Advances and Applications in Statistics
Volume 20, Number 1, 2011, Pages 25-41

FOCUSING ON THE LOWER SCORING DATA IN ORDER TO IMPROVE CREDIT SCORING MODEL SELECTION

ARTHUR L. DRYVER

Graduate School of Business Administration
National Institute of Development Administration
Bangkapi, Bangkok, 10240, Thailand
e-mail: dryver@gmail.com

Abstract

Under certain circumstances, such as with various credit scoring models, the model's performance in the tails is especially important. Standard two-sample tests statistics can be improved when the concerned difference is on the tails. A new methodology is developed, which splits the combined samples in half using the overall sample median of the $\hat{y}$'s as the cut point in order to place more focus on the tails. This new methodology can be used to aid in model selection for credit scoring models. Simulation results of the new methodology yield higher power when the population mean and standard deviation are the same for the two populations. Thus another application is to test if standardized scores have the same distribution.

1. Introduction

Credit scoring models are basically models that use credit data. These models can be used to estimate the probability of default on the re-payment of credit (risk), fraud and even the probability of a response to marketing campaigns. For risk, fraud and response models, the dependent variable can be broken out into two groups, one

---

2010 Mathematics Subject Classification: 91Gxx, 91G40.

Keywords and phrases: Anderson-Darling test statistic, credit scoring, goodness-of-fit tests, Kolmogorov-Smirnov test statistic, standardized test scores.

Received February 16, 2011

group comprised of the successes and the second group of the failures. For a risk model, successes are considered good loans and failures bad loans, and similarly for fraud and response models. There has been research comparing modeling techniques within credit scoring (e.g., Reichert et al. [10], Hand and Henley [7], Srinivasan and Kim [13], Doumpos and Pasiouras [3], Galindo and Tamayo [5], Mo and Yau [9]). A subject of equal importance is how to validate the model, as different test statistics can lead to different conclusions on model performance. In a general sense, a good model will produce scores for the successes and failures that differ significantly. In order to determine, if the model is good or not, then often goodness-of-fit tests are used.

Goodness-of-fit tests in the two-sample case are concerned with whether or not the two-samples are from the same probability distribution function. For this case, the tests determine if the scores of the successes and failures come from the same distribution or not. If the scores for the successes and failures have a similar distribution, then the scores cannot be used to differentiate between the successes and failures. Thus the concept in using a goodness-of-fit test is that the more different the distribution of scores is the better for using the scores to differentiate between the successes and failures. One such goodness-of-fit test statistic that is commonly used (Dryver and Sukkasem [4]) is the two-sample Kolmogorov-Smirnov statistic (K-S statistic), which was introduced by Smirnov [12] to test the equality between two c.d.f.s. The two-sample K-S statistic is simply the maximum difference between the two empirical distribution functions (e.d.f.s).

Unfortunately, the K-S statistic lacks sensitivity when the difference in the true probability distributions is in the tails of the c.d.f.s (Mason and Schuenemeyer [8]). This can be seen in Tables 1 and 2, where the K-S statistic is the same for the two models, model A and model B, but the models perform differently in the tails. The Anderson-Darling test statistic (A-D) is sensitive to differences in the tails (Scholz and Stephens [11]), but when the tails of the distribution are of interest, there is room for improvement. In the credit scoring arena, the modeler expects the distributions of scores for “goods” and “bads” to differ but the focus in terms of K-S statistic is for model comparison. Thus this short coming can yield misleading results when the K-S statistic is used for model selection.

The purpose of the new methodology proposed, leveraging the standard two-sample A-D or K-S statistic, is to propose a more powerful test than the standard statistics when two distributions are similar but differ in the tails. This new test not only aids in the comparison of credit scoring models, but also in the comparison of standardized test scores. Standardized test scores such as CEEB scores (Hanania and Shikhani [6]) will have the same mean and standard deviation, as they are standardized but may not have the same distribution. It may be of concern if the distributions of two standardized scores are the same for various reasons. In this case, the proposed new methodology is more powerful and should perform better than the standard K-S and A-D statistics, as shown in Section 5.

The new methodology proposed can be thought of as a partitioning technique, which ultimately yields two A-D or K-S statistics for testing the equality of distributions. The sample median of the combined scores for “goods” and “bads” is used to partition the two-samples into essentially four samples. Then the A-D or K-S statistic is calculated for below the sample median and above the sample median. In this manner, each individual test on the partitioned samples is more sensitive to the difference in the tails. Unfortunately, the statistics from the partitioned samples are dependent, as shown in Section 3, but from the simulation, as shown in Section 5, it can be seen that the correlation is very small. Due to the fact that there are now two hypothesis tests as opposed to a single test, the significance level for the individual tests must be adjusted in order to obtain the desired overall significance level, α. The adjustment used to α for the individual tests is straightforward, as shown in Section 3. Throughout this paper, we will set the overall $\alpha = 0.05$ for rejection of the null hypothesis.

In addition, it may be more beneficial to merely do a single hypothesis test on the partitioned samples as opposed to two hypothesis tests. For example, with a fraud detection model, only a small percentage of the applicants will be rejected due to risk of fraud. Thus for a fraud detection model, the only concern is the low scoring individuals with the highest potential of fraud. For example, model A in Table 1 performs better than model B in Table 2 when focusing on the bottom 20% and 10%, but the K-S is equal for both models. The A-D statistic shows that the model A is better than model B in terms of differentiating between goods and frauds as it does better when the difference is in the tails. The A-D statistic also shows that the model A is better than model C, see Tables 1 and 3, but when focusing on the bottom 20% and 10%, we can see the models actually perform the same. For this situation, the test statistic on the partitioned sample would be more informative (see Section 4) and in fact, a single test statistic would suffice. For this reason, a single hypothesis is also investigated in Section 5, using an adjusted α_p set equal to the desired α, as it is a single test. The limiting distribution of the test statistics is only known under the null hypothesis and is unknown under the alternative hypothesis (Capon [2], Scholz and Stephens [11]). Thus the power of this new testing methodology compared to the standard two-sample test statistics is investigated through simulation under various alternatives in Section 5.

Table 1. Validation of fraud model A

Score Category	Total #	Cum. %	# of Non-Fraud	# of Fraud	Cum. % Non-Fraud	Cum. % Fraud	The Difference
10	500,000	10%	497,500	2,500	10.2%	2.5%	7.65%
9	500,000	20%	496,000	4,000	20.3%	6.5%	13.78%
8	500,000	30%	494,500	5,500	30.4%	12.0%	18.37%
7	500,000	40%	493,500	6,500	40.4%	18.5%	21.94%
6	500,000	50%	493,000	7,000	50.5%	25.5%	25.00%
5	500,000	60%	492,500	7,500	60.6%	33.0%	27.55%
4	500,000	70%	491,500	8,500	70.6%	41.5%	29.08%
3	500,000	80%	488,250	11,750	80.5%	53.3%	27.30%
2	500,000	90%	483,250	16,750	90.4%	70.0%	20.41%
1	500,000	100%	470,000	30,000	100.0%	100.0%	0.00%
Total	5,000,000		4,900,000	100,000		AD=31433 & K-S=29.08%	

Table 2. Validation of fraud model B

Score Category	Total #	Cum. %	# of Non-Fraud	# of Fraud	Cum. % Non-Fraud	Cum. % Fraud	The Difference
10	500,000	10%	497,000	3,000	10.1%	3.0%	7.14%
9	500,000	20%	496,500	3,500	20.3%	6.5%	13.78%
8	500,000	30%	494,500	5,500	30.4%	12.0%	18.37%
7	500,000	40%	493,500	6,500	40.4%	18.5%	21.94%
6	500,000	50%	493,000	7,000	50.5%	25.5%	25.00%
5	500,000	60%	492,500	7,500	60.6%	33.0%	27.55%
4	500,000	70%	491,500	8,500	70.6%	41.5%	29.08%
3	500,000	80%	484,500	15,500	80.5%	57.0%	23.47%
2	500,000	90%	482,000	18,000	90.3%	75.0%	15.31%
1	500,000	100%	475,000	25,000	100.0%	100.0%	0.00%
Total	5,000,000		4,900,000	100,000		AD=27157 & K-S=29.08%	

Table 3. Validation of fraud model C

Score Category	Total #	Cum. %	# of Non-Fraud	# of Fraud	Cum. % Non-Fraud	Cum. % Fraud	The Difference
10	500,000	10%	497,000	3,000	10.1%	3.0%	7.14%
9	500,000	20%	496,500	3,500	20.3%	6.5%	13.78%
8	500,000	30%	494,500	5,500	30.4%	12.0%	18.37%
7	500,000	40%	493,500	6,500	40.4%	18.5%	21.94%
6	500,000	50%	493,000	7,000	50.5%	25.5%	25.00%
5	500,000	60%	492,500	7,500	60.6%	33.0%	27.55%
4	500,000	70%	491,500	8,500	70.6%	41.5%	29.08%
3	500,000	80%	488,250	11,750	80.5%	53.3%	27.30%
2	500,000	90%	483,250	16,750	90.4%	70.0%	20.41%
1	500,000	100%	470,000	30,000	100.0%	100.0%	0.00%
Total	5,000,000		4,900,000	100,000		AD=31326 & K-S=29.08%	

2. Two-sample Statistics

The K-S and A-D statistic tests if two independent random samples from unknown continuous cumulative distribution functions (c.d.f.s), F_X and F_Y, actually are from the same c.d.f.s (Smirnov [12], Scholz and Stephens [11]). For this paper, the two distributions are the distribution of scores for the successes, *X*, and the scores for the failures, *Y*. A very good model will produce very different scores for the successes and failures. Another way to view this is that the scores for the successes come from a different distribution than the failures. This is one reason why the K-S statistics are used to determine the strength of the credit scoring model. The A-D and K-S statistic tests null hypothesis

$$H_0 : F_X = F_Y \text{ versus } H_A : F_X \neq F_Y. \tag{1}$$

2.1. The two-sample K-S statistic

The K-S statistic equals

$$KS_{mn} = \max_t | \hat{F}_{Xm}(t) - \hat{F}_{Yn}(t) |, \tag{2}$$

where $\hat{F}_{Xm}(t)$ and $\hat{F}_{Yn}(t)$ are the e.d.f.s of F_X and F_Y evaluated at *t* from samples of size *m* and *n*, respectively. The limiting distribution of $\sqrt{\frac{mn}{m+n}}KS_{mn}$ under the null hypothesis (Capon [2]) is

$$\lim_{m, n \to \infty} Prob\left(\sqrt{\frac{mn}{m+n}} KS_{mn} \le t\right) = 1 - 2\sum_{j=1}^{\infty}(-1)^{j-1}\exp(-2j^2t^2),\ t \ge 0$$

$$= 0,\ t < 0, \tag{3}$$

from which the *p*-value can be calculated for large *m* and *n*.

2.2. The two-sample A-D statistic

The A-D statistic can test if two independent random samples from unknown continuous cumulative distribution functions (c.d.f.s), F_X and F_Y, actually are from the same c.d.f.s (Scholz and Stephens [11]). The A-D statistic can be used in the same manner as the K-S statistic in Subsection 2.1 to test null hypothesis 1. The following formula and general notation are from Scholz and Stephens [11], adjusted for the fact there are only two samples considered in this paper whereas the original formula allows for multiple, *k*, samples

$$AD = \frac{A_{2N}^2 - (2-1)}{\sigma_N},$$

$$A_{2N}^2 = \frac{1}{n}\sum_{j=1}^{L-1}\frac{l_j}{N}\frac{(NM_{ij} - nB_j)^2}{B_j(N - B_j)} + \frac{1}{m}\sum_{j=1}^{L-1}\frac{l_j}{N}\frac{(NM_{ij} - mB_j)^2}{B_j(N - B_j)},$$

$$\sigma_N^2 = \text{var}(A_{2N}^2) = \frac{aN^3 + bN^2 + cN + d}{(N-1)(N-2)(N-3)},$$

$$a = (4g - 6)(2 - 1) + (10 - 6g)H,$$

$$b = (2g - 4)2^2 + 8h(2) + (2g - 14h - 4)H - 8h + 4g - 6,$$

$$c = (6h + 2g - 2)2^2 + (4h - 4g + 6)2 + (2h - 6)H + 4h,$$

$$d = (2h + 6)2^2 - (4h)2,$$

$$H = \frac{1}{n} + \frac{1}{m}, \quad h = \sum_{i=1}^{N-1}\frac{1}{i},$$

$$g = \sum_{i=1}^{N-2} \sum_{j=i+1}^{N-1} \frac{1}{(N-i)\,j}, \quad M_{ij} = \sum_{u=1}^{j} f_{iu},$$

$$l_j = \sum_{i=1}^{2} f_{ij}, \quad B_j = \sum_{u=1}^{j} l_u. \tag{4}$$

Let $Z_1^* < Z_2^* < \cdots < Z_L^*$ represent the ordered combined sample after removing duplicates and f_{ij} equals the number of observations in the *i*th sample equal to Z_j^*. Thus *L* is the number of unique observations and $N = n + m$.

3. Split Sample Statistics

The new methodology stems from the following concept:

$$H_0 : F_X \ (X \le M) = F_Y \ (Y \le M) \text{ and } H_0 : F_X \ (X > M) = F_Y \ (Y > M)$$

versus

$$H_A : F_X \ (X \le M) \ne F_Y \ (Y \le M) \text{ and/or } H_A : F_X \ (X > M) \ne F_Y \ (Y > M), \tag{5}$$

where *M* is the population median, which is the same under the null hypothesis. In addition, the new methodology can be used to focus on a more important part of the distribution, as in credit scoring near the cutoff. In short, it is possible to investigate either

$$H_0 : F_X \ (X \le M) = F_Y \ (Y \le M) \text{ versus } H_A : F_X \ (X \le M) \ne F_Y \ (Y \le M)$$

or

$$H_0 : F_X \ (X > M) = F_Y \ (Y > M) \text{ versus } H_A : F_X \ (X > M) \ne F_Y (Y > M). \tag{6}$$

Null hypothesis 5 is false if and only if null hypothesis 1 is false. Thus a test on null hypothesis 5 can be used to test null hypothesis 1. In order to test null hypothesis 5, the data are split in half using the sample median of all the scores for successes and failures combined resulting from the credit scoring model. After splitting the sample scores in this manner, four e.d.f.s are created. There are two e.d.f.s created for scores above the sample median, one for successes and one for failures. In addition, two e.d.f.s are created for scores below the sample median. Finally, two K-S or A-D statistics are calculated, one using the top half scoring observations, KS_1 or AD_1, and another for the bottom half of the scoring individuals, KS_2 or AD_2. In the rare case, when the smallest data point in one sample is larger than the largest data point in the other data point, technically the statistics cannot be calculated. In this case, the *p*-values for all AD_* and KS_* are set to zero and KS_1 and KS_2 will be set to 100%. Assuming sufficient data, if one sample's minimum is larger than the other sample's maximum, then the latter is a reasonable approach as it is expected to reject null hypothesis 1.

In practice, *n* and *m* are known, but the values of n_1, m_1, n_2 and m_2 are obtained after the partitioning of the data by the sample median, $\hat{m}$, of the entire dataset. From Table 4, it can be seen that if n_1 is given, then the values of n_2, m_1 and m_2 can be determined.

Table 4. The breakout of the data, where $Z_{(i)}$ represents the ordered data points

	$Z_{(i)} \leq \hat{m}$	$Z_{(i)} > \hat{m}$	
X	n_1	n_2	$n_1 + n_2 = n$
Y	m_1	m_2	$m_1 + m_2 = m$
Total	$m_1 + n_1 = \frac{n+m}{2}$	$m_2 + n_2 = \frac{n+m}{2}$	$n + m$

From equation (3), it can be seen that the distribution of the two-sample K-S statistic depends upon the values of the sample sizes. Under null hypothesis 1, all of the observations were generated independently and were identically distributed (i.i.d.). Unfortunately, n_1, n_2, m_1 and m_2 are not independent (see Table 4). Thus the probability distribution functions for sample statistics, which rely on n_1, n_2, m_1 and m_2 are not independent. Knowing *m*, *n* and n_1, then m_1, n_2 and m_2 can be solved. Finally, under the null hypothesis, the distribution of n_1 before any data observed, is

$$n_1 \sim bin\left(n,\ \pi = \frac{1}{2}\right) \text{ if } n \text{ is even, and}$$

$$n_1 \sim bin\left(n,\ \pi = \frac{n/2+1}{n}\right) \text{ if } n \text{ is odd.} \tag{7}$$

3.1. The split sample K-S statistics

The split sample K-S statistics are defined using the sample sizes obtained after the sample is split. The e.d.f.s are used to investigate the c.d.f.s split at the population median. The KS_1 statistic equals

$$KS_1 = \max_t |\hat{F}_{1Xm_1}(t) - \hat{F}_{1Yn_1}(t)|, \tag{8}$$

where $\hat{F}_{1Xm_1}(t)$ and $\hat{F}_{1Yn_1}(t)$ are the e.d.f.s to investigate F_X $(X \leq M)$ and F_Y $(Y \leq M)$ evaluated at t from samples of sizes m_1 and n_1, respectively. The KS_2 statistic is defined similarly and equals

$$KS_2 = \max_t |\hat{F}_{2Xm_2}(t) - \hat{F}_{2Yn_2}(t)|, \tag{9}$$

where $\hat{F}_{2Xm_2}(t)$ and $\hat{F}_{2Yn_2}(t)$ are the e.d.f.s to investigate F_X $(X > M)$ and F_Y $(Y > M)$ evaluated at t from samples of sizes m_2 and n_2, respectively. The limiting distribution of $\sqrt{\frac{m_1 n_1}{m_1 + n_1}} KS_{1_{m_1 n_1}}$ given n_1 under the null hypothesis (Capon [2]) is

$$\begin{aligned} \lim_{m_1, n_1 \to \infty} Prob\left(\sqrt{\frac{m_1 n_1}{m_1 + n_1}} KS_{1_{m_1 n_1}} \leq t \,|\, n_1\right) &= 1 - 2\sum_{j=1}^{\infty} (-1)^{j-1} \exp(-2j^2 t^2),\ t \geq 0 \\ &= 0,\ t < 0, \end{aligned} \tag{10}$$

and for $\sqrt{\frac{m_2 n_2}{m_2 + n_2}} KS_{2_{m_2 n_2}}$ given n_1 under the null hypothesis is

$$\begin{aligned} &\lim_{m_2, n_2 \to \infty} Prob\left(\sqrt{\frac{m_2 n_2}{m_2 + n_2}} KS_{2_{m_2 n_2}} \leq t \,|\, n_1\right) \\ &= 1 - 2\sum_{j=1}^{\infty} (-1)^{j-1} \exp(-2j^2 t^2),\ t \geq 0 \\ &= 0,\ t < 0, \end{aligned} \tag{11}$$

from which the p-values can be calculated for large n_1, m_1, n_2 and m_2.

3.2. The split sample A-D statistics

The split sample Anderson-Darling statistics work in the same manner as the split K-S statistics in Subsection 3.1. The calculations of AD_1 and AD_2:

$$AD_1 = \frac{A^2_{2N_1} - (2-1)}{\sigma_{N_1}},$$

$$A^2_{2N_1} = \frac{1}{n_1}\sum_{j=1}^{L_1-1}\frac{l_{j1}}{N_1}\frac{(N_1 M_{ij1} - n_1 B_{j1})^2}{B_{j1}(N_1 - B_{j1})} + \frac{1}{m_1}\sum_{j=1}^{L_1-1}\frac{l_{j1}}{N_1}\frac{(N_1 M_{ij1} - m_1 B_{j1})^2}{B_{j1}(N_1 - B_{j1})},$$

$$\sigma^2_{N_1} = \mathrm{var}(A^2_{2N_1}) = \frac{a_1 N_1^3 + b_1 N_1^2 + c_1 N_1 + d_1}{(N_1 - 1)(N_1 - 2)(N_1 - 3)}, \tag{12}$$

$$AD_2 = \frac{A^2_{2N_2} - (2-1)}{\sigma_{N_2}},$$

$$A^2_{2N_2} = \frac{1}{n_2}\sum_{j=1}^{L_2-1}\frac{l_{j2}}{N_2}\frac{(N_2 M_{ij2} - n_2 B_{j2})^2}{B_{j2}(N_2 - B_{j2})} + \frac{1}{m_2}\sum_{j=1}^{L_2-1}\frac{l_{j2}}{N_2}\frac{(N_2 M_{ij2} - m_2 B_{j2})^2}{B_{j2}(N_2 - B_{j2})},$$

$$\sigma^2_{N_2} = \mathrm{var}(A^2_{2N_2}) = \frac{a_2 N_2^3 + b_2 N_2^2 + c_2 N_2 + d_2}{(N_2 - 1)(N_2 - 2)(N_2 - 3)}, \tag{13}$$

where $Z_1^{1*} < Z_2^{1*} < \cdots < Z_{L_1}^{1*} \le \hat{m}$ represents the ordered combined sample less than or equal to the sample median after removing duplicates and f_{ij1} equals the number of observations in the *i*th sample equal to Z_j^{1*}. Similarly, $\hat{m} < Z_1^{2*} < Z_2^{2*} < \cdots < Z_{L_2}^{2*}$ represents the ordered combined sample greater than the sample median after removing duplicates and f_{ij2} equals the number of observations in the *i*th sample equal to Z_j^{2*}. Thus L_1 is the number of unique observations less than or equal to the sample median, $N_1 = n_1 + m_1$, L_2 is the number of unique observations greater than the sample median, and $N_2 = n_2 + m_2$.

3.3. Adjusted alpha levels

For the split K-S statistics and split A-D statistics, when testing both KS_1 and KS_2 or AD_1 and AD_2, an adjusted rejection region is used. This is done in order to obtain an overall approximate $\alpha = 0.05$ for null hypothesis 5. The adjusted α, α_p used in the simulation are

$$\alpha = 1 - (1 - \alpha_p)^2,$$

$$1 - \alpha = (1 - \alpha_p)^2,$$

$$\alpha_p = 1 - \sqrt{(1 - \alpha)} = 1 - \sqrt{.95},$$

$$\alpha_p = 0.02532.$$

Note that for certain values of $n_1,$ it may be impossible to obtain the desired level of significance. For example, it is possible for $n_1 = 0,$ under which the null hypothesis is automatically rejected, as the probability of $n_1 = 0$ is very small under null hypothesis 5, where the distribution of n_1 under the null hypothesis is given by equation (7) in Section 3. In the case that $n_1 = 0$ or $n_1 = n,$ all of the scores of one sample are greater than all of the scores of the other sample, which would lead one to believe that even without a statistical test, the scores of the successes and failures do not come from the same distribution. Thus, for a sufficiently large n and m setting, $\alpha_p = 0.02532$ yields an approximate level of significance of 0.05, as shown in Section 5. In addition, it is investigated when the researcher is only interested in either scores less than or equal to the median or above the median, but not concerned with both, for example, using KS_2 or AD_2 to test only $H_0 : F_{2X} = F_{2Y}$. Under this circumstance, it is not necessary to adjust the rejection in order to obtain the desired significance level, as only a single test statistic is used.

3.4. Theoretical limitations

The newly proposed methodology has some limitations. For starters, the null hypothesis 5 splits the distributions by the population median, whereas the sample median is used to calculate the statistics proposed in this paper. This affects the breakout of m and n into $n_1,$ $m_1,$ n_2 and $m_2,$ which are used to calculate the statistics. Also, there is a lack of independence of the split sample statistics resulting from the split sample sizes (see Table 4). For larger samples, the sample median converges to the population median and making the latter concerns for sufficiently large sample sizes minor concerns to non-issues. For example, within credit scoring and model selection, often sample sizes are in the thousands and thus using the split K-S for model selection, the latter concerns are very minor. Finally, a small note is taken that the statistics are on e.d.f.s but the null hypothesis 5 merely splits the c.d.f.s.

4. Illustrative Example

Tables 5 and 7 are from model A in Table 1, partitioned by the sample median, from which the K-S statistic and A-D statistic were calculated for the upper 50% scoring individuals and separately for the bottom 50% scoring individuals. Tables 6 and 8 are from model B in Table 2, partitioned by the sample median; and the same calculations are performed for Tables 5 and 7. Table 3 produces results that mimic the top half of Table 2 and the bottom half of Table 1 and thus Tables 6 and 7 are the split samples for Table 3 as well.

As mentioned earlier, the K-S statistic is the same for both models A and B, even though model A performs better when looking at a bottom 20% and 10% cutoff point for rejecting applicants. The split sample K-S for the model A bottom scoring individuals is higher than that of the bottom scoring individuals for model B (Tables 7 and 8), indicating that model A is better than model B, breaking the tie that occurred with the standard K-S. This is an example of one major advantage of the split sample K-S statistic, in that it is more sensitive to the tails, where it is often more important to concentrate upon for credit scoring models, and it can distinguish between two models of similar overall performance. The benefit for the A-D statistic of splitting the samples is that it makes model A evidently better than model B in the bottom half, as AD_2 in Table 7 is slightly over 50% higher than AD_2 in Table 8. Finally, the split samples clarify the fact that model C performs the same as model A in the bottom half as shown in Tables 7 and 8.

Table 5. Validation of fraud model A, top scoring half

Score Category	Total #	Cum. %	# of Non-Fraud	# of Fraud	Cum. % Non-Fraud	Cum. % Fraud	The Difference
10	500,000	20%	497,500	2,500	20.1%	9.8%	10.30%
9	500,000	40%	496,000	4,000	40.1%	25.5%	14.66%
8	500,000	60%	494,500	5,500	60.1%	47.1%	13.07%
7	500,000	80%	493,500	6,500	80.1%	72.5%	7.53%
6	500,000	100%	493,000	7,000	100.0%	100.0%	0.00%
Total	2,500,000		2,474,500	25,500		$AD_1 = 1739$ & $KS_1 = 14.66\%$	

Table 6. Validation of fraud models B and C, their top scoring halves

Score Category	Total #	Cum. %	# of Non-Fraud	# of Fraud	Cum. % Non-Fraud	Cum. % Fraud	The Difference
10	500,000	20%	497,000	3,000	20.1%	11.8%	8.32%
9	500,000	40%	496,500	3,500	40.1%	25.5%	14.66%
8	500,000	60%	494,500	5,500	60.1%	47.1%	13.07%
7	500,000	80%	493,500	6,500	80.1%	72.5%	7.53%
6	500,000	100%	493,000	7,000	100.0%	100.0%	0.00%
Total	2,500,000		2,474,500	25,500		$AD_1 = 1586$ & $KS_1 = 14.66\%$	

Table 7. Validation of fraud models A and C, their bottom scoring halves

Score Category	Total #	Cum. %	# of Non-Fraud	# of Fraud	Cum. % Non-Fraud	Cum. % Fraud	The Difference
5	500,000	20%	492,500	7,500	20.3%	10.1%	10.24%
4	500,000	40%	491,500	8,500	40.6%	21.5%	19.09%
3	500,000	60%	488,250	11,750	60.7%	37.2%	23.45%
2	500,000	80%	483,250	16,750	80.6%	59.7%	20.89%
1	500,000	100%	470,000	30,000	100.0%	100.0%	0.00%
Total	2,500,000		2,425,500	74,500		$AD_2 = 13655$ & $KS_2 = 23.45\%$	

Table 8. Validation of fraud model B, bottom scoring half

Score Category	Total #	Cum. %	# of Non-Fraud	# of Fraud	Cum. % Non-Fraud	Cum. % Fraud	The Difference
5	500,000	20%	492,500	7,500	20.3%	10.1%	10.24%
4	500,000	40%	491,500	8,500	40.6%	21.5%	19.09%
3	500,000	60%	484,500	15,500	60.5%	42.3%	18.26%
2	500,000	80%	482,000	18,000	80.4%	66.4%	13.97%
1	500,000	100%	475,000	25,000	100.0%	100.0%	0.00%
Total	2,500,000		2,425,500	74,500		AD_2=9081 & KS_2=19.09%	

5. Simulation Results

There are 24 simulations under the null hypothesis in Table 9, and 24 simulations under various alternative hypotheses in Table 10, with varying sample sizes for *m* and *n*. The cut-off for when a single K-S, either the standard K-S or a split sample K-S, is $\alpha = 0.05$. For testing null hypothesis 5, if the *p*-value of either KS_1 or KS_2 was below $\alpha = 0.02532$, then the null was rejected. Ten thousand iterations were performed under each scenario in order to estimate the power of the test under the various sample sizes and alternatives. The results given are the number of iterations rejected divided by the number of iterations, 10,000. In addition, the sample correlation of KS_1 and KS_2, and the sample correlation of the *p*-value of AD_1 and *p*-value of AD_2, are calculated[1].

Table 9. Simulation results under the null hypothesis with various distributions

	$n = m$	K-S	KS_1 or KS_2	KS_1	KS_2	$r_{KS_1 KS_2}$	AD	AD_1 or AD_2	AD_1	AD_2	$r_{AD_1 AD_2}$
Normal(0,1)	50	0.033	0.042	0.047	0.045	-0.008	0.049	0.048	0.051	0.048	-0.012
	100	0.036	0.049	0.048	0.048	-0.004	0.050	0.051	0.054	0.047	0.005
	250	0.041	0.043	0.046	0.042	0.007	0.050	0.049	0.050	0.050	0.003
	500	0.050	0.044	0.044	0.045	0.005	0.052	0.052	0.050	0.051	-0.003
Log-normal(0,1)	50	0.040	0.046	0.046	0.048	0.002	0.048	0.053	0.053	0.050	0.002
	100	0.038	0.045	0.048	0.046	-0.002	0.050	0.050	0.051	0.049	-0.004
	250	0.041	0.041	0.041	0.043	0.016	0.048	0.053	0.048	0.050	0.013
	500	0.048	0.042	0.045	0.047	0.001	0.050	0.054	0.052	0.054	0.003
Uniform(0,1)	50	0.038	0.046	0.045	0.048	-0.019	0.049	0.053	0.051	0.051	-0.023
	100	0.036	0.043	0.046	0.044	0.008	0.054	0.050	0.048	0.051	0.011
	250	0.041	0.044	0.043	0.044	0.005	0.049	0.057	0.052	0.054	0.004
	500	0.050	0.042	0.042	0.043	-0.017	0.051	0.050	0.050	0.048	-0.020
T-distribution d.f.=10	50	0.043	0.046	0.046	0.050	0.003	0.052	0.053	0.047	0.055	0.000
	100	0.033	0.046	0.043	0.050	0.013	0.051	0.053	0.046	0.051	0.010
	250	0.047	0.043	0.043	0.042	-0.018	0.051	0.053	0.051	0.049	-0.005
	500	0.046	0.041	0.045	0.040	0.014	0.047	0.051	0.050	0.047	0.007

---

[1] The correlation was set to NA when there was no variation in the *p*-value of AD_1 or AD_2 and correlation was undefined as a result. The *p*-value was used to calculate correlation for AD_1 and AD_2 because AD_1 and AD_2 are undefined when $n_1 = n$ and $n_1 = 0$, but their *p*-values are set to zero and thus defined under those circumstances.

Table 10. Simulation results under the alternative hypothesis with various distributions

	$n = m$	K-S	KS_1 or KS_2	KS_1	KS_2	$r_{KS_1KS_2}$	AD	AD_1 or AD_2	AD_1	AD_2	$r_{AD_1AD_2}$
N(0,1) & T-dist d.f.=10	50	0.038	0.048	0.050	0.047	-0.018	0.053	0.058	0.059	0.057	-0.013
	100	0.038	0.055	0.051	0.051	0.012	0.058	0.064	0.058	0.062	0.011
	250	0.044	0.055	0.056	0.052	0.008	0.063	0.085	0.081	0.077	-0.007
	500	0.055	0.079	0.073	0.076	-0.005	0.089	0.137	0.118	0.124	-0.012
N(0,1.118) & T-dist d.f.=10	50	0.042	0.050	0.052	0.048	-0.002	0.053	0.059	0.058	0.054	-0.003
	100	0.042	0.067	0.056	0.062	-0.008	0.056	0.069	0.059	0.066	-0.009
	250	0.056	0.078	0.072	0.068	-0.002	0.062	0.091	0.080	0.080	-0.012
	500	0.078	0.116	0.107	0.102	0.007	0.076	0.135	0.118	0.111	0.006
N(0.5,0.2887) & U(0,1)	50	0.070	0.129	0.118	0.116	-0.016	0.074	0.146	0.126	0.127	-0.025
	100	0.097	0.250	0.203	0.202	-0.000	0.120	0.265	0.220	0.220	0.005
	250	0.258	0.531	0.417	0.425	-0.000	0.397	0.647	0.530	0.535	-0.002
	500	0.580	0.859	0.739	0.738	0.022	0.888	0.961	0.892	0.892	0.009
N(1.6487,2.1611) & LogN(0,1)	50	0.660	0.932	0.955	0.243	0.072	0.833	0.938	0.962	0.265	0.031
	100	0.984	1.000	1.000	0.436	0.083	0.999	1.000	1.000	0.477	0.030
	250	1.000	1.000	1.000	0.791	0.093	1.000	1.000	1.000	0.878	0.012
	500	1.000	1.000	1.000	0.979	0.086	1.000	1.000	1.000	0.997	NA

6. Discussion

The split sample test statistics yield the desired α under the null hypothesis, Table 9, even for smaller sample sizes despite some of the concerns expressed in Subsection 3.4. The split sample test statistics are consistently more powerful than their standard two-sample counterparts for testing null hypothesis 1 under the investigated alternative distributions and sample sizes, as shown in Table 10. Even in the case, where only a single split sample test statistic was used to reject null hypothesis 1, it was consistently considerably more powerful, with the only exception being under the lognormal distribution when compared to a normal distribution with the same mean and variance. When comparing a lognormal to a normal distribution, KS_1 and AD_1 outperformed the standard two-sample counterparts and KS_2 and AD_2 underperformed the standard two-sample statistics. It should be noted that under the alternative hypothesis, where the distributions were a lognormal and a normal distribution, the power was greatest for the old and the new methodology when compared to all other investigated alternatives. From the simulation results, it is believed that this methodology is more powerful in general when the mean and variances of the two distributions are the same but the c.d.f.s differ in the tails. Thus, from the simulation results, it is believed that it can be used for comparing the c.d.f.s of standardized scores from various situations where the means and variances are known to be the same but the distribution is believed to be different such as CEEB scores (Hanania and Shikhani [6]). In addition, this methodology can be very useful when developing credit scoring models where the difference in the tails can be of great importance (see Section 4).

One of the benefits of the K-S statistic is the relative ease in explaining it to non-statisticians. Ease of understanding can be an attractive feature for choosing which statistic to use within the finance arena, such as with value at risk (Beder [1]). Thus another benefit of the splitting the sample and then taking the K-S statistic on only the lower scoring individuals is its relative ease of understanding especially to those already familiar with the K-S statistic within the context of credit scoring. Also, within the credit scoring context, the real question is not if the score distribution of goods and bads are the same, but which model performs best. That is the K-S statistic and other statistics are used for model comparison and selection rather than as a test of significance. There is never really one statistic that tells the entire story in model selection, but given the findings in this paper the split sample K-S should definitely be considered as another tool used for credit scoring model selection.

For future research, it is possible that a more powerful test statistic can be derived that incorporates the information in count, n_1, under the same circumstances, especially for smaller sample sizes.

Acknowledgment

The author would like to thank the Graduate School of Business Administration, NIDA, Bangkok, Thailand for their support to this work.

References

[1] T. Beder, VAR: Seductive but dangerous, Financial Analysts J. 51(5) (1995), 12-24.

[2] J. Capon, On the asymptotic efficiency of the Kolmogorov-Smirnov test, J. Amer. Statist. Assoc. 60(311) (1965), 843-853.

[3] M. Doumpos and F. Pasiouras, Developing and testing models for replicating credit ratings: a multicriteria approach, Computational Economics 25(4) (2005), 327-341.

[4] A. Dryver and J. Sukkasem, Validating risk models with a focus on credit scoring models, J. Statist. Comput. Simul. 79(2) (2009), 181-193.

[5] J. Galindo and P. Tamayo, Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications, Computational Economics 15(1) (2000), 107-143.

[6] E. Hanania and M. Shikhani, Interrelationships among three tests of language proficiency: standardized ESL, cloze, and writing, TESOL Quarterly 20(1) (1986), 97-109.

[7] D. Hand and W. Henley, Statistical classification methods in consumer credit scoring: a review, J. Roy. Statist. Soc. Ser. A 160(3) (1997), 523-541.

[8] M. D. Mason and H. J. Schuenemeyer, A modified Kolmogorov-Smirnov test sensitive to tail alternatives, Ann. Statist. 11(3) (1983), 933-946.

[9] Leo S. F. Mo and Kelvin K. W. Yau, Survival mixture model for credit risk analysis, Asia-Pacific J. Risk and Insurance 4(2) (2010), Article 5.

[10] A. Reichert, C. Cho and G. Wagner, An examination of the conceptual issues involved in developing credit-scoring models, J. Bus. Econom. Statist. 1(2) (1983), 101-114.

[11] F. Scholz and M. Stephens, *K*-sample Anderson-Darling tests, J. Amer. Statist. Assoc. 82(399) (1987), 918-924.

[12] N. V. Smirnov, Estimate of deviation between empirical distribution functions in two independent samples, Bull. Moscow Univ. 2(2) (1939), 3-16 (in Russian).

[13] V. Srinivasan and Y. Kim, Credit granting: a comparative analysis of classification procedures, J. Finance 42 (1987), 665-681.