# Biostatistics
# Week I

Ege Ülgen, M.D.

7 October 2021


ACIBADEM
MEHMET ALİ AYDINLAR
ÜNİVERSİTESİ

# Statistics

- A discipline concerned with
  - **Collecting data** for a certain purpose
  - **Analysis** of the collected data
  - **Reaching conclusions** based on the analysis

# Statistics

Collection

Organization

Analysis

Interpretation

Presentation

# Biotatistics

- Does a novel drug affect survival in pancreatic cancer?
- Which mutation is most likely the cause of an inherited disease?
- Can health status of advanced AIDS patients be improved by a novel treatment?

# Descriptive/Inferential Statistics

- Descriptive Statistics
  - Organization of collected data, calculation of mean and dispersion, presentation as tables, graphics, etc.
- Inferential Statistics
  - Building hypothesis concerning the population based on sample findings, hypothesis testing, interpretation.

# Population vs. Sample

- Population
  - All subjects under consideration that have the same properties
  - E.g., everyone living in Istanbul

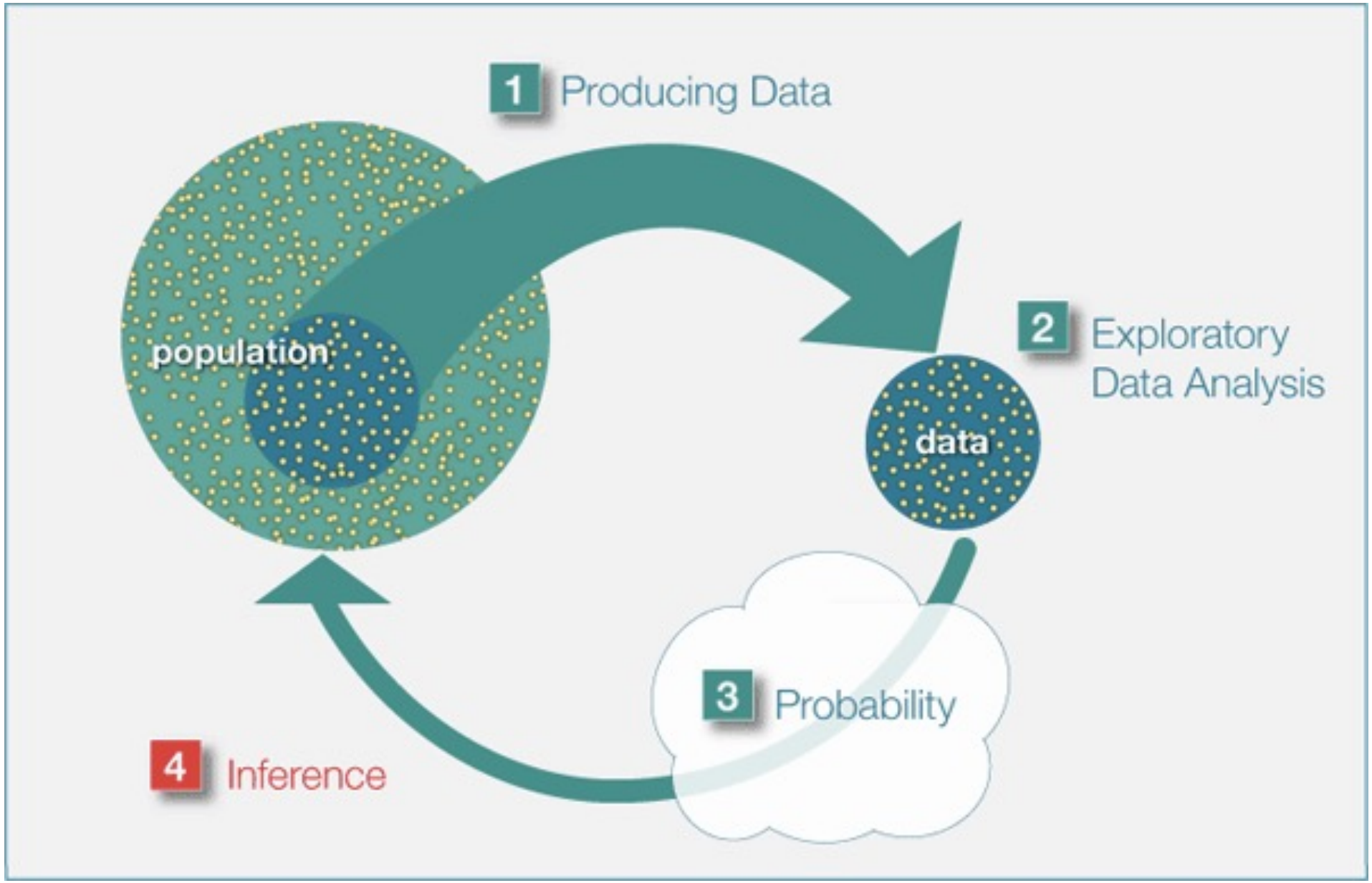
  **N** = 15.52 million (as of 31 Dec 2019)

- Sample
  - A proportion of the population (ideally randomly selected)
  - E.g., **n** = 500, 1000, 5000, …

  (n might be decided based on sample size calculations – Week 11)

# Terminology/Notation

| | Sample Statistic | Population Parameter |
|---|---|---|
| **Size** | $n$ | $N$ |
| **Mean** | $\bar{x} = \frac{\sum x}{n}$ | $\mu = \frac{\sum X}{N}$ |
| **Variance** | $s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$ | $\sigma^2 = \frac{\sum (X-\mu)^2}{N}$ |
| **Standard Deviation** | $s = \sqrt{s^2}$ | $\sigma = \sqrt{\sigma^2}$ |
| **Proportion** | $\hat{p} = \frac{n \text{ of successes}}{n \text{ of trials}}$ | $p = \frac{N \text{ of successes}}{N \text{ of trials}}$ |




*Unit 4a: introduction to statistical inference [Internet]. [cited 2021 Sep 27]. Available from: https://bolt.mph.ufl.edu/6050-6052/unit-4/*

# Data/Variable

- Items of information, often numeric, that are collected through observation
    - Age
    - Gender
    - Ethnicity
    - Systolic blood pressure
    - Treatment type

    …

# Example Study

- Main question
  - Can the health status of advanced AIDS patients be improved by a novel drug treatment?

- Sub-questions
  - Are there differences between treatments in terms of health benefits?
  - Do health benefits differ with respect to gender?
  - Do health benefits differ with respect to age?

# Example Study (cont.)

- Randomized clinical trial
- 1178 patients
  - 289, 288, 293, and 308 patients per treatment arm
- Data collection at baseline (week = 0)
- 5 more follow-ups with 8-week intervals

# Example Study (cont.)

- Variables
  - Identification number
  - Treatment arm
  - Age
  - Gender
  - CD4 cell count at each follow-up
  - Time of follow-up (in weeks since baseline)

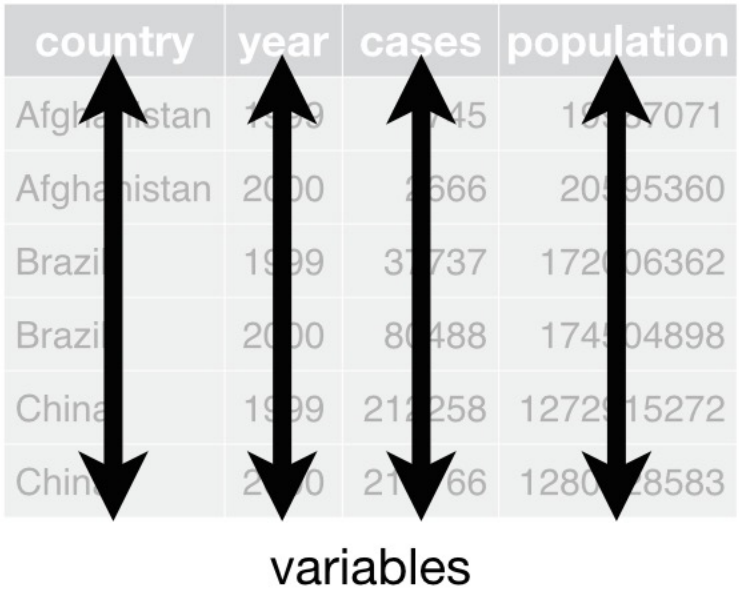
# Example Study (cont.)

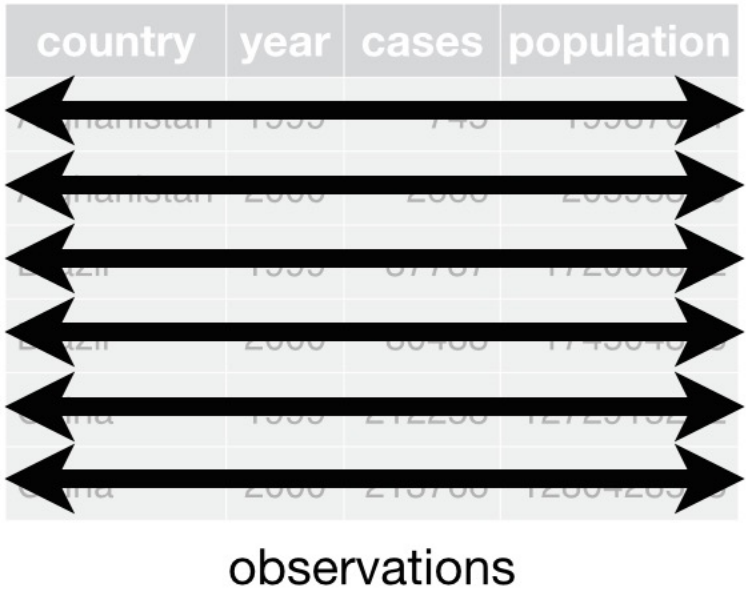
First 5 patients' data, out of 1,178 (only for the first two weeks)

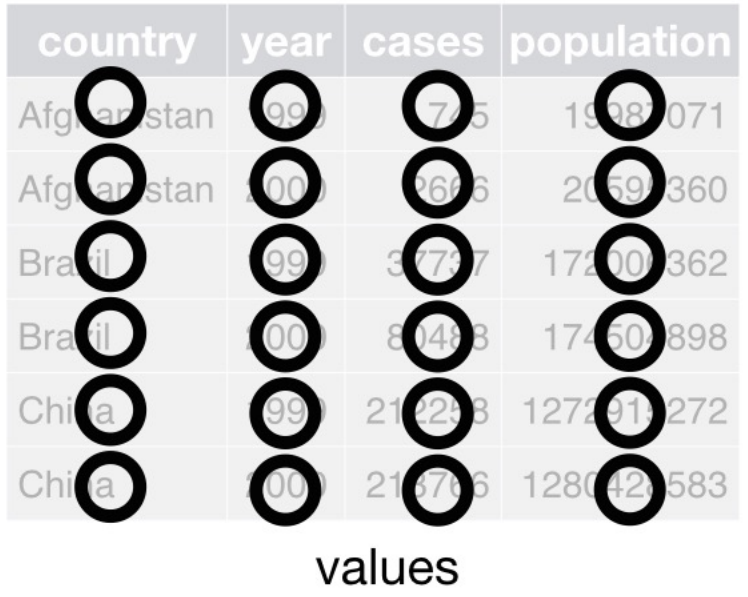
| id | treatment | age | gender | week_1 | cd4_1 | week_2 | cd4_2 |
|---|---|---|---|---|---|---|---|
| 1 | trt2 | 36.43 | male | 0 | 22 | 7.57 | 20 |
| 2 | trt4 | 47.85 | male | 0 | 21 | 8.00 | 48 |
| 4 | trt3 | 36.60 | male | 0 | 61 | 7.14 | 60 |
| 5 | trt1 | 35.95 | male | 0 | 35 | 8.00 | 30 |
| 6 | trt2 | 38.40 | male | 0 | 10 | 7.29 | 10 |

*Applied longitudinal analysis [Internet]. [cited 2021 Sep 27]. Available from: https://content.sph.harvard.edu/fitzmaur/ala2e/*

# «Clean» Data




*12 tidy data | r for data science [Internet]. [cited 2021 Sep 27]. Available from: https://r4ds.had.co.nz/tidy-data.html*

# Variable Types

- **Discrete/Categorical/Qualitative**
  - Measured in a discrete manner

  - **Nominal**: no natural ordering. E.g., eye color, zip-code
  - **Dichotomous/binary**: only takes two values. E.g., dead/alive, female/male
  - **Ordinal**: natural ordering. E.g., agree/neutral/disagree, bad/fair/good
  - **Count**: counted values. E.g., number of tumor occurrences in one month

# Variable Types

- **Continuous/Quantitative**
  - Measured in a continuous manner

  - **Interval:** real number (+/- including 0). E.g., temperature, location
  - **Ratio:** positive values (**0 indicates none**). E.g., height, age, daily calcium consumption (mg).

# Example Study (cont.)

| id | treatment | age | gender | week_1 | cd4_1 | week_2 | cd4_2 |
|---|---|---|---|---|---|---|---|
| 1 | trt2 | 36.43 | male | 0 | 22 | 7.57 | 20 |
| 2 | trt4 | 47.85 | male | 0 | 21 | 8.00 | 48 |
| 4 | trt3 | 36.60 | male | 0 | 61 | 7.14 | 60 |
| 5 | trt1 | 35.95 | male | 0 | 35 | 8.00 | 30 |
| 6 | trt2 | 38.40 | male | 0 | 10 | 7.29 | 10 |

Discrete - nominal

Contin.- ratio

Discrete – nominal /binary

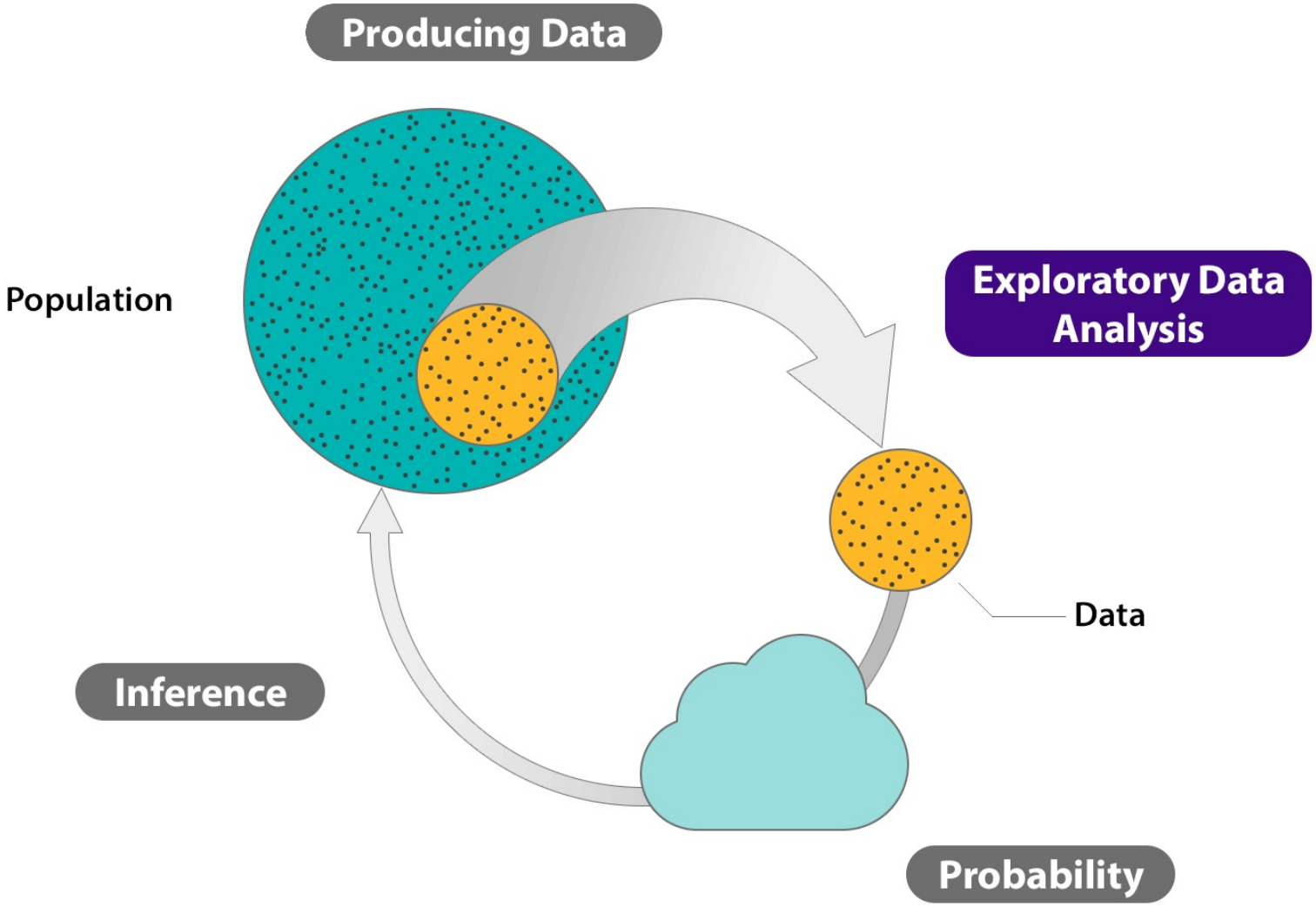
Discrete - count
Contin. - ratio

*Applied longitudinal analysis [Internet]. [cited 2021 Sep 27]. Available from: https://content.sph.harvard.edu/fitzmaur/ala2e/*

# Same variable – different classifications

1. Time after study entry
   0, 1.2, 2.5, 3.1, 4.6, 5.2, 6.6, 7.1, 8 weeks
   Continuous - ratio
2. Time after study entry
   < 4 weeks, ≥ 4 weeks
   Categorical - binary
3. Time after study entry
   < 2 weeks, ≥ 2 and < 4 weeks, ≥ 4 weeks
   Categorical - ordinal
4. Time after study entry
   -4.6, -3.4, -2.1, -1.5, 0, 0.6, 2, 2.5, 3.4 weeks
   Continuous - interval

# The Big Picture




*Unit 1: exploratory data analysis [Internet]. [cited 2021 Sep 27]. Available from: https://bolt.mph.ufl.edu/6050-6052/unit-1/*

# Exploratory Data Analysis (EDA)

- Examining Distributions — exploring data **one variable at a time**.
- Examining Relationships — exploring data **two variables at a time**.

# Frequency Tables – Categorical Variable

- Eye colors of 10 individuals:

  blue, green, brown, blue, brown, blue, blue, green, brown, brown

# Frequency Tables – Categorical Variable

- Eye colors of 10 individuals:

  blue, green, brown, blue, brown, blue, blue, green, brown, brown

| Eye Color | Frequency |
|---|---|
| Blue | 4 |
| Brown | 4 |
| Green | 2 |

# Frequency Tables – Categorical Variable

- Eye colors of 10 individuals:

  blue, green, brown, blue, brown, blue, blue, green, brown, brown

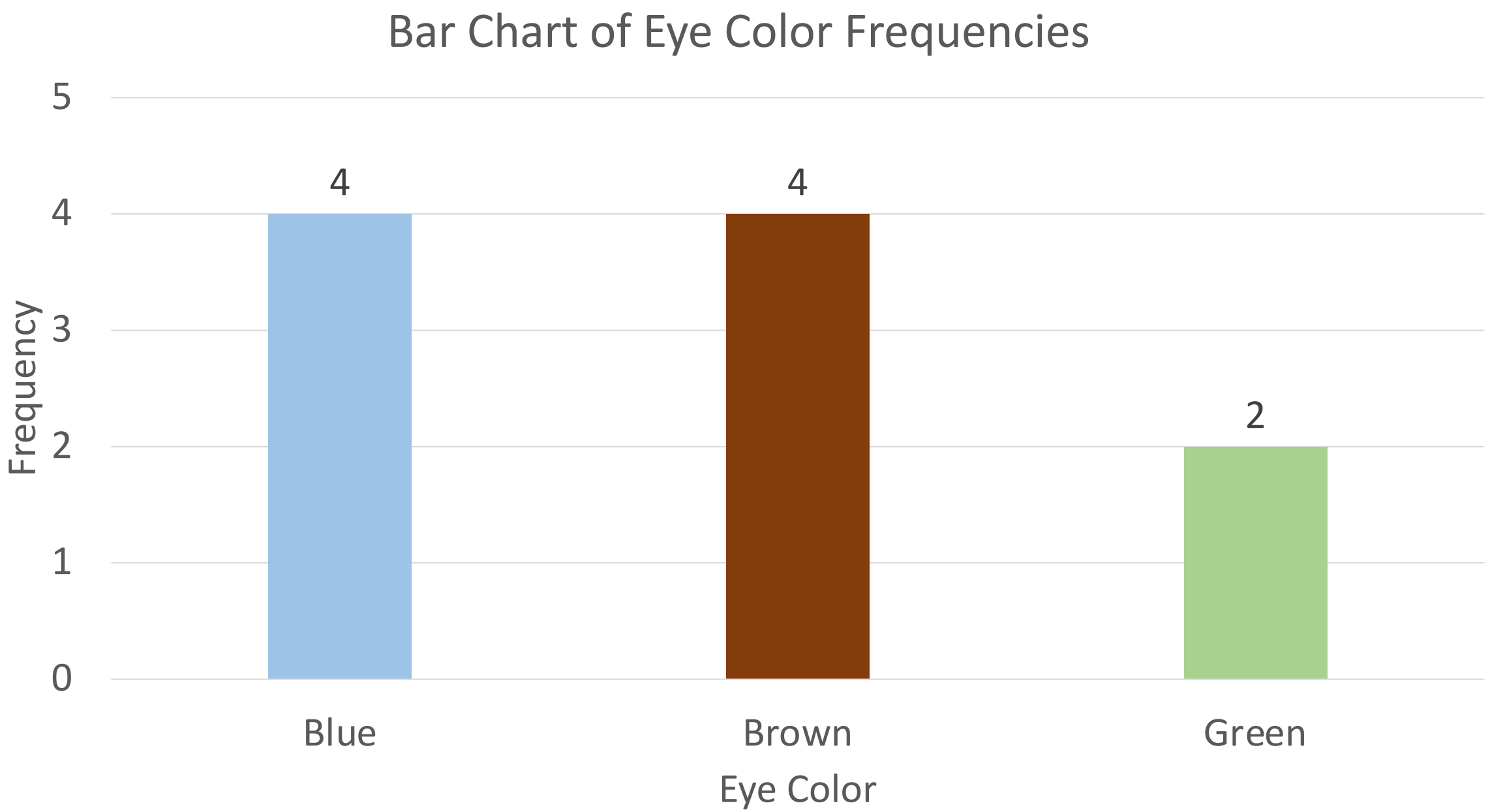
| Eye Color | Frequency | Relative Freq. |
|---|---|---|
| Blue | 4 | 4/10 = 0.4 |
| Brown | 4 | 4/10 = 0.4 |
| Green | 2 | 2/10 = 0.2 |

# Frequency Tables – Categorical Variable

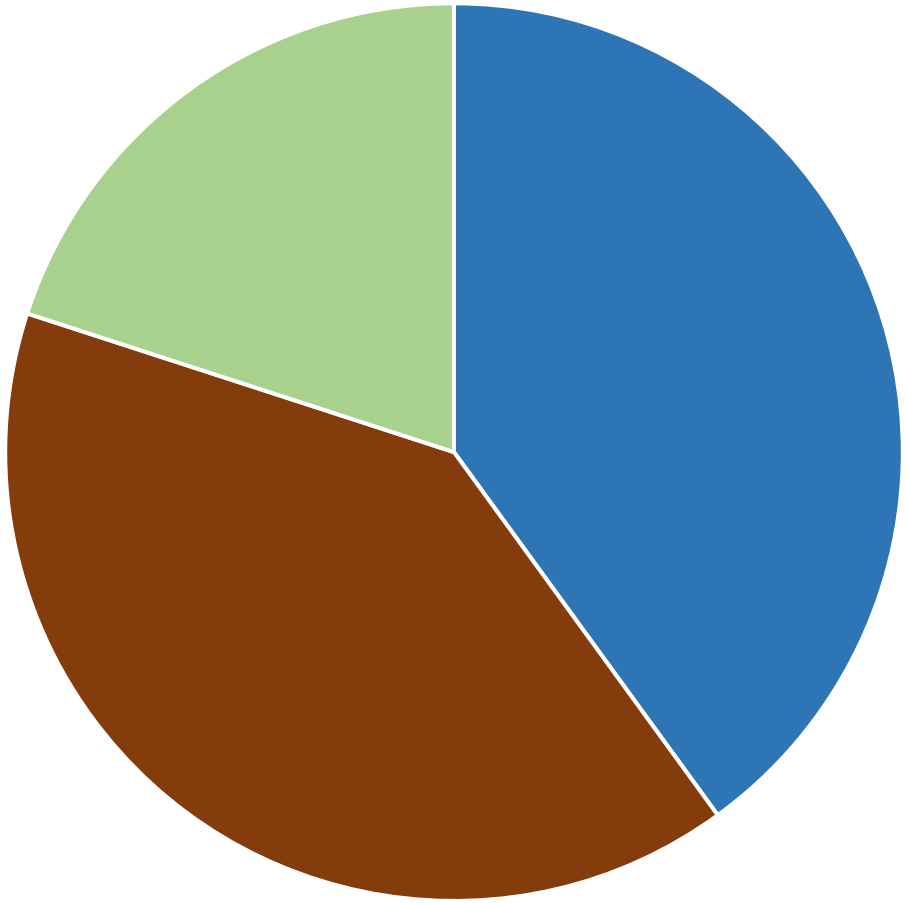
- Eye colors of 10 individuals:

  blue, green, brown, blue, brown, blue, blue, green, brown, brown

| Eye Color | Frequency | Relative Freq. | % |
|---|---|---|---|
| Blue | 4 | 4/10 = 0.4 | 40 |
| Brown | 4 | 4/10 = 0.4 | 40 |
| Green | 2 | 2/10 = 0.2 | 20 |


Bar Chart of Eye Color Frequencies
5
4
3
2
1
0
Frequency
4
4
2
Blue
Brown
Green
Eye Color

# Do not use pie charts!



# Contingency table/Cross tabulation/Crosstab

- Tables in which two categorical variables are investigated together

| | Male | Female |
|---|---|---|
| No education | 4 | 10 |
| Primary school | 3 | 5 |
| High school | 2 | 8 |
| Bachelor’s degree | 7 | 9 |

# Frequency Tables – Continuous Variable

Cholesterol levels of 40 patients:

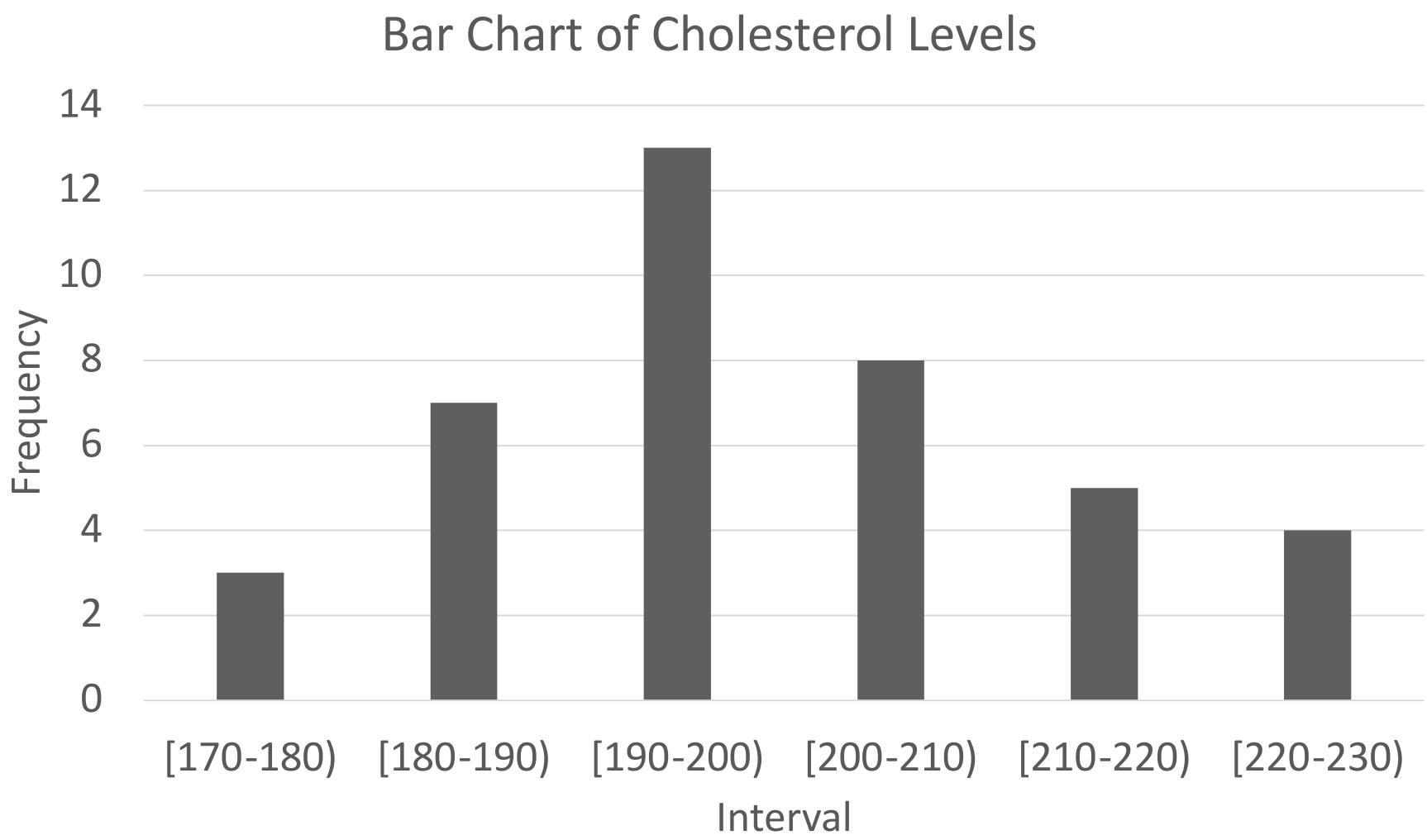
Original data

213, 174, 193, 196, 220, 183, 194, 200, 192, 200, 200, 199, 178, 183, 188, 193, 187, 181, 193, 205, 196, 211, 202, 213, 216, 206, 195, 191, 171, 194, 184, 191, 221, 212, 221, 204, 204, 191, 183, 227

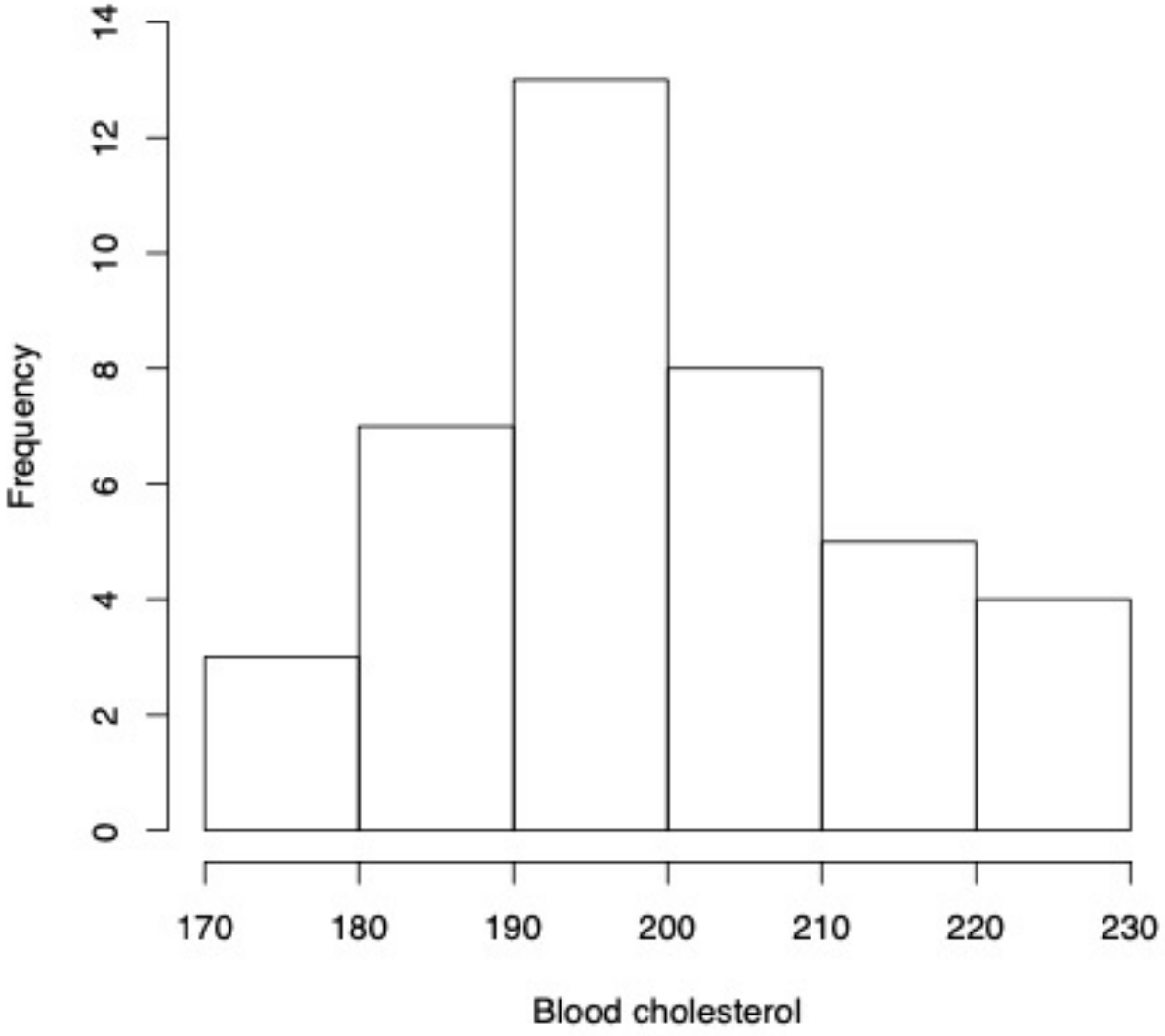
Sorted data

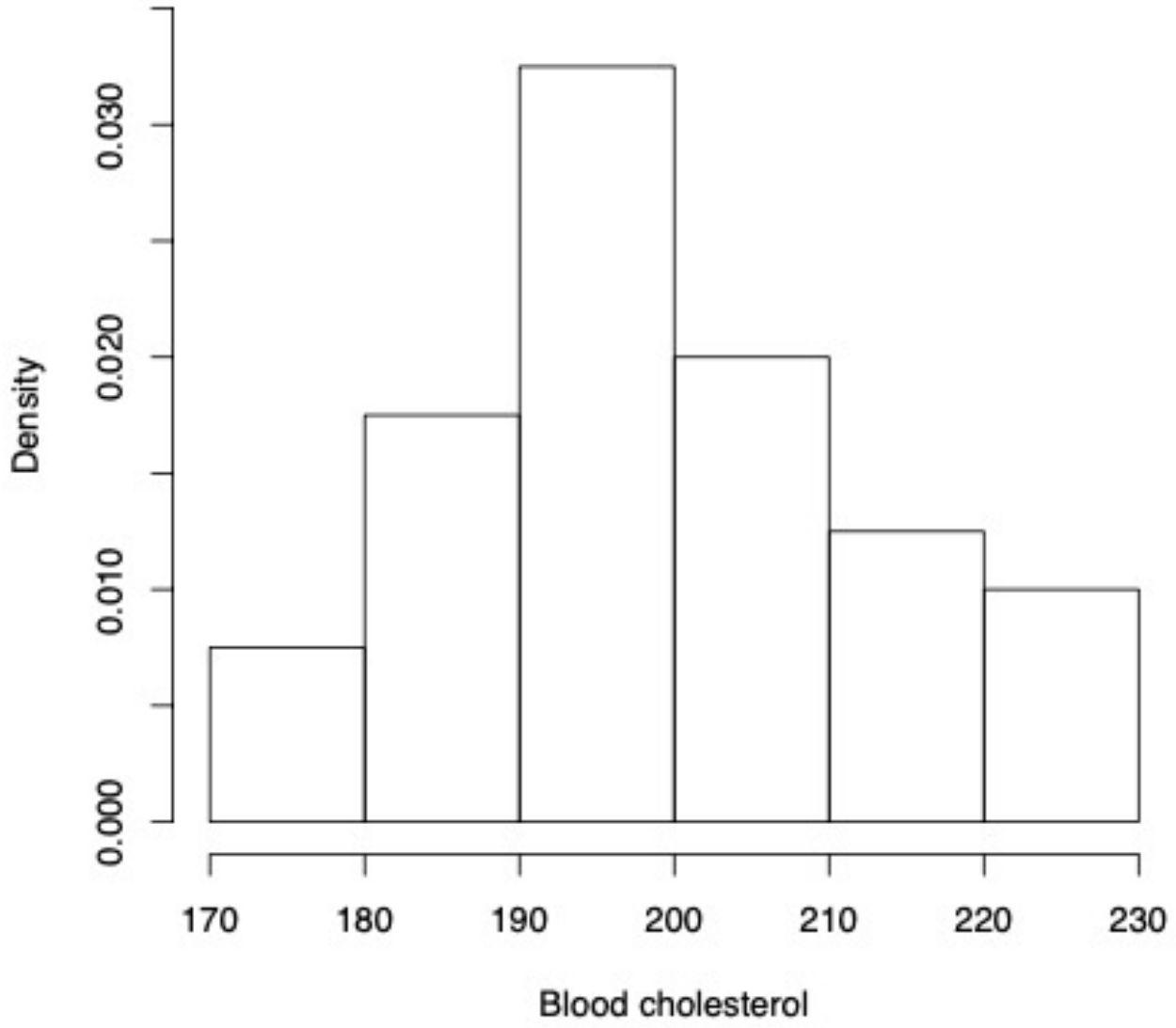
171, 174, 178, 181, 183, 183, 183, 184, 187, 188, 191, 191, 191, 192, 193, 193, 193, 194, 194, 195, 196, 196, 199, 200, 200, 200, 202, 204, 204, 205, 206, 211, 212, 213, 213, 216, 220, 221, 221, 227

| Interval | Frequency | Relative Freq. | % |
|---|---|---|---|
| [170-180) | 3 | 3/40 = 0.075 | 7.5 |
| [180-190) | 7 | 7/40 = 0.175 | 17.5 |
| [190-200) | 13 | 13/40 = 0.325 | 32.5 |
| [200-210) | 8 | 8/40 = 0.200 | 20.0 |
| [210-220) | 5 | 5/40 = 0.125 | 12.5 |
| [220-230) | 4 | 4/40 = 0.100 | 10.0 |

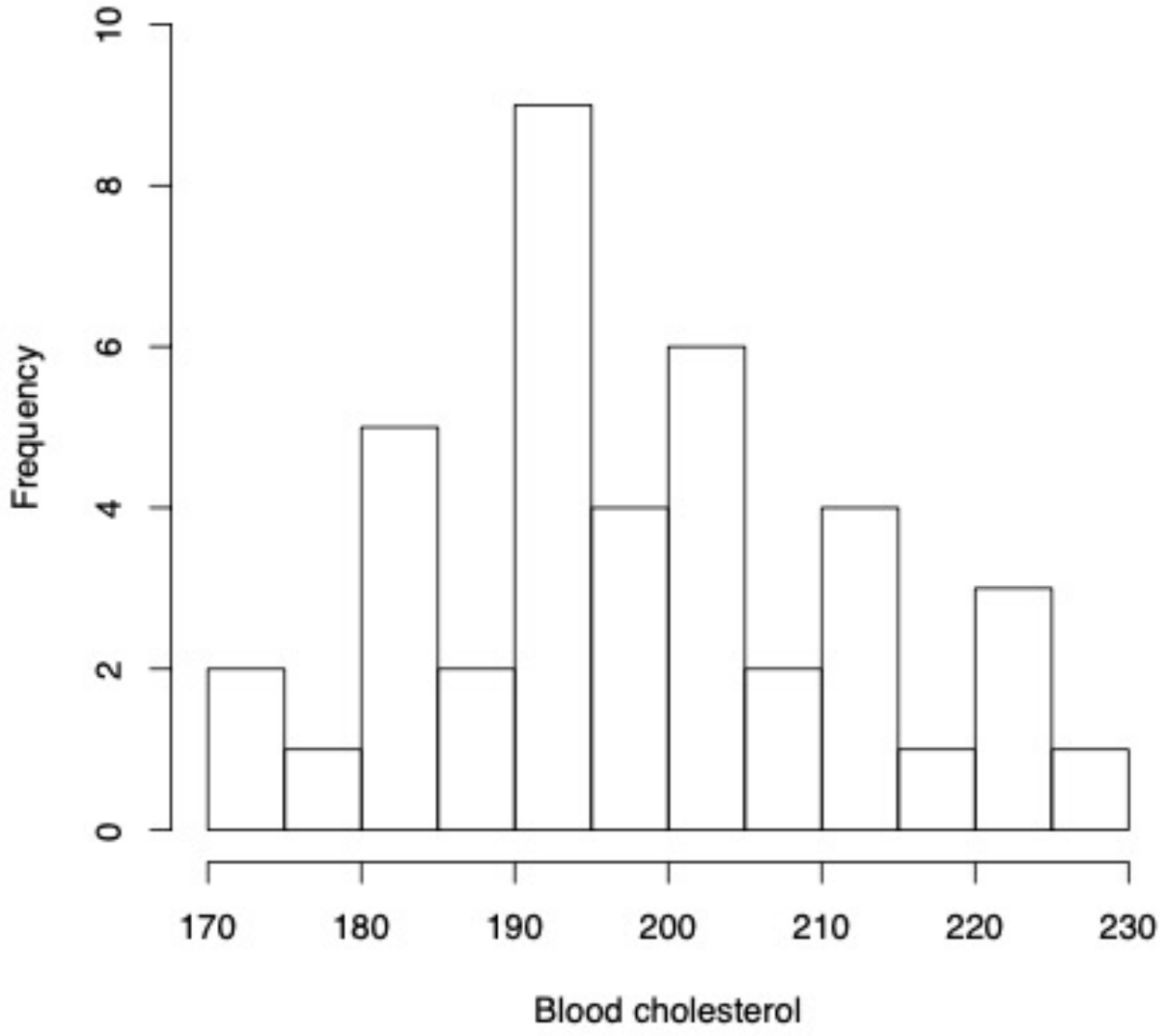

Bar Chart of Cholesterol Levels
14
12
10
8
6
4
2
0
Frequency
[170-180)
[180-190)
[190-200)
[200-210)
[210-220)
[220-230)
Interval

# Histogram


Frequency
0
2
4
6
8
10
12
14
170
180
190
200
210
220
230
Blood cholesterol


Density
0.000
0.010
0.020
0.030
170
180
190
200
210
220
230
Blood cholesterol

# Histogram


Frequency
0
2
4
6
8
10
170
180
190
200
210
220
230
Blood cholesterol

*Ross, S. M. (2017). Introductory statistics. Amsterdam: Elsevier/AP.*

# Histogram

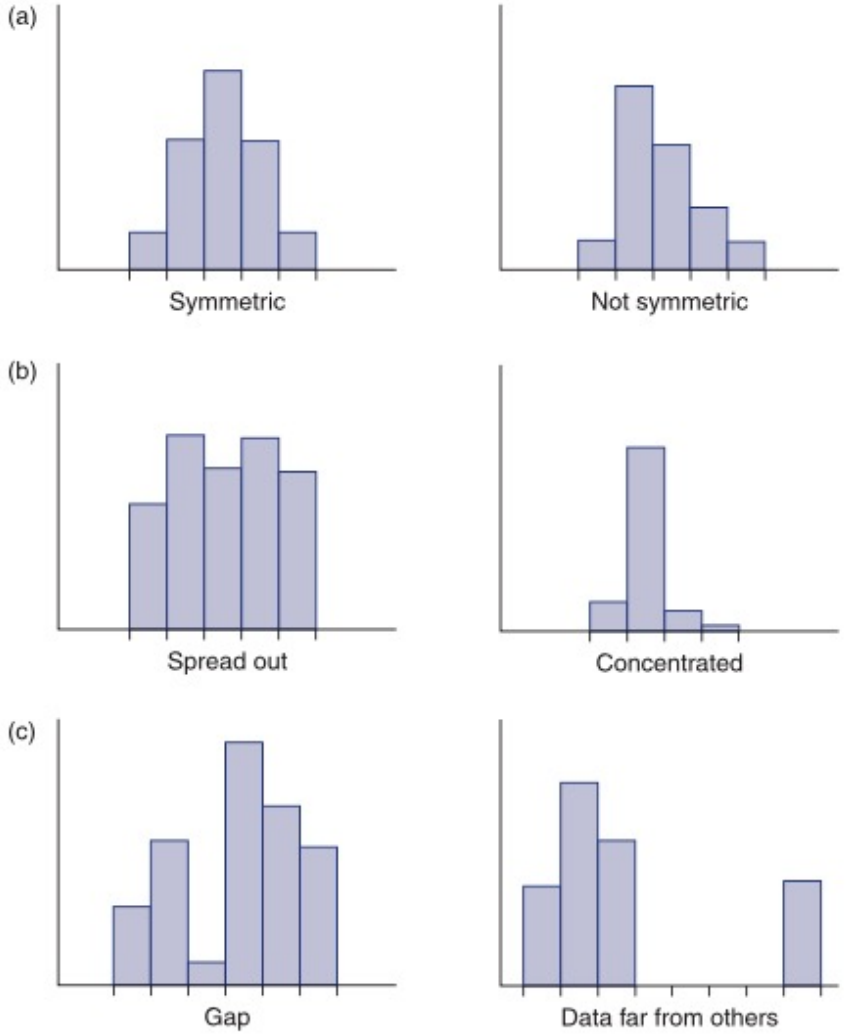



**FIGURE 2.8**

*Characteristics of data detected by histograms. (a) symmetry, (b) degree of spread and where values are concentrated, and (c) gaps in data and data far from others.*

*Ross, S. M. (2017). Introductory statistics. Amsterdam: Elsevier/AP.*

**Table 2.9** Class Frequencies of Systolic Blood Pressure of Two Groups of Male Workers

| Blood pressure | Number of workers | |
|---|---|---|
| | Aged 30–40 | Aged 50–60 |
| Less than 90 | 3 | 1 |
| 90–100 | 17 | 2 |
| 100–110 | 118 | 23 |
| 110–120 | 460 | 57 |
| 120–130 | 768 | 122 |
| 130–140 | 675 | 149 |
| 140–150 | 312 | 167 |
| 150–160 | 120 | 73 |
| 160–170 | 45 | 62 |
| 170–180 | 18 | 35 |
| 180–190 | 3 | 20 |
| 190–200 | 1 | 9 |
| 200–210 | | 3 |
| 210–220 | | 5 |
| 220–230 | | 2 |
| 230–240 | | 1 |
| **Total** | **2540** | **731** |

**Table 2.10** Relative Class Frequencies of Blood Pressures

| Blood pressure | Percentage of workers | |
|---|---|---|
| | Aged 30–40 | Aged 50–60 |
| Less than 90 | 0.12 | 0.14 |
| 90–100 | 0.67 | 0.27 |
| 100–110 | 4.65 | 3.15 |
| 110–120 | 18.11 | 7.80 |
| 120–130 | 30.24 | 16.69 |
| 130–140 | 26.57 | 20.38 |
| 140–150 | 12.28 | 22.84 |
| 150–160 | 4.72 | 9.99 |
| 160–170 | 1.77 | 8.48 |
| 170–180 | 0.71 | 4.79 |
| 180–190 | 0.12 | 2.74 |
| 190–200 | 0.04 | 1.23 |
| 200–210 | | 0.41 |
| 210–220 | | 0.68 |
| 220–230 | | 0.27 |
| 230–240 | | 0.14 |
| **Total** | **100.00** | **100.00** |

Ross, S. M. (2017). Introductory statistics. Amsterdam: Elsevier/AP.

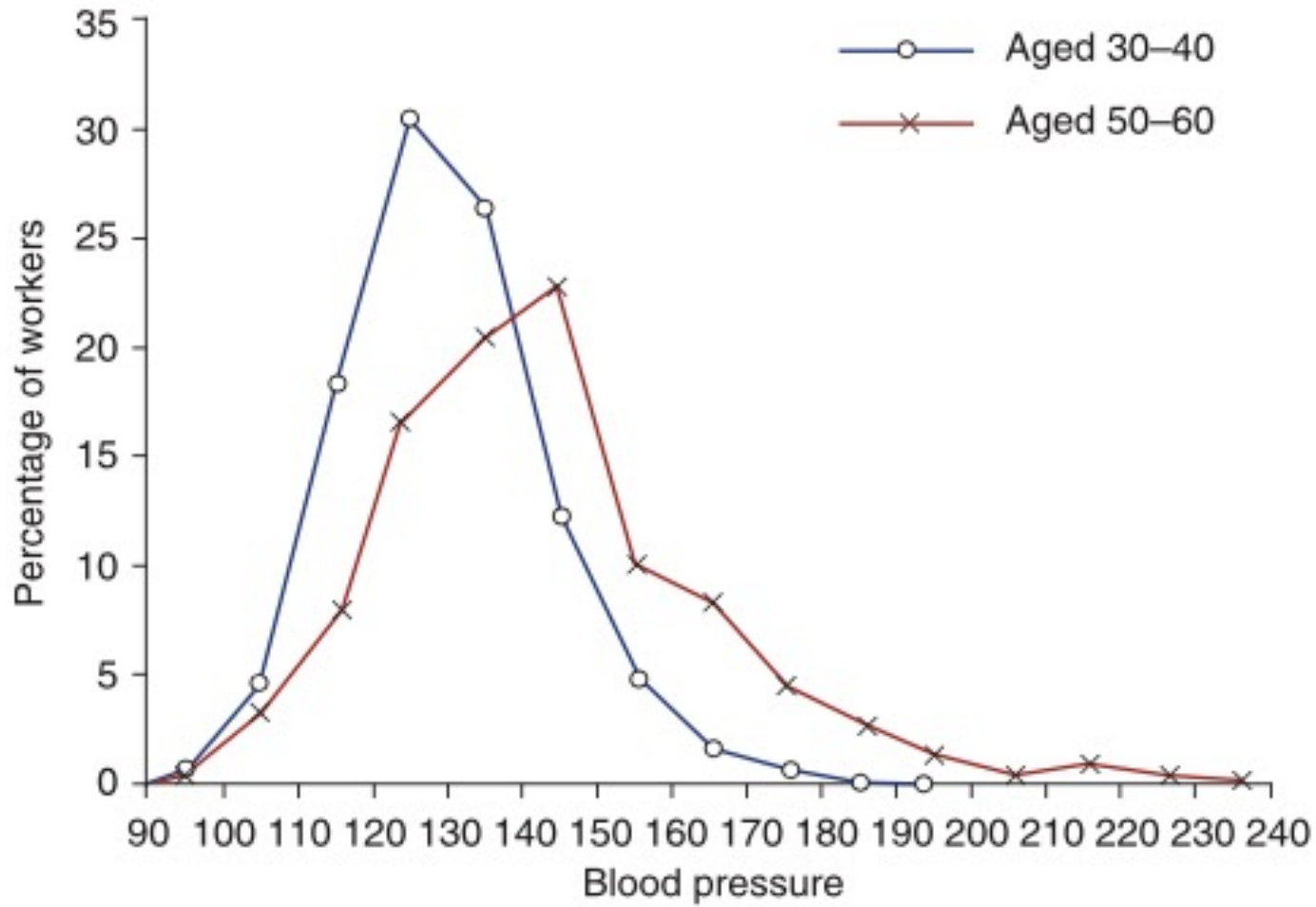



**FIGURE 2.10**

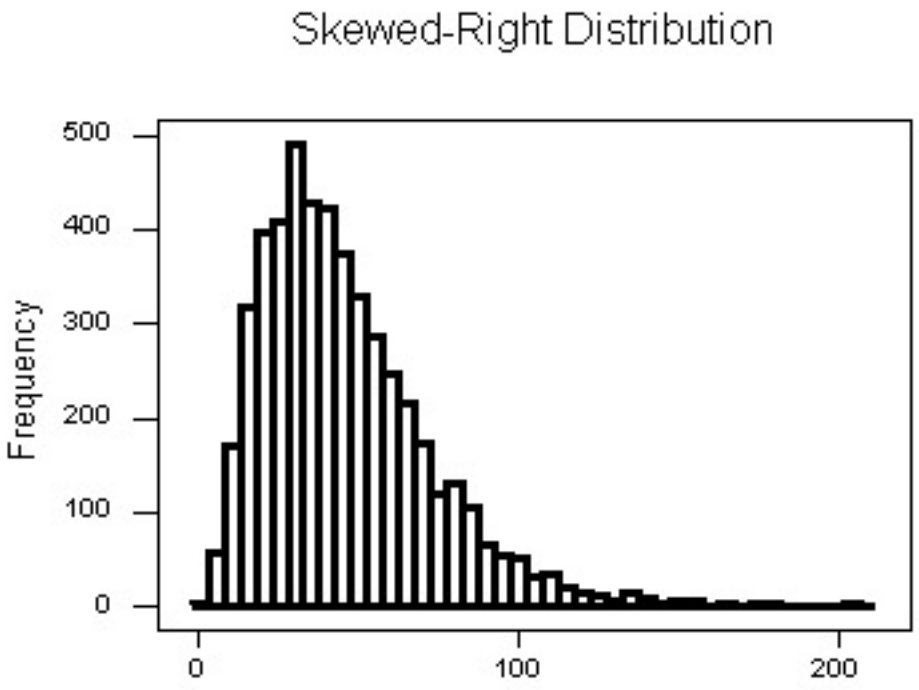
*Relative frequency polygons for the data of Table 2.10.*

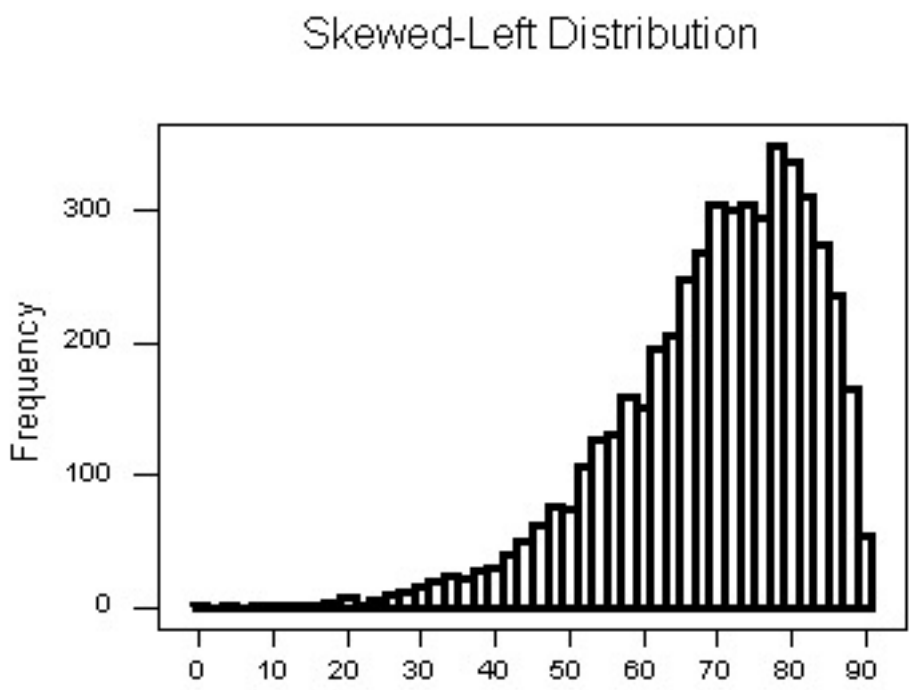
# Describing Distributions

- **Shape**
- **Center**
- Spread
- Outliers

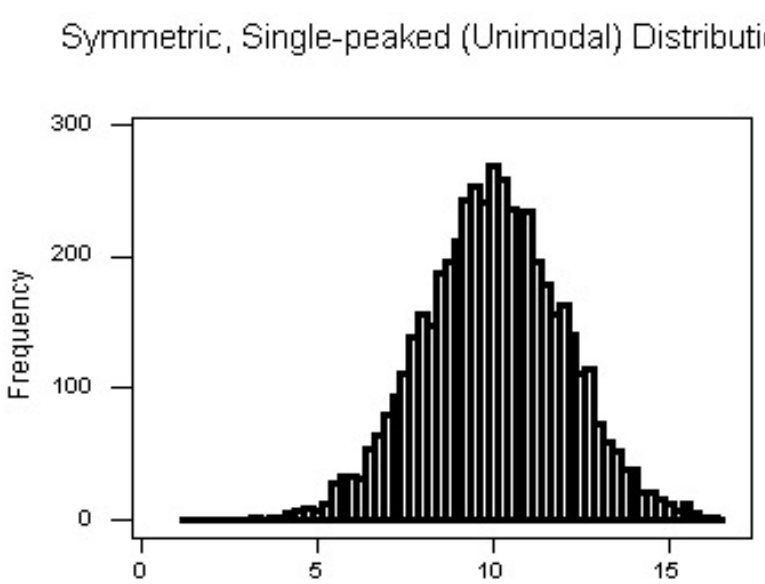
# Shape

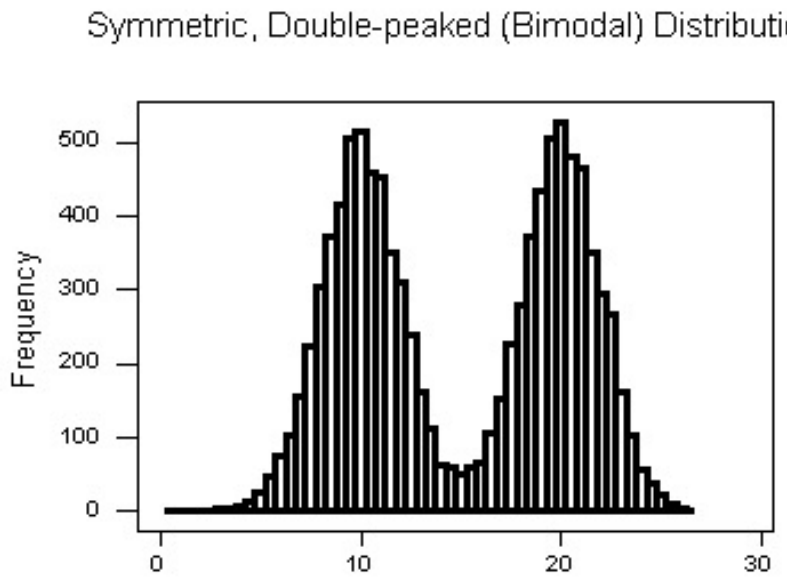
- **Symmetry/Skewness** of the distribution
- **Peakedness (modality)**
  - The number of peaks (modes) the distribution has

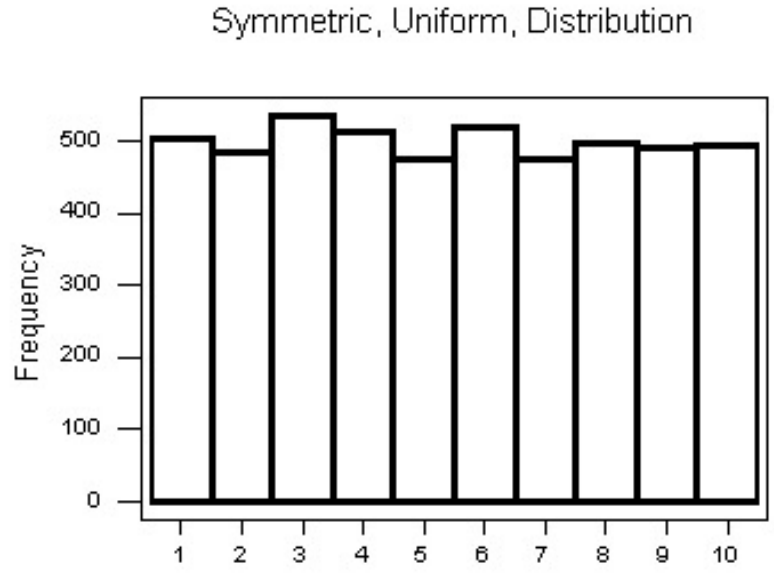



*Describing distributions [Internet]. [cited 2021 Oct 1]. Available from: https://bolt.mph.ufl.edu/6050-6052/unit-1/one-quantitative-variable-introduction/describing-distributions/*




*Describing distributions [Internet]. [cited 2021 Oct 1]. Available from: https://bolt.mph.ufl.edu/6050-6052/unit-1/one-quantitative-variable-introduction/describing-distributions/*

# Center

- Mean
- Median
- Mode

# Center - Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Cholesterol levels of 40 patients:

213, 174, 193, 196, 220, 183, 194, 200, 192, 200, 200, 199, 178, 183, 188, 193, 187, 181, 193, 205, 196, 211, 202, 213, 216, 206, 195, 191, 171, 194, 184, 191, 221, 212, 221, 204, 204, 191, 183, 227

$$\bar{x} = \frac{213+174+\ldots+227}{40} = 197.625$$

# Mean

If $y_i = x_i + c$ ($c$ is a constant) $\bar{y} = \bar{x} + c$

$$\bar{x} = \frac{213+174+\ldots+227}{40} = 197.625$$

$$\bar{y} = \frac{(213+5)+(174+5)+\ldots+(227+5)}{40} = 202.625$$

# Mean

If $y_i = x_i \; x \; c$ ($c$ is a constant) $\quad \bar{y} = \bar{x} \; x \; c$

x: 1, 2, 3, 4, 5

y: 3 (1 * 3), 6 (2 * 3), 9 (3 * 3), 12 (4 * 3), 15 (5 * 3)
$\Rightarrow c = 3$

$\bar{x} = 3, \; \bar{y} = 9 \Rightarrow \bar{y} = 3 * \bar{x}$

# Mean

- Even a small change in a single value affects the mean

213, 174, 193, 196, 220, 183, 194, 200, 192, 200, 200, 199, 178, 183, 188, 193, 187, 181, 193, 205, 196, 211, 202, 213, 216, 206, 195, 191, 171, 194, 184, 191, 221, 212, 221, 204, 204, 191, 183, 227

- If the maximal value was 700 (instead of 227), the mean would be 209.45 (instead of 197.625)

# Median

- It is calculated as the:
  - middle value of the sorted values (if n is odd)
  - average of two middle values of the sorted values (if n is even)

2, 5, 3, 10, 4

2, 3, 4, 5, 10 => median = 4

5, 3, 10, 4

3, 4, 5, 10 => median = 4.5

# Median

Cholesterol levels of 40 patients:

Original data

213, 174, 193, 196, 220, 183, 194, 200, 192, 200, 200, 199, 178, 183, 188, 193, 187, 181, 193, 205, 196, 211, 202, 213, 216, 206, 195, 191, 171, 194, 184, 191, 221, 212, 221, 204, 204, 191, 183, 227

Sorted dataa

171, 174, 178, 181, 183, 183, 183, 184, 187, 188, 191, 191, 191, 192, 193, 193, 193, 194, 194, 195, 196, 196, 199, 200, 200, 200, 202, 204, 204, 205, 206, 211, 212, 213, 213, 216, 220, 221, 221, 227

Mean = 197.625

Median = 195.5

# Median

171, 174, 178, 181, 183, 183, 183, 184, 187, 188, 191, 191, 191, 192, 193, 193, 193, 194, 194, 195, 196, 196, 199, 200, 200, 200, 202, 204, 204, 205, 206, 211, 212, 213, 213, 216, 220, 221, 221, **227**

Mean = 197.625

Median = 195.5

171, 174, 178, 181, 183, 183, 183, 184, 187, 188, 191, 191, 191, 192, 193, 193, 193, 194, 194, 195, 196, 196, 199, 200, 200, 200, 202, 204, 204, 205, 206, 211, 212, 213, 213, 216, 220, 221, 221, **700**

Mean = 209.45

Median = 195.5

# Mode

- The mode is the value that appears most often in a set of data values

- Systolic blood pressures of 12 patients:

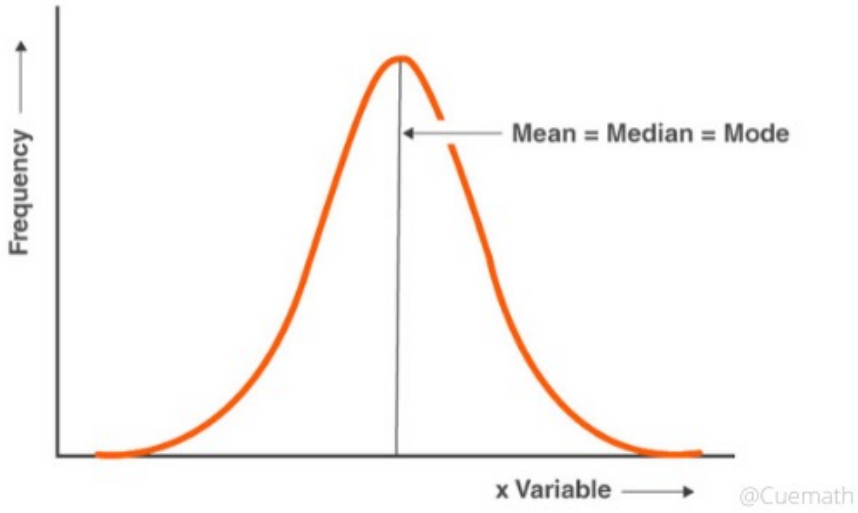
90, 80, **100**, 110, **100**, 120, **100**, 90, **100**, 110, 120, 110

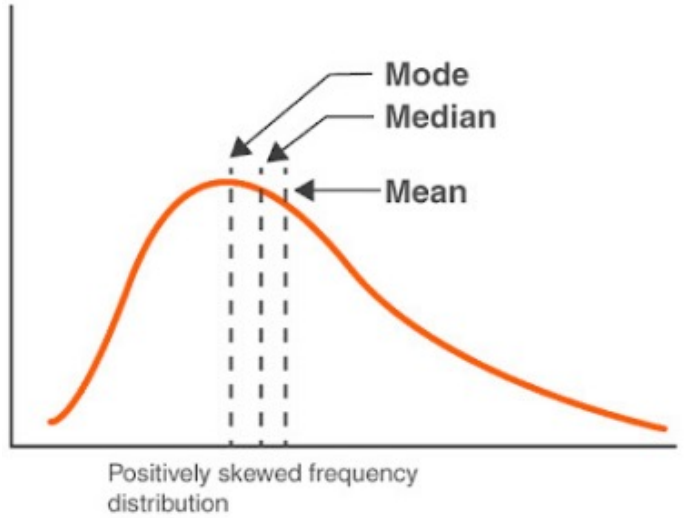
Mode = 100

Mean = 102.5

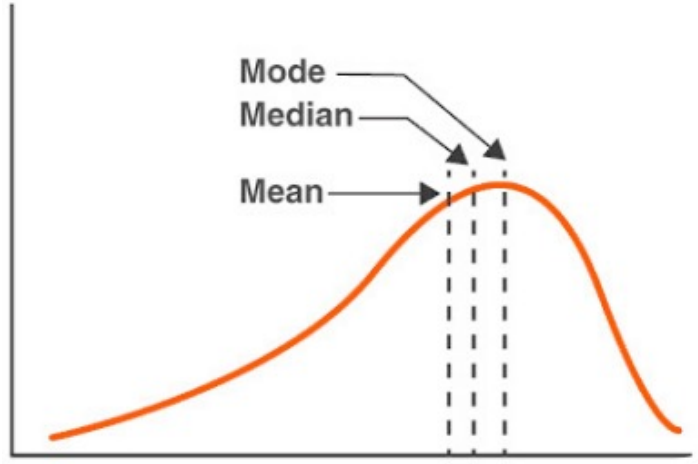
Median = 100

# Mean – Median – Mode Relationship




Positively skewed frequency distribution




Negatively skewed frequency distribution

# Brief Summary

- Statistics is a discipline concerned with collection, organization, analysis and interpretation of data
- The aim is to infer information regarding the population using sample data
- There are two kinds of variables:
  - Categorical – nominal, binary, ordinal, count
  - Continuous – interval, ratio
- We may summarize a categorical variable using frequency, relative frequency and/or percentage tables
- We may visually display a categorical variable using bar charts, etc.
- We may visually inspect the distribution of a continuous variable using histograms
- To determine the center of a continuous variable, one can use mean, median, mode
- The mean is very sensitive to outliers, while the median is robust to outliers