Descriptive statistics & measures of association

Lukáš Lehotský & Petr Ocelík

ESS401 Social Science Methodology / MEB431 Metodologie sociálních věd

27th February 2017

Outline

- Measures of central tendency, position, and variability
- Graphic displays of descriptive statistics
- Measures of association

Descriptive statistics

- The purpose is to **summarize data**.
- Quantitative variables have two key features:
  - The **center** of the data – a typical observation.
  - The **variability** of the data – the spread around the center.

Notation

	Mean	Standard Deviation	Variance
Population	μ	σ	σ^2
Sample	$\bar{x}$	s	s^2

$\sum$ = "the sum of ..."

n = number of pieces of data (population)

$n-1$ = number of pieces of data (sample)

$\bar{x}$ = mean (average) of data

x_i = each of the values in the data

$x_1, x_2, x_3, x_4, \ldots x_n$ (as i goes from 1 to n)

Kittel 2013

Central tendency

- The statistics that describe **the center of a frequency** distribution for a quantitative variable.
- Shows a **typical** observation/case.
- Most common measures: mean, mode, and median.

Central tendency: mode

– Value that **occurs most frequently** in the sample.
– Applicable at **all levels of measurement**.
– Used mainly for highly discrete variables such as **categorical data**.

– {“catholic”, “Muslim”, “Hindu”, “catholic”, “catholic”, “Muslim”, “catholic”, “catholic”}
– {1, 2, 3, 1, 1, 2, 1, 1}
– {“agree”, “agree”, “disagree”, “agree”, “neutral”, “disagree”, “disagree”, “disagree”, “agree”}
– {1, 1, -1, 1, 0, -1, -1, -1, 1}
– Years of education.
– {13, 9, 9, 18, 13, 9, 18, 13, 9, 13, 13}

Central tendency: median

- Observation that is in **the middle of the ordered sample** (between 50th bottom and 50th upper percentile).
- Splits data into **two parts with equal # of observations**.
- For even sized samples: average value of the two middle observations.
- Applicable **at least at ordinal level**.

Central tendency: median

- Identification of median: **(n + 1) / 2 ;**
  n = # of observations in the data

- **Odd** numbered *n*: {1, 1, 2, 2, 3, 3, **5**, 6, 6, 6, 7, 10, 39}
- Median = (13 + 1)/2 = 7th position = **5**

- **Even** numbered *n*: {1, 1, 2, 2, 3, **3, 5**, 6, 6, 6, 7, 10}
- Median = (12 + 1)/2 = 6.5th position
  = (6th + 7th position)/2 = (3 + 5)/2 = **4**

Central tendency: median

Set 1	8	9	**10**	11	12
Set 2	8	9	**10**	11	100
Set 3	0	9	**10**	10	10
Set 4	8	9	**10**	100	100

Finlan & Agresti 2009: 43

Central tendency: mean

- **Arithmetic mean**

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

- **Properties:**
  - Center of gravity of a distribution.
  - Can be used **only for metric scales.**
  - Strongly influenced by outliers.

Central tendency

- Mode
- Median
- Mean
- {1, 1, 2, 2, 3, 3, 5, 6, 6, 6, 7, 10, 39}

Central tendency

- Mode
- Median
- Mean
- {1, 1, 2, 2, 3, 3, **5, 6, 6, 6, 7,** 10, 39}

Position

- The measures of central tendency are not sufficient for description of data for a quantitative variable.
- Does not describe the **spread of the data**.

- **Position measures:** describe the point at which a given percentage of the data fall below or above that point.

Position: percentile

- **Percentile.** The *pth* percentile is the point such that *p%* of the observations fall below that point and (and 100 - p)% fall above it.

  - E.g. 89th percentile = indicates a point where 89% of observations lie below and 11% lie above it.
  - **Median is a 50th percentile**.
  - "Standard" percentiles: (25, 50, 75), or (10, 25, 50, 75, 90).

Position: IQR

- **Interquartile range**
  - Difference between the values of observations at **75%** (upper quartile) and **25%** (lower quartile).
  - Shows spread of middle half of the observations.

{1, 1, 2, 2, 3, 3, 5, 6, 6, 6, 7, 10, 39}

Median = (13 + 1)/2 = 7^{th} observation = 5

Q1 = (6 + 1)/2 = 3.5^{th} observation = (2 + 2)/2 = 2

Q2 = (6 + 1)/2 = 3.5^{th} observation = (6 + 7)/2 = 6.5

IQR = Q3 – Q1

IQR = 6.5 – 2 = **4.5**

Position: quartile

- **Quartile**
  – Values of observations at 25% (Q1), 50% (Q2), and 75% (Q3) of a distribution.

{1, 1, 2, 2, 3, 3, 5, 6, 6, 6, 7, 10, 39}

Q1 (25 %) = 2

Q2 (50 %) = 5

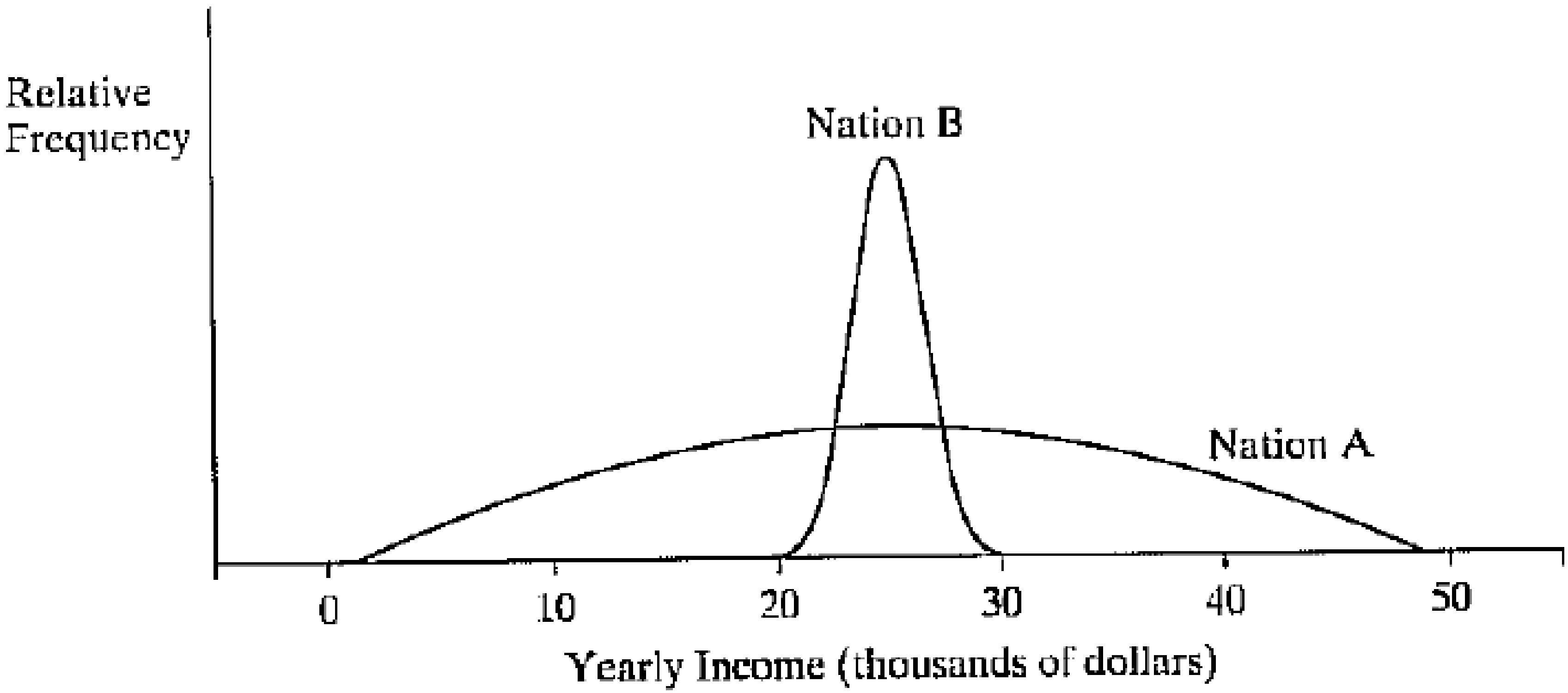
Q3 (75 %) = 6.5

Variability

- The measures of central tendency are not sufficient for description of data for a quantitative variable.
- Does not describe the **spread of the data**.

- **Variability measures:** describe the deviations of the data from a measure of center (such as mean).
  - With exception of a **range**.

Variability



Finlan & Agresti 2009: 46

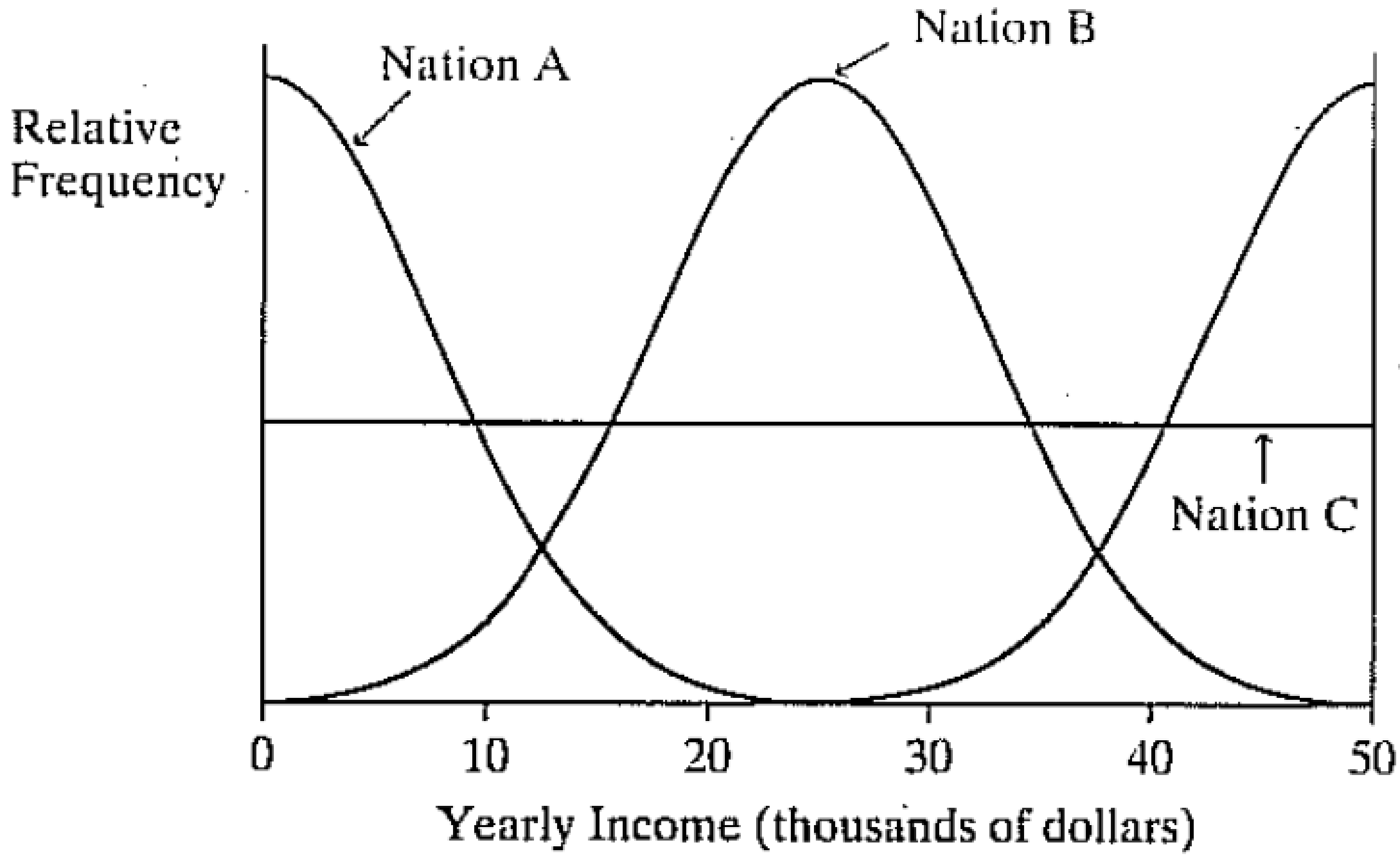
Variability: range

- **Range:** difference between largest and smallest value.
- The simplest measure of variability.
- Does not describe deviations from the mean.

{**1**, 1, 2, 2, 3, 3, 5, 6, 6, 6, 7, 10, **39**}

Range = 39 – 1 = 38

Variability



Finlan & Agresti 2009: 47

Variability: deviation

- **Deviation**
  - Difference between value of observation and mean.

$$(x_i - \mu)$$

$$(x_i - \bar{x})$$

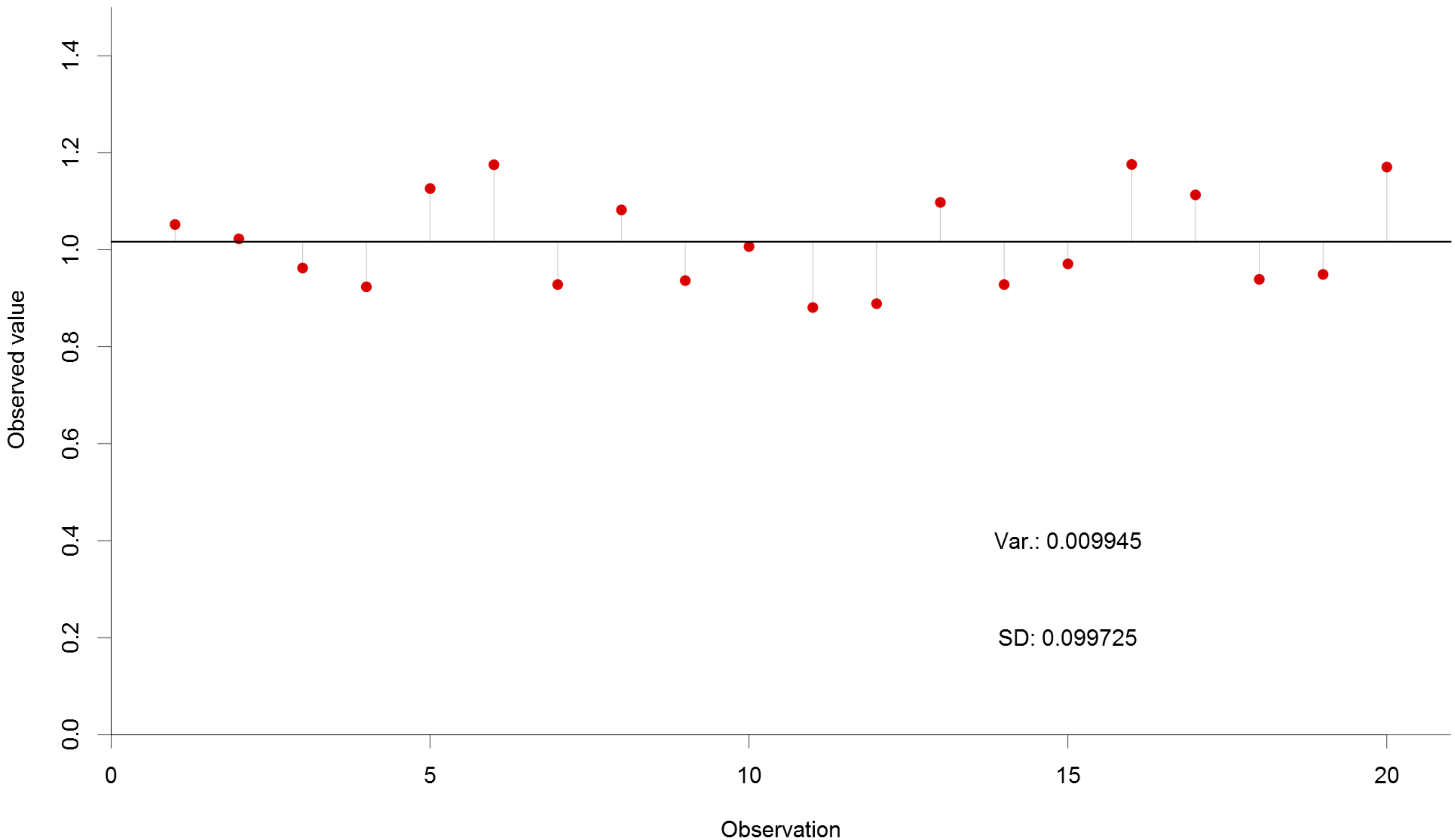
{1, 1, 2, 2, 3, 3, 5, 6, 6, 6, 7, 10, 39}

(1 - 7), (1 - 7), (2 - 7), … , (39 - 7)

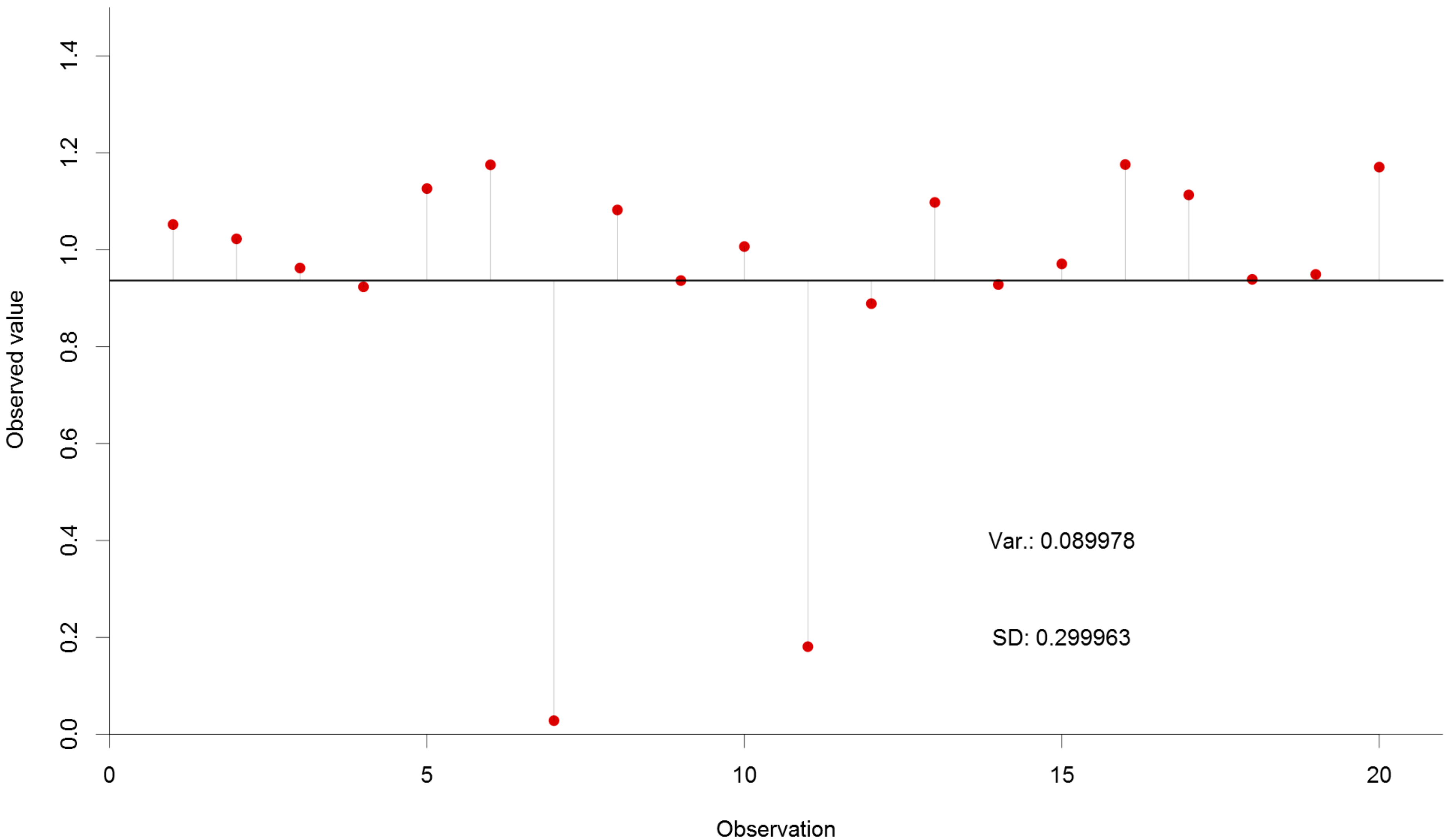
-6, -6, -5, -5, -4, -4, -2, -1, -1, -1, 0, 3, 32

Variability: deviation

- **Deviation**
  - Difference between value of observation and mean.
  - **Positive** deviation: observation value > mean
  - **Negative** deviation: observation value < mean
  - **Zero** deviation: observation value = mean.
  - Since **sum of deviations = 0**, the absolute values or the squares are used in measures that use deviations.



Lehotský 2016



Lehotský 2016

Variability: variance

- Mean is usually not very indicative for data dispersion:

{4, 4, 6, 6}; mean = 5; s^2 = 1.33

{0, 0, 10, 10}; mean = 5; s^2 = 33.33

- Therefore we need other measures such as **variance (s^2)**.

Variability: variance

- **Variance**
  - Squared **mean deviation** from mean.

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

population = {1, 3, 6, 10}

¼ * ((1 - 5)^2 + (3 - 5)^2 + (6 - 5)^2 + (10 - 5)^2)

¼ * ((-4)^2 + (-2)^2 + 1^2 + 5^2)

¼ * (16 + 4 + 1 + 25) = ¼ * 46 = **11.5**

Variability: variance

- **Variance**
  – Squared **approximate mean deviation** from mean.

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

sample = {1, 3, 6, 10}

1/3 * ((1 - 5)^2 + (3 - 5)^2 + (6 - 5)^2 + (10 - 5)^2)

1/3 * ((-4)^2 + (-2)^2 + 1^2 + 5^2)

1/3 * (16 + 4 + 1 + 25) = 1/3 * 46 = **15.33**

Variability: standard deviation

- **Standard deviation**
  - Measure of average deviation.
  - Typical distance of observation from the mean.
  - Sensitive to outliers.

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

sample = {1, 3, 6, 10}

s^2 = 15.33

s = sqrt(15.33) = 3.92

Variability: standard deviation

- **Properties**
  - *s* >= 0
  - *s* = 0 only when all observations have same value.
  - The greater variability around mean, the larger *s*.
  - If data are rescaled, the s is rescaled as well.
  - E.g. if we rescale s of annual income in $ = 34,000 to thousands of $ = 34, the *s* also changes by factor of 1000 from 11,800 to 11.8.

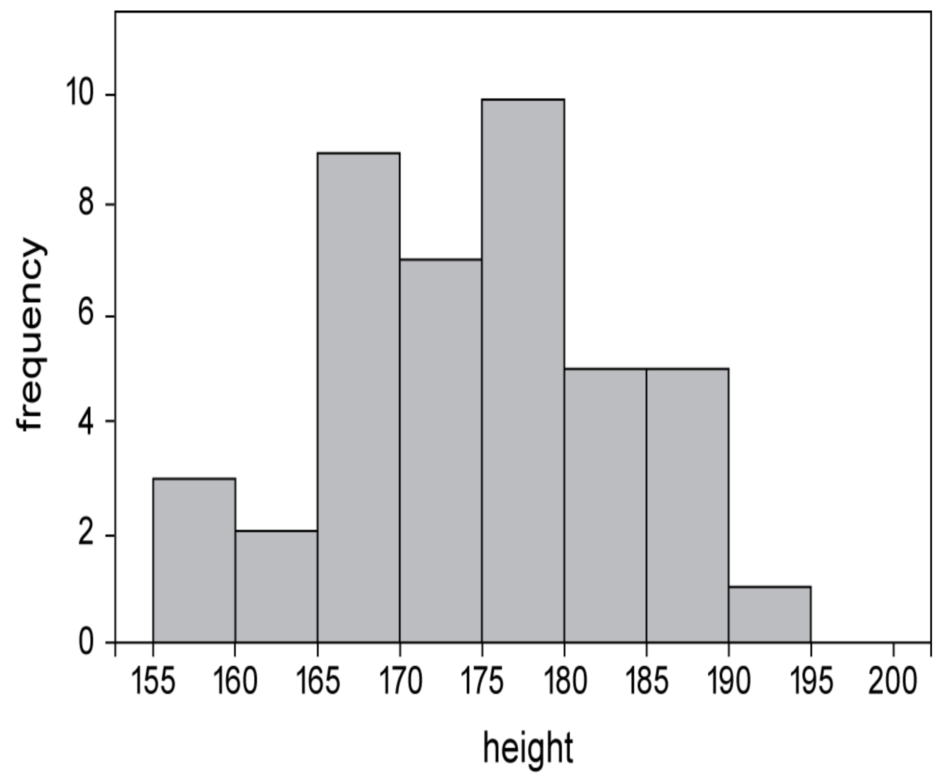
Variability: standard deviation

- **Interpretation**
  - Scale dependent.
  - E.g. assume that average amount of points received in this course is 35 points graded on a scale 0 to 40.
  - $s = 0$ extremely unlikely (no differences in performance).
  - As well as $m = 20$, $s > 15$ (huge differences in performance).

Frequency distribution

- Frequency distribution: table or visual display of the **frequency** of variable values.

155-160	3
160-165	2
165-170	9
170-175	7
175-180	10
180-185	5
185-190	5
190-195	1
195-200	0

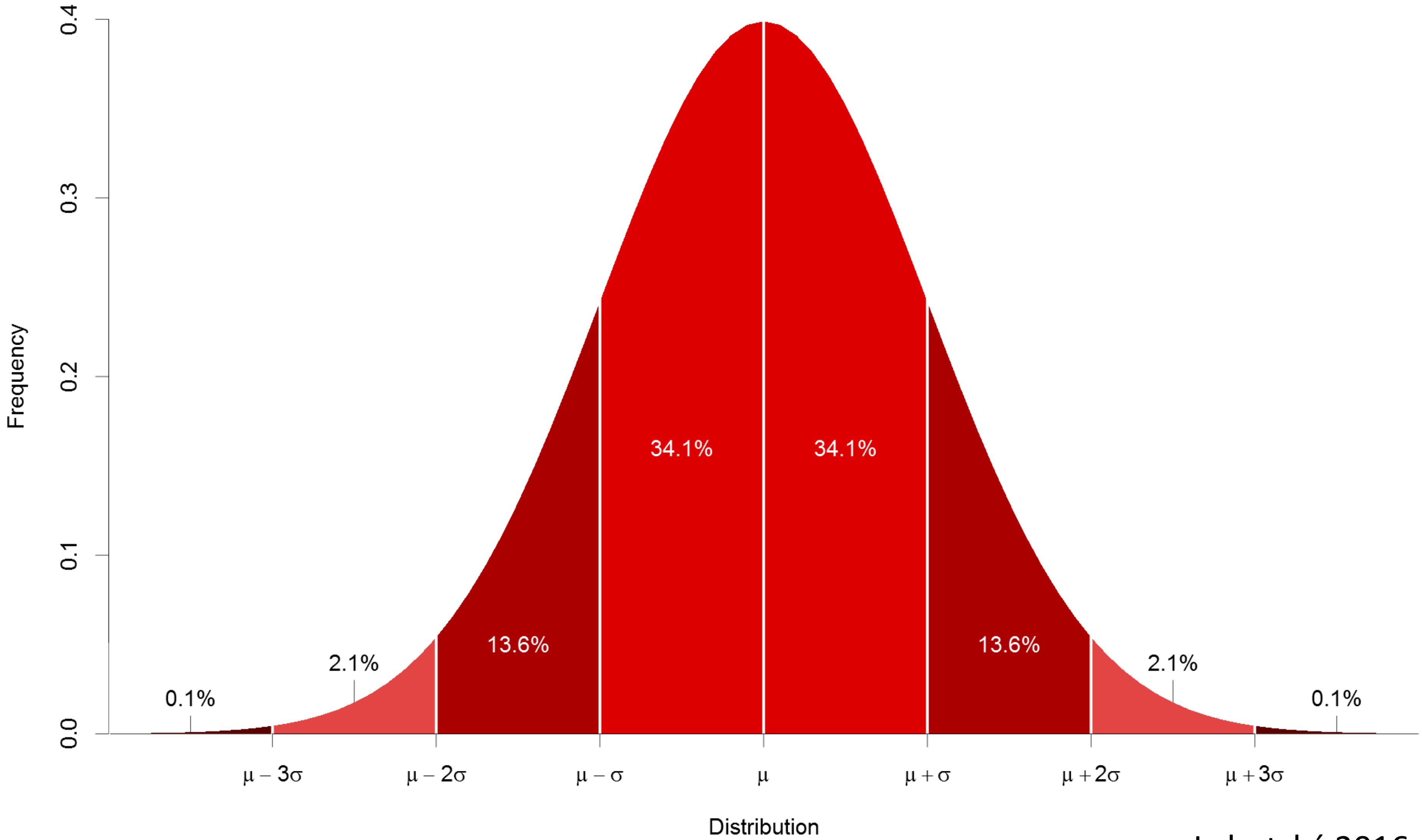


Frequency distribution

- **Absolute frequency:** # of the observations of a category.
- **Relative frequency:** proportion of the observations of a category over total # of observations.
- **Percentage:** proportion multiplied by 100.

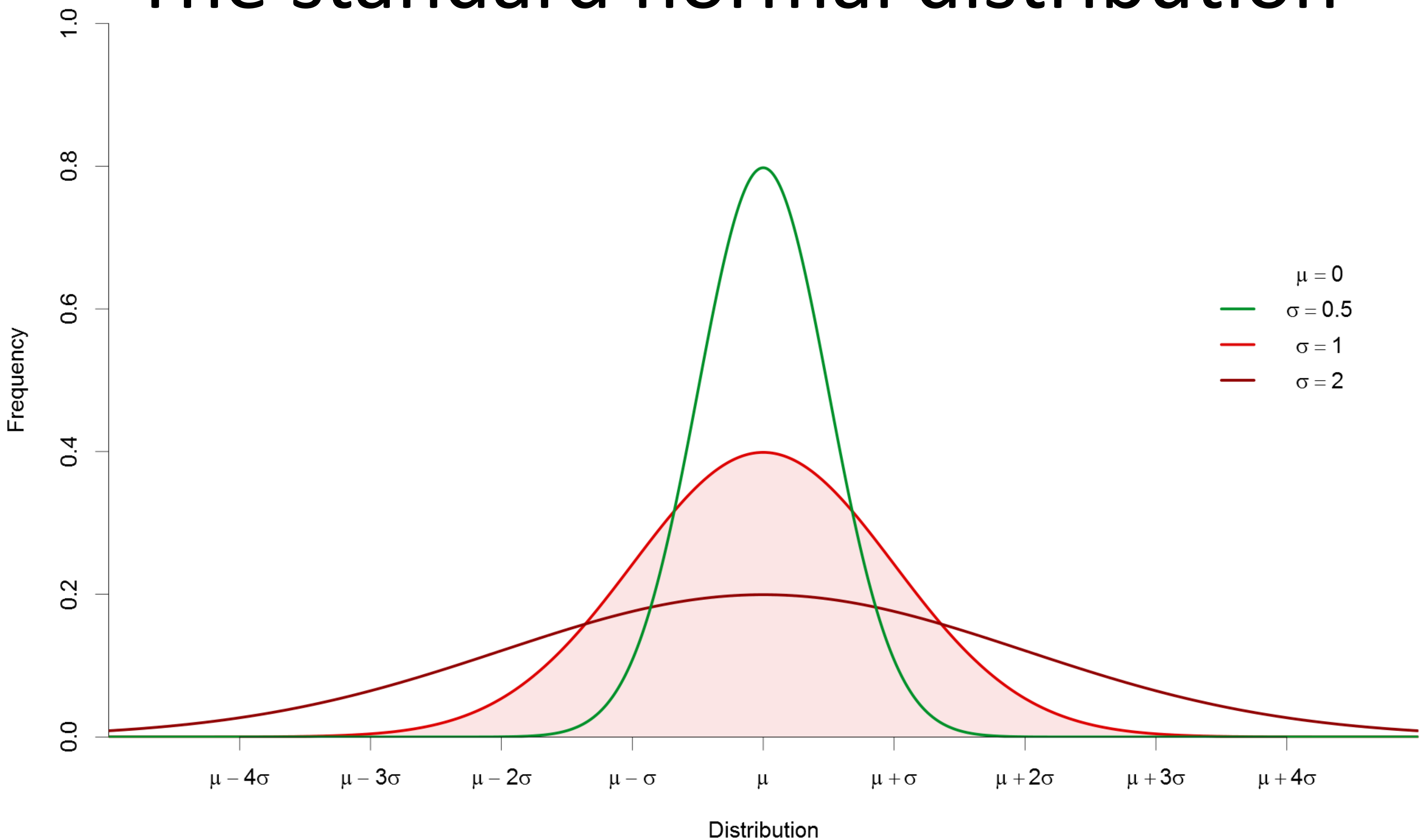
155-160	3	0.07	7%
160-165	2	0.05	5%
165-170	9	0.21	21%
170-175	7	0.17	17%
175-180	10	0.24	24%
180-185	5	0.12	12%
185-190	5	0.12	12%
190-195	1	0.02	2%
195-200	0	0	0%

The standard normal distribution



Lehotský 2016

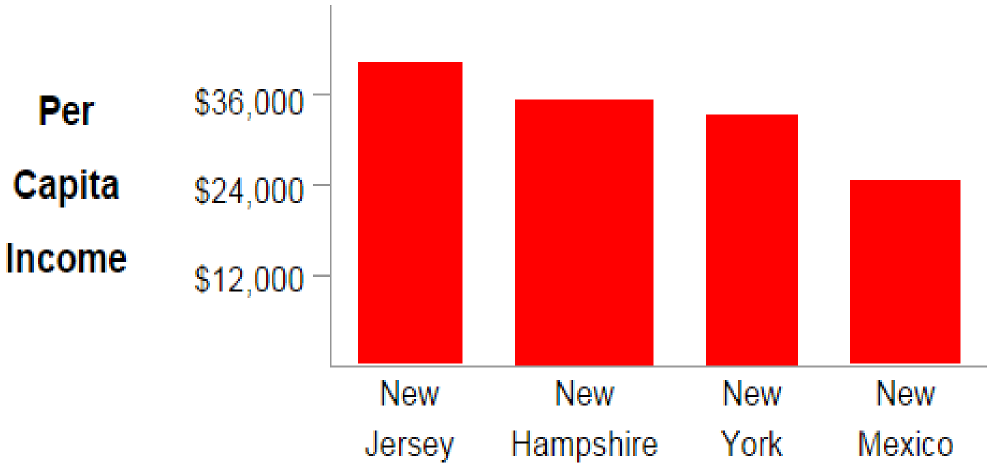
The standard normal distribution



Lehotský 2016

Bar chart

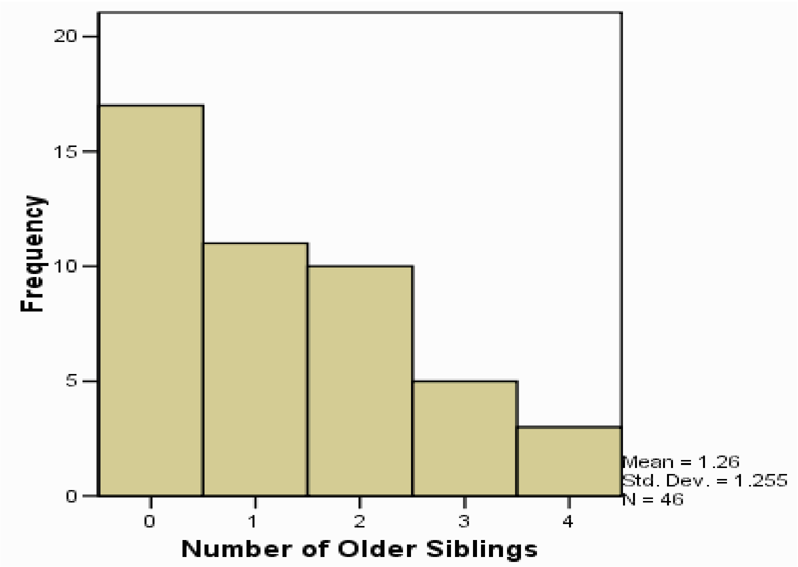
- The columns are positioned over values of **categorical variable** (U.S. states).
- The height of the column indicates the value of the variable (per capita income).

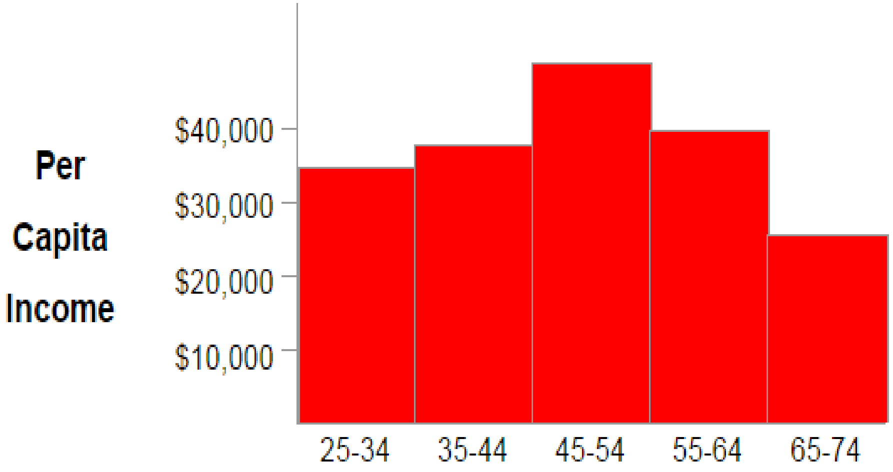


stattrek.com

Histogram

- The columns are positioned over a values of **quantitative variable.**
- The column label can be single value or range of values.
- The height of the column indicates the value of the variable.

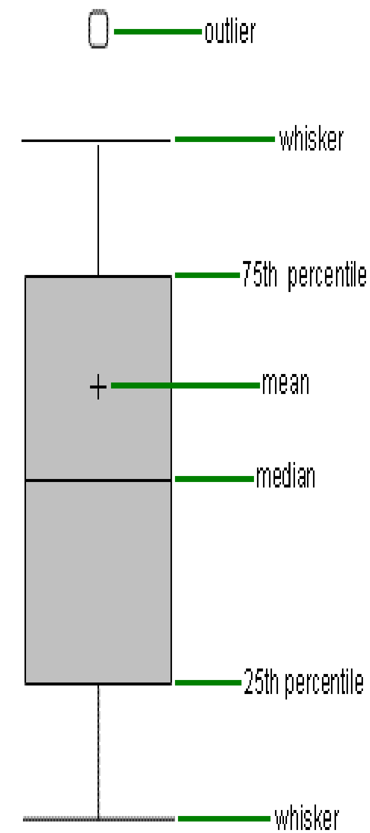




stattrek.com

Boxplot

- Splits data into quartiles (position measure).
- Box: from Q1 to Q3.
- Median (Q2): line within the box.
- Whiskers: indicate the range from:
  - Q1 to smallest non-outlier.
  - Q3 to largest non-outlier.
- Outlier > 1.5 * (Q3 – Q1) from Q1 or Q3
- Outliers are represented separately.



statmethods.com

Measures of association (MA)

- Examination of a single variable (distribution)
→ **univariate statistics**.
- Examination of associations among variables (distributions)
→ **bivariate** (and multivariate) **statistics**.

- **MA:** variety of coefficients that measure the size (and/or direction) of associations between the variables of interest.
- MA typically range within <0,1> or <-1,1> intervals.

Measures of association (MA)

level of measurement	**coefficient**
nominal	Jaccard’s index
ordinal	Kendall’s tau
metric (interval & ratio)	Pearson’s rho

Measures of association (MA)

- There are many measures of association.
- Correlation coefficients represent just one of the subsets of the MA.

- Correlation is not causation.
- Causation can be based on different types of associations.

Medzihorský 2016

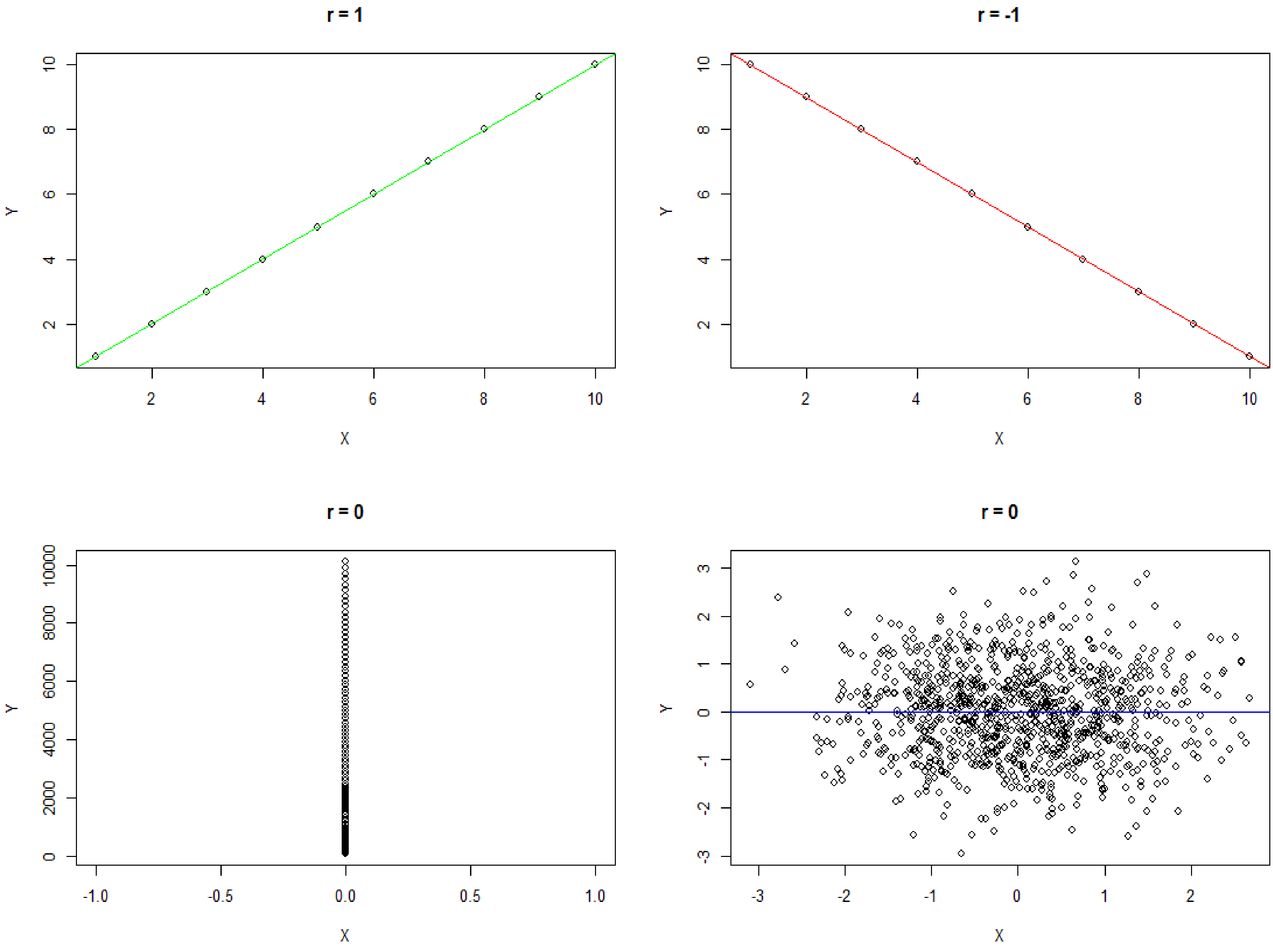
Pearson’s rho correlation coefficient

- Pearson’s product-moment correlation coefficient (**r**).
- Pearson’s r measures the **strength and direction of the linear relationship between two variables**.
- Ranges within <-1,1>
  - Perfect positive linear relationship = 1
  - Perfect negative linear relationship = -1
  - No linear relationship = 0
- Value does not depend on variables’ units.
- It is a **sample statistic.**

Pearson’s r: description

Pearson’s r strength	Description
0.00–0.19	very weak
0.20–0.39	weak
0.40–0.59	moderate
0.60–0.79	strong
0.80–1.00	very strong

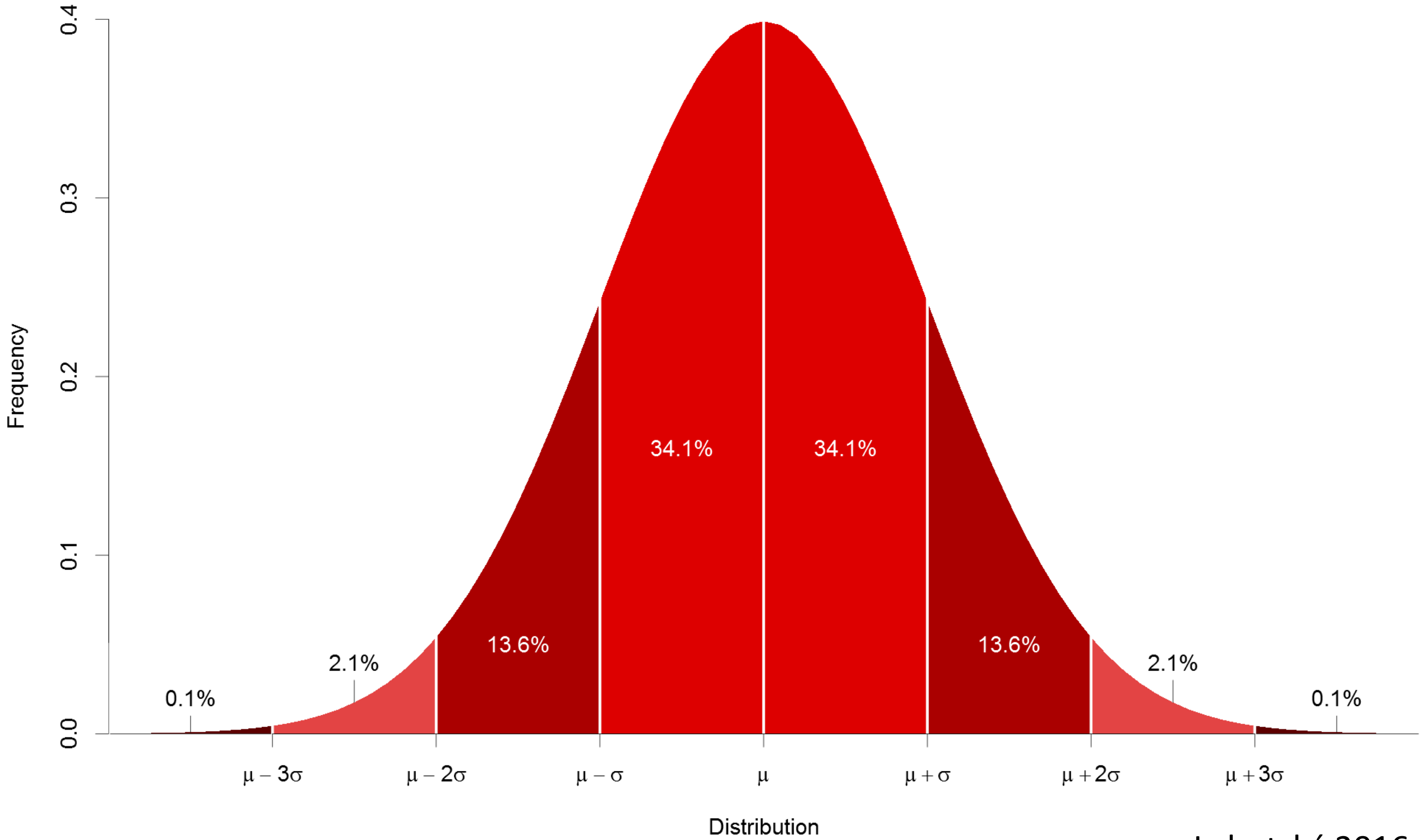
Evans 1996


r = 1
Y
X
2
4
6
8
10
r = -1
Y
X
2
4
6
8
10
r = 0
Y
X
0
2000
4000
6000
8000
10000
-1.0
-0.5
0.0
0.5
1.0
r = 0
Y
X
-3
-2
-1
0
1
2
3

Pearson's correlation

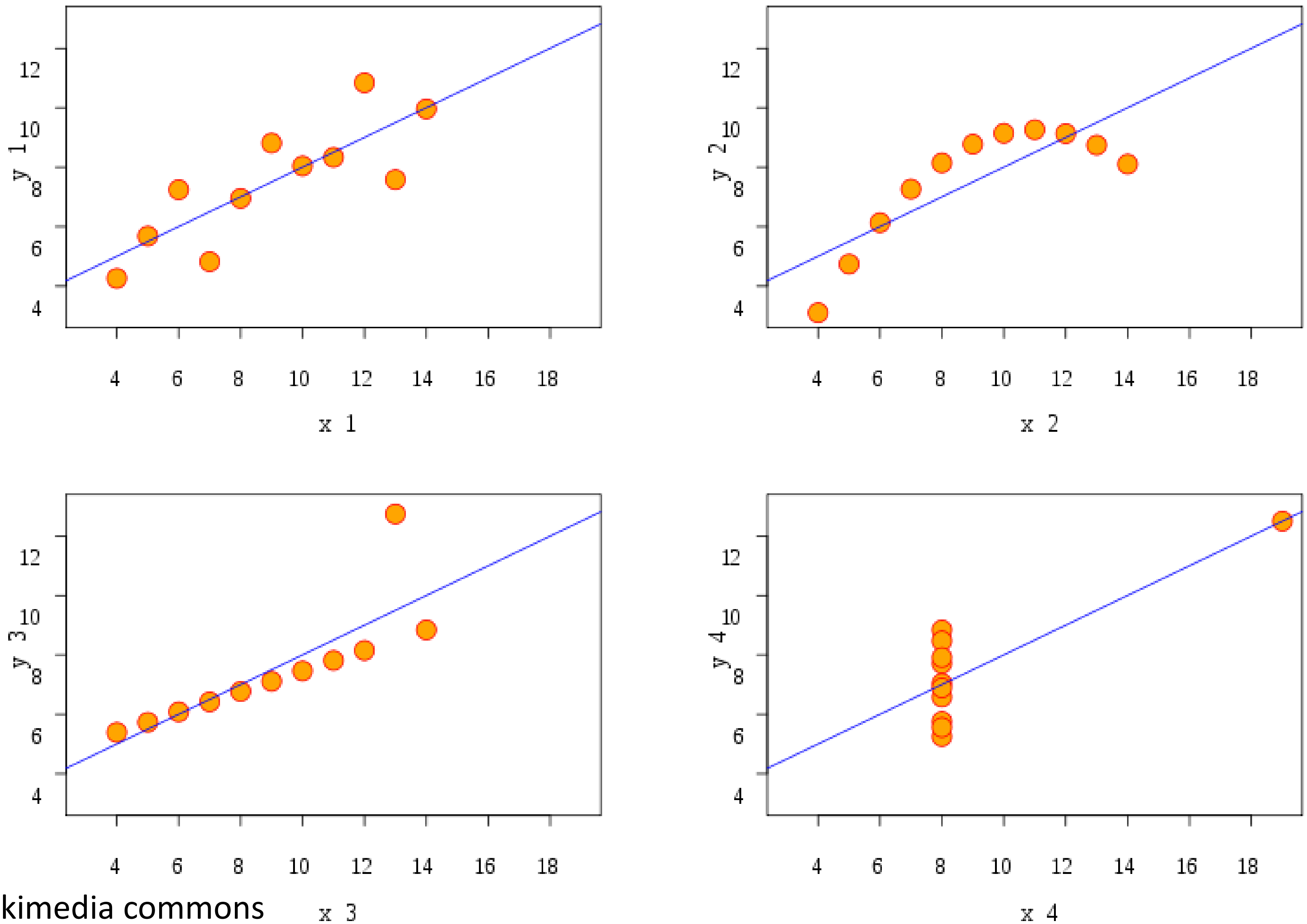
- Assumptions and limitations:
  - Metric (at least interval) level of measurement
  - Normal distribution of X and Y
  - Linear relationship between X and Y
  - Homoscedasticity
  - Sensitive to outliers

The standard normal distribution

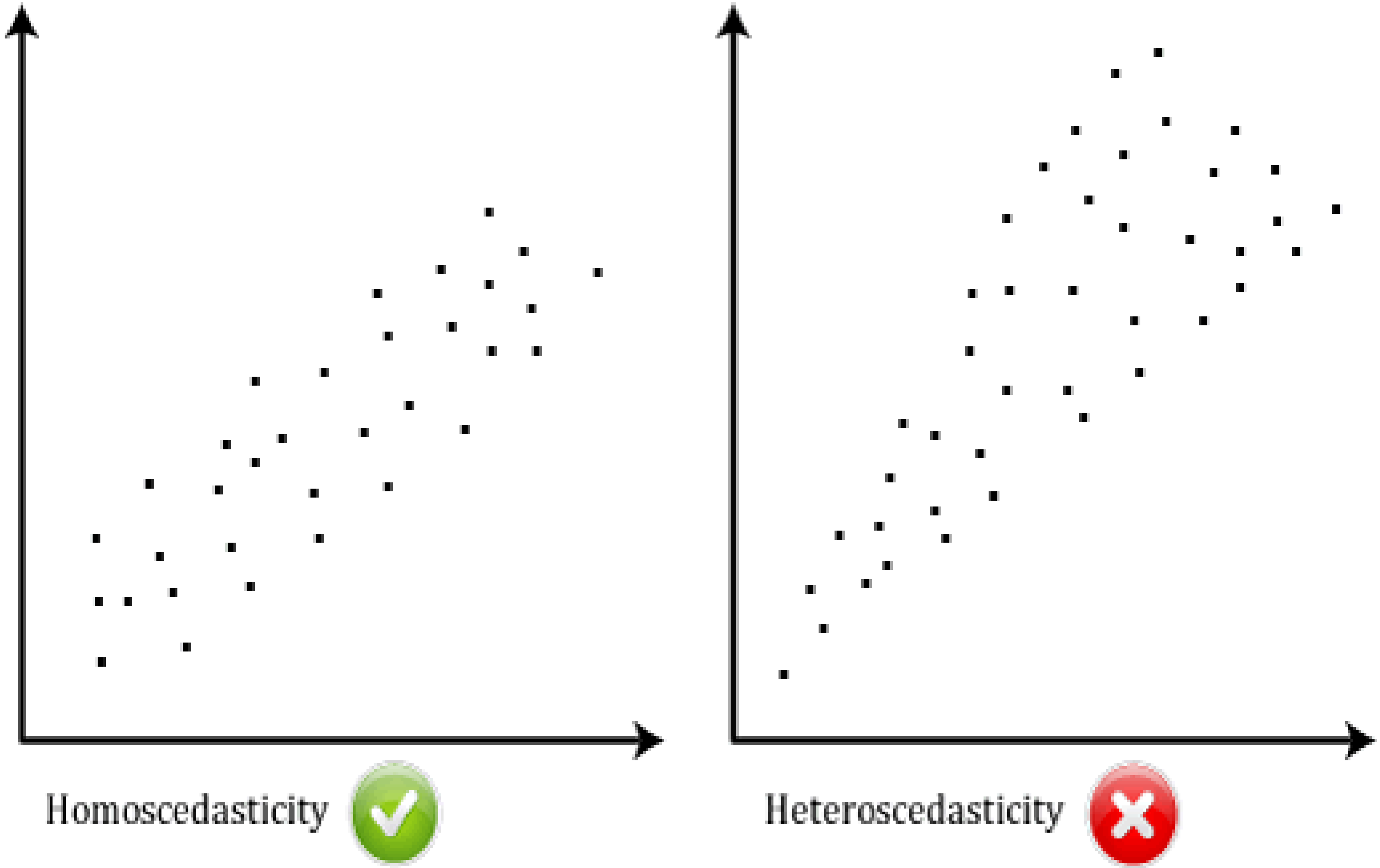


Lehotský 2016

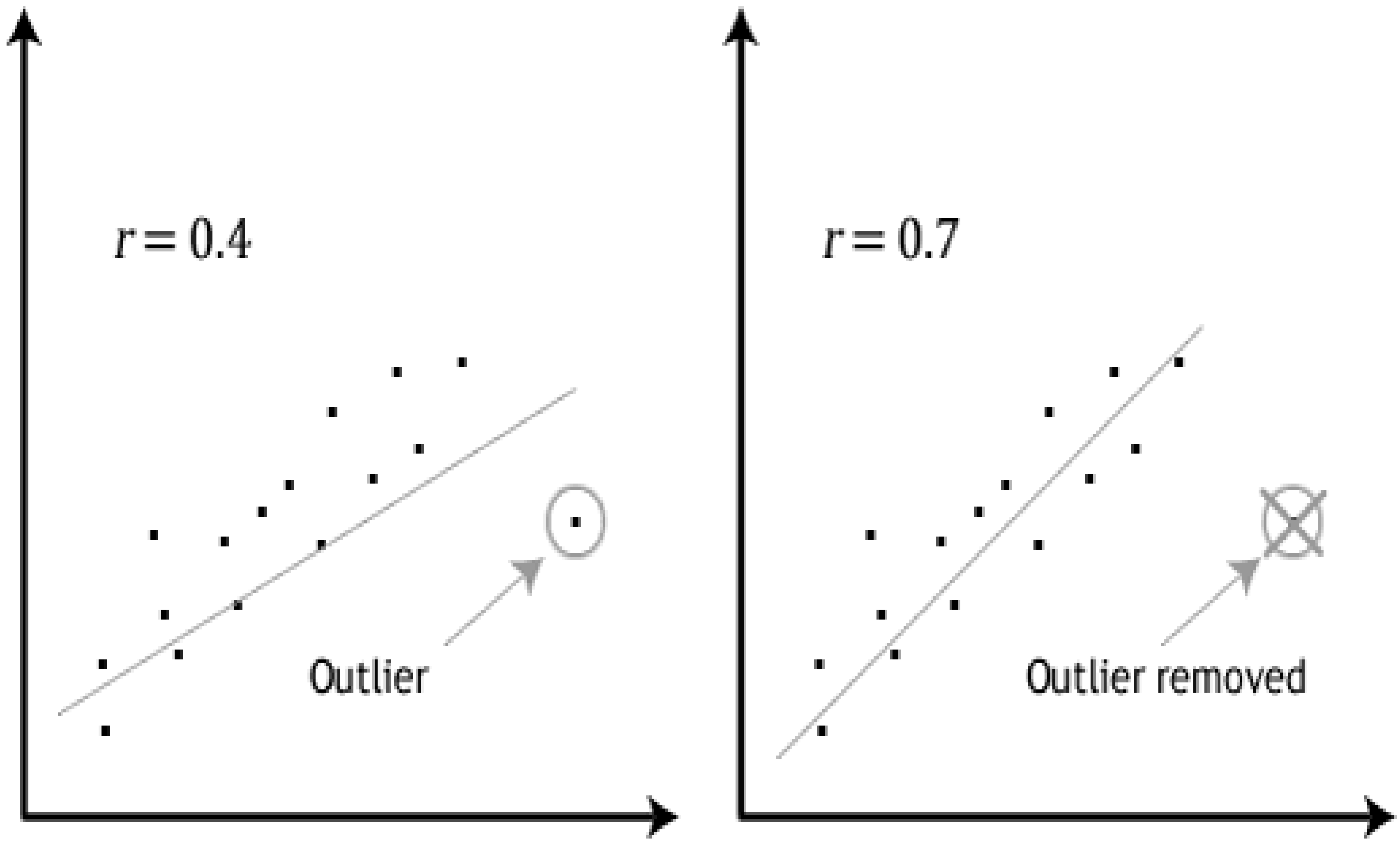
Anscombe's quartet



wikimedia commons



stats.stackexchange.com



statistics.leard.com

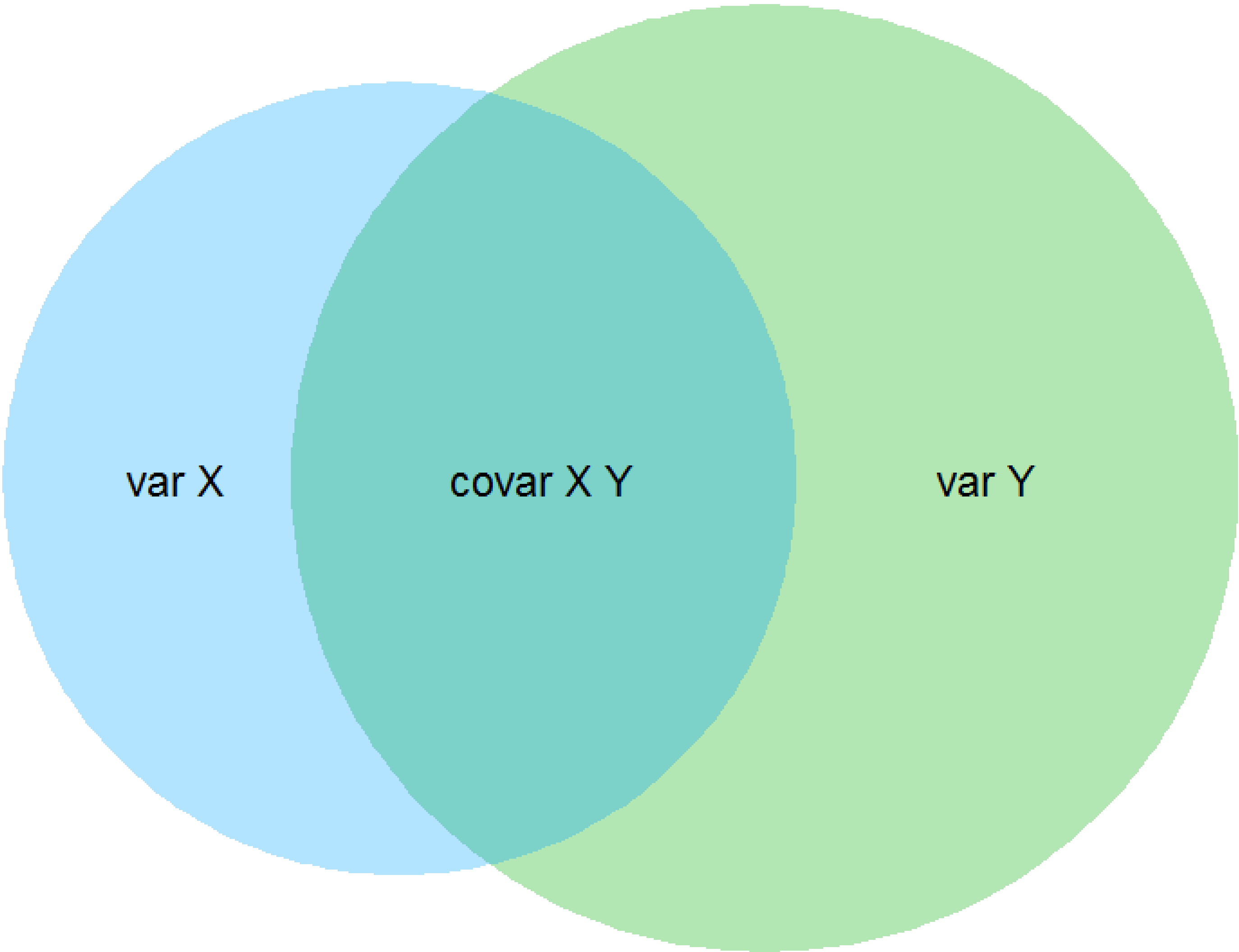
Pearson's correlation: example

- Assume we have 2 variables: X and Y.

X	Y
1	0
2	1
1	4
6	8
7	4

- What is correlation (r) of these two variables?

- r = covariance / combined total variance.



- First: we calculate **variance of variables**.
- *mean(x)* = 3.4; *mean(y)* = 3.4

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

X	(x – m)	dev.	dev.^2
1	(1 – 3.4)	-2.4	5.76
2	(2 – 3.4)	-1.4	1.96
1	(1 – 3.4)	-2.4	5.76
6	(6 – 3.4)	2.6	6.76
7	(7 – 3.4)	3.6	12.96
sum	0	0	33.2

Y	(y – m)	dev.	dev.^2
0	(0 – 3.4)	-3.4	11.56
1	(1 – 3.4)	-2.4	5.76
4	(4 – 3.4)	0.6	0.36
8	(8 – 3.4)	4.6	21.16
4	(4 – 3.4)	0.6	0.36
sum	0	0	39.2

- **s^2(X)** = 33.2 / 4 = **8.3**; **s^2(Y)** = 39.2 / 4 = **9.8**

- Second: we calculate **covariance of variables**.
- Covariance is a sum of deviation products of two variables divided by n–1.

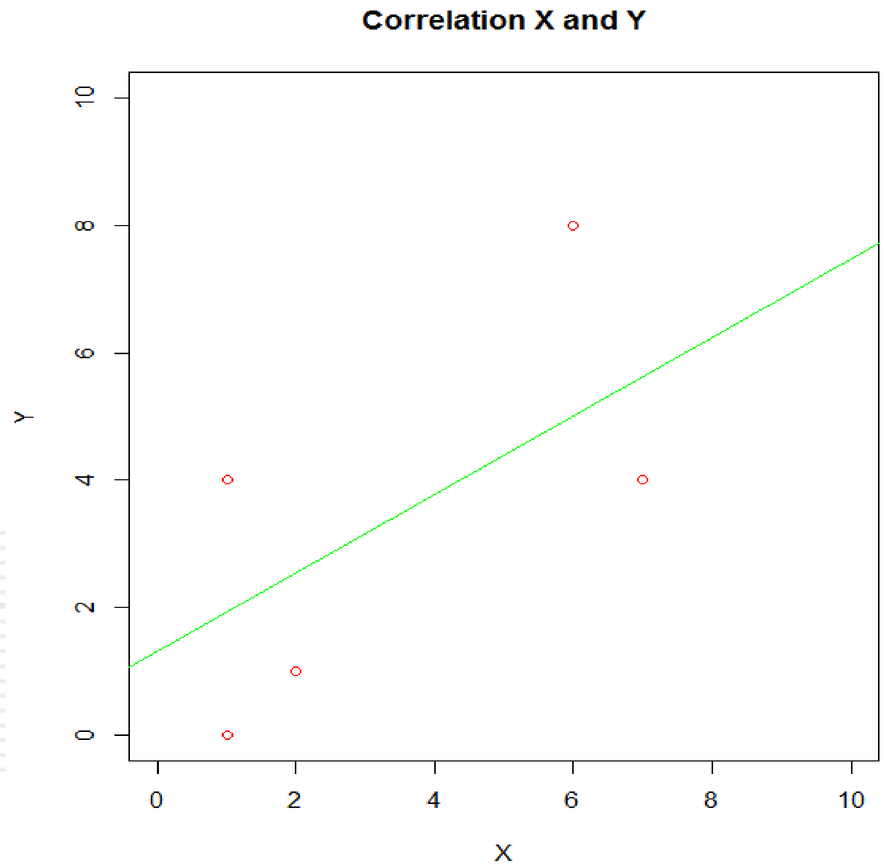
$$COV(x,y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

(x – m)	**(y – m)**	**cross-prod.**
(1 – 3.4)	(0 – 3.4)	8.16
(2 – 3.4)	(1 – 3.4)	3.36
(1 – 3.4)	(4 – 3.4)	-1.44
(6 – 3.4)	(8 – 3.4)	11.96
(7 – 3.4)	(4 – 3.4)	2.16
0	0	**24.2**

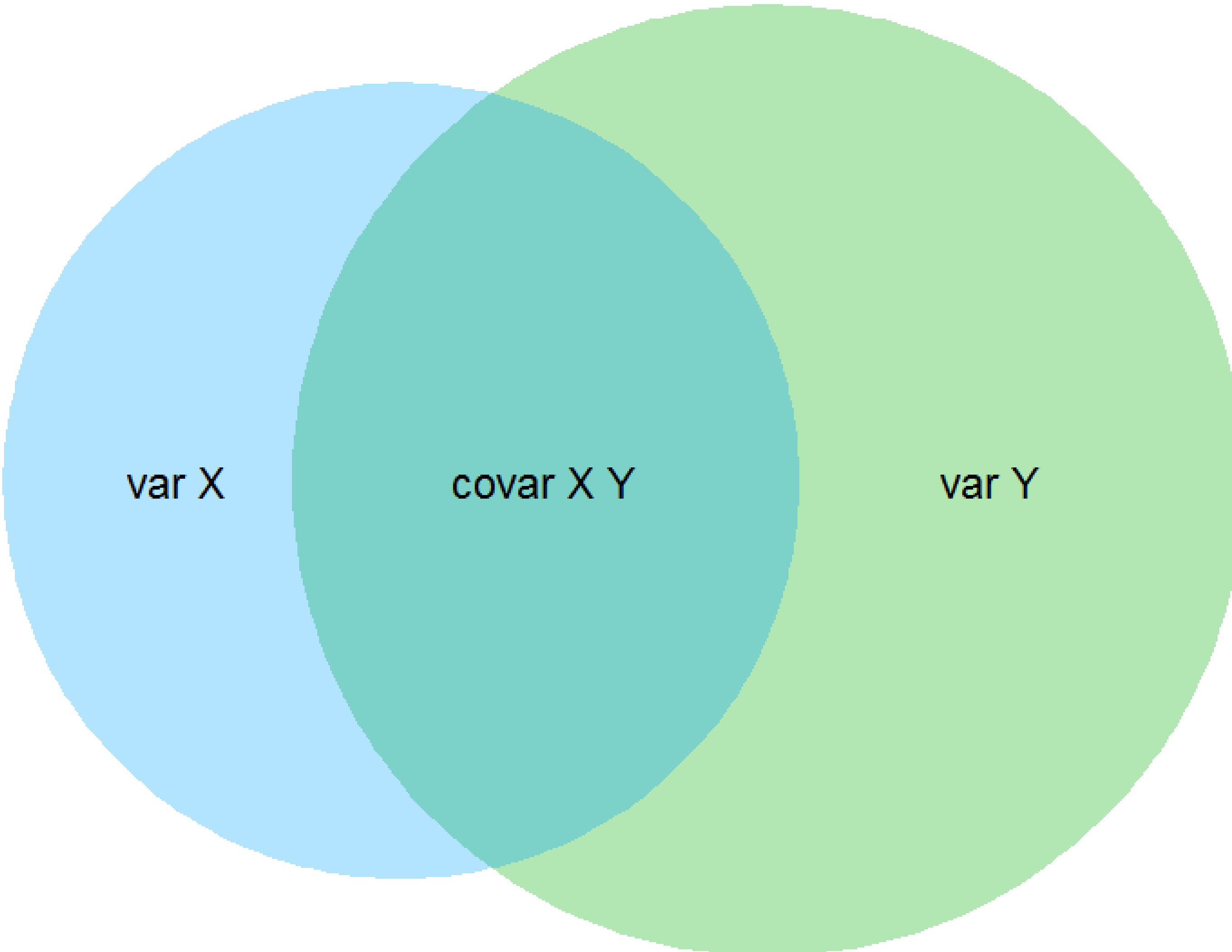
cov(X, Y) = 24.2 / 4 = **6.05**

- Third: we divide X, Y covariance by square rooted product of X and Y variances.
  - **r = cov(X, Y) / sqrt(var(X) * var(Y))**
  - **r** = 6.05 / sqrt(8.3 * 9.8) = **0.67**

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \times \sum (y_i - \bar{y})^2}}$$



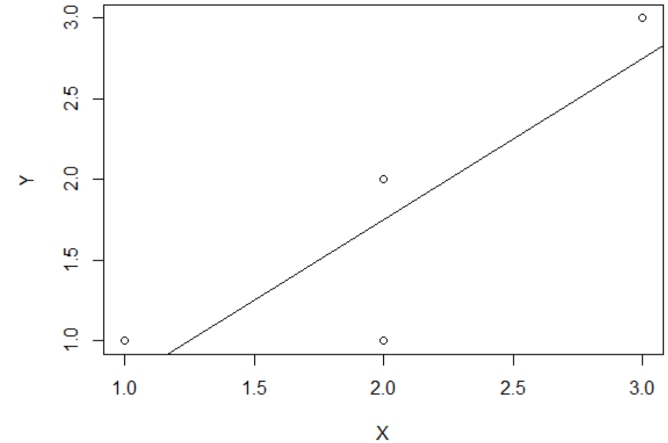
- r = covariance / combined total variance.



Kendall's tau correlation coefficient

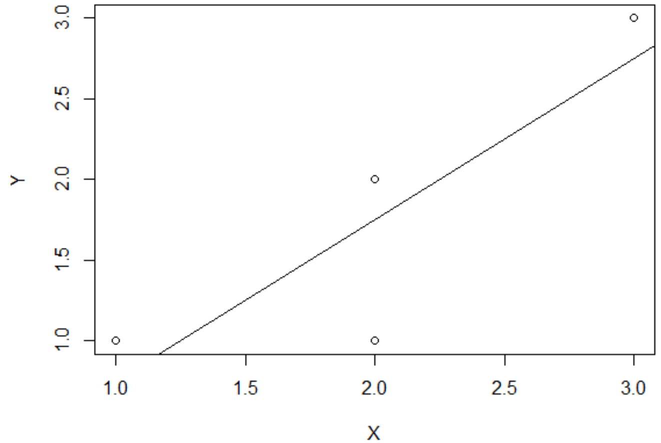
- **Kendall's tau** (τ) used for ordinal data (e.g. attitude scales).
- **A non-parametric** measure of association between two ordinal variables.
- Accommodates also small samples and many values with the same order/ranking.
- **Ranges within <-1,1>**
  - Perfect agreement (variables are identically ordered) = 1
  - Perfect inversion (variables are ordered in exactly reversed way) = -1
  - No ordered relationship = 0
- KT represents the degree of concordance between two ordinal variables.
  - $\boldsymbol{\tau_a}$ does not correct for tied values
  - $\boldsymbol{\tau_b}$ corrects for tied values
- **E.g.:** is there an ordered association between the income level and attitudes towards climate change?

cases (N)	X: income	Y: attitude
A	1 (low)	1 (disagree)
B	2 (middle)	1 (disagree)
C	2 (middle)	2 (neutral)
D	3 (high)	3 (agree)



- We have **n*(n – 1)/2 pair combinations**; i.e. 4*(4-1)/2 = **6.**
- Specifically: (A,B), (A,C), (A,D), (B,C), (B,D), (C,D).
- **Concordance:** $X_i > X_j$ AND $Y_i > Y_j$; or: $X_i < X_j$ AND $Y_i < Y_j$
- **Discordance:** $X_i > X_j$ AND $Y_i < Y_j$; or: $X_i < X_j$ AND $Y_i > Y_j$
- **Neither (tied values):** $X_i = X_j$ OR $Y_i = Y_j$
  - Pair (A,B) = neither (tied); $\mathbf{Y_A = Y_B}$
  - Pair (A,C) = concordant; $X_A < X_C$ & $Y_A < Y_C$
  - Pair (A,D) = concordant; $X_A < X_D$ & $Y_A < Y_D$
  - Pair (B,C) = neither (tied); $\mathbf{X_B = X_C}$
  - Pair (B,D) = concordant; $X_B < X_D$ & $Y_B < Y_D$
  - Pair (C,D) = concordant; $X_C < X_D$ & $Y_C < Y_D$

cases (N)	X: income	Y: attitude
A	1 (low)	1 (disagree)
B	2 (middle)	1 (disagree)
C	2 (middle)	2 (neutral)
D	3 (high)	3 (agree)



- We have **n*(n – 1)/2 pair combinations**; i.e. 4*(4-1)/2 = **6.**
  - Pair (A,B) = neither (tied)
  - Pair (A,C) = concordant
  - Pair (A,D) = concordant
  - Pair (B,C) = neither (tied)
  - Pair (B,D) = concordant
  - Pair (C,D) = concordant

$\boldsymbol{\tau_a}$ = (# of concordant pairs – # of discordant pairs) / # of all pairs

$\boldsymbol{\tau_a} = \mathbf{n_c - n_d / (n * (n - 1))}$

$\boldsymbol{\tau_a}$ = 4 – 0 / (4 * (4 - 1)) = 4 / 6 = **0.66**

- We have **n*(n – 1)/2 pair combinations**; i.e. 4*(4-1)/2 = **6.**
  - Pair (A,B) = neither (tied)
  - Pair (A,C) = concordant
  - Pair (A,D) = concordant
  - Pair (B,C) = neither (tied)
  - Pair (B,D) = concordant
  - Pair (C,D) = concordant

$\boldsymbol{\tau_b}$ = (# of concordant pairs – # of discordant pairs) / # of all pairs

$\boldsymbol{\tau_b = (n_c - n_d) / \mathrm{sqrt}((N - n_1) * (N - n_2))}$

N = (n * (n – 1))/2; total # of pairs

$n_1 = t_1 * (t_1 - 1))/2$; t_1 = # of tied values in the first set/variable

$n_2 = t_2 * (t_2 - 1))/2$; t_2 = # of tied values in the second set/variable

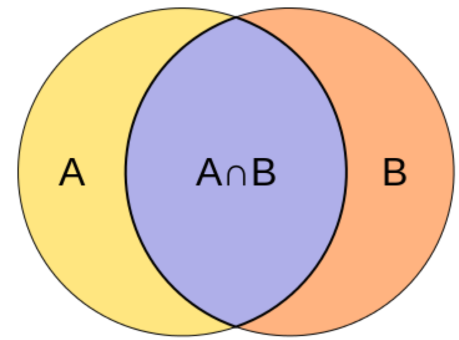
$n_1 = 2 * (2 - 1)/2 = 1$ (income var: middle/middle)

$n_2 = 2 * (2 - 1)/2 = 1$ (attitude var: disagree/disagree)

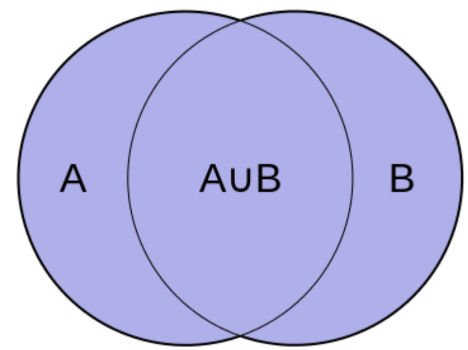
$\boldsymbol{\tau_b}$ = (4 – 0) / sqrt((6 - 1)*(6 - 1)) = 4 / sqrt(25) = 4 / 5 = **0.8**

Jaccard (similarity) index

- J used for **categorical binary data** (e.g. gender).
- Measures similarity between two samples.



		sample B	
		present	absent
sample A	present	a ($A \cap B$)	b
	absent	c	d



- J = the **size of the intersection** ($a = A \cap B$)
  by the **size of the union** ($a + b + c = A \cup B$) of the samples.
- $J = a / (a + b + c)$
- Does not account for observations missing in both samples (d).

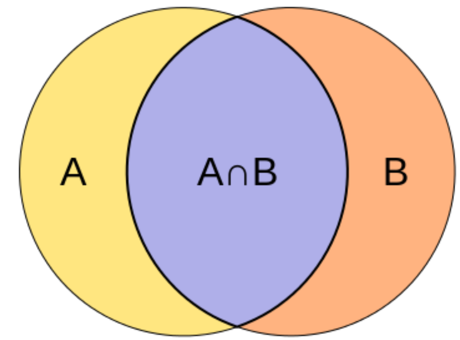
wikimedia commons

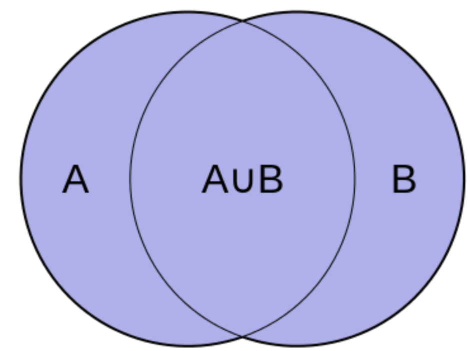
Jaccard (similarity) index: example

- Similarity of the CR and Germany based on presence/absence of int. environ. NGOs.

IENGOs		Czech Republic	
		present	absent
Germany	present	21 (a)	56 (b)
	absent	13 (c)	101 (d)

- J = a / (a + b + c)
- J = 21 / (21 + 56 + 13) = 21 / 90 = **0.23** = 23%





wikimedia commons