Aykırı Gözlemlerin Belirlenmesinde Kullanılan Bazı İstatistikler*

Fatih ÜÇKARDEŞ,****, Suat ŞAHİNLER***, Ercan EFE****
**K.S.Ü. Ziraat Fakültesi, Zootekni Bölümü, Kahramanmaraş
***Mustafa Kemal Üniversitesi, Ziraat Fakültesi, Zootekni Bölümü, Hatay

Geliş Tarihi: 02.03.2010 **Kabul Tarihi: 12.05.2010**

ÖZET: İstatistik çalışmaların en önemli unsurlarından biri verilerin normal dağılımlı olup olmadığıdır. Ancak, bazen araştırıcının kontrolü dışında, ortalamaya göre çok farklı bir veya birden fazla gözlem **v**erilerin normal dağılıştan sapmasına neden olabilmektedir. Verilerin normal dağılım gösterip göstermediğini belirlemek için birçok istatistik test vardır. Bu testler genellikle gözlemlerin tek etkilerini belirlemek yerine tamamını dikkate alarak hesaplandıklarından, gözlemlerin içerisindeki aykırı gözlem veya gözlemlerden kaynaklanabilecek bir sapma sıfır hipotezinin red edilmesi gerekirken edilmemesine neden olabilecektir. Bu nedenle bir veri seti içerisinde sonuca önemli etkileri olan aykırı değerlerin belirlenmesi büyük önem arz eder.

Bu çalışmada, aykırı değerlerin belirlenmesi amacıyla geliştirilmiş dixon, rosner, discordance, grubbs ve walsh testleri üzerinde durulmuş ve aykırı değerlerin olması durumunda nasıl bir yol izlenmesi gerektiği konusu tartışılmıştır.
Anahtar Sözcükler: Aykırı gözlemler, Dixon, Rosner, Discordance, Walsh

Some Statistics Using in Determination of Outliers

ABSTRACT: One of the most important components of statistical studies is whether normal distribution of data or not. However, sometimes, research can obtain some observation which can be significantly different from the mean. There are several tests to determine whether the data show normal distribution or not. These tests make calculation considering all data rather than individual observation. The error resulted from the outlier can lead to acceptance of H_0 hypothesis. Therefore, the determination of the outliers in datum is very important issue.

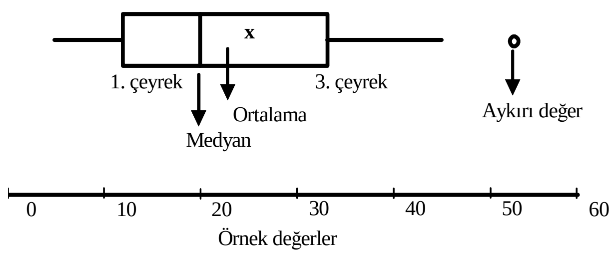
In this study, dixon, rosner, discordance, grubbs and walsh tests were evaluated to determine the outliers in data and in the presence of outliers in data the ways how to solve this problem were discussed.
Keywords: Outliers, Dixon, Rosner, Discordance, Walsh

GİRİŞ

Bilimsel çalışmalardan elde edilen verilerin büyük bir çoğunluğu normal dağılış gösterir. Birçok istatistiksel analiz yöntemi de verilerin normal dağılışlı olduğu varsayımı üzerine geliştirilmiştir. Bu nedenle asıl analizlere geçmeden önce verilere ilişkin normallik testlerinin yapılması gerekmektedir (Bek ve Efe, 1987; Akdeniz, 1998).

Dağılışının normal olması beklenen veri setlerinin normal dağılış göstermemesi durumunda ilk akla gelen nedenlerden biri de aykırı gözlemlerinin varlığıdır. Bu çalışmada aykırı gözlemlerin belirlenmesi için kullanılan bazı yöntemler incelenecektir (Alpar, 1997).

Aykırı gözlemler, veri setinin ortalamasının çok uzağına düşen gözlemler olarak ifade edilir. Bu değerler, bir tane olabileceği gibi birden fazla da olabilir. Bu değerler, verilerin standart sapmasını artırmanın dışında, dağılımın şeklini de değiştirebilir ve istatistik karar süreci sonucunda hatalı kararlar verilmesine neden olabilirler. Şekil 1'de aykırı bir değer durumu görülmektedir.



Şekil 1. Örnek bir aykırı değer durumu (West, 1999a).

Ortalamadan uzak bir noktada bulunan bu gözlemlerin aykırı olup olmadığını belirleyebilmek amacıyla geliştirilmiş birçok istatistik test mevcuttur. Bu testlerden bazıları sadece bir, bazıları da aynı anda birden fazla gözlemin istatistiksel olarak aykırı gözlem olup olmadığını belirleyebilirler. Bu durum Çizelge 1'de yer verilmiştir.

*Bu araştırma makalesi yüksek lisans tezinden üretilmiştir.
****Sorumlu yazar: Üçkardeş, F. fatihuckardes@ksu.edu.tr

Çizelge 1. Aykırı değerle ilgili bazı testler (Üçkardeş, 2006)

Testler	Normallik varsayımı	Örnek büyüklüğü	Aykırı gözlem sayısı	Test sınıfı
Dixon Test	Normal	$3 \le n \le 25$	Tek	Parametrik
Rosner Test	Normal	$n \ge 25$	Çok	Parametrik
Discordance Test	Normal	$n \le 50$	Tek	Parametrik
Grubbs Test	Normal	$n \le 50$	Tek/Çok	Parametrik
Walsh Test	Normal olmayan	$60 \le n \le 220 \quad \alpha = 0.10$ $n > 220 \quad \alpha = 0.05$	Çok	Parametrik olmayan

DİXON TESTİ

Dixon testi bir veri setinde ortalamadan uzakta yer alan bir gözlemin aykırı (extreme value) olup olmadığını tespit etmek için kullanılan bir testtir. Bu test örnek büyüklüğünün $3 \le n \le 25$ arasında olduğu ve bir tek aykırı gözlem olduğunun düşünüldüğü durumlarda kullanılır. Ayrıca bu test, aykırı olarak düşünülen gözlemin dışındaki gözlem değerlerinin normal dağılış gösterdiğini varsayar. Veriler artan sıraya göre dizildiğinde, aykırı gözlem olarak düşünülen gözlem veri setinin sağ ya da sol tarafında yer alan en uçtaki değer, bir diğer ifadeyle gözlemlerin en küçüğü ya da en büyüğü olacaktır. Dixon testi, aykırı gözlemin en küçük değere sahip gözlem mi yoksa en büyük değere sahip gözlem mi olmasına göre farklı formüller kullanır (Çizelge 2). Dixon test istatistiği aykırı değer, sıraya dizilmiş gözlemlerin en büyüğü ise büyük değer için, en küçüğü ise küçük değer için gerekli olan formül kullanılarak hesaplanır ve Dixon'un $d_{n,\alpha}$ Çizelge değeri ile karşılaştırılarak gözlemin aykırı gözlem olup olmadığına karar verilir.

Hipotez,
H_0: Veri setinde aykırı bir değer yoktur.
H_1: Veri setinde aykırı bir değer vardır.

Çizelge 2. Dixon testi için gerekli formüller (West, 1999a)

Örnek Büyüklüğü	En düşük değer için (Sol uç)	En yüksek değer için (Sağ uç)	Kritik değer
$3 \le n \le 7$	$\tau = \frac{X_{(2)} - X_{(1)}}{X_{(n)} - X_{(1)}}$	$\tau = \frac{X_{(n)} - X_{(n-1)}}{X_{(n)} - X_{(1)}}$	$\sim d_{n,\alpha}$
$8 \le n \le 10$	$\tau = \frac{X_{(2)} - X_{(1)}}{X_{(n-1)} - X_{(1)}}$	$\tau = \frac{X_{(n)} - X_{(n-1)}}{X_{(n)} - X_{(2)}}$	$\sim d_{n,\alpha}$
$11 \le n \le 13$	$\tau = \frac{X_{(3)} - X_{(1)}}{X_{(n-1)} - X_{(1)}}$	$\tau = \frac{X_{(n)} - X_{(n-2)}}{X_{(n)} - X_{(2)}}$	$\sim d_{n,\alpha}$
$14 \le n \le 25$	$\tau = \frac{X_{(3)} - X_{(1)}}{X_{(n-2)} - X_{(1)}}$	$\tau = \frac{X_{(n)} - X_{(n-2)}}{X_{(n)} - X_{(3)}}$	$\sim d_{n,\alpha}$

Çizelge 2'de,
$d_{n,\alpha}$: Dixon kritik cetvel değerini,
n: Toplam gözlem sayısını,
α : Önem düzeyini,
X_n: Veri setindeki aykırı olarak gözüken değeri göstermektedir.

ROSNER TESTİ

Rosner testi bir veri setinde 25 ve daha fazla gözlem sayısı olduğunda, verilerin normal bir dağılım gösterdiğinde ve $r_0 \le 10$'a kadar aykırı değerini hesaplayabilen bir testtir. Bu testte hem küçük hem de büyük aykırı değerler tanımlanabilir. Bu yüzden, test her zaman çift yönlüdür. Bu testin uygulanabilmesi için aykırı olarak düşünülen gözlemin ya da gözlemlerin dışındaki değerlerin normal dağılımdan gelmiş olması gerekir.

Rosner testi, veri setinde yer alan 10 aykırı değere kadar hesaplama yapabilmektedir. Rosner testinin aykırı değer sayısı arttıkça hesaplama işlemi zorlaşır. Her bir aykırı değer için, o aykırı değer işlem dışı tutulup, aritmetik ortalaması ve standart sapması hesaplanır ve formülde yerlerine koyulup çıkan değer cetvel değeri ile karşılaştırılır. Hesap değerinin cetvel değerinden büyük çıkması durumunda bu değerin aykırı olduğu kabul edilir.

Rosner testinde en uzak gözlemden başlayarak i = 0, 1, 2,…, 10 sıralama değerleri verilir. 10 ortalamaya en yakın aykırı değerin sıralamasını, 0 ortalamaya en uzak aykırı değerin sırasını ifade eder. Rosner tesinde kaç aykırı gözlemden şüpheleniliyorsa o sıra belirlenip (i), en yakın değer red edilene kadar işlem 0'a kadar yürütülür. En yakın değerin ilk kez red edildiği durumda o ve o değerin altındaki gözlemlerin aykırı değerler olduğuna karar verilir (West, 1999a).

Veriler

$$\left[\overline{X}^{(0)}, S^{(0)}, Y^{(0)}\right];.......;\left[\overline{X}^{(r_0-1)}, S^{(r_0-1)}, Y^{(r_0-1)}\right] \quad (1)$$

şeklinde düzenlendiğinde hipotezler,

H_0: Veri setinde aykırı değerler yoktur.

H_1: Veri setinde en az bir değer aykırıdır.

şeklinde yazılabilir. Ortalama ve standart sapma ise sırasıyla,

$$\overline{X}^{(i)} = \frac{1}{n-i}\sum_{j=1}^{n-i}(Xj) \quad (2)$$

ve

$$S^{(i)} = \left[\frac{1}{n-i}\sum_{j=1}^{n-i}(X_j - \overline{X}^{(i)})^2\right]^{1/2} \quad (3)$$

şeklinde ifade edilebilir. Buna göre test istatistiği,

$$R_r = \frac{\left|Y^{(r-1)} - \overline{X}^{(r-1)}\right|}{S^{(r-1)}} \sim \lambda_{n,r_{o,\alpha}} \quad (4)$$

şeklinde hesaplanır. Burada,

$\lambda_{n,r_{o,\alpha}}$: Rosner kritik cetvel değerini,

n: Toplam gözlem sayısını,

r_0: Aykırı değer sayısını,

α: Önem düzeyini,

Y: En uçtaki sapan değeri,

$\overline{X}$: Aritmetik ortalamayı,

S: Standart sapmayı ifade etmektedir.

DİSCORDANCE TESTİ

Discordance testi, veri setinin sağında ya da solunda bulunan tek bir aykırı değer için kullanılır. Eğer veri seti içindeki değer en küçük ise küçük test istatistiğini, eğer en büyük değer ise büyük test istatistiği kullanılır. Örnek sayısının $n \le 50$ ve daha aşağı olduğu durumlarda kullanılır. Bu testin uygulanabilmesi için gözlem değerlerin normal dağılımdan gelmiş olması gerekir (West, 1999a).

Test hipotezi,

H_0: Veri setinde aykırı bir değer yoktur.

H_1: Veri setinde aykırı bir değer vardır.

Test istatistiği,

$$D_k = \frac{\overline{X} - X_{(1)}}{S} \sim D_{n,\alpha} \quad (5)$$

$$D_b = \frac{X_{(n)} - \overline{X}}{S} \sim D_{n,\alpha} \quad (6)$$

şeklinde hesaplanır. Burada,

$D_{n,\alpha}$: Discordance kritik cetvel değerini,

n: Toplam gözlem sayısını,

α: Önem düzeyini,

D_k: Küçük sapma için discordance test değeri,

D_b: Büyük sapma için discordance test değeri,

$X_{(1)}$: Veri setindeki en küçük sapma değeri,

$X_{(n)}$: Veri setindeki en büyük sapma değerini,

S: Veri setinin standart sapması,

$\overline{X}$: Veri setinin aritmetik ortalamasını ifade etmektedir.

GRUBBS TESTİ

Frank Grubbs tarafından aykırı değerler için geliştirilen bu test verilerin normal dağılımdan geldiğini varsayar. Bu yüzden, bu testin uygulanabilmesi için aykırı olarak düşünülen gözlemin dışındaki değerlerin normal dağılımdan gelmiş olması gerekir. Veriler normal dağılımdan geliyorsa veriler artan sıraya göre dizilirler, ortalama ve standart sapma değerleri hesaplanır. Grubbs testinde farklı aykırı durumlara göre farklı test istatistikleri kullanılır. Test istatistiği kısmında, mümkün olabilecek üç farklı aykırı değer için kullanılacak test istatistikleri verilmiştir (Burke, 2001).

Test hipotezi,

H_0: Veri setinde aykırı değer yoktur.

H_1: Veri setinde en az bir değer aykırıdır.

Test istatistiği,

a) Sol uç ya da sağ uçta tek bir aykırı değer olduğunda,

$$G_1 = \frac{\left|\overline{X} - X_{(i)}\right|}{S} \sim G_{n,\alpha} \quad (7)$$

b) Hem sol uçta hem de sağ uçta birer aykırı değer olduğunda,

$$G_2 = \frac{X_{(n)} - X_{(1)}}{S} \sim G_{n,\alpha} \quad (8)$$

c) Her iki uçta birden fazla aykırı değer olduğunda,

$$G_3 = 1 - \left(\frac{(n-3)(S^2_{(n-2)})}{(n-1)S^2}\right) \sim G_{n,\alpha} \quad (9)$$

şeklinde hesaplanır.

Burada,

$G_{n,\alpha}$: Grubbs'un kritik cetvel değerini,

$\overline{X}$: Aritmetik ortalamayı,

X_1 : Veri setindeki en küçük gözlem değerini,

X_n: Veri setindeki en büyük gözlem değerini,

X_i: Şüpheli tek aykırı değeri (ortalamadan en uzakta olan veri),

S: Veri setinin standart sapmasını,

n: Toplam gözlem sayısını,

α: Önem düzeyini,

S^2_{n-2}: Sıraya dizilmiş veri setinden büyük değerlerinden iki verinin çıkartılarak elde edilen varyansı ifade etmektedir.

WALSH TESTİ

Walsh testi bir veri setinde çok sayıda aykırı değerleri test etmek için kullanılan parametrik olmayan bir testtir. Bu testteki veriler için herhangi bir normallik varsayımına ve dolayısı ile normallik testlerine ihtiyaç yoktur. Veri setindeki gözlenen değerler küçükten büyüğe doğru sıraya dizilir. Bu test örnek büyüklüğünün 60'tan küçük olduğu durumlarda uygulanmaz. Eğer örnek büyüklüğü $60 < n \le 220$ aralığında ise $\alpha = 0.10$, eğer örnek büyüklüğü $n > 220$ den fazla ise $\alpha = 0.05$ olarak kabul edilir (West, 1999a).

Test hipotezi,

H_0: Veri setinde aykırı değerler yoktur.

H_1: Veri setinde en az bir değer aykırıdır.

Test için bazı katsayılar,

$$c = \sqrt{2n}\ , \tag{10}$$

$$k = r + c\ , \tag{11}$$

$$b^2 = 1/\alpha\ , \tag{12}$$

$$a = \frac{\sqrt{(1 + b\left\{(c - b^2)/(c-1)\right\}}}{(c - b^2 - 1)} \tag{13}$$

olarak hesaplanır. Buna göre en küçük aykırı değer için,

$$x_r - (1+\alpha)x_{(r-1)} + ax_{(k)} < 0 \tag{14}$$

ise (hesap değeri 0'dan küçük çıktığında) en küçük değerin aykırı olduğu varsayılır.

En büyük aykırı değer için,

$$x_{(n+1-r)} - (1+a)x_{(n-r)} + ax_{(n+1-k)} > 0 \tag{15}$$

ise (hesap değeri 0'dan büyük çıktığında) en büyük değerin aykırı olduğu kabul edilir.

SONUÇ

Bu testlerin sonucunda örnek veri setinde aykırı gözlem değeri var ise ne yapmak gerekir ? Mümkünse aşırı derecede sapan değerler için sapmanın sebebi tanımlamaya çalışılır. İlk yapılması gereken en önemli şey veri girişi ve kayıtların doğru bir şekilde geçilip geçilmediğidir. Eğer buradan bir sonuç alınamaz ise, sapan gözlemleri etkileyecek iç ve dış faktörler araştırılmalıdır. Eğer buradan da yine sonuç alınamaz ise, aşağıdaki yöntemlerden biri kullanılabilir (Thode, 2001; Alpar, 1997; West, 1999b).

a) Aykırı değer veri seti aralığında ise interpolasyon yöntemi kullanılabilir. Eğer gözlem değeri ilk veya son gözlem değeri ise o zaman ekstrapolasyon yöntemi kullanılarak yeni bir değer belirlenebilir (Türkbal, 1981). Aykırı gözlemlerin yerine gözlem değerinin belirlenmesi için kullanılacak en iyi yöntemdir. Eğer bu yöntem yapılamaz ise aşağıdaki yöntemlerden herhangi biri kullanılabilir.

b) Örnek sayısı yeterli büyüklükteyse ve aykırı değerler çok değil ise (bir ya da birkaç tane) örnekten çıkartılabilirler.

c) Örnek sayısı yeterli büyüklükte değil ise, aykırı değer kendisine en yakın aykırı olmayan değere yuvarlanır.

d) Örnek setinde aykırı değer sayısı fazla ise ve bu büyük bir sorun teşkil ediyorsa, uygun bir transformasyon tekniği kullanılmalıdır.

e) Veri setinde aykırı değer var ise, ve bu çok büyük bir sorun teşkil etmiyorsa herhangi bir değişiklik yapılmadan analizler yapılır. Çalışmada bu aykırı değerlerin olduğu dipnot olarak düşülür.

f) Veriler transformasyon yöntemi ile normalliğe dönüşmüyorsa, verilerin yapısına uygun olan parametrik olmayan testlere başvurulmalıdır.

KAYNAKLAR

Akdeniz, F. 1998. Olasılık ve İstatistik, Baki Kitapevi, Adana, 546s.

Alpar, R. 1997. Uygulamalı Çok Değişkenli İstatistiksel Yöntemlere Giriş I, Spor Kitapevi, Ankara, 340s.

Bek, Y. ve EFE, E. 1987. Araştırma Deneme Metotları 1, Ç.Ü. Ziraat Fakültesi Ofset ve Teksir Atölyesi, Adana, 395s.

Burke, S. 2001. Missing Values, Outliers, Robust Statistics and Nonparametric methods. http://www.lcgceurope.com/lcgceurope/data/articlestandard/lcgceurope/502001/4509/article.pdf (01.12.2009).

Ergün, M. 1995. Bilimsel Araştırmalarda Bilgisayarla İstatistik Uygulamaları SPSS for Windows, Ocak Yayınları, Sıhhiye-Ankara, 292s.

Thode, C. H. 2001. Testing for Normality, Marcel Dekker Inc., New York, 368p.

Türkbal, A. 1981. Bilimsel Araştırma Metodları ve Uygulamalı İstatistik. Ankara Üniversitesi, İşletme Fakültesi, Yayın No: 76, Erzurum, 276

Üçkardeş, F. 2006. İstatistik Testler Üzerine Bir Çalışma. K.S.Ü. Fen Bil. Ens., Ziraat Fak., Zootekni Bölümü, Yüksek Lisans Tezi, 250s.

West, S. E. 1999a. Handbook for Statistical Analysis of Environmental BackGround Data, EPA Company, Washington, 147p.

West, S. E. 1999b. Guidance for Data Quality Assessment, EPA Company, Washington, 157p.