

사회통계연습

Measures of Dispersion Tendency

김현우, PhD¹

¹충북대학교 사회학과 조교수

September 10, 2021

- 1 데이터의 요약(II): 산포성향(Dispersion Tendency)
- 2 데이터 요약 연습

데이터의 요약(II): 산포성향(Dispersion Tendency)

데이터의 요약(II): 산포성향(Dispersion Tendency)

앞서 연습한 중심성향으로만 충분히 자료를 잘 요약할 수 있었을까? 사실은 그렇지 않다.

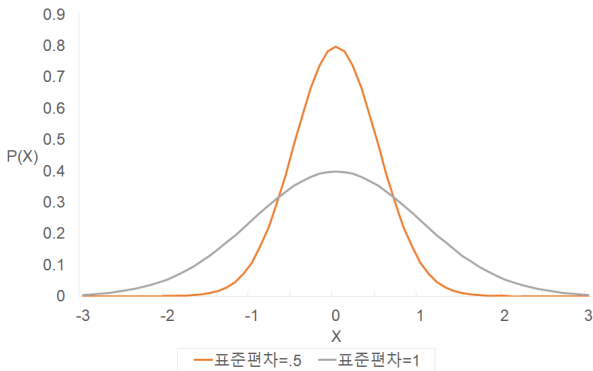
두 예제를 통해 중심성향의 근본적인 한계를 살펴보자.

- 다음의 데이터가 주어졌다: [-10, 0, 0, 10]
- 평균(mean), 중위수(median), 최빈치(mode)는 각각 얼마인가?
- 다음의 데이터가 주어졌다: [-100, 0, 0, 100]
- 평균(mean), 중위수(median), 최빈치(mode)는 각각 얼마인가?
- 두 데이터는 어떻게 다른가?
- 중심성향에 근거하여 두 데이터가 잘 요약되었나?

데이터의 요약(II): 산포성향(Dispersion Tendency)

다음 두 개의 분포는 0을 기준으로 모두 대칭이다.

- Find the means, medians, and modes.
- 두 그림의 차이는 세 통계에 의해 잘 설명되는가? 아니면 어떤 차이가 있나?



두 개의 상이한 분포들

데이터의 요약(II): 산포성향(Dispersion Tendency)

그러므로 관측치(observations)들이 얼마나 흩어져 있는가(산포; dispersion)를 측정하는 통계가 필요하다.

가장 간단한 산포성향의 통계는 범위(range)다.

- 최댓값(maximum) - 최솟값(minimum)
- e.g., 주가의 일일등락폭, 하루 온도의 일교차
- 다음의 데이터를 보라: $[-10, 0, 10]$. Find the range.

그러나 간단한 만큼 range는 몇 가지 치명적인 결함을 갖고 있다.

- 다음의 데이터를 보라: [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 100]. Find the range.
- 평균보다 더 outlier에 민감하다.

데이터의 요약(II): 산포성향(Dispersion Tendency)

이 문제에 대한 임시대응책은 이른바 사분위수간 범위(interquartile range; IQR)를 사용하는 것이다.

- 이에 앞서 **분위수(quantile)**라는 용어에 친숙해져야 한다.
- p번째 백분위수(p-th percentile): “이 값보다 작은 값들이 관측치들의 p%이고, 이 값보다 큰 값들이 (100-p)%인 값.”
- e.g., “정휴의 점수는 80 퍼센타일이야” 라고 말할 때, 그의 점수보다 낮은 점수들이 80%이고 높은 점수들이 20% 라는 의미. 즉 80 퍼센타일은 상위 20%.
- 당연히 그 범위는 0~100% percentile.

데이터의 요약(II): 산포성향(Dispersion Tendency)

사분위수(quartile)는 25번째, 50번째, 75번째, 100번째 백분위수(percentile)로 총 네 개의 분위수이다.

- 첫번째 사분위수(Q1), 25th percentile
- 두번째 사분위수(Q2), 50th percentile (=median)
- 세번째 사분위수(Q3), 75th percentile
- 네번째 사분위수(Q4), 100th percentile

(다시 돌아와서) IQR은 세번째 사분위수와 첫번째 사분위수 간의 차이를 의미한다.

- 따라서 관측치들의 중간 50%가 흩어져 있는 정도를 측정한다!
- IQR이 큰 값을 가진다는 것은 첫 번째 사분위수(Q1)와 세 번째 사분위수(Q3)가 멀리 떨어져 있어 변동성(variation)이 크다는 것을 의미.

데이터의 요약(II): 산포성향(Dispersion Tendency)

다시 airpollution.zip의 2020년 9월.xlsx를 열자.

[데이터]-[필터]를 사용해 다음의 조건을 특정하자: <지역>은 충남 대전시, <망>은 도시대기, 측정일시는 <2020.09.10 전체>.

필터링된 자료를 복사하여 다른 새 탭에 붙여넣고 작업하자.

- PM2.5의 range와 IQR를 구하기 위해 다음을 연습한다. $=\max()$, $=\min()$, $=\text{quartile}()$.
- range는 정의상 $=\max(:) - \min(:)$ 로 계산된다.
- IQR은 정의상 $=\text{quartile}(:, 3) - \text{quartile}(:, 1)$ 로 계산된다.

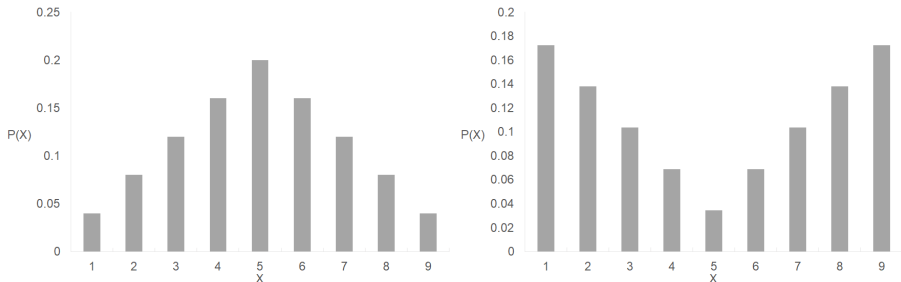
데이터의 요약(II): 산포성향(Dispersion Tendency)

IQR은 “첫번째 사분위수와 네번째 사분위수에 극단치가 있을 수 있다”라는 전제에 입각해 있다.

- 그러나 이런 방식으로든 여전히 range의 근본적인 결함은 해소되지 않는다.
- 문제의 본질에는 range나 IQR 모두 정보의 낭비가 심하다는 사실이 있다.
- 다시 말해, 크고 작은 두 값(최대/최소 또는 $Q3/Q1$)만 사용하고 나머지는 모두 버리기 때문이다.
- 말하자면 통계적으로 비효율적(inefficient)이다.

데이터의 요약(II): 산포성향(Dispersion Tendency)

그러다보니 range나 IQR로는 “범위(range)는 같으나 분포(distribution)가 다른 경우”를 제대로 요약하지 못한다.



범위는 같으나 분포의 형태가 상이한 두 개의 분포들

데이터의 요약(II): 산포성향(Dispersion Tendency)

보다 우월한 방식은 “모든 관측치(observations)를 하나도 버리지 않고” 활용하면서 변동성(variation)을 측정하는 것이다.

분산(variance)은 range나 IQR보다 훨씬 “정보를 효율적으로” 활용한다.

- 분산을 계산하기 위해 먼저 모든 관측치에 대해 편차(deviation)를 계산한다:
 $(x_i - \mu)$
- 편차가 크다면 평균에서 개별값들이 많이 이탈해 있다는 의미이므로 변동성이 높다고 할 수 있다!
- 모든 관측치에 대해 구해진 이 편차들을 그냥 더하면 곤란하다. 왜죠? “이 문제”를 피하기 위해 편차들을 제곱해서 더한다: $\sum_{i=1}^n (x_i - \mu)^2$
- 관측치들의 숫자(n)로 나누어준다.

$$\begin{aligned}\text{Var}(X) = \sigma^2 &= \frac{(x_1 - \mu)^2 + (x_2 - \mu)^2 + \cdots + (x_n - \mu)^2}{n} \\ &= \frac{1}{n} \sum_{i=1}^n (x_i - \mu)^2\end{aligned}$$

데이터의 요약(II): 산포성향(Dispersion Tendency)

분산은 평균처럼 수학적으로 우아하다. 하지만...

- 편차에 대해 제곱을 하는 과정에서 편차가 지나치게 과장스럽게 커진다!
- 수학에서 제곱의 반대는 무엇인가? 제곱근(square root)이다. 예컨대, $\sqrt{7^2} = 7$ 이다.
- 그러므로 구해진 분산에 대해서도 제곱근을 취하면 아까 커진 부분이 도로 작아지게 된다. 이것을 표준편차(standard deviation)라고 부른다.

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

데이터의 요약(II): 산포성향(Dispersion Tendency)

데이터의 요약(II): 산포성향(Dispersion Tendency)

조금 다른 이야기지만, 불확실한 수익이라는 측면에서 보면 평균은 크고 분산은 작은 것이 좋다.

- 그러나 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 지수(Sharpe Ratio)를 개발하기도 했다. 어떤 투자의 샤프지수가 높을수록 그 투자는 투자자에게 위험에 대해 더 잘 보상한다.

데이터 요약 연습

데이터 요약 연습

eCampus에서 queue.csv를 다운받자. 이것은 신영과 재현이 운영하는 두 패스트푸드점의 주문 후 음식 수령까지 대기시간을 기록한 자료의 샘플이다.

- 신영이 운영하는 1번 레스토랑의 평균과 표준편차를 구하시오.
- 재현이 운영하는 2번 레스토랑의 평균과 표준편차를 구하시오.
- 대기시간 외에 모든 요소가 같다고 가정한다면, 손님으로서 당신은 신영이 운영하는 1번 레스토랑과 재현이 운영하는 2번 레스토랑 사이 어느 쪽을 이용할 것인가? 왜죠?

데이터 요약 연습

eCampus에서 서울시아파트관리비.xlsx 을 다운받자. 이것은 서울시 두 개 자치구에서 임의로 선별된 아파트단지에서 수집된 관리비 자료이다.

- 중구와 중랑구에서 총 관리비 합계 평균, 중위값, 최빈치를 각각 구하시오. 어느 쪽이 높은가?
- 중구와 중랑구에서 총 관리비 합계의 range와 IQR을 각각 구하시오. 어느 쪽의 변동이 심한가?
- 앞서 구한 Range와 IQR은 각 자치구에서 총 관리비 합계를 잘 요약하고 있는가? 아니라면 왜죠?
- 중구와 중랑구에서 세대당 관리비 합계, 즉 총 관리비 합계/세대 수로 나눈 값을 새로 계산하고 이 값의 표준편차를 각각 구하시오. 어느 쪽의 변동이 심한가?

데이터 요약 연습

다시 airpollution.zip에서 2020년 6월.xlsx를 열자. 필터링을 적용하여 다음의 조건을 맞춘다: <지역>은 경기 군포시, <망>은 도시대기, <측정소명>은 산본동, 측정일시는 <2020.06.15 전체>. 필터링 된 내용을 새 탭에 복사하고 그 탭에서 작업하자.

stdev.p와 var.p를 사용하지 않고 엑셀에서 공식대로 계산하여 PM2.5의 표준편차를 구하시오.

끝!

와~ 또 숙제다~