

# 제 2장 데이터의 이해

## 2.1 표본과 데이터

## 2.2 데이터의 수집방법

### 2.2.1 실험

### 2.2.2 관찰(observe)

### 2.2.3 조사(survey)

## 2.3 데이터의 구분

### 2.3.1 속성에 따른 구분

### 2.3.2 형태에 따른 구분

|    | A          | B                | C     | D          | E    | F    | G    | H    | I   | J    | K   | L    |
|----|------------|------------------|-------|------------|------|------|------|------|-----|------|-----|------|
| 1  | Company    | Plant/Line       | Metal | Area       | Rate | Area | 0-20 | 21   | 22  | 23   | 24  | 25   |
| 2  | All Sakers | Lowest co-Silver | 3     | Colorado3  | 109  | 1286 | 297  | 8404 | 297 | 8404 | 297 | 8404 |
| 3  | All Sakers | Lowest co-Silver | 7     | Colorado7  | 196  | 4585 | 293  | 6229 | 293 | 6229 | 293 | 6229 |
| 4  | All Sakers | Lowest co-Silver | 8     | Colorado8  | 196  | 1095 | 309  | 8435 | 309 | 8435 | 309 | 8435 |
| 5  | All Sakers | Lowest co-Silver | 9     | Colorado9  | 245  | 7819 | 387  | 0579 | 387 | 0579 | 387 | 0579 |
| 6  | All Sakers | Lowest co-Silver | 10    | Colorado10 | 242  | 6912 | 362  | 1969 | 362 | 1969 | 362 | 1969 |
| 7  | Cigna      | Lowest co-Silver | 3     | Colorado3  | 158  | 0892 | 248  | 9533 | 248 | 9533 | 248 | 9533 |
| 8  | Colorado   | Lowest co-Silver | 2     | Colorado2  | 131  | 2135 | 206  | 6354 | 206 | 6354 | 206 | 6354 |
| 9  | Colorado   | Lowest co-Silver | 3     | Colorado3  | 146  | 2935 | 230  | 3834 | 230 | 3834 | 230 | 3834 |
| 10 | Colorado   | Lowest co-Silver | 4     | Colorado4  | 177  | 9689 | 282  | 265  | 282 | 265  | 282 | 265  |
| 11 | Colorado   | Lowest co-Silver | 6     | Colorado6  | 190  | 9649 | 285  | 0155 | 285 | 0155 | 285 | 0155 |
| 12 | Colorado   | Lowest co-Silver | 8     | Colorado8  | 153  | 82   | 237  | 5117 | 237 | 5117 | 237 | 5117 |
| 13 | Colorado   | Lowest co-Silver | 9     | Colorado9  | 177  | 9689 | 282  | 265  | 282 | 265  | 282 | 265  |
| 14 | COOP       | Lowest co-Silver | 1     | Colorado1  | 132  | 1425 | 208  | 1664 | 208 | 1664 | 208 | 1664 |
| 15 | COOP       | Lowest co-Silver | 3     | Colorado3  | 135  | 5909 | 213  | 529  | 213 | 529  | 213 | 529  |
| 16 | COOP       | Lowest co-Silver | 4     | Colorado4  | 148  | 962  | 234  | 4514 | 234 | 4514 | 234 | 4514 |
| 17 | COOP       | Lowest co-Silver | 6     | Colorado6  | 143  | 9383 | 226  | 6745 | 226 | 6745 | 226 | 6745 |
| 18 | COOP       | Lowest co-Silver | 9     | Colorado9  | 177  | 4174 | 279  | 3975 | 279 | 3975 | 279 | 3975 |
| 19 | COOP       | Lowest co-Silver | 10    | Colorado10 | 190  | 8379 | 306  | 5321 | 306 | 5321 | 306 | 5321 |
| 20 | COOP       | Lowest co-Silver | 15    | Colorado15 | 229  | 5643 | 351  | 4241 | 351 | 4241 | 351 | 4241 |
| 21 | COOP       | Lowest co-Silver | 1     | Colorado1  | 132  | 0919 | 239  | 5149 | 239 | 5149 | 239 | 5149 |
| 22 | COOP       | Lowest co-Silver | 2     | Colorado2  | 153  | 2844 | 241  | 3929 | 241 | 3929 | 241 | 3929 |
| 23 | COOP       | Lowest co-Silver | 3     | Colorado3  | 156  | 662  | 246  | 7512 | 246 | 7512 | 246 | 7512 |
| 24 | COOP       | Lowest co-Silver | 4     | Colorado4  | 171  | 3585 | 269  | 8435 | 269 | 8435 | 269 | 8435 |

## 2.1 표본과 데이터

- 표본 (sample) : 모집단의 부분집합  
표본은 그 자체로 데이터일 수 있고,  
데이터 값을 가지지 않을 때도 있다.

예) A 고등학교 학생의 신장을 알아보기 위해 50명을 랜덤하게 선발하는 경우 랜덤하게 선발된 학생은 표본이지만 그 자체가 데이터 값은 아님. 선발된 학생을 대상으로 키를 측정하여야 데이터 값을 얻게 됨.

예) 주사위를 던질 때 나타나는 표본(눈의 수) 그 자체는 데이터 임.

## 2.2 데이터의 수집방법

- 우리가 통계학의 영역에서 다루고자 하는 것은 모집단의 속성을 알기 위하여 적절한 방법으로 표본을 얻지만 궁극적으로는 이 표본에 근거한 결론으로부터 모집단에 관한 추론을 하는 것이다.
- 이러한 과정에서 표본은 아주 중요한 역할
- 표본(데이터)을 얻는 방법 : 실험, 관찰, 조사

## 2.2.1 실험(實驗, experiment)

- 데이터를 얻는 대표적인 방법은 실험
- 실험은 가설이나 이론이 실제로 들어맞는지를 확인하기 위해 다양한 조건 아래에서 여러 가지 측정을 실시하는 일이다.
- 실험은 관찰(측정도 포함)과 함께 과학의 기본적인 방법의 하나이다. 다만 관찰이 대상 그 자체를 있는 그대로 알아보는 일이라면, 실험은 어떤 조작을 가해 그에 따라 일어나는 변화를 조사하고 결론을 내는 일이다.  
(출처 – 위키 백과)

효율적 실험을 위하여 준비하고 계획하는  
내용을 연구하는 분야

### 실험계획과 분석(experimental design and analysis)

- 자료를 수집하기 전에 미리 어떻게 실험할 것인지를 **계획**하여 원하는 **데이터를 정확하게 수집**하고 기록할 수 있도록 하는 통계학의 한 분야
- 실험의 계획뿐 만 아니라 실험을 통한 자료의 분석까지도 포함되는 통계분야

## 실험의 종류

- 통계학에서의 실험이란  
자연과학적 실험, 사회과학적 실험  
인문과학적 실험 등을 모두  
포함한 실험을 통칭한다.



## 2.2.2 관찰(observe)

- 데이터를 얻는 또 다른 방법은 관찰(observe)이다.
- 대부분의 경우 표본을 자세히 **관찰**함으로써 데이터(관찰값)를 얻게 됨
- 관찰은 큰 의미에서 **측정(measurement)**을 포함
- 예) 몸무게의 측정, 키의 측정 등



### 2.2.3 조사(survey)

- 실험이나 관찰 혹은 측정을 통하여 데이터를 획득할 수 없는 경우
- 데이터를 얻는 방법은 **조사**(survey)
- 예) 여론조사나 의식조사 – 관찰이나 실험으로 데이터를 얻기 어려워 설문지나 조사지 등을 통해 데이터를 얻음





## 2.3 데이터의 구분

- 실험, 관찰 혹은 조사를 통해 얻어진 데이터를 속성이나 형태에 따라 구분하는 방법
- 속성에 따라 **질적** 데이터와 **양적** 데이터로 구분
- 형태에 따라 **연속형** 데이터와 **이산형** 데이터로 구분  
(연속형 : 연속인 어떤 구간에서 값을 취하는 자료,  
이산형 : 셀 수 있는 형태의 값을 취하는 자료)  
예) 몸무게-연속형, 한 페이지에 있는 오자의 수-이산형

## 속성에 따른 구분

- 질적 데이터(qualitative data)
  - 숫자로 나타낼 수는 있으나 아무 의미 없음
  - 범주형 데이터(categorical data)라고도 함
  - 예) 성별, 교육수준, 결혼상태, 종교, 혈액형 등
- 양적 데이터(quantitative data)
  - 데이터 자체가 숫자로 나타남
  - 데이터의 속성을 그대로 나타내고 있음
  - 예) 키, 몸무게, 지능지수(intelligent quotient), 감성지수(emotional quotient), 수능점수, 스마트폰 배터리 수명, 월수입, 월 보험료 등
  - 수치형 데이터(numerical data)라고도 함
  - 경우에 따라 질적 데이터의 형태로 변환 가능  
(집단화, 군화(grouping)를 통하여)

## 속성에 따른 구분(계속)

- **명목형** (nominal data) : 이름만이 의미가 있음, 대부분의 경우 범주형 데이터로 나타남. 예) 성별, 주민등록번호
- **순서형** (ordinal data) : 값의 차이는 의미 없지만 순서는 의미가 있음, 예) 학력수준, 학년, 소득수준 등
- **구간형** (interval data) : 대부분의 데이터, 값의 순서뿐만 아니라 차이까지 의미가 있음. 예) 시험점수, 지능지수, 물가지수, 온도 등
- **비율형** (ratio data) : 구간형과 비슷함. 원점(0)의 의미(비율의 의미)가 분명한 데이터. 예) 월 소득 (소득이 0이란 의미는 전혀 수입이 없음을 의미)

## 데이터의 구분(속성)

|                             |                     |                                                          |
|-----------------------------|---------------------|----------------------------------------------------------|
| ▶ 범주형<br>(categorical data) | 명목형 (nominal data)  | 대상의 범주적 속성만을 나타내는 데이터<br>예) 결혼상태(결혼, 미혼), 혈액형(A,B,AB,O), |
|                             | 순서형 (ordinal data)  | 대상의 속성을 순서로 구분할 수 있는<br>데이터<br>예) 학력, 학년, 생활 정도          |
| ▶ 수치형<br>(numerical data)   | 구간형 (interval data) | 대상의 속성을 숫자 값으로 나타낸 데이터<br>예) 사람의 키, 보험금                  |
|                             | 비율형 (ration data)   | 원점의 의미가 분명한 데이터<br>예) 소득                                 |

## 데이터의 구분(형태)

|                                        |                                                                                                                                                   |
|----------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------|
| <p>▶ 이산형 데이터<br/>(discrete data)</p>   | <p>자료가 이산인 값을 취하는 형태(하나씩 셀 수 있는 것)<br/>예) 사고건 수, 안타의 수, 하루 동안 택시에 탄 승객의 수,<br/>한쪽당 오자의 수, 하루당 은행방문자 수 등</p>                                       |
| <p>▶ 연속형 데이터<br/>(continuous data)</p> | <p>자료가 연속인 구간에서 값을 취하는 형태<br/>몸무게, 키 등 모든 가능한 측정값들이 연속인 구간에<br/>속하는 값으로 세기(counting)가 불가능 함<br/>예) 몸무게, 키, 나이, 혈당, 혈압, 콜레스테롤 수치,<br/>소변 납 농도,</p> |