Seoul National University

- 일반통계학 -
제 2장
모집단과 표본

2-1 모집단의 분포

모집단의 분포(distribution)
-모집단의 특성값이 흩어져 있는 상태를 합이 1인 양수로써 나타낸 것
- 표본조사의 목적 중 하나는 모집단의 분포를 추측하는 것

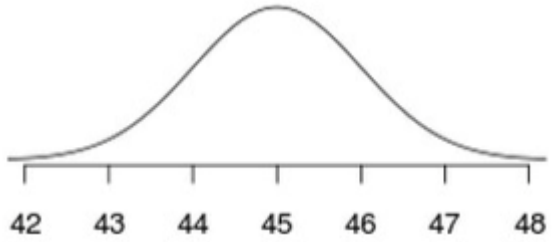
예) 1987년 대통령 선거 결과 (※ 유권자 전원이 투표함을 가정)

후보명	노태우	김영삼	김대중	김종필	합계
득표율	0.367	0.281	0.271	0.081	1.000

예) 특성값이 키인 모집단의 분포 – 상대도수 이용

계급	130~140	140~150	150~160	160~170	170~180	180~190	합계
상대도수	0.05	0.1	0.3	0.3	0.2	0.05	1

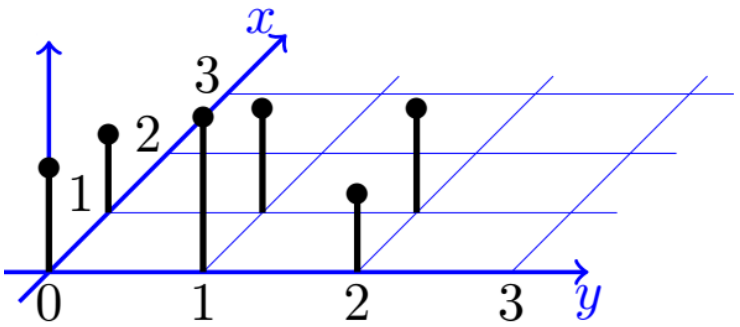
특성값이 연속인 경우 계급을 나누는 방법에 따라 모집단의 분포가 변할 수 있다. 따라서 연속적인 곡선으로 모집단의 분포를 나타낸다. 이를 밀도곡선이라고 하고 이 경우 곡선 아래의 면적의 합이 1이 된다.



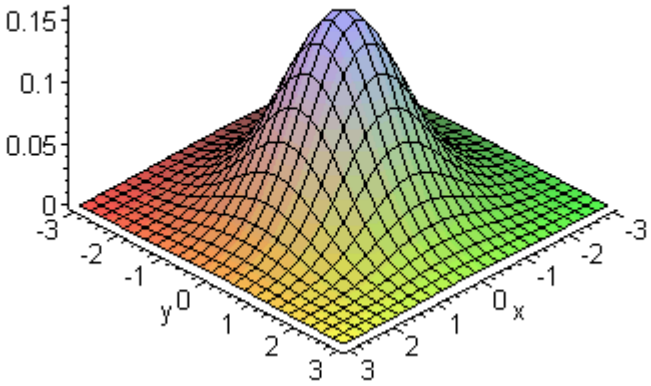
2-1 모집단의 분포

다차원 특성값의 모집단 분포를 결합분포라고 부른다. (Joint distribution)

•이산형 특성값



•연속형 특성값



2-2 모집단의 대표값

•대표값

-모집단의 특성값이 숫자로 표현되고 크기의 개념을 가지는 경우에만 정의
-모집단의 분포를 추측하는 것이 복잡하고 어렵기 때문에 모집단 분포의 특징인 대표값을 주로 추측

(1)모집단 분포의 위치를 나타내는 대표값
: 모평균, 모중앙값, 모사분위수, 모백분위수 등

a. 모평균(μ) : 모집단 분포의 무게중심

- 유한 모집단

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

- 무한 모집단

$$\mu = \begin{cases} \sum_{\text{모든 } x} xp(x) & \text{(이산적인 경우)} \\ \int_{-\infty}^{\infty} xp(x)\,dx & \text{(연속적인 경우)} \end{cases}$$

2-2 모집단의 대표값

b. 모사분위수, 모백분위수

- 모중위수 또는 모중앙값은 모평균과 마찬가지로 모집단 분포의 중심위치를 나타낸다. 이 때는 무게중심이 아니라 전체 특성값의 50%에 해당하는 위치이므로 분포의 모양에 따라 모평균과 일치할 수도 있고 다를 수도 있다.

2-2 모집단의 대표값

(2)모집단 분포의 산포를 나타내는 대표값

: 모분산 (σ^2), 모표준편차($\sigma = \sqrt{\sigma^2}$), 평균절대편차, 사분위수범위(교재 참고)

모분산 (모평균으로부터 특성값들이 흩어진 정도)

- 유한 모집단

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

- 무한 모집단

$$\sigma^2 = \begin{cases} \sum_{\text{모든 } x} (x-\mu)^2 p(x) & \text{(이산적인 경우)} \\ \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx & \text{(연속적인 경우)} \end{cases}$$

2-2 모집단의 대표값

(3) 연관성의 대표값 : 두 특성값의 선형관계 유무를 알려주는 대표값
: 모공분산, 모상관계수

a. 모상관계수(population correlation coefficient)

➢ 유한 모집단

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

➢ 무한 모집단

$$\rho = \begin{cases} \dfrac{\sum_{\text{모든 } x}\sum_{\text{모든 } y}(x-\mu_1)(y-\mu_2)p(x,y)}{\sigma_1\sigma_2} & \text{(이산적인 경우)} \\ \dfrac{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_1)(y-\mu_2)\,p(x,y)dxdy}{\sigma_1\sigma_2} & \text{(연속적인 경우)} \end{cases}$$

b. 모공분산 (population covariance): $\rho\sigma_1\sigma_2$

*모집단의 특성을 나타내는 대표값 : 모수 (parameter)

2-3 표본의 대표값

1. 위치

-자료들이 대략 어떠한 값을 갖는 지를 알아보기 위하여, 어느 위치를 중심으로 자료들이 모여 있는 지를 나타내는 척도

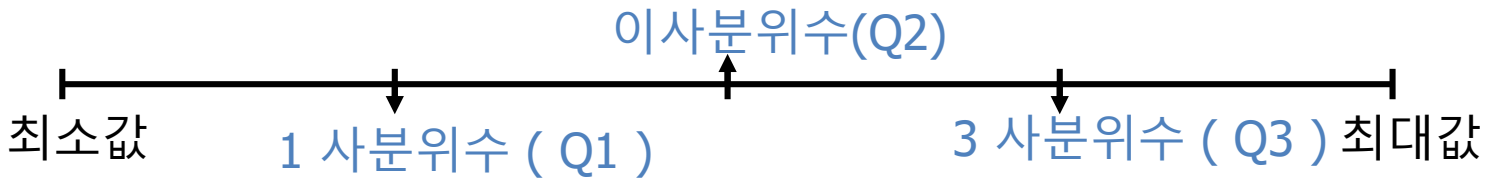
(1) 평균 - 1, 3, 5, 7, 9 → 평균 = 25 / 5 = 5
1, 3, 5, 7, 14 → 평균 = 30 / 5 = 6

(2) 중앙값 – 자료가 홀수 개 : 가운데 값 / 짝수 개: 가운데 두 값의 평균

(3) 최빈값 - 자료 중 그 빈도 수가 최대인 값 , 여러 개 나올 수 있음.
31, 34, 36, 33, 28, 34, 30, 34, 32, 40

(4) 사분위수 – 자료를 크기 순서로 나열 했을 때, 25%, 50%, 75%에 해당하는 값 (4분위수는 중앙값과 동일)
여러가지 공식이 존재 함.

2-3 표본의 대표값



$$Q_1 = (n+1)\frac{1}{4} = (n+1)\frac{25}{100}$$ 번째 순위 값

$$Q_3 = (n+1)\frac{3}{4} = (n+1)\frac{75}{100}$$ 번째 순위 값

예) 다음과 같은 자료를 얻었다고 하자.

1, 0, 7, 5, 3, 2, 0, 1, 8, 4

자료를 크기 순으로 나열 -> 0,0,1,1,2,3,4,5,7,8

Q1 : (10 + 1) * 25 / 100 = 2.75 번째에 해당

따라서 0 + (1 – 0) * 0.75 = 0.75.

Q2 : (10 + 1) * 50 / 100 = 5.5 번째에 해당

따라서 2 + (3 – 2) * 0.5 = 2.5.

Q3 : (10 + 1) * 75 / 100 = 8.25 번째에 해당

따라서 5 + (7 – 5) * 0.25 = 5.5.

2-3 표본의 대표값

2. 산포의 측도

- 자료가 변동하거나 퍼져 있는 정도를 나타냄

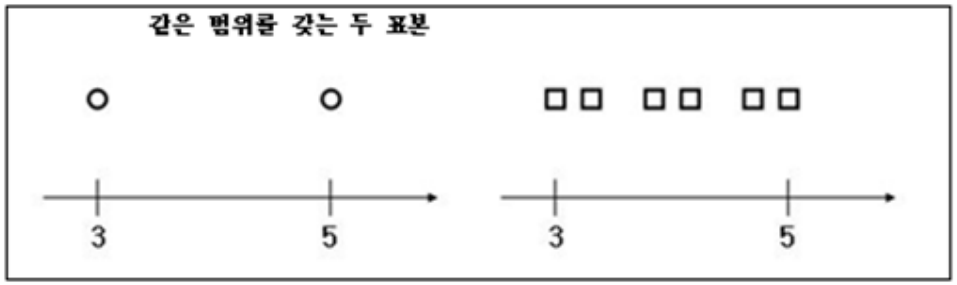
(1) 범위 : 최대값- 최소값

쉽고 빠르게 구할 수 있음

: 특이하게 크거나 작은 값이 있을 경우 자료의 범위가 왜곡됨

: 자료의 개수와 상관없이 같게 나올 수 있음

⇒ 자료의 변동성을 대표하지 못하는 경우가 많음



(2) 사분위수 범위 : 3사분위수-1사분위수

특이값의 영향을 거의 받지 않음

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

표본표준편차 : $s=\sqrt{s^2}$

(4)변동계수(coefficient of variance)=표준편차/ 평균
- 단위가 다른 두 그룹에 대한 산포 비교

예) 성인과 신생아의 몸무게에 대한 자료에서 어느 집단이 평균으로부터 더 퍼져 있을까?
성인 : 평균 67.1 표준편차 4.19 / 신생아 : 3.22 표준편차 1.08

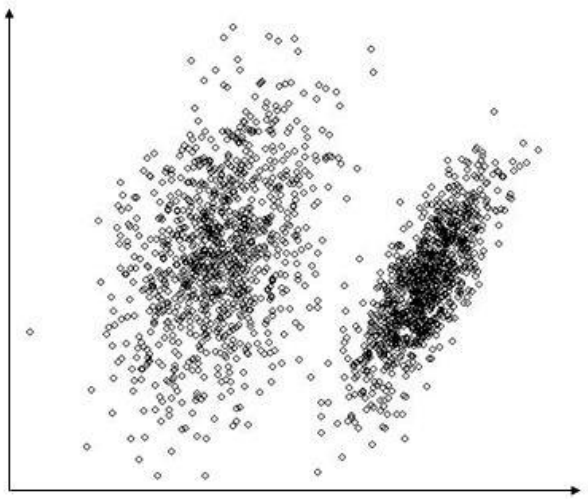
3. 선형 관계의 측도 (Measure of linear association)

표본: $(x_1, y_1), (x_2, y_2) \dots (x_n, y_n)$

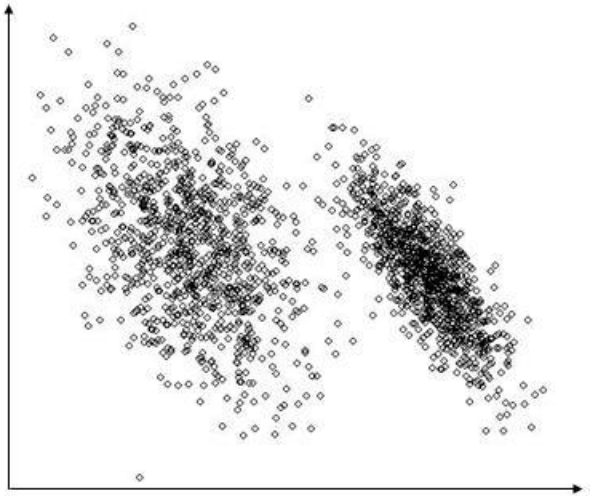
표본상관계수 : $r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$ $(0 \le |r| \le 1)$

표본공분산 : $s_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1} = r s_x s_y$

2-3 표본의 대표값



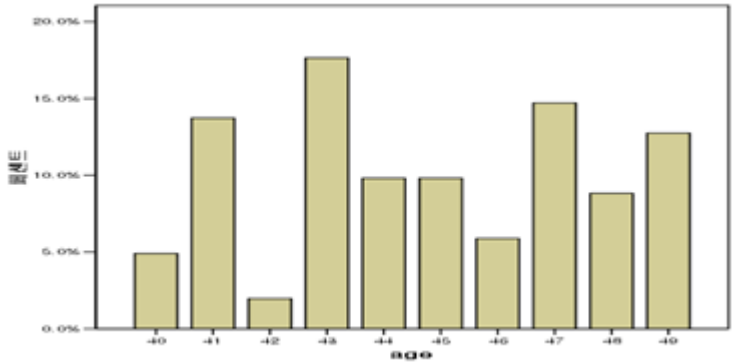
양의 공분산의 크기 비교



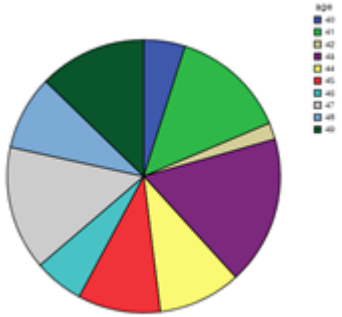
음의 공분산의 크기 비교

2-4 표본의 도표화 (자료의 정리)

1. 막대그래프 : 범주형 자료, 이산형 자료(특성값의 개수가 유한일 때)
: 수평축 위에 각 특성값의 위치를 잡고, 막대의 높이가 상대도수에 비례하게 그린 것.
예) 100명의 40대 남성의 나이



2. 원형그래프 : 범주형 자료, 이산형 자료(특성값의 개수가 유한일 때,)
: 원을 부채꼴 모양으로 나누는데, 각 부채꼴의 넓이가 상대도수에
비례하게 그린 것.
예) 100명의 40대 남성의 나이



2-4 표본의 도표화 (자료의 정리)

3. 히스토그램 : 연속형 자료

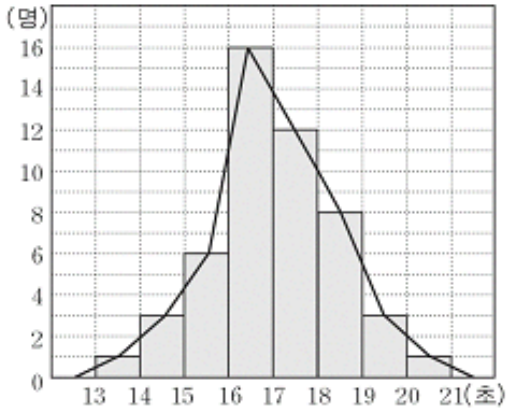
: 수평축위에 계급구간을 표시하고 그 위로 각 계급의 상대도수에 비례하는 넓이의 직사강형을 그린 것.

(1) 전체 직사각형의 넓이의 합은 1이 된다.

(2) 정보의 손실이 많지만 자료의 시각적 이해에 도움이 된다.

4. 도수다각형

: 히스토그램에서 윗변의 중점을 연결한 직선



2-4 표본의 도표화 (자료의 정리)

5. 줄기-잎 그림

(1) 줄기-잎 그림의 작성요령 및 예 (예제 2.6)

① 처음 두 자리의 수를 세로로 순서대로 나열한 후 그 오른쪽에 수직선을 그린다.

18 |
17 |
16 |
15 |
14 |
13 |

② 각 자료값에 대하여 마지막 한 자리수(1의자리수)를 해당되는 처음 두 자리수의 오른쪽으로 크기 순으로 나열한다.

18 | 0, 0, 3, 3, 3, 5
17 | 0, 0, 0, 0, 0, 1, 3, 3, 4, 6, 8, 8, 0
16 | 0, 0, 0, 3, 3, 3, 5, 5, 5, 5, 5, 8, 8, 8, 8, 8
15 | 0, 2, 4, 8
14 | 5
13 | 8

(2) 성질

① 자료에 대한 정보의 손실이 거의 없다
② 자료의 형태 파악이 쉽다.
③ 이상치 자료에 대한 정보를 제공한다.
④ 방대한 표본자료인 경우에는 그리기 어렵다.

2-4 표본의 도표화 (자료의 정리)

6. 분할표

(1) 특성값이 모두 범주형인 이차원 모집단의 도수분포표
(2) 각 이차원 특성값의 상대도수(또는 도수)를 이차원 표에 나열한 것
예) 400명의 학생들의 국어, 영어, 수학 세 과목의 선호도

	국어	영어	수학	합계
남자	0.18	0.24	0.18	0.60
여자	0.14	0.16	0.10	0.40
합계	0.32	0.40	0.28	1.00

7. 산점도, 리그레쏘그램

(1) 산점도: 특성값이 모두 연속형인 이차원 모집단인 경우 각 이차원 자료값에 대하여 좌표가 (특성 1의 값, 특성 2의 값)인 점을 좌표평면 위에 찍은 것.
(2) 수평축 위의 각 계급구간에서 속하는 Y의 평균이 수직축의 좌표가 되도록 그린 그림