3장. 여러 가지 확률분포

3.1 초기하분포

모집단(母集團 population):조사와 추측의 대상이 되는 전체

모집단분포(母集團分布 population distribution):전체가 흩어져 있는 상황을 묘사하는 분포

비복원추출(sampling without replacement):
N 개의 개체로 구성된 모집단에서 축차적으로 한 개씩 동일한 확률로 뽑아나가며 한 번 뽑힌 것은 되돌려 넣지 않고 n개를 추출하는 방법으로서 **단순랜덤추출(單純랜덤抽出 simple random sampling)**이라고도 하며, 추출된 n개를 **랜덤표본**(random sample) 또는 간단히 **표본**(標本 sample)이라고 한다.

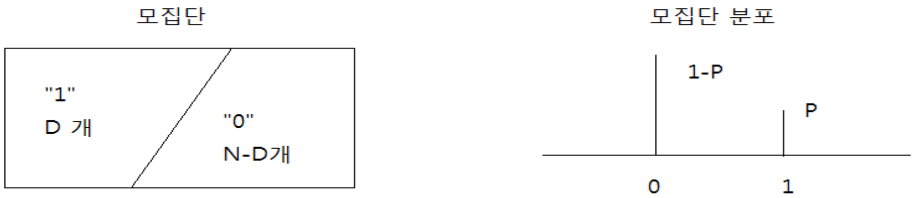


그림 3.1.1 두 가지 분류의 모집단과 모집단 분포($p = D/N$: 모비율)

초기하분포(超幾何分布 hypergeometric distribution): $\mathrm{X} \sim \mathrm{H}(n;N,D)$
  n개를 단순랜덤추출하여 얻은 표본 중 '1'의 개수를 X라고 하면

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

<u>초기하분포 $\mathrm{H}(n;N,D)$의 평균:</u>

$$\begin{aligned}\mathrm{E}(\mathrm{X}) &= \sum_x x\binom{D}{x}\binom{N-D}{n-x}\Big/\binom{N}{n}\\ &= D\sum_x \binom{D-1}{x-1}\binom{N-1-(D-1)}{n-1-(x-1)}\Big/\binom{N}{n}\\ &= D\binom{N-1}{n-1}\Big/\binom{N}{n}\quad (\because (1+t)^{D-1}(1+t)^{N-1-(D-1)} = (1+t)^{N-1})\\ &= nD/N\end{aligned}$$

<u>초기하분포 $\mathrm{H}(n;N,D)$의 분산:</u> 같은 방법으로

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

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

초기하분포의 평균과 분산:

$\mathrm{X} \sim \mathrm{H}(n;N,D)$일 때, 모비율을 $p = D/N$라고 하면

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

예 3.1.1 (초기하분포 적용의 예)

초기하분포의 근사 계산: $\mathrm{H}(n;N,D)$에서 $N \gg n$일 때

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

-----복원추출(復元抽出 sampling with replacement)에 의한 확률

-----모집단 크기 N이 커짐에 따라 비복원추출의 효과가 없어짐

3.2 이항분포와 다항분포

이항분포(이항분포 binomial distribution): $X \sim B(n,p)$

각 개체가 ‘0’ 또는 ‘1’의 두 가지로 분류되어 있고 ‘1’의 비율이 p인 모집단에서 한 개씩 동일한 확률로 뽑아 나가며 복원추출에 의해 뽑은 n개 중에서 ‘1’의 개수를 X라고 하면 $pdf(x) = \binom{n}{x} p^x (1-p)^{n-x}$, $x = 0,1,\cdots,n$

이항분포 $B(n,p)$의 평균:

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

이항분포 $B(n,p)$의 분산: 같은 방법으로

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

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

이항분포의 평균과 분산:

$X \sim B(n,p)$이면

$$E(X) = np, \quad Var(X) = np(1-p)$$

베르누이시행(Bernoulli trial): $Z_i \overset{iid}{\sim} \text{Bernoulli}(p)$

한 개씩 복원추출하는 경우에 한 개씩의 추출 결과를 $Z_1, Z_2, \cdots, Z_n, \cdots$ 이라고 하면 이들은 서로 독립이고 각각의 분포는

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

로서 동일하다.

이항분포의 대의적 정의(代意的 定義 representational definition)[1]:

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

---

1) $X \overset{d}{\equiv} Y$는 ‘확률변수 X와 Y가 같은 분포를 갖는다’ 는 뜻이며, iid는 independent and identically distributed에서 앞 글자를 따서 표기한 것으로 ‘서로 독립이고 같은 분포를 갖는다’ 는 뜻이다.

정리 3.2.1: 이항분포의 성질

(a) $X \sim B(n,p)$ 이면 그 적률생성함수는

$$mgf_X(t) = (pe^t + q)^n,\ -\infty < t < +\infty \ \ (q = 1-p)$$

(b) $X_1 \sim B(n_1,p), X_2 \sim B(n_2,p)$이고 X_1, X_2가 서로 독립이면

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

[KEY] (a) 이항분포의 대의적 정의와 독립인 확률변수의 합에 관한 정리 2.26으로부터

$$mgf_X(t) = mgf_{Z_1}(t) \cdots mgf_{Z_n}(t),\ Z_i \overset{\text{iid}}{\sim} \text{Bernoulli}(p)(i=1,\cdots,n)$$

$$mgf_{Z_i}(t) = E(e^{tZ_i}) = e^{t\cdot 1}P(Z_i = 1) + e^{t\cdot 0}P(Z_i = 0) = pe^t + q \quad (i = 1,\cdots,n)$$

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

(b) 서로 독립인 확률변수의 합에 관한 정리 2.26과 (a)로부터

$$mgf_{X_1+X_2}(t) = mgf_{X_1}(t) mgf_{X_2}(t) = (pe^t + q)^{n_1+n_2},\ -\infty < t < +\infty$$

적률생성함수의 분포 결정성으로부터

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

예 3.2.1 (베르누이 확률변수들을 이용한 초기하분포의 대의적 정의)

다항분포(multinomial distribution)[2]**:** $\mathrm{X}=(\mathrm{X}_1,\mathrm{X}_2,\cdots,\mathrm{X}_k)^t \sim \mathrm{Multi}(n,(p_1,p_2,\cdots,p_k)^t)$

각 개체가 세 가지 이상의 유형으로 분류되고 각 유형의 비율이 $p_1,p_2,\cdots,p_k$인 모집단에서한 개씩 동일한 확률로 뽑아 나가며 <u>복원추출한</u> n개의 랜덤표본에 있는 각 유형의 개수를 $\mathrm{X}_1,\mathrm{X}_2,\cdots,\mathrm{X}_k$라고 하면[3]

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

다항시행(多項試行 multinomial trial): $\mathrm{Z}_i \overset{iid}{\sim} \mathrm{Multi}(1,(p_1,p_2,\cdots,p_k)^t)$

여러 가지 유형으로 분류되는 모집단에서 한 개씩 복원추출한 결과를 $\mathrm{Z}_i=(\mathrm{Z}_{i1},\mathrm{Z}_{i2},\cdots,\mathrm{Z}_{ik})^t$ $(i=1,\cdots,n)$이라고 하면 이들은 서로 독립이고 각각의 분포는

$$\mathrm{P}(\mathrm{Z}_i=(1,0,\cdots,0))=p_1,\cdots,\mathrm{P}(\mathrm{Z}_i=(0,\cdots,0,1))=p_k\ \ (i=1,\cdots,n)$$

로서 동일하다. 그런데 이 분포를 하나의 식으로 나타내면

$$\mathrm{P}(\mathrm{Z}_{i1}=z_1,\mathrm{Z}_{i2}=z_2,\cdots,\mathrm{Z}_{ik}=z_k)=p_1^{z_1}p_2^{z_2}\cdots p_k^{z_k},\, z_i=0,1\,(i=1,\cdots,k), z_1+\cdots+z_k=1$$

다항분포의 대의적 정의:

$$\begin{aligned}&\mathrm{X}=(\mathrm{X}_1,\mathrm{X}_2,\cdots,\mathrm{X}_k)^t \sim \mathrm{Multi}(n,(p_1,p_2,\cdots,p_k)^t)\\ \Leftrightarrow\ &\mathrm{X}\overset{d}{\equiv}\mathrm{Z}_1+\cdots+\mathrm{Z}_n, \mathrm{Z}_i=(\mathrm{Z}_{i1},\mathrm{Z}_{i2},\cdots,\mathrm{Z}_{ik})^t \overset{iid}{\sim} \mathrm{Multi}(1,(p_1,p_2,\cdots,p_k)^t)\,(i=1,\cdots,n)\end{aligned}$$

정리 3.2.2:다항분포의 성질

(a) $\mathrm{X}=(\mathrm{X}_1,\mathrm{X}_2,\cdots,\mathrm{X}_k)^t \sim \mathrm{Multi}(n,(p_1,p_2,\cdots,p_k)^t)$이면

$$\mathrm{E}(\mathrm{X}_l)=np_l\ \ (l=1,\cdots,k)$$
$$\mathrm{Var}(\mathrm{X}_l)=np_l(1-p_l),\mathrm{Cov}(\mathrm{X}_l,\mathrm{X}_m)=-np_lp_m\,(l\neq m, l,m=1,\cdots,k)$$

(b) $\mathrm{X}=(\mathrm{X}_1,\mathrm{X}_2,\cdots,\mathrm{X}_k)^t \sim \mathrm{Multi}(n,(p_1,p_2,\cdots,p_k)^t)$이면 그 적률생성함수는

$$mgf_{\mathrm{X}}(t)=(p_1e^{t_1}+\cdots+p_ke^{t_k})^n,\ \ -\infty<t_l<+\infty\ (l=1,\cdots,k)$$

[KEY] 다항분포의 대의적 정의로부터

$$\mathrm{E}(\mathrm{X})=\mathrm{E}(\mathrm{Z}_1+\cdots+\mathrm{Z}_n)=n\mathrm{E}(\mathrm{Z}_1),\ \ \mathrm{Var}(\mathrm{X})=\mathrm{Var}(\mathrm{Z}_1+\cdots+\mathrm{Z}_n)=n\mathrm{Var}(\mathrm{Z}_1)$$
$$mgf_{\mathrm{X}}(t)=mgf_{\mathrm{Z}_1+\cdots\,\mathrm{Z}_n}(t)=(mgf_{\mathrm{Z}_1}(t))^n$$

한편, $Z_1=(Z_{11},Z_{12},\cdots,Z_{1k})^t \sim \mathrm{Multi}(1,(p_1,p_2,\cdots,p_k)^t)$ 이므로

$$\mathrm{P}(\mathrm{Z}_{1l}=1)=p_l,\mathrm{P}(\mathrm{Z}_{1l}=0)=1-p_l,\mathrm{Z}_{1l}\mathrm{Z}_{1m}=0\ \ (l\neq m, l,m=1,\cdots,k)$$
$$\therefore\begin{cases}\mathrm{E}(\mathrm{Z}_{1l})=p_l & (l=1,\cdots,k)\\ \mathrm{Cov}(\mathrm{Z}_{1l},\mathrm{Z}_{1m})=\mathrm{E}(\mathrm{Z}_{1l}\mathrm{Z}_{1m})-\mathrm{E}(\mathrm{Z}_{1l})\mathrm{E}(\mathrm{Z}_{1m})=-p_lp_m & (l\neq m, l,m=1,\cdots,k)\end{cases}$$
$$mgf_{\mathrm{Z}_1}(t)=\mathrm{E}(e^{t_1\mathrm{Z}_{11}+\cdots+t_k\mathrm{Z}_{1k}})=p_1e^{t_1}+\cdots p_ke^{t_k}$$

---

2) $k=2$인 경우의 다항분포 $\mathrm{Multi}(n,(p,1-p)^t)$는 이항분포 $\mathrm{B}(n,p)$를 따르는 확률변수 X에 대하여 X와 $n-\mathrm{X}$의 결합분포를 뜻하고 있는 것이다. 즉 $(\mathrm{X},n-\mathrm{X})^t \sim \mathrm{Multi}(n,(p,1-p)^t)$

3) $\binom{n}{x_1\,x_2\,\cdots\,x_k}$는 서로 다른 유형의 $x_1,x_2,\cdots,x_k$개를 나열하는 방법의 수인 $\dfrac{n!}{x_1!x_2!\cdots x_k!}$를 나타내는 다항계수(多項係數 multinomial coefficient)이다.

3.3 기하분포와 음이항분포

기하분포[4](幾何分布 geometric distribution): $W_1 \sim \text{Geo}(p)$

서로 독립이고 성공률이 p 인 베르누이시행 $X_1, \cdots, X_n, \cdots$ 을 관측할 때, 첫번째 성공을 관측할 때까지의 시행횟수를 W_1이라고 하면,

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

정리 3.3.1:기하분포의 성질

(a) $W_1 \sim \text{Geo}(p)$이면 그 적률생성함수는

$$mgf_{W_1}(t) = (1-qe^t)^{-1}e^t p,\ t \leftarrow \log q \quad (q = 1-p)$$

(b) $W_1 \sim \text{Geo}(p)$이면

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

[증명] (a) 기하급수의 공식

(b) 누율생성함수이용:

$$\begin{aligned} cgf_{W_1}(t) &= -\log(1-qe^t) + t + \log p \\ &= t - \log\{p - q(e^t - 1)\} + \log p \\ &= t - \log\left\{1 - \frac{q}{p}(e^t - 1)\right\},\quad t \leftarrow \log q \end{aligned}$$

로그함수의 멱급수 전개식:

$$-\log(1-A) = A + A^2/2 + A^3/3 + \cdots,\ (A \simeq 0)$$

$$\therefore E(W_1) = c\dot{g}f_{W_1}(0) = 1/p,\ Var(W_1) = c\ddot{g}f_{W_1}(0) = q/p^2$$

음이항분포[5](陰二項分布 negative binomial distribution):$W_r \sim \text{Negbin}(r, p)$

서로 독립이고 성공률이 p인 베르누이시행 $X_1, \cdots, X_n, \cdots$ 을 관측할 때 r번째 성공까지의 시행횟수를 W_r이라고 하면

$$(W_r = x) = ((x-1)\text{번의 시행 중에서 } (r-1)\text{번의 성공 그리고 } x\text{번째 시행은 성공})$$

이므로

$$P(W_r = x) = \binom{x-1}{r-1}p^{r-1}(1-p)^{(x-1)-(r-1)}p = \binom{x-1}{r-1}p^r(1-p)^{x-r},\ x = r, r+1, \cdots$$

---

4) 무한급수 $\sum_{x=1}^{\infty} q^{x-1}$을 기하급수라고 부르는 데에서 기하분포의 이름이 유래되었다.

5) 음의 정수를 지수로 갖는 함수 $(1+t)^{-r}$의 다음과 같은 멱급수 전개식을 음이항전개식(negative binomial expansion)이라고 부르는 데에서 음이항분포의 명칭이 유래하였다.

$$(1+t)^{-r} = \sum_{k=0}^{\infty} \frac{(-r)(-r-1)\cdots(-r-k+1)}{k!}t^k = \sum_{k=0}^{\infty} \frac{(r+k-1)!}{k!(r-1)!}(-t)^k,\ -1 < t < 1$$

우변의 무한합에서 $x = r+k$로 치환하고 $-t = q = 1-p$를 대입하면 다음이 성립함을 알 수 있다.

$$p^{-r} = \sum_{x=r}^{\infty}\binom{x-1}{r-1}(1-p)^{x-r},\quad \text{즉} \quad \sum_{x=r}^{\infty}\binom{x-1}{r-1}p^r(1-p)^{x-r} = 1$$

한편, 각각의 성공 후 다음 성공까지의 시행횟수를 나타내는 $W_1, W_2-W_1, \cdots, W_r-W_{r-1}$의 결합 확률에 대하여 다음이 성립하는 것을 알 수 있다.

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

따라서 $W_1, W_2-W_1, \cdots, W_r-W_{r-1}$은 서로 독립이고 동일한 기하분포를 따른다. 즉

$$W_1, W_2-W_1, \cdots, W_r-W_{r-1} \overset{iid}{\sim} \mathrm{Geo}(p)$$

그러므로 이들의 합인 W_r의 분포인 음이항분포를 다음과 같이 정의할 수 있다.

음이항분포의 대의적 정의:

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

정리 3.3.2: 음이항분포의 성질

(a) $X \sim \mathrm{Negbin}(r,p)$이면 $E(X)=r/p$, $\mathrm{Var}(X)=rq/p^2$

(b) $X \sim \mathrm{Negbin}(r,p)$ 이면 그 적률생성함수는

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

(c) $X_1 \sim \mathrm{Negbin}(r_1,p), X_2 \sim \mathrm{Negbin}(r_2,p)$이고 X_1, X_2가 서로 독립이면

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

3.4 포아송분포

이항확률의 포아송(Poisson) 근사[6]:

시행횟수 n이 크고 비율 p가 작은 경우에

$$\binom{n}{x}p^x(1-p)^{n-x} = n(n-1)\cdots(n-x+1)p^x(1-p)^{n-x}/x!$$
$$= \frac{n}{n}(1-\frac{1}{n})\cdots(1-\frac{x-1}{n})(np)^x(1-\frac{np}{n})^{n-x}/x!$$
$$\simeq (np)^x e^{-np}/x!$$

$$\lim_{\substack{n\to\infty \\ np_n\to\lambda}} \binom{n}{x}p_n^x(1-p_n)^{n-x} = e^{-\lambda}\lambda^x/x! \qquad (\lambda>0)$$

예 3.4.1(이항확률의 포아송 근사 예)

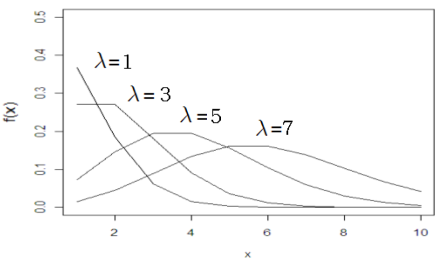
지수함수의 멱급수 전개: $e^a = 1+\sum_{n=1}^{\infty}\frac{1}{n!}a^n, \ -\infty < a < +\infty$

포아송분포: $X \sim \text{Poisson}(\lambda)$

이항확률의 근사 계산에서 주어지는 확률밀도함수

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

패키지 R을 이용하여 그린 포아송분포의 형태 (부록 III 참조):



정리 3.4.1:포아송분포의 성질

(a) $X \sim \text{Poisson}(\lambda)$이면 그 적률생성함수는

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

(b) $X \sim \text{Poisson}(\lambda)$이면

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

(c) $X_1 \sim \text{Poisson}(\lambda_1), X_2 \sim \text{Poisson}(\lambda_2)$이고 X_1, X_2가 서로 독립이면

$$X_1+X_2 \sim \text{Poisson}(\lambda_1+\lambda_2)$$

---

6) 이 근사 계산의 과정에서는 다음과 같은 극한 값 계산 결과를 이용하고 있다.

$$\lim_{\substack{n\to\infty \\ a_n\to a}}(1+\frac{a_n}{n})^n = e^a \quad \text{즉} \quad \lim_{\substack{n\to\infty \\ a_n\to a}} n\log(1+\frac{a_n}{n}) = a$$

[증명] (a) 지수함수의 멱급수 전개
(b)누율생성함수의 이용

$$cgf_{\mathrm{X}}(t) = -\lambda + \lambda e^t = \lambda t + \frac{\lambda}{2!}t^2 + \frac{\lambda}{3!}t^3 + \cdots$$

$$\therefore\ \mathrm{E}(\mathrm{X}) = \dot{cgf}_{\mathrm{X}}(0) = \lambda,\ \mathrm{Var}(\mathrm{X}) = \ddot{cgf}_{\mathrm{X}}(0) = \lambda$$

(c) $mgf_{\mathrm{X}_1+\mathrm{X}_2}(t) = mgf_{\mathrm{X}_1}(t)mgf_{\mathrm{X}_2}(t) = e^{-(\lambda_1+\lambda_2)+(\lambda_1+\lambda_2)e^t},\ -\infty < t < +\infty$
적률생성함수의 분포 결정성:

$$\mathrm{X}_1 + \mathrm{X}_2 \sim \mathrm{Poisson}(\lambda_1 + \lambda_2)$$

포아송과정(Poisson process):
시각 0에서 t까지 특정한 현상이 발생하는 횟수를 N_t라고 할 때, 다음의 조건들이 만족되면 $\{\mathrm{N}_t : t \ge 0\}$를 발생률(occurrence rate) λ인 포아송과정(Poisson process)이라고 한다.
(a)(정상성 stationarity) 현상이 발생하는 횟수의 분포는 시작하는 시각에 관계없다. 즉 N_t의 분포와 $\mathrm{N}_{s+t} - \mathrm{N}_s$의 분포가 같고, $\mathrm{N}_0 = 0$ 이다.
(b)(독립증분성 independent increment) 시각 0부터 t까지 현상이 발생하는 횟수와 시각 t 후부터 $t+h(h>0)$까지 발생하는 횟수는 서로 독립이다. 즉 N_t와 $\mathrm{N}_{t+h} - \mathrm{N}_t$는 서로 독립이다.
(c)(비례성 proportionality) 짧은 시간 동안에 현상이 한 번 발생할 확률은 시간에 비례한다. 즉

$$\mathrm{P}(\mathrm{N}_h = 1) = \lambda h + o(h),\ \ h \to 0$$

여기에서 λ는 양수의 비례상수이고, $o(h)$의 의미는 $\lim_{h\to 0} o(h)/h = 0$임을 뜻한다.
(d)(희귀성 rareness) 짧은 시간 동안에 현상이 두 번 이상 발생할 확률은 매우 작다. 즉
$\mathrm{P}(\mathrm{N}_h \ge 2) = o(h),\ \ h \to 0$

정리 3.4.2:포아송과정에서 발생횟수의 분포
발생률이 λ인 포아송과정 $\{\mathrm{N}_t : t \ge 0\}$에서 시각 t까지 발생횟수 N_t의 분포는 평균이 λt인 포아송분포이다. 즉

$$\mathrm{N}_t \sim \mathrm{Poisson}(\lambda t)$$

[증명] 포아송과정의 독립증분성, 정상성, 희귀성으로부터

$$\begin{aligned}
&\mathrm{P}(\mathrm{N}_{t+h} = x) \\
&= \mathrm{P}(\mathrm{N}_{t+h} = x, \mathrm{N}_t = x) + \mathrm{P}(\mathrm{N}_{t+h} = x, \mathrm{N}_t = x-1) + \mathrm{P}(\mathrm{N}_{t+h} = x, \mathrm{N}_t \le x-2) \\
&= \mathrm{P}(\mathrm{N}_t = x, \mathrm{N}_{t+h} - \mathrm{N}_t = 0) + \mathrm{P}(\mathrm{N}_t = x-1, \mathrm{N}_{t+h} - \mathrm{N}_t = 1) + \mathrm{P}(\mathrm{N}_{t+h} = x, \mathrm{N}_{t+h} - \mathrm{N}_t \ge 2) \\
&= \mathrm{P}(\mathrm{N}_t = x)\mathrm{P}(\mathrm{N}_h = 0) + \mathrm{P}(\mathrm{N}_t = x-1)\mathrm{P}(\mathrm{N}_h = 1) + \mathrm{o}(h),\ \ h \to 0
\end{aligned}$$

따라서 $g(x,t) = \mathrm{P}(\mathrm{N}_t = x)$ 라고 하면, 비례성과 희귀성으로부터 $h \to 0$ 일 때

$$g(x,t+h) = g(x,t)\{1 - (\lambda h + o(h)) - o(h)\} + g(x-1,t)\{\lambda h + o(h)\} + o(h)$$

임을 알 수 있다. 이로부터

$$\frac{d}{dt}g(x,t) = -\lambda g(x,t) + \lambda g(x-1,t)$$

$$\therefore \frac{d}{dt}\{e^{\lambda t}g(x,t)/\lambda^x\} = e^{\lambda t}g(x-1,t)/\lambda^{x-1}$$

한편 $\mathrm{N}_0 = 0$이므로

$$g(0,0) = 1, g(1,0) = g(2,0) = \cdots = 0$$

임을 이용하여 위의 미분방정식을 $x = 0, 1, \cdots$에 대하여 차례로 풀면

$$\mathrm{P}(\mathrm{N}_t = x) = g(x,t) = e^{-\lambda t}(\lambda t)^x / x!,\ \ x = 0, 1, \cdots$$

예 3.4.2(포아송 과정의 적용 예)

3.5 지수분포와 감마분포

지수분포(exponential distribution): $W_1 \sim \mathrm{Exp}(1/\lambda)\ (\lambda > 0)$

발생률이 λ인 포아송과정 $\{N_t : t \ge 0\}$에서 첫 번째 현상이 발생할 때까지의 시간을 W_1이라고 하면

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

$$P(W_1 > t) = P(N_t = 0) = e^{-\lambda t},\ t \ge 0$$

$$\begin{aligned} P(W_1 \le t) &= \begin{cases} 1 - e^{-\lambda t}, & t \ge 0 \\ 0, & t < 0 \end{cases} \\ &= \int_{-\infty}^{t} \lambda e^{-\lambda x} I_{(x \ge 0)} dx \end{aligned}$$

따라서 W_1의 확률밀도함수는

$$pdf(x) = \lambda e^{-\lambda x} I_{(\mathrm{x} \ge 0)}$$

정리 3.5.1:지수분포의 성질

(a) $W_1 \sim \mathrm{Exp}(1/\lambda)\ (\lambda > 0)$이면 그 적률생성함수는

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

(b) $W_1 \sim \mathrm{Exp}(1/\lambda)\ (\lambda > 0)$이면

$$E(W_1) = 1/\lambda,\ \ \mathrm{Var}(W_1) = 1/\lambda^2$$

[증명] (a) 지수함수의 이상 적분

(b) 누율생성함수의 이용

$$\begin{aligned} cgf_{W_1}(t) &= -\log(1 - t/\lambda),\ t/\lambda < 1 \\ &= t/\lambda + (t/\lambda)^2/2 + (t/\lambda)^3/3 + \cdots \end{aligned}$$

$$\therefore\ E(W_1) = c\dot{gf}_{W_1}(0) = 1/\lambda,\ \mathrm{Var}(W_1) = c\ddot{gf}_{W1}(0) = 1/\lambda^2$$

감마분포(gamma distribution): $W_r \sim \mathrm{Gamma}(r, 1/\lambda)$

발생률이 λ인 포아송과정 $\{N_t : t \ge 0\}$에서 r 번째 현상이 발생할 때까지의 시간을 W_r이라고 하면

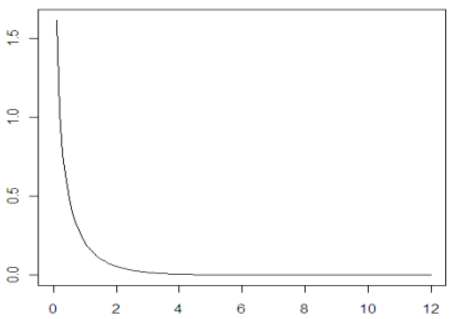
$$(\mathrm{W_r > t}) = (\mathrm{N_t \le r - 1})$$

$$P(W_r \le t) = 1 - P(W_r > t) = 1 - P(N_t \le r - 1) = 1 - \sum_{k=0}^{r-1} e^{-\lambda t} (\lambda t)^k / k!,\ \ t \ge 0$$

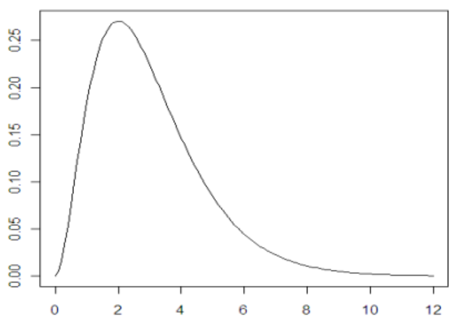
W_r의 확률밀도함수는

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

패키지 R을 이용하여 그린 감마분포 $\mathrm{Gamma}(r, 1/\lambda)$의 형태(부록 III 참조):



$r = 0.5, \lambda = 1$



$r = 3, \lambda = 1$

r; 형상모수(形狀母數 shape parameter)

발생률의 역수인 $\beta = 1/\lambda$; 척도모수(尺度母數 scale parameter)

부록 Ⅰ.6.2:감마함수

양수 α에 대하여

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

로 정의된 함수를 감마함수라고 하며 감마함수는 다음과 같은 성질을 갖는다.

(a) 임의의 양수 α 에 대하여 $\Gamma(\alpha)$는 실수이다.

(b) $\Gamma(\alpha) = (\alpha-1)\Gamma(\alpha-1),\ \alpha > 1$ 특히 $\Gamma(n) = (n-1)!, n = 1, 2, \cdots$

(c) $\Gamma(1/2) = \sqrt{\pi}$

[증명] (a) (ⅰ) $0 < \alpha < 1$ 일 때,

$$0 < \int_0^{+\infty} x^{\alpha-1} e^{-x} dx \le \int_0^1 x^{\alpha-1} dx + \int_1^{+\infty} e^{-x} dx = \frac{1}{\alpha} + e^{-1} < +\infty$$

(ⅱ) $\alpha \ge 1$일 때, α 보다 같거나 큰 최소의 자연수를 $\{\alpha\}$라고 하면 양수 x에 대하여

$$(x/2)^{\{\alpha\}-1} / (\{\alpha\}-1)! \le e^{x/2} = 1 + (x/2) + (x/2)^2/2! + \cdots$$

이므로

$$0 < \int_0^{+\infty} x^{\alpha-1} e^{-x} dx = 2^{\alpha-1} \int_0^{+\infty} (x/2)^{\alpha-1} e^{-x} dx \le 2^{\alpha-1} (\{\alpha\}-1)! \int_0^{+\infty} e^{x/2} e^{-x} dx < +\infty$$

(b) $\alpha > 1$일 때 부분적분을 이용하면

$$\begin{aligned} \Gamma(\alpha) &= \int_0^{+\infty} x^{\alpha-1} e^{-x}\, dx \\ &= [x^{\alpha-1}(-e^{-x})]_0^{+\infty} - \int_0^{+\infty} (-e^{-x}) dx^{\alpha-1} \\ &= (\alpha-1) \int_0^{+\infty} e^{-x} x^{\alpha-2} dx \\ &= (\alpha-1)\Gamma(\alpha-1) \end{aligned}$$

한편 $\Gamma(1) = \int_0^{+\infty} e^{-x} dx = 1$ 이므로 자연수 n에 대하여

$$\Gamma(n) = (n-1)\Gamma(n-1) = (n-1)(n-2)\Gamma(n-2) = (n-1)(n-2)\cdots\Gamma(1) = (n-1)!$$

(c) $\Gamma(1/2) = \int_0^{+\infty} x^{-1/2} e^{-x}\, dx$ 에서 $\sqrt{x} = z/\sqrt{2}$ 로 치환하면

$$\begin{aligned} \Gamma(1/2) &= \int_0^{+\infty} x^{-1/2} e^{-x}\, dx \\ &= \int_0^{+\infty} (z/\sqrt{2})^{-1} e^{-z^2/2}\, d(z^2/2) \\ &= \sqrt{2} \int_0^{+\infty} e^{-z^2/2} dz \\ &= \sqrt{2}/2 \int_{-\infty}^{+\infty} e^{-z^2/2} dz \end{aligned}$$

<u>그런데 예 Ⅰ.6.1에서(추후에)</u>

$$\int_{-\infty}^{+\infty} e^{-\frac{1}{2}x^2} dx = \sqrt{2\pi}$$

$$\therefore\ \Gamma(1/2) = \sqrt{2}/2 \int_{-\infty}^{+\infty} e^{-z^2/2} dz = \sqrt{\pi}$$

Gamma(α,β)**분포:**

$$pdf(x) = \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta} \mathrm{I}_{(x>0)} \quad (\alpha > 0, \beta > 0)$$

정리 3.5.2:감마분포의 성질

(a) $\mathrm{X} \sim \mathrm{Gamma}(\alpha,\beta)$이면

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

(b) $\mathrm{X} \sim \mathrm{Gamma}(\alpha,\beta)$이면 그 적률생성함수는

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

(c) $\mathrm{X}_1 \sim \mathrm{Gamma}(\alpha_1,\beta), \mathrm{X}_2 \sim \mathrm{Gamma}(\alpha_2,\beta)$이고 $\mathrm{X}_1,\mathrm{X}_2$가 서로 독립이면

$$\mathrm{X}_1 + \mathrm{X}_2 \sim \mathrm{Gamma}(\alpha_1+\alpha_2,\beta)$$

[증명] (a) 다음과 같이 $x/\beta = y$로 치환하여 적분하면 X와 X^2의 기댓값을 구할 수 있다.

$$\int_0^{+\infty} x \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta} dx = \frac{\beta}{\Gamma(\alpha)} \int_0^{+\infty} y^{\alpha} e^{-y} dy = \frac{\Gamma(\alpha+1)}{\Gamma(\alpha)}\beta = \alpha\beta$$

$$\int_0^{+\infty} x^2 \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta} dx = \frac{\beta^2}{\Gamma(\alpha)} \int_0^{+\infty} y^{\alpha+1} e^{-y} dy = \frac{\Gamma(\alpha+2)}{\Gamma(\alpha)}\beta^2 = \alpha(\alpha+1)\beta^2$$

$$\therefore\ \mathrm{E(X)} = \alpha\beta,\ \mathrm{Var(X)} = \mathrm{E(X^2)} - (\mathrm{E(X)})^2 = \alpha(\alpha+1)\beta^2 - (\alpha\beta)^2 = \alpha\beta^2$$

(b) 다음과 같이 $(1/\beta - t)x = y$로 치환하여 적분하면

$$\begin{aligned} mgf_{\mathrm{X}}(t) &= \int_0^{+\infty} e^{tx} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} x^{\alpha-1} e^{-x/\beta} dx \\ &= \frac{1}{\Gamma(\alpha)\beta^{\alpha}} (1/\beta - t)^{-\alpha} \int_0^{+\infty} y^{\alpha-1} e^{-y} dy,\ \ 1/\beta - t > 0 \\ &= (1-\beta t)^{-\alpha},\ t < 1/\beta \end{aligned}$$

(c) 서로 독립인 확률변수의 합에 관한 정리 2.26과 (b)로부터

$$mgf_{\mathrm{X}_1+\mathrm{X}_2}(t) = mgf_{\mathrm{X}_1}(t) mgf_{\mathrm{X}_2}(t) = (1-\beta t)^{-\alpha_1-\alpha_2},\ t < 1/\beta$$

우변은 $\mathrm{Gamma}(\alpha_1+\alpha_2,\beta)$분포의 적률생성함수이므로, 적률생성함수의 분포결정성으로부터

$$\mathrm{X}_1 + \mathrm{X}_2 \sim \mathrm{Gamma}(\alpha_1+\alpha_2,\beta)$$

형상모수가 자연수인 감마분포의 대의적 정의:

형상모수 r이 자연수인 경우에

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

한편 발생률이 λ인 포아송과정에서 각각의 현상 발생 후 다음 현상의 발생까지의 시간을 나타내는 $\mathrm{W}_1, \mathrm{W}_2 - \mathrm{W}_1, \cdots, \mathrm{W}_r - \mathrm{W}_{r-1}$은 서로 독립이고 동일한 지수분포 $\mathrm{Exp}(1/\lambda)$를 따르고, 이들의 합인 W_r은 감마분포 $\mathrm{Gamma}(r, 1/\lambda)$을 따르는 것이 알려져 있다.

3.6 정규분포

이항분포의 누적 확률을 적분으로 나타내는 근사식[7](부록 Ⅰ.7)

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

에서의 함수

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

는 그 적분 값이 1이 되는 함수이다.

---------------------- 부록 I.6.1 ----------------------------

예 Ⅰ.6.1: 극좌표 변환

이상적분

$$\mathrm{I} = \int_{-\infty}^{+\infty} e^{-\frac{1}{2}x^2} dx$$

의 값을 극좌표 변환

$$\omega : \begin{cases} x_1 = r\cos\theta \\ x_2 = r\sin\theta, & 0 \le \theta < 2\pi, 0 \le r < +\infty \end{cases}$$

을 이용하여 구해보자. 부채꼴의 넓이 공식으로부터 바닥 넓이의 증분은

$$\triangle x_1 \triangle x_2 \simeq \frac{1}{2}(r+\triangle r)^2 \triangle\theta - \frac{1}{2} r^2 \triangle\theta \simeq r \triangle r \triangle\theta$$

즉 $$dx_1 dx_2 = r\, dr\, d\theta$$

$$\mathrm{I}^2 = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} e^{-\frac{1}{2}(x_1^2+x_2^2)} dx_1 dx_2 = \int_0^{2\pi}\int_0^{+\infty} r e^{-\frac{1}{2}r^2} dr d\theta = 2\pi$$

$$\therefore \int_{-\infty}^{+\infty} e^{-\frac{1}{2}x^2} dx = \sqrt{2\pi}$$

표준정규분포(標準正規分布 standard normal distribution): $\mathrm{Z} \sim \mathrm{N}(0,1)$

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

정규분포: $\mathrm{N}(\mu,\sigma^2)$

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

정규분포의 형태:

$x=\mu$에 대칭인 종 모양의 곡선으로서, σ는 분포의 흩어진 정도를 나타내고 있다.

7) 드모아브르(De Moivre)-라플라스(Laplace) 정리로 알려진 이 근사식의 증명은 부록 Ⅰ.7에 주어져 있다.

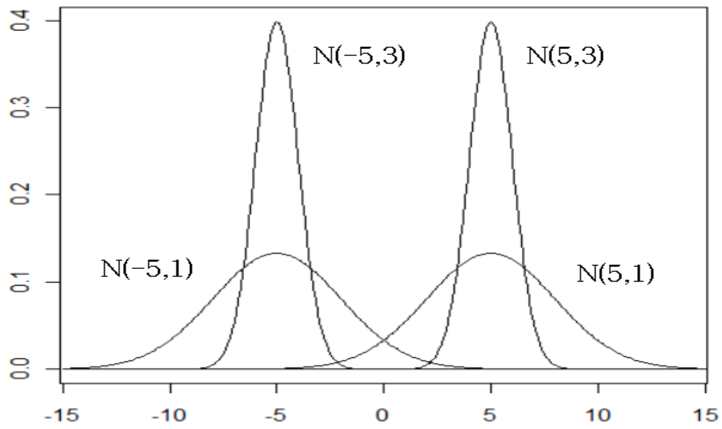


그림 3.6.1 정규분포 $\mathrm{N}(\mu,\sigma^2)$의 형태

정리 3.6.1:정규분포의 성질

(a) $\mathrm{X} \sim \mathrm{N}(\mu,\sigma^2)$이면

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

(b) $\mathrm{X} \sim \mathrm{N}(\mu,\sigma^2)$이면 그 적률생성함수는

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

(c) $\mathrm{X}_1 \sim \mathrm{N}(\mu_1,\sigma_1^2)$, $\mathrm{X}_2 \sim \mathrm{N}(\mu_2,\sigma_2^2)$이고 $\mathrm{X}_1,\mathrm{X}_2$가 서로 독립이면

$$\mathrm{X}_1 + \mathrm{X}_2 \sim \mathrm{N}(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$$

[증명] (a) $\dfrac{x-\mu}{\sigma} = z$로 치환

$$\mathrm{E}(\mathrm{X}) = \int_{-\infty}^{+\infty} x\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)dx = \sigma\int_{-\infty}^{+\infty} z\,\phi(z)\,dz + \mu$$

여기에서 $z^2/2 = y$ 로 치환하여 적분하면

$$\int_{-\infty}^{+\infty} z\,\phi(z)\,dz = -\frac{1}{\sqrt{2\pi}}\int_0^{+\infty} e^{-y}dy + \frac{1}{\sqrt{2\pi}}\int_0^{+\infty} e^{-y}dy = 0 \quad \therefore\ \mathrm{E}(\mathrm{X}) = \mu$$

같은 방법으로

$$\mathrm{Var}(\mathrm{X}) = \int_{-\infty}^{+\infty}(x-\mu)^2\frac{1}{\sigma}\phi\left(\frac{x-\mu}{\sigma}\right)dx = \sigma^2\int_{-\infty}^{+\infty} z^2\phi(z)\,dz$$

$$\int_{-\infty}^{+\infty} z^2\phi(z)dz = 2\int_0^{+\infty} z^2\phi(z)dz = \frac{2}{\sqrt{2\pi}}\int_0^{+\infty}\sqrt{2y}\,e^{-y}dy = \frac{2\Gamma(3/2)}{\sqrt{\pi}}$$

그런데 부록 Ⅰ의 예Ⅰ.6.2에서

$$2\Gamma(3/2) = 2\Gamma(1/2+1) = 2(1/2)\Gamma(1/2) = \sqrt{\pi}$$

$$\therefore \int_{-\infty}^{+\infty} z^2\phi(z)dz = 1,\ \ \mathrm{Var}(\mathrm{X}) = \sigma^2$$

(b) $\frac{x-\mu}{\sigma}=z$로 치환하여 $\phi(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}$에 관한 적분으로 나타내면

$$mgf_{\mathrm{X}}(t)=\int_{-\infty}^{+\infty}e^{tx}\frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma})dx=\int_{-\infty}^{+\infty}e^{t(\sigma z+\mu)}\phi(z)dz$$

여기에서 다음과 같이 $z-\sigma t=y$로 치환하여 적분하면

$$\int_{-\infty}^{+\infty}e^{t\sigma z}\phi(z)dz=\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(z-\sigma t)^2}dz\ e^{\frac{1}{2}\sigma^2t^2}=e^{\frac{1}{2}\sigma^2t^2}\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}y^2}dy=e^{\frac{1}{2}\sigma^2t^2}$$

$$\therefore\ mgf_{\mathrm{X}}(t)=\int_{-\infty}^{+\infty}e^{t(\sigma z+\mu)}\phi(z)dz=e^{\mu t+\frac{1}{2}\sigma^2t^2},\quad -\infty<t<+\infty$$

(c) 서로 독립인 확률변수의 합에 관한 정리 2.5.11과 (b)로부터

$$mgf_{\mathrm{X}_1+\mathrm{X}_2}(t)=mgf_{\mathrm{X}_1}(t)mgf_{\mathrm{X}_2}(t)=\exp\left\{(\mu_1+\mu_2)t+\frac{1}{2}(\sigma_1^2+\sigma_2^2)t^2\right\},\quad -\infty<t<+\infty$$

우변은 $\mathrm{N}(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$분포의 적률생성함수이므로, 적률생성함수의 분포 결정성으로부터

$$\mathrm{X}_1+\mathrm{X}_2\sim\mathrm{N}(\mu_1+\mu_2,\sigma_1^2+\sigma_2^2)$$

정리 3.6.2: 정규분포의 대의적 정의

(a) $\mathrm{X}\sim\mathrm{N}(\mu,\sigma^2)$이면 상수 a,b에 대하여

$$a\mathrm{X}+b\sim\mathrm{N}(a\mu+b,a^2\sigma^2)$$

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

[증명] (a) $mgf_{a\mathrm{X}+b}(t)=\mathrm{E}[e^{t(aX+b)}]=mgf_{\mathrm{X}}(at)e^{bt}$

$$mgf_{a\mathrm{X}+b}(t)=\exp\left\{bt+\mu(at)+\frac{1}{2}\sigma^2(at)^2\right\}=\exp\left\{(a\mu+b)t+\frac{1}{2}(a^2\sigma^2)t^2\right\},\quad -\infty<t<+\infty$$

우변은 $\mathrm{N}(a\mu+b,a^2\sigma^2)$분포의 적률생성함수이므로, 적률생성함수의 분포 결정성으로부터

$$a\mathrm{X}+b\sim\mathrm{N}(a\mu+b,a^2\sigma^2)$$

표준정규분포의 누적분포함수:

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

정규분포 $\mathrm{N}(\mu,\sigma^2)$를 따르는 X의 누적확률은

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

패키지 R(부록 III)을 이용하여 계산.

표준정규분포의 누적확률: 부록 Ⅳ에

$$\Phi(1.64)=0.9495,\ \ \Phi(1.65)=0.9505,\ \ \Phi(1.96)=0.9750$$

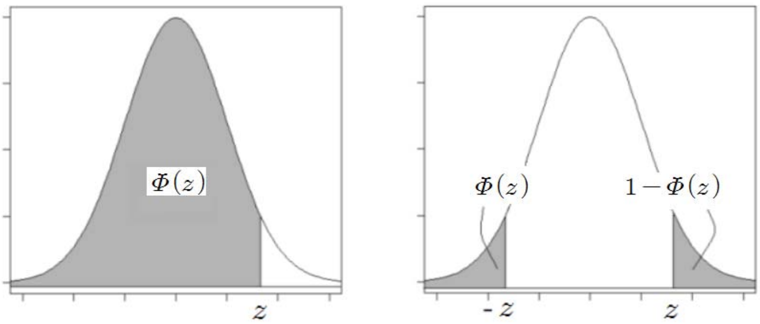


그림 3.6.2 표준정규분포의 누적확률

예 3.6.1

$X \sim N(3,4)$ 일 때 부록 Ⅳ의 표를 이용하여 다음 확률을 구하여라.

(a) $P(5 < X \le 7)$ (b) $P(1 < X \le 4)$ (c) $P(X > 0)$

[풀이] $Z = \dfrac{X-3}{\sqrt{4}} \sim N(0,1)$ 이므로 구하는 확률을 다음과 같이 계산할 수 있다.

(a) $P(5 < X \le 7) = \Phi(\frac{7-3}{2}) - \Phi(\frac{5-3}{2}) = \Phi(1.0) - \Phi(2.0)$

(b) $P(1 < X \le 4) = \Phi(\frac{4-3}{2}) - \Phi(\frac{1-3}{2}) = \Phi(0.5) - \Phi(-1.0) = \Phi(0.5) - (1-\Phi(1.0))$

(c) $P(X > 0) = 1 - P(X \le 0) = 1 - \Phi(\frac{0-3}{2}) = 1 - \Phi(-0.5) = \Phi(0.5)$

한편, $Z \sim N(0,1)$일 때

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

를 만족시키는 값 z_α를 표준정규분포의 상방 α 분위수(upper α quantile)라고 한다. 이러한 상방 분위수는 패키지 R이나 부록 Ⅳ의 표를 이용하여 구할 수 있다. 예를 들면

$$z_{0.025} = 1.96,\ z_{0.05} = 1.645$$

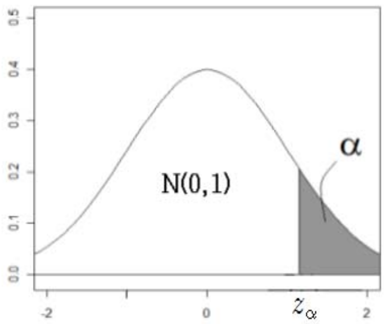


그림 3.6.3 표준정규분포의 상방 분위수

따라서 정리 3.6.2로부터 정규분포 $N(\mu,\sigma^2)$의 상방 α 분위수는 $\mu+\sigma z_\alpha$로 주어진다. 즉 $X \sim N(\mu,\sigma^2)$일 때

$$P(X > \mu + \sigma z_\alpha) = \alpha$$

예 3.6.2

$X \sim N(3,4)$ 일 때 다음 분위수 $q_{0.95}$와 $q_{0.025}$를 구하여라.

(a) $P(X \le q_{0.95}) = 0.95$ (b) $P(X \le q_{0.025}) = 0.025$

[풀이]

(a) $P(X > q_{0.95}) = 0.05$ 이므로 $q_{0.95} = 3 + 2z_{0.05} = 6.290$

(b) $P(X > q_{0.025}) = 0.975$ 이므로 $q_{0.025} = 3 + 2z_{0.975} = 3 - 2z_{0.025} = -0.92$

여러 가지 분포의 정의

분포의 명칭과 기호 (distribution)	확률밀도함수 (pdf)	대의적 정의 (representational definition)
이항분포 $\mathrm{B}(n,p)\ \ 0 \le p \le 1$	$\binom{n}{x}p^x(1-p)^{n-x}$, $x=0,1,\cdots,n$	$\mathrm{X} \sim \mathrm{B}(n,p) \Leftrightarrow \mathrm{X} \overset{\mathrm{d}}{\equiv} \mathrm{Z}_1 + \cdots \mathrm{Z}_n$, $\mathrm{Z}_i \overset{iid}{\sim} \mathrm{Bernoulli}(p)(i=1,\cdots,n)$
베르누이분포 $\mathrm{Bernoull}(p)$ $0 \le p \le 1$	$p^x(1-p)^{1-x}, x=0,1$	$\mathrm{B}(1,p)$
음이항분포 $\mathrm{Negbin}(r,p)$ $0<p<1$, r : 자연수	$\binom{x-1}{r-1}p^r(1-p)^{x-r}$ $x=r,r+1,\cdots$	$\mathrm{X} \sim \mathrm{Negbin}(r,p) \Leftrightarrow \mathrm{X} \overset{\mathrm{d}}{\equiv} \mathrm{Z}_1 + \cdots + \mathrm{Z}_r$ $\mathrm{Z}_i \overset{iid}{\sim} \mathrm{Geo}(p)(i=1,\cdots,r)$
기하분포 $\mathrm{Geo}(p)\ \ 0<p<1$	$(1-p)^{x-1}p, x=1,2,\cdots$	$\mathrm{Negbin}(1,p)$
포아송분포 $\mathrm{Poisson}(\lambda)\ \ \lambda \ge 0$	$\frac{e^{-\lambda}\lambda^x}{x!}, x=0,1,\cdots$	
다항분포 $\mathrm{Multi}(n,(p_1,p_2,\cdots,p_k)^t)$ $p_j \ge 0,\ \sum_{j=1}^{k} p_j = 1$	$\binom{n}{x_1 x_2 \cdots x_k} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$ $x_i = 0,\cdots,n\ (i=1,2,\cdots,k)$ $x_1+x_2+\cdots x_k = n$	$\mathrm{X}=(\mathrm{X}_1,\cdots,\mathrm{X}_k)^t, p=(p_1,\cdots,p_k)^t$ $\mathrm{X} \sim \mathrm{Multi}(n,p) \Leftrightarrow \mathrm{X} \overset{d}{\equiv} \mathrm{Z}_1 + \cdots + \mathrm{Z}_n$, $\mathrm{Z}_i \overset{iid}{\sim} \mathrm{Multi}(1,p)\ (i=1,\cdots,n)$ $\mathrm{Z}_i = (\mathrm{Z}_{i1},\cdots,\mathrm{Z}_{ik})^t$
감마분포 $\mathrm{Gamma}(\alpha,\beta)$ $\alpha>0, \beta>0$	$\frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta} \mathrm{I}_{(0,\infty)}(x)$	$\alpha = r$이 자연수인 경우: $\mathrm{X} \sim \mathrm{Gamma}(r,\beta) \Leftrightarrow \mathrm{X} \overset{d}{\equiv} \mathrm{Z}_1 + \cdots \mathrm{Z}_r$ $\mathrm{Z}_i \overset{iid}{\sim} \mathrm{Exp}(\beta)(i=1,\cdots,r)$
지수분포 $\mathrm{Exp}(1/\lambda)\ \ \lambda > 0$	$\lambda e^{-\lambda x} \mathrm{I}_{(0,\infty)}(x)$	$\mathrm{Gamma}(1,1/\lambda)$
정규분포 $\mathrm{N}(\mu,\sigma^2)$ μ : 실수, $\sigma > 0$	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$ $-\infty < x < +\infty$	$\mathrm{X} \sim \mathrm{N}(\mu,\sigma^2)$ $\Leftrightarrow \mathrm{X} \overset{d}{\equiv} \sigma \mathrm{Z} + \mu,\ \mathrm{Z} \sim \mathrm{N}(0,1)$

여러 가지 분포의 생성함수

분포의 명칭과 기호 (distribution)	적률생성함수 (mgf)	누율생성함수 (cgf)
이항분포 $\mathrm{B}(n,p)\ \ 0 \le p \le 1$	$(pe^t+q)^n,\ q=1-p$ $-\infty < t < +\infty$	$n\log\{1+p(e^t-1)\}$ $= n\sum_{k=1}^{\infty}\frac{1}{k}(-1)^{k-1}p^k\left(\sum_{j=1}^{\infty}\frac{t^j}{j!}\right)^k$
음이항분포 $\mathrm{Negbin}(r,p)$ $0<p<1,\ r$: 자연수	$\{pe^t(1-qe^t)^{-1}\}^r,\ q=1-p$ $t \leftarrow \log q$	$-r\log(1-p^{-1}(1-e^{-t}))$ $= r\sum_{k=1}^{\infty}\frac{1}{k}(p^{-1})^k\left(\sum_{j=1}^{\infty}(-1)^{j+1}\frac{t^j}{j!}\right)^k$
포아송분포 $\mathrm{Poisson}(\lambda)\ \ \lambda \ge 0$	$e^{-\lambda+\lambda e^t}$ $-\infty < t < +\infty$	$-\lambda+\lambda e^t = \sum_{k=1}^{\infty}\lambda\frac{t^k}{k!}$
다항분포 $\mathrm{Multi}(n,(p_1,p_2,\cdots,p_k)^t)$ $p_j \ge 0,\ \sum_{j=1}^{k}p_j=1$	$(p_1e^{t_1}+\cdots+p_ke^{t_k})^n$ $-\infty < t_j < +\infty\ (j=1,\cdots,k)$	$n\log\left\{1+\sum_{j=1}^{k}p_j(e^{t_j}-1)\right\}$
감마분포 $\mathrm{Gamma}(\alpha,\beta)\ \alpha>0,\beta>0$	$(1-\beta t)^{-\alpha}$ $t<1/\beta$	$-\alpha\log(1-\beta t)=\alpha\sum_{k=1}^{\infty}\frac{1}{k}\beta^k t^k$
정규분포 $\mathrm{N}(\mu,\sigma^2)$ μ는 실수, $\sigma>0$	$e^{\mu t+\frac{1}{2}\sigma^2t^2}$ $-\infty < t < +\infty$	$\mu t+\frac{1}{2}\sigma^2t^2$

정리 Ⅰ.7.2: 이항확률의 정규 근사

(a) $q=1-p$ 라고 할 때, $a \le \dfrac{x-np}{\sqrt{npq}} \le b$ 인 x에 대하여

$$\binom{n}{x} p^x (1-p)^{n-x} \sim \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-np}{\sqrt{npq}}\right)^2\right\} \frac{1}{\sqrt{npq}}, \quad n \to \infty$$

(b)

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

[증명] (a) 스털링의 공식

$$m! \sim m^{m+1/2} e^{-m} \sqrt{2\pi}, \quad m \to \infty$$

으로부터

$$\begin{aligned}
\binom{n}{x} p^x (1-p)^{n-x} &= \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\
&\sim \frac{n^{n+1/2} e^{-n} \sqrt{2\pi}}{x^{x+1/2} e^{-x} \sqrt{2\pi}\, (n-x)^{n-x+1/2} e^{-(n-x)} \sqrt{2\pi}} p^x (1-p)^{n-x} \\
&\sim \frac{1}{\sqrt{2\pi}} \frac{p^x (1-p)^{n-x}}{\left(\frac{x}{n}\right)^x \left(1-\frac{x}{n}\right)^{n-x}} \frac{1}{\sqrt{n\frac{x}{n}\left(1-\frac{x}{n}\right)}}
\end{aligned}$$

여기에서 $z = \dfrac{x-np}{\sqrt{npq}}$ 라고 하면

$$\begin{aligned}
&\log \frac{p^x (1-p)^{n-x}}{\left(\frac{x}{n}\right)^x \left(1-\frac{x}{n}\right)^{n-x}} \\
&= -x\left(\log\frac{x}{n} - \log p\right) - (n-x)\left(\log\left(1-\frac{x}{n}\right) - \log(1-p)\right) \\
&= -(np + z\sqrt{npq}) \log\left(1 + z\frac{\sqrt{q/p}}{\sqrt{n}}\right) - (nq - z\sqrt{npq}) \log\left(1 - z\frac{\sqrt{p/q}}{\sqrt{n}}\right) \\
&= -\frac{1}{2}z^2 + \frac{1}{\sqrt{n}}(\cdots) + \cdots \qquad (\because -\log(1-t) = t + t^2/2 + t^3/3 + \cdots,\ t \to 0)
\end{aligned}$$

$$\therefore\ \frac{p^x (1-p)^{n-x}}{\left(\frac{x}{n}\right)^x \left(1-\frac{x}{n}\right)^{n-x}} = \exp\left\{-\frac{1}{2}z^2 + \frac{1}{\sqrt{n}}(\cdots) + \cdots\right\} \sim \exp\left(-\frac{1}{2}z^2\right)$$

한편 $a \le \dfrac{x/n-p}{\sqrt{pq/n}} \le b,\ -b \le \dfrac{(1-x/n)-(1-p)}{\sqrt{pq/n}} \le -a$ 이므로

$$\frac{1}{\sqrt{n\frac{x}{n}\left(1-\frac{x}{n}\right)}} \sim \frac{1}{\sqrt{np(1-p)}} = \frac{1}{\sqrt{npq}}$$

그러므로 $a \le \dfrac{x-np}{\sqrt{npq}} \le b$ 일 때, $z = \dfrac{x-np}{\sqrt{npq}}$에 대하여

$$\begin{aligned}
\binom{n}{x} p^x (1-p)^{n-x} &\sim \frac{1}{\sqrt{2\pi}} \frac{p^x (1-p)^{n-x}}{\left(\frac{x}{n}\right)^x \left(1-\frac{x}{n}\right)^{n-x}} \frac{1}{\sqrt{n\frac{x}{n}\left(1-\frac{x}{n}\right)}} \\
&\sim \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}z^2\right) \frac{1}{\sqrt{npq}}
\end{aligned}$$

(b) $a \le \dfrac{x-np}{\sqrt{npq}} \le b$ 를 만족시키는 최소의 정수를 $x_{\min}$, 최대의 정수를 $x_{\max}$라고 하고,

$$z_1 = \frac{x_{\min}-np}{\sqrt{npq}}, \cdots\cdots, z_N = \frac{x_{\max}-np}{\sqrt{npq}}, \quad \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

라고 하면 $\quad z_1 \sim a,\ z_N \sim b,\ \dfrac{N-1}{\sqrt{npq}} = z_N - z_1 \sim b-a$

$$\begin{aligned} \therefore \sum_{x\,:\,a \le \frac{x-np}{\sqrt{npq}} \le b} \binom{n}{x} p^x (1-p)^{n-x} &\sim \frac{1}{\sqrt{npq}} \sum_{x\,:\,a \le \frac{x-np}{\sqrt{npq}} \le b} \phi\left(\frac{x-np}{\sqrt{npq}}\right) \\ &\sim \frac{b-a}{N-1} \sum_{j=1}^{N} \phi(z_j) \\ &\sim \int_a^b \phi(z)\,dz \end{aligned}$$

정리 Ⅰ.7.1 스털링(Stirling)의 공식

$$m! \sim m^{m+1/2} e^{-m} \sqrt{2\pi}, \quad m \to \infty$$

여기에서 $a_m \sim b_m$, $m\to\infty$ 의 의미는 $\lim\limits_{m\to\infty} a_m/b_m = 1$을 뜻한다.

[증명]

$$\begin{aligned} m! &= \Gamma(m+1) \\ &= \int_0^{+\infty} x^m e^{-x}\,dx \\ &= \int_0^{+\infty} e^{m\log x - x}\,dx \\ &= \int_0^{+\infty} e^{m\log my - my} m\,dy \\ &= m^{m+1} \int_0^{+\infty} e^{m(\log y - y)}\,dy \\ &= m^{m+1} e^{-m} \int_0^{+\infty} e^{m(\log y - y + 1)}\,dy \qquad (\log y = \log(1+(y-1))\text{의 전개}) \\ &= m^{m+1} e^{-m} \int_0^{+\infty} \exp\left[m\left\{-\frac{1}{2}(y-1)^2 + \frac{1}{3}(y-1)^3 - \frac{1}{4}(y-1)^4 + \cdots\right\}\right] dy \\ &= m^{m+1} e^{-m} \int_0^{+\infty} e^{-\frac{m}{2}(y-1)^2} \exp\left\{\frac{m}{3}(y-1)^3 - \frac{m}{4}(y-1)^4 + \cdots\right\} dy \end{aligned}$$

여기에서 $z = \sqrt{m}(y-1)$로 치환하면

$$\begin{aligned} m! &= m^{m+1} e^{-m} \int_{-\sqrt{m}}^{+\infty} e^{-\frac{1}{2}z^2} \exp\left(\frac{1}{\sqrt{m}}\frac{1}{3}z^3 - \frac{1}{m}\frac{1}{4}z^4 + \cdots\right) \frac{1}{\sqrt{m}}\,dz \\ &= m^{m+1/2} e^{-m} \sqrt{2\pi} \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \left\{1 + \frac{1}{\sqrt{m}}\frac{1}{3}z^3 + \frac{1}{m}\left(\frac{1}{2}\frac{1}{9}z^6 - \frac{1}{4}z^4\right) + \cdots\right\} dz \\ &\sim m^{m+1/2} e^{-m} \sqrt{2\pi}, \quad m \to \infty \end{aligned}$$