The Korean Journal of Applied Statistics (2019) DOI: http://dx.doi.org/10.5351/KJAS.2019.32.4.593
32(4), 593–605

AUC and VUS using truncated distributions

Chong Sun Hong[a,1] · Seong Hyuk Hong[a]

[a]Department of Statistics, Sungkyunkwan University

(Received May 16, 2019; Revised June 14, 2019; Accepted June 28, 2019)

Abstract

Significant literature exists on the area under the ROC curve (AUC) and the volume under the ROC surface (VUS) which are statistical measures of the discriminant power of classification models. Whereas the partial AUC is restricted on the false positive rate, the two-way partial AUC is restricted on both the false positive rate and true positive rate, which could be more efficient and accurate than partial AUC. The two-way partial AUC was suggested as more efficient and accurate than the partial AUC. Partial VUS as well as the three-way partial VUS were also developed for the ROC surface. A proposed AUC is expressed in this paper with probability and integration using two truncated distribution functions restricted on both the false positive rate and true positive rate. It is also found that this AUC has a relation with the two-way partial AUC. The three-way partial VUS for the ROC surface is also related to the VUS using truncated distribution functions. These AUC and VUS are represented and estimated in terms of Mann-Whitney statistics. Their parametric and non-parametric estimation methods are explored based on normal distributions and random samples.

Keywords: classification, discrimination, FPR, TPR, truncation

1. 서론

의학진단이나 신용평가 분야에서 세 범주를 분류하기 위한 X_1, X_2, X_3를 스코어 확률변수라 하면 이들의 누적분포함수를 각각 $F_1(\cdot), F_2(\cdot), F_3(\cdot)$ ($F_1(x) \geq F_2(x) \geq F_3(x)$, $x \in (-\infty, \infty)$)로 가정한다. 잘 알려진 receiver operating characteristic (ROC) 곡선(curve)은 임의의 x에 대하여 true positive rate (TPR) $F_1(x)$와 false positive rate (FPR) $F_2(x)$를 각각 이차원 평면의 Y축과 X축 좌표에 대응시킨 그래프로 식 (1.1)과 같이 표현된다 (Metz, 1978; Zweig와 Campbell, 1993; Greiner 등, 2000; Tasche, 2006).

$$(F_2(x), F_1(x)) = (p, \mathrm{ROC}(p)), \tag{1.1}$$

여기서 $\mathrm{ROC}(p) = F_1(F_2^{-1}(p))$, $p \in (0,1)$이다. ROC 곡선에 대한 판별력을 측정하는 통계량으로 ROC 곡선 아래의 면적을 나타내는 area under the ROC curve (AUC)는 식 (1.2)와 같이 확률과 적분식으로 표현된다 (Bradley, 1997; Krzanowski와 Hand, 2009).

$$\mathrm{AUC} = P(X_1 \leq X_2) = \int_0^1 F_1\left(F_2^{-1}(p)\right) dp. \tag{1.2}$$

[1]Corresponding author: Department of Statistics, Sungkyunkwan University, 25-2, Sungkyunkwan-Ro, Jongno-Gu, Seoul 03063, Korea. E-mail: cshong@skku.edu

부분 AUC (partial AUC; pAUC)는 절단점 $x_L, x_U (x_L < x_U)$ 사이의 면적을 나타내며 $F_2(x_L) = v_1$, $F_2(x_U) = v_2$에 대하여 다음과 같이 확률로 표현된다 (McClish, 1989; Thompson와 Zucchini, 1989; Jiang 등, 1996; Hong과 Cho, 2019).

$$\text{pAUC}(v_1, v_2) = P(X_1 \le X_2 \cap x_L \le X_2 \le x_U).$$

pAUC는 FPR의 범위를 제한하여 AUC 중의 일부 면적을 구한 통계량이다. 실제의 성능 평가 진단에서는 낮은 FPR과 높은 TPR을 함께 고려하는 것이 중요할 수 있다. Yang 등 (2019)은 이러한 문제점을 개선하고자 TPR과 FPR을 동시에 제한하여 성능을 평가하기 위해 양방향 부분 AUC (two-way partial AUC; tpAUC)를 제안하였고, Hong 등 (2019)은 tpAUC를 식 (1.3)과 같이 확률 개념으로 그리고 식 (1.4)와 같이 적분식으로 표현하였다.

$$\begin{aligned} \text{tpAUC}(v_1, v_2) &= P(x_L \le X_1 \le X_2 \le x_U) \qquad (1.3)\\ &= \text{pAUC}(v_1, v_2) - P(X_1 \le x_L \le X_2 \le x_U)\\ &= \int_{v_1}^{v_2} F_1\left(F_2^{-1}(p)\right) dp - F_1(x_L) \times (v_2 - v_1). \qquad (1.4) \end{aligned}$$

3차원으로 확장한 ROC 곡면(surface)은 두 개의 임의의 x_1과 x_2에 대한 각각의 누적분포함수 값들을 계산하여 각 범주에 대한 정분류율을 3차원 공간에 대응시킨 그래프로 식 (1.5)와 같이 표현한다 (Scurfield, 1996; Mossman, 1999; Dreiseitl 등, 2000; Heckerling, 2001; Fawcett, 2003; Nakas와 Yiannoutsos, 2004; Nakas 등, 2010; Wandishin와 Mullen, 2009).

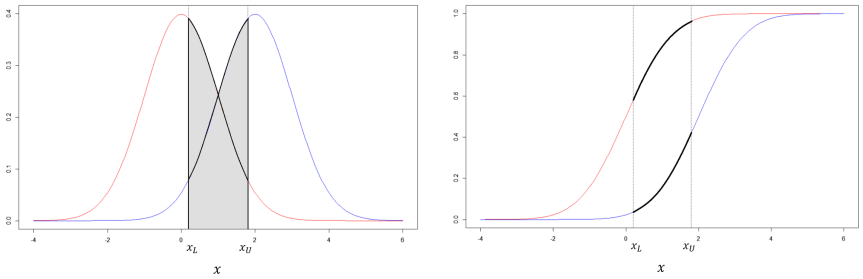
$$(F_1(x_1), 1 - F_3(x_2), F_2(x_2) - F_2(x_1)) = (p_1, p_3, \text{ROC}_s(p_1, p_3)), \qquad (1.5)$$

여기서 $\text{ROC}_s(p_1, p_3) = F_2(F_3^{-1}(1 - p_3)) - F_2(F_1^{-1}(p_1))$, $p_1, p_3 \in (0, 1)$. ROC 곡면 아래의 부피를 나타내는 volume under the ROC surface (VUS)는 식 (1.6)과 같이 확률과 적분식으로 표현된다 (Nakas와 Yiannoutsos, 2004; Xiong 등, 2006).

$$\begin{aligned} \text{VUS} &= P(X_1 \le X_2 \le X_3)\\ &= \int_0^1 \int_0^{F_1\left(F_3^{-1}(1-p_3)\right)} \left[F_2\left(F_3^{-1}(1 - p_3)\right) - F_2\left(F_1^{-1}(p_1)\right)\right] dp_1 dp_3. \qquad (1.6) \end{aligned}$$

Hong 등 (2019)은 ROC 곡선의 pAUC와 tpAUC를 ROC 곡면의 VUS로 확장하여 부분 VUS(partial VUS; pVUS)와 세 방향 부분 VUS(three-way partial VUS; tpVUS)를 식 (1.7), (1.8)과 같이 확률과 적분식으로 정의하였다. 네 개의 절단점들 $x_{L_1}, x_{U_1}, x_{L_2}, x_{U_2}$ $(x_{L_1} < x_{U_1} < x_{L_2} < x_{U_2})$ 그리고 $F_1(x_{L_1}) = u_1$, $F_1(x_{U_1}) = u_2$, $1 - F_3(x_{U_2}) = v_1$, $1 - F_3(x_{L_2}) = v_2$에 대해,

$$\begin{aligned} \text{pVUS}((u_1, u_2), (v_1, v_2)) &= P\left(X_1 \le X_2 \le X_3 \cap x_{L_1} \le X_1 \le x_{U_1} \cap x_{L_2} \le X_3 \le x_{U_2}\right),\\ \text{tpVUS}((u_1, u_2), (v_1, v_2)) &= P\left(X_1 \le X_2 \le X_3 \cap x_{L_1} \le X_1 \le X_2 \le x_{U_1} \cap x_{L_2} \le X_2 \le X_3 \le x_{U_2}\right) \qquad (1.7)\\ &= \text{pVUS}((u_1, u_2), (v_1, v_2)) - P(x_{L_1} \le X_1 \le x_{U_1} \le X_2 \le x_{L_2} \le X_3 \le x_{U_2})\\ &= \int_{v_1}^{v_2} \int_{u_1}^{u_2} \left[F_2\left(F_3^{-1}(1 - p_3)\right) - F_2\left(F_1^{-1}(p_1)\right)\right] dp_1 dp_3\\ &\quad - (u_2 - u_1)(v_2 - v_1)\left[F_2(x_{L_2}) - F_2(x_{U_1})\right]. \qquad (1.8) \end{aligned}$$



(a) Truncated pdfs (b) Truncated cdfs

Figure 2.1. Two truncated functions. pdf = probability density function; cdf = cumulative distribution function.

본 연구에서는 복잡하게 확률로 표현되고 어려운 적분식을 이용하여 구하는 양향 부분 AUC인 tpAUC와 세 방향 부분 VUS인 tpVUS를 보다 간단하게 표현하고 추정하는 방법을 제안한다. 우선 2절에서는 tpAUC와 tpVUS에 대응하는 ROC 곡선과 곡면을 구성하는 확률밀도함수와 누적분포함수를 절단함수(truncated function)로 표현한다. 그리고 이런 절단함수를 이용하는 AUC와 VUS를 정의하고, tpAUC와 tpVUS에 대하여 각각 비교하면서 설명한다. 3절에서는 tpAUC와 tpVUS를 절단함수를 이용하는 AUC와 VUS의 관계로 각각 유도하면서, tpAUC와 tpVUS의 모수적 추정 방법을 제안하고, 맨-휘트니 통계량을 사용하는 비모수적 추정 방법도 제안한다. 4절과 5절에서는 절단된 정규분포의 다양한 경우에 대하여 모수적 추정 방법으로 절단함수를 이용하는 AUC와 tpAUC와의 관계 그리고 절단함수를 이용하는 VUS와 tpVUS와의 관계에 대하여 탐색하고, 절단된 정규분포로부터 확률표본을 추출하여 비모수적인 추정 방법을 사용하여 절단함수를 이용하는 AUC와 VUS 그리고 tpAUC와 tpVUS와의 관계가 성립함을 발견한다. 마지막 6절에서는 결론을 서술하면서 향후 연구과제에 대하여 서술한다.

2. 절단함수를 이용한 AUC와 VUS

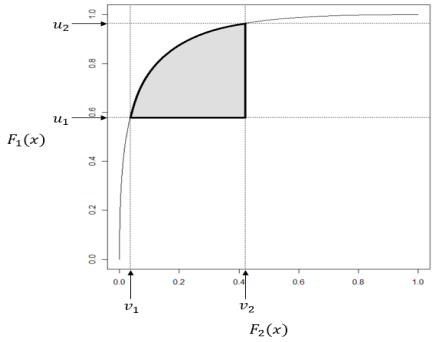
2.1. 절단함수를 이용한 AUC

Figure 2.1과 같은 확률밀도함수(probability density function; pdf) $f_1(x), f_2(x)$와 누적분포함수(cumulative distribution function; cdf) $F_1(x), F_2(x)$에서 두 개의 절단점 x_L, x_U $(x_L < x_U)$에 의하여 절단(truncated)된 경우에 대한 확률 변수를 X_1^*와 X_2^*로 정의하고 이에 대응하는 절단확률밀도함수(truncated pdf)를 다음과 같이 정의할 수 있다.

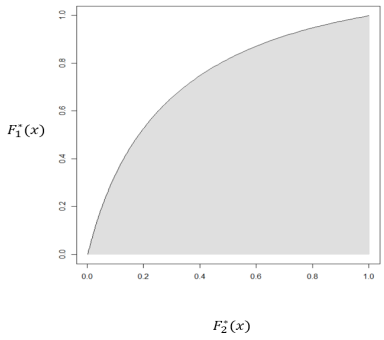
$$f_1^*(x) \equiv \frac{f_1(x)}{F_1(x_U) - F_1(x_L)} = \frac{1}{u_2 - u_1} f_1(x), \quad x_L < x < x_U,$$
$$f_2^*(x) \equiv \frac{f_2(x)}{F_2(x_U) - F_2(x_L)} = \frac{1}{v_2 - v_1} f_2(x), \quad x_L < x < x_U,$$

여기서 $F_1(x_L) = u_1$, $F_1(x_U) = u_2$, $F_2(x_L) = v_1$, $F_2(x_U) = v_2$이다. 그리고 확률변수 X_1^*와 X_2^*에 대응하는 절단누적분포함수(truncated cdf)는 다음과 같다.

$$F_1^*(x) = \int_{x_L}^{x} f_1^*(t)\, dt = \frac{F_1(x) - F_1(x_L)}{F_1(x_U) - F_1(x_L)} = \frac{F_1(x) - u_1}{u_2 - u_1},$$
$$F_2^*(x) = \int_{x_L}^{x} f_2^*(t)\, dt = \frac{F_2(x) - F_2(x_L)}{F_2(x_U) - F_2(x_L)} = \frac{F_2(x) - v_1}{v_2 - v_1}. \tag{2.1}$$

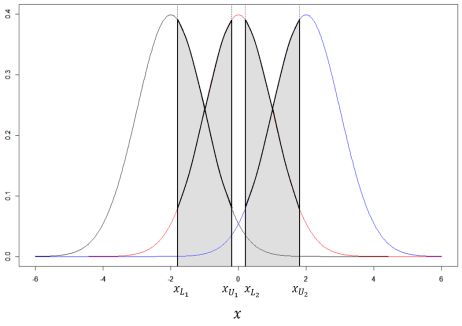


(a) tpAUC for ROC curve

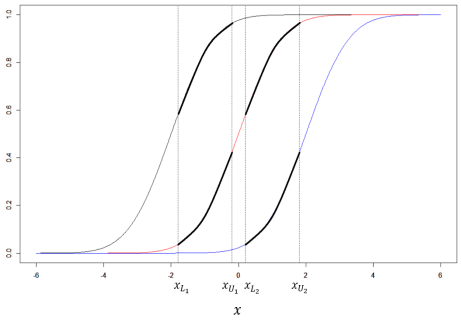


(b) ROC curve using truncated cdfs

Figure 2.2. ROC curves. tpAUC = two-way partial AUC; AUC = area under the ROC curve; ROC = receiver operating characteristic; cdf = cumulative distribution function.



(a) Truncated pdfs



(b) Truncated cdfs

Figure 2.3. Three truncated functions. pdf = probability density function; cdf = cumulative distribution function.

Figure 2.2(a)는 누적분포함수 $F_1(x)$, $F_2(x)$를 바탕으로 작성한 ROC 곡선이며 곡선 아래에서 어두운 색으로 표시한 영역이 식 (1.3)과 (1.4)의 tpAUC(v_1, v_2)를 나타낸다. 그리고 절단누적분포함수 $F_1^*(x)$, $F_2^*(x)$에 대하여 식 (1.1)과 같이 작성한 ROC 곡선과 이에 대응하는 AUC를 어두운 색으로 Figure 2.2(b)에 표현할 수 있다.

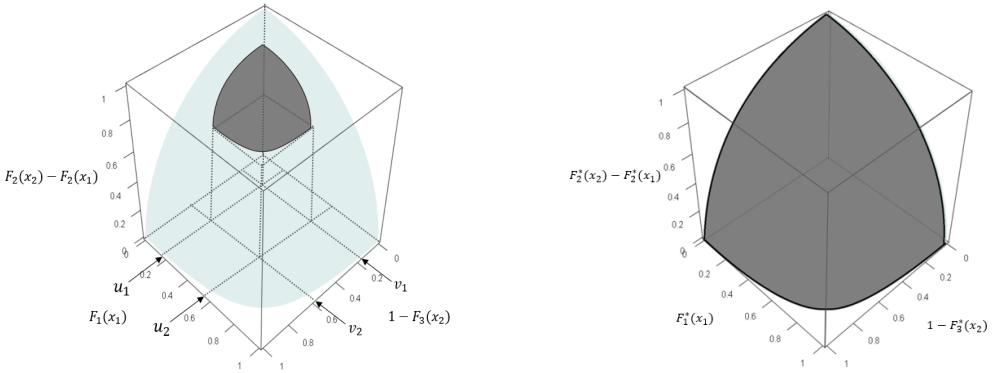
두 절단누적분포함수를 이용하여 ROC 곡선을 표현하였듯이 절단함수를 이용하여 AUC를 정의 2.1과 같이 제안한다.

정의 2.1 식 (2.1)의 절단누적분포함수에 대한 AUC를 AUC*로 표기하면 다음과 같이 정의한다.

$$\text{AUC}^* = P\left(X_1^* \leq X_2^*\right).$$

2.2. 절단함수를 이용한 VUS

2.1절에서 논의한 확률밀도함수와 누적분포함수를 세 종류로 확장하고, 절단점을 네 개로 확장하여, Figure 2.3과 같이 절단점들 $x_{L_1}, x_{U_1}, x_{L_2}, x_{U_2}$ $(x_{L_1} < x_{U_1} < x_{L_2} < x_{U_2})$에 대한 확률 변수를 X_1^*, X_2^*, X_3^*로 정의하고 이에 대응하는 절단확률밀도함수와 절단누적분포함수를 다음과 같이 정의할 수



(a) tpVUS for ROC surface (b) ROC surface using truncated cdfs

Figure 2.4. ROC surfaces. tpVUS = three-way partial VUS; VUS = volume under the ROC surface; ROC = receiver operating characteristic; cdf = cumulative distribution function.

있다.

$$
\begin{aligned}
f_1^*(x) &\equiv \frac{f_1(x)}{F_1(x_{U_1}) - F_1(x_{L_1})} = \frac{1}{u_2 - u_1} f_1(x), \quad x_{L_1} < x < x_{U_1}, \\
f_2^*(x) &\equiv \frac{f_2(x)}{F_2(x_{U_1}) - F_2(x_{L_1}) + F_2(x_{U_2}) - F_2(x_{L_2})}, \quad \begin{cases} x_{L_1} < x < x_{U_1} & \text{or} \\ x_{L_2} < x < x_{U_2}, \end{cases} \\
f_3^*(x) &\equiv \frac{f_3(x)}{F_3(x_{U_2}) - F_3(x_{L_2})} = \frac{1}{v_2 - v_1} f_3(x), \quad x_{L_2} < x < x_{U_2}, \\
F_1^*(x) &= \frac{F_1(x) - F_1(x_{L_1})}{F_1(x_{U_1}) - F_1(x_{L_1})} = \frac{1}{u_2 - u_1}(F_1(x) - u_1), \quad x_{L_1} < x < x_{U_1}, \\
F_2^*(x) &= \begin{cases} \dfrac{F_2(x) - F_2(x_{L_1})}{F_2(x_{U_1}) - F_2(x_{L_1}) + F_2(x_{U_2}) - F_2(x_{L_2})}, & x_{L_1} < x < x_{U_1}, \\ \dfrac{F_2(x_{U_1}) - F_2(x_{L_1}) + F_2(x) - F_2(x_{L_2})}{F_2(x_{U_1}) - F_2(x_{L_1}) + F_2(x_{U_2}) - F_2(x_{L_2})}, & x_{L_2} < x < x_{U_2}, \end{cases} \\
F_3^*(x) &= \frac{F_3(x) - F_3(x_{L_2})}{F_3(x_{U_2}) - F_3(x_{L_2})} = \frac{1}{v_2 - v_1}(F_3(x) - (1 - v_2)), \quad x_{L_2} < x < x_{U_2},
\end{aligned} \tag{2.2}
$$

여기서 $F_1(x_{L_1}) = u_1$, $F_1(x_{U_1}) = u_2$, $1 - F_3(x_{U_2}) = v_1$, $1 - F_3(x_{L_2}) = v_2$이다.

Figure 2.4(a)는 누적분포함수 $F_1(x), F_2(x), F_3(x)$을 바탕으로 작성한 ROC 곡면이며 곡면 아래에서 어두운 색으로 표시한 영역이 식 (1.7)과 (1.8)의 tpVUS$((u_1, u_2), (v_1, v_2))$를 나타낸다. 그리고 절단누적분포함수 $F_1^*(x), F_2^*(x), F_3^*(x)$에 대하여 식 (1.5)와 같이 작성한 ROC 곡면과 이에 대응하는 VUS를 어두운 색으로 Figure 2.4(b)에 표현할 수 있다.

세 종류의 절단누적분포함수 $F_1^*(x)$, $F_2^*(x)$, $F_3^*(x)$를 이용한 ROC 곡면으로부터의 VUS는 정의 2.2와 같이 제안한다.

정의 2.2 절단누적분포함수를 이용하여 VUS를 VUS*로 표기하면 다음과 같이 정의한다.

$$\text{VUS}^* = P(X_1^* \le X_2^* \le X_3^*).$$

3. tpAUC와 tpVUS의 추정 방법

3.1. 모수적 추정 방법: tpAUC와 tpVUS의 관계

Yang 등 (2019)의 tpAUC(v_1, v_2)와 2.1절에서 제안한 절단누적분포함수를 이용한 AUC (AUC*)와의 관계를 유도하고, 나아가 Hong 등 (2019)이 제안한 tpVUS$((u_1, u_2), (v_1, v_2))$와 2.2절에서의 절단누적분포함수를 이용한 VUS (VUS*)와의 관계를 확장하여 설명한다. 우선 AUC*의 확률을 이용하여 tpAUC(v_1, v_2)와의 관계를 정리 3.1에 설명한다.

정리 3.1 *tp*AUC(v_1, v_2)와 AUC*는 다음과 같은 선형 관계를 갖는다.

$$tp\text{AUC}(v_1, v_2) \equiv \text{AUC}^* \times (u_2 - u_1)(v_2 - v_1),$$

여기서 $F_1(x_L) = u_1,\ F_1(x_U) = u_2,\ F_2(x_L) = v_1,\ F_2(x_U) = v_2$0이다.

증명:

$$\begin{aligned}\text{AUC}^* &= P(X_1^* \le X_2^*) = P(X_1 \le X_2 \mid X_1^* \cap X_2^*) \\ &= P(X_1 \le X_2 | x_L \le X_1 \le x_U \cap x_L \le X_2 \le x_U) \\ &= \frac{P(x_L \le X_1 \le X_2 \le x_U)}{P(x_L \le X_1 \le x_U \ \cap \ x_L \le X_2 \le x_U)}.\end{aligned}$$

식 (1.3)에서 tpAUC$(v_1, v_2) = P(x_L \le X_1 \le X_2 \le x_U)$이므로

$$\text{AUC}^* = \frac{\text{tpAUC}(v_1, v_2)}{(u_2 - u_1)(v_2 - v_1)}.$$

□

정리 3.1의 선형 관계는 Figure 2.2를 통하여 설명하면 다음과 같다. Figure 2.2(a)에서의 양방향 부분 AUC, tpAUC(v_1, v_2)를 살펴보면, TPR, $F_1(x)$의 범위가 (u_1, u_2)이며 FPR, $F_2(x)$의 범위가 (v_1, v_2)임을 파악할 수 있다. 따라서 tpAUC(v_1, v_2)에 대응하는 TPR과 FPR의 폭을 tpAUC(v_1, v_2)에 나누어 주면, Figure 2.2(b)의 절단누적분포함수를 이용한 ROC 곡선에서의 AUC*의 면적임을 확인할 수 있다. 역으로 Figure 2.2(b)의 AUC*에서 $F_1(x)$와 $F_2(x)$의 절단된 각각의 크기 $(u_2 - u_1)$과 $(v_2 - v_1)$를 곱하면, tpAUC(v_1, v_2)임을 발견할 수 있다.

또한 Figure 2.2(a)에서의 tpAUC(v_1, v_2)에서 수평축까지 아래 부분에 해당하는 직사각형의 면적은 $(v_2 - v_1) \times u_1 = (v_2 - v_1)F_1(x_L)$이므로 tpAUC$(v_1, v_2)$에 이 면적을 합하면 pAUC$(v_1, v_2)$이 되는 관계식인 식 (1.4)를 설명한다.

정리 3.1을 확장하여 VUS*와 tpVUS$((u_1, u_2), (v_1, v_2))$와의 관계를 유도한다.

정리 3.2 *tp*VUS$((u_1, u_2), (v_1, v_2))$와 VUS*는 다음과 같은 선형 관계를 갖는다.

$$tp\text{VUS}((u_1, u_2), (v_1, v_2)) = \text{VUS}^* \times (u_2 - u_1)(v_2 - v_1) \times \left[F_2(x_{U_1}) - F_2(x_{L_1}) + F_2(x_{U_2}) - F_2(x_{L_2})\right],$$

여기서 $F_1(x_{L_1}) = u_1,\ F_1(x_{U_1}) = u_2,\ 1 - F_3(x_{U_2}) = v_1,\ 1 - F_3(x_{L_2}) = v_2$0이다.

증명: 정리 3.2의 증명도 정리 3.1의 증명에서와 같이 확률로 표현 가능하지만 이번에는 적분식을 이용

하여 증명한다.

$$\begin{aligned}
\mathrm{VUS}^* &= \int_0^1 \int_0^{F_1^*(x_3)} [F_2^*(x_3) - F_2^*(x_1)]\, dF_1^*(x_1) dF_3^*(x_3) \\
&= \int_{x_{L_2}}^{x_{U_2}} \int_{x_{L_1}}^{x_{U_1}} [\{F_2(x_{U_1}) - F_2(x_{L_1})\} + \{F_2(x_3) - F_2(x_{L_2})\} - \{F_2(x_1) - F_2(x_{L_1})\}]\, dF_1(x_1) dF_3(x_3) \\
&\quad (u_2 - u_1)(v_2 - v_1)[F_2(x_{U_1}) - F_2(x_{L_1}) + F_2(x_{U_2}) - F_2(x_{L_2})] \\
&= \int_{F_3(x_{L_2})}^{F_3(x_{U_2})} \int_{F_1(x_{L_1})}^{F_1(x_{U_1})} [\{F_2(x_3) - F_2(x_1)\} - \{F_2(x_{L_2}) - F_2(x_{U_1})\}]\, dF_1(x_1) dF_3(x_3) \\
&\quad (u_2 - u_1)(v_2 - v_1)[F_2(x_{U_1}) - F_2(x_{L_1}) + F_2(x_{U_2}) - F_2(x_{L_2})] \\
&= \left\{ \int_{v_1}^{v_2} \int_{u_1}^{u_2} \left\{ F_2\left(F_3^{-1}(1-p_3)\right) - F_2\left(F_1^{-1}(p_1)\right) \right\} dp_1 dp_3 - (u_2 - u_1)(v_2 - v_1)[F_2(x_{L_2}) - F_2(x_{U_1})] \right\} \\
&\quad (u_2 - u_1)(v_2 - v_1)[F_2(x_{U_1}) - F_2(x_{L_1}) + F_2(x_{U_2}) - F_2(x_{L_2})].
\end{aligned}$$

식 (1.8)에서의 $\mathrm{tpVUS}((u_1, u_2), (v_1, v_2))$의 적분식을 대체하여 다음을 유도할 수 있다.

$$\mathrm{VUS}^* = \frac{\mathrm{tpVUS}((u_1, u_2), (v_1, v_2))}{(u_2 - u_1)(v_2 - v_1)\,[F_2(x_{U_1}) - F_2(x_{L_1}) + F_2(x_{U_2}) - F_2(x_{L_2})]}.$$

□

정리 3.2의 관계는 Figure 2.4를 이용하여 설명하여 보자. Figure 2.4(a)에서의 세 방향 부분 VUS, $\mathrm{tpVUS}((u_1, u_2), (v_1, v_2))$에 대응하는 $F_1(x)$축의 범위는 (u_1, u_2) 그리고 $1-F_3(x)$축의 범위는 (v_1, v_2)이다. 그리고 수직축에서 최고높이는 $F_2(x_{U_2}) - F_2(x_{L_1})$이며 최저높이는 $F_2(x_{L_2}) - F_2(x_{U_1})$이므로 수직축 범위의 폭은 $[F_2(x_{U_1}) - F_2(x_{L_1}) + F_2(x_{U_2}) - F_2(x_{L_2})]$이다. 따라서 $\mathrm{tpVUS}((u_1, u_2), (v_1, v_2))$에서 각 축의 범위의 폭들의 곱 $(u_2 - u_1) \times (v_2 - v_1) \times [F_2(x_{U_1}) - F_2(x_{L_1}) + F_2(x_{U_2}) - F_2(x_{L_2})]$를 나누면 Figure 2.4(b)의 VUS^*임을 탐색할 수 있다.

그리고 Figure 2.4(a)에서 $\mathrm{tpVUS}((u_1, u_2), (v_1, v_2))$ 아래의 직육면체를 살펴보면, 직육면체 밑변 넓이는 $(u_2 - u_1) \times (v_2 - v_1)$이며 직육면체 높이는 $\mathrm{tpVUS}((u_1, u_2), (v_1, v_2))$의 최저 높이인 $F_2(x_{L_2}) - F_2(x_{U_1})$이다. 따라서 $\mathrm{tpVUS}((u_1, u_2), (v_1, v_2))$ 아래의 직육면체 부피는 $(u_2 - u_1) \times (v_2 - v_1) \times \{F_2(x_{L_2}) - F_2(x_{U_1})\}$이 된다. 따라서 $\mathrm{tpVUS}((u_1, u_2), (v_1, v_2))$에 이 부피 $(u_2 - u_1) \times (v_2 - v_1) \times \{F_2(x_{L_2}) - F_2(x_{U_1})\}$를 더하면 1절에서 논의한 $\mathrm{pVUS}((u_1, u_2), (v_1, v_2))$임을 확인할 수 있다.

3.2. 비모수적 추정 방법

3.1절에서는 $\mathrm{tpAUC}(v_1, v_2)$와 $\mathrm{tpVUS}((u_1, u_2), (v_1, v_2))$를 AUC^*와 VUS^*의 관계로 유도하였다. 이 관계를 이용하면 $\mathrm{tpAUC}(v_1, v_2)$와 $\mathrm{tpVUS}((u_1, u_2), (v_1, v_2))$의 모수적인 추정 방법으로 활용할 수 있다. 본 절에서는 비모수적인 추정 방법으로 연구를 확장한다. 우선 $\mathrm{tpAUC}(v_1, v_2)$를 추정하는 방법으로 Yang 등 (2019)은 맨-휘트니 통계량(Mann-Whitney statistic)을 이용한 $\mathrm{tpAUC}_{\mathrm{MW}}$를 다음과 같이 제안하였다. 확률표본 $\{X_{1,i}\}$, $\{X_{2,j}\}$의 표본크기가 각각 n_1, n_2인 경우에,

$$\begin{aligned}
\mathrm{tpAUC}_{\mathrm{MW}} = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \Bigg[& I(X_{1,i} < X_{2,j},\ X_{1,i} \geq x_L,\ X_{2,j} \leq x_U) \\
& + \frac{1}{2} I(X_{1,i} = X_{2,j},\ X_{1,i} \geq x_L,\ X_{2,j} \leq x_U) \Bigg].
\end{aligned}$$

$\{X_{1,i} \geq x_L,\ X_{2,j} \leq x_U\}$의 범위로 제한하는 절단된 확률표본을 $\{X^*_{1,i}\}$, $\{X^*_{2,j}\}$라고 하고, 이들의 표본크기를 각각 n_1^*, n_2^* $(n_1^* < n_1,\ n_2^* < n_2)$고 하면, 절단된 두 확률표본에 대하여 맨-휘트니 통계량을 이용한 AUC^*을 $\mathrm{AUC}^*_{\mathrm{MW}}$로 표기하고 Lemma 3.1에서와 같이 제안한다.

Lemma 3.1 $\mathrm{AUC}^*_{\mathrm{MW}}$의 비모수적 추정 방법으로 다음과 같이 설정한다.

$$\mathrm{AUC}^*_{\mathrm{MW}} = \frac{1}{n_1^* n_2^*} \sum_{i=1}^{n_1^*} \sum_{j=1}^{n_2^*} \left[I\left(X^*_{1,i} < X^*_{2,j}\right) + \frac{1}{2} I\left(X^*_{1,i} = X^*_{2,j}\right) \right].$$

정리 3.3 $tp\mathrm{AUC}_{\mathrm{MW}}$와 $\mathrm{AUC}^*_{\mathrm{MW}}$는 다음과 같은 관계를 갖는다.

$$tp\mathrm{AUC}_{\mathrm{MW}} = \mathrm{AUC}^*_{\mathrm{MW}} \times \frac{n_1^* n_2^*}{n_1 n_2}.$$

증명: $\mathrm{tpAUC}_{\mathrm{MW}}$의 확률표본 $\{X_{1,i}\}$, $\{X_{2,j}\}$에 대한 조건과 $\mathrm{AUC}^*_{\mathrm{MW}}$의 절단된 확률표본 $\{X^*_{1,i}\}$, $\{X^*_{2,j}\}$의 조건이 동일하므로 두 통계량에서 이중합(double summation)의 결과는 일치한다. 그리고 n_1^*/n_1와 n_2^*/n_2는 각각 $(u_2 - u_1)$와 $(v_2 - v_1)$의 추정값으로, 정리 3.1에서와 같이 $\mathrm{AUC}^*_{\mathrm{MW}}$에 n_1^*/n_1와 n_2^*/n_2를 곱해주면 $\mathrm{tpAUC}_{\mathrm{MW}}$의 값과 동일하다. □

Hong과 Cho (2015)는 맨-휘트니 통계량을 이용하여 세 확률표본일 때의 VUS의 비모수 추정 방법을 제안하였다. 이를 확장하여 $\mathrm{tpVUS}((u_1, u_2), (v_1, v_2))$의 비모수 추정량 $\mathrm{tpVUS}_{\mathrm{MW}}$은 다음과 같이 표현할 수 있다. 세 확률표본 $\{X_{1,i}\}$, $\{X_{2,j}\}$, $\{X_{1,i}\}$의 표본크기가 각각 n_1, n_2, n_3이고, $\{x_{L_1} \leq X_{1,i} \leq x_{U_1},\ X_{2,j} \leq x_{U_1},\ X_{2,j} \geq x_{L_2},\ x_{L_2} \leq X_{3,k} \leq x_{U_2}\}$의 범위에 대하여,

$$\begin{aligned}
&\mathrm{tpVUS}_{\mathrm{MW}} \times (n_1 n_2 n_3) \\
&= \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \sum_{k=1}^{n_3} \Big[I\left(X_{2,j} < X_{3,k} | X_{1,i} < X_{2,j},\ x_{L_1} \leq X_{1,i}, X_{2,j} \leq x_{U_1},\ x_{L_2} \leq X_{2,j}, X_{3,k} \leq x_{U_2}\right) \\
&\quad + \frac{1}{2} I\left(X_{2,j} = X_{3,k} | X_{1,i} < X_{2,j},\ x_{L_1} \leq X_{1,i}, X_{2,j} \leq x_{U_1},\ x_{L_2} \leq X_{2,j}, X_{3,k} \leq x_{U_2}\right) \\
&\quad + \frac{1}{2} I(X_{2,j} < X_{3,k} | X_{1,i} = X_{2,j},\ x_{L_1} \leq X_{1,i}, X_{2,j} \leq x_{U_1},\ x_{L_2} \leq X_{2,j}, X_{3,k} \leq x_{U_2}) \\
&\quad + \frac{1}{2^2} I(X_{2,j} = X_{3,k} | X_{1,\,i} = X_{2,\,j},\ x_{L_1} \leq X_{1,i}, X_{2,\,j} \leq x_{U_1},\ x_{L_2} \leq X_{2,j}, X_{3,k} \leq x_{U_2}) \Big],
\end{aligned}$$

여기서 $I(A|B)$는 사상 B 조건에서 사상 A의 지시함수이다 (Hong과 Cho, 2015). 절단된 확률표본을 $\{X^*_{1,i}\}$, $\{X^*_{2,j}\}$, $\{X^*_{3,k}\}$라고 하고, 이에 대응하는 표본크기를 각각 n_1^*, n_2^*, n_3^* $(n_i^* < n_i,\ i = 1, 2, 3)$라고 하자. 절단된 세 확률표본의 VUS를 $\mathrm{VUS}^*_{\mathrm{MW}}$로 표기하고 Lemma 3.2와 같이 제안한다.

Lemma 3.2 $\mathrm{VUS}^*_{\mathrm{MW}}$의 비모수적 추정 방법으로 다음과 같이 설정한다.

$$\begin{aligned}
\mathrm{VUS}^*_{\mathrm{MW}} \times n_1^* n_2^* n_3^* = \sum_{i=1}^{n_1^*} \sum_{j=1}^{n_2^*} \sum_{k=1}^{n_3^*} \Big[& I\left(X^*_{2,j} < X^*_{3,k} | X^*_{1,i} < X^*_{2,j}\right) + \frac{1}{2} I\left(X^*_{2,j} = X^*_{3,k} | X^*_{1,i} < X^*_{2,j}\right) \\
& + \frac{1}{2} I\left(X^*_{2,j} < X^*_{3,k} | X^*_{1,i} = X^*_{2,j}\right) + \frac{1}{2^2} I\left(X^*_{2,j} = X^*_{3,k} | X^*_{1,i} = X^*_{2,j}\right) \Big].
\end{aligned}$$

정리 3.4 $tp\mathrm{VUS}_{\mathrm{MW}}$와 $\mathrm{VUS}^*_{\mathrm{MW}}$는 다음과 같은 관계를 갖는다.

$$tp\mathrm{VUS}_{\mathrm{MW}} = \mathrm{VUS}^*_{\mathrm{MW}} \times \frac{n_1^* n_2^* n_3^*}{n_1 n_2 n_3}.$$

증명: tpVUS_{MW}의 확률표본 $\{X_{1,i}\}$, $\{X_{2,j}\}$, $\{X_{3,k}\}$에 대한 조건과 VUS^*_{MW}의 절단된 확률표본 $\{X^*_{1,i}\}$, $\{X^*_{2,j}\}$, $\{X^*_{3,k}\}$의 조건이 같으므로 두 통계량을 구할 때 삼중합(triple summation)의 결과는 동일하다. 따라서 정리 3.2에서와 유사하게 VUS^*_{MW}에 n^*_1/n_1, n^*_2/n_2와 n^*_3/n_3을 곱해준 값과 tpVUS_{MW} 값은 일치한다. □

4. 정규분포에서 모수적 추정 예제

본 절에서는 절단된 정규분포를 다양하게 설정하여 tpAUC와 AUC^*의 면적과 tpVUS와 VUS^*의 부피를 구하여 비교하면서 이들의 관계를 확인한다.

4.1. tpAUC의 모수적 추정

두 개의 스코어 확률변수 X_1과 X_2의 분포함수를 각각 정규분포 $N(0,1)$과 $N(1.5,1)$로 가정하고 절단점 x_L, x_U를 0.3과 1.2로 각각 설정하여, 이에 대응하는 $\text{tpAUC}(v_1, v_2)$와 AUC^*를 계산한다. 두 정규분포와 절단점을 Figure 2.1에 그리고 $\text{tpAUC}(v_1, v_2)$와 AUC^*를 Figure 2.2에 표현하였다.

가정한 분포에서 $\text{tpAUC} = 0.0434$로 Figure 2.2(a)에서 어두운 영역으로 표시한 면적이다. $u_2 - u_1 = F_1(1.2) - F_1(0.3) = 0.2670$, $v_2 - v_1 = F_2(1.2) - F_2(0.3) = 0.2670$이다. Figure 2.2(b)의 ROC 곡선 아래 면적 $\text{AUC}^* = 0.6094$이다. 따라서 $\text{tpAUC} = \text{AUC}^* \times (u_2 - u_1)(v_2 - v_1) = 0.6094 \times (0.2670 \times 0.2670) = 0.0434$로 정리 3.1의 관계를 확인하였다.

4.2. tpVUS의 모수적 추정

세 개의 스코어 확률변수 X_1, X_2, X_3의 분포를 각각 정규분포 $N(-1.5,1)$, $N(0,1)$, $N(1.5,1)$로 가정하고, 네 개의 절단점 x_{L_1}, x_{U_1}, x_{L_2}, x_{U_2}을 각각 -1.2, -0.3, 0.3, 1.2로 설정한다. 그리고 두 절단점 x_{L_1}과 x_{U_1} 그리고 x_{L_2}과 x_{U_2} 범위 사이의 부피에 대한 $\text{tpVUS}((u_1, u_2), (v_1, v_2))$를 구하고 VUS^*와 비교한다. 세 종류의 정규분포와 네 절단점들을 Figure 2.3에 그리고 $\text{tpVUS}((u_1, u_2), (v_1, v_2))$와 VUS^*를 Figure 2.4(a)와 (b)에 표현하였다.

가정한 분포에서 $\text{tpVUS}((u_1, u_2), (v_1, v_2)) = 0.0232$로 Figure 2.4(a)에서 어두운 영역으로 표시한 부피이다. $u_2 - u_1 = F_1(-1.2) - F_2(-0.3) = 0.2670$, $v_2 - v_1 = F_1(-1.2) - F_2(-0.3) = 0.2670$ 그리고 $[F_2(x_{U_1}) - F_2(x_{L_1}) + F_2(x_{U_2}) - F_2(x_{L_2})] = [F_2(-0.3) - F_2(-1.2) + F_2(1.2) - F_2(0.3)] = 0.5340$이다. Figure 2.4(b)의 ROC 곡면 아래 부피 $\text{VUS}^* = 0.6094$이다. 따라서 $\text{tpVUS}((u_1, u_2), (v_1, v_2)) = 0.6094 \times (0.2670 \times 0.2670 \times 0.5340) = 0.0232$로 정리 3.2의 관계를 확인하였다.

5. 표본자료에서 비모수적 추정

본 절에서는 4절에서 가정한 절단된 정규분포에서 확률표본들을 추출하여 $\text{tpAUC}(v_1, v_2)$와 AUC^* 그리고 $\text{tpVUS}((u_1, u_2), (v_1, v_2))$와 VUS^*를 비모수적인 방법으로 추정하여 비교하면서 이들의 관계를 확인한다.

5.1. tpAUC의 비모수적 추정

확률변수 X_1과 X_2의 분포함수를 4.1절과 동일하게 가정하고 각 분포에서 $n_1 = 200$, $n_2 = 300$개의 표본을 추출한다. 그리고 절단점 x_L, x_U를 각각 0.3과 1.2로 설정하면 절단된 확률표본은 각각 $n^*_1 = 59$,

Table 5.1. Two random samples and truncated random samples

X_1	−2.7318	−2.4883	−2.2189	⋯	2.0534	2.3101	2.5538	$n_1 = 200$
X_2	−0.9862	−0.6491	−0.5332	⋯	3.7586	3.7934	3.9402	$n_2 = 300$
X_1^*	0.3136	0.3218	0.3431	⋯	1.1319	1.1391	1.1998	$n_1^* = 59$
X_2^*	0.3058	0.3149	0.3223	⋯	1.1680	1.1819	1.1903	$n_2^* = 76$

Table 5.2. Three random samples and truncated random samples

X_1	−4.2394	−3.8165	−3.6567	⋯	1.0604	1.3810	2.2400	$n_1 = 200$
X_2	−2.6655	−2.3114	−2.2158	⋯	2.7097	2.8688	4.1226	$n_2 = 300$
X_3	−1.0976	−1.0453	−0.8644	⋯	4.2573	4.2944	7.8311	$n_3 = 400$
X_1^*	−1.1925	−1.1923	−1.1497	⋯	−0.3081	−0.3068	−0.3062	$n_1^* = 55$
X_2^*	−1.1816	−1.1695	−1.1691	⋯	1.1804	1.1848	1.1867	$n_2^* = 166$
X_3^*	0.3096	0.3735	0.3753	⋯	1.1727	1.1955	1.1997	$n_3^* = 111$

$n_2^* = 76$으로 Table 5.1에 일부분만을 크기순으로 나열하여 정리하였다.

맨-휘트니 통계량을 응용해 $\text{tpAUC}_{\text{MW}} = 0.0430$, $n_1^*/n_1 = 59/200 = 0.2950$, $n_2^*/n_2 = 76/300 = 0.2533$, 그리고 $\text{AUC}_{\text{MW}}^* = 0.5750$이다. 따라서

$$\text{tpAUC}_{\text{MW}} = \text{AUC}_{\text{MW}}^* \times \frac{n_1^* n_2^*}{n_1 n_2} = 0.5750 \times (0.2950 \times 0.2533) = 0.0430$$

으로 정리 3.3의 관계를 확인하였다. 참고로 모수적인 방법으로 구한 추정값과 비교하면 $\text{tpAUC} = 0.6094 \times (0.2670 \times 0.2670) = 0.0434$로 근소한 차이가 발생함을 파악할 수 있다.

5.2. tpVUS의 비모수적 추정

확률변수 X_1, X_2와 X_3의 분포함수를 4.2절과 동일하게 가정하고 각 분포에서 $n_1 = 200$, $n_2 = 300$, $n_3 = 400$개의 표본을 추출한다. 그리고 절단점 $x_{L_1}, x_{U_1}, x_{L_2}, x_{U_2}$를 각각 −1.2, −0.3, 0.3, 1.2로 설정하면 절단된 확률표본은 각각 $n_1^* = 55, n_2^* = 166, n_3^* = 111$로 Table 5.2에 일부분만을 크기순으로 나열하여 정리하였다.

맨-휘트니 통계량을 응용해 $\text{tpVUS}_{\text{MW}} = 0.0240$, $n_1^*/n_1 = 55/200 = 0.2750$, $n_2^*/n_2 = 166/300 = 0.5533$, $n_3^*/n_3 = 111/400 = 0.2775$, 그리고 $\text{VUS}_{\text{MW}}^* = 0.5666$이다. 따라서

$$\text{tpVUS}_{\text{MW}} = \text{VUS}_{\text{MW}}^* \times \frac{n_1^* n_3^* n_2^*}{n_1 n_3 n_2} = 0.5666 \times (0.2750 \times 0.2775 \times 0.5533) = 0.0240$$

으로 정리 3.4의 관계를 확인하였다. 모수적인 방법으로 구한 추정과 비교하면 $\text{tpVUS}((u_1, u_2), (v_1, v_2)) = 0.6094 \times (0.2670 \times 0.2670 \times 0.5340) = 0.0232$로 각각의 추정값들은 조금씩 차이가 발생하지만 전체적으로 보면 유사함을 탐색할 수 있다.

6. 결론

본 연구에서는 Yang 등 (2019)과 Hong 등 (2019)이 제안한 tpAUC와 tpVUS를 간단하게 표현하고 추정하는 방법을 제안한다. 양방향으로 제약을 받은 tpAUC와 세 방향으로 제약을 받은 tpVUS에 대응하는 ROC 곡선과 곡면을 구성하는 각각의 확률밀도함수와 누적분포함수를 절단함수로 표현하였다. 그리고 이러한 절단함수들을 이용하여 ROC 곡선과 곡면에 대한 AUC와 VUS를 각각 정의하였다. 절단함

수를 이용하는 AUC와 tpAUC의 관계 그리고 절단함수를 이용하는 VUS와 tpVUS의 관계가 선형적인 관계를 갖고 있음을 수리적으로 증명하였다.

이런 관계식을 바탕으로 복잡하게 표현되고 계산하는 tpAUC와 tpVUS를 간단한 식으로 표현하는 모수적 추정 방법을 제안하고, 확률표본에 대한 tpAUC와 tpVUS를 계산하는 경우에도 맨-휘트니 통계량을 사용하면서 비모수적 추정 방법도 제안하였다.

다양한 정규분포의 경우에 대하여 tpAUC와 tpVUS를 계산하는 예제에서 절단된 정규분포에 대한 AUC와 VUS를 간단하게 구하여 복잡한 tpAUC와 tpVUS를 추정하는 방법으로 활용할 수 있음을 탐색하였으며, 절단된 정규분포로부터 확률표본을 추출하여 비모수적인 추정 방법을 사용하여 AUC와 VUS를 계산하여 tpAUC 그리고 tpVUS와의 관계가 성립함을 발견하였다.

본 연구에서는 두 개 또는 세 개의 확률변수들과 이에 대응하는 ROC 곡선과 곡면에 대한 AUC와 VUS에 관한 연구를 하였지만, 네 개 이상의 확률변수들의 ROC manifold 그리고 이에 대한 hypervolume under the ROC manifold (HUM)에 관한 연구 즉 multi-way HUM (mwHUM)에 관한 연구를 향후 연구과제로 남겨놓는다.

References

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms, *Pattern Recognition*, **30**, 1145–1159.

Dreiseitl, S., Ohno-Machado, L., and Binder, M. (2000). Comparing three-class diagnostic tests by three-way ROC analysis, *Medical Decision Making*, **20**, 323–331.

Fawcett, T. (2003). ROC graphs: notes and practical considerations for data mining researchers, *HP Laboratories*, California.

Greiner, M., Pfeiffer, D., and Smith, R. D. (2000). Principles and practical application of the receiver operating characteristic analysis for diagnostic tests, *Preventive Veterinary Medicine*, **45**, 23–41.

Heckerling, P. S. (2001). Parametric three-way receiver operating characteristic surface analysis using mathematica, *Medical Decision Making*, **21**, 409–417.

Hong, C. S. and Cho, H. S. (2019). Partial AUC and optimal thresholds, *The Korean Journal of Applied Statistics*, **32**, 187–198.

Hong, C. S. and Cho, M. H. (2015). VUS and HUM represented with Mann-Whitney statistic, *Communications for Statistical Applications and Methods*, **22**, 223–232.

Hong, C. S., Jung, M. S., and Shin, H. S. (2019). Three-way partial VUS, *Journal of The Korean Data and Information Science Society*, **30**, 445–454.

Jiang, Y., Metz, C., and Nishikawa, R. (1996). A receiver operating characteristic partial area index for highly sensitive diagnostic tests, *Radiology*, **201**, 745–750.

Krzanowski, W. J. and Hand, D. J. (2009). *ROC Curves for Continuous Data* (1st ed.), CRC Press, New York.

McClish, D. (1989). Analyzing a portion of the ROC curve, *Medical Decision Making*, **9**, 190–195.

Metz, C. E. (1978). Basic principles of ROC analysis, *Seminars in Nuclear Medicine*, **8**, 283–298.

Mossman, D. (1999). Three-way ROCs, *Medical Decision Making*, **19**, 78–89.

Nakas, C. T. and Yiannoutsos, C. T. (2004). Ordered multiple-class ROC analysis with continuous measurements, *Statistics in Medicine*, **23**, 3437–3449.

Nakas, C. T., Alonzo, T. A., and Yiannoutsos, C. T. (2010). Accuracy and cut-off point selection in three-class classification problems using a generalization of the Youden index, *Statistics in Medicine*, **23**, 3437–3449.

Scurfield, B. K. (1996). Multiple-event forced-choice tasks in the theory of signal detectability, *Journal of Mathematical Psychology*, **40**, 253–269.

Tasche, D. (2006). Validation of internal rating systems and PD estimates, arXiv.org, eprint arXiv: physics / 0606071

Thompson, M. L. and Zucchini, W. (1989). On the statistical analysis of ROC curves, *Statistics in Medicine*, **8**, 1277–1290.

Wandishin, M. S. and Mullen, S. J. (2009). Multiclass ROC analysis, *Weather and Forecasting*, **24**, 530–547.

Xiong, C., Van Belle, G., Miller, J. P., and Morris, J. C. (2006). Measuring and estimating diagnostic accuracy when there are three ordinal diagnostic groups, *Statistics in Medicine*, **25**, 1251–1273.

Yang, H., Lu, K., Lyu, X., and Hu, F. (2019). Two-way partial AUC and its properties, *Statistical Methods in Medical Research*, **28**, 184–195.

Zweig, M. H. and Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine, *Clinical Chemistry*, **39**, 561–577.

절단함수를 이용한 AUC와 VUS

홍종선[a,1] · 홍성혁[a]

[a]성균관대학교통계학과

(2019년 5월 16일 접수, 2019년 6월 14일 수정, 2019년 6월 28일 채택)

요 약

ROC 곡선 아래 면적과 ROC 곡면 아래 부피를 이용하여 분류모형의 판별력을 측정하는 통계량인 AUC와 VUS에 관한 많은 연구가 있다. ROC 곡선을 구성하는 FPR과 TPR 모두에 제한을 두는 양방향 부분 AUC는 부분 AUC보다 더 효과적이고 정확하게 제안되었다. ROC 곡면에서도 부분 VUS 뿐만 아니라 세 방향 부분 VUS 통계량이 개발되었다. 본 연구에서는 ROC 곡선의 FPR과 TPR 모두에 제한된 두 개의 절단함수를 이용하여 확률 개념과 적분 표현으로 대안적인 AUC를 제안한다. 또한 이 AUC는 양방향 부분 AUC와 관계가 있음을 알 수 있다. ROC 곡면에서의 세 방향 부분 VUS도 절단함수를 이용하는 VUS와 관련되어 있음을 발견하였다. 그리고 이러한 대안적인 AUC와 VUS는 맨-휘트니 통계량으로 표현되고 추정된다. 정규분포와 확률표본을 기반으로 이들의 모수적인 추정 방법과 비모수적인 추정 방법을 탐색한다.

주요용어: 분류, 절단, 판별, FPR, TPR

[1]교신저자: (03063) 서울특별시 종로구 성균관로 25-2, 성균관대학교 통계학과. E-mail: cshong@skku.edu