제 0 강 통계학 Review

Part I. 확률론 (Probability Theory)

I. 확률변수 (Random Variable)와 확률분포

A. 확률변수 X 는 표본공간 Ω 상에서 정의되는 **real valued function** 임

i. 어떤 확률적 실험의 결과로 나올 수 있는 모든 가능한 결과에 대해 어떤 실수값이 대응되어야 함

ii. 하나의 실험에 대해 여러 가지의 확률변수가 정의될 수 있음

1. 주사위 던지는 실험: 던진 결과 나오는 값을 대응 시켜주는 확률변수 X, 짝수는 0 홀수는 1 을 대응시켜주는 확률변수 Y 등등

B. 확률(밀도)함수 (Probability (density) Function)

i. 확률함수 : 확률변수 X 가 취하는 값이 이산적(discrete)인 경우, X 가 취하는 각 값에 대해 취하는 확률을 대응시켜주는 함수

ii. 확률밀도함수 : 확률변수 X 가 취하는 값이 연속적(continuous)한 경우, X 가 취하는 값이 I 라는 구간에 속할 확률이 $P\left(X \in I\right) = \int_I f(x)dx$ 로 주어질 때, $f(x)$ 를 확률밀도함수라 함

C. (누적) 분포함수 ((Cumulative) Distribution Function): 확률변수 X 의 분포함수 : $F(x) \equiv \mathrm{P}\left(X \le x\right)$

D. 결합 확률(밀도)함수 (Joint Probability (density) Function):

i. 두 개의 확률변수 X, Y 가 이산적이 경우 결합확률함수는 $f_{xy}\left(x_i, y_j\right) \equiv P\left(X = x_i, Y = y_j\right)$ 와 같이 주어지며

ii. 두 개의 확률변수 X, Y 가 연속적이 경우 결합확률밀도함수 $f_{xy}\left(x, y\right)$ 는 $P\left(\left(X, Y\right) \in D\right) = \iint_D f_{xy}\left(x, y\right) dxdy$ 와 같이 주어진다.

E. 한계 확률(밀도)함수 (Marginal Probability (density) Function)

i. $f_x\left(x_i\right) = \sum_j f_{xy}\left(x_i, y_j\right),\ i = 1, 2, \ldots$: X 의 한계확률함수

ii. $f_x\left(x\right) = \int_{-\infty}^{\infty} f_{xy}\left(x, y\right) dy$: X 의 한계확률밀도함수

F. 조건부 확률(밀도)함수 (Conditional Probability (density) Function)

i. 이산적인 경우: $Y = y_j$ 로 주어졌을 때, X 의 조건부확률함수는

$$f_{x|y}\left(x_i \mid y_j\right) \equiv P\left(X = x_i \mid Y = y_j\right) = \frac{P\left(X = x_i, Y = y_j\right)}{P\left(Y = y_j\right)} = \frac{f_{xy}\left(x_i, y_j\right)}{f_y\left(y_j\right)} = \frac{f_{xy}\left(x_i, y_j\right)}{\sum_i f_{xy}\left(x_i, y_j\right)}$$

와 같이 주어짐

ii. 연속적인 경우: $Y = y$ 로 주어졌을 때, X 의 조건부확률밀도함수

$$f_{x|y}(x \mid y) == \frac{f_{xy}(x,y)}{f_y(y)} = \frac{f_{xy}(x,y)}{\int_{-\infty}^{\infty} f_{xy}(x,y)\,dx}$$ 와 같이 주어짐

G. 수학적 기대값 (Mathematical Expectation)

i. $g(X)$ 가 확률변수 X 의 함수라고 할 때,

$E(g(X)) \equiv \sum_i g(x_i) f(x_i)$ (이산적인 경우),

$E(g(X)) \equiv \int_{-\infty}^{\infty} g(x) f(x) dx$ (연속적인 경우)

1. $g(X) = X$ 일 때, $E(X)$ 를 확률변수 X 의 평균(보통 μ_X 또는 μ 로 표기)이라하고,

2. $g(X) = (X - E(X))^2$ 일 때, $V(X) \equiv E\left[(X - E(X))^2\right]$ 을 확률변수 X 의 분산 (보통 σ_X^2 또는 σ^2 로 표기)이라고 함

   a. $\sigma_X \equiv \sqrt{V(X)}$: X 의 표준편차(Standard Deviation)

ii. $g(X,Y)$ 가 확률변수 X, Y 의 함수라고 할 때,

$E(g(X,Y)) \equiv \sum_i \sum_j g(x_i, y_j) f(x_i, y_j)$ (이산적인 경우),

$$E\big(g(X,Y)\big) \equiv \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)f(x,y)dxdy$$ (연속적인 경우)

1. $g(X,Y)=(X-\mu_X)(Y-\mu_Y)$ 일 때, $\mathrm{E}\left[(X-\mu_X)(Y-\mu_Y)\right]$ 를 X 와 Y 의 공분산(Covariance) 이라고함 ($Cov(X,Y)$ 로 표기)

iii. 수학적 기대값와 관련된 사항 몇 가지

1. $V(X)=E(X^2)-\left[E(X)\right]^2$, $Cov(X,Y)=E(XY)-E(X)\cdot E(Y)$

2. $E(c)=c$, $V(c)=0$, c 가 상수일 때,

3. $E(aX+bY+c)=aE(X)+bE(Y)+c$,

   $V(aX+bY+c)=a^2V(X)+b^2V(Y)+2abCov(X,Y)$ (a, b, c 는 상수)

4. $E\left(\sum_i a_iX_i\right)=\sum_i a_iE(X_i)$, $V\left(\sum_i a_iX_i\right)=\sum_i\sum_j a_ia_jCov(X_i,X_j)$ a_i 는 상수, 단 $Cov(X_i,X_i)=V(X_i)$

5. $\rho_{XY} \equiv \dfrac{Cov(X,Y)}{\sqrt{V(X)}\cdot\sqrt{V(Y)}}$: 확률변수 X, Y 간의 상관계수(correlation coefficient)

H. 확률적 독립 (Stochastic Independence)

i. 일반적으로 n 개의 확률변수 $X_1, X_2, ..., X_n$ 가 서로 확률적으로 독립(mutually Stochastically Independent)이라는 것은 각 확률변수의 한계확률(밀도)함수를 $f_1(x_1), f_2(x_2), ..., f_n(x_n)$ 이라고 할 때, $X_1, X_2, ..., X_n$ 의 결합확률밀도함수가 $f\left(x_1, x_2, ..., x_n\right) = f_1\left(x_1\right) \times f_2\left(x_2\right) \times ... \times f_n\left(x_n\right) = \prod_i f_i\left(x_i\right)$ 와 같이 한계확률밀도함수의 곱으로 나타나는 경우로 정의된다.

ii. 확률적 독립과 관련된 몇 가지 사항

1. 확률변수 $X_1, X_2, ..., X_n$ 가 서로 확률적으로 독립일 경우,

   a. $E\left\{\prod_{i=1}^{n} u_i\left(X_i\right)\right\} = \prod_{i=1}^{n} E\left(u_i\left(X_i\right)\right)$, ex.) $E\left(X_1 X_2\right) = E\left(X_1\right) E(X_2)$

   b. $Cov\left(X_i, X_j\right) = 0,\ \rho_{X_i X_j} = 0$ (그러나 이 경우 逆은 일반적으로 성립하지 않는다)

   c. $Var\left(\sum_i a_i X_i\right) = \sum_i a_i^2 V\left(X_i\right)$, ex) $V\left(X_1 \pm X_2\right) = V\left(X_1\right) + V(X_2)$

II. 이산적 확률분포들 (Discrete Probability Distributions)

A. 베르누이 실험 (Bernouilli Experiment) : 가능한 결과가 오직 두 가지로만 나타나는 실험 (편의상 그 두 가지 결과 중 하나를 성공(S)로 나머지 하나를 실패(F)로 명명.

i. 확률변수 X 를 한 번의 베르누이 실험에서 나오는 성공의 횟수라고 하고, 성공의 확률을 p 라고 하면 X 의 확률분포는 다음과 같다.

1. $P(X=1)=p$, $P(X=0)=1-p\equiv q$ 또는 $f(i)=p^i q^{1-i}$, $i=0,1$

2. $E(X)=p$, $V(X)=pq$

확률분포명 및 표기	확률변수	확률함수	평균, 분산	기타
Binomial 분포 (이항분포) $\mathscr{B}(n,p)$	N 번의 독립적 베르누이 실험시, 성공의 횟수	$f(i)=\binom{N}{i}p^i q^{N-i}$, $i=0,1,2,\ldots,N$	$E(X)=Np$ $V(X)=Npq$	
Poisson 분포	시간당 ψ의 비율로	$f(i)=\dfrac{\lambda^i e^{-\lambda}}{i!}$, $i=0,1,2,\ldots$	$E(X)=\lambda$	λ_1, λ_2 의 모수를

$\mathscr{P}(\lambda)$ $\lambda = T \times \psi$	물고기가 잡힐 때, T 시간 동안 잡힐 물고기의 수 n 이 충분히 크고 p 가 충분히 작은 값일 때, $\mathscr{B}(n,p)$는 $\mathscr{P}(np)$로 근사		$V(X) = \lambda$	갖는 두 개의 독립적인 Poisson 분포의 합은 $\lambda_1+\lambda_2$ 의 모수를 갖는 포아송분포를 함
Multinomial 분포	각 실험이 3 개(혹은 그 이상)의 결과(A, B, C)가 가능한 실험을 독립적으로 행할 때, A 의 횟수(X)와 B 의 횟수(Y)	$f(i,j) = \binom{n}{i,j,k} p_X^i p_Y^j p_Z^k,$ $i = 0,1,\ldots n,\ j = 0,1,\ldots n-i$ 단, $k = n-i-j$, $p_Z = 1 - p_X - p_Y$	$Cov(X,Y)$ $= -np_X p_Y$	X 와 Y 의 한계확률분포는 각각 이항분포 $\mathscr{B}(n,P_X)$, $\mathscr{B}(n,P_Y)$임

III. 연속적 확률분포들 (Continuous Probability Distributions)

A. 기본 분포들

확률분포명 및 표기	확률밀도함수, 분포함수	평균, 분산	기타
Uniform 분포 $\mathcal{U}(a,b)$	$f(x)=\frac{1}{b-a},\ a<x<b$, zero elsewhere. $F(x)=0,\ x\le a,$ $=\frac{x-a}{b-a},\ a<x<b$ $=1, x\ge b$	$E(X)=\frac{a+b}{2}$ $V(X)=\frac{(b-a)^2}{12}$	
Exponential 분포 $\mathcal{E}(\beta)$, 단, β는 물고기 한 마리 낚는데 걸리는 시간	$f(x)=\frac{1}{\beta}e^{-\frac{x}{\beta}},\ x>0$, zero elsewhere $F(x)=0,\ x\le 0,$ $=1-e^{-\frac{x}{\beta}}\ \ x>0$	$E(X)=\beta$ $V(X)=\beta^2$	물고기 한마리 낚는데 걸리는 시간이 β (또는 시간당 ψ≡1/β 의 비율로 물고기가 낚임)일 때 첫 물고기를 낚을 때 까지 걸리는 시간이 Exponential 분포
Normal 분포(정규분포) $\mathcal{N}(\mu,\sigma^2)$	$f(x)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},\ -\infty<x<\infty$	$E(X)=\mu$ $V(X)=\sigma^2$	$X_1\sim N(\mu_1,\sigma_1^2)$, $X_2\sim N(\mu_2,\sigma_2^2)$ aX_1+bX_2 $\sim N(a\mu_1+b\mu_2, a^2\sigma_1^2+b^2\sigma_2^2+2ab\sigma_{12})$

B. 변환을 통해 도출되는 분포들

확률분포명 및 표기	변환	확률밀도함수	평균, 분산	기타
(표준) Cauchy 분포 $\mathcal{C}(0,1)$		$f(x)=\frac{1}{\pi\left[1+x^2\right]}$, $-\infty<x<\infty$	0 을 중심으로 대칭인 분포이나, 꼬리 부분이 표준정규분포에 비해 두꺼우며, 적률이 존재하지 않는다.	
Chi-square(카이제곱) 분포 χ_n^2 :자유도가 인 카이제곱분포	$X=\sum_{i=1}^{n}Z_i^2$ $Z_1,\ldots,Z_n$ 은 서로 독립인 표준정규분포	$f(x)=\frac{x^{\frac{n}{2}}e^{-\frac{x}{2}}}{2^{\alpha}\Gamma\left(\frac{n}{2}\right)}$, $x>0$, zero elsewhere (굳이 외울 필요는 없음)	$E(X)=n$ $V(X)=2n$	
(Student) t 분포 t_n :자유도가 n 인 t 분포	$X=\frac{Z}{\sqrt{V/n}}$ $Z\sim N(0,1)$, $V\sim\chi_n^2$, Z, V 는 서로 확률적 독립	$f(x)=\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\sqrt{n\pi}}\frac{1}{\left(1+\frac{x}{n}\right)^{\frac{n+1}{2}}}$, $-\infty<x<\infty$ (굳이 외울 필요는 없음)	$E(X)=0$, ($n\geq 2$ 인 경우에만 존재) $V(X)=\frac{n}{n-2}$, ($n\geq 3$ 인 경우에만 존재)	자유도가 1 인 t 분포는 표준 Cauchy 분포가 됨

F-분포 $F_{n,m}$: 분자의 자유도가 n 이고 분모의 자유도가 m 인 F 분포	$X=\dfrac{V_1/n}{V_2/m}$ $V_1 \sim \chi_n^2$, $V_2 \sim \chi_m^2$ V_1, V_2 는 서로 확률적 독립	$f(x)=\dfrac{\Gamma\left(\frac{n+m}{2}\right)}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)}\left(\dfrac{n}{m}\right)^{\frac{n}{2}}\dfrac{x^{\frac{n}{2}-1}}{\left(1+\frac{n}{m}x\right)^{\frac{n+m}{2}}}$, $x>0$, zero elsewhere (굳이 외울 필요는 없음)	$E(X)=\dfrac{m}{m-2}$, ($n\geq 3$ 인 경우에만 존재) $V(X)=\dfrac{2m^2(n+m-2)}{(m-2)^2(m-4)n}$, ($n\geq 5$ 인 경우에만 존재) (굳이 외울 필요는 없음)	X 가 $F_{n,m}$ 인 경우 1/X 는 $F_{m,n}$ 임 $t_n^2=F_{1,n}$

C. 기타 확률분포들

확률분포명 및 표기	변환	확률밀도함수	평균, 분산	기타
Lognormal 분포	$Y=\exp(X)$, $X\sim N(\mu,\sigma^2)$	$f(y)=\dfrac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\left(\frac{\ln y-\mu}{\sigma}\right)^2}\dfrac{1}{y}$, $y>0$		
Logistic 분포		$f(x)=\dfrac{e^{-x}}{\left(1+e^{-x}\right)^2}$, $-\infty<x<\infty$, $F(x)=\dfrac{1}{1+e^{-x}}$	0 을 중심으로 대칭인 분포이나, 꼬리 부분이 표준정규분포에 비해 두꺼움	