제2강 확률분포이론 개요

1. 확률변수, 확률분포, 확률함수
2. 두 확률변수의 분포
3. 확률변수의 기대값 및 분산
4. 이론적 확률분포

1. 확률변수, 확률분포, 확률함수

(1) 확률변수, 확률분포, 확률함수

① 확률변수(random variable) : 임의실험의 결과에 의해 그 값이 결정되는 변수를 말하며 기호로는 X,Y,Z 등으로 나타내고 그 값은 x,y,z 등으로 씀

- 이산확률변수(discrete r.v.) : 항상 정수의 값을 취하는 확률변수(방 수, 가족 수...)
- 연속확률변수(continuous r.v.) : 일정한 범위 내 모든 실수값을 연속적으로 택하는 확률변수(예:키, 몸무게..)

② 확률분포(probability distribution) : 확률변수 X의 취하는 모든 값과 이에 대응하는 확률을 나타낸 것

- 이산확률분포 : 이산확률변수의 확률분포(예: 베르누이분포, 이항분포, 포아송분포)
- 연속확률분포 : 연속확률변수의 확률분포(예 : 균등분포, (표준)정규분포, t-분포, χ^2-분포, F-분포)

③확률함수(probability function) : 확률의 크기를 표현하는 함수

- 확률질량함수(probability mass function: PMF) : 이산확률변수 X가 어느 특정한 값 x를 취할 확률이 어떻게 결정되는가를 보여주는 함수이며 다음의 2조건을 만족시킴

$$f(x) = P(X = x)$$

· $0 \le f(x) \le 1 \; for \; \forall x$

· $\sum f(f) = 1$

- 확률밀도함수(probability density function: PDF) : 연속확률변수 X가 어느 특정한 구간에 포함될 확률이 어떻게 결정되는가를 보여주는 함수이며 다음의 3조건을 만족시킴

· 모든 x값에 대해서 $f(x) \ge 0$

· 구간 (a, b)의 확률은 $\int_a^b f(x)dx$

· X의 모든 가능한 값의 확률은 $\int_{-\infty}^{\infty} f(x)dx = 1$

2. 두 확률변수의 분포

(1) 결합확률분포

① 결합확률질량함수(joint PMF) : 두 변수 X, Y가 이산확률변수라고 할 때, X는 x의 값을, Y는 y의 값을 동시에 갖는 확률이 어떻게 결정되는가를 보여주는 함수이며 다음의 2조건을 만족시킴

$$f(x,y) = P(X = x, Y = y)$$

· $f(x,y) \geq 0 \; for \; \forall x,y$

· $\sum_x \sum_y f(x,y) = 1$

② 결합확률밀도함수(joint PDF) : 두 변수 X, Y가 연속확률변수라고 할 때, 두 변수가 동시에 어느 특정한 구간에 포함될 확률이 어떻게 결정되는가를 보여주는 함수이며 다음의 3조건을 만족시킴

· $f(x,y) \geq 0$

· $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x,y) dx dy = 1$

· $\int_a^b \int_c^d f(x,y) dx dy = P(a \leq xb, c \leq y < d)$

(2) 주변확률분포

① 주변확률질량함수(marginal PMF) : 이산확률변수인 두 변수 X, Y의 결합분포에서 X 또는 Y의 어느 하나만의 확률질량함수

$f(x) = \sum_y f(x,y)$ (marginal PMF of X) / $f(y) = \sum_x f(x,y)$ (marginal PMF of Y)

② 주변확률밀도함수(marginal PDF) : 연속확률변수인 두 변수 X, Y의 결합분포에서 X 또는 Y의 어느 하나만의 확률밀도함수

$f(x) = \int_{\in y} f(x,y)\, dy$ (marginal PDF of X) / $f(y) = \int_{\in x} f(x,y)\, dx$ (marginal PDF of Y)

(3) 조건부확률분포

- 조건부확률함수(conditional PF) : 두 확률변수 X, Y 의 결합확률함수를 $f(x,y)$라고 할 때, 어느 한 변수가 특정 값을 취한다는 조건 하에 다른 변수가 값을 취할 확률을 나타내는 함수

$$f(x|y) = P(X = x\,|Y = y) = \frac{f(x,y)}{f(y)}$$ (conditional PF of X)

$$f(y|x) = P(Y = y|X = x) = \frac{f(x,y)}{f(x)}$$ (conditional PF of X)

단, $f(x)$ 및 $f(y)$는 각각 주변확률함수를 나타내고, $f(x,y)$는 결합확률함수를 나타냄.

(4) 두 확률변수의 독립성

- 두 확률변수 X,Y의 결합확률함수 $f(x,y)$의 주변확률함수를 각각 $f(x)$, $f(y)$라고 할 때,

$f(x,y) = f(x)f(y)$이면

두 확률변수는 서로 독립이라고 함

3. 확률변수의 기대값 및 분산

(1) 확률변수

① 기대값(Expected Value) : 확률분포에서 분포의 무게중심을 말하며, 확률값을 가중치로 하는 확률변수의 가능한 값에 대한 가중평균이라고 할 수 있음

- 이산확률변수 : $E(X) = \sum_x f(x) = \mu$
- 연속확률변수 : $E(X) = \int_{-\infty}^{\infty} xf(x)dx = \mu$

② 분산(Variance) : 확률분포의 기대값을 중심으로 흩어진 정도를 측정함

- 이산확률변수 : $Var(X) = E\left[(X-\mu)^2\right] = E\left(X^2\right) - \mu^2 = \sigma^2$
- 연속확률변수 : $Var(X) = \int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2 = \sigma^2$

(2) 두 확률변수

① 기대값

- 두 확률변수의 합 : $E(X+Y) = \sum_x \sum_y (x+y)\, f(x,y) = \mu_x + \mu_y$
- 두 확률변수의 곱 : $E(XY) = \sum_x \sum_y (xy)\, f(x,y)$
- 독립인 두 확률변수의 곱 : $E(XY) = E(X)E(Y) = \mu_X \mu_Y$

② 공분산(Covariance) : 두 확률변수 X와 Y의 선형관계를 나타냄

- $Cov(X,Y) = E[X - E(X)][Y - E(Y)] = E[X - \mu_X][Y - \mu_Y] = \sigma_{XY}$
  - $Cov(X,Y) > 0$ → X,Y는 正(+)의 선형관계, 즉, X가 커지면 Y도 커지고 X가 작아지면 Y도 작아짐
  - $Cov(X,Y) < 0$ → X,Y는 負(−)의 선형관계, 즉, X가 커지면 Y는 작아지고 X가 작아지면 Y는 커짐
  - $Cov(X,Y) = 0$ → X,Y는 선형의 관계를 가지고 있지 않음
  - $Cov(X,Y) = \frac{\sum_{i=1}^{n}(X_i-\bar{X})\sum_{i=1}^{n}(Y_i-\bar{Y})}{n-1}$
  - 공분산의 크기는 X, Y의 측정단위에 영향을 받으므로 상관관계의 크기를 측정하는 지표로서는 부적합

③ 상관계수(Correlation coefficient) : 두 확률변수 X와 Y의 선형관계를 나타냄

- $Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \rho_{XY}$
  - $0 < \rho_{XY} \leq 1$ → X,Y는 正(+)의 선형관계, 즉, X가 커지면 Y도 커지고 X가 작아지면 Y도 작아짐
    $\rho_{XY} = 1$인 경우 완전 정(+)의 상관관계를 가짐
  - $-1 \leq \rho_{XY} < 0$ → X,Y는 負(−)의 선형관계, 즉, X가 커지면 Y는 작아지고 X가 작아지면 Y는 커짐
    $\rho_{XY} = -1$인 경우 완전 부(−)의 상관관계를 가짐
  - $\rho_{XY} = 0$ → X,Y는 선형의 관계를 가지고 있지 않음
  - $\rho_{XY} = \frac{\sum_{i=1}^{n}(X_i-\bar{X})\sum_{i=1}^{n}(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2 \sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$
  - X, Y의 측정단위에 영향을 받지 않는(unit free) 상관계수가 상관관계의 정도를 측정하는데 적합

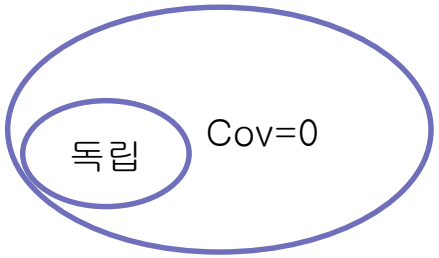
④ 독립성(independence) : 만약에 두 확률변수 X와 Y가 서로 독립이면 다음의 관계가 성립함

- $\boldsymbol{Cov(X,Y) = E(XY) - E(X)E(Y) = 0}$ $\because \boldsymbol{E(XY) = E(X)E(Y)}$ 만약에 X, Y가 독립이면
- $\boldsymbol{Var(X \pm Y) = Var(X) + var(Y)}$
- $\boldsymbol{\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = 0}$

(주의)

- 두 확률변수 X, Y가 서로 독립이면 $\boldsymbol{Cov(X,Y) = 0}$이지만 $\boldsymbol{Cov(X,Y) = 0}$ 이라고 해서 반드시 X, Y가 서로 독립인 것은 아님
- 즉, $\boldsymbol{Cov(X,Y) = 0}$ 은 독립이기 위한 필요조건이지 충분조건은 아님

독립 ⟷ (O →, ← X) $\boldsymbol{Cov(X,Y) = 0}$



(2) 연산자(operator)

① 기대 연산자

- $E(c) = c$
- $E(X \pm c) = E(X) \pm c$
- $E(cX) = cE(X)$
- $E(a \pm bX) = a \pm bE(X)$
- $E(aX \pm bY) = aE(X) \pm bE(Y)$

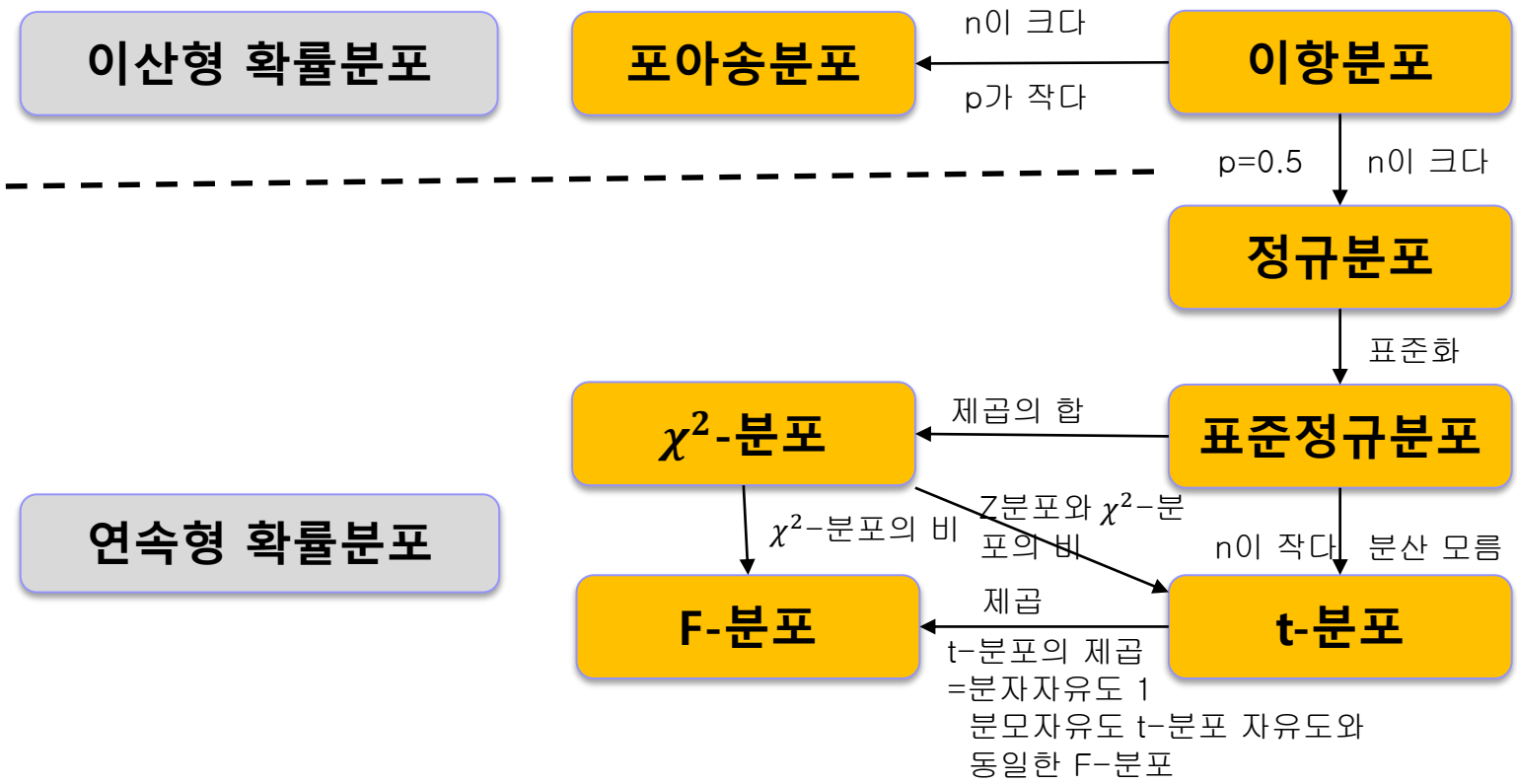
② 분산 연산자

- $Var(c) = 0$
- $Var(cX) = c^2 Var(X)$
- $Var(X + c) = Var(X)$
- $Var(X \pm Y) = Var(X) + Var(Y)$ 만약 X, Y가 독립이면
- $Var(aX \pm bY) = a^2 Var(X) + b^2 Var(Y) \pm 2abCov(X, Y)$ 만약 X, Y가 독립이 아니면

4. 이론적 확률분포

(1) 이론적 확률분포의 관계

-통계분석에서 자료를 수집하고 그 수집된 자료로부터 어떤 정보를 얻고자 하는 경우에는 항상 수집된 자료가 특정한 확률분포를 따른다고 가정



(2) 이산확률분포

① 베르누이분포(Bernoulli distribution)

- $P(X=x)=p^x(1-p)^{1-x}, x=0,1$
- $\boldsymbol{X \sim B(1,p)}$
- $\boldsymbol{E(X)=p, Var(X)=pq, q=1-p}$

② 이항분포(binomial distribution)

- $P(X=k)=\binom{n}{k}p^k q^{n-k}, k=0,1,2,\dots,k$
- $\boldsymbol{X \sim B(n,p)}$
- $\boldsymbol{E(X)=np, Var(X)=npq, q=1-p}$

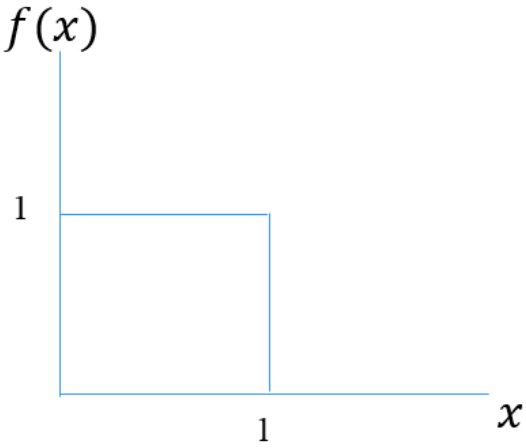
③ 포아송분포(poisson distribution)

- $P(X=k)=\frac{\mu^k}{k!}e^{-\mu}, k=0,1,2,3,\dots$
- $\boldsymbol{X \sim B(\mu)}$
- $\boldsymbol{E(X)=\mu, Var(X)=\mu}$

(3) 연속확률분포

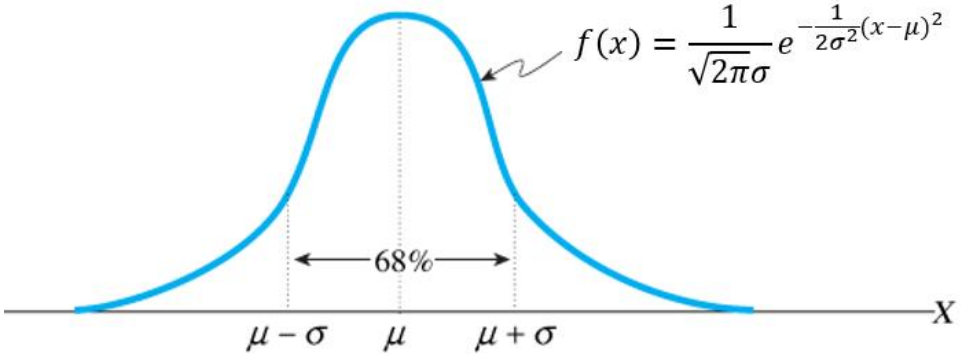
① 균등분포(uniform distribution)

- $f(x) = \begin{cases} \frac{1}{b-a}, a \le X \le b \\ 0, \text{다른 곳에서} \end{cases}$
- $\boldsymbol{X \sim U(a,b)}$
- $\boldsymbol{E(X) = \frac{a+b}{2}, Var(X) = \frac{(b-a)^2}{12}}$



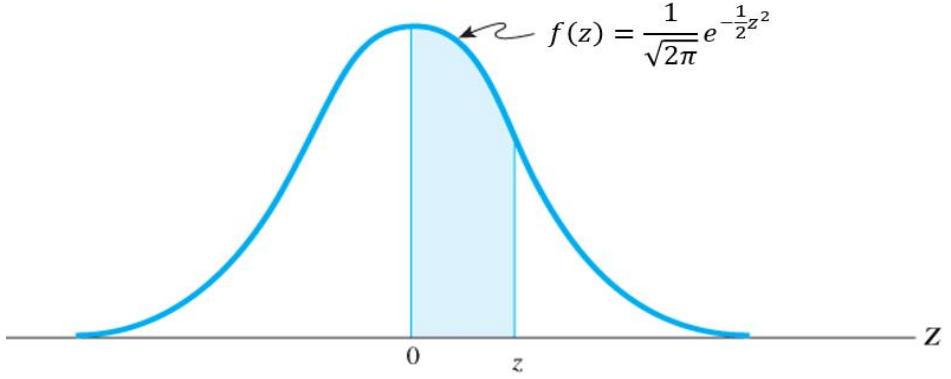
② 정규분포(normal distribution)

- $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}, -\infty < x < \infty$
- $\boldsymbol{X \sim N(\mu, \sigma^2)}$
- $\boldsymbol{E(X) = \mu, Var(X) = \sigma^2}$



③ 표준정규분포(standard normal distribution)

- $f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}, -\infty < z < \infty$
- $\boldsymbol{Z = \frac{X-\mu}{\sigma} \sim N(0,1)}$
- $\boldsymbol{E(Z) = 0, Var(Z) = 1}$

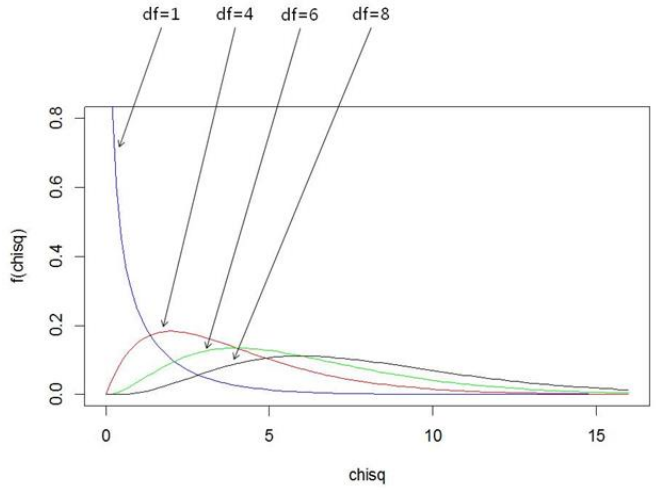


① χ^2 –분포(chi-square distribution)

– 확률변수 $Z_1, Z_2, \ldots, Z_n$이 서로 독립적으로 표준정규분포 $Z_i \sim N(0,1)$을 따를 때, $Z_1, Z_2, \ldots, Z_n$의 제곱합 $\sum_{i=1}^{n} Z_i^2$은 자유도가 n인 χ^2 –분포를 따름

– $X \sim \chi_v^2$일 때

– $E(X) = v, Var(X) = 2v$



② t –분포(Student's t-distribution)

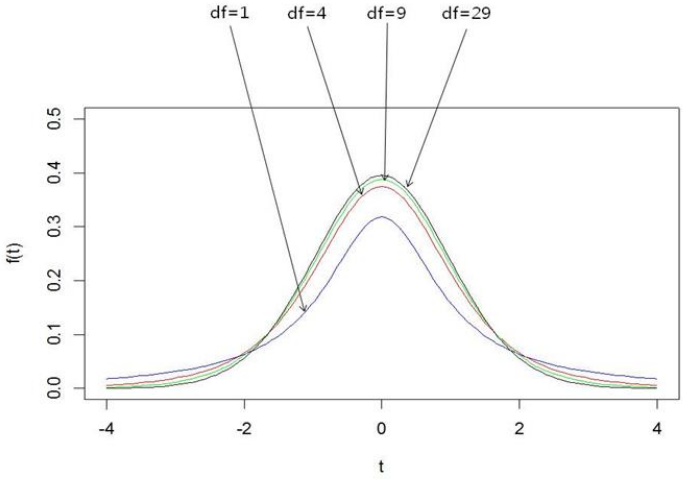
– $Z \sim N(0,1)$, $V \sim \chi_v^2$이고 Z와 V가 독립이면, $T = \frac{Z}{\sqrt{\frac{V}{v}}} \sim t_v$임

– $X \sim t_v$일 때

– $E(X) = 0, Var(X) = \frac{v}{v-2}$

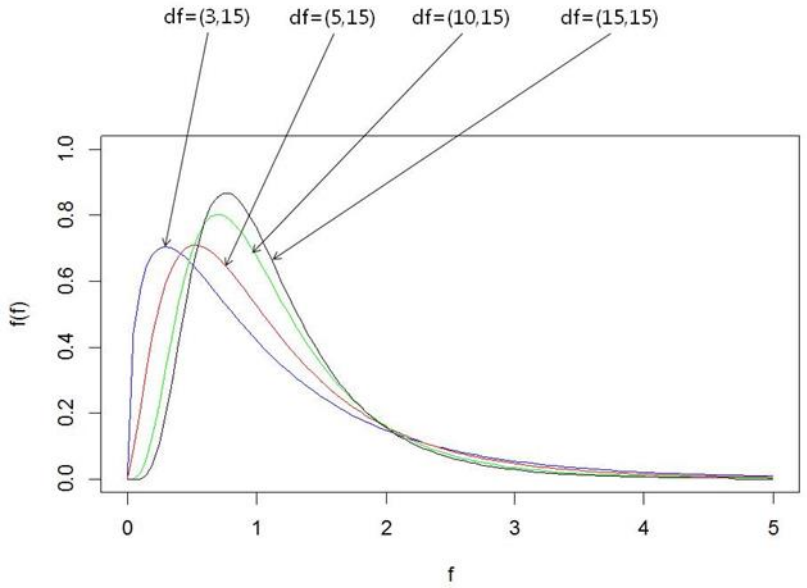
– 자유도가 무한히 커지면 t-분포는 표준정규분포에 접근함

– 예를 들어, 표준정규분포와 자유도가 4인 t분포를 비교한 그림으로 t분포가 표준정규분포보다 꼬리부분의 확률이 조금 더 큰 것을 알 수 있음



③ F-분포(Snedecor's F- distribution)

- $X_1 \sim \chi^2_{v_1}$, $X_2 \sim \chi^2_{v_2}$이고 X_1, X_2이 서로 독립이면, $F = \frac{\frac{X_1}{v_1}}{\frac{X_2}{v_2}} \sim F\binom{v_1}{v_2}$임
- $E(X) = \frac{v_2}{v_2 - 2}$
- 분산은 복잡함
- 자유도가 커지면서 분모와 분자의 자유도가 같아질수록 정규분포와 비슷하게 됨



- 만약에 $T \sim t_n$이면, $T^2 \sim F\binom{1}{n}$임

$$T^2 = \frac{Z^2}{\frac{V}{v}} = \frac{\frac{\chi^2}{1}}{\frac{V}{v}} \sim F\binom{1}{v}$$임

이론적 확률분포(요약)

분포표기 및 모수	확률함수$(f(x))$	평균	분산
$X\sim B(1,p)$	$p^x q^{1-x}, x=0,1$	p	$pq, (p+q=1)$
$X\sim B(n,p)$	$\binom{n}{x} p^x q^{n-x}, x=0,1,\dots,n$	np	$npq, (p+q=1)$
$X\sim P(\mu)$	$\frac{e^{-\mu}\mu^x}{x!}, x=0,1,\dots..$	μ	μ
$X\sim U(a,b)$	$\frac{1}{b-a}, a<x<b$	$\frac{a+b}{2}$	$\frac{(b-a)^2}{12}$
$X\sim N(\mu,\sigma^2)$	$\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}, -\infty<x<\infty$	μ	σ^2
$X\sim \chi^2(v)$	복잡함	v	$2v$
$X\sim t(v)$	복잡함	0	$\frac{v}{v-2}$
$X\sim F(v_1,v_2)$	복잡함	$\frac{v_2}{v_2-2}$	$\frac{2v_2^2(v_1+v_2-2)}{v_1(v_2-2)^2(v_2-4)}$