2023학년도 2학기 언어데이터과학

제5강 데이터과학 (1)

박수민

서울대학교 인문대학 언어학과

2023년 9월 18일 월요일

오늘의 목표

1. 데이터의 정의를 외울 수 있다.
2. 언어데이터(Linguistic data)의 범위를 설명할 수 있다.
3. 데이터의 네 가지 범주를 설명할 수 있다.
4. 데이터를 중심 경향성, 산포도, 상관관계로 기술할 수 있다.

교수자의 고민

[[언어][데이터]][과학]인가, [언어][[데이터][과학]]인가?

What is data?

Data

"entities used as evidence of phenomena for the purposes of research or scholarship"

— C. Borgman. (2015). Big Data, Little Data, No Data: Scholarship in the Networked World.

Two approaches to the analysis of language

Descriptive vs. Prescriptive

- Description ⇒ Collection of linguistic data that are directly observable

Two data types from observable linguistic behavior

Functionalist vs. Formalist

Functionalist Naturalistic instances of language in use

Formalist Constructed examples of language that can serve as prompts to collect grammaticality judgments from users of a language

Two classes of linguistic data

Naturalistic data

- Documentary linguistic data
- Corpora(말뭉치)

Specialized data

- Grammaticality judgments

데이터의 유형

- 범주형(Categorical)
  - 명목형(Nominal)
  - 순서형(Ordinal)
- 수치형(Numerical)
  - 이산형(Discrete)
  - 연속형 (Continuous)

예시: 언어학 데이터

- 범주형
  - 명목형: '밀덕'(밀리터리 덕후)을 어떻게 발음합니까?
    - [밀덕], [밀떡]
  - 순서형: 이 문장이 자연스럽습니까?
    - 매우 어색함/어색함/보통/자연스러움/매우…
- 수치형
  - 이산형: 각 단어에 장애음이 몇 개 있는가?
    - 0, 1, 2, 3, …
  - 연속형: 어두 자음의 VOT가 몇 ms인가?
    - -14.15, 3.60, 23.61, -7.42,…

데이터를 기술하는 방법

- 통계량
  - 중심 경향성: 평균, 중앙값, 최빈값 – 어디에 몰려 있는가?
  - 산포도: 표준편차, 사분위수 — 얼마나 흩어져 있는가?
  - 상관관계
- 시각화
  - 히스토그램: 한 가지 데이터의 분포
  - 산점도: 두 가지 데이터의 관계

중심 경향성

평균

- 계산이 간편하다.
- 데이터의 변화에 따라 변한다. $\Rightarrow$ 이상치에 민감하다.

중앙값

- 이상치가 포함되어도 큰 영향을 받지 않는다.
- 데이터를 크기순으로 정렬해야 한다. $\Rightarrow$ 계산량이 많아진다.

“퍼짐 경향성”

편차

(편차) = (관측치) − (평균)

- 편차의 합은 항상 0이다.

⇒ 데이터가 얼마나 퍼져 있는지를 반영할 수 없다.

⇒ 편차의 “크기”를 사용해야 한다.

표준편차

편차의 크기를 측정하는 방법

1. 절댓값을 취해서 더한다. ⇒ 미분을 할 수 없다.
2. 제곱을 해서 더한다. ⇒ 채택
   ⇒ 다 더한 뒤 제곱근을 취하여 원래의 값과 같은 1차로 만든다.

두 변수 사이의 관계

공분산

두 변수가 각각의 평균에서 얼마나 떨어져 있는지를 측정하는 통계량

cf. 분산: 하나의 변수가 평균에서 얼마나 떨어져 있는가?

편차를 곱해서 더한다. $\rightarrow$ 상관관계에 따라 음수가 될 수 있다.

(비편향)공분산

$$\frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}, \quad \bar{x} = (x\text{의 평균}), \bar{y} = (y\text{의 평균})$$

편향공분산: $(n-1)$ 대신 n으로 나눈 것

두 변수 사이의 관계

상관관계

$$\frac{(x\text{와 } y\text{의 공분산})}{(x\text{의 표준편차}) \times (y\text{의 표준편차})}$$

값의 범위 단위에 상관 없이 항상 -1에서 1 사이의 값을 가진다.

양의 상관관계 x가 증가할 때 y도 증가한다.

음의 상관관계 x가 증가할 때 y는 감소한다.

상관관계의 주의사항

상관관계 이외의 관계

상관관계가 0인 두 변수

```
x = [-2, -1, 0, 1, 2]
y = [2, 1, 0, 1, 2]
```

상관관계 $\neq$ 연관성

상관관계가 1인 두 변수

```
x = [-2, -1, 0, 1, 2]
y = [99.98, 99.99, 100, 100.01, 100.02]
```

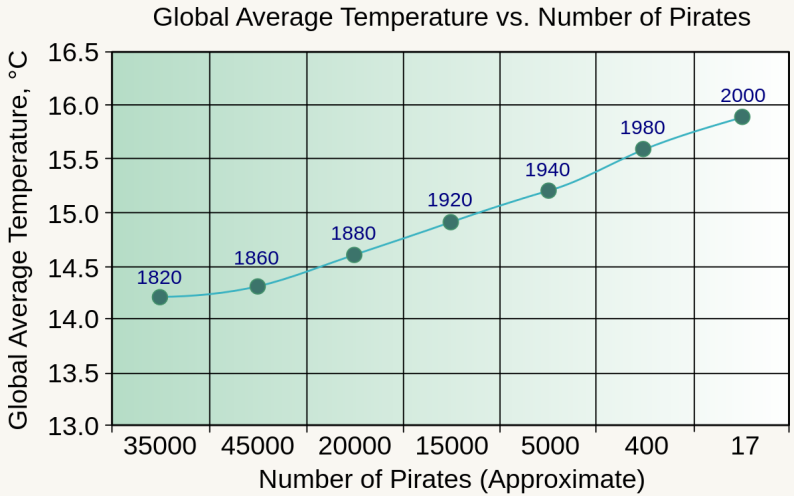
상관관계의 주의사항

상관관계 $\neq$ 인과관계

접속 시간과 친구 수 사이에 상관관계가 존재하는데...

- 친구가 많기 때문에 접속 시간이 늘어났는가?
- 오래 접속하다 보니 친구가 늘어났는가?
- ...이도 저도 아닌 우연인가?

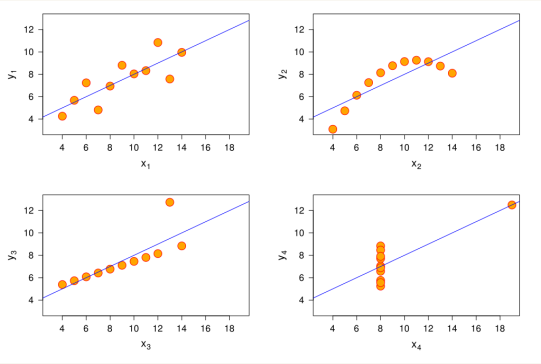
상관관계의 주의사항



https://commons.wikimedia.org/wiki/File:PiratesVsTemp(en).svg 가공

Anscombe's quartet

같은 평균, 같은 분산, 같은 상관관계



https://commons.wikimedia.org/wiki/File:Anscombe.svg

기술통계량

x의 평균	9
x의 분산	11
y의 평균	7.5
y의 분산	4.125
x와 y의 상관관계	0.816

네 데이터셋의 기술통계량은 같지만
실제 분포는 모두 다르다!

Simpson's paradox

전체의 대소 관계와 부분의 대소 관계가 달라지는 현상

전체 친구 수: 서부 > 동부

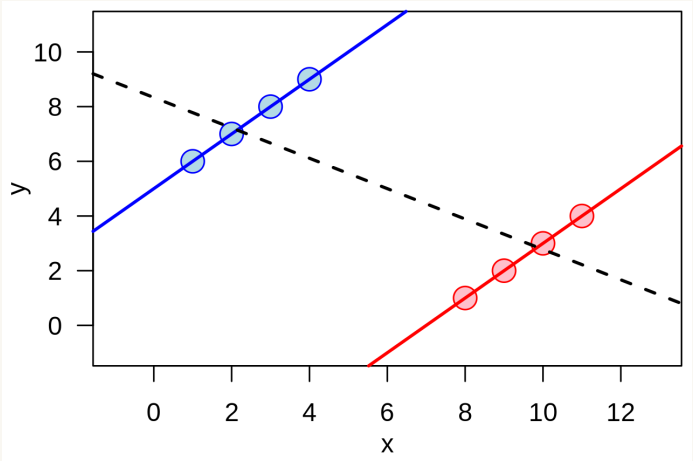
지역	사용자 수	평균 친구 수
서부	101	8.2
동부	103	6.5

학위별 친구 수: 서부 < 동부

지역	학위	사용자 수	평균 친구 수
서부	박사	35	3.1
동부	박사	70	3.2
서부	기타	66	10.9
동부	기타	33	13.4

Simpson’s paradox

전체의 대소 관계와 부분의 대소 관계가 달라지는 현상



교수자의 헛된 고민

Linguistics AND Data Science

이번 시간에 배운 것

1. 데이터의 정의와 범주
2. 언어데이터의 범위
3. 데이터의 기술통계량
4. 기술통계량의 한계

다음 시간에 배울 것

1. numpy 활용법