

## 1. sampling methods

모집단을 모두 조사하는게 불가능하기 때문에, **sample**(표본)을 모은다

- 1) 무작위 추출
- 2) 계통 추출(k번씩 건너뛰기)
- 3) 층화추출법(계층을 나누어서 계층별로 추출)
- 4) 집락추출법(집단을 나누어서 집단마다 추출)  
(층화와 집락은 집단의 수직/수평적 구조에 차이가 있습니다)

## 2. 변수의 종류

- 1) 범주형 변수 . ex) {강아지, 고양이}
- 2) 연속적 변수. 명확히 종류가 구분되지 않고 셀 수 없다
- 3) 불연속적 변수. 명확히 종류가 구분되지 않고 셀 수 있다

이렇게 변수의 종류마다 데이터를 다루는 방법과 시각화 하는 방법이 다르다

## 3. 시각화

- #Bar chart : 범주형 데이터를 bar형태로 보여준다
- #Pareto chart : Bar chart의 특수한 경우. 빈도의 역순
- #Histograms : 연속/불연속 변수를 다루는데에 쓴다

## 4. 평균과 기대값

- 1) 평균(mean & average)
  - a) 산술평균 . 일반적인 평균. average
  - b) 기하평균
  - c) 조화평균
- 2) 기대값(expectation)  
샘플을 통해 구한 모집단의 평균 값에 대한 확률적인 예상치
- 3) 평균의 일반화  
멱평균(멱함수를 이용)  
 $f$ -평균(임의의 함수를 이용) → ex)  $f(x) = x$  이면 산술,  $\log bx$ 이면 기하평균
- 4) 기타  
비대칭도(skewness) : 꼬리가 오른쪽이면 positive  
변수가  $x \rightarrow 2x$ 가 되면 평균과 표준편차는 2배. 분산은 2의 제곱배.

## 5. 확률기초

- 1)  $P(A \cap B) + P(A^c \cap B) = p(B)$
- 2)  $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- 3)  $P(A \cup B \cup C) = [P(A) + P(B) + P(C)] - [P(A \cap B) + P(A \cap C) + P(B \cap C)] + P(A \cap B \cap C)$
- 4)  $P(A|B) = P(A \cap B) / P(B)$
- 5)  $P(A \cap B) = P(A|B)P(B)$
- 6) 독립이면  $P(A|B) = P(A)$  ,  $P(A \cap B) = P(A)P(B)$
- 7) 전체확률법칙:  $P(B) = P(A_1)P(B|A_1) + P(A_2)P(B|A_2) + \dots + P(A_n)P(B|A_n)$
- 8) Bayes's Theorem. 알고자 하는 확률을 알고 있는 확률들에 대한 수식으로 바꾼다.  
(사실 암묵적으로 쓰이고 있다)

## 6. 확률변수

- 1) 이산확률변수 : 불연속적인 특정 값만 가지는 경우.  
PMF로 표현가능하다.  
ex) 동전 앞면 수
- 2) 연속확률변수 : 어떤 구간 내에서 임의의 실수 값을 가지는 경우  
PDF로 표현가능하다  
ex) 8월 최고 온도

## 7. 확률질량함수[PMF](코드 참조)

이산확률변수  $X$ 가 어떤 값  $x_i$ 를 가질 확률  $p_i$   
 $P(X=x_i) = p_i$

## 8. 누적분포함수[CDF](코드 참조)

이산,연속 확률변수 모두 표현 가능  
 $F(x) = P(X \leq x)$ .  $x$ 보다 작은 모든 경우의 확률 합  
참고)  $1 - F(x)$  를 CCDF라고 함. 파레토 분포 등을 볼 때 유용하게 쓰임

## 9. PMF와 CDF의 관계

$$P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$$

#앞으로의 이론들은 코드 중심(numpy,pandas,matplotlib)으로 구현하겠습니다.