Testing rating accuracy

As Basel II approaches the implementation stage, regulators have identified internal ratings validation as a key challenge for banks using this approach. Here, Bernd Engelmann, Evelyn Hayden and Dirk Tasche build upon previous research showing how to use the so-called receiver operator characteristic method in ratings validation, testing their results on a real database of small and medium-sized enterprise loans

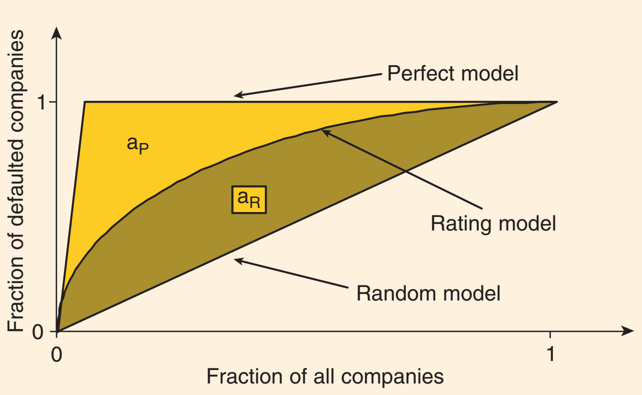
Various rating methodologies and credit risk modelling approaches have been developed in the past few years. Furthermore, in the second consultative document of its new capital adequacy framework, the Basel Committee on Banking Supervision (2001) has announced that an internal ratings-based approach could soon form the basis for setting capital charges with respect to credit risk. This is forcing banks and supervisors to develop statistical tools to evaluate the quality of internal rating models. The importance of sound validation techniques for rating systems stems from the fact that rating models of poor quality could lead to sub-optimal capital allocation. Therefore, the Basel Committee (2000) has emphasised that the field of model validation is one of the major challenges for financial institutions and supervisors.

The most popular validation technique currently used in practice is the cumulative accuracy profile (CAP) and its summary statistic, the accuracy ratio. A detailed explanation of this method can be found in Sobehart, Keenan & Stein (2000). The receiver operating characteristic (ROC) is a similar concept to the CAP. This method has its origin in signal detection theory, psychology and, especially, in medicine (see, for example, Hanley & McNeil, 1982).[1] Sobehart & Keenan (2001) explain how to use this concept for validating internal rating models. In their article, they concentrate on the qualitative features of ROC curves such as calculation and interpretation. The main conclusion of their article is that the size of the area below an ROC curve is an indicator of the quality of a rating model.

A single number, such as the accuracy ratio or the area below the ROC curve, contains little information from a statistical point of view. To get a feeling for the quality of a rating system, it is desirable to state confidence intervals. It is also insufficient to compare two rating systems that are calibrated on the same data set by just comparing two numbers only. A rigorous statistical test is the only way to obtain a sound decision about the superiority of one rating model over the other.

This article consists of four parts. In the first, to keep the article self-contained, we briefly review the concepts of the CAP and the ROC. For both concepts it is possible to summarise the information concerning the quality of a rating system with a single number, namely with the accuracy ratio and the area below the ROC curve. We will show that the accuracy ratio is just a linear transformation of the area below the ROC curve. In the second part, we discuss a simple method to calculate confidence intervals for the area under the ROC curve. Because of the relation between the accuracy ratio and the area below the ROC curve, this method is also applicable to the accuracy ratio. Compared with bootstrap analysis, the method is much faster. In the third part, we discuss a test to compare the area below the ROC curve of two rating models that are validated on the same data set. In the final part, we apply these concepts to real rating models.

1. Cumulative accuracy profiles



Both this test and the method to calculate confidence intervals rely on asymptotic normality. The reliability of these methods is not guaranteed in the case of a validation sample containing only a small number of defaults. From our experience, we find that as a rule of thumb around 50 defaults in the validation sample are enough for the asymptotic properties to hold. If the validation sample contains fewer than 50 defaults, the results should be interpreted with care or the use of alternative methods should be considered.

Throughout this article, we will assume rating systems that produce continuous rating scores. This is mainly because our examples are based on logit models and discriminant analyses that fulfil this assumption. However, all the results apply also to rating scores with a finite number of grades. We will not capture this case in great detail but we will give comments on modifications of the results for the continuous case where necessary.

Cumulative accuracy profiles

Consider an arbitrary rating model that produces a continuous rating score. The score under consideration could be a rating score such as Altman's Z-score (1968) or a score obtained from a logit model or from any other approach. A high rating score is usually an indicator of a low default probability. To obtain the CAP curve, all debtors are first ordered by their respective scores from riskiest to safest, that is, from the debtor with the lowest score to the debtor with the highest score. For a given fraction x of the total number of debtors, the CAP curve is constructed by calculating the percentage $d(x)$ of the defaulters whose rating scores are equal to or lower than the maximum score of fraction x. This is done for x ranging from 0% to 100%. Figure 1 illustrates CAP curves.

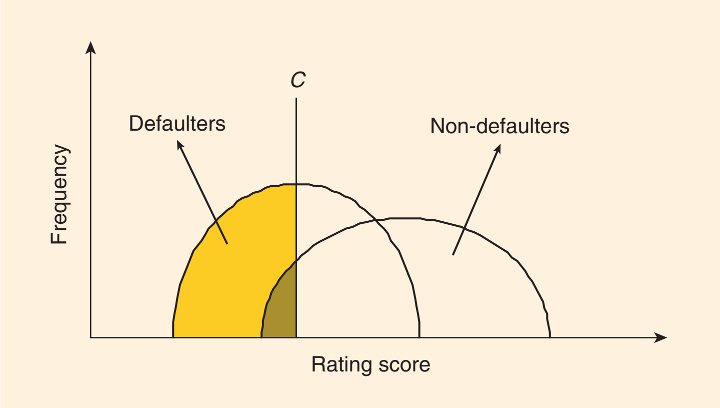
A perfect rating model will assign the lowest scores to the defaulters. In this case, the CAP is increasing linearly and then staying at one. For a random model without any discriminative power, the fraction x of all debtors with the lowest rating scores will contain x% of all defaulters. Real rating systems will be somewhere in between these two extremes. The quality of a rating system is measured by the accuracy ratio AR. It is defined as the ratio of the area a_R between the CAP of the rating model being validated and the CAP of the random model, and the area a_P between the CAP of the perfect rating model and the CAP of the random model, that is:

$$AR = \frac{a_R}{a_P}$$

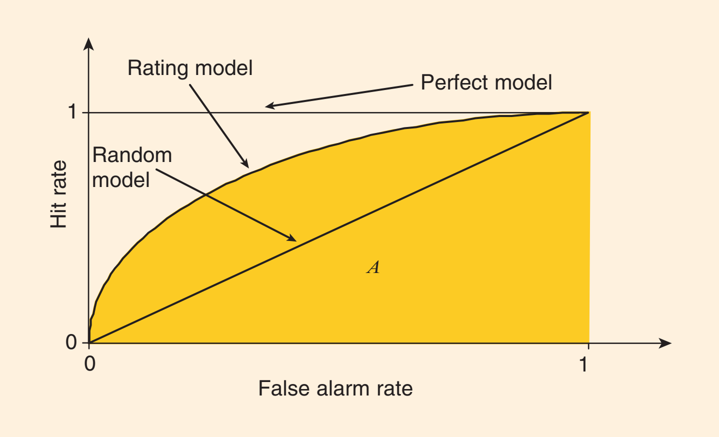
Thus, the rating method is the better the closer AR is to one.

[1] *An interesting overview of the variety of possible applications of ROC curves is given in Swets (1988)*

2. Distribution of rating scores for defaulting and non-defaulting debtors



3. Receiver operating characteristic curves



Receiver operating characteristic

We also briefly explain the concept of an ROC curve. The construction of an ROC curve is illustrated in figure 2, showing possible distributions of rating scores for defaulting and non-defaulting debtors. For a perfect rating model, the left distribution and the right distribution in figure 2 would be separate. For real rating systems, perfect discrimination in general is not possible. Both distributions will overlap, as illustrated in figure 2.

Assume someone has to find out from the rating scores which debtors will survive during the next period and which debtors will default. One possibility for the decision-maker would be to introduce a cutoff value C as in figure 2, and to classify each debtor with a rating score lower than C as a potential defaulter and each debtor with a rating score higher than C as a non-defaulter. Then four decision results would be possible. If the rating score is below the cutoff value C and the debtor defaults subsequently, the decision was correct. Otherwise the decision-maker wrongly classified a non-defaulter as a defaulter. If the rating score is above the cutoff value and the debtor does not default, the classification was correct. Otherwise, a defaulter was incorrectly assigned to the non-defaulters group.

Using the notation of Sobehart & Keenan (2001), we define the hit rate $HR(C)$ as:

$$HR(C) = \frac{H(C)}{N_D}$$

where $H(C)$ (equal to the light area in figure 2) is the number of defaulters predicted correctly with the cutoff value C, and N_D is the total number of defaulters in the sample. The false alarm rate $FAR(C)$ (equal to the dark area in figure 2) is defined as:

$$FAR(C) = \frac{F(C)}{N_{ND}}$$

where $F(C)$ is the number of false alarms, that is, the number of non-defaulters that were classified incorrectly as defaulters by using the cutoff value C. The total number of non-defaulters in the sample is denoted by N_{ND}. The ROC curve is constructed as follows. For all cutoff values C that are contained in the range of the rating scores the quantities $HR(C)$ and $FAR(C)$ are calculated. The ROC curve is a plot of $HR(C)$ versus $FAR(C)$. This is shown in figure 3.

A rating model's performance is better the steeper the ROC curve is at the left end and the closer the ROC curve's position is to the point (0, 1). Similarly, the larger the area below the ROC curve, the better the model. We denote this area by A. It can be calculated as:

$$A = \int_0^1 HR(FAR)d(FAR)$$

The area A is 0.5 for a random model without discriminative power and it is 1.0 for a perfect model. It is between 0.5 and 1.0 for any reasonable rating model in practice.

Connection between ROC curves and CAP curves

We prove a relation between the accuracy ratio and the area under the ROC curve (A) in order to demonstrate that both measures are equivalent. By a simple calculation, we get for the area a_P between the CAP of the perfect rating model and the CAP of the random model:

$$a_P = \frac{0.5N_{ND}}{N_D + N_{ND}}$$

We introduce some additional notation. If we randomly draw a debtor from the total sample of debtors, the resulting score is described by a random variable S_T. If the debtor is drawn randomly from the sample of defaulters only, the corresponding random variable is denoted by S_D, and if the debtor is drawn from the sample of non-defaulters only, the random variable is denoted by S_{ND}. Note that $HR(C) = P(S_D < \mathrm{C})$ and $FAR(C) = P(S_{ND} < C)$.

To calculate the area a_R between the CAP of the rating model being validated and the CAP of the random model, we need the cumulative distribution function $P(S_T < C)$, where S_T is the distribution of the rating scores in the total population of all debtors. In terms of S_D and S_{ND}, the cumulative distribution function $P(S_T < C)$ can be expressed as:

$$P(S_T < C) = \frac{N_D P(S_D < C) + N_{ND} P(S_{ND} < C)}{N_D + N_{ND}}$$

Since we assumed that the distributions of S_D and S_{ND} are continuous, we have $P(S_D = C) = P(S_{ND} = C) = 0$ for all attainable scores C.

Using this, we find for the area a_R:

$$\begin{aligned} a_R &= \int_0^1 P(S_D < C)dP(S_T < C) - 0.5 \\ &= \frac{N_D \int_0^1 P(S_D < C)dP(S_D < C) + N_{ND} \int_0^1 P(S_D < C)dP(S_{ND} < C)}{N_D + N_{ND}} - 0.5 \\ &= \frac{0.5N_D + N_{ND}A}{N_D + N_{ND}} - 0.5 = \frac{N_{ND}(A - 0.5)}{N_D + N_{ND}} \end{aligned}$$

With these expressions for a_P and a_R, the accuracy ratio can be calculated as:

$$AR = \frac{a_R}{a_P} = \frac{N_{ND}(A - 0.5)}{0.5N_{ND}} = 2(A - 0.5)$$

This means that the accuracy ratio can be calculated directly from the area below the ROC curve and vice versa.[2] Hence, both summary statistics contain the same information.

Calculation of confidence intervals for A

Here, we discuss a simple method of calculating confidence intervals for A, the area below the ROC curve. The same reasoning applies to the accuracy ratio by means of the relation proven above. The results presented are based on Bamber (1975). Derivations and proofs of the results we use in this article, as well as a more complete discussion of the limitations of these approaches and their assumptions, are given there. We start with a probabilistic interpretation of A.

Consider the following experiment. Two debtors are drawn at random, the first from the distribution of defaulters, the second from the distribution of non-defaulters. The scores of the defaulter and the non-defaulter determined in this way can be interpreted as realisations of the two independent continuous random variables S_D and S_{ND}. Assume someone has to decide which of the debtors is the defaulter. A rational decision-maker might suppose that the defaulter is the debtor with the lower rating score. The probability that he is right is equal to $P(S_D < S_{ND})$. A simple calculation shows that this probability is exactly equal to the area below the ROC curve A.

This interpretation relates to the statistic of the U-test of Mann-Whitney (1947).[3] If we draw a defaulter with score s_D from S_D and a non-defaulter with score s_{ND} from S_{ND} and define $u_{D,ND}$ as:

$$u_{D,ND} = \begin{cases} 1, if & s_D < s_{ND} \\ 0, if & s_D \geq s_{ND} \end{cases}$$

then the test statistic $\hat{U}$ of Mann-Whitney is defined as:

$$\hat{U} = \frac{1}{N_D N_{ND}} \sum_{(D,ND)} u_{D,ND}$$

where the sum is over all pairs of defaulters and non-defaulters (D, ND) in the sample.

Observe that $\hat{U}$ is an unbiased estimator for $P(S_D < S_{ND})$, that is:

$$A = E\left(\hat{U}\right) = P\left(S_D < S_{ND}\right)$$

Furthermore, we find that the area $\hat{A}$ below the ROC curve calculated from the empirical data is equal to $\hat{U}$. For the variance $\sigma^2_{\hat{U}}$ of $\hat{U}$ we find the unbiased estimator $\hat{\sigma}^2_{\hat{U}}$ as[4]:

$$\hat{\sigma}^2_{\hat{U}} = \frac{1}{4\left(N_D - 1\right)\left(N_{ND} - 1\right)} \times$$

$$\left[1 + \left(N_D - 1\right)\hat{P}_{D,D,ND} + \left(N_{ND} - 1\right)\hat{P}_{ND,ND,D} - 4\left(N_D + N_{ND} - 1\right)\left(\hat{U} - \frac{1}{2}\right)^2\right]$$

where $\hat{P}_{D,D,ND}$ and $\hat{P}_{ND,ND,D}$ are estimators for the expressions $P_{D,D,ND}$ and $P_{ND,ND,D}$ which are defined as:

$$\begin{aligned} P_{D,D,ND} &= P\left(S_{D,1}, S_{D,2} < S_{ND}\right) + P\left(S_{ND} < S_{D,1}, S_{D,2}\right) \\ &\quad - P\left(S_{D,1} < S_{ND} < S_{D,2}\right) - P\left(S_{D,2} < S_{ND} < S_{D,1}\right), \\ P_{ND,ND,D} &= P\left(S_{ND,1}, S_{ND,2} < S_D\right) + P\left(S_D < S_{ND,1}, S_{ND,2}\right) \\ &\quad - P\left(S_{ND,1} < S_D < S_{ND,2}\right) - P\left(S_{ND,2} < S_D < S_{ND,1}\right) \end{aligned}$$

The quantities $S_{D,1}$, $S_{D,2}$ are independent observations randomly sampled from S_D, and $S_{ND,1}$, $S_{ND,2}$ are independent observations randomly sampled from S_{ND}. This unbiased estimator $\hat{\sigma}^2_{\hat{U}}$ is implemented in standard statistical software packages.[5]

For $N_D, N_{ND} \to \infty$ it is known that $(A - \hat{U})/\hat{\sigma}_{\hat{U}}$ is asymptotically normally distributed with mean zero and standard deviation one. So confidence intervals at confidence level α can be calculated for $\hat{U}$ using the relation:

$$P\left(\hat{U} - \hat{\sigma}_{\hat{U}}\Phi^{-1}\left(\frac{1+\alpha}{2}\right) \leq A \leq \hat{U} + \hat{\sigma}_{\hat{U}}\Phi^{-1}\left(\frac{1+\alpha}{2}\right)\right) \approx \alpha$$

where Φ denotes the cumulative distribution function of the standard normal distribution. Our analysis below indicates that the number of defaults should be at least around 50 in order to guarantee that the above formula is a good approximation. There is no clear rule for which values of $\hat{U}$ the asymptotic normality of $\hat{U}$ is a valid approximation, because $\hat{U}$ can solely take values in the interval $[0, 1]$. If $\hat{U}$ is only a few (two, three or four) standard deviations away from one, it is clear that the normal approximation is problematic.[6] However, as illustrated in our examples below, even in this situation the normal approximation can lead to reasonable results. Nevertheless, one should keep in mind this potential problem and interpret the results of this method with care in these situations.

Comparing the areas below the ROC curves for two rating models

This part of the article is based on the work of DeLong, DeLong & Clarke-Pearson (1988). The aim of their work is to provide a test for the difference between the areas A^1 and A^2 below the ROC curves of two different rating models, 1 and 2. From the last section we know how to calculate the variance $\sigma^2_{\hat{U}^i}$ for an estimator $\hat{U}^i$ of A^i. For the covariance $\hat{\sigma}^2_{\hat{U}^1\hat{U}^2}$ between the estimators $\hat{U}^1$ and $\hat{U}^2$ of A^1 and A^2 we find[7]:

$$\hat{\sigma}^2_{\hat{U}^1,\hat{U}^2} = \frac{1}{4\left(N_D - 1\right)\left(N_{ND} - 1\right)} \times$$

$$\left(\tilde{P}^{12}_{D,D,ND,ND} + \left(N_D - 1\right)\tilde{P}^{12}_{D,D,ND} + \left(N_{ND} - 1\right)\tilde{P}^{12}_{ND,ND,D} - 4\left(N_D + N_{ND} - 1\right)\left(\tilde{U}^1 - \frac{1}{2}\right)\left(\tilde{U}^2 - \frac{1}{2}\right)\right)$$

where $\tilde{P}^{12}_{D,D,ND,ND}$, $\tilde{P}^{12}_{D,D,ND}$ and $\tilde{P}^{12}_{ND,ND,D}$ are estimators for $P^{12}_{D,D,ND,ND}$, $P^{12}_{D,D,ND}$ and $P^{12}_{ND,ND,D}$, which are defined as[8]:

$$\begin{aligned} P^{12}_{D,D,ND,ND} &= P\left(S^1_D > S^1_{ND}, S^2_D > S^2_{ND}\right) + P\left(S^1_D < S^1_{ND}, S^2_D < S^2_{ND}\right) \\ &\quad - P\left(S^1_D > S^1_{ND}, S^2_D < S^2_{ND}\right) - P\left(S^1_D < S^1_{ND}, S^2_D > S^2_{ND}\right), \\ P^{12}_{D,D,ND} &= P\left(S^1_{D,1} > S^1_{ND}, S^2_{D,2} > S^2_{ND}\right) + P\left(S^1_{D,1} < S^1_{ND}, S^2_{D,2} < S^2_{ND}\right) \\ &\quad - P\left(S^1_{D,1} > S^1_{ND}, S^2_{D,2} < S^2_{ND}\right) - P\left(S^1_{D,1} < S^1_{ND}, S^2_{D,2} > S^2_{ND}\right), \\ P^{12}_{ND,ND,D} &= P\left(S^1_D > S^1_{ND,1}, S^2_D > S^2_{ND,2}\right) + P\left(S^1_D < S^1_{ND,1}, S^2_D < S^2_{ND,2}\right) \\ &\quad - P\left(S^1_D > S^1_{ND,1}, S^2_D < S^2_{ND,2}\right) - P\left(S^1_D < S^1_{ND,1}, S^2_D > S^2_{ND,2}\right) \end{aligned}$$

The quantities S^i_D, $S^i_{D,1}$ and $S^i_{D,2}$ are independent draws from the sample of defaulters. The upper index i indicates whether the score of the rating model 1 or the score of the rating model 2 has to be taken. The meaning of S^i_{ND}, $S^i_{ND,1}$ and $S^i_{ND,2}$ is analogous.

To carry out the test for the difference between the two rating methods (where the null hypothesis is equality of both areas below the ROC curve), we have to evaluate the test statistic T, which is defined as:

$$T = \frac{\left(\hat{U}^1 - \hat{U}^2\right)^2}{\sigma^2_{\hat{U}^1} + \sigma^2_{\hat{U}^2} - 2\sigma^2_{\hat{U}^1,\hat{U}^2}}$$

[2] *This relation is valid for rating systems with a finite number of grades, too*

[3] *This U-test of Mann-Whitney can be used to assess if a rating system has discriminative power at all by testing the null hypothesis $P(S_D < S_{ND}) = 0.5$. It can also be applied to calculate confidence intervals, which is the application we discuss here*

[4] *In Bamber (1975), several upper bounds for the variance are given. These upper bounds are easy to evaluate and can be used to derive conservative estimates of confidence intervals. However, usually they are not helpful in determining the true confidence intervals because they rely on specific distributional assumptions that are, in general, not fulfilled by the data*

[5] *For rating systems with a finite number of grades only, $\hat{u}_{D,ND}$ has to be defined as 0.5 for $s_D = s_{ND}$. In the formula for the unbiased estimator of $\hat{\sigma}^2_{\hat{U}}$, the first term in the brackets ('1') has to be replaced by $P(S_D \neq S_{ND})$, which is equal to one in the continuous case*

[6] *Several methods for the computation of confidence intervals without relying on the assumption of asymptotic normality are known, which lead in general to very conservative confidence intervals. An overview of these methods is given in Bamber (1975). One could rely on these methods if the normal approximation is questionable as in the case of very few defaults in the validation sample*

[7] *This expression is also correct for rating systems with a finite number of rating categories*

[8] *The expressions given in DeLong, DeLong & Clarke-Pearson (1988) look very different from this expression. However, it can be shown that both are equivalent. We used this expression to be consistent with the notation of the previous section*

A. Results for $\hat{A}$, $\hat{\sigma}_{\hat{A}}$, 95% and 99% confidence intervals for the total portfolio

	$\hat{A}$	$\hat{\sigma}_{\hat{A}}$	95% confidence	95% confidence	99% confidence	99% confidence
			interval (normal)	interval (bootstrap)	interval (normal)	interval (bootstrap)
Z-score	0.72	0.007	[0.7061, 0.7336]	[0.7059, 0.7336]	[0.7018, 0.7378]	[0.7014, 0.7375]
Logit score	0.84	0.006	[0.8278, 0.8526]	[0.8294, 0.8507]	[0.8248, 0.8558]	[0.8258, 0.8542]

B. Results for $\hat{A}$, $\hat{\sigma}_{\hat{A}}$, 95% and 99% confidence intervals for sub-portfolio 1 (50 defaults)

	$\hat{A}$	$\hat{\sigma}_{\hat{A}}$	95% confidence	95% confidence	99% confidence	99% confidence
			interval (normal)	interval (bootstrap)	interval (normal)	interval (bootstrap)
Z-score	0.704	0.036	[0.6348, 0.7741]	[0.6332, 0.7707]	[0.6131, 0.7959]	[0.6083, 0.7892]
Logit score	0.777	0.037	[0.7046, 0.8485]	[0.7032, 0.8445]	[0.6820, 0.8711]	[0.6768, 0.8638]

C. Results for $\hat{A}$, $\hat{\sigma}_{\hat{A}}$, 95% and 99% confidence intervals for sub-portfolio 2 (20 defaults)

	$\hat{A}$	$\hat{\sigma}_{\hat{A}}$	95% confidence	95% confidence	99% confidence	99% confidence
			interval (normal)	interval (bootstrap)	interval (normal)	interval (bootstrap)
Z-score	0.696	0.048	[0.6018, 0.7899]	[0.6021, 0.7868]	[0.5729, 0.8187]	[0.5769, 0.8131]
Logit score	0.801	0.050	[0.7031, 0.8980]	[0.6953, 0.8861]	[0.6733, 0.9277]	[0.6578, 0.9050]

D. Results for $\hat{A}$, $\hat{\sigma}_{\hat{A}}$, 95% and 99% confidence intervals for sub-portfolio 1 (10 defaults)

	$\hat{A}$	$\hat{\sigma}_{\hat{A}}$	95% confidence	95% confidence	99% confidence	99% confidence
			interval (normal)	interval (bootstrap)	interval (normal)	interval (bootstrap)
Z-score	0.697	0.098	[0.5041, 0.8894]	[0.5004, 0.8651]	[0.4436, 0.9499]	[0.4367, 0.9004]
Logit score	0.855	0.063	[0.7317, 0.9777]	[0.7220, 0.9520]	[0.6931, 1.0000]	[0.6716, 0.9680]

E. Results for $\hat{A}$, $\hat{\sigma}_{\hat{A}}$, 95% and 99% confidence intervals for a rating system with seven categories (50 defaults)

$\hat{A}$	$\hat{\sigma}_{\hat{A}}$	95% confidence	95% confidence	99% confidence	99% confidence
		interval (normal)	interval (bootstrap)	interval (normal)	interval (bootstrap)
0.8116	0.0251	[0.7625, 0.8608]	[0.7603, 0.8576]	[0.7471, 0.8762]	[0.7434, 0.8702]

This test statistic is asymptotically $\chi^2(1)$-distributed with one degree of freedom.

Application to real rating systems

Here, we apply the concepts introduced in the previous two sections using a Bundesbank database[9] containing about 325,000 balance sheets for the years 1987–1999. The database includes about 3,000 defaults where default was defined as legal insolvency. To produce rating scores, we applied Altman's Z-score and the score of a logit model that we calibrated on the data from 1987–1993. To be precise, the formula of Altman's Z-score is:

$$\text{Z-score} = 0.717 \times \text{working capital/assets} + 0.847 \times \text{retained earnings/assets} + 3.107 \times \text{EBIT/assets} + 0.420 \times \text{net worth/liabilities} + 0.998 \times \text{sales/assets}$$

while the calibration for the logit model yielded:

$$\text{Logit score}^{10} = 5.65 - 0.98 \times \text{liabilities/assets} - 1.37 \times \text{bank debt/assets} + 2.42 \times \text{cash/current liabilities} + 2.08 \times \text{cashflow/(liabilities} - \text{advances)} - 0.81 \times \text{current assets/net sales} - 1.49 \times \text{current liabilities/assets} - 5.26 \times \text{accounts payable/net assets} + 0.19 \times \text{net sales/assets} + 0.28 \times \text{(net sales} - \text{material costs)/personnel costs} + 8.21 \times \text{ordinary business income/assets} - 0.17 \times \text{net sales/net sales one year ago}$$

To calculate the ROC curves, $\hat{A}$, and confidence intervals for $\hat{A}$ for both rating models, we used the data from 1994–1999, which contained about 200,000 balance sheets and about 825 defaults. In that way we performed an out-of-sample and out-of-time validation. The ROC curves for both rating methods are given in figure 4.

Furthermore, we compared the approach based on asymptotic normality to bootstrapping[11] to see how well the assumption of asymptotic normality is justified. Additionally, we draw three sub-portfolios randomly from our large portfolio to assess how the approach based on asymptotic normality works for smaller portfolios. Sub-portfolio 1 consists of 50 defaulters and 450 non-defaulters, sub-portfolio 2 contains 20 defaulters and 480 non-defaulters, and sub-portfolio 3 contains 10 defaulters and 490 non-defaulters. The results are summarised in tables A, B, C and D.

We see from tables A and B, where we had a sufficiently large number of defaults in our validation sample, that the confidence intervals based on asymptotic normality are close to the confidence intervals that resulted from bootstrapping. In tables C and D, where the number of defaults

[9] *This database consists only of small and medium-size enterprises. We have excluded all companies listed on stock exchanges from the database*

[10] *As the empirical relationship between the ratio net sales/net sales one year ago and its log odds were found to be non-linear, this variable was transformed using the approach described in Falkenstein, Boral & Carty (2000)*

[11] *The number of simulations was 25,000 to obtain the confidence intervals based on bootstrapping in tables A, B, C, D and E*

is small, we find that $\hat{A}$ is in three cases only three or four standard deviations away from one. In these cases, it is not clear how well the normal approximation is justified. We see that the boundaries of the confidence intervals differ by about 2–5 percentage points. However, for these cases with very few defaults in the validation sample, the approximation does not lead to completely misleading results. The central limit theorem, which is the basis of this approximation, appears to be very robust. Our examples provide clear evidence that asymptotic normality can be assumed even for small portfolios and that this concept is applicable to a wide range of practical situations. The main advantage of confidence intervals based on asymptotic normality compared with bootstrapping is the considerably lower computational time to obtain them. It took less than one minute to obtain the confidence interval for the large portfolio in table A using asymptotic normality, while the application of bootstrapping can take several hours for portfolios of realistic size.

To make clear that our analysis is not restricted to rating systems with continuous scores, we want to discuss the calculation of confidence intervals for rating systems with a finite number of grades in some detail. Most commercial banks assign their corporate clients to a small number of rating categories. With only a few rating grades available, one gets a very coarse ROC curve because one has to construct it from a small number of points. However, for the calculation of confidence intervals, it is not essential that the rating score is continuous, as the only driver for the central limit theorem to hold is the number of defaults in the validation sample.

To illustrate our argument with an example, we generate a rating system with seven categories. Each debtor is assigned a score value of one, two, three, four, five, six or seven. We randomly select 5,000 debtors from our sample, including 50 defaulters. We use the logit score to distribute them among the rating categories (each category contains approximately the same number of debtors). However, this distribution of debtors on the rating categories could also have been achieved by human judgement. The only information we use is that a debtor has a score between one and seven, and whether he is a defaulter or not. This is sufficient to apply the methods presented in our paper. Table E confirms that this method of calculating confidence intervals can be applied in very general situations.

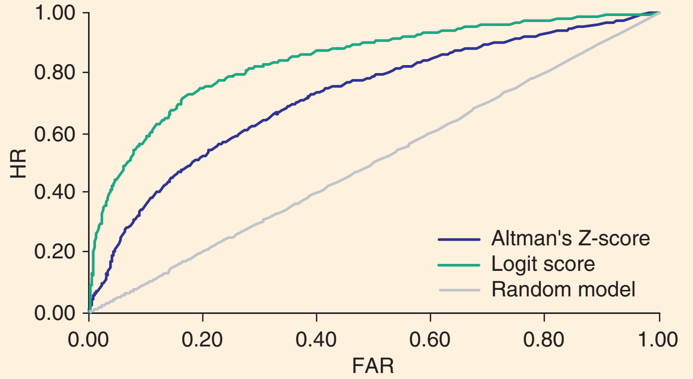
Finally, we apply the test for the difference of two rating systems to the examples in tables A, B, C and D. Again, the results for the two small portfolios have to be interpreted with care since the number of defaults in these examples is rather small. The results are summarised in table F.

Not surprisingly, for the total portfolio we find that the difference of both rating methods is highly significant. In all other cases, there is weak evidence that both models are different. The differences in the results for sub-portfolio 2 and sub-portfolio 3 can be explained by the estimate of the correlation between A^1 and A^2. For sub-portfolio 2 this correlation is 0.35 while for sub-portfolio 3 it is 0.80. This explains why the p-value for sub-portfolio 3 is much lower than for sub-portfolio 2 in spite of the fact that the confidence intervals in both cases are overlapping on a rather large range.

Conclusion

By demonstrating the correspondence of the area A under the ROC curve and the accuracy ratio, we have shown that these summary statistics of the CAP and the ROC are equivalent. Furthermore, this result enables us to use a simple analytical method, based on Bamber (1975), to obtain confidence intervals for these statistics. Additionally, by means of a methodology introduced by DeLong, DeLong & Clarke-Pearson (1988), we have a test at our disposal for comparing two different rating methods being validated on the same data set. Even though these methods rely on asymptotic normality, which can be questionable in practice, in the real data examples we have demonstrated that they can be reliable also for smaller portfolios. Although bootstrap methods appear to be more generally robust, the approach that we discuss here is computationally faster and provides good approximations in some cases. ■

4. ROC curves for Altman's Z-score and the logit model



F. Results of the test for the difference of the areas below the ROC curve of the logit model and the Z-score

Portfolio	$\chi^2(1)$	p-value
Total portfolio	331.14	< 0.0001
Sub-portfolio 1	5.87	0.0154
Sub-portfolio 2	3.53	0.0602
Sub-portfolio 3	6.96	0.0083

REFERENCES

Altman E, 1968
Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy
Journal of Finance 23, pages 589–609

Bamber D, 1975
The area above the ordinal dominance graph and the area below the receiver operating graph
Journal of Mathematical Psychology 12, pages 387–415

Basel Committee on Banking Supervision, 2000
Supervisory risk assessment and early warning systems
December

Basel Committee on Banking Supervision, 2001
The internal ratings-based approach
Consultative document, January

DeLong E, D DeLong and D Clarke-Pearson, 1988
Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach
Biometrics 44, pages 837–845

Falkenstein E, A Boral and L Carty, 2000
RiscCalc private model: Moody's default model for private firms
Moody's Investors Service

Hanley A and B McNeil, 1982
The meaning and use of the area under a receiver operating characteristics (ROC) curve
Diagnostic Radiology 143 (1), pages 29–36

Mann H and D Whitney, 1947
On a test of whether one of two random variables is stochastically larger than the other
Annals of Mathematical Statistics 18, pages 50–60

Sobehart J and S Keenan, 2001
Measuring default accurately
Risk March, pages S31–S33

Sobehart J, S Keenan and R Stein, 2000
Benchmarking quantitative default risk models: a validation methodology
Moody's Rating Methodology

Swets J, 1988
Measuring the accuracy of diagnostic systems
Science 240, pages 1,285–1,293

Bernd Engelmann and Dirk Tasche are members of the banking supervision research group at the Deutsche Bundesbank in Frankfurt. This paper represents their personal opinion and does not necessarily reflect the views of the Deutsche Bundesbank. Evelyn Hayden is assistant professor at the department of business studies at the University of Vienna. The authors thank two anonymous referees for valuable comments. e-mail: bernd.engelmann@bundesbank.de, evelyn.hayden@univie.ac.at, dirk.tasche@bundesbank.de