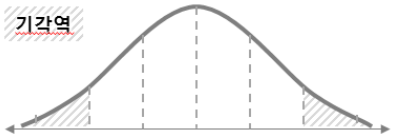
Bayesian (4)

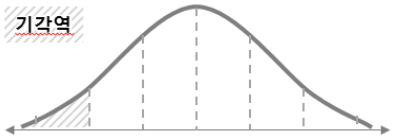
Statistical Hypothesis

- 빅데이터에서의 추론통계 (모집단에서 샘플링한 표본으로 모집단 특성 추론 → 결과 신뢰성 검정)
  - 표본 = 내가 가지고 있는 데이터 전체
  - 모집단 = 현실 세계 전체의 데이터나 미래에 대한 데이터
- 가설 검정의 절차

- 가설
  - **귀무가설** (null hypothesis; H0): 일반적으로 맞다고 가정하는 가설. “차이가 없다”, “영향력이 없다”, “연관성 없다”, “효과 없다”
  - **대립가설** (alternative hypothesis; H1): 새롭게 맞다고 증명하려는 가설. “차이가 있다”, “영향력 있다”, “연관성 있다”, “효과 있다”
- 검정 방법
  - 양측검정: 대립가설이 “~가 아니다 (크거나 작다)”인 경우 사용

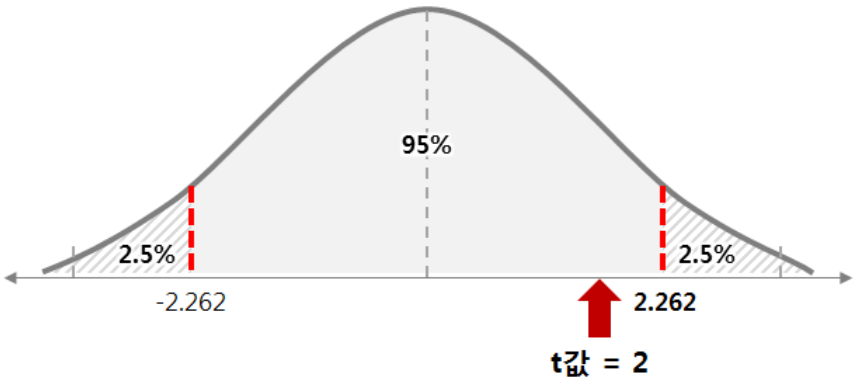


  - 단측검정: 대립가설이 “~보다 작다 or 크다”인 경우 사용



- 신뢰수준 (confidence level): 가설 검정 시 얼마나 빡빡하게 할건지 결정하는 수준. 일반적으로는 95%, 연구는 99%, 설문조사는 90%
- 유의수준 (significance level): 1-신뢰수준 (밑에 그림처럼 양측검정인 경우 기각역=유의수준/2)

- 검정통계량 (test statistic): 가설을 검정하기 위한 기준으로 사용하는 값 (ex. **t-값**). 이 값이 확률분포 상에 어디에 위치하는지에 따라 귀무가설 기각
- 유의확률 (p-value): 귀무가설의 신뢰구간을 벗어나는 확률, 즉 검정통계량에 대한 확률. p-value<기각역 이어야 귀무가설 기각 가능



t값이 기각역에 속하지 않음 → 귀무가설 기각X

T-test (t 검정) / student t-test

: 평균을 비교하여 두 집단이 같은지 다른지 비교하는 가설 검정

- 수행 조건

  1) 표본이 독립 (독립적이지 않다면 paired t-test)

  2) 수집된 데이터가 정규 분포를 따름

  3) 집단이 2개임 (3개인 경우 **ANOVA**)

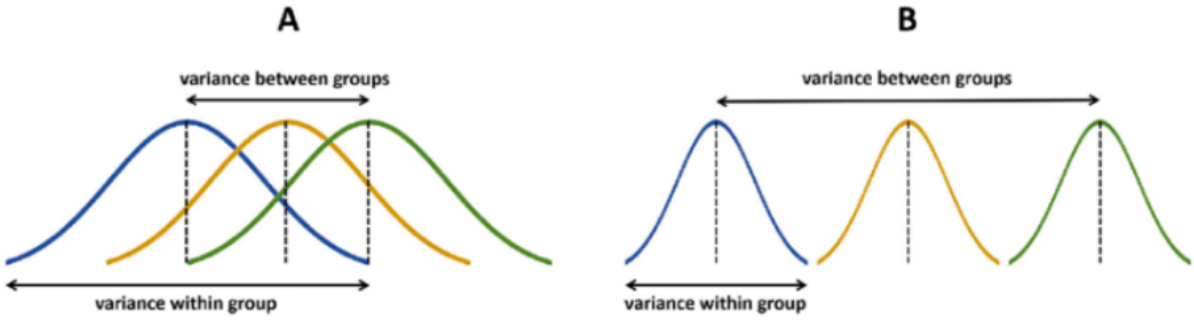
- t값

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

  - m=두 집단 차이의 평균
  - μ=모집단의 평균
  - s=두 집단 차이의 표준편차. $s/\sqrt{n}$=표준오차

- 양측검정인지 단측검정인지 판별한 후, t 분포표로 자유도 (표본수 - 1)에 따른 p-value 찾아서 가설 검증

ANOVA (Analysis of Variance)



: 세 집단 이상의 평균을 비교하는 방법 (분산분석), t-test와 원리 같음

- 귀무가설: 전체 그룹 중 하나 이상의 그룹에서 평균의 차이가 난다

Bonferroni (본페로니)

- 사후검정 (posh hoc analysis)의 한 종류
  - ANOVA로 통계적으로 유의하다는 결과 얻으면 (ex. p-value< 0.05) 집단별로 차이가 있다는 사실 도출 가능 but 어떤 집단간에 차이가 있는지 알 수 X → **사후 분석**. 두 그룹씩 짝을 지어 **다중비교 (multiple comparison)**
- 다중비교문제: 비교 횟수가 늘어날수록 우연에 의해 연관성이 있는 것처럼 나올 수 있는 확률이 100%에 가깝게 되는 현상
  - 다중비교 문제로 인해 검정의 p-value 조정 다르게 하는 Turkey, Duncan 등 방법 다수 존재. 그 중 본페로니가 해석이 직관적이고 적용이 간단해 많이 사용됨
  - Bonferroni correction이라고도 함
- p-value 비교하는 집단의 개수로 나누어줌. (ex. 0.05/3=약 0.017) → 검정시 각 군에 대한 p-value를 0.017로 설정
- 단점: 집단 수가 많으면 p-value가 크게 감소한다, 보수적인 결정을 내려 실제 유의한 마커 찾아내기 힘들어진다. 보완: **FDR**

FDR (False Discovery Rate)

구분	H_0 채택	H_0 기각	Total
H_0가 참	U	V	m_0
H_0가 거짓	T	S	m_1
Total	W	R	m

FDR = false positive (V) / total positive (R) (=false positive + true positive)

: 유의하다고 판단한 것 중 실제로 유의하지 않은 것의 비율을 조정하는 방법

Rank	P-value	i/m · α	Significance
1	0.0001	0.0033	S (Significant)
2	0.0004	0.0067	S
3	0.0019	0.0100	S
4	0.0095	0.0133	S
5	0.0201	0.0167	NS (Not Significant)
6	0.0278	0.0200	NS
7	0.0298	0.0233	NS
8	0.0344	0.0267	NS
9	0.0495	0.0300	NS
10	0.3240	0.0333	NS
11	0.4262	0.0367	NS
12	0.5719	0.0400	NS
13	0.6528	0.0433	NS
14	0.7590	0.0467	NS
15	1	0.0500	NS

- 각 가설에 대해 p-value 구한 후 오름차순으로 정렬 (i=rank, m=변수의 개수, α=목표 p-value)
- 기존 단일 가설검정: 1~9번이 중요변수
- Bonferroni: 1~3번이 중요변수 (0.05/15 기준으로)
- FDR: 1~4번이 중요변수 (p-value보다 $i/m * \alpha$이 작은 지점까지)