# 3장 데이터 분석기법-기초

- **데이터의 분포 파악**
  - Histogram
  - Boxplot
  - 평균값과 분산
- **2개 양적 데이터의 관계 파악**
  - 산포도(scatter plot)
  - 상관계수(correlation coefficient)

# 01 데이터의 분포 파악: 예제 데이터

- 히코네시의 30년간 매달 1일의 최저기온 데이터

〈표 2.1〉 히코네시의 1988년부터 2017년까지 매년 10월1일의 최저기온(℃)

| 1988 | 1989 | 1990 | 1991 | 1992 | 1993 | 1994 | 1995 | 1996 | 1997 |
|---|---|---|---|---|---|---|---|---|---|
| 15.2 | 12.0 | 19.8 | 17.5 | 15.8 | 14.4 | 20.3 | 18.2 | 15.4 | 11.9 |
| 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 |
| 22.6 | 20.2 | 17.7 | 18.6 | 18.1 | 11.3 | 14.9 | 19.9 | 16.2 | 17.6 |
| 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 |
| 17.7 | 17.7 | 14.7 | 14.7 | 19.4 | 20.0 | 20.3 | 15.4 | 19.8 | 12.1 |

〈표 2.2〉 히코네시의 1988년부터 2017년까지 매년 12월1일의 최저기온(℃)

| 1988 | 1989 | 1990 | 1991 | 1992 | 1993 | 1994 | 1995 | 1996 | 1997 |
|---|---|---|---|---|---|---|---|---|---|
| 3.6 | 3.0 | 11.8 | 6.8 | 7.6 | 10.4 | 5.1 | 3.0 | -1.3 | 4.4 |
| 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 |
| 7.3 | 6.2 | 5.8 | 5.8 | 6.3 | 12.1 | 3.9 | 4.1 | 5.7 | 5.8 |
| 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 |
| 2.6 | 4.6 | 5.7 | 7.3 | 4.5 | 3.2 | 8.3 | 5.2 | 8.6 | 6.2 |

# 01-1 데이터의 분포 파악: 데이터 분류

- **Quantitative vs Qualitative  data**
  - quantitative data (양적 데이터): 개수, 길이 등의 수량을 나타내는 데이터
  - qualitative data (질적 데이터): 수량이 아닌 분류항목을 나타내는 데이터 (예: 성별, 생물의 종 등)

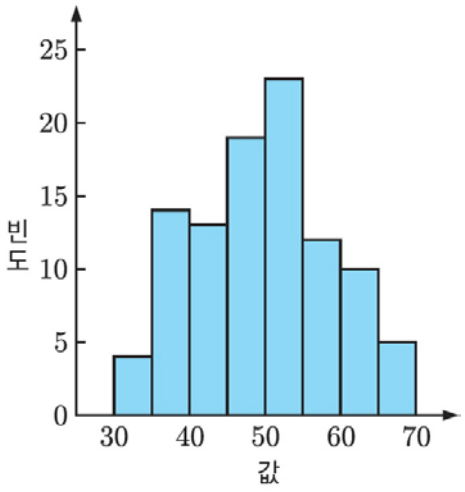
- **Quantitative 데이터의 연속성**
  - 이산 데이터(discrete data): 개수 등과 같이 정수로 표현
  - 연속 데이터(continuous data): 길이 등과 같이 실수로 표현

- **Quantitative 데이터의 척도**
  - Ratio scale: 데이터 간의 차이와 비율을 측정할 수 있는 척도 (예: 개수, 길이)
  - Interval scale: 데이터 사이의 차이는 의미를 갖지만 비율은 의미를 갖지 않음 (예: 기온)

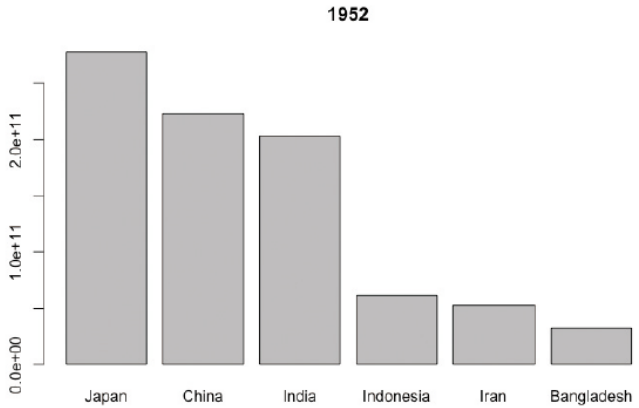
- **Qualitative 데이터의 척도**
  - ordinal scale: 대소/전후가 정해짐
  - nominal scale: 순서가 없음 (예: 남자/여자, 백조/비둘기/백로)

# 01-2 데이터의 분포 파악: 그래프 표현 (히스토그램)

- 히스토그램: 데이터의 전체적인 특징을 파악 가능. 양적 데이터의 분포를 나타내는 그래프로서  데이터의 값들이 어떻게 흩어져 있는지에 대한 경향을 살펴볼 수 있다.
- 각 구간들에 포함되는 데이터의 개수(도수, 빈도)를 막대의 길이로 표현한다.
- 막대의 길이는 상대빈도(빈도의 총합에 대한 각 구간의 빈도의 비율, 즉, 빈도/빈도의 총합계)를 나타내는 경우도 있음
- 연속 데이터의 히스토그램은 각각의 막대들을 간격없이 나열하여 표현한다.




〈그림 2.1〉 히스토그램의 예

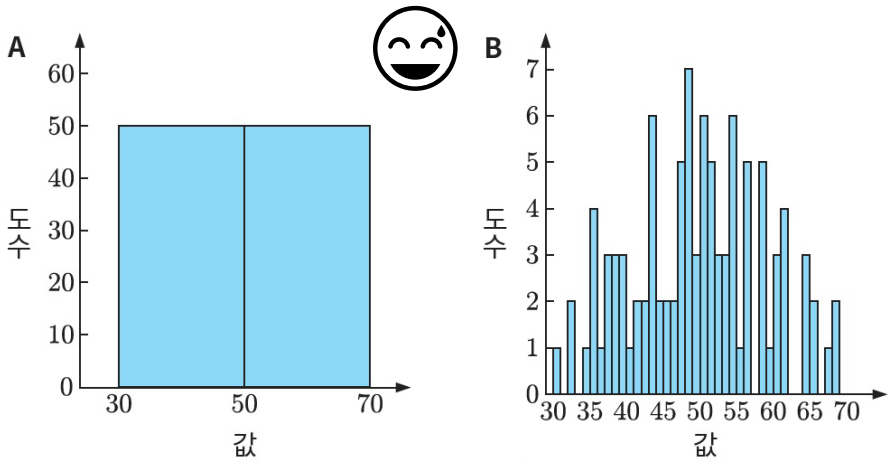



bar chart의 예

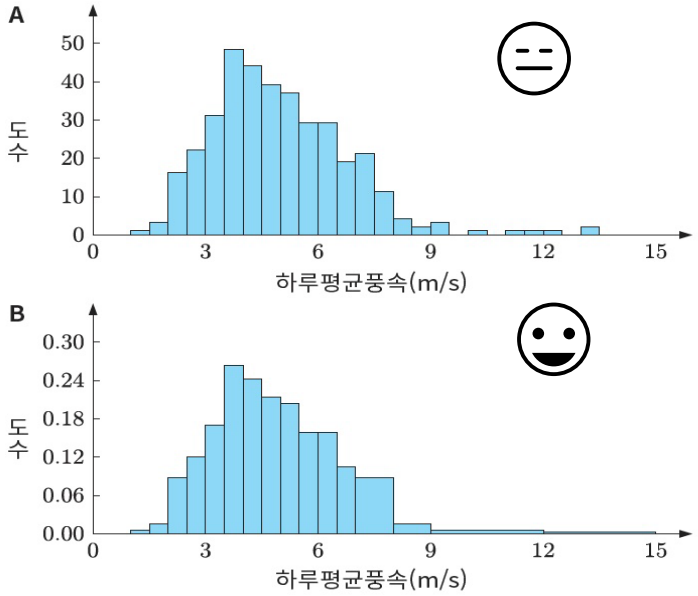
1952년 아시아 국가들의 gdp 구성과 순위

**Histograms** are sometimes confused with bar charts. A **histogram** is used for **continuous data**, where the bins represent ranges of **data**, while a bar chart is a plot of categorical variables.

# 01-2 데이터의 분포 파악: 히스토그램




〈그림 2.4〉 구간이 지나치게 넓은 히스토그램(A)과 너무 좁은 히스토그램(B)

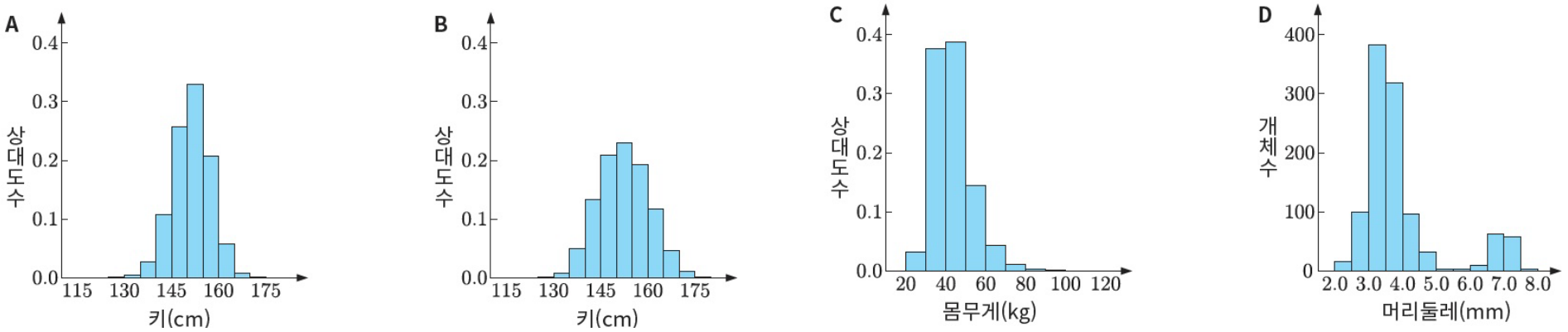



〈그림 2.5〉 구간의 폭이 일정한 히스토그램(A)과 구간의 폭을 적절하게 변경한 히스토그램(B). 일본 나하시(那覇市)의 2017년1월1일부터 2017년12월31일까지의 일평균풍속에 관한 데이터를 사용하였음.

- **구간의 갯수**
  - 간격이 너무 커지면 데이터의 분포를 파악하기 어려움
  - 간격이 너무 세밀해도 분포를 이해하기 어려움
  - 일반적으로 표본의 크기에 대한 제곱근 정도가 좋음

- **구간의 폭**
  - 기울어진 히스토그램은 구간의 폭을 일정하게 하면 극단적으로 데이터 개수가 매우 적은 구간이 발생하게 된다
  - 구간의 폭을 변경할 수 있으며, 막대의 면적이 도수에 비례하게 표현한다.

# 01-2 데이터의 분포 파악: 히스토그램 -모양분석




〈그림 2.2〉 다양한 형태의 히스토그램. A : 흩어진 정도가 적다. B : 흩어진 정도가 크다. C : 오른쪽으로 기울어져 있다. D : 쌍봉형. A는 12세 여자들의 키, B는 12세 남자들의 키, C는 12세 남자들의 몸무게로 2017년도 학교보건통계조사(https://www.e-stat.go.jp/stat-search/files?tstat=000001011648)에서 인용, D는 기가스 왕개미들 중에서 일개미의 머리부분 둘레에 대한 분포(Pfeiffer M. & Linsenmair K. E., 2000)로서 소형 일개미와 대형 일개미로 나누어져 있음을 알 수 있다.

- **흩어진 정도**
  - A: 흩어진 정도가 적음
  - B: 흩어진 정도가 큼
- **치우쳐진 모양새**
  - C: 왼쪽으로 기울어져 있음 (left-skewed)
- **산봉우리 개수**
  - D: 2개의 산봉우리가 있음, 쌍봉형(bimodal)
  - 2개 이상의 산봉우리를 갖는 경우: 다봉형 (multimodal) 히스토그램
  - A, B, C: 단봉형(unimodal)

# 01-2 데이터의 분포 파악: 히스토그램 - 예제 분석

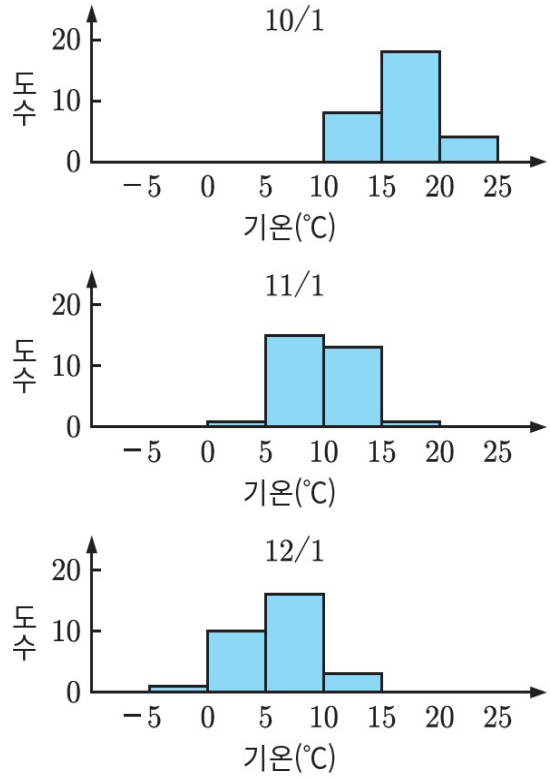
- 표 형태보다 히스토그램이 데이터의 경향을 파악하기 수월함
- 유사한 데이터를 비교하기 위하여 여러 개의 히스토그램을 동시에 나타내는 경우 가로/세로 축의 범위를 통일하여 나타내는 것이 좋다.

〈표 2.1〉 히코네시의 1988년부터 2017년까지 매년 10월1일의 최저기온(℃)

| 1988 | 1989 | 1990 | 1991 | 1992 | 1993 | 1994 | 1995 | 1996 | 1997 |
|---|---|---|---|---|---|---|---|---|---|
| 15.2 | 12.0 | 19.8 | 17.5 | 15.8 | 14.4 | 20.3 | 18.2 | 15.4 | 11.9 |
| 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 |
| 22.6 | 20.2 | 17.7 | 18.6 | 18.1 | 11.3 | 14.9 | 19.9 | 16.2 | 17.6 |
| 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 |
| 17.7 | 17.7 | 14.7 | 14.7 | 19.4 | 20.0 | 20.3 | 15.4 | 19.8 | 12.1 |

〈표 2.2〉 히코네시의 1988년부터 2017년까지 매년 12월1일의 최저기온(℃)

| 1988 | 1989 | 1990 | 1991 | 1992 | 1993 | 1994 | 1995 | 1996 | 1997 |
|---|---|---|---|---|---|---|---|---|---|
| 3.6 | 3.0 | 11.8 | 6.8 | 7.6 | 10.4 | 5.1 | 3.0 | -1.3 | 4.4 |
| 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 |
| 7.3 | 6.2 | 5.8 | 5.8 | 6.3 | 12.1 | 3.9 | 4.1 | 5.7 | 5.8 |
| 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 |
| 2.6 | 4.6 | 5.7 | 7.3 | 4.5 | 3.2 | 8.3 | 5.2 | 8.6 | 6.2 |

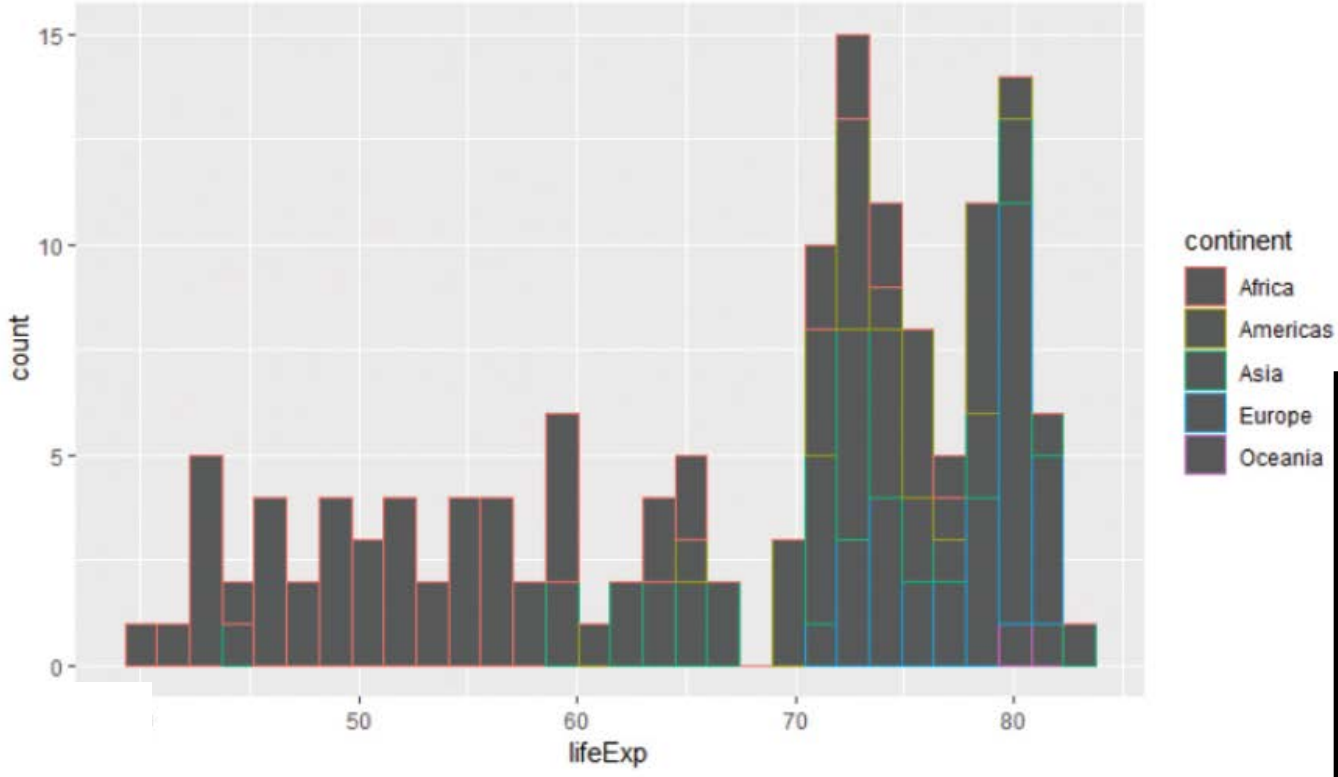



〈그림 2.3〉 1988년부터 2017년까지 30년간 히코네시의 10월1일, 11월1일, 12월1일의 최저기온에 대한 히스토그램

# 01-2 데이터의 분포 파악: 히스토그램 – ggplot2 라이브러리 이용

- geom_histogram 함수

```
> gapminder %>% filter(year == 2007) %>% ggplot(aes(lifeExp, col=continent)) +
geom_histogram()
```



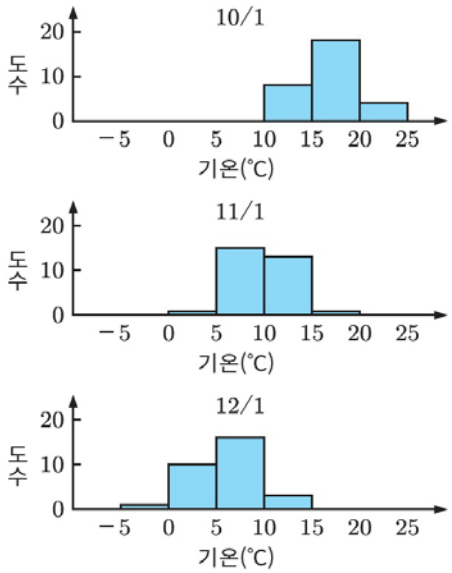

geom_histogram 함수를 이용한 히스토그램 : 그룹의 분포를 수직으로 쌓아 표시

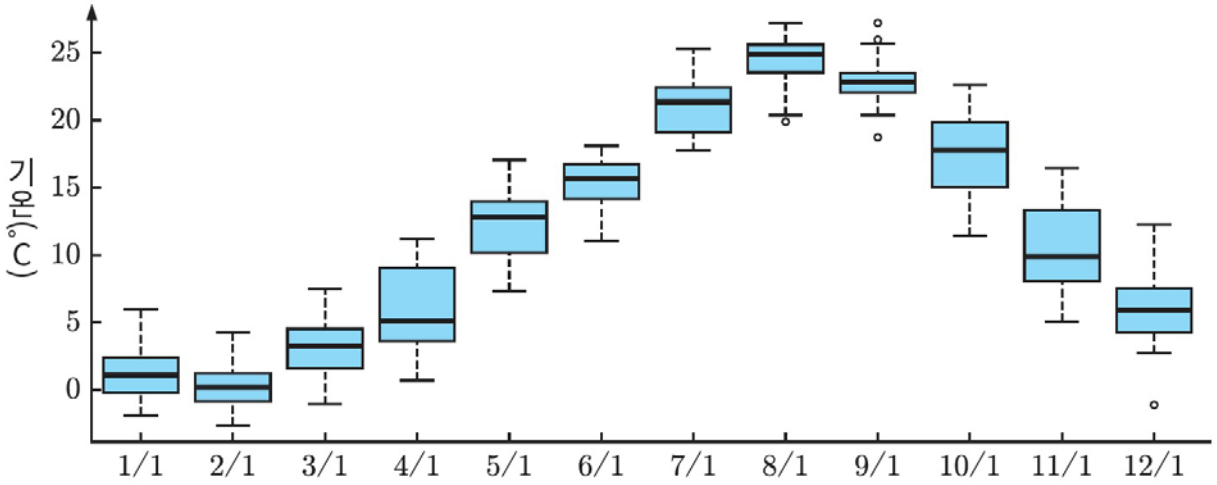
| country | continent | year | lifeExp |
|---|---|---|---|
| Swaziland | Africa | 2007 | 39.613 |
| Mozambique | Africa | 2007 | 42.082 |
| Zambia | Africa | 2007 | 42.384 |
| Sierra Leone | Africa | 2007 | 42.568 |
| Lesotho | Africa | 2007 | 42.592 |
| Angola | Africa | 2007 | 42.731 |
| Zimbabwe | Africa | 2007 | 43.487 |
| Afghanistan | Asia | 2007 | 43.828 |
| Central Africa | Africa | 2007 | 44.741 |
| Liberia | Africa | 2007 | 45.678 |
| Rwanda | Africa | 2007 | 46.242 |
| Guinea-Bissau | Africa | 2007 | 46.388 |
| Congo, Dem. | Africa | 2007 | 46.462 |

# 01-3 데이터의 분포 파악: 그래프 표현 (상자그림)

- **히스토그램은 너무 많은 정보량을 나타냄.**

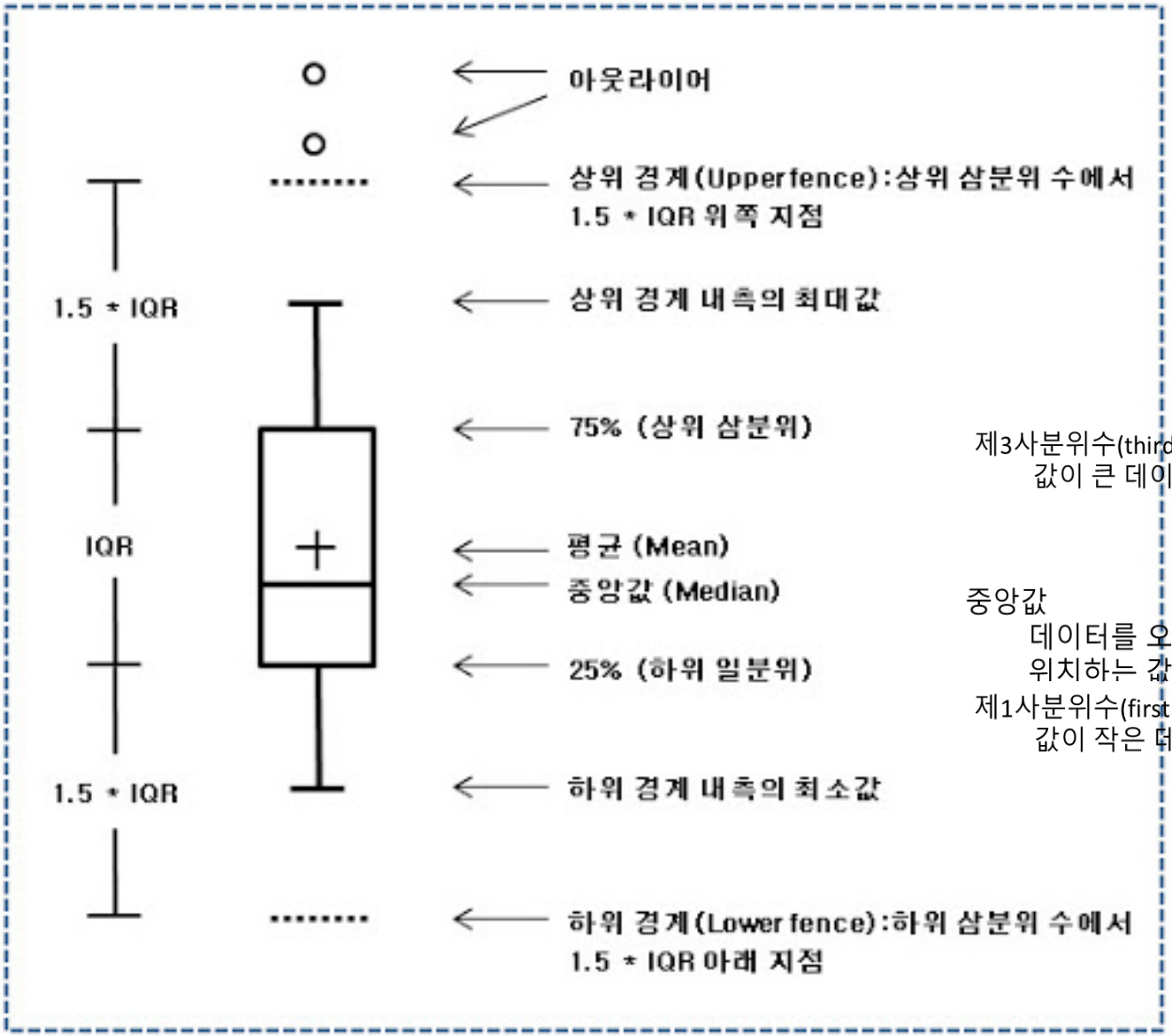
- **상자그림/상자수염그림 (box-and-whisker plot, box-and-whisker diagram):**
  - 보다 간편하게 요점 (데이터의 흩어진 모양새)만 알 수 있는 그래프 표현
  - 데이터의 흩어진 모양새를 상자에 수염이 달려있는 형태로 나타냄

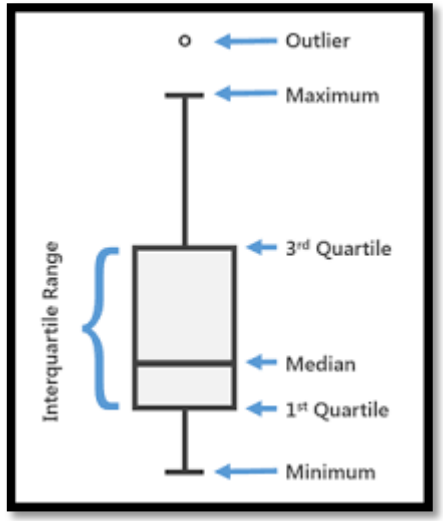



〈그림 2.7〉 1988년부터 2017년까지 30년간 히코네시의 매달 첫째날의 최저기온을 나타낸 상자수염그림

# 01-3 데이터의 분포 파악: 상자그림 – 해석 방법




제3사분위수(third quartile)
값이 큰 데이터의 중앙값

중앙값
데이터를 오름차순으로 정렬했을 때 중앙부분에 위치하는 값

제1사분위수(first quartile)
값이 작은 데이터의 중앙값

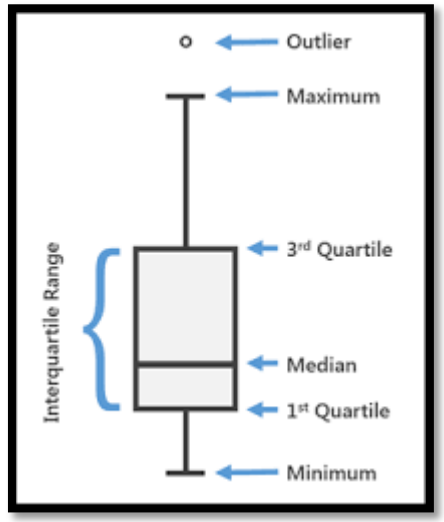
IQR(interquartile range): 제3사분위수와 제1사분위수의 차이

# 01-3 데이터의 분포 파악: 상자그림 – 해석 방법

중앙값
데이터를 오름차순으로 정렬했을 때 중앙부분에 위치하는 값. 데이터가 짝수개인 경우, 중앙부분에 위치하는 2개의 값에 대한 평균값

제1사분위수(first quartile)
값이 작은 데이터의 중앙값

제3사분위수(third quartile)
값이 큰 데이터의 중앙값




중앙값:
8

5, 7, 8, 11, 12

중앙값:
(7+8)/2=7.5

1, 5, 7, 8, 11, 12

제1사분위수:
5

제3사분위수:
11

1, 5, 7과 8, 11, 12로
분할해서
제1사분위수: 5
제3사분위수: 11

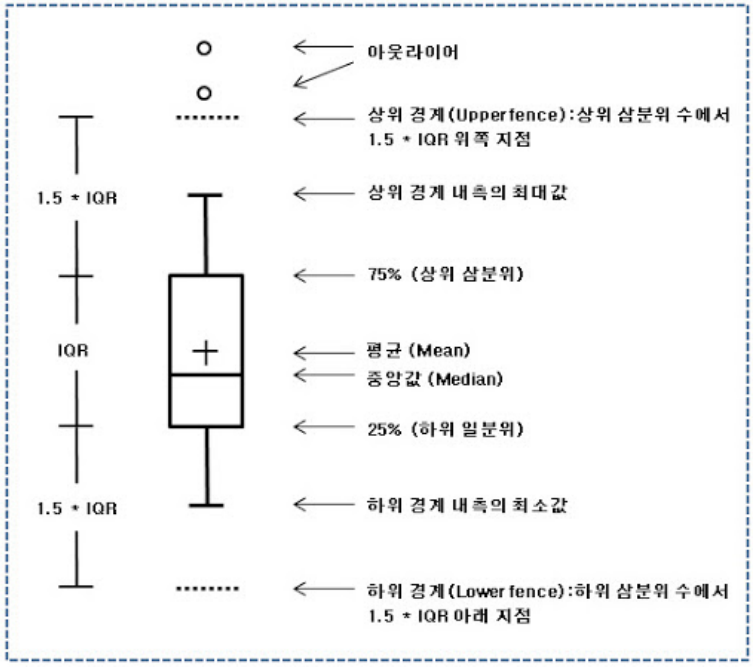
5, 7과 11, 12로 분할해서
제1사분위수: 6
제3사분위수: 11.5

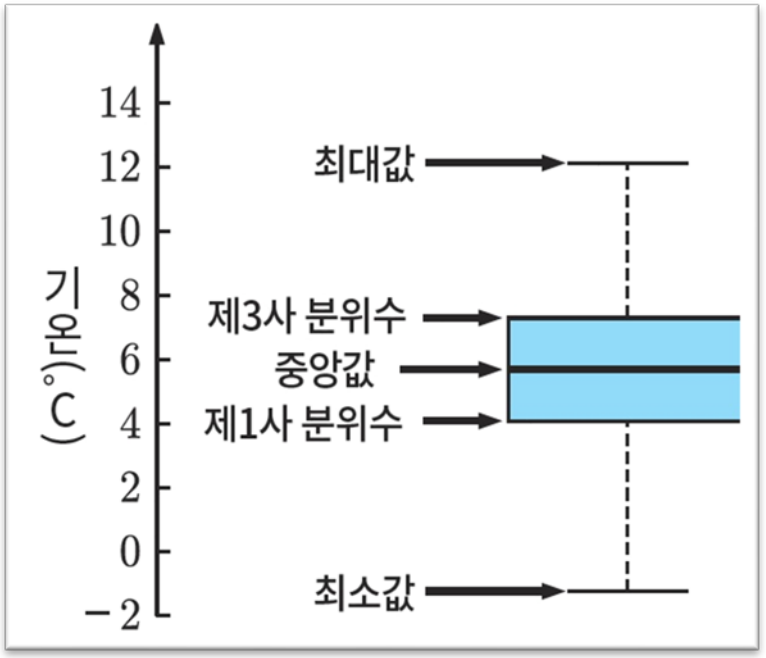
5, 7, 8과 8, 11, 12로
분할해서
제1사분위수:7
제3사분위수: 11

힌지법(Hinge)

# 01-3 데이터의 분포 파악: 상자그림 - 그리는 방식

- **튜키 방식**
  1. 데이터의 제1사분위수와 제3사분위수 사이에 상자를 그린다.
  2. 중앙값의 위치에 선분을 긋는다.
  3. 사분위범위의 1.5배 이상 떨어져 있는 값을 흰색 동그라미 점으로 그린다. (outlier 표시)
  4. Outlier가 아닌 데이터의 최대값과 최소값까지 수염을 그려 나간다.
- **간편법:** 모든 데이터들 중에서 최대값과 최소값까지 상자로부터 수염을 그려 나간다.



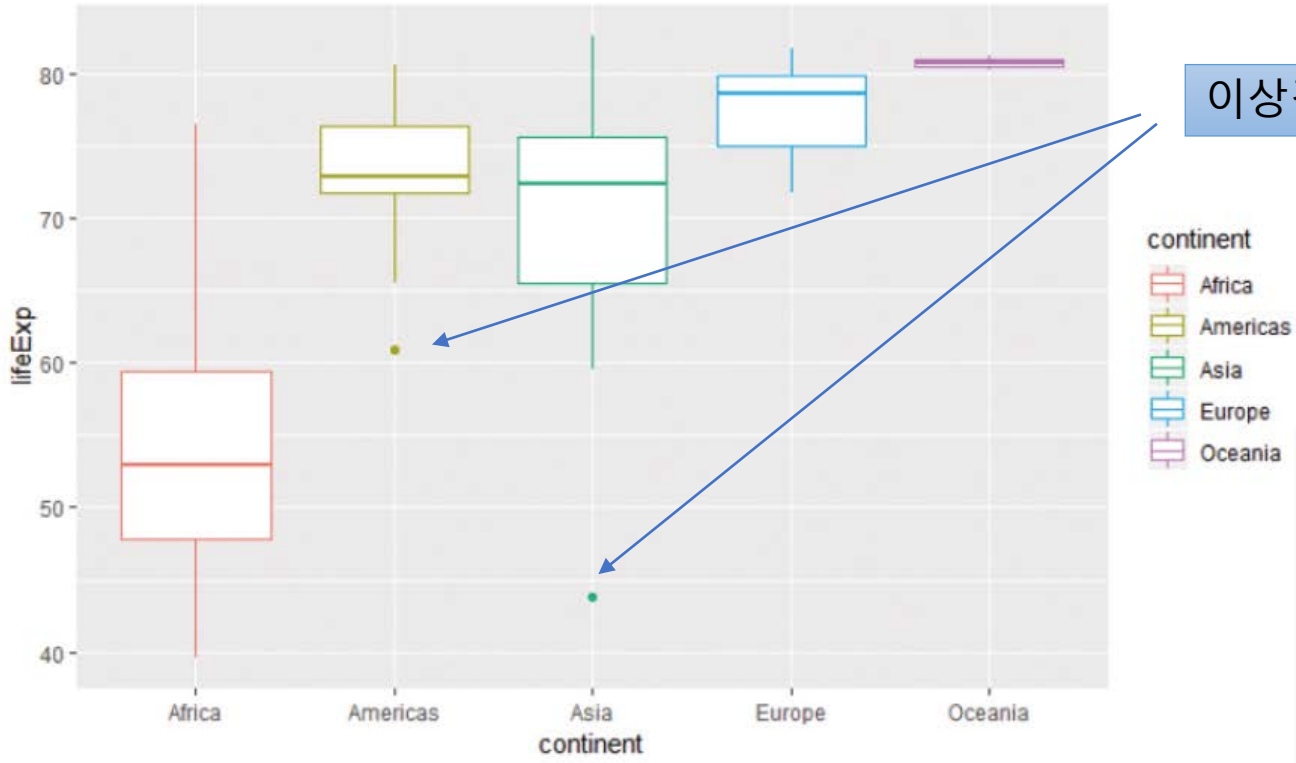




〈그림 2.6〉 A : 튜키(John W. Tukey)방식의 상자수염그림 , B : 간편법으로 표현한 상자수염그림

# 01-3 데이터의 분포 파악: 상자그림 -ggplot2 라이브러리 이용

- geom_boxplot 함수

```
> gapminder %>% filter(year == 2007) %>% ggplot(aes(continent,lifeExp, col=
continent)) + geom_boxplot()
```




이상값 (outlier)

geom_boxplot 함수를 이용한 박스플롯

| country | continent | year | lifeExp |
|---|---|---|---|
| Swaziland | Africa | 2007 | 39.613 |
| Mozambique | Africa | 2007 | 42.082 |
| Zambia | Africa | 2007 | 42.384 |
| Sierra Leone | Africa | 2007 | 42.568 |
| Lesotho | Africa | 2007 | 42.592 |
| Angola | Africa | 2007 | 42.731 |
| Zimbabwe | Africa | 2007 | 43.487 |
| Afghanistan | Asia | 2007 | 43.828 |
| Central Africa | Africa | 2007 | 44.741 |
| Liberia | Africa | 2007 | 45.678 |
| Rwanda | Africa | 2007 | 46.242 |
| Guinea-Bissau | Africa | 2007 | 46.388 |
| Congo, Dem. | Africa | 2007 | 46.462 |

# 01-4 데이터의 분포 파악: 평균값

- **대표값: 데이터의 값을 집약하여 1개의 숫자값으로 나타낸 것**
- **데이터의 대표값: 평균값, 중앙값**
- **평균값**
  - 데이터의 합을 표본의 크기로 나눈 값
  - 값이 전체적으로 커지면 평균값도 커지고, 값이 전체적으로 작아지면 평균값도 작아짐

$$\overline{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - 5cm, 10cm, 12cm, 13cm의 평균값

$$\frac{5 + 10 + 12 + 13}{4} = 10\ cm$$

# 01-4 데이터의 분포 파악: 분산과 표준편차

- 데이터의 흩어진 정도를 측정하는 지표: 분산, 표준편차
- 데이터의 흩어진 정도가 클 수록 분산과 표준편차의 값은 커지며, 반대로 데이터의 흩어진 정도가 작을 수록 분산과 표준편차의 값은 작아진다.
- 데이터: $x_1, x_2, \ldots, x_n$ (n개의 데이터 ), 평균값:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

- 분산 (variance: $S^2$)

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- 불편분산(unbiased variance: $\sigma^2$)

$$\sigma^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

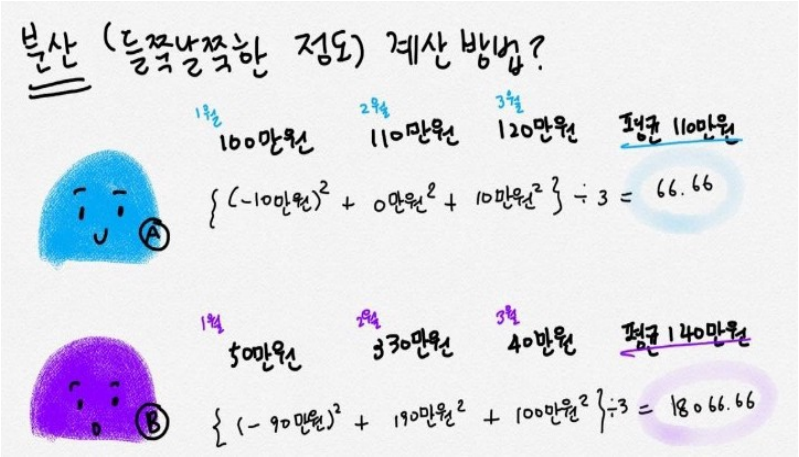
- 분산과 불편분산의 관계

$$\sigma^2 = \frac{n}{n-1}s^2$$

- 표준편차(standard deviation)
  - 분산의 제곱근$\sqrt{s^2}$(또는 불편분산의 제곱근 $\sqrt{\sigma^2}$)

# 01-4 데이터의 분포 파악: 분산 및 표준편차의 성질

- 실수의 제곱은 음수 값이 아니므로, 분산도 음수 값이 아님
- 표준편차는 분산의 제곱근이므로, 음수 값이 아님
- 모든 값이 같은 값인 경우에는 분산과 표준편차가 0이 됨




- 예제: 5cm, 10cm, 12cm, 13cm에 대한 분산과 표준편차
  - 평균 $\frac{5+10+12+13}{4} = 10\ cm$
  - 분산 $\frac{(5-10)^2+(10-10)^2+(12-10)^2+(13-10)^2}{4} = 9.5\ \mathrm{cm}^2$
  - 표준편차 $\sqrt{9.5} \approx 3.08\ \mathrm{cm}$
  - 평균값과 표준편차는 원래의 값과 같은 단위이지만, 분산은 단위가 다름. 따라서 단위가 같은 평균, 표준편차를 많이 사용함.

# 01-4 데이터의 분포 파악: 표준편차 비교

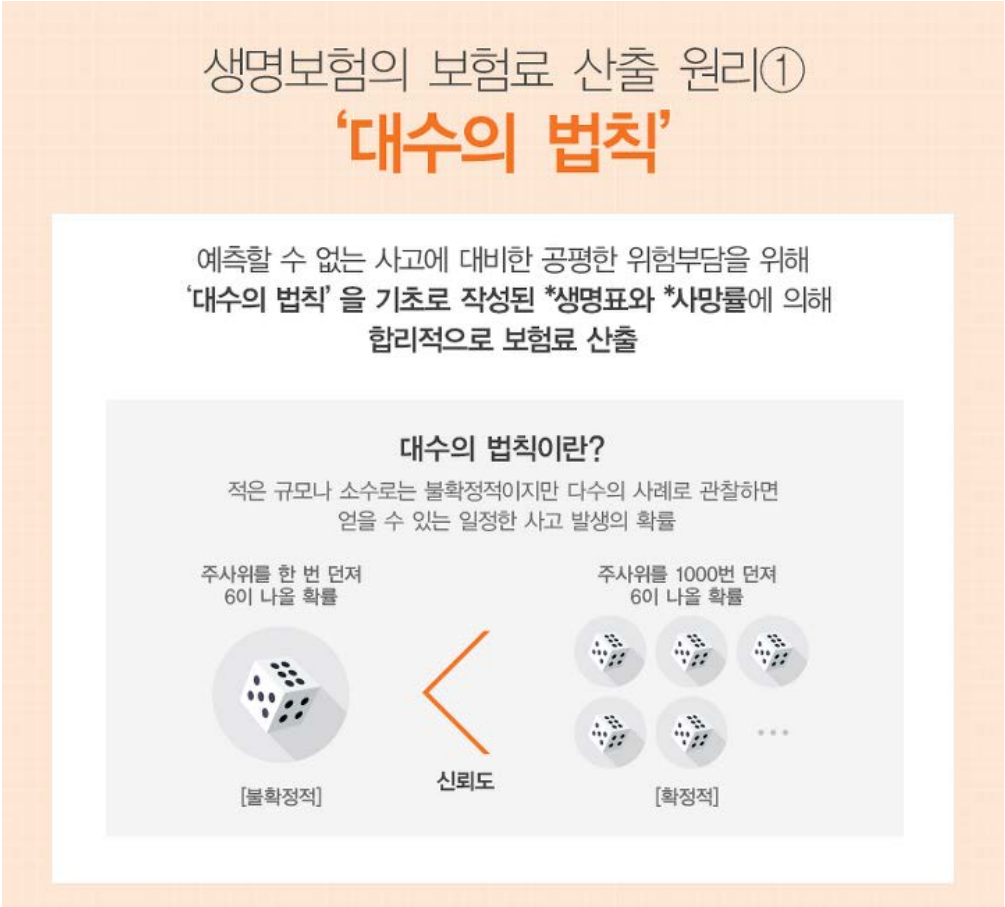
- 2011년부터 2017년까지 7년간 히코네시의 10월1일, 11월1일, 12월 1일의 각 최저기온과 평균값, 표준편차
- 최저기온의 흩어진 정도가 적은 날은 표준편차 값이 적어지고, 큰 날은 표준편차가 커짐

〈표 2.3〉 히코네시의 2011년부터 2017년까지의 10월1일, 11월1일, 12월1의 각 최저기온과 평균값, 표준편차(℃)

| | 2011 | 2012 | 2013 | 2014 | 2015 | 2016 | 2017 | 평균 | 표준편차 |
|---|---|---|---|---|---|---|---|---|---|
| 10/1 | 14.7 | 19.4 | 20.0 | 20.3 | 15.4 | 19.8 | 12.1 | 17.4 | 3.0 |
| 11/1 | 11.3 | 9.8 | 9.7 | 14.3 | 5.8 | 7.9 | 5.9 | 9.2 | 2.8 |
| 12/1 | 7.3 | 4.5 | 3.2 | 8.3 | 5.2 | 8.6 | 6.2 | 6.2 | 1.9 |

# 01-4 데이터의 분포 파악: 평균값의 성질

- 대수/큰수의 법칙 (Law of large Numbers): 모집단에서 무작위로 뽑은 표본의 평균은 표본의 크기가 커질수록 모집단의 평균에 근사하다는  통계적 의미



# 01-4 데이터의 분포 파악: 평균/분산 구하기

anscombe

Filter

| | x1 | x2 | x3 | x4 | y1 | y2 | y3 | y4 |
|---|---|---|---|---|---|---|---|---|
| 1 | 10 | 10 | 10 | 8 | 8.04 | 9.14 | 7.46 | 6.58 |
| 2 | 8 | 8 | 8 | 8 | 6.95 | 8.14 | 6.77 | 5.76 |
| 3 | 13 | 13 | 13 | 8 | 7.58 | 8.74 | 12.74 | 7.71 |
| 4 | 9 | 9 | 9 | 8 | 8.81 | 8.77 | 7.11 | 8.84 |
| 5 | 11 | 11 | 11 | 8 | 8.33 | 9.26 | 7.81 | 8.47 |
| 6 | 14 | 14 | 14 | 8 | 9.96 | 8.10 | 8.84 | 7.04 |
| 7 | 6 | 6 | 6 | 8 | 7.24 | 6.13 | 6.08 | 5.25 |
| 8 | 4 | 4 | 4 | 19 | 4.26 | 3.10 | 5.39 | 12.50 |
| 9 | 12 | 12 | 12 | 8 | 10.84 | 9.13 | 8.15 | 5.56 |
| 10 | 7 | 7 | 7 | 8 | 4.82 | 7.26 | 6.42 | 7.91 |
| 11 | 5 | 5 | 5 | 8 | 5.68 | 4.74 | 5.73 | 6.89 |

앤스콤의 4분할 데이터 셋

4개의 데이터 셋
data1=(x1, y1)
data2=(x2, y2)
data3=(x3, y3)
data4=(x4, y4)

```
> # 평균
> apply(anscombe, 2, mean)
      x1       x2       x3       x4       y1       y2       y3       y4
9.000000 9.000000 9.000000 9.000000 7.500909 7.500909 7.500000 7.500909
```

```
> # 분산
> apply(anscombe, 2, var)
       x1        x2        x3        x4        y1        y2        y3        y4
11.000000 11.000000 11.000000 11.000000  4.127269  4.127629  4.122620  4.123249
```

# 02-1 2개의 양적 데이터의 관계 파악: 데이터

- 2종류의 양적 데이터를 변수 X와 Y로 나타내면 전체 n개 쌍의 데이터로 표현

〈표 2.4〉 $n$개 쌍의 2종류 데이터

| 개체 | 1 | 2 | 3 | … | $n$ |
|---|---|---|---|---|---|
| 변수 $X$ | $x_1$ | $x_2$ | $x_3$ | … | $x_n$ |
| 변수 $Y$ | $y_1$ | $y_2$ | $y_3$ | … | $y_n$ |

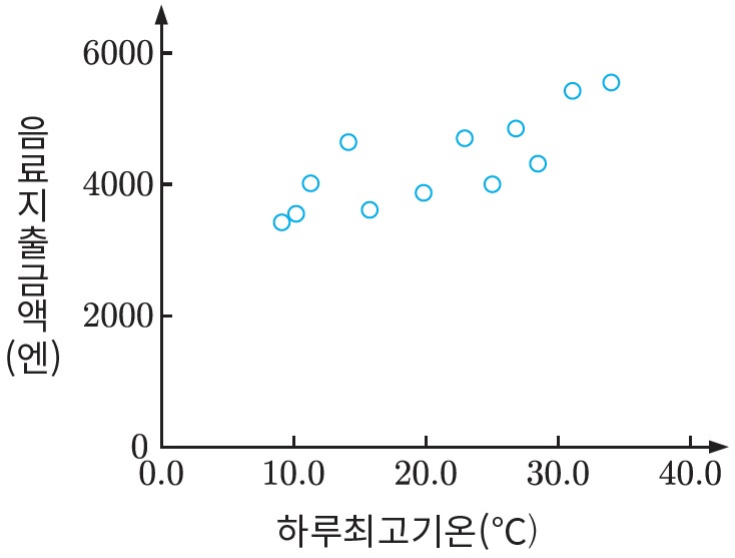
- 2016년 시가현 오오쯔시의 매월 하루최고기온에 대한 평균값과 2인 이상 가구의 세대별 음료지출금액

〈표 2.5〉 하루최고기온과 음료지출금액에 관한 데이터

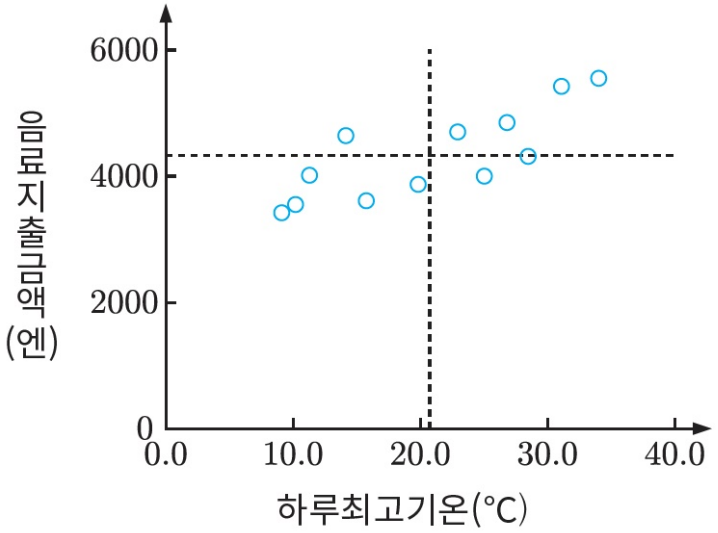
| 월 | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| 하루최고기온(℃) | 9.1 | 10.2 | 14.1 | 19.8 | 25.0 | 26.8 |
| 음료지출금액(엔) | 3416 | 3549 | 4639 | 3857 | 3989 | 4837 |
| 월 | 7 | 8 | 9 | 10 | 11 | 12 |
| 하루최고기온(℃) | 31.1 | 34.0 | 28.5 | 22.9 | 15.7 | 11.3 |
| 음료지출금액(엔) | 5419 | 5548 | 4311 | 4692 | 3607 | 4002 |

# 02-2 2개의 양적 데이터의 관계 파악: 산포도

- **Scatter plot (산포도): X-Y 평면에 표현**
- **예제:**
  - 가로축: 하루 최고 기온
  - 세로축: 음료 지출 금액
- **변수 X와 Y의 평균값을 나타내는 2개의 보조선을 추가**

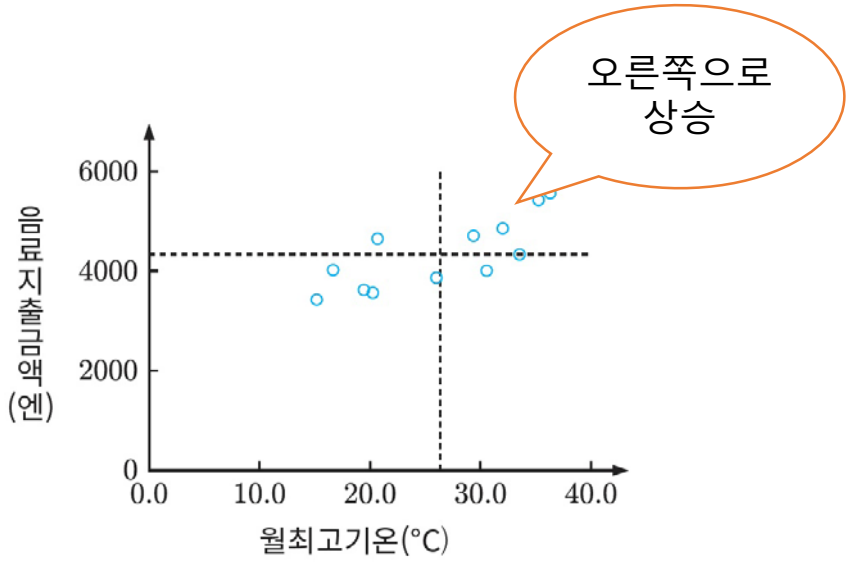



〈그림 2.8〉 하루최고기온과 음료지출금액의 산포도

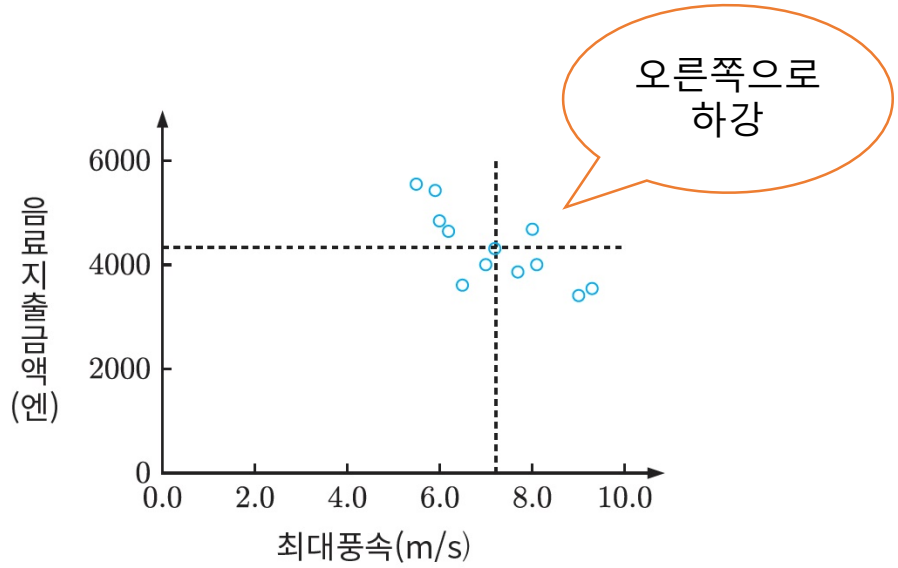



〈그림 2.9〉 그림 2.8의 산포도에 보조선(점선)을 추가한 산포도

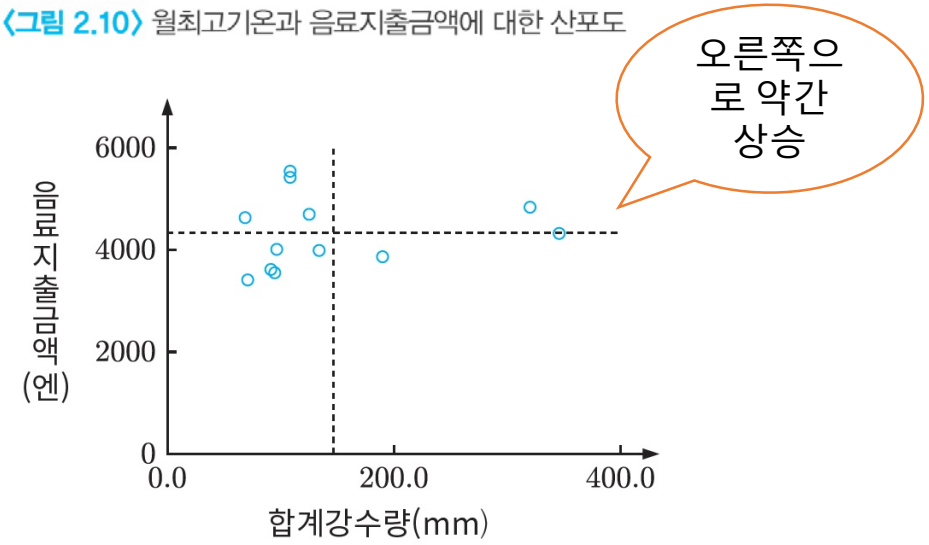
# 02-2 2개의 양적 데이터의 관계 파악: 산포도 관찰법



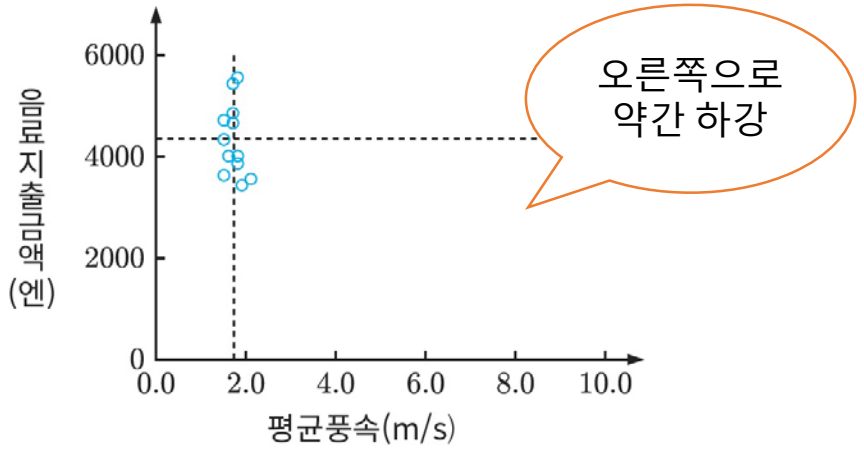

〈그림 2.10〉 월최고기온과 음료지출금액에 대한 산포도




〈그림 2.11〉 최대풍속과 음료지출금액에 대한 산포도



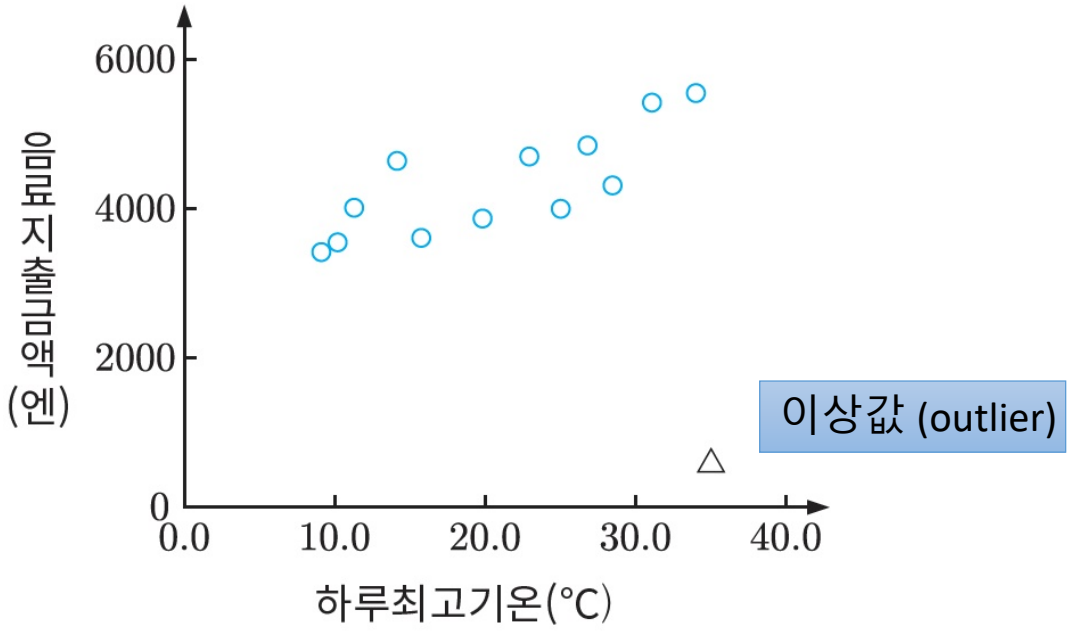

〈그림 2.12〉 합계강수량과 음료지출금액의 산포도




〈그림 2.13〉 평균풍속과 음료지출금액의 산포도

# 02-2 2개의 양적 데이터의 관계 파악: 산포도 (outlier 추출)

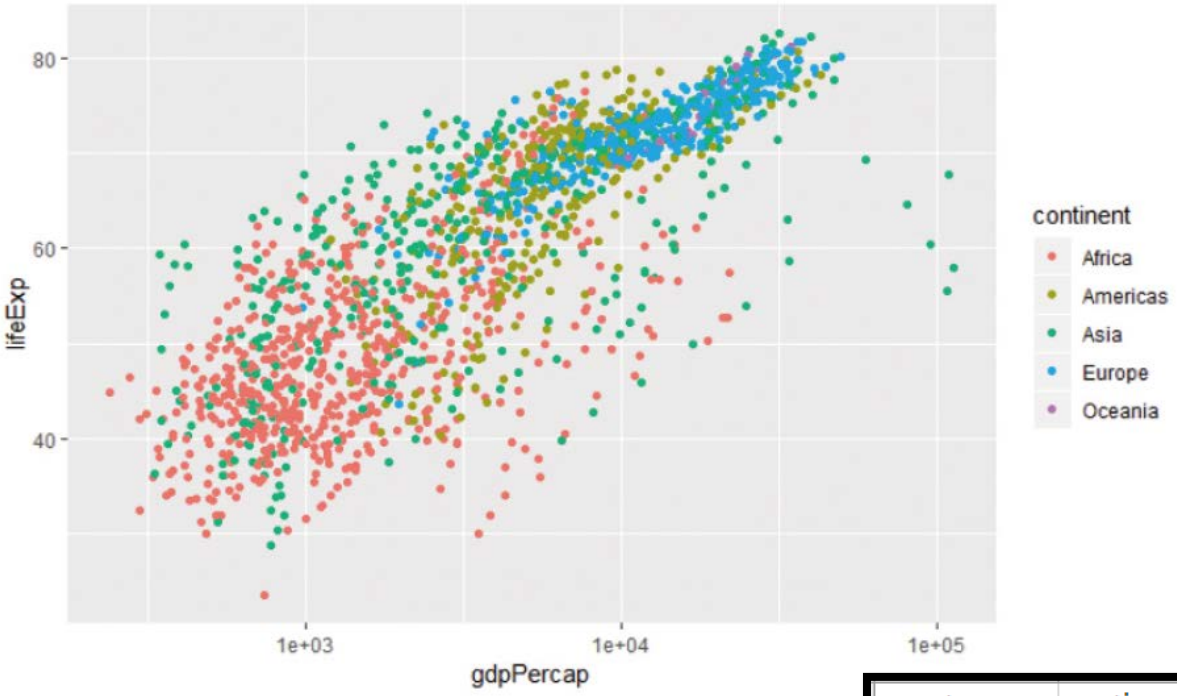
- △로 표시된 점은 ○로 표시된 다른 점들에 비해서 음료지출금액의 값이 극단적으로 낮기 때문에 이상값 (outlier)으로 설정
- 산포도로부터 이상값이 발견되는 경우
  - 입력에 오류가 없는지 확인
  - 데이터를 해석할 때 해당 값을 삭제할 것인지 검토




〈그림 2.14〉 가상의 데이터를 추가한 산포도

# 02-2 2개의 양적 데이터의 관계 파악: 산포도 그리기

```
> library(ggplot2)
> ggplot(gapminder, aes(x=gdpPercap, y=lifeExp, col=continent)) + geom_
point() + scale_x_log10()
```




| country | continent | year | lifeExp | pop | gdpPercap |
|---|---|---|---|---|---|
| Afghanistan | Asia | 1952 | 28.801 | 8425333 | 779.445315 |
| Afghanistan | Asia | 1957 | 30.332 | 9240934 | 820.85303 |
| Afghanistan | Asia | 1962 | 31.997 | 10267083 | 853.10071 |
| Afghanistan | Asia | 1967 | 34.02 | 11537966 | 836.197138 |
| Afghanistan | Asia | 1972 | 36.088 | 13079460 | 739.981106 |
| Afghanistan | Asia | 1977 | 38.438 | 14880372 | 786.11336 |
| Afghanistan | Asia | 1982 | 39.854 | 12881816 | 978.011439 |

# 02-3 2개의 양적 데이터의 관계 파악: 상관계수

- **공분산(covariance) $S_{XY}$: 2개의 확률변수의 상관정도를 나타냄.**

$$s_{XY} = \frac{1}{n}\{(x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})\}$$

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

- **상관관계(correlation):** 공분산을 2개 확률변수의 표준편차 값으로 나누어 준 값. 공분산 값을 정규화 하여 특정범위 값을 나오게 하는 상관계수의 의미

$$\text{상관계수 } r_{XY} = \frac{[X\text{와 } Y\text{의 공분산}]}{[X\text{의 표준편차}] \times [Y\text{의 표준편차}]} = \frac{S_{XY}}{S_X S_Y}$$

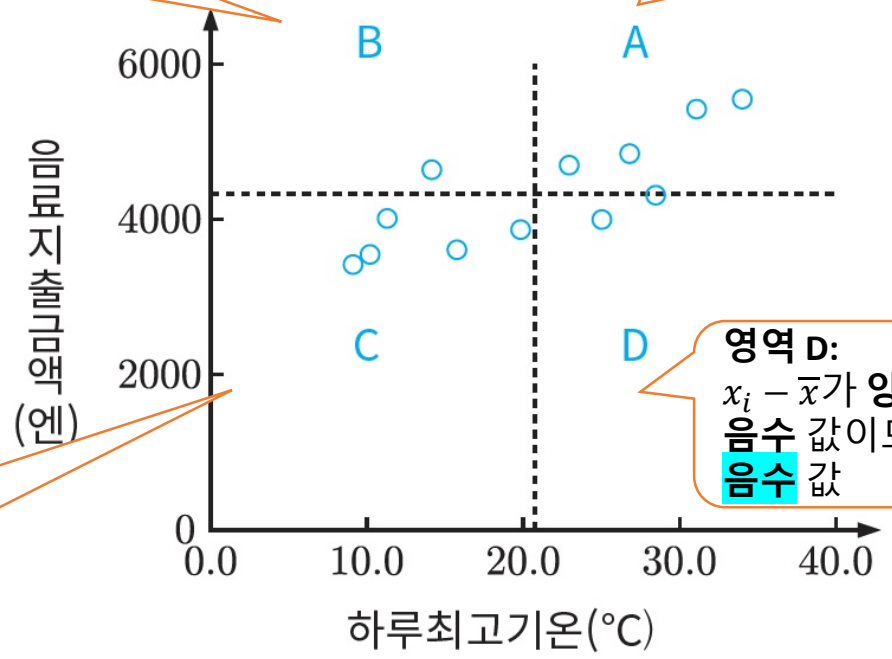
# 02-3 2개의 양적 데이터의 관계 파악: 상관계수의 부호

- **상관계수의 부호는 공분산을 구할 때 사용한 편차의 곱을 합한 값의 부호와 같다.**

$$s_{XY} = \frac{1}{n}\{(x_1 - \bar{x})(y_1 - \bar{y}) + \cdots + (x_n - \bar{x})(y_n - \bar{y})\}$$

**영역 B:**
$x_i - \bar{x}$가 **음수** 값이고 $y_i - \bar{y}$가 **양수** 값이므로 편차의 곱은 **음수** 값

**영역 A:**
$x_i - \bar{x}$와 $y_i - \bar{y}$가 **모두 양수** 값이므로 편차의 곱 $(x_i - \bar{x})(y_i - \bar{y})$도 **양수** 값




**영역 C:**
$x_i - \bar{x}$와 $y_i - \bar{y}$가 **모두 음수** 값이므로 편차의 곱은 **양수** 값

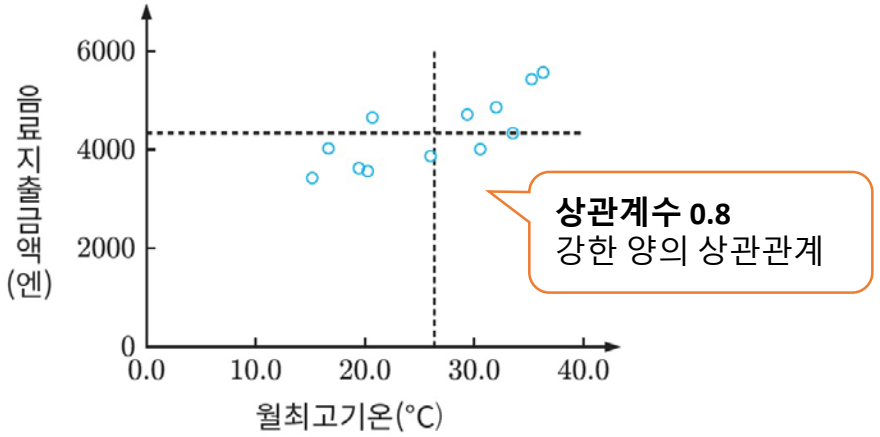
**영역 D:**
$x_i - \bar{x}$가 **양수** 값이고 $y_i - \bar{y}$가 **음수** 값이므로 편차의 곱은 **음수** 값

〈그림 2.15〉 4개 영역으로 분할한 산포도

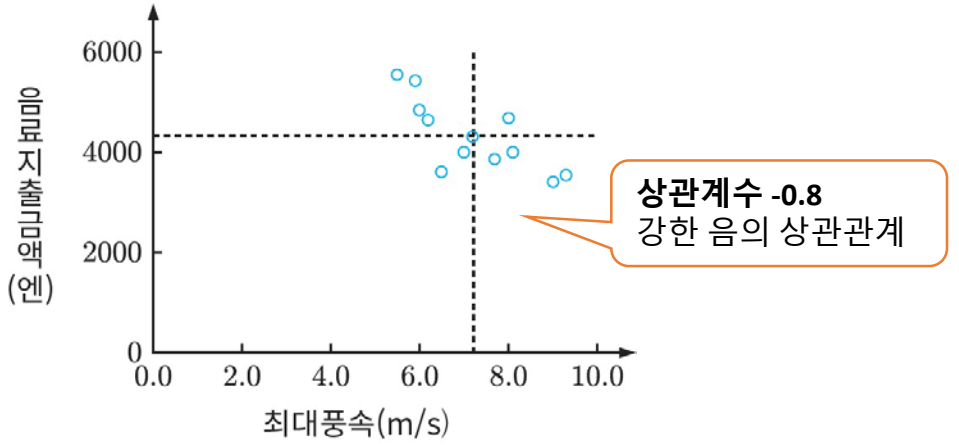
# 02-3 2개의 양적 데이터의 관계 파악: 상관계수의 값 평가

- **상관계수의 절대값에 따른 관계(직선 관계)**
  - 0 ~ 0.2 이하: 거의 관계가 없음, 0.2 ~ 0.4 이하 : 약한 관계
  - 0.4 ~ 0.7 이하 : 중간 정도, 0.7 ~ 1.0 : 강한 관계

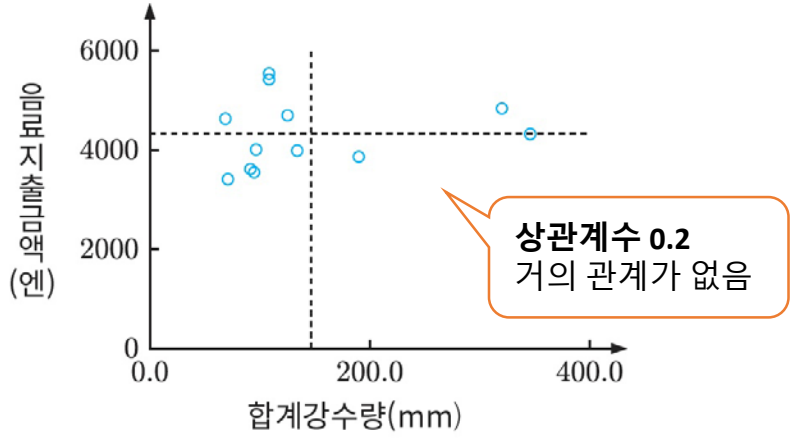
상관계수의 값에 대한 평가는 사용분야에 따라 다름



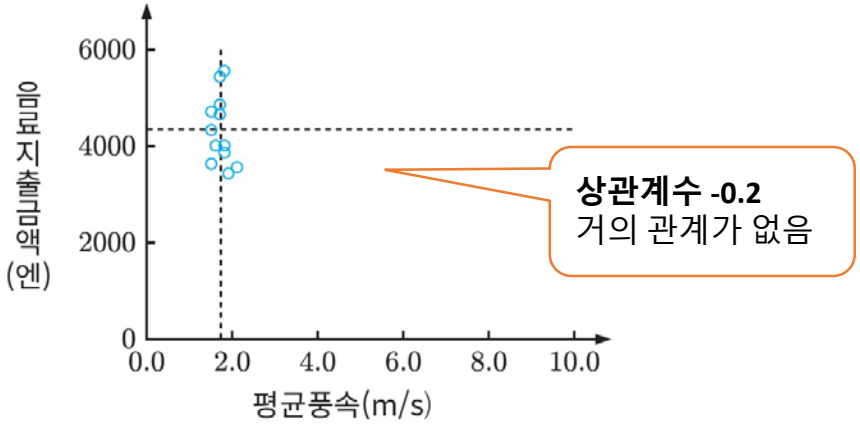

〈그림 2.10〉 월최고기온과 음료지출금액에 대한 산포도




〈그림 2.11〉 최대풍속과 음료지출금액에 대한 산포도



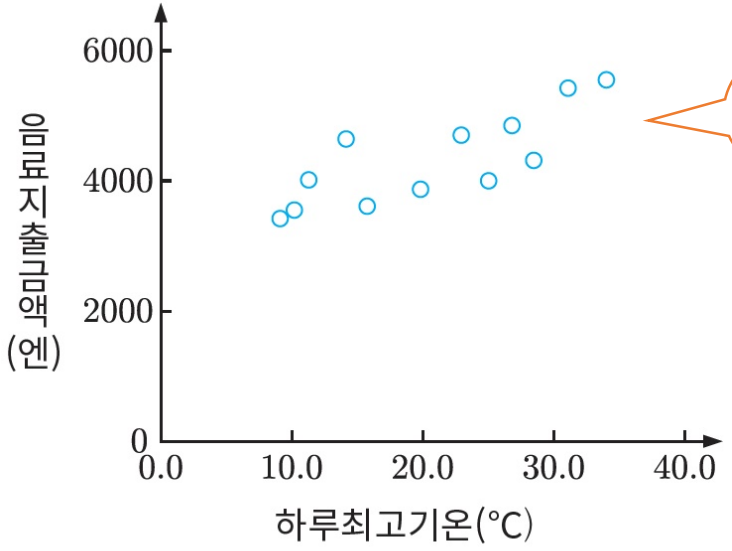

〈그림 2.12〉 합계강수량과 음료지출금액의 산포도



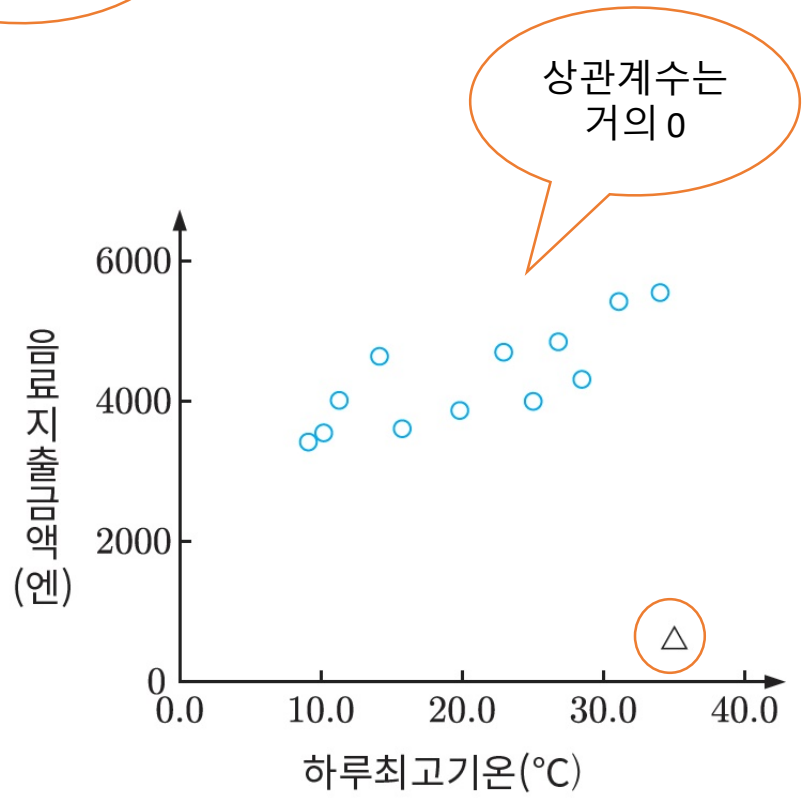

〈그림 2.13〉 평균풍속과 음료지출금액의 산포도

# 02-3 상관계수: outlier의 영향

- **데이터가 이상값 (outlier)을 포함하면 상관계수의 값이 크게 변하는 경우가 있음**




〈그림 2.8〉 하루최고기온과 음료지출금액의 산포도




〈그림 2.14〉 가상의 데이터를 추가한 산포도

# 02-3 2개의 양적 데이터의 관계 파악: 상관계수 구하기

```
> # 상관관계(상관계수)
> cor(anscombe$x1, anscombe$y1)
[1] 0.8164205

> cor(anscombe$x2, anscombe$y2)
[1] 0.8162365

> cor(anscombe$x3, anscombe$y3)
[1] 0.8162867

> cor(anscombe$x4, anscombe$y4)
[1] 0.8165214
```

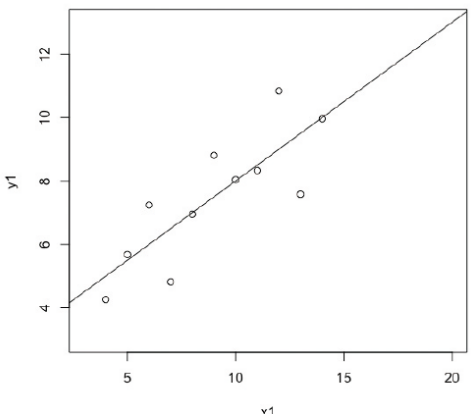
anscombe

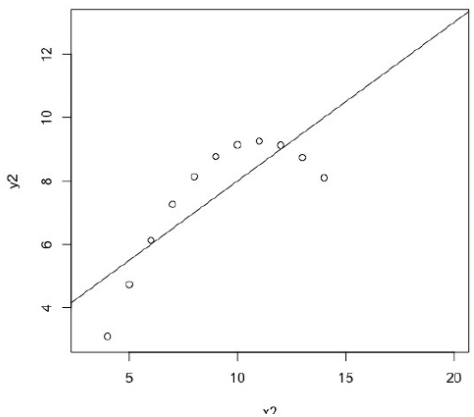
Filter

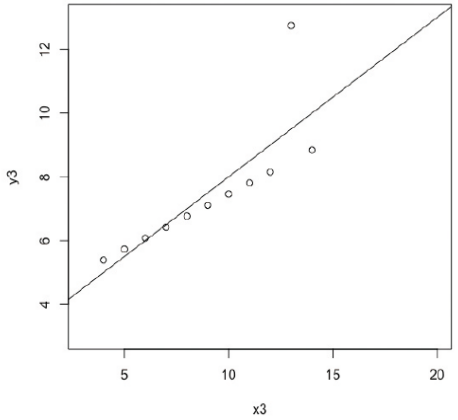
| | x1 | x2 | x3 | x4 | y1 | y2 | y3 | y4 |
|---|---|---|---|---|---|---|---|---|
| 1 | 10 | 10 | 10 | 8 | 8.04 | 9.14 | 7.46 | 6.58 |
| 2 | 8 | 8 | 8 | 8 | 6.95 | 8.14 | 6.77 | 5.76 |
| 3 | 13 | 13 | 13 | 8 | 7.58 | 8.74 | 12.74 | 7.71 |
| 4 | 9 | 9 | 9 | 8 | 8.81 | 8.77 | 7.11 | 8.84 |
| 5 | 11 | 11 | 11 | 8 | 8.33 | 9.26 | 7.81 | 8.47 |
| 6 | 14 | 14 | 14 | 8 | 9.96 | 8.10 | 8.84 | 7.04 |
| 7 | 6 | 6 | 6 | 8 | 7.24 | 6.13 | 6.08 | 5.25 |
| 8 | 4 | 4 | 4 | 19 | 4.26 | 3.10 | 5.39 | 12.50 |
| 9 | 12 | 12 | 12 | 8 | 10.84 | 9.13 | 8.15 | 5.56 |
| 10 | 7 | 7 | 7 | 8 | 4.82 | 7.26 | 6.42 | 7.91 |
| 11 | 5 | 5 | 5 | 8 | 5.68 | 4.74 | 5.73 | 6.89 |

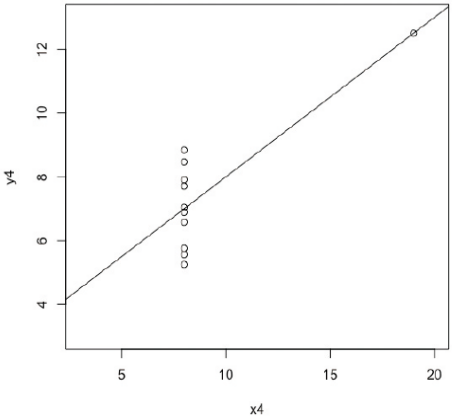
앤스콤의 4분할 데이터 셋

4개의 데이터 셋
data1=(x1, y1)
data2=(x2, y2)
data3=(x3, y3)
data4=(x4, y4)



# 데이터 분석기법-기초

## 데이터의 분포 파악

- Histogram
- Boxplot
- 평균값과 분산

## 2개 양적 데이터의 관계 파악

- 산포도(scatter plot)
- 상관계수(correlation coefficient)