

일반통계학

제 1장

자료의 생성

통계학 2016.1학기 정혜영

1-1 통계학이란 무엇인가

•통계학이란

주어진 문제에 대하여 합리적인 답을 줄 수 있도록 숫자로 표시되는 정보를 수집하고 정리하며, 이를 해석하고 **신뢰성 있는 결론**을 이끌어 내는 방법을 연구하는 과학의 한 분야이다.

Q1) 수집: 전체를 잘 대표할 수 있는가?

Q2) **신뢰성 있는 결론**: 어떻게 결론 내릴 것인가? 어떻게 신뢰성을 측정할 것인가?

예) 흡연여부에 따른 폐암 발병률 

	폐암발병 여부		발병률
	예	아니오	
흡연자	10	40	20%
비흡연자	5	45	10%

1-1 통계학이란 무엇인가

•통계학의 응용분야

공업통계 : 실험계획법, 분산분석법, 신뢰성이론

농업통계 : 표본론

의(약)학통계 : 생존분석, 임상실험, 범주형자료분석

경제(경영)통계 : 시계열분석, 회귀분석

사회통계 : 범주형자료분석

•용어

모집단(population) : 관심의 대상이 되는 모든 추출단위의 특성값을 모아 놓은 것

추출단위(sampling unit) : 전체를 구성하는 각 개체

특성값(characteristic) : 각 추출단위의 특성을 나타내는 값

표본(sample) : 실제로 관측한 추출단위의 특성값의 모임

유한모집단(finite population) : 유한개의 추출단위로 구성된 모집단

무한모집단(infinite population) : 무한개의 추출단위로 구성된 모집단

1-1 통계학이란 무엇인가

•자료의 종류

범주형 자료(categorical data), 질적자료(qualitative data)

- 관측결과가 몇 개의 범주 또는 항목의 형태로 나타나는 자료
- 명목자료(nominal data): 순위의 개념이 없다. 혈액형, 성별
- 순서자료(ordinal data): 순위의 개념을 갖는다. 학점, 선호도

수치형 자료(numerical data), 양적자료(quantitative data)

- 자료 자체가 숫자로 표현되며 숫자자체가 자료의 속성을 반영.
- 연속형자료(continuous data, uncountable): 키, 몸무게
- 이산형(discrete data), 계수형자료(counting data) : 교통사고 건수

※ 수치형 자료도 범주형자료로 변환 가능

1-1 통계학이란 무엇인가

•기술통계학

기술통계학(descriptive statistics)

- 표나 그림 또는 대표값 등을 통하여 수집된 자료의 특성을 쉽게 파악할 수 있도록 자료를 정리·요약하는 방법을 다루는 분야
- 2장


추측통계학 (inferential statistics)

- 표본에 내포된 정보를 분석하여 모집단의 여러 가지 특성에 대하여 과학적으로 추론하는 방법
- 3장~11장

1-2 표본 추출 방법

•단순랜덤추출

유한모집단에서 n 개의 추출단위로 구성된 모든 부분집합들이 표본으로 **선택될 확률이 같도록** 설계된 표본추출방법이다.

- 모집단을 잘 대표할 수 있는 표본을 추출할 수 있다.
 - 표본이 모집단의 한쪽 방향으로 치우치지 않게 하기 위한 방법이다.
 - 주로 난수표나 컴퓨터의 난수 발생 프로그램을 이용한다.
 - 단순랜덤복원추출, 단순랜덤비복원추출
- 복원추출(sampling with replacement): 한 번 뽑은 것을 다시 되돌려 놓고 다음 것을 뽑는 것 
- 비복원추출(sampling without replacement): 한 번 뽑은 것은 다시 집어넣지 않고 다음 것을 뽑는 것

•계통, 층화, 집락 등의 추출방법이 있다.-12장

1-3 자료생성 방법

•표본조사

표본조사시 유의사항 : 표본의 치우침 현상을 피하기 위한 주의사항

1. 완전한 모집단의 리스트 준비
2. 무응답의 적절한 관리
: 무응답자 집단과 응답자 집단 간의 비교작업이 필요하다.
3. 철저한 조사자의 훈련 및 감독
4. 정확한 설문지 작성

1-4 자료생성방법

•통계적 실험



실험단위와 처리

실험이 행해지는 개체를 **실험단위**라 하고, 각각의 실험단위에 특정한 실험환경 또는 실험조건을 가하는 것을 **처리**라고 한다.

반응변수와 인자 및 인자수준

통계적 실험에서, 실험환경이나 실험조건을 나타내는 변수를 **인자**라 하고, 이에 대한 반응을 나타내는 변수를 **반응변수**라고 한다. 인자가 취하는 값을 그 **인자의 수준**이라고 한다.



예) 고등학교 수학 교과서에서 문제의 위치가 교육에 미치는 영향에 관한 연구

위치: 전반부, 후반부

질문의 종류 : 간단한 사실, 계산, 개념

실험: 위치와 질문의 종류 조합으로 여섯 가지 종류의 시범교과서를 만들어 여섯 그룹의 학생들에게 동일 기간에 같은 양의 숙제 등으로 일정하게 교육한 후, 같은 시험문제를 통해 교육의 효과를 측정하고자 한다.



1-5 비교실험과 랜덤화

• 용어

- (1) 위약효과(placebo effect): 어떤 처방에도 긍정적으로 반응하는 것
- (2) 처리집단(treatment group): 처리를 받는 집단
대조집단(control group): 처리를 받지 않은 집단
- (3) 이중눈가림실험(double blind experiment): 실험자 및 피실험자 모두가 처리 여부를 모르는 실험
- (4) 교락(confounding): 관심인자와 외부인자의 효과가 분리되지 않은 상태
- (5) 랜덤화(randomization): 실험단위가 처리집단이나 대조집단에 들어갈 기회를 동등하게 부여하는 방법
- (6) 완전랜덤화계획(completely randomized design): 전체 실험단위를 처리의 개수만큼 나눈 후 그에 따라 모든 실험단위를 각 처리에 배정하는 실험계획

1-6 블록화 (Blocking)

•블록화

블록화란 실험 이전에 동일 처리에 대한 반응이 유사할 것으로 예상되는 실험 단위들끼리 모아서 군을 형성하는 것을 뜻하며, 이 때 각 군을 블록이라고 한다.

예) 흙의 종류와 비옥함의 정도는 농지에 따라 매우 다르다. 콩의 산출량에 대한 두 가지 품종과 세 가지 살충제의 처리 효과를 알아보기 위하여 실험을 실시한다고 하자. 이 때 각 단위 농지를 블록으로 잡고 각 단위 농지를 다시 6개의 작은 구획으로 나누어, 각 블록 안에서 랜덤화를 통해 여섯 가지의 처리를 6개의 작은 구획에 배정하여 실험할 수 있다.

통계 관련 영상 사이트

<https://www.youtube.com/watch?v=X4hMFym0-uo>

<https://www.youtube.com/watch?v=drhH5Wl419Q>

<https://www.youtube.com/watch?v=9ghBMTdJlVA>