7. 표본 분포

담당교수 : 김 덕 기

toby123@cbnu.ac.kr

표본 분포(Sample Distribution)

표본추출(sampling)

$\overline{X}$, $\overline{P}$, s^2 ~ 통계량의 확률분포

n개 표본(sample)의 가지 수 : $_N C_n$

모집단(N)
(population)

$X_1, X_2, \cdots, X_N$

모수
(parameter)

μ, σ

$X_1, X_2, \cdots, X_n$ $\xrightarrow{\text{관측}}$ $x_1, x_2, \cdots, x_n$

$T(X_1, X_2, \cdots, X_n)$

통계량(statistic)

$T(x_1, x_2, \cdots, x_n)$

통계치

$$\overline{X} = \sum_{i=1}^{n} \frac{X_i}{n}$$

$$\overline{x} = \sum_{i=1}^{n} \frac{x_i}{n}$$

$\overline{X_1}, \overline{X_2}, \cdots, \overline{X_n}$

$\overline{x_1}, \overline{x_2}, \cdots, \overline{x_n}$

표본 평균의 표본분포(평균, 표준편차) ➔ $\mu_{\overline{x}} = \mu,\ \sigma_{\overline{x}} = \dfrac{\sigma}{\sqrt{n}}$

비율들이 이루는 평균, 분산 따라잡

평균의 표본 분포에서 평균, 분산

주사위2개 던지는 실험

크기 n=2의 복원방식의 단순확률표본의 평균 $\overline{X}$ 분포

번호	표본	평균	번호	표본	평균	번호	표본	평균
1	(1,1)	1.0	13	(3,1)	2.0	25	(5,1)	3.0
2	(1,2)	1.5	14	(3,2)	2.5	26	(5,2)	3.5
3	(1,3)	2.0	15	(3,3)	3.0	27	(5,3)	4.0
4	(1,4)	2.5	16	(3,4)	3.5	28	(5,4)	4.5
5	(1,5)	3.0	17	(3,5)	4.0	29	(5,5)	5.0
6	(1,6)	3.5	18	(3,6)	4.5	30	(5,6)	5.5
7	(2,1)	1.5	19	(4,1)	2.5	31	(6,1)	3.5
8	(2,2)	2.0	20	(4,2)	3.0	32	(6,2)	4.0
9	(2,3)	2.5	21	(4,3)	3.5	33	(6,3)	4.5
10	(2,4)	3.0	22	(4,4)	4.0	34	(6,4)	5.0
11	(2,5)	3.5	23	(4,5)	4.5	35	(6,5)	5.5
12	(2,6)	4.0	24	(4,6)	5.0	36	(6,6)	6.0

크기 n=2의 복원방식의 단순확률표본의 평균 $\overline{X}$의 확률분포

$\overline{x}$	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0
$P(\overline{X}=\overline{x})$	$\frac{1}{36}$	$\frac{2}{36}$	$\frac{3}{36}$	$\frac{4}{36}$	$\frac{5}{36}$	$\frac{6}{36}$	$\frac{5}{36}$	$\frac{4}{36}$	$\frac{3}{36}$	$\frac{2}{36}$	$\frac{1}{36}$

$$\mu_{\overline{X}} = E(\overline{X}) = 1 \times \frac{1}{36} + \cdots + 6 \times \frac{1}{36} = 3.5$$

$$\sigma^2_{\overline{X}} = Var(\overline{X}) = (1-3.5)^2 \times \frac{1}{36} + \cdots + (6-3.5)^2 \times \frac{1}{36} = 1.46$$

$$\boxed{E(\overline{X}) = E(X) = 3.5 \ , \ \ Var(\overline{X}) = \frac{1}{n} Var(X) = \frac{2.92}{2} = 1.46}$$

주사위1개 던지는 실험 : **E(X)=3.5, V(X)=2.92**

비율도 결국 오른쪽과 비슷하려면 베르누이 시행 (Ber(0,1))
을 따르는 것이 전제되어 있다.

중심극한정리(central limit theorem, C.L.T.)

표본평균 $\overline{X}$의 평균과 분산

평균 μ와 분산 σ^2을 갖는 무한모집단으로부터 단순확률추출에 의해 크기 n인 표본을 추출한다면, 표본평균 $\overline{X}$의 기댓값과 분산은 각각 다음과 같다.

$$E(\overline{X}) = \mu \ , \ Var(\overline{X}) = \frac{\sigma^2}{n}$$

$\boxed{\text{중심극한정리}}$

평균 μ와 분산 σ^2을 갖는 무한모집단으로부터 단순확률추출에 의해 크기 n의 표본을 추출한다면, 표본평균 $\overline{X}$는 표본크기 n이 커짐에 따라 근사적으로 평균이 μ이고 분산이 σ^2/n인 정규분포를 따른다.

따라서 표준화 확률변수 $Z = \dfrac{\overline{X} - \mu}{\sigma / \sqrt{n}}$의 분포는 표본크기 n이 커짐에 따라 근사적으로 표준정규분포 $N(0,1)$을 따른다.

$X_1, X_2, \cdots, X_n$: 평균 μ, 분산 σ^2 인 임의의 모집단으로부터의 확률표본

➡ n이 충분히 클 때, $\overline{X} \stackrel{\bullet}{\sim} N\left(\mu, \dfrac{\sigma^2}{n}\right) \Leftrightarrow Z = \dfrac{\overline{X} - \mu}{\sigma / \sqrt{n}} \stackrel{\bullet}{\sim} N(0,1)$

주1) 대체적으로 $n \geq 25$(또는 30)이면, 근사 정도가 만족할 만 하다.

중심극한정리(central limit theorem, C.L.T.)

$X_1, X_2, \cdots, X_n$: 모평균 μ, 모분산 σ^2인 모집단으로부터의 확률표본

➡ 표본평균 $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$

● **표본평균의 분포에 대한 성질 :** $E(\overline{X}) = \mu$, $Var(\overline{X}) = \dfrac{\sigma^2}{n}$, $sd(\overline{X}) = \dfrac{\sigma}{\sqrt{n}}$

표본의 크기가 클 때 $\overline{X}$는 모집단의 평균인 μ 근처에 밀집되어 분포!

[1] 정규분포 $N(\mu, \sigma^2)$인 경우 ➡ $\overline{X} \sim N\left(\mu, \dfrac{\sigma^2}{n}\right)$

확률로 오류시 계산(X)라 한다.

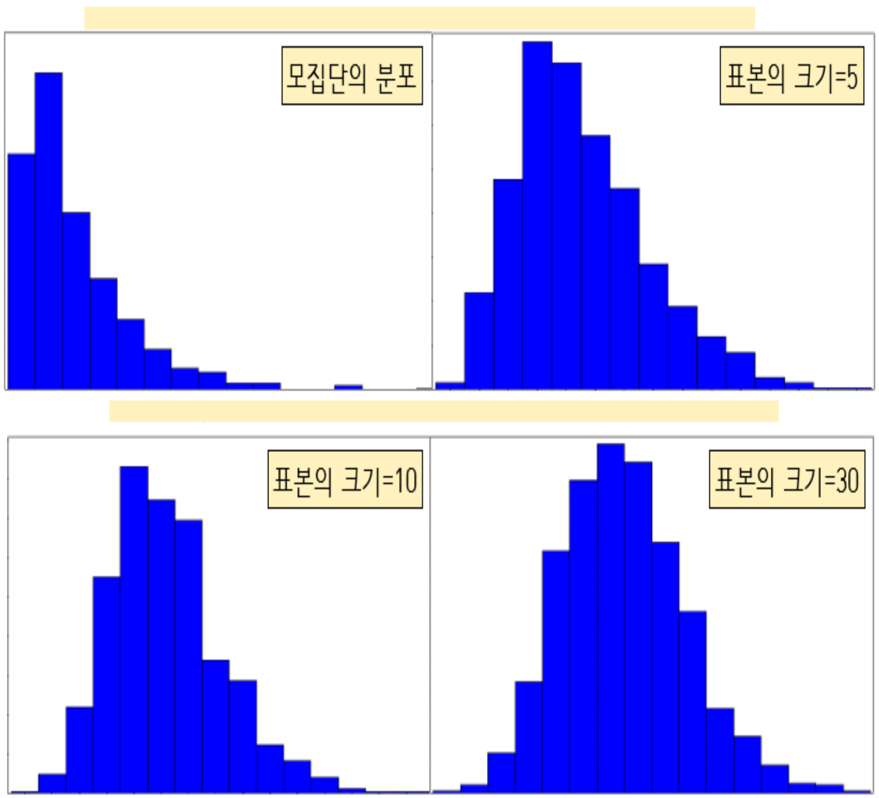
[2] 정규분포가 아닌 경우 ➡ ?

대안 : 비모수적 방법, 분포무관

예제 : 충북대학교의 전체 교수의 수는 800명이고 평균연령은 45.8세이고 표준편차가 15세이다. 만약 크기가 100명인 표본을 모두 추출할 경우 이 평균연령의 표본분포의 평균과 표준편차는? 또한 평균연령의 확률분포는 ?

표본 크기에 따른 표본 분포의 형태

- 표본의 크기 변화에 따른 표본평균의 분포



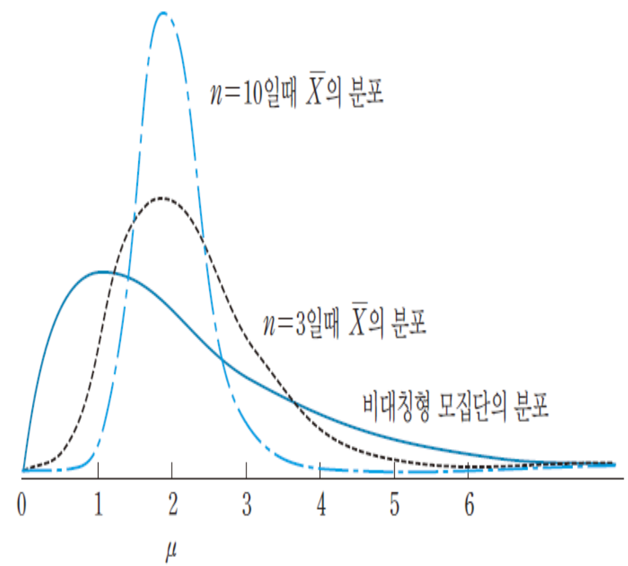


그림 7-2 비대칭형 모집단에서 $\bar{X}$의 분포($n=3$과 $n=10$)

ex) 이 문제에서 '정규분포'를 빼면 1) → 계산 X, 2) → 계산 O

2)에서 49개 → 20개 CLT 적용X 계산 X ... (좀 더 확실하지 않으므로 확인 필요)

모집단 분포와 표본분포 : 확률계산

예제2 신안전 타이어 공업주식회사에서는 새로운 형태의 광폭 타이어를 생산하고 있다. 이 타이어의 **평균수명이 50,000km이고 표준편차는 14,000km인 정규분포**를 하고 있다고 알려져 있다. $X \sim$ 제품수명 (km) $X \sim N(50000, 14000^2)$

(1) 수명이 60,500km이상인 타이어는 전체 생산량의 몇 %가 되는가?

(2) 만일 크기가 49개인 표본을 뽑는다면 이 표본의 평균이 48,000km 이하가 될 확률은 얼마인가?

모집단 :

(1) $$P(X \geq 60{,}500) = P\left(Z \geq \frac{X-\mu}{\sigma}\right) = P\left(Z \geq \frac{60{,}500 - 50{,}000}{14{,}000}\right)$$
$$= P(Z \geq 0.75) = 1 - 0.7734 = 0.2266\,(22.66\%)$$

표본집단 :

(2) $$\mu_{\bar{x}} = E(\bar{X}) = 50{,}000km, \quad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{14{,}000}{7} = 2{,}000km$$

$$P(\bar{x} \leq 48{,}000) = P\left(Z \leq \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}}\right) = P\left(Z \leq \frac{48{,}000 - 50{,}000}{2{,}000}\right)$$
$$= P(Z \leq -1.0) = 1 - 0.8413 = 0.1587\,(15.87\%)$$

CLT를 이용한 표본분포의 확률계산

예제 3) 평균이 82이고 분산이 144인 모집단으로부터

1. 크기 64인 표본의 표본평균이 80.8에서 83.2 사이에 있을 확률은?
2. 크기 100인 표본의 표본평균이 80.8에서 83.2 사이에 있을 확률은?

- $n = 64$일 때, 근사적으로 $\overline{X} \sim N\left(82, \frac{144}{64} = \left(\frac{3}{2}\right)^2\right)$이다. 따라서

$$P(80.8 \le \overline{X} \le 83.2) = P\left(\frac{80.8-82}{3/2} \le \frac{\overline{X}-82}{3/2} \le \frac{83.2-82}{3/2}\right)$$
$$= P(-0.8 \le Z \le 0.8) = 0.5762$$

- $n = 100$일 때, 근사적으로 $\overline{X} \sim N\left(82, \frac{144}{100} = \left(\frac{6}{5}\right)^2\right)$이다. 따라서

$$P(80.8 \le \overline{X} \le 83.2) = P\left(\frac{80.8-82}{6/5} \le \frac{\overline{X}-82}{6/5} \le \frac{83.2-82}{6/5}\right)$$
$$= P(-1 \le Z \le 1) = 0.6826$$

n이 증가함에 따라 표본 평균의 분포가 모집단의 평균을 중심으로 더 집중되어 나타남

이항 분포의 정규분포 근사1

- 이항분포 $Bin(n,p)$에서 n이 매우 큰 경우는 직접 확률을 구하는 것이 쉽지 않다.
  - n이 매우 크고 p가 충분히 작은 경우는 포아송 근사를 이용하여 확률을 구할 수 있다. 매우 희박한 확률

- 이항분포 $Bin(n,p)$에서 n이 매우 크고 p가 0이나 1에 가깝지 않아서 np와 $n(1-p)$ 모두 충분히 큰 경우에 (보통 $np \geq 10, n(1-p) \geq 10$) 이항분포는 정규분포에 가까워진다.

$B(n,p) \approx N(\mu, \sigma^2)$

$\mu = np$

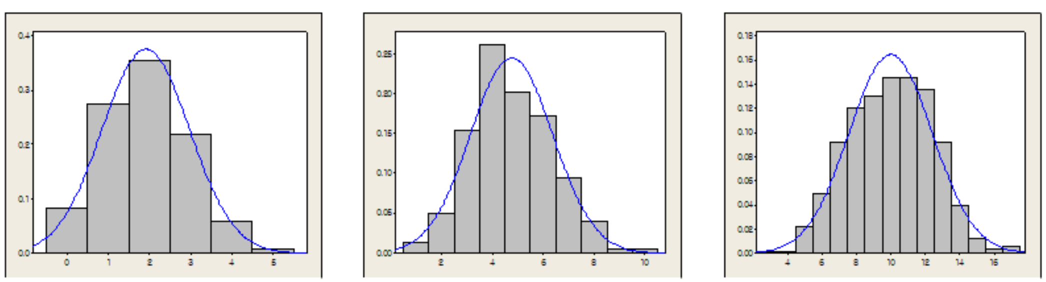
$\sigma^2 = np(1-p)$

① $n \uparrow$ by. CLT

② $p \approx 0.5$

$\Rightarrow$ 정규분포

- $p = 0.4$일 때, $n = 5, 12, 25$인 이항분포의 확률히스토그램



$n=5 \quad p=0.4$ $\qquad$ $n=12 \quad p=0.4$ $\qquad$ $n=25 \quad p=0.4$

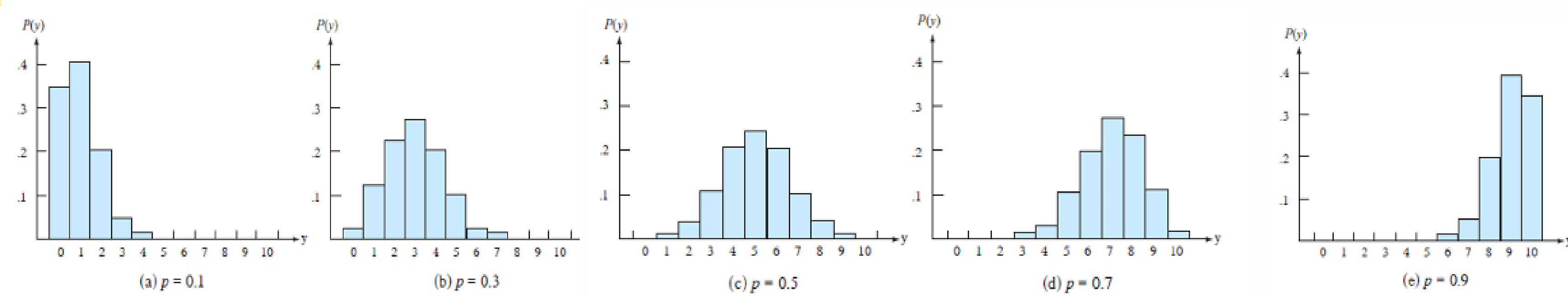
n이 커질수록 종 모양의 정규분포에 가까워진다.

이항 분포의 정규분포 근사2

[중심극한정리, central limit theorem]

$X \sim Bin(n, p)$이고 np와 $n(1-p)$ 모두 충분히 클 때, X는 근사적으로 평균이 np, 분산이 $np(1-p)$인 정규분포를 따른다. 즉

$$Z = \frac{X - np}{\sqrt{np(1-p)}} \sim N(0, 1)$$



n=10일 때 p=0.1~0.9에 따른 이항분포 ➔ n(커지고), p~1/2에 가까우면 정규분포에 가까워짐.

이항 분포의 정규분포 근사3

예) 어느 항공사 A노선의 예약자 중 노쇼(no-show)의 비율이 10%로 알려져 있다.
좌석 320석 비행기에 350명분의 예약을 받았을 때 비행기가 자리가 모자라 예약 승객이 탑승하지 못할 확률은?

[풀이] X를 350명 예약자 중에서 나타나는 사람의 수라고 하면 $X \sim Bin(350,\ 0.9)$ 이다.
따라서 원하는 확률은 $P(X \geq 321)$이다. $np = 315, n(1-p) = 35 \geq 10$ 이므로 정규근사 조건을 만족한다. 중심극한정리를 이용하면

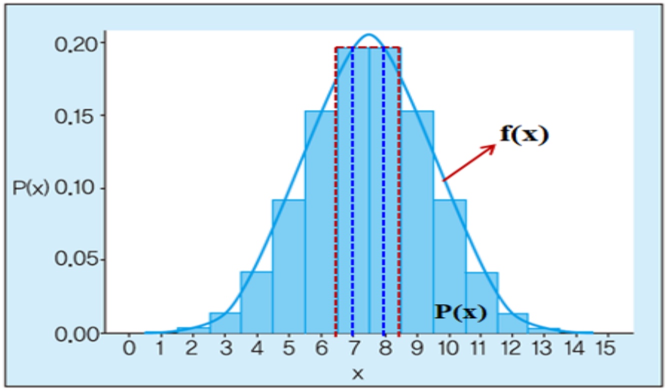
$$P(X \geq 321) = P\left(\frac{X - np}{\sqrt{np(1-p)}} \geq \frac{321 - 315}{\sqrt{350 * 0.9 * 0.1}}\right) \approx P(Z \geq 1.069) = 0.1425$$

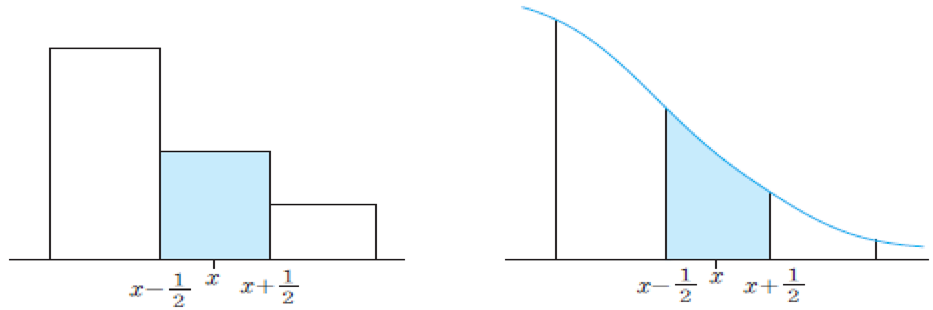
가 된다.

만약 원하는 확률을 $P(X > 320)$로 하여 정규근사를 하면 $P(Z > 0.89087) = 0.1865$

- 참고로 이항분포를 이용하여 원하는 정확한 확률을 구하면 0.163601

이항 분포의 정규분포 근사 : 연속성 수정





(i) (ii)

$$P(X = x) = P\left(x - \frac{1}{2} \le X \le x + \frac{1}{2}\right)$$

막대의 끝점

이항확률분포를 정규분포로 근사할 때,

$$P(a \le X \le b) = P\left(a - \frac{1}{2} \le X \le b + \frac{1}{2}\right)$$

와 같이 $\frac{1}{2}$ 씩 가감하여 확률의 근삿값을 구할 수 있다.

- $P(X = x) = P\left(x - \frac{1}{2} \le X \le x + \frac{1}{2}\right)$
- $P(a < X < b) = P\left(a + \frac{1}{2} \le X \le b - \frac{1}{2}\right)$
- $P(a < X \le b) = P\left(a + \frac{1}{2} \le X \le b + \frac{1}{2}\right)$

연속성 수정 : 예제

예제) 항공사 예약노쇼의 문제 $X \sim Bin(350, 0.9)$일 때,

$$P(X \geq 321) = P(X \geq 320.5) = P\left(\frac{X-np}{\sqrt{np(1-p)}} \geq \frac{320.5-315}{\sqrt{350*0.9*0.1}}\right)$$

$$\approx P(Z \geq 0.9799) = 0.16355$$

노쇼비율 : 10%

- 이항분포를 이용하여 원하는 정확한 확률을 구하면

$$P(X \geq 321)$$

$$= \sum_{321}^{350} \binom{350}{x} 0.9^x 0.1^{350-x} = 1 - \text{pbinom(320,350,0.9)} = 0.163601$$

예제8) $X \sim Bin(150, 0.6)$일 때,

- $P(82 \leq X \leq 101)$의 값은?
- $P(X > 97)$의 값은?

$np = 150 \times 0.6 = 90, np(1-p) = 36$로
$N(90, 6^2)$을 따른다.

$$P(82 \leq X \leq 101) = P(81.5 \leq X \leq 101.5)$$

$$= P\left(\frac{81.5-90}{6} \leq \frac{X-90}{6} \leq \frac{101.5-90}{6}\right)$$

$$= P(-1.42 \leq Z \leq 1.92) = 0.8948$$

$$P(X > 97) = P(X \geq 97.5) = P\left(\frac{X-90}{6} \geq \frac{97.5-90}{6}\right)$$

$$= P(Z \geq 1.25) = 0.1056$$

연속성 수정을 하지 않아도 되는 경우

- $np(1-p)$이 충분히 크면 연속성 수정에 큰 영향을 받지 않으므로 연속성 수정을 하지 않아도 된다. (why? 이항분포의 정규근사에서 분모에 있는 $\sqrt{np(1-p)}$의 값이 크므로)

- **예제9)** 성인의 30%가 알코올 음료 섭취
  X = 성인 1000명 중 알코올 음료 섭취하는 사람의 수 $\sim B(1000, 0.3)$
  - $P(X < 280)$의 값은?
  - $np = 1000 \times 0.3 = 300, np(1-p) = 210$로 충분히 크므로 X는 근사적으로 $N(300, 210)$을 따른다.

  1. (연속성 수정을 한 경우)
     $$P(X < 280) = P(X \le 279.5) = P\left(\frac{X-300}{\sqrt{210}} \le \frac{279.5-300}{\sqrt{210}}\right)$$
     $$= P(Z \le -1.41) = 0.0793$$
  2. (연속성 수정을 하지 않은 경우)
     $$P(X < 280) = P(X \le 280) = P\left(\frac{X-300}{\sqrt{210}} \le \frac{280-300}{\sqrt{210}}\right)$$
     $$= P(Z \le -1.38) = 0.0838$$

발산 ⇒ 개념만 이해 (계산 X)

보통 (μ, σ^2)

비율 $\left(p, \frac{p(1-p)}{n}\right)$

$N\begin{pmatrix} X_1 \cdots x_n \\ \theta = \mu, \sigma^2, p \end{pmatrix}$

비율의 표본분포

$X \sim \left(p, \frac{\sigma^2}{n}\right)$

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

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

표본추출(sampling)

X_i : 이항변수(0 또는 1), 표본군=m

모집단(N)
(population)

$X_1, X_2, \cdots, X_N$

모수
(parameter)

$P = \frac{\sum_{i=1}^{N} X_i}{N}$

$X_1, X_2, \cdots, X_n \xrightarrow{\text{관측}} x_1, x_2, \cdots, x_n$

$T(X_1, X_2, \cdots, X_n)$ → 통계량(statistic)

$T(x_1, x_2, \cdots, x_n)$ → 통계치

$Y = \sum_{i=1}^{n} X_i \, , \; \bar{P} = \frac{Y}{n}$ → $\bar{p} = \frac{y}{n}$

$\bar{P}_1 = \frac{Y_1}{n}, \cdots, \bar{P}_m = \frac{Y_m}{n}$ → $\bar{p}_1 = \frac{y_1}{n}, \cdots, \bar{p}_m = \frac{y_m}{n}$

표본비율의 평균과 분산 ➔ $\mu_{\bar{p}} = p \; , \; \sigma_{\bar{p}}^2 = \frac{p(1-p)}{n}$

비율의 표본분포 : example

$\bar{p} \sim N(p, \frac{p(1-p)}{n})$

[예제] 어느 감기약의 치유율은 90%라고 알려져 있다. 올해에 유행하고 있는 감기에 대해서도 90%의 치유율을 보장할 수 있는가를 알아보기 위해 100명의 감기환자에게 감기약을 투여하였다. 이들 중 감기로부터 완전히 회복된 사람들의 비율이 85%~95%내에 들어 있을 확률은 ?

$$by\ C.L.T\ :\ \hat{p} \sim N\left(0.9,\ \frac{0.9(1-0.9)}{100}\right) \rightarrow P(0.85 < \hat{p} < 0.95) \simeq P(-1.67 < Z < 1.67) = 0.905$$

➔ 이항 분포를 이용해 구해 보아라.

➔ 이항 분포의 정규분포 근사 방법을 이용해 구해 보아라.

➔ 이항 분포의 정규분포 근사 시 연속성 수정을 이용해 구해 보아라.