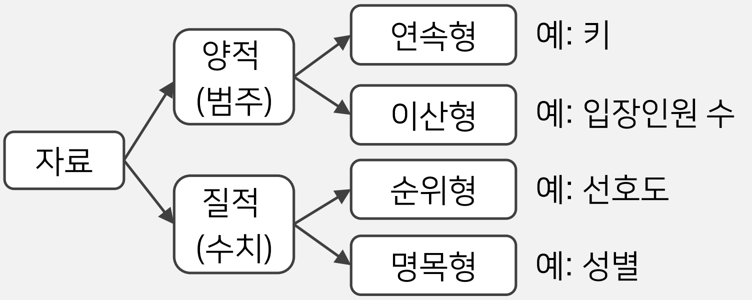
이전 내용

자료의 종류



자료를 요약하는 초기단계,

1. **양적 자료**: 수치 값을 이용해 분산과 같은 **통계적 계산**을 이용
2. **질적 자료**: 범주에 따른 **빈도수를 이용**해 자료를 요약 정리

1. 범주형 자료 요약

* 범주: 동일한 성질을 가진 부류나 범위

❶도수(frequency)

범주형 자료를 요약할 때는 각각 범주에 속하는 관측 값의 개수 파악

❷도수분포표(frequency table)

범주형 자료에 대해 각각의 범주와 그에 대응되는 도수를 나열한 표

▶ 도수분포표를 작성하는 것은 범주형 자료에 대한 가장 기본적인 요약기법

❸상대도수(relative frequency)

도수를 자료 전체 개수로 나눈 비율로, 백분율(%)로 표현할 수 있음

ex)

성별	❶도수	❷상대도수	❸상대도수(%)
여	45	0.45	45%
남	55	0.55	55%
합계	100	1.00	100%

▶ **범주형 자료를 요약할 때 사용할 수 있는 효과적인 그래프로는 원도표(pie chart)와 막대도표(bar chart)가 있음**

원도표

원을 그린 뒤, 상대도수에 비례하게 중심각을 나누어 그린 그림

범주가 차지하는 비율을 파악하기 용이 함

그러나, **범주상의 도수를 비교하거나 도수의 차이를 파악하기 힘듦**

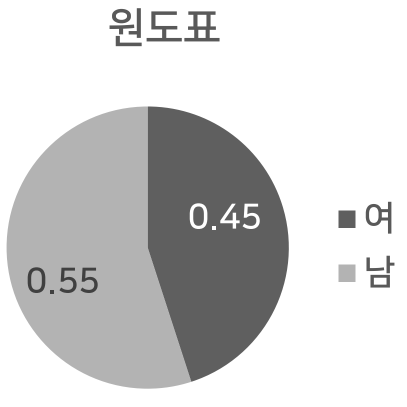
▶ 이 경우 막대도표 사용

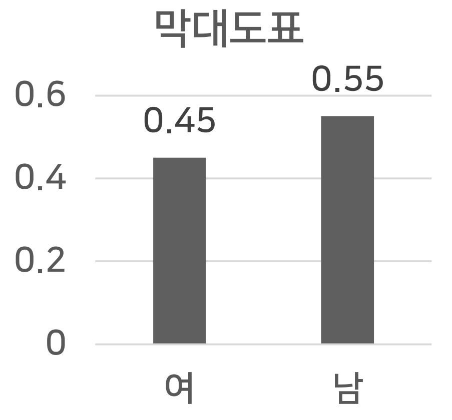
막대도표

범주에서 도수의 크기를 막대의 높이로 표현함

각 범주가 차지하는 비율을 확인하고 싶다면 도수보다 상대도수를 사용함

도수와 상대도수 그림의 모양은 같으며, **범주간의 도수를 비교하는데 용이**





2. 이산형 자료 요약

* 이산: 연속적이지 않음

관측값을 측정할 때 세어서 파악한 자료로, 관측값 중 중복 값이 많고 적음에 따라 요약하는 기법이 다름

- **중복되는 값이 많은 경우, 범주형 자료의 요약기법 사용**
- **중복되는 값이 적은 경우, 연속형 자료 요약기법 사용**

예시

각 가정의 자녀수를 조사하는 경우, 0, 1, 2, 3, 4의 범주를 가지고, **범주 별 중복된 값의 도수가 구해짐** ① 도수 분포표를 작성 ② 막대도표나 원도표로 요약 가능	100명 키 조사(소수점 둘째 자리) 도수분포표를 작성하면 **범주 별 중복 값 없을 수 있음.** 도수분포표의 범주 별 도수가 1인 도수분포표가 작성될 수 있음
★범주형 요약기법 적용	**★연속형 요약기법 적용**

3. 연속형 자료 요약(시각화)

연속형 자료는 반올림되어 정수 값으로 표현되기도 하지만, 이산형 자료와 달리 실제값은 실수 값으로 표현될 수 있음(ex. 키의 소수점 둘째 자리). 따라서, 연속형 자료는 관측 값들 중에서 중복되는 값이 많지 않을 수 있음

일반적으로 관측 값의 종류가 많기 때문에 최소값부터 최대값까지 범위를 구간으로 나누어 구간에 포함되는 관측 값의 개수를 도수로 표현함

계급(class)
관측 값이 몇 개의 구간으로 나누어진 부분

계급구간(class interval)
계급에 포함되는 구간

예시

계급구간 경계점 계급구간의 폭 계급의 개수: 2개

A B

키	161	162	163	164	165	166	167	168	169	170
명	1	3	1	0	0	2	1	1	1	0

▶

계급	계급구간	도수
A	161 ~ 165	5
B	166 ~ 170	5

계급구간 설정방법

① 최대값과 최소값의 차이를 계산하여 모든 관측 값의 범위를 파악함
② 계급의 개수로 나누어 계급구간의 폭을 결정함

계급 개수 결정에 특별한 법칙은 없음

- 계급의 개수 ↓, 계급 구간의 폭 ↑, 구간의 도수 ↑ ▶ 많은 정보를 잃음
- 계급의 개수 ↑, 계급 구간의 폭 ↓, 구간의 도수 ↓ ▶ 경향을 찾기 어려움

▶ 자료의 성향을 파악하고 도수 분포 경향이 드러날 수 있도록 선택해야 함
▶ 연속 자료형 도수분포표 표현에는 많은 시행착오를 거쳐야 함

추가

계급구간의 경계점

계급의 폭에 따라 모든 관측 값을 포함하도록 설정
관측값이 계급의 경계점에 놓이지 않도록 하는 것이 바람직
(관측 단위보다 한 단계 아래 단위로 잡기도 함)

계급구간 시작점↔최소값, 계급구간 종료점↔최대값의 거리가 비슷해야 함

첫번째 계급구간의 시작점에 최소값이 위치하게 되면 마지막 계급구간에 최대값이 포함되지 못하고. 한 단계 앞의 계급 구간에 최대값이 위치할 수도 있기 때문

히스토그램(histogram)

연속형 자료에서 계급에 대해 적용하며, 범주형의 막대도표와 유사한 모양의 그림

① 연속형 자료에 대해 **계급구간 사이의 도수 비교** 가능

② **계급구간에 따른 도수 변화의 경향(자료 분포)을 쉽게 알 수 있음**

* 막대도표는 막대의 높이가 도수 혹은 상대도수를 나타내 범주간의 도수 비교 가능

③ 막대도표와 달리 **막대의 넓이가 상대도수**를 나타냄. 따라서 **전체 면적은 항상 1**

④ **막대의 높이**는 상대도수를 계급구간의 폭으로 나눠 구할 수 있음
(막대 높이 = 상대도수 / 계급구간의 폭)

⑤ **계급 구간 폭이 모두 동일**한 경우, 상대도수가 높이를 대체하여도 **같은 모양으로 표현** 됨

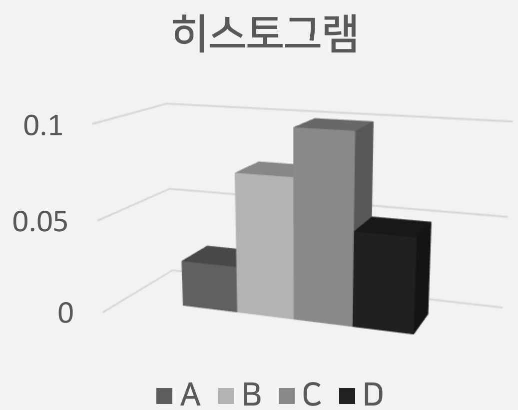
⑥ **계급구간의 폭이 일정하지 않으면 서로 다른 모양으로 표현**되므로 주의 필요

히스토그램의 막대 개수 = 도수분포표 상 계급의 개수

예시

* 계급구간 폭:4

계급	계급구간	도수	상대도수	높이
A	1 이상 5 미만	10	0.10	0.025
B	6 이상 10 미만	30	0.30	0.075
C	11 이상 15 미만	40	0.40	0.100
D	16 이상 20 미만	20	0.20	0.050
-	합계	100	1	0.250



4 (히스토그램 막대 개수) = 4 (도수분포표 상 계급 수)

4. 연속형 자료 요약(수치활용)

시각적 요약은 일관성과 객관성이 부족하고, 통계적 추론에서 요구되는 이론적 근거를 제시하는 것이 어려움

표본평균(sample mean)

중심위치 측도 중에서 가장 많이 사용되는 방법

표본평균 = 관측 값의 총합 / 관측 값의 개수

통계적 추론과정에서 광범위하게 사용되며, 통계적 분석에 서 가장 기초적인 수치임
그러나, 모든 관측값이 반영되기 때문에 극단적으로 아주 크거나 작은 값에 영향을 많이 받아 잘못된 중심 위치를 나타냄

중위수(median)

전체 관측값을 크기순으로 정렬했을 때, 가운데 위치하는 값

중위수 = 전체 관측값 * 0.5

예시 관측 값: 5, 8, 13, 7, 10, 15

표본평균
= 5 + 8 + 13 + 7 + 10 + 17 / 6
= **10**

중위수
5 7 8 10 13 17
= (8 + 10) / 2 = **9**

편차(deviation)

중심으로 각각의 관측 값들이 얼마나 흩어져 있는지 파악

편차 = 관측값 – 표본평균

표본분산(sample variance)

편차의 합은 항상 0이 되므로, 편차의 제곱합을 구한 뒤 관측값의 개수에서 1을 뺀 값으로 나누면, 단 하나의 수치로 얼마나 흩어져 있는지를 할 수 있음

n개의 표본자료 x_n개 존재 , 표본평균은 $\bar{x}$

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

표본표준편차(sample standard deviation)

표본분산의 단위는 관측값 측정 단위의 제곱이 되므로, 계산된 수치로 흩어짐 정도에 대한 크기 가늠이 어려움. 양의 제곱근을 통해 관측값의 단위와 일치 시킴

n개의 표본자료 x_n개 존재 , 분산은 S^2

$$S = +\sqrt{s^2}$$

예시

관측 값	5	8	13	7	10	15
표본평균	5 + 8 + 13 + 7 + 10 + 15 / 6 = 10					
편차	-5	-2	3	-3	0	5
편차제곱	25	4	9	9	0	25
표본분산	25 + 4 + 9 + 9 + 0 + 25 / 5 = 14.4					
표본표준편차	$+\sqrt{14.4}$ = 3.79					

* 통계에서 표본분산을 구할 때 n-1이 n보다 값의 정확도가 더 높아서 n-1로 나눔

제 100 x p 백분위수(percentile)

중위수는 전체 관측값의 반으로, 50%로 나눌 수 있는 경계 값임. 제 100 x p 백분위수는 전체 관측값을 (100xp)%로 나눌 수 있는 값을 뜻함. p는 위치 비율(1사분위 =1/4 , 2사분위 2/4 등)로, 0 <= p <=1을 만족함. n개의 값이 주어지면 제 100 x p 백분위수보다 작거나 같은 관측 값의 개수는 np개 이상이 됨

n은 관측값의 개수, p는 위치 비율
① np가 정수, 제 100 x p 백분위수는 np번째 관측값과 np+1 관측값의 평균
② np가 정수 아님, 제 100 x p 백분위수는 (np의 정수부분에 1을 더한 값)번째 관측값

예시

관측값(=n)	6					
관측 값 정렬	5	7	8	10	13	15

제 1사분위수 = 1/4 ▶ 6 * 1/4 = 1.5 ▶ 실수 ▶ 1+1 = 2 ▶ 2번째 관측 값 ▶ **7**
제 3사분위수 = 3/4 ▶ 6 * 3/4 = 4.5 ▶ 실수 ▶ 4+1 = 5 ▶ 5번째 관측 값 ▶ **13**
제 50백분위수 = 50/100 ▶ 6*0.5 = 3 ▶ 정수 ▶ 8과 10의 평균 값 ▶ **9**

사분위범위(inter-quartile range)

중심으로 각각의 관측 값들이 얼마나 흩어져 있는지 파악

사분위범위(IQR) = 제 3사분위수 – 제 1사분위수 = Q3 – Q1

5. 상자그림(box plot)

히스토그램은 자료가 모여 있는 위치나 자료의 분포의 대략적인 정보를 한 눈에 파악할 수 있지만, 수치정보를 쉽게 알아 볼 수 없음. 이 경우 상자그림(box plot)이 더 좋음

상자그림(inter-quartile range) 또는 상자-수염그림(box-whisker plot)
최소값과 제 1사분위수(Q1), 중위수, 제 3사분위수(Q3), 최대값의 다섯 가지 요약 수치를 이용한 그림

① 제 1사분위수(Q1)와 제 3사분위수(Q3)의 위치에 하나의 네모난 상자를 표현
② 상자 안 중위수의 위치에 수직선을 그음
③ 사분위 범위(IQR)의 1.5배를 계산
④ 양끝으로부터 1.5*IQR 크기의 범위를 펼쳐 울타리의 경계 값을 계산하고 그음
⑤ 울타리 범위 내에 포함되는 관측값 중 최대, 최소값에 수직선을 그어 상자와 연결
⑥ 울타리를 벗어나는 관측값은 이상값임