



HACETTEPE
ÜNİVERSİTESİ
İSTATİSTİK BÖLÜMÜ

İST156 İSTATİSTİĞE GİRİŞ II

DERS 8- KORELASYON KATSAYILARI

Ders sorumluları: Prof.Dr.Serpil AKTAŞ ALTUNAY (01 Şubesi)
Doç.Dr. Ayten YİĞİTER (02 Şubesi)

İlişki (Korelasyon-Correlation) Kavramı

Ölçülebilen nitelikler arasındaki ilişkiler her zaman araştırılır.

Örneğin:

- Sigara ile akciğer kanseri arasında ilişki var mıdır?
- COVID 19 ile obezite arasında ilişki var mıdır?
- COVID 19 ile hava kirliliği arasında ilişki var mıdır?
- Tüketim ile gelir arasında ilişki var mıdır?
- Faiz ile yatırım arasında ilişki var mıdır?
- Çalışma süresi ile başarı arasında ilişki var mıdır?

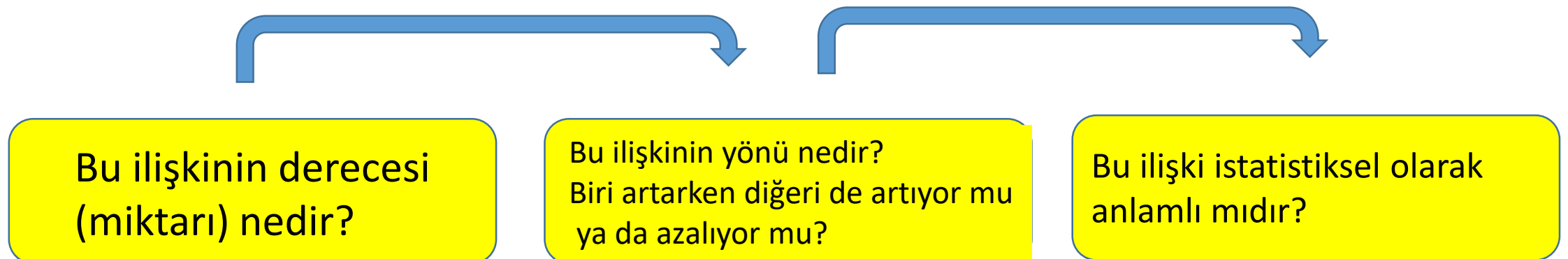
2. İlişki (Korelasyon) Katsayıları

Bu nedenle ilişki katsayıları hesaplanarak bu ilişkinin derecesi, yönü ve istatistiksel olarak anlamlı olup olmadığı incelenebilir.

Değişkenlerin türüne göre (aralıklı, oransal, nominal, ordinal ...) farklı ilişki katsayısı kullanılır.

Normal dağılıma sahip ve aralıklı ya da oransal ölçekte elde edilmiş iki değişken arasındaki ilişkinin derecesini ve yönünü gösterir.

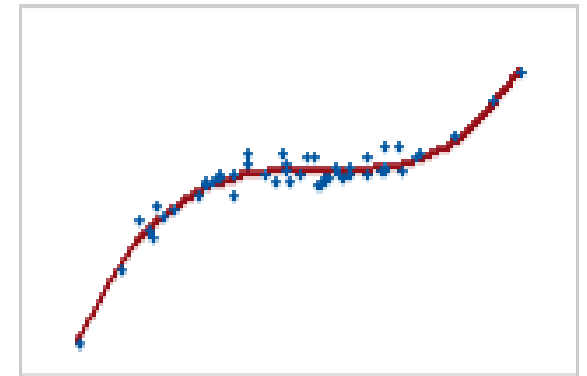
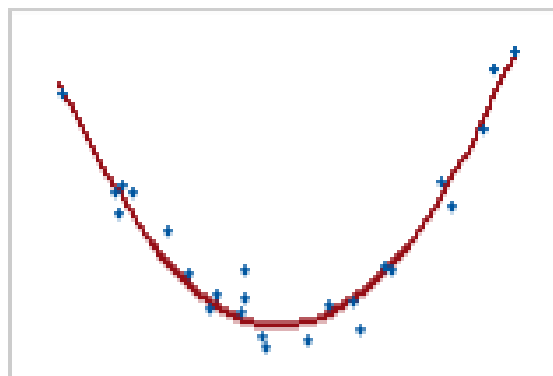
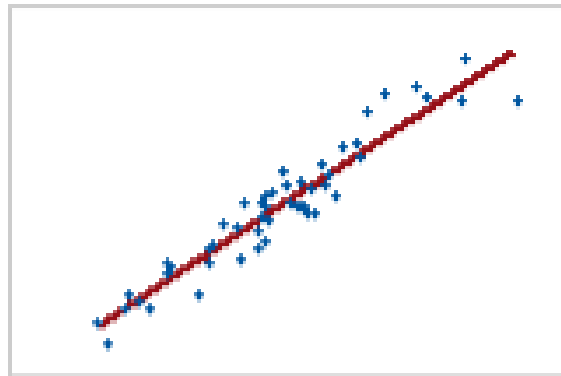
İlişki hesaplanacak gözlemler birbirini ile bağımsız olmalıdır.

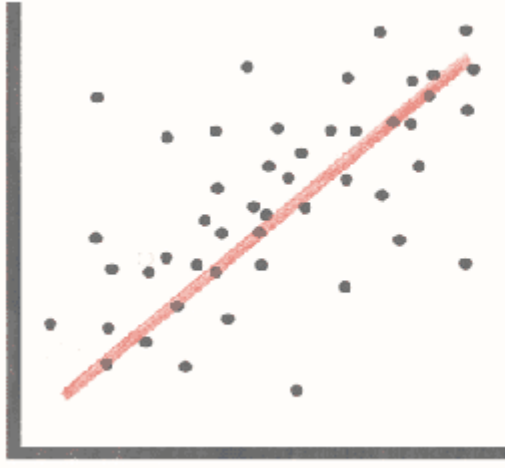


Doğrusal ilişki : İki değişken arasındaki ilişki doğrusaldır. İki değişken arasındaki ilişki bir doğru ile ifade edilir.

Doğrusal olmayan ilişki: İki değişken sabit bir oranda değişiklik göstermez. Grafik düz bir doğru değil parabol ya da hiperbol ile gösterilir.

Monoton ilişki: İki değişken de aynı yönde değişme eğilimindedir ancak sabit bir oranda değişmeyebilirler .

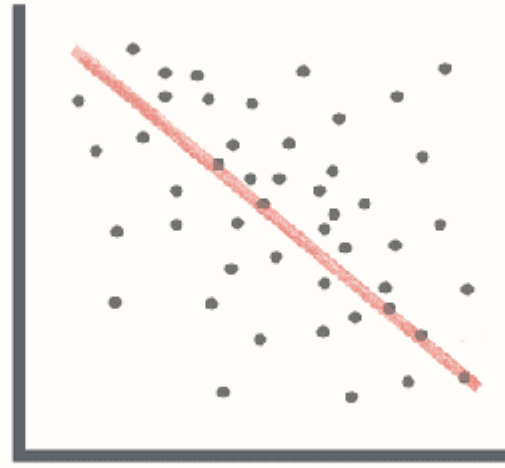




Pozitif ilişki



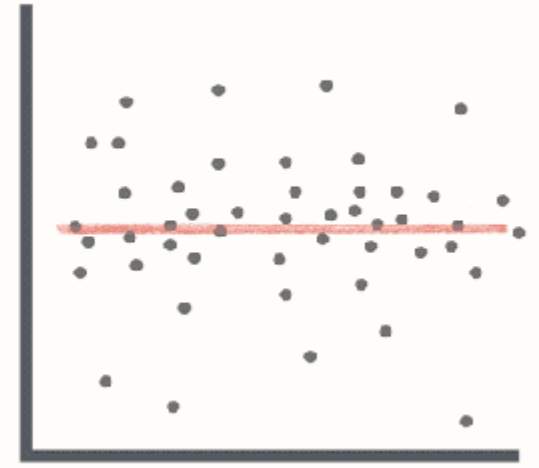
Çalışma süresi - başarı puanı



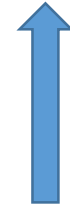
Negatif ilişki



Hava sıcaklığı – yakıt tüketimi



İlişki yok



faiz – hava sıcaklığı

2.1. Pearson'ın Çarpım Moment Korelasyon Katsayısı (Pearson's Product Moment Correlation Coefficient)

Normal dağılıma sahip X ve Y değişkeni arasındaki ilişkinin miktarını veren bir ilişki katsayısıdır. Örneklem korelasyon katsayısı r ile, kitle korelasyon katsayısı ρ ile gösterilir.

ζT_{XY} : X ve Y arasındaki çarpımlar toplamı

KT_X : X'e ait kareler toplamı, KT_Y : Y'ye ait kareler toplamı

$Cov(X,Y)$: X ve Y değişkenleri arasındaki kovaryans

S_X : X'a ait standart sapma,

S_Y : Y'ye ait standart sapma,

n: denek sayısı

olmak üzere,

$$\zeta T_{XY} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n} = \frac{\sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n}}{n} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{n}$$

$$KT_X = \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n} \quad KT_Y = \sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n}$$

Korelasyon Katsayısı



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

$$r = \frac{\zeta T_{XY}}{\sqrt{KT_X KT_Y}}$$

$$r = \frac{\text{Cov}(X, Y)}{S_X S_Y}$$

$$r = \frac{\sum_{i=1}^n x_i y_i - \frac{\sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n}}{\sqrt{\left(\sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}\right) \left(\sum_{i=1}^n y_i^2 - \frac{(\sum_{i=1}^n y_i)^2}{n}\right)}}$$

$$-1 \leq r \leq 1$$

r, +1'e yakın bir değer alıyorsa X ve Y arasında doğrusal artan bir ilişki vardır.

r, -1'e yakın bir değer alıyorsa X ve Y arasında doğrusal ters yönlü bir ilişki vardır.

r, 0'a yakın bir değer alıyorsa X ve Y ilişkisi yoktur.

2.2. Spearman'ın Rank Korelasyon Katsayısı

Değişkenlerin normal dağılıma sahip olmadığı durumlarda, verilerin kendi değerleri değil de rankları kullanılarak Spearman Rank korelasyon katsayısı tercih edilir.

$$r_s = 1 - \frac{6 \sum_{j=1}^n d_j^2}{n(n^2 - 1)}$$

$$d_j = \text{rank}(x_j) - \text{rank}(y_j)$$

$$-1 \leq r_s \leq 1$$

Yorumu Pearson'ın ilişki katsayısı ile aynıdır.

2.3. Kendall's Tau İlişki Katsayısı

$$\text{Kendall's Tau} = \frac{C - D}{C + D} = \frac{(\text{uyumlu gözlem çifti sayısı}) - (\text{uyumsuz gözlem çifti sayısı})}{\binom{n}{2}}$$

(x_i, x_j) ve (y_i, y_j) ($i < j$) çiftleri için

C: uyumlu çiftler (concordant):

Eğer $x_i < x_j$ ve $y_i < y_j$ ya da $x_i > x_j$ ve $y_i > y_j$ ise bu gözlem çifti uyumludur denir.

D: uyumsuz çiftler (discordant):

Eğer $x_i < x_j$ ve $y_i > y_j$ ya da $x_i > x_j$ ve $y_i < y_j$ ise bu gözlem çifti uyumsuzdur denir.

X	Y
x_1	y_1
x_2	y_2
\vdots	\vdots
x_n	y_n

Eğer $x_i = x_j$ ya da $y_i = y_j$ ise bu gözlem çifti ne uyumlu ne de uyumsuzdur denir.

Gözlemlerin kendisini değil rankları üzerinden ilişki katsayısını hesaplar.

İlk değişken doğal sıralamasında yazılır. C ve D sütunları ikinci değişkene göre belirlenir. İkinci değişkende i. gözleme bakılır i. gözlemden daha **büyük** olan gözlemler sayılarak C sütununa yazılır.

D için ise i. gözleme bakılır i. gözlemden daha **küçük** olanlar sayılarak D sütununa yazılır.

İki değişken arasında hesaplanan korelasyon (r) değeri:

$r \sim 0$ ise ilişki yoktur,

$0,10 \leq r < 0,20$ çok zayıf ilişki,

$0,20 \leq r < 0,40$ zayıf ilişki,

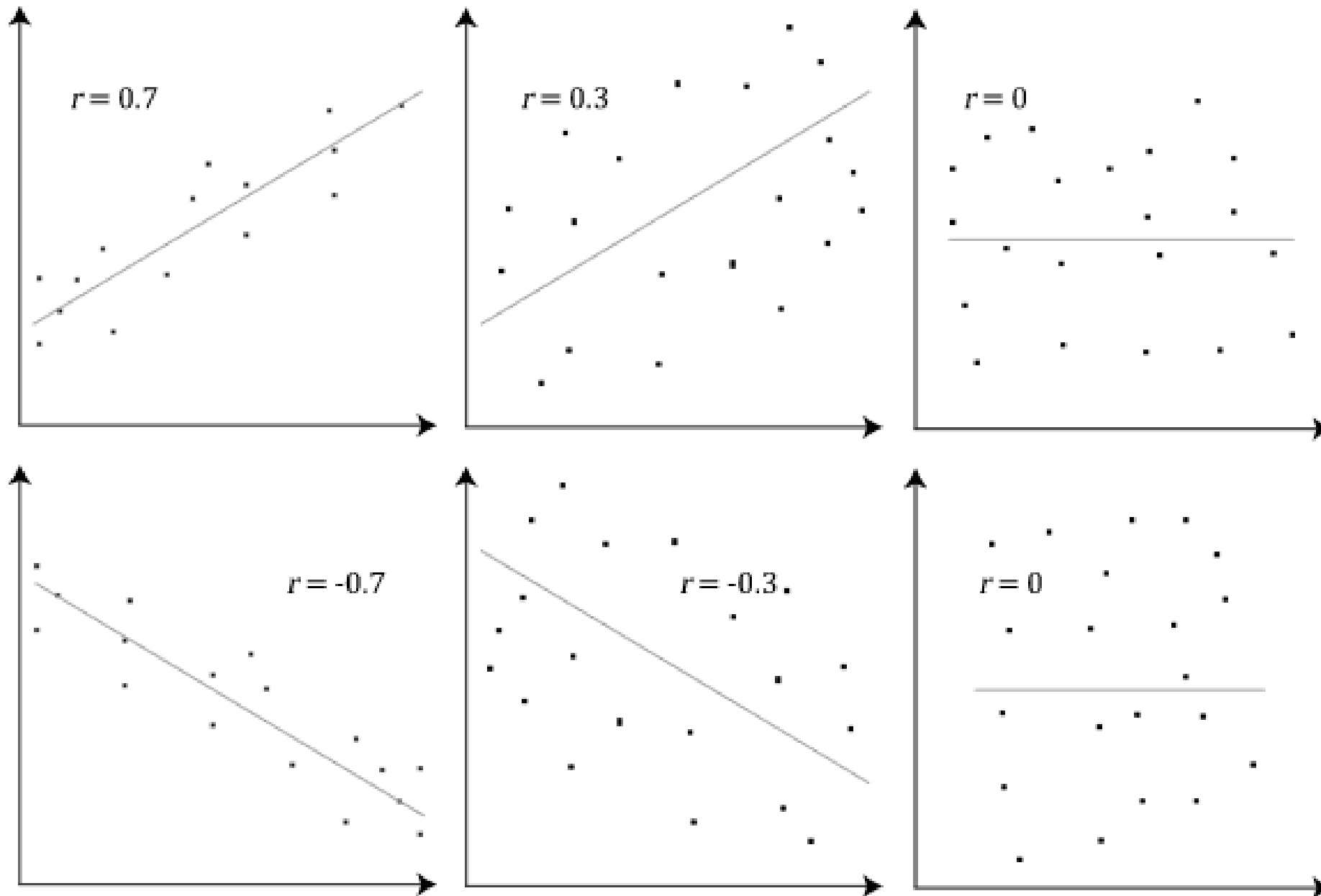
$0,40 \leq r < 0,60$ orta düzeyde ilişki

$0,60 \leq r < 0,80$ yüksek

$0,80 \leq r < 0,90$ çok yüksek

$0,90 \leq r < 1$ mükemmel

$r \sim 1$ mükemmel ilişki.



3. İlişki katsayısının önem kontrolü

Pearson ve Spearman Korelasyon Katsayıları İçin

Hesaplanan ilişki katsayısının istatistiksel olarak anlamlı olup olmadığı test edilir.

$$H_0: \rho=0$$

$$H_S: \rho \neq 0$$

hipotezi, $t = \frac{r}{SH(r)}$ ile test edilir, $SH(r) = \sqrt{\frac{1-r^2}{n-2}}$ olmak üzere

$$t = \frac{r \sqrt{n-2}}{\sqrt{1-r^2}}$$

$|t| \geq t_{(\alpha/2, n-2)}$ ise H_0 hipotezi reddedilir. Yani korelasyon katsayısı istatistiksel olarak anlamlıdır.

İstatistiksel yazılımlarda genellikle test istatistiğine ilişkin p değeri verilir.

Eğer $p \leq \alpha$ ise H_0 reddedilir. Yani korelasyon katsayısı istatistiksel olarak anlamlıdır denir.

$\rho \neq 0$ hipotezi için Fisher'in Z dönüşümü ile elde edilen Z istatistiği kullanılır.

Kendall Korelasyon Katsayısı İçin

$$H_0: \rho=0$$

$$H_S: \rho \neq 0$$

hipotezi, $z = \frac{\tau}{SH(\tau)}$ ile test edilir, $SH(\tau) = \sqrt{\frac{2(2n+5)}{9n(n-1)}}$ olmak üzere

$$z = \frac{3\tau\sqrt{n(n-1)}}{\sqrt{2(2n+5)}}$$

$|z| \geq z_{\alpha/2}$ ise H_0 reddedilir. Yani korelasyon katsayısı istatistiksel olarak anlamlıdır.

İstatistiksel yazılımlarda genellikle test istatistiğine ilişkin p değeri verilir.

Eğer $p \leq \alpha$ ise H_0 reddedilir. Yani korelasyon katsayısı istatistiksel olarak anlamlıdır denir.



NOT: Ranklar üzerinden hesaplanan ilişki katsayılarında eğer değişkenlerin sıralaması değil de ham verileri verilmişse, o zaman ilk değişken doğal sıralamasına göre sıraya koyularak, 1,2 3,4... gibi , diğer değişken de ilk değişkene göre sıralanır.

ÖRNEK 1: Bir meyve ağacının yaşı (Y-yıl) ile verdiği meyve verimi (X-kg.) arasındaki ilişki araştırılmak isteniyor.

- X ile Y arasındaki ilişki katsayısını hesaplayınız.
- Bulduğunuz ilişki katsayısının önem kontrolünü yapınız ($\alpha=0,05$).

Verim	Yaş
5	22
6	19
7	16
9	12
10	9
12	8

$$\text{ÇT}_{XY} = -72,33$$

$$\text{KT}_X = 34,83$$

$$\text{KT}_Y = 157,33$$

Ağacın yaşı ile meyve verimi arasında %97,7'lik negatif yönlü bir ilişki vardır. Yani ağacın yaşı arttıkça verim azalmaktadır.

$$r = \frac{-72,33}{\sqrt{34,83 \times 157,33}} = -0,977$$

Bulunan ilişki katsayısının α düzeyinde önem kontrolünü yapmak için hipotez testi yapılır.

$$H_0: \rho=0$$

$$H_S: \rho \neq 0$$

$$t = \frac{r \sqrt{n-2}}{\sqrt{1-r^2}} = \frac{-0,977 \sqrt{6-2}}{\sqrt{1-(-0,977)^2}} = -9,165$$

$$t_{(0,05/2,4)} = 2,776 \quad (p=0,001 < \alpha=0,05 \text{ } H_0 \text{ ret})$$

$|t| \geq t_{\text{tablo}}$ olduğu için H_0 reddedilir



Ağacın yaşı ile meyve verimi arasında %97,7'lik negatif yönlü ve istatistiksel olarak anlamlı bir ilişki olduğu %5 anlamlılık düzeyinde söylenebilir.

ÖRNEK 2: Bir iş sınavına giren 12 adayın, ilk mülakat ve ikinci mülakat sonucundaki başarı sırasına göre derecelendirmeleri verilmiştir. İki mülakat sıralaması arasındaki ilişkiyi hesaplayınız.

Aday	İlk mülakat	İkinci mülakat
A	1	1
B	2	2
C	3	4
D	4	3
E	5	6
F	6	5
G	7	8
H	8	7
I	9	10
J	10	9
K	11	12
L	12	11

Öğrenci	İlk mülakat	İkinci mülakat	C	D
A	1	1	11	0
B	2	2	10	0
C	3	4	8	1
D	4	3	8	0
E	5	6	6	1
F	6	5	6	0
G	7	8	4	1
H	8	7	4	0
I	9	10	2	1
J	10	9	2	0
K	11	12	0	1
L	12	11	-	-
Toplam			61	5

1'den **büyük**
kaç gözlem
var

1'den **küçük**
kaç gözlem
var

$$\begin{aligned}
 \text{Kendall's Tau} &= \frac{61 - 5}{61 + 5} \\
 &= \frac{56}{66} \\
 &= 0,848
 \end{aligned}$$

Adayların iki mülakat
Sonuçları arasında %85'lik
bir ilişki vardır.

ÖRNEK 3: 10 çocuk bir okulda yapılan resim ve spor seçmelerine katılarak bu aktivitelerden puan almışlardır. Verilerin normal dağılıma uymadığı varsayımı ile Spearman'ın korelasyon katsayısını hesaplayınız.

Resim	Spor
55	39
66	61
69	65
72	64
77	75
81	70
88	45
90	55
91	41
93	50



Her iki değişken kendi içinde küçükten büyüğe doğru sıralanır.

Resim	Spor	d_j	d_j^2
1	1	0	0
2	6	-4	16
3	8	-5	25
4	7	-3	9
5	10	-5	25
6	9	-3	9
7	3	4	16
8	5	3	9
9	2	7	49
10	4	6	36

$$r_s = 1 - \frac{6 \sum_{j=1}^n d_j^2}{n(n^2-1)}$$

$$= 1 - \frac{6 \times 194}{10(10^2-1)}$$

$$= 1 - \frac{1164}{990}$$

$$r_s = -0,1757$$

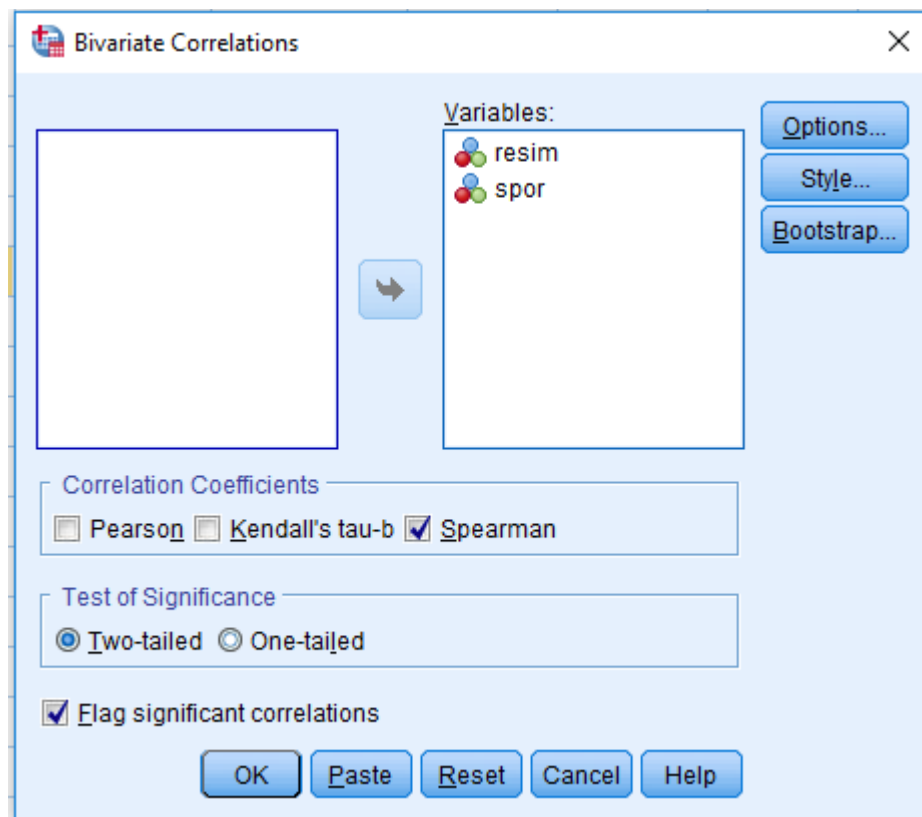
Öğrencilerin resim ve spor puanları arasında %17,57'lik negatif yönlü düşük bir ilişki vardır. $p=0,627$ olduğundan bu ilişki katsayısı istatistiksel olarak da anlamlı bulunmamıştır (sayfa 23'deki SPSS çıktısına bakınız).

SPSS UYGULAMASI (Örnek 3)

resim	spor
1,00	1,00
2,00	6,00
3,00	8,00
4,00	7,00
5,00	10,00
6,00	9,00
7,00	3,00
8,00	5,00
9,00	2,00
10,00	4,00

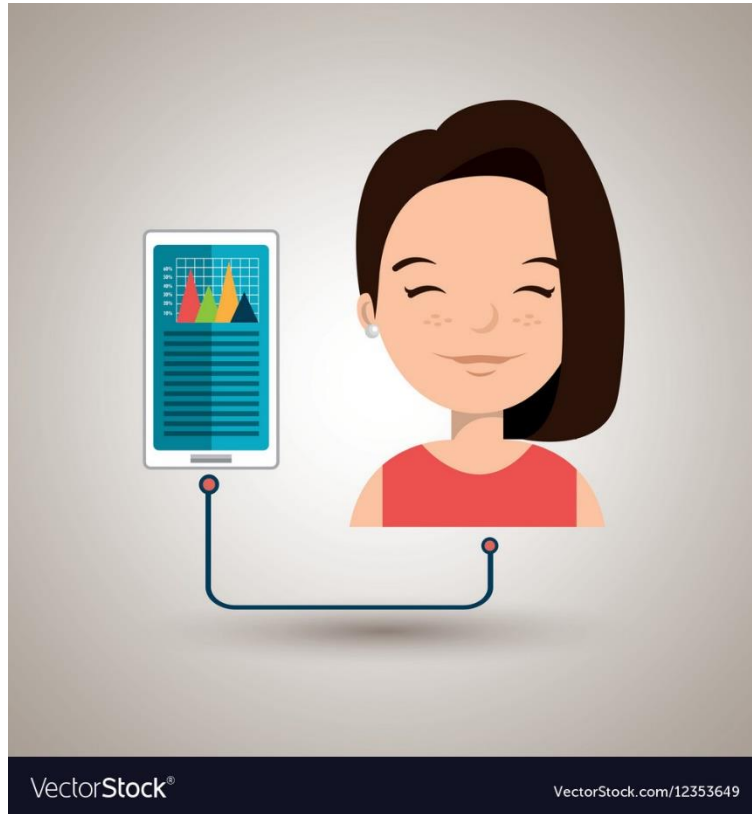
Correlations

			resim	spor
Spearman's rho	resim	Correlation Coefficient	1,000	-,176
		Sig. (2-tailed)	.	,627
		N	10	10
	spor	Correlation Coefficient	-,176	1,000
		Sig. (2-tailed)	,627	.
		N	10	10



ÖDEV: Bir iş yerinde çalışan 8 kişinin haftalık ücretleri ve iş memnuniyetlerine üzerine bir araştırma yapılmıştır. Bu iki değişken arasındaki ilişki katsayısını hesaplayarak %10 anlamlılık düzeyinde önem kontrolünü yapınız.

Haftalık Ücret (\$)	İş memnuniyeti
1537	7,6
1610	7,7
1322	6,6
910	6,3
1200	6,1
400	5,0
1100	6,1
890	4,2



Bir sonraki derste «Regresyon Çözümlemesi» incelenecek.



KAYNAKLAR

H. Demirhan, C. Hamurkaroğlu, “İstatistiksel Yöntemlere Giriş”, H.Ü. Yayınları, 2011.

S. Erbaş, «Olasılık ve İstatistik», Gazi Kitapevi, 2019.

Ö. Toktamış, S. Türkan, «R Programı ile Temel İstatistiksel Yöntemler», Seçkin Yayınevi, 2017.