13.2 들쥐 세기

학습지

문제

STEP II 2000, 13번 문제

더 읽어 나가기 전에 문제부터 읽어 보자. 이 문제는 STEP Support Programme의 문제 데이터베이스에 있다:

step.maths.org/questions/00-s2-q13

힌트

먼저 혼자 힘으로 문제를 풀어 보자. 막히면 힌트를 **하나씩** 읽고, 힌트를 하나 읽을 때마다 다시 문제로 돌아가자. **잠깐, 직접 해 보기**라고 표시된 힌트가 핵심 단계이다. 각 힌트 뒤에는 풀이를 적을 공간이 있다.

> ? **생각해 보기** 들쥐가 모두 N마리이고, 그 가운데 200마리에 표시가 되어 있다. 200마리를 임의로 표본으로 뽑는다. 정확히 11마리에 표시가 되어 있을 확률은 얼마인가?

1단계 가능도

다루는 내용: 확률 p_N

> ? **생각해 보기** 동일한 확률로 뽑히는 200마리짜리 모든 표본 가운데에서, 표시된 들쥐 중 11마리와 표시되지 않은 들쥐 중 189마리를 고른다. 그런 다음 N이 들어 있는 이항계수(binomial coefficient)를 계승(factorial)으로 풀어 쓴다.

2단계 자연수 범위에서 최대화하기

다루는 내용: p_N의 최대화

> **잠깐, 직접 해 보기** N은 자연수이므로 p_{N+1}과 p_N을 (또는 p_N과 p_{N-1}을) 비교한다. 계승은 대부분 약분된다. 그 비가 1보다 커지는 것은 언제인가?

3단계 세 번째 표본

다루는 내용: 세 번째 표본

? **생각해 보기** 두 번째 표본을 뽑은 뒤에는 표시된 들쥐가 몇 마리인가? 앞에서 구한 N의 추정값으로 보면, 표시되지 않은 들쥐는 몇 마리인가? 200마리짜리 세 번째 표본에 표시된 들쥐가 정확히 j마리 들어 있을 확률을 써 보아라. 이 확률들을 모두 더하면 얼마가 되는가?

숨은 수학

최대가능도

1910년대와 1920년대에 로널드 피셔(Ronald Fisher)가 발전시킨 **최대가능도** 원리는 다음과 같다. 미지의 모수를, 관측된 자료가 나올 확률을 가장 크게 만드는 값으로 추정하라. 이것은 단연 가장 널리 쓰이는 추정 방법으로, 자료에 곡선을 맞추는 일부터 기계 학습의 신경망을 훈련하는 일까지 두루 쓰인다. 신경망의 "손실 함수"는 대개 음의 로그 가능도이다. 피셔는 모형이 대부분의 일상적인 모형에서 성립하는 의미로 충분히 잘 정의되어 있다면, 표본이 클 때 최대가능도 추정값이 근사적으로 정규분포를 따르며 어떤 합리적인 추정량보다도 뒤지지 않을 만큼 정밀하다는 것을 보였다. 중요한 예외도 있다. 다음 절(Section 13.3)에서는 미지수가 가능한 가장 큰 값, 즉 분포의 끝점인데, 이때 최대가능도 추정값은 사뭇 다르게 행동한다. 그 오차는 근사적으로 정규분포를 따르지 않으며, $\frac{1}{\sqrt{n}}$이 아니라 $\frac{1}{n}$처럼 줄어든다.

추정은 얼마나 정밀한가?

그림 13.2.1 (온라인 전체 절에 수록)을 보면 가능도의 폭이 넓다. 11마리의 재포획은 그리 많은 정보가 아니다. 대략 계산해 보자. 두 번째 표본에 들어 있는 표시된 들쥐의 수는 평균이 $\frac{200\times 200}{N}$인 푸아송 분포(Poisson distribution)를 근사적으로 따르므로, 그 표준편차는 약 $\sqrt{11} \approx 3.3$이고, 상대 오차는 약 30%이다. 이 오차는 N의 추정값에도 그대로 이어진다.

N을 10% 이내로 추정하려면 생물학자들은 약 100마리를 재포획해야 한다. 통계학자들은 단순한 추정값의 작은 편향(bias)도 보정하는데, 이때 **채프먼 추정량**(*Chapman estimator*) $\frac{(M+1)(C+1)}{R+1} - 1$을 쓴다. 여기서 M과 C는 두 번 잡은 개체의 수이고, R은 재포획된 개체의 수이다.

라플라스, 프랑스 인구를 세다

라플라스(Laplace)는 1802년에 같은 아이디어를 이용하여 인구 조사 없이 프랑스의 인구를 추정했다. 그는 한 해 동안 나라 전체에 등록된 출생 수를 알고 있었고, 표본으로 고른 몇몇 코뮌(commune)에 대해서는 인구와 출생 수를 모두 알고 있었다. 인구와 출생 수의 비가 어디에서나 같다고 가정하여, 그는 프랑스의 인구를 약 28백만 명으로 추정했고, 불확실성의 척도까지 제시했다. 이 방법은 1890년대에 덴마크의 피오르에서 가자미에 표시를 한 C. G. J. Petersen과, 1930년대에 새 가락지를 이용한 Frederick Lincoln에 의해 동물 개체 수 추정에 다시 발견되었다. 그래서 지금도 흔히 **링컨–피터슨**(*Lincoln–Petersen*) 추정량이라고 부른다.

더 도전하기

1. **일반 공식.** M마리의 동물에 표시를 하고, C마리로 이루어진 두 번째 표본에 표시된 동물이 R마리 들어 있다면, N의 최대가능도 추정값이 $\frac{MC}{R}$의 정수 부분임을 보여라.

2. **방데르몽드.** 여자 a명과 남자 b명 가운데에서 c명으로 이루어진 위원회를 뽑는 방법의 수를 세어 $\sum_j \binom{a}{j}\binom{b}{c-j} = \binom{a+b}{c}$를 증명하여라.

3. **동전 던지기.** 앞면이 나올 확률 p를 모르는 동전을 n번 던져 앞면이 h번 나왔다. p의 최대가능도 추정값이 $\frac{h}{n}$임을 보여라.

4. **재포획이 없는 경우.** 두 번째 표본에 표시된 들쥐가 한 마리도 없다면 최대가능도 추정값은 어떻게 되는가? 이것은 이 방법에 대해 무엇을 말해 주는가?

더 읽을거리

- **David Spiegelhalter, *The Art of Statistics*** (Pelican). 추정과 불확실성에 관한 책이다.
- **George Seber, *The Estimation of Animal Abundance*** (Blackburn). 포획–재포획에 관한 표준 참고서이다.
- **John Rice, *Mathematical Statistics and Data Analysis*** (Duxbury). 최대가능도를 포함한 수리통계학 입문 교재이다.

안내에 따른 전체 풀이는 제 13.2절에 있다:

https://beyondstep.keithmcnulty.org/ko/ch13/voles.html

매주 STEP 문제 하나를 끝까지 풀어 본다. The STEP Up은 이메일로 무료로 받아 볼 수 있다: keithmcnulty.substack.com