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 지지율이

다. 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 N을 ∞로 간주한다. (반면에, 표본의 크기 n은 상대적으로 N보다 훨씬 작아야 된다.) 예를 들어, <사례 1.1>에서는 이 가정이 별로 무리가 없다. 즉, N이 약 이천만이면 $N \to \infty$라고 간주할만하다. (그러나, N이 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 모집단의 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.

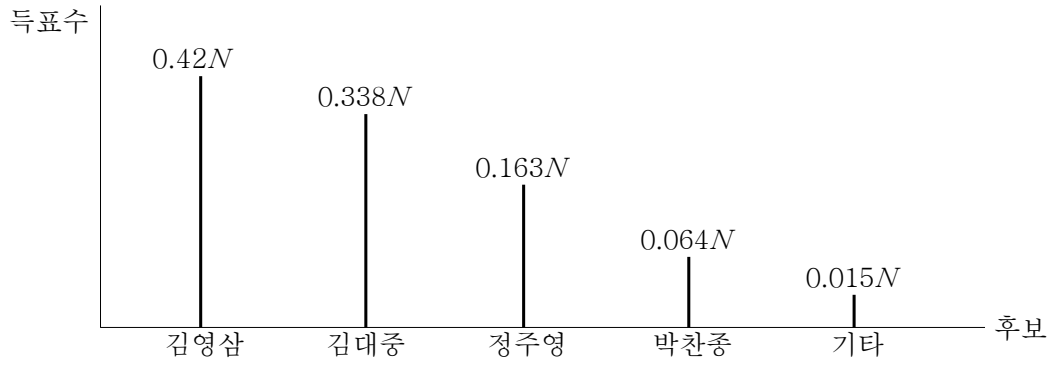
§1.3 모집단의 표현

<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 N표의 42%인 $0.42N$표를 얻었다.

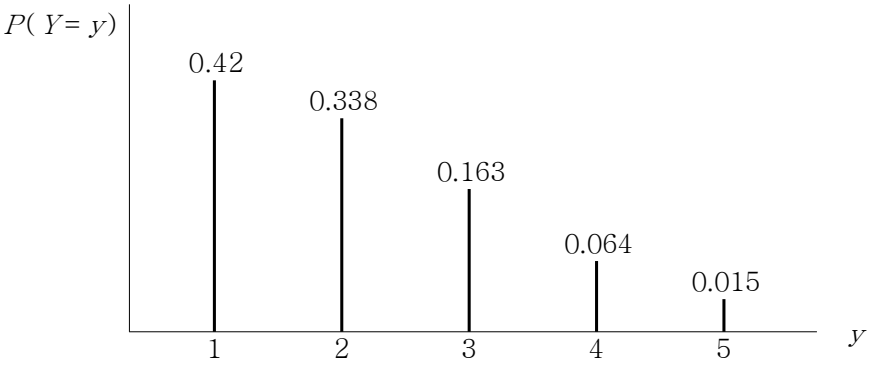


<그림 1.1> <사례 1.1>의 모집단의 분포

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 모

집단의 부분집합이다.)

확률변수 Y를 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, Y의 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 Y의 분포를 모집단의 "확률분포"라 부른 이유는 다음과 같다. 모집단의 N개의 요소 중에서 하나를 임의로 뽑는다고 하자. 임의로 뽑으므로, 뽑힐 확률은 N개 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) Y의 분포를 지칭하는 것이다.

앞으로는 Y의 분포를 모분포라 할 뿐더러, Y를 모집단의 임의요소라 한다. 그리고, $P(Y=y)$는 임의요소(에 대응된 실수)가 y일 확률을 의미한다. (비고: "$Y=y$"를 Y가

y로 구현(realize)되었다고 표현함.) 이때 유의할 점은 다음과 같다. 대문자 Y는 확률변수이지만 소문자 y는 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 Y로 표현하고 Y의 분포를 모분포라 하면 편리하다고 했다. 마찬가지로, 표본도 확률변수로 표현하면 사용하기에 편리하다.

§1.2에서, 표본을 모집단의 임의 부분집합으로 정의했다. 결론부터 말하자면, 모집단의 임의요소를 Y로 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 Y를 generic 확률변수라 한다.

크기가 N인 모집단의 부분집합은 모두 2^N개인데, 공집합과 전체집합을 제외하면 $2^N - 2$개이다. 이 중에서 크기가 n인 것은 $\binom{N}{n}$개가 있다 ($1 \le n \le N-1$). N개에서 n개를 "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 것이다.) 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 Y가 "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 Y의 분포인 모분포를 따른다. 즉, $Y_1, \cdots, Y_n$은 각각 모집단의 임의요소를 나타낸다. 이는 $n=1$인 표본 $\{Y_1\}$에서는 당연하다. 그러나, $n \geq 2$ 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_2라 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_1은 X_1일 수도 있고 X_2일 수도 있다. Y_2 역시 X_1일 수도 있고 X_2일 수도 있다. 단, 이 예제에서는 모집단의 요소가 모두 다르므로, $X_1 \neq X_2$ 이고 $Y_1 \neq Y_2$ 이다. 이제 X_1과 X_2의 분포를 구한다. X_1은 처음 뽑히는 요소이므로 당연히 $P(X_1=x)=1/5$이다 ($x=1,3,5,7,9$). 즉, X_1의 분포는 모분포와 같다. 그런데, X_2는 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_1에 관한 정보가 없으면 X_2의 (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_1일 수도 있고 X_2일 수도 있는 Y_1과 Y_2의 분포는 모두 모분포와 동일하다. 그러나, X_1과 X_2가 서로 종속이듯이 Y_1과 Y_2도 서로 종속이다. (비고: X_2가 X_1에 종속이면 X_1도 X_2에 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_2의 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, N이 커지면 비복원추출에 따른 종속성이 약해지고, 극단적으로 $N \to \infty$이면 종속성을 완전히 무시할 수 있게 된다. 그런데, N이 크더라도 n도 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, N이 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. n번에 걸쳐서 추출된 크기가 1인 표본들을 $\{Y_1\}, \{Y_2\}, \cdots, \{Y_n\}$이라 하면, 매번 N개에서 하나를 임의로 추출하므로 $Y_1, \cdots, Y_n$의 분포는 모두 모분포와 같다. 또한, n개의 표본이 각각 독립적으로 추출되었으므로 $Y_1, \cdots, Y_n$은 서로 독립이다. 예를 들어, i번째 결과는 j번째 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_i의 분포는 모분포와 같

다 ($i = 1, \cdots, n$). 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 N인 모집단의 임의요소를 Y라 하고, 편의상 Y의 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

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

가 된다. (물론, Y의 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 N이다. 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 Y를 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 N마리 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_i의 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. (S는 sum을 의미함.) 또한, $s=y_1+y_2+y_3$라 하자. S는 임의표본에 속한 꼬리표를 단 동물의 수를 의미하고, s는 관찰된 표본에 속한 꼬리표를 단 동물의 수를 의미한다.

이 문제에서는 s값이 1로 주어져 있다. 그리고 s가 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.

S의 분포는 초기하(hypergeometric)분포로 알려져 있다 (§2.2.1 참조). 즉,

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

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

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

<표 1.1> N과 $L(N)$

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, N에 대한 최우추정치(maximum likelihood estimate)는 11과 12이다. 즉, $L(N)$을 최대가 되게 하는 N값이 최우추정치이다. 사실 N의 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 N값에 따라 다른데, 만약 N이 11또는 12라면 그 확률이 최대가 된다. 따라서, 기왕이면 "세마리 중 한마리가 꼬리표를 달고 있을 가능성"이 가장 큰 경우인 11과 12를 N에 대한 추정치로 사용하자는 것이다.

LRT에 대한 개요는 다음과 같다. A,B,C 세사람이 각각 N이 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, N마리 중에 m마리가 꼬리표를 달고 있을 때, N마리에서 임의로 n마리를 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 n번을 (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 N값인데, 이는 식 (1.6.3)을 미분해서 얻는다. (비고: N을 양의 실수로 간주하여 미분을 함. 그러나, 이 문제에서는 $dL(N)/dN$을 0이 되게 하는 N값이 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 N에 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/s에 대응하는 최우추정량 mn/S를 얻는

다.

<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, m과 n도 아직 정해지지 않은 상태일 수 있다. 오히려, m과 n을 미리 정하는 것보다는 이들을 제어(control)용 모수(parameter)로 활용하는 것이 바람직하다. 최우추정량 mn/S의 확률분포로부터 추정의 정확도를 가름할 수 있는데, 이때 정확도를 어느 수준으로 올리기 위해서는 m, n이 얼마이어야 되는지를 계산할 수 있다. (비고 : m과 n의 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 $N=\widehat{N}$를 대입한 $L(\widehat{N})$의 배율인

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

가 기준치 이상이면 "$N=x$"라는 주장(또는 가설)을 받아들인다. 그리고, 식 (1.6.5)에서 $s=\sum_{i=1}^{n} y_i$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 Y라 하고, Y의 분포를 모분포라 한다. 모집단의 임의 부분집합인 표본은 $\{Y_1, \cdots, Y_n\}$으로 표현하는데, Y_i의 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.