모르고 해도 문제

오류 없는 통계 보도를 위한 세 가지 원칙

하상응 / 서강대 정치외교학과 교수

기자는 통계를 자주 인용한다. 객관성을 담보한 기사를 만들기 위해서다. 그러나 간혹 통계를 잘못 인용해 오보를 내기도 한다. 통계 수치의 의미를 제대로 파악하지 못해서다. 고의가 아니어도, 몰라서 잘못된 정보를 전하면 그것도 문제다. 선거여론조사를 예로 들어, 통계 보도 시 유의해야 할 원칙을 알아본다. 편집자 주

대의민주주의 제도를 채택하고 있는 우리나라 국민은 선거에 참여할 기회를 자주 얻는다. 4년 주기로 돌아오는 국회의원 선거와 지방 선거, 5년 주기로 돌아오는 대통령 선거가 있을 뿐만 아니라 드물지 않게 생기는 보궐선거까지 포함한다면 선거가 없는 해를 찾아보기 어려울 지경이다. 당장 내년 2021년만 해도 원래는 선거가 예정되지 않은 해이지만, 예상하지 못한 사건들로 인해 서울과 부산에서 시장 선거를 치러야 한다.

언론은 선거가 있을 때마다 다양한 정보를 유권자에게 제공해 준다. 후보들의 공약, 정당들의 정책 입장과 같이 투표 선택에 도움을 주는 정보뿐만 아니라, 유권자들이 현직자에 대해 어떤 평가를 하는지 혹은 어떤 후보가 당선될 가능성이 높은지를 설문조사에 따라 보도한다. 그런데 선거여론조사 결과를 보도하는 과정에서 오류들이 자주 발견된다. 이러한 오

류는 보통 설문조사에서 확인되는 통계 수치의 의미를 제대로 파악하지 못했기 때문에 발생한다.

모집단과 표본

설문조사 결과를 제대로 읽기 위해서는 모집단(population)과 표본(sample)이라는 개념을 이해해야 한다. 모집단은 우리가 '알고 싶은' 정보를 담고 있는 집단이다. 표본은 모집단의 한 부분으로서, 우리가 '알 수 있는' 정보를 제공해주는 집단이다. 가령 우리가 한국 유권자들의 기본소득제도에 대한 찬반 여부를 확인하고 싶어서 1,000명을 대상으로 설문조사를 수행했다면, '한국의 유권자 전체'가 모집단, '1,000명의 설문 응답자'가 표본이 되는 것이다.

여기서 두 가지 문제가 있다. 첫 번째 문제는 모집단의 규모가 너무 커서 우리가 현실적으로 모집단 수준의 정보를 직접 얻을 수 없다는 것이다. 한국의 모든 유권자와 동시에 접촉해 기본소득제도에 대한 의견을 물을 방법은 없다. 그렇기 때문에 우리는 모집단의 일부인 표본을 사용해서 정보를 얻는다. 여기서 두 번째 문제가 발생한다. 모집단의 일부인 표본은 모집단과 동일하지 않기 때문에, 모집단 수준의 정보를 정확하게 대변하지 못한다는 것이다. 1,000명의 유권자로 구성된 표본에서 560명, 즉 56%가 기본소득제도에 찬성한다고 대답을 했어도 그것이 곧 한국 유권자의 56%가 기본소득제도에 긍정적인 입장을 취한다는 것이 아니라는 말이다.

추론과 불확실성

우리가 아는 정보인 표본 수준의 정보를 이용해 우리가 알고 싶은 모집단 수준의 정보를 유추하는 행위를 통계적 추론(statistical inference)이라고 한다. 그런데 표본에서 얻은 정보가 정확하게 모집단 수준의 정보라는 확신이 없다. 그래서 모집단 수준의 정보를 추론하는 과정에서 우리는 겸손하게 불확실성을 표시해 준다. 보통, 이 불확실성 정보는 오차의 한계(margin of errors) 혹은 신뢰구간(confidence interval)이라는 개념으로 표현된다.

"우리가 아는 정보인 표본 수준의 정보를 이용해 우리가 알고 싶은 모집단 수준의 정보를 유추하는 행위를 통계적 추론(statistical inference)이라고 한다. 그런데 표본에서 얻은 정보가 정확하게 모집단 수준의 정보라는 확신이 없다. 그래서 모집단 수준의 정보를 추론하는 과정에서 우리는 겸손하게 불확실성을 표시해 준다. 보통, 이 불확실성 정보는 오차의 한계(margin of errors) 혹은 신뢰구간(confidence interval)이라는 개념으로 표현된다."

가상의 대통령 선거에서 두 후보가 경쟁하는 상황을 상정해 보자. 어떤 후보가 당선 가능성이 높은지를 파악하기 위해 1,000명의 유권자를 대상으로 설문조사를 했다. 여기서 A후보를 찍겠다고 한 응답자가 510명(51%)이고, 나머지 490명(49%)이 B후보를 찍겠다고 응답했다. 여기서 확인되는 2%포인트 차이는 표본 수준의 정보이다. 즉, 우리가 얻은 설문자료에서는 A후보가 B후보를 2%포인트 차이로 앞서는 것처럼 보이지만, 이것이 실제 모집단 수

준에서도 그러할지는 알 수 없다. 그렇기 때문에 설문조사 결과를 제시하는 과정에서 항상 요구되는 것이 불확실성 정보, 즉 오차의 한계이다.

위의 예에서 오차의 한계가 ±3%포인트라고 해보자. 오차의 한계를 고려한 신뢰구간을 활용해 추론된 모집단 수준의 정보는 다음과 같다.

A후보: 51%±3%포인트 = [48%, 54%]
B후보: 49%±3%포인트 = [46%, 52%]

A후보의 득표율이 이 특정 설문조사 표본에서는 51%이나, 불확실성을 고려하면 모집단 수준에서 48%와 54% 사이 어딘가에 있을 것이라는 말이다. 마찬가지로 B후보의 득표율은 이 표본에서는 49%이나, 불확실성을 나타내는 정보인 오차의 한계를 고려하면 모집단 수준에서 46%와 52% 사이 어딘가에 있을 것이다. 따라서 실제 선거에서 A후보가 54%, B후보가 46%의 득표율을 거둬 A후보가 이길 수도 있지만, A후보가 총 투표수의 48%, B후보가 52%를 얻어 B후보가 당선될 수도 있다는 말이다. 즉, 51% 대 49%로 나뉜 표본 정보에 근거해 섣불리 A후보가 실제 선거에서 승리할 것이라고 해석하면 안 된다.

표본의 가변성

현재 우리나라에는 다양한 여론조사기관이 있다. 이 기관들은 모두 자기 나름의 방식으로 설문조사를 수행해 여러 가지 유용한 정보를 유권자와 언론에 제공한다. 선거 때 여론조사기관은 전체 유권자 집단이라는 하나의 모집단을 공유한다. 하지만 각 여론조사기관이 이 모집단에서 추출한 표본은 서로 다를 가능성이 높다. 다시 말해 10개의 여론조사기관이 각각 1,000명의 유권자를 이용한 설문조사를 수행한다고 할 때, 두 개 이상의 여론조사기관 표본에 포함된 유권자는 거의 없을 것이다. 10개의 여론조사기관에서 사용하는 표본에 포함된 1,000명의 유권자가 서로 다르기 때문에 10개의 조사 결과는 다를 수밖에 없다. 이 현상을 표본의 가변성(sampling variability)이라고 부른다.

선거 때 여러 여론조사기관에서 예측하는 내용이 다르다고 해서 여론조사 자체를 신뢰하지 않는 것은 바람직하지 못하다. 그러면 이렇게 서로 다른 여론조사 결과 중 어느 것을 보도해야 하는가? 보도의 객관성을 유지하려면 이 모든 결과를 고려해 신중하게 보도해야 한다. 확인된 결과들의 평균을 내 보도하거나, 대표적인 패턴을 보이는 몇 가지 결과를 부각해 보도하면 보통 큰 문제가 되지 않는다. 하지만 특정 여론조사 결과는 보도하고, 그것과 상충하는 결과는 무시해서는 안 된다. 보도의 객관성이 보장되지 않기 때문이다.

대표성 있는 표본

전술한 바와 같이 여론조사에서는 모집단의 규모가 너무 커서 그 정보를 직접적으로 얻을 수 없기 때문에 표본을 사용하는 것이 일반적이다. 표본은 모집단의 일부이지만 모집단 그 자체일 수는 없기 때문에 표본으로부터 얻은 정보는 모집단 수준의 정보와 어느 정도 다르기 마련이다. 그런데 하나의 표본이 잘 구성돼 모집단의 성격을 정확히 반영하는 상황을 상상해 볼 수 있다. 이러한 표본을 대표성 있는 표본(representative sample), 즉 모집단을 대표하는 표본이라고 부른다.

대표성 있는 표본을 만들기 위해 필요한 조건들은 이미 잘 알려져 있다. 문제는 실제 여론조사를 수행하는 과정에 이러한 조건들을 만족시키기가 어렵다는 점이다. 대표성 있는 표본을 얻기 위해 필요한 가장 중요한 조건은 확률에 근거해 표본을 구축해야 한다는 것이다. 간단히 예를 들자면, 한국 유권자가 4,200만 명이라고 했을 때 표본에 포함될 확

률이 각각의 유권자에게 4,200만분의 1로 동일해야 한다는 말이다. 어떤 유권자는 표본에 포함될 확률이 4,200만분의 1인데 다른 유권자의 확률은 4,200만분의 1보다 작거나 크면, 그 표본은 대표성 있는 표본이 되기 어렵다.

할당표집의 문제

문제는 한국에서 통용되는 할당표집(quota sampling)이 확률에 근거한 표본의 구축 방법이 아니라는 것이다. 할당표집은 유권자의 연령, 성별, 그리고 거주 지역에 근거해 유권자 집단을 나누고, 인구조사에서 확인된 정도의 비율을 맞추어 표본을 만드는 방법을 의미한다. 따라서 할당된 특정 집단에 해당하는 응답자가 잘 채워지지 않는 경우, 그 응답자를 의도적으로 찾아 나서는 일이 빈번하다. 예를 들어 강원도에 사는 20대 여성 두명을 표본에 포함해야 하는데 이들을 찾기 어려우면 계속 이 조건에 만족하지 못하는 유권자들을 버리는 작업을 하게 된다. 이 과정에서 특정 유권자가 표본에 포함될 확률이 높아지거나 낮아지게 되기 때문에 대표성 있는 표본을 구축할 수 없게 된다.

할당표집 방식이 대표성 있는 표본의 구축을 불가능하게 한다는 사실은 또 다른 심각한 문제와 연관돼있다. 위에서 오차의 한계를 고려한 신뢰구간을 보고해 불확실성이 고려된 해석을 유도해야 한다고 말했다. 그런데 오차의 한계와 신뢰구간은 모두 확률에 근거한 표본 구축 방식을 사용하는 경우에만 의미가 있는 개념이다. 즉, 할당표집 방식을 통한 표본을 사용해 설문조사를 했다면, 엄격히 말해 그 결과를 해석하는 과정에서 오차의 한계와 신뢰구간을 적용할 수 없다. 그래서 양심적인 조사회사에서는 할당표집을 한 자료의 오차의 한계를 보고할 때 "확률 표집을 전제했을 때 ±3%포인트"와 같이 전제조건을 명확히 제시한다.

여론조사 결과 단정적 표현 지양해야

선거여론조사에서 모집단은 모든 유권자들이다. 이들을 동시에 접촉해 정보를 얻을 방법이 없기 때문에 우리는 표본을 사용해 모집단 정보를 추론한다. 모집단과 표본은 동일할 수 없기 때문에, 표본을 사용해 보고하는 정보는 '정확'할 수 없다. 다만 표본 정보에 오차의 한계 혹은 신뢰구간과 같은 불확실성을 덧붙여 조심스럽게 모집단 정보에 대한 언급을 하고자 할 뿐이다. 따라서 절대 여론조사 결과를 단정적으로 표현해서는 안 된다. 가능한 한 가장 조심스럽게, 다양한 해석의 여지를 두면서 보도해야 과학적 엄밀성과 보도 윤리에 해를 끼치지 않을 것이다.

<오류 없는 통계 보도를 위한 3가지 원칙>

❶ 설문조사 결과를 보도하려면 반드시 조사회사에서 제공해 주는 오차의 한계, 신뢰구간 정보를 병기해야 한다.

❷ 조사회사에서 확률표집 방식을 사용해 표본을 구축했는지 여부를 확인하고, 만약 그렇지 않았다면 결과 보고 시 대표성 있는 표본을 사용하지 않은 결과라는 사실을 명확히 기입해야 한다.

❸ 그 어떤 설문조사 결과도 단정적으로 해석해서는 안 된다.

> 선거 때 서로 다른 여론조사 결과 중 어느 것을 보도해야 하는가? 보도의 객관성을 유지하려면 이 모든 결과를 고려해 신중하게 보도해야 한다. 확인된 결과들의 평균을 내 보도하거나, 대표적인 패턴을 보이는 몇 가지 결과를 부각해 보도하면 보통 큰 문제가 되지 않는다. 하지만 특정 여론조사 결과는 보도하고, 그것과 상충하는 결과는 무시해서는 안 된다. 보도의 객관성이 보장되지 않기 때문이다.