European Commission

Step 4 Normalisation

18th JRC Annual training on Composite Indicators and Scoreboards

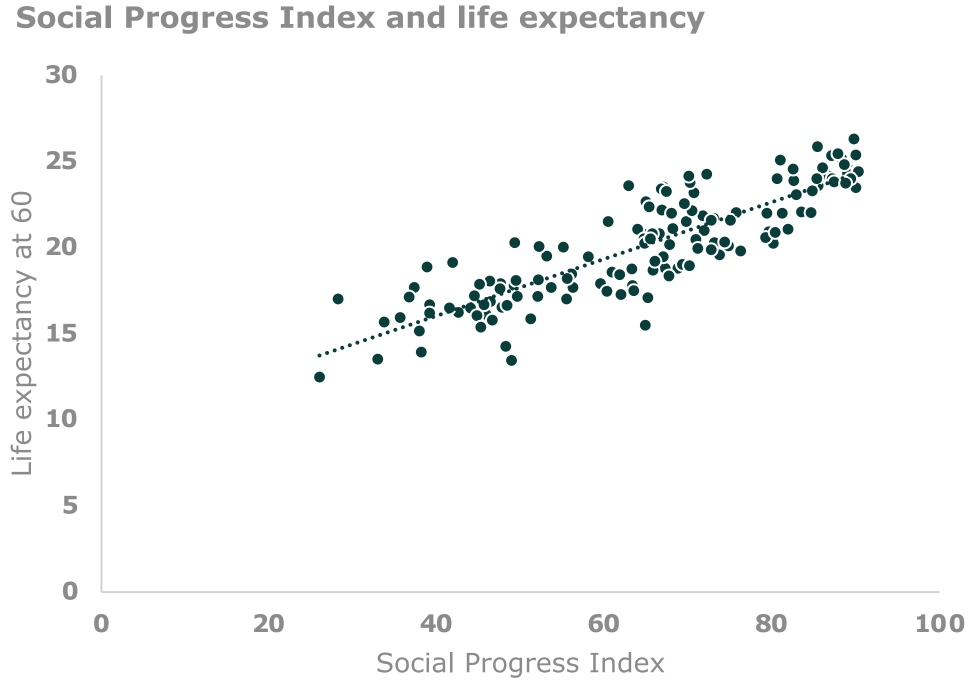
Pablo de Pedraza

Joint Research Centre

10 STEPS to build a Composite Indicator

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

Before normalising data

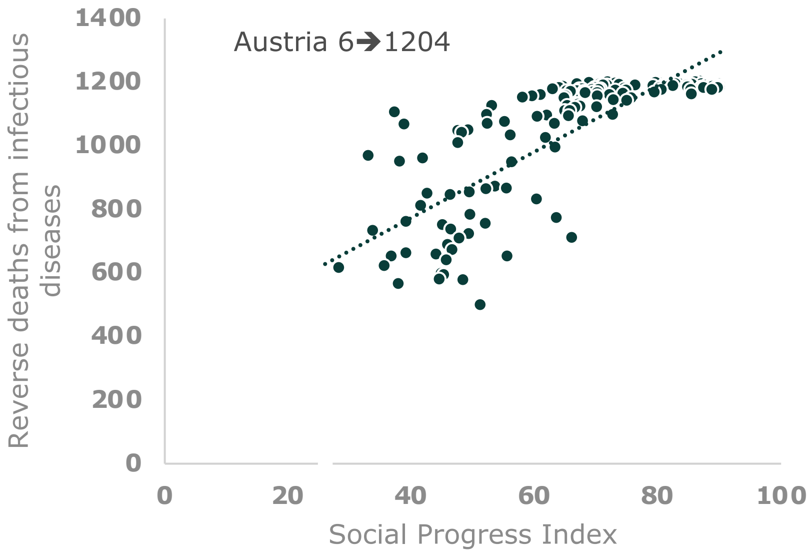
Social Progress Index and deaths from infectious diseases

Max = 1210 (not plotted)

Adjust for direction

Social Progress Index and reverse deaths from infectious diseases

Reverse I= (max-x)



Social Progress Imperative, Social Progress Index, 2018

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

What is data Normalisation?

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

Aim: achieve comparability of variables dealing with

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

Normalisation

Normalisation methods

1. Z-score
2. Min-max

Linear scale

3. Distance to a reference country

Ratio scale

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

Ordinal scale

Further methods

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

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]

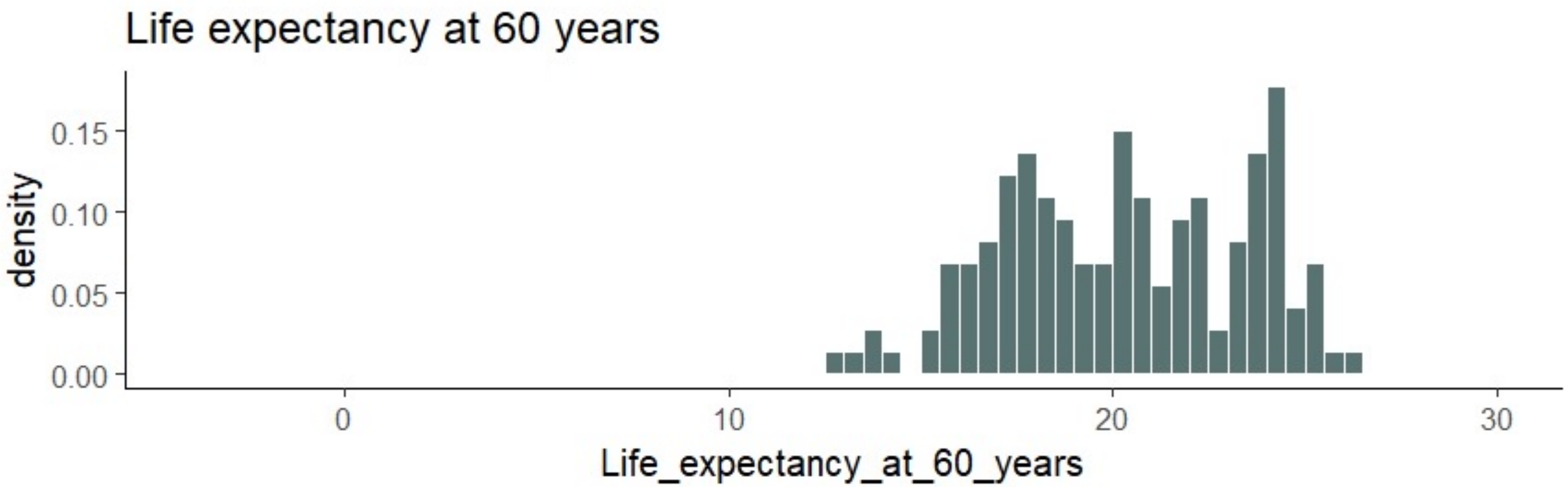
- How are the two indicators different?
  1. Units of measurement
  2. Ranges of variation

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

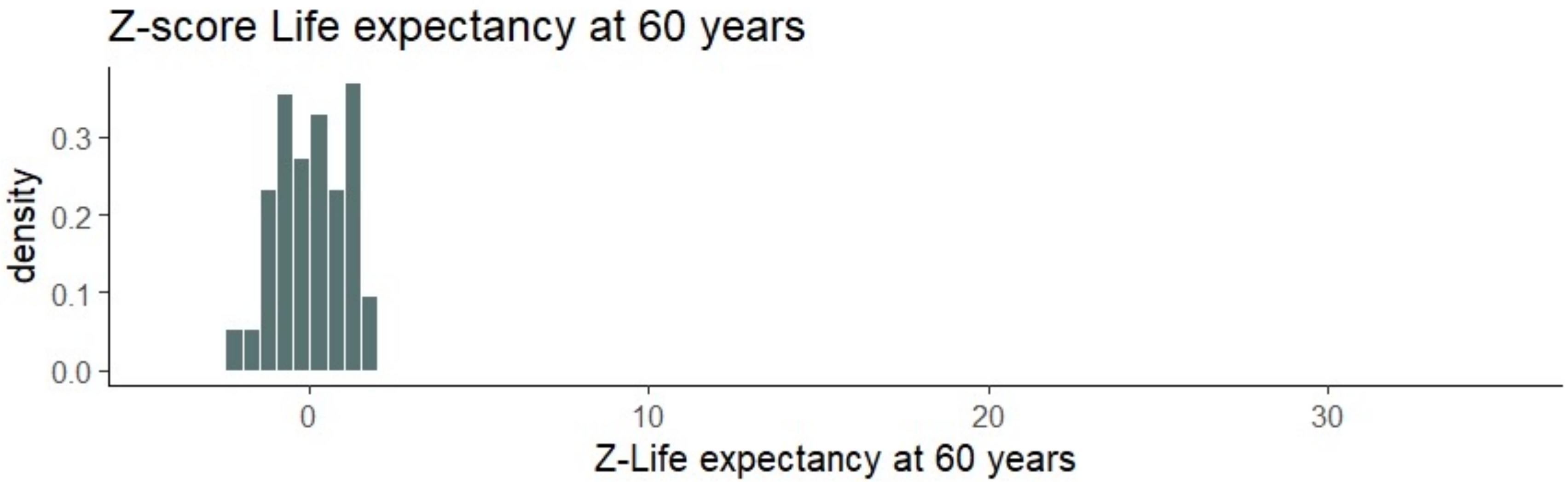
Z-score effects

- Unit of measurement
- I ~ $\mu = 0$, $\sigma^2 = 1$
- Variation range: not equal
- Extreme values: no adjustments
- Distribution: no adjustments

1. Z-score



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



$\boxed{\mu = 0 , \sigma^2 = 1}$

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]**

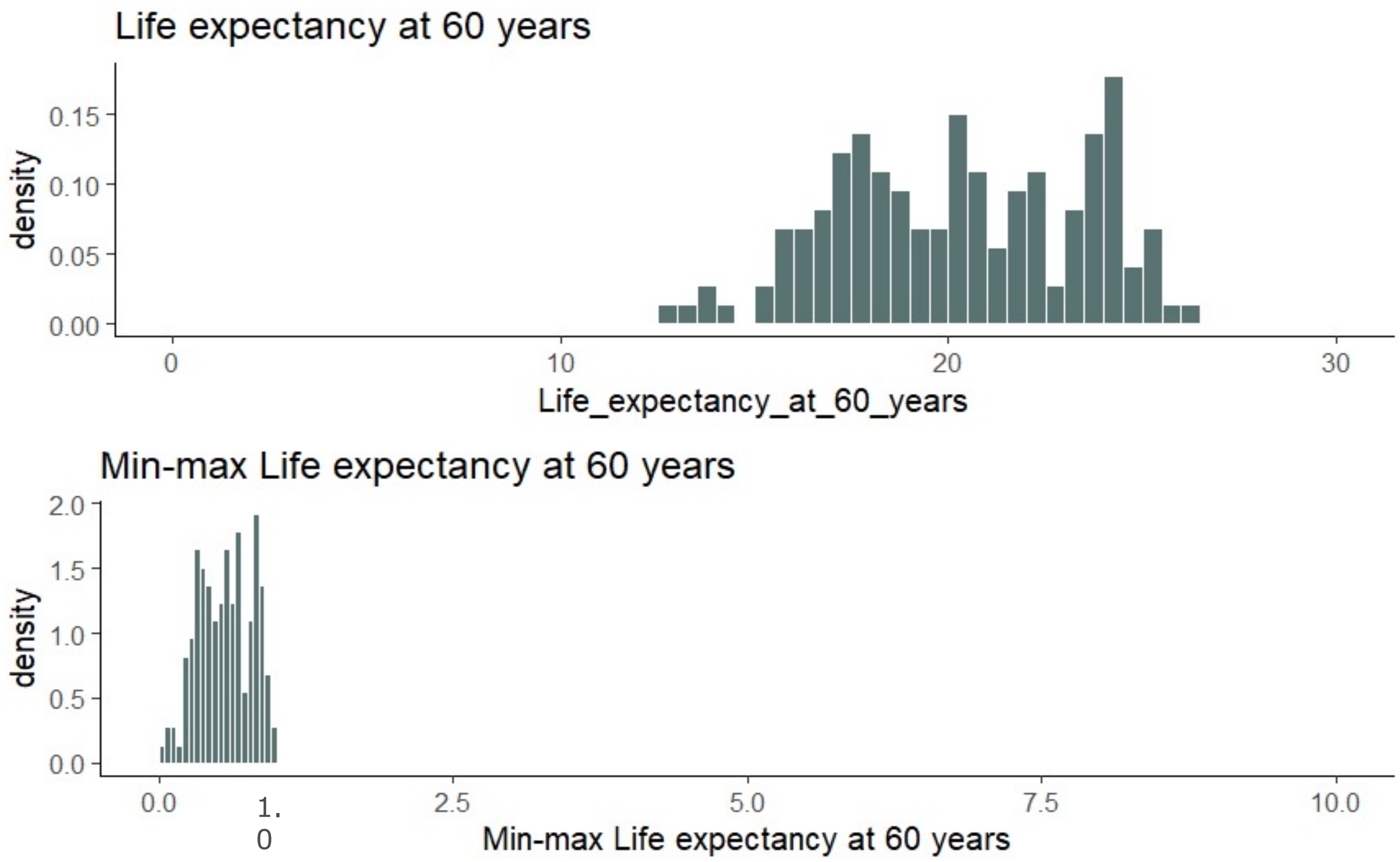
- How are the two indicators different?
  1. Units of measurement
  2. Ranges of variation

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

Min-max effects

- Unit of measurement
- μ, σ^2 not equal
- Variation range: [0, 1]
- Extreme values: no adjustments
- Distribution: no adjustments

2. Min-max



variation range = [12.5, 24.1]

variation range = [0, 1]

Rescaling eases communication, but be careful to aggregation formulas

3. Distance to the reference unit

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

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

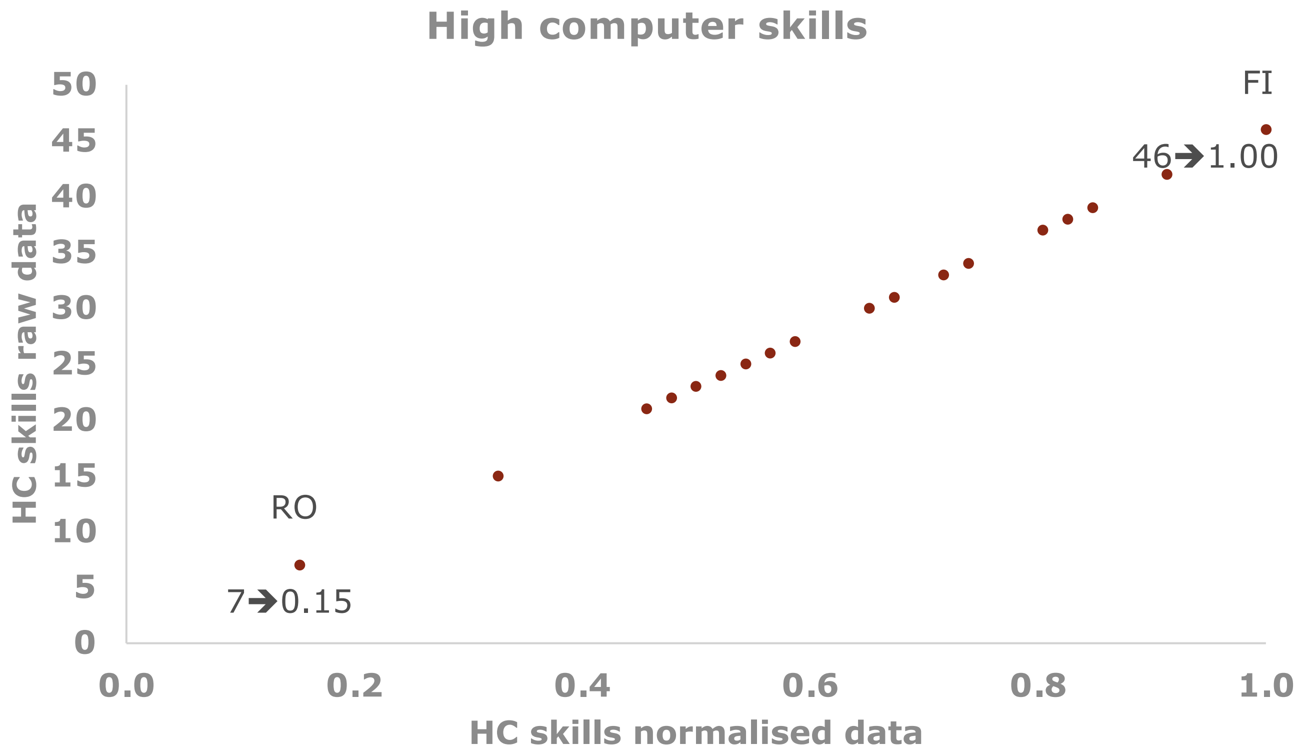
- group leader, external benchmark or hypothetical country, city etc. (target to be reached in a given timeframe)
- average (eg., EU28, world)

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

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

European Commission

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 (unless)
- Extreme values: no adjustments
- Distribution: no adjustments

4. Categorical scale

Ordinal scales

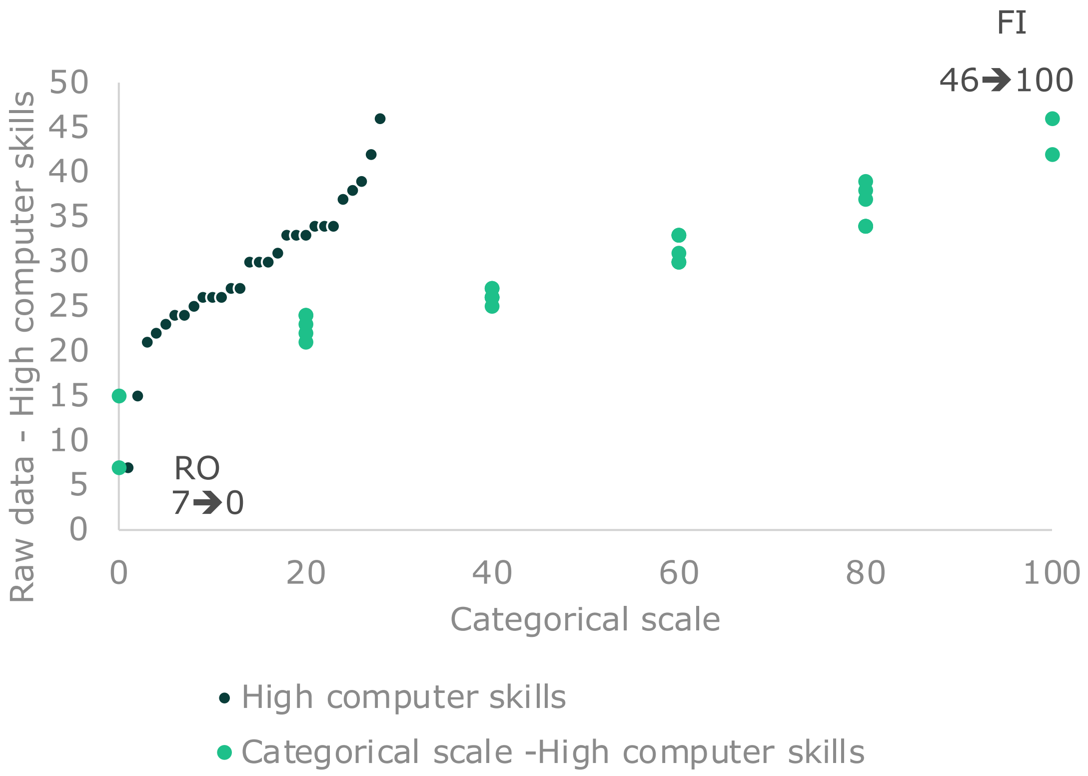
Indicator score based on categories: e.g. school grade, qualitative evaluations

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

Numerical scales

- Categories lie on a variation range portion
- Can be based on the percentile of the distribution of the indicator across countries (and/or time)
- Needed: Justify the choice of intervals and scores

4. Categorical scale



Indicators after categorical scaling

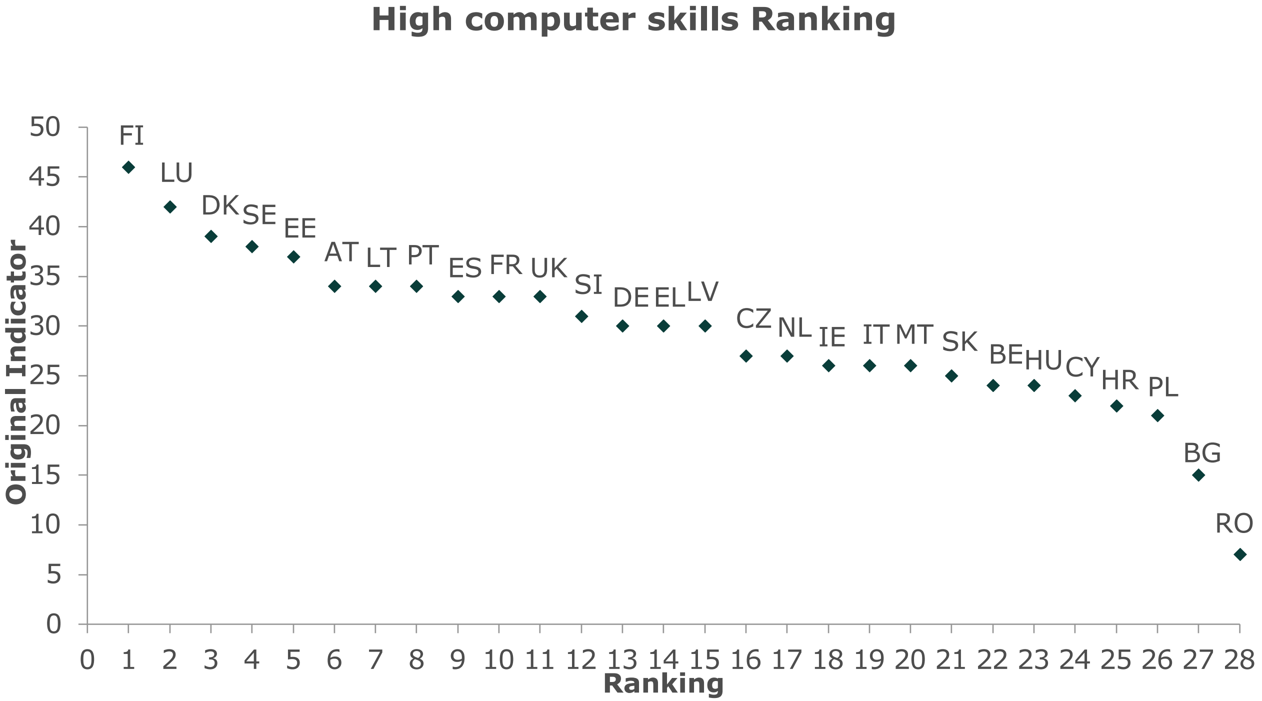
$$
I_{q,c}^{t} = \begin{cases} 0 & if \ p^{0} \leq x < p^{10} \\ 20 & if \ p^{10} \leq x < p^{25} \\ 40 & if \ p^{25} \leq x < p^{50} \\ 60 & if \ p^{50} \leq x < p^{75} \\ 80 & if \ p^{75} \leq x < p^{90} \\ 100 & if \ p^{90} \leq x \leq p^{100} \end{cases}
$$

Categorical scale effects

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

European Commission

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 $\sigma^2 = 65.25$
Immune to extreme values
Distribution: Uniform

European Commission

Normalization methods sum up

Normalisation effects	Quantile empirical distribution/Ranking	Categorical scale	Z-score	Min-max	Distance to a reference country
	Normalisation methods				
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				

European Commission

Step 4 key messages

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

Converting data onto a common **scale**

Prepare the data for the aggregation step (or comparison in dashboard)

How do we normalise data?

Five **normalisation methods➔** choice coherent with **data structur**e and **conceptual framework**

Check alternative normalisation within uncertainty/sensitivity analysis

European Commission

Thank you

pablo.depedraza@ec.europa.eu | jrc-coin@ec.europa.eu

composite-indicators.jrc.ec.europa.eu

European Commission