31

Market Parameter from Historical Time Series

Having shown in the previous sections how statistical parameters such as the volatility can be obtained *implicitly* from the prices of derivatives traded in the market (if they are not quoted directly anyway), we now proceed with presenting in the following section, how such statistical figures, one of which is the volatility, could be determined by analyzing the historical time series. These volatilities are called *real world* or *historic* volatilities in contrast to *implicit* or *risk-neutral* volatilities.

Depending on the usage, either risk-neutral or real-world volatilities are applied. The risk-neutral or arbitrage-free valuation of derivatives requires risk-neural volatilities (as one might have guessed), since otherwise it would be impossible to match the price quotes of (simple) options, which are used for dynamic hedging that is essential for the risk-neutral valuation approach. On the other hand, the purpose of risk management is to calculate the risk of a potential real-world loss especially of those risks, which are not explicitly (statically) hedged. Here, real-world volatilities are needed.

Time series analysis is a broad topic in the field of *statistics* whose application here will be limited to those areas which serve the purposes of this book. A much more general and wide-reaching presentation can be found in [86], for example.

H.-P. Deutsch, M. W. Beinker, *Derivatives and Internal Models*, Finance and Capital Markets Series, https://doi.org/10.1007/978-3-030-22899-6_31

31.1 Historical Averages as Estimates for Expected Values

Based on a *time series* of some risk factor (e.g. a stock price), we start with calculating for every price change the logarithm of the ratio of new and old price, which equals in linear approximation the *relative* price change, see Eq. 30.9. The statistical measures we estimate are then related to the logarithm of risk factor, e.g. the stock price. This is always handy, if we consider lognormally distributed risk factors. For stock prices, this is the case, at least in first order approximation. Interest rates used to be considered as lognormally distributed, too, as long as the interest rates does not become to low or even negative. For normally distributed risk factors (e.g. interest rates in domains with low or negative interest rates), the *absolute* difference of price for risk factor changes would be used. The statistical methods described below work in both cases, and for a historical as well as for a simulated time series.

From the expectation and the variance of this variable the *historical volatility* σ and the *historical mean return* μ can be determined by recording the relative changes over a period of length δt along a historical (or simulated) path:

$$\mu = \frac{1}{\delta t}\,\mathrm{E}[X]\ , \quad \sigma^2 = \frac{1}{\delta t}\,\mathrm{Var}[X]$$
$$\text{with}\ \ X = \ln\left(\frac{S(t+\delta t)}{S(t)}\right) \approx \frac{S(t+\delta t) - S(t)}{S(t)}\ . \tag{31.1}$$

The *historical correlation* between two price processes is likewise determined from the relative changes. With Y denoting the change of the second price (analogous to X as defined above), the correlation, as defined in Eq. A.14, is

$$\rho = \frac{\mathrm{cov}[XY]}{\sqrt{\mathrm{Var}[X]\,\mathrm{Var}[Y]}}\ .$$

These equations make it apparent that a procedure is required enabling the determination of the expectation and variance from time series data. It is well known from statistics that from n measurements X_i, which are realizations of a random variable X, the *mean* $\langle X\rangle$ (and more generally the mean $\langle f(X)\rangle$ of a function of X) can be computed in the following way

$$\langle X\rangle = \frac{1}{n}\sum_{i=1}^{n} X_i\ , \quad \langle f(X)\rangle = \frac{1}{n}\sum_{i=1}^{n} f(X_i)\ .$$

The *law of large numbers* states that these means approximate the desired *expectations* and variances as follows[1]:

$$\begin{aligned} \langle f(X) \rangle &\overset{n\to\infty}{\longrightarrow} \mathrm{E}[f(X)] \\ \left\langle f(X)^2 \right\rangle - \langle f(X) \rangle^2 &\overset{n\to\infty}{\longrightarrow} \frac{n-1}{n} \operatorname{var}[f(X)] \ . \end{aligned} \tag{31.2}$$

This result may seem trivial or the difference between the actual parameters (the right-hand side) and the measured approximation (left-hand side) may be unclear to the reader. This difference, however, is fundamental: the theoretical value (the right-hand side) can never be precisely known; it can only be *estimated* more or less exactly through the computation of means of measured data. Such "means of measured data" are called *estimators* in statistics to distinguish them from the true values. Examples of estimators are the expectations on the left-hand sides in Eq. 31.2. More estimators will be introduces in the following sections.

The expectation and variance of X are needed for the description of the risk factor X. For this, the means of the realizations of the random variables X and X^2 are needed. The determination of the *error* made in making these estimates requires X^4 as well (see Sect. 31.2). Therefore, for any time series analysis the computation of the following measures are especially helpful:

$$\langle X \rangle = \frac{1}{n} \sum_{i=1}^{n} X_i \ , \quad \left\langle X^2 \right\rangle = \frac{1}{n} \sum_{i=1}^{n} X_i^2 \ , \quad \left\langle X^4 \right\rangle = \frac{1}{n} \sum_{i=1}^{n} X_i^4 \ . \tag{31.3}$$

These measures are called the 1., 2. and 4. *moment* of the distribution. In general, the n-th moment m_k of a distribution is defined as

$$m_k = \frac{1}{n} \sum_{i=1}^{n} X_i^k \ .$$

If multiple correlated prices are involved, the following means for each pair of prices are also necessary:

$$\langle XY \rangle = \frac{1}{n} \sum_{i=1}^{n} X_i Y_i \ , \quad \left\langle (XY)^2 \right\rangle = \frac{1}{n} \sum_{i=1}^{n} X_i^2 Y_i^2 \ . \tag{31.4}$$

[1]The factor $(n-1)/n$ of the variance is necessary if the estimator for the variance is to be *unbiased*. See any introductory statistics textbook for more on this subject.

These are the *composite moments* $m_{1,1}$ and $m_{2,2}$ of the time series X and Y. From these values, historical estimates for the mean return, volatility and correlation can be obtained:

$$\mu = \frac{1}{\delta t}\,\mathrm{E}\,[X] \approx \frac{1}{\delta t}\,\langle X\rangle \tag{31.5}$$

$$\sigma = \frac{1}{\sqrt{\delta t}}\sqrt{\mathrm{var}\,[X]} \approx \frac{1}{\sqrt{\delta t}}\sqrt{\frac{n}{n-1}}\sqrt{\langle X^2\rangle - \langle X\rangle^2}$$

$$\rho = \frac{\mathrm{cov}\,[XY]}{\sqrt{\mathrm{var}\,[X]}\,\sqrt{\mathrm{var}\,[Y]}} \approx \frac{\langle XY\rangle - \langle X\rangle\,\langle Y\rangle}{\sqrt{\langle X^2\rangle - \langle X\rangle^2}\,\sqrt{\langle Y^2\rangle - \langle Y\rangle^2}}\,.$$

31.2 Error Estimates

Estimating the error in the measured values is essential for evaluating their meaningfulness. It is therefore insufficient to provide a value as the result of making observations since the actual theoretical value is not obtained (and will never be obtained) from measured data. It is more appropriate to find an interval on the basis of observed data within which the theoretical value lies.

A claim such as "the historical price volatility is 20% per year" is not very meaningful if nothing is said about the error associated with such a claim, for example 0.5% or 50%. A valid statement on the other hand might be "the historical price volatility is 20% $\pm$ 3% per year". This means that the actual volatility lies with high probability somewhere between 17% and 23% per year. The safer you need to be, i.e. the higher the probability should be that the value lies indeed in this interval, the greater the interval will be, if the range of parameter values is not restricted for other reasons (e.g., in case of a dice the result of a valid throw is with 100% probability between 1 and 6).

In this section, several simple methods for determining the statistical error are introduced and its calculation will be demonstrated explicitly for the volatility and correlation. We begin by assuming that data in a time series (referred to as *observed values* or *measurements*) are pair wise independent and in consequence uncorrelated. Finally, we will briefly show how to test whether measurements are independent and how to proceed in the case that they are not, i.e., how to account for autocorrelations. Naturally, the subject is quite technical. Error estimation is theoretically quite simple but lengthy. Nonetheless, anyone who wishes to conduct a serious analysis of historical or simulated data should understand and apply this material.

There are two different types of error. The first is the *statistical error*, which is a consequence of the fact that only a finite number of measurements are taken. The second is the *systematic error*. These are errors arising from a fundamental mistake (for example, a programming error in a Monte Carlo simulation). The failure to decrease with an increasing number of measurements is characteristic of a systematic error (as opposed to the statistical error). Only the statistical error will be dealt with in this section.

31.2.1 Uncorrelated Measurements

The determination of expectations is accomplished through calculating the mean of the observed values as in Eq. 31.5. For a sufficiently large number n of measurements

$$\mathrm{E}[X] \approx \langle X \rangle \ , \quad \mathrm{var}[X] \approx \frac{n}{n-1} \left(\left\langle X^2 \right\rangle - \langle X \rangle^2 \right) \tag{31.6}$$

holds. The central question is: what is the (statistical) error involved in estimating these parameters as above? The n observations are realizations of the random variables X_i, $i = 1, \ldots n$. The mean is the weighted sum over these random variables X_i and as such is again a random variable. The statistical error will be defined as the *standard deviation* of this new random variable, or equivalently, the error is the square root of the variance of the mean. A fundamental result from the field of statistics is

$$\begin{aligned} \langle X \rangle &= \frac{1}{n} \sum_{i=1}^{n} X_i \quad \Rightarrow \\ \mathrm{var}[\langle X \rangle] &= \frac{1}{n} \mathrm{var}[X] \quad \text{if } \mathrm{E}\left[X_i X_j\right] = 0 \quad \text{for } i \neq j \ . \end{aligned} \tag{31.7}$$

The variance of the *mean* of uncorrelated, identically distributed random variables is equal to the variance of the random variable itself divided by the number of observations. This result combined with the approximation in Eq. 31.6 for the variance yields the statistical error (denoted below by the symbol δ) defined as the standard deviation of the mean

$$\delta \langle X \rangle \equiv \sqrt{\mathrm{var}[\langle X \rangle]} = \sqrt{\frac{1}{n} \mathrm{var}[X]} \approx \sqrt{\frac{\left\langle X^2 \right\rangle - \langle X \rangle^2}{n-1}} \ . \tag{31.8}$$

This holds in general for the mean of a function of the random variable X:

$$\delta \langle f(X) \rangle = \sqrt{\frac{1}{n} \operatorname{var}[f(X)]} \approx \sqrt{\frac{\langle f(X)^2 \rangle - \langle f(X) \rangle^2}{n-1}} \,. \tag{31.9}$$

The statistical error in the estimation of the mean return as defined in Eq. 31.5 is thus[2]

$$\delta\mu \approx \frac{1}{\delta t \sqrt{n-1}} \sqrt{\langle X^2 \rangle - \langle X \rangle^2} \,. \tag{31.10}$$

The mean of a function is to be distinguished from the function of the mean if the function is not linear: $\langle f(X) \rangle \neq f(\langle X \rangle)$. Thus, the *error of a function* of an uncertain value z (in the case under discussion, $z = \langle X \rangle$) cannot be computed directly in general. A Taylor series expansion of the function can provide assistance in such cases. The *propagation of error* can be obtained from this Taylor series:

$$f = f(z) \text{ mit } z = z \pm \delta z \quad \Rightarrow \quad \delta f = \left| \delta z \frac{\partial f}{\partial z} \right| + \frac{1}{2} \left| (\delta z)^2 \frac{\partial^2 f}{\partial z^2} \right| + \cdots$$

The vertical lines in the above equation indicate the absolute value. This can be generalized for functions of multiple variables with associated errors. Such expressions quickly become very lengthy. Restricting our consideration to the first (linear) terms, we obtain what is known as the *quadratic error propagation.*

$$f = f(z_1, z_2, \ldots, z_k) \quad \text{with} \quad z_i = z_i \pm \delta z_i$$

$$\Rightarrow \quad \delta f \approx \left| \sum_{i=1}^{k} \delta z_i \frac{\partial f}{\partial z_i} \right| = \sqrt{\sum_{i=1}^{k} \left(\delta z_i \frac{\partial f}{\partial z_i} \right)^2} \,. \tag{31.11}$$

[2]Here, the δ in δt denotes the length of a time interval between two data points in the time series and *not* the "error in t".

This law will be required in order to determine the error involved in measuring the variance since

$$\begin{aligned}
\operatorname{var}[X] &\approx \frac{n}{n-1}\left(\left\langle X^2\right\rangle - \langle X\rangle^2\right) = f(z_1, z_2) \quad \text{with} \\
z_1 &= \langle X\rangle\,,\ z_2 = \left\langle X^2\right\rangle,\ f(z_1, z_2) = \frac{n}{n-1}\left(z_2 - z_1^2\right) \\
\Rightarrow \quad \frac{\partial f}{\partial z_1} &= -\frac{2n}{n-1} z_1\,,\quad \frac{\partial f}{\partial z_2} = \frac{n}{n-1}\,.
\end{aligned}$$

The error of f, calculated with quadratic error propagation, is then

$$\begin{aligned}
\delta f &\approx \sqrt{\left(\delta z_1 \frac{\partial f}{\partial z_1}\right)^2 + \left(\delta z_2 \frac{\partial f}{\partial z_2}\right)^2} \\
&= \sqrt{\left(\delta z_1 \frac{2n}{n-1} z_1\right)^2 + \left(\delta z_2 \frac{n}{n-1}\right)^2} \\
&= \frac{n}{n-1}\sqrt{4 z_1^2 (\delta z_1)^2 + (\delta z_2)^2}\,.
\end{aligned}$$

Since z_1 and z_2 are means of X and $f(X) = X^2$ respectively, their errors are respectively,

$$\begin{aligned}
z_1 = \langle X\rangle \quad &\Rightarrow \quad \delta z_1 = \delta \langle X\rangle \approx \frac{1}{\sqrt{n-1}}\sqrt{\left\langle X^2\right\rangle - \langle X\rangle^2} \\
z_2 = \left\langle X^2\right\rangle \quad &\Rightarrow \quad \delta z_2 = \delta \left\langle X^2\right\rangle \approx \frac{1}{\sqrt{n-1}}\sqrt{\left\langle X^4\right\rangle - \left\langle X^2\right\rangle^2}\,.
\end{aligned}$$

Substituting these into the expression for δf yields

$$\begin{aligned}
\delta f &\approx \frac{1}{\sqrt{n-1}} \frac{n}{n-1}\sqrt{4\langle X\rangle^2\left(\left\langle X^2\right\rangle - \langle X\rangle^2\right) + \left(\left\langle X^4\right\rangle - \left\langle X^2\right\rangle^2\right)} \\
&= \frac{1}{\sqrt{n-1}} \frac{n}{n-1}\sqrt{\left\langle X^4\right\rangle - \left\langle X^2\right\rangle^2 + 4\left\langle X^2\right\rangle\langle X\rangle^2 - 4\langle X\rangle^4}\,.
\end{aligned}$$

This is the statistical error made in measuring the variance $\mathrm{var}[X]$. In order to determine this error when analyzing historical time series, the means of X, X^2 and X^4 must be measured.

Likewise, the error of the volatility σ can be determined through the following deliberations

$$\sigma\,[X] \;\equiv\; \frac{1}{\sqrt{\delta t}}\sqrt{\mathrm{var}\,[X]} \approx \frac{1}{\sqrt{\delta t}}\sqrt{\frac{n}{n-1}}\sqrt{\langle X^2\rangle - \langle X\rangle^2} = \frac{1}{\sqrt{\delta t}}g(z_1, z_2)$$

$$\text{with}\;\; z_1 = \langle X\rangle \;\;,\quad z_2 = \left\langle X^2\right\rangle \;\;,\quad g(z_1, z_2) = \sqrt{\frac{n}{n-1}}\sqrt{z_2 - z_1^2} \;\;\Rightarrow$$

$$\frac{\partial\, g}{\partial\, z_1} = -\sqrt{\frac{n}{n-1}}\frac{z_1}{\sqrt{z_2 - z_1^2}}\;,\quad \frac{\partial\, g}{\partial\, z_2} = \frac{1}{2}\sqrt{\frac{n}{n-1}}\frac{1}{\sqrt{z_2 - z_1^2}}\;.$$

An analogous calculation as above yields

$$\delta\,\sigma \approx \frac{1}{\sqrt{\delta t}}\frac{1}{\sqrt{n-1}}\sqrt{\frac{n}{n-1}}\frac{\sqrt{\left\langle X^4\right\rangle - \left\langle X^2\right\rangle^2 + 4\left\langle X^2\right\rangle\langle X\rangle^2 - 4\langle X\rangle^4}}{2\sqrt{\langle X^2\rangle - \langle X\rangle^2}} \tag{31.12}$$

for the error in the measured volatility. The expression for the error of the correlation between two prices X and Y is even longer as it is represented by a function of *five* means:

$$\rho[X, Y] \equiv \frac{\mathrm{cov}[XY]}{\sqrt{\mathrm{var}[X]\mathrm{var}[Y]}} \approx \frac{\langle XY\rangle - \langle x\rangle\langle Y\rangle}{\sqrt{\langle X^2\rangle - \langle X\rangle^2}\sqrt{\langle Y^2\rangle - \langle Y\rangle^2}} = \rho(z_1, z_2, z_3, z_4, z_5)$$

$$\text{where}\;\; z_1 = \langle X\rangle,\quad z_2 = \langle Y\rangle,\quad z_3 = \langle X^2\rangle,\qquad z_4 = \langle Y^2\rangle,\quad z_5 = \langle XY\rangle$$

$$\rho = \frac{z_5 - z_1 z_2}{\sqrt{z_3 - z_1^2}\sqrt{z_4 - z_2^2}}$$

The derivatives of the correlation with respect to the z_i are

$$\frac{\partial\rho}{\partial z_1} = -\rho\left(z_1 + \frac{z_2}{z_5 - z_1 z_2}\right)$$

$$\frac{\partial\rho}{\partial z_2} = -\rho\left(z_2 + \frac{z_1}{z_5 - z_1 z_2}\right)$$

$$\frac{\partial \rho}{\partial z_3} = \frac{\partial \rho}{\partial z_4} = \frac{\rho}{2}$$
$$\frac{\partial \rho}{\partial z_5} = \frac{\rho}{z_5 - z_1 z_2}$$

The errors of the z_i are as in Eq. 31.9

$$\delta z_1^2 = \frac{\langle X^2 \rangle - \langle X \rangle^2}{n-1}, \quad \delta z_2^2 = \frac{\langle Y^2 \rangle - \langle Y \rangle^2}{n-1}$$
$$\delta z_3^2 = \frac{\langle X^4 \rangle - \langle X^2 \rangle^2}{n-1}, \quad \delta z_4^2 = \frac{\langle Y^4 \rangle - \langle Y^2 \rangle^2}{n-1}$$
$$\delta z_5^2 = \frac{\langle X^2 Y^2 \rangle - \langle XY \rangle^2}{n-1} .$$

All these results inserted into Eq. 31.11 yields the statistical error of the correlation

$$\partial \rho \approx \sqrt{\sum_{i=1}^{5} \left(\delta z_i \frac{\partial \rho}{\partial z_i} \right)^2} .$$

Table 31.1 illustrates the application of Eq. 31.5 for the measurement of the mean return and the volatility from a data set with $n = 250$ observations and the estimation of their statistical errors in accordance with Eqs. 31.10 and 31.12. In addition to the relative price changes X, the second and fourth powers are measured. Using these means, the above equations are used to compute the mean return and the volatility as well as the errors involved in estimating these two parameters.

The data set was generated by a simulated random walk with a yield of 6.00% and a volatility of 20.00%. The measured values could thus be compared with the "true values" (a luxury naturally not at our disposal when using historical data). The true values lie within the error of the measurement.

The error is naturally large since the number of measurements taken is so small. As the number of measurements gets bigger, we see that the error decreases as the inverse of square root of the number of measurements; a tenfold decrease in the statistical error can thus only be accomplished if a sample size 100 times as large is placed at our disposal.

Table 31.1 Measuring the yield, the volatility and their errors from a (simulated) data series of 250 "measurements". The parameters used to simulate the data set are shown for comparison

	Yield per δt		Vol per δt		
Simulated		6.00%		20.00%	
Measured		5.99%		19.70%	
Error		1.25%		1.00%	
x	$x2$		$x4$		
		Averages			
0.05986719	0.04223756		0.00502627		
		Data			***n***
−0.070,8944,77	0.005,026,027		2.526,09	$\times 10^{-05}$	1
−0.014,768,650	0.000,218,113		4.757,33	$\times 10^{-08}$	2
0.011,417,976	0.000,130,37		1.699,64	$\times 10^{-08}$	3
−0.066,321,993	0.004,398,607		1.934,77	$\times 10^{-05}$	4
−0.113,237,822	0.012,822,804		0.000,164,424		5
0.089,718,194	0.008,049,354		6.479,21	$\times 10^{-05}$	6
0.200,262,728	0.040,105,16		0.001,608,424		7
0.329,164,502	0.108,349,270		0.011,739,564		8
−0.003,557,309	1.265,44	$\times 10^{-05}$	1.601,35	$\times 10^{-10}$	9
−0.066,032,319	0.004,360,267		1.901,19	$\times 10^{-05}$	10
0.266,072,855	0.070,794,764		0.005,011,899		11
−0.173,980,700	0.030,269,284		0.000,916,23		12
0.193,003,141	3.725,02	$\times 10^{-02}$	0.001387578		13
−0.044,716,037	0.001,999,524		3.998,1	$\times 10^{-06}$	14
−0.038,558,758	0.001,486,778		2.2105,1	$\times 10^{-06}$	15
−0.033,664,767	0.001,133,317		1.284,41	$\times 10^{-06}$	16
...	...		...		...

It is essential to be aware that errors are also only statistical quantities. Therefore we can *not* be *sure* that the interval defined by the error actually contains the true value of the estimated parameter; there is only a certain *probability* that this is the case. If we have reason to believe that the measured estimator, such as the mean return, is normally distributed, then there is approximately 68% probability that the true parameter will lie within the error interval, since the error is defined as *one* standard deviation. Thus, the probability for the true value to lie outside the range obtained from the statistical error is approximately 32% in this case. If this uncertainty is too large, we could of course define the statistical error to correspond to two or three standard deviations or more. We only need to multiply the error in the derivations above by the desired multiplicative factor. For example, if we define the error as two standard deviations and the measured estimator is normally distributed, then the probability that the true value will lie within the new error interval is 95.4%.

As is clear from the above, we need the probability distribution of the measured estimator if we want to assign confidence levels to error intervals. It can by no means be taken for granted that the measured estimator has the same distribution as the random variables in the time series. If, for instance, the random variables X in the time series are *uniformly* distributed on a finite interval $[a, b]$ then the estimator for μ as defined in Eq. 31.5 is (for large n) approximately *normally* distributed (because of the *central limit theorem*, see Sect. A.4.3). If, however, the random variables X in the time series are *normally* distributed, then the estimator for μ is also *normally* distributed. However, the estimator for the *variance* is in this case a χ^2-distributed variable. This can be seen as follows: according to Eqs. 31.6 and 31.3, the estimator for the variance is

$$\mathrm{var}[X] \approx \frac{n}{n-1}\left(\left\langle X^2\right\rangle - \langle X\rangle^2\right) = \frac{n}{n-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2 - \langle X\rangle^2\right) .$$

The X_i are all normally distributed and the mean $\langle X\rangle$ is also normally distributed in this case. Therefore the estimator for the variance is a sum of *squared* normally distributed random variables and as such χ^2-distributed (see Sect. A.4.6).

31.2.2 Error of Autocorrelated Measurements

The methods described above for the determination of the statistical error hold only for *uncorrelated* measurements. That means, it has been tacitly assumed that the n measurements in a time series come from *independent* observations. However, independence is an assumption which often cannot be made, in particular in time series analysis (for example, in the case of moving average methods, see Sect. 33.4). Daily changes in a moving 30-day price average, for example (the mean of prices observed over the previous 30 days is computed) will remain small from one day to the next since in each daily adjustment, only the oldest price is replaced by the most recently observed value, the other 29 prices in the average remain the same. The measurement of such a variable is highly correlated with the measurement made on the previous day. In such a situation, error considerations must be modified significantly.

Autocorrelation and Autocovariance

A correlation of one and the same variable *with itself* is called *autocorrelation*.[3] Just as *correlation* measures the (linear) dependency between two *different* random processes, *auto*correlation measures the (linear) dependency between a process has *on itself*. The autocorrelation is defined by

$$\varrho(t,h) = \frac{\mathrm{cov}(X_{t+h}, X_t)}{\mathrm{cov}(X_t, X_t)} = \frac{\mathrm{cov}(X_{t+h}, X_t)}{\mathrm{var}(X_t)} \ , \tag{31.13}$$

where the definition of the *autocovariance* is completely analogous to the definition of the covariance between two different random variables (see Eq. A.10)

$$\begin{aligned} \mathrm{cov}(X_{t+h}, X_t) &= \mathrm{E}\left[(X_{t+h} - \mathrm{E}\left[X_{t+h}\right])\,(X_t - \mathrm{E}\left[X_t\right])\right] \\ &= \mathrm{E}\left[X_{t+h}X_t\right] - \mathrm{E}\left[X_{t+h}\right] E\left[X_t\right] \ . \end{aligned} \tag{31.14}$$

We arrive at the final equality by merely taking the product in the previous expression and using the linearity of the expectation. The resulting equation corresponds to A.11. Definition 31.14 immediately establishes the relation between the autocovariance and the variance

$$\mathrm{cov}(X_t, X_t) = \mathrm{E}[(X_t - \mathrm{E}\left[X_t\right])(X_t - \mathrm{E}\left[X_t\right])] = \mathrm{var}(X_t) \ .$$

which has already been used in establishing the last equality in Eq. 31.13.

The autocorrelation in Eq. 31.13 also corresponds to the ordinary correlation known from statistics. If the time series is stationary[4] (in particular, if the variance of the process remains constant), $\mathrm{var}(X_t) = \mathrm{var}(X_{t+h})$ holds and ρ can be written in a form corresponding to Eq. A.14:

$$\varrho(h) = \frac{\mathrm{cov}(X_{t+h}, X_t)}{\sqrt{\mathrm{var}(X_{t+h})}\sqrt{\mathrm{var}(X_t)}} \ .$$

[3] Autocorrelations do not appear merely in certain measurement methods but are in general inherent to non-Markov processes, i.e. for processes whose current value is influenced by past values. See Sect. 32.1 for more on this topic. It should be noted that correlation measures only linear dependencies, though.

[4] This means intuitively that the parameters describing the time series are time independent, see Chap. 32.

This is just the definition of the correlation of the random variable X_{t+h} with the random variable X_t, irrespective of whether these random variables belong to the same time series.

Autocorrelation Time and Error Estimates

With the autocorrelations above we can estimate an autocorrelation *time* which specifies the number of time steps needed between two measurements in order to guarantee that the two measurements are (at least approximately) uncorrelated. The *autocorrelation time* τ is defined through the autocorrelation Eq. 31.13 in the following manner

$$\tau(t) \equiv \frac{1}{2} \sum_{h=-\infty}^{\infty} \varrho(t,h) = \frac{1}{2} \sum_{h=-\infty}^{\infty} \frac{\mathrm{cov}(X_{t+h}, X_t)}{\mathrm{cov}(X_t, X_t)} \,. \tag{31.15}$$

Usually, only *stationary* time series are investigated (see Chap. 32). Then ϱ is only dependent on the time difference h between the observations and not on the time point t. The autocorrelation time is then likewise independent of t, i.e., a constant. For *un*correlated observations we have $\varrho(t,h) = \delta_{h,0}$ (where $\delta_{h,0}$ again denotes the Kronecker delta) and therefore simply $\tau = 1/2$.

If the number n of observations is much larger than the autocorrelation time τ it can be shown that

$$\delta \langle X \rangle \approx \sqrt{\frac{2\,\tau}{n} \mathrm{var}\,[X]} \approx \sqrt{\frac{2\,\tau}{n-1} \left[\langle X^2 \rangle - \langle X \rangle^2 \right]} \qquad \text{for} \quad n >> \tau \tag{31.16}$$

holds for the mean error of X. This reduces to Eq. 31.8 if the measurements are uncorrelated since then we have $\tau = 1/2$.

The autocorrelation time (and thus the autocorrelations) must be measured if the error in Eq. 31.16 is to be determined. An estimator for the *autocovariance* is given by

$$\mathrm{cov}(X_{t+h}, X_t) \approx \frac{1}{n-|h|} \sum_{\substack{i,j \\ i-j=|h|}} X_i X_j - \left(\frac{1}{n} \sum_{i=1}^{n} X_i \right)^2 \,. \tag{31.17}$$

For $h = 0$ this estimator is similar to Eq. 31.6 but the factor $n/(n-1)$ is not reproduced (for large n however, this factor is very close to 1):

$$\mathrm{var}\,[X_t] = \mathrm{cov}(X_t, X_t) \approx \frac{1}{n}\sum_{\substack{i,j \\ i=j}} X_i X_j - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2 = \left\langle X_t^2 \right\rangle - \langle X_t \rangle^2 \ .$$

As can be seen from Eq. 31.15, the autocorrelations between $h = -\infty$ and $h = -\infty$ must in theory be measured for the calculation of the autocorrelation *time*. In practice, a suitable *cutoff* $(n - \tilde{n})$ can be chosen to limit the sum in Eq. 31.15 to a finite one, neglecting the autocorrelations that are so small that they contribute almost nothing to the sum. Substituting the estimator for the autocovariance, Eq. 31.17, then yields the estimator for the autocorrelation time.

$$\begin{aligned}\tau \ &\approx \frac{1}{2}\sum_{h=-(n-\tilde{n})}^{(n-\tilde{n})} \frac{\mathrm{cov}(X_{t+h}, X_t)}{\mathrm{cov}(X_t, X_t)} \quad \text{with } \tau << \tilde{n} << n \\ &\approx \frac{1}{2}\frac{1}{\left\langle X^2 \right\rangle - \langle X \rangle^2}\sum_{h=\tilde{n}-n}^{n-\tilde{n}} \left[\frac{1}{n-|h|}\sum_{\substack{i,j \\ i-j=|h|}} X_i X_j - \langle X \rangle^2\right] .\end{aligned}$$

All this substituted into Eq. 31.16 finally provides a possibility of estimating the error of a mean taking autocorrelations into account:

$$\delta\,\langle X \rangle \approx \sqrt{\frac{1}{n-1}\sum_{h=\tilde{n}-n}^{n-\tilde{n}} \left[\frac{1}{n-|h|}\sum_{\substack{i,j \\ i-j=|h|}} X_i X_j - \langle X \rangle^2\right]} \ .$$

There are thus two possibilities of taking the autocorrelations into consideration:

- We wait for at least as long as the autocorrelation time before taking a second measurement; the measurements would then be uncorrelated.
- We do *not* wait for the autocorrelation time to pass (e.g., because it is to short, for instance) and use the above expression for the error of the correlated observations.

31.3 Return and Covariance Estimates

We will now present some commonly used estimators for returns and covariances. The importance of return estimates is unquestionable for any investment decision. Just as important are covariance estimates for quantifying the risk, as shown for instance in Sect. 21.5. In addition, estimates for other quantities like volatilities, correlations and Betas can be derived from the covariance estimates. For all estimates we will use historical risk factor market prices S_i, $i = 0, 1, \ldots K$ at times

$$t_n = t_0 + n\, dt \quad \text{with } n = 0, \ldots, N \ .$$

The estimates will be done at time

$$t = t_N > t_0$$

and the *window* used for the estimates ranges from t_0 until t_N. Regular sizes of such time windows range from ca. 30 day to ca. 2 years. We will denote by δt the time span over which the estimations will be made. This time span is also called *investment horizon*, *holding period* or *liquidation period*. The time span between to adjacent data in the time series will be denoted by dt. We will present a situation often occurring in practice, namely that the holding period does *not* have the same length as dt. To still have a clear presentation of the issues we will, however, assume that the holding period is a multiple of dt:

$$\delta t = m\, dt \ .$$

Using time series of *daily* settlement prices, for instance, one can calculate estimators for holding periods of m days (e.g., weekly or monthly estimators).

31.3.1 Return Estimates

The risk factor *returns* over the holding period δt, i.e., the logarithmic price changes will be denoted by r_i.

$$r_i(t) = \frac{1}{\delta t} \ln \left[\frac{S_i(t + \delta t)}{S_i(t)} \right] \quad \text{for } i = 0, \ldots, K \ . \tag{31.18}$$

The historical prices $S_i(t_n)$ in a time series with $N+1$ prices at times t_n with

$$t_n - t_{n-1} = dt \quad \text{for all } n = 1, \ldots, N$$

can be translated into *historical returns* for all past holding periods (all with length δt):

$$r_i(t_{n-m}) = \frac{1}{\delta t} \ln\left[\frac{S_i(t_n)}{S_i(t_{n-m})}\right] \quad \text{for } i = 0, \ldots, K \text{ and } n = m, \ldots, N \ . \tag{31.19}$$

The return at the *current* time $t = t_N$ over the *next* holding period δt (which still lies in the future) is only known for the risk free investment (this is the risk free rate). For all risky investments, however, the future prices $S_i(t_N + \delta t)$ are not yet known. Thus, the returns cannot be calculated but can only be estimated. We will use $\mu_i\ (t_N)$ to denote the estimator for $r_i(t_N)$.

The Moving Average (MA) Estimator

The common *moving average* estimator is simply the (equally weighted) mean of all historical returns over past time spans of length δt within the time window used for the estimation.[5]

$$\begin{aligned}\mu_i\ (t_N) &= \frac{1}{N-m+1} \sum_{n=m}^{N} r_i(t_{n-m}) \\ &= \frac{1}{(N-m+1)\,\delta t} \sum_{n=m}^{N} \ln\left[\frac{S_i(t_n)}{S_i(t_{n-m})}\right] \ .\end{aligned}$$

The logarithm appearing above can be written as a sum over historical dt-returns since everything except the first and the last term cancels in the

[5]Of course, this estimator is strongly autocorrelated since one time step later $N-m$ out of the $N-m+1$ values in the sum are still the same.

following sum:

$$\begin{aligned}\sum_{k=1}^{m} \ln\left[\frac{S_i(t_{n-k+1})}{S_i(t_{n-k})}\right] &= \ln\left[S_i(t_n)\right] - \ln\left[S_i(t_{n-1})\right] + \\ &\quad \ln\left[S_i(t_{n-1})\right] - \ln\left[S_i(t_{n-2})\right] + \\ &\quad \ldots + \\ &\quad \ln\left[S_i(t_{n-m+1})\right] - \ln\left[S_i(t_{n-m})\right] \\ &= \ln\left[S_i(t_n)\right] - \ln\left[S_i(t_{n-m})\right] = \ln\left[\frac{S_i(t_n)}{S_i(t_{n-m})}\right] .\end{aligned}$$

Inserting this into the expression for $\mu_i\ (t_N)$ and making the index substitution $x = n - k$ yields:

$$\begin{aligned}\mu_i\ (t_N) &= \frac{1}{(N-m+1)\,\delta t} \sum_{n=m}^{N} \sum_{k=1}^{m} \ln\left[\frac{S_i(t_{n-k+1})}{S_i(t_{n-k})}\right] \\ &= \frac{1}{(N-m+1)\,\delta t} \sum_{k=1}^{m} \sum_{x=m-k}^{N-k} \ln\left[\frac{S_i(t_{x+1})}{S_i(t_x)}\right] . \qquad (31.20)\end{aligned}$$

Now we have expressed $\mu_i\ (t_N)$ as a sum over dt-returns (in contrast to δt-returns). Again, almost everything cancels in the sum over x

$$\sum_{x=m-k}^{N-k} \ln\left[\frac{S_i(t_{x+1})}{S_i(t_x)}\right] = \ln\left[\frac{S_i(t_{N-k+1})}{S_i(t_{m-k})}\right] ,$$

leaving us with

$$\mu_i\ (t_N) = \frac{1}{(N-m+1)\ \ \delta t} \sum_{k=1}^{m} \ln\left[\frac{S_i(t_{N-k+1})}{S_i(t_{m-k})}\right] . \qquad (31.21)$$

The remaining sum contains only m terms now, instead of N. This clearly shows the fundamental problem with return estimates: only historical prices in the earliest and latest time period δt contribute. All prices in between are simply not used! This becomes especially severe for $dt = \delta t$, i.e., for $m = 1$. In this case only the very first and the very last price of the time window

contribute:

$$\mu_i\,(t_N) = \frac{1}{N\,\delta t}\ln\left[\frac{S_i(t_N)}{S_i(t_0)}\right]\,.$$

Two Alternatives for the Moving Average

One way to circumvent this problem is to use historical returns over time spans with different lengths $(N-n)\,dt$, all ending today. Based on Definition 31.18, we can construct an estimator using these historical returns in the following way:

$$\begin{aligned}\mu_i\,(t_N) &= \frac{1}{N}\sum_{n=0}^{N-1}\underbrace{\frac{1}{(N-n)\,dt}\ln\left[\frac{S_i(t_N)}{S_i(t_n)}\right]}_{\text{return from } t_n \text{ to } t_N}\\ &= \frac{1}{N\,dt}\sum_{n=0}^{N-1}\frac{1}{N-n}\ln\left[\frac{S_i(t_N)}{S_i(t_n)}\right]\,. \end{aligned}\tag{31.22}$$

Observe that in this estimator we have dt (and not δt) appearing in the denominator. All historical prices (even the ones way back in the past) enter with their influence still relevant today, namely with the return over the corresponding time span ending today. The historical returns over long time periods enter with the same weight as returns of shorter time periods. Thus, old prices are effectively under-weighted since the corresponding returns, although belonging to long time spans, have no more influence then returns over shorter time spans (resulting from more recent prices).

If older prices are to be as important as more recent ones, we can give the historical returns weights proportional to the length of the time spans they belong to: the longer the time span, the more weight the corresponding returns gets. An estimator with this feature is

$$\begin{aligned}\mu_i\,(t_N) &= \underbrace{\frac{1}{\sum_{k=0}^{N-1}(N-k)}}_{\text{numeraire}}\sum_{n=0}^{N-1}\underbrace{(N-n)}_{\text{time weight}}\cdot\underbrace{\frac{\ln\left[S_i(t_N)/S_i(t_n)\right]}{(N-n)\,dt}}_{\text{yield from } t_n \text{ to } t_N}\\ &= \frac{2}{N(N+1)dt}\sum_{n=0}^{N-1}\ln\left[\frac{S_i(t_N)}{S_i(t_n)}\right]\,, \end{aligned}\tag{31.23}$$

where we have used $\sum_{k=0}^{N-1}(N-k) = N(N+1)/2$, which can be shown easily.[6]

The Exponentially Weighted Moving Average (EWMA)

A very well known method is the *exponentially weighted moving average* estimator, or *EWMA* for short. In this estimator, the historical data are weighted less and less the further in the past they lie. This is accomplished by a damping factor λ in the following way:

$$\mu_i(t_N) = \frac{1}{M}\sum_{n=m}^{N}\lambda^{N-n} r_i(t_{n-m}) \text{ with } 0 < \lambda \leq 1 \text{ and } M := \sum_{k=m}^{N}\lambda^{N-k} \,. \tag{31.24}$$

With Eq. 31.19 for the historical returns this can be written as

$$\mu_i(t_N) = \frac{1}{M\,\delta t}\sum_{n=m}^{N}\lambda^{N-n}\ln\left[\frac{S_i(t_n)}{S_i(t_{n-m})}\right] \tag{31.25}$$

In contrast to the simple moving average, Eq. 31.21, the price logarithms of adjacent time periods dt do not cancels each other since they are differently weighted. Thus all prices influence the estimator.

By adjusting the parameter λ, the EWMA estimator can be made similar to the estimator in Eq. 31.22 as well as to the estimator in Eq. 31.23. For $\lambda = 1$ the EWMA estimator equals the simple moving average in Eq. 31.21.

---

[6] On one hand we get with the index transformation $i := N - k$

$$\sum_{k=0}^{N-1}(N-k) = \sum_{i=N}^{i=1} i = \sum_{i=1}^{N} i;\,.$$

On the other hand we have

$$\sum_{k=0}^{N-1}(N-k) = N^2 - \sum_{k=0}^{N-1} k = N^2 - \sum_{i=1}^{N}(i-1) = N^2 + N - \sum_{i=1}^{N} i\,.$$

Equating both results yields

$$\sum_{i=1}^{N} i = N^2 + N - \sum_{i=1}^{N} i \Longrightarrow \sum_{i=1}^{N} i = N(N+1)/2\,.$$

All the above return estimates are demonstrated in the Excel-Workbook ReturnEstimates.xls. Although neither the root mean square errors nor the Correlations with the ex post realized returns point out any clear favorite, one can see from the graphical presentation of the return time series that Eq. 31.23 looks like the best compromise between quite stable historical evolvement and still fast reaction to market movements; even when large time windows are used for the estimation.

31.3.2 Covariance Estimates

The entries $\delta\Sigma_{ij}$ in the covariance matrix, Eq. 21.22, i.e., the *risk factor* covariances can be determined via Eq. 21.28 using the *return* covariances

$$\delta\Sigma_{ij} \equiv \mathrm{cov}\left[\delta \ln S_i, \delta \ln S_j\right] = \delta t^2 \,\mathrm{cov}\left[r_i, r_j\right]$$

Similarly to the moving average estimators of the mean returns, the moving average estimators for the covariances of the returns over the holding period with length $\delta t = mdt$ at time t_N are

$$\mathrm{cov}\left[r_i, r_j\right](t_N) = \frac{1}{N-m}\sum_{n=m}^{n=N}\left[r_i(t_{n-m}) - \mu_i(t_N)\right]\left[r_j(t_{n-m}) - \mu_j(t_N)\right] ,$$

with r as in Eq. 31.19 and μ as in Eq. 31.21. Explicitly:

$$\begin{aligned}&\mathrm{cov}\left[r_i, r_j\right](t_N) \qquad (31.26)\\ &= \frac{1}{N-m}\sum_{n=m}^{n=N}\left(\frac{1}{\delta t}\ln\left[\frac{S_i(t_n)}{S_i(t_{n-m})}\right] - \mu_i(t_N)\right)\left(\frac{1}{\delta t}\ln\left[\frac{S_j(t_n)}{S_j(t_{n-m})}\right] - \mu_j(t_N)\right) .\end{aligned}$$

The EWMA estimator is analogously

$$\begin{aligned}&\mathrm{cov}\left[r_i, r_j\right](t_N) \qquad (31.27)\\ &= \frac{1}{M-1}\sum_{n=m}^{n=N}\lambda^{N-n}\left[\frac{1}{\delta t}\ln\left[\frac{S_i(t_n)}{S_i(t_{n-m})}\right] - \mu_i(t_N)\right]\left[\frac{1}{\delta t}\ln\left[\frac{S_j(t_n)}{S_j(t_{n-m})}\right] - \mu_j(t_N)\right]\end{aligned}$$

with μ from Eq. 31.25 and M from Eq. 31.24.

From these covariance estimates the *volatility* can be determined according to Eq. 21.30

$$\sigma_i\,(t_N) = \sqrt{\delta t\,\mathrm{cov}\,[r_i, r_i]\,(t_N)} \qquad \text{for} \quad i = 0, \ldots, K\ . \tag{31.28}$$

With the moving average estimate, Eq. 31.26, this yields the simply moving average volatility estimate. With Eq. 31.27, this yields the EWMA volatility estimate. The difference between these estimates and the ones presented in Sect. 33.4 is that in Sect. 33.4 we assume $\mu_i \equiv 0$.

The *correlations* resulting from the above covariances are according to their definition (see for example Eq. 31.5)

$$\rho_{i,j}\,(t_N) = \frac{\mathrm{cov}\left[r_i, r_j\right](t_N)}{\sqrt{\mathrm{cov}\,[r_i, r_i]\,(t_N)}\sqrt{\mathrm{cov}\left[r_j, r_j\right](t_N)}} \quad \text{for } i, j = 0, \ldots, K \tag{31.29}$$

Again, this yields the moving average or the EWMA estimate depending on which estimate is used for the covariance.

In the *Capital Asset Pricing Model* (*CAPM*) there is a ratio *Beta* , which relates the evolvement of risk factors $S_i(t)$ to the evolvement of a benchmark $S_0(t)$. As a rule, this benchmark represents a whole market and is usually an index. Beta is calculated from a regression of the risk factors time series with the benchmark time series, see Sect. 28.1. Estimates for Beta directly follow from the covariance estimates: the Beta of the i-th risk factor at time t_N is

$$\beta_i\,(t_N) = \frac{\mathrm{cov}\,[r_i, r_0]\,(t_N)}{\mathrm{cov}\,[r_0, r_0]\,(t_N)} = \rho_{i,0}\,(t_N)\,\frac{\sigma_i\,(t_N)}{\sigma_0\,(t_N)} \quad \text{for } i = 1, \ldots, K\ . \tag{31.30}$$

Again, this yields the moving average or the EWMA estimate for β, depending on which estimate is used for the covariance.