0. 통계학(Statistics)이란?

통계학 정의

- 통계학(統計學, statistics)은 산술적 방법을 기초로 하여, 주로 다량의 데이터를 관찰하고 정리 및 분석하는 방법을 연구하는 수학의 한 분야

  (독일에 있어서의 넓은 뜻의 국가학(國家學)의 한 부문. 일반 국가학과는 달리 낱낱의 국가를 대상으로 하여 그 비교적·종합적 연구를 하는 기술학(記述學)으로서, 정책학(政策學)의 한 분야임)

통계학 역사

- 19세기 중반 벨기에의 케틀레가
  - 독일의 "국상학(國狀學, Staatenkunde, 넓은 의미의 국가학)"과
  - 영국의 "정치 산술(Political Arithmetic, 정치 사회에 대한 수량적 연구 방법)"을
  - 자연과학의 "확률 이론"과 결합하여, 수립한 학문에서 발전

현대의 통계학

- 현대에 들어와 데이터 과학자들로 구성된 통계 조직은 기관과 단체 그리고 기업의 수익에 영향을 미치는 다양한 데이터를 입체적으로 분석하고 미래를 예측해 의사결정(decision making)에 반영
  - 예) 전사자원관리(전사적자원관리, ERP) · 고객관계관리(CRM) · 생산관리시스템(MES) · 경영 정보 시스템(MIS) · 전략적 기업 경영(SEM) 등 각종 시스템 (ref. Wiki)

통계학의 특징

- 불확실성을 계량적으로 측정해서 정확하게 만듬
- 불확실성의 정도를 확신하고 범주형으로 의사결정

통계라는 수학적 기법을 위한 세가지 근거

1. 데이터 기술:

- 데이터의 수집, 전시, 요약

2. 확률론:

- 어떤 사건이 실제로 일어날 것인지 혹은 일어났는지에 대한 지식 혹은 믿음을 표현하는 방법이며 같은 원인에서 특정한 결과가 나타나는 비율을 뜻

3. 통계적 추론:

- 확률 지식을 이용해 특정 데이터에서 통계적 결론을 이끌어내는 과학

---

1. 데이터의 기술(데이터의 집합)

- 데이터를 알기 쉽게 표현하는 방법
- 숫자 속의 숨어 있는 일정한 유형
- 데이터의 본질적 형태

1) 데이터 시각화

- 도수분포표: 구간으로 나누고, 해당구간에 포함되는 수를 세는것
  - 히스토그램
- 상대도수: 각 구간에 속하는 수를 전체의 수로 나는 것
  - 상대도수 히스토그램
- 산포도 측정
  - 줄기-잎 그림
  - 상자그림(box plot)

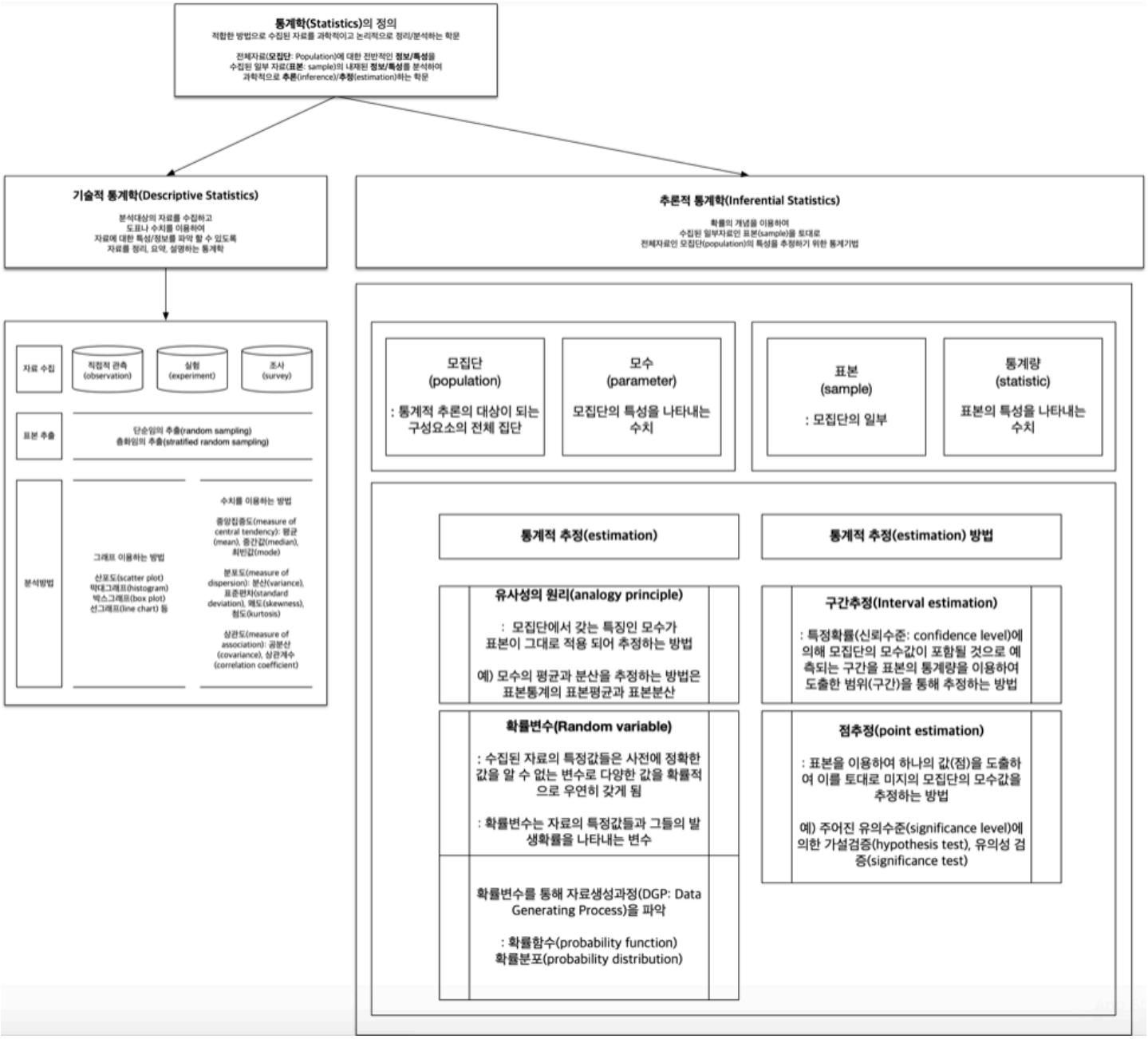
2) 기술통계

- 데이터 집합의 일반적 특성을 간단히 나타내는 방법

- 대표값: 중심으로 부터 흩어져 있는 정도(산포도)
  - 평균값(mean): 모든 데이터의 값을 더한 다음 데이터의 개수로 나누어 구함 $\sum_{i=1}^{n} x_i$
  - 중앙값(media): 순서대로 정리된 수의 가운데 값

- 산포도의 측정
  - 데이터가 대표값에서 얼마나 멀리 떨어져 있는가
  - 사분위 범위(IQR): 중앙값을 근거로 데이터를 4개의 동일 그룹으로 나눈 다음 양끝의 그룹이 얼마나 많이 떨어져 있는지 알아보는 것: $IQR = Q_3 - Q_1$

- 표준편차(standard deviation)
  - 산포도를 측정하는 표준 방법으로 평균으로 부터 측정
  - 평균($\bar{x}$) 에서 떨어져 있는 평균거리
  - 평균제곱거리 $= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}))$
  - 표본분산$(S^2) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2)$
  - 표준편차$(S) = \sqrt{S^2} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2)}$
  - Z-점수(z-score) : 평균에서 부터 표준편차 거리로 정의, $z_i = \frac{x_i - \bar{x}}{s}$
  - 경험법칙 : 64%(평균에서 표준편차의 한 배 이내), 97%(평균에서 표준편차의 두배 이내)


통계학(Statistics)의 정의
적합한 방법으로 수집된 자료를 과학적이고 논리적으로 정리/분석하는 학문
전체자료(모집단: Population)에 대한 전반적인 정보/특성을
수집된 일부 자료(표본: sample)의 내재된 정보/특성을 분석하여
과학적으로 추론(inference)/추정(estimation)하는 학문
기술적 통계학(Descriptive Statistics)
분석대상의 자료를 수집하고
도표나 수치를 이용하여
자료에 대한 특성/정보를 파악 할 수 있도록
자료를 정리, 요약, 설명하는 통계학
추론적 통계학(Inferential Statistics)
확률의 개념을 이용하여
수집된 일부자료인 표본(sample)을 토대로
전체자료인 모집단(population)의 특성을 추정하기 위한 통계기법
자료 수집
직접적 관측 (observation)
실험 (experiment)
조사 (survey)
표본 추출
단순임의 추출(random sampling)
층화임의 추출(stratified random sampling)
분석방법
그래프 이용하는 방법
산포도(scatter plot)
막대그래프(histogram)
박스그래프(box plot)
선그래프(line chart) 등
수치를 이용하는 방법
중앙집중도(measure of central tendency): 평균(mean), 중간값(median), 최빈값(mode)
분포도(measure of dispersion): 분산(variance), 표준편차(standard deviation), 왜도(skewness), 첨도(kurtosis)
상관도(measure of association): 공분산(covariance), 상관계수(correlation coefficient)
모집단 (population)
: 통계적 추론의 대상이 되는 구성요소의 전체 집단
모수 (parameter)
모집단의 특성을 나타내는 수치
표본 (sample)
: 모집단의 일부
통계량 (statistic)
표본의 특성을 나타내는 수치
통계적 추정(estimation)
통계적 추정(estimation) 방법
유사성의 원리(analogy principle)
: 모집단에서 갖는 특징인 모수가 표본이 그대로 적용 되어 추정하는 방법
예) 모수의 평균과 분산을 추정하는 방법은 표본통계의 표본평균과 표본분산
구간추정(Interval estimation)
: 특정확률(신뢰수준: confidence level)에 의해 모집단의 모수값이 포함될 것으로 예측되는 구간을 표본의 통계량을 이용하여 도출한 범위(구간)을 통해 추정하는 방법
확률변수(Random variable)
: 수집된 자료의 특정값들은 사전에 정확한 값을 알 수 없는 변수로 다양한 값을 확률적으로 우연히 갖게 됨
: 확률변수는 자료의 특정값들과 그들의 발생확률을 나타내는 변수
점추정(point estimation)
: 표본을 이용하여 하나의 값(점)을 도출하여 이를 토대로 미지의 모집단의 모수값을 추정하는 방법
예) 주어진 유의수준(significance level)에 의한 가설검증(hypothesis test), 유의성 검증(significance test)
확률변수를 통해 자료생성과정(DGP: Data Generating Process)을 파악
: 확률함수(probability function)
확률분포(probability distribution)

2. 확률론

확률론(영어: probability theory):

- 수학의 한 분야로 확률에 대해 연구하는 분야

확률론의 목적:

- 비결정론적 현상을 수학적으로 기술하는 것

확률론의 주요 연구 대상

- 확률 변수
- 확률 사건
- 확률 과정
- 기타 등등

확률론의 적용

- 변화하는 환경에 대처하여 결정의 내릴 때 의식적 혹은 무의식적 확률론에 기반
- 완전한 정보가 알려지지 않은 복잡계를 기술하는 데에도 확률론적 방법론
- 예) 당첨확률, 강수확률, 발병확률, 충동확률, 우승확률, 사고확률 등

--> 확률론은 통계학의 수학적 기초

확률이론의 역사

- 블레즈 파스칼(1623 ~ 1662), 17세기 프랑스의 수학자, 물리학자, 철학자 (도박의 문제)
- 피에르 페르마(1601 ~ 1665), 프랑스 수학자, 정수이론, 확률론, 좌표기하학에 기여

확률론의 다른 접근 방법

- 발생빈도주의적 확률론: 고전적 확률이론으로 모든 가능성을 지닌 경우의 수에 대한 원하는 경우의 수의 비
- 베이즈 확률론: 믿음의 정도(degree of belief)를 나타내는 것, 어떤 주어진 증거 혹은 어떤 상태에서 개인의 주관적 믿음의 정도
  - 주관적 믿음: 사전분포를 기본 바탕
  - 주관적 믿음의 변화: 사전분포에서 경험(데이터)를 통해 사후분포를 얻음

확률

- 하나의 사건이 일어날 수 있는 가능성을 수로 나타낸 것
- 어떤 사건이 실제로 일어날 것인지 혹은 일어났는지에 대한 지식 혹은 믿음을 표현
- 어떤 잠재적 사건이 일어날 경우의 가능성
- 같은 원인에서 특정 결과가 나타나는 비율

- 수학적 확률(mathmatical probability; 선험적 확률)
- 통계적 확률(emprical probability; 경험적 확률)

- 높은 확률: 가능
- 낮은 확률: 불가능

확률의 법칙

- 불확실한 상황에서 가능성의 법칙
- 확률실험(시행) : 우연이 지배하는 사건의 결과를 관찰하는 과정
- 근원사건: 어떤 시행에서 일어날 수 있는 모든 결과
- 표본공간: 모든 근원사건의 집합
  - 예) 동전 : 근원사건(앞면, 뒷면), 표본공간: {앞면, 뒷면}
  - 예) 1개의 주사위: 근원사건(1~6), 표본공간: {1, 2, 3, 4, 5, 6} --> 1/6
  - 예) 2개의 주사위: 근원사건(6*6개), 표본공간 {(1,1), (1,2), (1,3), ...(1,6), (2,1)(2,2), (2,3),...(2,6), ... , (6,4), (6,5), (6,6)} --> 1/36
  - 3개의 주사위, 표본공간(216= 6 x 6 x 6) --> 1/216

- **확률의 특성**
  - 확률은 음수가 아님, 양수
  - 모든 근원사건의 확률의 합은 1

확률 연산

- 사건은 근원사건의 집합

확률의 표현

- 사건의 확률은 그 집합에 속하는 근원사건들의 확률의 합
- 예) 2개의 주사위에서 나온합이 3인 경우
  - 사건에 속하는 근원사건 {(1,2), (2,1)}
  - 확률 P(A) = 1/6 * 1/6 + 1/6 * 1/6 = 1/36 + 1/36
- 사건 E와 F 둘 다 일어난다 : E and F
  - P(E and F)
- 사건 E또는 F가 일어난다(또는 둘 다 일어난다): E or F
  - P(E or F) = P(E) + P(F) - P(E and F)
- 사건 E는 일어나지 않느다: not E
  - P(not E) = 1 - P(E)

조건부 확률

- $P(A|C)$: C가 주어졌을 때, A의 확률
- $P(A|C) = \frac{P(AandC)}{P(C)}$
- $P(A|A)$ = 1
- 예) 주사위: 첫번째 3이 나오고 두번째 4가 나오는 경우: P(4|3)
- $P(4|3) = \frac{P(4and3)}{P(3)} = \frac{\frac{1}{36}}{\frac{1}{6}} = \frac{1}{6}$

곱셈정리

- $P(AandC) = P(A|C)P(C)$
- 독립사건의 곱셈정리 (주사위의 경우)
  - $P(AandC) = P(A)P(C)$

덧셈정리

- $P(AorC) = P(A) + P(C) - P(EandF)$
- A와 C가 배반일 때: $P(AorC) = P(A) + P(C)$

뺄셈정리

- $P(E) = 1 - P(notE)$

베이즈 정리

- $P(A|B) = \frac{P(A)P(B|A)}{P(A)P(B|A)+P(notA)P(B|notA)}$
- $\frac{P(AandB)}{P(AandB)+P(notAandB)} = \frac{P(AandB)}{P(B)} = P(A|B)$

베이즈 정리(예제)

- P(A) = .001 (1000개의 동일 장비중 한개는 바이러스가 있다.)
- P(B|A) = .99 (감염된 경우, 바이러스 테스트에 양성 반응이 나타날 확률은 0.99)
- P(B | not A) = .02 (바이러스가 없는 장비에서 바이러스 테스트 한경우 양성반응이 잘못 나타나는 경우 0.02)
- P(A|B) = ? , (양성반응이 나타난 경우 실제 바이러스에 감염되었을 확률)

	A	not A
B	A and B	not A and B
not B	A and not B	not A and not B

	A	not A	합계
B	P(A and B)	P(not A and B)	P(B)
not B	P(A and not B)	P(not A and not B)	P(not B)
	P(A)	P(not A)	1

- - P(A and B) = P(B|A)P(A) = (0.99)(0.001) = .00099
- - P(not A and B) = P(B|not A)P(not A) = (0.02)(0.999) = 0.1998

	A	not A	합계
B	0.00099	0.01998	0.02097
not B	0.00001	0.97902	0.97903
	0.001	0.999	1

- - P(A|B) : (양성반응이 나타난 경우 실제 바이러스에 감염되었을 확률)

$$P(A|B) = \frac{P(A and B)}{P(B)} = \frac{0.00099}{0.02097} = 0.0472$$

	바이러스 감염 장비	정상장비	합계
바이러스 검사장비 테스트에서 양성반응	1	20	21
바이러스 검사장비 테스트에서 음성반응	0	979	979
합계	1	999	1000

확률공간(probabilty sapce)

- 확률공간의 구성의 3요소: $(\Omega, \mathcal{F}, \mathbf{P})$
- Ω : 표본공간(sample space), 일어날 수 있는 모든 가능한 결과들(outcomes), 그 원소 ω는 결과(outcome)
- $\mathcal{F}$: 표본공간의 부분집합들의 모임(collection) 시그마집합체(σ-field), 시그마집합체의 원소는 표본공간의 부분집합으로서 사건(event)
- $\mathbf{P}$: 확률 P, 시그마집합체에서 정의되는 집합함수 P가 확률의 세조건(모든 사건에 대해 0보다 크고, 모든 합은 1이며, 배반사건(disjointed event)들인 경우 전체확률의 합은 모든 이벤트의 확률)을 만족하면 P를 확률이라고 함

확률변수(random variable)

- 표본공간을 정의역(domain)으로 하고 실수공간을 공역으로 하는 잴 수 있는 함수(measurable function) X를 확률변수라고 함
  - (실수공간 위의 보통위상(usual topology)에 의해 유도된 위상공간에서 정의되는 시그마집합체인 보렐-시그마집합체 중 하나의 보렐-시그마집합체)
- 확률변수 X는 각각의 결과(outcome) ω에 실수를 대응시키느 함수

In [ ]:

```
1
```