Biostatistics
Week II

Ege Ülgen, M.D.

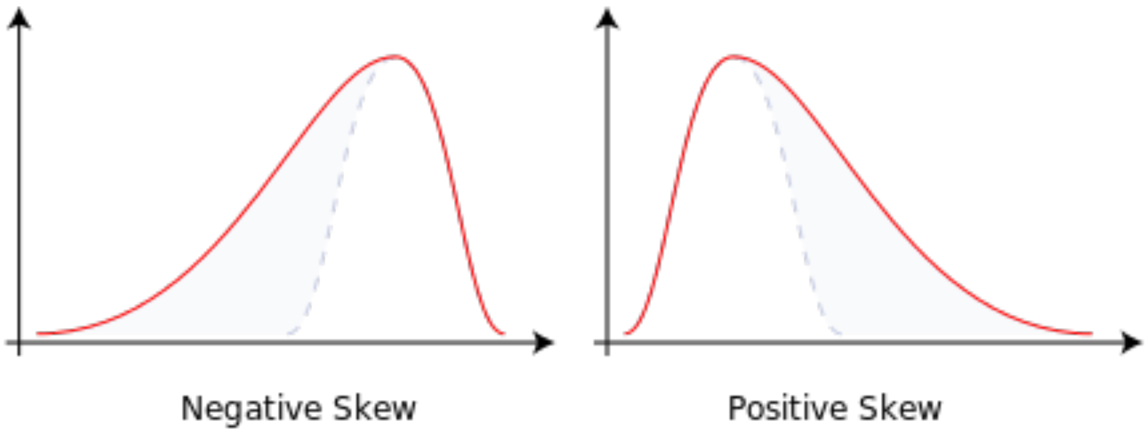
14 October 2021

ACIBADEM
MEHMET ALİ AYDINLAR
ÜNİVERSİTESİ

Describing Distributions

- **Shape – skewness, modality/*kurtosis***
- Center
- (Measures of position)
- Spread
- Outliers

Skewness



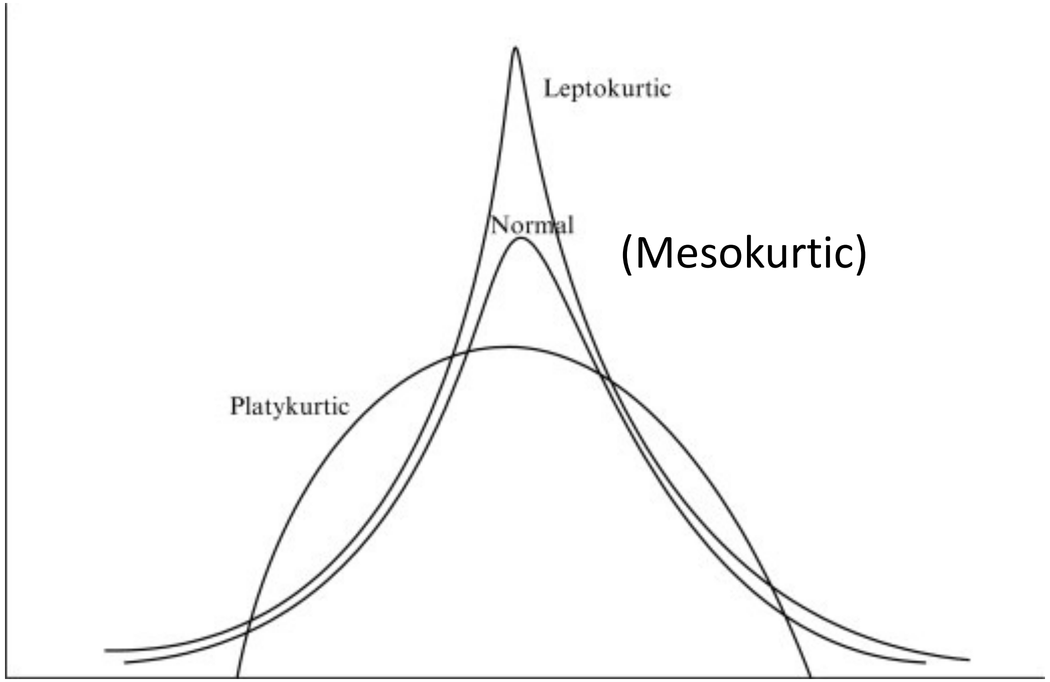
- There are several ways to measure skewness
- 2 common:
  - Pearson's first skewness coefficient (mode skewness) $= \frac{\bar{X} - Mode(X)}{s}$
  - Pearson’s second skewness coefficient (median skewness) $= \frac{3[\bar{X} - Median(X)]}{s}$

Ibe, Oliver C. (2014). Fundamentals of Applied Probability and Random Processes || Introduction to Descriptive Statistics. , (), 253–274. doi:10.1016/B978-0-12-800852-2.00008-0

Kurtosis

Sample excess kurtosis (biased)

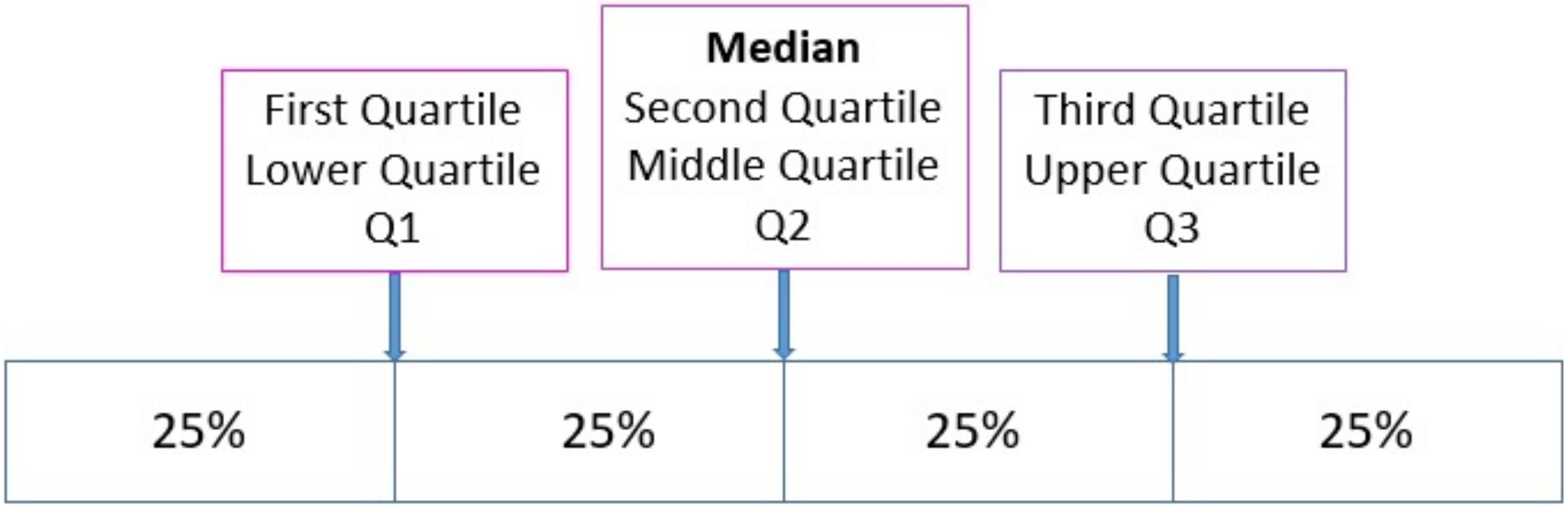
$$g_2 = \frac{m_4}{m_2^2} - 3 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^4}{\left[\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2\right]^2} - 3$$



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

$Q_2 = 35$

$$\begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \\ 10 & 20 & 24 & 30 \end{array} \qquad \begin{array}{cccc} x_5 & x_6 & x_7 & x_8 \\ 40 & 62 & 65 & 70 \end{array}$$

$$Q_1 = \frac{20+24}{2} = 22 \qquad Q_3 = \frac{62+65}{2} = 63.5$$

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

1. Sort data X in ascending order
2. Calculate $n \times p$
3. If np is not an integer, return $X_{ceiling(np)}$
4. Else (if $n \times p$ is an integer), return $(X_{np} + X_{np+1})/2$

Percentiles – simple example

- Original data: 13, 14, 12, 11, 19, 15, 18, 16, 17, 20 (n = 10)
- Sorted data: 11, 12, 13, 14, 15, 16, 17, 18, 19, 20

- 25th percentile (1st quartile): 13 (10 * 0.25 = 2.5 >> 3)
- 50th percentile (median): 15.5 (10 * 0.5 = 5 >> 5 & 6)
- 75th percentile (3rd quartile): 18 (10 * 0.75 = 7.5 >> 8)
- 90th percentile: 19.5 (10 * 0.9 = 9 >> 9 & 10)
- 95th percentile: 20 (10 * 0.95 = 9.5 >> 10)
- 97.5th percentile: 20 (10 * 0.975 = 9.75 >> 10)

Percentiles – another 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)

Quantiles – general formula

$Q(q) = (1-\gamma)X_j + \gamma X_{j+1}$ where:

- $\frac{j-m}{n} \leq q \leq \frac{j-m+1}{n}$
- $m \in \mathbb{R}$
- $0 \leq \gamma \leq 1$ and γ is a function of j and g
- $j = floor(qn+m)$ and $g = qn + m - j$

Type 7[2]: $\gamma = g$ and $m = 1 - p$

Type 2: $m = 0$ and $\gamma = 0.5$ when $g = 0$ and $\gamma = 1$ when $g > 0$

[2] By default, R uses **Type 7**

Quantiles – A simple example (type 7)

Original data: 13, 14, 12, 11, 19, 15, 18, 16, 17, 20 (n = 10)

Sorted data: 11, 12, 13, 14, 15, 16, 17, 18, 19, 20

$p = 0.25$,

- $m = 1 - p = 0.75$,
- $j = \text{floor}(n * p + m) = \text{floor}(10 * 0.25 + 0.75) = 3$
- $g = p * n + m - j = 0.25 * 10 + 0.75 - 3 = 0.25$
- $\gamma = g = 0.25$
- $Q(0.25) = (1 - 0.25) * 13 + 0.25 * 14 = 13.25$

Quantiles – A simple example (type 7)

Original data: 13, 14, 12, 11, 19, 15, 18, 16, 17, 20 (n = 10)

Sorted data: 11, 12, 13, 14, 15, 16, 17, 18, 19, 20

$p = 0.5$,

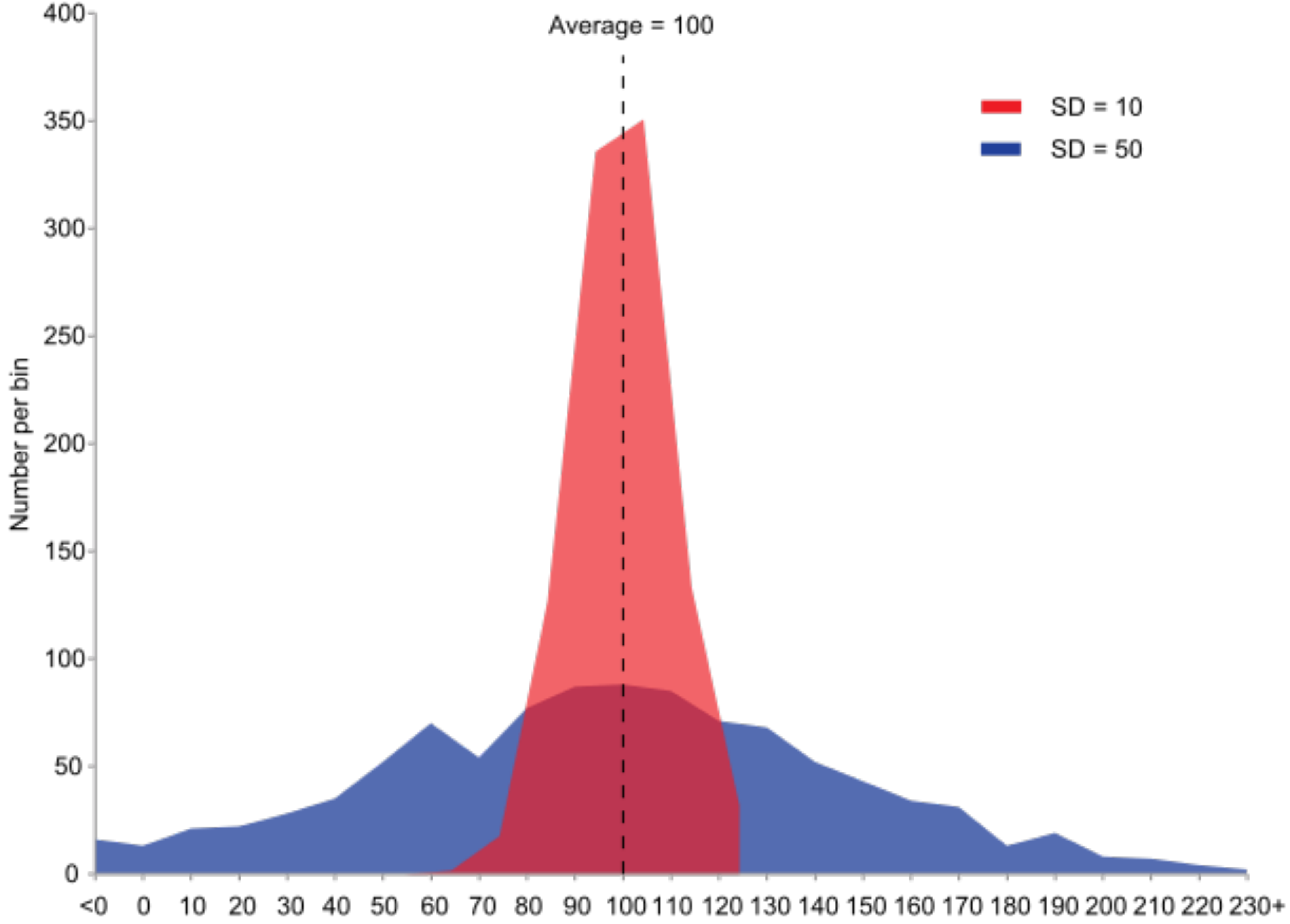
- $m = 1 - p = 0.5$,
- $j = \text{floor}(pn + m) = \text{floor}(10 * 0.5 + 0.5) = 5$
- $g = pn + m - j = 0.5 * 10 + 0.5 - 5 = 0.5$
- $\gamma = g = 0.5$
- $Q(0.5) = (1 - 0.5) * 15 + 0.5 * 16 = 15.5$

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
0
50
100
150
200
250
300
350
400
<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:

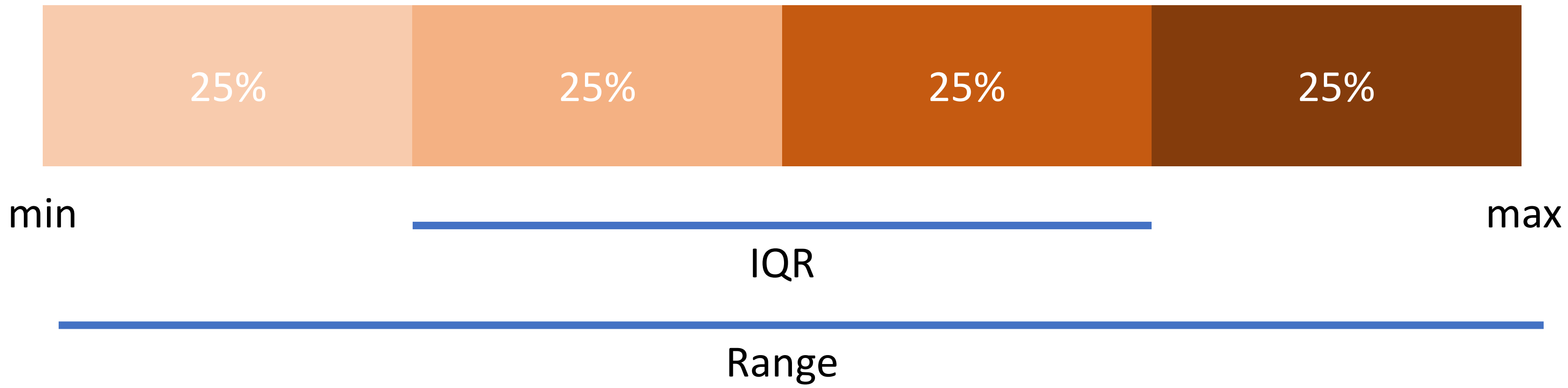
30, 12, 15, 22, 40, 55, 20, 58, 25, 60, 23, 72

$$R = 72 - 12 = 60$$

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

$$\begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \\ 10 & 20 & 24 & 30 \end{array} \qquad \begin{array}{cccc} x_5 & x_6 & x_7 & x_8 \\ 40 & 62 & 65 & 70 \end{array}$$

$$Q_1 = \frac{20+24}{2} = 22 \qquad Q_3 = \frac{62+65}{2} = 63.5$$

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

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

Sample Variance and Standard Deviation

Ages of 6 patients in a study:

10, 15, 22, 26, 31, 40

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

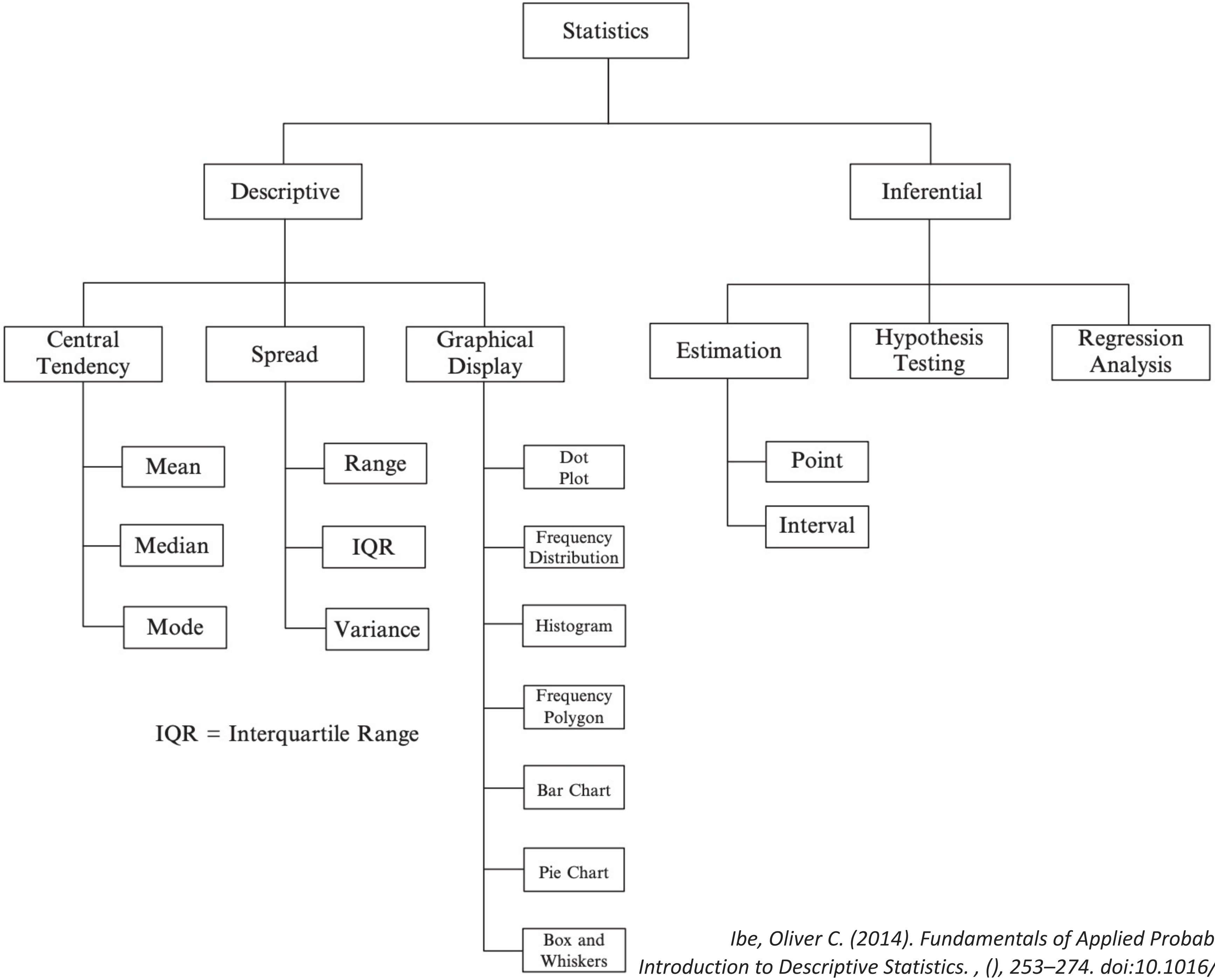
$$s^2 = \frac{(10-24)^2+(15-24)^2+(22-24)^2+(26-24)^2+(31-24)^2+(40-24)^2}{6-1} = 118$$

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

Sample Variance and Standard Deviation

If $y = x + c$, where c is a constant, $\text{var(y)} = \text{var(x)}$

If $z = x * c$, where c is a constant, $\text{var(z)} = c^2\text{var(x)}$



Ibe, Oliver C. (2014). Fundamentals of Applied Probability and Random Processes || Introduction to Descriptive Statistics. , (), 253–274. doi:10.1016/B978-0-12-800852-2.00008-0

Describing Distributions

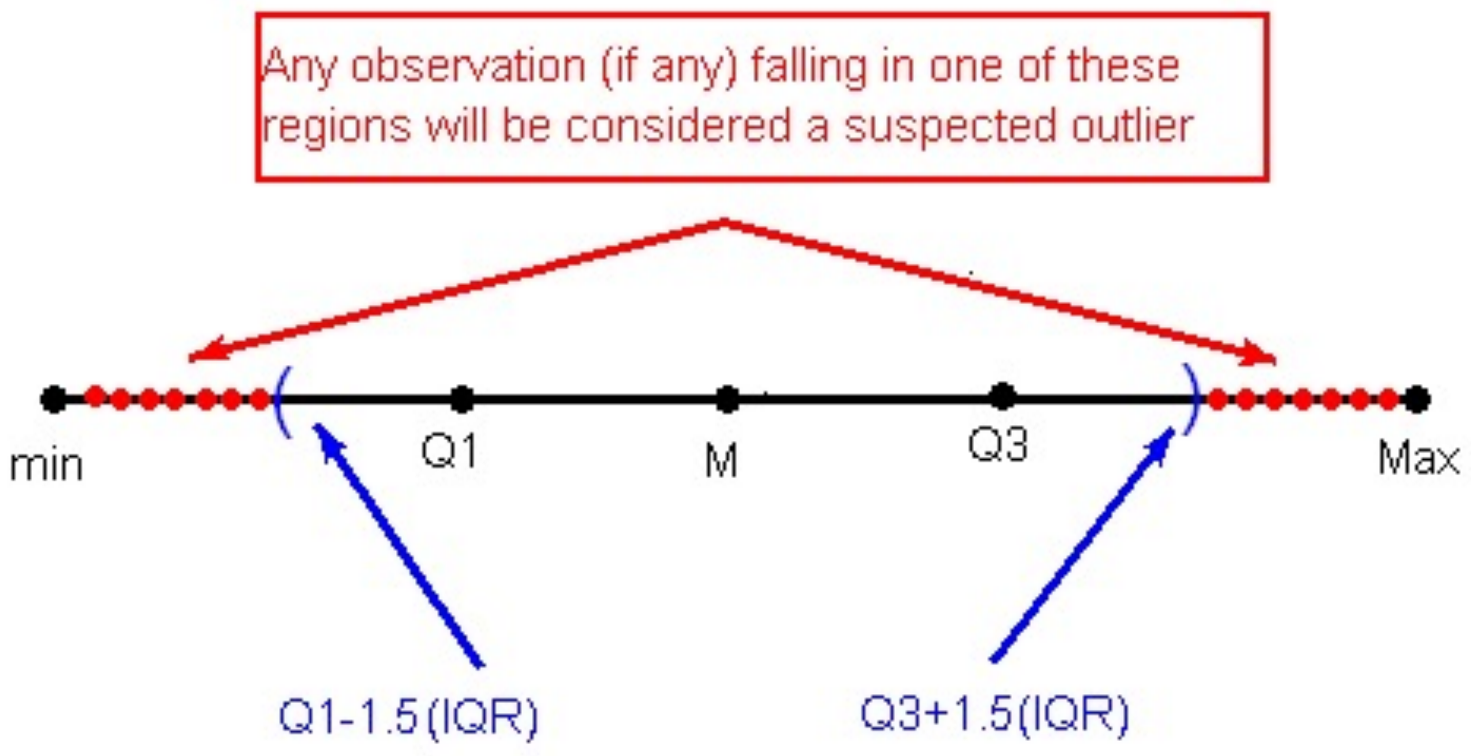
- Shape
- Center
- Spread
- **Outliers**

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.)

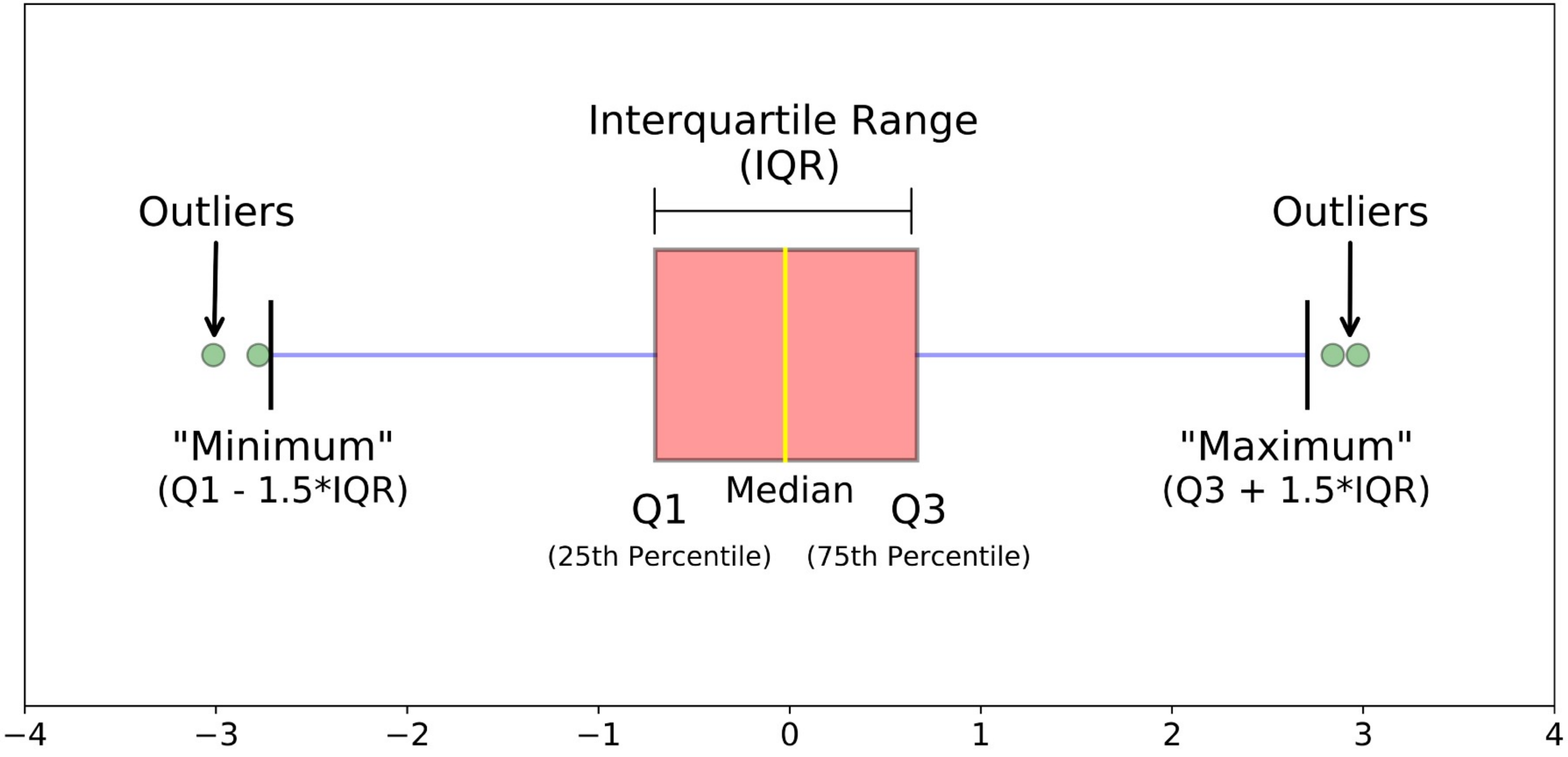
- Sorted data: 171, 174, 178, 181, 183, 183, 183, 184, 187, 188, 191, 191, 191, 192, 193, 193, 193, 194, 194, 195, 196, 196, 199, 200, 200, 200, 202, 204, 204, 205, 206, 211, 212, 213, 213, 216, 220, 221, 221, **300**

- 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**

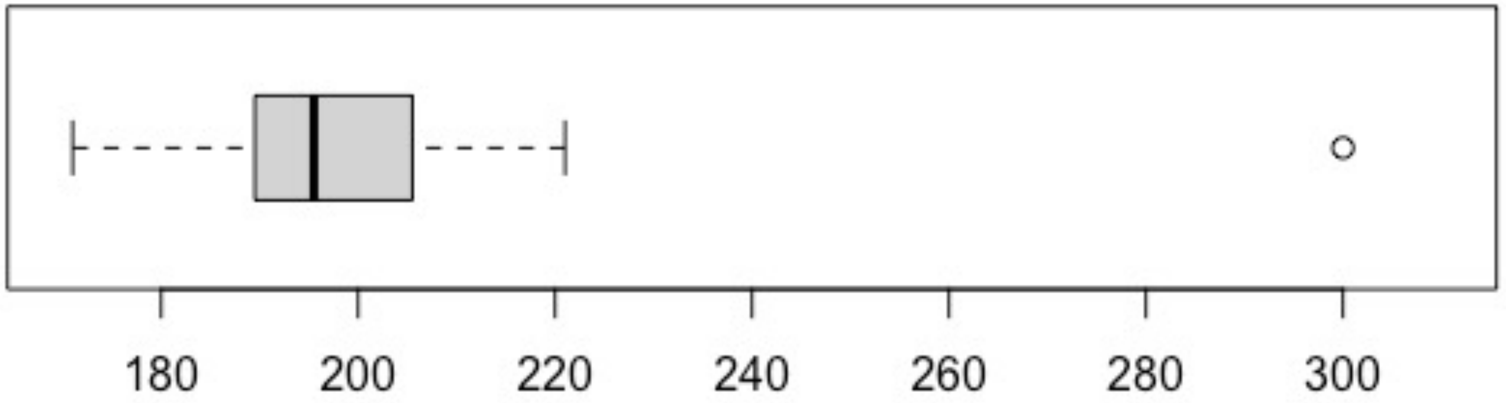
Box Plot

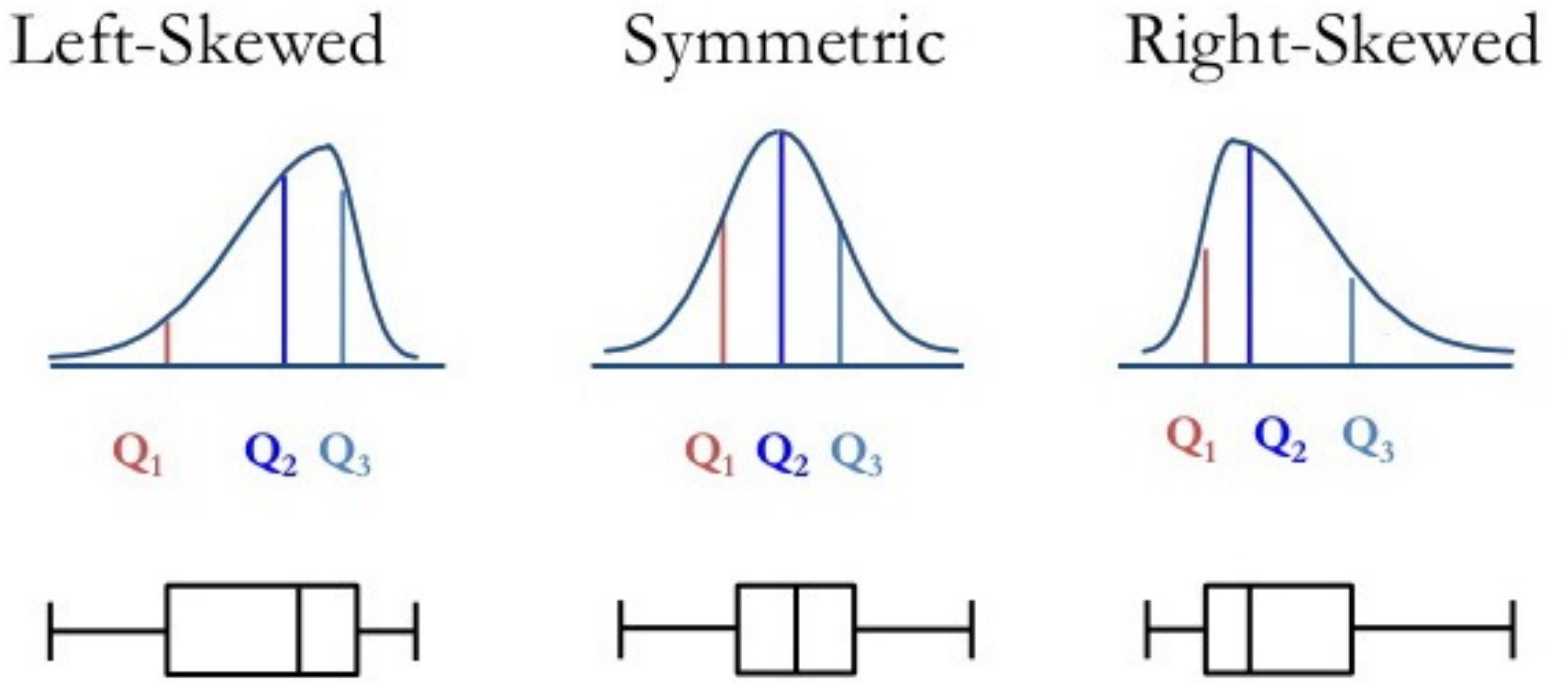


Galarnyk M. Understanding boxplots [Internet]. Medium. 2020 [cited 2021 Oct 4]. Available from: https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51

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**



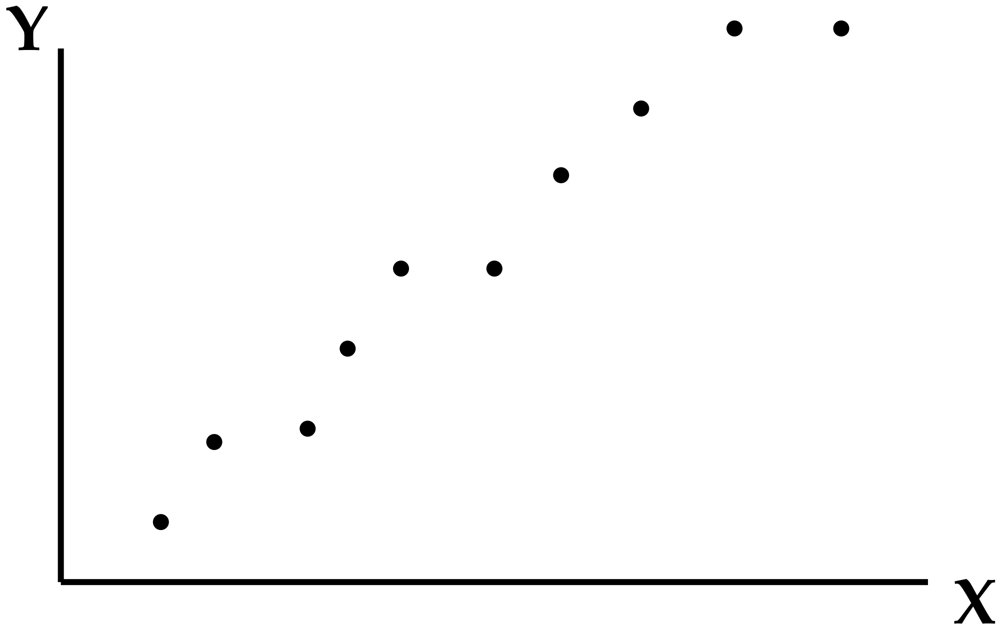


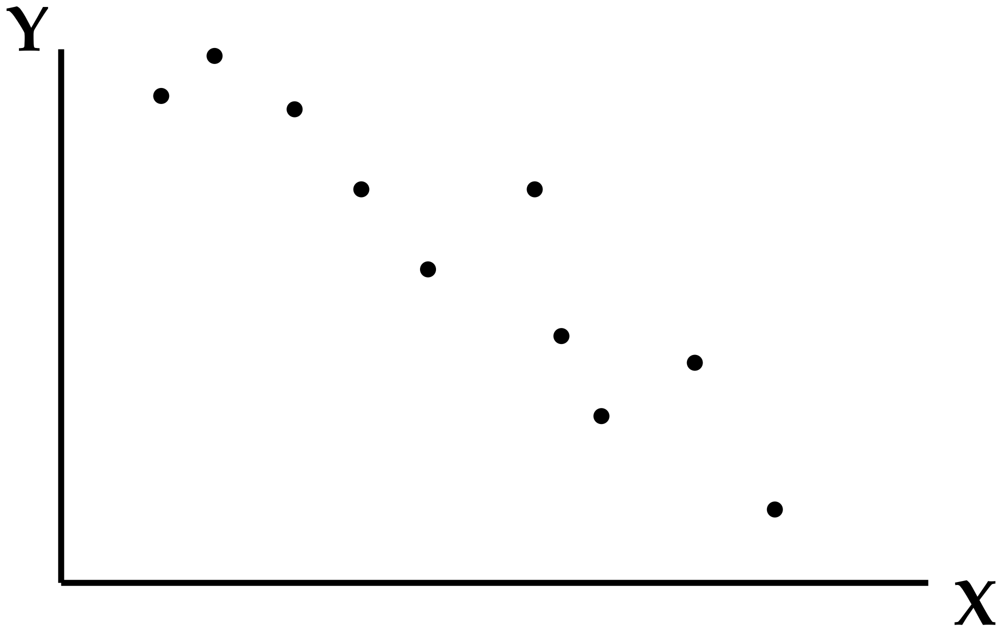
Box plot | simply psychology [Internet]. [cited 2021 Oct 4]. Available from: https://www.simplypsychology.org/boxplots.html

Exploratory Data Analysis (EDA)

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

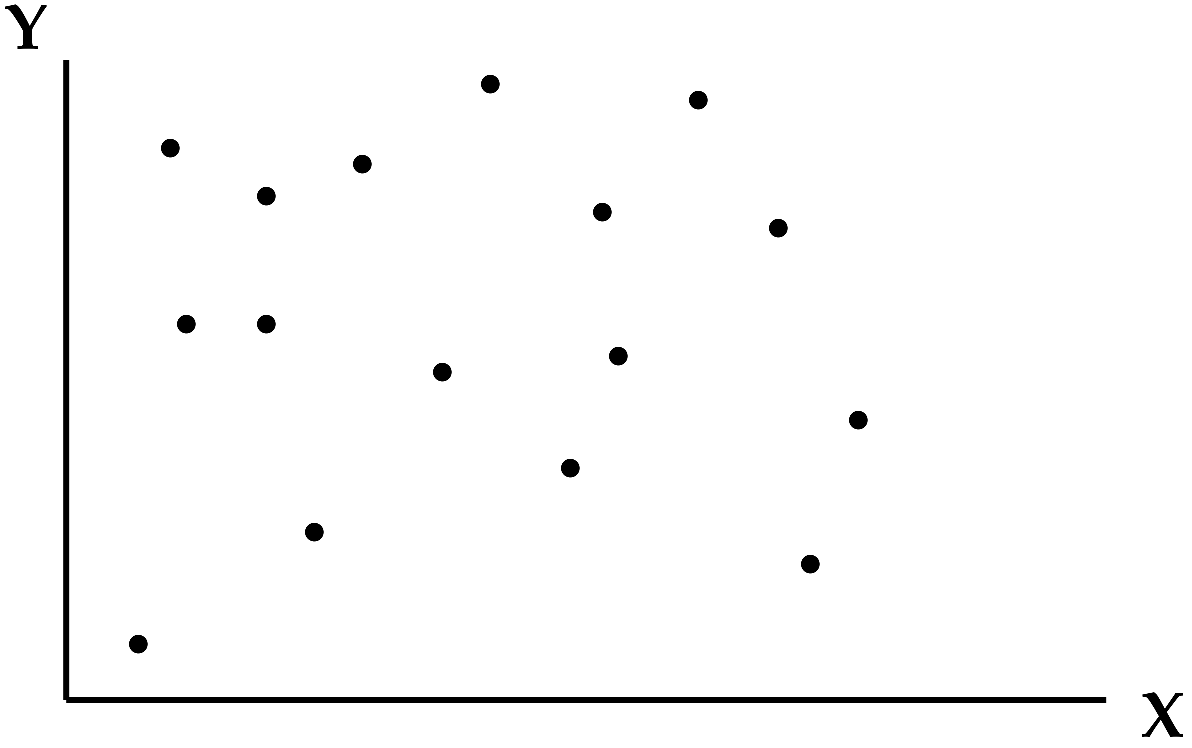
Relationship between two variables





Linear relationship

Relationship between two variables



Sample Covariance

- A measure of how two variables change together

$$Cov(X,Y) = \frac{1}{n-1}\sum_{i=1}^{n}(X - \bar{X})(Y - \bar{Y})$$

Sample Covariance

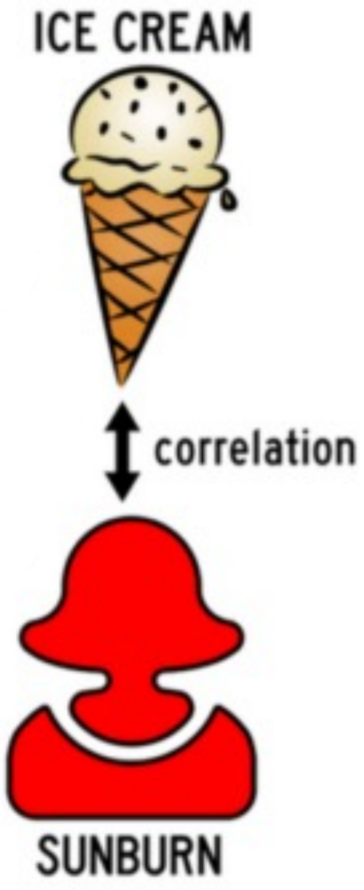
Proporties:

- $Cov(X,Y) \in \mathbb{R}$
- $Cov(X,Y) = Cov(Y,X)$
- $Cov(X,X) = Var(X)$
- $Cov(aX,bY) = abCov(X,Y),\ a,b \in \mathbb{R}$
- $Cov(X+a,Y+b) = Cov(X,Y),\ a,b \in \mathbb{R}$

Correlation

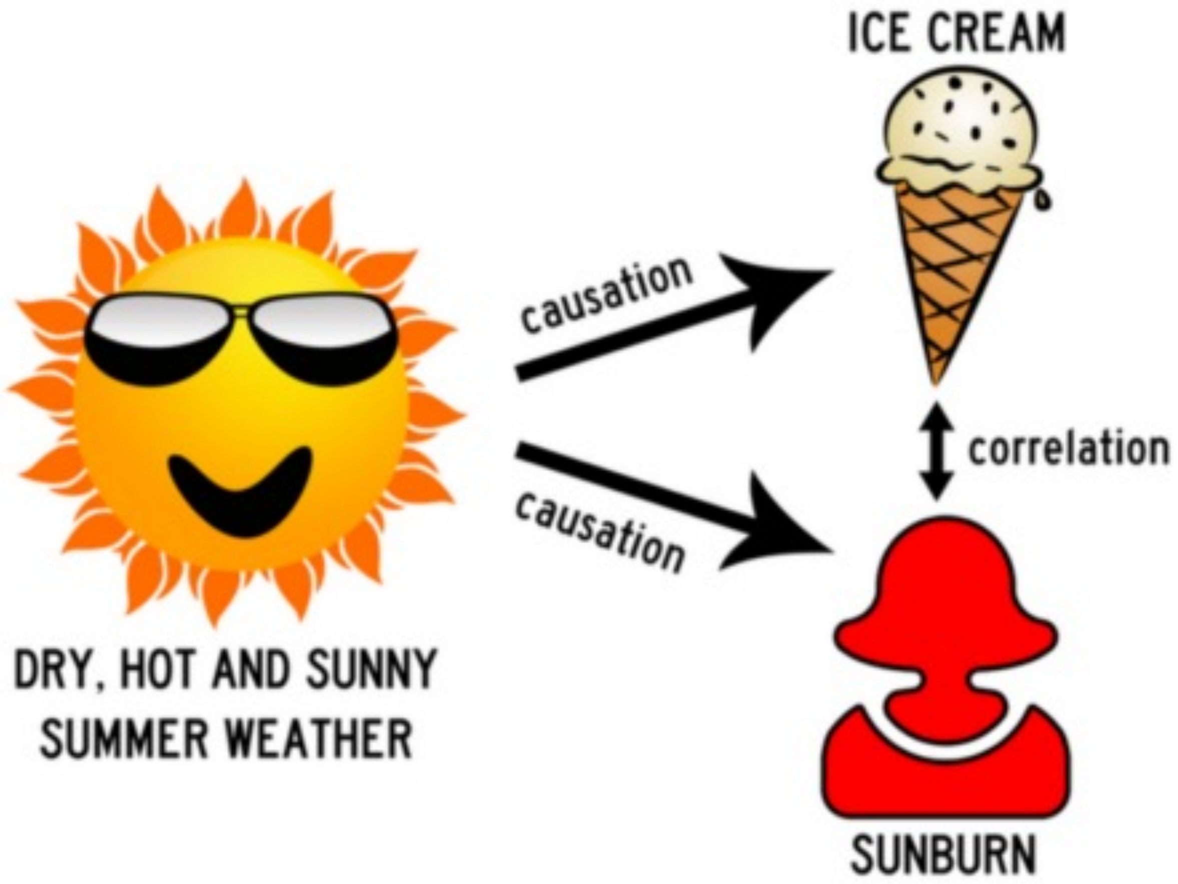
- Correlation is a bivariate analysis that measures **the strength of association** between two variables and **the direction** of the relationships
- In terms of the strength of relationship, the value of the correlation coefficient varies **between +1 and -1**
- **Correlation does not mean causation**

Correlation does not mean causation



Figueroa A. Correlation is not causation [Internet]. Medium. 2019 [cited 2021 Oct 8]. Available from: https://towardsdatascience.com/correlation-is-not-causation-ae05d03c1f53

Correlation does not mean causation



Figueroa A. Correlation is not causation [Internet]. Medium. 2019 [cited 2021 Oct 8]. Available from: https://towardsdatascience.com/correlation-is-not-causation-ae05d03c1f53

Correlation Coefficient

- A statistic that measures the relationship between two variables

- Pearson’s r
    - Measures **linear** relationship
    - Both variables have to be normally distributed
- Spearman’s ρ
    - Measures **monotonic** relationship
    - Based on rank – non-parametric

Pearson Correlation Coefficient

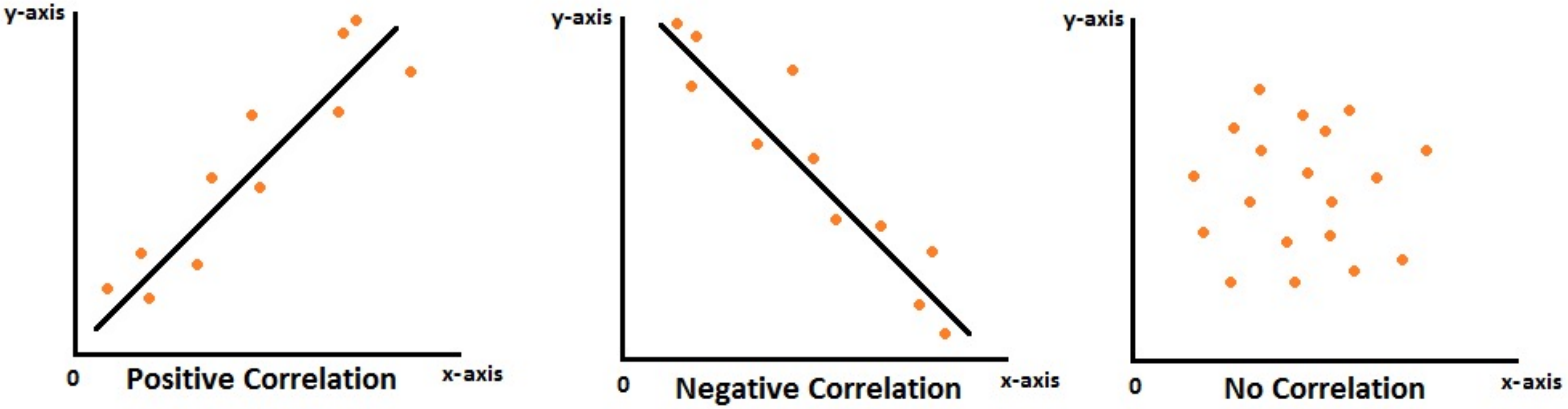
$$r_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

- A measure of the **linear** correlation between two variables X and Y
- takes values between -1 and 1
- unitless
- $r_{X,Y} = r_{Y,X}$
- $r_{X,Y} = 0$ means **no linear relationship**

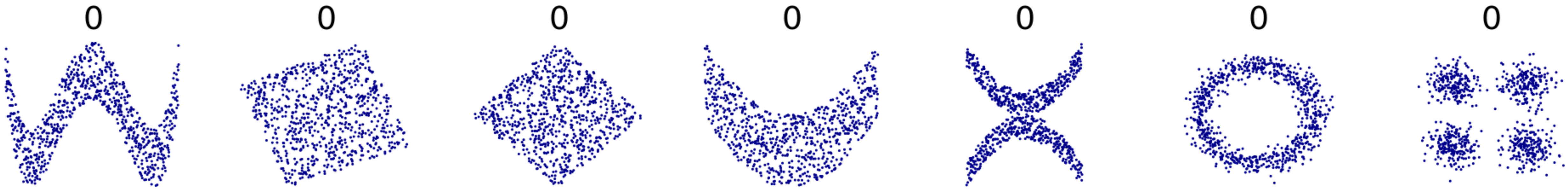
Pearson Correlation Coefficient

Cohen's (1988) conventions to interpret effect size:

- $|r| = 0.10 - 0.29$: Weak
- $|r| = 0.30 - 0.49$: Moderate
- $|r| \geq 0.50$: Strong

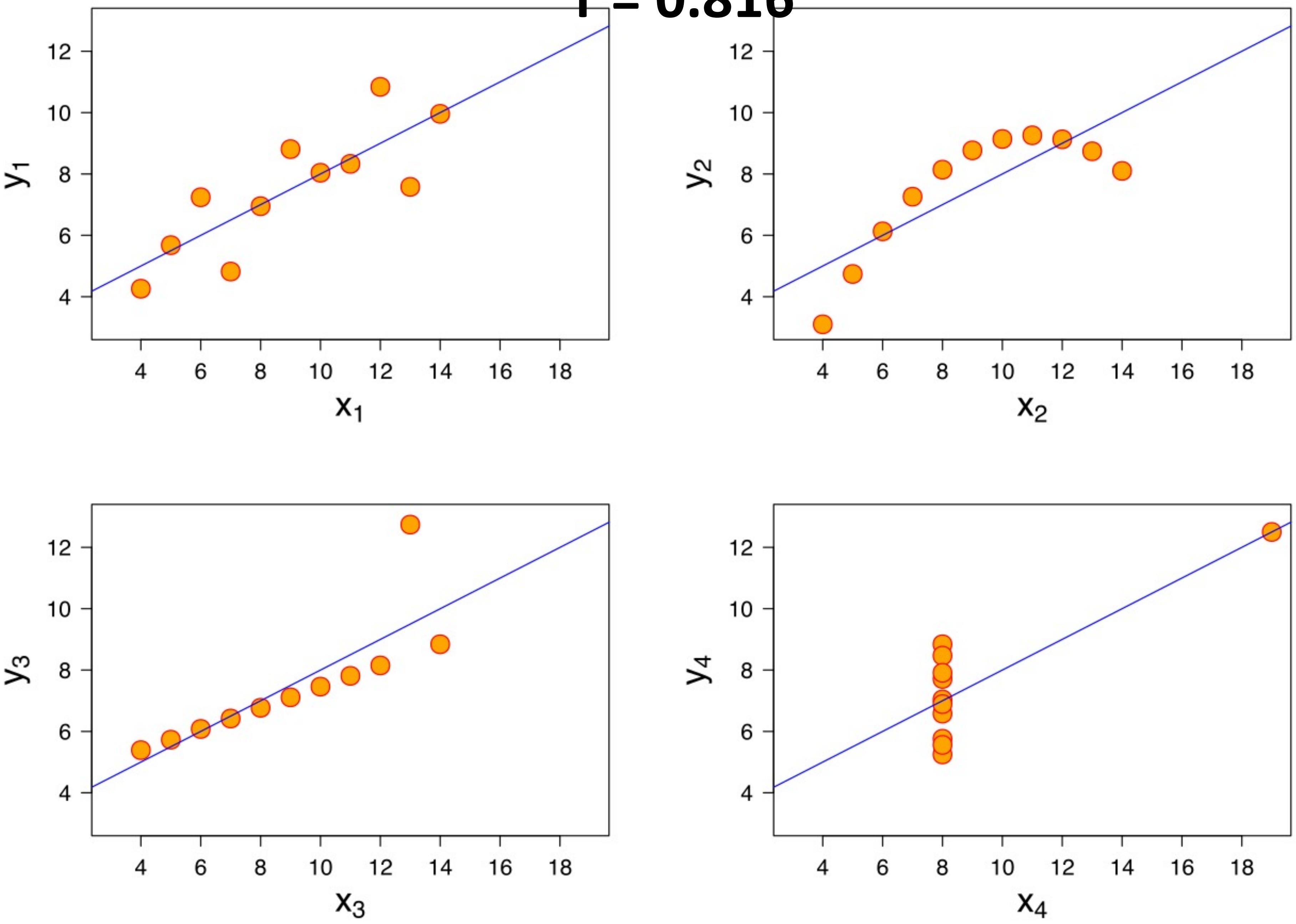


Suresh S. From correlation to causation [Internet]. Medium. 2020 [cited 2021 Oct 8]. Available from:
https://towardsdatascience.com/from-correlation-to-causation-49f566eea954



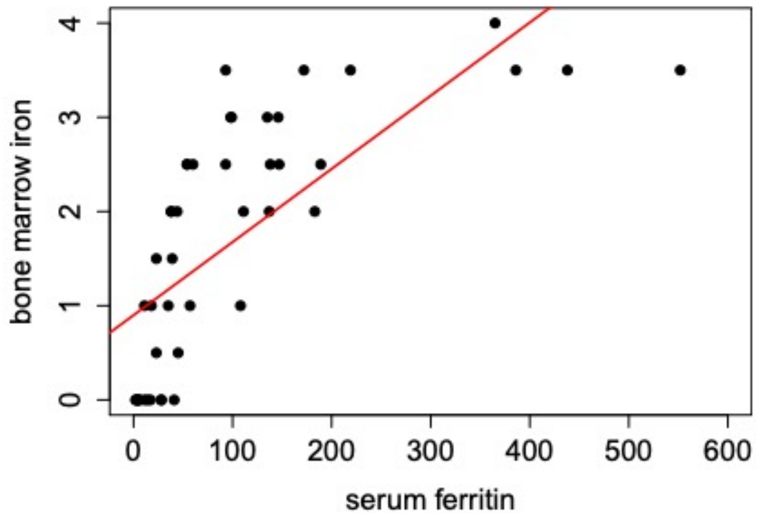
https://en.wikipedia.org/wiki/Correlation_and_dependence

r = 0.816



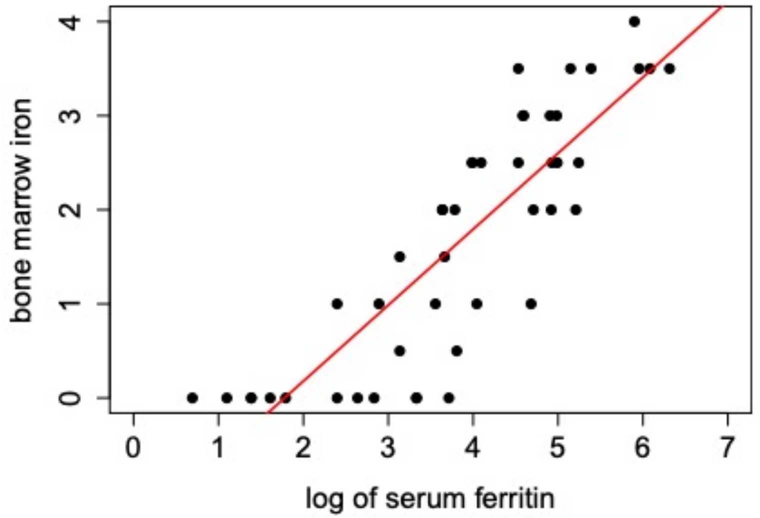
https://en.wikipedia.org/wiki/Correlation_and_dependence

Example: Relation between blood serum content of Ferritin and bone marrow content of iron.



$r = 0.72$

- Transformation to linear relation?
- Frequently a transformation to the normal distribution helps.



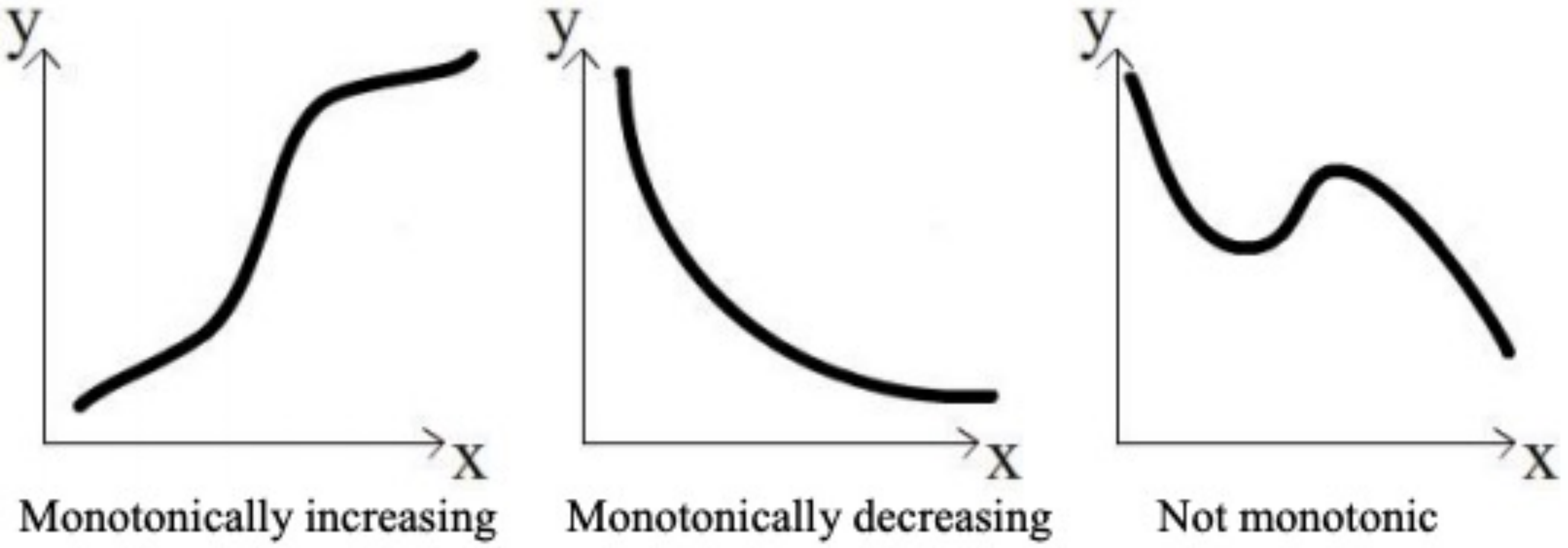
$r = 0.85$

https://www.ebpi.uzh.ch/dam/jcr:ffffffff-c1f2-5119-ffff-ffffacc68d53/slides-04.pdf

Spearman Rank Correlation

- It assesses how well the relationship between two variables can be described **using a monotonic function**
- It **does not carry any assumptions about the distribution** of the data and is the appropriate correlation analysis when the variables are measured on a scale that is at least ordinal

Spearman Rank Correlation



Spearman Rank Correlation

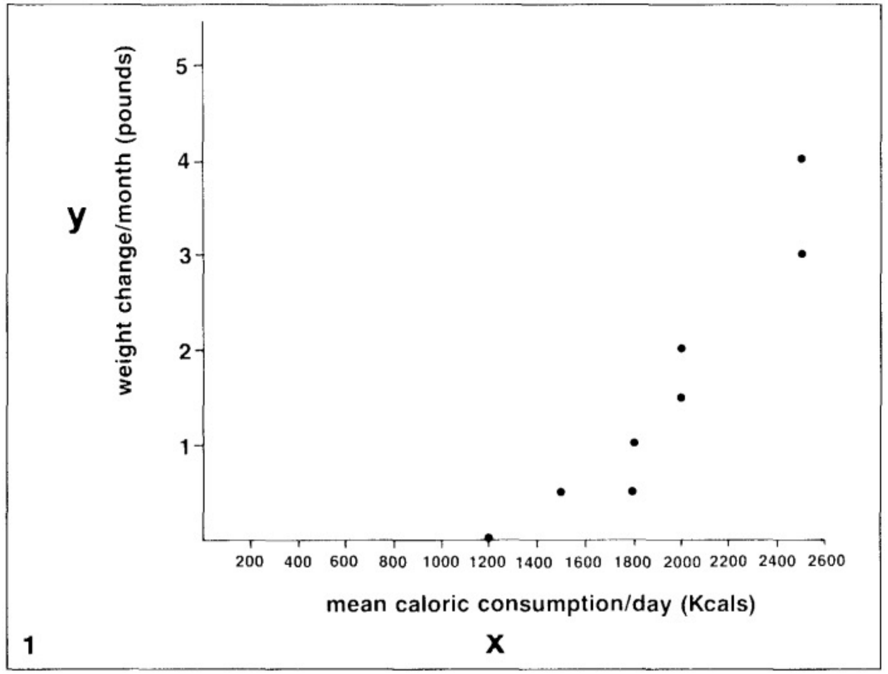
$$\rho = 1 - \frac{6 \sum_{i=1}^{n} d_i^2}{n(n^2 - 1)}$$

- d_i := the difference between the ranks of corresponding variables (i.e., $d = X_i - Y_i$)
- n := number of observations

TABLE 1. *Sample data: Caloric consumption versus weight change*

Patient	(X) Mean Caloric Consumption/Day	(Y) Weight Change/ Month
1	1,200	0.0
2	1,500	0.5
3	1,800	0.5
4	2,000	1.5
5	2,500	4.0
6	1,800	1.0
7	2,500	3.0
8	2,000	2.0

FIGURE 1. *Scatter diagram for sample data given in Table 1 (caloric consumption vs weight change).*



There is a strong positive relationship between mean caloric consumption/day and weight change/month

$r = 0.94$ or
$\rho = 0.97$

Gaddis ML, Gaddis GM. Introduction to biostatistics: Part 6, Correlation and regression. Ann Emerg Med. 1990 Dec;19(12):1462–8.

Units

- Mean: same unit with the data
- Median: same unit with the data
- Mode: same unit with the data
- Quantiles: same unit with the data
- Variance: square of the unit of the data
- Standard deviation: same unit with the data
- Covariance: square of the unit of the data
- Correlation: unitless

Brief Summary

- Quantiles can be used to partition the data and calculate specific positions
- The most used measures of spread are:
  - IQR
  - Variance and standard deviation
- 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
- The relationship between two variables can be visualized using scatter plots
- The relationship between two variables can be assessed using correlation
  - Pearson
  - Spearman