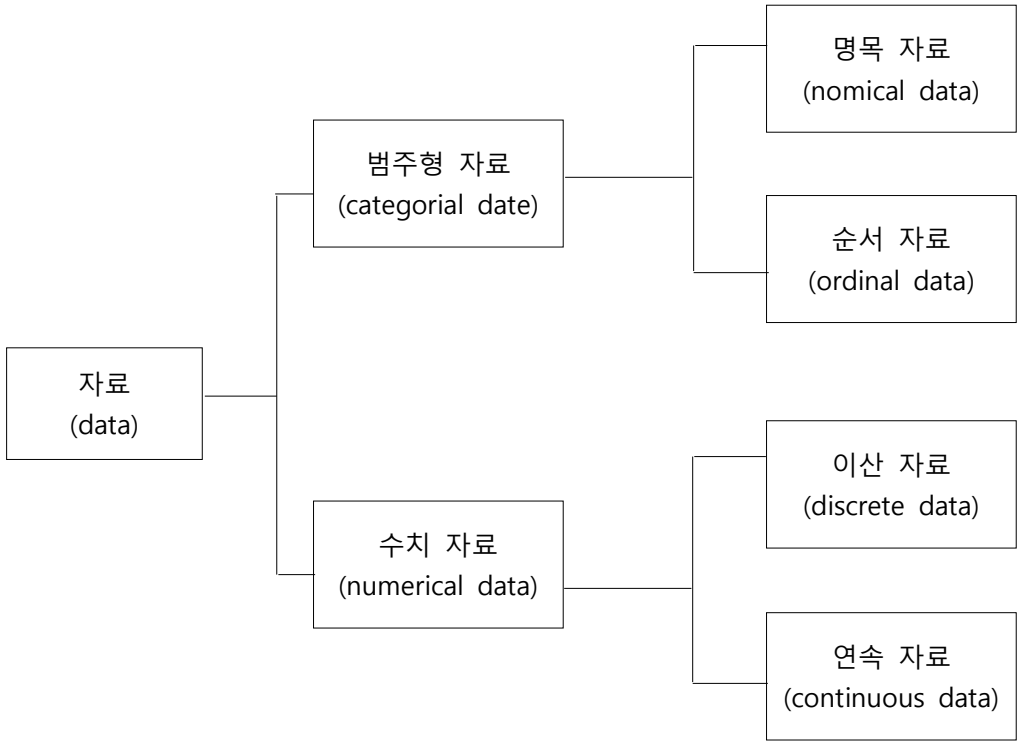
SOOKMYUNG WOMEN'S UNIVERSITY K-MOOC

■ 통계학이란?

- 관심 또는 연구의 대상인 모집단의 특성을 파악하기 위해
- 모집단부터 일부의 자료(표본)를 수집하고
- 수집된 표본을 정리, 요약, 분석하여 표본의 특성을 파악한 후
- 표본의 특성을 이용하여 모집단의 특성에 대해 추론하는 원리와 방법을 제공하는 학문

■ 자료의 분류


자료
(data)
범주형 자료
(categorial date)
수치 자료
(numerical data)
명목 자료
(nomical data)
순서 자료
(ordinal data)
이산 자료
(discrete data)
연속 자료
(continuous data)

SOOKMYUNG WOMEN'S UNIVERSITY K-MOOC

■ 도수분포표(Frequency table)

- 각 범주에 몇 개의 관측개체가 있는지를 정리한 표
  - 도수(frequency): 범주에 속한 관측개체의 수(=빈도)
  - 상대도수(relative frequency): 전체 자료 중 해당 범주에 속한 자료의 비율
  - 누적상대도수(cumulative relative frequency): 순서자료의 경우 전체 자료 중 해당 범주 이하(이상)에 속한 자료의 비율
- 모집단에서 각 범주의 구성비율을 알아보기 위해 사용

SOOKMYUNG WOMEN'S UNIVERSITY K-MOOC

■ 분할표(cotingency table)

- 두 개 이상의 변수를 동시에 고려하여 각각의 범주에 관측개체의 빈도를 정리한 교차표
- 비율(상대도수) 표시
  - 비율은 분석 목적 또는 자료가 어떻게 수집 되었는지에 따라 다르게 표시
  - 비교 vs 관계

SOOKMYUNG WOMEN'S UNIVERSITY K-MOOC

■ 그래프 이용

- 원도표(Pie chart): 원에 각 범주에 해당되는 비율만큼 각도로 표시
  - 해당 범주의 각도 = 비율 × 360°
- 막대그래프(Bar plot): 각 범주의 도수나 상대도수를 막대의 길이로 표시한 그림
- 히스토그램(Histogram): 계급의 상대도수를 면적으로 표시한 그림
  - 높이 = 상대도수/계급구간길이 = 밀도(density)
- 상자그림(Box plot): 사분위수 등 자료의 주요 위치 파악과 이상점 검출 등에 사용되는 그림
  - 그룹별 수치자료의 분포 비교

- 산점도(scatter plot): 쌍을 이룬 자료를 2차원 평면상에 점으로 표시한 그림
  - 목적: 수치 변수들 간의 관계를 고찰
  - 시계열그림(time series plot), 산점도 행렬(scatter matrix)

SOOKMYUNG WOMEN'S UNIVERSITY K-MOOC

■ 중심위치

- n 개의 수치자료: $x_1, x_2, \ldots, x_n$
- 표본평균(sample mean)

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - 표본비율 = 표본평균
  - 이상점(outlier)에 robust하지 않음
  - 가중평균(weighted mean)
- 기하평균(geometric mean)
- 조화평균(harmonic mean)

- 표본중앙값(sample median)

순서통계량(order statistics): 표본을 오름차순으로 정렬한 것

  - $x_{(i)}$: i 번째로 작은 값 $\Rightarrow x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$

- α% 표본절사평균(sample trimmed mean): 순서통계량에서 하위 α%부터 상위 α%까지의 자료를 이용한 표본평균
  - α 백분위수(percentile): 하위 α%에 해당하는 값
  - $p = \alpha/100$ 이면 p 분위수(quantile, 위수)
- 표본최빈값(sample mode): 자료 중 빈도가 가장 많은 값
- 관심 모수가 무엇인가에 따라 선택

SOOKMYUNG WOMEN'S UNIVERSITY K-MOOC

■ 산포(dispersion, 퍼짐)

- 자료들이 얼마나 퍼져 있는지를 나타내는 측도
- 중심위치가 얼마나 안정적인지에 대한 중요한 정보를 제공
- 범위(Range): 최댓값 - 최솟값
- 사분위(간) 범위(Inter-quartile range, IQR)

$$IQR = Q_3 - Q_1$$

  - 사분위수(quartiles): 25%, 50%, 75% 지점(백분위수)
  - 상자그림(Box plot)

- 표본분산(sample variance)

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

  - $x_i - \bar{x}$: i 번째 표본의 편차(deviation)
  - $n-1$: 자유도(degree of freedom)

- 표본표준편차(sample standard deviation)

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

- 변동계수(coefficient of variation)

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

- 표본분산(sample variance)

  - $\overline{x}$: 평균, s_x: 표준편차

$$z_i = \frac{x_i - \overline{x}}{s_x} \Rightarrow x_i = s_x z_i + \overline{x}$$

  - 평균 0, 표준편차 1 ⇒ 측정 단위에 영향을 받지 않게 중심위치와 척도 조정을 통해 절대비교 가능

  - 위치나 척도에 영향을 받지 않는 왜도, 첨도, 상관계수 등의 계산에 사용

SOOKMYUNG WOMEN'S UNIVERSITY K-MOOC

■ 분포의 형태

- 왜도(Skewness): 자료가 대칭적으로 분포되어 있는지는 한쪽으로 기울어져 있는지에 대한 측도

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

- 첨도(Kurtosis): 양쪽꼬리가 얼마나 두터운지를 나타내는 값

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

- 정규성 검정에도 사용

SOOKMYUNG WOMEN'S UNIVERSITY K-MOOC

■ 수치변수 간의 직선관계

- 표본공분산(sample covariance)

$$c = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

  - 측정 단위(척도)에 영향을 받음

- 표본상관계수(coefficient of correlation)

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

  - $-1 \le r \le 1$
  - 두 변수의 인과관계를 나타내는 것은 아님

- 표본의 어떤 특성을 알아볼 것인지?
- 모집단의 어떤 특성에 관심을 가지는가?
  - 모집단에 대한 가정에 따라 다름