13.4 평균인가, 중앙값인가?

학습지

문제

STEP II 2000, 14번 문제

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

step.maths.org/questions/00-s2-q14

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** 중앙값의 확률밀도함수(density)가 왜 $y^n(1-y)^n$에 비례할까? (Y가 y 근처에 있으려면 $2n+1$개의 값 가운데 몇 개가 y보다 작아야 하고, 몇 개가 커야 할까?)

1단계 상수

다루는 내용: 상수 k

? **생각해 보기** 확률밀도함수(density)를 적분하면 1이다. 문제에 주어진 적분에서 두 거듭제곱을 모두 n으로 두고 사용한다.

2단계 평균과 분산

다루는 내용: 평균(mean)과 분산(variance)

생각해 보기 밀도함수는 $\frac{1}{2}$에 대하여 대칭(symmetric)인가? 그러면 $\mathrm{E}(Y)$를 얻는다.

잠깐, 직접 해 보기 $\mathrm{E}(Y^2)$를 구하려면 주어진 적분을 다시 사용하되 거듭제곱을 $n+2$와 n으로 두고, 계승(factorial)을 정리한다.

3단계 평균과 비교하기

다루는 내용: 중앙값(median)과 평균의 비교

생각해 보기 $\bar{X}$의 평균과 분산은 무엇인가? n이 클 때 두 분산을 비교한다.

잠깐, 직접 해 보기 Y와 $\bar{X}$를 모두 평균이 $\frac{1}{2}$인 정규분포(normal distribution)로 취급한다. 각 확률을 Φ를 써서 나타낸다. 어느 쪽이 더 크며, 그 이유는 무엇인가?

숨은 수학

효율

두 불편추정량(unbiased estimator)의 **상대 효율**(*relative efficiency*)은 두 분산의 비이다. 이는 한 추정량이 다른 추정량과 같은 정밀도를 얻으려면 관측값이 몇 배 필요한지를 말해 준다. 균등분포 데이터에서 중앙값의 평균에 대한 효율은 $\frac{1}{3}$이다. 정규분포 데이터에서는 $\frac{2}{\pi} \approx 0.64$로, 중앙값은 관측값이 약 57% 더 필요하다. 그러나 꼬리가 두꺼운(heavy-tailed) 데이터에서는 상황이 뒤바뀐다. 코시 분포(Cauchy distribution, Section 12.22)에서는 관측값 n개의 평균이 관측값 하나보다 나을 것이 없으므로 평균은 쓸모가 없지만, 중앙값은 잘 작동한다. 라플라스 분포(Laplace distribution, Section 12.24)에서는 중앙값이 평균보다 두 배 효율적이다.

로버스트 통계

중앙값은 **로버스트**(*robust*)하다. 데이터의 거의 절반을 임의의 값으로 바꾸어도 중앙값은 크게 움직이지 않는다. 평균에는 이런 보호 장치가 없어서 잘못된 값 하나만으로도 망가질 수 있다. 1960년대와 1970년대에 존 튜키(John Tukey), Peter Huber 등은 **로버스트 통계**(*robust statistics*)를 발전시켰다. 이는 깨끗한 데이터에서는 평균에 거의 버금가는 효율을 보이면서도 이상값(outlier)에 잘 견디는 추정량들로, 예를 들어 **절사평균**(*trimmed mean*)(가장 큰 값과 가장 작은 값 10%씩을 버리고 나머지를 평균한 것)이 있다. 절사평균은 다이빙이나 체조 같은 스포츠의 채점에 쓰인다.

라플라스와 중앙값

라플라스(Laplace)는 1774년에 오차가 이중지수분포(라플라스 분포)를 따르면 참값의 가장 좋은 추정값은 중앙값임을 보였다. 가우스(Gauss)는 1809년에 오차가 정규분포를 따르면 가장 좋은 추정값은 평균임을 보였고, 이로부터 최소제곱법(method of least squares)에 이르렀다. 어느 쪽이 옳은지는 오차가 어떻게 행동하는지에 달려 있으며, 통계학자들은 새로운 종류의 데이터가 나올 때마다 여전히 이 문제를 논쟁한다.

더 도전하기

1 **중앙값의 밀도함수.** 값 하나는 $[y, y+\delta y]$에, n개는 그 아래에, n개는 그 위에 있는 경우를 세어 밀도함수 $\frac{(2n+1)!}{(n!)^2} y^n (1-y)^n$을 유도하라.

2 **관측값 세 개.** $n = 1$(균등분포 관측값 세 개)일 때 $\mathrm{Var}(Y)$와 $\mathrm{Var}(\tilde{X})$를 정확히 구하라. 어느 쪽이 더 작은가?

3 **중간범위.** 중심의 또 다른 추정량으로 중간범위(midrange) $\frac{1}{2}(\max + \min)$가 있다. 관측값이 $m = 2n+1$개인 균등분포 데이터에서 그 분산이 $\frac{1}{2(m+1)(m+2)}$임을 보여라. 평균과 비교하면 어떠한가? (Section 13.3와 비교해 보라.)

4 **이상값.** 측정값 열한 개가 9.8, 9.9, 10.0, 10.0, 10.1, 10.1, 10.2, 10.2, 10.3, 10.4, 104(입력 실수)이다. 평균과 중앙값을 비교하라.

더 읽을거리

- **John Rice, *Mathematical Statistics and Data Analysis*** (Duxbury). 순서통계량(order statistics)과 효율을 다룬다.
- **Rand Wilcox, *Fundamentals of Modern Statistical Methods*** (Springer). 로버스트 방법이 왜 중요한지를 비전문가에게 설명한다.

- **Stephen Stigler, *The Seven Pillars of Statistical Wisdom*** (Harvard University Press). 자료의 집계, 평균과 중앙값을 역사적 관점에서 살펴본다.

안내에 따른 전체 풀이는 제 13.4절에 있다:

https://beyondstep.keithmcnulty.org/ko/ch13/mean-or-median.html

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