

ARE THE PROBABILITIES RIGHT?

A First Approximation to the Lower Bound on the Number of Observations Required to Test for Default Rate Accuracy

TECHNICAL **REPORT** #030124

ABSTRACT

AUTHOR

Roger M. Stein

Researchers and practitioners have begun to investigate and adopt credit default models for practical applications. As a result, the issue of the accuracy of probability estimates has naturally arisen. Specifically, users of a default model that produces estimates of probabilities of default desire to know the accuracy of the probabilities produced by the model. There are a number of mechanisms for doing this, but one that has found favor due to its intuitive appeal is the estimation of goodness of fit between expected (under a particular hypothesis) and predicated default rates. While most experimenters readily acknowledge that large data sets are required to test these estimates, particularly when probabilities are small as in the case of higher credit quality borrowers, the question of *how large* often arises. In this short note we demonstrate, based on simple statistical relationships, how a lower bound on the size of a sample may be calculated for such experiments. It is a lower bound as it assumes no positive correlation among the data either in time or cross-sectionally, when in practice both of these assumptions are typically violated. In the presence of correlation, the bound can change considerably and we show this. That said, the bound is useful in that it can be helpful in determining when the data at hand are not sufficient to draw rigorous conclusions about the probability estimates of a model. In addition, where an experimenter has a fixed sample size, this approach provides a means for sizing the minimum difference between an estimated and an empirical default rate that should be observed in order not to conclude that the hypothesized and observed rates are statistically indistinguishable. We also discuss some of the circumstances under which the lower bound may be misleading.

© 2003 Moody's KMV Company. All rights reserved. Credit Monitor®, EDFCalc®, Private Firm Model®, KMV®, CreditEdge, Portfolio Manager, Portfolio Preprocessor, GCorr, DealAnalyzer, CreditMark, the KMV logo, Moody's RiskCalc, Moody's Financial Analyst, Moody's Risk Advisor, LossCalc, Expected Default Frequency, and EDF are trademarks of MIS Quality Management Corp.

ACKNOWLEDGEMENTS

I am grateful to Jeff Bohn, David Bren, Ahmet Kocagil, Matt Kurbat and Bill Morokoff of Moody's KMV, to Richard Cantor of Moody's Investors Service and to William Greene of NYU for their generous comments. All remaining errors are, of course, my own.

Published by:
Moody's KMV Company

To Learn More
Please contact your Moody's KMV client representative, visit us online at www.moodyskmv.com, contact Moody's KMV via e-mail at info@mkmv.com, or call us at:

NORTH AND SOUTH AMERICA, NEW ZEALAND AND AUSTRALIA, CALL:
1 866 321 MKMV (6568) or 415 296 9669

EUROPE, THE MIDDLE EAST, AFRICA AND INDIA, CALL:
44 20 7778 7400

FROM ASIA CALL:
813 3218 116

1 INTRODUCTION

It is common for users of credit models to assign probabilities of default to specific credit grades. Users of these models often wish to know the accuracy of these probabilities. It is common for researchers to run experiments in which they attempt to estimate the goodness of fit between expected (under a hypothesis) and predicated default rates.

There are a number of mechanisms for doing this. One of the more common is for a researcher to examine whether the observed default rate for borrowers of a certain credit grade is within the expected range for that credit grade. For example, a firm that were using a model to produce probability of default predictions might wish to know whether the predicted default rate in the "Pass - 2" rating category were correct. The firm might test this by examining all borrowers that were graded "Pass - 2" by the model over a period of time. By calculating the actual number of defaults observed and comparing this with the predicted or target number for "Pass - 2" borrowers, the firm could try to assess the accuracy of the model.

While most researchers readily acknowledge that in many cases large data sets are required to perform such tests, particularly when probabilities are small as in the case of higher credit quality borrowers, the question of *how large* often arises. In this short note we review some of the statistical machinery that can be used to help determine the number of records required to perform such tests.

The approach is based on elementary statistical relationships. We characterize it as a lower bound since the calculations shown assume no positive correlation among the data either in time or cross-sectionally, when in practice both of these assumptions are typically violated. We also show that effects of more realistic correlation structures can impact the calculation of such bounds significantly and we discuss this later in the paper. For example, in the case of zero correlation, with sufficient data, the bound can be made arbitrarily small. However, when correlation is non-zero, adding more observations does not necessarily produce a narrower confidence bound.

Nonetheless, the analytic bound discussed here is useful in that it can be used to determine when the data at hand are *not* sufficient to draw rigorous conclusions about the probability estimates of a model. Conversely, where an experimenter has a fixed sample size, this approach can be used to size the minimum difference between an estimated and an empirical default rate that must be observed in order to not to conclude the rates are statistically indistinguishable. Furthermore, in settings where the underlying assumptions are violated, simulation methods are available in many cases and we suggest approaches for this context.

The remainder of this note proceeds as follows: In Section 2, we discuss the mathematical tools we use to calculate these bounds. In Section 3 we provide an example of how the approach can be applied. Section 5 discusses the limitations of the approach due to the simplifying assumption of no correlation in the data and presents a simulation example that demonstrates this. Appendices present tables of bounds for various quantities at different levels of precision and a simple simulation algorithm for use in cases where the analytic approximation is unreliable.

2 MATHEMATICAL PRELIMINARIES

The mathematical machinery necessary for answering the questions set forth in this paper is well established in the form of the Law of Large Numbers and the Central Limit Theorem and can be found in most textbooks on probability (c.f., (Papoulis 1991; Grinstead and Snell 1997)).

In short, we take advantage of the limiting properties of the binomial distribution and assume it approaches normal distribution in the limit as the number of observations gets large. We can then formulate a standard hypothesis test, and with some algebra, solve for the sample size required to be sure that any difference between the true mean and the sample estimate will be small, with some level of confidence.

We start with an assumed predicted default probability, p , perhaps produced by a model, a rating system or as the target of some underwriting practice. We also have a data set containing n firms, d of which have defaulted. We wish to determine whether the assumed default rate is the reasonably close to the true default rate.

We can define the empirical frequency of defaults (the default rate) as:

$$f_d = d / n$$

We would like to be certain that

$$P(|f_d - p| < \varepsilon) \leq \alpha \tag{1}$$

where α is a significance level. For example, we might wish to be sure that the difference between the true default rate and our predicted default rate were less than 20 basis points.

In the case of defaults, we often make the assumption that the underlying distribution is binomial¹. If we do so we can appeal to the Central Limit Theorem (CLT) and obtain a convenient (Gaussian) limit which facilitates calculations. Using the CLT we get the familiar result:

$$P(np_L \leq np \leq np_U) \cong \frac{1}{\sqrt{2\pi}} \int_{\frac{n(p_L - p)}{\sqrt{npq}}}^{\frac{n(p_U - p)}{\sqrt{npq}}} e^{-x^2/2} dx = \Phi\left(\frac{n(p_U - p)}{\sqrt{npq}}\right) - \Phi\left(\frac{n(p_L - p)}{\sqrt{npq}}\right)$$

where $\Phi(\cdot)$ is the standard cumulative normal distribution. Since here we are assuming that $p_U - p = p - p_L = \varepsilon$, this simplifies to

$$2\Phi\left(\frac{n\varepsilon}{\sqrt{npq}}\right) - 1 \geq 1 - \alpha$$

or, more conveniently,

$$\Phi\left(\frac{n\varepsilon}{\sqrt{npq}}\right) - 1 \geq 1 - \alpha / 2$$

yielding the following inequality:

$$\frac{n\varepsilon}{\sqrt{npq}} \geq \Phi^{-1}(1 - \alpha / 2)$$

Rearranging terms gives:

$$n \geq \frac{pq}{\varepsilon^2} \left[\Phi^{-1}(1 - \alpha / 2) \right]^2 \tag{2}$$

Equation (2) gives the minimum required number of firms n , in the case that we wish to be certain that we will have enough firms to determine whether a probability p is accurate to within ε at the α level.

Conversely, given that we have n firms, we can ask, how big a difference between p and f_d do we need to observe in order to conclude at the α level that the assumed probability and the observed default rate differ. Rearranging terms, we get:

$$\varepsilon \geq \sqrt{\frac{pq}{n}} \Phi^{-1}(1 - \alpha / 2) \tag{3}$$

¹ We might instead assume the distribution to be hypergeometric, but for expository purposes, the binomial assumption will work about as well and will be more familiar to most readers.

So far we have assumed that we had a predicted probability and wished to determine how well it matched an empirical default rate². If, on the other hand, we were unsure of the true default rate and wished to estimate it, how many firms would we need? We can calculate this by observing that the quantity pq is maximized when $p=q=0.5$. Setting $p=0.5$ ensures that irrespective of what the true default rate is, we can measure it within ε with $100*(1-\alpha)\%$ confidence as long as we have at least n firms. Similarly, for a fixed number of firms, the estimate of the default rate that we obtain will be within $\varepsilon 100*(1-\alpha)\%$ of the time. This is just the standard confidence bound for a probability estimate.

To the extent that the sample size used is very much smaller than the actual population (on the order of 5-10% of the full population), (2) and (3) will generally suffice. However, to the extent that the sample is much larger relative to the overall population, it is often advisable to make a *finite population correction* (fpc) (c.f., Cochran 1977). The fcp serves to adjust for the heightened variance observed in the sample of empirical defaults. The recommended adjustment to the variance is $(N-n)/(N-1)$, where n is the sample size and the population size is N .

3 AN EXAMPLE

Assume that there is a firm using a rating system that produces scores in a number of pre-defined rating buckets and that it has assigned a probability of default to each bucket. The firm would like to know whether the predicted default rate in the "Pass - 2" rating category is correct. Assume that firms in this category, according to the model, have a default rate in the range of 25-75 bps. How many companies would the firm need in this category to determine if the empirical default rate was within this range?³

To evaluate this, we use (2), setting $p=0.0050$ (the mid-point between 25 and 75 bps) and $\varepsilon=0.0025$ (since $50 \text{ bps} \pm 25 \text{ bps} = [25 \text{ bps}, 75 \text{ bps}]$, the range of the "Pass-2" category. This yields an estimated sample size n of 3,058 to achieve 95% confidence level or 5,281 to achieve a 99% confidence level.

If the firm only had 1000 firms available to test on, it could try to determine how big the ε would have to be (i.e., how big a difference from 0.005 would the observed default rate, f_d , have to be to indicate that the model's predictions were incorrect). To do so, the firm would use (3) and find that ε would be 0.0044 and 0.0057, for the 95% and 99% confidence bounds. In other words, given the number of firms, it would not be possible to conclude that the bucket probability was incorrect unless it was well outside the 25 - 75 bps range. Since $\varepsilon = 44 \text{ bps}$ in this case, any empirical default rate observed in the range of 6 bps - 94 bps would still not provide evidence that the probabilities were misspecified for the bucket. gives some additional levels of ε for different sample sizes at the 95% confidence level.

TABLE 1 Required levels of ε for various sample sizes when $p = 0.005$ and $\alpha = 0.05$.

n	ε
1,000	0.0044
2,500	0.0028
5,000	0.0020
10,000	0.0014

² Importantly, an implicit assumption is that our dataset accurately captures all relevant defaults. To the extent that this is not true (i.e., there are hidden defaults) the observed empirical probability may not reflect the true frequency of defaults. See: Dwyer, D. and R. M. Stein (2003). Inferring the Default Rate in a Population by Comparing Two Incomplete Default Databases, Technical Report #021216, Moody's KMV, New York, for a discussion.

³ Here it is important to note that we are assuming that the distribution of true probabilities is distributed along the range of the upper and lower bounds of the bucket. An alternative assumption would be that the true probability is located somewhere in the range but that it could be represented as a single point (rather than a distribution). This latter assumption would require slightly different treatment. We chose the former as it seems more consistent with the formulation of the empirical problem.

4 REGIONS OF BREAKDOWN FOR THE ANALYTIC RESULTS

It is also useful to consider that when p and/or n is very small, the limiting properties on which this analysis relies may not be present. In such cases, it is not prudent to use the analytic results. Rather, simulation provides a less convenient but more reliable mechanism to determine more appropriate values of the quantities of interest. In Appendix B, we provide one such algorithm for doing so.

Table 2 gives examples of the simulated and analytic results for several selected cases.

TABLE 2 Analytic vs. simulated levels of ε for various sample sizes when $\alpha = 0.05$.

n	p	Analytic ε	Simulated ε	Percent Difference
100	0.001	0.0062	0.0090	45%
250	0.001	0.0039	0.0030	-23%
500	0.001	0.0028	0.0030	8%
1,000	0.001	0.0020	0.0030	2%
50	0.025	0.0433	0.0350	-19%
100	0.025	0.0306	0.0250	-18%
250	0.025	0.0194	0.0190	-2%
500	0.025	0.0137	0.0130	-5%

From the table, it is clear that the analytic result provides a reasonable estimate in cases where n or p are fairly large, or, more appropriately, not very small. However, the relative difference in predicted values ($[\varepsilon_{\text{simulated}} - \varepsilon_{\text{analytic}}] / \varepsilon_{\text{simulated}}$) can be quite large in cases where the values are too small. For example, even for moderately high default probabilities (e.g., $p=2.5\%$), the difference between the analytic approximation to the error bound ε and the simulation result is almost 1% in default probability (83 bps) for small samples ($n=50$).

This result is generally consistent with a common heuristic which recommends avoiding the approximation unless the quantity npq is a good deal larger than 2. These are also often the cases in which the distribution can become significantly skewed which complicates the interpretation of the results⁴. From our informal experiments, we recommend using simulation in cases where npq is less than about 4. In the experiments, this resulted in relative errors of less than about 10% when estimating ε .

Unfortunately, there does not appear to be a simple (not computationally intensive) mechanism for estimating n through simulation. However, as it is typical for analysts to have a fixed sample of firms (rather than a fixed value of ε), this does not pose as great a practical problem as it might. That said, we can, use the methodology above to test specific values of n , estimated using (2), to determine the degree to which they produce expected values of ε .

Though not reported here, we have tested several of the values of n generated by the analytic estimates. We found that in general, the approximations worked well except for very small values. Excluding these extremes, errors tended to be below 10%.

⁴The skewness of the binomial distribution is given, after simplification as: $\frac{1-6pq}{npq}$. For theoretical binomial distributions the skewness becomes significant just below $p=1\%$ and $p=2\%$ for $n=500$ and $n=200$, respectively, using a variant of Fisher's test.

5 COMPLICATIONS AND CAVEATS: WHY A LOWER BOUND?

We have described these measures as lower bounds. The statistical theory, however, provides that under the assumptions of the CLT, the limiting values should be *upper* bounds, given assumptions of independence and appropriate values of n , p , and ε .

In practice, however, it is rare that databases of corporate obligors and defaults strictly meet the assumptions of the CLT, as we have presented it, particularly as they relate to the independence of the observations. This is due largely to additional sources of variance and covariance, as well as skewness and other higher moments that may act to increase the estimates of n and ε .

5.1 Additional Sources of Variance

The first of these relates to the fact that the probability estimates produced by a model are typically not actual probabilities but rather probability estimates. For example, an EDFTM is actually a statistical estimate of the conditional probability of default given a particular value of a Distance-to-Default, a measure of credit quality⁵. As estimates, such measures are subject to sampling variation, which is not explicitly captured in the formulation given here.

The approach suggested in the example above is a special case that does minimize to some degree the impact of this estimation variability since the actual value of p that is being tested is in the example not an estimate. Rather it is the midpoint of the range of possible values of p associated with a particular rating class. We have not assumed that the classification of obligors into that class is done using any particular method so it is not necessary that the probabilities associated with the class be statistical estimates. They could, for example, be based on a particular underwriting target default rate. Unfortunately, this is typically a special case.

In practice, some banks use a probability model to assign borrowers to rating grades. In this situation, this extra variability would likely increase the overall variance of the estimates of n and ε . Furthermore, if the estimate for the target default rate within the bucket were itself an estimate (say an mean or median of the EDFs for borrowers within a particular rating class), the variance of this estimate would almost certainly affect the variance which (assuming zero correlation) in turn will increase both the estimated values of ε and n . Thus, even if we assume no correlation, the estimates produced by (2) and (3) as presented above are likely to be lower bounds for the maximum number of firms required.

5.2 Correlation Among Data

The estimates become further understated in the presence of correlation among the data. The analysis above assumes independent (i.i.d.) observations. However in the case of correlated observations, (e.g., if the firms in the sample are counted for multiple years, if the the firms are affected by similar economic factors, etc.) this assumption does not hold. This is because the financial statements (in the first case) and the credit quality of the firms (in the second case) may be correlated from observation to observation.

This is often the case when testing is done using "firm years" rather than unique firms. A firm year is an observation of a single firm for a single year. Thus if we had twenty years of data for XYZ Co. and we used all twenty observations, we would have twenty firm years (but only one firm). So in this case although we may count n firms, we are actually observing at a smaller number of independent firms due to the correlation.

⁵ An EDFTM is an Expected Default Frequency. This is an empirical estimate of the probability of default produced by Moody's KMV. The EDF is derived from a proprietary version of a structural model that calculates the Distance to Default for individual firms based on information in equity prices and the firm's capital structure. The Distance to Default is then mapped to an empirical default probability through calibration to an empirical default database.

To see the impact of this, consider the following toy example. Say we observe a variable that is *known to be normally distributed* and we observe this variable on same set of firms for two years but treat this as a single dataset. If we knew that the correlation among firms in the dataset were on the order of ρ , the adjusted variance ($\tilde{\sigma}^2$) would be given as:

$$\tilde{\sigma}^2 = \frac{\sigma^2}{1-\rho^2} \geq \sigma^2$$

where σ^2 is the variance assuming that the data are i.i.d.

Given this result, it is tempting to conclude that we should adjust the variance of our estimates by a factor of $(1-\rho^2)$ and then substitute this into (2).

However, this would be incorrect since the binomial distribution only approaches a normal distribution in the limit *when the observations are independent*. On the other hand, there is no analytic solution to this problem in general when the observations are not independent. Furthermore, in the non-zero correlation case, we have no guarantee that the bound goes to zero as the number of observations gets large.

To explore the impact of correlation, we performed a second set of simulations. This time, we assumed that there was a hidden factor (e.g., asset value⁶) that generated the probabilities of default for each firm and that this variable followed a Gaussian distribution. We then estimated ε assuming different levels of (constant) correlation among the companies⁷. We present the results in Table 3 and Table 4.

TABLE 3 Required 5% significance levels of ε when defaults are correlated to various degrees

Correlation	n	$p = 1\%$	$p = 3\%$	$p = 5\%$
0.0	500	0.008	0.011	0.018
0.1	500	0.020	0.048	0.070
0.2	500	0.030	0.063	0.108
0.3	500	0.036	0.083	0.142
0.1	1,000	0.006	0.008	0.011
0.2	1,000	0.020	0.046	0.067
0.3	1,000	0.029	0.061	0.102
0.4	1,000	0.034	0.081	0.135

⁶ Note that correlation among asset values is not equivalent to correlation among defaults, although the two are related. See: Gersback, H. and A. Lipponer (2000). The Correlation Effect. Heidelberg, Alfred-Weber-Institut, University of Heidelberg. for a discussion.

⁷ I am grateful to Bill Morokoff for suggestions that greatly improved this section of the paper.

TABLE 4 Required 5% significance levels of ε for various sample sizes when defaults are correlated ($corr = 0.03, p = 0.01$)

n	Correlation = 0.03	Correlation = 0.0	Analytic (Correlation = 0.0)
25	0.070	0.030	0.039
50	0.050	0.030	0.028
100	0.040	0.020	0.020
250	0.038	0.010	0.012
500	0.036	0.008	0.009
1,000	0.034	0.006	0.006
5,000	0.034	0.002	0.003

In Table 3, as expected, we see that the estimate of ε with zero correlation is significantly smaller than in the cases in which the correlation is non-zero. We observe that ε increases with the degree of positive correlation. For example, in Table 4, we see that the estimate of a 95% confidence level for ε using 1000 firms with a probability of default of 1% and no correlation is about 60 basis points. In contrast, ε turns out to be about six times greater, 3.4%, when the correlation is 0.3. In Table 4 we also see that the reduction in ε is extremely small even as the sample size increases. It turns out that for very large samples in the special case of constant correlation we can sometimes estimate the limit of this value analytically, but that in general it is not possible to do so and we must resort to simulation.'

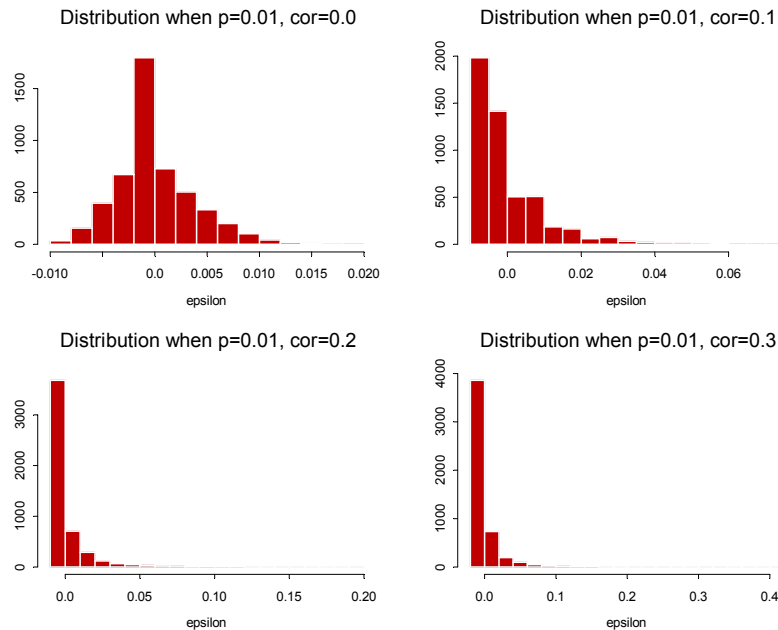


FIGURE 1 Distribution of ε at various levels of correlation when $p = 0.01$

This figure shows the distribution of ε when we assume constant correlation at various levels and a probability of default. We chose values of p at 0, 0.1, 0.2 and 0.3. Note that as the correlation increases, the distribution becomes increasingly skewed. As a result we observe two effects. First the values of ε increase significantly and secondly, as a result of the loss of symmetry, these values become more difficult to interpret.

However, it is important to note that as the correlation increases the distributions of ε resulting from the simulations becomes increasingly skewed. For example, the skewness of the zero correlation case is moderately low at about 0.48 for this set of simulations. In contrast, the skewness of the distributions for the cases of $\rho=0.1$ and $\rho=0.3$ are 2.2 and 6.3, respectively. As a result of this skewness, we observe two effects. First the values of ε increase considerably with the correlation as the right tails of the distributions lengthen and produce more extreme values. Secondly, as a result of the loss of symmetry, the values of ε become more difficult to interpret since they are mostly generated at the tail on the right side of the mean of the distribution.

Also note that even in the case of zero correlation, there is evidence in Table 4 that the distributions become quite skewed when n is small, thus making the symmetric interpretation of ε more involved. In this case it is not until n reaches about 500 that either the theoretical or simulated estimates of ε get smaller than p itself (see Footnote 5, above). Since negative values of p are not feasible, this implies that the distribution must be skewed and thus the largest ε are being generated on the right hand side of the distribution.

In practice it is sometimes extremely difficult to estimate ρ , even when we might know something about the asset correlations. This is particularly true in overlapping samples such as those typically used in testing default rates. In such samples, many of the firms will be present in the data set for more than one year and thus the sample overlaps to varying degrees from one year to the next. Covariance estimators for such situations have been developed for the study of census data but are less well known in the credit literature⁸.

In addition, default rates may not be constant in time. If the default rates themselves are shifting, the analysis can become more protracted. (Cantor and Falkenstein 2001), for example explore the case in which the correlation structure is more involved and the estimator of the variance becomes far more elaborate. These authors also show that failing to account for various correlation effects leads to significant underestimation of default rate variance.

Furthermore, because default rates are typically non-stationary, that is the level of defaults varies from year to year, this can manifest itself in the form of so-called "default clustering" which can also confound the calculation of sample size. In a simulation study, (Kurbat and Korablev 2002) make this point dramatically in a series of experiments in which they vary the levels of correlation among the defaults at various risk levels. As we have also shown here, the authors' results show that the level of correlation among defaults can considerably affect the skewness of the resulting default distributions.

In short there are a number of reasons to believe that estimates based on (2) and (3) will understate the actual required levels of the quantities ε and n in realistic settings. As a result, we feel that these formulae should be used as a hurdle test for examining default rate experiments.

While values that exceed the levels suggested by these formulae are necessary for experiments to have significance, they are almost certainly not sufficient. Thus, while they can be used to disqualify results in cases where an experiment produces an effect that is smaller than ε or that uses a sample smaller than n , they typically cannot be used to determine that a result is significant.

⁸ The effects of overlap correlation effects can be significant and special methodologies have been developed to address the problem of "births and deaths" in the populations (e.g., new firms and firms that are delisted, dropped from the database, etc.). This phenomenon requires special attention as we observe that it is uncommon for the composition of the firms to be identical from year to year, rather some percentage of the sample overlaps prior years from year to year with the addition of new firms and the loss of others.

6 CONCLUSION

Users of default models that produce estimates of probabilities of default frequently desire to know the accuracy of the probabilities produced by the model. It is common for researchers therefore to run experiments in which they attempt to estimate the goodness of fit between expected (under a model) and observed default rates. In this short note we have reviewed some of the statistical machinery that can be used to help determine a lower bound on the number of records required to perform such tests.

The approach is based on simple statistical relationships. It is a lower bound as it assumes no positive correlation among the data in either in time or cross-sectionally, when in most realistic situations both of these assumptions are typically violated. We provided a simple example of how even mild correlation among the firms in a sample can increase the required size of the effect that (and conversely the number of firms that must be observed).

That said, we feel that the bound is useful since it can be used to determine when the data at hand are not sufficient to draw rigorous conclusions about the probability estimates of a model. In addition, where an experimenter has a fixed sample size, this approach can be used to size the minimum difference between an estimated and an empirical default rate that must be observed in order to not to conclude that there is no evidence of a difference in the rates.

In closing, we reiterate that while values that exceed the levels calculated by these formulae are necessary for experiments to have significance, they are almost certainly not sufficient. Thus, while they can be used to disqualify results in cases where the experiment produces an effect that is smaller than ε or that uses a sample smaller than n , they cannot be used to determine that a result is significant.

APPENDIX A

TABLES FOR LOWER BOUNDS OF ε AND n

TABLE 5 Analytic lower bound on deviation, ε , that must be observed for level p and fixed sample size n .
 $\alpha = 0.05$

This table gives the lower bound on the difference, ε , between the predicted and actual default rates that must be observed to conclude that a significant difference exists at the $\alpha = 5\%$ level between the two, given n firms are available to test. Assumes no correlation in the default rates. Estimates are made assuming the limiting properties of the binomial including zero correlation.

$n \setminus p$	0.0005	0.0010	0.0050	0.0100	0.0250	0.0500	0.0750	0.1000	0.1250	0.1500	0.2000
50	*	*	*	*	*	*	*	0.0832	0.0917	0.0990	0.1109
100	*	*	*	*	*	0.0427	0.0516	0.0588	0.0648	0.0700	0.0784
250	*	*	*	*	0.0194	0.0270	0.0326	0.0372	0.0410	0.0443	0.0496
500	*	*	*	0.0087	0.0137	0.0191	0.0231	0.0263	0.0290	0.0313	0.0351
1,000	*	*	0.0044	0.0062	0.0097	0.0135	0.0163	0.0186	0.0205	0.0221	0.0248
2,500	*	*	0.0028	0.0039	0.0061	0.0085	0.0103	0.0118	0.0130	0.0140	0.0157
5,000	*	0.0009	0.0020	0.0028	0.0043	0.0060	0.0073	0.0083	0.0092	0.0099	0.0111

* indicates those cells in which the $npq < 4$ and tended to be unreliable.

TABLE 6 Analytic lower bound on deviation, ε , that must be observed for level p and fixed sample size n .
 $\alpha = 0.01$

This table gives the lower bound on the difference, ε , between the predicted and actual default rates that must be observed to conclude that a significant difference exists at the $\alpha = 1\%$ level between the two, given n firms are available to test. Assumes no correlation in the default rates. Estimates are made assuming the limiting properties of the binomial including zero correlation.

$n \setminus p$	0.0005	0.0010	0.0050	0.0100	0.0250	0.0500	0.0750	0.1000	0.1250	0.1500	0.2000
50	*	*	*	*	*	*	*		0.1205	0.1301	0.1457
100	*	*	*	*	*	0.0561	0.0678	0.0773	0.0852	0.0920	0.1030
250	*	*	*	*	0.0254	0.0355	0.0429	0.0489	0.0539	0.0582	0.0652
500	*	*	*	0.0115	0.0180	0.0251	0.0303	0.0346	0.0381	0.0411	0.0461
1,000	*	*	0.0057	0.0081	0.0127	0.0178	0.0215	0.0244	0.0269	0.0291	0.0326
2,500	*	*	0.0036	0.0051	0.0080	0.0112	0.0136	0.0155	0.0170	0.0184	0.0206
5,000	*	0.0012	0.0026	0.0036	0.0057	0.0079	0.0096	0.0109	0.0120	0.0130	0.0146

* indicates those cells in which the $npq < 4$ and tended to be unreliable.

TABLE 7 Analytic lower bound on sample size, n , for level p to consider deviation of ε significant
 $\alpha = 0.05$

This table gives the lower bound on the number of independent observations, n , that would be needed to be certain that an observed difference, ε , between the predicted and actual default rates were significant difference exists at the $\alpha= 5\%$ level. Assumes no correlation in the default rates. Estimates are made assuming the limiting properties of the binomial including zero correlation.

εp	0.0005	0.0010	0.0050	0.0100	0.0250	0.0500	0.0750	0.1000	0.1250	0.1500	0.2000
0.0001	191,977	383,762	1,911,126	3,803,044	9,363,556	18,246,929	26,650,121	34,573,129	42,015,956	48,978,600	61,463,341
0.0005	*	15,350	76,445	152,122	374,542	729,877	1,066,005	1,382,925	1,680,638	1,959,144	2,458,534
0.0010	*	*	19,111	38,030	93,636	182,469	266,501	345,731	420,160	489,786	614,633
0.0025	*	*	*	6,085	14,982	29,195	42,640	55,317	67,226	78,366	98,341
0.0050	*	*	*	*	3,745	7,299	10,660	13,829	16,806	19,591	24,585
0.0100	*	*	*	*	*	1,825	2,665	3,457	4,202	4,898	6,146

* indicates those cells in which the $npq < 4$ and tended to be unreliable.

TABLE 8 Analytic lower bound on sample size, n , for level of p to consider deviation of ε significant
 $\alpha = 0.01$

This table gives the lower bound on the number of independent observations, n , that would be needed to be certain that an observed difference, ε , between the predicted and actual default rates were significant difference exists at the $\alpha= 1\%$ level. Assumes no correlation in the default rates. Estimates are made assuming the limiting properties of the binomial including zero correlation.

εp	0.0005	0.0010	0.0050	0.0100	0.0250	0.0500	0.0750	0.1000	0.1250	0.1500	0.2000
0.0001	331,579	662,826	3,300,861	6,568,548	16,172,560	31,515,759	46,029,595	59,714,069	72,569,182	84,594,932	106,158,346
0.0005	*	26,513	132,034	262,742	646,902	1,260,630	1,841,184	2,388,563	2,902,767	3,383,797	4,246,334
0.0010	*	*	33,009	65,685	161,726	315,158	460,296	597,141	725,692	845,949	1,061,583
0.0025	*	*	*	10,510	25,876	50,425	73,647	95,543	116,111	135,352	169,853
0.0050	*	*	*	*	6,469	12,606	18,412	23,886	29,028	33,838	42,463
0.0100	*	*	*	*	*	3,152	4,603	5,971	7,257	8,459	10,616

* indicates those cells in which the $npq < 4$ and tended to be unreliable.

APPENDIX B

ALGORITHM FOR ESTIMATING ε THROUGH SIMULATION

The goal is to determine an upper bound on the size of the deviation, ε , between the assumed default rate, p , and the observed default rate f_d at a confidence level of α . The basic idea is to simulate N events with probability p of default a large number of times, to calculate the difference between the (known) p and observed f_d for each iteration, and to calculate the percentile of the distribution of these observed differences consistent with α . The simulation is straightforward and follows the algorithm below. An example of the implementation in S-Plus is given following the pseudo-code. The code shown is designed for two-sided tests, but it is trivial to modify the code to examine one-sided tests.

Algorithm: calc simulated default eps

1. *input* p, N, α
2. *for* $s=1$ *to* simus
 - i. *for* $i = 1$ *to* N
 - $u = \text{random_uniform}(0,1)$
 - if* $u \leq p$
 - $x[i] = 1$
 - else*
 - $x[i] = 0$
 - endif*
 - endfor* i
 - ii. $f_d[s] = \sum x[i] / N$
 - iii. $\varepsilon[s] = |f_d[s] - p|$
- endfor* s
- return* $\text{percentile}(\varepsilon, 1-\alpha)$

S-Plus code to implement calc simulated default eps

```
calc.simulated.default.eps.fn<-function(p = 0.05, N = 5000, simus = 100000, alpha =
0.05)
{
  x <- 1:simus
  for(j in 1:simus) {
    x[j] <- abs(mean(simulate.default.fn(p, N)) - p)
  }
  eps <- quantile(x, 1 - alpha)

  return(eps)
}

simulate.default.fn <- function(p = 0.05, N = 5000)
{
  x <- runif(N)
  ret <- ifelse(x < p, 1, 0)
  return(ret)
}
```


REFERENCES

1. Cantor, R. and E. Falkenstein (2001). "Testing for Default Consistency in Annual Default Rates." The Journal of Fixed Income (September).
2. Cochran, W. G. (1977). Sampling Techniques. New York, John Wiley & Sons.
3. Dwyer, D. and R. M. Stein (2003). Inferring the Default Rate in a Population by Comparing Two Incomplete Default Databases, Technical Report #021216, Moodys' KMV, New York
4. Gersback, H. and A. Lipponer (2000). The Correlation Effect. Heidelberg, Alfred-Weber-Institut, University of Heidelberg.
5. Grinstead, C. M. and J. L. Snell (1997). Introduction to Probability, American Mathematical Society.
6. Kurbat, M. and I. Korablev (2002). Methodology for Testing the Level of the EDFTM Credit Measure. San Francisco, Moody's KMV.
7. Papoulis, A. (1991). Probability, Random Variables and Stochastic Processes. New York, McGraw-Hill.