범주형과 수치형 자료

보건빅데이터통계분석

이새봄
삼육대학교 SW융합교육원

자료란?

자료

- 삼육대학교에서 수업할 수 있는 능력을 평가하기 위한 자료?

  -> 수학능력시험

자료

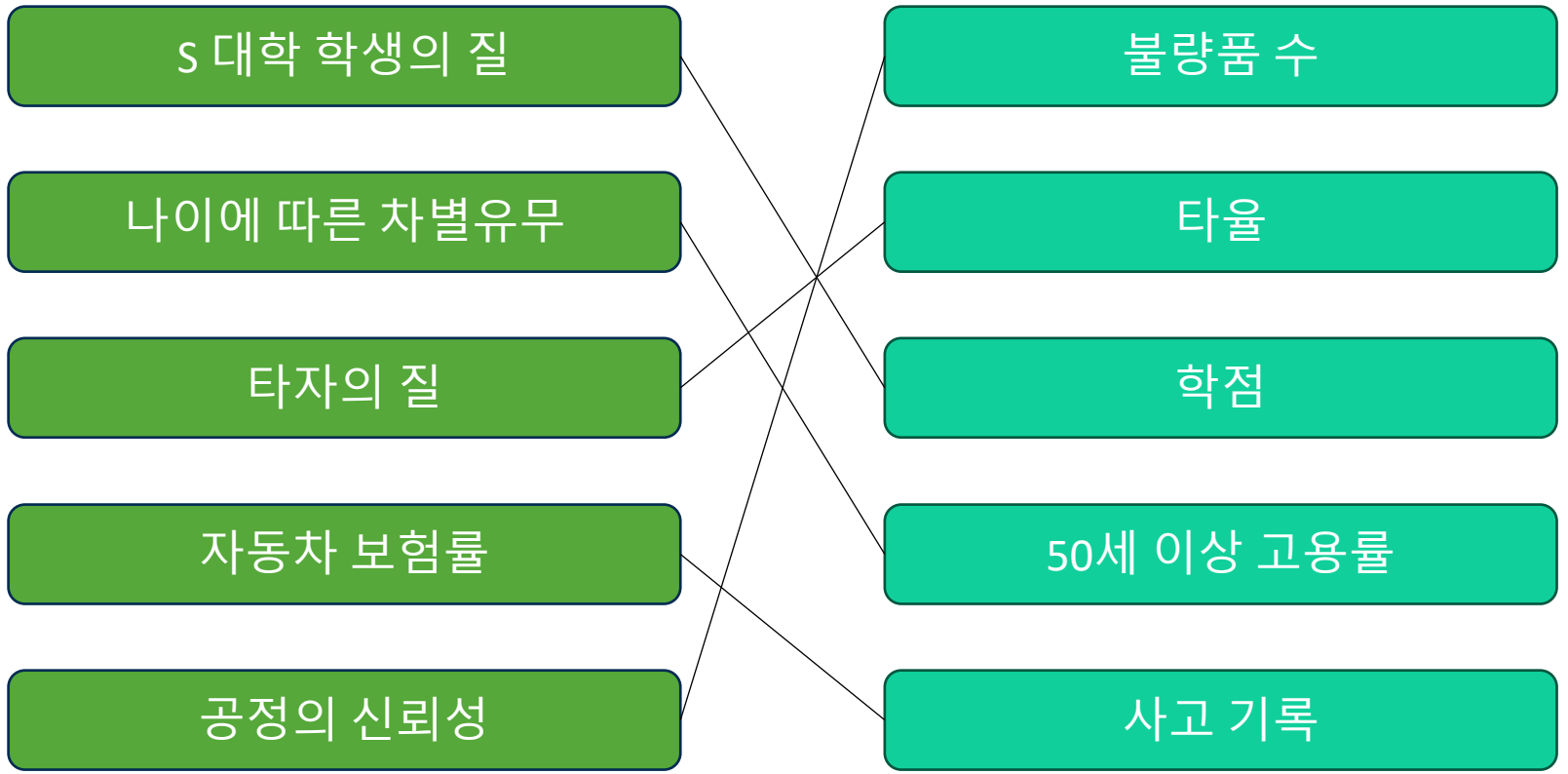
- 개체(Subject)
  - 관찰된 항목이나 대상, 예) 고등학교 3학년
- 요인(Factor)
  - 개체에 관한 특성 중 연구자가 관심을 갖는 특성, 예) 대학수학능력
- 변수(Variable)
  - 요인의 특성을 수치화 하기 위해 쓰이는 속성, 예) 대학수학능력시험점수
- 척도 (Measurement)
  - 일정한 규칙을 가지고 기호 또는 숫자로 나타낸 것 , 예) 점수, cm, kg 등
- 자료(Data)
  - 관심의 대상인 사물이나 속성을 측정, 관찰, 조사한 값들의 모음, 예) 각 학생들의 개별 수능점수

자료

- 관측대상(Subject)
  - 15개(1~15)
- 변수(Variable)
  - 7개(성별 ~기말)
- 다변량(multivariate)자료(여러 개의 변수로 구성)
- *일변량(Univariate)자료: 하나의 변수로만 구성*
- 관측치(Observations)
  - 15*7 = 105개

id	성별	분반	학년	몸무게	출석	중간	기말
1	남자	1	1	40	100	87	80
2	여자	2	2	50	100	83	60
3	남자	1	3	56	100	84	60
4	여자	2	4	51	100	73	60
5	남자	1	1	55	100	68	60
6	남자	2	2	61	100	77	50
7	여자	1	3	69	100	40	80
8	여자	2	2	44	100	73	30
9	여자	1	2	66	80	64	40
10	남자	2	2	60	100	66	40
11	여자	1	2	56	100	63	40
12	여자	2	4	72	100	76	20
13	여자	1	1	46	100	73	20
14	남자	2	2	63	100	73	20
15	남자	1	3	56	100	67	30

요인과 변수


S 대학 학생의 질
나이에 따른 차별유무
타자의 질
자동차 보험률
공정의 신뢰성
불량품 수
타율
학점
50세 이상 고용률
사고 기록

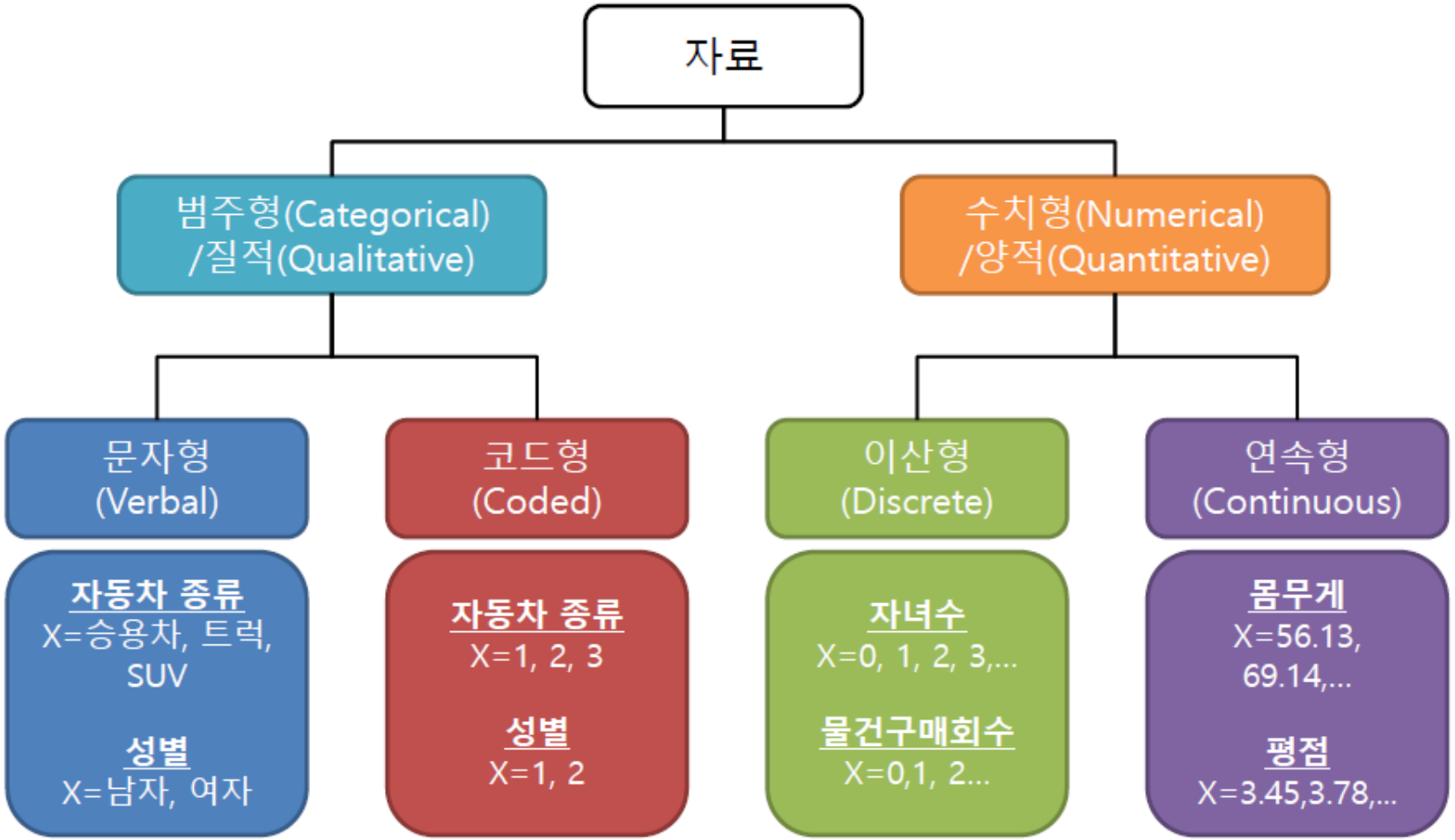
자료가 중요한 이유

자료와 분석방법

- 자료의 형태가 중요한 이유
  - 자료의 형태와 분석목적에 따라 통계분석이 결정

형태		요약방법	자료정리	그래프	분석방법
범주형	범주형	도표 그래프	도수분포표 분할표	막대도표 원도표	교차분석
범주형	수치형	도표+ 수치	그룹별 평균	그룹별 막대도표 그룹별 상자도표	t-test ANOVA
수치형	수치형	수치 그래프	산술평균 중앙값 조화평균	히스토그램 상자도표 산점도	상관분석 회귀분석 등

자료의 종류



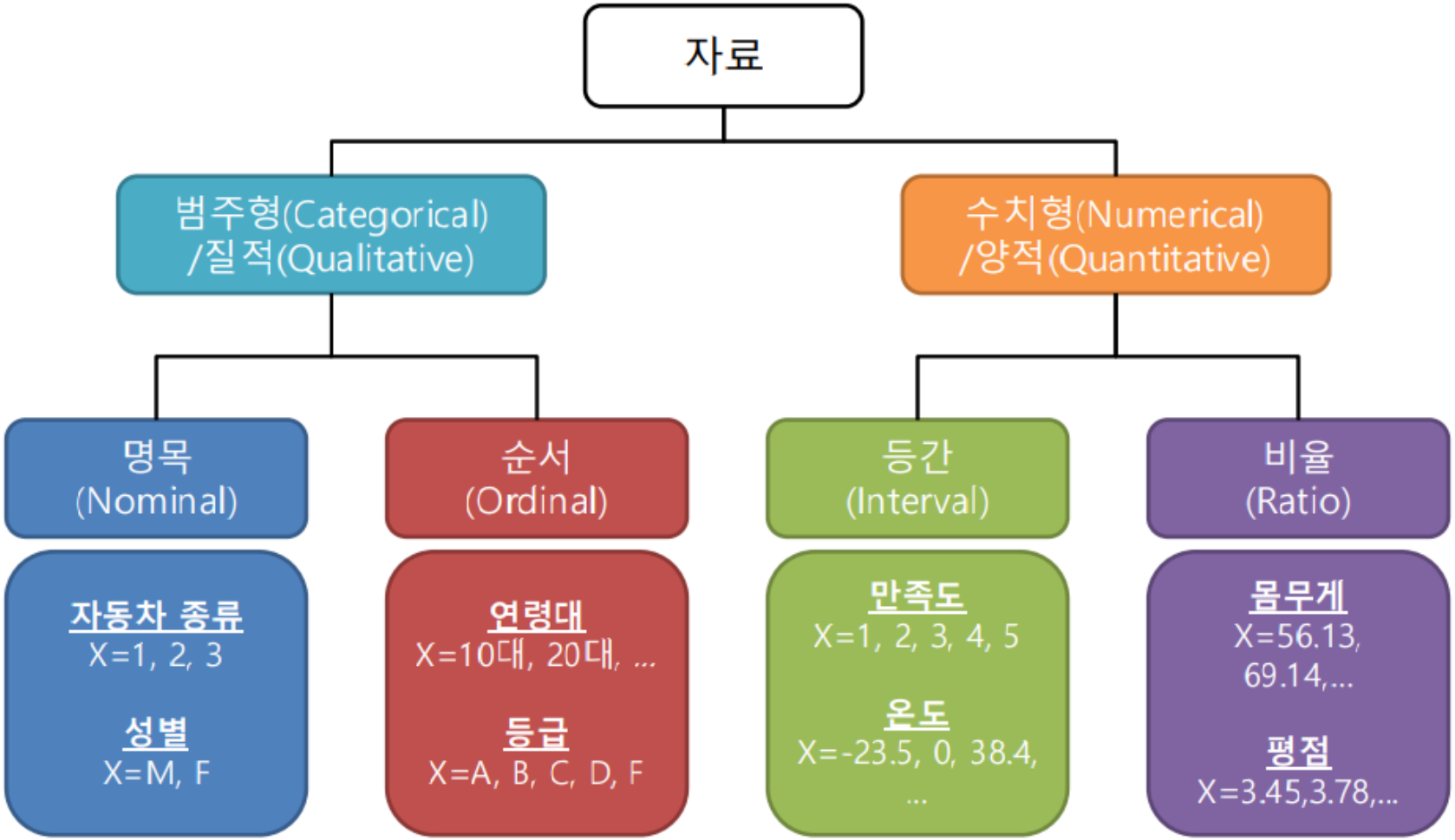
Source : Doane and Seward(2011), Applied Statistics in Business & Economics

자료의 종류

- 범주형(질적)자료
  - 명목(Categorical), 속성(Attribute) 자료 또는 변수(Variable)
  - 몇 개의 특성에 의한 범주를 나누어 코드숫자로 나타낸 자료
    - 예) 성별(남자= 1, 여자=2), 학력(고졸 = 1, 대졸 =2, 대학원졸=3)
  - 연산을 할 수 없음 (성별 = 1.5?)

- 순서형(연속)자료
  - 연속 자료 또는 변수(Continuous variables)
  - 이산형 자료(Discrete data): 하나, 둘 셀 수 있는 자료
    - 예) 도시의 아파트 층수, 썩은 치아 수
  - 연속형 자료(Continuous data): 구간에서 값을 모두 취할 수 있는 자료
    - 예) 신생아의 키, 세금액수

자료와 척도



Source : Doane and Seward(2011), Applied Statistics in Business & Economics

자료와 척도

- 척도(Measurement)의 종류
  - 척도: 측정을 하기 위해서 사용한 측정도구
  - 예) 키-cm, 몸무게-kg, 성별-M,F, 속도-km/h, 실업률-%
- 범주형 자료
  - 명목자료/척도 (Nominal measurement)
  - 측정 대상의 특성을 분류하거나 확인
  - 예) 성별, 혈액형, 직업구분
- 순서자료/척도(Ordinal measurement)
  - 측정대상의 특성을 몇 개의 범주로 구분할 뿐만 아니라 그 범주들 사이에 순서관계가 성립하는 경우
  - 예) 학력, 학점, 나이대 등
  - 예) 좋아하는 과목의 순서: 통계학=1위, 간호학=2위, 수학=3위
  - 예) 먹고 싶은 순서대로 순위를 정하세요
  - 아이스크림(3), 초콜렛(1), 솜사탕(4), 오렌지 (2)

자료와 척도

- 수치형 자료
  - 등간자료/척도 (Interval measurement)
  - 측정 대상의 양적인 차이를 나타내주는 변수
  - 절대영점이 존재하지는 않지만 균일한 간격을 두고 분할하여 측정
  - 예) 설문지의 설문문항(리커르트 5점), 온도, 아이큐지수

- 비율자료/척도(Ratio measurement)
  - 측정 대상의 양적인 차이를 나타내주는 변수
  - 절대영점이 존재하며, 비율계산이 가능
  - 예) 시험 점수, 스트레스 점수, 키, 몸무게

이름	연령	연령대
홍길동	21	20대
이길동	29	30대
백두산	35	30대

자료와 척도

- 리커르트 척도법(Likert Scaling) - 등간자료
  - 리커르트 척도법은 여러 개의 항목으로 측정대상의 속성을 측정하고 해당
  - 항목에 대한 점수를 합산하여 대상의 속성에 대한 측정치를 얻어내는 척도
  - 5점, 7점, 9점
  - 설문문항간 타당도 및 신뢰도가 높아야 함

	전혀 그렇지 않다	약간 그렇지 않다	그렇지 않다	보통이다	그렇다	약간 그렇다	매우 그렇다
	①	②	③	④	⑤	⑥	⑦
8. 이 병원은 약속한 의료서비스(진료, 수술, 면담, 투약, 회진)를 모두 제공한다.							
9. 환자가 어려운 일에 발생했을 때 이 병원의 의료종사자들은 관심을 가져주고 환자를 안심시키기 위해 노력한다.							
10. 이 병원은 믿고 의지할 만하다.							

자료와 척도

- 서스턴 척도법(Thurstone Scaling) - 등간자료
  - 척도의 등간성을 확보할 목적으로 개발된 척도
  - 예: 통증을 측정하기 위한 서스턴 척도법
  - 1) 욱신거린다(1.5) [ ]
  - 2) 뜨끔뜨끔하다(2.5) [ ]
  - 3) 바늘로 찌르듯이 아프다(5.7) [O]
  - 4) 망치로 때리듯이 아프다(7.5) [O]
  - 5) 까무러치게 아프다(8.9) [ ]
  - ==> 측정값 :(5.7+7.5)/2=6.6

자료와 척도

- 어의차이척도(Semantic differential scale) - 등간자료
  - 양극단에 서로 배타되는 속성을 놓고 평가

■ 다음 고속철도의 서비스 정도에 대한 본인의 느낌을 평가하여 주십시요.

문항1. 빠르다 : ④__ : ③__ : ②__ : ①__ : ②__ : ③__ : ④__ : 느리다
문항2. 소음이 적다 : ④__ : ③__ : ②__ : ①__ : ②__ : ③__ : ④__ : 소음이 많다.
문항3. 좌석이 넓다 : ④__ : ③__ : ②__ : ①__ : ②__ : ③__ : ④__ : 좌석이 좁다
문항4. 청결하다. : ④__ : ③__ : ②__ : ①__ : ②__ : ③__ : ④__ : 청결하지 않다.

- 고정적합척도(Constant sum scale) - 비율자료
  - ex) 자동차를 구매할 때 중요시하는 정도를 100점을 만점으로 분할하여 할당하시오.
  - 디자인 (10)점, 배기량(20)점, 색상 (20)점, 상표 (50)점
  - 합계 : 100점

자료의 종류

등간척도

비율척도

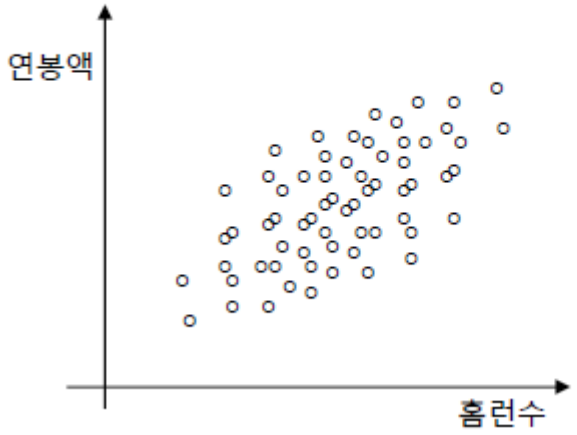
순서척도

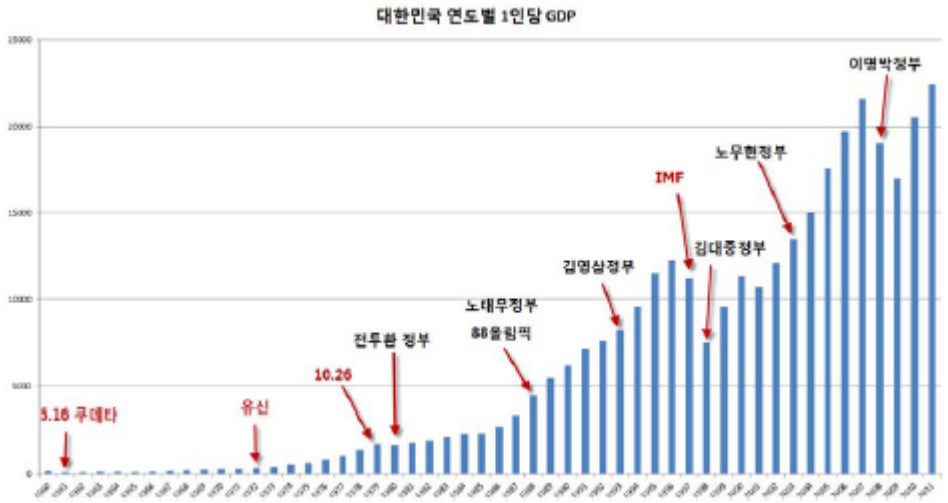
명목척도

자료의 종류(시간관점)

- **횡단 자료**(Cross-sectional data)
  - 한번의 시간에 얻어진 데이터

- **종단 자료** (Longitudinal data)
  - 시계열자료(Time series data)
  - 동일한 대상으로부터 여러 시간에 걸쳐 얻어진 데이터
  - 예) 주식, 2010-2014년까지의 GDP변화 등





범주형 자료(일변량)

도수분포표

- 도수분포표(frequency table)
  - 처음 조사된 원자료는 그 자료의 특징 및 분포를 파악하기 어려움
  - 데이터 각 값의 출현도수를 세거나 몇 개의 구간으로 나누어 각 구간에 속하는 데이터의 개수를 세어서 정리한 표
  - 수집된 자료를 적절한 계급으로 분류, 정리한 표
  - 도수(빈도): 범주에 속한 개체의 수
  - 상대도수: 범주에 속한 개체의 비율 (%)
  - 누적도수: 범주에 속한 개체의 누적비율(%) – 순서자료일 경우
  - 질적자료 정리 및 연속변수를 질적자료로 변환해서 사용할 경우에 사용

도수분포표

- 범주형 자료의 도수분포표

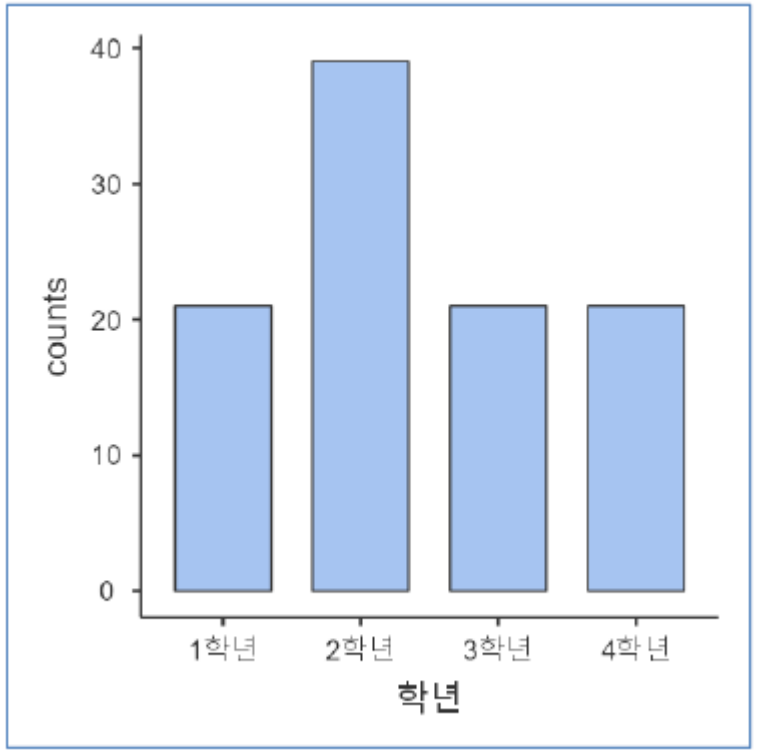
학점 (사례)

id	성별	분반	학년	몸무게	출석	중간	기말
1	남자	1	1	40	100	87	80
2	여자	2	2	50	100	83	60
3	남자	1	3	56	100	84	60
4	여자	2	4	51	100	73	60
5	남자	1	1	55	100	68	60
6	남자	2	2	61	100	77	50
7	여자	1	3	69	100	40	80
8	여자	2	2	44	100	73	30
9	여자	1	2	66	80	64	40
10	남자	2	2	60	100	66	40
			.....				
98	여자	1	3	63	100	56	40
99	남자	1	2	69	100	52	20
100	여자	1	2	66	100	67	70
101	남자	2	2	90	90	54	10
102	여자	1	2	120	100	49	40

도수분포표

- 도수분포표와 막대그래프 (Bar chart)
  - 누적비율: 순서형일 경우에 사용하면 편리

학점	빈도	비율(%)	누적비율(%)
1학년	21	21%	21%
2학년	39	38%	59%
3학년	21	21%	79%
4학년	21	21%	100%
합계	102	102	102



그래프의 중요성

그래프의 중요성

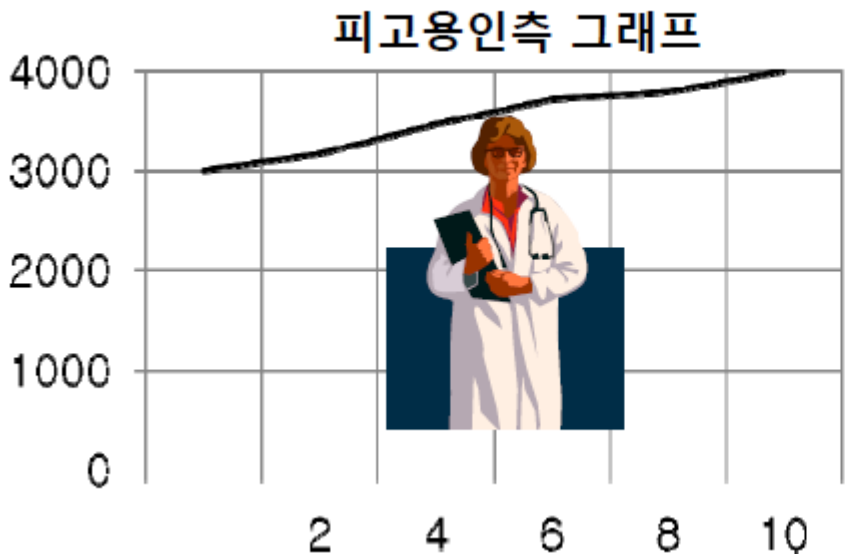
- 그래프
  - 인간이 지닌 시각적인 인지능력을 활용하여 직관적으로 그 현상을 쉽게 인식하도록 하는 방법
  - 통계적인 데이터를 요약하여 시각적으로 그 특징을 나타내는 것

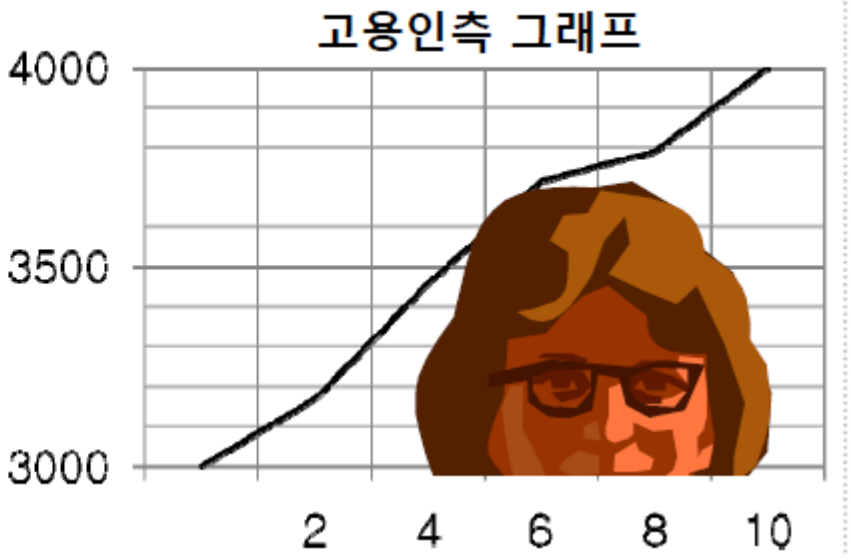
- 그래프의 문제점
  - 그래프는 자료가 가지고 있는 속성뿐만 아니라 강렬한 인상을 주게 되어 확대해석의 오류를 범할 위험이 있음

- 활용
  - 그래프를 통해 데이터의 특징 및 자료의 이상치를 점검하여 이후의 통계분석에 대비함

올바른 그래프 그리는 법

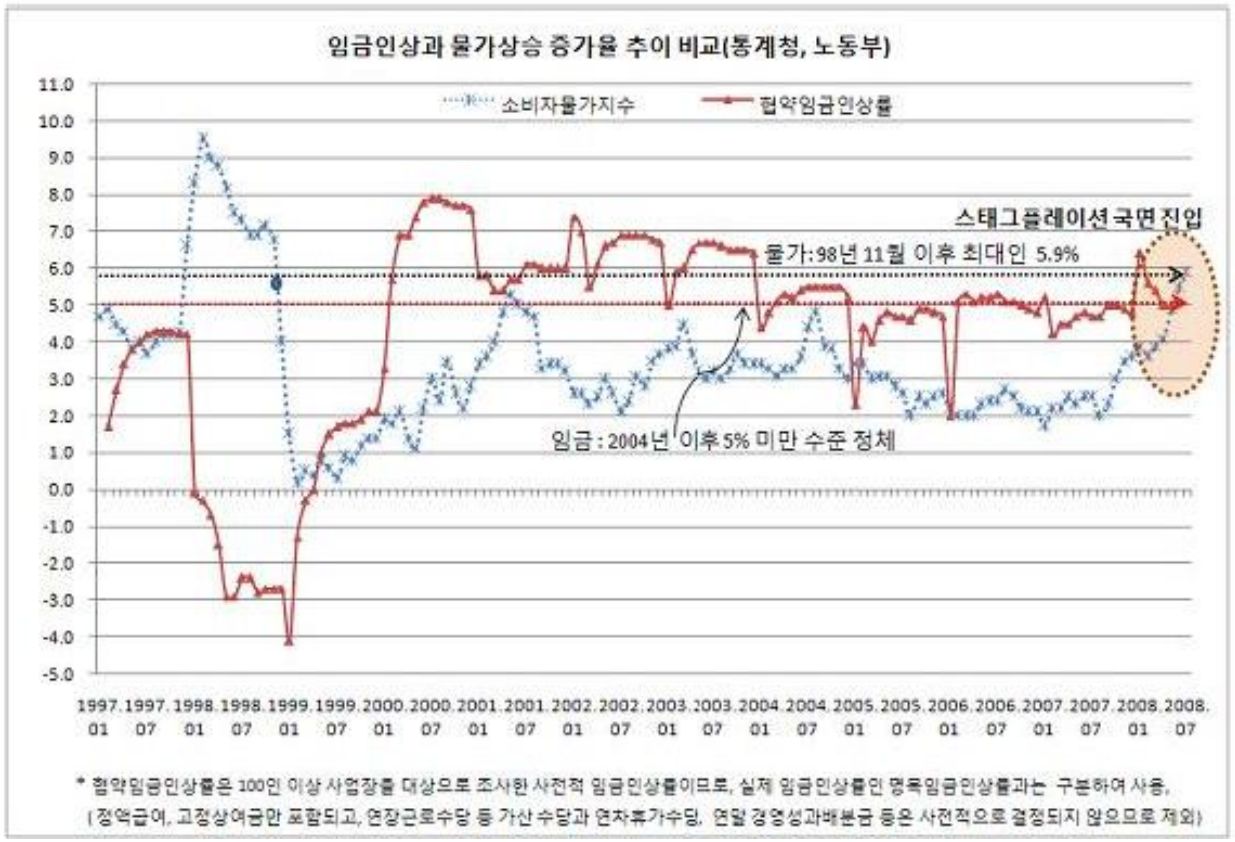
- 올바른 그래프의 작성
  - 그래프에 적합한 제목
  - 자료의 출처, 표본의 크기, 수집방법을 나타냄
  - 축에 대한 제목을 명확히 함
  - 도수비율, 퍼센트 등이 0에서 시작하는지 점검
  - 변수의 측정단위가 표시되어야 함





올바른 그래프 그리는 법

- Y축 단위가 0이 아닐 때
  - 단순히 하나의 수치만 적지 말고 비교되는 자료를 동시에 포함
  - 물가변동에 대한 임금상승율의 변화를 객관적으로 볼 수 있음



범주형 자료(다변량)

범주형자료(2개) - 분할표

- 분할표(Contingency table)
  - 관측치를 몇 개의 범주로 분할하여 그 해당도수로 자료를 정리해 놓은 표
  - 다변량 자료: 범주형 변수가 2개 일 때
  - 비율을 표시하는 방법이 중요(행, 열, 전체): 분석목적에 따라 비율표시
  - 분할표를 이용한 통계분석 : 교차분석(chi-square)

일원분할표

대조군	처리군	합계
60	40	100

일원분할표

생존	사망	합계
50	50	100

이원분할표

	대조군	처리군	계
생존	40	10	50
사망	20	30	50
계	60	40	100

사전설계 분할표

- 사전(실험)설계(코호트 연구)일 때
  - 사전에 그룹의 수를 결정해서 연구할 때 -> 해석: 그룹에 따른 차이
  - 코호트 연구
    - 예) 비타민과 감기에 대한 연구를 하기 위해, 비타민을 투여할 실험군과 가짜약을 투여할 대조군으로 사전에 구분하여 연구
  - 실험군과 대조군에 따른 차이를 검정: 교차분석(동질성 검정)
  - 비율기준: 그룹별 자료수

사후

사전	그룹	감기발병		합계
		유	무	
	실험군 (비타민)	17 (34.0%)	33 (66.0%)	50 (100.0%)
	대조군 (Placebo)	38 (76.0%)	12 (24.0%)	50 (100.0%)
	합계	55 (55.0%)	45 (45.0%)	100 (100.0%)

사전설계 분할표

- 감기발병을 기준으로 했을 경우: 실험군의 자료수 증가할 때

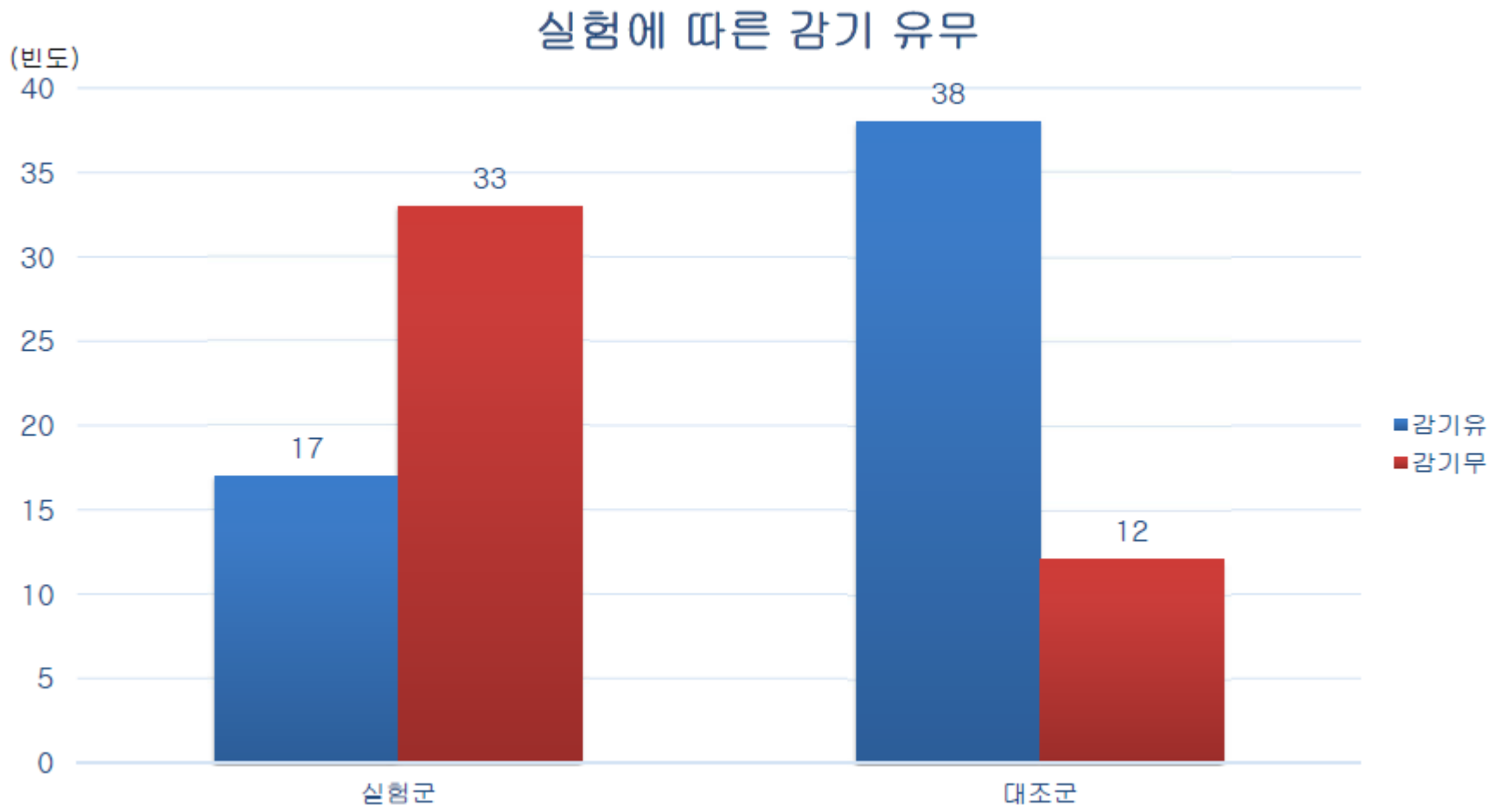
그룹	감기발병		합계
	유	무	
실험군 (비타민)	170 (34.0%)	330 (66.0%)	500 (100.0%)
대조군 (Placebo)	38 (76.0%)	12 (24.0%)	50 (100.0%)
합계	55 (37.8%)	45 (62.2%)	100 (100.0%)

변화없음

그룹	감기발병		합계
	유	무	
실험군 (비타민)	170 (81.7%)	330 (96.5%)	500 (91.0%)
대조군 (Placebo)	38 (18.2%)	12 (3.5%)	50 (9.0%)
합계	208 (100.0%)	342 (100.0%)	550 (100.0%)

자료수에 따라 변화

사전설계 분할표



사후설계 분할표

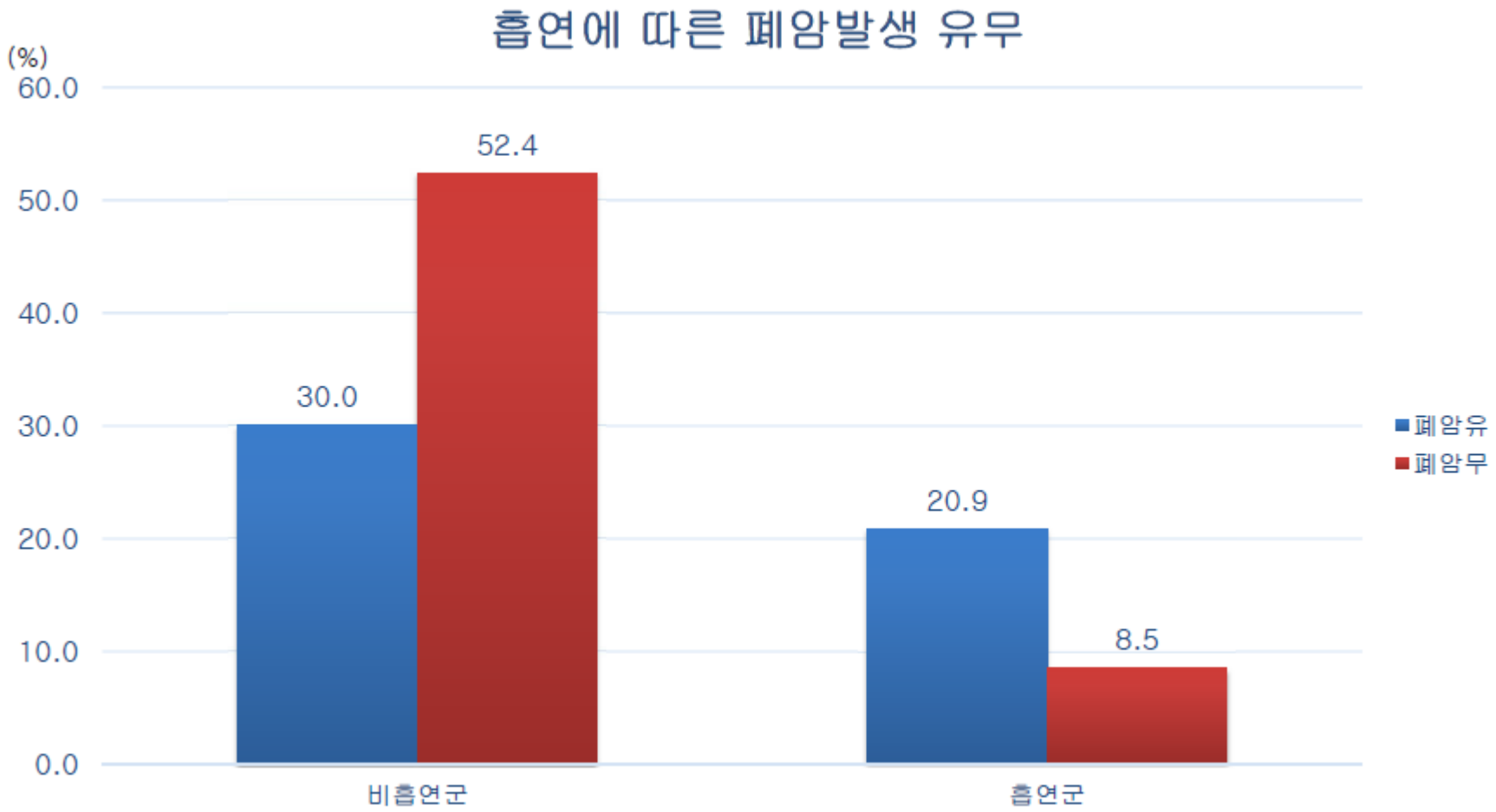
- 사후설계(사례대조) 분할표
  - 사전에 그룹의 인원수를 정하지 못하고 사후의 결과를 토대로 연구할 때 -> 해석: 두 변수간의 관련성
  - 예) 흡연이 폐암과 연관이 있는지를 연구하기 위해 흡연자와 비흡연자를 대 상으로 폐암발생여부를 사후에 조사
  - 폐암과 흡연간의 관련성을 검정: 교차분석(독립성검정)
  - 비율기준: 전체 자료수

사전

사후	폐암	흡연					합계
		비흡연군	장기금연군	단기금연군	재흡연군	흡연군	
	무	170,867 (52.0%)	51,690 (15.7%)	46,598 (14.2%)	29,178 (8.9%)	27,784 (8.5%)	326,117 (99.3%)
	유	723 (0.2%)	370 (0.1%)	497 (0.2%)	319 (0.1%)	504 (0.2%)	2,413 (0.7%)
	합계	171,590 (52.2%)	52,060 (15.8%)	47,095 (14.4%)	29,497 (9.0%)	28,288 (8.6%)	328,530 (100.0%)

한국인에서 흡연과 폐암의 상관관계 및 폐암의 위험인자 분석, 국민건강보험 일산병원 연구소 (2016)

사후설계 분할표


흡연에 따른 폐암발생 유무
(%)
60.0
50.0
40.0
30.0
20.0
10.0
0.0
52.4
30.0
20.9
8.5
비흡연군
흡연군
폐암유
폐암무

연습문제

- 07.온라인게임.csv를 이용하여 범주형 자료를 분석하세요.
- Q1.성별: 1=남자, 2=여자
- Q2. 학력: 1=고등학생이하, 2=고등학교졸업, 3=대학생, 4=대학생 이상
- Q3.성별(sex)의 돗수분포표, 막대그래프(돗수), 막대그래프(비율),
- Q4.학력(school)의 돗수분포표, 막대그래프(돗수), 막대그래프(비율),
- Q5.성별(sex)과 학력(school)의 이원분할표, 막대그래프

수치형 자료(일변량)

수치형 자료

- 수치형 자료 정리
  - 일반적으로 연속자료의 특성을 시각적으로 파악하기 보다는 숫자로 기술
  - 분포의 특성을 숫자로 표현하는 법
  - 지능지수 : IQ, 경제현상:GDP, 불쾌지수 등
  - 중심위치와 산포경향

- 중심위치(central location)
  - 관찰된 자료들이 어디에 집중되어 있는가를 나타냄
  - 종류 : 산술평균, 중앙값, 최빈값, 기하평균, 조화평균, 가중평균

- 산포경향 (변동)
  - 자료가 중심위치로부터 어느 정도 흩어져 있는가를 나타냄
  - 자료가 평균으로부터 떨어진 평균 차이(거리)
  - 종류 : 범위, 편차, 분산, 표준편차

수치형 자료

- S대학 기초통계 수강생의 몸무게 분석

(단위: kg)

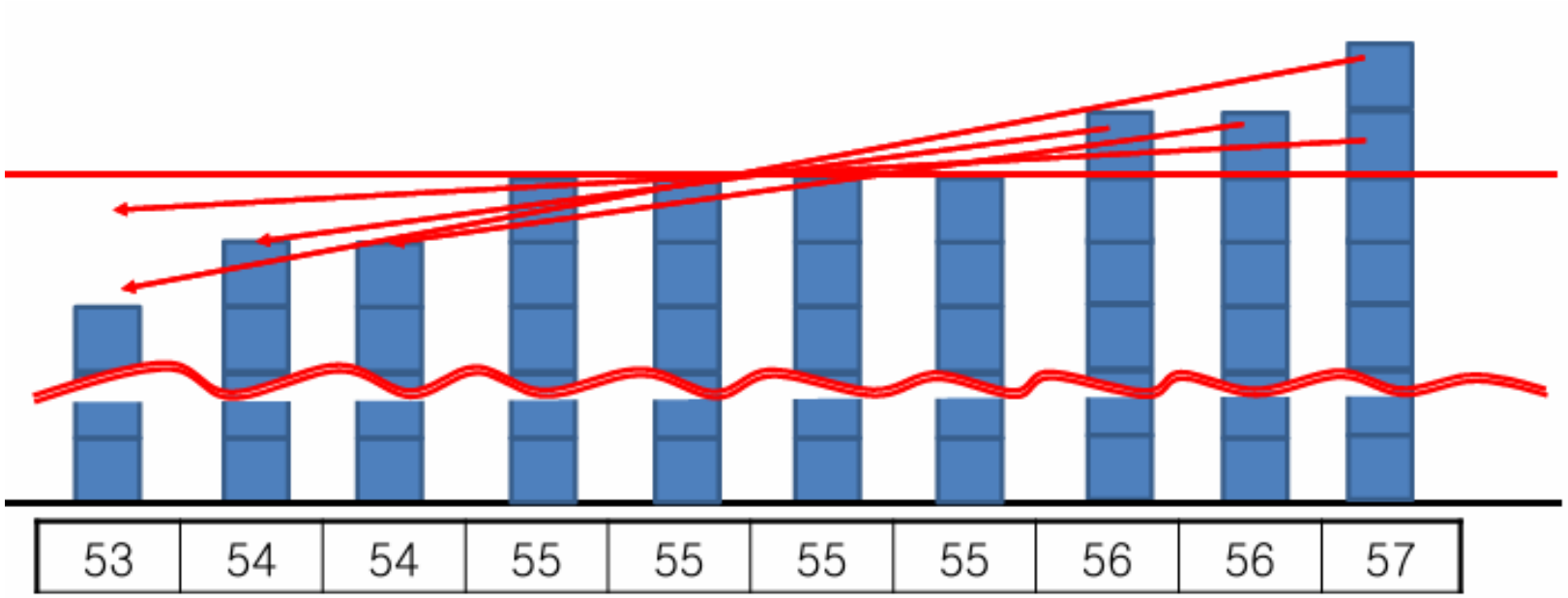
40	70	46	52	55	56	62	71	53	62
50	44	63	45	62	48	57	58	57	56
56	66	56	57	67	50	68	41	60	59
51	60	52	55	49	68	54	58	59	50
55	56	48	67	49	53	62	48	53	61
61	72	57	52	52	58	59	69	47	64

(산술) 평균 (Mean)

- 평균
  - 균등하게 나누다
  - 중심위치(central location)
  - 자료의 중심적인 경향을 나타내는 수치(무게중심)
  - 관찰된 자료들이 어디에 집중되어 있는가를 나타냄
  - 전체 자료를 대표
    - 예) 학생들의 평균 국어점수, S대학 통계 수강생의 몸무게 분석

(산술) 평균 (Mean)

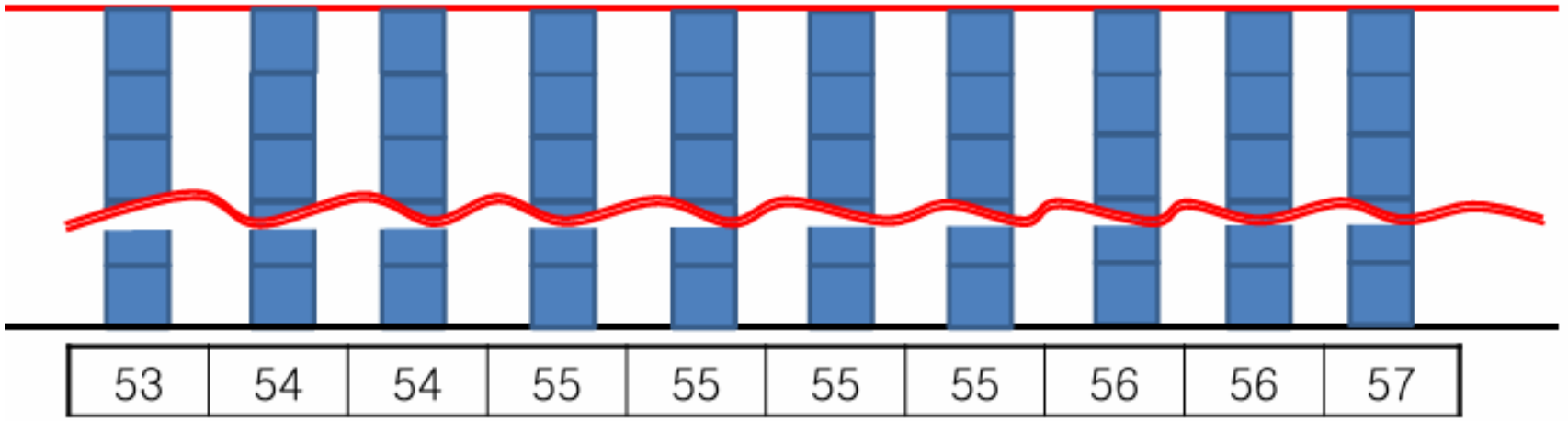
- 평균
  - 균등하게 나누다
  - S대학 경영통계 수강생의 몸무게 분석



(산술) 평균 (Mean)

- 평균
  - 균등하게 나누다
  - S대학 경영통계 수강생의 몸무게 분석

$$\bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n) = \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)$$
$$= \frac{1}{10}(53 + 54 + \cdots + 56 + 57)$$
$$= 55$$

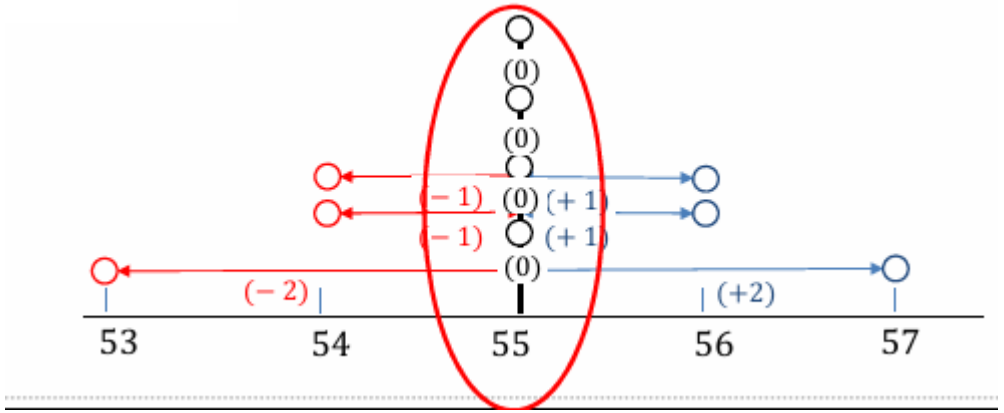


(산술) 평균 (Mean)

- 평균
  - 자료의 중심적인 경향을 나타내는 수치(무게중심)
  - 중심위치(central location): 분포상의 무게 중심
  - 중심위치: 평균을 중심으로 왼쪽 자료와 오른쪽 자료를 다 더하면 0
  - 편차: $\sum(x_i - \bar{x}) = 0$

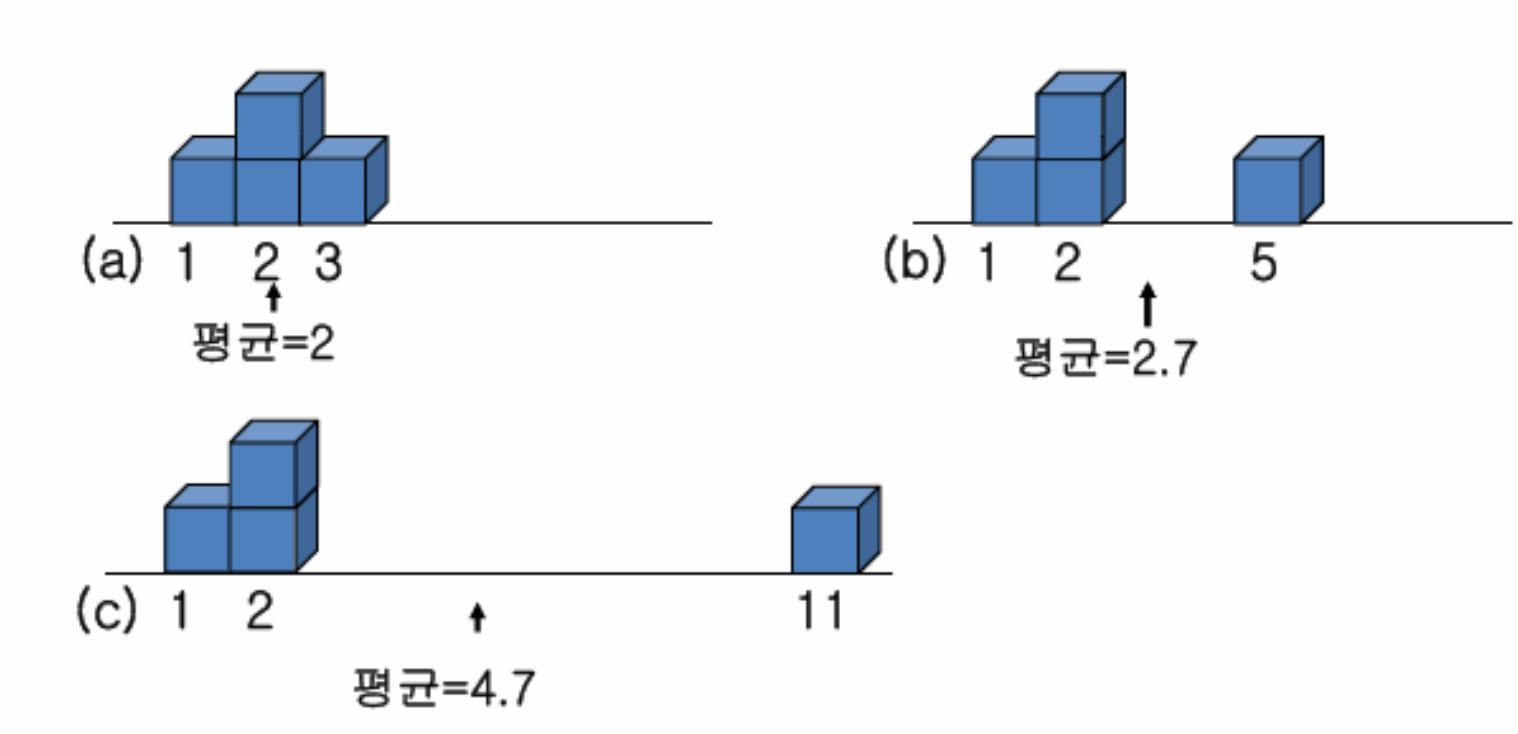
    $(53 - 55) + (54 - 55) + \ \ldots + (56 - 55) + (57 - 55) = 0$

  - 관찰된 자료들이 어디에 집중되어 있는가를 나타냄
  - 전체 자료를 대표



(산술) 평균 (Mean)

- 산술평균의 문제
  - 이상치(outlier)에 민감하게 반응함
  - 보완: 중앙값, 최빈값, 절사평균



중앙값(Median)

- 자료를 크기 순으로 나열할 때 가장 가운데 오는 값

$$\tilde{x} = x_{\frac{n+1}{2}}$$

- 특징
  - 이상치의 영향을 받지 않음
  - 중앙값을 중심으로 좌우 분포 면적이 같음
  - 원데이터를 크기 순서대로 재 배열
  - 짝수일 경우에는 가운데 두개 값의 평균

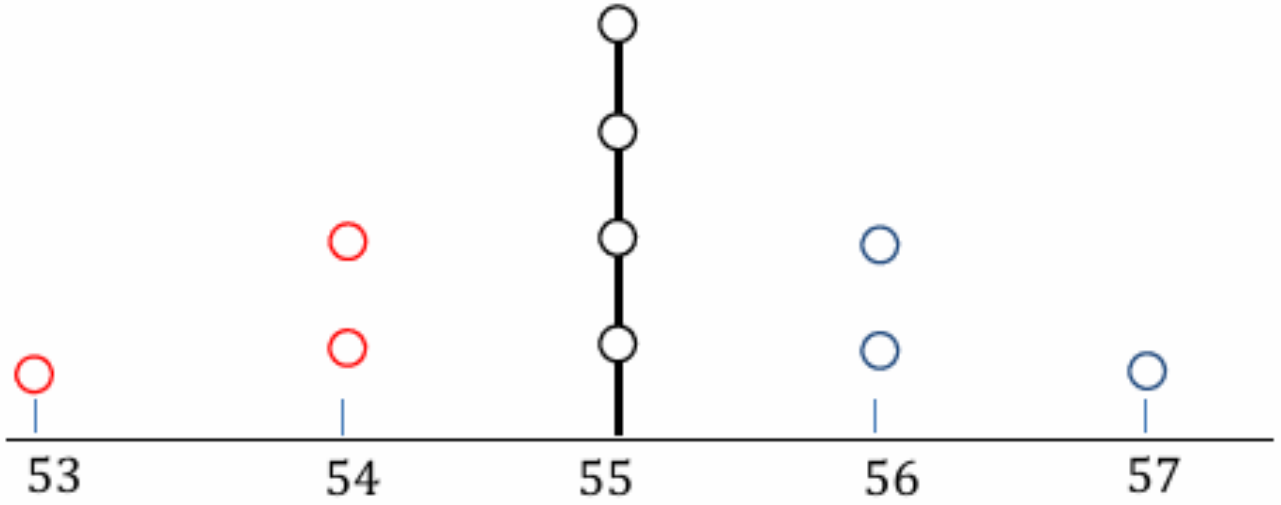
$$53, 54, \ldots, \underset{6}{55}, \underset{7}{55}, \ldots, 56, 57$$
$$1 \quad 2 \qquad\qquad 9 \quad 10$$

$$\tilde{x} = \frac{55 + 55}{2} = 55.5$$

최빈값(Mode)

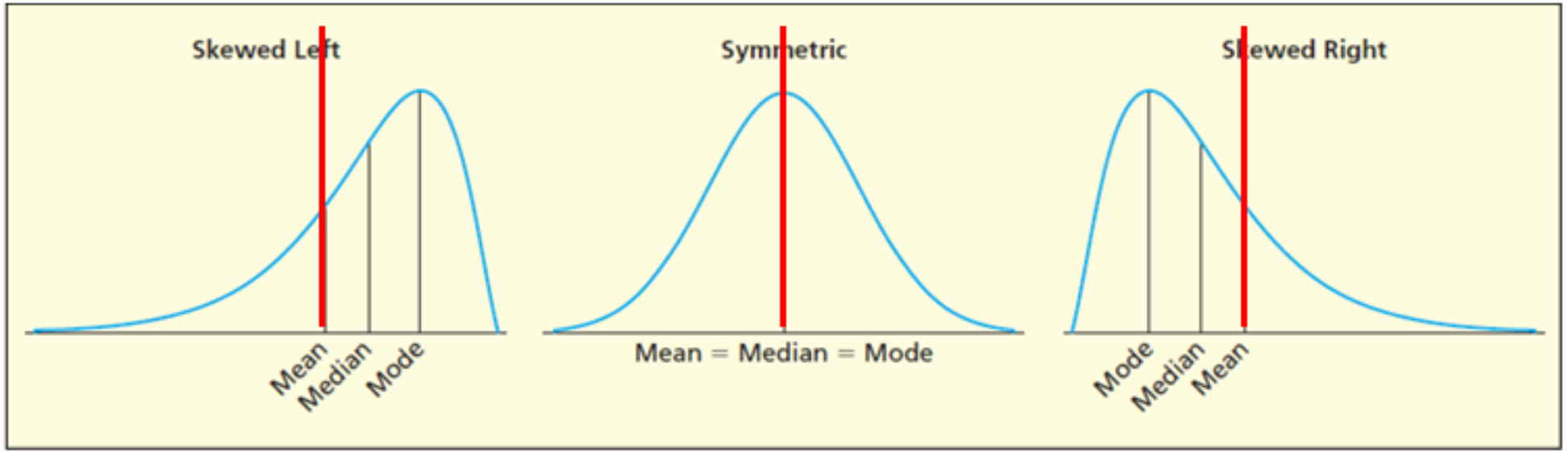
- 자료 중 발생빈도가 가장 높은 값
  - 빈도수에 의해 산출
  - 유일하지 않을 수도 있음

몸무게	빈도
53	1
54	2
55	4
56	2
57	1



중심위치의 모양

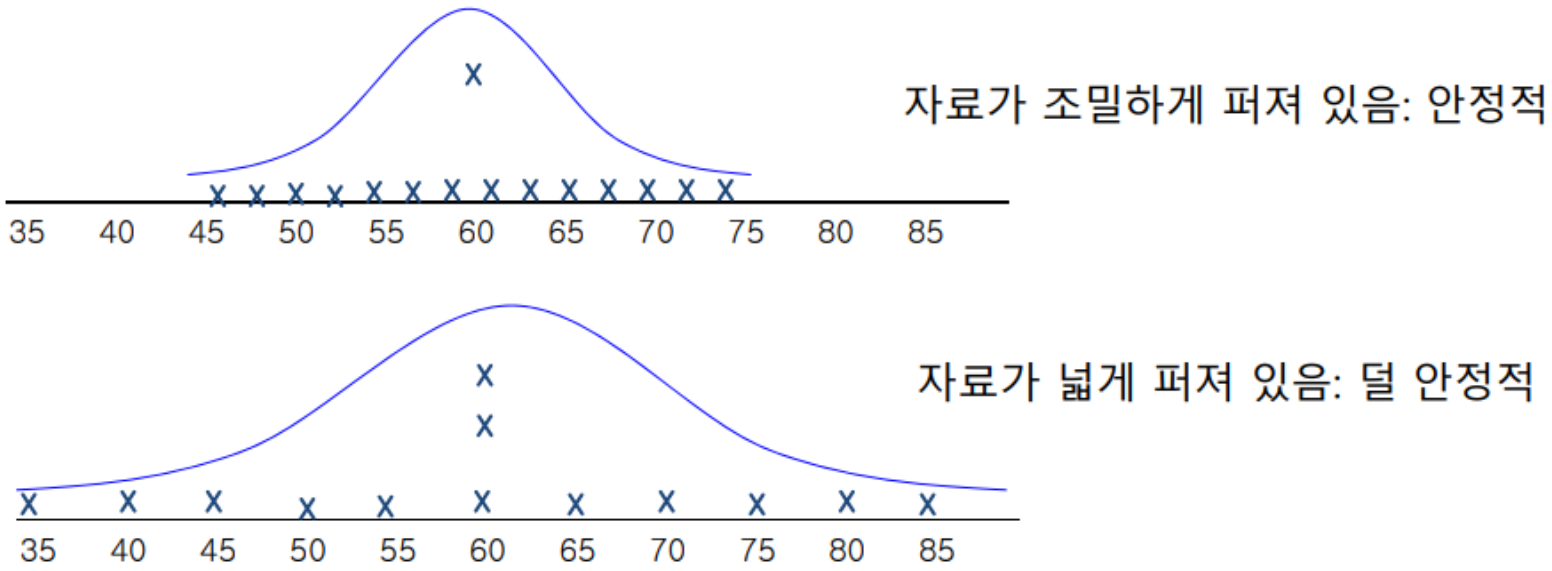
- 자료의 분포와 평균, 중앙값, 최빈값과의 관계



산포(변동)

산포

- 산포 (Dispersion, 변동)
  - 자료가 중심위치로부터 어느 정도 흩어져 있는가를 나타냄
  - 자료가 평균으로부터 떨어진 평균 차이(거리)
  - 수치 자료의 특징을 정리할 때 평균과 같이 제공
  - 종류 : 범위, 4분위, 편차, 분산, 표준편차
  - 중심위치(평균)이 얼마나 안정적인지에 대한 정보



범위

- 범위(Range)
  - 자료의 최대값에서 최소값을 뺀 것

$$R = X_{\max} - X_{\min}$$

  - 두 환자의 측정시간에 따른 맥박수 분포

	아침	점심	저녁
환자A	72	76	74
환자B	72	91	59

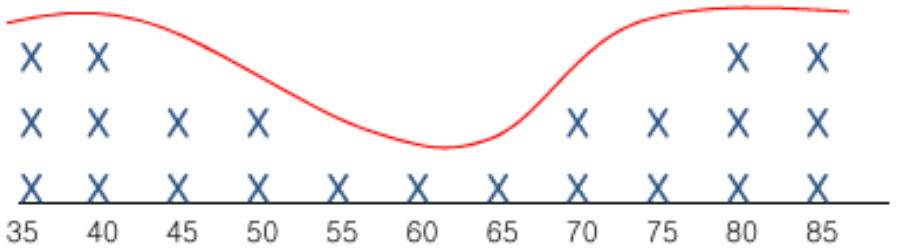
$$R_A = 76 - 72 = 4$$

$$R_B = 91 - 59 = 32$$

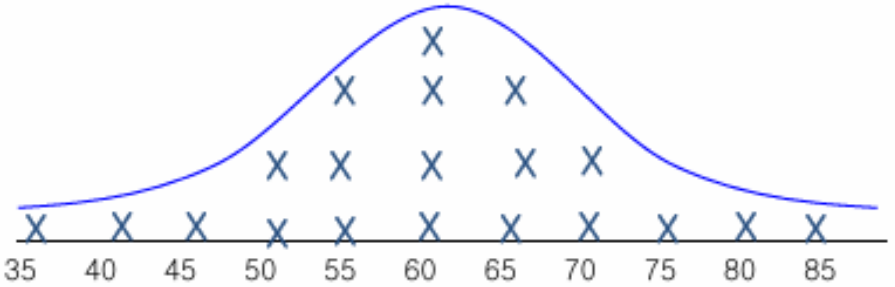
범위

- 범위(Range)의 문제
  - 같은 범위 값의 서로 다른 분포
  - 최대값과 최소값에 의해서만 영향 받음
  - 분포의 대표값으로 사용 못함

- 자료1

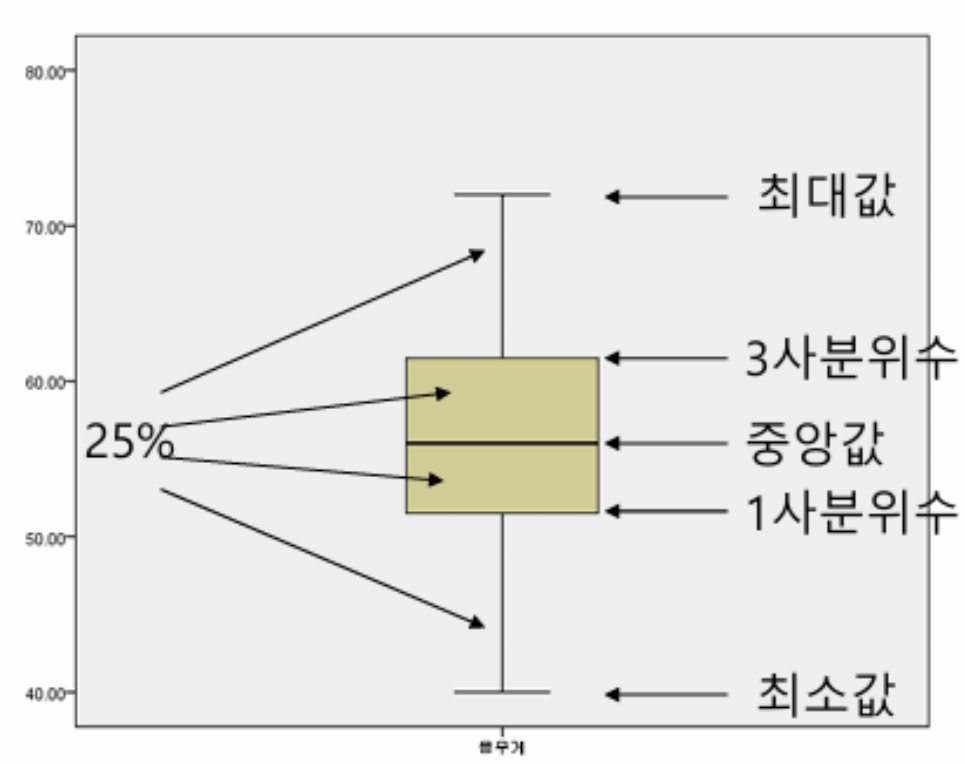


- 자료2



사분위

- 사분위(Interquartile-Range)
  - 자료를 동일한 비율로 4등분
  - 자료를 순서대로 정렬
  - 제1사분위수(Q1): 25%
  - 제2사분위수(Q2): 50%
  - 제3사분위수(Q3): 75%
  - 제4사분위수(Q4): 100%
  - 상자도표(box plots)에서 사용
  - 이상치(Outlier) 제거시 사용

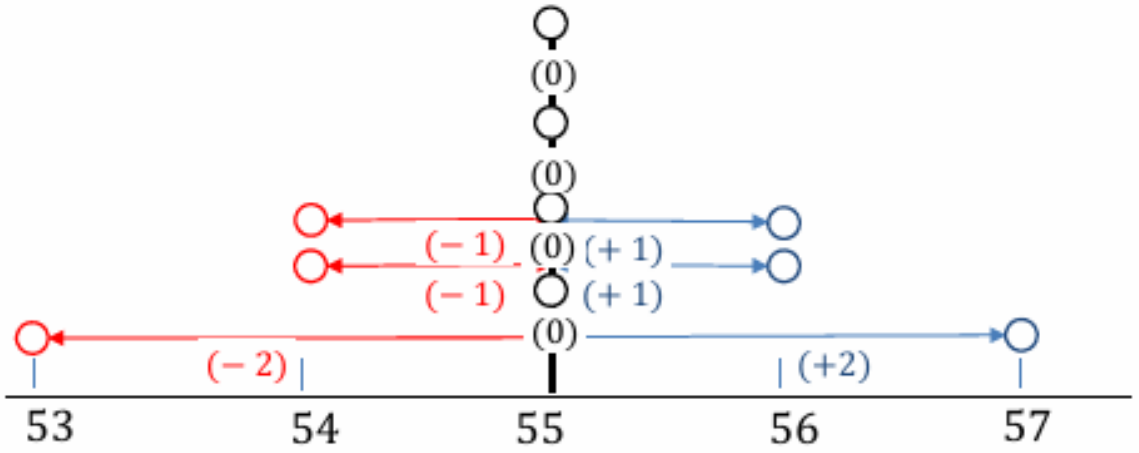


편차

- 편차
  - 데이터가 평균으로부터 얼마나 떨어져 있는지를 나타내는 지표
  - 평균을 중심으로 왼쪽 자료와 오른쪽 자료를 다 더하면 0

$$\sum (x_i - \bar{x}) = 0$$

$$(53 - 55) + (54 - 55) + \ \ldots + (56 - 55) + (57 - 55) = 0$$



분산과 표준편차

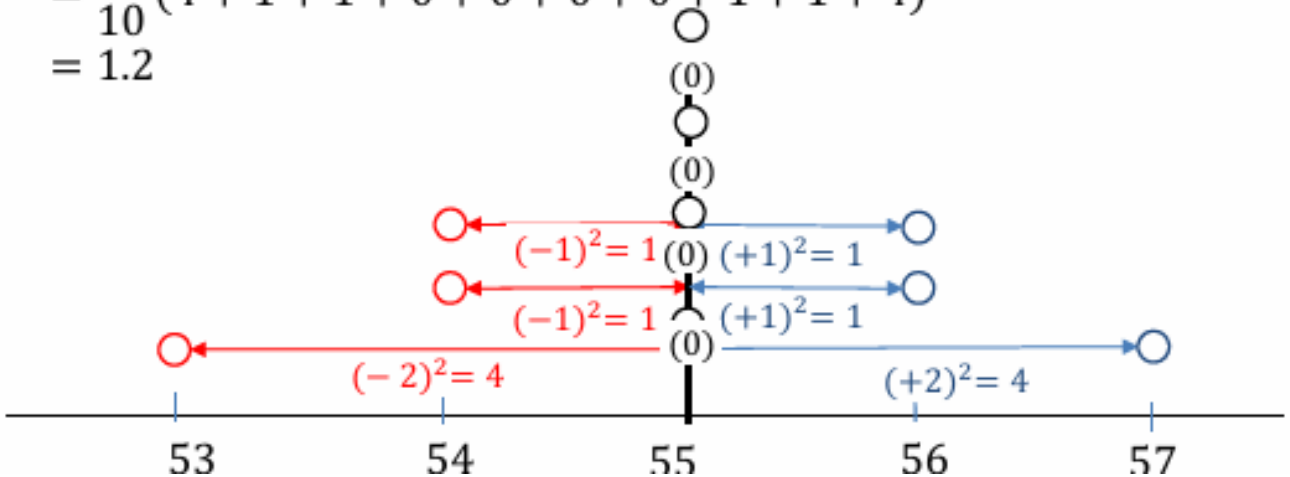
- 분산(variance)
  - 편차 -> 분산: "-"를 없애주기 위해
  - 자료가 평균을 중심으로 얼마나 광범위하게 분포하고 있는 가를 하나의 수치로 나타낸 통계량

$$var(\bar{x}) = \sigma^2 = \frac{1}{n}\sum (x_i - \bar{x})^2$$

$$\frac{1}{n}\sum (x_i - \bar{x})^2 = \frac{1}{10}(53-55)^2 + (54-55)^2 + \ \ldots + (56-55)^2 + (57-55)^2$$

$$= \frac{1}{10}(4+1+1+0+0+0+0+1+1+4)$$

$$= 1.2$$



분산과 표준편차

- 표준편차(Standard Deviation)
  - 분산을 원 자료의 측정단위로 다시 전환하기 위해
  - 자료가 평균으로부터 떨어진 평균 차이(거리)

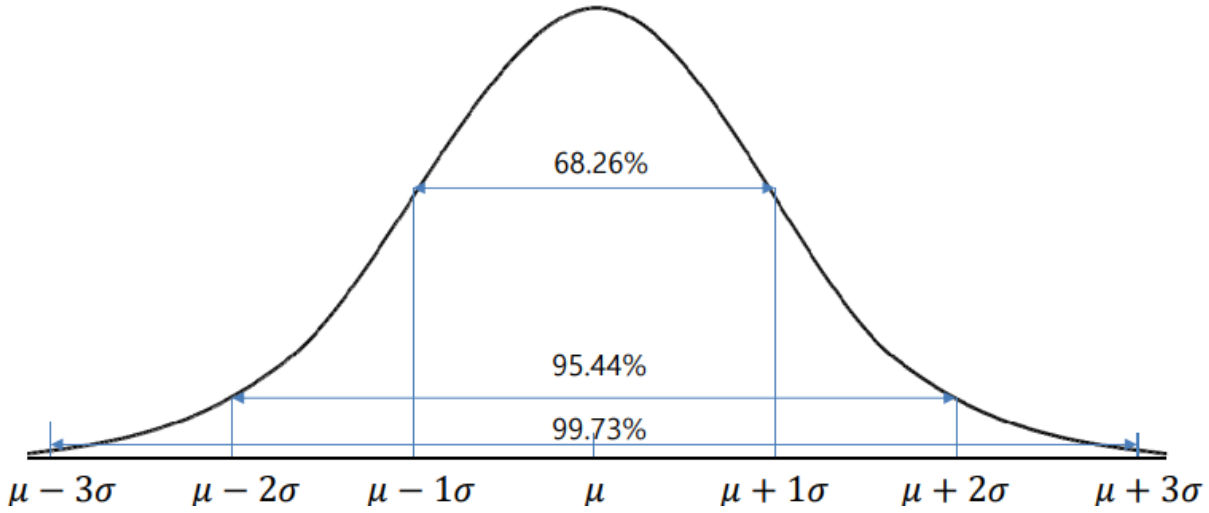
$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n}\sum(x_i - \bar{x})^2}$$

$$\sigma = \sqrt{1.2} = 1.1$$

- 편차 → 분산 → 표준편차
- $\sum(x_i - \bar{x}) = 0 \quad \rightarrow \quad \frac{1}{n}\sum(x_i - \bar{x})^2 = \sigma^2 \quad \rightarrow \quad \sigma = \sqrt{\sigma^2}$

분산과 표준편차

- 표준편차의 중요성
  - 자료의 분포와 변동에 대한 중요한 정보를 제공
  - 통계학의 중요한 규칙과 연결
  - Empirical Rule (경험적 법칙)
    - $k = 1,$ 68.26% 이상의 데이터가 $\mu \pm 1\sigma$ 사이에 있음
    - $k = 2,$ 95.44% 이상의 데이터가 $\mu \pm 2\sigma$ 사이에 있음
    - $k = 3,$ 99.73% 이상의 데이터가 $\mu \pm 3\sigma$ 사이에 있음



표준화

- 표준화(standardization)
  - 측정단위 등과 관계없이 자료를 표준화 시킨 값
  - z값으로 변환된 자료

$$z_i = \frac{x_i - \bar{x}}{s}$$

- 사례) $\bar{x} = 50kg,\ s = 5kg$일 때 40, 65kg인 자료의 표준화된 값은?

$$z = \frac{40 - 50}{5} = -2 \qquad z = \frac{65 - 50}{5} = 3$$

- 모든 자료가 $\bar{z} = 0,\ s_z = 1$ 로 표준화 됨 → 절대 비교가 가능

변동계수(Coefficient of Variation)

- 측정단위가 다르거나 자료 값의 차이가 큰 경우에 사용

$$CV = \frac{s}{\bar{x}} \times 100$$

- (예) 유치원 여자 어린이들의 몸무게와 50대 주부의 몸무게 분포 비교

	$\bar{x}$	s	CV
어린이	20	8	0.40
50대주부	55	13	0.23

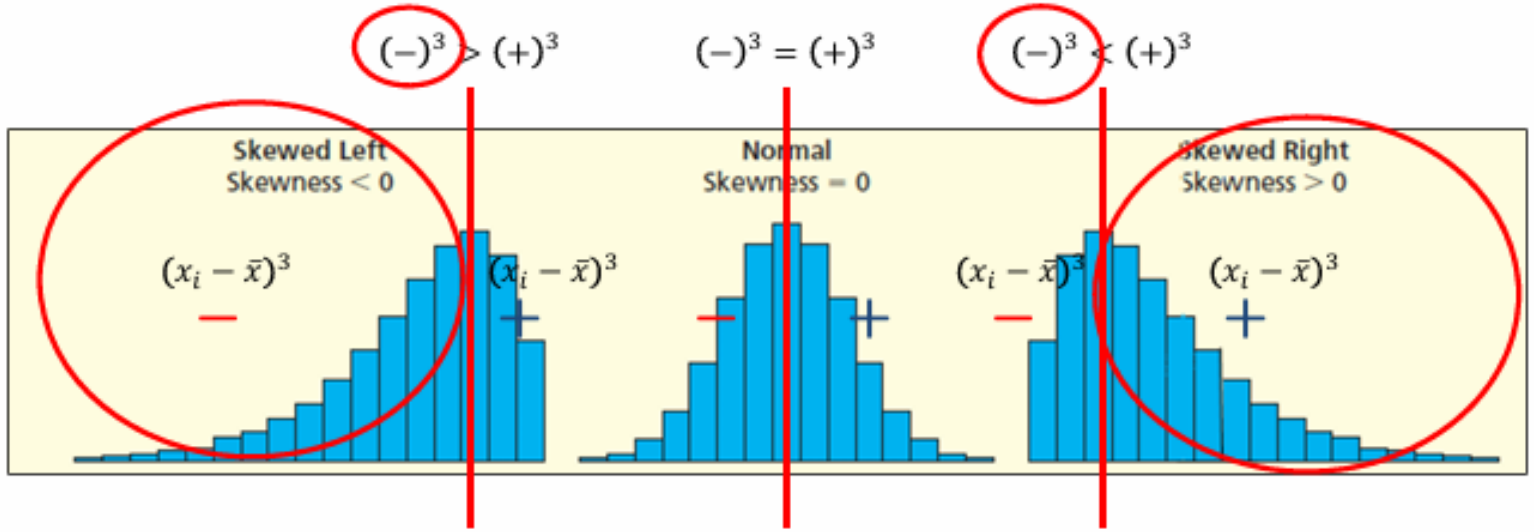
- 키(cm)와 몸무게(kg)

	$\bar{x}$	s	CV
키	175	15	0.09
몸무게	73	9	0.12

분포 평태

- 자료의 분포 형태
  - 왜도(skewed): 자료가 평균을 중심으로 대칭인지-> 정규분포인지 확인
  - 자료에 이상점이 있는지 점검

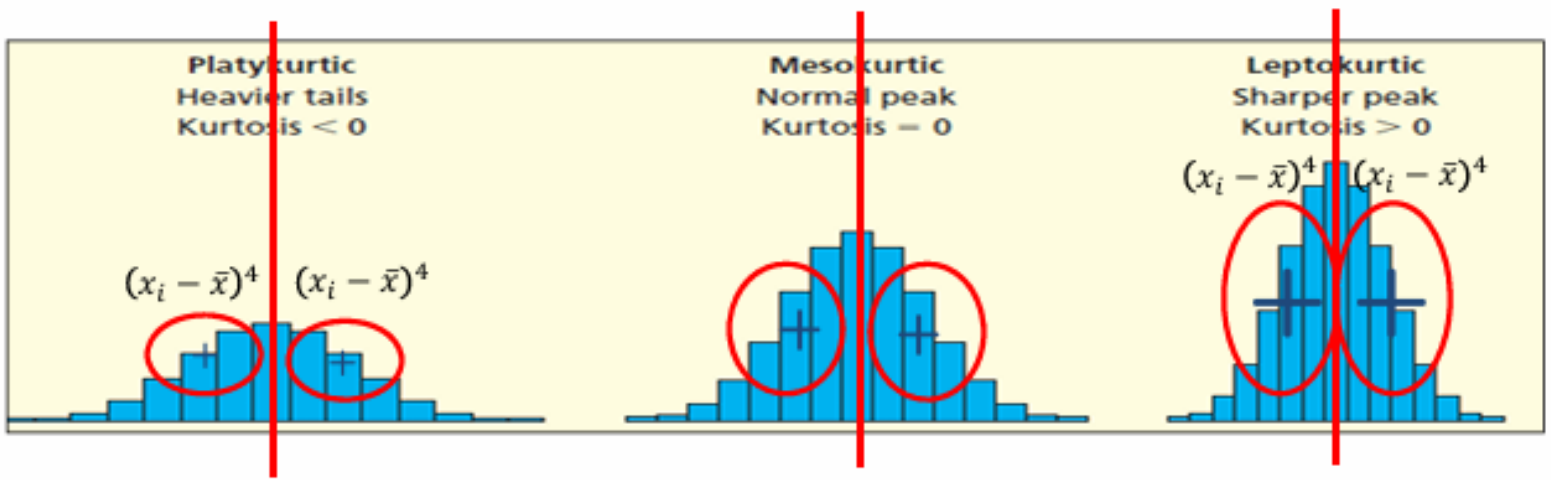
$$\sqrt{b_1} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^3$$



분포 평태

- 자료의 분포 형태
  - 첨도(kurtosis): 양쪽 꼬리가 얼마나 두터운지
  - 자료에 이상점이 있는지 점검

$$\sqrt{b_2} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s}\right)^4 - 3$$



수치형 자료의 범주화

수치형 자료의 범주화

- 계급(bin)의 수 결정 규칙
  - Sturges' Rule

$$k = 1 + 3.3\log(n)$$

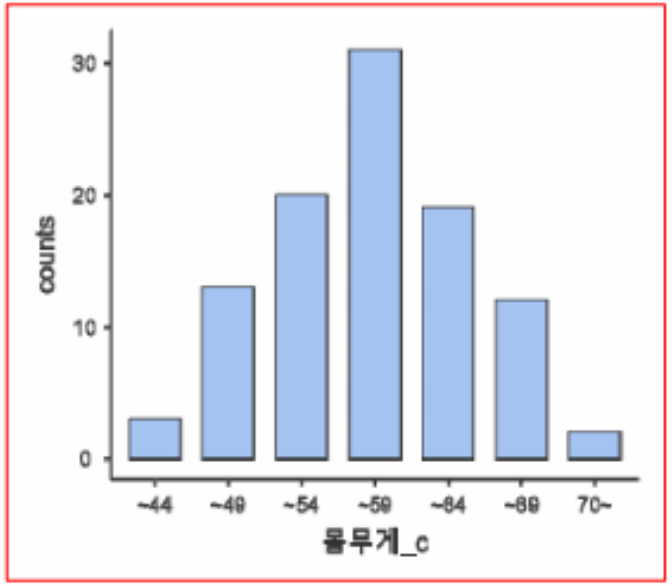
  - 자료의 개수 : $n = 60$
  - 계급의 수 : $1 + 3.3\log(60) = 6.87 \approx 7$
  - 자료의 최대값과 최소값: $72 - 40 = 32$
  - 계급의 폭 : $\frac{32}{7} = 4.57 \approx 5$

- 자료의 특성을 고려해 분석자가 결정(^^)
  - 같은 간격으로 하는 것이 기본
  - 소득 등 특정 구간에 자료가 많을 경우에는 특성을 고려해 구간 배분 가능

수치형 자료의 범주화

- 범주형 자료로 변환 후 정리

몸무게_범주	돗수	%	누적%
~44	3	3	3
~49	13	13	16
~54	20	20	36
~59	31	31	67
~64	19	19	86
~69	12	12	98
70~	2	2	100

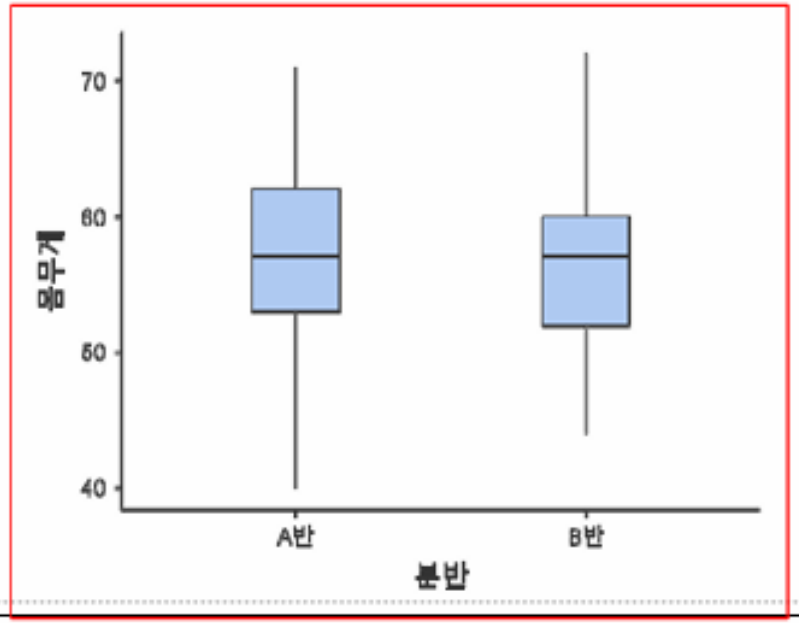


수치형 자료(다변량)

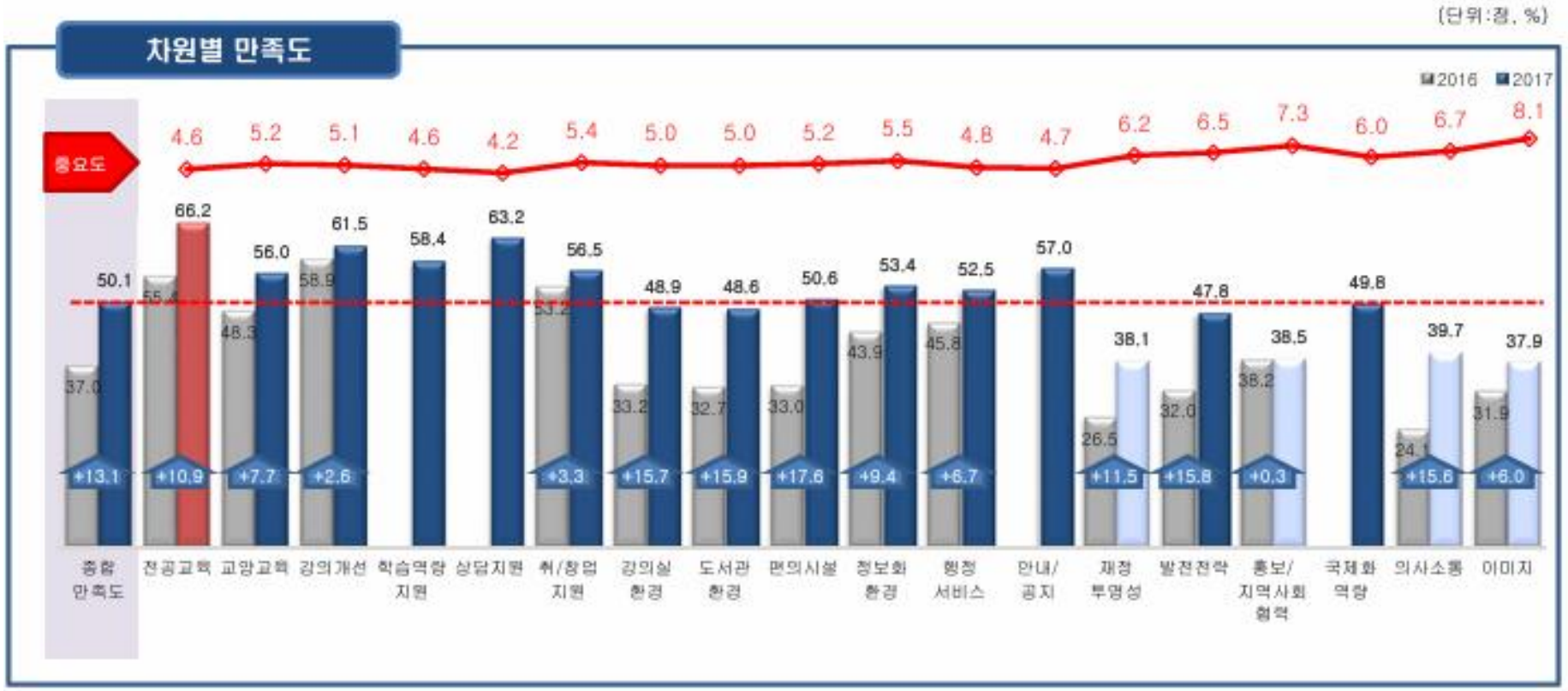
그룹별 수치자료 비교

- 그룹간 수치자료 비교
  - 범주형 자료+ 수치형 자료
  - 통계값: 표본크기, 평균, 표준편차

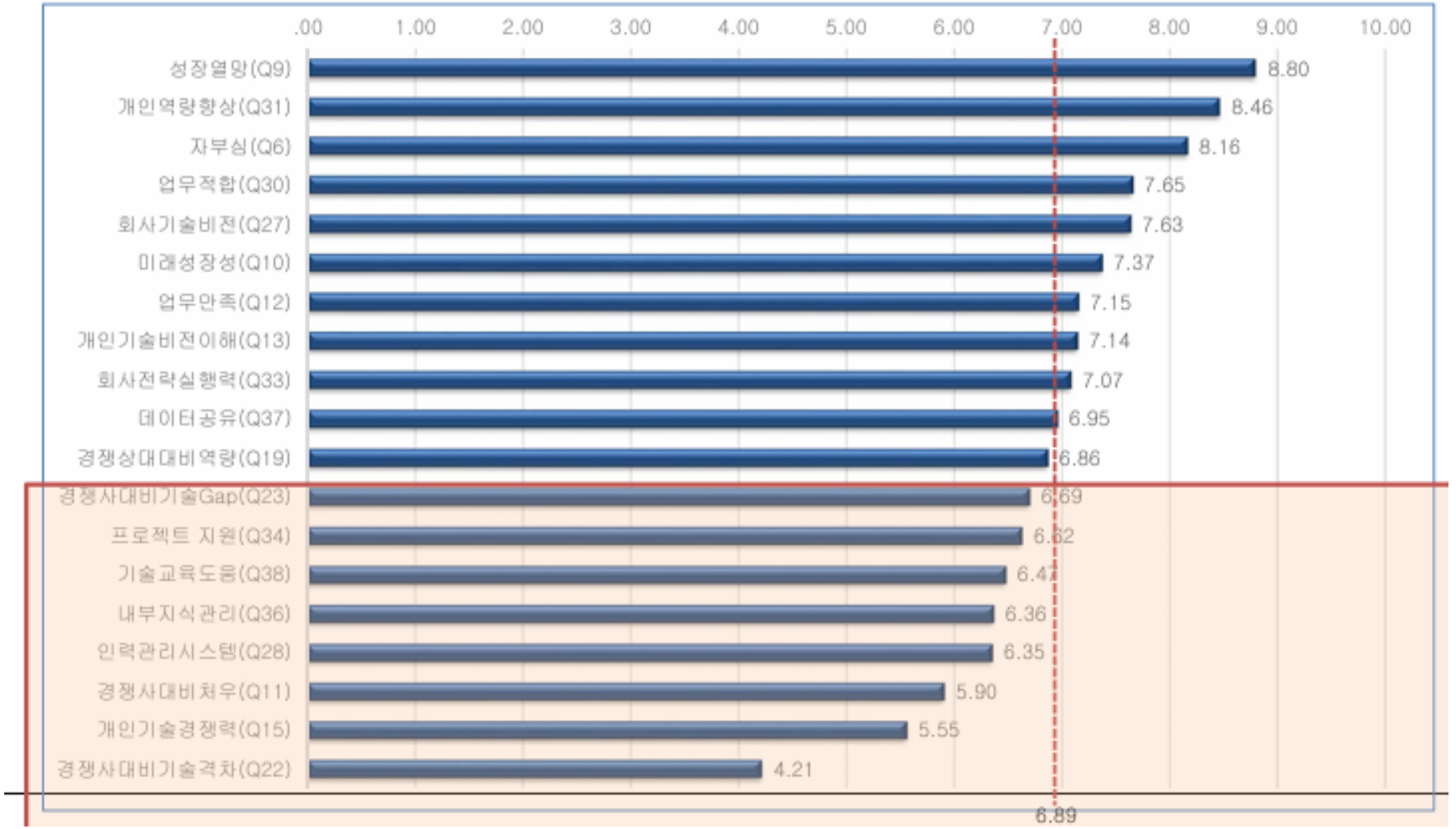
분반	N	Mean	Median	SD	Skewness	Kurtosis
A반	51	56.96	57	7.21	-0.17	-0.29
B반	49	56.59	57	6.41	0.22	-0.2



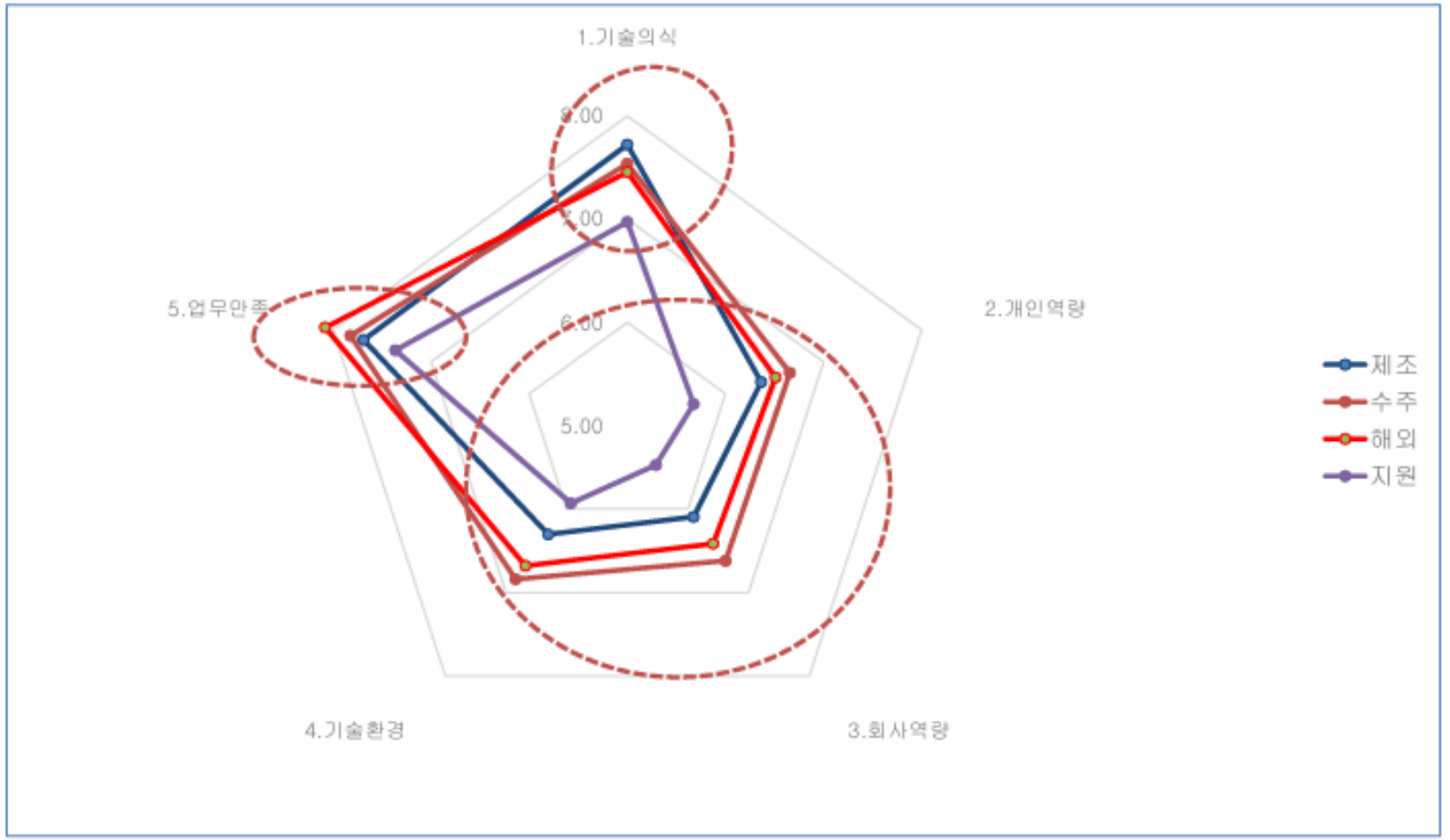
다변량 수치형 자료의 정리



다변량 수치형 자료의 정리

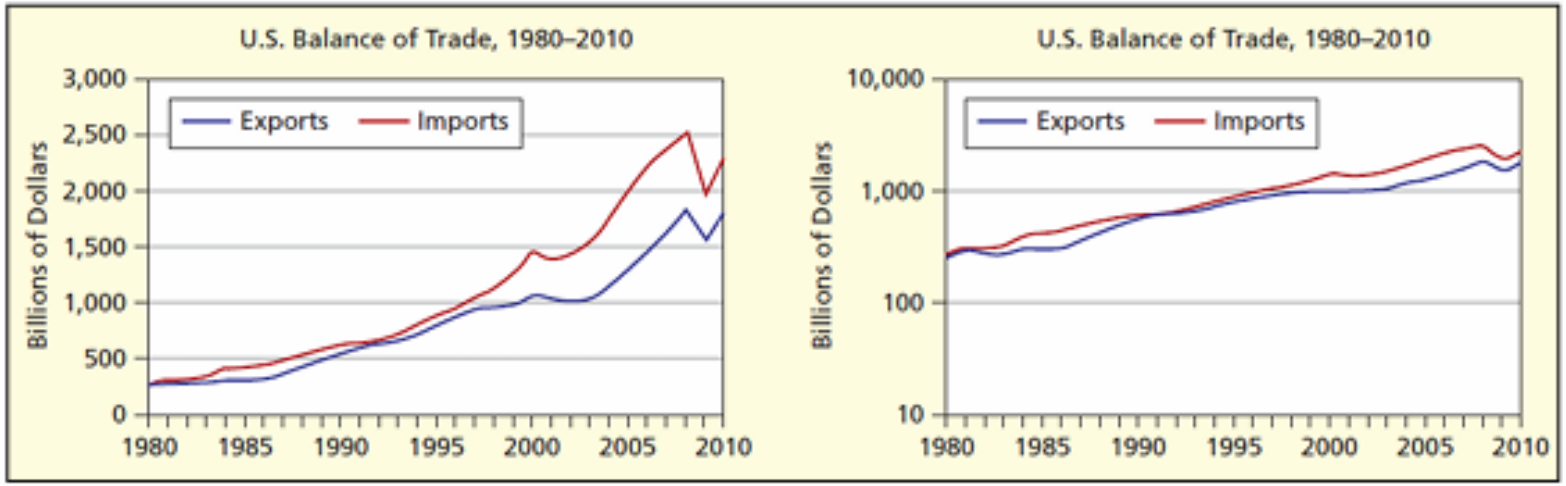


다변량 수치형 자료의 정리



시계열 자료

- Time-series
  - 시간의 변화에 따른 자료
  - 예) 주식, GDP변화, 수출 등 경제관련 지표에서 많이 사용
  - 자료의 값의 변화가 크기 때문에 log 값으로 변환해서 많이 사용



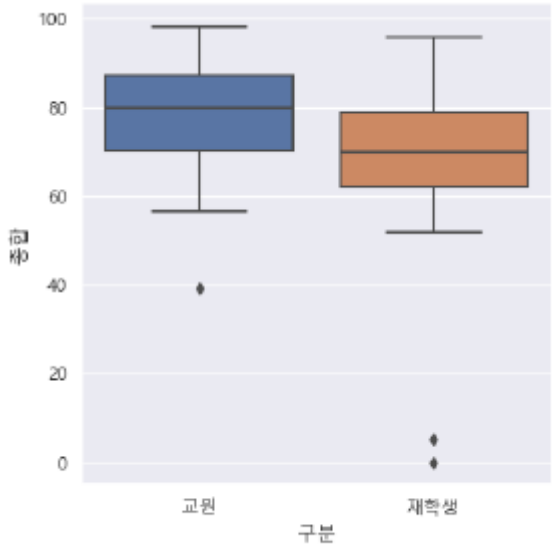
연습문제

- 07.온라인게임.csv를 이용하여 수치형 자료를 분석하세요.
  - Q1.나이(age)의 기초통계분석(rstatix), 히스토그램(bins=10), 상자도표
  - Q2.age의 이상치 제거(±3.0 이상만 제거)
  - Q3.age를 범주형으로 변환(10대, 20대, 30대 등으로 구분), 막대그래프 그리기
  - Q4.성별에 따른 age의 점수를 구하고, box plot 그리기
  - Q5.Design과 Flow의 값 구하고, 산점도로 표시
  - Q6.성별에 따른 Design과 Flow의 값 구하고, 산점도로 표시

데이터 분석 연습문제

- 문제의 정의
  - 대학에서는 재학생(1)과 교원(2)을 대상으로 교육과정에 대한 현황조사를 실시하였다.
  - 종합점수가 재학생과 교원이 차이가 있는 가?
  - 09.edu.csv

	구분	종합
0	재학생	0.0
1	재학생	5.0
2	재학생	70.8
3	재학생	71.6
4	재학생	71.7



	T	dof	alternative	p-val	CI95%	cohen-d	BF10	power
T-test	-3.462	95	two-sided	0.001	[-12.61, -3.42]	0.703	35.774	0.929

Q&A