The European Commission's science and knowledge service

Joint Research Centre

European Commission

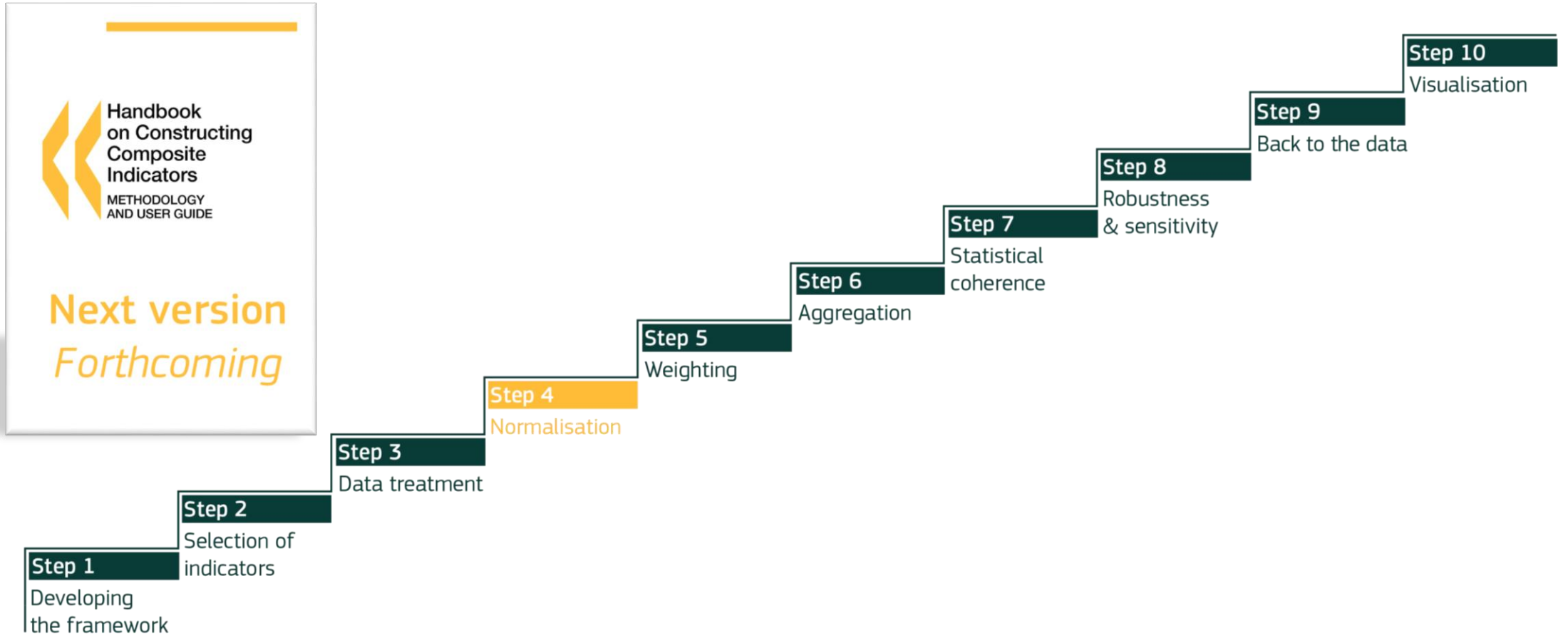
Step 4: Normalisation for Composite Indicators

Maria Del Sorbo

COIN 2019 - 17th JRC Annual Training on Composite Indicators & Scoreboards
04-06/11/2019, Ispra (IT)

European Commission

Ten steps



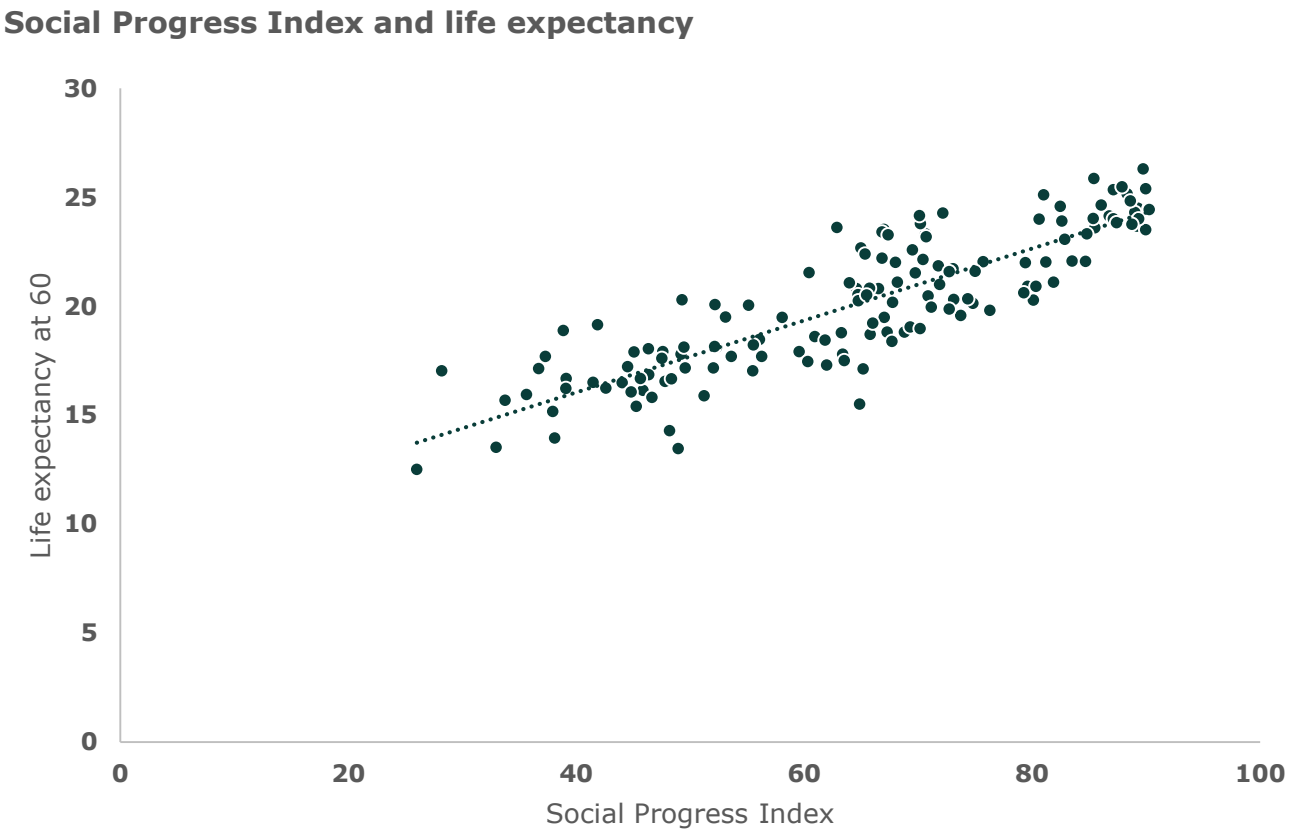
European Commission

Outline

- Before normalising data
- Definition — *What is data normalisation? /Why do we need it?*
- Normalisation methods — *How do we normalise data?*
  1. standardization (or z-score)
  2. min-max
  3. distance to a reference
  4. categorical scale
  5. ranking
  6. quantile empirical distribution
- Summary table
- Key messages

Before normalising data

Adjust for direction



Social Progress Imperative, Social Progress Index, 2018

Prior normalisation ***take properly into account the sign of the indicators***, i.e. positive vs. negative orientation towards the index

European Commission

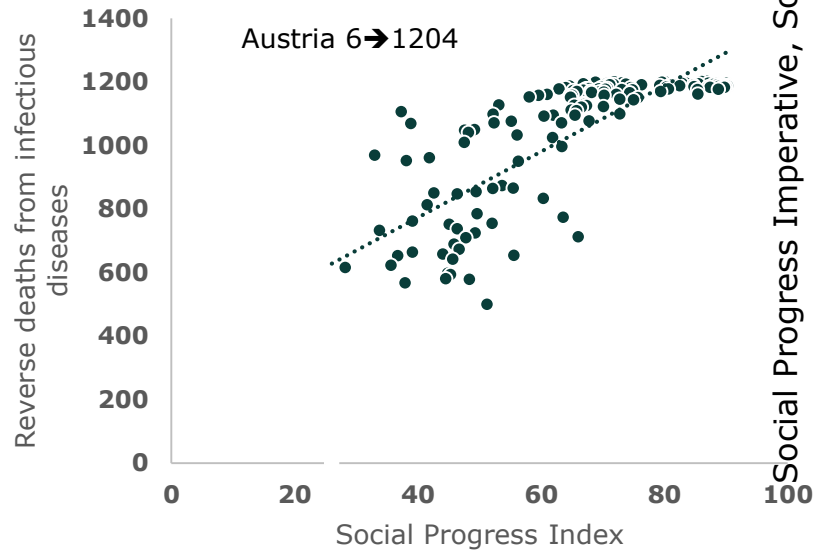
Before normalising data

Adjust for direction

Social Progress Index and deaths from infectious diseases

Social Progress Index and reverse deaths from infectious diseases

Reverse I= (max-x)
max=1210



Social Progress Imperative, Social Progress Index, 2018

Make sure that higher values in the dataset mean better results, if not, reverse the original direction

European Commission

What is data Normalisation?

Definition:
… is the adjustment of variables onto a common scale, *prior to any data aggregation.*

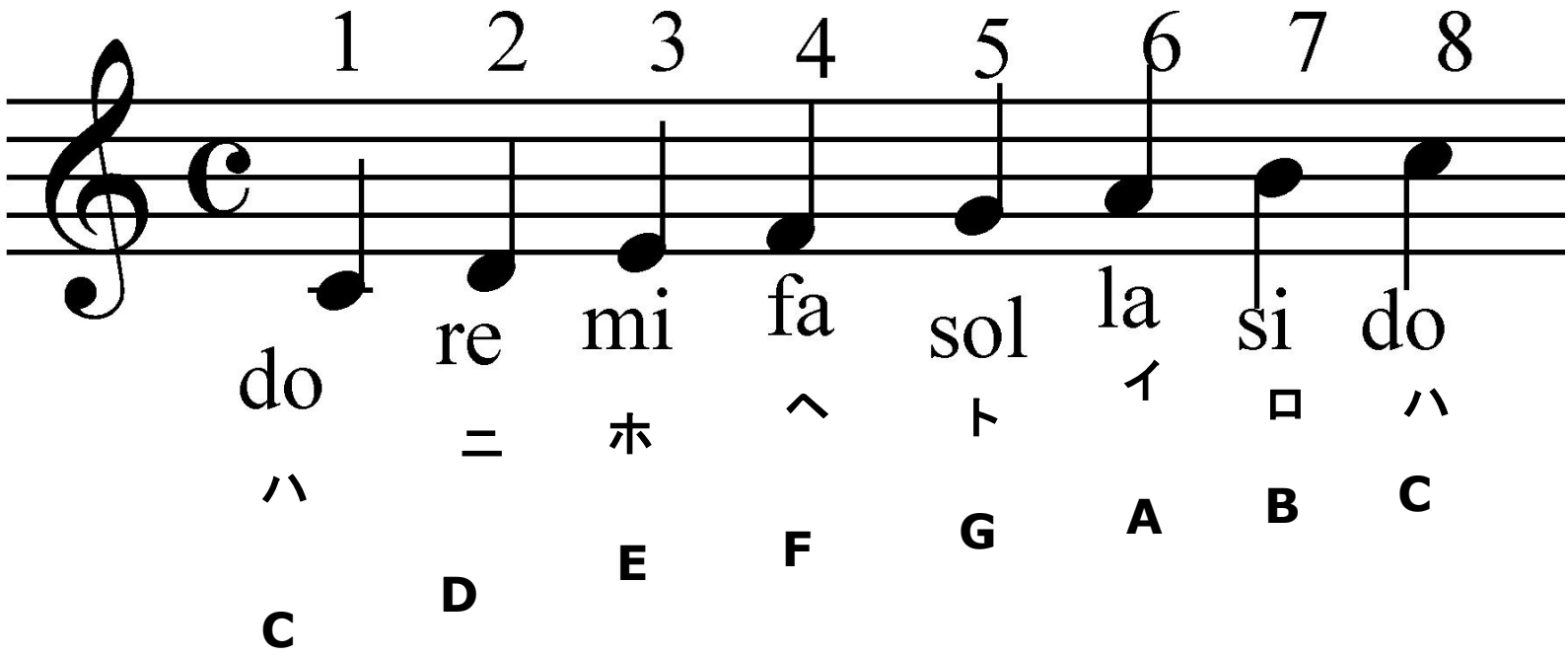
Aim: comparability across variables by dealing with

1. different units of measurement
2. different ranges of variation

What is data Normalisation?

"… avoid adding up apples and oranges"

Why do we need it?



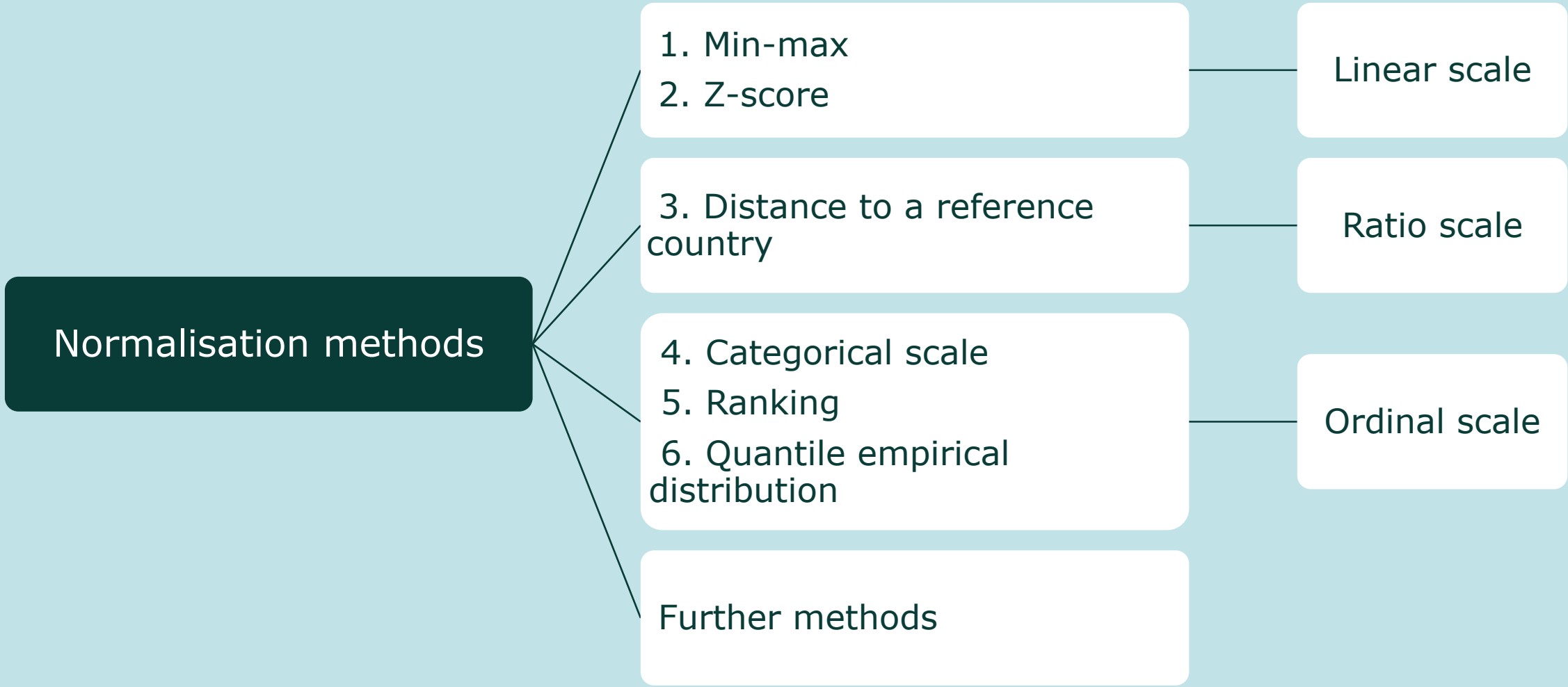
https://en.wikipedia.org/wiki/Musical_note

European Commission

Why do we need it?

Different units of measurement => common scale
Different ranges of variation => suitable range of variation

This allows variables to be combined in averages (i.e. composite indicators) without giving undue weight to variables with different scales



1. the normalisation method should respect the **conceptual framework** and **the data properties**
2. different normalisation methods may lead to different rankings

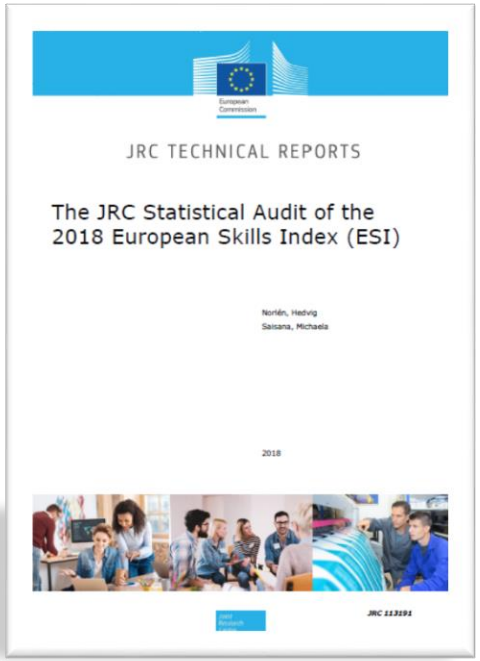
‘The goal is to turn data into information, and information into insight’

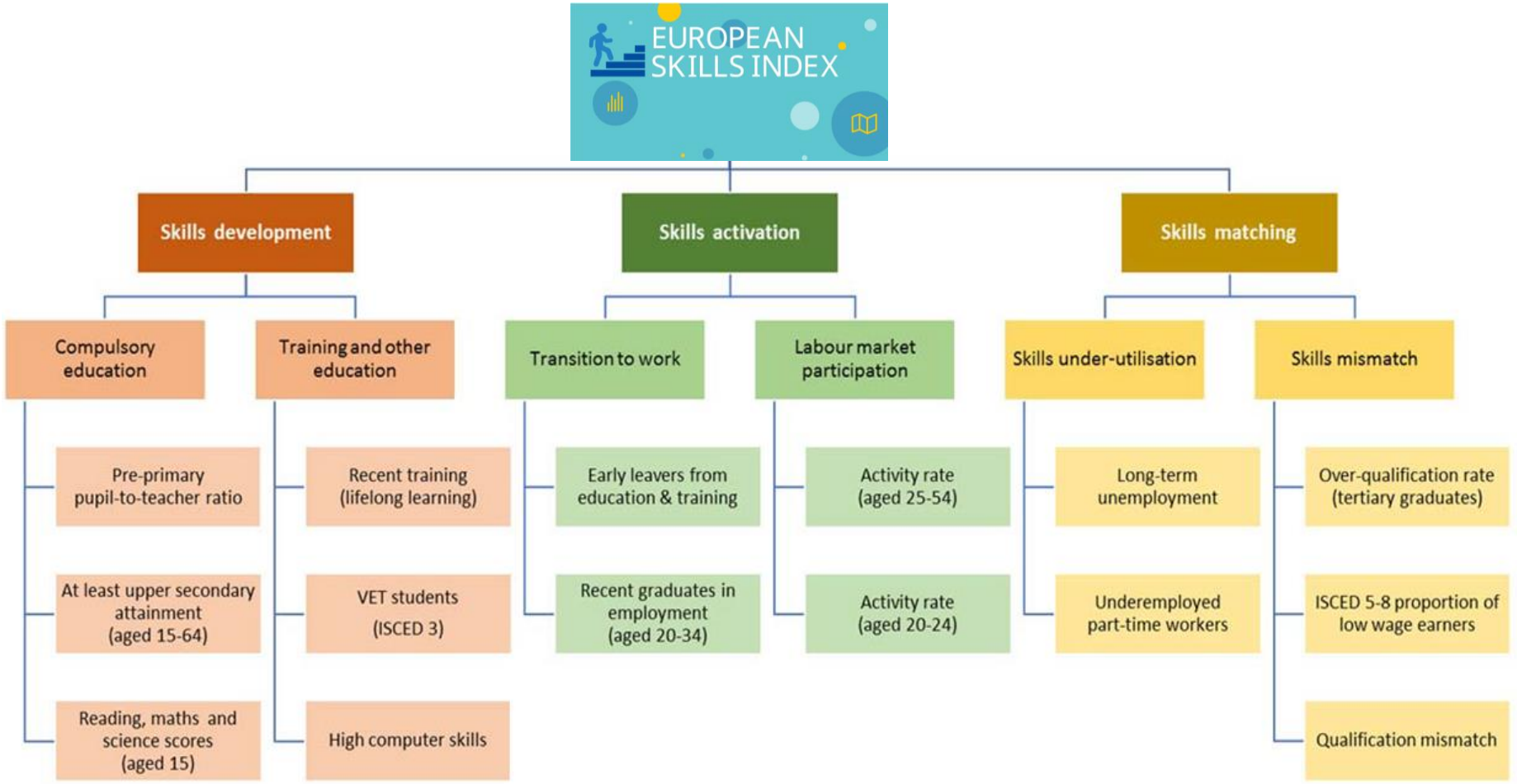
Carly Fiorina

Linear scale

1. Z-score
2. Min-max

Example: European Skills Index





Source: European Skills Index (2018), Cedefop

European Commission

1. Z-score

Indicators before and after z-score normalisation	3. Reading, maths & science scores aged 15 (Pisa score)	4. Recent training	6. High computer skills
Before normalisation			
Mean	486.94	10.85	29.18
Variance	23.44	7.61	7.93
Min	437.49	1.20	7.00
Max	524.29	29.60	46.00
Variation range	**[437.49, 524.29]**	**[1.2, 29.6]**	**[7, 46]**

After z-score normalisation			
Mean	**0**	**0**	**0**
Variance	**1**	**1**	**1**
Min	-2.11	-1.27	-2.80
Max	1.59	2.46	2.12
Variation range	[-2.11, 1.59]	[-1.27, 2.46]	[-2.8, 2.12]

How are the two indicators different?

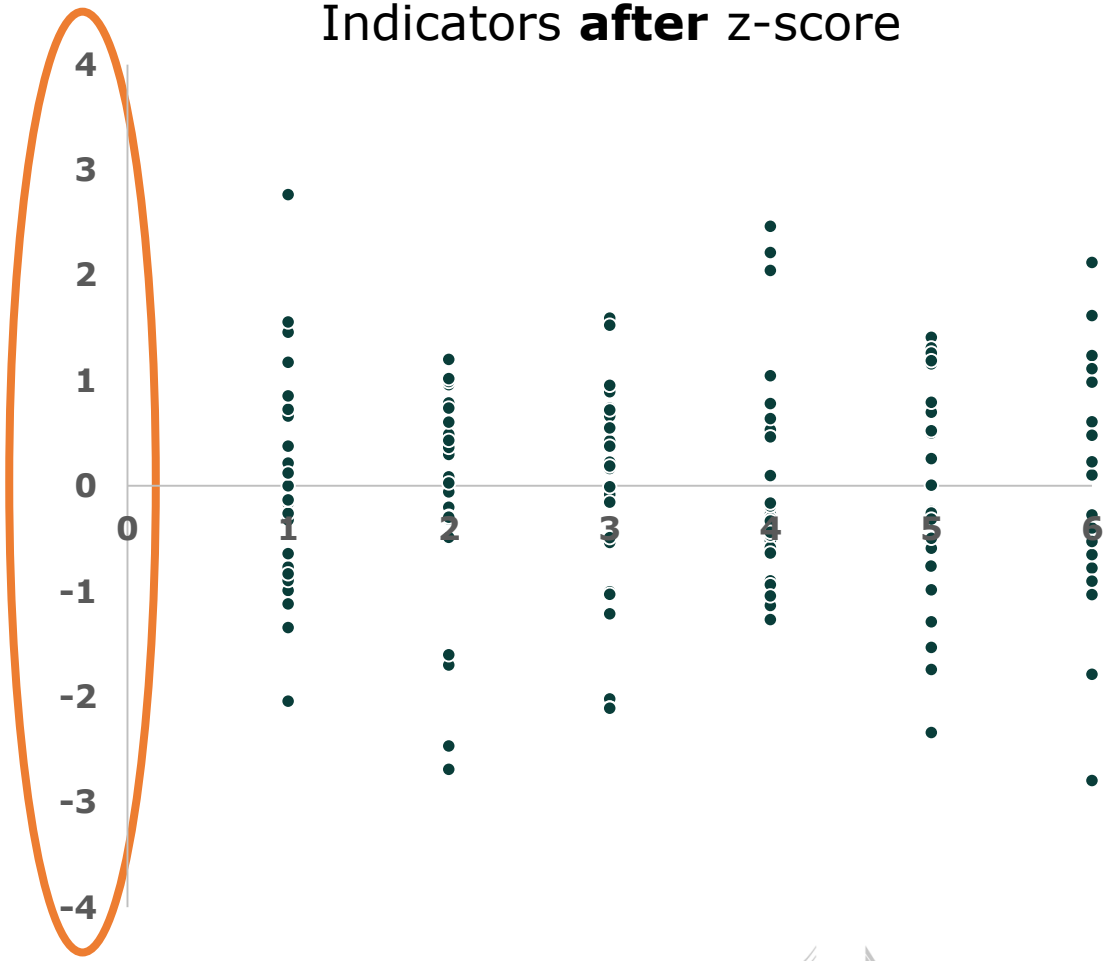
1. Units of measurement
2. Ranges of variation

$$z = \frac{x - \mu}{\sigma}$$

Z-score effects

- Unit of measurement
- $I \sim \mu = 0$, $\sigma^2 = 1$
- Variation range: no adjustments
- Extreme values: no adjustments
- Distribution: no adjustments

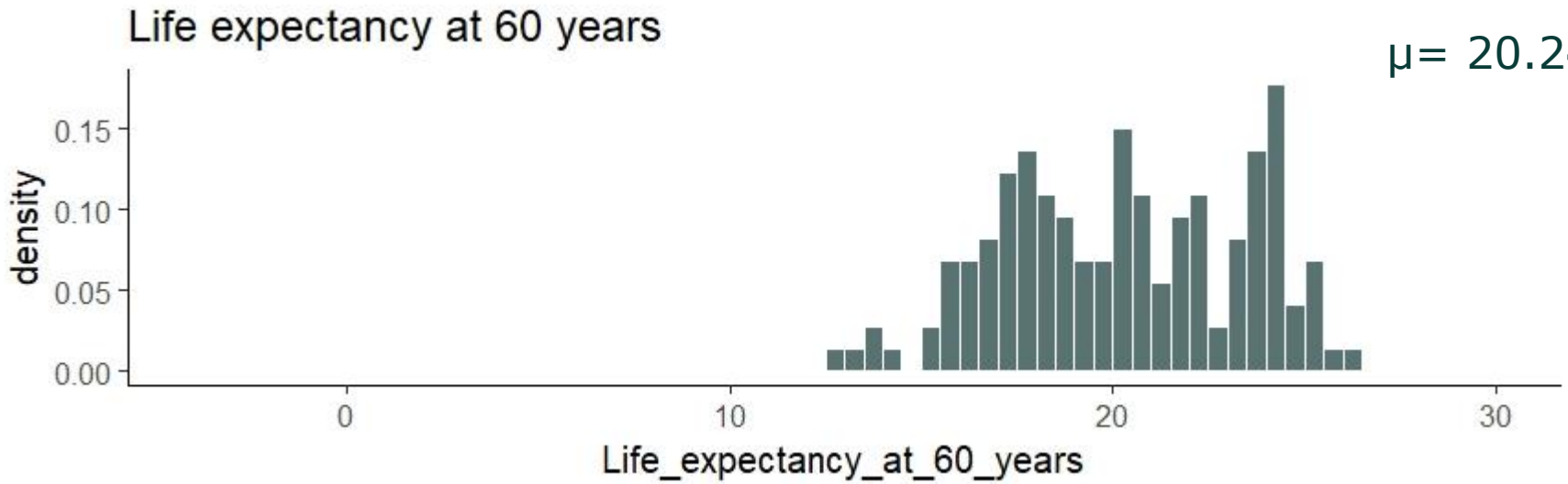
1. Z-score



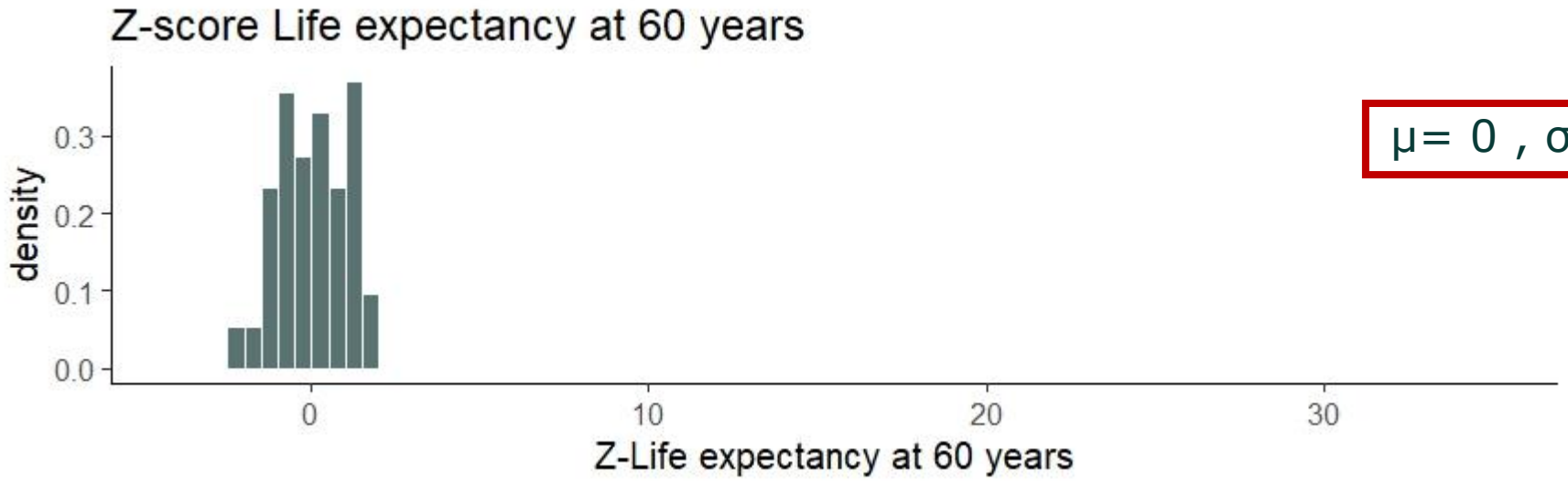
Source: European Skills Index (2018), Cedefop

European Commission

1. Z-score



$\mu = 20.24$, $\sigma^2 = 9.87$



$\mu = 0$, $\sigma^2 = 1$

European Commission

2. Min-max

Indicators before and after z-score normalisation	3. Reading, maths & science scores aged 15 (Pisa score)	4. Recent training	6. High computer skills
Before normalisation			
Mean	486.94	10.85	29.18
Variance	23.44	7.61	7.93
Min	437.49	1.20	7.00
Max	524.29	29.60	46.00
Variation range	**[437.49, 524.29]**	**[1.2, 29.6]**	**[7, 46]**
After normalisation using min-max			
Mean	0.55	0.34	0.57
Variance	0.08	0.07	0.04
Min	**0.00**	**0.00**	**0.00**
Max	**1.00**	**1.00**	**1.00**
Variation range	**[0, 1]**	**[0, 1]**	**[0, 1]**

How are the two indicators different?

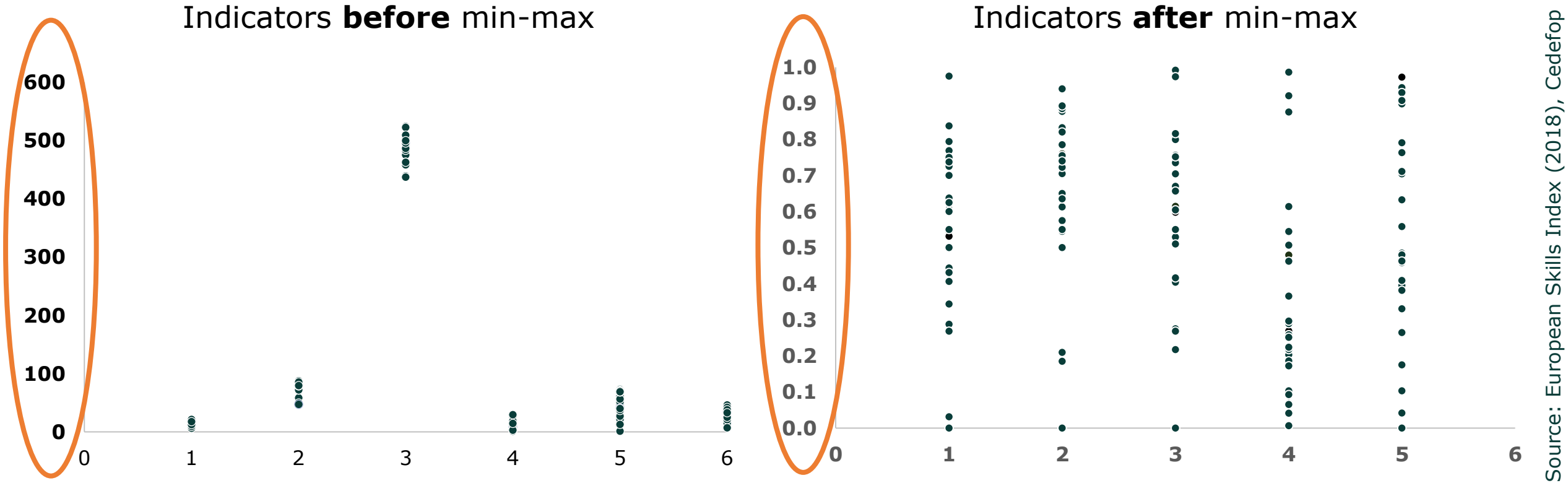
1. Units of measurement
2. Ranges of variation

$$I = \frac{x - \min}{\max(x) - \min(x)}$$

Min-max effects

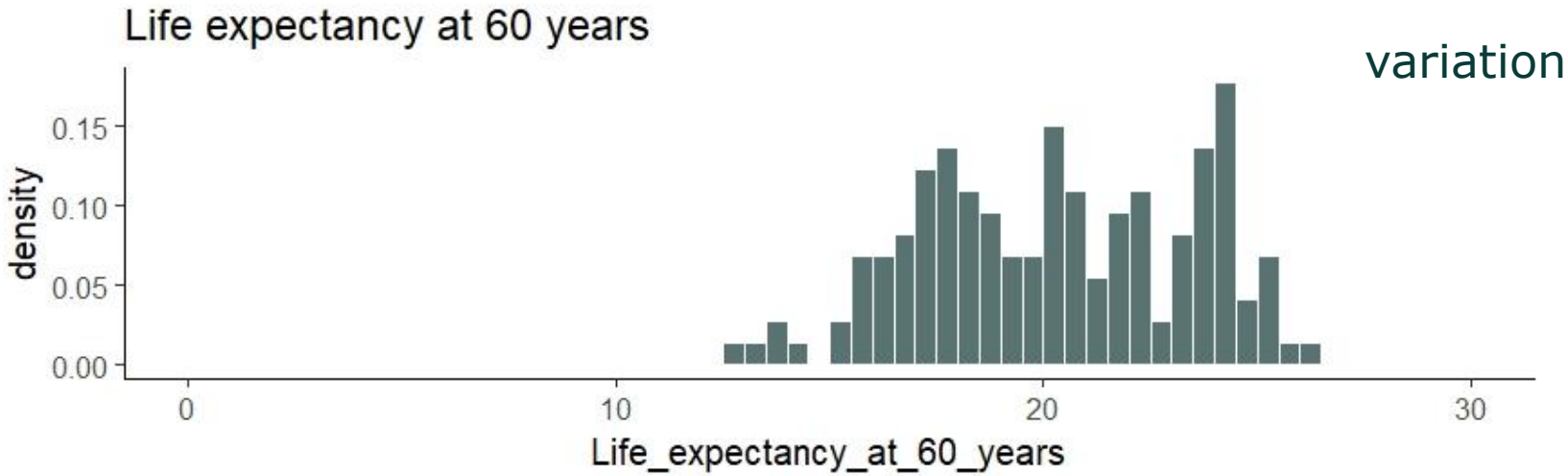
- Unit of measurement
- μ, σ^2 no adjustments
- I Variation range: [0, 1]
- Extreme values: no adjustments
- Distribution: no adjustments

2. Min-max

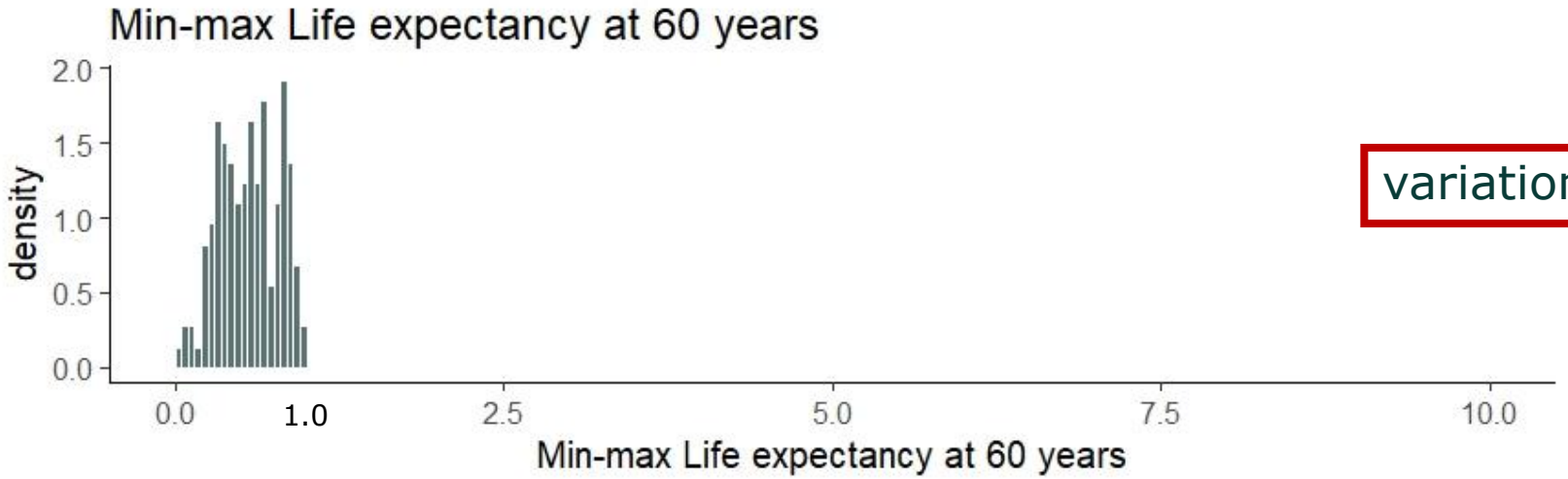


Rescaling eases communication to public: all indicators move in the same range [0, 1] or [0, 100], usually higher score represent better achievement

2. Min-max



variation range = [12.5, 24.1]



variation range = [0, 1]

European Commission

Correlation structure
before and after linear transformation

Unchanged Correlation* structure between 3 indicators - EU Skills Index

Raw indicators	Reading, maths & science scores aged 15 (Pisa score)	Recent training	High computer skills
Reading, maths & science scores (aged 15)	1		
Recent training	0.58	1	
High computer skills	0.63	0.77	1
Z-score normalised indicators			
Reading, maths & science scores (aged 15)	1		
Recent training	0.58	1	
High computer skills	0.63	0.77	1
Min-max normalised indicators			
Reading, maths & science scores (aged 15)	1		
Recent training	0.58	1	
High computer skills	0.63	0.77	1

* Linear Pearson Correlation

Sig. level 0.01

Ratio scale

3. Distance to a reference unit

3. Distance to a reference unit

The reference unity may be a country, city, region, company, etc.:

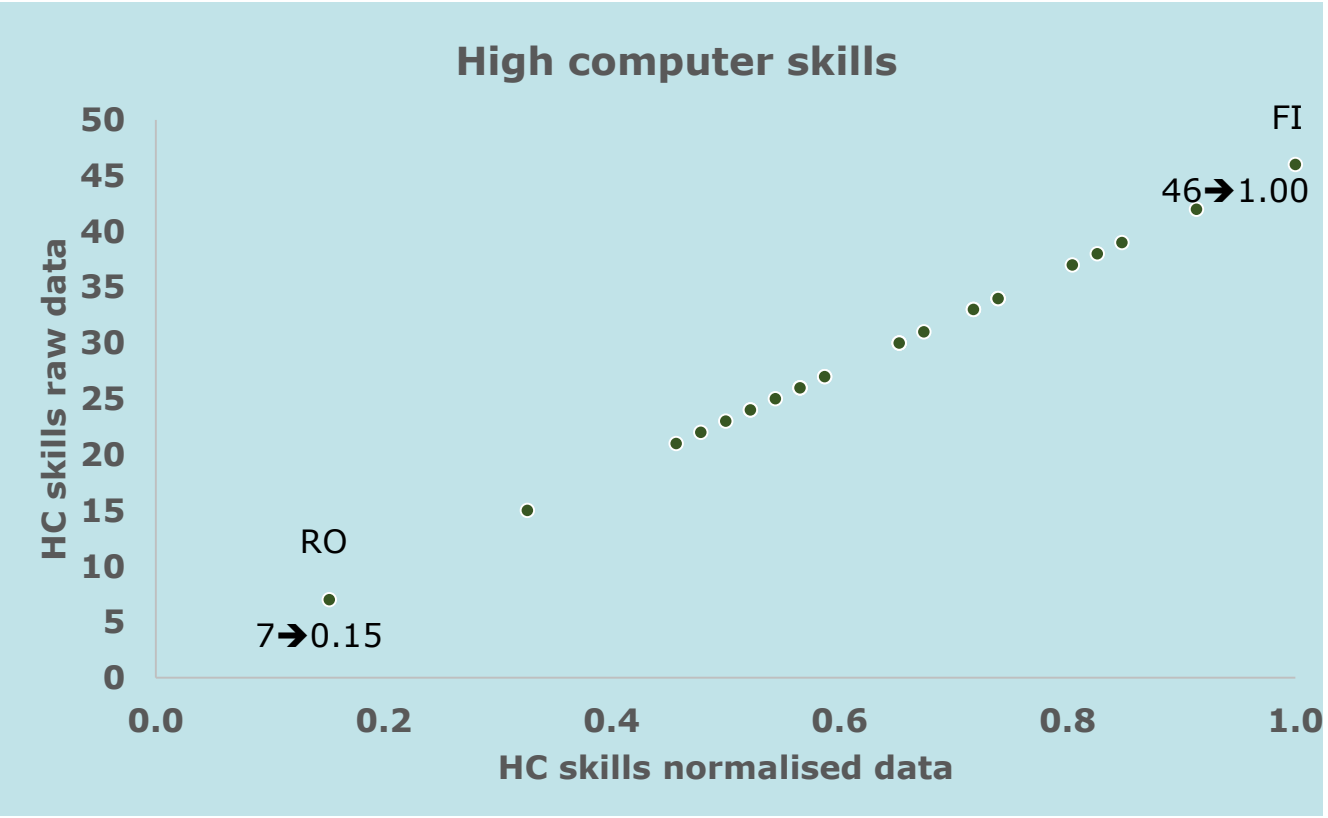
- group leader, external benchmark or hypothetical country, city etc. (target to be reached in a given timeframe)
- average (eg., EU28, world)

$$I_c = \frac{x_c}{x_{\bar{c}}}$$

Indicator evolution across time (e.g. reference time t_0)

$$I_c = \frac{x_c^t}{x_c^{t_0}}$$

3. Distance to a reference unit



Indicators after distance to a reference unit

$$I_c = \frac{x_c}{x_{\bar{c}}}$$

Distance to a reference unit effects

- Unit of measurement
- μ, σ^2 no adjustments
- Variation range no adjustments
- Extreme values: no adjustments
- Distribution: no adjustments

Ordinal scale

4. Categorical scale
5. Ranking
6. Quantile empirical distribution

4. Categorical scale

Ordinal scales

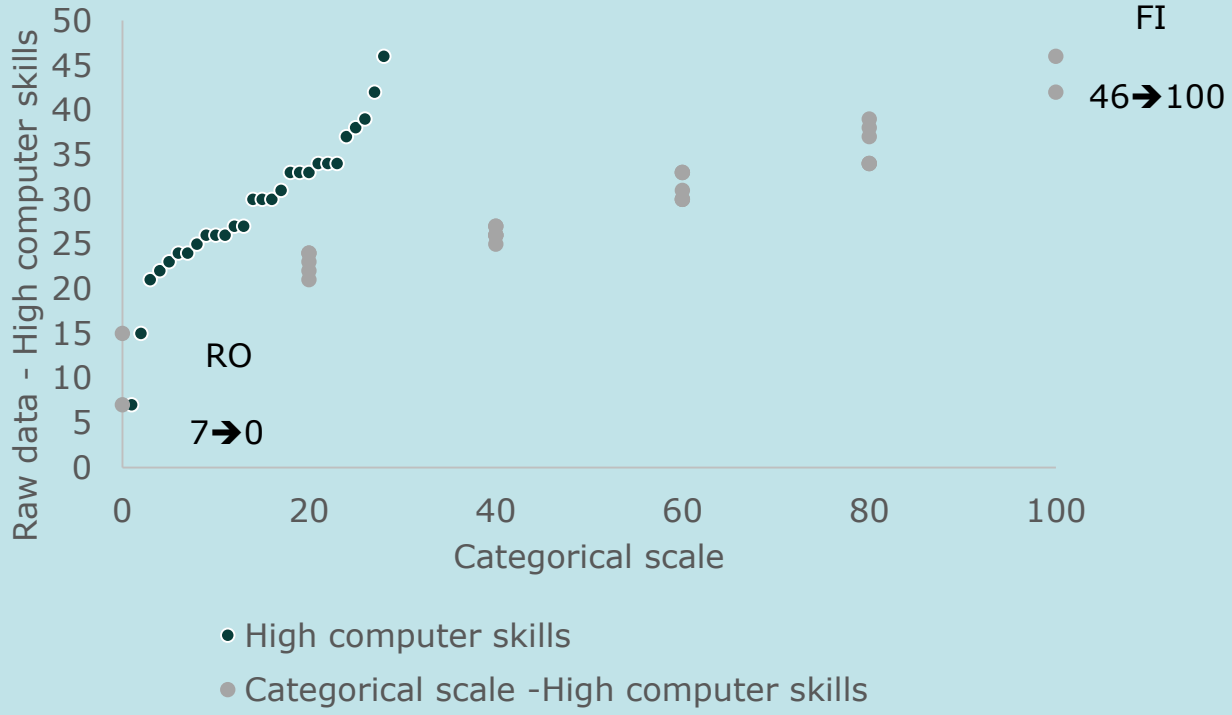
Indicator score based on categories: e.g. school grade: crèche, primary school, secondary school, high school, university

Numerical scales

- Categories lie on a variation range portion
- Categories can be based on the percentile of the distribution of the indicator across countries
- Justify the choice of intervals and scores

$$I_{q,c}^{t} = \begin{cases} 0 & if \ p^{0} \le x < p^{10} \\ 20 & if \ p^{10} \le x < p^{25} \\ 40 & if \ p^{25} \le x < p^{50} \\ 60 & if \ p^{50} \le x < p^{75} \\ 80 & if \ p^{75} \le x < p^{90} \\ 100 & if \ p^{90} \le x \le p^{100} \end{cases}$$

4. Categorical scale



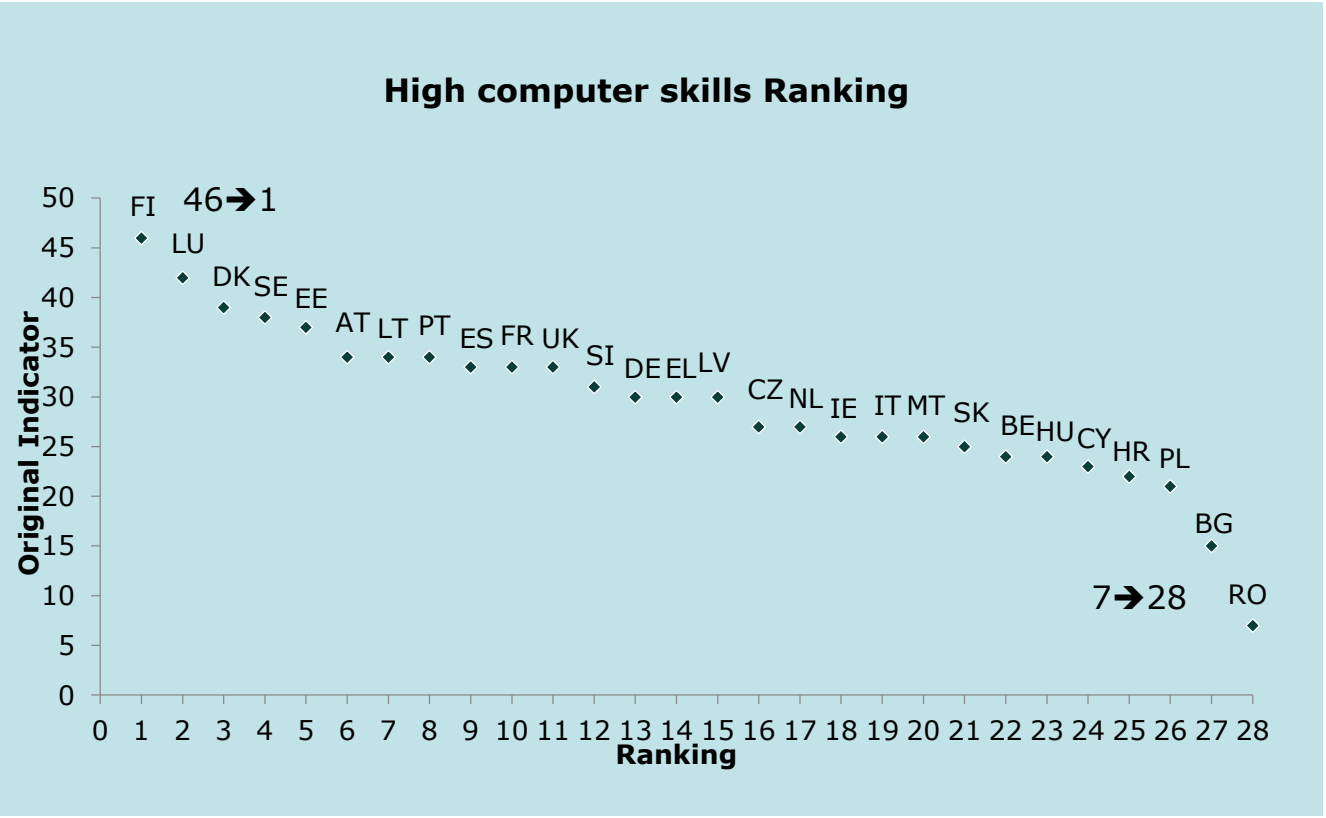
Indicators after categorical scaling

$$I_{q,c}^{t} = \begin{cases} 0 & if \ p^{0} \le x < p^{10} \\ 20 & if \ p^{10} \le x < p^{25} \\ 40 & if \ p^{25} \le x < p^{50} \\ 60 & if \ p^{50} \le x < p^{75} \\ 80 & if \ p^{75} \le x < p^{90} \\ 100 & if \ p^{90} \le x \le p^{100} \end{cases}$$

Categorical scale effects

- Unit of measurement
- Variation range [0, 100]
- Variance: *depends on the categories*
- Robust to extreme values
- Distribution: No uniform

5. Ranking



Indicators after ranking scale

- Scores are replaced by ranks, e.g. the highest score receives rank 1
- Uses only ordinal information, information on levels is not kept

$$I = rank(x)$$

Ranking scale effect

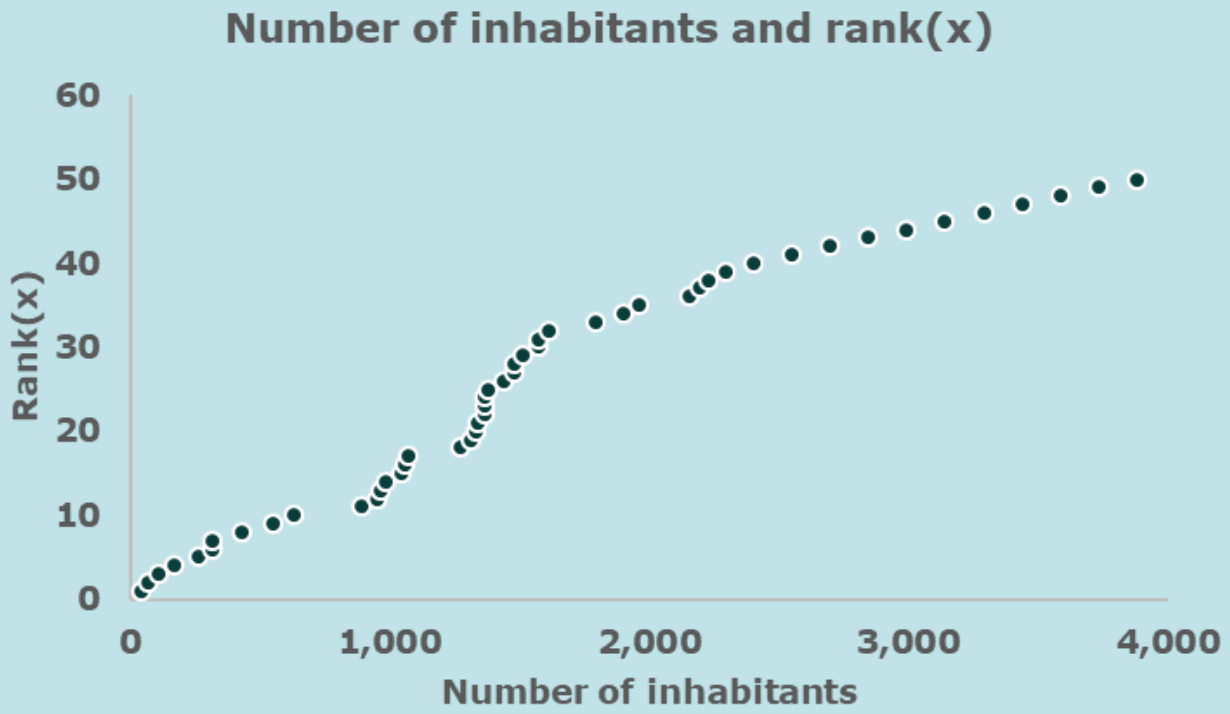
- Unit of measurement
- Range [1, n], our case n=28
- Same variance, our case $\sigma^2 = 65.25$
- Robust to extreme values
- Distribution: Uniform

Source: European Skills Index, 2018

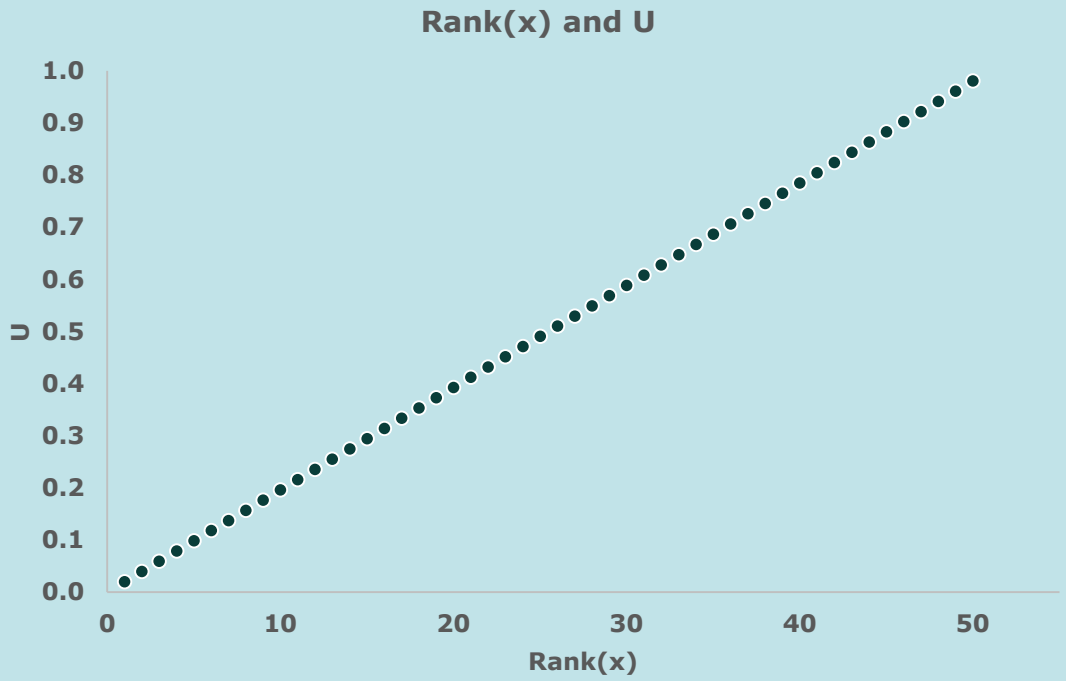
6. Quantile empirical distribution

Definition: the quantile normalisation makes two variables identical distributed (same σ^2 and same shape)

Rank(x) is the rank associated to the set of observations



6. Quantile empirical distribution



Indicators after quantile empirical distribution

x= million inhabitants	Rank(x)	U
37	1	0.02
68	2	0.04
110	3	0.06
167	4	0.08
259	5	0.10
....	....	...
3,290	46	0.90
3,438	47	0.92
3,586	48	0.94
3,734	49	0.96
3,882	50	0.98

$$u = \frac{rank(x)}{N+1}$$

U [0, 1]
μ equal to median= ½ (a+b)
= μ_u 0.5

Quantile empirical distribution effects:

- Unit of measurement
- Variation range [0, 1]
- Same variance $\sigma^2 = 1/12(b-a)^2$
- Robust to extreme values
- Distribution: Uniform

Normalisation methods effects
to sum up

	Normalisation methods				
Normalisation effects	Quantile empirical distribution/Ranking	Categorical scale	Z-score	Min-max	Distance to a reference country
Unit of measurement	Y	Y	Y	Y	Y
Variance	Y	Y/N*	Y	N	N
Range of variation	Y	Y	N	Y	N
Extreme values**	Y	Y	N	N	N
Distribution***	Y	Y/N*	N	N	N
*	Yes, only if there are not tied ranks				
**	Non-sensitive to extreme values				
***	The distribution will be the same for the normalised indicators				

Key messages
Step 4 - Normalisation

What is data normalisation? / Why do we need it?

Converting data onto a common **scale**
Effects on data, e.g. same range of variation across indicators ➔ min-max
Prepare the data for the aggregation step

How do we normalise data?

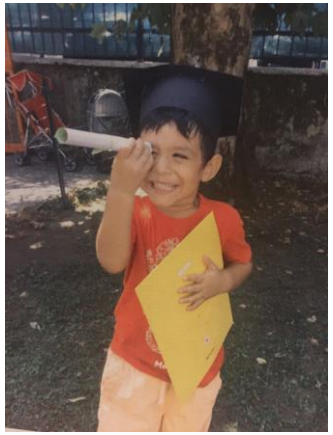
Six **normalisation methods➔** choice coherent with **data structur**e and **conceptual framework** (COIN tips)

Alternative normalisation methods within uncertainty/sensitivity analysis

European Commission

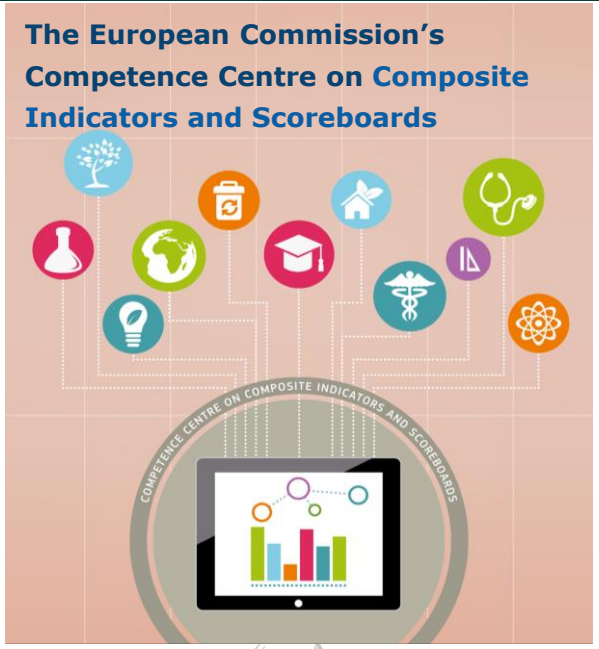
THANK YOU

Welcome to email us at: jrc-coin@ec.europa.eu



COIN in the EU Science Hub
https://ec.europa.eu/jrc/en/coin

COIN tools are available at:
https://composite-indicators.jrc.ec.europa.eu/



European Commission

References

- Becker, W., M. Saisana, P. Paruolo, and I. Vandecasteele. 2017. 'Weights and Importance in Composite Indicators: Closing the Gap. Ecological Indicators 80: 12–22.
- Bichindaritz I., Big Data, Genes, and Medicine, the State University of New York (SUNY), https://www.coursera.org/lecture/data-genes-medicine/data-normalisation-jGN7k
- Cohen, A., Saisana, M., 2014, Quantifying the Qualitative: Eliciting Expert Input to Develop the Multidimensional Poverty Assessment Tool, Journal of Development Studies 50(1): 35-50.
- European Union, European Centre for the Development of Vocational Training (Cedefop). 2018, European Skills Index technical report. Thessaloniki, Greece.
- OECD/JRC, 2008, *Handbook on Constructing Composite Indicators. Methodology and user Guide*, OECD Publishing, ISBN 978-92-64-04345-9.
- Paruolo P., Saisana M., Saltelli A., 2013, Ratings and Rankings: voodoo or science?. J Royal Statistical Society A 176(3), 609-634.
- Saisana, M., and Saltelli, A., 2011, Rankings and Ratings: Instructions for use, Hague Journal on the Rule of Law 3(2), 247-268.

References

- Saisana M., D'Hombres B., Saltelli A., 2011, Rickety Numbers: Volatility of university rankings and policy implications. Research Policy 40, 165–177
- Social Progress Imperative, Social Progress Index, 2018
- Unsplashed, photos. 2019, Icons made by: Sean Mungur, Twins fisch, https://unsplash.com/
- Western musical scale, wikipedia: https://en.wikipedia.org/wiki/Musical_note

Technical Appendix

Z-score – time-dependent studies

$$I_{q,c} = \frac{x_{q,c} - \bar{x}_q}{\sigma_q} \quad \Rightarrow \quad I^{t}_{q,c} = \frac{x^{t}_{q,c} - \bar{x}^{t_0}_q}{\sigma^{t_0}_q}$$

For time-dependent studies, in order to assess country performance across years, the average and the standard deviation across countries are calculated for a reference year, usually the initial time point.

Otherwise, we lose info on both trend and spread.

Min-Max – time-dependent studies

The expression $$\mathrm{I}_{\mathrm{q,c}}^{\mathrm{t}} = \frac{x_{q,c}^{t} - \min_c\left(x_q^{t_0}\right)}{\max_c\left(x_q^{t_0}\right) - \min_c\left(x_q^{t_0}\right)}$$

is sometimes used in time-dependent studies. However, if:

$$x_{q,c}^{t} > \max_c\left(x_q^{t_0}\right)$$

the normalized indicator would be larger than 1

Min-Max – time-dependent studies

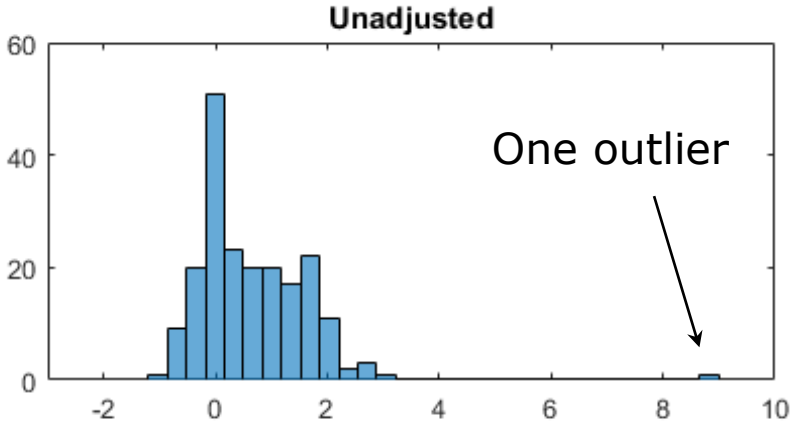
Minimum and maximum for each indicator are calculated across countries and time. Normalised indicators values [0, 1]

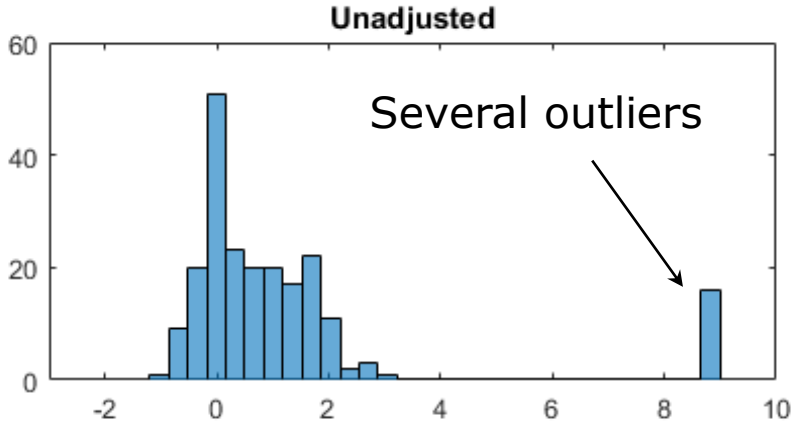
When **data for a new time point** become available the global minimum and/or the maximum may be affected. To keep comparability between previous and new data, the composite indicator for previous data must be re-calculated.

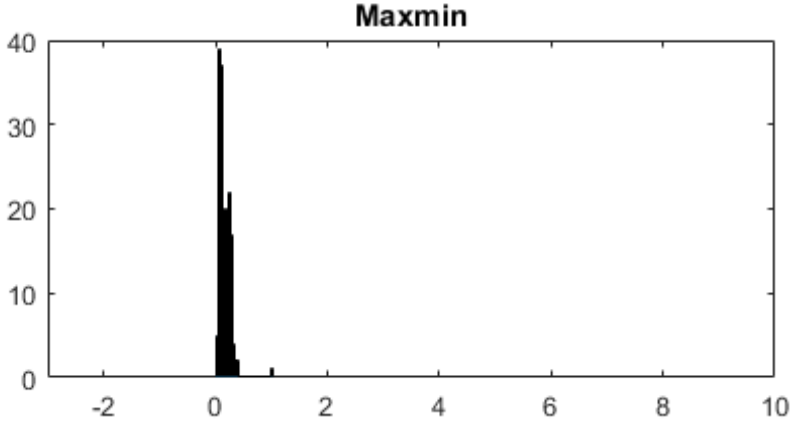
A simple alternative:

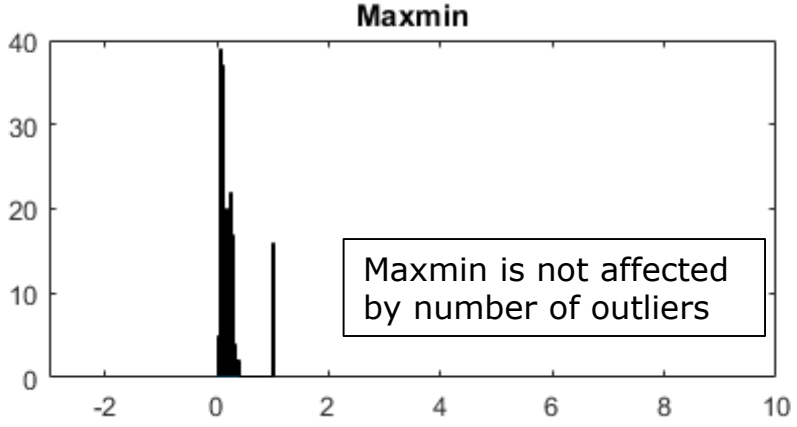
$$I_{q,c}^{t} = \frac{x_{q,c}^{t} - \min_{c,t}(x_q)}{\max_{c,t}(x_q) - \min_{c,t}(x_q)}$$

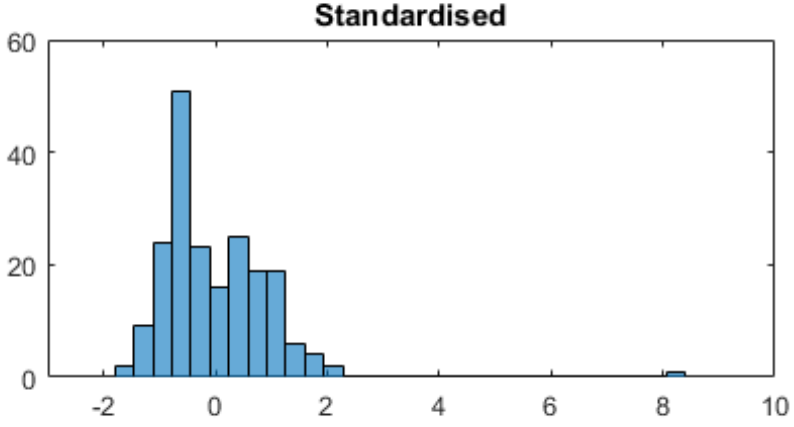
European Commission

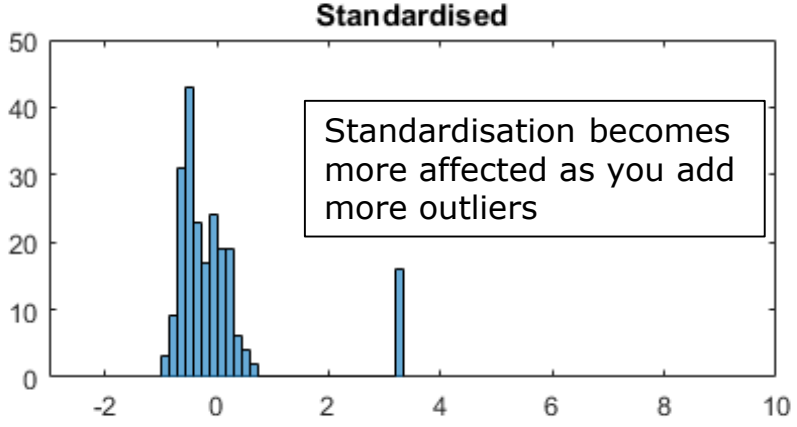












European Commission