STATISTICAL METHODS WITH APPLICATION TO FINANCE

a.y. 2025-2026

Behaviour of Returns: Exploratory analysis

R. Pappadà

rpappada@units.it

Deams

Department of Economics, Business, Mathematics and Statistics "B. de Finetti"

University of Trieste

April 1, 2026

Table of Contents

Time series plot

A **time series** can be defined as a sequence of observations of some quantity taken over time (e.g. prices, interest rates, log returns)

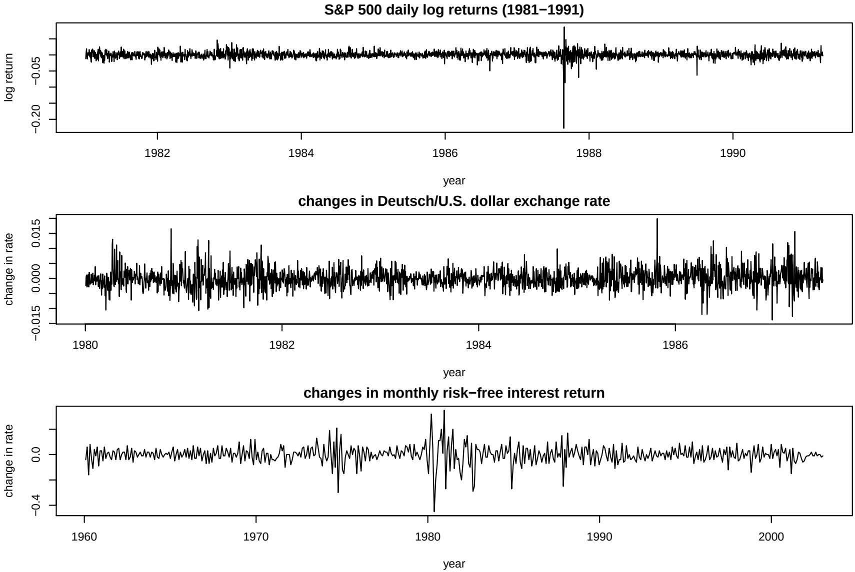
Graphs are useful tools in analyzing financial data.

A *time series plot* is a plot of a time series in chronological order

- when a time series appears *stationary*, then the nature of its random variation is assumed constant over time.

The choice of scales, the size of the intercept and the way that the points are plotted, may substantially affect the way the plot "looks".

Examples


S&P 500 daily log returns (1981−1991)
log return
−0.20
−0.05
1982
1984
1986
1988
1990
year
changes in Deutsch/U.S. dollar exchange rate
change in rate
−0.015
0.000
0.015
1980
1982
1984
1986
year
changes in monthly risk−free interest return
change in rate
−0.4
0.0
1960
1970
1980
1990
2000
year

Marginal distribution

Each of these time series will be modelled as a sequence $Y_1, Y_2, \ldots$ of continuous random variables, each with a marginal distribution function $F(y) = Pr(Y_t \leq y)$, that is assumed to be constant within each series.

The marginal distribution of a *stationary time series*, is the distribution of Y_t, given no information about the previous or future observations.

$\Rightarrow$ When modeling a marginal distribution, we disregard dependencies in the time series (*unconditional distribution*).

Dependencies such as autocorrelation and volatility clustering will be discussed later, for basic analysis of financial time series.

Other graphical tools

Consider the S&P 500 series of daily log returns (January 1981 - April 1991) shown before. The log returns $\{r_1, r_2, \ldots, r_T\}$ are computed as first differences of log closing index values, that is

$$r_t = \log(I_t) - \log(I_{t-1})$$

To gain a better visualization of the distribution of the returns, we can employ

- the histogram or empirical density function of the data
- quantile-quantile plots

Histograms

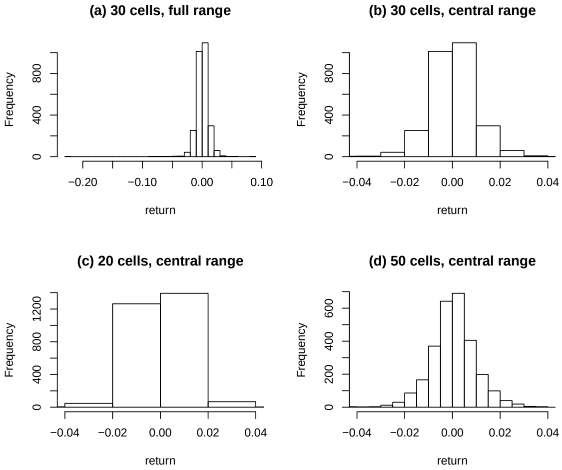


Figure 1: Histograms of the daily log returns on the S&P 500 index from January 1981 to April 1991. Panels (b), (c), and (d) differ only in the number of bins.

Kernel Density Estimation

The empirical density function can be regarded as a refined version of the histogram. It is obtained by a nonparametric smoothing method.

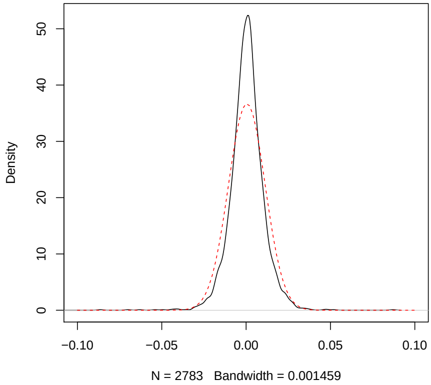


Figure 2: Kernel density estimate (solid) of the daily log returns on the S&P 500 index compared with normal density (dashed).

S&P 500 statistics

Minimum	-0.2280
Maximum	0.0871
1st Quartile	-0.0048
3rd Quartile	0.0058
Mean	0.0004
Median	0.0005
Stdev	0.0109
Skewness	-3.4957
Kurtosis	74.4002

Table 1: S&P 500 daily return summary statistics (1981–1991).

Table of Contents

QQ plots

A **probability plot** or **quantile-quantile (Q-Q) plot** is a standard visual tool for showing the relationship between empirical quantiles of the data and theoretical quantiles of a reference distribution, with a lack of linearity showing evidence against the hypothesized reference distribution.

Given an asset return series, we can check the normality assumption by means of

- *Normal Probability Plots* or Normal QQ plots, that are special cases of quantile-quantile plots
- formal statistical tests (e.g. Jarque-Bera, Shapiro-Wilk) when it is difficult to decide whether a normal plot is close enough to linear to conclude that the data are normally distributed

Empirical cumulative distribution function

Quantiles are points in data below which a certain proportion of the data fall. If we sort the values $\{r_1, r_2, \ldots, r_T\}$ from the smallest to largest observation we obtain the order statistics of the returns:

$$r_{(1)} \leq r_{(2)} \leq \cdots \leq r_{(T)}$$

The *empirical cumulative distribution function* (or sample CDF) $F_T(t)$ at the ordered values

$$r_{(1)} \quad r_{(2)} \quad \cdots \; \cdots \quad r_{(T-1)} \quad r_{(T)}$$

is evaluated

$$\frac{1}{T} \quad \frac{2}{T} \quad \cdots \; \cdots \quad \frac{T-1}{T} \quad \frac{T}{T}$$

that is, the fraction of observations that are less than or equal to the specified value:

$$F_T(t) = \frac{\text{number of } r_i \leq t}{T}$$

$\Rightarrow r_{(i)}$ can be viewed as the quantile of order $q = i/T$, for $i = 1, \ldots, T$.

ECDF: Example

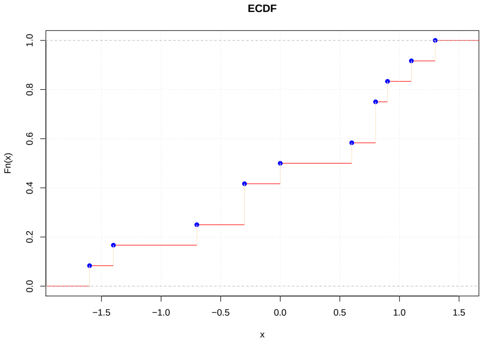


Figure 3: Empirical CDF (F_n) for a sample of size $n = 12$.

Normal Q-Q Plots

Normal QQ plots compare the quantiles of the normal distribution (theoretical quantiles) against the quantiles computed from the data (sample quantiles):

1. order statistics of returns: $r_{(1)} \leq r_{(2)} \leq \cdots \leq r_{(T)}$.

   These are the sample quantiles of order $(i - 0.5)/T$, where the subtraction of 0.5 from i is used to avoid $z = +\infty$, when $i = T$.

2. compare with quantiles of order $(i - 0.5)/T$ of the corresponding normal distribution, $y_i = \mu + \sigma z_{(i-0.5)/T}$, by plotting the points

$$\left(r_{(i)}, y_i\right), \text{ for } i = 1, \ldots T$$

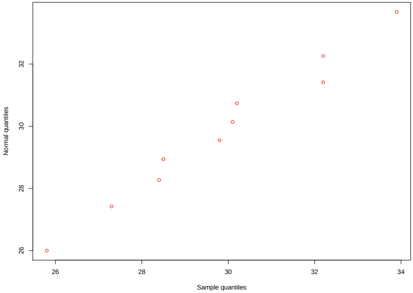
Normal Q-Q Plot: Example

Consider the observations

$$(29.8, 32.2, 30.2, 25.8, 32.2, 27.3, 33.9, 28.4, 28.5, 30.1)$$

We obtain the normal quantile-quantile plot by plotting the points with coordinates x and y in the table (the normal quantiles are computed using mean 29.84 and standard deviation 2.33).

i	$x_{(i)}$	$\frac{i-0.5}{10}$	$y_{(i-0.5)/10}$
1	25.80	0.05	26.00
2	27.30	0.15	27.42
3	28.40	0.25	28.27
4	28.50	0.35	28.94
5	29.80	0.45	29.55
6	30.10	0.55	30.13
7	30.20	0.65	30.74
8	32.20	0.75	31.41
9	32.20	0.85	32.26
10	33.90	0.95	33.68



Normally distributed data

The normality of the data can be evaluated by observing the extent in which the points appear on a reference line.

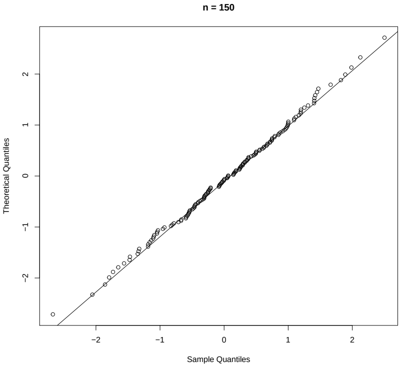


Figure 4: Normal Q-Q plot of a simulated sample of size 150 from an $\mathcal{N}(0, 1)$ population. The reference lines pass through the first and third quartiles.

Right-skewed data

Below is an example of data (200 observations) that are drawn from a distribution that is right-skewed (in this case, the lognormal with log-mean 0 and log-standard deviation $\sigma = 1, 1/2$).

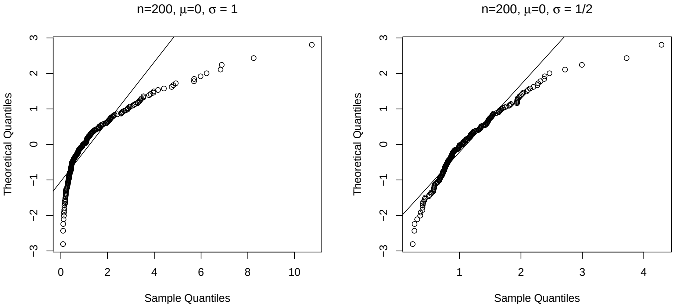


Figure 5: Normal Q-Q plots of samples of size 200 from lognormal distributions.

The concave shapes indicate right skewness. Note that the skewness is less pronounced when $\sigma = 1/2$.

Heavy-tailed data

Data are drawn from t-distributions with 3 and 10 degrees of freedom. If the data are plotted on the x-axis, the Q-Q plot of heavy-tailed data appears S shaped.

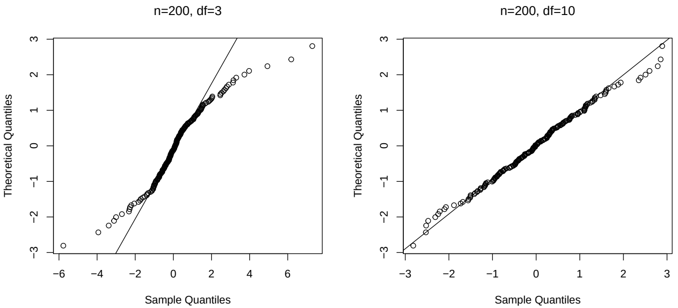


Figure 6: Normal Q-Q plots of samples of size 200 from t-distributions.

Tails are heavier in the sample with 3 degrees of freedom compared to the sample with 10 df.

Tests of Normality

Jarque-Bera test: compares the sample skewness and kurtosis to 0 and 3, their values under normality.

The **null hypothesis** for the test is that the data is normally distributed; the test statistic uses the *sample skewness and kurtosis coefficients*. If $\widehat{sk}$ and $\widehat{kur}$ are estimates of the theoretical skewness and kurtosis, respectively, then

$$JB = \frac{(\widehat{sk})^2}{6/T} + \frac{(\widehat{kur} - 3)^2}{24/T}$$

having an asymptotic chi-squared distribution with two degrees of freedom under the null hypothesis of normality.

Sample kurtosis values differing widely from three and skewness values differing widely from zero may lead to rejection of normality.