datax = T)
  print(shapiro.test(logR[ ,i]))
}
```

Problem 3 *Briefly describe the shape of each of the four normal plots and state whether the marginal distribution of each series is skewed or symmetric and whether its tails appear normal. If the tails do not appear normal, do they appear heavier or lighter than normal? What conclusions can be made from the Shapiro–Wilk tests? Include the plots with your work.*

The next set of R code creates t-plots with 1, 4, 6, 10, 20, and 30 degrees of freedom and all four indices. However, for the remainder of this lab, only the DAX index will be analyzed. Notice how the reference line is created by the `abline()` function, which adds lines to a plot, and the `lm()` function, which fits a line to the quantiles. The `lm()` function is discussed in Chap. 9.

```
 1 n=dim(logR)[1]
 2 q_grid = (1:n) / (n + 1)
 3 df_grid = c(1, 4, 6, 10, 20, 30)
 4 index.names = dimnames(logR)[[2]]
 5 for(i in 1:4)
 6 {
 7   # dev.new()
 8   par(mfrow = c(3, 2))
 9   for(df in df_grid)
10   {
11     qqplot(logR[,i], qt(q_grid,df),
12        main = paste(index.names[i], ", df = ", df) )
13     abline(lm(qt(c(0.25, 0.75), df = df) ~
14        quantile(logR[,i], c(0.25, 0.75))))
15   }
16 }
```

If you are running R from Rstudio, then line 7 should be left as it is. If you are working directly in R, then remove the “#” in this line to open a new window for each plot.

Problem 4 *What does the code* `q.grid = (1:n) / (n + 1)` *do? What does* `qt(q.grid, df = df[j])` *do? What does* `paste` *do?*

Problem 5 *For the DAX index, state which choice of the degrees of freedom parameter gives the best-fitting t-distribution and explain why.*

Run the next set of code to create a kernel density estimate and two parametric density estimates, t with `df` degrees of freedom and normal, for the DAX index. Here `df` equals 5, but you should vary `df` so that the t density agrees as closely as possible with the kernel density estimate.

At lines 5–6, a robust estimator of the standard deviation of the t-distribution is calculated using the `mad()` function. The default value of the argument `constant` is 1.4826, which is calibrated to the normal distribution since $1/\Phi^{-1}(3/4) = 1.4826$. To calibrate to the t-distribution, the normal quantile is replaced by the corresponding t-quantile and multiplied by `df/(df - 2)` to convert from the scale parameter to the standard deviation.

```
1 library("fGarch")
2 x=seq(-0.1, 0.1,by = 0.001)
```

```
par(mfrow = c(1, 1))
df = 5
mad_t = mad(logR[ , 1],
   constant = sqrt(df / (df - 2)) / qt(0.75, df))
plot(density(logR[ , 1]), lwd = 2, ylim = c(0, 60))
lines(x, dstd(x, mean = mean(logR[,1]), sd = mad_t, nu = df),
   lty = 5, lwd = 2, col = "red")
lines(x, dnorm(x, mean = mean(logR[ ,1]), sd = sd(logR[ ,1])),
   lty = 3, lwd = 4, col = "blue")
legend("topleft", c("KDE", paste("t: df = ",df), "normal"),
   lwd = c(2, 2, 4), lty = c(1, 5, 3),
   col = c("black", "red", "blue"))
```

To examine the left and right tails, plot the density estimate two more times, once zooming in on the left tail and then zooming in on the right tail. You can do this by using the `xlim` parameter of the `plot()` function and changing `ylim` appropriately. You can also use the `adjust` parameter in `density()` to smooth the tail estimate more than is done with the default value of `adjust`.

Problem 6 *Do either of the parametric models provide a reasonably good fit to the first index? Explain.*

Problem 7 *Which bandwidth selector is used as the default by* `density`*? What is the default kernel?*

Problem 8 *For the CAC index, state which choice of the degrees of freedom parameter gives the best-fitting t-distribution and explain why.*

4.10.2 McDonald's Prices and Returns

This section analyzes daily stock prices and returns of the McDonald's Corporation (MCD) over the period Jan-4-10 to Sep-5-14. The data set is in the file `MCD_PriceDail.csv`. Run the following commands to load the data and plot the adjusted closing prices:

```
data = read.csv('MCD_PriceDaily.csv')
head(data)
adjPrice = data[ , 7]
plot(adjPrice, type = "l", lwd = 2)
```

Problem 9 *Does the price series appear stationary? Explain your answer.*

Problem 10 *Transform the prices into log returns and call that series* `LogRet`*. Create a time series plot of* `LogRet` *and discuss whether or not this series appears stationary.*

The following code produces a histogram of the McDonald's log returns. The histogram will have 80 evenly spaced bins, and the argument `freq = FALSE` specifies the density scale.

```
hist(LogRet, 80, freq = FALSE)
```

Also, make a QQ plot of `LogRet`.

Problem 11 *Discuss any features you see in the histogram and QQ plot, and, specifically, address the following questions: Do the log returns appear to be normally distributed? If not, in what ways do they appear non-normal? Are the log returns symmetrically distributed? If not, how are they skewed? Do the log returns seems heavy tailed compared to a normal distribution? How do the left and right tails compare; is one tail heavier than the other?*

4.11 Exercises

1. This problem uses the data set `ford.csv` on the book's web site. The data were taken from the `ford.s` data set in `R`'s `fEcofin` package. This package is no longer on CRAN. This data set contains 2000 daily Ford returns from January 2, 1984, to December 31, 1991.
   (a) Find the sample mean, sample median, and standard deviation of the Ford returns.
   (b) Create a normal plot of the Ford returns. Do the returns look normally distributed? If not, how do they differ from being normally distributed?
   (c) Test for normality using the Shapiro–Wilk test? What is the p-value? Can you reject the null hypothesis of a normal distribution at 0.01?
   (d) Create several t-plots of the Ford returns using a number of choices of the degrees of freedom parameter (df). What value of df gives a plot that is as linear as possible? The returns include the return on Black Monday, October 19, 1987. Discuss whether or not to ignore that return when looking for the best choices of df.
   (e) Find the standard error of the sample median using formula (4.3) with the sample median as the estimate of $F^{-1}(0.5)$ and a KDE to estimate f. Is the standard error of the sample median larger or smaller than the standard error of the sample mean?
2. Column seven of the data set `RecentFord.csv` on the book's web site contains Ford daily closing prices, adjusted for splits and dividends, for the years 2009–2013. Repeat Problem 1 using these more recent returns. One of returns is approximately -0.175. For part (d), use that return in place of Black Monday. (Black Monday, of course, is not in this data set.) On what date did this return occur? Search the Internet for news about Ford that day. Why did the Ford price drop so precipitously that day?

3. This problems uses the `Garch` data set in `R`'s `Ecdat` package.
   (a) Using a solid curve, plot a kernel density estimate of the first differences of the variable `dy`, which is the U.S. dollar/Japanese yen exchange rate. Using a dashed curve, superimpose a normal density with the same mean and standard deviation as the sample. Do the two estimated densities look similar? Describe how they differ.
   (b) Repeat part (a), but with the mean and standard deviation equal to the median and MAD. Do the two densities appear more or less similar compared to the two densities in part (a)?
4. Suppose in a normal plot that the sample quantiles are plotted on the vertical axis, rather than on the horizontal axis as in this book.
   (a) What is the interpretation of a convex pattern?
   (b) What is the interpretation of a concave pattern?
   (c) What is the interpretation of a convex-concave pattern?
   (d) What is the interpretation of a concave-convex pattern?
5. Let `diffbp` be the changes (that is, differences) in the variable `bp`, the U.S. dollar to British pound exchange rate, which is in the `Garch` data set of `R`'s `Ecdat` package.
   (a) Create a 3 × 2 matrix of normal plots of `diffbp` and in each plot add a reference line that goes through the p- and $(1-p)$-quantiles, where $p = 0.25$, 0.1, 0.05, 0.025, 0.01, and 0.0025, respectively, for the six plots. Create a second set of six normal plots using n simulated $N(0,1)$ random variables, where n is the number of changes in `bp` plotted in the first figure. Discuss how the reference lines change with the value of p and how the set of six different reference lines can help detect nonnormality.
   (b) Create a third set of six normal plots using changes in the logarithm of `bp`. Do the changes in `log(bp)` look closer to being normally distributed than the changes in `bp`?
6. Use the following fact about the standard normal cumulative distribution function $\Phi(\cdot)$:
$$\Phi^{-1}(0.025) = -1.96.$$
   (a) What value is $\Phi^{-1}(0.975)$? Why?
   (b) What is the 0.975-quantile of the normal distribution with mean -1 and variance 2?
7. Suppose that $Y_1, \ldots, Y_n$ are *i.i.d.* with a uniform distribution on the interval (0,1), with density function f and distribution function F defined as
$$f(x) = \begin{cases} 1 \text{ if } x \in (0,1), \\ 0 \text{ otherwise}, \end{cases} \quad \textit{and} \quad F(x) = \begin{cases} 0 \text{ if } x \le 0, \\ x \text{ if } x \in (0,1), \\ 1 \text{ if } x \ge 1. \end{cases}$$
Use Result 4.1 to conclude which sample quantile q will have the smallest variance?

References

Abramson, I. (1982) On bandwidth variation in kernel estimates—a square root law. *Annals of Statistics*, **9**, 168–176.

Atkinson, A. C. (1985) *Plots, transformations, and regression: An introduction to graphical methods of diagnostic regression analysis*, Clarendon Press, Oxford.

Bolance, C., Guillén, M., and Nielsen, J. P. (2003) Kernel density estimation of actuarial loss functions. *Insurance: Mathematics and Economics*, **32**, 19–36.

Carroll, R. J., and Ruppert, D. (1988) *Transformation and Weighting in Regression*, Chapman & Hall, New York.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W., Eds. (1983) *Understanding Robust and Exploratory Data Analysis*, Wiley, New York.

Hoaglin, D. C., Mosteller, F., and Tukey, J. W., Eds. (1985) *Exploring Data Tables, Trends, and Shapes*, Wiley, New York.

Jones, M. C. (1990) Variable kernel density estimates and variable kernel density estimates. *Australian Journal of Statistics*, **32**, 361–371. (Note: The title is intended to be ironic and is not a misprint.)

Kleiber, C., and Zeileis, A. (2008) *Applied Econometrics with R*, Springer, New York.

Lehmann, E. L. (1999) *Elements of Large-Sample Theory*, Springer-Verlag, New York.

Scott, D. W. (1992) *Multivariate Density Estimation: Theory, Practice, and Visualization*, Wiley-Interscience, New York.

Serfling, R. J. (1980) *Approximation Theorems of Mathematical Statistics*, Wiley, New York.

Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London.

Tukey, J. W. (1977) *Exploratory Data Analysis*, Addison-Wesley, Reading, MA.

van der Vaart, A. W. (1998) *Asymptotic Statistics*, Cambridge University Press, Cambridge.

Wand, M. P., and Jones, M. C. (1995) *Kernel Smoothing*, Chapman & Hall, London.

Wand, M. P., Marron, J. S., and Ruppert, D. (1991) Transformations in density estimation, *Journal of the American Statistical Association*, **86**, 343–366.

Yap, B. W., and Sim, C. H. (2011) Comparisons of various types of normality tests. *Journal of Statistical Computation and Simulation*, **81**, 2141–2155.