Derleme

Aykırı Değer Yönetimi

Management of Outliers

Havva Didem OVLA[1], Bahar TAŞDELEN[1]

[1]Mersin Üniversitesi Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı, Mersin

Özet

Diğer değerlerle karşılaştırıldığında veri setine uygun olmadığı tespit edilen aşırı değerlere aykırı değer denir. Aykırı değerlerin fazla olması veri setinin normal dağılımdan sapmasına ve yapacağımız istatistiksel analizlerin etkilenmesine sebep olabilir. Hem seçilecek yöntem hem de kullanılan hesaplamalar, farklı örnek genişliği içeren veri setlerine bağlı olarak farklılık göstermektedir. Yapılan testler sonucunda normal dağılmadığı ve fazlasıyla aykırı değer içerdiği belirlenen veri setlerine, veri dönüşüm yöntemlerinden biri uygulanarak hem verinin bilgi sağlayacak hale gelmesi ve anlamlı özet değerler üretmesi, hem de analitik yöntemlerin kullanılabilir hale gelmesi sağlanır. Bu yazının amacı veri setinin dağılımını etkileyen aykırı değerleri tespit etmek için kullanılan tanımlayıcı ve teste dayalı yöntemleri incelemek ve veri setinin dağılımını normal dağılıma benzetip, parametrik yöntemler kullanmaya olanak sağlayan veri dönüşümlerini değerlendirmektir. Dixon testi en az duyarlı ancak en hızlı sonuç veren test olarak belirlenmiştir.

Anahtar Sözcükler: aykırı değer; veri dönüştürme; örnek büyüklüğü; Dixon testi; Grubbs t testi; Walsh testi

Abstract

The extreme values differing greatly from the majority of the data set upon comparison are called outliers. If the outliers are excessive in quantity, this can result in deviation of the data set from normal distribution and as well as have an influence on the statistical analysis to be carried out. Both the method to be chosen and the calculations used show differences based on the data sets with varying sample sizes. By applying one of the data transformation methods to the data sets possessing excess number of outliers and not showing normal distribution, the data do not only become informative and produce significant summary values, but also make analytical methods available. The aims of this article are to examine descriptive and test-based methods used for detection of outliers affecting distribution of a data set, and to assess data transformations allowing application of parametric methods upon the assumption of normal distribution. Dixon test was determined as the least sensitive, but the fastest result-giving test.

Keywords: outliers; data transformation; sample size; Dixon test; Grubbs t test; Walsh test

Bu makale 12-14 Eylül 2011 Ankara, XIII. Ulusal Biyoistatistik Kongresi'nde sunulmuştur.

Mersin Univ Saglık Bilim Derg, 2012;5(3):1-8
Geliş tarihi : 04.04.2013
Kabul tarihi : 01.07.2013
Yazışma adresi : Arş. Gör. Havva Didem OVLA, Mersin Üniversitesi Tıp Fakültesi, Biyoistatistik ve Tıbbi Bilişim Anabilim Dalı, Çiftlikköy Kampüsü, 33343, Mersin
Tel : 324 3610684/1032
Faks : 324 3412400
E-posta : didemovla@yahoo.com

Giriş

Gerçek hayatta elde edilen verilerin bilimsel yöntemler kullanılarak değerlendirilebilmesi için bazı ön işlemlerden geçmesi gerekebilir. Verinin eksik olması veya tutarsız olması gibi sorunların yanı sıra aşırı veya uç değerler içermesi özellikle istatistiksel analizlerin yapılması için dikkate alınması gereken başka bir sorundur.

Veri setindeki diğer değerlerle karşılaştırıldığında veri setine uygun olmadığı tespit edilen aşırı değerlere aykırı değer (outlier) denir. Bu aykırı değerler hatalı veri girişi sebebiyle olabileceği gibi ölçüm aletinin hatalı olmasından veya tamamen deneme materyalindeki farklılıktan kaynaklanabilir. Diğer ölçümlenen değerlerden oldukça farklı olduğu için aykırı olarak belirlenmiş olan değerlerin gerçekten aykırı olup olmadığı kontrol edilmelidir. Uç değerlerin fazla olması veri setinin normal dağılımdan sapmasına ve yapacağımız istatistiksel analizlerin etkilenmesine sebep olabilir (1). Bu nedenle veri setinin dağılımını normal dağılıma benzeterek, sonuçların daha güvenilir olması sağlanabilir.

Örnek genişliği yeterince büyük olan bir veri setinin istatistiksel yöntemlerle analizi sırasında aykırı değerler pek çok araştırmacı tarafından üzerinde fazla düşünmeden analiz dışı bırakılabilir. Ancak örnek genişliği küçük olduğunda tek bir gözlemin bile analiz sonuçlarına katkısı çok değerlidir. O nedenle aykırı değerlerin doğru tespiti ve giderilmesi küçük örnek genişlikleri için büyük önem taşımaktadır. Bunun yanı sıra, aykırı değer tespiti mikrodizin verileri ve klinik biyokimya verileri gibi büyük veri setlerinin kalite kontrolü ve ilaç endüstrisi için ayrı bir öneme sahiptir.

Aykırı değer tespit yöntemleri tanımlayıcı istatistik ve teste dayalı yöntemler olmak üzere ikiye ayrılmaktadır. Standart sapma yöntemi ve box-plot grafik yöntemi tanımlayıcı istatistiklerle aykırı değer tespit yöntemlerine girerken, test yöntemleri verinin dağılım şekline göre parametrik ve parametrik olmayan test yöntemleri olarak iki şekilde incelenmektedir. Eğer veri normal dağılıma uyumlu ise aykırı olduğundan şüphelenilen değer, Dixon, Grubbs t, Rosner, Discordance, Tietjen-Moore, Nalimov ve Weisberg t testleri ile değerlendirilir. Normal dağılıma uyumlu olmayan veri setinde ise Walsh testi ile aykırı değer tespiti yapılmaktadır.

Bu makalenin amacı veri setinin dağılımını etkileyen aykırı değerleri tespit etmek için kullanılan tanımlayıcı ve teste dayalı yöntemleri incelemek ve veri setinin dağılımını normal dağılıma benzetip, parametrik yöntemler kullanmaya olanak sağlayan veri dönüşümlerini değerlendirmektir.

Aykırı Değer Tespit Yöntemleri

Bir veri setindeki değerin aykırı değer olup olmadığını belirlemek için geliştirilmiş yöntemlerden bazıları; standart sapma yöntemi, box-plot grafiği, Dixon testi, Grubbs testi, Rosner testi, Discordance testi, Weisberg t testi, Tietjen-Moore, Nalimov ve Walsh testi'dir.

Standart Sapma Yöntemi

İlgili değişkenin populasyonda normal dağılım gösterdiği biliniyorsa, çalışılan konunun hassasiyetine göre ±2 ya da ±3 standart sapmanın altında ve üstünde kalan değerler aykırı değer olarak belirlenir. Bu yöntemin kullanılabilmesi için örnek genişliği 120 ve üzeri olmalıdır (2,3).

Box-Plot Yöntemi

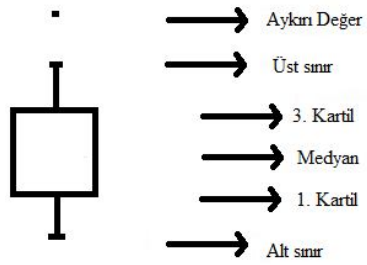
Değerlerin % 25'inin başladığı sınır 1. kartil ve % 75'inin başladığı sınır 3. kartil olmak üzere kartiller arası genişlik değeri (IQR) hesaplanır. Bu değer kullanılarak alt ve üst sınırlar belirlenir (4-6).

IQR = 3. kartil – 1. kartil

Üst sınır = 3. kartil + 3/2IQR

Alt sınır = 1. kartil – 3/2IQR

Bu sınırların dışındaki değerler aykırı değer olarak belirlenir. Box-plot grafiği ile tespit edilen aykırı değerler Şekil 1'deki gibi görünür.



Şekil 1. Aykırı değerin Box-Plot grafiğiyle tespit edilmesi

Teste dayalı yöntemler için hipotez;

H_0: Veri setinde aykırı değer yoktur

H_1: Veri setinde en az bir aykırı değer vardır

şeklinde oluşturulur.

Dixon Testi

Örnek genişliği 3 ile 25 arasında değişen veri setlerinde aykırı değer tespiti için geliştirilmiş bir yöntemdir. En düşük ve en yüksek birer değeri test edebilmektedir.

Veri seti küçükten büyüğe ya da büyükten küçüğe dizildikten sonra işlemler uygulanır. Test edilecek değerin en küçük veya en büyük olmasına göre ya da içinde bulunduğu örnek genişliğine göre uygulanacak test istatistiği değişmektedir. Tablo 1'deki formüller yardımıyla hesaplanan test istatistikleri Dixon kritik tablo değerleriyle karşılaştırılarak test edilen değerin aykırı değer olup olmadığına karar verilir (7,8).

Tablo 1. Dixon Test Formülleri

Örnek Büyüklüğü	En küçük değer için hesap	En büyük değer için hesap	Kritik Tablo Değeri
3≤n≤7	$d = \frac{x(2) - x(1)}{x(n) - x(1)}$	$d = \frac{x(n) - x(n-1)}{x(n) - x(1)}$	$\sim d_{n,\alpha}$
8≤n≤10	$d = \frac{x(2) - x(1)}{x(n) - x(1)}$	$d = \frac{x(n) - x(n-1)}{x(n) - x(2)}$	$\sim d_{n,\alpha}$
11≤n≤13	$d = \frac{x(2) - x(1)}{x(n-1) - x(1)}$	$d = \frac{x(n) - x(n-2)}{x(n) - x(2)}$	$\sim d_{n,\alpha}$
14≤n≤25	$d = \frac{x(3) - x(1)}{x(n-2) - x(1)}$	$d = \frac{x(n) - x(n-2)}{x(n) - x(3)}$	$\sim d_{n,\alpha}$

Burada n; toplam gözlem sayısını, x(n); veri setindeki en büyük değeri, x(1); veri setindeki en küçük değeri, α; anlamlılık seviyesini ve $d_{n,\alpha}$; Dixon kritik tablo değerini göstermektedir (9).

Nalimov Testi

Aynı anda bir X_i değerinin aykırı olup olmadığını test edebilir. Normal dağılım gösteren veri setlerinde uygulanılabilir. Veri setinin ortalaması hesaplanırken aykırı olduğundan şüphelenilen X_i değeri de dahil edilir (10).

S; veri setinin standart sapmasını ifade etmektedir.

Sonuca kritik değerler tablosuna bakarak karar verilir. Serbestlik derecesi f=n-2'dir.

$$q = \left| \frac{X_i - \overline{X}}{S} \right| \sqrt{\frac{n}{n-1}}$$

Rosner Testi

Örnek genişliği 25'ten fazla olduğu durumlarda tercih edilir. Rosner testi kullanılarak, bir veri setinde en fazla 10 aykırı değer belirlenebilir.

Rosner testinde, en uzak gözlemden başlayarak 0'dan 9'a kadar sıra numarası verilir. Ortalamaya en yakın değerin sıra numarası 9 en uzak değerin sıra numarası 0 ile gösterilir. En yakın değerden başlayarak ilk reddedilen değere kadar işlem yürütülür. İlk reddedilen değer ve daha sonra gelen gözlemlerin aykırı değer olduklarına karar verilir. Test istatistiğini hesaplayabilmek için aykırı değer olduğundan şüphelenilen her bir değer dışlanarak aritmetik ortalama ve standart sapma hesaplanır. Aykırı olduğundan şüphelenilen değerin sıra numarası i olmak üzere,

$$\overline{X}^{(i)} = \frac{1}{n-i} \sum_{j=1}^{n-i} (X_j)$$

ve

$$S^{(i)} = \left[\frac{1}{n-i} \sum_{j=1}^{n-i} (X_j - \overline{X}^{(i)})^2 \right]^{1/2}$$ hesaplanır.

Bu değerler yardımıyla test istatistiği;

$$R_r = \frac{\left| Y^{r-1} - \overline{X}^{r-1} \right|}{S^{r-1}} \sim \lambda_{n,r_0,\alpha}$$ şeklinde hesaplanır.

Eğer test istatistiği kritik tablo değerinden ($\lambda_{n,r_0,\alpha}$) büyükse aykırı değer olarak belirlenir (11).

Burada n; toplam gözlem sayısını, α; anlamlılık seviyesini, $\lambda_{n,r_0,\alpha}$; Rosner kritik tablo değerini, r; aykırı olduğundan şüphelenilen değer sayısını, Y; en uçtaki sapan değeri ifade etmektedir (7,8).

Discordance Testi

Küçükten büyüğe dizilmiş bir veri setinin en solunda ya da en sağında bulunan bir tek aykırı değeri tespit edebilen bir yöntemdir. Örnek genişliğinin 3 ile 50 arasında olması gereklidir. Büyük ve küçük aykırı değerler için farklı formüller kullanılır (7,8). Elde edilen sonuçlar Discordance test istatistiği kritik tablo değeriyle ($D_{n,\alpha}$) karşılaştırılarak test edilen değerin aykırı olup olmadığına karar verilir (12).

$$D_k = \frac{\overline{X} - X_{(1)}}{S} \qquad D_b = \frac{X_{(n)} - \overline{X}}{S}$$

Burada α; anlamlılık seviyesini, $D_{n,\alpha}$; Discordance kritik tablo değerini, X(n); veri setine ait en büyük gözlemi, X(1); veri setine ait en küçük gözlemi ve S;veri setinin standart sapmasını ifade etmektedir.

Weisberg t Testi

En az 15 değer içeren veri seti için aykırı değer belirlemede kullanılan bir yöntemdir. Veri setinin dağılımının normal dağılıma uyumlu olması gerekmektedir. Genellikle ilaç ve biyofarmasötik yöntem geçerliliğini sınayan veri setlerinin küçük örnek genişliğine sahip olduğu durumlar için güçlü bir aykırı değer tespit yöntemidir (13). Elde edilen test istatistiği sonucu, student t testi kritik tablo değeriyle karşılaştırılarak test edilen değerin aykırı olup olmadığına karar verilir (14).

$$t_w = \frac{(\frac{n-1}{n})^{1/2}\left(y_i - \bar{y}_{-i}\right)}{s_{-i}}$$

Burada; y_i; aykırı olduğundan şüphelenilen değeri, s_{-i}; test edilecek değerin veri setinden çıkarıldıktan sonra hesaplanan standart sapmasını ifade eder.

Grubbs T Testi

Normal dağılım gösteren, 3 ve 100 arasında gözlem içeren, veri setlerinde aykırı değer tespiti için kullanılan bir testtir. Aynı anda en fazla iki değeri test edebilir. İkiden fazla değer için testin tekrarlanması gerekir. Test işleminden önce veri seti küçükten büyüğe dizilir. Uç değer olup olmadığı test edilen her bir değer için bir T değeri hesaplanır. Hesaplanan değerler, tablo kritik değerini aştığı takdirde veri uç değer olarak kabul edilir (14-16).

Burada n; toplam gözlem sayısını, x(i); test edilen değeri, s; veri setinin standart sapmasını, x(n); veri setine ait en büyük gözlem değerini, s_{n-2}; ortalamadan en uzak iki değer veri setinden atıldıktan sonra hesaplanan standart sapma değerini, α; anlamlılık seviyesini, $G_{n,\alpha}$; Grubbs T kritik tablo değerini göstermektedir (16).

En küçük veya en büyük tek bir değerin aykırı olup olmadığını test etmek için,

$$T(1) = \frac{\left|\overline{X} - x(i)\right|}{s} > G_{n,\alpha}$$

En küçük ve en büyük birer değerin aykırı olup olmadığını test etmek için,

$$T(2) = \frac{\left|x(n) - x(1)\right|}{s} > G_{n,\alpha}$$

Tek bir uçtaki iki değerin aykırı olup olmadığını test etmek için,

$$T(3) = 1 - \left(\frac{(n-3)s_{n-2}^2}{(n-1)s^2}\right) > G_{n,\alpha}$$

istatistikleri kullanılır (15).

Tietjen-Moore Testi

Normal dağılım gösteren bir veri setinde birden fazla değerin aykırı olup olmadığını aynı anda test etmek için Grubbs testine bir alternatif olarak geliştirilmiştir. Tek bir aykırı değer test edildiğinde Grubbs ile eşdeğer sonuçlar verir.

Burada y, tüm ölçümlerin içinde olduğu durumu, y_k, aykırı olduğu düşünülen k adet ölçümün çıkarılmış halini göstermek üzere;

Veri seti küçükten büyüğe dizildikten sonra y_i 'den büyük değerlerin aykırı olup olmadığının testi için:

$$L_k = \frac{\sum_{i=1}^{n-k}(y_i - \bar{y}_k)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Benzer şekilde veri seti büyükten küçüğe dizildikten sonra y_i 'den küçük değerlerin aykırı olup olmadığının testi için:

$$L_k = \frac{\sum_{i=k+1}^{n}(y_i - \bar{y}_k)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

eşitliği kullanılmaktadır.

Elde edilen test istatistiği sonucu, Tietjen-Moore testi kritik tablo değeriyle karşılaştırılarak test edilen değerin aykırı olup olmadığına karar verilir (17).

Walsh Testi

Normal dağılım göstermeyen veri setlerinde aykırı değer tespiti için geliştirilmiş bir yöntemdir.

Örnek büyüklüğü, α=0.10 seviyesinde bir anlamlılık için 60'dan fazla, α=0.05 anlamlılık seviyesi için ise 220'den fazla olmalıdır (8). Örnek genişliğinin 60'dan küçük olduğu durumlarda uygulanması önerilmez (7).

Test için öncelikle r, aykırı olduğundan şüphelenilen değer sayısı olmak üzere;

$$c = \left|\sqrt{2n}\right|, \; b^2 = 1/\alpha \text{ ve } a = \frac{1 + b\sqrt{(c-b)^2/(c-1)}}{c - b^2 - 1}$$

hesaplanır. Daha sonra k=r+c olacak şekilde, en küçük değer için test istatistiği;

$x_r - (1+a)x_{r+1} + ax_k$ sıfırdan küçükse en küçük değerin aykırı değer olduğu varsayılır. En büyük değer için test istatistiği; $x_{(n+1-r)} - (1+a)x_{(n-r)} + ax_{(n+1-k)} > 0$ ise en büyük değerin aykırı olduğu varsayılır (8).

Yapılan testler sonucunda normal dağılmadığı ve fazlasıyla aykırı değer içerdiği belirlenen veri setlerine, veri dönüşüm yöntemlerinden biri uygulanarak gerek ham verinin bilgi sağlayacak hale gelmesi ve anlamlı özet değerler üretmesi, gerekse analitik yöntemlerin kullanılabilir hale gelmesi sağlanır (18).

Dağılım Dönüştürme Yöntemleri

Bir veri setinde aykırı değer varlığı veri setinin

dağılımını etkileyecektir. Bu nedenle normal dağılıma uyumlu olmayan veri setlerine daha güçlü olan parametrik yöntemler uygulanamayabilir. Ayrıca pek çok normal dağılım göstermeyen ve/veya varyansları homojen olmayan biyolojik ve tıbbi değişken parametrik test varsayımlarına uygun değildir. İşte bu nedenle dağılımı normal dağılıma benzetmek ve grup karşılaştırmalarındaki varyans homojenliğini sağlayabilmek için geliştirilmiş dağılım dönüştürme yöntemlerinin yanı sıra tahminlerin aşırı değerlere hassasiyetini gidermeyi amaçlayan, Kırpılmış Ortalama ve Winsorized Ortalama gibi sağlam (robust) kestiriciler de mevcuttur. Dağılım dönüştürme teknikleri uygulanırken, verideki her bir gözleme uygun matematiksel bir yöntem seçilerek dönüşüm işlemi gerçekleştirilir ve istatistiksel test dönüştürülmüş gözlemlerle yapılır. İstatistiksel testler dönüştürülmüş veriler üstünden yapılmış olsa da standart sapma, ortalama gibi özet istatistikleri ham veri üstünden hesaplanır.

Tıbbi ve biyolojik verilere uygun istatistiksel yöntemin uygulanabilmesi için seçilecek dönüşüm yöntemi önemli bir başlangıçtır. Çok sayıda gözlem içeren veri setlerinde farklı dönüşümlerin normallik ve sabit varyanslılık üstüne etkileri karşılaştırılır. Ancak az sayıda gözlem mevcut ise bu etkileri karşılaştırmak mümkün olmayabilir.

Logaritmik Dönüşüm

Sağa ya da sola çarpık, değişim aralığının çok geniş olduğu ve merkezi kısımdaki simetrinin önemli olduğu veri setlerinde varyans dengeleme dönüşümü olarak kullanılır. En yaygın olarak 10 tabanlı logaritmik dönüşüm kullanılırken, doğal logaritma olarak bilinen e tabanlı logaritmik dönüşüm de kullanılabilmektedir. Bakterilerle ilgili çarpık dağılımlarda 10 tabanlı logaritmik dönüşüm kullanılmasına karşı, genetik çalışmalarda elde edilen ekspresyon oranları için 2 tabanlı logaritmik dönüşüm tercih edilmektedir. Büyük değerler küçüklere oranla daha az önemli hale gelir. Veride sıfır ve negatif değerler varsa, tüm gözlemlere pozitif bir sayı eklenmeli ve dönüşüm öyle yapılmalıdır. Kesikli bir veri setiyle ilgileniliyor ve sıfır gözlemleri varsa her gözleme 0.5 eklenir. Veri setindeki tüm değerlerin logaritması alınarak işlem yapılır ancak sonuçlar yorumlanırken veriler eski değerlerine dönüştürülmelidir. Pek çok biyolojik değişken log-normal dağılım gösterdiğinden logaritmik dönüşüm en uygun dönüşüm yöntemidir (19). Öyle ki "Elli metrelik bir akarsuda ortalama balık miktarı 1.875'dir." demek anlamsız bir ifade olacaktır. Bunun yerine anti logaritması alınarak "Ortalama balık miktarı 75'dir" denilmelidir. Elde edilen regresyon eğimi ve regresyon sabiti gibi değerlerin sonuçları yazılırken kullanılan logaritmik dönüşüm tipi verilmelidir.

Kuvvet Dönüşümü

Gözlem değerlerinin p kuvvetinin alındığı dönüşüm şeklidir.

Verinin kaçıncı kuvvetinin alınacağını belirlemek için veri setindeki her bir değişken için 3.kartil-1.Kartil=IQR değerleri hesaplanarak logaritması alınır ve karşılık gelen medyan=(M) değerlerinin de logaritması alınarak saçılım grafiği çizilir.

Buradan log(IQR) = a + b.logM denklemine ait eğim katsayısından yararlanılarak "b=1-p" olacak şekilde kuvvet (p) değeri belirlenir.

Karekök Dönüşümü

Kuvvet dönüşümü sırasında elde edilen p değerinin 1/2'ye eşit olduğu durumdaki özel ismi karekök dönüşümüdür. Normalleştirme özelliğinin yanı sıra y arttıkça varyansın da arttığı veri setlerinde varyansı stabilize etme özelliği vardır (20). İçinde negatif ölçümler içeren veriye karekök dönüşümü uygulamadan önce tüm ölçümlere sabit bir sayı eklenerek pozitif hale getirilir. Veri setinin kuyruk kısımlarındaki simetrinin önemli olduğu Poisson dağılımına sahip kesikli veri tercih edilir (19).

Petri başına düşen bakteri kolonisi sayısı, çocuk sayısı, ölüm sayısı, gebelik sayısı gibi değişkenler için bu dönüşüm tercih edilir.

Reciprocal (Resiprok) Dönüşüm

Kuvvet dönüşümü sırasında elde edilen p değerinin -1'e eşit olduğu durumdaki özel ismi Resiprok dönüşümüdür.

Küçük sayılar büyür (0.001 1000), büyük sayılar küçülür (1000 0.001).

Y'nin artmasıyla varyansın da hızla arttığı durumda ve sağkalım verilerinde kullanılır. Veri setindeki sıfır değerlerine resiprok dönüşümü uygulanamaz.

Arcsin Dönüşümü

Binomiyal dağılım gösteren, varyansın ortalamanın bir fonksiyonu olduğu durumlarda kullanılan bir metottur. Sayıların karekökü alındıktan sonra arcsinüsünün alınmasıyla oluşur. Sonuçlar radyan cinsinden $(-\pi/2,\pi/2)$ aralığında yer alırlar (19). Bir parazitle enfekte olan dişi balıkların oranı gibi 0-1 aralığındaki değerlere açı dönüşümü uygulanır.

Kırpma Yöntemleri

Trimmed Ortalama

Ortalamanın aykırı değerlere karşı olan hassasiyetini gidermek için geliştirilmiş bir yöntemdir. Aykırı olduğu belirlenen en küçük ya da en büyük z. değerden itibaren atılır. Veri setinin büyüklüğüne göre, ister en düşük ve en büyük %5'lik isterse daha büyük (%10, %20) bir kısım veri setinden atıldıktan sonra kalan değerlerin ortalaması hesaplanır. Atılacak kısım %(2z/n) kadardır ve en çok %50'lik kısım atılabilir. Rastlantısal hatalara karşı dirençlidir.

$$\overline{X}_{tz} = \frac{1}{n-2z}\sum_{i=z+1}^{n-z} X_i$$

Burada n; veri setindeki değer sayısını, z; aykırı

olduğu tespit edilen değer sayısını ifade etmektedir.

Örneğin n=7 ve z=2 ise veri setinin %(2z/n) kadarı yani, %57'si atılarak veri setinin bu şekliyle ortalaması hesaplanmalıdır (21,22).

Winsorized Ortalama

Veri setinde aykırı değer varlığında hassaslaşan ortalamanın tahmininde kullanılan ve aykırı değerin örnekleme olan etkisini azaltan güçlü bir yöntemdir. Değerler arasındaki değişkenliğin fazla olduğu durumda %15-45 kadar aykırı değer yerine en yakın aykırı olmayan değer yazılır. Veri setinin dağılımı simetrikse, ortalamanın tahmini yansız kestirilebilir.

Winsor ortalama, veri seti küçükten büyüğe dizildikten sonra z. en küçük gözlem yerine en küçük z+1. gözlem ve en büyük k. gözlem yerine en büyük z+1. gözlem yazıldıktan sonra hesaplanır.

$$\overline{X}_{wz} = \frac{1}{n}\left(\sum_{i=z+1}^{n-z} X_i + z(X_{z+1}) + z(X_{n-z})\right)$$

Burada n; veri setindeki değer sayısını, z; aykırı olduğu tespit edilen değer sayısını ifade etmektedir.

Örneğin n=5 ve z=1 ise veri setinin %(2z/n) kadarı yani, %40'ı yerine en yakın aykırı olmayan değer yazılmalı ve veri setinin bu şekliyle ortalaması hesaplanmalıdır (21).

Uygulama

Normal Dağılıma Uyumlu Veri Setinde Aykırı Değer Tespiti

70 yaşındaki 25 erkek bireyden alınan kolesterol ölçümleri içinde aykırı olduğundan şüphelenilen en küçük ölçüm test edilecektir. Küçükten büyüğe dizilmiş olarak veriler; 75, 122, 122, 135, 136.5, 137, 141.32, 145, 146, 147.23, 151, 157, 159, 161.37, 164.92, 168, 171, 177.51, 180, 182, 184, 184.45, 190, 190.93, 206 şeklindedir.

İlk olarak Dixon testi için örnek büyüklüğü 25 olduğunda en küçük değerin aykırı değer kontrolünde kullanılan formül seçilir. Gereken bilinmeyenler yerine konulduğunda,

$$d = \frac{122-75}{190-75} = 0.409 \quad \text{olarak hesaplanır.}$$

Bu değer Dixon testi kritik tablo değeri (0.406) ile karşılaştırıldığında H_0 hipotezi red edilir (7,8). Yani 75 ölçüm değeri bu veri setine göre aykırı değer olarak tespit edilmiştir.

Grubbs t testine göre hesaplanan değer;

$$T(1) = \frac{\left|\overline{X} - x(i)\right|}{s} = \frac{157.37-75}{28.3} = 2.91$$

tablo değeri ile karşılaştırıldığında H_0 hipotezi reddedilir ve en küçük ölçümün diğer ölçümlere göre aykırı değer olduğu varsayılır (G_{tablo}: 2.66) (16).

Rosner testinde ilk olarak aykırı olup olmadığı test edilecek olan 75 ölçümüne 0 sıra numarası verilir ve bu ölçüm dışlanarak veri setinin ortalama ve standart sapması hesaplanır.

$$\overline{X}^{(1)} = \frac{1}{25-1}\sum_{j=1}^{24}(X_j) = 160.8$$

$$S^{(i)} = \left[\frac{1}{25-1}\sum_{j=1}^{24}(X_j - 160.8)^2\right]^{1/2} = 22.98$$

Elde edilen ortalama ve standart sapma değerleri kullanılarak test istatistiği;

$$R_r = \frac{|75-160.8|}{22.98} = 3.73$$

olarak hesaplanır. Bu değer Rosner kritik tablo değerinden (2.82) büyük olduğundan, 75 ölçümü diğer ölçümlere göre aykırı olarak tespit edilir (11).

Discordance testiyle en küçük ölçümün aykırı değer olup olmadığının tespiti için D_k,

$$D_k = \frac{\overline{X} - X_{(1)}}{S} = \frac{157.37-75}{28.3} = 2.91$$

olarak bulunur. Daha sonra, bu değer kritik tablo değerinden (2.663) büyük olduğundan 75 olarak ölçülen değerin aykırı değer olduğu kararına varılır (12).

Tietjen-Moore testi ile ilgili değerin aykırı olup olmadığı test edileceğinde hesaplanan L_k değeri;

$$L_k = \frac{\sum_{i=k+1}^{n}(y_i - \bar{y}_k)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{12151.01}{19218.39} = 0.632$$

olarak bulunur. Bu değer ilgili tablo değeri olan 0.696 ile karşılaştırıldığında bu ölçümün aykırı değer olduğuna karar verilir (17).

Diğer bir aykırı değer tespit yöntemi olan Nalimov testi en küçük ölçümün aykırı değer olup olmadığının tespiti için hesaplanan N değeri;

$$N = \left|\frac{X_i - \overline{X}}{S}\right|\sqrt{\frac{n}{n-1}} = \left|\frac{75-157.37}{28.29}\right|\sqrt{\frac{25}{24}} = 2.971$$

olarak hesaplanır. Bu değer Nalimov kritik tablo değerinden (1.94) büyük olduğundan, 75 ölçümü diğer ölçümlere göre aykırı olarak tespit edilir (10).

Son olarak Weisberg t testiyle aykırı değer tespiti için,

$$t_w = \frac{(\frac{24}{25})^{1/2}(75-160.8)}{22.98} = -3.66$$

olarak elde edilir. 24 serbestlik dereceli t tablosu değeri ile karşılaştırıldığında |-3.66|> 2.064 olduğundan bu

ölçümün aykırı değer olduğuna karar verilir.

Normal Dağılıma Uyumlu Olmayan Veri Setinde Aykırı Değer Tespiti (Walsh Testi)

Yetmiş yaşındaki 70 erkek bireyden alınan normal dağılıma uyumlu olmayan hdl ölçümleri içinde aykırı olduğundan şüphelenilen en küçük ve en büyük ölçümler test edilecektir. Küçükten büyüğe dizilmiş haliyle veriler; 14.80, 34.19, 34.40, 37.00, 37.98, 38.76, 38.90, 40.00, 40.20, 40.38, 40.38, 40.90, 41.00, 41.96, 41.96, 42.00, 42.00, 43.10, 43.10, 43.23, 43.24, 43.63, 43.64, 44.00, 44.00, 44.46, 45.00, 45.00, 46.39, 46.90, 47.00, 49.35, 50.20, 50.30, 50.60, 61.00, 63.00, 63.77, 64.24, 64.62, 65.00, 65.00, 65.40, 65.48, 65.98, 67.00, 67.00, 68.00, 68.00, 69.00, 69.60, 71.73, 71.83, 73.00, 73.09, 73.28, 73.28, 74.52, 74.52, 75.00, 75.00, 76.60, 76.90, 77.65, 78.00, 78.80, 79.10, 79.30, 80.00, 103.00 şeklindedir.

Örnek sayısı 60-220 aralığında olduğundan α=0.10 alınacaktır.

Buna göre,

$$c = \left[\sqrt{2*70}\right] = 12 , \qquad b^2 = \frac{1}{0.10} = 10 ,$$

$$a = \frac{1+3.16\sqrt{(12-10)(12-1)}}{12-10-1} = 2.347 \text{ ve}$$

$k = 1+12 = 13$ olarak hesaplanır.

En küçük ölçümün Walsh testi ile aykırı değer olup olmadığını test etmek için,

$$x_r - (1+a)x_{r+1} + ax_k = 14.8 - (1+2.347)*34.19 + 2.347*41 = -3.407$$

olarak bulunur. Hesaplanan bu değer sıfırdan küçük olduğu için 14.8 değerinin bu veri setine göre aykırı değer olduğu söylenebilir.

En büyük ölçümün aykırı değer kontrolünde ise;

$$x_{(n+1-r)} - (1+a)x_{(n-r)} + ax_{(n+1-k)} = 103 - (1+2.347)*80 + 2.347*74.52 = 10.14$$

olarak hesaplanmış ve bu değerin sıfırdan büyük olması sebebiyle ölçümün aykırı olduğuna karar verilmiştir.

Sonuç ve Öneriler

Veri girişi sırasında yapılan bir hata sonucu yanlış girilmiş bir veri eğer örnek genişliği yeterliyse veri setinin dağılımını etkilemeyebilir. Yine de ilgilenilen veri setinin dağılımı normal dağılıma uyum gösterse de göstermese de aykırı değer kontrolü yapılmalıdır. Bu kontrol için seçilecek yöntem örnek büyüklüğüne ve şüphe edilen aykırı değer sayısına göre değişiklik göstermektedir.

Aykırı değer tespitinde kullanılan parametrik testler karşılaştırıldığında Dixon testi haricindeki tüm test sonuçları daha duyarlı iken, Dixon testi diğer testlere göre daha az duyarlı sonuçlar vermiştir. Bunun yanı sıra hesaplama zamanı olarak en hızlı sonuç elde edilen test Dixon testi iken, Rosner ve Tietjen-Moore testleri için en yavaş sonuç elde edilen testlerdir diyebiliriz.

Aykırı değer sayısı bir ya da daha fazla ise aykırı değer olarak belirlenen değer, örnek genişliği yeterliyse dışlanabilir. Örnek genişliğinin yeterli olmadığı durumda ise regresyon denklemi kurularak, bağımsız X değişkeni kullanılarak aykırı değer olan bağımlı Y değişkeni tahmin edilir. Aykırı değer sayısı fazla olduğunda ise veri setine en uygun dönüşüm yöntemi seçilerek veri setinin dağılımı normal dağılıma uyumlu hale getirilmeye çalışılır. Normal dağılım göstermeyen veri setine uygulanacak parametrik olmayan yöntem testin gücünü düşürebileceğinden, öncelikle dönüşüm uygulayarak veriyi parametrik yöntemlerle analize uygun hale getirmek amaçlanmalıdır (27).

Ayrıca örnek genişliği arttırılarak aykırı değer yok edilemez. Veri seti küçük de olsa büyük de olsa aykırı değer hala aykırı değerdir.

Kaynaklar

1. Aktürk Z, Acemoğlu H. Sağlık Çalışanları için Araştırma ve Pratik İstatistik, Anadolu Matbaası, İstanbul, 2010:45-46,83-90.
2. Balcı Y. Laboratuvar hasta verileri kullanılarak biyokimya testlerinde referans aralıkları belirlenmesi. Uzmanlık Tezi, İstanbul, 2006.
3. Seo S. A review and comparison of methods for detecting outliers in univariate data sets. Master thesis. University of Pittsburgh, 2006.
4. Walfish S. A review of statistical outlier methods. *PharmTec* 2006;30(11):82-8.
5. McGill R, John W. Tukey and Wayne A. Larsen. Variations of box plots. *The American Statistician* 1978;32(1):12-6.
6. Tukey JW. Exploratory data analysis. Reading, Mass: Addison-Wesley Publishing Company, 1977.
7. Uckardes F, Sahinler S, Efe E. Aykırı gözlemlerin belirlenmesinde kullanılan bazı istatistikler. *KSÜ Doğa Bil Derg* 2010;13(1):42-5.
8. Yvonne J Prettyman-Beck. Environmental quality-environmental statistics, U.S. Army Corps of Engineers Washington, DC 20314-1000, 2008;253-265. Erişim: http://140.194.76.129/publications/eng-manuals/em1110-1-4014/toc.htm Erişim Tarihi: 05.02.2013.
9. Kanji GK. 100 Statistical Tests. SAGE Publication Ltd. London, 1993.
10. Lohninger H. Fundamentals of statistics. Erişim: http://www.statistics4u.com/fundstat_eng/ee_nalimov_outliertest.html# Erişim tarihi: 20.06.2013.

11. Rosner B. Percentage points for a generalized ESD many-outlier procedure. *Technometrics* 1983;25(2):165-72.

12. Navy. Handbook for Statistical Analysis of Environmental Background Data, Tech. rep Department of the Naval, Southwest Division, Naval Facilities Engineering Command, 1999, San Diego, California.

13. Seely RJ, Munyakazi L, Haury J, Simmerman H, Rushing WH, Curry TF. Demonstrating the consistency of small data sets. application of the Weisberg *t*-test for outliers. *BioPharm International 2003(1);36-58.*

14. Solak MK. Detection of Multiple Outliers ın Univarıate Data Sets, PharmaSUG 2009, Portland, Oregon. Erişim: http://www.pharmasug.org/download/papers/SP06.pdf Erişim Tarihi: 05.02.2013.

15. Burke S. Missing values, outliers, robust statistics and nonparametric methods. *LCGC Europe Online Supplement* 2001;14(2):19-24.

16. Grubbs F. Procedures for detecting outlying observations in samples. *Technometrics* 1969;11(1):1-21.

17. Tietjen G, Moore R. Some Grubbs-type statistics for the detection of several outliers. *Technometrics* 1972;14(3): 583-97.

18. Oguzlar A. Veri ön işleme. *Erciyes Üniversitesi İktisadi ve İdari Bilimler Fakültesi Dergisi* 2003;21(2):67-76.

19. McDonald JH. Handbook of Biological Statistics, 2nd ed. Sparky House Publishing, Baltimore, Maryland. 2009:160-5.

20. Petrie A, Sabin C. Medical Statistics at a Glance, 2005, Wiley-Blackwell Publishing, *Massachusetts, USA.*

21. Jose VRR, Winkler LR. Simple robust averages of forecasts: Some empirical results. *International Journal of Forecasting* 2008;24(1):163-9.

22. Beliakov G. Fast computation of trimmed mean. *Journal of Statistical Software* 2011;39(2):1-6.

23. Hawkins DM. Identification of Outliers, 1980, Chapman and Hall Ltd, New York. Erişim: http:// books.google.com. tr/ books?id= fb0OAAAAQAAJ&printsec= frontcover &hl=tr&source=gbs_ge_summary_r&cad=0#v=onepage &q&f=false Erişim Tarihi: 05.02.2013.

24. Quackenbush J. Microarray data normalization and transformation. *Nature Genetic Supplement* 2002;32(5): 496-501.

25. Jain A, Nandakumar K, Ross A. Score normalization in multimodal biometric systems. *Pattern Recognition* 2005;38(12):2270-85.

26. Chromıński K, Tkacz M. Comparison of Outlier Detection Methods in Biomedical Data. *Journal of Medical Informatics&Technologies* 2010;16(2):89-94.

27. Enli Y, Aslan D, Akalın N, Aydın Y, Yılmaztürk G, Göçhan İ, Tekintürk S, Demir S. Denizli'de yaşayan 18-40 yaş arası bireylerde farklı yöntemlerle referans aralıkların saptanması. *Turk J Biochem* 2003;28(4):228-45.