

## Veri Analizinde Grafiksel Yöntemler

### 1.Histogram

Histogram, gruplandırılmış bir veri dağılımının sütun grafiğiyle gösterimidir. Bir histogramın amacı, tek değişkenli bir veri kümesinin dağılımını grafiksel olarak özetlemektir. Histogram grafiksel olarak aşağıdakileri gösterir:

- Verilerin merkezini (yani konumunu).
- Verinin yayılımını
- Verilerin çarpıklığını.
- Aykırı değerlerin varlığı.

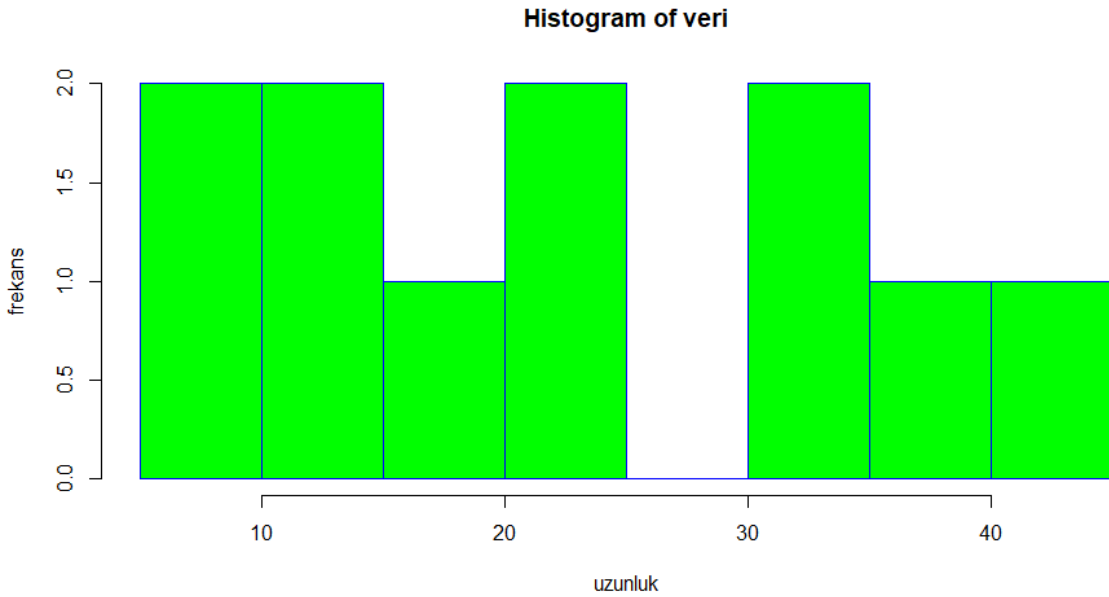
Bu özellikler, verilerin uygun dağılımı için güçlü göstergeleri sağlamaktadır.

Dersimizde kullandığımız R, *hist()* işlevini kullanarak histogram oluşturur. Bu işlev şu şekildedir; *hist(x, ...)* burada x kullandığımız verinin ismiyken noktalarla ifade edilen kısımlar çok sayıda parametreden oluşur. Bunları teker teker incelemek adına help kısmına *hist* yazarak ulaşılabilir.

**R üzerinden bir örnek:**

```
# grafik için bir veri seti oluşturalım.  
veri <- c(9,13,21,8,36,22,12,41,31,33,19)  
  
# histogram çizdirir.  
hist(veri,xlab = "uzunluk",ylab="frekans",col = "green",border = "blue")
```

Üstteki kodlarla temel olarak aşağıdaki gibi bir histogram çizdirilebilir.



## 2.Kutu Grafiği (Box-plot)

Veri kümeleri arasındaki benzerlik ya da farklılıkları görmek için kullanılır. Kutu-grafigi kullanılarak veri kümesinin

- Konumu
- Yayılımı
- Çarpıklığı
- Kuyruk uzunluğu
- Aykırı değerleri

tespit edilebilmektedir. Veri setinin yayılımı ve konum değerlerinin yardımıyla verilerin dağılımı hakkında ipucu da vermektedir.

Ayrıca, birden fazla veri seti olduğunda kutu grafikleri veri kümeleri arasındaki benzerlik ve farklılıkları görmek için kullanılır. Bu ilerleyen konularda gruplar arası kıyaslama yapılırken bizlere testlerimizi destekleyici argümanlar sunacaktır.

**Örnek:** Aşağıdaki veri setinde Türkiye’de nüfusu en çok olan 16 şehrin 2018 yılına ait nüfus sayıları 10 binlik olarak aşağıda verilmektedir. Bu verilere göre kutu grafiğini çizmek için gerekli olan noktaları hesaplayınız ve kutu grafiğinden çıkan sonucu yorumlayınız.

No	Şehir	Nüfus(×10000)
1	Samsun	134
2	Kayseri	139
3	Manisa	143
4	Hatay	161
5	Diyarbakır	173
6	Mersin	181
7	Kocaeli	191
8	Gaziantep	203
9	Şanlıurfa	204
10	Konya	221
11	Adana	222
12	Antalya	243
13	Bursa	299
14	İzmir	432
15	Ankara	550
16	İstanbul	1506

Bu veri setindeki toplam gözlem sayısının 16 olduğu görülür.16 elemanımız olduğu için medyanı bulurken veri setinde yapılan sıralamaya göre ortadaki iki elemanın ortalamasını alırız. Bunlar 8. Ve 9. şehirlerdir. Buna göre **medyan yani ortanca değer 203.5** bulunur.

Q1 değerini bulmak içinse 4. ve 5. değerlerin ortalaması alınır ve  $\frac{161+173}{2} = 167$  elde edilir.

Q3 değerini bulmak içinse 12. ve 13. değerlerin ortalaması alınır ve  $\frac{243+299}{2} = 271$  elde edilir.

Alt sinir yani minimum değeri için şu formül kullanılır,

$$\text{Min. Değer} = Q1 - 1.5 \times (Q3 - Q1)$$

Üst sinir yani Maksimum değeri için şu formül kullanılır,

$$\text{Maks. Değer} = Q3 + 1.5 \times (Q3 - Q1)$$

Bu formüllere göre alt sinir 11, üst sinir 427 bulunur.

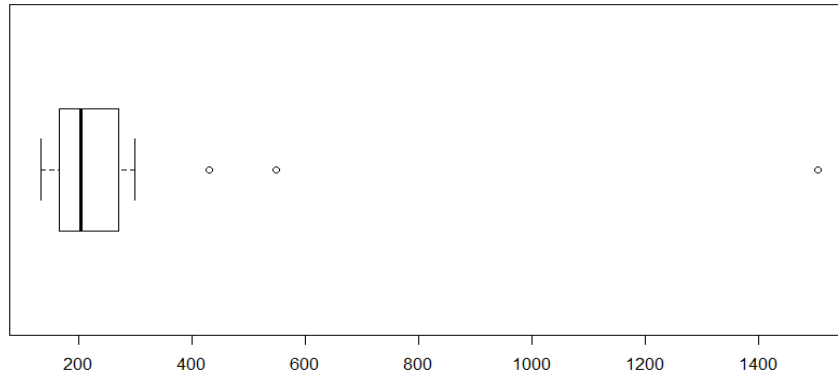
Bu veri setinde nüfusu 110.000'den az olan şehir bulunmamaktadır. Ancak, İstanbul, Ankara ve İzmir şehirlerinin nüfusları 4.270.000'den fazla olduğu için bu şehirler nüfus bakımından aykırı değer olarak belirlenir.

Yukarıdaki veri seti için bu adımlar izlenerek elde edilen kutu-grafiği aşağıdaki gibi elde edilir.

**R programında** bu kutu grafiğini çizdirmek istersek aşağıdaki kodu yazarız.

```
veri<- c(1506,550,432,299,243,222,221,204,203,191,181,173,161,143,139,134)
boxplot(veri,horizontal = TRUE)
```

ve aşağıdaki kutu grafiği şeklini elde ederiz.



### Yorum

- Kutu grafiğine göre 3 tane gözlem noktası aykırı gözlem olarak tespit edilmiştir. Buna göre; İstanbul, Ankara ve İzmir nüfus bakımından aykırı değerdir.
- Ayrıca bu grafik verilerin dağılımı açısından ipucu vermiştir ve verilerin sağa çarpık olduğunu söyleyebiliriz.

### 3.Q-Q Grafiği (Q-Q Plot)

Kantil kantil (q-q) grafiği, iki veri setinin ortak bir dağılıma sahip popülasyonlardan gelip gelmediğini belirlemek için kullanılan grafiksel bir tekniktir.

Bir q-q grafiği, ikinci veri kümesinin miktarlarına karşı birinci veri kümesinin miktarlarının grafiğidir. 45 derecelik bir referans çizgisi de çizilir. İki veri seti aynı dağılıma sahip bir popülasyondan geliyorsa, noktalar yaklaşık olarak bu referans çizgisi boyunca düşmelidir. Bu

referans çizgisinden ne kadar uzaklaşırsa, iki veri kümesinin farklı dağılımlara sahip popülasyonlardan geldiğine dair kanıtlar o kadar büyük olur.

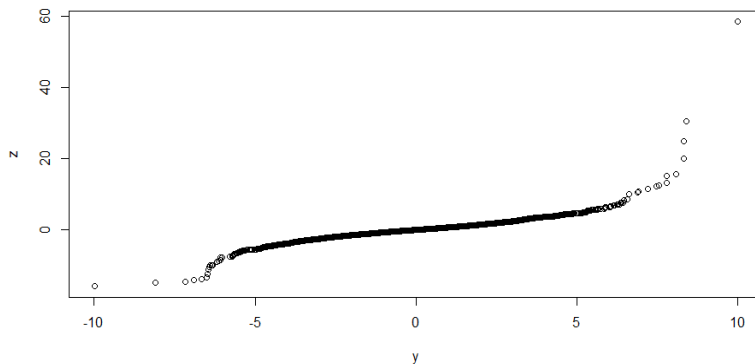
#### Q-q grafiğinin avantajları:

- Örnek boyutlarının eşit olması gerekmez.
- Dağılımla ilgili olarak bilgi verebilecek birçok argüman aynı anda test edilebilir. Örneğin, konumdaki değişimler, ölçek değişimleri, simetri değişiklikleri ve aykırı değerlerin varlığı bu grafikten tespit edilebilir. Örneğin, iki veri kümesi, dağılımları yalnızca konumdaki bir kayma ile farklılık gösteren popülasyonlardan geliyorsa, noktalar 45 derecelik referans çizgisinden yukarı veya aşağı yer değiştiren düz bir çizgi boyunca uzanmalıdır.
- Ayrıca normallik varsayım testi olarak da kullanılabilen bu grafik oldukça faydalıdır.

#### R üzerinden bir örnek:

```
y <- rlogis(10000) #lojistik dagilimdan rastgele elde edilmiş veriler
z <- rt(10000, 3) #serbestlik derecesi 3 olan t dagilimindan rastgele elde edilmiş 1000 veri
qqplot(y, z)
```

Burada ilk iki satırda yorum kısmında ifade edildiği gibi lojistik ve t dağılımlarından veri türetildi. Bu verilerin dağılımlarının ortusup ortusmedigine q-q grafiğiyle karar vermek istenirse 3. Satırdaki qqplot() işlevi uygulanır ve aşağıdaki grafik elde edilir.



Bu grafikte y ve z verileri 45 derecelik düz bir çizgi üzerinde buluşmamışlardır. Bu da bizlere verilerin dağılımlarının farklı olduklarını göstermektedir.

#### 4. Dal-Yaprak Grafiği (Stem-and-Leaf Plot)

Nicel bir değişkenin dal ve yaprak grafiği, veri öğelerini en önemli sayısal rakamlarına göre sınıflandıran metinsel bir grafikdir. Buna ek olarak, grafiği okunabilirlik için basitleştirmek amacıyla her bir alternatif satırı bir sonraki satırla birleştiririz.

Kodumuzu aşağıdaki gibi oluşturduktan sonra alttaki şekli elde ederiz.

```
# grafik için bir veri seti oluşturalım.
veri <- c(9,13,21,8,36,22,12,41,31,33,19)
stem(veri)
```

Eldeki verinin özetinin aşağıdaki gibi olduğu gözlenir.

The decimal point is 1 digit(s) to the right of the |

0		89
1		239
2		12
3		136
4		1

Kaynaklar:

- PROF. DR. BİRDAL ŞENOĞLU Ankara Üniversitesi Açık Kaynak Ders Notları.
- [https://www.tutorialspoint.com/r/r\\_boxplots.htm](https://www.tutorialspoint.com/r/r_boxplots.htm)
- Chambers, John, William Cleveland, Beat Kleiner, and Paul Tukey, (1983), Graphical Methods for Data Analysis, Wadsworth.
- Tukey, John (1977), EDA Exploratory Data Analysis, Addison-Wesley.