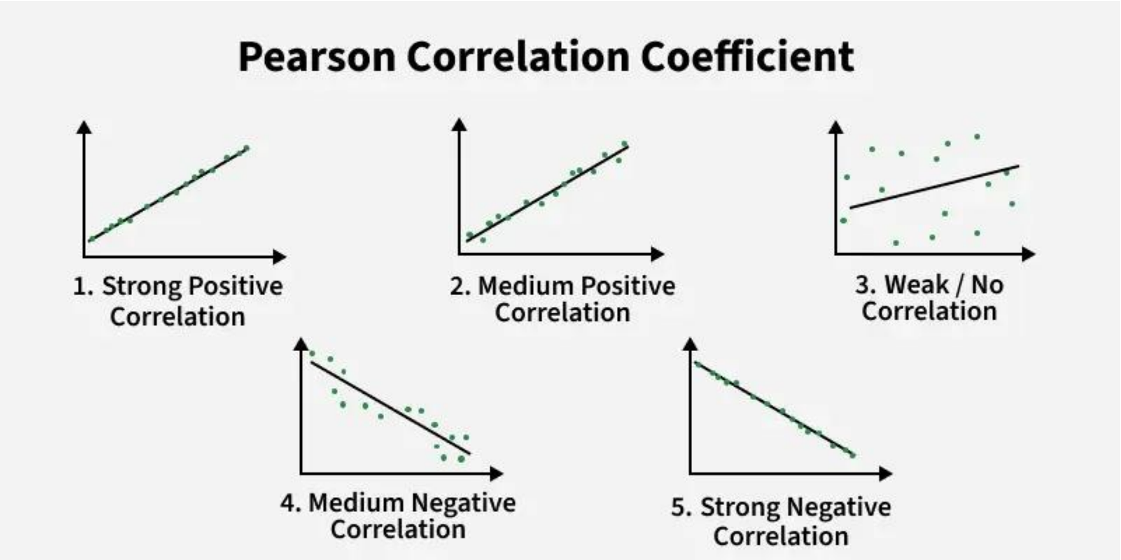
PSY 121- Statistics in Social Sciences

Correlation: Understanding the Relationship Between Variables-II

Pearson Correlation Coefficient

- The **Pearson Correlation Coefficient (r)** is a number that shows **how strong and in which direction** two **continuous variables** are related.
- It tells us **whether two things increase or decrease together**.
- The value of **r ranges from -1 to +1**.



Assumptions:

- The relationship between the two variables is assumed to be linear.
- Both variables must be continuous (interval or ratio scale).
- Both variables must be normally distributed.

Pearson Correlation Coefficient

Use Pearson correlation when:

- The variables are **continuous** (height, weight, score, age, time, etc.)
- The relationship is **linear**
- The variables follow a **normal distribution!**

Do you remember what the normal distribution is?

Pearson Correlation Coefficient

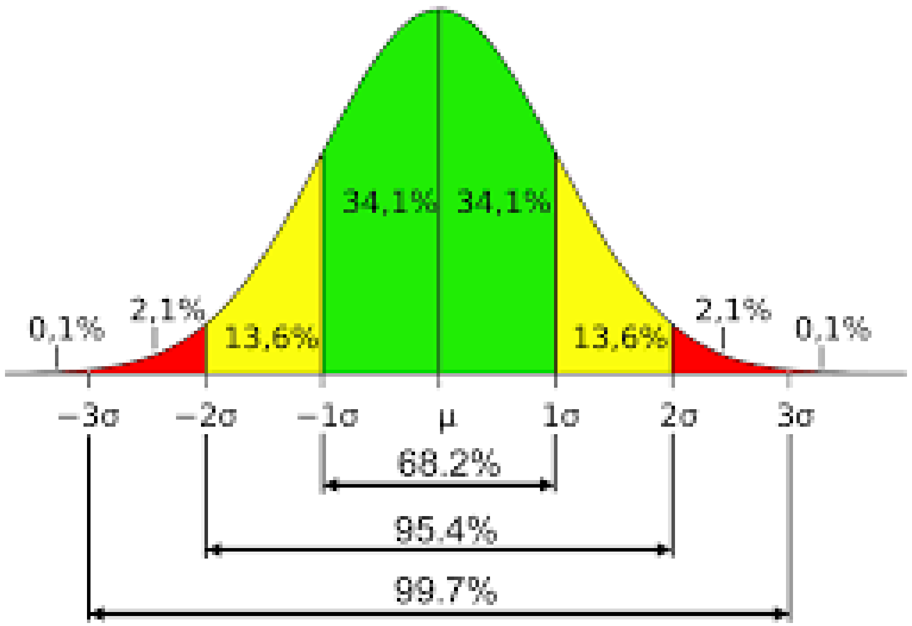
Normal dağılım nedir?

- Verilerin büyük çoğunluğunun **ortalama etrafında toplandığı**,
- Ortalamadan uzaklaştıkça değerlerin sayısının **giderek azaldığı**,
- Şekil olarak **çan eğrisi (bell curve)** gibi görünen bir olasılık dağılımıdır.

Özellikleri

Simetriktir: Sağ ve sol taraf birbirinin aynısıdır.
Ortalama, median ve mod birbirine eşittir.
Değerlerin yaklaşık:

- %68'i ortalamanın ±1 standart sapma
- %95'i ortalamanın ±2 standart sapma
- %99.7'si ortalamanın ±3 standart sapma içindedir.



Pearson Correlation Coefficient

- It is used to explain the linear relationship between two **continuous** variables measured at the ratio scale.
- **+1 → Perfect positive relationship** (one increases, the other also increases)
- 0 → No linear relationship
- **–1 → Perfect negative relationship** (one increases, the other decreases)

$$r = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Where, $\overline{X}$ = mean of X variable

$\overline{Y}$ = mean of Y variable

Pearson Correlation Coefficient

- Bir sınıfta **5 öğrencinin** iki sınavdan aldığı notlar:

Öğrenci	Matematik (X)	Fen (Y)
A	80	85
B	90	95
C	70	75
D	60	65
E	85	80

$$r = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Where, $\overline{X}$ = mean of X variable
$\overline{Y}$ = mean of Y variable

1. **Ortalama Hesapla**

Xmean= 77
Ymean= 80

$\Sigma(X-\bar{X})(Y-\hat{Y}) = 500$
$\Sigma(X-\bar{X})^2 = 580$
$\Sigma(Y-\hat{Y})^2 = 500$

X	Y	X−X̄	Y−Ŷ	(X−X̄)(Y−Ŷ)	(X−X̄)²	(Y−Ŷ)²
80	85	3	5	15	9	25
90	95	13	15	195	169	225
70	75	-7	-5	35	49	25
60	65	-17	-15	255	289	225
85	80	8	0	0	64	0

Pearson Correlation Coefficient

- Bir sınıfta **5 öğrencinin** iki sınavdan aldığı notlar:

Öğrenci	Matematik (X)	Fen (Y)
A	80	85
B	90	95
C	70	75
D	60	65
E	85	80

$$r = \frac{\Sigma(X-\overline{X})(Y-\overline{Y})}{\sqrt{\Sigma(X-\overline{X})^2}\sqrt{(Y-\overline{Y})^2}}$$

Where, $\overline{X}$ = mean of X variable
$\overline{Y}$ = mean of Y variable

r = 0.93

How do you interpret the result?

X	**Y**	**X–X̄**	**Y–Ŷ**	**(X–X̄)(Y–Ŷ)**	**(X–X̄)²**	**(Y–Ŷ)²**
80	85	3	5	15	9	25
90	95	13	15	195	169	225
70	75	-7	-5	35	49	25
60	65	-17	-15	255	289	225
85	80	8	0	0	64	0

Interpretation of the Correlation Coefficient

1. MAGNITUDE

- The absolute value of the correlation coefficient is interpreted based on the threshold values given in the “Strength of Relationship” section
  [0.01–0.30–0.70 → low, moderate, high].

2. DIRECTION

- If the *r* value is **negative**, it means that **one variable increases while the other decreases.**

- If it is **positive**, it indicates that **both variables change in the same direction** *(both increase or decrease together).*

3. EXPLAINED VARIANCE

- It expresses, as a percentage, how much of the variation in one variable is explained by the other variable.
- It is the square of the correlation coefficient.

Explained Variance (Açıklanan Varyans) -r^2

- It allows expressing, as a percentage, how much of the variation in one variable is explained by the other variable.
- The amount of variance that the variables explain in each other is equal to the square of the correlation coefficient, which is called the coefficient of determination.

Example: Let the correlation between Turkish achievement and reading speed be *r = 0.80*. Accordingly, the coefficient of determination is:

$$r^2 = 0.64$$

- This means that **64% of the total variability in students' reading speed** is explained by their achievement in Turkish. Alternatively, it can be interpreted as **64% of the variation in students' Turkish achievement** being explained by their reading speed.
- It can also be stated that the remaining **36%** of the variance is due to other variables.

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.709	46.367	46.367	3.709	46.367	46.367	2.501	31.260	31.260
2	1.478	18.471	64.838	1.478	18.471	64.838	2.045	25.566	56.826
3	1.361	17.013	81.850	1.361	17.013	81.850	2.002	25.024	81.850
4	.600	7.499	89.349						
5	.417	5.214	94.563						
6	.281	3.508	98.071						
7	.129	1.608	99.679						
8	2.569E-02	.321	100.000						

Extraction Method: Principal Component Analysis.

Spearman's Rank Correlation Coefficient (*r_s*)

- This coefficient is used to measure the **relationship between two variables when the data are in ordinal (ranked) form**.
- It is helpful when your data are **not normally distributed** or when you are working with **rank orders instead of exact numerical values**.

It is often used in situations such as:

- When scores are ranked (1st, 2nd, 3rd, etc.).
- When the assumption of <u>normal distribution is not met</u>.
- When you want a **non-parametric** measure of correlation.

$$r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

Spearman's Rank Correlation Coefficient (r_s)- *example*

- Tabloda, 6 öğrencinin iki sınavdaki sıraları
- Veriler (A sınavı sırası, B sınavı sırası):
- Her gözlem için sıra farkı $d = A - B$:

Öğrenci 1: $d = 1 - 2 = -1$
Öğrenci 2: $d = 2 - 1 = 1$
…
Her d için d^2:
$(-1)^2 = 1$
$1^2 = 1$
$(-1)^2 = 1$
$1^2 = 1$
$(-1)^2 = 1$
$1^2 = 1$
d^2toplamı: $\Sigma d^2 = 1 + 1 + 1 + 1 + 1 + 1 = 6$
n= 6

Öğrenci	A (sıra)	B (sıra)
1	1	2
2	2	1
3	3	4
4	4	3
5	5	6
6	6	5

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

Let’s do another example… ☺

- ***«Is there a relationship between the fruits students like and the frequency of eating those fruits?»***
- *(Öğrencilerin sevdiği meyveler ile o meyveleri yeme sıklıkları arasındaki ilişki var mı?)*

Data (Students’ answers):

$$r_s = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

Student	Favorite Fruit (1=en çok sevilen)	Yeme Sıklığı (1=en sık yenilen)
A	Elma (1)	2
B	Muz (2)	4
C	Çilek (3)	3
D	Portakal (4)	1

Pearson Correlation vs Spearman Correlation

Pearson Correlation	Spearman Correlation
Pearson Correlation can be us only in a linear relationship between two variables	Spearman Correlation can be used in a monotonic relationship between two variables, regardless of linearity
Data must be continuous and quantitative	Data must be ordinal and qualitative
Works on raw data	Works on ranked values for each variable
The calculation is based on the covariance of the two variables	The calculation is based on the ranks of the two variablesvariable
Formula: $r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 (y_i - \bar{y})^2}}$	Formula: $p = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$
Indicates the strength and direction of the linear relationship between the two variablescovariance of the two variables	Indicates the strength and direction of the monotonic relationship between the two variablesof the two variablesvariable
Pearson correlation is also more sensitive to outliers, so it is important to check your data for outliers before using Pearson correlation.	The Spearman correlation is less powerful than the Pearson correlation but more robust to outliers than the Pearson correlation.

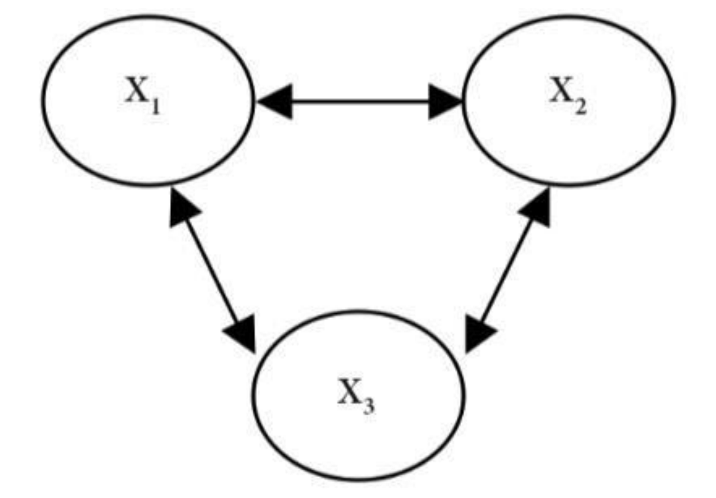
Partial Correlation (Kısmi Korelasyon)

- The relationship between two variables **after controlling for the variance caused by one or more additional variables** is called a **partial correlation**.
  (*iki değişken arasındaki ilişkiyi, **bir veya daha fazla başka değişkenin etkisini kontrol ederek** ölçer*)
- It shows the "pure" correlation between two variables when the influence of another variable is removed.

Partial correlation has **<u>two fundamental assumptions</u>**:

1. The variables whose relationship is being examined and the control variable must be **continuous**.

2. All variables included in the analysis must **show a normal distribution**.



Şekil : X_1, X_2 ve X_3 Değişkenleri Arasındaki İlişki

Multiple Correlation (Çoklu Korelasyon)

- **Multiple correlation** examines the relationship between **one dependent variable** and **a linear combination of two or more independent variables**.

- It is represented by the symbol **R**, not *r*.

- In other words, it shows **how well a set of predictors (X_1, X_2, X_3 ...)** together explain or predict a **dependent variable (Y)**.
- If **R is close to 1**, predictors explain Y very well.
- If **R is close to 0**, predictors do *not* explain Y well.

Tablo *: Çoklu Korelasyon İçin Hipotetik Veriler*

Hastalığın Şiddeti (X_1)	Sağlık Sigortası Ödemesi (X_2)	Tedavi Süresi (Y)
5	25	20
3	0	10
7	150	120
5	20	80
4	0	2
10	10	60
6	20	60
8	10	30
5	15	60
8	100	120
$\overline{X}_1 = 6.10$	$\overline{X}_2 = 35.00$	$\overline{Y} = 56.20$
$Sx_1=2.02$	$Sx_2=47.01$	$S_Y=39.77$
	$r_{x1Y}= .51$	
	$r_{x2Y}= .83$	
	$r_{x1x2}= .33$	

Multiple Correlation- *example*

- A school psychologist wants to predict students' **final exam scores (Y)** using:

X_1: Study hours per week
X_2: Class attendance rate
X_3: Motivation score
She runs a multiple correlation analysis and gets:
R = 0.82

- **Interpretation:**

An **R of 0.82** means:

- The combination of study hours, attendance, and motivation **strongly predicts students' exam scores**.
- This tells the psychologist that these three variables **together** explain a large portion of the variation in exam performance.

Any questions???