김은영

제1장 기초통계이론

1.3 확률변수와 확률표본

1.3.4 확률변수들의 선형결합

1. 적률생성함수(moment generating function: m.g.f.)

(1) 정의: 확률변수 Y의 적률생성함수

$$M_Y(t) = E\left[e^{tY}\right],\ \text{여기서}\ |t| < h,\ h > 0$$

※ 적률생성함수는 Y의 함수가 아니라, t의 함수다.

(2) 멱급수전개(power series expansion)에 의하면

$$M_Y(t) = \sum_{j=0}^{\infty} \frac{\mu_j t^j}{j!},\ \text{여기서}\ \mu_j = E(Y^j):\ j\text{번째 적률}$$

(3) 테일러전개(Taylor expansion)를 사용하면

$$M_Y(t) = E\left[e^{tY}\right] = E\left[1 + tY + \frac{t^2Y^2}{2!} + \frac{t^3Y^3}{3!} + \cdots\right] = 1 + tE(Y) + \frac{t^2}{2!}E(Y^2) + \cdots$$

$$= \sum_{j=0}^{\infty} \frac{\mu_j t^j}{j!}$$

$$\frac{\partial}{\partial t}M_Y(t) = M_Y^{(1)}(t) = E(Y) + tE(Y^2) + \frac{t^2}{2}E(Y^3) + \cdots$$

$M_Y^{(1)}(0) = E(Y) = \mu_1$: 1차 적률(the first moment)

$$\frac{\partial^2}{\partial t^2}M_Y(t) = E(Y^2) + tE(Y^3) + \cdots$$

$M_Y^{(2)}(0) = E(Y^2) = \mu_2$: 2차 적률(the second moment)

일반화하면, $M_Y^{(j)}(0) = E(Y^j) = \mu_j$: j차 적률(the j-th moment)

2. 확률변수들의 선형결합

(1) 확률변수 $Y_1, Y_2, \cdots, Y_n$의 선형결합: $\sum_{i=1}^{n} a_i Y_i = a_1 Y_1 + a_2 Y_2 + \cdots + a_n Y_n$

※ 이 선형함수도 하나의 확률변수다.

(2) 기댓값

$$E\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} a_i E(Y_i) = a_1 E(Y_1) + a_2 E(Y_2) + \cdots + a_n E(Y_n)$$

(3) 분산

$$Var\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j Cov(Y_i, Y_j)$$

만약 Y_i들이 서로 독립이면 $Cov(Y_i, Y_j) = 0\ (i \neq j)$이므로

$$Var\left(\sum_{i=1}^{n} a_i Y_i\right) = \sum_{i=1}^{n} a_i^2 Var(Y_i)$$

더 일반적인 경우로

확률변수 $Y_1, Y_2, \cdots, Y_n$들의 두 선형함수 $\sum_{i=1}^{n} a_i Y_i$와 $\sum_{i=1}^{n} b_i Y_i$ 간의 공분산은

$$Cov\left(\sum_{i=1}^{n} a_i Y_i, \sum_{j=1}^{n} b_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i b_j Cov(Y_i, Y_j)$$

만약 Y_i들이 서로 독립이면

$$Cov\left(\sum_{i=1}^{n} a_i Y_i, \sum_{i=1}^{n} b_i Y_i\right) = \sum_{i=1}^{n} a_i b_i Var(Y_i)$$

(3) 적률과의 관계

$$\begin{aligned} E[Y - E(Y)]^2 &= E[Y^2 - 2YE(Y) + E^2(Y)] \\ &= E(Y^2) - 2E(Y)E(Y) + E^2(Y) \\ &= E(Y^2) - E^2(Y) \end{aligned}$$

여기서, $E(Y^2)$: 2차 적률

$E^2(Y)$: 1차 적률의 제곱

1.3.5 이산확률분포

1. 이항분포(binomial distribution)

(1) 이항분포(binomial distribution)

n: 베르누이 시행의 횟수 p: 성공확률 X: n번 시행에서 성공한 횟수	확률변수 X: 이항확률변수 확률변수 X의 분포: 이항분포
$f(x)=P[X=x]=\binom{n}{x}p^x q^{n-x},\ x=0,1,\cdots,n$	
m.g.f. $M_X(t)=E\left[e^{tx}\right]=\sum_{x=1}^{n}e^{tx}\binom{n}{x}p^x(1-p)^{n-x}=\sum_{x=1}^{n}\binom{n}{x}(pe^t)^x(1-p)^{n-x}$ $=\left[pe^t+(1-p)\right]^n$	
➽ 평균(mean)=np ➽ 분산(variance)=npq 여기서, $q=1-p$ ➽ 표준편차(standard deviation)= $\sqrt{npq}$ cf) 1차 적률 $E(X)=M_X^{(1)}(0)$ $M_X^{(1)}(t)=n\left[pe^t+(1-p)\right]^{n-1}pe^t$ $M_X^{(1)}(0)=n\left[pe^0+(1-p)\right]^{n-1}pe^0=np$	

(2) 누적이항분포표(cumulative binomial distribution table), (n,p)가 주어지면 표에서 각 c에 해당하는 값은 아래 누적확률을 나타낸다.

$$P[X\le c]=\sum_{x=0}^{c}f(x)$$

이항분포

x	$f(x)$
0	$f(0)$
1	$f(1)$
2	$f(2)$
⋮	⋮
n	$f(n)$
합계	1

누적이항분포표

c	$P[X\le c]=\sum_{x=0}^{c}f(x)$
0	$f(0)$
1	$f(0)+f(1)$
2	$f(0)+f(1)+f(2)$
⋮	⋮
n	$f(0)+f(1)+f(2)+\cdots+f(n)=1$

[예제] 근육통을 치료하는 새로운 방법이 50%의 성공률을 거두는 것으로 알려져 있다. 15명의 환자를 대상으로 치료했을 때 다음의 확률은 얼마인가?

① 기껏해야 6명이 치료될 것이다.

② 6명 이상 10명 이하가 치료될 것이다.

③ 12명 이상 치료될 것이다.

【풀이】

확률변수 X: 치료된 환자의 수, $n=15$이고 $p=0.5$인 이항분포를 따른다.

① $P[X \le 6]=0.304$

② $P[6 \le X \le 10]=f(6)+f(7)+f(8)+f(9)+f(10)$

$$=P[X \le 10]-P[X \le 5]=0.941-0.151=0.790$$

③ $P[X \ge 12]=1-P[X \le 11]=1-0.982=0.018$

[예제] $n=3$이고 $p=0.5$인 이항분포의 평균과 표준편차를 구하라.

【풀이】

☺ 평균$=np=3\times 0.5=1.5$

☺ 표준편차$=\sqrt{npq}=\sqrt{3\times 0.5\times 0.5}=\sqrt{0.75}=0.866$

2. 포아송분포(Poisson distribution)

양의 실수 λ에 대하여 확률질량함수가 다음과 같은 확률변수 X의 확률분포 $X \sim P(\lambda)$
$$f(x)=\frac{e^{-\lambda}\lambda^x}{x!},\ x=0,1,\cdots$$
m.g.f. $M_X(t)=E[e^{tx}]=\sum_{x=0}^{\infty} e^{tx}\frac{e^{-\lambda}\lambda^x}{x!}=e^{-\lambda}\sum_{x=0}^{\infty}\frac{(\lambda e^t)^x}{x!}=e^{-\lambda}e^{\lambda e^t}=\exp[\lambda(e^t-1)]$
➨ 평균(mean) = λ ➨ 분산(variance) = λ ➨ 표준편차(standard deviation)= $\sqrt{\lambda}$ cf) 1차 적률 $E(X)=M_X^{(1)}(0)$ $M_X^{(1)}(t)=\lambda e^t \bullet \exp[\lambda(e^t-1)]$ $M_X^{(1)}(0)=\lambda e^0 \bullet \exp[\lambda(e^0-1)]=\lambda$

[예제] 어떤 공정라인에서 생산된 제품의 불량률은 0.05이다. 이 공정라인에서 생산된 제품 20개를 임의로 선정했을 때, 다음을 구하라.

① 1개 이하의 불량품이 나올 확률

② 포아송분포에 의한 ①의 근사확률

【풀이】

① X: 불량품의 수, $X \sim B(20, 0.05)$

$P(X=1) = P(X \le 1) - P(X=0) = 0.7358 - 0.3585 = 0.3773$

② $n=20$, $p=0.05$ → X는 $\mu=1$인 포아송분포에 근접한다.

⇒ 포아송분포에 의한 ①의 근사확률은

$P(X=1) = P(X \le 1) - P(X=0) = 0.736 - 0.368 = 0.368$

1.3.6 정규분포

1. 정규분포

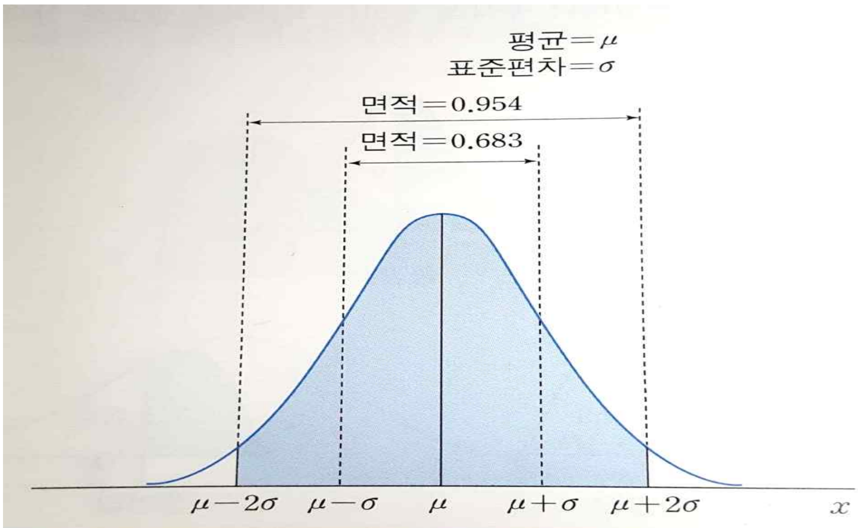
(1) 정규분포는 Pierre Laplace와 Carl Gauss에 의해 발전되었다.

(2) 정규분포의 확률밀도함수

$X \sim N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2},\ -\infty < x < \infty,\ \pi = 3.1416,\ e = 2.7183$$

(3) 정규분포의 그림



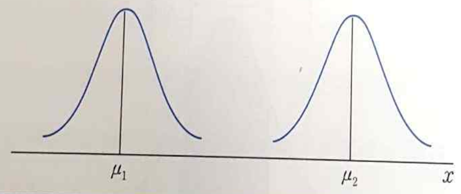


그림 5-6 | 평균($\mu_1 < \mu_2$)이 다르고 표준편차가 같은 두 정규분포

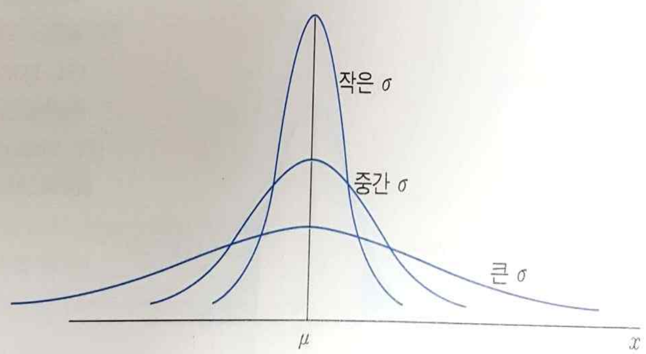


그림 5-7 | 평균이 같고 표준편차가 다른 정규분포들

2. 표준정규분포

(1) 표준정규분포의 정의

$$X \sim N(\mu, \sigma^2) \qquad Z = \frac{X-\mu}{\sigma} \sim N(0,1)$$

(2) 공식

$$X \sim N(\mu, \sigma^2)$$

$$P[a \le X \le b] = P\left[\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right] \quad \text{where} \ \ Z \sim N(0,1)$$

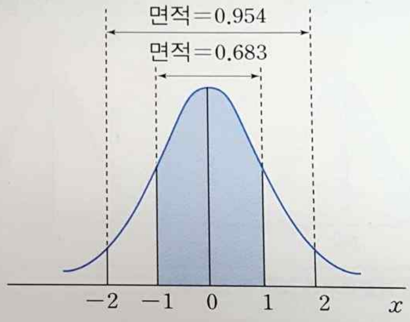


그림 5-8 | 표준정규곡선

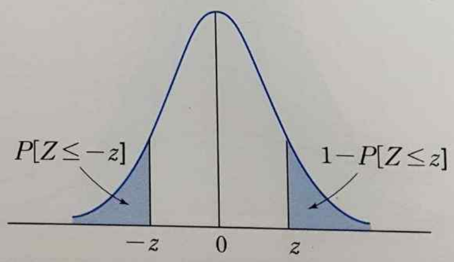


그림 5-9 | 표준정규분포

[예제]

$P[Z \le 1.37]$과 $P[Z > 1.37]$은 얼마인가?

【풀이】

$P[Z \le 1.37] = 0.9147$

$P[Z > 1.37] = 1 - P[Z \le 1.37] = 1 - 0.9147 = 0.0853$

[예제]

$P[Z > z] = 0.025$를 만족하는 z를 찾아라.

【풀이】

$P[Z < z] = 1 - 0.025 = 0.975$

표준정규분포표에서 0.975에 해당하는 주변값을 찾으면 $z = 1.96$이다.

3. 정규분포의 확률계산

> X가 $N(\mu, \sigma^2)$를 따르면 표준화된 변수
>
> $$Z = \frac{X - \mu}{\sigma}$$
>
> 는 표준정규분포 $N(0, 1)$을 따른다.

[예제]

X가 정규분포 $N(60,4^2)$을 따를 때 $P[55 \le X \le 63]$을 구하라.

【풀이】

표준화된 변수: $Z=\dfrac{X-\mu}{\sigma}=\dfrac{X-60}{4}$

$X=55$인 경우: $Z=\dfrac{X-\mu}{\sigma}=\dfrac{55-60}{4}=-1.25$

$X=63$인 경우: $Z=\dfrac{X-\mu}{\sigma}=\dfrac{63-60}{4}=0.75$

$P[55 \le X \le 63]=P[-1.25 \le Z \le 0.75]$

$=P[Z \le 0.75]-P[Z \le -1.25]=0.7734-0.1056=0.6678$

X가 $N(\mu,\sigma^2)$를 따르면 $$P[a \le X \le b]=P\left[\frac{a-\mu}{\sigma} \le Z \le \frac{b-\mu}{\sigma}\right]$$ 여기서 Z는 표준정규분포이다.

[예제]

점심식사로 곁들여 먹는 상추의 칼로리가 평균 200이고 표준편차가 5인 정규분포를 한다고 하자. 다음의 확률은 얼마인가?

(1) 208 칼로리가 넘는다.

【풀이】

X: 상추의 칼로리, $Z=\dfrac{X-\mu}{\sigma}=\dfrac{X-200}{5}$

$X=208$인 경우: $Z=\dfrac{X-\mu}{\sigma}=\dfrac{208-200}{5}=1.6$

$P[X>208]=P[Z>1.6]=1-P[Z \le 1.6]=1-0.9452=0.0548$

(2) 190과 200 칼로리 사이에 있다.

【풀이】

$X=190$인 경우: $Z=\dfrac{X-\mu}{\sigma}=\dfrac{190-200}{5}=-2.0$

$X=200$인 경우: $Z=\dfrac{X-\mu}{\sigma}=\dfrac{200-200}{5}=0.0$

$P[190 \le X \le 200]=P[-2.0 \le Z \le 0.0]=0.5-P[Z \le -2.0]=0.5-0.0228=0.4772$

4. 이항분포의 정규근사

np와 $n(1-p)$가 모두 클 때, 즉 15보다 클 때, 이항분포는 평균=np 와 표준편차=$\sqrt{np(1-p)}$를 갖는 정규분포에 의해 잘 근사된다. 즉, $$Z=\frac{X-np}{\sqrt{np(1-p)}}$$ 는 근사적으로 $N(0,1)$을 따른다.

1.3.7 정규분포와 관련된 연속확률분포

1. 카이제곱분포

(1) 카이제곱분포(χ^2-distribution)

$X_1, X_2, \cdots, X_n$: 정규모집단 $N(\mu,\sigma^2)$으로부터의 확률표본 $$\chi^2=\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{\sigma^2}=\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2/(n-1)}{\sigma^2/(n-1)}=\frac{(n-1)s^2}{\sigma^2}$$: 자유도 $n-1$인 χ^2분포   여기서, $s^2=\frac{\sum\left(X_i-\overline{X}\right)^2}{n-1}$

(2) 카이제곱분포의 확률밀도곡선의 성질

$P\left(\chi^2 \geq \chi_\alpha^2(r)\right)=\alpha$ 여기서, $r=n-1$: 자유도

① 양의 값만 가진다.

② 비대칭이고 오른쪽으로 긴 꼬리를 가진다.

③ 확률밀도곡선의 모양은 자유도에 따라 달라진다.

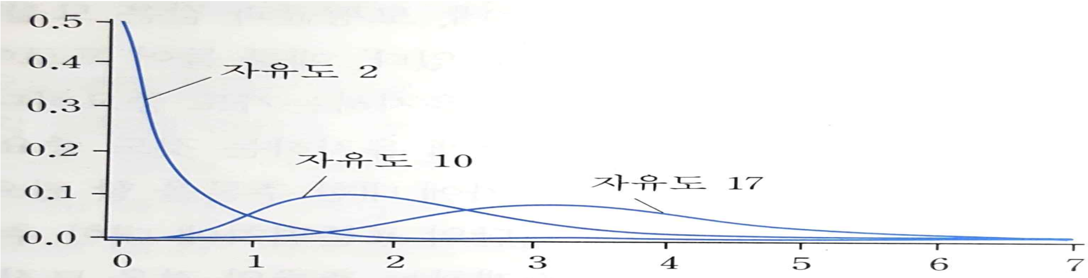


그림 11-5 자유도 2, 10, 17인 χ^2 분포

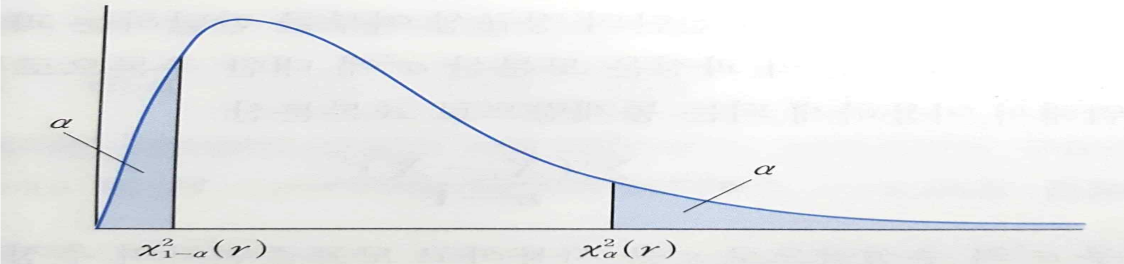


그림 11-6 상, 하위 α의 확률을 주는 값

예제 6 χ^2 분포표에서 자유도가 17인 χ^2 분포의 상, 하위 5%의 확률을 주는 값을 찾아라.

풀이▶ 아래와 같이 발췌된 χ^2 분포표로부터 상위 5%의 확률을 주는 값은 $\chi^2_{0.05}(17)=27.59$이고, 하위 5%의 확률을 주는 값은 $\chi^2_{0.95}(17)=8.67$이다.

자유도 \ α	· ·	0.95	· · ·	0.05
⋮		⋮		⋮
17	· ·	8.67	· · ·	27.59
⋮	· ·	⋮	· · ·	⋮

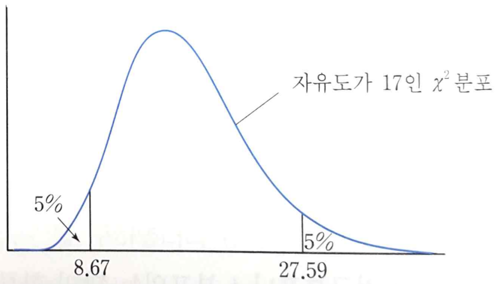


그림 11-7 상, 하위 5%의 확률을 주는 값

[예제]

자유도가 9인 카이제곱분포의 상위 5백분위수와 하위 5백분위수를 구하라.

【풀이】

$$\chi^2_{\alpha}(d.f.)=\chi^2_{0.05}(9)=16.92$$

$$\chi^2_{\alpha}(d.f.)=\chi^2_{0.95}(9)=3.33$$

2. $t-$분포

(1) t-분포 (W. S. Gosset, 1908)

❏ $X_1, X_2, \cdots, X_n$: 정규모집단 $N(\mu, \sigma^2)$에서 추출된 확률표본이고

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \ , \ s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} \text{ 이면,}$$

$$T = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$: 자유도가 $n-1$인 $t-$분포, $T \sim t(n-1)$

❏ $Z \sim N(0,1), X \sim \chi^2(n)$, Z, X: 독립

$$T = \frac{Z}{\sqrt{X/n}}$$: 자유도가 n인 $t-$분포, $T \sim t(n)$

(2) t-분포의 확률밀도곡선의 성질

❏ $t=0$을 중심으로 대칭

❏ $N(0,1)$의 확률밀도곡선에 비해 두꺼운 꼬리를 가진다.

❏ 자유도 $(n-1)$가 커짐에 따라 t-분포의 확률밀도곡선은 $N(0,1)$의 확률밀도곡선에 가까워진다.

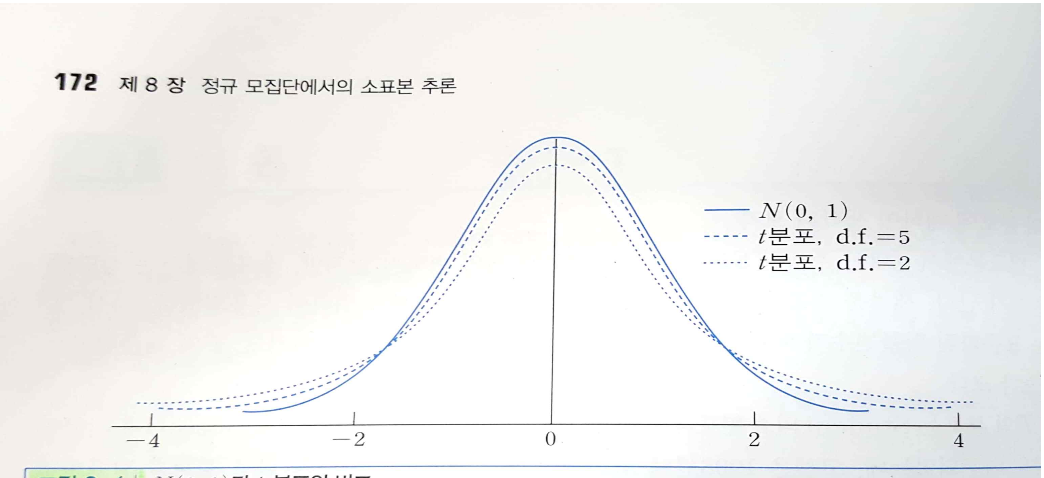


그림 8-1 $N(0, 1)$과 t-분포의 비교

[예제]

부록 [표 3]의 t-분포표를 이용하여 자유도 5인 t-분포의 상위 10%점과 하위 10%점을 구하라.

【풀이】

$t_{0.10}(5) = 1.476$, $-t_{0.10}(5) = -1.476$

[예제]

확률변수 T가 자유도 9인 t-분포를 따를 때 $P[-b<T<b]=0.90$을 만족시키는 값 b를 구하라.

【풀이】 $\alpha=0.05,\ d.f.=9\ \rightarrow\ t_{0.05}(9)=1.833\ \Rightarrow\ \therefore\ b=1.833$

3. $F-$분포

(1) 중심(central) $F-$분포

$X_1 \sim \chi^2(n_1), X_2 \sim \chi^2(n_2), \qquad X_1, X_2$: 독립

$$F=\frac{X_1/n_1}{X_2/n_2} \sim F(n_1,\ n_2)$$

(2) 비중심(non-central) $F-$분포

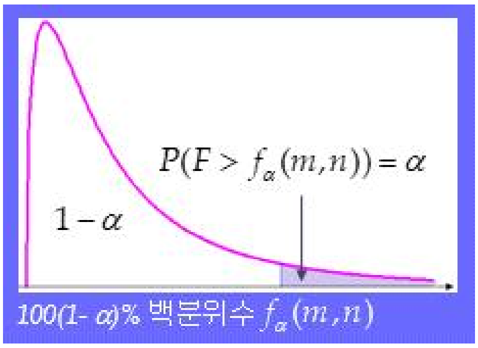
$X_1 \sim \chi^2(n_1 : \lambda)$: 비중심 카이제곱분포

$X_2 \sim \chi^2(n_2)$: 중심 카이제곱분포

X_1, X_2 : 독립

$$F=\frac{X_1/n_1}{X_2/n_2} \sim F(n_1,\ n_2)$$

: 자유도가 n_1, n_2이고 비중심성 모수가 λ인 비중심 F분포를 따른다.



1.3.8 확률벡터와 다변량정규분포

1. 확률벡터

(1) n개의 확률변수 $X_1, X_2, \cdots, X_n$이 있을 때 이들을 원소로 하는 n차원 벡터

$\underline{X}=(X_1, X_2, \cdots, X_n)^t$: n차원 확률벡터(random vector)

$E(\underline{X})=\underline{\mu}=[E(X_1), E(X_2), \cdots, E(X_n)]^t$

$Cov(\underline{X})=E[\{\underline{X}-E(\underline{X})\}\{\underline{X}-E(\underline{X})\}^t]$

$=E[\{\underline{X}-\underline{\mu}\}\{\underline{X}-\underline{\mu}\}^t]$

$=E[\underline{X}\,\underline{X}^t-\underline{\mu}\,\underline{X}^t-\underline{X}\,\underline{\mu}^t+\underline{\mu}\,\underline{\mu}^t]$

$=E(\underline{X}\underline{X}^t)-\underline{\mu}E(\underline{X}^t)-E(\underline{X})\underline{\mu}^t+\underline{\mu}\,\underline{\mu}^t$

$=E(\underline{X}\underline{X}^t)-\underline{\mu}\,\underline{\mu}^t$

$n=3$인 경우의 $Cov(\underline{X})=\begin{pmatrix} Var(X_1) & Cov(X_1, X_2) & Cov(X_1, X_3) \\ Cov(X_1, X_2) & Var(X_2) & Cov(X_2, X_3) \\ Cov(X_1, X_3) & Cov(X_2, X_3) & Var(X_3) \end{pmatrix}$

$A_{(m\times n)}$, $b_{(m\times 1)}$의 모든 원소가 상수

$E(A\underline{X}+b)=AE(\underline{X})+b$

$Cov(A\underline{X}+b)=ACov(\underline{X})A^t$

2. 다변량정규분포

(1) 확률벡터 $\underline{X}=(X_1, X_2, \cdots, X_n)^t$의

평균벡터: $E(\underline{X})=\underline{\mu}=[E(X_1), E(X_2), \cdots, E(X_n)]^t$

분산-공분산행렬 $Cov(\underline{X})=\Sigma$

확률밀도함수 $f(\underline{X}|\mu, \Sigma)=\dfrac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}}exp\left[-\dfrac{1}{2}(\underline{X}-\mu)^t\Sigma^{-1}(\underline{X}-\mu)\right]$

확률벡터 $\underline{X}=(X_1, X_2, \cdots, X_n)^t$는 평균벡터가 μ이고, 분산-공분산행렬이 Σ인 n차원 정규분포를 따른다. ⇒ $\underline{X}\sim N_n(\mu, \Sigma)$ → $X_i\sim N(\mu_i, \sigma_i^2)$

(2) 조건부 확률분포

n차원 확률벡터 $\underline{X}\sim N_n(\mu, \Sigma)$, $\underline{X_1}$: r차원 벡터 $(r<n)$

$\underline{X}=\begin{pmatrix}\underline{X_1}\\ \underline{X_2}\end{pmatrix}$, $\underline{\mu}=\begin{pmatrix}\underline{\mu_1}\\ \underline{\mu_2}\end{pmatrix}$, $\Sigma=\begin{pmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{21} & \Sigma_{22}\end{pmatrix}$

$\underline{X_2}$가 주어졌을 때 $\underline{X_1}$에 대한 조건부 확률분포

$\underline{X_1}|\underline{X_2}\sim N_r(\underline{\mu_1}+\Sigma_{12}\Sigma_{22}^{-1}(\underline{X_1}-\mu_2), \Sigma_{11}-\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})$