강원대학교
전자·AI시스템공학과

머신러닝1

- 기초통계 -

추론통계(평균의 분포)

추론통계

확률 변수는 : 확률로 결정되는 변수

- 확률의 종류는 이산확률변수, 연속확률변수 이를 표현하는 것은 이산확률분포, 연속확률분포
- 확률변수의 원소의 개수를 셀 수 있을 경우 이를 이산확률변수라고 함

예제)

한 개의 동전을 던지는 시행에서 표본공간에서 발생할 수 있는 것은

동전 = {앞면, 뒷면}

한 개의 주사위를 던지는 시행에서 표본공간에서 발생할 수 있는 것은

주사위 = {1, 2, 3, 4, 5, 6}

추론통계

확률 분포(probability distribution) : 표본공간의 각 원소에 대응된 확률변수에 각각의 값을 가질 확률을 대응시킨 관계

- **표본 공간에서 얼마만큼의 질량을 가지고 있는가를 궁금함**
- **이산인 경우 y축은 질량, 연속인 경우 y축은 밀도 부피는 구간**
- **확률변수 X에 대한 누적된 확률을 누적분포함수라고 함**

예제)

X**의 확률분포** $= P(X = x)$ or $f(x)$ **서로 다른 두 개의 동전을 던지는 시행에서 앞면의 개수를 확률변수 X**

$$P(X = 0) = \frac{1}{4}, P(X = 1) = \frac{1}{2}, P(X = 2) = \frac{1}{4},$$

추론통계

확률 변수는 : 확률로 결정되는 변수

- **확률의 종류는 이산확률변수, 연속확률변수 이를 표현하는 것은 이산확률분포 연속확률분포**
- **확률변수의 원소의 개수를 셀 수 없을 경우 이를 연속확률변수라고 함**

예제)

학교의 학생들의 키

2023년 지역별 강우량

추론통계

확률 분포(probability distribution) : 표본공간의 각 원소에 대응된 확률변수에 각각의 값을 가질 확률을 대응시킨 관계

- **표본 공간에서 얼마만큼의 질량을 가지고 있는가를 궁금함**
- **이산인 경우 y축은 질량, 연속인 경우 y축은 밀도 부피는 구간**
- **확률변수 X에 대한 누적된 확률을 누적분포함수라고 함**

예제)

X**의 확률분포** $= P(X \leq x)$ or $\mathbf{F}(\mathrm{x})$ **어떤 사람을 뽑았을 때, 키가 160인 학생이 뽑힐 확률은?**

$$P(\boldsymbol{a} \leq x \leq \boldsymbol{b}) = \int_{\boldsymbol{a}}^{\boldsymbol{b}} \boldsymbol{f}(\boldsymbol{x})\boldsymbol{dx}$$

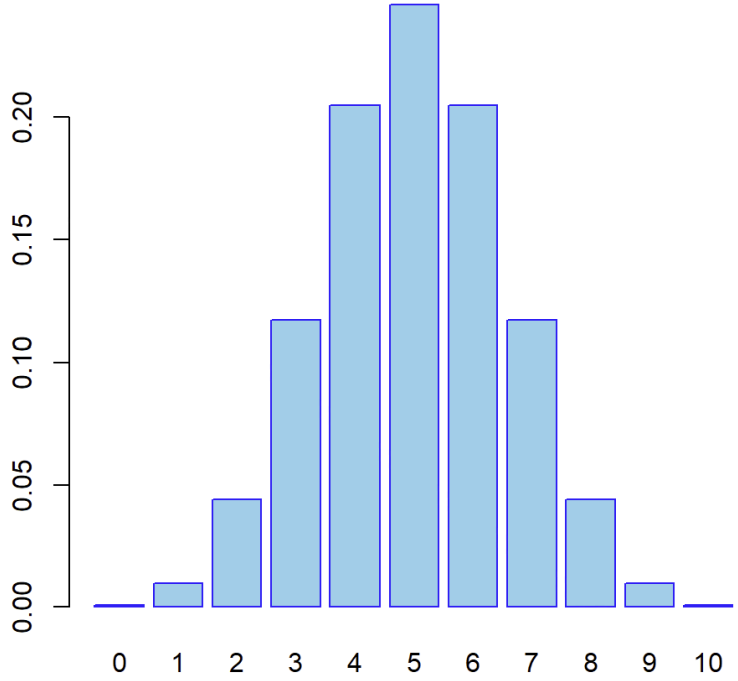
어떤 사람을 뽑았을 때, 키가 160에서 170사이에 학생이 뽑힐 확률은?

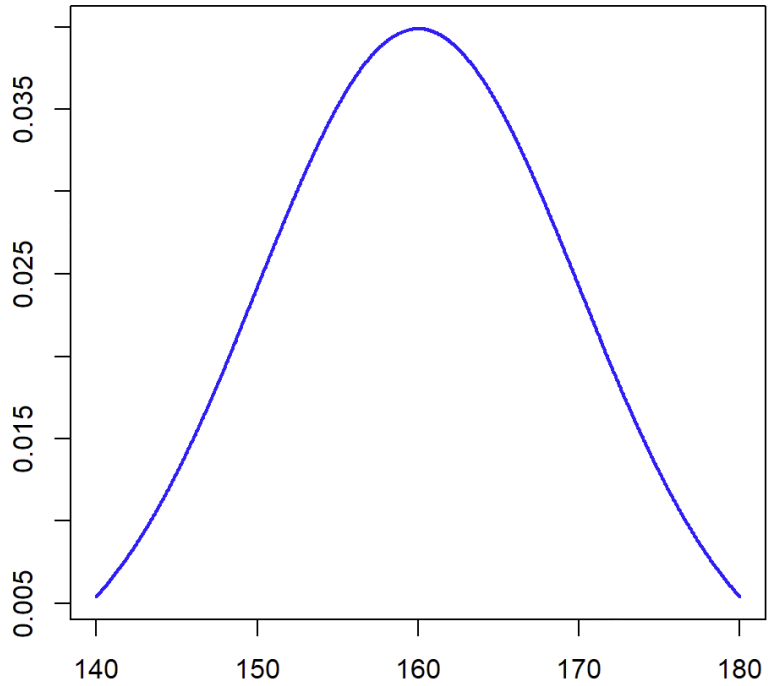
$$P(\mathbf{160} \leq x \leq \mathbf{170})$$

추론통계

확률 변수의 변수는 확률로 결정됨

- **확률의 종류는 이산확률변수, 연속확률변수 이를 표현하는 것은 이산확률분포 연속확률분포**
- **표본 공간에서 얼마만큼의 질량을 가지고 있는가를 궁금함**
- **이산인 경우 y축은 질량, 연속인 경우 y축은 밀도 부피는 구간**





추론통계

계산이 되는 데이터에 대해서 가능함

- 수치형 데이터에 주로 사용됨
- 왜 데이터 분석을 할까?? → 세상의 모든 데이터를 알 수 없음
- 세상의 모든 데이터를 바로 알 수 있으면 분석이 필요 없음
- 모든 데이터를 알 수 없고, 우리는 표본을 구해야 함
- 현실 세계에서 모든 데이터를 수집하는 것은 불가능하기 때문에 우리는 표본을 통해 모집단에 대한 결론을 내리려고 함
- 확률밀도함수는 이러한 결론을 내릴 때 필요한 확률적 배경을 제공해 줌

세상의 모든 사과는 특정 범주안의 크기를 가진다

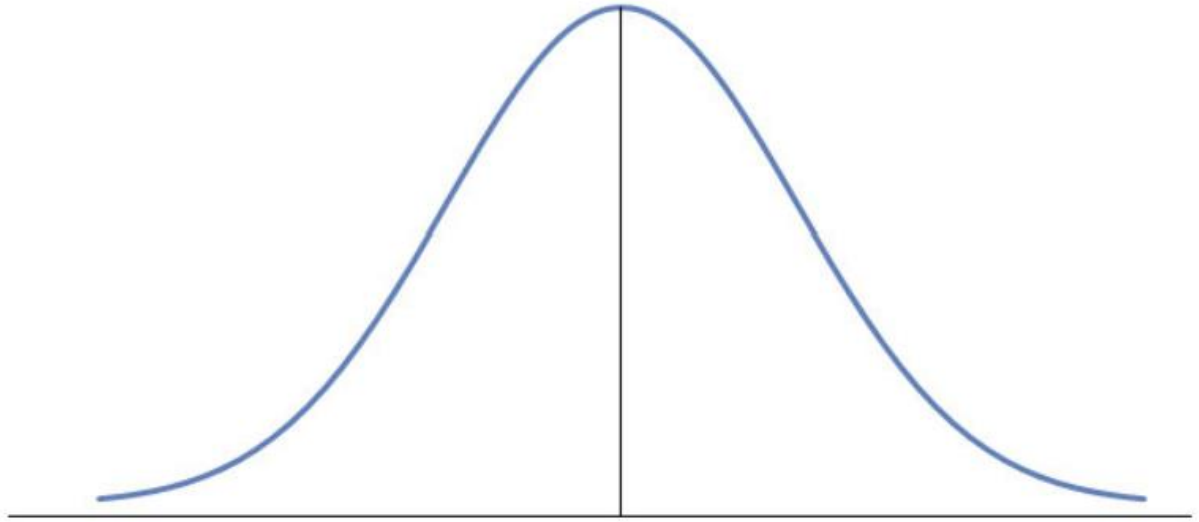
표본

모집단 예측

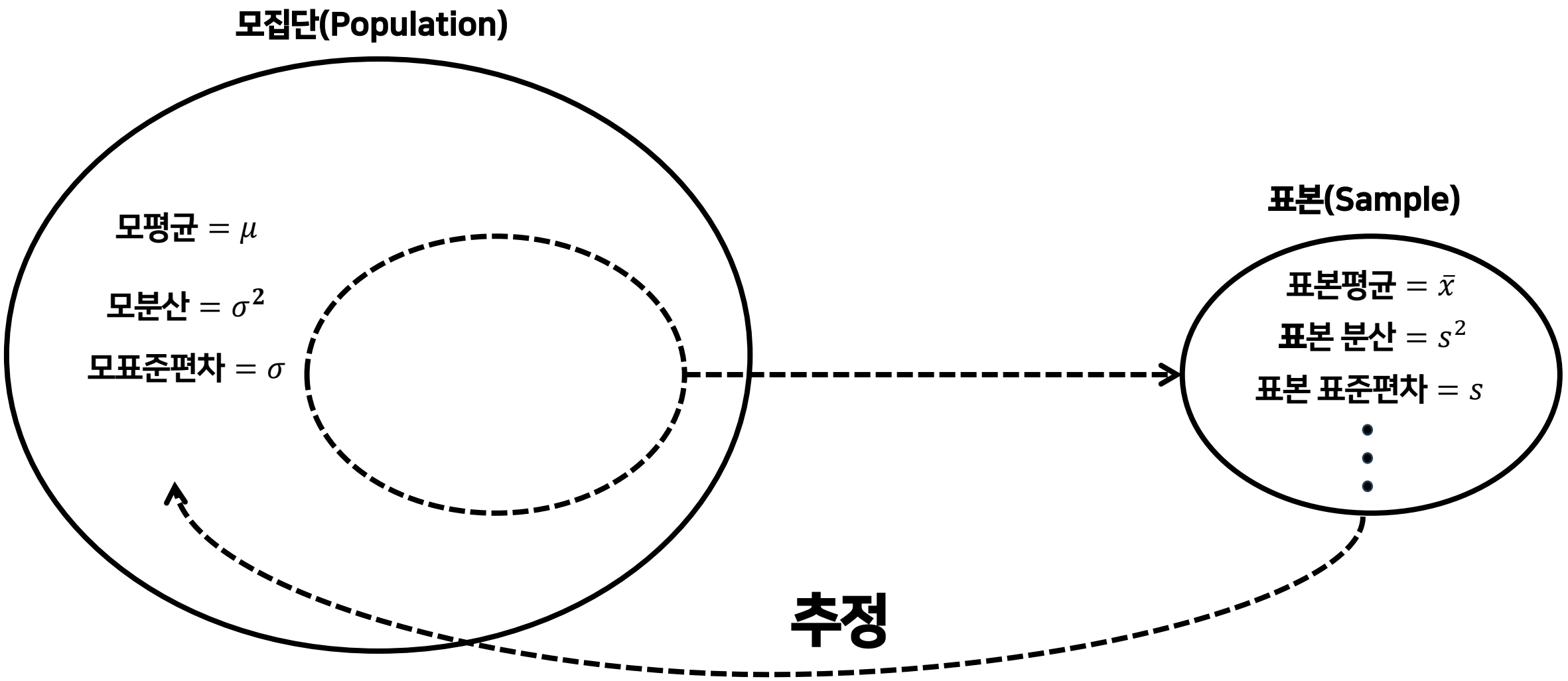
추론통계

계산이 되는 데이터에 대해서 가능함

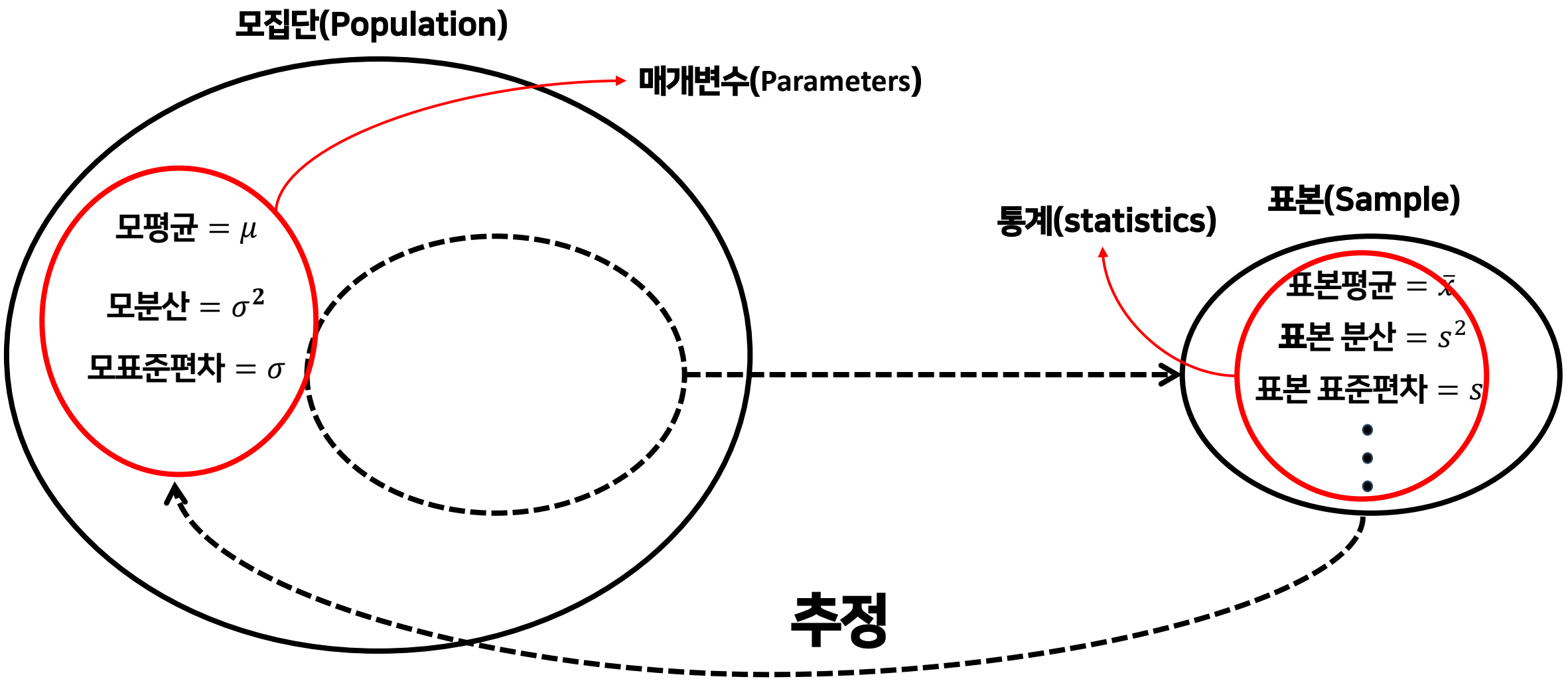
- 수치형 데이터에 주로 사용됨
- 중심 극한의 정리 : 표본의 크기가 커질수록 모집단의 분포와 상관없이 정규분포(Normal distribution)에 가까워진다는 것을 의미함
  - 표본의 크기(n≥30)는 평균의 샘플링 분포가 거의 정상
  - 모집단의 분산은 유한하고 알려져 있어야 함
  - 표본 관측치는 독립적이어야 함 → 하나의 관찰이 발생해도 다른 관찰의 발생에 영향을 미치지 않는다는 것을 의미



추론통계


모집단(Population)
모평균 $= \mu$
모분산 $= \sigma^2$
모표준편차 $= \sigma$
표본(Sample)
표본평균 $= \bar{x}$
표본 분산 $= s^2$
표본 표준편차 $= s$
추정

추론통계


모집단(Population)
매개변수(Parameters)
모평균 $= \mu$
모분산 $= \sigma^2$
모표준편차 $= \sigma$
통계(statistics)
표본(Sample)
표본평균 $= \bar{x}$
표본 분산 $= s^2$
표본 표준편차 $= s$
추정

추론통계

통계량

- 샘플들의 데이터로 통계량(Statistic)을 구할 수 있음
- 자유도(Degrees of freedom) : 매개변수를 추정하는 데 사용할 수 있는 독립적인 정보의 수를 반영함

$$Statistic = f(x_1, x_2, x_3, x_4, \dots, x_n)$$

표본집단 $\bar{X} = \dfrac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$ $\quad s^2 = \dfrac{\sum(x_i - \bar{X})^2}{n-1}$

모집단 $\mu = \dfrac{x_1 + x_2 + x_3 + \cdots + x_n}{N}$ $\quad \sigma^2 = \dfrac{\sum(x_i - \mu)^2}{N}$

통계량의 분포(Distribution of a Statistic) → 표본분포(Sampling Distribution)

- $\bar{X}$의 분포는 무엇인가? ➔ 평균의 기대값

$$\bar{X} = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}$$

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

$$E[\bar{X}] = \frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}$$
$$= \frac{1}{n}E[X_1] + \frac{1}{n}E[X_2] + \ldots + \frac{1}{n}E[X_n]$$
$$= \frac{1}{n}\mu + \frac{1}{n}\mu + \ldots + \frac{1}{n}\mu$$
$$= \mu$$

- s^2의 분포는 무엇인가? ➔ 분산의 기대값

$$V(X) = E[(X - \mu)^2]$$

$$s^2 = \frac{\sum(x_i - \bar{X})^2}{n - 1}$$

$$E[\bar{V}] = V\left(\frac{X_1 + X_2 + X_3 + \cdots + X_n}{n}\right)$$
$$= \frac{1}{n^2}\sigma^2 + \frac{1}{n^2}\sigma^2 + \frac{1}{n^2}\sigma^2 + \cdots +$$
$$= n\frac{1}{n^2}\sigma^2$$
$$= \frac{\sigma^2}{n}$$

추론통계

	모집단 (Population)	표본집단 (Sample)
평균 (Mean)	$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i$	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$
분산 (Variance)	$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2$	$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
표준편차 (Standard Deviation)	$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(X_i - \mu)^2}$	$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$

추론통계

통계량의 분포(Distribution of a Statistic) → 표본분포(Sampling Distribution)

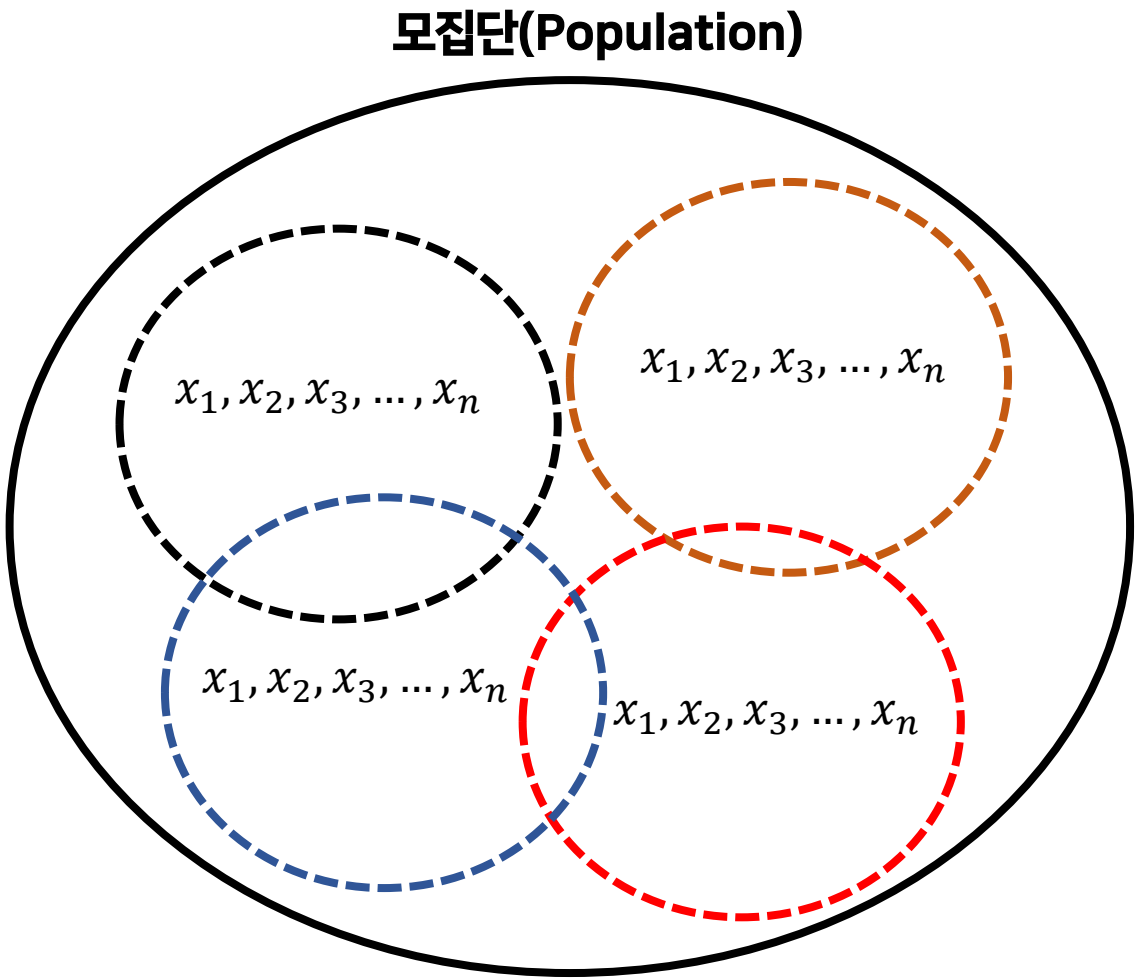
- 모집단의 분포가 정규분포를 따를 때 $\mathrm{N}(\mu, \sigma^2)$
- $X_1 + X_2 + X_3 \ldots, X_n$ 은 i.i.d. $\mathrm{N}(\mu, \sigma^2)$ → independent, identically, distributed
- $\mathrm{E}[\bar{X}] = \mu, V[\bar{X}] = \frac{\sigma^2}{n}$ ➔ **확률변수의 모집단의 평균, 모집단의 분산**
- 표본분포는? $\mathrm{N}(\mu, \sigma^2/n)$

검정통계량 : 모집단 매개변수에 대한 추론이나 결정을 내리기 위해 표본자료로부터 계산된 수치

- **표준정규분포(Standard Normal Distribution)로 변환** $\mathrm{Z} = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}}$
- **표준 정규분포를 따름** $Z \sim N(0,1)$

추론통계

임의의 모집단에서 표본의 크기가 n이 크면(n≥30), 표본평균 $\bar{X}$는 근사적으로 정규분포를 따름



$$\overline{X_1}, \overline{X_2}, \dots \overline{X_m}$$

표본분포 Sampling distribution

표준정규분포(Standard Normal Distribution)

$$\mathrm{N}(\mu, \sigma^2/n) \Rightarrow Z = \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \Rightarrow Z \sim N(0,1)$$

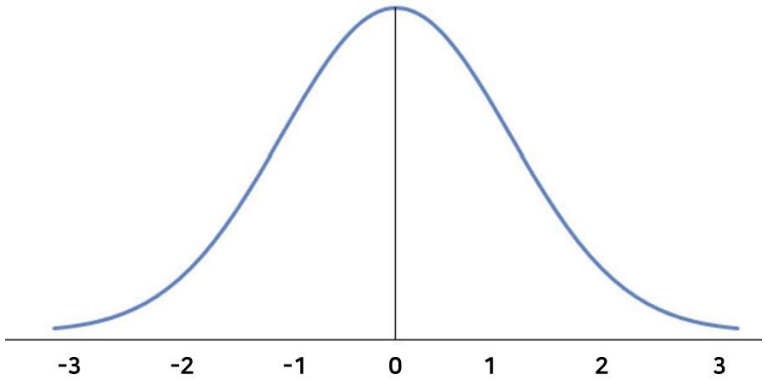
표본평균도 정규분포를 따름
→ 검정 통계량 Z: 정규 분포의 평균에서 얼마나 많은 표준 편차를 벗어났는지 계산하는 데 사용됨

추론통계

표준정규분포(Standard Normal Distribution)

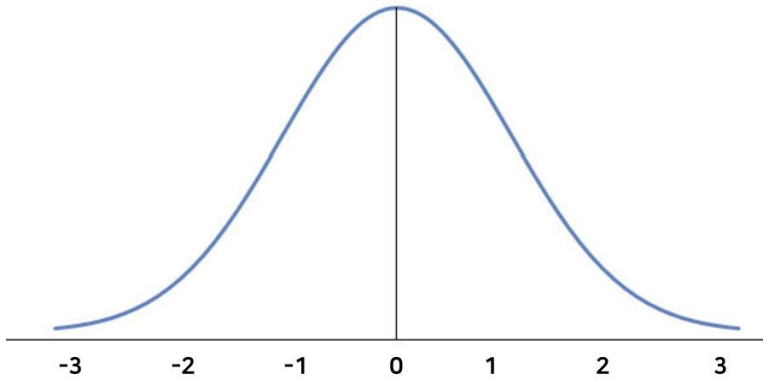
- 평균(μ)이 0이고 표준 편차(σ)가 1인 특정 유형의 정규 분포

Normal Distribution



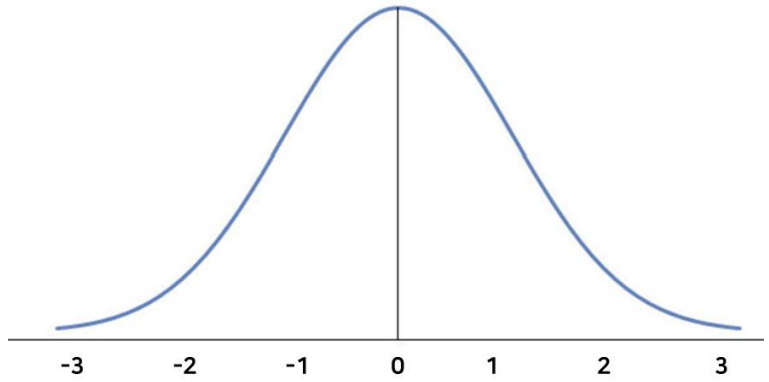
정규 분포는 항상 종 모양

Sampling Distribution



세상에 존재하는 실제 데
이터의 형태에 가까움

Standard Normal Distribution



항상 종 모양

추론통계

Standard Normal Distribution

예제 1)

모든 사람들의 키의 평균이 160cm이고 표준편차가 7cm인 정규분포를 따른다는 것이 알려져 있다고 가정할 때, 무작위 10명의 평균 키가 157cm 미만일 확률은 얼마입니까?

n **=10**

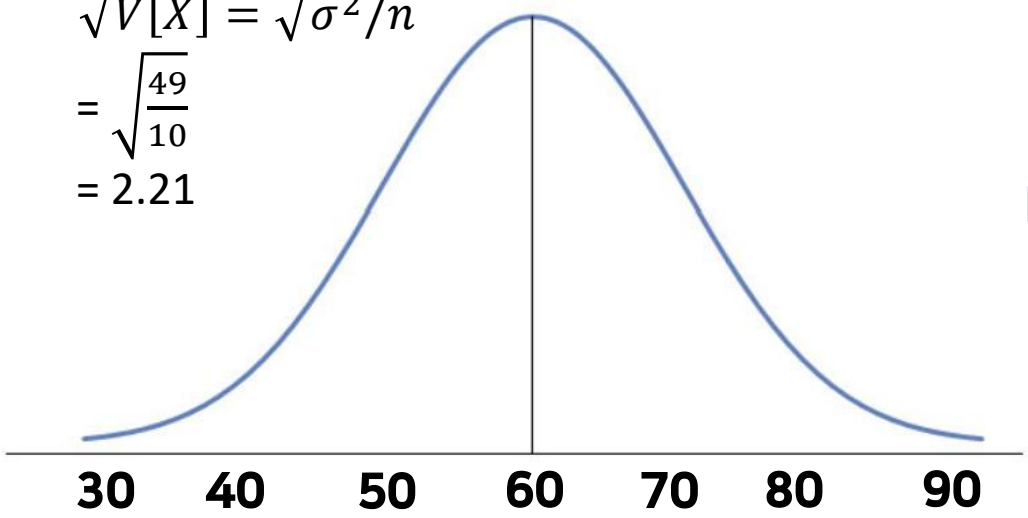
$\mathbf{E}[\bar{X}] = \mathbf{160}$

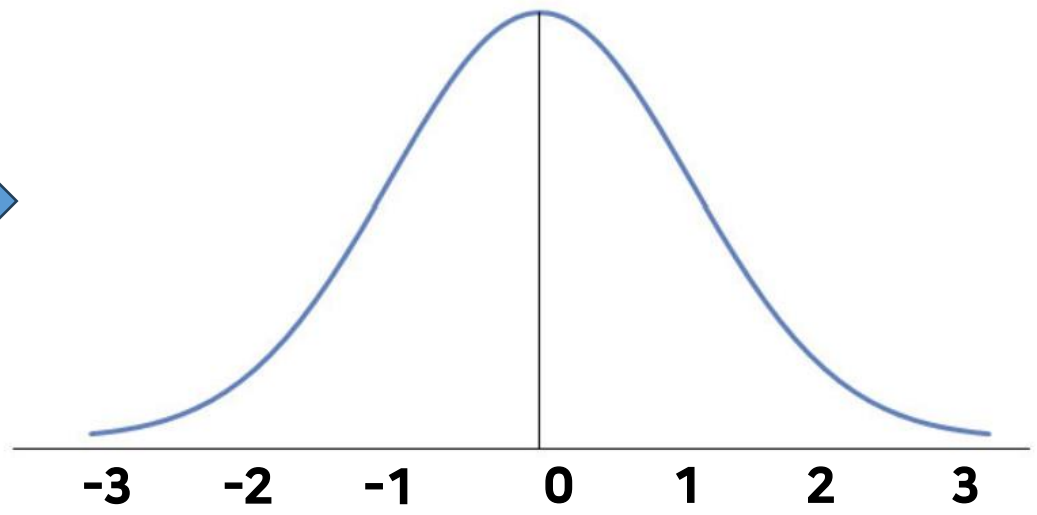
$\sqrt{V[\bar{X}]} = \sqrt{\sigma^2/n}$

$= \sqrt{\frac{49}{10}}$

$= 2.21$

$$\mathbf{Z} = \frac{\bar{\boldsymbol{X}} - \mu}{\sqrt{\sigma^2/n}} \qquad \mathbf{Z} = \frac{157 - 160}{2.21} = -1.36 \quad P(\mathbf{Z} < -1.36) = 0.0869$$





Standard Normal Distribution

예제 2)

25개의 샘플을 뽑았을 때, 모집단의 평균이 15, 분산이 100일 때, 표본평균이 20보다 작을 확률은? (P[$\bar{X} \leq 20$])

P[$\bar{X} \leq 20$] = P[$Z \leq 2.5$] = 0.9938

Standard Normal Distribution

예제 3)

모집단이 정규분포를 따를 때, 모집단의 평균이 25, 표준편차가 12일 때, 이때, A라는 35개의 표본 집단을 추출했을 때, 표본평균이 15보다 작을 확률은?

예제 4)

짐의 무게는 모평균이 18kg이고, 표준편차가 3kg인 정규분포를 따른다. 36명의 짐을 임의로 추출할 때, 짐의 평균 무게가 17kg 이상일 확률은?

Standard Normal Distribution

예제 5)

강원도 고등학생들의 평균 점수는 85이고, 표준편차가 10일 경우 해당 고등학생들 중 30명을 임의로 뽑았을 때, +5점 이상 차이가 날 확률을 구하여라.

예제 6)

강원도 고등학생들의 평균 점수는 85이고, 표준편차가 10일 경우 해당 고등학생들 중 30명을 임의로 뽑았을 때, +5점 이상 +8점 이하의 차이가 날 확률을 구하여라.

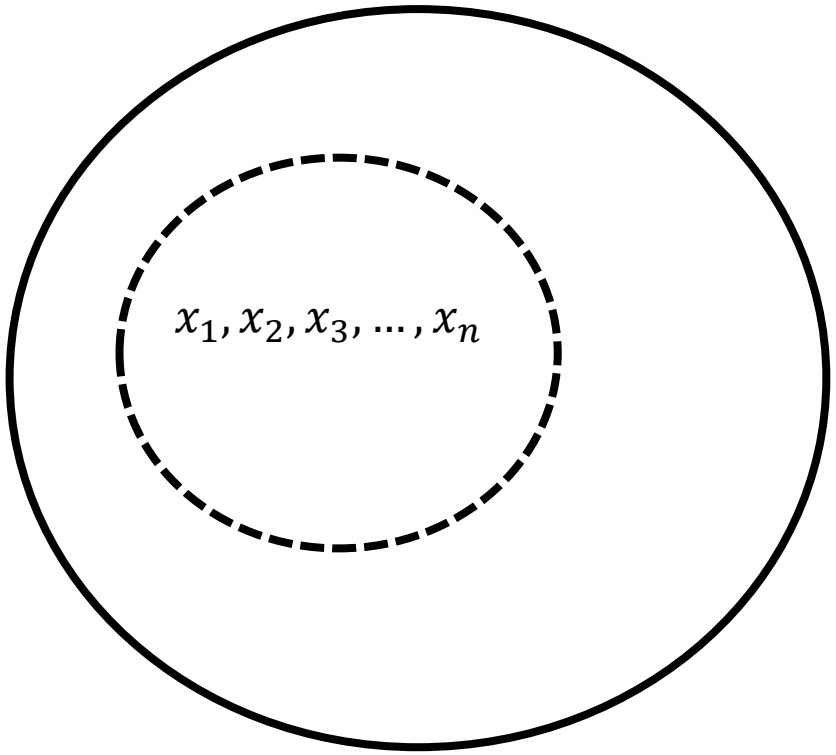
예제 7)

강원대학교 학생들의 평균 키는 175cm이고, 표준편차가 15인 정규분포를 따른다. 이때, 임의로 강원대학교 학생들을 49명을 선택할 때, 해당 학생들의 키의 평균이 173cm에서 178cm 사이일 확률은?

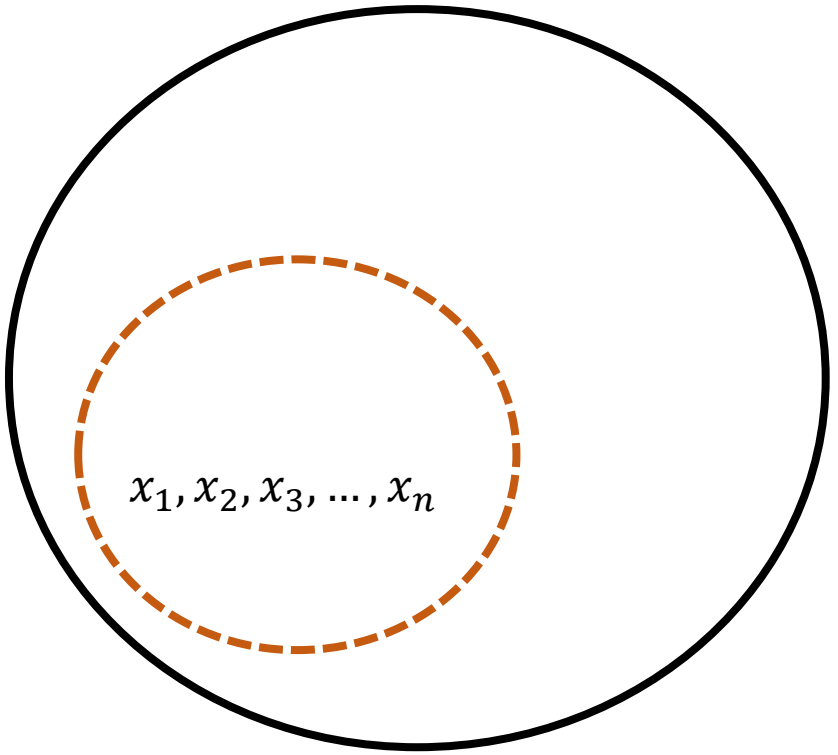
추론통계

모집단이 두개일 때?

1. 모집단(Population)



2. 모집단(Population)



$$\mathbf{E}[\overline{X_1} - \overline{X_2}] = \mathbf{E}[\overline{X_1}] - \mathbf{E}[\overline{X_2}] = \mu_1 - \mu_2$$

$$\boldsymbol{V}[\overline{X_1} - \overline{X_2}] = \boldsymbol{V}[\overline{X_1}] + \boldsymbol{V}[\overline{X_2}] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$\mathbf{N}(\mu, \sigma^2/n)$ $\quad Z = \dfrac{\overline{\boldsymbol{X}} - \mu}{\sqrt{\sigma^2/n}}$ $\qquad \mathbf{N}(\mu_1 - \mu_2, \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2})$ $\qquad Z = \dfrac{\overline{X_1} - \overline{X_2}}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}}$

추론통계

모집단이 두개일 때?

예제 1)

첫번째 모집단에서 샘플 64개의 평균이 4, 분산 16, 두번째 모집단에서 샘플 75개의 평균이 12, 분산 48 일 때, 두개의 샘플간의 평균의 차이가 -6보다 작을 확률을 구해라

$$\mathbf{P}[\overline{X_1} - \overline{X_2} \leq -6] = \mathbf{P}[\mathbf{Z} \leq -2.12] = 0.017$$

$$\mathbf{N}(\mu_1 - \mu_2,\ \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}) \qquad Z = \frac{4 - 12}{\sqrt{\frac{16}{64} + \frac{48}{75}}} = \frac{-8}{0.9434} = -8.48$$

$$Z = \frac{4 - 12 + 6}{\sqrt{\frac{16}{64} + \frac{48}{75}}} = \frac{-8 + 6}{0.9434} = -2.12$$

추론통계

모집단이 두개일 때?

예제 2)

평균이 100이고 분산이 25인 첫 번째 모집단에서 50개의 표본이 있고, 평균이 200이고 분산이 45인 두 번째 모집단에서 60개의 표본이 있다고 가정할 때, 두 표본 간의 평균 차이가 나는지를 판단해라

01 기초통계

추론통계

추론통계 : 샘플의 데이터를 기반으로 더 큰 모집단에 대해 예측하거나 결론을 도출할 수 있음

- 목적 : 추론 통계는 의미 있는 결론을 도출하고 샘플 데이터를 기반으로 모집단에 대한 예측을 수행하는 역할을 함 이러한 방법을 통해 모집단 매개변수에 대한 정보에 입각한 추측을 하고 가설을 테스트하며 결과의 신뢰성을 평가할 수 있음
- 측정 유형 : 추론 통계는 범주형 측정과 연속 측정 모두에 적용가능 데이터의 특성과 연구 질문에 적응하여 다양한 분야와 시나리오에 적용가능

- 표현 : 추론 통계는 수학 공식과 확률 모델을 활용하여 모집단 매개변수를 추정하거나 관찰된 결과의 가능성을 결정함 test values, 신뢰 구간, p-값 이러한 표현을 통해 불확실성을 정량화하고 증거를 기반으로 결정을 내림
- 사용 : 추론 통계는 사회 과학에서 자연 과학, 비즈니스 및 그 이상에 이르기까지 수많은 영역에서 사용됨
  - 이러한 기술을 사용하여 가설을 검증하고, 그룹을 비교하고, 변수 간의 관계를 평가하고, 미래 추세를 예측
  - 의사 결정, 정책 수립 및 데이터 내에 숨겨진 패턴 발견 및 추론 통계는 관찰된 샘플의 범위를 넘어서는 더 광범위한 의미를 갖는 통찰력을 발견할 수 있는 수단을 제공

추론통계

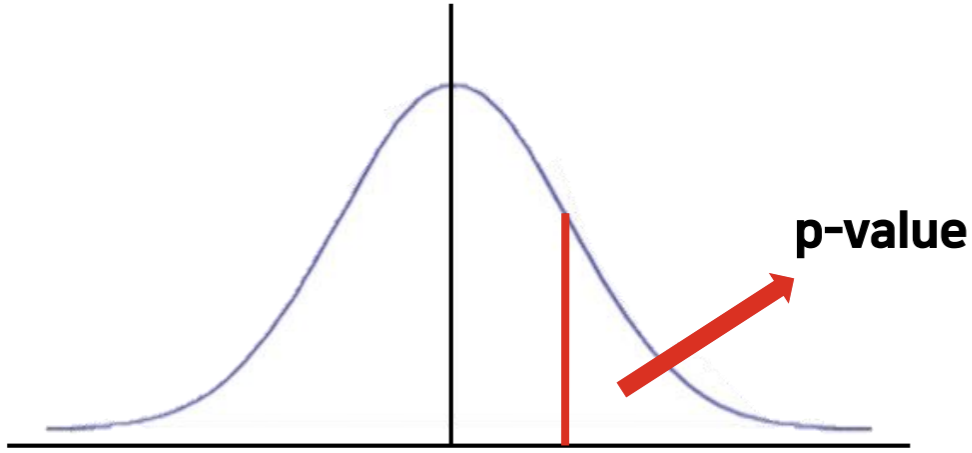
가설 검정 및 추론통계

- 데이터의 구조를 뜯어봤고, 이 데이터를 실제로 사용할 필요가 있는지 없는지를 판단하기 위해 가설 검정
- 가설 설정은 유의한 효과, 관계 또는 차이가 있는지 여부를 결정하기 위한 가설 검정의 초기 단계
- 모델이나 테스트를 사용한 후속 통계 분석은 가설에 대한 증거를 평가하는 데 도움이 되며 관찰된 데이터를 기반으로 연구 중인 모집단에 대한 구체적인 판단으로 이어짐
- 가설 검정에서는 표본(모집단의 하위 집합)을 가설 값이나 다른 표본과 비교하여 모집단에 대한 결론을 도출함
- 세상의 모든 데이터를 모집단으로 수집할 수 있으면 가설 검정을 진행할 필요가 없음

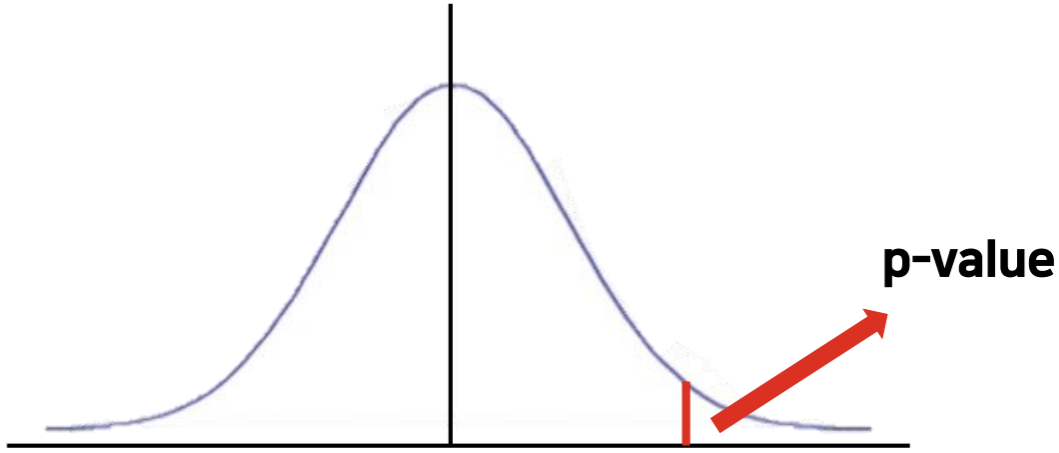
추론통계

가설 검정 및 추론통계

- 가설 검정에서 귀무가설과 대립가설로 이루어 짐
- 귀무가설(H_0) : 현상 유지 또는 영향이 없다는 가정
- 대립가설(H_1) : 귀무가설과는 반대로 효과가 있는 상황을 나타냄
- 귀무가설 내주장과 반대되는 가설 (H_0) : A약과 B약은 집중력 향상에 차이가 없다.
- 대립가설 내주장에 대한 가설 (H_1) : A약과 B약은 집중력 향상에 차이가 있다.



가까우면 귀무가설을 기각할 충분한 근거가 없음



멀면 귀무가설을 기각할 충분한 근거가 있음

추론통계

가설 검정 및 추론통계

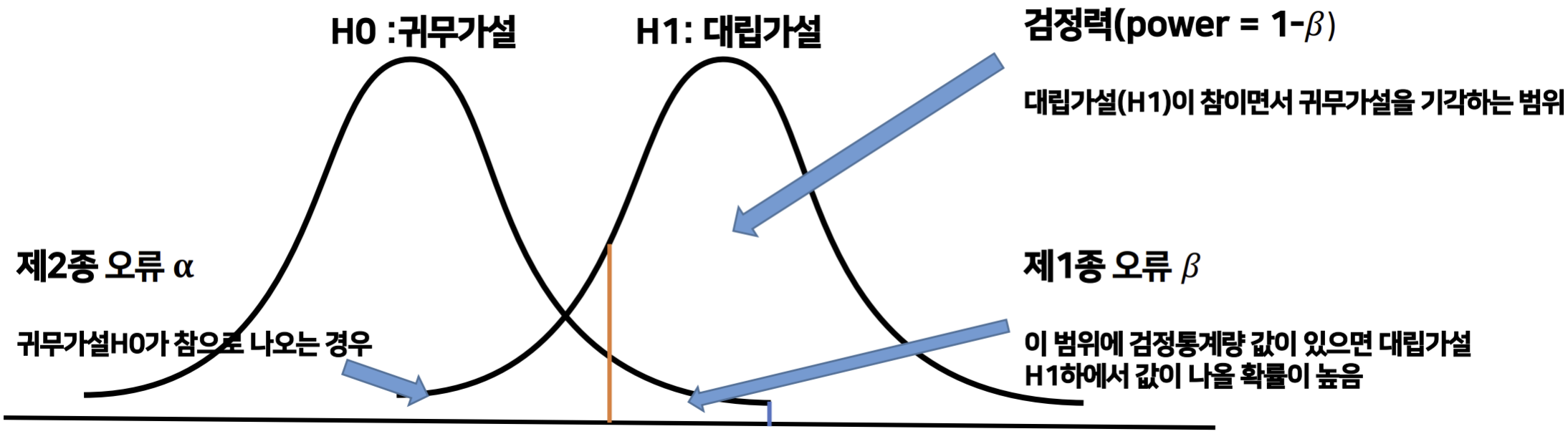
- p-value가 크다 작다의 기준을 세워야 하고, 1,2종 오류에 대해 살펴보면
- P-value가 1종 오류 확률의 최대 허용치, 즉 유의 수준(α) 보다 작으면 귀무가설을 기각 반대의 경우 채택
- 1종 오류의 최대 허용치가 바로 유의 수준(α)으로 p-value의 크기를 판단할 기준점

귀무가설 H0 진위	귀무가설 채택	귀무가설 기각
귀무가설 사실	올바른 의사 결정 확률 (1-α)	제 1종 오류 확률 α (유의수준)
귀무가설 거짓	제 2종 오류 확률 (β)	올바른 의사 결정 확률 1-β (검정력)

추론통계

가설 검정 및 추론통계

- 옳은 결정(1-β)이 검정력 이라고 함 → 얼마나 귀무가설이 잘못되었는가?
- 평균의 차이, 표본의 크기, 유의수준의 크기, 양방향 검정, 일방향 검정, 검사의 신뢰도
- p-value가 작을 수록 회귀계수가 0이라는 귀무가설을 기각함 → 검정력은 잘못된 귀무가설을 올바르게 기각할 확률



추론통계

신뢰구간

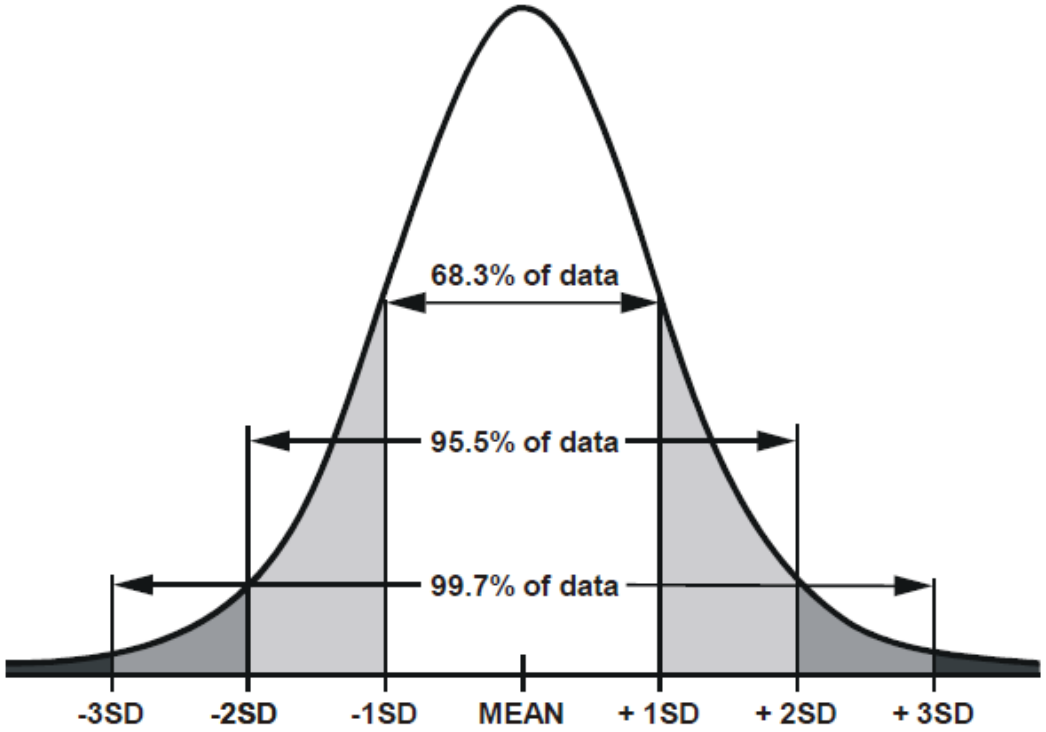
- α 귀무가설이 참일 때 귀무가설을 기각할 확률
- 모집단이 정규분포를 따를 때, μ의 100(1-α)% 신뢰구간 → 연구자가 정하는 것(0.05,0.01, 0.1)
- 제1종 오류를 범할 확률 5%, 1%, 10%
- 데이터의 약 68.27%가 평균(μ =0)에서 표준편차(σ =1) 범위, $-1 \le x \le 1$ 범위 내에 속함
- 약 95.45%는 평균에서 2표준편차(σ =1)범위 $(-2 \le x \le 2)$내에 속함
- 약 99.73%는 평균에서 3표준편차(σ =1)범위 $(-3 \le x \le 3)$내에 속함

$$[\bar{X} - z\frac{\alpha}{2}\frac{\sigma}{\sqrt{n}},$$

신뢰하한

$$\bar{X} + z\frac{\alpha}{2}\frac{\sigma}{\sqrt{n}}]$$

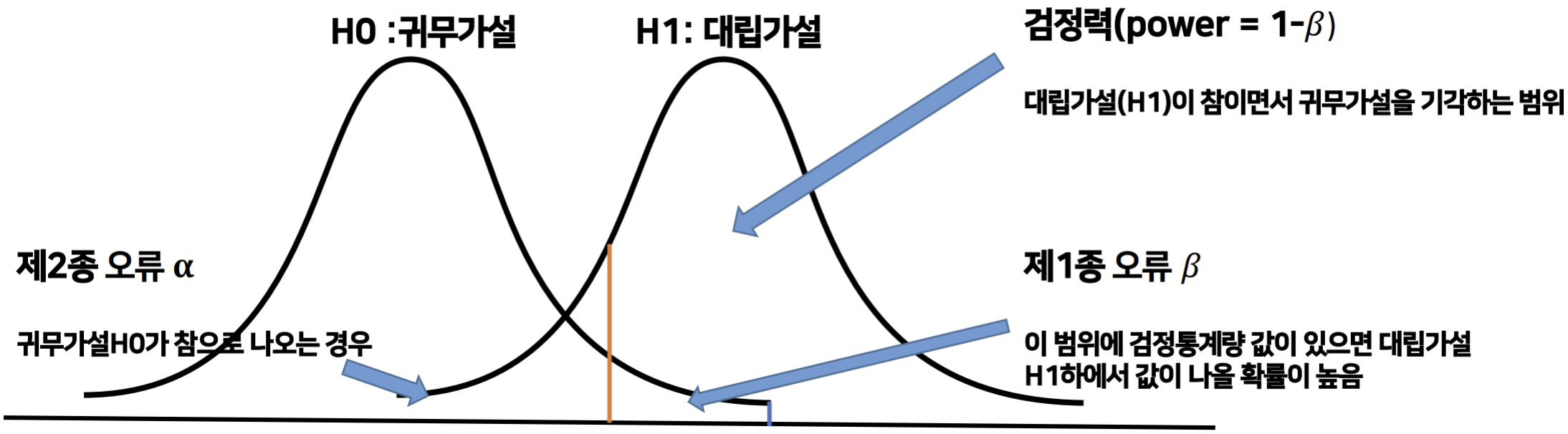
신뢰상한



추론통계

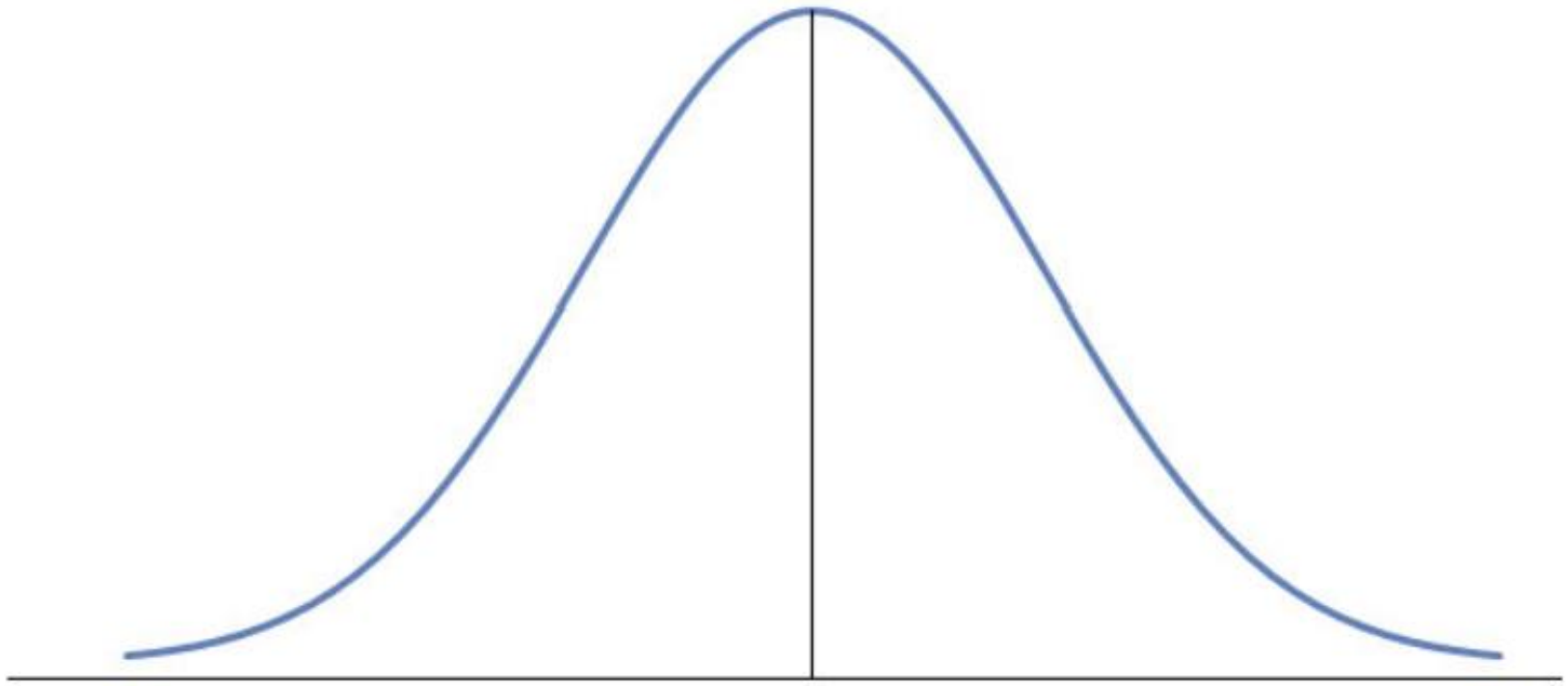
가설 검정 및 추론통계

- 효과 크기: 모집단 또는 그룹 간의 실제 차이를 나타냄
- 효과 크기가 클수록 귀무가설 분포와 대립 가설 분포가 더 많이 분리됨
- 표본 크기: 표본 크기가 클수록 일반적으로 분포가 좁아져(즉, 표준 오차가 감소) 1,2종 오류가 모두 감소함
- 변동성: 데이터의 변동성(또는 표준 편차)이 크면 분포가 넓어져 분포가 더 많이 겹치고 유형 II 오류의 가능성이 높아짐



가설 검정 및 추론통계

- 유의확률(p-value) : 가설에 대한 주장을 얼마나 정확하게 주장할 수 있는지에 대한 기준
- 제1종 오류를 범할 확률 → 귀무가설을 채택할 확률



추론통계

가설 검정 및 추론통계

- **p-value : 가설에 대한 주장을 얼마나 정확하게 주장할 수 있는지에 대한 기준**

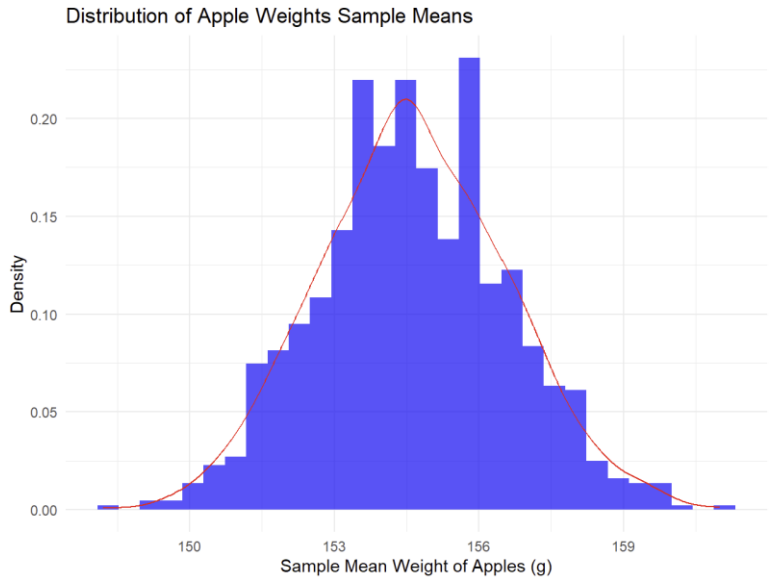
```
# 사과의 무게
Apple_weights <- c(150, 152, 147, 160, 155, 153, 158, 162, 149, 154, 159, 157)
# 샘플사이즈와 시행 횟수
sample_size <- 5
num_samples <- 1000

get_sample_mean <- function() {
    sample <- sample(apple_weights, sample_size, replace = TRUE)
    return(mean(sample))
}

sample_means <- replicate(num_samples, get_sample_mean())
```

추론통계

가설 검정 및 추론통계

- p-value : 가설에 대한 주장을 얼마나 정확하게 주장할 수 있는지에 대한 기준

```
ggplot(data.frame(sample_means = sample_means), aes(sample_means)) +
geom_histogram(aes(y=..density..), bins = 30, fill="blue", alpha=0.7) +
geom_density(color="red") +   theme_minimal() +  ggtitle("Distribution of Apple
Weights Sample Means") +  xlab("Sample Mean Weight of Apples (g)") +  ylab("Density")
```



추론통계

가설 검정 및 추론통계

- Z-test : Z-test는 t-test와 유사하지만 표본 크기가 크고(일반적으로 n ≥ 30) 모집단 표준 편차를 알고 있을 때 사용됨
- 그룹의 평균이 가설 값과 유의하게 다른지 테스트함
- 과거의 경험, 많은 샘플수로 모집단을 예측할 수 있으므로 모집단의 표준편차를 알고 있다고 할 수 있음
- 표준정규분포의 평균의 분포 ➔ 0을 기준으로 정규분포를 이룸

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

μ : 모집단의 평균
σ : 모집단의 표준편차

$$Z = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$$t = \frac{\bar{X} - \mu}{\frac{s}{\sqrt{n}}} \sim t_{(df=n-1)}$$

추론통계

가설 검정 및 추론통계(차이 파악)

- t-test : t-test는 두 그룹의 평균을 비교하거나 **모집단 표준 편차를 모를 때** 단일 그룹의 평균의 차이를 테스트하는 데 사용
  - 데이터가 대략적으로 정규분포를 이루고 표본크기가 작은 경우에 적용할 수 있음($n \leq 30$) → 평균의 분포

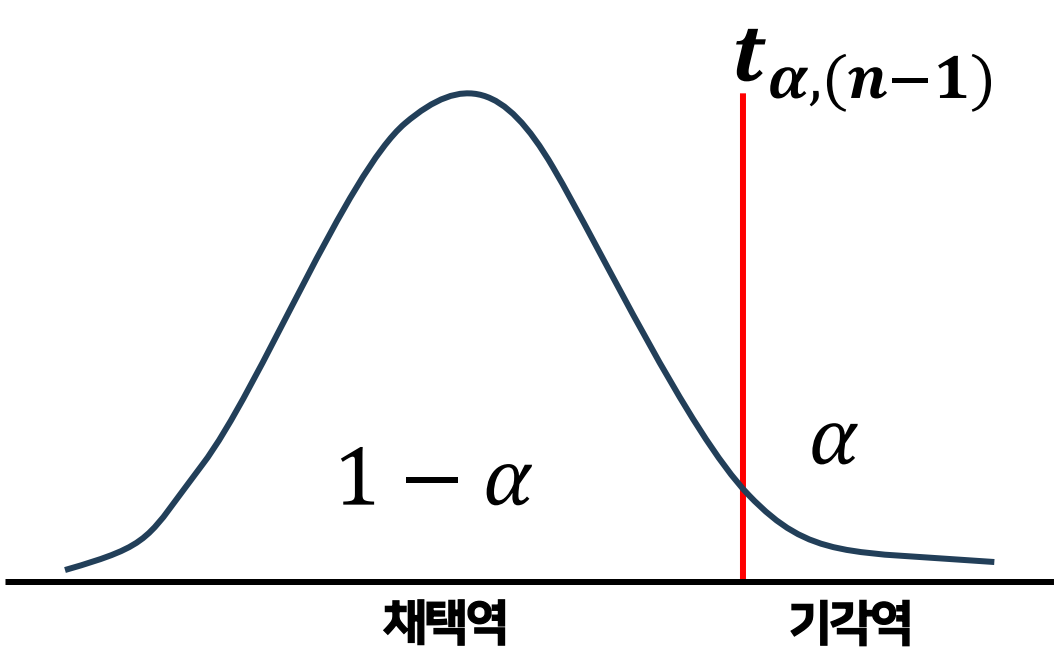
- Z-test : Z-test는 t-test와 유사하지만 표본 크기가 크고(일반적으로 $n > 30$) **모집단 표준 편차를 알고 있을 때** 사용
  - 그룹의 평균이 가설 값과 유의하게 다른지 테스트함 → 평균의 분포

추론통계

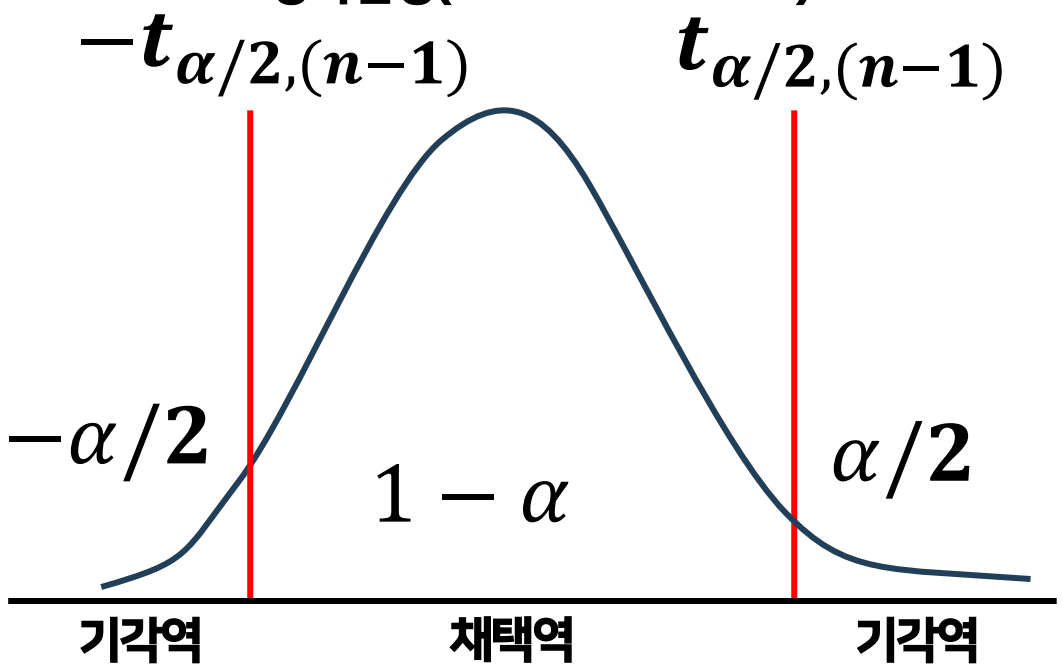
가설 검정 및 추론통계

- t-test : t-test는 두 그룹의 평균을 비교하거나 단일 그룹의 평균이 가설 값과 유의하게 다른지 테스트하는 데 사용됨
- 데이터가 대략적으로 정규분포를 이루고 표본크기가 작은 경우에 적용할 수 있음(n≤30)
- 모집단을 대표하는 표본으로부터 추정된 분산이나 표준편차를 가지고 검정하는 방법으로 두 모집단의 평균간의 차이를 검정

단측검정(one-tailed test)



양측검정(two-tailed test)



추론통계

t-test의 특징

- 관찰의 독립성: 비교되는 두 그룹의 데이터는 서로 독립적이어야 함
- 정규성: 비교되는 두 그룹 각각의 데이터는 대략 정규 분포를 따라야 함 & 무작위로 샘플링이 가능해야 함
- 이상값 없음: 데이터에 극단값이나 이상값이 없어야 함
- 연속성 : 수치값을 가지거나 연속적이여야 함
- 이 비율이 1에 가까우면 등분산(예: 0.5와 2 사이)

$$t = \frac{\overline{X_1} - \overline{X_2}}{s_p \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$$

두집단의 분산이 같을 때(등분산 성립)
0.5와 2사이(s_p : 합동 표준편차)

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

두집단의 분산이 같지 않을 때 (등분산 성립X)

추론통계

t-test

- 자유도(df) : 매개변수를 추정하는 데 사용할 수 있는 독립적인 정보의 수를 반영함
- 표본분산 : 모집단에서 추출한 여러 가능한 표본에 대한 통계의 변동성 또는 분산

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

$$s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}}$$

$$t = \frac{\overline{X_1} - \overline{X_2}}{s_p \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

등분산가정(O)

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

등분산가정(X)

t-test

- 평균 차이 검증 → 등분산가정(X)

A방법	B방법
85	78
90	80
88	77
83	82
87	79
89	75

$s_1^2 = \{(85-88.67)^2 + (90-88.67)^2 + \ldots + (89-88.67)^2\} / (6-1) = 6.8$
$s_2^2 = \{(78-78.50)^2 + (80-78.50)^2 + \ldots + (75-78.50)^2\} / (6-1) = 5.9$

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

$$t = \frac{87 - 78.5}{\sqrt{\frac{6.8}{6} + \frac{5.9}{6}}}$$

$t \approx 5.84$

$$df = n_1 + n_2 - 2 = 6 + 6 - 2 = 10$$

t-test

- 평균 차이 검증 → 등분산가정(O)

A방법	B방법
85	78
90	80
88	77
83	82
87	79
89	75

$s_1^2 = \{(85-88.67)^2 + (90-88.67)^2 + \ldots + (89-88.67)^2\} / (6-1) = 6.8$
$s_2^2 = \{(78-78.50)^2 + (80-78.50)^2 + \ldots + (75-78.50)^2\} / (6-1) = 5.9$

$$s_p = \sqrt{\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}} \quad t = \frac{\overline{X_1} - \overline{X_2}}{s_p \times \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$t \approx 4.26$

$\mathrm{df} = n_1 + n_2 - 2 = 6+6-2 = 10$

t-test

- **평균 차이 검증**

```
group_a <- c(85, 88, 90, 92, 91, 87, 89, 86, 84, 83)
group_b <- c(78, 82, 80, 85, 84, 87, 83, 81, 80, 79)

mean_a <- mean(group_a)
mean_b <- mean(group_b)
sd_a <- sd(group_a)
sd_b <- sd(group_b)
t_statistic <- (mean_a - mean_b) / sqrt((sd_a^2 / length(group_a)) +
(sd_b^2 / length(group_b)))

# 자유도
df <- length(group_a) + length(group_b) - 2
```

t-test

- **평균 차이 검증**

```
group_a <- c(85, 88, 90, 92, 91, 87, 89, 86, 84, 83)
group_b <- c(78, 82, 80, 85, 84, 87, 83, 81, 80, 79)

#집단 a의 평균과 집단 b의 평균의 차이가 존재한다
t_test_result <- t.test(group_a, group_b, alternative = “two.sided")

#집단 a의 평균이 집단 b보다 작다
t_test_result <- t.test(group_a, group_b, alternative = “less")

#집단 a의 평균이 집단 b보다 크다
t_test_result <- t.test(group_a, group_b, alternative = “greater")
```

Z-test

- **평균 차이 검증**

```
library(BSDA)

group_a <- c(85, 88, 90, 92, 91, 87, 89, 86, 84, 83, 85, 88, 90, 92, 91, 87, 89, 86, 84, 83,
85, 88, 90, 92, 91, 87, 89, 86, 84, 83, 85, 88, 90, 92, 91, 87, 89, 86, 84, 83)
group_b <- c(78, 82, 80, 85, 84, 87, 83, 81, 80, 79, 78, 82, 80, 85, 84, 87, 83, 81, 80, 79,
78, 82, 80, 85, 84, 87, 83, 81, 80, 79, 78, 82, 80, 85, 84, 87, 83, 81, 80, 79)

#표준편차
sd_a = sd(group_a)
sd_b = sd(group_b)
# Z-test
result <- z.test(x = group_a, y = group_b, sigma.x = sd_a, sigma.y = sd_b, alternative = "
two.sided")

t_test_result <- t.test(group_a, group_b, alternative = “two.sided")
```

추론통계

가설 검정 및 추론통계

- 모집단 표준 편차의 가정
  - Z-test : 모집단 표준 편차를 알고 있다고 가정
  - T-test : 모집단 표준 편차를 알 수 없다고 가정

- 샘플 크기
  - Z-test : 더 큰 샘플 크기($n \geq 30$)에서 잘 작동
  - T-test : 더 작은 샘플 크기와 더 큰 샘플 크기에서 잘 작동

- 테스트 통계 분포
  - Z-test : 테스트 통계는 표준 정규 분포(z-distribution)를 따름
  - T-test : 테스트 통계는 자유도(샘플 크기 관련)에 따라 달라지는 t-분포를 따름

- 테스트 통계량 계산
  - Z-test, T-test
  - 표본 평균과 모집단 평균의 차이를 비교
  - 표본의 변동성을 고려하여 표본 평균과 가설 평균의 차이(또는 쌍 표본 간의 차이)를 비교

추론통계

가설 검정 및 추론통계

- 가설 설정
- 유의수준 선택
- 자유도 계산
- 검정 통계량 계산
- P-value값 계산
- 결론 도출

문제

- **Score_2 Text파일을 csv파일로 변환하고, eng, math에 대한 Boxplot을 그리시오.**

```
a=read.csv("C:/Users/USER/Desktop/Score_2.csv")

df <- data.frame(
        group = c(rep("Group 1", length(a$eng)), rep("Group 2", length(a$math))),
        values = c(a$eng, a$math))

ggplot(df, aes(x = group, y = values)) +
  geom_boxplot(fill = c("lightblue","lightgreen"), outlier.color = "red") +
  labs(title = "Boxplot Example") +
  xlab("Group") +
  ylab("Values")
```

추론통계

문제

- 모집단이 정규분포를 따를 때, 모집단의 평균이 60, 표준편차가 10일 경우
  이 모집단에서 크기 40개의 표본 집단 B를 추출했을 때, 표본평균이 55보다 작을 확률은?

추론통계

문제

- 한 커피 전문점에서 판매하는 아메리카노 한 잔의 카페인 함량은 평균 150mg, 표준편차 20mg인 정규분포를 따른다. 이 커피숍의 품질관리팀은 당일 생산된 아메리카노 중 무작위로 36잔을 샘플링 하여 카페인 함량을 측정했는데, 표본 평균이 143mg이하일 확률은?