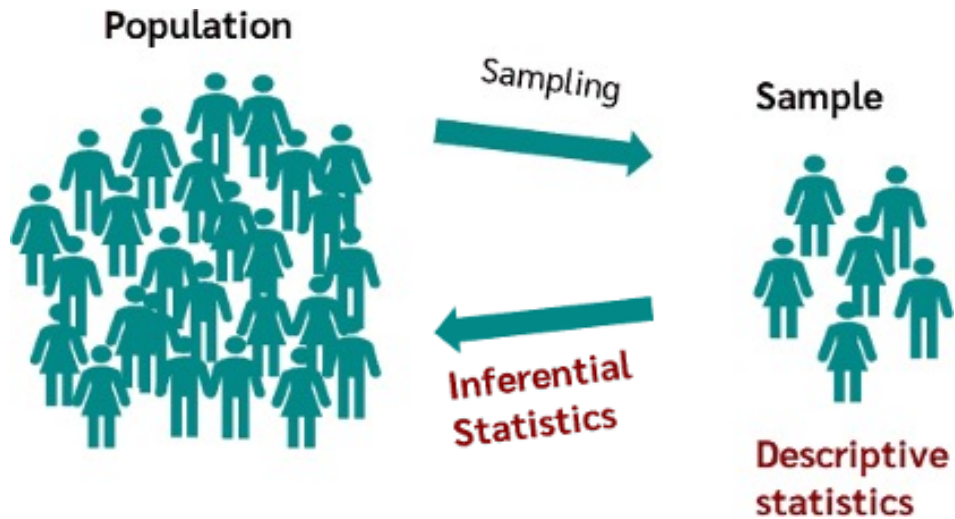


S_02 Hypothesis

Statistical Analysis

Inferential Statistics



- Descriptive statistics (기술 통계): Sample의 통계를 describe 하기 위해 사용
- Inferential statistics (추론 통계): Sample의 통계로부터 Population의 통계를 역으로 추론 (inference) 하기 위해 사용

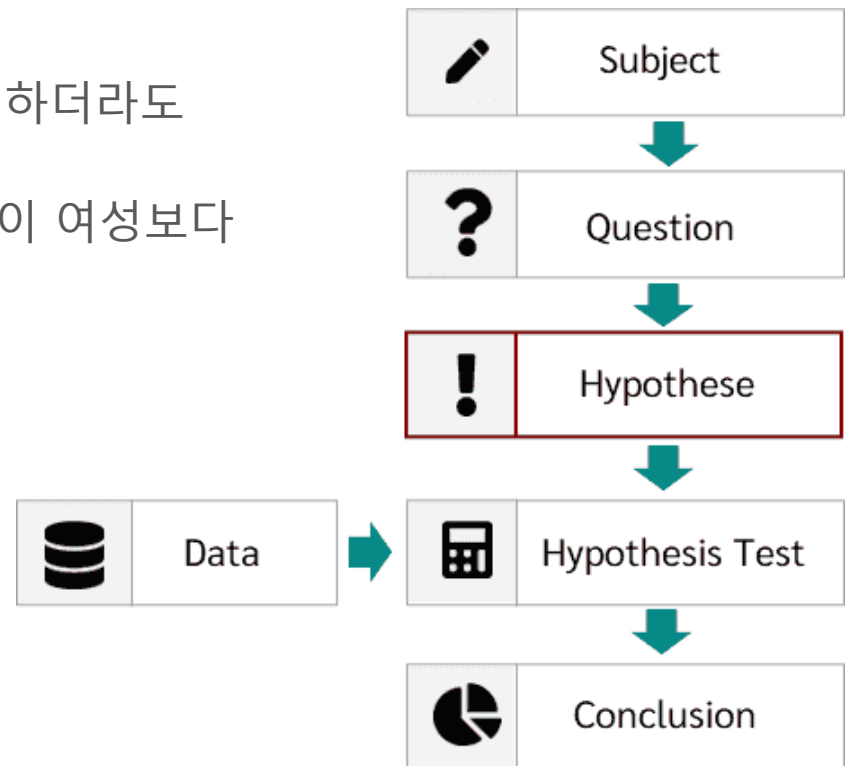
Hypothesis (가설)



- 입증 (proven) 되지도 반증 (disproven) 되지도 않은 가정
- 연구 초기에 만들어짐
- 연구 목표는 결국 Hypothesis를 **reject** (기각) 하거나 **retain** (유지, 기각하지 않음) 하는 것
- Hypothesis를 reject하거나 retain 하기 위해 **hypothesis test**를 사용
- Hypothesis는 variable들 간의 인과 관계 또는 연관성 (association) 에 대한 가정을 의미함

Research Question to Hypothesis

- **Subject:** 한국내의 성별간 수입 격차
- **Research Question (RQ):** 한국에서는 같은 일을 하더라도 성별간 수입의 격차가 존재할까?
- **Hypothesis:** 한국에서는 같은 일을 하더라도 남성이 여성보다 수입이 많다.

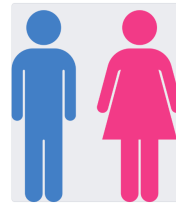


Variables in Hypothesis

- Property of an object or event that can take on different values

- ex) Social science research

- Gender
- Income
- Attitude towards environmental protection



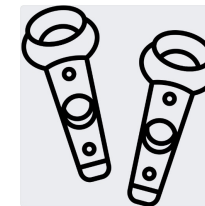
- ex) Medical research

- Body weight
- Smoking status
- Heart rate (심박수)



- ex) Virtual Reality

- Device의 signal, rendering parameters 등
- sense of Experience (immersion, presence, performance, usability, ...)



Independent vs Dependent Variable

- Independent variable (독립변수)
 - = Treatment variable
 - 실험 또는 연구에서 조작하거나 변화시키는 변수
 - 독립 변수를 조작함으로써 종속 변수에 미치는 영향을 관찰
 - 실험에서 원인 또는 자극
 - ex) 교육연구: class type (online, offline)
 - ex) 식물 성장 실험: 비료의 종류 (질소비료, 인산비료)
- Dependent variable (종속변수)
 - = Response variable
 - 독립 변수의 변화에 따라 변화하는 변수
 - 측정하거나 관찰하는 결과
 - ex) 교육연구: 시험점수
 - ex) 식물 성장 실험: 식물의 길이

Variables in XR Research

- Independent Variables
 - use of specific devices
 - rendering parameters
 - simultaneous perception of multiple different stimuli
 - system specific parameters
- Dependent Variables
 - level of immersion
 - level of presence
 - performance parameters
 - usability
 - sense of embodiment
 - cognitive load
 - cybersickness
 - social experience related levels 등

Variables in Correlational Research

- Independent, Dependent variable 이라는 명칭은 Correlational research에는 부적절
- 항상 인과관계인 경우가 아닐 수 있기 때문
- 그러나 대부분의 경우 선행, 후행의 관계는 존재
- Predictor variable (ex. 강우량), Outcome variable (ex. 진흙의 양)
- 넓은 의미로
 - Independent variable \subset Predictor variable
 - Dependent variable \subset Outcome variable

Other Types of Variables

- Control Variables
 - 실험 내내 일정하게 유지되는 변수, 교란 요인을 통제하는데 사용
 - ex) 식물 성장 실험
 - Independent variable: 비료의 종류
 - Dependent variable: 식물의 길이
 - Control variables: 온도, 빛, 물 – 모든 식물에게 동일하게
- Latent Variables
 - 직접적인 측정은 불가하지만 실험내의 다른 대리 (proxy) 변수에 의해 유추 가능
 - ex) 식물 염분 내성 직접 측정은 불가, 염분 첨가하고 식물의 건강 상태를 측정하여 유추
 - ex) immersion 을 biophysiological responses (뇌파, 심박수, 피부전도도, 눈움직임, 호흡) 에서 유추
- Composite Variables
 - 두 개 이상의 variable들을 combine하여 만들어지는 variable
 - 측정할 때가 아니라 data를 분석하면서 만들어 진다.
 - ex) 국어성적, 수학성적, 영어성적의 세 dependent variable들이 합쳐져 개인의 성적을 나타냄

Null Hypothesis vs Alternative Hypothesis

- Null Hypothesis (H_0 : 귀무가설)
 - 두 개 이상의 그룹 간에 어떤 특성과 관련한 **차이가 없다**는 가설
 - ex) 한국의 남성과 여성 근로자의 임금은 차이가 없다.
- Alternative Hypothesis (H_1 : 대립가설)
 - 두 개 이상의 그룹 간에 어떤 특성과 관련한 **차이가 있다**는 가설
 - ex) 한국의 남성과 여성 근로자의 임금은 차이가 있다.
- Hypothesis test에서는 H_0 만 test 됨, 즉, H_0 를 reject할지 retain할지를 결정
- H_0 와 H_1 은 서로 정확히 반대되는 가설이므로 H_0 를 reject하면 H_1 을 retain하는 것임
- Research Hypothesis는 "차이가 있다" 는 것을 보이려는 경우가 많으므로
 - 보통 Research Hypothesis를 H_1 으로 놓고
 - H_0 를 reject할 수 있음을 보인다.

Types of Hypothesis

- Difference hypothesis (차이 가설)
- Correlation hypothesis (상관관계 가설)
- Directional hypothesis (방향성 가설)
- Non-directional hypothesis (비방향성 가설)

Difference Hypotheses

- 두 개 이상의 그룹에 어떤 차이가 있다 (또는 없다) 는 가설
- ex)
 - 남성 그룹이 여성 그룹보다 수입이 높다
 - 흡연자는 비흡연자보다 심장마비 위험이 높다
 - 주당 근무시간에서 한국과 일본은 차이가 없다
- Variables
 - 한 variable은 categorical variable
 - ex) gender (male/female)
 - ex) 흡연여부 (흡연/비흡연)
 - ex) country (한국/일본/중국)
 - 다른 하나의 variable은 최소한 ordinal 이상 (metric)
 - ex) 급여
 - ex) 심장마비 비율
 - ex) 주당 근무 시간



Correlation Hypotheses

- 두 variable간 (ex. height, weight) 의 correlation (상관관계) 에 대한 가설
- ex)
 - 키가 클 수록 체중이 더 나간다
 - 차의 마력이 더 클수록 연비는 떨어진다
 - 수학 성적이 좋을 수록 미래 연봉이 높다
- 최소한 두 개의 ordinally scaled variable들이 관찰 되어야 함
- ex)
 - 키, 체중
 - 차의 마력 수, 차의 연비
 - 수학성적, 미래연봉



Probability Density Function (PDF)

- 연속확률분포에서 특정 구간 내의 값들이 나타날 확률을 표현하는 function: $f(x)$
- $f(x)$ 는 다음 특징을 가진다.
 - $f(x) \geq 0$ for all x
 - 확률변수 X 가 $[a, b]$ 구간에 존재할 확률 (즉, 구간 $[a, b]$ 에서 값이 나타날 확률):

$$P(a \leq X \leq b) = \int_a^b f(x) dx$$

- 전체확률은 1

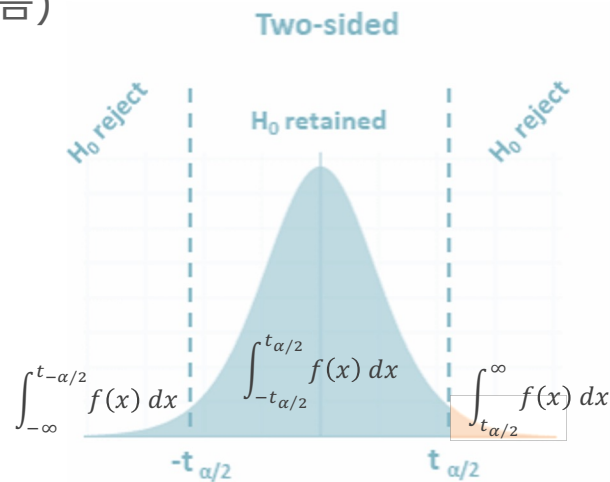
$$\int_{-\infty}^{\infty} f(x) dx = 1$$

- $f(x)$ 는 확률 자체를 나타내지는 않음.
- Example) Normal Distribution (정규분포), 평균: μ , 표준편차: σ

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right)$$

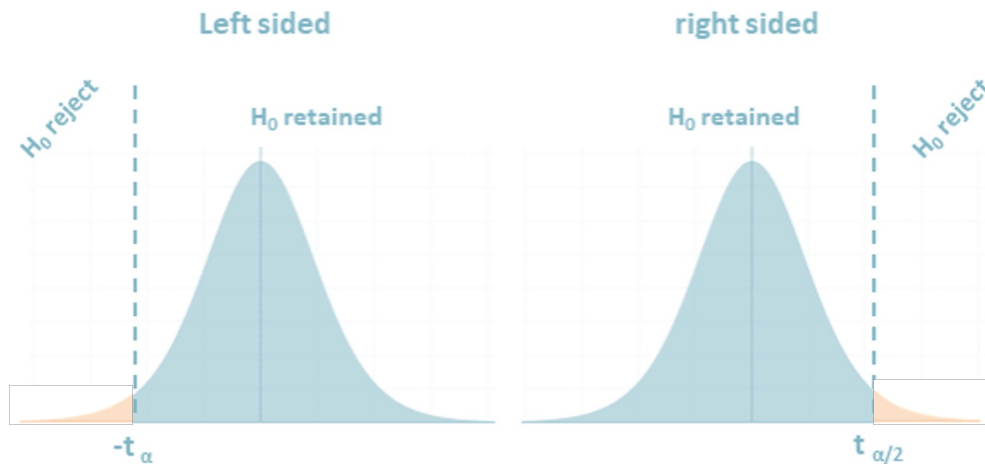
Non-directional Hypotheses

- “차이가 있다 (there is a difference between) ” 는 말이 포함되나, 차이가 있냐 없냐 만이 관심의 대상이고, 어느 쪽이 더 큰지는 상관을 하지 않음
- Difference 또는 Correlation Hypothesis 일 수 있음
- ex)
 - 남성과 여성의 급여에는 차이가 있다 (그러나 누가 더 많이 버는지는 말하지 않음)
 - 흡연자와 비흡연자 간에는 심장마비 위험에 차이가 있다 (그러나 누가 더 위험하다고 말하지는 않음)
 - 키가 차이 나면 몸무게가 차이 난다 (correlation: 키가 크고 작은 중에 어느 쪽이 더 몸무게가 무거운지 말하지 않음)



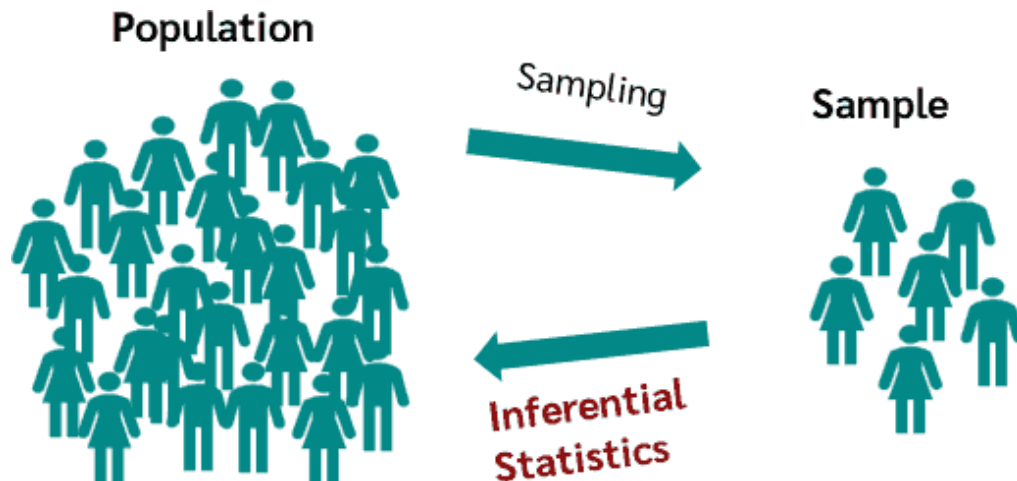
Directional Hypotheses

- Directional Hypotheses (방향성 가설)
 - 일반적으로 가설에 “더 좋다 (better than)” 또는 “더 나쁘다 (worse than)” 이 포함된 경우
 - ex)
 - 남자가 여자보다 수입이 높다
 - 흡연자는 비흡연자보다 심장마비 위험이 높다
 - 자동차의 마력수가 높을 수록 연비는 낮아진다 (correlation)



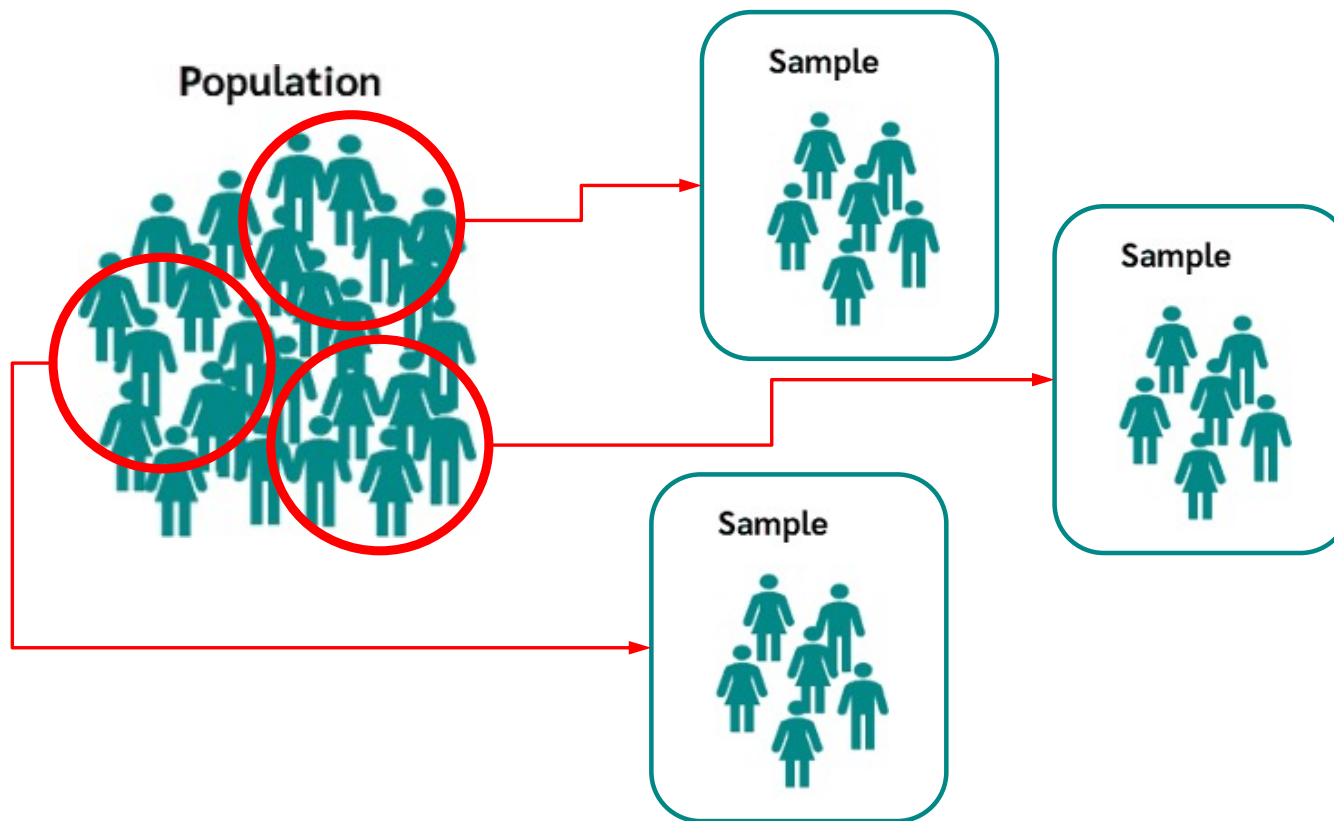
Hypothesis Testing

- Sample의 도움으로 Population에 대한 Hypothesis를 테스트 할 때 사용
- Test의 결과는 H_0 의 reject 또는 retain
- ex) Hypothesis: “오리온 초코파이의 무게는 35g이다”
 - Population: “생산된 모든 초코파이”
 - Sample: n개를 택해 무게의 평균을 낸 후
 - 이를 바탕으로 Population에 대한 가설을 테스트



Types of errors (1/2)

- Hypothesis test는 항상 정확한 것이 아니고 어느 정도의 error가 일어날 가능성이 있음
 - 고려하는 sample이 population의 분포를 충분히 반영하지 못하는 경우가 있기 때문



Types of errors (2/2)

- Error의 종류

Fact	Decision	
	H_0 true	H_0 false
H_0 true	Right	Type 1 error
H_0 false	Type 2 error	Right

- Type 1 Error: H_0 가 실제로 true (즉, H_1 이 false) 인데도 불구하고 H_0 를 reject하는 경우
- Type 2 Error: H_0 가 실제로 false (즉, H_1 이 true) 인데도 불구하고 H_0 를 retain하는 경우

p-Value

- “Population에서 H0 가 true인 Sample을 택하게 될 확률”
- p-value가 충분히 작다는 의미는?
 - Population에서 H0 가 true인 경우의 Sample을 택하게 될 확률이 충분히 작다
 - 즉, 어떻게 sampling해도 그 Sample에서 H0가 true가 될 확률이 충분히 작다
 - 즉, H0는 true가 아닌 것에 가깝다
 - 즉, H0는 reject 될 수 있다
- 보통 H0: 두 data의 평균의 차이가 없다, H1: 두 data의 평균의 차이가 있다 이고, 연구자는 보통 H0를 reject, 즉, H1을 retain하기를 원할 것입니다.
- 따라서 p-value를 구해 그것이 충분히 작다는 것을 보이면 될 것입니다.

Significance Level (α : 유의 수준)

- 일반적으로 p-value의 한계를 충분히 작은 significant level α 로 미리 정해 놓는데,
- α 는 5% (0.05) 또는 1% (0.01) 로 설정해 놓고
- α 를 H_0 의 reject 여부를 결정하는데 사용
 - p-value가 α 보다 작으면 significant level α 에서 H_0 를 reject, 즉 H_1 을 retain 할 수 있음
 - 그렇지 않으면 H_0 를 retain, 즉 H_1 을 reject
- Ex) 백신 A와 B의 효과 비교
 - 100명의 Sample에 A 백신을 접종한 경우, 발병된 개체수는 평균 23명이고
 - 또 다른 100명의 Sample에 B 백신을 접종한 경우, 개체수 평균 32명이 발병하였다.
 - H_1 : 백신 A의 효과가 B보다 좋다 라는 가설을 받아들일 수 있는가?
 - significance level $\alpha = 0.05$ 라 할 때, A와 B의 차이에 대한 p-value가 0.053 이었다 라고하면
 - $0.053 > 0.05$ 이므로
 - H_0 : 백신 A와 B의 효과에 차이가 없다 라는 가설을 retain 할 수 있다.
 - 즉, Sample stat에서 백신 A가 B보다 우수하지만,
 - 이 차이는 significance level $\alpha = 0.05$ 에서 유의미하지 않다.

p-Value and Significance Level

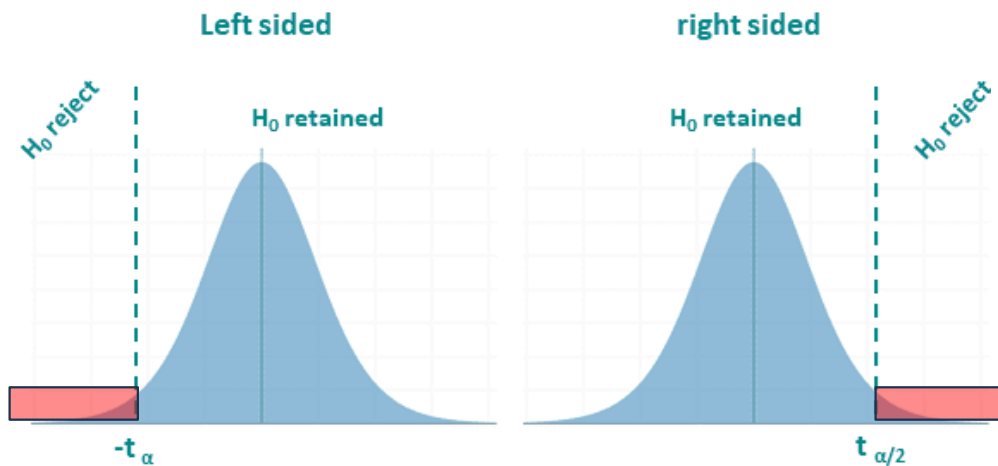
- p-Value ≤ 0.01
 - $\alpha = 1\%$ (0.01) 에서 H0는 reject 될 수 있다. 즉 H1은 retain
 - $\alpha = 1\%$ (0.01) 에서 두 population의 stat value (예: 평균) 의 차이는 "highly significant" (매우 유의미 하다)
- p-Value ≤ 0.05
 - $\alpha = 5\%$ (0.05) 에서 H0는 reject 될 수 있다. 즉 H1은 retain
 - $\alpha = 5\%$ (0.05) 에서 두 population의 stat value의 차이는 "significant" (유의미 하다)
- p-Value > 0.05
 - $\alpha = 5\%$ (0.05) 에서 H0는 retain 될 수 있다. 즉 H1은 reject
 - $\alpha = 5\%$ (0.05) 에서 두 population의 stat value의 차이는 "not-significant" (유의미 하지 않다)

Confidence Level, Confidence Interval, Margin of Error

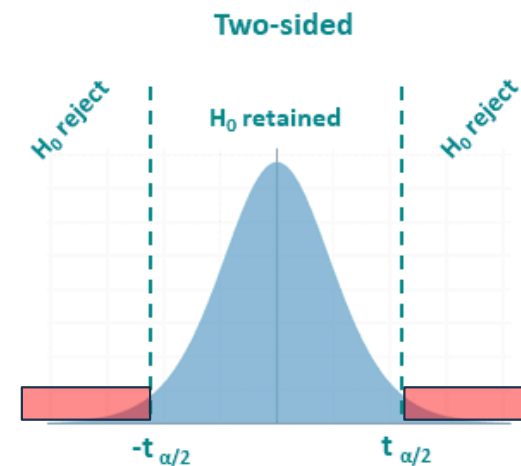
- Confidence Level (CL: 신뢰수준)
 - $1 - \text{Significance Level} = 1 - \alpha$
 - ex) $\alpha = 5\%$ (0.05) 이면 CL은 95%
- Margin of Error (MOE: 오차범위)
 - 추정된 Sample mean이 실제 Population mean에서 벗어날 수 있는 최대 범위
 - $\text{MOE} = \text{SE} * \text{Z-value at Confidence Level}$
 - SE (Standard Error, 표준오차) = $\frac{\sigma}{\sqrt{n}}$
 - Confidence Level에서의 z-value는 95% 일때 1.96, 99% 일때 2.576
- Confidence Interval (CI: 신뢰구간)
 - $\text{Sample mean} \pm \text{MOE}$
- Ex)
 - $n = 1000$ 명에 대한 95% 신뢰 수준의 Sample 조사에서 지지율 45%, 표준편차는 20% 였다.
 - Sample mean: $\mu = 0.45$, Sample standard deviation: $\sigma = 0.2$
 - $\text{MOE} = 1.96 * (0.2/\sqrt{1000}) = 0.012396128427860047$
 - $\text{CI} = 0.45 \pm 0.0124 = [0.4376, 0.4624]$, 또는 [43.76%, 46.24%]

One Tailed vs Two Tailed p-Value (1/2)

- One Tailed (Sided) p-value
 - Directional hypothesis의 경우 두 Sample들 간의 차이가 있느냐 외에도 어느 Sample쪽이 더 큰지 (혹은 작은 지)에 관심이 있음
 - 주어진 direction (크다 또는 작다) 을 만족하는 경우만 counting한 확률
 - Two Tailed p-value를 2로 나눈 값.



One Tailed p-value



Two Tailed p-value

One Tailed vs Two Tailed p-Value (2/2)

- Two Tailed (Sided) p-value
 - Non-Directional hypothesis의 경우 두 Sample들 간의 차이가 있는가에만 관심
 - Direction에 따라 positive와 negative한 경우를 모두 counting
- ex) H0와 H1은 Two Tailed p-value를 고려:
 - H0: Group A와 Group B의 반응시간은 차이가 없다
 - H1: Group A와 Group B의 반응시간은 차이가 있다
 - Two Tailed p-value: 0.04
 - “Group A의 반응시간이 Group B보다 더 길다” 의 p-value는 $0.04 / 2 = 0.02$
 - “Group B의 반응시간이 Group A보다 더 길다” 의 p-value도 $0.04 / 2 = 0.02$
 - Significance level α 도 2로 나누어 판단해야 함
 - $0.02 \leq 0.025$

Hypothesis Test vs Statistical Test

- Hypothesis Test
 - H_0 의 reject (or retain) 여부를 결정하는 test
 - 따라서 두 개 이상의 group (sample) 간의 comparison에 한정된 (difference hypothesis) 의 test를 의미
 - ex) t-test, ANOVA
- Statistical Test
 - Difference 뿐만 아니라 Correlation, Linearity 등 다양한 statistics 측면에서의 test를 망라
 - Hypothesis Test를 포함
 - ex) Regression, Correlation, ...