Journal of the Korean Data & Information Science Society
2009, **20**(2), 261–272

한국데이터정보과학회지

신용평가모형에서 두 분포함수의 동일성 검정을 위한 비모수적인 검정방법

홍종선[1] · 김지훈[2]

[1]성균관대학교 통계학과 · [2]성균관대학교 응용통계연구소

접수 2009년 1월 7일, 수정 2009년 2월 19일, 게재확정 2009년 3월 3일

요 약

신용평가모형에서 두 집단의 판별력 검정방법 중의 하나로 두 분포함수의 동일성 검정을 위한 비모수적인 Kolmogorov-Smirnov (K-S) 검정방법이 대표적으로 적용되고 있다. 본 연구에서는 신용평가모형에서 두 분포함수의 동일성 검정을 위하여 K-S 검정 방법 외에 Cramer-Von Mises, Anderson-Darling, Watson 검정방법들을 소개하고 Joseph (2005)의 기준에 대응하는 판단기준을 제안한다. 또한 신용평가 자료와 유사한 상황 하에서의 모의실험을 통해서 불량률, 표본크기 그리고 제Ⅱ종 오류율을 고려한 대안적인 판단기준을 제시하고 그 적용방법에 대해서 살펴본다.

주요용어: 부도율, 스코어, 임계값, 타당성, 판별력.

1. 서론

두 확률변수 $X_1, \cdots, X_n$과 $Y_1, \cdots, Y_m$의 누적분포함수 $F_X(x)$와 $G_Y(x)$의 동일성을 검정하는 $H_0 : F_X(\cdot) = G_Y(\cdot)$에 대한 비모수적 방법으로 Kolmogorov-Smirnov, Cramér-Von Mises, Anderson-Darling, Watson 검정방법들이 있다 (Lehmann, 1951; Rosenblatt, 1952; Darling, 1957; Fisz, 1960; Watson, 1961, 1962; Anderson, 1962; Pearson, 1963; Burr, 1964; Stephens, 1965, 1970, 1976; Pettitt, 1976). 신용평가모형에서는 분류하고자하는 대상이 '불량'과 '정상' 두 집단으로 구분되기 때문에 신용평가모형의 판별력에 관한 많은 연구 (예를 들어 Hong 과 Suh, 2008; Hong 과 Choi, 2008 등 다수)가 있지만, 특히 검정방법으로 두 집단에 대응하는 두 분포함수의 동일성 검정방법이 이용된다. 부도율 (probability of default)의 함수인 스코어 (score)를 불량과 정상기업으로 구분하고 이에 대응하는 두 분포함수의 동일성을 검정하기 위하여 비모수적 방법인 Kolmogorov-Smirnov (K-S) 검정방법을 많이 사용한다. K-S 검정방법은 두 경험적인 분포함수 (empirical distribution)간의 차이가 최대인 점을 이용하여 두 집단에 대응하는 분포함수의 동일성을 검정하는 방법이며 통계적인 임계값 (critical value)은 유의수준과 두 집단의 표본크기에 의존한다. 표본크기가 큰 경우에는 K-S 통계량의 임계값이 작은 값을 갖는다. 이는 모형 판별력에 대한 검정이 민감해지기 때문에 동일한 두 분포함수의 귀무가설을 쉽게 기각하게 하는 문제를 야기한다. 그런데 신용평가모형에서 이용되는 표본크기는 매우 크다 (일반적으로 5,000 이상). 그러므로 K-S 통계량은 모든 신용평가모형이 판별력이 좋다고 결론 내리게 된다. 이러한 이유로 신용평가모형에 대한 판별력을 검정하는 경우 통계적인 임계값 이외의 다른 판단기

[1] 교신저자: (110-745) 서울시 종로구 명륜동 3가 53, 성균관대학교 통계학과, 교수.
E-mail: cshong@skku.ac.kr
[2] (110-745) 서울시 종로구 명륜동 3가 53, 성균관대학교 응용통계연구소, 연구원.

준의 필요성이 요구되었고 현재는 Joseph (2005)의 판단기준이 많이 적용되고 있다. 불량과 정상의 분포가 동일한 표준편차를 갖는 정규분포 가정 하에서 생성된 Joseph (2005)의 판단기준 중에서 두 분포함수를 검정하는 평균차이와 K-S 검정통계량만을 표 1.1에 정리하였다 (Wilkie, 2004 참조). 여기서 불량과 정상에 대한 스코어의 평균을 뺀 값에 두 기업의 표준편차가 동일하다고 가정하여 σ로 나눈 것이 평균차이 (MD)이다.

표 1.1 정규분포 가정에 기초한 K-S 통계량의 판단기준

의미	평균차이(MD)	K-S
Random	0.00	0.00
Doubtful	0.25	0.10
Poor	0.50	0.20
Marginal	0.75	0.29
Satisfactory	1.00	0.38
Good	1.25	0.47
Very Good	1.50	0.55
Strong	1.75	0.62
Very Strong	2.00	0.68
Excellent	2.25	0.74
Excellent	2.50	0.79
Excellent	2.75	0.83
Superior	3.00	0.87

Joseph (2005)의 판단기준은 표본크기가 큰 경우에 타당하다고 생각할 수 있다. 여기서 고려되어야 하는 다른 하나의 문제는 검정할 때 이용되는 두 표본의 크기의 차이이다. 신용평가모형 수립을 위해 수집된 불량표본은 과거 심사에서 정상으로 판단된 대상에서 발생하였기 때문에 3% 또는 5% 비율 정도로 매우 작은 경우가 일반적이다. 따라서 정상과 불량의 표본크기의 차이가 커지면 두 집단의 표준편차가 같다는 가정을 만족하기 어렵다.

앞에서 언급한 K-S 통계량의 적용시 문제점은 다른 통계량에서도 같은 문제를 야기한다. 이러한 면에서 정상과 불량 두 집단의 현재 공통으로 적용되고 있는 평가모형의 판단기준과 다르게 두 집단의 표본크기의 차이 등을 고려한 대안적인 판단기준의 필요성이 제기된다.

본 연구에서는 Joseph (2005)이 제안한 표 1.1의 판단기준인 K-S 통계량 외에 추가적으로 Cramér-Von Mises, Anderson-Darling, Watson 통계량들을 고려하여 판단기준을제시한다. 또한 박용석과 홍종선 (2008)의 유의수준, 두 집단의 표본크기의 차이 그리고 제Ⅱ종오류율을 고려한 대안이 되는 판단기준을 Cramér-Von Mises, Anderson-Darling, Watson 통계량의 연구를 확장하여 비교 토론하고자 한다.

본 논문의 구성은 다음과 같다. 2절에서는 비모수적인 Cramér-Von Mises W^2, Anderson-Darling A^2, Watson U^2통계량들에 대해서 설명한다. 3절에서는 Joseph (2005)이 제안한 판단기준에 세 가지 통계량의 판단기준을 확장하기 위하여, 표 1.1을 구하는 과정을 설명하고 세 가지 통계량의 판단기준을 제시한다. 4절에서는 박용석과 홍종선 (2008)의 연구를 확장한 대안적인 판단기준을 구하기 위하여 모의실험의 절차와 결과를 제시하며 활용방법을 설명한다. 그리고 3절과 4절에서 제안한 판단기준을 통합하여 토론한다. 5절에서 실제 사례를 통하여 두 판단기준의 적용방법과 차이점을 고려해보고, 마지막 6절에서는 본 연구의 결과에 대해서 정리하고 토론한다.

2. 비모수 검정통계량들

2.1. Cramér-Von Mises 통계량

$X_1, \cdots, X_n$과 $Y_1, \cdots, Y_m$이 각각 독립이고 두 집단의 누적분포함수가 $F(x), G(x)$인 두 확률표본일 때 Cram ér-Von Mises W^2통계량은 다음과 같이 정의한다 (Anderson-Darling, 1952; Darling, 1957; Fisz, 1960; Anderson, 1962; Burr, 1964; Stephens, 1970).

$$W^2 = [nm/N] \int_{-\infty}^{\infty} [F_n(x) - G_m(x)]^2 dH_{n+m}(x). \tag{2.1}$$

여기서 $F_n(x) = \sum I(X_i \le x)/n$, $G_m(x) = \sum I(Y_j \le x)/m$, 표본크기 $N = n + m$의 통합 자료 $X_1, \cdots, X_n$과 $Y_1, \cdots, Y_m$의 누적분포함수 $H_{n+m}(x)$는 가중값 $1/(n+m)$이 주어진 $(n+m)H_{n+m}(x) = nF_n(x) + mG_m(x)$이며 식 (2.1)을 다시 표현하면 다음과 같다.

$$W^2 = [nm/N^2] \left\{ \sum_{i=1}^{n} [F_n(x_i) - G_m(x_i)]^2 + \sum_{j=1}^{m} [F_n(y_j) - G_m(y_j)]^2 \right\}. \tag{2.2}$$

여기서 $X_1, \cdots, X_n$과 $Y_1, \cdots, Y_m$의 자료를 통합한 전체 관측값의 순위를 정하여 $X_1, \cdots, X_n$의 i번째 순위값을 r_i, $Y_1, \cdots, Y_m$의 j번째 순위값을 s_j라고 지정하자. 그러면 $F_n(x_i) - G_m(x_i) = i/n - (r_i - i)/m$이고 $F_n(y_j) - G_m(y_j) = (s_j - j)/n - j/m$이므로 식 (2.2)를 이용하여 Cramér-Von Mises W^2통계량은 다음과 같이 요약된다.

$$W^2 = \left\{ n/m \sum_{i=1}^{n} (r_i - Ni/n)^2 + m/n \sum_{j=1}^{m} (s_j - Nj/m)^2 \right\} / N^2. \tag{2.3}$$

2.2. Anderson-Darling 검정통계량

Anderson-Darling 통계량은 Cram ér-Von Mises 통계량과 유사한 방법으로 통계량 값을 구할 수 있다. Anderson-Darling A^2통계량은 다음과 같이 정의한다 (Pettitt, 1976).

$$A^2 = nm/N \int_{-\infty}^{\infty} \{F_n(x) - G_m(x)\}^2 / \{H_{n+m}(x)(1 - H_{n+m}(x))\} \, dH_{n+m}(x). \tag{2.4}$$

그리고 식 (2.4)를 다음과 같이 요약할 수 있다.

$$A^2 = 1/nm \sum_{i=1}^{N-1} (M_i N - ni)^2 / \{i(N-i)\}. \tag{2.5}$$

여기서 $N = n + m$, $M_i = nF_n \circ H_N^{-1}(i/N)$, 그리고 $H_N^{-1}(t) = inf\{x : H_N(x) = t\}$이다.

Anderson-Darling A^2통계량은 분포의 꼬리부분에서 Cramér-Von Mises 통계량보다 더 높은 값을 갖는다.

2.3. Watson 검정통계량

Watson U^2검정통계량은 다음과 같이 정의한다 (Watson, 1961, 1962; Stephens, 1965, 1976).

$$U^2 = nm/N \int_{-\infty}^{\infty} \left\{ F_n(x) - G_m(x) - \int_{-\infty}^{\infty} [F_n(y) - G_m(y)] \, dH(y) \right\}^2 dH(x). \tag{2.6}$$

식 (2.6)을 요약하면 다음과 같다.

$$U^2 = nm/N^2 \left[\begin{array}{c} \sum_{i=1}^{n} [F_n(x_i) - G_m(x_i)]^2 + \sum_{j=1}^{m} [F_n(y_j) - G_m(y_j)]^2 \\ - \left\{ \sum_{i=1}^{n} [F_n(x_i) - G_m(x_i)] + \sum_{j=1}^{m} [F_n(y_j) - G_m(y_j)] \right\}^2 /N \end{array} \right]. \tag{2.7}$$

2.1절에서와 같이 $F_n(x_i) - G_m(x_i) = i/n - (r_i - i)/m$와 $F_n(y_j) - G_m(y_j) = (s_j - j)/n - j/m$으로 나타낼 수 있으므로 식 (2.7)을 다음과 같이 정리할 수 있다.

$$U^2 = nm/N^2 \left[\begin{array}{c} \sum_{i=1}^{n} [i/n - (r_i - i)/m]^2 + \sum_{j=1}^{m} [(s_j - j) - j/m]^2 \\ - \left\{ \sum_{i=1}^{n} [i/n - (r_i - i)/m] + \sum_{j=1}^{m} [(s_j - j)/n - j/m] \right\}^2 /N \end{array} \right].$$

3. 판단기준 제안 I

현재 신용평가모형의 판별력을 판단하는 기준은 Joseph (2005)이 제안한 판단기준이 많이 사용되고 그 중에서 분포함수의 동일성을 비교 검정하는 방법으로는 동일한 표준편차를 갖는 정규분포를 가정한 모수적인 접근 방법으로 K-S 검정을 표 1.1과 같이 사용한다. 여기서는 Joseph이 설정한 상황과 동일한 정규분포하에서 2절에서 소개한 Cramér-Von Mises, Anderson-Darling, Watson 검정통계량들의 판별력 판단기준을 제시한다.

우선 모의실험 과정은 다음과 같다.

1. 표준정규분포 $N(0,1)$로부터 10,000개의 난수를 생성하여 부도기업의 스코어로 간주한다.
2. 정규분포 $N(\Delta,1)$로부터 Δ의 값을 0에서 0.25씩 늘려가며 3까지 13구간의 평균값을 생성한다. 부도기업의 스코어와 같은 10,000개의 난수를 생성하여 정상기업의 스코어로 간주한다.
3. 2장에서 정의한 Cramér-Von Mises, Anderson-Darling, Watson 통계량을 계산하고 10,000번 반복한다. 여기서 대표본인 경우의 K-S 통계량은 임계값에 $\sqrt{(n+m)/nm}$을 곱한 대표본 근사 임계값을 사용하듯이 세가지 통계량의 평균과 백분위수에도 표본크기와 무관하게 N/nm을 곱하여 대표본 근사 임계값을 구한다.

위에서 언급한 과정에 따라 구한 통계량들의 결과는 표 3.1과 같다.

표 3.1 검정통계량들의 판단기준 I

의미	MD	W^2	A^2	U^2
Random	0.0000	0.0000	0.0000	0.0000
Doubtful	0.2500	0.0057	0.0299	0.0008
Poor	0.5000	0.0222	0.1152	0.0031
Marginal	0.7500	0.0479	0.2456	0.0070
Satisfactory	1.0000	0.0803	0.4069	0.0125
Good	1.2500	0.1165	0.5826	0.0194
Very Good	1.5000	0.1538	0.7590	0.0274
Strong	1.7500	0.1897	0.9245	0.0360
Very Strong	2.0000	0.2222	1.0709	0.0446
Excellent	2.2500	0.2502	1.1947	0.0529
Excellent	2.5000	0.2732	1.2940	0.0602
Excellent	2.7500	0.2912	1.3709	0.0665
Superior	3.0000	0.3048	1.4280	0.0715

평균차이 (MD)의 값이 1일 때의 의미는 Satisfactory로서 이 값에 대응하는 Cramér-Von Mises W^2통계량은 0.0803, Anderson-Darling A^2통계량은 0.4069, Watson U^2통계량은 0.0125의 값을 갖는다. 즉 이 값들보다 크면 신용평가모형의 판별력이 Satisfactory하다고 판단할 수 있다. 그리고 평균차이의 값 2를 기준으로 W^2통계량은 0.2222, A^2통계량은 1.0709, U^2통계량은 0.0446으로 이 값들보다 큰 값을 가지면, 신용평가모형의 판별력이 Very Strong하다고 판단할 수 있겠다.

4. 판단기준 제안 II

Joseph (2005)이 제시한 표 1.1과 3절에서 제시한 표 3.1의 판단기준 모두는 신용평가 자료의 특성인 이질적인 표준편차, 그리고 불량률과 표본크기 등을 고려하지 않는 정규분포의 평균차이만을 기반으로 한 판단기준이다. 여기서는 불량률, 표본크기 그리고 제II종 오류율을 고려한 박용석과 홍종선 (2008)의 연구를 확장하여 Cramér-Von Mises, Anderson-Darling, Watson 검정통계량들의 판단기준을 제시한다.

모의실험 과정은 다음과 같다.

1. 표준정규분포 $N(0,1)$으로부터 N개의 난수를 생성하여 N개의 스코어로 간주한다 ($N = 1,000, 5,000, 10,000$). 신용평가모형 구축에 사용되는 자료는 기업체의 경우 대기업은 약 5,000개 정도의 표본이 이용되고 그 밖의 소규모 업체나 개인에 대한 평가모형은 더 많은 자료를 이용한다. 그러므로 신용평가 모형의 표본크기를 대표하기 위해서 신용평가영역에서 비교적 작은 표본크기인 1,000, 5,000, 10,000까지의 표본크기를 설정 하였다.
2. 1에서 생성된 스코어를 크기 순서대로 나열했을 때 표본불량률 r에 대응하는 자료를 $x_r = \Phi^{-1}(r)$ 을 경계로 스코어가 x_r이하인 $n' \approx Nr$자료를 생성하고 n'중에서 $n \approx Np$ $(p < r)$개의 불량을 임의로 추출하여 실제불량으로 설정한다 ($p = 0.03, 0.05$ 이고 $r = 0.10$, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70). 예를 들면 1,000개의 자료에 불량률이 0.05인 경우 불량의 개수는 50개이다. r이 0.1인 경우 $n' = 100$개가 되므로 100개의 자료 중에 불량 50개와 정상 50개가 혼합되어 오분류를 만든다.
3. 통계량들을 계산하고 I = 0.05를 만족하는 "분류기준 스코어" x_c를 경계로 스코어가 x_c이하인 자료를 불량으로 예측하여 제II종 오류율을 생성한다. 여기서 p는 전체 기업중에 실제 불량 기업의 비율이고 제 I 종 오류율은 실제 불량을 정상으로 예측할 확률이며 제II종 오류율은 실제 정상을 불량으로 예측할 확률이다.
4. 위 1~3의 과정을 10,000번씩 수행한다.
5. 표본불량률 $r = 0.10 \sim 0.70$의 경우에 따라 자료를 생성한 후 생성된 자료들을 하나로 통합하고, 제II종오류율 10, 20, 30, 40, 50, 60%를 기준으로 Cramér-Von Mises, Anderson-Darling, Watson 통계량에 대한 각각의 평균과 90과 95 백분위수를 산출한다. 통계량의 평균과 백분위수에도 표본크기와 무관하게 N/nm을 곱하여 대표본 근사 임계값을 구한다.

실제 현장에서 불량의 수는 정상의 수보다 매우 작기 때문에 불량률 p를 3%와 5%로 설정하여 정상과 부도의 수를 97:3와 95:5의 비율로 고려하였으며, 제 I 종 오류의 위험 (risk)은 제II종 오류의 위험보다 매우 크기 때문에 제 I 종 오류율을 5%로 통제하였다. 모의실험에서 제II종오류율이 60%이상의 값도 구할 수 있지만 여기서는 60%까지 적용하기로 하겠다. 위의 모의실험 과정을 통해 구한 결과를 불량률 $p = 0.03$과 0.05에 따라 W^2은 평균을 $W^2_{.90}$은 90 백분위수를 $W^2_{.95}$는 95 백분위수로 표 4.1과 4.2에 정리하였다.

먼저 $p = 0.03$인 표 4.1의 결과를 살펴보자. Cramér-Von Mises W^2, Anderson-Darling A^2, Watson U^2통계량의 평균은 제Ⅱ종 오류율이 증가함에 따라서 선형적으로 감소한다. 예를 들어 W^2의 경우 표본크기 $N = 1,000$에서 제Ⅱ종 오류율이 10%인 경우 0.2866, 20%인 경우 0.2271, 60%인 경우 0.0599로 선형적으로 감소한다. A_2과 U^2통계량도 W^2통계량과 유사하게 제Ⅱ종 오류율이 증가함에 따라서 선형적으로 감소한다. 세 가지 통계량들의 평균은 표본크기에 따라서 큰 차이가 없지만 90과 95 백분위수는 표본크기가 커짐에 따라서 작아진다. 예를 들어 A^2통계량을 살펴보면 평균값은 제Ⅱ종 오류율 20%에서 $N = 1,000$일 때 1.2675, $N = 10,000$일 때 1.2553으로 약간 감소하지만 큰 차이가 없다. 반면에 동일한 조건에서 95 백분위수는 각각 1.6798, 1.3802로 그 차이가 큰 것을 알 수 있다. 따라서 백분위수를 대안적인 판단기준으로 고려하는 경우 표본크기를 고려해야 한다.

$p = 0.05$인 표 4.2의 결과는 $p = 0.03$일 때인 표 4.1에서의 통계량과 유사한 현상을 나타내지만, $p = 0.03$일 때 보다 통계량들의 평균은 크고 백분위수는 작다. 예를 들어 U^2통계량을 살펴보면 $N = 1,000$이고 제Ⅱ종 오류율이 30%인 경우 $p = 0.03$일 때의 평균과 95 백분위수는 각각 0.0449와 0.0594인 반면에 $p = 0.05$일 때의 평균과 95 백분위수는 각각 0.0458과 0.0572로 $p = 0.03$일 때 보다 통계량들의 평균은 크고 백분위수는 작다.

표 3.1은 표본크기와 불량률 등을 고려하지 않은 일괄적인 판단기준인 반면에 표 4.1과 표 4.2는 불량률, 표본크기 그리고 제Ⅱ종 오류율의 정보를 포함하기 때문에, 표 4.1과 표 4.2에 제시한 90과 95 백분위수를 각각 유의수준 10%와 5%에서의 임계값으로 간주하여 박용석과 홍종선 (2008)에서 제안한 가설검정을 할 수 있다. 대안적인 판단기준의 적용방법은 다음과 같다. 모형 생성을 위해 이용된 자료의 크기가 5,000이고 그 중에서 불량의 개수가 260개인 사례를 고려해보면 추정볼량률은 0.05에 근사함으로 $p = 0.05$인 표 4.2의 N=5,000에 대한 결과를 토대로 가설 검정한다. 예를 들어 신용평가모형에서 생성된 W^2, A^2, U^2통계량이 각각 0.12, 0.48, 0.03이라고 하면, W^2통계량은 표 4.2의 95 백분위수를 기준으로 제Ⅱ종 오류율 40%와 50% 사이의 값을 나타내므로 제Ⅱ종 오류율의 허용범위 50% 하에서 모형의 판별력이 좋다고 결론내릴 수 있다. 그리고 A^2과 U^2통계량도 같은 방법으로 제Ⅱ종 오류율 허용범위 50%하에서 모형의 판별력이 좋다고 결론내릴 수 있다. 또한 박용석과 홍종선 (2008)이 제안한 기대비용함수를 고려하여 제Ⅱ종 오류율을 설정하면 신용평가모형의 가설검정이 가능하다.

표 4.1 통계량들의 판단기준 Ⅱ(불량률 $p = 0.03$)

표본크기	제Ⅱ종오류율	W^2	$W^2_{.90}$	$W^2_{.95}$	A^2	$A^2_{.90}$	$A^2_{.95}$	U^2	$U^2_{.90}$	$U^2_{.95}$
1,000	10%	0.2866	0.3524	0.3711	1.9589	2.3972	2.5144	0.0720	0.0883	0.0930
	20%	0.2271	0.2803	0.2964	1.2675	1.5803	1.6798	0.0575	0.0706	0.0747
	30%	0.1756	0.2190	0.2329	0.8840	1.1206	1.1944	0.0449	0.0559	0.0594
	40%	0.1302	0.1651	0.1764	0.6226	0.8060	0.8659	0.0338	0.0425	0.0453
	50%	0.0916	0.1196	0.1294	0.4322	0.5758	0.6236	0.0244	0.0315	0.0340
	60%	0.0599	0.0828	0.0908	0.2883	0.3995	0.4411	0.0165	0.0224	0.0243
5,000	10%	0.2872	0.3164	0.3249	1.9553	2.1478	2.2051	0.0718	0.0791	0.0813
	20%	0.2269	0.2507	0.2575	1.2571	1.3937	1.4361	0.0569	0.0627	0.0645
	30%	0.1738	0.1926	0.1980	0.8666	0.9681	1.0001	0.0437	0.0484	0.0498
	40%	0.1279	0.1430	0.1479	0.6046	0.6830	0.7070	0.0322	0.0360	0.0372
	50%	0.0892	0.1017	0.1056	0.4150	0.4773	0.4968	0.0226	0.0258	0.0267
	60%	0.0572	0.0672	0.0702	0.2710	0.3186	0.3328	0.0146	0.0171	0.0179
10,000	10%	0.2872	0.3079	0.3137	2.0237	2.0916	2.1271	0.0718	0.0770	0.0784
	20%	0.2268	0.2435	0.2484	1.2553	1.3522	1.3802	0.0568	0.0609	0.0622
	30%	0.1737	0.1875	0.1911	0.8652	0.9398	0.9599	0.0435	0.0470	0.0479
	40%	0.1277	0.1388	0.1421	0.6028	0.6598	0.6768	0.0321	0.0348	0.0356
	50%	0.0888	0.0976	0.1002	0.4124	0.4560	0.4698	0.0223	0.0245	0.0252
	60%	0.0570	0.0641	0.0663	0.2694	0.3031	0.3138	0.0144	0.0162	0.0168

표 4.2 통계량들의 판단기준 Ⅱ(불량률 $p = 0.05$)

표본크기	제Ⅱ종오류율	W^2	$W^2_{.90}$	$W^2_{.95}$	A^2	$A^2_{.90}$	$A^2_{.95}$	U^2	$U^2_{.90}$	$U^2_{.95}$
1,000	10%	0.2993	0.3500	0.3646	2.0402	2.3532	2.4414	0.0749	0.0875	0.0913
	20%	0.2371	0.2796	0.2923	1.3180	1.5593	1.6365	0.0596	0.0701	0.0732
	30%	0.1811	0.2156	0.2262	0.9067	1.0883	1.1512	0.0458	0.0546	0.0572
	40%	0.1337	0.1618	0.1705	0.6356	0.7815	0.8281	0.0341	0.0413	0.0433
	50%	0.0935	0.1165	0.1236	0.4378	0.5523	0.5869	0.0242	0.0299	0.0319
	60%	0.0605	0.0792	0.0847	0.2892	0.3782	0.4068	0.0160	0.0207	0.0223
5,000	10%	0.2987	0.3214	0.3279	2.0330	2.1733	2.2140	0.0747	0.0804	0.0821
	20%	0.2363	0.2553	0.2606	1.3080	1.4162	1.4474	0.0592	0.0638	0.0652
	30%	0.1812	0.1964	0.2006	0.9024	0.9840	1.0058	0.0454	0.0492	0.0502
	40%	0.1333	0.1456	0.1493	0.6296	0.6925	0.7114	0.0335	0.0366	0.0374
	50%	0.0927	0.1028	0.1058	0.4305	0.4801	0.4945	0.0233	0.0258	0.0266
	60%	0.0596	0.0676	0.0699	0.2815	0.3284	0.3316	0.0150	0.0171	0.0177
10,000	10%	0.2993	0.3155	0.3202	2.0365	2.1370	2.1652	0.0748	0.0789	0.0800
	20%	0.2364	0.2495	0.2534	1.3080	1.3835	1.4053	0.0591	0.0624	0.0634
	30%	0.1810	0.1917	0.1948	0.9009	0.9575	0.9748	0.0453	0.0480	0.0487
	40%	0.1331	0.1418	0.1443	0.6280	0.6727	0.6857	0.0333	0.0355	0.0361
	50%	0.0925	0.0996	0.1016	0.4291	0.4641	0.4744	0.0232	0.0250	0.0255
	60%	0.0593	0.0650	0.0666	0.2799	0.3067	0.3150	0.0149	0.0163	0.0168

그림 4.1은 Joseph (2005)의 적정성 여부 판단기준인 표 3.1의 각 통계량의 'Satisfactory'의 값과 4절에서 제안된 판단기준인 표 4.1과 표 4.2의 95 백분위수 기준값이다. 각 그림에서 실선은 Joseph의 기준에 근거한 값이고, 검은색 점과 'x' 표시한 점은 $p = 0.03$과 $p = 0.05$에서 해당 통계량들의 95 백분위수를 나타낸다. 그림에서 보는 것처럼 4절에서 제시된 판단기준은 표본크기에 따라서 달라진다. 적정성 여부의 판단기준은 W^2통계량이 표 3.1에서의 통계량과 유사한 지점을 기준으로 선택된다. 따라서 현재 적용되고 있는 'Satisfactory' 지점은 약 60% 지점을 경계로 설정된다. 먼저 W^2통계량의 결과를 살펴보면 Joseph의 기준에 의한 판단기준은 0.08이고 표본크기가 비교적 작은 1,000개에서는 Joseph의 기준보다 높은 기준이 필요하고 표본크기가 큰 5,000과 10,000개에서는 Joseph의 기준보다 낮아야함을 나타낸다. A^2통계량의 결과는 W^2의 결과와 동일하다. 반면에 U^2통계량의 결과는 조금 다른데 모든 표본크기에서 Joseph의 기준보다 높은 기준이 필요함을 나타낸다. $p = 0.03$과 $p = 0.05$간의 판단기준의 차이는 소표본에서 크고 표본의 크기가 커짐에 따라서 두 불량률간의 판단기준 값의 차이는 거의 없다.

그림 4.2는 표 3.1에서 제시한 판단기준과 오분류율에 따라 모의실험을 하여 얻은 표본크기가 5,000개일 때의 표 4.1과 표 4.2의 판단기준을 각각 비교한 그림이다. 세 종류의 통계량에 대하여 왼쪽그림은 $p = 0.03$의 경우를 나타내고 오른쪽 그림은 $p = 0.05$인 경우이며, 표 3.1의 판단기준으로부터 얻은 각 통계량의 Satisfactory와 Very Strong에 해당하는 값을 참고선 (reference line)으로 나타내고, 표 4.1과 표 4.2에서 구한 각 통계량의 평균과 95 백분위수를 제Ⅱ종 오류율에 따라 각각 실선과 점선으로 나타내었다. 그림 4.2에서 보는 것처럼 불량률 $p = 0.03$과 0.05의 경우인 W^2, A^2, U^2통계량에 대응하는 곡선의 형태를 통해 표 4.1과 표 4.2의 결과가 유사하다는 것을 확인할 수 있다.

우선 Cramér-Von Mises W^2의 결과를 살펴보자. 표 3.1에서 Satisfactory와 Very Strong하다는 의미를 가질 때 W^2통계량은 0.0803과 0.2222 사이의 값을 가진다. 이 결과는 표 4.1과 표 4.2에서 얻은 W^2통계량의 평균값에 의한 제Ⅱ종 오류율 20%와 50% 사이의 값에 대응하고, 95 백분위수는 제Ⅱ종 오류율 25%와 55% 사이의 값에 대응한다. 그러므로 W^2통계량이 0.0803 이상의 값을 가진다면, 제Ⅱ종 오류율 55% 이상으로는 귀무가설을 기각할 수 없으며, 0.0803과 0.2222 사이의 값을 가진다면, 제Ⅱ종 오류율 25% $\sim$ 55% 정도에서 귀무가설을 기각할 수 있겠다. 마찬가지로 A^2의 결과를 살펴보면,

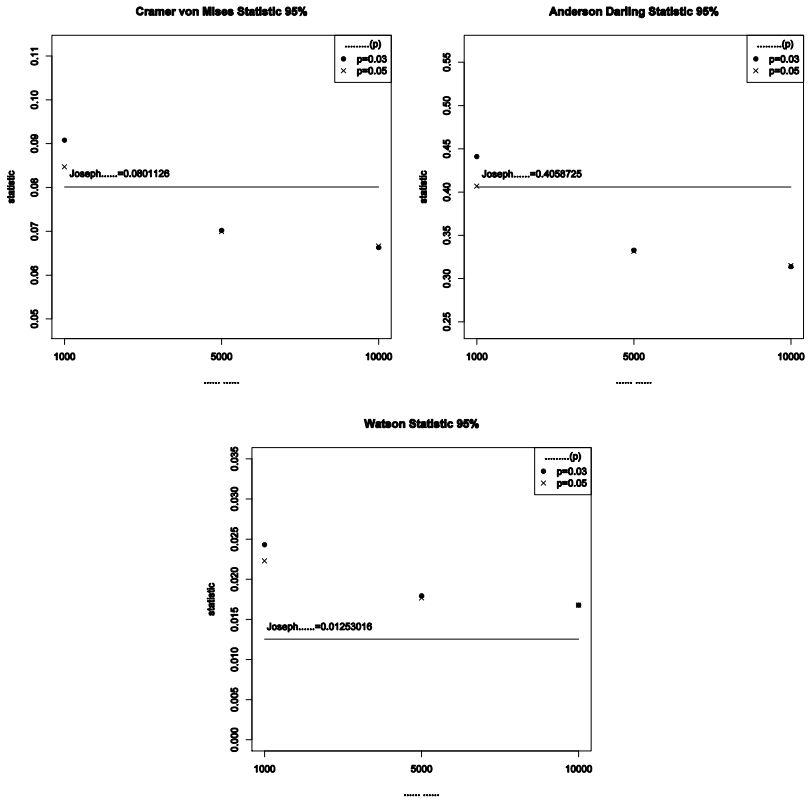

그림 4.1 비모수적 통계량들에 대한 적정성여부 판단기준점

표 3.1에서 Satisfactory와 Very Strong하다는 의미를 기준으로 할 때 A^2통계량은 0.4069와 1.0709 사이의 값을 가진다. 이 결과는 표 4.1과 표 4.2에서 얻은 A^2통계량의 평균값과 비교할 때 제Ⅱ종 오류율 25%와 50% 사이의 값에 대응하고, 95 백분위수는 제Ⅱ종 오류율 30%와 55% 사이의 값에 대응한다. 그러므로 A^2통계량이 0.4069 이상의 값을 가질 때 제Ⅱ종 오류율 55% 이상으로는 귀무가설을 기각할 수 없으며, 0.4069와 1.0709 사이의 값을 가질 때 제Ⅱ종 오류율 30% $\sim$ 55%정도에서 귀무가설을 기각할 수 있다.

표 3.1에서 Satisfactory와 Very Strong하다는 의미를 가질 때 U^2통계량은 0.0125와 0.0446 사이의 값을 가진다. 이 결과는 표 4.1과 표 4.2에서 얻은 U^2통계량의 평균값에 의한 제Ⅱ종 오류율 30%와 60% 사이의 값에 대응하고 95 백분위수는 제Ⅱ종 오류율 60% 이상의 값에 대응한다. U^2통계량이 0.0125 이상의 값을 가질 때 제Ⅱ종 오류율 60% 이하에서 귀무가설을 기각하며, U^2통계량이 0.0446 이하의 값을 가질 때 제Ⅱ종 오류율 30%에서 60% 보다 더 큰 오류율까지도 귀무가설을 기각할 수 있다. 그러므로 3절과 4절에서 제안한 판단기준은 통합적으로 표현한 그림 4.2를 통해 신용평가모형의 적합성을 설명하는데 활용할 수 있다.

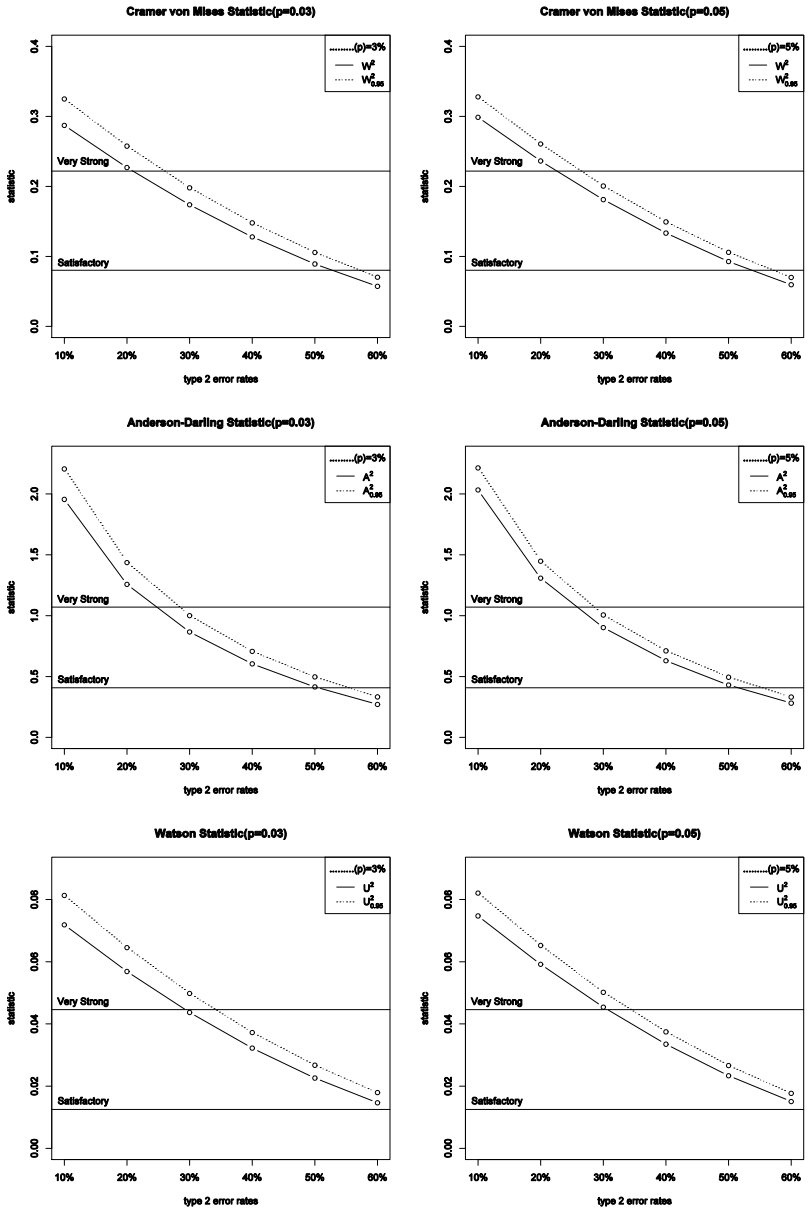

그림 **4.2** 판단기준 Ⅰ과 Ⅱ의 비교분석

5. 사례분석

국내 K금융회사에서는 1994년부터 2005년까지 은행연합회 불량발생 등록된 자료를 수집하여 신용평가모형을 개발하였다. 본 연구에서는 이 자료 중에서 매출액 1,000억 이상의 외감 대기업들

4,268건 (정상: 4101건, 부도: 167건)의 재무자료를 바탕으로 개발한 신용평가모형의 예측 부도율 자료를 살펴본다. 불량률은 모형생성을 위해 수집된 전체자료 중에서 불량이 차지하는 비율로써 $\hat{p} = 167/4,268 = 0.039$이므로 $p = 0.03$의 기준을 적용하자. 실제 신용평가모형에서 표본크기로 조정되어 산출된 Cramér-Von Mises W^2통계량은 0.1942, Anderson-Darling A^2통계량은 1.3753이고 Watson U^2통계량은 0.0413을 나타낸다.

본 연구의 서론에서 언급하였듯이 이 값들은 Watson (1962)과 Choulakian 등 (1994)에서 제시한 임계값보다 매우 큰 값을 갖기 때문에 두 집단의 분포함수가 동일하다는 귀무가설은 모두 기각할 수 있다. 그러나 표 3.1의 판단기준에 의하면, W^2과 U^2통계량은 Strong하고, A^2통계량은 Excellent이상이라고 판단할 수 있다. 또한 표 4.1의 $N = 5,000$인 경우와 비교하여 살펴보면, 유의수준 5%에서 W^2과 U^2통계량은 제Ⅱ종 오류율 40% 수준에서 귀무가설을 기각할 수 있으며, A^2통계량은 제Ⅱ종 오류율 30% 수준일 때 귀무가설을 기각하여 모형의 판별력이 좋다고 판단할 수 있다. 그림 4.2를 통해서 표 3.1과 표 4.1의 결과를 동시에 비교 분석할 수 있다. 그림 4.2에서 W^2과 U^2통계량은 제Ⅱ종 오류율의 30% ∼ 40% 사이의 값이므로 제Ⅱ종 오류율 40% 수준에서 귀무가설을 기각할 수 있으며, Joseph (2005)에 의한 판단기준인 표 3.1의 Satisfactory와 Very Strong사이의 값이라는 것을 식별할 수 있다. 이 결과는 박용석과 홍종선 (2008)의 연구에서 제안한 K-S 통계량의 판단기준을 바탕으로 유도한 결과와 동일하다. 또한 A^2통계량은 제Ⅱ종 오류율의 20% ∼ 30% 사이의 값이므로 제Ⅱ종 오류율 30% 수준에서 귀무가설을 기각할 수 있으며, Joseph (2005)에 의한 판단기준인 표 3.1의 Very Strong보다 큰 값을 가지므로 더 큰 판별력을 가진다는 것을 탐색할 수 있다.

6. 결론

본 연구에서는 신용평가모형 영역에서 두 집단의 판별력 검정방법 중의 하나로 분포함수의 동일성 검정을 위해 널리 사용되고 있는 비모수적인 검정통계량인 K-S 통계량에 추가적으로 Cramér-Von Mises, Anderson-Darling, Watson 통계량을 소개하였다. 이 통계량들의 판단기준을 설정하기 위하여 Joseph (2005)의 연구를 확장하여 Cramér-Von Mises, Anderson-Darling, Watson 통계량의 판단기준을 제시하였다. 그리고 신용평가 자료와 유사한 다양한 표본크기 N과 불량률 p그리고 제Ⅱ종 오류율 정보에 근거하여 자료를 생성한 박용석과 홍종선 (2008)의 연구를 확장하여 대안적인 판단기준을 제시하였으며 임계값으로 사용할 수 있는 90과 95 백분위수도 제시하였다. 불량률과 표본크기는 사전에 알 수 있고 제Ⅱ종 오류율의 정보는 각 사용자마다 허용범위를 고려하여 본 연구에서 제안한 기준을 추가적인 정보로 사용하면, 신용평가모형의 적합성 여부를 판단하는데 유용하게 활용할 수 있다.

참고문헌

박용석, 홍종선 (2008). 신용평가모형에서 콜모고로프-스미르노프 검정기준의 문제점. <한국통계학회 논문집>, **15**, 1013-1026.

송문섭, 박창순, 이정진 (2003). <S-LINK를 이용한 비모수통계학>, 자유아카데미.

홍종선, 이창혁, 김지훈 (2008). 범주형 재무자료에 대한 신용평가모형 검증 비교. <한국통계학회 논문집>, **15**, 615-631.

Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain "goodness-of-fit" criteria based on stochastic processes. *The Annals of Mathematical Statistics,* **23**, 193-212.

Anderson, T. W. (1962). On the distribution of the two-sample Cramer-von mises criterion. *The Annals of Mathematical Statistics,* **33**, 1148-1159.

Burr. E. (1964). Small-sample distributions of the two-sample Cramer-Von Mises' $W^2$and Watson's U^2. *The Annals of Mathematical Statistics,* **35**, 1091-1098.

Choulakian, V., Lockhart, R. A. and Stephens, M. A. (1994). Cramer-Von Mises statistics for discrete distributions. *The Canadian Journal of Statistics / La Revue Canadiennede Statistique*, **22**, 125-137.

Darling, D. A. (1957). The Kolmogorov-Smirno, Cramer-von Mises tests. *Annals of Mathematical Statistics,* **31**, 823-838.

Fisz, M. (1960). On a result by M. Rosenblatt concerning the Mises-Smirnov test. *Annals of Mathematical Statistics,* **31**. 427-429.

Hong, C. S. and Choi, J. M. (2008). Validation comparison of credit rating models using Box-Cox transformation. *Journal of the Korean data & Information Science Society*, **19**, 789-801.

Hong, Y. W. and Suh, J. S. (2008). Estimating the credit value-at risk of korean property and casuality insurers. *Journal of the Korean data & Information Science Society*, **19**, 1027-1036.

Joseph, M. P. (2005). A PD validation framework for basel Ⅱ internal ratings-based systems. *Credit Scoring and Credit Control Ⅳ.*

Lehmann, E. L. (1951). Consistency and unbiasedness of certain nonparametric tests. *Annals of Mathematical Statistics,* **22**, 165-179.

Pearson, E. S. (1963). Comparison of tests for randomness of points on a line. *Biometrika,* **50**, 315-325.

Pettitt, A. (1976). A two-sample Anderson_Darling rank statistic. *Biometrika,* **63**, 161-168.

Rosenblatt, M. (1952). Limit theorems associated with variants of the von Mises statistic. *Annals of Mathematical Statistics,* **23**, 617-623.

Stephens, M. A. (1965). Significance points for the two-sample statistic $U^2_{M,N}$, *Biometrika*, **52**, 661-663.

Stephens, M. A. (1970). Use of the Kolmogorov-Smirnov, Cramer-Von Mises and related statistics without extensive tables. *Journal of the Royal Statistical society,* **32**, 115-122.

Stephens, M. A. (1976). Asymptotic results for goodness-of-fit statistics with unknown parameters. *The Annals of Statistics,* **4**, 357-369.

Watson, G. S. (1961). Goodness-of-fit tests on a circle. *Biometrika*, **48**, 109-114.

Watson, G. (1962). Goodness-of-fit tests on a circle Ⅱ. *Biometrika,* **49**, 57-63.

Wilkie, A. D. (2004). Measures for comparing scoring systems. *In Readings in Credit Scoring-recent developments, advances, and aims,* Eds. Thomas, L. C., Crook, J. N, and Edelman, D. B. Oxford finance.

Journal of the Korean Data & Information Science Society
2009, **20**(2), 261–272

한국데이터정보과학회지

Nonparametric homogeneity tests of two distributions for credit rating model validation

Chong Sun Hong[1] · Ji Hoon Kim[2]

[1]Department of Statistics, Sungkyunkwan University
[2]Research Institute of Applied Statistics, Sungkyunkwan University

Received 7 January 2009, revised 19 February 2009, accepted 3 March 2009

Abstract

Kolmogorov-Smirnov (K-S) statistic has been widely used for testing homogeneity of two distributions in the credit rating models. Joseph (2005) used K-S statistic to obtain validation criteria which is most well-known. There are other homogeneity test statistics such as the Cramer-von Mises, Anderson-Darling, and Watson statistics. In this paper, these statistics are introduced and applied to obtain criterion of these statistics by extending Joseph (2005)'s work. Another set of alternative criterion is suggested according to various sample sizes, type Ⅱ error rates, and the ratios of bads and goods by using the simulated data under the similar situation as real credit rating data. We compare and explore among Joseph's criteria and two sets of the proposed criterion and discuss their applications.

Keywords: Credit rating model, discriminatory power, distribution function, nonparametric, validation.

[1] Corresponding author: Professor, Department of Statistics, Sungkyunkwan University, Seoul 110-745, Korea. E-mail: cshong@skku.ac.kr
[2] Researcher, Research Institute of Applied Statistics, Sungkyunkwan University, Seoul 110-745, Korea.