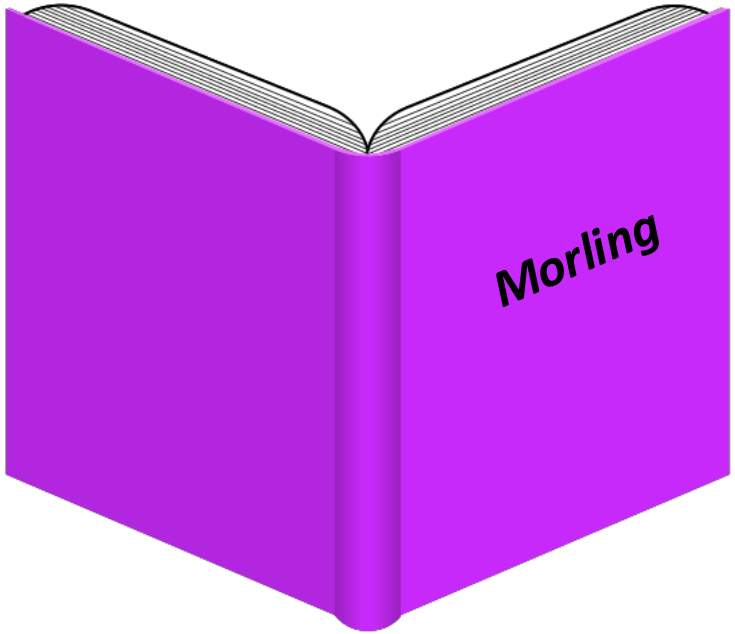
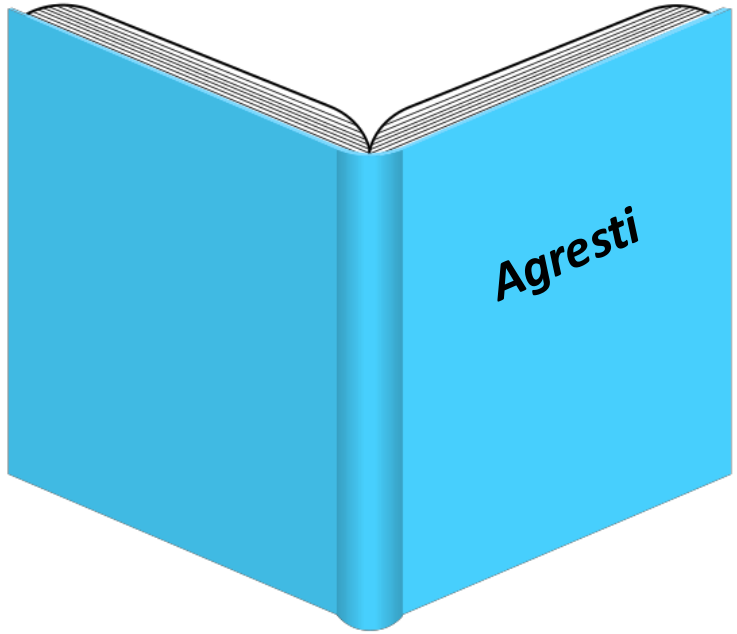




Research Methods and Statistics

Lecture 4: Variability in data + Association

Johnny van Doorn

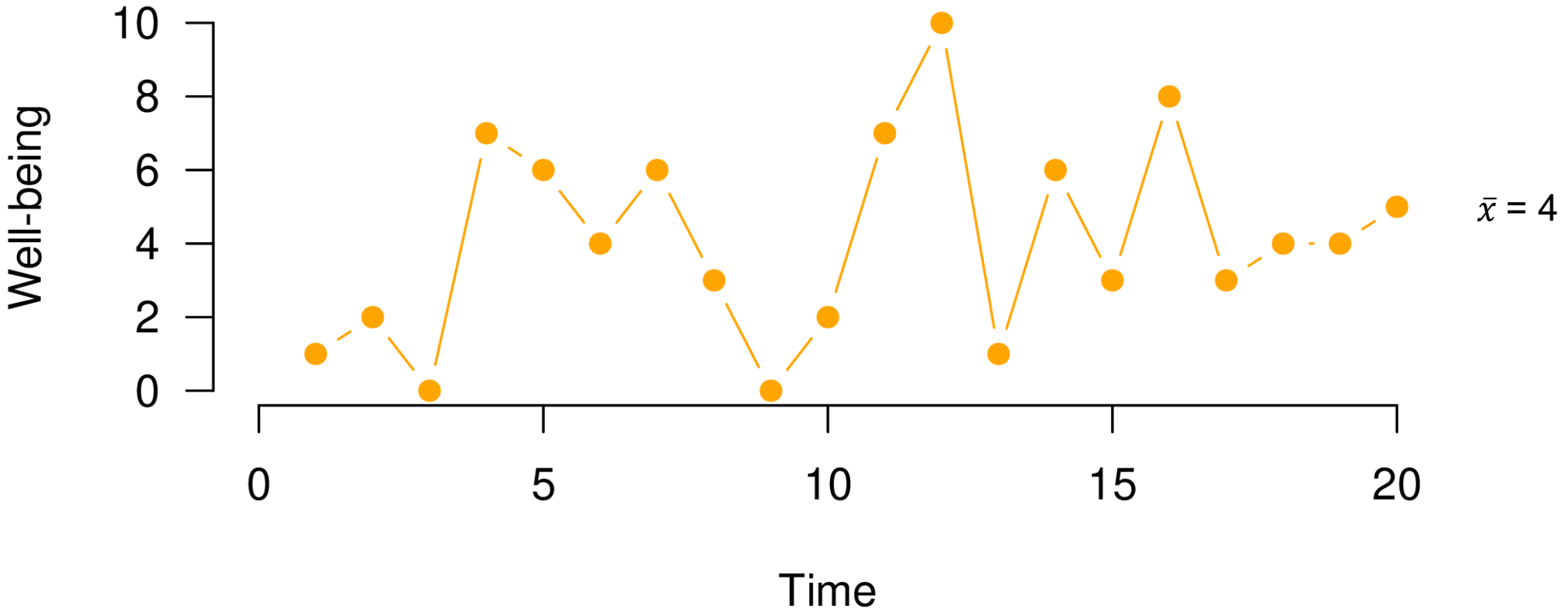
Pictures source: pixabay.org

Measuring Subjective Well-Being

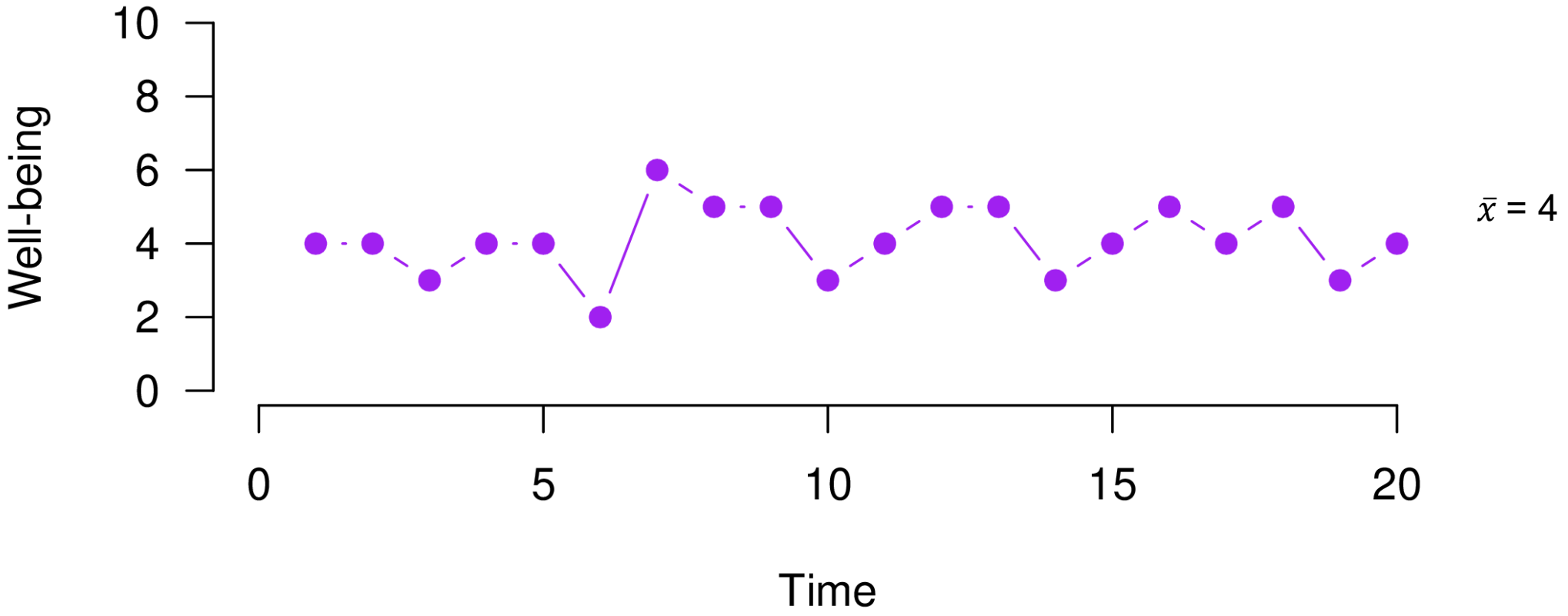
- Several measurements over time, on a discrete scale of 1-10
- 2 participants:
  - Person A: $\bar{x}$ = 4
  - Person B: $\bar{x}$ = 4

Source: https://www.ssplprints.com/image/97456/palmpilot-palmtop-computer-c-1998

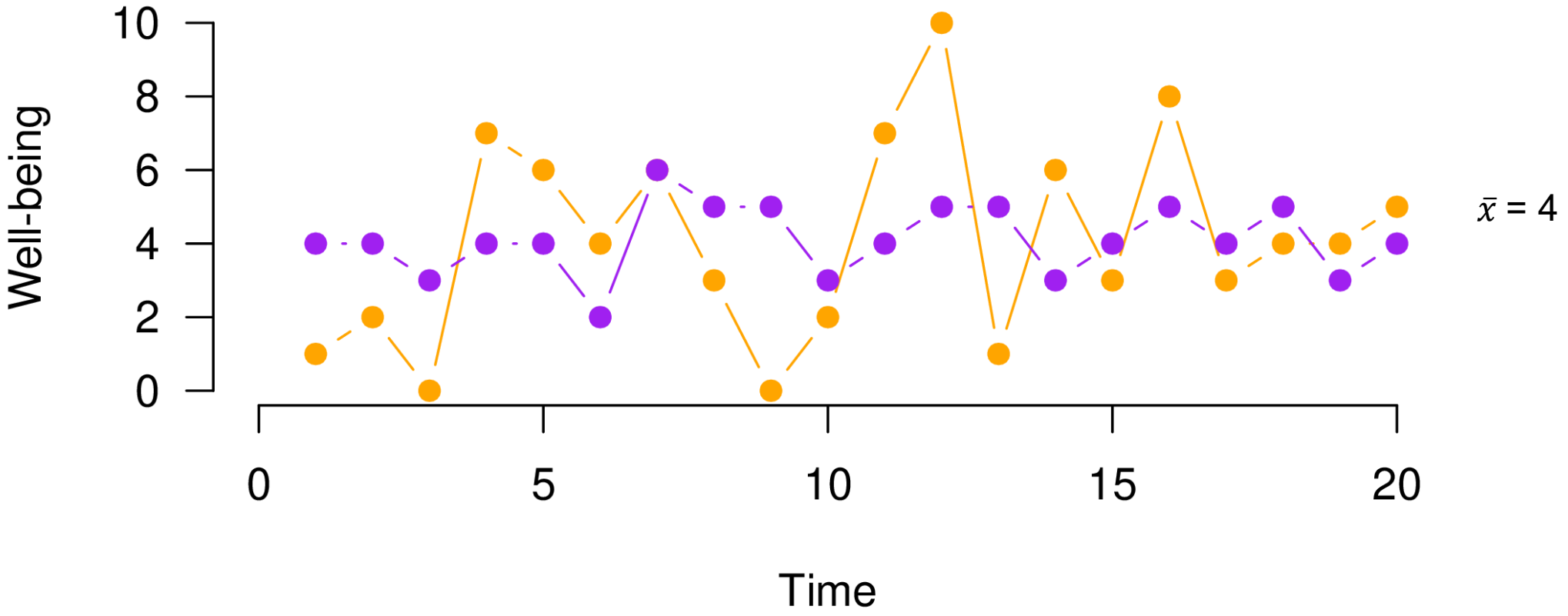
Person A



Person B



Person A & B



Today

1. **Variability in the data**
   - Standard deviation and variance
   - z-score
   - Quartiles
   - Boxplot
2. Associations
   - Between two categorical variables
   - Between two quantitative variables
3. Recap
   - Next time
   - Example exam question

How can we express/quantify variability of a variable?

How can we express/quantify an association between two variables?

Variability

- There are many ways to compute the *spread* of a distribution
  - Range
  - Absolute deviation
  - Variance + standard deviation
  - Inter Quartile Range
- Goal: Observe ***variability***
  - Complements the location measures
  - To detect outliers

Standard deviation and variance

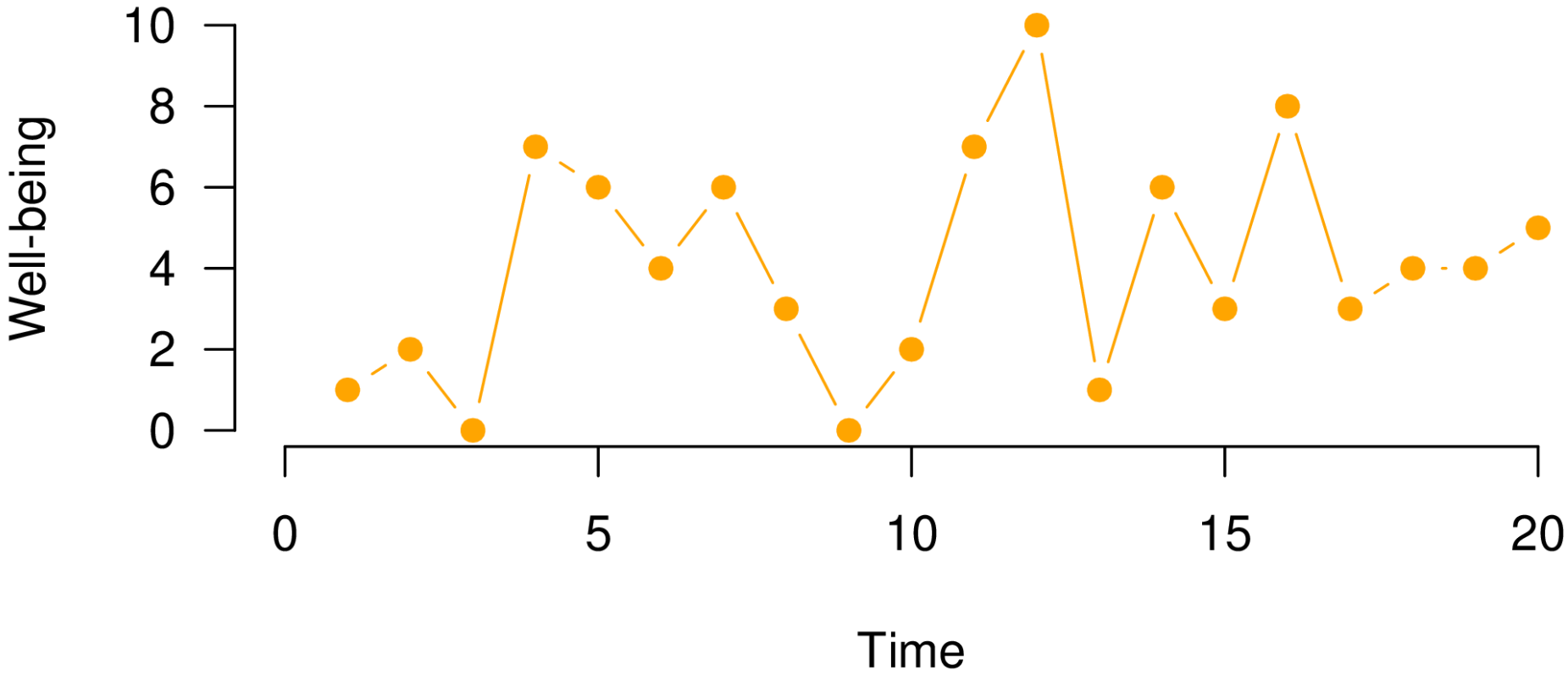
Variance: Average of the squared deviations

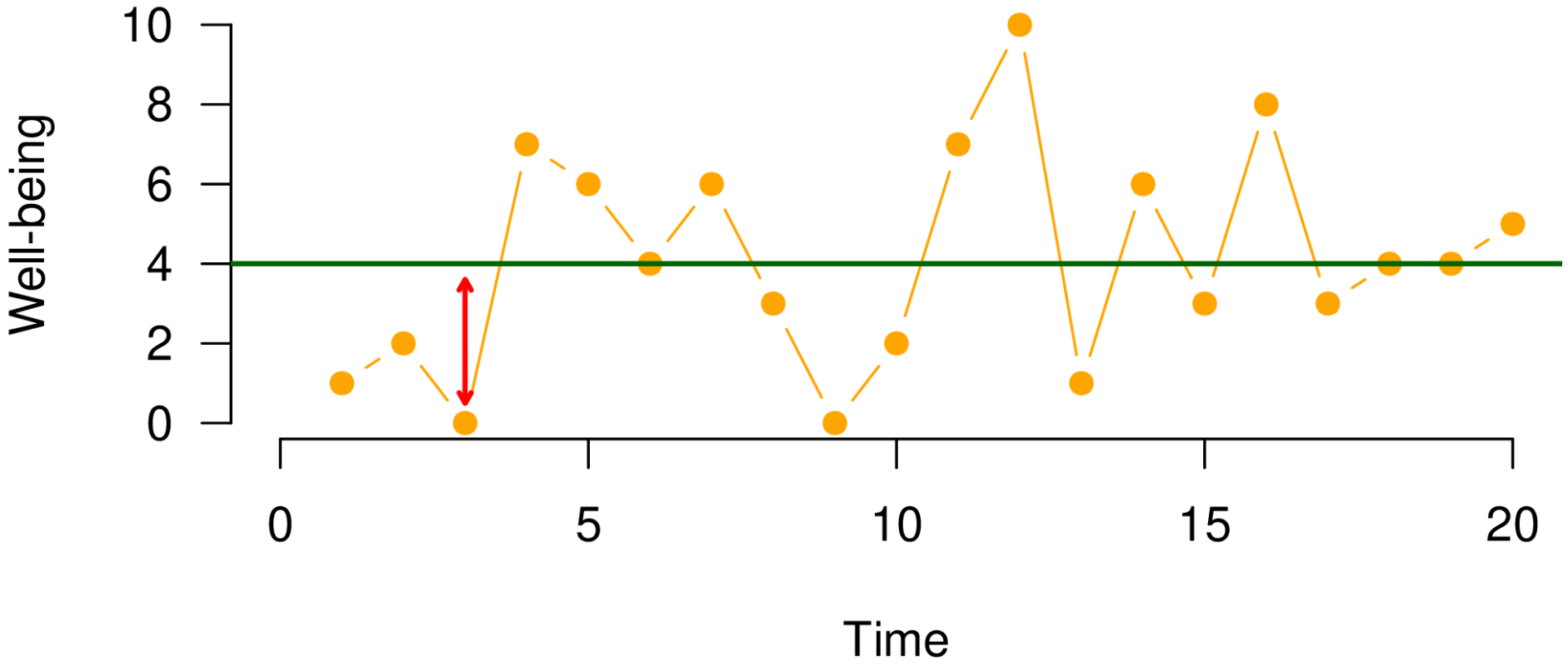
$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

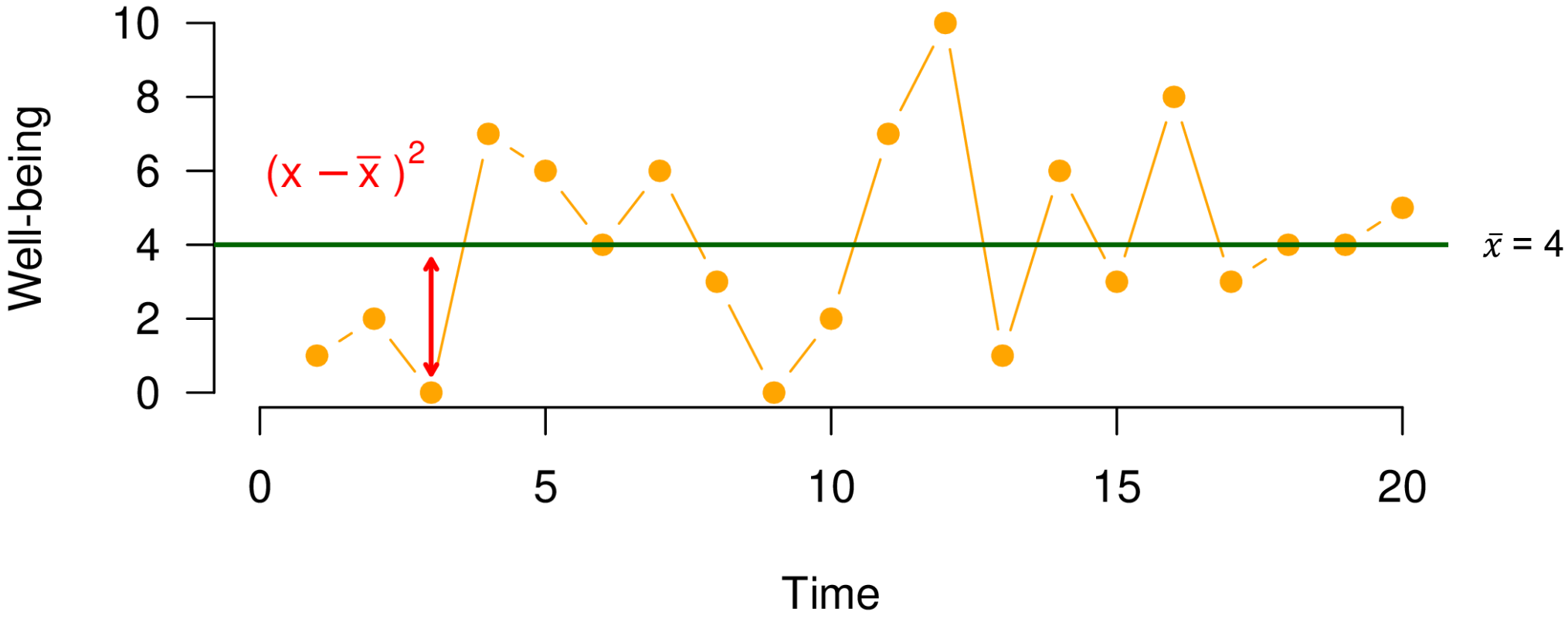
Standard deviation (sd): Square root of the variance

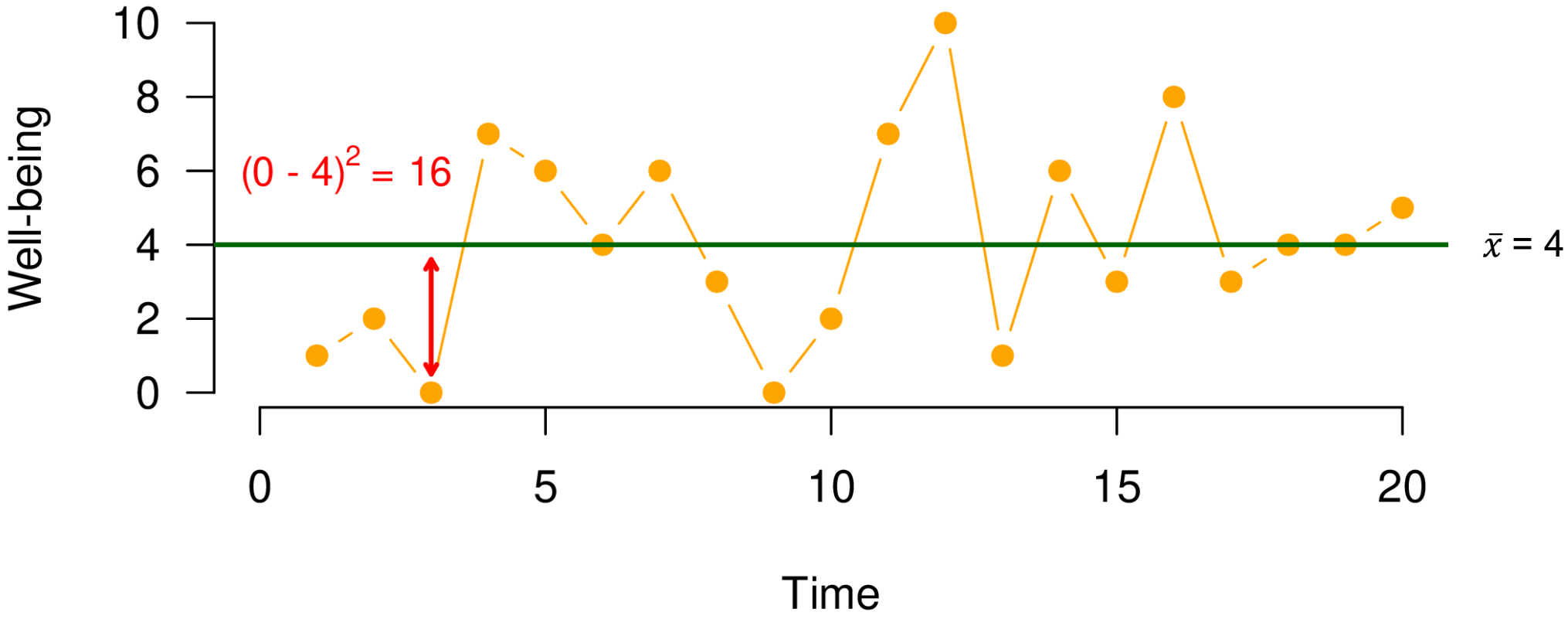
$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

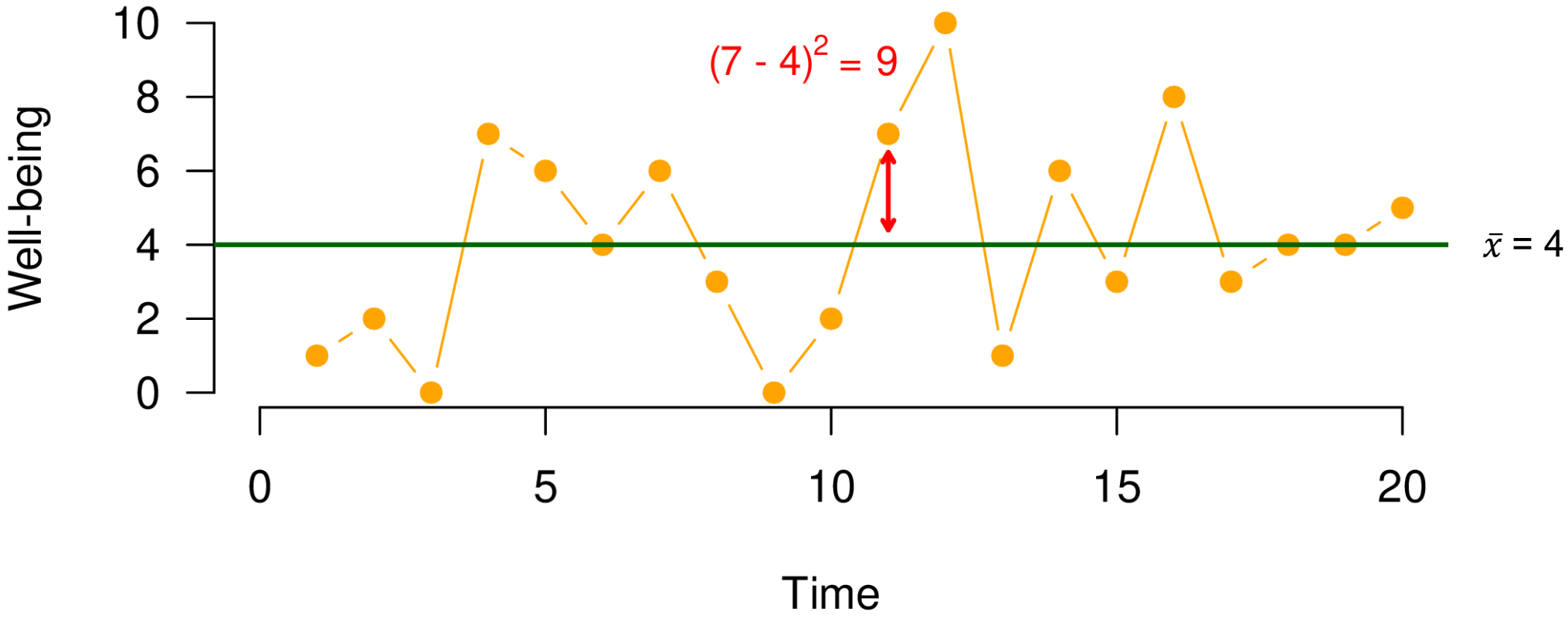
- Interpretation of *s* : A typical distance of an observation from the mean

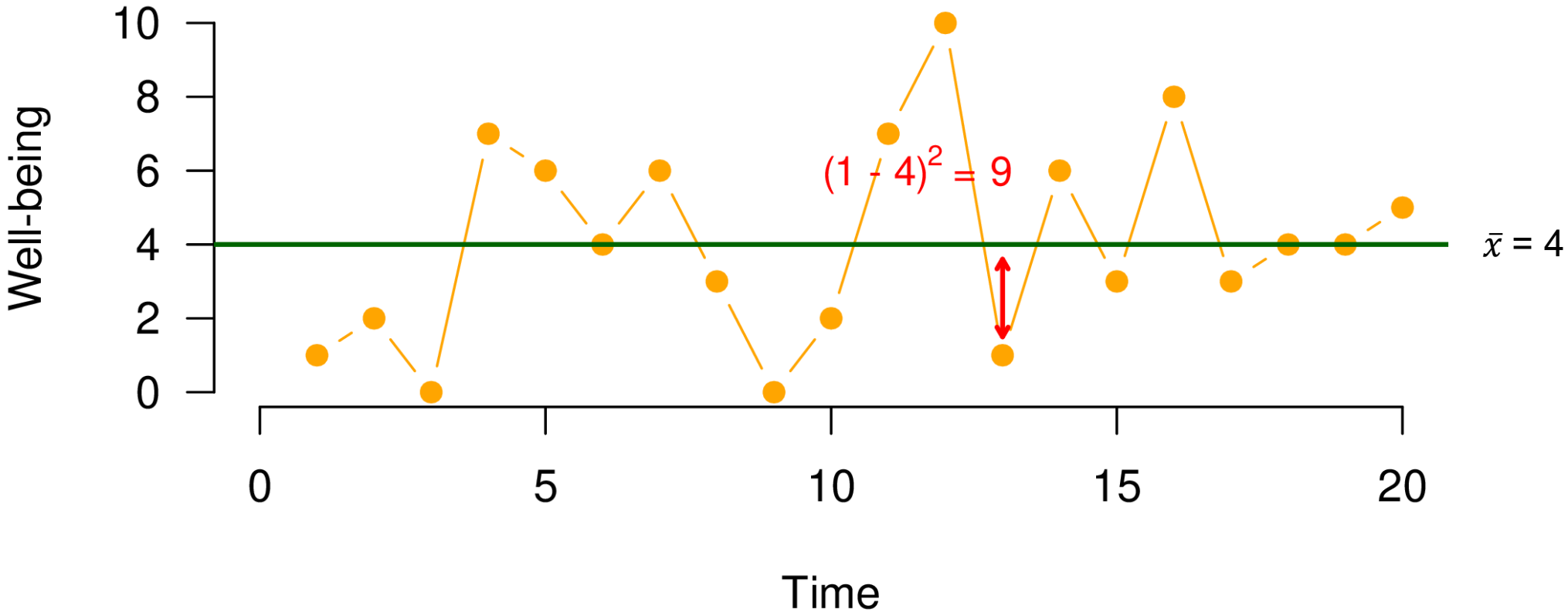

Well-being
10
8
6
4
2
0
0
5
10
15
20
Time

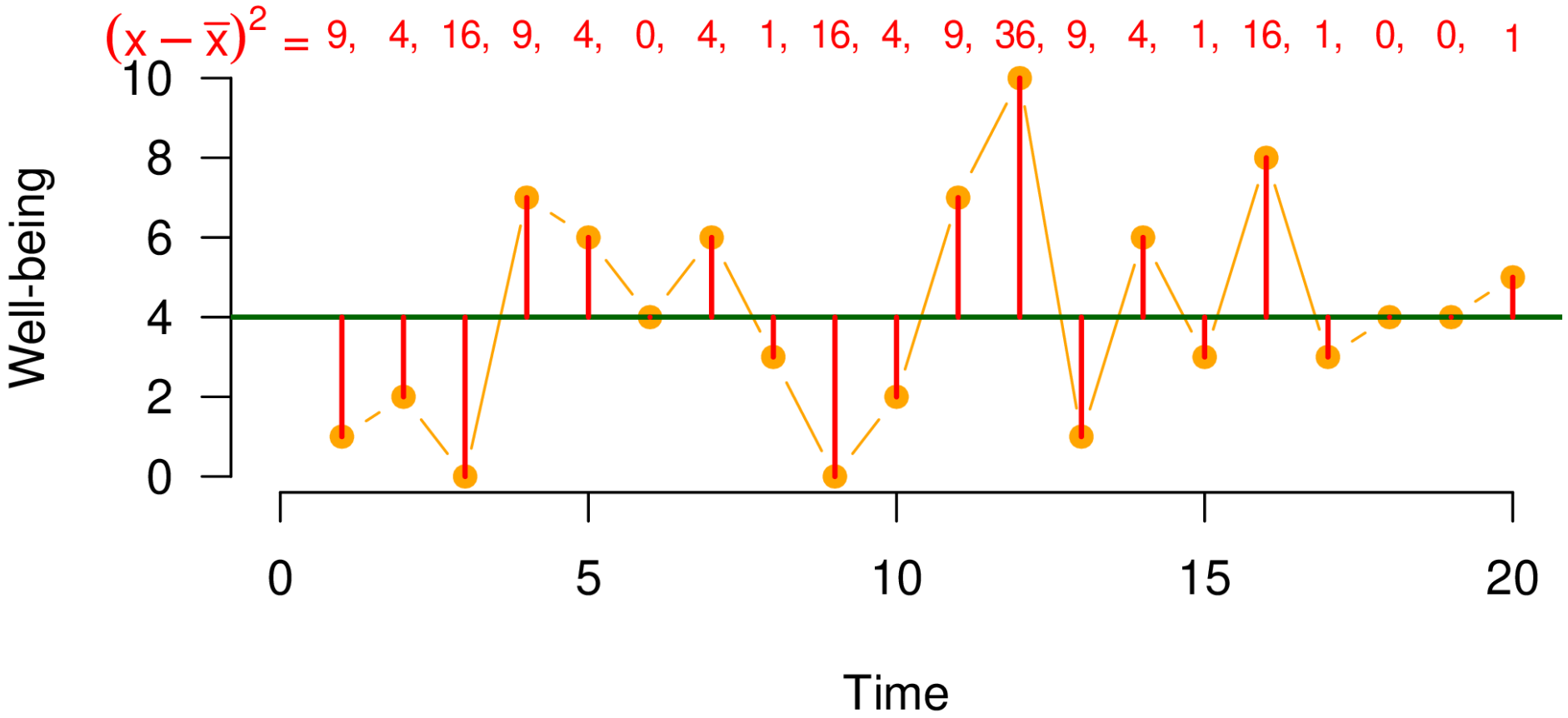

Well-being
10
8
6
4
2
0
0
5
10
15
20
Time

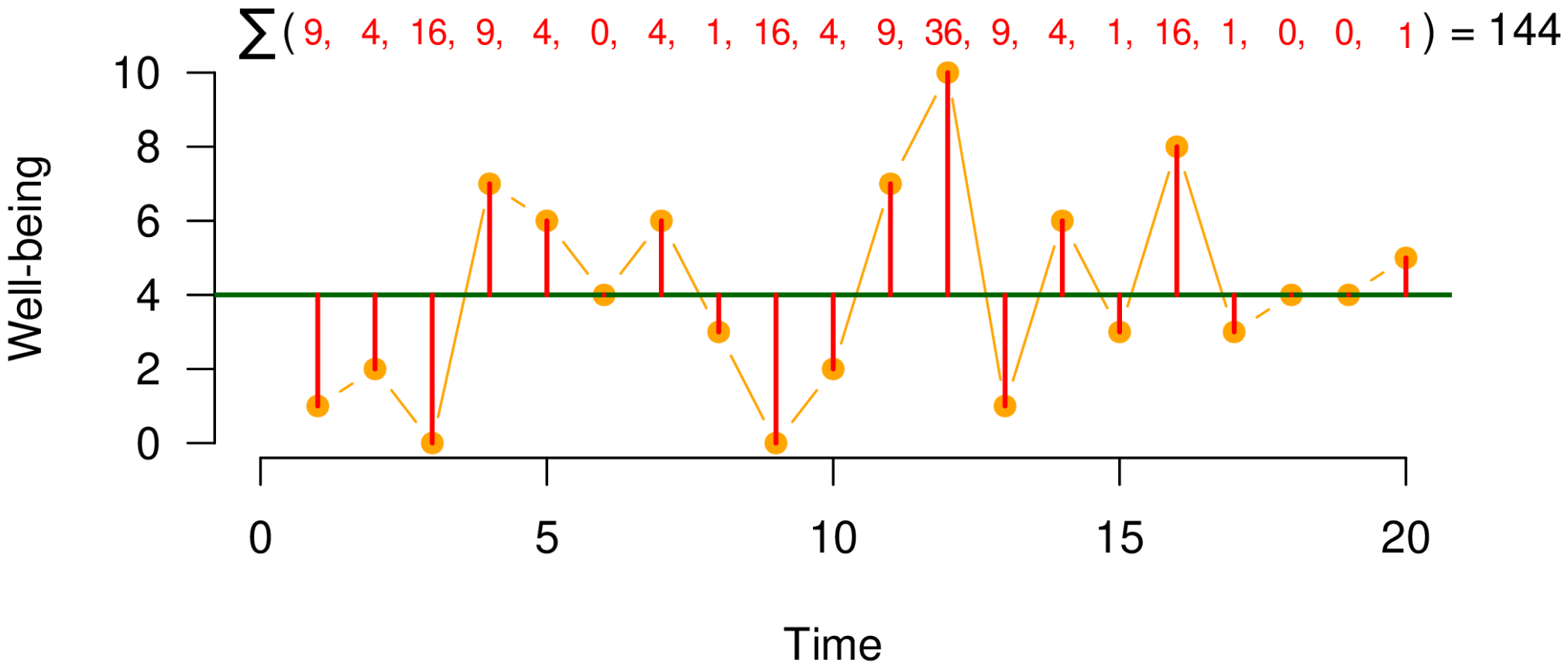

Well-being
10
8
6
4
2
0
$(x - \bar{x})^2$
$\bar{x} = 4$
0
5
10
15
20
Time

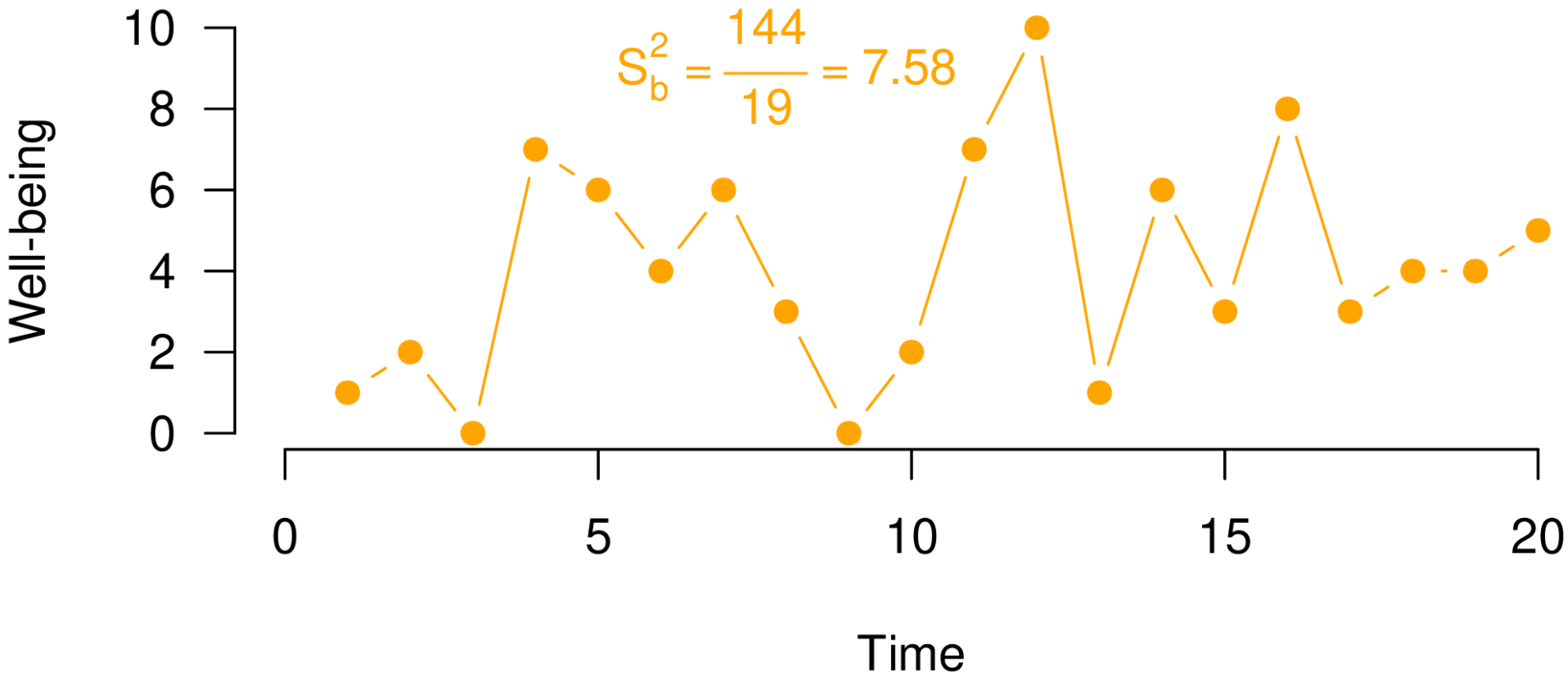

Well-being
10
8
6
4
2
0
$(0 - 4)^2 = 16$
$\bar{x} = 4$
0
5
10
15
20
Time

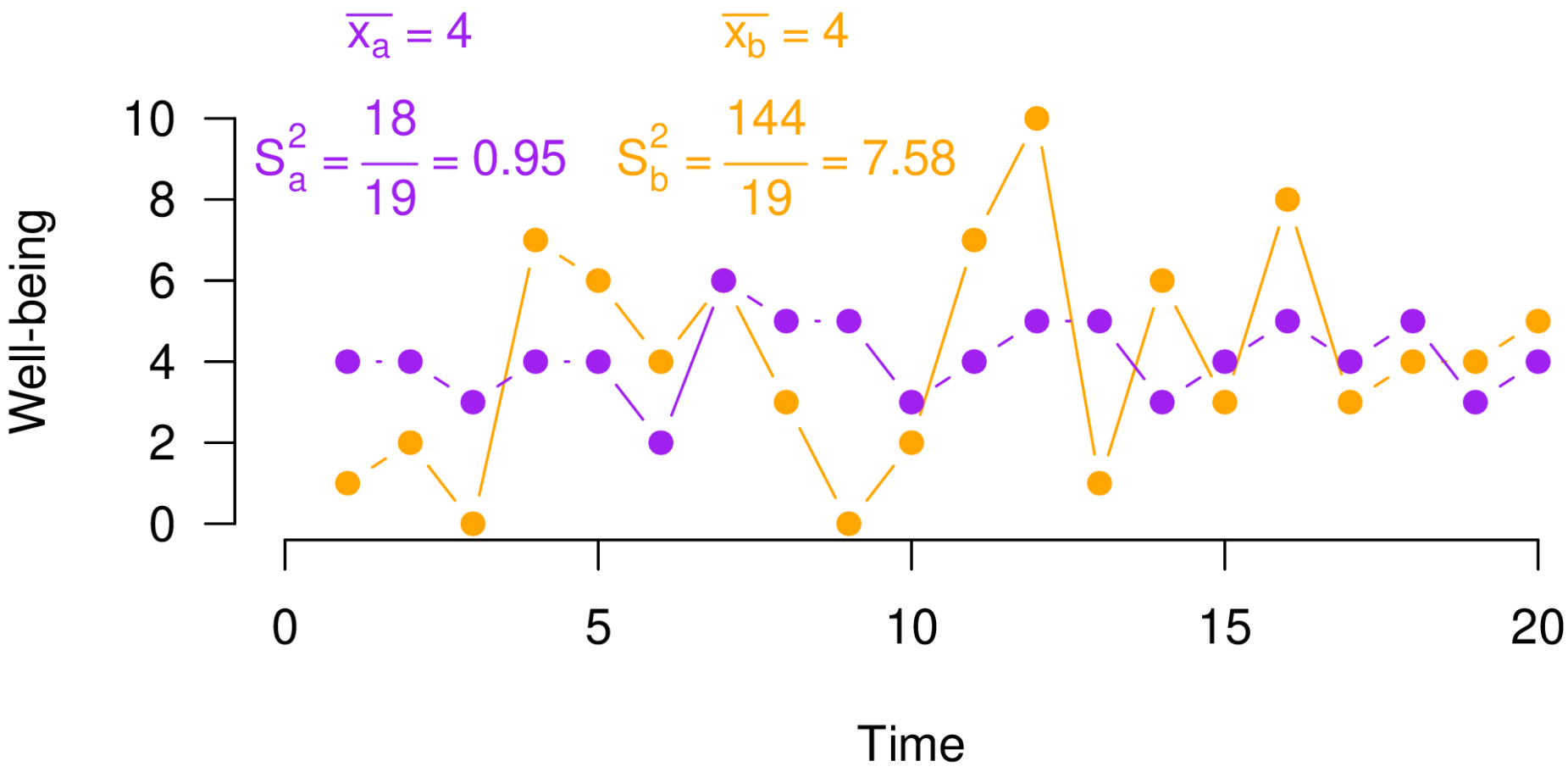

Well-being
10
8
6
4
2
0
0
5
10
15
20
Time
$(7 - 4)^2 = 9$
$\bar{x} = 4$


Well-being
10
8
6
4
2
0
0
5
10
15
20
Time
$(1 - 4)^2 = 9$
$\bar{x} = 4$


$(x - \bar{x})^2$ = 9, 4, 16, 9, 4, 0, 4, 1, 16, 4, 9, 36, 9, 4, 1, 16, 1, 0, 0, 1
Well-being
10
8
6
4
2
0
0
5
10
15
20
Time


∑(9, 4, 16, 9, 4, 0, 4, 1, 16, 4, 9, 36, 9, 4, 1, 16, 1, 0, 0, 1) = 144
Well-being
10
8
6
4
2
0
0
5
10
15
20
Time


$S_b^2 = \frac{144}{19} = 7.58$
Well-being
10
8
6
4
2
0
0
5
10
15
20
Time


$\overline{x_a} = 4$
$\overline{x_b} = 4$
$S_a^2 = \frac{18}{19} = 0.95$
$S_b^2 = \frac{144}{19} = 7.58$
Well-being
10
8
6
4
2
0
0
5
10
15
20
Time

Standard deviation and variance

Variance: Average of the squared deviations

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

Standard deviation: Square root of the variance

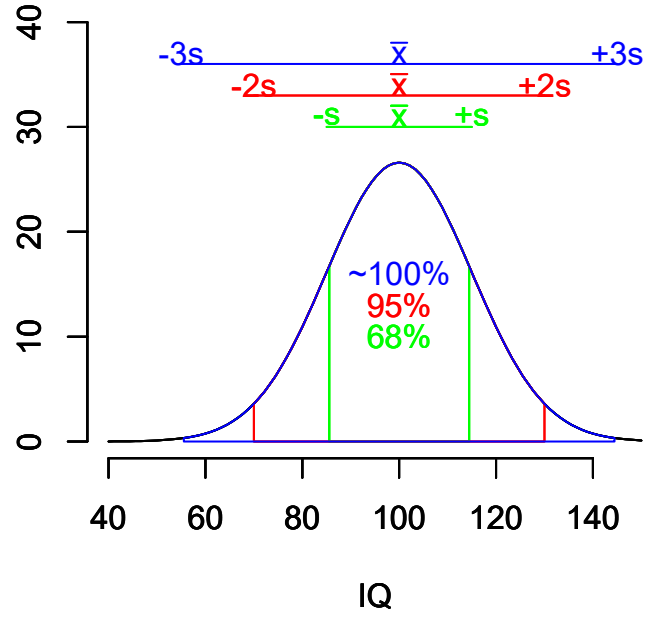
$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

- Interpretation of *s* : A typical distance of an observation from the mean

The variance and sd express the same thing, just on a different scale (because of the square root taking). By itself, the variance/sd does not mean much, but it can be informative to compare two variances/sd's

What can we now express?

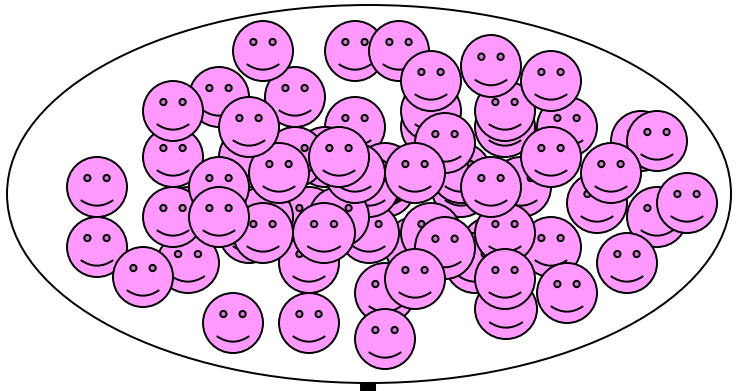
- *Proportion* of the data within a certain range
  - If the data is *Bell shaped*
  - *Empirical Rule*



Symbols

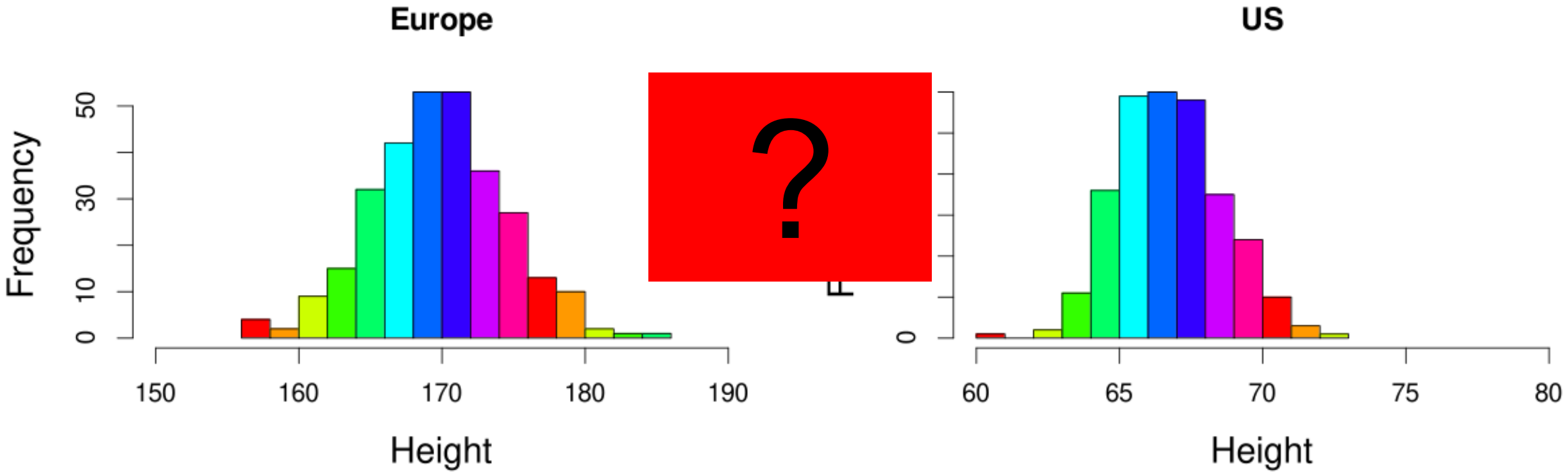
	Sample (statistics, in Latin)	Population (parameters, in Greek)
Location: mean	$\bar{x} = \frac{\sum x}{n}$	μ
Variability: variance	$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1}$	σ^2
Variability: standard deviation	$s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$	σ

Population
(μ, σ)

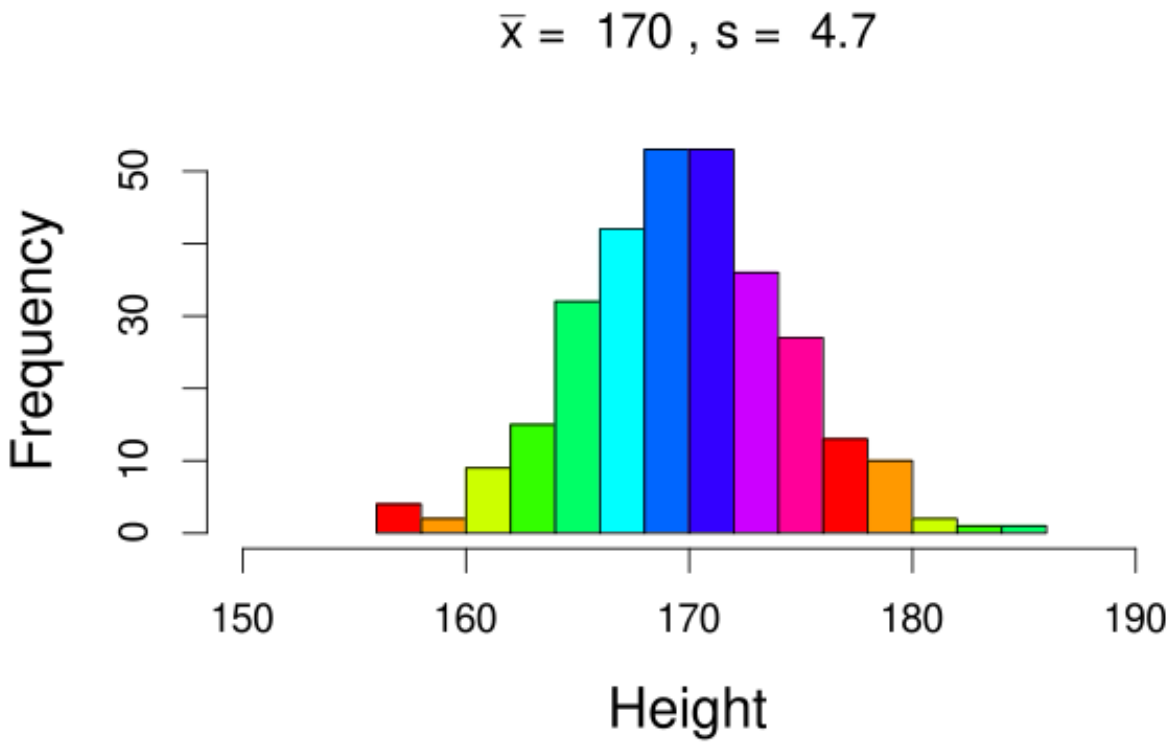


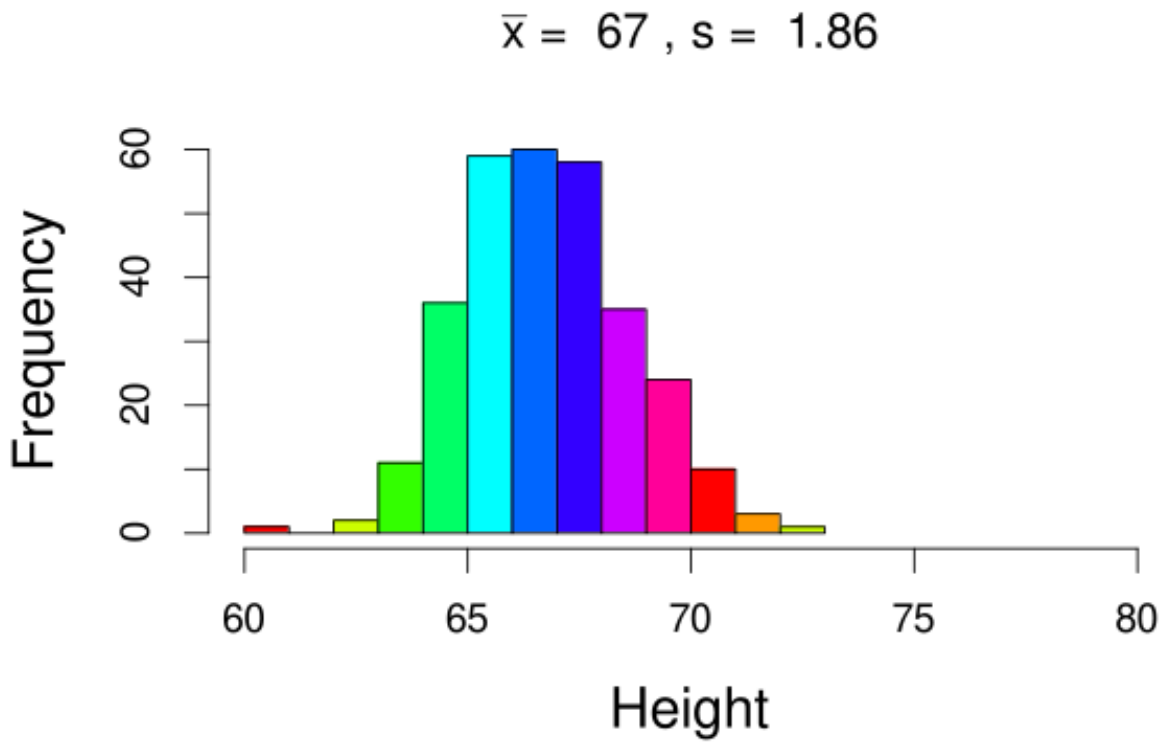
Sample
$(\bar{x}, s)$

Measuring Height Across Cultures



Measuring Height Across Cultures





Measuring Height Across Cultures

$\bar{x}$ = 170 , s = 4.7

Frequency

0 10 30 50

$\bar{x}$ = 67 , s = 1.86

Frequency

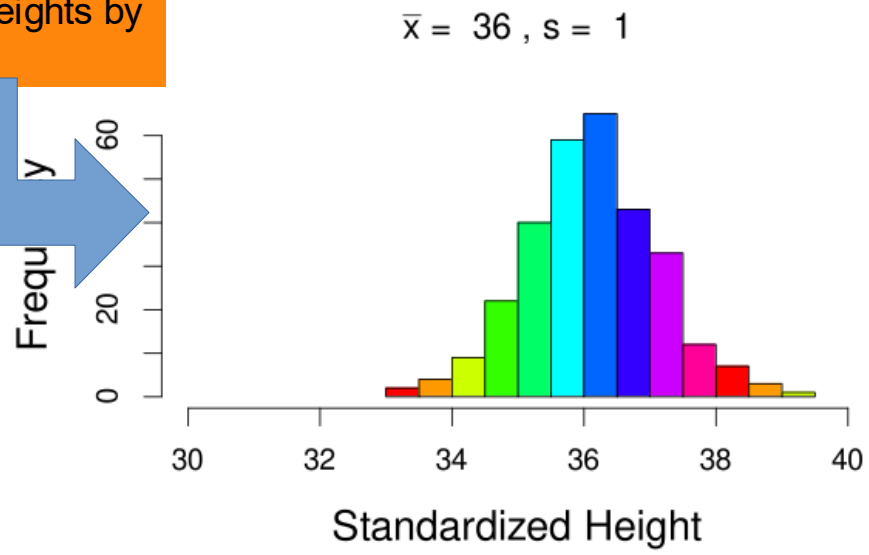
80

Source: https://pixabay.com/

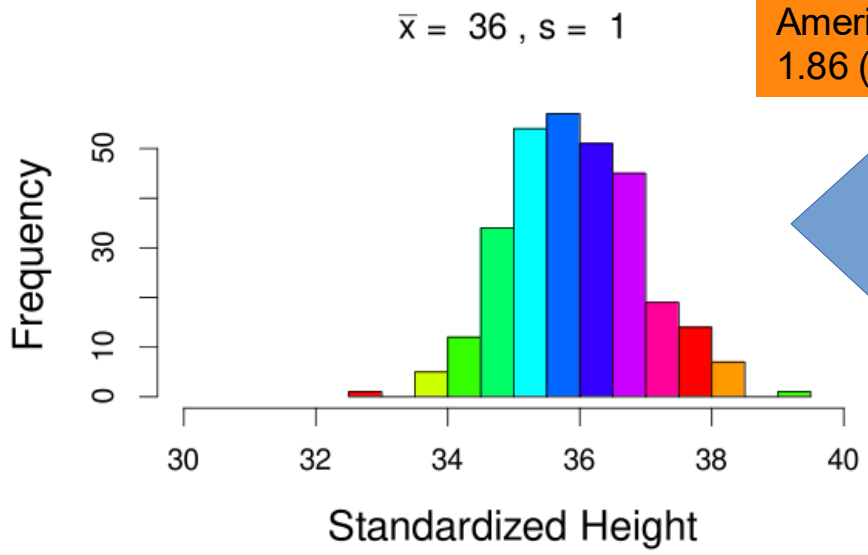
Measuring Height Across Cultures

We can use the standard deviation to **standardize** the scores, such that the original measurement scales do not matter:

We divided all European heights by 4.7 (= sd)



We divided all American heights by 1.86 (= sd)



Measuring Height Across Cultures

We can use the standard deviation to **standardize** the scores, such that the original measurement scales do not matter:

We divided all European heights b 4.7 (= sd)

e divided all nerican heights by 6 (= sd)

40

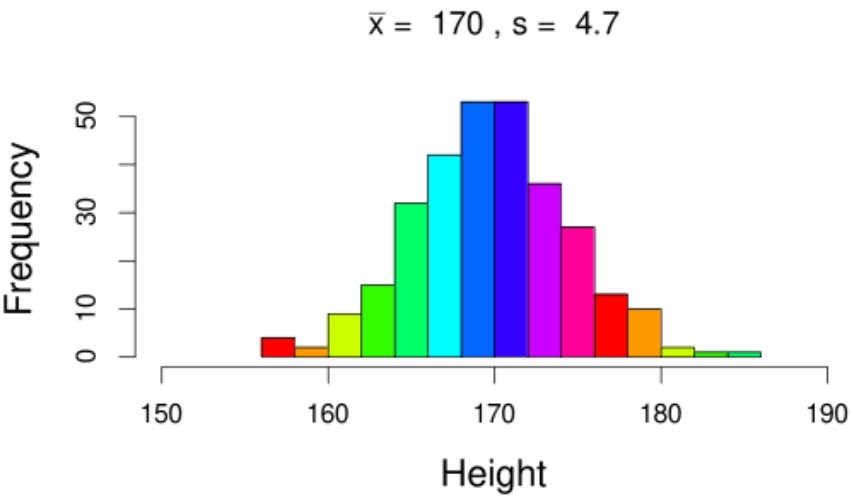
Z-score is a way to also standardize the mean

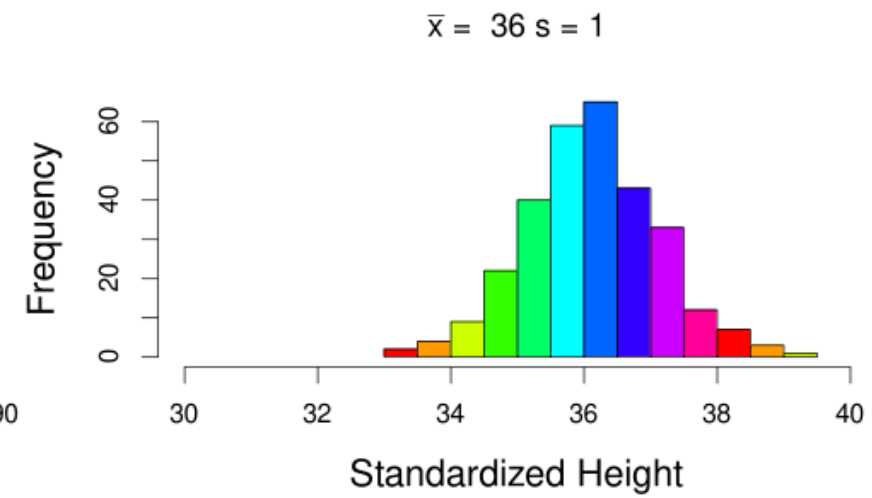
- You can recompute all observations to a z-score

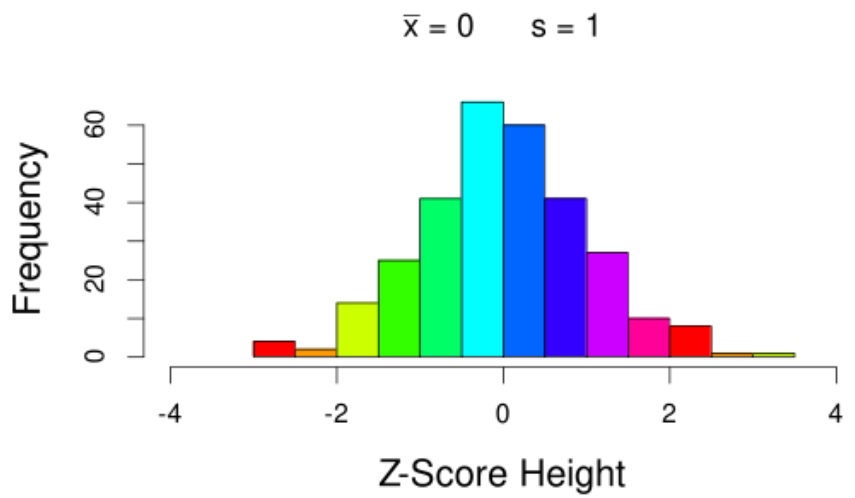
- Mean is always 0
- Standard deviation is always 1

z-score: $$z = \frac{\text{observation - mean}}{\text{standard devation}}$$

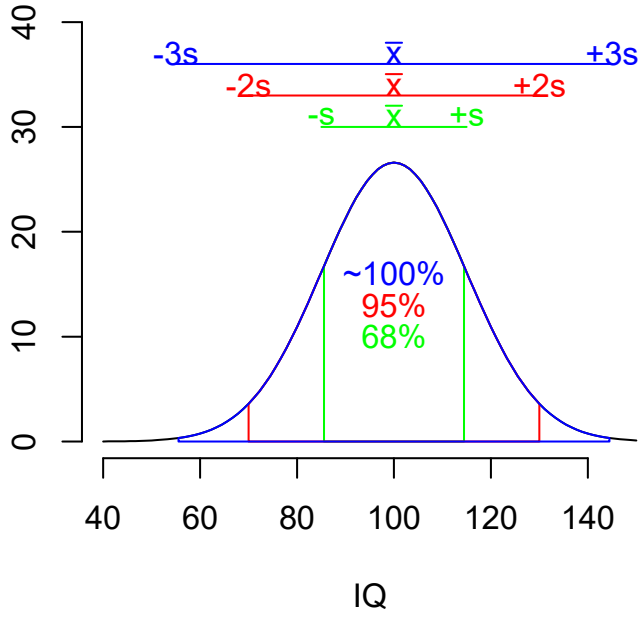
Z-score is a way to also standardize the location



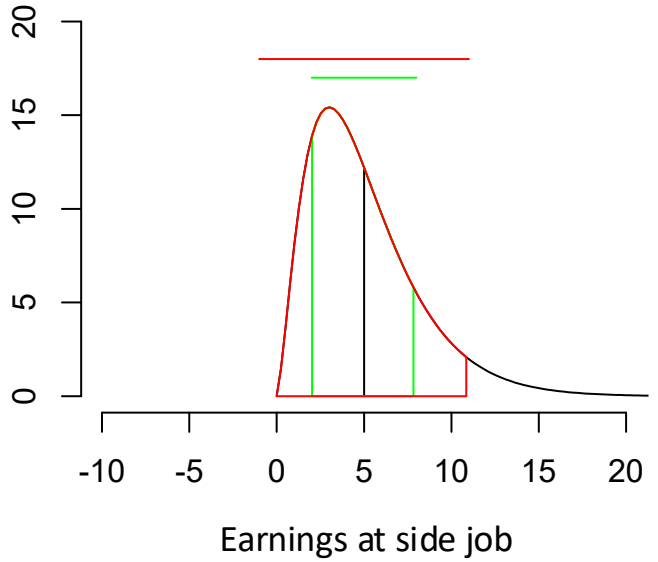




Empirical rule only applies to Bell shaped distributions



Symmetric: Equal proportion left and right of the mean

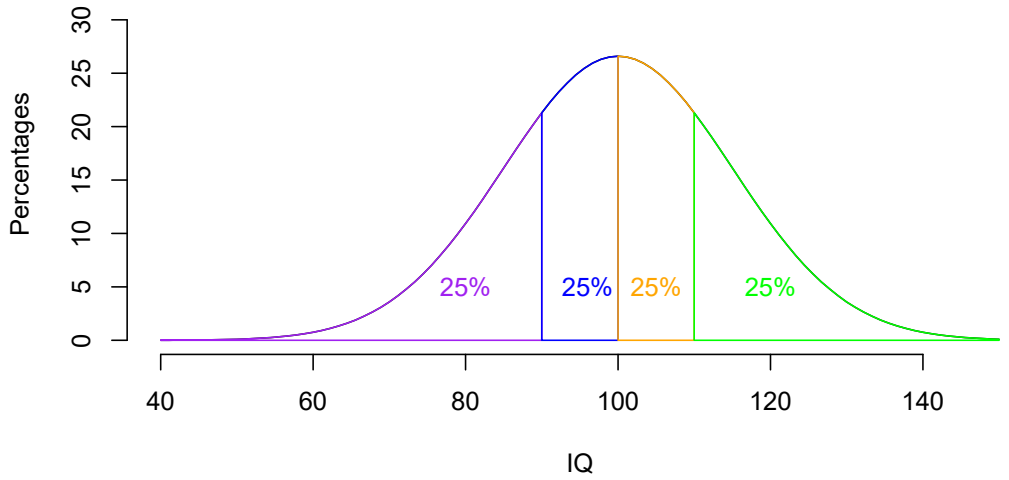


(Right) skewed: Higher proportion right than left

Other solution: Quartiles

- Proportion of the data within a certain range

Q1 = 242: 25% earns less then 242 euro
Q2 = 473: 50% earns less then 473 euro
= Median
Q3 = 821: 75% earns less then 821 euro

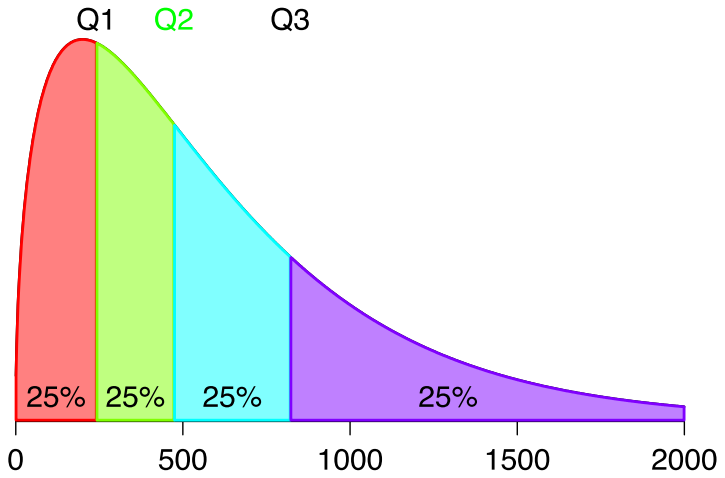


Agresti p. 65 explains how to find the quartiles
→ You do not need to find these yourself, but understand them

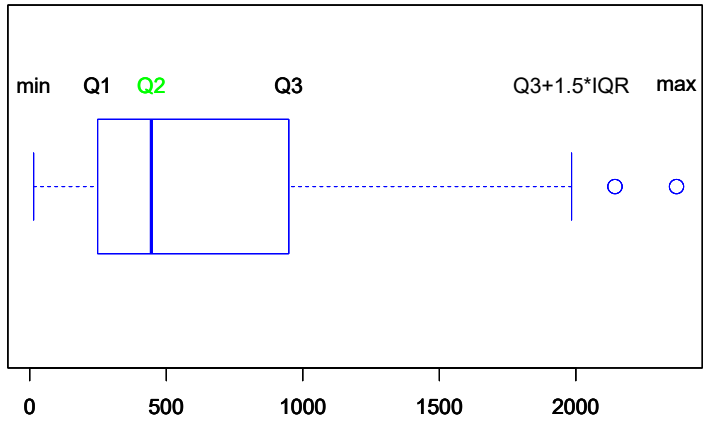
Box plot

- IQR: Interquartile range
  - IQR = Q3 - Q1
- Whiskers
  - Q1 – 1.5 x IQR (or minimum)
  - Q3 + 1.5 x IQR (or maximum)
  - Outliers?

Source pixabay.com



Earnings at side job



Earnings at side job

Today

1. Variability in the data
   - Standard deviation and variance
   - z-score
   - Quartiles
   - Boxplot
2. **Associations**
   - Between two categorical variables
   - Between two quantitative variables
3. Recap
   - Next time
   - Example exam question

Association

Association: There is an association if a particular value of one variable is more likely to occur with certain values of the other variable

- One variable *depends on* the other variable

- Two variable types:

Response variable: The variable for which you want to explain/predict the outcome
Explanatory variable: The variable that you use to predict

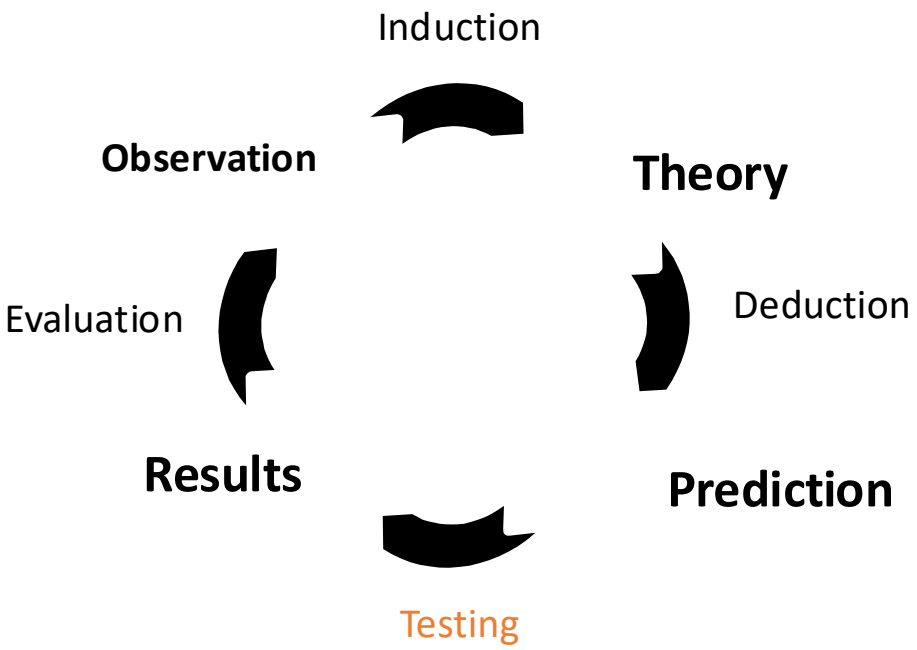
- Note different terminology:
  - Response variable → dependent variable
  - Explanatory variable → independent variable, predictor variable

Response or explanatory variable?

- Are older students generally happier?
  - “Age student”?
    - Explanatory
  - “Subjective well-being score”?
    - Response
- Does sitting in a dark room enhance cognitive ability?
  - “Reaction time”
    - Response
  - “Room illumination”
    - Explanatory

Role of variables in empirical cycle

Operationalize: determine how you will measure the conceptual variables from the prediction



Role of variables in empirical cycle

Explanatory variable

Response variable	Quantitative	Categorical
Quantitative	Correlation Regression	t-test ANOVA
Categorical	Logistic regression	Contingency table

Today

1. Variability in the data
   - Standard deviation and variance
   - z-score
   - Quartiles
   - Boxplot
2. Associations
   - **Between two categorical variables**
   - Between two quantitative variables
3. Closing
   - Next time
   - Example exam question

Association between two categorical variables

One flew over the Cuckoo's nest (1975)
source:
https://i.ytimg.com/vi/d_mASr1djMM/maxresdefault.jpg

Association between two categorical variables

- Research question: Do people that are treated with electroshock therapy recover from a psychotic disorder?

- Both the response and explanatory variables are categorical (yes/no)
  - Contingency table!

Explanatory variable

Response variable	Quantitative	Categorical
Quantitative	Correlation Regression	t-test ANOVA
Categorical	Logistic regression	Contingency table

Contingency Table

Counts

Treatment? \ Recovered?	No	Yes	Total
No	21	19	40
Yes	13	27	40
Total	34	46	**80(= n)**

Proportions

Treatment? \ Recovered?	No	Yes	Total
No	0.26	0.24	0.5
Yes	0.16	0.34	0.5
Total	0.42	0.58	**80(= n)**

(e.g., 21 / 80 = 0.26)

Conditional Proportions

Conditional proportion: The proportion of the response variable, for *one level* of the explanatory variable

Recovered?

Treatment?	No	Yes	Total
No	21	19	40
Yes	**13**	**27**	**40**
Total	34	46	**80(= n)**

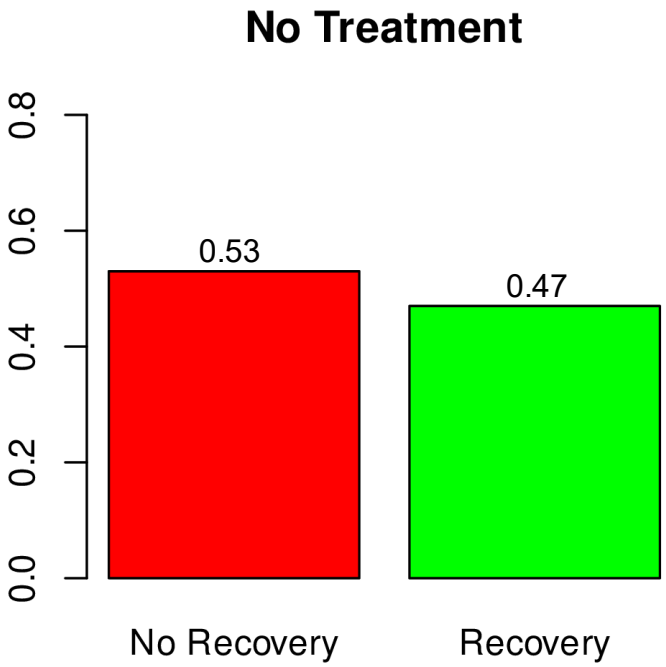
Recovered?

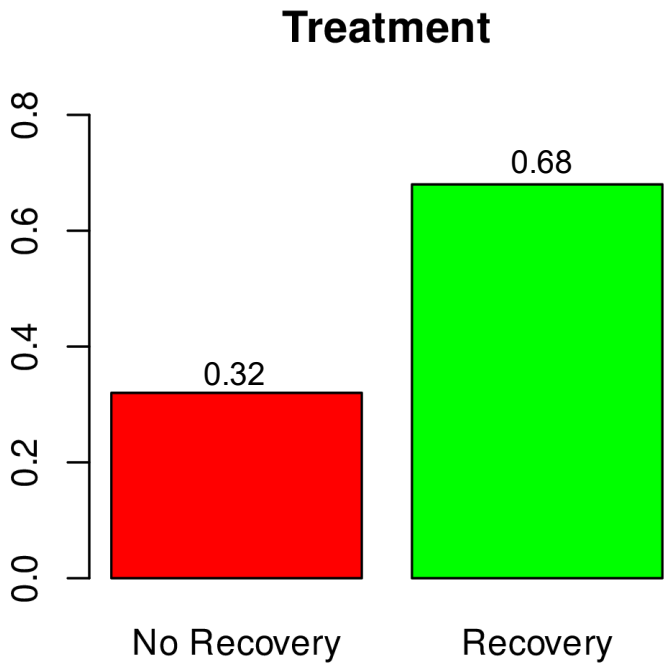
Treatment?	No	Yes	Total
No	0.53	0.47	1
Yes	0.32	0.68	1
Total			**80(= n)**

(e.g., 13 / 40 = 0.32)

Conditional Proportions

Conditional proportion: The proportion of the response variable, for *one level* of the explanatory variable





Recovered?

Treatment?	No	Yes	Total
No	0.53	0.47	1
Yes	0.32	0.68	1
Total			80(= n)

What can we now express?

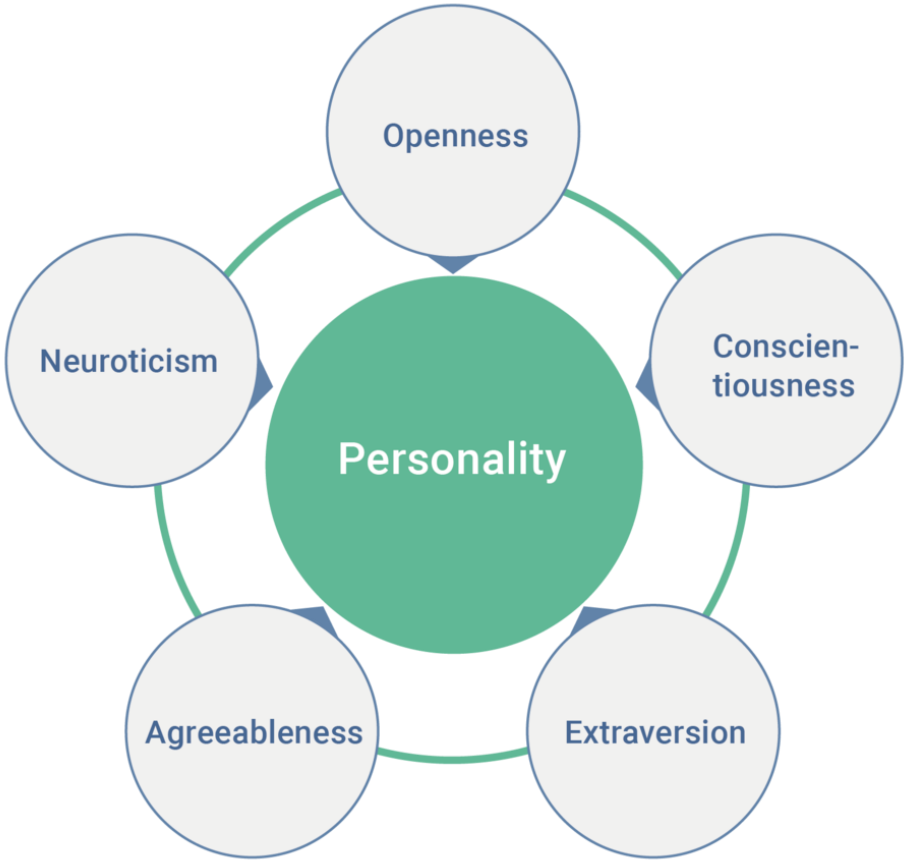
- Electroshock therapy seems to affect the chance of recovery
  - Treatment and recovery are *associated*
  - In this sample → Descriptive statistics
  - In week 9 we will discuss inferential statistics

Today

1. Variability in the data
   - Standard deviation and variance
   - z-score
   - Quartiles
   - Boxplot
2. Associations
   - Between two categorical variables
   - **Between two quantitative variables**
3. Recap
   - Next time
   - Example exam question

Association between two quantitative variables

Big 5 Personality Inventory

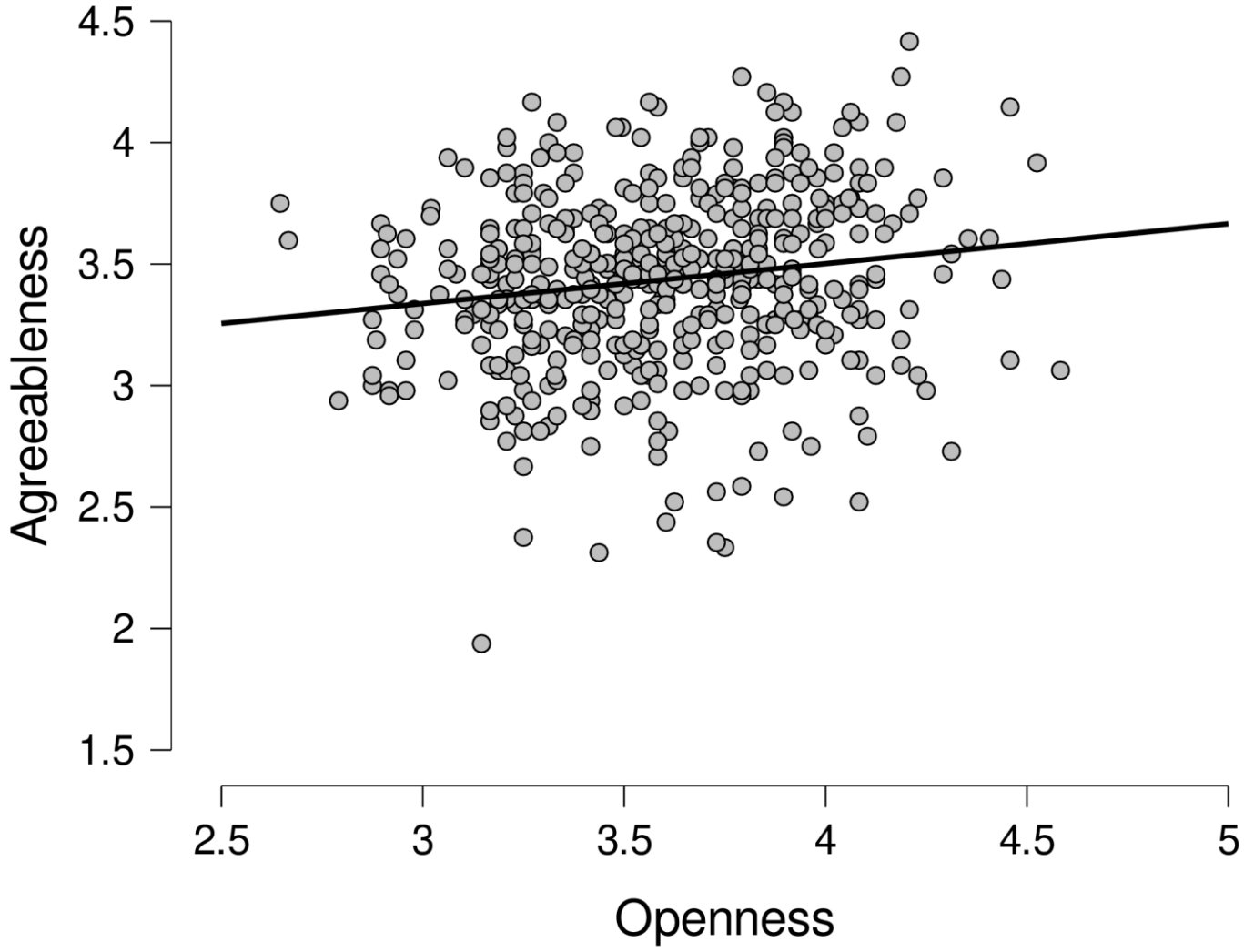


Source: https://en.wikipedia.org/wiki/Big_Five_personality_traits

Big 5 Personality Inventory

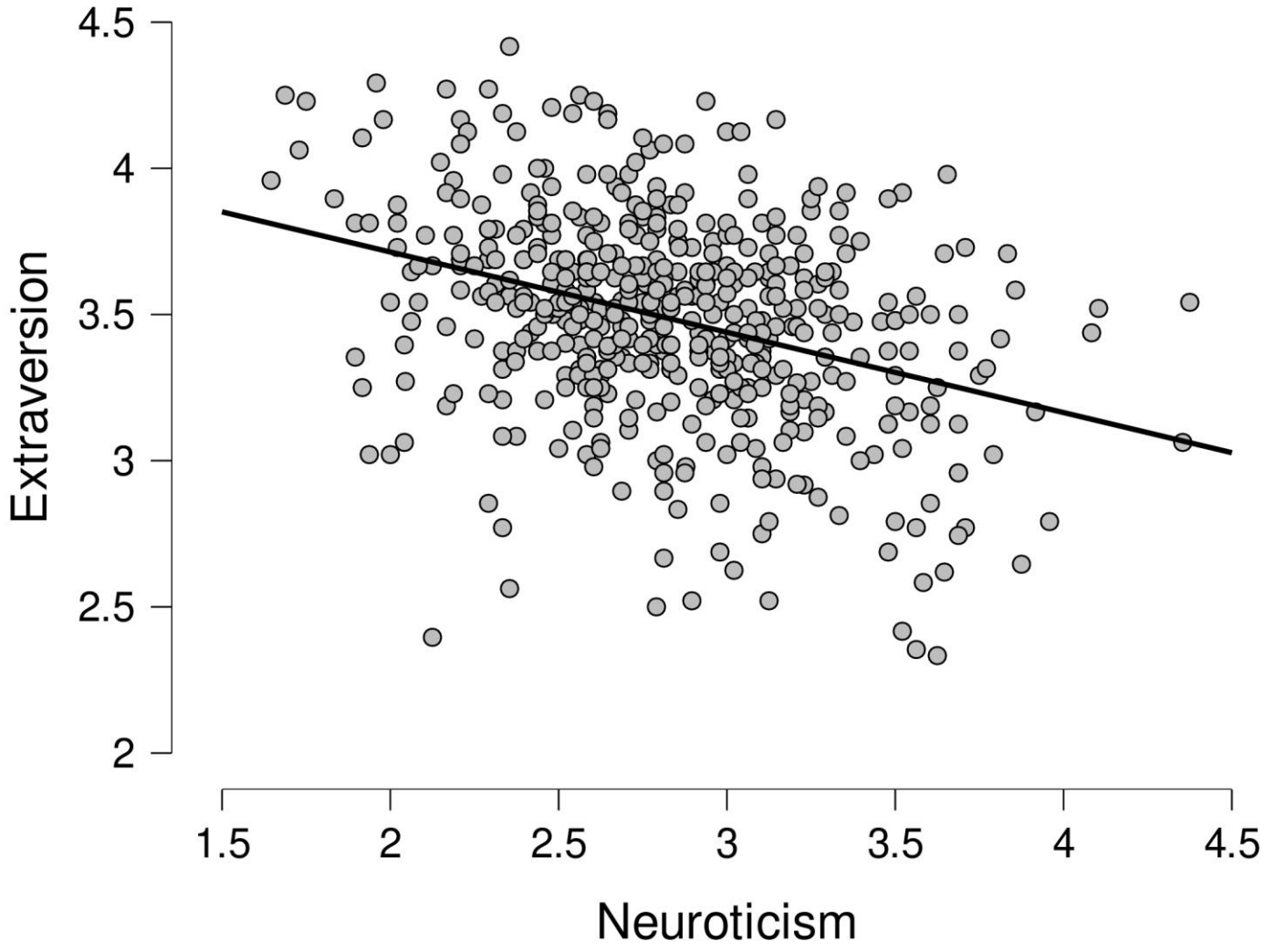
- Openness
  - “I am quick to understand things.”
- Conscientiousness
  - “I am always prepared.”
- Extraversion
  - “I am the life of the party.”
- Agreeableness
  - “I feel others' emotions.”
- Neuroticism
  - “I get stressed out easily.”

Association – Big 5 Personality Inventory



Data from Conor V. Dolan, Frans J. Oort, Reinoud D. Stoel & Jelte M. Wicherts (2009) Testing Measurement Invariance in the Target Rotated Multigroup Exploratory Factor Model, Structural Equation Modeling: A Multidisciplinary Journal, 16:2, 295-314

Association – Big 5 Personality Inventory



Data from Conor V. Dolan, Frans J. Oort, Reinoud D. Stoel & Jelte M. Wicherts (2009) Testing Measurement Invariance in the Target Rotated Multigroup Exploratory Factor Model, Structural Equation Modeling: A Multidisciplinary Journal, 16:2, 295-314

Correlation

$$r = \frac{1}{n-1}\sum z_x z_y$$ (Agresti p. 107)

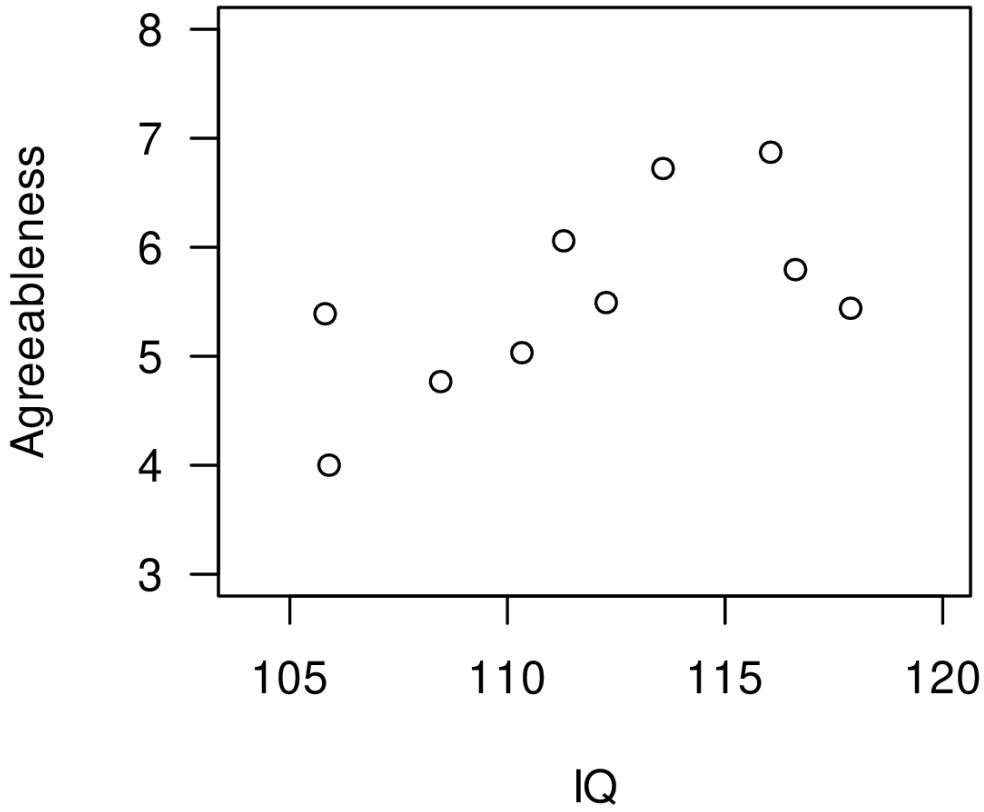
r is “average” contribution to the association of each pair of observations

Properties correlation:

- *r* between -1 and 1
- *r* > 0: positive association
- *r* < 0: negative association
- Does not depend on the scale of the variables
- Remains the same if the role of response and explanatory variables flips

Correlation

	Agree		IQ	
	6.7		113.6	
	5.4		105.8	
	5.4		117.9	
	4.8		108.5	
	5.8		116.6	
	6.9		116.0	
	5.0		110.3	
	5.5		112.3	
	4.0		105.9	
	6.1		111.3	
Mean	5.6		111.8	
s	0.87		4.3	



Correlation

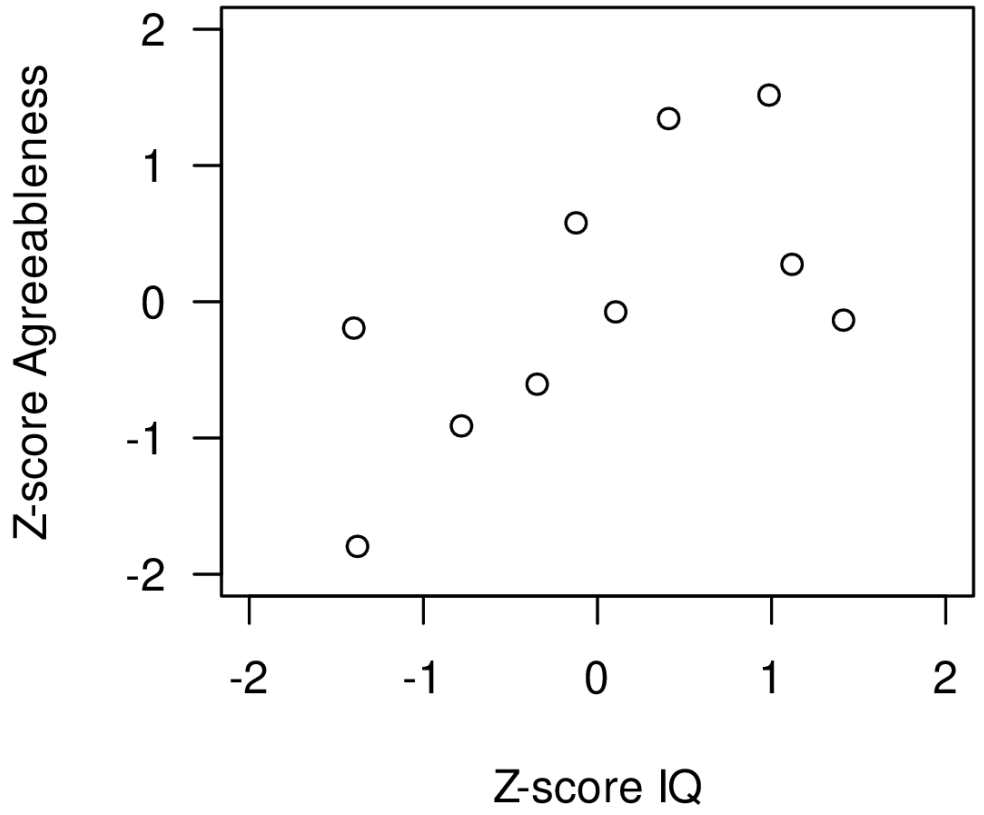
e.g., (6.7 – 5.6) / 0.87 = 1.3

	Agree	Z-agree	IQ	
	6.7	1.3	113.6	
	5.4	-0.2	105.8	
	5.4	-0.1	117.9	
	4.8	-0.9	108.5	
	5.8	0.3	116.6	
	6.9	1.5	116.0	
	5.0	-0.6	110.3	
	5.5	-0.1	112.3	
	4.0	-1.8	105.9	
	6.1	0.6	111.3	
Mean	5.6	0	111.8	
s	0.87	1	4.3	

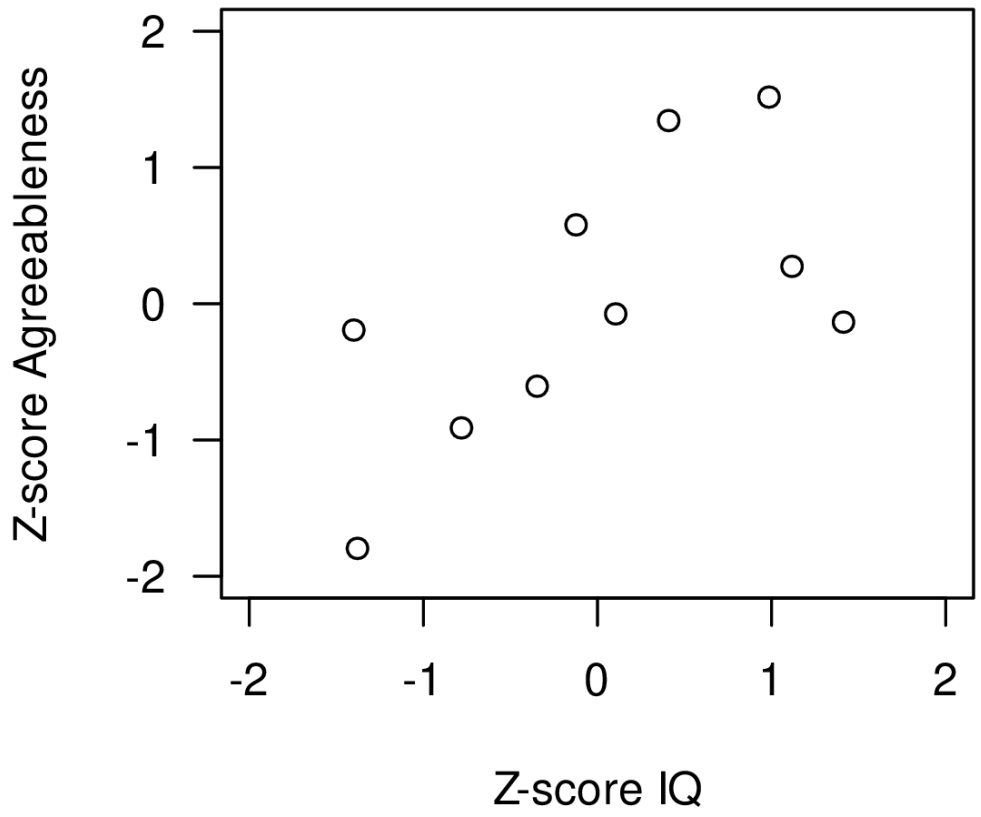
Correlation

e.g., (113.6 – 111.8) / 4.3 = 0.4

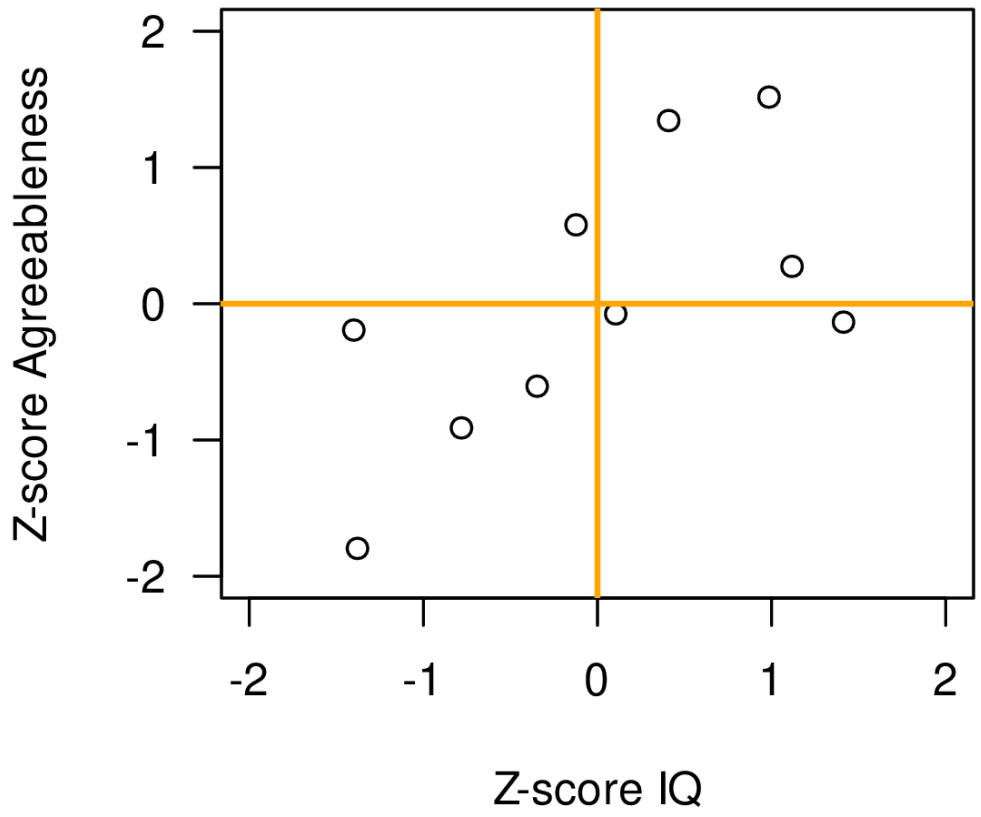
	Agree	Z-agree	IQ	Z-IQ
	6.7	1.3	113.6	0.4
	5.4	-0.2	105.8	-1.4
	5.4	-0.1	117.9	1.4
	4.8	-0.9	108.5	-0.8
	5.8	0.3	116.6	1.1
	6.9	1.5	116.0	1.0
	5.0	-0.6	110.3	-0.3
	5.5	-0.1	112.3	0.1
	4.0	-1.8	105.9	-1.4
	6.1	0.6	111.3	-0.1
Mean	5.6	0	111.8	0
s	0.87	1	4.3	1



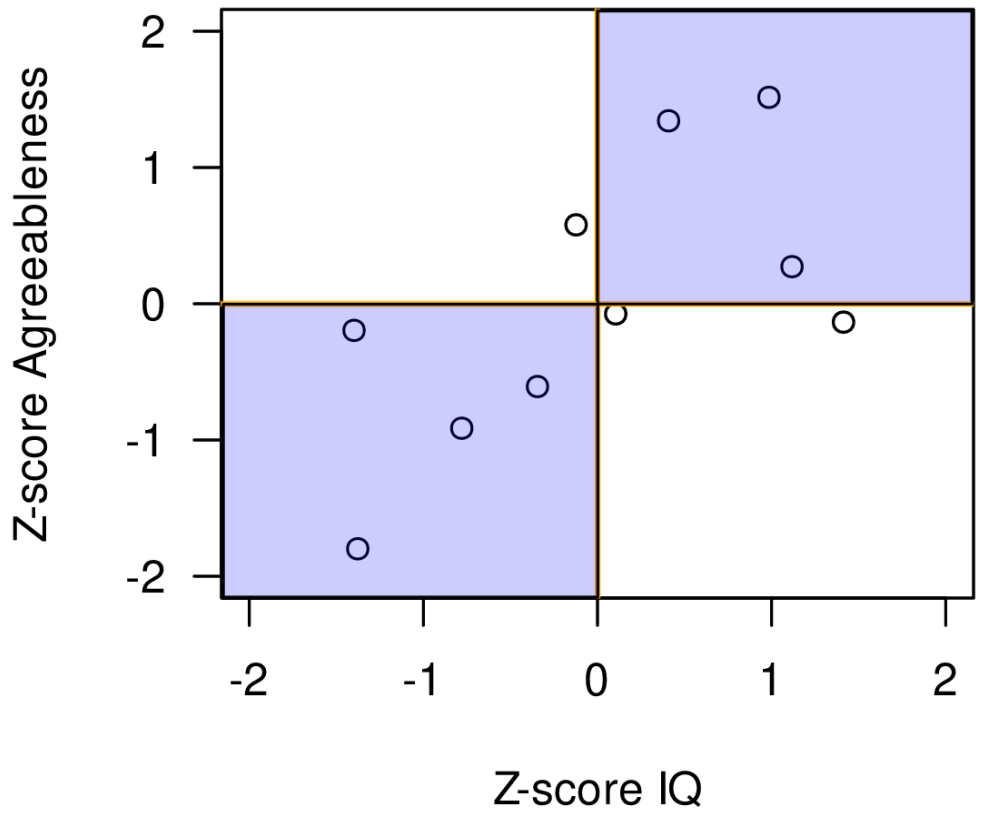
Correlation



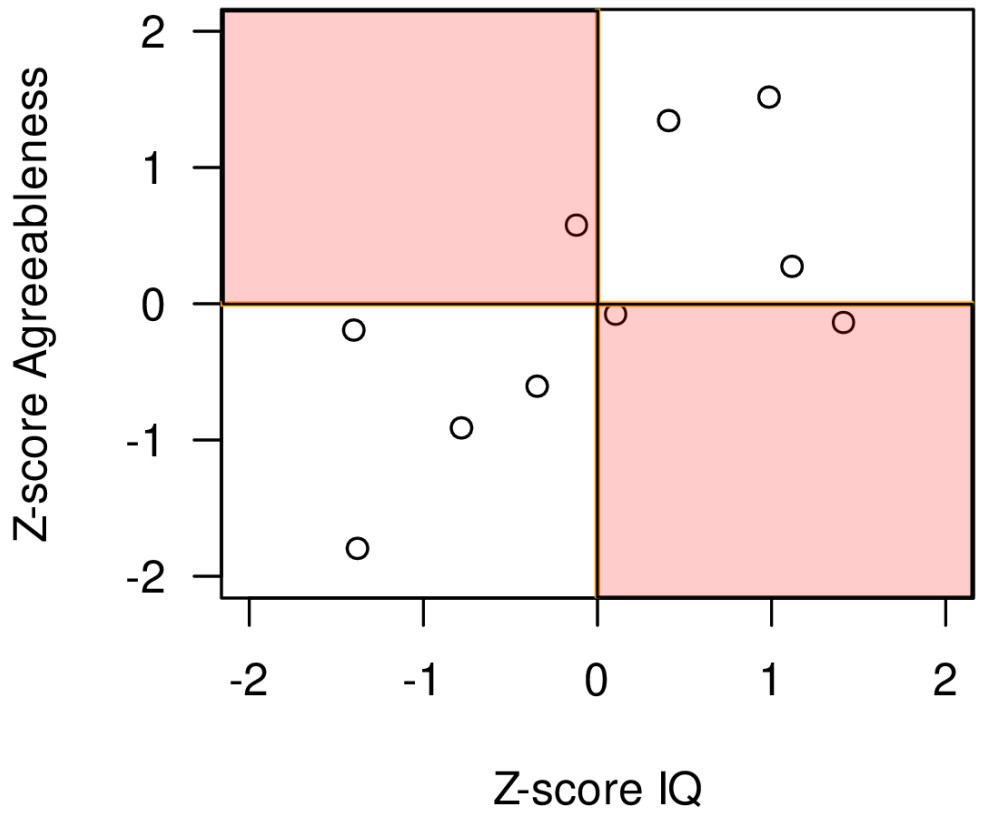
Correlation



Correlation

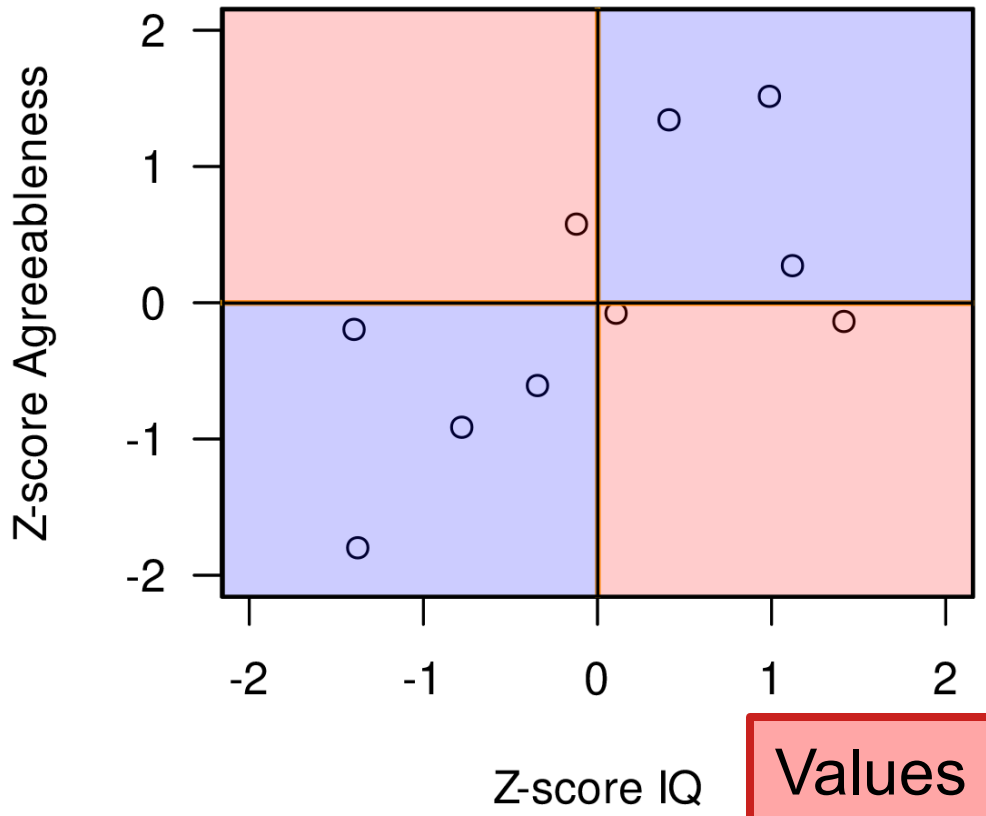


Correlation



Correlation

Values in the blue quadrant: $Z_x Z_y > 0$
Contribute to *positive* correlation

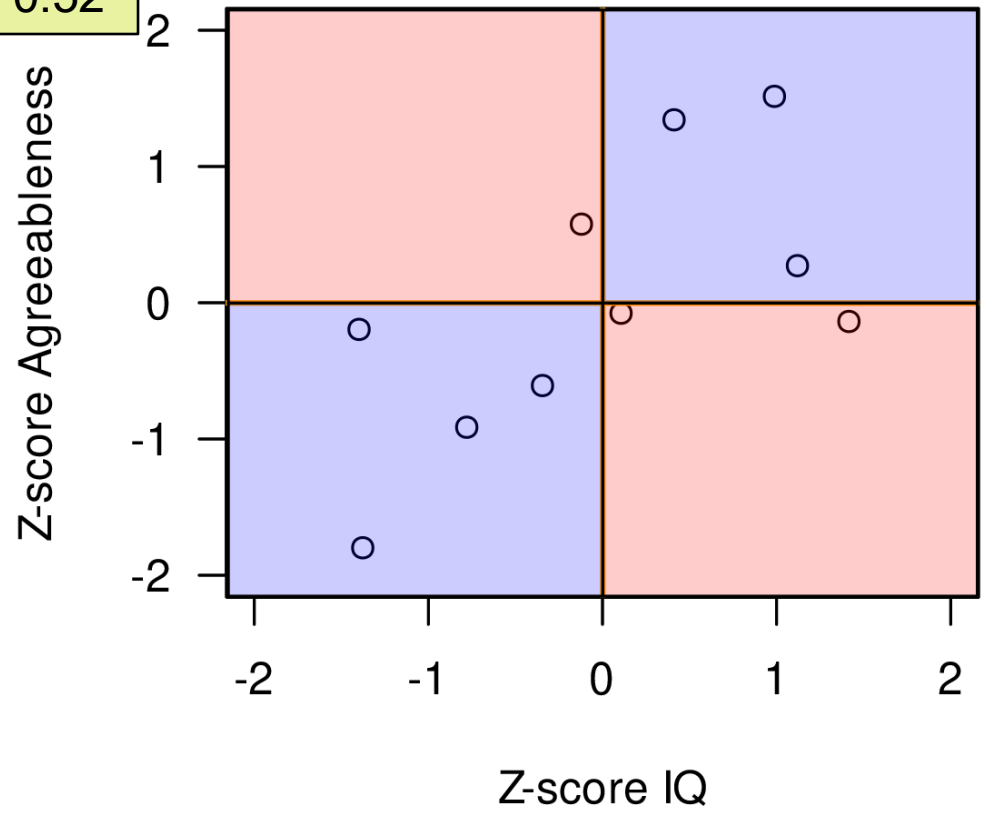


Values in the red quadrant: $Z_x Z_y < 0$
Contribute to *negative* correlation

Correlation

Z-agree	Z-IQ	$Z_x Z_y$
1.3	0.4	0.52
-0.2	-1.4	0.28
-0.1	1.4	-0.14
-0.9	-0.8	0.72
0.3	1.1	0.33
1.5	1.0	1.50
-0.6	-0.3	0.18
-0.1	0.1	-0.01
-1.8	-1.4	2.52
0.6	-0.1	-0.06

e.g., 1.3 * 0.4 = 0.52



Correlation

$$r = \frac{1}{n-1}\sum z_x z_y$$

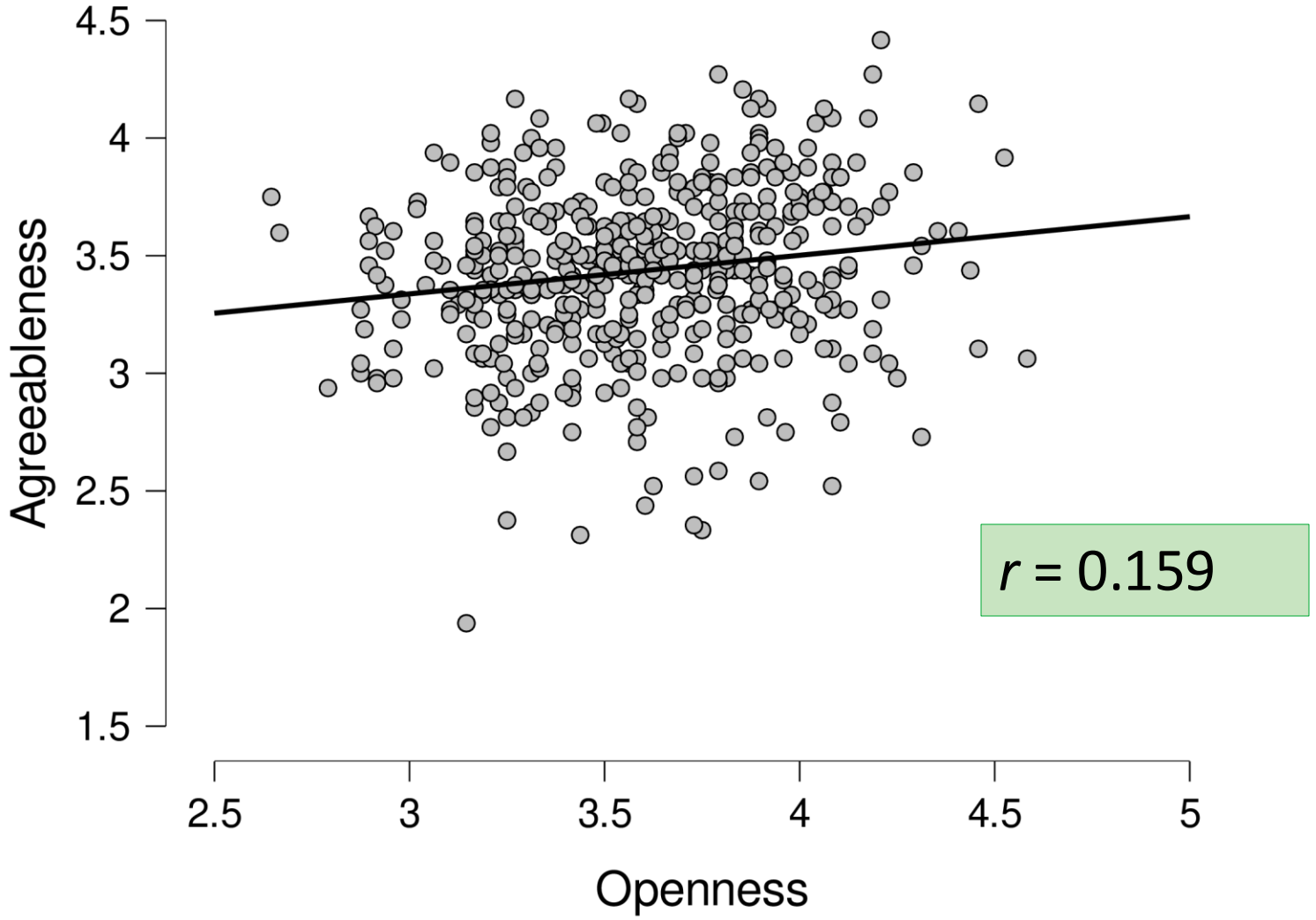
Z-agree	Z-IQ	$Z_x Z_y$
1.3	0.4	0.52
-0.2	-1.4	0.28
-0.1	1.4	-0.14
-0.9	-0.8	0.72
0.3	1.1	0.33
1.5	1.0	1.50
-0.6	-0.3	0.18
-0.1	0.1	-0.01
-1.8	-1.4	2.52
0.6	-0.1	-0.06
		5.84 **Sum**

n = 10

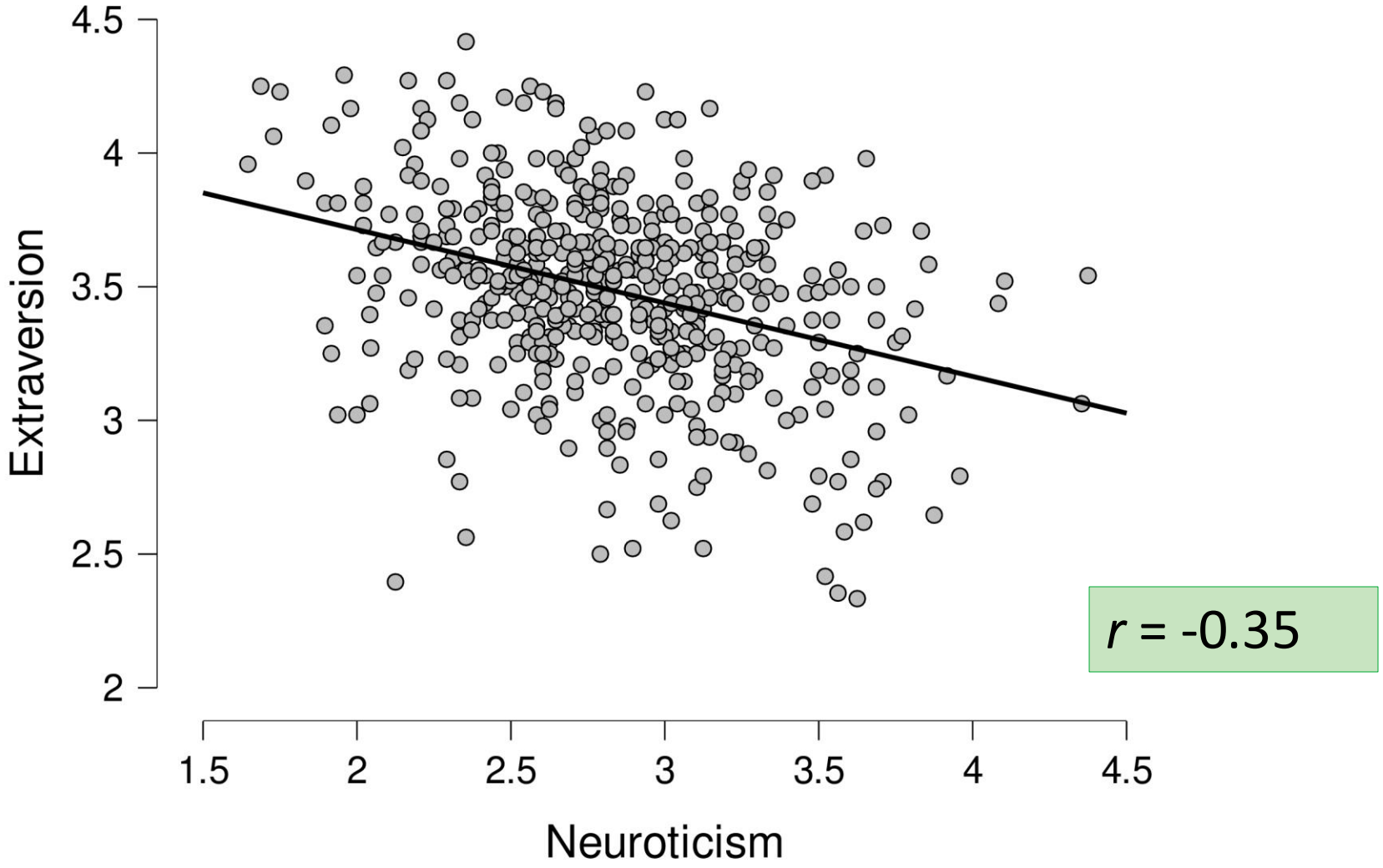
$\sum z_x z_y$ = 5.84

r = 5.84 / (10 - 1) = 0.65

Association – Big 5 Personality Inventory

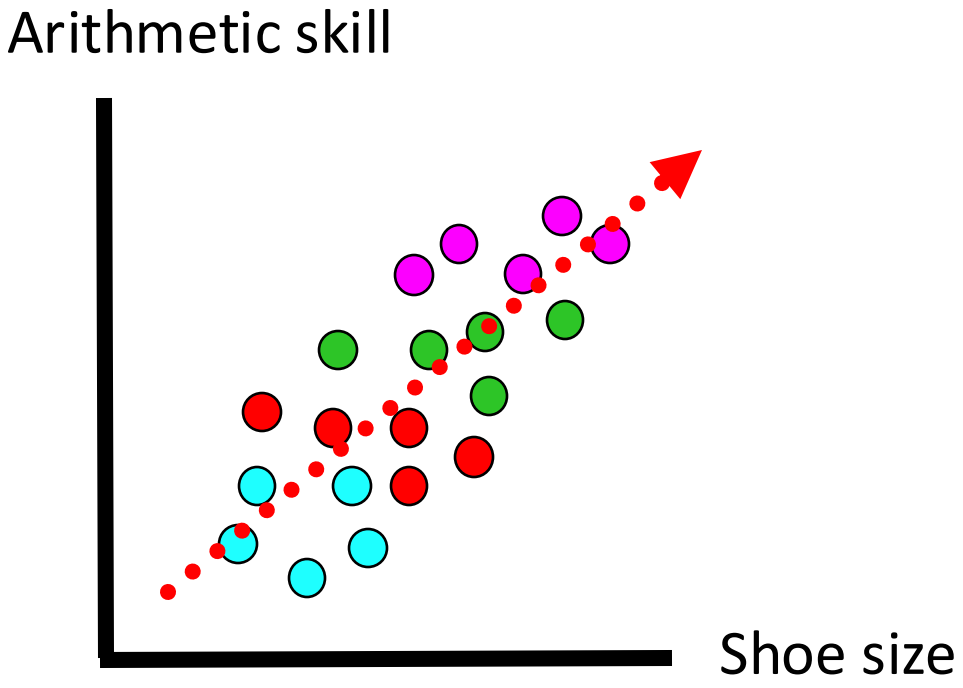


Association – Big 5 Personality Inventory



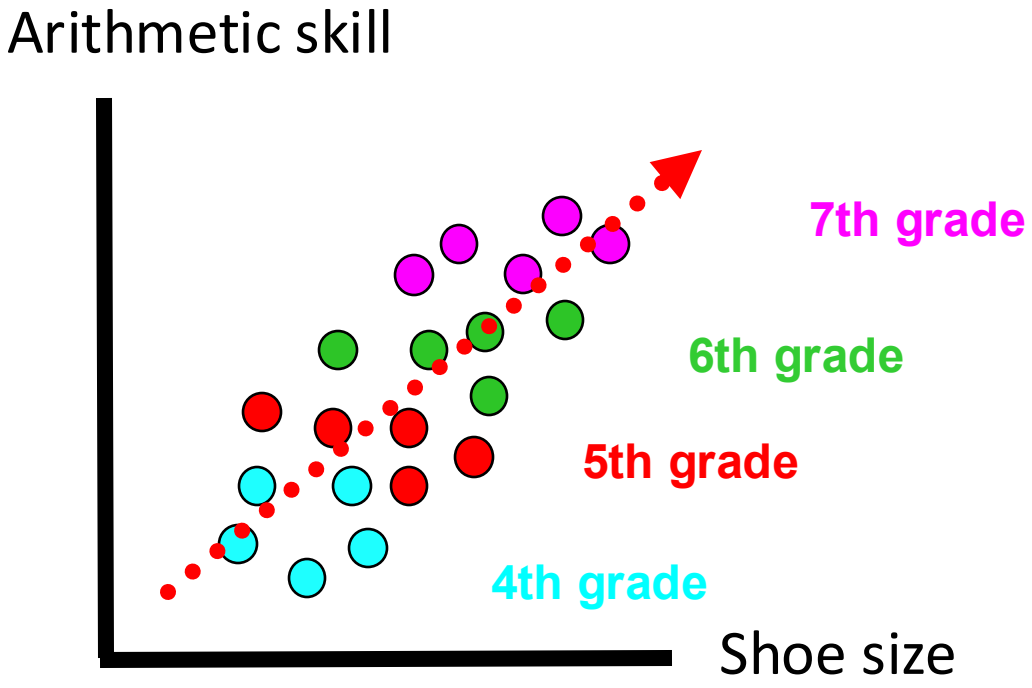
Wrong inference, beware of confounds!

- You can not conclude a causal relationship based on an association
  - More arithmetic skill does not cause larger shoe size
  - Larger shoe size does not cause more arithmetic skill

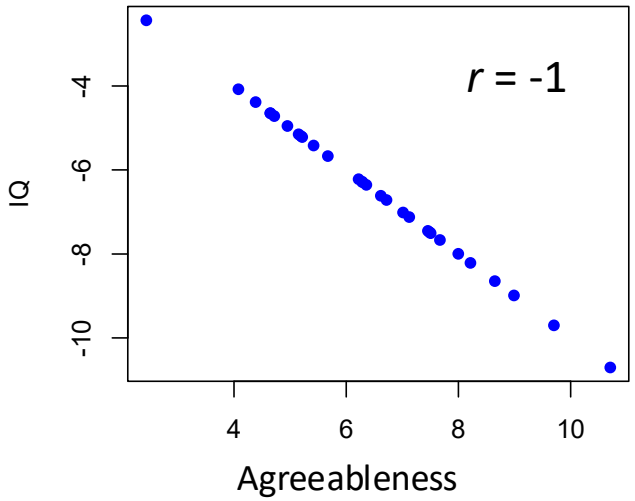


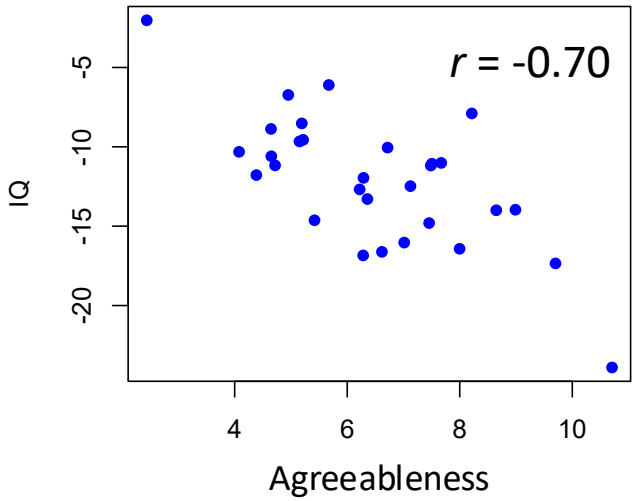
Wrong inference, beware of confounds!

- You can not conclude a causal relationship based on an association
  - More arithmetic skill does not cause larger shoe size
  - Larger shoe size does not cause more arithmetic skill
  - **We only observe an association!**
  - Lecture 8 covers requirements for causal claims

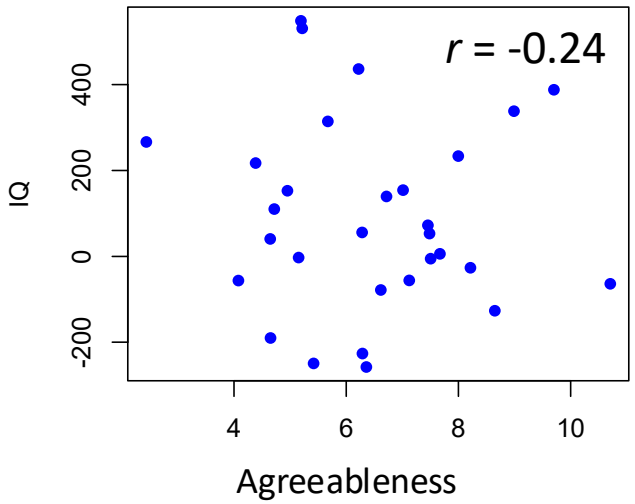


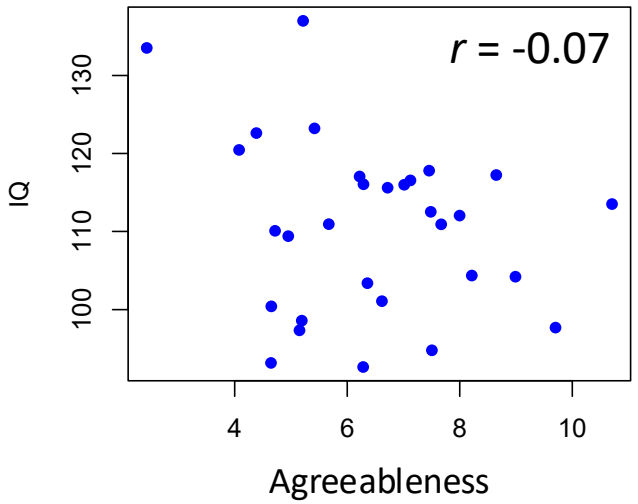
Correlation strength





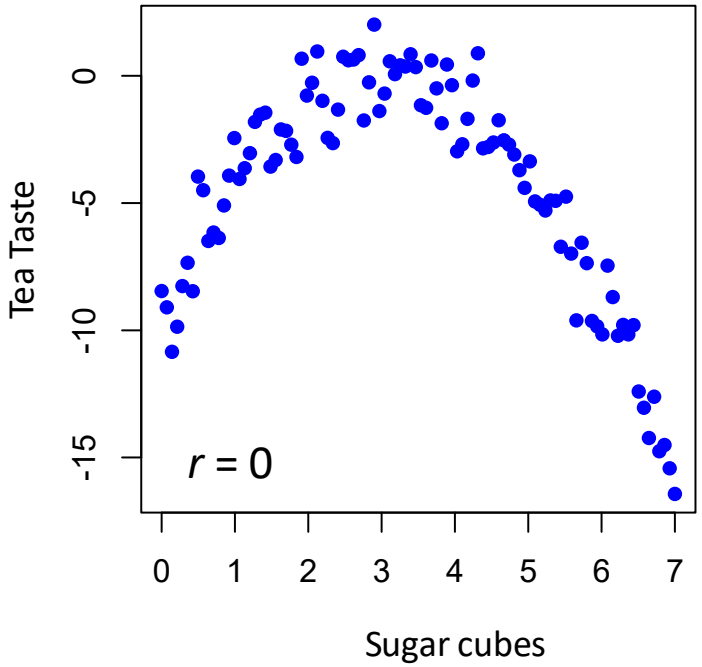
- Strong



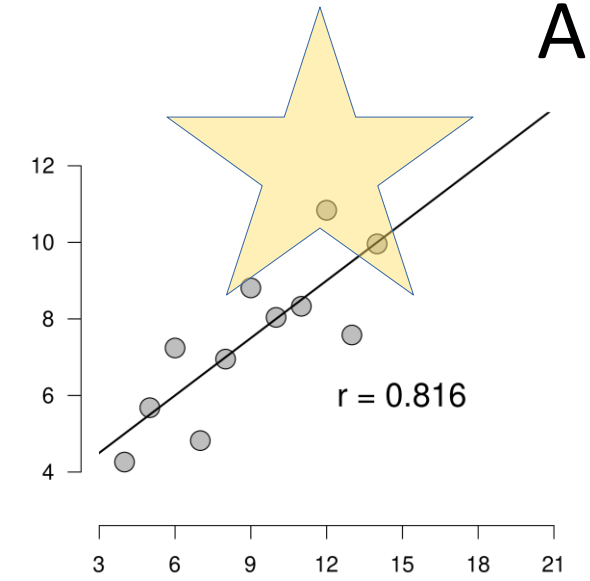


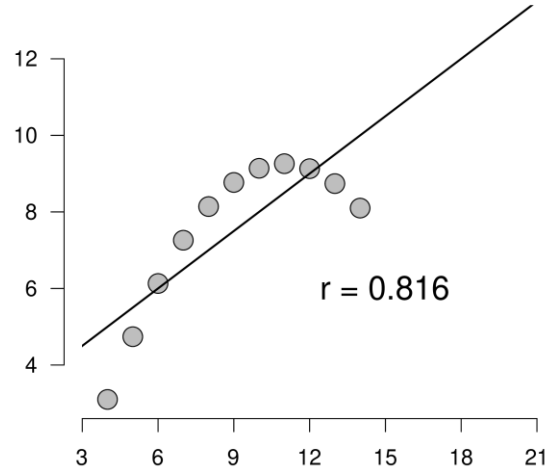
- Weak

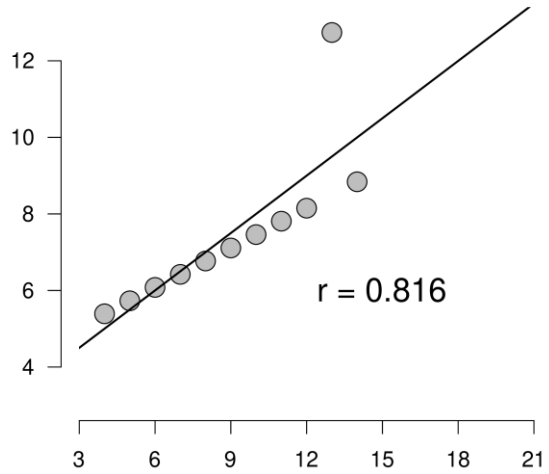
Correlation is a linear association

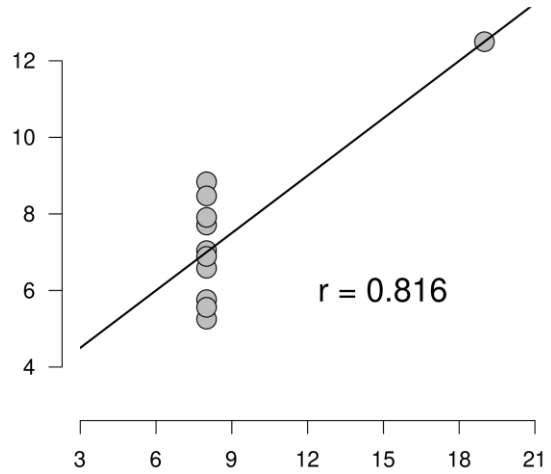


Anscombe's Quartet





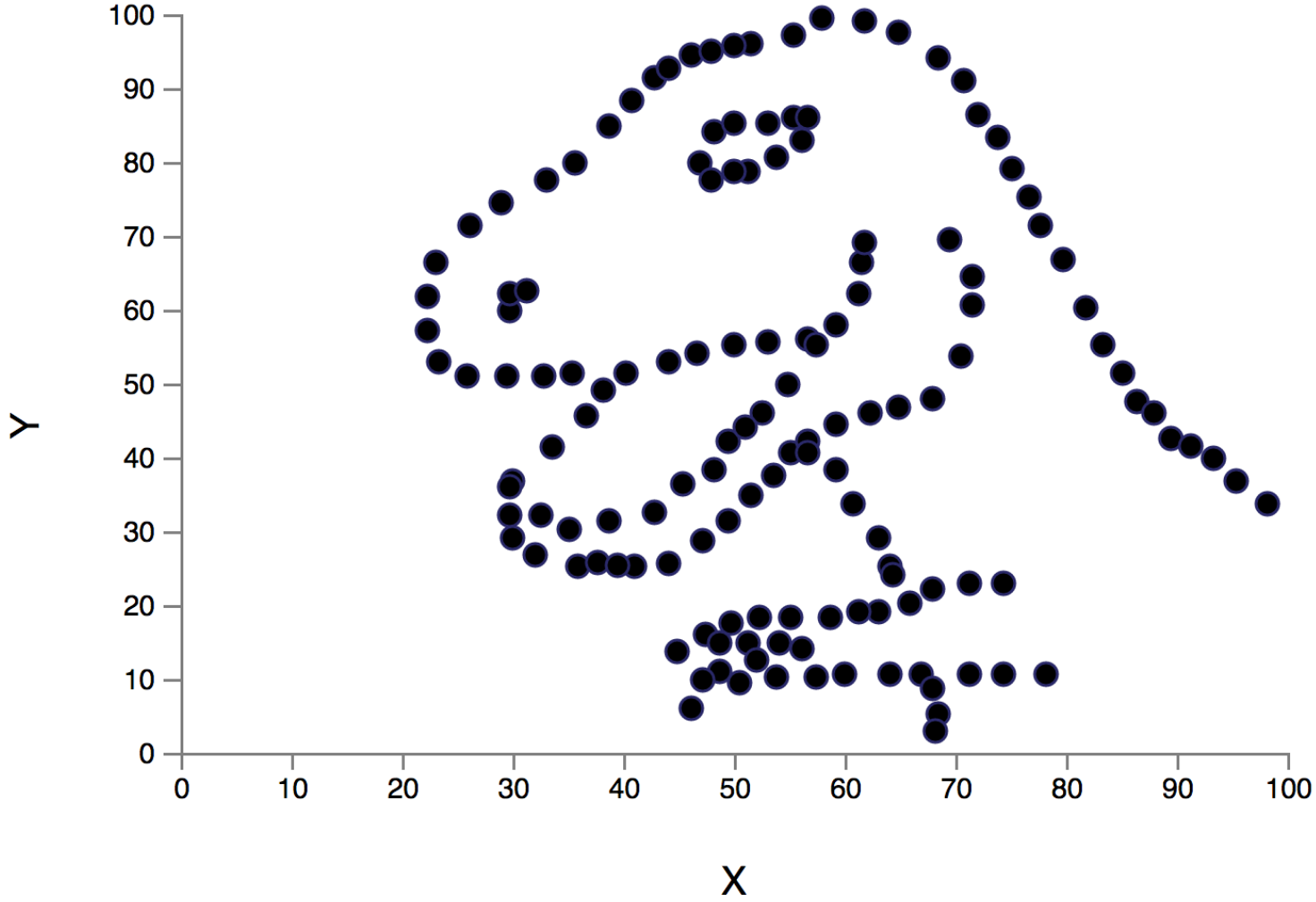




https://en.wikipedia.org/wiki/Anscombe%27s_quartet

Always Plot Your Data!

r = -0.06



Today

1. Variability in the data
   - Standard deviation and variance
   - z-score
   - Quartiles
   - Boxplot
2. Associations
   - Between two categorical variables
   - Between two quantitative variables
3. **Recap**
   - Next time
   - Example exam question

Recap of Today

- The distribution of a variable is determined by the mean, shape (skewness), and variability (standard deviation)
- We can standardize variables so we can compare them more easily
- For analyzing the association between two *quantitative* variables, we use the *correlation coefficient*
- For analyzing the association between two *categorical* variables, we use the *contingency table*

Next time

- Today we talked about variability, proportion etc
- These are concepts from Probability

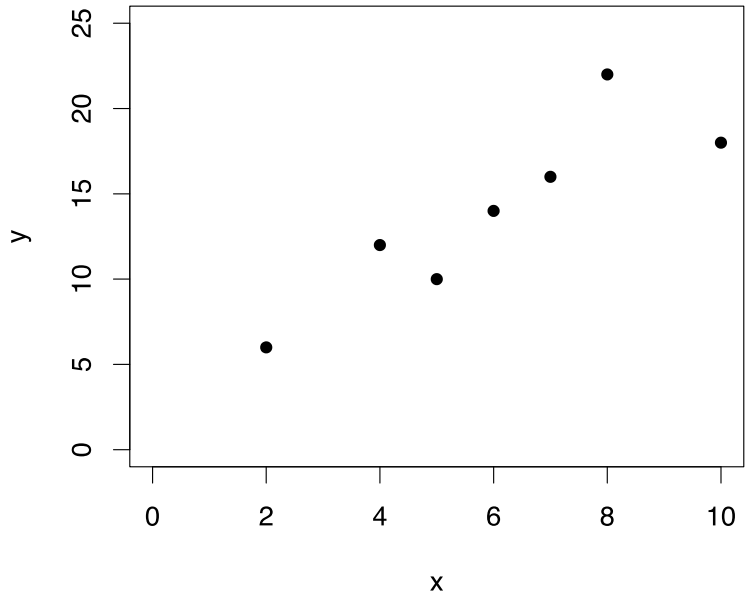
- To draw inferences, we also need to consider probability

- Next week: reasoning with probability
  - Basic probability rules
  - Conditional probability

Example exam question

- Using the data and the scatter plot below, compute the correlation between x and y

x	y
2	6
4	12
5	10
6	14
7	16
8	22
10	18



a) 0.23

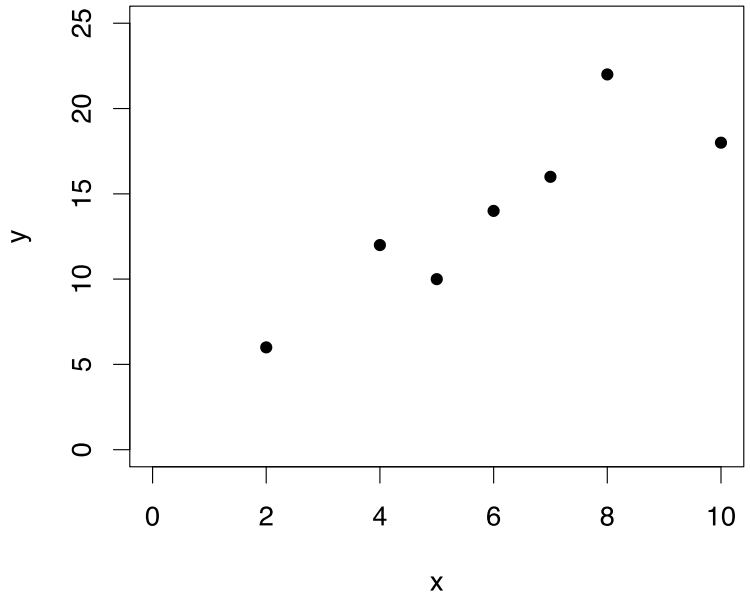
b) 0.76

c) 0.88

Example exam question

- Using the data and the scatter plot below, compute the correlation between x and y

x	y
2	6
4	12
5	10
6	14
7	16
8	22
10	18



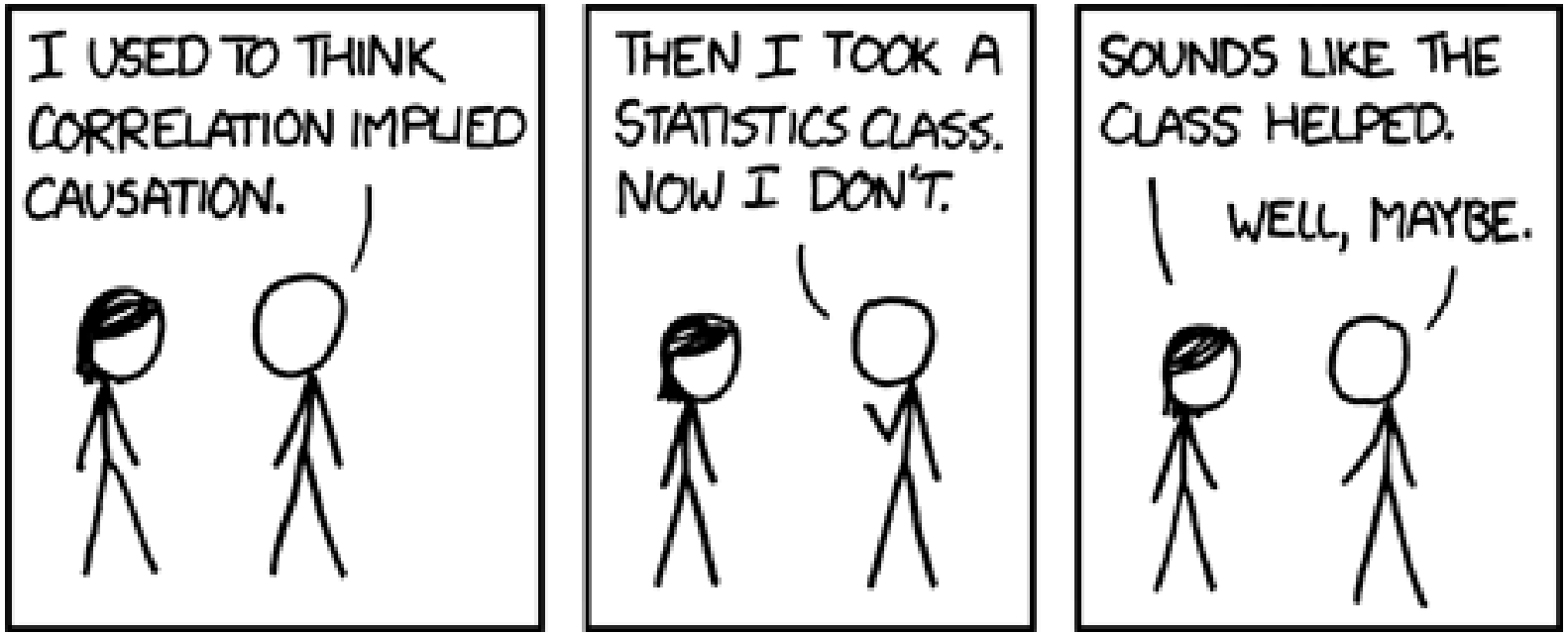
~~a) 0.23~~

~~b) 0.76~~

c) 0.88

Questions?

Thank you for your attention



https://www.xkcd.com/552/

Bonus Page

Spurious Correlations – Ridiculous things that correlate due to chance or confounding variables (or do they??)

http://www.tylervigen.com/spurious-correlations