표본 분포

들어가기

- 통계이론의 주된 활용은 연구대상에 대한 실험이나 관찰 등을 통해 구한 자료로부터 연구대상의 성격을 규명하는 것이다. 연구대상의 총체(totality)를 모두 조사하는 일은 많은 경우 시간 또는 경비 등의 제약으로 인해 불가능하다. 따라서 통계적 방법들은 그 총체의 일부인 표본을 조사함으로써 총체에 대한 성격을 귀납적(inductive)으로 추론하는 길을 택한다.

- 통계적 추론방법들은 불확실성을 계량화하기 위하여 확률을 사용한다.

모집단과 표본

- 연구 또는 관찰의 대상이 되는 총체(totality)를 모집단이라 한다.
- 랜덤표본(임의표본)의 수학적인 정의는?
  - 확률변수 $X_1, \dots, X_n$의 결합확률밀도함수가 $f_{X_1,\dots,X_n}(x_1, \dots, x_n) = f(x_1) \dots f(x_n)$ (단, $f(\cdot)$는 각 X_i의 공통 확률밀도함수)의 꼴로 나타나면, $X_1, \dots, X_n$을 모확률밀도함수가 $f(\cdot)$인 크기가 n인 랜덤표본 또는 임의표본이라고 한다. X_i는 i번째 관찰변수가 가질 값을 뜻한다. 일단 관찰된 알려진 값은 $x_1, \dots, x_n$으로 소문자를 사용하여 표기한다.
  - 랜덤표본을 표현할 때 "확률변수 $X_1, \dots, X_n$이 iid(independent and identically distributed)이다" 라고도 한다.

모집단과 표본 예제

- 모분포가 "성공"확률이 p인 베르누이 확률밀도함수는 $f(x) = p^x q^{1-x}$, $x = 0$ 또는 1, $(p + q = 1)$이다. 이 경우 서로 독립인 베르누이 시행을 10회 반복하였을 때의 표본을 $X_1, X_2, \dots, X_{10}$으로 표기하면, 이 표본의 결합 확률밀도함수는?
  - $f_{X_1,X_2,\dots,X_{10}}(x_1, x_2, \dots, x_{10}) = f(x_1)f(x_2) \dots f(x_{10}) = p^{\sum_{i=1}^{10} x_i} q^{10-\sum_{i=1}^{10} x_i}$ $(x_i = 0$ 또는 $1\ ;\ i = 1, \dots, 10)$
  - $X_1, \dots, X_{10}$을 모확률밀도함수가 $f(x) = p^x q^{1-x}$ $(x = 0$ 또는 1)인 크기 10의 랜덤표본이라고 한다.

통계량

- 미지의 모수를 포함하지 않는 랜덤표본의 함수 $T = T(X_1, \dots, X_n)$를 통계량(statistic)이라고 한다. 통계량 자체도 확률변수이다.
  - $X_1, \dots, X_n$ 이 확률밀도함수 $f(x; \theta)$ (θ는 알려지지 않은 모집단의 모수임)로부터 얻은 랜덤표본이라고 하자. 이 때, 표본평균 $\overline{X_n} = \sum_{i=1}^{n} \frac{X_i}{n}$ 또는 표본최댓값 $\max\{X_1, \dots, X_n\}$은 통계량이다. 그러나 $\overline{X_n} - \theta$ 또는 $\max\{\frac{X_1}{\theta}, \dots, \frac{X_n}{\theta}\}$ 등은 미지의 모수 θ에 의존하므로 통계량이 아니다.

표본평균과 표본분산

- 평균과 분산이 각각 μ와 σ^2인 확률밀도함수 $f(x)$로부터 랜덤표본 $X_1, \dots, X_n$을 얻었다. 이 때 표본평균은 $\overline{X_n} = \sum_{i=1}^{n} \frac{X_i}{n}$, 표본분산은 $S^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X_n})^2}{(n-1)}$ 이다.
- 대수의 법칙
  - 평균이 $\mu < \infty$ 인 확률밀도함수 $f(x)$로부터 랜덤표본 $X_1, \dots, X_n$을 얻었다면, 표본의 크기가 커짐에 따라 표본평균이 모평균으로 **확률적으로 수렴**한다.
  - 증명은 체비셰프의 부등식을 사용한다.

중심극한정리

- 중심극한정리(central limit theorem)는 확률론과 통계학에 있어서 가장 중요하다고 여겨지는 이론이다.
- 평균과 분산이 각각 μ와 $\sigma^2 < \infty$인 확률밀도함수 $f(x)$로부터의 랜덤표본 $X_1, \ldots, X_n$
  - 확률변수 $Z_n = \frac{\sum_{i=1}^{n} X_i - E(\sum_{i=1}^{n} X_i)}{\sqrt{Var(\sum_{i=1}^{n} X_i)}} = \frac{\sum_{i=1}^{n}(X_i - \mu)}{\sigma\sqrt{n}}$ 의 분포는 표본의 크기 n이 무한대에 접근함에 따라 표준정규분포 $N(0,1)$로 수렴한다.
- 모분포의 형태와 관계없이 유한한 평균과 분산만 존재하면 확률변수 Z_n의 분포는 표준정규분포로 수렴한다.

중심극한정리 예제 1

균등분포와 정규분포

- ✓ $X_1, \ldots, X_n$이 $U(0,1)$로부터 얻은 랜덤표본이라고 하자. 그러면 $E(X_i) = \frac{1}{2}$, $Var(X_i) = \frac{1}{12}$이다. 중심극한정리를 사용하여, $\sum_{i=1}^{n} X_i$의 분포를 $N(\frac{n}{2}, \frac{n}{12})$로 근사할 수 있다.
- ✓ 즉, 중심극한정리에 의해서 $\sum_{i=1}^{n} X_i$를 표준화한 $\frac{\sum_{i=1}^{n} X_i - \frac{n}{2}}{\sqrt{\frac{n}{12}}}$는 근사적으로 표준정규분포를 따른다.
- ✓ $P[a \le \sum_{i=1}^{n} X_i \le b] = P\left[\frac{a-\frac{n}{2}}{\sqrt{\frac{n}{12}}} \le \frac{\sum_{i=1}^{n} X_i - \frac{n}{2}}{\sqrt{\frac{n}{12}}} \le \frac{b-\frac{n}{2}}{\sqrt{\frac{n}{12}}}\right]$

$$\approx P\left[\frac{a-\frac{n}{2}}{\sqrt{\frac{n}{12}}} \le Z \le \frac{b-\frac{n}{2}}{\sqrt{\frac{n}{12}}}\right] = \Phi\left[\frac{b-\frac{n}{2}}{\sqrt{\frac{n}{12}}}\right] - \Phi\left[\frac{a-\frac{n}{2}}{\sqrt{\frac{n}{12}}}\right]$$

중심극한정리 예제 2

이항분포와 정규분포

- ✓ $X \sim B(n,p)$라고 하면 X는 서로 독립이며 "성공"확률이 p인 n개의 베르누이 확률변수 $X_1, X_2, \ldots, X_n$들의 합 $\sum_{i=1}^{n} X_i$와 같은 분포를 갖는다.
- ✓ $E(X_i) = p$, $Var(X_i) = pq$ $(q = 1 - p)$
- ✓ 중심극한정리에 의해서 n이 클 때 $\frac{(\sum_{i=1}^{n} X_i - np)}{\sqrt{npq}}$는 근사적으로 표준정규분포를 따른다.
- ✓ $P[a \le X \le b] = P[a \le \sum_{i=1}^{n} X_i \le b] = P\left[\frac{a-np}{\sqrt{npq}} \le \frac{\sum_{i=1}^{n} X_i - np}{\sqrt{npq}} \le \frac{b-np}{\sqrt{npq}}\right]$

$$\approx P\left(\frac{a-np}{\sqrt{npq}} \le Z \le \frac{b-np}{\sqrt{npq}}\right)$$

$$= \Phi\left[\frac{b-np}{\sqrt{npq}}\right] - \Phi\left[\frac{a-np}{\sqrt{npq}}\right]$$

중심극한정리 예제 2

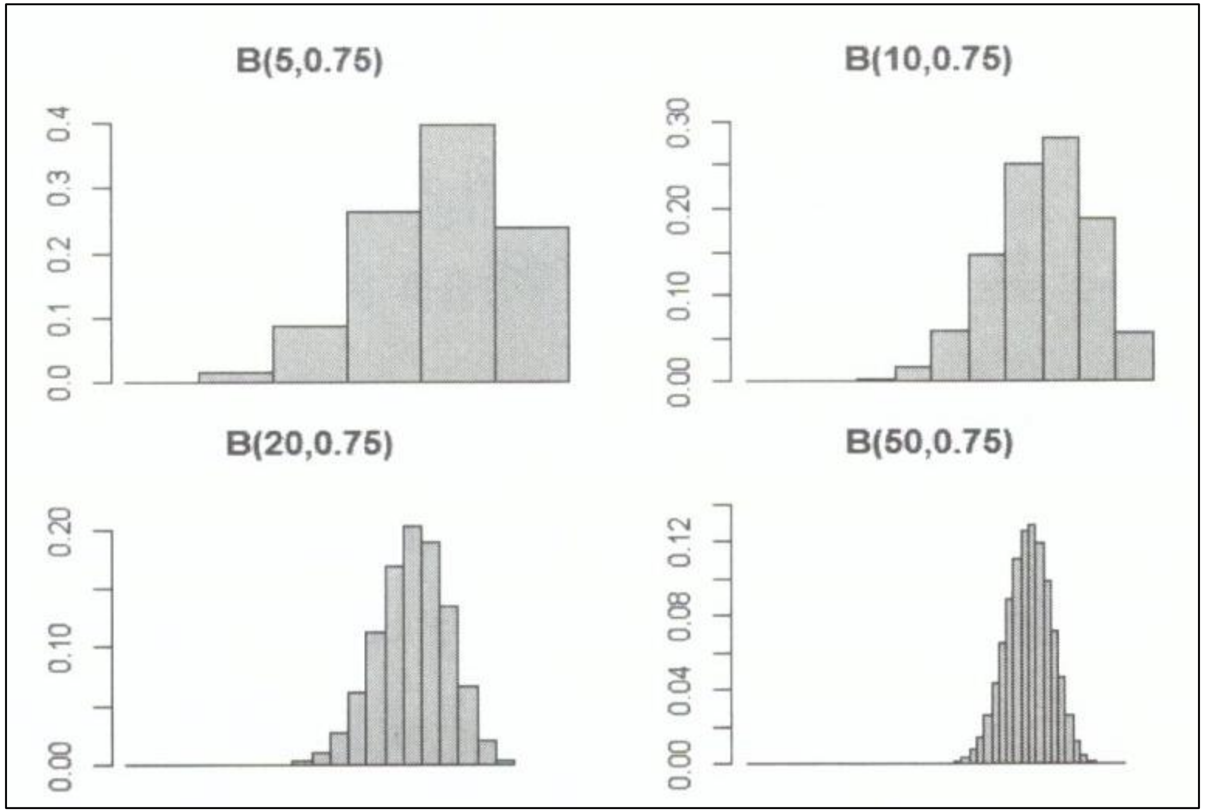
✓ 이항확률변수는 이산형인 반면 정규확률변수는 연속형이므로 이를 반영한 연속성 수정 (continuous correction)을 하여 근사의 정확성을 더 높일 수 있다.

✓ 연속성 수정

- 이항분포는 이산형으로서, 정수값만을 취할 수 있으므로 어떤 정수 k에 대해 $P(B(n,p) \le k) = P(B(n,p) < k+1)$이지만, 연속형인 정규분포는 같은 성질을 만족하지 않는다.
- 정규분포로 근사할 때, k나 $k+1$을 쓰지 않고, $k+\frac{1}{2}$을 쓰게 된다.
- $B(a \le B(n,p) \le b)$를 근사할 때, $B\left(a-\frac{1}{2} \le B(n,p) \le b+\frac{1}{2}\right)$를 이용하여 $\Phi\left[\frac{b-np}{\sqrt{npq}}\right] - \Phi\left[\frac{a-np}{\sqrt{npq}}\right]$로 근사한다.

중심극한정리 예제 2

✓ n의 크기에 따른 이항확률변수의 확률밀도함수

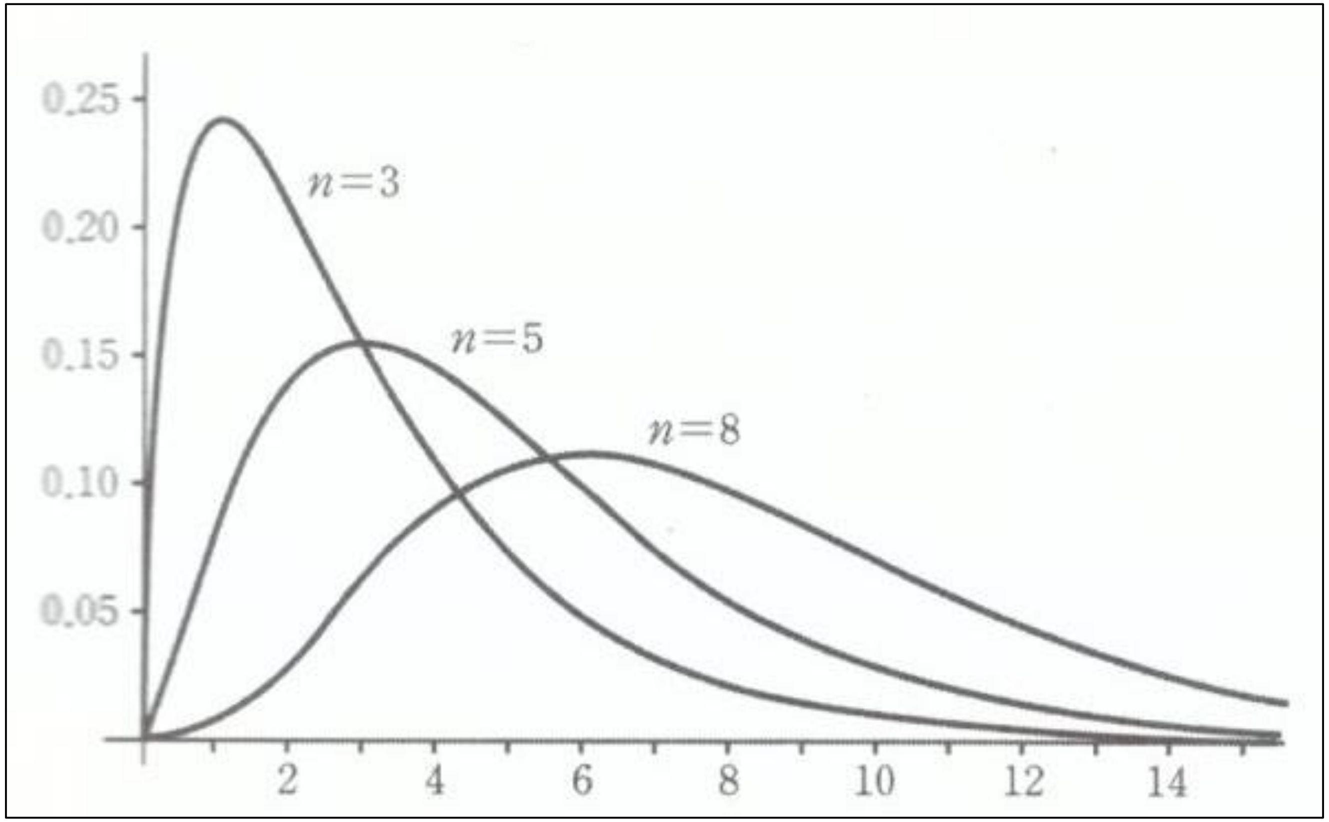


카이제곱분포

- 모수가 (k, θ)인 감마분포에서 $k = \frac{n}{2}$, $\theta = 2$ 인 확률변수 X, 즉, $X \sim GAM(\frac{n}{2}, 2)$는 자유도 (degrees of freedom)가 n인 카이제곱분포(chi-squared distribution)을 따른다고 한다.

- 자유도가 n인 카이제곱분포를 가지는 변수 X의 확률밀도함수는 $f_X = \frac{1}{\Gamma\left(\frac{n}{2}\right)\left(\frac{1}{2}\right)^{\frac{n}{2}}} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}$, $x > 0$ 이다.
  - $X \sim \chi^2(n)$ 일 때, X의 평균은 $E(X) = n$, $Var(X) = 2n$ 이 된다.
  - 서로 독립인 확률변수 $X_i (i = 1, \dots, n)$들이 각각 자유도가 k_i인 카이제곱분포를 따르면 그들의 합인 $Y = \sum_{i=1}^{n} X_i$는 자유도가 $\sum_{i=1}^{n} k_i$인 카이제곱분포를 따른다.

카이제곱분포의 pdf



자유도가 n인 카이제곱분포의 확률밀도함수

카이제곱분포

- 확률변수 Z가 $N(0,1)$ 분포를 따를 때, $Y = Z^2$은 $\chi^2(1)$ 분포를 가진다.

- 서로 독립인 확률변수 $X_i(i = 1, \dots, k)$ 들이 각각 정규분포 $N(\mu_i, \sigma_i^2)$을 따른다고 하면,
  - $V = \sum_{i=1}^{k} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2$은 **자유도가 k인 카이제곱분포**를 따른다.

- $X_1, \dots, X_n \sim N(\mu, \sigma^2)$이고, 크기가 n인 랜덤표본이라고 하면,
  - $\bar{X}_n$와 $S^2 = \sum_{i=1}^{n} \frac{(X_i - \bar{X}_n)^2}{(n-1)}$은 서로 독립이며,
  - $\frac{(n-1)S^2}{\sigma^2}$은 **자유도가 $n-1$인 카이제곱분포**를 따른다.

카이제곱분포 예제

- 어떤 종류의 제품을 1개 생산하는 데 걸리는 시간을 X라 하자. $X \sim N(6, 2^2)$라 하자. 10개의 랜덤표본의 표본분산이 5보다 클 확률은?
  - ✓ $\frac{9S^2}{4} \sim \chi^2(9)$
  - ✓ $P(S^2 > 5) = P\left[\frac{9S^2}{4} > \frac{45}{4}\right] = P[\chi^2(9) > 11.25] \approx 0.259$

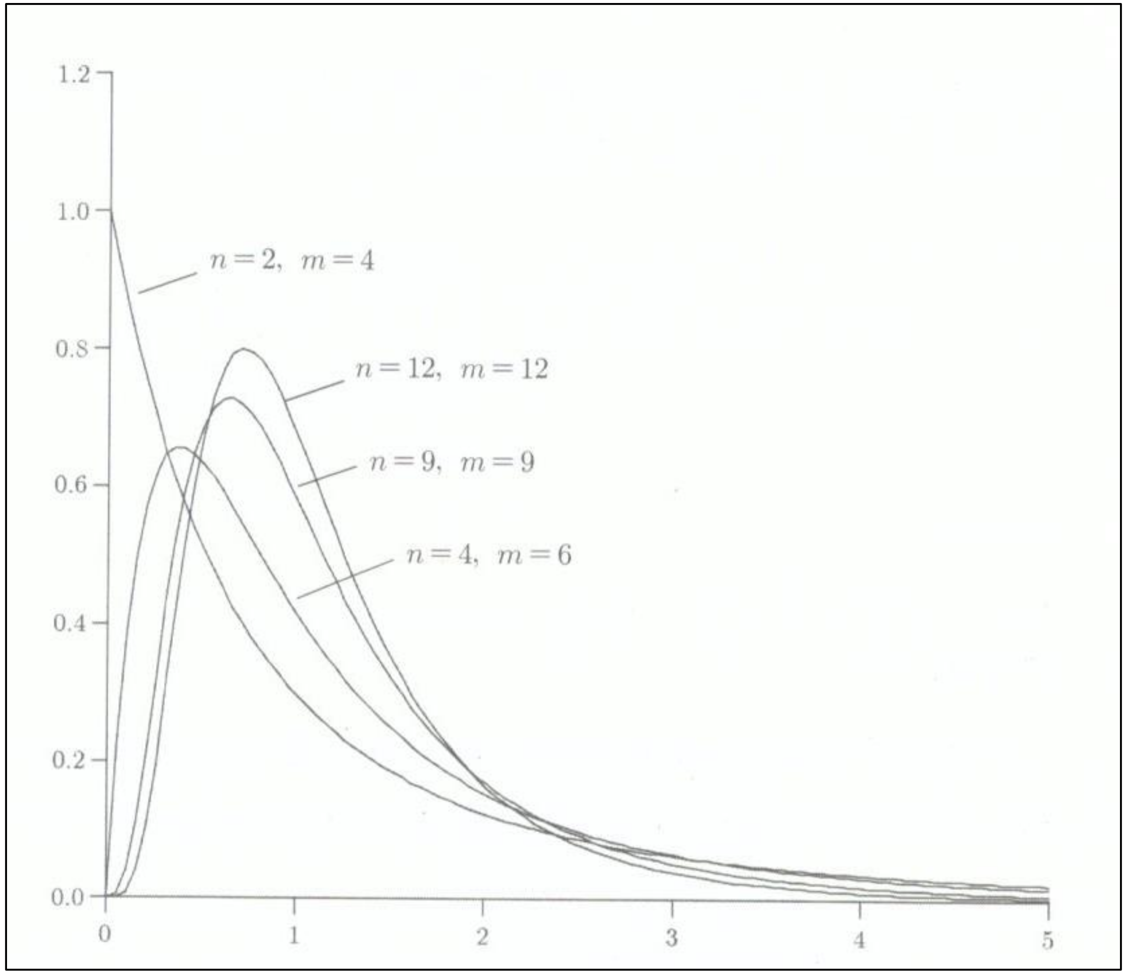
F분포

- F분포는 정규분포로부터 구한 독립인 **두 표본의 분산비**에 대한 분포를 설명하는 데 사용되며, 분산 분석에 활용한다.

- 서로 독립인 카이제곱확률변수 U와 V의 자유도가 각각 n, m이라고 할 때,
  - 변수 $\boldsymbol{X} = \frac{\boldsymbol{U}}{\boldsymbol{n}} / \frac{\boldsymbol{V}}{\boldsymbol{m}}$은 자유도가 (n, m)인 F분포를 따른다.
  - X의 확률밀도함수 $f_X(x) = \frac{\Gamma\left[\frac{n+m}{2}\right]}{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{m}{2}\right)}\left(\frac{n}{m}\right)^{\frac{n}{2}}\left(\frac{x^{\frac{n-2}{2}}}{\left[1+\frac{nx}{m}\right]^{\frac{n+m}{2}}}\right)$, $x > 0$ 이다. $X \sim F(n, m)$으로 표기한다.

- $X \sim F(n, m)$일 때,
  - X의 평균은 $E(X) = \frac{m}{m-2}$ $(m > 2)$
  - X의 분산은 $Var(X) = \frac{2m^2(m+n-2)}{n(m-2)^2(m-4)}$ $(m > 4)$ 이다.

F분포의 pdf



$F(n, m)$인 F분포의 확률밀도함수

F분포

- $\frac{1}{F_\alpha(n,m)} = F_{1-\alpha}(m,n)$

- $X_1, \dots, X_n$은 평균이 μ_X이고 분산이 σ_X^2인 정규분포로부터 얻은 크기가 n인 랜덤표본이고, $Y_1, \dots, Y_m$은 평균이 μ_Y이고 분산이 σ_Y^2인 정규분포로부터 얻은 크기가 m인 랜덤표본이다.
    - 두 표본이 서로 독립이면 확률변수 $F = \frac{S_X^2}{\sigma_X^2} / \frac{S_Y^2}{\sigma_Y^2}$은 **자유도가 $(n-1, m-1)$인 F분포**를 따른다.
    - $\sigma_X^2 = \sigma_Y^2$인 경우, 표본분산비 $\frac{S_X^2}{S_Y^2} = \frac{\sum_{i=1}^{n} \frac{(Y_i - \overline{Y_n})^2}{n-1}}{\sum_{i=1}^{m} \frac{(Y_i - \overline{Y_m})^2}{m-1}}$는 자유도가 $(n-1, m-1)$인 F분포를 따른다.

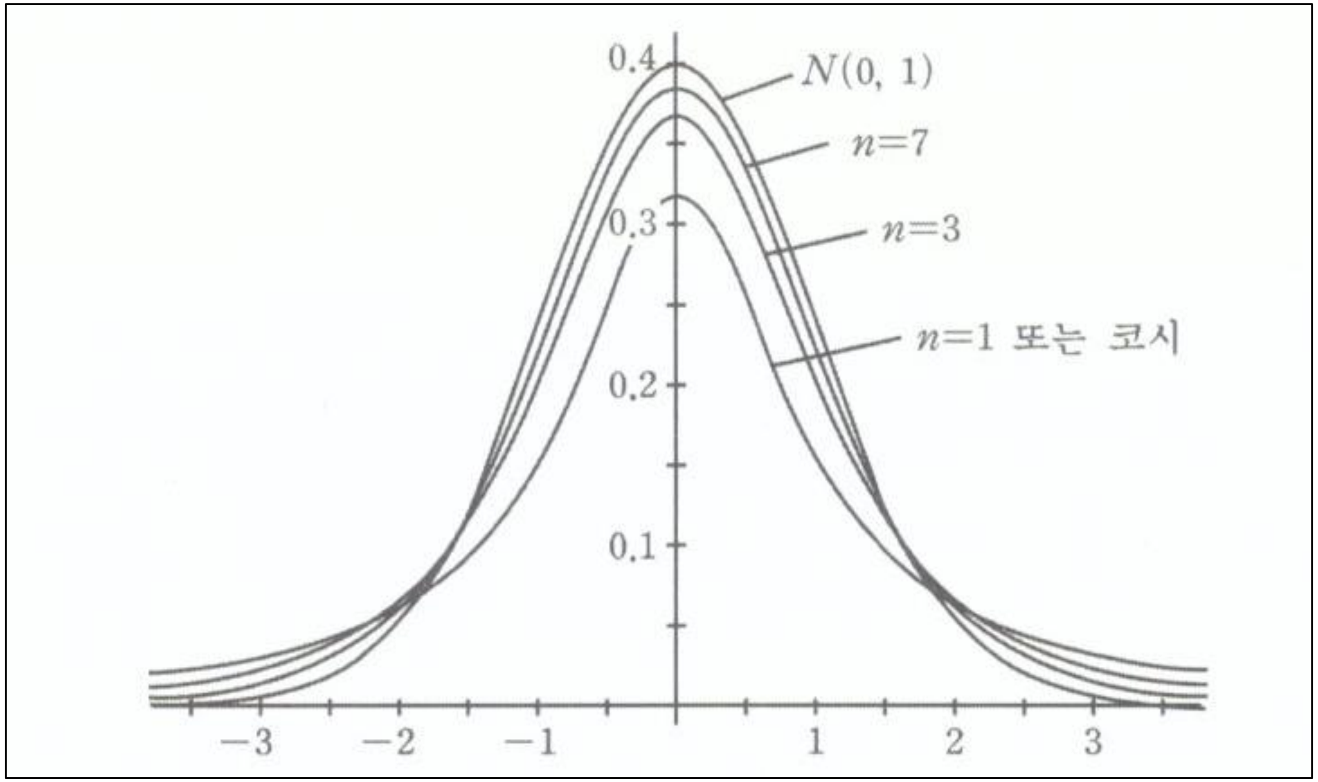
t분포

- t분포(Student's t distribution)는 정규분포의 모평균에 대한 가설검정 등을 포함한 검정론에서 매우 중요한 역할을 한다.

- 확률변수 Z가 표준정규분포를 따르며, U는 자유도 k인 카이제곱분포를 따르고, Z와 U가 서로 독립일 때, 확률변수 $\boldsymbol{X} = \frac{\boldsymbol{z}}{\sqrt{\boldsymbol{U}/\boldsymbol{k}}}$는 t분포를 따른다.

- t분포를 따르는 확률밀도함수 :
  - $f_X(x) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \frac{1}{\sqrt{k\pi}} \left(\frac{1}{\left(1+\frac{x^2}{k}\right)^{\frac{k+1}{2}}}\right)$, $(-\infty < x < \infty)$

t분포 pdf



$T(n)$인 t분포의 확률밀도함수

t분포

- 평균이 μ이고 분산이 σ^2인 정규분포로부터의 랜덤표본 $X_1, \ldots, X_n$을 고려
  - 확률변수 $T = \frac{\sqrt{n}(\overline{X_n}-\mu)}{\sqrt{\sum_{i=1}^{n}\frac{(X_i-\overline{X_n})^2}{n-1}}}$은 **자유도가 $(n-1)$인 t분포**를 따른다.
  - μ의 값이 알려져 있을 때 T를 스튜던트화 t통계량이라 한다.
  - μ의 값이 알려져 있는 않는 경우에도 T는 t분포를 따르지만 미지의 모수(μ)를 포함하므로 통계량은 아니다.
  - 스튜던트화 t통계량은 주로 모수의 추정량에 포함되는 장애모수(nuisance parameter)인 분산의 추정에서 비롯된다. (장애모수는 추정하고자 하는 모수 이외의 모수이다.)
  - 자유도 $(n-1)$이 커짐에 따라 표준정규분포로 가까워진다.