채 교수의

# 통계학 강의

한국과학기술원 산업공학과 확률모형 연구실
http://OSL7.kaist.ac.kr

채 교수의

# 통계학 강의

**한국과학기술원 산업공학과 확률모형 연구실**

(재학생) 김남기, 윤봉규, 김진동, 장석호, 조준, 최대원, 유준기, 이태경

(졸업생) 장영훈, 김태성, 안창원, 김영진, 박성준, 양원석, 김길환, 박연일, 박현민, 성기원

발행일: 2000년 10월 10일

등록번호: 한국과학기술원 산업공학과 Technical Report #00-12

# 머리말

이 책은 교재가 아니라 참고서입니다. 통계학 교재를 여러 가지 영양분을 두루 갖춘 그러나 양이 많은 음식에 비유한다면, 이 책의 역할은 과식한 독자의 소화를 돕고 편식한 독자에게는 결핍된 비타민을 보충해주는 것입니다.

이 책의 수준은 학부과정의 수리통계학 수준이고, 이 책의 목적은 학부 및 대학원과정의 관련과목에 대비해서 통계학의 핵심적인 개념을 확고히 하는 것입니다.

현재 KAIST 산업공학과의 학부과정 필수과목인 공업통계 I과 II의 교재는 Mathematical Statistics with Applications (5판)인데, 저자 3명의 initial을 따서 WMS라 부르겠습니다 (문헌 [9] 참조). 공업통계 I에서는 WMS의 전반부를 다루고 공업통계 II에서는 WMS의 후반부를 다룹니다. 저희 확률모형 연구실의 지도교수이신 채경철 교수님은 지금까지 WMS의 전반부를 7번, 후반부를 8번 강의하셨습니다. (비고: 채 교수님이 경영과학과에 계실 때에도 WMS를 경영통계 I과 II의 교재로 사용하셨습니다.)

WMS는 무척 좋은 책입니다만, 798쪽이나 되는 방대한 양이라서 (예제와 연습문제는 총 1,549개) 자칫 독자가 숲은 못 보고 나무만 보게 되는 경향이 있습니다. 또한, 원서로서의 장점도 많지만 반면에 뜻이 명확하게 와 닿지 않는 단점도 있습니다. 이에, 채 교수님이 WMS로 강의하실 때 강조하신 내용을 간추려서 (가려운 데를 긁어주는) 참고서 형태로 정리한 것이 이 책입니다.

이 책의 제목은 “The Feynman Lectures on Physics”에서 땄으며, 부제는 “뜻으로 본 통계학”입니다.

이 책을 교수님의 영애 채윤이, 채선이 양께 헌정합니다.

2000년 10월

KAIST 확률모형 연구실 일동

# 목차

# 제 1장 통계학이란



## §1.1 모집단과 표본

통계학이란 한 마디로 부분을 가지고 전체에 대해서 왈가왈부하는 것이다. 구체적으로, 표본(sample)에 담긴 정보를 사용하여 모집단(population)의 성질(characteristics)을 추론(inference)하는 것인데, 추론은 크게 추정(estimation)과 검정(hypothesis test)으로 나뉜다. 먼저, 추정에 관한 실감나는 사례를 들어보자.

<사례 1.1> '92 대통령선거 후보별 득표율 (%)

| | 김영삼 | 김대중 | 정주영 | 박찬종 | 기타 |
|---|---|---|---|---|---|
| 예측치 | 39.5 | 31.1 | 15.7 | 12.4 | 1.2 |
| 실제값 | 42.0 | 33.8 | 16.3 | 6.4 | 1.5 |

<사례 1.1>에서 예측치는 한국갤럽조사연구소가 투표함이 열리기 전에 언론기관에 발표한 것인데, 이에 사용된 표본의 크기는 약 이천이다 (문헌 [2] 참조). 반면에, 모집단의 크기는 전체 유권자 중에서 투표권을 행사한 사람의 수로서 어림잡아 이천만은 될 것이다.

<비고 1.1.1> “표본이 이천개”라 하지 않고, 표본은 하나인데 그 “크기가 이천”이라 표현함.

예측(prediction 또는 forecasting)도 일종의 추정이다. 예측의 경우 대개 시간이 지나면 실제값이 알려진다. <사례 1.1>에서도 개표가 끝난 후 실제 득표율이 알려졌다. 그러나, 일반적인 추정의 상황에서는 대체로 실제값이 알려지지 않는다.

<사례 1.2> ‘92 대통령선거 후보별 지지율

<사례 1.2>에서의 관심사는 (기권한 사람을 포함한) 전체 유권자의 후보별 지지율이다. 이 경우 실제 지지율은 알려지지 않는다. 다만, 이에 대한 추정치로 <사례 1.1>의 후보별 득표율을 사용할 수 있을 것이다. 이때, 약 이천명에 근거한 득표율 예측치보다는 약 이천만명에 근거한 실제 득표율을 (실제 지지율에 대한) 추정치로 사용하는 것이 바람직할 것이다.

<사례 1.2>에서 모집단의 크기는 전체 유권자의 수로서 약 삼천만이다. 그리고, 실제 득표율을 추정치로 사용하는 경우 표본의 크기는 약 이천만이다.

<비고 1.1.2> 모집단은 고정된 것이 아니라 상황에 따라 달라지는 것이다.

## §1.2 다루기 쉬운 표본

우리가 사용할 표본은 수학적으로 가장 다루기 쉬운 것으로써, 임의표본(random sample)이라 불리는 것이다. (비고: 책에 따라 확률표본 또는 랜덤표본이라고도 함.)

<사례 1.1>에서와 같이, 여론조사기관에서 예측치를 발표할 때에 흔히 오차의 범위도 함께 발표한다. 예를 들어, "95%의 신뢰수준에서 최대오차는 ±2.2"라고 발표한다 (§3.1.4, §3.1.5 참조). 그런데, 이러한 오차의 범위는 다음과 같은 가정 하에서 계산된 것이다.

첫째로, 표본을 모집단의 부분집합으로 가정한다. 이 가정을 <사례 1.1>에 적용하면, 여론조사에 응한 약 이천명의 유권자는 기권을 하지 않아야 되고 또한 반드시 여론조사 때 응답했던 대로 투표를 해야 된다. (출구조사의 경우에도 투표한 후보를 솔직하게 알려야 된다.) 이 가정에 완벽하게 일치하는 표본 및 예측치는 개표가 완료되기 전에 발표된 개표상황 및 이에 따른 후보별 지지율이다.

둘째로, 표본은 모집단에서 임의로 (또는 무작위로) 추출한다고 가정한다. 이 가정을 <사례 1.2>에 적용하면, 투표권을 행사한 약 이천만명은 전체 유권자 중에서 임의로 뽑힌 사람들이어야 된다. 만약 임의로 약 이천만명을 뽑았다면, 임의로 약 이천명을 뽑을 때에 비해서 오차의 범위는 1/100 정도로 줄어든다. (95%의 신뢰수준에서 최대오차는 ±0.022. 단, 아래의 세번째 가정 하에서.) 그러나, <사례 1.2>에서는 실제 지지율이 실제 득표율과 제법 차이가 날 가능성이 있는데, 이는 후보별로 지지자들의 기권율이 제법 차이가 날 수 있기 때문이다. (즉, 임의추출로 간주하기 어렵다.)

셋째로, 특별히 무리가 없는 한 모집단의 크기 $N$을 $\infty$로 간주한다. (반면에, 표본의 크기 $n$은 상대적으로 $N$보다 훨씬 작아야 된다.) 예를 들어, <사례 1.1>에서는 이 가정이 별로 무리가 없다. 즉, $N$이 약 이천만이면 $N \to \infty$라고 간주할만하다. (그러나, $N$이 약 삼천만인 <사례 1.2>에서는 오히려 이 가정이 다소 무리가 있는데, 그 이유는 $n \approx 2N/3$ 이기 때문이다.) 이 가정의 내막은 다음과 같다. 표본을 모집단

의 임의부분집합이라고 정의했는데, 부분집합이라는 말은 비복원추출(sampling without replacement)을 의미한다. 그런데, 비복원추출은 복원추출(sampling with replacement)에 비해서 다루기가 까다롭다. 그렇지만, $N \to \infty$이(고 $n << N$이)면, 비복원과 복원의 차이를 무시할 수 있게 된다 (§1.4, §2.7.1 참조). 따라서, $N \to \infty$이(고 $n << N$이)면 비복원으로 추출한 표본을 마치 복원으로 추출한 것처럼 취급함으로써 복원추출에 따른 수학적인 편의를 취할 수 있게 된다.

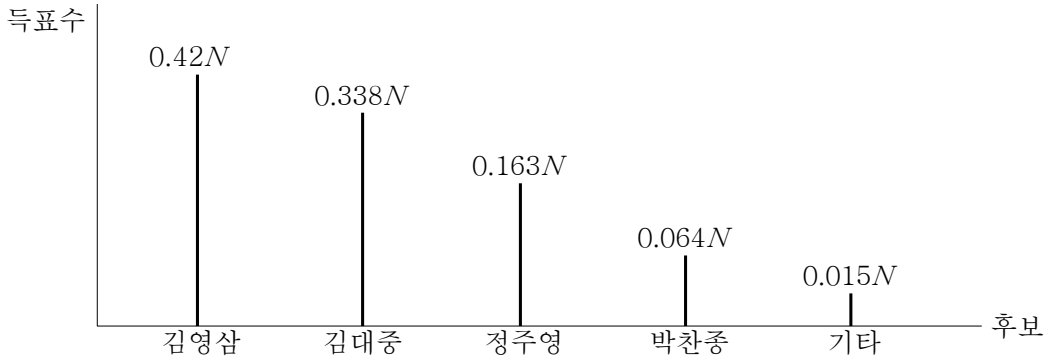
## §1.3 모집단의 표현

<비고 1.1.2>에서 모집단은 통계적 추론의 목적에 따라 달라진다고 했다. 추론의 목적은 <사례 1.1>에서는 득표율의 추정이고 <사례 1.2>에서는 지지율의 추정이며, 모집단의 크기는 각각 약 이천만과 약 삼천만이라고 했다. 그러나, 모집단의 실체에 대해서는 아직도 딱부러지게 언급되지 않았다.

사실 모집단이란 약간은 추상적인 개념으로써, 한 마디로 "정의하기 나름"이다. 결국 문제는 필요한 정보가 무엇인지 그리고 얻을 수 있는 정보가 무엇인지에 있다. 예를 들어, <사례 1.1>에서의 관심사는 득표율이다. 그러므로, 투표권을 행사한 약 이천만명의 유권자에 관한 각종 정보 중에서 유일하게 필요한 정보는 후보별 득표수이다. 누가 누구를 찍었는지는 (찍은사람 외에는) 알 수도 없고, 또한 알 필요도 없다. 누가 어느 투표장에서 투표했으며, 그 표가 어느 개표장에서 개표되었지는 알 수가 있으나, 이 역시 불필요한 정보이다.

<비고 1.3.1> "꼭 필요한 (최소한의) 정보"라는 개념은 (§3.4에서) 가장 효과적인 추정방법을 찾을 때에 쓰인다.

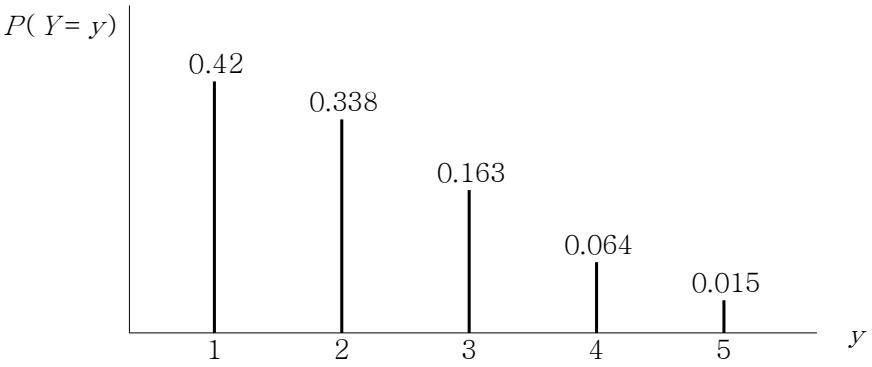
<사례 1.1>에서 후보별 (실제) 득표수는 <그림 1.1>과 같은데, 편의상 이를 모집단의 분포(distribution)라 하자. 이때, 분포란 도수(frequency)분포를 의미한다. 예를 들어, 김영삼 후보는 전체 $N$표의 42%인 $0.42N$표를 얻었다.




<그림 1.1> <사례 1.1>의 모집단의 분포

모집단의 분포를 확률변수(random variable)의 분포로 표현하면 사용하기에 편리하다. 확률변수란 전체집합(universal set)의 요소(element)들 각각에 실수를 하나씩 대응시키는 함수이다. 그러니까, 두 개 이상의 요소가 동일한 실수에 대응될 수는 있으나, 하나의 요소는 반드시 하나의 실수에만 대응되어야 된다. 물론 모집단도 전체집합이다. (그리고, 표본은 모집단의 부분집합이다.)

확률변수 $Y$를 다음과 같이 정의한다. <그림 1.1>의 순서대로, 김영삼 후보(가 얻은 표 또는 김영삼 후보를 찍은 유권자)는 1에, 김대중 후보는 2에,…, 기타 후보 (및 무효표)는 5에 대응시키자. 그러면, $Y$의 분포는 <그림 1.2>와 같다.




<그림 1.2> <사례 1.1>의 모집단의 확률분포

<비고 1.3.2> 관례상 확률변수는 대문자로 표기한다.

<비고 1.3.3> 확률변수의 분포를 확률분포라 하는데, 이는 합이 1이 되도록 정규화 (normalize) 되었기 때문이다. 즉, 확률분포는 상대도수(relative frequency)의 분포라 할 수 있다.

<그림 1.2>에서 $Y$의 분포를 모집단의 "확률분포"라 부른 이유는 다음과 같다. 모집단의 $N$개의 요소 중에서 하나를 임의로 뽑는다고 하자. 임의로 뽑으므로, 뽑힐 확률은 $N$개 모두 $N^{-1}$ 씩으로 동일하다. 그런데, 예를 들어, 1에 대응된 요소의 수는 모두 $0.42N$개이다. 따라서, 임의로 뽑힌 하나의 요소가 1에 대응된 요소일 확률은 $0.42N \cdot N^{-1} = 0.42$인데, 이를 $P(Y=1) = 0.42$로 표현한다.

<비고 1.3.4> 모집단의 분포를 모분포(population distribution)라 하는데, 이는 (편의상) $Y$의 분포를 지칭하는 것이다.

앞으로는 $Y$의 분포를 모분포라 할 뿐더러, $Y$를 모집단의 임의요소라 한다. 그리고, $P(Y=y)$는 임의요소(에 대응된 실수)가 $y$일 확률을 의미한다. (비고: "$Y=y$"를 $Y$가 $y$로 구현(realize)되었다고 표현함.) 이때 유의할 점은 다음과 같다. 대문자 $Y$는 확률변수이지만 소문자 $y$는 실수이다. 예를 들어, $P(Y=5)=0.015$이고 $P(Y=6)=0$이다. 또는, $P(Y=y)=0.42$, if $y=1;\cdots;P(Y=y)=0.015$, if $y=5$; $P(Y=y)=0$, if $y \notin \{1,2,3,4,5\}$로 표현하기도 한다.

## §1.4 표본의 표현

§1.3에서, 모집단의 임의요소를 확률변수 $Y$로 표현하고 $Y$의 분포를 모분포라 하면 편리하다고 했다. 마찬가지로, 표본도 확률변수로 표현하면 사용하기에 편리하다.

§1.2에서, 표본을 모집단의 임의 부분집합으로 정의했다. 결론부터 말하자면, 모집단의 임의요소를 $Y$로 표현하듯이 모집단의 임의 부분집합을 $\{Y_1, Y_2, \cdots, Y_n\}$으로 표현한다 ($1 \le n \le N-1$). 이때, $Y_1, \cdots, Y_n$은 각각 모집단의 임의요소를 의미한다. 따라서, $Y_1, \cdots, Y_n$의 분포는 모두 모분포와 같다. 그러나, 비복원추출에 따른 종속성 때문에 $Y_1, \cdots, Y_n$은 서로 독립이 아니다. 그렇지만, $n << N$이면 종속성을 무시할 수 있게 되어 $Y_1, \cdots, Y_n$을 서로 독립인 확률변수로 취급할 수 있다. (즉, 비복원으로 추출한 표본을 복원으로 추출한 것처럼 취급할 수 있게 된다.)

<비고 1.4.1> 서로 독립이고 동일한 분포를 따르는 $Y_1, Y_2, \cdots, Y_n$을 *iid* (independent and identically distributed) 확률변수라 하고, 이들을 대표하는 $Y$를 generic 확률변수라 한다.

크기가 $N$인 모집단의 부분집합은 모두 $2^N$개인데, 공집합과 전체집합을 제외하면 $2^N - 2$개이다. 이 중에서 크기가 $n$인 것은 $\binom{N}{n}$개가 있다 ($1 \le n \le N-1$). $N$개에서 $n$개를 "임의"로 뽑으므로, 총 $\binom{N}{n}$개의 특정(specific) 부분집합들이 표본으로 뽑힐 확률은 각각 $\binom{N}{n}^{-1}$씩이다. (이는 대칭성에 근거한 결과로써, 특별히 어느 부분집합이 다른 부분집합에 비해서 표본으로 뽑힐 확률이 커야될 이유가 없다는 것이

다.) 따라서, $\binom{N}{n}$개의 특정 부분집합들 중에서 하나를 뽑기 전까지는 (또는, 뽑았더라도 그 내용을 확인하기 전까지는) 표본을 확률적으로 표현할 수 밖에 없다. (단, $n=N$ 경우는 확률적이 아니라 확정적이므로 제외했음.)

앞으로 표본을 $\{Y_1, Y_2, \cdots, Y_n\}$으로 표현한다 ($1 \le n \le N-1$). $\{Y_1, \cdots, Y_n\}$은 "모집단의 임의의 부분집합"을 의미하는데, 이는 $Y$가 "모집단의 임의의 요소"를 의미하는 것과 같은 이유이다. 그러나, 일단 $\binom{N}{n}$개 중에서 하나가 뽑히고 그 내용이 확인되면 이를 관찰된(observed 또는 realized) 표본이라 하고 이를 $\{y_1, y_2, \cdots, y_n\}$으로 표현한다. 물론, $\{y_1, \cdots, y_n\}$은 임의의 부분집합이 아니라 운좋게(?) 뽑힌 특정 부분집합을 의미한다. (비고: $y_1, \cdots, y_n$은 모두 실수임.)

집합에서 요소들의 배열순서는 의미가 없다. 예를 들어, $\{1,2,3\}, \{1,3,2\}, \{2,1,3\}, \{2,3,1\}, \{3,1,2\}, \{3,2,1\}$은 모두 같은 집합이다. 따라서, 표본 $\{Y_1, \cdots Y_n\}$의 요소들은 모두 대등한 입장에 있다. 또한, 대등한 입장이므로 동일한 분포를 따르는데, 바로 $Y$의 분포인 모분포를 따른다. 즉, $Y_1, \cdots, Y_n$은 각각 모집단의 임의요소를 나타낸다. 이는 $n=1$인 표본 $\{Y_1\}$에서는 당연하다. 그러나, $n \ge 2$인 경우에는 수긍이 가지 않을 수도 있는데, 그 이유는 비복원추출에 따른 종속성 때문이다. 즉, $Y_1, \cdots, Y_n$은 서로 독립이 아니다. 간단한 예를 통해서 이를 설명한다.

모집단 $\{1,3,5,7,9\}$의 분포는 $P(Y=y)=1/5$이다 ($y=1,2,3,4,5$). $n=2$인 부분집합은 $\binom{5}{2}=10$개이므로, 예를 들어, $\{1,3\}$이 표본으로 뽑힐 확률은 1/10이다. 즉, $P(\{Y_1, Y_2\}=\{1,3\})=1/10$ 이다. 편의상, 2개를 한꺼번에 뽑지 않고 차례로 하나씩 뽑는다고 하자. 그리고, 처음 뽑히는 요소를 $X_1$, 두번째로 뽑히는 요소를 $X_2$라 하자. 이제 뽑히는 순서까지 따지므로, 경우의 수는 2!배로 늘어서 모두 20이 된다. 예를 들어, $P(X_1=1, X_2=3) = P(X_1=3, X_2=1)=1/20$이다. 그러나, $\{1,3\}$이

표본으로 뽑힐 확률은 여전히 1/10인데, 이는 $P(\{Y_1, Y_2\} = \{1,3\})$ $= P(X_1 = 1, X_2 = 3) + P(X_1 = 3, X_2 = 1)$ 이기 때문이다. 즉, $Y_1$은 $X_1$일 수도 있고 $X_2$일 수도 있다. $Y_2$ 역시 $X_1$일 수도 있고 $X_2$일 수도 있다. 단, 이 예제에서는 모집단의 요소가 모두 다르므로, $X_1 \neq X_2$ 이고 $Y_1 \neq Y_2$ 이다. 이제 $X_1$과 $X_2$의 분포를 구한다. $X_1$은 처음 뽑히는 요소이므로 당연히 $P(X_1 = x) = 1/5$이다 ($x = 1,3,5,7,9$). 즉, $X_1$의 분포는 모분포와 같다. 그런데, $X_2$는 두번째로 뽑히는 요소이므로 첫번째에 무엇이 뽑히는가에 따라 (조건부) 분포가 달라진다. 즉, $P(X_2 = x \mid X_1 \neq x) = 1/4$ 이고 $P(X_2 = x \mid X_1 = x) = 0$ 이다 ($x = 1,3,5,7,9$). 그러나, $X_1$에 관한 정보가 없으면 $X_2$의 (무조건: unconditional) 분포는 모분포와 동일하다. 즉,

$$
\begin{aligned}
P(X_2 = x) &= P(X_2 = x, X_1 \neq x) + P(X_2 = x, X_1 = x) \\
&= P(X_1 \neq x) \cdot P(X_2 = x \mid X_1 \neq x) + P(X_1 = x) \cdot P(X_2 = x \mid X_1 = x) \\
&= (4/5)(1/4) + (1/5)(0) = 1/5, \quad x = 1,3,5,7,9
\end{aligned}
$$

이다. (비고: 이 또한 일종의 대칭성으로 간주할 수 있음.) 따라서, $X_1$일 수도 있고 $X_2$일 수도 있는 $Y_1$과 $Y_2$의 분포는 모두 모분포와 동일하다. 그러나, $X_1$과 $X_2$가 서로 종속이듯이 $Y_1$과 $Y_2$도 서로 종속이다. (비고: $X_2$가 $X_1$에 종속이면 $X_1$도 $X_2$에 종속임.) 즉, $P(Y_2 = y \mid Y_1 \neq y) = 1/4 \neq P(Y_2 = y)$ 이고 $P(Y_2 = y \mid Y_1 = y) = 0 \neq P(Y_2 = y)$ 이다 ($y = 1,3,5,7,9$).

다음은 $N = 5000$이고 $n = 2$ 인 예이다. 모집단의 5000개 요소 중에 1,3,5,7,9가 각각 1000개씩 이면, 모분포는 여전히 $P(Y = y) = 1000/5000 = 1/5$ 이다 ($y = 1,3,5,7,9$). 그리고, $X_1, X_2, Y_1, Y_2$의 분포는 모두 모분포와 같다. 그러나,

$P(X_2 = x \mid X_1 \neq x) = 1000/4999 \approx 1/5$ 이고 $P(X_2 = x \mid X_1 = x) = 999/4999 \approx 1/5$ 이다. 즉, 1,3,5,7,9가 1000개씩이나 있으므로, 두번째에 뽑히는 요소는 첫번째에 무엇이 뽑히든 별로 영향을 받지 않는다.

이와 같이, $N$이 커지면 비복원추출에 따른 종속성이 약해지고, 극단적으로 $N \to \infty$이면 종속성을 완전히 무시할 수 있게 된다. 그런데, $N$이 크더라도 $n$도 크다면 상황이 달라진다. 위의 예에서, $n = 2000$이라 하고 차례로 뽑히는 요소를 $X_1, X_2, \cdots, X_{2000}$이라 하자. 그러면, $P(X_2 = x \mid X_1 \neq x)$ 와 $P(X_2 = x \mid X_1 = x)$는 여전히 각각 1000/4999 와 999/4999 이다 ( $x = 1,3,5,7,9$). 그러나, 첫번째로부터 1999번째에 이르기까지 무엇이 뽑히는가에 따라 $X_{2000}$의 (조건부) 분포는 상당히 영향을 받는다. 극단적인 예로, $i \geq 1001$에 대해서 $P(X_i = x \mid X_1 = x, \cdots, X_{1000} = x) = 0$ 이다 ( $x = 1,3,5,7,9$). 따라서, $N$이 아주 크더라도 $n << N$ 인 경우에 한해서 비복원추출에 따른 종속성을 무시할 수 있다.

이제, 복원추출을 설명한다. 복원추출은 크기가 1인 표본을 반복해서 추출하는 것인데, 매번 추출된 표본을 모집단에 다시 복원(replace)시킨다. 따라서, 매번의 결과는 확률적으로 동일할 뿐더러 (원천적으로) 서로 독립이다. 이는 결국 독립시행의 상황인데, 이를 위의 예제를 통해서 설명한다. $N = 5000$이고 이중에 1,3,5,7,9가 각각 1000개씩 이라고 하자. $n$번에 걸쳐서 추출된 크기가 1인 표본들을 $\{Y_1\}, \{Y_2\}, \cdots, \{Y_n\}$이라 하면, 매번 $N$개에서 하나를 임의로 추출하므로 $Y_1, \cdots, Y_n$의 분포는 모두 모분포와 같다. 또한, $n$개의 표본이 각각 독립적으로 추출되었으므로 $Y_1, \cdots, Y_n$은 서로 독립이다. 예를 들어, $i$번째 결과는 $j$번째 결과에 영향을 받지 않으므로 $P(Y_i = y_i \mid Y_j = y_j) = P(Y_i = y_i) = 1000/5000$ 이다 ( $y_i, y_j = 1,3,5,7,9$). (비고: 복원추출에서는 $n > N$도 가능함.)

이상을 종합하면 다음과 같다. 임의표본 $\{Y_1, \cdots, Y_n\}$에서 $Y_i$의 분포는 모분포와 같다 ( $i = 1, \cdots, n$). 그리고, $n << N$이면 $Y_1, \cdots, Y_n$을 $iid$ 확률변수로 취급할만

한데, 이는 마치 비복원추출을 복원추출로 간주하는 것과 같다.

## §1.5 표본분포

크기가 $N$인 모집단의 임의요소를 $Y$라 하고, 편의상 $Y$의 분포를 모분포라 한다고 했다. 표본은 모집단의 임의 부분집합인데, 이를 $\{Y_1,\cdots,Y_n\}$으로 표현한다고 했다 ($1\le n\le N-1$). 이때, $Y_1,\cdots,Y_n$은 모두 모집단의 임의요소를 의미하므로 이들의 분포는 모두 모분포와 같다고 했다. 그리고, $n<<N$ 이면 $Y_1,\cdots,Y_n$을 *iid* 확률변수로 취급할 수 있다고 했다.

$Y_1,\cdots,Y_n$의 함수를 통계량(statistic)이라 한다. 그러니까, 광의의 statistics는 통계학이지만 협의의 statistic(s)은 통계량(들)이라 할 수 있다. 예를 들어, 표본평균(sample mean)인 $(Y_1+\cdots+Y_n)/n$은 $Y_1,\cdots,Y_n$의 함수이므로 통계량이다. 구체적으로, 추정에 쓰이는 통계량을 추정량(estimator)이라 하고, 검정에 쓰이는 통계량을 검정통계량(test statistic)이라 한다.

$Y_1,\cdots,Y_n$이 확률변수이므로 이들의 함수인 통계량도 확률변수이다. 그리고, 통계량의 (확률)분포를 표본분포(sampling distribution)라 한다.

<비고 1.5.1> 모분포는 모집단의 분포를 일컫지만, 표본분포는 (관찰된) 표본의 분포를 일컫는 표현이 아님 (<비고 1.3.4> 참조).

표본분포는 통계량의 분포이고 통계량은 $Y_1,\cdots,Y_n$의 함수이므로, 표본분포는 $Y_1,\cdots,Y_n$의 결합(joint)분포로부터 얻을 수 있다. 결합분포란 $P(Y_1=y_1,\cdots,Y_n=y_n)$을 의미한다. 그런데, $Y_1,\cdots,Y_n$이 *iid* 확률변수이면 다음과 같은 이점이 있다. 서로 독립이므로 $P(Y_1=y_1,\cdots,Y_n=y_n)$ $=P(Y_1=y_1)\cdot P(Y_2=y_2)\cdots P(Y_n=y_n)$인데, 또한 서로 동일하므로

$$P(Y_1=y_1,\cdots,Y_n=y_n)=\prod_{i=1}^{n}P(Y=y_i) \qquad (1.5.1)$$

가 된다. (물론, $Y$의 분포는 모분포와 같다.)

통계학의 주요과제는 통계적 추론의 목적에 적합한 통계량을 찾은 다음, 그 분포를 (즉, 표본분포를) 구하는 것인데, 이 과정에서 식 (1.5.1)이 핵심적인 역할을 한다. 식 (1.5.1)은 이미 찾아 놓은 통계량의 분포를 구하는 데에만 쓰이는 것이 아니라, 적합한 통계량을 찾는 과정에서도 쓰인다.

<비고 1.5.2> 식 (1.5.1)을 우도함수(likelihood fuction)라 하는데, 앞으로 이를 LF로 표기한다.

## §1.6 MLE 와 LRT

대표적인 추정 방법은 MLE(maximum likelihood estimation)이고, 대표적인 검정 방법은 LRT(likelihood ratio test)인데, 두 방법 모두 LF로부터 얻는다 (<비고 1.5.2> 참조). WMS(문헌 [9])의 연습문제 3-80을 사례로 든다.

<사례 1.3> 관심사는 희귀한 야생동물의 모집단의 크기 $N$이다. 이 동물을 잡을 때마다 꼬리표를 달고 놓아주는데, 이렇게 해서 모두 네마리에 꼬리표를 달았다. 그리고는 얼마 후에 다시 세마리를 잡았더니 그 중 한마리가 꼬리표를 달고 있다고 한다.

모집단의 임의 요소를 나타내는 $Y$를 다음과 같이 정의하자.

$$Y=\begin{cases} 1, \text{ if 꼬리표가 있는 동물} \\ 0, \text{ if 꼬리표가 없는 동물} \end{cases} \tag{1.6.1}$$

임의로 잡은 동물이 꼬리표를 달고 있을 확률은 $P(Y=1)=4/N$이다. 즉, 잡힐 확률은 $N$마리 각각 $N^{-1}$씩인데, 네마리가 꼬리표를 달고 있으므로 이 중의 하나가 걸릴 확률은 $4\times N^{-1}$이다.

표본의 크기는 $n=3$인데, 일단 비복원추출이라고 하자. (이는 세마리를 동시에 잡든지 또는 한마리씩 잡더라도 한번 잡혔던 동물이 다시 잡히면 이를 안 잡은 걸로 친다는 뜻이다.) 임의표본을 $\{Y_1, Y_2, Y_3\}$라 하고, 관찰된 표본을 $\{y_1, y_2, y_3\}$라 하자. $Y_i$의 분포는 모분포와 같으므로 $P(Y_i=1)=4/N$ 이고 $P(Y_i=0)=(N-4)/N$ 이다 ($i=1,2,3$).

통계량 $S=Y_1+Y_2+Y_3$를 정의한다. ($S$는 sum을 의미함.) 또한,

$s=y_1+y_2+y_3$라 하자. $S$는 임의표본에 속한 꼬리표를 단 동물의 수를 의미하고, $s$는 관찰된 표본에 속한 꼬리표를 단 동물의 수를 의미한다.

이 문제에서는 $s$값이 1로 주어져 있다. 그리고 $s$가 1이 되게하는 $\{y_1,y_2,y_3\}$는 유일하게 {1,0,0}이다. (비고: 집합의 요소들의 배열순서는 의미가 없음.) 따라서, LF는 $P(\{Y_1,Y_2,Y_3\}=\{1,0,0\})$인데, 이는 $P(S=1)$과 같다.

$S$의 분포는 초기하(hypergeometric)분포로 알려져 있다 (§2.2.1 참조). 즉,

$$P(S=1)=\binom{4}{1}\binom{N-4}{2}\Big/\binom{N}{3}=\frac{12(N-4)(N-5)}{N(N-1)(N-2)} \qquad (1.6.2)$$

인데, $\binom{N}{3}$은 $N$마리에서 세마리를 뽑는 경우의 수이고, $\binom{4}{1}\binom{N-4}{2}$는 $N$마리에서 세마리를 뽑되 꼬리표가 있는 네마리 중에서 한마리 그리고 꼬리표가 없는 $N-4$마리 중에서 두마리를 뽑는 경우의 수이다.

식 (1.6.2)는 $N$의 함수이다. 이 경우 LF를 $L(N)$으로 표현한다. <표 1.1> 은 몇 가지 $N$값에 대해서 $L(N)$값을 구한 것이다.

<표 1.1> $N$과 $L(N)$

| $N$ | 5이하 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 15 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $L(N)$ | 0 | 0.2 | 0.343 | 0.429 | 0.476 | 0.5 | 0.509 | 0.509 | 0.503 | 0.484 | 0.421 |

결론부터 말하면, $N$에 대한 최우추정치(maximum likelihood estimate)는 11과 12이다. 즉, $L(N)$을 최대가 되게 하는 $N$값이 최우추정치이다. 사실 $N$의 참값은 아무도 모른다. 5이하는 불가능하지만 6이상은 모두 가능하다. (만약 $N=5$이면, 세마리 중 두마리 이상이 꼬리표를 달고 있음.) 다만, 알려진 표본정보(sample

information)는 세마리 중 한마리가 꼬리표를 달고 있다는 것이다. 그러나, (이미 알려지기는 했지만) 이러한 표본정보를 얻게 될 확률은 가능한 $N$값에 따라 다른데, 만약 $N$이 11또는 12라면 그 확률이 최대가 된다. 따라서, 기왕이면 "세마리 중 한마리가 꼬리표를 달고 있을 가능성"이 가장 큰 경우인 11과 12를 $N$에 대한 추정치로 사용하자는 것이다.

LRT에 대한 개요는 다음과 같다. A,B,C 세사람이 각각 $N$이 6,8,11이라고 주장한다고 하자. 그러면, A보다는 B 그리고 B보다는 C의 주장이 설득력이 강할 것이다. 이때, $N=x$라는 주장의 설득력의 강도를 $L(x)/L_{\max}$ 로 표현하는데, $L_{\max}$ 는 $L(11)$ 또는 $L(12)$를 의미한다. 그리고, $L(x)/L_{\max}$가 기준치 이상이면 $N=x$ 라는 주장을 받아들이는데, 이때 기준치는 유의수준(significance level)에 의해 결정된다 (식 (4.1.2) 참조).

이제 더욱 일반적인 상황에 대한 예를 든다. 지금까지는 통계량 $S=Y_1+Y_2+Y_3$를 사용하기로 미리 정해 놓았고, 또한 $s=y_1+y_2+y_3=1$ 이라는 표본정보까지 얻어 놓은 상태에서 MLE와 LRT를 논했다. 일반적인 상황이란 표본추출계획(sampling plan)만 세워 놓은 상황이다. 즉, $N$마리 중에 $m$마리가 꼬리표를 달고 있을 때, $N$마리에서 임의로 $n$마리를 잡아서 꼬리표 유무를 확인하겠다는 것이다. 그러나, 비복원추출에서는 수학적으로 까다롭기 때문에 편의상 복원추출 경우를 예로 든다. (이는 한마리씩 잡아서 꼬리표 유무를 확인하고는 다시 풀어준다는 뜻인데, 이렇게 $n$번을 (독립)시행하면 한마리가 여러번 잡힐 수도 있다.)

복원추출이므로 LF는 식 (1.5.1)에 의해서 $\prod_{i=1}^{n} P(Y=y_i)$이다. 편의상, □ $P(Y=1)=m/N$ □ 과 □ $P(Y=0)=1-(m/N)$ □을 하나로 묶어서 □ $P(Y=y_i)=(m/N)^{y_i}\{1-(m/N)\}^{1-y_i}$ □라 하면 ($y_i=0,1$),

$$L(N)=\prod_{i=1}^{n}P(Y=y_i)=\left(\frac{m}{N}\right)^{\sum_{i=1}^{n}y_i}\left(1-\frac{m}{N}\right)^{n-\sum_{i=1}^{n}y_i} \tag{1.6.3}$$

를 얻는다. 최우추정치 $\hat{N}$은 $L(N)$을 최대가 되게 하는 $N$값인데, 이는 식 (1.6.3)을 미분해서 얻는다. (비고: $N$을 양의 실수로 간주하여 미분을 함. 그러나, 이 문제에서는 $dL(N)/dN$을 0이 되게 하는 $N$값이 자연수로 떨어지므로 더 이상의 손질이 필요 없음.)

$$\hat{N}=mn\Big/\sum_{i=1}^{n}y_i \tag{1.6.4}$$

앞에서와 같이 $m=4, n=3, s=y_1+y_2+y_3=1$이면, $\hat{N}=12$를 얻는다. (비고: 이는 직감적으로도 수긍이 가는 결과이다. 만약 12마리 중에 네마리가 꼬리표를 달고 있다면, 이는 평균적으로 세마리당 한마리가 꼬리표를 달고 있는 것이기 때문이다.)

식 (1.6.4)와 관련해서 두가지 유의할 점이 있다. 첫째로, 식 (1.6.4)에는 $y_1,\cdots,y_n$이 $\sum_{i=1}^{n}y_i$의 형태로만 등장한다. 따라서, 필요한 표본정보가 $s\equiv\sum_{i=1}^{n}y_i$ 임을 알 수 있다. 그런데, 다시 짚고 넘어갈 점은 지금은 표본추출계획만 세워 놓은 상황이라는 점이다. 그러니까, 사실은 관찰된 표본 $\{y_1,\cdots,y_n\}$은 아직 없으며, 또한 $s=\sum_{i=1}^{n}y_i$값도 아직 모른다. 다만, 앞으로 $\{y_1,\cdots,y_n\}$을 얻으면 그때 $\hat{N}=mn/\sum_{i=1}^{n}y_i$를 $N$에 대한 최우추정치로 사용할 계획일 뿐이다.

표본은 추출하기 전까지는(또는, 추출했더라도 그 내용인 $\{y_1,\cdots,y_n\}$을 확인하기 전까지는) 표본을 $\{Y_1,\cdots,Y_n\}$으로 표현한다고 했다(§1.4 참조). 그리고, 표본을 $\{Y_1,\cdots,Y_n\}$으로 표현하면, $s=\sum_{i=1}^{n}y_i$에 대응하는 $S=\sum_{i=1}^{n}Y_i$가 자연스럽게

등장한다. $s=\sum_{i=1}^{n} y_i$를 필요한 표본정보라 불렀는데, 그렇다면 $S=\sum_{i=1}^{n} Y_i$는 바로 우리가 필요한 통계량인 셈이다. 같은 맥락으로, 최우추정치 $mn/s$에 대응하는 최우추정량 $mn/S$를 얻는다.

<비고 1.6.1> 추정량(estimator)은 확률변수이고, 추정치(estimate)는 실수이다. 최우추정량과 최우추정치를 모두 MLE라 부르는데, 이때 MLE의 "E"는 estimation, estimator, estimate 세 가지의 공통 약자이다.

둘째로, 식 (1.6.4)에서 $\sum_{i=1}^{n} y_i$ 값만 아직 관찰되지 않은 것이 아니라, $m$과 $n$도 아직 정해지지 않은 상태일 수 있다. 오히려, $m$과 $n$을 미리 정하는 것보다는 이들을 제어(control)용 모수(parameter)로 활용하는 것이 바람직하다. 최우추정량 $mn/S$의 확률분포로부터 추정의 정확도를 가름할 수 있는데, 이때 정확도를 어느 수준으로 올리기 위해서는 $m$, $n$이 얼마이어야 되는지를 계산할 수 있다. (비고 : $m$과 $n$의 상대적인 크기는 처음 꼬리표를 달 때와 나중에 꼬리표를 확인할 때에 한마리당 드는 비용을 따져서 결정할 수 있을 것임.)

LRT는 다음과 같이 시행한다. 식 (1.6.3)에 "$N=x$"를 대입한 $L(x)$와 $N=\hat{N}$를 대입한 $L(\hat{N})$의 배율인

$$\frac{L(x)}{L(\hat{N})}=\frac{(\frac{m}{x})^{s}(1-\frac{m}{x})^{n-s}}{(\frac{s}{n})^{s}(1-\frac{s}{n})^{n-s}} \tag{1.6.5}$$

가 기준치 이상이면 "$N=x$"라는 주장(또는 가설)을 받아들인다. 그리고, 식 (1.6.5)에

서 $s = \sum_{i=1}^{n} y_i$를 $S = \sum_{i=1}^{n} Y_i$로 대체하면 검정통계량을 얻는다. 사실, 기준치라는 것도 검정통계량의 분포(와 정해진 유의수준)에 의해서 결정되는 것이다. (자세한 내용은 §4.2 참조.)

## §1.7 이책의 구성

지금까지 거론된 통계학의 기본적인 틀을 요약하면 다음과 같다. 통계학은 표본을 가지고 모집단의 성질을 추론하는 것인데, 추론은 크게 추정과 검정으로 나뉜다. 모집단의 임의요소를 $Y$라 하고, $Y$의 분포를 모분포라 한다. 모집단의 임의 부분집합인 표본은 $\{Y_1, \cdots, Y_n\}$으로 표현하는데, $Y_i$의 분포는 모분포와 같다 ($i = 1, \cdots, n$). 그리고, $n << N$인 경우에는 $Y_1, \cdots, Y_n$을 $iid$ 확률변수로 취급한다. $Y_1, \cdots, Y_n$의 함수를 통계량이라 하고, 통계량의 분포를 표본분포라 한다. 추정용 통계량을 추정량이라 하고, 검정용 통계량을 검정통계량이라 한다. 통계량의 분포는 $Y_1, \cdots, Y_n$의 결합분포인 LF로부터 얻는다. 대표적인 추정 방법과 검정 방법은 MLE와 LRT인데, 이들의 근거는 물론 LF이다.

2장에서는 확률분포들을 소개하고 이들의 특성을 요약한다. 3장과 4장에서는 본격적으로 각각 추정과 검정을 다룬다. 이후, 추정과 검정을 묶어서 선형모형(linear model)의 틀로 발전시키는데, 대표적인 ANOVA(analysis of variance : 분산분석)와 회귀분석(linear regression)을 각각 5장과 6장에서 다룬다.

<비고 1.7.1> 이 책에서는 비모수적(non-parametric) 추론과 베이지안(Bayesian) 추론은 다루지 않는다.

# 제 2 장 확률분포



## §2.1 복원추출 관련 분포

### 2.1.1 *Bernoulli* 분포

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

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

로 표현한다 (식 (1.6.1) 참조). 그리고, $P(Y=1)$을 소문자 $p$로, $P(Y=0)$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, $S$는 $n$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 $Y$라 하면, 기하분포인 $Y$의 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, $S$는 $n$번째 성공이 발생할 때까지 시행하는 독립시행의 횟수를 의미한다. $S$의 분포는 다음과 같이 이항분포를 이용해서 구할 수

있다.

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

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

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

### 2.1.5 *Uniform* 분포

*Uniform* 분포는 독립시행과 다음과 같은 관계가 있다. $n$회의 독립시행 중에서 단 한번 성공이 발생했다고 하는 정보가 있을 때, 성공이 발생한 시행이 $i$번째의 ($i=1,2,\cdots,n$) 시행일 확률은 $1/n$로 모두 동일하다. 이는 대칭성에 의한 결과인데, $n$번의 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. $n$번의 시행 중에서 성공하는 횟수를 $B_n$이라 하고, $i$번째의 성공이 발생할 때까지 시행하는 횟수를 $N_i$라 하면, $B_n$과 $N_i$는 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 $p$와 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 $n$번의 시행에서 (성공이 발생하는) $j$번의 시행을 뽑는 경우의 수이다. 반면에, 분자는 $\binom{m-1}{i-1}\binom{1}{1}\binom{n-m}{j-i}$ 인데, 이는 $n$번에서 $j$번을 뽑되 $m-1$번째 이하에서 $i-1$번, $m$번째에서 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 다

항분포를 따른다. 후보 $i$의 실제 득표율을 $p_i$라 하자. 즉, $p_i = 0.42, \cdots, p_5 = 0.015$ 이다. 그리고, 투표함에서 임의로 복원추출한 $n$장의 투표지 중에서 후보 $i$가 득표한 수를 $S_i$라 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, $S_n$을 $n$번째 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 $n$을 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 Type Ⅰ, Type Ⅱ 계획이다.

음초기하분포는 이미 식 (2.1.5)로 등장했다. 자루 속에 탁구공이 $n$개 들어 있는데, $j$개는 흰색이고 $n-j$개는 노란색이라고 하자. 탁구공을 임의로 하나씩 자루에서 꺼내되, 꺼낸 공은 자루에 다시 넣지 않는다. 이때, 흰공을 $i$개 꺼낼때까지 총 $m$개를 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 $i$의 실제 득표수를 $N_i$라 하자. 그리고, 투표함에서 임의로 비복원추출한 $n$장의 투표지 중에서 후보 $i$가 득표한 수를 $S_i$라 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 ($N$은 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 $N$개에서 $n$개를 뽑는 경우의 수이고, $\binom{N_1}{s_1}\cdots\binom{N_5}{s_5}$는 $N$개에서 $n$개를 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 $S_i$의 (marginal : 주변) 분포가 이항분포이고 식 (2.2.2)에서는 $S_i$의 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 $n$은 점점 증가시키고 성공확률 $p$는 점점 감소시키되, 곱 $np$를 일정한 값으로 유지시키면 포아송 분포를 얻는다. 편의상, $dt$ 시간마다 한번씩 시행하여 $t$시간동안 총 $n=t/dt$회 독립시행을 한다고 하자. 그리고, $p=\lambda dt=\lambda t/n$이라 하자. (비고 : $np=\lambda t$.) 그러면, $t$시간 동안 (또는, $n$회 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, $t$는 반드시 시간일 필요는 없으며, 심지어 1차원일 필요도 없다. 예를 들어, $S(t)$는 면적이 $t$인 지역에 서식하는 어떤 식물의 수일 수도 있고, 체적이 $t$인 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 $\lambda dt$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 $t$시간 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 $T$의 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 $t$시간 동안의 결과와 독립이므로

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

를 얻는다. 그리고, 식 (2.4.1)의 좌변을 $dt$로 나눈 것을 $T$의 (확률)밀도함수(density function)라 하고 $f_T(t)$로 표기하는데, 관행상 다음과 같이 모든 $t$에 대해서 정의한

다.

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

### 2.4.3 *Erlang* 분포

음이항분포는 $n$번째 성공이 발생될 때까지의 독립시행 횟수의 분포인 반면에, *Erlang* 분포는 $n$번째 성공이 발생할 때까지 걸리는 시간 $Y=\sum_{i=1}^{n} T_i$의 분포이다. (비고 : $T_1,\cdots,T_n$은 *iid* 지수 확률변수임.) $Y$의 밀도함수를 구하는 방법은 ( $T$의 밀도함수를 구할 때와 동일한데) 다음과 같다.

$$\begin{aligned}P(t<Y<t+dt) &= P[t\text{시간 동안 } n-1\text{번 성공, } (t,t+dt)\text{에서 성공})]\\ &= P[S(t)=n-1]\cdot\lambda dt=\frac{(\lambda t)^{n-1}e^{-\lambda t}}{(n-1)!}\cdot\lambda dt\end{aligned}$$

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

### 2.4.4 감마(*gamma*) 분포

*Erlang* 분포에서 n은 자연수인데, 이를 양의 실수로 간주하면 감마분포가 된

다. 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

$n$회의 독립시행 중에서 단 한번 성공이 발생했다는 조건 하에, 성공이 발생한 시행이 $i$번째의 시행일 확률은 모든 $i$에 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, $0<t<s$) $dt/s$로 동일한데, 이를 연속 *Uniform* 분포라 한다.

이를 정식으로 유도하되 다음과 같이 확장한다.

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

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

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

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

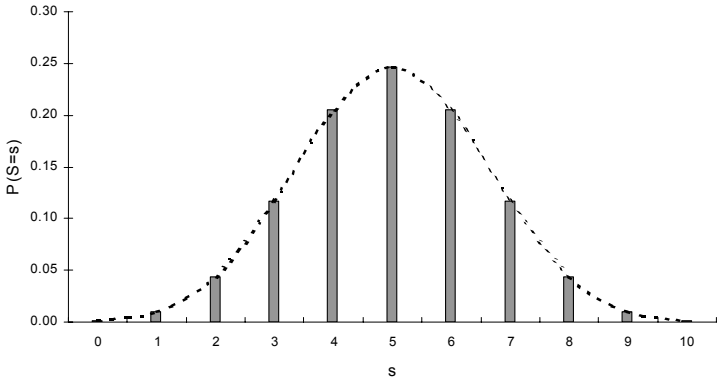
### 2.4.6 베타( *beta*) 분포

*Erlang* 분포에서 자연수 $n$을 양의 실수로 간주하면 감마분포가 되듯이, 식 (2.4.5)에서 자연수 $n_1$, $n_2$를 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.

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

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 $n$이 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 $n$이 커짐에따라 정규분포(의 밀도함수)에 가까워진다.

정규분포(의 밀도함수)는 다음과 같다.

$$f_S(s)=\frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2}\left(\frac{s-\mu}{\sigma}\right)^2},\ -\infty<s<\infty \qquad (2.5.1)$$

정규(분포를 따르는) 확률변수를 $S$라 하면, 식 (2.5.1)에서 $\mu=E(S)$이고 $\sigma^2=V(S)$이다. 즉, $\mu$는 $S$의 기대치이고 $\sigma^2$은 $S$의 분산이다. (따라서, $\sigma$는 $S$의 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 $Z$로 표기하는데, $Z$의 밀도함수는 다음과 같다.

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

<비고 2.5.1> 식 (2.5.1)이 $S$의 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 $d$이다 (§2.15.4 참조).

$C_d$의 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 $d_1$과 $d_2$일 때,

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

는 $F$분포를 따른다. 이때, $d_1$을 분자 자유도라 하고 $d_2$는 분모 자유도라 한다. (밀도함수는 복잡해서 생략함.)

### 2.6.3 $t$ 분포

$Z$와 $C_d$는 각각 표준정규분포와 (자유도가 $d$인) 카이제곱분포를 따르며 서로 독립일 때,

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

의 분포를 (자유도가 $d$인) $t$ 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, $Z^2$은 자유도가 1인 카이제곱분포를 따르므로

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

이다. 즉, $T_d^2$은 분자 자유도가 1이고 분모자유도가 $d$인 $F$분포를 따른다.

<비고 2.6.1> 식 (2.5.5)에서 $d \to \infty$이면 $T_d$의 분포가 $Z$의 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 $n$이 아무리 크더라도 $Y_1,\cdots,Y_n$은 *iid* 확률변수이(고 각각 모분포를 따른)다.

### 2.7.2 연속분포의 표현

편의상, 지수분포를 따르는 $T$와 기하분포를 따르는 $Y$를 예로 든다. $f(t)=\lambda\, e^{-\lambda t},\ t\geq 0$ 을 $T$의 "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 $Y$의 "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 $y$와 $t$에 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 $t\ge 0$에서는 연속적으로 증가한다. 구체적으로, $F_T(t)=P(T\le t)=1-P(T>t)$인데, $t\ge 0$이면 $P(T>t)=P(t$시간 동안 0번 성공$)=P[S(t)=0]=e^{-\lambda t}$이므로

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

이다.

수학적인 정의에 따르면, CDF가 불연속인 $Y$는 이산 확률변수이고, CDF가 연속인 $T$는 연속 확률변수이다. 그리고, 연속 확률변수의 CDF의 미분(또는, 중가율)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,

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

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

<비고 2.7.3> 이산 확률변수인 $Y$의 $dF_Y(y)$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 $q$의 확률로 전구가 터지고, $p(=1-q)$의 확률로 불이 들어온다. 그리고, 불이 들어오는 전구의 수명은 지수분포를 따른다고 하자.

이 전구의 수명을 $T$라 하면, $P(T<0)=0$, $P(T=0)=q$, $P(T>0)=p$ 이다. 그리고, $t>0$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.

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

## §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.

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

이제, $N$명의 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 $N$명의 점수를 모두 합친 것이고, 이를 학생수 $N$으로 나눈 것이 바로 평균 $\mu$ (의 정의) 이다. 반면에, $\sum_y y\cdot P(Y=y)$는 확률변수 $Y$가 구현(realize)될 수 있는 모든 $y$값들에 가중치(weight)인 ( $Y$가 $y$로 구현될) 확률 $P(Y=y)$를 곱해서 모두 더해 놓은 것으로써, 바로 $Y$의 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. $Y$ 가 이산 확률변수이면 $P(Y=y)$ 는 질량함수인데, 질량중심(center of mass)이 평균임을 다음과 같이 보일 수 있다.

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

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

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

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

<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

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

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.

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

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.

$$V(Y)=E(Y^2)-E(Y)^2=p-p^2=pq \qquad (2.8.6)$$

### 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 $\lambda t$ 와 $\lambda tq$ 인데 $q=\lim_{n\to\infty}(1-\frac{\lambda t}{n})=1$ 이므로 평균과 분산이 모두 $\lambda t$ 가 된다.

<비고 2.8.7> 포아송 확률변수 $S(t)$ 의 특징은 $E[S(t)]=V[S(t)]=\lambda t$ 임. 따라서, $\lambda$ 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 초

기하분포의 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: <사례1.3>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. 예

를 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 $n_1 n_2/\{(n_1 + n_2)^2(n_1 + n_2 - 1)\}$ 이다 (문헌 [5] 참조).

정규분포인 식 (2.5.1)은 이미 평균 $\mu$ 와 분산 $\sigma^2$ 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 여전히 $P(S=s)$ 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.

$$P(s<S<s+ds) = P(s<\sigma Z+\mu<s+ds)$$
$$= P(\frac{s-\mu}{\sigma} < Z < \frac{s+ds-\mu}{\sigma})$$

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 $\sigma$ 배로 확대한 다음 $\mu$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 $\sigma$ 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 0에서 $\mu$로 이동하고, 표준편차는 1에서 $\sigma$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 확률변수이면 $P(X=x)=P(Y=y)$이지

만, $Y$ 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 $y= g^{-1}(x)= \ln x$ , $x\geq 0$, 과 $dy/dx= x^{-1}$ 을 대입하면 식 (2.5.1)에 의해서

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

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 $y\geq 0$ 에 대해서만 따짐.) 식 (2.9.2)에 $y= g^{-1}(x)$

$= x^m$, $x \geq 0$, 과 $dy/dx = mx^{m-1}$ 을 대입하면

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

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 $\lambda_i^{-1}$ 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 $(y\geq 0)$, $r(y)=\lambda m y^{m-1}$ 이고 $P(Y>y)=e^{-\lambda y^m}$ 이다.

FR 의 의미는 다음과 같다.

$$r(y)dy=P(y<Y<y+dy \mid Y>y) \qquad (2.10.1)$$

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 $\lambda dy$ 를 얻는다 (§2.4.1 참조). 즉, 지수분포는 FR가 $\lambda$ 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

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

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 같

다.

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

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

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, "$Y_1=y_1$" 이라고 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

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

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.

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

<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.

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

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. ($J$는 Jacobian을 의미하며, $x_1$, $x_2$ 의 함수임.)

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

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 $Z$와 $C_d$는 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 (0에서 $\infty$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 분

포를 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 $n$개의 *iid* 확률변수 중에서 하나를 뽑는 경우의 수이고, $f(y)dy$는 뽑힌 것의 값이 $y$와 $y+dy$ 사이의 값을 가질 확률이며, $\{P(Y \le y)\}^{n-1}$은 나머지 $(n-1)$개가 모두 $y$이하의 값을 가질 확률이다.

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

$$f_{(k)}(y)\,dy \equiv P[Y_1, \cdots, Y_n \text{ 중에서 } k\text{번째로 작은 것의 값이 } (y, y+dy)]$$

$$= P[\, Y_1, \cdots, Y_n \text{ 중에서 } (k-1)\text{개는 } y \text{ 이하, 나머지 중에서}$$

$$\text{하나는 } (y, y+dy), \text{ 남은 } (n-k)\text{개는 모두 } y\text{보다 큼}\,]$$

$$= \binom{n}{k-1} \{P(Y \le y)\}^{k-1} \cdot \binom{n-k+1}{1} f(y)\,dy \cdot \{P(Y > y)\}^{n-k}$$

$$= \frac{n!}{(k-1)!(n-k)!} F(y)^{k-1} \{1-F(y)\}^{n-k} f(y) dy \qquad (2.11.9)$$

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 $\lambda^{-1}$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 $Y$의 MGF는 $E(e^{\Theta Y})$이다. 즉, MGF는 $Y$의 함수인 $e^{\Theta Y}$의 기대치이다. $E(e^{\Theta Y})$를 $\Theta$에 대해서 $n$번 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 $Y$의 $n^{th}$ moment라 함.)

### 2.12.2 기하분포의 평균과 분산

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

$$E(e^{\Theta y}) = \int_{-\infty}^{\infty} e^{\Theta y} dF(y) = \sum_{y=1}^{\infty} e^{\Theta y} q^{y-1} p$$

$$= pe^{\Theta} \sum_{y=1}^{\infty} (qe^{\Theta})^{y-1} = pe^{\Theta}/(1 - qe^{\Theta}) \qquad (2.12.1)$$

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 $Y$의 MGF는

$$E(e^{\Theta Y})=\int_{-\infty}^{\infty} e^{\Theta y} f(y)dy=\int_{0}^{\infty} e^{\Theta y}\lambda e^{-\lambda y}dy=\frac{\lambda}{\lambda-\Theta} \qquad (2.12.2)$$

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 $d_1$과 $d_2$), 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 $d_1+d_2$인) 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 $S$라 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 $e^{\Theta}$를 각각 $z$와 $e^{-\Phi}$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 $Y_1$과 $Y_2$간의 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, $\mu_1$과 $\mu_2$는 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.

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

$$= V(Y_1) + V(Y_2) + 2 \cdot Cov(Y_1, Y_2) \qquad (2.13.3)$$

좀더 일반적인 형태는

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

인데, 예를 들어 $V(Y_1 - Y_2)$는 다음과 같다.

$$V(Y_1 - Y_2) = V(Y_1) + V(Y_2) - 2 \cdot Cov(Y_1, Y_2) \qquad (2.13.5)$$

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

$Y_1$과 $Y_2$가 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. $Y_1$과 $Y_2$가 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, $X_1$의 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 $Y_1$ 과 $Y_2$ 가 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, $Y_1$ 과 $Y_2$ 가 동일한 분포를 따를 때 평균을 $\mu$ 분산을 $\sigma^2$ 이라 하

면, $E(2Y_1)$과 $E(Y_1+Y_2)$는 모두 $2\mu$이다. 그러나, $V(2Y_1)$는 $(2\sigma)^2=4\sigma^2$인 반면에 $V(Y_1+Y_2)$는 $Y_1$과 $Y_2$가 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 $Y_1$, $Y_2$억원이 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, $Y_1$과 $Y_2$가 독립인 경우에는 $V(X_2) = 2V(Y_1)$ 이다.

$V(X_2) > 2V(Y_1)$인 경우 $Y_1$과 $Y_2$는 양의 상관관계에 있다고 하고, $V(X_2)$ $< 2V(Y_1)$인 경우에는 $Y_1$과 $Y_2$가 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 $Y_1$과 $Y_2$의 상관관계를 표준화시킨 것이 상관계수(correlation coefficient)인데, 정의는 다음과 같다.

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

$\rho > 0$이면 $Y_1, Y_2$가 양의 상관관계에 있고, $\rho < 0$이면 $Y_1, Y_2$가 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, $Y_1$과 $Y_2$의 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 $\rho$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 $Y_1, Y_2$는 선형독립이다. 그러나, 선형독립이더라도 비선형(nonlinear)적으로는 독립이 아닐 수 있다 (<비고 2.13.2> 참조).

<비고 2.13.4> $Y_1, Y_2$가 정규분포를 따르는 경우에 $\rho = 0$ 또는 $Cov(Y_1, Y_2) = 0$ 이면 $Y_1, Y_2$는 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, $Y_1, Y_2$가 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 $n$차원 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 $S$를 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 $T$의 조건부 기대치는 $E(T|S=1)=0$, $E(T|S=0)=\lambda^{-1}$ 이다. 그리고, 수명 $T$의 (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, $T$의 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 $T$는 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 $E(T)$, $E(T^2)$, … 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,

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

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

$$P(A) \cdot \{E(Y|A) - E(Y)\}^2 + P(B) \cdot \{E(Y|B) - E(Y)\}^2 \tag{2.14.3}$$

$$=(0.8)\cdot(55-60)^2+(0.2)(80-60)^2=100$$

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 $P(Y_1 \le y_1 | Y_2=y_2)$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 $Y_1$의 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 $Y_1$이 이산 확률변수이면 $P(Y_1=y_1 | Y_2=y_2)$를 의미하고, $Y_1$, $Y_2$가 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 ( $Y_1$과 $Y_2$의) 결합밀도함수 $f(y_1,y_2)$와 $Y_2$의 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 $y_1$과 $y_2$의 함수이지만, ($y_1$을 곱하고) $y_1$에 대해서 적분한 값인 $E[Y_1 | Y_2 = y_2]$는 $y_2$만의 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, $y_2$의 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.

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

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

$$= \int_{y_2} \{E(Y_1 \mid Y_2 = y_2)^2 - E(Y_1)^2\}\, dF_{Y_2}(y_2)$$

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 $y_2$만의 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.

$$V(Y_1) = E[\,V(Y_1 \mid Y_2)\,] + V[\,E(Y_1 \mid Y_2)\,] \tag{2.14.8}$$

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

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

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

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

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 $N$인 모집단의 임의 요소를 $Y$라 하고 $Y$의 분포를 모분포라 한다. 모집단에서 무작위로 비복원추출한 $\{Y_1,\cdots,Y_n\}$을 크기가 $n$인 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 비

복원과 복원 모두에 동일하지만 $V(\overline{Y})$는 각각 $(\sigma^2/n)\{(N-n)/(N-1)\}$과 $\sigma^2/n$인데, 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

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

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

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

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 $y_i$를 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 $n$이 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,

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

이다 (§2.9.1 참조). 즉, $\overline{Y}$역시 점근적으로 정규분포를 따르며 기대치와 분산은 각각 $\mu$와 $\sigma^2/n$이다 (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 $Y_i$의 MGF는

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

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

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

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 $S^2$의 분포

표본평균 $\overline{Y}$ 다음으로 중요한 통계량은 표본분산(sample variance)인데, 관행상

이를 $S^2$으로 표기한다. (지금까지 종종 $S$를 □sum□으로 사용했는데, 이제부터 $S$는 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) $S^2$및 기타 통계량의 분포를 구하기 어렵기 때문이다. (비고: 중심극한정리는 $\overline{Y}$에 대해서만 유효함.)

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

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

로 (일단) 정의한다. 그러나, $\sigma^2$을 몰라서 $S^2$으로 추정하는 상황에서는 $\mu = E(Y)$조차 모르는 것이 더욱 현실적이다. $\mu$를 모르는 경우에는 $\mu$에 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 $n$대신 $(n-1)$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 $d$인 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,

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

이다. 따라서, 식 (2.15.8)의 $S^2$의 함수인 $nS^2/\sigma^2$은 자유도가 $n$인 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 $S^2$과 관련된 분포를 구하는 과정은 약간 복잡한데, 결론부터 언급하면

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

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.

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

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, $t$분포, $F$분포는 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, $\mu$가 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, $Y_1,Y_2,Y_3$를 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 $t$분포의 등장

앞으로 $\mu$에 대한 가설을 검정할 때, $\sigma$가 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, $Y\not\sim N(\mu,\sigma^2)$이면 중심극한정리에 의해서 $(\overline{Y}-\mu)/(\sigma/\sqrt{n})\overset{A}{\sim}N(0,1^2)$이다 (식 (2.15.5) 참조).

그러나, $\sigma$를 모르는 경우가 더 현실적인데, 이때 $\sigma$를 식 (2.15.9)의 $S$로 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 $t$분포를 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 $Z$와 $C_{n-1}$역시 독립이다. 따라서, 이들을 식 (2.5.5)에 대입

하여 다음을 얻는다.

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

<비고 2.15.3> 이때 유의할 점은 다음과 같다. $Y\not\sim N(\mu,\sigma^2)$인 경우 $(\overline{Y}-\mu)/(S/\sqrt{n})$은 $t$분포를 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 $\sigma$를 $S$로 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, $t$분포의 정의상 서로 독립인 $Z$와 $C_{n-1}$이 필요한데, 이는 $Y\sim N(\mu,\sigma^2)$일 때에만 가능하다는 것이다.

### 2.15.6 $F$분포의 등장

모집단이 2개(이상) 있을 때 $F$분포가 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 $F$ 분포를 따른다.

# 제 3 장 추정



## §3.1 비율추정

### 3.1.1 서론

1장에서는 통계학의 기본적인 틀을 개괄적으로 언급했고, 2장에서는 통계학에 필요한 확률이론을 다루었다. 이제부터 정식으로 통계학을 논하겠는데, 3장에서 다룰 주제는 추정(estimation)이다.

추정 중에서는 비율 추정을 먼저 다루는데, 1장에 등장한 <사례 1.1>을 예제로 사용한다. 실제 득표율을 모집단의 비율 또는 줄여서 모비율(population proportion)이라 한다. 예를 들어, 김영삼 후보는 전체 $N$표의 (비고: $N$은 모집단의 크기) 42%인 $.42N$표를 얻었는데 이때 모비율은 0.42이다 (<그림 1.1, 1.2> 참조). 반면에, 추정치(또는 예측치)로는 표본비율(sample proportion)을 사용했다. 표본의 크기 $n$은 약 2000인데, 이를 편의상 2000이라 하면 김영삼 후보는 2000표의 39.5%인 790표를 얻었으므로, 표본비율은 0.395이다.

<사례 1.1>을 예제로 사용하는 이유는 모비율이 알려진 특수한 사례이기 때문이다 (§1.1 참고). 따라서, 추정오차를 정확히 알 수 있다. 예를 들어 김영삼후보의

득표율에 대한 추정오차는 0.395-0.42=-0.025 이다.

그러나, <사례 1.2>와 같은 일반적인 상황에서는 모비율이 알려지지 않는다. 부분의 비율로 전체의 비율을 추정하기 때문에 추정오차는 불가피한데, 모비율을 모르기 때문에 추정오차도 정확히 알 수는 없다. 다만, 추정오차의 확률분포로부터 추정치를 얼마나 신뢰할 수 있는가를 가름할 수 있을 뿐이다. 사실, <사례1.1>에서도 추정치(또는 예측치)를 얻은 시점에서는 모비율이 알려져 있지 않았다. 이에 따라, 추정치를 발표할 때에 "95%의 신뢰수준에서 최대오차는 ±0.022 (또는 ±2.2%)"라고 오차의 범위도 함께 발표했다 (§1.2 참조). 이제, 추정치를 얻은 시점으로 돌아가서 (비고: 모비율이 알려지지 않은 시점임) 오차의 범위에 대한 근거를 알아본다.

### 3.1.2 MLE $\hat{p}$

§1.6에서 MLE를 대표적인 추정방법으로 꼽았다. MLE는 한마디로 "관찰된 표본을 얻게 될 확률을 최대가 되게하는 모비율값"이다.

임의표본을 $\{Y_1,\cdots,Y_n\}$이라 하고 관찰된 표본을 $\{y_1,\cdots,y_n\}$이라 할 때 $P(Y_1=y_1,\cdots,Y_n=y_n)$을 LF(likelihood function)라 불렀다. 즉, LF는 모집단의 특정 부분집합인 $\{y_1,\cdots,y_n\}$이 표본으로 뽑힐 확률이다. $Y_1,\cdots,Y_n$은 모두 모분포를 따르는데, <그림 1.2>에서 편의상 $p_i=P(Y=i)$라 하자. 즉, 모비율을 $p_1,\cdots,p_5$로 표현한다. LF는 $p_1,\cdots,p_5$의 함수인데, LF가 최대가 되게 하는 $p_1,\cdots,p_5$ 값이 바로 MLE이다 (<비고 1.6.1> 참조).

이미 여러번 언급했듯이, 편의상 표본을 복원추출한 것으로 간주한다. 그러면, LF는 다음과 같이 다항분포를 따른다 (§2.1.6 참조).

$$L(p_1,\cdots,p_5)=Kp_1^{n_1}p_2^{n_2}p_3^{n_3}p_4^{n_4}p_5^{n_5} \tag{3.1.1}$$

$$K=n!/(n_1!n_2!n_3!n_4!n_5!) \tag{3.1.2}$$

위의 식에서 $n=\sum_{i=1}^{5} n_i$인데, $n_1,\cdots,n_5$는 후보별 득표수이다 (예 : $n_1=790$). 또한, $1=\sum_{i=1}^{5} p_i$이므로 식 (3.1.1)에서 $p_5$는 $(1-p_1-p_2-p_3-p_4)$와 동일하다.

식 (3.1.1)을 $p_i$에 대해서 ($i=1,2,3,4$) 편미분한 식을 0으로 놓고 풀면 $(p_i/n_i)=(p_5/n_5)$ 을 얻는다. 이로부터 (LF를 최대가 되게 하는 $p_i$값인) 최우추정치

$$\widehat{p_i}=\frac{n_i}{n} \tag{3.1.3}$$

를 얻는데 ($i=1,\cdots,5$), 이것이 바로 표본비율이다 (예 : $\widehat{p_1}=790/2000=0.395$). (비고: 최대값의 충분조건 확인과정은 생략함.)

또한, 식(3.1.3)의 $n_i$를 이에 대응하는 확률변수 $N_i$로 대체하면 최우추정량으로

$$\widehat{p_i}=\frac{N_i}{n} \tag{3.1.4}$$

를 얻는다 (비고: $N_i$는 식 (2.1.6)의 $S_i$와 동일).

<비고 3.1.1> 일반적으로 확률변수는 대문자로 표기하고 (예 : $N_i$) 대응하는 관찰치는 소문자로 표기하지만 (예 : $n_i$), 흔히 추정량과 추정치는 동일하게 표기한다 (예 : $\widehat{p_i}$).

<비고 3.1.2> 식 (3.1.1)에서 $K$를 누락시키더라도 동일한 MLE를 얻음 (식 (1.6.3) 참조). 즉, 복원추출시 추출하는 순서를 따지든 안 따지든 결과는 동일함 (§1.4 참조).

### 3.1.3 오차의 분포

추정오차를 $(\widehat{p}_i - p_i)$로 정의한다. 이에 따라, 오차가 양이면 overestimation을 음이면 underestimation을 의미한다. $\widehat{p}_i$가 최우추정치이면 오차는 상수이다 (예 : $\widehat{p}_1 - p_1 = 0.395 - 0.42 = -0.025$). 반면에 $\widehat{p}_i$가 최우추정량이면 오차는 확률변수가 되므로 확률분포를 거론하게 된다 (<비고 3.1.1> 참조).

식 (3.1.4)에서 $1 = \sum_{i=1}^{5} \widehat{p}_i$ (또는 $n = \sum_{i=1}^{5} N_i$) 이므로, $\widehat{p}_1, \cdots, \widehat{p}_5$는 독립이 아니다. 따라서 $\widehat{p}_1, \cdots, \widehat{p}_5$의 결합분포는 복잡하다. 그러나, 각각의 (주변 : marginal) 분포는 복잡하지 않다. §2.2.3에서 언급했듯이 $N_i$는 이항분포를 따르는데, 기대치는 $np_i$이고 분산은 $np_iq_i$이다 (단, $q_i = 1 - p_i$ : §2.8.7 참조). 그런데, $n = 2000 \gg 1$이므로 중심극한정리를 활용하면 (§2.15.2 참조), $N_i \overset{A}{\sim} N(np_i, np_iq_i)$로 근사할 수 있다. 따라서,

$$\widehat{p}_i = \frac{N_i}{n} \overset{A}{\sim} N\left(p_i, \frac{p_iq_i}{n}\right) \tag{3.1.5}$$

$$\widehat{p}_i - p_i \overset{A}{\sim} N\left(0, \frac{p_iq_i}{n}\right) \tag{3.1.6}$$

$$\frac{\widehat{p}_i - p_i}{\sqrt{p_i q_i / n}} \overset{A}{\sim} N(0,1^2) \tag{3.1.7}$$

을 얻는다 (§2.9.1 참조).

### 3.1.4 오차의 범위

오차 $(\widehat{p}_i - p_i)$가 $\pm\varepsilon_i$이내에 들 확률이 0.95가 되게 하는 $\varepsilon_i$값을 구해보자.

$$\begin{aligned} 0.95 &= P(-\varepsilon_i < \widehat{p}_i - p_i < \varepsilon_i) \\ &= P\left(\frac{-\varepsilon_i}{\sqrt{p_i q_i / n}} < \frac{\widehat{p}_i - p_i}{\sqrt{p_i q_i / n}} < \frac{\varepsilon_i}{\sqrt{p_i q_i / n}}\right) \end{aligned} \tag{3.1.8}$$

이므로, 식 (3.1.7)과 표준 정규분포의 확률표로부터

$$1.96 \approx \varepsilon_i / \sqrt{p_i q_i / n} \tag{3.1.9}$$

를 얻는다.

식 (3.1.9)에 $n=2000$과 알려진 $p_i$값을 대입하면 아래의 결과를 얻는다 ($q_i = 1 - p_i$).

| $i$ | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $p_i$ | 0.42 | 0.338 | 0.163 | 0.064 | 0.015 |
| $\varepsilon_i$ | 0.0216 | 0.0207 | 0.0162 | 0.0107 | 0.0053 |

즉, $p_i$값이 0.5에 가까울수록 $\varepsilon_i$값이 커진다. 그런데, 추정치를 얻은 시점에서는 $p_i$ 값이 알려져 있지 않으므로, 식 (3.1.9)에 $p = q = 0.5$를 대입하여 $\varepsilon$의 최대값인 0.0219를 구한 것이 바로 "최대오차는 ±0.022"라고 언급한 것이다.

### 3.1.5 신뢰구간

"95% 신뢰수준에서 최대오차는 ±0.022"라는 표현 중에서 "최대오차는 ±0.022" 부분은 방금 설명했다. 이제 "95% 신뢰수준에서"부분을 설명한다.

신뢰구간(confidence interval)은 대표적인 구간 추정이다. 반면에, 앞에서 구한 MLE는 대표적인 점(point) 추정이다. 점 추정치인 최우추정치는 상수이고 점 추정량인 최우추정량은 확률변수이듯이, 구간 추정치는 상수이고 구간추정량은 확률변수이다.

구간 추정량은 첨 추정량으로부터 얻는다. 예를 들어, 신뢰수준(confidence level)이 95%인 구간 추정량은 다음과 같이 구한다. 식 (3.1.8)과 (3.1.9)로부터

$$0.95 \approx P(-1.96 < \frac{\widehat{p_i} - p_i}{\sqrt{p_i q_i / n}} < 1.96) \tag{3.1.10}$$

을 얻는데 (비고: "≈"를 사용한 이유는 중심극한정리에 따른 근사식이기 때문인데, 앞으로는 이를 무시하고 "="를 사용함), 우변의 괄호속을 정리하면

$$0.95 = P(\widehat{p_i} - 1.96\sqrt{p_i q_i / n} < p_i < \widehat{p_i} + 1.96\sqrt{p_i q_i / n}) \tag{3.1.11}$$

가 된다. 이때 유의할 점은 다음과 같다. 식 (3.1.10)에서는 부등식의 중간 항이 확률변수인 반면에, 식 (3.1.11)에서는 부등식의 첫 항과 끝 항이 확률변수이다 (비고: 중간

항 $p_i$는 확률변수가 아님).

구간 추정량은 바로 식 (3.1.11)의 두 확률변수를 의미하는데, 이들을 하나로 묶어서

$$\widehat{p}_i \pm 1.96\sqrt{p_i q_i / n} \tag{3.1.12}$$

로 표현한다. 그리고, 구간 추정치는 식 (3.1.12)의 점 추정량 $\widehat{p}_i = N_i/n$을 점 추정치 $\widehat{p}_i = n_i/n$으로 대체하여 얻는다 (<비고 3.1.1> 참조).

그런데, 문제는 $p_i$를 몰라서 추정을 하고 있는 상황이므로, 식 (3.1.12)의 $p_i$와 $q_i$도 물론 모르는 값들이다. 따라서, 두 번째의 근사가 필요하다. (비고: 첫 번째의 근사는 중심극한정리를 사용한 것임.) 이때 점 추정치인 $\widehat{p}_i$로 $p_i$를 대체하는데, 이는 <비고 2.15.3>과 유사한 상황이다.

결과적으로, 95% 신뢰구간이라고 부르는 구간 추정치는 다음과 같다.

$$\frac{n_i}{n} \pm 1.96\sqrt{\frac{n_i}{n}\left(1-\frac{n_i}{n}\right)/n} \tag{3.1.13}$$

예를 들어, $p_1$에 대한 95%신뢰구간은 0.395 ±0.0214 이고, $p_5$에 대한 95%신뢰구간은 0.012±0.0048 이다.

### 3.1.6 신뢰수준

95% 신뢰구간은 식 (3.1.11)로부터 얻었는데, 이때 95% (또는 0.95)는 엄연히 확률이다. 그런데, 이를 확률이라 부르지 않고 신뢰수준이라고 부르는 이유는 다음과 같다.

식 (3.1.10), (3.1.11), (3.1.12)에서는 $\widehat{p}_i$가 확률변수(인 점 추정량)이므로 확률을

운운할 수 있다. 즉, $\widehat{p}_i \pm 1.96\sqrt{p_i q_i / n}$가 상수 $p_i$를 포함할 확률은 0.95이다. 다시 말해서, 확률변수 $\widehat{p}_i - 1.96\sqrt{p_i q_i / n}$은 $p_i$보다 작은 값을 가지고 또한 확률변수 $\widehat{p}_i + 1.96\sqrt{p_i q_i / n}$은 $p_i$보다 큰 값을 가질 (결합) 확률이 바로 0.95이다.

반면에, 식 (3.1.13)은 확률변수가 아니므로 확률을 운운할 수 없다. 예를 들어, $p_1$에 대한 95% 신뢰구간인 0.395±0.0214가 $p_1$을 포함할 확률을 운운할 수는 없다. 만약, $p_1$이 확률변수라면 $p_1$이 0.395±0.0214에 포함될 확률을 운운할 수 있을 것이다. 그러나, 추정 당시 $p_1$=0.42가 알려져 있지 않았을 뿐이지 $p_1$은 어디까지나 상수인 것이다.

<비고 3.1.3> 베이지안 통계학에서는 $p_i$를 확률변수로 취급함 (<비고 1.7.1> 참조).

그럼에도 불구하고, 구간추정치를 얼마나 "신뢰"할 수 있는가를 나타내기 위해서 0.95를 (확률수준이 아니라) "신뢰수준"이라고 부르는 것이다.

그런데, 신뢰구간 및 신뢰수준 등의 표현이 널리 쓰이다 보니까 오히려 일반인에게는 확률이라는 용어 (또는 개념)보다 더 친숙하게 느껴지게 되었다. 따라서, 불필요한 (또는 부적합한) 곳에 까지 신뢰수준이라는 표현을 사용하기도 하는데, 바로 "95% 신뢰수준에서 최대오차는 ±0.022"라는 표현이 그 예이다. (비고: 식 (3.1.8)에서 0.95는 신뢰수준이 아니라 확률임.)

## §3.2 정규 모분포 관련

### 3.2.1 MLE $\hat{\mu}$, $\widehat{\sigma^2}$

지금까지 LF를 $P(Y_1=y_1,\cdots,Y_n=y_n)$으로 정의했는데, 이는 $Y_1,\cdots,Y_n$이 이산 확률변수이었기 때문이다. 반면에 $Y_1,\cdots,Y_n$이 연속 확률변수인 경우에는 이들의 결합밀도함수를 LF로 정의한다 (§2.11.1 참조).

모분포가 정규분포라고 가정하면 <비고 2.7.1>에 의해서 $Y_1,\cdots,Y_n \sim iid\ N(\mu,\sigma^2)$이다 (§2.5.12 참조). 따라서, LF는 다음과 같다 (식 (2.11.2), (2.5.1) 참조).

$$L(\mu,\sigma^2)=\prod_{i=1}^{n} f(y_i)=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{y_i-\mu}{\sigma}\right)^2} \tag{3.2.1}$$

편의상, 식 (3.2.1)에 자연대수(natural log)를 취하면

$$\ln L(\mu,\sigma^2)=K-\frac{n}{2}\ln\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2 \tag{3.2.2}$$

가 되는데, $K$는 미분할 때 0이 되는 항이다. (비고: 표본추출의 순서까지 따져서 식 (3.2.1)에 $n!$을 곱하면 이 또한 $K$에 포함됨.) 대수함수는 1:1 함수이므로 $\ln L(\mu,\sigma^2)$을 최대가 되게 하는 $\mu$와 $\sigma^2$의 값은 $L(\mu,\sigma^2)$도 최대가 되게 한다. 식 (3.2.2)를 $\mu$와 $\sigma^2$에 대해서 편미분한 식을 0으로 놓고 연립으로 풀면 최우추정치로 $\hat{\mu}=\sum_{i=1}^{n} y_i/n \equiv \overline{y}$와 $\widehat{\sigma^2}=\sum_{i=1}^{n}(y_i-\overline{y})^2/n$을 얻는다. (비고: 최대값의 충분조건 확인과정은 생략함.) 그리고, $y_i$와 $\overline{y}$를 각각 $Y_i$와 $\overline{Y}$로 대체하면 아래의 최우추정량을 얻

는다 (<비고 3.1.1> 참조).

$$\hat{\mu} = \overline{Y} \left( = \frac{\sum_{i=1}^{n} Y_i}{n} \right) \tag{3.2.3}$$

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{n} (Y_i - \overline{Y})^2}{n} \tag{3.2.4}$$

다소 비현실적이기는 하지만, 만약 $\sigma^2$이 알려져 있으면 식 (3.2.2)는 $\mu$에 대해서만 미분하면 되는데, 결과는 (3.2.3)과 같다. 반면에 $\mu$가 알려져 있는 경우에는 식 (3.2.2)를 $\sigma^2$에 대해서 미분해서 $\widehat{\sigma^2} = \sum_{i=1}^{n} (y_i - \mu)^2 / n$을 얻는데, 이에 대응하는 최우추정량

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{n} (Y_i - \mu)^2}{n} \tag{3.2.5}$$

은 식 (2.15.8)과 일치한다.

### 3.2.2 $\widehat{\sigma^2}$과 $S^2$

이제 근본적인 문제를 짚고 넘어갈 때가 되었다. 예를 들어, 모분산 $\sigma^2$에 대한 점 추정량으로 식 (3.2.4)의 $\widehat{\sigma^2}$과 식 (2.15.9)의 $S^2$중에서 어느 것을 선택할 것인가 하는 문제이다.

사실 이 문제에 대한 딱부러진 해답은 없는데, 그 이유는 이 문제가 결국 다기준(multi-criteria)의사결정 문제이기 때문이다. 즉, 서로 상충되는 기준들간의 절충이 필요한데, 이때 절충방법은 각자의 선호도에 따라 다를 수 있기 때문이다.

결과부터 언급하면 다음과 같다. (자세한 내용은 §3.3에서 다룸.)

$$E[(\widehat{\sigma^2}-\sigma^2)^2] < E[(S^2-\sigma^2)^2]$$

$$E(\widehat{\sigma^2}-\sigma^2) \neq 0, \quad E(S^2-\sigma^2) = 0$$

즉, 첫 번째 기준인 $\min E[(\text{추정오차})^2]$에 의하면 $\widehat{\sigma^2}$이 낫지만, 두 번째 기준인 $\min|E(\text{오차})|$에 의하면 $S^2$이 낫다.

## §3.3 점추정량의 선택

먼저 용어를 정의한다. 모집단의 특성치인 모비율($p_i$), 모평균($\mu$), 모분산($\sigma^2$) 등을 모수(population parameter)라 하는데, 이들은 물론 모분포를 표현할 때 사용되는 parameter 이기도 하다 (<비고 1.7.1> 참조).

편의상, 모수를 $\theta$로 표현하고, $\theta$에 대한 점 추정량을 $\hat{\theta}$이라 하자. 한마디로 추정오차는 0에 가까울수록 좋다. 그런데, 이 책에서 오차로 정의한 (§3.1.1 참조) $\square\ \hat{\theta}-\theta\ \square$는 확률변수이므로 오차가 0에 얼마나 가까운지는 확률적으로 (또는, 기대치적으로) 운운할 수 밖에 없다.

첫째로, 표본의 크기 $n$이 클수록 $E[(\hat{\theta}-\theta)^2]$또는 $E(|\hat{\theta}-\theta|)$가 점점 작아지다가, 극단적으로 $n\to\infty$이면 0이 되는 경우에 $\hat{\theta}$을 일치(consistent)추정량이라 하는데, 사실상 이 책에 등장하는 모든 추정량이 이에 해당된다.

둘째로, 같은 크기의 표본을 가지고도 $E[(\hat{\theta}-\theta)^2]$또는 $E(|\hat{\theta}-\theta|)$ 중에서 어느 것을 (또는, 제 3의 것을) 기준으로 사용할 것인가 하는 문제가 발생한다. 이는 마치 평균과 중앙값 중에서 어느 것을 모집단의 대표값으로 사용할 것인가 하는 상황과 같다 (§2.8.2, §2.8.3 참조). 모평균 $\mu$는 $E[(Y-y_0)^2]$을 최소가 되게 하는 $y_0$값이고, 중앙값 $m$은 $E(|Y-y_0|)$을 최소가 되게 하는 $y_0$값이므로, $\mu$와 $m$이 모두 나름대로 의미가 있기는 하지만 이미 우리는 은연중에 $\mu$를 선호하고 있다. 특히, 모분포가 정규분포라고 가정하는 상황에서는 정규분포의 parameter인 $\mu$(와 $\sigma^2$)의 사용이 당연시 되고 있다 (<비고 2.8.1> 참조). 이와 유사하게, 관행상 $E[(Y-y_0)^2]$에 대응하는 $E[(\hat{\theta}-\theta)^2]$을 기준으로 사용하는데, 이를 $MSE$(mean square error)라 부른다.

셋째로, $MSE$는 다음과 같이 표현할 수 있다 (식 (2.8.3) 참조).

$$E[(\hat{\theta}-\theta)^2]=V(\hat{\theta}-\theta)+\{E(\hat{\theta}-\theta)\}^2 \qquad (3.3.1)$$

즉, $MSE$= $V$(오차)+ $E$(오차)$^2$인데, 이때 $E$(오차)를 편의(bias)라 부른다. 그리고, $E$(오차)=0인 경우에 $\hat{\theta}$를 불편(unbiased)추정량이라 부른다. 즉, 불편추정량 $\hat{\theta}$의 정의는 다음과 같다.

$$E(\hat{\theta})=\theta \tag{3.3.2}$$

넷째로, $E(\hat{\theta}_1)=E(\hat{\theta}_2)=\theta$ 일 때 $V(\hat{\theta}_1)<V(\hat{\theta}_2)$이면, $\hat{\theta}_1$이 ($\hat{\theta}_2$에 비해서) 상대적으로 효율적(efficient)이라고 한다. (비고: $V(\hat{\theta}_i-\theta)=V(\hat{\theta}_i)$, <비고 2.9.1> 참조.) 그리고, $\theta$에 대한 모든 불편추정량들 중에서 분산이 최소인 것을 MVUE(minimum variance unbiased estimator)라 부른다. 즉, MVUE는 불편추정량이라는 제약(constraint) 하에 $MSE$를 최소화(minimize)하는 추정량이다. 반면에, 아무런 제약없이 $MSE$를 최소화하는 추정량을 min $MSE$ 추정량이라 한다.

투자 대안들(alternatives) 중에서 하나를 선택하는 문제와 비교해 보자. 투자의 위험도(risk)는 따지지 않고 무조건 기대 수익률이 최대인 대안을 선택하는 것은 min $MSE$ 방법과 유사하다. 반면에 정기예금 및 국채와 같이 위험도가 0에 가까운 대안들 중에서 기대 수익률이 최대인 대안을 선택하는 것은 MVUE 방법과 유사하다. 보수적인 의사결정이 보편적으로 선호되듯이, min $MSE$ 방법에 비해서 MVUE 방법이 선호된다.

MVUE를 구하는 방법은 §3.4에서 다룬다. 결과부터 언급하면, 식 (3.1.4)의 $\hat{p}_i=N_i/n$, 식 (3.2.3)의 $\hat{\mu}=\overline{Y}$, 식 (2.15.8)과 (2.15.9)의 $S^2$은 모두 MVUE 이다. 즉, 이미 MLE 방법으로 구한 추정량들 중에서 식 (3.2.4)의 $\hat{\sigma}^2$을 제외한 모든 것이 (결과론적으로) MVUE이다.

그렇다면, MLE는 무엇인가? 최우추정치는 ($\theta$의 함수인) LF를 최소화하는 $\theta$값이고, 최우추정량은 최우추정치의 $y_i$를 $Y_i$로 대체한 것이다. 그러니까, 의사결정의

기준 자체가 다르다. 즉, □ min $MSE$□를 기준으로 사용하는 것이 아니라 □max LF□를 기준으로 사용하는 것이다. MLE 방법을 투자 문제에 비유한다면 □max LF□라는 기준은 나름대로 설득력이 있는 투자지침에 해당 된다. 그리고, 이 투자지침을 따르면 $p$와 μ에 대해서는 MVUE 방법과 동일한 대안을 선택하게 되지만, $\sigma^2$에 대해서는 약간 다른 대안을 선택하게 된다.

아래의 표는 (μ가 알려지지 않은) 정규 모분포의 $\sigma^2$에 대한 추정량 3개를 비교한 것이다. (비고: 편의상 $\sum_{i=1}^{n}(Y_i-\overline{Y})^2$을 $SS$ (sum of squares)라 부르면, 식 (2.15.11)와 <비고 2.8.10>에 의해서 $E(SS)=(n-1)\sigma^2$, $V(SS)=2(n-1)\sigma^4$.)

| 추정량 | $\frac{SS}{n-1}$ | | $\frac{SS}{n}$ | | $\frac{SS}{n+1}$ |
|---|---|---|---|---|---|
| $E$(오차)$^2$ | 0 | < | $\frac{1}{n^2}\sigma^4$ | < | $\frac{4}{(n+1)^2}\sigma^4$ |
| $V$(오차) | $\frac{2}{n-1}\sigma^4$ | > | $\frac{2(n-1)}{n^2}\sigma^4$ | > | $\frac{2(n-1)}{(n+1)^2}\sigma^4$ |
| $MSE$ | $\frac{2}{n-1}\sigma^4$ | > | $\frac{2n-1}{n^2}\sigma^4$ | > | $\frac{2}{n+1}\sigma^4$ |

MVUE인 $SS/(n-1)$은 $E$(오차)=0 이지만 $MSE$는 제일 크다. 반면에, MLE인 $SS/n$의 $MSE$는 MVUE보다 작지만 $E$(오차)≠0 이다. 그리고, $SS/(n+1)$의 $MSE$는 MLE보다도 작지만 |$E$(오차)|는 더 커진다. (비고: $SS/(n+1)$은 $SS$에 $n$의 함수를 곱한 형태의 추정량 중에서 $MSE$가 최소인 것임.)

## §3.4 MVUE 구하는 방법

MVUE는 최소충분(minimal sufficient)통계량의 함수라고 알려져 있다. 그리고, 최소충분통계량을 구하는 방법도 알려져 있다. 그러나, 함수형태는 주먹구구식으로 찾을 수 밖에 없다. 즉, 최소충분통계량의 함수 중에서 불편추정량이 되는 것을 자동적으로 찾아 주는 방법은 없다.

최소충분통계량이란 모수 Θ를 추정하는데 필요한 최소한의 표본정보를 의미하는데 (<비고 1.3.1> 참조), 지금까지 등장한 예는 다음과 같다. §1.6 의 <사례 1.3>에서는 □임의 표본에 속한 꼬리표를 단 동물의 수□이고, §3.1 의 비율 추정에서는 식 (3.1.4)의 □ $N_i$□이며, §3.2 의 모평균 추정에서는 식 (3.2.3)의 □ $\sum_{i=1}^{n} Y_i$□이고 모분산 추정에서는 식 (2.15.9)의 $SS$ 속에 들어 있는 □ $\sum_{i=1}^{n} Y_i$ 와 $\sum_{i=1}^{n} {Y_i}^2$□이다.

<비고 3.4.1> $SS = \sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} {Y_i}^2 - (\sum_{i=1}^{n} Y_i)^2/n$

최소충분통계량의 의미를 표본 $\{Y_1, Y_2, Y_3\}$로 설명한다 (§2.15.4 참조).

$S_i = {Y_1}^i + {Y_2}^i + {Y_3}^i$라 하면, $\{Y_1, Y_2, Y_3\}$를 $\{S_1, S_2, S_3\}$로 대체하더라도 표본에 담긴 정보의 손실은 없다. 그런데, μ를 추정하기 위한 정보는 $\{S_1\}$만으로 충분하다. 즉, 일단 $S_1$이 있으면 $S_2$와 $S_3$는 불필요하다. 또한, $\sigma^2$을 추정하기 위한 정보는 $\{S_1, S_2\}$만으로 충분하다.

MVUE가 최소충분통계량의 함수로 알려져 있듯이, MLE는 충분통계량의 함수로 알려져 있다. 충분통계량에는 꼭 필요한 최소한의 정보 외에 불필요한 정보도 포함될 수 있다. (즉, 최소충분통계량은 충분통계량의 일종의 부분집합이다.) 그러나, 이

는 어디까지나 일반적인 경우일 뿐이고, 지금까지 등장한 MLE 모두 최소충분통계량의 함수이다. 이러한 이유로 MLE가 불편추정량이면 그 자체가 MVUE가 되기도 하고 (예: $\widehat{p}_i$ 와 $\hat{\mu}$), MLE가 불편추정량이 아닌 경우에는 이를 불편추정량이 되도록 손질하여 MVUE를 얻을 수도 있다 (예: $SS/n$의 분모 $n$을 $n-1$로 대체함).

마지막으로, 최소충분통계량을 구하는 방법을 정규 모분포를 예로 들어서 설명한다. 지금까지 임의표본을 $\{Y_1,\cdots,Y_n\}$이라 하고 관찰된 표본을 $\{y_1,\cdots,y_n\}$이라 했는데, 이제 또 다른 관찰된 표본을 $\{y_1',\cdots,y_n'\}$이라 하자. 식 (3.2.1)에 의해서 각각의 LF는

$$L=\prod_{i=1}^{n} f(y_i)=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2}$$

$$L'=\prod_{i=1}^{n} f(y_i')=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i'-\mu)^2}$$

이므로, 이들의 비율은 다음과 같다.

$$\frac{L}{L'}=e^{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n}y_i^2-\sum y_i'^2\right)}e^{-\frac{\mu}{\sigma^2}\left(\sum_{i=1}^{n}y_i-\sum y_i'\right)} \tag{3.3.3}$$

먼저, $\mu$의 값을 변화시키더라도 $\frac{L}{L'}$의 값이 변화하지 않는 필요충분조건을 구하면 $\sum_{i=1}^{n}y_i=\sum_{i=1}^{n}y_i'$을 얻는데, 이를 확률변수로 표현한 $\square\sum_{i=1}^{n}Y_i\square$가 바로 $\mu$에 대한 최소충분통계량이다. 다음, $\sigma^2$의 값을 변화시키더라도 $\frac{L}{L'}$의 값이 변화하지 않는 필요충분조건을 구하면 $\sum_{i=1}^{n}y_i^2=\sum_{i=1}^{n}y_i'^2$과 $\sum_{i=1}^{n}y_i=\sum_{i=1}^{n}y_i'$을 얻는데, 이들을 확률변

수로 표현한 □ $\sum_{i=1}^{n} Y_i^2$ 과 $\sum_{i=1}^{n} Y_i$ □가 바로 $\sigma^2$에 대한 최소충분통계량이다.

## §3.5 MLE 의 속성

### 3.5.1 MLE 와 MVUE

§3.2까지는 MLE를 대표적인 추정방법이라 했는데, §3.3에서 MVUE가 등장하면서부터는 MLE가 밀리는 것같은 인상을 준다. 그럼에도 불구하고 MLE는 여전히 가장 중요한 방법인데 그 이유는 다음과 같다.

첫째로, MLE는 구하기가 쉽다. 즉, LF 만 있으면 이를 미분해서 MLE를 구할 수 있다.

둘째로, $\theta$의 MLE $\hat{\theta}$만 구하면 $\theta$의 함수 $g(\theta)$의 MLE는 자동적으로 $g(\hat{\theta})$가 되는데 (단, $g(\cdot)$은 1:1함수), 이를 MLE의 불변성(invariance) 속성이라 한다. 예를 들어, $\sigma^2$의 MLE가 $SS/n$이면 $\sigma$의 MLE는 $\sqrt{SS/n}$이다.

셋째로, MVUE는 최소충분통계량의 함수라고 했는데, 함수형태에 대한 힌트를 MLE의 형태로부터 얻을 수 있다. 특히, §3.4에서 보았듯이 MLE가 최소충분통계량의 함수인 경우에는 MLE를 불편추정량이 되도록 간단히 손질만 하면 바로 MVUE를 얻는다. 그러니까, MVUE를 원하는 경우에조차 MLE가 도움이 된다.

넷째로, 중심극한정리를 확장하면 모든 MLE에 적용된다 (§2.15.2 참조). 즉, 모든 최우추정량은 점근적으로(asymptotically) 정규분포를 따른다. 더우기, 최우추정량은 점근적으로 불편추정량일뿐더러 분산은 점근적으로 이론적인 하한치와 일치한다 (구체적인 내용은 §3.5.2 참조).

이 속성은 사실 대학원 수준에 가서야 제대로 진가를 발휘하는데, 그 이유는 다음과 같다. 학부 수준인 이 책에 지금까지 등장한 모수인 $p_i$, $\mu$, $\sigma^2$에 대해서는 MVUE 뿐만 아니라 (MVUE의) 확률분포조차 비교적 쉽게 구할 수 있었다. 그러나, 문제가 복잡해지면 MVUE를 구하기 어려운 경우가 발생하기도 하고 또한 MVUE는 구하더라도 그 확률분포를 구하기 어려운 경우가 발생한다. (비고: 확률분포를 알아야 추정치에 대한 신뢰수준을 운운할 수 있음.) 그러나, MLE를 구하기만 하면 정규분포

를 근사분포로 사용할 수 있는데, 이는 MLE의 확률분포를 구하기 어려운 경우 뿐만 아니라 확률분포를 구하더라도 그 분포가 사용하기에 복잡한 경우에도 해당된다. (이는 예를 들어 이항분포를 정규분포로 근사하는 것과 유사하다: §2.15.2 참조.)

나아가서, LF로부터 MLE를 얻는 과정이 분석적으로(analytically) 어려운 경우조차 발생하는데, 이때에도 수치적으로(numerically) 최우추정치를 얻을 수 있을 뿐만 아니라, 최우추정치에 대한 신뢰수준도 거론할 수 있다. 신뢰수준을 거론할 때 (또는 신뢰구간을 구할 때) 사용되는 것은 최우추정량의 점근적 분산인데, 놀라운 사실은 이 점근적 분산이 단순한 근사치가 아니라 이론적으로 밝혀진 하한치라는 점이다.

### 3.5.2 MLE의 점근 분포

모수 $\Theta$의 최우추정량인 $\hat{\Theta}$의 점근(asymptotic) 분포는 다음과 같다.

$$\hat{\Theta} \overset{A}{\sim} N(\Theta,\ I(\Theta)) \tag{3.5.1}$$

즉, $\hat{\Theta}$의 분포는 $n$이 클수록 점점 정규분포에 가까와지는데, 평균은 $\Theta$이고 (따라서, $\hat{\Theta}$은 점근적으로 불편추정량이고), 분산은 관례상 $I(\Theta)$로 표기한다.

$I(\Theta)$는 사실 MVUE 와 관련이 있다. 모수 $\Theta$의 MVUE를 $\tilde{\Theta}$라 하면, $V(\tilde{\Theta})$는 $\Theta$에 대한 모든 불편추정량들 중에서 최소이다. 이때

$$V(\tilde{\Theta}) \geq I(\Theta) \tag{3.5.2}$$

가 성립하는데, 이를 Cramer-Rao 부등식이라 한다. 그러니까, $\tilde{\Theta}$가 MVUE 더라도 □ $V(\tilde{\Theta}) > I(\Theta)$ □가 가능하다. 반면에, 불편추정량 $\tilde{\Theta}$가 MVUE 이기 위한 충분조건

은 □ $V(\tilde{\Theta}) = I(\Theta)$ □이다.

<비고 3.5.1> §3.3에서 $E(\tilde{\Theta}_1) = E(\tilde{\Theta}_2) = \Theta$ 이고 $V(\tilde{\Theta}_1) < V(\tilde{\Theta}_2)$ 일 때 $\tilde{\Theta}_1$ 이 $\tilde{\Theta}_2$ 에 비해서 □상대적으로□ 효율적이라고 했다. 그런데, 만약 $E(\tilde{\Theta}_3) = \Theta$ 이고 $V(\tilde{\Theta}_3) = I(\Theta)$ 이면, $\tilde{\Theta}_3$ 를 (절대적으로) 효율적이라 한다. 이에 따라, 모든 MLE를 점근적으로 효율적(efficient)이라 일컫는다.

$I(\Theta)$ 의 정의는 다음과 같다.

$$I(\Theta) = \frac{-1}{n \cdot E[\frac{\partial^2 \ln f(Y)}{\partial \Theta^2}]} \tag{3.5.3}$$

$f(Y)$ 는 연속 모분포의 밀도함수 $f(y)$ 에서 $y$ 를 $Y$ 로 대체한 것이다. 예를 들어, 정규 모분포의 경우에는

$$\ln f(Y) = K - \frac{1}{2} \ln \sigma^2 - \frac{(Y-\mu)^2}{2\sigma^2} \tag{3.5.4}$$

인데, $K$ 는 미분할 때 0이 되는 항이다 (식 (3.2.2) 참조). $\Theta = \mu$ 인 경우에는 $\frac{\partial^2}{\partial \mu^2} \ln f(Y) = \frac{-1}{\sigma^2}$ 이므로 $I(\mu) = \sigma^2/n$ 을 얻는데, 이는 $V(\overline{Y})$ 와 동일하다. 반면에, $\Theta = \sigma^2$ 인 경우에는

$$\frac{\partial^2}{\partial(\sigma^2)^2}\ln f(Y)=\frac{1}{2\sigma^4}-\frac{(Y-\mu)^2}{\sigma^6}$$

이므로 $E[(Y-\mu)^2]=\sigma^2$임을 활용해서, $I(\sigma^2)=2\sigma^4/n$을 얻는다. (비고: $V(SS/(n-1))=2\sigma^4/(n-1)>I(\sigma^2)$.)

<비고 3.5.2> 식 (3.5.3)의 우변에서 2차 편미분 □$\partial^2\ln f(Y)/\partial\Theta^2$□이 존재하지 않으면 대신 □$-[\partial\ln f(Y)/\partial\Theta]^2$□를 사용할 수 있음.

또한, 식 (3.5.3)은 다음과 같이 확장된다. 모수 또는 모분포의 parameter가 두 개 이상일 때 이들의 최우추정량들 간의 점근적 공분산을 유사한 방법으로 구할 수 있다. 예를 들어, 정규 모분포의 경우에 $\hat{\mu}$와 $\widehat{\sigma^2}$간의 공분산은 점근적으로 다음과 같다.

$$I(\mu,\sigma^2)=\frac{-1}{n\cdot E[\frac{\partial^2\ln f(Y)}{\partial\mu\,\partial(\sigma^2)}]}=0$$

즉, <비고 2.15.2>에 의해서 $\overline{Y}$와 SS가 서로 독립이므로 $\hat{\mu}=\overline{Y}$ 와 $\widehat{\sigma^2}=SS/n$도 서로 독립이다. 따라서 $\hat{\mu}$와 $\widehat{\sigma^2}$간의 공분산은 0인데 이는 물론 $n\to\infty$일 때에도 유효하다.

마지막으로, $I(\Theta)$에서 $I$는 Information을 의미하는데, 이는 $I(\Theta)$가 추정오차에 대한 정보를 제공한다는 뜻이다. 또한, 공분산까지 포함시킨 행렬을 Information matrix 라 하는데, 예를 들면 $\begin{bmatrix} I(\mu) & I(\mu,\sigma^2) \\ I(\sigma^2,\mu) & I(\sigma^2) \end{bmatrix}$이다.

## §3.6 Moment 방법

MVUE와 MLE를 모두 구할 수 없을 때 시도해볼 만한 방법이 MMM(the method of matching moments) 인데, 이는 한마디로 모집단의 $k^{th}$moment 인 $E(Y^k)$와 표본의 $k^{th}$moment 인 $\sum_{i=1}^{n} y_i{}^k/n$을 같다고 놓은 식을 푸는 것이다. (단, $k=1,2,\cdots$ 의 순서로 식을 세우되 필요한 개수만 사용함.) 예를 들면

$$\mu = E(Y) = \sum_{i=1}^{n} y_i/n = \overline{y}$$

$$\sigma^2 + \mu^2 = E(Y^2) = \frac{\sum_{i=1}^{n} y_i{}^2}{n} = \frac{\sum_{i=1}^{n} (y_i - \overline{y})^2}{n} + (\overline{y})^2$$

으로부터 moment 추정치인 $\hat{\mu} = \overline{y}$와 $\widehat{\sigma^2} = \sum_{i=1}^{n} (y_i - \overline{y})^2/n$ 을 얻은 다음, $y_i$를 $Y_i$로 대체해서 moment 추정량인 $\hat{\mu} = \overline{y}$와 $\widehat{\sigma^2} = SS/n$ 을 얻는다.

moment 추정량에 대해서 유일하게 알려진 속성은 일치추정량이라는 것인데 (§3.3 참조), 물론 MVUE와 MLE도 일치추정량이다.

한가지 유의할 점은 다음과 같다. 위에서 moment 추정량 $\hat{\mu} = \overline{y}$와 $\widehat{\sigma^2} = SS/n$ 을 얻는 과정에서 모분포에 대한 아무런 가정을 하지 않았다. 그런데도 결과는 정규 모분포라는 가정하에 얻은 최우추정량과 일치한다. 이는 다음과 같은 가능성을 암시한다. 정규 모분포를 가정하고 얻은 최우추정량들은 정규 모분포라는 가정이 다소 무리가 있더라도 별로 영향을 받지 않는다는 점인데, 이를 MLE의 robustness 속성이라 한다. (즉, MLE는 모분포가 달라지더라도 이에 별로 민감하지 않다는 뜻이다.)

<비고 3.6.1> OR(operations research)의 용어를 빌리면, MVUE와 MLE는 최적해인 반면에 moment 추정량은 heuristic 해이다. 즉, MVUE는 불편추정량이라는 제약하에 $MSE$를 최소화시키는 최적해이고, MLE는 LF를 최대화 시키는 최적해인 반면에, MMM은 단지 일리가 있는 heuristic 방법이라고 할 수 밖에 없다.

## §3.7 신뢰구간

### 3.7.1 모집단 하나의 경우

이 책에는 두 종류의 구간추정이 등장하는데 첫째는 이미 §3.1.5에 등장했던 신뢰구간이고 둘째는 §6.4.7에 등장할 예측구간(prediction interval)이다.

먼저, PQ(pivotal quantity)를 정의한다. 예를 들어, 식 (3.1.10) 우변의 부등식에서 축(pivot)의 역할을 하고 있는 중간 항이 바로 PQ이다. 일반적으로, 모수 $\theta$에 대한 신뢰구간을 구할 때 사용하는 PQ는 확률변수인데 그 분포가 (정확하게 또는 근사적으로) 알려져 있어야 되고, 또한 $\theta$의 함수이어야 된다.

정규 모분포의 모평균 $\mu$에 대한 95% 신뢰구간을 구해보자. PQ로 $Z \equiv (\overline{Y} - \mu)/(\sigma/\sqrt{n})$를 사용하면, $Z \sim N(0,1^2)$이므로

$$0.95 = P(-1.96 < \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} < 1.96)$$

$$= P(\overline{Y} - 1.96\,\sigma/\sqrt{n} < \mu < \overline{Y} + 1.96\,\sigma/\sqrt{n})$$

을 얻는데, $\square\ \overline{Y} \pm 1.96\sigma/\sqrt{n}\ \square$이 바로 $\mu$에 대한 구간 추정량이다. 물론, $\overline{Y}$를 $\overline{y}$로 대체하면 구간 추정치인 95% 신뢰구간을 얻는다. 그러나 이는 모분산 $\sigma^2$이 알려진 경우에나 사용할 수 있다. $\sigma^2$을 모르는 경우에는 관행상 이를 MVUE인 $S^2 = SS/(n-1)$로 대체하는데 (§2.15.5 참조), 이에 따라 식 (2.15.13)의 $T_{n-1} = (\overline{Y} - \mu)/(S/\sqrt{n})$을 PQ로 사용하게 된다. 예를 들어, $n = 10$인 경우에는 $t$ 분포의 확률분포로부터 $0.025 = P(T_9 > 2.262) = P(T_9 < -2.262)$를 얻으므로, 결국 구간 추정량은 $\square\ \overline{Y} \pm 2.262\,S/\sqrt{n}\ \square$이 된다. 그리고, $\mu$에 대한 95% 신뢰구간은

□ $\overline{y} \pm 2.262\, s/\sqrt{n}$ □인데, $\overline{y} = \sum_{i=1}^{n} y_i/n$ 이고 $s^2 = \sum_{i=1}^{n} (y_i - \overline{y})^2/(n-1)$ 이다.

정규 모분포의 분산 $\sigma^2$에 대한 구간 추정시에는 모평균 $\mu$가 알려져 있으면 식 (2.15.10)의 $C_n$을 PQ로 사용하고, $\mu$를 모르는 경우에는 식 (2.15.11)의 $C_{n-1}$을 PQ로 사용한다. 후자의 경우가 더 현실적인데, 예를 들어 $n=10$인 경우에는 카이제곱 분포의 확률표로부터 $0.025 = P(C_9 > 19.0228) = P(C_9 < 2.7004)$를 얻으므로

$$0.95 = P(2.7004 < \frac{SS}{\sigma^2} < 19.0228)$$
$$= P(\frac{SS}{19.0228} < \sigma^2 < \frac{SS}{2.7004})$$

가 된다. (비고: $SS = \sum_{i=1}^{10} (Y_i - \overline{Y})^2$.) 따라서, $\sigma^2$에 대한 95% 신뢰구간은 □ $\sum_{i=1}^{10} (y_i - \overline{y})^2/19.0228$ 에서 $\sum_{i=1}^{10} (y_i - \overline{y})^2/2.7004$ 까지□ 이다.

정규 모분포가 아닌 (경우 또는 모분포를 모르는) 경우에도 중심극한정리를 이용해서 모평균 $\mu$에 대한 신뢰구간을 얻을 수 있는데, 이에 대표적인 사례인 비율 추정을 §3.1.5에서 다루었다. (비고: 이항분포의 parameter인 $p$는 *Bernoulli* 모분포의 평균임. §2.1.1 ,§2.1.2 ,§2.8.6 참조.) 또한, 소위 one-sided 신뢰구간이라는 것이 있는데, (지금까지 등장한 two-sided 신뢰구간에 비해서) 잘 쓰이지 않으므로 이를 생략한다.

### 3.7.2 모집단 두 개의 경우

§2.15.6에서 $F$분포를 등장시킬 때 두 개의 정규 모분포를 가정했는데, 그때 정

의한 확률변수들을 계속 사용한다.

먼저, 식 (2.15.14)의 $(S_1^2/\sigma_1^2)/(S_2^2/\sigma_2^2)$을 PQ로 사용해 보자. 예를 들어, $n_1=10$ 이고 $n_2=5$ 인 경우에는 $F$분포의 확률표로부터 $0.025=P(F_{9,4}>8.90)$ $=P(F_{9,4}<1/4.72)$를 얻으므로 (비고: $P(F_{4,9}>4.72)=0.025$),

$$0.95=P\left(\frac{1}{4.72}<\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}<8.90\right)$$

$$=P\left(\frac{S_2^2}{4.72S_1^2}<\frac{\sigma_2^2}{\sigma_1^2}<\frac{8.90S_2^2}{S_1^2}\right)$$

가 된다. 따라서, $s_1^2=\sum_{i=1}^{10}(y_i-\overline{y})^2/9$, $s_2^2=\sum_{j=1}^{5}(x_j-\overline{x})^2/4$ 라 하면 □$\sigma_2^2/\sigma_1^2$□에 대한 95% 신뢰구간은 □$s_2^2/(4.72s_1^2)$에서 $8.90s_2^2/s_1^2$까지□이다.

다음, $(\mu_1-\mu_2)$에 대한 신뢰구간은 다음과 같이 구한다. <비고 2.15.1>에 의해서 $\overline{Y}\sim N(\mu_1,\sigma_1^2/n_1)$이고 $\overline{X}\sim N(\mu_2,\sigma_2^2/n_2)$인데, $\overline{Y}$와 $\overline{X}$가 서로 독립이므로 다시 <비고 2.15.1>에 의해서

$$(\overline{Y}-\overline{X})\sim N\left(\mu_1-\mu_2,\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)$$

가 된다. 이제, $\sigma_1^2$과 $\sigma_2^2$이 알려진 경우에는 PQ로

$$Z=\frac{(\overline{Y}-\overline{X})-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

를 사용하면 (이후 과정은 앞에서와 동일함), $(\mu_1-\mu_2)$에 대한 95% 신뢰구간으로

$$(\overline{y}-\overline{x})\pm 1.96\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}} \tag{3.7.2}$$

을 얻는다.

그러나, 더욱 현실적인 경우는 $\sigma_1^2$과 $\sigma_2^2$을 모르는 경우인데, 조금 복잡하기는 하지만 어차피 4장 이후에 등장할 내용이므로 예습삼아 다룬다. 이제 정규 모분포라는 가정에다가 □$\sigma_1^2=\sigma_2^2$□이라는 가정을 추가한다. 그러면, $\sigma_1^2=\sigma_2^2$이므로 $\sigma_1^2$과 $\sigma_2^2$을 따로 추정하지 않고 묶어서 한꺼번에 추정하는데, 소위 pooled 추정량이라는 것은 다음과 같다.

$$S^2=\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{(n_1-1)+(n_2-1)} \tag{3.7.3}$$

즉, $S^2$은 $S_1^2$과 $S_2^2$의 가중평균인데, 이때 가중치로는 각각의 자유도를 사용한다.

<비고 3.7.1> 식 (3.7.3)은 식 (2.14.2)의 우변 또는 식 (2.14.8)의 $E[V(Y_1 \mid Y_2)]$에 해당된다. 단, 차이점은 식 (3.7.3)은 표본 통계량이고 $E[V(Y_1 \mid Y_2)]$는 모집단에 관련된 것이라는 점이다.

식 (3.7.1)의 $\sigma_1^2$과 $\sigma_2^2$을 식 (3.7.3)의 $S^2$으로 대체하면 PQ로

$$T_{n_1+n_2-2}=\frac{(\overline{Y}-\overline{X})-(\mu_1-\mu_2)}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}} \tag{3.7.4}$$

을 얻는데, 이는 자유도가 $(n_1+n_2-2)$인 $t$분포를 따른다.

<비고 3.7.2> $(n_1+n_2-2)S^2/\sigma^2$은 자유도가 $(n_1+n_2-2)$인 카이제곱분포를 따르며 (§2.12.6 참조), 정규분포를 따르는 $(\overline{Y}-\overline{X})$와 독립임.

예를 들어, $n_1+n_2=11$ 인 경우에는 $(\mu_1-\mu_2)$에 대한 95% 신뢰구간으로 식 (3.7.2) 대신에

$$(\overline{y}-\overline{x})\pm 2.262\ s\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}$$

을 얻는다. (단, $s$는 식 (3.7.4)의 $S$에 대응하는 표본 관찰치임.)

정규 모분포가 아닌 (경우 또는 모분포를 모르는) 경우에도 중심극한정리를 이용해서 $(\mu_1-\mu_2)$에 대한 근사적 신뢰구간을 얻을 수 있다. 이때, 식 (3.7.1)을 PQ로 사용하는데, 유의할 점은 다음과 같다. $\sigma_1^2$과 $\sigma_2^2$을 모르는 경우에 이들은 식 (3.7.1)에서는 각각의 추정량으로 그리고 식 (3.7.2)에서는 각각의 추정치로 대체하더라도 식 (3.7.1)은 여전히 근사적으로 $N(0,1^2)$이고 (<비고 2.15.3> 참조), 따라서 식 (3.7.2)는 여전히 유효하다. 예를 들어, 모평균이 각각 $p_1$과 $p_2$인 두 개의 *Bernoulli* 모집단에서 크기가 각각 $n_1$과 $n_2$인 표본을 (복원) 추출해서 얻은 최우추정치를 각각 $\widehat{p_1}$과

$\widehat{p_2}$이라 하면, $(p_1 - p_2)$에 대한 95% 신뢰구간은 다음과 같다.

$$(\widehat{p_1} - \widehat{p_2}) \pm 1.96\sqrt{\frac{\widehat{p_1}(1-\widehat{p_1})}{n_1} + \frac{\widehat{p_2}(1-\widehat{p_2})}{n_2}} \tag{3.7.5}$$

즉, 식 (3.7.2)에서 $\overline{y}$와 $\overline{x}$는 각각 $\widehat{p_1}$과 $\widehat{p_2}$에 해당되는데, $\sigma_1^2 = p_1(1-p_1)$과 $\sigma_2^2 = p_2(1-p_2)$는 알려지지 않은 값이므로 이들을 각각 $\widehat{p_1}(1-\widehat{p_1})$과 $\widehat{p_2}(1-\widehat{p_2})$으로 대체한 것이다 (식 (3.1.13) 참조).

# 제 4장 검정



## §4.1 서론 및 용어

1장에서는 통계학의 기본적인 틀을 언급했고, 2장에서는 통계학에 필요한 확률이론을 다루었다. 그리고, 3장부터 정식으로 통계학을 논하고 있는데, 3장에서는 미지의 모수에 대한 추정을 다루었고 4장에서는 미지의 모수에 대한 가설을 검정(檢定)한다.

<비고 4.1.1> 더러 가설검정 (test of hypothesis)을 가설검증으로 표현하기도 하는데, 공인된 용어는 가설검정이다.

통계학에 필요한 확률이론을 2장에서 다루었다고 했으나 실제로 3장에서 쓰인 것은 2장 내용의 일부에 지나지 않는다. 이는 통계학 과목의 전반부에서 다루는 확률이론이 후반부만을 위한 것이 아니라 기타 관련과목 (예: OR, 신뢰성 공학, 생산관리 등)을 위한 준비과정의 역할도 담당하고 있기 때문이다. 4장은 2장보다 오히려 3장과 관계가 깊다. 예를 들어, 3장에서 신뢰구간을 얻을 때 사용했던 PQ(pivotal quantity)가 이제 4장에서는 검정통계량(test statistic)의 역할을 한다. 또한, 3장에서와 같이 4장에서도 LF(likelihood function: 우도함수)가 핵심적인 역할을 한다. §1.6에서 이미

보았듯이, LF가 3장에서는 MLE 및 MVUE에 대한 근거가 되었고, 4장에서는 LRT(likelihood ratio test: 우도비검정)의 근거가 된다.

배심원 평결을 예로 들어서 가설검정의 배경에 깔린 기본적인 틀을 설명하고, 아울러 앞으로 사용할 용어를 정의한다. 첫째로, 가설은 두 가지인데, 귀무(null)가설은 $H_0$로 대립(alternative)가설은 $H_a$로 표기한다. 배심원 평결에서 $H_0$와 $H_a$는 다음과 같다.

$$\begin{array}{l} H_0 : (\text{피고는}) \text{ 무죄} \\ H_a : (\text{피고는}) \text{ 유죄} \end{array} \tag{4.1.1}$$

<비고 4.1.2> 고등학교 과정에서는 흔히 귀무가설 하나만 등장하는데, 이때 대립가설은 □ $H_a$ : $Not\ H_0$□이다.

<비고 4.1.3> 책에 따라 $H_a$를 $H_1$으로 표기하기도 한다.

재판은 피고 (또는 피고측 변호사)가 피고의 무죄를 입증하기 위해서 하는 것이 아니라, 원고 (또는 검사)가 피고의 유죄를 입증하기 위해서 하는 것이다. 따라서, 재판이 없으면 유죄판결도 있을 수 없으며, 또한 최종 판결이 내려질 때까지는 피고를 죄인취급하지 않는다. 일반적으로, 비용과 시간을 들여서 표본을 추출하고 이를 근거로 가설검정을 하는 이유는 귀무가설을 입증하기 위한 것이 아니라 대립가설을 입증하기 위한 것이다. 즉, 구태여 검정을 하지 않더라도 통념적으로 (또는 보편적으로) 받아들여지는 것이 귀무가설이다. 예를 들어, 품질관리에서 기계의 정상적인 가동상태는 귀무가설로, 비정상 가동상태는 대립가설로 설정한다.

배심원 평결의 결과는 □무죄평결□과 □유죄평결□의 두 가지 인데, 무죄평결을 □귀무가설을 채택□(accept)한다고 하고 유죄평결을 □귀무가설을 기각□(reject)한다고 표현한다.

<비고 4.1.4> □대립가설을 채택□한다는 표현은 잘 쓰이지 않음.

배심원 평결은 물론 100% 정확한 것이 아니다. 두 가지의 오류가 가능한데, 제 1종 오류(type Ⅰ error)라 불리는 것은 □무죄인 피고에게 유죄평결을 내리는 것□이고, 제 2종 오류(type Ⅱ error)라 불리는 것은 □유죄인 피고에게 무죄평결을 내리는 것□이다. 그리고 이러한 오류를 저지를 확률을 각각 α와 β로 표기한다. 즉,

$$\begin{aligned} \alpha &= p(\textit{reject } H_0 \mid H_0 \textit{ in true}) \\ \beta &= p(\textit{accept } H_0 \mid H_0 \textit{ in false}) \end{aligned} \tag{4.1.2}$$

인데, 이때 α를 유의수준(significance level)이라 한다.

α와 β는 작을수록 좋다. 배심원 평결에서는 물적 증거와 증언이 많을수록 그리고 가설검정에서는 표본의 크기 $n$이 클수록 α와 β가 작아진다. 그러나 표본의 크기가 일정할 때에는 α를 감소시키면 β가 증가하고 반면에 β를 감소시키면 α가 증가한다. 극단적으로, 모든 피고에게 무죄평결을 내리면 $\alpha = 0$이지만 이때 β는 최대가 된다. 반면에, 모든 피고에게 유죄평결을 내리면 $\beta = 0$이지만 이때 억울한 사람이 가장 많아진다. 따라서, 표본의 크기가 일정할 때에는 α와 β간의 절충이 필요한데, 보통 α를 더 작게 잡는다. (예: 10명의 죄인에게 무죄평결을 내리는 한이 있더라도 한명의 억울한 사람은 없도록 한다는 것이 현대 민주국가의 법철학임.)

그런데, 일반적인 가설검정의 상황에서는 β를 계산하기 어렵거나 또는 β를 딱 부러지게 정의하기가 어려운 경우가 대부분이다. 따라서, 관행상 α를 먼저 책정하는데, 주로 1%, 5%, 10%의 세가지 중에서 하나를 선택한다.

<비고 4.1.5> 유의수준으로 $\alpha = 0.05$가 가장 많이 쓰이는데, 이는 신뢰수준으로 95%가 가장 많이 쓰이는 것과 같은 맥락임.

증거 및 증언은 표본정보에 해당된다고 했는데, 이제 관계법규 및 판례에 해당되는 검정통계량이 필요하다. 또한, 예산과 시간적인 제약 때문에 효율적인 재판의 진행이 요구되는데, 이것이 바로 우리가 사용할 LRT에 해당된다. 즉, LRT방법은 필요한 검정통계량을 제공해 줄 뿐만 아니라 주어진 $n$과 $\alpha$에 대해서 $\beta$를 최소가 되게 하는 효율적인 검정법으로 알려져 있다.

<비고 4.1.6> "$1-\beta$"를 검정력(power of test)이라 한다. 즉, 검정력은 "틀린 귀무가설을 기각할 확률"인데, 주어진 $n$과 $\alpha$에 대해서 검정력이 최대(most powerful)인 검정법이 바로 LRT이다.

마지막으로, 유의할 점은 검정결과에 대한 해석이다. 피고가 무죄인지 유죄인지는 (본인 외에는?) 아무도 모른다. 다만, 관계법규 및 판례에 비추어 볼 때 증거 및 증언이 피고에게 충분히 (또는, 결정적으로) 불리하면 유죄판결을 내린다. 반면에, 증거 및 증언이 피고에게 (유리한 경우뿐만 아니라) 다소 불리하더라도 "증거불충분"인 경우에는 무죄판결을 내린다. 마찬가지로, 귀무가설이 참인지 아닌지는 아무도 모른다. 따라서, 귀무가설을 기각할 때에는 "귀무가설을 기각할 만한 충분한 근거가 (표본정보에 담겨) 있음"이라 하고, 귀무가설을 채택할 때에는 "귀무가설을 기각할 만한 충분한 근거가 없음"이라 표현한다.

## §4.2 LRT

β를 계산하기 쉬운 문제를 예로 들어서 LRT를 설명한다. 다음날 종합주가지수가 올라갈지 내려갈지를 잘 알아맞히기로 소문난 점쟁이가 있는데, 본인의 주장에 의하면 적중률이 70%라고 한다. 이를 확인하기 위해서 앞으로 열흘간에 걸쳐서 몇번을 맞추는지 알아보기로 하자. 적중률에 대한 가설은 다음과 같이 설정한다.

$$\begin{aligned} H_0 &: \ p=0.5 \\ H_a &: \ p=0.7 \end{aligned} \tag{4.2.1}$$

확률변수 $Y_i$를 다음과 같이 정의한다 ( $i=1,\cdots,10$ ).

$$Y_i = \begin{cases} 1, & \text{if } i\text{번째 날에 맞춤} \\ 0, & \text{if } i\text{번째 날에 못 맞춤} \end{cases}$$

$Y_1,\cdots,Y_{10}$은 독립이 아닐 가능성이 다분히 있지만, 편의상 독립으로 간주한다. 그러면, $X=\sum_{i=1}^{10} Y_i$는 이항분포를 따르므로, LF는 $P(X=x)=\binom{10}{x}p^x q^{10-x}$이다 (단, $q=1-p$). 이제, $L_0$와 $L_a$를 다음과 같이 정의한다.

$$\begin{aligned} L_0(x) &= P(X=x \mid H_0 \text{ is true}) = \binom{10}{x}(0.5)^{10} \\ L_a(x) &= P(X=x \mid H_a \text{ is true}) = \binom{10}{x}(0.7)^x(0.3)^{10-x} \end{aligned} \tag{4.2.2}$$

아래의 표는 모든 가능한 $x$값에 대해서 $L_0(x)$와 $L_a(x)$ 그리고 이들의 비율인 $L_0(x)/L_a(x)$를 구한 것이다.

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $L_0(x)$ | .001 | .010 | .044 | .117 | .205 | .246 | .205 | .117 | .044 | .010 | .001 |
| $L_a(x)$ | .000 | .000 | .002 | .009 | .036 | .103 | .200 | .267 | .234 | .121 | .028 |
| $\frac{L_0(x)}{L_a(x)}$ | 165 | 70.9 | 30.4 | 13.0 | 5.58 | 2.39 | 1.02 | .439 | .188 | .081 | .035 |

$L_0(x)$는 $x=5$일 때 최대이고, $L_a(x)$는 $x=7$일 때 최대이다. 반면에 LR(likelihood ratio)인 "$L_0(x)/L_a(x)$"는 $x$가 증가함에 따라 계속 감소한다.

LRT는 다음과 같다.

$$Reject\ H_0 \quad \text{if} \quad \frac{L_0(x)}{L_a(x)} < k \tag{4.2.3}$$

$k$는 α값에 의해 정해지는 양의 상수인데, 이러한 판정법(decision rule)을 따르면 β가 최소가 된다고 알려져 있다. 그런데, 이 문제에서는 $x$가 증가함에 따라 LR이 계속 감소하므로 판정법은 결국 $x$가 어떤 기준치보다 크면 귀무가설을 기각하고, 기준치 이하이면 채택하라는 것이 된다. 사실 이는 상식적으로 납득이 가는 당연한 결과이다.

<비고 4.2.1> 식 (4.2.2)에서 $\binom{10}{x}$를 누락시키더라도 식 (4.2.3)에는 변화가 없음. 즉, 복원추출시 추출하는 순서를 따지든 안 따지든 LRT의 결과와 동일함 (<비고 3.1.2> 참조).

그러나, 문제가 복잡해지면 일일이 위와 같은 표를 만들어서 LR의 값을 직접 눈으로 확인하기가 어렵다. 따라서, 일반적인 문제에서는 식 (4.2.3)로부터 구체적인 판정법을 얻어야 되는데, 다음과 같이 시범을 보인다. (표본의 크기를 $n$, 귀무가설을 $H_0$: $p=p_0$, 대립가설을 $H_a$: $p=p_a$라 함. 단, $p_0<p_a$이고 $q_0=1-p_0$, $q_a=1-p_a$임.)

$$\frac{L_0(x)}{L_a(x)}=\frac{\binom{n}{x}p_0^x\,q_0^{n-x}}{\binom{n}{x}p_a^x\,q_a^{n-x}}=\left(\frac{p_0q_a}{p_aq_0}\right)^x\left(\frac{q_0}{q_a}\right)^n<k$$

$$\rightarrow\left(\frac{p_0q_a}{p_aq_0}\right)^x<k'=\frac{k}{(q_0/q_a)^n}$$

$$\rightarrow x\ln\left(\frac{p_0q_a}{p_aq_0}\right)<k''=\ln k'$$

$$\rightarrow x>k'''=\frac{k''}{\ln\left(\frac{p_0q_a}{p_aq_0}\right)} \qquad (4.2.4)$$

(비고: 식 (4.2.4)에서 부등호의 방향이 바뀌는 이유는 $p_0<p_a$, $q_a<q_0$ 이므로 $(p_oq_a/p_aq_0)<1$ 이고 따라서 $\ln(p_0q_a/p_aq_0)<0$ 이기 때문임.) 즉, $x$값이 상수 $k'''$보다 크면 귀무가설을 기각한다는 판정법이 LRT로부터 도출된다. 아울러, (4.2.4)의 $x$에 대응되는 확률변수 $X$가 검정통계량이 되는데, 이는 바로 최소충분통계량이다 (§3.4 참조).

<비고 4.2.2> 식 (4.2.4)를 기각역(rejection region)이라 함.

이제, 식 (4.2.4)의 $k'''$값을 변화시킴에 따라 α와 β가 어떻게 달라지는지 알아보자. 식 (4.1.2)에 의해서

$$\begin{aligned} \alpha &= P(X > k''' \mid p=0.5) = \sum_{x=k'''+1}^{10} \binom{10}{x} (0.5)^{10} \\ \beta &= P(X \le k''' \mid p=0.7) = \sum_{x=0}^{k'''} \binom{10}{x} (0.7)^{x} (0.3)^{10-x} \end{aligned} \tag{4.2.5}$$

인데, 아래의 표는 $k'''=0,\cdots,10$에 대해서 α와 β를 계산한 것이다.

| $k'''$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| α | .999 | .989 | .945 | .828 | .623 | .377 | .172 | .055 | .011 | .001 | .000 |
| β | .000 | .000 | .002 | .011 | .047 | .150 | .350 | .617 | .851 | .972 | 1.00 |

예를 들어, "$x>6$"을 기각역으로 사용하면 (비고 <4.2.2> 참조) $\alpha=0.172$ 이고 $\beta=0.350$이다. 그러나, α를 더 줄이기 위해서 "$x>7$"을 기각역으로 사용하면 $\alpha=0.055$이지만 $\beta=0.617$이 된다.

역시, 문제가 복잡해지면 위와 같은 표를 만들기가 어렵다. 예를 들어, 위의 문제에서 $n=100$이라 하자. 즉, 앞으로 100일 간에 걸쳐서 점쟁이가 몇 번을 맞히는지 알아본다고 하자. 이 경우, 중심극한정리를 이용해서 $X$의 분포(인 이항분포)를 정규분포로 근사하면

$$Z \equiv \frac{X-np}{\sqrt{npq}} \overset{A}{\sim} N(0,1) \tag{4.2.6}$$

이 된다. 일반적인 문제에서는 α 값을 먼저 책정한다고 했다. 이제, α = 0.05 로 했을 때 (<비고 4.1.5> 참조) 기각역이 어떻게 되는지 알아보자. 식 (4.2.5)와 (4.2.6)에 의해서

$$0.05 = P(X > k''' \mid p = 0.5)$$
$$= P\left(Z > \frac{k''' - (100)(0.5)}{\sqrt{(100)(0.5)(0.5)}}\right)$$

인데, $P(Z > 1.645) \approx 0.05$ 이므로 $k''' \approx 58.225$ 를 얻는다. 따라서, 유의수준 5% 에서의 기각역은 $x \geq 59$ 이다. 즉, 100일 중에 59번 이상을 맞추면 귀무가설을 기각하고, 58번 이하를 맞추면 귀무가설을 채택한다. 이때 β 값은 식 (4.2.5)와 (4.2.6)에 의해서 다음과 같다.

$$\beta = P(X \leq 58 \mid p = 0.7)$$
$$= P\left(Z \leq \frac{58 - (100)(0.7)}{\sqrt{(100)(0.7)(0.3)}} = -2.62\right)$$
$$\approx 0.0044$$

만약, 유의수준 α 를 1% 로 책정했더라면 기각역은 $x \geq 62$ 가 되고 β 값은 2.5% 가 된다.

## §4.3 검정의 종류

먼저 모수공간(parameter space)과 표본공간(sample space)을 정의한다. 모수공간이란 미지의 모수가 가질 수 있는 모든 값들의 집합이다. 예를 들어, 식 (4.2.1)에서는 모수가 $p$ 하나인데, 그나마 $p=0.5$ 와 $p=0.7$ 두가지의 값만 고려하였으므로 모수공간은 $\{0.5,0.7\}$ 이다. 그러나, 일반적인 비율 검정 문제에서는 모수공간이 $\{p: 0\leq p\leq 1\}$ 이다. 또한, 정규 모분포의 경우 모수공간은 일반적으로 2차원 공간인 $\{\mu,\sigma^2: -\infty<\mu<\infty,\sigma^2>0\}$ 이다.

반면에, 표본공간은 $n$ 차원 공간인데, 관찰된 표본 $\{y_1,\cdots,y_n\}$ 은 $n$ 차원 공간에서 하나의 점에 해당된다. 그리고, 표본공간을 두 부분으로 나누어서 그 중 하나를 기각역으로 사용하는 것이다. 그런데, 이 $n$ 차원 공간을 좌표변환하면 기각역이 간단히 표현되기도 하는데, 예를 들어 §4.2에서는 기각역이 $x=\sum_{i=1}^{n} y_i$ 하나로만 표현되었다 (§3.4의 "최소충분통계량의 의미" 참조).

모수공간의 정의에 따라 검정의 종류가 달라지는데, 이를 §4.2의 예제로 설명한다. 귀무가설은 "$p=0.5$"이고 대립가설은 "$p=0.7$"인데, 이러한 가설을 단순(simple) 가설이라 한다. 만약에, 대립가설이 "$p>0.5$"였다면 이를 복합(composite) 대립가설이라 하는데, 이 경우 모수공간은 $\{p: p\geq 0.5\}$ 가 된다. 이때 유의할 점은 다음과 같다. 식 (4.2.4)는 $p_a=0.7$ 뿐만아니라 0.5 보다 큰 모든 $p_a$ 값에 대해서 성립한다. 따라서, 대립가설이 "$p>0.5$"이었더라도 기각역은 여전히 동일하다. (예: $n=100$, $\alpha=0.01$ 이면 기각역은 $x\geq 62$.) 다만, $\beta$ 값을 계산하기가 애매해진다. 대립가설이 단순가설일 때에는 $\beta$ 가 하나의 값이지만, 이제는 $p_a\in H_a$ 인 모든 $p_a$ 에 대해서 $\beta$ 값이 하나씩만 있으므로 $\beta$ 는 $p_a$ 의 함수가 된다. 이를 $\beta(p_a)$ 로 표현하면, 예를 들어 $n=100$, $\alpha=0.01$ 일 때 $\beta(0.7)=0.025$ 이다.

<비고 4.3.1> 식 (4.2.4)는 $p_0$보다 큰 모든 $p_a$에 대해서 성립하는데, 각각의 $p_a$에 대해 LRT에 의한 $\beta(p_a)$가 최소이므로 (<비고 4.1.6> 참조), 복합 대립가설 "$H_a$: $p > p_0$"의 경우에는 LRT를 UMPT(uniformly most powerful test)라 부른다. 이때 "uniformly"는 ($p_0$보다 큰) "모든 $p_a$에 대해서 골고루" 라는 뜻이다.

식 (4.2.4)와 같이 기각역이 "$x > c$" 형태인 경우를 UTT(upper-tail test)라 하는데, 이는 검정통계량인 $X$ (또는 식 (4.2.6)의 $Z$)의 분포에서 우측 꼬리부분이 기각역에 해당되기 때문이다. 반면에, 만약 귀무가설은 여전히 "$p = p_0$"이나 대립가설이 "$p < p_0$"였다면 기각역은 "$x < c$" 형태가 되는데, 이 경우는 LTT(lower-tail test)라 한다. 그리고, UTT와 LTT를 합쳐서 OTT(one-tail test)라 한다.

지금까지 등장한 OTT는 UMPT라는 장점이 있다 (<비고 4.3.1> 참조). 반면에, 모수공간이 모든 가능한 모수값을 포함하지 않는다는 단점이 있다. 예를 들어, UTT의 모수공간은 $\{p: p \geq p_0\}$이고 LTT의 모수공간은 $\{p: p \leq p_0\}$이다. 모수공간이 $\{p: 0 \leq p \leq 1\}$가 되는 대표적인 경우는 두가지가 있다. 첫째는 귀무가설을 복합가설로 설정하는 경우인데, 예를 들면 "$H_0$: $p \leq p_0$"이고 "$H_a$: $p > p_0$"인 경우이다. 이때, 등호 "=" 는 반드시 귀무가설에 포함시키는데, 이는 $\alpha$를 계산할 때 $p_0$값을 사용할 수 있게 하기 위해서이다. 물론, (복합 대립가설의 경우에 $\beta$가 $H_a$에 속한 $p$값들의 함수이듯이) 복합 귀무가설의 경우에는 $\alpha$도 $H_0$에 속한 $p$값들의 함수가 된다. 그러나, $p = p_0$일 때 $\alpha$값이 최대가 되는데, 이를 기준치로 사용한다. 예를 들어, $H_0$: $p \leq 0.5$, $H_a$: $p > 0.5$, $n = 100$인 경우에 $\alpha(0.5)$를 0.01로 책정하면 기각역은 $x \geq 62$가 되는데, 0.5보다 작은 $p$값에 대해서는 $\alpha(p) < 0.01$이다.

<비고 4.3.2> 복합 귀무가설 "$H_0$: $p \le p_0$"와 복합 대립가설 "$H_a$: $p > p_0$"의 경우에도 (최소한 이 책에 등장하는 모분포에 대해서는) 식 (4.2.4)의 LRT가 UMPT이다 (<비고 4.3.1> 참조). 따라서, UTT 경우 "$H_0$: $p = p_0$"를 "$H_0$: $p \le p_0$"로 간주해도 별로 무리가 없다.

모수공간을 $\{p: 0 < p < 1\}$이 되게하는 두번째 경우는 "$H_0$: $p = p_0$"이고 "$H_a$: $p \neq p_0$"인 경우이다. 이는 가장 많이 사용되는 형태의 가설검정인데 (<비고 4.1.2> 참조), 결론부터 언급하면 기각역이 "$x > c_1$ 또는 $x < c_2$" 형태라서 이를 TTT(two-tailed test)라 한다. 즉, 검정통계량의 분포에서 양쪽 꼬리부분이 모두 기각역에 해당된다.

TTT역시 LRT로 시행한다. 편의상, $H_a$에 포함된 $p$값을 $p_a$라 부르자. 즉, $p_a \in H_a$인데 $H_a$가 "$p \neq p_0$"이므로 $p_a$는 $p_0$보다 클수도 있고 작을수도 있다. 그런데, 식 (4.2.4)에서 $p_a > p_0$이면 기각역은 $x > c_1$의 형태이고 $p_a < p_0$이면 기각역은 $x < c_2$의 형태가 된다. 따라서, 식 (4.2.5) 대신에 (단, $c_1 > c_2$)

$$\begin{aligned} \alpha &= P(X > c_1 \text{ 또는 } X < c_2 \mid p = p_0) \\ &= P(X > c_1 \mid p = p_0) + P(X < c_2 \mid p = p_0) \end{aligned}$$

를 얻는데, 관행상 $\alpha$를 이등분하여

$$\frac{\alpha}{2} = P(X > c_1 \mid p = p_0) = P(X < c_2 \mid p = p_0) \tag{4.3.1}$$

를 사용한다. 예를 들어, $p_0 = 0.5$ 이고 $n = 100$일 때 $\alpha$를 10%로 책정하면 ($\alpha/2$는

앞에서와 같이 5%가 되므로) $c_1 \approx 50 + 8.225 = 58.225$ 이고 $c_2 \approx 50 - 8.225 = 41.775$ 이다. 따라서, 기각역은 "$x \geq 59$ 또는 $x \leq 41$"이다.

TTT는 가장 많이 사용되는 검정법이지만 UMPT는 아니다 (<비고 4.3.1>, <비고 4.3.2> 참조). 위의 예에서 보았듯이, TTT에서는 표본공간을 세 부분으로 나누어서 그 중 둘을 기각역으로 사용하는데, 이때 α를 이등분해서 기각역 별로 α/2씩 할당한다. 그러니까, OTT와 동일한 검정력을 얻기 위해서는 α값을 두배로 책정해야 된다. 따라서, (동일한 $n$과) 동일한 α값을 사용하면 TTT의 검정력은 OTT보다 못하다.

그러나, 이와 같이 TTT를 OTT와 비교하는 것은 사실 공평하지 못하다. 이들 두 방법은 엄연히 서로 다른 상황에 사용되는 서로 다른 방법일 뿐이지, 각각이 사용되는 각각의 상황에서는 각각 나름대로 최적(optimal)의 방법이라고 할 수 있다.

선거철에 각 정당에서 발표하는 여론조사 결과를 예로 들자. 만약에 정당마다 자기에게 유리한 결과를 발표한다면 이는 의심을 받을만하다. 그 이유는, 여론조사 결과가 마음에 들면 이를 발표하고, 마음에 들지 않으면 이를 묵살할 수도 있기 때문이다. 이와 유사한 현상이 TTT와 OTT간에도 발생할 수 있다. 가설검정의 목적과 상황에 비추어 TTT가 적합하다고 하자. 예를 들어, 보편적으로 받아들여지고 있는 사회적인 통념을 "$H_0$: $p = 0.5$"라 하고, 이 통념이 옳지 않다는 것을 보이기 위해서 "$H_a$: $p \neq 0.5$"인 TTT를 한다고 하자. 앞의 예에서와 같이 $n = 100$, $\alpha = 0.10$이면 기각역은 $x \geq 59$ 또는 $x \leq 41$이다. 그런데, 만약 표본을 먼저 추출해서 $x$값을 미리 확인한 다음에 $x$가 50보다 크면 대립가설을 "$p > 0.5$"로 고치고, $x$가 50보다 작으면 대립가설을 "$p < 0.5$"로 고친다고 하자. 즉, 관찰된 표본정보에 따라서 TTT를 UTT 또는 LTT로 바꾸는 것인데, 이렇게 하면 (동일한 검정력에 대해서는) α가 절반인 0.05로 줄거나 또는 (동일한 $\alpha = 0.10$에 대해서는) 검정력이 증가한다. 그러나, 이는 정당한 방법이 아니라 일종의 사기라고 할 수 있다.

## §4.4 정규 모분포 관련

### 4.4.1 진짜 LRT

지금까지 이 책에서 LRT라 부른 것을 (대부분의) 다른 책에서는 LRT라 부르지 않는다. 이 책에서는 LF(likelihood function)의 비율인 LR(likelihood ratio)에 근거한 검정법은 모두 LRT라 부르고 있지만, 공인된 용어에 따른 진짜 LRT는 §4.4.2에 처음으로 등장한다.

진짜 LRT는 주로 (정규 모분포 경우와 같이) 모수공간이 이차원(이상)인 경우에 사용하는데, LR의 형태는 식 (1.6.5)와 같다. 그러나, 이를 모수공간이 일차원인 경우에도 적용할 수 있다. 예를 들어, §4.2와 §4.3에 등장한 예제에서 식 (4.2.3) 대신에 식 (1.6.5) 형태의 LR를 사용해도 동일한 기각역을 얻는다.

먼저, "$H_0$: $p=0.5$, $H_a$: $p\neq 0.5$"이고 "$n=10$"인 경우를 다룬다. LF는 $P(X=x)=\binom{10}{x}p^x q^{10-x}$인데, $L(p_0)$와 $L(\hat{p})$을 다음과 같이 정의한다 (식 (4.2.2) 참조).

$$\begin{aligned} L(p_0)&=P(X=x \mid p=p_0)=\binom{10}{x}p_0^x\, q_0^{10-x} \\ L(\hat{p})&=P(X=x \mid p=\hat{p})=\binom{10}{x}\hat{p}^x\hat{q}^{10-x} \end{aligned} \tag{4.4.1}$$

식 (4.4.1)에서 $p_0$는 귀무가설 하에서의 $p$값인 0.5이고 ( $q_0=1-p_0=0.5$ ), $\hat{p}$은 $p$에 대한 최우추정치인 $x/10$이다 ( $\hat{q}=1-\hat{p}$ ). 아래의 표는 모든 가능한 $x$값에 대해서 $L(p_0)$와 $L(\hat{p})$ 그리고 이들의 비율을 구한 것이다.

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $L(p_0)$ | .001 | .010 | .044 | .117 | .205 | .246 | .205 | .117 | .044 | .010 | .001 |
| $\hat{p}$ | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
| $L(\hat{p})$ | 1 | .387 | .302 | .267 | .251 | .246 | .251 | .267 | .302 | .387 | 1 |
| $\frac{L(p_0)}{L(\hat{p})}$ | .001 | .025 | .146 | .439 | .818 | 1 | .818 | .439 | .146 | .025 | .001 |

최우추정치 $\hat{p}$는 (정의상) LF를 최대가 되게하는 $p$값이므로, LR인 $L(p_0)/L(\hat{p})$의 최대치는 1이다 (최소치는 0). 그리고, LR이 1에 가까울수록 귀무가설의 설득력이 강하고, 0에 가까울수록 약하다. 따라서, 기각역의 형태는

$$\frac{L(p_0)}{L(\hat{p})} \leq k \tag{4.4.2}$$

가 되는 것이 바람직한데, 이는 §4.3에서 다룬 TTT와 동일한 것이다. 예를 들어, $\alpha = 2 \cdot (0.172) = 0.344$ 이면 기각역은 "$x \geq 7$ 또는 $x \leq 3$"이고, $\alpha = 2 \cdot (0.055) = 0.11$ 이면 기각역은 "$x \geq 8$ 또는 $x \leq 2$"가 된다 (§4.2의 표 참조).

다음, "$H_0: p = 0.5$, $H_a: p > 0.5$"이고 "$n = 10$"인 경우를 다룬다. 이때 유의할 점은 다음과 같다. 모수공간은 $\{p: p \geq 0.5\}$인데, 이는 마치 "$p < 0.5$"가 불가능하다는 전제조건과 같다. 이에 따라, $p$에 대한 최우추정치인 $\hat{p}$에 대해서도 "$\hat{p} \geq 0.5$"만 고려대상이 된다. 구체적으로

$$\hat{p} = \max\left(\frac{x}{10}, 0.5\right) = \begin{cases} \frac{x}{10} & \text{if } x \geq 5 \\ 0.5 & \text{if } x < 5 \end{cases} \tag{4.4.3}$$

인데, 이를 식 (4.4.1)에 대입하면 아래의 표를 얻는다.

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $L(p_0)$ | .001 | .010 | .044 | .117 | .205 | .246 | .205 | .117 | .044 | .010 | .001 |
| $\widehat{p}$ | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
| $L(\widehat{p})$ | .001 | .010 | .044 | .117 | .205 | .246 | .251 | .267 | .302 | .387 | 1 |
| $\frac{L(p_0)}{L(\widehat{p})}$ | 1 | 1 | 1 | 1 | 1 | 1 | .818 | .439 | .146 | .025 | .001 |

따라서, 식 (4.4.2)를 적용하면 자동적으로 UTT가 된다. 예를 들어, α=0.172 이면 기각역은 "$x \geq 7$"이고, α=0.055 이면 기각역은 "$x \geq 8$"이 된다.

마지막으로, "$H_0 \leq 0.5$, $H_a > 0.5$"이고 "$n=10$"인 경우를 다룬다. 모수공간은 다시 TTT에서과 같이 $\{p: 0 \leq p \leq 1\}$이므로, 최우추정치 역시 $\widehat{p} = x/10$ 이다. 그런데, 귀무가설이 복합가설이라서 $L(p_0)$를 계산하기가 애매해진다. §4.3에서는 $p_0$값으로 복합 귀무가설의 경계치인 0.5를 사용한다고 했는데 (<비고 4.3.2> 참조), 이는 어디까지나 단순 귀무가설의 틀에 맞추어 보려는 임시변통이었을 뿐이다. 사실상 이 예제는 §4.4.2에 등장할 진짜 LRT의 범주에 속하는 것인데, 이를 위해서 식 (4.4.2)를 다음과 같이 확장시킨다.

$$\frac{L(\widehat{p_0})}{L(\widehat{p})} \leq k \tag{4.4.4}$$

식 (4.4.4)에서 $\widehat{p_0}$은 복합 귀무가설 하에서의 최우추정치로서 다음과 같다 (식 (4.4.3) 참조).

$$\widehat{p_0} = \min\left(\frac{x}{10}, 0.5\right) = \begin{cases} \frac{x}{10} & \text{if } x \leq 5 \\ 0.5 & \text{if } x > 5 \end{cases} \tag{4.4.5}$$

<비고 4.4.1> 단순 귀무가설 $H_0$: $p = p_0$ 경우 $\widehat{p_0} = p_0$ 인데, 이는 귀무가설 하에서 $p$가 취할 수 있는 값이 $p_0$하나 뿐이기 때문이다. 따라서, 식 (4.4.2)는 식 (4.4.4)의 특수한 경우이다.

아래의 표는 모든 가능한 $x$값에 대해서 $L(\widehat{p_0}) = \binom{10}{x} \widehat{p_0}^{\,x}\, \widehat{q_0}^{\,10-x}$와 $L(\widehat{p}) = \binom{10}{x} \widehat{p}^{\,x}\, \widehat{q}^{\,10-x}$ 그리고 이들의 비율을 구한 것이다.

| $x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\widehat{p_0}$ | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |
| $L(\widehat{p_0})$ | 1 | .387 | .302 | .267 | .251 | .246 | .205 | .117 | .044 | .010 | .001 |
| $\widehat{p}$ | 0.0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 |
| $L(\widehat{p})$ | 1 | .387 | .302 | .267 | .251 | .246 | .251 | .267 | .302 | .387 | 1 |
| $\frac{L(\widehat{p_0})}{L(\widehat{p})}$ | 1 | 1 | 1 | 1 | 1 | 1 | .818 | .439 | .146 | .205 | .001 |

그런데, 유의할 점은 LR인 "$L(\widehat{p_0})/L(\widehat{p})$"의 값이 단순 귀무가설 "$H_0$: $p = 0.5$, $H_a$: $p > 0.5$" 경우의 LR인 "$L(p_0)/L(\widehat{p})$"의 값과 동일하다는 점이다. 따라서, 동일한 UTT가 된다 (<비고 4.3.2> 참조).

### 4.4.2 μ에 대한 검정

정규 모분포의 평균 μ에 대한 검정을 T-test라 하는데, 이는 검정통계량인

$$T_{n-1}=\frac{\overline{Y}-\mu_0}{S/\sqrt{n}} \tag{4.4.6}$$

가 (자유도가 $n-1$인) $t$분포를 따르기 때문이다 (§2.15.5 참조). §4.4.1에서 언급했듯이, 검정통계량 $T_{n-1}$은 식 (4.4.4) 형태의 LRT로부터 얻는다.

먼저, "$H_0$: $\mu=\mu_0$, $H_a$: $\mu\neq\mu_0$" 경우를 다룬다. 이때 유의할 점은, 정규 모분포의 모수공간이 이차원이므로 (§4.3 참조), 사실상 "$\sigma^2>0$"가 $H_0$와 $H_a$에 포함되어 있는 것이나 마찬가지라는 점이다. 따라서, 귀무가설은 단순가설이 아니라 복합가설 "$H_0$: $\mu=\mu_0$ and $\sigma^2>0$"인 셈이다. 귀무가설 하에서의 $\mu$와 $\sigma^2$에 대한 최우추정치는 다음과 같다 (<비고 4.4.1>과 식 (3.2.5) 참조).

$$\widehat{\mu}_0=\mu_0, \quad \widehat{\sigma}_0^2=\frac{\sum_{i=1}^{n}(y_i-\mu_0)^2}{n} \tag{4.4.7}$$

반면에, "귀무가설 하에서"라는 제약(constraint) 없이 구한 최우추정치는 다음과 같다 (§3.2.1 참조).

$$\widehat{\mu}=\overline{y}, \quad \widehat{\sigma}^2=\frac{\sum_{i=1}^{n}(y_i-\overline{y})^2}{n} \tag{4.4.8}$$

식 (4.4.7)과 (4.4.8)을 LF인 식 (3.2.1)에 대입하면

$$L(\widehat{\mu}_0,\widehat{\sigma}_0^2)=\left(\frac{1}{\sqrt{2\pi}}\right)^n\left(\frac{1}{\widehat{\sigma}_0^2}\right)^{\frac{n}{2}}\exp\left(-\frac{n\widehat{\sigma}_0^2}{2\widehat{\sigma}_0^2}\right)$$

$$L(\hat{\mu},\widehat{\sigma^2})=(\frac{1}{\sqrt{2\pi}})^n(\frac{1}{\widehat{\sigma^2}})^{\frac{n}{2}}\exp(-\frac{n\widehat{\sigma^2}}{2\widehat{\sigma^2}})$$

를 얻으므로, 식 (4.4.4) 형태의 LRT는 다음과 같다.

$$\lambda\equiv\frac{L(\widehat{\mu_0},\widehat{\sigma^2})}{L(\hat{\mu},\widehat{\sigma^2})}=\left(\frac{\widehat{\sigma^2}}{\widehat{\sigma_0^2}}\right)^{\frac{n}{2}}\le k \tag{4.4.9}$$

<비고 4.4.2> 식 (4.4.9)에서 "$0\le\lambda\le1$"이다 ($0<k<1$). 그리고, 표본 관찰치 $y_i$를 이에 대응되는 확률변수 $Y_i$로 대체하면 $\lambda$는 확률변수가 되는데, 이를 (그대로 또는 손질해서) 검정통계량으로 사용한다.

식 (4.4.9)를 손질해서 아래의 식 (4.4.10)을 얻는 과정은 (식 (4.2.4)를 얻는 과정과 유사하지만 훨씬) 복잡하므로 생략한다 (문헌 [9] 참조).

$$t^2\equiv\left|\frac{\bar{y}-\mu_0}{\sqrt{\frac{\sum(y_i-\bar{y})^2}{n-1}}/\sqrt{n}}\right|^2\ge k'=(n-1)(k^{-\frac{2}{n}}-1) \tag{4.4.10}$$

따라서, LRT는 TTT로써 기각역은 $t\ge\sqrt{k'}$ 또는 $t\le-\sqrt{k'}$이다. 그리고, 식 (4.4.10)의 $t$에서 $y_i$를 $Y_i$로 대체하면 식 (4.4.6)을 얻는다.

<비고 4.4.3> 식 (4.4.6)의 $T_{n-1}$은 "$\mu=\mu_0$"일 때 자유도가 $n-1$인 $t$분포를 따른다.

예를 들어, $n=10$ 일 때 $0.025=P(T_{n-1}>2.262)=P(T_{n-1}<-2.262)$ 이므로, 유의수준 5% 에서의 기각역은 $t>2.262$ 또는 $t<-2.262$ 이다.

다음, "$H_0$: $\mu=\mu_0$, $H_a$: $\mu>\mu_0$" 경우를 다룬다. 이때 유의할 점은 모수공간이 $\{\mu,\sigma^2:\ \mu\geq\mu_0,\ \sigma^2>0\}$ 이므로 $\hat{\mu}=\max(\mu_0,\bar{y})$ 라는 점이다 (식 (4.4.3) 참조). 이에 따라, 식 (4.4.9)와 (4.4.10) 대신에

$$\lambda=\begin{cases}(\widehat{\sigma^2}/\widehat{\sigma_0^2})^{\frac{n}{2}} & \text{if } \bar{y}\geq\mu_0\\ 1 & \text{if } \bar{y}<\mu_0\end{cases} \tag{4.4.11}$$

$$t=\frac{\bar{y}-\mu_0}{\sqrt{\frac{\sum(y_i-\bar{y})^2}{n-1}}\Big/\sqrt{n}}\geq\sqrt{k'}\ (>0) \tag{4.4.12}$$

를 얻는데, 이는 UTT의 기각역이다. 그리고, 식 (4.4.12)의 $y_i$를 $Y_i$로 대체하면 역시 식 (4.4.6)을 얻는다 (<비고 4.4.3> 참조). 예를 들어, $n=10$ 일 때 $0.05=P(T_{n-1}>1.833)$ 이므로, 유의수준 5% 에서의 기각역은 $t>1.833$ 이다.

마지막으로, "$H_0$: $\mu\leq\mu_0$, $H_a$: $\mu>\mu_0$" 경우를 다룬다. 이때 달라지는 것은 귀무가설 하에서의 $\mu$ 에 대한 최우추정치인데, 이는 식 (4.4.5)와 유사한 형태인 $\widehat{\mu_0}=\min(\mu_0,\bar{y})$ 가 된다. (비고: 식 (4.4.8)은 유효함.) 그런데, 결과론적으로 식 (4.4.11)과 (4.4.12)는 여전히 유효하다. (비고:§4.4.1의 마지막 부분에서도 $L(\widehat{p_0})/L(\hat{p})=L(p_0)/L(\hat{p})$ 이었음.) 즉, 식 (4.4.6)이 여전히 검정통계량이고 또한 UTT가 된다. 다만, 유의수준과 검정통계량은 귀무가설 "$\mu\leq\mu_0$"에 속한 모든 $\mu$ 값들 중에서 경계치인 $\mu_0$를 기준으로 삼은 것으로 간주하면 된다 (<비고 4.3.2> 위의 문단과 <비고

4.4.3> 참조).

### 4.4.3 기타 정규 모분포 관련

진짜 LRT보다 먼저 등장한 UMPT(<비고 4.3.1> 참조)와 식 (4.2.4)까지 모두 LRT라 하면(§4.4.1 참조), 사실상 모든 검정이 LRT라 해도 과언이 아니다. 그런데, LR인 $\lambda$를 손질해서(<비고 4.4.2> 참조) 식 (4.2.6)의 $Z$와 식 (4.4.6)의 $T_{n-1}$ 같이 눈에 익은 검정통계량을 얻어(내고 아울러 기각역의 형태를 알아)내는 과정은 앞에서 보았듯이 제법 복잡하다. 따라서, 앞으로는 LRT의 결과만을 언급하기로 한다.

첫째로, $\sigma^2$이 알려진 경우에는 $\mu$에 대한 검정통계량으로 "$Z=(\overline{Y}-\mu_0)/(\sigma/\sqrt{n})$"을 사용하는데, 이는 물론 $\mu=\mu_0$일 때 $N(0,1^2)$분포를 따른다. 그런데, §4.4.2의 $T$-test가 이제 $Z$-test로 바뀌었으므로 $t$분포의 확률표 대신 $N(0,1^2)$분포의 확률표를 사용하는 것만 달라질 뿐이지, 가설의 형태에 따라 검정의 종류 및 기각역의 형태가 달라지는 양상은 $T$-test 때와 동일하다.

<비고 4.4.4> $N(0,1^2)$의 확률표는 $t$분포의 확률표에서 자유도가 $\infty$인 경우에 해당됨 (<비고 2.6.1> 참조).

다음, $\sigma^2$에 대한 검정통계량으로는 $\mu$가 알려진 경우에는 식 (2.15.10)을 그리고 $\mu$를 모르는 경우에는 식 (2.15.11)을 사용하는데 (단, $\sigma^2$을 귀무가설의 $\sigma_0^2$으로 대체함), 이들은 $\sigma^2=\sigma_0^2$일 때 자유도가 각각 $n$과 $n-1$인 카이제곱분포를 따른다. 그러나, 가설의 형태에 따라 검정의 종류 및 기각역의 형태가 달라지는 양상은 $\mu$에 대한 검정때와 동일하다. 예를 들어, $\mu$를 모르는 경우에 $n=10$이라 하자. 만약 귀무가설이 "$\sigma^2=\sigma_0^2$" 또는 "$\sigma^2\le\sigma_0^2$"이고, 대립가설이 "$\sigma^2=\sigma_a^2$" (단, $\sigma_a^2>\sigma_0^2$) 또는

"$\sigma^2>\sigma_0^2$"이면, 이는 UTT이고 유의수준 5% 에서의 기각역은 $c_{n-1}\geq 16.919$ 이다. 또한, 귀무가설이 "$\sigma^2=\sigma_0^2$"이고 대립가설이 "$\sigma^2\neq\sigma_0^2$"이면, 이는 TTT이고 유의수준 5% 에서의 기각역은 $c_{n-1}\geq 19.0228$ 또는 $c_{n-1}\leq 2.7004$ 이다. (비고: $c_{n-1}$ 은 검정통계량에서 확률변수 $Y_i$를 표본 관찰치 $y_i$로 대체한 것임. 즉, $c_{n-1}$ 은 카이제곱분포를 따르는 검정통계량이 구현된(realized) 값을 의미함.)

<비고 4.4.5> 지금 등장한 카이제곱 검정은 UTT, LTT, TTT가 모두 가능하다. 그러나, §4.6에 등장하는 카이제곱 검정은 모두 UTT이다.

결국 §3.7에서 신뢰구간을 얻을 때 사용되었던 PQ(pivotal quantity)가 이제는 LRT에 따른 검정통계량으로 쓰이고 있는 셈인데 (단, PQ의 $\mu$ 와 $\sigma^2$을 각각 $\mu_0$와 $\sigma_0^2$으로 대체함), 이는 나머지 PQ에 대해서도 성립한다.

두 개의 정규 모분포가 있을 때 귀무가설 "$\sigma_1^2=\sigma_2^2$"를 검정하는 통계량은 식 (2.15.14) 인데, 이는 귀무가설 하에서 (단, $\mu_1$과 $\mu_2$를 모를 때)

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

이다. 그런데, 귀무가설 "$\sigma_1^2=\sigma_2^2$"은 "$(\sigma_1^2/\sigma_2^2)=1$"과 동일하므로, 대립가설이 "$\sigma_1^2>\sigma_2^2$" 또는 "$(\sigma_1^2/\sigma_2^2)>1$"이면 UTT가 되고 대립가설이 "$\sigma_1^2\neq\sigma_2^2$" 또는 "$(\sigma_1^2/\sigma_2^2)\neq 1$"이면 TTT가 되는 것이 쉽게 수긍이 간다. 물론, 대립가설이 "$\sigma_1^2<\sigma_2^2$" 또는 "$(\sigma_1^2/\sigma_2^2)<1$"이면 LTT가 된다. 그리고, 예를 들어 $n_1=10$, $n_2=5$, $\alpha=0.05$ 일 때 TTT의 기각역은 8.90이상 또는 $(4.72)^{-1}$ 이하이다 (§3.7.2 참조).

<비고 4.4.6> 지금 등장한 $F$-test는 UTT, LTT, TTT가 모두 가능하다. 그러나, 5장 이후에 등장하는 $F$-test는 모두 UTT이다. (<비고 4.4.5> 참조.)

두 개의 정규 모분포가 있을 때 귀무가설 "$\mu_1=\mu_2$"를 검정하는 통계량으로는 $\sigma_1^2$과 $\sigma_2^2$이 알려져 있으면 식 (3.7.1)을 그리고 $\sigma_1^2$과 $\sigma_2^2$을 모르는 경우에는 식 (3.7.4)를 사용한다. 즉, 식 (3.7.1)과 (3.7.4)에 "$\mu_1-\mu_2=0$"를 대입하면 귀무가설 하에서 각각 $N(0,1^2)$과 자유도가 $n_1+n_2-2$ 인 $t$분포를 따른다. 그리고, 대립가설이 "$\mu_1<\mu_2$"이면 UTT, "$\mu_1>\mu_2$"이면 LTT, "$\mu_1\neq\mu_2$"이면 TTT가 된다. 물론, 대립가설이 "$\mu_1<\mu_2$" 또는 "$\mu_1>\mu_2$"일 때에는 귀무가설이 "$\mu_1\geq\mu_2$" 또는 "$\mu_1\leq\mu_2$"이더라도 각각 UTT와 LTT가 된다. 또한, 기각역을 구하는 방법은 앞에서 등장한 $Z$-test와 $T$-test의 경우와 동일하다.

<비고 4.4.7> $T$-test이고 TTT인 "$H_0$: $\mu_1=\mu_2$, $H_a$: $\mu_1\neq\mu_2$" 경우는 §5.2에서 상세히 다룬다.

사실, 가설이 다르더라도 검정통계량이 귀무가설 하에서 $N(0,1^2)$분포를 따르기만 하면 기각역은 동일하다. 예를 들어, 모든 $Z$-test에서 $\alpha=0.05$이고 UTT이면 기각역은 $z\geq1.645$이다. 이는 또한 $T$-test, 카이제곱 검정, $F$-test에서도 성립한다. 예를 들어, 모든 $T$-test에서 자유도가 9, $\alpha=0.05$, 그리고 UTT이면 기각역은 $t\geq1.833$이다.

## §4.5 중심극한정리와 검정

중심극한정리(central limit theorem)를 CLT라 부르자. CLT는 §2.5에서 정규분포를 소개할 때 처음 등장했는데, 이를 협의의 CLT라 하자. 즉, $Y_1, \cdots, Y_n$이 $iid$ 확률변수이면 $\overline{Y} = \sum_{i=1}^{n} Y_i/n$ 은 점근적으로 정규분표를 따른다는 것이 협의의 CLT인데(§2.15.2 참조), 이는 물론 $Y$의 분포와 무관하게 (그리고, $Y$의 분포를 모를 때에도) 성립한다.

반면에, 광의의 CLT는 다음과 같다. 첫째로, §3.5.2에서 모든 최우추정량은 점근적으로 정규분포를 따른다고 했다. 둘째도 이와 같은 맥락인데, LRT에서 LR를 $\lambda$라 하면 (<비고 4.4.2> 참조)

$$-2\ln\lambda \xrightarrow{A} \chi^2(d) \tag{4.5.1}$$

이라고 알려져 있다. 즉, "$-2\ln\lambda$"는 점근적으로 카이제곱분포를 따른다는 것인데, 이때 유의할 점은 검정의 대상이 무엇이든지 불문하고 성립한다는 점과 또한 모집단의 분포와 무관하다는 점이다. 그리고, 자유도 $d$ 는 귀무가설에 포함된 제약식의 개수인데, 예를들면 하나의 모집단의 경우에 "$\mu = \mu_0$", "$\sigma^2 = \sigma_0^2$" 그리고 두 개의 모집단의 경우에 "$\mu_1 = \mu_2$", "$\sigma_1^2 = \sigma_2^2$ " 등이 제약식이다. 그런데, 식 (4.4.9)에서 $0 \le \lambda \le 1$ 이고 기각역은 $\lambda \le k$이므로, 식 (4.5.1)에서는 $0 \le -2\ln\lambda < \infty$ 이고 기각역은 $-2\ln\lambda \ge k'(=-2\ln k)$가 된다. 즉, 광의의 CLT에 의한 카이제곱 검정은 항상 UTT이다 (<비고 4.4.5> 참조). 예를 들어 자유도가 1이고 유의수준이 5%이면 기각역은 $-2\ln\lambda > 3.84146$ 이다.

협의의 CLT의 대표적인 활용사례는 모비율에 대한 추정 및 검정이다. §3.1에서는 모비율에 대한 점 추정량 및 신뢰구간을 구했고, §4.2에서는 모비율에 대한 검정을 예로 들어서 LRT를 설명했다.

광의의 CLT의 대표적인 활용사례는 §4.6에 등장하는데, 이 역시 모비율에 대한 검정이다. 다만, 차이점은 이항분포 대신에 다항분포가 등장하고 또한 $Z$-test 대신에 카이제곱 검정이 되는 것이다.

<비고 4.5.1> 이항분포는 다항분포의 특수한 경우이고, ($Z$를 제곱하면 자유도가 1인 카이제곱 확률변수가 되므로 : §2.6.1 참조) $Z$-test는 카이제곱 검정의 특수한 경우이다(§4.6.2 참조).

## §4.6 분할표 분석

### 4.6.1 일차원 분할표

주사위 하나를 600번 굴려서 아래의 결과를 얻었다고 하자. (비고: 이 예제는 문헌[9]의 연습문제 14-2임.)

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 합계 |
|---|---|---|---|---|---|---|---|
| $n_i$ | 89 | 113 | 98 | 104 | 117 | 79 | 600(=$n$) |

이와 같은 표를 일차원 분할표 (contingency table)라 한다. (비고:이차원 분할표는 §4.6.4에 등장함.) 즉, 전체 600을 주사위 눈의 수에 따라 여섯 부분으로 분할해 놓은 표이다.

이 예제에서 관심사는 주사위가 과연 대칭인가 (또는, balanced 인가) 하는 것이다. 이를 가설검정의 틀로 표현하면

$$\begin{aligned} H_0 &: p_i = \frac{1}{6}\ ,\ i = 1,\cdots,6 \\ H_a &: Not\ H_0 \end{aligned} \tag{4.6.1}$$

인데, $p_i$는 주사위를 한번 굴릴 때 $i$개의 눈이 나올 확률이다.

$n_i$에 대응하는 확률변수를 $N_i$라 하자. 즉, 600번 굴릴 때 $i$개의 눈이 나오는 횟수가 $N_i$인데, 귀무가설 하에서 $\{N_1, \cdots, N_6\}$는 다항분포를 따른다(§3.1.2 참조). 아래의 표는 위의 표에 $E(N_i)$와 관련된 몇가지를 추가한 것이다. (비고: $E(N_i) = np_i$)

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 합계 |
|---|---|---|---|---|---|---|---|
| $n_i$ | 89 | 113 | 98 | 104 | 117 | 79 | 600 |
| $E(N_i)$ | 100 | 100 | 100 | 100 | 100 | 100 | 600 |
| $n_i - E(N_i)$ | -11 | 13 | -2 | 4 | 17 | -21 | 0 |
| $\{n_i - E(N_i)\}^2$ | 121 | 169 | 4 | 16 | 289 | 441 | |
| $\frac{\{n_i - E(N_i)\}^2}{E(N_i)}$ | 1.21 | 1.69 | 0.04 | 0.16 | 2.89 | 4.41 | 10.40 |

□contingency table□을 □분할표□라 부르기는 하지만, □contingency□는 □우발성□이라 부르는데 이는 위의 표에서 $\{n_i - E(N_i)\}$를 일컫는 표현이다. 귀무가설이 참이라고 해서 반드시 $n_i = 100 (= E(N_i))$이 되는 것은 아니다. 확률적으로 또는 □우발적□으로 $n_i$는 100보다 크기도 하고 작기도 한 것이다.

결론부터 언급하면, 이는 카이제곱 검정인데 UTT이고(<비고 4.4.5> 참조) 자유도는 5이다. 따라서, α =0.05에 의한 기각역은 11.0705 이상이다. 그런데, 위의 표에서

$$\sum_{i=1}^{6} \frac{\{n_i - E(N_i)\}^2}{E(N_i)} = 10.40 < 11.0705 \qquad (4.6.2)$$

이므로 귀무가설을 채택(accept)한다. 즉, 주어진 표본정보는 유의수준 5%에서 귀무가설을 기각할 만한 충분한 근거가 되지 못한다.

이 검정 역시 LRT에 의한 것인데, 2-단계 근사라고 할 수 있다(<비고 2.15.3> 참조). 즉, 1차적으로는 광의의 CLT인 식 (4.5.1)에 의한 근사이고, 2차적으로는 $-2\ln\lambda$를 멱급수(power series)로 전개하여 식(4.6.2)만 남기고 나머지는 무시함에 따른 근사이다. (비고 : □ $-2\ln(1-x) = 2\sum_{k=1}^{\infty} x^k/k$ □에서 □ $2x + x^2$□만 남긴

것임.) 참고로, LF는 식(3.1.1) 형태인 $K\prod_{i=1}^{6} p_i^{n_i}$인데, $p_i$에 대한 최우추정치는 귀무가설 제약 하에서는 1/6 이고(<비고 4.4.1> 참고) 제약없이는 $n_i/n$이다. 따라서, $\lambda = \prod_{i=1}^{6}\{(1/6)/(n_i/n)\}^{n_i}$이다. (이후 과정은 생략함.)

이제, 자유도가 5인 이유를 설명한다. 식(4.6.1)의 귀무가설에 포함된 제약식의 개수는 6개이지만, 이중에서 5개가 주어지면 남은 하나는 자동적으로 결정되므로(비고 : $\sum_{i=1}^{6} p_i = 1$ ) 사실상 5개인 셈이다. 이는 또한 정보의 개수라는 관점으로 해석할 수도 있다(§2.15.4 참조). 분할표에 있는 정보는 $\{n_1 = 89, \cdots, n_6 = 79\}$ 라고 할 수 있는데, 정보의 개수가 5인 이유는 $\sum_{i=1}^{6} n_i = 600$ 이기 때문에 $n_1, \cdots, n_6$ 중에서 5개만 주어지면 남은 하나는 자동적으로 결정되기 때문이다.

일차원 분할표 분석에 대한 별칭은 □chi-square test of GOF(goodness of fit)□이다. GOF란 주어진 표본의 분포가(<비고 1.5.1>참조) 귀무가설 하의 분포와 얼마나 잘 맞는가를 가늠하는 척도인데, □ $H_0 : p_i = 1/6$ □은 이산 uniform 분포를 의미한다(§2.1.5 참조).

<비고 4.6.1> 귀무가설 하의 분포에 미지의 모수가 표함되어 있으면 이를 MLE로 대체한다. 아울러, MLE로 대체된 모수의 개수만큼 자유도가 감소하는데, 이는 식(2.15.10)의 μ를 MLE인 $\overline{Y}$로 대체함에 따라 식 (2.15.11)에서는 자유도가 하나 감소한 이유와 같다.

GOF 검정에 대해서 마지막으로 짚고 넘어갈 것은 유의수준이다. §4.1에서 검정의 목적은 귀무가설을 기각하기 위한 것이라고 했다. 그리고, 귀무가설이 기각되었을 때에는 α 값이 작을수록 설득력이 강하다. 즉, α 값을 작게 책정함으로써 어지간하면

귀무가설이 채택되도록 했음에도 불구하고 귀무가설이 기각되었다는 것은 표본정보가 귀무가설에 결정적으로 불리했다는 것을 의미한다. 그러나, GOF 검정은 예외이다. 그 이유는 GOF 검정이 귀무가설을 (기각이 아니라) 채택하기 위해서 사용되는 경우가 많기 때문이다. 이는 주어진 표본의 분포가 귀무가설에서 주장하는 분포와 잘 맞는다는 것을 보이기 위한 검정을 의미하는데, 이 경우 귀무가설이 채택되었을 때에는 α 값이 클수록 오히려 설득력이 강해진다. 위의 예에서 $\alpha = 0.05$로 귀무가설을 채택했는데, 만약 $\alpha = 0.10$이었다면 기각역은 9.23635 이상이 되어 귀무가설을 (채택하지 못하고) 기각하게 된다.

### 4.6.2 $Z$-Test 와의 관계

§4.6.1의 예제에서 $p_1$하나만 관심의 대상이라고 하자. 이에 따라, 식 (4.6.1)을

$$H_o: p_1 = 1/6, \quad H_a: p_1 \neq 1/6 \tag{4.6.3}$$

로 고치면, 식(4.6.2) 대신

$$\frac{(89-100)^2}{100} + \frac{(511-500)^2}{500} = 1.452 \tag{4.6.4}$$

를 얻는다. 이 역시 카이제곱 검정이고 UTT지만 자유도는 1이 된다. 따라서, $\alpha = 0.10$에 의한 기각역은 2.70554 이상이므로 ($\alpha = 0.05$에 의한 기각역은 3.84146 이상), 유의수준 10%에서 조차 귀무가설을 채택하게 된다.

식 (4.6.3)은 협의의 CLT에 의한 $Z$-Test로 처리할 수도 있다(<비고 4.5.1> 참조). 귀무가설하에서 검정통계량인 식 (3.1.7)은 점근적으로 $N(0,1^2)$분포를 따르는데, 이에 표본관찰치를 대입하면

$$z = \frac{\widehat{p_1} - p_1}{\sqrt{p_1(1-p_1)/n}} = \frac{\frac{89}{600} - \frac{1}{6}}{\sqrt{\frac{1}{6} \cdot \frac{5}{6}/600}} = -1.205 \qquad (4.6.5)$$

를 얻는다. 그런데, 이는 TTT 이므로 $\alpha = 0.10$에 의한 기각역은 $z \geq 1.645$ 또는 $z \leq -1.645$ 이다. 따라서, 유의수준 10%에서 귀무가설을 채택한다.

위의 두가지 검정이 동일함을 수치적으로 보인다. 첫째, 식 (4.6.5)을 제곱하면 식 (4.6.4)가 된다. 둘째로, $\alpha = 0.10$에 의한 기각역 "$z \geq 1.645$ 또는 $z \leq -1.645$"는 "$z^2 \geq 2.70554$"와 동일하다(<비고 4.5.1> 참조>). 따라서, §4.6.1의 GOF 검정은 모비율에 대한 $Z$-Test를 확장한 것이라 할 수 있다.

### 4.6.3 $p$-value

§4.6.2의 예제에서 $\alpha = 0.10$으로 $H_0$를 채택했다. 만약 $\alpha < 0.10$ 이었더라도 (기각역은 오히려 더 좁아지므로) 여전히 $H_0$를 채택했을 것이다. 그러나, $\alpha$값을 점점 증가시키면 (기각역이 점점 넓어져서) 언젠가는 $H_0$를 기각하게 되는데, 이 때 경계치를 $p$-value 라 부른다. 구체적으로, TTT의 기각역을 "$z \geq 1.205$ 또는 $z \leq -1.205$"가 되게하는 $\alpha$값이 바로 $p$-value 인데, 이는 $P(Z \geq 1.205 \text{ 또는 } Z \leq 1.205) = 2(0.1141) = 0.2282$이다. 즉, $\alpha < 0.2282$였다면 $H_0$를 채택하고 $\alpha \geq 0.2282$였다면 $H_0$를 기각한다.

일반적으로, 모든 검정에서 $\alpha < p$-value 이면 $H_0$를 채택하고 $\alpha \geq p$-value 이면 $H_0$를 기각한다. 통계 관련 패키지 (package 또는 s/w)는 대부분 $p$-value를 제공하는데, 이는 사용자들이 각자가 선호하는 $\alpha$ 값과 비교해서 $H_0$를 채택 또는 기각할 수 있게하기 위함이다.

### 4.6.4 이차원 분할표

§4.6.1에 등장한 일차원 분할표는 전체 $n$을 한가지 요인(factor)에 따라 $n_1, n_2, \cdots$로 분할한 것이다 (단, $n = \sum_i n_i$). 반면에, 이차원 분할표는 전체 $n$을 두가지 요인에 따라 아래의 형태로 분할한 것이다. (비고: $r_i = \sum_j n_{ij}$, $c_j = \sum_i n_{ij}$, $n = \sum_i r_i = \sum_j c_j$)

| $i$ \ $j$ | 1 | 2 | $\cdots$ | $J$ | row sum |
|---|---|---|---|---|---|
| 1 | $n_{11}$ | $n_{12}$ | $\cdots$ | $n_{1J}$ | $r_1$ |
| 2 | $n_{21}$ | $n_{22}$ | $\cdots$ | $n_{2J}$ | $r_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ | | $\vdots$ | $\vdots$ |
| $I$ | $n_{I1}$ | $n_{I2}$ | $\cdots$ | $n_{IJ}$ | $r_I$ |
| column sum | $c_1$ | $c_2$ | $\cdots$ | $c_J$ | $n$ |

일차원 분할표의 분석을 “chi-square test of GOF”라고 부르듯이, 이차원 분할표의 분석은 “chi-square test of independence (또는, homogeneity)”라고 부른다. 여기에서, “independence”란 두가지 요인이 서로 영향을 미치지 않는다는 의미이다. 예를 들어, 담배를 피우는가($i=1$) 안 피우는가($i=2$)하는 것이 교회에 다니는가($j=1$) 안 다니는가($j=2$)하는 것과 상관이 있는지를 알아보는 것이다. 또한, “homogeneity”라는 표현은 다음과 같은 형태의 “동질성”을 의미한다.

$$\frac{n_{1j}}{r_1} \approx \frac{n_{2j}}{r_2} \approx \cdots \approx \frac{n_{Ij}}{r_I} \left( \approx \frac{c_j}{n} \right),\ j = 1, \cdots, J \tag{4.6.6}$$

<비고 4.6.2> 요인 $i$와 요인 $j$는 교환가능(interchangeable)하다. 따라서, 식(4.6.6)은 "$(n_{ij}/c_j) \approx (r_i/n),\ i = 1, \cdots, I,\ j = 1, \cdots, J$"와 같다.

이차원 분할표의 분석을 간단히 "독립성 검정"이라 부르자. 독립성 검정은 GOF 검정을 확장시킨 것이다. 구체적으로, GOF 검정에는 다항분포 하나가 등장하지만 독립성 검정에는 $I$개의 (또는, <비고 4.6.2>에 의해서 $J$개의) 다항분포가 등장한다. 예를 들어, 지난 총선 때 여론조사가 $I$번 시행되었는데, 그 중 $i$번째에서 $j$번째 정당을 지지한 유권자의 수를 $n_{ij}$라 하자. 이 경우, 귀무가설은 "시간이 흘러도 정당별 지지율은 변하지 않음"이다. 가설을 수학적으로 표현하기 위해서, $n_{ij}$에 대응하는 확률변수를 $N_{ij}$라 하면 $\{N_{i1}, \cdots, N_{ij}\}$는 다항분포를 따르는데, 이의 모수를 $\{p_{i1}, \cdots, p_{ij}\}$라 하자. 그러면 $H_0$와 $H_a$는 다음과 같다.

$$\begin{aligned} H_0 :\ & p_{11} = p_{21} = \cdots = p_{I1} \qquad H_a :\ Not\ \ H_0 \\ & p_{12} = p_{22} = \cdots = p_{I2} \\ & \qquad \vdots \\ & p_{1J} = p_{2J} = \cdots = p_{IJ} \end{aligned} \tag{4.6.7}$$

독립성 검정 역시 광의의 CLT에 의한 카이제곱 검정이고 UTT인데, 자유도는 $(I-1)(J-1)$이다. 즉, $H_0$에 포함된 제약식 중에서 자동적으로 결정되는 것을 제외하면 모두 $(I-1)(J-1)$개다. 구체적으로, "$p_{1j} = p_{2j} = \cdots = p_{Ij}$"는 "$p_{1j} = p_{2j}$, $p_{2j} = p_{3j}, \cdots, p_{I-1,j} = p_{Ij}$"와 같으므로 $(I-1)$개의 제약식에 해당된다. (비고:

"$p_{Ij}=p_{1j}$"는 자동적으로 결정됨.) 그리고, 이같이 $(I-1)$개의 제약식에 해당되는 것들은 모두 $J$개가 있으나 이중에서 $(J-1)$개가 주어지면 남은 하나는 자동적으로 결정된다 (이유 ?). 이는 자유도가 "사용된 정보의 개수"라는 관점으로도 설명이 가능하다. 모두 $IJ$개의 $n_{ij}$값들이 있으나 일차적으로 $r_1,\cdots,r_I$에 의해서 $I$개는 자동적으로 결정된다. (비고: $r_i$는 여론조사에서 표본의 크기에 해당됨.) 남은 $I(J-1)$개에서 이차적으로 $(J-1)$개를 빼는데 이는 귀무가설 하에서 필요한 MLE의 개수이다 (<비고 4.6.1> 참조). 즉, $I(J-1)$개의 가용 정보 중에서 $(J-1)$개는 미지의 모수를 추정하는데 쓰이고 나머지 $(I-1)(J-1)$개 만이 검정에 쓰이는 셈이다.

참고로, LF는 식 (3.1.1)를 확장한 형태인 $\prod_{i=1}^{I} K_i \prod_{j=1}^{J} p_{ij}^{n_{ij}}$인데, $p_{ij}$에 대한 최우추정치는 귀무가설 제약 하에서는 $\widehat{p}_j = c_j/n$ 이고 (식 (4.6.6) 참조), 제약없이는 $\widehat{p}_{ij} = n_{ij}/r_i$ 이다. 따라서,

$$\lambda = \frac{\prod_{i=1}^{I}\prod_{j=1}^{J}(c_j/n)^{n_{ij}}}{\prod_{i=1}^{I}\prod_{j=1}^{J}(n_{ij}/r_i)} \tag{4.6.8}$$

이다. 그리고, GOF 검정에서와 같이, $-2\ln\lambda$ 를 멱급수로 전개하여 3차항 이상을 무시하면 (비고: 1차항은 0이 되므로 사실상 2차항만 남기는 것임), 기각역으로

$$Reject\ H_0 \text{ if } \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\{n_{ij}-\widehat{E}(N_{ij})\}^2}{\widehat{E}(N_{ij})} \geq k \tag{4.6.9}$$

를 얻는데, 이는 식 (4.6.2)을 확장한 형태이다. 단, $\widehat{E}(N_{ij})$는 귀무가설 하에서의 "$E(N_{ij}) \equiv r_I p_{ij}$"에 대한 MLE 이다. 즉,

$$\widehat{E}(N_{ij}) = r_i \widehat{p}_j = \frac{r_i c_j}{n} \tag{4.6.10}$$

이다(§3.5.1에 등장한 MLE의 invariance 속성 참조).

<비고 4.6.3> 식 (4.6.9)를 3-차원으로 확장하면 (카이제곱 검정의) 기각역은

$$\sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{\{n_{ijk} - \widehat{E}(N_{ijk})\}^2}{\widehat{E}(N_{ijk})} \geq k$$

가 되고, 자유도는 $(I-1)(J-1)(K-1)$이다.

### 4.6.5 독립성 검정 예제

100명에게 설문한 결과 담배를 피우는 사람은 40명이고 안 피우는 사람은 60명인데, 그 중에서 교회에 다니는 사람은 각각 8명과 22명이라 하자. 이를 분할표로 표현하면 다음과 같다.

| $i$ \ $j$ | 1 | 2 | |
|---|---|---|---|
| 1 | $n_{11}=8$ | $n_{12}=32$ | $n_1 \equiv r_1 = 40$ |
| 2 | $n_{21}=22$ | $n_{22}=38$ | $n_2 \equiv r_2 = 60$ |
| | $c_1=30$ | $c_2=70$ | $n=100$ |

편의상, $p_i \equiv p_{i1}$, $q_i \equiv p_{i2}$ 라 하면 $H_0$와 $H_a$는

$$H_0 : p_1 = p_2 \quad (\text{비고: } \text{“}p_1 = p_2\text{”}\text{이면 자동적으로 } \text{“}q_1 = q_2\text{”})$$
$$H_a : p_1 \neq p_2 \quad (\text{비고: } \text{“}p_1 \neq p_2\text{”}\text{는 } \text{“}Not\ H_0\text{”}\text{와 같음}) \tag{4.6.11}$$

인데, 이는 식 (4.6.7)의 특수한 경우 ($I = J = 2$)일 뿐만 아니라 식 (3.7.5)에 대응되는 검정이기도 하다. 돌이켜보면, §3.7에서 신뢰구간용 PQ로 사용되었던 통계량들이 4장에서는 (LRT에 의한) 검정통계량으로 재등장하고 있는데, 아직 하나 남은 것이 식 (3.7.5)에 대한 PQ인

$$Z = \frac{(\widehat{p_1} - \widehat{p_2}) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \tag{4.6.12}$$

인 것이다.

식 (4.6.12)의 배경에 깔린 가정은 다음과 같다. $p_1$은 담배를 피우는 사람들의 모집단에서 교회에 다니는 사람의 모비율이고, $p_2$는 담배를 안 피우는 사람들의 모집단에서 교회에 다니는 사람의 모비율이다. 표본의 크기는 각각 $n_1 = 40$과 $n_2 = 60$인데, $n_{i1}$에 대응하는 확률변수를 $N_i$라 하면 $N_1$과 $N_2$는 서로 독립이고 (이유: 서로 다른 모집단에서 추출), 각각 이항분포를 따른다 (이유: 복원추출로 가정). 이에, 협의의 CLT를 적용하면 $N_i \overset{A}{\sim} N(n_i p_i, n_i p_i q_i)$이고 또한 $\widehat{p_i} \equiv (N_i/n_i) \overset{A}{\sim} N(p_i, p_i q_i / n_i)$이다. 그리고, $N_1$과 $N_2$가 독립이므로 $\widehat{p_1}$과 $\widehat{p_2}$도 독립이다 (<비고 2.11.2> 참조). 따라서, 식 (4.6.12)는 점근적으로 $N(0,1^2)$분포를 따른다.

귀무가설 하에서 식 (4.6.12)에 “$p_1 = p_2$”를 대입하면, 분자의 “$p_1 - p_2$”는 “0”이 되지만 분모에는 여전히 미지의 모수(인 $p_1$과 $p_2$)가 남는다. (비고: 식 (4.6.5)에서는 이러한 문제가 발생하지 않았는데, 이는 식 (4.6.3)의 귀무가설에서 구체적인 수치 ($p_1 = 1/6$)가 주어졌기 때문임.) 따라서, 2차적인 근사로 (1차는 CLT에 의한 근사),

미지의 모수를 최우추정치로 대체한다. 그런데, 유의할 점은 귀무가설에서 구체적인 수치가 주어지지는 않았으나, "$p_1 = p_2$"라고 했으므로 $p_1$과 $p_2$를 따로따로 추정하지 않고 한꺼번에 묶어서

$$\hat{p} = \frac{8+22}{40+60} = \frac{30}{100}\left(= \frac{c_1}{n}\right) \tag{4.6.13}$$

로 추정한다는 점이다. 즉, $\hat{p}$은 소위 pooled 추정치인데 (식 (3.7.3) 참조), 이는 물론 식 (4.6.8)의 $(c_j/n)$에 해당된다.

$Z$-test의 판정결과를 얻기 위해서 식 (4.6.12)의 "$\hat{p}_i = N_i/n_i$"를 이에 대응하는 관찰치인 "$n_{ij}/n_i$"로 대체하고, 또한 $p_1$과 $p_2$를 식 (4.6.13)으로 대체하면

$$z \equiv \frac{\left(\frac{8}{40} - \frac{22}{60}\right) - 0}{\sqrt{\left(\frac{30}{100}\right)\left(\frac{70}{100}\right)\left(\frac{1}{40} + \frac{1}{60}\right)}} = -1.782 \tag{4.6.14}$$

를 얻는다. 그런데, 이는 TTT이므로 $\alpha = 0.05$에 의한 기각역은 $z \geq 1.96$ 또는 $z \leq -1.96$이다. 따라서, 유의수준 5%에서 귀무가설은 채택된다. 반면에, $\alpha = 0.10$에 의한 기각역은 $z \geq 1.645$ 또는 $z \leq -1.645$이므로, 유의수준 10%에서는 귀무가설을 기각할 수 있다. 참고로, $p$-value는 $P(Z > 1.782 \text{ 또는 } Z \leq -1.782) \approx 0.075$이다(§4.6.3 참조).

이제 식 (4.6.11)의 가설을 식 (4.6.9)로 검정한다. 카이제곱 검정에서 자유도는 $(2-1)(2-1) = 1$이고 UTT이므로, 기각역은 (§4.6.2에서와 동일하게) $\alpha = 0.05$ 일 때는 3.84146이상이고 $\alpha = 0.10$일 때는 2.70554이상이다. 그런데, 식 (4.6.9)에 $n_{ij}$ 값과 식 (4.6.10)을 대입하면

$$\frac{(8-12)^2}{12}+\frac{(32-28)^2}{28}+\frac{(22-18)^2}{30}+\frac{(38-42)^2}{42}=3.1746 \qquad (4.6.15)$$

을 얻는데, 이는 $\alpha=0.05$일 때는 기각역에 속하지 않고 $\alpha=0.10$일 때는 기각역에 속한다. 또한, §4.6.2에서와 같이, 위의 두가지 검정은 동일하다. (비고: 식 (4.6.14)를 제곱하면 식 (4.6.15)를 얻음.) 따라서, 카이제곱 검정의 $p$-value 역시 ($Z$-test와 같이) 약 0.075 (인데, 이는 카이제곱 확률표에는 잘 등장하지 않는 값)이다.

# 제 5장 ANOVA



## §5.1 서론

ANOVA는 “analysis of variance”의 약자인데, 이를 직역하면 “분산분석”이 된다. 그러나, 관심의 대상은 모분산이 아니라 모평균이다. 다만, 모평균에 대한 가설을 검정할 때 표본의 분산을 도구로 사용할 따름이다.

ANOVA는 다수의 모집단이 있을 때 사용하는데, 기본적인 가정은 첫째로 모분포들이 모두 정규분포이고, 둘째로 모분산들이 (알려지지는 않았지만) 모두 동일하다는 것이다. 그러니까, 모평균 $\mu_1, \mu_2, \mu_3, \cdots$는 서로 다를 수도 있는데, ANOVA는 바로

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \mu_3 = \cdots \\ H_a &: Not\ H_0 \end{aligned} \tag{5.1.1}$$

를 검정하는 것이다. (비고: “$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \cdots$”를 $H_0$와 $H_a$에 포함시켜도 무방함.) 즉, ANOVA는 §4.4.3의 뒷부분에서 간단히 다루었던

$$H_0 : \mu_1=\mu_2 \ (\text{그리고, } \sigma_1^2=\sigma_2^2)$$
$$H_a : \mu_1\neq\mu_2 \ (\text{그리고, } \sigma_1^2=\sigma_2^2) \qquad (5.1.2)$$

경우를 확장한 것이다 (<비고 4.4.7> 참조).

§4.6.2와 §4.6.5에서, 모비율에 대한 TTT인 $Z$-test를 확장하면 UTT인 카이제곱 검정이 되는 것을 보았다. 이와 같이, 식 (5.1.2)에 대한 TTT인 $T$-test를 확장하면 식 (5.1.1)에 대한 UTT인 $F$-test가 된다 (<비고 4.4.6> 참조). $T$-test가 $F$-test로 확장되는 것은 식 (2.5.7)에 의한 것이다. 그리고, $F$-test가 UTT인 이유는 이미 등장한 식 (4.4.10)에서도 찾을 수 있다. 즉, 식 (4.4.10)에서 "$t\geq\sqrt{k'}$ 또는 $t\leq\sqrt{k'}$"은 "$t^2\geq k'$"과 동일한데, 전자는 $T$-test의 기각역이고 후자는 $F$-test의 기각역이다.

§5.2에서는 식 (5.1.2)에 대한 검정을 정식으로 다루고, §5.3에서는 식 (5.1.1)에 대한 검정을 다룬다. 그리고, §5.4 이후에서는 보다 효과적으로 식 (5.1.1)을 검정하는 방법을 소개한다. 사실, 이 책에서 ANOVA라고 부르는 것은 소위 (통계적) 실험계획법(experimental design)의 범주에 속하는 것인데, 이 책에서는 효과적인 검정을 위한 실험계획법 중에서 가장 기본적인 경우만 소개한다. 또한, ANOVA는 소위 선형(통계)모형(linear model)의 틀에 속하기도 하는데, 6장에서 다룰 Linear Regression 역시 선형모형이다.

## §5.2 $T$-test for Independent Samples

ANOVA를 거론하기에 앞서 준비작업 삼아 식 (5.1.2)에 대한 $T$-test를 정식으로 다룬다.

젖소용 사료가 두 종류가 있는데, 사료 $i$를 먹인 젖소의 우유생산량을 $Y_i$라 하자 ($i=1,2$). $Y_i$에 대한 가정은

$$Y_i \sim N(\mu_i, \sigma^2) \tag{5.2.1}$$

이다 (비고: $\sigma_1^2=\sigma_2^2\equiv\sigma^2$). 식(5.1.2)의 가설을 검정하기 위해서 모집단 $i$에서 크기가 $n_i$인 표본 $\{Y_{i1}, Y_{i2}, \cdots, Y_{i,n_i}\}$를 추출한다고 하자 ($i=1,2$).

<비고 5.2.1> 이 절의 제목에 "independent samples"라는 표현이 사용된 이유는 서로 다른 모집단에서 따로따로 추출된 두 표본은 서로 독립이기 때문이다.

두 표본의 관찰치를 $\{y_{i1}, y_{i2}, \cdots, y_{i,n_i}\}$, $i=1,2$ 라 하면 LF는 식 (3.2.1)을 확장한 형태인

$$L(\mu_1,\mu_2,\sigma^2)=\left\{\prod_{i=1}^{2}\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^{n_i}\right\}\cdot e^{-\frac{1}{2}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij}-\mu_i)^2/\sigma^2} \tag{5.2.2}$$

이다. 즉, LF는 식 $\{Y_{11}, \cdots, Y_{1,n_1}\}$의 결합 밀도함수와 $\{Y_{21}, \cdots, Y_{2,n_2}\}$의 결합 밀도함수의 곱이다 (<비고 5.2.1> 참조).

귀무가설 하에서는 "$\mu_1=\mu_2$"이므로, 식 (5.2.2)에서 $\mu_1$과 $\mu_2$를 $\mu$로 대체한 다음 ($\mu$에 대해서 편미분하여) $\mu$에 대한 최우추정치를 구하면

$$\hat{\mu}=\frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}y_{ij}}{\sum_{i=1}^{2}n_i}(\equiv\overline{y}) \tag{5.2.3}$$

을 얻는데, 이는 두 표본을 합친 평균 관찰치이다. 다음, 식 (5.2.2)에 식 (5.2.3)을 대입(하고나서 $\sigma^2$에 대해서 편미분)하면, $\sigma^2$에 대한 최우추정치로

$$\widehat{\sigma_0^2}=\frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2}{n},\ n=\sum_{i=1}^{2}n_i \tag{5.2.4}$$

를 얻는다. 반면에, (귀무가설의 제약이 없는) 전체 모수공간에서의 최우추정치는 식 (5.2.2)를 $\mu_1,\mu_2,\sigma^2$에 대해서 편미분해서 얻는데, 결과는 다음과 같다 (§3.2.1 참조).

$$\widehat{\mu_i}=\frac{\sum_{j=1}^{n_i}y_{ij}}{n_i}(\equiv\overline{y_i}),\ i=1,2 \tag{5.2.5}$$

$$\widehat{\sigma^2}=\frac{\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij}-\overline{y_i})^2}{n},\ n=\sum_{i=1}^{2}n_i \tag{5.2.6}$$

위에서 구한 최우추정치들을 LF인 식 (5.2.2)에 대입하여 LRT의 기각역을 구하면 다음과 같다.

$$\lambda=\left(\frac{\widehat{\sigma^2}}{\widehat{\sigma_0^2}}\right)^{\frac{n}{2}}\le k,\ n=\sum_{i=1}^{2}n_i \tag{5.2.7}$$

<비고 5.2.2> λ의 형태는 식 (4.4.9)와 같다. 다만, $\widehat{\sigma_0^2}$과 $\widehat{\sigma^2}$의 내용이 복잡해졌을 뿐이다.

λ를 손질하면 (자세한 내용은 §5.3.1 참조), 다음과 같이 TTT의 기각역을 얻는다.

$$t^2=\left\{\frac{\overline{y_1}-\overline{y_2}}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}\right\}^2 \geq k'=(n-2)(k^{-\frac{2}{n}}-1) \tag{5.2.8}$$

그리고, $\overline{y_i}$를 $\overline{Y_i}$로 대체하고 $s^2$을 식 (3.7.3)의 $S^2$으로 대체하면, 검정통계량으로

$$T_{n-2}=\frac{\overline{Y_1}-\overline{Y_2}}{S\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}},\ n=n_1+n_2 \tag{5.2.9}$$

을 얻는데, 이는 귀무가설 하에서 자유도가 $n-2$인 $t$분포를 따른다 (<비고 3.7.2> 참조). (비고: 식 (3.7.4)에 귀무가설 "$\mu_1=\mu_2$"를 대입하면 식 (5.2.9)를 얻음. 단, 지금은 ($\overline{X}, \overline{Y}$) 대신에 ($\overline{Y_1}, \overline{Y_2}$)로 표본평균을 표기하고 있음.)

예를 들어, $\{y_{11}, y_{12}, y_{13}, y_{14}\}=\{3,2,1,2\}$이고 $\{y_{21}, y_{22}, y_{23}, y_{24}\}=\{5,2,4,5\}$라 하자. 먼저, 표본평균과 표본분산을 따로따로 구하면 다음과 같다.

$$\overline{y_1}=\frac{3+2+1+2}{4}=\frac{8}{4}=2,\ \overline{y_2}=\frac{5+2+4+5}{4}=\frac{16}{4}=4$$

$$s_1^2=\frac{(3-2)^2+(2-2)^2+(1-2)^2+(2-2)^2}{4-1}=\frac{2}{3} \tag{5.2.10}$$

$$s_2^2 = \frac{(5-4)^2+(2-4)^2+(4-4)^2+(5-4)^2}{4-1} = \frac{6}{3}$$

그리고, 이로부터 $\overline{y}$와 $s^2$을 다음과 같이 얻는다.

$$\overline{y} = \frac{4\overline{y_1}+4\overline{y_2}}{4+4} = \frac{8+16}{4+4} = 3$$

$$s^2 = \frac{(4-1)s_1^2+(4-1)s_2^2}{(4-1)+(4-1)} = \frac{2+6}{3+3} = \frac{8}{6} \qquad (5.2.1)$$

다음, 식 (5.2.8)에 $\overline{y_1}=2$, $\overline{y_2}=4$, $s=\sqrt{8/6}$, $n_1=n_2=4$를 대입하면 $t=-2.449$를 얻는다. 그런데, 자유도가 $n-2=6$인 $T$-test에서 $\alpha=0.05$에 대한 TTT의 기각역은 "$t\geq 2.447$ 또는 $t\leq -2.447$"이므로 (또는, $t^2>-5.99$), 유의수준 5%에서 귀무가설을 (가까스로나마) 기각한다. (비고: 가까스로 기각하므로 $p$-value는 0.05보다 약간 작은 값이 될 것임. §4.6.3 참조.)

## §5.3 One-Way ANOVA

### 5.3.1 One-Way ANOVA에 대한 LRT

§5.2의 "$T$-test for Independent Samples"를 모집단이 셋 이상인 경우로 확장한 것을 "One-Way ANOVA"라 하는데, 이는 §5.5에 등장할 "Two-Way ANOVA"와 구별하기 위한 명칭이다.

모집단의 개수를 $I$라 하자. 그러면, 귀무가설은 "$\mu_1 = \mu_2$"에서 "$\mu_1 = \mu_2 = \cdots = \mu_I$"로 확장되는데, 이때 식 (5.2.1)~(5.2.7)에서 달라지는 것은 "$i = 1, 2$"가 "$i = 1, 2, \cdots, I$"로 달라지는 것 한가지 뿐이다. 즉, $\prod_{i=1}^{2}$와 $\sum_{i=1}^{2}$를 각각 $\prod_{i=1}^{I}$와 $\sum_{i=1}^{I}$로 바꾸기만 하면 된다. 이제 식 (5.2.7)을 손질해서 편리한 형태로 고치겠는데, 손질한 후에 $I = 2$를 대입하면 식 (5.2.8)이 된다.

먼저, 앞으로 사용할 용어를 다음과 같이 정의한다.

$SS$ : sum of squares

$$TSS \text{ (total } SS) = n\,\widehat{\sigma}_0^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \overline{y})^2 \tag{5.3.1}$$

$$SSE \text{ ($SS$ for error)} = n\,\widehat{\sigma}^2 = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij} - \overline{y_i})^2 \tag{5.3.2}$$

$$SSTr \text{ ($SS$ for treatments)} = \sum_{i=1}^{I}\sum_{j=1}^{n_i}(\overline{y_i} - \overline{y})^2 = \sum_{i=1}^{I} n_i(\overline{y_i} - \overline{y})^2 \tag{5.3.3}$$

<비고 5.3.1> §3.3에서는 $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$을 $SS$라 불렀음.

<비고 5.3.2> 식 (5.3.1)~(5.3.3)에서 $y_{ij}$ (및 $\overline{y_i}$ 와 $\overline{y}$)를 확률변수 $Y_{ij}$ (및 $\overline{Y_i}$ 와 $\overline{Y}$)로 대체하더라도 여전히 동일한 호칭 ( $TSS$, $SSE$, $SSTr$)을 사용함 (<비고 3.1.1> 참조).

<비고 5.3.3> "$TSS = SSE + SSTr$"가 성립할 뿐더러 (증명은 생략함), 이들이 확률변수일 때 (<비고 5.3.2> 참조) $SSE$와 $SSTr$은 서로 독립이다.

$SSE$와 $SSTr$이 독립인 이유는 정규 모분포가 하나일 때 $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$와 $\overline{Y}$가 독립인 (<비고 2.15.2> 참조) 이유와 동일하다. 또한, $I = 2$인 경우에는

$$SSE = (n_1 + n_2 - 2)S^2 \tag{5.3.4}$$

$$SSTr = (\overline{Y_1} - \overline{Y_2})^2 \Big/ \left(\frac{1}{n_1} + \frac{1}{n_2}\right) \tag{5.3.5}$$

가 되는데, 이들은 각각 식 (5.2.9)의 분모와 분자에 등장한다. (비고: $SSE$와 $SSTr$이 독립이면 $S$와 ($\overline{Y_1} - \overline{Y_2}$)도 독립. <비고 2.11.2> 참조.)

식 (5.3.1)과 (5.3.2)를 식 (5.2.7)에 대입하면

$$\lambda = \left(\frac{SSE/n}{TSS/n}\right)^{\frac{n}{2}} \le k,\ n = \sum_{i=1}^{I} n_i \tag{5.3.6}$$

가 되는데, 이를 $SSE$와 $SST_r$로 표현되도록 손질하면 (<비고 5.3.3> 참조),

$$f \equiv \frac{SST_r/(I-1)}{SSE/(n-I)} \geq k' = \frac{n-I}{I-1}(k^{-\frac{2}{n}}-1) \qquad (5.3.7)$$

를 얻는다. 그리고, 이에 식 (5.3.4)와 (5.3.5)를 대입하면 식 (5.2.8)이 된다.

식 (5.3.7)에서 $SST_r$과 $SSE$를 각각 $(I-1)$과 $(n-I)$로 나눈 이유는 다음과 같다. $SST_r$과 $SSE$가 확률변수인 경우에 (즉, $y_{ij}$를 $Y_{ij}$로 대체했을 때: <비고 5.3.2> 참조), $SST_r/\sigma^2$과 $SSE/\sigma^2$은 귀무가설 하에서 카이제곱 분포를 따르는데 자유도는 각각 $(I-1)$과 $(n-I)$ (이고 물론 서로 독립)이다. 따라서,

$$\frac{SST_r/(I-1)}{SSE/(n-I)} \sim F(I-1, n-I) \qquad (5.3.8)$$

이다 (식 (2.15.14) 참조). 그리고, $SST_r/\sigma^2 \sim \chi^2(I-1)$과 $SSE/\sigma^2 \sim \chi^2(n-I)$에 대한 증명 방법은 §2.15.3에서 식 (2.15.11)을 증명한 방법과 동일한데 복잡하므로 생략한다.

<비고 5.3.4> 귀무가설 하에서 $TSS/\sigma^2 \sim \chi^2(n-1)$이다. 그러나, 이는 $SST_r/\sigma^2$ 및 $SSE/\sigma^2$과 독립이 아니다.

### 5.3.2 One-way ANOVA 예제

§5.2에 등장한 예제를 $I=4$ 경우로 다음과 같이 확장한다.

| $j$ \ $i$ | 1 | 2 | 3 | 4 | |
|---|---|---|---|---|---|
| 1 | 3 | 5 | 2 | 4 | |
| 2 | 2 | 2 | 2 | 2 | |
| 3 | 1 | 4 | 4 | 5 | |
| 4 | 2 | 5 | 4 | 1 | |
| $\overline{y_i}$ | 2 | 4 | 3 | 3 | $\overline{y}=3$ |

식 (5.3.1)과 (5.3.3) 그리고 <비고 5.3.3>에 의해서

$$TSS=\sum_{i=1}^{4}\sum_{j=1}^{4}(y_{ij}-\overline{y})^2=30$$

$$SST_r=\sum_{i=1}^{4}4(\overline{y_i}-\overline{y})^2=8$$

$$SSE=TSS-SST_r=22$$

인데, 이를 식 (5.3.7)에 대입하면

$$f=\frac{SST_r/(4-1)}{SSE/(16-4)}=\frac{2.67}{1.833}=1.455$$

를 얻는다. 그런데, 분자와 분모 자유도가 각각 (4-1)과 (16-4)인 $F$-test에서 $\alpha=0.05$에 대한 UTT의 기각역은 $f\geq 3.49$ 이므로 귀무가설을 (기각하지 못하고) 채택한다. (비고: $\alpha=0.10$에 대해서도 기각역이 $f\geq 2.61$ 이므로 귀무가설을 채택함.)

### 5.3.3 ANOVA Table

ANOVA 결과를 일목요연하게 표로 만들면 편리하다. 이때 추가되는 용어는 MS(mean square)인데, 이는 $SS$를 자유도로 나눈 것이다. §5.3.2의 예제에 대한 ANOVA Table 은 다음과 같다. (비고: 통계 패키지는 $p$-value 도 제공함. §4.6.3 참조.)

| Source (of Variation) | $SS$ | 자유도 | $MS$ | $f$ |
|---|---|---|---|---|
| Treatment | 8 | 3 ( = $I$ - 1 ) | 2.6 | 1.45 |
| Error | 22 | 12 ( = $n$ - $I$) | 1.83 | |
| Total | 30 | 15 ( = $n$ - 1 ) | (2) | |

참고로, §5.2의 예제에 대한 ANOVA Table은 다음과 같다.

| Source | $SS$ | 자유도 | $MS$ | $f$ |
|---|---|---|---|---|
| Treatment | 8 | 1 | 8 | 6 |
| Error | 8 | 6 | 1.3 | |
| Total | 16 | 7 | (2.286) | |

그런데, 분자와 분모 자유도가 각각 1 과 6 인 $F$-test에서 $\alpha = 0.05$ 에 대한 UTT의 기각역은 $f \geq 5.99$ 이므로 귀무가설을 (가까스로) 기각한다. (비고: $f$ 값 6 과 경계치 5.99는 §5.2 에서 얻은 $t$ 값 -2.449 와 경계치 2.447 을 각각 제곱한 것임.)

## §5.4 실험계획

### 5.4.1 신호와 잡음

3 장에서는 추정을 다루었고 4장 이후에는 검정을 다루고 있는데, 지금까지는 주로 LF를 미분해서 MLE를 얻고 또한 LR을 손질해서 검정통계량을 얻는데에만 급급했다. 그래도, 추정량은 대부분 그 자체로 의미가 있어서 실감이 났다. (예: 모평균에 대한 추정량은 표본평균.) 이제 검정통계량에 대해서도 의미를 부여해 보기로 한다.

신호(signal)와 잡음(noise)의 비율을 SN비라고 한다. 신호가 어느정도 강해도 잡음이 더 강하면 신호를 감지하기 어렵다. 반면에, 잡음이 거의 없으면 약한 신호라도 감지할 수 있다. 검정통계량도 SN비로 해석하면 이해하기 쉽다. 예를 들어, 식 (5.2.9)와 (5.3.8)에서 분자는 신호에 그리고 분모는 잡음에 해당된다. 그리고, §5.3.3의 ANOVA Table에 있는 MS들의 비율인 $f$값은 바로 관찰된 SN비에 해당된다.

신호가 강하고 잡음이 약한 효과적인 검정법을 찾는 것이 바로 실험계획이다. 이제, §5.3의 One-way ANOVA 보다 더 효과적인 검정법을 찾기 위한 준비작업으로 먼저 지금까지 얻은 결과를 분석한다.

### 5.4.2 Source of Variation

§5.3.3의 ANOVA Table에서 Treatment와 Error를 Source of Variation이라 불렀다. 여기에서, Variation이란 $(y_{ij}-\overline{y})^2$를 의미하는데, Variation의 총량(total)은 TSS인 $\sum_{i=1}^{I}\sum_{j=1}^{n_i}(y_{ij}-\overline{y})^2$이다.

§5.2에서 예로 들었듯이, 사료 $i$를 먹인 젖소 $n_i$마리 중에서 $j$번째 젖소의 우유생산량을 $y_{ij}$라 하자. 그러면, 어떤 젖소는 우유생산량이 (평균치 $\overline{y}$보다) 많고

또 어떤 젖소는 우유생산량이 적은데, 이러한 Variation의 총량을 $TSS$라 하는 셈이다. 그리고, $TSS$중에서 서로 다른 사료를 먹인데 기인한 부분이 $SST_r$인 $\sum_i \sum_j (\overline{y_i} - \overline{y})^2$이고, 나머지 부분을 $SSE$라 하는 셈이다. 이때, $SST_r$을 사료(또는, Treatment)에 의해서 설명된(explained) Variation이라 하고, $SSE$를 설명안된(unexplained) Variation이라 한다.

$SST_r$이 클수록 신호가 강하고 $SSE$가 작을수록 잡음이 약하다. 그러나, $SST_r$과 $SSE$ 그 자체가 아니라 각각에 대응된 자유도로 나누어서 표준화시킨 것을 신호와 잡음으로 사용하는데, 이들이 바로

$$MST_r \equiv SST_r/(I-1) \tag{5.4.1}$$

$$MSE \equiv SSE/(n-I) \tag{5.4.2}$$

이다.

### 5.4.3 MS의 정체 (이 절은 생략해도 무방함)

ANOVA의 기본 가정인 "분산이 같은 $I$개의 정규 모분포"에 귀무가설인 "$\mu_1 = \mu_2 = \cdots = \mu_I$" 를 합치면

$$Y_{ij} \sim N(\mu, \sigma^2) \tag{5.4.3}$$

이 된다. (비고: $Y_{ij}$, $i = 1, \cdots, I$, $j = 1, \cdots, n_i$, 는 $iid$ 확률변수임.) 즉, $Y_{ij}$는 정규분포를 따르는데 평균 $\mu$와 분산 $\sigma^2$은 사료의 종류인 $i$와 무관하다. 따라서, $I$개의 표본을 하나로 합치면 $\mu$에 대한 MLE로 $\overline{Y}$를 얻는데 이는 또한 MVUE이다

(식 (5.2.3) 참조). 또한, $\sigma^2$ 에 대한 MLE는 $TSS/n$ 이고 (식 (5.2.4) 참조), MVUE는

$$TMS \equiv TSS/(n-1) \tag{5.4.4}$$

이다.

<비고 5.4.1> $TMS$ (total mean square)는 ANOVA Table에 잘 등장하지 않음. 또한, "$TMS \neq MST_r + MSE$" 임 (<비고 5.3.3>참조).

그런데, $MST_r$ 과 $MSE$ 도 (귀무가설 하에서는) $\sigma^2$ 에 대한 불편추정량이다. 다만, 이들의 분산은 $TMS$ 의 분산보다 크다. (즉, $MST_r$ 과 $MSE$ 는 $MVUE$ 가 아니다.) 구체적으로, $E(TMS) = E(MST_r) = E(MSE) = \sigma^2$ 이지만, $V(TMS) = 2\sigma^4/(n-1)$, $V(MST_r) = 2\sigma^4/(I-1)$, $V(MSE) = 2\sigma^4/(n-I)$ 이다 (식 (2.15.12) 참조).

반면에, 귀무가설의 제약이 없으면, $\mu_i$ 에 대한 MLE 겸 MVUE로 $\overline{Y_i}$ 를 얻고 (식 (5.2.5) 참조), $\sigma^2$ 에 대한 MLE로 $SSE/n$ 을 얻으며 (식 (5.2.6) 참조), $\sigma^2$ 에 대한 MVUE로는 바로 MSE를 얻는다. (비고: $I=2$ 일때의 MSE는 식 (3.7.3)과 (5.3.4)의 $S^2$ 임.) 그런데, 귀무가설은 단순가설이므로 이는 전체 모수공간에서 극히 일부분에 지나지 않는다. (비고; 모수공간은 $I+1$ 차원 공간인데, 이 속에서 귀무가설은 2차원을 차지함.) 따라서, 귀무가설의 제약없이 구한 추정량들은 사실상 "대립가설 하에서" 구한 것이라 해도 별로 무리가 없다. 즉, MSE는 사실상 "대립가설 하에서" $\sigma^2$ 에 대한 MVUE인 셈이다.

그렇다면, 귀무가설 하에서 MSE와 같이 $\sigma^2$ 에 대한 불편추정량이던 $MST_r$ 은

대립가설 하에서는 무엇이 되는가? 이를 $I=2$ 인 경우에 대해서 간단히 살펴보자 (자세한 내용은 실험계획법 교재 참조). $I=2$ 이면 $I-1=1$ 이므로, 식 (5.3.5)는 $SST_r$ 인 동시에 $MST_r$ 이다. 그런데, 기대치를 구하면 (과정 생략)

$$E(MST_r)=\sigma^2+\frac{(\mu_1-\mu_2)^2}{\frac{1}{n_1}+\frac{1}{n_2}} \tag{5.4.5}$$

이 되므로, $MST_r$ 속에는 $\sigma^2$ 에 대한 불편추정량 외에 다른 것들이 추가로 포함되어 있음을 알 수 있다. (비고: 귀무가설 하에서는 $\mu_1-\mu_2=0$ 임.)

$I=2$ 인 경우에는 식 (5.4.5)에서 $(\mu_1-\mu_2)^2$ 이 클수록 $E(MST_r)$ 또는 $E$(신호의 크기)가 커진다. 그런데, $(\mu_1-\mu_2)^2$ 은 우리가 제어(control)할 수 있는 것이 아니다. 따라서 우리는 $E$(신호의 크기)를 증가시키는 대신에 $E$(잡음의 크기)를 감소시키려고 노력한다.

## §5.5 Two-Way ANOVA

### 5.5.1 Two-way ANOVA

잡음의 크기를 감소시키기 위한 대표적인 방법은 TWA (Two-Way ANOVA)이다. §5.3.2의 OWA(One-Way ANOVA) 예제에서 Variation의 총량인 $TSS$는 30인데 그 중에서 사료의 차이에 의해서 설명된 부분인 $SST_r$은 8이고 설명안된 나머지인 SSE는 22라고 했다. (§5.4.2 참조). TWA란 OWA에서 설명안된 부분 중에서 일부를 추가로 설명하는 것이다.

§5.3.2의 예제에서는 모두 16마리의 젖소가 동원되었다. 즉 16마리를 4마리씩 4무리로 나누어서 각각 다른 사료를 먹인 것이다.

<비고 5.5.1> OWA를 □Completely Randomized Design□ 이라고도 부르는데, 그 이유는 전체 실험대상을 $I$개의 무리로 나누는 방법이 □무작위□이기 때문이다.

예를 들어, 전체 16마리 중에서 가장 어린 4마리에게는 사료 1을 먹이고 가장 잘 자란 4마리에게는 사료 2를 먹인다면 이는 공정한 실험이 아니다. 물론, 그 이유는 잘 자란 젖소의 우유생산량이 어린 젖소보다 많을 것이기 때문이다. 따라서, OWA에서는 16마리를 무작위로 4마리씩 4무리로 나눈다(<비고 5.5.1> 참조).

반면에, 젖소의 (나이, 체중, 품종 등의) 차이에 의해서 발생하는 Variation을 아예 제거하는 방법이 바로 TWA이다. 이 경우 젖소는 4마리만 필요하지만 시간은 4배가 소요된다. 예를 들어, 처음 한달간은 4마리 모두에게 사료 1을 먹인다음 우유생산량을 측정한다. 그리고, 다음 한달 간은 4마리 모두에게 사료 2를 먹인 다음 우유생산량을 측정하는 식으로 실험을 하는 것이다. (비고: 실제로는 사료의 순서를 무작위

로 결정함. <비고5.5.3> 참조.)

OWA와 TWA의 차이점을 쉽게 파악하기 위해서 편의상 §5.3.2 의 $y_{ij}$ 를 다음과 같이 순서만 바꾸어서 사용한다.

| 젖소 $j$ \ 사료 $i$ | 1 | 2 | 3 | 4 ( = $I$) | $\overline{y_{.j}} = \sum_{i=1}^{I} y_{ij}/I$ |
|---|---|---|---|---|---|
| 1 | 3 | 5 | 4 | 4 | 4 |
| 2 | 2 | 5 | 4 | 5 | 4 |
| 3 | 1 | 4 | 2 | 1 | 2 |
| 4 ( = $J$) | 2 | 2 | 2 | 2 | 2 |
| $\overline{y_{i.}} = \sum_{j=1}^{J} y_{ij}/J$ | 2 | 4 | 3 | 3 | $\overline{y} = 3$ |

$y_{ij}$ 의 순서만 바꾸었으므로 $TSS$ 는 여전히 30 이다. 그리고, $\overline{y_{i.}}$ 는 §5.3.2의 표에서 $\overline{y_i}$ 와 일치하므로 $SST_r$ 또한 여전히 8 이다. 즉, Variation 의 총량 30 중에서 사료의 차이에 의해서 설명된 Variation은 여전히 8 이다.

이제, 젖소의 차이에 의해서 설명된 Variation 을 구한다. 이를 $SSB$ ( $SS$ for Blocks)라 부르는데, 구하는 방법은 $SST_r$ 과 동일하다(식 (5.3.3) 참조). 즉,

$$SSB = \sum_{i=1}^{I}\sum_{j=1}^{J}(\overline{y_{.j}} - \overline{y})^2 = I\sum_{j=1}^{J}(\overline{y_{.j}} - \overline{y})^2 = 16 \qquad (5.5.1)$$

이다. 그리고, $SST_r$ 에 대응된 자유도가 ( $I$ - 1 ) 이듯이 (식 (5.4.1) 참조), $SSB$ 에 대응된 자유도는 ( $J$ - 1 ) 이다.

TWA에서는 $SSE$가 다음과 같다.

$$SSE = TSS - SST_r - SSB \tag{5.5.2}$$

즉, $SSE$는 전체 Variation 중에서 사료의 차이와 젖소의 차이에 의해서 설명된 부분들을 제외하고 남은 (설명안된) Variation이다. 또한, 식 (5.5.2)는 $SS$에 대응된 자유도 간에도 성립한다. 따라서, $SSE$에 대응된 자유도는 다음과 같다.

$$(IJ-1)-(I-1)-(J-1)=(I-1)(J-1) \tag{5.5.3}$$

이상의 결과를 ANOVA Table로 정리하면 다음과 같다.

| Source (of Variation) | $SS$ | 자유도 | $MS$ | $f$ |
|---|---|---|---|---|
| Treatment | 8 | 3 | 2.6̄ | 4 |
| Block | 16 | 3 | | |
| Error | 6 | 9 | 0.6̄ | |
| Total | 30 | 15 | | |

§5.3.3 의 ANOVA Table과 비교하면, 관찰된 신호의 크기는 같지만 ($MST_r = 2.6̄$), 관찰된 잡음의 크기인 $MSE$는 1.83̄에서 0.6̄으로 줄어들었다. 따라서 SN비인 $f$ 값은 1.45̄에서 4로 늘어났다.

분자와 분모의 자유도가 각각 3과 9인 $F$-test에서 $\alpha = 0.05$에 대한 UTT의 기각역은 $f \geq 3.86$이므로, 이제는 귀무가설 "$\mu_1 = \mu_2 = \mu_3 = \mu_4$"를 기각할 수 있다. (비고: $\alpha = 0.01$ 이면 기각역은 $f \geq 6.99$이므로 귀무가설을 채택함.)

### 5.5.2 TWA에 대한 LRT

OWA에서는 모집단이 $I$개인 반면에, TWA에서는 모집단이 $IJ$개이다. 그러니까, 예제에서는 사료별로 그리고 젖소별로 모집단이 다르다. 그리고, 각 모집단에서 크기가 1인 표본을 하나씩 추출하는 셈이다.

<비고 5.5.2> 이책에서 다루는 TWA는 □TWA with one observation per cell□에 해당된다.

<비고 5.5.3> TWA를 □Randomized Block Design□ 이라고도 하는데, 이는 예를 들어 사료를 먹이는 순서(예: 사료 3→1→4→사료2)를 젖소 별로 무작위로 결정하는 것이다.

$Y_{ij}$를 젖소 $j$에게 사료 $i$를 먹일 때의 우유생산량이라 하자. TWA의 기본적인 가정은 OWA와 같다. 즉, $Y_{ij}$는 정규분포를 따르고 모분산이 ($i$와 $j$에 무관하게) 모두 같다는 것이다. 그러나, 모평균 $\mu_{ij}$는 $i$뿐만 아니라 $j$에 따라서도 다를 수 있다. 편의상,

$$\mu_{ij}=\mu+\tau_i+\beta_j \qquad (5.5.4)$$

라 하자. (단, $\sum_{i=1}^{I}\tau_i=0$, $\sum_{j=1}^{J}\beta_j=0$.) 그러면,

$$Y_{ij} \sim N(\mu + \tau_i + \beta_j, \sigma^2) \tag{5.5.5}$$

가 TWA 의 가정이다. 이를 식 (5.2.1)과 비교하면, OWA에서는 "$\mu_i = \mu + \tau_i$" 이다. 즉, OWA에서는

$$\beta_1 = \beta_2 = \cdots = \beta_J (=0) \tag{5.5.6}$$

을 가정하는 셈이다. 그리고, 식 (5.1.1)의 가설은 이제 다음과 같이 표현된다.

$$\begin{aligned} &H_0 : \tau_1 = \tau_2 = \cdots = \tau_I (=0) \\ &H_a : Not\ H_0 \end{aligned} \tag{5.5.7}$$

<비고 5.5.4> 이 책에서는 미지의 모수인 $\tau_i$ 와 $\beta_i$ 가 (확률변수가 아니라) 상수인 경우만 취급한다.

LRT의 결과를 OWA 경우과 비교해서 요약하면 다음과 같다. 첫째, 식 (5.5.7)의 $H_0$ 와 $H_a$ 에서 공통으로 $\mu$ 와 $\beta_j$ 에 대한 최우추정량(겸 MVUE)은 각각 $\overline{Y}$ 와 ($\overline{Y_{\cdot j}} - \overline{Y}$) 이다. 둘째로, $\tau_i$ 에 대한 최우추정량(겸 MVUE)은 $H_a$ 하에서 (엄격히 하자면 "$H_o \cup H_a$" 하에서) ($\overline{Y_{i\cdot}} - \overline{Y}$) 이다. (비고: OWA의 $H_a$ 하에서 $\mu_i \equiv \mu + \tau_i$ 에 대한 최우추정량(겸 MVUE)는 $\overline{Y} + (\overline{Y_{i\cdot}} - \overline{Y}) = \overline{Y_{i\cdot}}$임.) 셋째로, $\sigma^2$ 에 대한 최우추정량은 $H_0$ 와 $H_a (\cap H_0)$ 하에서 각각

$$\widehat{\sigma_0^2} = \frac{SSE + SST_r}{IJ} \left(= \frac{TSS - SSB}{IJ}\right) \tag{5.5.8}$$

$$\widehat{\sigma^2} = \frac{SSE}{IJ} \tag{5.5.9}$$

이다. (비고: OWA에서는 $TSS = SSE + SST_r$.)

LR인 $\lambda$의 형태는 식 (5.2.7)과 같은데 (<비고 5.2.2> 참조), 식 (5.5.8)과 (5.5.9)를 대입해서 손질하면

$$\frac{MST_r}{MSE} \equiv \frac{SST_r/(I-1)}{SSE/(I-1)(J-1)} \sim F(I-1,(I-1)(J-1)) \tag{5.5.10}$$

을 얻는다 (식 (5.3.8) 참조). 참고로, 식 (5.5.2)에서 $\{SSE,\ SST_r,\ SSB\}$는 서로 독립이고, 귀무가설 하에서 $SSE/\sigma^2 \sim \chi^2((I-1)(J-1))$, $SST_r/\sigma^2 \sim \chi^2(I-1)$, $SSB/\sigma^2 \sim \chi^2(J-1)$ 이다. 또한 식 (5.5.8)과 (5.5.9)를 불편추정량이 되도록 손질하면 MVUE로 $(SSE + SST_r)/(IJ-J)$와 MSE를 얻는다. 마지막으로, $H_a(\cap H_0)$ 하에서 $E(MST_r)$과 $E(MSB)$는 다음과 같다 (식 (5.4.5) 참조).

$$E(MST_r) = \sigma^2 + \frac{J}{I-1}\sum_{i=1}^{I}\tau_i^2$$

$$E(MSB) = \sigma^2 + \frac{I}{J-1}\sum_{j=1}^{J}\beta_j^2$$

### 5.5.3 ANOVA 뒷처리

§5.5.1의 예제에서는 $H_0$인 "$\tau_1 = \tau_2 = \tau_3 = \tau_4 (= 0)$"을 (유의수준 5%에

서) 기각했다. 그러니까 $H_a$ 인 □Not $H_0$□을 채택한 셈인데 (<비고 4.4.1>참조), 이 때 유의할 점은 대립가설이 □모든 $\tau_i$ 가 0 이 아님□이 아니라 □ $\tau_i$ 중에서 최소한 하나는 0 이 아님□이라는 점이다. 이에 따라, 어느 $\tau_i$ 가 0 이 아닌지 알아보는 뒷처리가 필요한데, 이를 Post-ANOVA Analysis 라 한다.

$\tau_i$ 에 대한 최우추정량(겸 MVUE)는 $(\overline{Y_{i\cdot}} - \overline{Y})$ 인데, 예제에서는 $\tau_3$ 와 $\tau_4$ 에 대한 추정치인 $(\overline{y_{3\cdot}} - \overline{y})$와 $(\overline{y_{4\cdot}} - \overline{y})$ 가 모두 0 이므로 $\tau_3$ 와 $\tau_4$ 에 대한 검정은 해 볼 필요조차 없다. (비고: 관찰된 신호의 크기가 0 임.) 반면에, $\tau_1$ 과 $\tau_2$ 에 대한 검정은 해볼만한데, 유의할 점은 이 두 가지 검정이 서로 독립이 아니라는 점이다. 이는 "$\sum_{i=1}^{I} \tau_i = 0$"이라는 제약 때문인데, 예를 들어 "$H_0 : \tau_1 = 0$"를 기각하면 "$H_0 : \tau_2 = 0$"도 기각하게 된다. 이러한 경우에는 둘을 묶어서 "$H_0 : \tau_1 = \tau_2$, $H_a : \tau_1 \neq \tau_2$"를 검정하면 효과적이다. (비고:검정통계량도 더 간단하거니와 검정력도 더 강하다.)

"$H_0 : \tau_1 = \tau_2$"에 대한 검정 역시 LRT 이다. 그런데, 이번에는 검정통계량을 구하는 기계적인 과정은 생략하고 대신에 SN비의 관점으로 검정통계량을 해석하기로 한다. 귀무가설은 "$\tau_1 - \tau_2 = 0$"과 마찬가지이고, $(\tau_1 - \tau_2)$ 에 대한 최우추정량(겸 MVUE)는 $(\overline{Y_{1\cdot}} - \overline{Y_{2\cdot}})$ 이다. 신호의 역할을 하게될 $(\overline{Y_{1\cdot}} - \overline{Y_{2\cdot}})$ 의 분포는 다음과 같이 구한다. 식 (5.5.5)에서 $Y_{ij}$, $i = 1, \cdots, I$, $j = 1, \cdots, J$ 는 서로 독립이다. (비고: 평균이 다르므로 *iid* 확률변수는 아님.) 따라서, <비고 2.15.1>에 의해서 $\overline{Y_{i\cdot}} \sim N(\mu + \tau_i, \sigma^2/J)$ 이고, 또한 $(\overline{Y_{1\cdot}} - \overline{Y_{2\cdot}}) \sim N(\tau_1 - \tau_2, 2\sigma^2/J)$ 이다. (참고로, $\tau_i$ 에 대한 추정량인 $(\overline{Y_{i\cdot}} - \overline{Y})$ 의 분포는 $N(\tau_i, \sigma^2(I-1)/IJ)$ 이다.) 그러니까, 만약 $\sigma^2$ 이 알려졌더라면 귀무가설 하에서 $N(0, 1^2)$ 분포를 따르는

$$\frac{\overline{Y_{1.}} - \overline{Y_{2.}}}{\sqrt{2\sigma^2/J}} \tag{5.5.11}$$

를 검정통계량으로 사용했을 것이다. (비고: $\sqrt{2\sigma^2/J}$는 잡음의 역할을 함) 그런데, $\sigma^2$을 모르므로 이를 MVUE인 $MSE$로 대체하면 $t$분포를 따르는 검정통계량이 된다. 이때 유의할 점은 다음과 같다. 가설 "$\tau_1 = \tau_2$"가 $\tau_1$과 $\tau_2$에 대해서만 언급했다고 해서 $i=1,2$에 해당되는 표본만 사용하는 것은 아니다. (비고: $MSE$는 $i=1,2$ 뿐만 아니라 $i=3,4$에 해당되는 표본까지 사용해서 얻은 것임.) 따라서, 식 (5.5.11)의 $\sigma^2$을 $MSE$로 대체하면

$$\frac{\overline{Y_{1.}} - \overline{Y_{2.}}}{\sqrt{MSE(2/J)}} \sim t((I-1)(J-1)) \tag{5.5.12}$$

가 되는데, 이는 귀무가설 "$\tau_1 = \tau_2$"하에서 $t$분포를 따르며 자유도는 $MSE$의 자유도와 같다.

식 (5.5.12)에 $\overline{y_{1.}} = 2$, $\overline{y_{2.}} = 4$, $MSE = 0.6$, $I = J = 4$를 대입하면 $t = -3.464$를 얻는데, 자유도가 9인 T-검정에서 $\alpha = 0.05$에 대한 TTT의 기각역은 "$t > 2.262$ 또는 $t < -2.262$"이므로 귀무가설 "$\tau_1 = \tau_2$"를 기각한다. 또한, $\alpha = 0.01$에 대해서도 $t = -3.464 < -3.250$이므로 귀무가설을 기각한다. (참고로, "$H_0 : \tau_1 = 0$, $H_a : \tau_1 \neq 0$"에 대해서는 $t = -2.83$을 얻으므로 $\alpha = 0.05$로는 $H_0$를 기각하지만 $\alpha = 0.01$로는 $H_0$를 채택한다.)

아울러, $(\tau_1 - \tau_2)$에 대한 95%와 99% 신뢰구간으로

$$(\overline{y_{1.}} - \overline{y_{2.}}) \pm 2.262\sqrt{MSE(2/J)} = -2 \pm 1.306$$

$$(\overline{y_{1.}} - \overline{y_{2.}}) \pm 3.250\sqrt{MSE(2/J)} = -2 \pm 1.876 \qquad (5.5.13)$$

를 얻는다.

<비고 5.5.5> 식 (5.5.12)와 신뢰구간에서 동일한 $MSE$를 사용하므로, □귀무가설 □$\tau_1 - \tau_2 = 0$ □을 기각함□과 □($\tau_1 - \tau_2$)에 대한 신뢰구간이 0을 포함하지 않음□은 동치이다.

신뢰구간은 TTT의 대용품이 되기도 하고 (<비고 5.5.5> 참조), 또한 그 자체로도 의미가 있으므로 몇 가지 더 구해보기로 한다. 첫째, $\mu_i = \mu + \tau_i$에 대한 추정량은

$$\overline{Y} + (\overline{Y_{i.}} - \overline{Y}) = \overline{Y_{i.}} \sim N(\mu + \tau_i,\ \sigma^2/J)$$

이므로, 예를 들어 $\mu_1 \equiv \mu + \tau_1$에 대한 95% 신뢰구간은

$$\overline{y_{1.}} \pm 2.262\sqrt{MSE/J} = 2 \pm 0.923 \qquad (5.5.14)$$

이다. 둘째로

$$\mu_1 - \mu_2 \equiv (\mu + \tau_1) - (\mu + \tau_2) = \tau_1 - \tau_2$$

이므로 ($\mu_1 - \mu_2$)에 대한 신뢰구간은 ($\tau_1 - \tau_2$)에 대한 신뢰구간인 식 (5.5.13)과 같다.

참고로, 식 (5.5.13)과 (5.5.14)는 OWA에서도 유효한데, 다만 $MSE$의 값이 달라

질 뿐이다. 물론, OWA 에서는 $n_1 \neq n_2$ 일 수 있으므로 식 (5.5.14)의 $J$는 $n_1$ 으로 고치고, 식 (5.5.13)의 $2/J$는 $(1/n_1 + 1/n_2)$ 로 고치면 된다.

마지막으로, TWA 의 주 목적은 식 (5.5.7)을 검정하는 것이지만, 부산물로 식 (5.5.6)에 대한 검정결과도 얻는다. 식 (5.5.6)에 대한 검정통계량은 다음과 같다 (식 (5.5.10) 참조).

$$\frac{MSB}{MSE} = \frac{SSB/(J-1)}{SSE/(I-1)(J-1)} \sim F(J-1,(I-1)(J-1))$$

이에, $MSB = 16/3 = 5.3$ 과 $MSE = 6/9 = 0.6$ 을 대입하면 $f = 8 > 6.99$ 이므로 $\alpha = 0.01$ 에서도 식 (5.5.6)을 기각할 수 있다.

### 5.5.4 $T$-test for Matched Samples

OWA 에서 $I = 2$ 인 경우를 □T-test for Independent Samples□라 불렀듯이 (<비고 5.2.1> 참조), TWA 에서는 $I = 2$ 인 경우를 □T-test for Matched Samples□ 또는 □Pairwise T-test□라 부른다.

먼저, §5.5.1의 예제에서 $I = 1,2$ 에 대해서만 TWA 를 적용한 결과는 다음과 같다.

| Source (of Variation) | $SS$ | 자유도 | $MS$ | $f$ |
|---|---|---|---|---|
| Treatment | 8 | 1 | 8 | 8 |
| Block | 5 | 3 | | |
| Error | 3 | 3 | 1 | |
| Total | 16 | 7 | | |

유의수준을 5%로 잡으면, 분자와 분모 자유도가 각각 1과 3인 $F$-test에서 UTT의 기각역은 $f > 10.13$ 이므로 귀무가설 "$\tau_1 = \tau_2$"를 기각하지 못한다.

이와 동일한 결과를 $T$-test로도 얻을 수 있는데, 흥미로운 점은 (실제로는) 모집단이 8개인 문제를 마치 모집단이 하나인 문제처럼 처리할 수 있다는 것이다 (§ 4.4.2 참조). 먼저, $Y_{1j}$와 $Y_{2j}$를 짝(match)을 지어서 그 차이(difference)를 $D_j$라 하자. 예를 들어, $D_j = Y_{1j} - Y_{2j}$는 젖소 $j$의 우유생산량의 차이인데, 이 차이가 두 종류의 사료의 차이에 기인한다는 것이 대립가설의 입장이다. 다음, $\{D_1, \cdots, D_J\}$를 크기가 $J$인 표본으로 간주하는데, 모집단의 분포는 $N(\delta, \sigma_D^2)$이라 가정한다. 그리고는

$$H_0 : \delta = 0, \; H_a : \delta \neq 0 \tag{5.5.15}$$

를 검정하는 것이다. 따라서, 기각역은 식 (4.4.10)의 형태가 된다.

예제에서 관찰된 표본 $\{d_j = y_{1j} - y_{2j} \; ; \; j = 1,2,3,4\}$는 $\{-2, -3, -3, 0\}$이므로, 이로부터

$$\overline{d} = \sum_{j=1}^{4} d_j / 4 = -2$$

$$\sigma_D^2 = \sum_{j=1}^{4} (d_j - \overline{d})^2 / (4-1) = 2$$

를 얻은 다음, $t$값을 구하면

$$t \equiv \frac{\overline{d} - 0}{\widehat{\sigma}_D / \sqrt{J}} = \frac{-2}{\sqrt{2}/\sqrt{4}} = -\sqrt{8} (\approx 2.828)$$

을 얻는다. 유의수준을 5%로 잡으면, 자유도가 $J-1=3$ 인 $T$-test 에서 TTT의 기각역은 "$t>3.182$ 또는 $t<-3.182$"이므로 귀무가설 "$\delta=0$"를 기각하지 못한다. (비고: $f=8=t^2$, $10.13=(3.182)^2$.)

## §5.6 선형모형

이제 ANOVA를 마무리하는 동시에 6장의 회귀분석을 대비할 때가 되었다.

TWA의 기본가정은 다음과 같다. 모집단이 $IJ$개 있는데, 모분포는

$$Y_{ij} \sim N(\mu + \tau_i + \beta_j, \sigma^2)$$

이다 (식 (5.5.5) 참조). 단, $\tau_i$와 $\beta_j$는 확률변수가 아니며 (<비고 5.5.4> 참조), $\sum_{i=1}^{I} \tau_i = 0$ 과 $\sum_{j=1}^{J} \beta_j = 0$ 를 만족시킨다.

TWA의 기본가정을 선형모형(linear model)으로 표현하면 다음과 같다.

$$Y_{ij} = \mu + \tau_i + \beta_j + \varepsilon_{ij}\ ,\ \varepsilon_{ij} \sim iid\ N(0,\sigma^2) \qquad (5.6.1)$$

즉, 확률변수 $Y_{ij}$는 (미지의) 상수 $\mu$, $\tau_i$, $\beta_j$에 $iid$ 확률변수인 $\varepsilon_{ij}$를 합친 것인데, 정규분포를 따르는 $\varepsilon_{ij}$의 평균은 0이고 분산은 ($i$, $j$와 무관하게) $\sigma^2$이다. 한편, 귀무가설 "$\tau_1 = \cdots = \tau_I$" 하에서의 선형모형은 다음과 같다.

$$Y_{ij} = \mu + \beta_j + \varepsilon_{ij}\ ,\ \varepsilon_{ij} \sim iid\ N(0,\sigma^2) \qquad (5.6.2)$$

<비고 5.6.1> 식 (5.6.1)을 CM (complete model)이라 하고 식 (5.6.2)를 RM (reduced model)이라 한다.

반면에, OWA에 대한 CM과 RM은 다음과 같다(식 (5.2.1) 참조).

$$CM:\ Y_i=\mu+\tau_i+\varepsilon_i$$
$$RM:\ Y_i=\mu+\varepsilon_i,\ \varepsilon_i\sim iid\ N(0,\sigma^2)$$

사실 TTT인 $T$-test 들도 모두 선형모형의 틀에 속한다. 예를 들어, §4.4.2 에서 다룬 "$H_0:\mu=\mu_0,\ H_a:\mu\neq\mu_0$" 경우는

$$\begin{aligned} &CM:\ Y=\mu+\varepsilon \\ &RM:\ Y=\mu_0+\varepsilon,\ \varepsilon\sim N(0,\sigma^2) \end{aligned} \tag{5.6.3}$$

에 해당된다. (비고: $\mu$ 는 미지의 상수이고 $\mu_0$ 는 구체적인 수치임.) 또한, §5.5.4의 예제는 TWA 에 해당되기도 하지만

$$CM:\ D=\delta+\varepsilon$$
$$RM:\ D=0+\varepsilon,\ \varepsilon\sim iid\ N(0,\sigma_D^2)$$

에 해당되기도 한다( 식 (5.5.15) 참조).

6장에서 다룰 회귀모형 (regression model) 역시 선형모형이다. 그런데, 가장 큰 차이점은 다음과 같다. OWA에서는 $Y_i$ 의 $i$ 가 그리고 TWA 에서는 $Y_{ij}$ 의 $i$ 와 $j$ 가 자연수인 반면에, 회귀모형에서는 $Y_x$ 의 $x$ 가 연속적인 값을 가질 수 있다. 구체적으로, 단순(simple) 회귀모형에서는

$$Y_x=\beta_0+\beta_1 x+\varepsilon_x,\ \varepsilon_x\sim iid\ N(0,\sigma^2) \tag{5.6.4}$$

이고, 다중(multiple) 회귀모형에서는

$$Y_x=\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\varepsilon_x,\ \varepsilon_x \sim iid\ N(0,\sigma^2) \tag{5.6.5}$$

이다. (단, 식 (5.6.5)에서는 $x$가 벡터 $(x_1,x_2,\cdots)$를 의미함.)

예를 들어, 식 (5.6.4)에서 $x$는 학부과정의 성적(또는 평점)이고, $Y_x$는 학부성적이 $x$인 학생이 대학원에 진학할 경우에 대학원에서 취득할 성적을 의미한다. 또한, 식 (5.6.5)에서도 $Y_x$가 대학원에서 취득할 성적이라 하고 $x_1$을 학부성적이라 하면, $x_2,x_3,\cdots$는 대학원 성적에 영향을 미칠 수 있는 다른 변수들을 의미한다. 예를 들면

$$x_2=\begin{cases}1\ , \text{ if 모교출신}\\ 0\ , \text{ if 타교출신}\end{cases} \tag{5.6.6}$$

인데, 이 경우 $x_2$를 dummy variable이라 부른다.

<비고 5.6.2> 단순회귀모형은 다중회귀모형의 RM이다.

<비고 5.6.3> ANOVA는 dummy variable만 있는 다중회귀모형으로 표현할 수 있다.

예를 들어, OWA의 CM은 "$Y_x=\mu+\tau_1 x_1+\tau_2 x_2+\cdots+\tau_I x_I+\varepsilon_x$"에서

$$x_i=\begin{cases}1\ , & \text{if 사료 } i\text{를 먹인 젖소}\\ 0\ , & otherwise\end{cases}$$

인 경우에 해당된다. (단, $\sum_{i=1}^{I} \tau_i = 0$ 이므로 $\tau_I$대신에 $\square - \sum_{i=1}^{I-1} \tau_i \square$를 대입함.)

<비고 5.6.4> 이 책의 예제(§5.3.2,§5.5.1 참조)에서와 같이 동일한 표본을 사용하여 OWA와 TWA를 하는 경우에는 OWA의 CM이 TWA의 RM이 된다.

# 제 6 장 회귀분석



## §6.1 서론 및 용어

회귀(regression)라는 용어의 유래는 다음과 같다. 유전학자 Galton(1822-1911)의 연구결과에 의하면, 키가 작은 (큰) 아버지를 가진 아들의 키는 평균치보다 작지만 (크지만) 아버지의 키보다는 커서 (작아서) 평균치 쪽으로 "회귀"하는 경향이 있다고 한다(문헌 [4] 참조). 오늘날에는 2개 이상의 변수들간의 관계식을 찾아내고, (이는 추정에 해당됨) 이 관계식의 타당성과 정확성을 검토하는 (이는 검정에 해당됨) 통계적 방법을 회귀분석(regression analysis)이라 한다.

변수들 간의 관계식이 직선 또는 평면인 모형을 선형(linear) 회귀모형이라 하고 곡선 또는 곡면인 모형을 비선형(nonlinear) 회귀모형이라 하는데, 이 책에서는 선형 회귀모형만 다룬다.(§6.5.5 참조). 그리고, 직선 관계식을 회귀직선(regression line)이라 하고, 평면 관계식을 회귀평면(regression plane)이라 한다. 또한, 회귀직선에 관련된 추정 및 검정은 SLR(simple linear regression)이라 하고, 회귀평면에 관련된 추정 및 검정을 MLR(multiple linear regression)이라 한다.

아래의 표는 아버지와 아들의 키를 (㎝ 단위로) 나타낸 것이다.

| $i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 아버지의 키 ( $x_i$) | 156 | 159 | 168 | 177 | 183 | 183 | 192 |
| 아들의 키 ( $y_i$) | 166 | 166 | 169 | 175 | 183 | 186 | 187 |

( $x_i$, $y_i$), $i=1,\cdots,7$은 $xy$ 평면상의 7개의 점에 해당된다. 이 경우, 회귀직선은 $y=60+0.6\dot{3}x$ 인데, (§6.2.1 참조), 이 직선은 주어진 7개의 점에 가장 잘 들어맞는 직선이다.

<비고 6.1.1> "가장 잘 들어맞는 직선"이란 "주어진 점에서 직선까지의 수직 방향 (즉, $y$방향) 거리를 제곱한 값들의 총합을 최소가 되게 하는 직선"을 의미한다.

<비고 6.1.2> $x_i$의 평균은 $\overline{x}=174$ 이고 $y_i$의 평균은 $\overline{y}=176$ 인데, 회귀직선 $y=60+0.6\dot{3}x$는 점 $(\overline{x},\overline{y})$를 통과한다.

<비고 6.1.3> Galton의 연구결과는 회귀직선의 기울기가 0보다 크고 1보다 작음을 의미한다.

§5.6에서 SLR 모형을 식 (5.6.4)로 표현했다. 즉,

$$Y_x=\beta_0+\beta_1 x+\varepsilon_x,\ \varepsilon_x\sim iid\ N(0,\sigma^2) \qquad (6.1.1)$$

인데, 회귀직선의 계수인 60과 $0.6\dot{3}$은 각각 $\beta_0$와 $\beta_1$에 대한 최우추정치이다.

<비고 6.1.4> 회귀직선의 계수는 통계학적으로 최우추정치이지만 기하학적으로는 <비고 6.1.1>에 의한 것이다. 이에 따라, 회귀직선의 계수를 MLE인 동시에 LSE(least squares estimate)라 부른다. 또한, LSE의 "E"는 estimate 뿐만 아니라 estimator와 estimation을 의미하기도 한다 (<비고 1.6.1> 참조).

그리고, (§3.5.1에 등장한) MLE의 불변성(invariance)에 의해서 $(60+0.6x)$는 $(\beta_0+\beta_1 x)$에 대한 최우추정치인데, $(\beta_0+\beta_1 x)$는 바로 $E(Y_x)$이다. (비고: 식 (6.1.1)에서 $E(\varepsilon_x)=0$ 이므로 $E(Y_x)=\beta_0+\beta_1 x$임.)

회귀직선 $y=60+0.6x$가 $E(Y_x)$에 대한 최우추정치라면 $Y_x$는 과연 무엇인가? 식 (6.1.1)에 의하면, 키가 $x$인 아버지를 가진 아들의 키가 $Y_x$이다. 그러나, 키가 $x$인 아버지가 여럿인 경우에 (또는, 아버지가 같더라도), 아들들의 키가 모두 같지는 않다. 즉, $Y_x$는 상수가 아니라 확률변수인데, 식 (6.1.1)은 바로 $Y_x$가 평균이 $(\beta_0+\beta_1 x)$이고 분산이 $\sigma^2$인 정규분포를 따른다는 가정인 셈이다. 그렇다면, 무엇이 모집단이고 무엇이 표본인가? 사실 $Y_x$의 분포가 바로 모분포이다. 그러나, $x$ 값이 다르면 (모평균 "$\beta_0+\beta_1 x$"가 달라지기 때문에) 모분포도 다르고 또한 모집단도 다르다. 예제에서는 모두 6개의 모집단이 등장하고 (비고 : $x_5=x_6=188$), 관찰된 표본의 수 역시 6개인데 (<비고 5.2.1> 참조), 이 중에서 크기가 1인 표본은 5개이고 크기가 2인 표본은 하나이다.

회귀분석의 첫 단계는 회귀직선 또는 회귀평면을 추정하는 것이다. 그리고, 둘째 단계에서는 $\beta_0, \beta_1, \cdots$에 대한 검정을 한다. 또는 $\beta_0, \beta_1, \cdots$을 하나로 묶어서 회귀모형 자체에 대한 타당성(validity)을 검정하기도 한다. 그러나, 첫째와 둘째 단계는 준비작업일 뿐이고, 회귀분석의 주목적은 예측(prediction)이다. 예를 들어 아버지의 키

가 180인 경우 아들의 키에 대한 예측치는 $60+(0.6)(180)=168$이다. 또한, 아버지의 키가 174이면 아들의 키의 예측치는 176이다.(<비고 6.1.2) 참조). 물론, 예측의 정확성을 검정하는 것도 회귀분석에 포함된다.

<비고 6.1.5> 회귀분석에서의 예측은 “prediction"이고, 시계열분석(time-series analysis)에서의 예측은 ”forecasting" 임.

$x$와 $Y_x$에 대한 호칭은 여러 가지가 있다. (비고: 식 (5.6.5)의 MLR 모형에서는 $x$가 벡터 $(x_1, x_2, \cdots)$를 의미함.) 첫째, 아들의 키가 크고 작음이 아버지의 키에 의해서 설명된다는 의미에서 $x$를 설명변수(explanatory variable)라 하고, $x$에 반응하는 것이 $Y_x$라는 의미에서 $Y_x$를 반응변수(response variable)라 한다. 둘째로, $x$를 회귀변수 또는 예측변수라 하고 $Y_x$를 피회귀변수 또는 피예측변수라 하기도 한다. 그러나, 가장 흔히 사용되는 호칭은 독립변수($x$)와 종속변수($Y_x$)이다. 이는 $Y_x$가 $x$의 함수라는 점을 강조하는 것이기도 하지만, 이때 “독립”이라는 표현은 $x$가 확률변수가 아니라는 점을 암시하는 것이기도 하다. $x$는 (확률변수가 아닐뿐더러) 많은 경우에 제어변수(control variable)의 역할까지 한다. 예를 들어, $x$는 광고비이고 $Y_x$는 매출액이라 하자. 매출액을 직접 결정할 수는 없다. 다만, 직접 결정할 수 있는 광고비를 통해서 (즉, 광고비를 제어 또는 조절함으로써) 간접적으로 매출액에 영향을 끼칠 수 있다.

독립변수가 하나일 때, $E(Y_x)=\beta_0+\beta_1 x$에 대한 추정치(또는 추정식)를 회귀직선이라 했다. 그리고, 독립변수가 둘이면 $E(Y_x)=\beta_0+\beta_1 x_1+\beta_2 x_2$에 대한 추정치는 회귀평면이 된다. 그런데, 독립변수가 셋 이상인 경우에도 여전히 $E(Y_x)$에 대한 추정치를 회귀평면이라 한다 (엄격히 하자면 평면이 아니라 초평면(hyperplane)임).

## §6.2 LSE

### 6.2.1 SLR에 대한 LSE

§6.1에서 주어진 7개의 점에 가장 잘 들어맞는 직선은 $y=60+0.6x$라고 했는데 이를 먼저 확인해 보자 (<비고 6.1.1> 참조).

회귀직선을 $\hat{y}=\hat{\beta}_0+\hat{\beta}_1 x$라 하면

$$f_i \equiv y_i-\hat{y}_i \equiv y_i-(\hat{\beta}_0+\hat{\beta}_1 x_i) \tag{6.2.1}$$

는 점 $(x_i, y_i)$로부터 회귀직선까지의 수직 방향 (또는 $y$방향) 거리를 나타낸다. 그리고, 이들의 제곱합(SS: sum of squares)을 $SSE$라 하자(§6.3.1 마지막 문단 참조). 즉,

$$SSE=\sum_i f_i^2=\sum_i \{y_i-(\hat{\beta}_0+\hat{\beta}_1 x_i)\}^2 \tag{6.2.2}$$

이다. 그러면, LSE의 정의에 의해 (<비고 6.1.4> 참조), $\hat{\beta}_0$과 $\hat{\beta}_1$은 아래의 식을 만족시킨다.

$$\frac{\partial SSE}{\partial \hat{\beta}_0}=-2\sum_i f_i=0 \tag{6.2.3}$$

$$\frac{\partial SSE}{\partial \hat{\beta}_1}=-2\sum_i x_i f_i=0 \tag{6.2.4}$$

식 (6.2.3)으로부터는 다음의 관계식을 쉽게 얻을 수 있다 (<비고 6.1.2> 참조).

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \tag{6.2.5}$$

반면에, 식 (6.2.4)는 약간의 손질이 필요한데 결과는 다음과 같다.

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \tag{6.2.6}$$

§6.1의 예제에서는 $\bar{x} = 174$, $\bar{y} = 176$, $\sum_{i=1}^{7} (x_i - \bar{x})(y_i - \bar{y}) = 720$, $\sum_{i=1}^{7} (x_i - \bar{x})^2 = 1080$ 이므로, 먼저 $\hat{\beta}_1 = 720/1080 = 2/3$을 얻고 나서 이를 식 (6.2.5)에 대입하면 $\hat{\beta}_0 = 60$을 얻는다.

<비고 6.2.1> 수직방향 거리는 최단거리가 아니다. 최단거리의 제곱합을 최소가 되게 하는 직선은 직교(orthogonal)회귀직선이라 하는데, 이는 다변량 (multivariate) 분석의 범주에 속한다(§6.6.4 참조).

### 6.2.2 LSE의 역학적 해석

LSE 방법으로 얻은 회귀직선은 힘의 평형으로 해석할 수 있다(문헌[3]참조). 회귀직선을 단단한 막대기라 하고, 식 (6.2.1)의 $f_i$를 막대기에 작용하는 힘이라 하자. 즉, 막대기의 $(x_i, \hat{y}_i)$ 지점에 크기가 $f_i$인 힘이 수직방향으로 작용한다고 하자. (비고: $f_i > 0$ 이면 막대기를 위로 잡아 당기고, $f_i < 0$ 이면 아래로 잡아 당김.)

막대기가 (움직이지 않고) 평형상태에 있을 조건은 두 가지이다. 첫째는 $\sum_i f_i = 0$ 인데, 이는 식 (6.2.3)에 해당된다. 합력이 0이면 최소한 막대기의 중심인(?) $(\overline{x}, \overline{y})$는 고정된다 (<비고 6.1.2> 참조). 그러나, 여전히 $(\overline{x}, \overline{y})$를 축으로 회전운동은 할 수 있다. 이러한 회전운동을 방지하기 위한 조건은 바로 식 (6.2.4)인데, 이를

$$\sum_i (x_i - \overline{x}) f_i = 0 \tag{6.2.7}$$

으로 고치면 이해하기 쉽다. (비고: 식 (6.2.7)은 식 (6.2.3)과 (6.2.4)로부터 얻음.) 식 (6.2.7)에서 $(x_i - \overline{x}) f_i > 0$ 이면 시계반대방향으로 그리고 $(x_i - \overline{x}) f_i < 0$ 이면 시계방향으로 회전효과(torque)가 작용하는데, 이들의 합이 0이면 (회전하지 않고) 평형상태가 된다.

### 6.2.3 MLR에 대한 LSE

먼저, 독립변수가 두 개인 경우를 다룬다. 주어진 점 $(x_{i1}, x_{i2}, y_i)$, $i = 1, \cdots, n$는 이제 3-차원 공간에 있는 $n$개의 점이 되고, 회귀평면 $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2$는 3-차원 공간에서 2-차원을 차지한다.

점 $(x_{i1}, x_{i2}, y_i)$로부터 회귀평면까지의 수직방향 (또는 $y$방향) 거리는

$$f_i \equiv y_i - \hat{y}_i \equiv y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}) \tag{6.2.8}$$

가 된다. 그리고, LSE인 $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2$는 여전히 $SSE \equiv \sum_i f_i^2$을 최소가 되게 한다. LSE의 조건식은 이제

$$-\frac{1}{2}\frac{\partial SSE}{\partial \hat{\beta}_0}=\sum_i f_i=0 \tag{6.2.9}$$

$$-\frac{1}{2}\frac{\partial SSE}{\partial \hat{\beta}_1}=\sum_i f_i x_{i1}=0 \tag{6.2.10}$$

$$-\frac{1}{2}\frac{\partial SSE}{\partial \hat{\beta}_2}=\sum_i f_i x_{i2}=0 \tag{6.2.11}$$

인데, 이들에 대한 역학적 해석은 다음과 같다. 회귀평면을 딱딱한 널빤지라 하면, 식 (6.2.9)는 널빤지에 작용하는 (수직방향) 힘의 합이 0임을 의미한다. 또한 기하학적으로는 점 $(\overline{x_1}, \overline{x_2}, \overline{y})$가 널빤지 상에 있음을 의미한다 (<비고 6.1.2> 참조). 합력이 0이더라도 회전운동은 가능한데, 이를 방지하기 위한 조건이 식 (6.2.10)과 (6.2.11)이다. 구체적으로 식 (6.2.10)은 널빤지가 $x_2$ 축(axis)을 축(pivot)으로 회전하는 것을 방지하고, 식 (6.2.11)은 $x_1$ 축(axis)을 축(pivot)으로 회전하는 것을 방지한다.

일반적으로, 독립변수가 $k$개인 경우에

$$f_i \equiv y_i - \hat{y}_i \equiv y_i - (\hat{\beta}_0 + \sum_{j=1}^{k} x_{ij}\hat{\beta}_j),\ i = 1, \cdots n, \tag{6.2.12}$$

이라 하면, LSE인 $\hat{\beta}_0$와 $\hat{\beta}_j$, $j = 1, \cdots$ 를 구하는 방정식은

$$\begin{aligned}&\sum_{i=1}^{n} f_i = 0\\ &\sum_{i=1}^{n} f_i x_{ij} = 0,\ \ j=1,\cdots,k\end{aligned} \tag{6.2.13}$$

인데, 이를 벡터와 행렬로 표현하면 풀기가 쉽다.

다음과 같이 모든 벡터는 열(column)벡터로 정의한다.

$$f \equiv \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{bmatrix}, \quad y \equiv \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \hat{y} \equiv \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix}, \quad \hat{\beta} \equiv \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix} \tag{6.2.14}$$

그리고, 행렬 $X$는 다음과 같이 정의한다.

$$X \equiv \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & & & & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix} \tag{6.2.15}$$

그러면, 식 (6.2.12)과 (6.2.13)은 각각

$$f = y - \hat{y} = y - X\hat{\beta} \tag{6.2.16}$$

$$X'f = 0 \tag{6.2.17}$$

이 된다. (비고: 식 (6.2.17)에서 $X'$은 $X$의 transpose이고 우변의 0은 0벡터를 의미함.) 따라서 식 (6.2,16)을 식 (6.2.17)에 대입하면

$$X'y - X'X\hat{\beta} = 0 \tag{6.2.18}$$

인데, 이를 $\hat{\beta}$에 대해서 풀면 다음을 얻는다.

$$\hat{\beta} = (X'X)^{-1}(X'y) \tag{6.2.19}$$

참고로 $SSE = \sum_{i=1}^{n} f_i^2 = f'f$에 식 (6.2.19)를 대입(하여, 간단히)하면 다음 식을 얻는다.

$$SSE = y'y - \hat{\beta}'X'y \tag{6.2.20}$$

### 6.2.4 MLR 예제

문헌 [9]의 예제 11.12는 다음과 같다.

$$y = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 3 \end{bmatrix}, \quad X = \begin{bmatrix} 1 & -2 & 4 \\ 1 & -1 & 1 \\ 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \end{bmatrix} \tag{6.2.21}$$

먼저, 식 (6.2.6)의 분자와 분모에 해당되는 $X'y$와 $(X'X)^{-1}$는

$$X'y = \begin{bmatrix} 5 \\ 7 \\ 13 \end{bmatrix}, \quad (X'X)^{-1} = \begin{bmatrix} 17/35 & 0 & -1/7 \\ 0 & 1/10 & 0 \\ -1/7 & 0 & 1/14 \end{bmatrix} \tag{6.2.22}$$

이다. 다음, 이들을 식 (6.2.19)에 대입하면

$$\hat{\beta}=\begin{vmatrix} 4/7 \\ 7/10 \\ 3/14 \end{vmatrix} \approx \begin{vmatrix} 0.5714 \\ 0.7000 \\ 0.2143 \end{vmatrix} \tag{6.2.23}$$

이므로, 회귀평면은

$$\hat{y} \approx 0.571 + 0.7\, x_1 + 0.214\, x_2 \tag{6.2.24}$$

가 된다. 또한, 식 (6.2.20)으로부터 다음을 얻는다.

$$SSE \approx 0.4571 \tag{6.2.25}$$

## §6.3 회귀모형의 MLE

### 6.3.1 LSE와 MLE

§6.2에서 구한 LSE가 MLE와 동일함을 보인다(<비고 6.1.4>참조). §6.2에서와 같이 $x_{ij}$를 $j$번째 ($j=1,\cdots,k$) 독립변수의 $i$번째 ($i=1,\cdots,n$) 관찰치라 하고, $y_i$를 종속변수의 $i$번째 관찰치라 하자. 이에 따라, §6.1에서 $Y_x$로 표현하던 종속변수를 지금부터는 $Y_i$로 표기한다. 그러면, 회귀모형은

$$Y_i = \mu_i + \varepsilon_i,\ i=1,\cdots,n \tag{6.3.1}$$

$$where\ \mu_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} \tag{6.3.2}$$

인데, 확률변수 $\varepsilon_i$의 분포에 대한 가정은 *iid* $N(0,\sigma^2)$이므로 결국

$$Y_i \sim N(\mu_i, \sigma^2),\ i=1,\cdots,n \tag{6.3.3}$$

이 된다.

<비고 6.3.1> $Y_1,\cdots,Y_n$은 (평균이 다르므로) 동일하지는 않지만, ($\varepsilon_1,\cdots,\varepsilon_n$이 *iid* 확률변수이므로) 서로 독립이다.

$Y_1,\cdots,Y_n$이 독립이므로, 이들의 결합밀도함수인 LF는

$$L=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu_i)^2} \tag{6.3.4}$$

이다 (식 (3.2.1) 참조). 그리고, 식 (6.3.4)에 자연대수를 취하면

$$\ln L=K-\frac{n}{2}\ln\sigma^2-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu_i)^2 \tag{6.3.5}$$

가 된다 (식 (3.2.2)이하 부분 참조). 이때, 유의할 점은 다음과 같다. 식 (6.3.5)를 $\beta_0, \beta_1, \cdots, \beta_k$에 대해서 (비고: $\mu_1, \cdots, \mu_n$은 $\beta_0, \beta_1, \cdots, \beta_k$의 함수임. 식 (6.3.2) 참조) 편미분한 식을 (0으로 놓고) 푸는 대신에, 마지막 항에 있는

$$\sum_{i=1}^{n}(y_i-\mu_i)^2 \tag{6.3.6}$$

를 편미분한 식을 풀어도 같은 결과를 얻는다. 즉, 식 (6.3.5)를 최대가 되게하는 $\beta_0, \beta_1, \cdots, \beta_k$ 값들을 식 (6.3.6)을 최소가 되게하는 $\beta_0, \beta_1, \cdots, \beta_k$와 동일하다. 그런데, LSE는 바로 식 (6.3.6)이 최소가 되는 조건식을 풀어서 얻은 것이다 (식 (6.2.8) 참조). 따라서, $\beta_0, \beta_1, \cdots, \beta_k$에 대한 MLE를 $\widehat{\beta}_0, \cdots, \widehat{\beta}_k$라 하면 이는 MLE인 동시에 LSE이다. (비고: MLE의 표기인 $\widehat{\beta}_0, \cdots, \widehat{\beta}_k$를 편의상 LSE의 표기로도 사용했음.)

<비고 6.3.2> 모분포에 대한 가정이 없는 LSE 방법은 일종의 heuristic 방법이다 (<비고 3.6.1> 참조). 따라서, 정규분포의 가정 하에 얻은 MLE가 LSE와 일치한다는 사실은 MLE의 robustness를 뒷받침하는 것이라 할 수 있다 (<비고 3.6.1>의 윗 문단 참조).

### 6.3.2 *SSE*

MLE인 $\hat{\beta}_0, \cdots, \hat{\beta}_k$를 식 (6.3.6)에 대입하면 §6.2에서 *SSE*라 부르던 것이 된다. 즉,

$$SSE = \sum_{i=1}^{n} ( y_i - \hat{\mu}_i )^2 \tag{6.3.7}$$

$$where\ \hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik} \tag{6.3.8}$$

이다. 그리고, 이를 식 (6.3.5)에 대입한 다음 $\sigma^2$에 대해서 편미분하면, $\sigma^2$에 대한 MLE로

$$\hat{\sigma}^2 = \frac{SSE}{n} \tag{6.3.9}$$

를 얻는다 (식(3.2.4)참조).

이제 용어에 대해서 한가지 짚고 넘어갈 때가 되었다. 회귀분석의 주목적은 예측이라 했다(<비고 6.1.5> 참조). 독립변수의 값이 $x_{ij}$일 때 ($i = 1, \cdots, n$, $j = 1, \cdots, k$), $Y_i$에 대한 예측치로 식 (6.3.8)의 $\hat{\mu}_i$를 사용한다. (비고: $\hat{\mu}_i$는 식 (6.2.12)의 $\hat{y}_i$와 같음.) 이때, 실제 관측치 $y_i$와 예측치 $\hat{\mu}_i$의 차이인 ($y_i - \hat{\mu}_i$)를 잔차(residual)라 부른다. (비고: §6.2에서는 잔차를 $f_i$로 포기하고, 이를 힘으로 해석했음.) 이에 따라, 식 (6.3.7)을 잔차제곱합(residual SS)이라 부르는 책이 많다. 그런데도 이 책에서 식 (6.3.7)을 *SSE*라 부르는 이유는 회귀모형의 *SSE*가 ANOVA의 *SSE*와 동일한 역할을 하기 때문이다.

사실 *SSE*는 모든 선형모형에서 동일한 역할을 한다. 즉, 모든 선형모형에서

$\sigma^2$에 대한 MLE는 $SSE/n$ 이다. 예를 들어, 식 (5.6.3)에서 $\sigma^2$에 대한 MLE는 식 (4.4.8)인데, 이때 $SSE$는 바로 3장에서 SS라 불렀던 것이다 (<비고 3.4.1> 참조).

이제, $SSE$의 의미와 용도를 더욱 확장시킨다. 지금까지 $SSE$라 부른 것은 5장의 ANOVA에서 정의된 $SSE$이다. 즉, 전체 Variation인 $TSS$ 중에서 "선형모형"에 의해서 설명된 부분을 빼고 남은 부분을 $SSE$라 불렀는데, 이때 "선형모형"이라 함은 CM을 의미한다 (<비고 5.6.1> 참조). 예를 들어, TWA의 $SSE$는 식 (5.6.1) 하에서 "설명안된 부분"인데 이를 $SSE_{CM}$이라 하자. 반면에, TWA의 RM인 식 (5.6.2) 하에서 "설명안된 부분"을 $SSE_{RM}$이라 하면

$$SSE_{RM} = SSE_{CM} + SST_r \tag{6.3.10}$$

의 관계가 성립한다. 즉, 식 (5.5.8)은 RM 하에서 $\sigma^2$에 대한 MLE인 반면에, 식 (5.5.9)는 CM 하에서 $\sigma^2$에 대한 MLE이다. 따라서, 식 (5.5.10)은

$$\frac{(SSE_{RM} - SSE_{CM})/(d_{RM} - d_{CM})}{SSE_{CM}/d_{CM}} \sim F(d_{RM} - d_{CM}, d_{CM}) \tag{6.3.11}$$

으로 표현할 수 있다. (비고: <비고 5.6.4>의 경우에는 $SSE_{RM} = SSE_{CM} + SSB$.)

### 6.3.3 회귀모형과 LRT

회귀모형을 포함한 모든 선형모형에서

$$\begin{aligned} H_0 &: RM \\ H_a(\cup H_0) &: CM \end{aligned} \tag{6.3.12}$$

에 대한 검정통계량은 식 (6.3.11)이다. TWA를 예로 들어서 설명했던 식 (6.3.11)을 이제 회귀모형으로 설명한다. 회귀모형의 LF는 식 (6.3.4)인데, LR은 여전히 식 (5.2.7)의 형태인

$$\lambda=(\frac{SSE_{CM}/n}{SSE_{RM}/n})\leq k \tag{6.3.13}$$

이다. 즉, 식 (5.2.7)에서 $\hat{\sigma}_2$은 CM 하에서 $\sigma^2$에 대한 MLE이고, $\hat{\sigma}_0^2$는 RM 하에서 $\sigma^2$에 대한 MLE이다.

회귀모형에서 CM은 식 (6.3.1)이다. 반면에, RM은 다양하게 정의할 수 있다. 한마디로, $\{\beta_1,\cdots,\beta_k\}$ 중에서 하나 이상을 0으로 놓으면 RM이 된다. 구체적으로, $\{\beta_1,\cdots,\beta_k\}$ 중에서 하나만 0으로 놓으면 식 (6.3.13)은 $T$-test가 되고, 둘 이상을 0으로 놓으면 $F$-test가 된다.

<비고 6.3.3> 일반적으로, $\beta_1,\cdots,\beta_k$의 (함수를 0으로 놓는) 제약식의 개수가 1이면 $T$-test 이고 2 이상이면 F-test 인데, 이 책에는 " $\beta_j=0$" 형태의 제약식만 등장함.

<비고 6.3.4> 제약식이 $\beta_1=\cdots=\beta_k=0$ 인 경우의 RM을 "Null Model"이라 하고, 이에 대한 $F$-test를 "$F$-test for Model"이라 한다.

Null Model 하에서 SLR의 회귀직선은 수평선이 되고, MLR의 회귀평면은 수평면이 된다. 그리고, 회귀직선 또는 회귀평면은 여전히 점 $(\overline{x_1},\cdots,\overline{x_k},\overline{y})$를 포함한다(<비고 6.1.2> 참조). 따라서, Null Model 하에서의 $SSE$를 $SSE_0$라 하면

$$SSE_0 \equiv \sum_{i=1}^{n}(y_i-\overline{y})^2 \tag{6.3.14}$$

인데, 이는 종전에 $TSS$라 부르던 것이다. 즉, Null Model 하에서 $\sigma^2$에 대한 MLE는 $SSE_0/n$이고, $\sigma^2$에 대한 MVUE는 $SSE_0/(n-1)$이다. 또한, $SSE_0/\sigma^2 \sim \chi^2(n-1)$이므로, RM이 Null Model인 경우 식 (6.3.11)의 $d_{RM}$은 $(n-1)$이다.

<비고 6.3.5> 제약식이 모두 "$\beta_j=0$" 형태인 경우 (<비고 6.3.3> 참조) CM과 RM 하에서의 독립변수의 개수를 각각 $k$와 $k'$이라 하면, 식 (6.3.11)에서 "$d_{CM}=n-(k+1)$"이고 "$d_{RM}=n-(k'+1)$"이다. 따라서 $(d_{RM}-d_{CM})=(k-k')$ 인데, $(k-k')$은 바로 RM 하에서의 제약식의 개수이다.

앞으로 자주 사용될 $MSE$는 바로 식 (6.3.11)의 분모를 의미한다. 즉,

$$MSE \equiv SSE_{CM}/(n-k-1) \tag{6.3.15}$$

이다.

### 6.3.4 $\hat{\beta}_i$의 분포

식 (6.3.11) 하나로 회귀분석에 관련된 모든 검정을 할 수 있다. 그러나, 여전히 $\hat{\beta} \equiv (\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k)'$의 확률분포가 필요하다. (비고: 행(row)벡터의 transpose는 열(column)벡터임.)

식 (6.2.19)의 $\hat{\beta}$은 $\beta = (\beta_0, \beta_1, \cdots, \beta_k)'$에 대한 점추정치이므로, $y = (y_1, \cdots, y_n)'$을 $Y = (Y_1, \cdots, Y_n)'$으로 대체하면 점추정량이 된다. 점추정량의 분포는 첫째로 $\beta_j$에 대한 신뢰구간을 구할 때 필요하다. 둘째로, 소위 예측구간(prediction interval)을 구하기 위해서는 $\hat{\beta}_0, \cdots, \hat{\beta}_k$간의 공분산까지 필요하다. 셋째로, 신뢰구간 및 예측구간 같은 구간추정을 할 때뿐만 아니라, 검정을 할 때에도 $\hat{\beta}$의 분포를 알면 식 (6.3.11)의 검정통계량을 간단히 얻을 수 있다.

$\beta$에 대한 점추정량인 $\hat{\beta} = (X'X)^{-1}(X'Y)$는 한마디로 $Y_1, \cdots, Y_n$의 (선형)함수이다. 따라서, <비고 6.3.1>과 <비고 2.15.1>에 의해서

$$\hat{\beta} \sim MVN(\beta, (X'X)^{-1}\sigma^2) \tag{6.3.16}$$

임을 보일 수 있다(증명은 생략함). 식 (6.3.16)에서 "$MVN$"은 $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$의 결합분포가 "Multivariate Normal" 분포임을 의미한다. 따라서, $\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$은 각각 정규분포를 따른다(§2.2.3 참조). 그리고, $E(\hat{\beta}) = \beta$이므로, $\hat{\beta}_0, \cdots, \hat{\beta}_k$는 모두 불편추정량이다. 반면에, $(X'X)^{-1}\sigma_2$은 공분산행렬(covariance matrix)이라는 것인데, 대각선(diagonal) 요소는 차례대로 $V(\hat{\beta}_0), V(\hat{\beta}_1), \cdots, V(\hat{\beta}_k)$이고 나머지는 $\hat{\beta}_0, \cdots, \hat{\beta}_k$간의 공분산이다.

또한, $\beta_0, \cdots, \beta_k$의 선형함수를

$$\sum_{j=0}^{k} a_j \beta_j = (\, a_0 \;\; a_1 \cdots \;\; a_k \,) \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \equiv a'\beta \tag{6.3.17}$$

로 표현하면, MLE의 불변성(invariance)에 의해서 $a'\beta$에 대한 MLE는 $a'\hat{\beta}$ 인데 이의 분포는

$$a'\hat{\beta} \sim N(a'\beta,\ a'(X'X)^{-1}a\sigma^2) \tag{6.3.18}$$

임을 보일 수 있다(증명은 생략함). (비고: $\hat{\beta}$가 $Y_1, \cdots, Y_n$의 선형함수이므로, $a'\hat{\beta}$ 역시 $Y_1, \cdots, Y_n$의 선형함수임.)

<비고 6.3.6> LSE인 동시에 MLE인 $\hat{\beta}$은 MVUE이기도 하다. 그런데, $\hat{\beta}$이 $Y_1, \cdots, Y_n$의 선형함수임을 강조하기 위해서, $\hat{\beta}$을 BLUE(best linear unbiased estomator)라 부르기도 한다. 또한, <비고 2.15.2>에서 $\overline{Y}$와 $\sum_{i=1}^{n} (Y_i - \overline{Y})^2$이 서로 독립이듯이, $\hat{\beta}$와 $SSE_{CM}$은 서로 독립이다. 그리고, 식 (6.3.15)의 $MSE$는 $\sigma^2$에 대한 MVUE이다.

## §6.4 SLR의 분석

### 6.4.1 SLR과 LRT

§6.1의 예제는 SLR에서 $n=7$인 경우인데, §6.2.1에서 $(\beta_0, \beta_1)$에 대한 추정치로 $(60, 0.\tilde{6})$을 얻었다. 따라서, $(\hat{\beta}_0, \hat{\beta}_1) = (60, 0.\tilde{6})$을 식 (6.2.2)에 대입하면 $SSE_{CM}=40$을 얻고, 이를 식 (6.3.15)에 대입하면 $MSE=40/5=8$을 얻는다.

SLR에서 CM은 다음과 같다 (식 (6.3.1) 참조).

$$Y_i=\beta_0+\beta_1 x_1+\varepsilon_i,\ i=1,\cdots,n \tag{6.4.1}$$

SLR에서 유의할 점은 Null Model인 (<비고 6.3.4> 참조)

$$Y_i=\beta_0+\varepsilon_i,\ i=1,\cdots,n \tag{6.4.2}$$

가 유일한 RM이라는 점이다. 따라서, 식 (6.3.14)로부터 $SSE_0=\sum_{i=1}^{7}(y_i-176)=520$을 얻는다. 그리고, 이들을 식 (6.3.11)에 대입하면

$$f=\frac{(520-40)/1}{40/5}=60 \tag{6.4.3}$$

을 얻는다. 그런데, 분자 자유도가 1이고 분모 자유도가 5인 F-test에서, $\alpha=5\%$ 와 $\alpha=0.5\%$에 대한 UTT 기각역은 각각 6.61과 22.78이므로, 귀무가설인 RM을 (식 (6.3.12)참조) $\alpha=0.5\%$에서조차 기각할 수 있다.

통계 패키지에 의한 ANOVA Table은 다음과 같다.

| Source of Variation | SS | 자유도 | MS | $F$ | $PR>F$ | R-SQUARE |
|---|---|---|---|---|---|---|
| Model | 480 | 1 | 480 | 60 | 0.0006 | 0.9231 |
| Error | 40 | 5 | 8 | | | |
| Total | 520 | 6 | | | | |

위의 Table에서 "$PR>F$"는 $p$-value이다(§4.6.3 참조). 그리고, "R-SQUARE"는

$$R^2 \equiv \frac{SSM}{TSS} = \frac{480}{520} = 0.9231 \tag{6.4.4}$$

인데, 이는 전체 Variation인 TSS(또는 $SSE_0$) 중에서 회귀모형에 의해서 설명된 부분인 $SSM$(SS for Model)이 차지하는 비율이다. (비고: $SSM = SSE_0 - SSE_{CM}$.) 따라서, 전체 Variation 중에서 92.31%가 회귀모형에 의해서 설명되었다고 할 수 있다.

### 6.4.2 $\beta_1$에 대한 추론

SLR 경우 식 (6.3.16)의 $(X'X)^{-1}$는 다음과 같다(계산은 생략함).

$$(X'X)^{-1} = \begin{bmatrix} \frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2} & \frac{-\bar{x}}{\sum (x_i - \bar{x})^2} \\ \frac{-\bar{x}}{\sum (x_i - \bar{x})^2} & \frac{1}{\sum (x_i - \bar{x})^2} \end{bmatrix} \tag{6.4.5}$$

따라서, $\hat{\beta}_1$의 분포는

$$\hat{\beta}_1 \sim N(\beta_1, \sigma^2/\sum_{i=1}^{n}(x_i-\bar{x})^2) \tag{6.4.6}$$

인데, $\sigma^2$을 식 (6.3.15)의 MSE로 대체하면 (<비고 6.3.6> 참조)

$$T_{n-2}=\frac{\hat{\beta}_1-\beta_1}{\sqrt{\mathrm{MSE}/\sum(x_i-\bar{x})^2}} \tag{6.4.7}$$

를 얻는다. (비고: $t$분포의 자유도 $(n-2)$는 MSE의 자유도임.)

먼저, 식 (6.4.7)을 PQ(pivotal quantity)로 사용하면 $\beta_1$에 대한 95% 신뢰구간은

로

$$\begin{aligned}\hat{\beta}_1 \pm 2.571\sqrt{\mathrm{MSE}/\sum(x_i-\bar{x})^2} &= 0.\dot{6} \pm 2.571\sqrt{8/1080} \\ &= 0.\dot{6} \pm 0.2213\end{aligned} \tag{6.4.8}$$

을 얻는다. (비고 : $0.025 = P(T_5 > 2.571) = P(T_5 < -2.571)$.)

다음, "$H_0: \beta_1 = \beta_{10}$"에 대한 검정통계량은 식 (6.4.7)의 $\beta_1$을 $\beta_{10}$로 대체한 것이다. 예를들어,

$$H_0: \beta_1 = 0, \quad H_a: \beta_1 \neq 0 \tag{6.4.9}$$

에 대해서 $\alpha = 5\%$ 로 검정하면

$$t=\frac{\hat{\beta}_1-0}{\sqrt{\mathrm{MSE}/\sum(x_i-\bar{x})^2}}=\frac{0.\dot{6}}{\sqrt{8/1080}}=7.746>2.571 \tag{6.4.10}$$

이므로, 귀무가설 "$\beta_1 = 0$"을 기각한다.

위의 결과에 대한 해설은 다음과 같다. 첫째, <비고 5.5.5>에서 언급했듯이, 식 (6.4.8)의 "95% 신뢰구간이 0을 포함하지 않음"과 식 (6.4.9)의 "가설을 $\alpha = 5\%$로 기각함"은 동치이다. 둘째로, 식 (6.4.1)과 (6.4.2)에 의해서, SLR에서는 식 (6.4.9)의 가설이

$$H_0 : \text{Null Model} \qquad H_a (\cup H_0) : CM$$

과 동치인데, 이는 다음과 같이 확인할 수 있다. $\alpha = 5\%$에 대해서 식 (6.4.3)은 $f = 60 > 6.61$인데, 이는 식 (6.4.10)을 제곱한 $t^2 = (7.746)^2 > (\pm 2.571)^2$과 일치한다.

통계 패키지는 ANOVA Table과 함께 (§6.4.1 참조) 다음과 같은 정보를 제공한다.

| Parameter | Estimate | $T$ for $H$:PARA=0 | $PR > \|T\|$ | STD Error of EST |
|---|---|---|---|---|
| Intercept | 60 | 3.996 | 0.0104 | 15.014 |
| $X$ | 0.6667 | 7.746 | 0.0006 | 0.0861 |

위의 표에서 "Parameter"는 $\beta_0$와 $\beta_1$을 의미하고, "Estimate"는 $\widehat{\beta_0}$과 $\widehat{\beta_1}$을 의미한다. 그리고 셋째 열에서 "7.746"은 바로 식 (6.4.10)의 $t$값이다. 이와 같이 "4.300"은 "$H_0 : \beta_0 = 0$, $H_a : \beta_0 \neq 0$"에 대한 $t$값이다 (§6.4.3 참조). 넷째 열은 "$p$-value"인데, "$PR > |T|$"에서 $T$에 절대값을 취한 것은 $T$-test가 $TTT$임을 의미하는 것이다. (비고 : 0.0006은 ANOVA Table의 $PR > F$값과 같음.) 마지막으로, 다섯째 열은 예를 들어 식 (6.4.10)의 분모인 $\sqrt{8/1080} \simeq 0.0861$인데, 이는 $T$-test에서 소음(noise)에 해당된다.

### 6.4.3 $\beta_0$에 대한 추론

식 (6.3.16)과 (6.4.5)에 의해서, 추정량 $\widehat{\beta_0}$의 분포는

$$\widehat{\beta_0} \sim N(\beta_0, \sigma^2(\frac{1}{n} + \frac{\overline{x^2}}{\sum(x_i - \overline{x})^2})) \tag{6.4.11}$$

인데, $\sigma^2$을 식 (6.3.15)의 MSE로 대체하면 (<비고 6.3.6> 참조)

$$T_{n-2} = \frac{\widehat{\beta_0} - \beta_0}{\sqrt{\mathrm{MSE}\left(\frac{1}{n} + \frac{\overline{x^2}}{\sum(x_i - \overline{x})^2}\right)}}$$

를 얻는다 (식(6.4.7) 참조).

위의 식에서 분모를 계산하면

$$\sqrt{8\left(\frac{1}{7} - \frac{(174)^2}{1080}\right)} = 15.014$$

를 얻는다. 이는 §6.4.2의 표에서 다섯째 열에 있는 값으로서 다음과 같이 구간추정과 검정에 쓰인다. 예를 들어, $\beta_0$에 대한 95% 신뢰구간은

$$\widehat{\beta_0} \pm (2.571)(15.014) = 60 \pm 38.6 \tag{6.4.12}$$

이고, "$H_0$:$\beta_0 = 0$, $H_a$:$\beta_0 \neq 0$"에 대한 검정은

$$t=\frac{\widehat{\beta_0}-0}{15.014}=\frac{60}{15.014}=3.996>2.571$$

이므로 $\alpha=5\%$에서 귀무가설 "$\beta_0=0$"를 기각한다.

$t$값인 3.996은 §6.4.2의 표에서 넷째 열에 있는 값이고, 이에 대한 $p$-value는 0.0104이다.

### 6.4.4 $Cov(\widehat{\beta_0},\widehat{\beta_1})$

식 (6.3.16)과 (6.4.5)에 의해서, 추정량 $\widehat{\beta_0}$와 $\widehat{\beta_1}$간의 공분산은

$$Cov(\widehat{\beta_0},\widehat{\beta_1})=-\sigma^2\overline{x}/\sum(x_i-\overline{x})^2 \qquad (6.4.13)$$

이므로, $\widehat{\beta_0}$와 $\widehat{\beta_1}$간의 상관계수는 (식 (2.13.6) 참조)

$$\rho=\frac{Cov(\widehat{\beta_0},\widehat{\beta_1})}{\sqrt{V(\widehat{\beta_0})\,V(\widehat{\beta_1})}}=\frac{-\overline{x}}{\sqrt{\overline{x^2}+\sum(x_i-\overline{x})^2/n}}\left(=\frac{-\overline{x}}{\sqrt{\sum x_i^2/n}}\right)=-0.9975$$

이다. (비고 : $V(\widehat{\beta_0})$와 $V(\widehat{\beta_1})$은 식 (6.4.11)과 (6.4.6) 참조.) 따라서, 예제에서는 $\widehat{\beta_0}$와 $\widehat{\beta_1}$간에 음의 상관관계가 있으며, 극단적인 경우인 $\rho=-1$에 가깝다. 이는, $(\overline{x},\overline{y})$가 회귀직선 상에 있으므로 (<비고 6.1.2> 참조), 기울기가 증가하면 $y$-절편은 감소하기 때문이다 (단, $\overline{x}>0$일 때). 만약, $\overline{x}<0$이면 기울기가 증가함에 따라 $y$-절편도 같이 증가하므로 양의 상관관계로 바뀐다. 그리고, $\overline{x}=0$이면 기울기의 변화와 무

관하게 $y$-절편은 항상 $(0, \overline{y})$이므로 $\rho = 0$이 된다.

간혹, $\beta_0$가 특별한 의미를 가지는 경우가 있는데, 이 경우에는 $\beta_0$에 대한 추정과 검정이 의미가 있다. 그러나, 일반적으로는 $\widehat{\beta_0}$가 $\widehat{\beta_1}$에 의해서 결정되다시피 하므로 $\beta_0$에 대한 추론은 상대적으로 중요시하지 않는다.

### 6.4.5 $\sigma^2$에 대한 추론

$\sigma^2$에 대한 점추정량으로는 MVUE인 MSE를 사용한다 (<비고 6.3.6> 참조). 반면에, $\sigma^2$에 대한 구간추정이나 검정은 별로 중요하게 여겨지지 않기 때문에 간단히 언급하면 신뢰구간은 §3.7.1에서와 같고 검정은 §4.4.3에서와 같은데, 다만 식 (2.15.11)의 $\sum_{i=1}^{n}(Y_i - \overline{Y})^2$대신에 SSE를 사용하기만 하면 된다.

### 6.4.6 CLM

지금까지 식 (6.3.16)에 관련된 추론을 했는데, 이제 식 (6.3.18)에 관련된 추론 (중에서 가장 대표적인 것)을 한다.

SLR 경우 식 (6.3.17)은 "$a_0\beta_0 + a_1\beta_1$"이다. 이에 "$a_0 = 1$, $a_1 = x$"를 대입하면

$$\mu_x \equiv \beta_0 + \beta_1 x \tag{6.4.14}$$

가 되는데, 이를 $\mu_x$라 부르는 이유는 식 (6.1.1)에 의해서

$$Y_x \sim N(\mu_x, \sigma^2) \tag{6.4.15}$$

이기 때문이다.

먼저, 식 (6.1.1)과 식 (6.4.1)의 차이점을 지적한다. 식 (6.4.1)은 독립변수의 관찰치인 $x_1,\cdots,x_n$과 이에 대응하는 종속변수 $Y_1,\cdots,Y_n$만을 염두에 둔 것이다. 반면에, 식 (6.1.1)에서는 독립변수 $x$가 (실제로 관찰된 $x_1,\cdots,x_n$뿐만 아니라) 모든 실수값을 가질 수 있는 변수(variable)로 취급되고 있다.

<비고 6.4.1> 식 (6.1.1)의 $Y_x$는 (일종의) 조건부 확률변수이다. 즉, $Y_x$는 독립변수의 값이 $x$일 때의 종속변수를 의미한다.

식 (6.4.14)의 $\mu_x$가 $x$의 함수이므로, $\mu_x$에 대한 점 추정량인

$$\widehat{\mu}_x \equiv \widehat{\beta}_0 + \widehat{\beta}_1 x$$

도 $x$의 함수이고 또한 $\mu_x$에 대한 신뢰구간도 $x$의 함수가 되는데, $\mu_x$에 대한 신뢰구간을 흔히 CLM(confidence limit for the mean)이라 부른다.

$\widehat{\mu}_x$의 분포는 식 (6.3.18)에 $a' = (a_0\ a_1) = (1\ x)$와 식 (6.4.5)를 대입하여 다음과 같이 얻는다.

$$\widehat{\mu}_x \sim N\left(\mu_x, \sigma^2\left\{\frac{1}{n} + \frac{(x-\overline{x})^2}{\sum_i (x_i-\overline{x})^2}\right\}\right) \tag{6.4.16}$$

그리고, 종전과 같이 $\sigma^2$을 MSE로 대체하면 $PQ$로 사용할

$$T_{n-2}=\frac{\widehat{\mu}_x-\mu_x}{\sqrt{\mathrm{MSE}\left[\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_i(x_i-\bar{x})^2}\right]}} \tag{6.4.17}$$

를 얻는다.

사용중인 예제에서 신뢰수준이 95%인 CLM은 다음과 같다.

$$\begin{aligned}\widehat{\mu}_x \pm 2.571&\sqrt{\mathrm{MSE}\left(\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum(x_i-\bar{x})^2}\right)}\\ =(60+0.6x)&\pm 2.571\sqrt{8\left(\frac{1}{7}+\frac{(x-174)^2}{1080}\right)}\end{aligned} \tag{6.4.18}$$

<비고 6.4.2> $\mu_x \equiv \beta_0+\beta_1 x$에 대한 신뢰구간인 CLM은 $x$의 함수인데, 신뢰구간의 폭은 $x=\bar{x}$일 때 최소이고 $x$가 $\bar{x}$에서 멀어질수록 증가한다.

예를 들어,

$$\begin{aligned}&(95\%\ \text{CLM when } x=168)= 172\pm 3.052\\ &(95\%\ \text{CLM when } x=174)= 176\pm 2.749\\ &(95\%\ \text{CLM when } x=180)= 180\pm 3.052\\ &(95\%\ \text{CLM when } x=222)= 208\pm 10.971\end{aligned} \tag{6.4.19}$$

인데, 유의할 점은 다음과 같다. 첫째, $x=168$은 관찰치인 $x_3=168$과 일치하지만 나머지 $x$값들은 관찰치 $x_1,\cdots,x_7$과 다르다. 둘째로, 관찰치들의 범위인 "156 ~ 192"를 벗어나는 $x=222$ 경우에는 CLM의 폭이 (10.971로) 상당히 크다.

<비고 6.4.3> 식 (6.4.18)에 "$x=0$"을 대입하면 $\beta_0$에 대한 95% 신뢰구간인 식 (6.4.12)가 된다.

### 6.4.7 예측구간 (CLI)

갓 결혼한 남자의 키가 $x$라 하자. 만약, 아들이 태어난다면 아들이 성장했을 때 키는 얼마나 되겠는가? 아들의 키를 $Y$라 하면 <비고 6.4.1>에 의해서

$$Y=\beta_0+\beta_1 x+\varepsilon,\ \ \varepsilon\sim N(0,\sigma^2) \tag{6.4.19}$$

이다 (식 (6.1.1.) 참조). 즉, SLR에 근거한 아들의 키는 확률변수로서 분포는

$$Y\sim N(\beta_0+\beta_1 x,\sigma^2)$$

이다 (식 (6.4.15) 참조).

예측(prediction)도 일종의 추정이지만 차이점은 다음과 같다. 지금까지 점추정과 구간추정의 대상은 $\mu$, $\sigma^2$, $\beta_j$ 등의 모수였는데, 이들은 모르는(unknown)값들일 뿐 확률변수는 아니었다. (다만, 추정량이 확률변수인데, 이는 추정량이 확률변수 $Y_1,\cdots,Y_n$의 함수이기 때문이다.) 그러나, 예측에서는 예측의 대상인 $Y$ 자체가 확률변수이다.

식 (6.4.19)에서 $\beta_0$와 $\beta_1$을 각각 식 (6.4.11)의 $\widehat{\beta_0}$과 식 (6.4.6)의 $\widehat{\beta_1}$으로 대체한 것을 $\widehat{Y}$라 하자. 즉,

$$\widehat{Y}=\widehat{\beta_0}+\widehat{\beta_1}x+\varepsilon \tag{6.4.20}$$

인데, 이는 $Y$에 대한 점추정량인 셈이다. 이때 유의할 점은 식 (6.4.20)에서 $\varepsilon$이 $(\widehat{\beta_0}+\widehat{\beta_1}x)$와 독립이라는 점이다. $\beta_0$와 $\beta_1$에 대한 추정량인 $\widehat{\beta_0}$와 $\widehat{\beta_1}$는 확률변수 $Y_1,\cdots,Y_n$의 함수인데, $Y_1,\cdots,Y_n$이 서로 독립인 이유는 $\varepsilon_1,\cdots\varepsilon_n$이 *iid* 확률변수이기 때문이다 (<비고 6.3.1> 참조). 그런데, SLR 모형인 식 (6.1.1)에 의하면 $\varepsilon_x$는 표본과 관련된 $\varepsilon_1,\cdots\varepsilon_n$뿐만 아니라 표본과 무관한 $\varepsilon_x$들에 대해서도 *iid* 확률변수이다.

<비고 6.4.4> $\varepsilon_x$는 같은 $x$값에 대해서도 *iid*이다. 예를 들어, 같은 아버지의 여러 아들들의 키는 *iid* $N(\beta_0+\beta_1 x,\sigma^2)$이다.

식 (6.4.20)에서, $(\widehat{\beta_0}+\widehat{\beta_1}x)$과 $\varepsilon$은 서로 독립이고 또한 정규분포를 따르므로 (비고 : 식 (6.4.16)이 $(\widehat{\beta_0}+\widehat{\beta_1}x)$의 분포임), $\widehat{Y}$의 분포는 <비고 2.15.1>에 의해서

$$\widehat{Y}\sim N\left(\beta_0+\beta_1 x,\sigma^2\left\{1+\frac{1}{n}+\frac{(x-\widehat{x})^2}{\sum_i(x_i-\overline{x})^2}\right\}\right) \tag{6.4.21}$$

가 된다. 즉 $(\widehat{\beta_0}+\widehat{\beta_1}x)$의 분포인 식 (6.4.16)에서 분산만 $\sigma^2$만큼 증가시킨 것이 $\widehat{Y}$의 분포이다. (비고 : $E(\varepsilon)=0$이므로 평균은 증가하지 않음.)

이후 과정은 종전과 같다. 즉, 식 (6.4.21)의 $\sigma^2$을 MSE로 대체하면 $PQ$로 사용할

$$T_{n-2}=\frac{\hat{Y}-(\beta_0+\beta_1 x)}{\sqrt{\mathrm{MSE}\left[1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_i (x_i-\bar{x})^2}\right]}}$$

를 얻는다. 따라서, 사용중인 예제에서 $Y$에 대한 95% 신뢰구간은 다음과 같다.

$$\hat{Y}\pm 2.571\sqrt{\mathrm{MSE}\left[1+\frac{1}{n}+\frac{(x-\bar{x})^2}{\sum_i (x_i-\bar{x})^2}\right]} \qquad (6.4.22)$$

이제, $Y$에 대한 점추정량인 $\hat{Y}$를 점추정치 $\hat{y}$로 대체할 때가 되었다. 식 (6.4.20)에서 $Y_1,\cdots,Y_n$을 관찰치 $y_1,\cdots,y_n$으로 대체하면 $\widehat{\beta_0}$와 $\widehat{\beta_1}$은 추정량에서 추정치로 바뀐다. 그리고, 관행상 $E(\varepsilon)=0$를 $\varepsilon$에 대한 추정치로 사용한다. 따라서, $Y$에 대한 점추정치는

$$\hat{y}=\widehat{\beta_0}+\widehat{\beta_1}x \qquad (6.4.23)$$

가 되어 결국 식 (6.4.18)의 $\widehat{\mu_x}$와 일치한다.

<비고 6.4.5> $\hat{y}$이 바로 §6.1에서 예측치라 불렀던 것이다.

<비고 6.4.6> 식 (6.4.22)의 $\hat{Y}$을 식 (6.4.23)의 $\hat{y}$으로 대체한 것을 예측구간(prediction interval)이라 부른다. 또한, 식 (6.4.18)을 CLM이라 부르듯이, 예측구간을 CLI(confidence limit for an individual $Y$)라 부르기도 한다.

<비고 6.4.7> 흔히 CLM과 CLI를 혼동하는데, 가장 큰 이유는 점추정치가 같기 때문이다.

예를 들어,

$$(95\%\ \text{CLI when } x=168) = 172 \pm 7.887$$

$$(95\%\ \text{CLI when } x=174) = 176 \pm 7.774$$

$$(95\%\ \text{CLI when } x=180) = 180 \pm 7.887$$

$$(95\%\ \text{CLI when } x=222) = 208 \pm 13.162$$

인데, 유의할 점은 다음과 같다. 첫째, 식 (6.4.19)와 비교하면, 점추정치는 같지만 (<비고 6.4.7> 참조) 구간의 폭은 상당히 증가했다. 둘째로, $x=168$은 관찰치인 $x_3=168$과 일치하지만, 키가 $x_3$인 아버지의 아들의 키 $y_3$는 이미 관찰된 것인 반면에 키가 $x=168$인 아버지의 아들의 키 $Y$는 아직 관찰되지 않은 것이다.

## §6.5 MLR의 분석

### 6.5.1 예제

§6.4에서는 SLR에 관해서 자세히 다루면서 결과들을 직접 계산하거나 유도하기도 하였다. 그러나, MLR은 SLR보다 복잡해서 간단한 예제에서나 직접 계산을 할 뿐, 일반적으로는 통계 패키지에 의존하게 된다. 사실, SLR을 자세히 다룬 이유도 패키지에 의한 MLR의 분석결과를 잘 이해하기 위한 것이다.

§6.2.4의 예제에 대한 패키지의 결과(output)는 다음과 같다.

| Source of Variation | SS | 자유도 | MS | $F$ | $PR>F$ | R-SQUARE |
|---|---|---|---|---|---|---|
| Model | 5.5429 | 2 | 2.7714 | 12.13 | 0.0762 | 0.9238 |
| Error | 0.4571 | 2 | 0.2286 | | | |
| Total | 6.0000 | 4 | | | | |

| Parameter | Estimate | $T$ for $H$:PARA=0 | $PR>\|T\|$ | STD Error of EST |
|---|---|---|---|---|
| Intercept | 0.5714 | 1.71 | 0.2285 | 0.3332 |
| $X_1$ | 0.7000 | 4.63 | 0.0436 | 0.1512 |
| $X_2$ | 0.2143 | 1.68 | 0.2355 | 0.1277 |

첫째, §6.2.4에서 계산했던 추정치 $\hat{\beta}=(4/7,\ 7/10,\ 3/14)$과 $SSE=0.4571$을 확인할 수 있다. 둘째로,

$$R^2=\frac{\mathrm{SSM}}{\mathrm{TSS}}\left(=\frac{\mathrm{SSE}_0-\mathrm{SSE}_{\mathrm{CM}}}{\mathrm{SSE}_0}\right)=\frac{5.5429}{6.0000}=0.9238 \qquad (6.5.1)$$

이다 (식 (6.4.4) 참조). 즉, 회귀모형이 전체 variation의 92.38%를 설명한다. 이때, "TSS=6"은 식 (6.3.14)에 의한 것이고 자유도는 "$n-1=4$"이다. 또한, 식 (6.3.15)에서 $SSE_{CM}=0.4571$이고 자유도는 $n-k-1=5-2-1=2$ 이다. 따라서 $SSM=TSS-SSE_{CM}=5.5429$이고 자유도는 $k-0=2$이다 (<비고 6.3.5> 참조).

셋째로, <비고 6.3.4>에 의한 "F-test for Model"의 결과는

$$F_{2,2}=\frac{SSM/2}{TSS/4}=\frac{2.7714}{0.2286}=12.13$$

인데 (식 (6.3.11)참조), $p$-value가 0.0762이므로 귀무가설 "$\beta_1=\beta_2=0$"을 $\alpha=5\%$로는 채택하고 $\alpha=10\%$로는 기각한다 (§4.6.3 참조).

넷째로, "$H_0:\beta_j=0$, $H_a:\beta_j\neq 0$"에 대한 $T$-test결과는 아래쪽 표에 있다. 예를 들어,

$$t=\frac{\widehat{\beta_1}-0}{(SDT\ Error\ of\ EST)}=\frac{0.7}{0.1512}=4.63 \qquad (6.5.2)$$

인데 (식 (6.4.10) 참조), $p$-value가 0.0436이므로 $\alpha=5\%$로 귀무가설 "$\beta_1=0$"를 기각한다. 반면에, 귀무가설 "$\beta_0=0$"와 "$\beta_2=0$"에 대한 $p$-value는 20%이상이므로 $\alpha=10\%$에서조차 귀무가설을 채택한다. 참고로 식 (6.5.2)에서 0.1512는 다음과 같이 얻을 수 있다. 추정량 $\widehat{\beta_1}$은 평균이 $\beta_1$인 정규분포를 따르는데, 식 (6.2.22)에 의해서 분산은 $\sigma^2/10$이다. 따라서, 분산의 추정치는 MSE/10 = 0.02286 이고, $\sqrt{0.02286}=0.1512$ 이다.

위의 결과들 외에도 CLM과 예측구간 (또는 CLI) 등 다양한 결과를 통계패키지로 얻을 수 있다.

### 6.5.2 예제 뒷처리

§6.5.1의 예제에서 $\alpha = 5\%$로 "$F$-test for Model"에 대한 귀무가설 "$\beta_1 = \beta_2 = 0$"을 채택했다. 그렇다면 과연 "$\beta_1 = 0$"이고 "$\beta_2 = 0$"인가? 반면에, $\beta_1$에 대한 $T$-test에서는 $\alpha = 5\%$로 귀무가설 "$\beta_1 = 0$"를 기각했는데, 이는 과연 모순된 결과인가?

사실, MLR에서 가장 중요하고 또한 가장 어려운 주제는 소위 "Model Selection"이라는 것이다. Model Selection 이란 종속변수를 조금씩이나마 설명할 수 있는 모든 가능한 독립변수들 중에서 어떤 것은 취하고 어떤 것은 버릴 것인가를 결정하는 것이다. 즉, 모든 가능한 독립변수를 모두 포함시킨 모형을 CM이라 (하고 독립변수의 개수를 $k$라) 하면, $\{\beta_1, \cdots, \beta_k\}$ 중에서 하나 이상을 0으로 놓은 것이 RM인데 (§6.3.3 참조), RM들 중에서 가장 좋을 것을 선택하는 것이 Model Selection 이다.

<비고 6.5.1> RM의 총수는 $(2^k - 1)$인데, 이는 $k$개의 독립변수 각각이 모형에 포함될 수도 있고 안될 수도 있기 때문이다. (단, CM을 제외하기 위해서 1을 뺌.)

예제에서는 $k = 2$이므로 CM과 RM들은

$$
\begin{aligned}
&\text{CM} \ : Y_i=\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\varepsilon_i \\
&\text{RM}_1 : Y_i=\beta_0+\beta_1 x_{i1}+\varepsilon_i \\
&\text{RM}_2 : Y_i=\beta_0+\beta_2 x_{i2}+\varepsilon_i \\
&\text{RM}_3 : Y_i=\beta_0+\varepsilon_i
\end{aligned}
\tag{6.5.3}
$$

인데, $\varepsilon_i$의 분포는 모두 $iid\ N(0,\sigma^2)$이고 $\text{RM}_3$는 Null Model이라 부르던 것이다 (<비고 6.3.4> 참조).

유의수준 $\alpha$를 5%로 책정하자. "$F$-test for Model"은

$$H_0: \text{RM}_3,\ \text{H}_a: \text{CM}$$

이므로 (식 (6.3.12) 참조), $H_0$를 채택한다는 것은 CM과 비교했을 때 $\text{RM}_3$가 낫다는 뜻이다. 그렇지만, 아직 $\text{RM}_3$를 $\text{RM}_1$ 또는 $\text{RM}_2$와 비교하지는 않았다. 반면에, $\beta_1$에 대한 $T$-test는

$$H_0: \text{RM}_2,\ \text{H}_a: \text{CM}$$

이므로, $H_0$를 기각한다는 것은 CM이 $\text{RM}_2$보다 낫다는 뜻이다. 또한, $\beta_2$에 대한 $T$-test는

$$H_0: \text{RM}_1,\ \text{H}_a: \text{CM}$$

이므로, $H_0$를 채택한다는 것은 $\text{RM}_1$이 CM보다 낫다는 뜻이다.

이상의 결과를 통합하면 ("$<$"는 선호도를 나타냄)

$$RM_2 < CM < RM_1, RM_3$$

이므로, 아직도 $RM_1$과 $RM_3$를 비교해 봐야한다. 가설을

$$H_0 : RM_3,\ H_a : RM_1$$

이라 하면 (식 (6.3.12) 참조), 이는 바로 독립변수 $X_1$ 하나만 사용하는 SLR에서의 "$F$-test for Model" (인 동시에 $\beta_1$에 대한 $T$-test)이다. SLR에 대한 패키지의 결과는 다음과 같다.

| Source of Variation | SS | 자유도 | MS | $F$ | $PR>F$ | R-SQUARE |
|---|---|---|---|---|---|---|
| Model | 4.9 | 1 | 4.9 | 13.36 | 0.0354 | 0.816 |
| Error | 1.1 | 3 | 0.36 | | | |
| Total | 6.0 | 4 | | | | |

| Parameter | Estimate | $T$ for $H_0$:PARA=0 | $PR>\|T\|$ | STD Error of EST |
|---|---|---|---|---|
| Intercept | 1.0 | 3.69 | 0.0345 | 0.2708 |
| $X_1$ | 0.7 | 3.66 | 0.0354 | 0.1915 |

$p$-value인 0.0354가 $\alpha = 0.05$보다 작으므로 귀무가설 "$RM_3$"를 기각하고 대립가설 "$RM_1$"을 채택한다. 따라서, $RM_1$을 가장 좋은 모형으로 선택하게 된다.

### 6.5.3 $Cov(\widehat{\beta}_1, \widehat{\beta}_2)$

사용중인 예제의 특징은 식 (6.2.22)의 $(X'X)^{-1}$에 "0"이 많이 있다는 점이다. 먼저 1행에 있는 0을 설명한다. $(X'X)^{-1}$에서 3행과 3열을 제거하면 독립변수 $X_1$하나만 사용하는 SLR에서의 $(X'X)^{-1}$이 되는데, 이를 식 (6.4.5)와 비교하면 0이 발생한 이유가 "$\overline{X_1}=0$"임을 알 수 있다. (비고 : 이에 대한 기하학적인 해석은 §6.4.4 참조.)

사실, 중요한 것은 3행에 있는 0인데, 이는 식 (6.3.16)에 의하면 "$Cov(\widehat{\beta}_1, \widehat{\beta}_2) = 0$"을 의미한다. 그런데, 역시 식 (6.3.16)에 의하며 $\widehat{\beta}_1$과 $\widehat{\beta}_2$는 정규분포를 따르므로, "$Cov(\widehat{\beta}_1, \widehat{\beta}_2) = 0$"은 "$\widehat{\beta}_1$과 $\widehat{\beta}_2$가 독립"임을 의미한다 (<비고 2.13.4> 참조).

<비고 6.5.2> 추정량 $\widehat{\beta}_1$과 $\widehat{\beta}_2$가 독립일 때 독립변수 $X_1$과 $X_2$가 독립이라고 표현하기로 한다.

추정량은 확률변수이므로 독립운운하는 것이 자연스럽다. 반면에, 독립변수는 확률변수가 아니라고 했으므로 독립운운하기가 어색한데, 이는 §6.6.6에서 논하기로 하자.

$\widehat{\beta}_1$과 $\widehat{\beta}_2$가 독립이든 $X_1$과 $X_2$가 독립이든 이에 따른 중요한 결과는 다음과 같다. $X_1$만 사용하는 SLR에서 SSM은 4.9이고, $X_2$만 사용하는 SLR에서 SSM은 9/14인데, $X_1$과 $X_2$를 모두 사용하는 MLR에서 SSM은 (4.9+ 9/14)이다. 즉, Variation인 TSS=6 중에서 $X_1$ 혼자서 설명하는 부분은 4.9이고, $X_2$ 혼자서 설명하는 부분은 9/14인데, $X_1$과 $X_2$가 함께 설명하는 부분은 (4.9+ 9/14)이다.

이러한 결과는 당연히 그리고 항상 성립해야 되는 것으로 오해하기 쉽다. 그러나, 이러한 결과는 이상적인(?) 상황에서나 성립할 뿐, 일반적으로는 성립하지 않는다. 대체로, $X_1$과 $X_2$가 함께 설명하는 양이 따로따로 설명하는 양들을 합친 것보다 작다. 예를 들어, $X_1$은 아버지 키이고 $X_2$는 할아버지의 키라 하자. 종속변수인 아들의 키를 $X_2$보다는 $X_1$이 더 잘 설명하겠지만, $X_2$만으로도 어느정도 설명이 가능하다. 식 (6.4.4)에 의하면 $X_1$만으로 TSS의 92.31%를 설명한다. 그리고, 예를 들어, $X_2$만으로는 TSS의 60%를 설명한다고 하자. 그러나, $X_1$과 $X_2$를 모두 사용하는 MLR에서 TSS의 152.31%가 설명될 수는 없는 것이다.

<비고 6.5.3> 드물기는 하지만, $X_1$과 $X_2$가 함께 설명하는 양이 따로따로 설명하는 양들을 합친 것보다 클 경우도 있음 (문헌 [8] 참조).

### 6.5.4 Model Selection

§6.5.2의 예제에서는 $Cov(\widehat{\beta}_1, \widehat{\beta}_2) = 0$이라서 (§6.5.3 참조),

$$\mathrm{SSM}(X_1, X_2) = \mathrm{SSM}(X_1) + \mathrm{SSM}(X_2) \tag{6.5.5}$$

가 성립했고 따라서 Model Selection 과정이 비교적 간단했다. 그런데, 설사 식 (6.5.5)가 성립하지 않았다고 해도 $k=2$이기 때문에 모두 $2^k=4$개의 모형만 서로 비교하면 되었다 (식 (6.5.1) 참조). 그리고, 비교하는 횟수는 최대한 $\binom{4}{2}=6$회에 지나지 않는다. 그러나, 예를 들어 $k=10$이면, $2^{10}=1024$이고 $\binom{1024}{2}=523776$가 된다. 따라서, 효과적인 Selection 절차가 필요하다.

Model은 간단할수록 좋다. 즉, 독립변수의 수가 적을수록 좋다. 그러나, 최소한 모든 독립변수가 유의(significant)해야 한다. 즉, 모든 $j$에 대해서, $\beta_j$에대한 $T$-test의 $p$-value가 책정된 유의수준보다 작아야 된다. (비고: 모든 독립변수가 유의하면 "$F$-test for Model"의 결과도 유의함.)

흔히, "max $R^2$"를 기준으로 사용하기도 하는데, 이는 잘못된 것이다. 물론, 그 이유는 어떤 독립변수를 추가하더라도 최소한 $R^2$가 감소하지는 않기 때문이다. 그러나, 동일한 개수의 독립변수를 사용하는 RM들 중에서는 $R^2$가 클수록 좋다. 예를 들어, 식 (6.5.3)에서 $\mathrm{RM}_1$의 $R^2$가 $\mathrm{RM}_2$의 $R^2$보다 크므로 $\mathrm{RM}_1$이 $\mathrm{RM}_2$보다 낫다.

다음, "min MSE"를 기준으로 사용하기도 하는데, 이는 최소한 "max $R^2$" 기준보다는 낫다. 별로 도움이 되지 않는 독립변수들을 제거하면 (SSM이 약간이나마 감소하므로) SSE가 약간 증가하기는 하지만 아울러 자유도도 증가하기 때문에 결과적으로 MSE가 감소하기도 하기 때문이다.

통계 패키지에 의한 방법은 크게 세가지가 있다. 첫째로, Forward 방법은 독립변수의 개수를 하나씩 증가시키는 방법이다. 예를 들어, 식 (6.5.3)에서는 $\mathrm{RM}_3$로 시작한다. 1-단계에서는 $\mathrm{RM}_1$과 $\mathrm{RM}_2$ 중에서 하나를 선택하는데, 예제에서는 $\mathrm{RM}_1$을 선택하게 된다. (단, $\mathrm{RM}_1$의 성능이 기준치 이상인 경우에 한함.) 2-단계에서는 독립변수를 하나 더 추가하는데, 예제에서는 CM이 유일하다. 그러나, 새로 추가된 $X_2$의 성능이 기준미달이므로 $\mathrm{RM}_1$으로 낙착이 된다.

둘째로, Backward 방법은 CM으로 시작해서 독립변수를 하나씩 제거하는 방법이다. 예제에서는 1-단계에서 기준미달인 $X_2$를 제거한다. 그리고, 2-단계에서는 하나 남은 $X_1$의 제거여부를 결정한다.

사용중인 예제에서는 식 (6.5.5)가 성립하기 때문에 Forward와 Backward 방법의 결론은 동일하다. 그러나, 일반적으로는 그렇지 않다. 예를 들어, 처음에는 $X_5$가 힘

을 발휘하다가 나중에 $X_6$가 등장하고 나서는 ($X_5$의) 힘이 약해지는 수가 있다. 또한 반대되는 현상도 가능하다 (<비고 6.5.3> 참조). 즉, 처음에는 $X_5$가 유의하지 않다가 나중에 $X_6$가 등장하면서 함께 큰 힘을 발휘하는 수도 있다. 이러한 점을 고려한 방법이 세 번째인 Stepwise 방법인데, 이는 Forward와 Backward 방법을 합친 것이라 할 수 있다. 즉, 일방동행이 아니라 양방향으로 왔다갔다 하면서 적절한 모형을 찾아내는 것이다.

### 6.5.5 선형모형의 범위

사용중인 예제는 다음과 같은 특징도 있다 (식 (6.2.21) 참조).

$$x_{i2} = x_{i1}^2 \ , \ i = 1, \cdots, n$$

즉, 두 번째 독립변수는 사실상 첫번째 독립변수를 제곱한 것이다. 그러니까, 회귀평면 (식 (6.2.24) 참조)

$$\hat{y} = 0.571 + 0.7x_1 + 0.214x_2$$

는 사실상 2-차원 상의 포물선

$$\hat{y} = 0.571 + 0.7x_1 + 0.214x_1^2$$

인 셈이다.

이와 같이, “선형”모형이라고 해서 반드시 “직선” 또는 “평면”에만 해당되는 것은 아니다. 문헌 [9]의 예제 11.10은 다음과 같다. SLR인

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i,\ i = 1, \cdots, n \tag{6.5.6}$$

에서 $Y_i \equiv \ln W_i$, $\beta_0 = \ln \alpha_0$, $x = \ln I$, $\varepsilon = \ln \varepsilon'$이면 식 (6.5.6)은 사실상

$$W_i = \alpha_0 I_i^{\beta_1} \varepsilon_i',\ i = 1, \cdots, n \tag{6.5.7}$$

이다. 다만, 식 (6.5.6)에 대한 가정이 $\varepsilon_i \sim iid\ N(0, \sigma^2)$이므로, 식 (6.5.7)에서는 $\varepsilon_1', \cdots, \varepsilon_n'$이 ( *iid* 확률변수이고) 대수 정규분포를 따르게 된다 (§2.9.3 참조).

식 (6.5.7)의 형태뿐만 아니라 어떤 형태라도 적절한 변환(transform)을 통해서 식 (6.5.6)의 형태가 (또는, 일반적으로 식 (6.3.1)의 형태가) 되기만 하면 이를 선형회귀모형으로 간주할 수 있다.

## §6.6 상관관계 분석

### 6.6.1 서론: 독립변수도 확률변수?

SLR 대신에 상관분석(correlation analysis)으로 독립변수와 종속변수의 관계를 분석하기도 한다. 상관분석은 간편해서 중학교 과정에조차 등장하지만, 그 배경에 깔린 가정은 제법 복잡한 편이다.

상관관계란 "두개의 확률변수" 간의 선형종속성을 의미하는데, 이에 대한 표준화된 척도가 식 (2.13.6)의 상관계수이다. 따라서, 상관분석에서는 "독립변수도 확률변수"로 취급해야 된다.

SLR 대신에 상관분석을 할 때에는 독립변수 $X$와 종속변수 $Y$의 결합분포가 BVN(bivariate normal : 이변량 정규)분포라고 가정한다 (식 (6.3.16) 참조). 그런데, BVN 분포는 2장에서 다루지 않았으므로, 이 기회에 정식으로 다룬다. 또한 이 기회에, 지금까지 독립변수를 확률변수가 아니라고 해왔던 점에 대해서도 명확히 한다.

### 6.6.2 BVN 분포

지금까지 등장한 다변량(multivariate)분포는 §2.1.6의 다항분포와 §2.2.3의 다변량 초기하 분포인데, 이들은 모두 이산분포이다. 반면에, 연속분포에 대해서는 §6.3.4에서 $\{\widehat{\beta}_0, \cdots, \widehat{\beta}_k\}$의 결합분포가 $MVN$(다변량 정규)분포라고 언급만 했을 뿐, $MVN$분포에 대한 설명은 없었다. 그러나 $MVN$은 복잡하므로 $BVN$만 다룬다.

$X \sim N(\mu_X, \sigma_X^2)$, $Y \sim N(\mu_Y, \sigma_Y^2)$이라 하고, 각각의 밀도함수를 $f_X(x)$, $f_Y(y)$라 하자. 그리고 $X$와 $Y$의 결합밀도함수를 $f(x,y)$라 하면, $X$와 $Y$가 독립인 경우에는 $f(x,y) = f_X(x) \cdot f_Y(y)$이다. 그러나 일반적으로는

$$f(x,y)=f_X(x)\cdot f_{Y|X}(y\mid x) \tag{6.6.1}$$

인데, $f_{Y|X}(y\mid x)$는 □$X=x$□라는 조건 하에서 $Y$의 (조건부)밀도함수이다 (<비고 2.14.1> 참조).

식 (6.6.1)에서 $f_X(x)$와 $f(x,y)$는 다음과 같다.

$$f_X(x)=\frac{1}{\sqrt{2\pi}\sigma_X}e^{-\frac{1}{2}(\frac{x-\mu_X}{\sigma_X})^2} \tag{6.6.2}$$

$$f(x,y)=\frac{1}{2\pi\sigma_X\sigma_Y}e^{-\frac{Q}{2}} \tag{6.6.3}$$

$$\text{where}\quad Q=\frac{1}{1-\rho^2}\{(\frac{x-\mu_X}{\sigma_X})^2+(\frac{y-\mu_Y}{\sigma_Y})^2-2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\}$$

식 (6.6.3)의 $Q$에 들어있는 $\rho$는 바로 $X$와 $Y$간의 상관계수인데, □$\rho=0$□일 때 □$f(x,y)=f_X(x)\cdot f_Y(y)$□가 성립함을 쉽게 알 수 있다 (<비고 2.13.4> 참조).

$BVN$분포를 따르는 모집단에서 추출한 크기가 $n$인 표본을 $\{(X_1,Y_1),\cdots,(X_n,Y_n)\}$이라 하고, 표본의 관찰치를 $\{(x_1,y_1),\cdots,(x_n,y_n)\}$이라 하자. 그러면, <비고 2.7.1>에 의해서 LF는

$$L(\rho,\mu_X,\mu_Y,\sigma_X^2,\sigma_Y^2)=\prod_{i=1}^{n}f(x_i,y_i)$$

인데, 이로부터 $\rho$에 대한 MLE로

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} \tag{6.6.4}$$

을 얻는다. (비고: $-1 \le r \le 1$ )

<비고 6.6.1> $r$은 표본상관계수라 불리는데, 식 (6.6.4)의 분모와 분자를 $n$으로 나누면 각각 식 (2.13.6)의 분모와 분자에 대응된다.

### 6.6.3 상관분석

§6.1의 예제에서 $\{(x_1, y_1), \cdots, (x_n, y_n)\}$을 관찰된 표본이라 하면,

$$r = \frac{720}{\sqrt{1080}\sqrt{520}} = 0.9608$$

을 얻는다. 이때 유의할 점은

$$r^2 = (0.9608)^2 = 0.9231 = R^2 \tag{6.6.5}$$

이다. 즉, SLR에서는 식 (6.6.4)를 제곱한 것이 식 (6.4.4)의 $R^2$와 일치한다 (계산은 생략함).

또한, $r$은 회귀직선의 기울기인 식 (6.2.6)의 $\hat{\beta}_1$과 다음과 같은 관계가 있다.

$$r = \hat{\beta}_1 \frac{\sqrt{\sum(x_i - \bar{x})^2}}{\sqrt{\sum(y_i - \bar{y})^2}} \tag{6.6.6}$$

<비고 6.6.2> $\widehat{\beta_1}$은 □$-1 \le r \le 1$□인 $r$을 $x$방향으로는 $\sqrt{\sum(x_i-\overline{x})^2/(n-1)}$만큼 늘리고, $y$방향으로는 $\sqrt{\sum(y_i-\overline{y})^2/(n-1)}$만큼 늘린 것과 같다. 즉, $r$은 일종의 표준화된 기울기인데, 이에 표본 표준편차의 비율을 곱하면 $\widehat{\beta_1}$이 된다.

마지막으로, 식 (6.6.5)와 (6.6.6)에 의해서, 식 (6.4.10)의

$$t = \frac{\widehat{\beta_1}-0}{\sqrt{MSE/\sum(x_i-\overline{x})^2}} = \frac{0.6}{\sqrt{8/1080}} = 7.746 > 2.571 \qquad (6.6.7)$$

은 다음과 일치한다.

$$t = \frac{r-0}{\sqrt{(1-r^2)/(n-2)}} = \frac{0.9608}{\sqrt{0.0769/5}} = 7.746 > 2.571 \qquad (6.6.8)$$

즉, □$H_0$: $\beta_1 = 0$, $H_a$: $\beta_1 \neq 0$□에 대한 검정과 □$H_0$: $\rho = 0$, $H_a$: $\rho \neq 0$□에 대한 검정은 일치한다. 그런데, 전자는 □$F$-test for model□과 일치하고 후자는 □$X$, $Y$의 독립성 검정□과 일치하므로(<비고 2.13.4> 참조), 결국 SLR에서는 위의 네가지가 모두 일치한다.

### 6.6.4 직교회귀

상관분석을 하기 위해서 $X$와 $Y$의 결합분포가 $BVN$분포라 가정하였다. 그리

고, 관찰된 표본을 $\{(x_1,y_1),\cdots,(x_n,y_n)\}$이라 하였다. 그러면, 관찰된 표본에 가장 잘 들어맞는 직선은 무엇인가?

SLR에서 회귀직선은 <비고 6.1.1>에 따른 것인데, 그때 □수직방향 거리□의 제곱합만 따진 이유는 $Y$만 확률변수라고 가정했었기 때문이다. 그러나, 이제는 $X$와 $Y$가 대등한 확률변수로 취급되고 있으므로, □최단거리□의 제곱합을 따진다 (<비고 6.2.1> 참조).

최단거리 제곱합이 최소가 되게하는 직교(orthogonal)회귀직선을 구하는 방법은 선형대수학과 관련이 있는데, 그 결과를 요약하면 다음과 같다 (문헌 [3] 참조). 첫째, 직교 회귀직선도 $(\overline{x},\overline{y})$를 통과한다 (<비고 6.1.2> 참조). 따라서, 기울기만 구하면 되는데, 편의상 좌표축을 옮겨서 $(\overline{x},\overline{y})=(0,0)$가 되게 하자. 그리고,

$$W \equiv \begin{vmatrix} x_1 & y_1 \\ x_2 & y_2 \\ \vdots & \vdots \\ x_n & y_n \end{vmatrix}$$

이라 하자. 둘째로, $W'W$의 고유치(eigen value)를 $\lambda_1,\lambda_2$라 하고 (단, $\lambda_1>\lambda_2$), 대응하는 고유 벡터(eigen vector)를 각각 $V_1$, $V_2$라 하자. 그러면, $V_1$의 연장선이 직교 회귀직선이 되고, $\lambda_2$는 최단거리 제곱합이 된다.

예를 들어,

$$W' = \begin{pmatrix} -5.2 & -3 & 1.2 & 3 & -1.4 & 1 & 4.4 \\ -3.6 & -4 & -3.4 & -1 & 4.8 & 3 & 4.2 \end{pmatrix}$$

이면, $W'W = \begin{pmatrix} 68.8 & 38.4 \\ 38.4 & 91.2 \end{pmatrix}$ 로부터

$$\begin{pmatrix}\lambda_1\\ \lambda_2\end{pmatrix}=\begin{pmatrix}120\\ 40\end{pmatrix},\quad V_1=\begin{pmatrix}0.6\\ 0.8\end{pmatrix},\quad V_2=\begin{pmatrix}-0.8\\ 0.6\end{pmatrix}$$

을 얻으므로, 직교 회귀직선은 $y=1.3x$이고 SSE=40이다. 참고로, $\lambda_1=120$은 7개의 점으로부터 ($V_2$의 연장선인) $y=-0.75x$까지의 최단거리 제곱합이다. 또한, 직교 회귀직선에 대해서도 역학적인 해석이 가능하나 이를 생략한다 (§6.2.2 참조).

### 6.6.5 회귀모형의 재해석

§6.5 이전에는 독립변수를 확률변수가 아니라고 했는데, 이제와서 $X$를 확률변수라 하고 $(X,Y)$가 $BVN$분포를 따른다고 하면 이는 과연 모순인가?

관찰된 표본 $\{(x_1,y_1),\cdots,(x_n,y_n)\}$을 얻는다는 것은 확률변수 쌍(pair)인 $(X,Y)$를 독립적으로 $n$번 구현(realize)시킨다는 뜻이다. 즉, $X$와 $Y$를 동시에 구현시키는 것인데, 이러한 관점에서 상관분석과 직교회귀분석을 한 셈이다.

반면에, §6.5 이전에서의 관점은 다음과 같다. 한마디로, $X$를 먼저 구현시켜서 $\{x_1,\cdots,x_n\}$을 얻은 다음, $i=1,\cdots,n$에 대해서 □$X=x_i$□라는 조건 하에서 $Y$를 구현시키는 것이다. 이에 <비고 6.4.1>에서 암시했듯이, §6.5 이전에 등장한 종속변수는 □조건부 확률변수□이다. 그리고, 이 조건부 확률변수의 밀도함수는 식 (6.6.1)의 $f_{Y|X}(y|x)$인데, 이 역시 정규분포를 따름을 보일 수 있다.

구체적으로, 식 (6.4.1)에서 □$Y_i$□는 사실상 □$Y|X=x_i$□이고, $(\beta_0+\beta_1 x_i)$는 $E(Y|X=x_i)$이며, $\varepsilon_i$의 분산인 $\sigma^2$은 $V(Y|X=x_i)$이다. 그리고, 이러한 관계는 식 (6.3.1)의 MLR에 대해서도 성립한다. 단, MLR 에서는 $(X_1,\cdots,X_k,Y)$가 $MVN$분포를 따르고, $i=1,\cdots,n$에 대해서 □$X_1=x_{i1},\cdots,X_k=x_{ik}$□라는 조건 하에서 $Y$를 구현시킬 따름이다.

### 6.6.6 독립변수가 독립?

$(X_1, \cdots, X_k, Y)$가 $MVN$분포를 따른다고 하자. 그러면, 이상적인 경우는 $X_1, \cdots, X_k$가 서로 독립인 경우인데, 이때 식 (6.5.5)가 다음과 같이 확장된다.

$$SSM(X_1, \cdots, X_k) = \sum_{j=1}^{k} SSM(X_j)$$

$X_1, \cdots, X_k$는 서로 독립이더라도 각각은 $Y$와 (선형) 종속이다. (비고: $MVN$ 가정하에서는 비선형 종속은 존재하지 않으므로 따질 필요가 없음. <비고 2.13.4> 참조.) 그러나, $Y$와 종속인 독립변수를 모두 모형에 포함시키는 것은 아니다. 모형은 간단할수록 좋다고 했는데, 이는 유의(significant)한 독립변수만 모형에 포함시키는 것을 의미한다 (§6.5.4 참조). (만약에, $Y$와 독립인 $X_j$를 실수로 CM에 포함시켰더라도 $SSM(X_j) = 0$이므로 $X_j$는 Model Selection 과정에서 제거된다.)

그런데, 실제 문제에서는 $X_1, \cdots, X_k$가 서로 독립이 아니다. 예를 들어, $Y$가 아들의 키일 때, $X_1, X_2, X_3$는 각각 아버지, 할아버지, 어머니의 키일 수 있다. 이 경우 $Y$가 $X_1$에 종속이면 $X_1$은 다시 $X_2$에 종속이다. 그리고, 예를 들어 키가 큰 (작은) 사람끼리 결혼하는 경향이 있다면 $X_1$과 $X_3$도 종속이다.

독립변수끼리 독립이 아니더라도 Model Selection 과정에서 어느정도까지는 불필요한 독립변수를 제거할 수 있다. 예를 들어, $X_1$이 $Y$에 대해서 설명하지 못한 부분 중에서 $X_2$가 추가로 설명하는 부분의 크기를 따져서 $X_2$의 유의성을 판단할 수 있다. 이때 한가지 주의할 점은 다음과 같다. 예를 들어, 아버지의 키인 $X_1$과 어머니의 키인 $X_3$간의 상관계수가 1에 가깝다고 하자. 이 경우, $X_1$이 (또는, $X_3$가) 먼저

$Y$를 설명하고 나서 $X_3$가 (또는, $X_1$이) 추가로 설명하는 부분의 크기는 미미하므로 $X_3$는 (또는, $X_1$은) 당연히 모형에서 제거되어야 한다. 그러나, 이때 기술적인 (technical) 문제가 발생할 수 있다. 첫째, 극단적으로 $X_1$과 $X_3$간의 상관계수가 $\pm 1$이면 식 (6.2.15)의 $X$의 rank가 하나 부족해서 $(X'X)^{-1}$가 존재하지 않는다. 둘째로, $X_1$과 $X_3$간의 상관계수가 $\pm 1$은 아니더라도 $\pm 1$에 가까우면 MSE가 상당히 커진다. 그런데 MSE는 회귀분석에 등장하는 모든 검정통계량에서 소음(noise)에 해당되는 분모에 포함되어 있으므로 MSE가 커지면 모든 검정의 검정력이 약해진다. 이러한 현상을 다중공선성(multicolinearity)이라 하는데, 이를 방지하기 위해서는 $X_1$과 $X_3$중에서 하나를 처음부터 CM에서 제거해야 된다. (예를 들어, $X_1$과 $X_3$간의 표본 상관계수가 $\pm 1$에 가까우면 하나를 제거하는데, 물론 $Y$를 조금이나마 잘 설명하는 것을 남긴다.)

참고로, §6.5.2의 예제에서는 $X_2 = X_1^{\,2}$이므로 $X_1$과 $X_2$가 (최소한 비선형적으로는) 종속인데, $X_1$의 관찰치를 원점에 대해서 대칭이 되도록 $(-2,-1,0,1,2)$라 했기 때문에 결과적으로 선형적으로는 $X_2$와 독립이 되었다. 만약, $(X_1, X_2, Y)$가 $MVN$분포를 따른다는 가정이 합당하다면 선형독립은 독립을 의미한다 (<비고 6.5.2> 참조). 그러나 $(-2,-1,0,1,2)$는 정규분포를 따르는 $X_1$의 관찰치라 하기 어렵고, 오히려 인위적으로 설정한 값이 분명하다. 따라서, 예제에서는 $X_1$과 $X_2$가 확률변수라하기 어렵다.

회귀모형에서 독립변수는 확률변수일 수도 있고 아닐 수도 있는데, 특히 인위적으로 제어(control)할 수 있는 독립변수는 확률변수라 하기 어렵다.

사실 독립변수가 확률변수인지 아닌지에 대한 논의는 5장의 ANOVA로 거슬러 올라간다. §5.5.2 에서 $Y_{ij}$를 젖소 $j$에게 사료 $i$를 먹일 때의 우유생산량이라 했다. 그런데, 많은 젖소 중에서 $J$마리를 (무작위로) 뽑았다면, 식 (5.5.4)에서 $\beta_j$는 확률변

수가 되어야 마땅하다. 또한, 여러 종류의 사료 중에서 $I$종류를 뽑았다면 식 (5.5.4)의 $\tau_i$까지도 확률변수가 되는데, $\beta_j$와 $\tau_i$가 확률변수가 되면 분석방법이 복잡해지므로 편의상 이들을 상수로 취급한 것이다. (비고: 자세한 내용은 실험계획법 교재 참조.)

# 참고 문헌

비고: [9]는 다년간 교재로 사용되었던 책으로서, 이 책에 대한 전반적인 참고문헌의 역할을 함. 그리고, [10]~[15]는 이 책에서 직접 인용되지는 않았지만 간접적으로 참고가 된 문헌임.


[1] 김태성, 채경철 (1995), □일곱장 포커 약의 확률□, 응용통계연구, 8권 2호, pp.163-177.

[2] 정한영, 이기원 (1994), □실례를 이용한 통계학 교육방법에 대한 제언□, 한국통계학회논문집, 1권 1호, pp.184-191.

[3] 채경철 (1990), □직교회귀의 역학적 고찰□, 응용통계연구, 3권 1호, pp.47-57.

[4] 한국통계학회 (1987), 통계용어사전, 자유아카데미.

[5] Chae, K.C. (1990), □A Geometric Interpretation of the PERT Assumptions on the Activity Time□, Int. J. of Math. Educ. in Sci. and Tech., Vol. 21, pp. 283-288.

[6] Chae, K.C. (1993), □Presenting the Negative Hypergeometric Distribution to the Introductory Statistics Courses□, Int. J. of Math. Educ. in Sci. and Tech., Vol.24, pp. 523-526.

[7] Johnson, N.L. & Kotz, S. (1969), Discrete Distributions, Wiley.

[8] Schey, H.M. (1993), □The Relationship between the Magnitudes of $SSR(x_2)$ and $SSR(x_2 \mid x_1)$: A Geometric Description□, The American Statistician, Vol. 47, pp. 26-30.

[9] Wackerly, D.D., Mendenhall, W. & Scheaffer, R.L. (1996), Mathematical Statistics with Applications (Fifth Edition), Duxbury.

[10] Hoel, P.G., Port, S.C. & Stone, C.J. (1971), Introduction to Statistical Theory, Houghton Mifflin.

[11] Hogg, R.V. & Craig, A.T. (1978), Introduction to Mathematical Statistics (Fourth Edition), Macmillan.

[12] Lehmann, E.L. (1983), Theory of Point Estimation, Wiley.

[13] Montgomery D.C. (1976), Design and Analysis of Experiments, Wiley.

[14] Neter, J. & Wasserman, W. (1974), Applied Linear Statistical Models, Irwin.

[15] Schmidt, P. (1976), Econometrics, Dekker.