4장 데이터 분석기법-기초(2)

- **회귀분석(Regression Analysis)**
  - 단순선형회귀분석(Simple Linear Regression Analysis)
  - 최소제곱법(Least Square Method, LSM)
  - 반응변수의 변동과 결정계수

- **데이터 분석에서 주의사항**
  - 상관관계와 인과관계
  - 관찰연구와 실험연구
  - 데이터 수집 방법: 표본조사
  - 적절한 그래프 사용법

01 회귀분석

●회귀란?

·두 변수 간의 상관관계를 기본으로 하여

하나의 1차 선형식으로 두 변인의 관계를 일반화하는 분석방법

●회귀(Regression)

·평균으로의 회귀현상을 의미하며,

두 변수의 관계가 어떤 일반화된 선형관계의 평균으로 돌아간다는 의미

●선형성(Linearity)

·두 변수의 관계가 하나의 직선의 형태로 설명 될 수 있는 관계를 지닌다는 것

●선형관계의 중심

·예측치와 관측치의 차이의 제곱$(Y_i - Y)$^2의 합이

최소가 되는 직선(최소제곱법)

●개념

·독립변수와 종속변수 간의 1차 선형적 관계를 도출하여

독립 변수가 종속변수에 미치는 영향, 혹은 예측 정도를 분석하는 방법

·변수는 모두 연속형 자료이어야 함(더미변수 제외)

●결과 적용

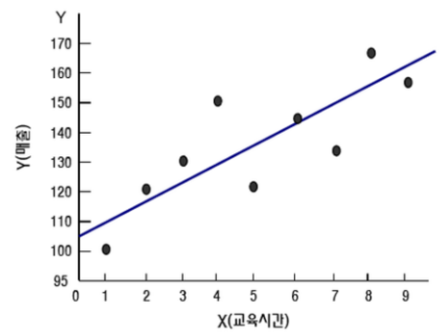
·독립변수가 종속변수에 영향을 미치는가?

·어느 정도의 영향을 미치는가?(영향력의 크기)

·어느 독립변수가 가장 큰 영향력을 미치는가?(영향의 상대적 크기/중요도)

·독립변수가 1증가할 때, 종속변수는 얼마나 증가할 것인가?(예측)

• 예) 종업원 교육시간이 매출성장에 미치는 영향



매출=기울기×교육시간+상수

$y = \beta \times x + c$

·독립변수 : 교육시간 → 종속변수 : 매출액

영향을 받는가 관계설명 or 매출액 향상에 기여할 수 있는가 예측

▫ 회귀분석과정

·두 변수가 선형의 관계를 가지는가를 알아보기 위해 산점도를 작성

·최소자승법으로 최적의 직선식을 구함

·'선형관계가 없다'는 귀무가설을

기각할 것인가를 결정하기 위하여 분산분석

·'기울기가 0이다'는 귀무가설을 기각할 것인지

각 독립변수에 대해 t/z검정

·이상의 분석에 기초하여 의사결정 진행

▫ 회귀분석의 기본구조

회귀식

$y = \beta x + c$ ······ 단일회귀분석

$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + c$ ······ 다중회귀분석

01-1 단순선형회귀 분석

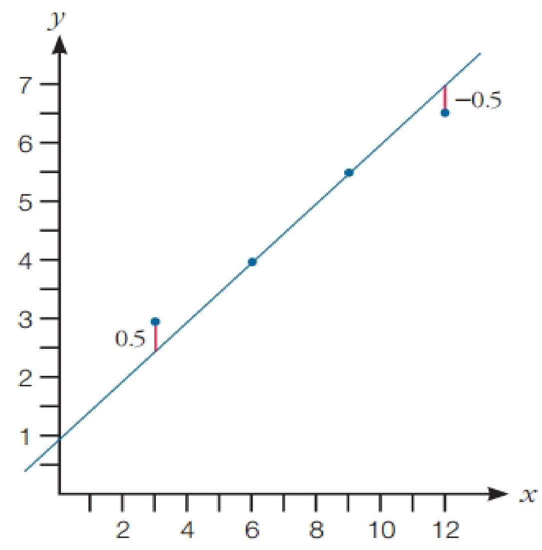
- **회귀직선(Regression Line)**
  - 2종류의 양적 데이터 X와 Y
    - X: 설명변수(explanatory variable), 독립변수
    - Y: 반응변수(response variable), 종속변수
  - 2개의 변수에 직선관계가 예상되는 경우에, 이에 근사하는 직선을 회귀직선 이라고 부름
  - 식: $\hat{y} = b_0 + b_1 x$

    $\hat{y}$: Y의 예측값

    b_0: 회귀직선의 절편

    b_1: 회귀직선의 기울기

X = {3.0, 6.0, 9.0, 12.0}, Y = {3.0, 4.0, 5.5, 6.5}

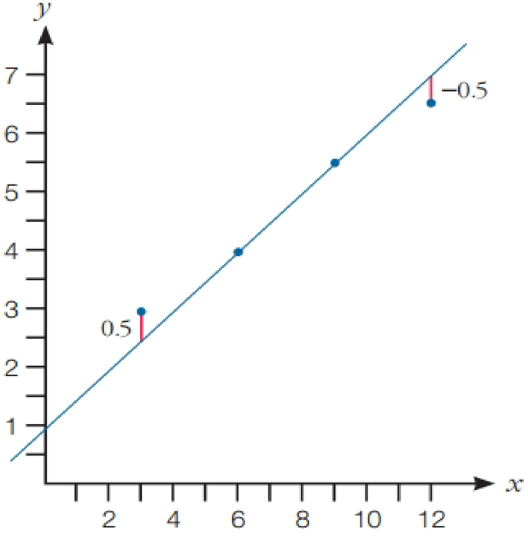


01-2 회귀직선의 오차분석

X = {3.0, 6.0, 9.0, 12.0}, Y = {3.0, 4.0, 5.5, 6.5}

$y=0.5x+1.0$의 오차 분석

x_i	3.0	6.0	9.0	12.0
예측값 $f(x_i)$	2.5	4.0	5.5	7.0
그라운드 트루스 y_i	3.0	4.0	5.5	6.5
오차	0.5	0.0	0.0	−0.5



평균 제곱 오차(MSE$_{\text{Mean squared error}}$)

$$E=\frac{1}{4}((0.5)^2+(0.0)^2+(0.0)^2+(-0.5)^2)=0.125$$

MSE: 그 값이 작을수록 오차가 적다

$$E=\frac{1}{n}\sum_{i=1}^{n}(y_i-f(x_i))^2$$

01-3-1 최소 제곱법: 회귀직선의 계수 구하기

- 선형 회귀에서는 최적화 문제를 풀어야 함
  - 최적화는 미분을 이용하여 해결함 (최소 제곱법)
  - R은 이 문제를 푸는 `lm` (linear model) 함수를 제공함

Derivation of linear regression equations

The mathematical problem is straightforward:

given a set of n points (X_i,Y_i) on a scatterplot,
find the best-fit line, $\boxed{\hat{Y}_i = a + bX_i}$
such that the sum of squared errors in Y, $\sum\left(Y_i - \hat{Y}_i\right)^2$ is minimized

The derivation proceeds as follows: for convenience, name the sum of squares "Q",

$$Q = \sum_{i=1}^{n}\left(Y_i - \hat{Y}\right)^2 = \sum_{i=1}^{n}\left(Y_i - a - bX_i\right)^2 \tag{1}$$

Then, Q will be minimized at the values of a and b for which $\boxed{\partial Q / \partial a = 0 \text{ and } \partial Q / \partial b = 0}$. The first of these conditions is,

$$\frac{\partial Q}{\partial a} = \sum_{i=1}^{n} -2\left(Y_i - a - bX_i\right) = 2\left(na + b\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} Y_i\right) = 0 \tag{2}$$

which, if we divide through by 2 and solve for a, becomes simply,

$$\boxed{a = \overline{Y} - b\overline{X}} \tag{3}$$

INFINITYBOOKS 인피니티북스

$$\frac{\partial Q}{\partial b} = \sum_{i=1}^{n} -2X_i(Y_i - a - bX_i) = \sum_{i=1}^{n} -2\left(X_iY_i - aX_i - bX_i^2\right) = 0 \qquad (4)$$

$$\boxed{a = \overline{Y} - b\overline{X}}$$

If we substitute the expression for a from (3) into (4), then we get,

$$\sum_{i=1}^{n}\left(X_iY_i - X_i\overline{Y} + bX_i\overline{X} - bX_i^2\right) = 0 \qquad (5)$$

We can separate this into two sums,

$$\sum_{i=1}^{n}\left(X_iY_i - X_i\overline{Y}\right) - b\sum_{i=1}^{n}\left(X_i^2 - X_i\overline{X}\right) = 0 \qquad (6)$$

$$\boxed{b = \frac{\sum_{i=1}^{n}\left(X_iY_i - X_i\overline{Y}\right)}{\sum_{i=1}^{n}\left(X_i^2 - X_i\overline{X}\right)} = \frac{\sum_{i=1}^{n}\left(X_iY_i\right) - n\overline{X}\,\overline{Y}}{\sum_{i=1}^{n}\left(X_i^2\right) - n\overline{X}^2}} \qquad (7)$$

We can translate (7) into a more intuitively obvious form, by noting that

$$\sum_{i=1}^{n}\left(\overline{X}^2 - X_i\overline{X}\right) = 0 \quad \text{and} \quad \sum_{i=1}^{n}\left(\overline{X}\,\overline{Y} - Y_i\overline{X}\right) = 0 \qquad (8)$$

so that b can be rewritten as the ratio of Cov(x,y) to Var(x):

$$b = \frac{\sum_{i=1}^{n}\left(X_iY_i - X_i\overline{Y}\right) + \sum_{i=1}^{n}\left(\overline{X}\,\overline{Y} - Y_i\overline{X}\right)}{\sum_{i=1}^{n}\left(X_i^2 - X_i\overline{X}\right) + \sum_{i=1}^{n}\left(\overline{X}^2 - X_i\overline{X}\right)} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = \boxed{\frac{Cov(X,Y)}{Var(X)}} \qquad (9)$$

01-3-2 최소제곱법: 회귀직선의 계수

- 절편

$$b_0 = \bar{y} - b_1\bar{x}$$

  - 회귀직선이 각 변수의 평균값을 좌표로 하는 점을 반드시 통과한다는 점을 의미함.

- 기울기

$$b_1 = \frac{[X\text{와 } Y\text{의 공분산}]}{[X\text{의 표준편차}]^2} = \frac{s_{XY}}{s_X{}^2}$$

$$= [X\text{와 } Y\text{의 상관계수}] \times \frac{[Y\text{의 표준편차}]}{[X\text{의 표준편차}]} = r_{XY}\frac{s_Y}{s_X}$$

$$b = \frac{\sum_{i=1}^{n}\left(X_iY_i - X_i\overline{Y}\right) + \sum_{i=1}^{n}\left(\overline{X}\,\overline{Y} - Y_i\overline{X}\right)}{\sum_{i=1}^{n}\left(X_i^2 - X_i\overline{X}\right) + \sum_{i=1}^{n}\left(\overline{X}^2 - X_i\overline{X}\right)} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2} = \frac{Cov(X,Y)}{Var(X)}$$

$$\text{상관계수 } r_{XY} = \frac{[X\text{와 } Y\text{의 공분산}]}{[X\text{의 표준편차}] \times [Y\text{의 표준편차}]} = \frac{s_{XY}}{s_X s_Y}$$

01-3-3 단순선형회귀분석의 예

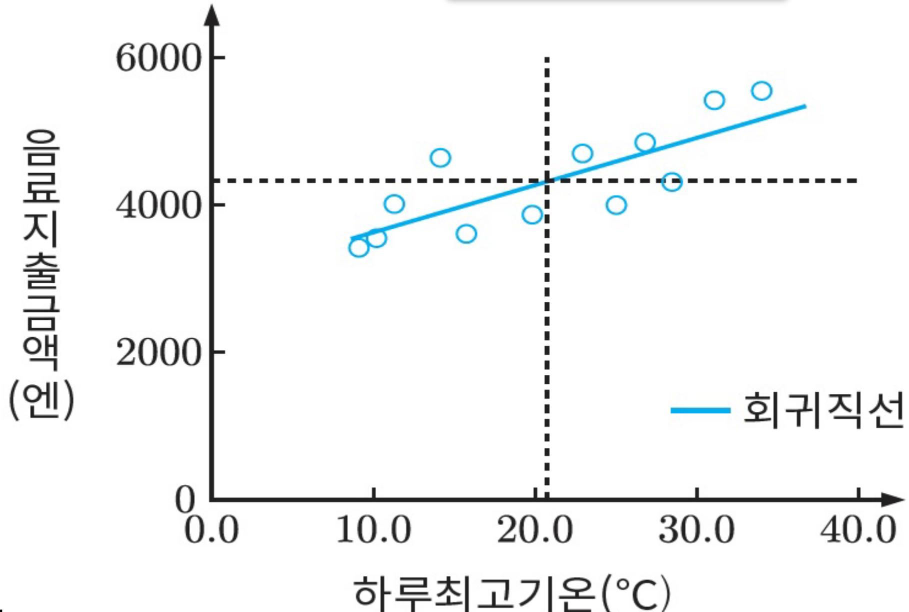
- 하루최고기온으로부터 음료지불금액을 예측

월	1	2	3	4	5	6
하루최고기온(℃)	9.1	10.2	14.1	19.8	25.0	26.8
음료지출금액(엔)	3416	3549	4639	3857	3989	4837
월	7	8	9	10	11	12
하루최고기온(℃)	31.1	34.0	28.5	22.9	15.7	11.3
음료지출금액(엔)	5419	5548	4311	4692	3607	4002

$$\boxed{b_0 = \bar{y} - b_1 \bar{x}}$$

- 최소제곱법을 사용해 회귀직선을 구함

$$\boxed{\hat{y} = 2947.8 + 66.4 \times x}$$



하루최고기온: 9.1~34.0˚C

하루최고기온이 10˚C인 경우 음료지불금액

: $2947.8 + 66.4 \times 10 = 3611.8$

기온이 1˚C 상승하면 평균적으로 음료지불금액이 66.4엔씩 증가

회귀직선은 각 변수의 평균값을 좌표로 갖는 점(20.7, 4322.2)를 통과

01-3-4 반응변수의 변동

- **회귀직선이 데이터를 적절하게 모델링(fitting)하고 있는가를 평가하기 위한 반응변수의 변동**

- **데이터의 변동:** $S_y^2 = (y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2$

  데이터와 평균값의 차에 대한 제곱합

- **예측값의 변동:** $S_{\hat{y}}^2 = (\hat{y}_1 - \bar{y})^2 + \cdots + (\hat{y}_n - \bar{y})^2$

  예측값과 평균값의 차에 대한 제곱합

- **잔차의 변동:** $S_e^2 = (y_1 - \hat{y}_1)^2 + \cdots + (y_n - \hat{y}_n)^2$

  데이터와 예측값의 차에 대한 제곱합

01-3-5 결정계수

- **결정계수(coefficient of determination)**
  - 회귀직선의 적합성 측정계수
  - 결정계수의 계산

$$R^2 = \frac{S_{\hat{y}}^2}{S_y^2} = 1 - \frac{S_e^2}{S_y^2}$$

$$S_y^2 = (y_1 - \bar{y})^2 + \cdots + (y_n - \bar{y})^2$$

데이터와 평균값의 차에 대한 제곱합

$$S_{\hat{y}}^2 = (\hat{y}_1 - \bar{y})^2 + \cdots + (\hat{y}_n - \bar{y})^2$$

예측값과 평균값의 차에 대한 제곱합

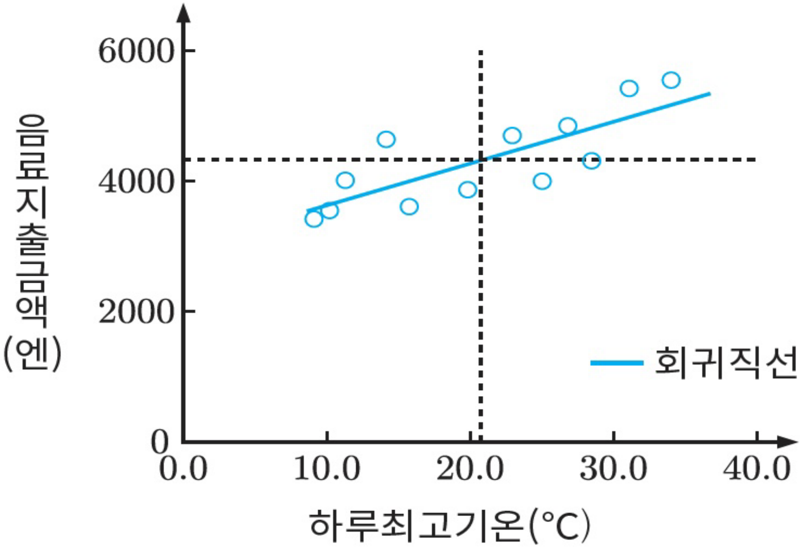
$$S_e^2 = (y_1 - \hat{y}_1)^2 + \cdots + (y_n - \hat{y}_n)^2$$

데이터와 예측값의 차에 대한 제곱합

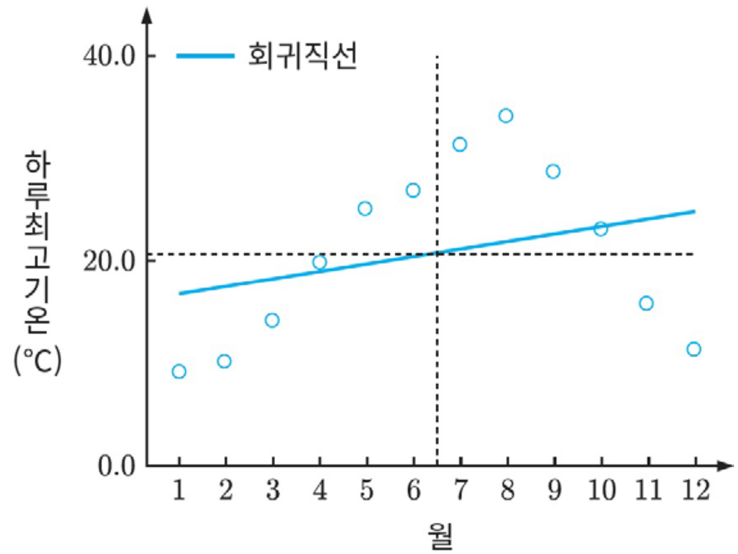
  - 잔차와 결정계수
    - 잔차의 변동이 0에 가까워지면
      - 결정계수는 **1**에 가까워짐
      - "실제 데이터에 대한 회귀직선의 **적합성이 좋다**"
      - 회귀직선의 유용성이 높아짐
    - 잔차의 변동이 커지게 되면
      - 결정계수는 **0**에 가까워짐
      - "실제 데이터에 대한 회귀직선의 **적합성이 나쁘다**"
      - 회귀직선의 유용성이 낮아짐

01-3-6 결정계수의 예

- **하루최고기온으로부터 음료지출을 예측**
  - 상관계수 0.8
  - 결정계수 0.64
  - 회귀직선의 실제 데이터에 대한 적합성은 나쁘지 않음

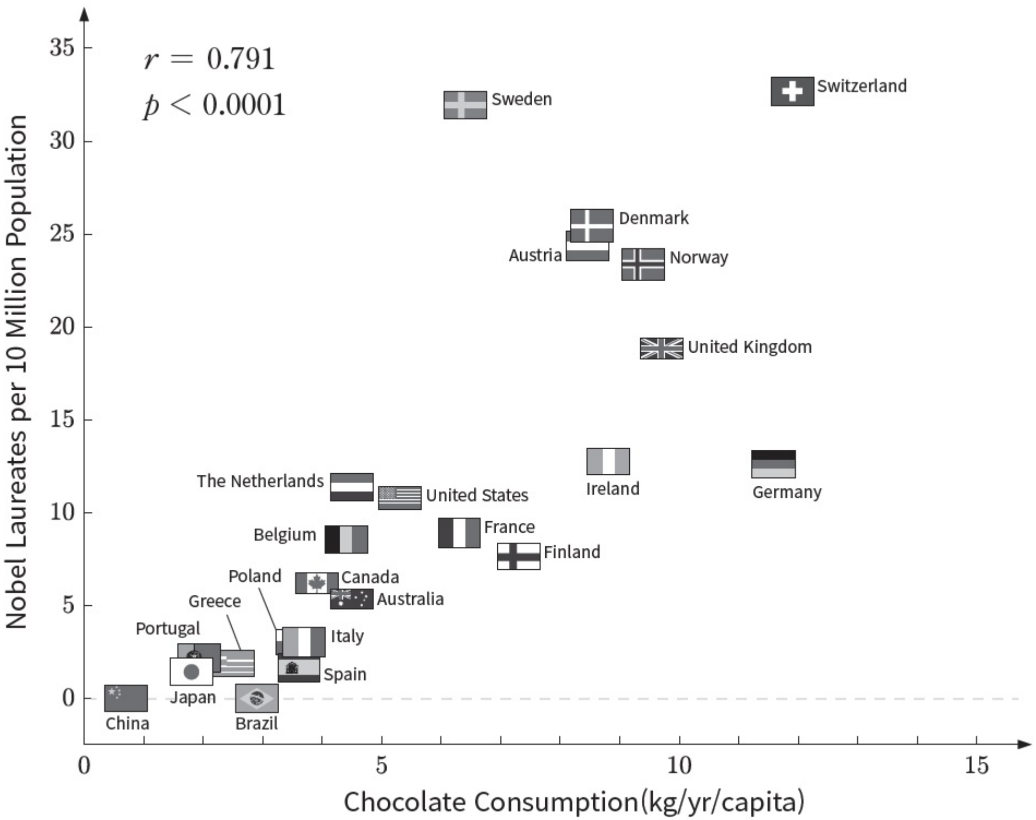


- **오오쯔시의 매달 하루 최고기온을 월단위로 예측**
  - 상관계수 0.309
  - 결정계수 0.10
  - 회귀직선의 적합성은 나쁨



01-3-6 결정계수의 예 (계속)

- **초콜릿 소비량으로부터 노벨상 수상자수를 예측**
  - 상관관계 0.791
  - 결정계수 0.63
  - 회귀직선의 적합성은 나쁘지 않으나...
  - 허위적 관계



〈그림 2.17〉 초콜릿소비량과 노벨상수상자 수의 관계(출전: Messeili(2012))

02 데이터 분석에서 주의 사항

- **올바른 데이터 분석**
  - 적절한 데이터 수집
  - 분석 결과의 올바른 해석

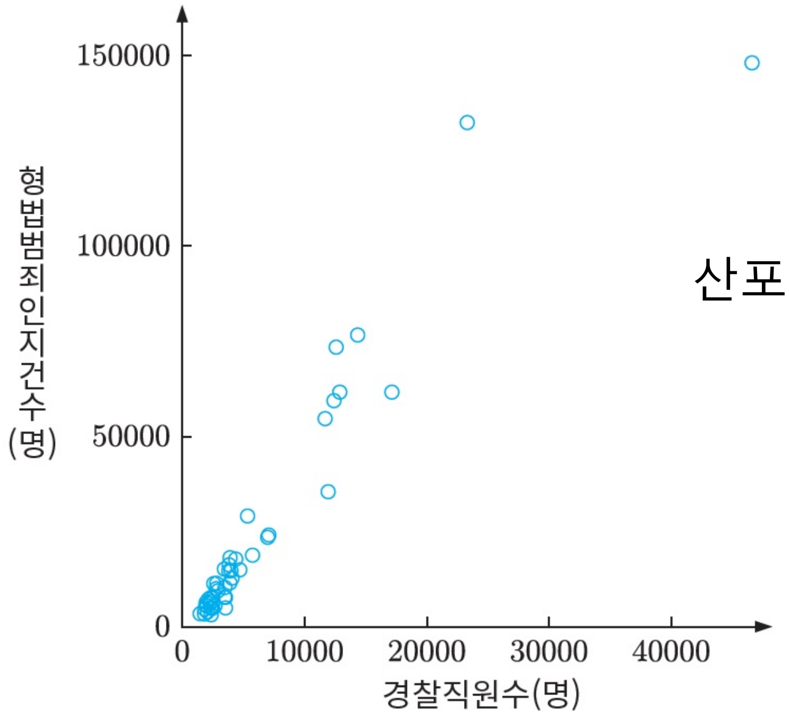
- **상관관계와 인과관계**
- **데이터 수집 방법**
- **적절한 그래프 사용법**

02-1 상관관계와 인과관계

- **상관관계: 산포도, 상관계수**
- **인과관계**
  - 한쪽 변수가 또다른 쪽의 변수의 원인이 됨
  - 원인이 되는 변수를 조정함으로써 다른 쪽의 변수를 어느정도 조작하는 것이 가능
- **2개 변수들 사이에 상관관계가 존재하더라도 인과관계가 존재한다고는 말할 수 없음**

02-2 상관관계와 인과관계: 예

- **경찰직원수와 형법범죄 인지건 수**



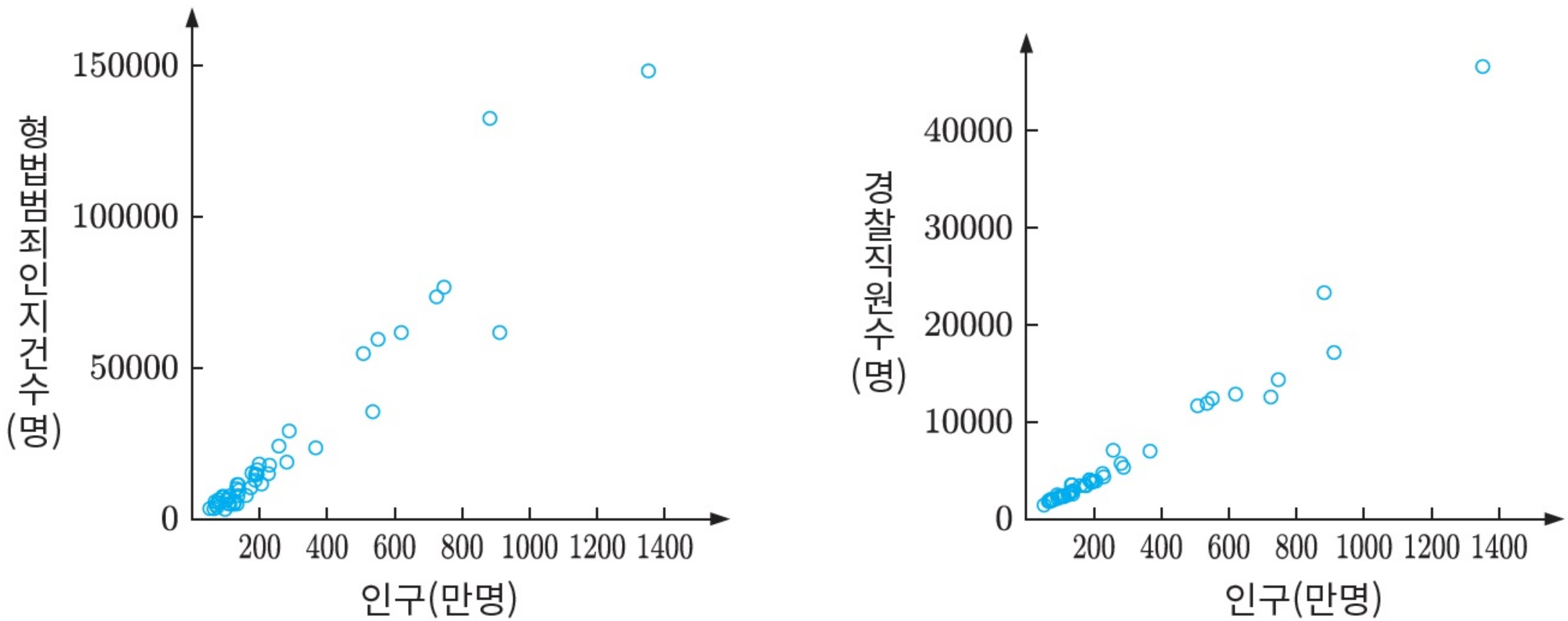
〈그림 2.19〉 2015년 지방자치단체별 경찰직원 수와 형법범죄 인지건 수의 산포도

산포도에 대한 상관계수는 0.95

- 경찰직원이 많아질 수록 형법 범죄가 증가하는가?
- 형법범죄가 많아질 수록 경찰직원이 늘어나는가?

02-3 허위상관관계

- **허위상관관계(Spurious correlation):** 살펴보려는 2개의 변수 각각과 강한 상관관계를 갖는 또다른 변수가 존재하는 경우, 원래의 2개 변수들의 상관이 강하게 되어버리는 현상
- **제3변수: 허위상관관계의 원인이 되는 변수**
- **잠재변수(latent variable): 수집되어 있지 않은(또는 입수할 수 없는) 제3변수**



〈그림 2.20〉 2015년 지자체별 인구와 형법 범죄 인지 건수의 산포도(왼쪽), 인구와 경찰직원수의 산포도(오른쪽)

INFINITYBOOKS 인피니티북스

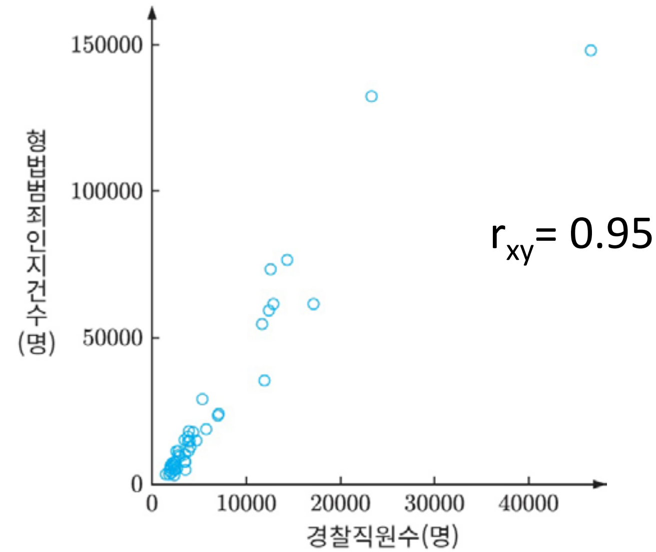
02-4 제3변수의 영향을 제거하는 방법

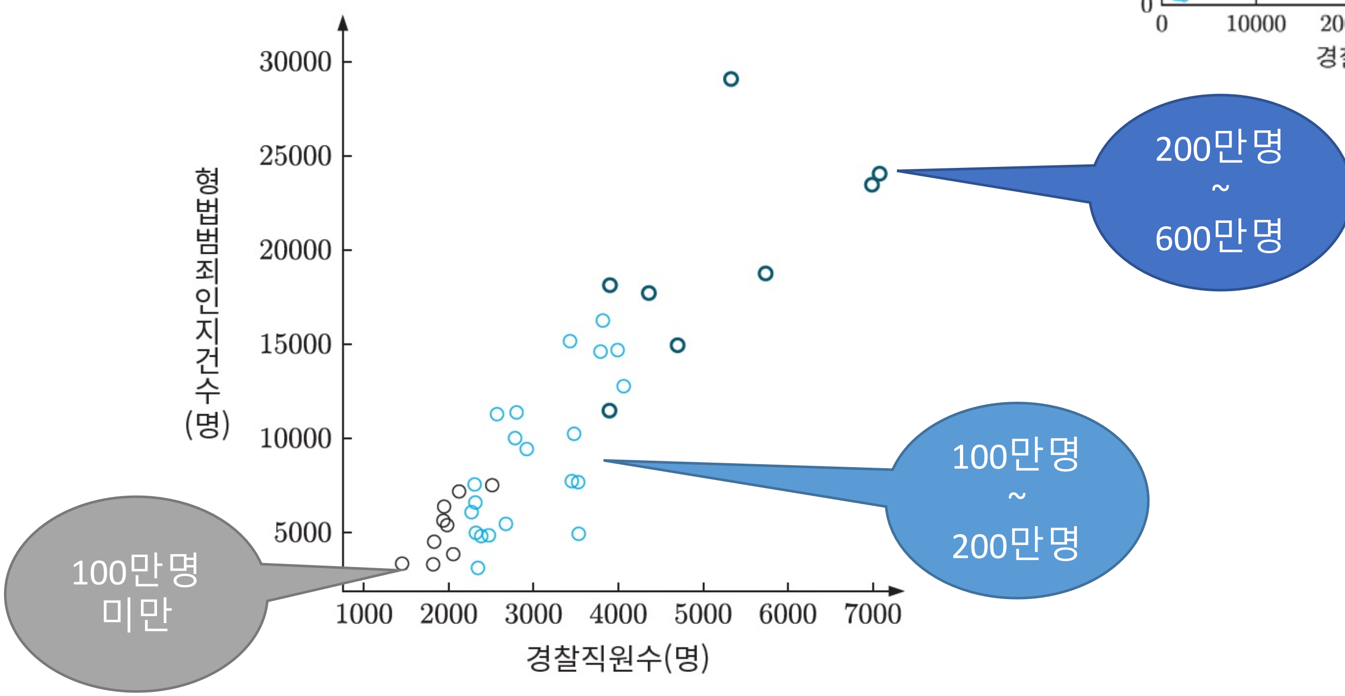
- **제3변수에 의한 계층화**
- **각 변수들을 제3변수의 단위량으로 변환**
- **편상관계수(Partial Coefficient of Correlation)를 계산**

- **어느 것이 좋은지는 상황에 따라 다름**
- **제3변수가 확보되어야 가능**
  - 데이터를 수집하는 단계에서 잠재변수를 간과하지 않도록 주의

02-4-1 제3변수에 의한 계층화

- **인구 600만명 이하의 지자체로 한정시켜서 계층화**
  - 100만명 미만 (상관계수 0.77)
  - **100만명 이상 200만명 미만 (상관계수 0.70)**
  - 200만명 이상 600만명 미만 (상관계수 0.91)

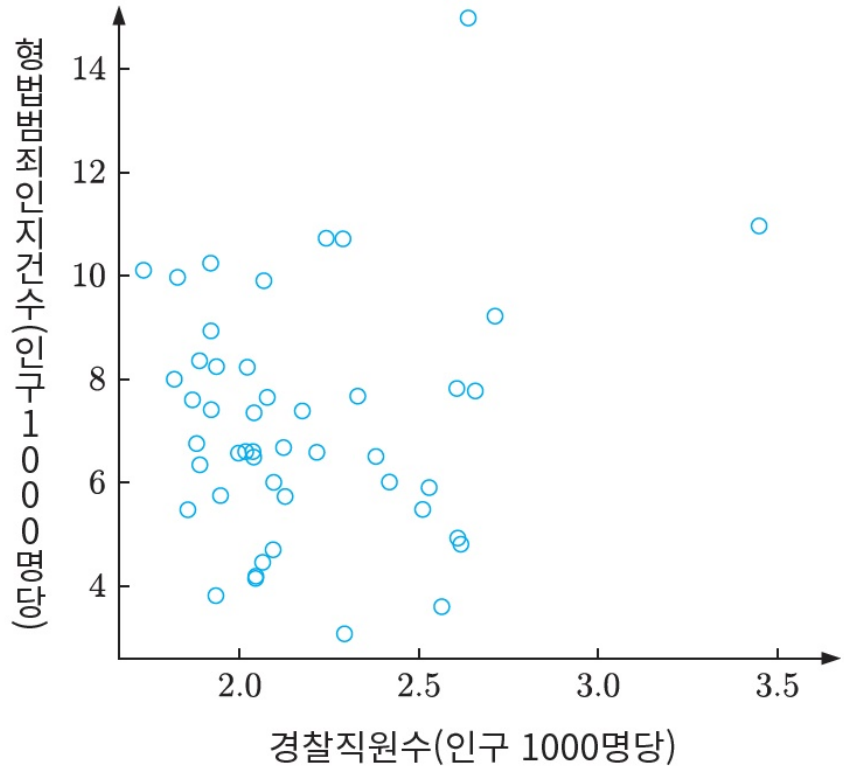




〈그림 2.21〉 2015년 지차체별 경찰직원수와 형법 범죄 인지건수에 대한 계층화된 산포도

02-4-2 각 변수들을 제3변수의 단위량으로 변환

- **인구 1000명당 경찰직원수와 형법 범죄 인지건수에 대한 산포도**
  - 상관계수 0.12
  - 인구의 영향을 제거하면 상관이 거의 없어짐



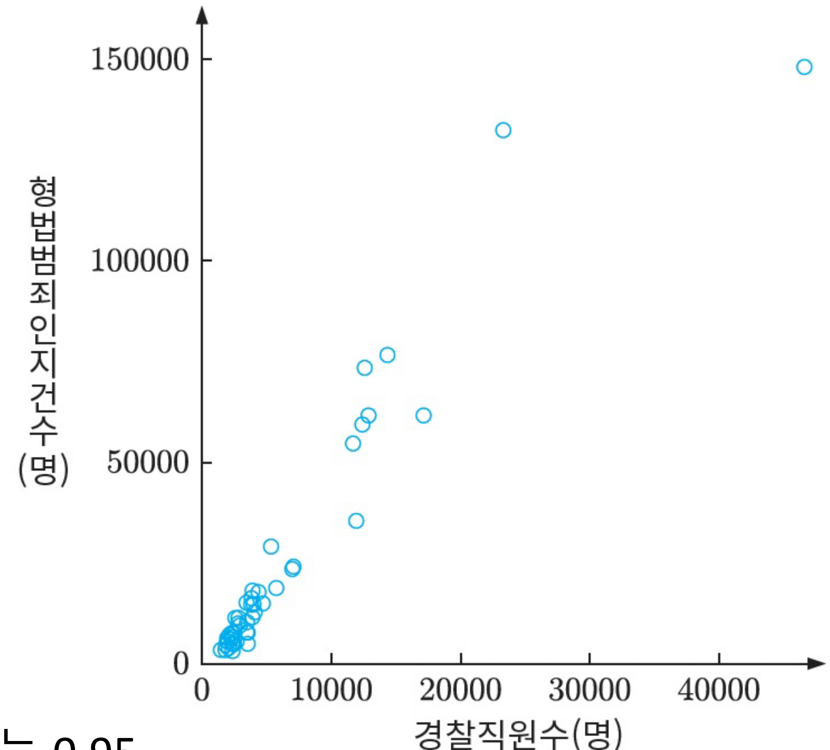
〈그림 2.22〉 2015년 지자체별 인구 1000명당 경찰직원수와 형법 범죄 인지건수에 대한 산포도

02-4-3 편상관계수를 이용

- **편상관계수:**
  - 관계를 조사하고자 하는 2개의 변수들에 대해서, 다른 변수의 영향을 제거한 상관계수
  - 회귀직선에 관한 개념을 이용: Z의 영향을 제거한 X와 Y의 편상관계수

$$\frac{r_{XY} - r_{XZ} r_{YZ}}{\sqrt{(1 - r_{XZ}^{2})(1 - r_{YZ}^{2})}}$$

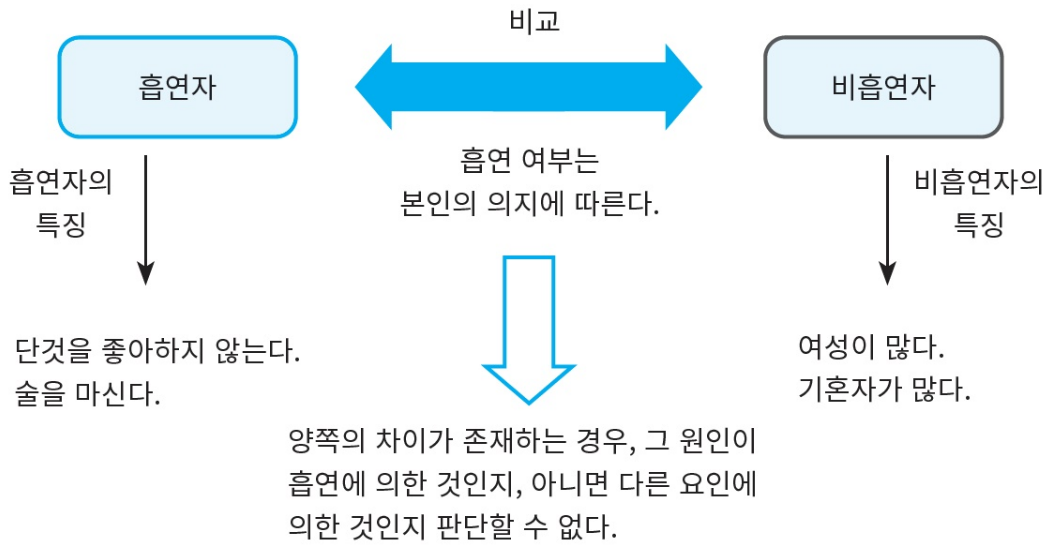
  - 인구의 영향을 제거한 경찰직원수와 형법범죄인지건수의 편상관계수는 0.37



산포도에 대한 상관계수는 0.95

02-5 사건의 영향조사: 관찰연구와 실험연구

- **관찰연구**
  - 어떤 사건을 수행하는지 여부를 본인이 결정할 수 있는 상황에서, 그 사건의 결과를 비교하는 연구
  - 예: 담배를 피우면 폐암발생률이 증가하는지 여부를 조사하기 위해서 흡연자와 비흡연자 사이의 폐암발생률을 조사. 이때 흡연여부는 각자의 의지에 따름.
  - 조사하려는 사건 이외의 조건들을 가능한 한 살펴볼 필요가 있음

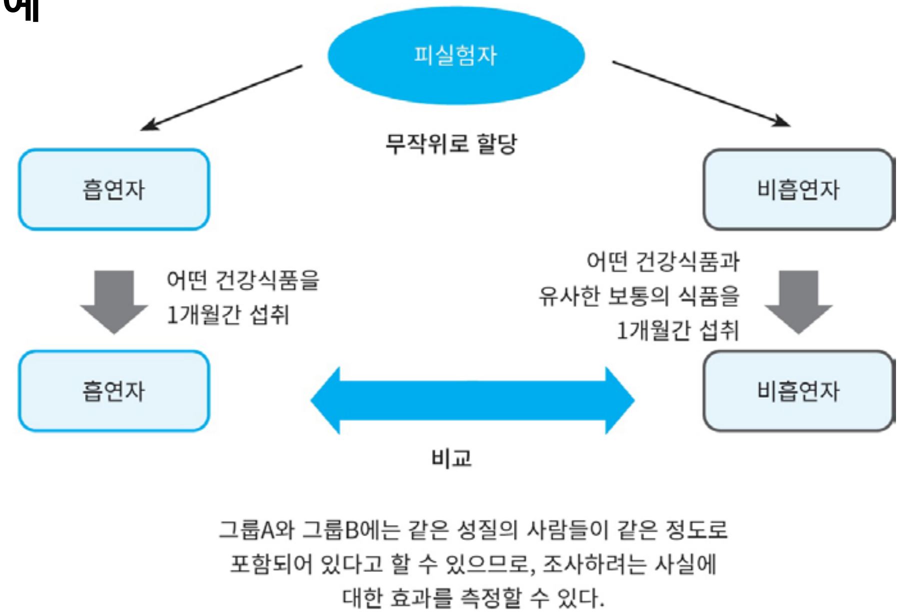


〈그림 2.23〉 관찰연구의 예

02-5 사건의 영향조사: 관찰연구와 실험연구 (계속)

● 실험연구

- **어떤 사건의 영향을 조사할 때, 그 사건을 적용할 것인가 여부를 연구자가 할당한 상태에서 서로 간의 차이점에 대해서 조사**
- **피실험자를 무작위로 할당**
  - 다양한 피실험자가 존재하더라도 유사한 성질을 갖는 피실험자가 각 그룹에 같은 정도로 포함될 것으로 기대
- **예: 건강식품의 효과를 조사하기 위한 실험연구의 예**
  - 그룹B에 소속된 사람들이게 "건강식품과 유사한 제품을 먹도록" 함
  - 피실험자가 어떤 그룹에 소속되어 있는지를 알 수 없도록 하는 것이 중요



〈그림 2.24〉 실험연구의 예

02-6 데이터 수집 방법: 표본조사

- **용어**
  - 모집단(population 또는 universe): 조사 대상의 전체 집합
  - 표본(sample): 모집단으로부터 조사를 위해서 추출한 대상들의 집합
  - 샘플 사이즈(sample size) 또는 표본 크기: 표본을 구성하는 대상의 개수

- **TV 시청률 조사**
  - 모집단: TV를 보유하고 있는 모든 세대
  - 표본: 시청률을 조사하는 장치를 설치하고 있는 세대

- **정당의 지지율 조사**
  - 모집단: 유권자 전체
  - 표본: 전화조사를 수행한 대상자 전원

02-6-1 표본 추출의 필요성과 방법

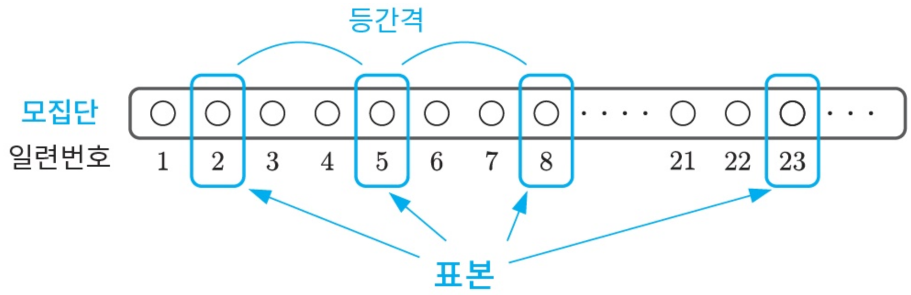
- **모집단 전체에 대한 조사의 어려움**
  - 비용 및 시간 문제
  - 물리적 문제

- **단순 무작위 추출(simple random sampling)**
  - 모집단이 방대한 경우 비용이 듦

- **조사 비용을 줄일 수 있는 표본 추출 방법**
  - 계통추출법
  - 클러스터 추출법
  - 층화 추출법
  - 다단 추출법

02-6-2 표본추출: 계통추출법

- **계통추출법(systematic sampling)**
  - 모집단의 대상 전체에 일련번호를 부여
  - 적당한 대상부터 같은 간격으로 표본을 추출
  - 일련번호가 무작위로 부여되면 표본조사 비용은 그다지 줄어들지 않음



〈그림 2.25〉 계통추출법

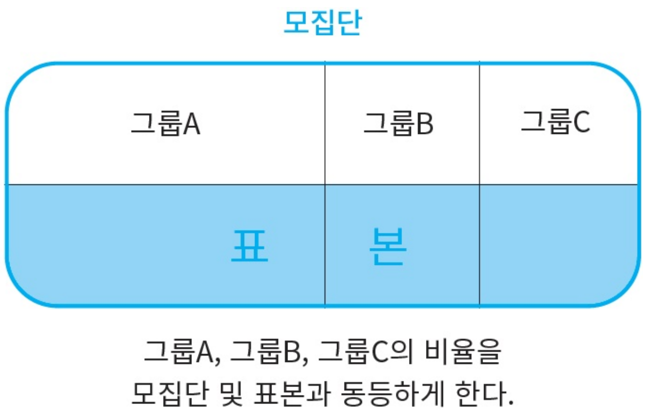
02-6-3 표본추출: 클러스터 추출법

- **클러스터 추출법(cluster sampling) 또는 군집표본추출법**
  - 모집단을 몇 개의 그룹으로 분할
  - 무작위로 추출한 1개 또는 여러 개의 그룹을 표본으로 선택
  - 특수한 치우침이 있는 그룹을 만들지 않도록 해야 함

〈그림 2.26〉 클러스터 추출법

02-6-4 표본추출: 층화 추출법

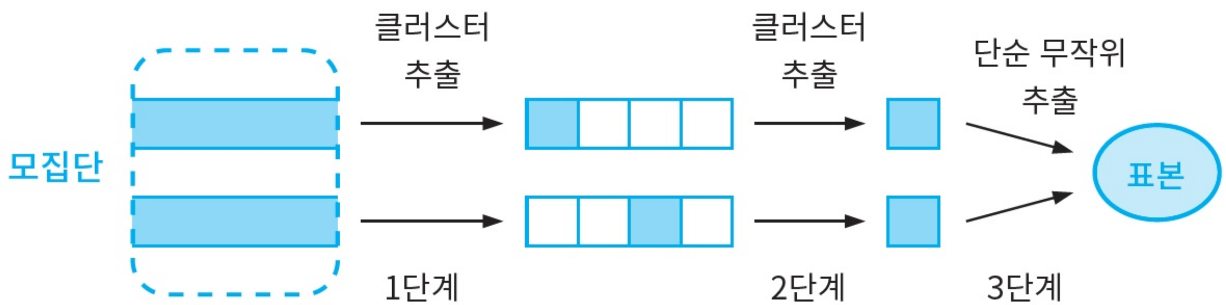
- **층화 추출법(stratified sampling)**
  - 모집단 속에서 유사한 성질을 갖는 그룹(계층)으로 나눔
  - 각 그룹에서 표본을 추출
  - 각 그룹에 비슷한 사람을 모이게 하는 것이 중요



〈그림 2.27〉 층화 추출법

02-6-5 표본추출: 다단 추출법

- **다단 추출법(multi-stage sampling)**
  - 클러스터 추출법을 반복수행
  - 마지막 단계에서 단순 무작위 추출법을 수행
  - 계층이 늘어날 수록 모집단과 표본의 차이가 커지기 쉽다는 점에 주의



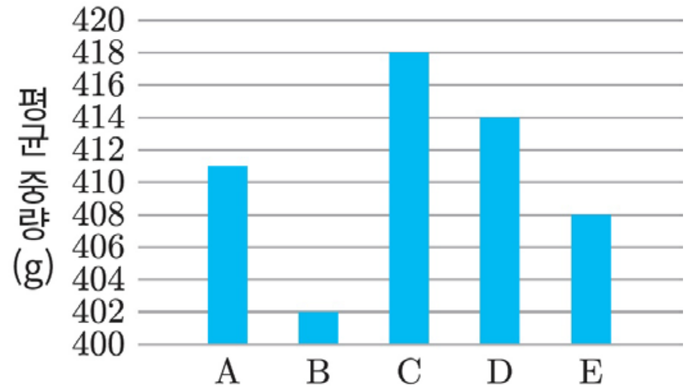
〈그림 2.28〉 다단추출법

02-7 적절한 그래프 사용법

- 막대그래프 (barplot)
- 히스토그램 (histogram)
- 꺾은선그래프 (line graph)
- 파이 그래프(pie graph)
- 띠 그래프(band graph)
- 누적 막대 그래프(stacked bar graph)
- 클러스터형 막대 그래프(clustered bar graph)
- 산포도(scatter graph)

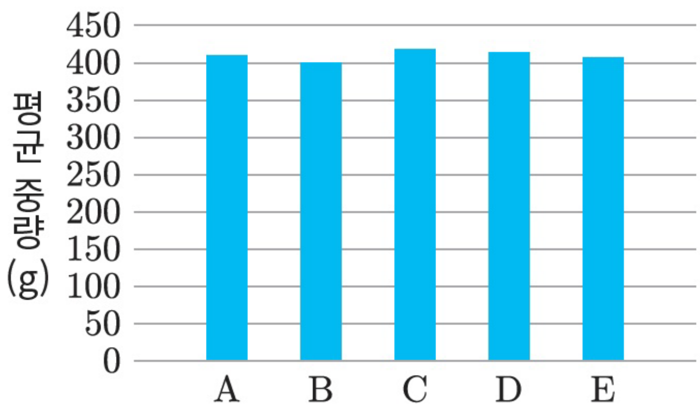
02-7-1 막대 그래프: 각 항목의 양을 비교하는 경우

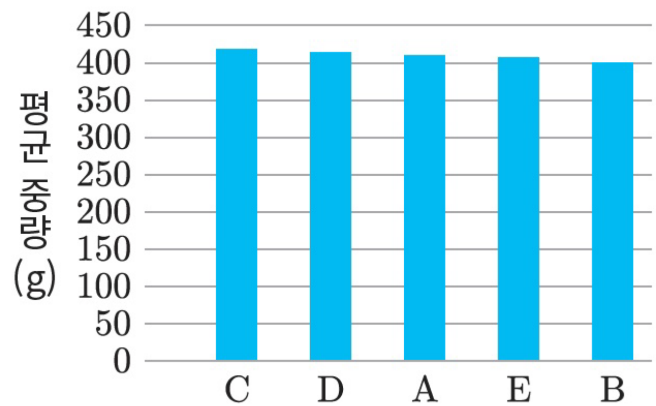
- **부적절한 막대 그래프의 예**



- **막대 그래프는 눈금을 0부터 시작하는 것이 중요**
  - 단, 품질관리 분야 등에서 어떤 기준량과의 차이를 나타내려는 경우는 기준량과의 차이에 대한 그래프를 작성

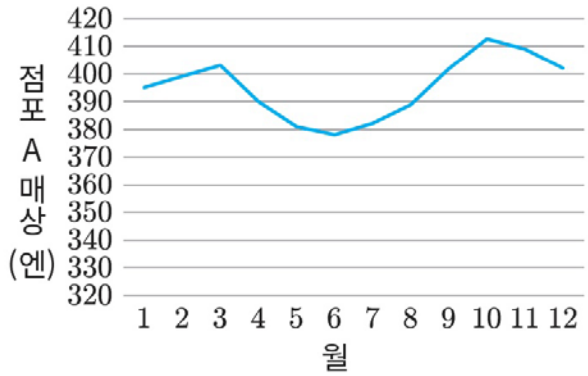
- **순서가 특별한 의미를 갖지 않는 경우에는 정렬하여 파악하기 수월하도록 표현할 수 있음**

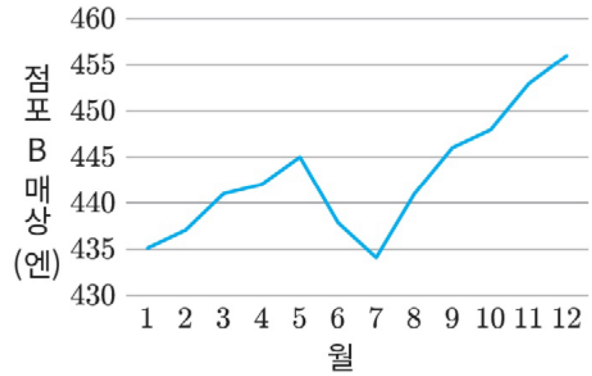




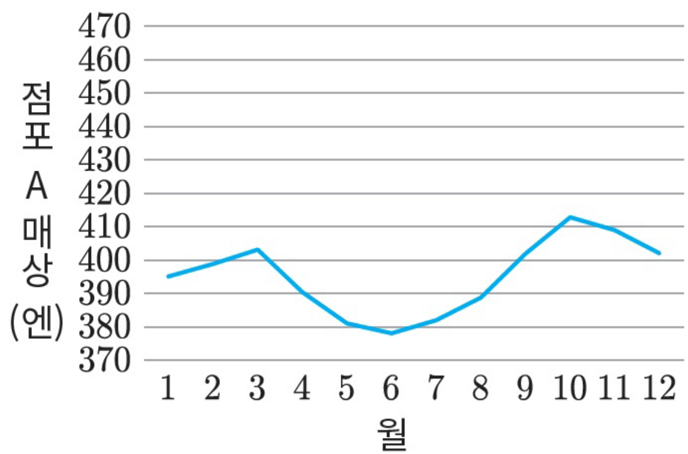
02-7-2 꺾은선 그래프: 데이터의 시간적 변화를 관측하는 경우

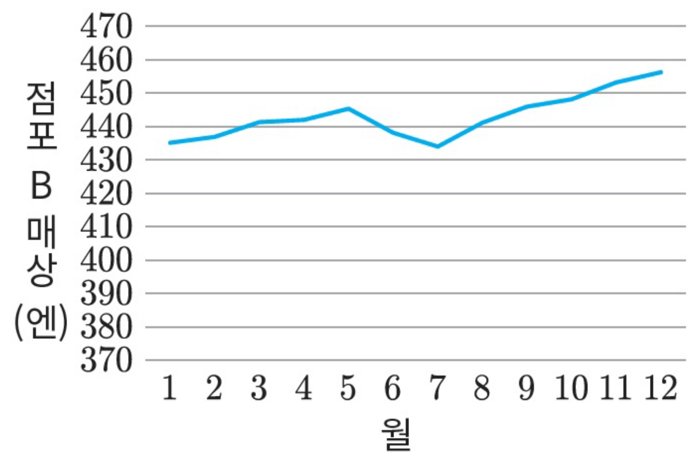
- 부적절한 꺾은선 그래프의 예





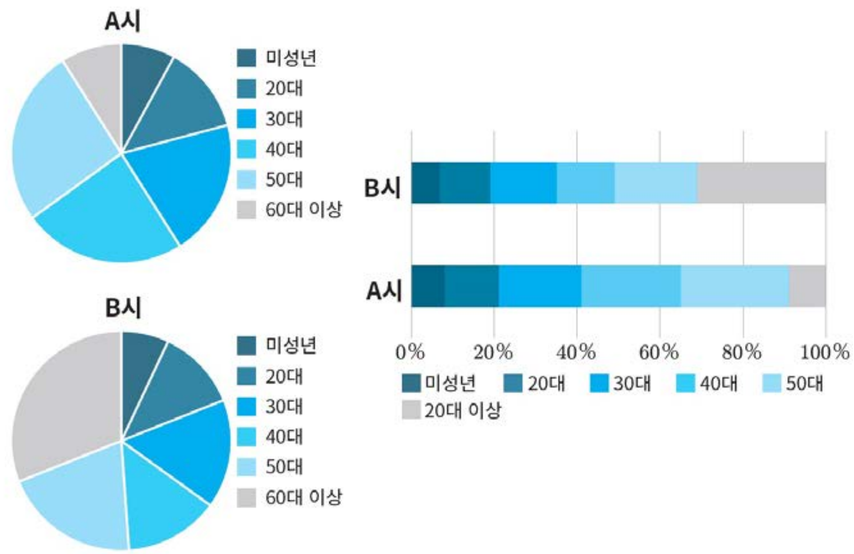
- 여러 그래프를 비교하려는 경우에는 눈금을 통일시키는 것이 좋음





02-7-3 파이/띠 그래프: 여러 데이터의 비율을 비교하는 경우

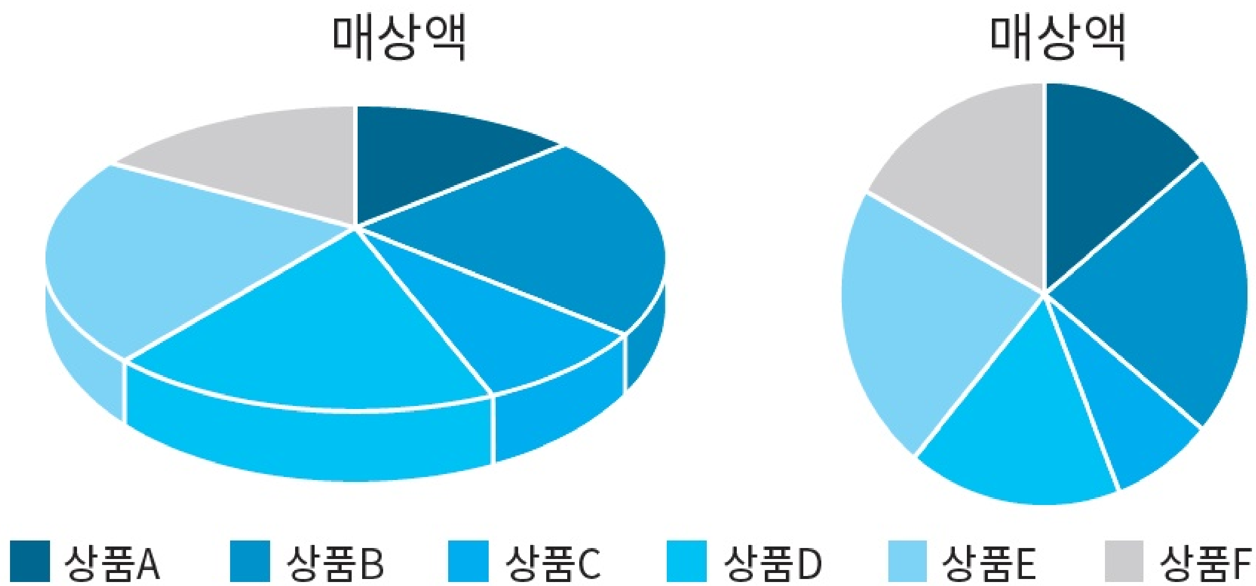
- **파이 그래프**
  - 어떤 데이터에 포함되는 비율을 파악하려는 경우에 적절
  - 2개 이상의 그래프에서 비율을 비교하기에는 부적절
- **띠 그래프**
  - 그래프들간의 비율을 비교하기 수월함



〈그림 2.34〉 파이 그래프와 띠 그래프

02-7-4 3D 파이 그래프

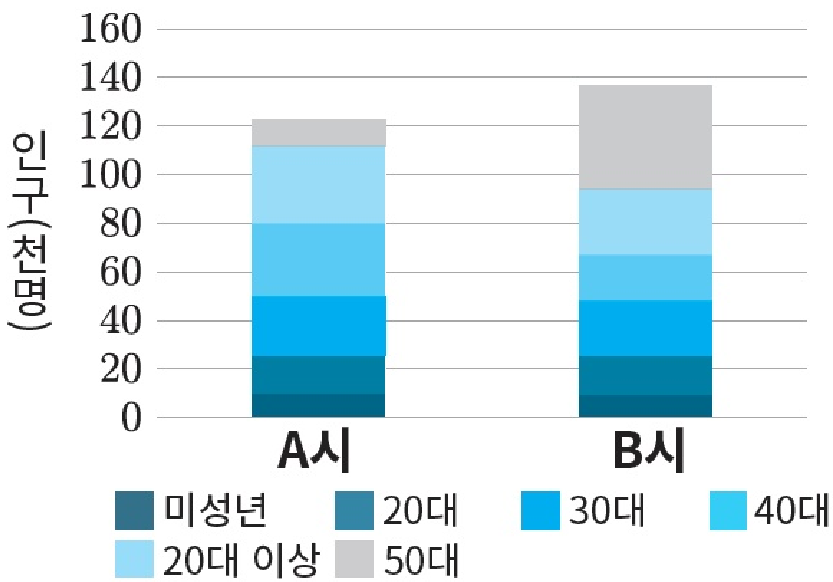
- 비율에 대한 수치가 기록되어 있지 않은 경우, 비율을 시각적으로 파악할 수 없기 때문에 착각을 유발할 수 있음

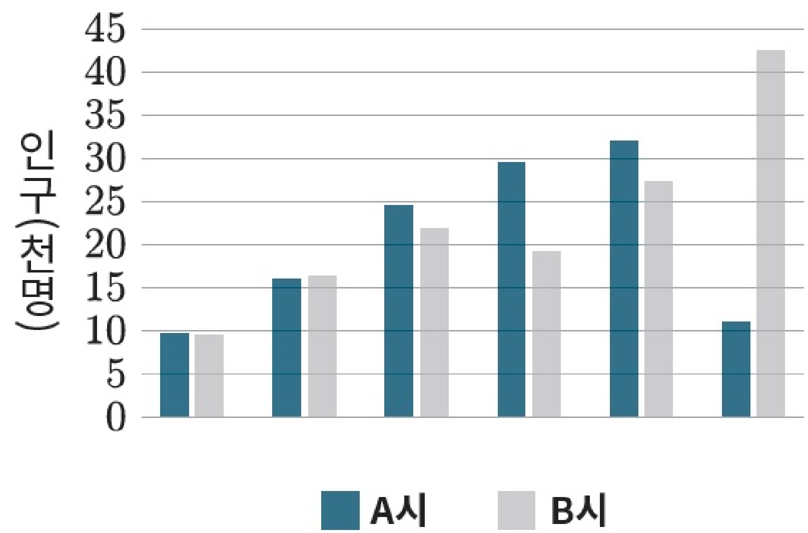


〈그림 2.36〉 3D 파이 그래프와 파이 그래프

02-7-5 누적/클러스터형 막대 그래프: 여러 데이터의 비율 및 총량 비교

- 비율과 양을 동시에 파악





〈그림 2.35〉 누적 막대 그래프와 클러스터형 막대 그래프

데이터 분석기법-기초(2)

- **회귀분석(Regression Analysis)**
  - 단순선형회귀분석(Simple Linear Regression Analysis)
  - 최소제곱법(Least Square Method, LSM)
  - 반응변수의 변동과 결정계수

- **데이터 분석에서 주의사항**
  - 상관관계와 인과관계
  - 관찰연구와 실험연구
  - 데이터 수집 방법: 표본조사
  - 적절한 그래프 사용법