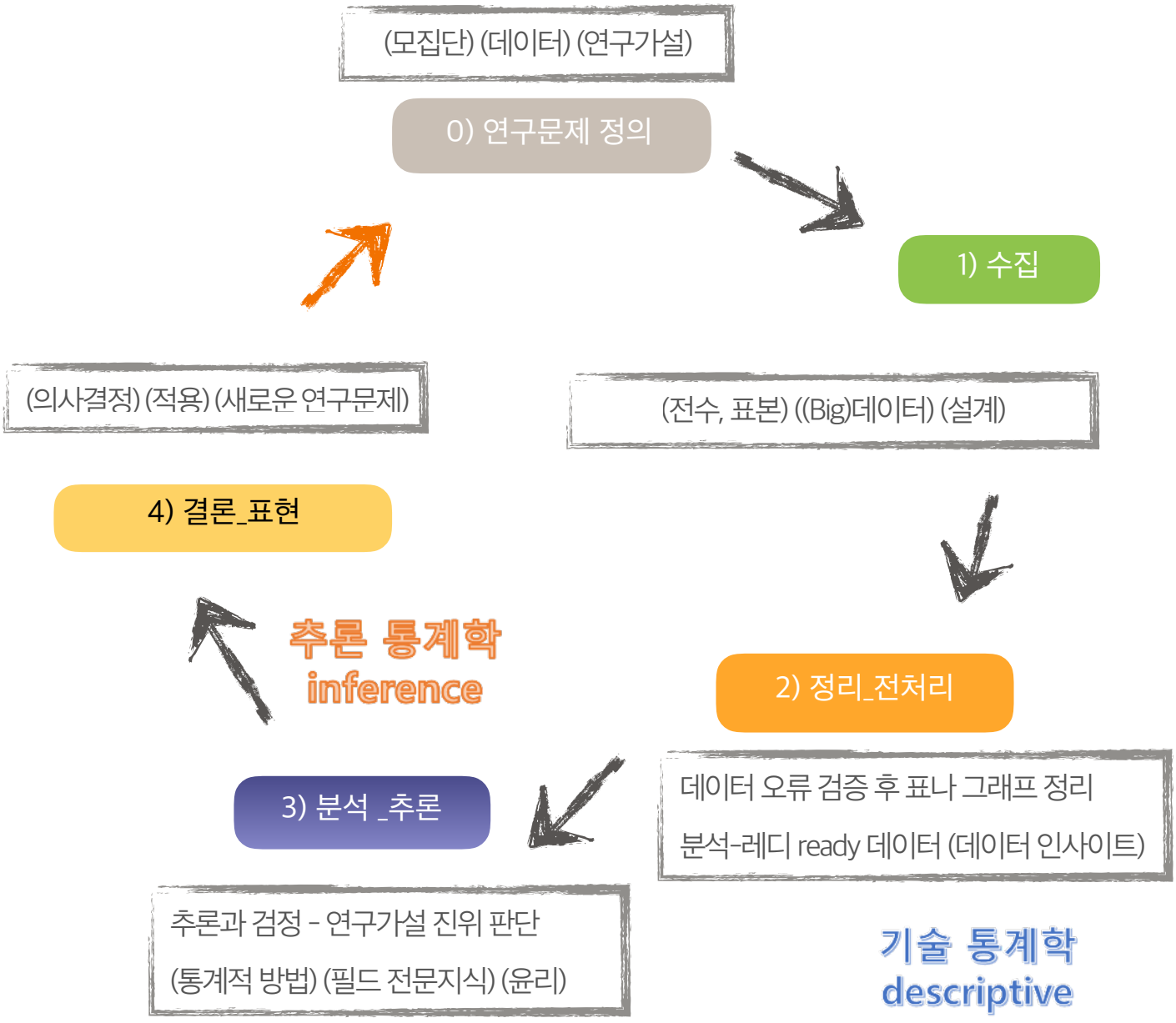
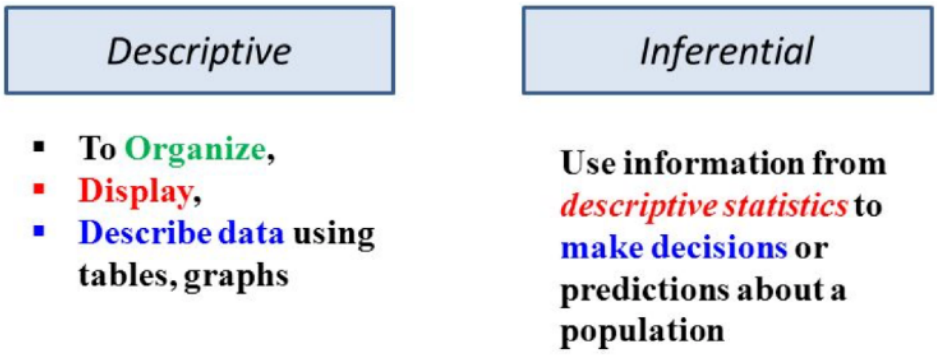
통계학 정의

- 통계학은 데이터에 관한 학문 Statistics is about Data. (ott)
- 통계학은 사회, 자연 현상을 데이터와 추론으로 설명하는 학문
- 통계는 아트이다. 통계적 방법론은 수학을 기초로 하지만 수학은 물감일 뿐 통계학은 새로운 세계를 만들어낸다.
- 수집 collect - 정리(전처리) summarize - 분석 analysis - 표현 conclusion & presentation 의 일련의 과정을 거친다. (Webster's Dictionary)



통계학 종류



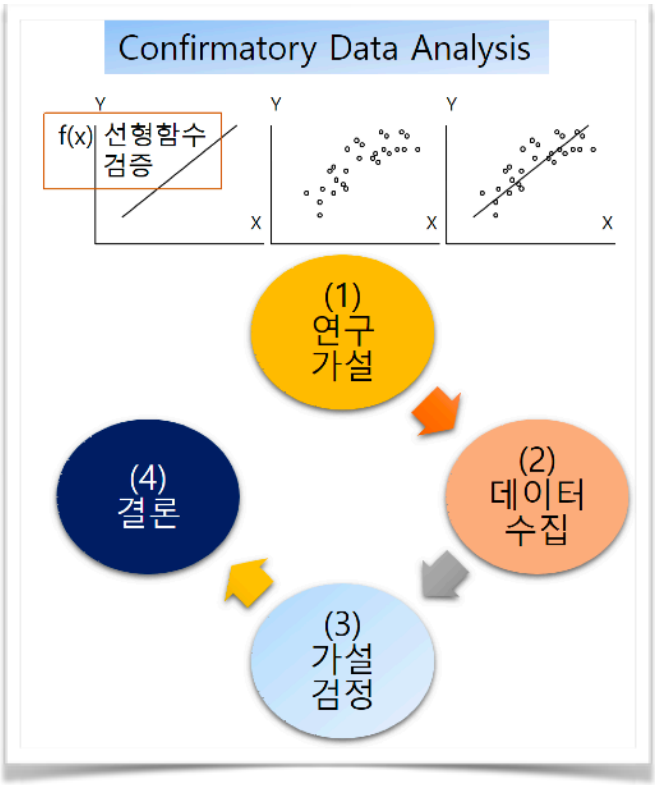
1) 기술 descriptive통계학 = 국가통계

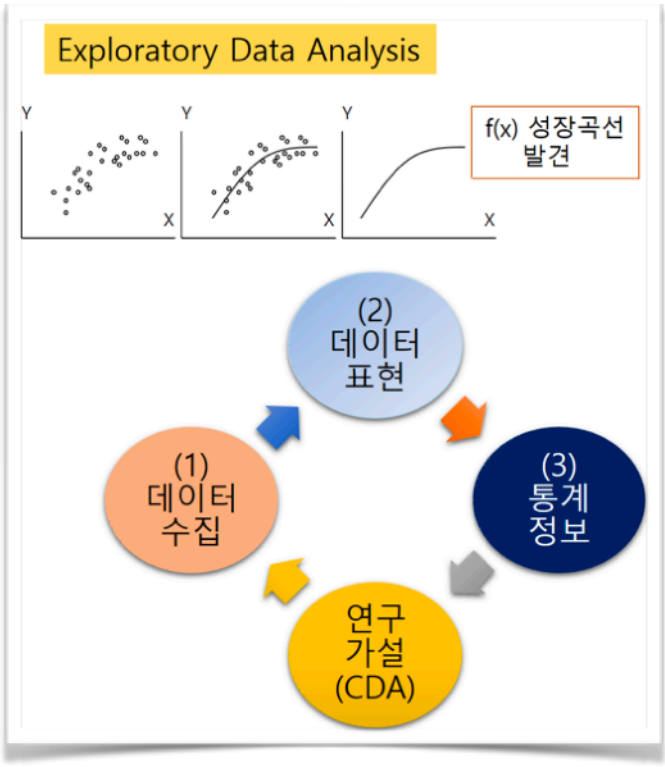
- 구약의 민수기(numbers)에 이스라엘 백성 인구 조사, 광야 생활 전과 후에 각 한 번씩 조사
- 로마 황제 Tullis: 세금 징수를 위하여 5년마다 인구 조사, Caesar 가 로 마 제국 전역으로 확대
- 센서스(census, 전수 조사): 어원 censura(라틴어 세금tax), statistics(통계학, 라틴어 status국가)
- 17C 영국 출생률과 사망률 조사: 나이팅게일도 통계학자 - polar diagram (원인별 비율)

2) 추론 inference 통계학

- 확률 , 게임이론, 수학을 기반으로 발전 : Fermat, Pascal(1754): 드멜라라는 친구의 요청으로2인 게임에서 경기 조합 방법을 계산하기 위하여 파스칼의 삼각형을 제안하였다.
- 확률 이론 발전 (J. Bernoulli, A. de Moivre, Komogorov), 사회현상에 대한 고찰이나 과학적 사고로 얻어진 논리에 대한 연구와 수학의 접목을 통하여 통계학은 더욱 발전하게 되었다.
- 정규분포 Normal Distribution : 이항실험(동전 던져 앞뒤 나오는 결과 실험)을 무한히 하는 경우 앞면의 회수에 대한 분포 고민(중심극한 정 리, DeMoivere, 1667-1754) - Laplace(1949-1827)는 다른 분포의 중 심극한 정리
- Gauss, (1777~1855) 정규분포 분포, 행성간 거리 오차에 대한 히스토그 램으로부터 식을 유도 - 측정오차의 대부분은 이 분포를 따르고 있어 normal 정규분포라 함. Gaussian 분포
- W.S. Gosset (1908): 독일 양조장 공장장, t-분포, 소표본 평균의 분포 가 정규분포에 따르지 않음으로 인하여 발견
- F. Galton(1885): 회귀분석(유전학자, 완두콩과 부모자녀 키의 관계), Karl Pearson 수리적 접근 => 인과관계
- Fisher: 농업 통계 분야 분산분석 방법론 적용

통계적 방법론 : 확증적 연구 vs. 탐색적 연구





▪연역적 방법 (deductive reasoning)

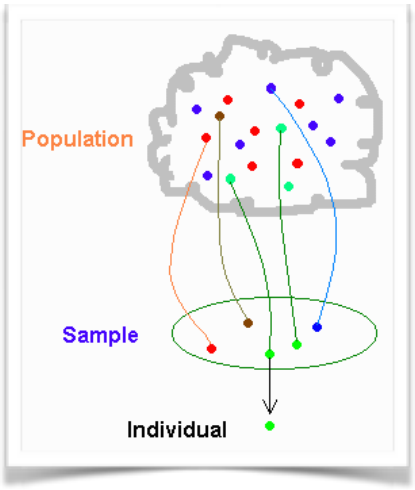
- Confirmatory(확증적) Data Analysis 과학철학자 Popper(11955)는 "이론은 직관에 의해서만 얻어질 수 있다"고 주장해 연역적 방법의 타당성을 강조하였다.
- 연구가설 설정 -> 데이터 수집 -> 가설 검증 및 결론

▪귀납적 방법 (inductive reasoning)

- 1977년 John W. Tukey 제안 탐색적 데이터 분석(EDA: Exploratory Data Analysis) 방법
- (1)수집된 데이터가 가진 정보를 숫자 요약과 그래프를 이용하여 찾아내거나 (2)데이터를 보다 유용하게 만들기 위하여 데이터를 재표현(re-expression) 하여 정보 획득
- => Data Mining => Big Data

통계분석 절차

1) 연구문제 정의 (분석목적)



- 가능하면 통계적 용어를 사용하지 않으며, 분석 대상 분야의 "언어"(개념)으로 표현한다.
- 모집단과 표본, 조사방법을 정리한다.

모집단 population, 표본 sample

- (정의) 모집단 : 분석 관심의 대상(subject, individual)이 되는 모든 개체(사람, 기업, 국가)의 모임 모집단 표현 : X_i (i번째 모집단 개체의 관심 측정값)
- (정의) 표본 sample : 모집단 일부 개체 , (표현) x_i
- 확률표본 random sample : 확률적으로 (모집단 개체가 표본으로 추출될 가능성이 동일 equally likely) 추출된 표본

1) 평생 담배 5갑(100개비) 이상 피웠고 현재 담배를 피우는 분율, 만19세이상 (1998년: 만20세이상) 국건영 2016년 발표에 성인 남자의 흡연율 40.7%, 여자는 6.4%라고 한다. 한남대학교 학생들의 흡연율을 알아보고 기독교 대학, 3C 운동의 효과가 있는지 알아하고자 한다. - **비율**

	1998	2001	2005	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
전체	35.1	30.2	28.8	25.3	27.8	27.3	27.5	27.1	25.8	24.1	24.2	22.6	23.9
남자	66.3	60.9	51.7	45.1	47.8	47.0	48.3	47.3	43.7	42.2	43.2	39.4	40.7
여자	6.5	5.2	5.7	5.3	7.4	7.1	6.3	6.8	7.9	6.2	5.7	5.5	6.4

(중알일보 기사 2017) 한남대학교 학생들은 과제 발표 준비 시간을 알아보고 전국 평균과 비교하여 학업에 대한 열정을 비교해 보고자 한다. - **평균**

2018년 2학기 등록한 한남대학교 학부 학생 (모집단 N=12,000명) 중 200명(표본)을 조사대상으로 하여 흡연여부를 조사한다.

1) 조사대상 선택 (표본추출) : (대학, 학년, 성별)을 층화하여

층 크기 비례배분하여 표본추출한다.(**확률적 표본추출 방법**이지만 비용과 시간적으로 비효율적임) 실제로는 2018년 9월 17일 오전 9시 정문으로 등교하는 학생 1분 단위로 조사대상을 정한다.

2) 조사방법 : 학과 사무실에 의뢰하여 (학년*성별) 배분 학생 수만큼 무작위 조사하게 한다.

3) 조사항목 (데이터) : 학과, 학년, 성별, 흡연여부, 일주일 과제/발표 준비시간

데이터 data 관측값 observation

- 표본 개체의 관심 특성(변수 variable)을 측정하거나 관측한 숫자, 문자 값
- 관측크기 n인 관측치 표현 : $(x_1, x_2, \ldots, x_n)$
- (예) $x_1 = 2.5, x_2 = 3.2, \ldots$ (예) $x_1 = Y, x_2 = N, \ldots$

2) 통계적 가설

연구문제는 관심 모수로 표현되는 통계적 가설, 귀무가설과 대립가설(연구가설)로 나타낸 후 데이터(확률표본)로부터 계산된 통계량을 이용하여 모수를 추정하고 귀무가설의 기각, 채택으로 연구문제에 대한 가설을 검증한다.

귀무가설	한남대학교 학생 흡연율은 전국 흡연율과 같다.
	$p = 0.239$
대립가설	한남대학교 학생 흡연율은전국 흡연율보다 낮다.
	$p < 0.239$

귀무가설	한남대학교 경상대학 과제/발표 준비시간 전국 평균과 같다.
	$\mu = 11.7$
대립가설	한남대 경상대학 학생들은 과제/발표 준비 시간은 전국 평균과 다르다.
	$\mu \neq 11.7$

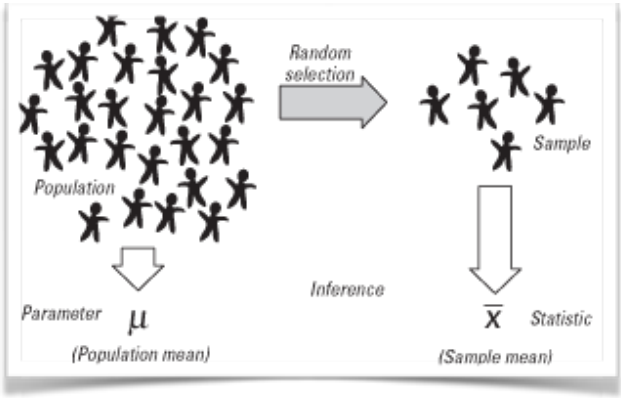
모수와 통계량

- (정의) 모수 parameter : 연구자가 관심을 갖는 모집단 개체 특성, θ (예 : H대학생의 흡연율=비율 p, 일주일 공부시간=평균 μ)

✳ (정의) 통계량 statistic : 확률표본 데이터로부터 계산된 값, $\hat{\theta}$ (표본평균, 표본비율, 표분분산, 최대값, 중위값) (예 : H대학생 200명 조사데이터 : 최대 공부시간 x_{max}, 평균 공부시간 $\bar{x}$)

3) 추론 inference

✳ 통계량이 추정에 사용되면 추정량 estimator, 추정값 ($\hat{\theta}$)estimate 이라 하고 통계적 가설 검정 사용되면 이를 검정통계량($t(\hat{\theta})$)이라 한다 (예 : 표본 200명이 일주일 공부한 시간을 조사하여 평균, 분산(공부 시간 산포)을 계산하면 이를 통계량이라 함



✳ (정의) 추론 inference : 통계량

✳ 추론을 통계적 분석이라고 한다.

4) 결론

통계추론의 결과는 일반적으로 모수의 95% 신뢰구간을 제공하거나, 유의수준 5%에서의 귀무가설 기각(혹은 채택)으로 표현된다.

95% 신뢰구간 <=> 유의수준 5% 가설 검정과 동일하다.

데이터를 이용하여 귀무가설을 기각하면 (귀무가설에 설정된 모수 값을 95% 신뢰구간이 포함하지 못하면) 연구가설이 채택된다.

연구내용이 채택되면 95% 신뢰구간 혹은 적절한 통계량 표를 제시하고 연구문제 결론을 리포팅한다.