1장. 기본적인 확률분포 5개

통계학(statistics, statistical science)의 주요영역 중 하나는 불확실한 현상에 대한 과학적 인식과 처리 방법론입니다. 예를 들어볼까요? A라는 사람이 암에 걸렸습니다. 그는 1년을 버티지 못하고 죽었습니다. B도 같은 암에 걸렸습니다. 그는 1년 넘게 살았습니다. C도 같은 암에 걸렸습니다. 그도 1년 넘게 살았습니다. 결과는 이렇게 경우마다 달랐습니다. 그러니 암의 예후는 불확실하다고 할 수 있습니다. 이렇게 불확실한 현상에 대한 과학적 접근은 다음 두 가지 관점에서 나옵니다.

첫째, 왜 어떤 사람은 죽고 어떤 사람은 사는가? 즉 결과를 결정짓는 원인을 탐구하는 것입니다.

둘째, 그것을 임의적 현상(random phenomenon)으로 인식하고 불확실성을 계량화하는 것입니다.

이 두 관점은 방향이 다릅니다. 첫 번째 관점은 적극적인 것입니다. 작업이 성공한다면 놀랄만한 과학적 성취하고 할 수 있겠지요. 그러나 인체, 환경, 또는 사회현상과 같은 복잡계에 대하여는 결과를 결정짓는 원인들을 완전히 밝히는 것은 어렵습니다. 단순계와는 다른 것입니다. 그러므로 어느 시점에서도 일정부분 만큼에 대하여는 우리가 모른다는 것으로 인정하지 않을 수 없습니다. 그러나 그것을 단순히 어떻게 될지 알 수 없다고 하는 것이 아니라 불확실성을 계량화함으로써 합리적으로 대처하자는 것이 다소 소극적인 두 번째 관점입니다. 그러므로 통계학의 영역에서는 접근방향이 다른 두 관점이 공존한다고 하겠습니다.

이제 두 번째 관점에서, 우리가 빈번히 사용하는 확률분포를 알아 둘 필요가 있습니다. 이 장에서는 그 중에서도 기본적인 다섯 개의 기본분포를 소개합니다.

차례:

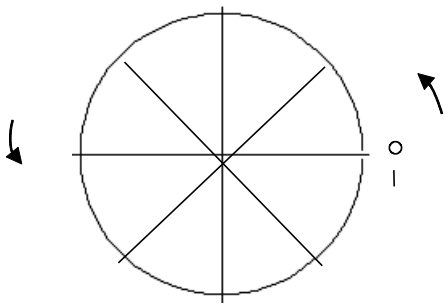
1.1 균일분포(uniform distribution)

여러 분은 뻔데기 판을 찍어 본 일이 있습니까? 다트 판은? 아니면 미국 TV에서 'Wheel of Fortune'을 본 일이 있겠지요. <그림 1>과 같이 둘레에 0부터 1까지 눈금이 새겨진 원판을 힘차게(↔살금살금) 돌린다고 합시다. 어디에서 멈출까요? 누구도 0과 1 사이의 어디에서 멈출지 전혀 예측할 수 없을 것입니다. 따라서 확률변수 X에 대한 밀도함수가 다음과 같습니다 .

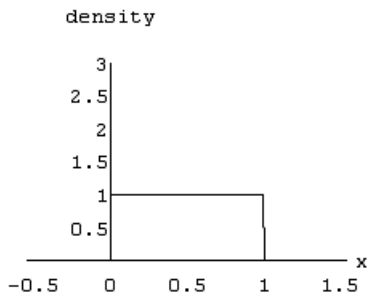
$$f_X(x) = 1, \quad 0 \le x \le 1. \tag{1}$$

여기서 $x = 1$을 포함시키지 말아야 하는가 (원판에서 0과 1은 일치하기 때문에), 포함시켜도 되는가는 사실상 중요하지 않습니다. <그림 2>를 보세요.

이것은 직관적으로 그래야 매우 마땅하기 때문에 왜 그러냐고 따지지 말기를 바랍니다 (따지지 않고 알 수 있는 사람은 행복합니다). 그래도 철저한 엄밀성을 요구하는 것이 현대수학의 못된 전통이기 때문에 굳이 따지고 싶은 학생은 다음에 제시된 논거를 보세요 (꼭 따져 보고야 직성이 풀리는 사람도 훌륭합니다).



<그림 1> 회전 원판



<그림 2> 균일분포의 밀도함수

균일분포의 유도

어떤 사람이 <그림 1>의 원판을 한 바퀴 이내로 살금살금 돌린다고 합시다. 그리고 그의 밀도함수를

$$h_1(t) > 0, \quad 0 \le t \le 1$$

라고 합시다. 여기서 t는 기선으로부터 회전된 바퀴 수(실수 값)입니다. 이 밀도함수는 전혀 균일분포가 아닙니다만 미분가능함수로 가정하겠습니다. 그가 점점 힘을 줘서 원판을 돌린다고 합시다. 즉, n 바퀴까지 돌아가도록 그의 새 밀도함수를

$$h_n(t) = \frac{1}{n} h_1\left(\frac{t}{n}\right), \quad 0 \le \frac{t}{n} \le 1$$

로 표현합시다. 눈금 x $(0 \le x \le 1)$와 회전된 바퀴 수 t $(0 \le t \le n)$ 사이의 관계는

$$x = t \pmod 1,$$

즉, x는 t를 1로 나누었을 때의 나머지입니다. 따라서

$$P_n(X \le x) = \sum_{k=0}^{n-1} \int_k^{k+x} h_n(t)\, dt$$

$$= \sum_{k=0}^{n-1} \int_k^{k+x} \frac{1}{n} h_1\left(\frac{t}{n}\right) dt = \sum_{k=0}^{n-1} \int_{\frac{k}{n}}^{\frac{k+x}{n}} h_1(u)\, du$$

입니다. 그런데 어떤 $\xi_{k,x,n}\left(\frac{k}{n} \le \xi_{k,x,n} \le \frac{k}{n}+\frac{x}{n}\right)$에 대하여

$$\int_{\frac{k}{n}}^{\frac{k+x}{n}} h_1(u)\, du = \frac{x}{n} h_1\left(\frac{k}{n}\right) + \frac{1}{2}\left(\frac{x}{n}\right)^2 h_1'(\xi_{k,x,n})$$

이 성립합니다. 따라서

$$\lim_{n\to\infty} P_n(X \le x) = \lim_{n\to\infty} \sum_{k=0}^{n-1} \int_{\frac{k}{n}}^{\frac{k+x}{n}} h_1(u)\, du$$

$$= \lim_{n\to\infty} \sum_{k=0}^{n-1} \left\{ \frac{x}{n} h_1\left(\frac{k}{n}\right) + \frac{1}{2}\left(\frac{x}{n}\right)^2 h_1'(\xi_{k,x,n}) \right\}$$

입니다 (여기서 사실은, 함수 $h_1(\theta)$의 1계 미분 $h_1'(t)$이 $0 \le t \le 1$사이에서 상하로 유계(bounded)되어 있다는 가정이 추가로 필요합니다). 그러므로 아주 세게 돌려진 원판은, 누가 돌리느냐에 관계없이 (함수 $h_1(t)$에 관계없이), 0과 1사이 임의의 눈금에서 멈춘다고 할 수 있겠습니다. ■

이와 같은 원리는 컴퓨터에서 0과 1사이의 임의수(random number, 亂數)를 생성시키는 보편적 방법인 선형합동법의 기본원리이기도 합니다.

선형합동법(linear congruential method)

단계 0: 0과 m 사이의 임의 정수 c_0를 선택합니다. 그리고 n을 1로 둡니다.

단계 1: $c_n = a \cdot c_{n-1} + b \pmod m$을 계산합니다. 여기서 a는 양의 정수, b는 비음의 정수이고 $\bmod m$은 해당 숫자를 m으로 나누었을 때의 나머지입니다.

단계 2: c_n을 m으로 나누어 0과 1 사이의 수 x_n을 산출합니다. 그리고 n을 1 만큼 증가시킵니다.

단계 3: 단계 1과 단계 2를 N번 반복하여 원하는 만큼의 0과 1 사이의 난수들을 생성시킵니다. 즉 $x_1, \cdots, x_N$을 얻습니다.

여기서 m은 예컨대 $2^{31}-1$(=2,147,483,647)이고 a는 397,204,094이며 b는 0입니다. 선형합동법이 실제로는 진짜 임의수를 만들지는 못하지만 m과 a와 b를 잘 잡으면 마치 임의적인 것처럼 보입니다 (이에 관하여는 많은 정수론과 실증적 연구가 있습니다. 최영훈・이승천(1995) 또는 손건태(1996) 참조). 하여튼, 선형합동법의 핵심은 앞 단계의 임의정수 c_{n-1} 다음에 어떤 임의정수 c_n이 나올 것인지 손가락에 힘주어 계산하기 전에는 알기 어렵다는 데 있습니다. 여러분에게 간단한(?) 계산문제 하나를 내겠습니다. c_0를 32,902,235 (내 연구실 전화번호)로 하고 m과 a와 b를 앞과 같이 잡는다면, 다음 임의정수 c_1과 0과 1 사이의 임의수 x_1은 얼마입니까? (답: c_1=2,015,391,483, x_1=0.9385). ■

이제부터, (1)의 밀도함수를 표준균일분포라고 하겠습니다. 표준균일분포를 축을 따라 늘리거나 이동시킴으로써 균일분포군을 만들어 볼 수 있습니다. 예컨대

Uniform$(0,\theta)$ 분포군:

$$f_X(x) = \frac{1}{\theta}, \quad 0 \le x \le \theta.$$

Uniform(θ_1,θ_2) 분포군:

$$f_X(x) = \frac{1}{\theta_2 - \theta_1}, \quad \theta_1 \le x \le \theta_2.$$

1.2 지수분포(exponential distribution)

확률변수 $X \geq 0$가 다음 조건을 만족한다고 합시다.

$$P(X > u+v \mid X > u) = P(X > v), \quad \text{모든 } u \geq 0, v \geq 0 \text{에 대하여} \qquad (2)$$

가령 X가 수명(壽命, lifetime)을 나타낸다고 할 때, 위 식의 좌변은 $u \geq 0$ 이상 생존한 개체가 $v \geq 0$ 이상 더 생존할 확률을 의미하고 우변은 개체가 $v \geq 0$ 이상 생존할 확률을 의미합니다. 그러므로 위 식은 잔여수명의 분포가 현재 나이 $u \geq 0$에 관계없음을 뜻합니다. 매우 특수한 경우이긴 합니다. 이제 확률변수 X의 밀도함수를 $f_X(x) > 0,\ 0 \leq x < \infty$, 분포함수를 $F_X(x),\ x \geq 0$ $(F_X(0) = 0)$라고 하고, $f_X(x)$와 $F_X(x)$를 다음과 같이 유도해볼 수 있겠습니다 $(f_X(x) = F_X'(x))$.

지수분포의 유도

(2)를 분포함수 $F_X(.)$를 써서 재표현하면

$$\frac{1-F_X(u+v)}{1-F_X(u)} = 1-F_X(v),\ u \geq 0,\ v \geq 0$$

즉

$$F_X(u+v) = 1-(1-F_X(u))(1-F_X(v)) = F_X(u)+F_X(v)-F_X(u)F_X(v)$$

입니다. 따라서, $F_X(.)$가 미분가능하다는 전제하에

$$\begin{aligned} f_X(u) &= \lim_{v\to 0} \frac{F_X(u+v)-F_X(u)}{v} \\ &= \lim_{v\to 0} (1-F_X(u)) \frac{F_X(v)}{v} \\ &= (1-F_X(u)) \lim_{v\to 0} \frac{F_X(v)-F_X(0)}{v} = (1-F_X(u))\, f_X(0) \end{aligned}$$

가 성립합니다 ($X \geq 0$이므로 $F_X(0) = 0$이기 때문). 다시 쓰면,

$$\frac{f_X(u)}{1-F_X(u)} = k, \quad u \geq 0$$

입니다 (여기서 상수 $k = f_X(0)$). 그러므로

$$\int_0^x \frac{f_X(u)}{1-F_X(u)}\, du = \int_0^x k\, du, \quad x \geq 0.$$

그런데 $f_X(u) = F_X'(u)$, $F_X(0) = 0$이므로, 따라서

$$-\log_e(1-F_X(x)) = kx,$$

즉

$$F_X(x) = 1-\exp[-kx],\ x \geqq 0$$

을 얻게 됩니다. 초기조건으로 $f_X(0) = 1\,(=k)$ 이라고 하면, 밀도함수로

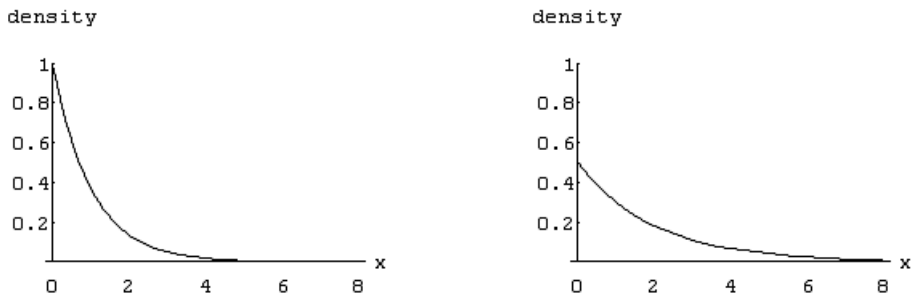
$$f_X(x) = \exp(-x),\ x \geq 0$$

의 형태가 유도됩니다. 이것이 표준지수분포의 밀도입니다. ■

표준지수분포의 밀도함수를 x축을 따라 늘리거나 이동하여 다음과 같은 지수분포군을 얻을 수 있습니다. <그림 3>을 보십시오.

Exponential(θ) : $f_X(x) = \frac{1}{\theta}\exp\left(-\frac{x}{\theta}\right),\ x \geqq 0,\ \theta > 0.$

Exponential(α,θ) : $f_X(x) = \frac{1}{\theta}\exp\left(-\frac{x-\alpha}{\theta}\right),\ x \geqq \alpha,\ \theta > 0.$



$\theta = 1$ $\theta = 2$

<그림 3> 지수분포의 밀도함수 ($\alpha = 0$)

1.3 정규분포(normal distribution)

정규분포는 일명 가우스 분포(Gaussian distribution)이라고 하여 통계학에서 가장 많이 언급되는 확률분포입니다. 중심극한정리(2.2절)로부터 이 분포를 유도하는 것이 정석이겠지만, 여기서는 앞서 소개된 균일분포와 지수분포를 가지고 정규분포를 만들어보도록 하겠습니다.

다음 두 조건을 만족하는 확률변수(X_1, X_2)의 결합밀도함수를 찾기로 합시다. 먼저 직교좌표 (X_1, X_2)를 극좌표 (R, Ψ)로 표현하기로 하겠습니다. 즉,

$$X_1 = R\cos\Psi,\ X_2 = R\sin\Psi \quad (R \geq 0, 0 \leq \Psi \leq 2\pi).$$

조건 1: $R^2 = X_1^2 + X_2^2$ 이 $\theta = 2$인 지수분포 Exponential(2)를 따른다.

조건 2: Ψ는 균일분포 Uniform$(0, 2\pi)$를 따른다.

조건 3: R과 Ψ는 독립이다.

이로부터 확률변수(X_1, X_2)의 결합밀도가 다음과 같음을 유도할 수 있습니다 (조금 아래를 참조하세요). <그림 4>를 보십시오.

$$f_{X_1,X_2}(x_1,x_2) = \frac{1}{2\pi} exp\left\{-\frac{x_1^2+x_2^2}{2}\right\},\ -\infty < x_1, x_2 < \infty. \tag{3}$$

그러므로 두 확률변수 X_1과 X_2가 각각 밀도함수

$$f_X(x) = \frac{1}{\sqrt{2\pi}} exp\left\{-\frac{x^2}{2}\right\},\ -\infty < x < \infty. \tag{4}$$

로부터 독립적으로 생성됨을 알 수 있습니다. (4)의 확률밀도에 의한 분포를 표준정규분포(standard normal distribution)라고 하지요. <그림 5>를 보십시오.

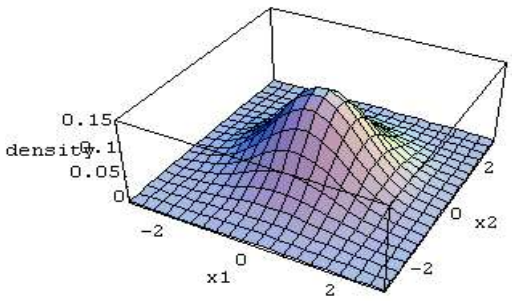
표준정규분포를 x축을 따라 이동하고 늘여서 정규분포군의 밀도함수를 다음과 같이 정의할 수 있습니다.

$$N(\mu, \sigma^2) \ :\ f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma} exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\},\ -\infty < x < \infty.$$

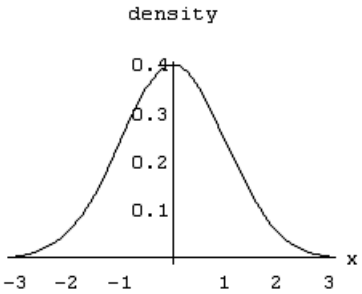
<u>이변량 정규확률밀도함수의 유도</u>

$W = R^2$이 Exponential(2) 분포를 따르므로

$$f_W(w) = \frac{1}{2}\exp\left(-\frac{w}{2}\right),\ w \geq 0$$



<그림 4> 이변량 정규분포의 밀도함수



<그림 5> 표준정규분포의 밀도함수

입니다. 따라서 $R = \sqrt{W}\ (\geq 0)$의 밀도는

$$f_R(r) = r\exp\left(-\frac{r^2}{2}\right),\ r \geq 0$$

이 됩니다. R과 독립적으로 Ψ가 Uniform$(0, 2\pi)$ 분포를 따르므로

$$f_{R,\Psi}(r,\psi) = \frac{r}{2\pi}\exp\left(-\frac{r^2}{2}\right),\ r \geq 0,\ 0 \leq \psi \leq 2\pi$$

라는 것을 알 수 있습니다. 그런데

$$J = \begin{vmatrix} \partial x_1/\partial r & \partial x_1/\partial \psi \\ \partial x_2/\partial r & \partial x_2/\partial \psi \end{vmatrix} = \begin{vmatrix} \cos\psi & -r\sin\psi \\ \sin\psi & r\cos\psi \end{vmatrix} = r$$

이므로

$$f_{X_1,X_2}(x_1,x_2)\ |J|\ = f_{R,\Psi}\ (r,\psi) = \frac{r}{2\pi}\exp\left(-\frac{r^2}{2}\right)$$

입니다. 따라서

$$f_{X_1,X_2}(x_1,x_2) = f_{R,\Psi}\ (r,\psi)\ \frac{1}{r} = \frac{1}{2\pi}\exp\left(-\frac{x_1^2+x_2^2}{2}\right)$$

가 유도되어 나옵니다. ■

두 확률변수 X_1과 X_2가 독립적으로 각각 표준정규분포를 따른다고 합시다. 이 때, 확률변수 Y_1과 Y_2를 다음과 같이 정의하기로 합니다.

$$Y_1 = X_1,\quad Y_2 = \rho X_1 + \sqrt{1-\rho^2}\ X_2,$$

여기서 ρ는 -1과 1 사이의 상수입니다. 그러면

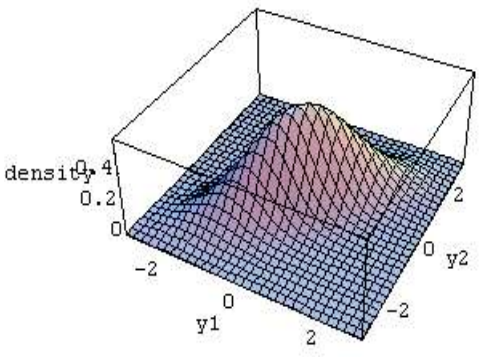
$$\mathrm{Cov}\,(Y_1, Y_2) = \mathrm{Cov}\,(X_1, \rho X_1 + \sqrt{1-\rho^2}\ X_2) = \mathrm{Cov}\,(X_1, \rho X_1) = \rho,$$

$$\mathrm{Var}\,(Y_1) = 1,\ \mathrm{Var}\,(Y_2) = \mathrm{Var}\,(\rho X_1 + \sqrt{1-\rho^2}\ X_2) = \rho^2 + (1-\rho^2) = 1$$

이므로, ρ는 Y_1과 Y_2 사이의 상관(積率相關, product-moment correlation)이 됩니다. 그리고, (Y_1, Y_2)의 확률분포가

$$f_{Y_1,Y_2}(y_1,y_2) = \frac{1}{2\pi\ \sqrt{1-\rho^2}}\exp\left\{-\frac{y_1^2 - 2\rho y_1 y_2 + y_2^2}{2\,(1-\rho^2)}\right\},\quad -\infty < y_1, y_2 < \infty \qquad (5)$$

로 유도됩니다 (다음 쪽). <그림 6>을 보십시오. 그리고 <그림 4>와 비교하세요.



<그림 6> 상관 $\rho = 0.5$인 이변량 정규분포의 밀도함수

상관이 ρ인 이변량 정규분포의 유도

(X_1, X_2)의 밀도함수가 (1)입니다. 그런데

$$f_{Y_1,Y_2}(y_1,y_2)\ |J| = f_{X_1,X_2}(x_1,x_2) = \frac{1}{2\pi}\exp\left(-\frac{x_1^2+x_2^2}{2}\right)$$

에서,

$$J = \begin{vmatrix} \partial y_1/\partial x_1 & \partial y_1/\partial x_2 \\ \partial y_2/\partial x_1 & \partial y_2/\partial x_2 \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{vmatrix} = \sqrt{1-\rho^2}\ ,$$

$$x_1^2 + x_2{}^2 = y_1^2 + \frac{(y_2-\rho y_1)^2}{1-\rho^2} = \frac{y_1^2 - 2\rho y_1 y_2 + y_2^2}{1-\rho^2}$$

이므로

$$f_{Y_1,Y_2}(y_1,y_2) = \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left\{-\frac{y_1^2-2\rho y_1 y_2+y_2^2}{2(1-\rho^2)}\right\}$$

입니다. 참고로 덧붙이자면, Y_1과 Y_2의 주변분포는 여전히 표준정규분포라는 것입니다. 즉,

$$f_{Y_1}(y_1) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y_1^2}{2}\right),\quad f_{Y_2}(y_2) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y_2^2}{2}\right)$$

입니다 (증명은 각자 해보세요). 따라서, $\rho \neq 0$인 경우

$$f_{Y_1,Y_2}(y_1,y_2) \neq f_{Y_1}(y_1)\, f_{Y_2}(y_2)$$

입니다. ■

(5)를 더욱 확장하여 Y_1과 Y_2가 각각 $N(\mu_1,\sigma_1^2)$ 분포와 $N(\mu_2,\sigma_2^2)$ 분포를 따르면서 상관이 ρ인 이변량 정규분포의 밀도함수를 다음과 같이 얻을 수 있습니다.

$BN(\mu_1,\mu_2,\sigma_1^2,\sigma_2^2,\rho)$:

$$f_{Y_1,Y_2}(y_1,y_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\ \exp[-Q(y_1,y_2)/2],\quad -\infty < y_1, y_2 < \infty\,.$$

여기서

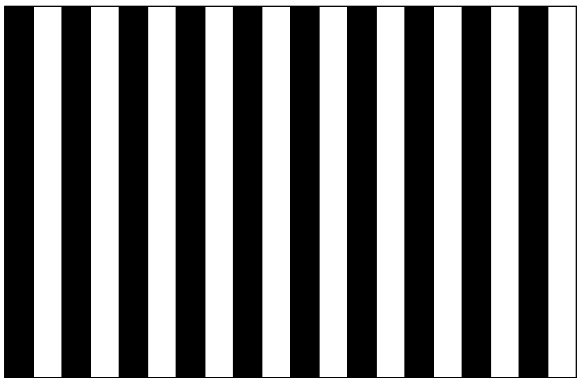
$$Q(y_1,y_2) = \frac{1}{1-\rho^2}\left\{\left(\frac{y_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{y_1-\mu_1}{\sigma_1}\right)\left(\frac{y_2-\mu_2}{\sigma_2}\right) + \left(\frac{y_2-\mu_2}{\sigma_2}\right)^2\right\}.$$

꽤 복잡해 보이지요. 그래도, 통계학을 직업으로 할 생각이 있는 학생들은 외워 두세요. '프로'라면 무엇인가 달라야 할 것입니다.

1.4 베르누이 분포(Bernoulli distribution)

여러분들 초등학교 때 교실에서 분필 던지기 장난을 해본 적이 있습니까? 분필을 작게 잘라서 그것으로 친구들을 맞추고 맞기도 하다 보면 교실 바닥과 분위기가 엉망이 되는 놀이이지요. 그러나 흥분은 잠시일 뿐, 벌이 훨씬 길다는 것을 명심해야 합니다. 하여튼 우리도 그것을 해보자고요.

교실 앞 칠판에 <그림 7>과 같이 반복 띠를 만들어 놓고 교실 뒤에 서서 분필조각 하나를 까만 띠를 향해 세게 던져 봅시다. 까만 띠와 하얀 띠 중 어디에 맞았습니까? 그 짓을 9번 더 해봅시다. 그리고 나서 90번 더 반복해봅시다. 이제까지 모두 100번, 실컷 했나요? 그 중에서 까만 띠에 모두 몇 번 맞았습니까?



<그림 7> 반복 띠의 칠판

까만 띠에 맞을 확률은 1/2이라고 해야겠지요. 왜 그런가요? 직관적으로 그렇지요. 물론 분필을 잘 던지는 학생에게는 그 확률이 1/2 보다 클 것입니다. 그러나 그 학생에게도 띠 폭을 줄이고 띠 수를 많게 함에 따라 까만 띠에 맞을 확률이 점차 1/2에 접근해갈 것입니다. 그 이유를 보기로 합시다.

베르누이 확률 1/2의 유도

칠판의 왼쪽 끝을 0, 오른 쪽 끝을 1이라고 하고 가로 폭을 $2n$ 등분하여 흑(B), 백(W), 흑(B), 백(W), …, 흑(B), 백(W)으로 칠하였다고 합시다. 교실

뒤에서 던진 분필조각이 칠판에 맞은 곳의 좌표를 $x\ (0 \le x \le 1)$라고 하고 어느 학생의 경우 그것이 밀도함수 $h(x) > 0\ (0 \le x \le 1)$를 따라 분포한다고 합시다. 그러면 흑(B) 띠에 맞을 확률은

$$P_n(B) = \sum_{k=0}^{n-1} \int_{\frac{2k}{2n}}^{\frac{2k+1}{2n}} h(x)\,dx$$

입니다. 이제 $h(x)$가 미분가능하다는 전제하에 $P_n(B)$의 극한이 1/2임을 증명할 것입니다. 어떤 $\xi_{k,n}\left(\frac{2k}{2n} \le \xi_{k,n} \le \frac{2k+1}{2n}\right)$에 대하여

$$\int_{\frac{2k}{2n}}^{\frac{2k+1}{2n}} h(x)\,dx = \frac{1}{2n} h\left(\frac{2k}{2n}\right) + \left(\frac{1}{2n}\right)^2 h'(\xi_{k,n})$$

이므로 (추가로 $h(\cdot)$의 1계 미분의 절대값이 유계되어 있다는 전제하에)

$$\lim_{n\to\infty} \sum_{k=0}^{n-1} \int_{\frac{2k}{2n}}^{\frac{2k+1}{2n}} h(x)\,dx = \lim_{n\to\infty} \sum_{k=0}^{n-1} \left\{ \frac{1}{2n} h\left(\frac{2k}{2n}\right) + \left(\frac{1}{2n}\right)^2 h'(\xi_{k,n}) \right\}$$

$$= \frac{1}{2} \lim_{n\to\infty} \sum_{k=0}^{n-1} \frac{1}{n} h\left(\frac{k}{n}\right)$$

$$= \frac{1}{2} \int_0^1 h(x)\,dx = \frac{1}{2}$$

을 얻게 됩니다. ■

만약 흑 띠와 백 띠의 폭을 $\alpha : \beta$의 비로 하면, 흑(B)의 확률로 $\theta\,(=\alpha/(\alpha+\beta))$, 백(W)의 확률로 $1-\theta\ (=\beta/(\alpha+\beta))$를 얻을 것입니다. 일반적으로, $X=1$(성공), 또는 0(실패)이 나타나고 $X=1$(성공)의 확률이 θ인 경우, 간결히

$$P(X=x) = \theta^x\ (1-\theta)^{1-x}, \quad 0 < \theta < 1, \quad x = 0, 1$$

로 표현가능합니다. 이 이산형 분포를 베르누이 분포 Bernoulli(θ)라고 합니다.

$X_1, X_2, X_3, \cdots$을 베르누이 분포 Bernoulli(θ)로부터 독립적으로 생성된다고 합시다. 예컨대 동전 던지기를 거듭 반복하는 것이 되겠습니다. 이런 것을 베르누이 과정(Bernoulli process)이라고 하는데, 이로부터 여러 재미있는 응용문제가 파생됩니다. 예를 들어, 첫번째 성공($X=1$)이 발생할 때까지 베르누이 과정이 지속된다고 합시다. 몇 번의 실패 N이 수반되어야 할까요?

실패 수 $N=n$이기 위해서는 처음 n 시행이 모두 실패이고 그 다음은 성공이어야 합니다. 따라서

$$P(N=n) = (1-\theta)^n \theta, \quad n = 0, 1, 2, \cdots$$

가 유도됩니다. 이것을 기하분포 Geometric(θ)라고 합니다.

더 일반화하여, r번째 성공이 발생할 때까지 베르누이 과정이 지속된다고 합시다. 몇 번의 실패 N이 선행되어야 할까요? $N=n$이기 위해서는 처음 $n+r-1$ 회의 시행 결과 n번이 실패이고 $r-1$번이 성공이며, 그 다음엔 성공이어야 합니다. 따라서, $r = 1, 2, 3, \cdots$ 에 대하여

$$P(N=n) = \binom{n+r-1}{r-1} (1-\theta)^n \theta^{r-1} \cdot \theta, \quad n = 0, 1, 2, \cdots$$

가 유도됩니다.[1] 이것을 음이항분포(陰二項分布, negative binomial distribution), 기호로 NB(r,θ)라고 합니다. 그러므로 기하분포는 음이항분포의 한 특수한 경우입니다 ($r = 1$).

N을 NB(r,θ) 분포로부터의 확률변수, 즉 N이 r번째 성공이 일어나기까지의 실패 수라고 합시다. r번째 성공이란

첫번째 성공 → 두번째 성공 → ⋯ → r번째 성공

의 마지막 것인데, 두번째 성공은 첫번째 성공 이후의 첫번째 성공이고, ⋯, 이런 식으로 r번째 성공은 $r-1$번째 성공 이후의 첫번째 성공이므로, N이 다음과 같이 표현된다고 하겠습니다.

$$N = N_1 + N_2 + \cdots + N_r, \tag{6}$$

여기서 $N_1, N_2, \cdots, N_r$은 독립적인 Geometric(θ) 변수들입니다. 식 (6)을 이용하여 NB(r,θ) 변수 N의 확률적 행태를 쉽게 파악할 수 있습니다. 예를 들어 N의 기대값을 계산해보도록 하겠습니다.

$$E(N) = E(N_1) + E(N_2) + \cdots + E(N_r) = r\,E(N_1)$$

인데, Geometric(θ) 변수 N_1의 기대값이

$$E(N_1) = \sum_{n=0}^{\infty} n\,(1-\theta)^n \theta = (1-\theta)\,\theta \sum_{n=0}^{\infty} n\,(1-\theta)^{n-1}$$

$$= (1-\theta)\,\theta \frac{d}{d\theta}\{\sum_{n=0}^{\infty} -(1-\theta)^n\}$$

$$= (1-\theta)\,\theta \frac{d}{d\theta}\left\{-\frac{1}{\theta}\right\} = \frac{1-\theta}{\theta}$$

---

1) $\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$, $k = 0,1, \cdots, n-1, n$: $0! = 1$, $k! = 1 \cdots k$ (for $1 \le k \le n$).

이므로,

$$E(N) = \frac{r(1-\theta)}{\theta}$$

입니다. 이와 같은 방법으로 $Var(N)$을 구해볼 수 있습니다 (연습문제 1.3).

한 응용 예를 들어보겠습니다. 어느 지독히 남아선호적인 사회에서 남자아이 r명을 출산할 때까지 모든 부부가 계속 아이를 출산해야 한다고 합시다. 그것이 사회적으로 뿐만 아니라 생물적으로도 가능하다고 합시다. 순수한 자연 상태에서 남아의 출산비율이 θ라고 할 때 이 사회의 성비(性比)는 얼마일까요?

그 사회에서 남아의 출산이 '성공'입니다. 이런 성공이 r번 있을 때까진 실패를 거듭해야만 합니다. 그러므로 r명의 남자아이가 출산할 때까진 생긴 여자아이의 수 N은 음이항분포 NB(r,θ)를 따릅니다. 그러므로 기대되는 남·녀 성비는

$$r : E(N) = r : \frac{r(1-\theta)}{\theta} = \theta : (1-\theta)$$

입니다. 그러므로 순수한 남아선호만으로는 성비가 자연출산시의 성비인 $0.51 : 0.49$를 바꾸지 못합니다. 실제로 남하선호적인 사회에서 자연성비가 바뀌는 이유는 태아의 성별감식으로 남자아이의 출생확률 자체가 높기 때문입니다. 1997년도 통계에 따르면, 우리나라에서 첫째 아이의 성비는 $0.51 : 0.49$이지만, 둘째 아이는 $0.52 : 0.48$이고 셋째 아이의 경우엔 $0.58 : 0.42$로 높아집니다.

베르누이 과정에서 시행수를 유한값 $n\,(=1,2,3,\cdots)$으로 고정한 것이 바로 이항분포(二項分布, binomial distribution) B(n,θ)인데, 이것은 성공의 확률이 θ인 베르누이 시행을 n번 독립적으로 하였을 때 총 성공의 수 S가 따르는 확률법칙입니다:

$$P(S=s) = \binom{n}{s}\theta^s(1-\theta)^{n-s},\ 0<\theta<1,\ s=0,1,\cdots,n-1,n.$$

이항분포에 대하여는 고등학교에서 지겹게 많이 접해보았을 것이므로 여기서 되풀이하고 싶지는 않습니다. 하나만 덧붙이기로 하지요.

이항분포 변수 S는 독립적인 Bernoulli(θ) 변수들인 $X_1, \cdots, X_n$의 합, 즉

$$S = X_1 + \cdots + X_n$$

로 표현 가능합니다. 이것을 이용하면 S의 기대값과 분산이

$$E(S) = n\theta, \quad Var(S) = n\theta(1-\theta)$$

임을 쉽게 보일 수 있습니다. 왜냐하면

$$E(X_1) = \theta,\ E(X_1^2) = \theta,\ Var(X_1) = E(X_1^2) - \{E(X_1)\}^2 = \theta(1-\theta)$$

이기 때문입니다.

이항분포의 역할 중 중요한 한 가지는 이것이 오차(error)의 확률적 행태를 이해하는 데 중요한 실마리가 된다는 것입니다. 역사적으로 오차 문제는 천문학에서 나왔지만 다른 분야로도 금방 확산되었습니다. 오차에 대한 프랑스의 수학자 드무아브르(de Moivre, 1667-1754)의 1733년 생각을 살펴보기로 합시다.

천문관측에 있어서 오차는 여러 원인에서 나옵니다. 측정한 날의 온도나 습도, 기압 같은 기상적 원인으로부터도 나오고 측정자의 개인적 편향 때문에도 나올 수 있고 망원경의 기기적 불균일성으로부터도 나올 수 있습니다. 이런 각각의 원인에 의한 영향을 $D_1, \cdots, D_n$이라고 하고 각각이 -1과 +1의 값을 1/2의 확률로 취한다고 합시다 (매우 거친 가정이지요). 이 때, 총체적 오차를 뜻하는

$$S_n = D_1 + \cdots + D_n$$

의 확률분포를 구해보기로 하겠습니다. n이 홀수이면 S_n이 0이 될 수 없으므로 이하 n이 짝수라고 가정하겠습니다 (즉 $n = 2m$, m은 자연수). S_n이 s $(= 2k, k$는 $-m$과 m 사이의 정수)이기 위해서는 +1의 값을 취하는 D가 $(n+s)/2\,(=m+k)$개, -1의 값을 취하는 D가 $(n-s)/2\,(=m-k)$개여야 하므로

$$P\{S_n = s\} = \binom{n}{(n+s)/2}\frac{1}{2^n} = P\{S_n = 0\} \cdot \frac{\binom{n}{(n+s)/2}}{\binom{n}{n/2}}.$$

$$\therefore\ P\{S_{2m} = 2k\}/P\{S_{2m} = 0\} = \frac{\binom{2m}{m+k}}{\binom{2m}{m}} = \frac{(m!)^2}{(m+k)!\,(m-k)!}.$$

여기서

$$\frac{(m!)^2}{(m+k)!\,(m-k)!} = \frac{(m-k+1)(m-k+2)\cdots m}{(m+1)(m+2)\cdots(m+k)} \equiv A_m$$

으로 놓으면, 정해진 k에 대하여 m이 커짐에 따라

$$A_m \sim e^{-k^2/m}$$

로 근사됩니다. 왜냐하면, 작은 수 h에 대하여 $\log_e(1+h) \sim h$ 이므로

$$\begin{aligned}\log_e A_m &= \log(1-(k-1)/m) + \log(1-(k-2)/m) + \cdots + \log 1 \\ &\quad -\log(1+1/m) - \log(1+2/m) - \cdots - \log(1+k/m) \\ &\sim -\{(k-1)/m + (k-2)/m + \cdots + 0 + 1/m + 2/m + \cdots + k/m\} \\ &= -(1/m)\{\,k(k-1)/2 + k(k+1)/2\,\} = -k^2/m\ (= s^2/2n)\end{aligned}$$

이기 때문입니다. 따라서

$$P\{S_n = s\} / P\{S_n = 0\} \sim e^{-\frac{s^2}{2n}}$$

입니다. 여기서 스털링(Stirling)의 공식

$$n! \sim \sqrt{2\pi}\ n^{n+0.5} e^{-n}, \quad \text{큰 정수 } n \text{에 대하여}$$

를 적용하면, n이 짝수인 경우($n = 2m$)

$$P\{S_n = 0\} = \binom{2m}{m}\left(\frac{1}{2}\right)^{2m}$$

$$\sim \frac{\sqrt{2\pi}\ (2m)^{2m+0.5} e^{-2m}}{\sqrt{2\pi}\ m^{m+0.5} e^{-m} \cdot \sqrt{2\pi}\ m^{m+0.5} e^{-m}} \left(\frac{1}{2}\right)^{2m}$$

$$= \frac{1}{\sqrt{2\pi}}\left(\frac{2}{m}\right)^{0.5} = \frac{2}{\sqrt{2\pi\, n}}$$

입니다. 따라서, 짝수인 n과 $s\,(0 \le s \le n)$에 대하여

$$P\{S_n = s\} \sim \frac{2}{\sqrt{2\pi\, n}} e^{-\frac{s^2}{2n}}$$

입니다. 그런데, $P\{S_n = s\} = P\{s-1 \le S_n \le s+1\}$이기 때문에

$$P\{s-1 \le S_n \le s+1\}/2 \sim \frac{1}{\sqrt{2\pi\, n}} e^{-\frac{s^2}{2n}}, \quad \text{짝수 } n \text{과 } s \text{에 대하여}$$

즉, 정규분포의 한 형태인 $N(0,n)$이 유도되어 나오는군요 (위 식의 좌변을 2로 나눈 이유는 구간의 폭이 2이기 때문). 역사적으로는 이것이 정규분포에 대한 첫 번째 유도입니다.

1.5 포아송 분포(Poisson distribution)

시구간 $(0,t]$에서 몇 건의 사건이 터지는가를 세어보기로 합시다. 예컨대 불 자동차가 어느 특정구역에 출동하는 사건이라고 하지요. 여기서 관측되는 사건의 수를 확률변수 $N(t)$로 표기합니다 ($N(0) \equiv 0$).

이 때, 다음과 같은 세 가지 가정을 하기로 하겠습니다.

① 독립증분(獨立增分, independent increments)의 가정:

$t_0\,(=0) < t_1 < t_2 < \cdots < t_{n-1} < t_n$ 일 때,

$$N(t_1) - N(t_0),\ N(t_2) - N(t_1),\ \cdots,\ N(t_n) - N(t_{n-1})$$

은 독립이다.

② 정상증분(定常增分, stationary increments)의 가정:

$t > 0, s > 0$일 때, $k = 0, 1, 2, \cdots$ 에 대하여

$$P\{N(t) - N(0) = k\} = P\{N(s+t) - N(s) = k\}$$

이다. 즉 $N(s+t) - N(s)$의 확률분포는 시구간의 폭 t에만 관련 있을 뿐 시작시점 s와는 무관하다.

③ 미시적 확률현상에 관한 가정:

작은 시구간 $(0, h]$에서 발생하는 사건수 $N(h)$에 대하여

$$P\{N(h) = 1\} = \lambda h + o(h),$$

$$P\{N(h) \geq 2\} = o(h)$$

이다.[2] 따라서

$$P\{N(h) = 0\} = 1 - \lambda h + o(h).$$

이와 같은 세 가정하에서 $N(t)$는 확률분포 Poisson(λt)를 따르게 됩니다. 여기서 Poisson(θ) 분포의 확률함수는 다음과 같습니다.

$$P\{X = x\} = e^{-\theta}\frac{\theta^x}{x!}, \quad x = 0,1,2,\cdots \quad .$$

즉 $p_n(t) \equiv P\{N(t) = n\}$는 다음과 같습니다.

$$p_n(t) = e^{-\lambda t}\frac{(\lambda t)^n}{n!}, \; n = 0, 1, 2, \cdots \quad . \tag{7}$$

이제부터 식 (7)을 유도해 보이겠습니다. 그 과정이 수학적으로 매우 아름답기 때문에 여러분들도 몇 번쯤은 혼자서 되새겨보기 바랍니다.

<u>포아송 분포의 유도</u>

먼저 $p_0(t) \equiv P\{\mathrm{N(t)} = 0\}$에 대하여 풀어봅니다. $p_0(t+h)$가 시점 $t+h$까지 전혀 사건이 발생하지 않을 확률인데 그러기 위해서는 시구간 $(0, t]$와 $(t, t+h]$에서 모두 사건이 발생하지 않아야 합니다. 따라서

---

2) 여기서 "little" $o(h)$는 h에 비하여 상대적으로 매우 작은 함수를 말합니다. 공식적인 정의는 $\lim_{h \to 0}\frac{o(h)}{h} = 0$ 입니다. 예컨대 h^2는 $o(h)$입니다.

$$p_0(t+h) = p_0(t)\ P\{\mathrm{N(t+h)-N(t)=0}\} \quad (\because \text{ 가정 ①})$$
$$= p_0(t)\ p_0(h) \quad (\because \text{ 가정 ②})$$
$$= p_0(t)\ (1-\lambda h + o(h)) \quad (\because \text{ 가정 ③})$$

라고 할 수 있습니다. 그러므로

$$p_0(t+h) - p_0(t) = -\lambda h\ p_0(t) + o(h).$$
$$\therefore \quad \frac{1}{h}\{p_0(t+h) - p_0(t)\} = -\lambda\ p_0(t) + \frac{o(h)}{h}.$$
$$\Rightarrow \lim_{h\to 0} \frac{1}{h}\{p_0(t+h) - p_0(t)\} = -\lambda\ p_0(t) + \lim_{h\to 0}\frac{o(h)}{h}.$$
$$\Rightarrow \ p_0'(t) = -\lambda\ p_0(t),\ t \geq 0.$$

이것은 아주 초보적인 미분방정식인데 풀이가 무지하게 쉽습니다.

$$\frac{p_0'(t)}{p_0(t)} = -\lambda \quad \Rightarrow \quad \int_0^t \frac{p_0'(s)}{p_0(s)}\,ds = -\int_0^t \lambda\,ds$$
$$\Rightarrow \ \log_e[p_0(t)] - \log_e[p_0(0)] = -\lambda t$$
$$\Rightarrow \ p_0(t) = \exp(-\lambda t) \quad (\because\ p_0(0) = 1)$$

가 유도됩니다. 그러므로 $n=0$인 경우의 (7)은 증명이 된 것입니다.

이제부터는 수학적 귀납법을 쓰도록 하겠습니다. 즉 (7)이 n에 대하여 성립한다고 합시다. 그런데, $p_{n+1}(t) \equiv P\{N(t) = n+1\}$은

$$p_{n+1}(t+h) = p_{n+1}(t)\ (1-\lambda h) + p_n(t)\ (\lambda h) + o(h)$$

로 표현 가능합니다. 왜냐하면 $N(t+h) = n+1$이기 위하여는 $N(t) = n+1$이거나 $N(t) = n$이어야 하고 나머지 가능성은 거의 무시할 만하기 때문입니다. 따라서

$$\frac{p_{n+1}(t+h) - p_{n+1}(t)}{h} = -\lambda\ p_{n+1}(t) + \lambda\ p_n(t) + \frac{o(h)}{h}$$
$$\Rightarrow \lim_{h\to 0}\frac{p_{n+1}(t+h) - p_{n+1}(t)}{h} = -\lambda\ p_{n+1}(t) + \lambda\ p_n(t) + \lim_{h\to 0}\frac{o(h)}{h}$$
$$\Rightarrow \ p_{n+1}'(t) = -\lambda\ p_{n+1}(t) + \lambda\ p_n(t).$$

$p_n(t) = e^{-\lambda t}\frac{(\lambda t)^n}{n!}$ 임을 가정하였으므로

$$p_{n+1}'(t) = -\lambda\, p_{n+1}(t) + \lambda e^{-\lambda t}\frac{(\lambda t)^n}{n!}.$$

$$\Rightarrow\quad p_{n+1}'(t)\, e^{\lambda t} + p_{n+1}(t)\,\lambda\, e^{\lambda t} = \lambda\frac{(\lambda t)^n}{n!}.$$

$$\Rightarrow\quad \frac{d}{dt}\{p_{n+1}(t)\, e^{\lambda t}\} = \lambda\frac{(\lambda t)^n}{n!}.$$

$$\Rightarrow\quad p_{n+1}(t)\, e^{\lambda t} = \int_0^t \lambda\frac{(\lambda s)^n}{n!}\,ds = \frac{(\lambda t)^{n+1}}{(n+1)!}.$$

이 됩니다 ($\because\ p_{n+1}(0) = 0$). 그러므로

$$p_{n+1}(t) = e^{-\lambda t}\frac{(\lambda t)^{n+1}}{(n+1)!}$$

을 얻게 됩니다. 이것으로써, 수학적 귀납법에 의하여, (7)이 모든 비음의 정수 n에 대하여 성립함을 보인 것입니다. ■

우리가 앞서 유도한 지수분포도 사실은 포아송 과정으로부터 끌어낼 수 있습니다. T를 첫 번째 사건이 발생하는 시점이라고 합시다. 그런데 $T > t$ 라는 것은 시구간 $(0,t]$에 아무 사건이 발생하지 않았음, 즉 $N(t) = 0$임을 의미합니다. 따라서

$$P(T > t) = P(N(t) = 0) = e^{-\lambda t},\ t \geq 0$$

입니다. 그러므로

$$P(T \leq t) = 1 - e^{-\lambda t},\ t \geq 0,$$

$$\text{즉}\quad f_T(t) = \frac{d}{dt}P(T \leq t) = \lambda e^{-\lambda t},\ t \geq 0$$

인 것입니다. 따라서 첫 발생시간 T는 Exponential$(1/\lambda)$를 따릅니다. 분포와 분포 사이에 이렇게 여러 관계가 얽혀 있습니다. 이런 것을 수학적 재미라고 하지요.

앞의 결과를 조금 확장할 수도 있습니다. T_n을 n 번째 사건이 발생하는 시점이라고 합시다. 그런데 $T_n > t$ 라는 것은 시구간 $(0,t]$에 최대 $n-1$개의 사건이 발생하였음, 즉 $N(t) \leq n-1$임을 의미합니다. 따라서

$$P(T_n > t) = P(N(t) \leq n-1) = \sum_{k=0}^{n-1} e^{-\lambda t}\frac{(\lambda t)^k}{k!},\ t \geq 0$$

입니다. 그러므로

$$P(T_n \leq t) = 1 - \sum_{k=0}^{n-1} e^{-\lambda t}\frac{(\lambda t)^k}{k!},\ t \geq 0,$$

$$\therefore \quad f_{T_n}(t) = \frac{d}{dt} P(T_n \le t) = -\frac{d}{dt}\sum_{k=0}^{n-1} e^{-\lambda t}\frac{(\lambda t)^k}{k!}$$

$$= \lambda e^{-\lambda t} + \sum_{k=1}^{n-1} \lambda e^{-\lambda t}\frac{(\lambda t)^k}{k!} - \sum_{k=1}^{n-1} \lambda e^{-\lambda t}\frac{(\lambda t)^{k-1}}{(k-1)!}$$

$$= \lambda e^{-\lambda t}\frac{(\lambda t)^{n-1}}{(n-1)!}, \quad t \ge 0$$

입니다. T_n의 분포를 이후에 감마분포 Gamma$(n, 1/\lambda)$라고 하게 될 것입니다. 지수분포 Exponential$(1/\lambda)$은 $n=1$인 Gamma$(n, 1/\lambda)$ 분포입니다.

포아송 분포는 이항분포와도 관계가 있습니다. 변수 X가 이항분포 B(n,θ)를 따른다고 합시다. 그런데 n이 상당히 크고 θ가 상당히 작다고 합시다. 예컨대 $n=365$이고 $\theta=1/365$라고 합시다. 이런 경우

$$P(X=x) = \binom{n}{x}\theta^x(1-\theta)^{n-x}, \quad x=0,1,2,\cdots,n$$

을 계산하는 것은 결코 쉬운 일이 아닙니다 (컴퓨터로 하더라도). 한 대안적 방법은 다음과 같습니다.

먼저 $\lambda = n\theta$를 고정시킵니다. 그리고, 구간$(0,1]$을 n등분함으로써 폭 $1/n$인 작은 구간$(k/n, (k+1)/n]$을 얻습니다. 이 구간에서 성공 (1개의 사건 발생) 확률을 $\theta(=\lambda/n)$라고 합시다. 2개 이상의 사건이 발생할 확률은 무시합니다. 그러면 구간 $(0,1]$에서 발생할 총 사건의 수 X는 이항분포 B(n,θ)를 따릅니다. 그런데 n이 커지고 구간의 폭 $1/n$이 작아짐에 따라 X는 포아송 분포 Poisson(λ)를 따른다고 볼 수 있습니다. 따라서 이항분포 B(n,θ)는 포아송 분포 Poisson(λ)로 근사됩니다 ($n\theta=\lambda$이며 n이 상당히 크고 θ가 상당히 작은 경우. 연습문제 1.4 참조).

간단한 연습문제를 하나 풀어 봅시다. 회원이 366명의 고등학교 동창회에서 생일이 내 생일과 같은 사람이 나 빼고 모두 몇 명 있을까요? 정확히 계산하려면, $n=365$이고 $\theta=1/365$인 이항분포 B(n,θ)를 생각해야 합니다. 또는 앞서 근거를 제시한대로 포아송 분포 Poisson(1)을 근사하는 것도 요령입니다. 다음 수치적 결과를 보십시오 (연습문제 1.5).

X =	0	1	2	3	4	5+
이항분포	0.3674	0.3684	0.1842	0.0612	0.0152	0.0036
포아송 분포	0.3679	0.3679	0.1839	0.0613	0.0153	0.0037

1.A 연습문제

1.1 X_1과 X_2를 Uniform(0,1) 분포로부터의 독립적인 확률변수라고 할 때, $X_1 + X_2$의 확률분포를 구하세요..

답 : $f(x) = \begin{cases} x, & 0 \le x \le 1 \\ 2-x, & 1 \le x \le 2 \end{cases}$

1.2 다음 조건을 만족하는 연속형 확률분포의 밀도함수를 유도하세요..

$$\frac{f(x)}{1-F(x)} = \alpha\, x^{\gamma-1}, \quad \alpha > 0, \gamma > 0 \text{은 상수}, \quad x \ge 0. \tag{8}$$

답: $f(x) = \frac{\gamma}{\beta}\, x^{\gamma-1} \exp\left(-\frac{1}{\beta}\, x^{\gamma}\right), \quad \gamma > 0, \quad \beta\,(=\gamma/\alpha) > 0, \quad x \ge 0.$ (9)

[Comment: 문제에서 주어진 식의 좌변을 생존분석에서 위험함수(hazard function)라고 합니다. $\gamma = 1$인 경우, 즉 위험함수가 상수인 경우 지수분포가 유도됩니다. 따라서 (8)로부터 생성되는 분포는 지수분포의 확장입니다. 분포 (9)를 와이블 분포(Weibull distribution) Weibull(γ,β)라고 합니다 ($\beta = \gamma/\alpha$).]

1.3 음이항분포 NB(r,θ)를 따르는 확률변수 N의 분산을 구하세요.

답: $Var(N) = \frac{r(1-\theta)}{\theta^2}$.

1.4 이항분포 B(n,θ)에서 $n \to \infty$, $p \to 0$ (such that $np = \theta, \theta$는 고정상수)에 따라 이항분포의 확률함수 $f(x) = \binom{n}{x} p^x (1-p)^{n-x}$가 Poisson($\theta$) 분포의 확률함수로 수렴함을 증명하세요.

1.5 이항분포 B(365, 1/365)의 확률과 포아송분포 Poisson(1)의 확률을 직접 계산하여 비교하세요.

[Comment: 여러분이 아는 컴퓨터 언어로 프로그램을 짜보세요. 이 때 중간과정에서 가급적 $\infty \times 0$ 형의 연산이 나타나지 않도록 하여야 합니다. 답은 본문중에 있습니다.]

1.B 읽을만한 책

확률에 대한 공부가 일반적으로 미진하거나 더 관심이 많다면 다음 책을 보기 바랍니다.

- 전종우・손건태 (2000)「확률의 개념 및 응용」 자유 아카데미.
- Ross, S. (1998) *A First Course in Probability*, Fifth Edition. Prentice Hall.

임의수 발생(random number generation)에 관하여는 다음 책을 보기 바랍니다.

- 손건태 (1996)「전산통계개론」 자유 아카데미. (3장)
- 최영훈・이승천 (1995)「C에 의한 전산통계」 자유 아카데미. (5장)

<그림 4>와 <그림 6>은 수학계산 소프트웨어인 Mathematica를 써서 얻은 것입니다. Maple도 유사하다고 합니다. 꼭 그런 소프트웨어들이 있어야 3-D 그림을 그릴 수 있는 것은 아닙니다. SAS 같은 것으로도 됩니다. 하지만 고급 수학 계산을 하려면 Mathematica나 Maple이 좋지요. Mathematica에 관한 설명서로는 다음 책이 있습니다. 적당히 두껍기 때문에 낮잠 잘 때 베개로 쓰기에 안성맞춤입니다.

- Wolfram, S. (1996) *The Mathematica Book*, 3rd Edition. Cambridge University Press.