2017-07-05 (3주차-1)

1. 난수발생과 카운팅
   컴퓨터에서 모든 무작위와 관련된 알고리즘은 사실 무작위가 아니라 시작숫자를 정해주면 그 다음에는 알고리즘에 의해 마치 난수처럼 보이는 수열을 생성함
   다만, 출력되는 숫자들 사이에 상관관계가 없을뿐임. 이러한 시작 숫자를 seed라고 한다.
   - np.random.rand : 난수 생성
   - np.random.shuffle : 기존의 데이터의 순서 바꾸기
   - np.random.choice : 기존의 데이터 일부를 선택하는 것
   - np.unique : 중복제거 (@return_counts = True 일때는 index, count 를 리턴)
   - np.bincount : 없는 데이터의 count를 0으로 세어줄때 (@minlength = number 일때 (0, number] 까지를 세어줌)

2. scipy를 이용한 확률분포 분석
   scipy 각종 수치 해석기능을 제공하는 파이썬 패키지
   a. 확률분포객체
   b. 모수지정
   c. 확률분포 메서드
      https://datascienceschool.net/view-notebook/e6c0d4ff9f4c403c8587c7d394bc930a/
3. 분포의 대표값
   분포의 특성은 집합전체의 특성을 나타내는 것으로 히스토그램등 시각화 또는 분포함수등의 수식으로 사용하는 것이 일반적이나 전체적인 특성을 나타낼 수 있는 몇개의 대표값을 구할 수 있다면 분포의 특성을 빠르게 파악하거나 분포간의 비교가 가능하다
   분포를 하나의 숫자로 대표할 수 있는 숫자가 대표값임
   - 평균 : 보통 샘플평균을 의미
   - 기댓값 : 확률 변수가 따르고 있는 확률모형, 정확히는 확률밀도함수를 알고 있는 경우의 평균
   - 중간값: 전체 자료를 크기별로 정렬하였을때 가장 중앙에 위치하게 되는 값
   - 최빈값: 가장 많이 나오는 값

   분포가 정규분포가 아닌 경우 평균, 중간값, 최대값은 같지 않다. 일반적으로 평균을 많이 이용하는 이유는 계산량이 상대적으로 적고 이용할 수 있는 수학적툴이 많기때문
   중간값은 정렬비용, 최빈값은 최적화 비용이 필요

4. 분산과 표준편차
   분포에서 폭(width)을 대표하는 값

5. 베르누이 확률 분포
   결과가 성공 혹은 실패 두 가지 중 하나로만 나오는 것을 베르누이 시도, 이 시도의 결과를 확률변수 X로 나타낼 때는 X = 1, X = 0 으로 정함, 수식이 간단해지기 때문

   - 베르누이 확률변수는 0,1 두 가지 값만을 가질 수 있음으로 이산확률변수
   - 확률 질량 함수
     Bern(x ;θ)=θ^x(1θ) ^(1x)

     만약 어떤 확률 변수 X 가 베르누이 분포에 의해 발생된다면 "확률 변수 X 가 베르누이 분포를 따른다"라고 말하고 다음과 같이 수식으로 쓴다.

     X ~Bern(x ;θ)

6. 이항 확률 분포

   성공확률이 θ θ 인 베르누이 시도를 N 번 하는 경우를 생각해 보자. 가장 운이 좋을 때에는 N N번 모두 성공할 것이고 가장 운이 나쁜 경우에는 한 번도 성공하지 못할 것이다. N번 중 성공한 횟수를 확률 변수 X 라고 한다면X의 값은 0 부터 N 까지의 정수 중 하나가 될 것이다.

   이러한 확률 변수를 이항 분포(binomial distribution)를 따르는 확률 변수라고 하며 다음과 같이 표시한다.

   X ~Bin(x ;N ,θ)

   이항 확률 분포를 수식으로 쓰면 다음과 같다.

   Bin(x ;N ,θ)=combi(N x)θ^ x(1θ)^(N x)

7. 정규분포(가우시안 정규분포)

   가우시안 정규 분포(Gaussian normal distribution), 혹은 그냥 간단히 정규 분포라고 부르는 분포는 자연 현상에서 나타나는 숫자를 확률 모형으로 모형화할 때 가장 많이 사용되는 확률 모형이다.

   정규 분포는 평균 μ μ와 분산 σ 2 σ2 이라는 두 개의 모수만으로 정의되며 확률 밀도 함수는 다음과 같은 수식으로 표현된다.

   정규 분포 중에서도 평균이 0 이고 분산이 1 인 (μ=0 μ=0, σ 2 =1 σ2=1) 정규 분포를 표준 정규 분포(standard normal distribution)라고 한다.

   - Q-Q 플롯

     정규 분포는 여러가지 연속 확률 분포 중에서도 가장 유용한 특성을 지니며 널리 사용되는 확률 분포이다. 따라서 어떤 확률 변수의 분포가 정규 분포인지 아닌지 확인하는 것은 정규 분포 검정(normality test)은 가장 중요한 통계적 분석 중의 하나이다. 그러나 구체적인 정규 분포 검정을 사용하기에 앞서 시작적으로 간단하게 정규 분포를 확인하는 Q-Q 플롯을 사용할 수 있다.

     Q-Q(Quantile-Quantile) 플롯은 분석하고자 하는 샘플의 분포과 정규 분포의 분포 형태를 비교하는 시각적 도구이다. Q-Q 플롯은 동일 분위수에 해당하는 정상 분포의 값과 주어진 분포의 값을 한 쌍으로 만들어 스캐터 플롯(scatter plot)으로

그린 것이다. Q-Q 플롯을 그리는 구체적인 방법은 다음과 같다.

a. 대상 샘플을 크기에 따라 정렬(sort)한다.
b. 각 샘플의 분위수(quantile number)를 구한다.
c. 각 샘플의 분위수와 일치하는 분위수를 가지는 정규 분포 값을 구한다.
d. 대상 샘플과 정규 분포 값을 하나의 쌍으로 생각하여 2차원 공간에 하나의 점(point)으로 그린다.
e. 모든 샘플에 대해 2부터 4까지의 과정을 반복하여 스캐터 플롯과 유사한 형태의 플롯을 완성한다.
f. 비교를 위한 45도 직선을 그린다.

- 중심극한정리
  세계에서 발생하는 현상 중 많은 것들이 정규 분포로 모형화 가능하다. 그 이유 중의 하나는 다음과 같은 중심 극한 정리(Central Limit Theorem)이다. 중심 극한 정리는 어떤 분포를 따르는 확류 변수든 간에 해당 확률 변수가 복수인 경우 그 합은 정규 분포와 비슷한 분포를 이루는 현상을 말한다.

  중심 극한 정리를 수학적인 용어로 쓰면 다음과 같다.
  X 1 ,X 2 ,...,X n X1,X2,...,Xn가 기댓값이 μ μ이고 분산이 σ 2 σ2으로 동일한 분포이며 서로 독립인 확률 변수들이라고 하자. 분포가 어떤 분포인지는 상관없다.

  이 분포의 합

  $$S_n = X_1 + \cdots + X_n \quad Sn=X1+\cdots+Xn$$

  도 마찬가지로 확률 변수이다. 이 확률 변수 S n Sn의 분포는 n n이 증가할 수록 다음과 같은 정규 분포에 수렴한다.

  $$S_n n\mu_n \sqrt{} \rightarrow d \quad N(0, \sigma 2)$$

- 대수의 법칙
  큰 수의 법칙(큰 의 , 영어: law of large numbers) 또는 대수의 법칙, 라플라스의 정리는 큰 모집단에서 무작위로 뽑은 표본의 평균이 전체 모집단의 평균과 가까울 가능성이 높다는 통계와 확률 분야의 기본 개념

9. 스튜던트 t 분포

   가우시안 정규 분포와 달리 정수값을 가지는 자유도(degree of freedom)라는 모수(parameter) ν를 추가적으로 가짐

   정규분포와 달리 양끝단의 비중이 더 큰 분포

   중심극한 정리에서 n이 무한대에 가까워지면 정규분포를 따른다 하였다.
   유한한 샘플에 대해서는 가우시안 정규 분포로부터 얻은 n 개의 샘플 x 1 ,,x n x1,,xn 로부터 얻은 샘플 평균을 샘플 표준편차로 나눈 값은 자유도가 n1 n1인 스튜던트 t 분포를 이룸
10. 정규분포의 샘플평균이 이루는 분포
    중심극한의 정리에서 모든 확률변수의 합(또는 평균)은 샘플의 수가 증가할수록 가우시안 정규분포에 가까워진다고 하였음. 하지만 샘플이 유한하다면?
    가우시안 정규분포로 부터 얻은 n개의 샘플 x1 ~ xn의 샘플평균을 샘플표준편차로 나눈 값은 자유도가 n-1인 t분포를 이룸
    샘플평균의 분포는 정규분포이지만 샘플표준편차라는 확률변수로 나누는 과정에서 t분포를 따르게 된다.
    추후에 정규분포의 기댓값에 대한 각종 검정에서 사용
11. 카이제곱분포
    가우시안 정규분포를 따르는 확률변수 X의 n개의 샘플 x1 ~ xn의 합(또는 평균)은 샘플분산으로 정규화하게 되면 t분포를 따르게 되는데 제곱을 하여 더하게 되면 양수값만을 가지는 분포가 되고 이 분포를 카이제곱분포라함
    카이제곱 분포도 자유도를 갖는다.

12. F분포
    카이제곱 분포를 따르는 독립적인 두개의 확률변수의 샘플로 부터 생성할수 있음
    두 카이제곱 분포의 샘플을 각각 x1, x2이라고 할때 이를 각각 n1,n2로 나누어 비율을 구하면 F(n1,n2) 분포가 됨
    n1,n2는 F분포의 자유도 인수

13. 카테고리 분포
    카테고리 분포는 베르누이 분포의 확장판, 카테고리 분포는 1부터 K개의 정수 값중 하나가 나오는 확률변수 분포 예를 들어 주사위라면 K=6인 카테고리 분포

    카테고리 분포의 모수 θ θ 는 베르누이 분포와 달리 다음과 같은 제약 조건을 가지는 벡터값이 된다.

    $$\theta = (\theta 1,,\theta K)$$
    $$0 \le \theta i \le 1$$
    $$\sum k=1\ K\ \theta k = 1$$

    확률변수가 벡터값임으로 각 자리에 따라 모멘트를 구한다.

14. 다항분포
    베르누이 시도를 여러번 하여 얻은 총 성공 횟수 합이 이항 분포를 이루는 것처럼 독립적인 카테고리 분포를 여러번 시도하여 얻은 각 원소의 성공횟수 값은 다항 분포(Multinomial distribution)을 이룬다.
    다항 분포는 확률 모수가 θ=(θ1,,θK)인 독립적인 카테고리 시도를 N번 반복해서 k가 각각 xk번 나올 확률 즉, 벡터 x=(x1,,xK)가 나올 확률 분포를 말한다.