통계학개론
- 제 2 장 자료의 구조(1)

0011 0010 1010 1101 0001 0100 1011

12
45

자료의 종류

- 자료의 형태에 따른 종류
  - 자료의 형태에 따른 대분류
    - 양적 자료 (quantitative data)
      - 관측된 값의 크기에 관심 – 수치 자료 (numerical data)
      - 사칙연산이 가능
      - 측도(measure)를 이용하여 측정
    - 질적 자료 (qualitative data)
      - 관측된 값의 내용에 관심 – 범주형 자료 (categorical data)
      - 개체가 속한 범주 판별

- 자료의 형태에 따른 소분류
  - 명목척도자료 (nominal scaling data)
    - 단지 구분하기 위한 부호로 표시된 자료
    - 예) 성별, 혈액형, 주거형태, 운동선수의 등번호 등
  - 서수척도자료 (ordinal scaling data)
    - 자료들 사이의 크기를 비교하여 순서를 나타낸 자료
    - 예) 학점, 학년, 선호도 등
  - 구간척도자료 (interval scaling data)
    - 자료들 사이의 크기가 의미를 갖는 자료
    - 절대 0의 의미가 없음
    - 예) 성적, 온도 등
  - 비율척도자료 (ratio scaling data)
    - 자료들 사이의 크기가 의미를 갖는 자료
    - 절대 0의 의미가 있음
    - 예) 키, 몸무게 등

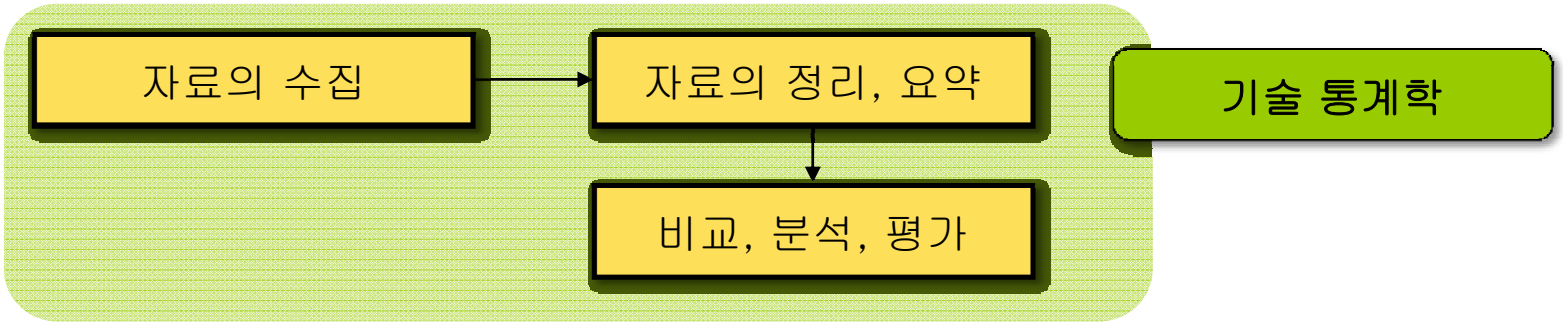
자료의 구성

- 자료의 구성 요소
  - 개체(observation)
    - 조사/실험을 통한 자료의 수집의 기본 단위
    - 측정 단위는 일치 되어야 함
  - 변수(variable)
    - 통계적으로 관심이 되는 어떠한 특성
    - 변수값 : 변수가 취할 수 있는 값
    - 관측값
      - 개체로부터 실제로 측정한 변수값
      - 변수값 중 일부를 취하며 같은 값을 반복하여 취할 수 있음

개 체	변 수	변 수 값
개 인	종 교	불교, 천주교, 기독교, …
라 면 상 자	불 량 품 수	0, 1, 2, …, 20
기 업	종 업 원 수	1, 2, 3, …

자료의 요약

- 자료의 정리 및 요약



- 자료에 포함된 정보를 쉽고 빠르게 파악 가능
- 도표에 의한 자료의 요약
  - 도수분포표 활용
- 그림에 의한 자료의 요약
  - 질적 자료의 경우 : 막대 그래프, 원형 그래프 등을 활용
  - 양적 자료의 경우 : 히스토그램, 줄기-잎 그림 등을 활용
- 수치에 의한 자료의 요약
  - 사칙연산이 가능한 양적 자료의 요약 기법

질적 자료의 요약

- 도표에 의한 질적 자료의 요약
  - 도수분포표(frequency table)
    - 각 범주에 속하는 관측값의 개수를 파악
    - 전체에서 차지하는 각 범주의 비율 파악
    - 범주간 비교 용이
    - 범주와 그 범주에 대응하는 도수와 상대도수 등을 나열한 표
      - 도수 :
        특정 범주에 속한 관측값의 수(빈도, frequency)
      - 상대도수 :
        해당 범주의 도수를 자료 전체 개수로 나눈 비율
      - 누적도수 :
        이전 범주들과 현재 범주의 도수 합
      - 누적상대도수 :
        해당 범주의 누적도수를 자료 전체 개수로 나눈 비율

- 예 – [A 고등학교 학생들의 혈액형 분포]
  - 한 학급 임의로 선택하여 60명에 대한 혈액형 검사 결과

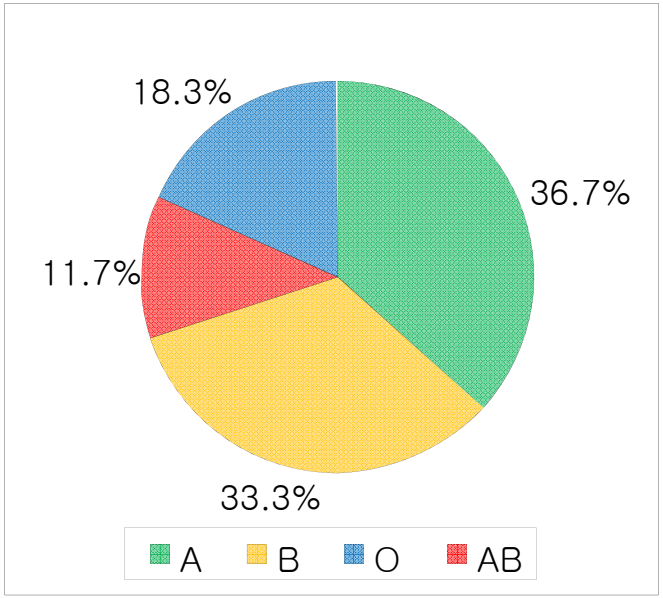
A	B	B	A	A	O	A	AB	O	O	A	B	O	AB	B
B	A	O	B	A	B	B	O	AB	B	A	AB	A	B	A
O	A	A	B	AB	A	O	B	A	B	B	A	B	A	B
AB	B	A	O	AB	O	B	A	B	A	O	B	A	A	A

  - 도수분포표 작성

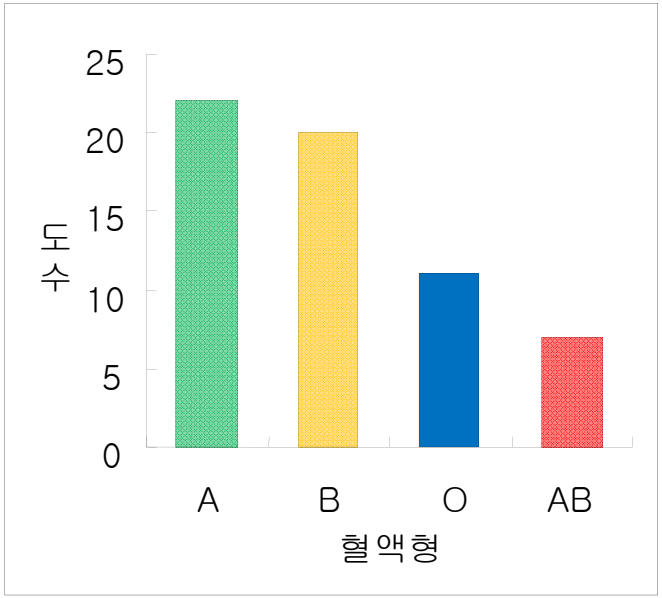
혈 액 형	도 수	상 대 도 수	누 적 도 수	누 적 상 대 도 수
A	22	0.367	22	0.367
B	20	0.333	42	0.700
O	11	0.183	53	0.883
AB	7	0.117	60	1.000
합계	60	1.000		

- 그림에 의한 질적 자료의 요약
  - 원형 그래프(pie chart)
    - 원을 상대도수에 비례하여 중심각을 나눈 그래프
    - 각 범주가 차지하는 비율 파악 및 비교 용이
    - 각 범주의 상대도수에 360을 곱하여 원조각의 각을 계산
  - 막대 그래프(bar chart)
    - 각 범주가 가지는 도수 또는 상대도수를 같은 폭의 직사각형 막대모양으로 그린 그래프
    - 막대의 높이가 해당 범주의 도수 또는 상대도수를 나타냄

- 예 – [A 고등학교 학생들의 혈액형 분포]
  - 원형 그래프



  - 막대 그래프



양적 자료의 요약

- 도표에 의한 양적 자료의 요약
  - 도수분포표
    - 도수분포표 작성 요령
      - 최대값과 최소값을 찾아 그 차이를 계산
      - 자료의 크기와 속성에 따라 계급의 개수를 5~15개로 결정
        - 계급의 개수가 너무 적으면 정보 손실이 많음
        - 계급의 개수가 너무 많으면 계급의 도수 경향 파악 어려움
        - 자료의 속성 감안하여 주관적으로 개수 결정
      - 계급의 개수만큼 동일 간격의 계급구간 계산
        - 계급구간의 폭 $= \dfrac{\text{최대값} - \text{최소값}}{\text{계급 개수}}$
        - 최소값과 최대값을 반드시 포함하여야 함
        - 하나의 관측값이 두 개의 계급구간에 포함되면 안됨
        - 관측값이 계급구간의 경계점에 놓이지 않도록 설정
      - 각 계급구간에 속하는 관측값의 개수를 세어 도수/상대도수 파악

- 예 – [통계학과 신입생의 키]
  - 신입생 51명을 임의 추출 후 측정 (단위 : cm)

181 M	161 F	170 F	160 F	158 F	169 F	162 F	179 M
183 M	178 M	171 M	177 M	163 F	158 F	160 F	160 F
158 F	174 M	160 F	163 M	167 F	165 F	163 M	173 M
178 M	170 M	167 M	177 M	176 M	170 M	152 F	158 F
160 F	160 F	159 F	180 M	169 M	162 F	178 M	173 M
173 M	171 M	171 M	170 M	160 F	167 M	168 F	166 F
164 F	174 M	180 M					

    - 최대값 : 183, 최소값 : 152, 최대값–최소값 : 31
    - 계급의 개수를 7개로 한다면 : 31 / 7 = 4.4 (계급구간의 폭 = 5)
    - 계급 구간의 시작값 = 152 – (5 / 2) = 149.5

- 예 – [통계학과 신입생의 키]
  - 도수분포표

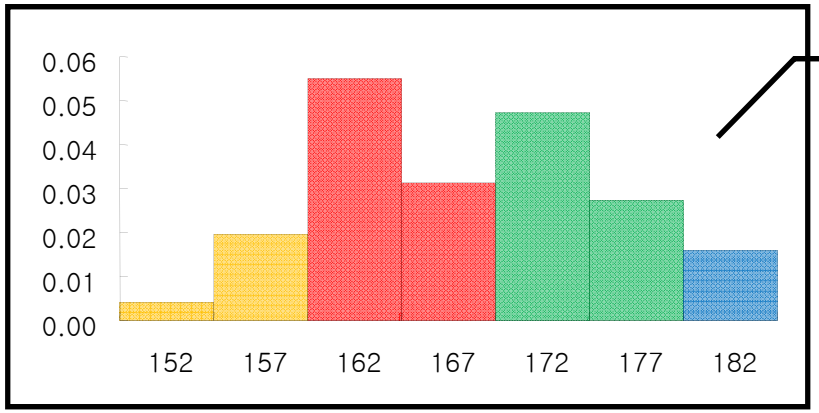
계 급 구 간	도 수	상대도수	누적 상대도수
149.5 이상 154.5 미만	1	0.020	0.020
154.5 이상 159.5 미만	5	0.098	0.118
159.5 이상 164.5 미만	14	0.275	0.392
164.5 이상 169.5 미만	8	0.157	0.549
169.5 이상 174.5 미만	12	0.235	0.784
174.5 이상 179.5 미만	7	0.137	0.922
179.5 이상 184.5 미만	4	0.078	1.000
합 계	51	1.000	

- 그림에 의한 연속형 자료의 요약
  - 히스토그램(histogram)
    - 막대그래프와 유사
    - 막대를 서로 붙여 밑변을 채워서 그림
      - 히스토그램은 막대의 넓이가 의미를 가짐 (전체 면적 = 1)
      - 히스토그램의 높이 : 해당 상대도수 / 계급구간의 폭
      - cf) 막대그래프의 높이 : 해당 범주의 도수
    - 두 개의 히스토그램을 연결하여 두 개의 자료 분포 비교 가능

  - 예 – [통계학과 신입생의 키]
    - 계급구간의 폭을 5로 할 경우 히스토그램의 높이

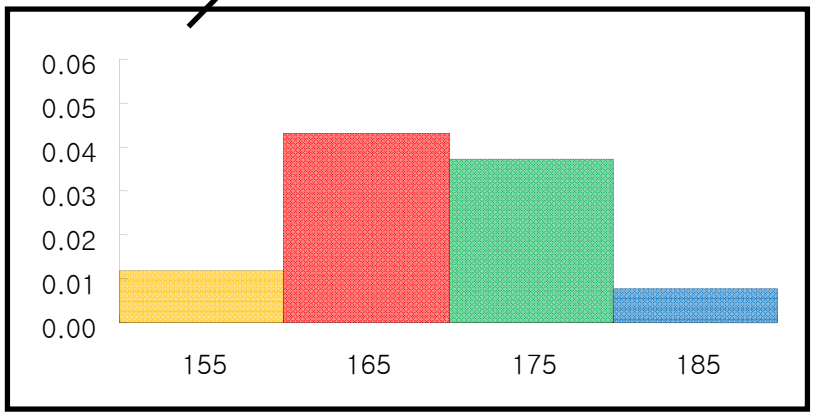
계급 구간	149.5–154.5	154.5–159.5	159.5–164.5	164.5–169.5	169.5–174.5	174.5–179.5	179.5–184.5
높이	0.004	0.020	0.055	0.031	0.047	0.027	0.016

- 예 – [통계학과 신입생의 키]

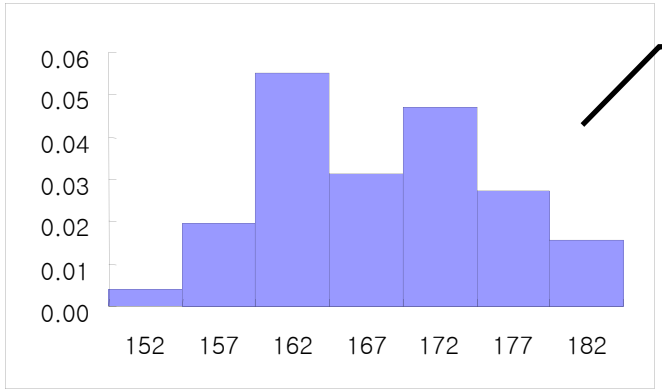


계급구간 폭이 5인 경우

계급구간 폭이 10인 경우

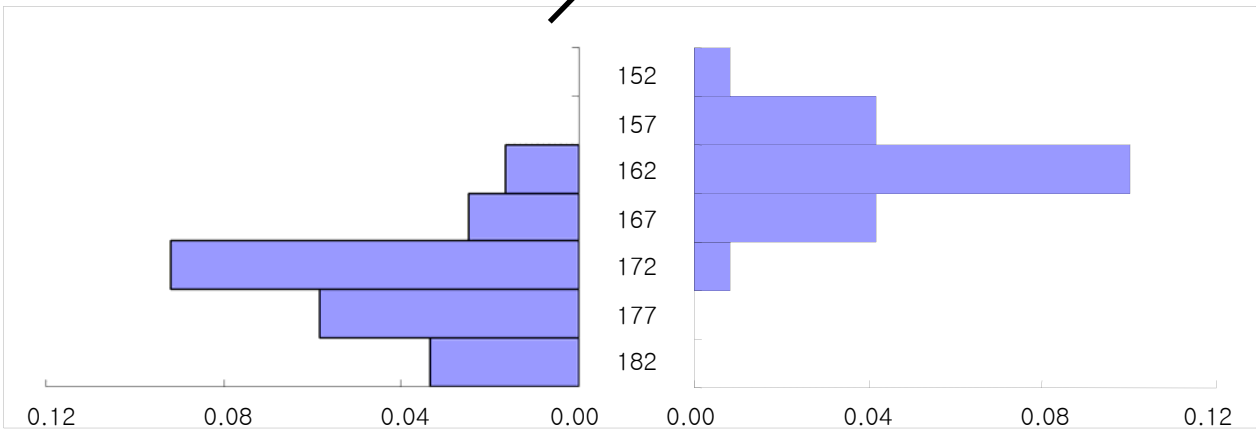


- 예 – [통계학과 신입생의 키]



51명 전체 대상인 경우

성별에 따른 분류를 한 경우



- 줄기-잎 그림(stem-and-leaf plot)
  - 자료의 분포를 시각적으로 쉽게 파악
  - 각 관측값에 대한 정보를 유지
  - 줄기-잎 그림 작성 요령
    - 관측값의 자리수를 고려하여 줄기와 잎을 결정
    - 줄기 값을 오름차순으로 세로 배열
    - 줄기 값 옆에 수직선 그림
    - 수직선 옆에 각 줄기의 잎에 해당하는 숫자를 오름차순으로 배열
  - 줄기의 간격이 너무 넓다고 판단될 경우 줄기 세분화
    - - : 잎에 해당하는 숫자가 0 ~ 4
    - + : 잎에 해당하는 숫자가 5 ~ 9
  - 두 개의 줄기-잎 그림을 연결하여 두 개의 자료 분포 비교 가능
    - 관측값 개수의 차이가 많이 나는 두 개의 집단을 비교할 겨우 히스토그램을 이용하는 것이 좋음

- 예 – [통계학과 신입생의 키] 결과 III

15	288889
16	00000001223334 56777899
17	000011133344 6778889
18	0013

15^-	2
15^+	88889
16^-	00000001223334
16^+	56777899
17^-	000011133344
17^+	6778889
18^-	0013

	15^- 2
	15^+ 88889
33	16^- 000000012234
977	16^+ 56789
44433111000	17^- 0
9888776	17^+
3100	18^-