Chapter03

통계분석 I

I 중심극한정리

FASTCAMPUS ONLINE

금융공학/퀀트 I

강사. 장순용

I 키포인트

- 중심극한정리 (CLT: Central Limit Theorem).
- 표준오차.

I 동전던지기 실험

- 동전을 **두** 번씩 던져서 평균을 구해본다. 즉, 크기 $n = 2$인 표본을 여러번 추출한다.

$$\overline{x_1}, \overline{x_2}, \overline{x_3}, \ldots$$

i	표본	$\overline{x_i}$
1	1,1	1
2	0,1	0.5
3	1,0	0.5
4	0,0	0
⋮	⋮	⋮

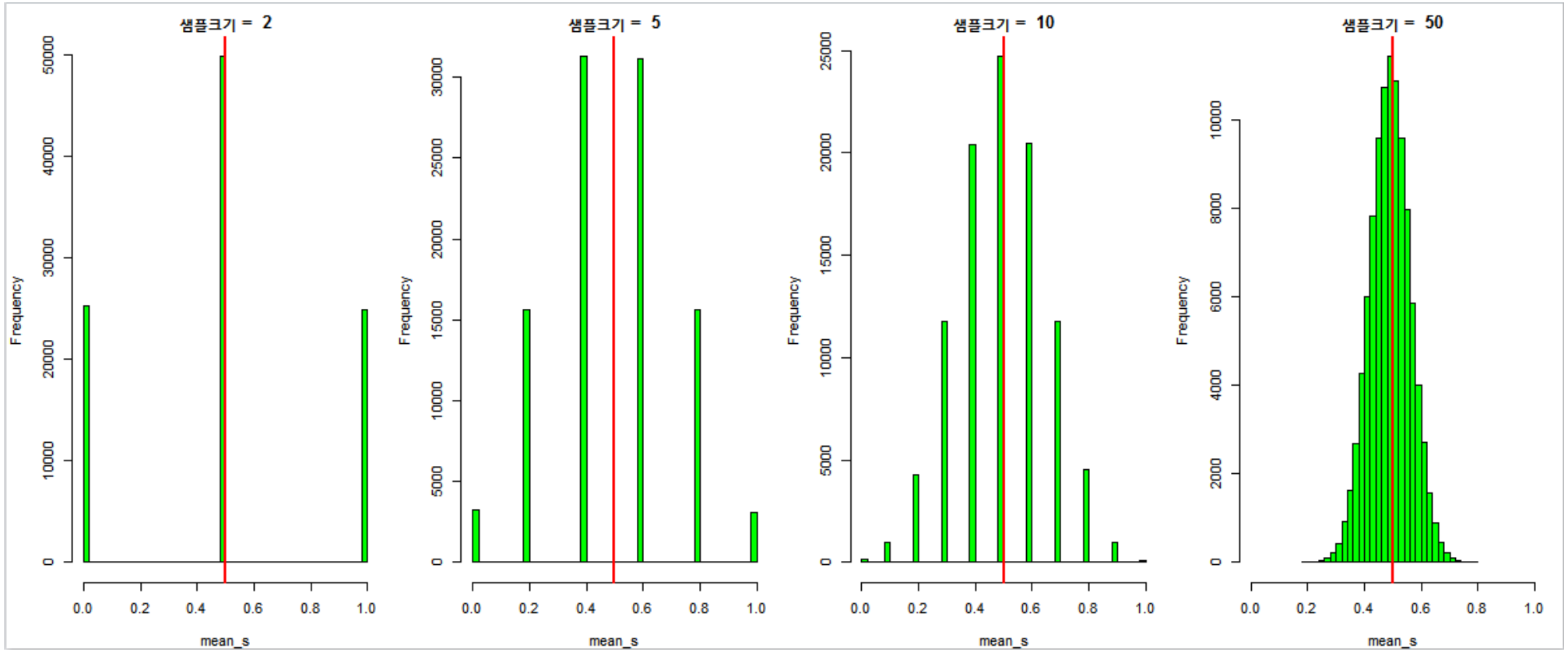
I 동전던지기 실험

- 동전을 **세** 번씩 던져서 평균을 구해본다. 즉, 크기 $n = 3$인 표본을 여러번 추출한다.

$$\overline{x_1}, \overline{x_2}, \overline{x_3}, \ldots$$

i	표본	$\overline{x_i}$
1	1,0,1	2/3
2	0,1,0	1/3
3	1,0,0	1/3
4	0,0,0	0
⋮	⋮	⋮

I 동전던지기 실험



표본평균의 히스토그램. 표본크기 n은 각각 2, 5, 10, 50이다.

I 동전던지기 실험

- 임의의 크기 n에 해당하는 표본평균 $\bar{x}$ 는 확률적으로 분포되어 있다.
- 그러므로 $\bar{X}$ (대문자)를 새로운 확률변수로 취급하여 이것의 평균과 분산을 계산한다.

  → 평균 : $E[\bar{X}] = \mu$

  → 분산 : $Var(\bar{X}) = \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} = \frac{p(1-p)}{n} = \frac{0.25}{n}$

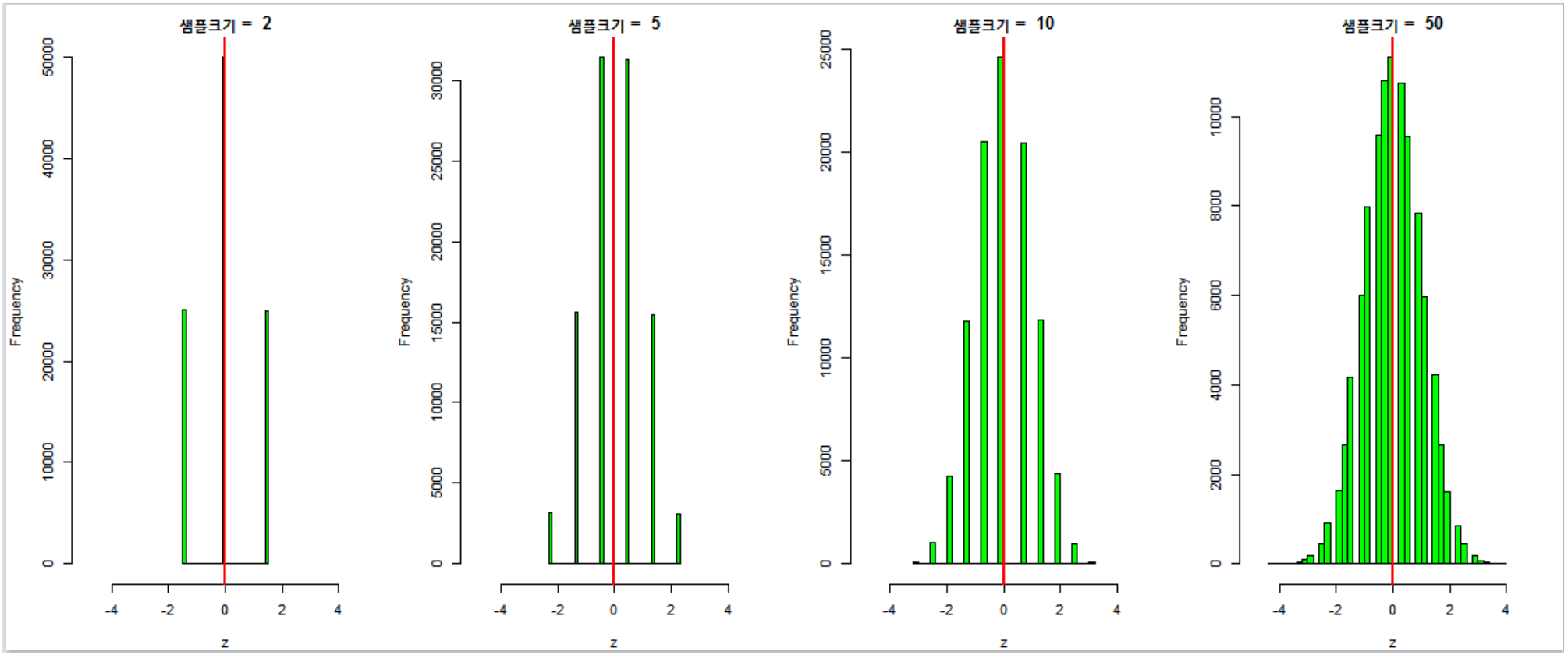
  → 표준편차 : $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \frac{0.5}{\sqrt{n}}$

- μ는 모평균이고 σ^2는 모분산임에 주의한다.
- 또한 $\sigma_{\bar{X}}^2$는 표본평균 $\bar{X}$의 분산이고 s^2는 단 하나의 표본 안의 분산이다.
- 표준편차 $\sigma_{\bar{X}}$는 모평균 μ를 추정할 때 발생하는 오차이며 표준오차라고 부른다.

I 동전던지기 실험

- 모집단의 확률분포는 베르누이의 특별 케이스 ($p = 0.5$) 이다.
- 그런데 $\bar{X}$의 확률분포는 근사적으로 정규분포인 것을 알수 있다. 특히 표본의 크기가 커질수록 폭이 좁아짐과 동시에 정규분포와 더욱 비슷해 진다.
- $Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$를 적용한 표준화로 일정한 구간을 유지시키면 시각화에 유리하다.
- 위에서 정의된 통계량은 표준정규분포를 따른다: $\frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \sim N(0,1)$

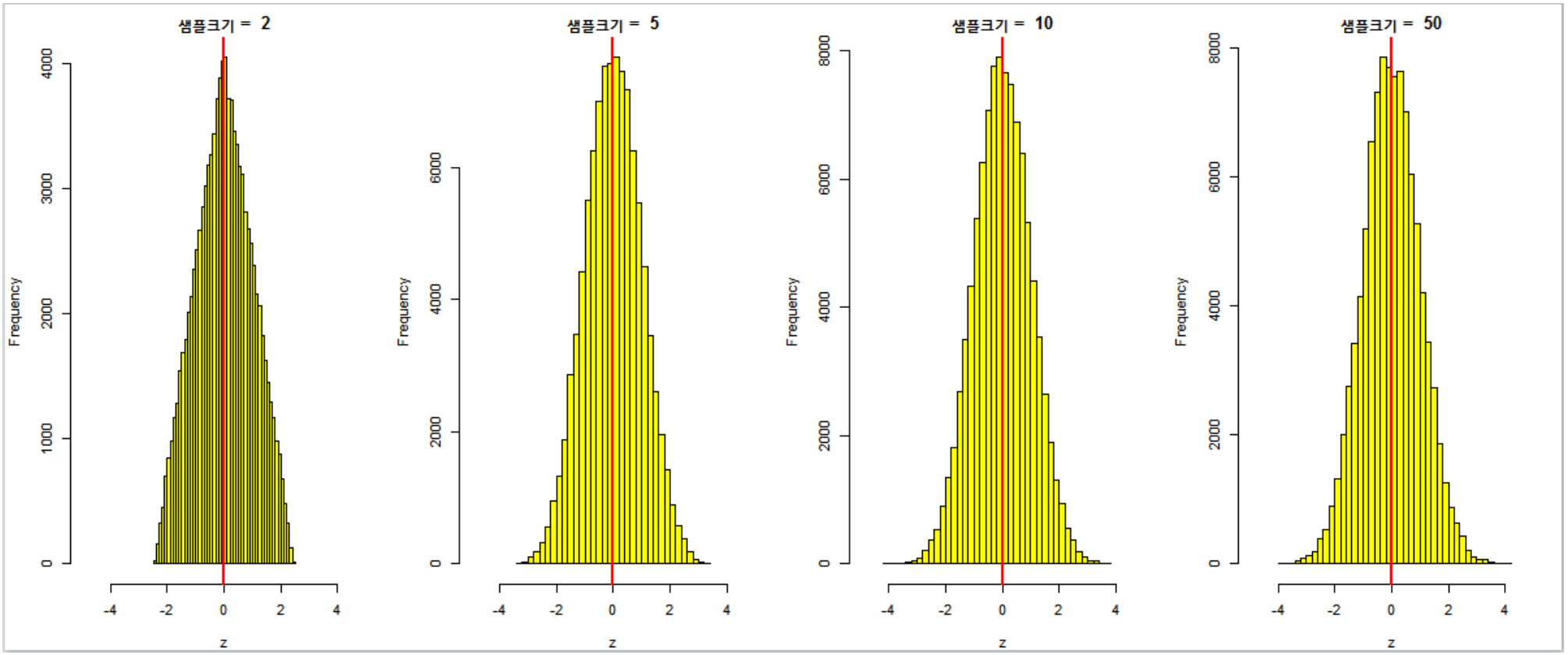
I 동전던지기 실험



표준화된 결과이다: $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$.

FAST CAMPUS
ONLINE
장순용 강사.

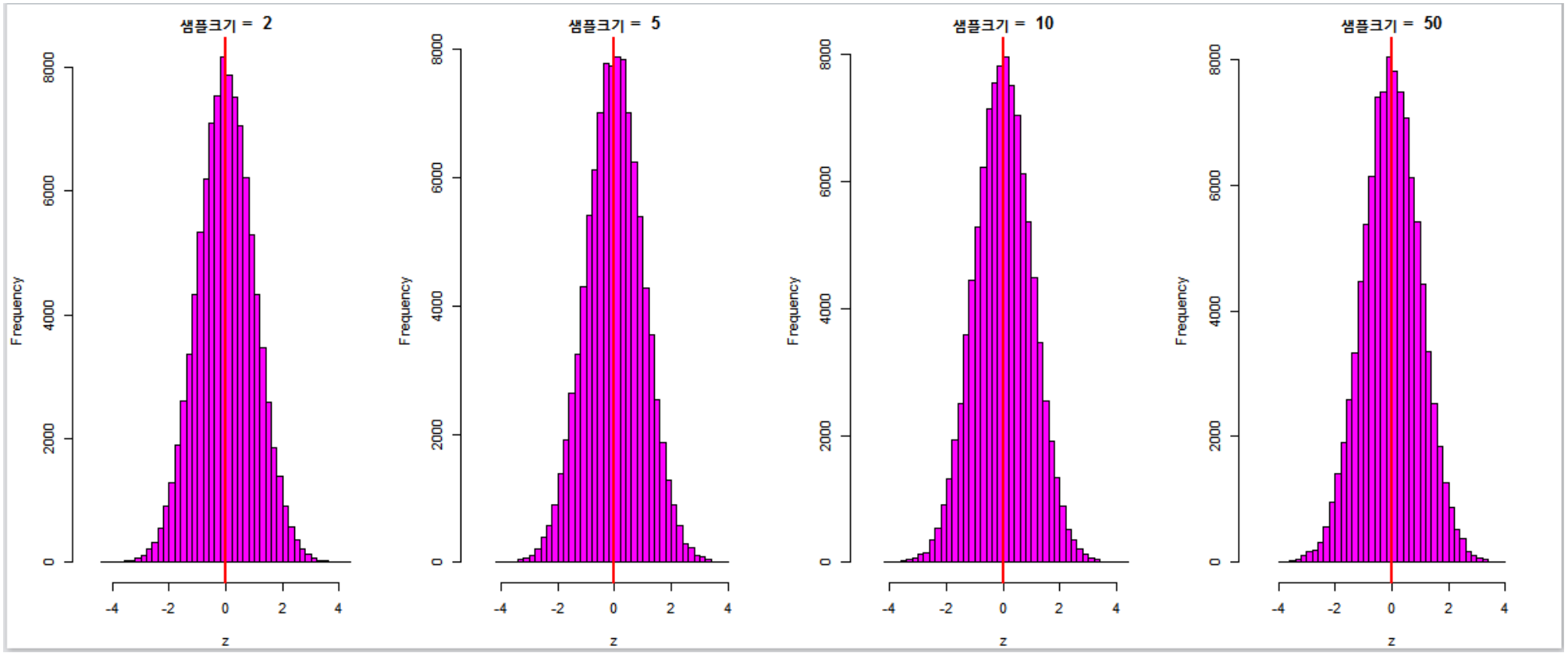
I 연속균등분포 실험



표준화된 결과이다: $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$.

FAST CAMPUS
ONLINE
장순용 강사.

I 정규분포 실험



표준화된 결과이다: $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$.

FAST CAMPUS
ONLINE
장순용 강사.

I 표본평균의 중심극한정리 결론

- 중심극한정리는 모집단의 확률분포와 무관하게 성립된다.
- 표본크기 n이 충분히 크다면 표본평균 $\bar{X}$의 분포는 근사적으로 정규분포이다.
- 보통 $n > 30$이면 중심극한정리가 성립된다고 인정함.
- 모집단의 확률분포가 정규분포이면 표본평균 $\bar{X}$의 분포는 정확하게 정규분포이다. 표본크기와 무관하게 성립된다.

  → 정규확률변수의 합은 또다른 정규확률변수이기 때문.

I 현실적 고려

- 현실에서는 표본은 단 한개이고 표본평균도 단 한개임.
- 하지만, CLT를 믿고 표본평균이 근사적으로 정규분포에 의해서 생성되었음을 전제한다.
- 그러면 정규분포의 특성을 응용하여 추정을 할 수 있다.

I 표준화

- 표본의 크기 n이 충분히 크다면 $\bar{X}$를 표준화할 수 있다.

  $\Rightarrow$ Z 통계량: 근사적으로 표준정규분포를 따른다.

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

- 만약에 모표준편차 σ를 모른다면, 대신해서 s를 사용한다.

  $\Rightarrow$ t 통계량: 자유도 $= n - 1$ 인 스튜던트 t 분포를 근사적으로 따른다.

$$t = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$

I 표본비율의 분포

- 동전은 베르누이 확률분포의 특별 케이스이다 ($p = 0.5$).
- 모집단이 일반적인 베르누이 확률분포를 따르는 경우를 전제해 본다.
- 성공확률이 p인 모집단을 전제하면 표본평균 $\bar{X}$의 기대값과 오차는 다음과 같다.

$\rightarrow$ 평균 : $E[\bar{X}] = p$

$\rightarrow$ 표준오차 : $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \sqrt{\frac{p(1-p)}{n}}$

I 표본비율의 분포

- 이 경우 $\bar{X}$를 **표본비율** 이라고 부르며 $\hat{P}$ 와 같이 표기한다:

  $\rightarrow \sigma_{\bar{X}}$ 을 $\sigma_{\hat{p}}$와 같이 표기하기로 한다.

- 보통 $np > 10$ and $n(1-p) > 10$이면 $\hat{P} \sim N(p, \frac{p(1-p)}{n})$으로 간주한다.

  $\Leftarrow$ 중심극한정리에 의함.

- 즉, 다음 통계량이 표준정규분포를 따른다:

$$\frac{\hat{P} - p}{\sigma_{\hat{P}}} \sim N(0,1)$$

I 통계량 사이의 차이 또는 합의 분포

- 두개의 모집단을 가정한다 (1과 2).
- 각각 모집단에서 크기가 n_1과 n_2인 표본을 추출한다.
- 각각의 표본평균 사이의 차이에는 다음과 같은 특성이 있다.

→ 평균 : $E[\overline{X_1} - \overline{X_2}] = E[\overline{X_1}] - E[\overline{X_2}] = \mu_1 - \mu_2$

→ 표준오차 : $\sigma_{\overline{X_1}-\overline{X_2}} = \sqrt{\sigma^2_{\overline{X_1}} + \sigma^2_{\overline{X_2}}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

I 통계량 사이의 차이 또는 합의 분포

- 다음 통계량은 근사적으로 표준정규분포를 따른다:

$$\frac{\overline{X_1} - \overline{X_2} - (\mu_1 - \mu_2)}{\sigma_{\overline{X_1} - \overline{X_2}}} \sim N(0,1)$$

- 표본평균의 합에도 유사한 규칙이 적용됨:

$$\frac{\overline{X_1} + \overline{X_2} - (\mu_1 + \mu_2)}{\sigma_{\overline{X_1} + \overline{X_2}}} \sim N(0,1)$$

- 이외의 통계량에도 유사한 규칙이 적용됨 (+ 또는 -).

I 통계량과 표준오차

통계량	표준오차	설명
평균	$\frac{\sigma}{\sqrt{n}}$	$n \geq 30$이면 표본평균은 근사적으로 정규분포를 따른다.
비율	$\sqrt{\frac{p(1-p)}{n}}$	보통 $np > 10$ and $n(1-p) > 10$이면 표본비율은 근사적으로 정규분포를 따른다.
중앙값	$\sigma\sqrt{\frac{\pi}{2n}}$	$n \geq 30$이면 표본중앙값은 근사적으로 정규분포를 따른다.
표준편차	a). $\frac{\sigma}{\sqrt{2n}}$ b). $\sqrt{\frac{\mu_4-\sigma^4}{4n\sigma^2}}$	a). 는 모집단이 정규분포를 따르는 경우, b)는 아닌 경우. $n \geq 30$이면 표본표준편차는 근사적으로 정규분포를 따른다.
분산	a). $\sigma^2\sqrt{\frac{2}{n}}$ b). $\sqrt{\frac{\mu_4-\sigma^2}{n}}$	a). 는 모집단이 정규분포를 따르는 경우, b)는 아닌 경우. 표본분산은 카이제곱 분포를 따른다.
상관계수	$\sqrt{\frac{1-r^2}{n-2}}$	r은 표본으로 계산한 상관계수. 근사적 정규분포.

I 끝.

감사합니다.

FAST CAMPUS
ONLINE
장순용 강사.