Cukurova Medical Journal
ÇUKUROVA ÜNİVERSİTESİ TIP FAKÜLTESİ

Cukurova Med J 2018;43(1):141-147
DOI: 10.17826/cumj.309430

ARAŞTIRMA / RESEARCH

Çarpık dağılımlı verilerde ROC eğrisi altında kalan alan tahmininde transformasyon etkili mi?

Is the transformation useful to estimate the area under the ROC curve with skewed data?

İlker Ünal

Çukurova Üniversitesi Tıp Fakültesi Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı, Adana, Turkey

Cukurova Medical Journal 2018;43(1):141-147.

Abstract

Purpose: The aim of this study is to compare the methods used for the estimation of the area under the ROC curve (AUC) with skewed data.

Materials and Methods: The area under the ROC curve can be estimated by using parametric or nonparametric methods. In the parametric method, the distributions of biomarker in both patient and non-patient groups are assumed to be normal distribution or transformed to the normal distribution. In non-parametric methods, there are two approaches: using the rank statistics or using the kernel smoothing methods. In this study, these methods are compared by using simulated skewed data and real data.

Results: According to the simulation results, the estimates of the investigated methods are affected by the real AUC value, the sample size and the degree of the skewness. For the scenarios with the skewed normal distribution, Mann Whitney method gets the smallest bias value and for the scenario with the gamma distribution, Binormal model with a Box-Cox transformation gets the smallest bias value.

Conclusion: For skewed data, after applying transformation techniques, the Binormal model can be used to estimate the AUC. After Box-Cox transformation, the estimate of the Binormal model is very close to the real AUC value for highly skewed data. It is recommended to use this method for this situation.

Key words: ROC curve, the area under the curve, skewed data

Öz

Amaç: Hasta ve sağlıklı grupta ölçülen sayısal ölçümlü tanı testinin dağılımının çarpık olması halinde ROC eğrisi altında kalan alan tahmini için kullanılan yöntemleri karşılaştırmaktır.

Gereç ve Yöntem: ROC eğrisi altında kalan alan tahmini parametrik ve parametrik olmayan yöntemler kullanılarak yapılabilmektedir. Parametrik yaklaşımlarda sayısal ölçümün hasta ve sağlıklı gruplarında normal dağılım gösterdiği veya dönüşüm teknikleri ile dağılımın normal dağılıma dönüştürülebileceği varsayılmaktadır. Parametrik olmayan yaklaşımlarda ise eğri altında kalan alan sıra istatistikleri kullanılarak hesaplanabildiği gibi, eğrinin çekirdek düzgünleştirme yöntemleri ile elde edilen olasılık yoğunluk fonksiyonu kullanılarak da hesaplanabilmektedir. Bu çalışmada parametrik ve parametrik olmayan yöntemler çarpık olarak üretilmiş veriler ve bir gerçek veri seti kullanılarak karşılaştırılmıştır.

Bulgular: Simülasyon çalışmalarında, yöntemlerin gerçek alan değeri tahminleri gerçek alan değeri, örneklem büyüklüğü ve çarpıklık derecesine göre değişim göstermektedir. Normal dağılımdan üretilen çarpık verilerde Mann Whitney yöntemi en az hatayı yaparken, Gamma dağılımından üretilen çarpık verilerde ise normale dönüşüm sonrası Binormal modeli en az hatayı yapmıştır.

Sonuç: Çarpık dağılımlı verilerde dönüşüm teknikleri ile parametrik yöntemler kullanılarak ROC eğrisi altında kalan alan değeri tahmin edilebilir. Çarpıklık derecesi fazla olan verilerde parametrik yöntemlerin alan tahmin değerleri gerçek değere daha yakın olduğundan çarpık dağılımlı verilere Box-Cox dönüşümünü uygulamak önerilir.

Anahtar kelimeler: ROC eğrisi, eğri altında kalan alan, çarpık dağılımlar

Yazışma Adresi/Address for Correspondence: Dr. İlker Ünal, Çukurova Üniversitesi Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı, Adana, Turkey. E-mail: ilkerun@cu.edu.tr
Geliş tarihi/Received: 13.04.2016 Kabul tarihi/Accepted: 27.04.2016

GİRİŞ

Sayısal ölçümlü testler, klinik uygulamalarda sıklıkla elde edilmektedir. Bu ölçüm değerleri, uygun kesim noktaları kullanılarak niteliksel ölçüme (hasta/hasta değil gibi) dönüştürülebilir ve böylece hasta hakkında klinik karar verilebilir. Kararın doğruluğunda sayısal verinin niteliksel veriye dönüştürülmesinde kullanılan kesim noktası oldukça etkilidir. Burada önemli olan uygun kesim noktasının belirlenmesidir. İşlem karakteristiği eğrisi (Receiver Operating Characteristic Curve – ROC Eğrisi) uygun kesim noktasını belirlemede kullanılan bir grafiktir[1].

ROC eğrisi ilk olarak radar sinyallerinin analizinde kullanılmıştır[2]. Radar sinyallerinin alıcı tarafından doğru belirlenip belirlenmemesinin görsel olarak gösterildiği ROC eğrisi, eğri altında kalan alanın değerlendirilmesi, uygun kesim noktasının belirlenmesi gibi amaçlarla kullanılan bir analiz yönteminin önemli bir bileşenidir. Sağlık alanında ise tanı testlerinin değerlendirmesinde sıklıkla kullanılan ROC analizinin, özellikle radyoloji alanında yapılan çalışmalarla popülaritesi artmıştır[3,4].

Açılımından net olarak anlaşılmamasına rağmen, ROC analizi iki ampirik dağılımın karşılaştırmasında kullanılan grafiksel bir yöntemdir. ROC eğrisi, farklı kesim noktaları için belirlenen duyarlılık (doğru pozitiflik oranı) ile hatalı pozitif oranları kullanarak çizilen bir grafiktir[3]. Sayısal ölçümlü testte elde edilen her bir değer kesim noktası olarak varsayılarak testin o değerdeki duyarlılık ve hatalı pozitif oranı değerleri hesaplanır. Bunlar koordinat düzlemine yerleştirilir. Yerleştirilen noktaların birleştirilmesiyle de ROC eğrisi elde edilir[3,4].

ROC analizinde, sayısal ölçümün hasta ve sağlıklı gruplarında parametreleri farklı iki dağılımdan geldiği varsayılır. Genellikle bilinmeyen bu dağılımların uygun dönüşüm teknikleri ile normal dağılıma dönüştürülebileceği kabul edilir[1]. Bu varsayım ile eğri altındaki kalan alan parametrik yöntem ile hesaplanabilir. Bu varsayımın sağlanamadığı durumlarda ise parametrik olmayan yöntemlerle eğri altında kalan alan değeri elde edilebilir[5].

Bu çalışmada, ROC eğrisi altında kalan alan tahmininde kullanılan parametrik ve parametrik olmayan yaklaşımlar çarpık bir dağılımdan üretilen sayısal ölçümlü veriler kullanılarak karşılaştırılmıştır. Bu amaçla, farklı eğri altında kalan alan değerlerine sahip çarpık veriler üretilmiş ve alan tahmini için parametrik olmayan (Mann Whitney, Çekirdek Düzgünleştirme-Kernel Smoothing) ve parametrik olan (Binormal model, Normale Dönüşüm ile Binormal model) yaklaşımlar kullanılmıştır.

GEREÇ VE YÖNTEM

Bir tanı testinin tanısal yeterliliğini belirlemede tek bir değer üzerinden yorum yapmak sıklıkla tercih edilen bir durumdur. Testin niteliksel ölçümlü olması halinde toplam doğruluk değeri (total accuracy), sayısal ölçümlü olması halinde ise ROC eğrisi altında kalan alan değeri bu amaçla kullanılan iki ölçüttür. İki ölçüt için de, değerin 1'e yakın olması tanı koymada testin çok başarılı olduğu anlamına gelir[1].

Bu ölçütlerden biri olan eğri altında kalan alan değerinin hesaplanmasında farklı yöntemler kullanılmaktadır. Bu yöntemler hakkında bilgi aşağıda başlıklar halinde verilmiştir.

Parametrik yaklaşım

Binormal model

Binormal model yaklaşımı ilk defa McClish tarafından önerilmiştir[6]. Bu yöntemde, hastalıklı ve sağlıklı olmak üzere iki ayrı popülasyon olduğu ve ölçülen bu sayısal ölçümün iki popülasyonda da farklı dağılış gösterdiği varsayılır. Sağlıklı popülasyondaki ölçüm X ve hastalıklı gruptaki ölçüm Y ile gösterilmek üzere olasılık dağılış ifadeleri aşağıdaki şekilde gösterilebilir[6,7].

$$X \approx N(\mu\_X, \sigma\_X\text{^}2), \qquad Y \approx N(\mu\_Y, \sigma\_Y\text{^}2)$$

Burada μ_X ve μ_Y X ve Y'nin geldiği dağılımın ortalama değerlerini, $\sigma\_X\text{^}2$ ve $\sigma\_Y\text{^}2$ ise varyanslarını göstermektedir. ROC eğrisi altında kalan alan değeri (A ile gösterilir), eğrinin çiziminde kullanılan Doğru Pozitiflik Oranı (DPO) ve Hatalı Pozitiflik Oranı (HPO) değerleri kullanılarak şu şekilde elde edilebilir.

$$A = \int_{-\infty}^{\infty} DPO(c)HPO(c)dc$$
$$= \int_{-\infty}^{\infty} \left[\Phi\left(\frac{\mu_Y - c}{\sigma_Y}\right)\Phi\left(\frac{\mu_X - c}{\sigma_X}\right)\right]dc$$
$$= \Phi\left(\frac{a}{\sqrt{1+b^2}}\right)$$

Burada Φ standart normal dağılımın kümülatif yoğunluk fonksiyonunu gösterirken, belirtilen a ve b değerleri şu şekilde elde edilir[6,7].

$$a = (\mu\_X - \mu\_Y)/\sigma\_Y\,, \quad b = \sigma\_X/\sigma\_Y$$

Parametrik olmayan yaklaşımlar

Mann Whitney istatistiği (MW)

ROC eğrisi verinin ölçüm sırası üzerinden de ampirik olarak oluşturulabilir. Ölçümün değer aralığındaki her bir ölçüm değeri bir kesim noktası alınarak, bu değerin altındaki ve üstündeki birey sayıları ile ROC eğrisi için gerekli olan duyarlılık ve seçicilik değerleri elde edilebilir. Bu sıralama mantığı Mann-Whitney tarafından önerilen U istatistiğinde de görülmektedir. İlk defa Bamber[8] tarafından önerilen Mann-Whitney U istatistiği ile ROC eğrisi altında kalan alan tahmini şu şekilde elde edilebilir.

$$A = U/(m * n)$$

Burada U, Mann Whitney istatistiği, m ve n ise hasta ve kontrol gruplarındaki birey sayısını göstermektedir[8].

Çekirdek düzgünleştirme (Kernel Smoothing-KS)

Parametrik olmayan MW yaklaşımını çekirdek düzgünleştirme yöntemleri ile daha iyi hale getirmek mümkündür. Bu yaklaşımda hasta popülasyonun ölçümlerinin F kümülatif fonksiyonuna sahip bir dağılımdan, sağlıklı popülasyonun ölçümlerinin ise G kümülatif fonksiyonuna sahip bir dağılımdan geldiği varsayılır. Bu noktadan sonra, Zhou ve arkadaşları Gauss çekirdeğini (normal dağılım fonksiyonunu) seçerek, f için olasılık yoğunluk fonksiyonunu şu şekilde tanımlamıştır[9].

$$\hat{f}(t) = \frac{1}{m}\sum_{i=1}^{m}\frac{1}{h_x}\varphi\left(\frac{t - x_i}{h_x}\right)$$

Buradaki φ standart normal dağılımın olasılık yoğunluk fonksiyonu, m ise hasta sayısını göstermekte ve h_x seçimi düzeltme işlemindeki dereceyi belirlemektedir. Olasılık yoğunluk fonksiyonundan kümülatif olasılık fonksiyonuna şu şekilde geçilebilir:

$$\hat{F}(t) = \frac{1}{m}\sum_{i=1}^{m}\Phi\left(\frac{t - x_i}{h_x}\right)$$

Buradaki Φ standart normal dağılımın kümülatif yoğunluk fonksiyonunu göstermektedir. Benzer hesaplama G fonksiyonu (sağlıklı popülasyonun ölçümleri) için de yapılabilir. Bu iki fonksiyondan hareketle Lloyd eğri altında kalan alanı şu şekilde hesaplamıştır[9].

$$A = \frac{1}{nm}\sum_{j=1}^{m}\sum_{i=1}^{n}\Phi\left(\frac{x_i - y_i}{\sqrt{h_x^2 + h_y^2}}\right)$$

Burada x değerleri hastaları ve y değerleri de sağlıklıları göstermektedir. Zou ve arkadaşları h_x değeri seçimi için şunu önermiştir[9].

$$h_x = 0.9 * \min\left(s_x, \frac{iqr_x}{1.34}\right) * m^{-1.5}$$

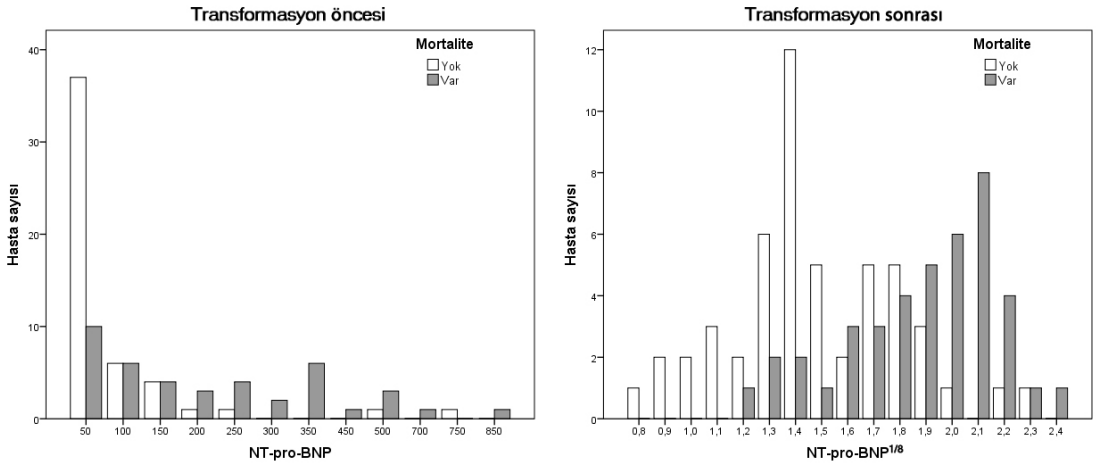
Buradaki s_x değeri ölçümdeki standart sapmayı, iqr_x değeri çeyrekler arası mesafeyi ve m ise hasta sayısını göstermektedir[9].

Verilerde çarpıklık

Popülasyon dağılımı normal dağılım olan verilerden alınan örneklemlerde bile örneklemdeki bireylerin özelliklerinden kaynaklanan nedenlerden dolayı çarpıklık gözlenebilir. Özellikle hasta popülasyonunda gözlenebilecek uç değerler dağılımda simetrinin bozulmasına yol açmaktadır. Bazı ölçümler için ise sağlıklı popülasyonda değer aralığı daha dar bir aralık olabilmektedir. Bu durumda da veri çarpık bir görünümde olabilmektedir. Çarpık verilerde Binormal model gibi parametrik yöntemleri kullanmak ROC eğrisi altında kalan alan tahmininde ciddi yanlılıklara yol açabilmektedir. Bu gibi durumlarda verinin normal dağılama yakınsaması için dönüşümler kullanılabilir. Bu dönüşümlerden en sık kullanılanı Box-Cox dönüşümüdür.

Çarpık veriler sağlık alanında yapılan ölçümlerde yapısal olarak sık gözlenmektedir. Koc ve arkadaşları[10] yaptıkları çalışmada kalp yetmezliği olan hastalarda dinlenme halinde bir nörohormon olan NT-pro-BNP düzeyini ölçmüşlerdir. Bu ölçüm yaş, cinsiyet ve vücut ağırlığı gibi fizyolojik faktörlerden etkilenir[10]. Yaş artışı ve bireyin cinsiyetinin kadın olması bu değeri arttırırken, kilo artışı düşürebilir. Bu durum da ölçümün geniş bir aralıkta değer almasına ve dağılımının çarpıklaşmasına neden olur. Koç ve arkadaşları[10] çalışmalarında NT-pro-BNP ölçümünün mortalite üzerindeki etkisini incelemiştir.

Elde ettikleri NT-pro-BNP ölçümünün bir alt grubu için ölen ve yaşayan hastalardaki dönüşüm öncesi ve sonrası dağılımı Şekil 1'de verilmiştir. Bu şekilden de görüldüğü üzere dönüşüm öncesi NT-pro-BNP düzeyi oldukça çarpık bir dağılım göstermektedir.



Şekil 1. Kalp yetmezliği hastalarındaki NT-pro-BNP düzeyinin mortaliteye göre Box-Cox dönüşümü öncesi ve sonrası dağılımları

Box-Cox dönüşümü

Box-Cox dönüşümü üstel dönüşümler ailesinden gelen ve ilk olarak George E.P. Box ve David Cox tarafından önerilen λ parametreli üstel bir dönüşümdür[11]. Bu dönüşümde y sayısal değişkenin λ kuvveti alınarak normal dağılıma yaklaştırılmaya çalışılır. Dönüşüm şu şekilde tanımlanır[11]:

$$y^{(\lambda)} = \begin{cases} \dfrac{y^{\lambda}-1}{\lambda} & eğer\ \lambda \neq 0 \\ ln(y) & eğer\ \lambda = 0 \end{cases}$$

Bu dönüşüm kullanılarak çarpık dağılımlar normal dağılıma dönüştürülebilir.

Veri üretme

Tüm analizlerde R programı 3.3.3 versiyonu kullanılmıştır. Bu programda AID[12], sROC[13] ve pROC[14] kütüphaneleri kullanılarak veri üretimi, veri dönüşümü ve ROC eğrisi altında kalan alan değerleri hesaplanmıştır. Alan tahmininde kullanılacak yaklaşımların tahmindeki başarılarını ölçebilmek amacıyla, önce normal dağılımlı ve gamma dağılımlı popülasyon verileri (n=1x107) üretilmiştir. Üretilen verilerde değişen eğri altında kalan alan değerlerini sağlayabilmek için farklı dağılım parametreleri (değişen ortalamalar) kullanılmıştır. Normal dağılım senaryosunda dağılım parametreleri; sağlıklı grupta ortalama=1.0 ve standart sapma=0.15 alınırken, hasta grubunda ortalama sırasıyla 0.91, 0.85 ve 0.77 ve standart sapma değeri=0.1 alınmıştır. Gamma dağımı için ise ölçek parametresi sağlıklı grupta 1 ve hasta grubunda 2 olarak alınırken, şekil parametresi sağlıklı grupta 0.5 ve hasta grubunda ise sırasıyla 2.06, 2.86 ve 4.2 olarak alınmıştır. Normal dağılım olarak üretilen veriler bir kuvveti (3 ve 8) alınarak çarpıklaştırılmıştır. Bu kuvvet değerleri, verideki çarpıklık derecesi az ve çok olacak şekilde seçilmiştir. Daha sonra hem gamma dağılımlı verilerden hem de çarpıklaştırılan normal dağılımlı verilerden rasgele olarak örneklemler çekilmiştir.

Bu çalışmada, önce örneklem alınan verilerin dağılım parametreleri kullanılarak Binormal model ile ROC eğrisi altında kalan alan değeri (VGA) belirlenmiştir. Daha sonra çarpıklaştırılan veri için parametrik olmayan yöntemlerden Mann Whitney yöntemi (MW) ve Çekirdek düzgünleştirme yöntemi (KS) kullanılarak eğri altında kalan alan tahmini yapılmıştır. Son olarak Box-Cox dönüşümü ile normal dağılama yaklaştırılan veriler için Binormal

modeli (TS) ile tekrar alan tahmini yapılmıştır.

Çarpıklaştırma işlemi öncesi ROC eğrisi altında kalan alan değeri 0.9, 0.8 ve 0.7 olacak şekilde belirlenmiştir. Üretilen veride çarpıklık hasta ve sağlıklı tüm bireylerin ölçümleri bir arada iken gözlendiği düşünüldüğünden veri setindeki hasta oranı 0.5 olarak sabit alınmıştır. Ayrıca çarpıklığın sonuç üzerindeki etkisini daha net görebilmek için örneklem büyüklüğü göreceli olarak küçük değerlerde alınmıştır. Karşılaştırmalarda gruplardaki hasta sayıları 25 ve 100 olacak şekilde veri setleri çekilmiş ve tüm karşılaştırmalar 1000'er tekrarla yapılmıştır. Alan tahmin yöntemlerinin gerçek değeri tahmin etme başarılarını karşılaştırmak için tahmindeki göreceli hata payları $\frac{A-\hat{A}}{A}$ ve hata kareleri ortalaması $HKO = \left(A-\hat{A}\right)^2$ hesaplanmıştır. (A gerçek alan değerini ve $\hat{A}$ ilgili yöntemin alan tahminini göstermektedir.)

BULGULAR

Simülasyon sonuçları

Yapılan karşılaştırmalara göre normal dağılımdan üretilen heterojen varyanslı çarpık dağılımlı veriler (çarpıklık yüksek değil) için elde edilen sonuçlar Tablo 1'de sunulmuştur. Bu senaryoda yöntemlerin gerçek alan değerini tahmin etmedeki başarıları benzer düzeyde bulunmuştur. Alan değerinin artması ve örneklem büyüklüğünün artması hata düzeylerinin azalmasına neden olmuştur.

Yöntemler içinde en yüksek hatayı KS yöntemi yaparken, bu yöntemi dönüştürülmemiş veriye Binormal modelin uygulandığı VGA yaklaşımı izlemiştir. Verinin normale dönüştürüldükten sonra Binormal modelin kullanıldığı TS yaklaşımı daha az hata yaparken en düşük hatayı Mann Whitney yaklaşımı yapmıştır.

Tablo 1. Heterojen varyanslı çarpık dağılım senaryosu. $X^{1/3}\sim N(1, 0.15)$, $Y^{1/3}\sim N(\mu_Y, 0.1)$

AUC	N	VGA		MW		KS		TS	
		Hata	HKO	Hata	HKO	Hata	HKO	Hata	HKO
A=0.7	25	-0,0179	0,0055	-0,0046	0,0061	-0,0819	0,0062	0,0071	0,0046
	100	-0,0180	0,0014	-0,0058	0,0014	-0,0447	0,0019	0,0109	0,0011
A=0.8	25	-0,0125	0,0044	-0,0031	0,0042	-0,1033	0,0096	-0,0036	0,0033
	100	-0,0094	0,0011	0,0003	0,0010	-0,0543	0,0027	-0,0025	0,0007
A=0.9	25	-0,0133	0,0017	0,0040	0,0019	-0,1034	0,0110	-0,0014	0,0019
	100	-0,0185	0,0007	0,0024	0,0005	-0,0558	0,0032	-0,0005	0,0005

μ_Y AUC değerlerine göre değişmektedir. Sırasıyla AUC=0.7, 0.8 ve 0.9 için μ_Y=0.91, 0.85 ve 0.77 alınmıştır. VGA: Dönüşüm öncesi alan tahmini, MW: Mann Whitney yöntemi ile alan tahmini, KS: Çekirdek düzgünleştirmeile alan tahmini, TS: Box-Cox dönüşümü sonrası alan tahmini, HKO: Hata Kareleri Ortalaması

Çarpıklığın yüksek olduğu ikinci senaryo olan heterojen varyanslı çarpık dağılım senaryosunda ise ilk senaryoya benzer şekilde, çekirdek düzgünleştirme yöntemi ve çekilen örneklem üzerinden direk Binormal modelle tahminde bulunmak diğer yöntemlere göre daha yüksek hataya neden olmaktadır. Binormal modelini veriyi Box-Cox dönüşümü ile dönüştürdükten sonra uygulamak tahmindeki hatayı azaltsa da Mann Whitney yaklaşımının doğru tahmin başarısı düzeyine ulaşamamıştır. (Tablo 2).

Tablo 2. Heterojen varyanslı çarpık dağılım senaryosu. $X^{1/8}\sim N(1, 0.15)$, $Y^{1/8}\sim N(\mu_Y, 0.1)$

AUC	N	VGA		MW		KS		TS	
		Hata	HKO	Hata	HKO	Hata	HKO	Hata	HKO
A=0.7	25	-0,1175	0,0154	-0,0040	0,0057	-0,0495	0,0040	0,0012	0,0028
	100	-0,1157	0,0114	-0,0037	0,0014	-0,0139	0,0012	-0,0106	0,0008
A=0.8	25	-0,0545	0,0041	-0,0005	0,0040	-0,0821	0,0062	-0,0108	0,0042
	100	-0,0785	0,0047	0,0003	0,0010	-0,0287	0,0012	-0,0093	0,0011
A=0.9	25	-0,0135	0,0027	0,0062	0,0020	-0,1074	0,0108	0,0007	0,0021
	100	-0,0438	0,0077	0,0012	0,0005	-0,0507	0,0024	-0,0017	0,0005

μ_Y AUC değerlerine göre değişmektedir. Sırasıyla AUC=0.7, 0.8 ve 0.9 için μ_Y=0.91, 0.85 ve 0.77 alınmıştır. VGA: Dönüşüm öncesi alan tahmini, MW: Mann Whitney ile alan tahmini, KS: Çekirdek düzgünleştirme ile alan tahmini, TS: Box-Cox dönüşümü sonrası alan tahmini, HKO: Hata Kareleri Ortalaması

Gamma dağılımında üretilen heterojen varyanslı çarpık dağılım senaryosunun sonuçları Tablo 3'te verilmiştir. Burada önceki senaryolara göre tüm yöntemler daha çok hata yapmıştır. Bu senaryoda çekilen örneklem üzerinden direk Binormal modelle tahminde bulunmak en yüksek hataya neden olmaktadır. Bu yöntemi yüksek hata düzeyi yönünden Mann Whitney yöntemi izlemektedir. KS yaklaşımı bu iki yöntemden daha az hata yapmakla birlikte, veriye dönüşüm uyguladıktan sonra Binormal modelini uygulamak en az hatanın yapılmasını sağlamaktadır.

Tablo 3. Heterojen varyanslı çarpık dağılım senaryosu. $X \sim G(0.5, 1)$, $Y \sim G(\alpha_Y, 2)$

AUC	N	VGA		MW		KS		TS	
		Hata	HKO	Hata	HKO	Hata	HKO	Hata	HKO
A=0.7	25	0,1072	0,0103	0,1105	0,0106	0,0093	0,0035	0,0141	0,0069
	100	0,1091	0,0071	0,1120	0,0073	0,0613	0,0029	0,0057	0,0018
A=0.8	25	0,0740	0,0067	0,0723	0,0063	-0,0260	0,0034	0,0090	0,0048
	100	0,0771	0,0046	0,0726	0,0041	0,0259	0,0012	0,0056	0,0013
A=0.9	25	0,0384	0,0026	0,0305	0,0023	-0,0454	0,0039	0,0016	0,0024
	100	0,0408	0,0017	0,0301	0,0011	-0,0050	0,0005	0,0014	0,0006

α_Y AUC değerlerine göre değişmektedir. Sırasıyla AUC=0.7, 0.8 ve 0.9 için α_Y=2.06, 2.86 ve 4.2 alınmıştır. VGA: Dönüşüm öncesi alan tahmini, MW: Mann Whitney ile alan tahmini, KS: Çekirdek düzgünleştirme ile alan tahmini, TS: Box-Cox dönüşümü sonrası alan tahmini, HKO: Hata Kareleri Ortalaması

TARTIŞMA

Koç ve arkadaşları[10] çalışmalarındaki NT-pro-BNP düzeyinin çarpık dağılımlı olması nedeniyle, bu değer için yapılacak ROC analizinde eğri altında kalan alan tahmini bu çalışmada verilen yöntemlerle yapılabilir. Buna göre; orijinal verinin bir alt grubu kullanılarak yapılan analizler sonucu; dönüşüm uygulamadan Binormal modelinin uygulanmasında alan tahmini 0.743, Mann Whitney yaklaşımı ile alan tahmini 0.814, çekirdek düzgünleştirme yöntemi ile alan tahmini 0.765 ve dönüşüm uygulama sonrası Binormal modeli ile alan tahmini 0.817 olarak elde edilmiştir. Şekil 1'den de görüldüğü gibi NT-pro-BNP düzeyine dönüşüm uygulansa bile tam olarak normal dağılımı elde etmek mümkün olmadığından, bu ölçümün normal dağılımdan çarpıklaştırılmış bir dağılımdan gelmediği düşünülmektedir. Gamma dağılım senaryosu sonucundan hareketle, dönüşüm sonrası Binormal modeli ile yapılan alan tahmini hata değerinin en düşük olması nedeniyle, bu veri seti için alan tahminin 0.817 olduğu sonucuna varılmıştır.

Eğri altında kalan alan değerine göre sonuçlar büyük ölçüde değişim göstermemektedir. Alan değerinin artması hata düzeylerini tüm yöntemlerde azaltmıştır. Bununla birlikte örneklem büyüklüğünün artması da hata düzeyinin azalmasını sağlamıştır.

Normal dağılımdan üretilen her iki çarpık veride de Mann Whitney yöntemi başarılı sonuçlar alırken, diğer yaklaşımlar daha hatalı sonuçlar vermektedir. Özellikle dönüşüm yapılmadan uygulanan Binormal model oldukça hatalı sonuç verirken, Box-Cox dönüşümü sonrası uygulanan Binormal modelinin daha iyi sonuçlar verdiği görülmüştür. KS yöntemi her iki senaryoda da yüksek hatalı alan tahminleri yapmıştır.

Gamma dağılımında ise önceki iki senaryoda olduğu gibi veriye dönüşüm yapmadan Binormal modelinin uygulanmasının hatayı arttırdığı gözlenmiştir. Burada önceki senaryolardan farklı olarak Mann Whitney yaklaşımının hata düzeyinin arttığı görülmüştür. KS bu yaklaşıma göre daha az hata yaparken, Gamma dağılımlı verilere dönüşüm yaparak Binormal modelinin uygulanmasının hata düzeyini azalttığı gözlenmiştir. Normal dağılımdan üretilen çarpık verilerde yapılacak ROC analizinde eğri altında kalan alanı belirlemede Mann Whitney yaklaşımı alan değerini belirlemede oldukça başarılıdır. Ancak altta yatan dağılımın yapısal olarak daha çarpık olması durumunda verinin normalleştirilmesi sonucu daha iyi hale getirmektedir.

Teşekkür

Çalışmanın gerçek veri analizinde kullanılan verileri sağlayan Dr. Mevlüt Koç'a teşekkür ederim.

KAYNAKLAR

1. Zhou X, Obuchowski N, McClish D. Statistical Methods in Diagnostic Medicine. New York, Wiley, 2002.
2. Green DM, Swets JA. Signal Detection Theory and Psychophysics. New York, NY, Wiley, 1966.
3. Zweig MH, Campbell G. Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. Clin Chem. 1993;39:561-77.
4. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. Radiology. 2003;229:3-8.
5. Faraggi D, Reiser B. Estimation of the area under the ROC curve. Stat Med. 2002;21:3093-106.
6. McClish DK. Analyzing a portion of the ROC curve. Med. Decis Making. 1989;9:190-5.
7. Metz CE, Herman BA, Shen JH. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. Stat Med. 1998;17:1033-53.
8. Bamber DC. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. J Math Psychol. 1975;12:387-415.
9. Zou KH, Tempany CM, Fielding JR, Silverman SG. Original smooth receiver operating characteristic curves estimation from continuous data: statistical methods for analyzing the predictive value of spiral CT of ureteral stones. Acad Radiol. 1998;5:680–7
10. Koç M, Bozkurt A, Acartürk E, Şahin DY, Ünal İ. Usefulness of N-terminal pro-B-type natriuretic peptide increase with exercise for predicting cardiovascular mortality in patients with heart failure. Am J Cardiol. 2008;101:1157-62.
11. Box GEP, Cox DR. An analysis of transformations. J R Stat Soc Series B Stat Methodol. 1964;26:211-43.
12. Dag O, Asar O, Ilk O. A methodology to implement Box-Cox transformation when no covariate is available. Commun Stat Simul Comput. 2014;43:1740-59.
13. Xiao-Feng Wang (2012). sROC: Nonparametric Smooth ROC Curves for Continuous Data. R package version 0.1-2. https://CRAN.R-project.org/package=sROC.
14. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, Müller M. "pROC: an open-source package for R and S+ to analyze and compare ROC curves". BMC Bioinformatics. 2011;12:77.