Biostatistics
Week II

Ege Ülgen, M.D.

14 October 2021

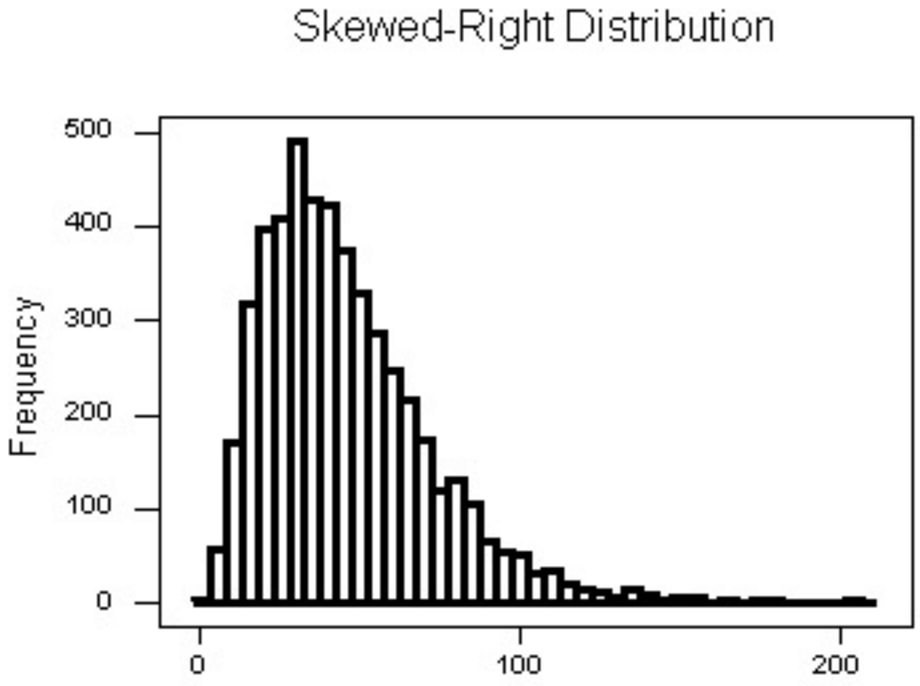
ACIBADEM
MEHMET ALİ AYDINLAR
ÜNİVERSİTESİ

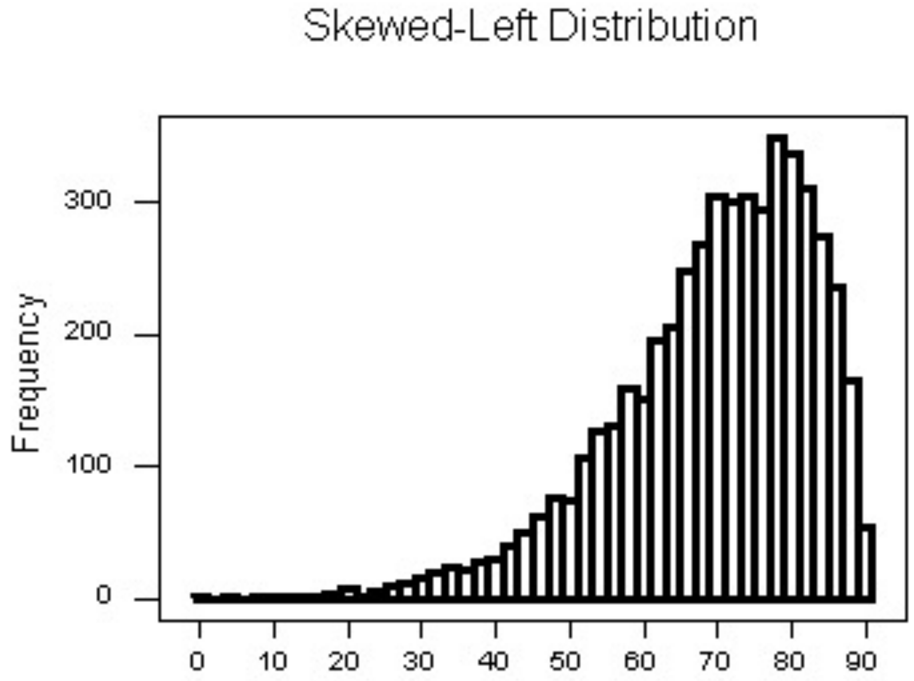
Describing Distributions

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

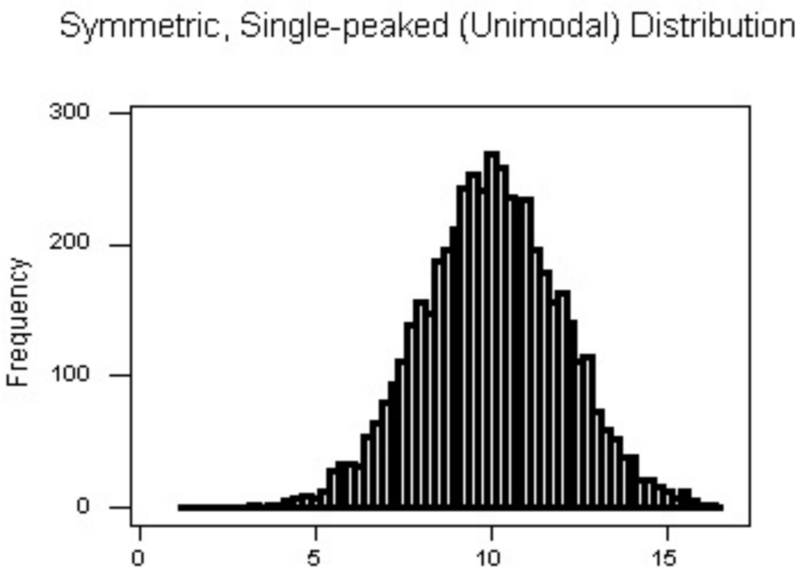
Shape

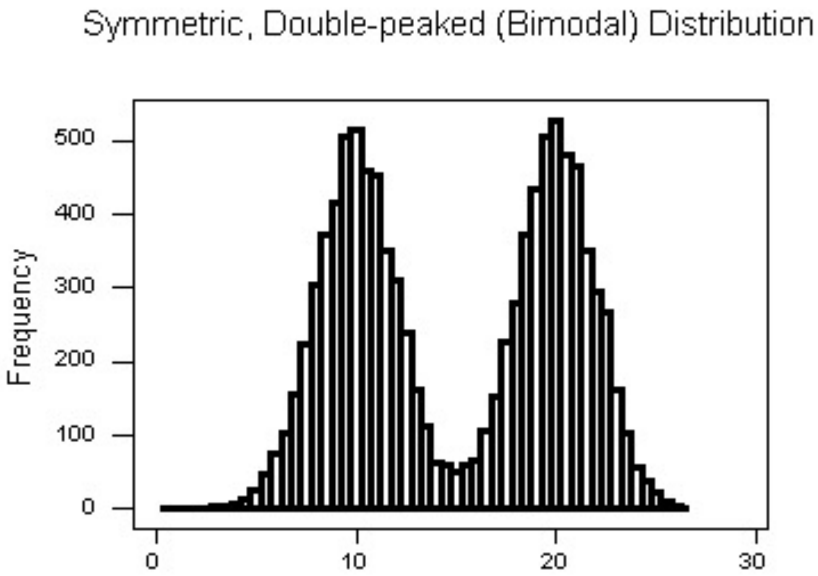
- **Symmetry/Skewness** of the distribution
- **Peakedness (modality)**
  - The number of peaks (modes) the distribution has

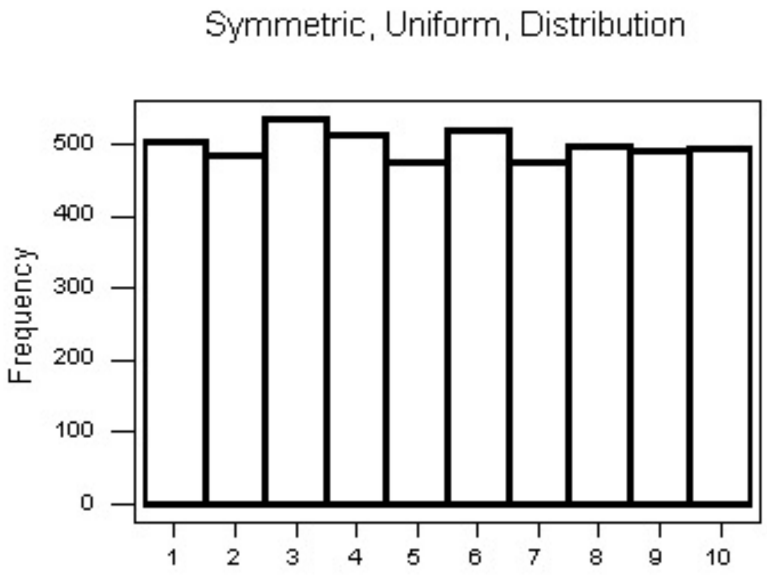




Describing distributions [Internet]. [cited 2021 Oct 1]. Available from: https://bolt.mph.ufl.edu/6050-6052/unit-1/one-quantitative-variable-introduction/describing-distributions/







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

Describing Distributions

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

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) $\bar{y} = \bar{x} + c$

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

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

Mean

If $y_i = x_i$ x c (c is a constant) $\bar{y} = \bar{x}$ 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, <u>4</u>, 5, 10 => median = 4

5, 3, 10, 4

    3, <u>4</u>, <u>5</u>, 10 => median = 4.5

Median

Cholesterol levels of 40 patients:

<u>Original data</u>
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

<u>Sorted dataa</u>
171, 174, 178, 181, 183, 183, 183, 184, 187, 188, 191, 191, 191, 192, 193, 193, 193, 194, 194, <u>195</u>, <u>196</u>, 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:

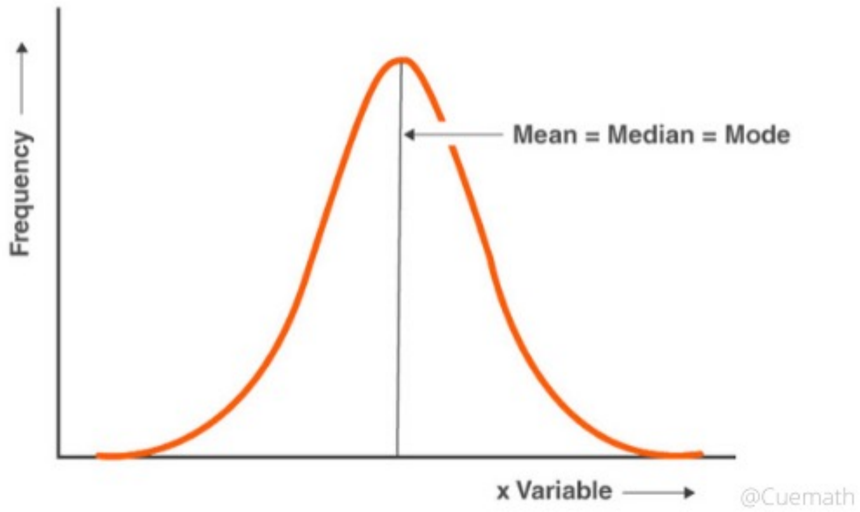
90, 80, **100**, 110, **100**, 120, **100**, 90, **100**, 110, 120, 110

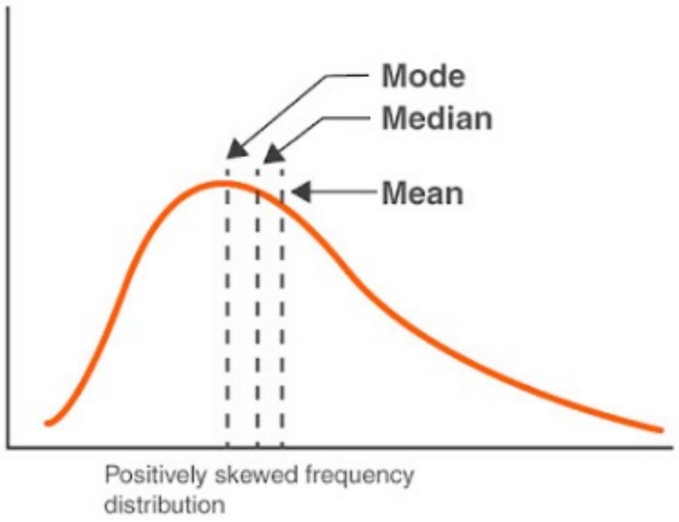
Mode = 100

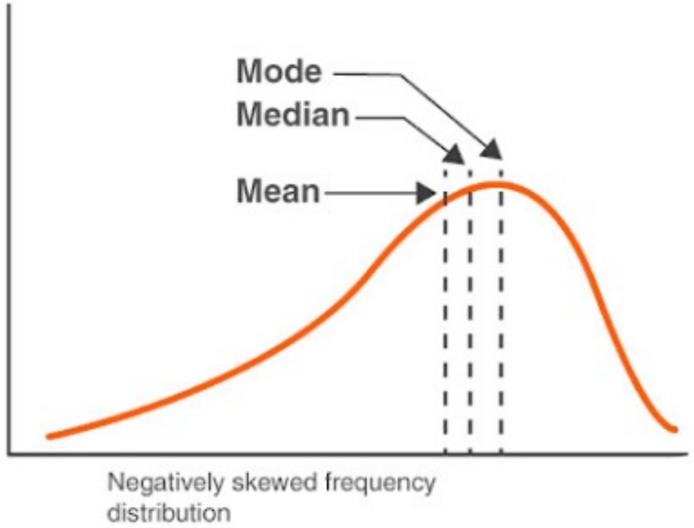
Mean = 102.5

Median = 100

Mean – Median – Mode Relationship



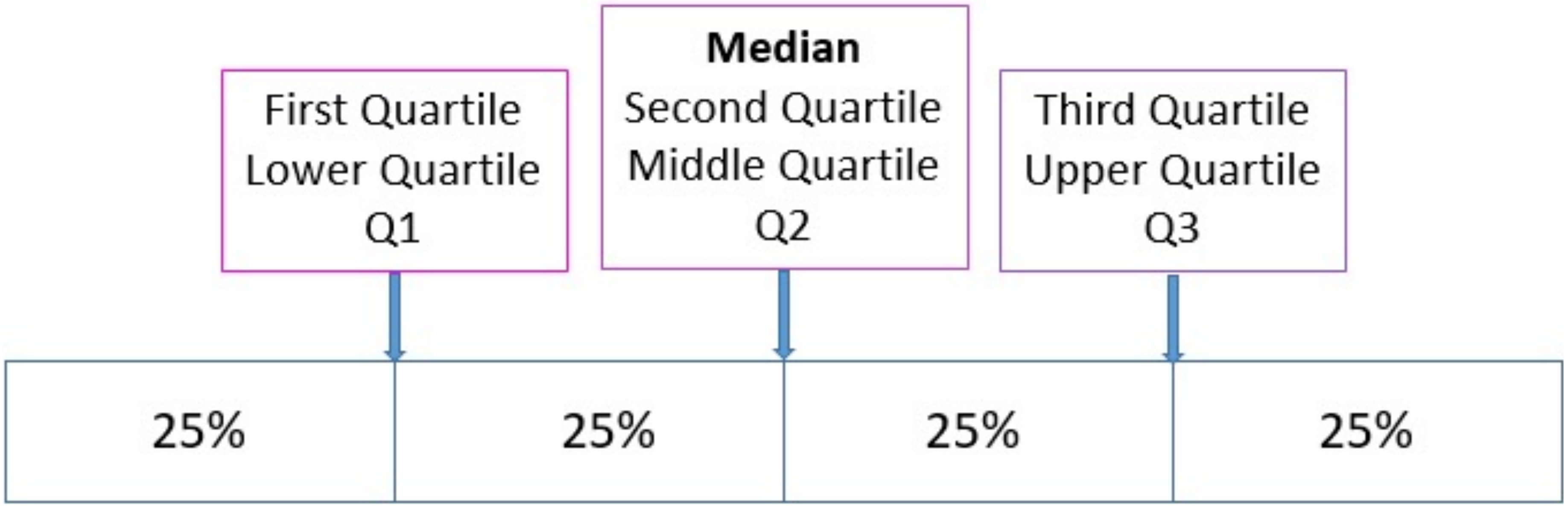




Describing Distributions

- Shape
- Center
- **(Measures of position)**
- Spread
- Outliers

Quartiles



Median, quartiles, percentiles(Video lessons, examples, solutions) [Internet]. www.onlinemathlearning.com. [cited 2021 Oct 4]. Available from: https://www.onlinemathlearning.com/quartile.html

Quartiles

- Recovery duration of 8 patients treated with a novel drug:

30, 20, 24, 40, 65, 70, 10, 62

10, 20, 24, $\underline{30, 40}$, 62, 65, 70

$Q_2 = 35$

x_1	x_2	x_3	x_4
10	20	24	30

$$Q_1 = \frac{20+24}{2} = 22$$

x_5	x_6	x_7	x_8
40	62	65	70

$$Q_3 = \frac{62+65}{2} = 63.5$$

Quartiles

- Systolic blood pressure measurements of 9 patients:

151, 124, 132, 170, 146, 124, 113, 111, 134

x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
111	113	124	124	132	134	146	151	170

Q_2

$$Q_1 = \frac{113 + 124}{2} = 118.5 \qquad Q_3 = \frac{146 + 151}{2} = 148.5$$

Percentiles - Definition

$100 * p$ percentile $(0 \le p \le 1)$ is the data value for which:

- at least $100 * p$ of the data values are less than or equal to it
- at least $100 * (1 - p)$ of the data values are greater than or equal to it

* If there are two values that satisfy the above conditions, the average of these values is taken as the $100 * p$ percentile

Percentiles - Algorithm

- Sort values in ascending order
- Calculate n * p
- If n * p is not an integer, take the smallest integer greater than n * p
- If n * p is an integer take the average of $n * p^{th}$ and $(n * p + 1)^{th}$ values

Percentiles - Example

- 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

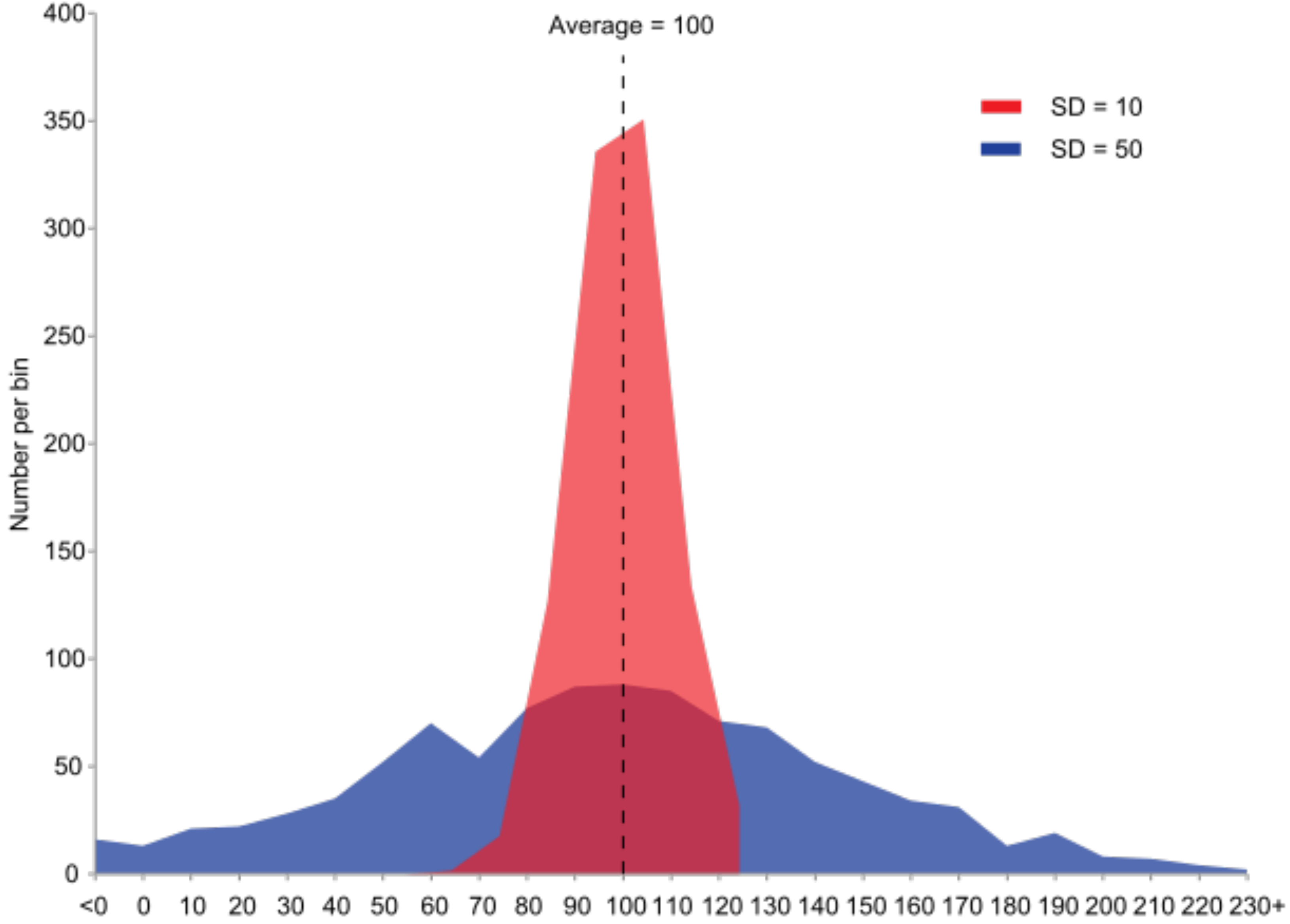
- 25th percentile (1st quartile, Q1): 189.5 (40 * 0.25 = 10)
- 50th percentile (median, Q2): 195.5 (40 * 0.5 = 20)
- 75th percentile (3rd quartile, Q3): 205.5 (40 * 0.75 = 30)
- 90th percentile : 218 (40 * 0.9 = 36)
- 95th percentile: 221 (40 * 0.95 = 38)
- 97.5th percentile: 224 (40 * 0.975 = 39)

Describing Distributions

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

Measures of Spread

- The distances of the values to the center differ
  - The degree of these differences constitute the spread of the distribution
- Two distributions may have the same mean/median/mode and differ in terms of spread



Average = 100
SD = 10
SD = 50
Number per bin
400
350
300
250
200
150
100
50
0
<0 0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230+

Range

- The difference between the maximal and minimal value

$$R = maximum - minimum$$

e.g., The ages of 12 arthritis patients:

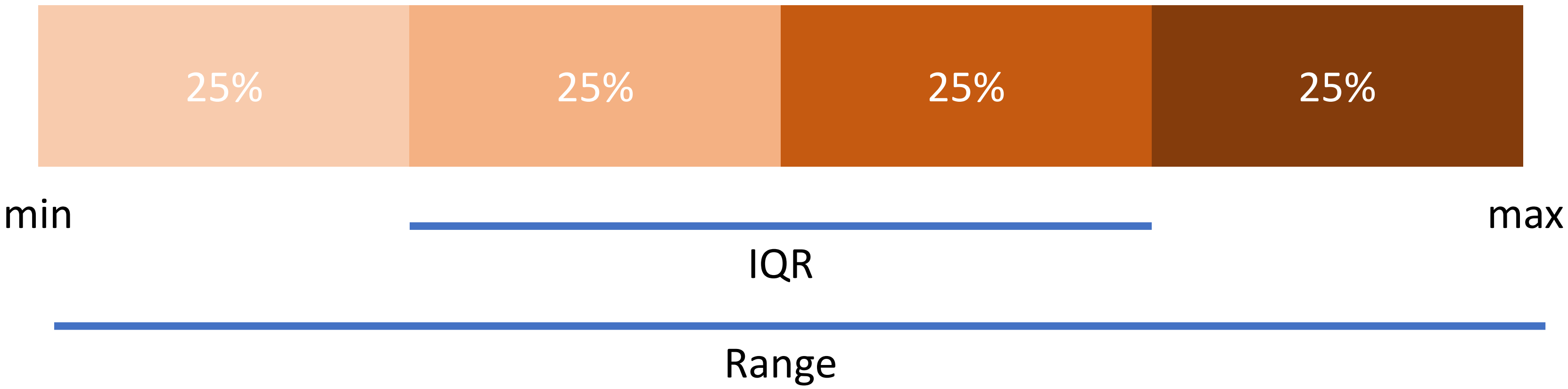
30, 12, 15, 22, 40, 55, 20, 58, 25, 60, 23, 72

$$R = 72 - 12 = 60$$

Inter-Quartile Range

- The range quantifies the variability by using the range covered by **all** the data
- the **Inter-Quartile Range (IQR)** measures the spread of a distribution by describing the range covered **by the middle 50%** of the data

$$\boldsymbol{IQR} = Q3 - Q1$$



Inter-Quartile Range

- Recovery durations of 8 patients in days:

  30, 20, 24, 40, 65, 70, 10, 62

  10, 20, 24, <u>30</u>, <u>40</u>, 62, 65, 70

$$\begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \\ 10 & 20 & 24 & 30 \end{array}$$

$$Q_1 = \frac{20+24}{2} = 22$$

$$\begin{array}{cccc} x_5 & x_6 & x_7 & x_8 \\ 40 & 62 & 65 & 70 \end{array}$$

$$Q_3 = \frac{62+65}{2} = 63.5$$

IQR = 63.5 - 22= 41.5

Variance and Standard Deviation

- Variance
  - A measure of how distant observations are from the mean
  - Population variance: σ^2
  - Sample variance: s^2
- Because **the unit of variance is quadratic**, standard deviation is more widely used

- Standard deviation (sd)
  - Defined as the square-root of variance
  - Population sd: σ
  - Sample sd: s

Sample Variance and Standard Deviation

$$s^2 = \frac{\sum_{j=1}^{n}(x_j - \bar{x})^2}{n-1}$$

Variance and Standard Deviation

Ages of 6 patients in a study:

10, 15, 22, 26, 31, 40

$\overline{x}$ = (10 + 15 + 22 + 26 + 31 + 40) / 6 = 24

$$s^2 = \frac{(10-24)^2+(15-24)^2+(22-24)^2+(26-24)^2+(31-24)^2+(40-24)^2}{6-1} = 118$$

$$s = \sqrt{s^2} = \sqrt{118} = 10.863$$

Units

- Mean: same unit with the data
- Median: same unit with the data
- Mode: same unit with the data
- Quartiles: same unit with the data
- Percentiles: same unit with the data
- Variance: square of the unit of the data
- Standard deviation: same unit with the data

Describing Distributions

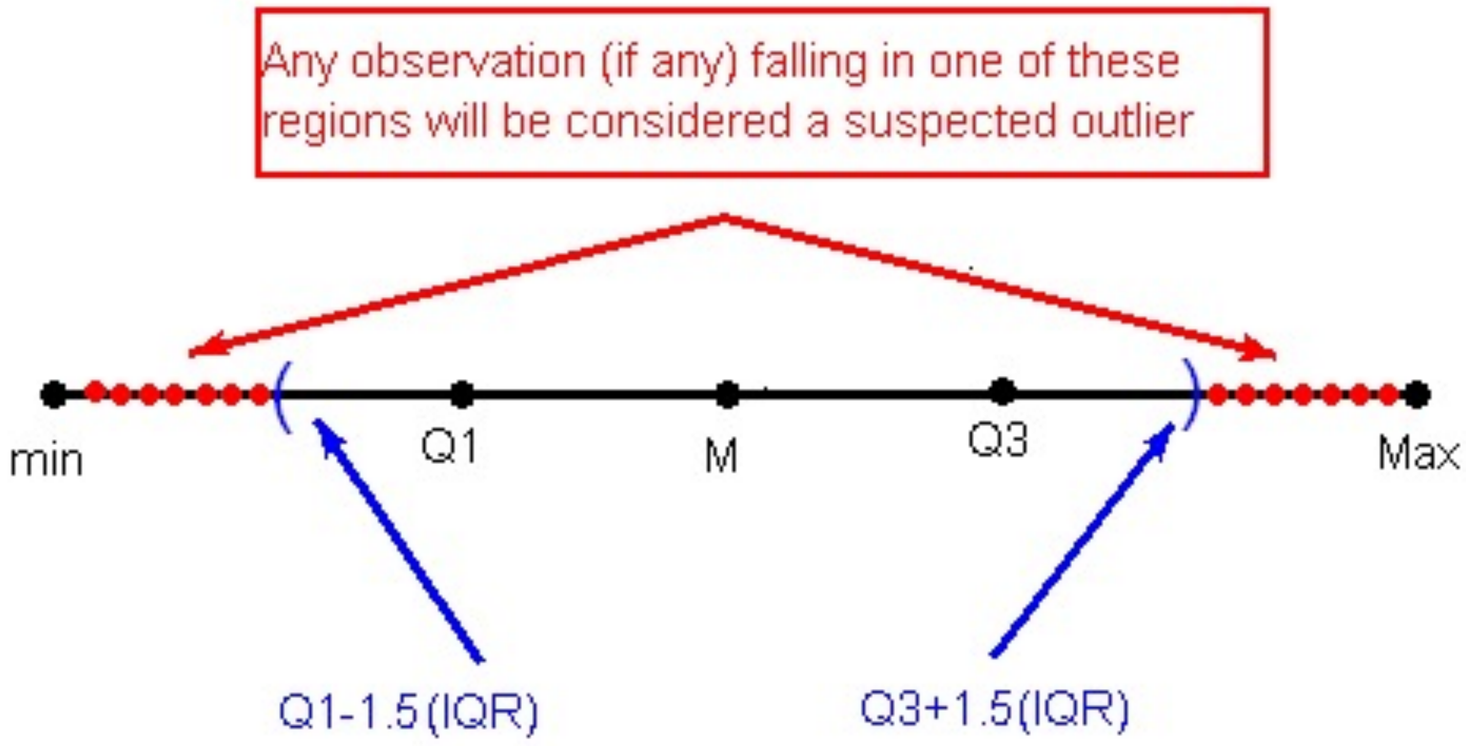
- Shape
- Center
- Spread
- **Outliers**

Outliers

- Extreme observations that are distant from the rest of the data

- For
  - Lower Limit = Q_1 - 1.5 * IQR
  - Upper Limit = Q_3 + 1.5 * IQR
- Outliers are defined as any value(s) larger than the upper limit or smaller than the lower limit

Outliers



Outliers [Internet]. [cited 2021 Oct 4]. Available from: https://bolt.mph.ufl.edu/6050-6052/unit-1/one-quantitative-variable-introduction/understanding-outliers/

Outliers – Cholesterol Level Example

- 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

- 25^{th} percentile (1st quartile, Q_1): 189.5 (40 * 0.25 = 10)
- 75th percentile (3rd quartile, Q_3): 205.5 (40 * 0.75 = 30)
- IQR = 205.5 - 189.5 = 16

- LL = Q_1 - 1.5 * IQR = 189.5 - 1.5 * 16 = 165.5
- UL = Q_3 + 1.5 * IQR = 205.5 + 1.5 * 16 = 229.5

- **No outliers**

Outliers – Cholesterol Level Example (cont.)

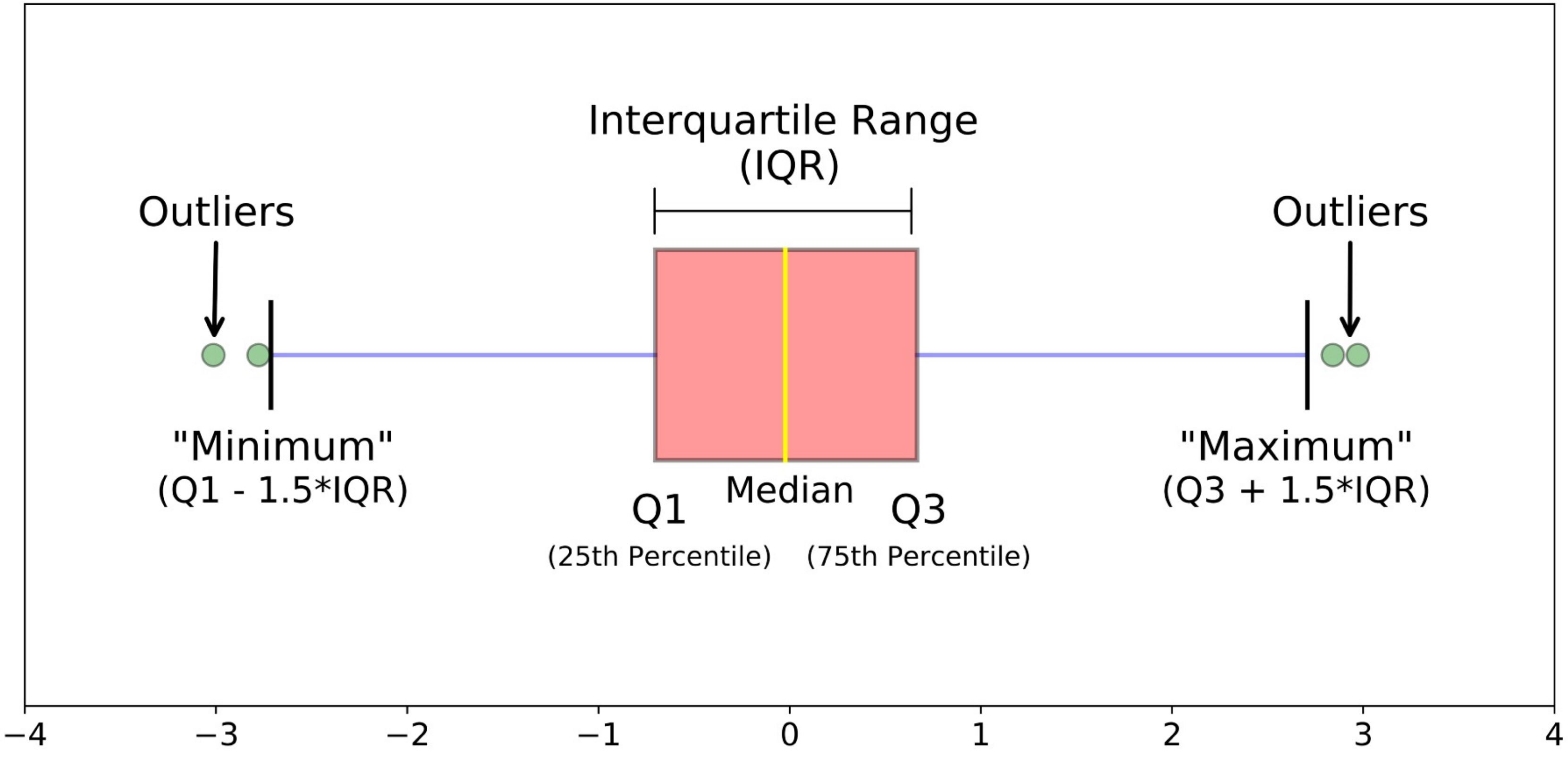
- Sorted data: 171, 174, 178, 181, 183, 183, 183, 184, 187, 188, 191, 191, 191, 192, 193, 193, 193, 194, 194, 195, 196, 196, 199, 200, 200, 200, 202, 204, 204, 205, 206, 211, 212, 213, 213, 216, 220, 221, 221, **300**

- 25^{th} percentile (1st quartile, Q_1): 189.5 (40 * 0.25 = 10)
- 75th percentile (3rd quartile, Q_3): 205.5 (40 * 0.75 = 30)
- IQR = 205.5 - 189.5 = 16

- LL = Q_1 - 1.5 * IQR = 189.5 - 1.5 * 16 = 165.5
- UL = Q_3 + 1.5 * IQR = 205.5 + 1.5 * 16 = 229.5

- **300 > UL => outlier**

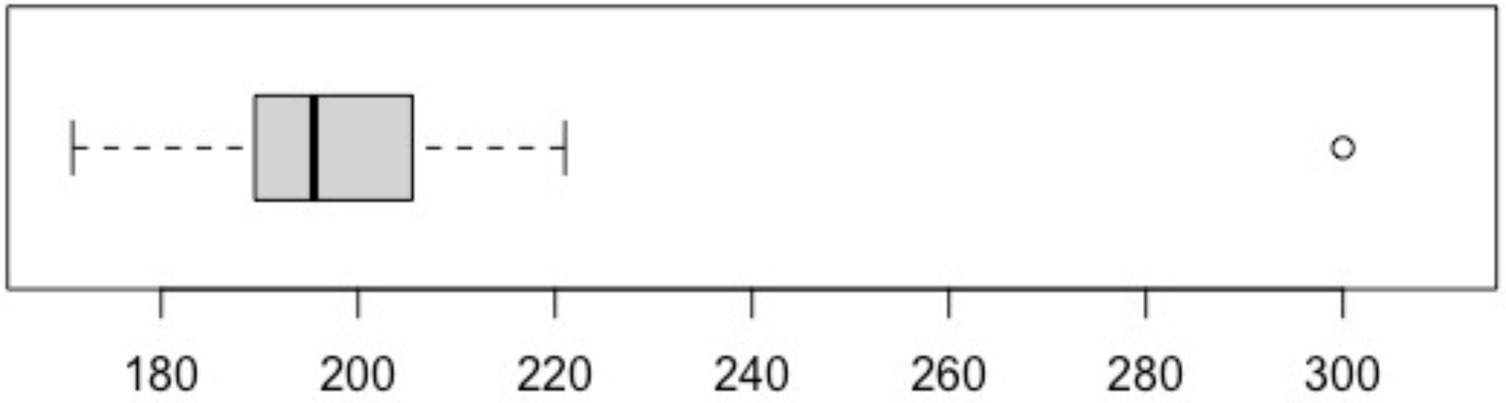
Box Plot

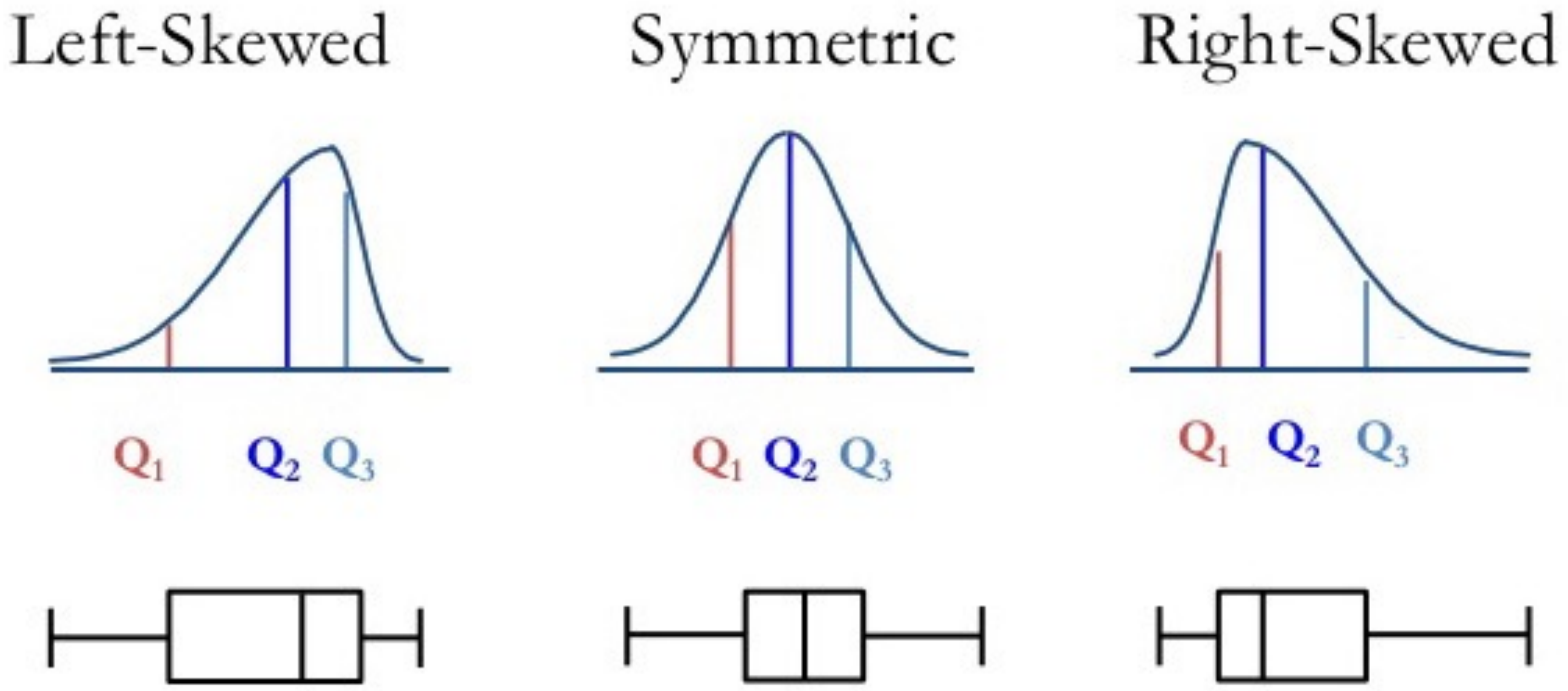


Galarnyk M. Understanding boxplots [Internet]. Medium. 2020 [cited 2021 Oct 4]. Available from: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

Box Plot – Example

- 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, **300**





Box plot | simply psychology [Internet]. [cited 2021 Oct 4]. Available from: https://www.simplypsychology.org/boxplots.html

Brief Summary

- Shape of a distribution can be described using skewness and modality
- Center of a distribution can be described using mean, median, mode
  - Median is more robust to outliers
- Quartiles and percentiles can be used to partition the data
- Variance and standard deviation are the most frequently used measures of spread
- Outliers can be defined based on Q1, Q3 and IQR
- Box plots can be used to display the distribution of a continuous variable
  - displays Q1, median, Q3, outliers