Reading statistical texts

Lukáš Lehotský & Petr Ocelík

Reliability and validity

Validity

- Analysis leads to **true** conclusions

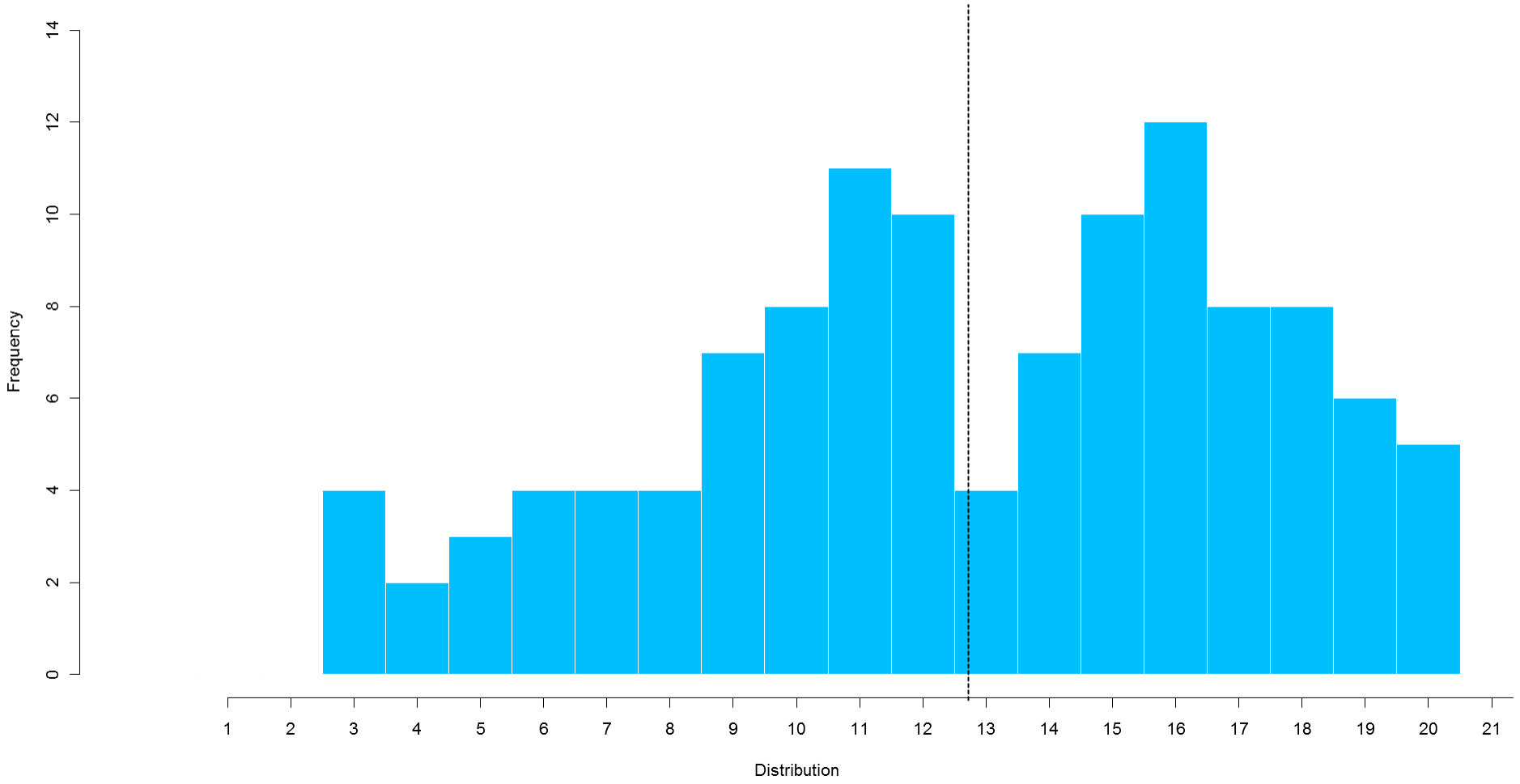
- Internal validity
  - Construct
  - Concept
- External/ecological validity

Reliability

- Repeating **research steps yields same outcomes**
- Replicable research
  - Reinhart, C. M., & Rogoff, K. S. (2010). *Growth in a Time of Debt* (Working Paper Series). http://doi.org/10.3386/w15639
  - Herndon, T., Ash, M., & Pollin, R. (2014). Does high public debt consistently stifle economic growth? A critique of Reinhart and Rogoff. *Cambridge Journal of Economics*, *38*(2), 257. http://doi.org/10.1093/cje/bet075

Beyond description

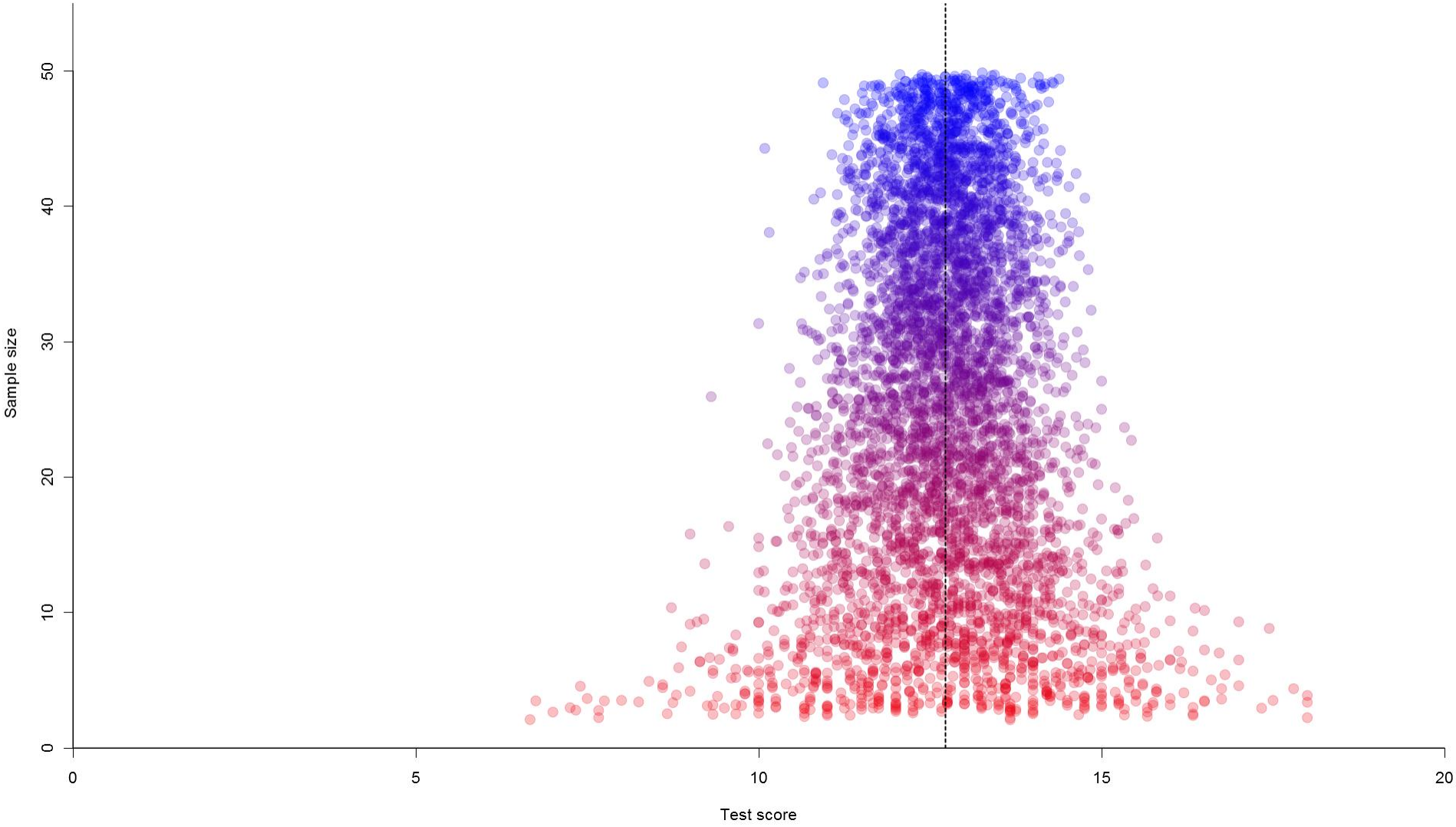
Real data


Frequency
0
2
4
6
8
10
12
14
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
Distribution

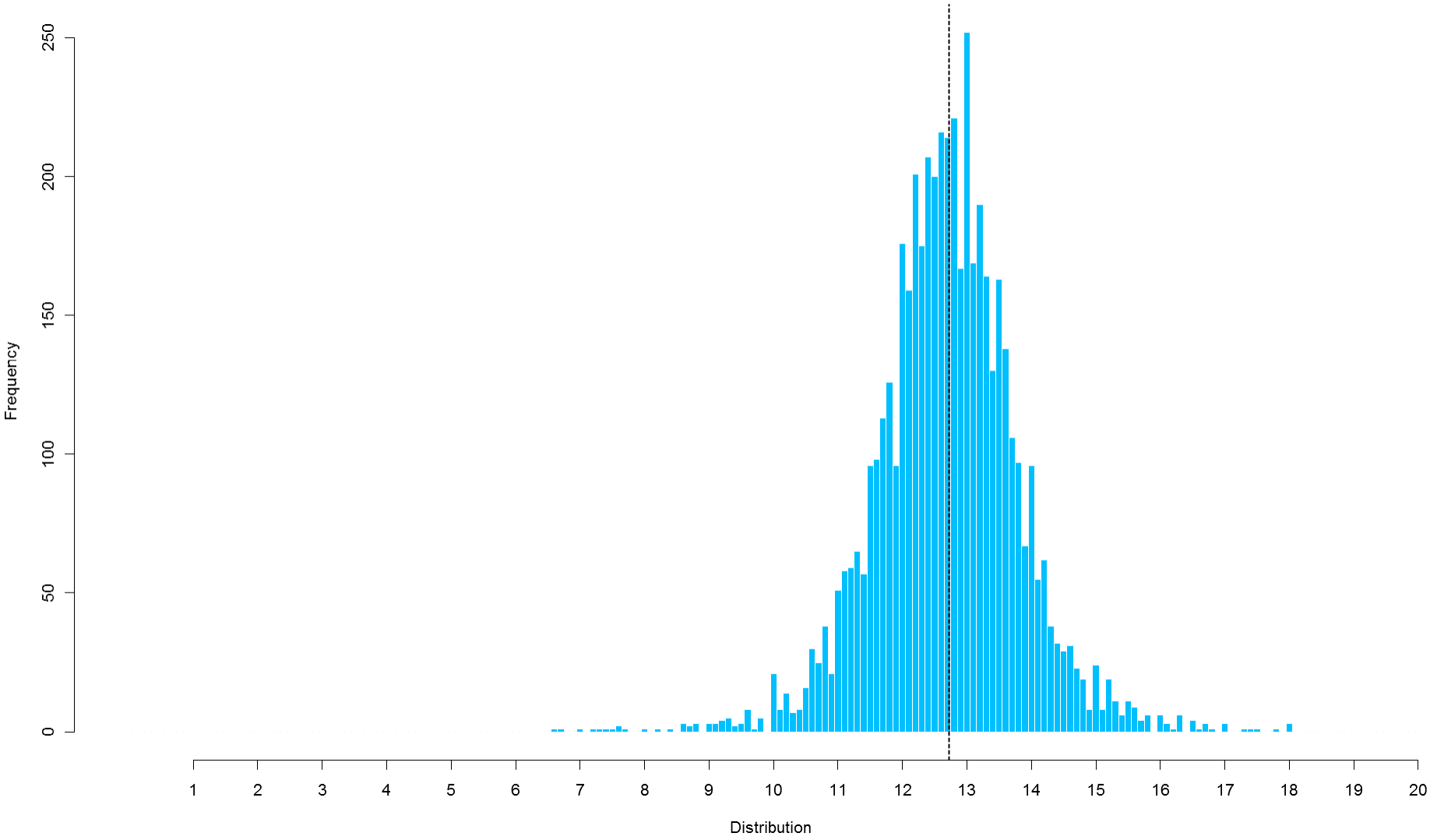
Description vs. inference

- What if collecting real data not feasible?
  - Attitude of Czech population to building NPP
  - Influence of distance from NPP to NPP acceptance
  - …
- Rely on sampling
  - How can we be sure we can generalize from sample to population?
- Central limit theorem
  - If we take a sample from population which **large enough**, it **approximates the mean of population**

Central limit theorem



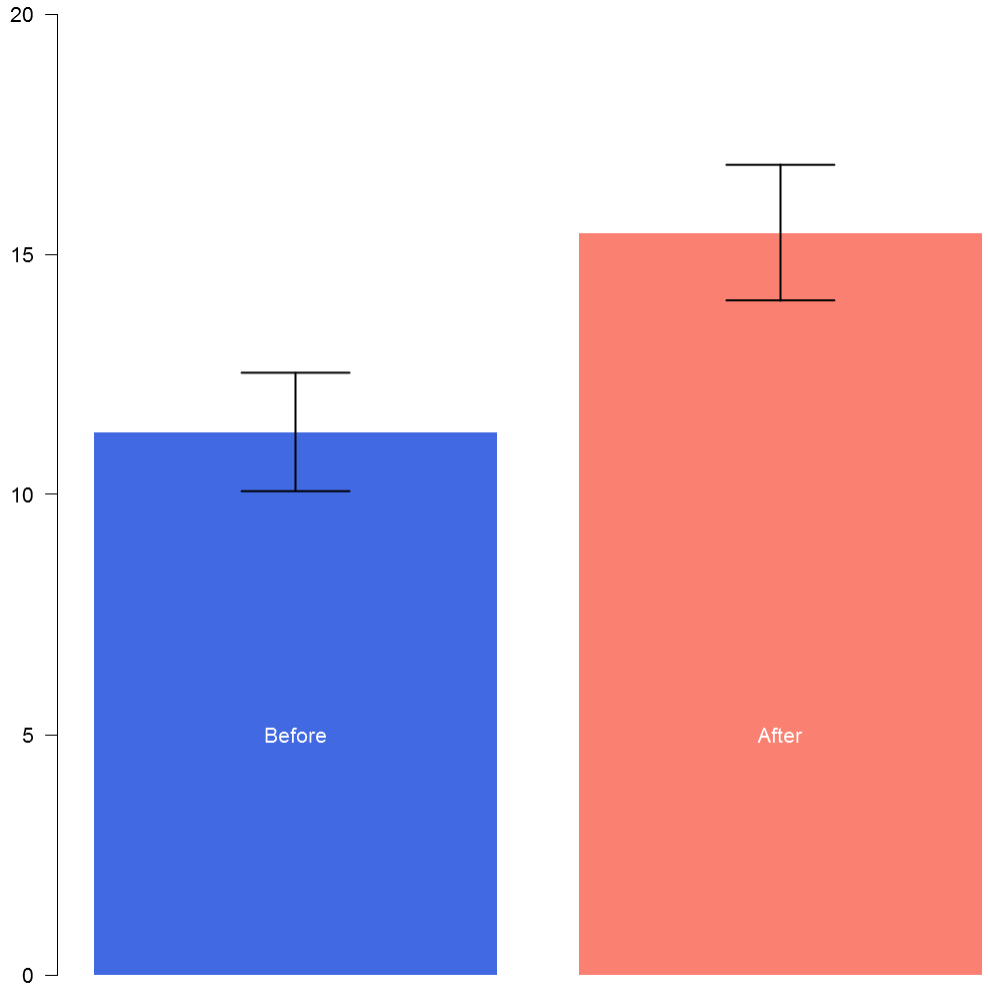
Central limit theorem



Standard error of the mean

- Don’t know the mean, just approximate
- Standard error of the mean
  - Approximation **how close** our sample mean $\bar{x}$ is to the true population mean μ
- Ratio of standard deviation of the sample and number of sample observations
- $SEM = \frac{s}{\sqrt{n}}$
- More observations → smaller SEM
- s shows dispersion of sample data, SEM describes quality of the sample

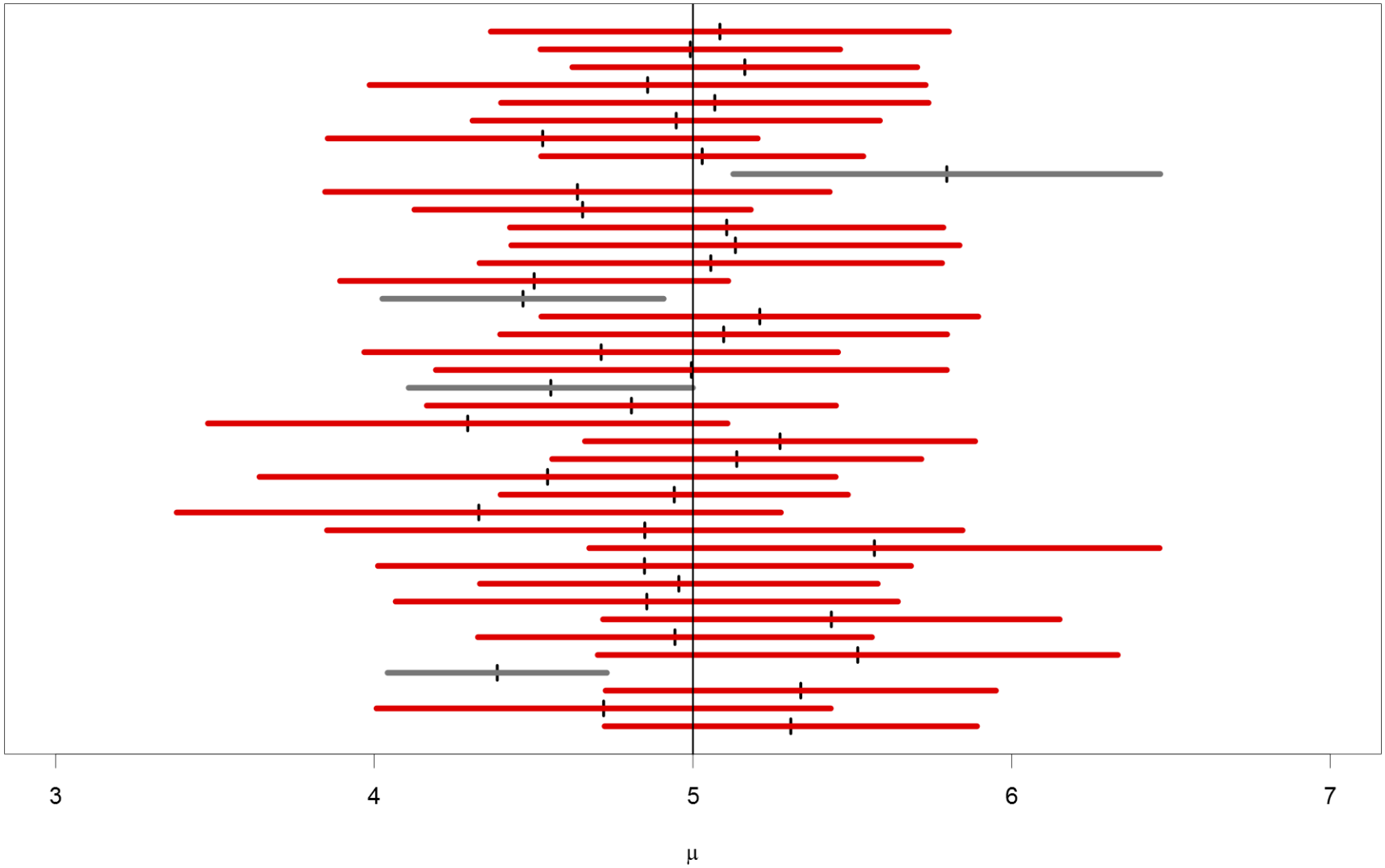
Standard error of the mean



Confidence intervals

Confidence interval

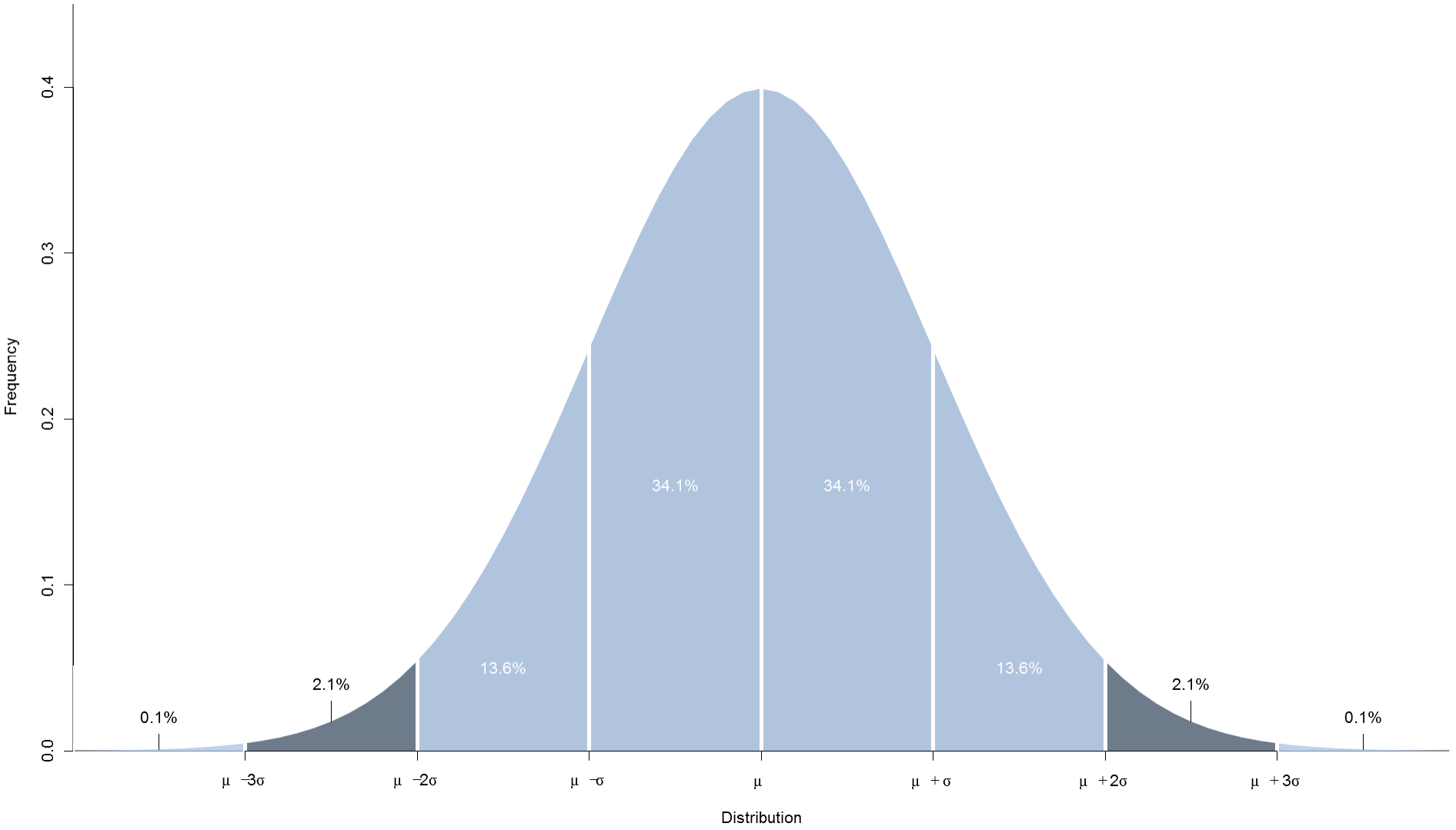
- Some dots further, some closer to the real mean of population
- Confidence interval over **parameter**
  - Interval of confidence that random samples will contain the real mean of the population
  - E.g. 95% confidence interval for μ – in 95% of cases, mean will lie between lower and upper bound of the interval
  - E.g. 99% confidence interval for σ – in 99% of cases, population standard deviation lies within the interval


3
4
5
6
7
μ

Confidence interval

- How likely is our sample mean $\bar{x}$ equal to the population mean μ?
- But we don't know the real mean!
- However, we assume
  - Normal distribution of data
  - Normal distribution of data samples
- We can calculate confidence interval of the sample mean $\bar{x}$ thanks to knowledge of the SEM
- 95% confidence interval "industry standard"

Confidence interval



Null hypothesis

Null hypothesis

- We can’t prove any hypothesis based on the sample
- We may only prove there is little chance the relation is random
- Null hypothesis – observed relation is **result of random variation**
- Thus, we aim to prove that it is highly unlikely the relationship between variables is generated by chance – reject the null hypothesis

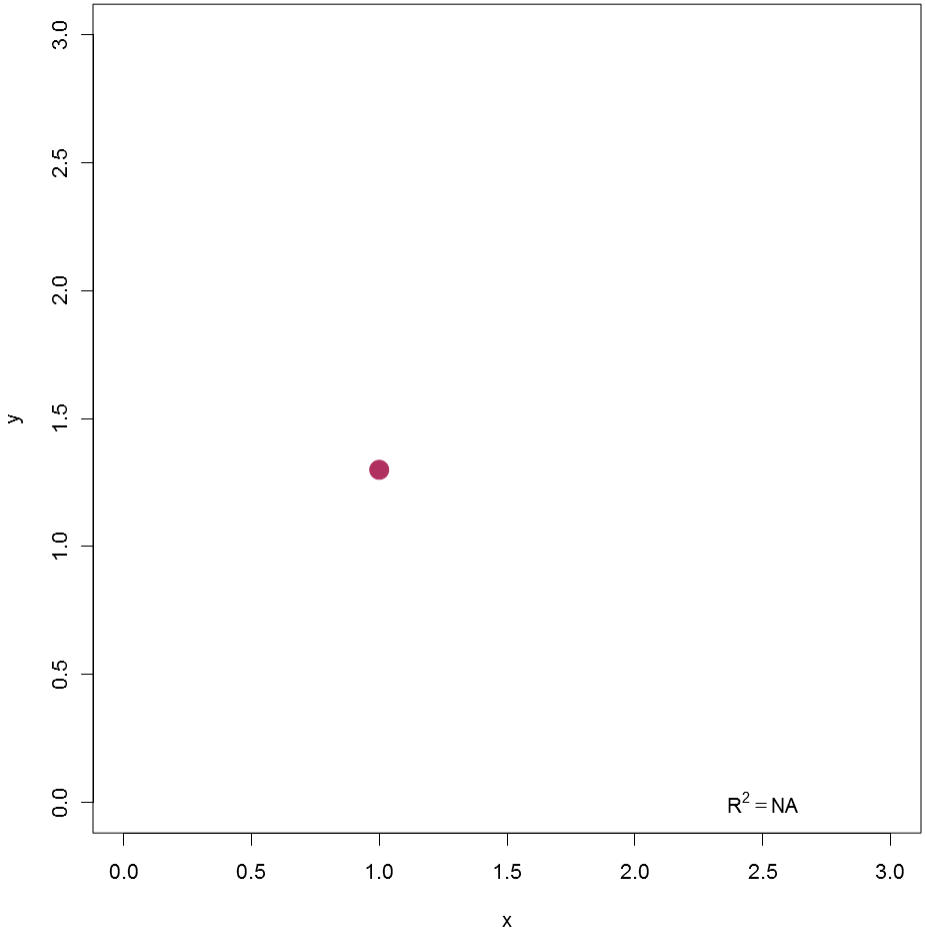
Statistical significance

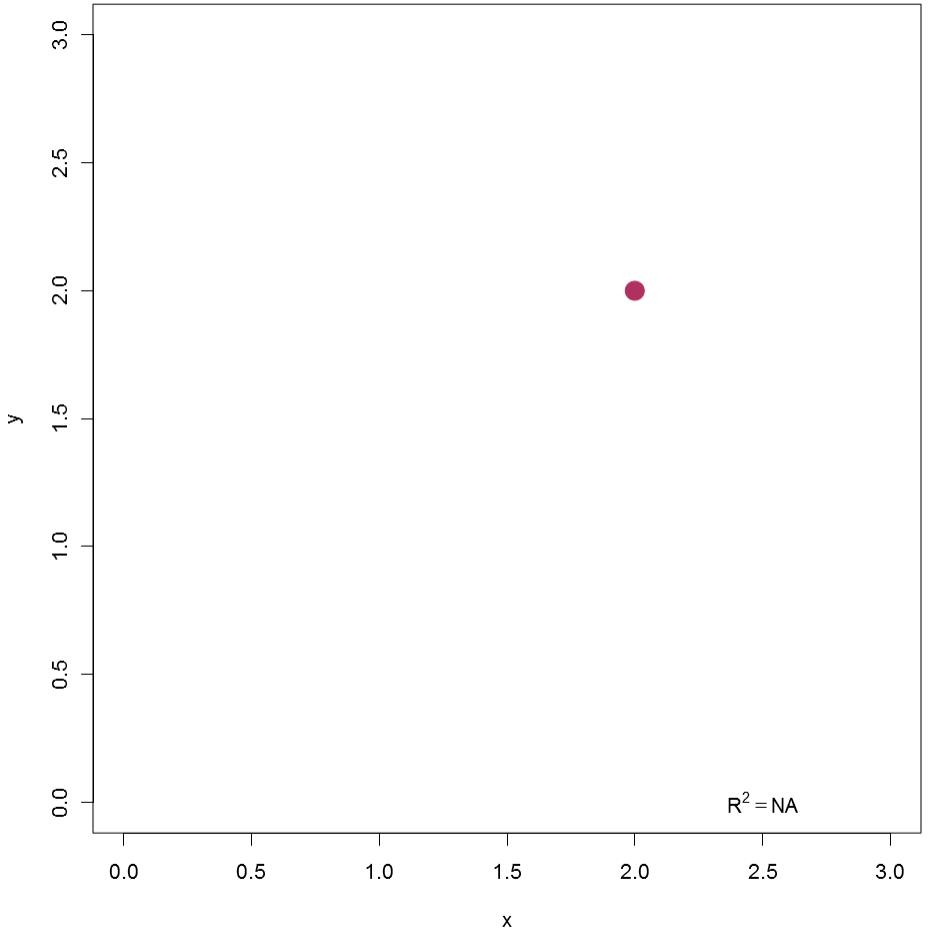
Statistical significance

- Probability that the sample comes from the population where the effect happens by chance
- E.g. probability the null hypothesis is valid
- Denoted as p
  - $p \leq 0.05$ – 95% statistical significance (1 in 20)
  - $p \leq 0.01$ – 99% statistical significance (1 in 100)
  - $p \leq 0.001$ – 99.9% statistical significance (1 in 1000)

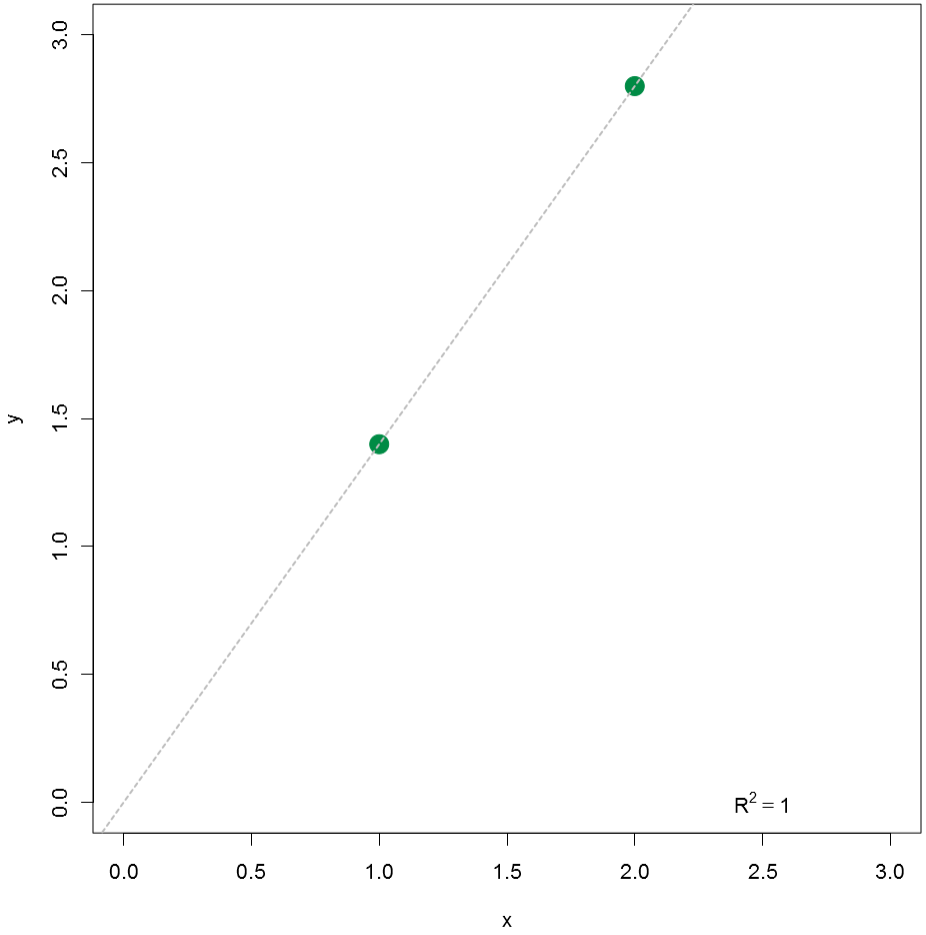
Degrees of freedom

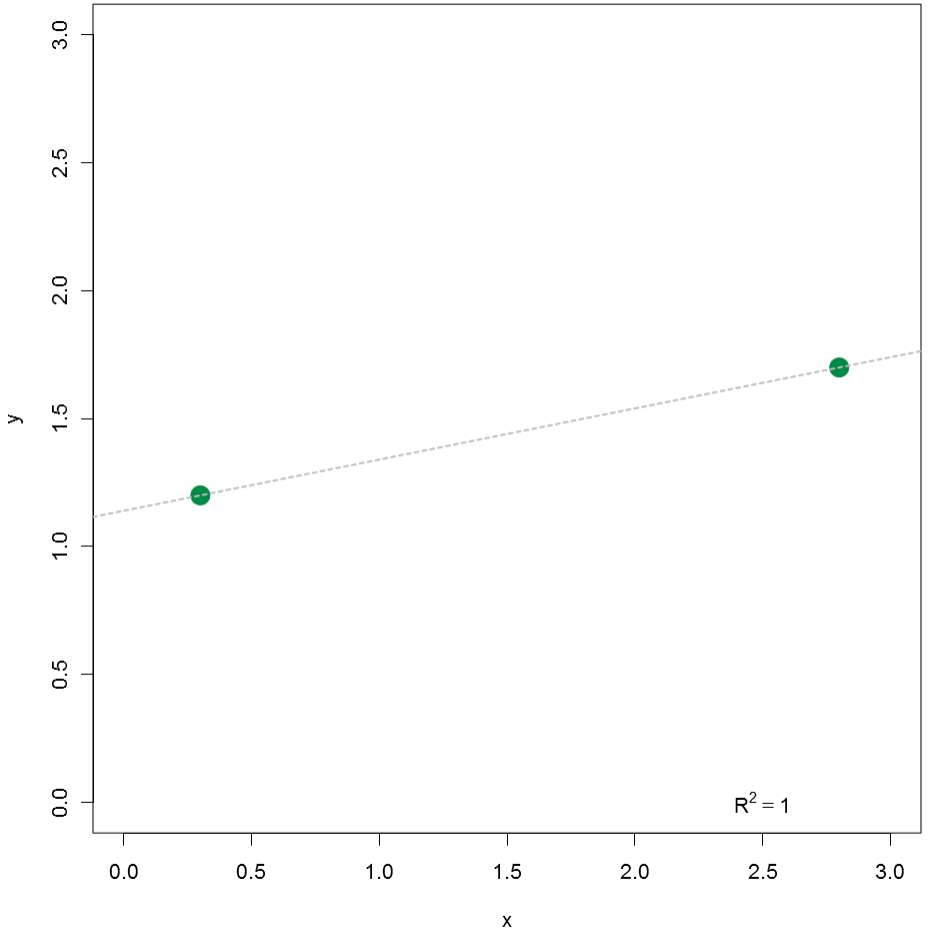
Degrees of freedom



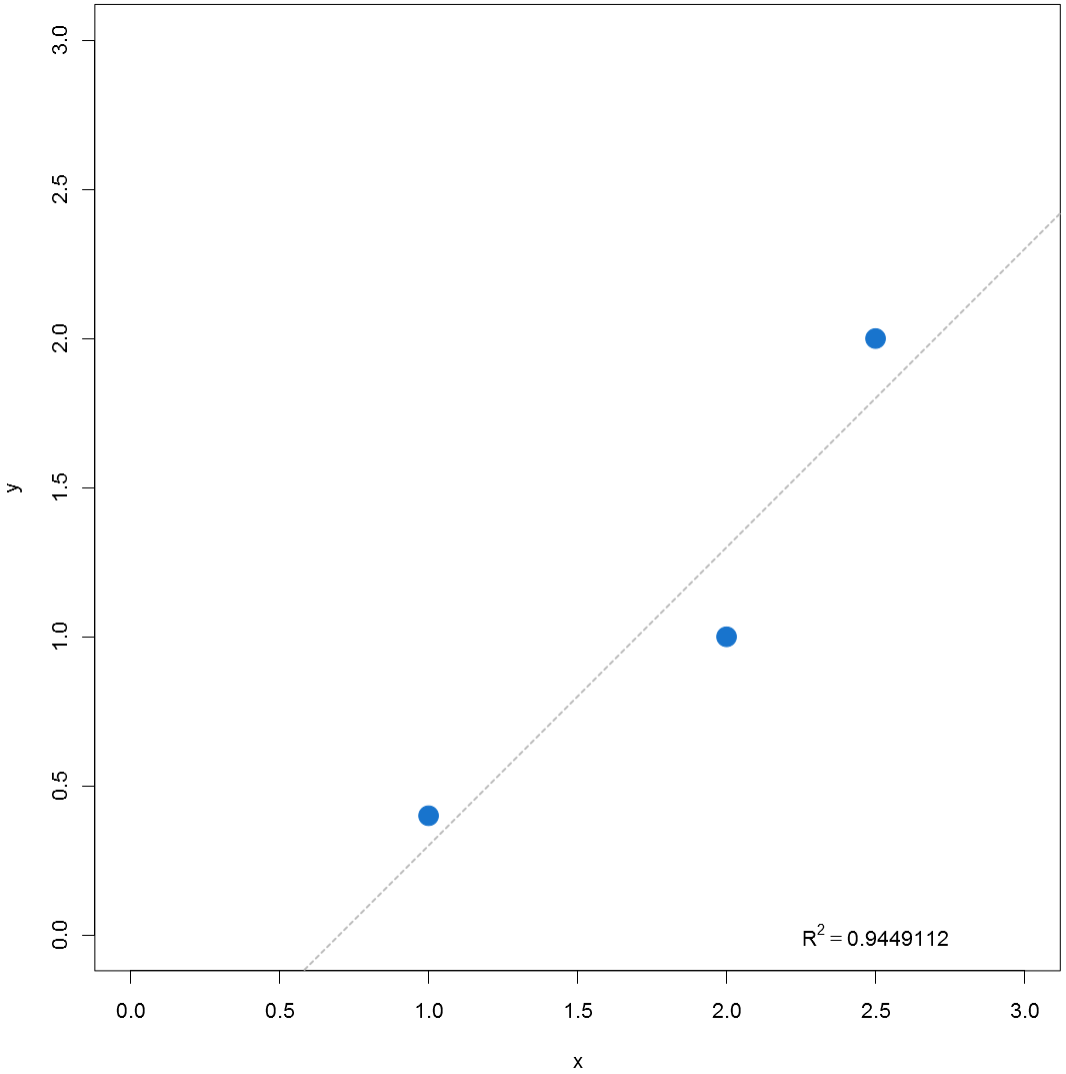


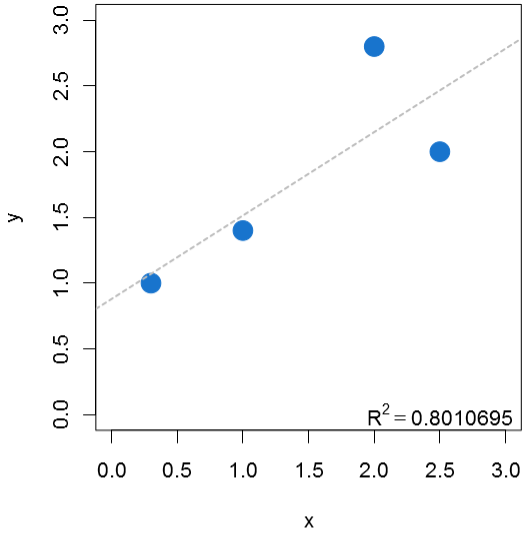
Degrees of freedom

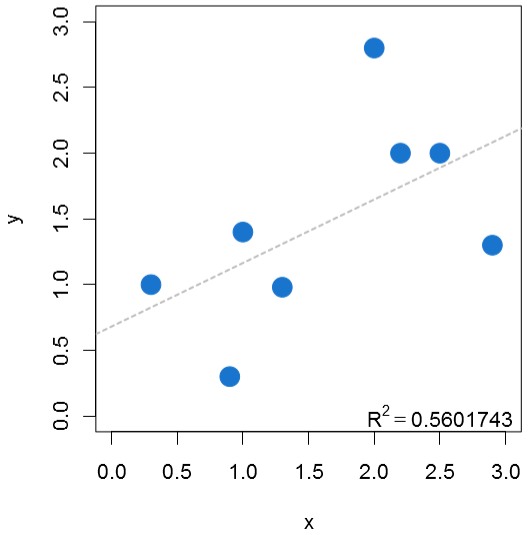




Degrees of freedom







Degrees of freedom

- Same with sum of different numbers
  - $a + b + c + d = 100$
  - If $c = 5$, than $a + b + 5 + d = 100$

- Number of observations $n \in \{a, b, c, d\}$
- Number of variables $k \in \{c\}$
- Degrees of freedom $n - k - 1$

Why so complicated?

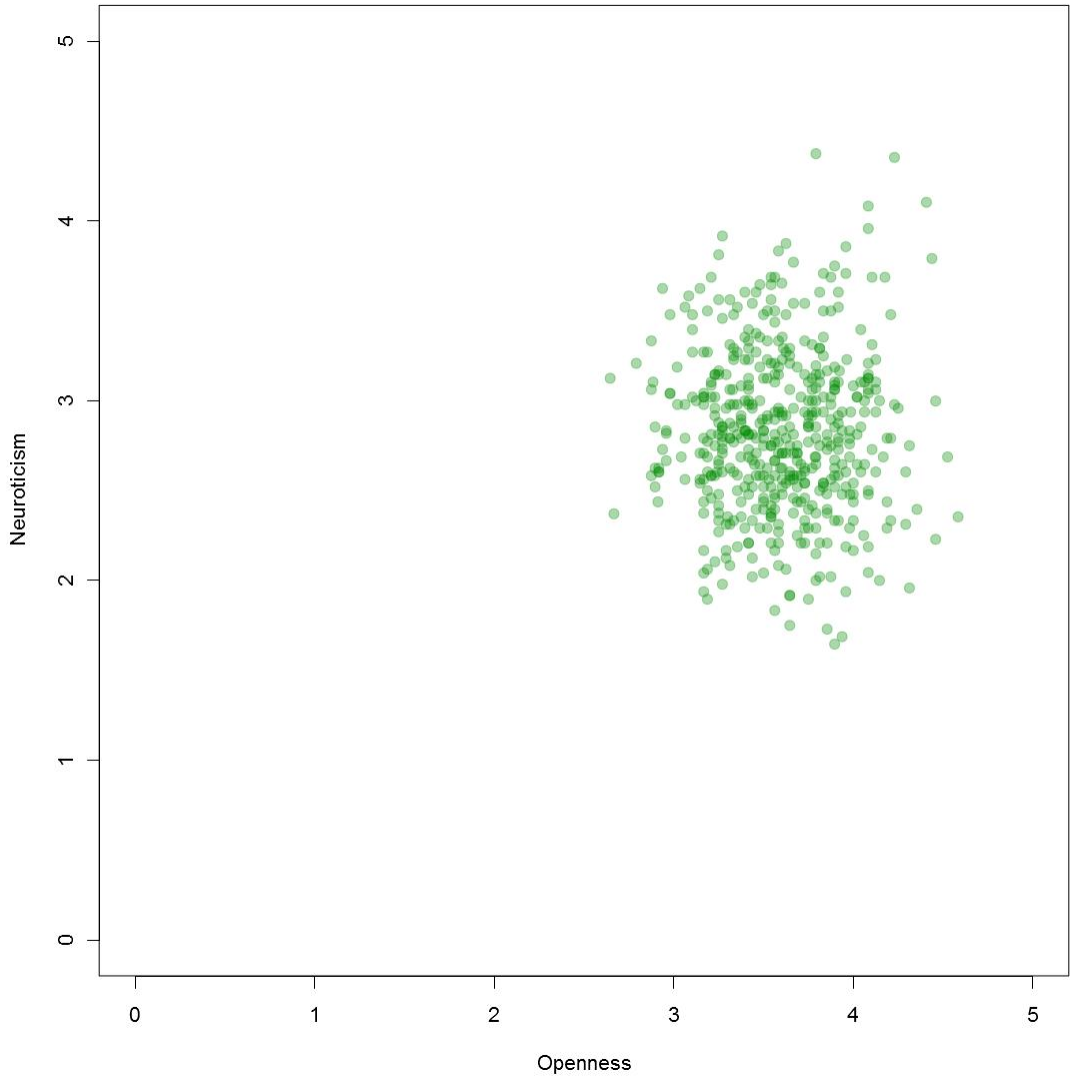
Reliability and validity!

Correlation

Dolan et. al. – “big 5”

- 500 observations – 500 first year psychology students
- Measurement of the Dutch translation of the NEO-PI-R – NEO Personality Inventory
- Big 5 personality traits
  - Agreeableness
  - Neuroticism
  - Conscientiousness
  - Extraversion
  - Openness

Correlation



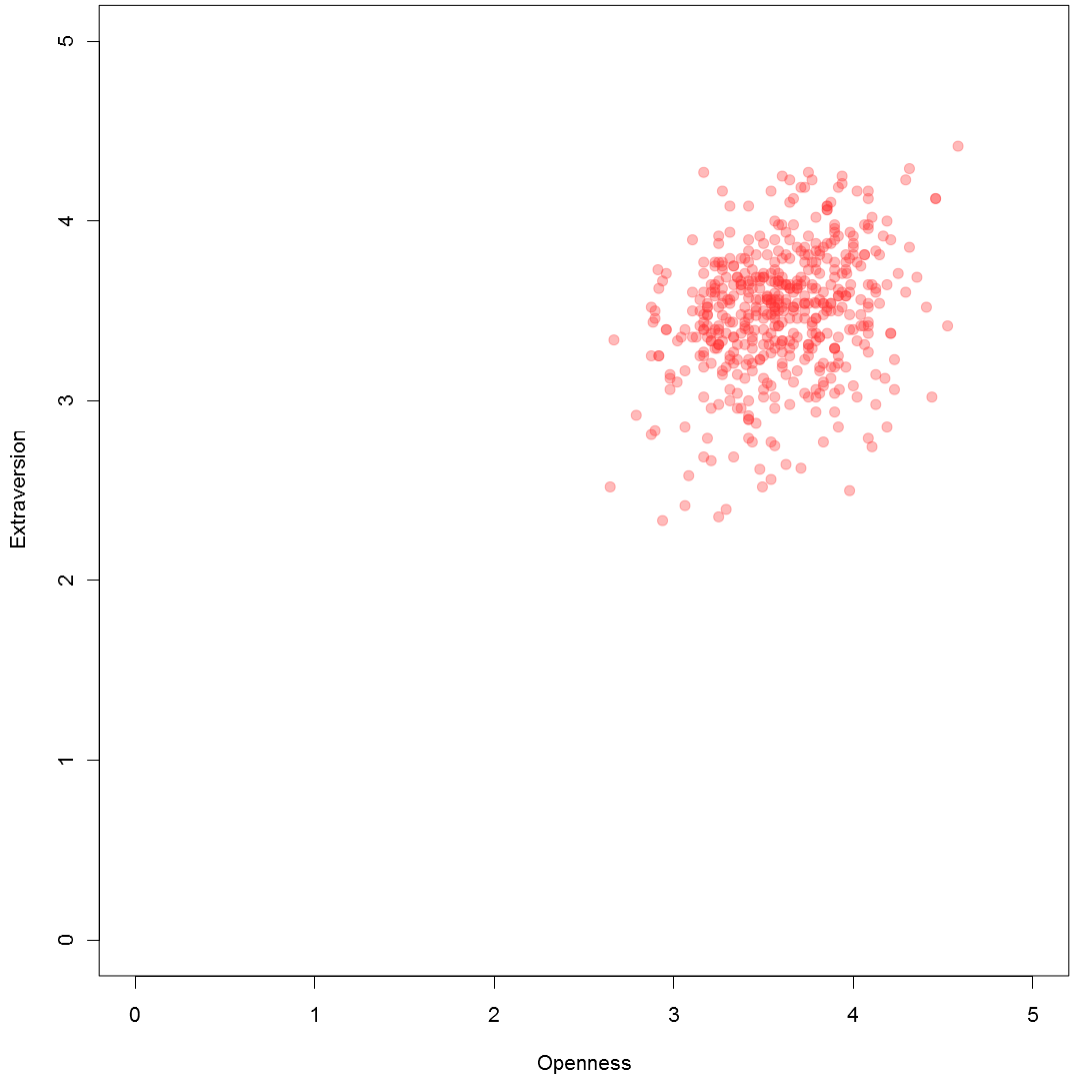
Correlation

Pearson Correlations

		Openness	Neuroticism
Openness	Pearson's r	—	-0.010
	p-value	—	0.817
Neuroticism	Pearson's r		—
	p-value		—

(Dolan – Oort – Stoel – Wicherts, 2009)

Correlation



Correlation

Pearson Correlations

		Extraversion	Openness
Extraversion	Pearson's r	—	0.267
	p-value	—	< .001
Openness	Pearson's r		—
	p-value		—

(Dolan – Oort – Stoel – Wicherts, 2009)

Geographical and sociodemographic differences in attitudes to coal mining.

Variable		Attitude [%]			Value of correlation[1]
		Convinced pro-coal	Reserved	Anti-coal	
Residence	Horní Jiřetín	7	32	61	0.359**
	Janov	19	32	49	
Place attachment	Low	22	34	44	0.262*
	Medium	15	30	55	
	High	0	31	69	
Gender	Males	15	34	51	n.s.
	Females	14	31	55	
Age	<20	0	50	50	n.s.
	20–29	20	13	67	
	30–39	18	27	55	
	40–49	20	30	50	
	50–59	11	44	45	
	60+	10	35	55	
Education	Elementary	10	35	55	n.s.
	Secondary	16	34	50	
	Tertiary	7	29	64	
Employment in coal industry	Yes	25	67	8	0.465**
	No	4	32	64	
Total		14	32	54	

[1]The values of correlation (Pearson's r) are significant at the level ** <0.01 or *<0.05; n.s. means a non-significant correlation.

(Frantál, 2016)

Bivariate correlations between outcome variables and predictor variables (N=248).

		1	2	3	4	5	6	7	8
1	General acceptance	–							
2	Local acceptance	.70***	–						
3	Affect	.45***	.45***	–					
4	Perceived risk	−.46***	−.41***	−.46***	–				
5	Perceived benefit	.64***	.43***	.36***	−.39***	–			
6	Support for renewables	−.09	−.16*	−.18**	.16*	−.08	–		
7	Acceptance of energy transition	−.08	−.16*	−.12	0.09	−.09	.63***	–	
8	House ownership (a)	0.12	.20**	.13*	−.09	−.02	.14*	.16*	–
9	Gender (b)	−.08	0.03	−.18**	0.07	−.18**	0.07	.15*	.23***

(a) House ownership was coded 1=yes, 2=no.
(b) Gender was coded 1=male, 2=female.
* p<.05, ** p<.01, *** p<.001

(Lienert - Suetterlin - Siegrist, 2015)

Linear regression

Linear regression

- Regression output
- Dolan et. al. – “big 5”
- Let’s test if “openness” → “agreeableness”

Linear regression output

Model Summary

Model	R	R^2	Adjusted R^2	RMSE
1	0.159	0.025	0.023	0.346

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	1.555	1	1.555	12.95	< .001
	Residual	59.777	498	0.12		
	Total	61.332	499			

Coefficients

Model	Agreeab.	Unstand.	Standard Error	Stand.	t	p	CI 2.5%	CI 97.5%
1	intercept	2.845	0.165		17.291	< .001	2.522	3.169
	Openness	0.164	0.046	0.159	3.599	< .001	0.075	0.254

(Dolan – Oort – Stoel – Wicherts, 2009)

Model fit

Model Summary

Model	R	R^2	Adjusted R^2	RMSE
1	0.159	0.025	0.023	0.346

ANOVA

Model		**Sum of Squares**	df	Mean Square	F	p
1	Regression	**1.555**	1	1.555	12.95	< .001
	Residual	**59.777**	498	0.12		
	Total	**61.332**	499			

(Dolan – Oort – Stoel – Wicherts, 2009)

Model fit

- R^2 - sum of squares of **explained** variation to **total** variation
- $R^2 = \frac{SS_{model}}{SS_{total}} = \frac{SS_{regression}}{SS_{total}}$
- From R^2, we may get R – comparable to Pearson's rho – **correlation** between indep. and dep. variable
- R^2 explains how much of the variance of dependent variable **can be explained** by variance of independent variable

Model fit

Model Summary

Model	R	R^2	Adjusted R^2	RMSE
1	0.159	**0.025**	0.023	0.346

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	**1.555**	1	1.555	12.95	< .001
	Residual	59.777	498	0.12		
	Total	**61.332**	499			

$$R^2 = \frac{SS_{regression}}{SS_{total}}$$

(Dolan – Oort – Stoel – Wicherts, 2009)

Model fit

- F-test
- $F = \frac{MSS_{model}}{MSS_{residual}}$
- Mean sum of squares of the model vs. mean sum of squares of residuals
- F explains the average increase of the prediction of the model compared to average model error
- F tells us if regression is of **any** use - if we can reject null hypothesis at all

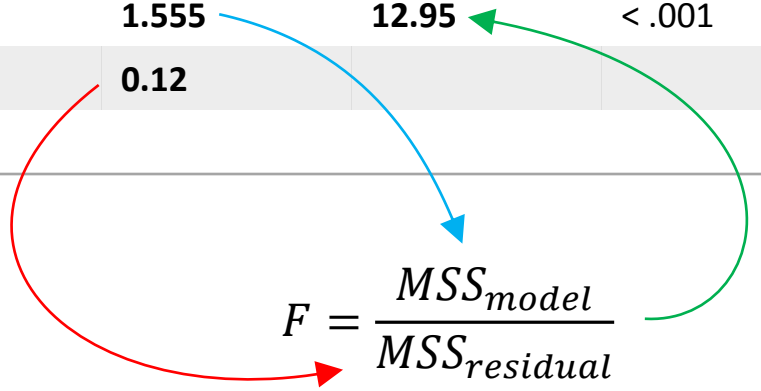
Model fit

Model Summary

Model	R	R^2	Adjusted R^2	RMSE
1	0.159	0.025	0.023	0.346

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	1.555	1	**1.555**	**12.95**	< .001
	Residual	59.777	498	**0.12**		
	Total	61.332	499			



(Dolan – Oort – Stoel – Wicherts, 2009)

Model fit

Model Summary

Model	R	R^2	Adjusted R^2	RMSE
1	0.159	0.025	0.023	0.346

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	1.555	1	1.555	12.95	**< .001**
	Residual	59.777	498	0.12		
	Total	61.332	499			

(Dolan – Oort – Stoel – Wicherts, 2009)

Model fit

Model Summary

Model	R	R^2	Adjusted R^2	**RMSE**
1	0.159	0.025	0.023	**0.346**

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	1.555	1	1.555	12.95	< .001
	Residual	59.777	498	0.12		
	Total	61.332	499			

(Dolan – Oort – Stoel – Wicherts, 2009)

Model fit

- RMSE
- Mean square error of residuals
  - **Mean error of each observation from the model** – average distance of observations from the model
- Useful to understand the model fit – higher the RMSE, lower fit the model has

Model fit

Model Summary

Model	R	R^2	**Adjusted R^2**	RMSE
1	0.159	0.025	**0.023**	0.346

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	1.555	1	1.555	12.95	< .001
	Residual	59.777	498	0.12		
	Total	61.332	499			

(Dolan – Oort – Stoel – Wicherts, 2009)

Adjusted R^2

- Adding variables to the model might help with explanation
- R^2 increases with more variables – more significant variables may occur to explain the variance of dependent variable

# of observations	# of predictors	R^2
10	4	0.7
10	5	0.71
10	6	0.73
10	7	0.79

Adjusted R^2

- R^2 asumes each independent variable has effect on the dependent variable
- $Adj.R^2$ explains variation by independent variables that actually affect the dependent variable
- $Adj.R^2$ penalizes adding independent variables not explaining the variation of dependent variable

# of obs.	# of predictors	R^2	df	Adj. R^2
10	4	0.7	5	0.46
10	5	0.71	4	0.3475
10	6	0.73	3	0.19
10	7	0.79	2	0.055

Model fit

Model Summary

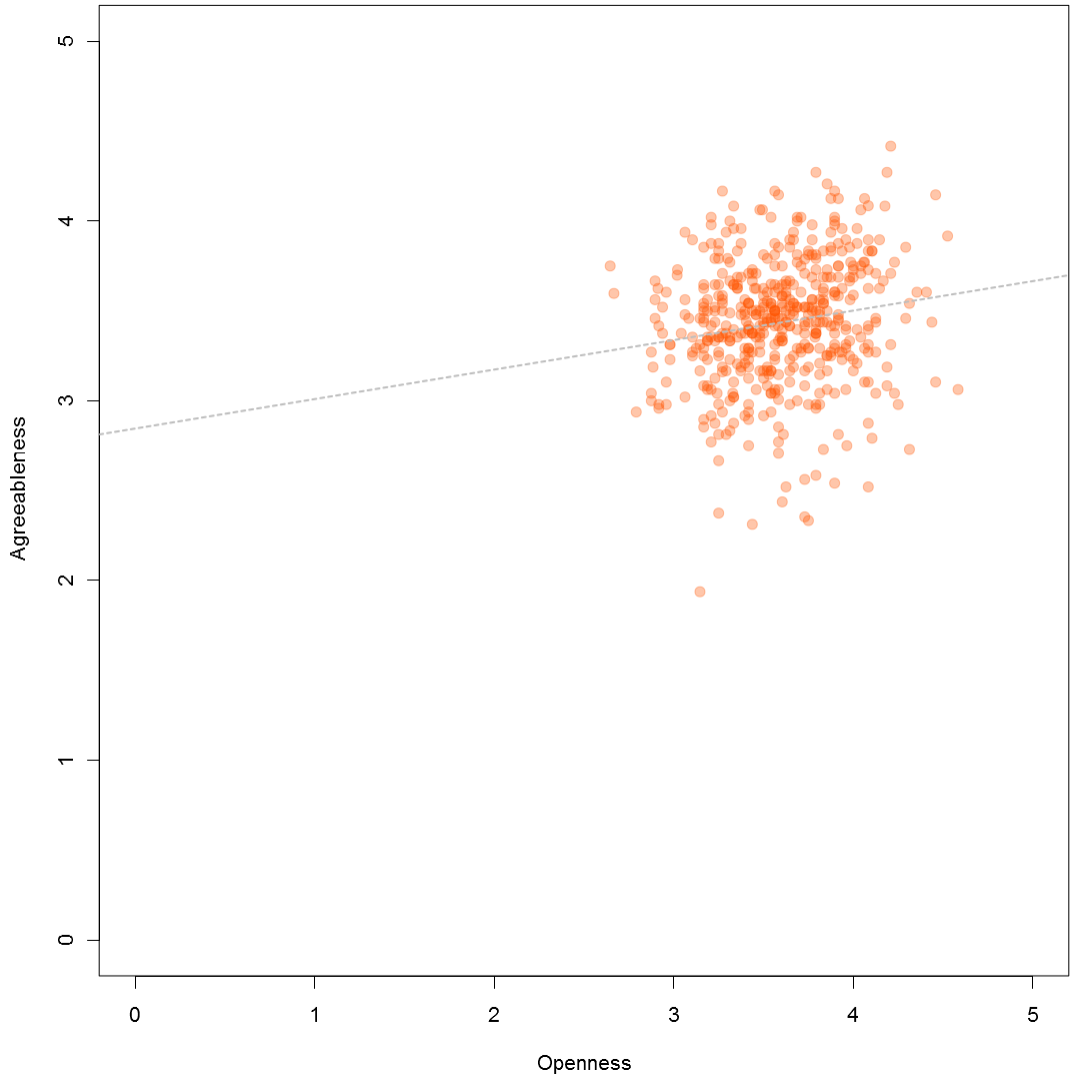
Model	R	R^2	Adjusted R^2	RMSE
1	0.159	0.025	**0.023**	0.346

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	1.555	1	1.555	12.95	< .001
	Residual	59.777	498	0.12		
	Total	61.332	**499**			

(Dolan – Oort – Stoel – Wicherts, 2009)

Model fit



Linear regression

- Fitting a straight line the model
- Line of best fit – ordinary least squares (OLS)
- $y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon$

- Test if “openness” → “agreeableness”

Regression line

Coefficients

Model	***Agreeab.***	Unstand.	Standard Error	Stand.	t	p	CI 2.5%	CI 97.5%
1	intercept	2.845	0.165		17.291	< .001	2.522	3.169
	Openness	0.164	0.046	0.159	3.599	< .001	0.075	0.254

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon$$

Regression line

Coefficients

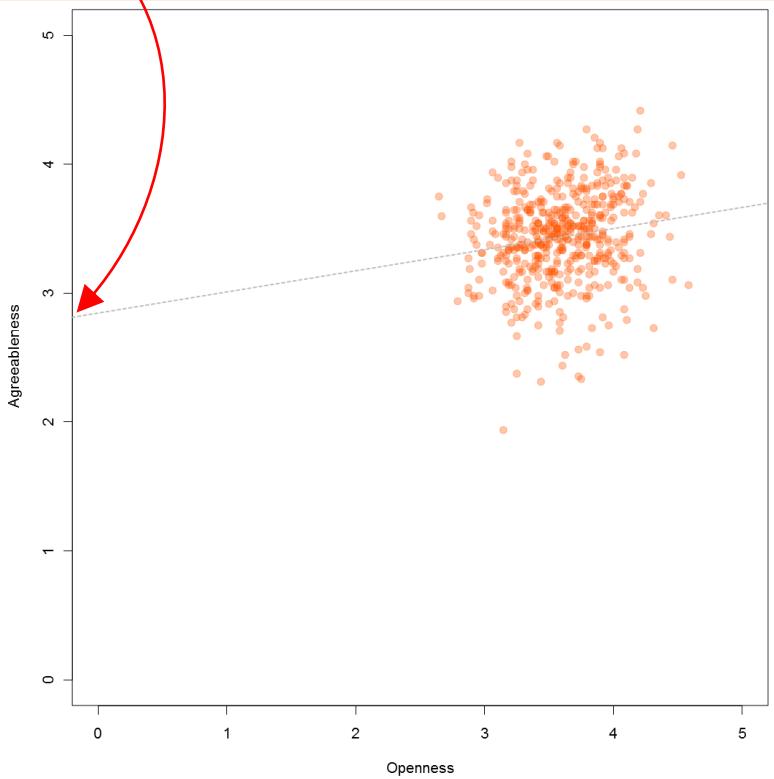
Model	*Agreeab.*	Unstand.	Standard Error	Stand.	t	p	CI 2.5%	CI 97.5%
1	intercept	**2.845**	0.165		17.291	< .001	2.522	3.169
	Openness	**0.164**	0.046	0.159	3.599	< .001	0.075	0.254

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon$$

Regression line

Coefficients

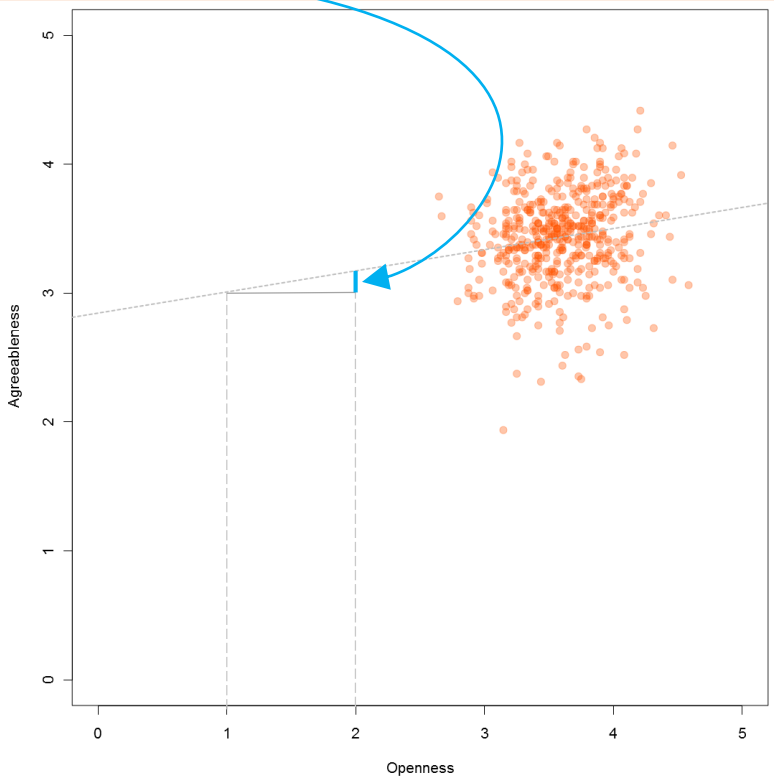
Model	*Agreeab.*	Unstand.	Standard Error	Stand.	t	p	CI 2.5%	CI 97.5%
1	**intercept**	**2.845**	0.165		17.291	< .001	2.522	3.169
	Openness	0.164	0.046	0.159	3.599	< .001	0.075	0.254



Regression line

Coefficients

Model	*Agreeab.*	Unstand.	Standard Error	Stand.	t	p	CI 2.5%	CI 97.5%
1	intercept	2.845	0.165		17.291	< .001	2.522	3.169
	Openness	**0.164**	0.046	0.159	3.599	< .001	0.075	0.254



Regression line

- At $x = 0$, intercept (line start) is $y = 2.845$
  - Intercept does not necessarily have a real-life explanation
- For each $x = 1$, $y = 0.164x$
- Each additional x will yield additional $y = 0.164$
- Allows us to do predictions!

Regression line

Coefficients

Model	*Agreeab.*	Unstand.	**Standard Error**	Stand.	t	p	CI 2.5%	CI 97.5%
1	intercept	2.845	**0.165**		17.291	< .001	2.522	3.169
	Openness	0.164	**0.046**	0.159	3.599	< .001	0.075	0.254

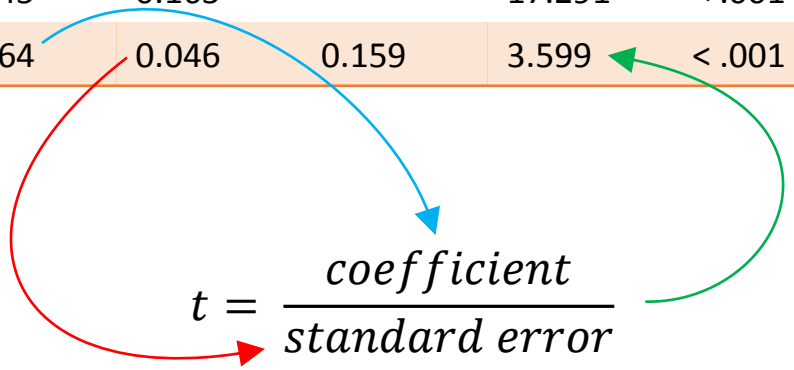
Regression line

- T statistic
- $t = \frac{coefficient}{standard\ error}$
- The **higher** the t, the **more reliable/significant** the coefficient is – the more variation the coefficient explains

Regression line

Coefficients

Model	*Agreeab.*	Unstand.	Standard Error	Stand.	t	p	CI 2.5%	CI 97.5%
1	intercept	2.845	0.165		17.291	< .001	2.522	3.169
	Openness	0.164	0.046	0.159	3.599	< .001	0.075	0.254



$$t = \frac{coefficient}{standard\ error}$$

Regression line

Coefficients

Model	*Agreeab.*	Unstand.	Standard Error	Stand.	t	**p**	CI 2.5%	CI 97.5%
1	intercept	2.845	0.165		17.291	**< .001**	2.522	3.169
	Openness	0.164	0.046	0.159	3.599	**< .001**	0.075	0.254

Regression line

- t statistic is important to get the **significance** value of our coefficient
- Statistical significance shows us to what extent there is a probability of acquiring the value of t statistic as a **result of a random chance**
- Statistical significance p
  - $p \leq 0.05$ – 95% – (at most 1 in 20)
  - $p \leq 0.01$ – 99% – (at most 1 in 100)
  - $p \leq 0.001$ – 99.9% – (at most 1 in 1000)

Regression line

Coefficients

Model	*Agreeab.*	Unstand.	Standard Error	Stand.	t	**p**	CI 2.5%	CI 97.5%
1	intercept	2.845	0.165		17.291	**< .001**	2.522	3.169
	Openness	0.164	0.046	0.159	3.599	**< .001**	0.075	0.254

Regression line

Coefficients

Model	*Agreeab.*	Unstand.	Standard Error	Stand.	t	p	**CI 2.5%**	**CI 97.5%**
1	intercept	2.845	0.165		17.291	< .001	**2.522**	**3.169**
	Openness	0.164	0.046	0.159	3.599	< .001	**0.075**	**0.254**

Regression line

- Point estimates are not best way how to report on regression line
- Confidence intervals should be taken into account
- Confidence interval of the regression line should not include 0
  - Otherwise, there may be a chance null hypothesis can’t be rejected

Regression line

Coefficients

Model	*Agreeab.*	Unstand.	Standard Error	Stand.	t	p	CI 2.5%	CI 97.5%
1	intercept	2.845	0.165		17.291	< .001	2.522	3.169
	Openness	0.164	0.046	0.159	3.599	< .001	**0.075**	**0.254**

Regression line

- Prediction
  - Suppose new observation $x = 2.85$
  - Point estimate is $y = 2.845 + (0.164 * 2.85) = 3.3124$
  - **Taking into account confidence interval**, the point estimate may be within range $y = \langle 3.2229 \ - 3.4019 \rangle$
  - E.g. point estimate of $x = 2.85$ should be reported as $y = 3.3124 \ \pm 0.0895$

Multiple regression

Model Summary

R	R^2	Adjusted R^2	RMSE
0.164	0.027	0.023	0.336

ANOVA

	Sum of Squares	df	Mean Square	F	p
Regression	1.552	2	0.776	6.865	0.001
Residual	56.189	497	0.113		
Total	57.742	499			

Coefficients

Openness	Unstand.	Standard Error	Stand.	t	p	0.025	0.975
intercept	3.153	0.18		17.498	< .001	2.799	3.507
Conscientiousness	-0.035	0.039	-0.04	-0.886	0.376	-0.112	0.042
Agreeableness	0.161	0.043	0.166	3.693	< .001	0.075	0.246

Sources

- Lienert, P., Suetterlin, B., & Siegrist, M. (2015). Public acceptance of the expansion and modification of high-voltage power lines in the context of the energy transition. *Energy Policy*, *87*, 573–583. http://doi.org/10.1016/j.enpol.2015.09.023
- Frantál, B. (2016). Living on coal: Mined-out identity, community displacement and forming of anti-coal resistance in the Most region, Czech Republic. *Resources Policy*, *49*, 385–393. http://doi.org/10.1016/j.resourpol.2016.07.011
- Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, A. J., Ly, A., Gronau, Q. F., Smira, M., Epskamp, S., Matzke, D., Wild, A., Knight, P., Rouder, J. N., Morey, R. D., & Wagenmakers, E.-J. (2015). JASP (Version 0.7.5)[Computer software].
- Dolan, C. V, Oort, F. J., Stoel, R. D., & Wicherts, J. M. (2009). Testing Measurement Invariance in the Target Rotated Multigroup Exploratory Factor Model. *Structural Equation Modeling: A Multidisciplinary Journal*, *16*(2), 295–314. http://doi.org/10.1080/10705510902751416