확률과 통계

섹션 - 3

강사 : James 쌤

기술통계

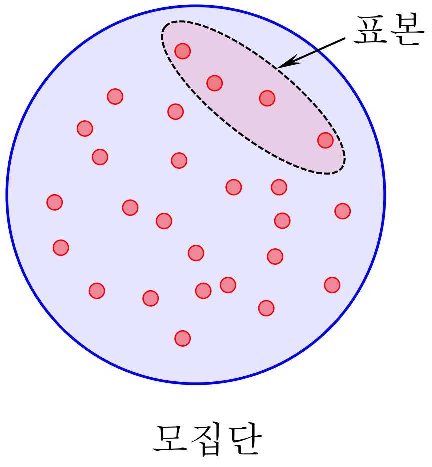
키포인트

- 모집단과 표본.
- 기술통계와 통계적 추론.
- 분위수.
- 상자그림 (Boxplot).

모집단과 표본

- 모집단 (population): 통계 분석 대상 전체. 실존 또는 개념적 존재.
- 표본 (sample): 모집단에서 추출한 일부. 데이터!

예). 대한민국 20세 이상 남성의 체질량지수 BMI 평균을 구하기 위해서 500명을 표본으로 뽑는다.



기술통계와 통계적 추론

- **기술통계**: 데이터의 통계적 특성을 있는 그대로 묘사한다.

  ⇒ 표본의 특성 즉 **통계량**을 계산한다.

- **통계적 추론**: 표본의 특성을 가지고 모집단의 특성 즉 **모수**를 알아낸다.

  ⇒ 일반화를 의미한다.

기술통계와 통계적 추론

- **기술통계**: 데이터의 통계적 특성을 있는 그대로 묘사한다.

  ⇒ 표본의 특성 즉 **통계량**을 계산한다.

- **통계적 추론**: 표본의 특성을 가지고 모집단의 특성 즉 **모수**를 알아낸다.

  ⇒ 일반화를 의미한다.

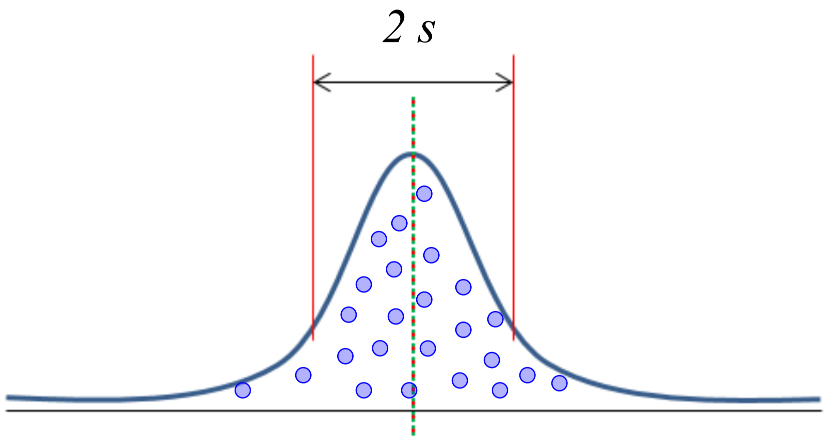
표본의 특성: 통계량

- 평균 (mean value): $\bar{x}$
- 중위수 (median): m
- 분산 (variance): s^2

  $\Rightarrow$ 표준편차 (standard deviation): $s = \sqrt{s^2}$
- 공분산 (covariance): s_{XY}

  $\Rightarrow$ 상관계수 (correlation): r
- 왜도 (skewness): ς

확률분포의 형상

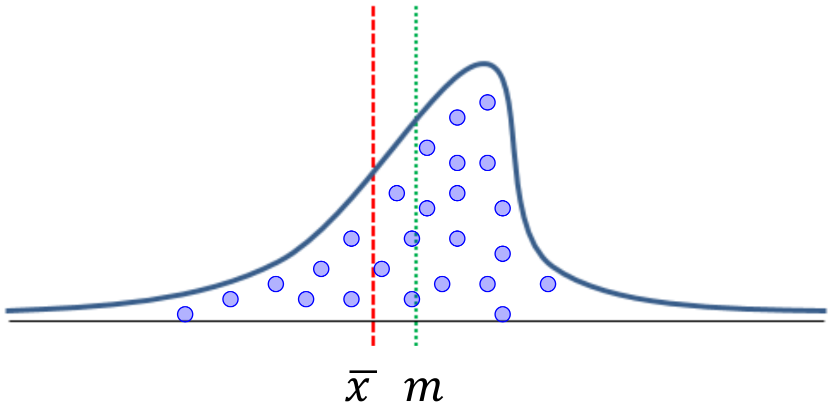


$\bar{x} = m$

$\zeta = 0$

좌우 대칭

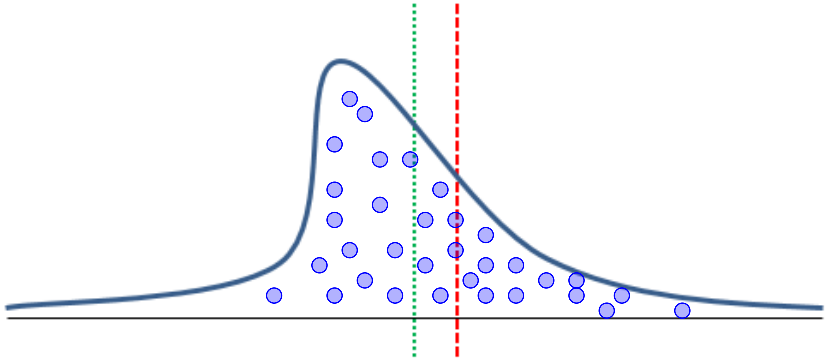
확률분포의 형상



$\zeta < 0$

왼 쪽으로 길게 뻗음

확률분포의 형상



$\zeta > 0$

오른 쪽으로 길게 뻗음

통계량 계산 수식

- 평균: $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
- 분산: $s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$
- 공분산: $s_{XY} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$
- 상관계수: $r = \frac{s_{XY}}{s_X s_Y}$이며 -1과 1 사이의 수치이다.

NOTE: 중위수를 계산하는 방법은 나중에 분위수와 관련하여 알아본다.

분위수

- 분위수 (quantile): 확률 α 에 해당하는 “α 분위수”는 누적확률이 α와 같은 지점을 일컫는다. (α 는 0과 1사이의 수치).
- **모집단**의 α 분위수는 확률분포와 누적확률을 안다면 다음과 같이 계산할 수 있다.

$$CDF(\alpha\ \text{분위수}) = \alpha$$

$$\alpha\ \text{분위수} = CDF^{-1}(\alpha)$$

분위수

- 분위수 (quantile): 확률 α 에 해당하는 "α 분위수"는 누적확률이 α와 같은 지점을 일컫는다. (α 는 0과 1사이의 수치).
- **모집단**의 α 분위수는 확률분포와 누적확률을 안다면 다음과 같이 계산할 수 있다.

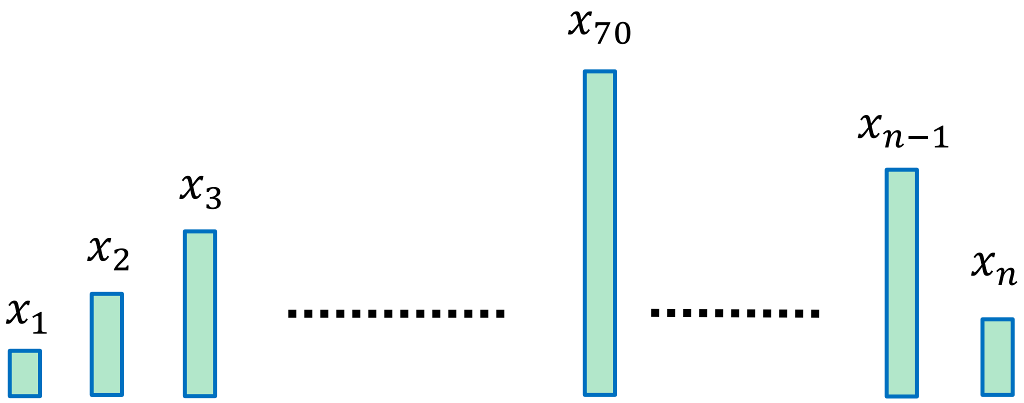
$$CDF(\alpha\text{ 분위수}) = \alpha$$

$$\alpha\text{ 분위수} = CDF^{-1}(\alpha)$$

*그럼, **표본**의 분위수는?*

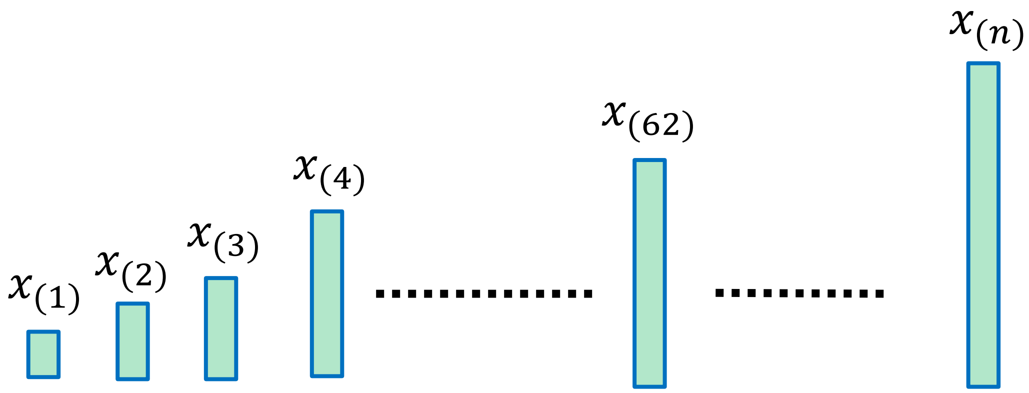
표본의 분위수

- $x_1, x_2, x_3, \cdots, x_n$ 와 값으로 이루어진 표본이 있다. 이 값들은 각양각색이다.



표본의 분위수

- 데이터를 **소→대** 순서대로 정렬한다.
- 정렬된 데이터를 $x_{(1)}, x_{(2)}, x_{(3)}, \cdots, x_{(n)}$와 같이 표기한다.



- 그러면, α 분위수는 $\alpha \times 100$ % 위치의 값이다. (α 는 0과 1사이의 수치).

백분위수, 사분위수

- 백분위수 (percentile): α 분위수와 같은데 α를 백분율 (0% ~ 100%)로 나타낸 경우.
- 사분위수 (quartile): α 를 4개의 구간으로 나눈 분위수.

$\Rightarrow$ 제1사분위수 (Q1) : $\alpha = 25\%$에 해당하는 백분위수.

$\Rightarrow$ 제2사분위수 (Q2) : $\alpha = 50\%$에 해당하는 백분위수.

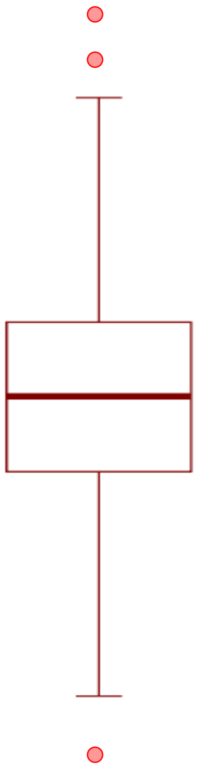
$\Rightarrow$ 제3사분위수 (Q3) : $\alpha = 75\%$에 해당하는 백분위수.

중위수, 최저값, 최고값

- 중위수 (median) = 50% 백분위수.
- 최고값 (maximum) = 100% 백분위수.
- 최저값 (minimum) = 0% 백분위수.

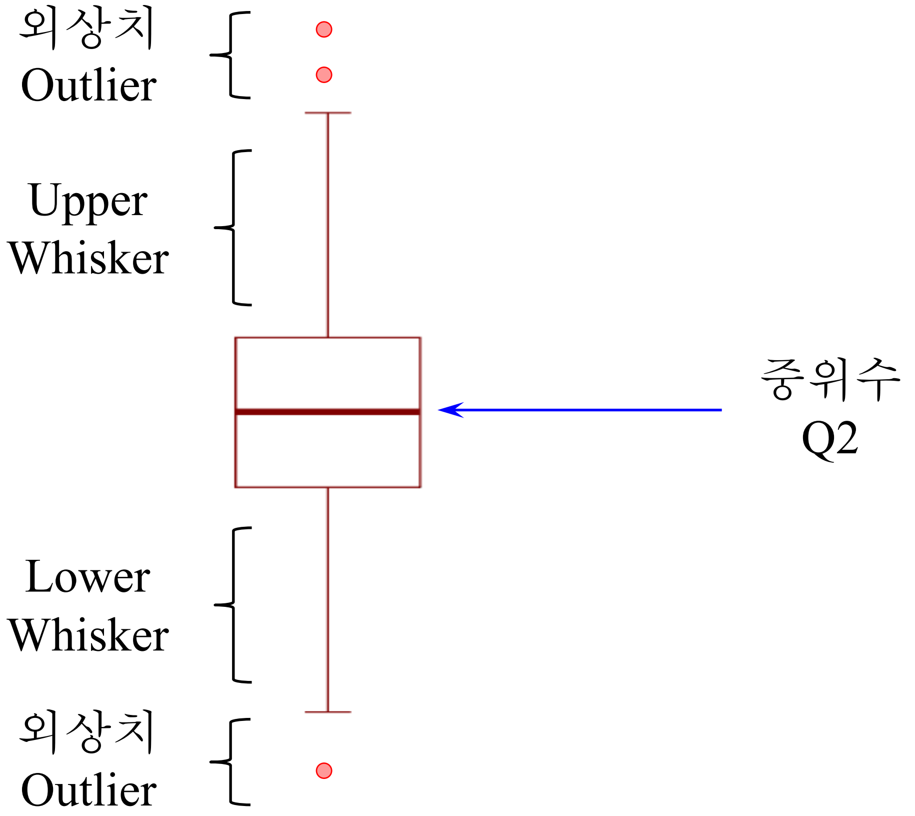
상자그림

- 상자그림 (Boxplot)은 표본의 분위수와 밀접한 시각화 유형이다.



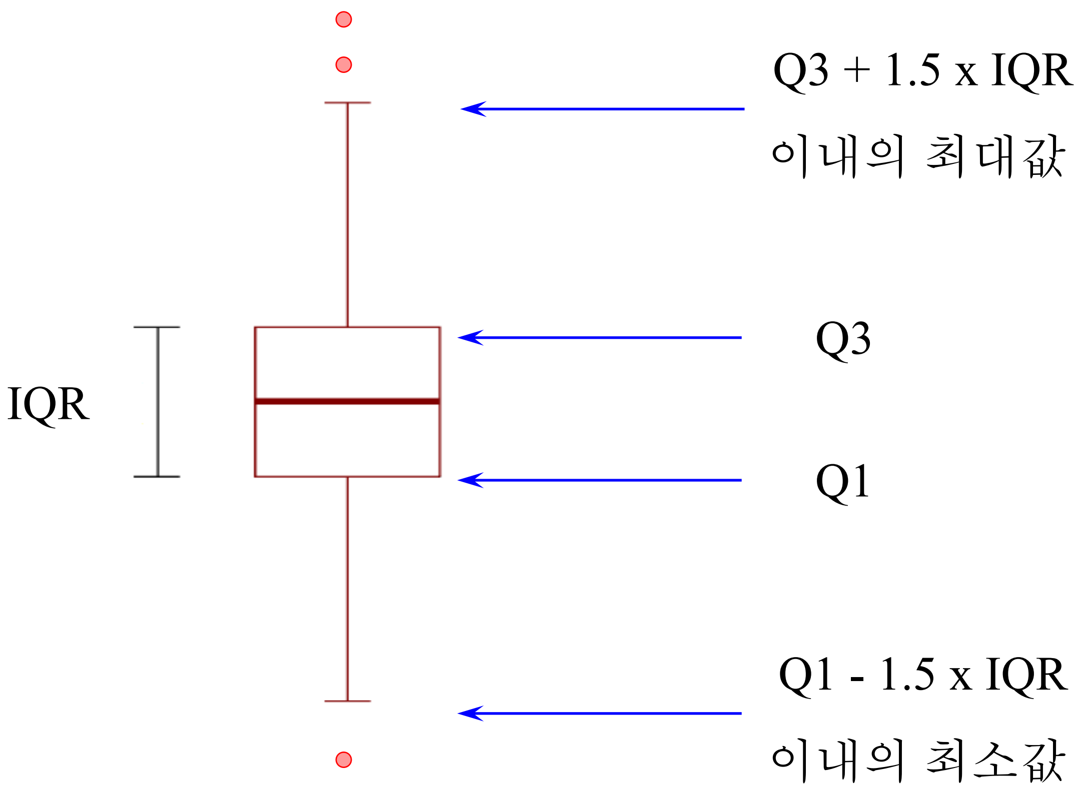
상자그림

- 상자그림 (Boxplot)은 표본의 분위수와 밀접한 시각화 유형이다.



상자그림

- 상자그림 (Boxplot)은 표본의 분위수와 밀접한 시각화 유형이다.



전수조사와 표본조사

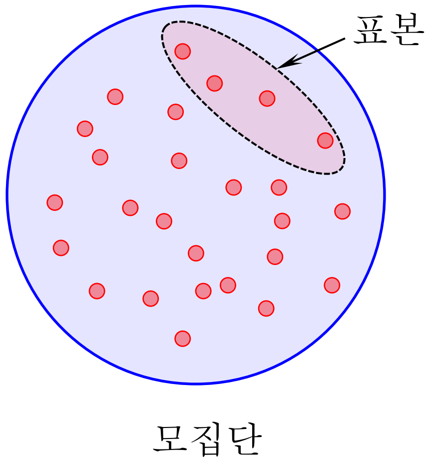
키포인트

- 전수조사와 표본조사.
- 통계량과 모수.
- 표본추출 방법.

모집단과 표본

- 모집단 (population): 통계 분석 대상 전체. 실존 또는 개념적 존재.
- 표본 (sample): 모집단에서 추출한 일부. 데이터!

예). 대한민국 20세 이상 남성의 체질량지수 BMI 평균을 구하기 위해서 500명을 표본으로 뽑는다.



기술통계와 통계적 추론

- **기술통계**: 데이터의 통계적 특성을 있는 그대로 묘사한다.

  ⇒ 표본의 특성 즉 **통계량**을 계산한다.

- **통계적 추론**: 표본의 특성을 가지고 모집단의 특성 즉 모수를 알아낸다.

  ⇒ 일반화를 의미한다.

통계적 추론의 당위성

- 모집단 대상 전수조사의 문제점:

  ⇒ 현실적으로 불가능할 수 있다.

  ⇒ 과다한 비용과 시간이 소요될 수 있다.

- 표본조사를 통해서 전수조사에 근접한 효과를 낼 수 있다.
- 표본조사는 실용적이고 비용면에서 이점이 있는 반면에 불확실성에 대한 고려가 필요하다.

  ⇒ 통계학 적용.

모수와 통계량

- 모수 (population parameter): 모집단을 사용하여 계산한 모집단의 특성.

  ⇒ 모집단은 단 하나. 모수도 단 한번 계산한다 (전수조사).

- 통계량 (sample statistic): 표본을 가지고 계산한 데이터의 특성.

  ⇒ 표본은 여러 번 추출할 수 있고 통계량도 여러 번 계산할 수도 있다. (실용적 아님).

  ⇒ 가급적이면 단 한번의 표본 추출로 통계량을 계산한다.

  ⇒ 궁극적으로는 모집단의 모수를 추정하기위한 목적으로 사용된다.

모수와 통계량

- 다음과 같이 모수와 통계량을 계산하는 방법을 요약해 본다.

	모수 $\boldsymbol{P(x)}$ 사용	모수 전수조사	통계량 데이터 사용
크기	N	N	n
평균	$\mu = \sum_{all\,x} x\,P(x)$	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
분산	$\sigma^2 = \sum_{all\,x} (x-\mu)^2\,P(x)$	$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i-\mu)^2}{N}$	$s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}$
표준편차	$\sigma = \sqrt{\sigma^2}$	$\sigma = \sqrt{\sigma^2}$	$s = \sqrt{s^2}$

표본추출

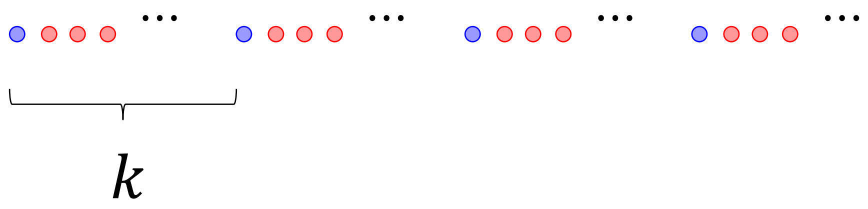
- 다음과 같은 방법들이 있다:
  - ✓ 단순임의추출 (simple random sampling).
  - ✓ 계통추출 (systematic sampling).
  - ✓ 가중치를 고려한 표본추출 (weighted random sampling).
  - ✓ 층화추출 (stratified sampling).
  - ✓ 집락추출 (cluster sampling).

단순 임의추출

- 모집단의 개개 값을 동일 확률로 추출하는 방법이다.
- 복원추출 (sampling with replacement)과 비복원추출 (sampling without replacement) 방법으로 세분화 할 수 있다.
- 복원추출을 통해서 무한대 크기의 모집단 효과를 낼 수 있다.

계통추출

- 모집단에서 임의의 위치에서 시작해 매 k번째 항목을 표본으로 추출하는 방법이다.
- 데이터가 정렬된 경우에는 단순 임의추출보다 좋은 방법이다.
- 데이터에 주기성이 있는 경우에는 부적절하다.



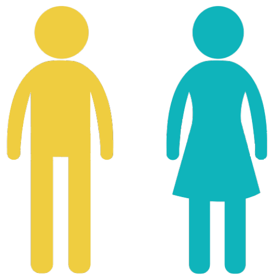
가중치를 고려한 표본추출

- 모집단의 개개 값에 가중치를 적용하여 동일하지 않은 확률로 추출하는 방법이다.

층화추출

- 계층의 비율을 고려한 표본 추출법.
- 데이터 값들이 중첩없이 분할될 수 있는 경우 적용한다 (교집합 없음).

예). 모집단에 남자 2000명, 여자 8000명이 있는 경우 남녀 분할해 놓고 각각 20명과와 80명을 표본으로 추출한다. 계층안에서는 단순임의추출.



집락추출

- 다단계 표본 추출 방법이다.

  ⇒ 모집단에서 군집을 일차적으로 추출한다 (1개 또는 다수).

  ⇒ 다음은 선정된 각 군집에서 일부 구성원 또는 전체를 표본으로 추출한다.

- 군집의 대표성을 고려한 표본추출 방법이다.

  예). A 고등학교 2학년이 모집단인 경우, 전체를 조사하지 않고 1반과 5반만을 표본으로 추출하는 경우.

중심극한정리

키포인트

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

동전 던지기 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
⋮	⋮	⋮

NOTE: 실용적인 상황은 아닙니다. 중심극한정리를 설명하기 위한 실험 입니다.

동전 던지기 실험

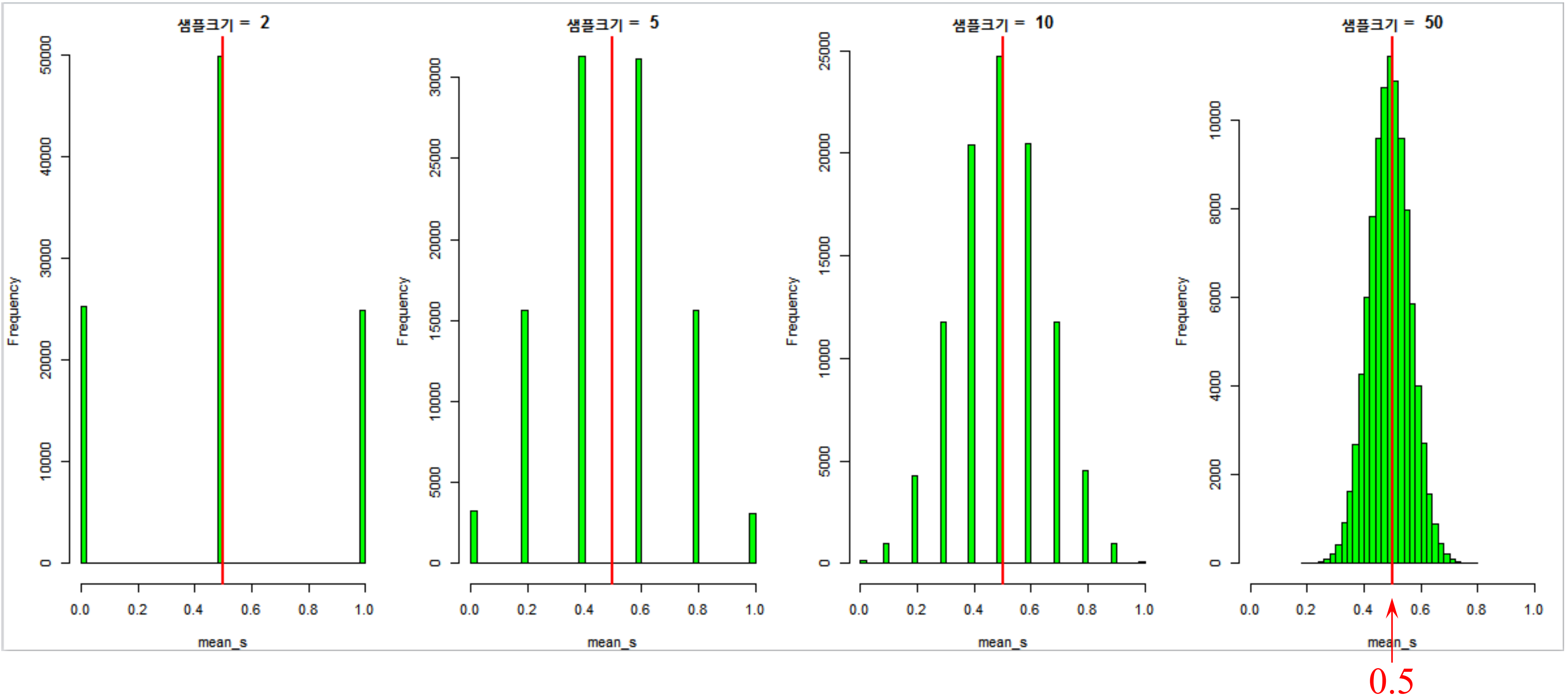
- 동전을 **세** 번씩 던져서 평균을 구해본다. 즉, 크기 $n = 3$인 표본을 여러번 추출한다.

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

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
⋮	⋮	⋮

NOTE: 실용적인 상황은 아닙니다. 중심극한정리를 설명하기 위한 실험 입니다.

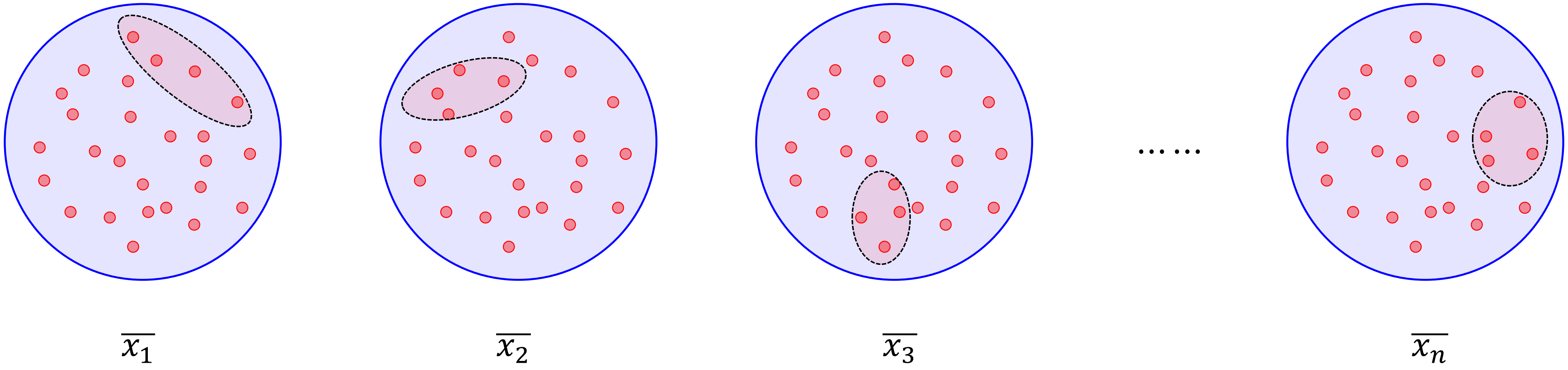
동전 던지기 실험



표본평균 $\overline{x_1}, \overline{x_2}, \overline{x_3}, \ldots$의 히스토그램. 표본크기 n은 각각 2, 5, 10, 50이다.

동전 던지기 실험

- 임의의 크기 n에 해당하는 표본평균 $\bar{x}$는 **확률적**으로 분포되어 있다.
- 그러므로, 표본평균 $\bar{X}$ (대문자)를 새로운 확률변수로 취급하여 이것의 평균과 분산을 계산해 보도록 한다.



동전 던지기 실험

- 표본평균 $\bar{X}$를 새로운 확률변수로 취급하여 이것의 평균과 분산을 계산해 본다.

  ⇒ 평균 : $E[\bar{X}] = \mu$. 그리고, 동전이기 때문에 $\mu = 0.5$.

  ⇒ 분산 : $Var(\bar{X}) = \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}$. 그리고, 동전이기 때문에 $\sigma_{\bar{X}}^2 = \frac{p(1-p)}{n} = \frac{0.25}{n}$

  ⇒ 표준편차 : $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$. 그리고, 동전이기 때문에$\sigma_{\bar{X}} = \frac{\sqrt{p(1-p)}}{\sqrt{n}} = \frac{0.5}{\sqrt{n}}$

- μ는 모평균이고 σ^2는 모분산임에 주의한다.

- 또한 $\sigma_{\bar{X}}^2$는 표본평균 $\bar{X}$의 분산이다. 참고로 s^2는 단 하나의 표본 안의 분산이다.

- 표준편차 $\sigma_{\bar{X}}$는 모평균 μ를 추정할 때 발생하는 오차이며 "표준오차"라 부른다.

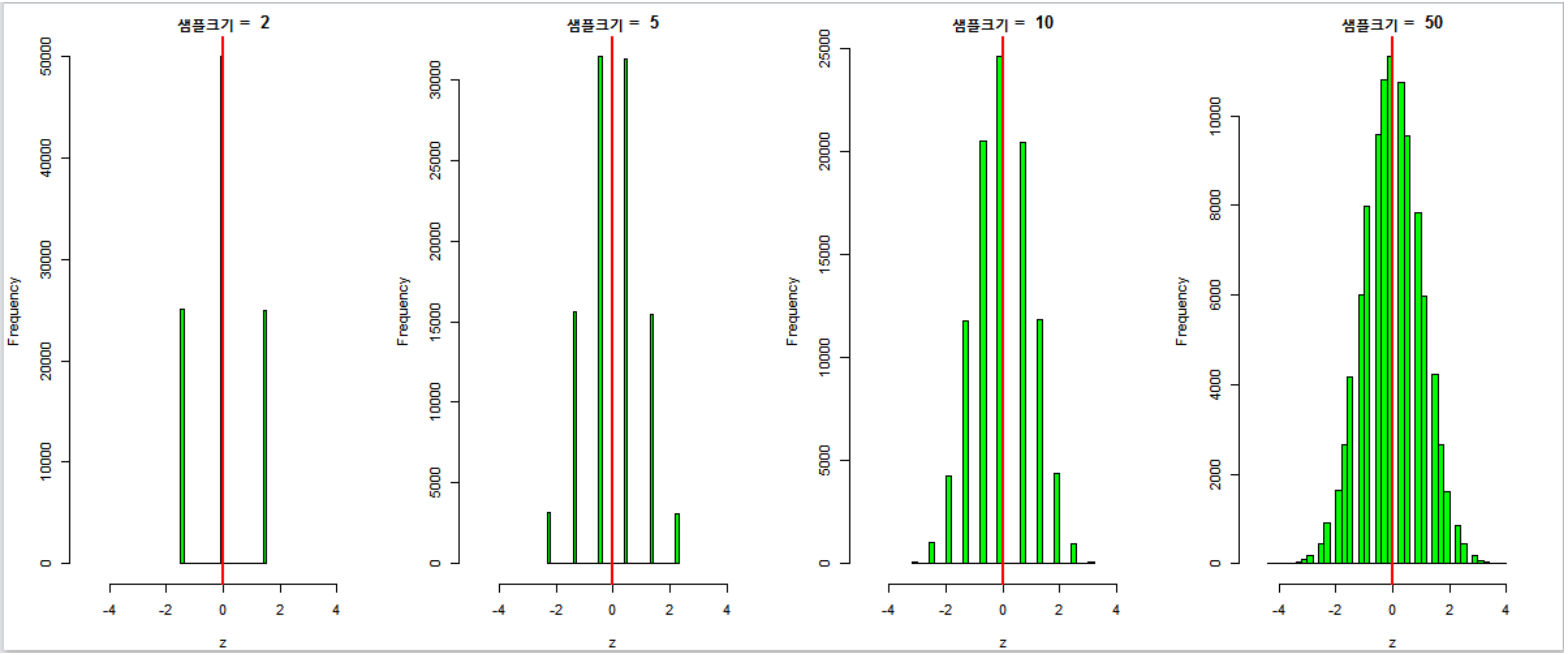
동전 던지기 실험

- 동전 모집단의 확률분포는 $p = 0.5$인 베르누이 확률분포의 특별 케이스이다.
- 그런데 $\bar{X}$의 확률분포는 근사적으로 정규분포인 것을 알 수 있다 (히스토그램).

  ⇒ 표본의 크기가 커질수록 너비는 좁아지면서 정규분포와 더욱 비슷해 진다.

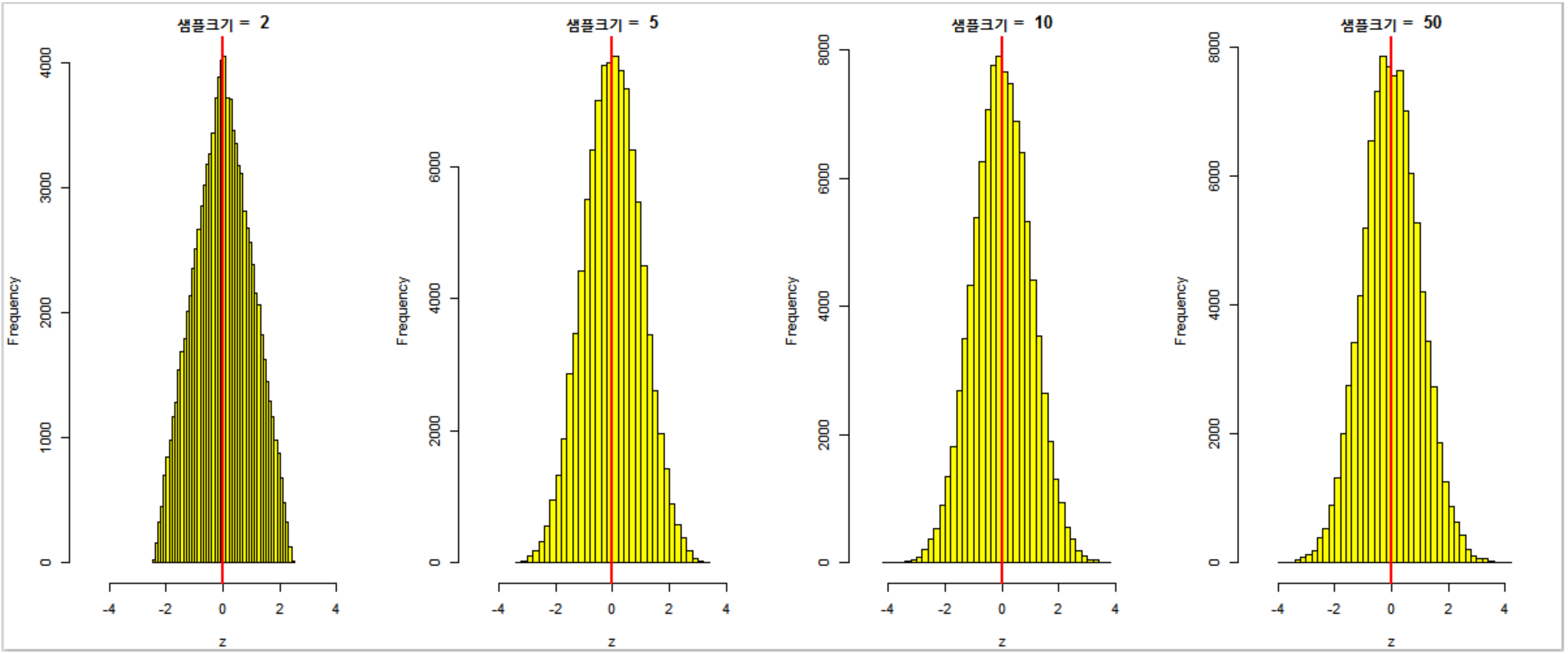
- $Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$를 적용한 표준화로 일정한 스케일을 유지시키면 시각화에 유리하다.
- 위의 "표준화된 통계량"은 표준정규분포를 따른다: $Z \sim N(0,1)$

동전 던지기 실험



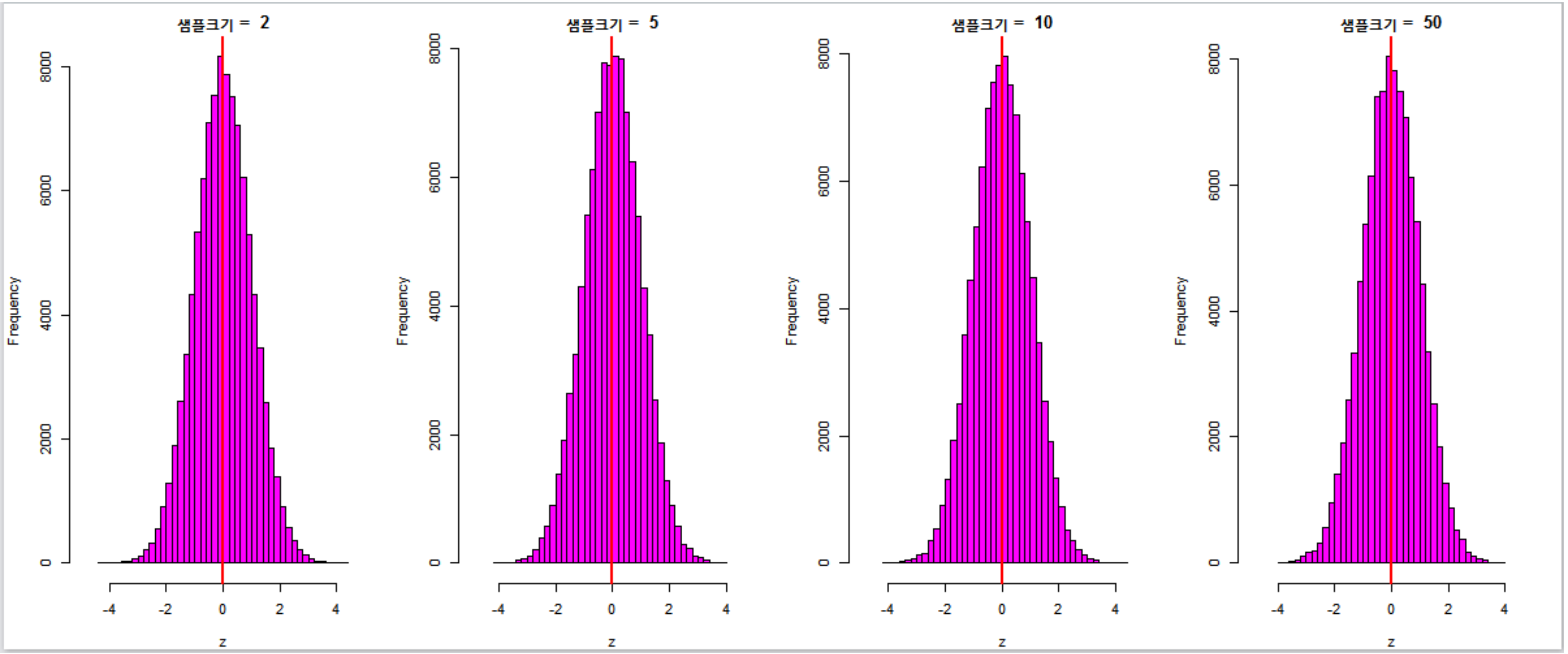
표준화된 통계량 Z의 확률분포를 보여주는 히스토그램.

연속균등분포 실험



표준화된 통계량 Z의 확률분포를 보여주는 히스토그램.

정규분포 실험



표준화된 통계량 Z의 확률분포를 보여주는 히스토그램.

표본평균의 중심극한정리 결론

- 중심극한정리는 모집단의 확률분포와는 무관하게 성립된다.
- 표본크기 n이 충분히 크다면 표본평균 $\bar{X}$의 분포는 근사적으로 정규분포이다.

  $\Rightarrow$ 보통 $n > 30$이면 중심극한정리가 성립된다고 인정한다.
- 모집단의 확률분포가 정규분포이면 표본평균 $\bar{X}$의 분포는 정확하게 정규분포이다.

  $\Rightarrow$ 이 경우에는 표본크기 n과는 무관하게 성립된다.

  $\Rightarrow$ 여러 정규확률변수의 합은 또다른 정규확률변수이기 때문이다.

현실적 고려

- 현실에서는 표본은 단 한 개이고 표본평균도 단 한 개이다. (표본크기 n은 임의)

  ⇒ 이전 실험과 같이 여러 개의 표본을 추출하는 방법은 현실적이지 않다.

  ⇒ 이전 실험은 중심극한정리를 설명하기 위해서 가정했을 뿐이다.

- 표본이 단 한 개인 현실적 상황에서는, 중심극한정리의 결과를 믿고 표본평균이 근사적으로 정규분포를 따른다고 전제한다.

- 그러면 정규분포의 특성을 활용하여 추정을 할 수 있게된다.

표준화

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

$\Rightarrow$ Z 통계량: 표준정규 확률분포를 근사적으로 따른다.

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

- 만약에 **모**표준편차 σ를 모른다면, 대신해서 **표본**표준편차 s를 사용한다.

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

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

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

표본비율의 분포

- 이 경우 $\bar{X}$를 **표본비율** 확률변수이고 $\hat{P}$ 와 같이 고쳐서 표기하도록 한다:

  ⇒ 또한 $\sigma_{\bar{X}}$ 을 $\sigma_{\hat{P}}$와 같이 고쳐서 표기하기로 한다.

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

  ⇒ 중심극한정리에 의함.

- 즉, 다음과 같이 정의된 통계량은 표준정규분포를 따르게 된다:

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

통계량 사이의 차이/합의 분포

- "1" 과 "2"로 칭하는 두 개의 모집단을 가정해 본다.
- 각각 모집단에서 크기가 n_1과 n_2인 표본을 추출한다.
- 표본평균 사이의 차이 $\overline{X_1} - \overline{X_2}$ 에 대해서 평균과 표준오차를 계산할 수 있다.

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

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

통계량 사이의 차이/합의 분포

- 다음 표본평균의 차이로 만든 통계량은 근사적으로 표준정규분포를 따른다:

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

- 또한 표본평균의 합으로 만든 통계량도 근사적으로 표준정규분포를 따른다:

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

통계량과 표준오차

통계량	표준오차	설명
평균	$\frac{\sigma}{\sqrt{n}}$	$n \geq 30$이면 표본평균은 근사적으로 정규분포를 따른다.
비율	$\sqrt{\frac{p(1-p)}{n}}$	대략 $n\,p > 10$ 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 경우이며 b)는 그렇지 아닌 경우에 해당한다. $n \geq 30$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 경우이며 b)는 그렇지 아닌 경우에 해당한다. 표본분산은 근사적으로 정규분포를 따른다.
상관계수	$\sqrt{\frac{1-r^2}{n-2}}$	r은 표본으로 계산한 상관계수를 나타낸다. 이것 또한 근사적으로 정규분포를 따른다.

점추정

키포인트

- 통계적 추정 방법.
- 점추정.
- 추정량의 조건.

통계적 추정 방법

- 다음 예를 살펴보자:

  예). 20세 이상 성인의 1일 평균 수면시간을 파악하기 위해서 표본을 대상으로 조사하였다. 다음과 같은 결과를 생각해 볼 수 있다.

  a) 8.0시간.

  b) 6.5 시간 ~ 8.5 시간.

  c) 7.2 시간 ~ 8.9 시간.

통계적 추정 방법

- 질문은 한가지 였는데, 여러가지 방식의 답변이 나왔다.
- 한개의 값을 제시하는 것을 점추정 (point estimation)이라고 한다.

  ⇒ "추정량"을 사용해서 계산한다. 추정량 ≈ 계산 방법 또는 수식.

- 구간을 제시하는 것을 구간추정 (interval estimation)이라고 한다.

  ⇒ 그런데, 구간이 여러 방식으로 제시되는 것을 알 수 있다.

통계적 추정 방법

- 질문은 한가지 였는데, 여러가지 방식의 답변이 나왔다.
- 한개의 값을 제시하는 것을 점추정 (point estimation)이라고 한다.

  ⇒ "추정량"을 사용해서 계산한다. 추정량 ≈ 계산 방법 또는 수식.
- 구간을 제시하는 것을 구간추정 (interval estimation)이라고 한다.

  ⇒ 그런데, 구간이 여러 방식으로 제시되는 것을 알 수 있다.

추정량의 조건

- 좋은 추정방식, 즉 "추정량"이 되기 위해서는 다음 조건들이 충족되어야 한다:

a) 불편성 (unbiasedness).

b) 효율성 (efficiency).

c) 일치성 (consistency).

추정량의 불편성

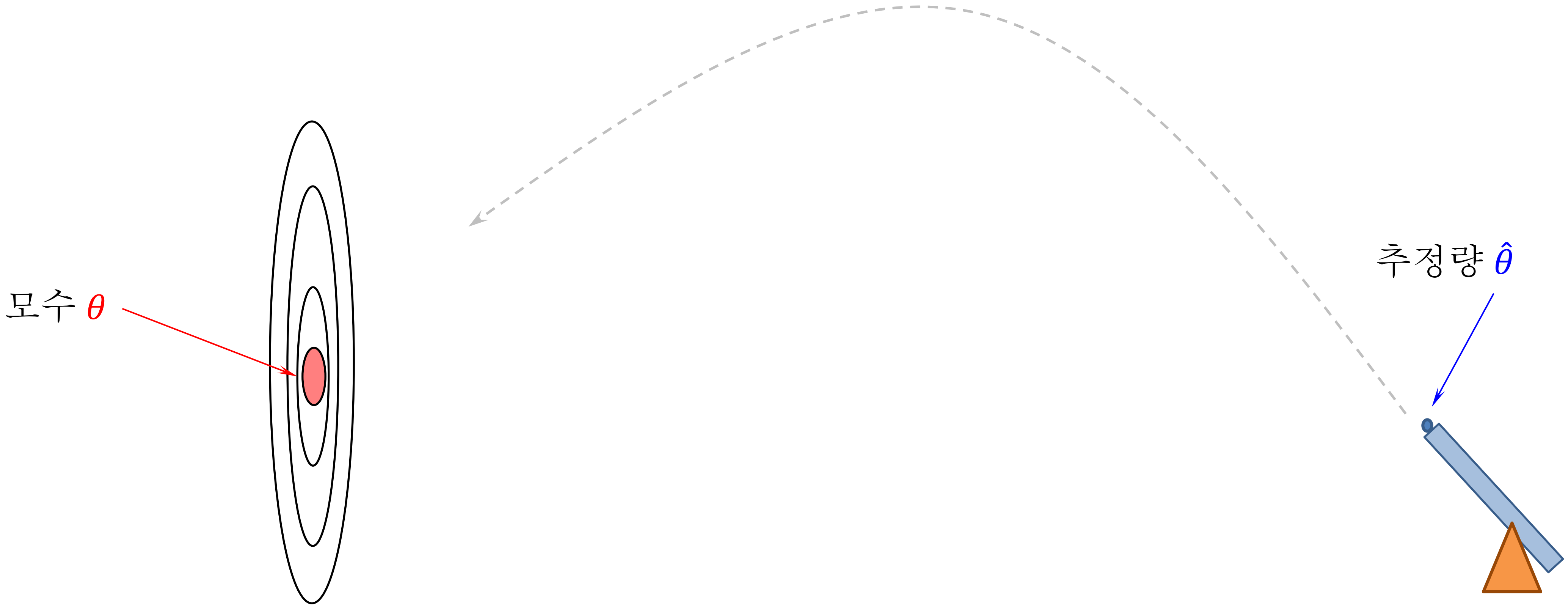
- 먼저 θ가 추정 대상인 모수이고 $\hat{\theta}$가 해당 추정량이라고 정의 한다.
- 추정량 $\hat{\theta}$가 다음 조건을 충족한다면 "불편 추정량"이라 부른다.

$$E\left[\hat{\theta}\right] = \theta$$

$\Rightarrow \frac{\sum_{i=1}^{n} x_i}{n}$는 μ의 "불편 추정량"이다.

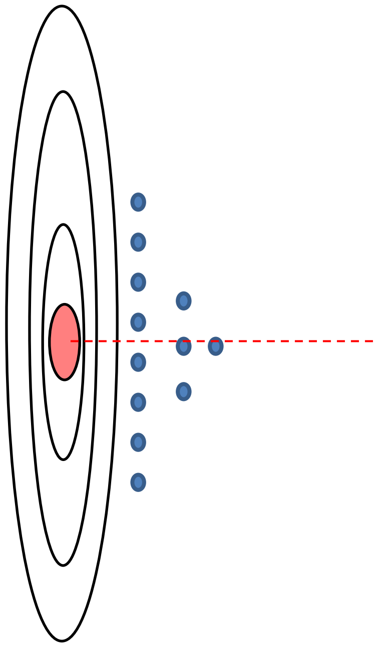
$\Rightarrow$ 하지만 σ^2의 불편 추정량은 $\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\boldsymbol{n}}$가 아니라 $\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\boldsymbol{n-1}}$임에 주의한다.

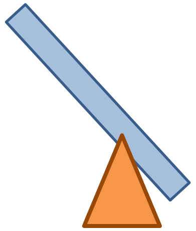
추정량의 불편성



추정량의 불편성

$E[\hat{\theta}] = \theta$ 충족!





추정량의 효율성

- 불편성 조건을 충족하는 두 개의 추정량이 있다고 가정한다:

$$E[\hat{\theta}_1] = \theta$$

$$E[\hat{\theta}_2] = \theta$$

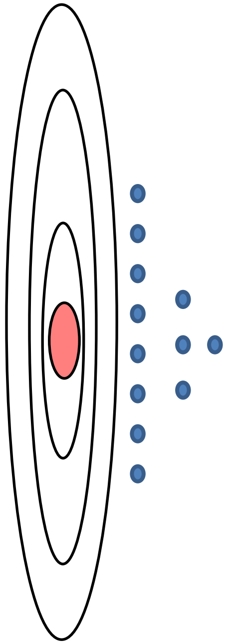
- 이 둘 중에서 불확실성이 적은 추정량을 "효율적" 추정량 이라고 부른다.

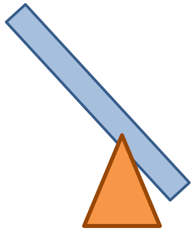
$\Rightarrow$ 즉, $Var(\hat{\theta}_1) > Var(\hat{\theta}_2)$라면 $\hat{\theta}_2$가 효율적 추정량이 된다.

추정량의 효율성

$E[\hat{\theta}_1] = \theta$ 충족

$Var(\hat{\theta}_1)$ 상대적으로 크다

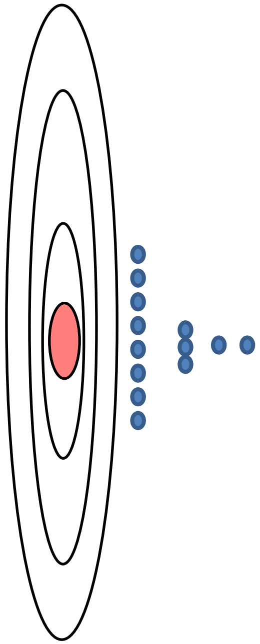


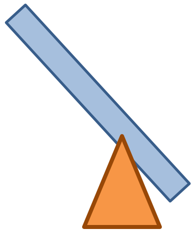


추정량의 효율성

$E[\hat{\theta}_2] = \theta$ 충족

$Var(\hat{\theta}_2)$ 상대적으로 작다





추정량의 일치성

- 표본크기가 커짐에 따라서 불편성이 잘 충족되는 방향으로 이동하는 특성.
- 작은 표본의 경우에는 불편성을 충족하지 못했던 추정량이 일치성에 의해서 큰 표본의 경우에 불편성을 충족할 수도 있다.

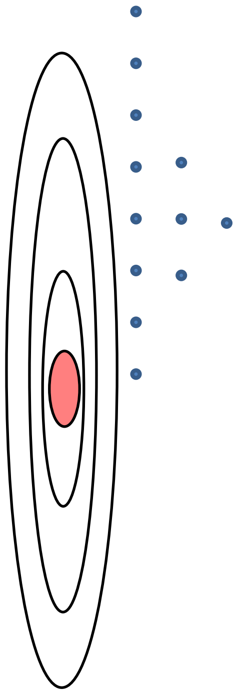
예). $\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\boxed{\boldsymbol{n}}}$ ← $\boxed{n-1\ \text{아님!}}$

- 일치성은 표본크기를 임의로 키울 수 있을 때에 유용한 기준이다.

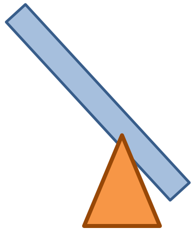
추정량의 일치성

$E[\hat{\theta}] > \theta$

불편성 충족 X



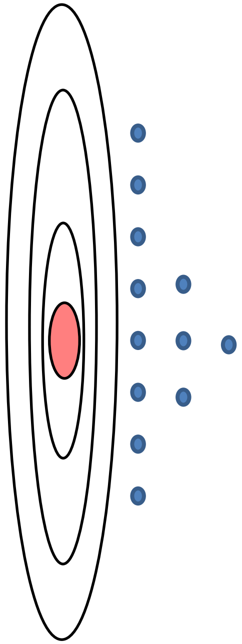
n *small*



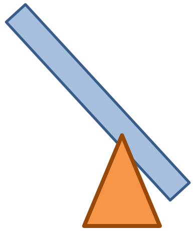
추정량의 일치성

$$E[\hat{\theta}] \cong \theta$$

불편성 충족 O



$n' > n$



구간추정

키포인트

- 구간추정의 원리.
- 신뢰구간.

모평균의 구간추정

- 통계량을 바탕으로 모평균의 신뢰구간 (confidence interval)을 계산하고자 한다.
- 신뢰구간: 표본평균의 확률분포에 모평균이 신뢰수준 확률로 포함되는 구간.
- **중심극한정리**에 의하면 표본평균 $\bar{X}$는 근사적으로 정규분포를 따르고 표준화된 Z는

  표준정규분포를 따른다: $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$.
- 그러면 다음을 정의한다:

  ⇒ 신뢰수준 확률: $(1-\alpha)$.

  ⇒ 오차율: α.

모평균의 구간추정

- 모평균 μ의 95% 신뢰구간을 만들어 본다. 이때, 모표준편차 σ를 안다는 전제를 한다.

$$P(-1.96 < Z < 1.96) = 0.95$$

$$\Downarrow$$

$$P\left(-1.96 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < 1.96\right) = 0.95$$

$$\Downarrow$$

$$P\left(-1.96\,\sigma/\sqrt{n} < \bar{X}-\mu < 1.96\,\sigma/\sqrt{n}\right) = 0.95$$

모평균의 구간추정

$$P\left(-1.96\,{}^{\sigma}\!/_{\sqrt{n}} < \bar{X} - \mu < 1.96\,{}^{\sigma}\!/_{\sqrt{n}}\right) = 0.95$$

$$\Downarrow$$

$$P\left(-\bar{X} - 1.96\,{}^{\sigma}\!/_{\sqrt{n}} < -\mu < -\bar{X} + 1.96\,{}^{\sigma}\!/_{\sqrt{n}}\right) = 0.95$$

$$\Downarrow$$

$$P\left(\bar{X} + 1.96\,{}^{\sigma}\!/_{\sqrt{n}} \geq \mu \geq \bar{X} - 1.96\,{}^{\sigma}\!/_{\sqrt{n}}\right) = 0.95$$

$$\Downarrow$$

$$\boxed{P\left(\bar{X} - 1.96\,{}^{\sigma}\!/_{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\,{}^{\sigma}\!/_{\sqrt{n}}\right) = 0.95}$$

모평균의 구간추정

- 그러면 모평균 μ의 95% 신뢰구간은 다음 상한과 하한으로 구성되어 있다.

하한: $\bar{X} - 1.96\, {}^{\sigma}/_{\sqrt{n}}$ 상한: $\bar{X} + 1.96\, {}^{\sigma}/_{\sqrt{n}}$

[$\leftarrow$ <신뢰구간> $\rightarrow$]

- 1.96이라는 수치는 어디에서 나온 것인가?

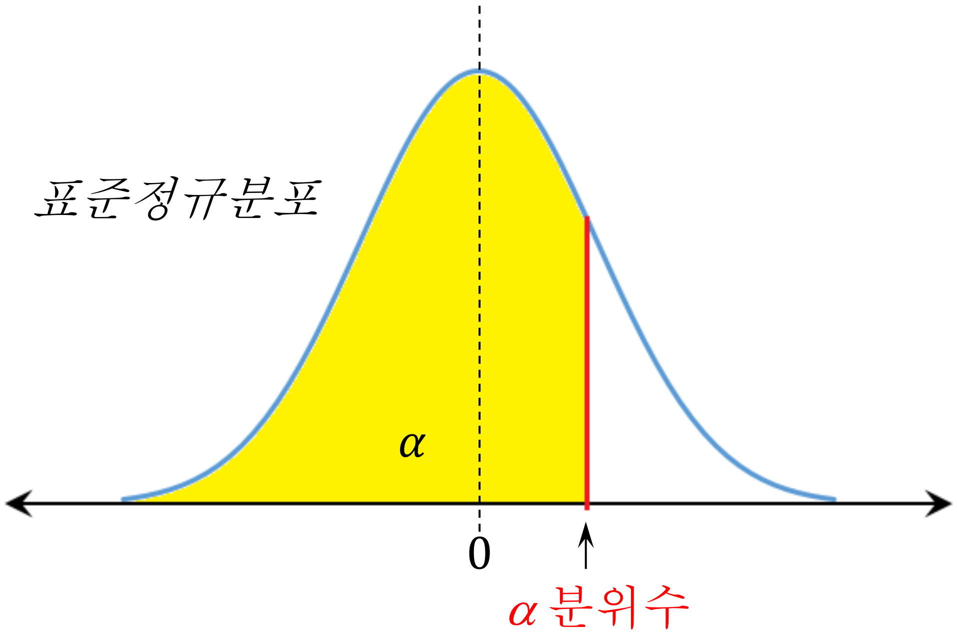
$\Rightarrow z_{0.025}$에 해당하는 수치이다.

$\Rightarrow z_{0.025}$는 표준정규확률분포에서 누적확률(CDF)가 0.975에 해당하는 위치이다.

표준정규분포의 분위수

- "α 분위수"는 누적확률 (CDF)가 확률 α와 같은 지점을 일컫는다.

$$P(Z < \alpha\ 분위수) = \alpha$$

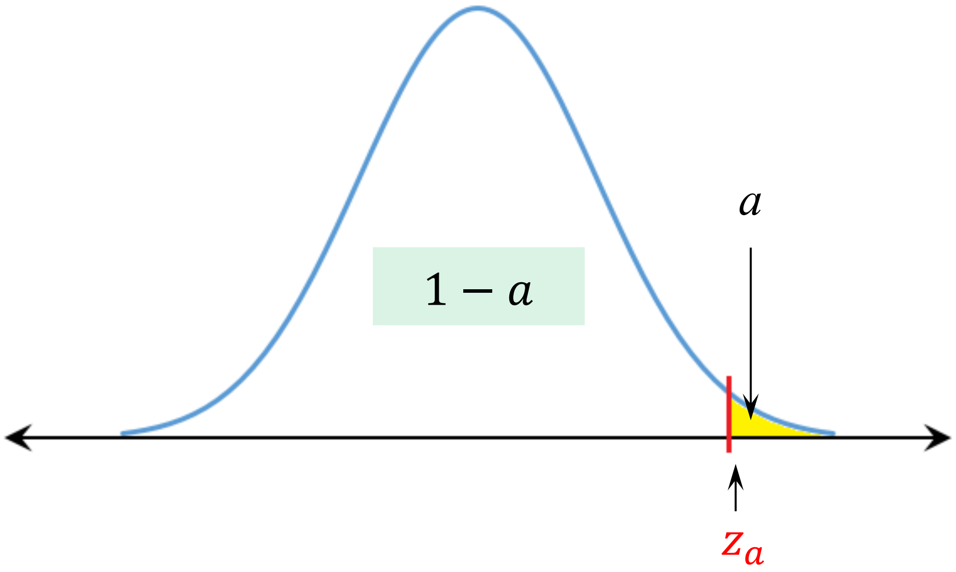


표준정규분포의 분위수

- z_a는 우측 꼬리의 확률이 a와 같은 위치를 의미한다.

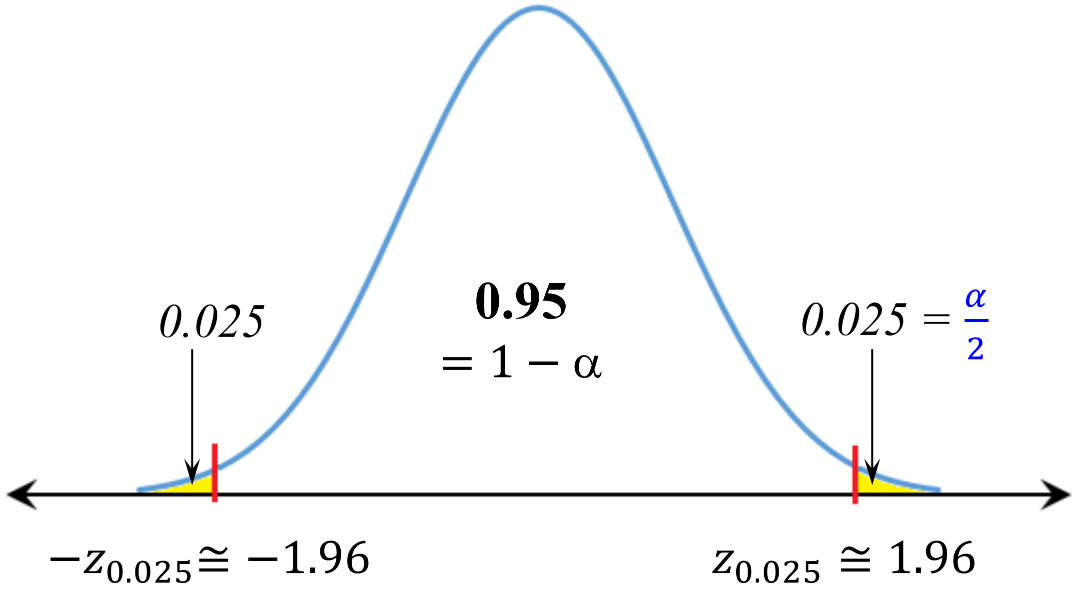
$$P(z_a < Z) = a$$

- 그러므로 $z_a = (1 - a)$분위수와 같다.



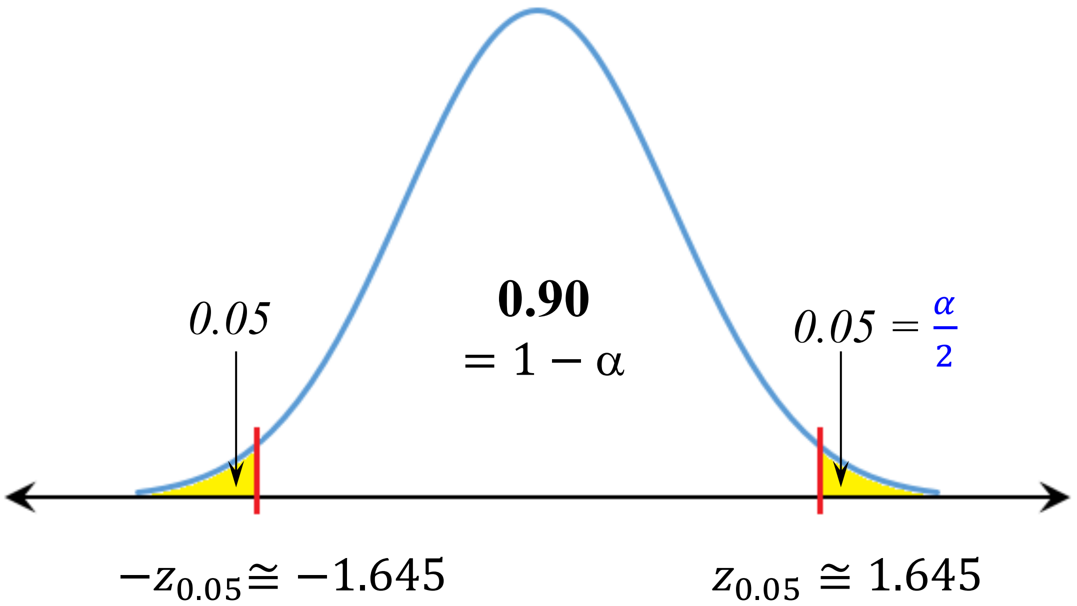
표준정규분포의 분위수

- 아래에서 오차율 $= \alpha = 0.05$이다.
- 그러므로 신뢰수준 $= (1 - \alpha) = 0.95$이다.



표준정규분포의 분위수

- 아래에서 오차율 $= \alpha = 0.10$이다.
- 그러므로 신뢰수준 $= (1 - \alpha) = 0.90$이다.



모평균의 구간추정

- 다음과 같이 임의의 신뢰수준 확률 $1-\alpha$에 해당하는 모평균의 신뢰구간을 만들어서 일반화 할 수 있다. 이전과 마찬가지로, 모표준편차 σ를 안다는 전제는 유지한다.

하한: $\bar{X} - z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$ 상한: $\bar{X} + z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$

[← <신뢰구간> →]

모평균의 구간추정

Question: 신뢰수준 확률은 무조건 높아야 좋은 것 아닌가?

모평균의 구간추정

(　　　)

신뢰수준 99% 신뢰구간.

(　　)

신뢰수준 95% 신뢰구간.

(　)

신뢰수준 90% 신뢰구간.

모평균의 구간추정

(⟵⟶)

99.9% 신뢰구간 : 성인 남성의 신장은 0m ~ 3m 사이이다.

(⟵⟶)

95% 신뢰구간 : 성인 남성의 신장은 1.60m ~ 1.90m 사이이다.

(⟵⟶)

90% 신뢰구간 : 성인 남성의 신장은 1.70m ~ 1.80m 사이이다.

모평균의 구간추정

Answer: 신뢰수준이 높으면서 신뢰구간이 좁아야 좋다.

모평균의 구간추정

- 신뢰구간의 상한과 하한은 다음과 같이 계산하였다: $\bar{X} \pm z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}}$.
- 오차율 α가 클수록 신뢰구간은 좁다 (컨트롤 가능하지만 그대로 둔다).
- 표준편차 σ가 작을수록 신뢰구간은 좁다 (컨트롤 불가능).
- 표본크기 n이 클수록 신뢰구간은 좁다 (컨트롤 가능).

표본크기를 키우면 오차율을 키우지 않고 (신뢰수준 유지) 신뢰구간을 좁힐 수 있다!

모평균의 구간추정

- W가 목표하는 신뢰구간의 폭이라고 한다면: $\bar{X} \pm W$
- 다음 관계를 사용해서 표본크기를 정한다.

$$\bar{X} \pm z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} = \bar{X} \pm W \quad \Rightarrow \quad z_{\frac{\alpha}{2}} \times \frac{\sigma}{\sqrt{n}} = W \quad \Rightarrow \quad \boldsymbol{n} = \left[\frac{\boldsymbol{z_{\alpha/2} \times \sigma}}{\boldsymbol{W}}\right]^{\boldsymbol{2}}$$

모평균의 구간추정

- 그런데, 이제는 모표준편차 σ를 모르는 경우를 전제하고 임의의 신뢰수준 확률 $1-\alpha$에 해당하는 모평균의 신뢰구간을 만들어본다.

하한: $\bar{X} - t_{\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}$ 상한: $\bar{X} + t_{\frac{\alpha}{2}} \times \frac{s}{\sqrt{n}}$

[← <신뢰구간> →]

- 모표준편차를 아는 경우와 비교해서 바뀐 것은:

$\Rightarrow$ $z_{\alpha/2}$ 대신에 $t_{\alpha/2}$를 사용한다. 자유도 $n-1$인 스튜던트 t 분포에 해당한다.

$\Rightarrow$ σ 대신에 s를 사용한다.

상관성 분석

키포인트

- 피어슨, 스피어맨, 켄달 상관계수.
- 피어슨 상관계수의 신뢰구간.

피어슨 상관계수

- 피어슨 상관계수 (Pearson’s correlation)은 “일상적인 상관계수”이고 다음과 같은 수식으로 계산할 수 있다.

$$Cor(X,Y) = \frac{Cov(X,Y)}{s_X s_Y}$$

- 피어슨 상관계수의 값은 -1과 1사이의 수치이다.

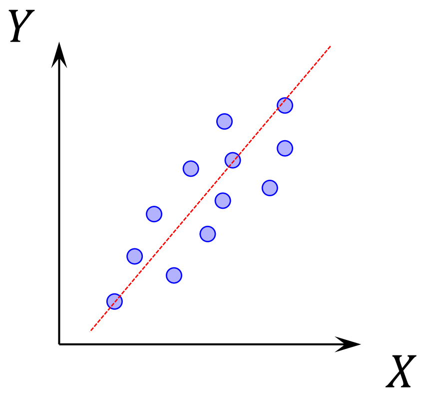
피어슨 상관계수

- 피어슨 상관계수는 선형관계의 방향과 강도를 나타낸다.

  $\Rightarrow Cor(X,Y) > 0$: X와 Y사이에 양의 선형관계가 있음.

  $\Rightarrow Cor(X,Y) < 0$: X와 Y사이에 음의 선형관계가 있음.

  $\Rightarrow Cor(X,Y) = 0$: X와 Y사이에 선형관계가 없음.



양의 선형관계

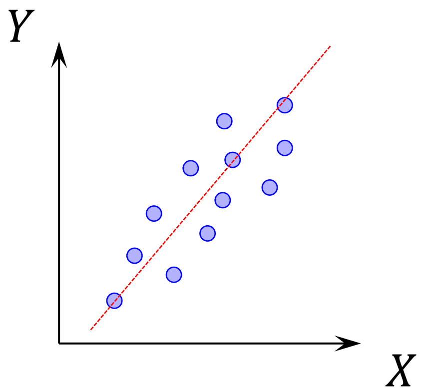
피어슨 상관계수

- 피어슨 상관계수는 선형관계의 방향과 강도를 나타낸다.

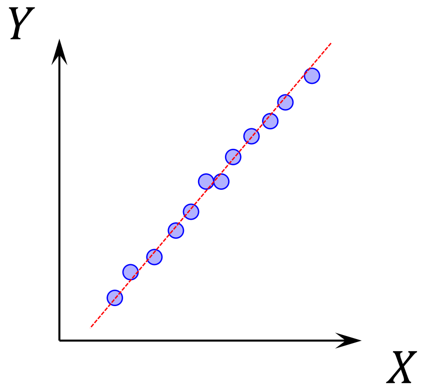
  $\Rightarrow Cor(X,Y)\ >\ 0$: X와 Y 사이에 양의 선형관계가 있음.

  $\Rightarrow Cor(X,Y)\ <\ 0$: X와 Y 사이에 음의 선형관계가 있음.

  $\Rightarrow Cor(X,Y)\ =\ 0$: X와 Y 사이에 선형관계가 없음.



<



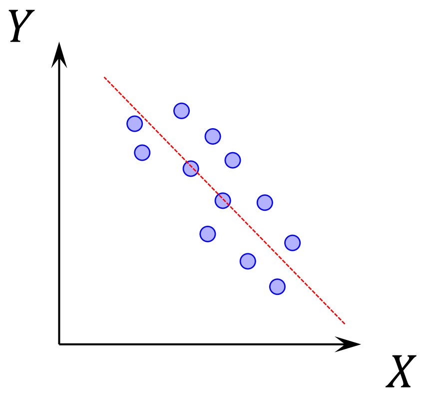
피어슨 상관계수

- 피어슨 상관계수는 선형관계의 방향과 강도를 나타낸다.

$\Rightarrow Cor(X,Y) > 0$: X와 Y사이에 양의 선형관계가 있음.

$\Rightarrow Cor(X,Y) < 0$: X와 Y사이에 음의 선형관계가 있음.

$\Rightarrow Cor(X,Y) = 0$: X와 Y사이에 선형관계가 없음.



음의 선형관계

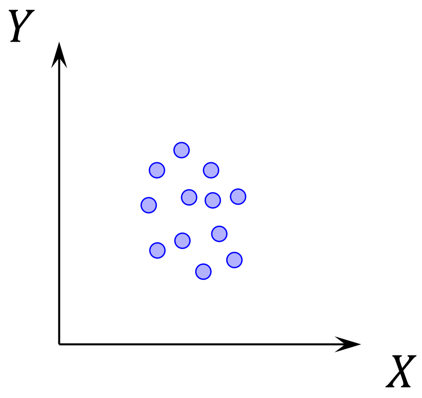
피어슨 상관계수

- 피어슨 상관계수는 선형관계의 방향과 강도를 나타낸다.

$\Rightarrow Cor(X,Y) > 0$: X와 Y사이에 양의 선형관계가 있음.

$\Rightarrow Cor(X,Y) < 0$: X와 Y사이에 음의 선형관계가 있음.

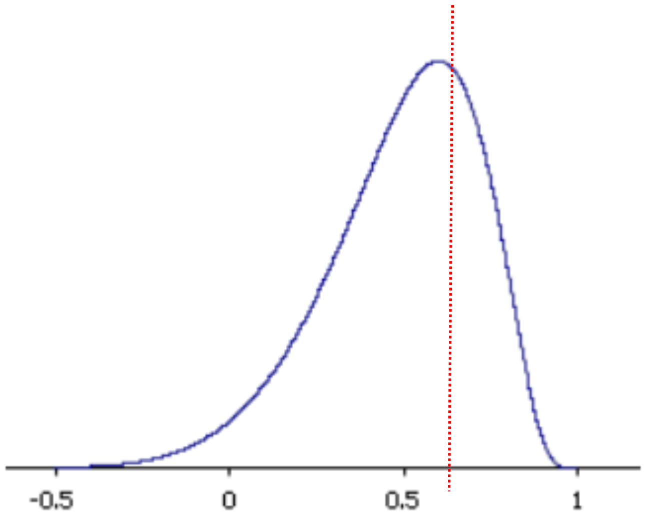
$\Rightarrow Cor(X,Y) = 0$: X와 Y사이에 선형관계가 없음.



선형관계 없음

피어슨 상관계수

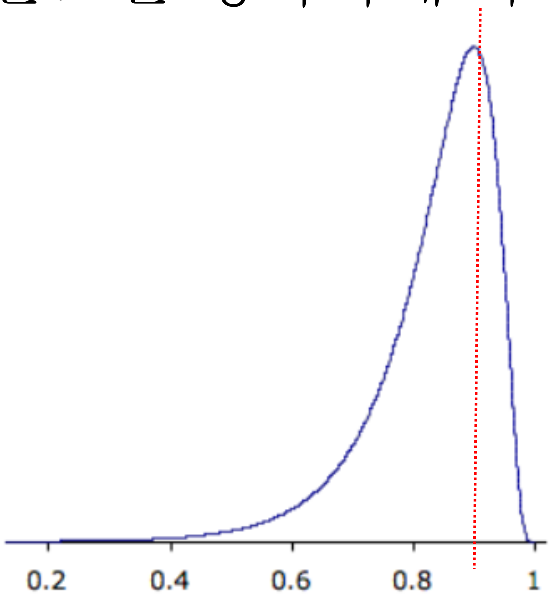
- 그런데 표본 상관계수 r은 정규분포를 정확하게 따르지 않는다.



$n = 12$, 모상관계수 $= 0.6$

피어슨 상관계수

- 그런데 표본 상관계수 r은 정규분포를 정확하게 따르지 않는다.



$n = 12$, 모상관계수 $= 0.9$

피어슨 상관계수

- 그런데 표본 상관계수 r은 정규분포를 정확하게 따르지 않는다.
- 다음과 같이 변환된 수치는 정규분포를 따른다: “피셔의 z 변환”

$$\Rightarrow z = 0.5\, ln\left(\frac{1+r}{1-r}\right) = \ arctanh(r)$$

$$\Rightarrow \sigma_z = \frac{1}{\sqrt{n-3}} \qquad \Leftarrow \ \text{“표준오차”}$$

- 역변환:

$$\Rightarrow r = \frac{e^{2z}-1}{e^{2z}+1} = \ tanh(z) \qquad \Leftarrow \ \text{“Hyperbolic Tangent”}$$

피어슨 상관계수

- 임의의 신뢰수준 확률 = $1-\alpha$에 해당하는 피어슨 상관계수의 신뢰구간을 만들 수 있다. 여기에서 $z_{\alpha/2} = (1-\alpha/2)$ 분위수와 같다. "피셔의 z" 와 혼동 주의!

하한: $tanh\left(z - z_{\frac{\alpha}{2}} \times \sigma_z\right)$

상한: $tanh\left(z + z_{\frac{\alpha}{2}} \times \sigma_z\right)$

[← <신뢰구간> →]

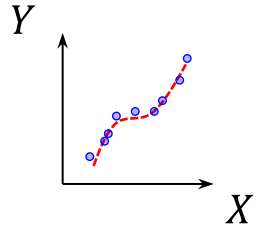
$\Rightarrow$ 95% 신뢰구간: $[tanh(z - 1.96\ \sigma_z)\ ,\ tanh(z + 1.96\ \sigma_z)]$

스피어맨 상관계수

- 스피어맨 상관계수 (Spearman's correlation)은 X와 Y 변수의 순위($rank$) 사이의 상관성을 나타낸다:

$$r_s = \frac{Cov(X_r, Y_r)}{s_{X_r} s_{Y_r}}$$

- 데이터의 순위만 정할 수 있다면 수치형 변수가 아니어도 적용이 가능하다.
- 스피어맨 상관계수의 값도 -1과 1사이의 수치이다.



- 스피어맨 상관계수는 X와 Y사이의 단조로움 (monotonicity)의 관계를 표현한다.

켄달 순위 상관계수

- 켄달 순위 상관계수 (Kendall's rank correlation)은 다음과 같은 원리로 계산할 수 있다.
- 먼저 (x, y) 테이블 형태로 주어진 데이터가 있다고 전제하고 i번째와 j번째를 비교해서 "부합"과 "비부합"을 가려낸다.

⇒ 부합: $x_i < x_j$ and $y_i < y_j$ 또는 $x_i > x_j$ and $y_i > y_j$

⇒ 비부합: $x_i < x_j$ and $y_i > y_j$ 또는 $x_i > x_j$ and $y_i < y_j$

$\boldsymbol{x}$	$\boldsymbol{y}$
x_1	y_1
x_2	y_2
x_3	y_3
$\vdots$	$\vdots$

켄달 순위 상관계수

- 켄달 순위 상관계수 r_k는 다음과 같이 계산한다:

$$r_k = \frac{(\text{부합 짝의 수}) - (\text{비부합 짝의 수})}{\frac{1}{2}n(n-1)}$$

- 켄달 순위 상관계수의 값도 -1과 1사이의 수치이다.

시각화와 분석

키포인트

- 시각화의 목적.
- 탐색적 시각화.
- 시각화 유형별 용도.
- 착시현상.

시각화의 목적

- 가설간의 비교 목적.
- 인과관계, 상응관계, 구조적 관계를 보여주는 목적.
- 다변량 데이터를 요약해서 보여주는 목적.
- 분석 결과를 요약해서 보여주는 목적.
- 기초 데이터와 결과를 뒷바침하는 증거를 정리하여 보여주는 목적.
- 결국은 컨텐츠 (스토리)의 유/무가 시각화의 효과를 결정한다!

탐색적 시각화 - EDA

- 데이터의 통계적 특성을 알아본다.
- 데이터에 패턴이 있는지 육안으로 확인해 본다.
- 향후 분석 방향을 정하는데 도움이 된다.
- 결과를 보고하는 목적으로도 사용된다.

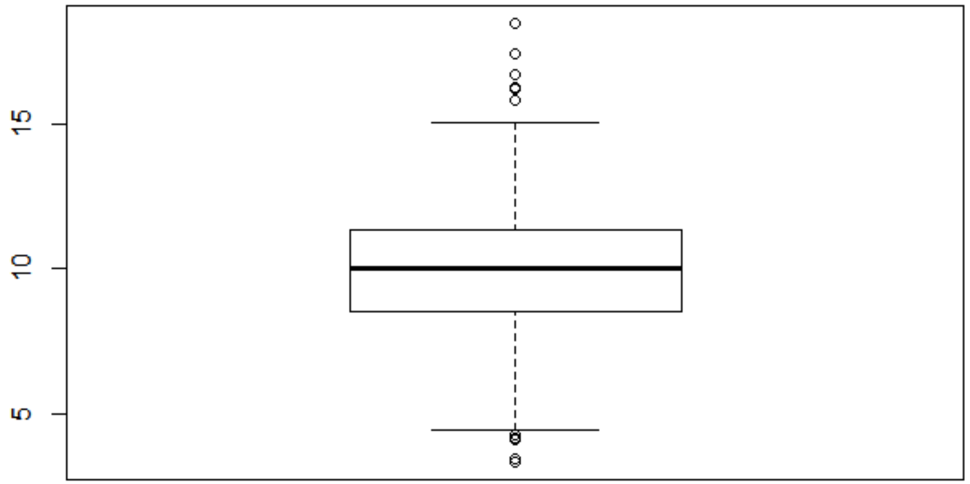
탐색적 시각화의 특징

- 많은 노력을 들이지 않고 짧은 시간안에 완성한다.
- 많은 수의 그래프를 생성한다.
- 분석가 본인의 직관적인 이해를 우선시 한다.
- 타이틀, 레이블 등은 크게 중요하지 않다.
- 컬러, 글꼴 등은 꼭 필요할 때만 사용한다.

단변량 기술통계 요약 시각화

- 상자그림 (Boxplot):

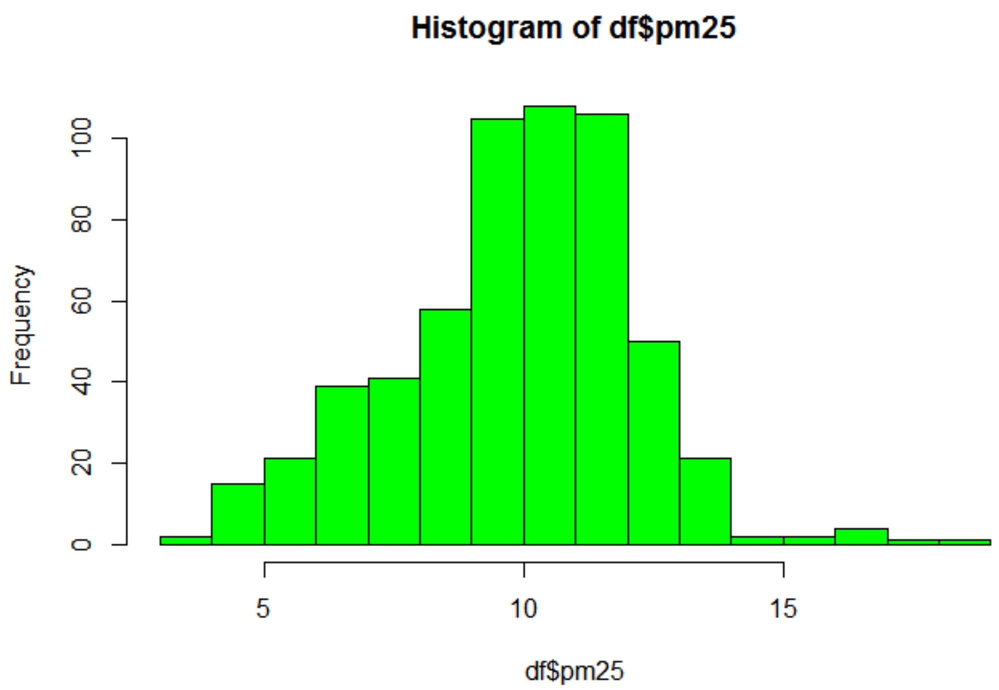
⇒ Box, whiskers, outlier로 구성됨. 연속형 변수 사용.



단변량 기술통계 요약의 용도

- 히스토그램 (Histogram):

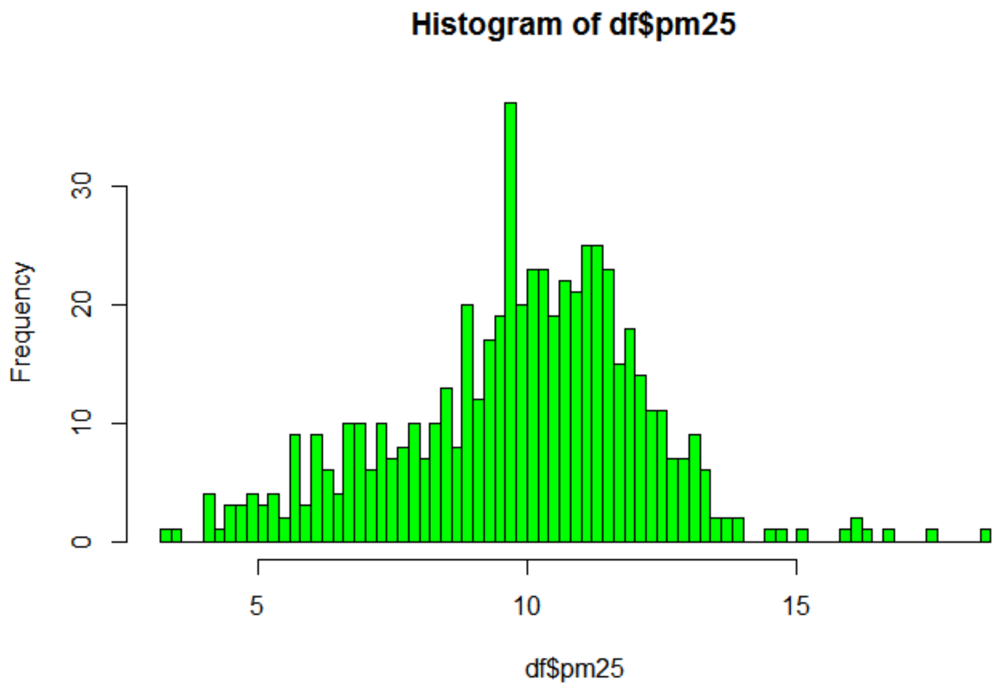
⇒ 계급에 해당하는 도수를 보여줌. 연속형 변수 사용.



단변량 기술통계 요약의 용도

- 히스토그램 (Histogram):

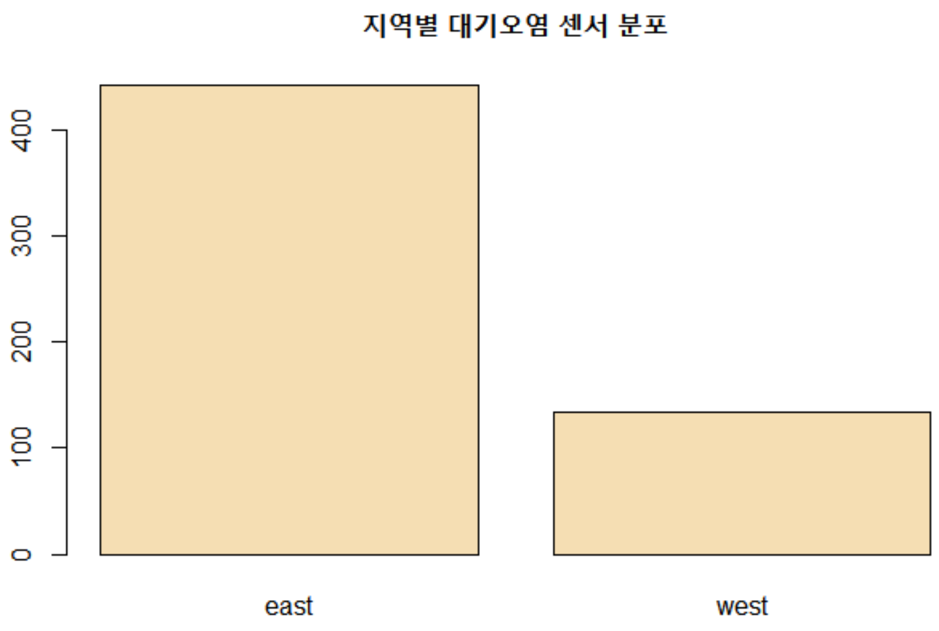
⇒ 계급의 크기를 조정할 수 있다. 연속형 변수 사용.



단변량 기술통계 요약의 용도

- 막대그림 (Bar Chart):

  ⇒ 명목형 변수의 도수를 보여준다.



단변량 기술통계 요약의 용도

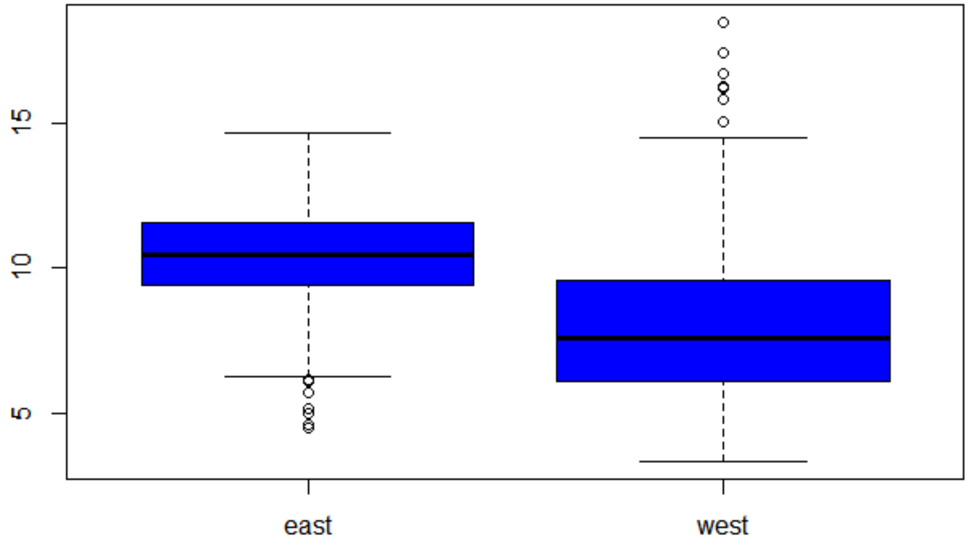
- 파이 (Pie Chart):

  ⇒ 명목형 변수의 상대도수를 보여준다.

비교, 상관관계 확인의 용도

- 다중 상자그림 (Multiple Boxplot):

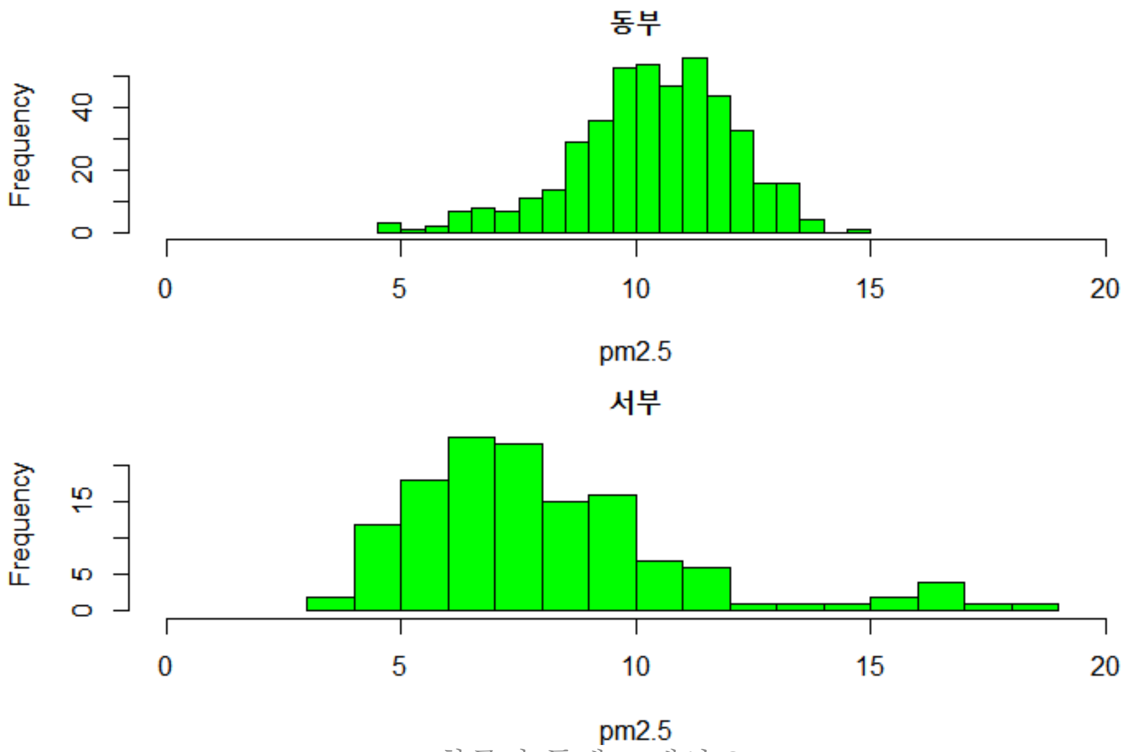
⇒ 제2의 명목형 변수에 의해서 여러 개의 상자그림이 생성된다.



비교, 상관관계 확인의 용도

- 다중 히스토그램 (Multiple Histogram):

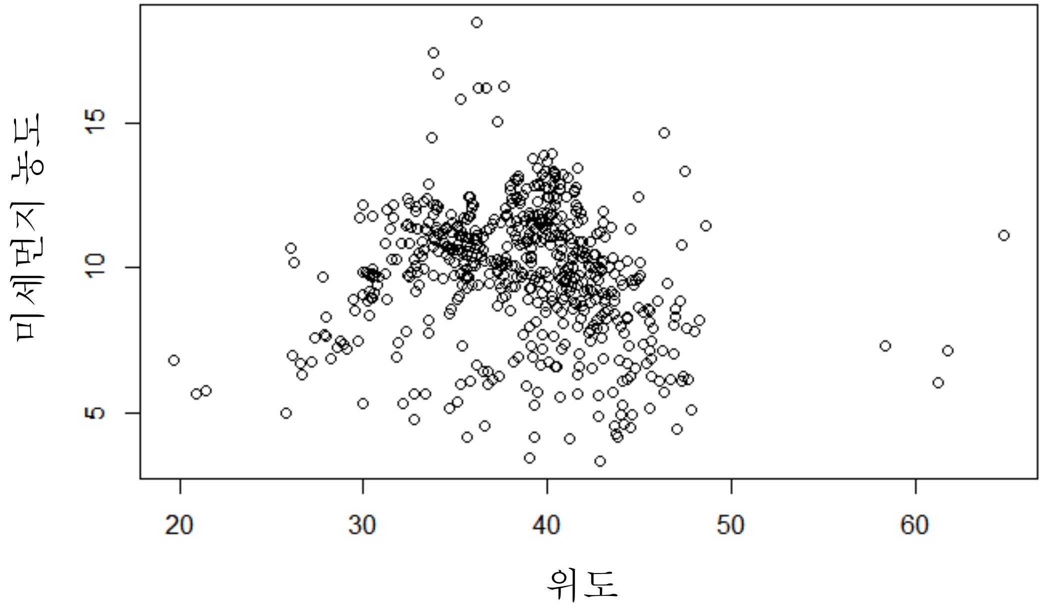
⇒ 제2의 명목형 변수에 의해서 여러 개의 히스토그램이 생성된다.



비교, 상관관계 확인의 용도

- 산점도 (Scatter Plot):

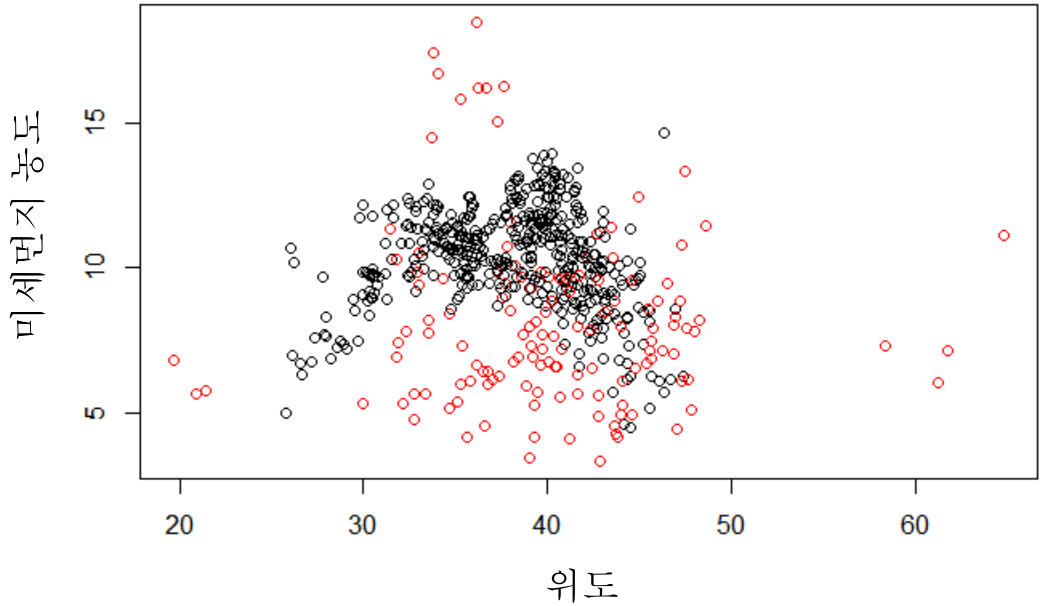
$\Rightarrow$ X와 Y 두 개의 연속형 변수 사이의 관계를 보여준다.



비교, 상관관계 확인의 용도

- 산점도 (Scatter Plot):

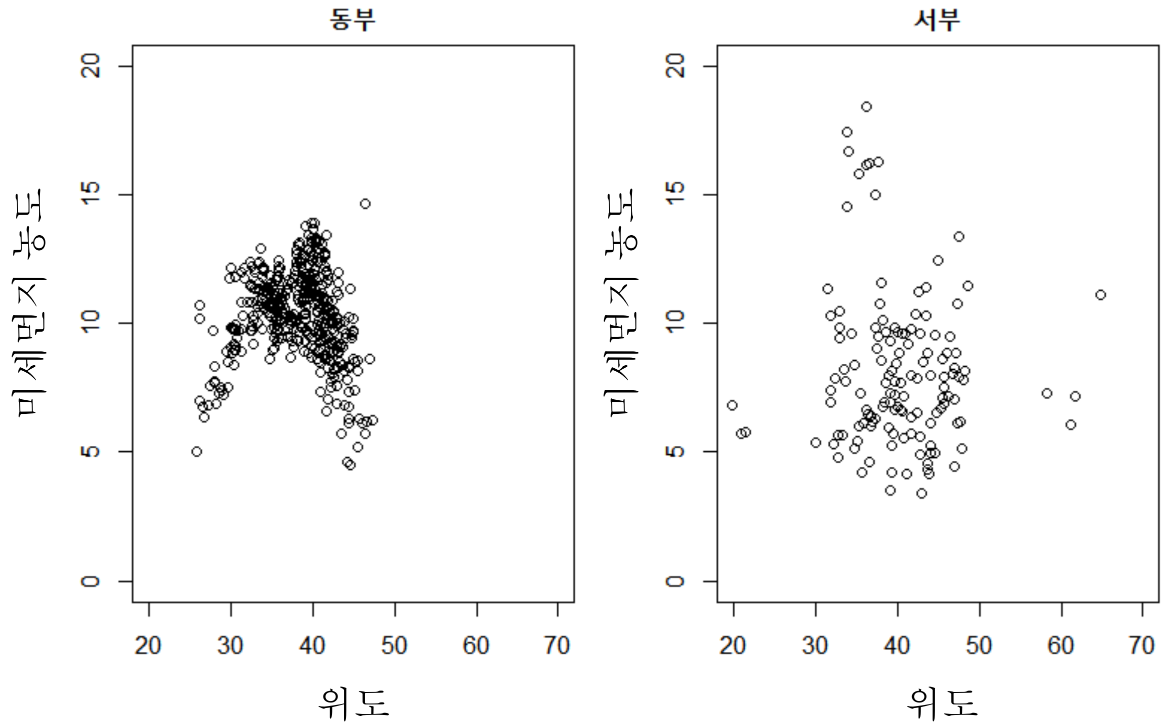
⇒ 제3의 명목형 변수의 유형을 컬러를 사용해서 표현할 수 있다.



비교, 상관관계 확인의 용도

- 다중 산점도 (Multiple Scatter Plot):

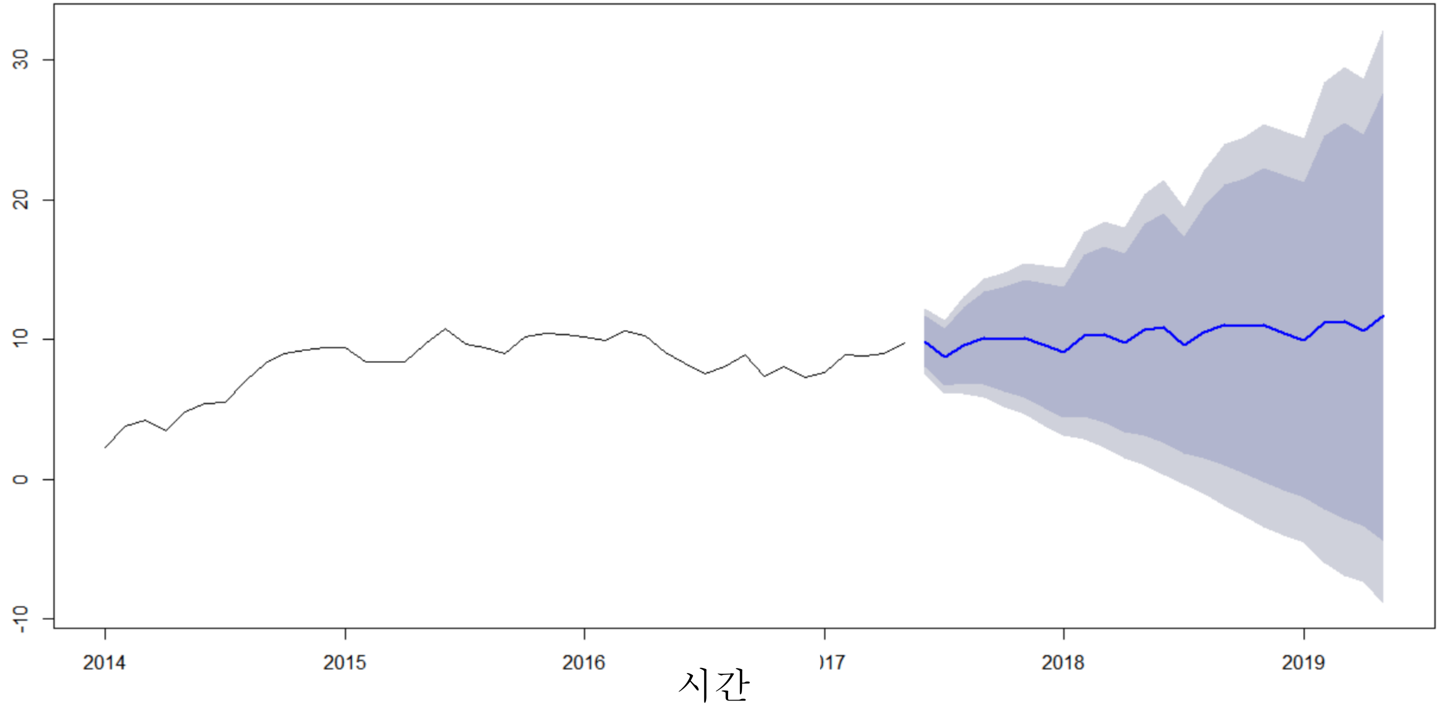
  ⇒ 제3의 명목형 변수의 유형별 구별된 산점도로 표현할 수 있다.



시간 추세 확인의 용도

- 시계열 그래프 (Time Series Plot):

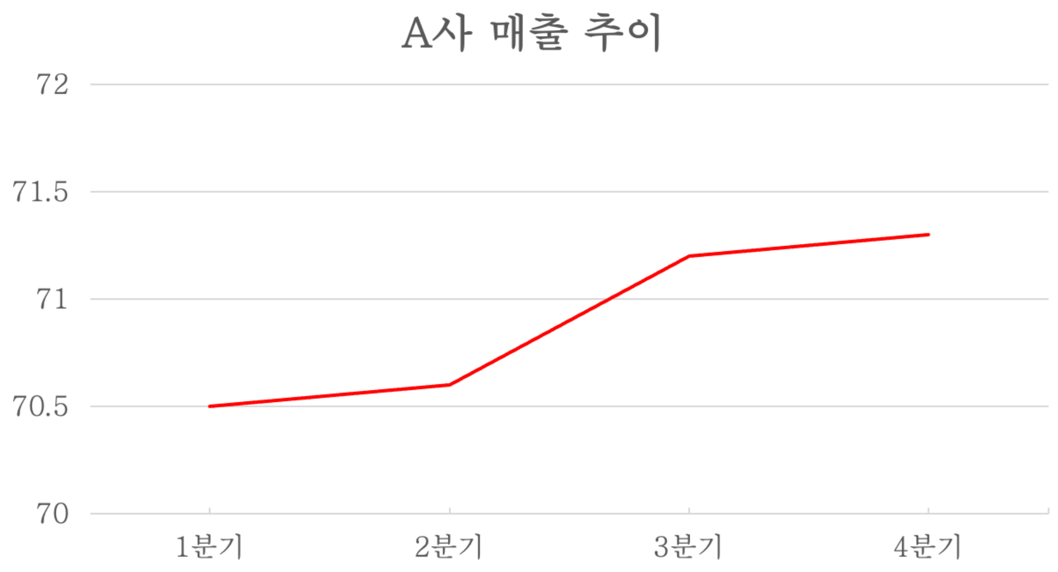
⇒ 시간이 지남에 따라서 변하는 수치를 나타낸다. 예측 신뢰구간도 표현 가능.



착시현상

- 다음은 A사의 분기별 매출 현황을 나타낸 그래프이다.

  ⇒ 큰 폭의 매출 증가?



착시현상

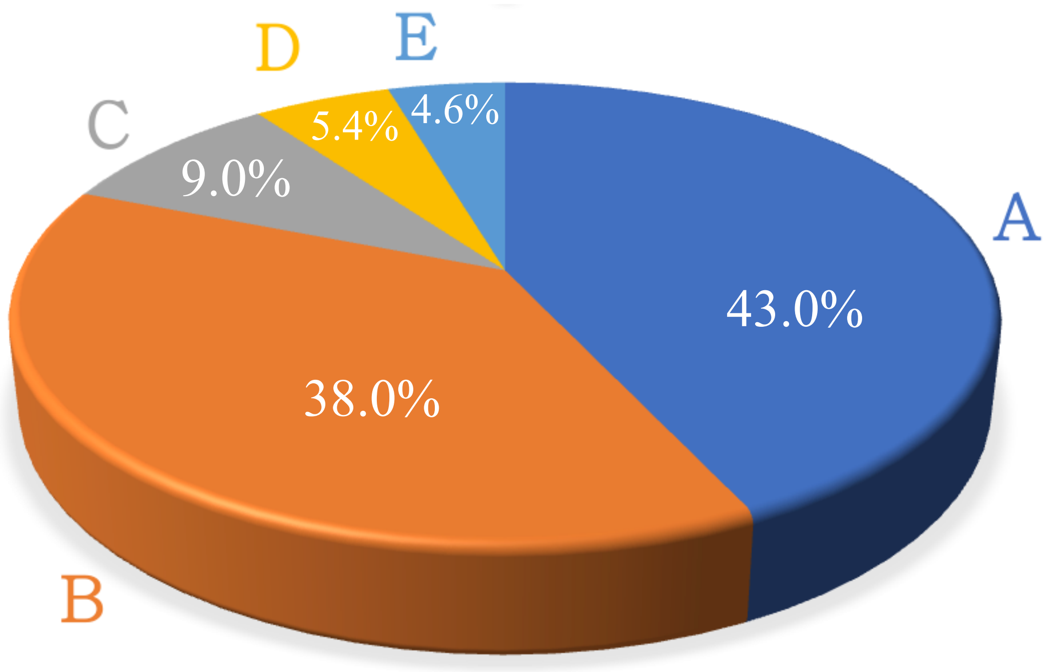
- 다음은 A사의 분기별 매출 현황을 나타낸 그래프이다.

  ⇒ 세로축 스케일을 조정하여 전체를 본다. 대략 2% 매출 증가.

착시현상

- 파이차트는 주의하여 사용한다.

⇒ 특히 3D 파이차트는 원근법 때문에 큰 착시 현상이 발행한다!



끝