25

Backtesting: Checking the Applied Methods

A comparison of the value at risk figures delivered by a risk management system with the actual value changes of a portfolio allows an estimation of the qualitative and quantitative “goodness” of the risk model. Comparisons of realized values with previously calculated values are called *backtesting* procedures.

25.1 Profit and Loss Computations

There are several different *profit & loss* (or *P&L*) methods which can be used for comparison with the value at risk. The differences in these methods reflect the differences in the fundamental “philosophy” behind them.

- The *Dirty* Profit & Loss: The actual P&L of the portfolio, including all changes in position, fees paid and received, commission, etc. over the value at risk period are compiled and compared with the value at risk previously calculated. Position changes arise from continued trading during the value at risk period, the maturing of positions in the portfolio (for example, futures and options), the knock-in or knock-out of barrier options, coupon payments of bonds, etc. The effect of continued trading is not, in general, contained in the value at risk model. The dirty P&L is therefore only suitable for evaluating trading performance and not for the evaluation of model performance.

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_25

- The *Cleaned* Profit & Loss: The cleaned P&L is calculated in the same way as the dirty P&L but *without* taking position changes into account which result from continued trading during the value at risk period. Furthermore, the payment and receipt of fees and commissions are also omitted from the calculation. However, the cleaned P&L still contains the position changes resulting from the maturity of instruments occurring during the value at risk period (such as options and futures) or other position effects caused by the market (as opposed to the trader) such as the knock-out or knock-in of barrier options, coupon payments of bonds, etc. The cleaned P&L is therefore suitable to evaluate the model performance of risk models which take account of such maturity effects. The Monte Carlo simulation, for example, allows for such effects, the Variance-Covariance method, on the other hand, does not.
- The *Clean* Profit & Loss: Finally, the clean P&L is calculated in the same ways as the cleaned P&L but with reversing the effects of the maturing of positions during the value at risk period. In calculating the clean P&L, the value of the exact same portfolio as that existing upon initial calculation of the value at risk is re-calculated with the new market data observed at the conclusion of the value at risk period. Of course, a record of the initial portfolio positions at the time of the value at risk computation must have been kept. The clean P&L is thus suitable for evaluating the model performance of risk models such as the Variance-Covariance method which do not account for *aging* effects of the positions.

Independent of the chosen profit & loss method, the profit & loss per backtesting period is recorded for the evaluation of the goodness of the value at risk. Additionally, a record is kept for the calculated value at risk of the portfolio per backtesting-period. The data required for backtesting is thus not very large: neither historical time series nor a history of the portfolio positions must be maintained. Only two values per portfolio must be recorded in order to save the history of the portfolio, namely the value at risk and the associated P&L of the portfolio to be compared with the value at risk. In most cases, the P&L will be more favorable than the value at risk, and in others less favorable. Counting the number of times that the P&L is less favorable than the value at risk enables statistical conclusions about the goodness of the utilized model to be drawn. This is the fundamental idea behind backtesting. The superversing authorities require banks to perform such a backtesting procedure to validate the value at risk calculation (see Sect. 21.1). In the following, we present a standard method for implementing a backtesting procedure.

25.2 The Traffic Light Approach of the Supervising Authorities

25.2.1 Adjusting the Value at Risk (Yellow Zone)

The value at risk is a statistical statement. In general, some of the changes in the portfolio's values will be less favorable than the calculated value at risk. Such changes are referred to as *outliers* in the following discussion. At a confidence level c the probability of such an outlier is $1 - c$. The *expected* number of outliers in n backtesting periods with respect to this level of confidence is thus

$$\mathrm{E}[k] = n(1 - c) \; , \tag{25.1}$$

where k denotes the number of outliers observed in the n backtesting periods. The value k is not always equal to its expectation; it is a random variable. A deviation of the observed k from the expected number of outliers does not necessarily imply that a model is incorrect. Such an observation may be the result of pure chance. This is particularly true for small deviations from the expected value. In such cases, one speaks of the *yellow zone*, in which the supervising authorities will accept the model, but require that the value at risk is increased in the following manner.

The realization of k outliers actually observed in backtesting allows the definition of a new confidence level c' with respect to which the observed number is equal to the expectation:

$$k = n(1 - c') \quad \Rightarrow \quad c' = 1 - k/n \; . \tag{25.2}$$

From the observed outliers, it can be concluded that the value at risk from the model does not correspond to the confidence level c, but to a confidence level c'; or at least that the experimental basis for a confidence c' is greater than for c. The given VaR is then interpreted with respect to a confidence level of c'. A VaR which better corresponds to the *claimed* confidence c is then obtained from the given value of the VaR through multiplication by the ratio of the percentiles (confidence interval bounds) Q_{1-c} and $Q_{1-c'}$ in accordance with Eq. 21.20, respectively Eq. 22.15:

$$\mathrm{VaR}(c, t, T) \approx \frac{Q_{1-c}}{Q_{1-c'}} \; \mathrm{VaR}(\underbrace{1 - \frac{k}{n}}_{c'}, t, T) \; . \tag{25.3}$$

In order to use this equation, all assumptions and approximations of the delta-normal method must be made. These are:

- The risk factors are random walks, i.e., they are lognormally distributed implying that $Q_{1-c} = Q_{1-c}^{\mathrm{N}(0,1)}$.
- The drifts of the risk factors are neglected in the calculation.
- The exponential time evolutions of the risk factors are linearly approximated.
- The dependence of the portfolio value on the risk factors is linearly approximated. In particular, the portfolio value is also assumed to be lognormally distributed.

Since, for logarithmic changes, all of the variables under consideration are assumed to be normally distributed, the $Q_{1-c'}$ percentile can be calculated as

$$c' = 1 - \frac{k}{n} = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Q_{1-c'}} e^{-x^2/2} dx \ .$$

For example, if backtesting over 250 periods is performed resulting in the observation of 6 logarithmic portfolio value changes outside of the claimed confidence interval at a confidence level of 99%, the percentiles Q_{1-c} and $Q_{1-c'}$ are given by

$$0,99 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Q_{1-c}} e^{-x^2/2} dx \Rightarrow Q_{1-c} \approx 2,326$$

$$1 - \frac{6}{250} = 0,976 = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{Q_{1-c'}} e^{-x^2/2} dx \Rightarrow Q_{1-c'} \approx 1,972 \ .$$

The confidence level of the value at risk calculated by the model is now assumed to be not 99% as claimed but rather 97.6% as calculated on the basis of the actual events. The value at risk must thus be adjusted by a factor of $Q_{1-c}/Q_{1-c'} = 2.326/1.972 = 1.18$. This now larger value at risk is the value which can, based on actual observations, be relied upon with a confidence of 99%.

25.2.2 Criteria for Rejecting a Model (Red Zone)

Adjusting the value at risk as described above may not be applied for *arbitrary* values of k. It is allowed only if the difference between the observed value of k and its expectation with respect to the claimed confidence c can be reasonably explained by fluctuations due to the randomness involved. If the number of outliers is too far removed from the expected value, chance is no longer a plausible explanation and the reasons for the deviation lie in all probability on fundamental errors in the model itself. Here, the notion is used that the deviation of the measured results from the expected results are *significant*. The supervising authorities say that the model is in the *red zone*.

The field of statistics provides *hypothesis tests* which serve to check whether or not the observed deviation from a claimed value can be plausibly explained by random fluctuations. If not, such a deviation is considered significant and the tested *hypothesis* is in all probability not true. But again, absolute statements cannot be made on the basis of statistics. From such a hypothesis test we can only conclude *with a certain probability* that it was correct to reject (or accept) the hypothesis. The possibility remains that the hypothesis is rejected (accepted) although it is true (false). In statistics, these kind of errors are referred to as *type-I error* (rejection of a correct hypothesis) and *type-II error* (acceptance of a wrong hypothesis).

The hypothesis made when backtesting an internal model is that the observed value changes of a portfolio lie within the confidence interval specified by the calculated value at risk with a probability of c, in other words, that with a probability c the observed value changes are more favorable than the computed value at risk. To check this claim, we test whether the observed portfolio's value changes actually lie within the respective confidence interval. This is done for every VaR period over the entire backtesting time span.

The observed results can be categorized into two possible outcomes: the change in the portfolio's value is either more favorable than the respective VaR or not. This corresponds to the *Bernoulli experiment* described in Sect. A.4.2 in full detail. We could associate the event that an actual portfolio change is *more favorable* (or equally favorable) than the VaR with the outcome "tails" when tossing a coin and likewise the event that the portfolio change is *less favorable* than the VaR to the outcome "heads". The probability of observing k "heads" in n trials (less favorable than the VaR) is binomially distributed (see Eq. A.41) where the binomial probability p is the probability of the outcome "heads". In our case here p is then equal to the claimed probability that the portfolio change lies outside of the confidence interval, i.e., $p = 1 - c$. The number of outliers should therefore be binomially distributed with a density

$$B_{n,p}(k) = \binom{n}{k} p^k (1-p)^{n-k} \quad \text{with} \quad p = 1 - c = 1 - \mathrm{N}(Q_{1-c}) \ , \tag{25.4}$$

where $\mathrm{N}(Q_{1-c})$ denotes the probability that a standard normally distributed variable is $\leq 1 - c$, see Eq. A.53. Assuming that the model is correct, $B_{n,p}(k)$ is the probability that precisely k outliers are observed. The probability that *at most k* outliers are observed is then

$$\sum_{i=0}^{k} B_{n,p}(i) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-i} \ . \tag{25.5}$$

Again, assuming that the model is correct, the probability of observing *more than k* outliers is then

$$1 - \sum_{i=0}^{k} B_{n,p}(i) = \sum_{i=k+1}^{n} \binom{n}{i} p^i (1-p)^{n-i} \ . \tag{25.6}$$

This is equal to the probability that the model is *correct* assuming that k or more outliers are observed. Therefore, this is the probability of making a type-I error (the rejection of a correct model) when the hypothesis is rejected if k or more outliers are observed.

The determination of a Type-II error (acceptance of a *false* hypothesis) requires that the *true* probabilities for outcomes of the *false* model have to be known—a luxurious situation which almost never happens in practice. To be more specific: If a hypothesis test accepts a model for up to k outliers, then the probability that a *false* model is accepted equals the *true* probability for the event that this *false* model produces k or fewer outliers. Let's consider a simple example: Assume that a false model claims a is 99% confidence for its calculated VaR-numbers while the *true* confidence for these VaR-number is only 95%. If one accepts that model's VaR-numbers as a 99% VaR as long as only up to 9 out of 250 backtesting periods produce an outlier, one makes a Type-II error (see Eq. 25.5 with $n = 250$, $p = 5\%$) with a probability of

$$\sum_{i=0}^{9} B_{n,p}(i) = \sum_{i=0}^{9} \binom{250}{i} 0,05^i * (0,95)^{250-i} \approx 19,46\% \ .$$

It should be clear from these considerations that type-II errors only play a minor role in practice since it is very rarely the case that they can be determined in a sensible way.

The supervising authorities make their decision on establishing the limits for the red zone based on the probability of a type-I error (rejection of a correct model). The model is said to be in the red zone for a number of outliers k if the rejection of the model for *more than* k outliers has a probability for a type-I error of less than 0.01%. Using Eqs. 25.6 and 25.5, we find this to be the case when the probability (calculated with the model under consideration) of *at most* k outliers is $\geq$ 99.99% under the assumption that the model under consideration is correct. For $n = 250$ backtesting periods, the probability of at most 9 outliers equals 99.975% (see the Excel Workbook BinomialBacktest.xls from the download section [50]). The probability of at most 10 outliers is equal to 99.995% and is thus larger than 99.99%. The red zone established by the supervising authorities for 250 backtesting periods therefore begins at 10 outliers although the probability of making a type-I error (which is the probability for *more than 9* outliers) is 0.025%, i.e., greater than 0.01%. The probability of a type-I error when deciding to reject the model if more than $k = 10$ (in other words, 11 or more) outliers are observed is 0.005%, smaller than the required 0.01%. Nonetheless, 10 outliers within 250 backtesting periods is already deemed to belong to the red zone.[1]

25.2.3 The Green Zone

The boundary of the red zone is the upper boundary of the yellow zone. Analogously, a lower boundary of the yellow zone has been defined. No add-on is required if the observed number of outliers lies below this boundary. This zone is called the *green zone*. The model is said to be in the yellow zone for a number of outliers k if the rejection of the model for *more than* k outliers has a probability for a type-I error of less than 5.00%. Using Eqs. 25.6 and 25.5, we find this to be the case when the probability of *at most* k outliers is $\geq$ 95% under the assumption that the model is correct. For $n = 250$ backtesting periods for example, the probability of at most 4 outliers is equal to 89.22% (see the Excel workbook BinomialBacktest.xls from the download section [50]). The probability of at most 5 outliers on the other hand is 95.88%. Thus, the supervising authorities establish the boundary for the

[1] It may seem inconsistent that the supervising authorities establish k as the boundary for the red zone although this k corresponds to a type-I error of greater than 0.01%. This is the rule, however.

yellow zone as $k = 5$ for 250 backtesting periods, although the probability of a type-I error in this case is 10.78%, i.e., greater than 5%. For the rejection of the model with *more* than 5 outliers, the probability of a type-I error is 4.12%.

These three zones established by the supervising authorities, motivate the name *traffic light approach.*

25.2.4 Multiplication Factor and Add-On

As a rule, the value at risk calculated with the model must be multiplied by a factor of three even when it is found to be in the green zone for the simple reason that it has not been computed in accordance with the standard methods but by means of an internal model (this is just a rule of thumb, though). In the yellow zone, the VaR must *additionally* be multiplied by the ratio of the two percentiles Q_{1-c} and $Q_{1-c'}$ as prescribed in Eq. 25.3 where $Q_{1-c'}$ is the value established from Eq. 25.2. The multiplication factor for the yellow zone is thus $3Q_{1-c}/Q_{1-c'}$. The amount by which the multiplication factor exceeds the factor 3 is referred to as the *add-on*

$$\text{Add-on} = 3\frac{Q_{1-c}}{Q_{1-c'}} - 3 \approx \frac{6,978}{Q_{1-c'}} - 3 \ ,$$

where in the last step the confidence level $c = 99\%$ (required in SolvV (Germany) resp. CRR) for the standard normal distribution, $\mathrm{N}(Q_{1-c}) = 1\%$, and consequently $Q_{1-c} \approx 2.326$ was used.

The concepts described above are illustrated in detail in the Excel workbook BinomialBacktest.xls. All of the probabilities mentioned above, the boundaries between the different zones and the add-ons in the yellow zone are computed. The number of backtesting periods as well as the required confidences and the probability thresholds for the boundaries between the zones can be modified and the subsequent effects of these modifications immediately computed. In Fig. 25.1, these values are presented for $n = 250$ backtesting periods and a confidence of $c = 99\%$.

In Table 25.1, the add-ons are again explicitly displayed for the situation in Fig. 25.1 (rounded in increments of 0.05). This table can be found in the SolvV resp. the (CRR). For the red zone, an add-on of one is set and the model is later subjected to a new test.

Despite the multiplication factor, the value at risk computed with internal models is often lower than that calculated in accordance with the standard

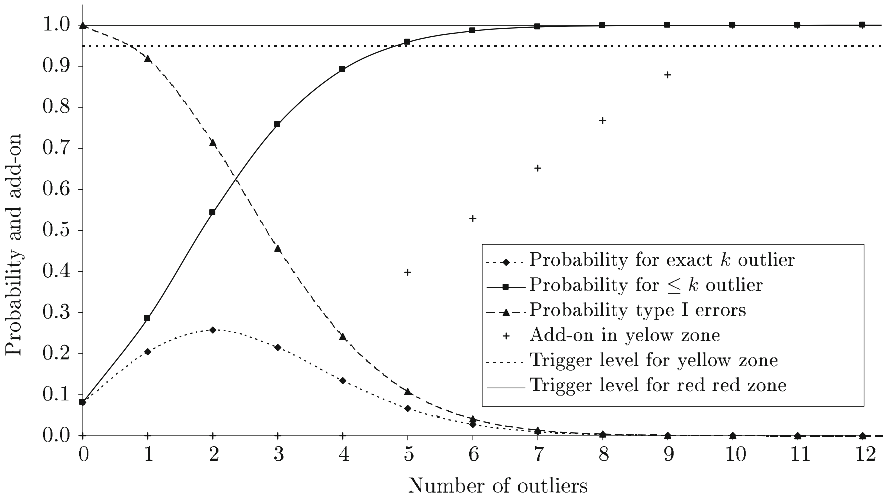

Fig. 25.1 Value at risk backtesting by means of a binomial test for 250 backtesting periods and 99% VaR confidence

Table 25.1 The table of add-ons for 250 backtesting periods as shown in the text of the German law

k	Add-on	Zone
4	0,00	Green
5	0.40	Yellow
6	0.50	Yellow
7	0.65	Yellow
8	0.75	Yellow
9	0.85	Yellow
≥ 10	1.00	Red

methods, since the correlation and compensation effects are more accurately taken into account. In many cases, the VaR of a portfolio computed according to an internal model is, despite the multiplication factor, is significantly lower than that computed with the standard methods.