1. HISTORY

(1) 회귀(regress)의 사전적 의미 “go back to an earlier and worse condition” (옛날 상태로 돌아감) – “평균”으로 회귀

(2) 영국의 유전학자 Francis Galton(1822-1911)의 연구 -(처음에는 sweat pea)

Diameter of Parent Seed (0.01 inch)	Diameter of Daughter Seed (0.01 inch)	Frequency
21.00	14.67	22
21.00	15.67	8
21.00	16.67	10

Height of adult child in inches	Height of midparent in inches											
	<64.0	64.5	65.5	66.5	67.5	68.5	69.5	70.5	71.5	72.5	>73.0	Totals
>73.7	—	—	—	—	—	—	5	3	2	4	—	14
73.2	—	—	—	—	—	3	4	3	2	2	3	17
72.2	—	—	1	—	4	4	11	4	9	7	1	41
71.2	—	—	2	—	11	18	20	7	4	2	—	64
70.2	—	—	5	4	19	21	25	14	10	1	—	99
69.2	1	2	7	13	38	48	33	18	5	2	—	167
68.2	1	—	7	14	28	34	20	12	3	1	—	120
67.2	2	5	11	17	38	31	27	3	4	—	—	138
66.2	2	5	11	17	36	25	17	1	3	—	—	117
65.2	1	1	7	2	15	16	4	1	1	—	—	48
64.2	4	4	5	5	14	11	16	—	—	—	—	59
63.2	2	4	9	3	5	7	1	1	—	—	—	32
62.2	—	1	—	3	3	—	—	—	—	—	—	7
<61.7	1	1	1	—	—	1	—	1	—	—	—	5
Totals	14	23	66	78	211	219	183	68	43	19	4	928

(3) 부모의 키와 자녀의 키 사이 관계 연구 : 928명의 성인 자녀 키(여자는 키에 1.08 배, 행)와 부모 키(아버지 키와 어머니 키의 평균, 열)를 조사하여 다음 표를 얻었음

(4) Galton은 표를 관찰한 결과 : 키는 일정한 수준이 이상이면 무한정 커지는 것이 아니고 이정 수준 이하이면 무한정 작아지는 것이 아니라 전체 키 평균으로 돌아가려는 경향

(5) 표에서 직선이 중앙 일정 구간이외에는 중심으로 향해 낮아짐 – 즉 중심으로 회귀하려는 경향

(6) 그가 제안한 분석 방법의 이름을 “회귀”분석이라 명명하였다.

(7) Karl Pearson(1903) : Galton의 아이디어 수리적 모형화 및 추정방법 OLS(ordinary least square **최소자승법**) 제안, 1,078명 부자 키 데이터 활용

$$Son\_H = 33.73 + 0.516 * F\_H$$

(8) 아버지의 키가 1인치 커지면 아들 키는 0.516인치 커짐 – 기울기 개념

2. 데이터 DATA

(1) 데이터 행렬 – 행은 개체 subject, 열은 변수 variable, 셀은 관측치 값 observation, observed value

(2) 예제 데이터 : SMSA.xls

변수	변수명	변수내용
종속변수	Mortality	사망률
기후	JanTemp	1월기온
	JulyTemp	7월기온
	RelHum	상대습도
	Rain	강우량
사회경제	Education	교육수준
	PopDensity	인구밀도
	NonWhite	비백인비율
	WC	화이트칼라 비율
	pop/house	가구당 가족수
	income	수입
환경	HCPot	오염물질1
	NOxPot	오염물질2
	SO2Pot	오염물질3

city	Mortality	JanTemp	JulyTemp	RelHum	Rain	Education	PopDensity	NonWhite
Akron, OH	921.87	27	71	59	36	11.4	3243	8.8
Albany-Schenectady-Troy, NY	997.87	23	72	57	35	11	4281	3.5
Allentown, Bethlehem, PA-NJ	962.35	29	74	54	44	9.8	4260	0.8
Atlanta, GA	982.291	45	79	56	47	11.1	3125	27.1
Baltimore, MD	1071.289	35	77	55	43	9.6	5441	24.4

3. 변수

(1) 수리 통계

- 이산형(discrete): 측정 결과를 셀 수 있는 경우이다. 성별, 직업, 교통량, 나이 등이 여기에 해당한다.
- 연속형(continuous): 측정 결과가 무한이(infinite) 많이 발생, 즉 변수의 범위(range) 중 어떤 구간을 설정하더라도 측정치가 발생할 수 있는 경우로 키, 몸무게, IQ, 소득 등이 여기에 해당된다.

(2) 자료 분석

- 측정형 변수(metric, measurable, quantitative): 측정 가능한 특성 – 키, 몸무게, 평점, IQ, 교통량, 사망자 수가 그 예이다. 연속형 변수는 모두 측정형 변수, 교통사고 건수 등은 이산형이며 측정형 변수

  (a) 비율 ratio – 값의 배율이 의미 있음, 몸무게 : 100kg은 50kg의 두 배

  (b) 구간 interval – 배율 의미 없음, 온도 : 온도 20도는 10도보다 2배 덥다고 할 수 없음

- 분류형(범주형) 변수(non-metric, classified, categorical, qualitative): 개체를 분류하기 위해 설정된 변수

  (a)명목형(nominal): 개체를 분류만 한다. 성별, 결혼여부, 학력

  (b)순서형(ordinal): 순서를 가진다. 성적(A, B, ..) 소득수준(상, 중, 하), 리커트 척도(5점, 매우 만족, 만족, 보통, 불만족, 매우 불만족)

(3) 시간

- 데이터가 시간적 순서를 가지면 이를 시계열 자료(time series)라 하고 그렇지 않은 경우를 횡단면 자료(cross-section: 일정 시간에 한꺼번에 조사)라 한다.
- 경제 지표(환율, 수출량)나 기업의 연차별 자료(연도별 매출액), 주가 등이 시계열 자료에 해당된다.
- 분석 방법으로는 시계열 자료 분석, 계량 경제(econometrics) 방법(시계열 데이터에 대한 회귀분석) 등이 있다.

(4) 인과 관계 Causal Relationship

- 변수명 X - 통계 모형의 인과 관계(casual relationship)에서 원인이 되는(영향을 주는) 변수를 설명변수(exploratory variable), 독립변수(independent), 목표변수(target), 외생변수(exogenous)
- 변수명 Y - 결과나 영향을 받는 변수를 종속변수(dependent), 반응변수(response), 내생변수(endogenous)
- 분산 분석에서 설명 변수는 처리 효과, 요인으로 불리어진다.
- 일반적 통계 모형 - $Y = f(X_1, X_2, ..., X_p) + e$ - 함수 f 형태 중 해석이 용이하고 간단한 모형이 선형
- 인과 관계는 이론적, 경험적 타당성에 근거하여 데이터 수집 전에 이론이나 경험에 의해 설정되는 것이지 분석 후 설정되는 것은 아님
- 예를 들어 보자. 통계학과 학생 40명의 수능성적과 용돈을 조사하였다. 수능성적-Y, 용돈을 X로 하여 회귀분석을 실시한 결과 유의한(significant)가 있다고 해서 용돈이 수능성적에 영향을 미친다고 할 수 있나?
- 고전적인 예제 - 흡연과 암의 인과관계 : 횡단시점 분석, 암 환자의 흡연비율과 일반인 흡연비율 차이 비교 - 그러나 숨은 변인 효과
- association (연관)이 인과 (causation)를 의미하는 것은 아님 - 두 변수의 관계를 분석할 때 고려하지 않았으나 관계에 영향을 미치는 변수를 lurking(숨은, 잠재) 변수로 인한 것임, 흡연과 암의 관계에서 소득수준이 잠재변인일 수 있음, 소득수준 효과를 제외하면 흡연과 암의 (인과)관계가 존재하지 않을 수 있음

4. 추론과 데이터 분석

(1) 개념

(a) 연구문제를 통계적 모수와 모수에 대한 통계적 가설로 변환

(b) 모수 parameter : 연구문제의 관심 집단의 특성 값, unknown but fixed

(c) 모수가 확률변수(분포함수를 가짐)이면 추정 방법 : Bayesian 베이지안 추정

(d) 추정 estimation : 모수의 값을 하나의 값으로 추정하면 점추정 point estimator, 구간으로 추정하면 interval estimator

(e) 통계적 가설 : 관심 모수가 하나의 값으로 설정된 귀무가설(null hypothesis)과 그에 대립되는 대립가설(alternative)로 나뉨

(f) 통계량 statistic : 모수에 대한 추정과 가설검정을 위하여 확률표본 random sample(독립이고 동일한 분포 independently and identically distributed)로부터 계산된 값 – 통계량이 추정에 사용되면 추정치, 가설검정에 사용되면 검정통계량이라 함

(g) 샘플링분포 sampling distribution : 통계량의 확률분포함수, 구간추정과 가설검정을 위해서는 샘플링분포를 알아야 함 – 통계량의 샘플링분포를 알지 못하면 "비모수추론" (nonparametric 혹은 distribution free test) 실시함

연구문제	20대 청년의 흡연률? 미국 흡연율 2%보다 높은가?	직장인 남녀 한달 지출 용돈은 동일한가?
통계적 문제	흡연율 p 구간추정? 흡연율 p 〉 0.24	남자 용돈평균과 여자 용돈 평균의 같나?
데이터 (확률표본)	20대 청년 200명 무작위 추출하여 흡연여부 조사	직장인 남자 40명, 여자 50명 무작위 추출하여 한달 용돈 조사
추정치	$\hat{p}$ = 200명 중 흡연한 청년 비율	$\bar{x}_1 - \bar{x}_2$ 남자 표본평균 – 여자 표본평균
샘플링분포	$\sim N(p, \frac{pq}{\sqrt{n}})$	$\sim N(\mu_1 - \mu_2, \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}})$
귀무가설	$H_0 : p = 0.24$	$H_0 : \mu_1 = \mu_2$
대립가설	$H_0 : p > 0.24$	$H_0 : \mu_1 \neq \mu_2$
검정통계량	$\frac{\hat{p} - 0.24}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim N(0,1)$	$\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t(df)$

(2) 귀무가설 null hypothesis H_0

- (정의) 모수의 값이 하나의 값으로 설정하고 관심의 대상이 아닌 가설, 영 가설이라고 함
- 차이가 없다, 영향을 미치지 않는다, 모수의 값은 2이다 등으로 설정됨

(3) 대립가설 alternative hypothesis $H_1 or H_a$

- (정의) 귀무가설에 설정된 모수의 값이 이외 모든 값의 모임 가설로 관심 대상인 가설, 연구가설이라고 함
- 차이가 있다, 영향을 미친다, 모수의 값은 2가 아니다 등으로 설정됨
- 단측 가설 one sided : 대립가설 모수 공간이 귀무가설 값 중심으로 한 쪽만, 모수>2 혹은 모수<2
- 양측 가설 two sided : 귀무가설 설정 값 이외 모든 모수 공간, 모수$\neq 2$

(4) 통계적 검정 statistical test

- 무죄 추정 원칙 – 적절한 데이터를 수집하여 유죄(대립가설 사실)임을 밝히는 과정

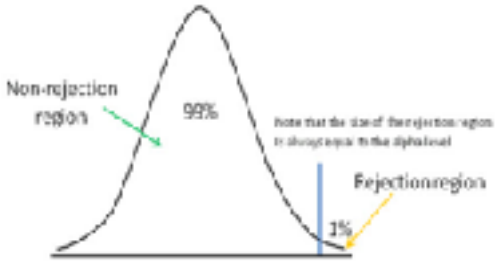
	귀무가설 진실	대립가설 진실
귀무가설 기각	1종 오류 type I error (α)	옳은 판단
귀무가설 채택	옳은 판단	2종 오류 type II error (β)

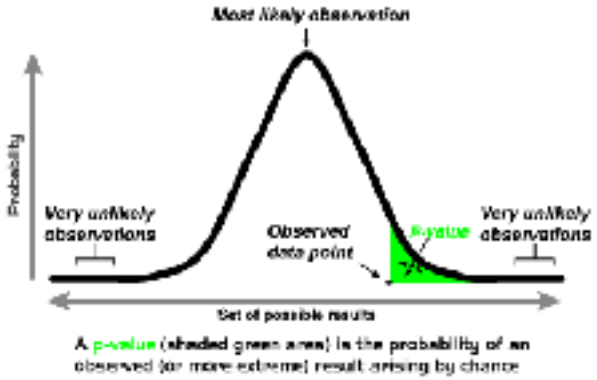
(5) 유의수준 significant level

- (정의) 설정된 (permitted) 1종 오류 (귀무가설이 사실인데 검정결과 귀무가설을 기각하는 확률)
- 1종 오류 값을 설정한 것은 우리의 관심이 대립가설에 있기 때문임
- 일반적으로 1%(highly significant), 5%(significant), 10%(little significant) 사용

(6) 기각역 critical region

- (정의) 귀무가설이 사실 (설정된 모수 값) 하에 얻어진 샘플링분포의 꼬리부분의 확률이 설정된 유의수준과 동일한 영역을 기각역이라 함
- 양측 대립가설이면 양측 꼬리 부분의 합이 유의수준 α와 동일하고, 단측이면 한 쪽 꼬리의 확률이 α인 영역임
- 표본 데이터로부터 계산된 검정통계량이 기각역에 속하면 귀무가설을 기각하게 된다.

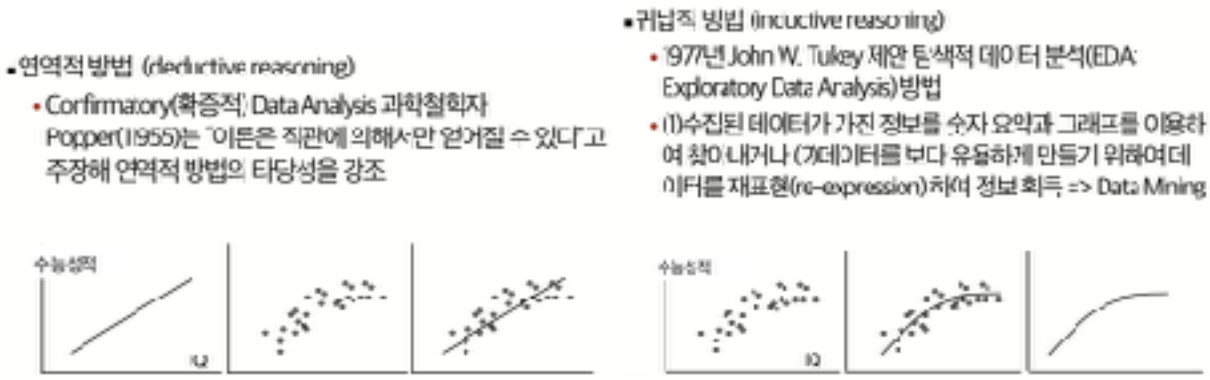




(7) 유의확률 p-value

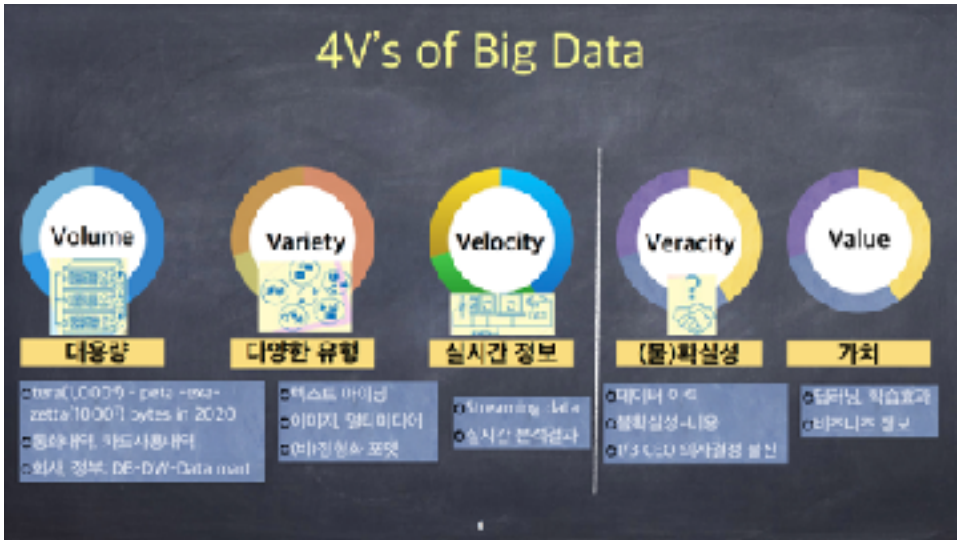
- (정의) 샘플링분포함수에서 계산된 검정통계량에 의해 구성된 꼬리 부분의 확률, 계산된 obtained 유의수준
- 유의확률이 유의수준보다 작으면 귀무가설이 기각된다 〈=〉 계산된 오류가 설정된 오류보다 작으므로
- 양측 대립가설의 경우에는 꼬리부분의 확률을 2배하여 유의확률로 활용

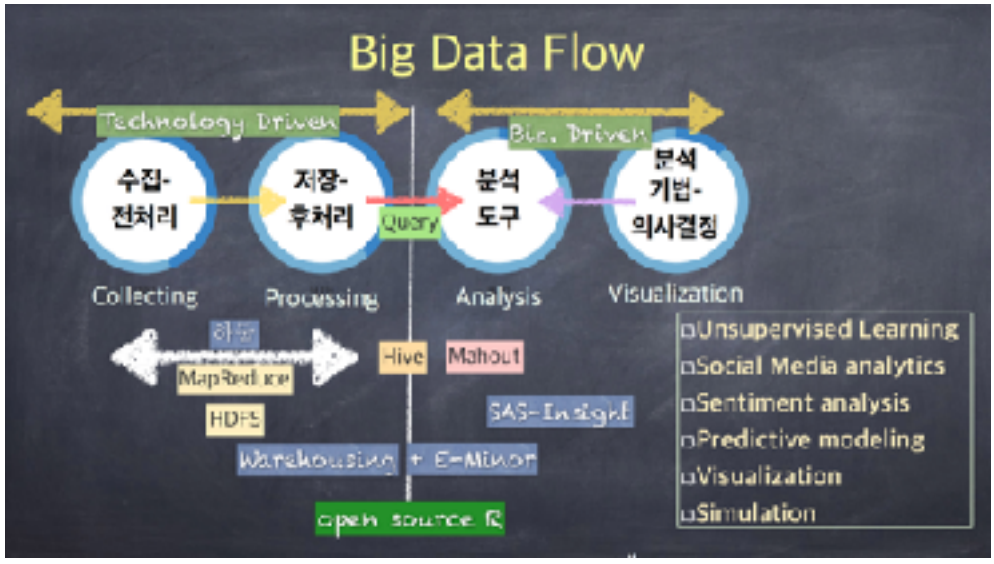
(8) 사회 과학 방법론

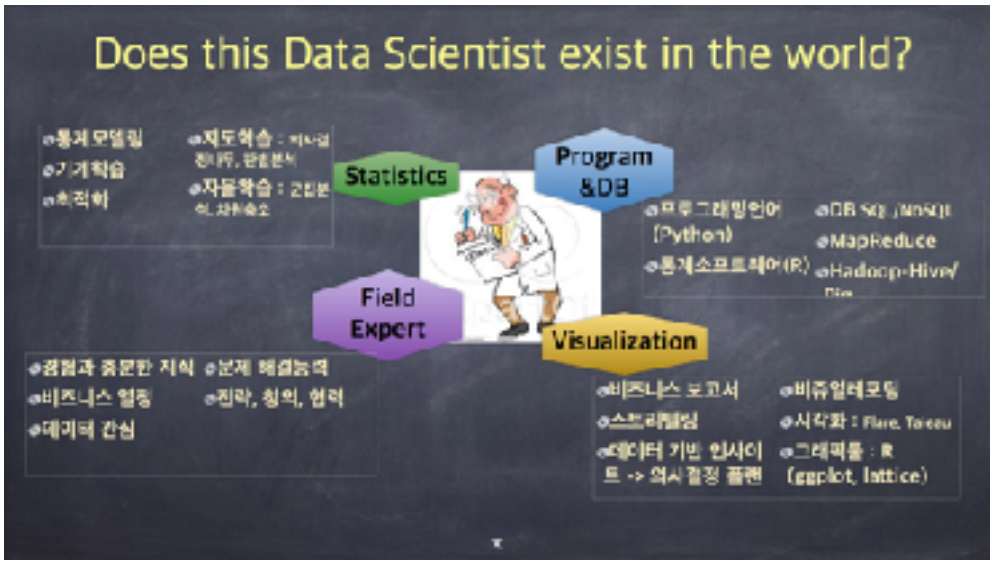


- 통계적 가설검정은 (1. 통계적 가설 설정 -〉 2. 데이터 수집 및 가설 검정 -〉 3. 가설 관련 결론) 연역적 방법, 확증적 방법론임
- 탐색적 자료분석은 데이터 요약과 적절한 그래픽 표현을 통하여 데이타가 말하는 정보를 얻는 방법론 – 얻은 정보에 대해서는 확증적 분석방법으로 확인하게 됨

- Modern EDA = Data Mining -> Big Data







5. 데이터와 통계적 방법

- 변수 종류

  (a) 일변량 vs. 다변량 (이변량 포함)
  (b) 측정형 vs. 범주형
  (c) 종속변수(Y) vs. 독립변수(X)

- 일변량 분석 univariate 분석

  (a) 분석 대상인 변수가 하나인 경우
  (b) 변수는 하나이지만 집단은 2개인 경우

	일집단		2집단	
변수 종류	그래프 요약	숫자 요약	그래프 요약	숫자 요약
측정형	히스토그램 나무상자 그림	평균, 표준편차 중앙값, IQR	나무상자 그림	평균 차이 분산 차이
범주형	파이 차트 바 차트	비율	집단별 바 차트	비율 차이

- 다변량 분석 multi-variate 분석 (인과분석)

  (a) 분석 대상인 변수가 2개 이상인 경우
  (b) 종속변수가 1개인 경우 - 종속변수가 2개 이상인 경우는 본 강의에서 제외
  *) 인과분석이 아닌 다변량분석 : 주성분분석, 요인분석, 군집분석, 판별분석 (종속변수 개념 없음)

종속변수(Y) 설명변수(X)	측정형	범주형
측정형	상관분석 correlation 회귀분석 regression	로지스틱 회귀분석 logistic, probit
범주형	분산분석 ANOVA	로그-선형 모형 log-linear

연습문제 1) 💾BANK2.CSV

여성 CEO들은 은행이 대출에 있어 남성 CEO에 비해 차별을 하고 있다고 주장한다. 남성 CEO 1,050명, 여성 CEO 115명(총 1,165명)을 조사하여 대출여부와 대출 받은 CEO의 이자율(rate)을 조사하였다.

은행은 이자율 결정은 기업의 설립연수, 기업형태(1=개인 2=파트너 3=기업), 기업의 매출액에 의해 결정된다고 주장하였다.

1) 성별에 따른 은행 승인율의 차이가 있나?
2) 성별에 따른 은행 이자율 차이는 있나?
3) 기업형태에 따른 은행 이자율의 차이는 있나?
4) 기업 매출액에 따른 은행 이자율 차이는 있나?
5) 성별에 따른 매출액의 차이가 있나?
6) 성별에 따른 기업형태의 차이가 있나?