# 경영통계

원광대학교 경영학부

담당교수: 정호일

# 제8장 표본분포

- 표본추출방법
- 표본분포
- 평균의 표본분포
- 표본분포의 형태
- 비율의 표본분포

# 본장의 개요

본 장은 기술통계의 개념과 확률분포의 개념을 적절히 연결시키는 것이다.
표본정보를 적절히 종합하여 주요 표본통계량 개념을 설정한 다음 이들 표본통계량들이 따르는 확률분포를 확인하는 과정을 종합 정리하여 추론 통계분석의 기초를 마련하고자 하는 것이다.

# Ⅰ. 표본추출

1. 표본을 추출하는 방법
   1) 확률추출방법(probability sampling)
   2) 비확률추출방법(nonprobability sampling)

2. 확률추출방법
   모집단 내의 각 구성원소가 표본에 포함될 가능성이 동일하며 독립적인 기회를 가진다는 요건하에 표본을 객관적으로 추출하는 방법
   1) 단순무작위 추출
   2) 층별 추출
   3) 체계적 추출
   4) 군집 추출 등이 있다.

   확률추출방법은 객관적이므로 표본오차의 측정이 가능하다.

# Ⅰ. 표본추출

## 3. 비확률추출방법

조사자의 주관에 의하여 표본을 임의로 추출하는 방법으로서 표본오차를 객관적으로 측정할 수 없다는 결점을 갖는다.

1) 할당추출
2) 편의추출

# Ⅰ. 표본추출

## [ 확률추출방법 ]

1. 단순무작위 추출(Simple random sampling)
   N개의 요소로 구성된 모집으로부터 표본크기 n개의 요소를 선정한다고 할 때 단순무작위방법은 n개의 가능한 각 표본이 똑같이 1/N의 확률로 선정될 수 있도록 설계된 방법.
   예) 복권추첨, 아파트추첨
2. 층별추출 (stratified sampling)
   모집단을 부, 지역, 연령, 성별, 교육 같은 일정한 기준에 의하여 동질적인 그룹(층)으로 분류한 다음 각 그룹으로부터 표본을 단순무작위로 추출하는 방법.
   각 그룹에서 추출하는 표본의 수가 모집단의 구성비율을 따를 때 비례적 층별추출이라고 한다.
   층별추출은 모집단의 특성을 더욱 정확하게 반영한다는 장점을 갖는다. 층별추출이 효과적인 때는 특성에 있어 층간에는 차이가 크지만 층 내에서는 차이가 별로 없는 경우이다.

# Ⅰ. 표본추출

## [ 확률추출방법 ]

3. 체계적 추출
   모집단이 큰 경우 단순무작위 추출방식을 사용하면 시간과 비용상 비경제적이므로 체계적 방법(systematic sampling)을 사용할 수 있다. 모집단의 크기가 100이고 표본크기가 5이면 표본간격을 20=100/5으로 정하고 모집단을 순서대로 번호를 부여한 후 첫 20명 중에서 1명을 무작위 추출한 후 20의 간격으로 5개의 표본을 추출하는 방법이다. 예를 들어 9, 29, 49, 69, 89

4. 군집추출(cluster sampling)
   모집단을 군집(그룹)으로 구분하고 이 중에서 단순무작위방식으로 조사대상인 군집을 선정하는 방식이다.선정된 군집에 대해서 전수조사를 하거나 일부의 표본을 추출하게 된다. 군집추출이 효과적인 경우는 층별추출과 반대의 경우로 군간에는 동질적이고 군 내에서는 이질적인 특성을 갖는 경우이다.

# Ⅰ. 표본추출

## [ 비확률 추출 방법 ]

1. 편의추출

비확률 추출기법의 하나인 편의추출방법은 표본이 조사자의 편의에 의해서만 선정되는 방법이다. 비교적 쉽게 표본을 선정하고 자료를 수집하는 장점이 있으나 모집단을 제대로 대표할 수 없다는 단점을 갖는다.

2. 판단추출

판단추출방법은 모집단의 특성을 잘 아는 전문가가 모집단을 가장 잘 대표하리라고 믿는 요소들을 표본으로 추출하는 방법이다

# II. 오차의 종류

오차의 종류

## 1) 표본추출오차

모집단을 대표할 수 있는 전형적인 구성요소를 표본으로 선택하지 못했기 때문에 발생하는 오류이다.

① 표본의 크기 때문에 발생하는 우연한 오류

■ 1학년 신입생IQ를 알아보기 위해 1학년 중 임의로 4명을 선정 조사한 결과 평균이 124로 높은 수준이라고 생각했는데 실제 신입생700명 전체 평균IQ는 98에 지나지 않았다.

② 모집단을 대표할 수 없는 비전형적인 구성요소를 뽑았기 때문에 발생하는 오류

- 조사자: 리터러리 다이제스트 - 기간: 1936년
민주당의 프랭클린 루즈벨트와 공화당의 알프레드 랜든의 선거 결과 예측을 위해 수백만명 샘플링- 랜든의 압도적 승리예상 결과:루즈벨트 당선

◆ 분석:

- ❖ 다이제스트사의 표본프레임-다이제스트사의 독자와 전화번호부
- ❖ 미국시대적 상황-1930년 경제공황의 시기
- ❖ 다이제스트 독자-정기구독자(부유층)
- ❖ 전화를 보유하고 있는 가정-부유층
- ❖ 부유층-공화당(랜든)지지 경향이 있었음

◆ 결과:표본추출계획의 오류(표본프레임오차)

## 2) 비표본추출오차

1) 조사과정과 집계과정에서 발생하는 오차로서 표본선택이 잘못된 것이 아니라 주로 표본의 성격을 측정하는 방법이 부정확해서 발생하는오차(측정오차)

2) 예:한 사람에게 두 사람이 같은 질문을 하는 경우 각각 다른 대답을 얻는 경우(질문방법, 설명부족이 원인)

3) 질문하는 사람과 응답자의 편견, 계층의식 등에 의해 발생하는 오류

- ■ 잘못된 저울로 무게를 측정했을 경우

# III. 표본분포 (Sampling Distribution)

## 표본분포의 개념

주어진 모집단으로부터 크기 n의 가능한 모든 표본집단으로부터 얻은 표본통계량(표본의 평균, 표본의 분산, 표본의 비율)의 확률분포를 말한다.

## 표본분포의 종류

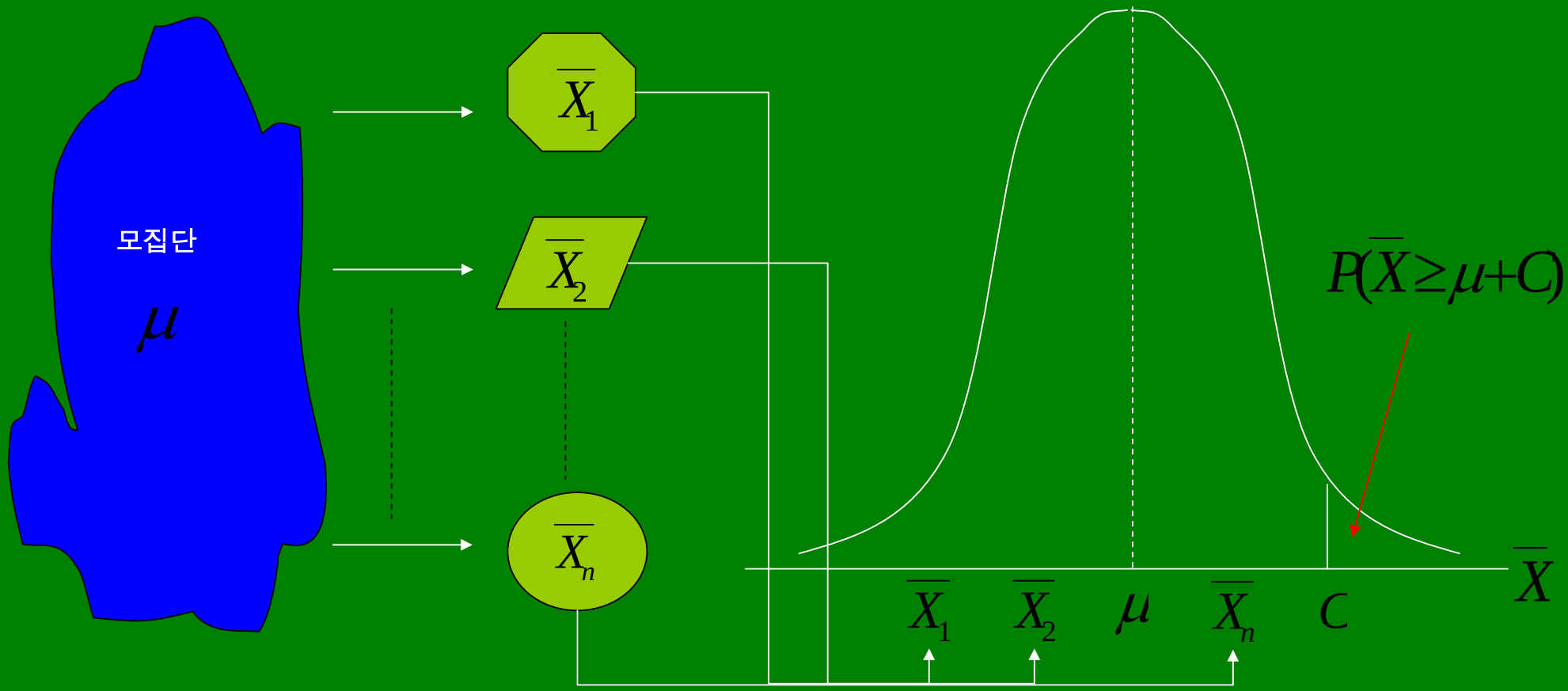
- 표본평균의 확률분포
- 표본분산의 확률분포
- 표본비율의 확률분포

## 표본분포의 유용성

- 표본의 크기 n의 표본을 수없이 추출하지 않아도 가능한 모든 표본통계량들의 분포를 이용하여 통계량이 모수에 어느 정도 근접되어 있는가에 대한 확률을 알 수 있다.

# Ⅳ. 표본평균의 확률분포

## 표본평균들의 확률분포



# Ⅳ. 표본평균의 확률분포

## 표본평균의 확률분포

평균 μ와 표준편차 σ를 갖는 정규모집단으로부터 동일한 크기 n의 표본을 수없이 추출하여 그들의 평균을 계산하였을 때 이 표본 평균들의 확률분포를 말한다.

이는 $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$ 으로 표기한다.

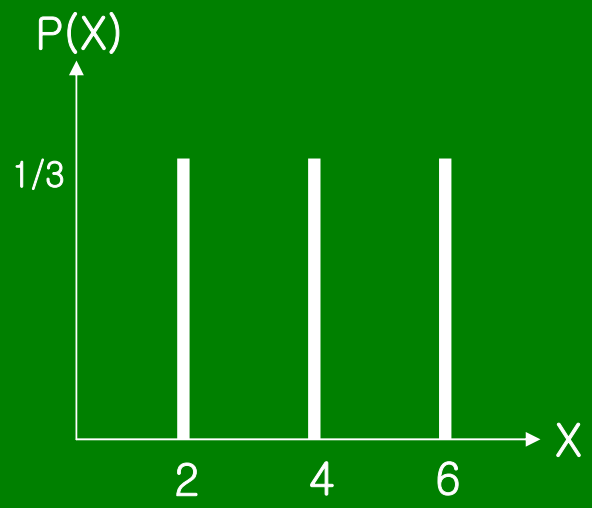
X:근무연수

2,4,6

(모집단)

모집단 확률분포

| X | P(X) |
|---|---|
| 2 | 1/3 |
| 4 | 1/3 |
| 6 | 1/3 |




$$\mu = E(x) = 2\times\frac{1}{3} + 4\times\frac{1}{3} + 6\times\frac{1}{3} = 4$$

$$\sigma^2 = (2-4)^2\times\frac{1}{3} + (4-4)^2\times\frac{1}{3} + (6-4)^2\times\frac{1}{3} = 2\frac{2}{3}$$

$$\sigma = \sqrt{2\frac{2}{3}} = 1.633$$

# Ⅳ. 표본평균의 확률분포

## 표본평균의 확률분포

X:근무연수

2,4,6

(모집단)

복원 추출 (n=2)

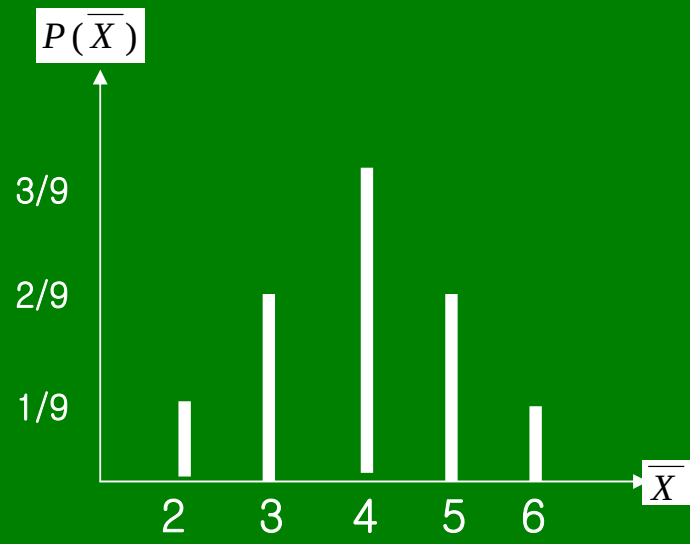
| 표본 집단 | 표본 평균($\overline{X}$) |
|---|---|
| 2,2 | 2 |
| 2,4 | 3 |
| 2,6 | 4 |
| 4,2 | 3 |
| 4,4 | 4 |
| 4,6 | 5 |
| 6,2 | 4 |
| 6,4 | 5 |
| 6,6 | 6 |

모집단 확률분포

| X | P(X) |
|---|---|
| 2 | 1/3 |
| 4 | 1/3 |
| 6 | 1/3 |

표본평균의 확률분포

| $\overline{X}$ | P(X) |
|---|---|
| 2 | 1/9 |
| 3 | 2/9 |
| 4 | 3/9 |
| 5 | 2/9 |
| 6 | 1/9 |




$$\mu_{\overline{X}} = E(\overline{X}) = 2\times\frac{1}{9}+3\times\frac{2}{9}+4\times\frac{3}{9}+5\times\frac{1}{9}+6\times\frac{1}{9}=4$$

$$\sigma_{\overline{X}}^{\ 2} = (2-4)^2\frac{1}{9}+(3-4)^2\frac{2}{9}+(5-4)^2\frac{2}{9}+(6-4)^2\frac{1}{9}=1\frac{1}{3}$$

$$\sigma_{\overline{X}} = \sqrt{1\frac{1}{3}}=1.155$$

# Ⅳ. 표본평균의 확률분포

## 표본평균의 기댓값과 분산

$$E(\overline{X}) = E\left(\frac{1}{n}X_1 + \frac{1}{n}X_2 + \cdots + \frac{1}{n}X_n\right) = \frac{1}{n}\mu + \frac{1}{n}\mu + \cdots + \frac{1}{n}\mu = \mu.$$

$\because X_1 = X_2 = \cdots = X_n$으로 동일한 확률분포를 가지는 확률변수이고 각 확률변수의 기대값 $E(X_i) = \mu$

$$Var(\overline{X}) = Var\left(\frac{X_1 + X_2 + \cdots + X_n}{n}\right)$$

$$= \frac{1}{n^2}Var(X_1 + X_2 + \cdots + X_n)$$

$$= \frac{1}{n^2}(Var(X_1) + Var(X_2) + \cdots + Var(X_n))$$

$$= \frac{1}{n^2}(\sigma^2 + \sigma^2 + \cdots + \sigma^2) = \frac{\sigma^2}{n}.$$

- 표본평균의 평균은 모평균 μ와 같고,
- 표본평균의 분산은 모분산 $\sigma^2/n$이며,
- 표본평균의 표준편차는 $\sigma/\sqrt{n}$ 이다.
  (표본평균의 표준편차를 평균의 표준오차라고 함.)

# Ⅳ. 표본평균의 확률분포

## 표본평균의 표준오차

1. 평균의 표본분포의 기대값(또는 모평균)과 표본평균들 간의 차이 즉 편차제곱의 기대값의 제곱근으로 평균의 표본분포의 표준편차를 의미한다.
2. 표준오차는 표본평균을 모평균의 추정치로 사용할 때 예상되는 부정확성(오류)의 크기를 나타낸다.
3. 표준오차가 커지면 표본평균을 가지고 의사결정을 내릴 때 오류가 커지고 반대로 표준오차가 작으면 오류가 작아진다.
4. 표본의 크기 n이 커지면 표준오차는 작아진다.

$\sqrt{n}$

# V. 표본분포의 형태

## 표본크기의 영향

1. 평균의 표준오차는 표본평균을이 모평균에서 어느 정도 떨어져 있는지를 나타낸다.
2. 표본의 크기 n이 커지면 표준오차는 작아져 모평균을 추정하는 정확성이 증가한다.
3. 표본의 크기를 증가시키는 것은 시간과 비용을 수반하게 된다.
4. 표본의 크기가 증가할 수록 표본분포는 정규분포에 근접한다.

$\sqrt{n}$

# V. 표본분포의 형태

## 모집단 분포와 표본분포와의 관계

[ 모집단이 정규분포를 따를 때 ]

$X \sim N(\mu, \sigma^2)$을 따르는 모집단 분포에 대하여 확률표본을 복원추출하는 경우 표본평균($\overline{X}$)의 표본분포는 표본의 크기에 관계없이 $\overline{X} \sim N(\mu_{\overline{X}}, \sigma_{\overline{X}}^2) = N(\mu, \frac{\sigma^2}{n})$인 정규분포를 따르며 이를 정규표본분포라고 한다.

◆ 표본평균 $\overline{X}$의 분포를 Z값으로 표준화시키는 공식은 다음과 같다.

$$Z = \frac{\overline{X} - \mu_{\overline{X}}}{\sigma_{\overline{X}}} = \frac{\overline{X} - \mu_{\overline{X}}}{\frac{\sigma}{\sqrt{n}}}$$

# V. 표본분포의 형태

## 표본분포 예제

예제1)엑셀은행의 개인저축예금은 평균 2,000만원, 표준편차 600만원의 정규분포를 따른다고 한다.

1) 개인 저축예금(X)이 2,240만원 이하일 확률을 구하라.
(풀이) P(X≤2,240만원)=P(Z≤0.4)=0.5+P(0≤Z≤0.4)
=0.5+0.1554 =0.6554

2) 100명씩 무작위로 표본을 추출할 때 그의 평균예금 $\overline{X}$의 분포는 어떤 모양일까?
(풀이) X~N(2,000만원, $\frac{600}{\sqrt{100}}$ )

3) 평균예금($\overline{X}$)이 1,900만원 이상일 확률을 구하라.
(풀이) P($\overline{X}$ ≥1,900만원)=P(Z ≥ $\frac{1900 - 2,000}{600 / \sqrt{100}}$ )=P(Z ≥-1.67)
=0.5+0.4525=0.9525

# V. 표본분포의 형태

## 모집단 분포와 표본분포와의 관계

[ 모집단이 정규분포를 따르지 않을 때 ]

모집단의 분포가 정규분포가 아닌 경우 평균의 표본분포의 모양은 전적으로 표본크기에 달려 있다.

- 표본의 크기가 작을 때는 평균의 표본분포를 쉽게 규명할 수 없다
- 표본 크기가 클수록 모집단이 어떤 분포를 하든 상관없이 표본평균의 표본분포는 정규분포에 근접한다.(=중심극한 정리)
- 중심극한 정리(Central Limit Theorem)
  모집단을 이루는 확률변수 X의 분포가 정규분포가 아니더라도 표본의 크기 $n \geq 30$이면 평균의 표본분포는 정규분포에 근접한다는 정리이다.

# V. 표본분포의 형태

## 표본분포 예제

(예제2) 평균 15.3 표준편차 4.1인 모집단으로부터 n=36의 표본을 무작위로 추출하였을 때 표본평균 $\overline{X}$가 14이하일 확률을 구하라.

(풀이) n=36으로 대표본으로 평균의 표본분포는 중심극한 정리에 의해 정규분포를 따른다. 그러므로

$$\mu_{\overline{X}} = 15.3 \quad \sigma_{\overline{X}} = \frac{4.1}{\sqrt{36}} = 0.683$$

$$P(\overline{X} \le 14) = P(Z \le \frac{14 - 15.3}{0.683}) = P(Z \le -1.9) = 0.0287$$

1) P(a≤$\overline{X}$≤b)=0.95일 때 a와 b의 값을 구하라. 단, $\overline{X}$는 두 점 a와 b의 중간에 위치한다.

(풀이)

$$P(\frac{a-15.3}{0.683} \le Z \le \frac{b-15.3}{0.683}) = P(0 \le Z \le \frac{b-15.3}{0.683}) \times 2 = 0.95$$

$0.475$에 해당하는 $Z$값은 $1.96$이므로

$$\frac{b-15.3}{0.683} = 1.96 \ \therefore b = 16.64 \quad \frac{a-15.3}{0.683} = -1.96 \ \therefore a = 13.96$$

# V. 표본분포의 형태

## 모집단 분포와 표본분포와의 관계

[ 모집단 크기가 작을 때 ]

모집단이 작거나 유한하거나 비복원추출하는 경우 평균의 표준오차를 계산하는 경우(보통 n/N≥0.05인 경우 n:표본의 수 N:모집단 수) 표준오차에 유한모집단 조정계수를 곱해주어야 한다.

◆ 평균의 표준오차:유한모집단

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}} \quad N:\text{모집단 크기}\ \ n:\text{표본크기}$$

◆ 정규확률 변수 $\overline{X}$의 표준화

$$Z = \frac{\overline{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\overline{X} - \mu_{\bar{X}}}{\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}}$$

# VI. 비율의 표본분포

◆ 경영/경제문제에서는 성공비율을 분석해야 하는 경우가 있다. 예를 들면 시장점유률, 제품의 불량률, 정당 또는 후보의 지지율 등이 있는데 이를 모비율이라고 한다.

## 모비율과 표본비율

모비율 : $p = P(\text{성공}) = \frac{x}{N} = \frac{\text{모집단에서 발생하는 성공횟수}}{\text{모집단을 구성하는 모든 요소}}$

표본비율 : $\hat{p} = \frac{x}{n} = \frac{\text{표본에서의 성공횟수}}{\text{표본크기}}$

## 비율의 표본분포

비율의 표본분포란 모집단으로부터 동일한 표본크기 n을 무작위로 수없이 추출하여 그들의 비율을 구했을 때 나타나는 표본비율들의 확률분포를 말한다.

## 비율의 표본분포의 평균과 표준오차

$$E(\hat{p}) = p$$

$$\sigma_{\hat{p}}^{2} = \frac{p(1-p)}{n}$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$$

예제1) 빨간공 4개와 흰공 6개가 들어있는 상자에서 5개의 표본을 복원추출하는 경우 빨간공이 나타나는 비율의 표본분포와 기대값, 표준편차를 구하라.
(풀이)

| 성공횟수(X) | 표본비율(x/n) | 확률 |
|---|---|---|
| 0 | 0/5=0.0 | 0.078 |
| 1 | 1/5=0.2 | 0.259 |
| 2 | 2/5=0.4 | 0.346 |
| 3 | 3/5=0.6 | 0.230 |
| 4 | 4/5=0.8 | 0.077 |
| 5 | 5/5=1.0 | 0.010 |

$$E(\hat{p}) = p = 0.4$$

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.4(0.6)}{5}} = 0.219$$

## 이항분포의 정규근사법

성공비율의 연속성 조정계수

$$\hat{p} \pm \frac{1}{2n}$$

성공비율의 표준화 계수

$$Z = \frac{(\hat{p} - p)}{\sigma_{\hat{p}}}$$

(예2) 여론조사결과 대통령선거에서 강남 유권자의 50%가 투표에 참여하리라 한다. 유권자 20명을 무작위로 추출하였을 때 이 가운데 12명 이상이 투표에 참여할 확률을 정규근사법으로 구하라.

(풀이)

$$p = 0.5 \qquad \sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.5(0.5)}{20}} = 0.112$$

$$\hat{p} = \frac{12}{20} = 0.6$$

$$P(\hat{p} \geq 0.6) = P\left(\hat{p} \geq 0.6 - \frac{1}{2(20)}\right) = P(\hat{p} \geq 0.575)$$

$$P\left(Z \geq \frac{0.575 - 0.5}{0.112}\right) = P(Z \geq 0.67) = 0.2514$$