1. 통계학이란?

- 통계학 정의

  (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 "정치산술(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

  전사자원관리(전사적자원관리, ERP)・고객관계관리(CRM)・생산관리시스템(MES)・경영 정보 시스템(MIS)・전략적 기업 경영(SEM) 등 각종 시스템

- 통계학의 특징: 불확실성을 계량적으로 측정해서 정확하게 만드는 특징
  - 불확실성의 정도를 확신하고 범주형으로 의사결정

- 통계라는 수학적 기법을 위한 세가지 근거
  - 데이터 분석 : 데이터의 수집, 전시, 요약
  - 확률
  - 통계적 추론 : 확률 지식을 이용해 특정 데이터에서 통계적 결론을 이끌어내는 과학

2. 데이터의 기술(데이터 집합)

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

데이터 시각화

- 도수분포표: 구간으로 나누고, 해당구간에 포함되는 수를 세는것
    - 히스토그램
- 상대도수: 각 구간에 속하는 수를 전체의 수로 나는 것
    - 상대도수 히스토그램
- 줄기-잎 그림

요약통계량

- 데이터 집합의 일반적 특성을 간단히 나타내는 방법
- 대표값: 중심으로 부터 흩어져 있는 정도(산포도)
    - 평균값: 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: $IQR = Q_3 - Q_1$
    - 상자그림(box plot)
- 표준편차
    - 산포도를 측정하는 표준 방법으로 평균으로 부터 측정
    - 평균($\bar{x}$) 에서 떨어져 있는 평균거리
    - $\text{평균제곱거리} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}))$
    - $\text{표본분산}\,(S^2) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2)$
    - $\text{표준편차}\,(S) = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2)}$
    - 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)

3. 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

- 확률의 의미
    - 고전적 확률 개념: 도박에 바탕, 모든 근원사건은 동일한 확률을 가진다고 가정 (객관주의자)
    - 통계적 확률 개념: 반복 가능한 시행에서, 한 사건이 일어날 확률은 오랫동안 관찰할 때 그 사건이 일어날 횟수의 비율 (객관주의자)
    - 주관적 확률 개념: 어떤 사건이 일어날 가능성을 개인이 평가한 것 : (주관주의자, 베이즈 주의자)

- 연산
    - 사건은 근원사건의 집합
    - 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(A and C) = P(A|C)P(C)
    - 독립사건의 곱셈정리 (주사위의 경우)
        - P(A and C) = P(A)P(C)
- 덧셈정리
    - P(A or C) = P(A) + P(C) - P(E and F)
    - A와 C가 배반일 때: P(A or C) = P(A) + P(C)

- 뺄셈정리
  - P(E) = 1 - P(not E)

- 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)

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

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

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

4. 확률변수

- 확률 변수의 개념
  - 확률변수는 대문자로 작성 예) X
  - 확률변수는 시행의 수치 결과로 정의
    - 예) 어느 섹터에서 임의로 한 직원을 추출한다고 하고, 그 직원의 키, 몸무게, 가족수입, 직무평가 점수는 그 직원의 특징을 나타내는 수치로 된 변수들 모두 확률변수
  - X의 한 값은 소문자 x로 작성
  - 확률변수 X가 x값을 가질 확률 Pr(X = x)를 간단히 p(x)로 씀
  - 예) 두 주사위의 합의 경우

y	2	3	…	7	…	11	12
Pr(Y = y)	$\frac{1}{36}$	$\frac{2}{36}$	…	$\frac{6}{36}$	…	$\frac{2}{36}$	$\frac{1}{36}$

- 확률모델로 현상을 기술

상대도수 히스토그램 --> 무어떤 시행을 무수히 반복하면 --> 확률변수의 확률 히스토그램

- 확률변수의 평균과 분산
  - 데이터의 속성은 표본의 속성
    - 데이터의 경우 : 평균($\bar{x}$), 표준편차(s)
    - 표본평균 = $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

        - $\bar{x} = \frac{1}{n}\sum_{allx} n_x x$
        - $\bar{x} = \sum_{allx} x\frac{n_x}{n}$
    - 확률분포의 속성은 모델 또는 모집단 속성으로 불림
        - 모집단의 평균(μ), 모집단의 표준편차(σ)
        - 확률변수 X의 평균 정의
            - $\mu = \sum_{allx} xp(x)$
            - X의 기대값 $E[X]$
            - 확률이 가중된 가능한 모든 값의 합
    - 분산:
        - 데이터가 평균에서 떨어져 있는 편차의 제곱을 평균
        - $s^2 = \frac{1}{n-1} sum_{i=1}^{n}(x_i - \bar{x})^2$
        - $s^2 = \sum_{allx}(x_i - \bar{x})^2 \frac{n_x}{n-1}$
    - 확률변수 X의 분산
        - 모평균에서 얻은 편차를 제곱한 기대값
        - $\sigma^2 = \sum_{allx}(x_i - \mu)^2 p(x)$
        - 분산(σ^2)의 제곱근이 표준편차(σ)
- 여기까지 이산확률변수
    - 결과들이 모두 서로 떨어진 이산 값의 집합
    - 동전던지기, 주사위 예시

- 연속확률변수
    - 확률변수 X가 0과 1사의의 무한개의 값을 갖는 경우
    - 연속확률변수 X의 확률밀도 f(x) : Pr(a <= X <= b)
    - 면적을 계산하는데 a에서 b까지 f의 적분 : $\int_a^b f(x)dx$
- 연속확률변수 두가지 성질
    - $f(x) \geq 0$
    - $\int_{-\infty}^{\infty} f(x)dx$
- 연속확률변수의 평균과 분산
    - $\mu = \int_{-\infty}^{\infty} xf(x)dx$
    - $\sigma^2 = \int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx$

- 확률변수의 덧셈
    - 예) 동전던지기

x	**0**	**1**
p(x)	0.5	0.5

- 평균
  - $E[X] = 0 * p(0) + 1 * p(1) = 0 + 0.5 = 0.5$
- 분산
  - $\sigma = (0 - E[X])^2 p(0) + (1 - E[X])^2 p(1) = (0 - 0.5)^2 p(0) + (1 - 0.5)^2 p(1) = 0.25$
- 예) 상황 : 도박게임에서 처음 6달러를 내고, 동전을 던저 앞면이 나오면 10달러를 가져가고, 뒷면이 나오면 6달러를 잃는 것, 수익은 W
  - W = 10X - 6
  - $E[W] = E[10X - 6] = 10E[X] - 6$

x	**0**	**1**
w	-6	4
p(w)	0.5	0.5

    10(0.5) - 6 = -1
  - E[aX + b] = aE[X] + b
  - $\sigma^2(aX + b) = a^2\sigma^2(X)$
  - 도박게임의 가능한 결과 -6, 4, W의 분산은 X의 분산보다 더 크게 되므로
  - $\sigma^2(W) = \sigma^2(10X + 6) = 100\sigma^2(X) = 25$
  - $\sigma(W) = 5$
- 확률변수의 계산에서의 일반화
  - $E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} E[X_i]$
  - X_i가 모두 독립일 때
    - $\sigma^2(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \sigma^2(X_i)$