08 강

이산형 확률분포 2

통계·데이터과학과
이긍희 교수

학습목표

1. 포아송 분포를 이해할 수 있다.
2. 초기하분포를 이해할 수 있다.

01

8강 이산형 확률분포 2

포아송 분포

포아송 분포의 정의

- 이항분포에서 발생 가능성(p)은 매우 작지만 시행 횟수(n)는 충분히 큰 경우 포아송 분포 유용

- 고속도로상에서 하루 동안 발생하는 교통사고 사망자수, 회귀질병에 의한 사망자수 등

포아송 분포의 확률질량함수

- 포아송 분포 $X \sim Poisson(\lambda)$의 확률질량함수

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

포아송 분포와 이항분포

◆ 포아송 분포는 이항 분포의 근사로 표현 가능

$$ {}_nC_x\, p^x (1-p)^{n-x} = \frac{n(n-1)\cdots(n-x+1)}{x!}\left(\frac{\lambda}{n}\right)^x \left(1-\frac{\lambda}{n}\right)^n \Big/ \left(1-\frac{\lambda}{n}\right)^x $$

$$ = \underbrace{\frac{\lambda^x}{x!}1\left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{x-1}{n}\right)}_{①}\underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{②} \Big/ \underbrace{\left(1-\frac{\lambda}{n}\right)^x}_{③} $$

포아송 분포와 이항분포

- 포아송 분포는 이항 분포의 근사로 표현 가능

① $= 1\left(1-\frac{1}{n}\right)\cdots\left(1-\frac{x-1}{n}\right) \xrightarrow[n\to\infty]{} 1$

② $= \left(1-\frac{\lambda}{n}\right)^{n} \xrightarrow[n\to\infty]{} e^{-\lambda}$

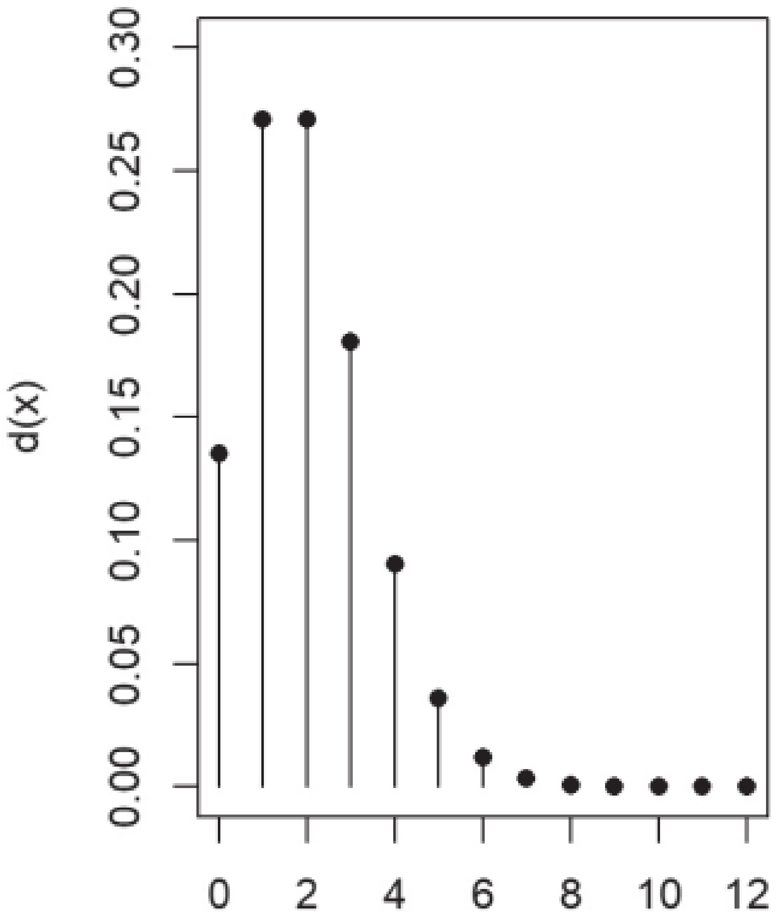
③ $= \left(1-\frac{\lambda}{n}\right)^{x} \xrightarrow[n\to\infty]{} 1$

포아송 분포와 이항분포

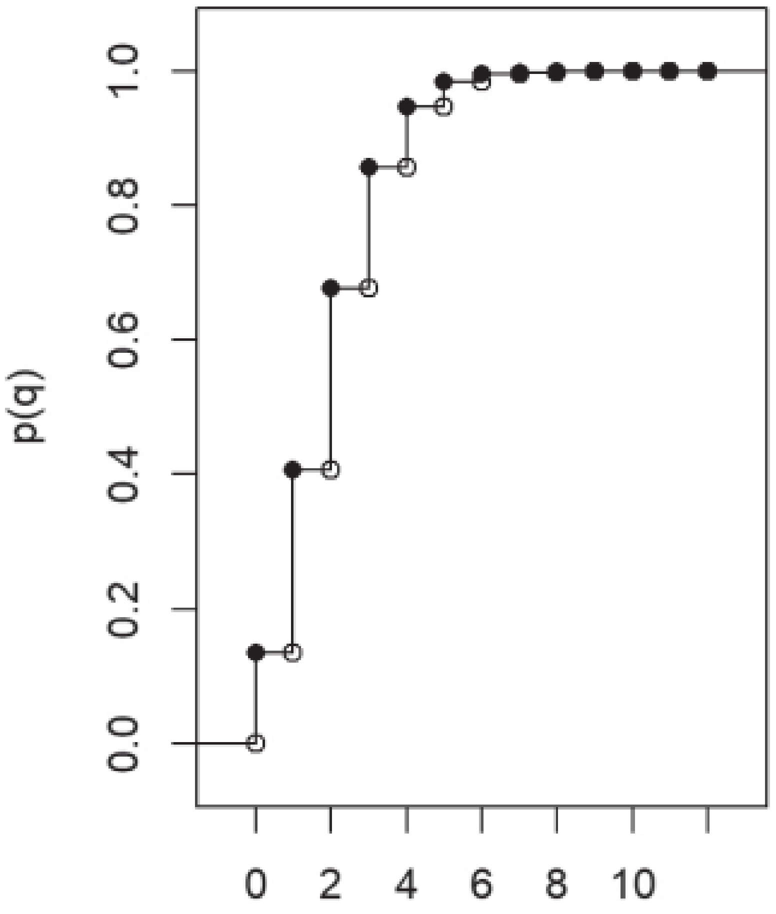
- 포아송 분포는 이항 분포의 근사로 표현 가능

$$ {}_nC_x\, p^x (1-p)^{n-x} = \underbrace{\frac{\lambda^x}{x!}\, 1 \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{x-1}{n}\right)}_{\text{①}} \underbrace{\left(1-\frac{\lambda}{n}\right)^n}_{\text{②}} \Big/ \underbrace{\left(1-\frac{\lambda}{n}\right)^x}_{\text{③}} $$

포아송 분포의 확률질량함수

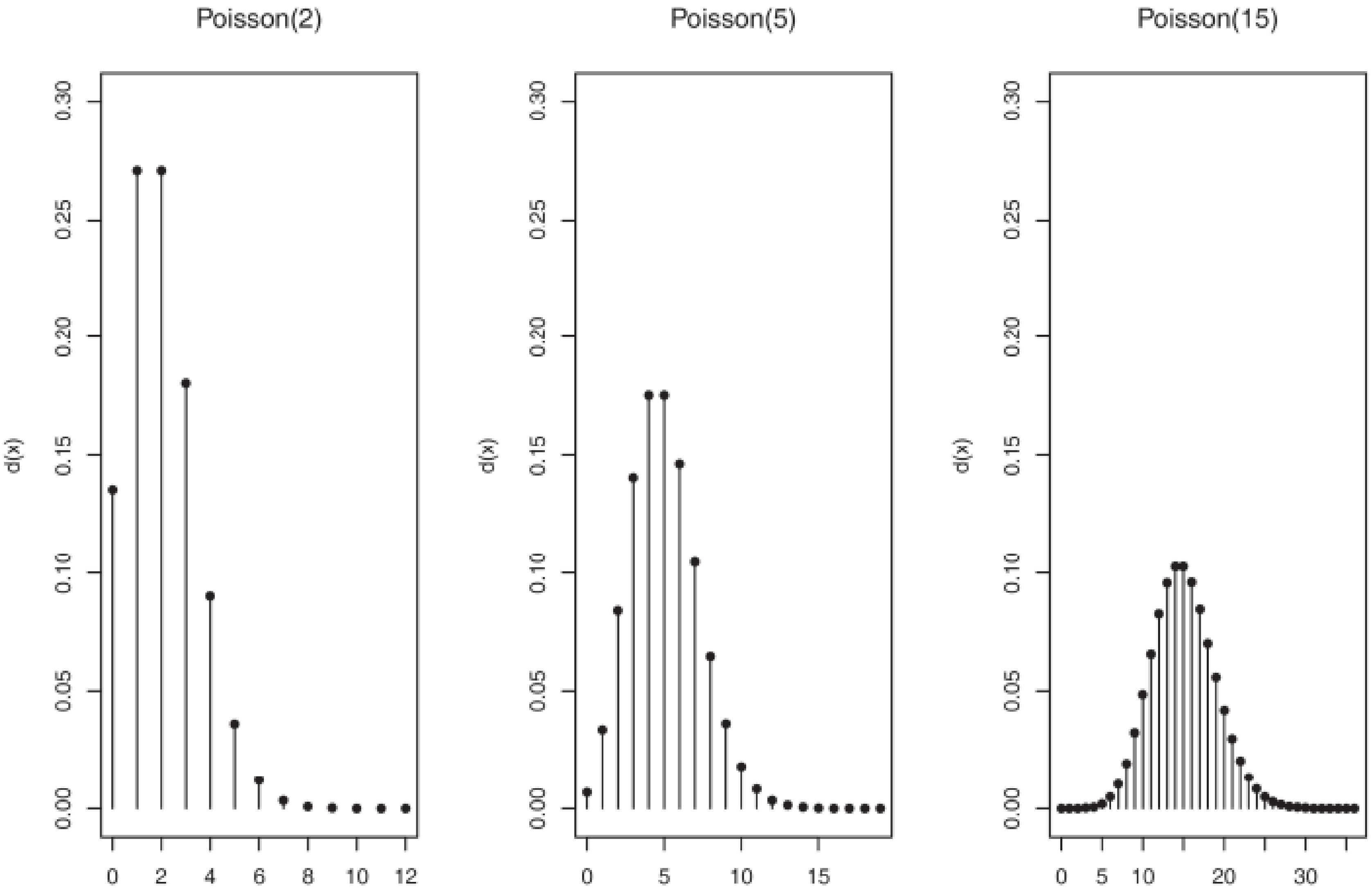


(a) 확률질량함수



(b) 누적분포함수

기댓값 변화에 따른 포아송 분포



포아송 분포의 예

예

1년 동안 위암 사망 확률 0.001, 1,000명 중 1년간 3명 이상 위암으로 사망할 확률을 포아송 분포로 구하면?

포아송 분포의 기댓값과 분산

◆ 기댓값과 분산

$$E(X) = \lambda, \quad Var(X) = \lambda$$

포아송 분포의 기댓값

- $E(X) = \lambda$의 증명

포아송 분포의 분산

- $Var(X) = \lambda$ 의 증명

포아송 분포의 분산

◆ $Var(X) = \lambda$ 의 증명

포아송 분포의 예

예 1년 동안 지진 발생 횟수(규모 3이상)는 평균 2회인 포아송 분포

(1) 1년 동안 지진이 전혀 발생하지 않을 확률은?

포아송 분포의 예

예

1년 동안 지진 발생 횟수(규모 3이상)는 평균 2회인 포아송 분포

(2) 1년 동안 지진이 1회 이상 발생할 확률은?

포아송 분포의 예

예

1년 동안 지진 발생 횟수(규모 3이상)는 평균 2회인 포아송 분포

(3) 1년 동안 지진의 발생 횟수의 분산은?

02

8강 이산형 확률분포 2

초기하분포

초기하분포 예

예

100개의 제품을 생산. 10개가 불량품.
10개를 공정하게 추출.
불량품의 개수가 1개인 경우 확률은?

초기하분포의 정의

- 모집단이 두 개의 집단으로 구분되어 있고 여기서 표본을 비복원추출할 때 추출된 한 집단 수의 분포

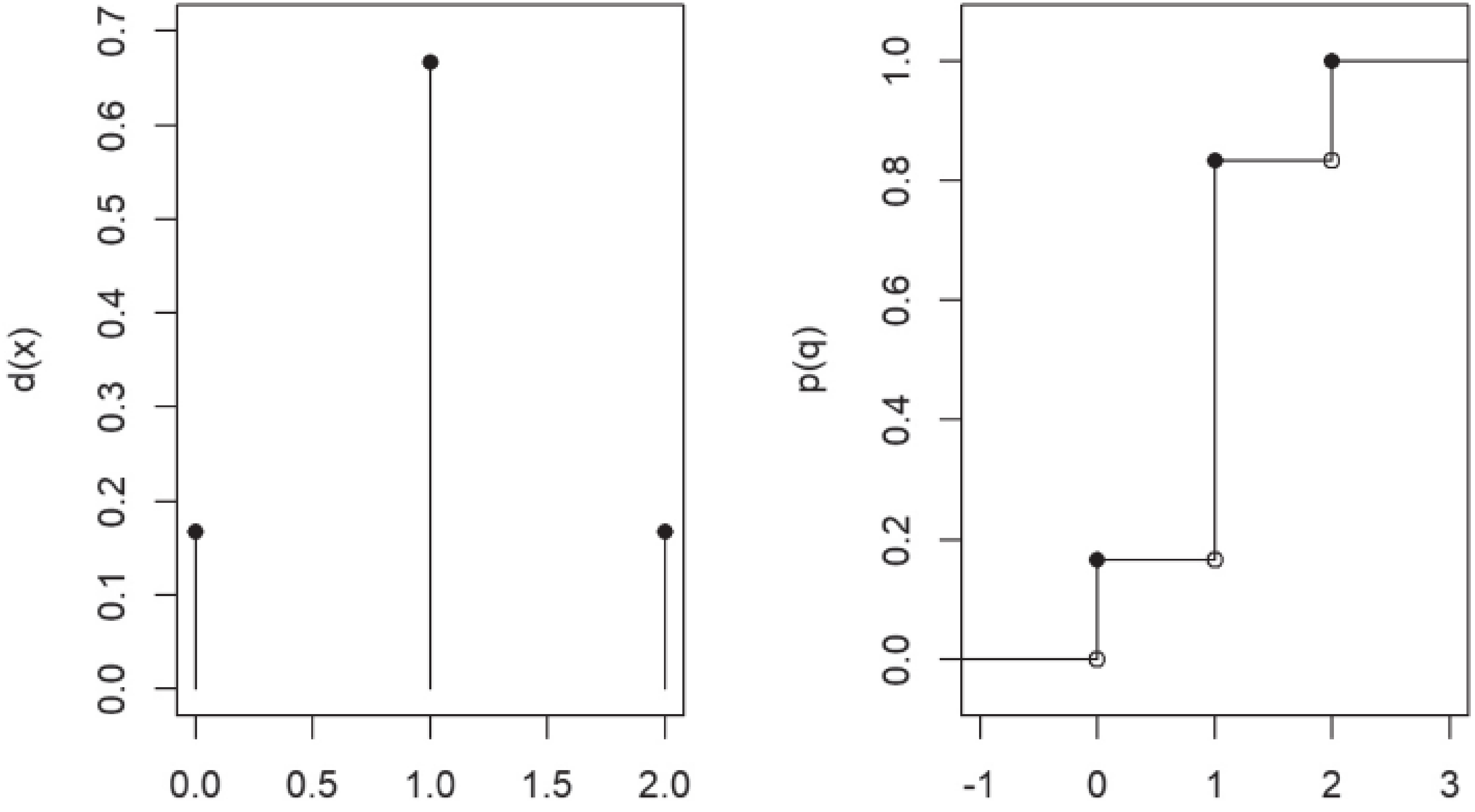
초기하분포의 확률질량함수

- 모집단수 N, 표본수 n, 모집단 불량품 수 D인 경우 표본에서 불량품 수의 분포

$$P(X=x)=\frac{{}_{D}C_{x}\ {}_{N-D}C_{n-x}}{{}_{N}C_{n}}$$

$$x=0,1,2,\cdots,n, \qquad n\le D, N-D$$

초기하분포의 확률질량함수



(a) 확률질량함수　　(b) 누적분포함수

초기하분포의 기댓값과 분산

◆ 기댓값과 분산

$$E(X) = n \cdot \frac{D}{N}$$

$$Var(X) = n \cdot \frac{D}{N} \cdot \frac{N-D}{N} \cdot \frac{N-n}{N-1}$$

초기하분포의 기댓값

- $E(X) = n \cdot \dfrac{D}{N}$ 의 증명

초기하분포의 기댓값

- $E(X) = n \cdot \dfrac{D}{N}$ 의 증명

초기하분포의 예

예

양호품 2개, 불량품 2개 중에서 임의로 2개를 비복원추출. 불량품 개수의 분포는?

초기하분포와 이항분포

◆ N이 클 때 초기하분포는 이항분포로 근사

$$P(X=x)=\frac{{}_{D}C_{x}\,{}_{(N-D)}C_{(n-x)}}{{}_{N}C_{n}}$$

$$=\frac{n!}{x!\,(n-x)!}\,\frac{D(D-1)\cdots(D-x+1)\,(N-D)\cdots(N-D-n+x+1)}{N(N-1)(N-2)\cdots(N-n+1)}$$

초기하분포와 이항분포

- N이 클 때 초기하분포는 이항분포로 근사

$$= {}_nC_x \left(\frac{D}{N}\times\frac{D-1}{N}\cdots\frac{D-x+1}{N}\right)\times\left[\left(1-\frac{D}{N}\right)\cdots\left(1-\frac{D}{N}-\frac{n-x-1}{N}\right)\right]\times\frac{N^n}{N(N-1)\ \cdots\ (N-n+1)}$$

03

8강 이산형 확률분포 2

R 실습

이항분포의 계산

예

치료율 0.8, 환자 4명, 2명 완치 확률,
완치 환자수의 기댓값, 분산은?

```
> library(distrEx)
```

```
> # 이항분포를 이용한 확률계산
> X=Binom(4,0.8)
> d(X)(2)
[1]   0.1536
```

```
> # 기댓값과 분산
> X=Binom(4,0.8)
> E(X)
[1]  3.2
> var(X)
[1]  0.64
```

포아송분포의 계산

예

위암 사망 확률 0.001, 1,000명 중 3명 이상 위암 사망 확률, 사망자 수의 기댓값과 분산은?

```
> library(distrEx)
```

```
> # 포아송 분포를 이용한 확률계산
> X=Pois(1)
> 1-p(X)(2)
[1]  0.0803014

> # 이항분포를 이용한 확률계산
> X1=Binom(1000,0.001)
> 1-p(X1)(2)
[1]  0.0802934
```

```
> # 기댓값과 분산
> E(X)
[1]  1
> var(X)
[1]  1
```

초기하분포의 계산

예

양호품 2개, 불량품 2개 중에서 임의로 2개 비복원추출. 불량품 개수의 분포, 기댓값, 분산은?

```
> library(distrEx)
> X = Hyper(2,2,2)

> # 초기하분포의 확률질량함수
> d(X)(0)
[1]   0.1666667
> d(X)(1)
[1]   0.6666667
> d(X)(2)
[1]   0.1666667
```

```
> # 기댓값과 분산
> E(X)
[1]   1
> var(X)
[1]   0.3333333
```

학습정리

- 포아송 분포는 일정기간 동안 회귀한 사건 발생 건수의 분포이다.

- 초기하분포는 유한 모집단이 두 개의 집단 (우량품, 불량품)으로 구분되어 있고, 모집단에서 표본을 비복원추출했을 때 불량품수의 확률분포이다.

정리하기

수고하셨습니다.

다음시간에는

08 강 이산형 확률분포 2

09 강 연속형 확률분포