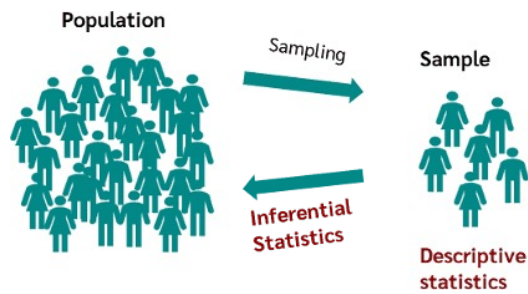


S_02 Hypothesis

Statistical Analysis

Statistical Analysis 두번째로 Hypothesis에 대해 강의합니다.

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

2

이전 강의에서 다룬 Descriptive statistics, 즉, 기술통계는 Sample의 통계를 describe하기 위해 사용하는 것입니다.

그에 반하여 Inferential statistics은 Sample로 구한 통계 정보로부터 Population의 통계를 역으로 추론 (inference) 해 내는 것을 말합니다.

Hypothesis (가설)



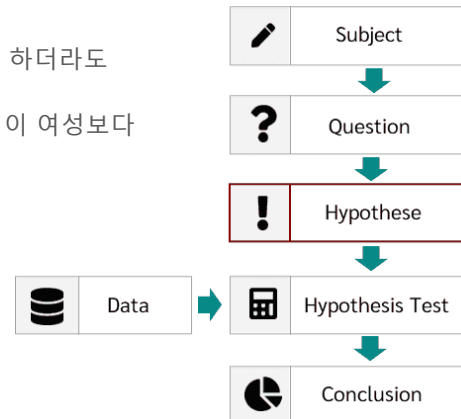
- 입증 (proven) 되지도 반증 (disproven) 되지도 않은 가정
- 연구 초기에 만들어짐
- 연구 목표는 결국 Hypothesis를 **reject** (기각) 하거나 **retain** (유지, 기각하지 않음) 하는 것
- Hypothesis를 reject하거나 retain 하기 위해 **hypothesis test**를 사용
- Hypothesis는 variable들 간의 인과 관계 또는 연관성 (association) 에 대한 가정을 의미함

3

Hypothesis, 즉, 가설이란, true라고 입증되지도, 또, false라고 reject되지도 않은 가정을 말합니다. 가설은 흔히 연구를 시작하는 초기 단계에 만들어 집니다. 연구의 목표는 결국 이 가설을 reject, 즉, 기각하거나, 혹은 retain, 즉, 유지하는 것 이라 할 수 있습니다. 이렇게 가설을 reject하거나 retain하기위해 사용되는 것이 hypothesis test 입니다. 가설을 variable들의 관점으로 바라 본다면, 가설은 variable들 간의 인과 관계 또는 연관성에 대한 가정을 의미한다고 할 수 있습니다. A가 값을 바꿀 수 있는 independent variable, B가 A 값에 의해서 값이 정해지는 dependent variable이라 할 때, "A = x 이면 B = y 이다" 라는 인과관계, 또는 연관성 (association) 에 대한 가정을 의미한다고 볼 수 있습니다.

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.



4

연구를 하는 단계들의 측면에서 보면,

- 1) 먼저 연구의 주제 (Subject)를 정하고
- 2) 연구질문 (Research Question: RQ)를 정의하며
- 3) RQ에 대응하는 가설 (Hypothesis)를 세우고
- 4) User test를 통해 수집한 데이터를 기반으로 Hypothesis test를 실행
- 5) 결론을 내리는

순서로 생각할 수 있습니다.

Subject는 연구의 주제, title이라 할 수 있습니다.

Research Question (RQ)과 Hypothesis는 보통 1:N 으로 대응합니다.

RQ의 경우 좀 더 자유로운 형식의 질문으로

연구에 대한 흥미를 이끌어 내는 역할을 한다고 볼 수 있습니다.

또한 RQ는 여러 개의 Hypothesis를 내포하고 있는 포괄적인 질문이라 볼 수 있습니다.

Hypothesis는 뒤에서 보게 되듯이 어느정도 만들어지는 규칙이 존

재하므로
이를 감안하여 조심스럽게 기술해야 합니다.

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



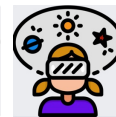
- ex) Medical research

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



- ex) Virtual Reality

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



5

가설에서 사용되는 서로 다른 variable들은
서로 다른 value들을 가지게 됩니다.

예를 들어, 사회과학 분야의 연구에서는

Gender, Income, 환경에 대한 태도와 같은 것들이
variable이 될 수 있습니다.

의학 분야의 연구에서라면, 체중, 흡연상태, 심박수 등이
변수가 될 수 있을 것입니다.

가상현실 분야에서는 어떤 device로 측정되는 신호
렌더링 파라미터들, 몰입감, 현존감, 성능, 사용성 등의
경험 수치들이 될 수 있을 것입니다.

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

6

이번에는 변수들의 종류를
그 역할에 따라 구분해 보도록 합니다.
먼저 Independent variable, 즉, 독립변수는
treatment variable 이라고도 불리는데,
실험 또는 연구에서 그 값을
조작하거나 변화시키는 변수를 말합니다.
독립변수의 값의 변화가
종속 변수의 값에 미치는 영향을 관찰하고
이에 대한 규칙을 찾아내는 것이
결국 연구에서 하는 일입니다.
독립변수는 실험에서의 원인 또는 자극을 말하는데,
예를 들어, 교육 연구에서는 수업 진행 방식, 즉,
online 인지 offline 인지과 같은 것이 될 것이고,
식물 성장 실험에서는 질소비료,
인산비료와 같은 값을 가지는
비료의 종류 같은 것을 독립변수라 할 수 있을 것입니다.

종속변수는 response variable 이라고도 불리는데,
독립변수의 변화에 따라서
그 값이 수동적으로 변하게 되는 변수라 할 수 있습니다.
종속변수는 측정하거나 관찰하는 결과를 지칭하며,
예를 들면, 교육연구에서는 시험점수,
식물 성장 실험에서는 식물의 길이 같은 것을 말합니다.

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

7

XR 분야에서의 독립변수로는
특정 디바이스의 사용 여부
렌더링 파라미터들,
여러 개의 다른 자극들의 동시적 인식 여부
시스템에 특화된 파라미터 등이 있을 수 있으며
종속변수로는
각종 experience, 즉, 몰입감, 현존감, 성능, 유용성,
embodiment (구체화), 인지부하, 사이버멀미 등의 레벨과
각종 social experience 지표들 등을 고려할 수 있습니다.

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

8

그렇다면 모든 연구에서 독립, 종속 변수라는 정의가 적절할까요?
Correlation에 대한 연구에서는 부적절할 수도 있습니다.
Correlation에 대한 연구는 variable A와 B의 관계가
어떠한지를 관찰하는 연구인데,
A와 B가 항상 원인과 결과의 관계를 지칭하는 것은 아닙니다.
Correlation에 대한 연구에서는
variable A와 B가 서로 대등한 위치일 수도 있습니다.
나중에 Correlational Research에 대해서는
좀 더 자세히 들여다 볼 기회가 있습니다.
하지만 그렇다 하더라도 대부분의 경우에는
선행, 후행의 관계는 존재합니다.
그래서 Correlational Research에서는
독립, 종속 변수라는 말보다는
Predictor, Outcome variable이라는 용어를 사용합니다.
예를 들면 " 강우량" 이 predictor variable이고
"진흙의 양" 이 outcome variable일 수 있습니다.

대체로 predictor/outcome variable 이라는 용어는 independent/dependent variable 의 개념을 포함한다고 볼 수 있습니다.
따라서 모든 형태의 다양한 연구들 에서 variable들을 지칭할 때에는 predictor variable, outcome variable 이라는 용어를 쓰는 것이 적절합니다.

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

9

predictor, outcome variable 이외에도
research에는 여러가지 다른 역할을 하는
variable들이 존재합니다.

Control variable은

실험 내내 일정하게 유지되는 변수를 말하는데,
흔히 교란 요인을 통제하는데 사용됩니다.

예를 들어 비료의 종류에 따른 식물 성장 실험에서는
온도, 빛의 양, 물의 양 등

다른 요인들은 모든 식물에게 똑같이 적용하여
교란 요인을 차단합니다.

이런 것들을 control variable이라 합니다.

Latent variable은 직접적으로 측정은 불가하지만,
실험 내에서 다른 대리 (proxy) 변수에 의해
유추가 가능한 variable을 말합니다.

예를 들면 "식물의 염분 내성"은 직접 측정할 수는 없습니다.
그러나 염분을 첨가하는 실험에서

식물의 건강상태를 측정할 수 있다면,
이로부터 엽분 내성을 유추할 수 있을 것입니다.
VR에서 immersion을 뇌파, 심박수, 피부전도도, 눈움직임,
호흡 등의 생리학적 반응으로만 측정하려할 때
immersion은 latent variable이 됩니다.
immersion이 높은가 낮은가를 직접 사용자에게
설문조사로 물어본 것이 아니라
다른 생리학적 반응으로 부터 유추하기 때문입니다.
Composite variable은 두 개 이상의 variable들을
combine하여 만들어지는 variable입니다.
예를 들면 (국어성적 * 수학성적 * 영어성적)
혹은 (국어성적 + 영어성적) 과 같이
기존의 variable들을 조합하여
새로운 variable을 만들어 낼 수 있는데,
이것은 측정되는 것은 아니고
분석단계에서 만들어 지게 되는 것입니다.

Null Hypothesis vs Alternative Hypothesis

- Null Hypothesis (H0: 귀무가설)
 - 두 개 이상의 그룹 간에 어떤 특성과 관련한 차이가 없다는 가설
 - ex) 한국의 남성과 여성 근로자의 임금은 차이가 없다.
- Alternative Hypothesis (H1: 대립가설)
 - 두 개 이상의 그룹 간에 어떤 특성과 관련한 차이가 있다는 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 “차이가 있다” 는 것을 보이려는 경우가 많으므로
 - 보통 Research Hypothesis를 H1으로 놓고
 - H0를 reject할 수 있음을 보인다.

10

이제 Hypothesis의 개념에서 가장 중요하다고 볼 수 있는 Null hypothesis와 Alternative Hypothesis의 개념에 대해 알아보니다.

Null hypothesis, 즉, 귀무 가설은 흔히 H0로 표시하는데, 두 그룹 간의 통계량 (대부분 mean) 의 차이가 없다는 가설입니다. 예를 들면 “한국의 남성과 여성 근로자의 임금 차이는 없다”, “비료의 종류를 달리 주어 기른 두 식물 그룹 간의 생장 정도에는 차이가 없다” 같은 것들입니다.

이 H0와 정확히 반대의 가설을 Alternative Hypothesis (대립가설) 이라 하는데 H1으로 표시합니다.

H1은 H0와는 달리 두 그룹 간의 통계량에 차이가 있다는 가설이 될 것입니다.

예를 들면 “한국의 남성과 여성 근로자의 임금에는 차이가 있다” 같은 것들입니다.

그런데 가설의 진실 여부를 test하는 Hypothesis test 에서는 H_0 만 test됩니다. 즉, H_0 를 reject할지 아니면 retain할지를 test로 결정하는 것입니다.

그럼 H_1 의 진위여부는 어떻게 결정할까요?

H_0 가 reject되었다는 것은 H_1 이 retain되었다는 것과 정확히 같은 의미 입니다.

반대로 H_0 가 retain되었다는 것은 H_1 이 reject되었음을 의미합니다.

Research에서는 흔히 H_1 형식의 (차이가 있다는) 가설을 보이려는 경우가 더 많은 것 같습니다.

따라서 Research Hypothesis를 H_1 으로 놓고, H_0 를 reject할 수 있음을 보이면 됩니다.

Types of Hypothesis

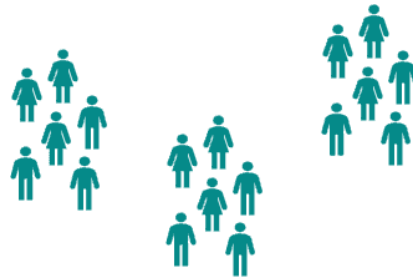
- Difference hypothesis (차이 가설)
- Correlation hypothesis (상관관계 가설)
- Directional hypothesis (방향성 가설)
- Non-directional hypothesis (비방향성 가설)

11

이제 H_0 와 H_1 의 의미를 알았으니
좀 다른 방향에서 Hypothesis를 분류해 보겠습니다.
Hypothesis는 Difference, Correlation,
Directional, Non-directional hypothesis 의 type들로 구분해 볼 수
있습니다.
그러나 이 네 가지의 type들이 서로 exclusive하지도 않고
그 union이 전체 가설의 set을 모두 cover하지도 않습니다.
다만 directional과 non-directional은 서로 exclusive한 개념입니다.

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



12

Difference hypothesis는

두 개 이상의 group에 어떤 차이가 있다 (또는 없다) 는 가설을 말합니다.

예를 들면 "남성 그룹이 여성 그룹보다 수입이 높다",

"흡연자는 비흡연자보다 심장마비 위험이 높다",

"주당 근무시간에서 한국과 일본은 차이가 없다" 같은 것들입니다.

이 type의 가설에서 한 variable 은 categorical variable입니다.

예를 들면 gender (value: male/female),

country (value: 한국/일본/중국) 같은 것들입니다.

다른 하나의 variable은 차이를 나타낼 수 있어야 하니까

최소한 ordinal variable이거나 metric 이어야 합니다.

예를 들면, 급여, 심장마비 비율, 주당 근무 시간 같은 것들입니다.

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



13

Correlation Hypothesis는 두 variable들 간의 correlation, 즉, 상관관계에 대한 가설입니다.
예를 들면 "키가 클 수록 체중이 더 나간다",
"차의 마력 수 (horsepower) 가 더 높을수록 연비는 떨어진다",
"수학 성적이 좋을 수록 미래 연봉이 높다" 같은 것들입니다.
상관관계 가설을 만들기 위해서는
두 variable이 최소한 ordinally scaled 또는 metric 이어야 합니다.

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

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

14

여기에서 잠시 probability density function (PDF) 의 개념을 복습해 보겠습니다.

PDF $f(x)$ 는 x 가 연속된 값 (실수) 인 연속확률분포에서 특정 구간 내의 값이 나타날 확률을 표현하는 function 입니다. $f(x)$ 는 다음의 특징을 가집니다.

먼저 $f(x)$ 는 모든 x 에 대해 nonnegative 입니다.

$f(x)$ 가 확률을 표현하기 때문에 당연히 그러할 것입니다.

확률변수 X 가 $[a, b]$ 구간에 존재할 확률 (즉, 구간 $[a, b]$ 에서 값이 나타날 확률)은

a 에서부터 b 까지의 구간에서 $f(x)$ 의 x 에 대한 정적분값입니다.

따라서 전체확률은 $f(x)$ 를 마이너스 무한대에서 플러스 무한대까지의

x 에 관하여 적분한 값이며, 1이 됩니다.

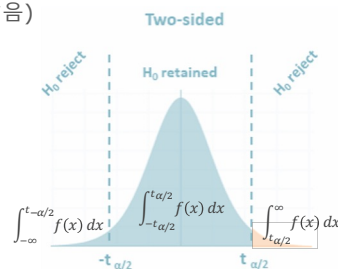
그러나 $f(x)$ 는 확률 자체를 나타내지는 않습니다.

한가지 대표적인 예로 평균 μ , 표준편차 σ 를 가지는 normal distribution, 즉 정규분포의 PDF $f(x)$ 는

여기에 제시된 식입니다.

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)



15

Non-directional 가설은 “차이가 있다”
 혹은 “차이가 없다” 만이 관심의 대상일 뿐,
 어느 쪽이 더 크고, 어느 쪽이 작은 지에는
 관심이 없는 가설을 말합니다.

Non-directional 가설은 Difference

혹은 Correlation Hypothesis 중에 하나 일 수 있습니다.

예를 들면 “남성과 여성의 급여에는 차이가 있다” 는 가설은
 차이만을 언급할 뿐, 남성과 여성 중 누가 더 많이 버는지는 말하지
 않습니다.

또, “흡연자와 비흡연자 간에는 심장마비 위험에 차이가 있다” 는
 가설도

차이가 있다는 언급을 하지만, 흡연자나 비흡연자 중
 누가 더 심장마비 위험도가 높은지는 말하지 않습니다.

이런 형태의 가설을 non-directional 이라 합니다.

그림에서 표준정규분포 (standard normal distribution) 의
 확률밀도함수 (pdf: Probability Density Function) 의 그래프를 볼

수 있는데,

H0가 retained 되는 확률을 가지는 horizontal axis의 t 의 구간은 $t_{-\alpha/2} \leq t \leq t_{\alpha/2}$ 입니다.

H0가 retain된다는 것은 "차이가 없다" 는 것이 인정된다는 뜻입니다.

그렇다면, H0를 reject할 수 있는 t 의 구간은

$(t < t_{-\alpha/2})$ 또는 $(t > t_{\alpha/2})$ 인 두 구간을 말합니다.

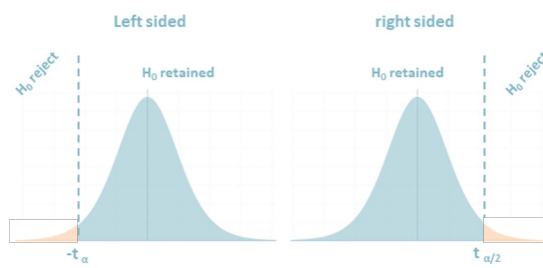
Non-directional 가설에서는 이 두 구간 모두가

H0를 reject할 수 있는 조건으로 간주될 수 있습니다.

이러한 이유로 Non-directional 가설을 test하는 것을 two-tailed test 또는 two-sided test라 부릅니다.

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)



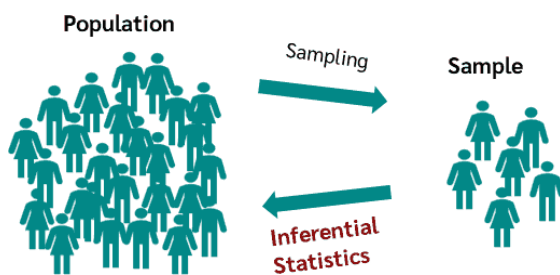
16

Directional Hypothesis는 Non-directional Hypothesis의 반대 개념으로,
일반적으로 가설에 “더 좋다 (better than)” 또는 “더 나쁘다 (worse than)” 와 같은 표현이 포함됩니다.
예를 들면, “남자가 여자보다 수입이 높다”,
“흡연자는 비흡연자보다 심장마비 위험이 높다” 같은 형태 입니다.
Directional hypothesis의 test에는
“one-tailed test” 또는 “one-sided test” 를 하게 됩니다.
또, correlational hypothesis 관점에서 보면
두 variable간에는 “positive correlation이 있다”,
또는 “negative correlation이 있다” 와 같은 방향성이 포함됩니다.
positive correlation 의 예는 “키가 크면 몸무게가 무겁다” 같은 것
들이 있고
negative correlation 의 예는
“노동자의 평균 수면시간이 길어지면 제조 상품의 불량률이 떨어진
다”

같은 것이 있습니다.

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

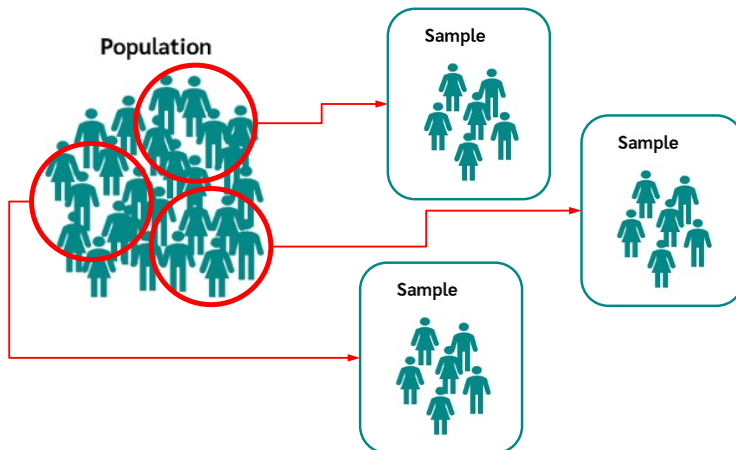


17

Hypothesis testing은 Sample의 도움으로 Population에 대한 Hypothesis를 테스트 할 때 사용합니다. Test의 결과는 H_0 의 reject 또는 retain입니다. 예를 들어 Hypothesis가 "오리온 초코파이의 무게는 35g이다" 일 때 Population은 "생산된 모든 초코파이" 가 되고 Sample의 경우 n개를 택해 무게의 평균을 낸 후 이를 바탕으로 Population에 대한 가설을 테스트하게 됩니다.

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



18

Hypothesis test, 즉, sample의 hypothesis에 대한 결과가 population의 hypothesis에 대한 결과로 확장될 수 있다는 것을 테스트하는 과정은

항상 정확한 것은 아니고, 어느 정도 error가 일어날 가능성이 있습니다.

그 이유는 random하게 고른 sample이 population의 분포의 특징을

항상 충분히 반영하는 것은 아니기 때문입니다.

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

19

때문에 항상 어떤 판단에 error가 생길 가능성을 대비해야 합니다.
Error의 종류로 Type 1 error는 H_0 가 실제로 true인데 H_0 를 reject하는 경우이며,
Type 2 error는 H_0 가 실제로는 false인데 H_0 를 retain하는 경우입니다.

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.

20

이제 p-Value의 개념에 대해 알아보겠습니다.

p-Value는 "Population에서 H0가 true인 Sample을 택하게 될 확률" 을 말합니다.

그렇다면 "p-Value가 충분히 작다"는 의미는 무엇일까요?

p-Value의 정의에 의해,

"Population에서 H0 가 true인 경우의 Sample을 택하게 될 확률이 충분히 작다" 는 뜻입니다.

다른 표현으로는 어떻게 sampling을 하더라도

그 sample에서 H0가 true가 될 확률이 충분히 작다" 라는 뜻이며, 그것은 H0는 true가 아닌 것에 가깝다는 뜻이 됩니다.

즉, H0는 reject될 수 있다 라는 뜻입니다.

보통 H0는 "두 data의 평균의 차이가 없다" 이고

"H1은 두 data의 평균의 차이가 있다" 이므로,

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 0.053이었다 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.

21

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 사용합니다.
즉 p-value를 구해 그 값이 α 보다 작거나 같으면
significant level α 에서 H_0 를 reject 할 수 있습니다.
즉 이 경우에는 H_1 을 retain 할 수 있습니다.
반면에 구해진 p-value가 α 보다 크면,
 H_0 를 retain, 즉, H_1 을 reject 할 수 있습니다.
백신 A와 B의 효과 비교를 예로 들어 보겠습니다.
일정 인원의 Sample에 백신 A를 접종한 경우, 발병된 개체수가 평균 23명이고
같은 인원의 Sample에 백신 B를 접종한 경우, 개체수 평균 32명이 발병하였다.
이 사실 만으로 보면 백신 A가 B보다 더 우수하다고 볼 수 있습니다.

정말 그럴까요?

A가 B가 우수하다면, H_1 을 백신 A의 효과가 B보다 좋다 라는 가설로 두고

이를 retain할 수 있어야 할 것입니다.

일단 p-value를 구할 수 있다고 가정하고,

A와 B의 차이에 대한 p-value를 구해보니, 0.053 이었습니다.

$0.053 > 0.05$ 이므로

H_0 : 백신 A와 B의 효과에 차이가 없다 라는 가설을 retain 할 수 있습니다.

즉, Sample stat에서 백신 A가 B보다 우수하지만,

이 차이는 significance level $\alpha = 0.05$ 에서 유의미하지 않다고 말할 수 있습니다.

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)

22

p-value와 significance level alpha의 관계로
가설에 대해 내릴 수 있는 결론을 정리해 보았습니다.
먼저 p-value가 0.01보다 작거나 같을 때에는
alpha = 1%, 즉, 0.01에서 H0를 reject할 수 있습니다.
따라서 H1을 retain할 수 있으며, 두 population의 평균 차이는
highly significant (매우 유의미) 합니다.
p-value가 0.05보다 작거나 같을 때에는
alpha = 5%, 즉, 0.05에서 H0를 reject할 수 있습니다.
따라서 H1을 retain할 수 있으며, 두 population의 평균 차이는
significant (유의미) 합니다.
마지막으로 p-value가 0.05보다 클 때에는
alpha = 5%, 즉, 0.05에서 H0를 retain할 수 있습니다.
따라서 H1을 reject할 수 있으며, 두 population의 평균 차이는
not-significant, 즉, 유의미하지 않습니다.

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%]

23

이번에는 조금 다른 측면으로 confidence level (신뢰 수준), confidence interval (신뢰 구간), margin of error (오차 범위)에 대해 알아보겠습니다.

이 용어들은 research의 statistical analysis보다는 여론조사 결과를 발표할 때에 많이 쓰이는 용어들입니다.

먼저 confidence level, 즉, 신뢰 수준, CL은 1에서 significance level인 alpha를 뺀 값입니다.

예를 들어 alpha가 5% 일 때, CL은 95% 입니다.

Margin of Error (MOE), 즉, 오차범위는 추정된 sample mean이 실제 population mean에서 벗어날 수 있는 최대 범위를 나타냅니다.

MOE는 $\text{SE} * \text{Confidence Level의 Z-value}$ 인데 SE는 standard error, 즉, 표준 오차로서 standard deviation을 Sample의 개체수 n 의 root값으로 나눈 것입니다.

Confidence level에서의 z-value는

95%일 때 1.96, 99%일 때는 2.576 입니다.

마지막으로 confidence interval (CI), 즉, 신뢰구간은
Sample mean 플러스 마이너스 MOE 입니다.

한가지 예를 보겠습니다.

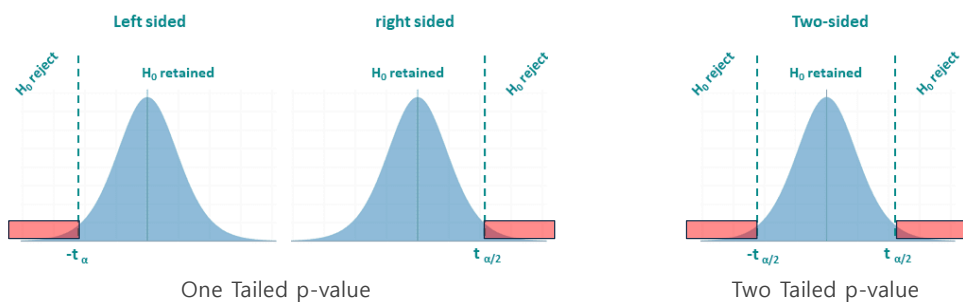
n = 1000명의 sample에 대하여 95% 신뢰 수준의 조사에서
지지율, 즉, sample mean이 45%, 표준편차가 20% 라고 가정합시다.

MOE는 $1.96 * (0.2 / \sqrt{1000})$ 으로써 0.012396... 입니다.

CI는 0.45 플러스/마이너스 0.0124 이므로
구간 [43.76%, 46.24%] 가 됩니다.

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.



24

One-tailed 또는 one-sided p-value는 Directional hypothesis 으로서 두 Sample들 간의 차이가 있느냐 외에도 어느 Sample쪽이 더 큰지 (혹은 작은 지)에 관심이 있습니다. 주어진 direction (크다 또는 작다) 을 만족하는 경우만 counting한 확률이 되는데, Two Tailed p-value를 2로 나눈 값을 p-value로 사용합니다. One Tailed p-value는 right sided와 left sided 로 나눌 수 있으며 Two tailed p-value의 경우 양쪽 끝의 확률을 모두 합한 값이 됩니다.

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
 - $0.02 \leq 0.025$

25

Two Tailed p-value 또는 Two Sided p-value는
Non-directional hypothesis의 경우
두 sample들 간의 차이가 있는가에만 관심이 있습니다.
Direction에 따라 positive와 negative한 경우를
모두 counting하고 있습니다.

One Tailed p-value와 Two Tailed p-value간의 관계는
One Tailed p-value의 경우 Two Tailed p-value를
2로 나눈 값이 됩니다.

예를 들자면, H0와 H1은 원래 Two tailed p-value 만을 고려합니다.
H0가 Group A와 Group B의 반응시간은 차이가 없다는 가설이며
H1은 Group A와 Group B의 반응시간은 차이가 있다는 가설이 됩니다.

Two tailed p-value가 0.04일 경우
"Group A의 반응시간이 Group B보다 더 길다" 또는
"Group B의 반응시간이 Group A보다 더 길다" 의 p-value는
0.04를 2로 나눈 0.02가 됩니다.

이 때 significance level α 도 2로 나누어 판단해야 하며
이 예의 경우 0.02가 0.05를 2로 나눈 0.025보다 작으므로
한쪽의 반응시간이 다른쪽보다 더 길다라는 가설은
significant하다고 말할 수 있습니다.

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, ...

26

Hypothesis test와 Statistical Test라는 용어들의 의미에 대해 알아보겠습니다.

Hypothesis test는 H0의 reject, 또는 retain 여부를 결정하는 test입니다.

따라서 두 개 이상의 group (sample) 간의 비교에 한정된 test를 의미합니다.

Hypothesis test에는 t-test나 ANOVA가 해당합니다.

반면 Statistical Test는 hypothesis test 뿐만 아니라 difference 뿐만 아니라 correlation, linearity 등 다양한 측면에서의 test를 망라 합니다.

Statistical Test에는 t-test, ANOVA 외에도 regression, correlation 등의 test가 포함됩니다.