AI기반 데이터 분석 및 AI Agent 개발 과정

『1과목 : 』
AI기반 데이터 분석

2025.09.22-10.02(9일, 62시간)

Prepared by DaeKyeong Kim

Ph.D.

삼성전자 구미사업부 G.제조팀

목차

『1-7』 상관관계 및 연관성 이해

변수 간의 관계

연관성 분석

Syllabus

학습목표

- 변수 간의 관계를 이해하고 측정할 수 있다
- 다양한 연관성 분석 방법을 습득한다
- 상관관계와 인과관계의 차이를 구분할 수 있다

눈높이 체크

- 변수 간의 관계
- 연관성 분석

1. 변수 간의 관계

변수란?

- 변수(Variable): 연구 대상의 특성이나 속성을 나타내는 값
- 관찰이나 측정을 통해 얻어지는 데이터의 기본 단위
- 1과 2 는 무엇인가? 숫자 > 키, 몸무게
- 홍길동은 무엇인가? 글씨 ? 이름 > 고객이름, 직원이름
  - <u>바나나, 사과 > 과일 이제서야 도메인이 보이나요?</u>
  - "ColombianMilds","EthiopianHarrar","EthiopianYirgacheffe","HawaiianKona","JamaicanBlueMountain" > CoffeeBeen
  - 학생 데이터는 이름, 나이, 성별, 키, 수학 점수 등
- 변수(variable) 는 값을 저장하는 이름.
  - 프로그래밍에서는 데이터를 저장하고 처리하기 위해 변수를 사용.

1. 변수 간의 관계

변수란?

- **변수의 유형**

◦ **질적 변수(범주형)**

– **명목변수: 성별, 혈액형, 거주지역**

– **순서변수: 만족도(상/중/하), 학년, 등급**

◦ **양적 변수(수치형)**

– **이산변수: 자녀 수, 결혼 횟수, 사고 건수**

– **연속변수: 키, 몸무게, 온도, 소득**

1. 변수 간의 관계

변수와 차원

- **변수: 데이터의 속성 하나하나**
- **차원: 변수들이 모여 만들어내는 공간의 크기**
  - 데이터셋에서 변수가 차지하는 축의 개수 (= 변수의 개수)
  - 표현:
    - 데이터 = 행렬 형태로 표현 가능
      - 행(Row) = 개체/샘플
      - 열(Column) = 변수
  - 예시:
    - 학생 100명의 [국어, 수학, 영어] 점수 데이터
      - 샘플 수 = 100 (행)
      - 변수 수 = 3 (열)
      - 차원 = 3차원 데이터

1. 변수 간의 관계

변수 간 관계의 종류

- **독립관계 (Independence)**
  - 한 변수의 변화가 다른 변수에 전혀 영향을 주지 않음
  - 예: 주사위 던지기 결과와 동전 던지기 결과

- **상관관계 (Correlation)**
  - 두 변수가 함께 변화하는 관계
  - 선형관계와 비선형관계로 구분

- **인과관계 (Causation)**
  - 한 변수의 변화가 다른 변수의 변화를 직접적으로 야기
  - 상관관계가 있다고 반드시 인과관계가 있는 것은 아님

2. 상관관계의 방향과 강도

상관관계의 방향

- 양의 상관관계 (Positive Correlation)
  - 한 변수가 증가할 때 다른 변수도 증가
  - 예: 공부시간과 성적, 키와 몸무게

- 음의 상관관계 (Negative Correlation)
  - 한 변수가 증가할 때 다른 변수는 감소
  - 예: TV 시청시간과 성적, 가격과 수요량

- 상관관계의 강도
  - 완전상관: $|r| = 1.0$
  - 강한상관: $0.7 \leq |r| < 1.0$
  - 보통상관: $0.3 \leq |r| < 0.7$
  - 약한상관: $0.1 \leq |r| < 0.3$
  - 무상관: $|r| \approx 0$

2. 상관관계의 방향과 강도

상관분석이란?

- 두 변수 간의 관계의 정도를 알아보기 위한 분석방법
- 두 변수의 상관관계를 알아보기 위해 사용

- 상관관계의 특성

상관계수 범위	해석
$0.7 < \gamma \le 1$	강한 양(+)의 상관이 있다.
$0.3 < \gamma \le 0.7$	약한 양(+)의 상관이 있다.
$0 < \gamma \le 0.3$	거의 상관이 없다.
$\gamma = 0$	상관관계(선형, 직선)가 존재하지 않음
$-0.3 \le \gamma < 0$	거의 상관이 없다.
$-0.7 \le \gamma < -0.3$	약한 음(-)의 상관이 있다.
$-1 \le \gamma < -0.7$	강한 음(-)의 상관이 있다.

2. 상관관계의 방향과 강도

상관분석이란?

- **상관분석과 상관관계**
  - 상관분석이란 두 변수 간에 관계가 있는지를 알아보고자 할 때 실시하는 분석방법
  - 상관관계란 두 변수(대상)이 서로 관련성이 있다고 추측되는 관계
- **상관계수(r)**
  - 상관분석에서 두 변수의 관련된 정도를 나타내주는 값

2. 상관관계의 방향과 강도

상관분석이란?

- X라는 양적 자료가 증가할수록 Y라는 양적 자료가 증가하면 양의 상관관계가 있다 고 하며, X가 증가할수록 Y가 감소하면 음의 상관관계가 있다고 한다. 하지만 X가 증 가하더라도 Y의 값이 별개로 있으면 상관관계가 없다고 할 수 있다. 두 양적 자료의 관련성 정도는 상관계수(coefficient of correlation)인 r로 표현되고, r은 -1 ~ +1의 사 이의 값을 가지고, r의 절대값이 1에 가까울수록 관련성이 높고, r의 절대값이 0에 가까울수록 관련성이 없다고 판단한다.

상관계수의 값	$-1 \le r \le 1$

2. 상관관계의 방향과 강도

공분산

- 두 확률변수 사이의 관계를 선형관계로 나타낼 때 두 변수 사이 상관의 정도를 나타내며 다음과 같이 구한다.

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))]$$

$$= E[(X - \mu_X)(Y - \mu_Y)], \quad E(X) = \mu_X, E(Y) = \mu_Y$$

◦ 두 확률변수 X, Y의 공분산은 $Cov(X, Y)$로 표기하고, 공분산이 갖는 값에 따라 두 확률변수의 관계를 확인할 수 있다.
◦ $Cov(X, Y) > 0$: 두 확률변수 X, Y 의 변화가 같은 방향임을 나타냅니다. 즉 X 증가하면 Y 도 증가하고, 반대로 한 변수가 감소하면 같이 감소한다.
◦ $Cov(X, Y) > 0$: 두 확률변수 X, Y 의 변화가 같은 방향임을 나타냅니다. 즉 X 증가하면 Y 도 증가하고, 반대로 한 변수가 감소하면 같이 감소한다.
◦ $Cov(X, Y) = 0$: 두 확률변수 간에 어떠한 (선형) 관계가 없음을 나타냄.

2. 상관관계의 방향과 강도

상관계수(correlation coefficient)

- 상관도
  - 두 변량 사이의 관계를 대략적으로 파악할 수 있는 그래프
- 단순상관계수
  - 상관도의 양상이 대체로 직선인 경우
- 상관분석
  - 변수 사이의 직선 관계를 상관계수를 이용하여 분석하는 것

- 상관분석은 변수들 간의 단순한 상호 관계성의 정도를 분석하는 통계적 기법
- 상관분석은 두 변수의 순서쌍으로 구성된 표본요소에 대해서 두 변수간의 상관관계를 표본 상관계수를 통해 나타냄

2. 상관관계의 방향과 강도

상관계수(correlation coefficient)

- **두 확률변수 *X*, *Y*의 공분산을 각 확률변수의 표준편차의 곱으로 나눈 값을 (모)상관계수라 하고, 기호로 *ρ_XY* (혹은 *ρ*)로 나타낸다.**

$$\rho_{XY} = \frac{Cov(X,\ Y)}{\sigma_X \sigma_Y} = \frac{E[(X-E(X))(Y-E(Y))]}{\sigma_X \sigma_Y}$$

$$\rho_{XY} \equiv Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

◦ **(모)상관계수는 -1부터 1사이의 값을 가진다.**

– 공분산의 경우 자료의 단위에 따라 값의 크기가 일정하지 않아 비교하기 힘들다.

– 공분산의 성질을 그대로 이어 받아 두 변수 간의 변화의 방향이 같으면 양수, 반대이면 음수를 갖는다.

◦ **(모)상관계수는 모집단의 특성 중에 하나로 일반적으로 알 수 없으며, 두 확률변수로부터 추출한 표본의 특성을 통해 구하는 (피어슨의) 표본상관계수를 이용하여 추정한다.**

2. 상관관계의 방향과 강도

상관계수(correlation coefficient)

- **상관계수의 특성**

① ρXY의 범위는 $-1 \le \rho XY \le 1$

② 두 변수가 서로 독립이면 두 변수 간에 상관관계가 없으며, ρXY =0

③ ρXY =0 이면 두 변수 간에 상관관계(선형관계)가 없다. 그러나 비선형관계는 있을 수 있기 때문에 두 변수가 서로 독립이라는 보장은 없다.

④ X와 Y가 정규분포를 따르는 경우, ρXY =0 이면 X와 Y는 독립

⑤ 양의 상관은 1에 가까워지고, 음의 상관은 -1에 가까워지고, 무상관은 0

$$
\begin{aligned}
r_{xy} &= \frac{S_{xy}}{S_x S_y} \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \overline{x}}{S_x} \right) \left(\frac{y_i - \overline{y}}{S_y} \right)
\end{aligned}
$$

2. 상관관계의 방향과 강도

표본상관계수(sample correlation coefficient)

- **모상관계수**
  - **표준화된 공분산을 두 변량 X 와 Y 사이의 모상관계수 p 라고 하며, 다음과 같다.**

$$\rho = \frac{\mathrm{Cov}(X,\ Y)}{\sigma_X \sigma_Y}$$

2. 상관관계의 방향과 강도

표본상관계수(sample correlation coefficient)

- 두 확률변수 X, Y로 부터 추출한 n개의 표본 쌍 $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$에서 확률변수 X로 부터 추출한 표본 $x_1, x_2, \dots, x_n$의 평균을 $\bar{x}$, 표준편차를 s_x, 확률변수 Y로 부터 추출한 표본 $y_1, y_2, \dots, y_n$의 평균을 $\bar{y}$, 표준편차를 s_y라 하면, 표본공분산 $cov(x, y)$는 다음과 같이 두 표본의 편차의 곱을 모두 합하고 이를 자료의 개수(표본 쌍의 개수) – 1 로 나누어 구한다.

$$cov(x, y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

2. 상관관계의 방향과 강도

표본상관계수(sample correlation coefficient)

- 표본공분산을 각 표본의 표준편차의 곱으로 나누어 구한다.
- 표본을 통하여 상관계수를 추정하는 통계량

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

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

$$r_{XY} = \frac{S_{XY}}{\sqrt{S_{XX}S_{YY}}}$$

$$S_{XX} = \sum_{i=1}^{n}(X_i-\bar{X})^2 = \sum_{i=1}^{n}X_i^2 - \frac{(\sum_{i=1}^{n}X_i)^2}{n}$$

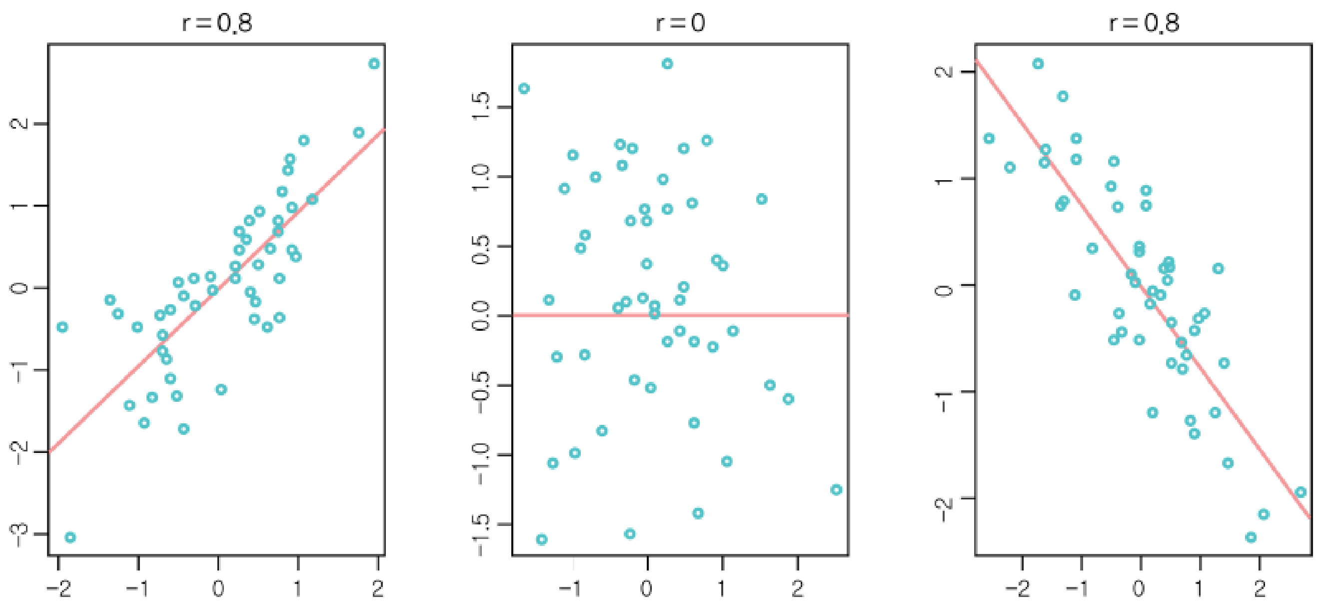
$$S_{YY} = \sum_{i=1}^{n}(Y_i-\bar{Y})^2 = \sum_{i=1}^{n}Y_i^2 - \frac{(\sum_{i=1}^{n}Y_i)^2}{n}$$

$$S_{XY} = \sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y}) = \sum_{i=1}^{n}X_iY_i - \frac{(\sum_{i=1}^{n}X_i)(\sum_{i=1}^{n}Y_i)}{n}$$

2. 상관관계의 방향과 강도

표본상관계수(sample correlation coefficient)

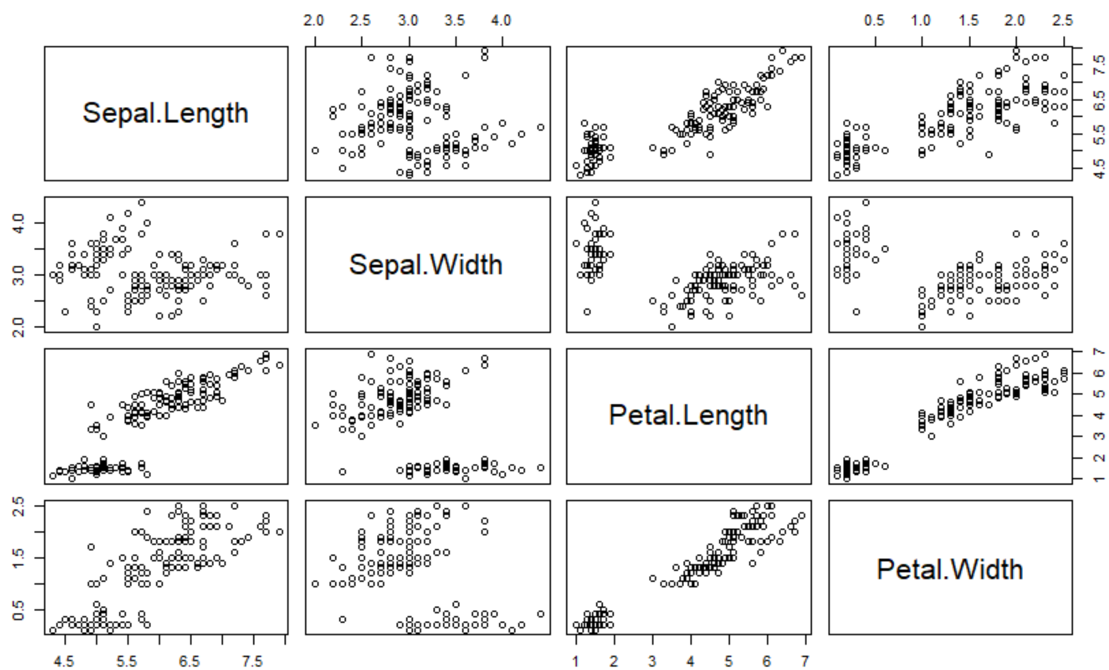
- 표본상관계수는 모상관계수와 동일한 성질을 가져
  - -1 혹은 1에 가까울수록 강한 상관을 나타내고,
  - 0에 가까이 갈수록 약한 상관을 나타낸다.
  - 양수일 경우 두 변수의 값의 변화는 같은 방향으로 진행되고, 음수일 경우 값의 변화는 서로 반대가 된다.



2. 상관관계의 방향과 강도

2차원 데이터의 시각화

- **산점행렬도**
  - 여러 개의 양적 자료에 대한 산점도를 하나의 그래프로 보여주는 것을 산점행렬도 (scatter plot matrix)라고 한다. plot(x, y) 형태로 산점도를 작성하는 번거로운 일이다. 어떻게 산점행렬도를 작성하는지 학습해 보자.
  - iris는 붓꽃이며 5개의 변수를 가지고 있다.



2. 상관관계의 방향과 강도

피어슨 상관계수 (Pearson Correlation Coefficient)

- **피어슨의 상관계수**
  - 정식 명칭은 피어슨의 곱적률 상관(Pearson's product-moment correlation)은 두 변수의 선형 관계가 존재할 경우 그 관계가 얼마나 강한지 알 수 있는 값이며 두 변수가 연속형 양적 변수일 경우에 사용가능한 방법이다.
  - 두 변수의 선형 관계를 측정

$$r = \frac{COV(X,Y)}{\sigma_x \times \sigma_y} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2} \times \sqrt{\sum (y_i - \bar{y})^2}}$$

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

2. 상관관계의 방향과 강도

피어슨 상관계수 (Pearson Correlation Coefficient)

- **특성:**
  - 범위: $-1 \le r \le +1$
  - 단위에 무관한 표준화된 측도
  - 선형관계만 측정 가능

- **해석 주의사항**
  - 상관관계 ≠ 인과관계
  - 비선형관계는 포착하지 못함
  - 이상값(outlier)에 민감

  - 예시: 키와 몸무게 (키가 커질수록 몸무게도 증가 → 양의 상관)

2. 상관관계의 방향과 강도

스피어만 순위상관계수

- 스피어만 순위상관계수 (Spearman's Rank Correlation)
- 스피어만 상관분석은 두 변수가 순서형 변수일 경우에 사용가능한 방법이며 두 변수가 정규성을 따르지 않는 경우에도 사용할 수 있는 비모수적 방법이다.
- 데이터 값을 순위(rank)로 바꾼 뒤, 두 순위 간의 단조(monotonic) 관계를 측정

$$\rho = 1 - \frac{6\sum d_i^2}{n^3 - n} \qquad t = \frac{\rho\sqrt{n-2}}{\sqrt{1-\rho^2}}$$

◦ 가설

– 귀무가설(H0): 두 변수간 선형관계가 존재하지 않는다. (rho = 0)

– 대립가설(H1): 두 변수간 선형관계가 존재한다. (rho ≠ 0)

2. 상관관계의 방향과 강도

스피어만 순위상관계수

- **적용 상황:**
  - 순서변수 간의 관계 분석
  - 비선형 단조관계 측정
  - 이상값의 영향을 줄이고자 할 때

  - 예시: 학생 시험 점수 순위와 운동 실력 순위 (점수는 선형이 아닐 수 있지만, 순위 간에는 관계가 있을 수 있음)

- **피어슨 vs 스피어만**
  - 피어슨: 선형관계, 연속변수
  - 스피어만: 단조관계, 순서변수 포함

2. 상관관계의 방향과 강도

켄달의 타우 (Kendall's Tau)

- **Kendall Correlation**
  - 상관분석은 두 변수가 순서형 변수일 경우에 사용가능한 방법이며 두 변수가 정규성을 따르지 않는 경우에도 사용할 수 있는 비모수적 방법이다.
  - 두 변수의 순위 쌍(pair)이 일치하는지 불일치하는지를 비교

$$\tau_A = \frac{n_c - n_d}{n_0}, \ \tau_B = \frac{n_c - n_d}{\sqrt{(n_0 - n_1)(n_0 - n_2)}}$$

$$\begin{aligned}
n_0 &= n(n-1)/2 \\
n_1 &= \sum t_i(t_i - 1)/2 \\
n_2 &= \sum u_j(u_j - 1)/2 \\
n_c &= Number\, of\, concordant\, pairs \\
n_d &= Number\, of\, discordant\, pairs \\
t_i &= Number\, of\, tied\, values\, in\, the\, i^th\, group\, of\, ties\, for\, the\, first\, quantity \\
u_j &= Number\, of\, tied\, values\, in\, the\, j^th\, group\, of\, ties\, for\, the\, second\, quantity
\end{aligned}$$

$$\tau_C = \frac{2(n_c - n_d)}{n^2 \frac{m-1}{m}}$$

$$\begin{aligned}
n_c &= Number\, of\, concordant\, pairs \\
n_d &= Number\, of\, discordant\, pairs \\
r &= Number\, of\, rows \\
c &= Number\, of\, columns \\
m &= min(r, c)
\end{aligned}$$

2. 상관관계의 방향과 강도

켄달의 타우 (Kendall's Tau)

- Kendall Correlation
  - 검정통계량 z

$$z_A = \frac{3(n_c - n_d)}{\sqrt{n(n-1)(2n+5)/2}}, \quad z_B = \frac{n_c - n_d}{\sqrt{v}}$$

$$\begin{aligned}
v &= (v_0 - v_t - v_u)/18 + v_1 + v_2 \\
v_0 &= n(n-1)(2n+5) \\
v_t &= \sum t_i(t_i-1)(2t_i+5) \\
v_u &= \sum u_j(u_j-1)(2u_j+5) \\
v_1 &= \sum t_i(t_i-1) \sum u_j(u_j-1)/2n(n-1) \\
v_2 &= \sum t_i(t_i-1)(t_i-2) \sum u_j(u_j-1)(u_j-2)/(9n(n-1)(n-2))
\end{aligned}$$

τ = (일치쌍 수 - 불일치쌍 수) / 전체 쌍의 수

2. 상관관계의 방향과 강도

켄달의 타우 (Kendall's Tau)

- **가설**
  - 귀무가설(H0): 두 변수간 선형관계가 존재하지 않는다. (tau = 0)
  - 대립가설(H1): 두 변수간 선형관계가 존재한다. (tau ≠ 0)

- **특징:**
  - 일치쌍과 불일치쌍의 비율로 계산
  - 스피어만보다 해석이 직관적
  - 표본 크기가 작을 때 더 안정적

- **적용 예시**
  - 두 심사위원의 순위 평가 일치도
  - 브랜드 선호도 순위와 구매 순위 관계
  - 예시: 영화 평점 순위와 관객 선호도 순위

2. 상관관계의 방향과 강도

상관분석의 비교

- 상관분석의 유형

	피어슨	스피어만
개념	등간척도 이상으로 측정된 두 변수들의 상관관계측정 방식	서열척도인 두 변수들의 상관관계 측정방식
특징	연속형 변수, 정규성 가정 대부분 많이 사용	순서형 변수, 비모수적 방법 순위를 기준으로 상관관계 측정
상관계수	피어슨 γ(적률상관계수)	순위상관계수(ρ, 로우)

2. 상관관계의 방향과 강도

상관계수 종류별로 계산하는 파이썬 코드 예시

```
import numpy as np
import pandas as pd
from scipy import stats
import seaborn as sns

# 샘플 데이터 생성
np.random.seed(42)
n = 10
x = np.arange(1, n+1)                    # 연속형 변수 (예: 공부 시간)
y = x + np.random.normal(0, 2, n)     # 연속형 변수 (예: 시험 점수, 약간의 노이즈 추가)
rank_y = np.argsort(np.argsort(y)) + 1 # 순위형 변수

binary = np.random.choice([0, 1], n)   # 이분형 변수 (예: 합격/불합격)
category1 = np.random.choice(['M','F'], n)   # 범주형 변수 1
category2 = np.random.choice(['A','B','C'], n) # 범주형 변수 2

# 1. 피어슨 상관계수
pearson_corr, _ = stats.pearsonr(x, y)
print("피어슨 상관계수:", round(pearson_corr, 3))

# 2. 스피어만 상관계수
spearman_corr, _ = stats.spearmanr(x, y)
print("스피어만 상관계수:", round(spearman_corr, 3))
```

2. 상관관계의 방향과 강도

상관계수 종류별로 계산하는 파이썬 코드 예시

```
# 3. 켄달의 타우
kendall_corr, _ = stats.kendalltau(x, y)
print("켄달 타우:", round(kendall_corr, 3))

# 4. 포인트-바이시리얼 상관계수 (연속형 vs 이분형)
pointbiserial_corr, _ = stats.pointbiserialr(x, binary)
print("포인트-바이시리얼 상관계수:", round(pointbiserial_corr, 3))

# 5. 크래머의 V (범주형 vs 범주형)
from sklearn.metrics import confusion_matrix

# 교차표 생성
conf_matrix = pd.crosstab(category1, category2)
chi2 = stats.chi2_contingency(conf_matrix)[0]
n_total = conf_matrix.sum().sum()
phi2 = chi2 / n_total
r, k = conf_matrix.shape
cramers_v = np.sqrt(phi2 / min(r-1, k-1))
print("크래머의 V:", round(cramers_v, 3))
```

『1-7』
상관관계 및 연관성 이해

변수 간의 관계

연관성 분석

1. 연관성 분석 개념

연관성 분석이란?

- 변수들 간의 관련성을 탐지하고 측정하는 통계적 방법
- 데이터에서 숨겨진 패턴과 관계를 발견
- 예측 모델링의 기초가 되는 탐색적 분석

연관성 분석의 목적

- 탐색적 목적: 데이터의 구조와 패턴 이해
- 예측적 목적: 한 변수로부터 다른 변수 예측
- 인과적 목적: 변수 간 인과관계 추론의 단초 제공

연관성의 유형

- 선형 vs 비선형
- 강한 연관성 vs 약한 연관성
- 직접적 vs 간접적 연관성

1. 연관성 분석 개념

범주형 변수의 연관성 - 카이제곱 검정

- 목적: 두 범주형 변수가 독립인지 연관이 있는지 검정
- 검정통계량:
  - $\chi^2 = \Sigma[(\text{관찰빈도} - \text{기대빈도})^2 / \text{기대빈도}]$

- 분할표 (Contingency Table)

	변수B		
변수A	B1	B2	합계
A1	n11	n12	n1.
A2	n21	n22	n2.
합계	n.1	n.2	n

- 크래머의 V (Cramér's V)
  - 카이제곱 통계량을 표준화한 연관성 척도
  - 범위: $0 \leq V \leq 1$
  - 표본 크기에 무관한 연관성 강도 측정

1. 연관성 분석 개념

범주형 변수의 연관성 - 카이제곱 검정

- **Cramér's V (크래머의 V)는 "두 범주형 변수 사이의 관계 강도"를 0~1 사이 값으로 표현하는 지표이며,카이제곱 검정 결과를 보완해 관계의 크기(Effect size)를 직관적으로 알려줌.**

1. 정의

•카이제곱(χ^2) 검정을 기반으로 계산된 연관성 척도
•두 변수 모두 범주형(nominal, 예: 성별, 지역, 제품 종류)일 때 사용
•값의 범위:

$$0 \le V \le 1$$

- **0** → 전혀 관계 없음
- **1** → 완벽한 관계 (하나의 변수가 다른 변수를 완벽하게 설명)

범주형 변수의 연관성 - 카이제곱 검정

2. 계산 공식

$$V = \sqrt{\frac{\chi^2/n}{\min(r-1, k-1)}}$$

- χ^2: 카이제곱 통계량
- n: 전체 표본 수
- r: 교차표의 행(row) 개수
- k: 교차표의 열(column) 개수

1. 연관성 분석 개념

범주형 변수의 연관성 - 카이제곱 검정

- **예제 데이터:**
  - **변수1: 성별 (남, 여)**
  - **변수2: 전공 (공학, 인문, 예술)**
- **교차표 (가상의 빈도수):**

	공학	인문	예술	합계
남	30	10	10	50
여	10	20	20	50
합계	40	30	30	100

- 이 교차표로 카이제곱 검정을 수행하면 χ2\chi^2χ2 값이 계산.
- 해당 값을 위 공식에 대입해 **Cramér's V**를 구하면, 성별과 전공 선택 간에 어느 정도의 연관성이 있는지 수치화할 수 있다.

1. 연관성 분석 개념

범주형 변수의 연관성 - 카이제곱 검정

4. 해석 기준 (경험적)

- 0.0 ~ 0.1: 거의 없음
- 0.1 ~ 0.3: 약한 연관
- 0.3 ~ 0.5: 중간 정도 연관
- 0.5 *이상*: 강한 연관

5. 활용 사례

- **마케팅:** 성별 vs 제품 선호도 관계
- **교육:** 학년 vs 과목 선택 관계
- **사회과학:** 지역 vs 투표 성향 관계

1. 연관성 분석 개념

연관규칙 분석 (Association Rule Mining)

- **대용량 데이터에서 항목들 간의 연관관계를 찾는 기법으로 좀 더 자세한 내용은 기타 데이터 마이닝에서 살펴봄.**
- **기본 개념:**
  - 장바구니 분석 (Market Basket Analysis)
  - "X를 구매한 고객은 Y도 구매할 가능성이 높다"

- **주요 측도**
  - 지지도 (Support)
    - $P(X \cap Y)$: X와 Y가 함께 발생할 확률
  - 신뢰도 (Confidence)
    - $P(Y|X)$: X가 발생했을 때 Y가 발생할 확률
  - 향상도 (Lift)
    - $P(Y|X)/P(Y)$: X 발생이 Y 발생 확률을 얼마나 높이는가

2. 연관규칙의 실제 적용

응용 분야

- **마케팅**
  - 교차판매 (Cross-selling) 전략
  - 상품 배치 최적화
  - 고객 세분화

- **추천시스템**
  - "이 상품을 본 고객이 함께 본 상품"
  - 협업 필터링의 기초

- **웹 마이닝**
  - 웹 사이트 구조 최적화
  - 사용자 행동 패턴 분석

- **실무 예시**
  - 맥주와 기저귀의 연관성
  - 아마존의 "함께 구매한 상품" 추천
  - 넷플릭스의 영화 추천 시스템

실습24 : 상관 분석-1

문제

[데이터변환축약데이터.xlsx] 첨부

전반적인 데이터의 상관관계 분석을 해주세요.

실습

실습25 : 상관 분석-2

문제
[정리된_통합_데이터_인덱스제거.csv] 파일 첨부
koreanize_matplotlib-0.1.1-py3-none-any.whl,
NanumBarunGothic.ttf 업로드한 라이브러리를 설치하고
Matplotlib 한글 사용 환경을 설정 한 다음 나눔체로 한글을 표현해 줘. 신선식품과 기온과의 상관관계를 분석하고 산점도를 그려줘.

THANK YOU.

앞으로의 엔지니어는 단순한 ‘코더’나 ‘기계 조작자’가 아니라 뇌-기계 인터페이스를 통해 지식과 능력을 즉각 확장하는 존재(뉴로-인터페이스: Neuro Interface)가 될 수 있습니다.

🎯 목표 달성을 위한 여정이 시작됩니다.
✨ 궁금한 점이 있으시면 언제든 문의해주세요!
🚀 함께 코더와 프롬프트 전문가로 성장해 나갑시다!