AYKIRI DEĞERLER (OUTLIERS)

Tanım: Verinin geri kalanından oldukça farklı olan gözlemler aykırı değer olarak adlandırılır.

Aykırı değerler tahmin edicilerin ve bu tahmin edicilere dayanan testlerin etkinliklerini olumsuz olarak etkiler.

Aykırı Değerleri Belirleme Yöntemleri

Aykırı değerleri belirleme yöntemlerini iki ana grupta incelemek mümkündür.

1. Grafiksel yöntemler
2. İstatistiksel testler

1. Grafiksel Yöntemler

Bu bölümde istatistiksel analizlerde yaygın olarak kullanılan bazı grafiksel yöntemler anlatılmıştır.

Kutu-Grafiği (Box-plot)

Kutu grafiği aykırı değer belirleme yöntemlerinin en basitlerinden bir tanesidir.

Kutu grafiği medyan ve dörtlükler kullanılarak elde edilir. Medyan ve dörtlüğün derinlikleri

$$Medyanın\ derinliği = \frac{n+1}{2}$$

$$Dörtlüğün\ derinliği = \frac{[\![Medyanın\ derinliği]\!] + 1}{2}$$

kullanılarak hesaplanır. Buradan, dörtlüğün yayılımı

$$d_F = F_U - F_L$$

eşitliği yardımıyla bulunduktan sonra alt ve üst kesim noktaları

$$c_L = F_L - 1.5d_F \text{ ve } c_U = F_U + 1.5d_F$$

hesaplanır. (c_L, c_U) aralığının dışına düşen gözlem değerleri aykırı değer olarak belirlenir.

Örnek: Bu ve diğer örneklerde aykırı değer belirleme yöntemleri incelenirken aşağıdaki veri seti kullanılacaktır.

174 166 128 175 188 187 182 171 189 178 171 195 192 180 183 235 166
170 178 168 193 166 169 166 169 177 173 198 180 185 213 208 182

Aykırı değerler belirlenirken öncelikle veriler küçükten büyüğe doğru sıralanır.

128 166 166 166 166 168 169 169 170 171 171 173 174 175 177 178 178
180 180 182 182 183 185 187 188 189 192 193 195 198 208 213 235

Bu veri seti için medyan, alt ve üst dörtlükler aşağıdaki gibi bulunur.

$$Medyanın\ derinliği = \frac{33}{2} = 16.5 => Medyan = \frac{178 + 178}{2} = 178$$

$$Dörtlüğün\ derinliği = \frac{[\![16.5]\!] + 1}{2} = 8.5$$

$$F_U = 188.5, F_L = 169.5.$$

dörtlüğün yayılımı ise

$$d_F = 188.5 - 169.5 = 19$$

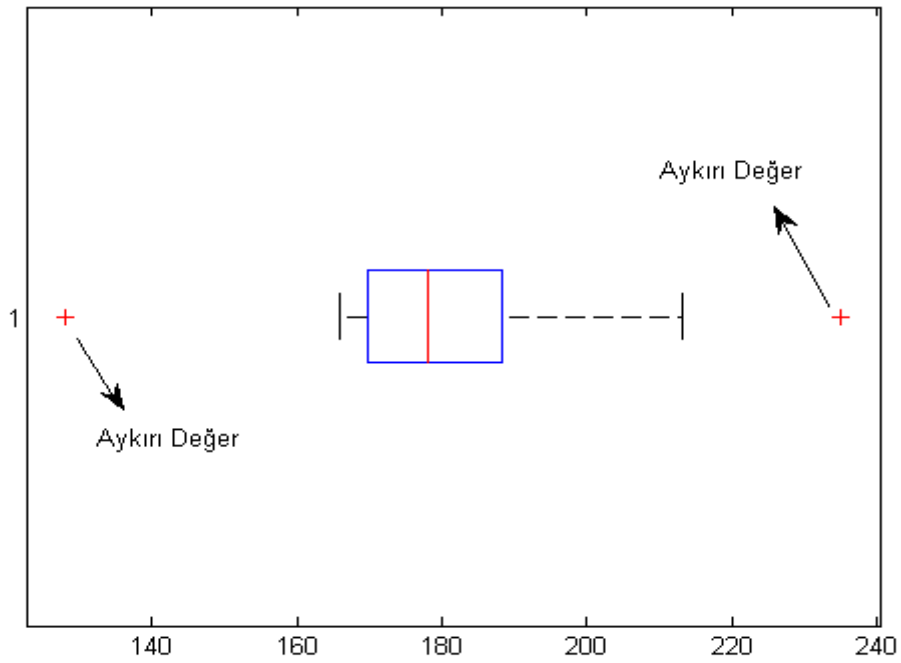
olarak bulunur.

Bu değerler kullanılarak alt ve üst kesim noktaları sırasıyla

$$C_L = F_L - 1.5d_F = 169.5 - 1.5 * 19 = 141$$

$$C_U = F_U + 1.5d_F = 188.5 + 1.5 * 19 = 217$$

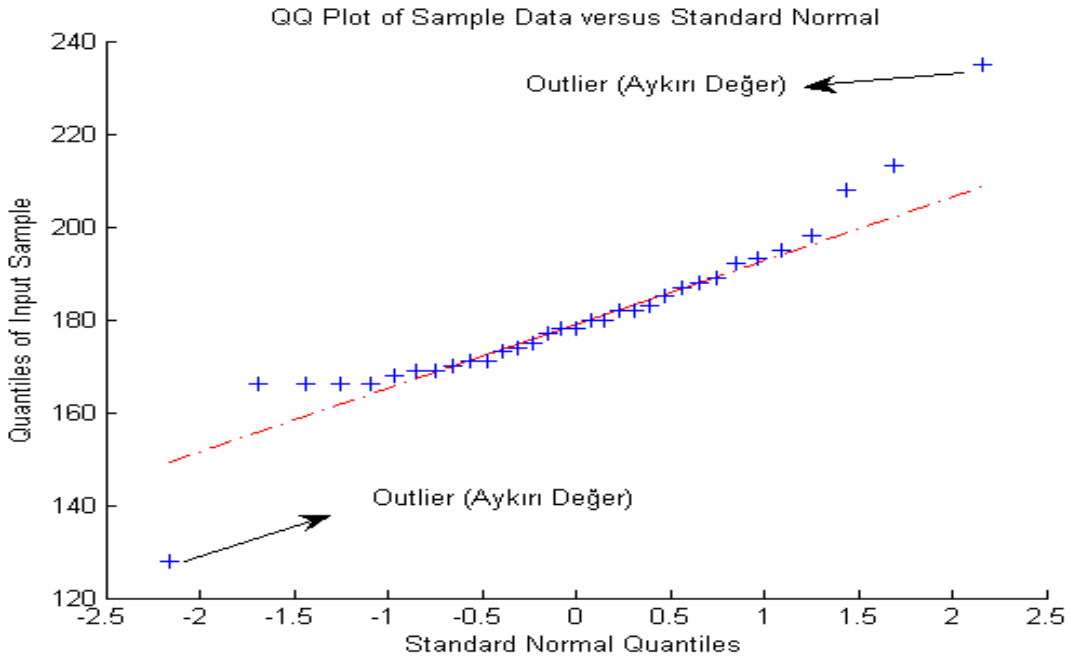
olarak hesaplanır. $235 > 217\ ve\ 128 < 141$ olduğundan bu veri setinde 128 ve 235 aykırı değer olarak belirlenir.



Q-Q Grafik Tekniği

Q-Q grafiği aykırı değer belirlenirken uygulamada yaygın olarak kullanılan grafiksel bir yöntemdir. Kullanılmasının kolay olması bir avantaj olmakla beraber subjektif bir yöntem olması bu tekniğin bir dezavantajıdır.

Aynı örnek için normal Q-Q grafiği aşağıdaki gibidir.



Histogram

Histogram kullanarak veri setinin

1. Konumu (Merkezi)

2. Yayılımı

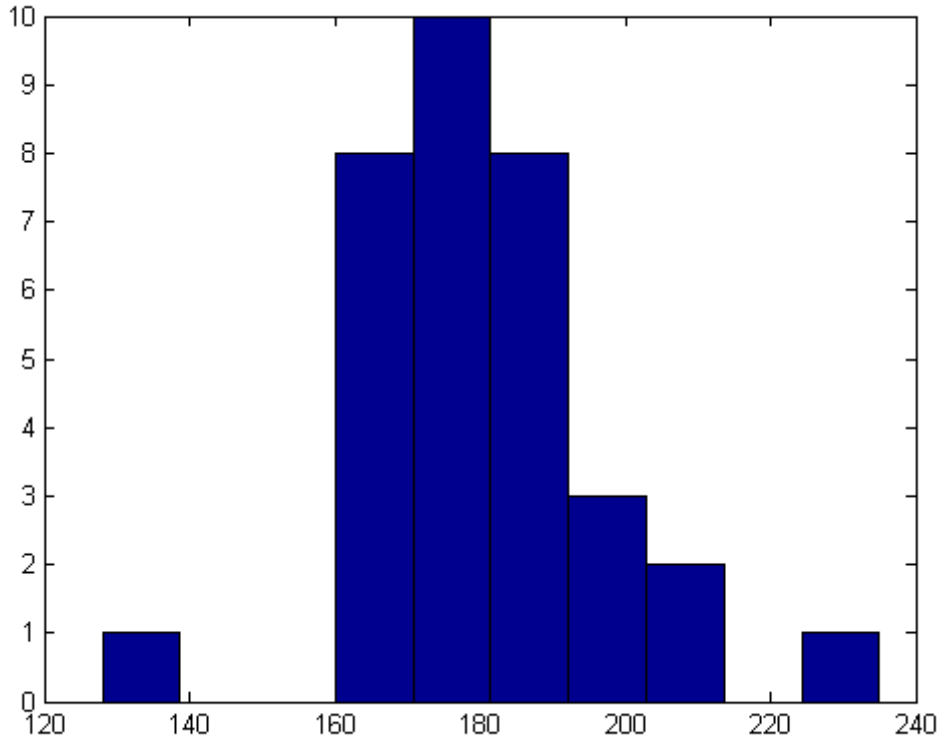
3. Çarpıklığı

4. Aykırı Değerler

5. Mod (Tepe Değer)

gibi özellikleri hakkında bilgi sahibi olunabilir. Bu nedenle, histogram aykırı değer belirlenirken kullanılan grafiksel bir yöntemdir.

Aynı örnek için histogram aşağıdaki gibidir.



Görüldüğü üzere verinin büyük bir kısmı 160 ile 215 aralığında değişmektedir. Buradan, 128 ve 235'in aykırı değer olabileceği sonucu ortaya çıkar.

Akış Dizisi (Run Sequence) Grafiği

Akış dizisi grafiği de aykırı değer belirleme yöntemlerinden bir tanesidir. Grafiğin X eksenine gözlem değerleri Y eksenine ise bu gözlem değerlerinin sıra numaraları konularak koordinat sisteminde gözlem değerleri işaretlenir. Grafiğin akışını bozan değerler aykırı değer olarak belirlenir.

Aynı örnek için akış dizisi grafiği aşağıdaki gibi elde edilir. Bu grafikten de görülebileceği gibi 128 ve 235 değerleri aykırı değer olarak belirlenir.