2장. 기본적인 확률도구 5개

목수가 솜씨를 발휘하려면 좋은 도구 몇 개가 필요하듯이 우리가 확률을 다루는 데도 최소한 도구 몇 개는 필요합니다. 확률변수와 확률분포를 이해하는 데 쓸만한 다섯 개의 도구(tool)를 소개하기로 합니다.

그 중 첫 번째의 적률생성함수는 확률분포를 특성화하는 도구입니다. 언듯 보면 별 관계가 없어 보이는 이상한 함수인데 사실은 그 안에 확률분포에 관한 모든 것이 들어 있습니다. 뿐만 아니라, 독립적인 변수들의 합이 어떤 분포를 따르는가를 밝히는 데도 유용하답니다.

대수의 법칙과 중심극한정리 등 극한이론은 관측변수의 수가 커짐에 따라 관측변수들의 평균이 확률적으로 어떻게 되는가에 관한 탐구입니다. 표본평균이 모(母)평균으로 수렴해간다는 대수의 법칙은 일견 당연해 보이지만, 당연해 보이는 것이 당연하다는 것을 논증해내야 하는 것이 수학자의 역할이기도 하지요. 한편, 중심극한정리는 사실 매우 놀라운 결과를 담고 있습니다. 만약 오차가 여러 독립적 원인들의 가법적 합성으로 간주될 수 있다면 그런 오차는 필연적으로 정규분포를 따르게 된다는 내용이니까요. 이것을 증명하는 데 적률생성함수가 요긴하게 쓰일 것입니다. 또한 변수변환 기법과 적률함수를 써서 t-분포, F-분포, 감마 분포 등을 유도해보겠습니다.

몬테칼로(Monte Carlo)란 도박으로 유명한 도시이름인데 도박도 밤낮으로 하면 도박에 관한 확률문제가 감(感)으로 대충 풀어질 것입니다. 수학적 확률문제의 풀이에도 이런 방식으로 접근해볼 수 있다는 것이 몬테칼로 기법입니다. 컴퓨터의 계산능력을 최대한 활용하게 될 것입니다.

마지막으로 유한과 무한에서는 정확성과 근사성을 탐구합니다. 유한에서 무한으로 가는 길목에서 에지워스 근사를 만나게 될 것이고 굽은 길에선 델타 방법으로 확률적 행태가 어떻게 변형되는지를 예견할 수 있답니다. 기대해보세요.

차례:

2.1 적률생성함수(moment generating function, mgf)

적률생성함수(積率生成函數)란 수학적으로는 라플라스 변환(Laplace transform)입니다. 정의는 다음과 같이 되지요.

확률분포가 연속형인 경우 밀도함수 $f_X(x)$에 대한 적률생성함수(mgf)는

$$m_X(t) = \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx$$

로 정의되고, 확률분포가 이산형인 경우엔 확률함수 $f_X(x)$에 대한 mgf가

$$m_X(t) = \sum_x e^{tx} f_X(x)$$

로 정의됩니다. 단, 적분이나 합이 $t=0$을 포함하는 개구간(開區間, open interval)에서 존재하여야 의미가 있습니다.

먼저, 적률생성함수의 계산을 몇 개 해보기로 합니다.

1) 균일분포 Uniform$(0,1)$에 대한 mgf $m_X(t)$를 계산해 볼까요?

$$m_X(t) = \int_0^1 e^{tx}\, dx = \begin{cases} \left[\dfrac{e^{tx}}{t}\right]_0^1 = \dfrac{e^t-1}{t}, & \text{if } t \neq 0, \\ \displaystyle\int_0^1 1\, dx = 1, & \text{if } t = 0. \end{cases}$$

$e^t = 1 + t + t^2/2 + \cdots$ 이므로 $m_X(t)$는 $t=0$에서 연속입니다. 일반적으로, 균일분포 Uniform$(0,\theta)$에 대한 mgf $m_X(t)$는 다음과 같습니다.

$$m_X(t) = \begin{cases} \dfrac{e^{t\theta}-1}{t\theta}, & \text{if } t \neq 0, \\ \displaystyle\int_0^\theta \frac{1}{\theta}\, dx = 1, & \text{if } t = 0. \end{cases}$$

증명은 각자 해보세요.

2) 지수분포 Exponential(1)에 대한 mgf $m_X(t)$

$$m_X(t) = \int_0^\infty e^{tx} e^{-x}\, dx = \int_0^\infty e^{-(1-t)x}\, dx = \left[-\frac{e^{-(1-t)x}}{1-t}\right]_0^\infty$$

$$= \frac{1}{1-t}, \text{ for } t < 1.$$

$t \geq 1$에 대하여는 $m_X(t)$가 존재하지 않습니다. 그러나 별 문제 없어요. $t=0$ 근처에서만 존재하면 됩니다.

일반적으로 지수분포 Exponential(θ)에 대한 mgf $m_X(t)$는 다음과 같습니다.

$$m_X(t) = \frac{1}{1-t\theta}, \text{ for } t < \frac{1}{\theta}.$$

3) 정규분포 N(0,1)에 대한 mgf $m_X(t)$

$$\begin{aligned} m_X(t) &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x-t)^2-t^2}{2}\right\} dx \\ &= \exp\left(\frac{t^2}{2}\right) \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(x-t)^2}{2}\right\} dx \\ &= \exp\left(\frac{t^2}{2}\right), \quad \text{for } -\infty < t < \infty. \end{aligned}$$

일반적으로, 정규분포 N(μ,σ^2)에 대한 mgf $m_X(t)$은 다음과 같습니다.

$$m_X(t) = \exp\left(t\mu + \frac{t^2\sigma^2}{2}\right), \quad \text{for } -\infty < t < \infty.$$

4) 베르누이 분포 Bernoulli(θ)에 대한 mgf $m_X(t)$

$$m_X(t) = \sum_{x=0,1} e^{tx} f_X(x) = (1-\theta) + e^t\theta, \text{ for } -\infty < t < \infty.$$

기하분포 Geometric(θ)에 대한 mgf $m_X(t)$

$$\begin{aligned} m_X(t) &= \sum_{x=0}^{\infty} e^{tx} (1-\theta)^x \theta \\ &= \sum_{x=0}^{\infty} \{e^t(1-\theta)\}^x \theta \\ &= \frac{\theta}{1-e^t(1-\theta)}, \quad \text{for } t < -\log_e(1-\theta). \end{aligned}$$

음이항분포 NB(r,θ)에 대한 mgf $m_X(t)$

$$\begin{aligned} m_X(t) &= \sum_{x=0}^{\infty} e^{tx} \binom{x+r-1}{r-1} (1-\theta)^x \theta^r \\ &= \sum_{x=0}^{\infty} \binom{x+r-1}{r-1} \{e^t(1-\theta)\}^x \theta^r \end{aligned}$$

$$= \left\{ \frac{\theta}{1 - e^t (1-\theta)} \right\}^r, \quad \text{for } t < -\log_e (1-\theta).$$

이항분포 B(n,θ)에 대한 mgf $m_X(t)$

$$m_X(t) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

$$= \sum_{x=0}^{n} \binom{n}{x} (e^t \theta)^x (1-\theta)^{n-x}$$

$$= (e^t \theta + 1 - \theta)^n, \quad \text{for } -\infty < t < \infty.$$

5) 포아송 분포 Poisson(θ)에 대한 mgf $m_X(t)$

$$m_X(t) = \sum_{x=0}^{\infty} e^{tx} e^{-\theta} \frac{\theta^x}{x!}$$

$$= e^{-\theta} \sum_{x=0}^{\infty} \frac{(e^t \theta)^x}{x!}$$

$$= e^{-\theta} \exp(e^t \theta)$$

$$= \exp\{(e^t - 1)\theta\}, \quad -\infty < t < \infty.$$

적률생성함수 $m_X(t)$의 첫 번째 가치는 말뜻 그대로 이것이 적률(積率, moment) $E(X^k), k = 1, 2, 3, \cdots$ 를 생성해낸다는 데 있습니다. 그 이치는 다음과 같습니다.

$$m_X'(t) = \frac{d}{dt} m_X(t) = \frac{d}{dt} \int_{-\infty}^{\infty} e^{tx} f_X(x)\, dx = \int_{-\infty}^{\infty} \frac{d}{dt} e^{tx} f_X(x)\, dx$$

$$= \int_{-\infty}^{\infty} x\, e^{tx} f_X(x)\, dx$$

입니다 (이산형 분포의 경우는 적분 $\int$ 대신 합 $\sum$로 대치하면 됩니다). 따라서

$$m_X'(0) = \int_{-\infty}^{\infty} x\, f_X(x)\, dx = E(X)$$

입니다. 같은 방법으로

$$m_X''(0) = \int_{-\infty}^{\infty} x^2 f_X(x)\, dx = E(X^2)$$

을 얻게 되고, 아주 확장하면

$$m_X^{(k)}(0) = \int_{-\infty}^{\infty} x^k f_X(x)\, dx = E(X^k), \quad k = 1, 2, 3, \cdots$$

을 얻습니다. 그러므로

$$Var(X) = E(X^2) - E(X)^2 = m_X''(0) - \{m_X'(0)\}^2$$

입니다. 이런 방법으로 균일분포, 지수분포, 정규분포, 베르누이 분포 (지수분포, 음이항분포, 이항분포 포함), 포아송 분포의 기대값과 분산 등을 구할 수 있습니다.

여기서는 정규분포 $N(0,1)$의 경우만 기대값 μ, 분산 σ^2, 왜도(歪度, skewness) ν_3, 첨도(尖度. kurtosis) ν_4등을 계산해보도록 하겠습니다.

<u>$N(0,1)$의 기대값, 분산, 왜도, 첨도</u>

$m_X(t) = \exp\left(\frac{t^2}{2}\right)$이므로

$$m_X'(t) = t\exp\left(\frac{t^2}{2}\right),$$

$$m_X''(t) = \exp\left(\frac{t^2}{2}\right) + t^2\exp\left(\frac{t^2}{2}\right),$$

$$m_X^{(3)}(t) = 3t\exp\left(\frac{t^2}{2}\right) + t^3\exp\left(\frac{t^2}{2}\right),$$

$$m_X^{(4)}(t) = 3\exp\left(\frac{t^2}{2}\right) + 6t^2\exp\left(\frac{t^2}{2}\right) + t^4\exp\left(\frac{t^2}{2}\right).$$

따라서 $E(X)=0$, $E(X^2)=1$, $E(X^3)=0$, $E(X^4)=3$입니다. 그러므로

$$\mu = 0,\ \sigma^2 = 1,$$

$$\nu_3 = \frac{E(X-\mu)^3}{\sigma^3} = 0,\ \nu_4 = \frac{E(X-\mu)^4}{\sigma^4} - 3 = 0$$

입니다. ■

적률생성함수 $m_X(t)$의 두 번째 가치는 첫 번째 가치보다 더 중요하다고 할 수 있습니다. 그것은 확률(밀도)함수 $f_X(x)$와 적률생성함수 $m_X(t)$가 1:1 대응한다는 사실입니다. 이것의 증명은 이 책의 수준을 훨씬 넘으므로 아쉽게도 여기서는 증명할 수 없습니다. 그러나 이 사실을 빈번히 사용할 것이므로 믿어 두는 것이 좋겠습니다.

이런 목적으로 적률생성함수를 응용하는 대표적인 예는 소위 독립동일분포(iid, independently identically distributed) 변수들의 합의 분포를 구하는 것입니다. X_1, $X_2, \cdots, X_n$이 독립적으로 확률(밀도)함수 $f_X(x)$로부터 생성되고 그것에 대한 적률생성

함수를 $m_X(t)$ 라고 한다면,

$$S = X_1 + X_2 + \cdots + X_n$$

에 대한 적률생성함수는

$$\begin{aligned} m_S(t) &= E[\exp(tS)] = E[\exp(tX_1 + tX_2 + \cdots + tX_n)] \\ &= E[\exp(tX_1)\exp(tX_2)\cdots\exp(tX_n)] \\ &= E[\exp(tX_1)]E[\exp(tX_2)]\cdots E[\exp(tX_n)] \\ &= m_{X_1}(t)\, m_{X_2}(t)\cdots m_{X_n}(t) = \{m_{X_1}(t)\}^n \end{aligned}$$

이 됩니다. 따라서 우변이 어떤 알려진 mgf인 경우에는 S의 분포를 알게 되는 것이지요. 예를 들어 $X_1, X_2, \cdots, X_n$이 독립적으로 포아송 분포 Poisson(θ)를 따른다고 합시다. 즉

$$m_{X_1}(t) = \exp\{(e^t - 1)\theta\}, \quad -\infty < t < \infty.$$

$$\therefore \quad m_S(t) = \{m_{X_1}(t)\}^n = \exp\{(e^t - 1)n\theta\}.$$

우변이 Poisson$(n\theta)$의 mgf이므로, $S = X_1 + X_2 + \cdots + X_n$이 Poisson$(n\theta)$를 따른다는 것을 알 수 있습니다 (이 문제에서는 어렵지 않게 직접 세 개의 포아송 가정으로부터도 S의 분포를 짐작해낼 수 있기도 합니다).

다음 문제는 사소하지 않은 예입니다. $X_1, X_2, \cdots, X_n$이 독립적으로 지수 분포 Exponential(1)을 따른다고 합시다. 즉

$$m_{X_1}(t) = \frac{1}{1-t}, \text{ for } t < 1.$$

$$\therefore \quad m_S(t) = \{m_{X_1}(t)\}^n = \frac{1}{(1-t)^n}, \quad t < 1. \tag{1}$$

그런데 문제는 우변이 어느 확률밀도함수의 mgf인지 모른다는 것이지요. 적어도 앞에 유도해놓은 mgf에는 이와 비슷한 것이 없습니다. 그러나 1.5절에서 잠시 소개하였던 Gamma$(n,1)$ 분포의 mgf를 구해보도록 해보지요.

감마분포 Gamma$(n,1)$에 대한 mgf $m_X(t)$

$$\begin{aligned} m_X(t) &= \int_0^\infty e^{tx} e^{-x} \frac{x^{n-1}}{(n-1)!}\, dx \\ &= \int_0^\infty e^{-(1-t)x} \frac{x^{n-1}}{(n-1)!}\, dx \end{aligned}$$

$$= \int_0^\infty e^{-y} \frac{y^{n-1}}{(1-t)^{n-1}(n-1)!} \frac{dy}{1-t}, \quad \text{for } 1-t > 0$$

$$= \frac{1}{(1-t)^n}, \quad \text{for } t < 1.$$

$t \geq 1$에 대하여는 $m_X(t)$를 정의하지 않습니다.

마침, Gamma$(n,1)$의 mgf가 (1)과 일치하는군요. 따라서 지수분포 Exponential(1)의 iid 변수들의 합 S는 Gamma$(n,1)$ 분포를 따릅니다.

일반적으로, 감마분포 Gamma(α,θ)의 밀도함수와 적률생성함수는 $m_X(t)$는 다음과 같습니다.

$$f_X(x) = \frac{1}{\Gamma(\alpha)} \frac{x^{\alpha-1}}{\theta^\alpha} \exp\left(-\frac{x}{\theta}\right), \quad \alpha > 0,\ \theta > 0,\ x > 0.$$

$$m_X(t) = \frac{1}{(1-t\theta)^\alpha}, \quad \text{for } t < \frac{1}{\theta}.$$

2.2 극한이론 (limit theory) - 대수의 법칙과 중심극한정리 -

$X_1, X_2, \cdots, X_n$을 어떤 확률분포 F로부터의 iid 변수들이라고 합시다. 이때 통계량 $T_n = T(X_1, X_2, \cdots, X_n)$의 확률적 행태를 이해하는 한 방법으로서 n이 무한히 커지는 경우 T_n이 어떻게 될 것인가를 보자는 것이 극한이론입니다. 예를 들어

$$T_n = \overline{X}$$

인 경우, n이 무한히 커짐에 따라

$$\overline{X} \rightarrow \mu \qquad \text{(대수의 법칙)},$$
$$\sqrt{n}\,(\overline{X} - \mu)/\sigma \rightarrow N(0,1) \qquad \text{(중심극한정리)}$$

이 됩니다. 여기서 μ는 $E(X_1)$이고 $\sigma^2 = E\{(X_1 - \mu)^2\}$입니다. 수렴에 관한 화살표 '→'의 의미는 앞으로 명확히 하겠습니다.

약대수의 법칙 (弱大數의 法則, weak law of large numbers)

$X_1, X_2, \cdots, X_n$을 평균이 μ이고 분산이 σ^2인 확률분포 F로부터의 iid 변수들이라고 하면, $\overline{X}$는 μ에 확률적으로(in probability) 수렴합니다. 즉,

$$\lim_{n\to\infty} P_n\{|\overline{X} - \mu| \geq \epsilon\} = 0 \text{ , 임의의 } \epsilon > 0 \text{에 대하여}$$

입니다.

이것의 증명은 다음과 같습니다. 체비세프(Chebyshev) 부등식에 의하여

$$P_n\{\mid \overline{X}-\mu \mid \geqq \epsilon\} \leqq \frac{1}{\epsilon^2} E\{(\overline{X}-\mu)^2\} = \frac{1}{\epsilon^2}\frac{\sigma^2}{n}$$

인데, n이 무한히 커짐에 따라 위 식의 우변이 0으로 수렴하므로 좌변도 0으로 수렴하게 됩니다. ■

이 정리는 그 자체로서 중요합니다. 예를 들어 확률변수 $X_1, X_2, \cdots, X_n$이 지수분포 Exponential(θ)로부터 독립적으로 생성된다고 합시다. 이 분포의 평균은 θ입니다. 따라서 약대수의 법칙은, 매우 큰 표본에서는, 표본평균 $\overline{X}$가 θ로부터 약간이라도 멀리 떨어질 확률이 거의 0임을 말해줍니다. 그러므로 관측을 무한히 계속하면 $\overline{X}$는 θ에 가깝게 될 것입니다 (여기서 '가까운' 거리와 '약간이라도 먼' 거리의 경계는 ϵ입니다). 그러므로 $\overline{X}$를 θ에 대한 추정량으로 쓸 수 있는 것이지요.

약(弱)대수의 법칙이 있으면 강(强)대수의 법칙이 있지 않겠습니까? 물론이지요

강대수의 법칙 (强大數의 法則, strong law of large numbers)

$X_1, X_2, \cdots, X_n$을 평균이 μ이고 분산이 σ^2인 확률분포 F로부터의 iid 변수들이라고 하면, $\overline{X}$는 μ에 거의 확실히(almost surely) 수렴합니다. 즉,

$$P\left\{\lim_{n\to\infty} \overline{X} = \mu\right\} = 1$$

입니다. 무한히 큰 표본에서는, 거의 틀림없이 $\overline{X}$는 μ에 수렴함을 말합니다.

이 법칙의 증명은 고급 확률론에서나 가능하기 때문에 여기서는 다루지 않겠습니다. [사실은 F가 유한한 σ^2을 가져야 한다는 조건이 필요 없습니다.]

약대수의 법칙과 강대수의 법칙을 모두 앞으로는 그냥 대수의 법칙(law of large numbers)이라고 하겠습니다. 대수의 법칙의 응용 예를 하나 제시하기로 합니다.

몬테칼로 계산

(X_1, X_2)가 2변량 정규분포 BN$(0,0,1,1,\rho)$를 따른다고 합시다. 즉 밀도함수가

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

와 같습니다. 이 때 $h(\rho) = P\{X_1 \geq 1, X_2 \geq 1\}$ 을 계산하고 싶다고 합시다. 이 경우에는

$$\int_1^\infty \int_1^\infty f_{X_1,X_2}(x_1,x_2)\,dx_2\,dx_1 \tag{1}$$

을 해석적으로 계산하는 것은 가능하지 않습니다. 어떻게 해야 할까요?

한 방법은 2변량 정규분포 $\mathrm{BN}(0,0,1,1,\rho)$ 로부터 n 쌍의 유사난수

$$(X_{1i}, X_{2i}),\ i = 1,\cdots,n$$

을 컴퓨터로 생성시켜

$$p = \frac{1}{n} \sum_{i=1}^{n} 1\{X_{1i} \geq 1, X_{2i} \geq 1\}$$

을 $h(\rho)$ 에 대한 추정량으로 쓰는 것이겠습니다. 왜냐하면 대수의 법칙에 의하여 p 가

$$E\{1(X_1 \geq 1, X_2 \geq 1)\} = P\{X_1 \geq 1, X_2 \geq 1\} = h(\rho)$$

에 수렴할 것이기 때문입니다. 예를 들어 $\rho = 0.5$ 인 경우 $h(0.5)$ 를 n = 10,000번의 몬테칼로(Monte-Carlo) 시행을 통해 추정해본 결과 다음과 같이 결과가 나왔습니다: $p = 0.064$. 이후 2.4절에서 이런 방법에 대하여 구체적으로 다루기로 하겠습니다. (다른 한 방법은 정적분 (1)을 수치적으로 계산하는 것입니다.)

다음으로 중심극한 정리를 설명하기로 합니다. $X_1, X_2, \cdots, X_n$ 을 평균이 μ 이고 분산이 σ^2 인 확률분포 F 로부터의 iid 변수들이라고 하면,

$$E(\overline{X}) = \mu, \quad Var(\sqrt{n}\,(\overline{X} - \mu)/\sigma) = 1$$

이 모든 n 에 대하여 성립합니다. 여기서

$$T_n = \sqrt{n}\,(\overline{X} - \mu)/\sigma$$

의 분포가 어떻게 될 것인가를 생각해 보기로 합시다. 사실 T_n 의 분포는 모분포인 F 에 따라 다르므로 의미 있는 결론은 나오지 않습니다. 그러나 T_n 의 극한분포는 모분포 F 에 관계없이 표준정규분포 $\mathrm{N}(0,1)$ 이 된다는 것이 바로 중심극한정리, 즉 으뜸이 되는 극한정리입니다.

중심극한정리(中心極限定理, central limit theorem; CLT)

$X_1, X_2, \cdots, X_n$을 평균이 μ이고 분산이 σ^2인 확률분포 F로부터의 iid 변수들이라고 하면,

$$\lim_{n\to\infty} P\{\sqrt{n}\,(\overline{X}-\mu)/\sigma \le z\} = \Phi(z)$$

입니다. 여기서 $\Phi(z)$는 정규분포 N(0,1)의 분포함수, 즉

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)\, du$$

입니다. 이것은 n이 무한히 커짐에 따라 $T_n = \sqrt{n}\,(\overline{X}-\mu)/\sigma$의 분포가 표준정규분포에 수렴함을 의미합니다. 그러므로 큰 표본에서는 T_n의 분포를 표준정규분포로 간주할 수 있지요.

이 정리를 증명하는 가장 쉬운 방법은 T_n의 적률생성함수를 구한 다음, 그것의 극한이 N(0,1) 분포의 적률생성함수임을 보이는 간접적인 것입니다. 어렵지 않으니까 다음 증명을 따라오세요.

$Y_1 = (X_1-\mu)/\sigma$에 대한 적률생성함수를 $m_{Y_1}(t)$라고 합시다. 그런데 Y_1의 평균이 0, 분산이 1이므로

$$m_{Y_1}(t) = 1 + t^2/2 + o(t^2)$$

으로 나타내집니다. $T_n = n^{-1/2}\sum_{i=1}^{n} Y_i$이므로, 따라서

$$\begin{aligned} m_{T_n}(t) &= E\{\exp(t\,T_n)\} = \left\{m_{Y_1}(n^{-1/2}t)\right\}^n \\ &= \{1 + (n^{-1/2}t)^2/2 + o((n^{-1/2}t))^2\}^n \\ &= \{1 + n^{-1}(t^2/2) + o(n^{-1})\}^n \end{aligned}$$

이 됩니다. 그리고 이것의 극한은

$$\lim_{n\to\infty} m_{T_n}(t) = \lim_{n\to\infty} \{1 + n^{-1}(t^2/2) + o(n^{-1})\}^n = \exp(t^2/2)$$

입니다. 그런데, 위 식의 우변은 N(0,1) 분포의 적률생성함수입니다. 그러므로 T_n의 분포는 N(0,1) 분포로 수렴합니다. ■

중심극한정리(CLT)는 표본평균 $\overline{X}$의 분포가 대략 $N(\mu, \sigma^2/n)$이라는 것을 의미합니다. 물론 이 근사는 표본크기 n이 클수록 정확해집니다.

중심극한정리의 적용 예를 1개 들기로 하겠습니다. $X_1, X_2, \cdots, X_n$을 균일분포 Uniform$(-1,1)$로부터의 독립적인 확률변수들이라고 합시다. 이 분포의 평균과 분산은 각각 $\mu = 0$, $\sigma^2 = 1/3$입니다. 그러므로 중심극한정리(CLT)에 의하여

$$T_n = \sqrt{n}\,(\overline{X} - \mu)/\sigma = \sqrt{n}\;\overline{X} \;/\; \sqrt{1/3} = \sqrt{3n}\;\overline{X} = \sqrt{3}\;n^{-1/2}\sum_{i=1}^{n} X_i$$

의 분포는 n이 커짐에 따라 점점 N(0,1) 분포에 가까워집니다. 즉

$$n^{-1/2}\sum_{i=1}^{n} X_i \text{ 이 } N(0, 1/3) \text{로 분포수렴한다}$$

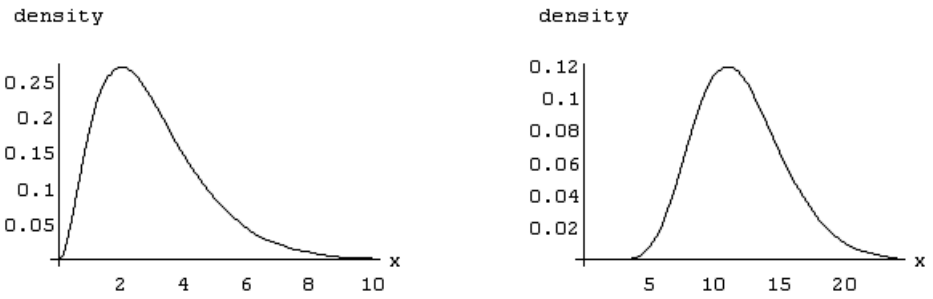
고 하겠습니다. 여기서 우리는 정규분포(normal distribution)를 다시 한 번 유도한 셈이 되었습니다.

다른 한 예를 들기로 합니다. $X_1, X_2, \cdots, X_n$이 지수 분포 Exponential(1)을 독립적으로 따른다고 합시다. Exponential(1) 분포의 평균과 분산은 모두 1입니다. 즉 $\mu = \sigma = 1$. CLT에 따르면

$$T_n = \sqrt{n}\,(\overline{X} - \mu)/\sigma = \sqrt{n}\,(\overline{X} - 1) = n^{-1/2}\sum_{i=1}^{n}(X_i - 1)$$

이 근사적으로 N(0,1) 분포를 따르게 됩니다. 이에 따라, $S = \sum_{i=1}^{n} X_i$는 근사적으로 평균이 n, 분산이 n인 정규분포를 따른다고 할 수 있습니다. 그런데 바로 앞 절에서는 S가 Gamma$(n,1)$ 분포를 따른다고 했었는데 모순되는 것이 아닐까요?

그렇지 않습니다. Gamma$(n,1)$ 분포의 평균과 분산이 각각 n일 뿐만 아니라, CLT는 n이 커질수록 Gamma$(n,1)$ 분포가 정규분포에 근사해짐을 말합니다. <그림 1>을 보십시오. 왼쪽 그림에 비해 오른 쪽 그림이 훨씬 정규분포처럼 보이지요.



<그림 1> Gamma$(3,1)$ 분포와 Gamma$(12,1)$ 분포

2.3 정규분포로부터 파생되는 확률분포들

정규분포로부터 통계학이론의 절반 이상이 파생되었다고 말해도 무리는 아닐 것으로 생각됩니다. 그 이유는 중심극한정리가 말하여 주듯이, iid 변수들의 합이 정규분포를 따르도록 되어있다는 사실에 있습니다. 그러므로 정규분포를 따르는 변수에는 그 안에 '오차'의 개념이 내재되어 있습니다.

X가 정규분포 N(0,1)을 따른다고 할 때, $V = X^2$은 어떤 분포를 따를까요?

$$F_V(v) = P\{X^2 \le v\} = \int_{-\sqrt{v}}^{\sqrt{v}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$$

$$= 2\int_0^{\sqrt{v}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$$

이므로

$$f_V(v) = \frac{d}{dv} F_V(v) = \frac{1}{\sqrt{v}} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{v}{2}\right)$$

$$= \frac{1}{\sqrt{\pi}} \frac{v^{1/2-1}}{2^{1/2}} \exp\left(-\frac{v}{2}\right), \quad v \ge 0$$

입니다. 따라서 $V = X^2$는 Gamma(1/2,2) 분포를 따릅니다. 그러므로 V의 분포에 대한 mgf는 다음과 같습니다.

$$m_V(t) = \frac{1}{(1-2t)^{1/2}}, \quad \text{for } t < \frac{1}{2}.$$

이를 일반화하여 봅시다. $X_1, X_2, \cdots, X_n$이 정규분포 N(0,1)로부터의 iid 변수들이라고 할 때, $W = \sum_{i=1}^{n} X_i^2$이 어떤 분포를 따르는지 조사해보기로 합시다. 따라서 $V_1 = X_1^2, V_2 = X_2^2, \cdots, V_n = X_n^2$은 각각 Gamma(1/2,2) 분포로부터 독립적으로 생성된다고 볼 수 있겠습니다. 그러므로 $W = \sum_{i=1}^{n} V_i$의 mgf는

$$m_W(t) = \{m_{V_1}(t)\}^n = \frac{1}{(1-2t)^{n/2}}, \quad \text{for } t < \frac{1}{2}$$

입니다. 그런데 이 mgf가 Gamma$(n/2,2)$ 분포의 것이므로 W의 밀도함수가 다음과 같음을 알게 됩니다.

$$f_W(w) = \frac{1}{\Gamma(n/2)} \frac{w^{n/2-1}}{2^{n/2}} \exp\left(-\frac{w}{2}\right), \quad w \ge 0.$$

이 분포를 자유도 n의 카이제곱(chi-square) 분포 $\chi^2(n)$이라고 합니다. $\chi^2(n)$분포의 평균과 기대값은 각각 n과 $2n$입니다. 즉,

$$E(W) = n, \quad Var(W) = 2n.$$

X가 N(0,1) 분포를 따르고 독립적으로 W가 $\chi^2(n)$ 분포를 따른다고 할 때

$$T = \frac{X}{\sqrt{W/n}}$$

의 분포를 구해보기로 합시다. $S = W$로 정의하면

$$X = n^{-1/2} S^{1/2} T, \; W = S$$

이므로 (X, W)로부터 (T, S)로의 변환에서

$$J = \begin{vmatrix} \frac{\partial x}{\partial t} & \frac{\partial x}{\partial s} \\ \frac{\partial w}{\partial t} & \frac{\partial w}{\partial s} \end{vmatrix} = \begin{vmatrix} n^{-1/2} s^{1/2} & n^{-1/2} s^{-1/2} t/2 \\ 0 & 1 \end{vmatrix} = n^{-1/2} s^{1/2}$$

입니다. 따라서

$$\begin{aligned} f_{T,S}(t,s) &= f_{X,W}(x,w) \cdot |J| \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \cdot \frac{1}{\Gamma(n/2)} \frac{w^{n/2-1}}{2^{n/2}} \exp\left(-\frac{w}{2}\right) \cdot n^{-1/2} s^{1/2} \\ &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2n} s\right) \cdot \frac{1}{\Gamma(n/2)} \frac{s^{n/2-1}}{2^{n/2}} \exp\left(-\frac{s}{2}\right) \cdot n^{-1/2} s^{1/2} \\ &= \frac{1}{\sqrt{2\pi n}} \frac{1}{\Gamma(n/2)} \frac{s^{(n+1)/2-1}}{2^{n/2}} \exp\left[-\left(\frac{1}{2}+\frac{t^2}{2n}\right) s\right], \quad s \geqq 0 \end{aligned}$$

이 됩니다. 그러므로

$$\begin{aligned} f_T(t) &= \int_0^\infty f_{T,S}(t,s)\, ds \\ &= \frac{1}{\sqrt{2\pi n}} \frac{1}{\Gamma(n/2)\, 2^{n/2}} \int_0^\infty s^{(n+1)/2-1} \exp\left[-\left(\frac{1}{2}+\frac{t^2}{2n}\right) s\right] ds \\ &= \frac{1}{\sqrt{2\pi n}} \frac{1}{\Gamma(n/2)\, 2^{n/2}} \cdot \Gamma((n+1)/2) \left(\frac{1}{2}+\frac{t^2}{2n}\right)^{-(n+1)/2} \\ &= \frac{\Gamma((n+1)/2)}{\sqrt{\pi n}\, \Gamma(n/2)} \left(1+\frac{t^2}{n}\right)^{-(n+1)/2}, \quad -\infty < t < \infty \end{aligned}$$

가 됩니다. 이것이 바로 자유도 n의 t 분포 $t(n)$입니다. 좀 복잡해 보이지요.

$t(n)$ 분포에서 자유도 $n = 1$인 경우, 즉 X가 N(0,1) 분포를 따르고 독립적으로 W가 $\chi^2(1)$ 분포를 따르는 경우 (마찬가지로 Y가 X와 독립적으로 N(0,1) 분포를 따르는 경우)

$$T = \frac{X}{\sqrt{W}} = \frac{X}{Y}$$

의 밀도함수는

$$f_T(t) = \frac{1}{\pi\,(1+t^2)}\,, \quad -\infty < t < \infty$$

가 됩니다. 이 분포를 코쉬(Cauchy) 분포라고 하는데, 이 분포에서는 평균이 정의되지 않습니다 (밀도함수의 꼬리가 t^{-2}에 비례하여 작아질 뿐이기 때문입니다). $t\,(n)$ 분포에서 자유도가 $n \geq 2$이면 평균은 0입니다.

W_1이 $\chi^2\,(n_1)$ 분포를 따르고 독립적으로 W_2가 $\chi^2\,(n_2)$ 분포를 따를 때

$$Z = \frac{W_1/n_1}{W_2/n_2}$$

의 분포를 구해보기로 합시다. $Y = W_2$로 정의하면

$$W_1 = \frac{n_1}{n_2}\,YZ, \quad W_2 = Y$$

이므로 (W_1, W_2)로부터 (Z, Y)로의 변환에서

$$J = \begin{vmatrix} \dfrac{\partial w_1}{\partial z} & \dfrac{\partial w_1}{\partial y} \\ \dfrac{\partial w_2}{\partial z} & \dfrac{\partial w_2}{\partial y} \end{vmatrix} = \begin{vmatrix} \dfrac{n_1}{n_2}y & \dfrac{n_1}{n_2}z \\ 0 & 1 \end{vmatrix} = \frac{n_1}{n_2}y$$

입니다. 따라서

$$\begin{aligned} f_{Z,Y}(z,y) &= f_{W_1,W_2}(w_1,w_2)\;|\,J\,| \\ &= \frac{1}{\Gamma(n_1/2)}\;\frac{w_1^{n_1/2-1}}{2^{n_1/2}}\;\exp\!\left(-\frac{w_1}{2}\right)\cdot\frac{1}{\Gamma(n_2/2)}\;\frac{w_2^{n_2/2-1}}{2^{n_2/2}}\;\exp\!\left(-\frac{w_2}{2}\right)\cdot\frac{n_1}{n_2}\,y \\ &= \frac{\left(\dfrac{n_1}{n_2}\right)^{n_1/2} z^{n_1/2-1}}{\Gamma(n_1/2)\;\Gamma(n_2/2)\;2^{(n_1+n_2)/2}}\;\; y^{(n_1+n_2)/2-1}\exp\!\left(-\frac{1+\dfrac{n_1}{n_2}\,z}{2}\,y\right),\;\; y \geqq 0 \end{aligned}$$

이 됩니다. 그러므로

$$\begin{aligned} f_Z(z) &= \int_0^\infty f_{Z,Y}(z,y)\,dy \\ &= \frac{\left(\dfrac{n_1}{n_2}\right)^{n_1/2} z^{n_1/2-1}}{\Gamma(n_1/2)\;\Gamma(n_2/2)\;2^{(n_1+n_2)/2}}\;\int_0^\infty\; y^{(n_1+n_2)/2-1}\exp\!\left(-\frac{1+\dfrac{n_1}{n_2}\,z}{2}\,y\right)dy \\ &= \frac{\left(\dfrac{n_1}{n_2}\right)^{n_1/2} z^{n_1/2-1}}{\Gamma(n_1/2)\;\Gamma(n_2/2)\;2^{(n_1+n_2)/2}}\;\cdot\;\Gamma\left[(n_1+n_2)/2\right]\left(\frac{1+\dfrac{n_1}{n_2}\,z}{2}\right)^{-(n_1+n_2)/2} \end{aligned}$$

$$= \frac{\Gamma[(n_1+n_2)/2]}{\Gamma(n_1/2)\,\Gamma(n_2/2)} \left(\frac{n_1}{n_2}\right)^{n_1/2} z^{n_1/2-1}\left(1+\frac{n_1}{n_2}\,z\right)^{-(n_1+n_2)/2}, \quad z \geqq 0$$

이 됩니다. 이것이 자유도 (n_1, n_2) 인 F 분포 $\mathrm{F}(n_1, n_2)$ 의 밀도함수입니다.

마지막으로, W_1 이 $\chi^2(n_1)$ 분포를 따르고 독립적으로 W_2 이 $\chi^2(n_2)$ 분포를 따를 때

$$U = \frac{W_1}{W_1 + W_2}$$

의 분포를 구해보기로 합시다. $V = W_1 + W_2$ 로 정의하면

$$W_1 = U\,V, \quad W_2 = (1-U)\,V$$

이므로 (W_1, W_2) 로부터 (U, V) 로의 변환에서

$$J = \begin{vmatrix} \frac{\partial w_1}{\partial u} & \frac{\partial w_1}{\partial v} \\ \frac{\partial w_2}{\partial u} & \frac{\partial w_2}{\partial v} \end{vmatrix} = \begin{vmatrix} v & u \\ -v & 1-u \end{vmatrix} = v$$

입니다. 따라서

$$f_{U,V}(u,v) = f_{W_1,W_2}(w_1,w_2)\;|J|$$

$$= \frac{1}{\Gamma(n_1/2)}\;\frac{w_1^{n_1/2-1}}{2^{n_1/2}}\exp\left(-\frac{w_1}{2}\right)\cdot\frac{1}{\Gamma(n_2/2)}\;\frac{w_2^{n_2/2-1}}{2^{n_2/2}}\exp\left(-\frac{w_2}{2}\right)\cdot v$$

$$= \frac{1}{\Gamma(n_1/2)}\;\frac{u^{n_1/2-1}}{2^{n_1/2}}\cdot\frac{1}{\Gamma(n_2/2)}\;\frac{(1-u)^{n_2/2-1}}{2^{n_2/2}}\,v^{(n_1+n_2)/2-1}\exp\left(-\frac{v}{2}\right)$$

가 됩니다. 그러므로

$$f_U(u) = \int_0^\infty f_{U,V}(u,v)\,dv$$

$$= \frac{1}{\Gamma(n_1/2)}\;\frac{u^{n_1/2-1}}{2^{n_1/2}}\frac{1}{\Gamma(n_2/2)}\frac{(1-u)^{n_2/2-1}}{2^{n_2/2}}\int_0^\infty v^{(n_1+n_2)/2-1}\exp\left(-\frac{v}{2}\right)dv$$

$$= \frac{1}{\Gamma(n_1/2)}\;\frac{u^{n_1/2-1}}{2^{n_1/2}}\frac{1}{\Gamma(n_2/2)}\frac{(1-u)^{n_2/2-1}}{2^{n_2/2}}\,\Gamma[(n_1+n_2)/2]\,2^{(n_1+n_2)/2}$$

$$= \frac{\Gamma[(n_1+n_2)/2]}{\Gamma(n_1/2)\,\Gamma(n_2/2)}\;u^{n_1/2-1}(1-u)^{n_2/2-1}, \quad 0 \leqq u \leqq 1.$$

이것을 일반적으로 나타낸 것이 베타분포 $\mathrm{Beta}(\alpha,\beta)$ 입니다:

$$f_U(u) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\,\Gamma(\beta)}\;u^{\alpha-1}(1-u)^{\beta-1}, \quad \alpha > 0,\ \beta > 0,\ 0 \leqq u \leqq 1.$$

$\alpha=\beta=1$인 경우의 베타분포는 균일분포 Uniform(0,1)이고, Beta(α,β) 분포의 평균과 분산은 다음과 같습니다 (연습문제 2.5).

$$E(U)=\frac{\alpha}{\alpha+\beta},\quad Var(U)=\frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}.$$

이들 분포의 통계적 응용은 다음과 같은 데서 나옵니다.

1) $X_1, X_2, \cdots, X_n$을 N(μ,σ^2) 분포로부터의 iid 변수들이라고 할 때,

$$T=\frac{\overline{X}-\mu}{\sqrt{S^2/n}}$$

은 t($n-1$)분포를 따릅니다 (여기서, $\overline{X}$는 표본평균이고 S^2는 표본분산임). 증명은 각종 수리통계학 서적을 참조하길 바랍니다 (연습문제 2.6).

2) $X_1, X_2, \cdots, X_{n_1}$이 N($\mu_1,\sigma_1^2$) 분포로부터 독립적으로 생성되고, 이와 독립적으로 $Y_1, Y_2, \cdots, Y_{n_2}$가 N($\mu_2,\sigma_2^2$) 분포로부터 독립적으로 생성될 때,

$$F=\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

은 F(n_1-1, n_2-1) 분포를 따릅니다 (여기서, S_1^2은 X 표본의 분산이고 S_2^2은 Y 표본의 분산임). 증명은 각종 수리통계학 서적을 참조하십시오 (연습문제 2.7).

2.4 몬테칼로 모의시행 (Monte Carlo simulation)

2.2절에서 보았듯이 어떤 확률 및 확률분포에 있어서는 해석적 풀이가 매우 어렵습니다. 때문에, 컴퓨터로 각종 확률분포로부터의 임의수(random number)를 발생시켜 실제 확률현상을 흉내내는 방법을 생각해봅니다. 이런 기법이 몬테칼로 모의시행(Monte Carlo simulation)인데, '몬테칼로'는 도박으로 유명한 도시입니다. 도박이라는 것이 확률현상을 인위적으로 반복하는 행위니까 이런 이름이 붙은 것입니다.

몬테칼로 기법의 기초는 균일분포 Uniform(0,1)로부터 임의수 $U_1, U_2, \cdots, U_n$을 발생시키는 것입니다. 이에 대하여는 이미 1.1절에서 언급한 바 있습니다. 균일분포를 바탕으로 지수분포, 정규분포, 베르누이 분포, … 등등의 수없이 많은 분포로부터 임의수가 발생하도록 할 수 있습니다.

다음 정리는 연속분포로부터의 임의수 발생원리를 말해줍니다. 일반적인 정리이니까 적용 폭이 매우 넓습니다.

모든 연속 분포함수는 균일분포를 따릅니다.

확률변수 X가 연속분포를 따른다고 하고 분포함수를 $F(x)$라고 합시다. 그러면 $Y = F(X)$는 균일분포 Uniform(0,1)을 따릅니다.

이것의 증명은 쉽습니다. 분포함수 $F(x) = P\{X \leq x\}$가 연속이고 단조증가하는 경우 역함수를 쉽게 정의할 수 있습니다. 즉

$$F(x) = u \quad \Leftrightarrow \quad F^{-1}(u) = x, \quad \text{for } 0 \leq u \leq 1$$

따라서

$$P\{Y \leq u\} = P\{F(X) \leq u\}$$
$$= P\{X \leq F^{-1}(u)\} = F(F^{-1}(u)) = u.$$

그러므로 $Y = F(X)$는 균일분포 Uniform(0,1)을 따릅니다. ■

예를 들어 어떤 확률변수의 분포함수가 $F(x) = x^2$, $0 \leq x \leq 1$이라고 합시다. 그러면 X^2는 Uniform(0,1) 분포를 따릅니다. 따라서 Uniform(0,1)로부터 확률변수 U를 발생시킨 다음,

$$X^2 = U, \quad \text{즉} \quad X = \sqrt{U}$$

로 변환하면 주어진 분포함수 $F(x)$를 따르는 확률변수 X가 실현됩니다.

이런 방법으로 지수분포 Exponential(θ)로부터 임의수를 발생시킬 수 있습니다.

지수분포 Exponential(θ)를 따르는 임의수 발생시키기

확률변수 X가 Exponential(θ) 분포를 따른다면

$$F(x) = 1 - \exp\left(-\frac{x}{\theta}\right), \quad x \geq 0$$

입니다. 따라서 Uniform(0,1)로부터 확률변수 U를 발생시킨 다음,

$$1 - \exp\left(-\frac{X}{\theta}\right) = U$$

로 놓으면 되겠군요. 즉

$$X = -\theta \log_e (1-U) \tag{2}$$

로 변환하면 됩니다.

실제 Exponential(2) 임의수의 산출 예를 들어보지요. Uniform(0,1)로부터 확률

<표 1> 지수분포로부터 임의수 발생시키기

```
/* Exponential(2) Random Number Generation  */
/* exp.iml                                  */

proc iml;
    Nrepeat = 100;
    sum = 0;
    do repeat=1 to Nrepeat;
        u = uniform(0);
        x = -2*log(1-u);
        print u x;
        sum = sum + x;
    end;
    mean = sum / Nrepeat;
    print Nrepeat mean;
quit;
```

변수U를 발생시키니까

0.1452 0.1994 0.6977 0.0424 0.5333 …

이 나왔습니다. 그러면 변환 (2)를 통하여 (단,$\theta = 2$), 다음과 같이 Exponential(2) 임의수가 발생됩니다.

0.3137 0.4447 2.3925 0.0867 1.5240 … .

이런 임의수를 100개 만들어보고 평균을 구하니 1.8874가 나오는군요 (참고로, $\theta = 2$인 지수분포의 평균은 2입니다). <표 1>의 컴퓨터 프로그램을 보십시오.

이것은 SAS/IML(interactive matrix language)로 쓰인 것인데, 기본 알고리즘은 마찬가지니까 다른 컴퓨터 언어로도 쉽게 바꿔 쓸 수 있을 것입니다. [SAS/IML에서 uniform(0)는 Uniform(0,1) 임의수를 발생시키는 함수입니다. 이 때 초기 수를 컴퓨터 시계가 결정하므로 같은 프로그램을 작동시키더라도 수치적 결과는 매번 다르게 됩니다.]

다음으로 표준정규분포 N(0,1)로부터 임의수를 발생시켜보도록 합시다. 1.3절에서 우리가 처음으로 정규분포를 유도하였을 때 어떻게 하였습니까? 그것을 활용하면 되지 않을까요?

<표 2> 표준정규분포로부터 임의수 발생시키기

```
/* Standard Normal Random Number Generation */
/* normal.iml                                */

proc iml;
    Nrepeat = 100;
    pi = 3.141592;
    sum = 0;
    do repeat=1 to Nrepeat/2;
       R2 = -2*log(1-uniform(0));
       theta = uniform(0);
       x1 = sqrt(R2)*cos(2*pi*theta);
       x2 = sqrt(R2)*sin(2*pi*theta);
       print x1 x2;
       sum = sum+x1+x2;
    end;
    mean = sum/Nrepeat;
    print Nrepeat mean;
quit;
```

정규분포 N(0,1)을 따르는 임의수 발생시키기

1) R^2을 지수분포 Exponential(2)로부터 생성시키고
2) 독립적으로 Θ를 균일분포 Uniform(0,1)로부터 생성시킨 뒤
3) $X_1 = R\cos(2\pi\Theta)$, $X_2 = R\sin(2\pi\Theta)$를 계산합니다.

이렇게 하면 표준정규분포를 따르는 임의수가 한번에 2개씩 만들어집니다. <표 2>는 이것을 컴퓨터 프로그램으로 만들어 본 것입니다. 그 결과 다음과 같은 임의수가 만들어집니다 (100개 평균 -0.0566).

0.7390 -0.2659 0.4851 2.706 0.1519 0.4743 …

정규분포 $N(\mu,\sigma^2)$으로부터의 임의수 Y 발생은 N(0,1) 임의수 X를 단순히

$$Y = \mu + \sigma X$$

로 선형 변환하면 됩니다.

베르누이 분포 Bernoulli(θ)로부터의 임의수 발생은 균일분포 Uniform(0,1)로부터 임의수 U를 만든 뒤 그것이 θ보다 작으면 성공($X=1$)으로, 그렇지 않으면 실패($X=0$)로 하면 될 것입니다.

<표 3> 음이항분포로부터 임의수 발생시키기

```
/* Negative Binomial Random Number Generation */
/* nb.iml                                     */

proc iml;
   r = 4;  theta=0.25;
   Nrepeat = 100;
   sum = 0;
   do repeat=1 to Nrepeat;
      sum1=0;  sum0=0;
      do while (sum1 < r);
         if uniform(0) < theta then success=1;
         else success=0;
         sum1 = sum1 + success;
         sum0 = sum0 + 1- success;
      end;
      x = sum0;
      print x;
      sum = sum + x;
   end;
   mean = sum/Nrepeat;
   print Nrepeat mean;
quit;
```

음이항분포 NB(r,θ)로부터의 임의수 발생은 어떻게 하면 될까요? 음이항변수가 r번째 성공이 있기까지의 실패 수이므로 Bernoulli(θ) 분포로부터 독립적으로 임의수를 반복 발생시키면서 성공이 r번 누적될 때까지 실패 수를 세고 있으면 될 것입니다. 때에 따라 좀 지루할 수도 있겠지요 (그러나 그런 염려는 아예 하지 마세요. 컴퓨터는 그런 것을 좋아하니까요. 너무 자기 위주로만 생각하지 맙시다!) <표 3>이 NB(4,0.25) 분포로부터 임의수를 발생시키기 위한 컴퓨터 프로그램입니다. 그 결과로

2, 13, 5, 16, 9, … (100개 평균 11.52)

를 얻습니다 (참고로, NB(4,0.25) 분포의 평균은 12입니다).

<표 4> 이항분포로부터 임의수 발생시키기

```
/* Binomial Random Number Generation */
/* b.iml                             */

proc iml;
   n = 10;  theta=0.25;
   Nrepeat = 100;
   sum = 0;
   do repeat=1 to Nrepeat;
      sum1=0;
      do j = 1 to n;
         if uniform(0) < theta then success=1;
         else success=0;
         sum1 = sum1 + success;
      end;
      x = sum1;
      print x;
      sum = sum + x;
   end;
   mean = sum/Nrepeat;
   print Nrepeat mean;
quit;
```

이항분포 $B(n,\theta)$로부터의 임의수 발생은 어떻게 하면 될까요? 이항확률변수는 단순히 n개의 Bernoulli(θ) 변수를 합한 것이므로 아주 쉽습니다. <표 4>의 컴퓨터 프로그램을 보십시오 ($n=10, \theta=0.25$인 경우). 그 결과로

1, 0, 1, 4, 2 … (100개 평균 2.69)

를 얻습니다 (참고로, $n=10$, $\theta=0.25$인 이항분포의 평균은 2.5입니다).

마지막으로, 포아송 분포 Poisson(θ)로부터 임의수를 생성시키는 방법을 생각해 봅시다. Poisson(θ) 변수는 0부터 무한대까지의 정수 값을 모두 가질 수 있기 때문에, Uniform(0,1) 확률변수 U의 값과 포아송 분포의 누적분포와 비교해가면서 포아송 임의수를 하나씩 찾는 것은 무척 소모적인 일이 됩니다. 그것보다는 포아송 분포와 지수분포와의 관계를 활용하는 방법이 좋습니다 (1.5절 참조). 즉

포아송 과정의 모형에서 사건이 발생하기까지 걸리는 시간인 $T_1, T_2, T_3, \cdots$를 Exponential($1/\theta$)로부터 거듭 발생시키면서 시구간 (0,1] 사이에 모두 몇 개의 사건이 발생하는지를 세어보는 것입니다. 수식으로 표현하면, Poisson(θ) 변수 N이

<표 5> 포아송 분포로부터 임의수 발생시키기

```
/* Poisson Random Number Generation */
/* poisson.iml                      */

proc iml;
   Nrepeat = 100;
   theta = 2;
   sum = 0;
   do repeat=1 to Nrepeat;
      count = -1;
      sum1 = 0;
      do while (sum1 < 1);
         t = -(1/theta)*log(1-uniform(0));
         sum1 = sum1 + t;
         count = count + 1;
      end;
      N = count;
      print N;
      sum = sum + N;
   end;
   mean = sum / Nrepeat;
   print Nrepeat mean;
quit;
```

n이라는 것은

$$T_1 + T_2 + \cdots + T_n \leqq 1 < T_1 + T_2 + \cdots + T_{n+1}$$

입니다. <표 5>의 컴퓨터 프로그램을 보십시오 ($\theta = 2$인 경우). 100번의 시행결과, 평균이 2.02로 나오는군요 (참고로, $\theta = 2$인 포아송 분포의 평균은 2입니다).

다시 한번 말하는데, 위의 프로그램들은 매번 작동될 때마다 다른 수치적 결과를 냅니다. 다른 결과가 나왔다고 이상해하지 마십시오. 간혹 크게 다른, 그러나 대부분은 조금씩 다른 결과가 나오는 것이 자연스러운 것입니다.

이제 몬테칼로 임의시행이 요구되는 실제 예를 하나 들겠습니다. 2.2절에서 다루었던 이변량 정규분포에서의 확률계산 문제입니다. 다시 옮기자면, (X_1, X_2)가 이변량 정규분포 $\mathrm{BN}(0,0,1,1,\rho)$에서 $\rho = 0.5$인 경우

$$h(\rho) = P\{X_1 \geq 1, X_2 \geq 1\}$$

을 계산하는 문제입니다.

<표 6> 이변량정규분포 확률의 몬테칼로 계산

```
/* Monte-Carlo Computing of Bivariate Normal Probability */
/* FileName bn.iml                                       */

proc iml;
   rho = 0.5;
   Nrepeat = 10000;
   sum = 0;
   do repeat=1 to Nrepeat;
      z1 = normal(0);
      z2 = normal(0);
      x1 = z1;
      x2 = rho*z1 + sqrt(1-rho*rho)*z2;
      if x1 >= 1 & x2 >= 1 then success = 1;
      else success = 0;
      sum = sum + success;
   end;
   p = sum/Nrepeat;
   print Nrepeat p;
quit;
```

이 문제에 대하여는 해석적 풀이가 불가능해 보이지만 몬테칼로 계산으로는 쉽습니다. 즉

1) $(X_1, X_2) \sim \mathrm{BN}(0,0,1,1,\rho)$
2) $X_1 \geq 1$ and $X_2 \geq 1$이면 계수기(counter)를 1만큼 증가
3) 단계 1과 2를 n번 반복
4) $h(\rho)$ 에 대한 추정치 p를 산출 : $p \leftarrow \mathrm{counter}/n$.

BN(0,0,1,1,ρ) 분포로부터 임의수를 발생시키기 위해서는

1) N(0,1) 분포로부터 독립적으로 Z_1과 Z_2를 발생시킨 뒤
2) $X_1 = Z_1$, $X_2 = \rho\, Z_1 + \sqrt{1-\rho^2}\; Z_2$를 계산합니다 (1.3절 참조).

이를 구현한 <표 6>의 컴퓨터 프로그램을 보십시오. 표준정규분포로부터 임의수 Z_1과 Z_2를 발생시키기 위하여 SAS/IML 함수인 normal(0)를 활용하였습니다. 프로그램을 쉽게 짜기 위해서 그랬던 것뿐입니다. 원천 기술을 그대로 쓰려면 normal(0) 대신 <표 2>의 프로그램을 넣으면 됩니다.

$n = 10{,}000$번의 몬테칼로 시행을 다시 한 결과, $p = 0.0653$을 얻었습니다.

2.4.1 t, t, t 분포에 관한 몬테칼로 연구

우리는 2.3절에서 t 분포를 유도하였습니다. 그리고 $X_1, X_2, \cdots, X_n$이 $\mathrm{N}(\mu,\sigma^2)$ 분포를 따르면

$$T = \frac{\overline{X}-\mu}{\sqrt{S^2/n}} \tag{2}$$

이 자유도 $n-1$인 t 분포를 따른다고 하였습니다. 증명은 빼먹었었고 대신 참고문헌을 제시하였었지요. 그렇게 되었던 까닭은 어려워서가 아니라 지겨웠기 때문입니다. 그래도 일생에 한두 번쯤은 꼼꼼히 처음부터 끝까지 증명해봐야 할 것입니다. 이 절에서 우리는 더 야심적인 문제에 도전해보기로 하겠습니다.

T의 분포에 관한 연구문제 : $X_1, X_2, \cdots, X_n$이 정규분포가 아니라

1) Uniform$(-1,1)$ 분포로부터 발생하는 경우
2) 라플라스(Laplace) 분포[1]로부터 발생하는 경우

에 T는 어떤 분포를 따릅니까? 그 분포의 상위 2.5%, 상위 5% 분위수, 상위 10% 분위수를 제시하세요.

모분포가 정규분포인 경우에는 수학적인 풀이가 가능하지만 (쉽지는 않았어도), 다른 분포의 경우에는 거의 불가능해 보입니다. 몬테칼로 방법, 즉 계산적인 방법으로 T의 분포를 구해볼 수는 없을까요?

몬테칼로 방법으로 문제풀이를 시도하기로 하고 다음 한 가지 경우를 추가하기로 하지요.

3) 정규분포 N(0,1)을 따르는 경우.

물론 이 경우에 대하여는 수학적 답이 있고 우리가 그 답을 알고 있습니다 (통계분포표를 통하여). 역설적으로, 바로 그렇기 때문에 이 경우를 포함시키고자 하는 것입니다. 이 경우에 대하여 우리가 구한 몬테칼로 해와 수학적 해가 얼마나 일치하는가를 봄으로써, 다른 경우에 대한 우리의 몬테칼로 해가 타당한가, 신뢰로운가를 간접적으로 확인할 수 있을 것입니다.

---

1) 이중지수분포(double exponential distribution)라고도 하는 라플라스 분포는 U, V가 독립적인 지수분포 Exponential(1)로부터 생성될 때, $X = U - V$가 따르는 확률분포입니다. 다음과 같이 대칭적인 밀도함수를 가집니다. 증명해보세요 (연습문제 2.9).

$$f_X(x) = \frac{1}{2} e^{-|x|}, \quad -\infty < x < \infty.$$

<표 7>이 몬테칼로 방법으로 t 분포의 분위수 T_α를 구하는 컴퓨터 프로그램입니다 ($n=10$인 경우). 프로그램에서 쓴 알고리즘은 다음과 같습니다.

1) 각 경우의 분포로부터 임의수 $x_1, x_2, \cdots, x_n$을 발생시킵니다.

2) 식 (2)로부터 t 통계값을 계산합니다 (세 경우 모두 $\mu=0$).

3) 단계 1과 2를 N_1번 반복하여 t 분포의 표본 α 분위수 t_α를 산출합니다:
모의생성된 N_1개의 t 통계값을 순서대로 정렬할 때 표본 α 분위수는 $(N_1+1)(1-\alpha)$ 번째 값으로 정의합니다.

4) t 분포의 분위수 T_α를 추정하기 위하여 앞의 단계 1·2·3을 N_2번 반복합니다. 그렇게 하여 $\{ t_\alpha^{(k)} \mid k=1,\cdots,N_2 \}$를 얻게되면 T_α의 추정치와 그것의 표준오차(standard error)를 다음과 같이 산출합니다.

$$\hat{T}_\alpha = \sum_{k=1}^{N_2} t_\alpha^{(k)} / N_2 \ (\equiv \bar{t}^{(.)}_{)\alpha},$$

$$s.e.\hat{T}_\alpha = \sqrt{\sum_{k=1}^{N_2} (t_\alpha^{(k)} - \bar{t}_\alpha{}^{(.)})^2 / (N_2(N_2-1))} \ .$$

$n=10$인 경우에서 반복시행수 $N_1=999$, $N_2=100$으로 하여, 그러니까 각 경우마다 총 99,000개의 t를 발생시켜 분위수 산출에 투입한 결과가 다음과 같습니다 (맨 아래 줄은 자유도 9의 t 분포표가 제시하는 분위수입니다).

	t 분포		
모(母)분포	10% 분위수	5% 분위수	2.5% 분위수
균일분포	1.3687 (0.0044)	1.8361 (0.0064)	2.3254 (0.0089)
라플라스 분포	1.3932 (0.0037)	1.7995 (0.0052)	2.1693 (0.0074)
정규분포	1.3789 (0.0046)	1.8247 (0.0068)	2.2580 (0.0092)
t 분포표	1.3830	1.8331	2.2622

먼저 정규분포의 경우를 보면, 몬테칼로 t 2.5% 분위수가 2.2580 ± 0.0092인데 이론적인 t 분위수는 2.2622입니다. 그런대로 일치하는 결과라고 하겠습니다. 균일분포의 t 2.5% 분위수는 정규분포의 t 2.5% 분위수에 비해 약간 큰 것으로 보입니다. 그리고 라플라스 분포의 경우는 이와 반대로 보입니다.

<표 7> 몬테칼로 방법으로 t 분포의 분위수 구하기

```
/* t, t, t distribution for Uniform, Laplace and Normal Cases */
/* ttt.iml                                                    */

proc iml;
   n = 10;
   Nrepeat1 = 999;
   Nrepeat2 = 100;
   t025 = j(Nrepeat2,1,0);
   t050 = j(Nrepeat2,1,0);
   t100 = j(Nrepeat2,1,0);
   do repeat2=1 to Nrepeat2;
      T = j(Nrepeat1,1,0);
      S = j(Nrepeat1,1,0);
      do repeat1=1 to Nrepeat1;
         sum1 = 0;  sum2 = 0;
         do sample=1 to n;
①           x = uniform(0)*2-1;
②           /*  case2:  x = -2*log(1-uniform(0))+2*log(1-uniform(0));  */
③           /*  case3:  x = normal(0);                                 */
            sum1 = sum1 + x;
            sum2 = sum2 + x*x;
         end;
         xbar = sum1/n;
         s2 = (sum2 - n*xbar*xbar)/(n-1);
         t[repeat1] = abs(xbar/sqrt(s2/n));
      end;
      R = rank(T);
      do repeat1=1 to Nrepeat1;
         S[R[repeat1]] = T[repeat1];
      end;
      t025[repeat2] = S[0.95*(Nrepeat1+1)];
      t050[repeat2] = S[0.90*(Nrepeat1+1)];
      t100[repeat2] = S[0.80*(Nrepeat1+1)];
   end;
   t100m = sum(t100)/Nrepeat2;
   t100s = sqrt((ssq(t100)-Nrepeat2*t100m*t100m)/(Nrepeat2*(Nrepeat2-1)));
   t050m = sum(t050)/Nrepeat2;
   t050s = sqrt((ssq(t050)-Nrepeat2*t050m*t050m)/(Nrepeat2*(Nrepeat2-1)));
   t025m = sum(t025)/Nrepeat2;
   t025s = sqrt((ssq(t025)-Nrepeat2*t025m*t025m)/(Nrepeat2*(Nrepeat2-1)));
   print Nrepeat1 Nrepeat2 t100m[format=8.4] t100s[format=8.4];
   print Nrepeat1 Nrepeat2 t050m[format=8.4] t050s[format=8.4];
   print Nrepeat1 Nrepeat2 t025m[format=8.4] t025s[format=8.4];
quit;
```

이제까지의 컴퓨터 계산(수행시간)은 티 석 잔 마실 동안이면 충분합니다. 수학으로는 불가능해 보이기 때문에 컴퓨터의 능력이 한층 돋보이는군요. 우리의 불쌍한 컴퓨터가 똑 같은 일을 각 경우에 거의 10만 번씩 · 3 경우를 처리해내느라고 무척 고생한 덕분입니다.

※ <표 7>의 프로그램은 경우 1 (균일분포)에 대한 것입니다. 경우 2 (라플라스 분포)를 하려면 ①을 없애고 대신 ②를 살려놓으면 됩니다. 마찬가지로 경우 3 (정규분포)을 하려면 ①과 ②를 없애고 ③을 살려놓으면 됩니다.

※ 모든 경우에서 T의 분포가 대칭이므로 표본 분위수의 산출에서도 그것을 활용하였습니다. 즉, t 분포의 상위 α 분위수를 직접 추정하는 대신 우회적으로 절대값 t 분포의 상위 2α 분위수를 추정하였습니다.

※ 분위수에 대한 몬테칼로 추정값의 정밀도(precision)를 높이기 위해서는 모의시행수 N_1과 N_2를 늘려야 합니다. 구체적으로, 추정값의 표준오차(s.e.)를 절반으로 하기 위해서는 N_2를 4배인 400으로 하여야 합니다 ($N_1 = 999$). 또는 N_2를 고정시키고 N_1을 4배인 3999로 해도 될 것입니다 ($N_2 = 100$).

2.5* 유한과 무한 - 극단값 분포, 에지워스 근사, 델타 방법

우리는 2.2절에서 극한이론(limit theory, asymptotic theory)을 다루었습니다만 처음이어서 그 중에서도 가장 굵은 2개의 줄기(대수의 법칙과 중심극한정리)만 보았습니다. 이 절에서는 극한이론의 좀 더 다양한 응용과 함께 유한과 무한 사이에 어떤 중간점이 있는가를 보도록 하겠습니다.

예로 시작하기로 하지요. $X_1, X_2, \cdots, X_n$을 지수분포 Exponential(θ)로부터의 iid 변수들이라고 합시다. 이 때,

$$S_n = \min\{X_1, X_2, \cdots, X_n\}$$

의 분포는 다음과 같이 쉽게 구할 수 있습니다.

$$\begin{aligned} P\{S_n \geq s\} &= P\{\min\{X_1, X_2, \cdots, X_n\} \geq s\} \\ &= \prod_{i=1}^{n} P\{X_i \geq s\} \\ &= \prod_{i=1}^{n} \exp\left(-\frac{s}{\theta}\right) = \exp\left(-\frac{n s}{\theta}\right), \text{ for } s \geq 0. \end{aligned}$$

따라서, S가 Exponential(θ/n) 분포를 따름을 알 수 있습니다. 즉 $E(S)$가 θ/n으로 $E(X_1) = \theta$의 $1/n$ 입니다. 그러면, 이제부터는

$$T_n = \max\{X_1, X_2, \cdots, X_n\}$$

의 분포를 구해 봅시다. T_n은 최장 수명, 최대 대기시간, … 등과 관계 있습니다.

$$\begin{aligned} P\{T_n \le t\} &= P\{\max\{X_1, X_2, \cdots, X_n\} \le t\} \\ &= \prod_{i=1}^{n} P\{X_i \le t\} \\ &= \prod_{i=1}^{n} \left\{1 - \exp\left(-\frac{t}{\theta}\right)\right\} = \left\{1 - \exp\left(-\frac{t}{\theta}\right)\right\}^n \end{aligned}$$

이고, 밀도함수는

$$\begin{aligned} f_{T_n}(t) &= \frac{d}{dt} P\{\max\{X_1, X_2, \cdots, X_n\} \le t\} \\ &= \frac{n}{\theta} \exp\left(-\frac{t}{\theta}\right)\left\{1 - \exp\left(-\frac{t}{\theta}\right)\right\}^{n-1}, \ t \ge 0 \end{aligned} \qquad (3)$$

입니다. 그러니, 이제까지 알고 있던 어떤 분포도 아닙니다. 그리고 어떤 $t \ge 0$에 대하여도

$$\lim_{n\to\infty} P\{T_n \le t\} = \lim_{n\to\infty} \left\{1 - \exp\left(-\frac{t}{\theta}\right)\right\}^n = 0$$

입니다. 그러므로 n이 커짐에 따라 T_n이 상당히 커질 것임을 알 수 있습니다. 그러므로 T_n이 어떻게 커질 것인가를 보기 위하여 다음의 표준화를 생각해봅시다:

$$Z_n = (T_n - a_n)/b_n,$$

여기서 a_n과 b_n은 우리가 정해야 할 적절한 n의 함수입니다.

$$\begin{aligned} P\{Z_n \le z\} &= P\{T_n \le a_n + b_n z\} \\ &= \left\{1 - \exp\left(-\frac{a_n + b_n z}{\theta}\right)\right\}^n = \left\{1 - \exp\left(-\frac{a_n}{\theta}\right)\exp\left(-\frac{b_n z}{\theta}\right)\right\}^n \end{aligned}$$

이므로

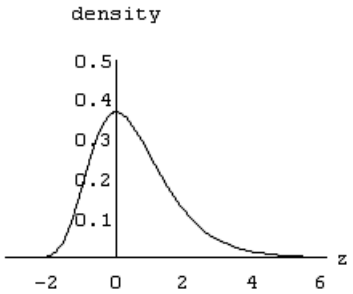
$$\exp\left(-\frac{a_n}{\theta}\right) = \frac{1}{n}, \quad b_n = 1$$

로 놓음으로써

$$\lim_{n\to\infty} P\{Z_n \le z\} = \lim_{n\to\infty} \left\{1 - \frac{1}{n}\exp\left(-\frac{z}{\theta}\right)\right\}^n = \exp\left\{-\exp\left(-\frac{z}{\theta}\right)\right\}$$

가 됩니다. 즉, $a_n = \theta \log_e n$, $b_n = 1$로 잡음으로써 $Z_n = T_n - \theta \log_e n$이 극한적으로 분포함수

$$F_Z(z) = \exp\left\{-\exp\left(-\frac{z}{\theta}\right)\right\}, \quad -\infty < z < \infty$$



<그림 2> 극단값 분포의 밀도함수 ($\theta = 1$)

를 따르고 그것의 밀도함수는

$$f_Z(z) = \frac{d}{dz} F_Z(z) = \frac{1}{\theta} exp\left\{ -\frac{z}{\theta} - \exp\left(-\frac{z}{\theta}\right)\right\}, \quad -\infty < z < \infty \qquad (4)$$

가 됩니다. 이 분포를 극단값 분포(extreme value distribution)라고 합니다. <그림 2>를 보십시오. 이 예가 주는 좋은 교훈은, 매우 당연하게도, 극한분포에는 정규분포 하나만 있는 것이 아니라는 것이지요.

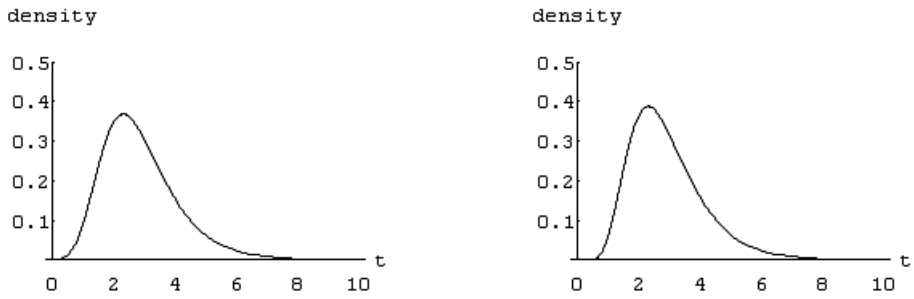
따라서 $T_n = \max\{ X_1, X_2, \cdots, X_n \}$은 대략 $\theta \log_e n \ (= a_n)$을 따라 증가하되 그 외에 밀도함수 $f_Z(z) > 0$를 따르는 확률변수 Z가 덧붙여진다고 하겠습니다. 그러므로 극한이론에 의하면

$$f_{T_n}(t) \simeq \frac{1}{\theta} \exp\left\{ -\frac{t-a_n}{\theta} - \exp\left(-\frac{t-a_n}{\theta}\right)\right\}, \quad t \geqq 0$$

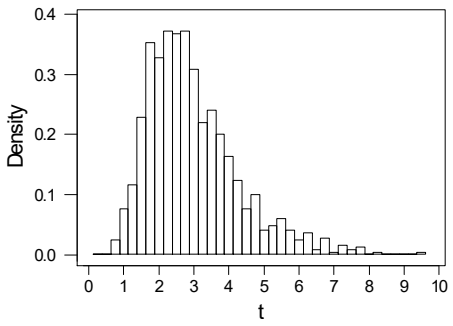
으로 근사됩니다. 예를 들어 $n = 10$, $\theta = 1$인 경우, 이 근사밀도를 정확한 확률밀도인 (3)과 비교해보기로 하지요. <그림 3>을 보십시오. 어떻습니까? 거의 비슷하지 않습니까? 이런 것을 수학의 힘이라고 말할 수 있겠습니다.

참고로, <그림 4>는 $T_n = \max\{ X_1, X_2, \cdots, X_n \}$의 분포를 몬테칼로 모의시행으로 근사해본 것입니다 ($n = 10$, $\theta = 1$, 시행 수 = 1,000). 역시 비슷한 형태의 밀도를 볼 수 있습니다. 이것은 컴퓨터의 힘이라고 하겠습니다.

"수학이냐? 컴퓨터냐?", 글세 꼭 그렇게 생각해야 할 필요가 있을까요? 통계학에서는 두 개가 모두 좋은 도구입니다. 우리가 오른 손만 쓸 것이 아니라 왼 손도 균형 있게 써야 좋은 것처럼 두 도구를 모두 활용할 필요가 있지 않을까요?



<그림 3> 최대값 T_n의 극단값 분포 근사(왼쪽)와 정확한 밀도(오른쪽)



<그림 4> 최대값 T_n의 몬테칼로 분포 (시행수 = 1000)

극단값 분포보다도, 중심극한정리(CLT)가 극한정리 중 '중심'이므로 극한분포의 응용으로서 정규분포가 가장 빈번합니다. 중심극한정리는, $X_1, X_2, \cdots$ 가 모분포 F로부터의 iid 변수들인 경우

$$P\{\sqrt{n}\,(\overline{X}-\mu)/\sigma \le z\} \to \Phi(z), \quad \text{as } n \to \infty$$

으로 기술되지만, 사람들은 그 극한적 행태에 대한 관심보다는 실용적 관점에서

$$Z_n = \sqrt{n}\,(\overline{X}-\mu)/\sigma$$

에 대한 근사분포로서 표준정규분포를 쓰고 싶어합니다 (유한한 n에 대하여). 그러나 문제는 유한한 n과 무한(∞)의 차이입니다. 아무리 n이 커도 무한이 될 수 없으며 특히 $n=20$ 정도라면 이것이 극한정리를 적용할 수 있는 경우인지 의문입니다. 다

음의 에지워스 전개가 바로 이런 의문에 대한 부분적인 답을 제공합니다 (이 정리의 증명은 생략하기로 합니다).

에지워스 전개(Edgeworth expansion):

$$P\{ Z_n \le z \} \simeq \Phi(z) - \phi(z) \left[\frac{\gamma_1}{6\sqrt{n}} H_2(z) + \frac{\gamma_2}{24n} H_3(z) + \frac{\gamma_1^2}{72n} H_5(z) \right],$$

$$f_{Z_n}(z) \simeq \phi(z) \left[1 + \frac{\gamma_1}{6\sqrt{n}} H_3(z) + \frac{\gamma_2}{24n} H_4(z) + \frac{\gamma_1^2}{72n} H_6(z) \right].$$

여기서 γ_1과 γ_2는 각각 모분포 F의 왜도(skewness)와 첨도(kurtosis)이며

$$H_1(z) = z, \qquad H_2(z) = z^2 - 1,$$

$$H_3(z) = z^3 - 3z, \qquad H_4(z) = z^4 - 6z^2 + 3,$$

$$H_5(z) = z^5 - 10z^3 + 15z, \; H_6(z) = z^6 - 15z^4 + 45z^2 - 15$$

는 에르미트(Hermite) 다항식들입니다:

$$\frac{d}{dz^k} e^{-z^2/2} = (-1)^k H_k(z) e^{-z^2/2}, \; k = 1,2,3,\cdots .$$

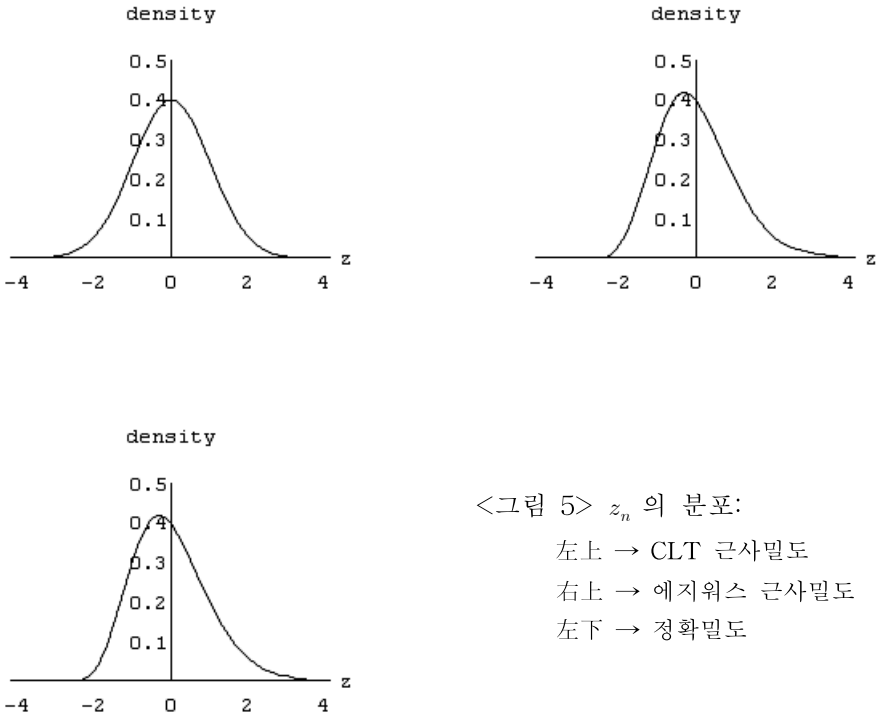
그러므로 그리 크지 않은 n의 경우라도, 모분포 F의 밀도함수가 대칭적일수록 ($\gamma_1 = 0$), 그리고 중앙에 밀도가 높고 양 꼬리의 밀도가 낮은 종모양(bell shape)일수록($\gamma_2 = 0$), Z_n에 대한 정규근사가 정확하다는 것을 알 수 있습니다.

응용 예를 보도록 하겠습니다. $X_1, X_2, \cdots, X_n$을 Exponential(θ) 분포로부터의 iid 변수들이라고 합시다 ($\mu = \theta, \sigma = \theta$). 이 때 $\overline{X}$의 분포를 알고 싶다고 합시다.

사실 이 경우에는 $\overline{X}$의 정확한 분포가 알려져 있습니다. 즉

$$S_n = \sum_{i=1}^{n} X_i \; (= n\overline{X})$$

가 감마분포 Gamma(n, θ)를 따르지요 (2.1절 참조). 그러나 잠시 그런 사실을 모르는 척하고, CLT 근사와 에지워스 근사를 시험해보기로 하겠습니다. Z_n에 대한 밀도함수로서 근사적인 두 가지와 정확한 한 가지는 다음과 같습니다.



<그림 5> z_n 의 분포:
左上 → CLT 근사밀도
右上 → 에지워스 근사밀도
左下 → 정확밀도

1) CLT 근사:

$$f_{Z_n}(z) \simeq \phi(z).$$

2) 에지워스 근사: 지수분포의 경우 $\gamma_1 = 2, \gamma_2 = 6$ 이므로

$$f_{Z_n}(z) \simeq \phi(z)\left[1 + \frac{\gamma_1}{6\sqrt{n}} H_3(z) + \frac{\gamma_2}{24n} H_4(z) + \frac{\gamma_1^2}{72n} H_6(z)\right].$$

3) 정확 밀도: $S_n = n\theta + n^{0.5}\theta Z_n$ 이므로 감마분포로부터

$$f_{Z_n}(z) = \frac{n^{n-0.5}}{\Gamma(n)}\left(1+\frac{z}{n^{0.5}}\right)^{n-1}\exp\left\{-n\left(1+\frac{z}{n^{0.5}}\right)\right\}.$$

$n = 10$, $\theta = 1$ 의 경우에 대하여 세 밀도를 비교한 <그림 5>를 보십시오. 이 경우에서는 에지워스 근사가 무척 근사하군요.

마지막으로, 델타 근사(delta approximation)를 설명하기로 합니다. CLT와 테일러 정리를 이용하는 매우 유용한 방법입니다.

$X_1, X_2, \cdots, X_n$가 평균과 분산이 각각 μ와 σ^2인 iid 변수인 경우, $h(\overline{X})$의 분포에 대하여 무엇을 알 수 있는가를 생각해보고자 합니다. (여기서 $h(t)$는 또한 몇 번이라도 미분가능할 정도로 매끄러운(smooth) 함수라고 가정하겠습니다.) 예를 들어 $X_1, X_2, \cdots, X_n$을 Exponential(θ) 분포로부터의 iid 변수들인 경우에서

$$T_n = \log_e \overline{X}$$

의 분포를 알고 싶다고 합시다.

$h'(\mu) \neq 0$인 경우, 테일러 정리로부터

$$h(\overline{X}) \simeq h(\mu) + h'(\mu)(\overline{X} - \mu)$$

이므로

$$\sqrt{n}\ \frac{\overline{X} - \mu}{\sigma} \simeq \sqrt{n}\ \frac{h(\overline{X}) - h(\mu)}{h'(\mu)\ \sigma}$$

입니다. 그런데 극한적으로 좌변이 N(0,1) 분포에 수렴하므로 우변도 그러할 것입니다. 그러므로 $h(\overline{X})$은 근사적으로 정규분포

$$\mathrm{N}(h(\mu), \{h'(\mu)\}^2\ \sigma^2/n)$$

을 따른다고 할 수 있습니다.

예를 들어 앞의 지수분포의 사례에서 $\mu = \theta$, $\sigma^2 = \theta^2$이고

$$h(\theta) = \log_e \theta,\ h'(\theta) = \frac{1}{\theta}$$

이므로, 근사적으로

$$T_n (= \log_e \overline{X}) \sim \mathrm{N}(\log_e \theta,\ 1/n)$$

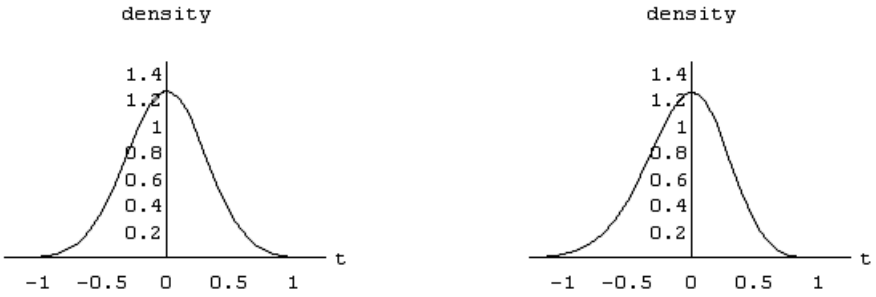
입니다. 따라서 T_n의 밀도에 대한 델타 근사는 다음과 같습니다.

$$f_{T_n}(t) \simeq \sqrt{\frac{n}{2\pi}}\ \exp(-\frac{n}{2}(t - \log_e \theta)^2).$$

이 경우도 T_n의 밀도를 정확히 유도할 수 있습니다. $\sum_{i=1}^{n} X_i = n \exp(T_n)$가 감마분포 Gamma($n, \theta$)를 따르므로

$$f_{T_n}(t) = \frac{\exp\{\, n \log_e n + n t - \frac{n e^t}{\theta}\}}{\Gamma(n)\theta^n},\ -\infty < t < \infty$$

가 정확한 밀도입니다. <그림 6>에서 두 밀도를 비교해보십시오 ($n = 10, \theta = 1$).



<그림 6> T_n의 밀도함수: 델타 방법에 의한 근사밀도(左)와 정확밀도(右)

델타 방법이 변수변환에 의하여 분산이 어떻게 달라지는지를 말해주기 때문에 소위 분산안정화 변환(variance stabilizing transform)을 유도하는 데에도 유용합니다. 예를 들어 $X_1, X_2, \cdots, X_n$을 Poisson(θ) 분포로부터의 iid 변수들이라고 합시다. 그러면

$$\overline{X} \sim \frac{1}{n} \text{Poisson}(n\theta)$$

라고 할 수 있습니다. 따라서 $E(\overline{X}) = \theta$, $Var(\overline{X}) = \theta/n$로서 평균과 분산이 θ에 의하여 연동(連動)되고 있읍니다. 이 때, $\overline{X}$를 $Y_n = h(\overline{X})$로 변환함으로써 변환변수 Y_n의 분산이 θ에 무관하게 되는 변환함수 $h(.)$를 찾아봅시다.

$h'(\theta) \neq 0$인 경우, 테일러 정리로부터

$$h(\overline{X}) \simeq h(\theta) + h'(\theta)(\overline{X} - \theta)$$

이므로

$$Var\{h(\overline{X})\} \simeq \{h'(\theta)\}^2 \; Var(\overline{X})$$

가 됩니다. 그런데 $Var\{h(\overline{X})\}$은 θ에 무관해야 하고 $Var(\overline{X}) = \theta/n$이므로

$$\text{constant} \simeq \{h'(\theta)\}^2 \frac{\theta}{n}$$

여야 합니다. 따라서

$$\{h'(\theta)\}^2 \simeq k_1 \frac{1}{\theta} \quad \Rightarrow h'(\theta) \simeq k_2 \theta^{-1/2}$$

$$\Rightarrow h(\theta) \simeq k_3 \theta^{1/2} \qquad (\text{여기서 } k_1, k_2, k_3 \text{는 상수}).$$

즉 제곱근 변환 $h(\overline{X}) \simeq k_3 \overline{X}^{1/2}$이 유도됩니다. 상수 k_3를 1로 잡으면, 델타 방법에 의하여

$$Y_n (= \sqrt{\overline{X}}) \sim \text{근사적으로 } \mathrm{N}(\theta^{1/2},\ 1/(4n)),$$

이 됩니다. 이와 같은 방법으로 분산안정화 변환을 지수분포와 베르누이 분포에 대하여도 유도할 수 있는데 그 결과는 다음과 같습니다 (연습문제 2.12와 2.13).

모(母)분포	분산안정화 변환	
지수분포	로그 변환	$\log_e \overline{X}$
베르누이 분포	역사인(arcsin) 변환	$\sin^{-1}\sqrt{p}$ $(p = \overline{X})$
포아송 분포	제곱근 변환	$\sqrt{\overline{X}}$

2.A 연습문제

2.1 확률변수 X에 대한 적률생성함수(mgf)를 $m_X(t)$라고 하면, $Y = a + bX$의 mgf가 $m_Y(t) = e^{ta} m_X(bt)$임을 보이세요.

2.2 ① 균일분포 Uniform(θ_1, θ_2)의 mgf를 구하세요.
② 지수분포 Exponential(θ)의 mgf를 구하세요.
③ 정규분포 N(μ,σ^2)의 mgf를 구하세요.

답: ① $\frac{e^{t\theta_2} - e^{t\theta_1}}{t(\theta_2 - \theta_1)}$ (if $t \neq 0$), $= 1$ (if $t = 0$). ② $\frac{1}{1 - t\theta}$, for $t < \frac{1}{\theta}$.
③ $\exp(\mu t + \sigma^2 t^2 / 2)$.

2.3 $X_1, \cdots, X_{30}$이 NB(10,0.5)로부터의 iid 확률변수라고 할 때, $P\{\overline{X} \leq 11\}$을 계산하고자 합니다. ① 정확한 확률값과 ② CLT 근사값을 계산하세요.

답: ① 0.8916 ② 0.8935 (연속성 보정값)

[Comment: $30\overline{X} \sim$ NB(300,0.5)임에 착안하여 정확한 확률값 계산을 위한 컴퓨터 프로그램을 짜보세요. 이 때 중간과정에서 가급적 $\infty \times 0$형의 연산이 나타나지 않도록 하여야 합니다.]

2.4 U_1, U_2를 Uniform$(-1,1)$로부터의 iid 확률변수라고 할 때, $X = U_1 / U_2$의 확률밀도함수를 유도하세요.

답: $f_X(x) = 1/4$ (for $|x| \leq 1$), $= 1/(4x^2)$ (for $|x| \geq 1$).

2.5 Beta(α,β) 분포의 평균과 분산이 다음과 같음을 보이세요.

$$E(U) = \frac{\alpha}{\alpha+\beta}, \quad Var(U) = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}.$$

2.6 $X_1, X_2, \cdots, X_n$을 N(μ, σ^2) 분포로부터의 iid 변수들이라고 할 때,

$$T = \frac{\overline{X} - \mu}{\sqrt{S^2/n}}$$

이 t$(n-1)$ 분포를 따름을 보이세요.

2.7 $X_1, X_2, \cdots, X_{n_1}$이 N$(\mu_1, \sigma_1^2)$ 분포로부터 독립적으로 생성되고, 이와 독립적으 로 $Y_1, Y_2, \cdots, Y_{n_2}$가 N$(\mu_2, \sigma_2^2)$ 분포로부터 독립적으로 생성될 때,

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$$

이 $F(n_1-1, n_2-1)$ 분포를 따름을 보이세요.

2.8 어떻게 와이블 분포 Weibull(γ,β)로부터 임의수를 만들 수 있습니까? 알고리즘을 제시하세요. 그리고 Weibull(2,1)로부터 임의수 100개를 발생시키고 그 평균을 구해보세요.

[Comment: 와이블 분포의 분포함수 $F(x)$를 구하고 $F(X)$가 Uniform(0,1)을 따름을 이용하세요. 참고로, Weibull(γ,β) 분포의 평균은 $\Gamma(1/\gamma+1)\,\beta^{1/\gamma}$입니다. 이것도 마저 보이세요.]

2.9 U, V가 독립적인 지수분포 Exponential(1)로부터 생성될 때, $X = U - V$의 밀도함수가 다음과 같음을 보이세요.

라플라스 분포 : $\quad f_X(x) = \frac{1}{2}\, e^{-|x|}, \; -\infty < x < \infty \quad .$

2.10 $X_1, \cdots, X_n$이 Uniform(0,1) 분포로부터의 iid 확률변수일 때,

$$Y_n = n\{1 - \max(X_1, \cdots, X_n)\}$$

의 극한분포를 유도하세요 $(n \to \infty)$.

답: Exponential(1).

2.11 확률변수 Y_n이 음이항분포 NB$(r, 1/n)$를 따를 때, n이 커짐에 따라 Y_n/n이 점차 Gamma$(r,1)$로 수렴함을 보이세요.

[Comment: $r=1$인 경우를 보이고, NB(r,θ) 확률변수가 r개의 Geometric(θ) 변수들의 합으로 표현됨을 활용하여, 앞에서 얻은 결과를 일반화하세요. 또는 mgf 기법을 활용해보세요.]

2.12* 로그변환이 Exponential(θ) 확률변수들의 평균에 대한 분산안정화변환임을 보이세요.

2.13* 역사인 제곱근 변환$\sin^{-1}\sqrt{p}$이 Bernoulli(θ) 확률변수들의 평균에 대한 분산안정화변환임을 보이세요.

[Comment: $\frac{d}{dx}\sin^{-1} p = \frac{1}{\sqrt{1-p^2}}$ 임을 활용하세요.]

2.14* $X_1, \cdots, X_n$을 Bernoulli(θ)로부터의 iid 확률변수들이라고 할 때 델타방법을 활용하여 $T_n = \overline{X}(1-\overline{X})$의 분포를 근사적으로 구해보세요.

답: $\theta \neq 0.5$인 경우, $\mathrm{N}\left(\theta(1-\theta), \frac{1}{n}\theta(1-\theta)(1-2\theta)^2\right)$.

$\theta = 0.5$인 경우, $\frac{1}{4} - \frac{1}{4n}\chi^2(1)$.

2.15 Poisson(1) 분포로부터의 iid 확률변수 $n(=10)$개의 평균 $\overline{X}$의 분포함수를 CLT 근사, 에지워스 근사, 정확한 방법으로 구하여 비교해보세요.

탐구문제 $X_1, \cdots, X_n$이 N(0,1) 분포로부터의 iid 확률변수일 때, 큰 n에 대하여 $T_n = \max(X_1, \cdots, X_n)$의 밀도함수를 구해 보세요. $n = 10$의 경우에 수리적 방법(if possible)과 몬테칼로 방법을 적용하여 비교해보세요.

2.B 읽을만한 책

적률생성함수, 대수의 법칙, 중심극한정리, 몬테칼로 모의시행 등에 대하여 보충이 필요하다면 다음 책을 보십시오.

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

정규분포로부터 파생되는 분포들에 대하여는 다음 수리통계학 책을 보십시오.

- Hogg, R.V. and Craig, A.T. (1995) *Introduction to Mathematical Statistics*, 5th Edition. Prentice Hall. (Chapter 4)

에지워스 근사나 델타방법에 대하여는 다음 책을 보십시오.

- Bickel, P.J. and Doksum, K.A. (1977) *Mathematical Statistics: Basic Ideas and Selected Topics*. Holden-Day. (Section 1.5)
- Cox, D.R. and Hinkley, D.V. (1974) *Theoretical Statistics*. Chapman and Hall. (Appendix 1)