

빅 데이터 혁신 공유 대학

통계학

한동대학교
창의융합교육원/전산전자
석좌교수 손중권



01 통계학 소개

1장 소개

- 1.1 Statistics란 무엇인가?
- 1.2 통계학의 분야
- 1.3 모집단과 표본
- 1.4 기본 용어들
- 1.5 변수의 형태

모든 분야에서는 고유의 용어가 있다. 통계학도 예외가 아니다. 이
장에서는 기본용어를 설명하고자 한다.

1-1 통계학의 정의와 역사에 대한 설명과 이해 1

- 통계란 데이터의 요약을 일컫는 일반적인 용어이다 즉 생활에서 듣는 많은 숫자들은 통계로 일컫는다.
- 그러나 통계학이란 어떤 상황이든 불확실성이 존재하는 모든 경우에 있어 데이터를 수집하고 때로는 그 수집의 방법 또한 연구하여 알맞은 데이터를 수집하며, 이 데이터를 기초로 불확실성에 대한 탐구를 하고 추정을 할 수 있다.
- 또한 통계학의 가장 큰 목적인 예측을 하기도 한다.

1-1 통계학의 정의와 역사에 대한 설명과 이해 1

- 사회 현상에 대한 추론 즉 추정과 검정을 하기도 한다.
- 이런 것들을 위해 수학을 기초로 한 과학적 방법을 통계학이라고 하며 유전학이나 과학적 문제에서 시작하였으나 사회과학에 활발하게 적용되어 자칫 사회과학으로 더러 오해를 하기도 한다.
- 21세기 지식시대와 빅데이터 문제를 푸는데도 그 사고 방식이 중요하게 적용되며 인공지능이 나오기는 1950년대이지만 통계적 방법이 활용되면서 극적인 발전을 가져왔다.
- 머신러닝에서도 역시 사용되며 따라서 통계적 사고 즉 귀납적 사고가 중요한 역할을 한다.

통계(統計, Statistic)

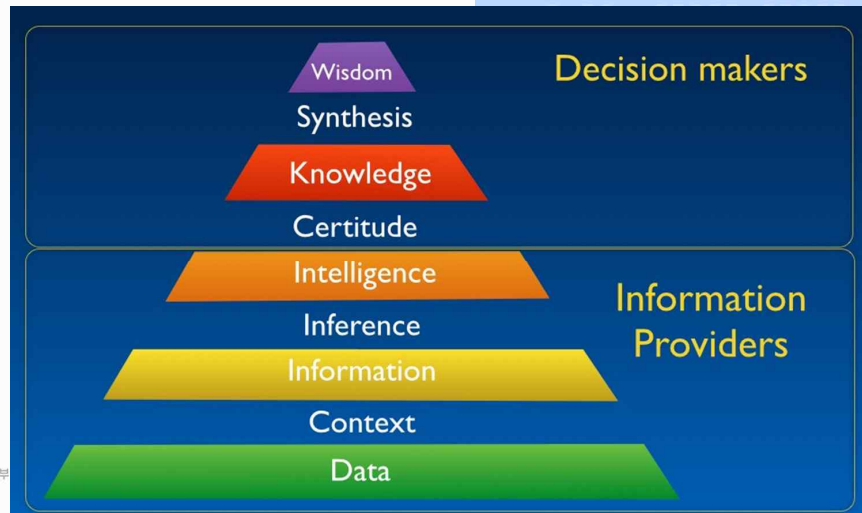
- 사회적 현상 혹은 자연 현상을 규명하기 위해 수집된 각종 데이터를 요약하거나 적절한 방법을 통하여 일차적으로(혹은 상대적으로 쉬운 방법으로) 가공 되어 나오는 정보를 통계라고 할 수 있으며 통계는 숫자나 그래프, 도표 혹은 그림 등 적절한 표현 방법을 통하여 나타낼 수 있다.
- 동일한 현상에 대하여 얻어지는 데이터에 따라 다양한 통계를 얻을 수 있으며 이를 종합하거나 따로 따로 해석함으로써 현상을 이해하는데 사용될 수도 있고 또한 예측하는데도 활용이 가능하다.

통계(統計, Statistic)의 종류

- | | |
|--------------|------------|
| • 소비자물가지수 | • 야구선수의 타율 |
| • 실업률 | • 팀의 승률 |
| • 조이혼률(粗離婚率) | • 시청률 |
| • 출산율 | • 정당별 지지율 |
| • 후보별 지지율 | • 각종 경제 지수 |
| • 한 해 출생자 수 | • 종합 주가 지수 |
| • 한 해 사망자 수 | • KOSDAC |
| • 수입 수출액 | |

통계(統計, Statistic)

- 어원 Status
- 뜻 State arithmetic



1-1 통계학의 정의와 역사에 대한 설명과 이해 2

• 통계학(統計學, Statistics)

- 자연과 사회적 집단 혹은 인간사회 등에서 나타나는 현상 등에서 보이는 불확실성(uncertainty)을 규명하기 위해 다양한 데이터를 기초로 수학 또는 확률론적 수단을 통해 학문적으로 분석하기 위한 설계.조사.분석.처리.추론에 대한 방법 또는 의사결정방법 또는 적절한 데이터의 가공을 통해 지식을 창출하는 방법을 연구하는 학문

통계학(統計學, Statistic^ㅸ)

- 기술 통계학 Descriptive statistics

도표나 그래프와 요약 측도 등을 이용하여 데이터를 구성하고 나타내며
설명하는 방법들로 구성되어 있음

=>원래의 자료세트는 매우 크므로 결론을 내리거나 결정을 하는데 별로
도움이 되지 않으므로 요약 테이블과 도표 등으로부터 결론을 쉽게 낼 수 있다.

통계학(統計學, Statistics)

- 추론 통계학 Inferential statistics
- 데이터로부터 예측을 하거나 가설 검정 등의 방법에 대한 연구를 하는 분야 :
- ex) 전형적인 대학 졸업생의 초기임금을 알기 위해 2,000명의 최근
대학졸업자를 뽑아서 이 정보를 근거로 결정

통계학에 대한 이해

	통계학	수학
동질성	수리적 방법의 접근 자연과 사회를 대상	
이질성	데이터를 기초로 함 귀납적 사고(deduction)	추상적 개념 연역적 사고(induction)

숫자는 살아있다!

모든 관찰은 숫자로 표현될 수 있으며, 이를 근거로 관심에 대한 어떤 것이든 예측 가능하다.

빅 데이터 혁신공유대학 |   교육부  한국연구재단

인류 최초의 통계

- BC 1440년경 모세에 의해 씌여졌다고 알려진 구약성서 중 모세5경(창세기, 출애굽기, 레위기, 민수기, 신명기) 가운데 시나이산 광야에서 인구를 조사한 내용이 민수기(The Numbers) 1장 1절에서 4장 49절까지와 26장 1절에서 65절
- 두 번째: BC 1030년경 역대상 21장에서 다윗이 인구조사를 함(21:1 사단이 일어나 이스라엘을 대적하고 다윗을 격동하여 이스라엘을 계수하게 하니라)

빅 데이터 혁신공유대학 |   교육부  한국연구재단

고대 사회에서의 통계(관청통계)

- 인구 · 농지 등의 사회사실의 수량적 조사 · 관찰
- 국가의 경제 · 사회적 규모에 대한 기술을 목적으로 한 국가학(國家學, Statskunde)이 생겨났다. 인구조사(Census)의 어원인 라틴어 Censere는 세금부과(taxation)의 의미를 지니며 이를 보면 당시 인구수가 조세능력에 직결

중근세에서의 통계

- 카알대제(742-814)에 의해 수행된 “王領일람표 (Capitulare de Villis)”와 같은 대규모 조사
- 중세말기가 되면서 여러 도시가 생성되어 조세와 징집 또는 장정수의 조사를 목적으로 시민부(市民簿, Burgerrolen)이 작성

우리나라 역사 속의 통계

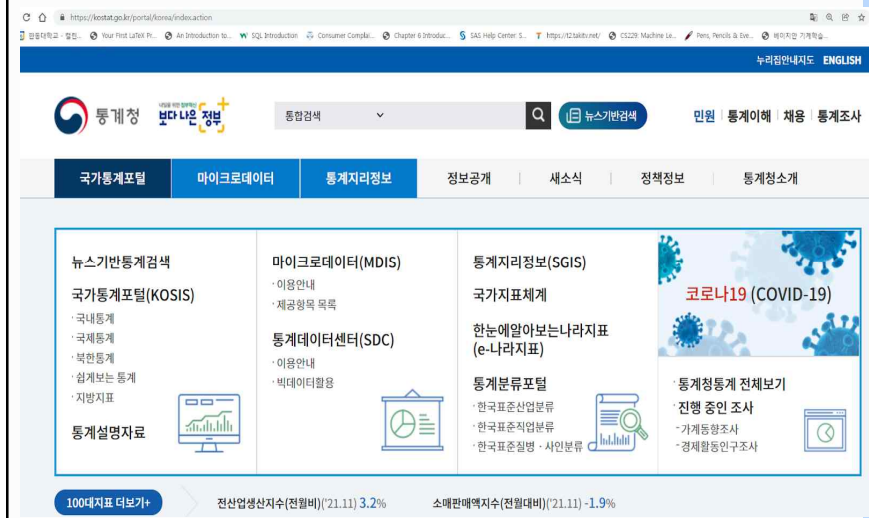
- 전한시대 낙랑군은 25현, 62,812호, 406,748명, 현도군은 3현, 4,500호, 22만여 명의 인구
- 삼국유사를 보면 신라 전성기 지금의 경주인 경중에는 178,936호가 있었던 것으로 기록
- 고구려 전성기 210,508호, 백제는 152,200호-동이보도(東史補道)

우리나라 역사 속의 통계

- AD 750년 경 경덕왕 민정문서(일본에 있음)에 의하면 신라의 촌락구조를 보면 인구로는 사해점촌이 남자 64명, 여자 78명, 살하지촌이 남자 47명, 여자 78명이며 또 서원경은 남자 47명, 여자 60명으로 나와 있다. 또 토지로는 사해점촌은 밭 102결, 논 63결, 소 22두, 말 25두이며 살하지촌은 밭 63결, 논 119결, 소 12두, 말 18두이고 서원경은 밭 29결, 논 78결, 소 8두, 말 10두

통계로 본 우리나라

참고: 통계청(www.kostat.go.kr)



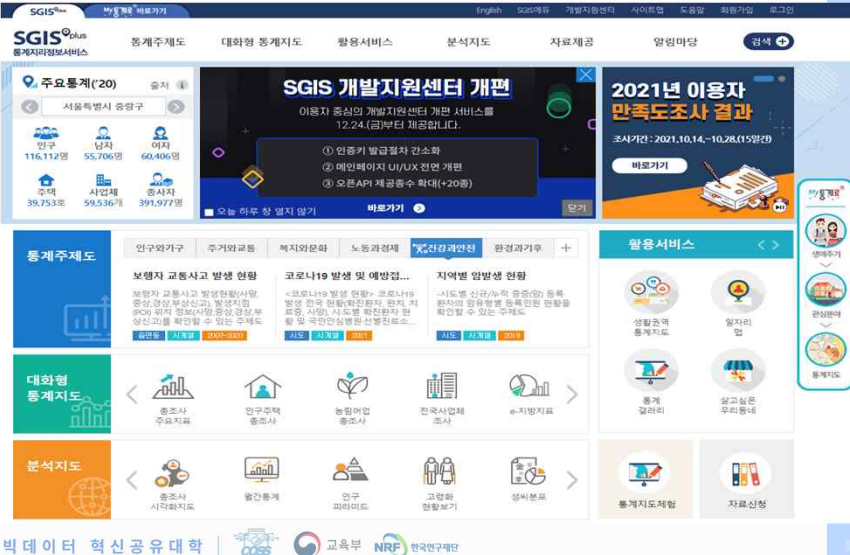
통계청은 우리나라
국가공인통계를
수집 정리하여
공표하고
또 중앙정부부처와
각 지방행정기관에
제공하여
정책에 반영하도록
하는 기관이다.

통계로 본 우리나라



통계 데이터
센터SDC는 각종
행정자료와
민간자료를
한 곳에 모으며
이용자 교육과
더불어 micro data를
민간에게 제공하는
곳이다.

통계로 본 우리나라

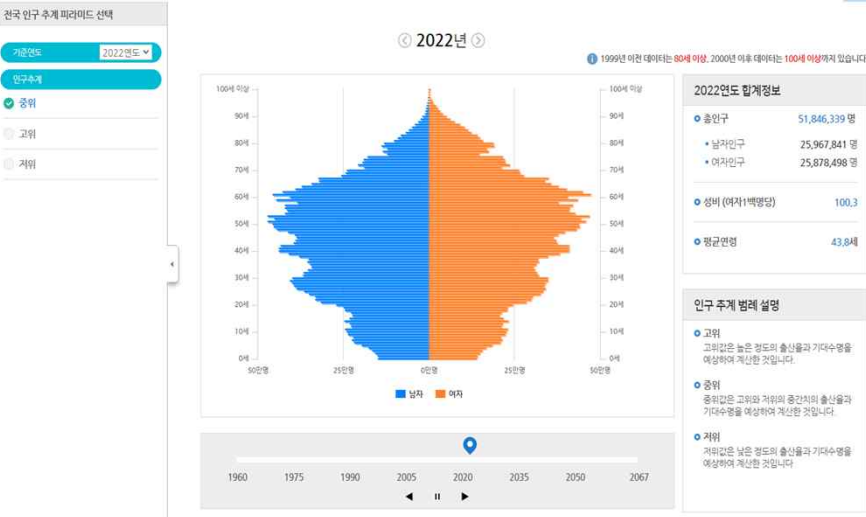


빅 데이터 혁신공유대학 | 교육부 | NRF | 한국연구재단

통계 지리정보 서비스는 각 주제별 지리들과 연관된 데이터를 제공하며 특히 대화형 데이터를 제공함으로써 사용자에게 맞춤형 데이터를 제공도 하며 일차적인 분석지도도 제공한다.

특히 움직이는 인구 피라미도도 제공하여 향후 인구 구성에 대한 정보도 제공한다.

통계로 본 우리나라



빅 데이터 혁신공유대학 | 교육부 | NRF | 한국연구재단

특히 움직이는 인구 피라미드도 제공하여 향후 인구 구성에 대한 정보도 제공한다.

연도는 2267년간 까지 또 각 광역시와 시도별 인구 피라미드를 보여줌으로써 시간의 흐름에 따른 인구구조의 변화를 볼 수도 있으며 지역별 비교도 가능하다.

단 인구감소가 당초보다 빨라 장기 예측은 오차를 감안해야 한다.

확률(Probability)

- 어떤 결과가 얻어질 가능성의 척도를 제공한다.
- 확률의 역사는 인류의 역사와 거의 비등할 정도로 오래되어 왔으며 중동지역 메소포타미아에서 많은 종류의 주사위가 발견되고 또 우리나라에서도 경주 동궁월지 (옛 안압지)에서도 목제 주령구가 발견되어 그 성질이 아직 다 탐구되자 않고 있다.
- 이와 같은 확률은 수학적 배경으로 21세기 초 수학자이자 통계학자인 콜모그로프에 의해 공리가 완성이 되었으며 그 전에 라플라스 등의 학자들이 정의했던 경험적 확률을 수리적 확률로 완성시켰다.
- 하지만 오늘날에는 확률의 해석이 다시 수리적인 것에서 경험적 확률로의 회귀 또는 재해석으로 이어지며 베이저안 통계학의 근본이 과거 고전통계학의 근본과는 다른 부분때문에 학자들 사이에 논란의 여지가 있다.
- 하지만 최근에는 베이저안적 해석인 경험적 확률 또는 가능성으로 재해석이 되면서 수학적 배경에는 위배가 될 수는 있으나 인공지능 등 인간의 학습과정을 감안하면 경험적 가능성으로 해석할 수도 있다.

빅 데이터 혁신공유대학 |   교육부  한국연구재단

통계가 왜 필요해?

- 사례 . 신의 섭리**
- 18세기 중엽 프러시아의 종군목사인 쥘스밀흐(John P. Sussmilch)는 신의 섭리를 증명하는데 통계를 활용
- 도박의 예**
- 16세기 이탈리아의 도박사인 카르다노는 3개의 주사위 던지기를 해서 합이 9에 거는 것이 10에 거는것 보다 불리(모두 다 6가지 방법)함을 경험적으로 증명

빅 데이터 혁신공유대학 |   교육부  한국연구재단

사례 : 기성복의 감추어진 이야기

아내는 키가 작다 아내는 키가 작다.
자꾸만 작다.
딸 셋아들 하나를 낳는 동안 아내의 키는 조금도 자라지 않고
놀랍게 변모한 가위질 솜씨 번번히 아니 맞는 기성복
아랫 도릴 자르는 대목에 가면 낮 꿈 만리 호랑나비 나른다.

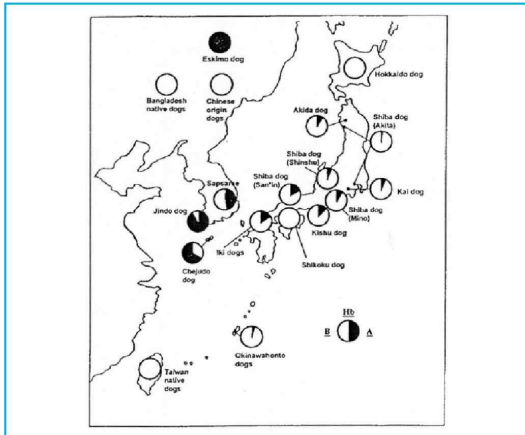
밤이 되면 아내는 사각사각 가위질에 신들린다.
그러면 나는 그만 자자커니 아내는
나머지 한 쪽마저 베이고 말겠다커니
우리들의 이견은 순전히 아내의 작은 키
때문에 철 철 철 생혈(生血)은 차라리 꽃이다.

빅 데이터 혁신공유대학 |   교육부  한국연구재단

사례 : 삽살개 우리 토종개 맞아?



한 때 삽살개가 우리나라 토종이 맞느냐로 논쟁이 일어난 적이 있다.
삽살개는 우리나라 천연 기념물로서 지정되어 있으며 신라시대 궁중에서 키웠던 기록이 남아 있다.
반면 진돗개는 일제 강점기 내선일체의 부산물로 일본개인 아끼다와 흡사한 모습을 띄고 있다. 그 논쟁은 다음의 통계분석 결과가 말해 주고 있다.



위의 그림은 각 견종별 혈액을 분석한 것으로 왼쪽은 헤모글로빈의 종류별로 분류한 것인데 삼살개나 제주개 등은 일본 개와는 완전히 다르게 구성되어 있다. 오른쪽의 그림은 통계분석을 한 결과 삼살개는 비록 외모는 다르나 진돗개와 한 그룹으로 묶이는 것을 볼 수 있다. 이런 통계분석이 논쟁을 마치게 할 것이다

빅데이터 혁신공유대학 |   교육부 |  NRF 한국연구재단

통계학이 왜 필요해?

- 웰즈(Herbert G. Wells,1866-1946)
- "Statistical thinking will be one day be as necessary for efficient citizenship as the ability to read and write."
- 당시 영국은 시민권 즉 투표권이 있는 시민과 자유민이나 없는 신분과 노예 신분으로 나누어져 있었는데 읽고 쓰는 능력이 없으면 투표권을 줄 수 없다는 사회 분위기가 있을 때 타임머신 등의 소설로 유명한 웰즈는 통계적 사고 능력이 쓰고 읽는 능력만큼 중요할 것이다 라고 예언처럼했다. 나이팅게일이 옥스포드대학에서 통계학으로 박사학위를 받고 보건행정에 일생을 헌신한 통계학 분야의 앞선 나라였다.

빅데이터 혁신공유대학 |   교육부  NRF 한국연구재단

21세기에서의 통계학

- From a Keynote Presentation by Hal Varian - Chief Economist, Google, to the 2008 Almaden Institute –
- Statistics - Dream Job of the next decade



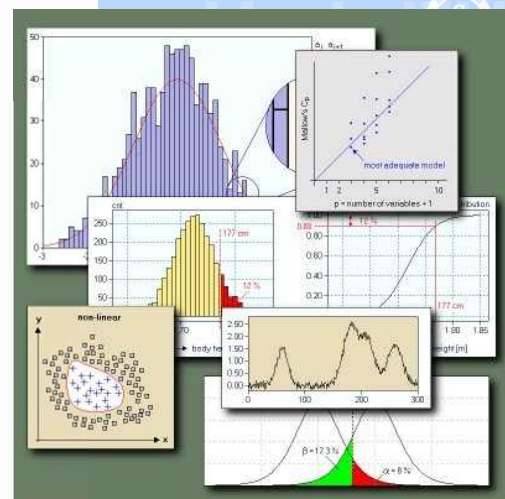
I am an emeritus professor in the School of Information, the Haas School of Business, and the Department of Economics at the University of California at Berkeley.

빅 데이터 혁신공유대학

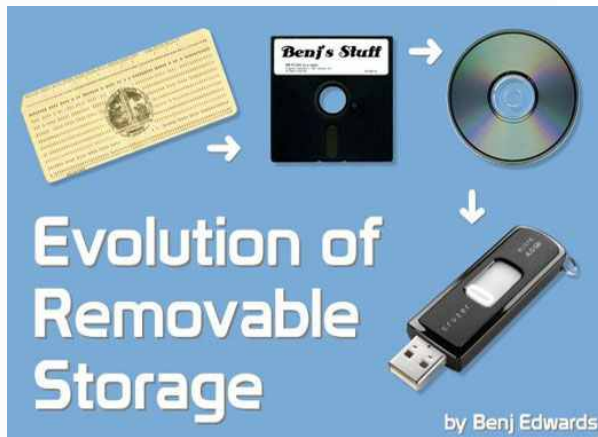
빅데이터 BIG DATA



빅 데이터 혁신공유대학 | 교육부 NRF 한국연구재단



대용량 데이터 시대



빅 데이터 혁신공유대학 | 한국연구재단

대용량 데이터 시대

1 The accelerating pace of change ...

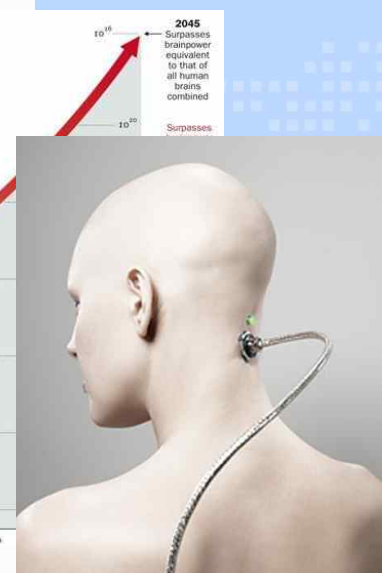
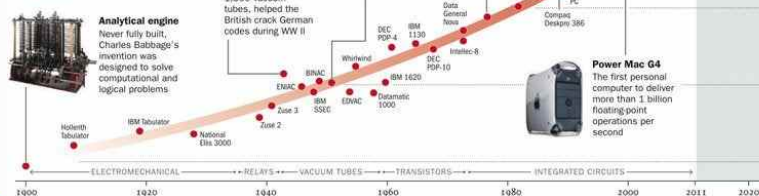


2 ... and exponential growth in computing power ...

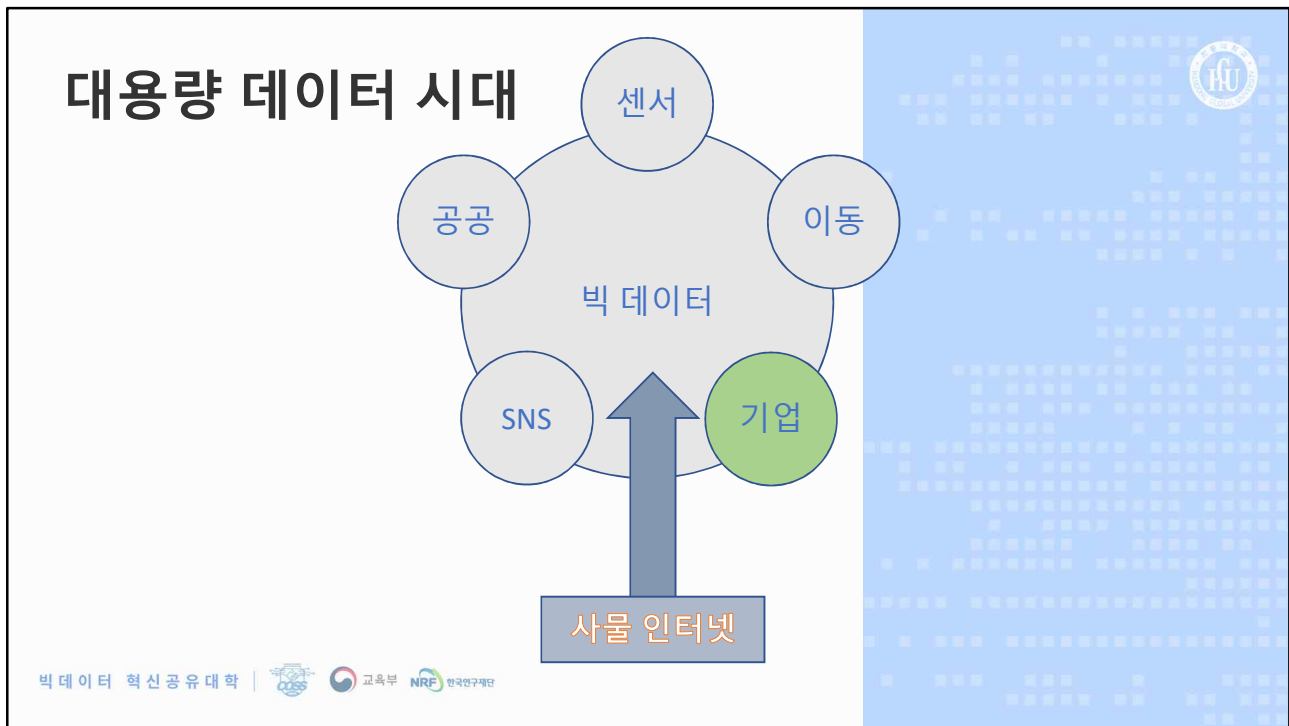
Computer technology, shown here climbing dramatically by powers of 10, is now progressing more each hour than it did in its entire first 90 years

COMPUTER RANKINGS

By calculations per second per \$1,000



빅 데이터 혁신공유대학 | 한국연구재단



Ordenes de magnitud de la Información (datos):

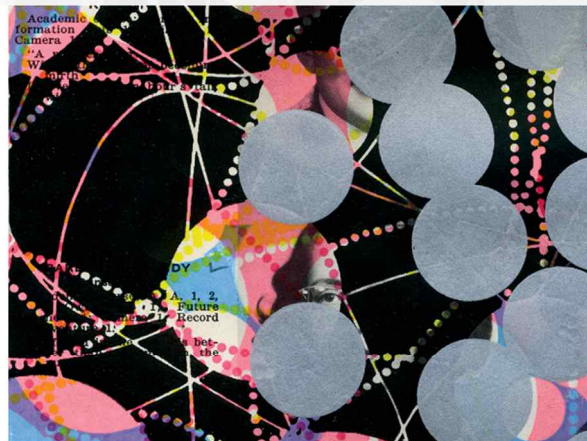
- 1 **Bit** (es la unidad mínima de almacenamiento, 0/1)
- 8 Bits = 1 **Byte**
- 1024 Bytes = 1 **Kilobyte** (un archivo de texto plano, 20 kb)
- 1024 Kilobytes = 1 **Megabyte** (un mp3, 3 mb)
- 1024 Megabytes = 1 **Gigabyte** (una película en DivX, 1 gb)
- 1024 Gigabytes = 1 **Terabyte** (800 películas, 1 tb)
- 1024 Terabytes = 1 **Petabyte** (toda la información de Google, entre 1 y 2 petabytes)
- 1024 Petabytes = 1 **Exabyte** (Internet ocupa entre 100 y 300 Exabytes)
- 1024 Exabytes = 1 **Zettabyte** (a partir de aquí no existen comparativas reales)
- 1024 Zettabytes = 1 **YottaByte**
- 1024 YottaBytes = 1 **Brontobyte**
- 1024 Brontobytes = 1 **GeopByte**
- 1024 GeopBytes = 1 **Saganbyte**
- 1024 Saganbytes = 1 **Jotabyte**

DOOGB

빅 데이터 혁신공유대학



Harvard
Business
Review

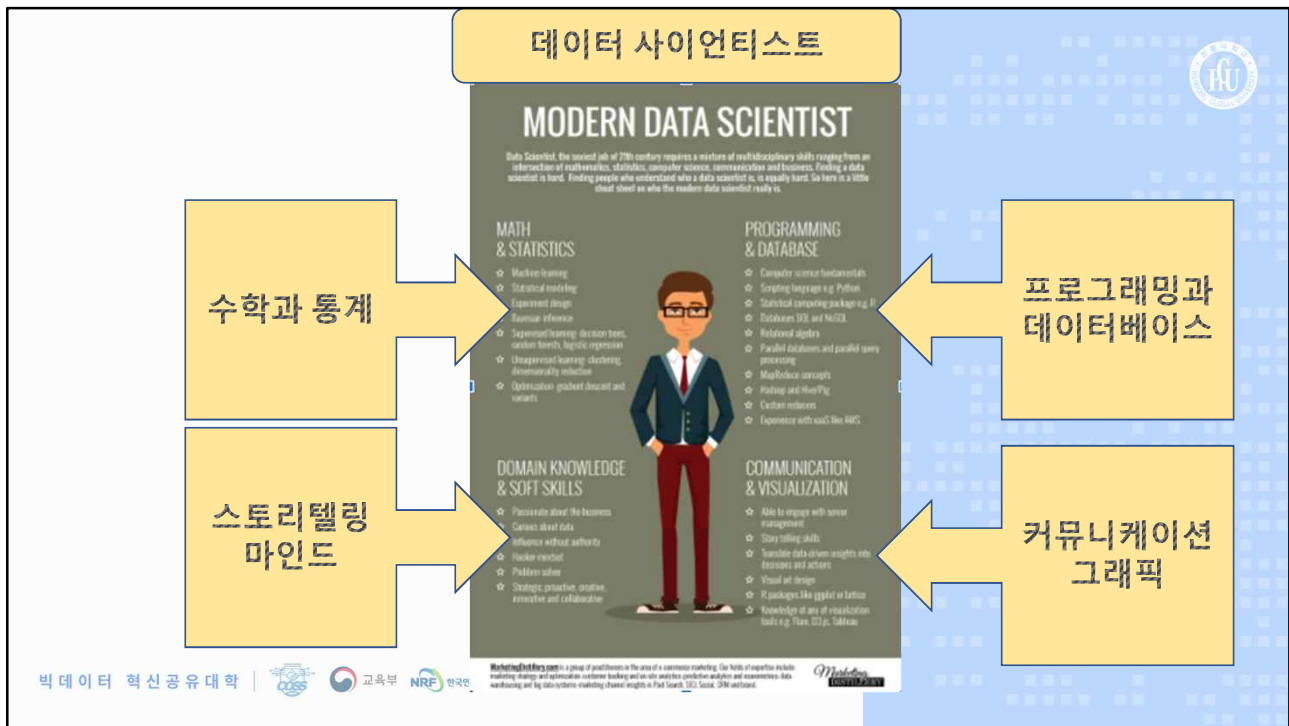


DATA

Data Scientist: The Sexiest Job of the 21st Century

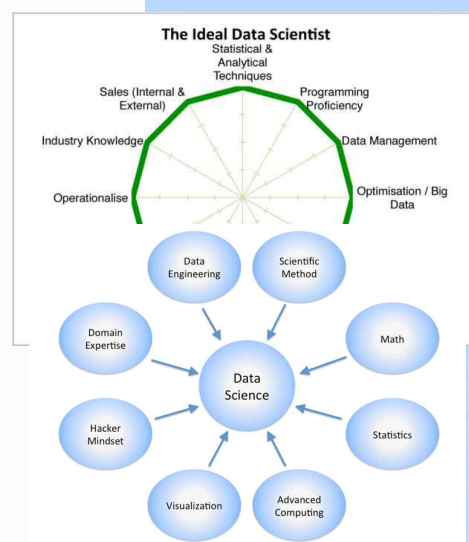
by Thomas H. Davenport and D.J. Patil

빅 데이터 혁신공유대학



좋은 데이터 사이언티스트란?

1. 통계 분석 능력
2. 프로그래밍 능력 -R, Python
3. Data Management
4. 수리통계학 배경
5. 수학적 이론
6. Big Data Analysis 기술
7. 최적화 수법
8. 프로젝트 운영 능력
9. 소통 능력
10. 앞을 내다 보는 능력



모집단과 표본의 정의와 개념

- 모집단 혹은 목표 모집단 (population or target population)
 - : 조사대상이 되는 모든 대상 전체를 일컫는 말
 - 예) 조사대상이 되는 각종 형태의 시장 전체
 - 각종 선거에 투표권이 있는 대한민국 국민
- 표본(sample)
 - : 조사를 위해 모집단에서 취해진 일부

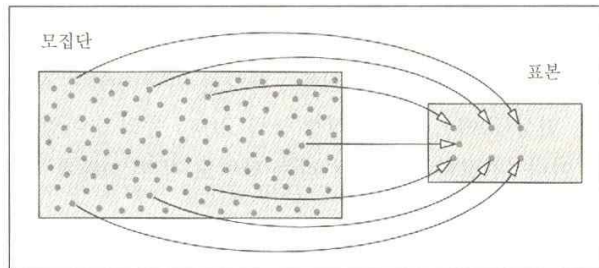


그림 1-1. 모집단과 표본

빅 데이터 혁신공유대학 |   교육부  한국연구재단

모집단과 표본의 정의와 개념

- 원소(element)
 - 모집단이나 표본을 구성하는 최소한의 객체로 사람이나 사회, 자연 등 이 있다.
- 변수(variable)
 - 각 원소를 대상으로 측정하고 자는 또는 관심이 있는 특성을 일컫고 주로 미지의 값들이다.
 - 변수는 일변량, 이변량 등의 변수와 값이 일정한 상수가 있다.
- 관찰값(observation)
 - : 각 원소들로부터 측정되거나 얻어진 변수들의 값으로 하나의 관찰값에는 여러 변수의 값으로 구성될 수도 있다.
- 데이터세트(data set)
 - : 한 가지 혹은 그 이상의 변수에 대한 관찰값들을 모아 둔 것

빅 데이터 혁신공유대학 |   교육부  한국연구재단

표본 추출 방법

- 복원추출(with replacement)
 - 표본을 뽑을 때 한 번 채택된 원소를 다시 모집단에 포함시킨 뒤
 - 뽑는 경우로 중복추출이 가능하다.
- 비복원추출(without replacement)
 - 한 번 뽑힌 원소는 제외하고 나머지 가운데서 표본을 뽑는 경우
- ※ 여론조사의 경우 복원추출 시 한번 의견을 제시한 사람을 다시 뽑는 것은 대표성을 상실한다.

2.1 모집단과 표본 -1

- 모집단(母集團, population)
- 표본추출단위(標本抽出單位, sampling unit)
 - 원소도 될 수 있으며 일단의 집합 또한 단위가 될 수 있음
- 표본(標本, sample)
- 모수(母數, parameter)
 - 관심있는 모집단의 특성으로 대부분 미지(unknown)임
- 변수(變數, variable)는 실험단위(experimental unit)로부터
 - 측정되어진 특성

- **전수조사(Census)**

- 모집단의 모든 원소를 대상으로 조사하는 방법으로 주로 인구조사에 사용된다

- **표본조사(sample survey)**

- 모집단을 효율적으로 조사하기 위해 적절하게 일부를 선택하여 조사하는 방법
- 예) 각종 여론 조사의 경우 1,000명을 대상으로 조사한다

- **대표표본(representative sample)**

- 모집단의 특성을 가능한 한 근접한 특성을 지니는 표본을 일컫는 말로써 선거의 예측 등에 활용된다

- **단순랜덤표본(simple random sampling)**

- : 모집단에서의 각 원소가 표본으로 뽑힐 확률이 동일하게 주어진 상태에서
- 표본을 택한 경우 단순랜덤표본(simple random sample)이라고 함

- **측정 : 관심의 대상에 숫자를 부여하는 작업**

- **1) 명목척도(nominal scale)**

분류에 목적이 있는 있음 예) 성별, 국적별

- **2) 순서척도(ordinal scale)**

순서 부여에 목적이 있는 척도 (대소관계가능)

- **3) 구간척도(interval scale)**

일정한 구간의 개수만큼 나타내는 경우 (+, - 연산가능)

예) 길이, 무게, 섭씨 혹은 화씨 온도

- **4) 비율척도(ratio scale)**

일정한 양의 비례로 표현되는 경우 (\times , \div 연산가능)

예) 절대온도(켈빈온도), 길이와 무게는 구간척도인 동시에 비율척도임

- **양적변수(quantitative variable)**

숫자로 표현될 수 있는 변수

- **1) 이산형(discrete type)**

헤아릴 수 있는 값을 가지는 양적 변수로 주로 정수값을 가짐

예) 가구당 자동차 보유 대수, 일주일 교통 사고건수

- **2) 연속형(continuous type)**

주어진 구간에서 임의의 값을 가질 수 있는 변수로 실수값을 가짐

예) 사람의 키, 몸무게

- **질적 또는 범주형 변수(qualitative or categorical variable)**

2가지 이상의 범주로 나누어 측정하며 관찰된 횟수로 분석이 가능한 변수

예) 성별, 혈액형

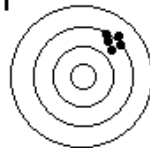
- **표본조사의 측정 결과**

- **편의(偏倚: bias)**는 반복해서 표본을 추출할 때 각 표본에서 구해지는 통계치가 참값으로부터 계속 벗어나는 양

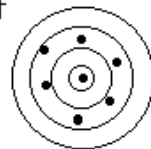
- **정도(精度: precision)**는 반복해서 표본을 추출할 때 각 표본에서 구해지는 통계값과 모수 사이의 변동

2.2 모집단과 표본 -2

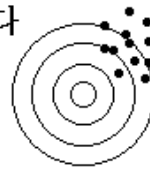
- 편의와 정도는 무관하다. 다음의 그림을 보자.
- 과녁 중앙은 참값으로 보고 각각의 점들은 관찰치로
- 간주하면 정도와 편의에 대해 생각해 볼 수 있다. 가



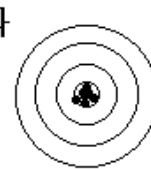
나



다



라



빅 데이터 혁신공유대학 |   교육부  한국연구재단

- 가의 경우는 과녁의 중심으로 부터 벗어나 있으나 값들이 모여 있음으로 정도는 비교적 높다고 할 수 있다.
- 나의 경우는과녁의 중심부로 부터 치우침이 없이 양 사방으로 골고루 흩어져 있어 편의 즉 치우침은 없으나 너무 많이 흩어져 있기에 정도는 낮다고 할 수 있다.
- 다의 경우는 한 방향으로 치우쳐 있으면서 값들이 흩어져 있기 때문에 정도도 낮으며 편의도 있다고 할 수 있다.
- 라의 경우는 과녁 정 중앙에 값들이 좁게 모여 있음으로 정도는 높고 편의는 없다고 할 수 있다.

빅 데이터 혁신공유대학 |   교육부  한국연구재단

• 측정의 타당성과 신뢰성

- 타당성(validity) : 측정된 데이터가 주어진 목적에 적합한가
- 내적 타당성
- 신뢰성(credibility): 한 대상에 대한 측정값이 얼마나 일관성 있게
- 그리고 안정적으로 얻어지는가 의 여부

2.3 모집단과 표본 -3

• 표본추출의 필요성

- 어느 후보가 시장 혹은 국회의원이 될지를 모든 유권자에게 다 물어 보자. 어휴, 이러다간 날 새겠네.
- 국 간볼까? 휘이 저어서. 그런데 왜 져나?
- 김장김치 맛보기. 음식 맛은 어머니 손맛이라는데. 무우를 살 때 다 잘라봐 어디? 배추도 맛있나 다 먹어보고?
- 사과 한 상자의 맛이나 수박 한 덩이는 한 입만 먹거나 한 조각만으로도 맛을 알 수 있는가? 왜 다 먹어야지 맛본다고 하지.

표본추출의 필요성

- 그 집 주부의 부지런함과 청결함은 그 집 화장실을 보라는데
- 달에도 물은 있다. 그럼 화성엔?
- 대통령 선거의 유권자가 3,500만명이 더 되는데 다 어찌 찾아가서 물어볼까?
- 팔의 개수를 정확히 센 사람이 누구지? 한 흙의 개수로 나머지를 헤아린 사람 - 오성과 한음 이야기
- 시간이 많이 드네. 중국은 어쩌나? 인구가 14억이라 인구센서스 후 합치는 데만 해도 3년은 족히 걸리는데.
- 돈은 어찌고?
- 표본으로 조사해도 아주 정확한데 뭘.
- 전수조사보다 오히려 더 정확할 수가 있다?

빅 데이터 혁신공유대학 |   교육부  한국연구재단

표본추출의 필요성_3

- 성범죄가 몇 배나 늘었다는데?
- 일본뇌염 비상! 어떻게 비상인줄 알지?
- 실업률은 얼마나 되나?
- 미사일 실험을 위해 30기를 쏘아 올려야 합니다. 흠, 그 정도로 되나? 마구 쏘아 올려 봐야지. 그런데 한 기당 100만불이 넘어요! 우리나라는 한 기에 30억원 하는 개발중인 미사일을 결국 10발 다 쏘고나서야 비로서 무기화 할 수 있었다. 반면 미국은 크루즈미사일에 4기만으로 충분했다.

빅 데이터 혁신공유대학 |   교육부  한국연구재단

2.4 모집단과 표본 -4

- **표본오차(sampling error)**
- **편의추출(便宜抽出, convenient sampling)**
일정한 절차를 무시한 채 임의로 조사의 편리성에 치우친 방법
- **편의추출(偏倚biased sampling)**
모집단의 대표성을 나타내기 보다는 특정한 집단을 선택함으로써
참값을 얻을 수 없는 방법
- **모집단의 동질성**
모집단은 마치 수박처럼 전체가 동질일 경우가 거의 없다. 따라서 어떤 부분을
조사하는가에 따라 그 값이 달라지며 대표성의 유무가 문제가 된다. 이런 동질성을 최대한
확보하는 것이 가장 중요하다.

빅 데이터 혁신공유대학 |   교육부  한국연구재단

- **표본오차(sampling error)의 축소**
- 편의를 제거한다. 즉 표본을 택할 때 모집단의 특정부분에 치우치지 않게 한다.
- 표본을 얻을 때마다 결과는 서로 다르며-표본추출변동(sampling variation)-때로는
모집단을 잘 대표하지 못하는 표본을 얻을 수도 있으므로 표본추출 과정에서
조심해야 한다.
- 표본의 크기를 크게 하면 되나 이런 경우 표본조사의 의의를 상실할 수 있다.

빅 데이터 혁신공유대학 |   교육부  한국연구재단

- **비표본오차(Nonsampling Error)**

- **무응답오차(nonresponse error)**

결측치는 전수조사에서도 일어날 수 있다

- **응답오차(response error)**

응답 할 때 표기 실수 등을 말한다

- **처리오차(processing error)**

컴퓨터의 입출력 오류 등을 일컫고 OMR 카드의 입력 오류 등이 있다

표본 조사 방법

- **단순랜덤추출법(SRS)**

- 모집단 N개, 표본 n개
- 모든 원소들이 표본으로 뽑힐 가능성이 동일하게 함
- 모집단이 큰 경우에는 곤란하다
- 모집단의 구성 형태에 따라 최선이 아닐 수도 있다
- 예)공장에서 계속해서 생산되는 전구
- 주로 난수표(컴퓨터)를 이용한다

- 장점: 쉽고 경비나 시간이 가장 적게 소요되며 모집단이 동질인 경우 가장 좋은 방법이 될 수 있다
- 단점: 모집단이 동질이 아니면 편의 즉 치우침이 일어날 가능성이 많아 기에 많이 사용되는 방법은 아니다

표본 조사 방법

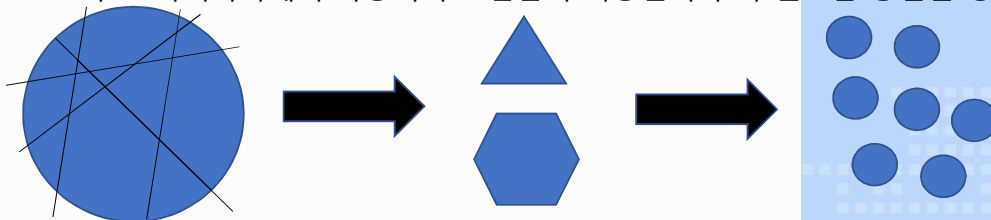
• 계통추출법(systematic sampling)

- 모집단 1, 2, ..., N 을 n 개 구간으로 나누어서 추출
- 예 : 20000개 중 $20000/100=200$ 개씩 100개 구간으로 나누어서,
- 1 구간(1, 2, ..., 200) : 51
- 2 구간(201, 202, ..., 400) : 251
- 100 구간(19801, 19802, 20000) : 19851
- 표본 선택과정이 SRS보다 간단하나 주기적으로 일어나는 제품 생산의 경우는 표본으로서의 대표성이 문제가 될 수 있다

표본 조사 방법

• 집락추출법(cluster random sampling)

- 모집단을 몇 개의 집락(cluster)으로 구성된 경우
- 주로 사회과학에서 사용되며 모집단이 비동질이나 각 원소는 동질인 경우 좋다



- 예 : 서울시내 가구를 대상으로 조사할 경우
- (1) SRS 불가(일련번호 부가가 불가능)
- (2) 서울시 25개의 구 중에서 5개 구를, 각 구에서 4개의 동을, 각 동에서 50개 가구를

표본 조사 방법

• 층화추출법(stratified random sampling)

- 모집단이 이질적 원소들로 구성되어 있을 때
- 유사한 것끼리 몇 개의 층(stratum)으로 나눈 후 각 층에서 SRS
- 표본의 크기는 각 층별 크기에 비례한다
- 시간과 경비 등 모든 면에서 일반적으로 적절하다고 판명되어 많이 사용되는 방법으로 주로 2단계 층화 추출법을 한다. 예를 들어 1단계에는 지역 인구수 비례대로 나누고 2단계에서는 해당 지역에서 남녀 비율대로 다시 표본 수를 정하여 SRS 한다.

