# CHAP.2
# 데이터 기술
# data description

# 도수 분포표

**아래 40개의 자료에 대하여 계급의 수가 6인 도수분포표를 작성하라**

41 32 30 23 24 32 11 39 24 46 50 18 41 14 33 50 38 25 32 16
43 19 35 22 46 43 10 22 17 47 66 48 25 43 28 31 12 25 12 48

# 도수 분포표

**아래 40개의 자료에 대하여 계급의 수가 6인 도수분포표를 작성하라**

41 32 30 23 24 32 11 39 24 46 50 18 41 14 33 50 38 25 32 16
43 19 35 22 46 43 10 22 17 47 66 48 25 43 28 31 12 25 12 48

| 계 급 | 계급간격 | 도수 | 상대도수 | 누적도수 | 누적상대도수 | 계급값 |
|---|---|---|---|---|---|---|
| 제1계급 | | 9 | 0.225 | 9 | 0.225 | 14.5 |
| 제2계급 | | 9 | 0.225 | 18 | 0.450 | 24.5 |
| 제3계급 | | 9 | 0.225 | 27 | 0.675 | 34.5 |
| 제4계급 | | 10 | 0.250 | 37 | 0.925 | 44.5 |
| 제5계급 | | 2 | 0.050 | 39 | 0.975 | 54.5 |
| 제6계급 | | 1 | 0.025 | 40 | 1.000 | 64.5 |
| 합 계 | | 40 | 1.00 | | | |

# 도수 분포표

**아래 40개의 자료에 대하여 계급의 수가 6인 도수분포표를 작성하라**

41 32 30 23 24 32 11 39 24 46 50 18 41 14 33 50 38 25 32 16
43 19 35 22 46 43 10 22 17 47 66 48 25 43 28 31 12 25 12 48

| 계 급 | 계급간격 | 도수 | 상대도수 | 누적도수 | 누적상대도수 | 계급값 |
|---|---|---|---|---|---|---|
| 제1계급 | 9.5 ~ 19.5 | 9 | 0.225 | 9 | 0.225 | 14.5 |
| 제2계급 | 19.5 ~ 29.5 | 9 | 0.225 | 18 | 0.450 | 24.5 |
| 제3계급 | 29.5 ~ 39.5 | 9 | 0.225 | 27 | 0.675 | 34.5 |
| 제4계급 | 39.5 ~ 49.5 | 10 | 0.250 | 37 | 0.925 | 44.5 |
| 제5계급 | 49.5 ~ 59.5 | 2 | 0.050 | 39 | 0.975 | 54.5 |
| 제6계급 | 59.5 ~ 69.5 | 1 | 0.025 | 40 | 1.000 | 64.5 |
| 합 계 | | 40 | 1.00 | | | |

- 누적도수(cumulative frequency)는 이전 계급까지의 모든 도수를 합한 도수
- 누적상대도수(cumulative relative frequency)는 이전 계급까지의 모든 상대도수를 합한 상대도수
- 계급값(class mark)은 각 계급의 중앙값

## [양적자료의 도수분포표 작성요령]

1. 계급의 수를 결정한다.
   자료의 수가 200 이하인 경우 : $k = \sqrt{n} \pm 3$
   자료의 수가 200을 넘는 경우 : Sturges 공식 $k = 1 + 3.3\log_{10} n$

2. 일정하게 주어지는 각 계급의 간격을 결정한다.
   $w$ = (최대 자료값 – 최소 자료값) / $k$

3. 각 계급이 중복되지 않도록 계급의 하한과 상한을 결정한다.
   제1계급의 하한 결정 : 최소 자료값 – (기본단위)/2

4. 끝으로 도수분포표 안에 각 계급의 도수, 상대도수, 누적도수, 누적상대도수, 계급값 등을 기입한다.

**[예제 6]**
다음 40개의 자료에 대하여 계급의 수가 6인 도수분포표를 작성하라.

41 32 30 23 24 32 11 39 24 46 50 18 41 14 33 50 38 25 32 16
43 19 35 22 46 43 10 22 17 47 66 48 25 43 28 31 12 25 12 48

**풀이**

계급의 수가 6이고 최댓값 66, 최솟값 10이므로 계급 간격 : $w = \frac{66-10}{6} = 9.3 \approx 10$

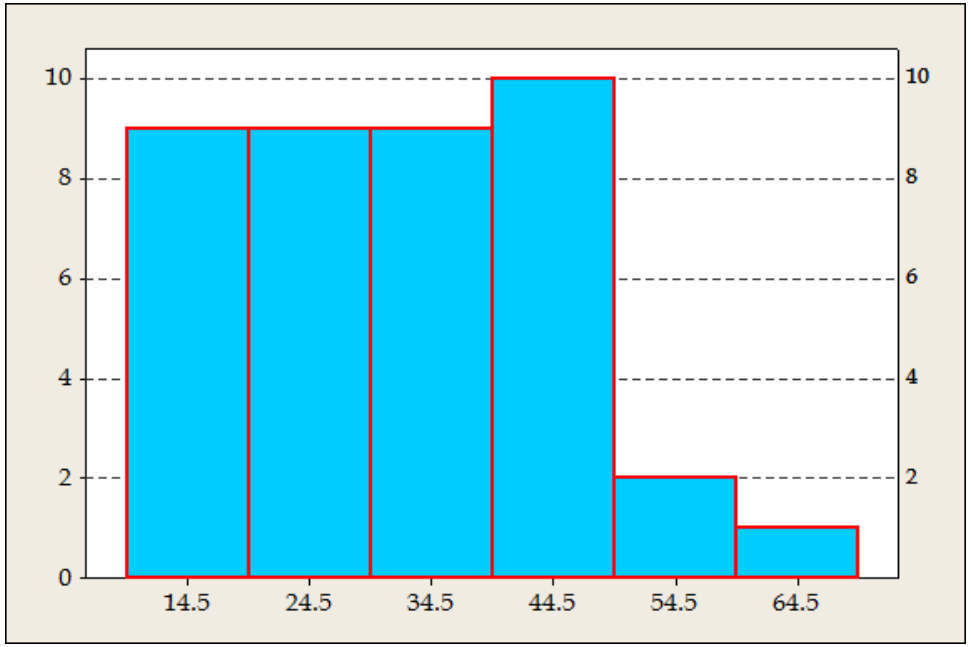
기본단위가 1이고 최솟값이 10이므로 제1계급의 하한 : 9.5

| 계 급 | 계급간격 | 도수 | 상대도수 | 누적도수 | 누적상대도수 | 계급값 |
|---|---|---|---|---|---|---|
| 제1계급 | 9.5 ~ 19.5 | 9 | 0.225 | 9 | 0.225 | 14.5 |
| 제2계급 | 19.5 ~ 29.5 | 9 | 0.225 | 18 | 0.450 | 24.5 |
| 제3계급 | 29.5 ~ 39.5 | 9 | 0.225 | 27 | 0.675 | 34.5 |
| 제4계급 | 39.5 ~ 49.5 | 10 | 0.250 | 37 | 0.925 | 44.5 |
| 제5계급 | 49.5 ~ 59.5 | 2 | 0.050 | 39 | 0.975 | 54.5 |
| 제6계급 | 59.5 ~ 69.5 | 1 | 0.025 | 40 | 1.000 | 64.5 |
| 합 계 | | 40 | 1.00 | | | |

# 도수 히스토그램

- 예제 6의 도수분포표를 이용하여 도수 히스토그램을 그려라.

- 각 계급의 계급값은 14.5, 24.5, 34.5, 44.5, 54.5, 64.5 이며,
- 각 계급의 도수는 9, 9, 9, 10, 2, 1이다.



# 코딩

예제6에 대해

자료 리스트: 하드코딩 또는 파일입력
기본단위: 하드코딩
계급의 수: 하드코딩 또는 Sturges 공식

min, max 찾기
- 함수 작성
- 기존 함수 활용; min( ), max( )

도수 분포표 그리기
상대도수 히스토그램 그리기
누적 상대도수 히스토그램 그리기

data=[41,32,30,23,24,32,11,39,24,46,50,18,41,14,33,50,38,25,
32,16,43,19,35,22,46,43,10,22,17,47,66,48,25,43,28,31,12,25, 12,48]

(numpy 사용하지 않음)

# 줄기-잎 그림; Stem-and-leaf plot

도수히스토그램을 나타내면서 동시에 정확한 자료값을 제공

**줄기-잎 그림을 작성하는 방법**

1. 줄기와 잎을 구분한다. 이때, 변동이 작은 부분을 줄기, 변동이 많은 부분을 잎으로 지정한다.

2. 수직방향으로 줄기부분을 작은 수부터 순차적으로 나열하고, 오른 쪽에 수직선을 긋는다.

3. 각 줄기부분에 해당하는 잎 부분을 원자료의 관찰 순서대로 나열한다.

4. 잎 부분의 자료값을 크기순으로 재배열한다.

5. 전체 자료를 크기순으로 나열하여 중앙에 놓이는 자료값이 있는 행의 왼쪽에 괄호( )를 만들고, 괄호 안에 그 행에 해당하는 잎의 수(도수)를 기입한다.

6. 괄호가 있는 행을 중심으로 괄호와 동일한 열에 누적도수를 위와 아래방향에서 각각 기입하고, 최소단위와 자료의 전체 개수를 기입한다.

| | |
|---|---|
| 1 | 50098 |
| 2 | 91990201898 |
| 3 | 0989831750811 |
| 4 | 9810290999555107 16 |
| 5 | 01 |
| 6 | 9 |

29 30 49 21 39 38 15 39 48 41 50 38 33 40 51 29 31 42 29 69
37 20 49 40 10 49 49 49 35 45 22 45 20 45 30 41 40 38 10 31
47 19 31 21 41 46 28 29 18 28

50개

1. 10의 자릿수를 줄기 그리고 1의 자릿수를 잎으로 구분하여, 줄기와 잎 사이에 수직선을 긋는다.
2. 수직방향으로 줄기부분을 작성하고 관찰된 순서대로 잎부분을 기록한다.

| | |
|---|---|
| 1 | 5 0 0 9 8 |
| 2 | 9 1 9 9 0 2 0 1 8 9 8 |
| 3 | 0 9 8 9 8 3 1 7 5 0 8 1 1 |
| 4 | 9 8 1 0 2 9 0 9 9 9 5 5 5 1 0 7 1 6 |
| 5 | 0 1 |
| 6 | 9 |

3. 잎 부분의 자료값을 크기순으로 재배열하고, 가장 가운데 놓이는 자료값이 있는 행의 맨 왼쪽에 잎의 수를 괄호 안에 기입한다.

| | | |
|---|---|---|
| | 1 | 0 0 5 8 9 |
| | 2 | 0 0 1 1 2 8 8 9 9 9 9 |
| (13) | 3 | 0 0 1 1 1 3 5 7 8 8 8 9 9 |
| | 4 | 0 0 0 1 1 1 2 5 5 5 6 7 8 9 9 9 9 9 |
| | 5 | 0 1 |
| | 6 | 9 |

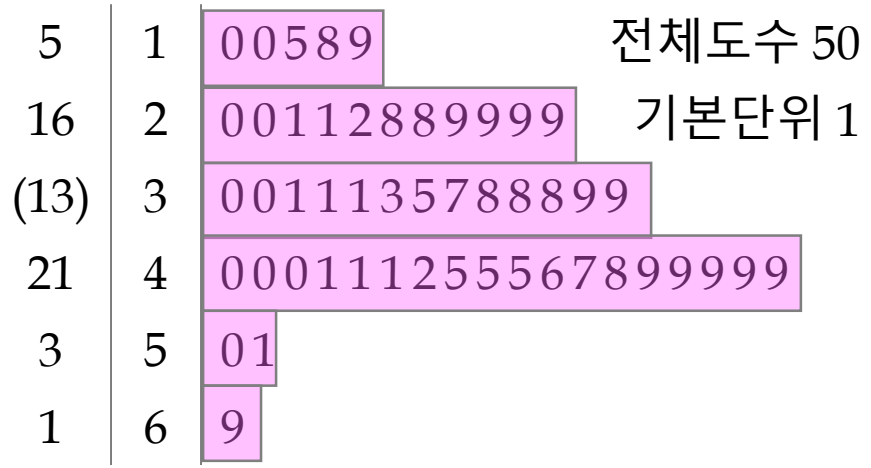
4. 괄호 안의 수를 중심으로 위쪽과 아래쪽으로부터 누적도수를 기입하고, 전체 도수와 기본단위를 기입한다.

※ 가장 가운데 놓이는 자료값이 서로 다른 행으로 분리된다면, 그 두 행을 기준으로 누적도수를 기입한다.

| | | | |
|---|---|---|---|
| **5** | **1** | **0 0 5 8 9** | **전체도수 50** |
| **16** | **2** | **0 0 1 1 2 8 8 9 9 9 9** | **기본단위 1** |
| **(13)** | **3** | **0 0 1 1 1 3 5 7 8 8 8 9 9** | |
| **21** | **4** | **0 0 0 1 1 1 2 5 5 5 6 7 8 9 9 9 9 9** | |
| **3** | **5** | **0 1** | |
| **1** | **6** | **9** | |

## 줄기-잎 그림의 해석

첫 번째 행 **1 | 00589**는 자료값이 10, 10, 15, 18, 19이고, 계급간격 10 ~ 19 사이에 도수가 5임을 나타낸다.



| | |
|---|---|
| A 고등학교 | 40 57 44 74 45 77 47 57 51 80 57 90 54 85 53 82 60 94 55 67 42 63 44 76 56<br>78 48 60 52 81 60 93 55 86 53 49 60 54 55 87 61 67 64 69 63 69 62 68 66 69 |
| B 고등학교 | 55 81 57 85 65 95 71 75 69 98 73 66 96 72 78 57 84 65 89 46 82 59 88 85 81<br>78 85 66 96 72 76 70 99 75 68 97 73 79 87 84 65 92 56 83 62 88 86 82 58 87 |

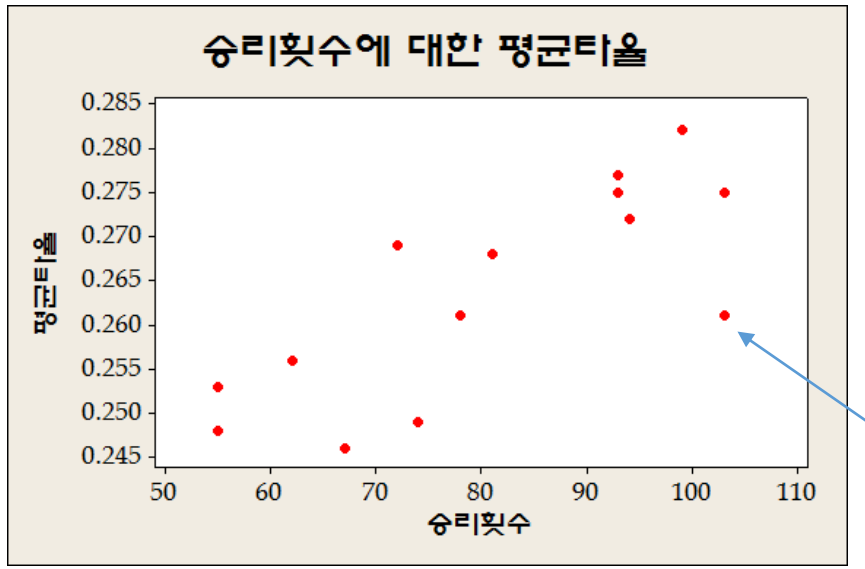
# 산점도; scatter diagram

두 종류의 자료가 **독립변수**와 **응답/종속변수**의 관계를 가짐으로써 *(x, y)* 형태의 쌍으로 주어진 자료를 나타내는 그림

| 팀 | 승리 게임 수 | 평균 타율 |
|---|---|---|
| Yew York | 103 | 0.275 |
| Toronto | 78 | 0.261 |
| Baltimore | 67 | 0.246 |
| Boston | 93 | 0.277 |
| Tampa Bay | 55 | 0.253 |
| Cleveland | 74 | 0.249 |
| Detroit | 55 | 0.248 |
| Chicago | 81 | 0.268 |
| Kansas City | 62 | 0.256 |
| Minnesota | 94 | 0.272 |
| Anaheim | 99 | 0.282 |
| Texas | 72 | 0.269 |
| Seattle | 93 | 0.275 |
| Oakland | 103 | 0.261 |

2002년 미국 MLB에 출전한 14팀 데이터

**산점도**




응답/종속변수

Oakland

독립변수

[관찰 결과] 승리횟수가 많을수록 평균타율이 좋고, 반대로 승리횟수가 적을수록 평균타율이 낮다는 결론을 얻는다. 그러나 Oakland 팀은 승리횟수가 많지만 평균타율은 높지 않은 편이다.

# 정리

- 도수/빈도(frequency)분포표
- 도수분포그래프; 히스토그램
  - 계급의 수, 계급 간격, 1계급 시작 값 (기본단위)
- 줄-잎 그림
- 산점도
  - 독립변수, 종속변수