# 통계분석

Statistical Analysis

# Data (자료)

# Data (데이터, 자료)

- 자료(Data) = “문제 해결을 위한 원재료로서 **처리되지 않은** 문자나 숫자 또는 일련의 사실이나 기록들의 **모임**”

  Collections of unprocessed text, numbers, a series of facts, or documents.

- 원소(element) = “자료를 이루는 기본 구성 단위” 혹은 최소 단위

  Basic single or minimal unit that constitutes data

- 모집단(population) = “관심의 대상이 되는 **모든** 원소들의 집합”

  A set of all items or events which is of interest for some experiment

# Variables (변수, 변인)

- Any characteristic or feature that can vary element by element in a population
- The characteristic can be either qualitative or quantitative.
- 모집단의 원소마다 다를 수 있는 속성들을 **변수** 혹은 **변인**이라 한다.
- 변수는 수치적인 (혹은 양적)변수일 수도 있고 질적 변수도 가능하다.

- Data = a set of variables measured in a set of elements of our interest
- 자료는 관심대상들을 상대로 측정한 변수값 혹은 변인값들의 모임이다.

- 자료가 한 종류의 변수를 가지는 경우, 단일 변수(univariate variable)를 가진다고 한다.
- 반면, 자료가 두 종류의 변수로 구성되어 있다면, 이원 변수 (bivariate variable)를 가진다고 한다.
- 이를 일반화하면, 두 종류 이상의 변수를 가지는 자료는 다원 변수(multivariate variable)를 가진다고 한다.

햄버거의 중량, 칼로리, 고기함량 등으로 측정을 한다면
이 자료는 다변량자료이다.

# 변수의 종류: Kinds of Variables

- The characteristic can be either qualitative or quantitative.
- 변수는 **수치적인 (혹은 양적)변수**일 수도 있고 **질적 변수**도 가능하다.

| 질적 변수 (qualitative variables)<br>질적 자료 (qualitative data) | 양적 변수 (quantitative variables)<br>양적 자료 (quantitative data) |
|---|---|
| 대통령 선거에서 투표를 한 후보자가 누구인가?<br>Which candidate do you elect in the presidency election? | 2018년 대전에서 팔린 빅맥의 무게<br>Weight of Big Mac sold in Daejeon 2018 |
| • 막대그래프: Bar chart<br>• 파이 차트: Pie chart | • 도수분포표 (frequency distribution table)<br>• 히스토그램 (histogram)<br>• 꺾은선그래프(frequency polygon) |

# 양적 변수: 이산형 vs 연속형

양적 변수 (quantitative variables)
양적 자료 (quantitative data)

이산적 변수 (discrete variables)

2010년에 입학한 UST학생의 나이

Ages of students who entered UST 2010

연속적 변수 (continuous variables)

2018년 대전에서 팔린 빅맥의 무게 (g)

Weight of Big Mac sold in Daejeon 2018

# 모집단과 표본
# Population and Sample

# Population (모집단)

- 모집단(population) = “관심의 대상이 되는 **모든** 원소들의 집합”

  A set of all items or events which is of interest for some experiment

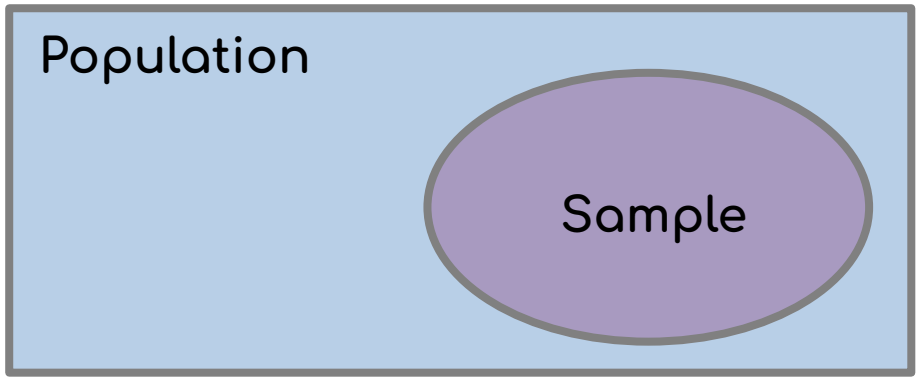
  - Set of all stars within the Milky Way galaxy
  - All individuals who received a B.S. in engineering last year
  - All automobiles sold in Korea last year
  - All burgers sold at McDonald’s in Daejeon

# Population (모집단) and Sample (표본)

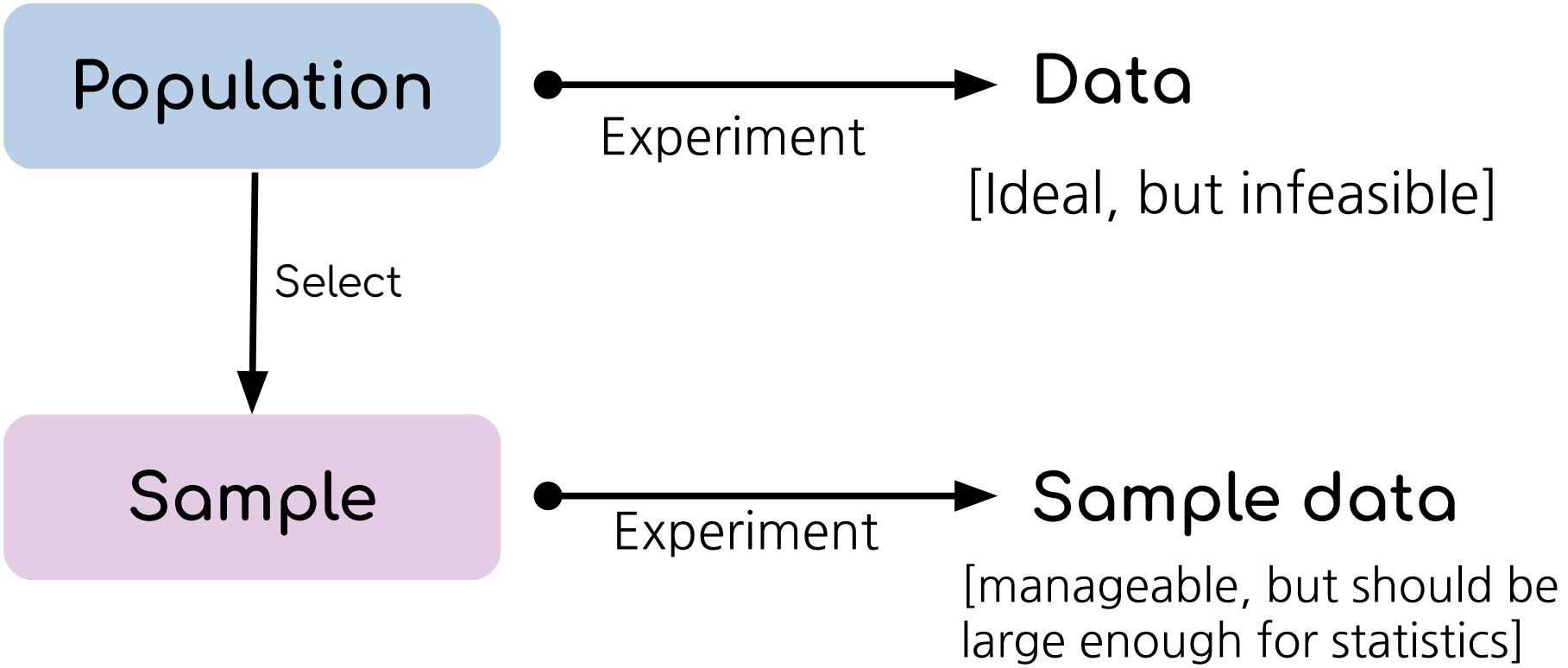
- It would be better to collect data or information from a population of our interest. But, due to limitations on time, money, and other resources, the **"population survey" (전수조사)** is not the ideal one.

- Instead, some part (subset) of the population is collected, and related survey is performed on the subset. Such a subset is called a **sample (표본)**, and the survey on the sample is the **"sample survey" (표본조사).**

- **Sample =** A subset of the population, on which real experiments or surveys will be conducted.



# Population (모집단) and Sample (표본)



# 기술통계 : Descriptive Statistics

Statistical techniques to describe population or sample data qualitatively and quantitatively by using numbers or graphs

모집단 혹은 표본 자료를 질적 혹은 양적으로 기술하기 위하여 수치나 그래프 등을 이용하여 나타내는 것

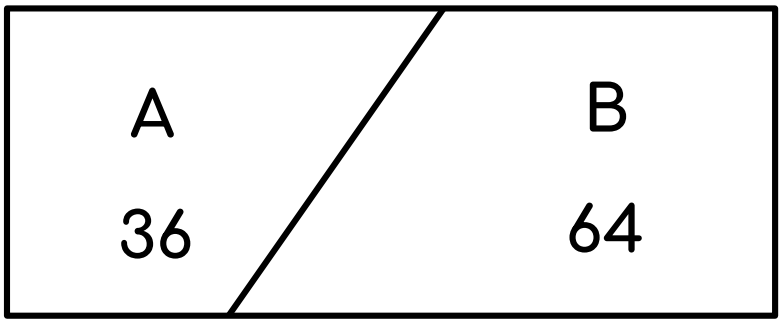
# Frequency

- **도수(frequency)**

  The number of elements included in a category
  특정 범주 혹은 구간에 속하는 원소들의 수

- **상대도수(relative frequency)**

  frequency divided by the total number of elements in data




| | | |
|---|---|---|
| A | 36 | 0.36 |
| B | 64 | 0.64 |

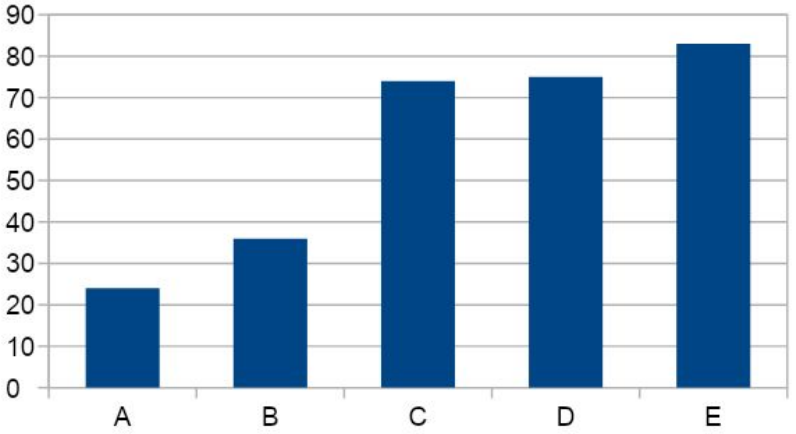
# Cumulative Frequency Distribution Table

| Range | Frequency | Relative Frequency | Percentage (%) | Cumulative Frequency | Relative Cumulative Frequency | Cumulative Percentage (%) |
|---|---|---|---|---|---|---|
| 0 ~ 10 | 1 | 0.02 | 2 | 1 | 0.02 | 2 |
| 10 ~ 20 | 1 | 0.02 | 2 | 2 | 0.04 | 4 |
| 20 ~ 30 | 2 | 0.04 | 4 | 4 | 0.08 | 8 |
| 30 ~ 40 | 2 | 0.04 | 4 | 6 | 0.12 | 12 |
| 40 ~ 50 | 4 | 0.08 | 8 | 10 | 0.2 | 20 |
| 50 ~ 60 | 7 | 0.14 | 14 | 17 | 0.34 | 34 |
| 60 ~ 70 | 10 | 0.2 | 20 | 27 | 0.54 | 54 |
| 70 ~ 80 | 13 | 0.26 | 26 | 40 | 0.8 | 80 |
| 80 ~ 90 | 7 | 0.14 | 14 | 47 | 0.94 | 94 |
| 90 ~ 100 | 3 | 0.06 | 6 | 50 | 1 | 100 |

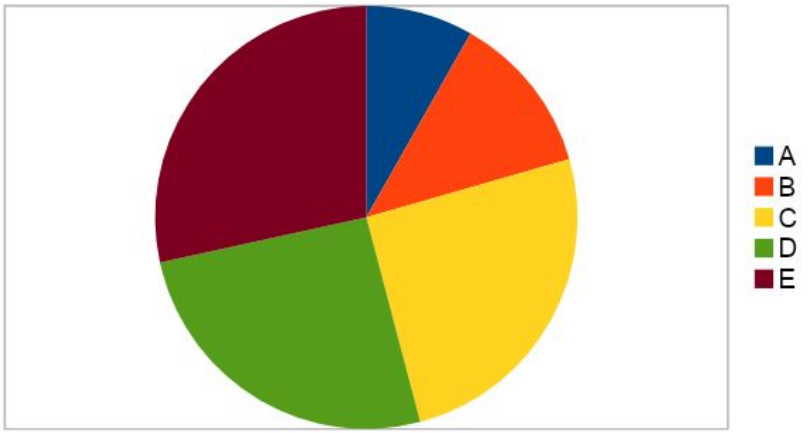
# Data Visualization Using Graphs

| Category | Frequency |
|---|---|
| A | 24 |
| B | 36 |
| C | 74 |
| D | 75 |
| E | 83 |

- 막대 그래프 (Bar Chart)



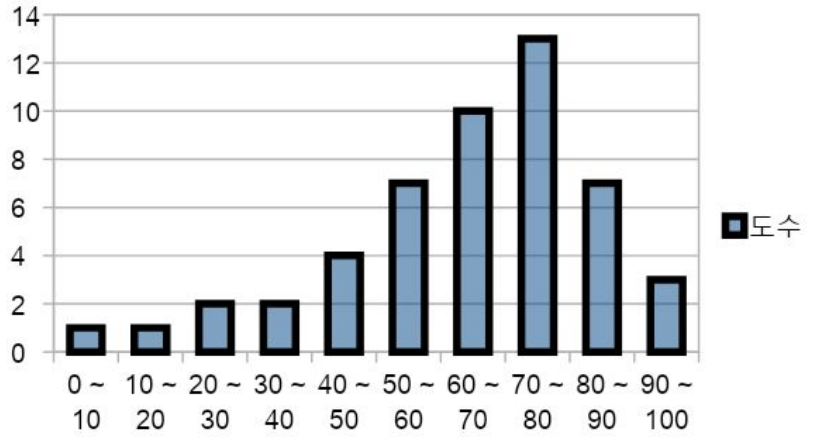

- 원 그래프 (Pie Chart)



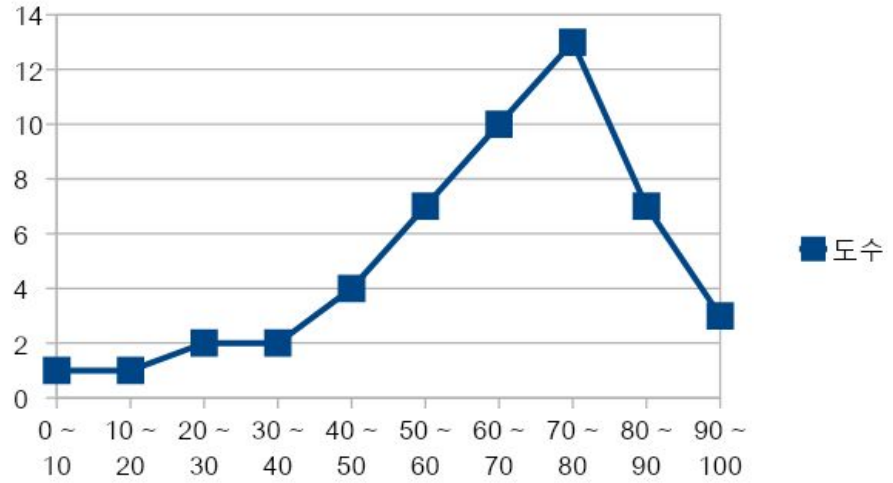
# Data Visualization Using Graphs

| Range | Frequency |
|---|---|
| 0 ~ 10 | 1 |
| 10 ~ 20 | 1 |
| 20 ~ 30 | 2 |
| 30 ~ 40 | 2 |
| 40 ~ 50 | 4 |
| 50 ~ 60 | 7 |
| 60 ~ 70 | 10 |
| 70 ~ 80 | 13 |
| 80 ~ 90 | 7 |
| 90 ~ 100 | 3 |

- 히스토그램 (Histogram)




- 꺾은선 그래프 (Line Graph)

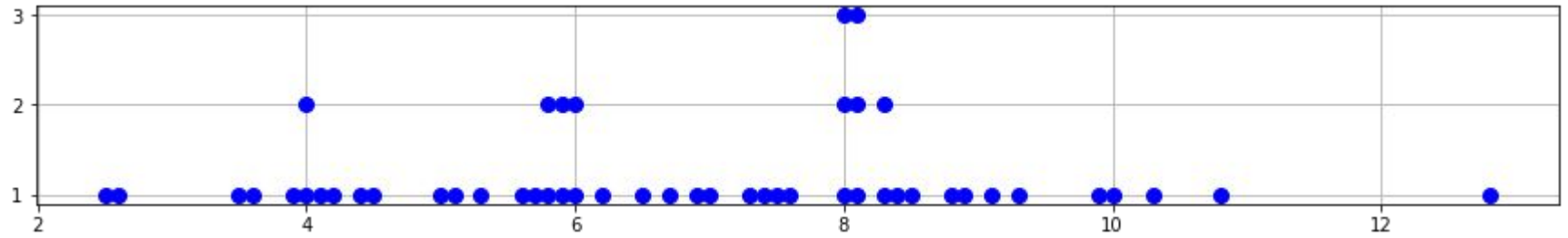


# Dot Plot : 점 도표

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 10.8 | 6.9 | 8.0 | 8.8 | 7.3 | 3.6 | 4.1 | 6.0 | 4.4 | 8.3 |
| 8.1 | 8.0 | 5.9 | 5.9 | 7.6 | 8.9 | 8.5 | 8.1 | 4.2 | 5.7 |
| 4.0 | 6.7 | 5.8 | 9.9 | 5.6 | 5.8 | 9.3 | 6.2 | 2.5 | 4.5 |
| 12.8 | 3.5 | 10.0 | 9.1 | 5.0 | 8.1 | 5.3 | 3.9 | 4.0 | 8.0 |
| 7.4 | 7.5 | 8.4 | 8.3 | 2.6 | 5.1 | 6.0 | 7.0 | 6.5 | 10.3 |

→

2.5, 2.6, 3.5
3.6, 3.9
4.0, 4.0, 4.1, 4.2, 4.4, 4.5
5.0, 5.1, 5.3, 5.6, 5.7, 5.8, 5.8, 5.9, 5.9
6.0, 6.0, 6.2, 6.5, 6.7, 6.9
7.0, 7.3, 7.4, 7.5, 7.6
8.0, 8.0, 8.0, 8.1, 8.1, 8.1, 8.3, 8.3, 8.4, 8.5, 8.8, 8.9
9.1, 9.3, 9.9
10.0, 10.3, 10.8
12.8



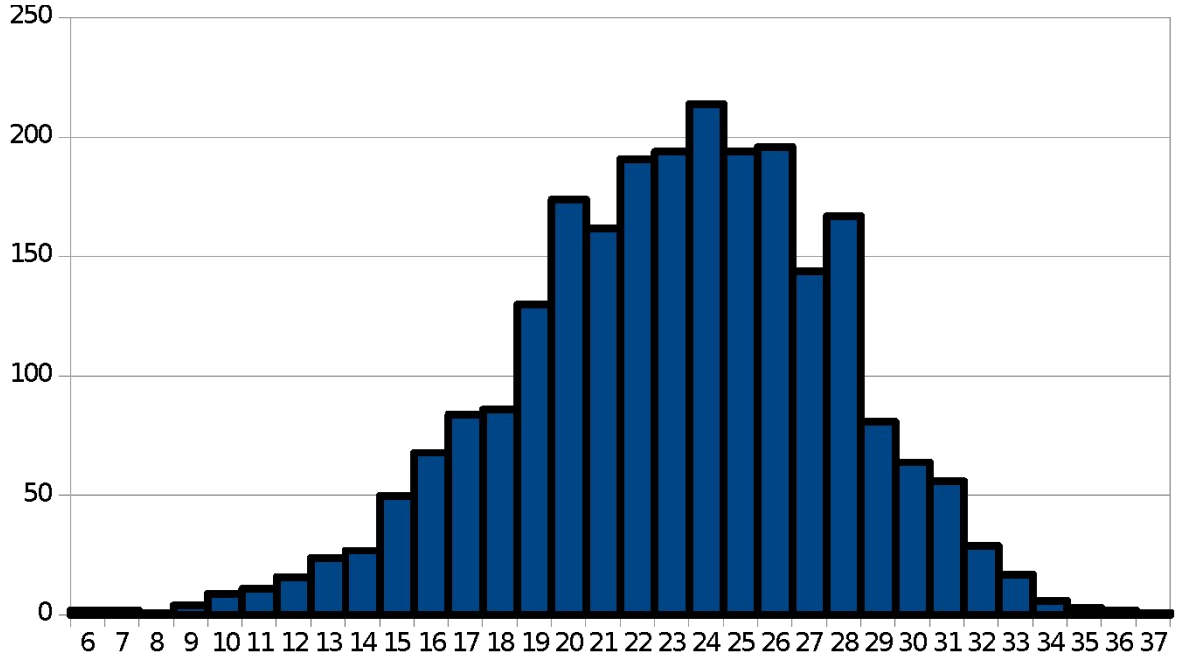

Dot plots:

- When a value occurs more than once, there is a dot for each occurrence, and these dots are stacked vertically.
- As with a stem-and-leaf display, a dot plot gives information about **location, spread, extremes, and gaps**.

# 기술통계 : Descriptive Statistics

- 중심경향도 (집중경향도) : Central Tendency
- 산포도 : Measure of Dispersion
- 왜도 및 첨도 : Skewness and Kurtosis

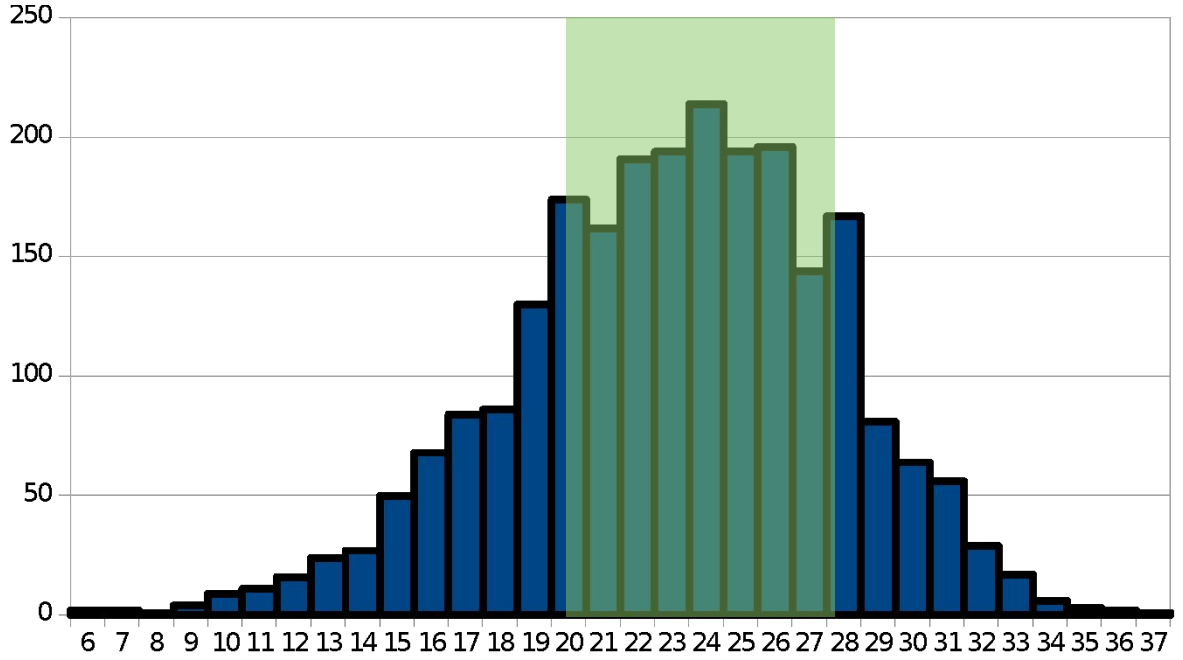
# How to express data in terms of numbers



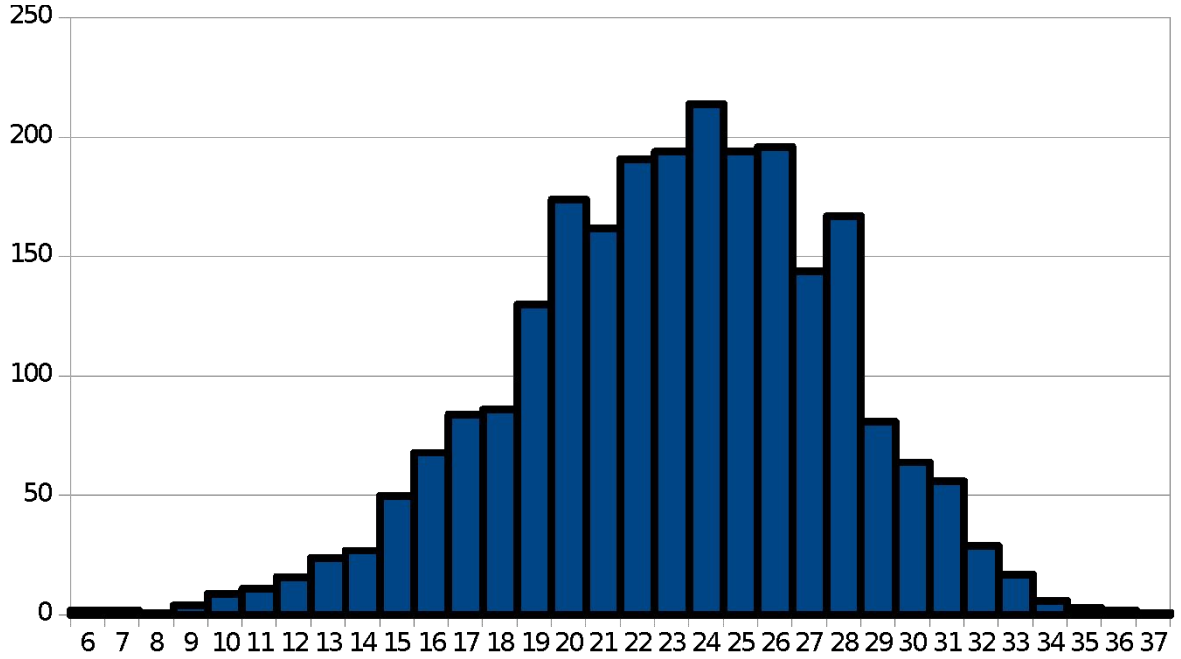

- 그래프에서 자료가 어떻게 분포하고 있는가?
- 그래프로 표현한 자료를 몇 개의 수치값으로 나타낼 수 있는가?
- 자료가 어디에 분포하고 있는가?: 자료의 분포 위치(location), 자료의 분포 범위 (range) 등등
- 자료의 분포 형태는 어떠한가?

# 집중경향도(중심경향도)
# Measure of Central Tendency
# Measure of Location

# 집중경향도: Measure of Central Tendency

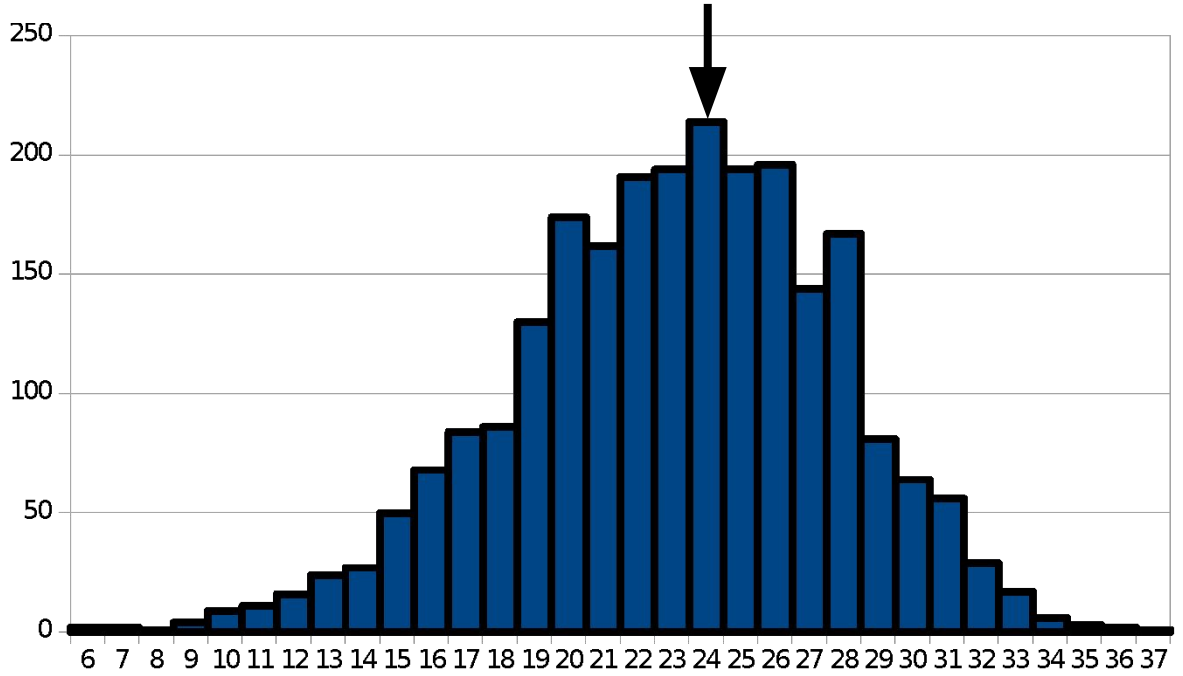


# 집중경향도: Measure of Central Tendency




- **최빈값 (mode)**
  - 가장 도수가 높은 측정치
  - a data point with the highest frequency.
- **중앙값 (median)**
- - 자료를 서열대로 나열하였을 때, 중앙에 위치하는 값
  - a data point located at the center when data is ordered from the smallest to the largest
- **평균값 (mean)**
- - 자료들의 무게 중심에 해당하는 값 또는 산술적 평균에 해당하는 값
  - an arithmetic average of a data set

# 집중경향도 I: 최빈값 (mode)

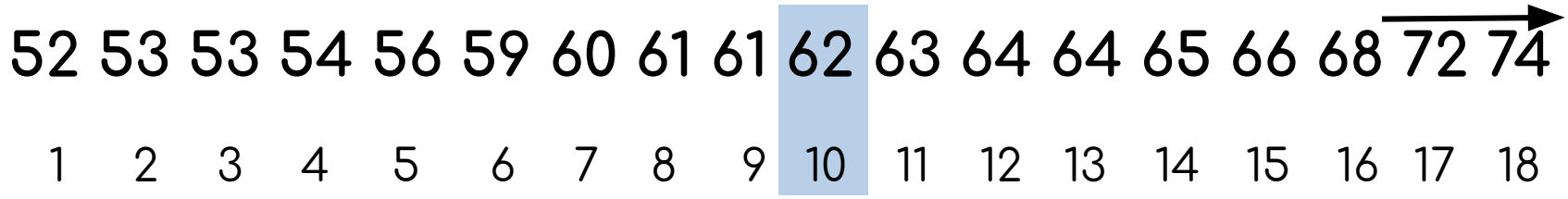



- **최빈값 (mode)**

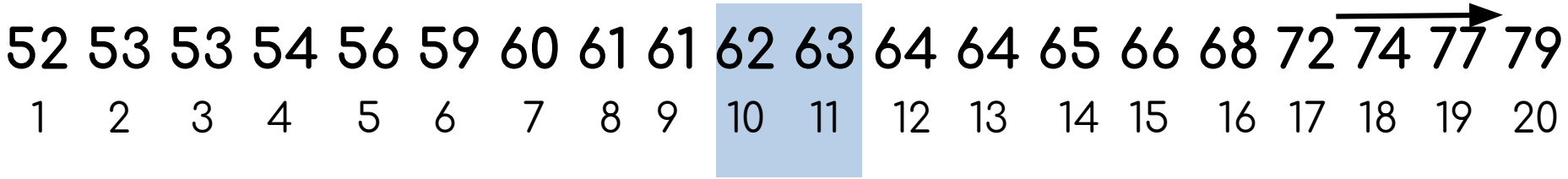
  - 가장 도수가 높은 측정치 : a data point with the highest frequency.

- Data can have more than one mode.

# 집중경향도II : 중앙값 (median)




중앙값(median) = 62




중앙값(median) = (62+63)/2 = 62.5

# 집중경향도III : 평균값 (mean)

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{: Population Mean}$$

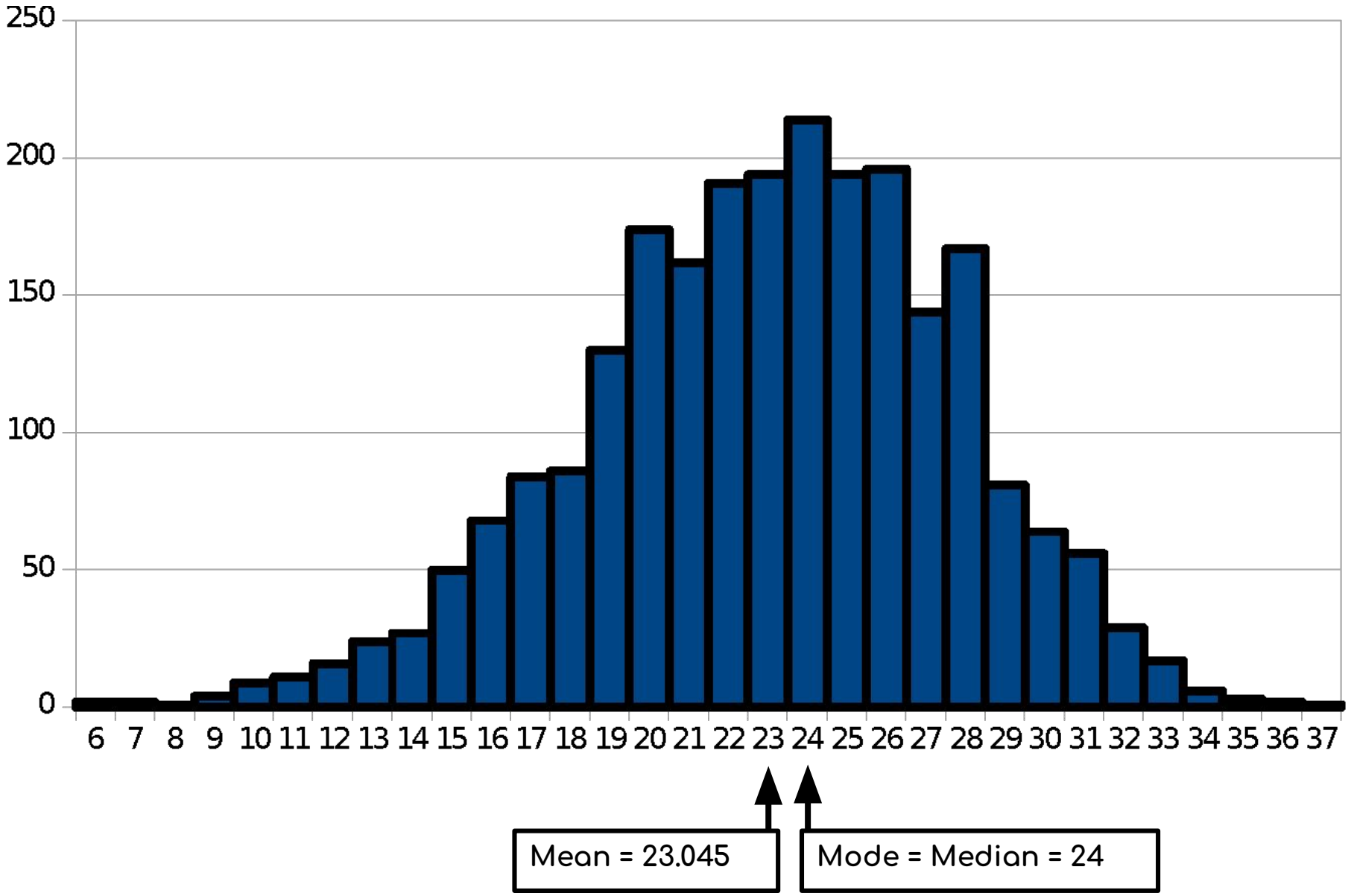
$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{: Sample Mean}$$

- **평균값 (mean)**
  - 자료들의 무게 중심에 해당하는 값 또는 산술적 평균에 해당하는 값
  - an arithmetic average of a data set

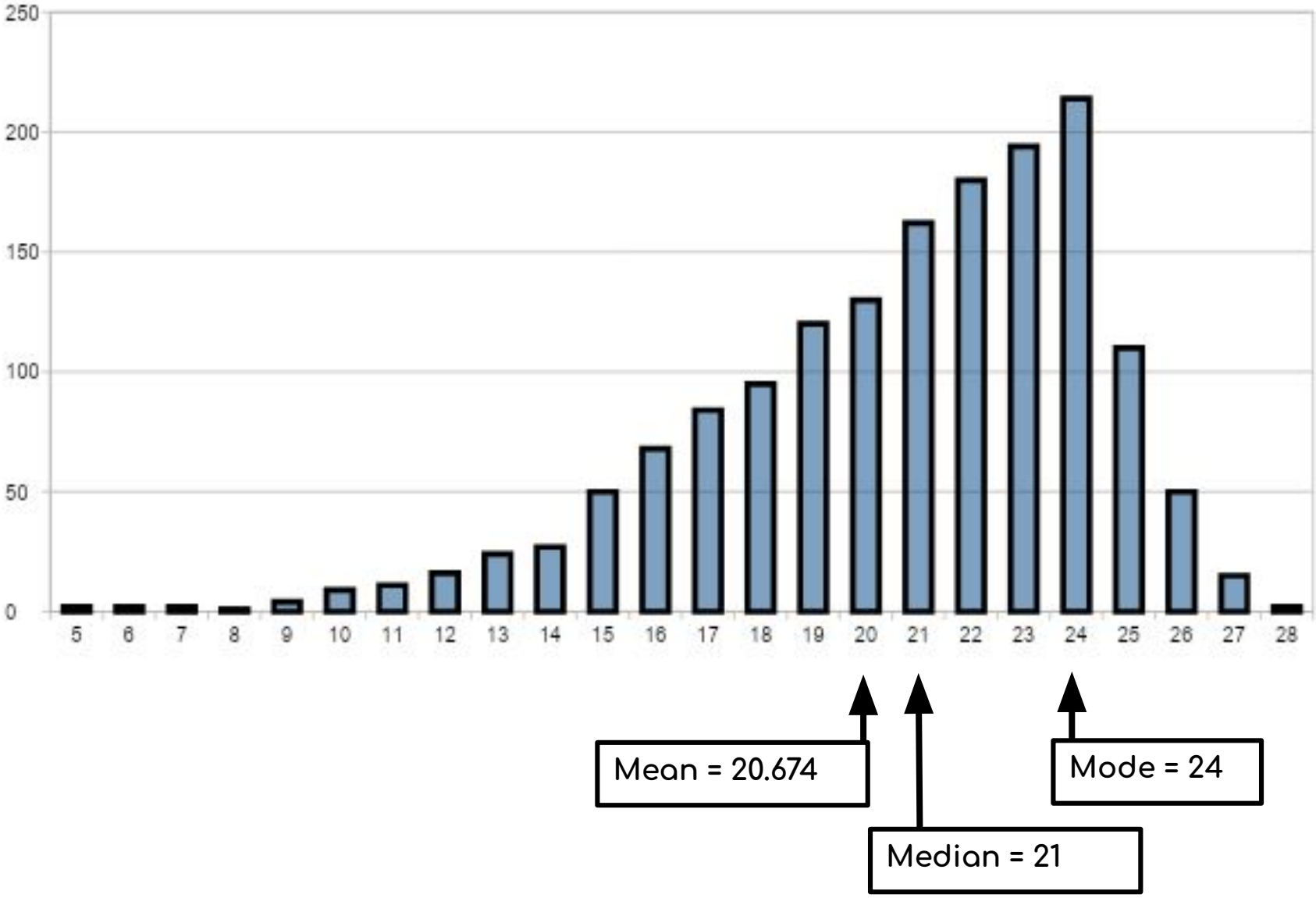
$\mu, N$ : (Population)

$\bar{x}, n$ : (Sample)

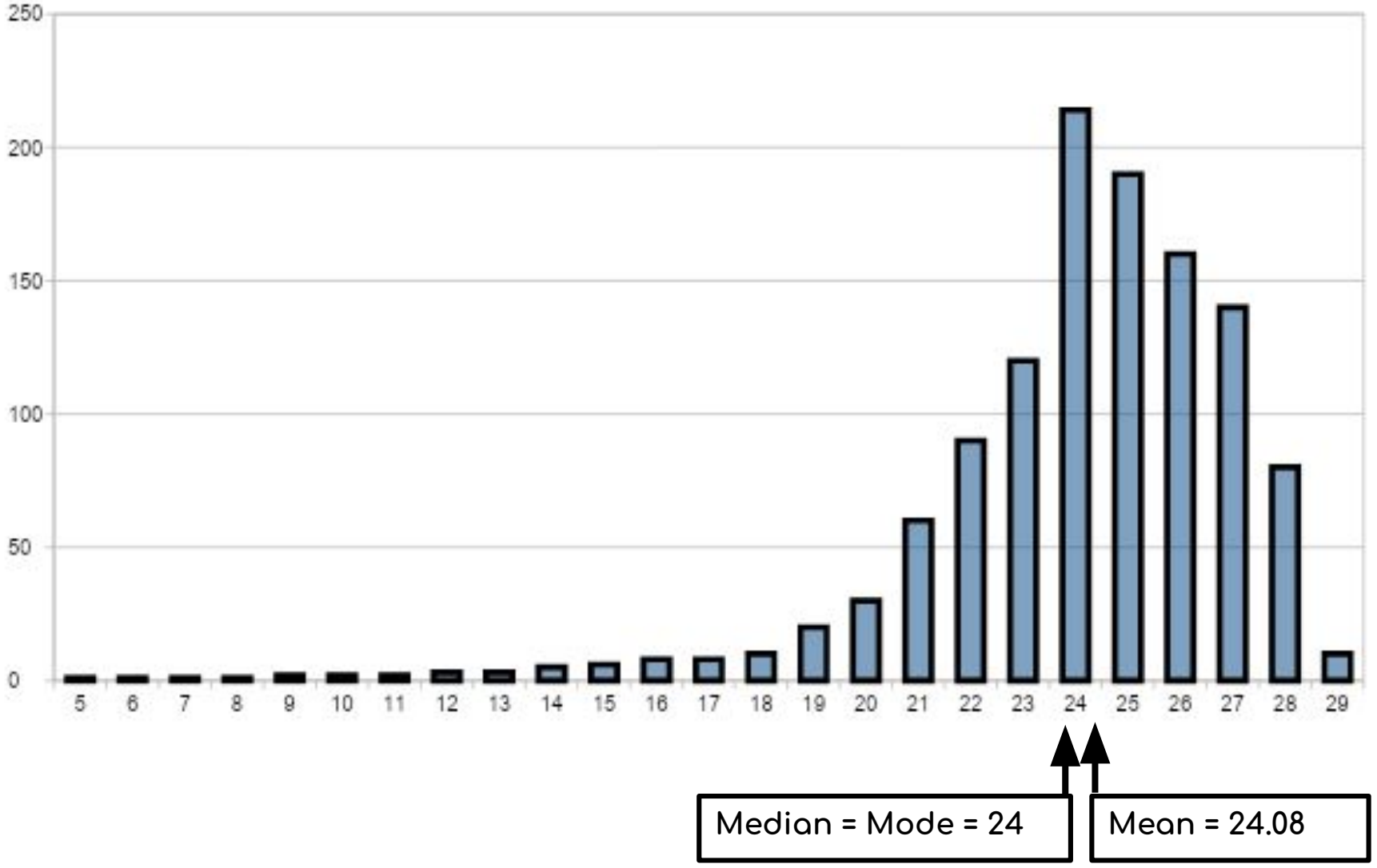
# 집중경향도: Example I



# 집중경향도: Example II

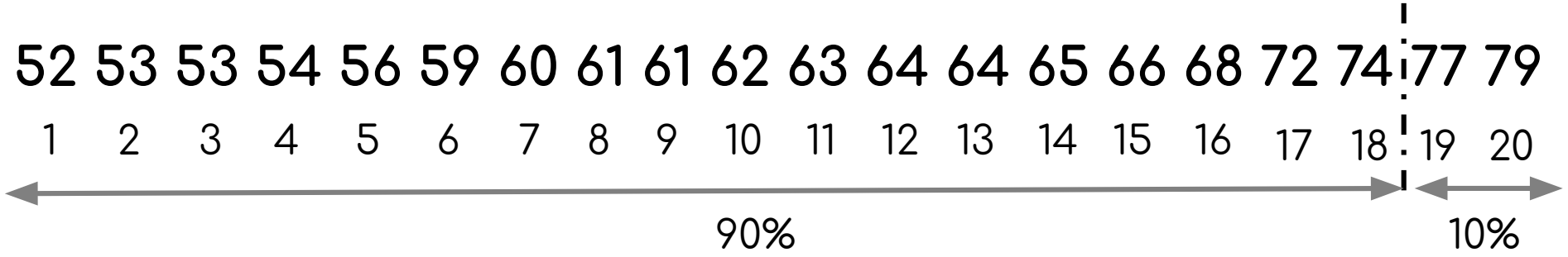


# 집중경향도: Example III



# Other Measures of Location

# Percentiles (백분위수)

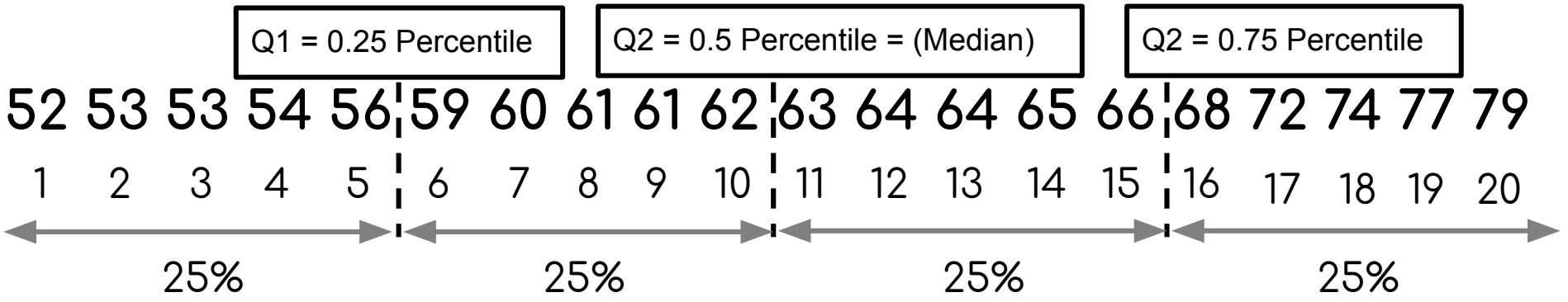



- Q: When data is sorted in terms of magnitude, what is a boundary corresponding to top 10%?

- Data from 52 to 74 correspond to 90% of total data. Data point 74 is a boundary to the cumulative percentage 90%. Here 74 is 90% or 0.9 percentile.

- We can generalize the above example. X percentile is the data point at which the cumulative percentage is X.

# Quartiles (사분위수)



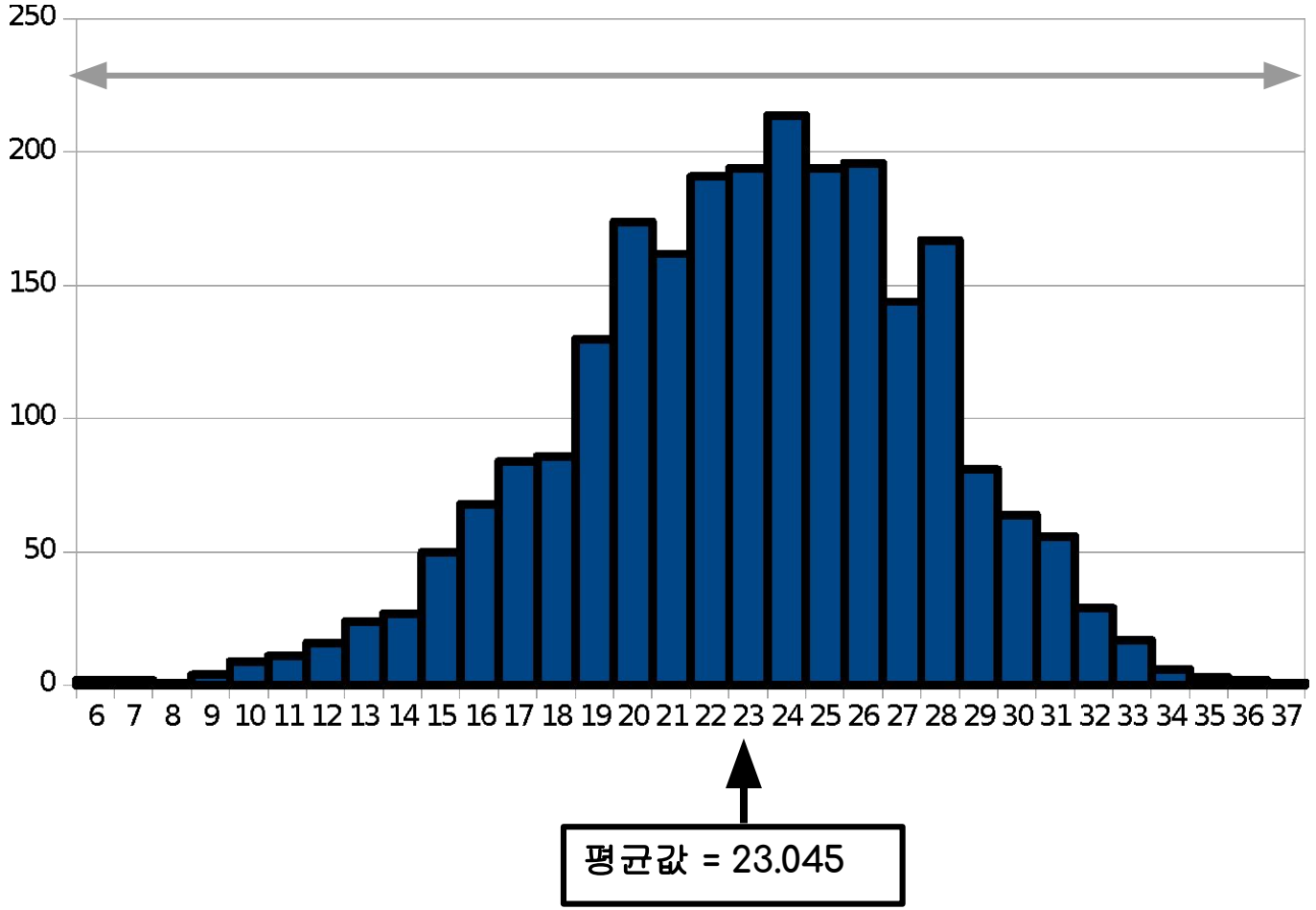

- Quartiles = Data corresponding to cumulative percentage 25%, 50%, and 75%.
- Cumulative percentages 25%, 50%, and 75% are called as the first quartile ($Q_1$), the second quartile ($Q_2$), and the third quartile ($Q_3$), respectively.

$$Q_1 = \text{the 1st Quartile}$$
$$Q_2 = \text{the 2nd Quartile}$$
$$Q_3 = \text{the 3rd Quartile}$$

- The second quartile($Q_2$) is median.

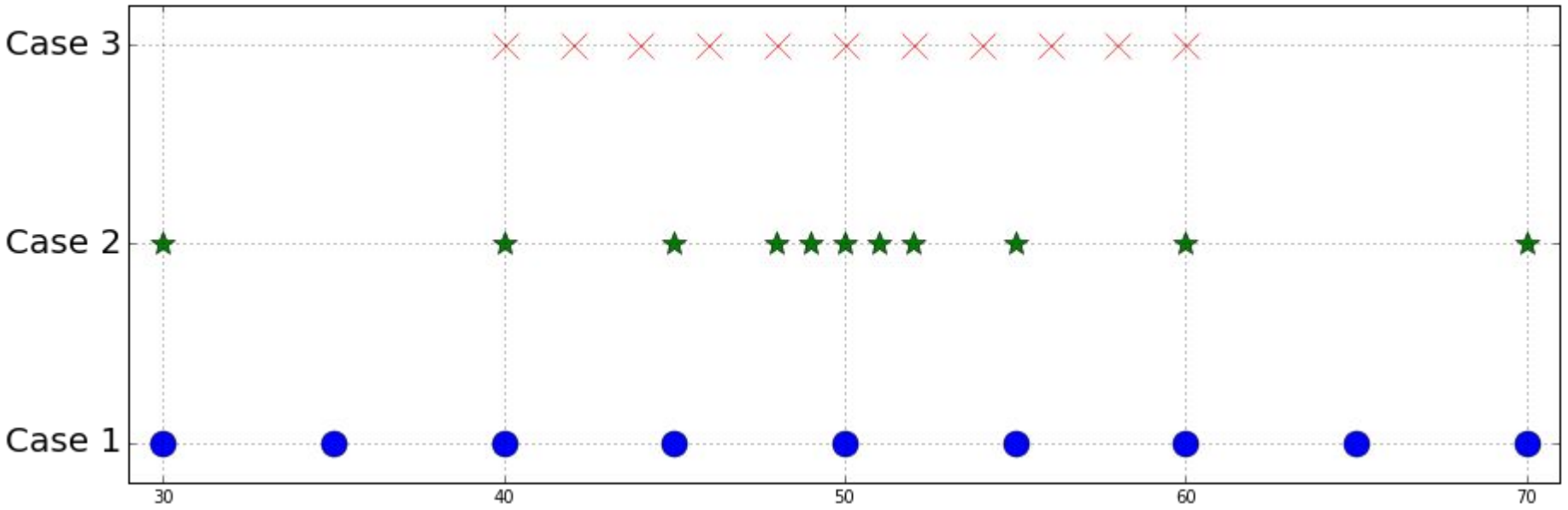
# 산포도
# Measure of Dispersion (Variability)

# Quantity to Describe Data Distribution?



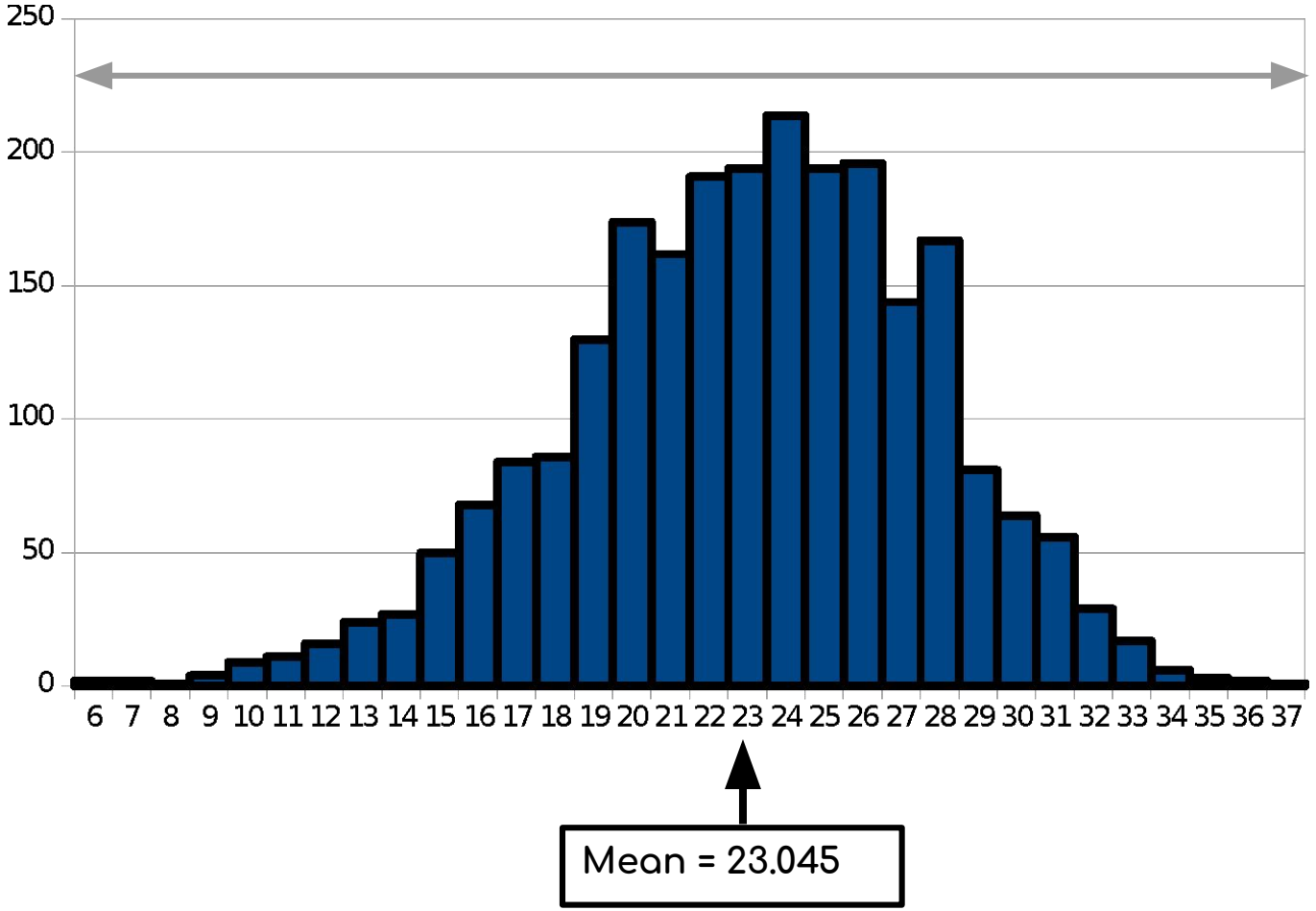

- Mean is one of quantities to represent the central tendency of data, but it cannot describe the range over which total data is distributed.
- 평균값은 자료의 중심을 나타내는 하나의 값이지만 전체 자료가 분포한 범위를 나타내지는 못한다.

# Mean vs Data Distribution



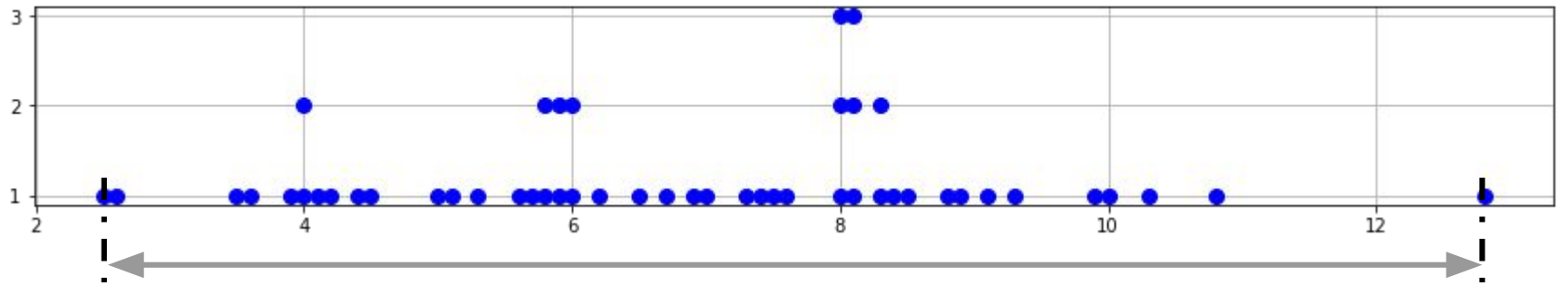

- Three cases have mean 50, but they have different distributions.

# 자료의 범위: Range




- Data Range = (Data Maximum) - (Data Minimum) = 37 - 6 = 31
- 자료의 범위 = (자료 중 최대값) - (자료 중 최소값) = 37 – 6 = 31

# 이상치 (Outlier)

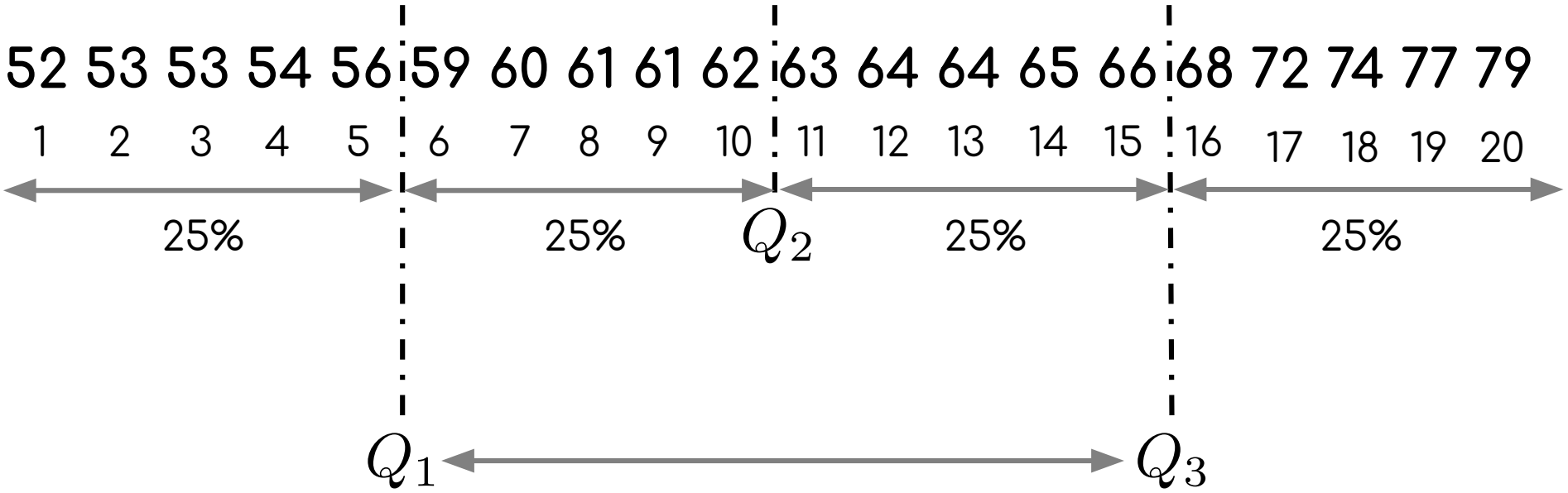



range = 12.8 – 2.5 = 10.3


range = 35.5 – 2.5 = 33.0

- Range can change by adding one data point which distinctly differs the other data.
- Outlier = A data point that significantly differs from other data points.
- Outliers can occur possibly due to measurement errors.

# 사분위 간 범위: Inter-Quartile Range



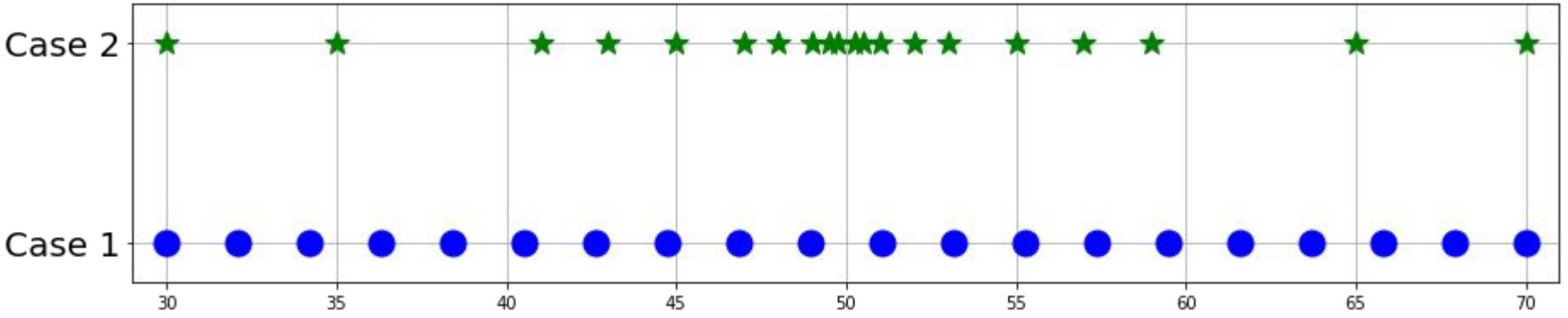

$$\text{Inter-quartile range} = Q_3 - Q_1$$

사분위 간 범위

$$\text{Semi-interquartile range} = \frac{Q_3 - Q_1}{2}$$

사분위편차

# Range vs Inter-Quartile Range

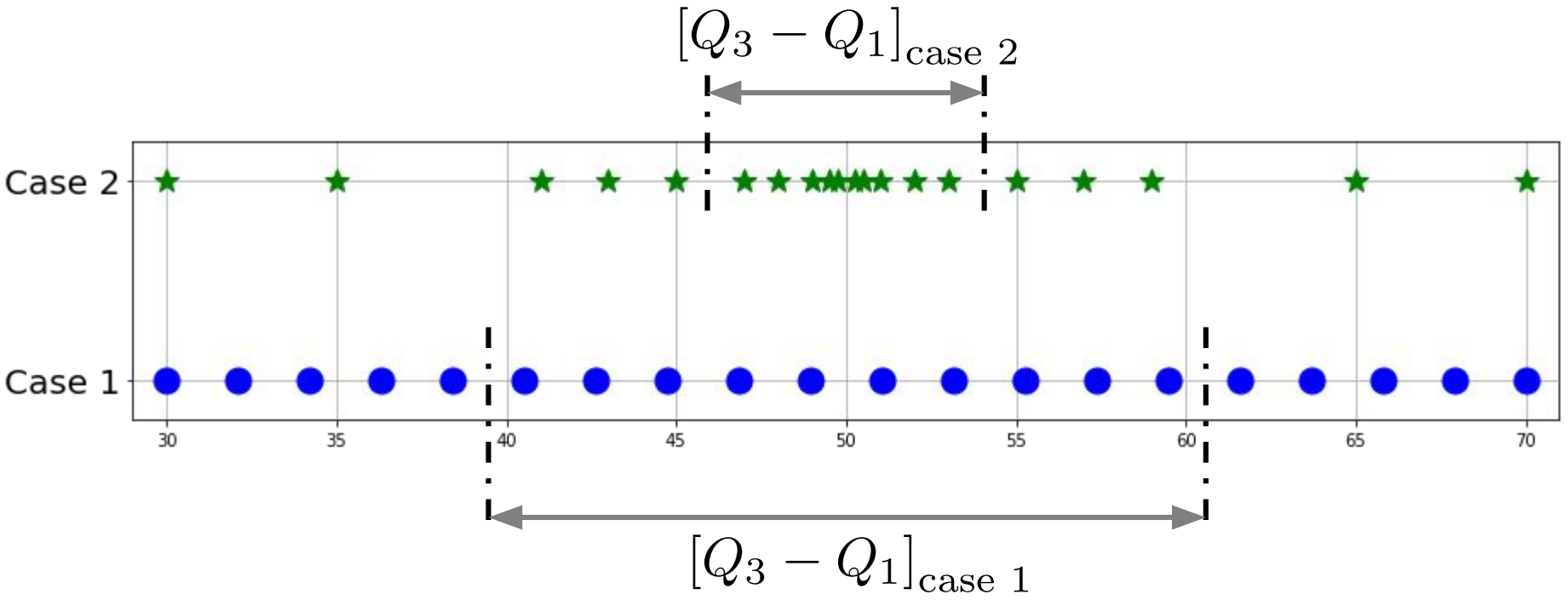



- Two cases have the same range.

$$(\text{range}) = 70 - 30 = 40$$

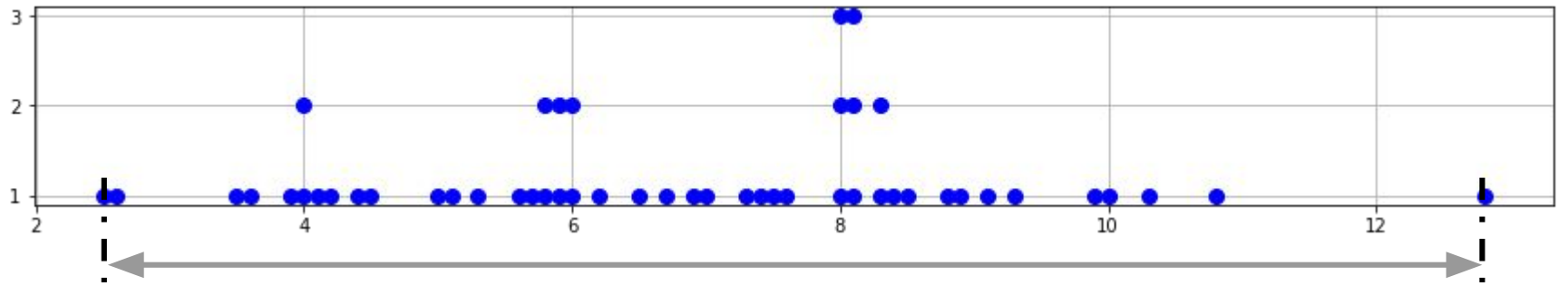
- However, their distributions are different; Case I is uniformly distributed, but the Case II data are concentrated with respect to the median 50.
- Range cannot describe this distribution of data.

# Range vs Inter-Quartile Range

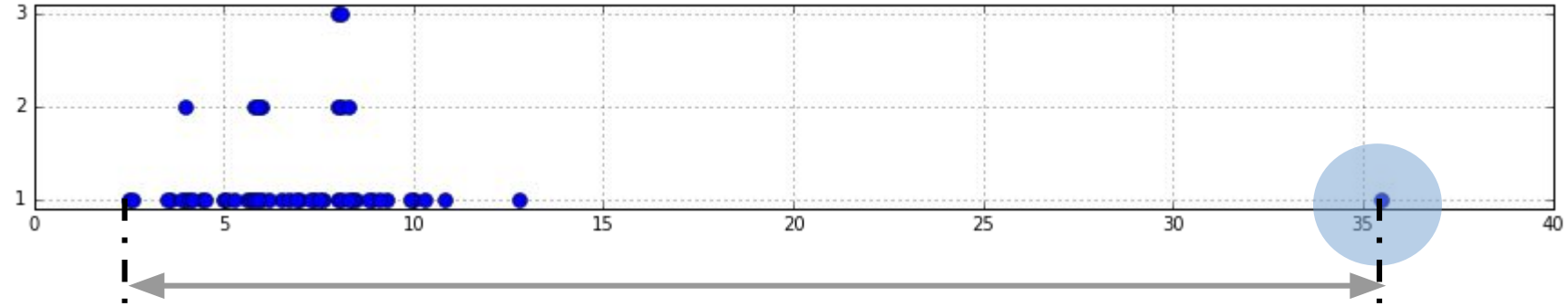



- Two cases have the same range, but their inter-quartile ranges are different.
- For Case 2, data is highly concentrated to the median 50, so the first and third quartiles are located closer to the second quartile.

# Inter-Quartile Range and Outliers



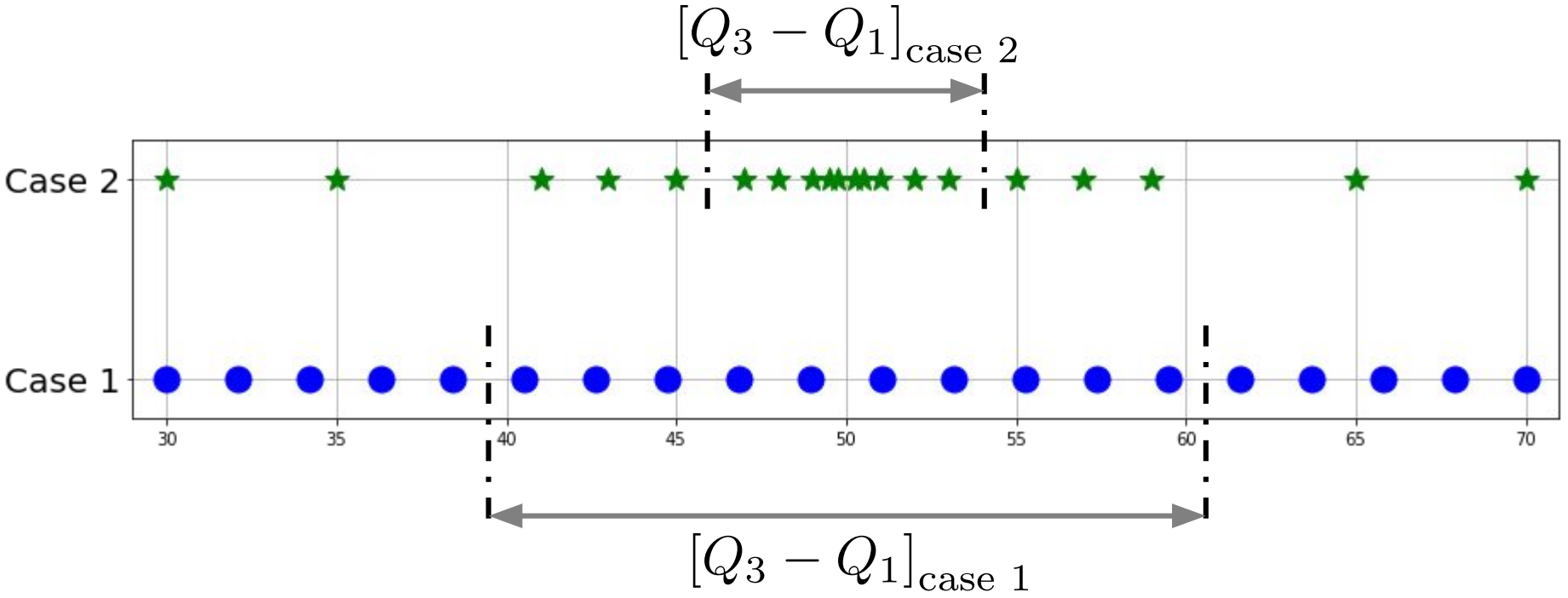

range = 12.8 – 2.5 = 10.3




range = 35.5 – 2.5 = 33.0

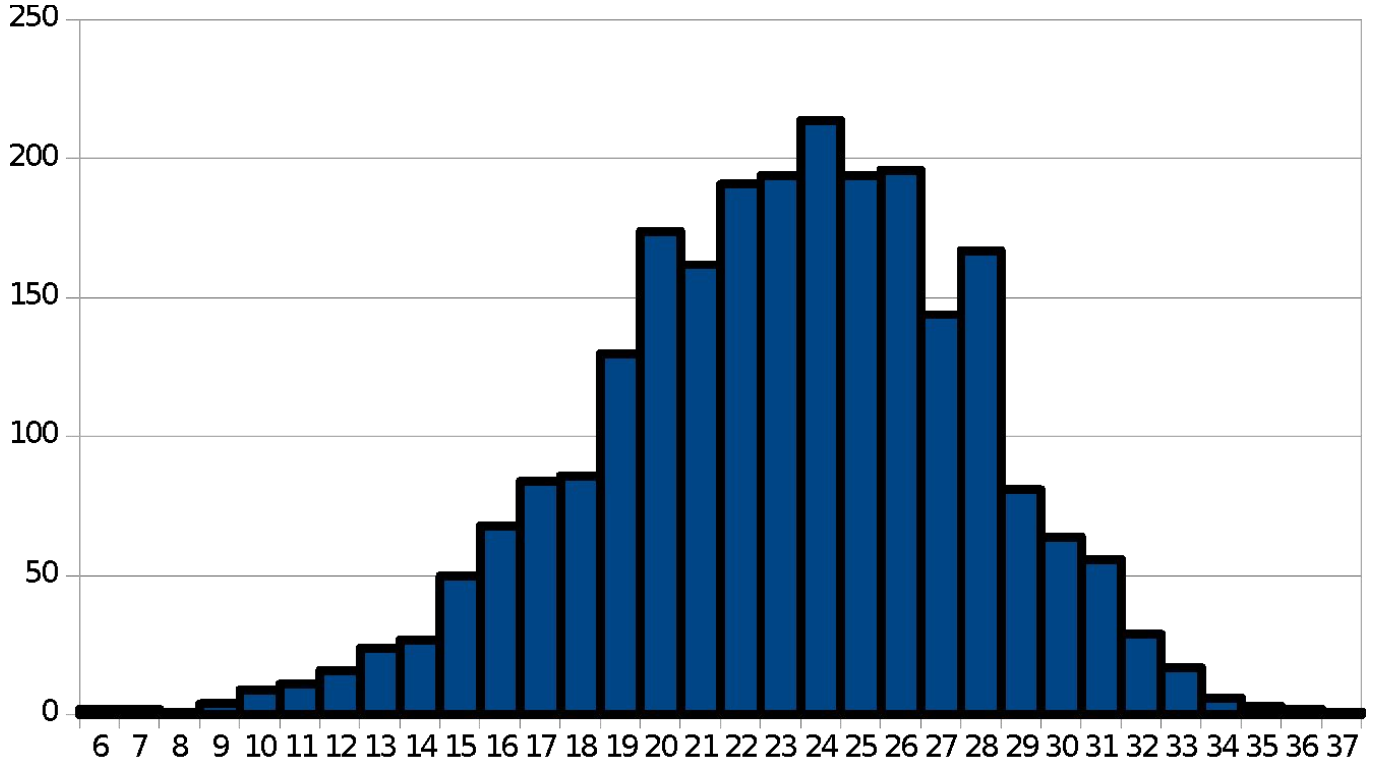
- Even if there is a few outliers, quartiles do not change significantly.
- Outliers are located either at the rightmost or at the leftmost region, which means that they are not included in the inter-quartile range.
- Hence the inter-quartile range is not affected by outliers.

# Range and Inter-Quartile Range




- When it comes to describing data distributions, the inter-quartile range is much better than the range.
- However, the two types of ranges are still defined by two particular data points.
- Is there any other quantity including all data points to describe a data distribution?

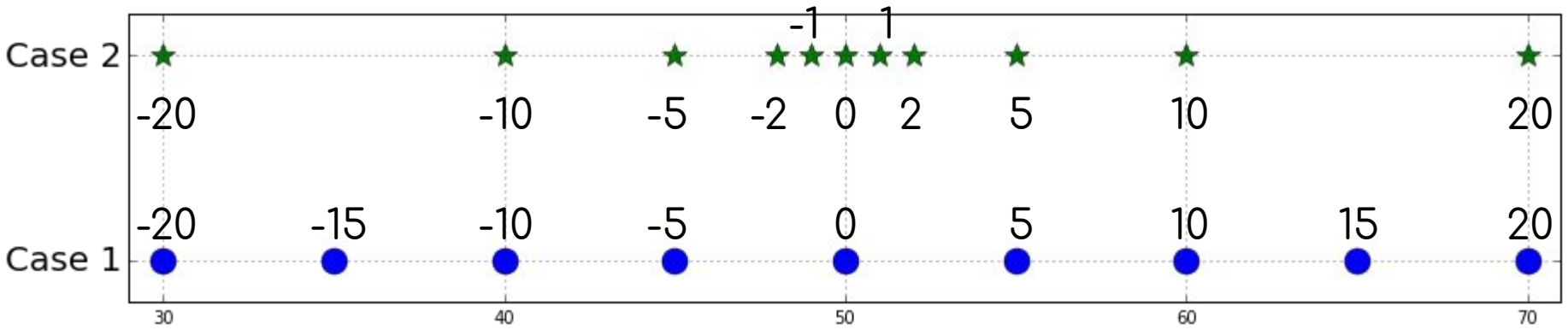
# 자료의 분포: Deviation from Mean




$$\Delta x_i \equiv x_i - \mu$$ [Population]

$$\Delta x_i \equiv x_i - \bar{x}$$ [Sample]

# 자료의 분포: Sum of Deviations



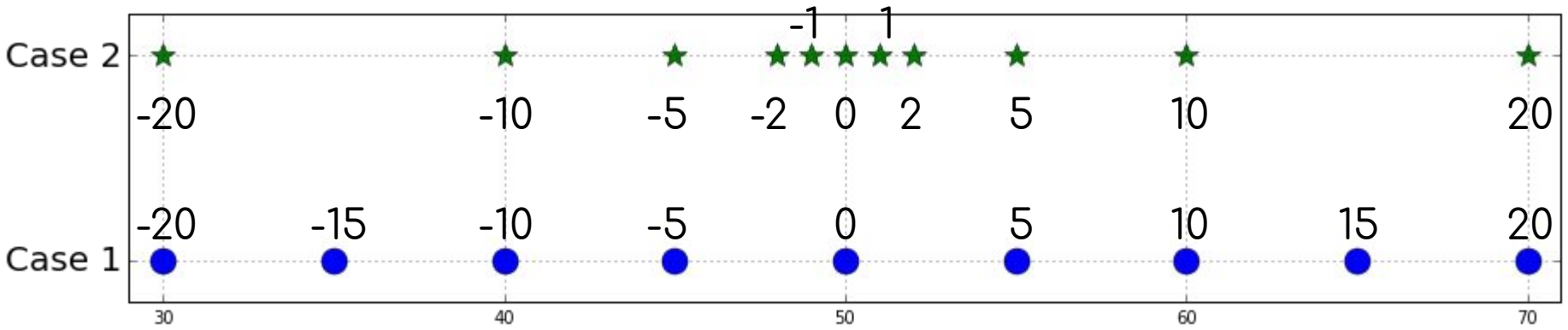

- Can summing deviations from mean describe the distribution of data?

$$\sum_{i=1}^{N} \Delta x_i = \sum_{i=1}^{N} (x_i - \mu) = 0$$

- Sum of deviations is always zero, so it cannot represent the distribution of data.

$$\sum_{i=1}^{N} \mu = N\mu = N\frac{1}{N}\sum_{i=1}^{N} x_i = \sum_{i=1}^{N} x_i$$

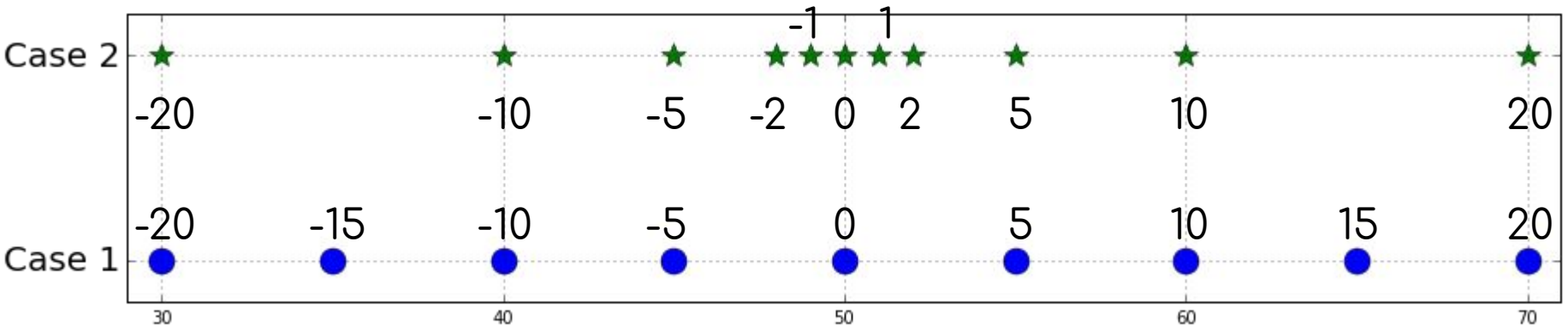
# Absolute deviations vs Squared deviations




- Sum of absolute deviations or Sum of squared deviations?

| | Case 1 | Case 2 |
|---|---|---|
| $\sum_{i=1}^{N} \lvert x_i - \mu \rvert$ | 100 | 76 |
| $\sum_{i=1}^{N} (x_i - \mu)^2$ | 1500 | 1060 |

# 자료의 분포: Deviation from Mean

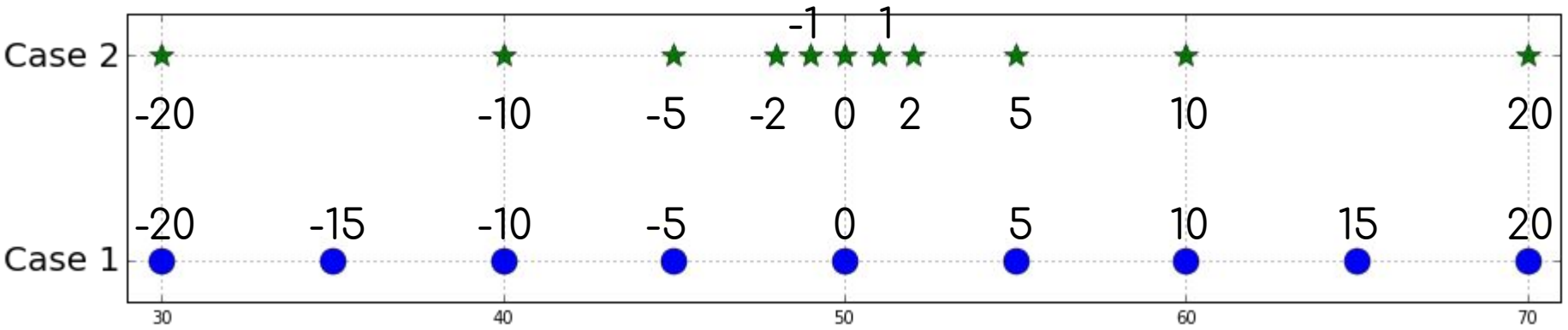



- Sum of absolute deviations or Sum of squared deviations?

$$\sum_{i=1}^{N} |x_i - \mu| \qquad \sum_{i=1}^{N} (x_i - \mu)^2$$

- Absolute deviations are difficult to mathematically manipulate, for example, differentiation.
- Sum of squared deviations has the difference unit against the original data.

# 자료의 분포: 분산 (Variances)




$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$ Population Variance

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$ Sample Variance

- The prefactor of the Sample Variance is different from that of the population variance. Why? It is related to statistical estimation. (Will be discussed in Inference).

# 자료의 분포: 표준편차(Standard Deviation)

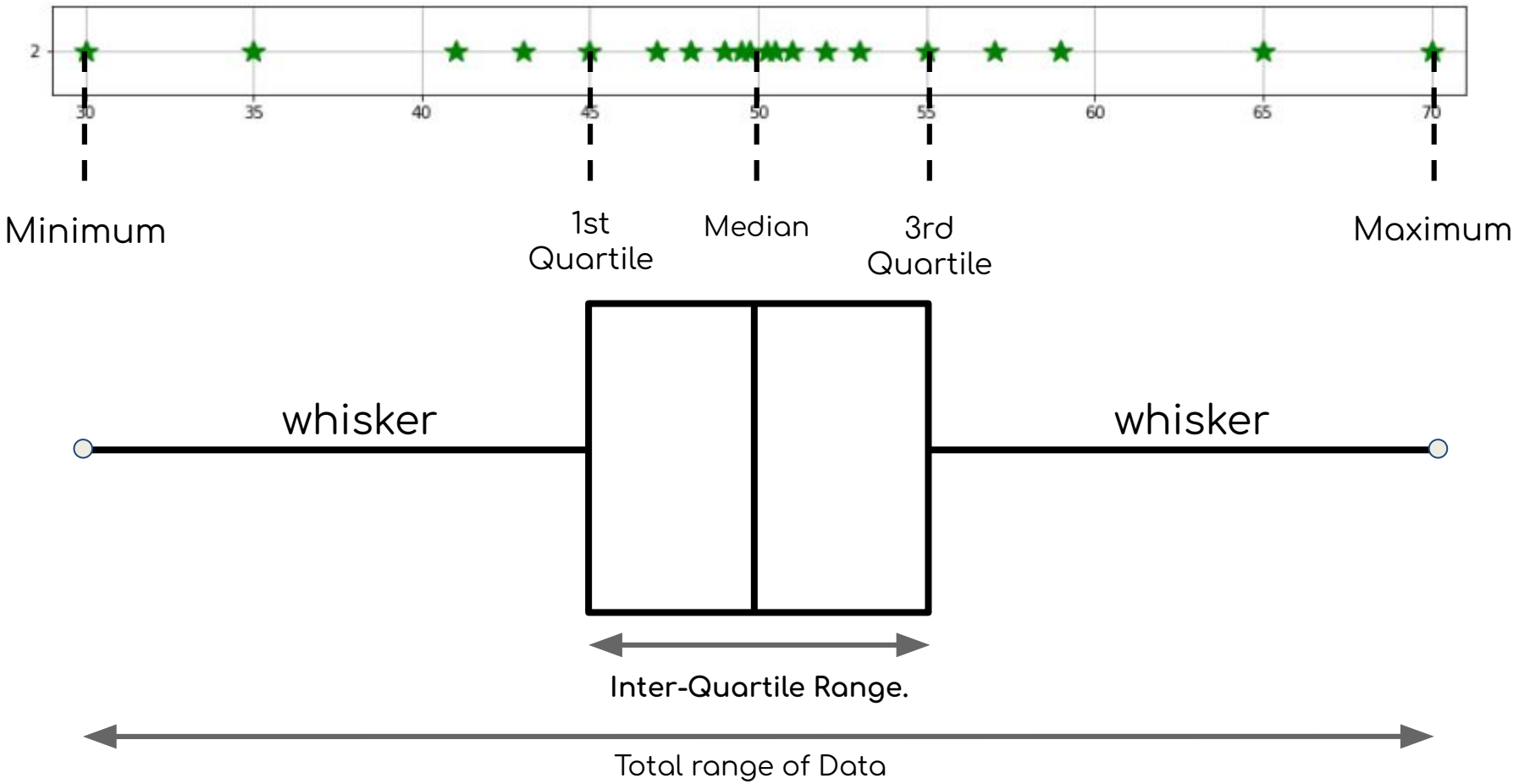
$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Population Standard Deviation

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Sample Standard Deviation

# 자료의 분포: Boxplots




https://en.wikipedia.org/wiki/Box_plot

# 자료의 분포: Boxplots


초
값

en.wikipedia.org/wiki/Whiskers

Article | Talk | Read | Edit | View history | Search Wikipedia

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here

# Whiskers

https://en.wikipedia.org/wiki/Whiskers

From Wikipedia, the free encyclopedia

*"Whisker" redirects here. For other uses, see Whisker (disambiguation).*

**Whiskers** or **vibrissae** (/vaɪbrɪsi/; singular: **vibrissa**; /vaɪbrɪsə/) are a type of mammalian hair that are typically characterised, anatomically, by their large length, large and well-innervated hair follicle, and by having an identifiable representation in the somatosensory cortex of the brain.[1]

They are specialised for tactile sensing (other types of hair operate as more crude tactile sensors). Vibrissae grow in various places on most mammals, including all primates except humans.[2]

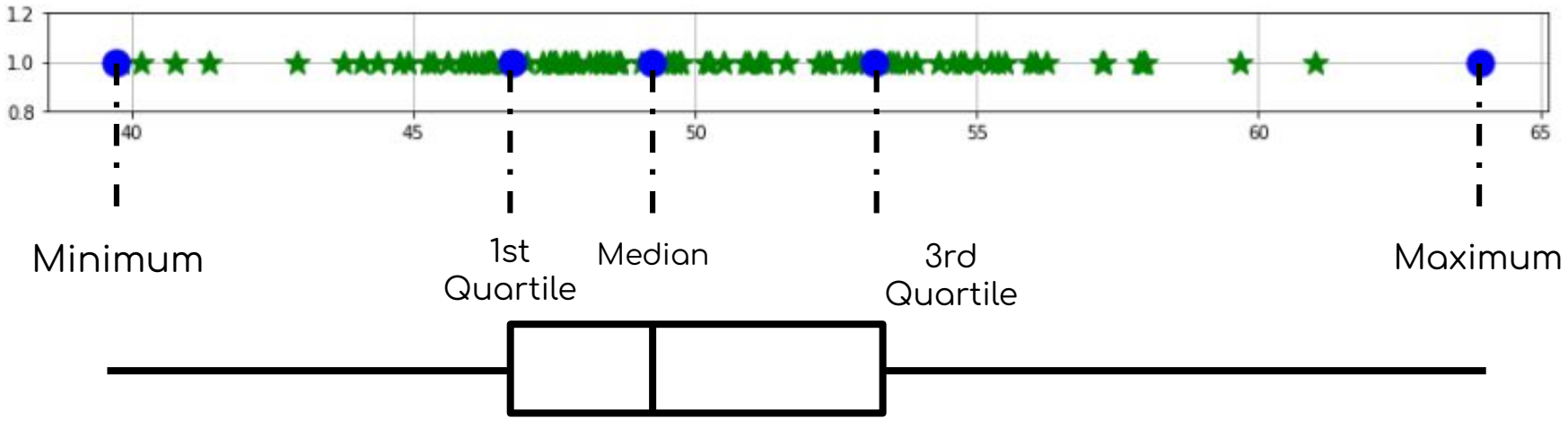
In medicine, the term vibrissae also refers to the thick hairs found inside human nostrils.[3]

**Contents** [hide]

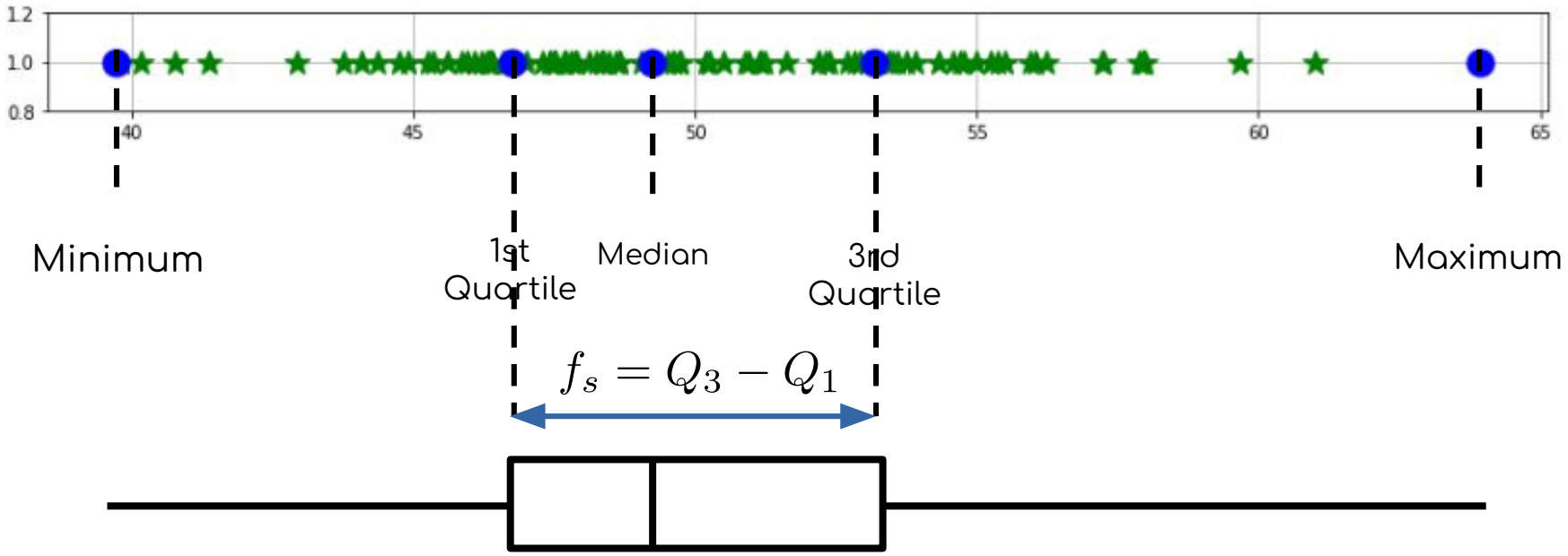
A Patagonian fox showing four major cranial groups of vibrissae: supraorbital (above the eye), mystacial (where a moustache would be), genal (on the cheek, far left), and mandibular (pointing down, under the snout).

https://en.wikipedia.org/wiki/Box_plot

# 자료의 분포: Boxplots

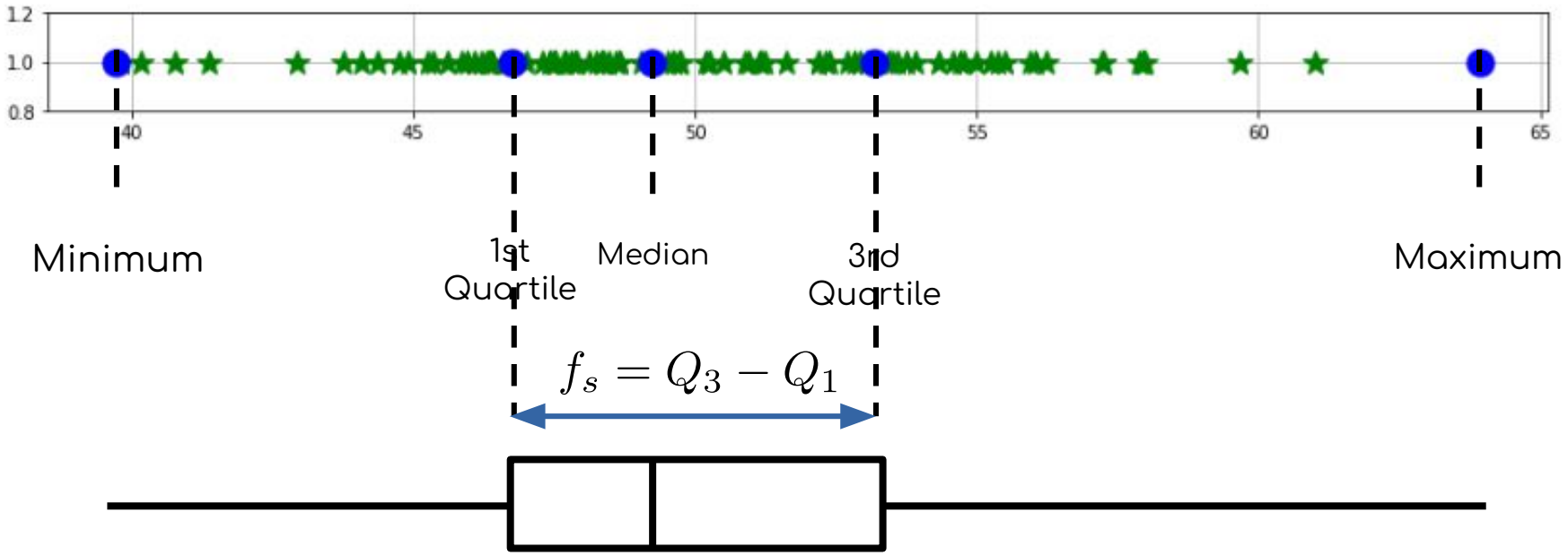


# Boxplots and Outliers




Inter-quartile range = fourth spread = $f_s = Q_3 - Q_1$

# Boxplots and Outliers




## Tukey's fences

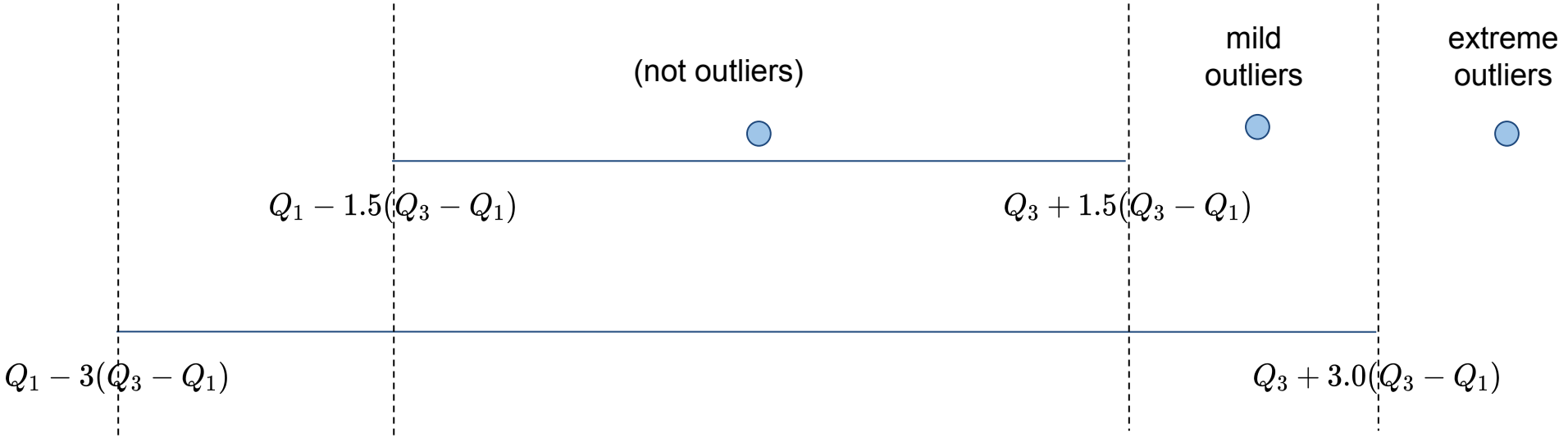
- Outliers are defined if data points are located outside the following range with k=1.5:

$$[Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)] \quad k = 1.5 \text{ for outliers}$$

- In particular, if outliers are located outside the above range of k=3.0, such outliers are called extreme outliers.

https://en.wikipedia.org/wiki/Outlier#Tukey's_fences

# Tukey's Fences : Boxplot and Outliers



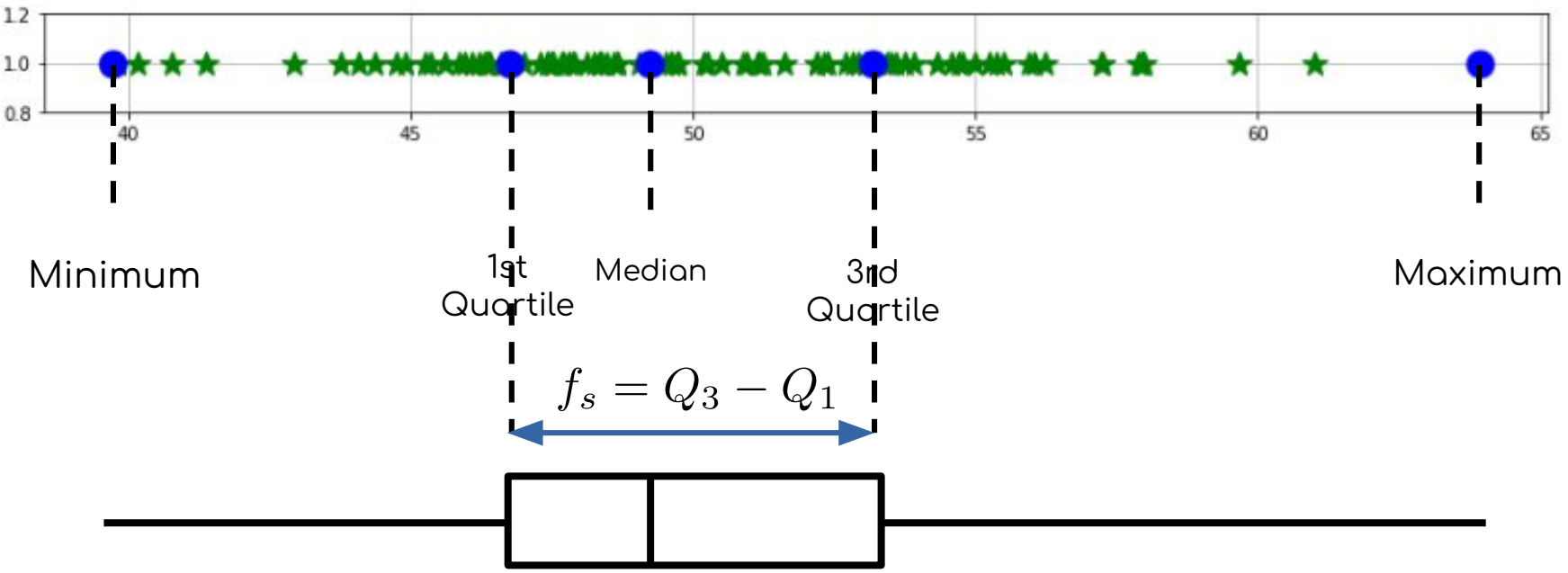

- Outliers are defined if data points are located outside the following range with k=1.5:

$$[Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)] \quad k = 1.5 \text{ for outliers}$$

- In particular, if outliers are located outside the above range of k=3.0, such outliers are called extreme outliers.

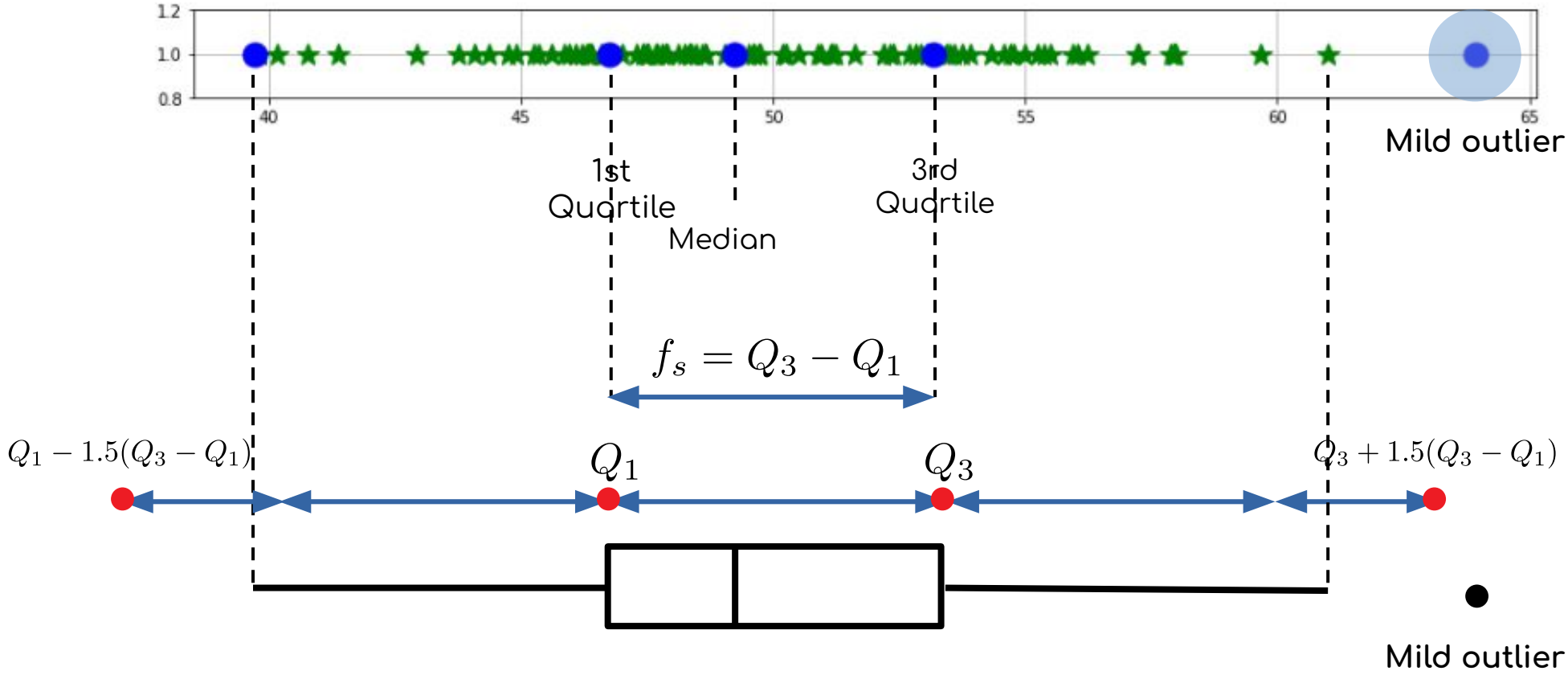
https://en.wikipedia.org/wiki/Outlier#Tukey's_fences

# Boxplots and Outliers



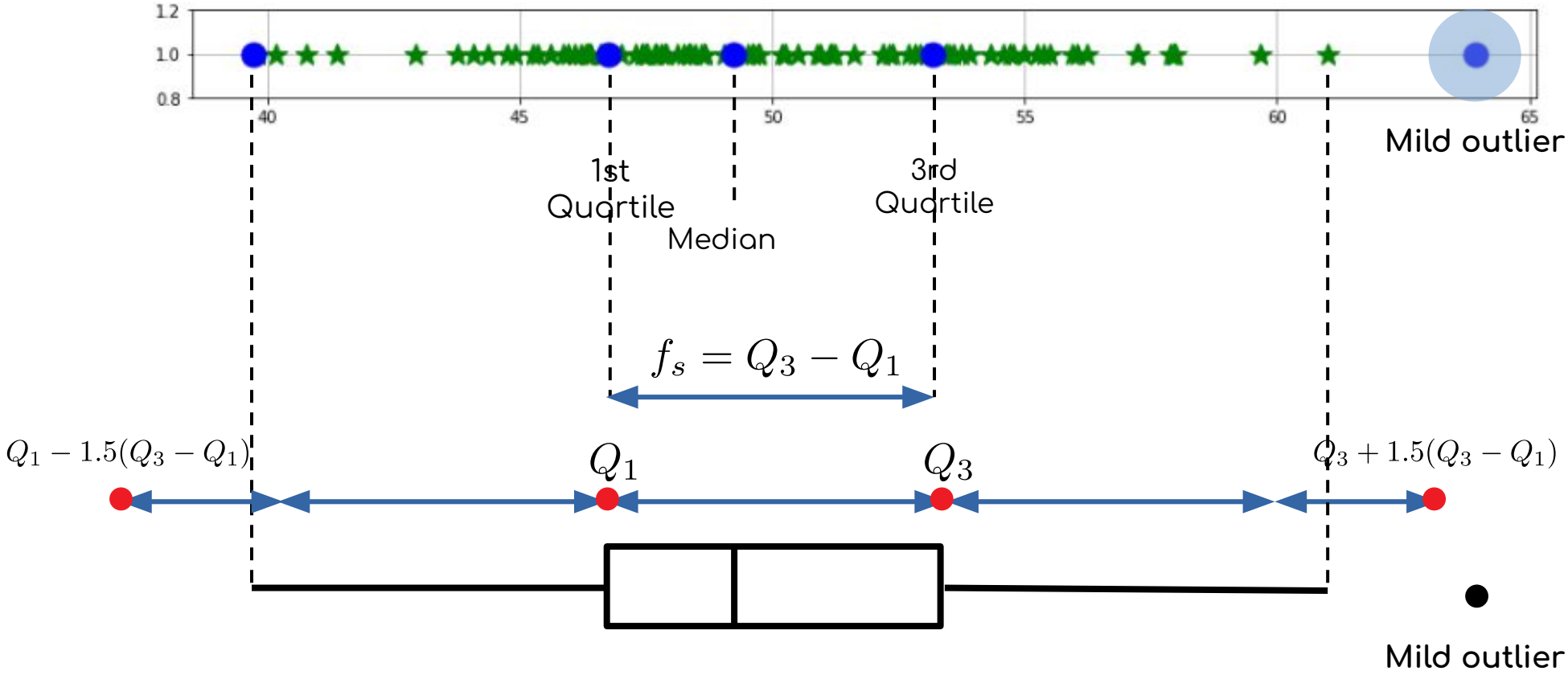

- In order to represent outliers in the boxplot, whiskers of the boxplot can be changed from minimum (maximum) of total data to minimum (maximum) of data in [Q1-1.5(Q3-Q1), Q3+1.5(Q3-Q1)].

# Boxplots and Outliers




- Boxplot with outliers
- Mild outliers (closed circle), Extreme outliers (open circles)

# Boxplots and Outliers




- [Q1-1.5(Q3-Q1), Q3+1.5(Q3-Q1)] : Not outliers
- [Q1-3.0(Q3-Q1), Q1-1.5(Q3-Q1)] or [Q1+1.5(Q3-Q1), Q1+3.0(Q3-Q1)]: Mild outliers
- outside [Q1-3.0(Q3-Q1), Q3+3.0(Q3-Q1)] : Extreme outliers

# Boxplots and Outliers

Mild outlier Extreme outlier

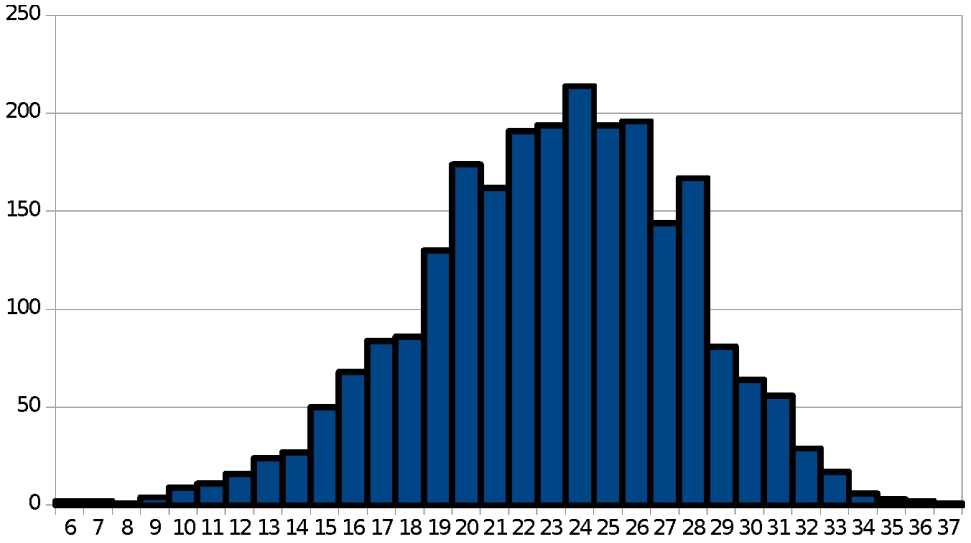
- [Q1-1.5(Q3-Q1), Q3+1.5(Q3-Q1)] : Not outliers
- [Q1-3.0(Q3-Q1), Q1-1.5(Q3-Q1)] or [Q1+1.5(Q3-Q1), Q1+3.0(Q3-Q1)]: Mild outliers
- outside [Q1-3.0(Q3-Q1), Q3+3.0(Q3-Q1)] : Extreme outliers

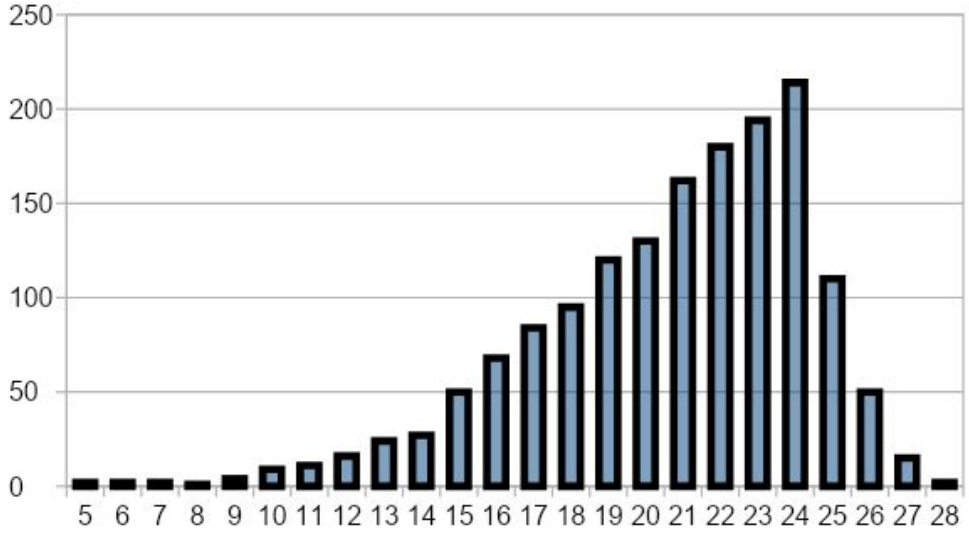
# 기술통계 : Descriptive Statistics

- Central Tendency : mode, median, mean
- Measure of Dispersion : range, inter-quartile range,  variance, standard deviation, boxplot

Anything else to describe data properties quantitatively?

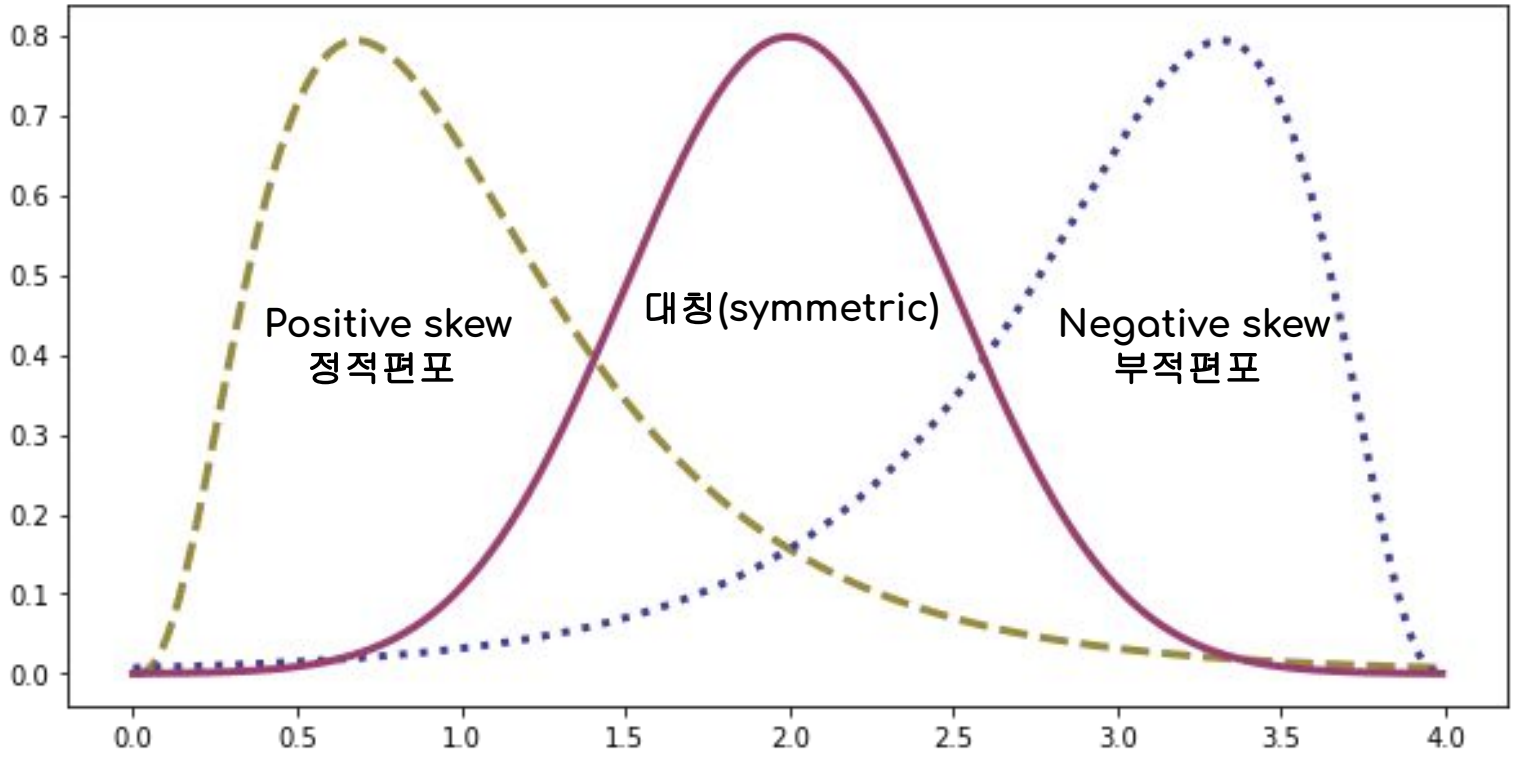
# Data Distributions




Difference between two distributions?

# 왜도(Skewness): Symmetry of Distributions



# 왜도(Skewness): Symmetry of Distributions

- Population skewness (모집단 왜도)

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^3}{\left[\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2\right]^{3/2}}$$

- Sample skewness (표본집단 왜도)

$$b_1 = \frac{m_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{3/2}}$$

$$G_1 = \frac{n^2}{(n-1)(n-2)}\frac{m_3}{s^3} = \frac{\sqrt{n(n-1)}}{n-2}\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{3/2}}$$

https://en.wikipedia.org/wiki/Skewness

why are there several definitions of sample skewness? Depending on whether they give good estimations of parameters

Joanes, D. N.; Gill, C. A. (1998). "Comparing measures of sample skewness and kurtosis". Journal of the Royal Statistical Society, Series D. 47 (1): 183–189

# 왜도(Skewness): Symmetry of Distributions

https://en.wikipedia.org/wiki/Skewness

- Population skewness (모집단 왜도)

$$\gamma_1 = \frac{\mu_3}{\sigma^3} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^3}{\left[\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2\right]^{3/2}}$$

- Sample skewness (표본집단 왜도)

$$b_1 = \frac{m_3}{s^3} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left[\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{3/2}}$$

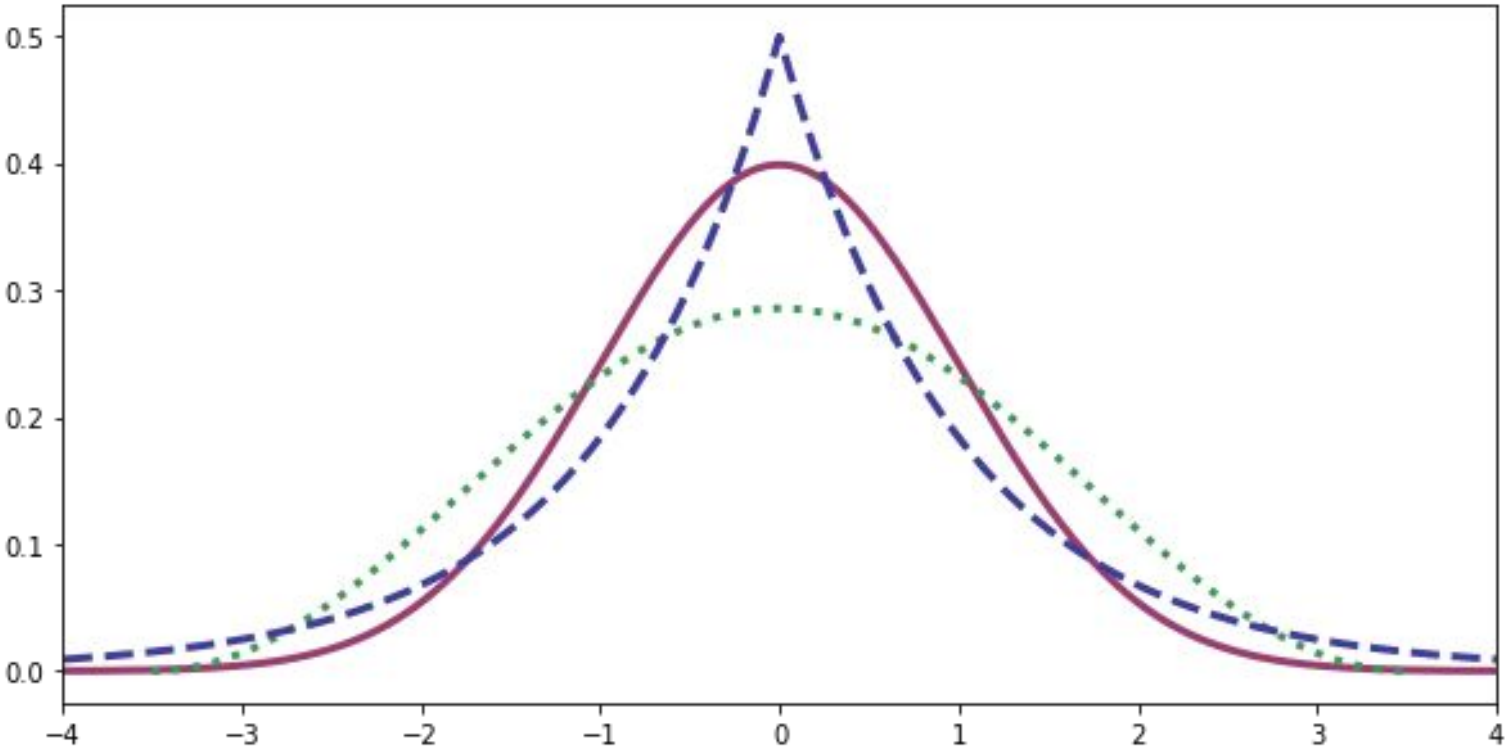
$$G_1 = \frac{n^2}{(n-1)(n-2)}\frac{m_3}{s^3} = \frac{\sqrt{n(n-1)}}{n-2}\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]^{3/2}}$$

$\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^k$ : kth central moment

- 1st moment = mean
- 2nd moment = variance
- 3rd moment = numerator of skewness

Joanes, D. N.; Gill, C. A. (1998). "Comparing measures of sample skewness and kurtosis". Journal of the Royal Statistical Society, Series D. 47 (1): 183-189

# 첨도(Kurtosis): Sharpness of Distributions ~ the 4th moment of your data




- Kurtosis measures sharpness of data distribution, or how much data is concentrated on the center.
- Precisely, kurtosis is the statistical quantity to describe that data distribution is sharper or not than the normal distribution.

# 첨도(Kurtosis)의 정의

- **Population kurtosis (모집단 첨도)**

$$\gamma_2 = \frac{\mu_4}{\sigma^4} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4}{\left[\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2\right]^2}$$

- **Sample kurtosis (표본 첨도)**

$$g_2 = \frac{m_4}{m_2^2} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^2}$$

- Kurtosis is proportional to $4^{th}$ central moment (numerator).
- If the distribution is normal, kurtosis is 3.
- If the distribution is less sharp than the normal distribution, kurtosis is smaller than 3.
- If the distribution is sharper than the normal distribution, kurtosis is larger than 3.
- (Some statistics program (python scipy) gives 0 if your data is normal.)