S_02 Hypothesis

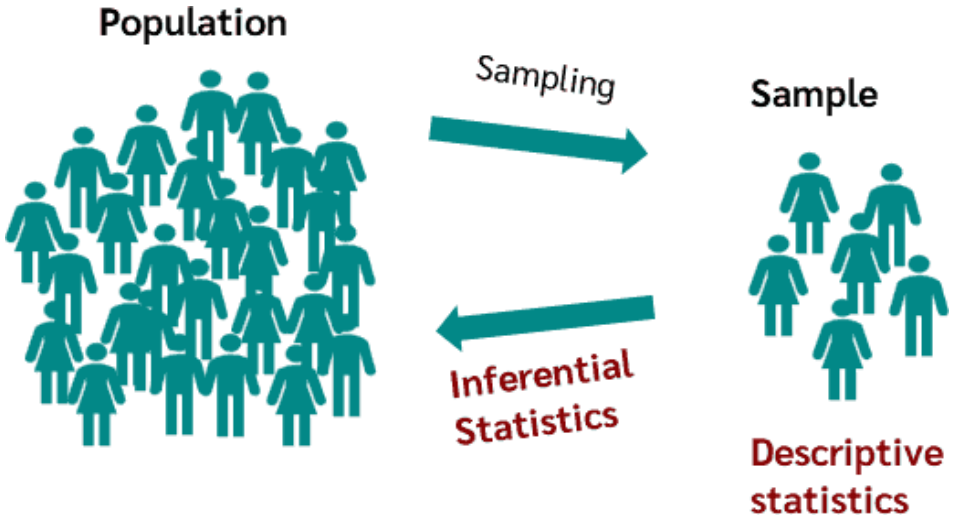
Statistical Analysis

References

- Text and figures from the DATAtab site (https://datatab.net/) and the book "Statistics made easy" published by DATAtab.

- Statistics Page from Scribbr site (https://www.scribbr.com/category/statistics/)

Inferential Statistics



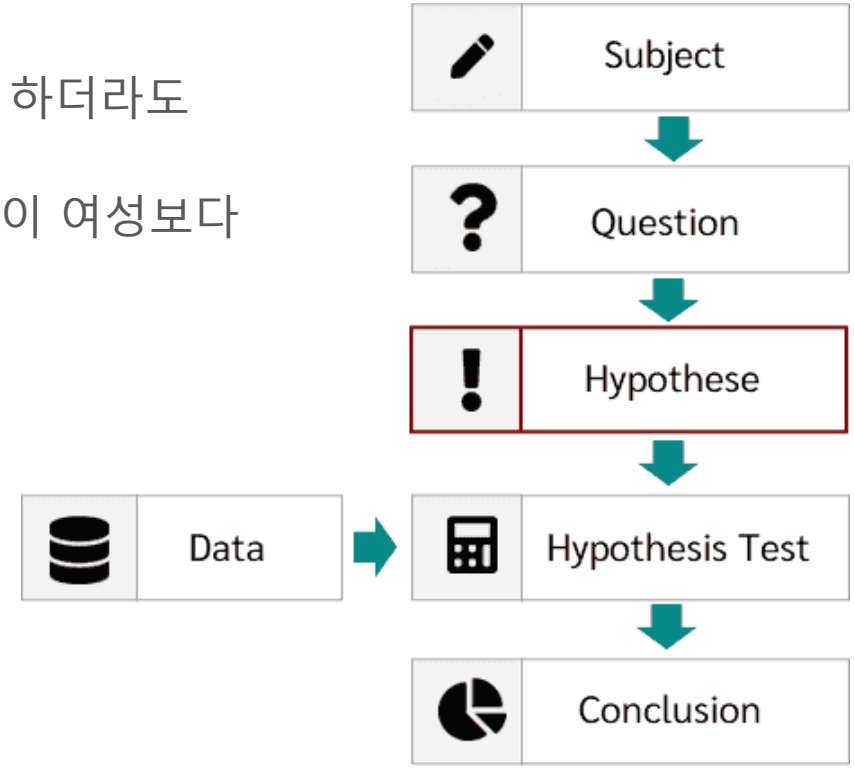
- Descriptive statistics (기술 통계): Sample의 통계를 describe 하기 위해 사용
- Inferential statistics (추론 통계): Sample의 통계로부터 Population의 통계를 역으로 추론 (inference) 하기 위해 사용

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

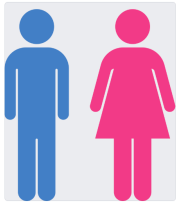
Research Question to Hypothesis

- **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)



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 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 ⊂ Predictor variable
  - Dependent variable ⊂ 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) 식물의 염분 내성은 직접 측정할 수는 없지만, 염분 첨가 실험에서 식물의 건강 상태를 측정하여 유추 가능
- 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 (H0: 귀무가설)
  - 두 개 이상의 그룹 간에 어떤 특성과 관련한 **차이가 없다**는 가설
  - ex) 한국의 남성과 여성 근로자의 임금은 차이가 없다.

- Alternative Hypothesis (H1: 대립가설)
  - 두 개 이상의 그룹 간에 어떤 특성과 관련한 **차이가 있다**는 가설
  - ex) 한국의 남성과 여성 근로자의 임금은 차이가 있다.

- Hypothesis test에서는 H0만 test 됨, 즉, H0를 reject할지 retain할지를 결정
- H0와 H1은 서로 정확히 반대되는 가설이므로 H0를 reject하면 H1을 retain하는 것임
- Research Hypothesis는 H1의 형식이 더 많으나 H0를 보이려는 연구도 있음

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

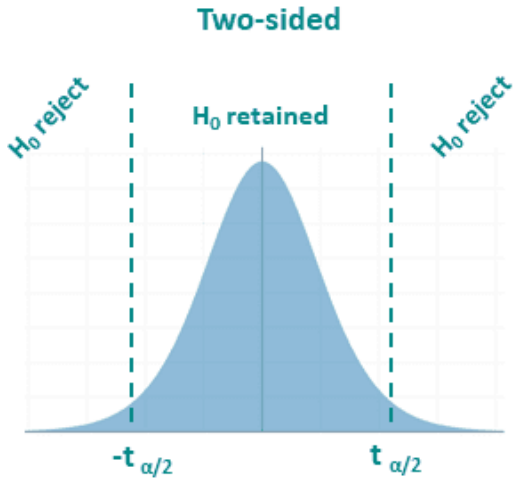
Correlation Hypotheses

- 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

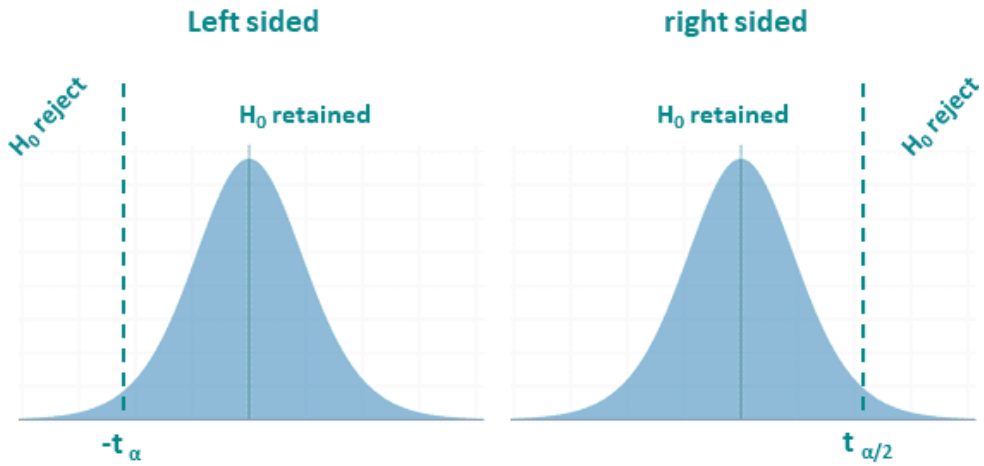
Non-directional Hypotheses

- "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)



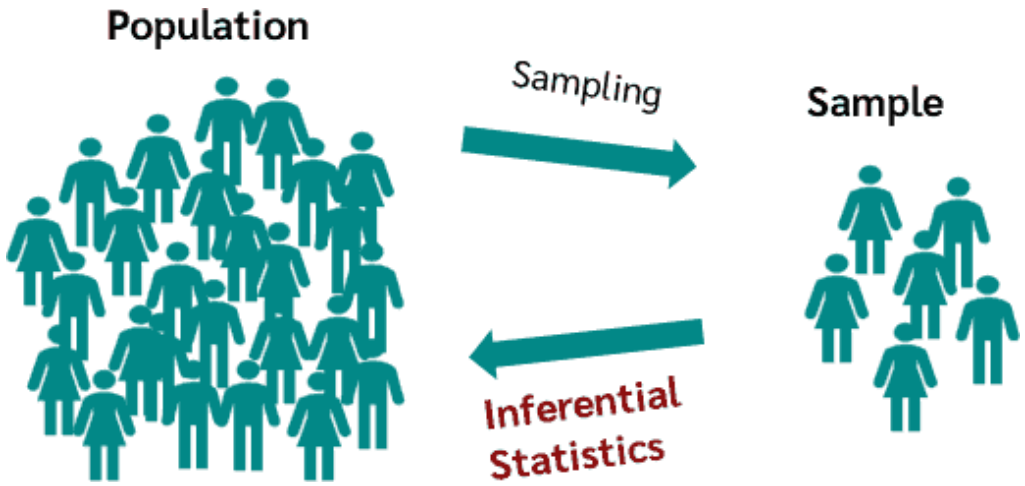
Directional Hypotheses

- 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 H0의 reject 또는 retain
- ex) Hypothesis: "오리온 쵸코파이의 무게는 35g이다"
  - Population: "생산된 모든 쵸코파이"
  - Sample: n개를 택해 무게의 평균을 낸 후
  - 이를 바탕으로 Population에 대한 가설을 테스트



Significance Level (α: 유의 수준)

- Hypothesis test만으로 H0를 100% reject 또는 retain 할 수 없음
- Sample들 마다 data의 구성이 다르고 편향성이 다를 수 있음
- **H0가 실제로는 true이지만 Hypothesis test로는 reject될** (Type 1 error) **확률의 한계를** Significance Level (α) 으로 정해 놓음
  - 어떤 실험에서 Sample을 택할 때, H0를 오판하여 Type 1 error가 날 확률 (p-value) 을 구해 보았음.
  - 만약 이 확률이 매우 작아서 α 보다 작다면, 그 실험에서 H0는 마땅히 reject되어야 할 것임.
- 일반적으로 α는 5% (0.05) 또는 1% (0.01) 로 설정해 놓음.
- α는 H0의 reject 여부를 결정하는데 사용
  - p-value가 α 보다 작으면 H0를 reject
  - 그렇지 않으면 H0를 retain

Significance Level, Confidence Level, Confidence Interval, Margin of Error

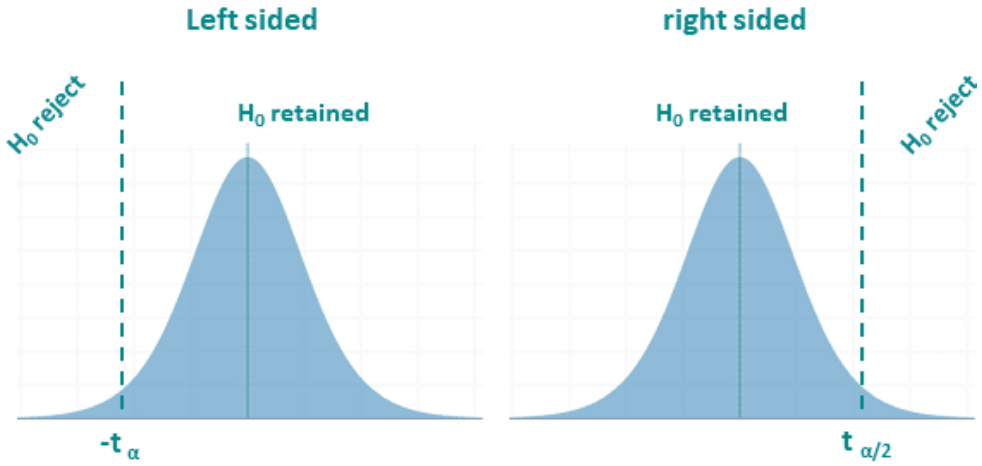
- Confidence Level (CL: 신뢰수준)
  - 1 – Significance Level = $1 - \alpha$
  - ex) α = 5% (0.05) 이면 CL은 95%
- Margin of Error (MOE: 오차범위)
  - 추정된 Sample mean이 실제 Population mean에서 벗어날 수 있는 최대 범위
  - MOE = SE * Z-value at Confidence Level
  - SE (Standard Error, 표준오차) = $\frac{\sigma}{\sqrt{n}}$
  - Confidence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

p-Value

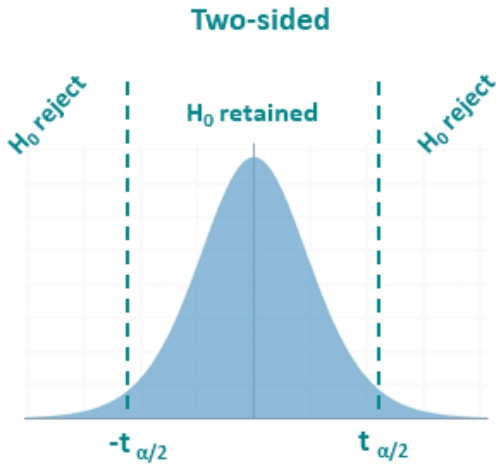
- "Population에서 H0 가 true인 Sample을 택하게 될 확률"
- p-value가 충분히 작다는 의미는?
  - Population에서 H0 가 true인 경우의 Sample을 택하게 될 확률이 충분히 작다
  - 즉, 어떻게 sampling해도 그 Sample에서 H0가 true가 될 확률이 충분히 작다
  - 즉, H0는 true가 아닌 것에 가깝다
  - 즉, H0는 reject 될 수 있다
- p-Value와 Significance Level
  - p-Value ≤ 0.01
    - α = 1% (0.01) 에서 H0는 reject 될 수 있다.
    - α = 1% (0.01) 에서 두 population의 stat value의 차이는 "highly significant" (매우 유의미 하다)
  - p-Value ≤ 0.05
    - α = 5% (0.05) 에서 H0는 reject 될 수 있다.
    - α = 5% (0.05) 에서 두 population의 stat value의 차이는 "significant" (유의미 하다)
  - p-Value > 0.05
    - α = 5% (0.05) 에서 H0는reject 될 수 없다.
    - α = 5% (0.05) 에서 두 population의 stat value의 차이는 "not-significant" (유의미 하지 않다)

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
  - 만약 "Group B가 Group A 보다 반응시간이 더 길다 였다" 면
  - p-value는 1 – 0.02 = 0.98

Types of errors

- Hypothesis test는 항상 정확한 것이 아니고 어느 정도의 error가 일어날 가능성이 있음
- Error의 종류는 다음 표와 같이 나뉨

	Decision	
	for H_0	against H_0
H_0 true	Right	Type 1 error
H_0 false	Type 2 error	Right

- Type 1 Error: H0가 실제로 true (즉, H1이 false) 인데도 불구하고 H0를 reject하는 경우
- Type 2 Error: H0가 실제로 false (즉, H1이 true) 인데도 불구하고 H0를 retain하는 경우

Significance vs Effect Size (1/2)

- Effect Size
  - Significance 이외에 실제로 관찰된 effect의 크기나 강도를 측정
  - 연구 결과의 실질적 중요성 (경제성 등 다른 면에서) 을 평가하는데 중요한 지표
  - 두 Sample 간의 mean 차이가 있지만, 한쪽 Sample의 경우 방법에 비용이 많이 드는 경우라면?
- Effect Size의 중요성
  - 연구에서는 significance와 effect size를 모두 고려해야 함
  - 통계적으로 유의미한 결과 (p-value가 significance level 보다 작다) 라 하더라도 effect size가 작다면 실질적인 의미가 없을 수 있음
  - 반대로, effect size가 크더라도 p-value가 significance level보다 크다면 (즉, 유의미하지 않다면) 해당 결과가 우연일 가능성을 배제할 수 없음

Significance vs Effect Size (2/2)

- Cohen's d
  - 두 sample의 mean 차이를 standard deviation으로 나눈 값
  - 두 sample을 비교하는 test에서 사용
- Pearson's r
  - 두 variable간의 correlation을 나타내는 값. $-1 \leq r \leq 1$
  - Correlation test에서 사용

Effect size	Cohen's *d*	Pearson's *r*
Small	0.2	.1 to .3 or -.1 to -.3
Medium	0.5	.3 to .5 or -.3 to -.5
Large	0.8 or greater	.5 or greater or -.5 or less

Hypothesis Test vs Statistical Test

- Hypothesis Test
  - H0의 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, ...