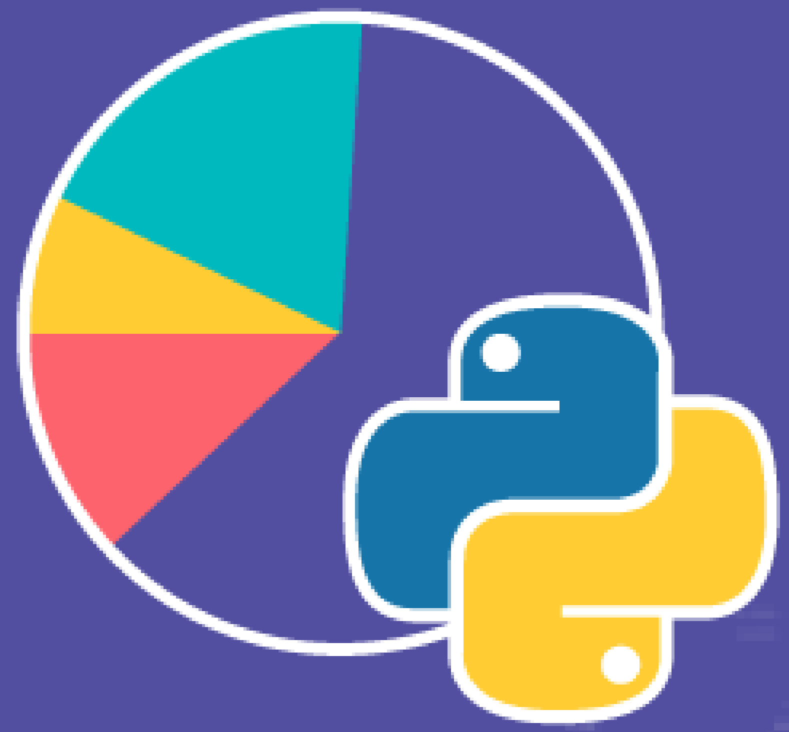
/* elice */

파이썬으로 배우는 기초 통계

추론 및 가설검정



목차

여러 가지 확률분포

확률 분포

이산 확률 분포

1) 베르누이 분포
2) 이항 분포
3) 기하 분포
4) 포아송 분포

연속 확률 분포

5) 균일 분포
6) 정규 분포

1) 베르누이 분포

베르누이 시행

1) 각 시행은 성공과 실패 두 가지 중 하나의 결과를 가짐

2) 각 시행에서 성공할 확률은 p, 실패할 확률은 1-p

3) 각 시행은 서로 독립으로 각 시행의 결과가 다른 시행의 결과에 영향을 미치지 않음

1) 베르누이 분포

베르누이 시행의 확률변수 X

확률분포

x	0	1
P(X=x)	1-p	p

확률질량함수

$$f(x) = \begin{cases} p\,, & x = 1 \\ 1 - p\,, & x = 0 \end{cases}$$

2) 이항 분포

베르누이 시행을 반복했을 때, 성공하는 횟수의 확률분포

이항 실험

성공확률이 동일한 베르누이 시행을 독립적으로 반복하는 실험

이항 확률변수

전체 시행 중 성공의 횟수에 따른 확률변수

2) 이항 분포

이항 확률변수 X 의 확률 질량 함수

$$p(x) = \binom{n}{x} \times p^x \times (1-p)^{n-x}, x = 0,1,2 \dots, n$$

시행 횟수 n은 자연수이며,
성공확률 p는 $0 \le p \le 1$을 만족

X ~ B(n,p)

시행 횟수가 n, 성공확률이 p인 이항 분포

2) 이항 분포

```
stat_bin  = scipy.stats.binom(n, p) #이항 분포 확률 변수
stat_bin.pmf(x축) # 확률 질량 함수 시각화
stat_bin.cdf(x축) # 누적 분포 함수 시각화
np.random.binomial(n, p, size) # 이항 분포 랜덤 샘플
```

- n : 시행 횟수
- p : n=1이 나올 확률
- size = 표본 추출 작업 반복 횟수

3) 초기하 분포

유한한 모집단에서 비복원 추출 시, 성공 횟수의 분포

X : 표본 내에서 **관심있는 범주**(예: 불량품 개수)에

속하는 **구성원소의 수**

불량률 계산 등에서 많이 사용

3) 초기하 분포

```
X ~ Hyper(M,n,N)
```

모집단의 크기가 M이고,

표본의 크기가 n,

관심이 있는 범주 (예: 불량품 개수)에 속하는

구성원소의 수가 N인 초기하분포

3) 초기하 분포

초기하 확률 변수 X 의 확률 질량 함수

$$p(X=x)=\frac{\binom{D}{x}\times\binom{N-D}{n-x}}{\binom{N}{n}}, x=0,1,2\ldots,n$$

여기서 n은 D 혹은 (N-D)보다 작거나 같은 수로 가정

3) 초기하 분포

상자 안에 **흰색 공 6개**와 **검은색 공 4개**가 있을 때

5개의 공을 꺼낸 결과 **흰 공이 3개**인 확률은?

3) 초기하 분포

10개 중 5개를 뽑는 경우의 수 가운데

흰색 공 **6개 중 3개**를 뽑고

검은색 공 **5개 중 2개**를 뽑을 확률

3) 초기하 분포

$$p(X=3) = \frac{\binom{6}{3} \times \binom{4}{2}}{\binom{10}{5}} = \frac{10}{21}$$

3) 초기하 분포

```
stat_hyp = scipy.stats.hypergeom(M, n, N) #초기하 분포 확률 변수
stat_hyp.pmf(x축) # 확률 질량 함수 시각화
stat_hyp.cdf(x축) # 누적 분포 함수 시각화
np.random.hypergeometric(ngood, nbad, nsample, size)
#초기하 분포 랜덤 샘플
```

ngood(= n) : 모집단 중 관심 있는 범주에 속하는 구성원소 수

nbad(=M-n) : 관심있는 표본 이외의 개수(ngood + nbad = M)

nsample(=N) : 표본의 크기

size : 표본 추출 작업 반복 횟수

4) 포아송 분포

연속된 시간 상에서
발생하는 사건은 매 순간 발생 가능

↓

시행 횟수가 많고 순간의 성공확률은
작기 때문에 이항분포로 설명하기 어려움

4) 포아송 분포

단위시간/공간에 드물게 나타나는 사건의 횟수에 대한 확률 분포

연속적인 시간에서 **매 순간에 발생할 것**으로

기대되는 평균 발생 횟수를 이용해

주어진 시간에 실제로 발생하는 사건의 횟수 분포

4) 포아송 분포

포아송 분포의 예시

일정 시간동안 발생하는 불량품의 수

일정 시간동안 톨게이트를 지나는 차량의 수

일정 페이지의 문장을 완성했을 때 발생하는 오타의 수

4) 포아송 분포

X ~ Poi(λ)

평균적으로 λ 회 발생하는 사건의
발생 횟수에 대한 포아송분포

포아송 확률 변수 X 의 확률 질량 함수

$$p(X = x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, x = 0,1,2 \ldots$$

4) 포아송 분포

이항분포 B(n,p)에서

n이 매우 크고 p가 매우 작은 경우

λ=np인 포아송 분포로 **근사 가능**

5) 균일 분포

구간 [a,b] 에 속하는 값을 가질 수 있고 그 확률이 균일한 분포

```
X ~ U(a,b)
```

5) 균일 분포

정육면체 주사위의 한 면이 나올 확률은 모두 $\frac{1}{6}$ 로 같다

$$P(X = 1,2,3,4,5,6) = \frac{1}{6}$$

5) 균일 분포

균일 확률 변수 X의 확률 밀도 함수

$a < b$를 만족하는 임의의 두 실수 a,b 에 대해 함수

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b \\ 0, & x < a \ or \ x > b \end{cases}$$ 를 정의하면,

$f(x)$를 확률 밀도 함수로 갖는 연속 확률 변수가 존재

5) 균일 분포

```
stat_uni  = scipy.stats.uniform(a, b) #균일 분포 확률 변수
stat_uni.pmf(x축) # 확률 질량 함수 시각화
stat_uni.cdf(x축) # 누적 분포 함수 시각화
np.random.uniform(a,b,n) # 균일 분포 랜덤샘플
```

a,b : 균일 분포의 구간

n : 표본 추출 작업 반복 횟수

6) 정규 분포

가장 많이 사용되고 유명한 분포

종형 곡선의 분포

평균 뮤(μ)와 표준편차 시그마(σ) 두 모수로 정의

`N(m, σ²)` 로 표시

정규 분포를 나타내는 확률 밀도 함수

$$f(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{(x-\mu)^2}{2\sigma^2}\right)$$

6) 표준 정규 분포

정규 분포의 표준 분포

평균 뮤(μ) = 0, 표준 편차 시그마(σ) = 1로 둔 정규 분포 Z

표준 정규 분포의 확률 밀도 함수

$$Z = \sigma Z + \mu$$

$$Z = \frac{(X - \mu)}{\sigma}$$

$$f(z|\mu = 0, \sigma = 1) = \frac{1}{\sqrt{2\pi}}\exp(\frac{(x)^2}{2})$$

6) 정규 분포

```
stat_nor = scipy.stats.norm(μ, σ) #정규 분포 확률 변수
stat_nor.pmf(x축) # 확률 질량 함수 시각화
stat_nor.cdf(x축) # 누적 분포 함수 시각화
np.random.normal(μ, σ, n) # 정규 분포 랜덤 샘플
```

μ : 평균

σ : 표준 편차

n : 표본 추출 작업 반복 횟수