2

Exploratory Data Analysis

2.1 Introduction

This book is concerned with the 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 2.1, 2.2, and 2.3 show, respectively, daily log returns on the S&P 500 index, daily changes in the Deutsch Mark (DM) to U.S. dollar exchange rate, and monthly changes in the risk-free interest rate. We will discuss returns in more detail in Chapter 6, but for now it is enough to know that if P_t is the price on day t, then $\log(P_t/P_{t-1}) = \log(P_t) - \log(P_{t-1})$ is the daily log return on day t.

Despite the large random fluctuations in all three time series,[1] we can see that each series appear **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 Chapter 14.

Each of these time series will be modeled as a sequence $X_1, X_2, \ldots$ of random variables with a CDF equal to F.[2] F will vary between series but,

[1] A time series is a sequence of observations of some quantity or quantities, e.g., equity prices, taken over time.

[2] See Appendix A.3.3 for definitions of CDF, PDF, and other terms used in elementary probability theory.

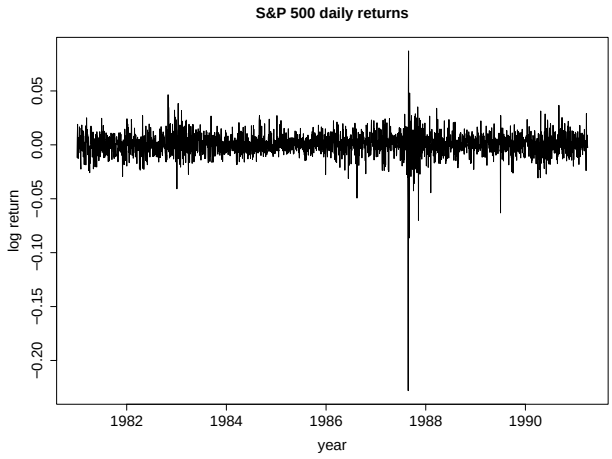


Fig. 2.1. *Daily log returns on the S & P 500 index from Jan 1981 to Apr 1991. This data set is the* `SP500` *series in the* `Ecdat` *package in* `R`.

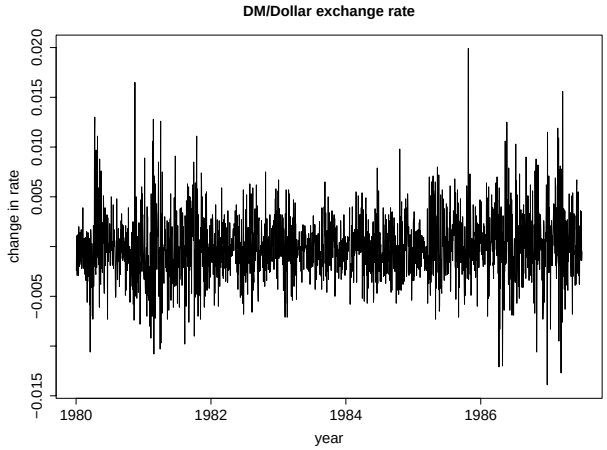


Fig. 2.2. *Daily changes in the DM/dollar exchange rate, Jan 2, 1980 to May 21, 1987. The data come from the* `Garch` *series in the* `Ecdat` *package in* `R`.

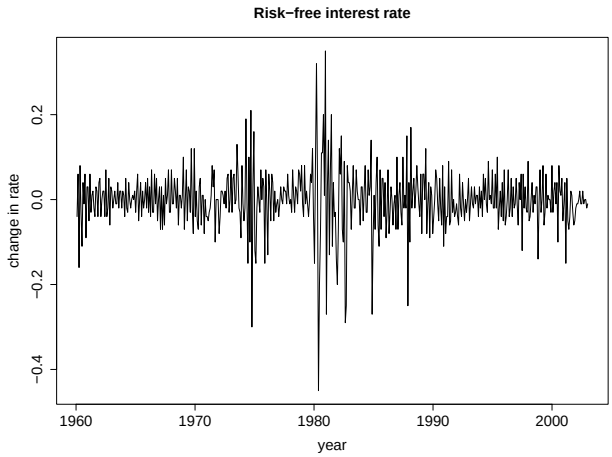


Fig. 2.3. *Monthly changes in the risk-free rate, Jan 1960 to Dec 2002. The data are in the* `Capm` *series in the* `Ecdat` *package in* `R`.

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 time series, we mean the distribution of X_t given no knowledge of the other observations, that is, of X_s for $s \neq t$. Thus, when modeling a marginal distribution, we disregard serial correlation,[3] volatility clustering, and other types of dependency in the time series.[4]

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 and maximum likelihood estimation.

2.2 Histograms and 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 Figure 2.4 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

[3] Serial correlation, also called autocorrelation, is correlation between X_t and X_s for $s \neq t$; see Chapter 7 for further discussion.

[4] However, the marginal distribution of a multivariate time series does include cross-sectional correlations, e.g., the correlation between two equity returns on the same day. See Chapter 5.

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.[5] Panel (b) of Figure 2.4 zooms in on the high probability region. Note that only a few of the 30 cells are in this area.

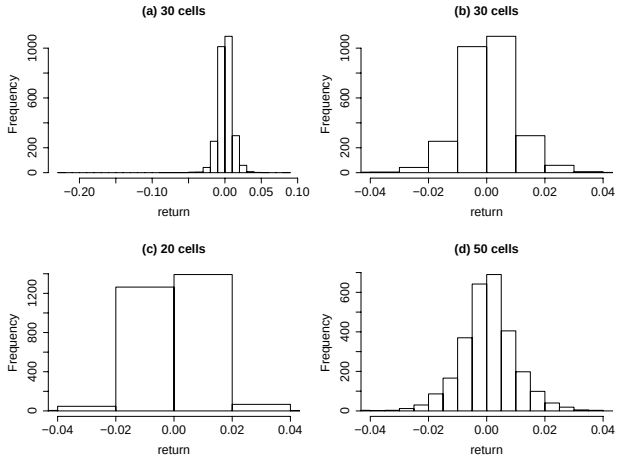


Fig. 2.4. *Histograms of the daily log returns on the S & P 500 index from Jan 1981 to Apr 1991.*

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 Figure 2.4 where panels (b), (c), and (d) differ only in the number of cells. A much better estimator is the kernel density estimator. 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 normal density function is a common choice for K and will be used here. The kernel density estimator based on a $X_1, \ldots, X_n$ is

$$\widehat{f}(x) = \frac{1}{nb} \sum_{i=1}^{n} K\left(\frac{X_i - x}{b}\right),$$

[5] 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.

where b, which is called the bandwidth, determines the resolution of the estimator. A small value of b allows the density estimator to detect fine features in the true density but it also permits a high amount of random variation. 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 trade-off between bias and variance. Fortunately, a large amount of research has been devoted to automatic selection of b. The solid curve in Figure 2.5 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 under-smoothed. The dotted curve is very smooth but under-estimates the peak near 0, a sign of bias. Such a curve is called over-smoothed.

Automatic bandwidth selectors are very useful, but there is nothing magical about them, and often one will use an automatic selector as a starting 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 before was to emphasize the effects of under- and over-smoothing.

The density estimates in Figure 2.5 are bell-shaped suggesting that a normal distribution might be a suitable model for F.[6] Figure 2.6 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 normal density to be too dispersed. 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.[7] 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 the normal distribution. The need for robust estimators is itself a sign of non-normality.

[6] Though we will soon see that there are better models, e.g., t-distributions.
[7] See Section A.15.4 for more discussion of robust estimation.

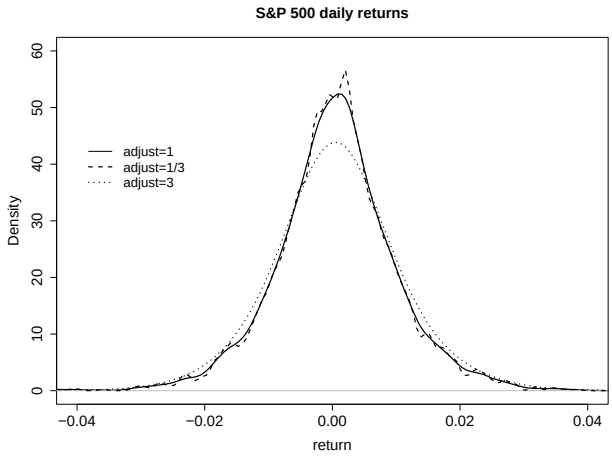


Fig. 2.5. *Kernel density estimates of the daily log returns on the S & P 500 index using three bandwidths. Each bandwidth is the default bandwidth times* `adjust` *and* `adjust` *is 1/3, 1, and 3.*

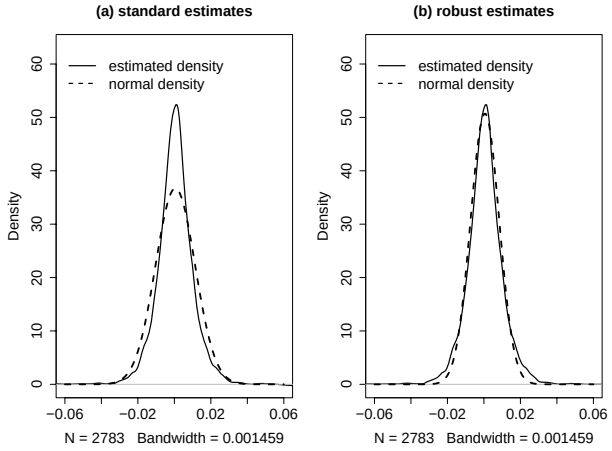


Fig. 2.6. *Kernel density estimates of the daily log returns on the S & P 500 index compared with normal densities. (a) The normal density uses the sample mean and standard deviation. (b) The normal density uses the sample median and MAD estimate of standard deviation.*

We have just seen a problem with using a kernel density estimator 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 Sections 2.3.1 and 2.3.2, are better methods for comparing a sample with a theoretical distribution.

2.3 Order Statistics, the Sample CDF, and Sample Quantiles

Suppose that $X_1, \ldots, X_n$ 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* $F_n(x)$ is defined to be the proportion of the sample that is less than or equal to x. For example, if 10 out of 40 $(= n)$ elements of a sample are 3 or less, then $F_n(3) = 0.25$. More generally,

$$F_n(x) = \frac{\sum_{i=1}^{n} I\{X_i \le x\}}{n},$$

where $I\{X_i \le x\}$ is 1 if $X_i \le x$ and is 0 otherwise. Figure 2.7 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 *order statistics* $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$ are the values $X_1, \ldots, X_n$ ordered from smallest to largest. The subscripts of the order statistics are in parentheses to distinguish them from the unordered sample. For example, X_1 is simply the first observation in the original sample while $X_{(1)}$ is the smallest observation in that sample. The *sample quantiles* are defined in various ways by different authors, but roughly the q-sample quantile is $X_{(k)}$ where k is qn rounded to an integer. Some authors round up, others round to the nearest integer, and still others round in both directions and then interpolate the two results.

quantile vs. percentile?

Example 2.1. Suppose the sample is 6, 4, 8, 2, −3, 4. Then $n = 6$, the order statistics are −3, 2, 4, 4, 6, 8, and $F_n(x)$ equals 0 if $x < -3$, equals 1/6 if $-3 \le x < 2$, equals 2/6 if $2 \le x < 4$, equals 4/6 if $4 \le x < 6$, equals 5/6 if $6 \le x < 8$, and equals 1 if $x \ge 8$. Suppose we want the 25th sample percentile. Note that $.25n = 1.5$ which could be rounded to either 1 or 2. Since 16.7% of the sample equals $X_{(1)}$ or less and 33.3% of the sample equals $X_{(2)}$ or less, either $X_{(1)}$ or $X_{(2)}$ or some number between them can be used as the 25th sample percentile.

The qth quantile is also called the $100q$th percentile. Certain quantiles have been given special names. The 0.5 sample quantile is the 50th percentile

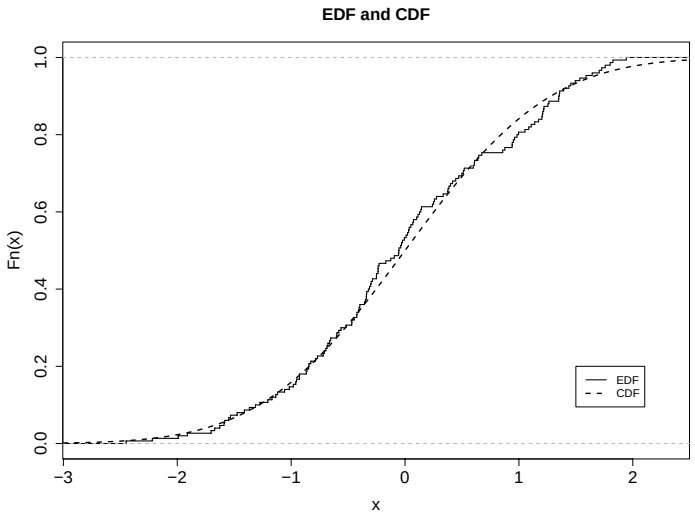


Fig. 2.7. *The EDF F_n (solid) and the true CDF (dashed) from an $N(0,1)$ population. The sample size is* 150.

and is called the median. The 0.25 and 0.75 sample quantiles are called the 1st and 3rd quartiles, and the median is also called the 2nd quartile. The 0.2, 0.4, 0.6, and 0.8 quantiles are the quintiles, and the 0.1, 0.2, . . ., 0.9 quantiles are the deciles.

2.3.1 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. The normal probability plot is a plot of $X_{(i)}$ versus $\Phi^{-1}\{i/(n+1)\}$. (These are the $i/(n+1)$ sample and population quantiles, respectively.) Systematic deviation of the plot from a straight line is evidence of nonnormality.

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. `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.

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 the top, left plot in Figure 2.8. Conversely, if the slope decreases as one moves from left to right, then the curve is concave; see the top, right plot in Figure 2.8. 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 the bottom plots in Figure 2.8.

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).

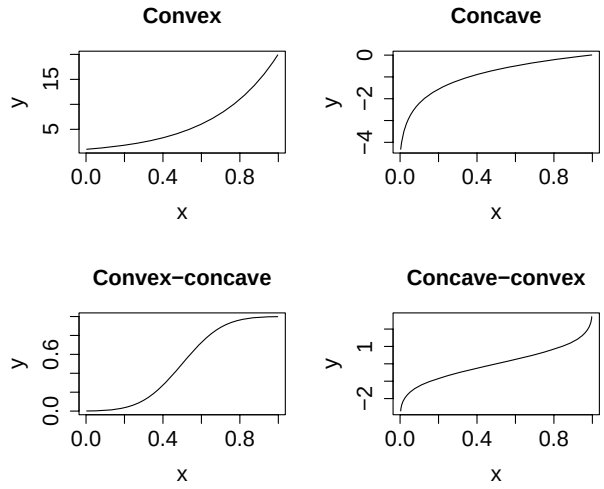


Fig. 2.8. *As one moves from top, left to bottom, right the curves are convex, concave, convex-concave, and concave-convex. Normal plots with these patterns indicate, left-skewness, right-skewness, heavy-tails, and light-tails, respectively, assuming that the data are on the x-axis and the normal quantiles on the y-axis.*

Figure 2.9 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, for each sample size two independent samples are shown. The plots are close to linear, but not exactly linear because of random variation. Even for normally distributed data, some deviation from linearly 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.

Often, a reference line is added to the normal plot to help the viewer determine whether the plot is reasonably linear. Various lines can be used. One choice is the line going through the pair of 1st quartiles and the pair of 3rd quartiles.

Figure 2.10 contains normal probability plots of samples of size 150 from lognormal $(0, \sigma^2)$ distributions,[8] with $\sigma = 1$, $1/2$, and $1/5$ The concave shapes in Figure 2.10 indicates 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 samples sizes.

Figure 2.11 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 are outlier-prone, meaning that the extreme observations on both the left and right sides are significantly more extreme than they would be 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 10 degrees of freedom, and the tails of the t-distribution with 30 degrees of freedom is not that much different than the tails of a normal distribution. These are 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[9], so none of the samples are 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.

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 Figure 2.9 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 this case, a test of normality is very helpful. These tests are discussed in Section 2.4.

[8] See Section A.8.3 for an introduction to the lognormal distribution.

[9] However, t-distributions have been generalized to the so-called skewed-t distributions which need not be symmetric.

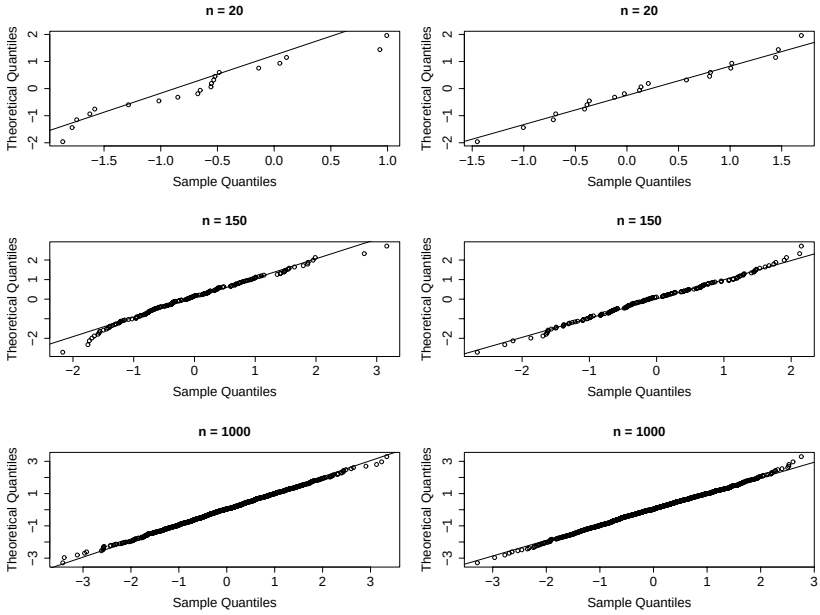


Fig. 2.9. *Normal probability plots of random samples of size* 20*,* 150*, and* 1000 *from a* $N(0,1)$ *population.*

2.3.2 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 Section 3.2.2. 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 2.12 contains a normal plot of the S&P 500 log returns in panel (a) and t-plots with 2, 4, and 15 df in panels (b)–(d). 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 9 returns in the left tail and 4 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

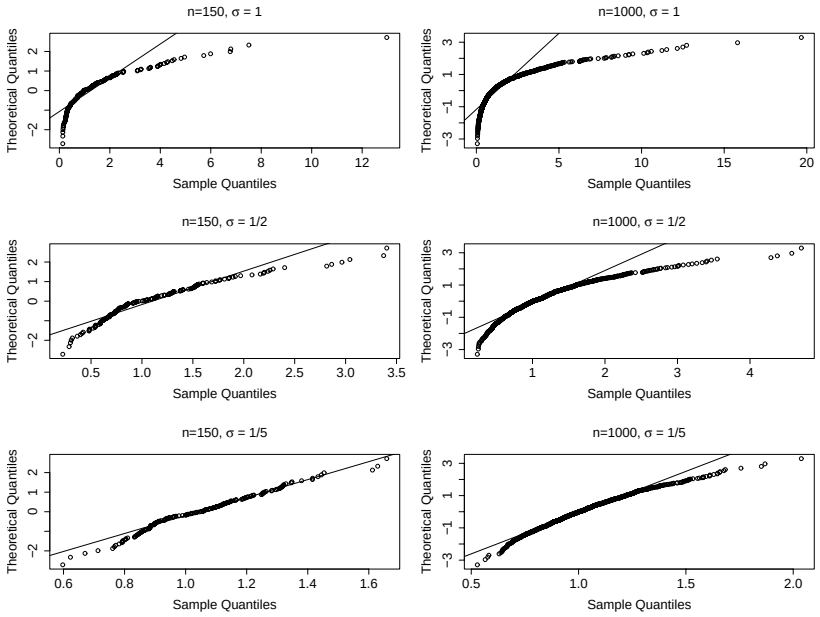


Fig. 2.10. *Normal probability plots of random samples of sizes* 150 *and* 1000 *from lognormal populationa with* $\mu = 0$ *and* $\sigma = 1,\ 1/2,\ or\ 1/5$.

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[10] implies that a daily log return of -0.223, 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 should not have occurred, and anyone using that model should be mindful that it did.

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 a range of samples 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-

[10] See Section A.15.1.

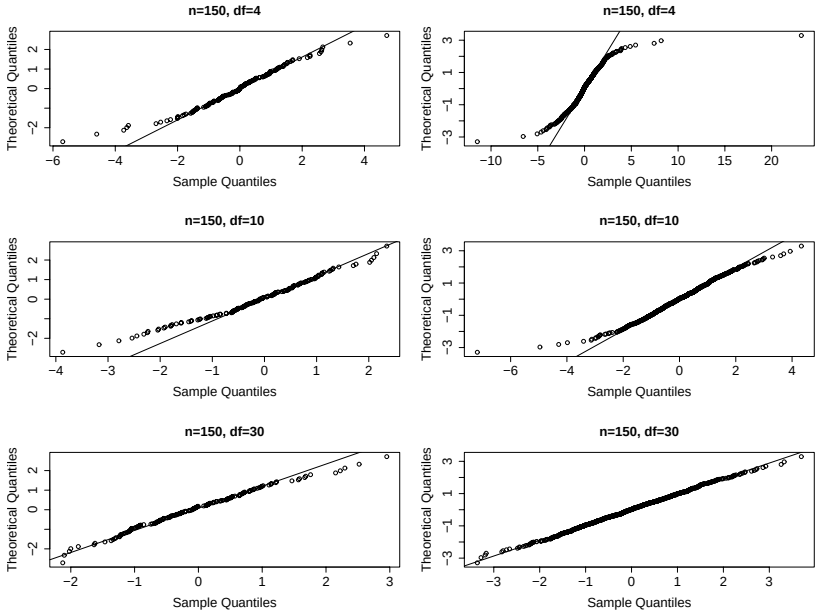
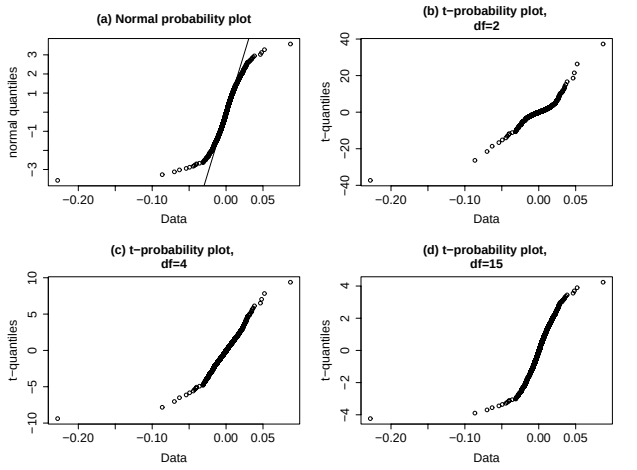


Fig. 2.11. *Normal probability plot of a random sample of size* 150 *and 1000 from a t-distribution with 4, 10, and 30 degrees of freedom.*



where are the lines?

Fig. 2.12. *Normal and t probability plots of the daily returns on the S & P 500 index from Jan 1981 to Apr 1991. This data set is the* `SP500` *series in the* `Ecdat` *package in* `R`.

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 1st and 3rd quartiles, is often added for reference.

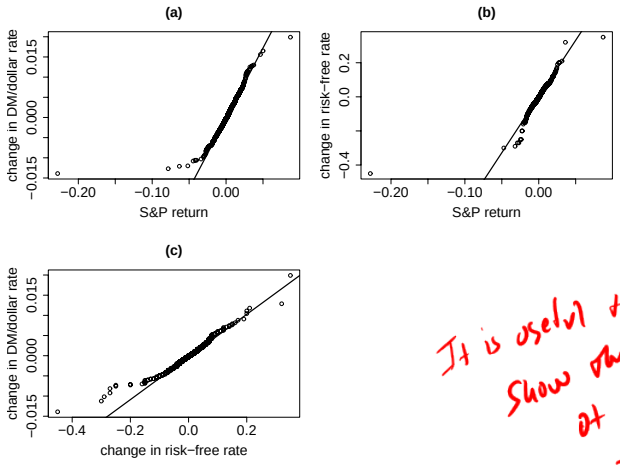


Fig. 2.13. *Sample QQ plots. The straight lines pass through the 1st and 3rd sample quantiles.*

Figure 2.13 contains sample QQ plots for all three combinations of the three time series, S&P 500 returns, changes in the DM/dollar rate, and changes in the risk-free rate, used as examples in this chapter. One sees that the S&P 500 returns have heavier tails than the other two series. The changes in DM/dollar and risk-free rates have somewhat similar shapes, but the changes in the risk-free rate have a slightly heavier left tail.

2.4 Tests of normality

When viewing a normal probability plot, it is often difficult to judge whether any deviation from linearity is systematic or 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 uses the normal probability plot to test these hypotheses. Specifically, the Shapiro-Wilk test is based on the correlation between $X_{(i)}$ and $\Phi^{-1}\{i/(n+1)\}$, which are the i/n

quantiles of the sample and of the standard normal distribution, respectively. Under normality, the correlation should be close to 1 and the null hypothesis of normality is rejected for small values of the correlation coefficient.

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 $\overline{X}$ and variance equal to s_X^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.

For the S&P 500 returns, the Shapiro-Wilks test rejects the null hypothesis of normality with a p-value less than 2.2×10^{-16}. The Shapiro-Wilks also strongly rejects normality for the changes in DM/dollar rate and for the changes in risk-free rate. With large sample sizes, e.g., 2783, 1866, and 515, for the S&P 500 returns, changes in DM/dollar rate, and change is risk-free rate, respectively, it is quite likely that normality will be rejected. In such cases, 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 of practical significance.[11]

2.5 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 Figure 2.14 were created by `boxplot` with default choice of tuning parameters. The "box" in the middle of each plot extends from the 1st to the 3rd quartiles and thus gives the range of the middle half of the data, often called the interquartile range or IQR. The "whiskers" are the vertical 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.[12] The ends of the whiskers are indicated by horizontal lines. All observations between the whiskers are plotted with a "o". The most obvious differences between the three boxplots in Figure 2.14 are differences in scale, and these obscure differences in shape.

[11] See Chapter 13 for discussion on how tail weight can greatly affect risk measures such as VaR and expected shortfall.

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

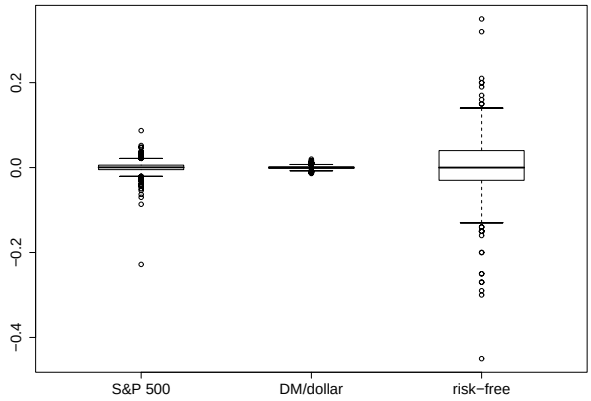


Fig. 2.14. *Boxplots of the S&P 500 daily log returns, daily changes in the DM/dollar exchange rate, and monthly changes in the risk-free rate.*

In Figure 2.15 the three series have been standardized by subtracting the median and then dividing that difference by the MAD estimate of the standard deviation. Now, differences in shape are much 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 is quite detached from the remaining data.

When comparing several samples, boxplots and QQ plots provide different looks at the data. It is best to use both. However, if there are N samples, then the number of QQ plots is $N(N-1)/2$.[13] This number can quickly get out of hand, so, for large values of N, one might use boxplots augmented with a few selected QQ plots.

2.6 Summary

2.7 Bibliographic Notes

2.8 References

2.9 Problems

[13] It is $N(N-1)$ if one includes both plots that are possible for each pair of samples.

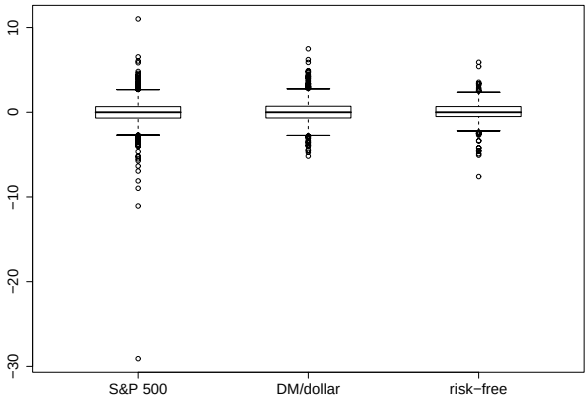


Fig. 2.15. *Boxplots of the standardized S&P 500 daily log returns, daily changes in the DM/dollar exchange rate, and monthly changes in the risk-free rate.*

Thoughts:

Distributions shapes change with data aggregation

Daily – non normal

Monthly – approx normal