이론 통계학

2장 확률변수와 확률분포

2021년 봄학기

전북대학교 통계학과

이산형 확률변수와 분포

베르누이 분포 (Bernoulli distribution)

베르누이 시행 (Bernoulli trial)

- 실험에서 나타날 수 있는 결과를 두 가지로 분류 - '성공'(S), '실패'(F)
- 성공과 실패 두 가지 결과만 나타나는 실험

베르누이 확률분포: $X \sim$**Bernoull**(p)

X=성공확률이 p인 베르누이 시행에서 결과가 성공이면 1, 실패면 0

$$f(x) = p^x(1-p)^{1-x}, \ x = 0, 1$$

이항 분포 (Binomial distribution)

이항 분포: $X \sim B(n, p)$

$X=$ 베르누이 시행을 독립적으로 n번 반복할 때 나오는 성공의 수

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \ x = 0, 1, \ldots, n$$

$f(x) = P(X = x) = P(n$번의 베르누이 시행 중 성공의 수$=x)$

$= (n$번의 베르누이 시행 중 성공의 수가 x인 경우의 수$) \times p^x (1-p)^{n-x}$

- $B(1, p) = Bernoulli(p)$
- $\sum_{x=0}^{n} \binom{n}{x} p^x (1-p)^{n-x} = 1$
- $M(t) = (pe^t + 1 - p)^n$
- $E(X) = np, \ Var(X) = np(1-p)$

기하 분포 (Geometric Distribution)

기하 분포: $X \sim Geometric(p)$

X=서로 독립인 베르누이 시행을 반복할 때 첫번째 성공이 나올 때까지 시행 횟수

$$f(x) = (1-p)^{x-1}p, \ x = 1, 2, \ldots$$

- $\sum_{x=1}^{\infty}(1-p)^{x-1}p = 1$
- $M(t) = pe^t/(1-qe^t)$ for $t < -\ln q$
- $E(X) = 1/p, \ Var(X) = q/p^2$

음이항 분포(Negative Binomial Distribution)

음이항 분포: $X \sim NB(r, p)$

$X=$ 서로 독립인 베르누이 시행을 r번째 성공이 나올 때까지의 시행 횟수

$$f(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r} \quad x = r, r+1, r+2, \ldots$$

$$P(X = x) = P(\ x\text{번째 시행에서 } r\text{번째 성공})$$

$$= P(\text{처음 } (x-1)\text{번 시행에서 성공이 } (r-1)\text{번,}$$

$$\text{마지막 } x\text{번째 시행에서 성공})$$

- $NB(1, p) = Geometric(p)$
- $M(t) = \left(\frac{pe^t}{1-qe^t} \right)^r$ for all $t < -\ln q$
- $E(X) = r/p, \quad Var(X) = rq/p^2$

$Bernoulli(p), B(n,p), Geometric(p), NB(r,p)$

- 네 분포 모두 성공확률이 p로 일정한 베르누이 시행을 독립적으로 반복한 결과
- $Bernoulli(p), B(n,p)$- 시행 횟수가 미리 고정됨
- $Geometric(p), NB(r,p)$ - 기다리는 시간이 미리 고정됨

초기하 분포 (Hypergeometric Distribution)

초기하 분포: $X \sim Hyper(n, m, r)$

X=r개의 원소로 이루어진 하나의 그룹 A와 $(n-r)$개의 원소로 이루어진 또 다른 그룹 B가 섞여 있을 때, 이 중에서 m개를 임의로 선택할 때 A그룹에서 선택된 원소의 수

$$f(x) = \frac{\binom{r}{x}\binom{n-r}{m-x}}{\binom{n}{m}}, \ \max(0, m-n+r) \le x \le \min(r, m)$$

- $\sum_x f(x) = 1$ (check)
- $E(X) = m\left(\frac{r}{n}\right), \quad Var(X) = m\left(\frac{r}{n}\right)\left(1-\frac{r}{n}\right)\frac{n-m}{n-1}$

포아송 분포 (Poisson Distribution)

시간 또는 공간에 따른 특정 사건의 발생 건수에 대한 확률분포로 이용 됨

예) 하루동안 걸려오는 전화통화 수, 1년동안 발생하는 지진의 수...

포아송 분포: $X \sim Poisson(\lambda)$

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

$\lambda =$ 단위 (시간 또는 공간)당 평균적으로 발생하는 사건의 수

- $\sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} = 1$
- $M(t) = \exp[\lambda(e^t - 1)]$
- $E(X) = \lambda, \quad Var(X) = \lambda$

포아송 분포 (Poisson Distribution)

포아송확률과정 조건

(1) 아주 짧은 구간에서 사건이 2회 이상 발생할 확률은 거의 0에 가깝다

(2) 아주 짧은 구간에서 사건이 발생할 확률은 구간은 길이에 비례

(3) 서로 중첩되지 않는 구간에서 발생하는 사건의 수는 서로 독립

$$P(X = x) \approx \binom{n}{x} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x}$$
$$\rightarrow \frac{\lambda^x e^{-\lambda}}{x!}$$

연속형 확률 변수와 분포

균일 분포 (Uniform Distribution)

일정한 구간에서 임의로 선택된 수에 대한 확률분포로 이용

균등 분포$X \sim U(a, b)$

$$f(x) = \frac{1}{b-a} I(a < x < b)$$

- $E(X) = \frac{a+b}{2} \quad Var(X) = \frac{(b-a)^2}{12}$

베타 분포(Beta Distribution)

유한 구간에서 값을 가지는 변수의 분포로 이용

예) 비율의 분포

베타 분포 : $X \sim Beta(\alpha, \beta)$

$$f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1-x)^{\beta-1} I(0 < x < 1),$$

여기에서

$$B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

- $Beta(1, 1) = Uniform(0, 1)$
- $E(X) = \frac{\alpha}{\alpha+\beta} \quad Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$

지수 분포 (Exponential Distribution)

포아송분포를 따르며 발생하는 사건에 대하여 첫번째 사건이 발생할 때까지 대기 시간 (waiting time)의 분포

지수 분포: $X \sim Exp(1/\lambda)$

$$f(x) = \lambda e^{-\lambda x} I(x > 0)$$

$$P(X \leq t) = 1 - P(X > t) = 1 - P(\text{첫 번째 사건이 발생하는 시간} > t)$$

$$= 1 - P([0, t] \text{ 사이에 발생하는 사건의 수} = 0)$$

- 비기억성 (memoryless property): $P(X > t + s | X > s) = P(X > t)$
- $E(X) = 1/\lambda, \ Var(X) = 1/\lambda^2$

감마 분포 (Gamma Distribution)

감마 분포: $X \sim Gamma(\alpha, \beta)$

$$f(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta} I(x > 0)$$

Gamma 함수

$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1} e^{-x} dx$$

- $\Gamma(\alpha+1) = \alpha\Gamma(\alpha),\ \Gamma(n) = (n-1)!, \Gamma(1/2) = \sqrt{\pi}$
- $Gamma(1, \beta) = Exp(\beta)$
- $M(t) = \left(\frac{1}{1-\beta t}\right)^{\alpha}$
- $E(X) = \alpha\beta,\ Var(X) = \alpha\beta^2$

감마 분포 (Gamma Distribution)

- 평균이 λ인 포아송분포를 따르며 발생하는 사건에 대하여 n번째 사건이 발생할 때까지 대기 시간 $\sim Gamma(n, 1/\lambda)$

$$P(X \le t) = 1 - P(X > t) = 1 - P(n \text{ 번째 사건이 발생하는 시간} > t)$$
$$= 1 - P([0, t] \text{ 사이에 발생하는 사건의 수} \le n - 1)$$

- 자유도가 n인 카이제곱분포: $\chi^2_n \sim Gamma\left(\frac{n}{2}, 2\right)$

$$f(x) = \frac{1}{\Gamma\left(\frac{n}{2}\right)} \frac{x^{n/2-1}}{2^{n/2}} e^{-x/2} I(x > 0)$$

정규분포 (Normal Distribution)

정규분포:, $X \sim N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

- $M(t) = e^{\mu t + \sigma^2 t^2/2}$
- $E(X) = \mu, \ Var(X) = \sigma^2$
- 표준정규분포 $N(0,1)$의 누적분포함수 $\Phi(\cdot)$로 표기

이변량 정규분포 (Bivariate Normal Distribution)

이변량 정규분포:, $\begin{pmatrix} X \\ Y \end{pmatrix} \sim N_2\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}\right)$

$$f(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}$$
$$\exp\left[-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x-\mu_X}{\sigma_X}\right)^2 + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right)\right)\right]$$

- (X, Y) 이변량 정규분포를 따르는 경우, $\rho = 0 \Leftrightarrow X, Y$ 독립

변수변환

정리 2.1
$f_X(x)$는 연속형 확률변수 X의 확률밀도함수이며 함수 $g(\cdot)$의 역함수가 존재하고 미분 가능할 때 $Y = g(X)$의 확률밀도함수는 다음과 같다.

$$f_Y(y) = f_X(g^{-1}(y))\left|\frac{d}{dy}g^{-1}(y)\right|$$

정리 2.2
연속형 확률변수 X의 누적확률분포함수가 $F(\cdot)$일 때 $Y = F(X)$라고 하면, Y는 균일 분포를 따른다. 즉 다음이 성립한다.

$$Y = F(X) \sim Uniform(0, 1)$$

정리 2.3
$U \sim Uniform(0,1)$을 따르는 확률변수이고 $F(\cdot)$은 연속형 확률분포의 누적확률분포함수일 때 $Y = F^{-1}(U)$라고 하면, Y는 누적분포함수 $F(\cdot)$을 갖는 확률변수가 된다.

지수분포 $Exp(1/\lambda)$를 따르는 random number 생성

- 누적분포함수 $F(x) = 1 - \exp^{-\lambda x}$, 역함수 $F^{-1}(x) = -\frac{1}{\lambda}\log(1-x)$
- $U \sim Uniform(0,1)$에 대하여 $Y = -\frac{1}{\lambda}\log(1-U) \sim Exp(1/\lambda)$

  $-\frac{1}{\lambda}\log(1-U_1), -\frac{1}{\lambda}\log(1-U_2), \ldots, -\frac{1}{\lambda}\log(1-U_n)$

Homework Week 5

- 0부터 1사이의 실수 중에서 임의로 20개를 선택하였을 때, 이 중 0.7보다 큰 숫자가 나올 횟수에 대한 확률분포는?
- 사과 20개가 들어있는 한 상자에서 5개를 임의로 골라 상태를 보고 불량 사과가 없으면 구입할 것이다. 만약 검사할 상자에 4개의 불량 사과가 들어있다고 하자. 그 상자를 구입하지 않을 확률은?
- 기계 A가 1시간 동안 오작동할 확률은 0.02이다. 이 기계가 두 시간 동안 오작동하지 않을 확률은?
- 헌혈 지원자의 80%는 헌혈 가능자라고 한다. 5명의 헌혈지원자 중 적어도 한 명이 헌혈 가능자일 확률은?

Homework Week 5

- 지질학 연구 결과 석유 탐사 중 석유 발견 확률은 0.2이라고 한다. 석유가 3번 발견될 때까지 탐사를 계속하기로 하였다. 몇 번은 탐사해야할까?
- 만약 낙하산이 A지점과 B지점 임의의 지점에 떨어진다고 하자. 낙하산이 B보다 A지점에 더 가까이 떨어질 확률은?
- 어느 사람에게 걸려오는 전화 통화 수가 평균적으로 1시간에 0.5통인 포아송분포를 따른다고 할 때 다음 통화가 걸려올 때까지의 대기 시간이 1시간 이상일 확률은?
- 어느 사람에게 걸려오는 전화 통화 수가 평균적으로 1시간에 0.5통인 포아송분포를 따른다고 할 때 전화 3통이 걸려올 때까지의 대기 시간이 1시간 이상일 확률은?