ASTROİSTATİSTİK
12. KONU

Hazırlayan: Doç. Dr. Tolgahan KILIÇOĞLU

12. KORELASYON

Bir deneyin veya gözlemin sonucunda elde edilen iki değişken arasında bir ilişki olup olmadığı ortaya konmak istenebilir. Aşağıda birbirleri arasında ilişki aranan durumlara ilişkin birkaç örnek verilmektedir;

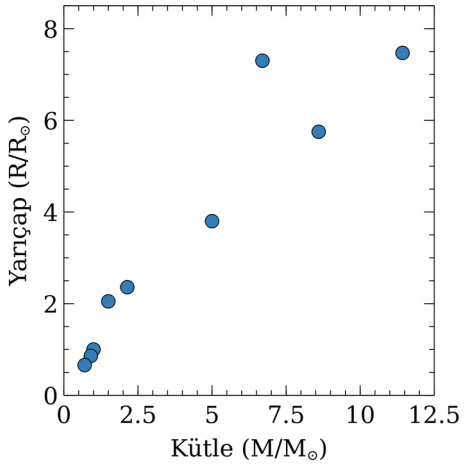
- Sigara kullanma miktarı ile yaşam süresi
- Öğrencilerin televizyon izleme miktarı ile üniversite sınavındaki başarısı
- Bir insanın IQ puanı ile kazandığı maaş
- Bilgisayar kullanma seviyesi ile ALES sınavında alınan not
- İnsanların kiloları ve boyları
- Bir lisede matematik sınavında alınan notlar ile Türkçe sınavından alınan notlar
- Anne-babanın ortalama IQ puanı ile çocuğun IQ puanı
- Belirli bir türden ağaçların yaşları ve boyları
- Otomobillerin ağırlıkları ile benzin tüketimi
- Yıldızların kütleleri ile ışınım güçleri
- Galaksilerin dikine hızları ile kırmızıya kaymaları
- Değişen yıldızların pulsasyon dönemleri ve parlaklıkları
  vb...

İki değişken arasındaki olması muhtemel ilişkiye **korelasyon** adı verilir. İki değişken arasında bir korelasyon olup olmadığının ortaya konulması için matematiksel bir yönteme ihtiyaç vardır. Bu bölümde korelasyona ilişkin ifadelerin nasıl hesaplanacağını ve yorumlanacağını göreceğiz.

12.1 Değişkenlerin Görsel Olarak İrdelenmesi

İki değişken arasında ilişki olup olmadığı irdelenirken öncelikle değişkenlerin bir grafiğe aktarılması faydalı olacaktır. Bunun için genellikle bir **saçılma grafiği** kullanılır. Aralarında korelasyon olup olmadığı belirlenecek iki değişkenden biri saçılma grafiğinin x eksenine diğeri ise saçılma grafiğinin y eksenine yerleştirilir.

Örneğin Şekil 12.1'de parlak 9 anakol yıldızının kütleleri ve yarıçapları bir saçılma grafiği üzerinde gösterilmektedir. Şimdi konuya devam etmeden bu saçılma grafiğini inceleyerek anakolda bulunan yıldızların kütleleri ile yarıçapları arasında bir ilişki olup olmadığını tartışınız.

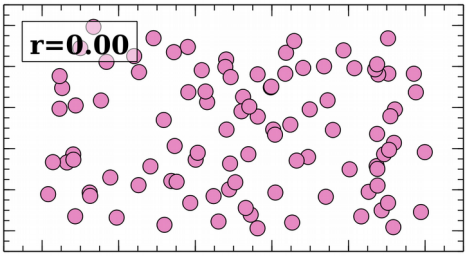


Şekil 12.1 Bazı anakol yıldızlarının kütleleri ile yarıçaplarının saçılma grafiği

12.2 Korelasyon Katsayısı (r) ve Anlamı

İki değişken arasında doğrusal bir ilişkinin olup olmadığı korelasyon katsayısı kullanılarak ortaya konabilmektedir. Korelasyon katsayısı (*r*) -1 ile 1 arasında değerler almaktadır. Bu katsayının nasıl hesaplandığına geçmeden önce aldığı değerlerin ne anlama geldiğini tartışalım.

***r* = 0 Durumu:** Eğer korelasyon katsayısı sıfıra eşitse (veya çok yakınsa) iki değişken arasında bir ilişki olmadığı eğer varsa da bu ilişkinin zayıf olması gerektiği söylenebilir. Ancak bu yorum yapılırken dikkatli olunmalıdır. Çünkü ele alınan verilerin sayısının yetersiz olması, yeterli aralığa dağılmamış olması ve/veya hatalarının yüksek olması da iki değişken arasında mevcut olan bir ilişkinin tespit edilmesinde engel teşkil ediyor olabilir. Şekil 12.2'de korelasyon katsayısı sıfır olan iki değişkenin saçılma grafiği gösterilmektedir.



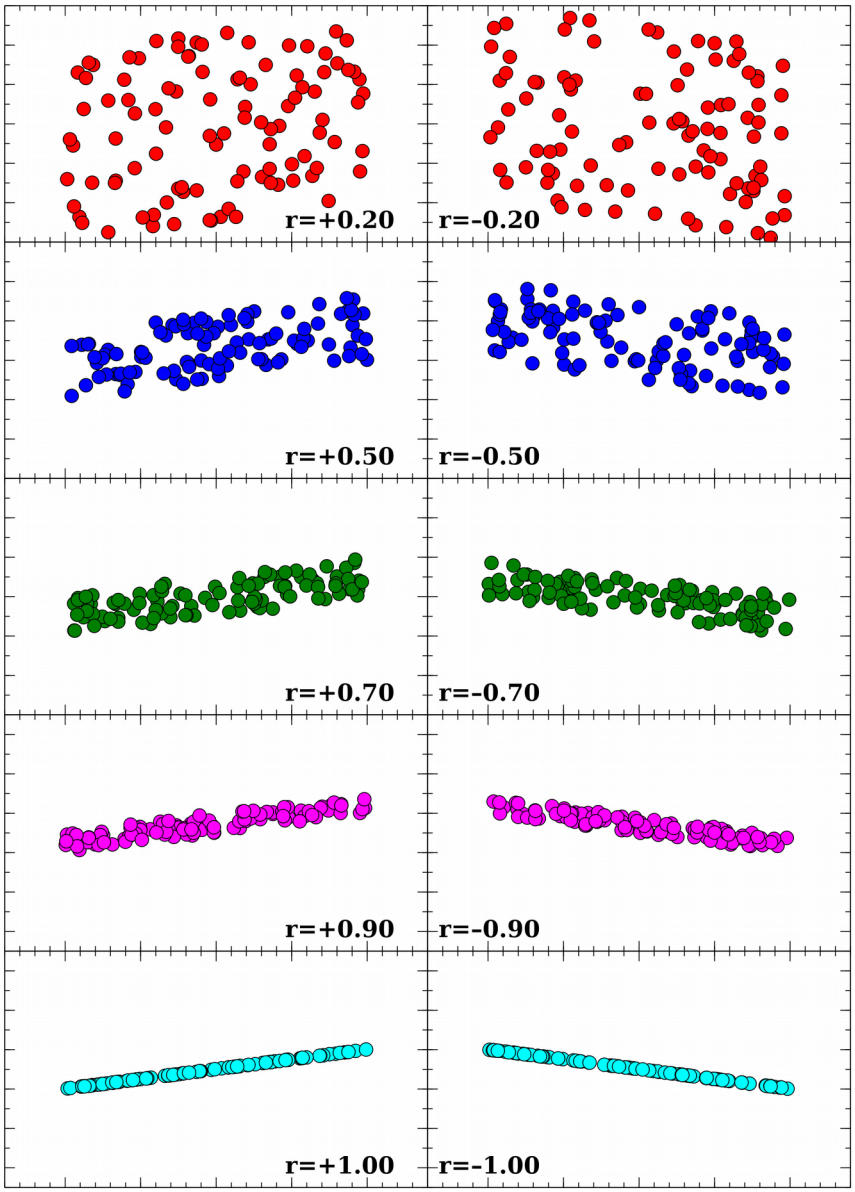
Şekil 12.2 Korelasyon katsayısı sıfır olduğu hesaplanan iki değişkenin saçılma grafiği

0 < *r* < 1 Durumu: Eğer korelasyon katsayısı 0 ile 1 arasında bir değere sahipse bu durum iki değişen arasında **pozitif** bir ilişki olduğunu işaret eder. Bir başka deyişle, değişkenlerden birinin değeri arttıkça diğeri de artmaya eğilimlidir. Bu ilişkinin ne kadar net şekilde ortaya konabildiği r'nin değerine bağlıdır. Eğer r'nin değeri 1'e yakınsa iki değişken arasındaki ilişki iyi bir şekilde ortaya konabilmektedir. Ancak r'nin değeri 0'a doğru yaklaştıkça iki değişken arasında yanlızca zayıf bir ilişkinin olabileceği veya eldeki verilerin (veya verilerin duyarlılığının) böyle bir ilşkiyi ortaya koymada yetersiz olduğu söylenebilir.

-1 < *r* < 0 Durumu: Eğer korelasyon katsayısı 0 ile –1 arasında bir değere sahipse bu durum iki değişen arasında **negatif** bir ilişki olduğunu işaret eder. Burada önceki durumun tersine, değişkenlerden birinin değeri arttıkça diğeri azalmaya eğilimlidir. Yine benzer şekilde r'nin değerinin 0'a yakın olması ilişkinin olmadığı veya mevcutsa da zayıf olduğu sonucunu işaret ederken –1'e yakın olması da ilişkinin varlığının daha belirgin olduğunu göstermektedir.

r=1 veya r=–1 Durumu: Korelasyon katsayısının tam 1'e veya –1'e eşit olması (veya bu değerlereçok yakın olması) iki değişkenin birbirlerine tamamen bağlı olduğunu gösterir. r=1 olması durumunda pozitif bir ilişki olduğu, r=–1 için ise negatif bir ilşkinin olduğu görülür. Ancak gerçek hayatta iki değişken arasında böyle bir ilşkinin olması çok olası değildir.

Şekil 12.3'de farklı korelasyon katsayılarına sahip olan değişken çiftlerinin saçılma grafikleri verilmektedir.

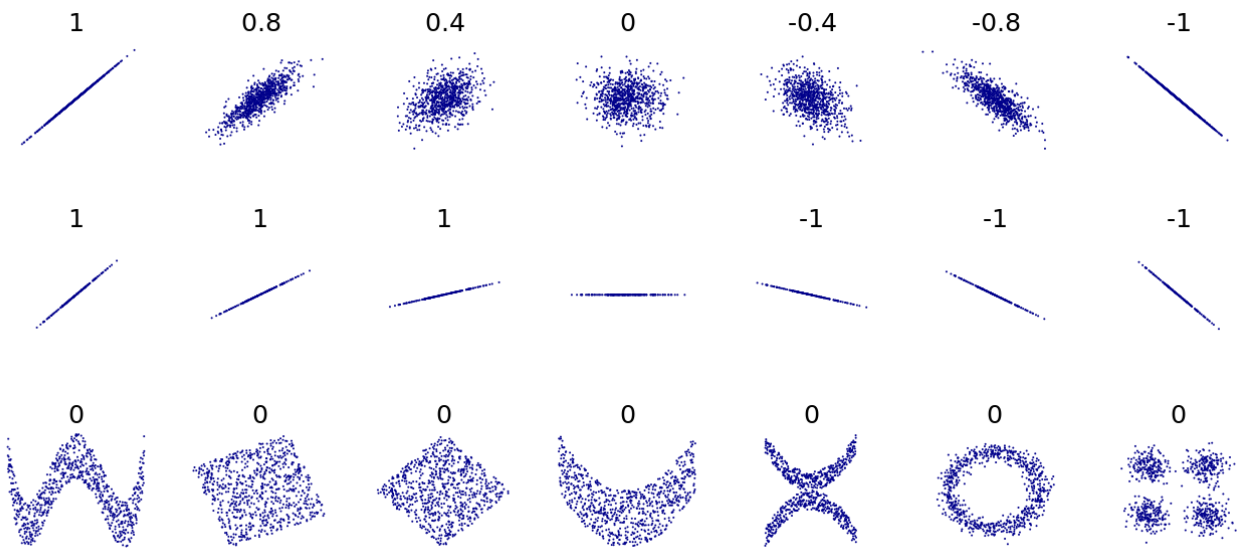


Şekil 12.3 Çeşitli değişkenlerin saçılma grafikleri ile hesaplanan *r* korelasyon katsayıları

12.3 Korelasyon Katsayısı ile Temsil Edilmeyen Durumlar

Korelasyon katsayısı birbiriyle ilişkili olan iki değişkenin şaçılma grafiğinde oluşan eğrinin eğiminden bağımsızdır. Şekil 12.4'ün orta satırında farklı eğime sahipolan saçılma grafikleri gözükmektedir. Sol üç saçılma grafiğinin korelasyon katsayısının r = 1 sağ üç grafiğin ise r = –1 olduğu görülmektedir. Görüldüğü gibi korelasyon katsayısı eğimden etkilenmemiştir. Bu gerçekte korelasyon katsayısının en güçlü özelliklerinden biridir. Çünkü ele aldıığımız değişkenlerin birbirlerin bağımlılıkları hep aynı eğime sahip olmak zorunda değildir. Korelasyon katsayısı bu anlamda sadece saçılma grafiğinin ne kadar **saçıldığını** temsil etmektedir.

Şekil 12.4'ün en alt satırına bakıldığında birbirlerine sıradışı şekilde bağımlı bazı değişkenlerin saçılma grafikleri verilmektedir. Bu değişkenler birbirlerine bağımlı olmalarına karşın korelasyon katsayıları hesaplandığında $r = 0$ olduğu görülür. Bu durum dağılımdaki ilişkinin eğrisel olmasından (soldan 1. ve 4. grafik), eş dağılımlı olmasından (soldan 2. ve 3. grafik) veya simetrik yapıda (5., 6. ve 7. grafik) olmasından kaynaklanmaktadır. Bu gibi verilerde başka tür korelasyon belirteçlerinin kullanılması gerekmektedir. Bu gibi sıradışı dağılımların olup olmadığının kontrol edilmesi için değişkenlerin saçılma grafiklerinin analizin başında çizdirilmesi oldukça önemlidir.

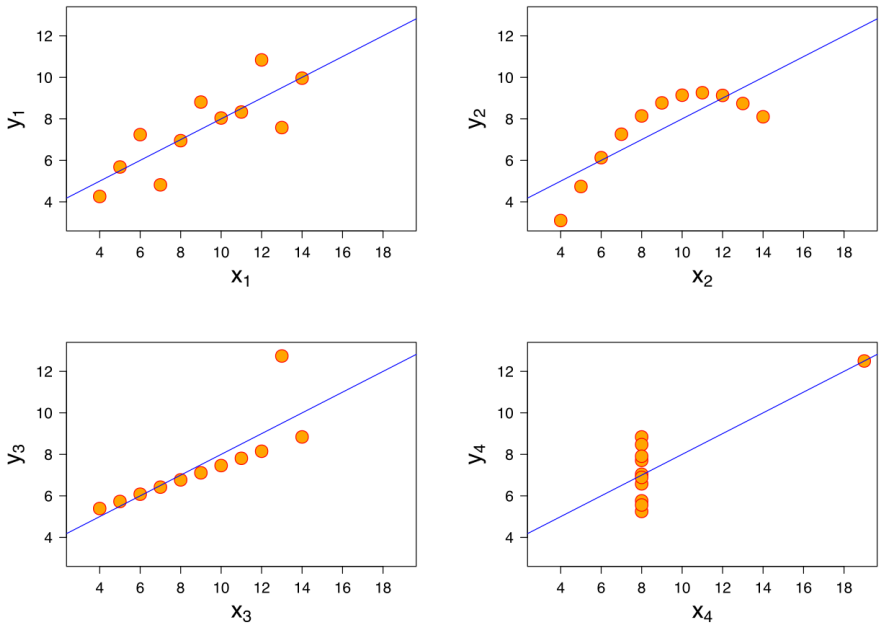


Şekil 12.4 Birbirleriyle farklı şekillerde ilişkili olan değişkenlerin saçılma grafikleri ve r korelasyon katsayıları
Kaynak: https://en.wikipedia.org/wiki/Correlation_and_dependence#/media/File:Correlation_examples2.svg

Son olarak korelasyon katsayısı bir yüzde gibi düşünülmemelidir. Örneğin, korelasyon değerinin %85 olması iki değişkenin birbiriyle %85 oranında uyumlu olduğu anlamına gelmez.

12.4 Dağılımların Farklı Korelasyonun Aynı Olduğu Durumlar

Şimdi Şekil 12.5'te verilen saçılma grafiklerini inceleyerek korelasyon katsayılarını tahmin etmeye çalışın.



Şekil 12.5 Dört farklı değişken setinin saçılma grafiği
Kaynak: https://en.wikipedia.org/wiki/Correlation_and_dependence#/media/File:Anscombe%27s_quartet_3.svg

Şekil 12.4'te verilen her grafik için farklı korelasyon katsayıları tahmin etmiş olabilirsiniz. Ancak bu dört grafiğin de korelasyon katsayısı r = 0.816 dır! Şimdi bu grafiklerde karşımıza çıkan durumları tek tek irdeleyelim;

i) Sol üstte yer alan grafikte verilerin homojen ve doğrusal olarak dağıldığı klasik bir saçılma grafiği gözükmektedir. Bu grafikteki değişkenlerin arasındaki ilşkinin ortaya konması için korelasyon katsayısı oldukça iyi bir ölçüttür.

ii) Sağ üstteki grafikte ise değişkenlerin birbirlerine eğrisel olarak bağımlı olduğu görülmektedir. Ancak, korelasyon katsayısı sadece doğrulsal korelasyonu tespit edebildiğinden bu eğrisel değişimin sadece doğrusal bileşenindeki uyumu ölçebilmiştir. İlişkinin eğrisel olduğu düşünüldüğünde gerçekte bu iki değişken arasında oldukça iyi bir korelasyon bulunmaktadır.

iii) Sol alttaki grafikte ise verilerin birbirleriyle sıkı ilişkili olduğu görülmekle birlikte sadece tek bir aykırı değer korelasyon katsayısının 1 yerine 0.816 çıkmasına neden olmaktadır. Astronomide bazen bu grafiğe benzer durumlar ile karşılaşılabilmektedir. Eğer veride tek bir değerin büyük bir sapma gösterdiği görülüyorsa korelasyon hakkında yorum yapmadan önce verinin gerçekten doğru olup olmadığının, bu farkılığı gözlemdeki bir hatadan kaynaklanıp kaynaklanmadığının tespit edilmesi gerekir.

iv) Sağ alttaki grafikte de bir grup verinin aynı x değişkeninde toplandığı görülmektedir. Bu grafiğe benzer durumlar da astronomide oldukça yaygındır. Örneğin bir hava kütlesi hesabında kullanılacak gözlemler günün belirli saatlerinde yapıldıysa verilerin belirli hava kütlesi değerlerinde yoğunlaştığı görülür. Korelasyon katsayısı burada yine veriler ışığında uyumun ne kadar olup olmadığını olabilecek en iyi ihtimalle vermektedir. Ancak ilşkinin olup olmadığının daha net şekilde ortaya konması daha geniş bir "x" aralığında veri alınmasını gerektirir.

12.5 Korelasyon Katsayısının Matematiksel İfadesi

Korelasyon katsayısının ne anlama geldiğini ve nasıl yorumlanması gerektiğini açıkladığımıza göre artık nasıl hesaplandığından söz edebiliriz. Bu hesaplama için öncelikle Fark Kareler Toplamı (KT) ve Fark Çarpımlar Toplamı $(ÇT)$ ifadelerini tanımlayalım. x ve y bir veride korelasyonu incelenecek değişkenler olmak üzere;

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

$$KT_y = \sum_{n=1}^{n} (y_i - \bar{y})^2$$

$$ÇT_{xy} = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Bir veri için yukarıda verilen ifadeler hesaplandığında iki değişken arasındaki korelasyon katsayısı aşağıdaki gibi hesaplanır;

$$r = \frac{ÇT_{xy}}{\sqrt{KT_x KT_y}}$$

Bu korelasyon katsayısı Pearson Korelasyon Katsayısı olarak da isimlendirilir.

12.6 Bir Örnek Veri İçin Korelasyon Katsayısının Belirlenmesi

Şimdi bu konunun en başında verdiğimiz anakol yıldızlarının kütleleri ile yarıçapları değişkenleri için korelasyon katsayısını hesaplayalım. Çizelge 12.1'de dokuz parlak anakol yıldızının kütle ve yarıçap değerleri sunulmaktadır.

Çizelge 12.1 Seçilmiş bazı parlak anakol yıldızlarının kütleleri ve yarıçapları

Yıldız Adı	Kütle ($M/M_\odot$)	Yarıçap ($R/R_\odot$)
Spica	11.43	7.47
Bellatrix	8.60	5.75
Achernar	6.70	7.30
2 Lac	5.00	3.78
Vega	2.14	2.36
Procyon	1.50	2.05
Güneş	1.00	1.00
Alpha Centauri B	0.91	0.86
61 Cyg A	0.70	0.66

Öncelikle ortalamaları hesaplayalım;

$$\bar{x}=\frac{\sum_{i=0}^{n} x_i}{n}=\frac{11.43+8.60+6.70+5.00+2.14+1.50+1.00+0.91+0.70}{9}=4.22$$

$$\bar{y}=\frac{\sum_{i=0}^{n} y_i}{n}=\frac{7.47+5.75+7.30+3.78+2.36+2.05+1.00+0.86+0.66}{9}=3.47$$

Şimdi fark kareler toplamlarını ve fark çarpım toplamını hesaplayalım;

$$KT_x=\sum_{n=1}^{n}(x_i-\bar{x})^2=(11.43-4.22)^2+...+(2.14-4.22)^2+...+(0.70+4.22)^2=123.367$$

$$KT_x=\sum_{n=1}^{n}(x_i-\bar{x})^2=(7.47-3.47)^2+...+(2.36-3.47)^2+...+(0.66+3.47)^2=60.021$$

$$ÇT_{xy}=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=(11.43-4.22)(7.47-3.47)+...+(0.70-4.22)(0.66-3.47)=81.2215$$

Artık korelasyon katsayısını hesaplamaya hazırız:

$$r=\frac{ÇT_{xy}}{\sqrt{KT_x KT_y}}=\frac{81.2215}{\sqrt{123.367\cdot 60.021}}=0.94$$

olarak elde edilir. Buradan şu yoruma gidilebilir: *İncelediğimiz anakol yıldızlarının kütleleri ile yarıçapları arasında belirgin bir pozitif ilişki olduğu saptanmıştır. Anakoldaki yıldızların kütleleri arttıkça yarıçapları da artma eğilimindedir.*

Ancak burada sadece 9 yıldızdan bu sonuca vardığımızı ve bu unutmayın. Bu genellemeyi doğrulayabilmek için gerçekte galaksinin farklı bölgelerinde gözlenmiş çok daha fazla yıldızın gözlemine ihtiyaç vardır.