제 2 장 확률분포

§2.1 복원추출 관련 분포

2.1.1 *Bernoulli* 분포

<사례 1.3>의 모분포가 바로 *Bernoulli* 분포인데, 관행상

$$Y = \begin{cases} 1, & \text{if 성공} \\ 0, & \text{if 실패} \end{cases} \tag{2.1.1}$$

로 표현한다 (식 (1.6.1) 참조). 그리고, $P(Y=1)$을 소문자 p로, $P(Y=0)$을 소문자 $q(=1-p)$로 표기한다. <사례 1.3>에서는 $p=m/N$이고 $q=1-(m/N)$이다.

2.1.2 이항(*binomial*) 분포

<사례 1.3>에서 복원추출의 경우 $S=\sum_{i=1}^{n} Y_i$의 분포를 이항분포라 한다. 즉, S는 n

회의 독립(이고 동일한)시행 중에서 성공하는 횟수를 의미한다. 요즈음은 중학교 과정에서도 등장하는 이항분포는 다음과 같다.

$$P(S=s)=\binom{n}{p} p^{s} q^{n-s}\ ,\ s=0,1,\cdots,n \tag{2.1.2}$$

2.1.3 기하(*geometric*) 분포

첫 번째 성공이 발생할 때까지 시행하는 독립시행의 횟수를 Y라 하면, 기하분포인 Y의 분포는 다음과 같다.

$$P(Y=y)=q^{y-1}p\ ,\ y=1,2,\cdots \tag{2.1.3}$$

2.1.4 음이항(*negative binomial*) 분포

$Y_1,\cdots,Y_n$이 기하분포를 따르는 *iid* 확률변수일 때, $S=\sum_{i=1}^{n} Y_i$의 분포를 음이항(또는 *Pascal*) 분포라 한다. 즉, S는 n번째 성공이 발생할 때까지 시행하는 독립시행의 횟수를 의미한다. S의 분포는 다음과 같이 이항분포를 이용해서 구할 수 있다.

$$\begin{aligned} P(S=s) &= P(s\text{번째 시행에서 } n\text{번째 성공이 발생}) \\ &= P(s-1\text{번의 시행 중에서 } n-1\text{번 성공, } s\text{번째 시행은 성공}) \end{aligned}$$

인데, s번째 시행의 결과는 이전의 ($s-1$번의) 시행결과와 독립이므로

$$\begin{aligned} P(S=s) &= P(s-1\text{번의 시행 중에서 } n-1\text{번 성공}) \cdot P(s\text{번째 시행은 성공}) \\ &= \binom{s-1}{n-1} p^{n-1} q^{(s-1)-(n-1)} \cdot p \\ &= \binom{s-1}{n-1} p^{n} q^{s-n}\ ,\ s=n,n+1,\cdots \end{aligned} \tag{2.1.4}$$

2.1.5 *Uniform* 분포

Uniform 분포는 독립시행과 다음과 같은 관계가 있다. n회의 독립시행 중에서 단 한번 성공이 발생했다고 하는 정보가 있을 때, 성공이 발생한 시행이 i번째의

($i=1,2,\cdots,n$) 시행일 확률은 $1/n$로 모두 동일하다. 이는 대칭성에 의한 결과인데, n번의 시행이 모두 동일한 여건 (동일한 성공확률) 하에서 시행되었기 때문에 특별히 어느 시행이 다른 시행과 달라야 될 이유가 없는 것이다.

이를 정식으로 유도하되 다음과 같이 확장한다. n번의 시행 중에서 성공하는 횟수를 B_n이라 하고, i번째의 성공이 발생할 때까지 시행하는 횟수를 N_i라 하면, B_n과 N_i는 각각 이항분포와 음이항분포를 따른다. 그러면, $B_n=j$라는 조건 하에서 $N_i=m$ 일 ($1\le i\le j,\ m\le n$) 확률인 $P(N_i=m \mid B_n=j)$ 는 조건부 확률의 정의에 의해서 $P(N_i=m,\ B_n=j)/P(B_n=j)$ 인데, 분모는 이항분포인 $\binom{n}{j}p^j q^{n-j}$이고 분자는 다음과 같다.

$$
\begin{aligned}
&P(m-1\text{번 중에 } i-1\text{번 성공, } m\text{번째는 성공, 이후 } n-m\text{번 중에 } j-i\text{번 성공})\\
&=\binom{m-1}{i-1}p^{i-1}q^{m-i}\cdot p\cdot\binom{n-m}{j-i}p^{j-i}q^{(n-m)-(j-i)}\\
&=\binom{m-1}{i-1}\binom{n-m}{j-i}p^j q^{n-j}
\end{aligned}
$$

따라서 다음과 같이 “성공확률 p와 무관한” 결과를 얻는다.

$$P(N_i=m \mid B_n=j)=\binom{m-1}{i-1}\binom{n-m}{j-i}\Big/\binom{n}{j} \qquad (2.1.5)$$

식 (2.1.5)에 $i=j=1$을 대입하면 *uniform* 분포인 $1/n$을 얻는다. 또한, 예를 들어, $i=j=2$를 대입하면 $2(m-1)/\{n(n-1)\}$을 얻는데 ($2\le m\le n$), 이는 $m-1$에 비례해서 증가하는 삼각형 형태의 분포이다.

식 (2.1.5)에 대한 해석은 다음과 같다. 분모는 총 n번의 시행에서 (성공이 발생하는) j번의 시행을 뽑는 경우의 수이다. 반면에, 분자는 $\binom{m-1}{i-1}\binom{1}{1}\binom{n-m}{j-i}$ 인데, 이는 n번에서 j번을 뽑되 $m-1$번째 이하에서 $i-1$번, m번째에서 한번, 그리고 $m+1$번째 이후에서 $j-i$번을 뽑는 경우의 수이다.

2.1.6 다항(*multinomial*) 분포

<사례 1.3>의 모집단에서 “복원추출”하는 경우 꼬리표가 있는 동물의 수는 이항분포를 따르듯이, <사례 1.1>의 모집단에서 “복원추출”하는 경우 후보별 득표수는 다항분포를

따른다. 후보 i의 실제 득표율을 p_i라 하자. 즉, $p_i = 0.42, \cdots, p_5 = 0.015$이다. 그리고, 투표함에서 임의로 복원추출한 n장의 투표지 중에서 후보 i가 득표한 수를 S_i라 하자. 그러면

$$P(S_1 = s_1, \cdots, S_5 = s_5) = \frac{n!}{s_1!\, s_2! \cdots s_5!} p_1^{s_1} p_2^{s_2} \cdots p_5^{s_5} \qquad (2.1.6)$$

인데, 물론 $\sum s_i = n$이고 $\sum p_i = 1$이다.

이항분포는 다항분포의 특수한 경우로써, $p_1 = p$, $p_2 = 1 - p = q$, $S_1 = S$, $S_2 = n - S$인 경우이다. 또한, 식 (2.1.6)의 $n!/(s_1! \cdots s_5!)$을 조합(combination)으로 표현하면 다음과 같다.

$$\binom{n}{s_1}\binom{n-s_1}{s_2}\binom{n-s_1-s_2}{s_3}\binom{s_4+s_5}{s_4}\binom{s_5}{s_5}$$

복원추출 또는 독립시행에 관해서 한가지 주의할 점이 있다. 독립시행과 관련된 모든 것이 독립인 것은 아니다. 예를 들어, 식 (2.1.6)에서 $S_1, \cdots, S_5$는 독립이 아닌데, 이는 어느 후보가 많이 득표하면 다른 후보는 적게 득표하기 때문이다 (비고 : $\sum_{i=1}^{5} S_i = n$). 또한, S_n을 n번째 성공이 발생할 때까지 시행하는 독립시행의 횟수라 하면 $S_1, S_2, S_3, \cdots$는 각각 음이항 분포를 따르는데, 이때 $S_1, S_2, S_3, \cdots$는 독립이 아니다 (비고 : $S_1 < S_2 < S_3 \cdots$). 이때 독립인 것은 $S_1, S_2 - S_1, S_3 - S_2, \cdots$인데, 이들은 기하분포를 따르는 iid 확률변수이다.

§2.2 비복원추출 관련 분포

2.2.1 초기하(*hypergeometric*) 분포

<사례 1.3>에서 □비복원추출□의 경우 $S=\sum_{i=1}^{n} Y_i$의 분포를 초기하분포라 한다. 식 (1.6.2)를 일반적으로 표현하면

$$P(S=s)=\binom{m}{s}\binom{N-m}{n-s}\Big/\binom{N}{n} \tag{2.2.1}$$

인데, $s=1, m=4, n=3$ 을 대입하면 식 (1.6.2)를 얻는다. 단, $0\le s\le m$이고 $0\le n-s\le N-m$이므로, 이로부터 다음을 얻는다.

$$\max[0, n-(N-m)]\le s\le \min(n, m)$$

이항분포와 초기하분포의 차이점은 물론 복원추출과 비복원추출이다. 그리고, 공통점은 표본추출계획에서 n을 상수(또는 제어 모수)로 취급함에 따라서, 성공횟수인 $S=\sum_{i=1}^{n} Y_i$는 확률변수가 된다는 점인데, 이를 Tpye Ⅰ계획이라 하자. 반면에, 음이항분포에서와 같이 성공횟수를 상수로 취급하면 총 시행횟수가 확률변수가 되는데 이를 Type Ⅱ 계획이라 하자. 그러니까 이항분포와 음이항분포의 공통점은 복원추출이고 차이점은 Type Ⅰ, Type Ⅱ 계획이다.

2.2.2 음초기하(*negative hypergeometric*)분포

Type Ⅱ 계획하에 비복원추출하면 총 시행횟수는 음초기하분포를 따른다. 따라서, 초기하분포와 음초기하분포의 공통점은 비복원추출이고 차이점은 TypeⅠ, Type Ⅱ 계획이다.

음초기하분포는 이미 식 (2.1.5)로 등장했다. 자루 속에 탁구공이 n개 들어 있는데, j개는 흰색이고 $n-j$개는 노란색이라고 하자. 탁구공을 임의로 하나씩 자루에서 꺼내되, 꺼낸 공은 자루에 다시 넣지 않는다. 이때, 흰공을 i개 꺼낼때까지 총 m개를 꺼내야 될 확률이 식 (2.1.5)이다(문헌[6] 참조).

2.2.3 다변량(*multivariate*) 초기하분포

복원추출의 경우에 이항분포를 다항분포로 확장하듯이 비복원추출의 경우에는 초기하분포를 다변량 초기하분포로 확장한다. <사례 1.1>의 모집단에서 비복원추출하는 경우 후보별 득표수는 다변량 초기하분포를 따른다. 후보 i의 실제 득표수를 N_i라 하자. 그리고, 투표함에서 임의로 비복원추출한 n장의 투표지 중에서 후보 i가 득표한 수를 S_i라 하자. 그러면

$$P(S_1=s_1,\cdots,S_5=s_5)=\binom{N_1}{s_1}\binom{N_2}{s_2}\cdots\binom{N_5}{s_5}\Big/\binom{N}{n} \qquad (2.2.2)$$

인데 (N은 모집단의 크기), 물론 $0\le s_i\le N_i$이고 $\sum s_i=n$, $\sum N_i=N$이다.

식 (2.2.2)에서, $\binom{N}{n}$은 N개에서 n개를 뽑는 경우의 수이고, $\binom{N_1}{s_1}\cdots\binom{N_5}{s_5}$는 N개에서 n개를 뽑되 $N_1,\cdots,N_5$개에서 각각 $s_1,\cdots s_5$개씩 뽑는 경우의 수이다 (문헌[1] 참조).

다변량 분포의 의미는 다음과 같다. 첫째, 식 (2.1.6)과 (2.2.2)는 모두 $S_1,\cdots S_5$의 결합분포이다. 둘째로, 식 (2.1.6)에서는 S_i의 (marginal : 주변) 분포가 이항분포이고 식 (2.2.2)에서는 S_i의 분포가 초기하분포이다. (비고 : 다항분포를 다변량 이항분포라 부를 수도 있음.) 즉, 식 (2.1.6)으로부터는 $P(S_i=s_i)=\binom{n}{s_i}p_i^{s_i}(1-p_i)^{n-s_i}$을 얻고, 식 (2.2.2)로부터는 $P(S_i=s_i)=\binom{N_i}{s_i}\binom{N-N_i}{n-s_i}\Big/\binom{N}{n}$을 얻을 수 있다($i=1,\cdots,5$). (비고 :

조건부 분포 $P(S_i = s_i \mid S_j = s_j)$ 역시 각각 이항분포와 초기하분포를 따름.)

§2.3 포아송($Poisson$) 분포

이항분포에서 시행횟수 n은 점점 증가시키고 성공확률 p는 점점 감소시키되, 곱 np를 일정한 값으로 유지시키면 포아송 분포를 얻는다. 편의상, dt 시간마다 한번씩 시행하여 t시간동안 총 $n=t/dt$회 독립시행을 한다고 하자. 그리고, $p=\lambda dt=\lambda t/n$이라 하자. (비고 : $np=\lambda t$.) 그러면, t시간 동안 (또는, n회 시행 중에) 발생하는 성공 횟수인 $S(t)$의 분포는 다음과 같다 (문헌[9] 참조).

$$
\begin{aligned}
P[S(t)=s] &= \lim_{n\to\infty}\binom{n}{s} p^s q^{n-s} = \lim_{n\to\infty}\binom{n}{s}\left(\frac{\lambda t}{n}\right)^s\left(1-\frac{\lambda t}{n}\right)^{n-s} \\
&= (\lambda t)^s e^{-\lambda t}/s!\ ,\ s=0,1,2,\cdots
\end{aligned}
\tag{2.3.1}
$$

포아송 분포는 연속(continuous) 시간축 상에서 임의로 (또는, random하게) 발생하는 이산(discrete) 사건을 묘사할 때 활용된다. 예를 들면, 안전사고의 발생, 전자부품의 고장, 고객의 도착 등이 있다. 그러나, t는 반드시 시간일 필요는 없으며, 심지어 1차원일 필요도 없다. 예를 들어, $S(t)$는 면적이 t인 지역에 서식하는 어떤 식물의 수일 수도 있고, 체적이 t인 유리잔에 있는 기포의 수일 수도 있다.

포아송 분포의 또 다른 역할은 이산분포를 연속분포로 연결해 주는 징검다리의 역할이다.

§2.4 포아송 과정 관련 연속분포

2.4.1 포아송 과정

dt 시간마다 성공이 한번 발생할 수도 있고 아니할 수도 있는데, 발생할 확률은 λdt 이고 아니할 확률은 $1-\lambda dt$라 하자. 그리고, 어느 dt 구간에서의 성공 발생 여부는 다른 dt 구간에서의 발생 여부와 무관하다 (즉, 독립이다).
이런 식으로 연속 시간축 상의 이산 시점들에서 성공이 발생하는 확률과정(stochastic process)을 포아송 과정(Poisson process)이라 한다. 물론 t시간 동안 발생하는 성공 횟수인 $S(t)$는 포아송 분포를 따른다.

2.4.2 지수(*exponential*) 분포

기하분포는 첫 성공이 발생할 때까지의 독립시행 횟수의 분포인 반면에, 지수분포는 첫 성공이 발생할 때까지 걸리는 시간 T의 분포이다. 처음으로 등장하는 연속분포인 지수분포는 다음과 같이 포아송 분포를 이용해서 구할 수 있다.
$P(t<T<t+dt)=P[t$시간 동안 0번 성공, $(t,t+dt)$에서는 성공$]$인데, $(t,t+dt)$에서의 성공 발생여부는 이전의 t시간 동안의 결과와 독립이므로

$$\begin{aligned} P(t<T<t+dt) &= P(t\text{시간 동안 0번 성공}) \cdot P[(t,t+dt)\text{에서 성공}] \\ &= P[S(t)=0] \cdot \lambda dt \\ &= \lambda e^{-\lambda t} dt,\ t \ge 0 \end{aligned} \tag{2.4.1}$$

를 얻는다. 그리고, 식 (2.4.1)의 좌변을 dt로 나눈 것을 T의 (확률)밀도함수(density function)라 하고 $f_T(t)$로 표기하는데, 관행상 다음과 같이 모든 t에 대해서 정의한다.

$$f_T(t)=\begin{cases} \lambda e^{-\lambda t}, & \text{if } t\ge 0 \\ 0, & \text{if } t<0 \end{cases} \tag{2.4.2}$$

2.4.3 *Erlang* 분포

음이항분포는 n번째 성공이 발생될 때까지의 독립시행 횟수의 분포인 반면에,

Erlang 분포는 n번째 성공이 발생할 때까지 걸리는 시간 $Y=\sum_{i=1}^{n} T_i$의 분포이다. (비고 : $T_1,\cdots,T_n$은 *iid* 지수 확률변수임.) Y의 밀도함수를 구하는 방법은 (T의 밀도함수를 구할 때와 동일한데) 다음과 같다.

$$P(t<Y<t+dt)=P[\,t\text{시간 동안 } n-1\text{번 성공}, (t,t+dt)\text{에서 성공}\,]$$

$$=P[\,S(t)=n-1\,]\cdot\lambda dt=\frac{(\lambda t)^{n-1}e^{-\lambda t}}{(n-1)!}\cdot\lambda dt$$

$$f_Y(t)=\begin{cases}\dfrac{(\lambda t)^{n-1}e^{-\lambda t}}{(n-1)!}\cdot\lambda\ , & \text{if } t\ge 0\\ 0 & ,\ \text{if } t<0\end{cases} \qquad (2.4.3)$$

2.4.4 감마(*gamma*) 분포

Erlang 분포에서 n은 자연수인데, 이를 양의 실수로 간주하면 감마분포가 된다. 단, 식 (2.4.3)의 $(n-1)!$은 감마함수 □$\Gamma(n)=\int_0^\infty e^{x}x^{n-1}dx$□로 대체해야 되는데, $\Gamma(n)$의 주요 속성은 $\Gamma(n)=(n-1)\Gamma(n-1)$이다.

2.4.5 연속 *Uniform* 분포

n회의 독립시행 중에서 단 한번 성공이 발생했다는 조건 하에, 성공이 발생한 시행이 i번째의 시행일 확률은 모든 i에 대해서 (단, $i=1,\cdots,n$) $1/n$로 동일한데, 이를 (이산) *Uniform* 분포라 했다. 이와 같이, $(0,s)$ 동안에 단 한번 성공이 발생했다는 조건 하에, 성공이 $(t,t+dt)$에서 발생했을 확률은 모든 $(t,t+dt)$에 대해서 (단, $0<t<s$) dt/s로 동일한데, 이를 연속 *Uniform* 분포라 한다.

이를 정식으로 유도하되 다음과 같이 확장한다.

$$P[\,n_1\text{번째 성공이 }(t,t+dt)\text{에서 발생}\mid(0,s)\text{에서 } n_1+n_2-1\text{번 성공}\,]$$
$$=P[\,(0,t)\text{에서 } n_1-1\text{번 성공}, (t,t+dt)\text{에서 성공}, (t,s)\text{에서 } n_2-1\text{번 성공}\,]$$
$$/P[\,(0,s)\text{에서 } n_1+n_2-1\text{번 성공}\,] \qquad (2.4.4)$$

인데, 이는 조건부 확률의 정의를 이용한 결과이다. 이후의 과정은 음초기하분포인 식

(2.1.5)를 유도할 때와 유사한데, 차이점은 이항분포 대신에 포아송 분포를 활용한다는 점이다. 물론, 독립시행이므로 식 (2.4.4)의 분자는 세 가지 확률의 곱인 $P[S(t)=n_1-1]\cdot\lambda dt\cdot P[S(s-t)=n_2-1]$이고, 분모는 $P[S(s)=n_1+n_2-1]$이다. 식 (2.3.1)을 대입하여 간단히 하면

$$\frac{(n_1+n_2-1)!}{(n_1-1)!(n_2-1)!}\left(\frac{t}{s}\right)^{n_1-1}\frac{dt}{s}\left(1-\frac{t}{s}\right)^{n_2-1},\ 0<t<s \qquad (2.4.5)$$

를 얻는데, $n_1=n_2=1$인 경우가 연속 *Uniform* 분포이다.

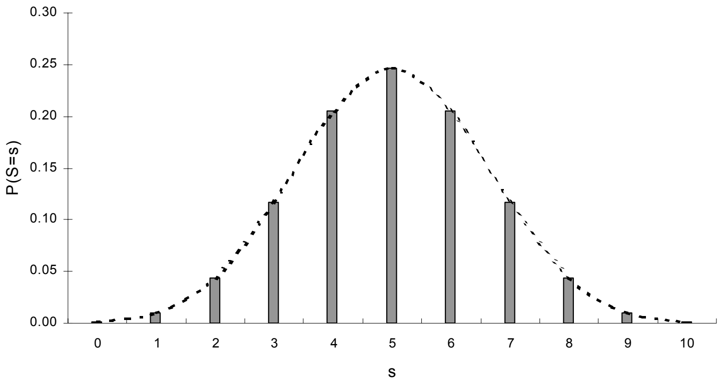
2.4.6 베타(*beta*) 분포

Erlang 분포에서 자연수 n을 양의 실수로 간주하면 감마분포가 되듯이, 식 (2.4.5)에서 자연수 n_1, n_2를 양의 실수로 간주하면 베타분포가 된다. 단, $(n_1+n_2-1)!$, $(n_1-1)!$, $(n_2-1)!$은 각각 $\Gamma(n_1+n_2)$, $\Gamma(n_1)$, $\Gamma(n_2)$로 대체해야 된다. 특히, $s=1$인 경우를 표준(standard) 베타분포라 하는데, 이의 밀도함수는 다음과 같다.

$$f(t)=t^{n_1-1}(1-t)^{n_2-1}/B(n_1,n_2) \qquad (2.4.6)$$

식 (2.4.6)에서 $B(n_1,n_2)=\Gamma(n_1)\Gamma(n_2)/\Gamma(n_1+n_2)=\int_0^1 t^{n_1-1}(1-t)^{n_2-1}dt$를 베타함수라 한다.

§2.5 정규(*normal*) 분포



<그림 2.1> $P(S=s)=\binom{10}{s}(\frac{1}{2})^{s}(\frac{1}{2})^{10-s}$

<그림 2.1>은 $n=10$, $p=1/2$인 이항분포의 그래프이다. 굵은 선은 확률인데, 예를 들어 $P(S=5)=\binom{10}{5}/2^{10}=252/1024\approx 0.2461$이다. 반면에, 점선은 포락선(envelope)인데, 확률이 기둥이라면 포락선은 지붕과 같다.

$n\to\infty$이면 포락선은 곡선이 되는데, 이것이 바로 정규분포의 밀도함수이다. 이 사실은 18세기 초에 De Moivre가 밝혔는데, 19세기 초에 Laplace는 $p\neq 1/2$이더라도 이 사실이 여전히 성립함을 보였다. 또한, 이 사실은 20세기 초에 중심극한정리(central limit theorem)로 확장되는데, 그 요지는 다음과 같다. 이항 확률변수 $S=\sum_{i=1}^{n}Y_i$에서 $Y_1,\cdots,Y_n$은 *Bernoulli* 분포를 따르는 *iid* 확률변수이다. 그런데, 중심극한정리에 의하면, $Y_1,\cdots,Y_n$가 *iid* 확률변수이기만 하면 그 분포가 무엇이든지 불문하고 이 사실이 성립한다는 것이다. 구체적으로, $Y_1,\cdots,Y_n$이 *iid* “이산” 확률변수이면 $S=\sum_{i=1}^{n}Y_i$의 분포의 “포락선”이 n이 커짐에 따라 정규분포(의 밀도함수)에 가까워지고, $Y_1,\cdots,Y_n$이 *iid* “연속” 확률변수이면 $S=\sum_{i=1}^{n}Y_i$의 “밀도함수”가 n이 커짐에따라 정규분포(의 밀도함수)에 가까워진다.

정규분포(의 밀도함수)는 다음과 같다.

$$f_s(s)=\frac{1}{\sigma\sqrt{2\pi}}\,e^{-\frac{1}{2}(\frac{s-\mu}{\sigma})^2},\ -\infty<s<\infty \tag{2.5.1}$$

정규(분포를 따르는) 확률변수를 S라 하면, 식 (2.5.1)에서 $\mu = E(S)$이고 $\sigma^2 = V(S)$이다. 즉, μ는 S의 기대치이고 σ^2은 S의 분산이다. (따라서, σ는 S의 표준편차인데, $f_s(s)$의 그래프에서 변곡점의 위치가 바로 $\mu \pm \sigma$이다.)

정규분포 중에서 $\mu = 0$, $\sigma = 1$인 경우를 표준(standard) 정규분포라 한다. 관행상, 표준정규 확률변수는 Z로 표기하는데, Z의 밀도함수는 다음과 같다.

$$f_Z(z) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}z^2},\ -\infty < z < \infty \tag{2.5.2}$$

<비고 2.5.1> 식 (2.5.1)이 S의 밀도함수이면 식 (2.5.2)는 $Z = (S - \mu)/\sigma$의 밀도함수임 (§ 2.9.1 참조).

§2.6 정규분포 관련 분포

2.6.1 카이제곱(*chi-squared*) 분포

$Z_1, Z_2, \cdots$ 가 표준정규분포를 따르는 iid 확률변수일 때, $C_d = \sum_{i=1}^{d} Z_i^2$는 카이제곱분포를 따르는데, 이때 자유도(degree of freedom)는 d이다 (§2.15.4 참조).

C_d의 밀도함수는 감마분포의 특수한 경우인데, 식 (2.4.3)에 $\lambda = 1/2$, $n = d/2$를 대입하여 다음과 같이 얻는다.

$$f(t) = \begin{cases} \dfrac{(t/2)^{\frac{d}{2}-1} e^{-t/2}}{\Gamma(d/2)} \dfrac{1}{2} & , t \geq 0 \\ 0 & , t < 0 \end{cases} \tag{2.5.3}$$

2.6.2 F 분포

C_{d_1}과 C_{d_2}가 카이제곱분포를 따르고 서로 독립이며 자유도는 각각 d_1과 d_2일 때,

$$F_{d_1, d_2} = \frac{C_{d_1}/d_1}{C_{d_2}/d_2} \tag{2.5.4}$$

는 F분포를 따른다. 이때, d_1을 분자 자유도라 하고 d_2는 분모 자유도라 한다. (밀도함수는 복잡해서 생략함.)

2.6.3 t 분포

Z와 C_d는 각각 표준정규분포와 (자유도가 d인) 카이제곱분포를 따르며 서로 독립일 때,

$$T_d = \frac{Z}{\sqrt{C_d/d}} \tag{2.5.5}$$

의 분포를 (자유도가 d인) t 분포라 하는데, 밀도함수는 다음과 같다 (유도과정은 §2.11.4 참조).

$$f(t) = \frac{\Gamma[(d+1)/2]}{\sqrt{\pi d}\,\Gamma(d/2)}\left(1+\frac{t^2}{d}\right)^{-\frac{1}{2}(d+1)}, \quad -\infty < t < \infty \tag{2.5.6}$$

t 분포는 사실상 F 분포의 특수한 경우라 할 수 있는데, 그 관계는 다음과 같다. 식 (2.5.5)를 제곱하면 $T_d^2 = Z^2/(C_d/d)$인데, Z^2은 자유도가 1인 카이제곱분포를 따르므로

$$T_d^2 = \frac{C_1/1}{C_d/d} = F_{1,d} \tag{2.5.7}$$

이다. 즉, T_d^2은 분자 자유도가 1이고 분모자유도가 d인 F분포를 따른다.

<비고 2.6.1> 식 (2.5.5)에서 $d \to \infty$이면 T_d의 분포가 Z의 분포와 같아진다.

§2.7 연속분포의 특징

2.7.1 연속 모분포

지금까지 등장한 이산분포 중에서 §2.1의 복원추출 관련 분포와 §2.3의 포아송 분포는 모두 독립시행과 관련이 있다. 그러나 이산분포 중에는 §2.2의 비복원추출 관련 분포같이 독립시행과 관련이 없는 분포도 있다. 그런데, 지금까지 등장한 연속분포는 모두 (직접 또는 간접적으로) 독립시행과 관련이 있다.

사실, 연속분포가 등장하면서부터는 아예 복원과 비복원을 거론하지 않았다. 통계학에서 모분포로 가장 많이 쓰이는 분포는 정규분포이다. 그런데, 모분포가 정규분포같은 연속분포이면 복원과 비복원의 차이는 아예 없어진다. 이는, 단순히 모집단의 크기를 $N \to \infty$ 이라 가정하는 것과는 근본적으로 다르며, 표본의 크기가 아무리 크더라도 성립한다. (비고 : 표본의 크기가 아무리 커도 셀 수 있는 것(countable)인 반면에, 연속분포를 따르는 모집단은 연속체(continuum)이므로 그 크기를 셀 수 없음.)

<비고 2.7.1> 모분포가 연속이면, 표본 $\{Y_1, \cdots, Y_n\}$에서 n이 아무리 크더라도 $Y_1, \cdots, Y_n$은 iid 확률변수이(고 각각 모분포를 따른)다.

2.7.2 연속분포의 표현

편의상, 지수분포를 따르는 T와 기하분포를 따르는 Y를 예로 든다. $f(t) = \lambda e^{-\lambda t}$, $t \ge 0$ 을 T의 "밀도함수"라 했는데, 이때 밀도라는 표현은 확률을 질량(mass)에 비유한 표현이다. 이에 따라. $P(Y=y) = q^{y-1}p$, $y = 1, 2, \cdots$ 를 Y의 "질량함수"라 부르기로 한다. 또한, 분포를 표현할 때, CDF(cumulative distribution function)를 사용하기도 하는데, 정의는 $F_T(t) = P(T \le t)$, $F_Y(y) = P(Y \le y)$이다.

<비고 2.7.2> (확률)분포라는 표현은 포괄적인 것으로써, 이산 확률변수의 경우에는 CDF 또는 질량함수를 그리고 연속 확률변수의 경우에는 CDF 또는 밀도함수를 지칭하는 표현이다.

관행상, $F_Y(y)$와 $F_T(t)$는 각각 모든 y와 t에 대해서 정의한다. 따라서,

$F_Y(y)$는 $y<1$에서는 0이다가 $y=1,2,\cdots$에서 각각 $q^{y-1}p$만큼씩 점프(jump)하는 계단식 함수(step function)가 된다. 반면에, $F_T(t)$는 $t<0$에서는 0이다가 $t\geq 0$에서는 연속적으로 증가한다. 구체적으로, $F_T(t)=P(T\leq t)=1-P(T>t)$인데, $t\geq 0$이면 $P(T>t)$ $=P(t$시간 동안 0번 성공$)=P[S(t)=0]=e^{-\lambda t}$이므로

$$F_T(t)=\begin{cases} 1-e^{-\lambda t} & ,t\geq 0 \\ 0 & ,t<0 \end{cases}$$

이다.

수학적인 정의에 따르면, CDF가 불연속인 Y는 이산 확률변수이고, CDF가 연속인 T는 연속 확률변수이다. 그리고, 연속 확률변수의 CDF의 미분(또는, 증가율)을 밀도함수로 정의한다. 즉, $f(t)=\frac{d}{dt}F_T(t)$이다. CDF는 확률이지만 CDF의 증가율인 밀도함수는 확률이 아니다. 그러나 $dF_T(t)=f(t)dt$는 확률인데, 이미 이를 $P(t<T<T+dt)$라 표현했다. 즉,

$$dF_T(t)=dP(T\leq t)=P(T\leq t+dt)-P(T\leq t)=P(t<T\leq t+dt)$$

이다. (비고: T가 연속 확률변수이면 모든 t에 대해서 $P(T=t)=0$ 이므로, $P(t<T\leq t+dt)=P(t<T<t+dt)$임.)

<비고 2.7.3> 이산 확률변수인 Y의 $dF_Y(y)$는 다음과 같다.

$$dF_Y(y)=P(Y=y)=\begin{cases} q^{y-1}p & , \text{ if } y=1,2,3,\cdots \\ 0 & , \text{ if } y\neq 1,2,3,\cdots \end{cases}$$

2.7.3 혼합(mixed) 분포

이산분포와 연속분포가 혼합된 분포의 예는 다음과 같다. 전구를 갈아 끼우고 나서 스위치를 켜면 q의 확률로 전구가 터지고, $p(=1-q)$의 확률로 불이 들어온다. 그리고, 불이 들어오는 전구의 수명은 지수분포를 따른다고 하자.

이 전구의 수명을 T라 하면, $P(T<0)=0$, $P(T=0)=q$, $P(T>0)=p$ 이다. 그리

고, $t>0$에 대해서 $P(T>t \mid T>0) = P[S(t)=0] = e^{-\lambda t}$이다. 따라서, $t>0$에 대해서 $P(T>t) = P(T>0, T>t) = P(T>0)P(T>t \mid T>0) = p\,e^{-\lambda t}$이다. 이를 CDF로 표현하면 다음과 같다.

$$F(t) = P(T \le t) = \begin{cases} 0 & , t<0 \\ q & , t=0 \\ q + p\,e^{-\lambda t} & , t>0 \end{cases}$$

§2.8 기대치

2.8.1 기대치와 평균

시험지를 채점하고나서 점수분포를 얻으면, 제일 먼저 계산하는 것이 평균이다. 평균(mean)은 점수분포의 특성치로써, 모든 점수를 하나의 숫자로 대표하는 값이다. 평균 다음으로 중요한 특성치는 분산(variance)인데, 이는 점수분포가 평균을 중심으로 얼마나 퍼져(또는, 흩어져) 있는가를 알려준다.

N명의 학생을 모집단이라 하고 이들 학생을 점수에 대응시키는 함수를 확률변수 Y라 하자. 즉, Y는 무작위로 뽑히는 학생의 점수를 의미한다. (편의상, 종전대로 Y의 분포를 모분포라 한다.) 점수가 y인 학생의 수를 n_y라 하자. 그러면, 성적이 y인 학생이 무작위로 뽑힐 확률은 $P(Y=y)=n_y/N$ 인데, 이것이 Y의 분포(또는, 모분포)이다 (§1.3 참조).

이제, N명의 평균점수를 다음과 같이 구한다.

$$\mu=\frac{\sum_y y\cdot n_y}{N}=\sum_y y\cdot P(Y=y)=E(Y) \tag{2.8.1}$$

식 (2.8.1)에서 $\sum_y y\cdot n_y$는 N명의 점수를 모두 합친 것이고, 이를 학생수 N으로 나눈 것이 바로 평균 μ (의 정의) 이다. 반면에, $\sum_y y\cdot P(Y=y)$는 확률변수 Y가 구현(realize)될 수 있는 모든 y값들에 가중치(weight)인 (Y가 y로 구현될) 확률 $P(Y=y)$를 곱해서 모두 더해 놓은 것으로써, 바로 Y의 기대치(expected value)인 $E(Y)$의 정의이다.

2.8.2 평균과 중심

분산은 분포가 평균을 "중심"으로 얼마나 퍼져있는가를 알려준다고 했는데, 사실 평균은 "중심"과 동일한 개념이다. Y가 이산 확률변수이면 $P(Y=y)$는 질량함수인데, 질량중심(center of mass)이 평균임을 다음과 같이 보일 수 있다.

$$\begin{aligned}\sum_{y<\mu}(\mu-y)\cdot P(Y=y) &= \mu\{1-\sum_{y\geq\mu}P(Y=y)\}-\{\mu-\sum_{y\geq\mu}y\cdot P(Y=y)\}\\ &= \sum_{y>\mu}(y-\mu)\cdot P(Y=y)\end{aligned}$$

이 식에서 $P(Y=y)$ 는 y 지점에 위치한 질점(mass point)의 질량인데, μ 보다 좌측에 ($y<\mu$) 위치한 질점의 질량에 μ 로부터 떨어진 거리 $\mu-y$ 를 곱한 값들의 합은 μ 보다 우측에($y>\mu$) 위치한 질점의 질량에 μ 로부터의 거리 $y-\mu$ 를 곱한 값들의 합과 같다.

<비고 2.8.1> y_0 의 함수 $\sum_y (y-y_0)^2 P(Y=y)$ 는 $y_0=\mu$ 일 때 최소가 되는데, 이때 최소 값인 $\sum_y (y-\mu)^2 P(Y=y)$ 를 분산이라고 하고 σ^2 으로 표기한다.

2.8.3 평균과 중앙값

모든 점수를 하나의 숫자로 대표하는 값으로 평균외에 중앙값(median)이라는 것이 있다. 중앙값을 m 이라 하면, y_0 의 함수 $\sum_y |y-y_0| \cdot P(Y=y)$ 는 $y_0=m$ 일 때 최소가 된다 (<비고 2.8.1> 참조).

편의상, N 개의 점수를 $y_1<y_2<\cdots<y_{N-1}<y_N$ 이라 하자. (비고: $P(Y=y_i)=1/N$, $i=1,\cdots,N$.) 그러면, μ 는 y_i 들로부터의 거리를 제곱한 값의 합이 최소가 되는 값이고, m 은 y_i 들로부터의 (절대값) 거리의 합이 최소가 되는 값이다. 그리고, m 의 값은 N 이 홀수이면 $y_{(N+1)/2}$ 이지만, N 이 짝수인 경우에는 $y_{N/2}$ 와 $y_{(N/2)+1}$ 의 (사이에 있는 아무 값이라도 상관없지만 관례상) 평균으로 정의한다.

2.8.4 Y 의 함수의 기대치

Y 가 이산 확률변수이면 Y 의 함수 $X=g(Y)$ 도 이산 확률변수이다. 그런데, X 의 기대치 $E(X)=\sum_x x \cdot P(X=x)$ 는 $E[g(Y)]=\sum_y g(y) \cdot P(Y=y)$ 로 구할 수도 있다 (증명 생략). 예를 들면, $E(Y^2)=\sum_y y^2 \cdot P(Y=y)$ 이고, $E[(Y-\mu)^2]=\sum_y (y-\mu)^2 \cdot P(Y=y)$ 이다.

$E[(Y-\mu)^2]$ 을 $V(Y)$ 로 표기하는데, 이는 바로 분산인 σ^2 이다. (<비고 2.8.1> 참조). 즉,

$$V(Y)=E[(Y-\mu)^2]=\sum_y (y-\mu)^2 P(Y=y)=\sigma^2 \qquad (2.8.2)$$

이다. $V(Y)$ 를 계산할 때 편리한 공식은 다음과 같다 (증명 생략).

$$V(Y)=E(Y^2)-\{E(Y)\}^2 \tag{2.8.3}$$

<비고 2.8.2> 식 (2.8.3)을 □ $E(Y^2)=\sigma^2+\mu^2$ □으로 표현하면, □ E(0으로부터의 거리의 제곱) = E(μ 로부터의 거리의 제곱)+ (0으로부터 μ 까지의 거리의 제곱)□ 으로 해석할 수 있다.

2.8.5 연속 확률변수의 기대치

Y 가 이산 확률변수이면 $E[g(Y)]=\sum_y g(y)\cdot P(Y=y)$ 라 했다. (비고: $E(Y)=\sum_y y\cdot P(Y=y)$ 는 $g(Y)=Y$ 인 특수한 경우임.) 이를 CDF $F(y)=P(Y\le y)$로 표현하면

$$E[g(Y)]=\int_{-\infty}^{\infty} g(y)dF(y) \tag{2.8.4}$$

인데(<비고 2.7.3>참조), 이는 Y 가 연속 확률변수일 때에도 성립한다. 물론, Y 가 연속 확률변수이면, $dF(y)=f(y)dy$ 이므로 흔히 다음과 같이 표현한다.

$$E[g(Y)]=\int_{-\infty}^{\infty} g(y)f(y)dy$$

<비고 2.8.3> 식(2.8.4)는 혼합분포의 경우에도 성립함 (§2.7.3 참조).

$P(Y=y)$ 가 y 지점에 위치한 질점의 질량이라면, $f(y)$ 는 일차원 연속체의 y 지점에서의 밀도이다. 그리고, 이 일차원 연속체를 아주 잘게 잘랐을 때, y 와 $y+dy$ 사이에 위치한 부분의 질량이 바로 $dF(y)=f(y)dy$ 이다.

불연속적으로 위치한 질점들의 질량대신에 연속체의 밀도로 표현을 했을 뿐이지, 평균, 중앙값, 분산 등의 의미는 동일하다. 즉, 평균은 연속체의 질량중심이고 중앙값은 ($P(Y=m)=0$ 이기 때문에 오히려 표현하기기가 편리한데) $P(Y<m)=\int_{-\infty}^{m} f(y)dy=\int_{m}^{\infty} f(y)dy=P(Y>m)$ 을 만족시키는 m 값이다. 또한, $\int_{-\infty}^{\infty}(y-y_0)^2 f(y)dy$ 를 최소가 되게 하는 y_0 값이 $\mu=E(Y)$ 이고, 이때 최소값은 분산이다. 즉,

$$V(Y) = E[(Y-\mu)^2] = \int_{-\infty}^{\infty}(y-\mu)^2 f(y)\,dy = \sigma^2$$

이다. 그리고, $\int_{-\infty}^{\infty}|y-y_0|f(y)dy$ 를 최소가 되게 하는 y_0 값이 중앙값 m 이다.

2.8.6 0-1 확률변수의 기대치

0 또는 1의 값을 가지는 확률변수를 *Bernoulli* 확률변수라 불렀다 (§2.1.1 참조). *Bernoulli* 분포 $P(Y=1)=p$, $P(Y=0)=q$ $(=1-p)$ 의 특징은 다음과 같다.

$$E(Y) = \sum_y y \cdot P(Y=y) = 1 \cdot p + 0 \cdot q = p \tag{2.8.5}$$

즉, Y 의 기대치는 $P(Y=1)$ 과 같다.

<비고 2.8.4> 식 (2.8.5)는 확률을 기대치에 연결해 주는 역할을 한다. 사실, 확률변수를 적절히 정의하기만 하면 모든 확률을 기대치로 해석할 수 있다.

그런데, $E(Y)$ 뿐만 아니라 $E(Y^2)$ 역시 $P(Y=1)$ 이다. 즉, $E(Y^2) = \sum_y y^2 \cdot P(Y=y) = 1^2 \cdot p + 0^2 \cdot q = p$ 이다. 따라서, 식 (2.8.3)에 의해서 분산을 다음과 같이 얻는다.

$$V(Y) = E(Y^2) - E(Y)^2 = p - p^2 = pq \tag{2.8.6}$$

2.8.7 이항분포의 평균과 분산

식 (2.8.4) 하나로 모든 기대치를 구할 수 있다. §2.1.2 의 이항분포의 경우 $dF(s) = \binom{n}{s} p^s q^{n-s}$, $s = 0, 1, \cdots, n$, ($dF(s) = 0$ if $s \neq 0, 1, \cdots, n$)을 식 (2.8.4)에 대입하면

$$E(S) = \sum_{s=0}^{n} s \binom{n}{s} p^s q^{n-s}$$

$$V(S) = E[\{S - E(S)\}^2] = \sum_{s=0}^{n} \{s - E(S)\}^2 \cdot \binom{n}{s} p^s q^{n-s}$$

가 된다. 그런데, $E(S)$ 와 $V(S)$ 를 계산하는 과정은 복잡하(거니와 약간의 트릭(trick)을 요하기도 한)다.

그러나, $S=\sum_{i=1}^{n} Y_i$ 의 관계를 이용하면 $E(S)$ 와 $V(S)$ 의 계산이 쉬어진다. (비고: $Y_1, \cdots, Y_n$ 은 *Bernoulli* 분포를 따르는 *iid* 확률변수임.) $Y_1, \cdots, Y_n$ 이 *iid* 확률변수이면

$$E(\sum_{i=1}^{n} Y_i)=\sum_{i=1}^{n} E(Y_i)=nE(Y) \tag{2.8.8}$$

$$V(\sum_{i=1}^{n} Y_i)=\sum_{i=1}^{n} V(Y_i)=nV(Y) \tag{2.8.9}$$

가 성립하는데 (§2.12.6 참조), 식 (2.8.5)와 (2.8.6)을 대입하면 $E(S)=np$ 와 $V(S)=npq$ 를 얻는다.

<비고 2.8.5> $Y_1, \cdots, Y_n$ 이 연속 확률변수일 때도 식 (2.8.8)과 (2.8.9) 는 성립함.

<비고 2.8.6> 식 (2.8.8)은 $Y_1, \cdots, Y_n$ 이 독립이 아니더라도 (동일하기만 하면) 성립함.

2.8.8 이산분포의 평균과 분산

§2.8.6에서 구한 *Bernoulli* 분포의 평균과 분산은 $E(Y)=p$, $V(Y)=pq$ 인데, 이로부터 §2.8.7에서 이항분포의 평균 $E(S)=nE(Y)=np$ 와 $V(S)=nV(Y)=npq$ 를 얻었다.

이와 같이, 기하분포의 평균과 분산 $E(Y)=1/p$ 와 $V(Y)=q/p^2$ 로부터 (§2.12.2 참조), 음이항분포의 평균 $E(S)=nE(Y)=n/p$ 와 분산 $V(S)=nV(Y)=nq/p^2$ 을 얻는다.

포아송 분포는 이항분포로부터 얻었으므로, 포아송 분포의 평균과 분산도 이항분포의 평균 np 와 분산 npq 로부터 얻을 수 있다. $np=\lambda t$ 를 대입하면 각각 λt 와 λtq 인데 $q=\lim_{n\to\infty}(1-\frac{\lambda t}{n})=1$ 이므로 평균과 분산이 모두 λt 가 된다.

<비고 2.8.7> 포아송 확률변수 $S(t)$ 의 특징은 $E[S(t)]=V[S(t)]=\lambda t$ 임. 따라서, λ 는 단위시간당 발생하는 성공횟수의 기대치인 동시에 분산임.

<비고 2.8.8> WMS (문헌 [9])를 포함한 일부 통계학 교재에서는 포아송 분포에서 $t=1$ 인 경우만 취급하고 있음.

비복원추출 관련 분포의 평균과 분산은 복잡하므로 (문헌 [7] 참조), 대표로 초기하분포의 평균과 분산만 제시한다. (유도과정은 문헌 [9] 참조.)

$$E(S) = n \cdot \frac{m}{N} \tag{2.8.8}$$

$$V(S) = n \cdot \frac{m}{N}\left(1 - \frac{m}{N}\right)\frac{N-n}{N-1} \tag{2.8.9}$$

식 (2.8.8)의 형태는 복원추출 관련 분포인 이항분포의 평균 np 의 형태이다. (비고: <사례 1.3>에서 m/N 은 임의로 잡은 동물이 꼬리표를 달고 있을 확률임.) 반면에, 식 (2.8.9)는 이항분포의 분산 npq 와 약간 다른 형태이다. 차이점은 $(N-n)/(N-1)$ 인데, 이는 바로 비복원추출에 따른 종속성에 기인한 것이다. 그러나, 비복원추출이더라도 $N \gg n$ 이면 $(N-n)/(N-1) \approx 1$ 이 되어 차이를 무시할 수 있게 된다 (§1.2, §1.4 참조).

2.8.9 연속분포의 평균과 분산

지수분포의 평균과 분산은 $E(T) = \lambda^{-1}$ 과 $V(T) = \lambda^{-2}$ 이다 (§2.12.3 참조).

<비고 2.8.9> 지수 확률변수 T 의 특징은 $E(T) = SD(T) = \lambda^{-1}$ 이다. 즉, 기대치와 표준편차(standard deviation)가 같다.

음이항분포의 평균과 분산이 각각 기하분포의 평균과 분산의 n 배씩이듯이, Erlang 분포의 평균과 분산은 각각 지수분포의 평균과 분산의 n 배씩이다. 즉, $E(Y) = n/\lambda$ 이고 $V(Y) = n/\lambda^2$ 이다. 그런데, 자연수 n 을 양의 실수로 간주하여 Erlang 분포를 감마분포로 확장하더라도 평균과 분산은 여전히 동일한 형태이다. 예를 들어, 카이제곱분포는 감마분포의 특수한 경우로써 $\lambda = 1/2$, $n = d/2$ 인 경우인데, 이를 대입하면 $E(C_d) = n/\lambda = (d/2)/(1/2) = d$ 와 $V(C_d) = n/\lambda^2 = (d/2)/(1/2)^2 = 2d$ 를 얻는다.

<비고 2.8.10> 카이제곱 확률변수 C_d 의 특징은 $E(C_d) = d$ 와 $V(C_d) = 2d$ 이다. 즉, 기대치는 자유도와 같고, 분산은 자유도의 2배이다.

표준 베타분포인 식 (2.4.6)의 평균은 $n_1/(n_1+n_2)$ 이고 분산은 $n_1n_2/\{(n_1+n_2)^2(n_1+n_2-1)\}$ 이다 (문헌 [5] 참조).

정규분포인 식 (2.5.1)은 이미 평균 μ 와 분산 σ^2 으로 표현되어 있다. 정규분포 관련 분포 중에서 카이제곱분포는 감마분포의 특수한 경우로 다루었으므로, 이제 남은 것은 F 분포와 t 분포이다. F 분포의 평균은 $d_2/(d_2-2)$ 이다 ($d_2 \geq 3$). 즉, 분모 자유도인 d_2 만의 함수인데, $d_2 \to \infty$ 이면 1 에 수렴한다. F 분포의 분산은 복잡해서 생략한다 (문헌 [9] 참조). t 분포인 식 (2.5.6)은 표준정규분포인 식 (2.5.2)같이 0을 중심으로 좌우 대칭이다. 따라서, 평균과 중앙값이 모두 0이다. (단, 자유도가 $d \geq 2$ 인 경우에만 평균이 정의됨.) t 분포의 분산은 F 분포의 평균과 같은 형태로써 $d/(d-2)$ 인데 ($d \geq 3$), 그 이유는 다음과 같다. 평균이 0 이면 식 (2.8.3)에 의해서 $V(T_d)=E(T_d^2)$ 인데, 식 (2.5.7)에 의해서 $E(T_d^2)=E(F_{1d})$ 이므로 $V(T_d)=E(F_{1d})=d/(d-2)$ 이다.

§2.9 $g(Y)$ 의 분포

2.9.1 $aY+b$ 의 분포

§2.5의 <그림 2.1>은 이항분포 $P(S=s)=\binom{10}{s}/2^{10}$ 의 그래프인데, 이로부터 S 의 선형(linear)함수인 $X=2S+10$ 의 분포를 구해보자. 예를 들면, $P(X=20)=$ $P(2S+10=20)=P(S=5)=\binom{10}{5}/2^{10} \approx 0.2461$ 이다. 일반적으로, $P(X=x)=$ $P(2S+10=x)=P[S=(x-10)/2]$ 이므로 편의상 $x=2s+10$ 이라 하면 $P(X=x)$ $=P(S=s)$ 이다. 그렇지만, $s=0,1,2,\cdots,10$ 일 때에 한해서 $P(S=s)=\binom{10}{s}/2^{10}$ 이고, $s \neq 0,1,2,\cdots,10$ 이면 $P(S=s)=0$ 이다. 따라서, $x=10,12,\cdots,30$ 일 때에 한해서 $P(X=x)=\binom{10}{(x-10)/2}/2^{10}$ 이고, $x \neq 10,12,\cdots,30$ 이면 $P(X=x)=0$ 이다.

$P(X=x)$ 의 그래프는 $P(S=s)$ 의 그래프를 수평방향으로 2배로 확대한 다음에 10만큼 수평이동시킨 것이다. 그러나 수직방향으로는 변화가 없다. 즉, $P(X=x)$ 의 크기는 여

전히 $P(S=s)$ 의 크기와 동일하다 (단, $x=2s+10$). (비고: $1=\sum_x P(X=x)$ $=\sum_s P(S=s)$.) 따라서, 분포의 평균도 수평방향으로 2배로 확대되고 나서 10만큼 수평이동된다. 즉, $E(X)=E(2S+10)=2E(S)+10=2\cdot 5+10=20$ 이다. 반면에, 평균을 중심으로 분포가 퍼진 정도는 2배로 증가한다. 즉, $SD(X)=SD(2S+10)=SD(2S)=2\cdot SD(S)$ 이다. 따라서, $V(X)=V(2S+10)=V(2S)=2^2\cdot V(S)$ 가 된다.

<비고 2.9.1> 분포의 수평이동은 표준편차와 분산에 영향을 미치지 않는다.

이제, 연속분포인 정규분포를 예로 든다. 식 (2.5.2)는 표준정규 확률변수 Z 의 밀도함수인데, 이로부터 Z 의 선형함수인 $S=\sigma Z+\mu$ 의 밀도함수를 구해보자.

$$\begin{aligned} P(s<S<s+ds) &= P(s<\sigma Z+\mu<s+ds) \\ &= P\left(\frac{s-\mu}{\sigma}<Z<\frac{s+ds-\mu}{\sigma}\right) \end{aligned}$$

인데, 편의상 $s=\sigma z+\mu$ 라 하면 (비고: $ds=\sigma dz$), $f_S(s)ds=P(s<S<s+ds)$ $=P(z<Z<z+dz)=f_Z(z)dz$ 를 얻는다. 따라서, $f_S(s)=f_Z(z)dz/ds$ 인데, $f_Z(z)=e^{-z^2/2}/\sqrt{2\pi}$ 에 $z=(s-\mu)/\sigma$ 를 대입하면 $e^{-\{(s-\mu)/\sigma\}^2}/\sqrt{2\pi}$ 가 되고 $dz/ds=1/\sigma$ 이므로 $f_S(s)$ 는 식 (2.5.1)과 일치한다.

$f_S(s)$ 의 그래프는 $f_Z(z)$ 의 그래프를 수평방향으로 σ 배로 확대한 다음 μ 만큼 수평이동한 것이다. 그런데, 수직방향으로는 (이산분포에서는 변화가 없었지만) $1/\sigma$ 배로 축소된다. 즉, $f_S(s)=f_Z(z)/\sigma$ 이다 (단, $s=\sigma z+\mu$). 이는 1차원 연속체를 수평방향으로 σ 배로 늘림으로 인해서 밀도가 $1/\sigma$ 배로 줄어든 것이다. (비고: $1=\int_{-\infty}^{\infty} f_S(s)\,ds$ $=\int_{-\infty}^{\infty} f_Z(z)\,dz$.) 물론, 평균의 위치는 0 에서 μ 로 이동하고, 표준편차는 1에서 σ 로 늘어난다.)

일반적으로, 확률변수 Y 의 선형함수인 $X=aY+b$ 의 기대치는 $E(X)=aE(Y)+b$ 이고 분산은 $V(X)=a^2V(Y)$ 이다. 그리고, 표준편차는 $SD(X)=|a|\cdot SD(Y)$ 인데, a 의 절대값을 사용하는 이유는 표준편차의 정의상 비음(non-negative)이기 때문이다. (비고: X 의 분포와 $-X$ 의 분포의 표준편차는 동일함.) 또한, $x=ay+b$ 라 할 때, Y 가 이산 확률

변수이면 $P(X=x)=P(Y=y)$ 이지만, Y 가 연속확률변수이면 $f_X(x)=f_Y(y)/|a|$ 인데, a 의 절대값을 사용하는 이유는 밀도함수가 비음이기 때문이다.

2.9.2 $g(Y)$ 의 분포

확률변수 Y 의 분포로부터 Y 의 함수인 $X=g(Y)$ 의 분포를 얻는 방법은 다음과 같다. 먼저, 함수 $g(\cdot)$ 가 1:1 함수인 경우를 다룬다. 편의상 $x=g(y)$ 라 하면, Y 가 이산 확률변수인 경우에는

$$P(X=x)=P[g(Y)=g(y)]=P(Y=y) \tag{2.9.1}$$

이고, Y 가 연속 확률변수인 경우에는

$$f_X(x)=f_Y(y)\cdot\left|\frac{dy}{dx}\right| \tag{2.9.2}$$

이다. 단, 우변도 좌변과 같이 x 로 표현하려면 $y=g^{-1}(x)$ 를 대입하면 된다 (§2.9.3, §2.9.4 참조). (비고: 1:1 함수인 $g(\cdot)$ 가 증가함수이면 $(dy/dx)>0$ 이고, 감소함수이면 $(dy/dx)<0$ 임.)

다음, 함수 $g(\cdot)$ 가 1:1이 아닌 경우에는 식 (2.9.1)과 (2.9.2)같은 일반적인 관계식을 만들 수 없으므로 문제 별로 다루어야 되는데, 이때 질량함수와 CDF를 사용하면 편리하다. 예를 들어, Y 가 이산 확률변수인 경우에 $X=g(Y)=Y^2$ 이면, $x\geq 0$ 에 대해서 $P(X=x)$ $=P(Y^2=x)=P(Y=\pm\sqrt{x})=P(Y=+\sqrt{x})+P(Y=-\sqrt{x})$ 이다. 또한, Y 가 연속 확률변수인 경우에 $X=g(Y)=Y^2$ 이면, $x\geq 0$ 에 대해서 $F_X(x)=P(X\leq x)=P(Y^2\leq x)$ $=P(-\sqrt{x}\leq Y\leq\sqrt{x})=P(Y\leq\sqrt{x})-P(Y<-\sqrt{x})=F_Y(\sqrt{x})-F_Y(-\sqrt{x})$ 이다 (비고: $P(Y<-\sqrt{x})=P(Y\leq-\sqrt{x})$). 그리고, 양변을 x 에 대해서 미분하면 X 의 밀도함수를 얻는다. 즉,

$$f_X(x)=\{f_Y(\sqrt{x})+f_Y(-\sqrt{x})\}/2\sqrt{x} \tag{2.9.3}$$

이다. 예를 들어, Y 가 표준정규분포를 따르면 식 (2.5.2)에 의해서 $f_Y(\sqrt{x})=f_Y(-\sqrt{x})$ $=e^{-\frac{1}{2}x}/\sqrt{2\pi}$ 이므로 $f_X(x)=e^{-\frac{1}{2}x}/\sqrt{2\pi x}$ 는 자유도가 1인 카이제곱 밀도함수가 된다 (식 (2.5.3) 참조. 비고: $\Gamma(1/2)=\sqrt{\pi}$.)

2.9.3 대수 정규분포

Y 가 정규분포를 따를 때, $X = g(Y) = e^{Y}$ 의 분포를 대수 정규(*lognormal*)분포라 한다. (역으로, X 가 대수 정규분포를 따르면, $Y = g^{-1}(X) = \ln X$ 는 정규분포를 따른다.) 식 (2.9.2)에 $y = g^{-1}(x) = \ln x$, $x \ge 0$, 과 $dy/dx = x^{-1}$ 을 대입하면 식 (2.5.1)에 의해서

$$f_X(x) = \frac{1}{x\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2}, \quad x \ge 0 \tag{2.9.4}$$

를 얻는다. (비고: $f_X(x) = 0$, *if* $x < 0$.)

2.9.4 *Weibull* 분포

Y 가 지수분포를 따를 때, $X = g(Y) = Y^{1/m}$, $m > 0$, 의 분포를 *Weibull* 분포라 한다. (단, $f_Y(y) = \lambda e^{-\lambda y}$ 인 $y \ge 0$ 에 대해서만 따짐.) 식 (2.9.2)에 $y = g^{-1}(x) = x^{m}$, $x \ge 0$, 과 $dy/dx = mx^{m-1}$ 을 대입하면

$$f_X(x) = \lambda m x^{m-1} e^{-\lambda x^{m}}, \quad x \ge 0 \tag{2.9.5}$$

을 얻는다. (비고: $f_X(x) = 0$, if $x < 0$.)

Weibull 분포의 평균과 분산은 비교적 간단한 형태인데, 각각 $E(X) = \Gamma(1 + m^{-1})/\lambda$ 와 $V(X) = \Gamma(1 + 2m^{-1})/\lambda^{2}$ 이다.

<비고 2.9.2> Y_1 , Y_2 가 독립이고 Y_i 가 (평균이 λ_i^{-1} 인) 지수분포를 따르면 $\min(Y_1, Y_2)$은 (평균이 $(\lambda_1 + \lambda_2)^{-1}$ 인)지수분포를 따른다. 이러한 속성은 *Weibull* 분포로 확장되는데, 단 독립인 X_1 , X_2 가 동일한 m 값을 가질 때에 한해서 유효하다 (비고: $m = 1$ 이면 지수분포).

§2.10 수명분포

기계의 수명, 즉 새 기계가 고장날 때까지의 시간을 Y 라 하자. 신뢰성 공학에서는 Y 의 분포로 *Weibull*, 대수 정규분포, 감마분포 등을 사용한다. 이들의 공통점은 물론 $P(Y<0)=0$ 이다. 이 중에서 특히 *Weibull* 분포가 애용되는 이유는 다음과 같다.

첫째로, 기계의 고장은 가장 취약한 부분(또는 부품)에서 발생한다. 그런데, 부품들의 수명이 독립이고 (동일한 m 값을 가지는) *Weibull* 분포를 따르면, 기계의 수명도 Weibull 분포를 따르게 된다 (<비고 2.9.2> 참조).

둘째로, 신뢰성 공학에서는 $P(Y>y)$ 를 신뢰도(reliability)라 하고, $r(y)=f(y)/P(Y>y)$ 를 FR(failure rate)라 하는데, *Weibull* 분포의 경우 이들의 형태가 간단하다. 즉, $f(y)=\lambda m y^{m-1}e^{-\lambda y^m}$ 에서 $(y\geq 0)$, $r(y)=\lambda m^{y-1}$ 이고 $P(Y>y)=e^{-\lambda y^m}$ 이다.

FR 의 의미는 다음과 같다.

$$r(y)dy=P(y<Y<y+dy\mid Y>y) \tag{2.10.1}$$

즉, 이미 y 시간 동안 고장나지 않고 작동중인 기계가 이후 dy 동안에 고장날 확률이 $r(y)dy$ 이다. 예를 들어, *Weibull* 분포에서 $m=1$ 이면 지수분포가 되는데, $r(y)dy=\lambda m y^{m-1}dy$ 에 $m=1$ 을 대입하면 λdy 를 얻는다 (§2.4.1 참조). 즉, 지수분포는 FR가 λ 로 일정하다. 반면에, $m>1$ 이면 y 가 증가함에 따라 FR가 증가하고, $0<m<1$ 이면 y 가 증가함에 따라 FR 가 감소한다.

수명 분포와 관련하여 간혹 FR 와 혼동이 되는 것은 절단(truncated) 분포이다. 이미 c 시간 동안 고장나지 않고 작동중인 기계가 앞으로 언제 고장이 날 것인가를 알기 위해서는

$$P(y<Y<y+dy\mid Y>c)=\begin{cases}\dfrac{f(y)dy}{P(Y>c)} & ,\ y\geq c\\[2mm] 0 & ,\ y<c\end{cases} \tag{2.10.2}$$

를 사용하는데, 이는 Y 의 분포에서 $y<c$ 부분은 잘라 버리고 나서 (남은 부분의 확률 합이 1 이 되도록) 남은 부분을 $\{P(Y>c)\}^{-1}$ 배로 확장시킨 것이다. 식 (2.10.2)도 식 (2.10.1) 같이 조건부 확률이고 y 의 함수인데, 차이점은 식 (2.10.2)에서는 조건 " $Y>c$" 가 y 와 무관한 반면에 식 (2.10.1)에서는 조건 " $Y>y$"가 y 의 함수라는 점이다.

§2.11 결합분포

2.11.1 결합분포의 정의

확률변수 $Y_1, Y_2, \cdots, Y_n$ 의 결합(joint) CDF의 정의는 다음과 같다.

$$F(y_1, y_2, \cdots, y_n) = P(Y_1 \le y_1, Y_2 \le y_2, \cdots, Y_n \le y_n)$$

$Y_1, \cdots, Y_n$ 이 이산 확률변수인 경우에는 결합 질량함수 $P(Y_1 = y_1, \cdots, Y_n = y_n)$ 을 그리고 연속 확률변수인 경우에는 결합 밀도함수 $f(y_1, \cdots, y_n)$ 을 사용하면 편리한데, 결합 밀도함수의 의미는 다음과 같다.

$$f(y_1, \cdots, y_n)\, dy_1 \cdots dy_n = P(y_1 < Y_1 < y_1 + dy_1, \cdots, y_n < Y_n < y_n + dy_n)$$

결합분포는 지금까지 세 번 등장했는데, 모두 이산 확률변수인 경우이다. 첫째는 우도함수 (<비고 1.9>, §1.6 참조)이고, 둘째는 다항분포 (§2.1.6) 그리고 셋째는 다변량 초기하분포 (§2.2.3)이다.

2.11.2 독립 속성

결합분포와 관련된 가장 중요한 속성인 독립의 정의는

$$F(y_1, \cdots, y_n) = \prod_{i=1}^{n} P(Y_i \le y_i) = \prod_{i=1}^{n} F_{Y_i}(y_i)$$

인데, 이를 (이산 경우의) 질량함수와 (연속 경우의) 밀도함수로 표현하면 다음과 같다.

$$P(Y_1 = y_1, \cdots, Y_n = y_n) = \prod_{i=1}^{n} P(Y_i = y_i) \tag{2.11.1}$$

$$f(y_1, \cdots, y_n)\, dy_1 \cdots dy_n = \prod_{i=1}^{n} P(y_i < Y_i < y_i + dy_i) = \prod_{i=1}^{n} f_{Y_i}(y_i)\, dy_i \tag{2.11.2}$$

독립의 의미는 다음과 같다. 예를 들어 Y_1 과 Y_2 가 이산 확률변수이면 곱의 법칙에 의해

$$P(Y_1 = y_1, Y_2 = y_2) = P(Y_1 = y_1) P(Y_2 = y_2 \mid Y_1 = y_1)$$

인데, Y_1 과 Y_2 가 독립이면 $P(Y_2 = y_2 \mid Y_1 = y_1) = P(Y_2 = y_2)$ 이다. 즉, “ $Y_1 = y_1$ ” 이라

고 하는 주어진 정보가 " $Y_2 = y_2$ " 일 확률에 아무런 영향을 미치지 못한다는 뜻이다.

2.11.3 결합분포와 기대치

$Y_1, \cdots, Y_n$ 의 함수를 $X = g(Y_1, \cdots, Y_n)$ 이라 하면 X 의 기대치 $E(X)$ 를 구하는 방법은 두 가지이다. 첫째는 X 의 CDF $F_X(x) = P(X \le x)$ 를 구한 다음 $E(X) = \int_{-\infty}^{\infty} x\, dF_X(x)$ 를 이용하는 방법인데, $F_X(x)$ 를 구하는 방법은 다음 절에서 다룬다. 둘째는 $Y_1, \cdots, Y_n$ 의 결합분포를 사용하는 방법인데, 이산 확률변수인 경우에는

$$E(X) = E[g(Y_1, \cdots, Y_n)] = \sum_{y_1} \cdots \sum_{y_n} g(y_1, \cdots, y_n) P(Y_1 = y_1, \cdots, Y_n = y_n) \qquad (2.11.3)$$

이고, 연속 확률변수인 경우에는 다음과 같다 (증명은 생략).

$$E(X) = E[g(Y_1, \cdots, Y_n)] = \int_{y_1} \cdots \int_{y_n} g(y_1, \cdots, y_n) f(y_1, \cdots, y_n)\, dy_1 \cdots dy_n \qquad (2.11.4)$$

편의상, 이산 확률변수 Y_1, Y_2 를 고려한다. 먼저, $X = g(Y_1, Y_2) = g_1(Y_1) + g_2(Y_2)$ 인 경우에는 다음이 성립한다.

$$\begin{aligned} E(X) &= E[g_1(Y_1) + g_2(Y_2)] \\ &= \sum_{y_1} g_1(y_1) \sum_{y_2} P(Y_1 = y_1, Y_2 = y_2) + \sum_{y_2} g_2(y_2) \sum_{y_1} P(Y_1 = y_1, Y_2 = y_2) \\ &= \sum_{y_1} g_1(y_1) P(Y_1 = y_1) + \sum_{y_2} g_2(y_2) P(Y_2 = y_2) \\ &= E[g_1(Y_1)] + E[g_2(Y_2)] \end{aligned} \qquad (2.11.5)$$

<비고 2.11.1> $E[g_1(Y_1) + g_2(Y_2)] = E[g_1(Y_1)] + E[g_2(Y_2)]$ 는 Y_1, Y_2 가 독립이 아니더라도 성립함 (<비고 2.8.6>, <비고 2.11.2> 참조).

다음, $X = g(Y_1, Y_2) = g_1(Y_1) \cdot g_2(Y_2)$ 인 경우에는 $E(X) = E[g_1(Y_1) g(Y_2)] = \sum_{y_1} \sum_{y_2} g_1(y_1) g_2(y_2) P(Y_1 = y_1, Y_2 = y_2)$ 인데, 만약 Y_1 과 Y_2 가 독립이면 $P(Y_1 = y_1, Y_2 = y_2) = P(Y_1 = y_1) \cdot P(Y_2 = y_2)$ 이므로 다음이 성립한다.

$$E[g_1(Y_1)g_2(Y_2)] = \sum_{y_1} g_1(y_1)P(Y_1=y_1)\sum_{y_2} g_2(y_2)P(Y_2=y_2)$$
$$= E[g_1(Y_1)] \cdot E[g_2(Y_2)]$$

<비고 2.11.2> Y_1 과 Y_2 가 독립이면 $g_1(Y_1)$ 과 $g_2(Y_2)$ 도 독립이다.

2.11.4 $g(Y_1,\cdots,Y_n)$ 의 분포

$Y_1,\cdots,Y_n$ 의 함수 $X=g(Y_1,\cdots,Y_n)$ 의 CDF $F_X(x)=P(X\le x)$ 는 다음과 같이 구한다.

$$F_X(x)=P[g(Y_1,\cdots,Y_n)\le x] = \begin{cases} \displaystyle\sum_{g(y_1,\cdots,y_n)\le x}\cdots\sum P(Y_1=y_1,\cdots,Y_n=y_n) \\ \displaystyle\int\cdots\int_{g(y_1,\cdots,y_n)\le x} f(y_1,\cdots,y_n)\,dy_1\cdots dy_n \end{cases}$$

즉, $g(y_1,\cdots,y_n)\le x$를 만족시키는 $y_1,\cdots,y_n$에 대해서, 결합 질량함수를 (또는 결합 밀도함수를) 모두 더한 (또는 적분한) 것이다.

또한, 식 (2.9.2)를 확장한 공식은 다음과 같다.

$$f_{X_1,\cdots,X_n}(x_1,\cdots,x_n)=f_{Y_1,\cdots,Y_n}(y_1,\cdots,y_n)\cdot |J| \qquad (2.11.7)$$

예를 들어. $n=2$ 인 경우, $X_1=g_1(Y_1,Y_2)$, $X_2=g_2(Y_1,Y_2)$ 일 때 편의상 $x_1=g_1(y_1,y_2)$, $x_2=g_2(y_1,y_2)$ 라 하면, 함수 $g_1(\cdot)$ 와 $g_2(\cdot)$ 에 의해서 (x_1,x_2) 가 (y_1,y_2) 에 1:1 로 대응되는 경우에 한해서 식 (2.11.7)이 유효한데, 이때 X_1, X_2 의 결합 밀도함수인 $f_{X_1,X_2}(x_1,x_2)$ 와 Y_1, Y_2 의 결합 밀도함수인 $f_{Y_1,Y_2}(y_1,y_2)$ 의 비율인 $|J|$ 는 다음과 같다. (J는 Jacobian을 의미하며, x_1, x_2 의 함수임.)

$$|J| = \left| det\begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{pmatrix} \right| = \left| \frac{\partial y_1}{\partial x_1}\frac{\partial y_2}{\partial x_2} - \frac{\partial y_1}{\partial x_2}\frac{\partial y_2}{\partial x_1} \right| \qquad (2.11.8)$$

연습문제 삼아 t 분포인 식 (2.5.6)을 유도한다. $X_1=T_d$, $Y_1=Z$, $Y_2=C_d$ 라 하면, 식 (2.5.5)에 의해서 $X_1=g_1(Y_1,Y_2)=Y_1/\sqrt{Y_2/d}$ 이다. 또한 Z와 C_d는 독립이므로,

$f_{Y_1,Y_2}(y_1,y_2)=f_{Y_1}(y_1)f_{Y_2}(y_2)$ 이다. 구하려는 것은 $f_{X_1}(x_1)$ 인데, 이는 식 (2.11.7)로 $f_{X_1,X_2}(x_1,x_2)$ 를 얻은 다음에 $f_{X_1}(x_1)=\int_0^\infty f_{X_1,X_2}(x_1,x_2)dx_2$ 로 구한다. 그러니까 부득이 $X_2=g_2(Y_1,Y_2)$ 를 정의해야 되는데, 예를 들어 $X_2=Y_2$ 라 하면 계산이 간단하다. 즉, $x_1=y_1/\sqrt{y_2/d}$ 와 $x_2=y_2$ 로부터 $y_1=x_1\sqrt{x_2/d}$ 와 $y_2=x_2$ 를 얻으므로 식 (2.11.8)은 간단하게 $|J|=|\partial y_1/\partial x_1|=|\sqrt{x_2/d}|=\sqrt{x_2/d}$가 된다. 따라서, 식 (2.11.7)의 우변은 $f_{Y_1}(y_1)f_{Y_2}(y_2)\sqrt{x_2/d}$ 인데, 식 (2.5.2)에 의해서 $f_{Y_1}(y_1)=e^{-y_1^2/2}/\sqrt{2\pi}$ 이고, 식 (2.5.3)에 의해서 $f_{Y_2}(y_2)=(y_2/2)^{d/2-1}\ e^{-y_2/2}/2\Gamma(d/2)$이다. 마지막으로, $y_1=x_1\sqrt{x_2/d}$ 와 $y_2=x_2$ 를 대입하면 식 (2.11.7)의 좌우변이 모두 x_1 과 x_2 로 표현된다. (비고: $-\infty<x_1<\infty, x_2>0$.) 이를 x_2 에 대해서 (0 에서 ∞ 까지) 적분하면 식 (2.5.6)을 얻는데, 적분과정은 복잡하므로 생략한다.

2.11.5 순서 통계량 (Order Statistics)

$Y_1,\cdots,Y_n$의 함수인 $Y_{(n)}\equiv\max(Y_1,\cdots,Y_n)$과 $Y_{(1)}\equiv\min(Y_1,\cdots,Y_n)$의 분포를 구하되, $Y_1,\cdots,Y_n$이 *iid* 연속 확률변수인 경우를 다룬다. 먼저, $Y_{(n)}$의 CDF는 다음과 같다.

$$F_{(n)}(y)\equiv P(Y_{(n)}\le y)=P(Y_1\le y,\cdots,Y_n\le y)$$

즉, “$Y_{(n)}\le y$”라는 사건(event)은 “$Y_1\le y,\cdots,Y_n\le y$”라는 사건과 동일하다. 그런데, $Y_1,\cdots,Y_n$ 은 서로 독립이므로 $P(Y_1\le y,\cdots,Y_n\le y)=\prod_{i=1}^n P(Y_i\le y)$ 이고, 또한 동일한 분포를 따르므로 $\prod_{i=1}^n P(Y_i\le y)=\{P(Y\le y)\}^n\equiv F(y)^n$이 된다 (<비고 1.4.1> 참조).

$Y_{(n)}$의 밀도함수는 $f_{(n)}(y)\equiv\frac{d}{dy}F_{(n)}(y)=nF(y)^{n-1}f(y)$ 인데, 이에 대한 해석은 다음과 같다.

$$
\begin{aligned}
f_{(n)}(y)\,dy &= P(y < Y_{(n)} < y+dy) \\
&= P[\,Y_1,\cdots,Y_n \text{ 중에서 하나는 } (y, y+dy),\text{ 나머지는 모두 } y \text{ 이하}\,] \\
&= \binom{n}{1} \cdot f(y)dy \cdot \{P(Y \le y)\}^{n-1}
\end{aligned}
$$

이 식에서 $\binom{n}{1}$은 n개의 iid 확률변수 중에서 하나를 뽑는 경우의 수이고, $f(y)dy$는 뽑힌 것의 값이 y와 $y+dy$ 사이의 값을 가질 확률이며, $\{P(Y\le y)\}^{n-1}$은 나머지 $(n-1)$개가 모두 y이하의 값을 가질 확률이다.

다음, $Y_{(k)}$의 밀도함수는 다음과 같이 정의한다, $k=1,\cdots,n$.

$$
\begin{aligned}
f_{(k)}(y)\,dy &\equiv P[\,Y_1,\cdots,Y_n \text{ 중에서 } k\text{번째로 작은 것의 값이 } (y, y+dy)\,] \\
&= P[\,Y_1,\cdots,Y_n \text{ 중에서 } (k-1)\text{개는 } y \text{ 이하, 나머지 중에서} \\
&\qquad \text{하나는 } (y, y+dy),\text{ 남은 } (n-k)\text{개는 모두 } y\text{보다 큼}\,] \\
&= \binom{n}{k-1}\{P(Y\le y)\}^{k-1} \cdot \binom{n-k+1}{1} f(y)\,dy \cdot \{P(Y>y)\}^{n-k} \\
&= \frac{n!}{(k-1)!\,(n-k)!}F(y)^{k-1}\{1-F(y)\}^{n-k}f(y)dy \qquad (2.11.9)
\end{aligned}
$$

식 (2.11.9)에 $k=1$을 대입하면 $Y_{(1)}$의 밀도함수로 $f_{(1)}(y)=n\{1-F(y)\}^{n-1}f(y)$를 얻는다. 예를 들어, iid 확률변수 $Y_1,\cdots,Y_n$이 평균이 λ^{-1}인 지수분포를 따르면 $f(y)=\lambda e^{-\lambda y}$, $F(y)=1-e^{-\lambda y}$이므로 $f_{(1)}(y)=n\lambda e^{-n\lambda y}$를 얻는데, 이는 평균이 $(n\lambda)^{-1}$인 지수분포이다 (<비고 2.9.2> 참조).

또 다른 예로, iid 확률변수 $Y_1,\cdots,Y_n$이 $(0,1)$ 구간에서 *uniform* 분포를 따르면 $f(y)=1$, $F(y)=y$인데 (단, $0<y<1$), 이를 식 (2.11.9)에 대입하면 표준 베타분포를 얻는다 (비고: 식 (2.4.6)에서 $n_1=k$, $n_2=n-k+1$ 에 해당).

순서 통계량 $Y_{(1)},\cdots,Y_{(n)}$ 중의 일부분 또는 전체의 결합밀도함수를 구하는 방법도 위와 동일한데, 복잡하므로 생략한다 (비고: $n!$은 항상 등장함).

§2.12 MGF

2.12.1 MGF의 정의

적분보다 미분이 쉽다. 기대치의 정의는 적분인데 (비고: $E[g(y)] = \int_{-\infty}^{\infty} g(y)\,dF(y)$), 적분 대신에 미분으로 기대치를 구하기 위해서 MGF(moment generating function)를 다음과 같이 정의한다. 확률변수 Y의 MGF는 $E(e^{\Theta Y})$이다. 즉, MGF는 Y의 함수인 $e^{\Theta Y}$의 기대치이다. $E(e^{\Theta Y})$를 Θ에 대해서 n번 미분하면 ($n = 1, 2, \cdots$)

$$\frac{d^n}{d\Theta^n} E(e^{\Theta Y}) = \frac{d^n}{d\Theta^n} \int_{-\infty}^{\infty} e^{\Theta y} dF(y) = \int_{-\infty}^{\infty} y^n e^{\Theta y} dF(y)$$

를 얻는데, 이에 $\Theta = 0$ 을 대입하면 $E(Y^n)$이 된다. (비고: $E(Y^n)$을 Y의 n^{th} moment라 함.)

2.12.2 기하분포의 평균과 분산

기하분포 $P(Y = y) = q^{y-1}p$, $y = 1, 2, \cdots$,를 따르는 Y의 MGF는

$$\begin{aligned} E(e^{\Theta y}) &= \int_{-\infty}^{\infty} e^{\Theta y} dF(y) = \sum_{y=1}^{\infty} e^{\Theta y} q^{y-1} p \\ &= pe^{\Theta} \sum_{y=1}^{\infty} (qe^{\Theta})^{y-1} = pe^{\Theta}/(1 - qe^{\Theta}) \end{aligned} \tag{2.12.1}$$

인데, 이를 미분해서 $E(Y) = 1/p$과 $E(Y^2) = (1+q)/p^2$를 얻고 또한 $V(Y) = E(Y^2) - E(Y)^2 = q/p^2$ 를 얻는다.

2.12.3 지수분포의 평균과 분산

지수분포 $f(y) = \lambda e^{-\lambda y}, y \geq 0,$을 따르는 Y의 MGF는

$$E(e^{\Theta y}) = \int_{-\infty}^{\infty} e^{\Theta y} f(y)\,dy = \int_0^{\infty} e^{\Theta y} \lambda e^{-\lambda y} dy = \frac{\lambda}{\lambda - \Theta} \tag{2.12.2}$$

인데, 이를 미분해서 $E(Y) = 1/\lambda$ 과 $E(Y^2) = 2/\lambda^2$를 얻고 또한 $V(Y) = E(Y^2) - E(Y)^2$

$=1/\lambda^2$ 을 얻는다 (<비고 2.8.9> 참조).

2.12.4 Convolution

Y_1, Y_2 가 독립인 확률변수일 때, Y_1+Y_2 의 CDF를 Y_1 의 CDF와 Y_2 의 CDF의 (협의의) convolution이라 한다. 반면에, 광의의 convolution은 Y_1+Y_2 에 관련된 모든 것을 의미한다.

Convolution은 MGF로 표현하면 간단하다. Y_1+Y_2 의 MGF는 $E[e^{\Theta(Y_1+Y_2)}]$ $=E(e^{\Theta Y_1}e^{\Theta Y_2})$ 인데, 식 (2.11.6)에 의해서 $E(e^{\Theta Y_1}e^{\Theta Y_2})=E(e^{\Theta Y_1})E(e^{\Theta Y_2})$ 가 된다. 즉, Y_1 과 Y_2 가 독립이면 Y_1+Y_2 의 MGF는 Y_1 의 MGF와 Y_2 의 MGF의 곱이 된다.

나아가서, $Y_1, \cdots, Y_n$ 이 *iid* 확률변수이면 $\sum_{i=1}^{n} Y_i$ 의 MGF는

$$E(e^{\Theta \sum_{i=1}^{n} Y_i})=\prod_{i=1}^{n} E(e^{\Theta Y_i})=\{E(e^{\Theta Y})\}^n \qquad (2.12.3)$$

이 된다. 예를 들어, 식 (2.12.1)로부터 음이항 확률변수의 MGF인 $\{pe^{\Theta}/(1-qe^{\Theta})\}^n$ 을 얻고, 식 (2.12.2)로부터는 *Erlang* 확률변수의 MGF인 $\{\lambda/(\lambda-\Theta)\}^n$ 을 얻는다.

또한, *Bernoulli* 확률변수 Y 의 MGF는 $E(e^{\Theta Y})=\sum_{y=0}^{1} e^{\Theta y}P(Y=y)=q+pe^{\Theta}$ 이므로, 이항 확률변수의 MGF는 $(q+pe^{\Theta})^n$ 이 된다.

2.12.5 기타 MGF

이항분포로부터 포아송 분포를 얻듯이, 이항 확률변수의 MGF인 $(q+pe^{\Theta})^n$ 에 $q=1-p$, $p=\lambda t/n$ 을 대입한 $\{1+\lambda t(e^{\Theta}-1)/n\}^n$ 에 $n\to\infty$ 를 취하면 포아송 확률변수의 MGF로 $e^{\lambda t(e^{\Theta}-1)}$ 을 얻는다.

다음, *Erlang* 분포에서 자연수 n 을 양의 실수로 간주하면 감마 분포가 되듯이, *Erlang* 확률변수의 MGF인 $\{\lambda/(\lambda-\Theta)\}^n$ 에서 n 을 양의 실수로 간주하면 감마 확률변수의 MGF가 된다. 또한, 카이제곱 분포는 $\lambda=1/2$, $n=d/2$ 인 감마분포이므로, 카이제곱 확률변수의 MGF는 $(1-2\Theta)^{-d/2}$ 이다.

마지막으로, 정규 확률변수의 MGF는 $e^{\mu\Theta+\sigma^2\Theta^2/2}$ 이다 (증명생략).

2.12.6 MGF의 기타 용도

확률분포와 MGF 간에 성립하는 1:1 대응관계를 이용해서 새로 정의된 확률변수의 분포를 파악할 수 있다. 특히, 서로 독립인 확률변수들의 합의 분포를 파악하는데에 유용하다. 예를 들어, C_{d_1}과 C_{d_2}가 서로 독립인 카이제곱 확률변수이면 (각각의 자유도는 d_1과 d_2), 각각의 MGF는 $(1-2\Theta)^{-d_1/2}$와 $(1-2\Theta)^{-d_2/2}$이므로 $C_{d_1}+C_{d_2}$의 MGF는 $(1-2\Theta)^{-(d_1+d_2)/2}$인데 (§2.14.4 참조), 이는 (자유도가 d_1+d_2인) 카이제곱 분포를 따른다. 또 다른 예로, 중심극한정리(§2.5 참조)의 증명에도 MGF를 이용한다 (문헌 [9] 참조).

또한, 예를 들어, 식 (2.8.8)과 (2.8.9)를 MGF로 증명할 수 있다. *iid* 확률변수인 $Y_1, \cdots, Y_n$의 합을 S라 하면 식 (2.12.1)에 의해서 $E(e^{\Theta S})=\{E(e^{\Theta Y})\}^n$ 이므로

$$E(S)=\frac{d}{d\Theta}E(e^{\Theta S})|_{\Theta=0}=\frac{d}{d\Theta}\{E(e^{\Theta Y})\}^n|_{\Theta=0}=nE(Y)$$

$$E(S^2)=\frac{d^2}{d\Theta^2}E(e^{\Theta S})|_{\Theta=0}=\frac{d^2}{d\Theta^2}\{E(e^{\Theta Y})\}^n|_{\Theta=0}=n(n-1)E(Y)^2+nE(Y^2)$$

$$V(S^2)=E(S^2)-E(S)^2=n\{E(Y^2)-E(Y)^2\}=nV(Y)$$

를 얻는다.

공학에서는 MGF 대신에 PGF(probability generating function)와 LT(Laplace transform)을 애용하는데, 이들을 기대치 형태로 표현하면 각각 $E(z^Y)$와 $E(e^{-\Phi Y})$이다. 즉, MGF $E(e^{\Theta Y})$에서 e^{Θ}를 각각 z와 $e^{-\Phi}$로 교체한 것이다. PGF와 LT의 용도 역시 MGF와 유사한데, 추가적으로 PGF를 미분해서 확률을 발생시킬 수 있으며 LT는 미분방정식을 푸는데 쓰이기도 한다.

§2.13 공분산과 상관계수

2.13.1 공분산의 정의

확률변수 Y_1과 Y_2간의 공분산(covariance) $Cov(Y_1, Y_2)$의 정의는

$$Cov(Y_1, Y_2) = E[(Y_1 - \mu_1)(Y_2 - \mu_2)] \tag{2.13.1}$$

인데, μ_1과 μ_2는 각각 $E(Y_1)$과 $E(Y_2)$이다.

<비고 2.13.1> 식 (2.13.1)에 $Y_1 = Y_2 = Y$를 대입하면 분산 $V(Y) = E[(Y - \mu)^2]$이 된다. 즉, 분산은 공분산의 특수한 경우이다.

분산을 구할 때 " $V(Y) = E(Y^2) - \mu^2$ "이 편리하듯이, 공분산을 구할 때는 다음 관계를 사용하면 편리하다 (증명 생략).

$$Cov(Y_1, Y_2) = E(Y_1 Y_2) - \mu_1 \mu_2 \tag{2.13.2}$$

공분산은 다음과 같이 $V(Y_1 + Y_2)$에 등장한다.

$$\begin{aligned} V(Y_1 + Y_2) &= E[(Y_1 + Y_2)^2] - \{E(Y_1 + Y_2)\}^2 \\ &= E(Y_1^2 + 2Y_1 Y_2 + Y_2^2) - (\mu_1 + \mu_2)^2 \\ &= E(Y_1^2) + 2E(Y_1 Y_2) + E(Y_2^2) - (\mu_1^2 + 2\mu_1\mu_2 + \mu_2^2) \\ &= V(Y_1) + V(Y_2) + 2 \cdot Cov(Y_1, Y_2) \end{aligned} \tag{2.13.3}$$

좀더 일반적인 형태는

$$V(aY_1 + bY_2) = a^2 V(Y_1) + b^2 V(Y_2) + 2abCov(Y_1, Y_2) \tag{2.13.4}$$

인데, 예를 들어 $V(Y_1 - Y_2)$는 다음과 같다.

$$V(Y_1 - Y_2) = V(Y_1) + V(Y_2) - 2 \cdot Cov(Y_1, Y_2) \tag{2.13.5}$$

<비고 2.13.2> Y_1과 Y_2가 독립이면 $Cov(Y_1, Y_2) = 0$ 이다.

Y_1과 Y_2가 독립이면 식 (2.11.6)에 의해서 "$E(Y_1 Y_2) = E(Y_1)E(Y_2)$"이므로, 이를 식 (2.13.2)에 대입하면 "$Cov(Y_1, Y_2)=0$"을 얻는다. 그리고, "$Cov(Y_1, Y_2) = 0$"을 식 (2.13.3), (2.13.4), (2.13.5)에 대입하면 "$V(Y_1 \pm Y_2) = V(Y_1) + V(Y_2)$"와 "$V(aY_1 + bY_2) = a^2 V(Y_1) + b^2 V(Y_2)$"를 얻는다.

2.13.2 $2Y_1$과 $Y_1 + Y_2$

분산과 공분산의 관계를 잘 이해하기 위하여 간단한 예를 든다. Y_1과 Y_2가 동일한 분포 "$P(Y_i=1) = P(Y_i=2) = P(Y_i=3) = 1/3$"을 따르면 $E(Y_i) = (1+2+3)/3 = 2$, $V(Y_i) = \{(-1)^2 + 0^2 + 1^2\}/3 = 2/3$, $i = 1, 2$, 이다.

$X_1 = 2Y_1$, $X_2 = Y_1 + Y_2$라 하자. 그러면, X_1의 분포는 $P(X_1=2) = P(X_1=4) = P(X_1=6) = 1/3$ 이고 (§2.9.1 참조), $E(X_1) = (2+4+6)/3 = 4$, $V(X_1) = \{(-2)^2 + 0^2 + 2^2\}/3 = 8/3$ 이다.

반면에, $X_2 = Y_1 + Y_2$의 분포는

$$P(X_2=2) = P(Y_1=1, Y_2=1)$$
$$P(X_2=3) = P(Y_1=1, Y_2=2) + P(Y_1=2, Y_2=1)$$
$$P(X_2=4) = P(Y_1=1, Y_2=3) + P(Y_1=2, Y_2=2) + P(Y_1=3, Y_2=1)$$
$$P(X_2=5) = P(Y_1=2, Y_2=3) + P(Y_1=3, Y_2=2)$$
$$P(X_2=6) = P(Y_1=3, Y_2=3)$$

이므로 결합분포 $P(Y_1=y_1, Y_2=y_2)$를 알아야만 $X_2 = Y_1 + Y_2$의 분포를 구할 수 있다. 그러나, 만약 Y_1과 Y_2가 독립이면 $P(Y_1=y_1, Y_2=y_2) = P(Y_1=y_1)\,P(Y_2=y_2)$이므로

$$P(X_2=2) = P(Y_1=1)P(Y_2=1) = (1/3)^2 = 1/9$$
$$P(X_2=3) = (1/3)^2 + (1/3)^2 = 2/9$$
$$P(X_2=4) = (1/3)^2 + (1/3)^2 + (1/3)^2 = 3/9$$
$$P(X_2=5) = (1/3)^2 + (1/3)^2 = 2/9$$

$$P(X_2=6)=P(Y_1=3)P(Y_2=3)=(1/3)^2=1/9$$

이고, $E(X_2)=E(Y_1)+E(Y_2)=4$, $V(X_2)=V(Y_1)+V(Y_2)=4/3$ 이다.

일반적으로, Y_1과 Y_2가 동일한 분포를 따를 때 평균을 μ 분산을 σ^2이라 하면, $E(2Y_1)$과 $E(Y_1+Y_2)$는 모두 2μ이다. 그러나, $V(2Y_1)$는 $(2\sigma)^2=4\sigma^2$인 반면에 $V(Y_1+Y_2)$는 Y_1과 Y_2가 독립이면 $2\sigma^2$이다.

사실, $V(2Y_1)$은 $V(Y_1+Y_2)$의 특수한 경우이다. 식 (2.13.3)에 $Y_2=Y_1$을 대입하면 $V(2Y_1)=2\cdot V(Y_1)+2\cdot Cov(Y_1,Y_1)$을 얻는데, 이는 <비고 2.20>에 의해서 $Cov(Y_1,Y_1)=V(Y_1)$이므로 결국 $V(2Y_1)=4V(Y_1)$을 얻는다.

2.13.3 공분산의 의미

Covariance는 “co”와 “variance”의 복합어인데, “co”는 “같이” 또는 “함께”를 의미한다. 식 (2.13.1)에서, $Y_1>\mu_1$일 때 $Y_2>\mu_2$일 확률이 크거나 또는 $Y_1<\mu_1$일 때 $Y_2<\mu_2$일 확률이 클수록 $Cov(Y_1,Y_2)$가 커진다. 반대로, $Y_1>\mu_1$일 때 $Y_2<\mu_2$일 확률이 크거나 또는 $Y_1<\mu_1$일 때 $Y_2>\mu_2$일 확률이 클수록 $Cov(Y_1,Y_2)$는 작아진다.

§2.13.2의 예를 계속한다. 퇴직금으로 받은 3억원을 A와 B 두 군데에 각각 1억 5천씩 투자하면 1년 후에 각각 Y_1, Y_2억원이 된다고 하자. 즉, 1년 후의 총 회수금은 $X_2=Y_1+Y_2$ 억원이다. $E(X_2)=E(Y_1)+E(Y_2)=2+2=4$ 억원이므로 기대수익율(expected rate of return)은 33.3%이다. 그런데, 기대수익율이 높은 만큼 위험도 따르는데, 예를 들어 $P(X_2=2)=P(Y_1=Y_2=1)$의 확률도 1억원을 손해본다.

흔히 $V(X_2)$를 위험도(risk)의 척도로 사용한다. $V(X_2)$가 최대인 경우는 $P(Y_1=Y_2=y)=1/3$, $y=1,2,3$인 경우인데, 이때 $Cov(Y_1,Y_2)=V(Y_1)$이고 $V(X_2)=4V(Y_1)$이 된다. 이는 마치 A 또는 B 한 군데에 3억원을 모두 투자하는 것과 같은 결과이다. 반대로, $V(X_2)$가 최소가 되는 경우는 $P(Y_1=1,Y_2=3)=P(Y_1=Y_2=2)=P(Y_1=3,Y_2=1)=\frac{1}{3}$인 경우인데, 이때 $Cov(Y_1,Y_2)=-V(Y_1)$이고 $V(X_2)=0$이 된다. 이는 A와 B의 위험요인이 서로 반대방향으로 작용하여 완전히 상쇄시키는 경우이다. 일반적으로 $V(X_2)$는 두 극단인 $4V(Y_1)$과 0사이의 값을 가진다. 예를 들어, Y_1과 Y_2가 독립인 경우에는 $V(X_2)=2V(Y_1)$이다.

$V(X_2) > 2V(Y_1)$ 인 경우 Y_1과 Y_2는 양의 상관관계에 있다고 하고, $V(X_2) < 2V(Y_1)$ 인 경우에는 Y_1과 Y_2가 음의 상관관계에 있다고 한다. 그러니까, 투자를 할 때에는 음의 상관관계에 있는 곳에 나누어서 투자하는 것이 가장 바람직하다. 그리고, 어떠한 경우더라도 한 군데에 모두 투자하는 것에 비해서 여러 곳에 나누어 투자하는 것이 못하지는 않다.

2.13.4 상관계수

앞 절에서는 비교의 편의상 $V(Y_1) = V(Y_2)$ 인 경우를 예로 들었다. $V(Y_1) \neq V(Y_2)$ 인 일반적인 경우에 대해서 Y_1과 Y_2의 상관관계를 표준화시킨 것이 상관계수(correlation coefficient)인데, 정의는 다음과 같다.

$$\rho = \frac{Cov(Y_1, Y_2)}{SD(Y_1) \cdot SD(Y_2)}, \quad -1 \le \rho \le 1 \tag{2.13.6}$$

$\rho > 0$ 이면 Y_1, Y_2가 양의 상관관계에 있고, $\rho < 0$ 이면 Y_1, Y_2가 음의 상관관계에 있다. 극단적인 경우는 $\rho = \pm 1$ 인 경우인데, 이는 $Y_1 = aY_2 + b$ (또는 $Y_2 = cY_1 + d$)일 때 발생한다. 즉, Y_1과 Y_2의 관계가 선형(linear)함수관계일때 이들을 선형종속이라 하는데, 이때 $a > 0$ (또는, $c > 0$)이면 $\rho = 1$ 이고 $a < 0$ (또는 $c < 0$)이면 $\rho = -1$ 이다.

한마디로 ρ 는 선형종속성의 척도이다. $|\rho|$가 클수록 선형종속성이 강하고 $\rho = 0$ 이면 선형종속성이 없다.

<비고 2.13.3> $\rho = 0$ 또는 $Cov(Y_1, Y_2) = 0$ 이면 Y_1, Y_2는 선형독립이다. 그러나, 선형독립이더라도 비선형(nonlinear)적으로는 독립이 아닐 수 있다 (<비고 2.13.2> 참조).

<비고 2.13.4> Y_1, Y_2가 정규분포를 따르는 경우에 $\rho = 0$ 또는 $Cov(Y_1, Y_2) = 0$ 이면 Y_1, Y_2는 독립이다.

2.13.5 공분산 공식

공분산에 어느정도 익숙해지고나면 아예 분산도 공분산의 일종으로 간주하면 편리하다

(<비고 2.13.1> 참조). 예를 들어, 식 (2.13.4)를 확장하면

$$V(\sum_{i=1}^{n} a_i Y_i) = Cov(\sum_{i=1}^{n} a_i Y_i, \sum_{j=1}^{n} a_j Y_j) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j Cov(Y_i, Y_j) \qquad (2.13.7)$$

가 되고, 이를 더욱 확장하면 다음을 얻는다.

$$Cov(\sum_{i=1}^{n} a_i Y_i, \sum_{j=1}^{m} b_j X_j) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j Cov(Y_i, X_j) \qquad (2.13.8)$$

식 (2.13.7)은 초기하분포의 분산을 구할 때 쓰이며 (식 (2.8.9) 참조), 식 (2.13.8)은 다항분포(식 (2.1.6) 참조)와 다변량 초기하분포(식 (2.2.2) 참조)에서 $Cov(S_i, S_j)$를 구할 때 쓰인다.

또한, 앞으로 유용하게 쓰일 관계를 식 (2.13.8)로부터 다음과 같이 얻는다. $Cov[(Y_1+Y_2),(Y_1-Y_2)] = Cov(Y_1,Y_1) - Cov(Y_1,Y_2) + Cov(Y_2,Y_1)$ $- Cov(Y_2,Y_2)$인데, $Cov(Y_1,Y_2) = Cov(Y_2,Y_1)$이므로

$$Cov[(Y_1+Y_2),(Y_1-Y_2)] = V(Y_1) - V(Y_2) \qquad (2.13.9)$$

를 얻는다. 따라서, Y_1, Y_2가 동일한 분포를 따르면 ($V(Y_1) = V(Y_2)$이므로), $Cov[(Y_1+Y_2),(Y_1-Y_2)] = 0$ 이다. 그리고, 이를 일반화하면 다음과 같다 (증명생략).

<비고 2.13.5> $Y_1, \cdots, Y_n$이 iid 확률변수일 때, $Cov(\sum_{i=1}^{n} a_i Y_i, \sum_{i=1}^{n} b_i Y_i) = 0$ 이 성립하기 위한 필요충분조건은 $\sum_{i=1}^{n} a_i b_i = 0$ 이다.

<비고 2.13.6> $\sum_{i=1}^{n} a_i Y_i$와 $\sum_{i=1}^{n} b_i Y_i$를 n차원 공간의 벡터라 하면, $\sum_{i=1}^{n} a_i b_i = 0$ 일 때 이들을 직교(orthogonal)관계에 있다고 한다. 이에 따라, 통계학에서는 $Cov(\sum a_i Y_i, \sum b_i Y_i) = 0$ 일 때 $\sum a_i Y_i$와 $\sum b_i Y_i$가 직교관계에 있다고 한다.

§2.14 조건부 기대치

2.14.1 예 #1

§2.7.3의 예를 계속한다. 편의상 S를 다음과 같이 정의하자.

$$S=\begin{cases} 1 \ , & \text{if 스위치를 켰을 때 전구가 터지면} \\ 0 \ , & \text{if 스위치를 켰을 때 불이 들어오면} \end{cases}$$

그러면, 전구의 수명 T의 조건부 기대치는 $E(T|S=1)=0$, $E(T|S=0)=\lambda^{-1}$이다. 그리고, 수명 T의 (무조건: unconditional) 기대치는 다음과 같이 얻는다.

$$\begin{aligned} E(T) &= P(S=1)\cdot E(T|S=1)+P(S=0)\cdot E(T|S=0) \\ &= q\cdot 0+p\cdot\lambda^{-1}=p/\lambda \end{aligned}$$

또한, T의 MGF도 같은 방법으로 구할 수 있다. 즉,

$$E(e^{\Theta T})=P(S=1)\cdot E(e^{\Theta T}|S=1)+P(S=0)\cdot E(e^{\Theta T}|S=0)$$

인데, $S=1$이면 $T=0$이므로 $E(e^{\Theta T}|S=1)=e^{\Theta\cdot 0}=1$ 이고, $S=0$이면 T는 지수분포를 따르므로 $E(e^{\Theta T}|S=0)=\lambda/(\lambda-\Theta)$이다. 따라서 $E(e^{\Theta T})=q+p\lambda/(\lambda-\Theta)$를 얻는데, 물론 이를 미분하면 $E(T)$, $E(T^2)$, $\cdots$ 를 얻을 수 있다.

2.14.2 예 #2

A반의 학생수는 80명이고 이들의 평균성적과 표준편차는 각각 55점과 12점이다. 반면에, B반의 학생수는 20명이고 평균과 표준편차는 각각 80점과 7점이다. 이제, A,B반을 합친 총 100명의 평균성적과 표준편차를 구하려고 한다.

전체의 평균성적은 다음과 같이 구한다.

$$E(Y)=P(A)E(Y|A)+P(B)E(Y|B) \tag{2.14.1}$$

$$=\frac{80}{100}\cdot 55+\frac{20}{100}\cdot 80=60$$

즉, 전체평균 60점은 반별 평균인 55점과 80점의 가중평균인데, 이때 가중치는 0.8과 0.2이다. (또는, A반의 총점 80・55와 B반의 총점 20・80을 합친 전체 총점을 총 학생수 100으로 나눈 것이다.)

그러나, 주의할 점은 전체의 분산은 반별 분산의 가중평균이 아니라는 점이다. 즉,

$$V(Y)\neq P(A)\cdot V(Y|A)+P(B)\cdot V(Y|B) \qquad (2.14.2)$$
$$=(0.8)(12^2)+(0.2)(7^2)=125$$

인데 (예외는 $E(Y|A)=E(Y|B)=E(Y)$ 경우), 그 이유는 다음과 같다. A반의 분산은 "A반의 평균을 중심으로"성적분포가 퍼진 정도를 나타내고, B반의 분산은 "B반의 평균을 중심으로" 성적분포가 퍼진 정도를 나타낸다. 반면에, 전체의 분산은 "전체의 평균을 중심으로" 퍼진 정도를 나타낸다.

식 (2.14.2)의 우변에 추가해야 되는 것은

$$P(A)\cdot\{E(Y|A)-E(Y)\}^2+P(B)\cdot\{E(Y|B)-E(Y)\}^2 \qquad (2.14.3)$$
$$=(0.8)\cdot(55-60)^2+(0.2)(80-60)^2=100$$

인데, 이는 반별 평균들이 전체평균을 중심으로 퍼진 정도를 나타낸다. 따라서, $V(Y)=125+100=225$ 이고 전체 표준편차는 15점이다.

2.14.3 조건부 기대치

이제 조건부 기대치를 정식으로 정의한다. 먼저 조건부 CDF인 $F(y_1|Y_2=y_2)$를 $P(Y_1\leq y_1|Y_2=y_2)$로 정의하면, 조건부 기대치의 정의는 다음과 같다.

$$E(Y_1|Y_2=y_2)=\int_{y_1} y_1\, dF(y_1|Y_2=y_2) \qquad (2.14.4)$$

즉, 무조건 기대치인 $E(Y_1)=\int_{y_1} y_1\, dF_{Y_1}(y_1)$와 같이, 조건부 기대치인 $E(Y_1|Y_2=y_2)$도 어디까지나 Y_1의 기대치이다. 다만, 가중치로 사용하는 확률이 무조건 확률인 $dF_{Y_1}(y_1)=P(y_1<Y_1<y_1+dy_1)$이 아니라 조건부 확률인 $dF(y_1|Y_2=y_2)=P(y_1<Y_1<y_1+dy_1|Y_2=y_2)$일 따름이다.

$dF(y_1|Y_2=y_2)$는 Y_1이 이산 확률변수이면 $P(Y_1=y_1|Y_2=y_2)$를 의미하고,

Y_1, Y_2가 연속 확률변수인 경우에는 $f_{Y_1|Y_2}(y_1|y_2)dy_1$으로 표기하는데, 이때 조건부 밀도함수인 $f_{Y_1|Y_2}(y_1|y_2)$를 (Y_1과 Y_2의) 결합밀도함수 $f(y_1,y_2)$와 Y_2의 밀도함수 $f_{Y_2}(y_2)$로 나타내면 $f_{Y_1|Y_2}(y_1|y_2)=f(y_1,y_2)/f_{Y_2}(y_2)$이다 (<비고 2.14.1> 참조).

<비고 2.14.1> $f_{Y_1|Y_2}(y_1|y_2)dy_1 = P(y_1 < Y_1 < y_1 + dy_1 | y_2 < Y_2 < y_2 + dy_2)$

$$= P(y_1 < Y_1 < y_1 + dy_1, y_2 < Y_2 < y_2 + dy_2)/P(y_2 < Y_2 < y_2 + dy_2)$$

$$= f(y_1,y_2)dy_1 dy_2 / f_{Y_2}(y_2)dy_2$$

<비고 2.14.2> 식 (2.14.4)에서 $dF(y_1|Y_2=y_2)$는 y_1과 y_2의 함수이지만, (y_1을 곱하고) y_1에 대해서 적분한 값인 $E[Y_1|Y_2=y_2]$는 y_2만의 함수이다.

2.14.4 무조건 기대치

식 (2.14.1)을 일반적인 형태로 나타내면 다음과 같다.

$$E(Y_1) = \int_{y_2} E(Y_1|Y_2=y_2)\,dF_{Y_2}(y_2) \qquad (2.14.5)$$

이를 간단히 $E(Y_1) = E[E(Y_1|Y_2)]$로 표현하기도 하는데, 그 이유는 다음과 같다. 식 (2.8.4)에 의해 $E[g(Y_2)] = \int_{y_2} g(y_2)\,dF_{Y_2}(y_2)$인데, y_2의 함수인 $E(Y_1|Y_2=y_2)$를 $g(y_2)$라 하면 (<비고 2.27> 참조), 이에 대응하는 $g(Y_2)$는 $E(Y_1|Y_2=Y_2)$, 즉 $E(Y_1|Y_2)$가 된다.

<비고 2.14.3> " $E(Y_1) = E[E(Y_1|Y_2)]$ "에 대한 해석은 다음과 같다.

Unconditional Mean = Mean of Conditional Means

식 (2.14.2)의 우변과 식 (2.14.3)을 일반적인 형태로 나타내면 다음과 같다.

$$E[V(Y_1|Y_2)] = \int_{y_2} V(Y_1|Y_2=y_2)\,dF_{Y_2}(y_2) \qquad (2.14.6)$$

$$V[E(Y_1 \mid Y_2)] = \int_{y_2} \{E(Y_1 \mid Y_2 = y_2) - E(Y_1)\}^2 dF_{Y_2}(y_2) \qquad (2.14.7)$$
$$= \int_{y_2} \{E(Y_1 \mid Y_2 = y_2)^2 - E(Y_1)^2\} dF_{Y_2}(y_2)$$

식 (2.14.6)에서 조건부 분산인 $V(Y_1 \mid Y_2 = y_2)$의 정의는 다음과 같다.

$$V(Y_1 \mid Y_2 = y_2) = \int_{y_1} \{y_1 - E(Y_1 \mid Y_2 = y_2)\}^2 dF(y_1 \mid Y_2 = y_2)$$

$E(Y_1 \mid Y_2 = y_2)$ 같이 $V(Y_1 \mid Y_2 = y_2)$ 역시 y_2만의 함수인데, 이를 $h(y_2)$라 하면 식 (2.14.6)은 $E[h(Y_2)] = \int_{y_2} h(y_2)\, dF_{Y_2}(y_2)$ 형태이다. 그리고, $g(y_2) = E(Y_1 \mid Y_2 = y_2)$라 하면 식 (2.14.7)은 $V[g(Y_2)] = \int_{y_2} \{g(y_2) - E[g(Y_2)]\}^2 dF_{Y_2}(y_2)$ 형태이다. (비고: 식 (2.14.5)에 의해서 $E[g(Y_2)] = E(Y_1)$ 임). 이들을 묶어서 다음을 얻는다.

$$V(Y_1) = E[V(Y_1 \mid Y_2)] + V[E(Y_1 \mid Y_2)] \qquad (2.14.8)$$

<비고 2.14.4> 식 (2.14.8)에 대한 해석은 다음과 같다.

$$\begin{pmatrix} Unconditional \\ Variance \end{pmatrix} = \begin{pmatrix} Mean\ of \\ Condtional\ Variances \end{pmatrix} + \begin{pmatrix} Variance\ of \\ Conditional\ Means \end{pmatrix}$$

연습문제 삼아 공식들을 증명하되, 편의상 Y_1, Y_2를 이산 확률변수라 하자. 식 (2.11.3)에서 $Y_1 = g(Y_1, Y_2)$라 하면

$$E(Y_1) = \sum_{y_1}\sum_{y_2} y_1 P(Y_1 = y_1, Y_2 = y_2) = \sum_{y_2} P(Y_2 = y_2) \sum_{y_1} y_1 \cdot P(Y_1 = y_1 \mid Y_2 = y_2)$$
$$= \sum_{y_2} P(Y_2 = y_2) E(Y_1 \mid Y_2 = y_2) = E[E(Y_1 \mid Y_2)]$$

를 얻는다. (비고: $P(Y_1 = y_1, Y_2 = y_2) = P(Y_2 = y_2) \cdot P(Y_1 = y_1 \mid Y_2 = y_2)$.) 또한, $Y_1^2 = g(Y_1, Y_2)$라 하면 같은 방법으로 $E(Y_1^2) = E[E(Y_1^2 \mid Y_2)]$를 얻고, 이들로부터 다음을 얻는다.

$$
\begin{aligned}
V(Y_1) &= E(Y_1^2) - E(Y_1)^2 = E[E(Y_1^2 \mid Y_2)] - E[E(Y_1 \mid Y_2)]^2 \\
&= E[V(Y_1 \mid Y_2) + E(Y_1 \mid Y_2)^2] - E[E(Y_1 \mid Y_2)]^2 \\
&= E[V(Y_1 \mid Y_2)] + \{E[E(Y_1 \mid Y_2)^2] - E[E(Y_1 \mid Y_2)]^2\} \\
&= E[V(Y_1 \mid Y_2)] + V[E(Y_1 \mid Y_2)]
\end{aligned}
$$

§2.15 대표적인 표본분포

2.15.1 비복원과 복원의 차이

이제 본격적으로 표본분포를 다룬다(§1.5 참조). 크기가 N인 모집단의 임의 요소를 Y라 하고 Y의 분포를 모분포라 한다. 모집단에서 무작위로 비복원추출한 $\{Y_1, \cdots, Y_n\}$을 크기가 n인 표본이라 하는데, $Y_1, \cdots, Y_n$은 모두 모분포와 동일한 분포를 따르지만 독립은 아니다. $Y_1, \cdots, Y_n$의 함수를 통계량이라 하고, 통계량의 분포를 표본분포라 한다.

$N \gg n$인 경우, 비복원을 복원으로 간주하여 $Y_1, \cdots, Y_n$을 *iid* 확률변수로 취급하면 간편하기 때문에(§1.4 참조), 다음 절부터는 $N \gg n$을 가정하거나 또는 아예 모분포를 연속분포라 가정한다 (<비고 2.7.1> 참조). 그러나, 본 절에서는 마지막으로 비복원추출을 복원추출로 간주함에 따른 차이점을 알아본다.

가장 대표적인 통계량은 표본평균(sample mean) $\overline{Y} = \sum_{i=1}^{n} Y_i/n$이다. $Y_1, \cdots, Y_n$이 *iid* 확률변수이면 모분포로부터 $\overline{Y}$의 분포를 비교적 쉽게 구할 수 있고(§2.12.4 참조), 또한 중심극한정리를 이용해서 정규분포를 $\overline{Y}$의 근사(approximate)분포로 사용할 수도 있다(§2.5 참조). 그러나, $Y_1, \cdots, Y_n$이 (동일하기는 하되) 독립이 아니면 $\overline{Y}$의 분포를 구하기도 쉽지 않고 또한 중심극한정리도 사용할 수 없다. (비고: $\sum_{i=1}^{n} Y_i$이 초기하분포를 따르는 경우가 가장 쉬운데, 이 경우에조차도 §1.6에서 근사분포로 이항분포를 사용했음.) 따라서, 분포의 차이점은 생략하고, $E(\overline{Y})$와 $V(\overline{Y})$의 차이점만 알아본다.

편의상, $\mu = E(Y)$, $\sigma^2 = V(Y)$ 라 하자. 결론부터 언급하면, $E(\overline{Y}) = \mu$ 는 비복원과 복원 모두에 동일하지만 $V(\overline{Y})$는 각각 $(\sigma^2/n)\{(N-n)/(N-1)\}$과 σ^2/n인데, 이때 차이점인 $(N-n)/(N-1)$은 이미 식 (2.8.9)에 등장했다.

먼저, $E(\sum_{i=1}^{n} Y_i) = \sum_{i=1}^{n} E(Y_i) = n\mu$ 이므로 (<비고 2.10> 참조),

$$E(\overline{Y}) = E(\sum Y_i/n) = E(\sum Y_i)/n = \mu \tag{2.15.1}$$

이다. 다음, 식 (2.13.7)에 $a_i = a_j = n^{-1}$ 을 대입하면

$$V(\overline{Y}) = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} Cov(Y_i, Y_j) \qquad (2.15.2)$$

를 얻는데, 복원추출의 경우에는 <비고 2.13.1>과 <비고 2.13.2>에 의해서

$$V(\overline{Y}) = n^{-2} \sum_{i=1}^{n} V(Y_i) = n\sigma^2/n^2 = \sigma^2/n \qquad (2.15.3)$$

을 얻는다. 반면에, 비복원의 경우에는 $i \neq j$일 때 $Cov(Y_i, Y_j) = -\sigma^2/(N-1)$인데 (아래 참조), 이를 식 (2.15.2)에 대입하면 $V(\overline{Y}) = (\sigma^2/n)\{(N-n)/(N-1)\}$을 얻는다.

편의상, 모집단을 $\{y_1, \cdots, y_N\}$이라 하자. 그러면, $\mu = \sum_{i=1}^{N} y_i/N$이고 $\sigma^2 = \sum_{i=1}^{N} (y_i - \mu)^2/N$이다. (비고: $\sum_{i=1}^{N} y_i = N\mu$, $\sum_{i=1}^{N} y_i^2 = N \cdot (\sigma^2 + \mu^2)$.) 대칭성에 의해서 (§1.4 참조) $Cov(Y_i, Y_j)$는 $Cov(Y_1, Y_2)$와 동일하다 (단, $i \neq j$). 그리고, 식 (2.13.2)에 의해서 $Cov(Y_1, Y_2) = E(Y_1 Y_2) - \mu^2$인데, $E(Y_1 Y_2)$는 다음과 같이 식 (2.14.5)와 유사한 방법으로 구할 수 있다.

$$E(Y_1 Y_2) = \sum_{i=1}^{N} P(Y_1 = y_i) \cdot E(Y_1 Y_2 | Y_1 = y_i) \qquad (2.15.4)$$

식 (2.15.4)에서 $P(Y_1 = y_i) = N^{-1}$이고, $E(Y_1 Y_2 | Y_1 = y_i) = E(y_i Y_2 | Y_1 = y_i) = y_i E(Y_2 | Y_1 = y_i)$인데, $E(Y_2 | Y_1 = y_i)$는 모집단에서 y_i를 제외한 나머지 $N-1$개의 요소의 평균이므로 $(N\mu - y_i)/(N-1)$이다. 이를 식 (2.15.4)에 대입하고 $\sum_{i=1}^{N} y_i = N\mu$와 $\sum_{i=1}^{N} y_i^2 = N(\sigma^2 + \mu^2)$을 사용해서 간단히 하면 $E(Y_1 Y_2) = \mu^2 - \{\sigma^2/(N-1)\}$을 얻는다.

2.15.2 $\overline{Y}$의 분포

이제부터는 $N \gg n$을 가정하거나 또는 연속 모분포를 가정하여 $Y_1, \cdots, Y_n$을 iid 확률변수로 취급한다. 아예 한걸음 나아가서 모분포를 정규분포로 가정하든지 또는 중심극한정리를 사용한다. (비고: 통계 교재에 정규분포가 아닌 모분포가 더러 예제나 연습문제로 등장하지만 본문에는 잘 등장하지 않음.)

중심극한정리는 다음과 같다 (§2.5 참조). $Y_1, \cdots, Y_n$이 iid 확률변수이기만 하면

$\sum_{i=1}^{n} Y_i$의 분포는 n이 클수록 점점 정규분포에 가까워지는데, 이를 $\sum_{i=1}^{n} Y_i \overset{A}{\sim} N(n\mu, n\sigma^2)$ 으로 표현하자. 즉, $\sum_{i=1}^{n} Y_i$은 점근적으로 (□A□는 □asymptotically□를 의미함) 정규분포를 따르는데, 기대치와 분산은 각각 $n\mu$와 $n\sigma^2$이다 (식 (2.8.8), (2.8.9) 참조). 따라서,

$$\overline{Y} \overset{A}{\sim} N(\mu, \frac{\sigma^2}{n}) \tag{2.15.5}$$

이다 (§2.9.1 참조). 즉, $\overline{Y}$ 역시 점근적으로 정규분포를 따르며 기대치와 분산은 각각 μ와 σ^2/n이다 (식 (2.15.1), (2.15.3) 참조).

식 (2.15.5)는 모분포를 몰라서 $\overline{Y}$의 분포를 구할 수 없는 경우뿐만 아니라, $\overline{Y}$의 분포를 구할 수 있는 경우에도 쓰인다. 대표적인 예는 $\sum_{i=1}^{n} Y_i$가 이항분포를 따르는 경우이다 (§2.5 참조). (비고: §1.6에서는 초기하분포의 근사분포로 이항분포를 사용했는데, 이제부터는 이항분포의 근사분포로 정규분포를 사용한다.)

모분포가 정규분포라는 가정을 $Y \sim N(\mu, \sigma^2/n)$으로 표현하자. (비고: □ ~ N□은 □is distributed Normal□을 의미함.) $Y \sim N(\mu, \sigma^2)$이면 자동적으로 $\overline{Y} \sim N(\mu, \sigma^2/n)$인데, 이는 다음과 같이 확장된다.

<비고 2.15.1> $Y_1, \cdots, Y_n$이 독립이고 $Y_i \sim N(\mu_i, \sigma_i^2)$이면

$$\sum_{i=1}^{n} a_i Y_i \sim N(\sum a_i \mu_i, \sum a_i^2 \sigma_i^2)$$ 이다.

<비고 2.15.1>에서 $a_i = n^{-1}, \mu_i = \mu, \sigma_i^2 = \sigma^2$ 인 경우가 바로 $\overline{Y} \sim N(\mu, \sigma^2/n)$이다. <비고 2.15.1>의 증명은 MGF로 하면 비교적 간단하다 (§2.12.6 참조). 먼저 Y_i의 MGF는

$$E(e^{\Theta Y_i}) = \exp(\mu_i \Theta + \sigma_i^2 \Theta^2/2) \tag{2.15.6}$$

이다 (§2.12.5 참조). 다음, $a_i Y_i$의 MGF는 $E[e^{\Theta(a_i Y_i)}] = E[e^{(\Theta a_i) Y_i}]$인데, 이는 $E(e^{\Theta Y_i})$에 □Θ□대신 □Θa_i□를 대입한 것이므로 (비고: Θ와 a_i는 확률변수가 아님), 식(2.15.6)에서 Θ를 Θa_i로 대체하여

$$E[e^{\Theta(a_iY_i)}] = \exp[(a_i\mu_i)\Theta + (a_i^2\sigma_i^2)\Theta^2/2] \quad (2.15.7)$$

를 얻는데, 이는 $a_iY_i \sim N(a_i\mu_i, a_i^2\sigma_i^2)$을 의미한다. 마지막으로, $Y_1, \cdots, Y_n$이 서로 독립이면 $a_1Y_1, \cdots, a_nY_n$도 서로 독립이므로 (<비고 2.19> 참조), $\sum_{i=1}^{n} a_iY_i$의 MGF는

$$E[e^{\Theta\sum_{i=1}^{n}(a_iY_i)}] = \prod_{i=1}^{n} E[e^{\Theta(a_iY_i)}]$$

이다 (§2.12.4 참조). 이에 식 (2.15.7)을 대입하면

$$E[e^{\Theta\left(\sum_{i=1}^{n} a_iY_i\right)}] = \exp[(\sum a_i\mu_i)\Theta + (\sum a_i^2\sigma_i^2)\Theta^2/2]$$

를 얻는데, 이는 $\sum_{i=1}^{n} a_iY_i \sim N(\sum a_i\mu_i, \sum a_i^2\sigma_i^2)$을 의미한다. (비고: 확률분포와 MGF간에는 1:1 대응관계가 있음.)

2.15.3 S^2의 분포

표본평균 $\overline{Y}$ 다음으로 중요한 통계량은 표본분산(sample variance)인데, 관행상 이를 S^2으로 표기한다. (지금까지 종종 S를 □sum□으로 사용했는데, 이제부터 S는 표본표준편차(sample standard deviation)를 의미한다. 그러나, 앞으로 등장할 □sum of squares□는 SS로 표기한다.)

이제부터는 $Y \sim N(\mu, \sigma^2)$을 가정하는데, 이는 모분포가 정규분포가 아니면 (일반적으로) S^2및 기타 통계량의 분포를 구하기 어렵기 때문이다. (비고: 중심극한정리는 $\overline{Y}$에 대해서만 유효함.)

이제 S^2을 정의한다. 앞으로 S^2을 σ^2에 대한 추정량으로 사용할 것이므로, $\sigma^2 = V(Y) = E[(Y-\mu)^2]$에 대응하는

$$S^2 = \frac{\sum_{i=1}^{n}(Y_i-\mu)^2}{n} \quad (2.15.8)$$

로 (일단) 정의한다. 그러나, σ^2을 몰라서 S^2으로 추정하는 상황에서는 $\mu = E(Y)$조차 모르는 것이 더욱 현실적이다. μ를 모르는 경우에는 μ에 대한 추정량인 $\overline{Y}$로 대체하여

$$S^2 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}{n-1} \tag{2.15.9}$$

로 정의한다. 이때, 분모에 n대신 $(n-1)$을 사용하는 이유는 $E(S^2) = \sigma^2$관계를 유지하기 위함이다 (식 (3.2.7) 참조).

§2.6.1에서 $C_d = \sum_{i=1}^{d} Z_i^2$는 자유도가 d인 카이제곱분포를 따른다고 했는데, 이를 $C_d \sim \chi^2(d)$로 표현하자. (이 역시 MGF로 증명이 가능하나 생략함.) $Z_i = (Y_i - \mu)/\sigma$ 라 하면, $Y_1, \cdots, Y_n$이 $iid\ N(\mu, \sigma^2)$이므로 $Z_1, \cdots, Z_n$은 $iid\ N(0, 1^2)$이다. 그러므로,

$$C_n = \sum_{i=1}^{n} Z_i^2 = \sum_{i=1}^{n} \frac{(Y_i - \mu)^2}{\sigma^2} \sim \chi^2(n) \tag{2.15.10}$$

이다. 따라서, 식 (2.15.8)의 S^2의 함수인 nS^2/σ^2은 자유도가 n인 카이제곱분포를 따른다. 그리고, <비고 2.8.10>에 의해서 $E(nS^2/\sigma^2) = n$, $V(nS^2/\sigma^2) = 2n$이므로, $E(S^2) = \sigma^2$과 $V(S^2) = 2\sigma^4/n$을 얻는다.

반면에, 앞으로 주로 쓰일 식 (2.15.9)의 S^2과 관련된 분포를 구하는 과정은 약간 복잡한데, 결론부터 언급하면

$$C_{n-1} = \sum_{i=1}^{n} \frac{(Y_i - \overline{Y})^2}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1) \tag{2.15.11}$$

이다. 그리고, <비고 2.8.10>에 의해서 $E[(n-1)S^2/\sigma^2] = n-1$, $V[(n-1)S^2/\sigma^2] = 2(n-1)$ 이므로 다음을 얻는다.

$$E(S^2) = \sigma^2, \quad V(S^2) = 2\sigma^4/(n-1) \tag{2.15.12}$$

편의상, $n = 3$ 경우에 대해서 식 (2.15.11)을 증명한다. $\sum_{i=1}^{3}(Y_i - \overline{Y})^2/\sigma^2$에 $\overline{Y} = (Y_1 + Y_2 + Y_3)/3$을 대입하면 $(2/3\sigma^2)(Y_1^2 + Y_2^2 + Y_3^2 - Y_1Y_2 - Y_2Y_3 - Y_3Y_1)$이 되는데, 이를 정리하면 $\{(Y_1 - Y_2)^2/2\sigma^2 + (Y_1 + Y_2 - 2Y_3)^2/6\sigma^2\}$가 된다. 그런데, <비고 2.15.1>에 의해서 $(Y_1 - Y_2) \sim N(0, 2\sigma^2)$이고 $(Y_1 + Y_2 - 2Y_3) \sim N(0, 6\sigma^2)$이다. 또한 <비고 2.13.5>에 의해서 $Cov(Y_1 - Y_2, Y_1 + Y_2 - 2Y_3) = 0$이므로, <비고 2.13.4>에 의해서 $(Y_1 - Y_2)$와 $(Y_1 + Y_2 - 2Y_3)$는 독립이다. 따라서, $Z_1 = (Y_1 - Y_2)/\sqrt{2}\sigma$와

$Z_2 = (Y_1 + Y_2 - 2Y_3)/\sqrt{6}\sigma$는 $iid\ N(0,1^2)$이므로, $\sum_{i=1}^{3}(Y_i - \overline{Y})^2/\sigma^2 = Z_1^2 + Z_2^2 \sim \chi^2(2)$가 된다.

일반적으로는 $\sum_{i=1}^{n}(Y_i - \overline{Y})^2/\sigma^2 = \sum_{i=1}^{n-1} Z_i^2 \sim \chi^2(n-1)$인데, 이때 $Z_1, \cdots, Z_{n-1}$은 $iid\ N(0,1^2)$이고 $Z_i = (Y_1 + \cdots + Y_i - iY_{i+1})/\sqrt{i+i^2}\sigma$이다($i = 1, 2, \cdots, n-1$).

2.15.4 자유도

§2.6에서 소개한 카이제곱분포, t분포, F분포는 모두 자유도가 있다. 자유도란 한마디로 사용된 정보의 개수인데, 편의상 $n = 3$인 경우로 이를 설명한다.

$Y_1, Y_2, Y_3 \sim iid\ N(\mu, \sigma^2)$인 표본 $\{Y_1, Y_2, Y_3\}$에 담긴 (독립적인) 정보는 3개이다. 그런데, 표본에 담긴 정보는 정보의 손실없이 다양한 형태로 변형시킬 수 있다. 예를 들어, μ가 알려진 상수인 경우에 $\{Y_1 - \mu, Y_2 - \mu, Y_3 - \mu\}$로 변형시켜도 여전히 (독립적인) 정보는 3개이다. 따라서, 식 (2.15.10)에서 $\sum_{i=1}^{3}(Y_i - \mu)^2$은 3개의 (독립적인) 정보인 $Y_1 - \mu, Y_2 - \mu, Y_3 - \mu$를 사용하므로 자유도는 3이다.

표본정보를 다음과 같이 세가지 형태로 변형시킨다. 첫째, $\{Y_1, Y_1 + Y_2, Y_1 + Y_2 + Y_3\}$에서는 $Y_1, Y_1 + Y_2, Y_1 + Y_2 + Y_3$가 독립은 아니지만 그래도 정보의 손실은 없다. (즉, Y_1, Y_2, Y_3를 다시 얻어낼 수 있다.) 둘째로, $\{Y_1 - \overline{Y}, Y_2 - \overline{Y}, Y_3 - \overline{Y}, \overline{Y}\}$에서는 $Y_1 - \overline{Y}, Y_2 - \overline{Y}, Y_3 - \overline{Y}$끼리는 독립이 아니지만 이들 모두 $\overline{Y}$와는 독립이다 (<비고 2.13.4>, <비고 2.13.5> 참조). 이때 유의할 점은 $\overline{Y}$가 하나의 (독립적인) 정보이므로($\overline{Y}$를 제외한) $\{Y_1 - \overline{Y}, Y_2 - \overline{Y}, Y_3 - \overline{Y}\}$에 담긴 정보는 3개가 아니라 2개라는 점이다. 따라서, 식 (2.15.11)에서 $\sum_{i=1}^{3}(Y_i - \overline{Y})^2$은 2개의 정보를 사용하(는 것과 같으)므로 자유도는 2이다. 셋째로, $\{Y_1 - Y_2, Y_1 + Y_2 - 2Y_3, Y_1 + Y_2 + Y_3\}$에서는 3개의 정보가 모두 독립이다(비고: $Y_1 + Y_2 + Y_3 = 3\overline{Y}$). 그리고, §2.15.3에서 보았듯이 $\sum_{i=1}^{3}(Y_i - \overline{Y})^2 = \{(Y_1 - Y_2)^2/2 + (Y_1 + Y_2 - 2Y_3)^2/6\}$이므로, 이는 (독립적인) 3개의 정보 중에서 2개만 사용하는 것과 같다.

<비고 2.15.2> $\overline{Y}$와 $\sum_{i=1}^{n}(Y_i-\overline{Y})^2$는 독립이다 ($n \geq 2$).

2.15.5 t분포의 등장

앞으로 μ에 대한 가설을 검정할 때, σ가 알려진 경우에는 $(\overline{Y}-\mu)/(\sigma/\sqrt{n})$을 검정통계량으로 사용한다. 이때, $Y\sim N(\mu,\sigma^2)$이면 $(\overline{Y}-\mu)/(\sigma/\sqrt{n})\sim N(0,1^2)$이고, $Y\nsim N(\mu,\sigma^2)$이면 중심극한정리에 의해서 $(\overline{Y}-\mu)/(\sigma/\sqrt{n}) \xrightarrow{A} N(0,1^2)$이다 (식 (2.15.5) 참조).

그러나, σ를 모르는 경우가 더 현실적인데, 이때 σ를 식 (2.15.9)의 S로 대체하여 $(\overline{Y}-\mu)/(S/\sqrt{n})$을 검정통계량으로 사용한다. 그러면, $Y\sim N(\mu,\sigma^2)$인 경우에 한해서 (<비고 2.15.3> 참조) $(\overline{Y}-\mu)/(S/\sqrt{n})\sim t(n-1)$이다. 즉, $(\overline{Y}-\mu)/(S/\sqrt{n})$은 자유도가 $(n-1)$인 t분포를 따르는데, 증명은 다음과 같다. $Y\sim N(\mu,\sigma^2)$이면, $Z=(\overline{Y}-\mu)/(\sigma/\sqrt{n})\sim N(0,1^2)$이고, 식 (2.15.11)에 의해서 $C_{n-1}=(n-1)S^2/\sigma^2\sim\chi^2(n-1)$이다. 또한, <비고 2.15.2>에 의해서 $\overline{Y}$와 $(n-1)S^2$은 독립이므로 Z와 C_{n-1} 역시 독립이다. 따라서, 이들을 식 (2.5.5)에 대입하여 다음을 얻는다.

$$T_{n-1}=\frac{Z}{\sqrt{C_{n-1}/(n-1)}}=\frac{\overline{Y}-\mu}{S/\sqrt{n}}\sim t(n-1) \tag{2.15.13}$$

<비고 2.15.3> 이때 유의할 점은 다음과 같다. $Y\nsim N(\mu,\sigma^2)$인 경우 $(\overline{Y}-\mu)/(S/\sqrt{n})$은 t분포를 따르지 않는다. 이 경우에도 $N(0,1^2)$을 근사분포로 사용할 수 밖에 없는데, 이는 2-단계 근사라고 할 수 있다. 즉, 1차적으로는 중심극한정리에 의한 근사이고, 2차적으로는 σ를 S로 대체함에 따른 근사이다.

사실 n이 크면 어차피 $t(n-1)$이나 $N(0,1^2)$이나 별로 차이가 없다 (<비고 2.6.1> 참조). 다만, t분포의 정의상 서로 독립인 Z와 C_{n-1}이 필요한데, 이는 $Y\sim N(\mu,\sigma^2)$일 때에만 가능하다는 것이다.

2.15.6 F분포의 등장

모집단이 2개(이상) 있을 때 F분포가 자연스럽게 등장한다. 모분포가 $Y \sim N(\mu_1, \sigma_1^2)$인 모집단에서 추출한 표본을 $\{Y_1, \cdots, Y_{n_1}\}$이라 하고, 모분포가 $X \sim N(\mu_2, \sigma_2^2)$인 또 다른 모집단에서 추출한 표본을 $\{X_1, \cdots, X_{n_2}\}$라 하자. 그리고, 각각의 표본평균, 표본분산을 $\overline{Y} = \sum_{i=1}^{n_1} Y_i / n_1$, $S_1^2 = \sum (Y_i - \overline{Y})^2 / (n_1 - 1)$과 $\overline{X} = \sum_{i=1}^{n_2} X_i / n_2$, $S_2^2 = \sum (X_i - \overline{X})^2 / (n_2 - 1)$이라 하자. 그러면, 식 (2.15.11)에 의해서 $C_{n_1 - 1} = (n_1 - 1) S_1^2 / \sigma_1^2 \sim \chi^2(n_1 - 1)$이고 $C_{n_2 - 1} = (n_2 - 1) S_2^2 / \sigma_2^2 \sim \chi^2(n_2 - 1)$이다. 또한, 서로 다른 모집단에서 추출한 표본들은 서로 독립이므로, $C_{n_1 - 1}$과 $C_{n_2 - 1}$은 서로 독립이다. 따라서, 이들을 식 (2.5.4)에 대입하여 다음을 얻는다.

$$F_{n_1 - 1, n_2 - 1} = \frac{C_{n_1 - 1} / (n_1 - 1)}{C_{n_2 - 1} / (n_2 - 1)} = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} \sim F(n_1 - 1, n_2 - 1) \qquad (2.15.14)$$

즉, $(S_1^2/\sigma_1^2)/(S_2^2/\sigma_2^2)$는 분자 자유도가 $(n_1 - 1)$이고 분모 자유도가 $(n_2 - 1)$인 F분포를 따른다.