기초통계학 I

Basic Statistics

기초통계학 I	기초통계학 II
○ 통계학이란? ○ 기술통계 - 표와 그래프, 수치적 해석 ○ 확률 ○ 확률변수와 확률분포 ○ 표집분포	○ 통계적 추론 - 추정, - 검정 ○ 분산분석 ○ 회귀분석 ○ 범주형 자료분석

sampling distribution

통계 (statistic)

특정 집단을 대상으로 한 조사나 실험에 의하여 얻은 자료(data)를 요약한 결과 (수치, 표, 그래프, 그림 ...)

예) 와플메이커 판매계

경제통계: 실업률, 물가지수, 가구 소득/지출, 작물재배면적 등

보건사회통계: 흡연율, 평균의료비, 사교육비, 평균수명, 생활시간 등

실험통계: 암치료제 임상실험 결과, 철강제품 강도, 식빵 당도 등

통계학과 타학문의 융합

o Biometrics, Econometrics, Thchnometrics, Psychometrics, sociometrics

※ "**Data Science**", "**Big Data**", "**Data Mining**" ...

통계학 응용 분야

o 인구, 경제, 경영 통계

o 사회, 심리, 교육 통계

o 보건, 의학, 생물, 환경, 농업 통계

o 일기예보, 지진, 대기오염 ...

o 여론조사, 스포츠, 시청률 ...

통계학(statistics) 어원

"State Arithmetic" (국가산술)

o 국가의 경제, 인구, 정치에 관련된 자료

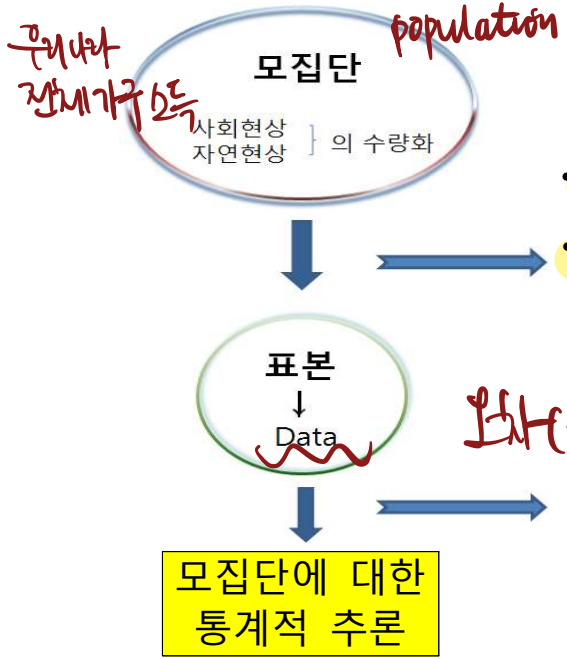
- 세금징수, (군)인력동원, 지가산출 등 목적
- 고전적인 통계학: 기술 통계학 (descriptive statistics) 잘정리

o 17, 18 세기

- 천문학, 물리 : 측정값의 변이성 (불확실성) 오차
- 도박 관련 확률 계산 (예) 한 쌍의 주사위 24번 ⇨ (6, 6) 1번 이상
  1654년, Pascal & Fermat ☞ "Theory of Probability"

※ **기술통계학**(descriptive statistics) & **추측통계학**(Inferential statistics)

통계학이란?



- Sample Survey: 사회과학 자료수집
- Experimental Design: 자연과학(실험) 자료수집

※ Sample survey: 조사대상 control 불가능

- 표본에서 얻은 정보의 일반화:
  - 추정, 검정 등

전수조사(Census) vs 표본조사(sample survey)

□ 전수조사(census), 총조사

o 모집단 전체를 대상으로 조사하는 경우

o "census": 추정하다(to estimate)라는 뜻의 라틴어 "censere"에서 유래

- 로마제국의 인구조사(Census)를 위해 출생 도시로 가서 호적 등록하라는 명령을 내렸으며, 이에 따라 예수님의 부모는 고향을 가는 과정에서 예수님 탄생
- 144년에 중국 한나라에서 실시한 조사에서는 994만 가구에 4973만 명이 사는 것으로 기록

o 우리나라의 Census "인구주택총조사"

- 1949년 정부수립 후 대규모 조사를 시작, 한국전쟁 중 자료 모두 소실
- 1955년 간이총인구조사를 거쳐 1960년까지 국세(國勢)조사 실시
- 1963년 일본식 용어인 국세조사 사용 금지, 이후 '센서스'와 '총조사'라는 용어를 번갈아 사용하다가 1990년 이후 총조사로 통일
- 2010년 '인구주택총조사'에서는 인터넷을 병행 실시
- 2015년부터 전수조사 대신 행정자료를 기반으로 한 **등록센서스** 실시

[예제 1.1] 조선시대 임금의 수명과 재위기간

- 27개 수명/재위기간으로 구성된 모집단 (모집단의 평균 계산 가능)
  ⇨ 조선시대 백성들의 평균 수명? ☞ **모집단?**

모집단과 표본

◉ **우리나라 가구당 평균 소득(자산/부채/지출)?** ☞ 가계금융복지조사

▷ 모집단(population)

o 우리나라 2,000만 가구의 소득(수치) 전체 집합 ☞ 유한? 무한?

- 실제 전체 가구 소득 파악은 불가능/비현실적

▷ 표본(sample), 표본자료(sample data)

o 표본추출 20,000가구 ☞ 실제로는 측정(measurement)이 필요함

※ 표본추출방법, 추정방법, 표본추출오차... & 자료 요약 및 분석

모집단 (population)

o 연구목적에 따라 개념적으로 규정한 관심 대상 **수치(자료)/개체** 들의 전체 집합을 모집단(popuation)이라고 한다.

- 숙명여대 2016년 신입생의 통학 시간
- 우리나라 실업률 (경제활동인구조사) 취업/실업여부 — 경제활동인구조사
- 우리나라 가구 평균 소득 (가계금융복지조사) — 가계금융복지조사
- 우리나라 ~~가구~~ 고등학생 평균 사교육비 (사교육비조사)

※ 목표모집단(target population) & 조사모집단(survey population)
연구목적에 따라 정의 / 실제 조사대상이됨 (ex. 도서·산간지역제외)

※ 유한모집단(finite population) & 무한모집단(infinite population) 이론적개념의 용이
모집단을 구성하는 수치의 개수 → 모집단이 큰경우 근사적으로 취급

표본조사(Sample Survey)

: 특정집단(**모집단**)에서 일부분(**표본**)을 추출하여, 표본에서 얻은 정보를 이용하여 모집단의 특성을 파악하기 위한 조사

- 정부 : 경제활동인구조사, 가계금융복지조사, 작물생산량조사, 국민건강영양조사, 사교육비조사, 생활시간조사
  - 목적 : 경제·사회 실태 파악 및 정책 수립, 자원의 효율적 배분 등

- 기업체 : 고객만족도조사, 시장조사, 품질관리 등
  - 목적 : 경영의사결정 (유통, 고객관리, 생산관리 등)

- 기타 : 일반여론조사, 선거여론조사(정당, 언론기관 등), 시청률조사 등

(표본조사)

Sample survey를 하는 이유?

- Why not a census (complete enumeration) ?

(전수조사)

- 경제성 (비용절감) : 조사비용, 관리비용, 자료처리비용 등

  (예) 2010년 인구주택총조사 예산: 1800억 수준

  (1월일조사 → 결과는 2011년가을 (1년걸림)

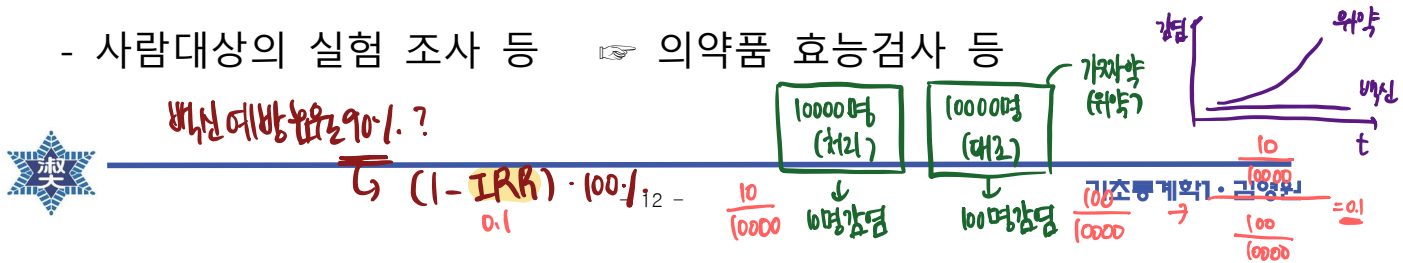
- 시간단축 (신속한 결과 & 적시성)

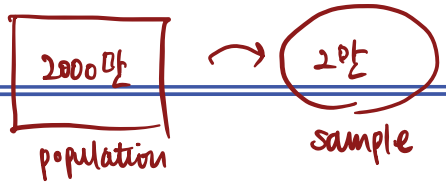
  (예) 쌀 수확량 -> 정부대책 수립, 선거여론조사 -> 선거전략 수정

- 전수조사 불가능

  - 파괴적인 조사인 경우 적용가능 ☞ 전구수명조사 등 타이어마모,..

  - 사람대상의 실험 조사 등 ☞ 의약품 효능검사 등

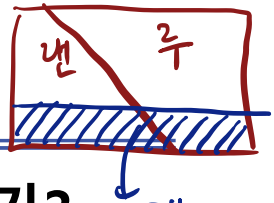




- 정확성의 확보 (???)
  - 표본추출오차(sampling error) : 일부분만 조사
    - ☞ 과학적 오차(정확도) 설명
  - 비표본추출오차(non-sampling error) : 조사원, 자료 관리 등
    - ☞ 전수조사의 문제점

※ 과학적, 합리적 표본조사 : 전수조사보다 정확한 결과 도출 가능.

▶ 소규모 모집단의 경우, 전수조사가 정확성 측면에서 유리함



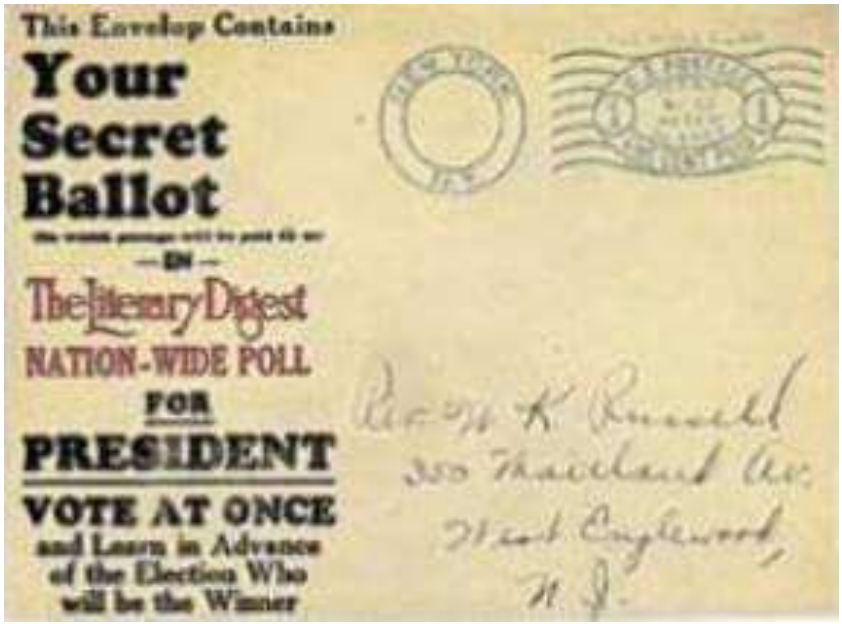
표본이 모집단 특성을 잘 설명할 수 있는가?

- 표본조사의 정확성: 과학적인 **표본추출 및 분석**이 요구됨

예) **1936년 미국대통령 선거**

- 공화당의 랜던(Landon)과 민주당의 루즈벨트(Roosevelt)
- 'Literary Digest' 구독자, 전화기, 자동차 보유자 대상 **천만명**에게 엽서 발송해 회수된 **2,376,523**명 의견 분석결과 **랜던 57%**, **루즈벨트 43%**
- 'Gallup poll' 표본조사를 토대로 루즈벨트 56%, 랜던 44%
- 선거결과에서 **루즈벨트 63%**, **랜던 37%**

※ Gallup이 루즈벨트의 당선을 예측했지만 예측한 득표율과 실제 득표율 간에 차이가 현재의 여론조사방법에 비해 상당히 큼 (할당추출법)


This Envelop Contains
Your
Secret
Ballot
—IN—
The Literary Digest
NATION-WIDE POLL
FOR
PRESIDENT
VOTE AT ONCE
and Learn in Advance
of the Election Who
will be the Winner

확률추출(probability sampling)

o 연구목적에 필요한 표본자료(sample data)를 여건이나 상황, 정확성 등을 고려하여 수집

☞ 표본론(sampling theory), 실험계획법(experimental design)

o 모집단을 이론적으로 설명할 수 있는 표본추출법: "확률추출법"

(- 단순확률추출, 계통표본추출, 층화확률추출, 집락표본추출)

o 어떤 표본이 선택되는가에 따라 결과에 차이가 발생

☞ 표본추출오차(sampling error)

o 표본에서 얻은 정보를 근거로 모집단의 특성에 대해 추론

☞ 통계적 추론(statistical inference) "확률이론"을 기초로 전개됨

☞ 확률추출방법에 의해 얻어진 표본자료가 필요함!!!

확률추출(probability sampling)

random number 추출프로그램?

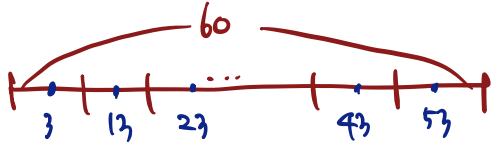
o 단순확률추출 (simple random sampling)

- 모든 개체가 추출될 확률이 동일하도록

o 계통추출 (systematic sampling) 체계적추출법

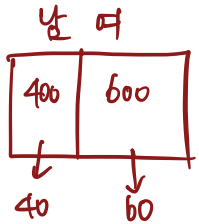
ex. 선거 출구조사

- 정렬 후 일정한 간격으로 추출



o 층화확률추출 (stratified random sapling)

- 모집단을 몇 개의 그룹(부모집단)으로 나눈 후 단순확률추출



o 집락추출 (cluster sampling)

- 개체들로 구성된 집락(clustet)를 기준으로 추출

ex. 고등학생의 사교육비조사 → 학급 (덩어리뽑기)

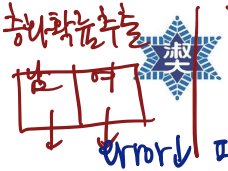
※ 비확률추출(non-probability sampling): 편의추출, 판단추출, 할당추출...

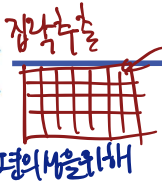
뽑히는 표본으로 조사자가 판단

(추출의추출)

20대남성중 편의로 100명

ex. 사교육비 층화확률추출





집락추출

학급

→ 학급3개뽑기

판단추출

층화집락추출

남/여로 집단구분 후 학급으로 집락추출

층화는 무작위 할당은 조사자마음

- 1948년 미국 대선: Truman vs Dewey (non-probability sampling)

Truman Wins CONTINUED

Chicago Daily Tribune Home

DEWEY DEFEATS TRUMAN

G.O.P. Sweep Indicated in State; Bonic Leads in City

THE JUBILANT PRESIDENT HOLDS UP A COPY OF NEWSPAPER EXTRA WHICH OVERCONFIDENT REPUBLICAN EDITORS RUSHED INTO PRINT ON NIGHT OF THE ELECTION

Kick Me!

ALL US POLITICAL EXPERTS

TALBURT IN SCRIPPS-HOWARD NEWSPAPERS

NEW YORK "WORLD-TELEGRAM'S" APOLOGY

PRESS AND POLLS WERE WRONG IN A LOUD VOICE

Almost without exception the political experts were abysmally wrong. Most agreed that they simply had not been as thorough in their analyses as they might have been. Many traced their downfall to the public-opinion polls of George Gallup, Elmo Roper and Archibald M. Crossley which, from convention days to election morning, unanimously predicted a Dewey victory. Harry Truman and a few of his most enthusiastic supporters said the polls were wrong, but columnists and commentators were so cocksure that they spent their time speculating about the lineup of Dewey's cabinet. The professional politicians were also fooled; many high-ranking Democrats were so depressed by the polls that they lifted not a finger to help their boss. Some Republicans were so elated that they too neglected to act. Local surveys showed that Republican candidates here and there were in trouble, but Dewey strategists never awakened to the fact that their ticket was in danger not merely in Minnesota or West Virginia, but everywhere.

A few prophets had the truth in their grasp but refused to recognize it. The Staley Milling Company of Kansas City ran a "pullet poll" among farmers, who bought chicken feed in sacks labeled Democratic or Republican. When Democratic feed pulled into a 54–46 lead, the company abandoned its poll in confusion. This also happened in Denver where a research institute discovered that Truman would carry the state by 3%, then spurned its findings and predicted a Dewey victory.

Washington MERRY-GO-ROUND

By Drew Pearson

(The author of this column is given credit for originating the term "inside dope" back of the mirror.)

IN NEW YORK, Nov. 2—I surveyed the close knit group around Tom Dewey, who will take over the White House 86 days from now.

The Lineup

© 1948, NEA SYNDICATE, INC.

DREW PEARSON, who has a fairly high batting average, struck out with this analysis of Dewey's "cabinet," which he had written on election day for the next morning's papers. Like wrong guessers Walter Winchell, Elmo Roper and George Gallup, Pearson was on a television broadcast and actually had to face his audience on election night.

MATTER OF FACT

By JOSEPH and STEWART ALSOP

WASHINGTON.

The first post-election question is how the government can get through the next ten weeks. Nowadays, unhappily, time's winged chariot has been equipped with jet propulsion. Events will not wait patiently until Thomas E. Dewey officially defeats Harry S. Truman. Precisely in the fields of foreign and defense policy, somebody somewhere in Washington must have authority to give answers that will still be valid after Jan. 20.

Oddly enough, the members of the outgoing administration seem to be more concerned with the urgent problem than the entou-

© 1948, NEW YORK HERALD TRIBUNE, INC.

THE ALSOP BROTHERS, Joseph and Stewart, also had an embarrassing column on the wires to newspapers for publication on Nov. 3. Most editors failed to kill it. Brother Joseph (right) cheerfully admitted their huge error, although he did point out that the Alsops had long ago predicted—but vastly underestimated—"a swing to the left."

42

통계학이란

- 관심 또는 연구의 대상이 되는 **모집단의 특성**을 파악하기 위해
- 모집단부터 일부의 **표본**자료를 수집하고
- 수집된 표본을 **정리, 요약, 분석**하여 표본의 특성을 파악한 후
- 표본의 특성을 이용해 모집단의 특성에 대해 **추론하는 이론과 방법**을 다루는 학문

※ 현대 통계학

- 자료에 근거를 둔 "**불확실성(uncertainty)**"에 대한 추론을 다루는 과학