☒ Teacher's Corner

잘못쓰기 쉬운 통계의 함정에 관한 고찰

박성현
서울대학교 계산통계학과

A study on misused statistical pitfalls

Sung H. Park
Dept. of Computer Science and Statistics Seoul National University

1. 서론

통계학은 20세기에 시작된 새로운 학문분야이다. 통계학(statistics)이란 용어는 원래 국가산술(state arithmetic)이란 말에서 나왔다고 하며, 역사적으로 정치가들이 국가의 살림을 꾸려가기 위하여 필요한 숫자를 체계적이고 과학적으로 산출해 내기 위한 방법을 연구하는 학문으로 출발하였다. 그러나 현대적인 통계학은 "국가산술"의 영역을 벗어나서 의사결정과학(decision-making science)으로 발전하고 있다. 통계학의 올바른 정의는 사회, 자연 및 인간생활의 온갖 현상을 연구하기 위하여 불확실성(uncertainty)이 내포된 자료의 선택, 관찰, 분석 및 추정을 통하여 의사결정에 필요한 정보의 획득과 처리방법을 연구하는 학문이라고 말할 수 있다.

사회가 발전함에 따라서 통계학의 필요성은 더욱 커질 것으로 예상된다. 사회가 발전함에 따라서 우리 주위에는 엄청난 양의 데이터가 매일 매일 발생하고 있으며, 이를 분석하여 정보를 얻는 것은 "정보화 시대"를 살아가는 우리들에게 필요불가결한 요소이다. 이러한 역할을 통계학이 해줄 것이다. 통계학은 정보를 획득하고, 불확실성 하에서 의사결정을 도우며, 얻어진 정보를 체계적으로 보는 사고의 틀을 제공하고 있다.

이 글은 통계학의 각종 분석방법을 사용하여 정보를 얻고, 그 결과를 해석하는 과정에서 자칫 범하기 쉬운 몇 가지의 오류인 "통계의 함정"에 대하여 다루려고 한다. 이 글은 통계의 유용성을 독자들에게 일깨워 주면서, 또한 빠지기 쉬운 통계의 함정을 알게 하여 통계를 보는 안목을 높이고자 하는데 목적이 있다.

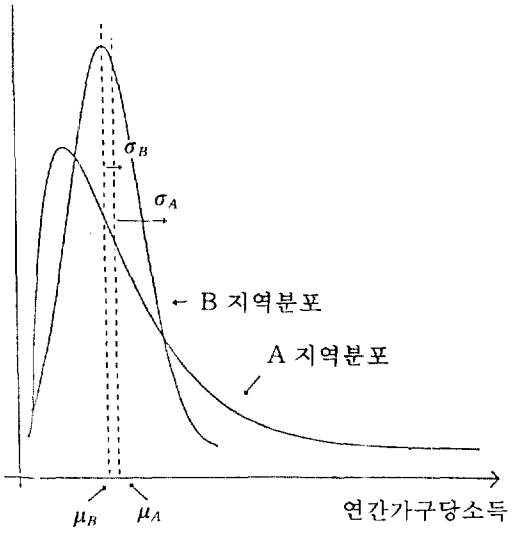
2. 평균이 주는 함정

보통 우리는 평균으로만 말한다. 예를 들면, 한국의 국민 일인당 GNP가 만 불에 도달했다던가, 출근할 때 걸리는 시간이 60분이라든가 하는 것들이다. 그러나 평균만으로는 전체의 상황을 나타낼 수 없는 경우가 많다. 집단의 평균은 집단내의 구성원들이 어떻게 서로 다르며, 또 서로 어느 정도 흩어져 있는지를 말해주지 못함으로 인하여, 평균만에 의존한 결정은 잘못되기 쉽다. 예를 들어보자. 어느 은행에서는 하나의 지점을 개설할 계획으로 있는데, 두 지역 A와 B가 그 대상으로 거론되고 있다. A 지역 주민의 가구당 평균소득이 연간 1500만원이고, B 지역 주민의 평균소득은 1450만원이다. 이 때 평균소득이 A 지역이 높다고 해서 A 지역에 지점을 개설하기로 결정한다면 잘못된 결정일 수도 있다. 만약 A 지역은 대부분의 사람이 가난하고 일부 가구만 부유하여 평균소득이 높아지고, B 지역은 대부분의 가구가 중산층이라면, 평균소득에서는 B 지역이 A 지역보다 떨어지나 저축률이 더 높을 수 있고, B 지역에 지점을 개설하는 것이 더 바람직하다고 볼 수 있다.

평균소득만으로는 가구당 소득의 분포상황을 알 수 없으며, 이 경우에는 표준편차라는 소득의 산포를 나타내는 통계숫자를 사용하면 편리하다. A, B 지역의 소득분포를 조사하여 보니 <표 1> 과 <그림 1>과 같다

〈표 1〉 가구당 소득의 평균과 표준편차

	A 지역	B 지역
연간 가구당소득의 평균	1500 만원	1450 만원
연간 가구당소득의 표준편차	700 만원	200 만원



〈그림 1〉 가구당 수입의 분포

3. 비율이 주는 함정

우리는 일상생활에서 흔히 비율에 관한 통계를 접하게 되며 또한 필요로 하게 된다. 가령 어느 정당에 대한 지지율이라든가, 경제활동인구의 실업률이라든가, 공산품의 불량률 등, 이루 셀 수 없이 많은 종류의 비율이 있다. 그러나 비율의 사용상에도 함정이 있으며, 대표적인 예를 다음에 두 가지만 보이기로 한다. 이러한 함정을 통하여 비율에 관한 통계적 감각을 가질 수 있을 것이다.

3.1 대표성의 파라독스

어느 신문의 칼럼니스트가 두 동종 대기업 '(주)수통' 과 '(주)계통'이 대졸사원을 신규채용할 때 출신지역별 차별이 있었다는 것을 주장하기 위하여 다음의 자료를 제시하였다고 하자.

〈표 2〉 회사에 따른 지역별 지원자와 입사자 비율

(주) 수통		
출신지역	지원자 비율	입사자 비율
A	70 %	50 %
B	20 %	28.6 %
C	10 %	21.4 %
합계	100 %	100.0 %

(주) 계통		
출신지역	지원자 비율	입사자 비율
A	10 %	3.8 %
B	20 %	15.4 %
C	70 %	80.8 %
합계	100 %	100.0 %

이 칼럼니스트의 주장은 "(주)수통은 B지역의 지원자를 우대하였고, (주)계통은 B지역 지원자를 차별하였다" 고 하는 것이다. 이 주장은 어느 정도 타당성이 있는 것인가? 이 질문에 답하기 위해서는 지원자수와 합격률에 관한 자료를 조사해 볼 필요가 있다.

〈표 3〉 회사에 따른 지원자수, 합격률, 입사자수

(주) 수통

출신지역	지원자수(비율)	합격률	입사자수(비율)
A	700 명(70 %)	10 %	70 명(50 %)
B	200 명(20 %)	20 %	40 명(28.6 %)
C	100 명(10 %)	30 %	30 명(21.4 %)
합계	1000 명(100 %)		140 명(100 %)

(주) 계통

출신지역	지원자수(비율)	합격률	입사자수(비율)
A	100 명(10 %)	10 %	10 명(3.8 %)
B	200 명(20 %)	20 %	40 명(15.4 %)
C	700 명(70 %)	30 %	210 명(80.8 %)
합계	1000 명(100 %)		260 명(100 %)

<표 3>을 보면 (주)수통과 (주)계통 간에는 B지역출신에 대한 아무런 차이가 없음을 알 수 있다. 이러한 현상을 대표성의 파라독스(representation paradox)라고 부른다. 이러한 파라독스는 지원자수가 상이할 때에 발생되는 현상으로, 지원자수와 입사자수의 비율만을 가지고 출신지역별 차별이 행해졌다고 말하는 것은 잘못된 판단이다.

3.2 심슨의 파라독스

의약품을 개발하고 있는 어느 회사에서 임상실험(clinical experiment)을 실시하고 있다. 예를 들어, 어떤 암의 치료를 위하여 새로이 개발된 '자사항암제'와 이미 시판되고 있는 '타사항암제'의 임상효과를 실증적으로 비교하고자 한다. 100 명의 암환자에게 자사항암제를 투약하고, 다른 100 명의 암환자에게 타사항암제를 투약하여, 1년의 치료기간이 지난 후 그 결과가 다음과 같이 나타났다고 하자.

〈표 4〉 항암제 생존율의 단순 비교

항암제	생존	사망	합계	생존율
자사	50 명	50 명	100 명	50 %
타사	30 명	70 명	100 명	30 %
합계	80 명	120 명	200 명	

<표 4>에서 보면 자사항암제의 생존율 50%가 타사항암제의 30%보다 월등 높으므로, 이 회사는 성공의 축배를 들려고 할 것이다. 그러나 좀더 세밀하게 환자의 남녀별로 조사하여 보면, <표 4>와 전혀 다른 결과를 얻을 수도 있음을 발견하게 된다.

〈표 5〉 성별 항암제 생존율의 비교

남자(명)				
항암제	생존	사망	합계	생존율
자사	48	32	80	60 %
타사	14	6	20	70 %
합계	62	38	100	

여자(명)				
항암제	생존	사망	합계	생존율
자사	2	18	20	10 %
타사	16	64	80	20 %
합계	18	82	100	

<표 5>에서 보면 남자의 경우에도 타사항암제가 생존율이 높고, 여자의 경우에도 타사항암제의 생존율이 높다. 따라서 자사 것이 타사 것 보다 오히려 나쁘다는 것을 말해 주고 있다. 그러면 어째서 이러한 결과가 발생한 것일까? <표 4>의 단순비교 결과가 어떻게 얻어 졌는가 살펴보자.

$$\begin{aligned}\text{자사항암제 생존율} &= (80\text{명}/100\text{명})(60\%) + (20\text{명}/100\text{명})(10\%)\\ &= (0.80)(60\%) + (0.20)(10\%)\\ &= 50\ \%\\ \text{타사항암제 생존율} &= (20\text{명}/100\text{명})(70\%) + (80\text{명}/100\text{명})(20\%)\\ &= (0.20)(70\%) + (0.80)(20\%)\\ &= 30\ \%.\end{aligned}$$

즉, 각 항암제의 전체적인 생존율은 남자와 여자의 가중평균이지만, 동일한 가중치가 적용되지 않음으로 인하여 비교상의 불일치가 발생된 것이다. 이러한 결과는 임상실험의 대상이 되었던 남자와 여자의 수가 항암제에 따라 큰 차이가 있을 때 발생될 수 있다. 이와 같은 현상을 심슨의 파라독스라고 부르며, 실지 사회에서 종종 발생될 수 있는 문제이므로 주의하여야 한다.

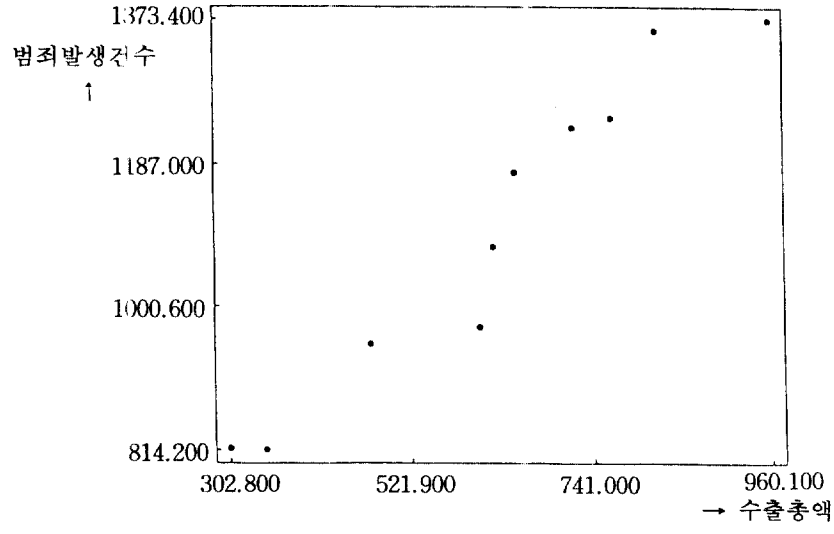
4. 상관관계가 주는 함정

우리는 종종 두 변수간에 산점도(scatter diagram)를 그리고 상관계수 r을 구한 후에 r의 절대값이 크면 두 변수간에 상관관계가 있다고 쉽게 결론 짓는다. 그러나 이와 같은 성급한 결론은 금물인 경우가 허다하다. 예를 들어보자.

〈표 6〉 연도별 수출총액과 범죄발생건수(1995년도 통계연감, 통계청)

연도	수출총액(억불)	범죄발생건수(천건)
1985	302.8	814.8
1986	347.1	814.2
1987	472.8	950.3
1988	607.0	972.6
1989	623.8	1078.1
1990	650.2	1175.2
1991	718.7	1233.4
1992	766.3	1245.2
1993	822.4	1359.1
1994	960.1	1373.4

<표 6>에는 과거 10년간 우리 나라의 수출총액과 범죄발생건수가 나타나 있으며 이를 산점도로 그려보면 <그림 2>가 얻어진다.



〈그림 2〉 수출총액과 범죄발생건수의 산점도

<그림 2>에서 보면 수출총액과 범죄발생건수간에 높은 상관관계가 있어 보인다 (r = 0.967). 즉, 수출이 증가하면 범죄건수가 증가한다고 속단할 수도 있을 것이다. 그러나 실제로는 우리 나라가 경제성장이 되면서 수출총액이 매년 늘어나고 있고, 여러 가지 사회적 문제로 인하여 범죄발생 건수가 매년 늘어나는 것이지, 수출이 범죄발생의 원인이라고 볼 수는 없는 것이다. 만약 그렇다면 수출을 줄이면 범죄발생이 준다고 볼 수 있겠는가? 이것은 시간의 흐름이라는 또 하나의 변수가 작용한 것이지, 이 두 변수간에는 직접적인 상관관계가 있다고 보기는 어렵다.

5. 여론조사의 함정

여론조사는 모집단(population)이 가지고 있는 의견을 파악하기 위하여, 표본(sample)을 취하여 여론을 객관적으로 조사하는 통계적 조사수단으로, 오늘날 국민의 여론을 빠르게 파악하는 중요한 수단으로 활용되고 있다. 최근 여론조사는 정당의 지지율, 어떤 정책의 찬성률, 국민의 의중을 파악하는 정치적 도구로도 사용되고 있다. 그러나 여론조사는 신중한 계획과 치밀한 준비 및 실행이 이루어지지 않으면 그 결과에 대한 신뢰성이 문제가 된다.

신뢰성 있는 여론조사 결과를 얻으려면 다음과 같은 구비여건이 필요하다.

(1) 오차가 적게 발생될 수 있는 좋은 표본설계(sample design)
(2) 잘 훈련된 조사원
(3) 조사항목의 내용이 확실하여 오해의 소지가 없을 것.
(4) 피조사자가 성의 있게 조사에 응하는 것.

이러한 구비여건 없이 엉성하게 실시된 여론조사의 결과는 통계의 신뢰성만을 떨어뜨리는 결과를 초래하게 될 것이다.

표본설계가 잘못되어 오차가 크게 발생한 전형적인 예로서, 1936년 미국 대통령선거에 대한 Literary Digest 회사의 여론조사 실패담이 있다. 이 회사는 전화번호부와 자동차 등록대장에서 추출한 200만명 이상의 유권자를 대상으로 여론조사를 실시한 결과, Landon 후보가 Roosevelt 후보를 압도적으로 승리하리라는 것이었으나, 투표결과는 오히려 Roosevelt가 압도적으로 승리한 것으로 판명되었다. 이 당시 미국에서 전화를 갖거나 자동차를 소유한 유권자는 상류층이었으며, Landon은 상류층에 인기가 있었고, Roosevelt는 서민층에 인기가 있었기 때문에, 이 회사의 표본설계는 근본적으로 잘못된 것이었다. 이 표본설계는 그 당시 전화나 자동차가 없는 전 인구의 75%에 해당하는 유권자를 표본조사의 대상에서 제외시킨 결과가 되었다. 따라서 오차가 엄청나게 커진 것이다.

6. 결론

아직까지 일상생활에서 만나기 쉬운 몇 가지의 통계의 함정에 대하여 기술하였다. 곧 도래하게 되는 21세기 정보화시대에 대비하기 위하여, 우리는 올바른 통계적 사고를 할 필요가 있으며, 또한 통계정보를 바르게 이해하는 것은 21세기를 살아가는 문명인의 바람직한 소양이라고 하겠다. 이 글에 실린 내용들과 관련이 있는 글을 더 읽기를 원하는 독자는 참고문헌(1, 2, 3, 4, 5)을 참조하여 주기 바란다.

참고문헌

[1] American Statistical Association. (1972), Statistics: A Guide to the Unknown, edited by Judith M. Tanur, Holden-Day, San Francisco.

[2] Campbell, S.K. (1974), "Flaws and Fallacies in Statistical Thinking," Prentice-Hall, New Jersey.

[3] Huff, D. (1954), "How to Lie with Statistics," Penguin Books, London.

[4] 이재창 외 5인 편역 (1984), "쉽게 읽는 생활 속의 통계학," 세경사, 서울.

[5] 한국통계학회 (1991), "알고 보면 재미있는 통계이야기," 자유아카데미, 서울.