Estimation (point & intervals)

Contents

1. 표본을 통한 모집단 추정

표본을 통한 모집단 추정

Definitions

모집단(population)

- 의사결정을 도출하기 위해 관심을 갖고 연구해야 할 대상, 즉 통계분석의 연구대상이 되는 모든 개체들의 집합을 의미

모수(parameter)

- 모집단의 특성을 나타내는 수치로서 평균, 표준편차, 분산, 비율 등 다양하게 있으며, 통계학에서는 의사결정을 위해 특별히 관심을 갖는 모수만을 선택하여 분석하게 된다.
- 예를 들면, 정규분포에서 모평균 μ를 추정하고자 한다면 모평균이 모수가 된다.

표본(sample)

- 모집단의 특성을 파악하기 위해 모집단으로부터 일정한 규칙에 의해 추출한 모집단의 부분집합
- 특히, 확률표본(random sample)이란 독립적이며 동일한 분포를 따르는(iid: independent and identically distributed) 확률 변수들의 집합을 의미한다.

표본을 통한 모집단 추정

기술통계 vs. 추론통계

기술통계(descriptive statistics)

- 수집된 데이터를 정리하고 그 내용을 특정 짓는 몇 가지의 대표 값으로 산출하며, 그래프로도 표현하여 모집단의 특성을 파악하는 방법을 다루는 분야이다.

추론통계(statistical inference)

- 실험이나 조사를 통해 얻은 자료를 어떤 모집단에서 얻어낸 확률표본 이라고 보고, 그 통계량으로부터 모집단의 분포를 특정 짓는 모수를 추측하려는 것이다. 즉, 추론통계학은 데이터에 내포되어 있는 정보를 분석하여 불확신한 사실에 대한 추론을 하는 분야이다.

표본을 통한 모집단 추정

점 추정 vs. 구간 추정

점 추정(point estimation)

- 추정량의 관측된 값을 통해 모수의 참값을 추정하는 절차

구간 추정(interval estimation)

- 모수의 참값을 포함할 확률이 신뢰구간 $1-\alpha$가 되는 신뢰구간을 결정하는 절차

표본을 통한 모집단 추정

통계량과 추정량

통계량(statistic)

- 미지의(unknown) 모수를 포함하지 않는 확률표본의 함수

추정량(estimator)

- 미지의 모수를 추정하기 위한 통계량

추정치(estimate)

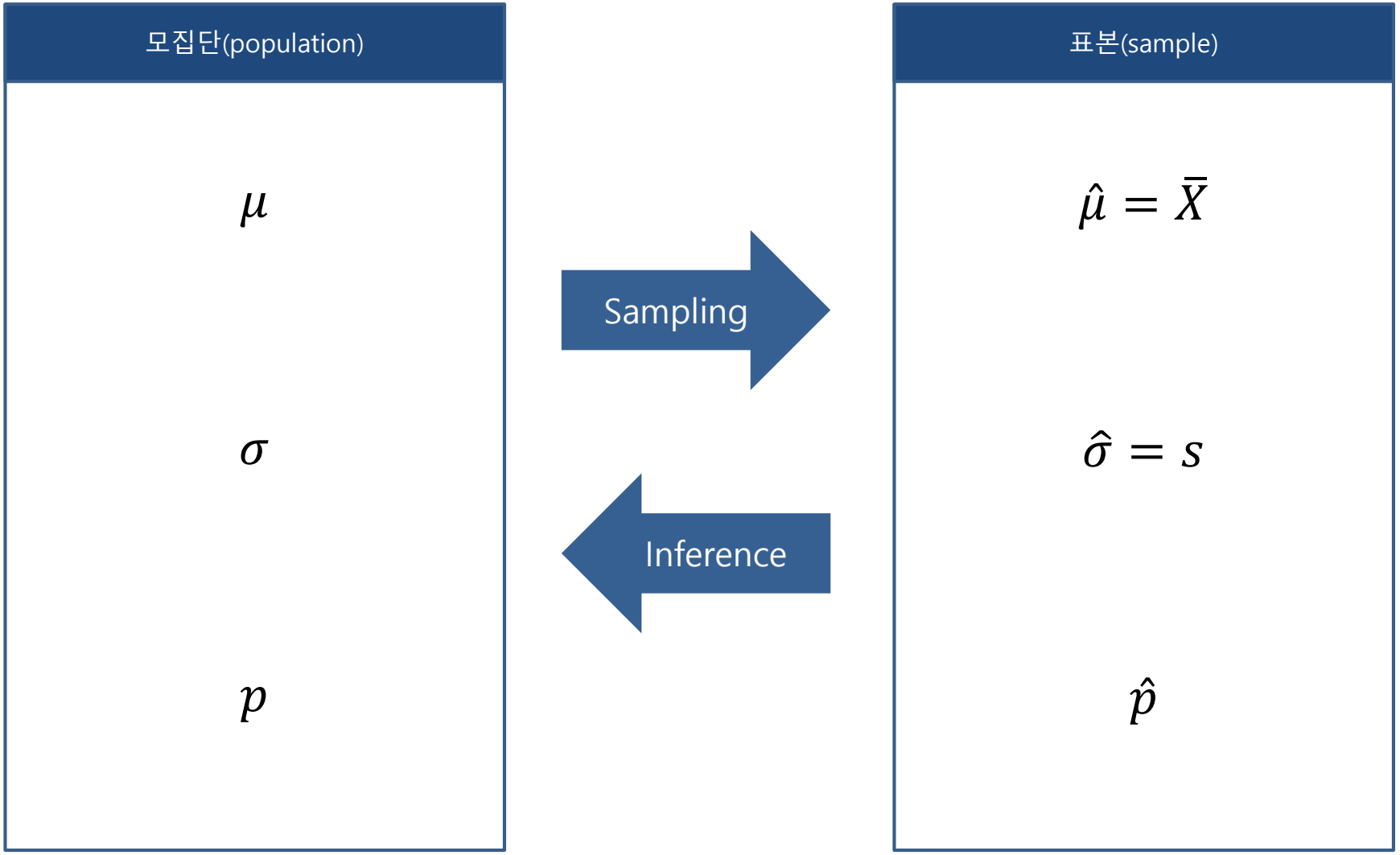
- 자료를 통해 산출된 추정량을 추정치로 정의
- 추정량 $\bar{X}$는 본 연구자료를 통해 10.25 값을 가지게 되는데 이를 추정치라고 부른다.

불편성(unbiasedness)

- 추정량의 기댓값이 추정하고자 하는 모수와 같아지는 특성으로서, 좋은 추정량이 되기 위한 첫 번째 요건
- 예를 들어, 추정량 $\bar{X}$는 모수 μ의 불편성을 가지므로 불편추정량이라고 부른다.

표본을 통한 모집단 추정

집단에서 사용되는 통계량



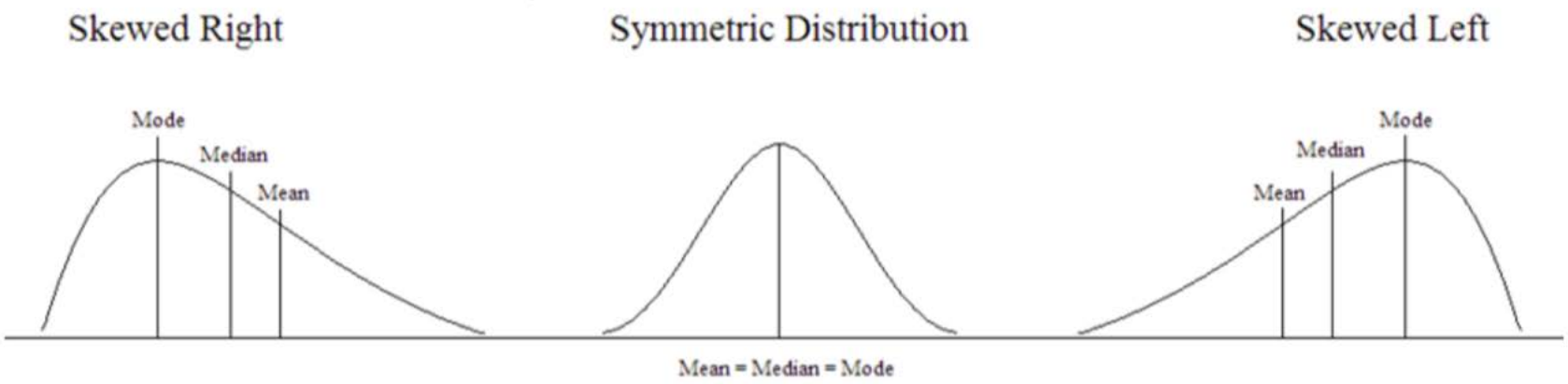
2. 중심위치 및 산포의 척도

중심위치 및 산포의 척도

1) 중심위치의 척도

- 평균(mean): $\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- 중앙값(median): 데이터를 크기 순으로 정렬하여 가운데 위치하는 값으로 정의한다.

  자료가 짝수 개인 경우는 가운데 위치한 2개의 값의 평균을 중앙 값으로 정의한다.
- 최빈값(mode): 가장 많이 발생한 값
- 절사평균(trimmed mean): 10% 절사평균이란 상하위 10%씩을 제외한 나머지 자료의 평균
- 분포가 대칭이고 이상점이 존재하지 않으면 표본평균을 사용한다.
- 비대칭이거나 이상점이 존재하면 중앙값을 사용한다. (생존기간 등)
- 순서척도는 중앙값을, 명목척도는 최빈값을 주로 사용한다.



중심위치 및 산포의 척도

1) 중심위치의 척도(예제)

2 7 3 5 5 7 6 1 4 5

- 평균(mean) = 4.5

  > mean(c(2,7,3,5,5,7,6,1,4,5))

  # [option] na.rm=TRUE ## Hmisc::describe

- 중앙값(median) = (5번째 + 6번째 값)/2 = 5

  ✓ 1 2 3 4 5 5 5 6 7 7

  > median(c(2,7,3,5,5,7,6,1,4,5))

- 최빈값(mode) = 5

  > names(which.max(table(c(2,7,3,5,5,7,6,1,4,5))))

- 10% 절사평균(trimmed mean) = (2+3+4+5+5+5+6+7)/8 = 4.625

  > mean(c(2,7,3,5,5,7,6,1,4,5), trim=0.1)

중심위치 및 산포의 척도

2) 산포의 척도

- 표본분산(sample variance): $\hat{\sigma}^2 = s^2 = \frac{\sum_{i-1}^{n}(x_i - \bar{x})^2}{n-1}$
  - ✓ 표본분산 s^2을 계산할 때, $\sum_{i=1}^{n}(x_i - \bar{x})^2$ 을 n으로 나누지 않고 $n-1$로 나누는 이유는 바로 불편추정량을 만들기 위함이다.
  - ✓ 불편추정량이란 $E(s^2) = \sigma^2$를 만족하는 추정량을 의미한다.
- 표본표준편차(sample standard deviation): $s = \sqrt{s^2}$
- 범위(range): $R = x_{max} - x_{min}$
- 사분위수 범위(inter-quartile range): $IQR = Q_3 - Q_1$

  여기서, Q_1는 제1사분위수를 Q_3는 제3사분위수를 의미한다.
- 변동계수(coefficient of variation, CV): $\frac{s}{\bar{x}}$

중심위치 및 산포의 척도

2) 산포의 척도(예제)

2 7 3 5 5 7 6 1 4 5

- 평균(mean) = 4.5
- 표본분산(sample variance): $s^2 = \frac{\sum_{i-1}^{n}(x_i - \bar{x})^2}{\boldsymbol{n-1}} = \frac{(2-4.5)^2+(7-4.5)^2+\cdots+(5-4.5)^2}{9} = 4.056$

  > var(c(2,7,3,5,5,7,6,1,4,5))

- 표본표준편차(sample standard deviation): $s = \sqrt{s^2} = 2.014$

  > sd(c(2,7,3,5,5,7,6,1,4,5))

- 범위(range): $R = x_{max} - x_{min} = 7 - 1 = 6$

  > range(c(2,7,3,5,5,7,6,1,4,5))

중심위치 및 산포의 척도

2) 산포의 척도(예제)

2 7 3 5 5 7 6 1 4 5

- 사분위수 범위(inter-quartile range, IQR): $IQR = Q_3 - Q_1 = 5.75 - 3.25 = 2.5$
  - 1 2 3 4 5 5 5 6 7 7

  > ?quantile #해당 함수를 통해서 다양한 quantile 정의 확인

  - R에서는 Q1을 1+0.25*(n-1)번째 데이터 값으로, Q3는 1+0.75*(n-1)번째 데이터 값으로 계산
  - 기본적인 통계량 값은 summary 함수를 통해서 확인 가능

  > summary(c(2,7,3,5,5,7,6,1,4,5))

  - Baseline characteristics table에서의 IQR로 기재함은 "median(Q1, Q3)"을 의미함
- 변동계수(coefficient of variation, CV): $\frac{s}{\bar{x}} = \frac{2.014}{4.5} = 0.45$

  > sd(c(2,7,3,5,5,7,6,1,4,5)) / mean(c(2,7,3,5,5,7,6,1,4,5))

중심위치 및 산포의 척도

3) 비율(proportions)

- 이산형(binary or dichotomous nominal) 변수에 대한 척도로 비율을 주로 사용한다.
- P = # of respondents / sample size
- 논문 상에서는 해당 범주 대상자 수(=count)와 비율을 함께 기재한다.

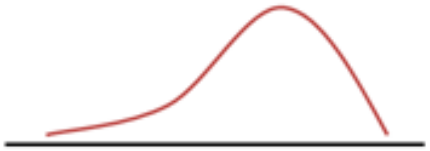
Table 3. Most Common Adverse Events in the Safety Population.*

Event	Daratumumab Group (N=243)		Control Group (N=237)	
	Any Grade	Grade 3 or 4	Any Grade	Grade 3 or 4
	number of patients (percent)			
Common hematologic adverse event				
Thrombocytopenia	143 (58.8)	110 (45.3)	104 (43.9)	78 (32.9)
Anemia	64 (26.3)	35 (14.4)	74 (31.2)	38 (16.0)
Neutropenia	43 (17.7)	31 (12.8)	22 (9.3)	10 (4.2)
Lymphopenia	32 (13.2)	23 (9.5)	9 (3.8)	6 (2.5)
Common nonhematologic adverse events				
Peripheral sensory neuropathy	115 (47.3)	11 (4.5)	89 (37.6)	16 (6.8)
Diarrhea	77 (31.7)	9 (3.7)	53 (22.4)	3 (1.3)
Upper respiratory tract infection	60 (24.7)	4 (1.6)	43 (18.1)	2 (0.8)
Fatigue	52 (21.4)	11 (4.5)	58 (24.5)	8 (3.4)
Cough	58 (23.9)	0	30 (12.7)	0
Constipation	48 (19.8)	0	37 (15.6)	2 (0.8)
Dyspnea	45 (18.5)	9 (3.7)	21 (8.9)	2 (0.8)
Insomnia	41 (16.9)	0	35 (14.8)	3 (1.3)
Peripheral edema	40 (16.5)	1 (0.4)	19 (8.0)	0
Asthenia	21 (8.6)	2 (0.8)	37 (15.6)	5 (2.1)
Pyrexia	38 (15.6)	3 (1.2)	27 (11.4)	3 (1.3)
Pneumonia	29 (11.9)	20 (8.2)	28 (11.8)	23 (9.7)
Hypertension	21 (8.6)	16 (6.6)	8 (3.4)	2 (0.8)
Secondary primary cancer†	6 (2.5)	NA	1 (0.4)	NA

중심위치 및 산포의 척도

4) 첨도(kurtosis) & 왜도(skewness)

왜도 Skewness

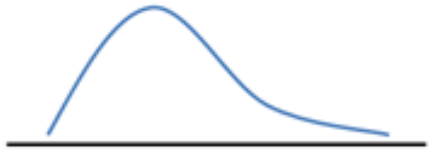


Negative Skew (S<0)

왼쪽으로 긴 꼬리

오른쪽으로 치우친 분포

예) 시험성적 분포



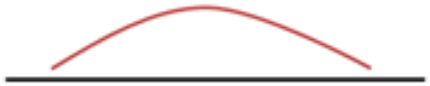
Positive Skew (S>0)

오른쪽으로 긴 꼬리

왼쪽으로 치우친 분포

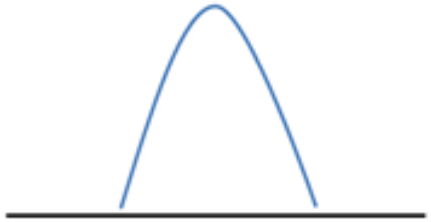
예) 소득자료

첨도 Kurtosis



Negative Kurtosis (K<0)

정규분포보다 납작한 분포



Positive Kurtosis (K>0)

정규분포보다 뾰족한 분포

3. 확률변수의 기댓값

확률변수의 기댓값

확률변수의 기댓값(expected value)

- 이산형: $\mu_x = E(X) = \sum_x xf(x)$
- 연속형: $\mu_x = E(X) = \int_{-\infty}^{\infty} xf(x)dx$
- 기댓값의 특성
  - ✓ $E(aX + b) = aE(X) + b$ (a, b는 상수)
  - ✓ $E(X + Y) = E(X) + E(Y)$
  - ✓ 확률변수 X와 Y가 서로 독립이면 $E(XY) = E(X)E(Y)$

확률변수의 기댓값

확률변수의 기댓값(expected value)

- (예제) 동전을 세 번 던져 나온 (앞면의 개수-뒷면의 개수)만큼 100원씩 주고받는 게임에서의 수익을 확률변수 X라 할 때, X의 기댓값을 구하시오.

Sample space		확률변수 X
HHH	→	300
HHT, HTH, THH	→	100
HTT, THT, TTH	→	-100
TTT	→	-300

- ✓ $f(300) = f(-300) = \frac{1}{8}$, $f(100) = f(-100) = \frac{3}{8}$
- ✓ $E(X) = \sum_x xf(x) = (300 - 300) \times \frac{1}{8} + (100 - 100) \times \frac{3}{8} = 0$
- ✓ 따라서 이 게임은 공정하다고 할 수 있다.

확률변수의 기댓값

확률변수의 분산(variance)

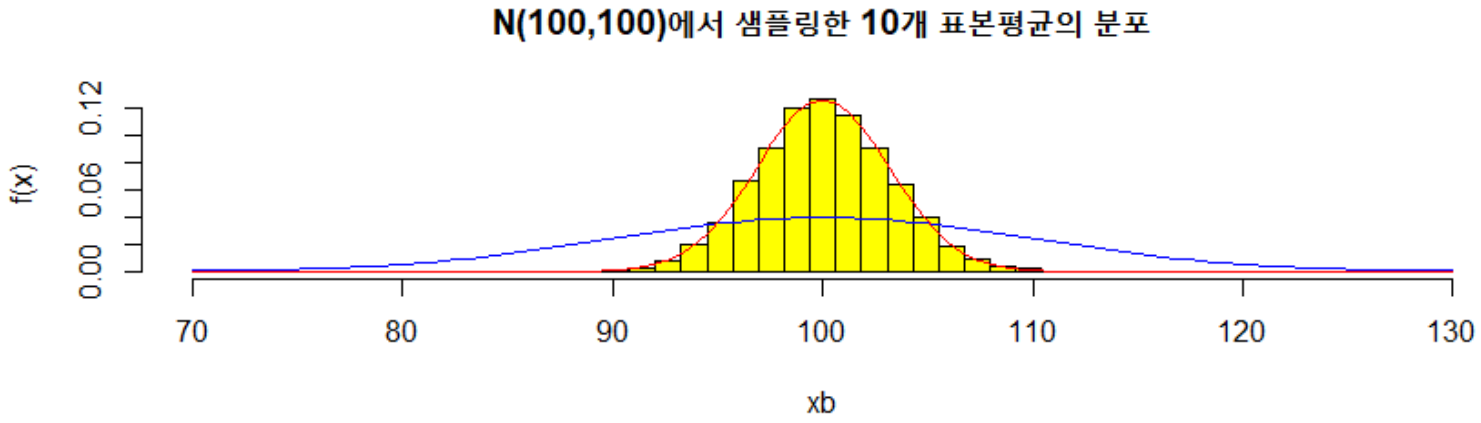
- $Var(X) = \sigma_x^2 = E[(X - E(X))^2]$
  - ✓ 분산의 특성: $Var(aX + b) = a^2 Var(X)$, (여기서 a, b는 상수)

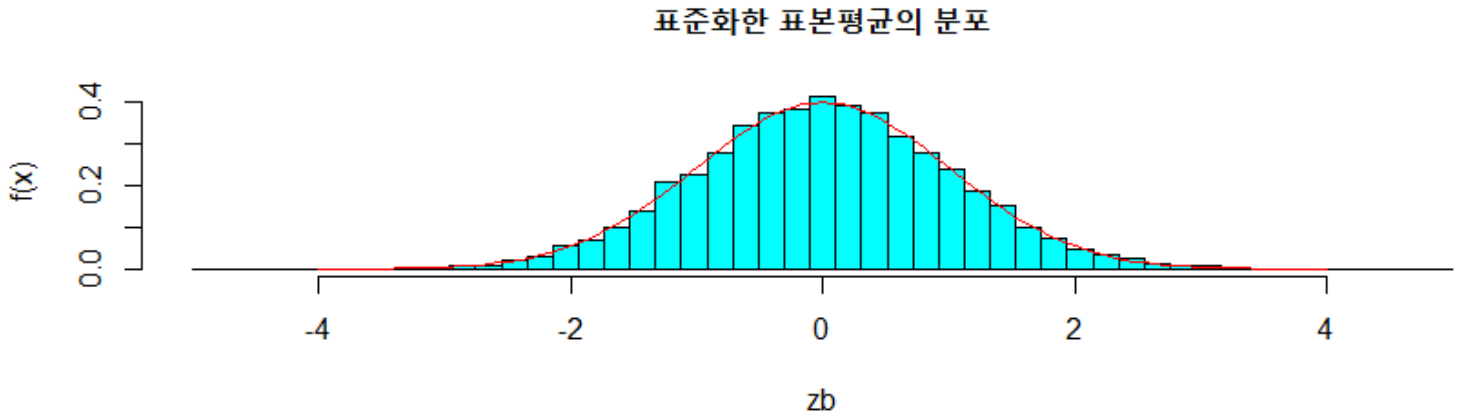
- 두 확률변수의 공분산(covariance)
  - ✓ $Cov(X, Y) \equiv \sigma_{XY} = E[(X - \mu_X)(Y - \mu_Y)]$
  - ✓ Covariance는 linear mixed model에서의 반복측정된 값들의 상관구조를 확인, 또는 변수들간의 interaction term 확인

확률변수의 기댓값

표본평균 vs. 표준화된 표본평균의 분포

- 평균 100, 분산 100인 정규분포 모집단에서 10개씩 표본을 랜덤 추출하여 표본평균을 구하는 작업을 10,000회 반복하시오. (R 실습!)

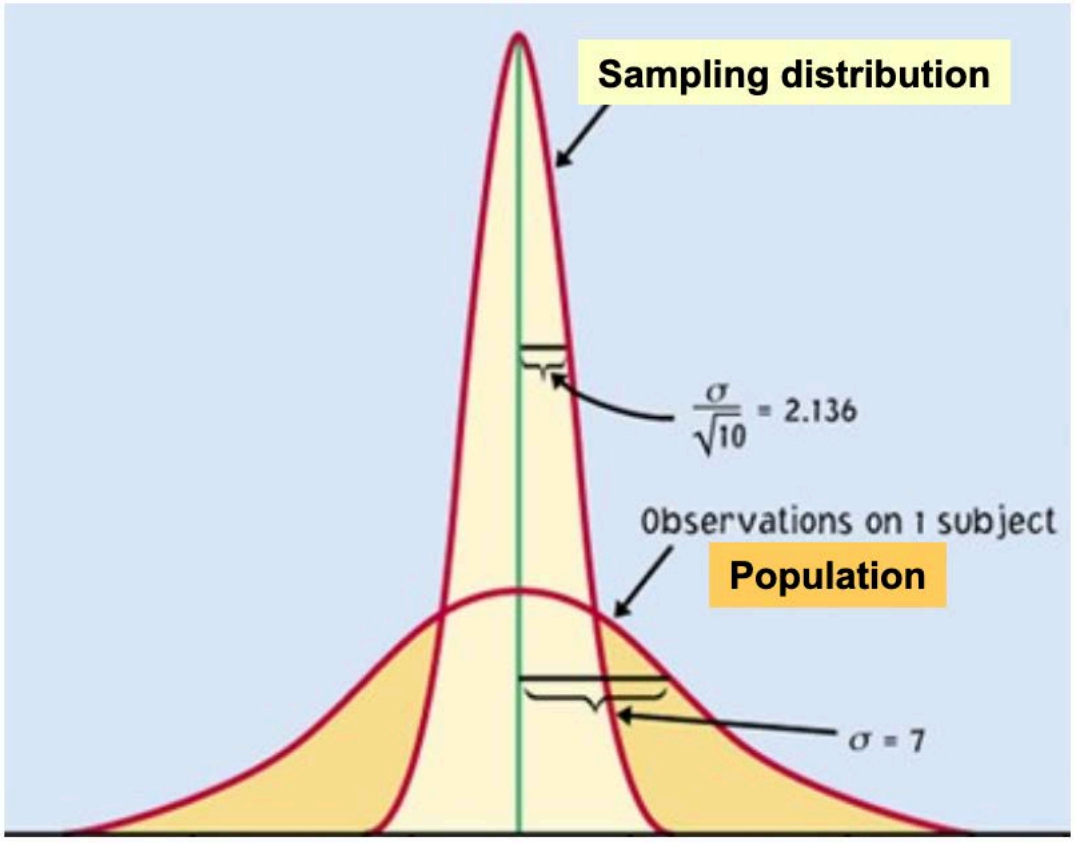




확률변수의 기댓값

X의 분포 vs. $\bar{X}$의 분포

- Standard deviation vs. standard error



구간 추정

확률변수의 구간 추정

$$\boldsymbol{Point\ estimate} \pm \boldsymbol{margin\ of\ error}$$

$$\bar{X} \pm Z(\frac{s}{\sqrt{n}})$$

$$\bar{X} \pm t(\frac{s}{\sqrt{n}})$$

- 하나의 통계량인 $\bar{X}$만으로 추정하지 않고 구간을 통해서 모수 μ를 추정할 수 있다.
- Large sample에 대해선 Z와 t 값은 거의 동일하게 되므로, 신뢰구간(confidence intervals)도 거의 유사한 값을 가지게 된다.

R 실습

Baseline characteristics에서 통계량

- *R-실습!*

Table 1. **Baseline Demographic and Clinical Characteristics of the Patients.***

Characteristic	CPAP (N=307)	Intubation (N=303)	P Value†
Gestational age (wk)	26.91±1.0	26.87±1.0	0.63
Gestational age of 25 or 26 wk (%)	33	35	0.59
Birth weight (g)	964±212	952±217	0.48
Use of antenatal corticosteroids (%)	94	94	0.76
Cesarean section (%)	66	69	0.51
Mother in labor (%)	65	66	0.82
Rupture of membranes (days before birth)			
Median	0	0	0.65
Interquartile range	0–2	0–1	
Male sex (%)	49	56	0.05
Multiple births (%)	35	32	0.57
Resuscitation device used (%)‡			0.13
None	19.5	13.9	
Self-inflating bag	14.7	16.5	
Self-inflating bag plus CPAP	13.0	14.2	
Flow-inflating bag	5.2	6.6	
Neopuff or bubble CPAP	47.2	46.5	
Apgar score at 5 minutes			
Median	9	8	0.001
Interquartile range	8–9	8–9	

* Plus–minus values are means ±SD. CPAP denotes continuous positive airway pressure.
† P values were calculated by the t-test, the chi-square test, or the Mann–Whitney test.
‡ In this category, eight infants (1.3%) were excluded because the resuscitation method was classified as "other."

Thank you!