Part.02

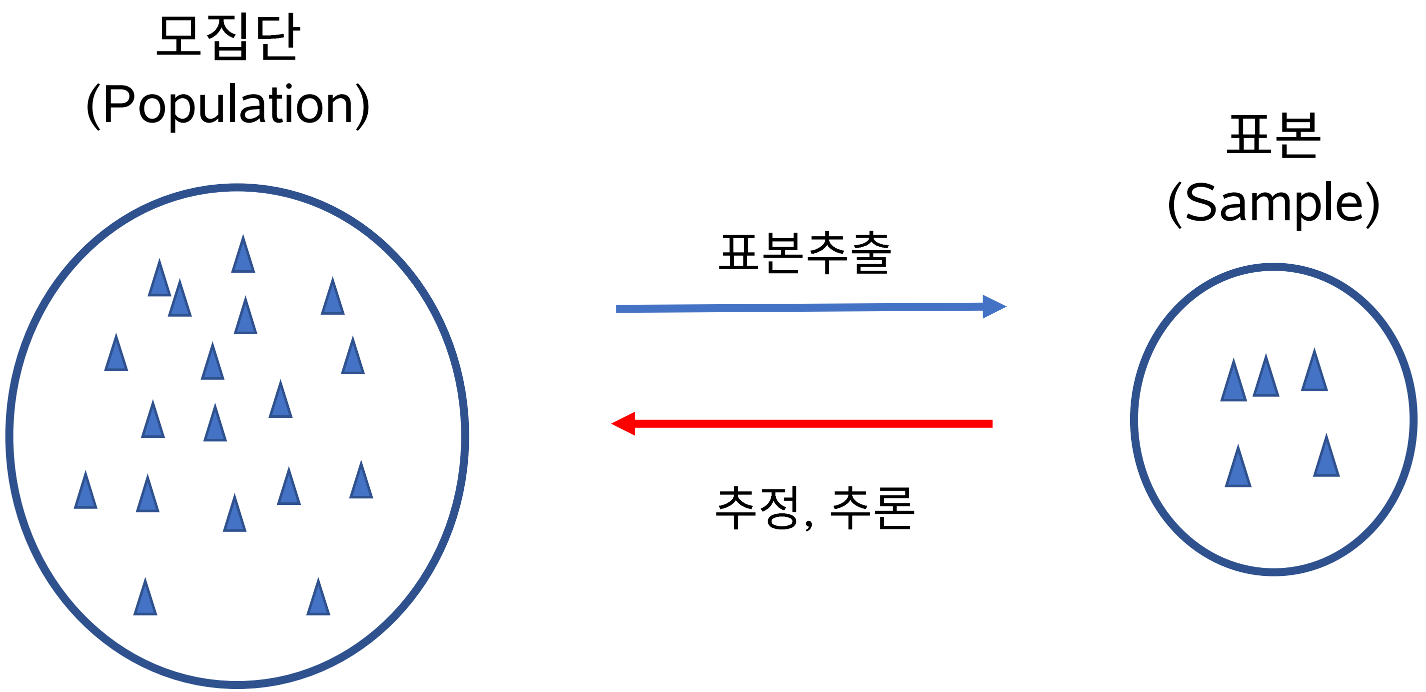
회귀분석

회귀분석을 위한 통계 수학적 개념이해 – 통계학 기초

FASTCAMPUS
ONLINE
머신러닝과 데이터분석 A-Z

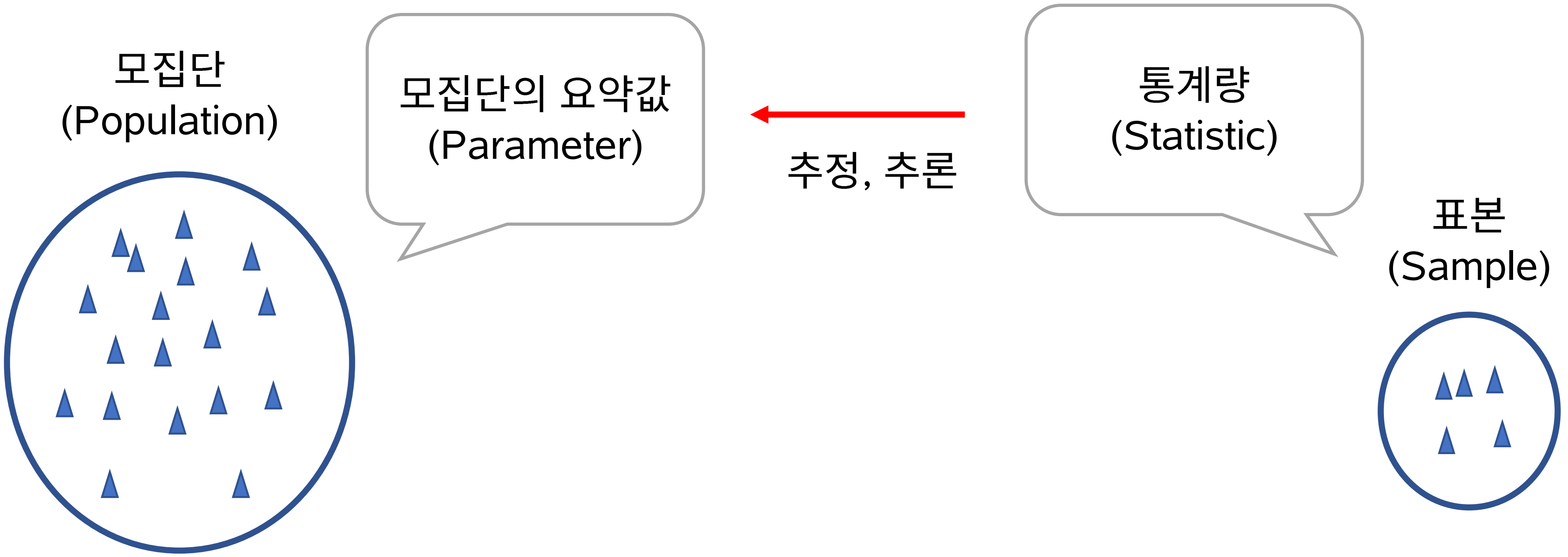
강사. 김강진

I 통계학이란?



- 모집단 (Population): 연구의 대상이 되는 모든 개체들을 모은 집합.
  - 일반적으로 시간적, 공간적 제약으로 인해 모집단 전체를 대상으로한 분석은 불가능.
- 표본 (Sample): 모집단의 일부분의 관측값들.

I 통계학이란?



- 모수 (Parameter): 수치로 표현되는 모집단의 특성.
- 통계량 (Statistic): 표본의 관측값들에 의해서 결정되는 양.

I 자료의 종류

- 수치형 (양적자료)
  - 연속형 (예: 몸무게, 키)
  - 이산형 (예: 전화 통화 수)

- 범주형 (질적자료)
  - 순위형 (예: 학점)
  - 명목형 (예: 성별)

I 자료의 종류

반응변수	설명변수	
	범주형	연속형
범주형 (이분형)	범주형 자료분석 (카이스퀘어 검정)	로지스틱 회귀분석
연속형	분산분석	회귀분석

I 자료의 요약 – 그림, 표

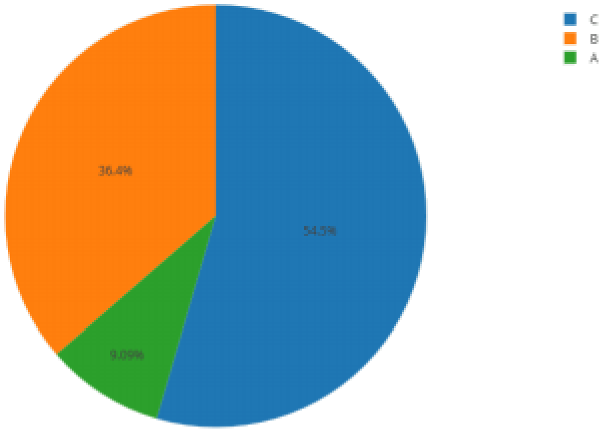
▪ 범주형 자료

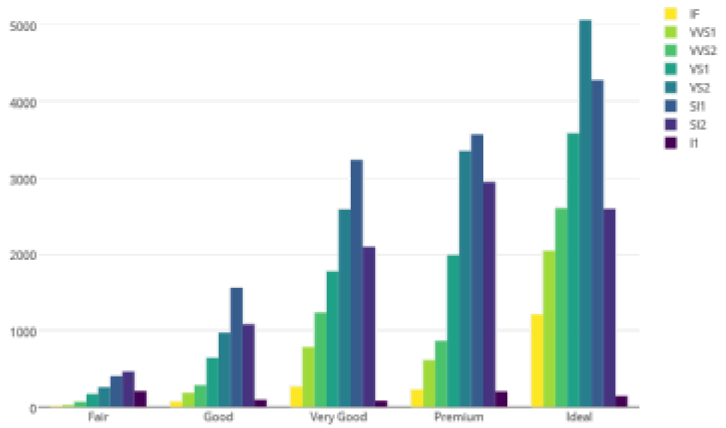
• 도수 분포표

체중(kg)	학생 수(명)	계급값(kg)
40 ~ 50	2	45
50 ~ 60	4	55
60 ~ 70	2	65
70 ~ 80	1	75

45kg이 2명
55kg이 4명
65kg이 2명
75kg이 1명

• 막대 / 원형 그래프





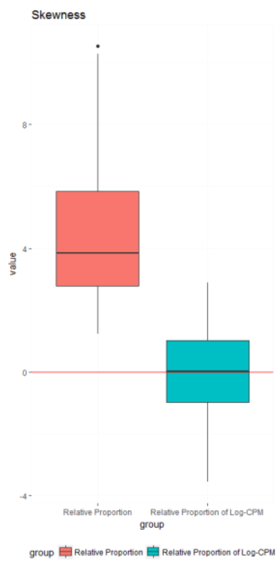
FAST CAMPUS ONLINE
김강진 강사.

이미지출처: https://datamarket.kr

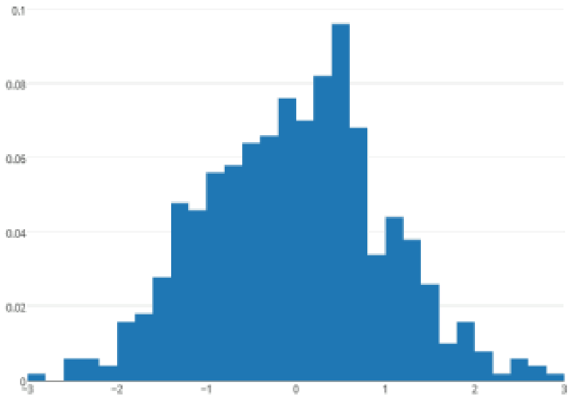
Life Changing Education

I 자료의 요약 – 그림, 표

- 연속형 자료
  - Box plot



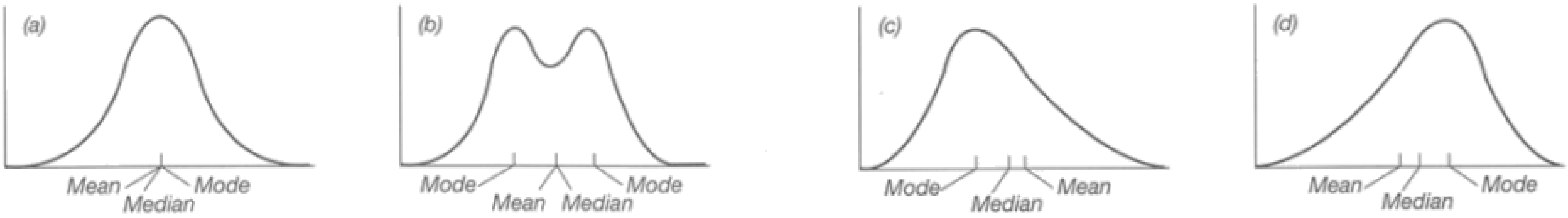
  - 히스토그램 (Histogram)



이미지출처: https://datamarket.kr

I 자료의 요약 - 수치

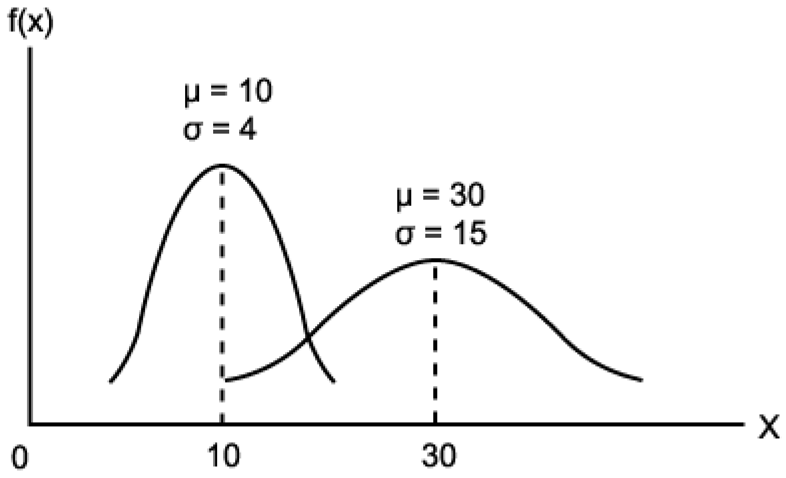
- 모집단 개체의 수: N
- 중심 경향값 (대표값)
  - 평균 (Mean): $\mu = \frac{x_1 + \ldots + x_N}{N} = \frac{\sum_{i=1}^{N} x_i}{N}$.
  - 중앙값 (Median): 크기순으로 정렬시켜 중앙에 위치한 값.
  - 최빈값 (Mode): 가장 자주 나오는 값.



[이미지출처: https://datamarket.kr](https://datamarket.kr)

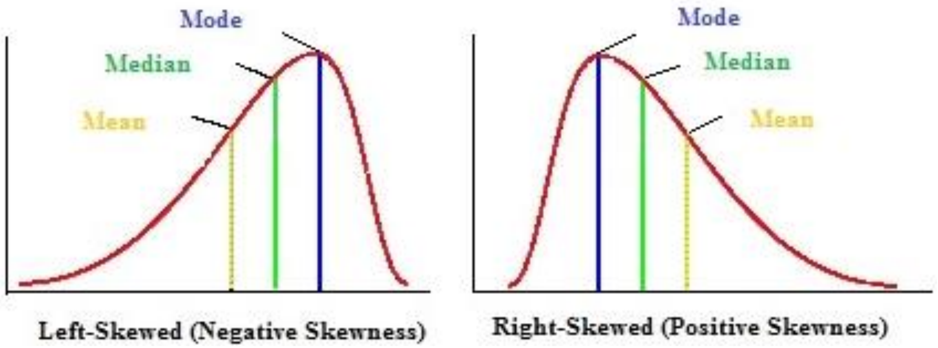
| 자료의 요약 - 수치

- 모집단 개체의 수: N
- 산포도 (퍼진 정도)
  - 분산 (Variance): $\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$
  - 사분위수 범위 (Inter quartile range)
    - 전체 관측값을 크기순으로 정렬했을 때 중앙에 위치한 50%의 관측치가 가지는 범위.
- 정규분포
  - 자연과학 현상을 설명할 때 가장 널리 쓰이는 분포.
  - 위치는 평균에 의해, 모양은 분산에 의해 결정.

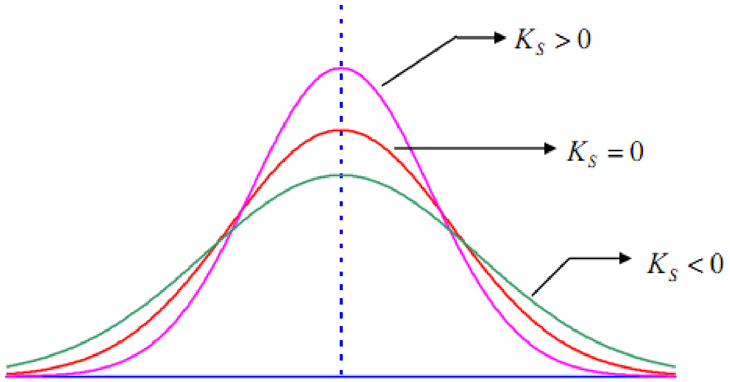


I 자료의 요약 - 수치

- 분포도
  - 왜도 (Skewness)
    - 분포의 비대칭 정도
    - Left-skewed를
      Negative skewed로 표현하기도 함.

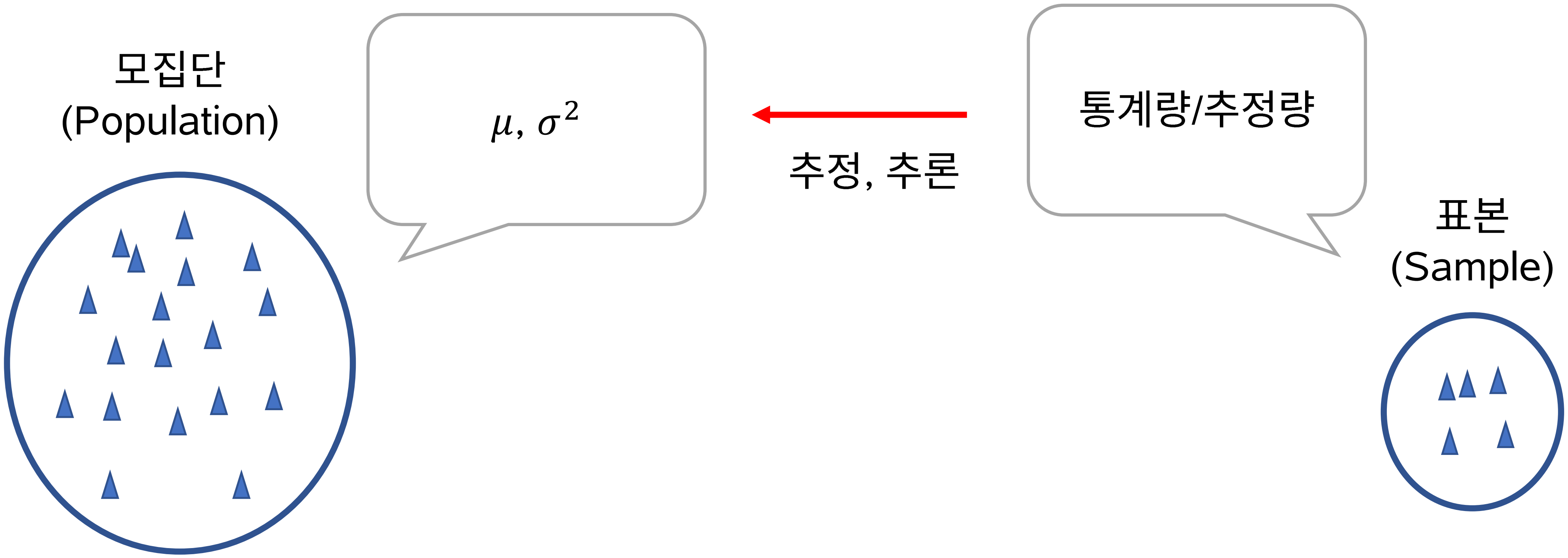


  - 첨도 (Kurtosis)
    - 분포의 꼬리 부분의 비중에 대한 측도
    - $K_s = 0$
      - 뾰족한 정도가 정규분포와 동일



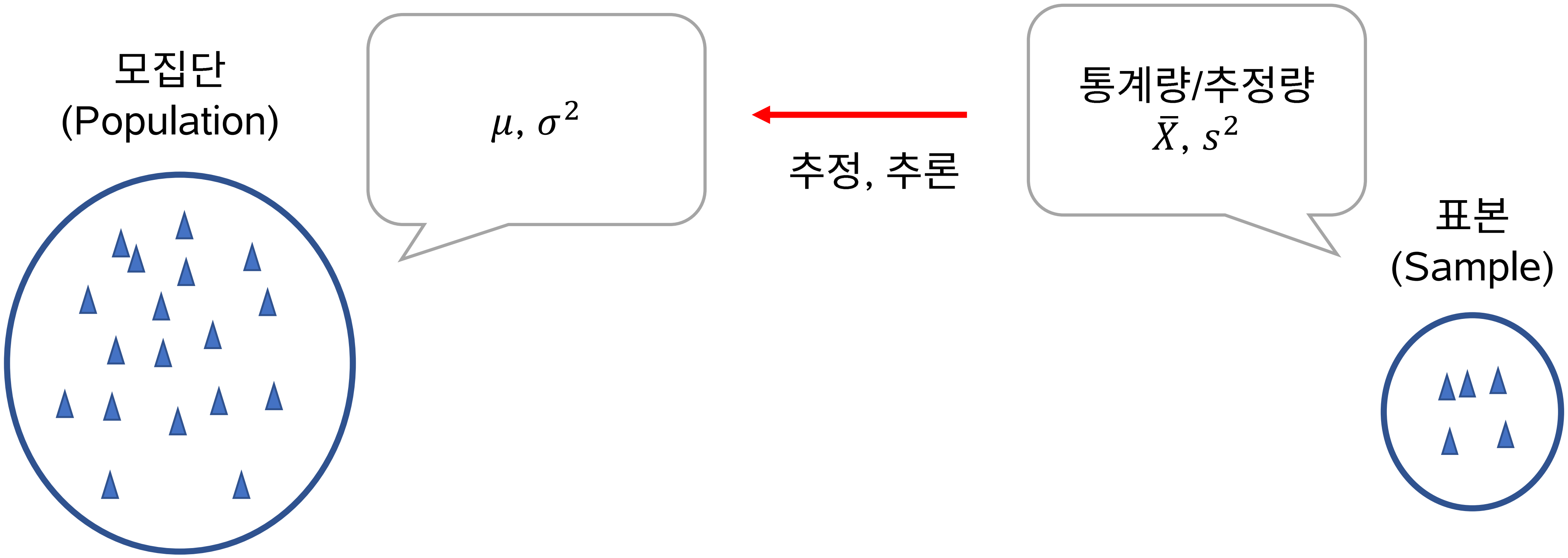
이미지출처: https://www.statisticshowto.datasciencecentral.com/pearson-mode-skewness/

I 통계학이란? - 리뷰



- 모수 (Parameter): 수치로 표현되는 모집단의 특성.
- 통계량 (Statistic): 표본의 관측값들에 의해서 결정되는 양.
- 추정량 (Estimator): 모수를 추정하고자 하는 목적을 지닌 통계량.

통계량, 추정량



- 추정량의 종류 (표본 관측치의 개수: n)
  - 표본평균: $\bar{X} = \frac{\sum_{i=1}^{n} x_i}{n}$
  - 표본분산 (Sample variance) : $s^2 = \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n-1}$

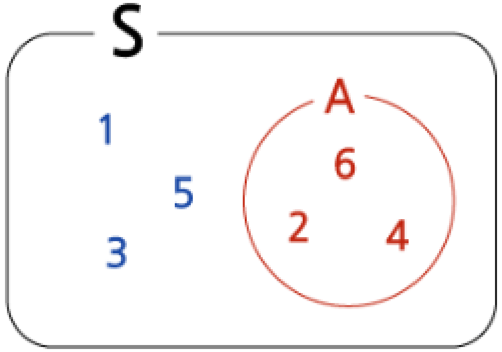
I 확률

- 확률실험 (Random experiment): 다음과 같은 속성을 지닌 관찰이나 인위적인 실험
  - 실험의 결과는 미리 알 수 없다.
  - 실험에서 일어날 수 있는 모든 결과는 사전에 알려져 있다.
  - 이론적으로는 실험을 반복할 수 있다.

- 표본공간 (Sample space): 모든 결과들의 모임.
- 근원사건 (Sample outcome) : 표본 공간의 원소.
- 사건 (Event): 표본 공간의 부분집합. 근원사건의 집합.
  - 배반 사건 (Mutually exclusive events): 서로 교집합이 공집합인 사건.

이미지출처: https://m.blog.naver.com/PostView.nhn?blogId=vollollov&logNo=220964820415&proxyReferer=https%3A%2F%2Fwww.google.com%2F

| 확률



$$P(A) = \frac{n(A)}{n(S)} = \frac{3}{6} = \frac{1}{2}$$

- 확률실험 (Random experiment): 주사위를 던지는 시행. 주사위 눈의 숫자로 결과를 표시.
- 표본공간 (Sample space) : $S = \{1,2,3,4,5,6\}$
- 근원사건 (Sample outcome) : $1, \ldots, 6$
- 사건 (Event) : 짝수가 나오는 사건 $A = \{2,4,6\}$

이미지출처: https://m.blog.naver.com/PostView.nhn?blogId=vollollov&logNo=220964820415&proxyReferer=https%3A%2F%2Fwww.google.com%2F,
https://dictionary.cambridge.org/ko/images/thumb/dice_noun_002_10624.jpg?version=4.0.91

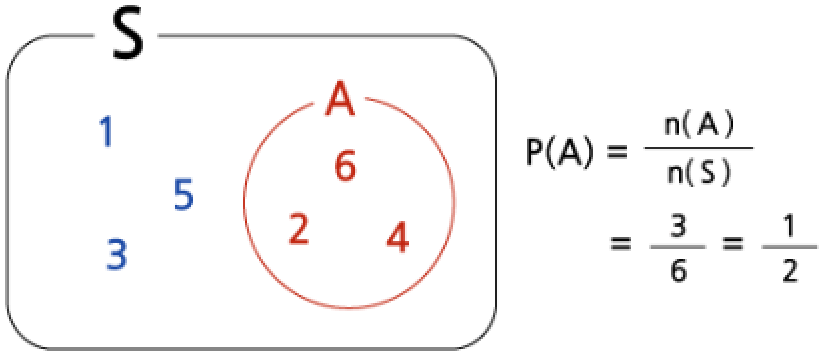
| 확률

- 확률실험 (Random experiment) : 두 동전을 던지는 시행. H,T의 쌍으로 결과를 표시.
- 표본공간 (Sample space) : $S = \{(H,H),(H,T),(T,H),(T,T)\}$
- 근원사건 (Sample outcome) : $(H,H),(H,T),(T,H),(T,T)$
- 사건 (Event) : 앞면이 한번이라도 나오는 사건 $A = \{(H,H),(H,T),(T,H)\}$

이미지출처: https://p5cdn4static.sharpschool.com/UserFiles/Servers/Server_3061305/Image/Lombard/Two%20Coins.jpg

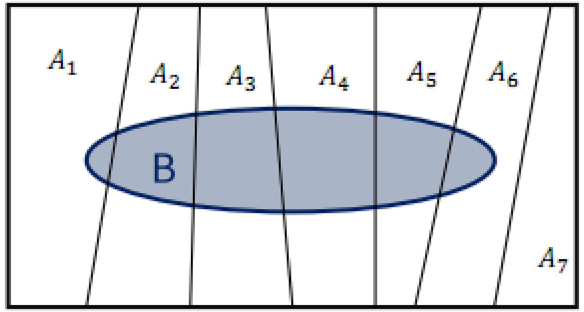
| 확률

- 확률
  - 어떠한 사건이 일어날 가능성의 정도
    - $P(\{2,4,6\}) = P(A)$로 표기
  - 근원사건이 일어날 가능성이 동일할 때의 계산,
    - $P(A) = \frac{|A|}{|S|} = \frac{3}{6} = \frac{1}{2}$
  - 확률의 공리
    - $0 \leq P(A) \leq 1$
    - $P(S) = 1$
    - 어떠한 사건들(A_i, i=1,…,n)이 서로 배반사건일 때, 이 사건들의 합사건의 확률은 각각의 사건이 일어날 확률의 합과 같다.
      - $P(\bigcup_{i=1,\dots,n} A_i) = \sum_{i=1}^{n} P(A_i)$

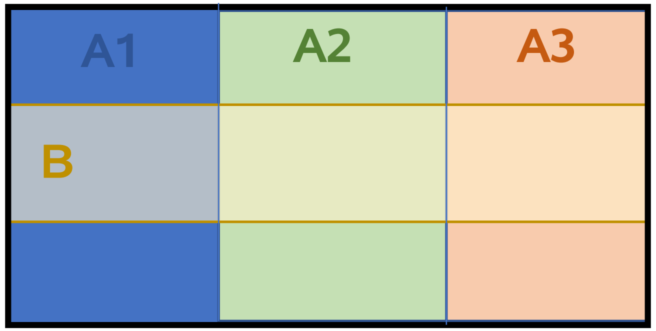


I 확률

- 조건부 확률
  - 사건 B에 대한 정보가 주어졌을 때 사건 A의 교정된 확률
  - B가 주어졌을 때 사건 A의 조건부 확률: $P(A|B) = \dfrac{P(A \cap B)}{P(B)}$



- 독립
  - 사건 A와 B가 서로에게 아무런 영향을 미치지 않을 때.
  - $P(A|B) = P(A)$, $P(B|A) = P(B)$
    - $P(A \cap B) = P(A)P(B)$



I 확률

- 확률변수
  - 각각의 근원사건들에 실수값을 대응시키는 함수.
  - 예) 두 쌍의 동전을 던지는 확률 실험에서, X: 동전 앞면의 개수.
  - $X\big((H,H)\big) = 2$
  - $X\big((H,T)\big) = 1$
  - $X\big((T,H)\big) = 1$
  - $X\big((T,T)\big) = 0$

I 확률

- 확률분포
  - 확률변수에서 확률값으로의 함수. 주로 $f(x)$로 표기.
  - $f(2) = P(X = 2) = P(\{(H, H)\}) = \frac{1}{4}$
  - $f(1) = P(X = 1) = P(\{(H, T), (T, H)\}) = \frac{2}{4} = \frac{1}{2}$
  - $f(0) = P(X = 0) = P(\{(T, T)\}) = \frac{1}{4}$

| 확률

- 확률변수의 기대값
  - 확률변수의 중심 경향값. 흔히 평균이라 칭함.
  - $E(X) = \mu = \sum_{i=1}^{n} x_i f(x_i)$

- 확률변수의 분산
  - $Var(X) = E(X - \mu)^2 = \sum_{i=1}^{n} (x_i - \mu)^2 \cdot f(x_i)$

| 확률

- 공분산
  - $Cov(X,Y) = E(X-\mu_X)(Y-\mu_Y) = \sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)f(x_i, y_i)$
  - 두개의 확률변수 X, Y가 상호 어떤 관계를 가지며 변화하는가를 나타낸 척도
  - X, Y가 독립이면 $Cov(X,Y)$ =0

- 상관계수
  - $\rho = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}, -1 \le \rho \le 1$
  - 공분산은 X, Y 단위의 크기에 영향을 받음.
  - 상관계수는 공분산을 단위화한 값.

I 이산형 확률분포

- 베르누이 시행
  - 실험의 결과의 범주가 2가지인 경우 (성공/실패)
  - $X = 1$ (성공) / $X = 0$ (실패)
    - $f(x) = p^x(1-p)^{1-x}$
  - 예) 앞면이 성공인 동전 던지기

- 이항분포
  - 성공확률이 p인 베르누이 시행을 독립적으로 n번 시행했을 때 성공한 횟수의 분포
    - $f(x) = \frac{n!}{x!(n-x)!} \cdot p^x(1-p)^{n-x}$
      - $n \geq x \geq 0$, 정수
  - 예) 동전 n번 던져 앞면의 횟수

I 이산형 확률분포

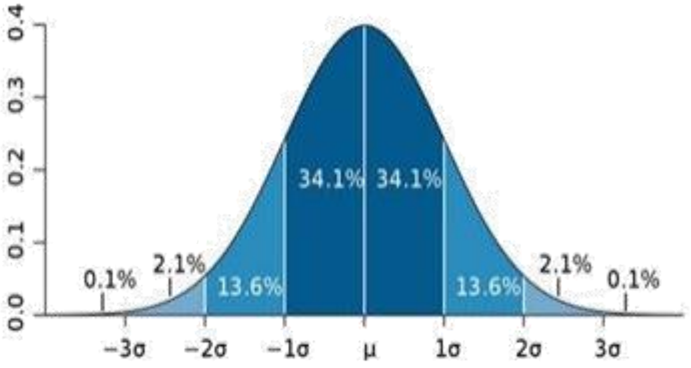
- 다항분포
  - 다항시행: 1회의 시행결과로 나올 수 있는 범주가 3개 이상이 되는 확률 시험.
  - K개 범주의 다항 시행을 n번 반복했을 때, 각 범주가 나타나는 횟수의 분포
    - $f(x_1, \dots, x_K) = \frac{n!}{x_1! \dots x_K!} p_1^{x_1} \dots p_K^{x_K}$
      - $x_K = \left(n - \sum_{k=1}^{K-1} x_k\right), p_K = 1 - \sum_{k=1}^{K-1} p_k, 0 \le x_k \le n,$ 정수
  - 예) 주사위 n번 던져 각 눈이 나온 횟수

I 이산형 확률분포

- 포아송분포
  - 주어진 단위 구간 내에 평균적으로 발생하는 사건의 횟수가 정해져 있을 때, 동일 단위에서의 발생 횟수.
    - 사건의 평균 발생횟수는 단위 구간에 비례.
    - 두개 이상의 사건이 동시에 발생할 확률은 0에 가깝다.
    - 어떤 단위구간의 사건의 발생은 다른 단위 구간의 발생으로부터 독립적.
  - 평균이 μ 인 포아송 분포
    - $f(x) = \frac{\mu^x e^{-\mu}}{x!}$
    - $x \geq 0$, 정수
  - 예) 1시간동안 걸려온 전화의 수. 100페이지안에 있는 오타의 수.

I 연속형 확률분포

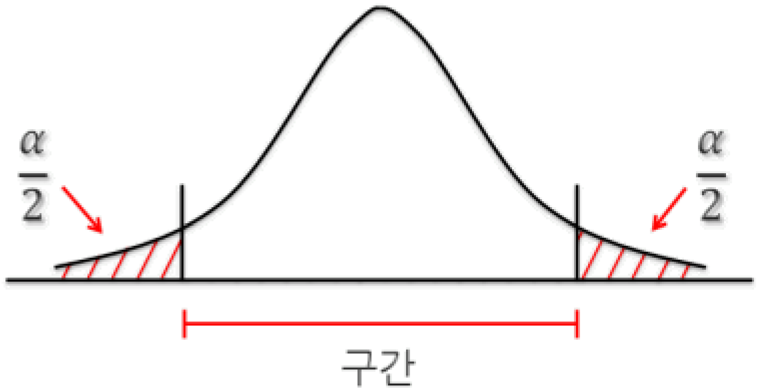
- 지수분포
  - 평균 소요시간이 μ인 사건이 발생하기까지 걸리는 소요시간
    - $f(x) = \frac{1}{\mu} e^{-\frac{1}{\mu}x}$
      - $x \geq 0$
- 정규분포
  - $f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
    - $x \geq 0$



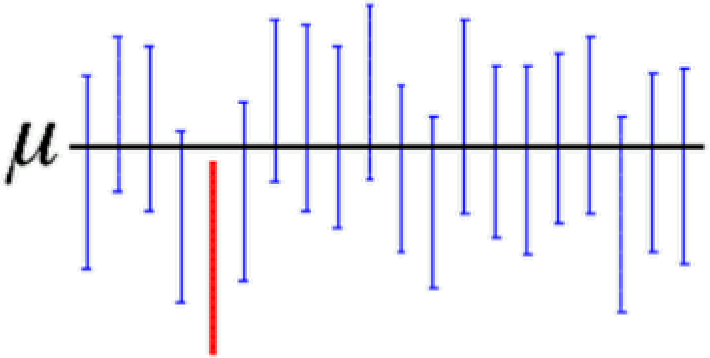
- 표준정규분포
  - 평균이 0이고 분산이 1인 정규분포

l 통계적 추론

- 점추정 (Point estimation)
  - 추정량을 통해 모수를 추정.
  - 예) $\bar{X}$, s^2 -> μ, σ^2
- 구간 추정 (Interval estimation)
  - 일정 신뢰수준 하에서 모수를 포함할 것으로 예상되는 구간을 제시.
  - 신뢰 수준과 구간의 길이는 반비례.



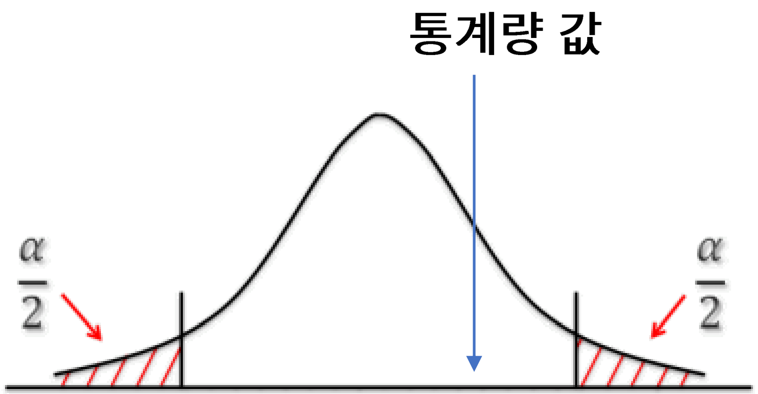
신뢰수준 : $1-\alpha$

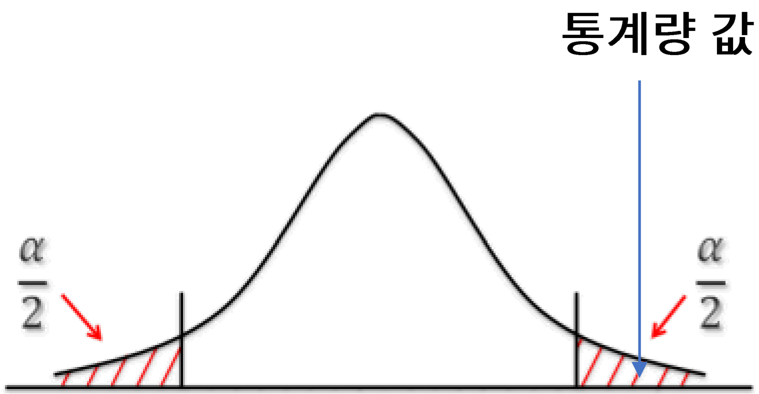


I 통계적 검정

- 대립가설 (H1)
  - 입증하여 주장하고자하는 가설

- 귀무가설 (H0)
  - 대립가설의 반대가설.
  - 귀무가설이 아니라는 충분한 증거를 데이터로부터 보임으로써 대립가설을 입증.
  - 귀무가설 하에서 통계량의 분포를 아는 것이 검정의 핵심.





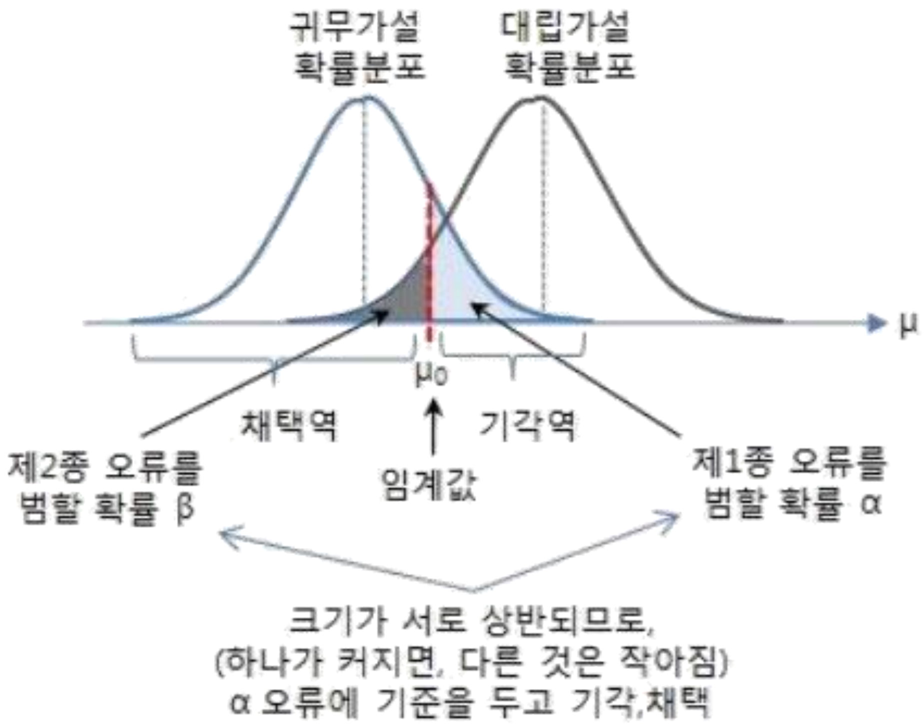
ㅣ오류의 종류

- 1종 오류
  - 귀무가설이 맞을 때, 귀무가설을 기각하는 오류.

- 2종 오류
  - 귀무가설이 틀렸을 때 귀무가설을 기각하지 않는 오류.

I 검정통계량, 기각역

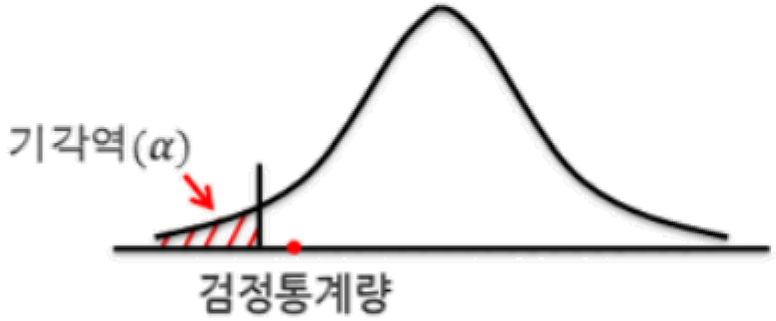
- 검정 통계량
  - 표본에서 구해낼 수 있는 함수. 이 값을 기준으로 귀무가설 기각여부를 결정.
- 기각역
  - 검정통계량이 취하는 구간 중 귀무가설을 기각하는 구간.

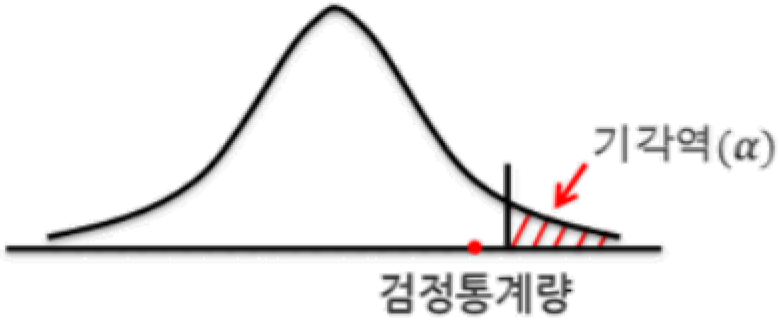


I 검정통계량, 기각역

- 단측검정

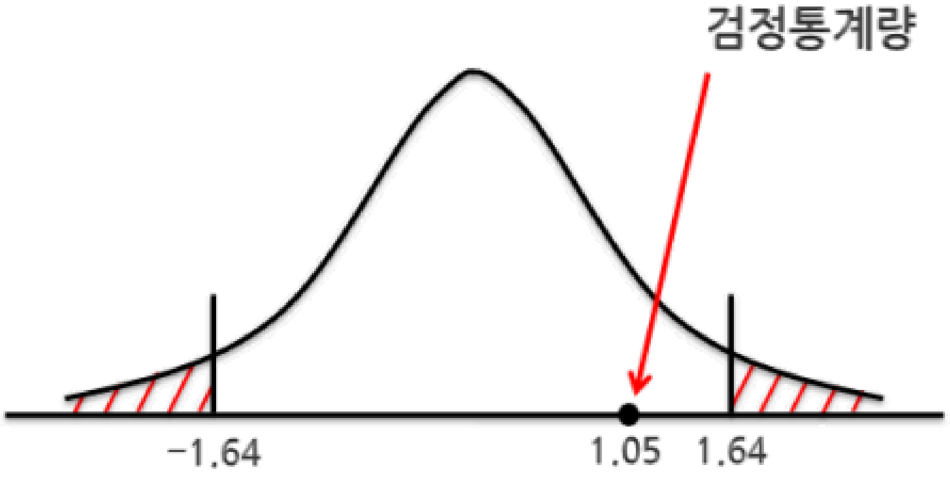
$H_1 : \mu > \mu_0$





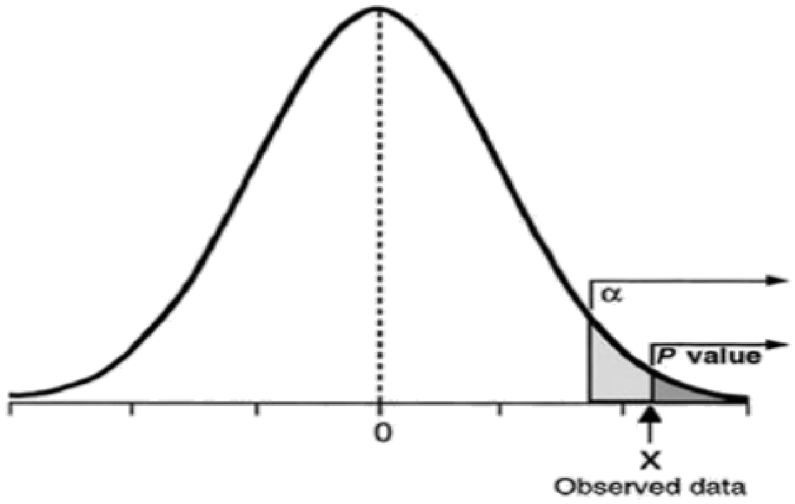
- 양측검정

$H_1 : \mu \neq \mu_0$



I 유의확률

- 유의확률 (P-value)
  - 주어진 검정통계량값을 기준으로 해당 값보다 대립가설을 더 선호하는 검정통계량 값이 나올 확률.
  - 이 값이 유의수준보다 낮으면 귀무가설을 기각.

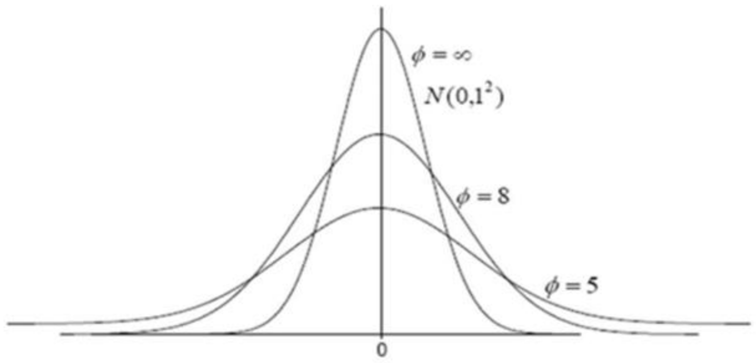


P-value < α	H_0을 기각할 수 있다.
P-value > α	H_0을 기각할 수 없다.

I 검정통계량과 관련된 분포

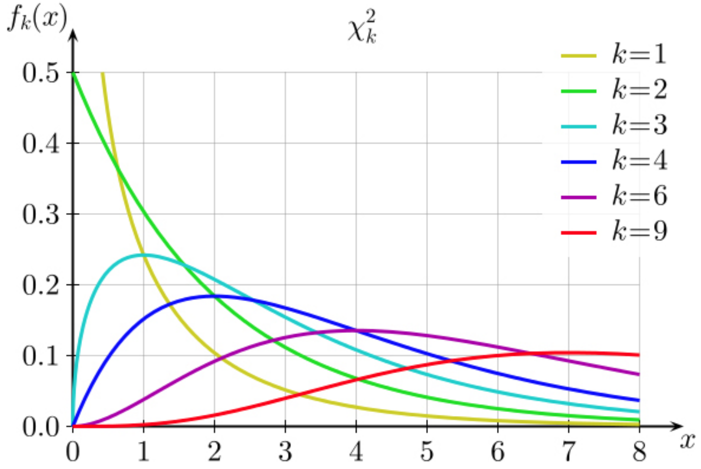
- Z 통계량
  - 귀무가설: X의 평균이 μ_0이다.
  - $Z = \frac{\bar{X}-\mu_0}{\sqrt{\frac{\sigma^2}{n}}} \sim N(0,1)$
    - 이 때 관측치의 수가 충분하다면 (30개 이상) σ^2을 s^2으로 대체 가능.

- t 분포
  - $\mathrm{t} = \frac{\bar{X}-\mu_0}{\sqrt{\frac{s^2}{n}}} \sim t(n-1)$
  - 자유도가 커질수록 정규분포에 근사



| 검정통계량과 관련된 분포

- 카이제곱 분포
  - $Z \sim N(0,1)$일 때,
    - $Z^2 \sim \chi^2_{(1)}$, $\sum_{i=1}^{k} Z_i^2 \sim \chi^2_{(k)}$
  - $f(x;k) = \frac{1}{2^{k/2}\Gamma(k/2)} x^{k/2-1} e^{-x/2}$
    - $x \geq 0$
  - 확률변수의 제곱합으로 이루어진 통계량



I 검정통계량과 관련된 분포

- F 분포
  - 두 확률변수 V_1, V_2가 자유도 k_1, k_2 이고 서로 독립인 카이제곱 분포를 따를 때,
  - $F = \frac{V_1/k_1}{V_2/k_2} \sim F(k_1, k_2)$
  - 확률변수의 제곱합을 관측치로 나눈 것의 비율로 이루어진 통계량.