

견해의 수정

텐트 너머로 총소리가 울렸다. 총탄이 뿔뿔거리며 날아와 모래땅을 스치고 지나갔다. 그는 자신의 정체에 발각될 확률이 얼마일까 계산해보았다. 그리고 확률이 50대 50이라고 결론내렸다.

존 업다이크(John Updike), 『쿠테타(The Coup)』

“노이로제에 대한 나의 이론은 놀랄 정도로 간단하지. 노이로제 환자는 자신이 위험에 노출되어 있는 확률에 대해 왜곡된 생각을 가지고 있는 사람이야. 예를 들면 그 또래의 사람들에게 심장마비가 올 확률이 1/1,000 정도일 경우, 노이로제 환자는 그 확률이 자신에게는 1/2이라고 생각하지. 비행기 사고가 날 확률은 1/100,000인데 노이로제 환자는 비행기를 탈 때 그 확률이 1/3이라고 생각하기도 한단 말이야.”

영 교수는 자신에게 쏟아질 찬사를 들으려고 잠시 말을 끊었다. (분명한 것은, 그는 조바심을 내면서 자신의 능력에 의문을 가지는 노이로제 환자가 아니라는 사실이다.)

엘렌은 미소를 지었다. 바로 영 교수가 추측했던 대로. 그렇기 때문에 그녀가 다음에 내뱉은 말은 두 배의 충격으로 다가왔다.

“영 교수님, 제 생각으로는 통계학자들이 사람들을 제대로 이해할 확률은 언제나 1/3에 불과합니다. 하지만 당신이 그 명석한 두뇌로 고통을 겪고 있는 사람들의 문제를 차가운 숫자놀음으로 격하시키신 다음에는, 통계학자들이 사람들을 이해하거나 그들의 아픔에 동감할 확률은 1/3,000,000에 불과하다고 생각하시게 되었습니다. 당신과 같은 부류의 사람들이 이야기하듯이 통계학자들과 보통 사람들은 서로 배타적인 범주에 속하는군요.”

엘렌은 울음을 터뜨리면서 문을 박차고 나가버렸습니다.

영 교수는 혼잣말로 중얼거렸다. “엘렌은 통계학과 통계학자 일반에 대해서 많은 것을 배워야겠군. 아마도 나 같은 통계학자가 가르쳐야만 할 거야.”

잉그램 카스웰(Ingram J. Carswell), 『숙녀와 통계학자(The Lady and the Statistician)』

우리는 어떤 특정한 사건에 대해 새로운 정보를 얻게 되었을 경우 그 사건이 일어날 확률에 대한 우리의 견해를 수정하기도 한다. 몇 가지 예를 들어 보자.

1. 비가 올 확률 : 당신은 밤 11시 15분에 내일 비가 올 확률이 60%라는 일기예보를 듣고 잠자리에 들었다. 그런데 아침 7시에 일어나보니 하늘이 먹구름으로 뒤덮여 있고 멀리서 천둥치는 소리가 들렸다고 하자. 그러면 아마도 당신은 비가 올 확률이 60%라는 이전의 추정값을 수정하고 싶어 할 것이다. 의심할 여지없이 당신은 그 추정값을 더 높일 것이다.

2. 안타를 칠 확률 : 랜스 스타는 로스앤젤레스 다저스 팀의 야구선수이다. 올해 그는 공식적으로 390 '타수'에 130 안타를 기록하고 있다. 야구용어로 말하면 타율이 3할 3푼 3리인 것이다. 이 선수가 **다음** 타석에서 안타를 칠 확률이 얼마인가라는 질문을 받게 되면 $1/3$ 이라고 대답하는 것이 합리적인 것처럼 보인다. 그러나 여러분이 다음과 같은 정보를 더 얻게 되었다고 해보자. 로스앤젤레스 다저스 팀의 다음 상대는 샌프란시스코 자이언츠팀이다. 이 팀의 선발투수로 지명된 쉐 매그넛은 다저스 '킬러'로 알려져 있고 랜스 스타도 그에게는 숨방망이였다. 올해 랜스 스타는 매그넛에게 10타수 1안타 밖에 기록하지 못했다. 이런 정보에 비추어보면 $1/3$ 이라는 여러분의 추정값을 수정하는 편이 합리적일 것 같다.

3. 병에 걸릴 확률 : 클라크 박사는 폐결핵의 발병률이 대략 0.001인 뉴욕 시에서 병원을 개업하고 있다. 즉 클라크 박사에게 정기검진을 받으러 온 사람들 중에서 어떤 사람이 실제로 폐결핵에 걸려 있을 확률에 대한 추정값이 0.001이라는 것이다. 클라크 박사는 검진을 할 때 보조자료로 삼기 위해 관례적으로 흉부 X레이 촬영을 요구했다. X레이 결과가 양성으로 나오면 폐결핵을 앓고 있다는 징조이다. 하지만 X레이 결과가 양성으로 나왔다고 해서 그 사람이 틀림없이 폐결핵 환자라는 것은 아니다. 그러나 X레이 결과가 양성으로 나오면 클라크 박사는 그 사람이 폐결핵에 걸려 있을 확률을 0.001보다 높게 수정할 것이다.

이러한 예들로부터 분명하게 알 수 있는 사실은, 관련된 정보를 추가로 얻게 되면 처음에 추정했던 확률에 대해 다시 생각해보게 되고 그것을 수정하게 될 것이라는 점이다. 추가 정보를 관찰한 다음에 그것을 어떻게 활용하여 처음에 추정했던 확률(**사전확률**, prior probability)을 뒤집고 수정된 확률(**사후확률**, posterior probability)을 계산할 수 있을까?

새로 얻은 정보를 이용하여 사전확률을 수정하는 것은 베이즈의 통계학적 방법론(Bayesian statistical methods)에 속한다. 이런 이름이 붙은 것은 18세기 중반의 철학자였던 토머스 베이즈(Thomas Bayes) 목사를 기리기 위해서이다.

베이즈의 방법은 원리상으로는 명쾌하다. 하지만 그것을 실제로 사용하기는 매우 어렵다. 베이즈의 통계학적 방법을 상세하게 설명하는 것은 여기서의 범위를 벗어난다. 그 대신 우리는 어떻게 확률을 합리적으로 수정할 수 있는가에 대한 간단한 예로, 앞에서 말한 검진의 예를 들어보기로 하겠다.

진찰에 오류가 전혀 없다면 확률을 수정하기는 쉬울 것이다. 그런 이상적인 세계에서는 양성반응이 나오면 폐결핵일 확률을 1로 수정하면 될 테니까 말이다. 이와 마찬가지로 음성반응이 나오면 폐결핵일 확률을 0으로 수정하면 된다. 그러나 진찰에 전혀 오류가 없을 수는 없다. 흉부 X레이를 촬영하는 경우에 발생할 수 있는 오류는 두 가지이다.

한 가지 실제로 폐결핵에 걸려 있는데도 음성반응이 나오는 것이다. 그러면 그 사람은 건강하다는 잘못된 진찰결과가 나오게 된다. 이런 오류를 **잘못된 음성반응(false-negative)**이라고 부른다. 또 한가지 오류는 폐결핵에 걸려 있지 않은데도 양성반응이 나오는 것이다. 그러면 그 사람은 건강한데도 폐결핵이라는 진찰결과가 나오게 된다. 이러한 오류를 **잘못된 양성반응(false-positive)**이라고 부른다.

잘못된 양성반응과 잘못된 음성반응이 나올 확률은, 어느 정도 X레이를 판독하는 사람의 기술에 달려 있다. 경험이 별로 없는 판독자가 필름을 판독하면 일시적인 폐의 이상을 폐결핵으로 해석하는 경우가 있는 것이다. 또 폐결핵의 진짜 증거를 놓치는 경우도 있다.

어떤 특정한 X레이 판독자가 오류를 범할 확률을 추정하는 몇 가지 방법이 있다. 예를 들면 판독자에게 무작위로 폐결핵에 걸린 사람의 X레이와 그렇지 않은 사람의 X레이를 제시해보는 방법이 있다. 판독자에게 어느 것이 폐결핵에 걸린 사람의 필름이고 어느 것이 그렇지 않은가를 물은 후에 그 결과를 보고 판독자가 오류를 범할 확률을 직접 추정할 수 있다.

예러셀미(Yerushalmy), 하크니스(Harkness), 코프(Cope), 케네디(Kennedy) 등은 1950년에 여섯 명의 판독자가 X레이를 판독하면서 저지른 오류에 대해 연구했다. 표에는 그 여섯 명의 판독자 중 한 사람에 대한 조사 결과가 나타나 있다.

폐결핵에 걸리지 않은 1,790명의 X레이 필름 중에서 이 판독자는 51개를 잘못 판독했다. 양성반응으로 잘못 처리된 비율은 $51/1,790=0.028$ 로 약 3%에 해당한다. 폐결핵에 걸린 30명의 X레이 필름에 대해서는 8개를 잘못 판독해서 음성반응으로 잘못 처리된 비율이 $8/30=0.267$, 그러니까 약 27%였다.

표. X레이 판독에 대한 연구 결과

	폐결핵에 안걸린 사람	폐결핵에 걸린 사람	합계
X레이 음성반응	1,739	8	1,747
X레이 양성반응	51	22	73
합 계	1,790	30	1,820

우리가 표에 나타난 1,820장의 X레이 필름 중에서 무작위로 한 장을 뽑는다고 가정해보자. 이렇게 뽑은 X레이 필름이 폐결핵에 걸린 사람의 것일 확률은 얼마인가? 여러분은 이렇게 생각해 볼 수 있을 것이다. 1,820장 중에서 30장이 실제로 폐결핵에 걸린 삶의 필름이니까 그 확률은 $30/1,820=0.016$ 이다. 즉 약 2% 정도가 될 것이다. (0.016이라는 수치가 폐결핵의 발병률이라고 보기에는 너무 큰 것이라는 사실을 주목하라. 그러나 표에 나타난 자료는 전형적인 집단에 대한 것이 아니다. 예러셀미 등의 연구는 판독자의 오류의 비율을 추정하기 위해서 수행된 것이다. 특히 잘못된 음성반응에 대해서 더 나은 추정값을 얻으려면 실제 상황에서 발견되는 것보다 병에 걸린 경우가 많은 편이 유리하다.)

1,820장의 X레이 필름 중에서 무작위로 선택된 하나가 판독자에 의해서 양성으로 판정되었다고 가정해보자. 여러분은 그 필름이 실제로 폐결핵에 걸린 사람의 것일 확률에 대한 추정값을 어떻게 수정하겠는가? 73장이 양성으로 판독되었는데 그중에서 겨우 22장만 진짜 폐결핵에 걸린 사람의 것이었다면, 이렇게 수정된 확률에 비추어볼 때 그가 제대로 판정했을 확률은 $22/73=0.301$ 이다. 이것은 약 30%에 해당하는 수치이다. 따라서 X레이 필름이 양성으로 판정되었다는 추가 정보는, 그 X레이 필름이 실제로 폐결핵에 걸린 사람의 것일 확률에 대한 여러분의 추정값을 0.016에서 0.301로 수정하게 해 준 셈이다.

요 약

1. 추가 정보를 얻게 되면 특정한 사건이 일어날 확률에 대한 견해가 수정된다.
2. 사전확률을 사후확률로 바꾸기 위해 자료를 사용하는 것은 베이즈의 통계학적 방법론에 속하는 문제이다.

출처 : 통계마인드 길들이기
마일즈 홀랜더, 프랭크 프로쉬안 지음, 도서출판 새날