# 단순 선형 회귀
# (Simple Linear Regression)

1. Correlation

부산대학교 정보·의생명공학대학
정보컴퓨터공학부

# 두 변수 사이의 연관성 이해

❖ Explanatory Variable (설명 변수) & Response Variable (반응 변수)

- Explanatory Variable → $X$, Response Variable → $Y$
- Ex) 아버지 키와 아들의 키, 수면제 종류와 수면 시간, 온도에 따른 장비의 고장 여부
    - Explanatory Variable ? Response Variable ?
- Explanatory Variable → Independent Variable (독립 변수), Response Variable → Dependent Variable (종속 변수)

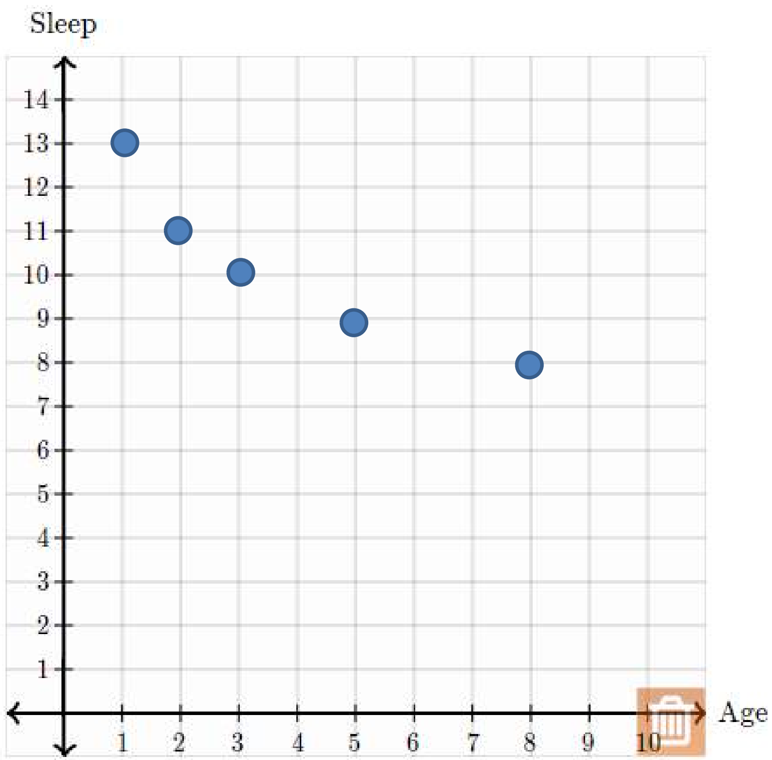
❖ 두 변수 사이의 관계와 연관성의 이해를 위한 도구들

- Scatter Plot (산점도)
- Correlation Coefficient (상관계수)
- Linear Regression (선형 회귀)

# Drawing a Scatter Plot

- Scatter Plot
  - Scatter Graph, Scatter Chart, Scattergram, Scatter diagram
  - X-axis : Explanatory Variable
  - Y-axis : Response Variable
- Colab
  - Matplotlib
    - Import matplotlib.pyplot as plt
    - plt.scatter()
  - Seaborn
    - Import seaborn as sns
    - sns.scatterplot()
    - sns.regplot()
- (Google)Spreadsheet

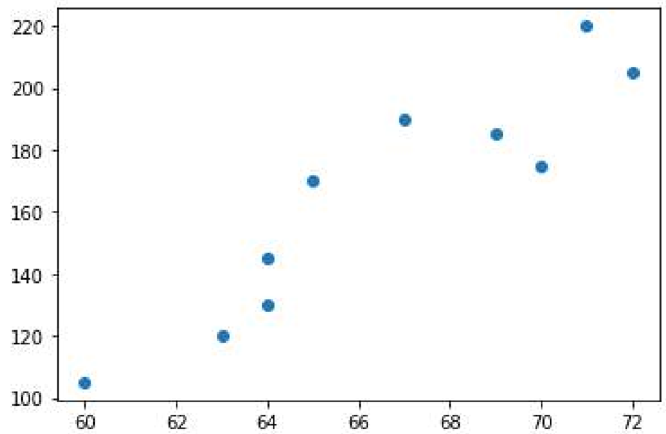
| Age (years) | 1 | 2 | 3 | 5 | 8 |
|---|---|---|---|---|---|
| Sleep (hours) | 13 | 11 | 10 | 9 | 8 |



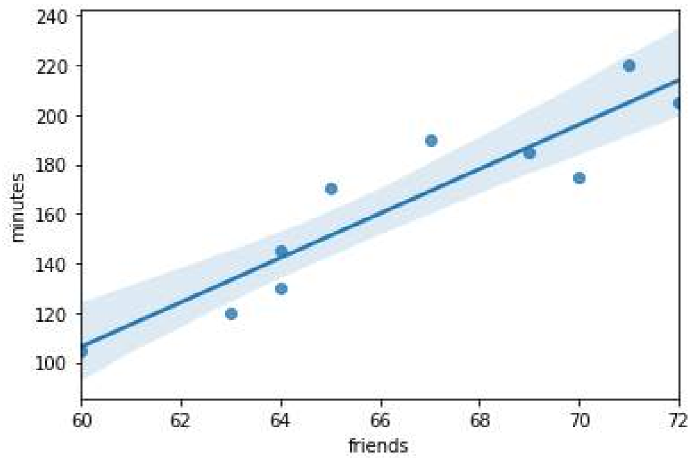
```python
import pandas as pd
datum = pd.read_csv('https://raw.githubusercontent.com/inetguru/IDS-CB35533/main/datum.csv', index_col='id')
datum.head()
```

| id | name | friends | minutes |
|---|---|---|---|
| 0 | Hero | 70 | 175 |
| 1 | Dunn | 65 | 170 |
| 2 | Sue | 72 | 205 |
| 3 | Chi | 63 | 120 |
| 4 | Thor | 71 | 220 |

```python
import matplotlib.pyplot as plt
plt.scatter(datum['friends'], datum['minutes'])
plt.show()
```



```python
import seaborn as sns
#sns.scatterplot(x='friends',y='minutes', data=datum[['friends','minutes']])
sns.regplot(x='friends',y='minutes', data=datum[['friends','minutes']])
plt.show()
```



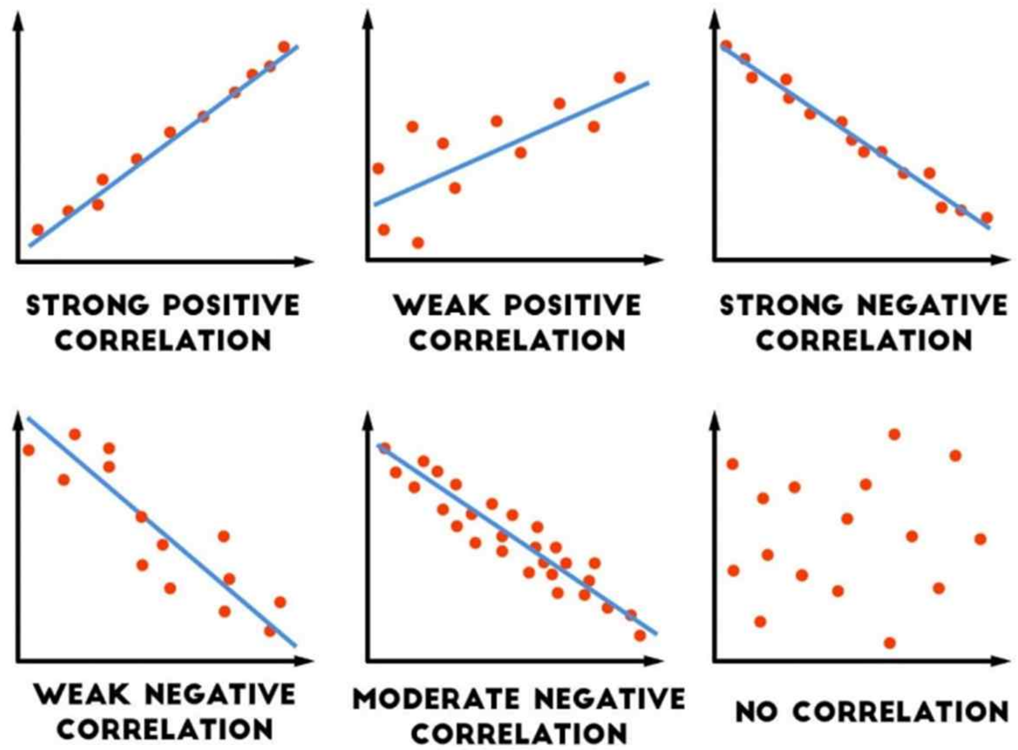


| id | name | friends |
|---|---|---|
| 0 | Hero | 70 |
| 1 | Dunn | 65 |
| 2 | Sue | 72 |
| 3 | Chi | 63 |
| 4 | Thor | 71 |
| 5 | Clive | 64 |
| 6 | Hicks | 60 |
| 7 | Devin | 64 |
| 8 | Kate | 67 |
| 9 | Klein | 69 |

# Patterns or Relationships in Scatterplot

❖ **Correlation** or dependence is any **statistical relationship**, whether causal or not, between two random variables or bivariate data.

- it commonly refers to the degree to which a pair of variables are linearly related.




= Uncorrelated

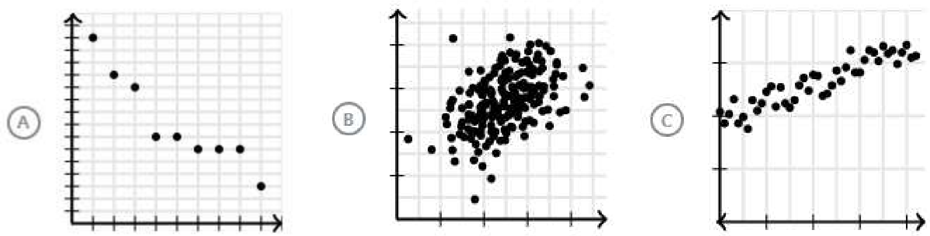
# Describing Scatterplots

❖ Form, Direction, Strength, Outliers

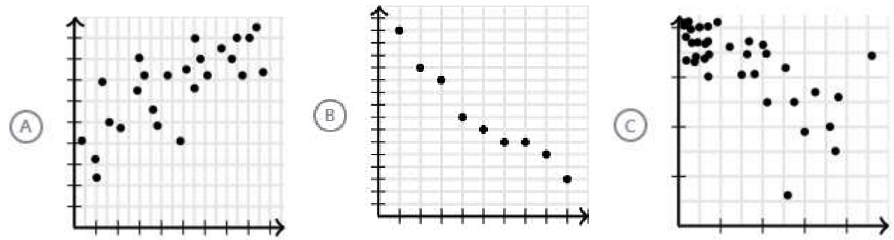
- **Form**: Is the association linear or nonlinear?
- **Direction**: Is the association positive or negative?
- **Strength**: Does the association appear to be strong, moderately strong, or weak?
- **Outliers**: Do there appear to be any data points that are unusually far away from the general pattern?

❖ Practice : choose the scatterplot that best its this description

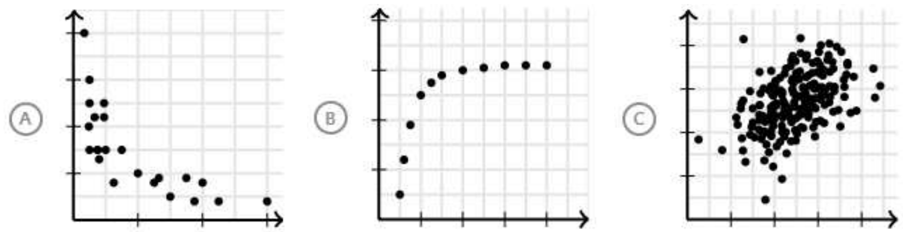
- A **strong, positive, linear** association between 2 variables




- **A moderately strong, negative, linear** association between the two variables with a few potential outliers.




- **A strong, negative, nonlinear association** between the two variables.



# Correlation Coefficient (상관 계수)

❖ (Pearson) Correlation Coefficient :

a measure of linear correlation (direction and strength) between two sets of data.

- also referred to as Pearson's $r$, or the bivariate correlation.

❖ Definition for a population

- Given a pair of random variables $(X, Y)$

$$\rho_{X,Y} = \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \cdot \sigma_Y}$$

- $\sigma_X$ : the standard deviation of $X$, $\sigma_Y$ : the standard deviation of $Y$, $\mu_X$ : is the mean of $X$, $\mu_Y$ : is the mean of $Y$

❖ **Definition for a sample**

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

- $n$ : sample size, $x_i, y_i$ are the individual sample points indexed with $i$. $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ , $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, the sample mean

# Calculating Correlation Coefficient

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

| id | name | friends | minutes | $x_i - \bar{x}$ | $y_i - \bar{y}$ |
|---|---|---|---|---|---|
| 0 | Hero | 70 | 175 | 3.5 | 10.5 |
| 1 | Dunn | 65 | 170 | -1.5 | 5.5 |
| 2 | Sue | 72 | 205 | 5.5 | 40.5 |
| 3 | Chi | 63 | 120 | -3.5 | -44.5 |
| 4 | Thor | 71 | 220 | 4.5 | 55.5 |
| 5 | Clive | 64 | 130 | -2.5 | -34.5 |
| 6 | Hicks | 60 | 105 | -6.5 | -59.5 |
| 7 | Devin | 64 | 145 | -2.5 | -19.5 |
| 8 | Kate | 67 | 190 | 0.5 | 25.5 |
| 9 | Klein | 69 | 185 | 2.5 | 20.5 |

- $\bar{x} = 66.5, \bar{y} = 164.5$
  - data['friends'].mean()
  - =AVERAGE(C2:C11)

- $\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = 1242.5$
  - sum( (data['friends']-data['friends'].mean()) * (data['minutes']-data['minutes'].mean()) )
  - =SUMPRODUCT(E2:E11,F2:F11)
- $\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2} = 11.7686023$
  - math.sqrt(sum((data['friends']- data['friends'].mean())**2))
  - =SQRT(SUMSQ(E2:E11))
- $\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 114.1161689$
  - math.sqrt(sum((data['minutes']-data['minutes'].mean())**2))
  - =SQRT(SUMSQ(F2:F11))

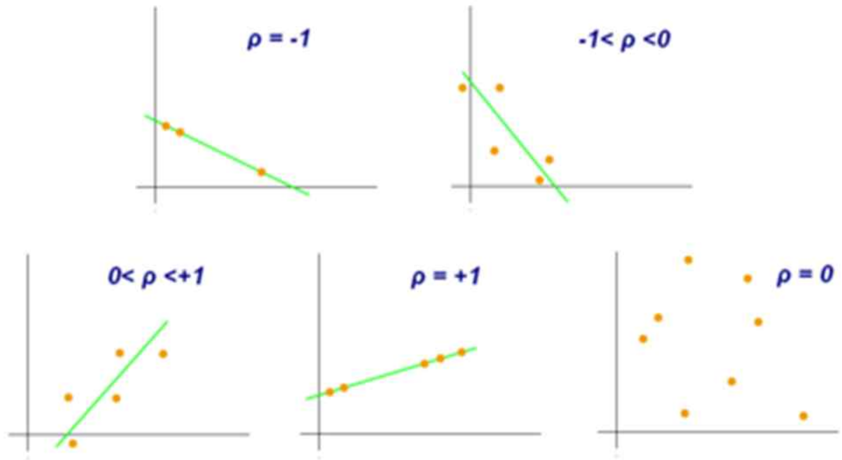
❖ $r_{xy} = 0.9251759349$

❖ Built-in Functions

- data['friends'].**corr**(data['minutes'])
- =**CORREL**(C2:C11,D2:D11)

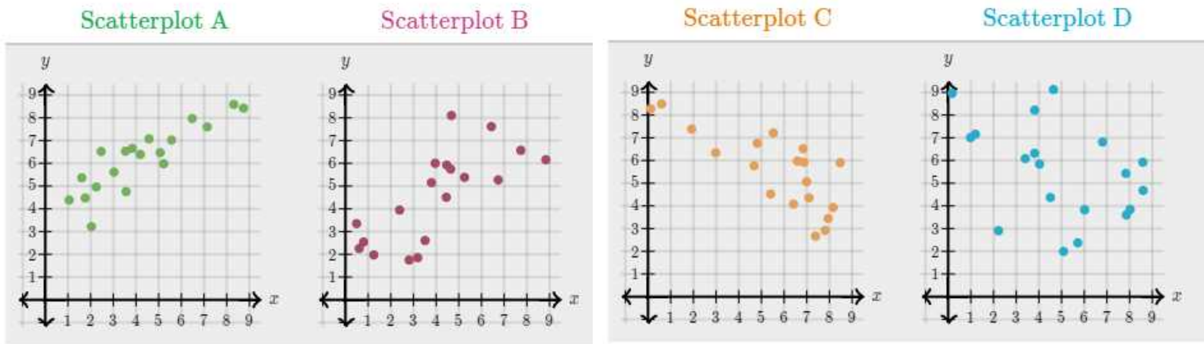
# Properties of Correlation Coefficient

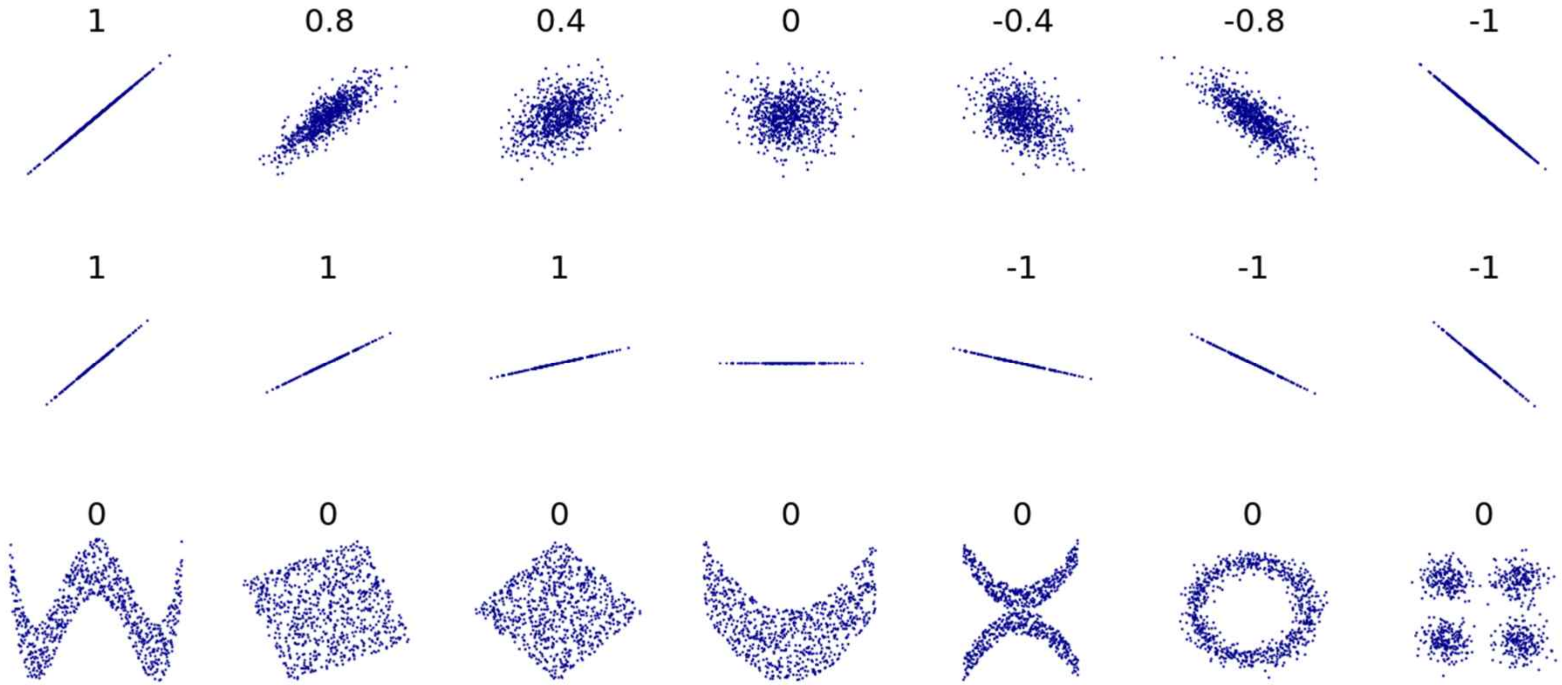
- It always has a value between $\boldsymbol{-1 \leq r \leq 1}$.
- Strong positive linear relationships have values of $r$ closer to 1.
- Strong negative linear relationships have values of $r$ closer to $-1$
- Weaker relationships have values of $r$ closer to 0




❖ Practice Example

- $r_1 = -0.42,\ r_2 = 0.73,\ r_3 = 0.87,\ r_4 = -0.77$
- Scatterplot A :
- Scatterplot B :
- Scatterplot C :
- Scatterplot D :






Source : https://upload.wikimedia.org/wikipedia/commons/thumb/d/d4/Correlation_examples2.svg/400px-Correlation_examples2.svg.png

Several sets of (*x*, *y*) points, with the correlation coefficient of *x* and *y* for each set. Note that the correlation reflects the strength and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of *Y* is zero.

# Various Coefficients

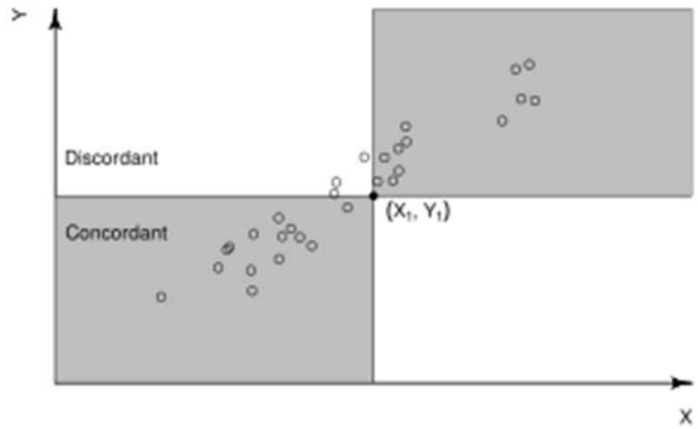
- 상관 계수는 이상치(Outlier Values)의 영향을 많이 받음
  - 이상치에 Robust한 상관 계수들이 개발됨

- Kendall's Tau ($\tau$) correlation coefficients

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{n(n-1)/2}$$

- Spearman's **Rank** correlation coefficients or Spearmans's $\rho$
  - The Spearman correlation coefficient is defined as the Pearson correlation coefficient **between the rankings of two variables**, or two rankings of the same variable

- Corr() function in Pandas
  - method = 'pearson', 'kendall', 'spearman'




All points in the gray area are concordant and all points in the white area are discordant with respect to point $(X_1, Y_1)$. With $n = 30$ points, there are a total of $\binom{30}{2} = 435$ possible point pairs. In this example there are 395 concordant point pairs and 40 discordant point pairs, leading to a Kendall correlation coefficient of 0.816.