CHAPTER

06

표본의 분포

1

제6장
표본의 분포

2/28

2

1. 통계량과 추정량

[정의 9-1] 확률표본(random sample)
독립적이며 동일한 분포를 따르는 (iid: independent and identically distributed) 확률변수들의 집합

[정의 9-2] 통계량(statistic)
미지의(unknown) 모수를 포함하지 않는 확률표본의 함수

[정의 9-3] 추정량(estimator)
미지의 모수를 추정하기 위한 통계량

- 표본평균 $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
- 표본분산 $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$

3/28

3

1. 통계량과 추정량

[정의 9-4] 불편성(unbiasedness) $E(\hat{\theta}) = \theta$
: 추정량의 기대값이 추정하고자 하는 모수와 같아지는 특성으로서, 좋은 추정량이 되기 위한 첫 번째 요건

- 표본평균의 불편성

$$E(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n} \mu = \mu$$

- 표본평균의 분산

$$Var(\bar{X}) = Var\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i) = \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{\sigma^2}{n}$$

✓ 정규 모집단에서 일정한 개수의 확률표본으로 만들 수 있는 모평균에 대한 불편추정량 중 표본평균의 분산이 최소(UMVUE)

4/28

4

2. 정규모집단 표본평균의 분포

[정리 9-1] 정규모집단 표본평균의 분포

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2) \;\Rightarrow \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n})$$

[증명] $$Y = \sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2) \;\Rightarrow \bar{X} = \frac{1}{n} Y \sim N(\mu, \frac{\sigma^2}{n})$$

[따름정리 9-1] 정규모집단 표준화된 표본평균의 분포

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2) \;\Rightarrow \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1)$$

[증명] $$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \;\Rightarrow\; E(Z) = 0, Var(Z) = 1$$

5/28

5

2. 정규모집단 표본평균의 분포

[예 9-1] 평균 100, 분산 100인 정규분포 모집단에서 10개씩 표본을 랜덤 추출하여 표본평균을 구하는 작업 10,000회 반복

(1) 표본평균 히스토그램, [정리 9-1] 확인

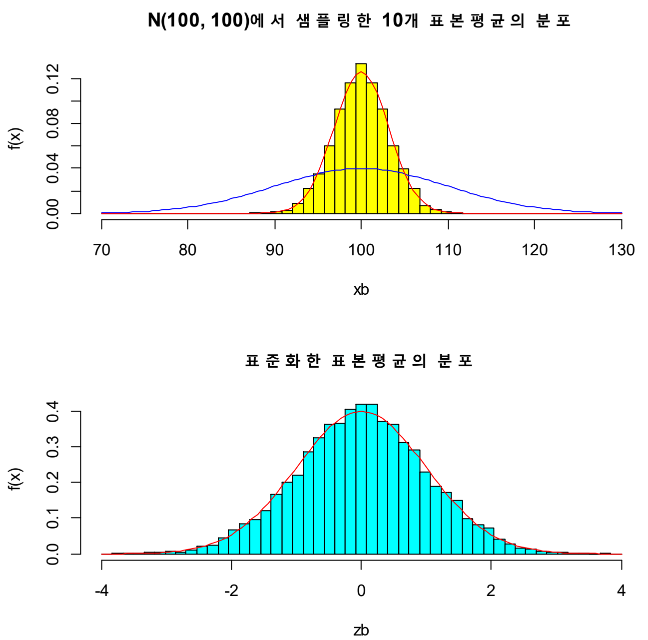
$$\Rightarrow \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(100, \frac{100}{10})$$

(2) 표본평균 표준화 히스토그램, [따름정리 9-1] 확인

$$\Rightarrow \frac{\bar{X} - 100}{10 / \sqrt{10}} \sim N(0,1)$$

6/28

6



2. 정규모집단 표본평균의 분포

[예 9-2] 초콜릿 한 개의 무게는 모평균 μ(g), 모표준편차 5(g)인 정규분포를 따름. 초콜릿 16개를 랜덤하게 샘플링하여 무게를 측정했을 때, 표본평균과 모평균과의 차이가 2(g) 이내일 확률?

$$X \sim N(\mu,\ 5^2) \qquad \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} = \frac{\bar{X}-\mu}{5/\sqrt{16}} \sim N(0,1)$$

$$P(|\bar{X}-\mu| \le 2) = P(-2 \le \bar{X}-\mu \le 2)$$

$$= P\left(-\frac{2}{5/\sqrt{16}} \le \frac{\bar{X}-\mu}{5/\sqrt{16}} \le \frac{2}{5/\sqrt{16}}\right)$$

$$= P(-1.6 \le Z \le 1.6) \doteq 0.8904$$

```
n <- 16
# 방법 1 (직접계산), mu=200 (임의의 수 입력)
pnorm(200+2, 200, 5/sqrt(n)) - pnorm(200-2, 200, 5/sqrt(n))
[1] 0.8904014
# 방법 2 (표준화)
pnorm(2/(5/sqrt(n))) - pnorm(-2/(5/sqrt(n)))
[1] 0.8904014
```

2. 정규모집단 표본평균의 분포

[예 9-3] [예 9-2]에서 표본평균과 모평균과의 차이가 2(g) 이내가 될 확률이 0.95 이상이 되려면 최소 몇 개의 표본이 필요한가?

$$P(|\bar{X}-\mu|\le 2) = P(-2\le \bar{X}-\mu \le 2) \ge 0.95$$

$$P\left(-\frac{2}{5/\sqrt{n}} \le \frac{\bar{X}-\mu}{5/\sqrt{n}} \le \frac{2}{5/\sqrt{n}}\right) = P(-0.4\sqrt{n} \le Z \le 0.4\sqrt{n})$$

$$= \Phi(0.4\sqrt{n}) - \Phi(-0.4\sqrt{n}) = 2\Phi(0.4\sqrt{n}) - 1 \ge 0.95$$

$$\Rightarrow \Phi(0.4\sqrt{n}) \ge 0.975 \;\Rightarrow 0.4\sqrt{n} \ge \Phi^{-1}(0.975) \equiv z_{0.975} \doteq 1.96$$

$$\Rightarrow n \ge (1.96/0.4)^2 \ge 24.01 \;\Rightarrow n \ge 25$$

```
# 직접 계산
ceiling((qnorm(0.975)/0.4)^2)
[1] 25
# 간단한 함수 정의 (p=신뢰수준, sigma=오차한계(표준편차의 배수))
findsn <- function(p, sigma) ceiling((qnorm(1-(1-p)/2)/sigma)^2)
findsn(0.95, 0.4); findsn(0.99, 0.4); findsn(0.95, 0.2); findsn(0.99, 0.2)
[1] 25    [1] 42    [1] 97    [1] 166
```

9/28

9

2. 정규모집단 표본평균의 분포

[예 9-4] 정규모집단에서 추출한 표본평균과 모평균과의 차이가 kσ 이내가 될 확률이 1-α 이상이 되기 위한 최소 표본개수의 식을 구하고, 그래프를 작성하여 그 변화에 대해 분석

$$P(|\bar{X}-\mu|\le k\sigma) = P(-k\sigma\le \bar{X}-\mu \le k\sigma) \ge 1-\alpha$$

$$P\left(-\frac{k\sigma}{\sigma/\sqrt{n}} \le \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \le \frac{k\sigma}{\sigma/\sqrt{n}}\right) = P(-k\sqrt{n} \le Z \le k\sqrt{n})$$

$$= \Phi(k\sqrt{n}) - \Phi(-k\sqrt{n}) = 2\Phi(k\sqrt{n}) - 1 \ge 1-\alpha$$

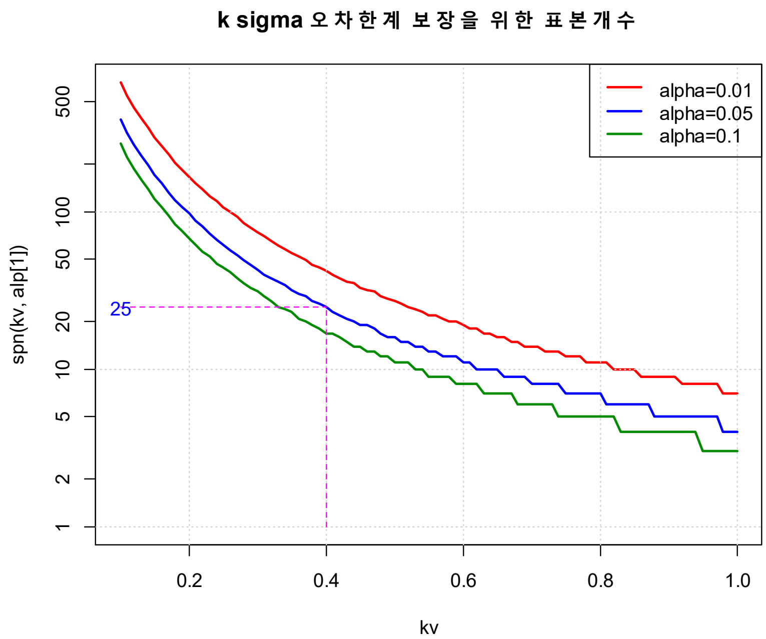
$$\Rightarrow \Phi(k\sqrt{n}) \ge 1-\alpha/2$$

$$\Rightarrow k\sqrt{n} \ge \Phi^{-1}(1-\alpha/2) \equiv z_{1-\alpha/2}$$

$$\Rightarrow n \ge \left[\frac{\Phi^{-1}(1-\alpha/2)}{k}\right]^2$$

10/28

10



2. 정규모집단 표본평균의 분포

[정리 9-2] 정규모집단 표본평균의 분포(σ를 모르는 경우)

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2) \;\Rightarrow \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$$

$$Y = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} \sim \chi^2(n-1)$$

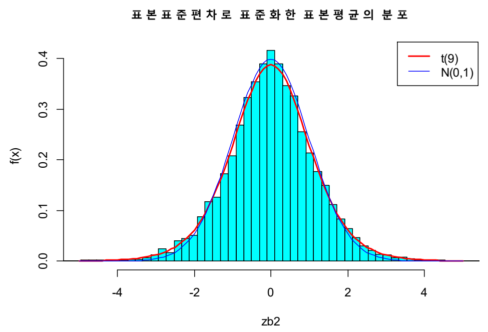
$$T = \frac{\overline{X}-\mu}{S/\sqrt{n}} = \frac{\dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)S^2}{(n-1)\sigma^2}}} = \frac{Z}{\sqrt{\chi^2_{(n-1)}/(n-1)}} \sim t(n-1)$$

2. 정규모집단 표본평균의 분포

[예 9-5] 평균 100, 분산 100인 정규분포 모집단에서 10개씩 표본을 랜덤 추출하여 표본평균과 표본분산 구하기 10,000회 반복

(1) 모분산을 모른다는 가정 하에서 표본평균들을 표준화하여 히스토그램을 그리고, [정리 9-2] 확인

(2) 표준정규분포와 어떤 차이가 있는지 비교 $\frac{\bar{X}-100}{S/\sqrt{10}} \sim t(9)$



3. 정규모집단 표본분산의 분포

[정의 8-3] 카이제곱분포(chi-square distributiion)

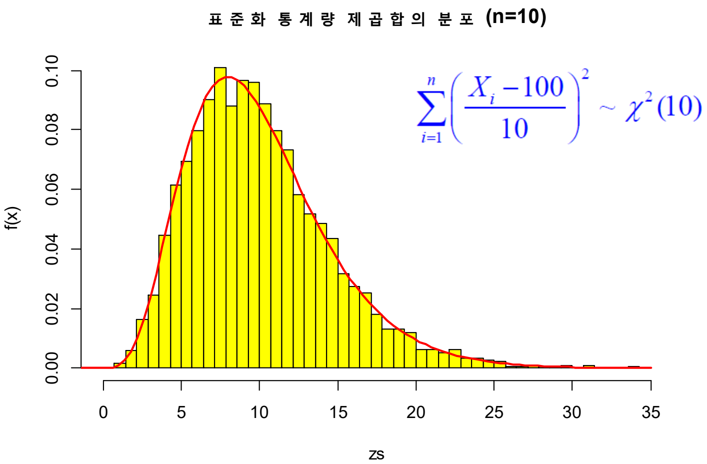
$$Z_1, Z_2, \ldots, Z_n \overset{iid}{\sim} N(0,1) \Rightarrow \sum_{i=1}^{n} Z_i^2 \sim \chi^2(n)$$

- 기대값 $E(X) = \nu$
- 분산 $Var(X) = 2\nu$

[정리 9-3] 표준정규 확률표본 제곱합의 분포

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2) \Rightarrow \sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2(n)$$

[예 9-6] N(100,10^2)에서 10개씩 표본 랜덤 추출, 표준화 제곱합 통계량 계산 10,000회 반복 → 히스토그램 작성 및 [정리 9-3] 확인



15/28

3. 정규모집단 표본분산의 분포

[정리 9-4] 정규모집단 표본분산의 분포

$$X_1, X_2, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2) \Rightarrow \frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 \sim \chi^2(n-1)$$

[증명]

$$\sum_{i=1}^{n}(X_i-\mu)^2 = \sum_{i=1}^{n}\left[(X_i-\bar{X})+(\bar{X}-\mu)\right]^2 \qquad S^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$$

$$= \sum_{i=1}^{n}(X_i-\bar{X})^2 + \sum_{i=1}^{n}(\bar{X}-\mu)^2 + 2(\bar{X}-\mu)\sum_{i=1}^{n}(X_i-\bar{X})$$

$$= \sum_{i=1}^{n}(X_i-\bar{X})^2 + n(\bar{X}-\mu)^2 \qquad \sum_{i=1}^{n}(X_i-\bar{X}) = 0$$

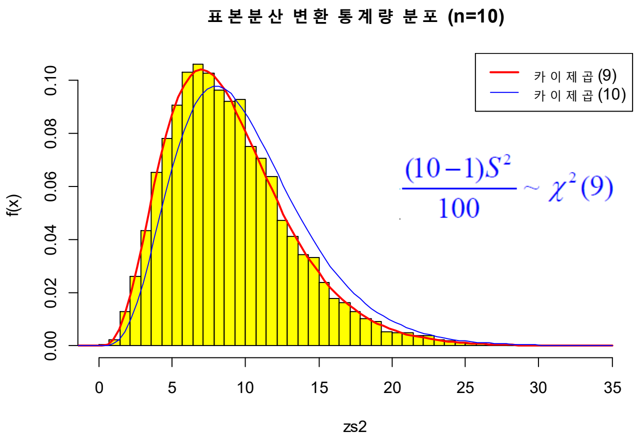
$$\Rightarrow \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\bar{X})^2 + \frac{(\bar{X}-\mu)^2}{\sigma^2/n}$$

$$\Rightarrow (1-2t)^{-n} = m(t)\times(1-2t)$$

$$\Rightarrow m(t) = (1-2t)^{-(n-1)}$$

16/28

[예 9-7] 평균 100, 분산 100인 정규분포 모집단에서 10개씩 표본 랜덤 추출 $(10-1)S^2/100$ 계산 10,000회 반복 → 히스토그램 작성 및 [정리 9-4] 확인



3. 정규모집단 표본분산의 분포

[예 9-8] 정규분포 모집단으로부터 n=5, 10, 30, 100개의 확률표본을 추출하여 표본분산 계산 → $P(S^2 \le c\sigma^2) = 0.95$

$$P(S^2 \le c\sigma^2) = P\left(\frac{(n-1)S^2}{\sigma^2} \le (n-1)c\right) = 0.95$$

$$\Rightarrow (n-1)c = \chi^2_{0.95;n-1}$$

```
# 자유도, 누적확률 입력
nu <- c(5, 10, 30, 100)
p <- 0.95

# 분위수/자유도 (상수 c)
qchisq(p, nu-1) / (nu-1)
[1] 2.371932 1.879886 1.467482 1.244699
```

4. 두 정규모집단 표본분산비의 분포

[정의 8-5] F-분포

$$U \sim \chi^2(\nu_1) \perp V \sim \chi^2(\nu_2) \Rightarrow F \equiv \frac{U/\nu_1}{V/\nu_2} \sim F(\nu_1, \nu_2)$$

[정리 9-5] 두 정규모집단 표본분산비의 분포

$$S_1^2 \perp S_2^2 \Rightarrow \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1-1, n_2-1)$$

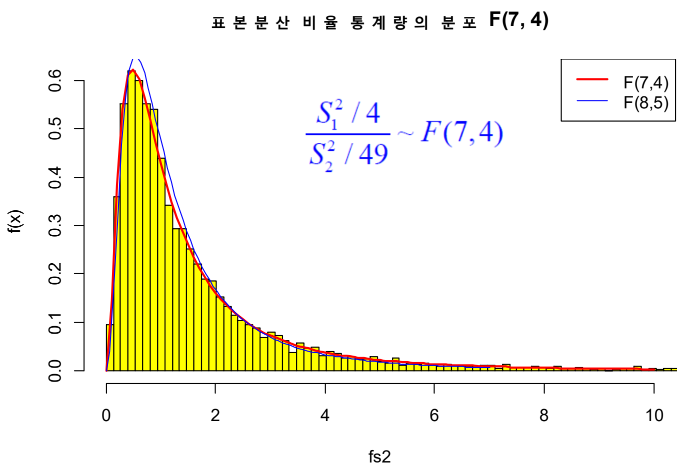
[증명] $U = \dfrac{(n_1-1)S_1^2}{\sigma_1^2}, V = \dfrac{(n_2-1)S_2^2}{\sigma_2^2} \Rightarrow U \sim \chi^2(n_1-1) \perp V \sim \chi^2(n_2-1)$

$$\frac{\dfrac{S_1^2}{\sigma_1^2}}{\dfrac{S_2^2}{\sigma_2^2}} = \frac{\dfrac{(n_1-1)S_1^2}{(n_1-1)\sigma_1^2}}{\dfrac{(n_2-1)S_2^2}{(n_2-1)\sigma_2^2}} = \frac{\dfrac{U}{n_1-1}}{\dfrac{V}{n_2-1}} \sim F(n_1-1, n_2-1)$$

19/28

19

[예 9-9] 분산이 각각 4와 49인 독립적인 정규분포 모집단에서 각각 8개, 5개씩 표본을 랜덤 추출하여 표본분산비 계산 10,000회 반복. 히스토그램 작성 및 [정리 9-5] 확인



20/28

20

4. 두 정규모집단 표본분산비의 분포

[예 9-10] 모분산이 같다고 알려져 있는 독립적인 두 정규모집단으로부터 각각 n1=12, n2=16인 두 확률표본 추출

(1) $P(S_1^2 / S_2^2 > a) = 0.05$

$$\sigma_1^2 = \sigma_2^2 \Rightarrow \quad F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2}{S_2^2} \sim F(11,15)$$

```
qf(0.05, 11, 15, lower.tail=FALSE)
[1] 2.506806
```

(2) 표본분산의 비율이 2배 이상 차이 날 확률

$$P(F \le 0.5) + P(F \ge 2)$$

```
pf(0.5, 11, 15) + 1-pf(2, 11, 15)
[1] 0.2307009
```

21/28

21

5. 중심극한정리

[정리 9-6] 중심극한정리(central limit theorem)

: 기대값과 분산이 일정한 확률표본에 대하여

$$S_n = X_1 + X_2 + \cdots + X_n \overset{a}{\sim} N(n\mu, n\sigma^2)$$

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \overset{a}{\sim} N(\mu, \frac{\sigma^2}{n})$$

- 표준화 $\dfrac{S_n - n\mu}{\sqrt{n}\sigma} = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$

[예 9-11]* 다음의 분포에 대해 중심극한정리를 적용하여 표준화 통계량의 히스토그램과 정규확률도를 작성하고, 수렴결과를 비교

$$t(3),\ Exp(5),\ \Gamma(0.5,10),\ Poi(1) \qquad (n = 10, 30, 50)$$

22/28

22

5. 중심극한정리

[예 9-12] 게임 기능사 자격시험 점수가 B(100, 0.7)을 따른다고 할 때, 랜덤하게 추출된 40명의 표본평균이 69점 이하일 확률

$$E(X) = np = 70,\ Var(X) = np(1-p) = 21 \Rightarrow \bar{X} \overset{a}{\sim} N(70,\ 21/40)$$

$$P(\bar{X} \le 69) \simeq P(Z \le \frac{69-70}{\sqrt{21/40}}) \doteq \Phi(-1.380) \doteq 0.084$$

```
pnorm(69, 70, sqrt(21/40))
[1] 0.08377314
```

[예 9-13] [예 9-12]의 확률을 정확하게 구하시오.

$$Y \sim B(4000,\ 0.7)$$

$$P(\bar{X} \le 69) = P(Y \le 40 \times 69) = P(Y \le 2760)$$

```
pbinom(2760, 4000, 0.7)
[1] 0.08676363
```

23/28

23

5. 중심극한정리

[예 9-14] 어느 커피가게의 대기시간은 평균이 10분인 지수분포.
랜덤하게 선택된 고객 40명의 평균 대기시간이 9분 이내일 확률

$$\bar{X} \sim N(10,\ 100/40)$$

$$P(\bar{X} \le 9) \simeq P(Z \le \frac{9-10}{\sqrt{2.5}}) \doteq \Phi(-0.6325) \doteq 0.2635$$

```
pnorm(9, 10, sqrt(100/40))
[1] 0.2635446
 pnorm(9.5, 10, sqrt(100/40))
[1] 0.3759148
```

[예 9-15]* [예 9-14]의 확률을 정확하게 구하시오. ([정리 7-2])

$$Y \sim \Gamma(40,\ 10) \qquad P(\bar{X} \le 9) = P(Y \le 40 \times 9) = P(Y \le 360)$$

```
pgamma(360, 40, scale=10)
[1] 0.2736964
```

24/28

24

6. 정규모집단이 아닌 경우의 표본분포

[정리 9-7] 표본평균의 분포(표본이 매우 큰 경우)

: 일정한 기댓값과 분산을 갖는 확률표본

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \overset{a}{\sim} N(\mu, \frac{\sigma^2}{n}) \qquad \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \overset{a}{\sim} N(0,1)$$

[예 9-16] 주사위를 랜덤하게 40번 굴렸을 때 나온 눈의 평균이 3과 4 사이에 들어갈 확률

$$\mu = \frac{6+1}{2} = 3.5 \quad \sigma^2 = \frac{(6+1)(6-1)}{12} = \frac{35}{12} \quad \Rightarrow \bar{X} \overset{a}{\sim} N(3.5, \frac{35}{12\times 40})$$

$$P(3 < \bar{X} < 4) \approx P\left(\frac{3-3.5}{\sqrt{35/480}} < Z < \frac{4-3.5}{\sqrt{35/480}}\right)$$

$$\doteq P(-1.8516 < Z < 1.8516) \doteq 0.9359$$

```
pnorm(4, 3.5, sqrt(35/480)) - pnorm(3, 3.5, sqrt(35/480))
[1] 0.9359225
```

25/28

25

6. 정규모집단이 아닌 경우의 표본분포

[따름정리 9-7] 표본비율의 분포

: 성공확률이 p인 n회의 시행에서 추출된 확률표본

$$\frac{X}{n} \overset{a}{\sim} N(p, \frac{p(1-p)}{n}) \qquad \frac{X/n-p}{\sqrt{p(1-p)/n}} = \frac{X-np}{\sqrt{np(1-p)}} \overset{a}{\sim} N(0,1)$$

[예 9-17]* 다음의 이항분포에서 구한 성공 횟수 표준화 통계량의 히스토그램과 정규확률도를 작성하고, 수렴결과 비교

$$B(n, 0.5) \quad B(n, 0.3) \quad B(n, 0.7) \quad B(n, 0.1)$$

26/28

26

6. 정규모집단이 아닌 경우의 표본분포

[예 9-18] 주사위를 랜덤하게 40번 굴렸을 때, 3이나 4가 나온 회수가 10 이상 15이하일 확률을 구하시오.

$$X \sim B(40, \frac{1}{3}) \qquad P(10 \le X \le 15) = \sum_{x=10}^{15} \binom{40}{x} (\frac{1}{3})^x (\frac{2}{3})^{40-x} \doteq 0.6723$$

```
pbinom(15, 40, 1/3) - pbinom(9, 40, 1/3)
[1] 0.6722687
```

근사적 계산
연속성 보정

$$np = 40 \times \frac{1}{3} = \frac{40}{3}, \; np(1-p) = 40 \times \frac{1}{3} \times \frac{2}{3} = \frac{80}{9}$$

$$P(10 \le X \le 15) \simeq P\left(\frac{9.5 - 40/3}{\sqrt{80/9}} < Z < \frac{15.5 - 40/3}{\sqrt{80/9}} \right)$$

$$\doteq P(-1.2857 < Z < 0.7267) \doteq 0.6670$$

```
pnorm(15.5, 40/3, sqrt(80/9)) - pnorm(9.5, 40/3, sqrt(80/9))
[1] 0.6670348
```

```
pnorm(15, 40/3, sqrt(80/9)) - pnorm(10, 40/3, sqrt(80/9))
[1] 0.5801487
```

27/28

27

6. 정규모집단이 아닌 경우의 표본분포

[정리 9-8] 표본평균의 분포 (모분산을 모르지만 표본이 매우 큰 경우)

$$\frac{\bar{X} - \mu}{S / \sqrt{n}} \overset{a}{\sim} N(0,1)$$

[예 9-19]* 다음의 분포에 대해 [정리 9-8]을 적용하여 비교

$$t(3) \quad Exp(5) \quad \Gamma(0.5, 10) \quad Poi(5)$$

28/28

28