강원대학교
AI 소프트웨어학과

머신러닝2

– Intro –

KNU
강원대학교

통계란?

01 기초통계

통계란?

데이터란?

- 이론을 세우는 데 기초가 되는 사실, 또는 바탕이 되는 자료
- 관찰이나 실험, 조사로 얻은 사실이나 자료
- 컴퓨터가 처리할 수 있는 문자, 숫자, 소리, 그림 따위의 형태로 된 자료
- 데이터는 신호,기호,숫자,문자 등으로 기록 됨
- 정보를 위한 기초적인 자료를 말함
- 정보는 데이터를 가공하지 않은 경우

01 기초통계

통계란?

정보란 무엇일까?

- 정보란? → 구성, 해석 및 맥락화 과정을 통해 데이터에서 파생

선수들의 수치

PLAYER	DPM	GOLDDIFFAT15	분당 K+A	GPM	분당 골드 차이	분당 데미지 차이
FAKER	375	232.21	0.220	396	11.732	36.575
SHOWMAKER	488	82.86	0.294	404	19.901	31.411
CHOVY	466	352.44	0.223	410	27.059	-30.201
BDD	461	-1.41	0.269	389	4.989	53.180
GORI	459	-400.44	0.239	397	3.552	-29.395
FATE	472	236.74	0.238	406	21.573	-8.689

선수들의 신체 조건에 따른 적성 진단

운동선수-일반인 기초스포츠적성진단 비교

자료: 국민체육진흥공단 체육과학연구원

운동선수 (일반인 평균)	체지방률 (16.1%)	팔굽혀펴기 (60.6회)	농구공던지기 (661.1cm)	윗몸일으키기 (67.7회/2분)	제자리멀리뛰기 (224.7cm)	오래달리기 (465.5초/1600m)
역도	16.9	76.3	1010.7	66.3	285.0	480.9
야구	17.0	86.4	1004.2	88.0	268.2	357.3
축구	13.0	116.7	1025.6	97.2	258.8	311.0
핸드볼	12.2	108.6	1065.0	91.7	263.9	334.6
양궁	14.1	54.3	919.3	71.6	249.3	382.6
농구	18.2	61.7	1346.1	94.4	256.9	354.1
유도	15.2	118.5	926.1	82.7	246.7	347.8
수영	11.2	63.1	942.1	91.2	251.1	436.1
체조	8.8	65.2	745.3	96.1	233.9	459.6

통계란?

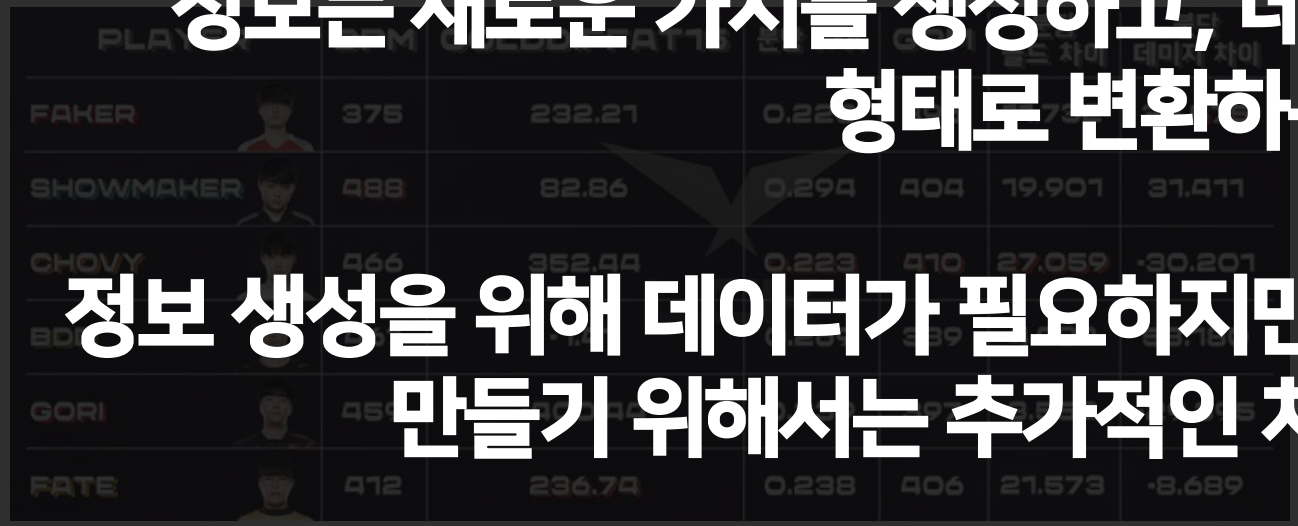
정보란 무엇일까?

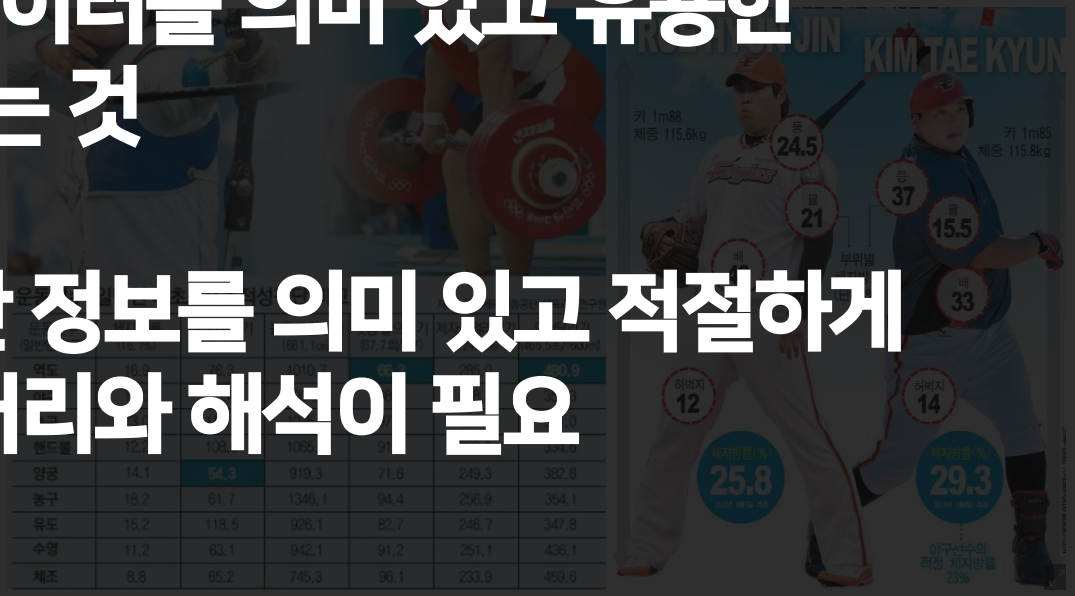
- 정보란? → 구성, 해석 및 [illegible]

데이터는 정보가 생성되는 원재료

정보는 새로운 가치를 생성하고, 데이터를 의미 있고 유용한 형태로 변환하는 것

정보 생성을 위해 데이터가 필요하지만 정보를 의미 있고 적절하게 만들기 위해서는 추가적인 처리와 해석이 필요





기술통계

중심 경향 측정

- 평균 : 데이터 세트에 있는 모든 데이터 포인트의 산술 평균
- 중앙값 : 데이터 세트에서 가장 작은 것부터 큰 순서로 정렬할 때 중간 값
- 최빈값 : 데이터 세트에서 가장 자주 발생하는 값
- 최대값/최소값 : 데이터 세트에서 가장 큰 값/ 데이터 세트에서 가장 작은 값

기술통계

변동성 측정

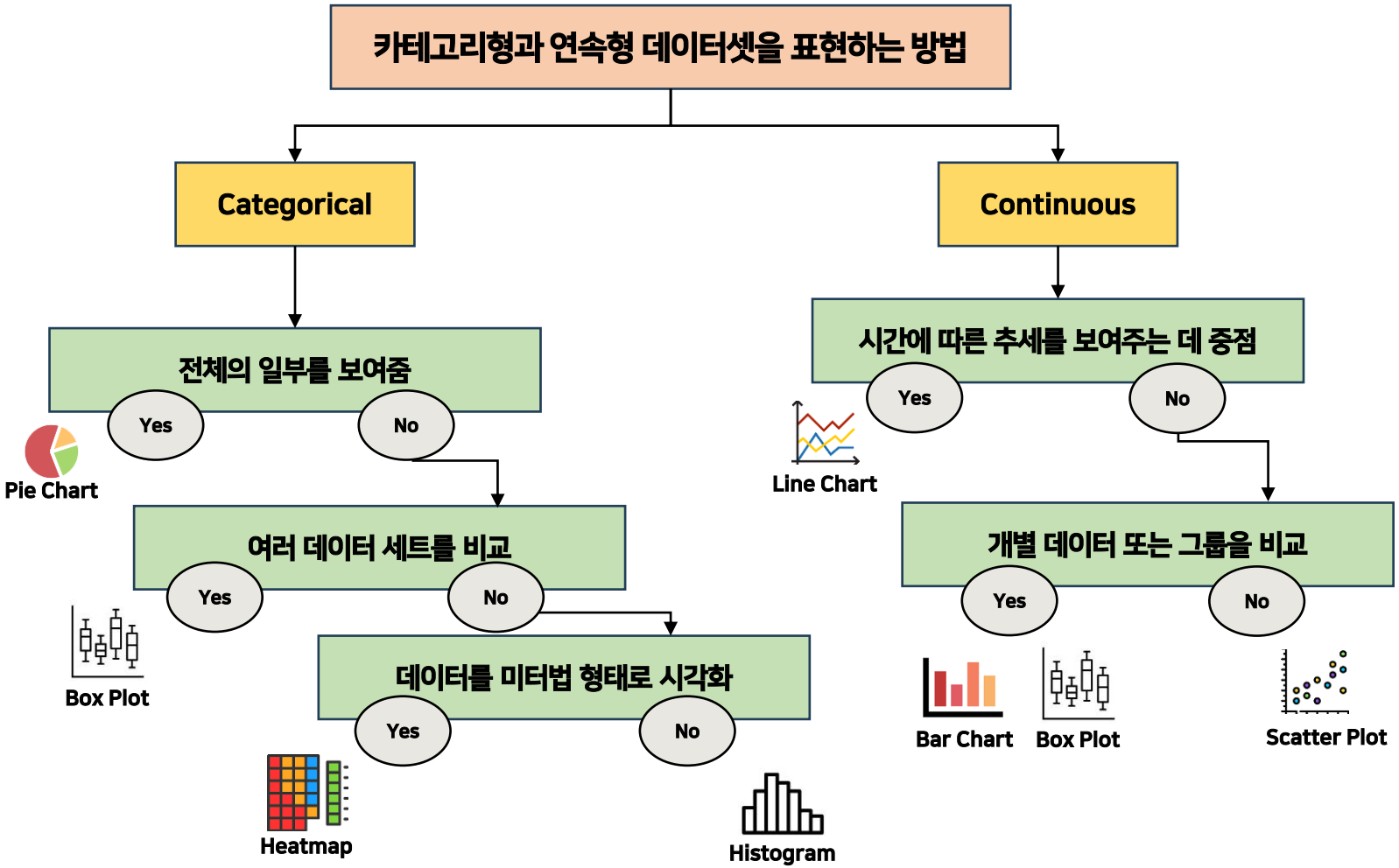
- 범위: 데이터 세트의 최대값과 최소값의 차이
- 사분위수 범위(IQR): 데이터의 중간 50%를 나타내는 첫 번째 사분위수(25% 백분위수)와 세 번째 사분위수(75% 백분위수) 사이의 값 범위
- 사분위수(Q1) : 아래쪽 절반에 짝수 개의 관측치가 있는 경우 Q1은 이 절반의 가운데 두 숫자의 평균
- 중앙값(Q2) : 짝수인 경우 중앙값은 가운데 두 숫자의 평균
- 사분위수(Q3) : 위쪽 절반에 짝수 개의 관측치가 있는 경우 Q3은 이 절반의 가운데 두 숫자의 평균

- 분산: 각 데이터 포인트와 평균 사이의 평균 제곱 차이
- 표준 편차: 데이터가 평균에서 얼마나 퍼져 있는지를 측정함

기술통계-탐색적 데이터 분석(EDA)

- EDA는 Exploratory Data Analysis의 약자로, 탐색적 데이터 분석을 의미함
- 데이터 분석을 시작하기 전에 데이터를 다양한 각도에서 관찰하고 이해하는 과정
- 데이터의 기본적인 특성, 구조, 패턴, 이상치, 변수 간의 관계 등을 파악함으로써 분석가가 보다 유익한 인사이트를 얻음
- 데이터에 대한 이해를 바탕으로 더 효율적인 분석 계획을 세울 수 있도록 하는 과정

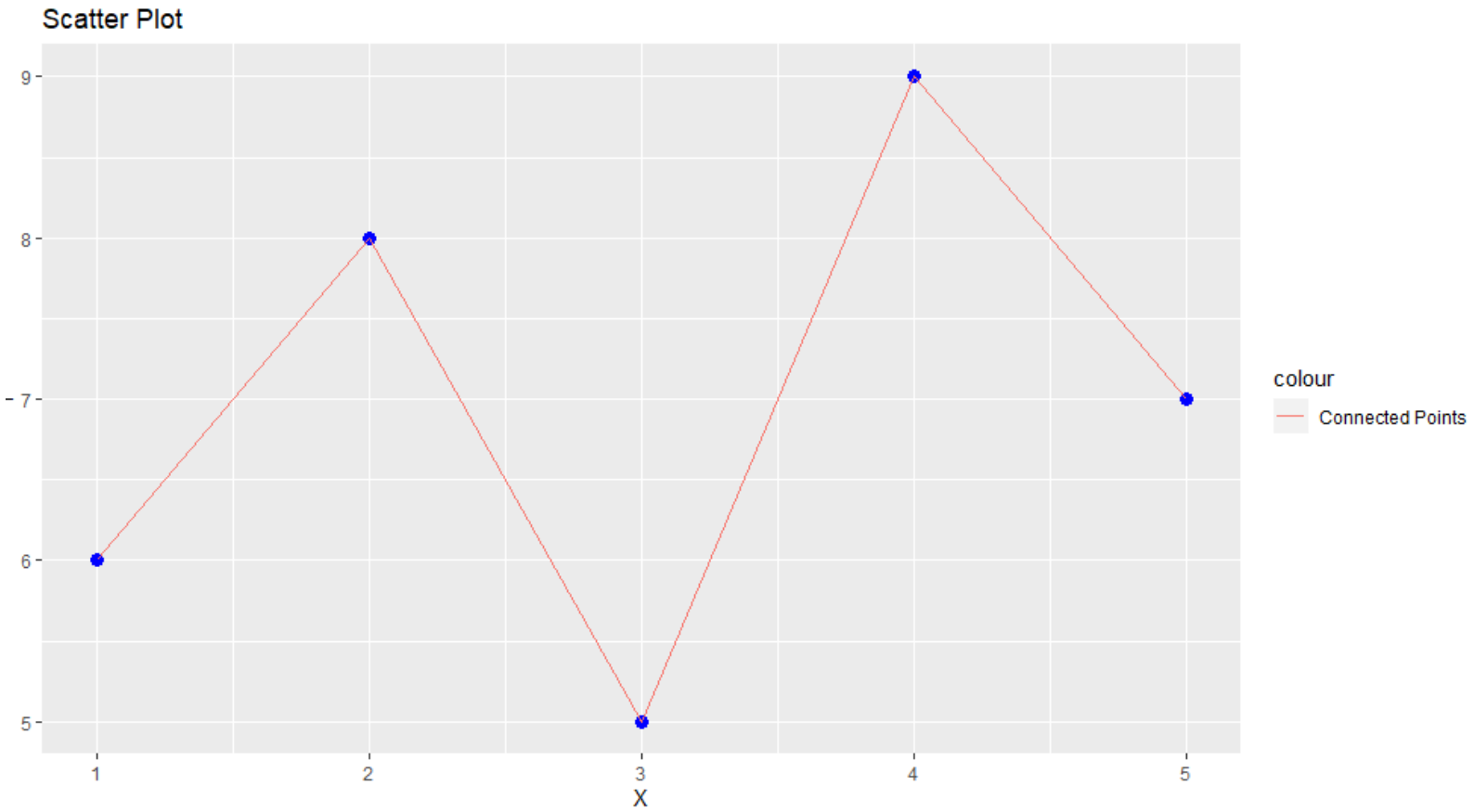
기술통계-탐색적 데이터 분석(EDA)



기술통계-탐색적 데이터 분석(EDA)

탐색적 데이터 분석(EDA) – Continuous

- 산점도(Scatter Plot) : 산점도는 직교 좌표계를 이용해 좌표상의 점들을 표시함으로써 두 개 변수 간의 관계를 나타내는 그래프 방법

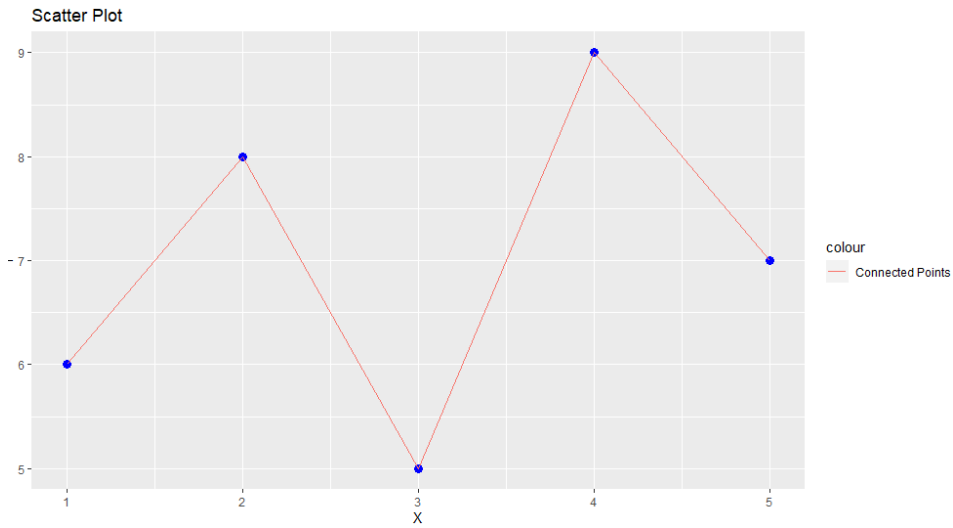


기술통계-탐색적 데이터 분석(EDA)

탐색적 데이터 분석(EDA) – Continuous

- 산점도(Scatter Plot) : 산점도는 직교 좌표계를 이용해 좌표상의 점들을 표시함으로써 두 개 변수 간의 관계를 나타내는 그래프 방법

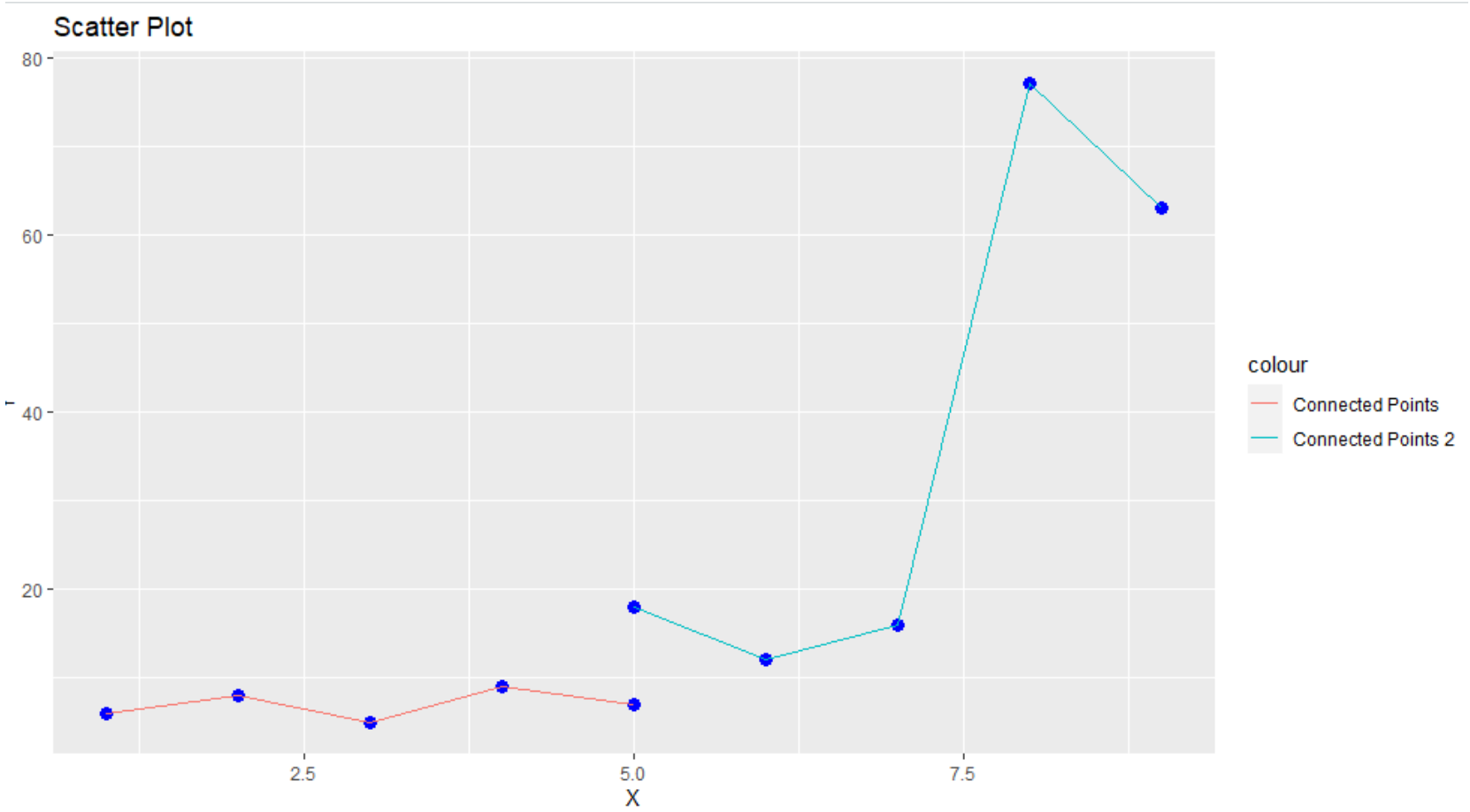
```
df <- data.frame( x = c(1, 2, 3, 4, 5),  y = c(6, 8, 5, 9, 7))

ggplot(df, aes(x = x, y = y)) +
        geom_point(color = "blue", size = 3) +
        geom_line(aes(color = "Connected Points"), size = 0.5) +
        labs(title = "Scatter Plot") +
        xlab("X") +
        ylab("Y")
```



기술통계-탐색적 데이터 분석(EDA)

탐색적 데이터 분석(EDA) – Continuous

- 산점도(Scatter Plot) : 산점도는 직교 좌표계를 이용해 좌표상의 점들을 표시함으로써 두 개 변수 간의 관계를 나타내는 그래프 방법

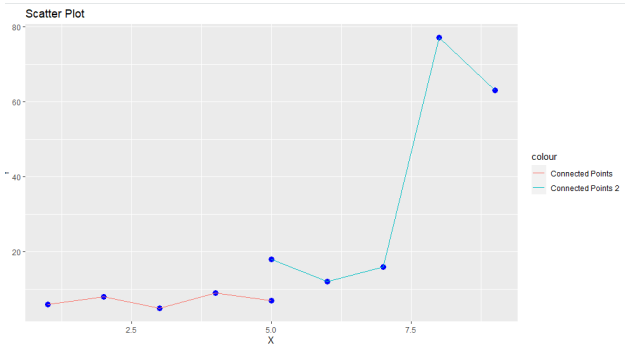


기술통계-탐색적 데이터 분석(EDA)

탐색적 데이터 분석(EDA) – Continuous

- **산점도(Scatter Plot) : 산점도는 직교 좌표계를 이용해 좌표상의 점들을 표시함으로써 두 개 변수 간의 관계를 나타내는 그래프 방법**

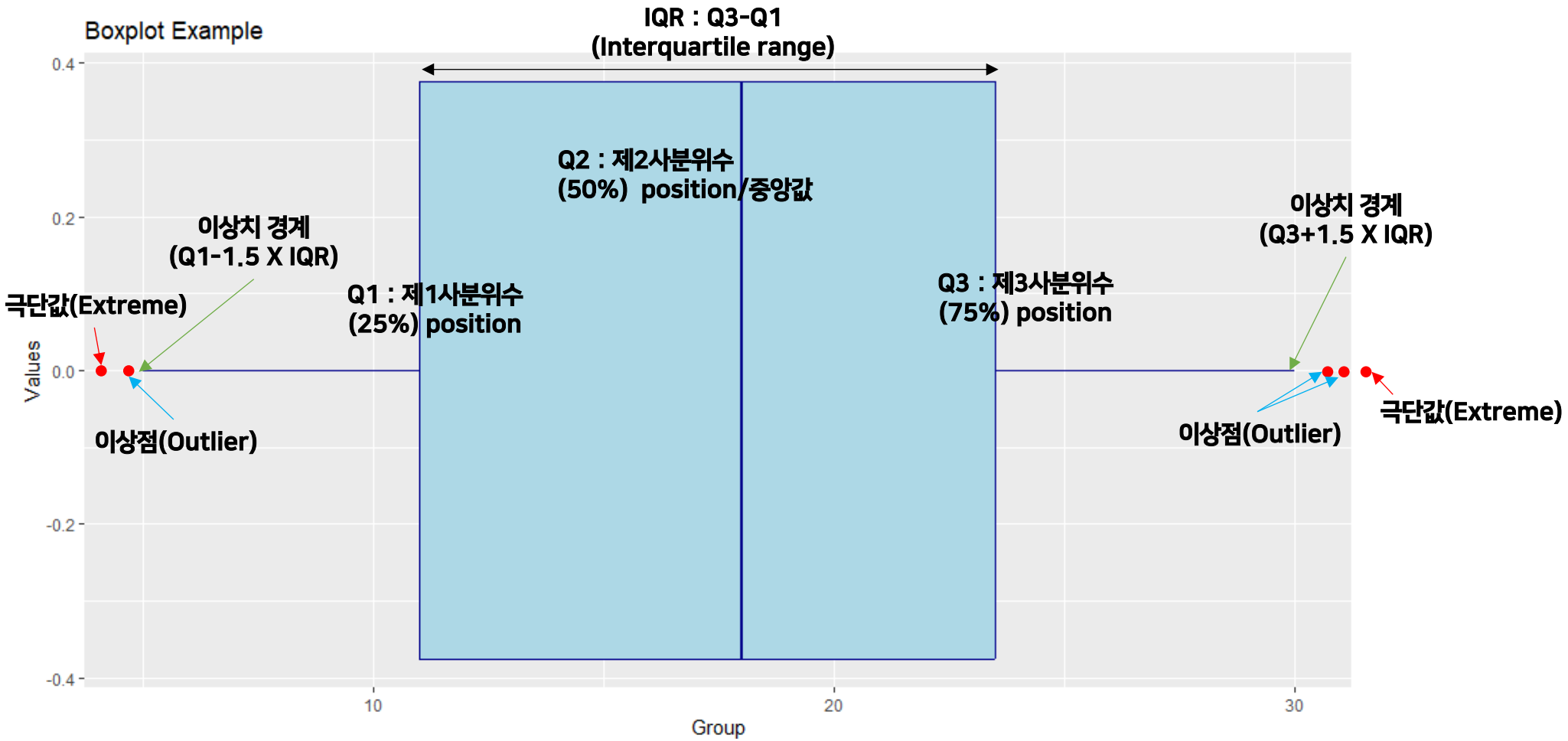
```
df <- data.frame(x = c(1, 2, 3, 4, 5), y = c(6, 8, 5, 9, 7))
df2 <- data.frame(x = c(5, 6, 7, 8, 9), y = c(18, 12, 16, 77, 63))

# Create the plot
ggplot() +
  geom_point(data = df, aes(x = x, y = y), color = "blue", size = 3) +
  geom_line(data = df, aes(x = x, y = y, color = "Connected Points"), size = 0.5) +
  geom_point(data = df2, aes(x = x, y = y), color = "blue", size = 3) +
  geom_line(data = df2, aes(x = x, y = y, color = "Connected Points 2"), size = 0.5) +
  labs(title = "Scatter Plot") +
  xlab("X") +
  ylab("Y")
```



기술통계-탐색적 데이터 분석(EDA)

탐색적 데이터 분석(EDA) – Continuous & Categorical

- 상자수염그림(Boxplot) : box-and-whisker plot이라고도 하는 box plot은 데이터 집합의 분포를 요약하는 그래프 표현

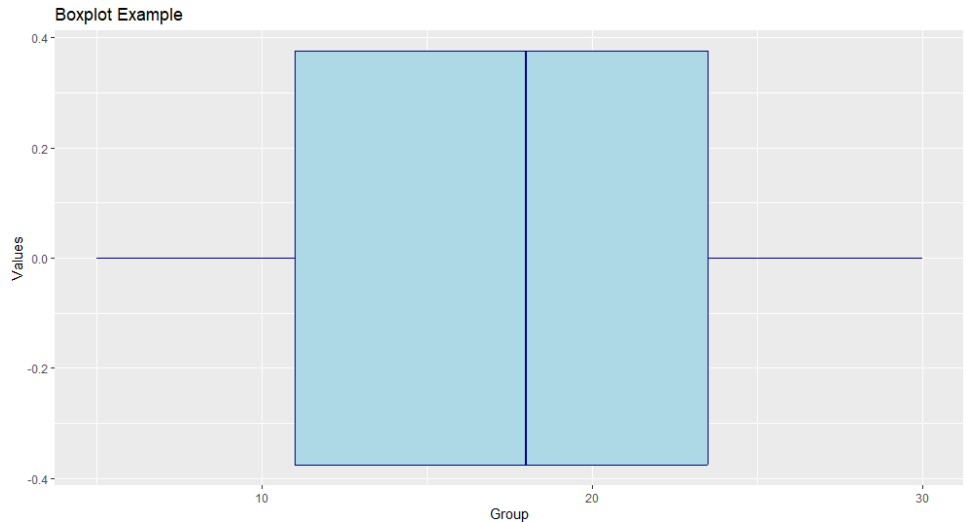


기술통계-탐색적 데이터 분석(EDA)

탐색적 데이터 분석(EDA) – Continuous & Categorical

- **상자수염그림(Boxplot) : box-and-whisker plot이라고도 하는 box plot은 데이터 집합의 분포를 요약하는 그래프 표현**

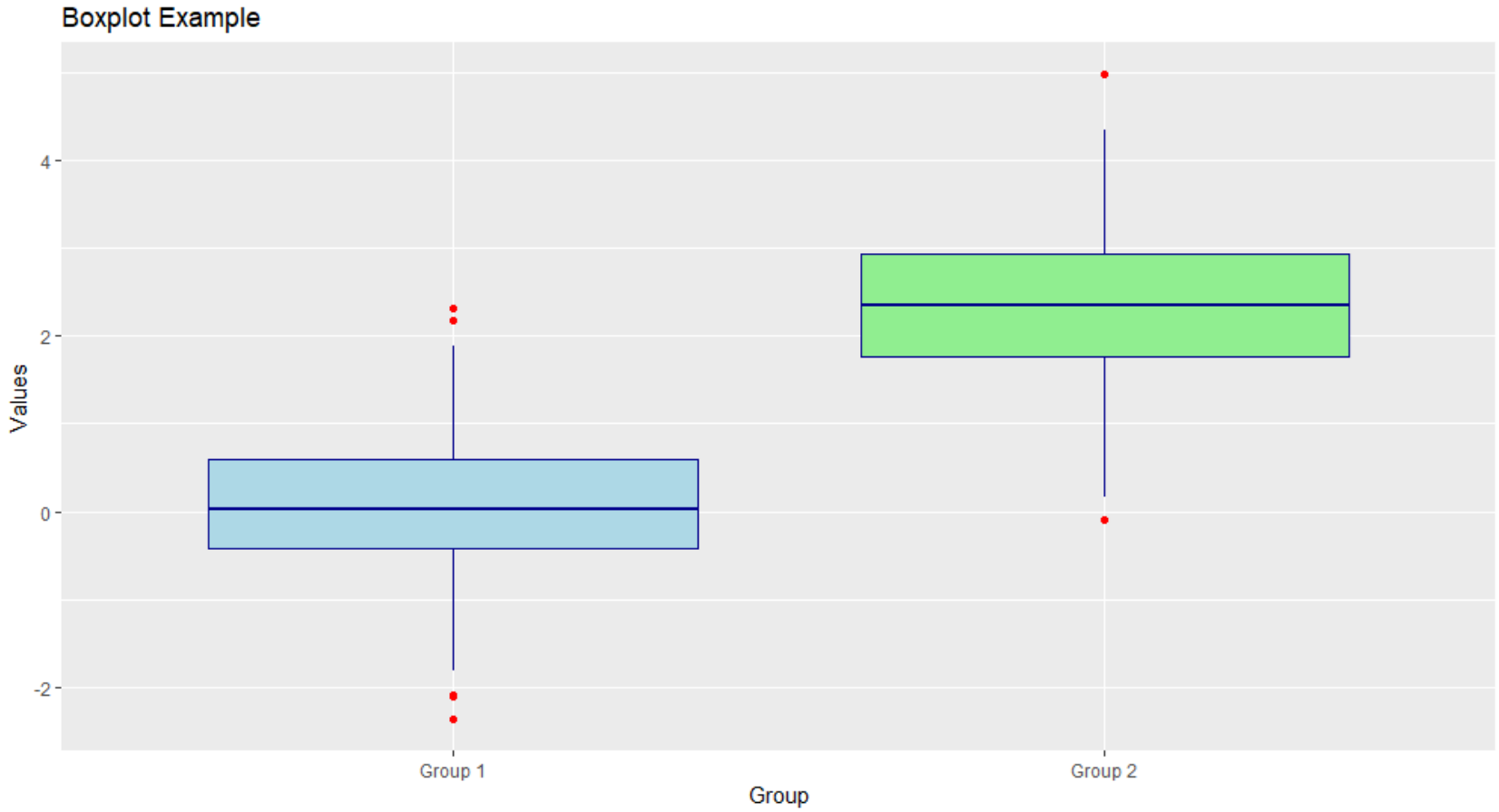
```
df <- data.frame(values = c(5, 7, 10, 12, 14, 18, 20, 22, 25, 27, 30))

ggplot(df, aes(x = values)) +
  geom_histogram(binwidth = 5, fill = "steelblue", color = "white") +
  labs(title = "Histogram of Values") +
  xlab("Values") +
  ylab("Frequency")
```



기술통계-탐색적 데이터 분석(EDA)

탐색적 데이터 분석(EDA) – Continuous & Categorical

- 상자수염그림(Boxplot) : box-and-whisker plot이라고도 하는 box plot은 데이터 집합의 분포를 요약하는 그래프 표현

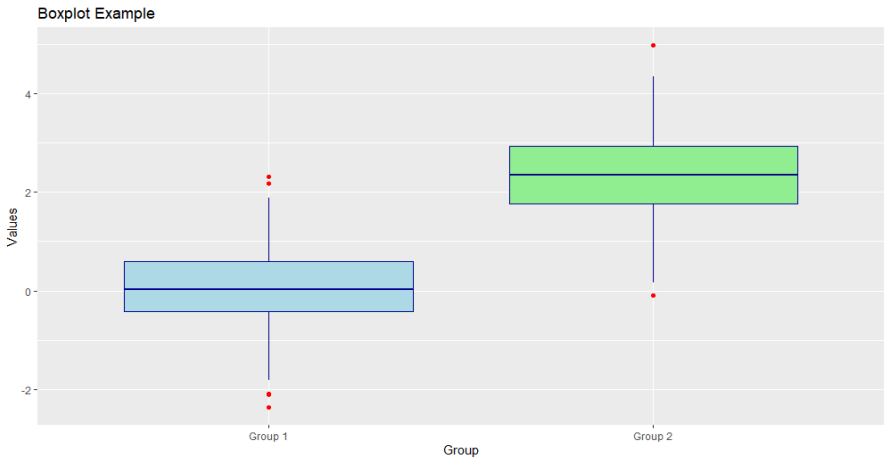


기술통계-탐색적 데이터 분석(EDA)

탐색적 데이터 분석(EDA) – Continuous & Categorical

- 상자수염그림(Boxplot) : box-and-whisker plot이라고도 하는 box plot은 데이터 집합의 분포를 요약하는 그래프 표현

```
df <- data.frame(
  group = c(rep("Group 1", 60), rep("Group 2", 60)),
  values = c(rnorm(60, mean = 0, sd = 1), rnorm(60, mean = 2, sd = 1)))

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

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

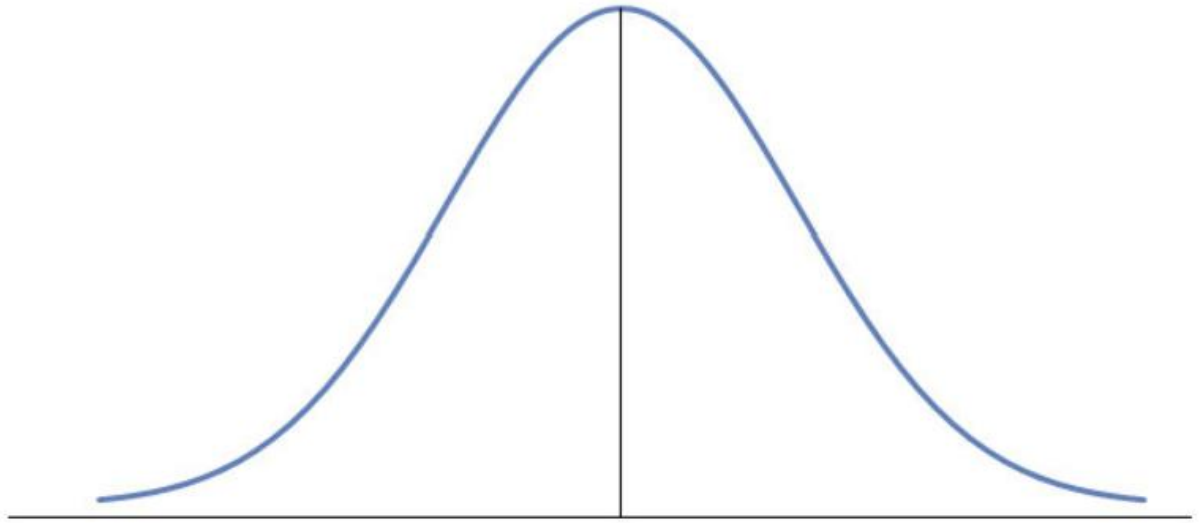
표본

모집단 예측

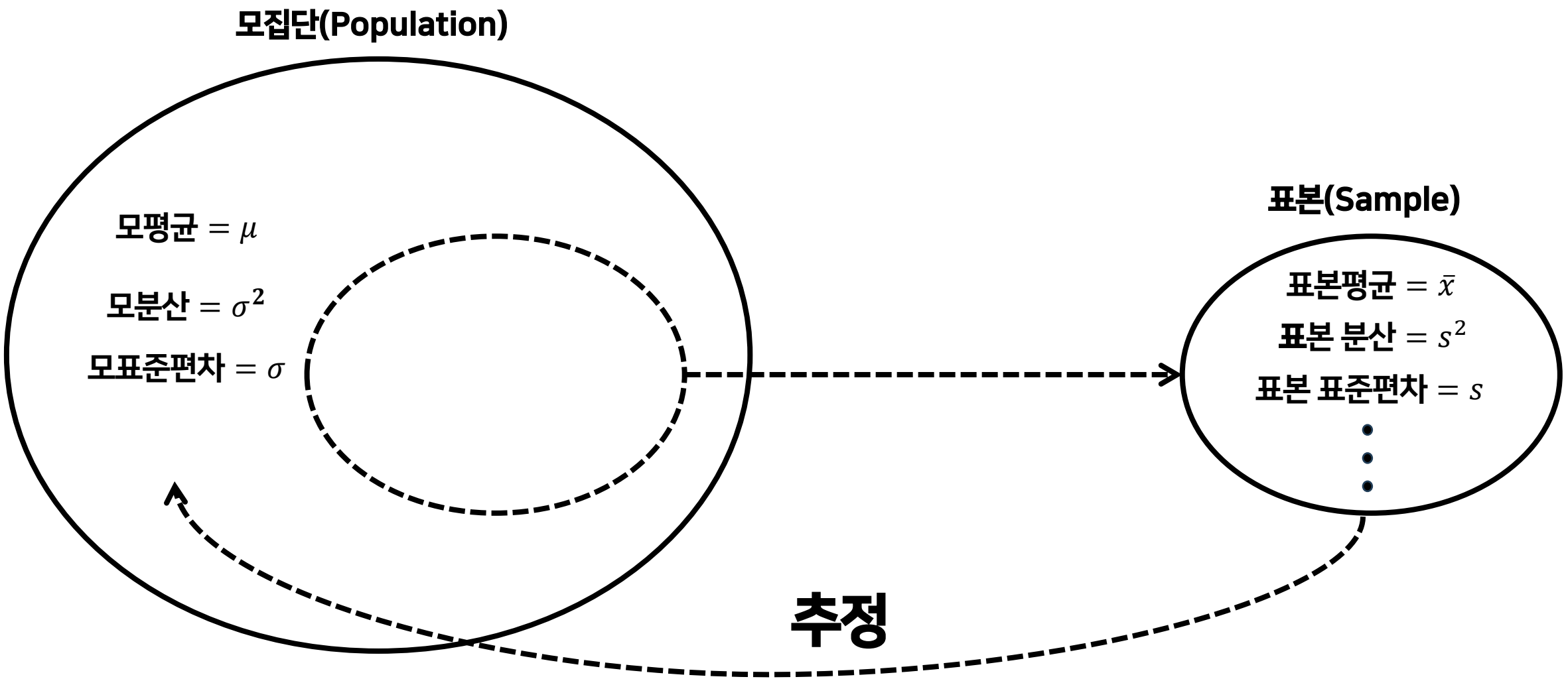
추론통계

계산이 되는 데이터에 대해서 가능함

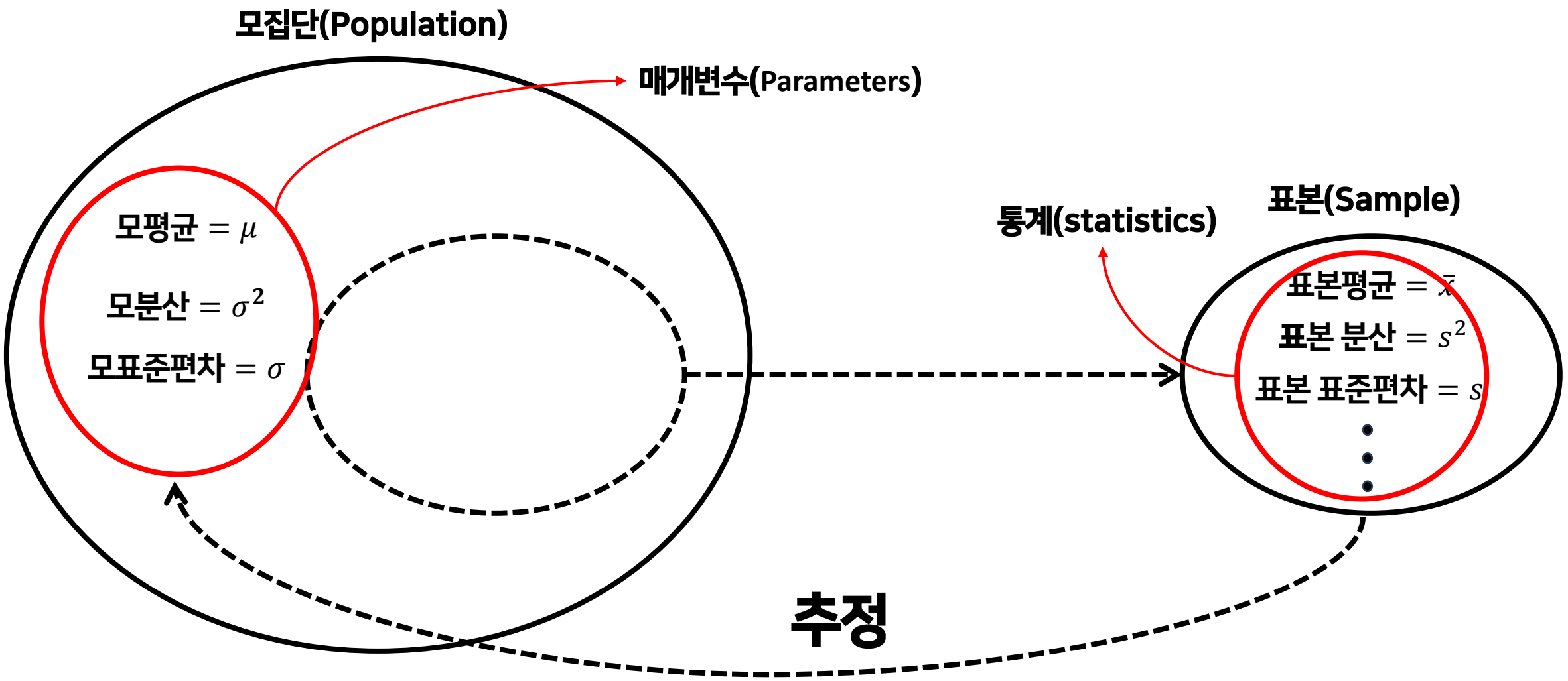
- 수치형 데이터에 주로 사용됨
- 중심 극한의 정리 : 표본의 크기가 커질수록 모집단의 분포와 상관없이 정규분포(Normal distribution)에 가까워진다는 것을 의미함
  - 표본의 크기(n≥30)는 평균의 샘플링 분포가 거의 정상
  - 모집단의 분산은 유한하고 알려져 있어야 함
  - 표본 관측치는 독립적이어야 함 → 하나의 관찰이 발생해도 다른 관찰의 발생에 영향을 미치지 않는다는 것을 의미



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

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

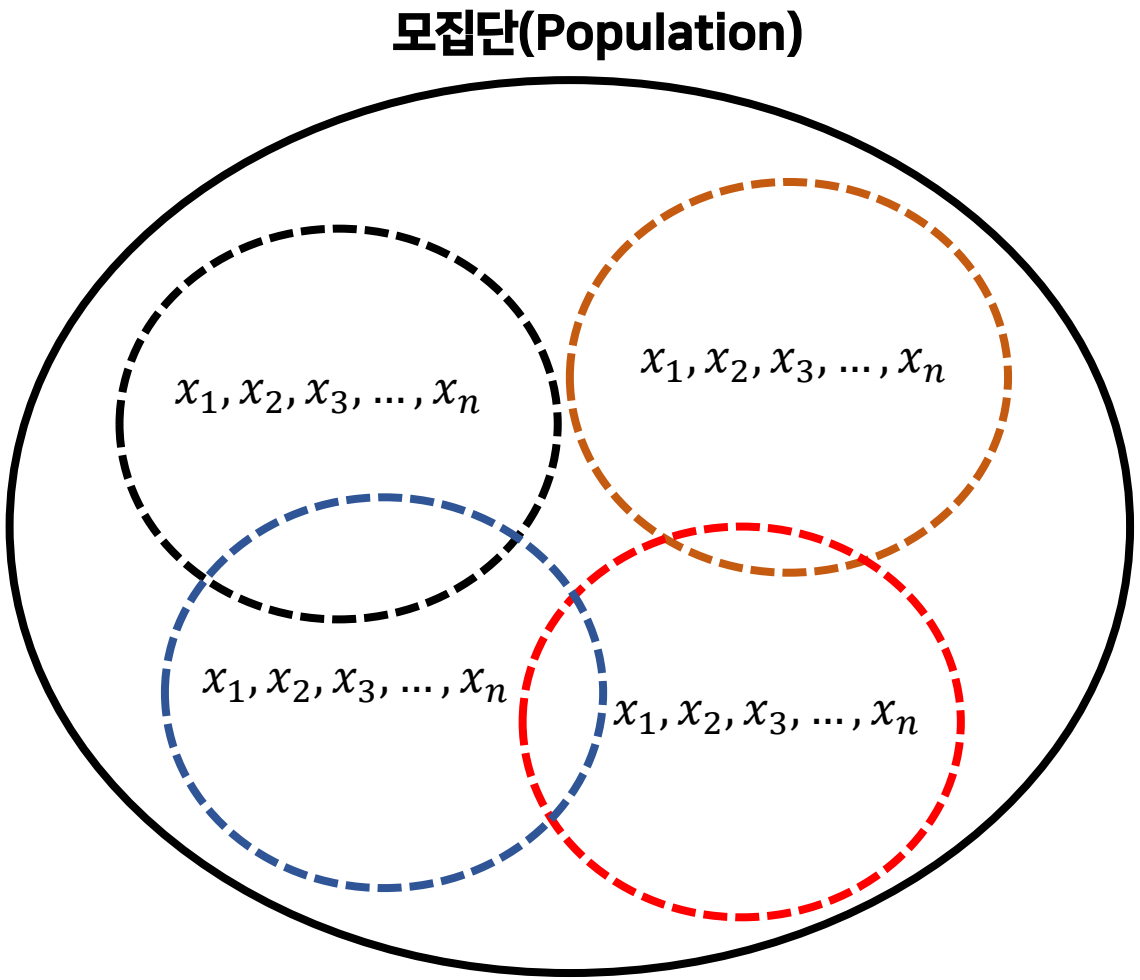
- 모집단의 분포가 정규분포를 따를 때 $\mathrm{N}(\mu, \sigma^2)$
- $X_1 + X_2 + X_3 \ldots, X_n$ 은 i.i.d. $\mathrm{N}(\mu, \sigma^2)$ → independent, identically, distributed
- $\mathrm{E}[\bar{X}] = \mu, V[\bar{X}] = \frac{\sigma^2}{n}$ ➔ **확률변수의 모집단의 평균, 모집단의 분산**
- 표본분포는? $\mathrm{N}(\mu, \sigma^2/n)$

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)

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

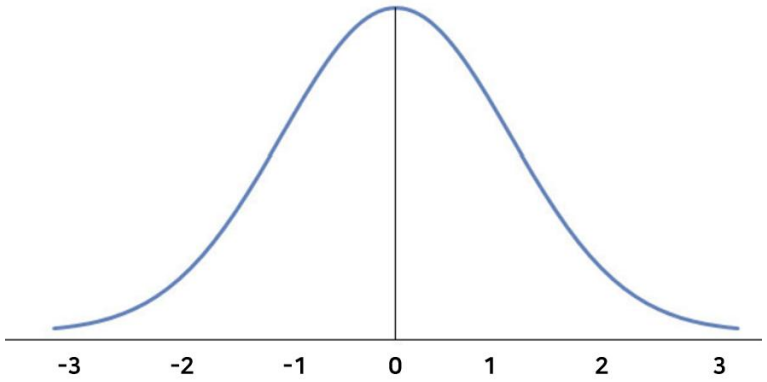
표본평균도 정규분포를 따름
→ 검정 통계량 Z: 정규 분포의 평균에서 얼마나 많은 표준 편차를 벗어났는지 계산하는 데 사용됨

추론통계

표준정규분포(Standard Normal Distribution)

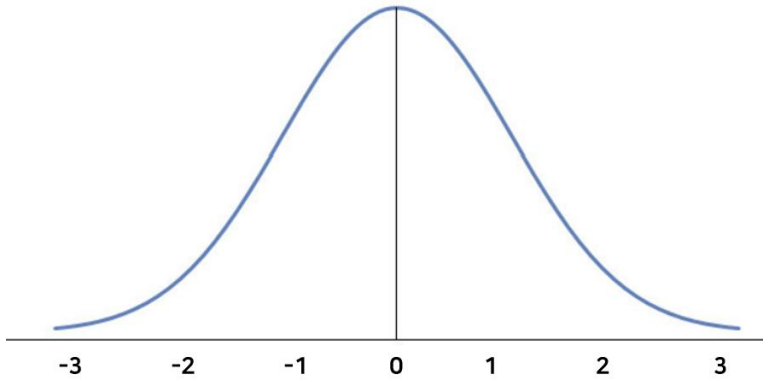
- 평균(μ)이 0이고 표준 편차(σ)가 1인 특정 유형의 정규 분포

Normal Distribution



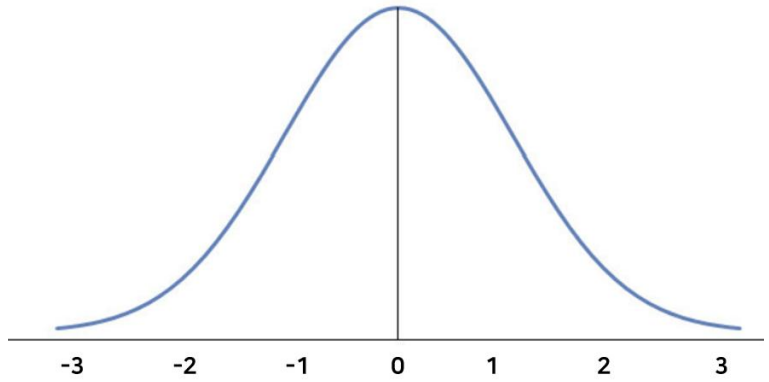
정규 분포는 항상 종 모양

Sampling Distribution



세상에 존재하는 실제 데
이터의 형태에 가까움

Standard Normal Distribution

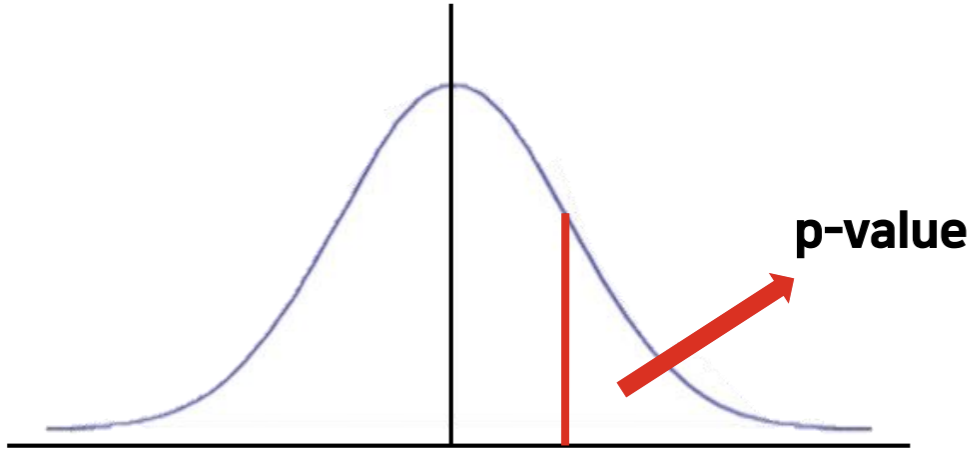


항상 종 모양

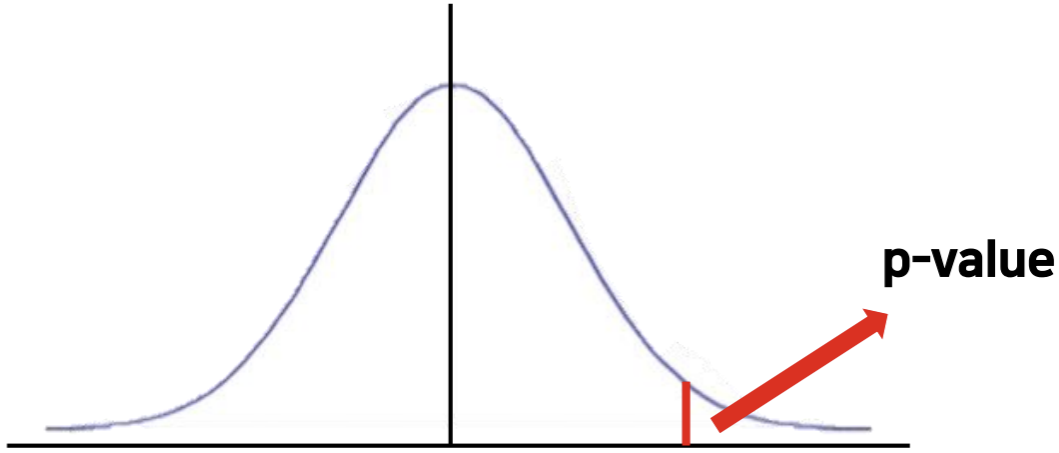
추론통계

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

- 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

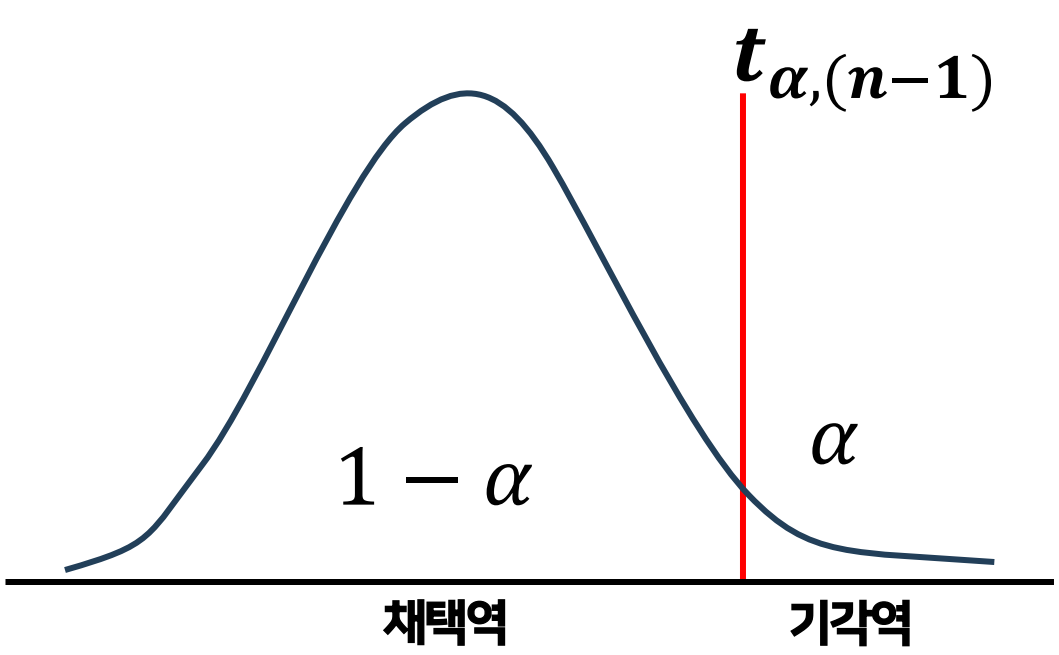
- Z-test : Z-test는 t-test와 유사하지만 표본 크기가 크고(일반적으로 $n > 30$) **모집단 표준 편차를 알고 있을 때** 사용
  - 그룹의 평균이 가설 값과 유의하게 다른지 테스트함 → 평균의 분포

추론통계

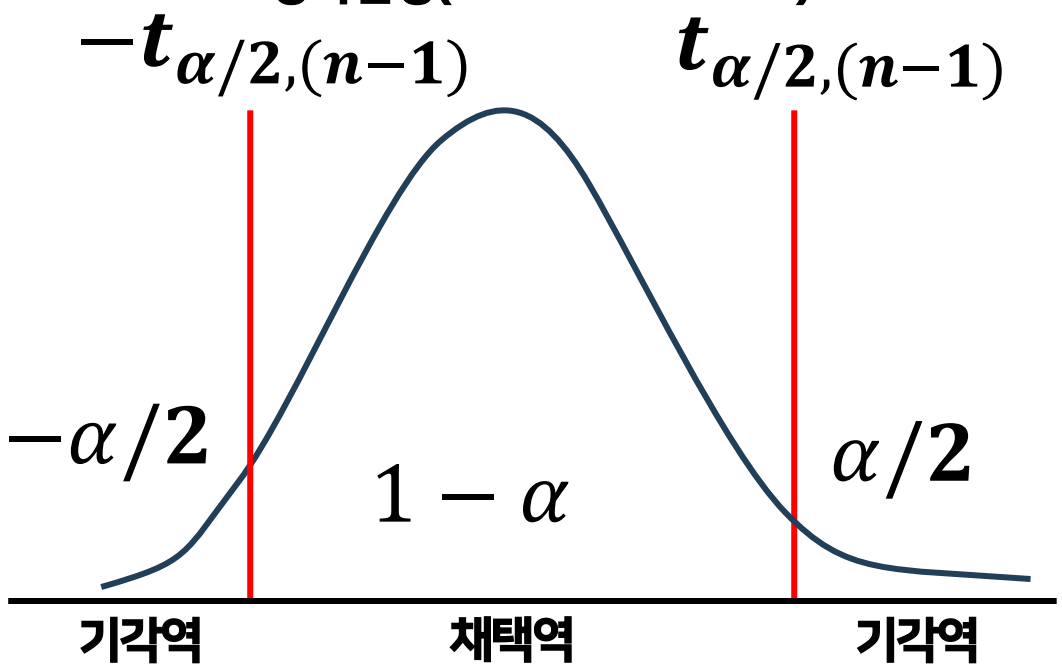
가설 검정 및 추론통계

- 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-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)

추론통계

분산의 분포

- 분산 분포는 확률 변수의 분산(또는 동등하게 표준 편차의 제곱)을 설명하는 통계적 분포를 나타냄 ➔ 변동성
- 통계적 추론에서, 특히 작은 표본을 다룰 때 분산 분포를 아는 것은 가설 검정과 모집단 분산에 대한 신뢰 구간 구성에 중요
  - 임상 시험: 두 가지 치료법의 효과를 비교하는 경우 어떤 치료법이 더 나은 평균 결과를 나타내는지 뿐만 아니라 어떤 치료법이 더 일관된(더 낮은 분산) 결과를 나타내는지 알고 싶음
  - 교육: 서로 다른 두 가지 교육 중재 간의 시험 점수를 비교할 때 일관된 차이를 이해하면 해당 중재가 학생 전체에 걸쳐 얼마나 잘 작동하는지 나타낼 수 있음
  - 제조: 품질 관리에서는 단순히 높은 평균 품질이 아닌 일관되게 높은 품질의 제품을 원하기 때문에 평균 품질 수준 뿐만 아니라 편차도 아는 것이 중요한 경우가 많음
  - 재무: 포트폴리오 관리에서 평균은 기대 수익을 제공할 수 있지만 분산 또는 표준 편차는 관련 위험에 대한 아이디어를 제공함

추론통계

가설 검정 및 추론통계 (연관관계 파악)

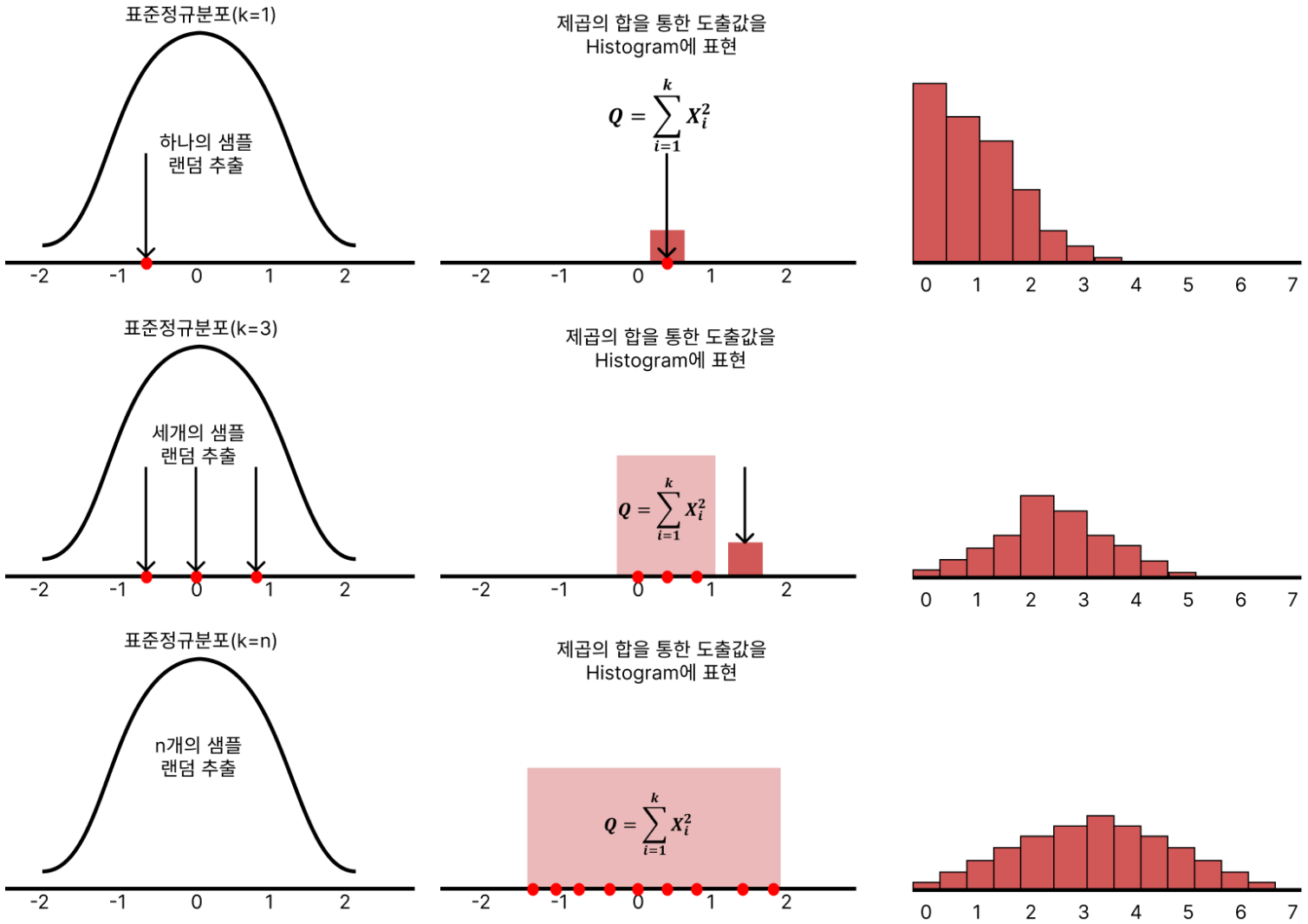
- 카이제곱분포(Chi-Square Distribution) : 카이제곱 검정은 두 범주형 변수 사이에 유의미한 연관성을 확인하는 데 사용
  - 분할표를 분석하고 관찰된 빈도 분포가 예상 빈도 분포와 다른지 여부를 테스트하는 데 자주 사용

- 분산 분석(ANOVA) : 여러 그룹을 비교하여 이러한 그룹 간의 분산(변동성)에 통계적으로 유의미한 차이가 있는지 확인
  - 관찰된 평균 차이가 무작위 변동으로 인한 것인지, 아니면 그룹 간의 실제 차이를 반영하는지를 판단
  - 목표는 그룹 간 변동성이 그룹 내 변동성보다 훨씬 큰지 확인하는 것

- 상관관계 테스트 : 상관 테스트는 두 연속 변수 사이의 관계의 강도와 방향을 측정하는 데 사용
  - Pearson 상관관계는 선형 관계에 적합하고 Spearman 순위 상관관계는 비선형 관계에 사용

추론통계

Chi-Square Test



K값이 증가하면 정규분포와 유사한 형태를 가짐

추론통계

Chi-Square Test

- 카이제곱 분포는 오차 혹은 편차를 분석할 때 사용함
- 카이제곱 분포를 이용해 오차나 편차 검증하면 → 우연히 발생하는 오차인지 숨겨진 의미가 있는 오차나 편차 인지 알 수 있음
- 두 가지 검정
  - 적합도 검정(goodness-of-fit-test) → 기대되는 빈도의 분포와 관찰한 빈도의 분포를 비교
  - 독립성 검정(chi-square independence test) → 범주형 변수가 여러 개인 경우에 사용하는 분석방법
- 두 경우 모두 다음의 아래의 통계량 공식을 사용함

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \sim \chi^2(k-1)$$

추론통계

Chi-Square Test : 적합도 검정(goodness-of-fit-test)

- 적합도 검정(goodness-of-fit-test) → 기대되는 빈도의 분포와 관찰한 빈도의 분포를 비교
- 분석하려는 범주형 변수의 각 범주에 대해 기대되는 빈도를 계산함
- 실제 데이터에서 각 범주의 관찰된 빈도와 예상 빈도를 비교하고, 이를 통해 각 범주 간의 차이를 계산함

가설 설정 방법

- 귀무가설(H_0) : 주어진 데이터는 분포가 적합하다.→ 실제 관찰된 빈도와 기대되는 이론적인 빈도 간에는 유의미한 차이가 없다.
- 대립가설(H_1) : 주어진 데이터는 분포가 적합하다. → 실제 관찰된 빈도와 기대되는 이론적인 빈도 간에 유의미한 차이가 있다.

추론통계

Chi-Square Test : 독립성 검정(chi-square independence test)

- 교차 분석(cross tabulation analysis) → 범주형 변수가 여러 개인 경우에 사용하는 분석방법
- 여러 범주형 변수의 범주 간 차이가 기대값에서 유의하게 벗어나는지를 판단 → 변수 간의 연관관계 파악

Subject ID	Age Group	Sex
GW	young	F
JA	middle	F
TJ	young	M
JMA	young	M
JMO	middle	F
JQA	old	F
AJ	old	F
MVB	young	M
WHH	old	F
JT	young	F
JKP	middle	M

변수를 가지는 데이터 셋

Age Group / Sex	Female	Male	Total
young	2	3	5
middle	2	1	3
old	3	0	3
Total	7	4	11

교차 테이블(Cross-tabulation)

추론통계

```
data <- data.frame(Gender = c("Male", "Female", "Male", "Male", "Female",
"Female", "Male", "Male", "Female", "Female"),  Food = c(“국밥”, “마라탕”, “국밥”,
“피자”, “피자”, “국밥”, “국밥”, “마라탕”, “피자”, “피자”))

cross_tab <- table(data$Gender, data$Food)
cross_tab

chi_square_test_result <- chisq.test(cross_tab)
print(chi_square_test_result)
```

추론통계

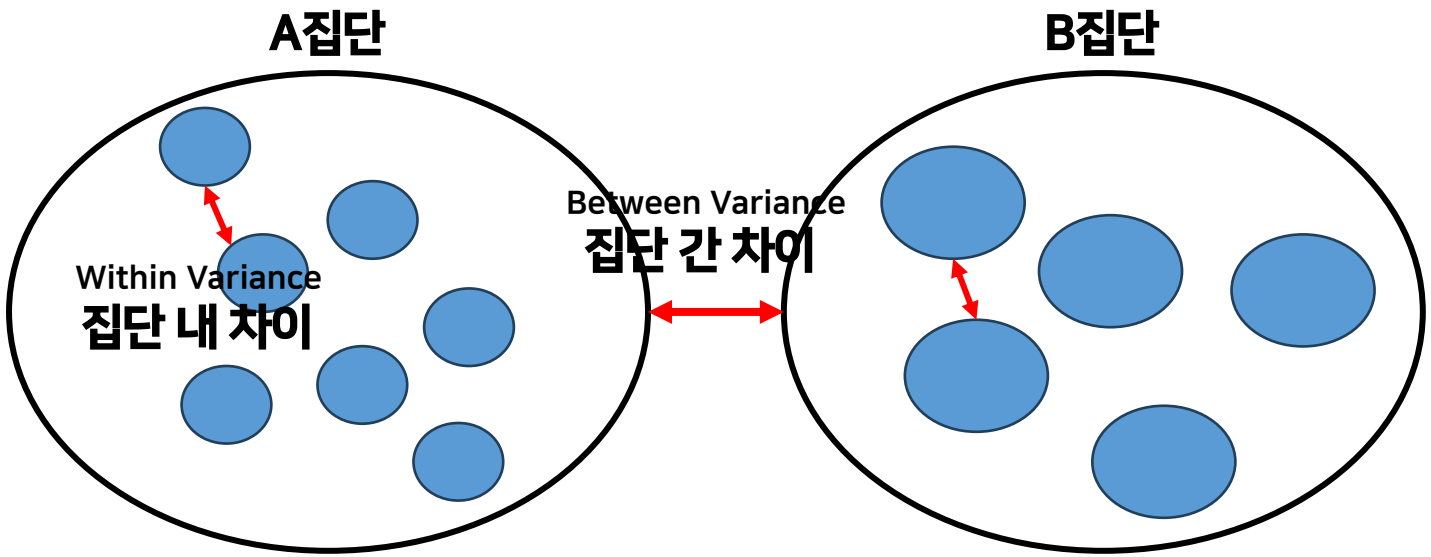
분산분석(F-test)

- F-분포(Fisher-Snedecor distribution) : F-value는 분산의 비율
- 따라서 분산분석이라 부름
- 전체 평균으로부터 각 집단의 평균 까지의 분산(Between Variance) → 집단 간의 차이
  - 전체 평균으로부터 각 집단의 평균값이 멀리 떨어져 있음 → 적어도 하나의 집단은 한 개는 다른 집단과 평균이 다를 수 있음
- 각각의 집단의 한 지점이 해당 집단으로부터 얼마나 떨어져 있는지의 분산(Within Variance) → 집단 내의 차이
- Between Variance가 Within Variance보다 커야 Between Variance가 통계적으로 유의하다 말할 수 있음
- 이것이 한 그룹의 평균값이 전체 평균값과는 다르다고 할 수 있음

추론통계

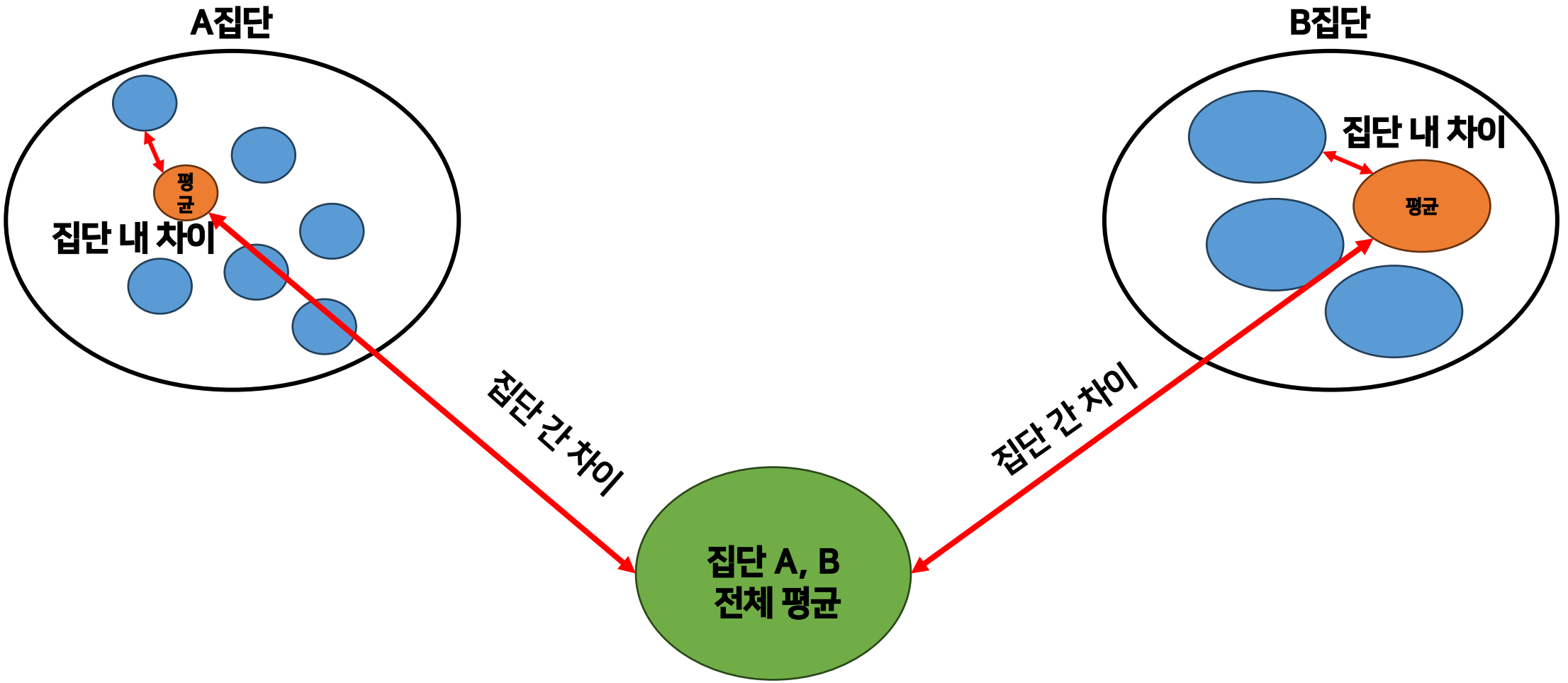
분산분석(F-test)

- F-분포(Fisher-Snedecor distribution) : 두 표본의 분산비에 대한 분포
- 집단 간 분산(Between Variance)을 집단 내 분산(Within Variance)으로 나눈 것 → 분산을 활용해 평균을 비교
- F-분포도 양수를 가짐
- 기준이 1 이고, 값이 커지면 집단간 분산과 집단 내 분산의 차이가 큰 것



추론통계

분산분석(F-test)



추론통계

가설 검정 및 추론통계

- 두 연속 집단의 평균 차이 비교(T-Test, Z-Test)
- 두 명목 집단의 연관성 비교(chi square-Test)
- 두 연속 집단의 분산 차이 비교(F-Test)

- 그룹을 비교할 때 세개 이상의 그룹을 비교할 수는 없을까?
- 분산분석(ANOVA) 세 개 이상의 그룹의 분산을 활용해 평균을 비교하는데 사용

추론통계

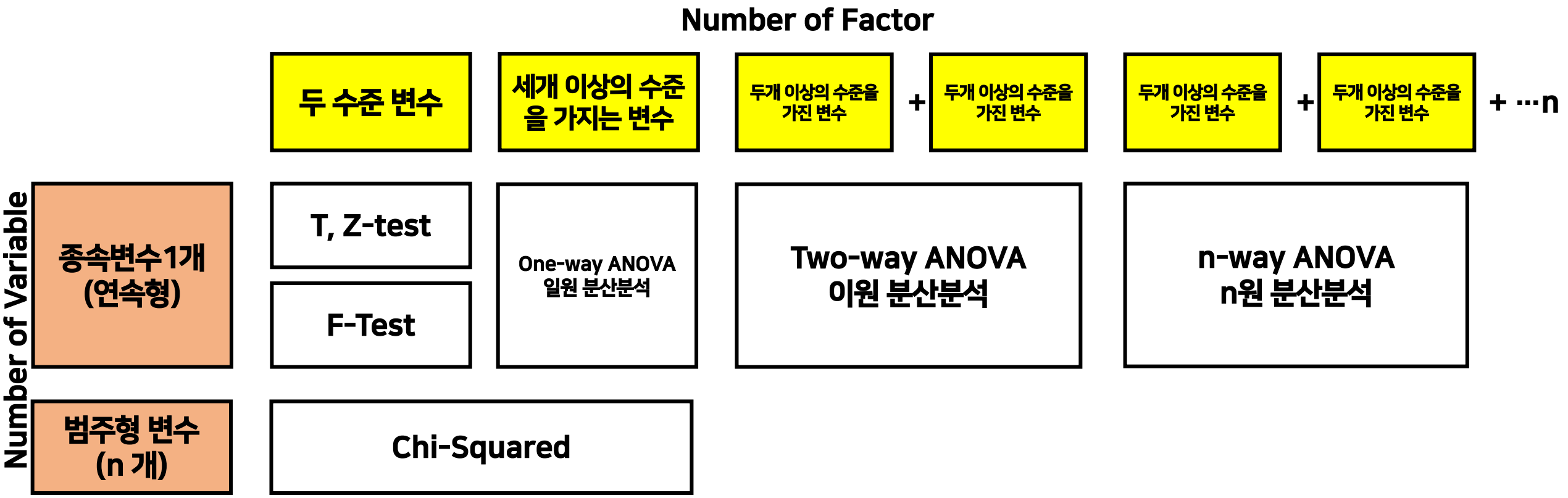
가설 검정 및 추론통계

- 독립변수 : 가설의 원인이 되는 변수, 종속변수에 영향을 미치는 변수
- 종속변수 : 가설의 결과가 되는 변수, 독립변수로 영향을 받는 변수

원인	결과
독립 변수(Independent Variable)	종속 변수(Dependent Variable)
설명 변수(Explanatory Variable)	반응 변수(Response Variable)
예측 변수(Predictor Variable)	결과 변수(Outcome Variable)
	표적 변수(Target Variable)

추론통계

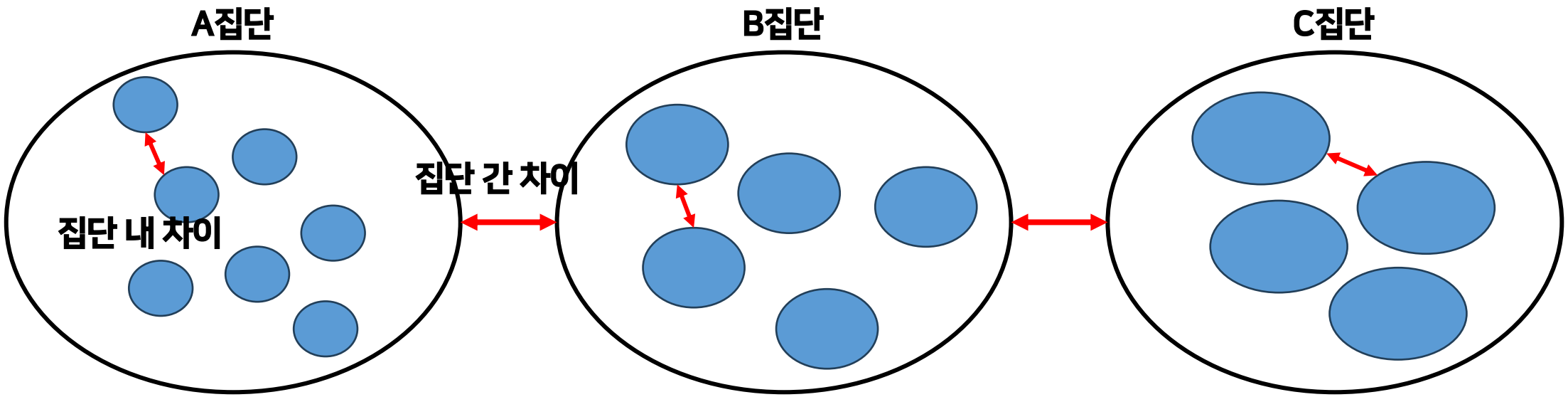
가설 검정 및 추론통계



추론통계

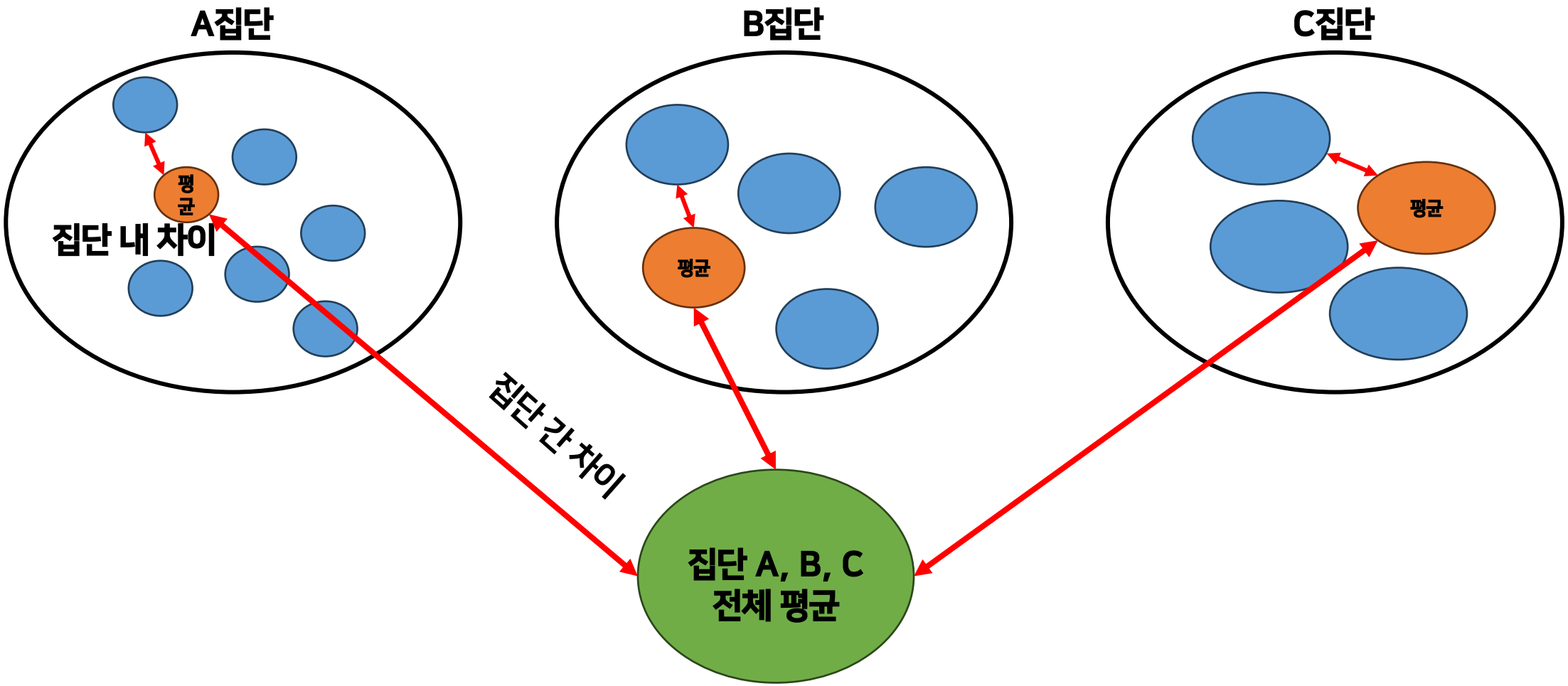
분산분석(ANOVA)

- 분산분석(ANOVA) : 세 개 이상의 그룹 평균을 비교하여 평균의 차이가 존재하는지 판단하는 방법
- 집단 간 분산을 집단 내 분산으로 나눈 것 → 분산을 활용해 평균을 비교
- F-분포도 양수를 가짐
- 기준이 1 이고, 값이 커지면 집단간 분산과 집단 내 분산의 차이가 큰 것



추론통계

분산분석(ANOVA)



추론통계

분산분석(ANOVA)

- 일원 분산분석(one-way ANOVA): 두 개 이상의 수준 또는 범주가 있는 하나의 독립변수가 있고, 이를 종속변수와 비교
- 이원 분산분석(Two-way ANOVA): 두 개의 독립 변수가 존재하며, 이 변수들은 개별적으로, 연관적으로 종속변수에 영향을 미치는지 판단하는 것(ex) 식이요법과 운동 수준이 체중 감량에 어떤 영향을 미치는지 알고 싶을 때

종속변수	독립변수
50	A 학원
60	B 학원
70	A 학원
85	B 학원
67	C 학원
88	A 학원
54	C 학원

일원분산분석

종속변수	독립변수 1	독립변수 2
50	A 학원	5(명목/순서형)
60	B 학원	5(명목/순서형)
70	A 학원	6(명목/순서형)
85	B 학원	8(명목/순서형)
67	C 학원	5(명목/순서형)
88	A 학원	8(명목/순서형)
54	C 학원	4(명목/순서형)

이원분산분석

추론통계

```
grow <- read.csv("C:/Users/USER/Desktop/cafe.csv", stringsAsFactors = TRUE) #문자형
변수를 요소로 변환

#분산분석(독립변수들간 상호작용X)
anova_result <- aov(Satisfaction ~ CoffeeType, data = grow)

summary(anova_result)

#사후검정
tukey_result <- glht(anova_result, linfct = mcp(CoffeeType = "Tukey"))

summary(tukey_result)
```

추론통계

상관관계 분석

- 상관관계 테스트(Correlation Test) : 상관 테스트는 두 연속 변수 사이의 연관성이 있는지 확인하는 데 사용됨
- Pearson 상관관계는 선형 관계에 적합하고 Spearman 순위 상관관계는 비선형 관계에 사용됨
- 상관 분석: 상관 분석은 두 연속 변수 간의 선형 관계의 강도와 방향을 측정하는 데 사용됨
- Pearson의 상관 계수(선형 관계의 경우) 또는 Spearman의 순위 상관 계수(단조 관계의 경우)를 사용하여 평가됨
- 상관 계수는 -1과 1 사이의 값을 가지며, 여기서 -1은 완벽한 음의 선형 관계를 나타내고, 1은 완벽한 양의 선형 관계를 나타내고, 0은 선형 관계가 없음을 나타냄

상관관계 분석

- 공분산(Covariance) : 두개의 변수사이의 관계를 숫자로 알려줄 수 있는 값
- 두 변수의 독립일 때, Cov(X, Y) = 0
- Cov(X, Y) = 0일 때, 반드시 독립이라고 할 순 없음
- H_0 : 상관계수가 0이다.
- H_1 : 상관계수가 0이 아니다.

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

$$\begin{aligned} Cov(x,y) &= E[X - E[X])(Y - E[Y])] \\ &= E[XY - XE[Y] - Y[E[X] + E[X]E[Y]] \\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

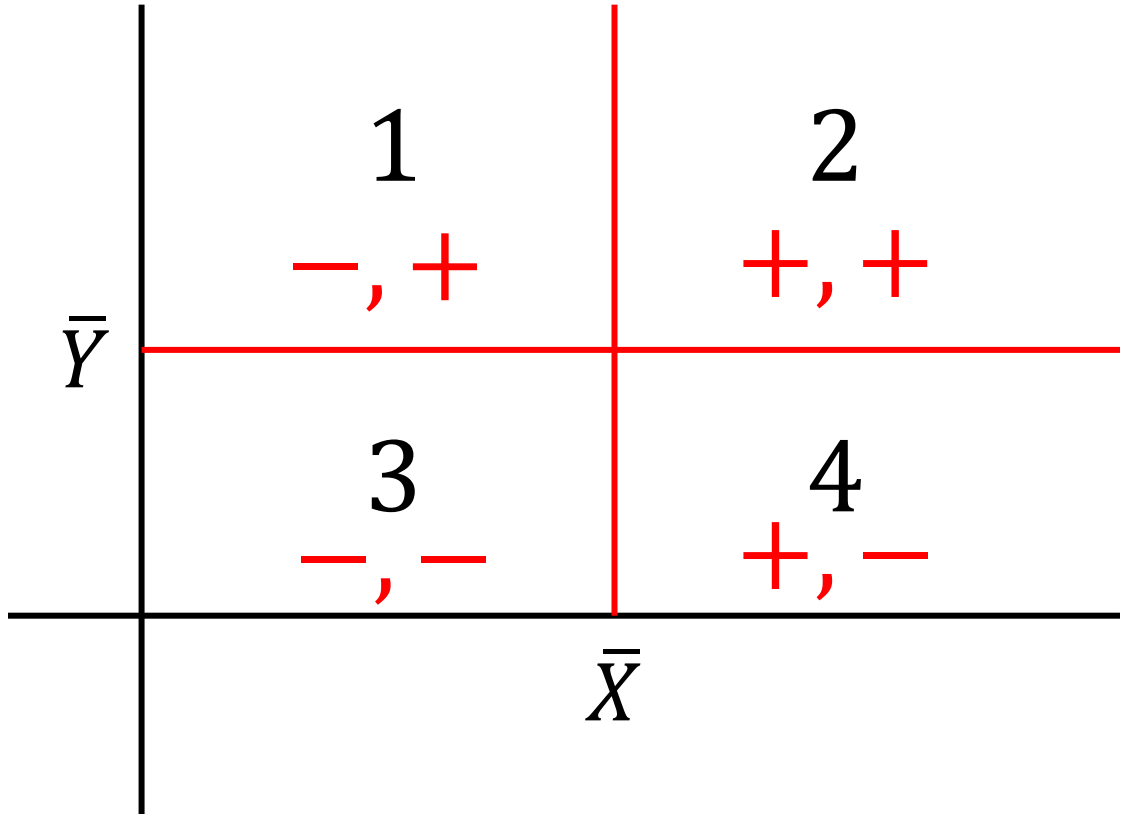
$$Cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

상관관계 분석

- 공분산(Covariance) : 두개의 변수사이의 관계를 숫자로 알려줄 수 있는 값
- 두 변수의 독립일 때, Cov(X, Y) = 0
- Cov(X, Y) = 0일 때, 반드시 독립이라고 할 순 없음

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

$$Cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$



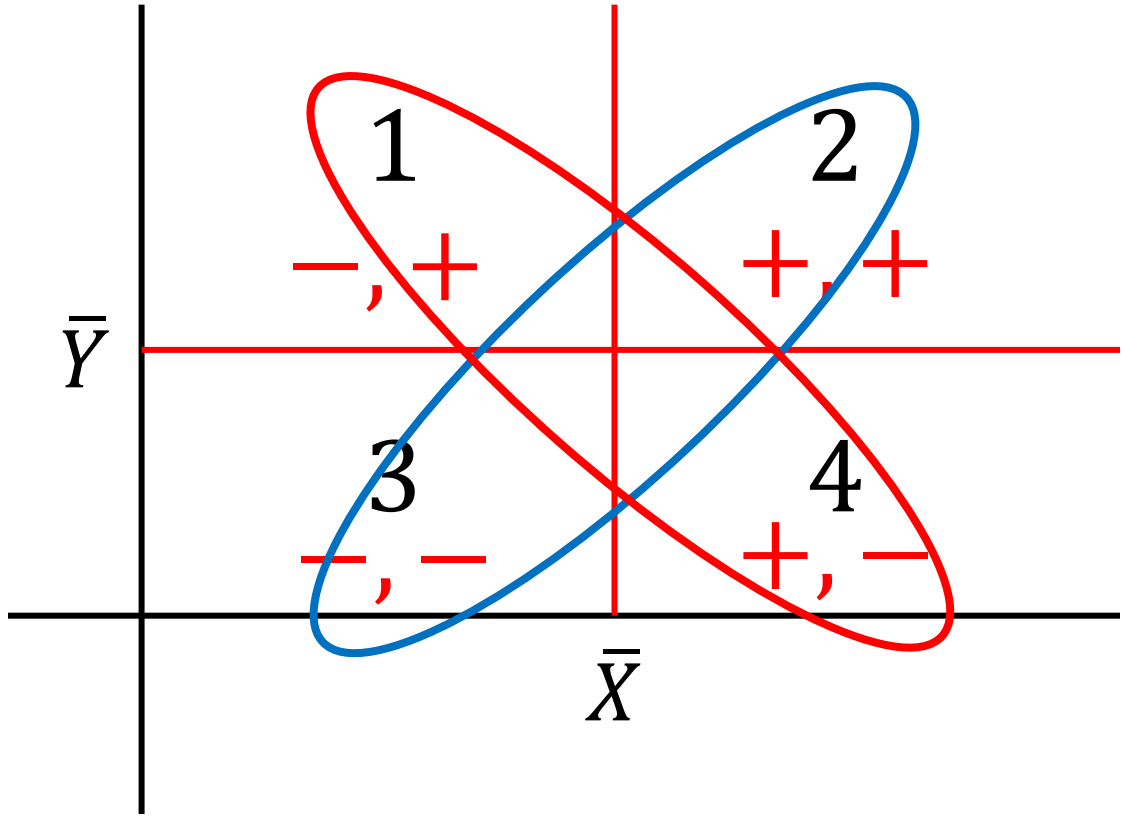
상관관계 분석

- 공분산(Covariance) : 두개의 변수사이의 관계를 숫자로 알려줄 수 있는 값
- 두 변수의 독립일 때, Cov(X, Y) = 0
- Cov(X, Y) = 0일 때, 반드시 독립이라고 할 순 없음

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

$$Cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

$Cov(x,y) > 0 (2,3\text{구간})$
$Cov(x,y) < 0 (1,2\text{구간})$



추론통계

상관관계 분석

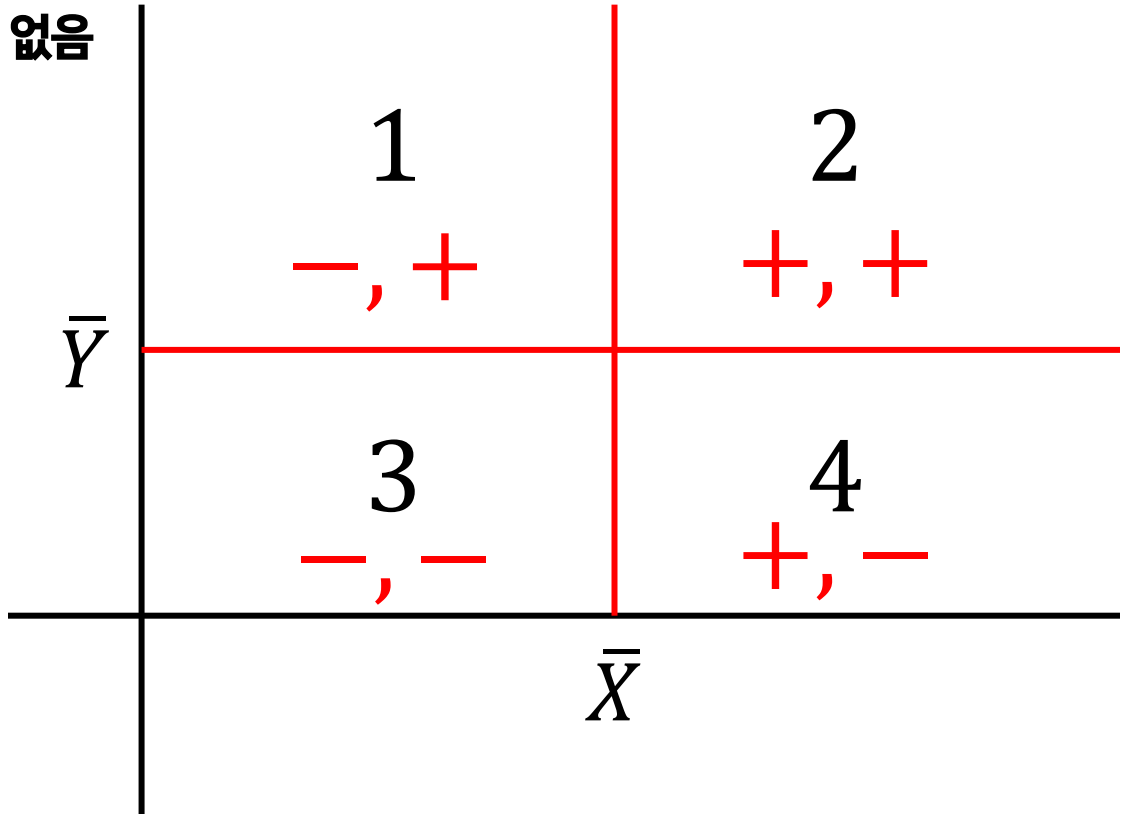
- 공분산(Covariance) : 두개의 변수사이의 관계를 숫자로 알려줄 수 있는 값
- 두 변수의 독립일 때, Cov(X, Y) = 0
- Cov(X, Y) = 0일 때, 반드시 두 변수는 독립이라고 할 순 없음

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

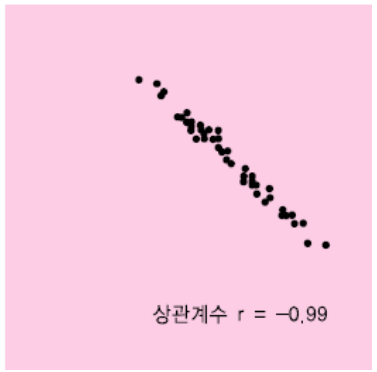
$$Cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})$$

극단치에 영향을 많이 받을 수 있음

$$\sqrt{Var(X)Var(Y)}$$



추론통계



(1) 표본상관계수의 범위는 $-1 \le r \le 1$이다.

(2) $0 < r \le 1$이면 양의 직선적 상관관계를 갖는다.

(3) $-1 \le r < 0$이면 음의 직선적 상관관계를 갖는다.

(4) $r = 0$이면 직선적 상관관계를 갖지 않는다.

공분산 : 두 변수가 얼마나 같이 움직이는지

$$\rho(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \ \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

두 변수의 크기(변동성=흩어진 정도)을 고려해 정규화

공분산을 정규화(normalize)하기 위해 두 변수의 표준편차로 나누어 줍니다. 이렇게 하면 상관계수는 -1에서 1까지의 범위를 가지게 됨

추론통계

```
# 키와 몸무게 데이터 생성
heights <- c(160, 162, 155, 180, 170, 175, 165, 171, 177, 172)
weights <- c(55, 60, 53, 72, 70, 73, 62, 64, 69, 65)

library(psych)
# 피어슨상관계수 도출 및 p-value
result_pearson=corr.test(heights, weights, method="pearson")

result_pearson$p #p-value
result_pearson$r #상관관계 계수
```

추론통계

가설 검정 및 추론통계

T-test, Z-test

A집단	B집단
연속형	연속형
연속형	연속형
연속형	연속형
연속형	연속형

카이제곱 분석

A요인 / B요인	B요소_1	B요소_2
A요소_1	범주형	범주형
A요소_2	범주형	범주형

F-test

A집단	B집단
연속형	연속형
연속형	연속형

ANOVA(일원분산분석)

집단(종속)	B요인(독립)
연속형	범주형_요소1
연속형	범주형_요소2
연속형	범주형_요소3

ANOVA(이원분산분석)

집단(종속)	A요인(독립)	B요인(독립)
연속형	범주형_요소1	범주형_요소1
연속형	범주형_요소2	범주형_요소2
연속형	범주형_요소3	범주형_요소3

추론통계

- 추론통계 : 모집단 전체를 조사하기는 어렵기 때문에 일부 표본만 뽑아서 분석함
  - 표본 데이터를 이용해 모집단을 일반화하는 것
  - 차이가 있는가 없는가? 라는 통계적 유의성을 판단
  - 모집단 간 차이 유무를 설명/검정
- 머신러닝 : 데이터로부터 패턴을 학습하고, 새로운 데이터에 대해 예측/분류/생성 등을 수행하는 방법
  - 추론통계 보다 많은 데이터가 필요함
  - 변수 간 관계를 직접 정의하지 않고, 모델이 스스로 패턴을 학습
  - 새로운 데이터에서 얼마나 잘 맞추는가? 라는 실용적 성능이 중요 → 설명보다는 예측력이 더 강조
  - 새로운 데이터 예측/분류