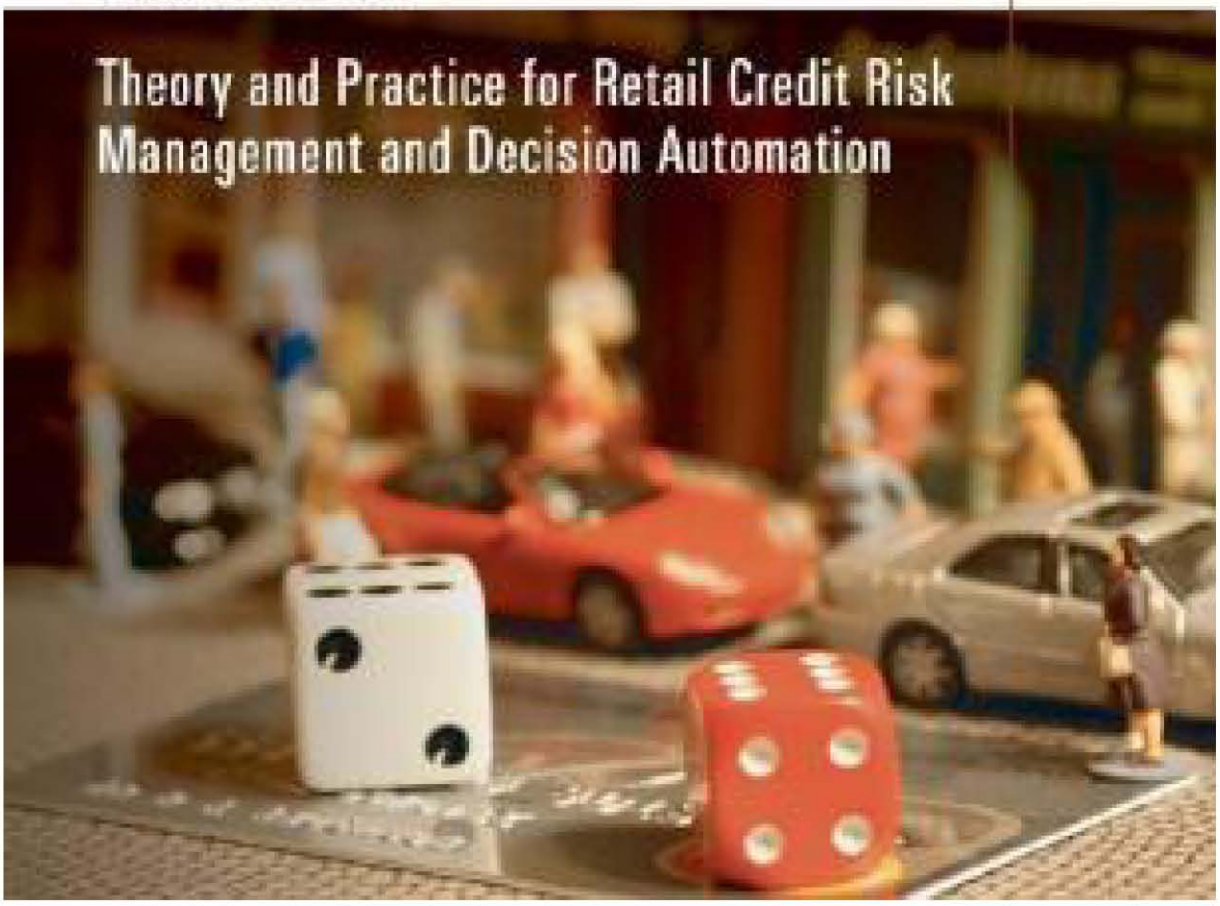
OXFORD

THE CREDIT SCORING TOOLKIT

Raymond Anderson

Theory and Practice for Retail Credit Risk Management and Decision Automation



8 Measures of separation/ divergence

> These measures of goodness of fit have a fatal attraction. Although it is generally conceded among insiders that they do not mean a thing, high values are still a source of pride and satisfaction to their authors, however hard they may try to conceal these feelings.
>
> *J.S. Cramer* (1987)

On a recent visit to Paris, my partner and I visited the Eiffel Tower, and opted to go to the top, level 3. This is 295 metres above the ground, and once at the top, I automatically felt slightly giddy, as though there were a slight sway in the tower. I had never felt this way before though, in spite of having looked over precipices of more than a kilometre on hiking trips in South Africa. When I mentioned the sway to my partner, she said, 'Nonsense!' I insisted that there must be some sway in the tower, whether because of wind (it was a calm day), the movement of lifts up and down its centre or legs, or the shifting of the many and gawking tourists on any of the three observation platforms. She conceded that there might be some sway, but thought it imperceptible, and insisted I was suffering from vertigo. I swore the contrary, but had no means of proving my point. No doubt it can be measured, but we had neither the tools nor access to information to settle the argument.

Credit scoring also needs tools to measure movements. The tools are called measures of separation, measures of divergence, or power/divergence statistics, which are used to determine differences in data distributions, whether during: (i) coarse classing; (ii) variable selection; (iii) segmentation; (iv) result evaluation; (v) post-development validation; or (vi) post-implementation monitoring. These are *bivariate statistics*, that originated in a variety of different disciplines, including mathematics, economics, psychology, and electronics. In some cases, like the Pearson's correlation coefficient, they have little application in credit scoring, but are covered here as part of a larger conceptual framework. For the most part, these statistics are used to assess:

Power—Measure of ranking ability, or dependence between a characteristic, score, or grade and a binary outcome. It shows the extent (random↔perfect), and usually the direction (positive/negative), of the correlation. Measures used include rank-order correlation measures, as well as the information value, KS statistic, and chi-square.

Drift—Measure of variation between expected and actual results. In this instance, direction will seldom be of interest, as the primary concern is deviation from expected. Measures used include the stability index, KS statistic, chi-square statistic, binomial test (and its normal approximation), and the Hosmer–Lemeshow statistic.

According to Thomas et al. (2002:155), in credit scoring the primary distinction is where the statistics are used. *Power* is of greatest interest when scorecard performance is being measured, and is used to prove that the scorecard will add value; big is good, and implies better risk ranking ability. In contrast, when monitoring *drift*, the hope is for minimal change; small is good, and means less variation from the baseline.

The value of these tools comes from their ability to collapse information, whether into a single number, graphic, or table. Furthermore, the statistics often provide a scientific basis for hypothesis testing, which requires the formulation of a null (H_0) and alternate (H_A) hypothesis, and use of a statistical test, to determine whether to reject the former.[1] Care must, however, be taken, because specific ranges are sometimes of interest (especially when dealing with cut-offs), and not the entire distribution. As a result, the KS curve, Lorenz curve, Receiver Operating Characteristic curve, Misclassification graph, and other graphical tools are presented, to aid visualisation of what is happening across the range of values.

As shown in Table 8.1, at a high level, the types of divergence statistics can be described by what is being compared: (i) *frequencies*, of classed characteristics; (ii) *rankings*, of raw or classed characteristics; and (iii) *cumulative percentages*, the percentage of the total that falls below each ranked value, also called an empirical cumulative distribution function (ECDF). Each of them has different strengths and weaknesses, depending on the situation, and although valuable, they should not be used in isolation. Cognizance must always be taken of: (i) peculiarities specific to each situation; (ii) other measures and tools; (iii) the scorecard developer's intuition; and (iv) any insight that can be provided by experts within the business.

Measures of separation—use and interpretation

When used to measure ranking ability, all of the tools mentioned here share the common trait, that the flat maximum that can be achieved depends upon the problem; in particular the portfolio being assessed, and its relationship to the outcome. If a *risk-heterogeneous* group is being assessed, whether using a single characteristic or the final scorecard, the results will always be

Table 8.1. Measures of separation

	Frequency	Ranking	Cumulative percentage
Chi-square	✓		
Kullback divergence	✓		
Spearman's rank		✓	
Gini coefficient		✓	✓
KS statistic			✓

[1] A null hypothesis can never be accepted. Either the evidence is sufficient to reject it at a given confidence level, or it is insufficient.

higher than for a *risk-homogenous* group. There are four primary sources of the homogeneity:

(i) It may be an inherent feature of the population. Anderson (2003b) and Mays (2004) both refer to the seemingly poorer results that are obtained in *sub-prime markets*, which results because the individuals are clustered at the high end of the risk spectrum (this is compounded by the next point, as they have traditionally been financially excluded, and hence not credit active).
(ii) It may result from data deficiencies, either because of poor data quality or lack of relevant data. The greatest leaps in predictive power occur when new data sources become available, that: (i) provide greater transparency, and new insight into the behaviour of individuals (like shared-performance data); and (ii) have a low correlation with that already available.
(iii) It may be the result of truncation by a selection process. Two unrelated examples are: (i) admission-test scores, versus academic grades in universities; and (ii) application scorecard developments, versus post-implementation monitoring of accepts. Assuming that rejects would have performed worse than accepts, any measure of separation based solely on accepts, will be lower than that for the full through-the-door population.
(iv) It may be a by-product of the segmentation, especially when separate scorecards are applied at different levels in the risk spectrum. As in the maxim, 'the sum of the parts is greater than the whole', so too may the apparently poor performance of an individual scorecard be more than offset by its contribution as part of a scorecard set.

As a final note, measures of separation should be used with caution. Falkenstein (2002:184) warns, 'one should not be focused on any single statistical test (of ranking ability), as this indirectly encourages overfitting'. It must also be ensured that 'apples are compared with apples', especially as regards the binary outcomes (good/bad definition), time frames (outcome window and censoring), and truncation (score cut-off and policies). Where firm conclusions have to be drawn, then appropriate statistical tests should be used, as not all measures are suited for hypothesis testing.

Divergence statistic

Perhaps the most straightforward summary measure of separation is the divergence statistic. It is a parametric statistic assumes the values for both groups are normally distributed, and is calculated as the squared difference between the means of two groups, divided by their average variance. The greater the spread of possible values, the greater the difference has to be before the two distributions are considered different. For the instance where the two groups are goods and bads, the formula for a given characteristic can be presented as:

Equation 8.1. Divergence statistic $$D^2 = \frac{(\pi_G - \pi_B)^2}{(\sigma_G^2 + \sigma_B^2)/2}$$

It can be applied to any continuous characteristic, including scores. It suffers, however, because it says nothing about the shape of the distribution. According to Mays (2004), this statistic is closely related to the information value, and it is also mentioned in Siddiqi (2006). It is only covered briefly here, because it is seldom encountered in practice, probably because of: (i) its limited focus on continuous characteristics; (ii) the assumption that the two distributions are normally distributed; (iii) the potential distorting effect of outliers; and (iv) other measures are more common. Even so, it is probably appropriate for many, if not most, score distributions provided by logit and probit models.

8.1 Misclassification matrix

A very simple way of evaluating how well a predictive model has worked, or at least those with binary outcomes, is to calculate the percentage of accounts that have been correctly classified. This was used for Table 7.4, to compare the results from various predictive modelling techniques. The percentage correctly classified is derived from a misclassification matrix that is created by:

(i) choosing a score cut-off;
(ii) marking all accounts below the cut-off as expected bads, and all those above as expected goods;
(iii) cross-tabulating the expected goods and bads against the actuals, using the development definition, or any other definition of interest;
(iv) determining the percentage of accounts that fall into each cell;
(v) calculating the various ratios that can be derived from the model.

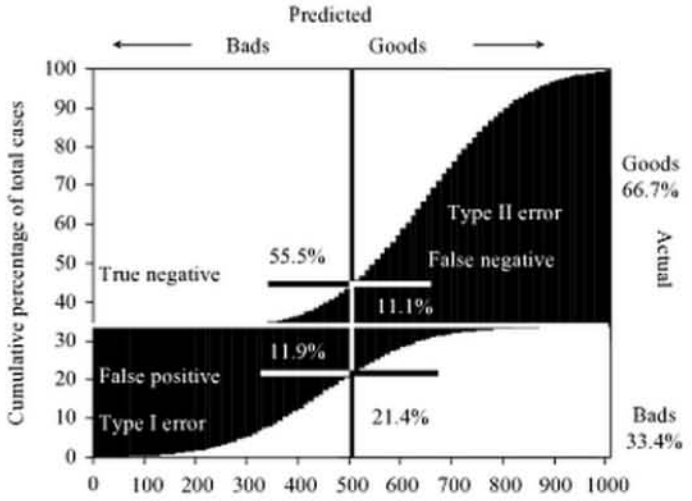
The correctly classified cases are the *true positives* (bads) *and negatives* (goods). If they do not correspond, they are labelled false positives (type I error, expected bad that is good) and negatives (type II error, expected good that is bad). The use of 'negative' for goods and 'positive' for bads is consistent with other tests used to identify rare events, such as in the medical profession; positive means that a patient has been diagnosed with the disease (e.g. HIV positive) whether correctly or incorrectly. In some cases, such as for the construction of the ROC curve discussed in Section 8.4.5, the focus is restricted to those predicted as bad; both true positives (hits), and false positives (false alarms).

The example in Table 8.2 presents the results for a through-the-door population of 150,000 applicants, with a bad rate (admittedly very high) of 33.3 per cent. The cumulative percentage of applicants by score was then reviewed, and it was found that a cut-off of 500 provides almost the same percentage. When split out into the four groups, the false positives and negatives (11.9 and 11.1 per cent respectively) become evident, indicating a misclassification rate of 23.0 per cent. It did, however, have 77 per cent correctly classified.

An issue is the choice of cut-off. The most common choice is one where the number of accounts below the cut-off is equal to the sample bad rate. Lenders may, however, wish to measure changes at different reject rates, and can construct a graph like that in Figure 8.1.

Table 8.2. Misclassification matrix

Actual	Predicted		Totals
	Negative/good	Positive/bad	
Negative/good	83,275	17,850	100,125
	55.5%	11.9%	66.4%
Positive/bad	16,700	32,175	48,875
	11.1%	21.5%	33.6%
Totals	99,975	50,025	150,000
	66.6%	33.4%	100.0%



Here it can be seen that at a cut-off of 500, 64.2 per cent of the bads (21.4/33.4) have been identified, at the expense of misclassifying only 16.7 per cent of the goods (11.1/66.7).

It must also be noted that although the misclassification rate is commonly used to measure model accuracy, it is seldom sufficient, unless the total misclassification costs can also be calculated. Unfortunately, 'per case' figures are difficult to derive, and analysts often use a significant degree of latitude when assuming Type I and II error costs. A similar type of analysis can also be used for comparing scorecards against each other, except in that instance, swap sets between the two are identified. This is particularly useful for comparing recently developed scorecards against those currently in place.

8.2 Kullback divergence measure

Kullback's divergence measure is used to gauge the difference between two frequency distributions. It is used in many disciplines, but as with many other statistics, it has become

masked under other names, that do not give proper credit to its author. In credit scoring, it is referred to either as the: (i) *information value*, when measuring power, to compare good and bad distributions; or (ii) *stability index*, when measuring drift, to compare a distribution at two points in time. It is based upon the *weight of evidence* (WoE), which provides a simple, and theoretically well-grounded, tool for assessing relative risk based upon available information. The WoE is covered first here, before delving into the information value and stability index in more detail.

8.2.1 Weight of evidence

Each and every day we make decisions based upon the probability of some event occurring. We decide on whether or not to cross the street, based on how much traffic there is, and how fast it is going; and on whether or not to take a raincoat, based on the morning weather report, or a look outside to see if it is sunny or cloudy. The probability is far from empirical, as it relies upon personal experiences, or information gained from others.

In 1950, Irving John (Jack) Good, an Englishman who was a Second World War code-breaker, published a book that addressed these personal and subjective probabilities. For any decision, one assesses the circumstances and determines a weight of evidence. Basically, this converts the risk associated with a particular choice onto a linear scale that is easier for the human mind to assess, which for credit scoring can be expressed as:

Equation 8.2. Weight of evidence $$W_i = \text{Ln}\left(\left(\frac{N_i}{\sum N}\right)\middle/\left(\frac{P_i}{\sum P}\right)\right)$$

where P = occurrence (positive), N = non-occurrence (negative), and i = index of the attribute being evaluated (such as 'Income < X'). The precondition is, of course, non-zero values for all N_i and P_i (small adjustments can be made to ensure this).

The WoE formula above is that most often used. It can be restated as: $W_i = \text{Ln}(N_i/P_i) - \text{Ln}(\Sigma N/\Sigma P)$, which illustrates two components: a variable portion for the odds of that group, and a constant portion for the sample or population odds. The WoE for any group with average odds is zero. Note that the two natural log odds values are both restatements of $\ln(p_N/(1-p_N))$. For a characteristic transformation, the WoE variable has a linear relationship with the logistic function, making it well suited for representing the characteristic when using logistic regression (logit).

The WoE is used: (i) to assess the relative risk of different attributes for a characteristic, to get an indication of which are most likely to feature within a scorecard; and (ii) as a means of transforming characteristics into variables. Some software packages provide it as a value or graph (see Figure 8.2), as it is a very useful tool for binning. Attributes with similar relative risks are usually merged. Unfortunately, the WoE does not consider the proportion of accounts with that attribute, only the relative risk. Other tools are used to determine the relative contribution of each attribute, and the total information value (covered next).

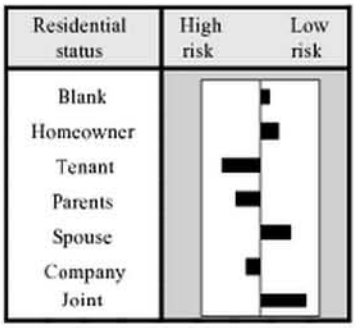


Figure 8.2. Weight of evidence.

With the advent of Basel II and the PD/EAD/LGD framework, there is a trend towards speaking in terms of probabilities, instead of risk grades. While this should not pose a problem for the cognoscenti, it could cause confusion for banks' rank and file. Jack Good's work effectively showed that people can relate to risk grades of 3, 6, 9, and 12, better than percentages of 0.04, 0.32, 2.5 and 17, as the human mind processes risk on a log scale.

8.2.2 Information value

In the early days of credit scoring, Fair Isaac (FI) adopted a measure that they dubbed the information value, to measure the predictive power of a characteristic. Very few people give proper credit to Solomon Kullback, who first published it in 1958, at the time when FI was first finding its feet. It is technically referred to as the Kullback divergence measure, and is used to measure the difference between two distributions. When applied to test results it is expressed as:

Equation 8.3. Information value $$F = \sum_{i=1}^{n}\left[\left(\frac{N_i}{\sum N} - \frac{P_i}{\sum P}\right) \times \mathrm{WoE}_i\right]$$

where N = negative identification (goods), P = positive identification (bads), WoE = the weight of evidence, i = index of the attribute being evaluated, and n = total number of attributes. The result for each attribute reflected between the square brackets is called the 'contribution'.

Values for F will always be positive, and may be above 3 when assessing scores provided by highly predictive behavioural scorecards. Characteristics with values of less than 0.10 are typically viewed as weak, while values over 0.30 are sought after, and are likely to feature in scoring models. Table 8.3 provides a very simple example for applicants' income. The predictive power is marginal; and if included in a scorecard, the point allocations would be low.

Please note, that *weak characteristics* may: (i) provide value in combination with others; or (ii) have individual attributes that could provide value as dummy variables. They should

Table 8.3. Information value calculation

Income	Outcome		Good/bad odds	Column (%)		WoE	Contribution
	Good	Bad		Goods	Bads		
Low	5,000	2,000	2.5	14.3	33.3	−0.847	0.161
Middle	10,000	2,000	5.0	28.6	33.3	−0.154	0.021
High	20,000	2,000	10.0	57.1	33.3	0.539	−0.074
Totals	35,000	6,000	5.8	100.0	100.0	Info value =	0.109

thus not be discarded indiscriminately. Further, even if not considered for the model, they should still be retained for future analysis (Mays 2004).

Like the Gini coefficient, the information value is also sensitive to how the characteristic is grouped, and the number of groups. Unlike the Gini coefficient, however, the information value will provide the same result, irrespective of how the attributes are ordered. It can, however, be difficult to interpret, because there are no associated statistical tests. As a general rule, it is best to use the information value and/or chi-square to assess individual characteristics, and the Gini coefficient (in combination with other measures) for the final scorecard.

8.2.3 Stability Index

The Kullback divergence measure is also used to measure drift, in much the same way as the chi-square statistic. In this instance however, it is called a population stability index, as shown in Equation 8.4, which measures the difference between the development sample and a more recent distribution:

Equation 8.4. Population stability $$F=\sum_{i=1}^{n}\left[\left(\frac{O_i}{\sum O}-\frac{E_i}{\sum E}\right)\times \mathrm{Ln}\left(\frac{O_i}{\sum O}\Big/\frac{E_i}{\sum E}\right)\right]$$

where O and E are the observed (recent population) and expected (development sample) frequencies. The result will always be positive, and a traffic-light approach is used to provide warnings: (i) *green*, less then 0.10, no cause for concern; (ii) *yellow*, between 0.10 and 0.25, some cause for concern, and (iii) *red*, greater than 0.25, concern! It can be used on the final score to provide a measure of score drift, or on individual characteristics. Please note that, once again, the precondition is positive values for all O_i and E_i.

According to Thomas et al. (2002:155), the stability index lacks sophistication and consistency, but it can be considered in combination with other measures, like the Gini coefficient and K-S statistic, which allow significance testing.

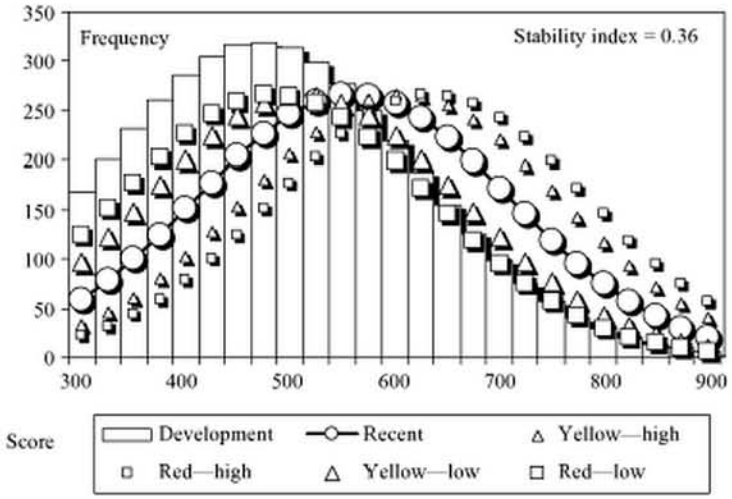


Figure 8.3. Population stability.

The graph in Figure 8.3 illustrates the score distributions for a recent through-the-door population (columns), a development sample (circles), and the warning lights either side of the development sample. The recent and development distributions indicate that not only volumes have increased since development, but also the applicant risk as measured by the score. The population stability index of 0.36 has gone past both the yellow and red lights to the left, indicating that special attention is required (the yellow and red lines are shown for illustration purposes only, and are hypothetical, assuming that neither the volume nor variance changes, only the mean).

Please note, that the scorecard could still be working well, regardless. The only way really to know is to have sufficient performance to assess its ranking capability directly. Otherwise, all that can be done—other than immediately fine-tuning or redeveloping the scorecards—is to get a better understanding of what is causing the shifts. That is the purpose of the score shift report (see Section 25.3.2).

8.3 Kolmogorov–Smirnov (KS)

It seems no conversation on credit scoring is complete without mention of the KS statistic. Even senior executives of financial services companies are familiar with it and in fact try to impress each other by claiming their scoring model has a bigger KS than the next guy's.

Mays (2004:121)

One of the statistics commonly used in credit scoring, as well as countless other disciplines, is the KS statistic. This was developed by two Soviet mathematicians, A.N. Kolmogorov and N.V. Smirnov. Kolmogorov first proposed it in an Italian actuarial journal in 1933.[2] Smirnov

[2] See also the biographical sketch later in this section. No biographical information could be found for Smirnov.

built upon it 1939, and tabulated it in 1948. It is one of several statistics, like the Gini coefficient, that is built upon an analysis of the empirical cumulative distribution function (ECDF). It is a better non-parametric measure for assessing error (or 'goodness of fit') in curve fitting than many others.

According to Mays (2004), the KS statistic is the most widely used statistic within the United States for measuring the predictive power of rating systems. This does not appear to be the case in other environments, where the Gini or AUROC seem to be more prevalent. In any event, it is dangerous to use any single measure in isolation. Mays also indicates that as a measure of ranking ability, KS values range from 20 per cent, below which the model's value should be questioned, to 70 per cent, above which it 'is probably too good to be true'.

The KS Curve (also known as the fish-eye graph) is a data-visualisation tool used to illustrate scorecard effectiveness. It charts the ECDF percentages for goods and bads against the score. In the left-hand chart within Figure 8.4, it can be seen that 56.6 per cent of the bads fall under the score of 470, but only 12.2 per cent of the goods.

The statistic of interest is where the difference is greatest.[3] This is the KS statistic, being the maximum absolute difference between the two curves: $0 < D_{KS} < 1$. In the right-hand graph of Figure 8.4, the distance at the score of 470 is 44.4 per cent (56.6 less 12.2 per cent), but this increases to 49.4 per cent at a score of 550.

Equation 8.5. KS statistic $D_{KS} = \max\{\text{abs}(cpY - cpX)\}$

While this is a very simple measure to understand, it may be too simple. The KS statistic often applies to a point on the curve that has no relevance to the problem at hand, especially where

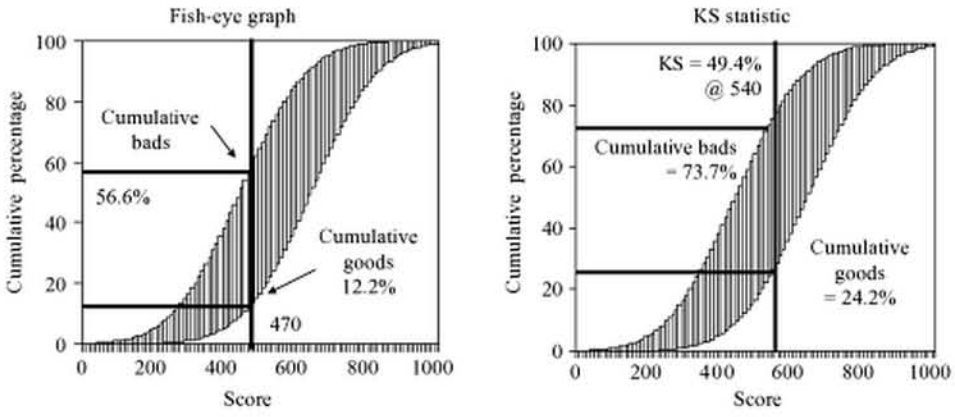


Figure 8.4. Kolmogorov–Smirnov.

[3] The treatment differs depending upon whether one or two samples were used to generate the values.

it is a long way above or below an application scorecard cut-off. It is thus usually used in conjunction with other measures.

The most common uses of the KS statistic are as a measure of predictive power, and to determine whether or not two distributions differ. Hypothesis tests can also be applied, by comparing it to $KS_{critical}$. If it is less, then there is a good chance that the two distributions are the same. $D_{KS\text{-}critical}$ is calculated as $c/\sqrt{n}$, where c varies according to the significance level, and type of distribution being assessed, and n is the sample size. In most instances, it is sufficient to assume that the distribution is normal, in which case 'c' is 1.36 at a 0.05 significance level.

For an exponential and Weibull distribution, the values would be 1.08 and 0.874 respectively at the 0.05 significance level. If the type of distribution is not known up-front, this can be determined using the expected probabilities for different possible distributions, and applying a chi-square test to determine which is most appropriate. If that is not feasible, the more conservative 0.874 should be used.

Note however, that $D_{KS\text{-}critical}$ is very sensitive to changes in the value of n. If two samples of 10,000 each are being compared, then $D_{KS\text{-}critical}$ is 0.0136 (1.36%). The logic is that as sample size increases, sampling error decreases, and the test has to be more stringent.

Note that this formula assumes that the two populations are the same size. Mays (2004) provides a formula for two sample sizes, where $D_{KS\text{-}critical}$ is $c/\sqrt{(n_1+n_2)/(n_1n_2)}$, '$n$' is the size of the respective populations, and 'c' is 1.22 at the 95 per cent confidence interval, for a one-sided test of significance. She does not indicate how the value for 'c' was obtained.

Like with the other statistics, care must be taken when using the KS statistic for comparisons. Mays (2004) warns that when assessing application scorecards, the KS statistic for the accept population will differ depending upon the cut-off, because of the truncating effect of rejects, especially if compared to the full development sample. In the particular example she provides, the KS statistic for high risk, low risk, and the full sample are 27, 36, and 49 per cent respectively. Likewise, when assessing variation in scorecard performance over time, it must be ensured not only that the good/bad definition and outcome windows are the same, but also that the same score cut-offs and policy rules are applied, or are at least similar.

Andrei Nikolaevich Kolmogorov (1903–1987)—Biographical sketch

A.N. Kolmogorov was a renowned Soviet mathematician, who wrote over 300 papers on practically every aspect of mathematics, and spent much time advancing mathematics in their school system, especially for the gifted. His father was the agronomist son of a clergyman, and his mother was of aristocratic stock. Kolmogorov was born out of wedlock in Tambov, during a delay while his mother was returning from the Crimea, and she died in childbirth. He was brought up by her sister, a woman of high social ideals, on his grandfather's estate in Tunoshna, near Yaroslavl, and took his maternal grandfather's surname.

As a teenager, Kolmogorov worked briefly as a railway conductor, prior to starting at Moscow University in 1920.

Besides mathematics, he also had interests in poetry and history, and briefly spent time researching fifteenth and sixteenth century manuscripts on agrarian relationships in ancient Novgorod. He stuck with mathematics though, and before graduation in 1925, he had already published eight papers. His interest in probability theory started in 1924, and in 1929—besides completing his doctorate—he published a paper titled *A general theory of measure and the calculus of probabilities*, that started providing foundations where none had existed before. His 1931 paper, *Analytical methods in probability theory*, built on Markov's work to develop the modern theory of Markov processes (diffusion theory). In 1933, his *Foundations of the Calculus of Probabilities* was published in German, which became the definitive work on probability theory, and established it as a formal branch of mathematics. He is also renowned for several articles on the theory of poetry and the statistics of text, where he analysed Pushkin's poetry. Kolmogorov was highly recognised by the Soviet state, receiving seven orders of Lenin, the Order of the October Revolution, and the Hero of Socialist Labour. Of all Soviet mathematicians, he was the most recognised outside of the USSR, and besides receiving a number of honorary doctorates, he was also elected as full or honorary member of many foreign mathematical and other societies.

8.4 Correlation coefficients and equivalents

It is part and parcel of human nature to notice and strive towards an understanding of patterns and relationships, whether because of idle curiosity or an innate desire to control the environment. If X varies with Y then maybe, just maybe, Y can be influenced if X is controlled. The problem, of course, is that correlation does not imply causation—the observed relationship is, more often than not, related to one or more other factors. Even so, knowledge of these correlations provides direction, and much time is invested in studying them.

As a result, the workhorse of modern quantitative analysis is the correlation coefficient, which measures the degree of association, or covariance (ratio of shared variation to independent variation), between two variables X and Y. Here the two most commonly used types of correlation coefficient are discussed: product-moment, which measures how well a linear formula, of the form $Y = a + bX$, can describe the relationship, and for which both X and Y must be scalar; and rank-order, which measures the extent to which the relationships are monotonic, where X and Y can be either scalar or ordinal. Where it is a linear relationship, it is assumed that both values are normally distributed. If the assumption is seriously violated, then a rank-order correlation may be the better option.

The strength of the correlation is usually represented by a coefficient, in the range $-1 \leq r \leq +1$. The sign indicates whether the two variables move in the same, or opposite directions. If the value is at or near zero, the variables are statistically independent; but at the

Table 8.4. Correlation

Value range	Strength & direction	
$r = +1.0$	Perfect	Positive
$+0.9 < r < +1.0$	Strong	
$+0.5 < r \leq +0.9$	Moderate	
$0.0 < r \leq +1.5$	Weak	
$r = 0$	Uncorrelated	
$-0.5 \leq r < 0.0$	Weak	Negative
$-0.9 \leq r < -0.5$	Moderate	
$-1.0 < r < -0.9$	Strong	
$r = -1.0$	Perfect	

two extremes, they are statistically dependent. The labels used in Table 8.4 are highly subjective, whereas in truth the strength of each measure will vary according to circumstances. A key use of correlation coefficients is for hypothesis testing, which also requires:(i) the formulation of a null and alternate hypothesis to be tested (ii) calculation of the standard deviation or variance; and (iii) degrees of freedom (see sections following). This section covers all correlation measures that provide values in the –1 to +1 range, or equivalent, and associated tools:

8.4.1 Pearson's product-moment

Although Galton first proposed the concept, it was Karl Pearson who came up with the original and most widely-used formula, which bears his name—the Pearson product-moment correlation coefficient. It is related to ordinary least squares regression, but rather than deriving a beta coefficient, it instead measures the extent to which there is a linear relationship between the two variables. The correlation for a population is usually represented using the Greek character ρ (*rho*), but for a sample the Roman letter r is used. The formula varies from textbook to textbook, but the one most commonly used is:

Equation 8.6. Pearson's correlation

$$r = \frac{N\sum XY - \sum X \sum Y}{N\sqrt{(\sum X^2 - (\sum X)^2)(\sum Y^2 - (\sum Y)^2)}}$$

It does have some restrictions. Both X and Y must be continuous characteristics that are approximately normally distributed, which makes it inappropriate for use with binary, ordinal, and discrete characteristics. Furthermore, if there are outliers, the results can be highly distorted. If the product-moment coefficient proves infeasible, it is still possible to consider a rank-order correlation statistic as an alternative.

The formula can be restated as $r = (\sum z_X z_Y)/N$ if the two variables are represented by standardised Z-scores. These are obtained by substituting X and Y for V in the formula $z_V = (V - \overline{V})/\sigma_V$ to obtain new values with a mean of zero ($\overline{z}_V = 0$) and standard deviation of one ($\sigma_{z_V} = 1$). This requires that X and Y are normally distributed, or can be transformed into something that is normally distributed.

A related statistic is the coefficient of determination, or r-squared (r^2) which can be interpreted as the proportion of variance in Y that is contained in X. Thus, if $r_{x,y}$ has a value of 0.9, then $r^2_{x,y}$ indicates that 81 per cent of the variance in Y is explained by changes in X, and vice versa. It is recommended that r-squared be used as a measure of association, as the correlation coefficient overstates the relationship, especially at lower values of r.

Karl Pearson—biographical sketch[4]

Karl Pearson (1857–1936) could well be called the founder of modern quantitative analysis. He was an English mathematician, renowned for dealing in symbols and formal truths. His interest was in the analysis of large samples to determine correlations, and he is credited with the product-moment correlation coefficient, the chi-square statistic, and the 1894 coining of the term 'standard deviation'.

After graduating from Cambridge University in 1879, he dabbled briefly in German literature and law, but by 1883 he was a professor of mathematics at University College, London (UCL), where he spent the rest of his working career. In his book, *The Grammar of Science* (1892), he anticipated some ideas later proposed by Einstein's relativity theory. He also developed a keen interest in heredity and evolution, and over the years 1893 to 1918, he wrote 18 papers, that are collectively referred to as the *Mathematical Contribution to the Theory of Evolution*. Although he claimed to be a socialist, and supported their causes, he was a believer in eugenics, and openly advocated 'war with inferior races'.

[4] Department of Statistical Science, University College London, England.
http://www.ucl.ac.uk/Stats/department/pearson.html
School of Mathematics and Statistics, University of Saint Andrews, Scotland.
http://www-history.mcs.st-andrews.ac.uk/Mathematicians/Pearson.html
O'Connor J.J. and Robertson E.F.
http://www-groups.dcs.st-and.ac.uk/%7Ehistory/Mathematicians/Pearson.html

In 1901, Pearson, Weldon, and Galton co-founded the journal Biometrika to develop statistical theory further. In 1911, Pearson founded the UCL's Department of Applied Statistics, the first university statistics department in the world, and was the first Galton Professor of Eugenics, chairing from 1911 to 1933. He had serious theoretical disputes with Ronald Fisher, who focused on small samples and finding causes. Fisher refused an offered post of Chief Statistician at the Galton Laboratory in 1919, because he would have reported to Pearson.

Hypothesis testing

While it is very nice to have a measure of association, in many cases it is needed to draw a conclusion. With correlation coefficients, tests can be done for dependence, independence, or direction in the correlation. A null and alternative hypothesis, H_0 and $H_A$respectively, are stated of the form:

H_0—The two variables are not linearly related, $\rho = 0.0$.

H_A—They are linearly related, $\rho <> 0.0$

The other possibility is to compare the correlation coefficients that have been calculated for two samples. For example, if the lender wants to determine whether the correlation between income and age for both loan and card applicants is the same, then the hypotheses would be stated as:

H_0—The two variables have the same correlation, $\rho_{LOAN} = \rho_{CARD}$.

H_A—They do not have the same correlation, $\rho_{LOAN} <> \rho_{CARD}$.

In either case, if the value of r calculated for the sample does not fall within the range demanded by the test, then the null hypothesis must be rejected. There is a problem however, because the t-tests and z-tests demand that the values tested are—at least approximately—normally distributed, and r is not. If multiple samples are taken from the same population, the distributions will only be approximately normal if the correlations are relatively low, say less than 0.5. For higher values, the distribution tends to be skewed to the right, with the skewness increasing as *rho* approaches 1.0.

In order to correct this, *rho* is converted using Fisher's z' transformation, which if applied to the r from repeated samples, would provide a value for z' that is normally distributed, and has a standard error $\sigma_{z'} = 1/\sqrt{(N-3)}$, where N is the number of observations. The transformation is:

Equation 8.7. Fisher's z' transformation $$z' = 0.5 \ln\left(\frac{1+r}{1-r}\right)$$

The transformation has minimal impact for values of r below 0.4, but starts growing thereafter. Hypothesis testing can be done using a Student t-test to: (i) determine whether or not the

correlation is dependent, independent, or equal to a certain value; or (ii) to compare correlations for data taken from independent samples.

Example: There is a correlation of 0.2 between income and age in a sample of 1,000. We wish to determine with 95 per cent confidence that they are not independent:

H_0: The two variables are independent, $\rho = 0.0$.
H_A: The two variables are not independent, $\rho <> 0.0$.

The value for z' is also 0.2, so the next task is to obtain $z_{critical}$ from the Student t-test table in Appendix B. This can be tricky, because this is a two-tail test—H_0 states that the variables are independent, so ρ must fall in a restricted range around zero. The ρ-value used is then 0.975, and the associated $z_{critical}$ is 0.3. Given that $[-z_{critical} < z' < z_{critical}]$, the null hypothesis can be rejected.

This example was quite simple, as hypothesis tests often require much more complex calculations to come up with a z-score that can be used in the Student's t-test. To look for help on the Internet, do a search on (Pearson correlation Fisher transformation).

8.4.2 Spearman's rank-order

While using a product-moment correlation is first prize, it is often precluded because the relationship is not linear, or the distributions are not normal. Approximations are still possible however, if the two characteristics are at least ordinal. It was Charles Spearman who came up with a formula, to measure the monotonic relationship between the rank-ordering of two variables. This is effectively the same as Pearson's correlation coefficient, except it is a non-parametric test. It has the distinct advantages that it: (i) can assess non-linear relationships; and (ii) is not affected by outliers. The formula used is:

Equation 8.8. Spearman's rank-order correlation

$$r_S = 1 - 6\frac{\sum(x_R - y_R)^2}{N^3 - N}$$

Within the formula, the term $(x_R - y_R)$ refers to the difference in the respective ranks for the same observation, and N refers to the total number of cases that are being ranked. There is a complication here for ties, in which instance the average rank for the tied cases should be used.

How would this statistic be used in credit scoring? Its primary use is for the comparison of different scores or grades, provided for the same group of cases. Comparisons can be made of new versus old, option A versus option B, or internal versus external. Its use is especially prevalent for benchmarking a lender's own internal credit-risk grades against those provided by a rating agency, or a model developed to assess the same cases. In any case, if the two are perfectly correlated, then the extra one provides no value, but there is usually a much less-than-perfect correlation.

Charles Spearman—biographical sketch

Charles Spearman (1863–1945), a British behavioural psychologist and statistician, spent 20 years in the army before completing his Ph.D. degree in 1904, at age 41. Besides being known for the rank-order correlation coefficient and factor analysis, both introduced within months of each other in the same journal in 1904, he also formulated the classical mental tests, and came up with a two-factor theory of intelligence, distinguishing between 'general intelligence' and 'specific factors of intelligence'.

8.4.3 Pareto principle and Lorenz curve

Surprisingly, several of the tools used to analyse and illustrate the result of scorecard developments stem from the field of economics. These include the Lorenz curve and the Gini coefficient. The late 1800s and early 1900s saw the emergence of both Marxist and Fascist ideologies in Europe, and a focus on income distributions in different countries by various academics. Vilfredo Pareto (1848–1923) was an Italian engineer, who later became an economist, and later yet took to sociology. In 1896, he noted how 80 per cent of the land in Italy was owned by 20 per cent of the population, and saw that this ratio also applied to land ownership and income in other countries. The ratio also applied in many other instances, and is now known as the 'Pareto principle', or '80/20 principle'.

In 1905, the American mathematician Max Otto Lorenz (1876–1959) went further, to develop what is today called the 'Lorenz curve', as a data-visualisation tool for displaying income inequality within society. The income data is sorted in decreasing order, and the cumulative percentages of both income and population are calculated as $\mathrm{cp}V_i = \sum_{j=1}^{i} V_i / \sum V$, where i refers to the rank in an ordered list.

The results are both then plotted on an XY graph, like that shown in Figure 8.5 (the graph will be inverted if income is sorted in ascending order). From this, it can be seen that

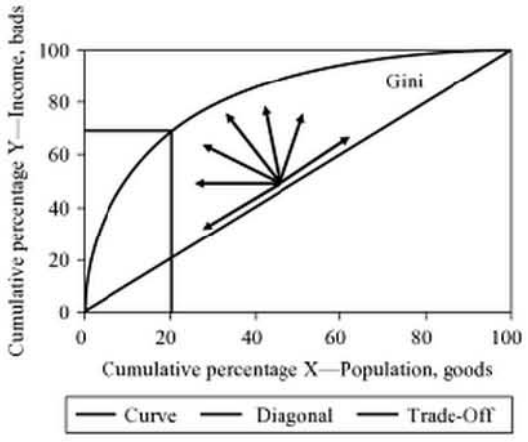


Figure 8.5. Lorenz curve.

20 per cent of the population earns 70 per cent of the income. The space within the bow above the diagonal represents the extent of inequality. Perfect equality lies exactly on the diagonal, and perfect inequality would cover the entire space above (or below) the diagonal.

This same curve is applied in the scoring world, to illustrate a model's ability to separate good and bad accounts, and may be called a *power curve*, *trade-off curve*, *efficiency curve*, or *Receiver Operating Characteristic* curve. Cumulative bads are plotted on one axis, and cumulative goods on another. A model that has no predictive power implies perfect equality, and a model that is perfectly predictive implies perfect inequality.

8.4.4 Gini (rank correlation) coefficient

One of Pareto's contentions was that income inequality would reduce in richer societies. In 1910, Corrado Gini proved him wrong by comparing income inequalities between countries, using what is today known as the Gini coefficient, which is the area between the curve and the diagonal, as a percentage of the area above the diagonal (for interest, the average value for most developed countries is about 40 per cent, and the greatest inequalities are in Brazil and South Africa, with values around 60 per cent). It is calculated as:

Equation 8.9. Gini coefficient $$D = 1 - \sum_{i=1}^{n}((\text{cp}Y_i - \text{cp}Y_{i-1})(\text{cp}X_i + \text{cp}X_{i-1}))$$

where cp$Y$ is the cumulative percentage of ranked income, and cp$X$ is the cumulative percentage of people. The result is a rank correlation coefficient, which is exactly the same as the Somer's D statistic provided by many statistical software packages. The Gini coefficient is not used for hypothesis testing, but does provide a powerful measure of separation, or lack of it.

The same calculation has been co-opted into a lot of other disciplines, including credit scoring, where it is often referred to as an accuracy ratio or power ratio. The Gini coefficient is used as a measure of how well a scorecard is able to distinguish between goods and bads, by having the bads as P and goods as N as in Table 8.6. The end result is again a value representing the area under the curve (also see Section 8.4.5).

The Gini coefficient does have some sensitivity: (i) it can be exaggerated by increasing the *indeterminate range*; and (ii) it is sensitive to the category definitions, in terms of contents, number, and ordering. As a result, some care must be taken in how it is used and interpreted.

Table 8.5. Income inequality

Income class	Totals		Per capita income	Cum (%)		$\text{cp}X_i + \text{cp}X_{i-1}$ (%)	$\text{cp}Y_i - \text{cp}Y_{i-1}$ (%)	Z (%)
	People	Income (mn)		People	Income			
Rich	1,000	60	60,000	20.0	50.6	20.0	50.6	10.1
Middle	1,500	36	24,000	50.0	81.0	70.0	30.4	21.3
Poor	2,500	22.5	9,000	100.0	100.0	150.0	19.0	28.5
	5,000	118.5	23,700	Gini coefficient = 1 − Sum (Z)				40.1

Table 8.6. Scorecard effectiveness

Score	Outcome		G/B odds	Cum (%)		$cpN_i +$ cpN_{i-1} (%)	cpP_{i-1} cpP_{i-1} (%)	Z (%)
	Goods	Bads		Goods	Bads			
Low	5,000	2,000	2.5	2.0	33.3	2.0	33.3	0.7
Middle	45,000	2,000	22.5	20.0	66.7	22.0	33.3	7.3
High	200,000	2,000	100.0	100.0	100.0	120.0	33.3	40.0
	250,000	6,000	41.7	Gini coefficient = (1−Sum(Z))				52.0

When measuring scorecard power, the exaggeration associated with a potentially oversized indeterminate range can be avoided by assessing bads versus not bads, possibly even using a stricter bad definition. The same applies to any other correlation measure, when used to rate the final model's predictive power.

The Gini coefficient provides a single value that represents predictive power over the entire range of possible values. There are a lot of instances though, especially in application scoring, where lenders' greater interest is in the model's power around the cut-off. It is thus always wise also to consider some other measure when comparing model results, such as the percentage of bads at different accept rates.

What is an acceptable Gini coefficient? There are no hard and fast rules, and those rules of thumb that do exist vary, depending upon the development type. In most retail application scoring, a Gini coefficient of 50 per cent plus is more than satisfactory, while less than 35 per cent is suspect, and 30 per cent possibly unacceptable. In contrast, for behavioural scoring with a one-year outcome window, Gini coefficients of over 80 per cent are possible, while anything below 60 per cent might raise suspicions. In all cases, these values apply to resulting scorecard sets, and not individual scorecards.

Corrado Gini—biographical sketch

Gini's association with a measure of income disparity might make one fallaciously think his interest was in social welfare, but in truth, Corrado Gini (1884–1965) was a keen fascist theorist who published *The Scientific Basis of Fascism* in 1927, and was the leader of Italy's eugenics movement under Mussolini from 1934. He was born into a family of landed gentry in the Treviso area of Italy. He started studying law at the University of Bologna, but his interests directed him into the social sciences (especially demography, sociology, and economics), and statistics—the latter being used to complement and support other research. He took over the Chair of Statistics at the University of Cagliari in 1910 and at the University of Padua in 1913, and

provided several significant contributions to the field of statistics before the end of First World War. His ideas were not, however, well accepted in the statistical arena, because he did not explore their mathematical basis.

In 1920, Gini founded the journal *Metron*[5] (which he directed for the rest of his life), the focus of which was restricted to ideas that could be practically applied. He was active politically, and highly regarded within Italian political circles. In 1923, he moved to the University of Rome, where he later became a professor, founded a sociology course, set up the School of Statistics (1928), and founded the Faculty of Statistical, Demographic, and Actuarial Sciences (1936). In 1926, he became president of the Central Institute of Statistics, and founded the journal *La Vita Economica Italiana*. In 1929, he also founded the Italian Committee for the Study of Population Problems, and in 1934, its journal *Genus*. The committee survived the Second World War and the fall of fascism, primarily due to the quality of its work. Over the following years, he was president of many professional societies, and received several awards before his death in 1965.

8.4.5 Receiver operating characteristic

While many of the statistics used in credit scoring have their origins in the social sciences, one statistic's origin is totally different. The Receiver Operating Characteristic (ROC) was developed in the 1940s, to measure radar operators' ability to distinguish between a true signal and noise. In the 1950s and 1960s, it was adopted in the field of psychology, for the study of behavioural patterns that were barely discernible, and could not be explained using existing theories. Today, the ROC is used widely in medicine, engineering, and other fields—including credit scoring. It falls under the heading of 'signal detection theory', two key concepts of which are: (i) *sensitivity*, ability to mark true positives; and (ii) *specificity*, ability to identify true negatives. Using the example illustrated in both Table 8.2 and Figure 8.1, at a score of 500, the sensitivity and specificity are 83.3 and 64.2 per cent respectively.

Building upon this, the ROC curve is the plot of $X = \Pr[S_{FP} \leq S_{Cut-off}]$ against $Y = \Pr[S_{TP} \leq S_{Cut-off}]$ as the cut-off is varied, where X= sensitivity, the true positive rate, or hit rate; and Y = 1– specificity, which is the false positive rate, or false alarm rate. The resulting curve looks much like that in Figure 8.6, which uses the same data as Figure 8.5 for the Lorenz curve.

In most of the literature, especially for medicine and psychology, the ROC curve is very jagged, as it is usually used in instances where the signal is weak or non-existent. A concave curve above the diagonal occurs only where the likelihood ratio $LR_i = p_i^+/p_i^-$ has a monotonic relationship with the measure being evaluated. If the curve goes below the diagonal, the model is getting it wrong, but a reversal of sign will correct it.

[5] Most of this information was obtained from Metron's web page, http://www.metronjournal.it/storia/ginibio.htm

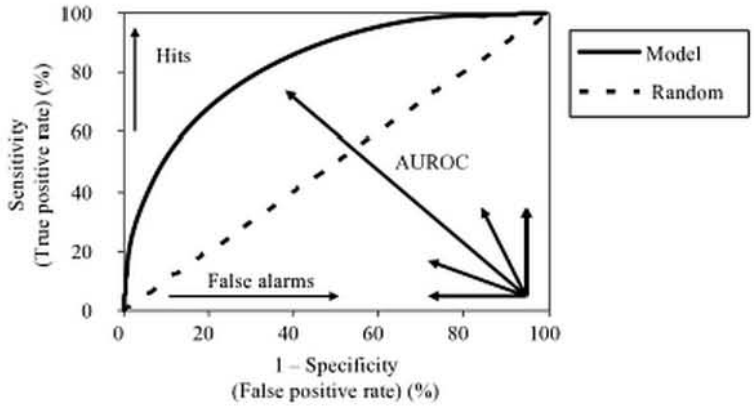


Figure 8.6. ROC curve.

There is also a summary statistic, much like the Gini coefficient, except it provides the percentage of the total area under the ROC curve (AUROC), as opposed to the area above the diagonal. This may either be called AUROC, or the c-statistic. The relationship between it and the Gini coefficient is $c \approx (D+1)/2$; for example, a Gini coefficient of 52 per cent equates approximately to an AUROC of 76 per cent. The formula usually used is:

Equation 8.10. AUROC $c_{P,N} = \Pr[S_{TP} < S_{TN}] + 0.5 \Pr[S_{TP} = S_{TN}]$

In English, it states that the area under the curve is equal to the probability that the rating for a true positive will be less than that for a true negative, plus 50 per cent of the probability that the two ratings will be equal. An AUROC of 50 per cent implies that the model is no better than making a random guess; and a value of 100 would indicate the unlikely occurrence of perfectly correct predictions. Likewise, any value less than 50 per cent implies that the model is getting it wrong with some consistency, and 0 per cent means the predictions are perfectly wrong.

This value can then be used for confidence testing, but the maths are complex and seldom used within credit scoring. With some patience, anybody who is interested will be able to find the formulae on the Internet. One possible source is Engelmann et al. (2003), who use the Mann-Whitney U test, and present several formulae for the variance that differ according to the hypothesis. The simplest is for testing whether a model has any predictive power: $\sigma^2 = (N_D + N_{ND} + 1)/(12N_DN_{ND})$.

As a final note, when measuring scorecard performance: (i) the Lorenz curve and ROC curve are almost exactly the same; and (ii) the values for the Gini coefficient and AUROC are extremely highly correlated. Yet, in spite of Lorenz and Gini being first on the block, they have been supplanted by ROC and AUROC. The reason is that Messrs. Lorenz and Gini were economists, who developed tools for univariate analysis of a ranked variable (income) versus

its count, which tools are only by chance being applied more broadly. In contrast, ROC and AUROC were developed by radio-analysts, for bivariate analysis of a ranked signal-detection measure versus a binary outcome, 'Was there a signal or not?' Credit scoring is a form of signal detection mechanism that fits neatly into the latter camp, and as a result, the ROC/AUROC concepts have taken precedence. The Gini coefficient is still often referred to, but as another way of measuring the area under the ROC curve.

8.5 Chi-square (χ^2) tests

You can predict nothing with zero tolerance. You always have a confidence limit, and a broader or narrower band of tolerance.

Dr Werner Karl Heisenberg, German physicist and Nobel laureate (1901–1976)

And now, another ancient Greek character! This time their equivalent of 'x', which is called 'chi', pronounced like 'sky' without the 's'. The chi-square (χ^2) test looks for a linear relationship between two characteristics, and the resulting *p*-value provides a measure of reliability—the probability that the *similarity (goodness of fit)* or *difference (independence)* between them is not a chance occurrence. It is usually used to evaluate a theory or hypothesis, by comparing *observed (actual)* and *expected (estimated)* distributions. There are several variations of the chi-square calculation, but the original and most commonly used is Pearson's chi-square test:

Equation 8.11. Pearson's chi-square $\chi^2 = \sum_{i=1}^{n}\left((O_i - E_i)^2 / E_i\right)$

where O = the observed frequencies and E = the expected frequencies for each class i. Basically, a χ^2 value of zero indicates a perfect fit, and χ^2 increases as the distributions become dissimilar, eventually becoming so high that one can only conclude that the two distributions bear no relationship to each other (independent).

Expected frequencies should be large. A rule is that no E value should be zero, and no more than 1/5th should have values less than 5 (some people insist on at least 10 or 20), otherwise the test can be considered invalid. Categories can be collapsed to accommodate this. If there are only two categories (d.f. = 1), Yates' correction should be applied, which deducts 0.5 from each $(O-E)$.

The chi-square is then converted into a *p*-value—a percentage that indicates whether or not the fit is a random occurrence: as χ^2 approaches 0, the *p*-value approaches 100 per cent; and as χ^2 increases, the *p*-value approaches zero. The task of converting this into an exact probability is where the test starts getting complicated, if not painful. For those less technically inclined, the rest of this section should be skipped.

The conversion depends upon the degrees of freedom (d.f.), meaning the number of independent pieces of information contained in a statistic. In most instances, the d.f. is calculated as $(n-1-a)$, where n is the number of classes, and 'a' is the number of assumptions, if any, made in the null hypothesis.

What usually happens is that null and alternative hypotheses, H_0 and H_A respectively, are stated of the form:

H_0—The observed distribution fits a certain distribution.

H_A—The observed distribution does not fit a certain distribution.

A test is performed to determine whether the null hypothesis is true at a given threshold significance level (SL); the higher the SL, the less the chance that the null hypothesis will be wrongly rejected. For the above instance, the threshold p-value is equal to the SL, α or $p_{critical}$; but if the hypothesis is turned around, then α is equal to one minus the SL. Nowadays, most spreadsheets and statistical packages are capable of calculating the p-value directly from the two distributions, or alternatively, can calculate the p-value using only the χ^2 and d.f. The null hypothesis is rejected if $p < \alpha$ for goodness of fit, or if $p > \alpha$ for independence.

In the absence of such tools, it is necessary to revert to the dark ages, and the use of tables. The p-value and d.f. are used to select $\chi^2_{critical}$ from a table, like that in Appendix A, and $\chi^2_{critical}$ is then compared against the calculated χ^2 value. The null hypothesis is rejected if $\chi^2 > \chi^2_{critical}$ for goodness of fit, or $\chi^2 < \chi^2_{critical}$ for independence.

Problem: Complaints have been received from the application processing area that high application volumes during certain quarters are affecting service levels. If this is true, then management will have to assign extra resources for peak periods. Application volumes per quarter (000s) have been averaged over several years, and the results are provided in Table 8.7, which yields a χ^2 value of 2.33. Management wishes to ascertain with 80 per cent certainty that these fluctuations imply meaningful differences before doing anything.

H_0—The number of applications is evenly spread over the quarters.

H_A—The number of applications is not evenly spread over the quarters.

From Appendix A, it can be found that $\chi^2_{critical}$ for d.f. = 3 and α = 20% is 4.642, which is more than the calculated χ^2 of 2.33. The hypothesis is rejected, and things are left alone. If the confidence level was reduced to less than 50 per cent (unlikely!), the decision might change.

Table 8.7. Chi-square calc

Group	Actual	Expected	Chi-square
Q1	292	300	0.21
Q2	320	300	1.33
Q3	285	300	0.75
Q4	303	300	0.03
Totals	1200	1200	2.33
$N = 4$, d.f. = 3			p = 50.74%

These functions are a bit easier to work with, if they can be visualised. Figure 8.7 illustrates the relationship between χ^2 and the associated probabilities, where there are 11 categories, and independence is being tested at an 80 per cent SL ($p = 0.2$). The points to the right are those above the $\chi^2_{critical}$ value of 13.442. As can be seen, the test becomes more demanding as $\chi^2_{critical}$ increases, and p decreases.

A similar pattern exists for other d.f. values, but the shape of the distributions changes. In Figure 8.8, it can be seen that as the number of categories increases, so too does $\chi^2_{critical}$ for each confidence level. Also, where the d.f. is low, the probability distribution is highly skewed to the left; but as it increases, the distribution starts looking like a normal distribution.

In credit scoring, the most obvious uses of the chi-square test are: (i) to measure the drift in score or characteristic distributions over time; and (ii) to measure the differences between the good and bad distributions. In these cases, contingency tables are being compared, and the rules for calculating the d.f. changes. The number of assumptions being made is affected by

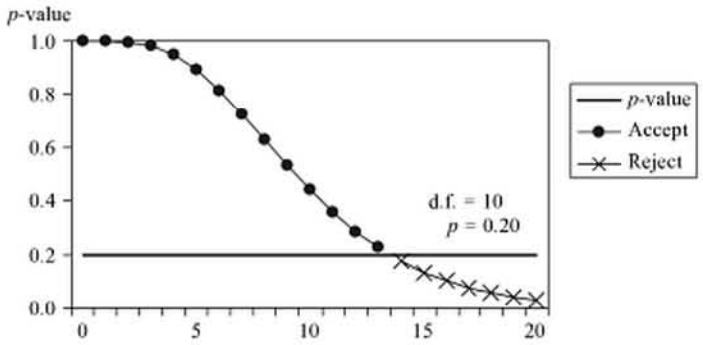


Figure 8.7. Chi-square distribution.

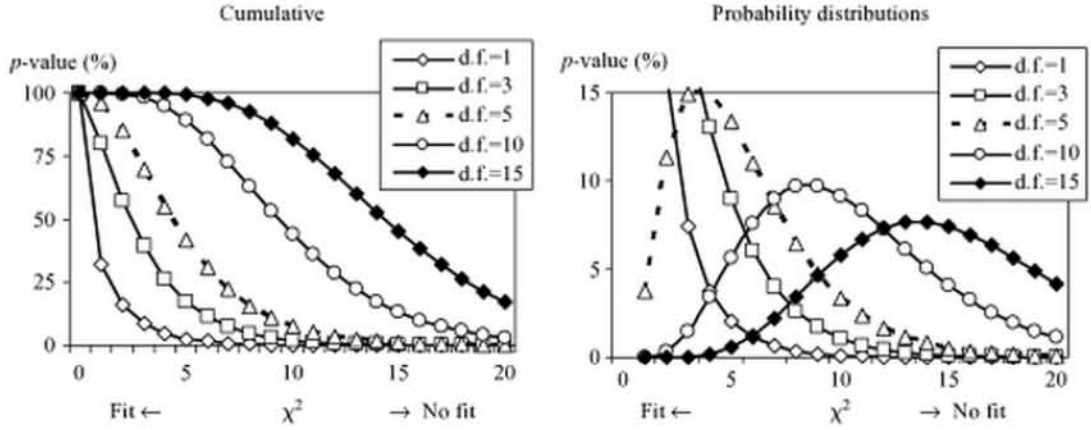


Figure 8.8. Degrees of freedom.

whether or not the row and column totals (expected and actual) have been anchored:

If both	d.f. = $(n_{\text{rows}}-1)^*(n_{\text{columns}}-1)$
If rows only	d.f. = $n_{\text{rows}}{}^*(n_{\text{columns}}\text{-}1)$
If columns only	d.f. = $(n_{\text{rows}}-1)^*\ n_{\text{columns}}$
If neither	d.f. = $n_{\text{rows}}{}^*\ n_{\text{columns}}-1$

Most academic works on the use of the chi-square statistic limit themselves to the 'both' formula, since it is the most common case. The following provide various examples of different tests done, using the historical and recent data in Table 8.8:

(a) Differences between historical and current odds (illustrated): estimates are the recent row totals, apportioned according to the historical good/bad odds. At a χ^2 value of 22.9 with d.f. = 4 = 4 * (2 – 1), the differences cannot be considered random at any probability level—χ^2_{critical} is already 13.3 at a *p*-value of 1 per cent.
(b) Differences between historical and current good/bad distributions: estimates are the recent total (14,190), apportioned by the historical cell percentages. The resulting χ^2 value is a massive 495.1, indicating that the variations are not random, no matter what the degrees of freedom (which in this instance is 7 = (4 * 2 – 1)).
(c) Characteristic's current predictive power: estimates are the current row totals, apportioned by average good/bad odds (8.5). In this case, χ^2 is 5.5 and d.f. = 3 = (4 – 1)*(2 – 1). This time there is some question about the result, since there is still at least a 10 per cent chance that these differences are random.

While this provides an understanding of the mechanics, software packages can often calculate the *p*-value, given a specified confidence level. Care must still be taken to ensure that the degrees of freedom used are appropriate for the problem.

There are two particular instances where the chi-square statistic can be used as part of the scorecard development process: *coarse classing* and *variable selection*. In both cases, the calculation is the same as '(c)' above. It can also be used to audit whether a model's scores adequately represent its characteristics' predictive power, which would use approach '(b)'

Table 8.8. Chi-square—good/bad odds test

Group	Historical			Current			Expected			Chi-square	
	Good	Bad	Odds	Good	Bad	Odds	Good	Bad	Odds	Good	Bad
A	2,313	252	9.2	1,805	220	8.2	1,826	199	9.2	0.25	2.01
B	3,773	443	8.5	4,950	592	8.4	4,960	582	8.5	0.02	0.16
C	2,565	304	8.4	2,888	303	9.5	2,853	338	8.4	0.43	4.07
D	2,999	295	10.2	3,051	381	8.0	3,125	307	10.2	1.78	14.23
Totals	11,650	1,294	9.0	12,694	1,496	8.5	12,763	1,427	8.9	2.47	20.48
	12,944			14,190			14,190			22.95	

above.[6] Scores are calibrated onto probability estimates, which are then used to 'fuzzy parcel' (see Section 19.3) each record into good and bad portions (irrespective of what the actual status is) that are then used as expected values.[7] A chi-square test is then done to compare actual versus expected for each characteristic, and low p-values may reflect a problem.

8.6 Accuracy tests

After one has tortured the data until it confesses, get a clean set of data and see if the confession was valid.

cretog8 (a.k.a. *John Ashcroft*), www.everything2.com

Most of the tests that are used in credit scoring test the models' power, or ranking ability. This is entirely appropriate when scorecards are used only for ranking, but they are being used increasingly to provide default probability estimates, whether for pricing, forecasting, or capital allocation purposes. There are also accuracy tests available, which can be applied: (i) in-sample, out-of-sample, or out-of-time; (ii) to the raw or calibrated scores; or (iii) to the entire risk spectrum, or a part thereof. The types of tests covered here are:

(i) **Binomial test**—Used to compare observed and estimated success rates for a single group.
(ii) **Hosmer–Lemeshow test**—Based upon the binomial test, but is applied across groups.
(iii) **Log-likelihood**—Provides measures of both power and accuracy for ungrouped cases.

Please note that these tests must be used with caution. No matter how accurate the scorecard is, the moment it is implemented the business tries to break it! Customers are punted or shunned according to their deemed risk, with high-risk cases receiving collections priority when problems arise. Also, credit is a dynamic environment, influenced by infrastructural, competitive, and economic changes. Many of these tests assume that the phenomena being observed are independent, whereas in truth the outcomes are correlated—both within the population, and over time.

8.6.1 Probability theory

The basic approach used for testing estimates' accuracy is the binomial test, which is applicable only to dichotomous outcomes—bad versus good, default versus not default, bankrupt versus not bankrupt. In order to understand the mathematics behind it, a brief review of probability theory is required. This is the realm of Bernoulli trials, which have three properties: (i) there are only *two possible outcomes*, success and failure, where the label is arbitrary; (ii) there is the *same success probability* for all trials; and (iii) the *results are random*, and each trial is independent of other trials.

[6] The audit may be done on all cases, or alternatively only on accepts, to ensure that the reject inference has not adversely affected the model's applicability to the accept population.

[7] It is easiest where the model provides a value that either is, or can be easily converted into, a reliable probability.

Jacob (Jacques) Bernoulli (1654–1705)

Jacob Bernoulli originally studied philosophy and Calvinist theology, and it was some dismay to his parents when he turned to mathematics, but he went on to become Professor of Mathematics at the University of Basel, in 1687. He was the first in his family of mathematicians to become famous, before his younger brother Johann (1667–1748), and his nephews Nicolaus (1687–1759) and Daniel (1700–1782). Both Jacob and Johann were renowned in Europe for their contributions to calculus, yet their relationship in later years was acrimonious. Today, Jacob is even more famous for his *Ars Conjectandi* (The Art of Conjecturing). He had formulated most of the ideas between 1684 and 1689, but because of the work's ambitious scope it was never completed, and was published as an *opus posthumus* by Nicolaus, in 1713.

Part I started with a review of Christiaan Huygen's 1657 tract, 'On Rationalisation in Dice Games', which complemented the theory of equity for pricing gambles, presented by Blaise Pascal in 1665 (see box below) and Jan De Witt in 1671. Parts II and III looked at combanitorics and games of chance respectively. Only Part IV was unfinished, which covered the possible application of the theory of equity to probabilities, and probabilities' practical uses in politics, law, and business. Unfortunately, he could not find the data that would provide him with real life examples outside of gambling. Johann was asked by other academic luminaries to finish and publish the work, but was prevented by Jacob's widow and son, who distrusted his intentions. As a result, some of the mathematical ideas were already obsolete by the time it was published. In 1708, Pierre Rémond de Montmort published the first edition of his *Essay d'analyse sur les jeux de hazard*, and later in 1713, the second edition was published with some of the latest ideas from Montmort, and both Johann and Nicolaus Bernoulli.

Ars Conjectandi was the first substantial work on probability theory, and covered the general theory on permutation and combination, the law of large numbers, and the binomial and multinomial theorems.[8] Such works were novel, in an era when mankind and science were searching for deterministic and mechanistic answers for everything, and believed that chance could only exist because of human ignorance. Even Bernoulli's theorem (law of large numbers) states that certainty can be determined with sufficient trials. Probability theory was not of scientific interest, but was used as a means of pricing uncertain future outcomes (economics, gambling, and contracts).

Pascal's original pricing problem related to the apportionment of monies from an unfinished gamble, but the concept was extended to explain reasonable expectations and behaviour. He eventually returned to the church, and in 1669 he published *Pensées*, which contained three wagers, one of which is today known as Pascal's Wager. People's belief in God was supported through purely logical argument—'If you gain, you gain all; if you lose, you lose nothing. Wager, then, without hesitation that He is'.

The first step in understanding probability theory is to understand factorials, which involves repeated multiplication of an incrementing integer, as shown in Equation 8.12. Note that

[8] Wolfram Research, scienceworld.wolfram.com

factorials only work for non-negative integers and, when working with fractions, only the integer portion is used.

Equation 8.12. Factorial $n! = \begin{cases} 1 & \text{if } n = 0 \\ 1 \times \cdots \times n & \text{if } n > 0, n \in 1, 2, 3, \ldots \end{cases}$

Thus, 2! = 2, 3! = 6, 4! = 24, 5! = 120, and so on. The increases are exponential, and beyond a value 170! most spreadsheets fall over. Even so, it is sufficient for many problems. This is then used to calculate the number of possible combinations that can be created from a set of unique items, as shown in Equation 8.13:

Equation 8.13. Number of combinations $${}_nC_k = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Where C is the number of possible combinations, n is the total number of cases, and k is the number selected. Thus, if there are nine unique items in a box, and three are selected, there are

$${}_9C_3 = \binom{9}{3} = 9!/(6!\,3!) = 3{,}62{,}280 \times 720 \times 6 = 84$$

possible combinations. No matter what the value of n, the possible values of ${}_nC_k$ will always be highest at, or immediately around, $n/2$; and will appear normally distributed, but lumpy, for small values of n. The probability of any one particular combination is then the inverse, $1/{}_nC_k$.

8.6.2 Binomial distributions

The number of combinations is only of interest because it is a key component of calculating the binomial distribution, meaning the frequency distribution of successes for any given number of Bernoulli trials, and a success rate estimate. For any given situation, the number of observed successes X will have a binomial distribution, denoted as $B(n, \hat{p})$, where n is the number of trials, and $\hat{p}$ the estimated probability of success. If there are nine items that fall into two categories, with an estimated proportion of 30 per cent for one of them (success), then the counts can range from 0 to 9, with a distribution represented as $B(9, 30\%)$. The probability of observing exactly k successes is then:

Equation 8.14. Binomial probability $$\Pr(X = k) = {}_nC_k \times \hat{p}^k \times (1-\hat{p})^{n-k}$$

Thus, if the expected success rate is 30 per cent, then the probability of exactly 3 successes out of 9 trials is $\Pr(X = 3) = {}_9C_3 \times 0.33 \times 0.7^{9-3} = 84 \times 0.007412 = 26.68\%$. In similar fashion, the probability of having at most three successes, is the cumulative probability for all integers 0 to 3, or $\Pr(X \leq 3) = \sum_{i=0}^{3} \Pr(X = i) = 72.96\%$.

The binomial test is typically associated with medical trials, where researchers are trying to determine if the observed and expected rates are consistent, especially for rare events. The problem is stated as a null and alternative hypothesis, of the form:

H_0—the observed and expected probabilities are the same, $p = \hat{p}$;

H_A—the observed and expected probabilities are not the same, $p \neq \hat{p}$;

For equality, a two-tailed test is used, where both upper and lower bounds are determined. For a confidence level of 99 per cent, critical values at SLs of both 0.005 and 0.995 are required. If the observed value lies outside the resulting range, then the null hypothesis is rejected. A one-tailed test is used to determine whether the observed value is greater or less than an estimate.

The example in Table 8.9 illustrates not only the two-tailed test, but also the HL statistic (covered next). The goal is to determine a range of observed default rates that would be considered acceptable, given previous estimates and a confidence level. To do this, the minimum number of successes that would provide probabilities at the lower and upper bounds for the significance level are calculated, as per Equation 8.15:

Equation 8.15. Critical binomial $k_\alpha = \min(k \mid \Pr(X \leq k) > \alpha)$

where k_α is the critical value for k, and α is the significance level. For the score range 'Low to 302' in Table 8.9, the number of trials was 2,600 with 600 observed successes (defaults), as compared to an estimated 594 at a rate of 22.87 per cent. At a confidence level of 99 per cent, the confidence interval is between $k_{0.5\%} = 540$ and $k_{99.5\%} = 650$,[9] or alternatively, a bad rate of between 20.77 and 25.00 per cent.

Table 8.9. Binomial accuracy tests

Score	Counts			Bad rate (%)		Boundaries			z-Stat	HL-Stat
	Total	Good	Bad	Obs.	Est.	0.5%	99.5%	Critical		
Low–302	2,600	2,000	600	23.08	22.87	20.77	25.00		0.25	0.06
303–348	4,100	3,500	600	14.63	14.85	13.44	16.29		−0.39	0.15
349–380	5,600	5,000	600	10.71	9.30	8.32	10.32	X	3.63	13.20
381–406	9,600	9,000	600	6.25	5.69	5.09	6.31		2.36	5.59
407–430	18,600	18,000	600	3.23	3.43	3.09	3.77		−1.52	2.30
431–454	22,600	22,000	600	2.65	2.05	1.81	2.29	X	6.47	41.91
455–479	50,600	50,000	600	1.19	1.21	1.09	1.34		−0.57	0.32
480–510	80,600	80,000	600	0.74	0.72	0.64	0.80		0.91	0.83
511–555	140,600	140,000	600	0.43	0.42	0.38	0.47		0.20	0.04
556–High	200,600	200,000	600	0.30	0.25	0.22	0.28	X	4.47	19.94
Total	535,500	529,500	6,000	1.12	1.06	1.02	1.09	X		84.35
		Gini coefficient = 51.5%					z-stat critical = 2.58			

[9] These values were calculated using the CRITBINOM function in Microsoft Excel.

The same calculation was repeated for all of the score ranges, and the default rates were found to be significantly different from the estimates for three of them, as well as the overall estimate. The latter is especially disconcerting, as the difference between 1.06 and 1.12 per cent is small. Remember though, that these tests are conservative, and often fail to recognise the dynamic environments in which lenders operate.

Normal approximation to binomial distribution

While the binomial formula provides exact probabilities, it is unfortunately not computationally feasible for larger values. The alternative is to use the normal approximation, which can be used as long as both the expected number of successes and failures are greater than 10 ($10 \leq n\hat{p} \leq (n - 10)$). Its base assumption is that the probabilities are normally distributed. The number of standard deviations that a value lies away from the mean is represented by the z-statistic, which is usually calculated as $z = (X-\mu)/\sigma$, where X is a value that lies somewhere within the distribution, μ is the mean, and σ is the standard deviation.

In this case: (i) X and μ are the observed and expected number of successes respectively, where $X = k$ and $\mu = n\hat{p}$; and (ii) the variance for a binomial distribution is $\sigma^2 = n\hat{p}(1-\hat{p})$. Thus, the z-statistic calculation changes to:

Equation 8.16. Binomial normal approximation $$z = \frac{k-n\hat{p}}{\sqrt{n\hat{p}\,(1-\hat{p})}}$$

As an example, for the score range '349 to 380' in Table 8.9, there are 600 successes out of 5,600 trials, whereas the estimated success rate of 9.3 would have provided 521.0 successes. The z-statistic is:

$$z = (600 - 5600 \times 9.3\%)/\sqrt{5600 \times 9.3\% \times (1-9.3\%)} = (600 - 521.0)/21.7 = 3.63$$

If the expression $\Phi^{-1}(\alpha)$ is used to denote the inverse standard normal cumulative distribution (NORMSINV function in MSExcel), which is used to provide the critical z-statistics given the required significance levels (α), then for a confidence level of 99 per cent, the threshold values of $\Phi^{-1}(0.005)$ and $\Phi^{-1}(99.5\%)$ are -2.58 and 2.58 respectively. In this instance, the hypothesis that the observed and expected success rates are equal must be rejected.

Rather than testing hypotheses, the task is often to calculate specific probabilities. If $\Phi(z)$ is used to denote the standard normal cumulative distribution function (NORMSDIST function in MSExcel), and $z_{n,\hat{p},k}$ the z-statistic (for a number of trials, success probability estimate, and number of successes), then the probabilities are:

Equality	$\Pr(X = k) = \Phi(z_{n,\hat{p},k+0.5}) - \Phi(z_{n,\hat{p},k-0.5})$
Less than	$\Pr(X < k) = \Phi(z_{n,\hat{p},k-0.5})$
Less than or equal to	$\Pr(X \leq k) = \Phi(z_{n,\hat{p},k+0.5})$
Less than	$\Pr(X > k) = 1 - \Phi(z_{n,\hat{p},k+0.5})$
Less than or equal to	$\Pr(X \geq k) = 1 - \Phi(z_{n,\hat{p},k-0.5})$

Note the ±0.5 adjustments to the observed successes, made because it is a continuous distribution. This adjustment should also have been reflected in Equation 8.16, but its omission has little impact where the number of observed successes is large.

For the case where 50 trials are performed with an estimated success rate of 20 per cent, the probability of exactly 5 successes using the normal approximation is: $\Pr(X=5) = \Phi(z_{50,\,20\%,\,4.5}) - \Phi(z_{50,\,20\%,\,4.5}) = \Phi(-1.591) - \Phi(-1.945) = 5.58\% - 2.59\% = 2.99\%$ In contrast, if the exact binomial test, as per Equation 8.14, were used, the result would be 2.95 per cent.

8.6.3 Hosmer–Lemeshow statistic

Both the binomial distribution and its normal approximation focus on individual groups or ranges, but an estimate's accuracy can also be tested across the entire risk range. The most commonly known approach is the Hosmer–Lemeshow statistic shown in Equation 8.17.

Equation 8.17. Hosmer–Lemeshow statistic $$\mathrm{HL} = \sum_{k=1}^{g}\left(n_k \times \frac{(p_k - \hat{p}_k)^2}{\hat{p}_k \times (1-\hat{p}_k)}\right) = \sum_{k=1}^{g} z_k^2$$

where k is an index for each group, and g the total number of groups. Note that all the authors did was sum the squared z-statistics for each of the ranges, much like squaring the error terms, only in this instance it is being used like a goodness of fit measure. The hypothesis is:

H_0—The observed and expected probabilities are not the same;

H_A—The observed and expected probabilities are the same;

The resulting values fit a chi-square distribution, but with a d.f. of g–2.

Although the effect is small, it may be wise to enforce the $10 \leq n\hat{p} \leq (n-10)$ constraint prior to calculating this statistic. For risk grades, those with insufficient numbers are collapsed with their neighbours. For scores, it usually sufficient to split the range into deciles using the number of trials, but it may be problematic for groups where the estimated success rates are very low. The number of successes should instead be used, as has been done in Table 8.9 (p. 215). For that example, the final HL statistic is 84.35, and for the 10 categories the d.f. is 8. For a significance level of 1 per cent, χ^2_{Critical} is 20.09, which leads to the conclusion that the estimated and observed rates are inconsistent even at that lax level. Indeed, the hypothesis could not be accepted at any significance level.

It is worth noting that the HL statistic was originally presented in a textbook on logistic regression, as a means of testing the estimates provided by the resulting models. When used to assess model performance on out-of-time data, it should only be used to test recalibration, as in dynamic environments, it is unlikely that the estimates will be reliable to the levels expected by these tests.

8.6.4 Log-likelihood

Model reliability can be broken down into two dimensions, power and accuracy. Power is more important than accuracy, because the latter can be provided after the fact, through

calibration. As a result, most of the measures focus solely upon model power. Irrespective, there may be times when lenders wish to compare models in terms of both. One means of doing this is to use a log-likelihood measure.

These measures are usually used in hypothesis testing, in much the same way as chi-square tests, except the focus is to determine the distribution that best fits a known distribution. It also forms the basis for maximum likelihood estimation (MLE). In the current instance, however, the error term needs to be represented in a manner similar to the chi-square statistic. The result shows how far away an expected distribution is from the actual, but has the added advantage that it can work for individual cases—not just a classed distribution. It is shown in Equation 8.18, in a form that can be applied to both positive and negative test results.

Equation 8.18. Total log-likelihood

$$\text{TLL} = \sum_{i=1}^{T} \begin{cases} P_i \times \ln(P_i/\hat{P}_i) & | \; P_i \neq 0, \hat{P}_i \neq 0 \\ N_i \times \ln(N/\hat{N}_i) & | \; N_i \neq 0, \hat{N}_i \neq 0 \end{cases}$$

where P is a positive indicator (0 or 1), N a negative indicator $(1-P)$, $\hat{P}$ and $\hat{N}$ are probability estimates, T is the total number of cases being evaluated, and i is the index for the case being evaluated. The results can then be converted into a likelihood figure as:

Equation 8.19. Likelihood

$$L = \exp\left(\frac{\text{TLL}}{T/2}\right)$$

The final result is an error term, which indicates the likelihood that the probability estimates are NOT a proper representation of the values; the lower the value, the better the fit. Unfortunately, it provides no indication of whether the error comes from problems with power or accuracy. In order to split it out, likelihoods are calculated for two naïve models; one using the estimates, and another using the actuals. For naïve models, the total log-likelihood formula in Equation 8.18 simplifies to:

Equation 8.20. Naïve TLL

$$\text{TLL}_{\text{Naïve}} = P \times \ln\left(\frac{T}{P}\right) + N \times \ln\left(\frac{T}{N}\right)$$

where P, N, and T are the total number of positives, negatives, and records respectively. The value of T will be the same for each naïve model, but P and N will almost always be different for the expected and observed totals.

Equation 8.21. Accuracy

$$\text{Accuracy} = 1 - \frac{(L_{\bar{E}} - L_{\bar{O}})}{L_{\bar{E}}}$$

where: $L_{\bar{E}}$ is the likelihood for a naïve model using estimates, $L_{\bar{O}}$ is the same, but for observed values. According to this formula, if the likelihood figures for the two naïve models happen to be equal, the accuracy is 100 per cent, irrespective of the model's power. Note here that for most models this figure should be very high, and the minimum required threshold may be 95 per cent or higher. The counterpoint to this is power, which can be calculated as:

Equation 8.22. Power

$$\text{Power} = \frac{L_{\bar{E}} - L_E}{L_{\bar{E}} - 1}$$

Table 8.10. Log likelihood

	Actuals		Estimates		Log Lik
#	P	N	P	N	
1	1	0	0.90	0.10	0.105
2	0	1	0.10	0.90	0.105
3	1	0	0.80	0.20	0.223
4	1	0	0.70	0.30	0.357
5	0	1	0.50	0.50	0.693
6	0	1	0.40	0.60	0.511
7	1	0	0.70	0.30	0.357
8	0	1	0.20	0.80	0.223
9	1	0	0.80	0.20	0.223
Total	5	4	5.10	3.90	2.797
Model		TLL	LL	Lik	Power
Test		2.797	0.622	1.862	70.8%
Naïve		6.185	1.374	3.953	99.9%
Naïve		6.183	1.374	3.951	Accuracy

where $L_{\bar{E}}$ is the likelihood for a naïve model using estimates, and L_E the likelihood for the model being tested.

In this case, power will approach 100 per cent as L_E approaches 1, which represents a perfect model, and will also provide accuracy of 100 per cent. At the other end, power will approach 0 per cent if the same estimate has been applied to every case—as happens with any naïve estimate—and will be negative, where the estimates are tending towards randomness. Please note that this makes no reference to the actual rankings, and as such is not a rank-order correlation coefficient.

An example of this calculation is provided in Table 8.10, for a small group of nine cases. The total log-likelihood for the sample is 2.797, which provides a log-likelihood of 0.622 and likelihood of 1.862. In contrast, the likelihood calculated using the observed average, provides a likelihood of 3.953, and using the estimated average, provides 3.951. The accuracy is 99.9 per cent, which anybody would be comfortable with. In contrast, the power is 70.8 per cent, which means that the model being tested explains 70.8 per cent of what its naïve equivalent cannot. Please note that this approach should be used with caution. While it can be used to compare models, no confidence intervals can be provided for use in hypothesis testing.

8.7 Summary

While Chapter 7 (Predictive Statistics 101) focused on the predictive-modelling techniques used to develop credit scoring models, this chapter has moved on to the mathematical

formulae used both as part of the scorecard development process, and to assess the results. Lenders' primary interests are in prediction (high power) and stability (low drift), and as a result there are several power/drift statistics that focus on these aspects. Power measures are used throughout the model development process, including coarse classing, variable selection, segmentation, and final result evaluation. The results can, however, be affected by real or apparent homogeneity, whether it is: (i) an inherent feature of the population; (ii) the result of data deficiencies; (iii) selection process truncation; or (iv) the product of segmentation. In contrast, drift (or divergence) measures are used primarily for post-development validation and post-implementation monitoring, albeit they could also be calculated against a recent sample and used as part of the scorecard development. A further aspect is accuracy, to ensure that the overall probabilities are more or less in line with those expected.

The measures used to assess power and drift are not mutually exclusive, and many can be used for both. The tools presented were: (i) *misclassification matrix*, the most basic tool, a 2×2 contingency table detailing true/false and positive/negative for both predicted and actual, which also provides the 'per cent correctly classified'; (ii) *Kullback divergence measure*, which is based upon the weight of evidence, and used to calculate the information value (power) and stability index (drift); (iii) *Kolmogorov–Smirnov curve* and *statistic*, the former displays two ECDFs in a 'fish-eye graph', and the latter provides the maximum percentage difference; (iv) *correlation coefficients*, including Pearson's product-moment, Spearman's rank-order, Gini coefficient (area between Lorenz curve and diagonal), and AUROC (area under receiver operating characteristic); (v) the *chi-square test*, used to examine contingency tables, where the number of cells affects the d.f.; and (vi) *accuracy tests*, including the chi-square test, binomial test (and its normal approximation), Hosmer–Lemeshow test, and log-likelihood.

Exactly how these measures are used is covered in Module E (Scorecard Development Process). A brief summary can be provided here though (Table 8.11 provides a high-level overview, but is for guidance purposes only, and should not be interpreted strictly). When doing *coarse classing* (binning characteristics in an optimal fashion), *variable selection* (choice of those that will provide value in a predictive model), *segmentation* (determine whether and which separate models are required), and *final performance assessment* (to rate the models' predictiveness out-of-sample and/or out-of-time) the goal is to extract maximum power. The most commonly used measures are: (i) for *predictors*, the information value and chi-square statistic; and (ii) for *scores*, the AUROC, Gini coefficient and KS statistic. Drift assessments, whether of characteristics or the final score, rely mostly on the stability index and chi-square statistic. The latter's advantage is that there are specific confidence thresholds, which do not exist for the stability index. When assessing the stability of the final score, both the stability index and KS statistic may be used, and there are guidelines for each.

The Gini coefficient—also called an accuracy ratio, Somer's D, or power statistic—is widely referred to, and often suggested for broader use, but assumes a rank ordering. As a result, it cannot be used to assess non-monotonic (especially categorical) characteristics, which applies to many of the characteristics used in retail credit. It is also not possible to use it for hypothesis testing. Its primary use is to assess the rank ordering of the final score or grade. The KS statistic is also commonly used for that purpose, and can be used for statistical tests, but unfortunately focuses upon a single point in the score range. In either case, care must be taken, because too heavy a focus upon a specific measure may lead to an overfitted model, and poor

Table 8.11. Use of statistical measures

	Predictive power		Stability	
	Predictors	Scores	Predictors	Scores
Chi-square	✓		✓	
Kullback divergence	✓		✓	✓
AUROC/Gini coefficient		✓		
KS statistic		✓	✓	

out-of-sample performance. Hence it is wise to use more than one measure, and perhaps also data-visualisation tools, like the misclassification graph or strategy curve.

Other measures were covered, but they tend to have more specific uses. First, *Spearman's rank-order correlation* is used primarily to assess the differences between different scores or grades calculated, or available, for the same set of cases (often for benchmarking internal versus external grades). The *chi-square test* is used to assess both power and drift, through a comparison of contingency tables, and is heavily influenced by the number of classes (d.f.). The *binomial test* is used to assess predictive accuracy for a single group, with a binary outcome. An extension of its normal approximation is the *Hosmer–Lemeshow statistic*, which can be used to assess the full model. And finally, the *log-likelihood* calculation forms the basis of MLE and logistic regression, but can also be used to assess both power and accuracy at the same time. Unfortunately, it is not possible to use it for hypothesis testing.