Seoul National University

제 2장. 모집단과 표본

모집단의 분포 – 범주형 자료

- 모집단의 분포
  : 모집단의 특성값이 흩어져있는 상태를 합이 1인 양수로써 나타낸 것

- 범주형 자료

(1) 유한 모집단 : 각 특성값의 상대도수를 이용

- 2017년 19대 대선 결과

후보명	M	H	A	기타	합계
득표율	0.411	0.240	0.214	0.135	1.00

(2) 무한 모집단 : 각 특성값의 상대도수의 극한을 이용

<예 2.1> 멘델의 이론

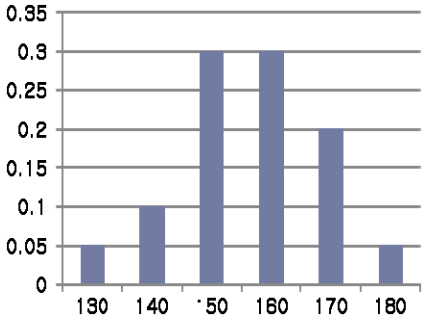
둥글고 노란 완두	둥글고 녹색인 완두	모나고 노란 완두	모나고 녹색인 완두
9/16	3/16	3/16	1/16

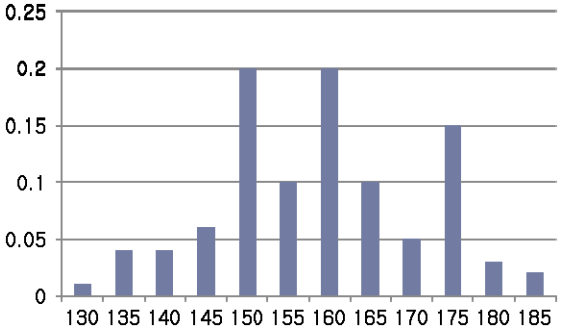
- 이산형 자료의 경우에도 모집단의 분포는 상대도수 또는 그 극한을 이용하여 나타낼 수 있음

모집단의 분포 – 연속형 자료

- 연속형 자료

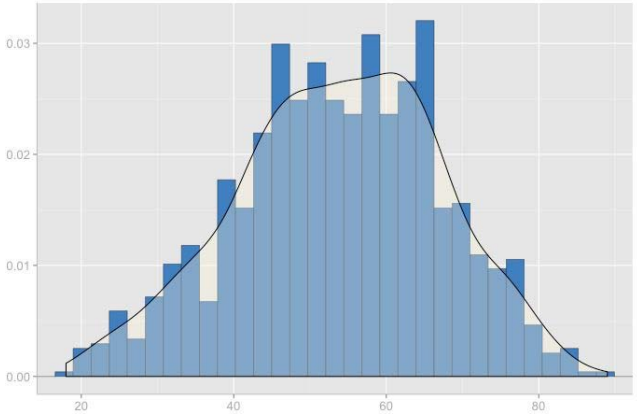
  : 연속형 자료를 계급으로 나누어 생각하는 경우, 계급을 나누는 방법에 따라 모집단의 분포가 여러 모양을 가질 수 있기 때문에 적절하지 않음





- 밀도곡선 (density curve)

  : 연속형 자료의 경우 무한 모집단의 분포를 나타내는 곡선



모집단의 대표값

- 표본 추출의 목적은 모집단의 분포를 추측하기 위한 것
- 하지만 모집단의 분포를 추측하는 것은 매우 어렵기 때문에 모집단 분포의 특징을 나타내는 대표값을 추측하는 방법을 사용
- 이러한 대표값은 특성값이 숫자로 표현되고 크기의 개념을 가지는 경우에만 정의됨

- **모수 (parameter)** : 모집단의 특징을 나타내는 대표값

- 위치의 측도 : 자료의 중심위치를 나타내는 측도
  - 평균 (mean)
  - 백분위수 (percentile) : 사분위수(quartile), 중앙값(median)

- 산포의 측도 : 자료의 흩어짐을 나타내는 측도
  - 분산 (variance) , 표준편차 (standard deviation)
  - 사분위수범위 (interquartile range : IQR)

- 유한모집단 : $\{c_1, c_2, \ldots, c_N\}$, $c_i = i$ 번째 추출단위의 특성값

모집단 위치의 측도

- 평균 : 위치를 나타내는 대표값으로서 분포의 균형점으로서의 중심위치

  - 유한 모집단의 모평균

$$\mu = \frac{1}{N}\sum_{i=1}^{N} c_i = \sum_{i=1}^{k} c_i^* \frac{f_i}{N}$$

  - 무한 모집단의 모평균

$$\mu = \begin{cases} \sum_{all\ x} xp(x) & \text{(이산형)} \\ \int_{-\infty}^{\infty} xp(x)dx & \text{(연속형)} \end{cases}$$

where $p(x)$: 상대도수의 극한 또는 밀도함수

- 제 $(100\times p)$ 백분위수 (제 p 분위수)
  : 특성값을 작은것부터 순서대로 나열하였을 때 $(100\times p)\,\%$ 이상의 특성값이 그 값보다 같거나 작고, $100\times(1-p)\%$ 이상의 특성값이 그 값보다 같거나 크게 되는 값

  - 제 25 백분위수 = 제 1 사분위수 (first quartile), Q_1
  - 제 50 백분위수 = 중앙값 (median), Q_2
  - 제 75 백분위수 = 제 3 사분위수 (third quartile), Q_3

모집단 산포의 측도

- 사분위수범위 (IQR)
  : 모집단 가운데 50% 특성값의 범위 ($IQR = Q_3 - Q_1$)
  : 양쪽 극단 값에서 자료의 25%씩 안쪽으로 들어와 있는 값의 거리
  ⇒ 특이값의 영향을 거의 받지 않음

- 분산 : 자료 각각이 그 평균으로부터 떨어져있는 거리(편차)를 제곱한 것의 평균값
  : 자료 하나하나의 값이 전부 고려되어 구해진 변동성 척도

  - 유한 모집단의 모분산

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(c_i - \mu)^2 = \sum_{i=1}^{k}(c_i^* - \mu)^2 \frac{f_i}{N}$$

  - 무한 모집단의 모분산

$$\sigma^2 = \begin{cases} \sum_{all\ x}(x-\mu)^2 p(x) & \text{(이산형)} \\ \int_{-\infty}^{\infty}(x-\mu)^2 p(x)dx & \text{(연속형)} \end{cases}$$

이차원 모집단의 분포

- <표 2.4> 두 이차원 모집단의 결합 분포

모집단 A

V1 / V2	0	1	2	합
0	.1	.2	.1	.4
1	.0	.2	.1	.3
2	.0	.0	.3	.3
합	.1	.4	.5	1.0

모집단 B

V1 / V2	0	1	2	합
0	.0	.0	.4	.4
1	.0	.2	.1	.3
2	.1	.2	.0	.3
합	.1	.4	.5	1.0

- 모집단 A와 모집단 B의 각 특성별 모평균과 모표준편차는 동일함
- 주변확률분포는 동일하지만 결합분포는 동일하지 않음
- 이차원 모집단의 경우, 두 특성의 변화관계를 나타내는 또 다른 대표값이 필요

이차원 모집단의 대표값 - 상관계수

- **상관계수 (correlation coefficient, ρ)**
  : 두 특성의 변화관계(선형관계)를 나타내는 대표값

  - 유한모집단의 모상관계수

$$\rho = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{c_{1i}-\mu_1}{\sigma_1}\right)\left(\frac{c_{2i}-\mu_2}{\sigma_2}\right) = \sum_{i=1}^{k}\sum_{j=1}^{l}\left(\frac{c_{1i}^{*}-\mu_1}{\sigma_1}\right)\left(\frac{c_{2j}^{*}-\mu_2}{\sigma_2}\right)\frac{f_{ij}}{N}$$

  - 무한모집단의 모상관계수

$$\rho = \begin{cases} \dfrac{1}{\sigma_1\sigma_2}\sum_{all\ x}\sum_{all\ y}(x-\mu_1)(y-\mu_2)\,p(x,y) & \text{(이산형)} \\ \dfrac{1}{\sigma_1\sigma_2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_1)(y-\mu_2)\,p(x,y)\,dxdy & \text{(연속형)} \end{cases}$$

  where $p(x,y)$: 결합밀도함수 (joint density function)

표본의 도표화

- 표본 (sample) : 실제 관측한 것들의 모임. 모집단의 부분집합

- 자료의 요약
  : 표본(자료)의 요약을 통해 모집단의 개형을 파악
  : 자료의 대략적인 모습을 보여주기 위한 단계
  : 섬세한 분석을 위한 기초단계

- **기술통계학(descriptive statistics)**
  : 자료의 특성을 쉽게 파악할 수 있도록 정리, 요약하는 방법을 다루는 분야
  - 표나 그래프를 이용한 요약
  - 통계량을 이용한 수치적 요약

도수분포표

- 자료의 전체적인 구성 형태를 도수 (Frequency) 로 표현함

- 범주형 자료 : 범주(category)에 따른 빈도(또는 도수, frequency)가 주어짐
  : 빈도표(frequency table, 또는 도수분포표)를 이용하여 자료 정리
- 연속형 자료 : 전체 범위를 몇 개의 계급(class)으로 나눈 뒤 각 계급에 속하는 자료의 수를 도수로 표현
  : 계급의 개수와 계급구간은 자료의 크기와 성질에 따라 달라짐

- <예> 완두콩의 분류 결과

완두콩의 형태	둥글고 노란 완두	둥글고 녹색인 완두	모나고 노란 완두	모나고 녹색인 완두
관측도수	315	108	101	32

- <예> 나이자료의 도수분포표

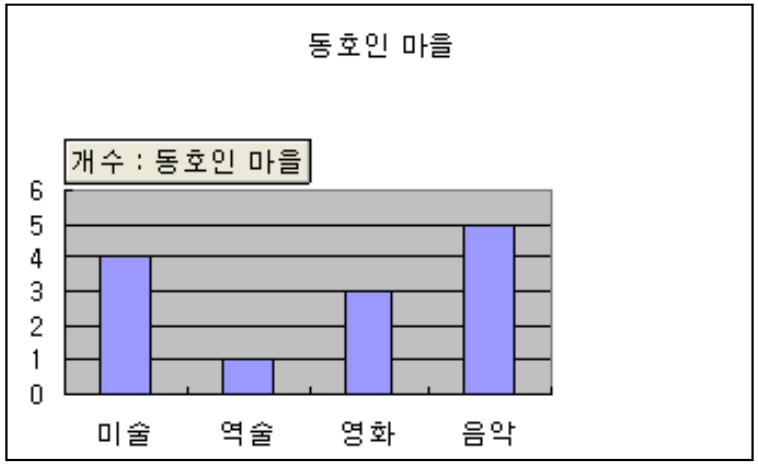
나이	~29	30~39	40~49	50~59	60~
도수	4	9	18	16	3

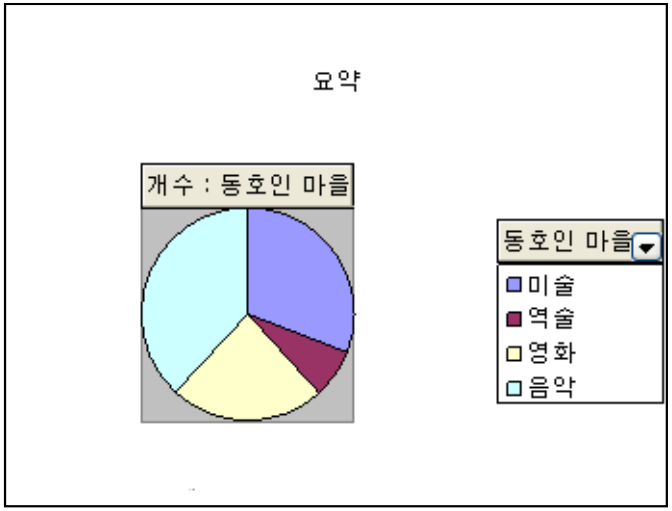
막대그래프와 원형그래프

- 범주형 자료의 표현

동호인 마을의 도수분포표

동호인 마을	도수	상대도수
음악	5	0.38
미술	4	0.31
영화	3	0.23
역술	1	0.08
합계	13	1.00





그래프를 이용한 자료의 정리 - 히스토그램

- **히스토그램 (histogram)**

  : 데이터의 분포상태를 일목요연하게 그림으로 나타내는 방법

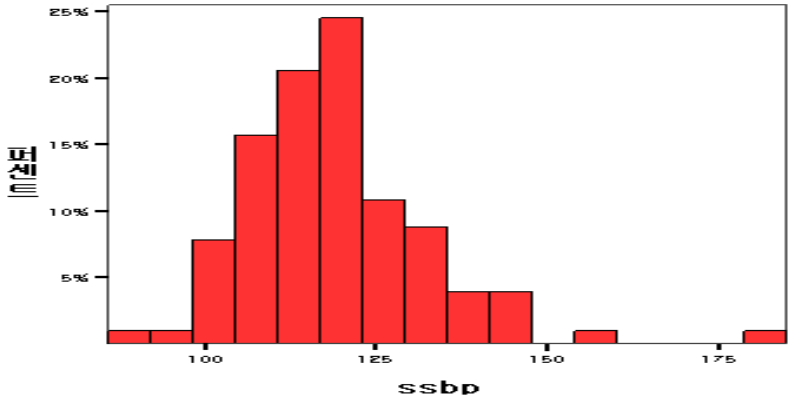
  : 데이터가 어떤 모양으로 분포하고 있는지를 파악하기 위한 가장 기초적인 단계

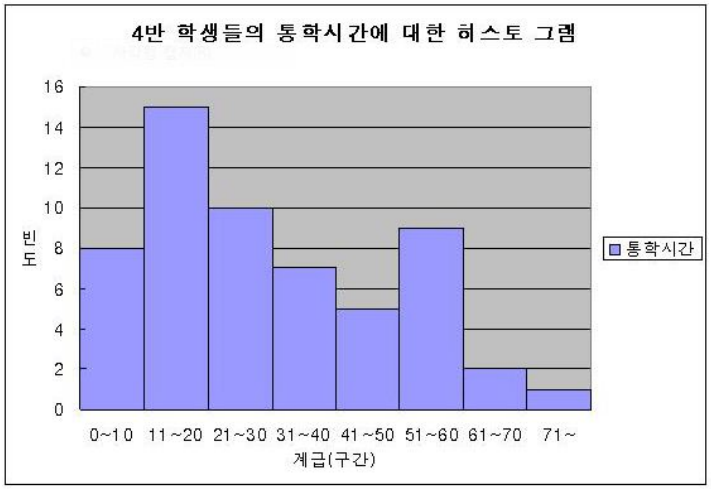
  : 히스토그램의 모양을 통해 데이터의 기본적인 특성 파악 가능

- 수평축 위에 계급구간을 표시하고 그 위로 각 계급의 상대도수에 비례하는 넓이의 직사각형을 그린 것

  : 전체 직사각형의 넓이의 합은 1

  : 도수분포표와 히스토그램은 관측값을 계급구간으로 변환시키는 방법이므로 일정부분 정보의 손실(loss of information)을 초래하게 됨

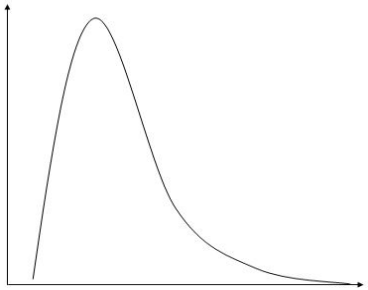




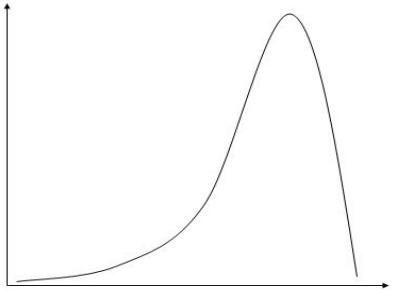
히스토그램

- 히스토그램의 모양을 통해 데이터의 기본적인 특성 파악 가능
  : 분포의 좌우 대칭(symmetry) 여부, 이상점(outlier)의 존재 유무

- 왜도 (Skewness) : 좌우로 쏠려있는 정도
  - 양의 왜도 (Right skewed) : 오른쪽으로 길게 늘어짐
  - 음의 왜도 (Left skewed) : 왼쪽으로 길게 늘어짐



Right Skewed



Left Skewed

- (참고) 분포에 따른 평균과 중앙값의 위치

그래프를 이용한 자료의 정리 – 줄기-잎 그림

- **줄기-잎 그림 (stem-and-leaf display)**

  : 데이터를 줄기(구간)와 잎(관찰값)으로 분리하여 나열한 그림

  : 원래의 데이터가 그대로 보존되어 있으므로 데이터의 분포 뿐만 아니라 각 관찰값들의 순위까지 알 수 있음

  : 히스토그램을 90도 회전시킨 모양

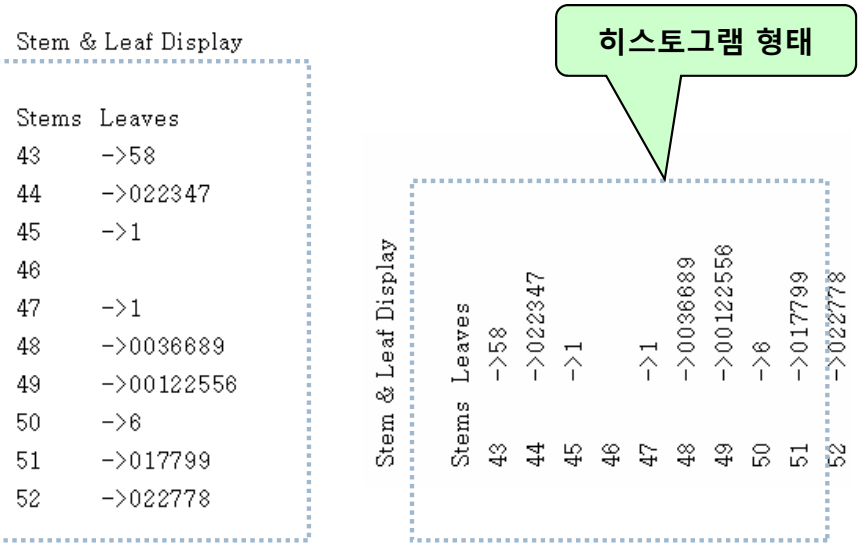
  : 히스토그램이 생략한 정보까지 추가로 나타냄

  : 데이터가 소규모일때 매우 유용함

  : 자료의 특성을 잘 나타낼 수 있도록 줄기의 수를 적당히 정해야 함

Stem & Leaf Display

Stems	Leaves
43	->58
44	->022347
45	->1
46	
47	->1
48	->0036689
49	->00122556
50	->6
51	->017799
52	->022778



줄기-잎 그림

- (예) 혈압 자료에서 '나이' 변수에 대한 줄기-잎 그림

46	53	53	48	53	58	48	66	67	40	…	42	46	73	51	36	43	47	67	40

줄기-잎 그림(Stem-and-Leaf Plot)
변수명: 나이

Stem Unit: 10 Leaf Unit: 1

6	3	455668
22	4	0022345566677888
(16)	5	1122333456677788
12	6	4456677
5	7	13567

이차원 표본자료의 도표화

- **분할표 (contingency table)**

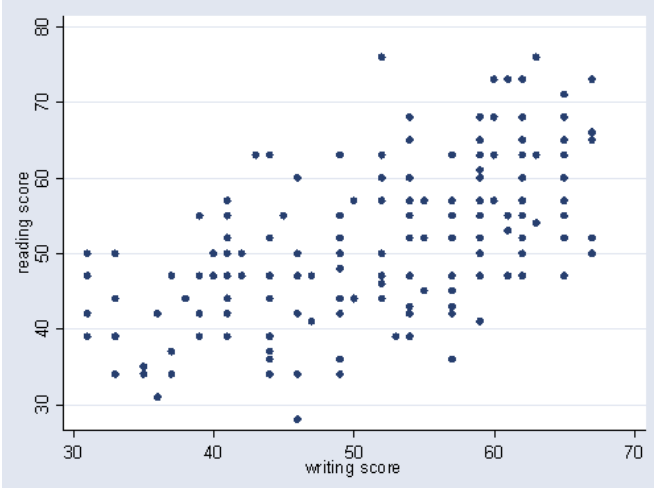
  : 특성값이 모두 범주형인 이차원 자료의 도수분포표

  <표 2.6> 400명의 학생들의 과목별 선호도 조사 결과

	국어	영어	수학	합계
남자	.18	.24	.18	.60
여자	.14	.16	.10	.40
합계	.32	.40	.28	1.00

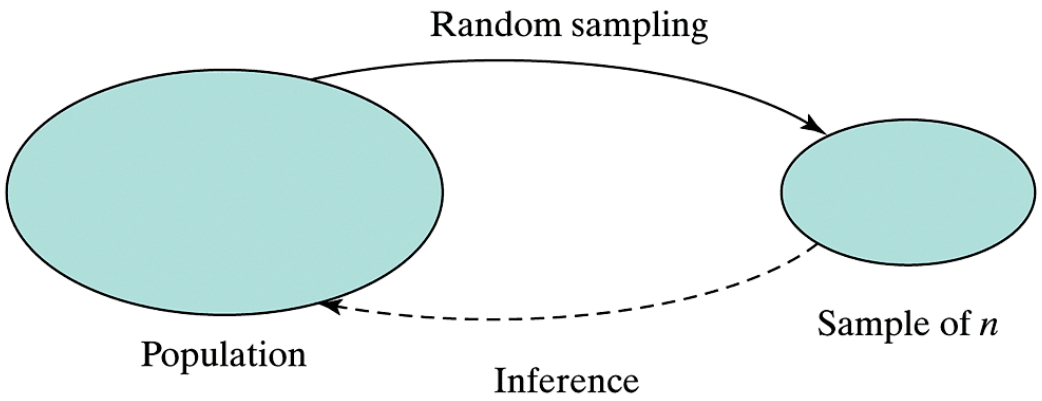
- **산점도 (scatter plot)**

  : 특성값이 모두 연속형인 이차원 자료를 좌표평면 위에 나타낸 그림



모집단과 표본

- **통계량 (statistic)** : 표본에서 얻은 표본의 대표값
- **추정량 (estimator)** : 모수의 추측에 사용되는 통계량



- 모수와 추정량

표본의 대표값

- $\{x_1, x_2, ..., x_n\}$: 크기가 n 인 표본의 특성값
- 표본평균 (sample mean) : $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- 표본의 제 $(100 \times p)$ 백분위수 $\hat{\eta}_p$: $x_{(1)} \le x_{(2)} \le \cdots \le x_{(m)} \le \cdots \le x_{(n)}$ 일 때, $x_{(np)} \le \hat{\eta}_p \le x_{(np+1)}$
  - np 가 자연수 일 때 : $\hat{\eta}_p = \frac{x_{(np)} + x_{(np+1)}}{2}$
  - np 가 자연수가 아닐 때 : $\hat{\eta}_p = x_{([np]+1)}$
- 표본의 제 1사분위수, 중앙값, 제 3사분위수 : $(\hat{Q}_1, \hat{Q}_2, \hat{Q}_3)$
- 예제 2.8 : 사분위수의 계산

표본의 산포

- 표본의 사분위수 범위 : $\widehat{IQR} = \hat{Q}_3 - \hat{Q}_1$

- 표본분산 (sample variance)

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

- 표본 표준편차 (sample standard deviation)

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

- 표본 상관계수 (sample correlation coefficient)

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_1}\right)\left(\frac{y_i - \overline{y}}{s_2}\right) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$