여러 가지 확률분포

hypergeometric distribution

> 초기하분포

정의

모집단에서 n개를 단순랜덤추출하여 얻은 표본 중 1(참)의 개수를 X라고 할 때, X의 확률 밀도함수는 아래의 식과 같고,
이러한 X의 분포를 hypergeometric distribution라고 한다. 기호는 $X\ H(n;N,D)$이다.

$$f(x) = \mathrm{P}(X=x) = \binom{D}{x}\binom{N-D}{n-x}/\binom{N}{n},\ 0 \le x \le D,\ 0 \le n-x \le N-D$$

hypergeometric distribution 평균

> 기호 $H(n;N,D)$

$$E(X) = np$$

유도

$$\begin{aligned} E(X) &= \sum_x \binom{D}{x}\binom{N-D}{n-x}/\binom{N}{n} \\ &= D\sum_x \binom{D-1}{x-1}\binom{N-1-(D-1)}{n-1-(x-1)}/\binom{N}{n} \\ &= D\binom{N-1}{n-1}/\binom{N}{n} \\ &= nD/N \end{aligned}$$

- 위 유도과정중 3번째 등식은 아래의 다항식에서 t^{n-1}의 계수를 비교하여 밝힐 수 있다.

$$(1+t)^{D-1}(1+t)^{N-1-(D-1)} = (1+t)^{N-1}$$

hypergeometric distribution 분산

$$Var(X) = \frac{N-n}{N-1}np(1-p)$$

유도

$$E[X(X-1)] = D(D-1)\binom{N-2}{n-2}/\binom{N}{n} = n(n-1)D(D-1)/N(N-1)$$

$$\begin{aligned} Var(X) &= E(X^2) - (E(X))^2 \\ &= E[X(X-1)] + E(X) - (E(X))^2 \\ &= \frac{N-n}{N-1}n\frac{D}{N}(1-\frac{D}{N}) \end{aligned}$$

N이 n보다 충분히 큰 hypergeometric distribution

> 전체크기가 표본크기보다 충분히 큰 초기하분포

모집단의 크기 N이 표본 크기 n에 비해 충분히 큰경우 다름과 같이 초기하분포 확률에 대한 근사계산이 가능하다.

$$\begin{aligned}\binom{D}{x}\binom{N-D}{n-x}/\binom{N}{n} &= \frac{D!}{x!(D-x)!}\frac{(N-D)!}{(n-x)!}\frac{n!(N-n)!}{N!} \\ &= \binom{n}{x}\frac{D(D-1)\cdots(D-x+1)}{N(N-1)\cdots(N-x+1)}\frac{(N-D)\cdots(N-D-n+x+1)}{(N-x)\cdots(N-n+1)} \\ &\fallingdotseq \binom{n}{x}(\frac{D}{N})^x(1-\frac{D}{N})^{n-x}\end{aligned}$$

위와 같은 계산은 복원추출에 의한 확률과 같다.
여기에서 모집단의 크기 N이 커짐에 따라 비복원추출의 효과가 없어지고 마치 복원추출하는 것과 같게 된다는 것을 알 수 있다.

binomial distribution와 multinomial distribution

binomial distribution의 일반적 정의

> 이항 분포, 기호: $X \sim B(n,p)$

모집단을 구성하는 각 개체의 특성이 '0' 또는 '1'로 분류되어 있고 '1'의 비율이 p일 때, 모집단에서 한 개씩 동일 확률의 복원추출에 의해 뽑은 n개 중에서 '1'의 개수를 X라고 하면 X의 확률밀도함수는 다음과 같다. 이때의 분포를 **binomial distribution** 라고 한다.

binomial distribution의 평균

$X\ B(n,p)$이면

$$E(X) = np$$

유도

> 이항정리 사용

$$\begin{aligned}E(X) &= \sum_{x=0}^{n} x\binom{n}{x}p^x(1-p)^{n-x} = \sum_{x=1}^{n} n\binom{n-1}{x-1}p^{(x-1)+1}(1-p)^{n-1-(x-1)} \\ &= np\sum_{k=0}^{n-1}\binom{n-1}{k}p^k(1-p)^{n-1-k} = np(p+(1+p))^{n-1} = np\end{aligned}$$

binomial distribution의 분산

$X\ B(n,p)$이면

$$Var(X) = np(1-p)$$

유도

먼저 $E[X(X-1)]$에 대한 값을 다음과 같이 구한다.

$$E[X(X-1)] = n(n-1)p^2$$

그후 위를 이용하여 분산을 구하면 다음과 같다.

$$\begin{aligned} Var(X) &= E(X^2) - (E(X))^2 = E[X(X-1)] + E(X) - (E(X))^2 \\ &= np(1-p) \end{aligned}$$

Bernoulli distribution

> 베르누이 분포, 기호: $Z_i \sim Bernoulli(p)$

각각의 복원추출이 독립인 추출결과를 $Z_1, Z_2, \cdots, Z_n, \cdots$이라고 하자.
각각의 분포가 아래와 같이 같을 때, 모집단의 분포(Z_i의 분포)를 Bernoulli distribution라고 한다.

$$P(Z_i = 1) = p, P(Z_i = 0) = 1 - p \ \ (i = 1, 2, \cdots, n)$$

Bernoulli trial

> 베르누이시행

두 가지로 분류된 모집단을 관측하는 것을 이른다.

- 서로 독립인 베르누이시행이란 서로 독립이고 베르누이분포를 따르는 확률변수들을 관측하는 것을 뜻한다.
- '1'과 '0'을 각각 '성공'과 '실패'로 부르고, 이항분포를 서로 독립인 Bernoulli trial에서 나오는 성공횟수의 분포라고도 한다.

binomial distribution의 대의적 정의

$$X \sim B(n, p) \Leftrightarrow X \stackrel{d}{\equiv} Z_1 + \cdots + Z_n, Z_i \stackrel{iid}{\sim} Bernoulli(p)(i = 1, \cdots, n)$$

- Bernoulli trial적용한 이항분포의 정의이다.

binomial distribution의 성질

1. $X \sim B(n, p)$이면 그 moment generating function은 다음과 같다.

$$mgf_X(t) = (pe^t + q)^n, -\infty < t < +\infty$$

2. $X_1 \sim B(n, p), X_2 \sim B(n_x, p)$이고 X_1, X_2가 서로 독립이면 다음관계가 성립한다.

$$X_1 + X_2 \sim B(n_1 + n_2, p)$$

multinomial distribution

> 다항분포, 기호: $X = (X_1, X_2, \cdots, X_k)^t \sim Multi(n, (p_1, p_2, \cdots, p_k)^t)$

모집단을 구성하는 각 개체의 특성이 3개 이상의 유형으로 분리되는 경우, 각 유형의 비율이 $p_1, p_2, \cdots, p_k$일 때를 생각해보자.
이때 동일 확률로 복원추출한 n개의 표본의 유형의 개수를 $X_1, X_2, \cdots, X_k$라고 하면 $X = (X_1, X_2, \cdots, X_k)^t$의 결합확률 밀도는 다음과 같고, 이 분포를 mutinomial distribution이라고 한다.

$$f(x_1, x_2, \cdots, x_k) = \binom{n}{x_1 x_2 \cdots x_k} p_1^{x_1} p_2^{x2} \cdots p_k^{x_k}$$
$$x_i = 0, \cdots, n(i = 1, 2, \cdots, k) x_1 + x_2 + \cdots + x_k = n$$

multinomial trial

> 다항시행

여러 가지 유형으로 분류되는 모집단에서 한 개씩 추출하여 관측하는 것.

- 복원 추출하는 것은 서로 독립인 multinomial trial을 뜻한다.

$$X = (X_1, X_2, \cdots, X_k)^t \sim Multi(n, (p_1, p_2, \cdots, p_k)^t)$$
$$\Leftrightarrow X \overset{d}{\equiv} Z_1 + \cdots + Z_n, Z_i = (Z_{i1}, Z_{i2}, \cdots, Z_{ik})^t \overset{iid}{\sim} Multi(1, (p_1, p_2, \cdots, p_k)^t)(i = 1, \cdots, n)$$

multinomial distribution의 성질

1. $X = (X_1, X_2, \cdots, X_k)^t \sim Multi(n, (p_1, p_2, \cdots, p_k)^t)$이면

$$E(X_i) = np_l (l = 1, \cdots, k)$$
$$Var(X_l) = np_l(1 - p_l), Cov(X_l) = -np_l p_m (l \neq m, l, m = 1, \cdots, k)$$

2. $X = (X_1, X_2, \cdots, X_k)^t \sim Multi(n, (p_1, p_2, \cdots, p_k)^t)$이면 moment generating function 은

$$mgf_X(t) = (p_1 e^{t_1} + \cdots + p_k e^{t_k})^n,\ -\infty < t_1 < +\infty\ (l = 1, \cdots, k)$$

geometric distribution

geometric distribution의 일반적 정의

> 기하분포, 기호: $W_1 \sim Geo(p)$

서로 독립이고 성공률이 p인 bernoulli trial $X_1, \cdots, X_n, \cdots$을 관측할 때, 첫번째 성공을 관측할 때까지의 시행횟수를 W_1라고 하면 이때의 확률은 다음과 같고, 이러한 확률분포를 geometric distribution 이라고 한다.

$$P(W_1 = x) = (1-p)^{x-1}p, x = 1, 2, \cdots$$

geometric distribution의 성질

1. $W_1 \sim Geo(p)$이면 그 moment generating function은

$$mgf_{W_1}(t) = (1-qe^t)^{-1}e^tp, \;\; t < -\log q \; (q = 1-p)$$

2. $W_1 \sim Geo(p)$ 이면

$$E(W_1) = 1/p, Var(w_1) = q/p^2 \; (q = 1-p)$$

negative binomial distribution

negative binomial distribution의 일반적 정의

> 음이항분포, 기호: $W_r \sim Negbin(r, p)$

서로 독립이고 성공률이 p인 Bernoulli trial $X_1, \cdots, X_n, \cdots$을 과나측할 때 r번째 성공까지의 시행횟수를 W_r이라고 하면 다음이 성립하고, 이 분포를 negative binomial distribution이라고 한다.

$$\begin{aligned} P(W_r = x) &= \binom{x-1}{r-1}p^{r-1}(1-p)^{(x-1)-r-1}p \\ &= \binom{x-1}{r-1}p^r(1-p)^x - r, \;\; x = r, r+1, \cdots \end{aligned}$$

negative binomial distribution의 대의적 정의

- 아이디어

$$\begin{aligned} &P(W_1 = x_1, W_2 - W_1 = x_2, \cdots, W_r - W_{r-1} = x_r) \\ &P(연속된(x_i)번의실패후성공, i = 1, \cdots, r) \\ &= (1-p)^{x_i-1}p(1-p)^{x_2-1}p\cdots(1-p)^{x_r-1}p, \; x_i = 1, 2, \cdots \; (i = 1, \cdots, r) \end{aligned}$$

- 정의

$$X \sim Negbin(r, p) \Leftrightarrow X \overset{d}{\equiv} Z_1 + \cdots + Z_r, Z_i \overset{iid}{\sim} Geo(p)(i = 1, \cdots, r)$$

negative binomial distribution의 성질

1. $X \sim Negbin(r, p)$이면 $E(X) = r/p, \; Var(X) = rq/p^2$
2. $X \sim Negbin(r, p)$이면 moment generating function은

$$mgf_X(t) = \{pe^t(1-qe^t)^{-1}\}^r, t < -\log q \ (q = 1-p)$$

3. $X_1 \sim Negbin(r_1, p) X_2 \sim Negbin(r_2, p)$이고 X_1, X_2가 서로 독립이면

$$X_1 + X_2 \sim Negbin(r_1 + r_2, p)$$

Poisson distribution

Poisson approximation

> 푸아송 근사

binomial distribution $B(n,p)$에서 시행횟수 n이 크고 확률 p가 작은 경우는 아래 식1과 같은 근사식이 성립하고,
이를 아래식 2로 나타낸 것을 binomial probability의 **Poisson approximation** 이라고 한다.

1. $$\begin{aligned} \binom{n}{x} p^x (1-p)^{n-x} &= n(n-1)\cdots(n-x+1)p^x(1-p)^{n-x}/x! \\ &= \frac{n}{n}(1-\frac{1}{n})\cdots(1-\frac{x-1}{n})(np)^x(1-\frac{np}{n})^{n-x}/x! \\ &= (np)^x e^{-np}/x! \end{aligned}$$
2. $$\lim_{\substack{n\to\infty \\ np_n \to \lambda}} \binom{n}{x} p_n^x (1-p_n)^{n-x} = e^{-\lambda}/x! \ (\lambda > 0, \lambda = np)$$

Poisson distribution's probability density function

> 푸아송 분포의 확률밀도함수, 기호: $X \sim Poisson(\lambda)$

$$f(x) = e^{-lambda} \lambda^x / x!, \ x = 0, 1, 2, \cdots \ \ (\lambda > 0)$$

Poisson distribution의 성질

1. $X \sim Poisson(\lambda)$ 이면 그 moment generating function은

$$mgf_X(t) = e^{-\lambda + \lambda e^t}, -\infty < t < +\infty$$

2. $X \sim Poisson(\lambda)$ 이면

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

3. $X_1 \sim Poisson(\lambda_1), X_2 \sim Poisson(\lambda_2)$이고 X_1, X_2가 서로 독립이면

$$X_1 + X_2 \sim Poisson(\lambda_1 + \lambda_2)$$

Poisson process

> 푸아송 과정

시각 0에서 t까지 특정한 현상이 발생하는 횟수를 N_t이라고 할 때,
다음의 조건들이 만족되면 $\{N_t : t \geq 0\}$를 occurrence rate(발생률) λ인 **Poisson process** (포아송 과정)이라고 한다.

1. Stationarity(정상성)
   현상이 발생하는 횟수의 분포는 시작하는 시각에 관계없다.
   즉, N_t의 분포와 $N_{s+t} - N_s$의 분포가 같고, $N_0 = 0$이다.
2. Independent Increment(독립증분성)
   시각 0부터 t까지 현상이 발생하는 횟수와 시각 t 후부터 $1 + h(h > 0)$까지 발생하는 횟수는 서로 독립이다.
   즉, N_t와 $N_{t+h} - N_t$는 서로 독립니다.
3. Proportionality(비례성)
   짧은 시간 동안에 현상이 한 번 발생할 확률은 시간에 비례한다.
   즉, $P = (N_h = 1) = \lambda h + o(h), h \to 0$ 이 성립한다.
   위에서 λ는 양수의 비례상수이며 $o(h)$의 의미는 $\lim_{h\to 0} o(h)/h = 0$을 말한다.
4. Rareness(희귀성)
   짧은 시간 동안에 현상이 두 번 이상 발생할 확률은 매우 작다.
   즉, $P(N_h \geq 2) = o(h), h \to 0$

Poisson process 에서 발생횟수의 분포

occurrence rate이 λ인 Poisson process $\{N_t : t \geq 0\}$에서 시각 t까지 발생횟수 N_t의 분포는 평균이 λt인 Poisson distribution이다.

$$N_t \sim Poisson(\lambda t)$$

exponential distribution

exponential distribution의 정의

> 지수분포, 기호:$W_1 \sim Exp(1/\lambda)$ $(\lambda > 0)$

occurrence rate이 λ인 Poisson process $N_t : t \geq 0$에서 첫 번째 현상이 시각 t후에 발생한다고 하자.
이때 까지의 시간을 W_1라고 하면 다음과 같은 등식이 성립하고

$$(W_1 > t) = (N_t = 0)$$

이를 이용해 다음과 같은 cumulative distribution funciton을 구할 수 있다.

$$P(W_1 > t) = P(N_t = 0) = e^{-\lambda t}, t \geq 0$$
$$P(W_1 \leq t) = \begin{cases} 1 - e^{\lambda t}, t \geq 0 \\ 0, \qquad t < 0 \end{cases}$$

따라서 첫 번째 현상이 발생할 때 까지의 시간 W_1의 확률밀도함수는 다음과 같고, 이를 **exponential distribution** 이라고 한다.

$$f(x) = \lambda e^{-lambdax} I_{(x \geq 0)}$$

exponential distribution의 성질

1. $W_1 \sim Exp(1/\lambda)\ (\lambda > 0)$이면 그 moment generating funciton은

$$mgf_{W_1}(t) = (1 - t/\lambda)$$

2. $W_1 \sim Exp(1/\lambda)\ (\lambda > 0)$이면

$$E(W_1) = 1/\lambda,\ Var(W_1) = 1.\lambda^2$$

gamma distribution

gamma distribution의 정의

> 감마분포

occurrence rate이 λ인 Poisson process $N_t : t \geq 0$ 에서 r 번째 현상이 시각 t후에 발생한다고 하자. 이때까지의 시간을 W_r라하면 다음 관계가 성립한다.

$$(W_r > t) = (N_r \leq r - 1)$$
$$P(W_r \leq t) = 1 - P(W_r > t) = 1 - P(N_t \leq r - 1) = 1 - \sum_{k=0}^{r-1} e^{-\lambda t} (\lambda t)^k / k!$$
$$t \geq 0$$

이로부터 양변을 미분하여 W_r의 probability density function을 구하면 다음과 같고, 이를 gamma distribution이라고 부른다.

$$pdf_{W_r}(t) = \frac{d}{dt} cdf_{W_r}(t)$$
$$= -\sum_{k=0}^{r-1} \{(-\lambda) e^{\lambda t} (\lambda t)^k / k! +$$
$$e^{-\lambda t} k \lambda (\lambda t)^{k-1}\}$$
$$= \lambda e^{-\lambda t} \{\sum_{k=0}^{r-1} (\lambda t)^k / k! -$$
$$\sum_{k=1}^{r-1} (\lambda t)^{k-1} / (k-1)!\}$$
$$= \lambda^r t^{r-1} e^{-\lambda t} / (r-1)!,\ t > 0$$

shape parameter

> 형상모수, $\Gamma(\alpha)$

위의 감마분포

$$\lambda^r t^{r-1} e^{-\lambda t} / (r-1)!,\ t > 0$$

에서 r의 값에 따라서 분포의 형태가 바뀐다. 따라서 r을 shape parameter라고 부른다.
일반적으로는 양수일 수 있으며, 이러한 경우에 흔히 α로 나타내기도 한다.
이와 관련해

$$\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1}e^{-x}dx,\ \alpha > 0$$

처럼 정의되는 감마함수를 이용해 $Gamma(\alpha, (\beta))$분포의 pdf를 다음과 같이 나타낸다.

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

- 위의 β는 scale parameter(척도모수)라고 하며 occurrence rate의 역수인 $1/\lambda$이다.

gamma distribution의 성질

1. $X \sim Gamma(\alpha, \beta)$이면

$$E(X) = \alpha\beta, Var(X) = \alpha\beta^2$$

2. $X \sim Gamma(\alpha, \beta)$ 이면 그 moment generating function은

$$mgf_X(t) = (1-\beta t)^{-alpha}, t < 1/\beta$$

3. $X_1 \sim Gamma(\alpha_1, \beta), X_2 \sim Gamma(\alpha_2, \beta)$이고 X_1, X_2가 서로 독립이면

$$X_1 + X_2 \sim Gamma(\alpha_1 + \alpha_2, \beta)$$

gamma distribution의 대의적 정의

exponential distribution은 shape parameter가 1인 gamma distribution이다.
이것으로부터 gamma distribution을 다음처럼 정의 할 수 있다.

shape parameter r이 자연수인 경우에

$$X \sim Gamma(r, \beta) \Leftrightarrow X \stackrel{d}{\equiv} Z_1 + \cdots + Z_r, Z_i \stackrel{iid}{\sim} Exp(\beta)$$

normal distribution

standard normal distribution

> 표준정규분포

binomial distribution의 cumulative probability를 적분으로 나타내는 근사식은 아래와 같다.

$$\sum_{x: a \le \frac{x-np}{\sqrt{np(1-p)}} \le b} \binom{n}{x} p^x (1-p)^{n-x} \sim \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz,\ \ n \to \infty$$

이때, 아래의 함수는 그 적분 값이 1이 되는 함수로서 standard normal distribution의 pdf라고 한다.

$$\phi(z) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}, \ -\infty < z < +\infty$$

normal distribution

> 정규분포, $N(\mu, \sigma^2)$

일반적으로 아래의 식을 normal distribution의 pdf라고 한다.

$$\frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma}) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}, \ -\infty < x < +\infty$$
$$(\mu\text{는 실수}, \ \sigma\text{는 양수})$$

normal distribution의 성질

1. $X \sim N(\mu, \sigma^2)$ 이면

$$E(X) = \mu, Var(X) = \sigma^2$$

2. $X \sim N(\mu, \sigma^2)$이면 그 mgf는

$$mgf_X(t) = e^{\mu t + \frac{1}{2}\sigma^2 t^2}, \ -\infty < t < +\infty$$

3. $X_1 \sim N(\mu_1, \sigma_1^2), X_2 \sim N(\mu_2, \sigma_2^2)$이고, X_1, X_2가 서로 독립이면

$$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

normal distribution의 대의적정의

1. $X \sim N(\mu, \sigma^2)$ 이면 상수 a, b에 대하여

$$aX + b \sim N(a\mu + b, a^2\sigma^2)$$

2. $X \sim N(\mu, \sigma^2) \Leftrightarrow \frac{X-\mu}{\sigma} \sim N(0, 1) \Leftrightarrow X \overset{d}{\equiv} \sigma Z + \mu, \ Z \sim N(0, 1)$

normal distribution의 cumulative probability

> 정규분포의 누적확률

아래와 같은 standard normal distribution의 cumulative distribution function는 아래와 같다.

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}}e^{-e^2/2}dz$$

이때, 일반적인 normal distribution $N(\mu, \sigma^2)$를 따르는 X의 cumulative distribution은 다음과 같다.

$$P(X \leq x) = P(\frac{X-\mu}{\sigma} \leq \frac{x-\mu}{\sigma}) = \Phi(\frac{x-\mu}{\sigma})$$

###quantile

> 분위수

$Z \sim N(0,1)$ 일때 아래의 식을 만족하는 값 z_α를
standard normal distribution의 $upper\ \alpha\ quantile$이라고 한다.

$$P(Z > z_\alpha) = \alpha(0 < \alpha < 1)$$

기타 필요 정의

모집단

> population

통계 조사에서 관심의 대상이 N개의 개체일 때 이들 중에서 n개를 '랜덤'하게 택하여 조사한 후 전체에 대한 추측을 한다고 하자.
이때, 조사와 추측의 대상이 되는 전체를 모집단이라한다.

비복원추출

> sampling without replacement

축자적으로 한개씩 동일한 확률로 뽑아나가며 한 번 뽑힌 것은 되돌려 넣지 않는 추출 방법

단순랜덤추출

> sample random sampling

N개의 개체로 구성된 모집단에서 '랜덤'하게 n개를 비복원추출방식으로 추출하는 것

랜덤 표본

> random sample

단순랜덤추출 로 추출된 n개를 지칭한다. 간단히 표본(sample)이라고도 한다.

모비율

> population proportion

각 객체의 특성에 대한 분류를 0또는 1의 두가지 분류로 나타낼 때, 조사와 추측 대상이 되는 전체에 0과 1이 각각 $N-D$개, D개 있다고 하자. 이 때 1의 비율인 $p=D/N$를 모비율이라고 한다. 이때 모집단분포는 1과 0에 각각 p와 $1-p$를 대응시키는 분포이다.

간단히 말해서 전체 N개중에 값이 1(혹은 참)인 D개의 비율 D/N을 모비율이라고 한다.

$X \overset{d}{\equiv} Y$의 의미

확률변수 X와 Y가 같은 분포를 갖는다.

iid의 의미

> independent and identically distributed

위 영어 문장의 약어로, 서로 독립이고 같은 분포를 같는다는 뜻이다.