X. BÖLÜM

KORELASYON (İLİŞKİ KATSAYISI)

X.1 Giriş

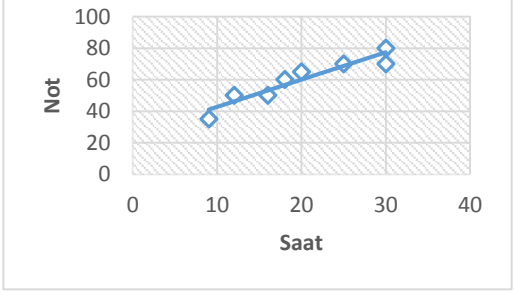
Oranlama ya da eşit aralıklı ölçme düzeyindeki deişkenler arası ilişki derecesinin belirlenmesine çoğu kez ihtiyaç duyulur. Örneğin; sınava hazırlanma sürei ile sınavda alınan not arasında bir ilişki var mıdır? Eklam harcaması ile satış miktarı arasında bir ilişki var mıdır? Bunun gibi örnekler çoğaltılabilir. İşte böyle amaçlarla sıkça karşılaşılır. Bu türden amaçlarda iki değişken arasındaki birlikte değişimin ölçüsü olan kovaryans, $Cov(X,Y) = E[(X-\bar{X})(Y-\bar{Y})]$ eşitliği ile tanımlanır ve $cov(x,y) = \frac{\sum[(x-\bar{x})(y-\bar{y})]}{N-1}$ kitle için ve $cov(x,y) = \frac{\sum[(x-\bar{x})(y-\bar{y})]}{n-1}$ örnek için eşitlikleri ile hesaplanır. Burada $Cov(X,Y)$, X ile Y değişkenleri arasındaki birlikte geğişimi gösteren tahmin edici kovaryanstır. Değişkenler arasındaki ilşşki derecesinin belirlenmesinde kovaryans çok sık kullanılmaz. Bunun nedeni ise kovaryansın değişkenlerin ölçü birimine bağlı olmasıdır. Kovaryansın ölçü birimine bağlılığını ortadan kaldırmak için korelasyon (ilişki katsayısı) kullanılır. Böylece X ve Y gibi iki değişken arasındaki korelasyon katsayısı, ρ_{XY} ile gösterilir ve $\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{V(X)V(Y)}}$ eşitliği ile tanımlanır. Bu nicelik kitle için $\rho_{XY} = \frac{\sum[(x-\bar{x})(y-\bar{y})]}{\left(\sqrt{\sum(x-\bar{2})^2}\right)\left(\sqrt{\sum(y-\bar{y})^2}\right)}$ ile ve örneklem için ise r_{XY} $r_{XY} = \frac{\sum[(x-\bar{x})(y-\bar{y})]}{\left(\sqrt{\sum(x-\bar{2})^2}\right)\left(\sqrt{\sum(y-\bar{y})^2}\right)}$ ile hesaplanır. Ayrıca korelasyon $r_{xy} = \frac{\sum xy - n\bar{x}\bar{y}}{\left(\sqrt{\sum x^2 - n(\bar{x})^2}\right)\left(\sqrt{\sum y^2 - n(\bar{y})^2}\right)}$ eşitliği yarmımı ile hesaplanabilir. Bu ilişki katsayısına Pearson korelasyon katsayısı adı da verilir. $-1 \leq \rho_{XY} \leq 1$ arasında tanımlıdır. Eğer $\rho_{XY} > 0$ ise değişkenler arasında aynı yönde, $\rho_{XY} < 0$ ise ters yönde bir ilişki olduğu, $\rho_{XY} = 0$ ilişki olmadığı ve $\rho_{XY} = \pm 1$ (ters ya da aynı yönde) tam ilişkiden söz edilir.

♦ **Örnek:** 50 öğrencinin bulunduğu bir sınıftan tesadüfen sekiz öğrenci seçilerek sınava hazırlanma süresi, X ve sınavda alınan puan Y takipteki tabloda verildiği gibi tespit edilmiş olsun.

Çalışma Süresi (x_i, saat)	30	30	25	20	18	16	12	9
Alınan not (y_i, 100)	80	70	70	65	60	50	50	35

Veriye ilişkin; a) Serpme grafiği oluşturunuz, b) Pearson ilşki katsayısını hesaplayınız ve sonucu yorumlayınız.

♦ **Çözüm:** a) Veriye ilişkin serpme grafiği, takipteki şekilde verildiği gibi sergilenir.



Şekil: Derse devam süresi-Alınan not serpme grafiği

Serpme grafiğinden de görüldüğü gibi derse çalışma süresi attıkça alınan not da artmaktadır. Bu ise değişkenler arasında aynı yönde bir ilişki olduğunu göstermektedir. b) Bazı hesaplamalar: $\sum xy = 10345,\ \bar{x} = \frac{160}{8} = 20,\ \bar{y} = \frac{480}{8} = 60,\ \sum x^2 = 3630,\ \sum y^2 = 30250$ olarak hesaplanır. Buradan $r_{xy} = \frac{\sum xy - n\bar{x}\bar{y}}{\left(\sqrt{\sum x^2 - n(\bar{x})^2}\right)\left(\sqrt{\sum y^2 - n(\bar{y})^2}\right)} = \frac{10345 - 8(20)60}{\sqrt{[3630 - 8(20)^2][30250 - 8(60)^2]}} = \frac{745}{\sqrt{430(1450)}} \cong \frac{745}{789.6202} \cong 0.9435$ bulunur. **Yorum:** Sınava hazırlanma süresi ile alınan not arasında aynı yönde oldukça güçlü bir ilişki olduğu söylenebilir.

X.2 Korelâsyon İçin Güven Aralığı

Kitle için X ile Y değişkenleri arasındaki ilişki katsayısı ρ_{XY} ile gösterilir. ρ_{XY} parametresinin istatistiği ise örneklemden hesaplanan r_{XY} istatistiğidir. Bu istatistiğin (tahmin edicinin) beklenen değeri ya da ortalaması $E(r_{XY}) = \rho_{XY}$ ve varyansının $V(r_{XY}) = S_r^2 = \frac{1-r_{XY}^2}{n-2}$ olduğu ve dağılımının ise $r_{XY} \sim N(\rho_{XY};\ S_r^2)$ olduğu gösterilmiştir. Böylece kitle korelasyon katsayısı için $100(1-\alpha)\%$ güve aralığı tahmin edicisi, $P\{r_{XY} - T_T S_r \le \rho_{XY} \le r_{XY} + T_T S_r\} = 1 - \alpha$ ve buna uygun aralık tahmini ise, $P\{r_{xy} - t_T s_r \le \rho_{xy} \le r_{xy} + t_T s_r\} = 1 - \alpha$ olasılık beyanı ile ifade edilir. Burada $t_T = t_{(n-2);\frac{\alpha}{2}}$ ($(n-2)$ s.d. ve $\frac{\alpha}{2}$ önem seviyesinde) t dağılımının tablo değeridir. $s_r = \sqrt{s_r^2}$ eşitliği ile hesaplanır.

♦ **Örnek:** Bir önceki örnekteki 50 birimlik öğrenci kitlesine ilişkin korelasyon katsayısı için % 95 güven aralığı tahmin ediniz ve sonucu yorumlayınız.

♦ **Çözüm:** Örnekte $r_{xy} = 0.9435$ bulunmuştu. Buradan $s_r = \sqrt{\frac{1-r_{xy}^2}{n-2}} = \sqrt{\frac{1-(0.9435)^2}{8-2}} \cong 0.1353$ bulunur. $t_T = t_{6;\,0.025} = 2.969$ (t tablosundan) bulunur. Bu değerler tahmin beyanında değerlendirilirse $P\{0.9435 - 2.969(0.1353) \le \rho_{xy} \le 0.9435 - 2.969(0.1353)\} = 0.95 \Rightarrow P\{0.5418 \le \rho_{xy} \le \mathbf{1}\ (1.3452\} = 0.95$ bulunur. Yorum: (0.5418; **1**) aralığının ρ_{xy} parametresini içinde bulundurma olasılığı % 95'dir, denilebilir.

♦ **Ödev:** Tesadüfî olarak seçilen yedi örneklem biriminden gelir (X) ve kültürel harcama (Y) değişkenlerine ilişkin veriler aşağıdaki tabloda verildiği gibi tespit edilmiştir.

Gelir (x_i, 1000 TL)	3	4	4	5	8	12	13
Kültürel Harcama (y_i, 1000 TL)	0.5	0.5	0.5	1.2	2	2	1

% 95 güvenle kitle korelasyon katsayısı için güven aralığı tahmin ediniz ve sonucu yorumlayınız. (C: $r_{xy} = 0.6427;\ \ s_r = 0.3426;\ t_T = t_{5;\,0.025} = 3.1634$ olup (-0.4411; **1**(1.7265)) olarak tahmin edilir.)

X.3 Korelâsyon İçin Hipotez Testi

Korelâsyon katsayılarına ilişkin hipotez testlerinde çeşitli test istatistikleri kullanılmaktadır. Bunun nedeni ise istatistiklerin örnekleme dağılımlarının farklı olmasıdır.

X.3.1 Tek Kitle Korelasyonu İçin Hipotez Testi

X.3.1.1 Test İstatistiği Olarak t-Dağılımının Kullanılması

Araştırma hipotezi, sadece; " değişkenler arasında ilişki vardır" şeklinde ise (yani; ilişkinin kuvveti ile ilgili değil ise) H_0 hipotezinin test edilmesinde $(n-2)$ s.d.'li t-dağılımı kullanılır. Test algoritması ise aşağıdaki gibi verilir.

a) **Hipotezler:** $H_0: \rho_{XY} = 0$ hipotezine karşılık üç farklı iddiada bulunulabilir. Bunlar: i) $H_A: \rho_{XY} < 0$ (ilişki ters yöndedir), ii) $H_A: \rho_{XY} > 0$ (ilişki aynı yöndedir) ve $H_A: \rho_{XY} \neq 0$ (ilişki yoktur) şeklinde kurulur.

b) **Test İstatistiği:** H_0'ın doğruluğu varsayımı altında test istatistiği, $t_H = \frac{r_{xy}}{s_r} \sim t_{(n-2)}$'dir.

c) **Karar:** Eğer araştırma hipotezi tek yönlü (i ve ii'de verildiği gibi) ise $t_T = t_{(n-2);\,\alpha}$ ve eğer araştırma hipotezi çift yönlü (iii'de verildiği gibi) ise $t_T = t_{(n-2);\,\frac{\alpha}{2}}$ kritik tablo değerini göstersin. Eğer $|t_H| \leq |t_T| \Rightarrow H_0$ hipotezi kabul edilir ve eğer $|t_H| > |t_T| \Rightarrow H_0$ hipotezi ret edilir.

d) **Yorum:** Karara ve iddiaya göre yorum yapılır.

♦ **Örnek:** Derse devamsızlık (X) arttıkça başarı notunun (Y) azaldığı iddia edilmektedir. Öğrenciler arasından tesadüfi olarak seçilen dokuz tanesinin ilgili değişkenlere ilişkin gözlem değerleri aşağıda verildiği gibi tespit edilmiş olsun.

Derse devamsızlık (x_i, gün)	0	1	1	3	4	6	6	7	8
Başarı notu (y_i, 100)	90	90	80	70	70	50	50	45	40

%5 önem seviyesinde iddianın doğruluğu hakkındaki kararınız ne olur.

♦ **Çözüm:**

a) Hipotezler: $H_0: \rho_{XY} = 0$ ve $H_A: \rho_{XY} < 0$ şeklinde kurulur.

b) Test İstatistiği: $t_H = \frac{r_{xy}}{s_r} \cong \frac{-0.9852}{0.0647} \cong -15.227$ bulunur. Burada bazı hesaplamalar ise $r_{xy} = \frac{\sum xy - n\bar{x}\bar{y}}{\left(\sqrt{\sum x^2 - n(\bar{x})^2}\right)\left(\sqrt{\sum y^2 - n(\bar{y})^2}\right)} = \frac{1895 - 9(4)65}{\sqrt{[212 - 9(4)^2][41025 - 9(65)^2]}} \cong -0.9852$ bulunur ve böylece $s_r = \sqrt{\frac{1 - r_{xy}^2}{n-2}} \cong 0.0647$ elde edilir.

c) Karar: $t_T = t_{7;\,0.05} = -1.895$ (t tablosundan) bulunur. Öyle ise $|t_H| > |t_T|$ olduğundan H_0 hipotezi ret edilir.

d) Yorum: % 95 güvenilirlikle derse devamsızlık arttıkça başarı notunun azaldığı söylenebilir. Yani, iddia geçerlidir.

♦ **Ödev:** Tamamlanan eğitim yılı (X) arttıkça kültürel harcamanın (Y) da arttığı iddia edilmektedir. Tesadüfî olarak seçilen yedi örneklem biriminden ilgili değişkenlere ilişkin aşağıda verilen gözlem değerleri elde edilmiş olsun.

Tamamlanan Eğitim Süresi (x_i, yıl)	5	5	8	11	15	15	18
Kültürel Harcama (y_i, 100 TL)	2	2	3	7	10	12	12

%5 önem seviyesinde iddianın geçerliliğini araştırınız.

(C: $H_0: \rho_{XY} = 0$ ve $H_A: \rho_{XY} > 0$; $r_{xy} \cong 0.9836$; $s_r \cong 0.0806$ ve $t_H \cong 12.204$; $t_T = t_{5;\,0.05} = 2.015$ olup $|t_T| < |t_H|$ olduğundan H_0hipotezi ret. İddia geçerlidir.)

X.3.1.2 Test İstatistiği Olarak z-Dağılımının Kullanılması

Araştırma hipotezi her zaman "değişkenler arasında ilişki vardır" şeklinde olmayabilir. Ayrıca bu ilişkinin miktarı da hedeflenirse ve $\rho_{XY} = \rho_0$ ve $\rho_0 \neq 0$ ile gösterilirse test istatistiği olarak z_H istatistiği kullanılır. Bu durumda test algoritması ise şöyle verilir.

a) **Hipotezler:** $H_0: \rho_{XY} = \rho_0$ hipotezine karşılık üç farklı iddiada bulunulabilir. Bunlar: i) $H_A: \rho_{XY} < \rho_0$ (ilişki iddia edilen değerden küçüktür), ii) $H_A: \rho_{XY} < \rho_0$ (ilişki iddia edilen değerden büyüktür) ve $H_A: \rho_{XY} \neq \rho_0$ (ilişki iddia edilen değerden farklıdır) şeklinde kurulur.

b) **Test İstatistiği:** H_0 hipotezinin doğruluğu varsayımı altında test istatistiği, $z_H = \frac{z_r - E(z_r)}{\sigma_{z_r}} \sim N(0;1)$'dir. Burada $z_r = \frac{1}{2} ln\left(\frac{1+r_{xy}}{1-r_{xy}}\right)$, $E(z_r) = \frac{1}{2} ln\left(\frac{1+\rho_0}{1-\rho_0}\right)$ ve $\sigma_{z_r} = \sqrt{\frac{1}{n-3}}$ ile verilir. Burada z_r, Fisher'in standart normal dağılım dönüşümü olarak bilinir.

c) **Karar:** Eğer araştırma hipotezi tek yönlü (i ve ii'de verildiği gibi) ise $z_T = z_{\alpha}$ ve eğer araştırma hipotezi çift yönlü (iii'de verildiği gibi) ise $z_T = z_{\frac{\alpha}{2}}$ kritik tablo değerini göstersin. Eğer $|z_H| \leq |z_T| \Rightarrow H_0$ hipotezi kabul edilir ve eğer $|z_H| > |z_T| \Rightarrow H_0$ hipotezi ret edilir.

d) **Yorum:** Karara ve iddiaya göre yorum yapılır.

♦ **Örnek:** Gelir (X) ve ödenen kira (Y) değişkenleri arasındaki korelâsyonun %80'den daha büyük olduğu iddia edilmektedir. Tesadüfi olarak seçilen sekiz kiracının söz konusu değişkenlerine ilişkin aşağıda verilen gözlem değerleri elde edilmiş olsun.

Aylık Gelir (x_i, 1000 TL)	4	4	5	7	9	12	14	17
Aylık Kira (y_i, 1000 TL)	1	1	2	3	4	5	5.5	6.5

%5 önem seviyesinde iddianın geçerliliğini araştırınız.

♦ **Çözüm:**

a) Hipotezler: $H_0: \rho_{XY} = 0.80$ ve $H_A: \rho_{XY} > 0.80$ şeklinde kurulur.

b) Test İstatistiği: $z_H = \frac{z_r - E(z_r)}{\sigma_{z_r}} \cong \frac{2.4393 - 1.0986}{0.4472} \cong 2.998$ bulunur. Burada bazı hesaplamalar ise $r_{xy} = \frac{\sum xy - n\bar{x}\bar{y}}{\left(\sqrt{\sum x^2 - n(\bar{x})^2}\right)\left(\sqrt{\sum y^2 - n(\bar{y})^2}\right)} = \frac{7050}{\sqrt{[16800][3050]}} \cong 0.9849$ bulunur ve böylece $s_r = \sqrt{\frac{1}{n-3}} = \sqrt{\frac{1}{5}} \cong 0.4472$ elde edilir. Ayrıca $z_r = \frac{1}{2} ln\left(\frac{1+0.9849}{1-0.9849}\right) \cong 2.4393$ ve $E(z_r) = \frac{1}{2} ln\left(\frac{1+0.8}{1-0.8}\right) \cong 1.0986$

c) Karar: $z_T = z_{0.05} = 1.645$ (z tablosundan) bulunur. Öyle ise $|z_H| > |z_T|$ olduğundan H_0 hipotezi ret edilir.

d) Yorum: % 95 güvenilirlikle aylık gelir arttıkça ödenen kira artışındaki korelasyon değerinin % 80'de daha yüksek olduğu söylenebilir. Yani, iddia geçerlidir.

♦ **Ödev:** Gelir (X) ve karbon hidratlı yiyeceklere yapılan harcama (Y) değişkenleri arasındaki korelâsyonun -%30'dan daha küçük olduğu iddia edilmektedir. Tesadüfi olarak seçilen 20 örneklem biriminden söz konusu değişkenlerine ilişkin örneklem korelasyon tahmin değeri $r_{xy} = -0.5$ olarak hesaplanmış olsun. %1 önem seviyesinde iddianın geçerliliği hakkında ne söylenebilir. (**C:** Hipotezler: $H_0: \rho_{XY} = -0.30$ ve $H_A: \rho_{XY} < -0.30$; $z_H \cong -0.989$; $z_T = z_{0.01} = -2.33$ olup $|z_H| < |z_T|$ olduğundan H_0 hipotezi kabul edilir. İddia geçersizdir.)

X.3.2 İki Kitle Korelasyon Katsayısına İlişkin Hipotez Testi

$\rho_{XY(1)}$, birinci kitlenin ve $\rho_{XY(2)}$ de ikinci kitlenin aynı X ile Y değişkenleri arasındaki kitle korelasyon katsayılarını göstersin. Bu durumda test algoritması aşağıdaki gibi verilir.

a) **Hipotezler:** $H_0: \rho_{XY(1)} = \rho_{XY(2)}$ (veya $H_0: \rho_{XY(1)} - \rho_{XY(2)} = 0$) hipotezine karşılık üç farklı iddiada bulunulabilir. Bunlar: i) $H_A: \rho_{XY(1)} < \rho_{XY(2)}$ (veya $H_A: \rho_{XY(1)} - \rho_{XY(2)} < 0$; birinci kitlenin ilişki katsayısı ikinci kitlenin ilişki katsayısından daha küçüktür), ii) $H_A: \rho_{XY(1)} > \rho_{XY(2)}$ (veya $H_A: \rho_{XY(1)} - \rho_{XY(2)} > 0$; birinci kitlenin ilişki katsayısı ikinci kitlenin ilişki katsayısından daha büyüktür) ve iii) $H_A: \rho_{XY(1)} \neq \rho_{XY(2)}$ (veya $H_A: \rho_{XY(1)} - \rho_{XY(2)} \neq 0$; birinci kitlenin ilişki katsayısı ile ikinci kitlenin ilişki katsayısı arasında fark yoktur) şeklinde kurulur.

b) **Test İstatistiği:** H_0 hipotezinin doğruluğu varsayımı altında test istatistiği, $z_H = \frac{z_{r_1} - z_{r_2}}{\sigma_{z_{r_1} - z_{r_2}}} \sim N(0;1)$'dir. Burada $z_{r_1} = \frac{1}{2} ln\left(\frac{1+r_{xy(1)}}{1-r_{xy(1)}}\right)$, $Ez_{r_2} = \frac{1}{2} ln\left(\frac{1+r_{xy(2)}}{1-r_{xy(2)}}\right)$ ve $\sigma_{z_{r_1} - z_{r_2}} = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$ ile verilir. Burada z_{r_j}, $j = 1, 2$ Fisher'in standart normal dağılım dönüşümü olarak bilinir.

c) **Karar:** Eğer araştırma hipotezi tek yönlü (i ve ii'de verildiği gibi) ise $z_T = z_{\alpha}$ ve eğer araştırma hipotezi çift yönlü (iii'de verildiği gibi) ise $z_T = z_{\frac{\alpha}{2}}$ kritik tablo değerini göstersin. Eğer $|z_H| \leq |z_T| \Rightarrow H_0$ hipotezi kabul edilir ve eğer $|z_H| > |z_T| \Rightarrow H_0$ hipotezi ret edilir.

d) **Yorum:** Karara ve iddiaya göre yorum yapılır.

♦ **Örnek:** İki farklı iş dalından birincisinde işçilerin hizmet süresi (X) ve gelirleri (Y) arasındaki korelasyonu diğer iş dalındaki aynı değişkenler arasındaki korelasyon katsayısından daha büyük olduğu iddia edilmektedir. Birinci ve ikinci iş dallarından sırası ile $n_1 = 15$ ve $n_2 = 20$ adet işçi tesadüfî olarak seçilmiş ve örneklem korelâsyonları da sırası ile $r_{xy(1)} = 0.90$ ve $r_{xy(2)} = 0.80$ olarak tahmin edilmiştir. %5 önem seviyesinde iddianın geçerliliğini araştırınız.

♦ **Çözüm:**

a) Hipotezler: $H_0: \rho_{XY(1)} = \rho_{XY(2)}$ ve $H_A: \rho_{XY(1)} > \rho_{XY(2)}$ şeklinde kurulur.

b) Test İstatistiği: $z_H = \frac{z_{r_1} - z_{r_2}}{\sigma_{z_{r_1} - z_{r_2}}} \cong \frac{1.4722 - 1.0968}{0.3770} \cong 0.996$ bulunur. Burada bazı hesaplamalar ise $z_{r_1} = \frac{1}{2} ln\left(\frac{1+0.9}{1-0.9}\right) \cong 1.4722$; $z_{r_2} = \frac{1}{2} ln\left(\frac{1+0.8}{1-0.8}\right) \cong 1.0968$ ve $s_r = \sqrt{\frac{1}{17} + \frac{1}{12}} \cong 0.3770$ bulunur.

c) Karar: $z_T = z_{0.05} = 1.645$ (z tablosundan) bulunur. Öyle ise $|z_H| < |z_T|$ olduğundan H_0 hipotezi kabul edilir.

d) Yorum: % 95 güvenilirlikle iki kitle korelâsyonları arasında fark olmadığı söylenebilir. Yani, iddia geçersizdir.