사회통계
산포경향 측도

김현우, PhD[1]

[1] 충북대학교 사회학과 조교수

March 13, 2024

진행 순서

1. 데이터의 요약(II): 산포경향
2. 표준편차의 해석

데이터의 요약(II): 산포경향

데이터의 요약(II): 산포경향

과연 중심경향으로만 충분히 자료를 잘 요약할 수 있을까?

- 다음의 예제를 통해 중심경향의 근본적인 한계를 살펴보자.
- 다음의 두 데이터에서 평균(mean), 중앙값(median), 최빈값(mode)을 각각 계산해보자.

$$D_1 = \{-10, 0, 10\}$$
$$D_2 = \{-100, 0, 100\}$$

- 두 데이터는 어떻게 다른가? 중심경향에 근거하여 두 데이터가 잘 요약되었나?

데이터의 요약(II): 산포경향

관측치들이 얼마나 흩어져 있는가를 측정하는 통계량이 필요하다.

- 가장 간단한 산포경향(dispersion tendency) 통계량은 범위(range)이다.

$$\text{범위} = \text{최대값(maximum)} - \text{최소값(minimum)}$$

- 일상에서도 꽤 많이 쓰인다(e.g., 주가의 일일등락폭, 하루 온도의 일교차).
- 이제 중심경향을 보완할 수 있다! 다시 다음의 자료로부터 범위를 구해보자.

$$D_1 = \{-10, 0, 10\}$$
$$D_2 = \{-100, 0, 100\}$$

- 그러나 범위는 극단치에 민감하다는 결함을 갖고 있다. 다음의 자료로부터 범위를 구해보자.

$$D = \{0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 100\}$$

데이터의 요약(II): 산포경향

이 문제에 대한 임시대응책은 이른바 사분위수간 범위를 사용하는 것이다.

- 이에 앞서 분위수(quantile)라는 용어에 친숙해져야 한다.
- 분위수와 사분위수(quartile)의 영어 단어는 살짝 다르다.
- p번째 백분위수(pth percentile): "이 값보다 작은 값들이 관측치들의 p%이고, 이 값보다 큰 값들이 $(100-p)$%인 값."
- 가령 "유림의 점수는 80 퍼센타일이야" 라고 말할 때, 그의 점수보다 낮은 점수들이 80%이고 높은 점수들이 20% 라는 의미(즉 80 퍼센타일은 상위 20%)이다.
- 당연히 백분위수의 범위는 0~100% 사이에 놓인다.

데이터의 요약(II): 산포경향

- 사분위수는 다음 4개의 분위수를 지칭한다.

  (1) 첫번째 사분위수(Q_1; 25번째 백분위수)
  (2) 두번째 사분위수(Q_2; 50번째 백분위수; median)
  (3) 세번째 사분위수(Q_3; 75번째 백분위수)
  (4) 네번째 사분위수(Q_4; 100번째 백분위수)

- (다시 돌아와서) 사분위수간 범위(interquartile range; IQR)는 세번째 사분위수(Q_3)와 첫번째 사분위수(Q_1) 간의 차이를 의미한다.

$$\text{IQR} = Q_3 - Q_1$$

- 따라서 관측치들의 중간 50%가 흩어져 있는 정도만 측정한다!
- IQR이 큰 값을 가진다는 것은 첫번째 사분위수(Q_1)와 세번째 사분위수(Q_3)가 멀리 떨어져 있어 자료의 변동성(variation)이 크다는 것을 의미한다.

데이터의 요약(II): 산포경향

예제 4. airpollution.zip의 2023년 3월.xlsx에서 [데이터]-[필터]를 사용해 다음의 조건을 특정하시오: 〈지역〉은 충북 청주시, 〈망〉은 도시대기, 측정일시는 〈2023031114〉. 초미세먼지(PM2.5)의 사분위수를 구하시오.

데이터의 요약(II): 산포경향

- 필터링된 자료를 복사하여 다른 새 탭에 붙여넣고 작업하자.
- PM2.5의 범위와 사분위간 범위를 구하기 위해 다음의 세 가지 함수를 사용한다: MAX(·), MIN(·), QUARTILE(·).
- 범위는 정의상 =MAX(·)-MIN(·) 로 계산한다.
- 사분위간 범위(IQR)는 정의상 =QUARTILE(·, 3)-QUARTILE(·, 1) 로 계산한다.

데이터의 요약(II): 산포경향

범위나 IQR 모두 정보의 손실이 심하다.

- IQR은 "첫번째 사분위수와 네번째 사분위수에 극단치가 있을 수 있다"라는 전제에 입각해 있다.
- 크고 작은 두 값(최대–최소 또는 $Q_3 - Q_1$)만 사용하고 나머지는 모두 버리므로 낭비가 심하다!
- 말하자면 통계적으로 비효율적(inefficient)이다.
- IQR로도 여전히 범위가 가진 근본적인 결함은 해소되지 않는다.

데이터의 요약(II): 산포경향

분산은 범위나 IQR보다 훨씬 정보를 효율적으로 활용한다.

- 분산(variance)을 계산하기 위해 모든 관측치에 대해 편차(deviation)를 먼저 계산한다.

$$(x_i - \mu)$$

- 편차가 크다면 평균에서 개별값들이 많이 이탈해 있다는 의미이므로 변동성이 높다고 할 수 있다!
- (IQR과는 달리) 분산은 모든 측정치들에 대해 편차를 전부 구하므로 "모든 관측치를 하나도 버리지 않고" 산포경향을 측정한다.
- 그렇다고 모든 관측치에 대해 구해진 이 편차들을 그냥 더하면 곤란하다(Why?). 이 문제를 피하기 위해 편차들을 제곱해서 더한다.

$$\sum_{i=1}^{n}(x_i - \mu)^2$$

데이터의 요약(II): 산포경향

- 편차의 제곱합을 모집단의 크기(N)로 나누어 모분산(population variance)을 구할 수 있다.

$$Var(X) \equiv \sigma^2 = \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_N - \mu)^2}{N}$$

$$= \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

- 표본에 대해서도 똑같은 발상을 적용하여 표본분산(sample variance)을 계산할 수 있다.

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2$$

데이터의 요약(II): 산포경향

- 분산의 단점 하나는 편차의 제곱합을 하는 과정에서 그 값이 과장스럽게 커진다는 점이다. 가령 원래 단위가 cm였다면 분산의 단위는 cm square가 된다(Why?).
- 수학에서 제곱(square)의 반대는 제곱근(square root)이다. 예컨대, $\sqrt{7^2} = 7$ 이다.
- 그러므로 구해진 분산에 대해서도 제곱근을 취하면 아까 커진 부분이 도로 작아지게 된다. 다시 말해, "분산의 제곱근"을 일컫어 표준편차(standard deviation)라고 부른다.
- (분산과 마찬가지로) 모표준편차(population standard deviation)와 표본표준편차(sample standard deviation)을 구분할 수 있다.

$$\sigma = \sqrt{\sigma^2} = \sqrt{Var(X)}$$
$$S = \sqrt{S^2}$$

데이터의 요약(II): 산포경향

예제 5. hsb2.csv와 hsb2.pdf를 참고하여 사회 점수(socst)의 모분산, 모표준편차, 표본분산, 표본표준편차를 각각 구하시오.

데이터의 요약(II): 산포경향

- 먼저 socst의 평균을 구하자. 평균을 구하는 함수는 AVERAGE(·)이다.
- 마우스로 드래그할 때, $을 활용하면 편리하게 셀 참조(cell reference)를 고정시킬 수 있다.
- 개별 socst의 관측치에서 평균을 빼고 그것들의 제곱을 구한다(괄호에 주의!)
- 편차의 제곱들을 모두 합한 뒤, 표본 크기(N)로 나누어준다. 이것이 모분산이다.
- 모분산에 제곱근하여 모표준편차를 구하자. 제곱근의 함수는 SQRT(·)이다.
- 엑셀 함수로 모분산과 모표준편차를 구하기 위해 각각 VAR.P(·)과 STDEV.P(·)를 사용할 수 있다. 답이 일치하는지 확인해보자.
- 표본분산과 표본표준편차는 각각 VAR.S(·)과 STDEV.S(·)로 구할 수 있다.

데이터의 요약(II): 산포경향

불확실한 수익이라는 측면에서 본다면 평균은 크고 분산은 작은 것이 좋다.

- 그러나 현실에서는 대체로 평균이 크면 분산도 크고(high risk, high return), 평균이 작으면 분산도 작다(low risk, low return).
- 노벨상 수상자 William Sharpe는 변동성 대비 보상을 측정하는 샤프 지수(Sharpe Ratio)를 개발하기도 했다. 어떤 투자의 샤프지수가 높을수록 투자자는 위험에 대해 더 잘 보상받는다.
- 결국 변동계수(coefficient of variation)가 작은 투자상품이 좋다(교재 67-68).

표준편차의 해석

표준편차의 해석

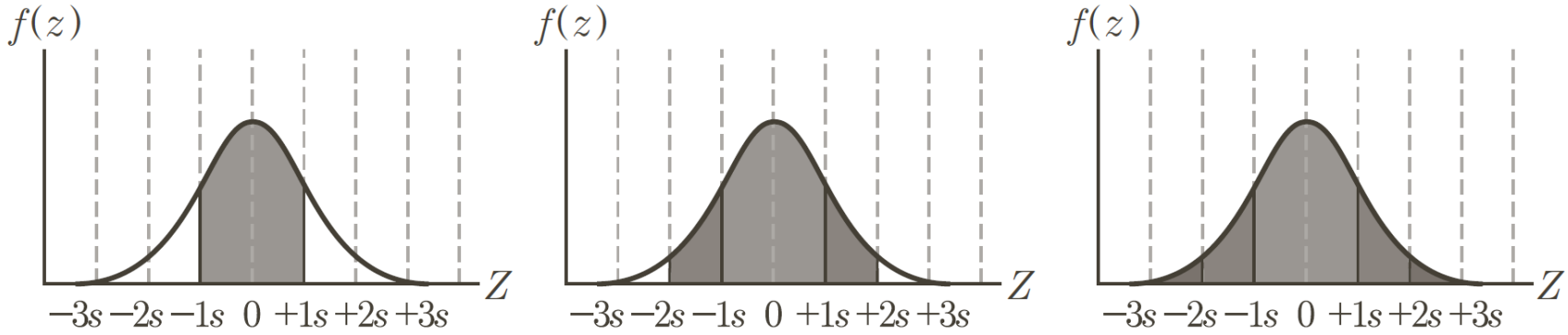
표준편차의 해석에 대해 곰곰히 고민해 보아야 한다.

- 표준편차가 크다는 것은 자료 안에 관측치들이 여기저기 흩어져 있다는 것을 뜻한다.
- 표준편차의 구체적인 값은 어떻게 해석될 수 있을까?
- 한 가지 방법은 (1 표준편차, 2 표준편차, 3 표준편차에 특별한 의미를 부여하는) 이른바 경험법칙(empirical rule)을 활용하는 것이다.

표준편차의 해석

자료가 정규분포한다면 아래의 세 가지 경험법칙이 성립한다.

- 모든 관측치의 약 68%는 평균으로부터 1 표준편차 안에 속한다(왼쪽).
- 모든 관측치의 약 95%는 평균으로부터 2 표준편차 이내에 속한다(중앙).
- 모든 관측치의 약 99.7%는 평균으로부터 3 표준편차 이내에 속한다(오른쪽).



표준편차의 해석

자료가 정규분포하지 않으면 이 경험법칙이 성립하지 않을 수도 있다.

- 자료가 정규분포하지 않더라도 다음의 체비쇼프 부등식(Chebyshev's inequality)은 성립한다.

$$P(\mu - k\sigma \leq X \leq \mu + k\sigma) = 1 - \frac{1}{k^2}$$

- 가령 $k = 2$를 계산해보고 이를 정규분포 경험법칙과 비교해보자.
- 체비쇼프 정리는 경험법칙의 하한값(lower bound)을 보여준다(Why?).

표준편차의 해석

예제 6. hsb2.csv를 사용하여 과학 점수가 정규분포한다는 가정 아래 모든 관측치의 95%는 몇 점과 몇 점 사이에 있는지 보고하시오.

표준편차의 해석

- 자료의 평균 μ과 표준편차 σ를 계산한 뒤, 아래 공식을 사용한다.

$$P\left(\mu - 2\sigma \leq X \leq \mu + 2\sigma\right) = 0.95$$

- $\mu \pm 2\sigma$은 각각 32.05와 71.65이다.
- "(과학 점수가 정규분포한다면) 200명의 학생 중 95% (약 190명)의 학생들은 32.05점과 71.65점 사이의 과학 점수를 받았다고 볼 수 있다."

표준편차의 해석

예제 7. hsb2.csv를 사용하되 이번엔 과학 점수가 정규분포한다는 가정 없이 모든 관측치의 75%는 몇 점과 몇 점 사이에 있는지 보고하시오.

표준편차의 해석

- 아래 공식에서 k를 구해야 한다(Why?).

$$P\left(\mu - k\sigma \leq X \leq \mu + k\sigma\right) = 1 - \frac{1}{k^2} = 0.75$$

- $k = 2$임을 알 수 있다. STDEV.S$(\cdot)$를 사용해보자.
- "(과학 점수가 정규분포하지 않는다면) 200명의 학생 중 적어도 75% (약 150명)의 학생들은 32.05점과 71.65점 사이의 과학 점수를 받았다고 볼 수 있다."
- 정규분포 가정이 없다면 우리는 훨씬 소극적으로 추정하게 된다(Why?).

2주차 과제

문제 2번, 3번, 4번
단 2(3), 3(3) 제외