5장 표본의 분포

1. 함수로 주어진 확률변수들의 분포
2. 독립인 확률변수들의 합의 분포
3. 중심극한정리
4. 중심극한정리의 응용
5. 순서 통계량
6. 많이 쓰이는 몇 가지 분포의 결정

1. 함수로 주어진 확률변수들의 분포

[정의 5.1-1]

(a) $E[u(X_1, X_2)] = \sum_{x_1}\sum_{x_2} u(x_1, x_2) f(x_1) f(x_2)$를 함수 $u(X_1, X_2)$의 기댓값이라고 한다.

(b) 일반적인 경우로 확장하면 $f_1(x_1) \cdots f_n(x_n)$이 서로 독립인 확률변수 $X_1, \cdots, X_n$의 결합밀도함수일 때,

$$E[u(X_1, \cdots, X_n)] = \sum_{x_1} \cdots \sum_{x_n} u(x_1, \cdots, x_n) f(x_1) \cdots f(x_n)$$

을 $u(X_1, \cdots, X_n)$의 기댓값으로 정의한다.

[예 5.1-1] $u(X_1, X_2) = X_1 + X_2$라고 하면 위의 정의에 따라 기댓값은 다음과 같다.

$$\begin{aligned} E[u(X_1, X_2)] &= \sum_{x_1}\sum_{x_2} (x_1 + x_2) f(x_1) f(x_2) \\ &= \sum_{x_1}\sum_{x_2} x_1 f(x_1) f(x_2) + \sum_{x_1}\sum_{x_2} x_2 f(x_1) f(x_2) \\ &= E(X_1) + E(X_2). \end{aligned}$$

[정리 5.1-1] (1) X_1과 X_2가 서로 독립이면 $E(X_1 X_2) = E(X_1)E(X_2)$이다.

(2) $X_1, \cdots, X_n$이 서로 독립이고 $E[u_i(X_i)]$가 존재한다면,

$$E[u_1(X_1) \cdots u_n(X_n)] = E[u_1(X_2)] \cdots E[u_n(X_n)].$$

증명> (1)

$$E[X_1 X_2] = \sum_{x_1}\sum_{x_2} x_1 x_2 f(x_1) f(x_2) = \sum x_1 f(x_1) \sum x_2 f(x_2) = E(X_1)E(X_2).$$

(2)의 증명은 (1)의 증명을 확장

[참고] 확률변수 $X_i,\ i = 1, 2, \cdots, k$들이 iid이고 확률밀도함수 $f(x) = e^{-x},\ x > 0$라고 하자.

① 적률생성함수는 $M_X(t) = \int_0^\infty e^{tx} \cdot e^{-x} dx = \frac{1}{1-t}$, $0 < t < 1$이다.

② 따라서, $Y = X_1 + \cdots + X_k$의 적률생성함수는 $M_Y(t) = (1-t)^{-k}$,

$E[Y] = M_Y'(t)|_{t=0} = k$

$E[Y^2] = M_Y''(t)|_{t=0} = k(k+1)$

[정리 5.1-2] 확률변수 $X_1, X_2, \cdots, X_n$에서 각 확률변수 X_i의 적률생성함수가 $M_{X_i}(t)$, $i=1, 2, \cdots, n$이고 서로 독립일 때, $Y=\sum_{i=1}^{n} a_i X_i$의 적률생성함수는 $M_Y(t)=\prod_{i=1}^{n} M_{X_i}(a_i t)$이다.

증명> $M_Y(t) = E[e^{tY}] = E[e^{t(a_1 X_1 + \cdots + a_n X_n)}] = E[e^{a_1 t X_1}] \cdots E[e^{a_n t X_n}]$

$$= M_{X_i}(a_1 t) \cdots M_{X_i}(a_n t) = \prod_{i=1}^{n} M_{X_i}(a_i t)$$

[따름정리] 확률변수 $X_1, X_2, \cdots, X_n$에서 각 확률변수의 적률생성함수가 $M(t)$ 일 때, 다음이 성립한다.

(1) $Y = \sum_{i=1}^{n} X_i$의 적률생성함수는 $M_Y(t) = \prod_{i=1}^{n} M(t) = [M(t)]^n$이다.

(2) $\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$의 적률생성함수는 $M_{\overline{X}}(t) = \prod_{i=1}^{n} M\left(\frac{t}{n}\right) = \left[M\left(\frac{t}{n}\right)\right]^n$이다.

[정리 5.1-3] 확률변수 X_i, $i=1, 2, \cdots, k$들이 iid이면, X_i들의 선형결합 $Y=X_1+X_2+\cdots+X_n$의 평균과 분산은 다음과 같다.

$$E[Y] = k \cdot E[X], \qquad Var[Y] = k \cdot Var[X].$$

증명 X_i, $i=1, 2, \cdots, k$들이 iid이므로 $M_Y(t) = (M_X(t))^k$임은 분명하다.

따라서,

$$E[Y] = k \cdot M_X'(t) \cdot [M_X(t)]^{k-1} \mid_{t=0} = k \cdot M_X'(0) \cdot [M_X(0)]^{k-1} = k \cdot E[X],$$

$$E[Y^2] = k(k-1) \cdot [M_X(t)]^{k-2}[M_X'(t)]^2 \mid_{t=0} + k \cdot M_X''(t) \cdot [M_X(t)]^{k-1} \mid_{t=0}$$

$$= k(k-1) \cdot [E[X]]^2 + k \cdot E[X^2]$$

$Var[Y] = E[Y^2] - [E[Y]]^2 = k \cdot Var[X]$이다.

[예 5.1-2] [예 5.1-1]에서와 같이 $Y = X_1 + X_2$라 하고 다음을 구해본다.

(1) $E[Y] = E[X_1 + X_2] = E[X_1] + E[X_2]$.

(2) $$\begin{aligned} Var[Y] &= E\left[(Y-\mu_Y)^2\right] = E\left[(X_1 + X_2 - \mu_1 - \mu_2)^2\right] \\ &= E\left[(X_1-\mu_Y)^2 + 2(X_1-\mu_1)(X_2-\mu_2) + (X_2-\mu_2)^2\right] \\ &= E[(X_1-\mu_Y)^2] + 2E[(X_1-\mu_1)(X_2-\mu_2)] + E[(X_2-\mu_2)^2]. \end{aligned}$$

X_1과 X_2가 서로 독립이라면 $E[(X_1-\mu_1)(X_2-\mu_2)] = E[(X_1-\mu_1)]E[(X_2-\mu_2)] = 0$이므로, $Var(X_1) = \sigma_1^2$, $Var(X_2) = \sigma_2^2$으로 두면 $Var(Y) = Var(X_1 + X_2) = Var(X_1) + Var(X_2) = \sigma_1^2 + \sigma_2^2$이 된다.

(3) $M_Y(t) = E\left[e^{tY}\right] = E\left[e^{t(X_1+X_2)}\right] = E\left[e^{tX_1}e^{tX_2}\right] = E\left[e^{tX_1}\right]E\left[e^{tX_2}\right]$이다. X_1과 X_2가 서로 독립이고 정규분포 $N(\mu, \sigma^2)$를 따른다면,

$$M_Y(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}} e^{\mu t + \frac{\sigma^2 t^2}{2}} = e^{2\mu t + \sigma^2 t^2}$$

이다. 이는 정규분포 $N(2\mu, 2\sigma^2)$의 적률생성함수이므로 Y의 분포가 $N(2\mu, 2\sigma^2)$임을 보여준다.

[예 5.1-3] 두 확률변수 X_1, X_2가 서로 독립이고 각각의 평균과 분산이 μ_1, σ_1^2, μ_2, σ_2^2일 때, 확률변수 $Y = X_1 X_2$에 대하여 다음을 구하라.

(1) $E[Y] = E[X_1 X_2] = E[X_1]E[X_2] = \mu_1 \mu_2$,

(2) $$\begin{aligned} Var[Y] &= E\left(X_1^2 X_2^2\right) - \mu_1^2\mu_2^2 = E(X_1^2)E(X_2^2) - \mu_1^2\mu_2^2 \\ &= (\sigma_1^2 + \mu_1^2)(\sigma_2^2 + \mu_2^2) - \mu_1^2\mu_2^2 \\ &= \sigma_1^2\sigma_2^2 + \mu_1^2\sigma_2^2 + \mu_2^2\sigma_1^2 \end{aligned}$$

이다. 위의 두 번째 등식에서 $E\left(X_1^2\right) = \sigma_2^2 + \mu_1^2$, $E\left(X_2^2\right) = \sigma_2^2 + \mu_2^2$가 사용되었다.

2. 독립인 확률변수들의 합의 분포

[정리 5.2-1] 확률변수 $X_1, X_2, \cdots, X_n$이 서로 독립이며 각각의 평균이 $\mu_1, \mu_2, \cdots, \mu_n$이고 분산이 $\sigma_1^2, \sigma_2^2, \cdots, \sigma_n^2$이라고 하자. $a_1, a_2, \cdots, a_n$이 상수일 때, 확률변수 $X_1, X_2, \cdots, X_n$의 선형 결합 $Y = \sum_{i=1}^{n} a_i X_i$의 평균과 분산은 다음과 같다.

$$E[Y] = \sum_{i=1}^{n} a_i \mu_i, \quad Var[Y] = \sum_{i=1}^{n} a_i^2 \sigma_i^2.$$

증명
$$E[Y] = E\left[\sum_{i=1}^{n} a_i X_i\right] = \sum_{i=1}^{n} a_i E(X_i) = \sum_{i=1}^{n} a_i \mu_i,$$

$$\begin{aligned} Var[Y] &= E[(Y-\mu_y)^2] = E\left[\left(\sum_{i=1}^{n} a_i X_i - \sum_{i=1}^{n} a_i \mu_i\right)^2\right] \\ &= E\left[\left\{\sum_{i=1}^{n} a_i (X_i - \mu_i)\right\}^2\right] \\ &= E\left[\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j (X_i - \mu_i)(X_j - \mu_j)\right] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j E[(X_i - \mu_i)(X_j - \mu_j)] \\ &= \sum_{i=1}^{n} a_i^2 E[(X_i - \mu_i)^2] = \sum_{i=1}^{n} a_i^2 \sigma_i^2. \end{aligned}$$

[따름정리] 각 확률변수 X_i, $i = 1, 2, \cdots, n$이 평균이 μ, 분산이 σ^2일 때, $\overline{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$의 평균과 분산은 다음과 같다.

$$\mu_{\overline{X}} = \sum_{i=1}^{n} \frac{1}{n} \mu = \mu, \quad \sigma_{\overline{X}}^2 = \sum_{i=1}^{n} \left(\frac{1}{n}\right)^2 \sigma^2 = \frac{\sigma^2}{n}.$$

[정리 5.2-1] $X_1, X_2, \cdots, X_n$이 정규분포 $N(\mu, \sigma^2)$을 따르는 확률표본일 때, 표본 평균 $\overline{X}$의 분포 $N(\mu, \sigma^2/n)$이다.

증명
$$\begin{aligned} M_{\overline{X}}(t) &= \left[M\left(\frac{t}{n}\right)\right]^n \\ &= \left[\exp\left\{\mu\left(\frac{t}{n}\right) + \sigma^2\left(\frac{t}{n}\right)^2/2\right\}\right]^n \\ &= \exp\left[\mu t + \frac{\sigma^2}{n} \cdot \frac{t^2}{2}\right]. \end{aligned}$$

이는 정규분포 $N(\mu, \sigma^2/n)$의 적률생성함수이므로 위의 정리가 성립한다.

[예 5.2-1] 확률변수 $X_1, X_2, \cdots, X_n$이 각각이 서로 독립인 베르누이 시행을 나타내는 확률변수라면, $Y=\sum_{i=1}^{n} X_i$의 적률생성함수는

$$M_Y(t)=\prod_{i=1}^{n}(1-p+pe^t)=(1-p+pe^t)^n$$

이며, 이는 이항분포 $B(n, p)$의 적률생성함수와 같음을 알 수 있다.

[예 5.2-2] 확률변수 $X_1, X_2, \cdots, X_n$이 서로 독립이고, 자유도가 $r_1, r_2, \cdots, r_n$인 카이제곱분포를 따를 때, $Y=\sum_{i=1}^{n} X_i$의 분포를 구해보자.

풀이 카이제곱분포의 적률생성함수는 $M_{X_i}(t)=(1-2t)^{-\frac{r_i}{2}}$, $t<\frac{1}{2}$ 이므로,

$$\begin{aligned} M_Y(t) &= M_{X_1}(t)M_{X_2}(t)\cdot \cdots \cdot M_{X_n}(t) \\ &= (1-2t)^{-\frac{r_1}{2}}(1-2t)^{-\frac{r_2}{2}}\cdot \cdots \cdot (1-2t)^{-\frac{r_n}{2}} \\ &= (1-2t)^{-\frac{1}{2}(r_1+\cdots+r_n)} \end{aligned}$$

이다. 이는 자유도가 $\sum_{i=1}^{n} r_i$인 카이제곱분포의 적률생성함수이다.

3. 중심극한정리

[정리 5.3-1] 중심극한정리 (Lindeberg and Levy)

확률변수 $X_1, X_2, \cdots$가 평균이 $E(X_n)=\mu$이고 분산이 $Var(X_n)=\sigma^2 \neq 0$이며 iid인 확률변수들이라고 할 때, $W_n = \dfrac{\overline{X}_n-\mu}{\sigma/\sqrt{n}} = \dfrac{\sum_{i=1}^n X_i - n\mu}{\sigma\sqrt{n}}$의 분포는 n의 값이 커짐에 따라 표준정규분포를 하게 된다.

증명 $Z_i = \dfrac{X_i-\mu}{\sigma}$, $i=1,2,\cdots,n$의 적률생성함수를 $m(t)=E\left[\exp\left\{t\left(\dfrac{X_i-\mu}{\sigma}\right)\right\}\right]$.

$-h < t < h$라고 두면 W_n의 적률생성함수는

$$\begin{aligned} E[\exp(tW_n)] &= E\left[\exp\left\{\left(\frac{t}{\sqrt{n}\sigma}\right)\left(\sum_{i=1}^n X_i - n\mu\right)\right\}\right] \\ &= E\left[\exp\left\{\frac{t}{\sqrt{n}}\cdot\frac{X_1-\mu}{\sigma}\right\}\cdot\cdots\cdot\exp\left\{\frac{t}{\sqrt{n}}\cdot\frac{X_n-\mu}{\sigma}\right\}\right] \\ &= E\left[\exp\left\{\left(\frac{t}{\sqrt{n}}\right)\left(\frac{X_1-\mu}{\sigma}\right)\right\}\right]\cdot\cdots\cdot E\left[\exp\left\{\left(\frac{t}{\sqrt{n}}\right)\left(\frac{X_n-\mu}{\sigma}\right)\right\}\right] \\ &= \left[m\left(\frac{t}{\sqrt{n}}\right)\right]^n \end{aligned}$$

$E(Z_i)=0$, $E(Z_i^2)=1$이므로 $m(0)=1$, $m'(0)=E\left(\dfrac{X_i-\mu}{\sigma}\right)=0$,

$$m''(0) = E\left[\left(\frac{X_i-\mu}{\sigma}\right)^2\right] = 1$$

이제 $m(t)$를 2차까지 테일러 급수전개를 하면

$$\begin{aligned} m(t) &= m(0) + m'(0)t + \frac{m''(t_1)}{2!}t^2, \quad 0 < t_1 < t \\ &= 1 + \frac{t^2}{2} + \frac{[m''(t_1)-1]}{2}t^2 \end{aligned}$$

이 되므로 W_n의 적률생성함수는 다음과 같이 표현할 수 있다.

$$\begin{aligned} E[\exp(tW_n)] &= \left[1 + \frac{1}{2}\left(\frac{t}{\sqrt{n}}\right)^2 + \frac{1}{2}\{m''(t_1)-1\}\left(\frac{t}{\sqrt{n}}\right)^2\right]^n \\ &= \left[1 + \frac{t^2}{2n} + \frac{\{m''(t_1)-1\}}{2n}t^2\right]^n, \quad -\sqrt{n}h < t < \sqrt{n}h. \end{aligned}$$

$0 < t_1 < \dfrac{t}{\sqrt{n}}$이므로 $n\to\infty$일 때, $t_1 = 0$이고 $m''(t)$는 $t=0$에서 연속이므로 $\lim_{n\to\infty}[m''(t_1)-1] = 1-1 = 0$이다. 이를 이용하면

$$\begin{aligned} \lim_{n\to\infty} E[\exp(tW_n)] &= \lim_{n\to\infty}\left[1 + \frac{t^2}{2n} + \frac{\{m''(t_1)-1\}}{2n}t^2\right]^n \\ &= \lim_{n\to\infty}\left\{1 + \frac{t^2/2}{n}\right\}^n = e^{t^2/2} \end{aligned}$$

이다. 이는 W_n의 적률생성함수가 표준정규분포 $N(0,1)$의 적률생성함수와 같음을 보여준다. 즉, W_n의 분포는 표준정규분포를 한다.

[참고]

(1) Liapounov의 중심극한정리

$X_1, X_2, \cdots$이 독립인 확률변수들의 수열이라고 하자. $E(\overline{X_n} - \mu_n)^2 = \sigma_n^2 \neq 0$, $E(X_n) = \mu_n$이고, 모든 n에 대하여 $E|X_n - \mu_n|^3 = \beta_n$이 존재하며 $B_n = \left(\sum_{i=1}^{n} \beta_i \right)^{\frac{1}{3}}$, $C_n = \left(\sum_{i=1}^{n} \sigma_i^2 \right)^{\frac{1}{2}}$ 으로 둘 때, $\lim_{n \to \infty} \left(\frac{C_n}{B_n} \right) = 0$이 성립하면 $Y_n = \frac{\sum_{i=1}^{n} (X_i - \mu_i)}{C_n}$ 의 극한분포는 표준정규분포를 따른다.

(2) Lindeberg - Feller의 중심극한정리

$X_1, X_2, \cdots$ 이 서로 독립인 확률변수들의 수열이라고 하고, G_n을 X_n의 분포함수라고 하자. $E(X_n) = \mu_n$, $Var(X_n) = \sigma_n^2 \neq 0$이고, $Y_n = \frac{\sum_{i=1}^{n} (X_i - \mu_i)}{C_n}$, $C_n = \left(\sum_{i=1}^{n} \sigma_i^2 \right)^{\frac{1}{2}}$ 으로 둘 때, 다음의 조건 $\lim_{n \to \infty} \frac{1}{C_n^2} \sum_{i=1}^{n} \int_{|X - \mu_i| > \varepsilon C_n} (x - \mu_i)^2 dG_i(x) = 0$, (이 조건을 Lindeberg의 조건이라고 함) $\lim_{n \to \infty} \max_{1 \le i \le n} \frac{\sigma_i}{C_n} = 0$이 성립하면 Y_n의 극한분포는 표준정규분포를 따른다.

예 5.3-1 $\overline{X} = \overline{X}_{15}$를 밀도함수가 $f(x) = \frac{3}{2}x^2,\ -1 < x < 1$인 분포로부터 표본의 크기가 $n = 15$인 확률표본의 평균이라고 하자. $\mu = 0,\ \sigma^2 = \frac{3}{5}$은 쉽게 얻을 수 있으므로 다음의 확률을 구해본다.

$$P[0.03 \le \overline{X} \le 0.15] = P\left[\frac{0.03-0}{\sqrt{3/5}/\sqrt{15}} \le \frac{\overline{X}-0}{\sqrt{3/5}/\sqrt{15}} \le \frac{0.15-0}{\sqrt{3/5}/\sqrt{15}}\right]$$

$$\approx \Phi(0.75) - \Phi(0.15)$$

$$= 0.2138.$$

예 5.3-2 확률표본 $X_1, X_2, \cdots, X_n$이 표준균일분포 $U(0, 1)$에서 추출되었다면 $Y = X_1 + X_2 + \cdots + X_n$의 분포를 구해보도록 한다. Y의 평균은 $E(Y) = \frac{n}{2}$이고 분산은 $Var(Y) = \frac{n}{12}$이므로 중심극한정리에 의하여 Y는 정규분포 $N\left(\frac{n}{2}, \frac{n}{12}\right)$을 따르게 된다. 표현을 바꾸면 $Y = \sum_{i=1}^{12} X_i - 6$은 표준정규분포를 따르게 된다. 부연해서 설명을 하면 표준균일분포에서 추출한 확률표본 12개를 더한 값에서 6을 빼면 표준정규분포를 따르는 난수(Random number)가 된다.

4. 중심극한정리의 응용

확률표본 $X_1, X_2, \cdots, X_n$이 베르누이 시행 $B(1, p)$, $0<p<1$에서 추출되었다고 하면 $Y_n = \sum_{i=1}^{n} X_i$는 이항분포 $B(n, p)$를 따르게 됨을 이미 앞에서 설명하였다. 중심극한정리에 의하여 Y_n의 표준화된 변수 $W_n = \frac{Y_n - np}{\sqrt{np(1-p)}} = \frac{\overline{X}_n - p}{\sqrt{p(1-p)/n}}$은 n의 값이 커짐에 따라 표준정규분포 $N(0, 1)$을 따르게 된다. 특히 $np \geq 5$이거나 $n(1-p) \geq 5$인 경우에는 이항분포 $B(n, p)$는 정규분포 $N(np, np(1-p))$에 근사적으로 가깝게 된다.

예 5.4-1 Y_{36}을 이항분포 $B(36, 1/2)$에서의 확률표본이라고 하면 다음의 확률을 계산할 수 있다.

$$
\begin{aligned}
P[Y_{36} = 20] &= P[19.5 \leq Y_{36} \leq 20.5] \\
&= P\left[\frac{19.5-18}{\sqrt{9}} \leq \frac{Y_{36}-18}{\sqrt{9}} \leq \frac{20.5-18}{\sqrt{9}}\right] \\
&\approx \Phi(0.8330) - \Phi(0.5) \\
&= 0.1060.
\end{aligned}
$$

예 5.4-2 (포아송 분포의 정규분포로의 근사)

포아송 분포의 밀도함수는 $f(x) = \frac{\lambda^x e^{-\lambda}}{x!}$이며 평균은 $E(X) = \lambda$이고 분산은 $Var(X) = \lambda$ 임은 앞에서 이미 구하였다. 중심극한정리에 의하여 표준화된 변수 $W = \frac{Y-\lambda}{\sqrt{\lambda}}$는 표준정규분포 $N(0, 1)$을 따르게 된다.

예 5.4-3 (이항 분포의 포아송 분포로의 근사)

이번에는 적률생성함수를 이용하여 이항 분포가 포아송 분포로 근사하게 됨을 알아보도록 한다. $np = \lambda$로 고정시키고 $p \to 0$일 때,

$$\begin{aligned} M(t) &= (1-p+pe^t)^n, \qquad p = \frac{\lambda}{n} \\ &= \left(1 - \frac{\lambda}{n} + \frac{\lambda}{n} e^t\right)^n \\ &= \left[1 + \frac{\lambda(e^t - 1)}{n}\right]^n \\ &= e^{\lambda(e^t - 1)} \end{aligned}$$

이다. 이는 모수가 λ인 포아송 분포의 적률생성함수이다.

예 5.4-4 (카이제곱분포의 정규분포로의 근사)

확률변수 X_n이 자유도가 n인 카이제곱분포를 따른다면 X_n의 평균은 $E(X_n) = n$이고 분산은 $Var(X_n) = 2n$이므로 이를 표준화한 변수 $\frac{X_n - n}{\sqrt{2n}}$은 표준정규분포 $N(0, 1)$을 따르게 된다.