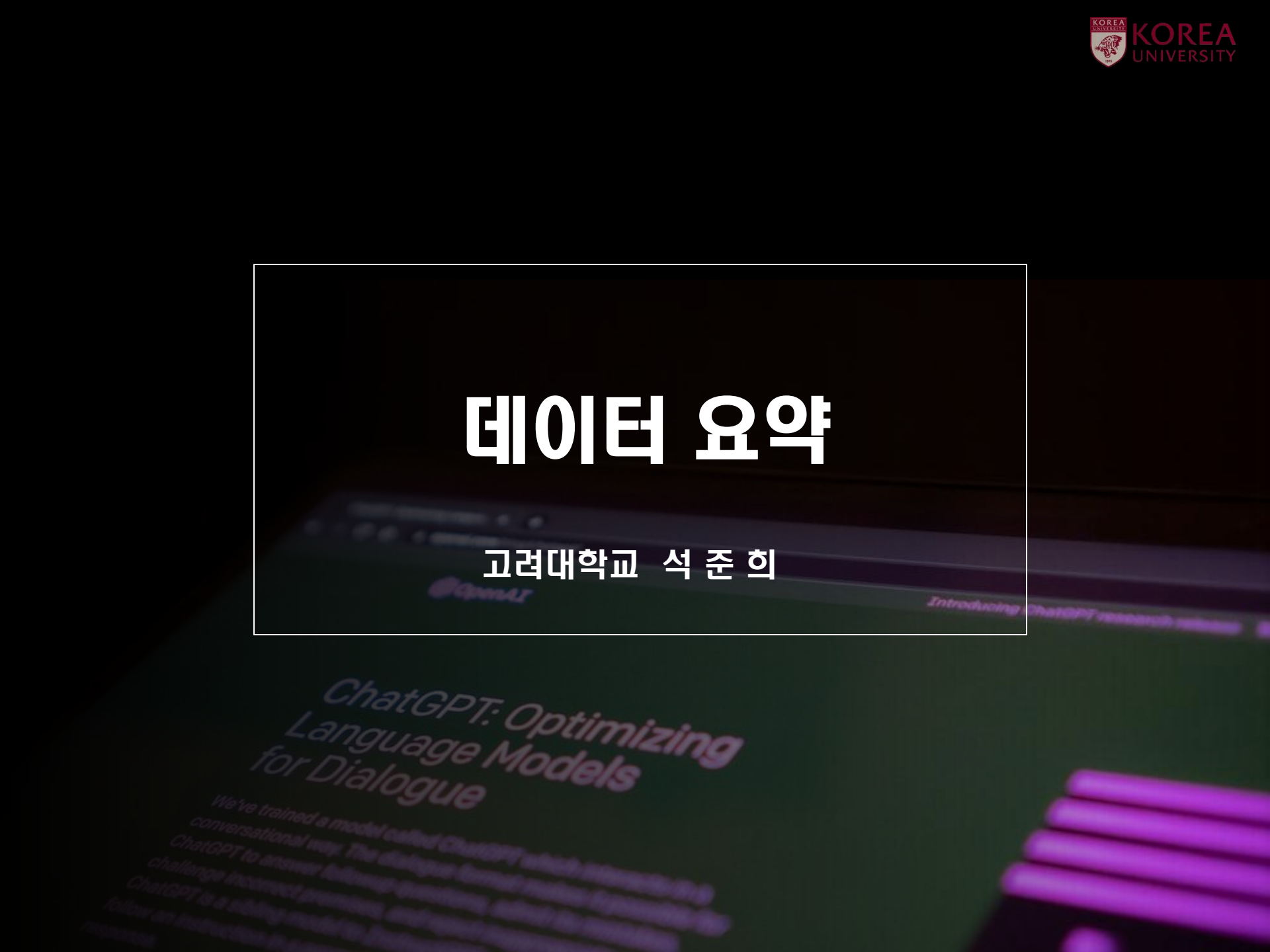
데이터 요약

고려대학교 석 준 희



목차

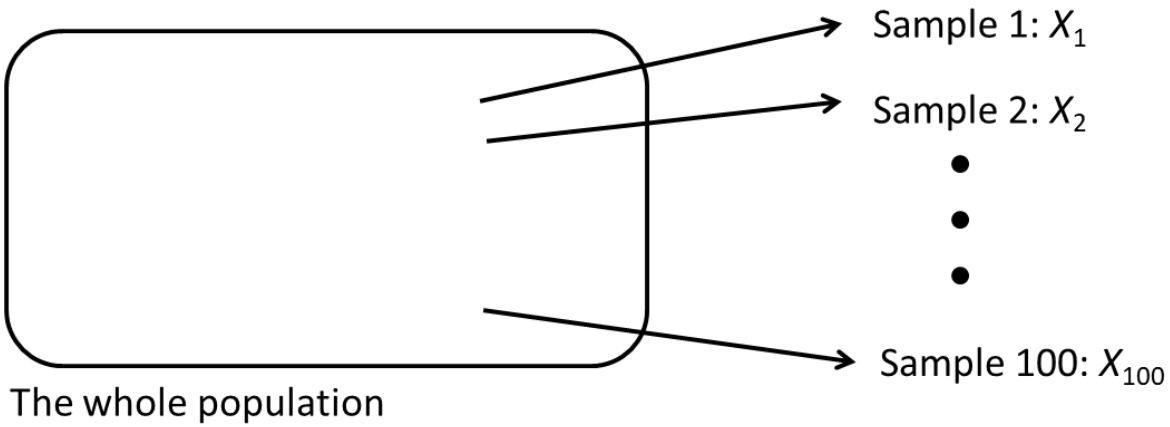
- 데이터의 구분
- 수치형 데이터 요약
- 범주형 데이터 요약
- 이변량 데이터 요약
- 다변량 데이터 요약

데이터 요약

데이터의 구분

임의 관측 데이터

- 데이터란 무엇인가?



- 임의 관측 데이터 (Observed random data)
  - 모든 관측 데이터는 모집단으로부터 임의로 추출된 데이터
  - 데이터를 다시 수집하면 기존과는 비슷한 성질을 지니는 다른 데이터가 수집됨
  - 모집단: 과거부터 현재, 미래까지 발생할 수 있는 무수히 많은 샘플을 갖고 있는 집단으로 생각함. 관측 데이터의 일반화라고도 생각할 수 있음.

- 데이터 과학의 목표: 관측된 데이터로부터 모집단의 정보(일반화된 정보)를 찾는 것

데이터의 종류

- 정형 데이터 (structured data)
  - 일반적으로 표현되는 숫자, 범주 등의 데이터
  - 각 변수가 고유의 특징을 갖고 있음
  - 데이터 행렬의 형태로 표현이 용이 (엑셀로 볼 수 있음) → Tabular data
  - 예: 성별, 판매량, 키, 주소 등
  - 통계, 기계학습, 딥러닝 등으로 분석 가능

- 비정형 데이터 (unstructured data)
  - 정형이 아닌 다른 모든 형태의 데이터
  - 각 변수의 의미를 찾기가 어려움
  - 데이터를 정형의 형태로 변환하여 사용
  - 예: 이미지, 음성, 텍스트 등
  - 주로 딥러닝으로 분석

- 오랫동안 비정형 데이터는 분석이 어려웠지만 최근 딥러닝의 발전으로 분석이 가능해짐

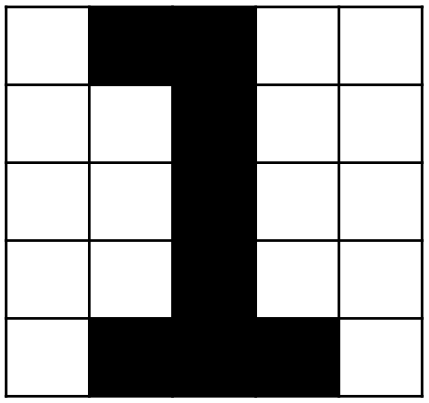
데이터의 종류

- 정형 데이터 (structured data)

	A	B	C	D	E
1	환자ID	성별	나이	키	몸무게
2	1001	남	35	174	76
3	1002	여	29	160	54
4	1003	여	40	163	62
5	1004	남	22	182	91

$$\begin{bmatrix} 1 & 35 & 174 & 76 \\ 2 & 29 & 160 & 54 \\ 1 & 40 & 163 & 62 \\ 2 & 22 & 182 & 91 \end{bmatrix}$$

- 비정형 데이터 (unstructured data)



$$\begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \end{bmatrix}$$

KOREA UNIVERSITY

데이터 행렬 (Data Matrix / Dataframe)

- (정형) 데이터 행렬
  - 일반적인 데이터의 표현
  - 행: 샘플, 표본, 개체
  - 열: 변수, 피처, 항목

- 변수의 수 (p): 데이터 차원
- 샘플의 수 (n)
  - 좁은 의미의 데이터의 크기
- 넓은 의미의 데이터의 크기
  - n x p

	X1	X2	X3	X4	X5	X6
s1	78.4	160.9	M	3	2	상
s2	70.4	167.8	M	2	1	상
s3	56.1	173.5	F	3	1	중
s4	58.8	166.1	F	3	1	하
s5	75.2	174.0	M	3	1	하
s6	63.2	160.0	F	1	2	중
s7	64.4	174.5	F	3	2	하
s8	59.9	179.2	F	2	3	상
s9	50.8	161.0	M	3	1	상
s10	60.7	169.4	M	1	3	하

- 날씬한(skinny) 행렬 (n>>p): 일반적이고 분석이 쉬움
- 뚱뚱한(fat) 행렬 (n<p): 특수한 경우에 발생하고 분석이 어려움

- 차원의 저주 (curse of dimensionality)
  - 데이터의 차원이 높아질 수록 분석이 매우 어려워 짐

데이터의 형태

- 보통은 변수별로 데이터의 형태를 구분

- 수치형 데이터 (numeric data)
  - 데이터가 숫자로 되어 있음
  - 연속형 데이터 (continuous data): 모든 실수값이 가능 (예: 키, 몸무게)
  - 이산형 데이터 (discrete data): 특정 값(보통은 정수)만 가능 (예: 문자 수신 횟수)

- 범주형 데이터 (categorical data)
  - 데이터가 범주(클래스 class, 팩터 factor 등)로 되어 있음
  - 명목형 데이터 (nomial data): 순서가 없음 (예: 성별, 지역, 오류코드 등)
  - 순서형 데이터 (ordinal data): 순서가 있음 (예: 상/중/하, 등급 등)

- 순서-범주형과 이산-수치형은 서로 다름
  - 상/하 vs. 2/1 은 수학적으로 유사하게 다룰 수 있음
  - 상/중/하 vs. 3/2/1 은 불가능
  - 일반적으로 서로 구별됨: 상과 하의 차이는 중과 하의 차이와 같지 않음

데이터의 형태

- 예시
  - 각 데이터는 어떤 형태인가?
  - X4가 "자격증의 수"라면? "내신등급"이라면? "지역코드"라면?

	X1	X2	X3	X4	X5	X6
s1	78.4	160.9	M	3	2	상
s2	70.4	167.8	M	2	1	상
s3	56.1	173.5	F	3	1	중
s4	58.8	166.1	F	3	1	하
s5	75.2	174.0	M	3	1	하
s6	63.2	160.0	F	1	2	중
s7	64.4	174.5	F	3	2	하
s8	59.9	179.2	F	2	3	상
s9	50.8	161.0	M	3	1	상
s10	60.7	169.4	M	1	3	하

데이터 형태에 따른 분석

- 수치형 데이터
  - 연속/이산 모두 동일하게 연속형 데이터로 분석
- 범주형 데이터
  - (기초단계) 명목/순서 모두 동일하게 명목형으로 분석
  - (고급단계) 순서형 데이터를 순서를 고려하는 방법론을 쓰거나 이산-수치형으로 변환하여 분석
    - 변수에 대한 이해가 필요
- 본 강의에서는
  - 연속-수치형 데이터
  - 명목-범주형 데이터
  - 로 나누어 살펴볼 예정

수치형과 범주형 데이터 구분

- 어떻게 구분하나? 보면 안다!
- 변수가 너무 많아 일일이 볼 수 없다면?

- 변수별로 데이터의 형태를 구분
- 보통은 Python에서 데이터를 읽는 함수가 고유의 알고리즘으로 판단
- 일반적인 판단의 절차
  - 값이 문자나 기호의 형태: 범주형
  - 숫자 형태인데 겹치는 값이 거의 없을 때: 수치형 데이터
    - 100개의 표본이 98개의 서로 다른 값을 갖는다
  - 숫자 형태인데 겹치는 값이 많은 때: 범주형 데이터에 대한 코드인지 확인
    - 100개의 표본이 1과 2 두 개의 값 만을 갖는다

- Rule of thumb: 데이터에 대해서 잘 알고 있는 사람에게 물어본다!!

데이터 요약 (Data Summary)

- 데이터 요약의 목적
  - 데이터의 크기(n x p)는 매우 크기 때문에 데이터를 하나하나 확인하기 어려움
  - 데이터를 단순하게 요약해서 분석가가 이해하기 쉽게 만들어야 함
  - 일반적으로 p가 작기 때문에 각 변수 별로 큰 n을 요약하는 방식으로 진행
    - 큰 n을 1~2개의 숫자로 요약하는 것이 목적 (평균, 분산 등)
  - p가 큰 경우 다수의 변수를 차원축소를 통해 요약하는 것도 가능

- 단변량(univariate) 데이터 요약: 하나의 변수에 대한 요약
- 이변량(bivariate) 데이터 요약: 두 개의 변수 사이의 관계에 대한 요약

- 요약값의 계산: 다수(n)의 데이터를 1~2개의 숫자로 표현
- **시각화(visualization)**: 데이터의 형태를 시각적으로 이해하기 쉽게 표현

데이터 요약

수치형 데이터 요약

대표값 (Representative Value)

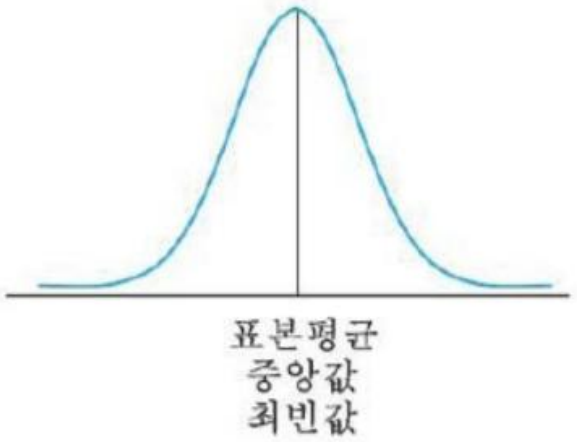
- 대표값: 하나의 변수에 다수(n)의 데이터가 있을 때 이들 데이터를 대표하는 하나의 값

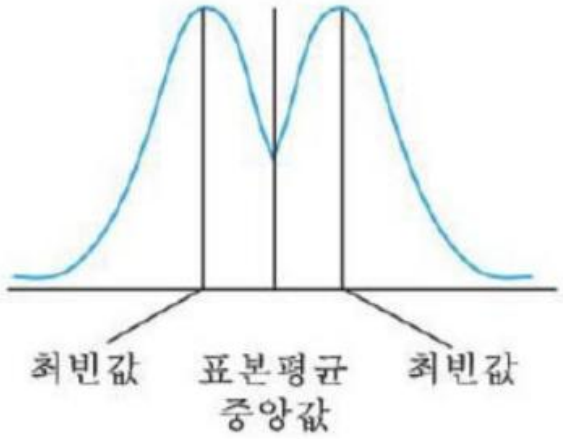
- (표본)평균 (sample mean, average): 각 데이터의 산술 평균, $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
  - 선형적으로 계산되어 이론 연구나 수학적 조작에 용이
  - 극단적인 값들에 영향을 많이 받음
- 중간값, 중앙값 (median): 데이터를 크기 순으로 나열했을 때 중앙에 있는 값
  - 극단적인 값들의 영향이 적어 값이 안정적임 (e.g. 신용카드 사용액 등)
  - 서열이 있는 명목형 데이터에도 사용 가능
- 최빈값 (mode): 가장 많이 나타나는 값
  - 찾기 쉽지만 포함하는 정보가 적음
  - 명목형 데이터에도 사용 가능

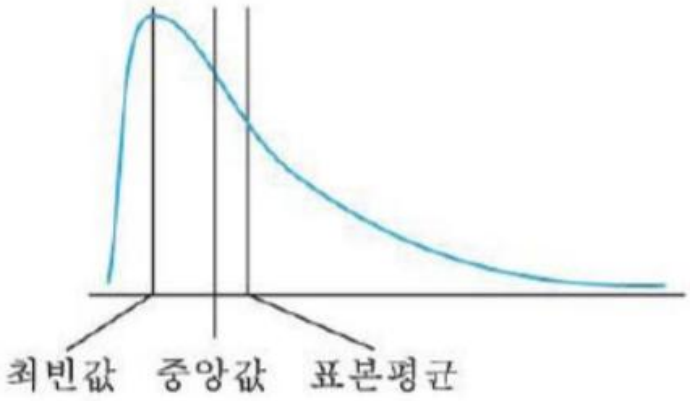
- 예제
  - 2, 3, 5, 7, 7 의 데이터에서 평균, 중간값, 최빈값은?
  - 2, 3, 5, 7의 중간값은?
  - 1, 2, 3의 평균과 중간값은? 1, 2, 3, 1000의 평균과 중간값은?

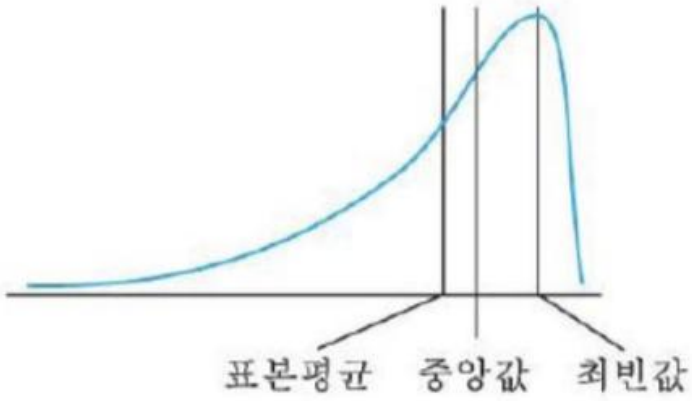
대표값 (Representative Value)

- 데이터 분포에 따른 대표값









분포의 표현

- 데이터 변화의 정도를 표현하는 값

- (표본)분산(sample variance): 각 값이 평균으로부터 얼마나 벗어나 있는지 표현

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$ S: 표준편차 (standard deviation)

- 최대값(max), 최소값(min)
- 백분위수(percentile) / 사분위수(quantile): 값들을 크기 순으로 세웠을 때 하위 %에 해당하는 값
  - 25% Percentile (1st Quartile): 아래에서부터 25%에 해당하는 값
  - 50% Percentile (2nd Quartile): 아래에서부터 50%에 해당하는 값
  - 75% Percentile (3rd Quartile): 아래에서부터 75%에 해당하는 값
  - 0%, 100% Percentile: 최소값, 최대값

- 예제
  - 2, 3, 5, 7, 7 의 데이터에서 최대/최소, 각 percentile은?
  - 1, 2, 3의 데이터에서 분산은?

시각화 – 히스토그램 (Histogram)

- 도수분포표 (frequency table): 수치형 데이터의 개수를 구간별로 나누어 표시한 표

학생들의 키

(단위 : cm)

147	169	145	128
163	132	153	156
169	135	128	141
139	159	149	145
138	151	146	153

<정리되지 않은 자료>

학생들의 키

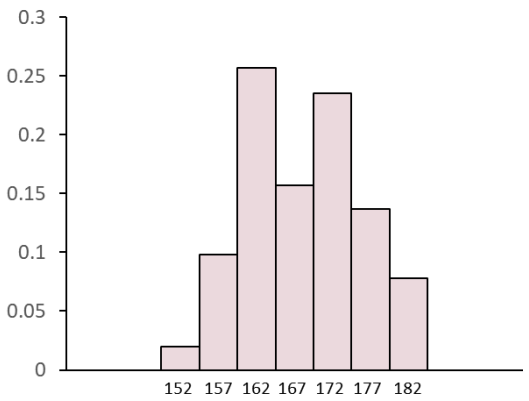
키 (cm)	학생 수 (명)
120 이상 ~ 130 미만	2
130 ~ 140	4
140 ~ 150	6
150 ~ 160	5
160 ~ 170	3
합 계	20

<도수분포표>

- 히스토그램 (histogram): 수치형 데이터의 도수분포표를 막대 그래프로 표현

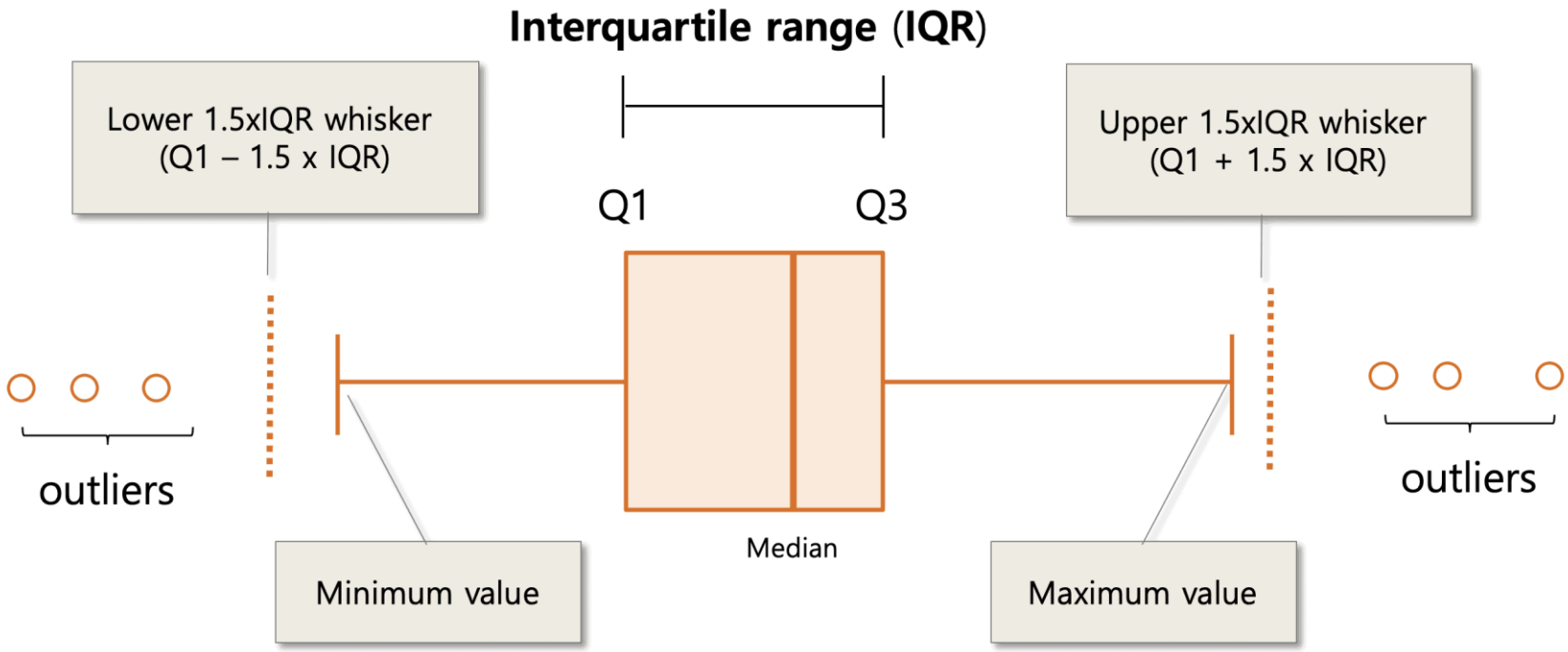
통계학과 신입생의 키에 관한 도수분포표

계급	계급구간(cm)	도수	상대 도수
1	149.5 ~ 154.5	1	0.020
2	154.5 ~ 159.5	5	0.098
3	159.5 ~ 164.5	14	0.257
4	164.5 ~ 169.5	8	0.157
5	169.5 ~ 174.5	12	0.235
6	174.5 ~ 179.5	7	0.137
7	179.5 ~ 184.5	4	0.078
합 계		51	1.000



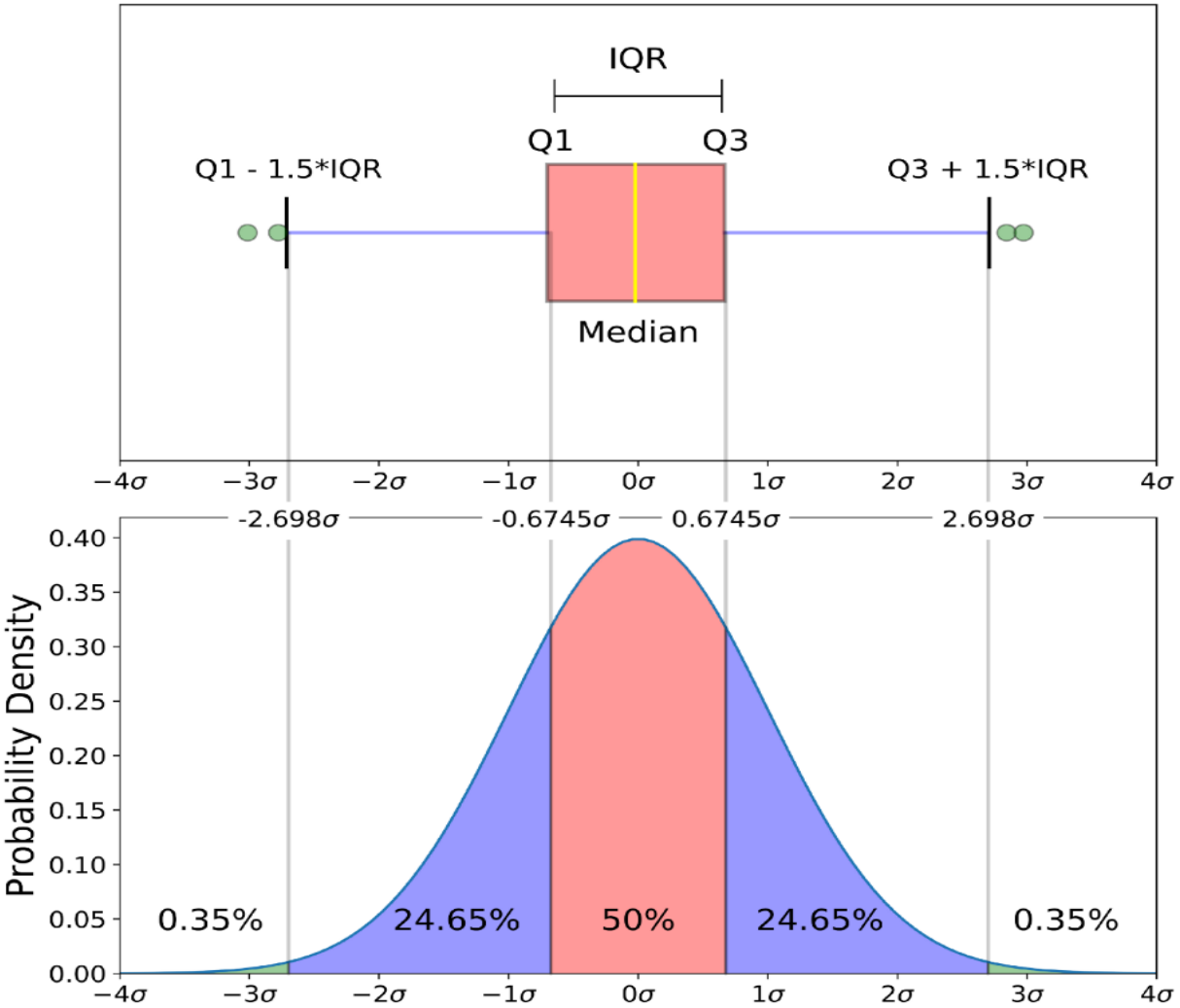
시각화 – 박스플롯 (Boxplot)

- 연속형 변수의 분포를 박스의 형태로 표현하는 시각화
  - 다양한 백분위수와 이상치(outlier)들이 표현
  - 평균도 같이 표현되기도 함



시각화 – 박스플롯 (Boxplot)

- 이상치는 정규분포에서 99.3% 바깥 쪽에 해당하는 데이터들



데이터 요약

범주형 데이터 요약

대표값 (Representative Value)

- 대표값: 하나의 변수에 다수(n)의 데이터가 있을 때 이들 데이터를 대표하는 하나의 값
- 범주형 데이터의 경우 일반적으로 최빈값을 사용
  - 평균의 계산이 불가, 순서가 있는 범주형의 경우 중간값의 사용이 가능
- 예제
  - A, B, B, A, D, C, B, A, B 의 데이터에서 대표값은?

분포의 표현

- 데이터 변화의 정도를 표현하는 값
- 도수분포표 (frequency table)
  - 범주형 데이터의 개수를 범주별로 나누어 표시한 표

혈액형	도수	상대 도수 (%)	각도 (◦)
A	22	36.67	132
B	20	33.33	120
AB	7	11.67	42
O	11	18.33	66
합	60	10.00	360

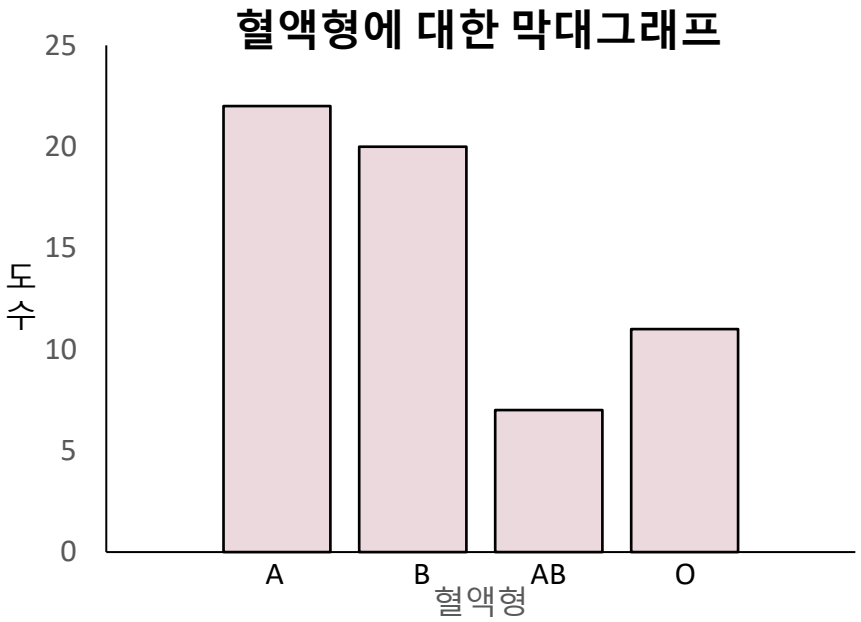
- 예제: 위의 표에서 대표값은?

시각화 – 막대 그래프 (Bar Graph)

- 범주형 데이터의 도수 분포를 막대 그래프로 표현
  - 범주의 순서는 상관없음
- 히스토그램: 수치형 데이터의 구간별 도수 분포를 막대 그래프로 표현
  - 구간의 순서는 변할 수 없음

혈액형에 대한 도수분포표

혈액형	도수	상대 도수 (%)	각도 (°)
A	22	36.67	132
B	20	33.33	120
AB	7	11.67	42
O	11	18.33	66
합	60	100.00	360



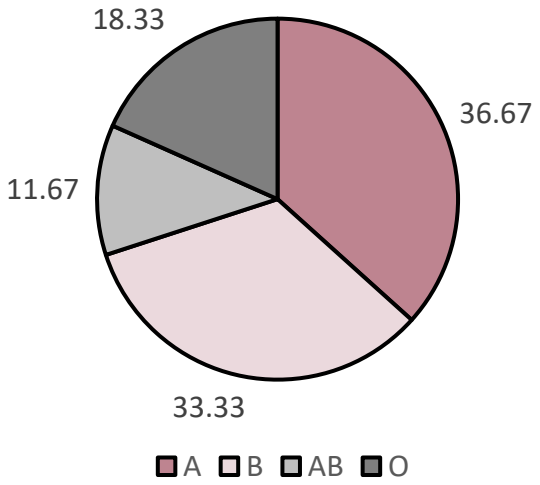
시각화 – 원 그래프 (Pie Chart)

- 범주의 도수에 비례하는 부채꼴로 도수 분포를 표현

혈액형에 대한 도수분포표

혈액형	도수	상대 도수 (%)	각도 (°)
A	22	36.67	132
B	20	33.33	120
AB	7	11.67	42
O	11	18.33	66
합	60	100.00	360

혈액형에 대한 원형 그래프



데이터 요약

이변량 데이터 요약

이변량 데이터 요약 (Bivariate Data Summary)

- 변수 자체보다는 변수와 변수 사이의 관계를 요약하는 것이 중요
- 이변량(bivariate) 데이터 요약: 두 개의 변수 사이의 관계를 요약
- 다변량(multivariate) 데이터 요약: 3개 이상의 변수 사이의 관계를 요약

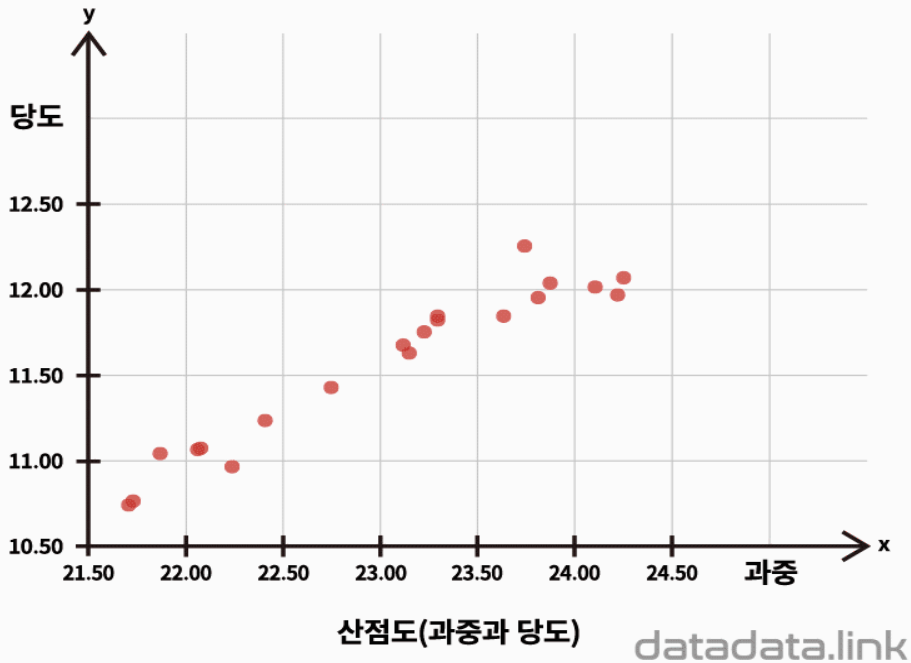
- 이변량 데이터 요약의 분류 (위: 시각화, 아래: 요약값)

	수치형	범주형
수치형	산점도 상관계수	박스플롯 SMD
범주형	박스플롯 SMD	모자이크 플롯 오즈비

수치-수치: 산점도 (Scatter Plot)

- 두 변수의 데이터를 이차원 좌표 상에 점으로 표현
- 두 수치형 변수 사이의 관계를 직관적으로 확인 가능

딸기ID	과중	당도
1	24.21	11.98
2	24.28	12.08
3	23.88	12.03
4	23.85	11.89
5	23.73	12.24
6	23.17	11.60
7	24.14	12.02
8	23.63	11.85
9	23.37	11.85
10	23.37	11.80
11	23.27	11.73
12	23.19	11.68
13	22.78	11.41
14	22.25	10.96
15	22.13	11.18
16	21.86	11.08
17	22.10	11.16
18	21.72	10.75
19	21.69	10.68
20	22.42	11.21



수치-수치: 상관계수 (Correlation Coefficient)

- 공분산(covariance): 두 변수 사이의 관계를 단순히 나타내는 값으로, 두 변수가 얼마나 같이 변하는지를 표현
  - 두 변수가 같이 방향으로 변함 (큰 양의 값), 두 변수가 서로 다른 방향으로 변함 (큰 음의 값), 두 변수의 변화가 관련 없음 (0에 가까운 값)
  - 스케일에 따라 값이 변함. 한 변수가 10배가 되면 공분산도 10배가 됨

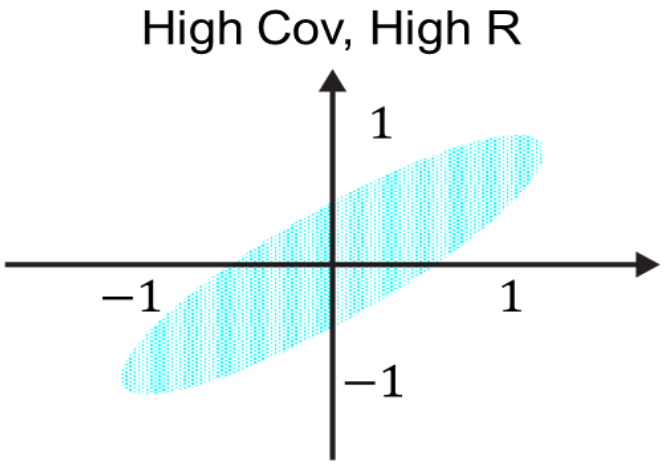
- **상관계수(correlation coefficient)**
  - 공분산의 분산으로 스케일하여 변수의 스케일에 따라 값이 변하지 않음
  - 일반적으로 -1과 1사이에 값을 가짐

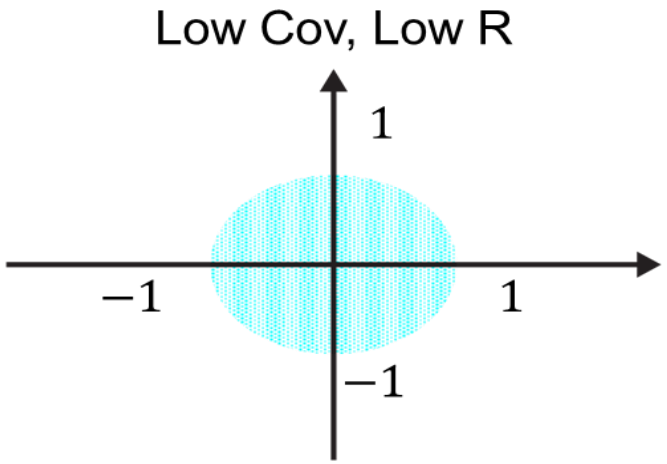
$$r_{X,Y} = \frac{S_{XY}}{S_X S_Y} = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

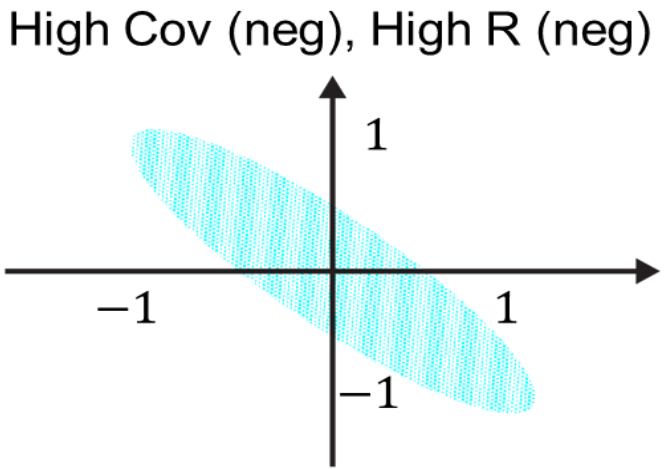
- 공분산 vs. 상관계수
  - Cov[Height(cm), Weight(kg)] vs. Cov[Height(m), Weight(g)]
  - 공분산은 다르지만 상관계수는 같음

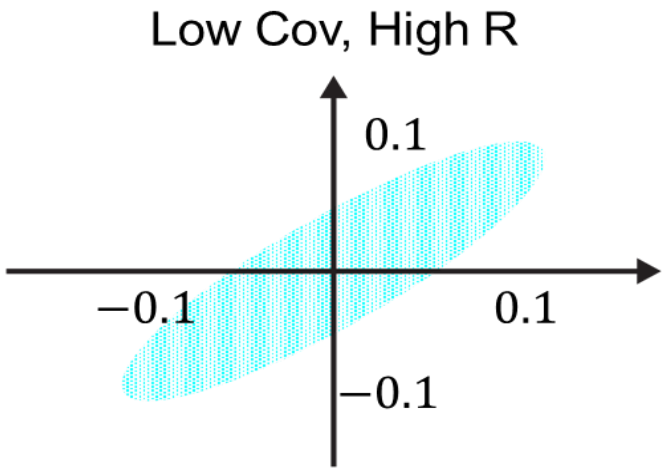
수치-수치: 상관계수 (Correlation Coefficient)

- 공분산은 스케일의 영향을 받지만 상관계수는 그렇지 않음









수치-범주: 데이터의 표현

- 아이리스(iris, 붓꽃) 데이터 셋
  - 꽃잎(petal)과 꽃받침(petal)의 크기를 나타내는 수치형 변수 4개와 품종(species)을 나타내는 범주형 변수 1개로 구성

- 데이터 행렬 (첫 4개의 표본)

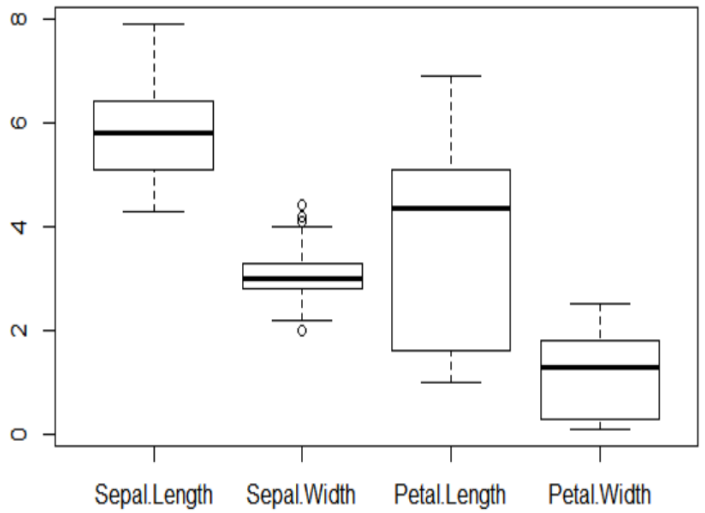
	A	B	C	D	E
1	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
2	5.1	3.5	1.4	0.2	setosa
3	4.9	3	1.4	0.2	setosa
4	4.7	3.2	1.3	0.2	setosa
5	4.6	3.1	1.5	0.2	setosa

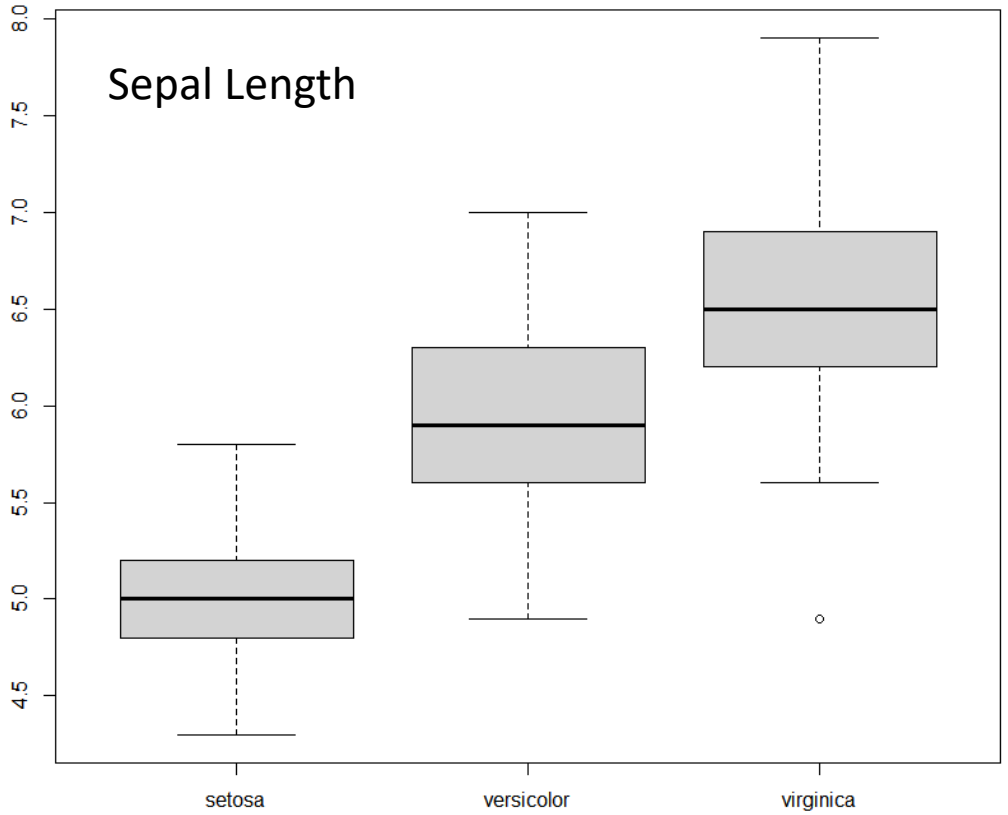
- 단변량 데이터 요약

```
 Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa     :50
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
```

수치-범주: 시각화

- 각 범주별로 수치형 데이터를 각각 시각화
  - Sepal Length는 Species와 어떤 관계가 있는가?





수치-범주: SMD (Standardized Mean Difference)

- 일반적으로 두 범주의 수치형 값의 평균이 크게 차이가 날수록 두 변수의 관련성이 높다고 생각됨
  - 성별-키 와 성별-IQ 중에서 어떤 것이 관련성이 높은가? 그 이유는?
  - 얼마나 큰 것이 큰 것인가?

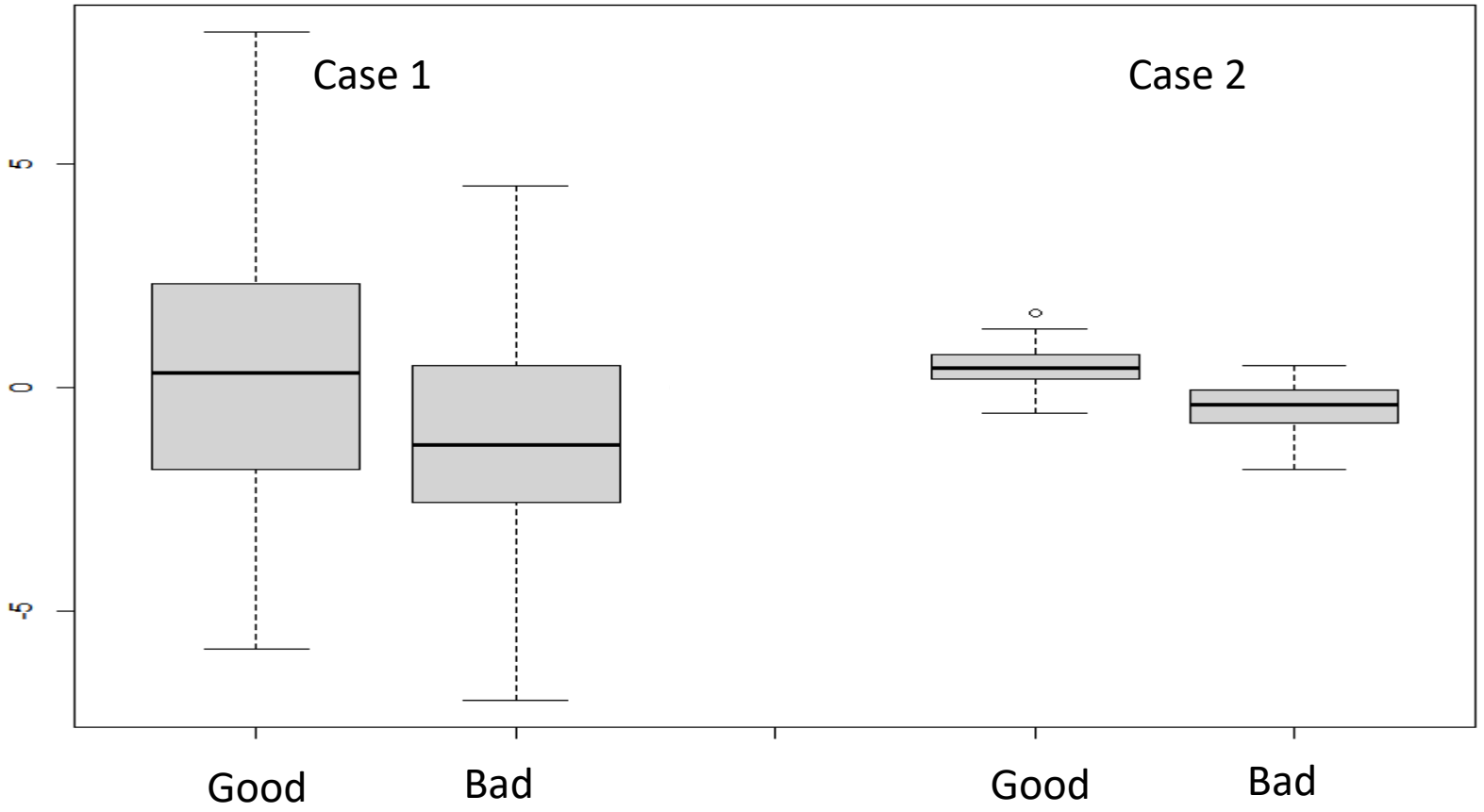
- SMD (Cohen's d): 평균의 차이를 표준편차로 스케일한 값
  - $\bar{X}_1$, S_1^2: 첫 번째 범주의 수치형 값의 평균과 분산
  - $\bar{X}_2$, S_2^2: 두 번째 범주의 수치형 값의 평균과 분산

$$d = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}}$$

- SMD는 범주가 2개일 때만 사용 가능

수치-범주: SMD (Standardized Mean Difference)

- 예제: 어느 쪽에 더 관련성이 높은가?
  - Case 1: 평균의 차이=1.5, SMD=0.5
  - Case 2: 평균의 차이=0.5, SMD=1.0



범주-범주: 크로스 테이블 (Cross Table)

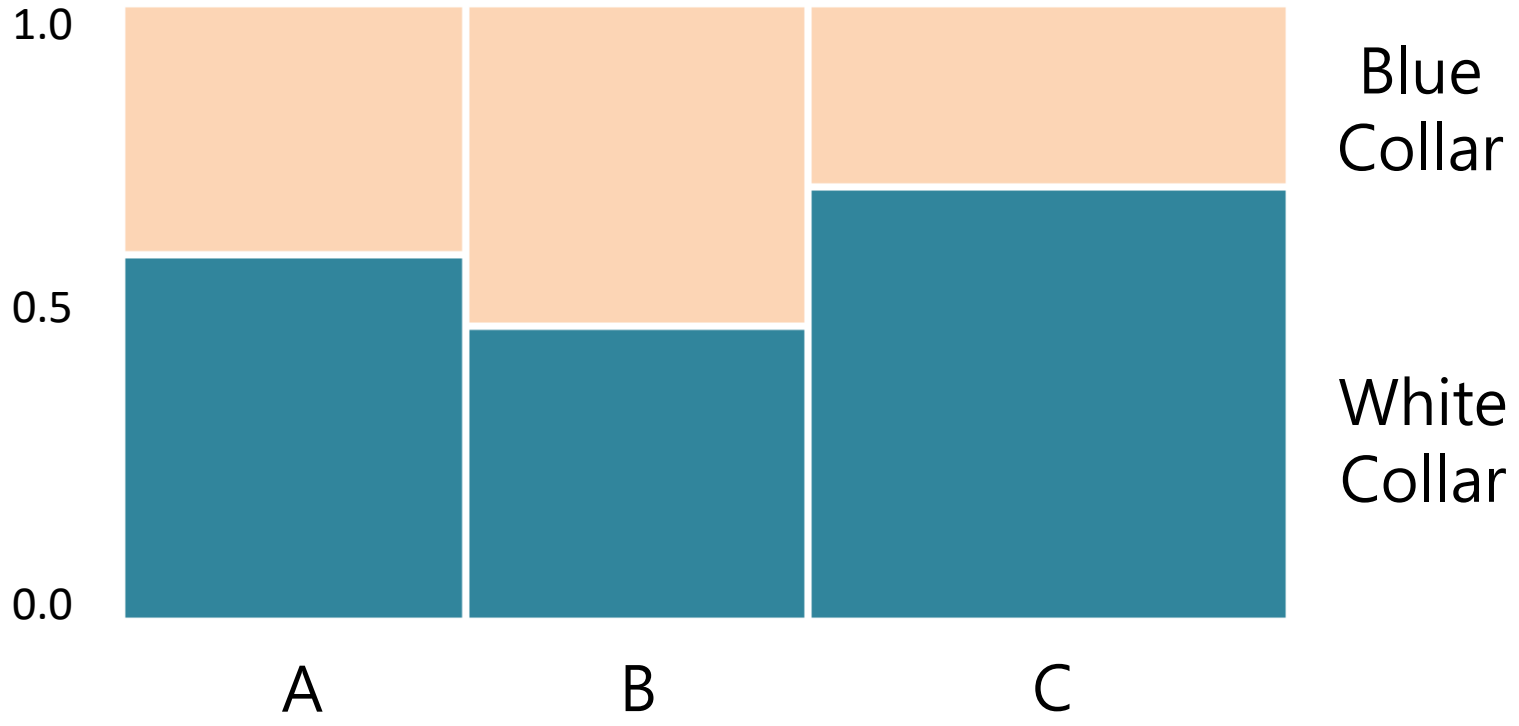
- 크로스 테이블: 두 범주형 데이터에 대한 결합(joint) 도수분포표

- 예제: 소비자 그룹 vs. 직업의 형태
  - X1: 소비 패턴에 따른 그룹 A, B, C
  - X2: 직업의 형태 (화이트/블루)

	A	B	C	Total
White collar	90	70	130	290
Blue collar	60	80	70	210
Total	150	150	200	500

범주-범주: 시각화

- 모자이크 플롯(mosaic plot): 두 범주형 데이터에 대한 시각화
  - 기준 변수(소비자 그룹)의 각 범주별로 다른 변수의 범주를 비례적으로 표현
  - 기준 변수의 도수에 따라 너비가 결정



범주-범주: 요약 (2x2의 경우)

- 범주형 변수가 2개의 범주를 갖고 있을 때
  - **승산비** (odds ratio): "두 범주형 변수가 서로 관련이 있는가?"에 대한 수치적 값
  - 1에 가까울 수록 관련이 없고, 1에서 멀어질 수록 관련이 높음
  - Log(승산비), log odds ratio or logit 가 일반적으로 사용

Above Average	Yes	No
Yes	A – Positives	C – Negatives
No	B - Negatives	D - Positives

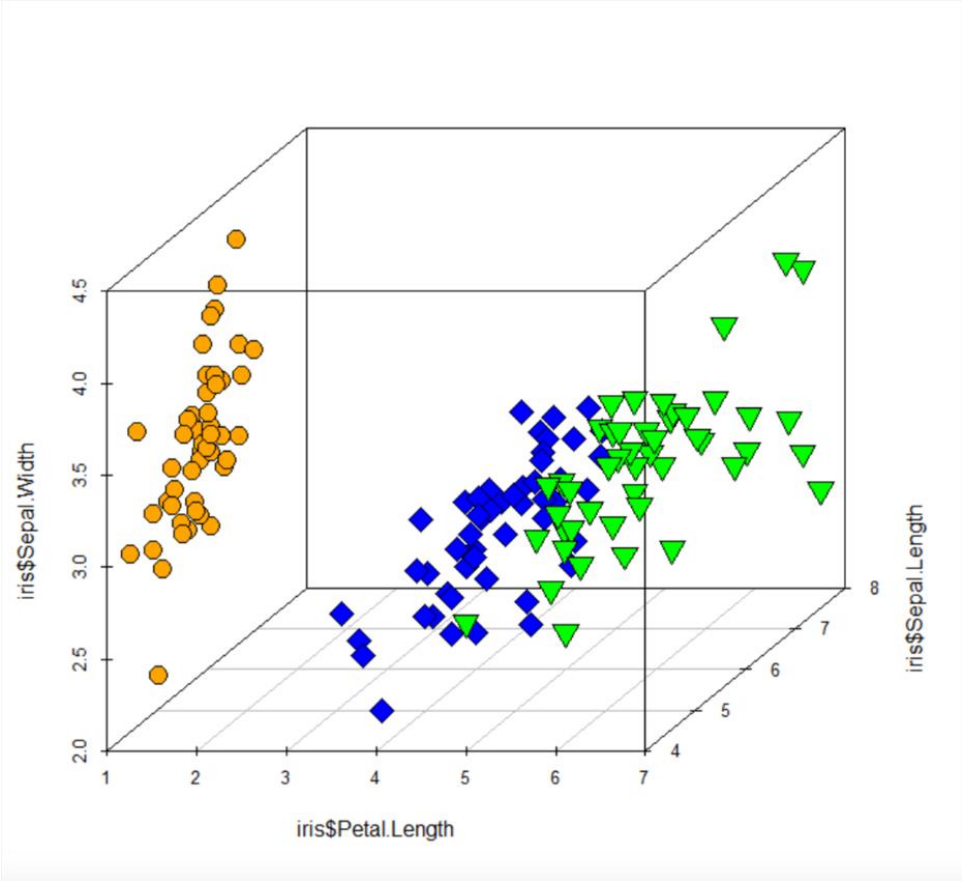
$$\widehat{OR} = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{a \cdot d}{b \cdot c}$$

다변량 데이터 요약

다변량 데이터 요약

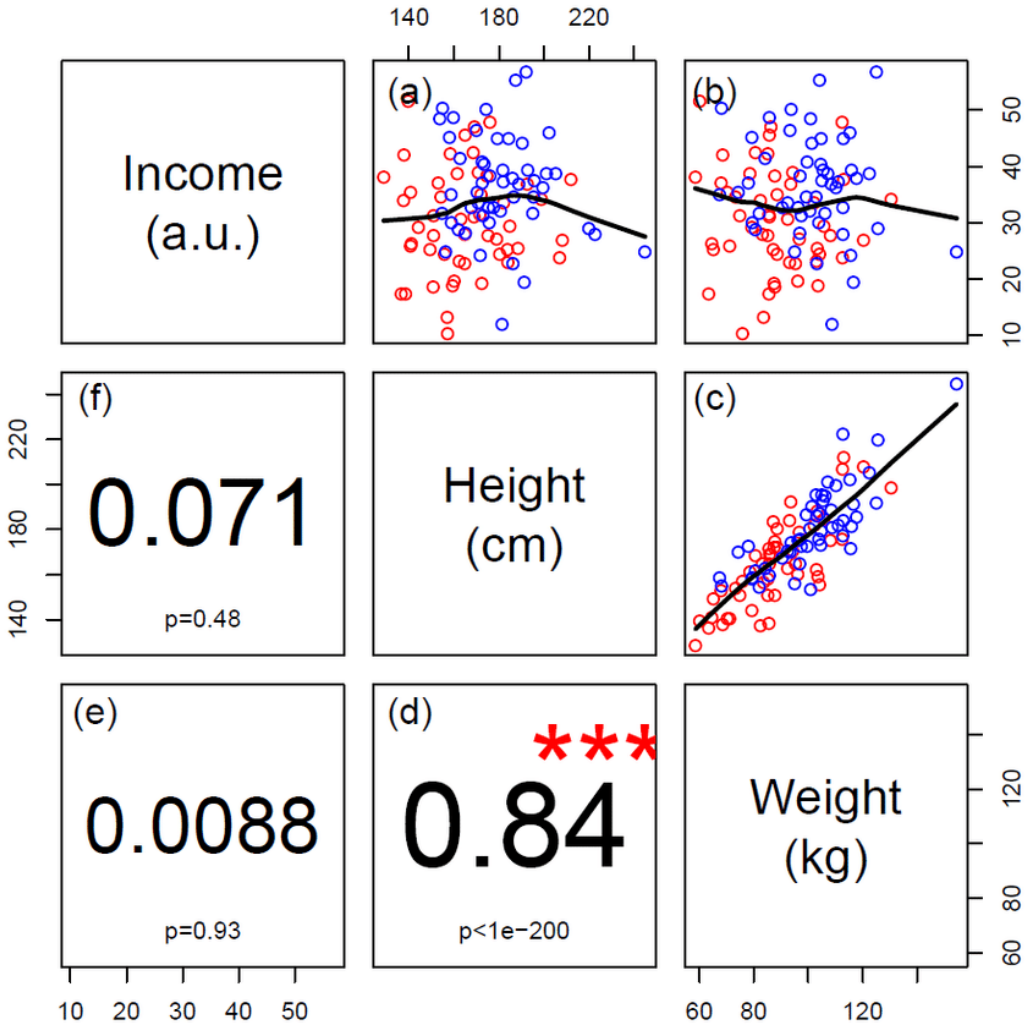
- 앞선 방법들을 응용하여 3개 이상의 변수를 요약

- 3차원 산점도: 3개의 수치형 변수의 시각화
- 산점도 포인트의 색과 모양: 2개의 범주형 변수를 표현



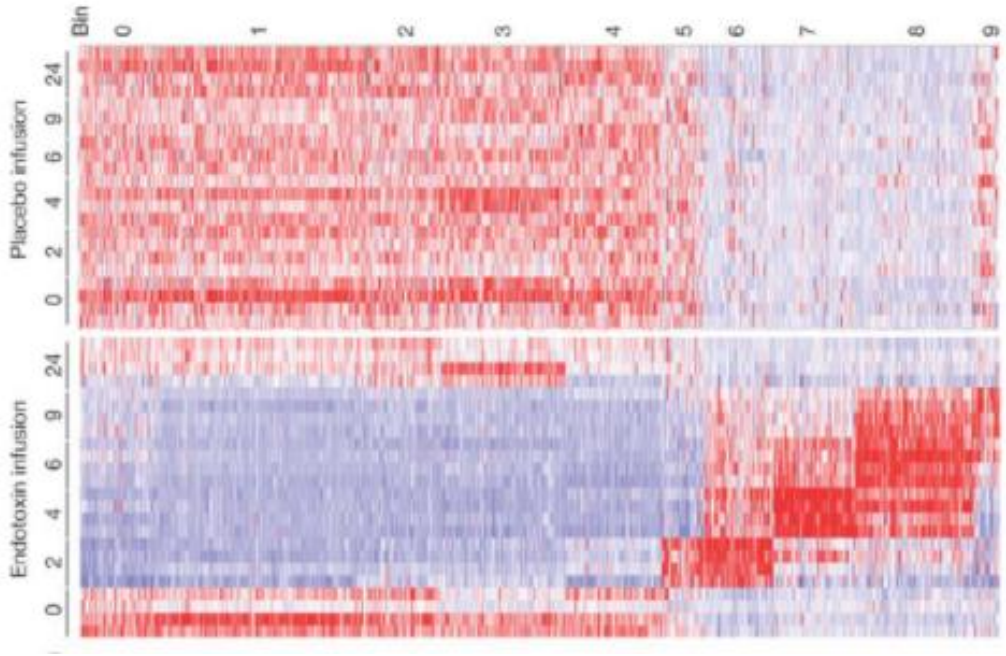
다변량 데이터 요약

- Pairwise plot: 여러 변수를 각 쌍에 대해서 시각화

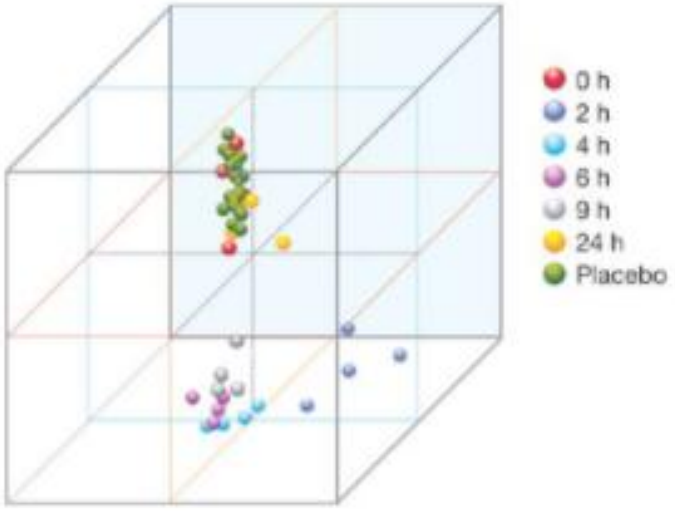


다변량 데이터 요약

- 데이터를 간단화하여 요약 및 시각화 진행
- 차원축소 (dimension reduction): 변수의 수를 축소
  - 예: 48x6000 데이터 행렬 → 48x3 데이터 행렬



히트맵



PCA 플롯

KOREA UNIVERSITY

감사합니다