Grade	AAA	AA	A	BBB	BB	B	CCC/C
Binomial	-	-	42.8871 (G)	0.004155 (R)	0.048299 (R)	1.0675 (Y)	2.76e-06 (R)
Normal	-	99.2014 (G)	53.196 (G)	2.04e-05 (R)	0.008213 (R)	0.833726 (R)	3.87e-07 (R)
One Factor	-	-	14.4642 (G)	1.74539 (Y)	6.5889 (G)	21.4833 (G)	2.04992 (Y)

Table VIII.39 Testing underestimation of default probabilities using Basel Traffic Light colors to the p-values of the tests

To obtain the default count from the observed default rates, we round the product of default rates and number of issuers. The asset correlation is $\rho = 7\%$. Running the script and examining the results, with the binomial test, we would classify three rating-specific PDs as underestimating the true default rate at a significance of 1%; and the number increases to four with the normal approximation.
Once we assume an asset correlation of 7%, however, the significance levels rise as we allow for the possibility that the year under scrutiny was a bad year in general. Now we can no longer reject a PD at a significance of 1%; we could, however reject two PDs at a significance of 5%. Let us note that the tests return error or null values if the realized default rate is zero. This makes sense because no evidence can be found for underestimating a default probability, if the realized default rate is at its minimum.
Clearly, decisions on significance levels are somewhat arbitrary. Therefore, in the Basel "Traffic lights" approach, two rather than one significance level must be chosen. If the p-value of a test is below α_{RED}, we assign an observation to the red zone (R), meaning that an underestimation of the default probability is very likely. If the p-value is above α_{RED}, but below α_{YELLOW}, we interpret the result as a warning that a PD might be underestimated, i.e. a yellow zone (Y). Otherwise, we assign it to the green zone (G).

Portfolio credit risk models produce a probability distribution for portfolio credit losses. To validate the quality of a given model, we can examine whether observed losses are consistent with the model's predictions. Certain analysts argue that portfolio models are difficult to validate empirically. Usually, such an opinion is justified by a comparison to market risk models. Market risk models produce loss forecasts for a portfolio (for instance the trading book of a bank) as well, but the underlying horizon is much shorter - often, it is restricted to a single day.
A standard validation procedure consists in checking the frequency with which actual losses exceeded the Value at Risk (VaR). In a market risk setting, risk managers usually examine 99% VaR. Over one year containing roughly 250 trading days, the expected number of exceedances of the 99% VaR is $250 \times (1 - 0.99) = 2.5$, provided that the VaR forecasts are correct.
When we observe the number of exceedances differing significantly from the expected number, we can conclude that the predictions were incorrect. Significance can be assessed with a traditional binomial test.
Obviously, such a test is not very useful for the validation of credit portfolio models, which mostly have a one-year horizon. There is a way out though: ifwe do not confine a test to the prediction of the extreme events but rather test the overall fit of the predicted loss distribution, we can make better use of information and possibly

learn valuable information about a model's validity with only 5 or 10 years of data.
We then introduce the Berkowitz test, which is a powerful method that has been used both for credit risk and for market risk. For each period (usually a length of one year), the information needed for applying the Berkowitz test is: a loss figure and a forecast of the loss distribution made at the start of the period. As a result, for a given loss L, we can compute the probability $F(L)$ with which this loss is not exceeded.
Obviously, the distribution can differ from year to year because of changes in portfolio composition or changes in the risk parameters of the portfolio constituents. The basic idea behind the Berkowitz (2001) test is to evaluate the entire distribution. The test involves a double transformation of observed losses, with the two transformations as follows:

- **1st transformation**: replace L_t, the loss in t, with the predicted probability of observing this loss or a smaller one. We obtain this probability by inserting the loss L_t into a cumulative distribution function $F(L_t)$: $p_t = F(L_t)$.

- **2nd transformation**: transform p_t by applying $\Phi^{-1}(x)$, the inverse cumulative standard normal distribution function. Formally: $z_t = \Phi^{-1}(p_t)$.

The 1st transformation produces numbers between 0 and 1. If the predicted distribution is correct, we have even more information: the numbers should be uniformly distributed between 0 and 1. To check this, we start by looking at the median of the distribution. If the model is correct, 50% of observed losses would be expected to end up below the median loss, which has F(median loss) = 0.5.
Thus, the transformed variable pt should be below 0.5 in 50% of all cases. We can continue in this way.
The 25th percentile, which has F(25% percentile) = 0.25, splits the first half into another pair of two halves, and again observations will be evenly spread on expectation. Similarly, we can conclude that there should be as many p_ts below 0.25 as there are p_ts above 0.75.
We can use finer partitions and still conclude that the p_t should be evenly spread across the intervals. Theoretically, we could stop after the 1st transformation and test whether the pts are actually uniformly distributed between 0 and 1. However, tests based on normally distributed numbers are more widespread in the scientific community and they are more powerful. This is the reason why we perform another transformation. If the model summarized by $F(L)$ is correct, transformed losses z_t will be normally distributed with zero mean and unit variance. The intuition behind this is similar to the 1st transformation. If p_t is uniform between 0 and 1, 2.5% of all observations will be below 2.5%, for example. Consequently, 2.5% of all z_t will be below -1.96, i.e. $\Phi^{-1}(0.025)$ and this is what we expect for a standard normal variable.
Berkowitz suggested the restriction of the test to the hypothesis that z_t have zero mean and unit variance. We could additionally test whether they are normally distributed, but tests of normality tend not to be very powerful if the number of observations is small, so we do not lose much information if we do not test for this property on z_t as well.
A convenient and powerful way of testing the joint hypothesis of zero mean and unit variance is a **likelihood ratio test**. The likelihood is the probability that we observe given data with a given model. With a likelihood ratio test, we test whether imposing a restriction (i.e. z_t have zero mean and unit variance) leads to a significant

loss in the likelihood. The test statistics is based on the log-likelihood function of the transformed series z_t. Since the z_t are normally distributed under the hypothesis that the model is correct, the likelihood is obtained through the normal density:

$$\text{Likelihood} = \prod_{t=1}^{T} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(z_t-\mu)^2}{2\sigma^2}\right] \textit{ (Eq. VIII.155)}$$

That is, if we have T observations, we multiply the probabilities of having individual observations z_t to reach the likelihood to have the set of T observations. This is correct if unexpected losses, which are captured here by $z_t - \mu$, are independent across time. Although this assumption may be violated in some situations, it should be fulfilled if the loss forecasts make efficient use of information. It should be noted that this is not the same as assuming that losses themselves are independent across time.
Also, there is no need to abandon the concept of credit cycles, as long as the notion of credit cycles relates to losses, not unexpected losses. It is more convenient to work with $\ln L$, the logarithm of the likelihood:

$$\ln L = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln(\sigma^2) - \sum_{t=1}^{T} \frac{(z_t-\mu)^2}{2\sigma^2} \textit{ (Eq. VIII.156)}$$

To evaluate the log-likelihood, we calculate the maximum likelihood (ML) estimators for the mean and variance of the transformed variable z_t: $\hat{\mu}_{ML} = \frac{1}{T}\sum_{t=1}^{T} z_t$, $\hat{\sigma}_{ML}^2 = \frac{1}{T}\sum_{t=1}^{T}(z_t - \hat{\mu}_{ML})^2$.
The likelihood ratio test is then structured to test the joint hypothesis that the z_t have zero mean and unit variance. It is given by: $\lambda = 2[\ln L\,(\mu = \hat{\mu}_{ML}, \sigma^2 = \hat{\sigma}_{ML}^2) - \ln L(\mu = 0, \sigma^2 = 1)]$.
If imposing the hypothesis $\mu = 0$ and $\sigma^2 = 1$ leads to large loss in likelihood, λ will be large. Therefore, the larger λ, the more evidence we have that the z_t do not have mean zero and unit variance. Under usual regularity conditions, the test statistics λ will be asymptotically distributed as a χ^2 variable with two degrees of freedom.
The CreditPortfolioValidation script allows the user to perform the Berkowitz test.
We assume in this example to have five years of loss data and for focusing on the core of the procedure we also assume that the predicted loss distribution was the same for every year and the specification of the loss distribution is such that we can immediately determine the exact probability of each loss.
The illustrative data are reported directly in the code. Running the script, we reach a p-value for the example equal to 0.329%, as a result we have to reject the hypothesis that the model which produced the outcome is correct.

FURTHER READINGS

Christodoulakis G., Satchel S. – "The validity of credit risk model validation methods" – The Analytics of Risk Model Validation in Quantitative Finance, Elsevier (2008).
De Laurentis G., Maino R., Molteni L. – "Developing, Validating and using Internal Ratings: methodologies & case studies"– Wiley (2010).
Löffler G., Posch P. N. – "Credit risk modeling" – Wiley (2011).
Sobehart J. R., Keenan S. C., Stein R. M. – "Benchmarking Quantitative Default Risk Models: a Validation