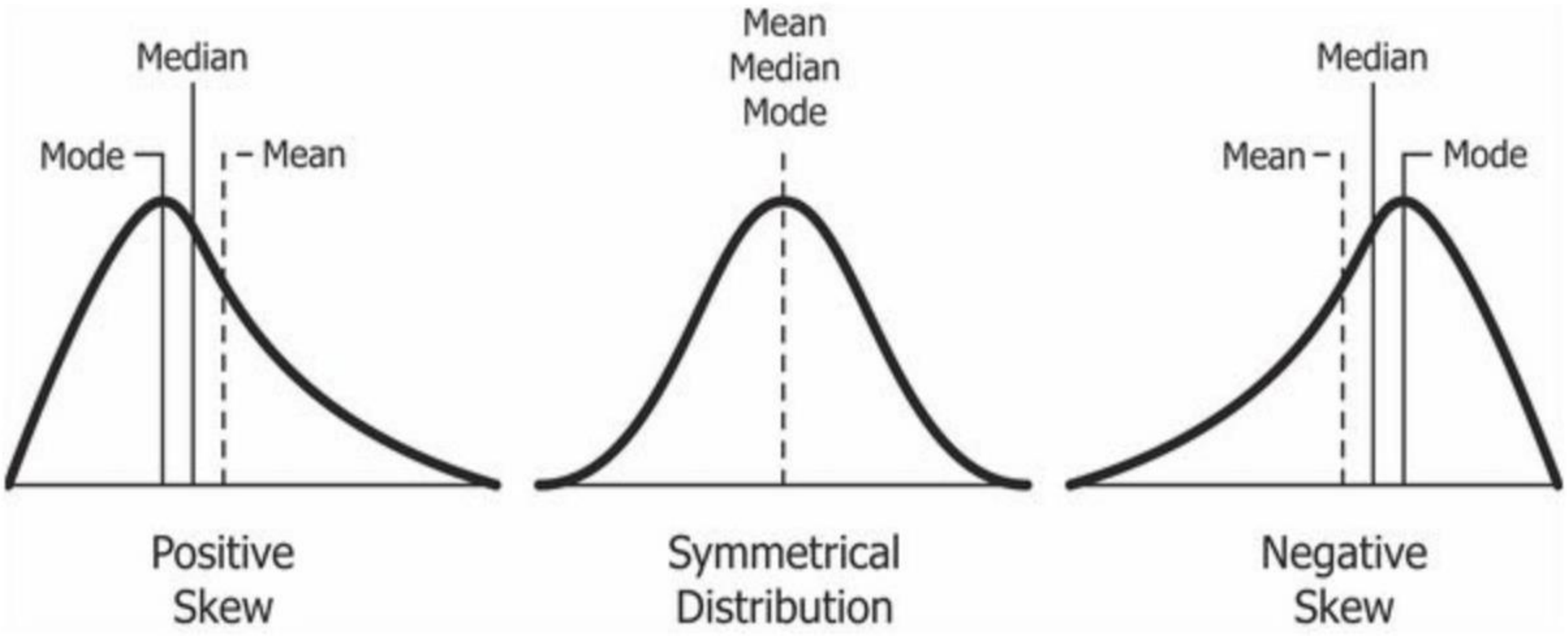
IST2084/ IST104.1/ IST104.2 Biyoistatistik 4. Hafta

Doç. Dr. Fatih KIZILASLAN

http://mimoza.marmara.edu.tr/~fatih.kizilaslan/

Aritmetik ortalama, mod ve medyan arasındaki ilişki



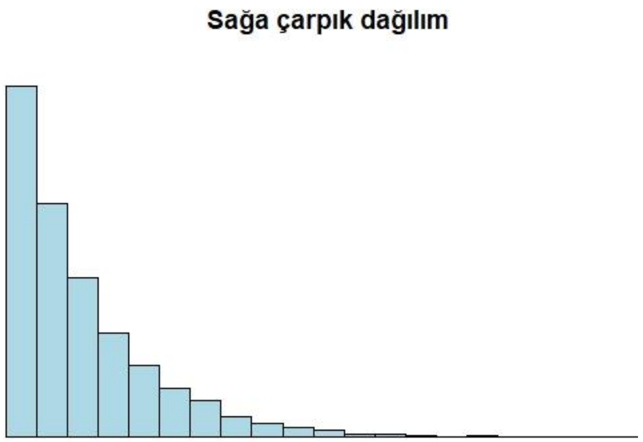
Sağa (pozitif)çarpık dağılım
Mod<Medyan<Ortalama

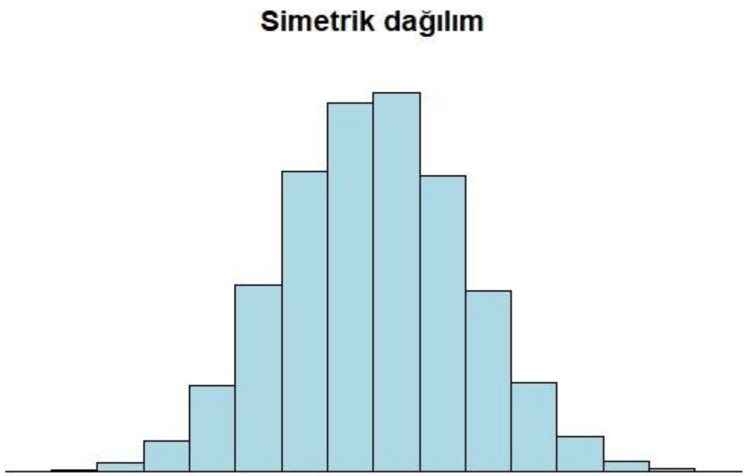
Simetrik dağılım
Ortalama=Medyan=Mod

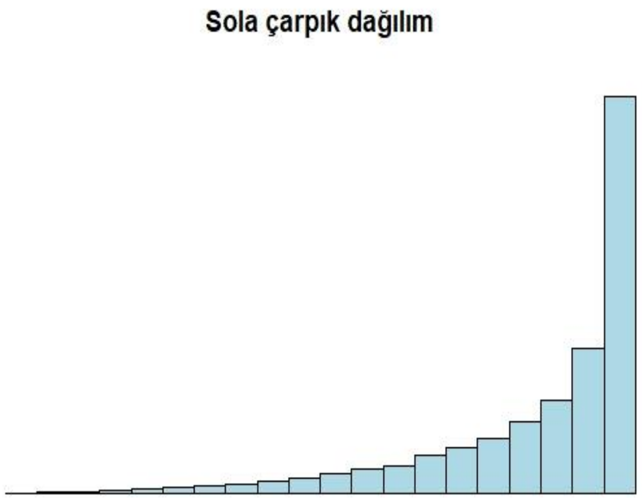
Sola (negatif) çarpık dağılım
Ortalama<Medyan<Mod

Grafik kaynağı: https://en.wikipedia.org/wiki/Skewness

- Çan eğrisi (normal dağılım) grafiğine benzer simetrik bir grafiğe veya histogram grafiğine sahip bir veri için ortalama, mod ve medyan değerleri birbirine eşittir.
- Dağılım veya histogram grafiği simetrik değil ise bu verinin dağılımının grafiği çarpıktır denir. Çarpıklık simetriden uzaklaşmayı anlatmak için kullanılır.
- Merkezi dağılım ölçüleri bölümünde verinin dağılımının grafiği için çarpıklık ve basıklıktan bahsedeceğiz.







Merkezi Dağılım Ölçüleri

Varyans ve Standart Sapma

- **Varyans** gözlem değerlerinin aritmetik ortalamadan sapmalarının (ayrılışlarının) karelerinin ortalaması olarak tanımlanır.
- Varyansın pozitif kareköküne **Standart Sapma** denir.
- En sık kullanılan dağılım ölçüleridir.
- Popülasyon (kitle) için varyansı σ^2 ve standart sapmayı σ (Sigma) ile gösteririz.
- Örneklem için varyansı S^2 ve standart sapmayı S ile gösteririz.

Örneklem için varyans ve standart sapma

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

n: toplam gözlem sayısı, x_i: i. gözlem değeri, $\bar{x}$: gözlemlerin ortalaması

Standart sapma, varyans S^2 nin pozitif kareköküdür, yani

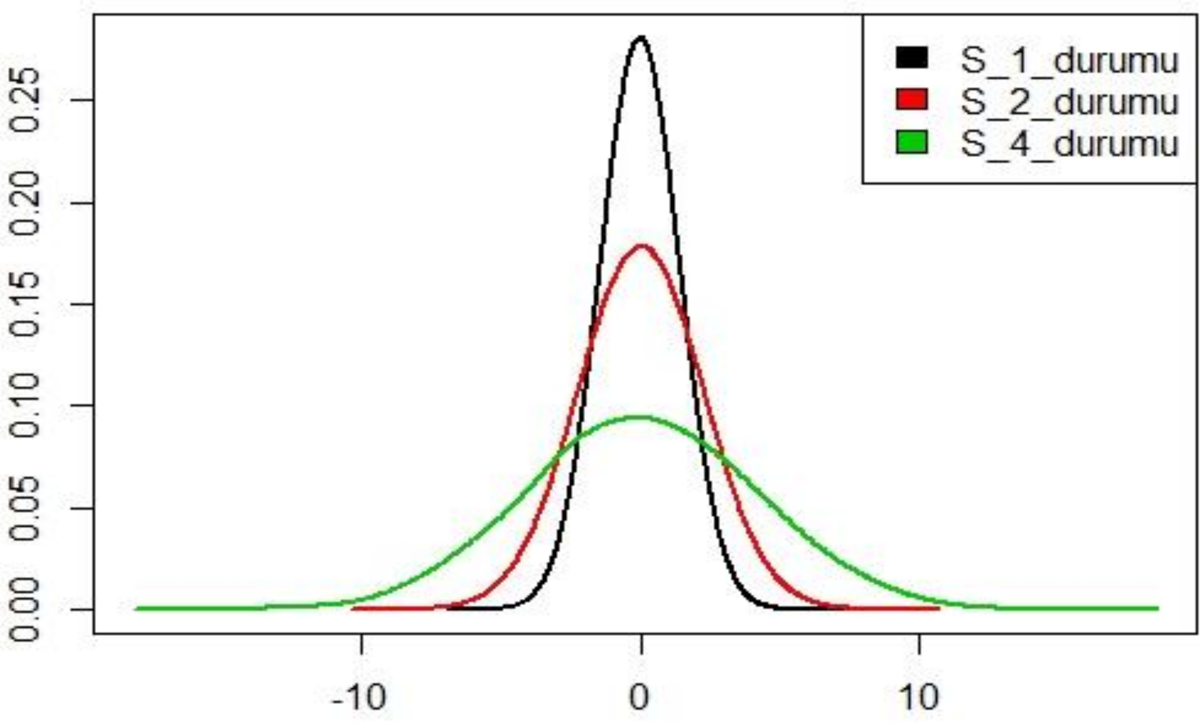
$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

biçiminde hesaplanır.

- Varyansın birimi gözlemlerin birimlerinin karesidir. Örneğin cm olan uzunluk verileri için varyansın birimi cm^2 olacaktır.
- Standart sapmanın birimi ise gözlemlerin birimleri aynıdır. Bu nedenle verinin yayılımı için genellikle tercih edilir.
- Gözlemlerin ortalama etrafında yayılımı genişledikçe yani ortalamadan farklılaştıkça standart sapma (ve de varyans) büyür. Tersine eğer gözlem değerleri birbirine benzer ise ortalamadan sapma büyük olmayacağı için standart sapma küçük olur.
- Sonuç olarak

Eğer standart sapma (veya varyans) küçük olduğunda gözlemlerin birçoğu birbirine benzerdir yani gözlemler homojendir diyebiliriz.

Eğer standart sapma büyük ise gözlemlerin çoğunluğu birbirlerinden farklıdır yani gözlemler heterojendir diyebiliriz.



Not: Buradaki gibi bir karşılaştırma yapabilmek için verilerin ortalamaları aynı olmalıdır. Dolayısıyla, standart sapma ile ortalamaları aynı olan verilerin yayılımları hakkında yorum yapabiliriz.

- Grafikte **ortalamaları aynı** 0 olan 3 farklı verinin dağılım grafikler çizdirilmiştir.
- Siyah ile çizilen I. veri için standart sapma $S_1 = 1$.
- Kırmızı ile çizilen II. veri için $S_2 = 2$.
- Siyah ile çizilen III. veri için $S_3 = 4$.
- Standart sapması büyük olan verinin yayılımı standart sapması küçük olan verilere göre daha fazladır.
- Bu grafikte $S_1 < S_2 < S_3$ olduğundan bu verilerin homojenlikleri için I. veri diğerlerine göre daha homojendir ve II. veri III. göre daha homojendir diyebiliriz.

Değişim Katsayısı

- Farklı veri setlerinin ortalamaları farklı olduğunda varyans veya standart sapmayı kullanarak bu verilerin değişkenliklerini karşılaştırmak mümkün değildir.
- Ayrıca, karşılaştırılan verilerin birimleri farklı olduğunda (örneğin yaş ve yıllık maaş gibi) birimsiz bir ölçüte ihtiyaç duyulur.
- Bu gibi durumlarda değişim katsayısı kullanılır. Değişim katsayısını DK ile göstereceğiz ve standart sapma / ortalama dır.
- Popülasyon için $DK = \frac{\sigma}{\mu}$ ve örneklem için $DK = \frac{S}{\bar{x}}$ olarak tanımlanır.
- Değişim katsayısı **küçük** olan verinin gözlem değerleri arasındaki değişimin az yani **homojen** olduğunu, değişim katsayısı **büyük** olanın ise gözlem değerleri arasındaki değişimin daha çok yani **heterojen** bir veri olduğunu söyleyebiliriz.

Örnek: Bir tansiyon hastasının 10 günlük büyük ve küçük tansiyon değerleri (mm) olarak aşağıda verilmiştir.

Büyük tansiyon: 125, 140, 130, 136, 150, 135, 134, 155, 140, 145

Küçük tansiyon: 65, 85, 75, 80, 90, 65, 80, 95, 85, 80

Bu verilerin her biri için ortalama, standart sapmayı bularak verilerin homojenliklerini karşılaştırınız.

Çözüm: Büyük tansiyon için $x_1, \ldots, x_{10}$ ve küçük için $y_1, \ldots, y_{10}$ kullanalım.

$\bar{x} = \frac{1}{10}\sum_{i=1}^{10} x_i = \frac{1}{10}[125 + 140 + \cdots + 145] = 139$

$$S_X^2 = \frac{1}{10-1}\sum_{i=1}^{10}\{(-14)^2 + (1)^2 + \cdots + (6)^2\} = \frac{742}{9} = 82.44$$

$S_X = 9.08$ ve $DK_1 = \frac{9.08}{139} = 0.06$ olarak bulunur.

$$\bar{y} = \frac{1}{10}\sum_{i=1}^{10} y_i = \frac{1}{10}[65 + 85 + \cdots + 80] = 80$$

$$S_Y^2 = \frac{1}{10-1}\sum_{i=1}^{10}\{(-15)^2 + (5)^2 + \cdots + (0)^2\} = \frac{850}{9} = 94.44$$

$S_Y = 9.72$ ve $DK_2 = \frac{9.72}{80} = 0.12$ olarak bulunur.

Böylece, $DK_1 < DK_2$ olduğundan büyük tansiyon değerleri küçük tansiyon değerlerine göre daha homojendir.

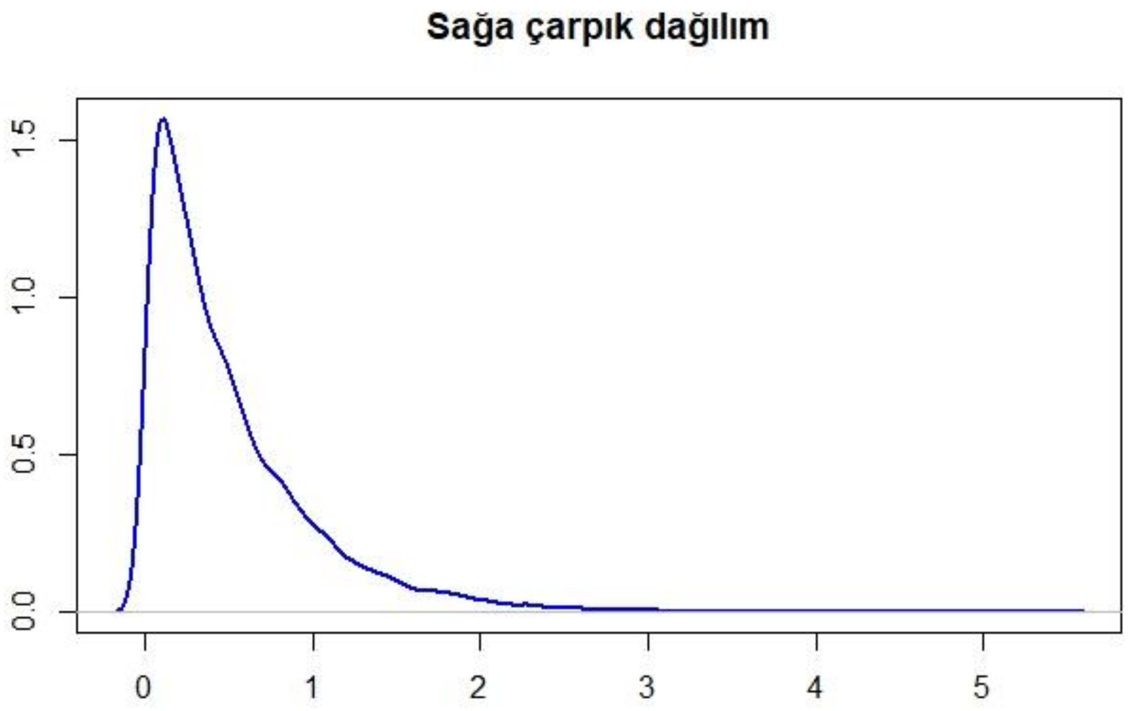
Çarpıklık (Skewness) Katsayısı

- Bir veri setinin en önemli özelliklerinden biri de verilerin hangi biçimde bir dağılıma sahip olduğunu belirlemektir. Verinin tek tepeli mi yoksa çok tepeli mi ve tek tepeli ise simetrik mi çarpık mı olduğu araştırılır.
- Eğer veri simetrik ise

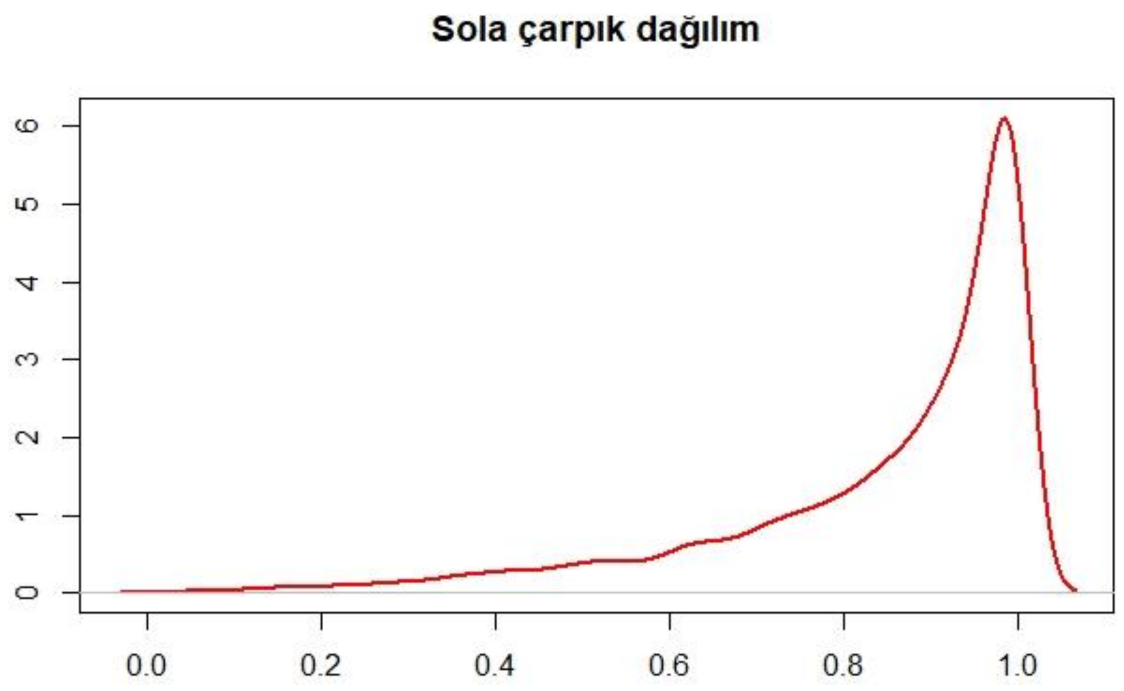
ortalama=mod=medyan olur.

- Eğer bir grafikte veriler sol tarafta

toplammışsa ve sağ kuyruk daha uzun ise dağılımın **sağa çarpık** olduğu söylenir.



- Eğer bir grafikte veriler sağ tarafta toplanmışsa ve sol kuyruk daha uzun ise dağılımın **sola çarpık** olduğu söylenir.



- Dağılımın çarpıklığı hakkında bilgi veren bir ölçü **çarpıklık** katsayısıdır. S standart sapma olmak üzere

çarpıklık katsayısı $\gamma_1 = \dfrac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{S^3}$ biçiminde tanımlanır.

Eğer $\gamma_1 > 0$ ise dağılım sağa çarpıktır.

Eğer $\gamma_1 < 0$ ise dağılım sola çarpıktır.

Eğer $\gamma_1 = 0$ ise dağılım simetriktir.

Basıklık (Kurtosis) katsayısı

- Bir verinin dağılım grafiğinin normal dağılıma göre daha sivri mi yoksa daha basık mı olduğunu bilgisini basıklık katsayısı verir.
- **Basıklık katsayısı** $\gamma_2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{s^4} - 3$ biçiminde tanımlanır.

- Eğer $\gamma_2 > 0$ ise dağılım normal dağılıma göre daha sivridir.
- Eğer $\gamma_2 < 0$ ise dağılım normal dağılıma göre daha basıktır.
- Eğer $\gamma_2 = 0$ ise dağılım normal dağılıma sahiptir.

- Pozitif basıklık durumunda (siyah) normal dağılım (çan eğrisi, kırmızı) grafiğe göre daha sivri bir grafik oluşur. Negatif basıklık durumunda (yeşil) ise normale göre daha basık bir grafik oluşur.