European Commission

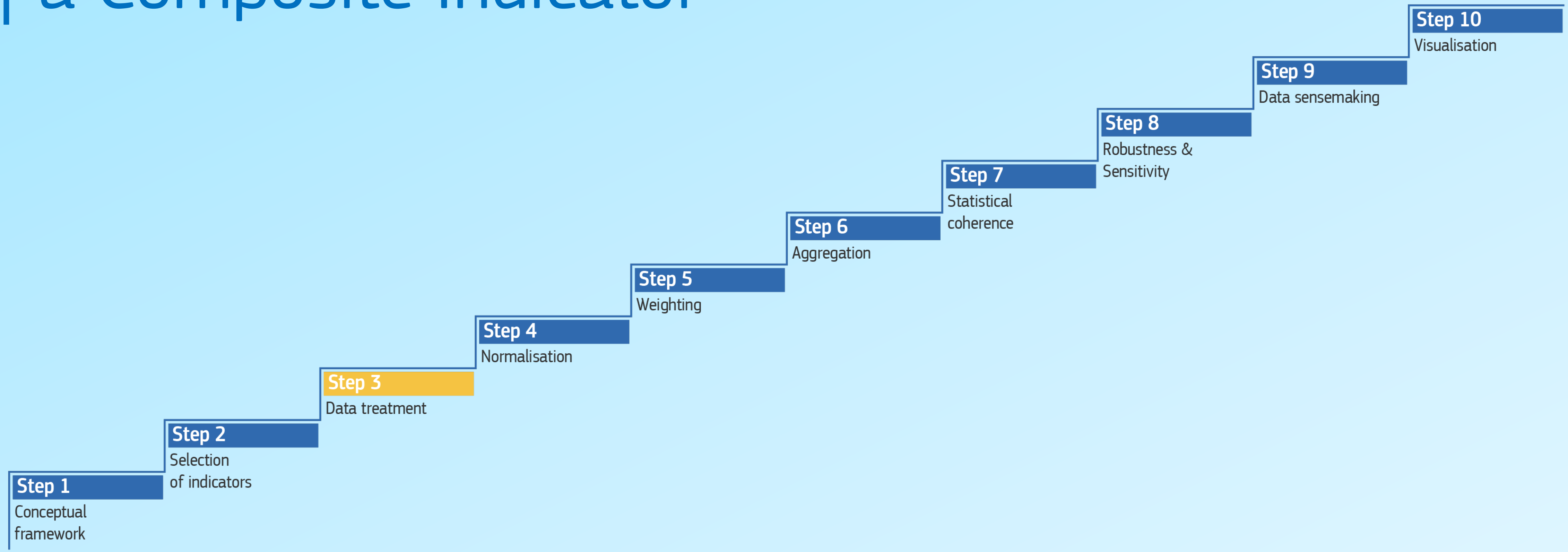
Step 3
Data treatment

18th JRC Annual training on Composite Indicators and Scoreboards

Marcos Dominguez-Torreiro

Joint Research Centre

10 STEPS to build a Composite Indicator



European Commission

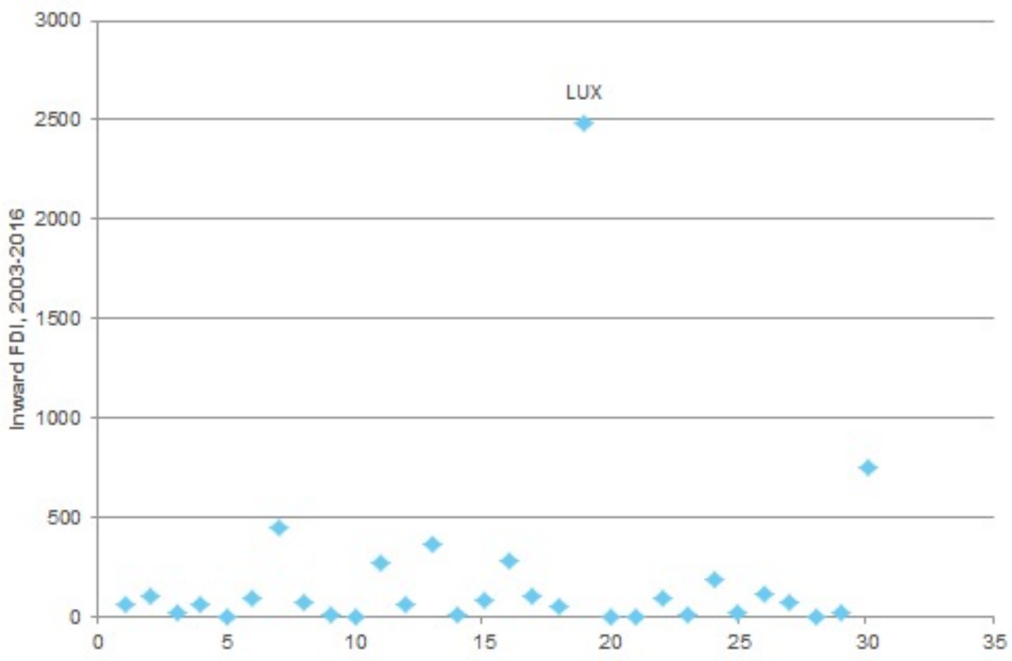
Outliers - outline

- Definition
- Identification
- Implications for CIs
- Treatment
- Takeaways

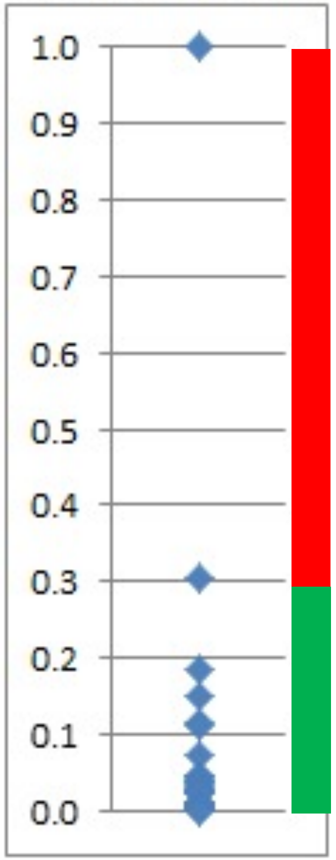
Definition

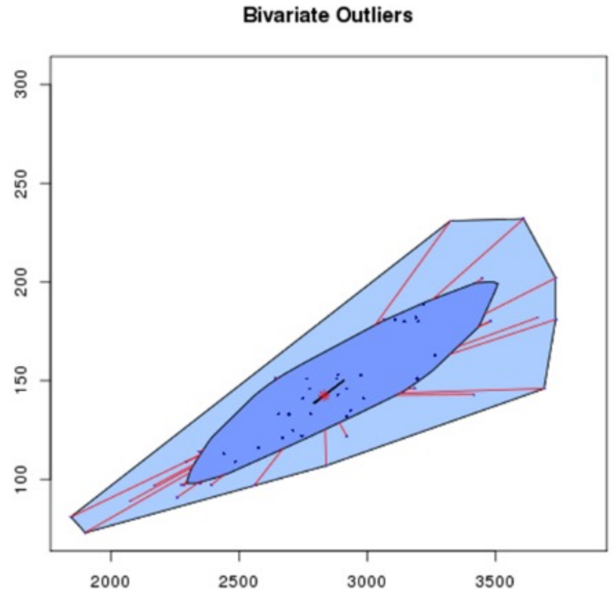
- (1) ***Outlier-univariate***: an extreme value of an indicator, i.e. an observed value that deviates markedly or stands apart from the rest; (2) ***Outlier-multivariate (e.g. bivariate)***: an unusual combination of indicator values which falls at the edge of the cloud of data-points (as shown on a scatterplot)
- Outliers could result from either ***heavy-tailed distribution*** of values in the population / phenomenon captured by the indicator or ***measurement errors***

Identification (I): plot the data



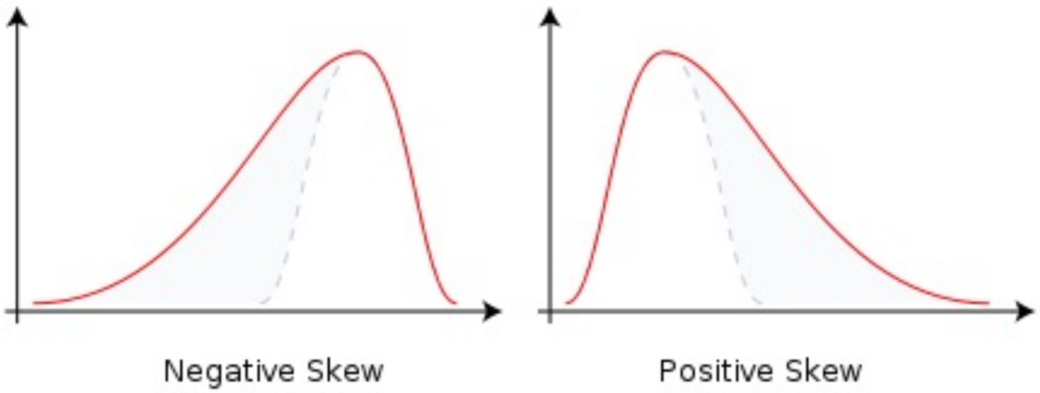
(min-max normalised data)





Identification (II): critical values of skewness and kurtosis

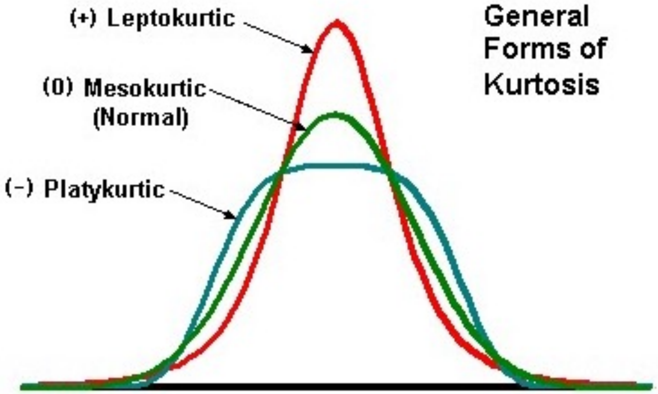
Presence of outliers in a variable if **|skewness| > 2 & kurtosis > 3.5**



Skewness: measure of the asymmetry of a distribution

(+) fatter tails

(-) thinner tails



Kurtosis: measure of the weight of the tails relative to the centre of the distribution ("tailedness" of the distribution)

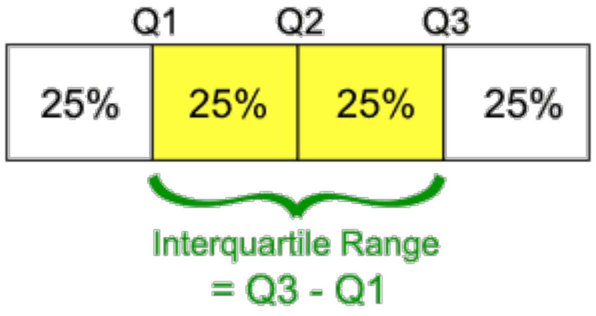
Identification (III): others

- Converting variable to ***z-scores:*** $z_i = \frac{x_i - \mu}{\sigma}$

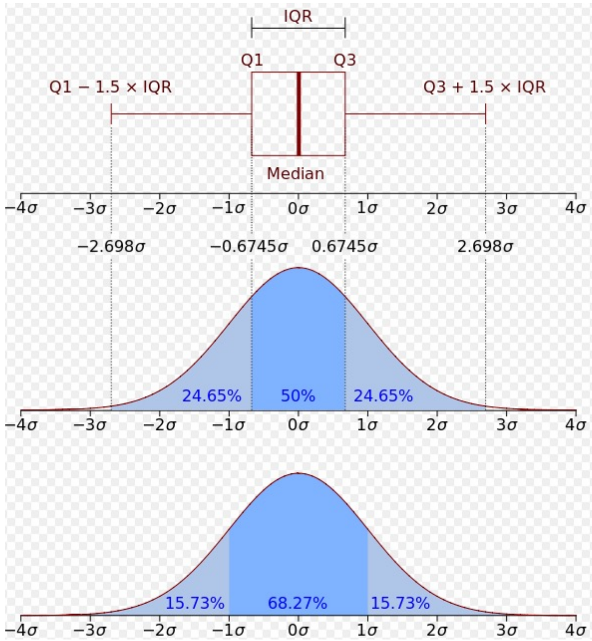
  - small sample size (80 or fewer obs.): a case is an outlier if $|z_i| \geq 2.5$

  - larger sample size (more than 80 obs.): a case is an outlier if $|z_i| \geq 3$

- A case is an outlier if outside ± ***1.5 * Interquartile range***



lower boundary $Q_1 - 1.5(Q_3 - Q_1)$

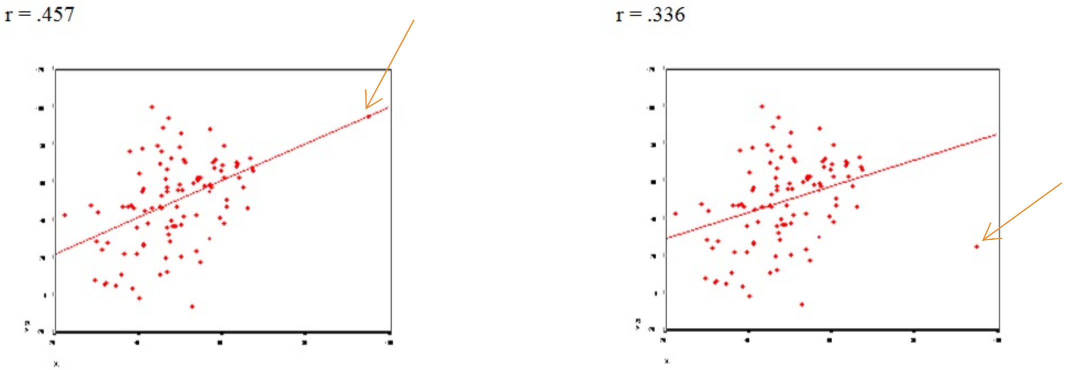


upper boundary $Q_3 + 1.5(Q_3 - Q_1)$

Implications for CIs

Indicators with outliers (heavy-tailed distributions) depart from the ideal of normality (bell-shaped distributions). This may have an impact on:

- (1) descriptive statistics: means and standard deviations/variances unrepresentative summary measures
- (2) statistical coherence analysis: biased pairwise correlations



- (3) normalisation step (e.g. min-max): i) large portion of the theoretical range of normalised values might remain empty; ii) could result in highly unequal variances across normalised indicators & unbalanced influence on aggregate scores

Treatment (I): winsorisation

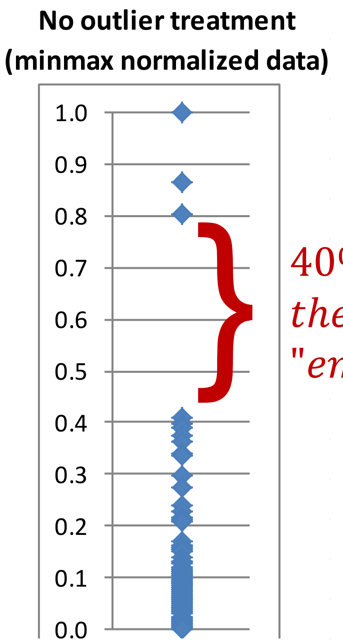
Winsorisation aims to mitigate the impact of extreme values by ***treating only potentially problematic observations*** (i.e. keep them but not take them too literally)

- “Capping” numeric outliers so they fall precisely at the edge of the main distribution (i.e. make them closer to the other observed values)
- Values distorting the indicator distribution are ***replaced by the next highest*** (pos. skew) / lowest (neg. skew) ***value***, up to the point where skewness *or* kurtosis enter within our desired ranges (i.e. |skewness| < 2 *or* kurtosis < 3.5).
- Winsorization does NOT preserve order relations for the units treated

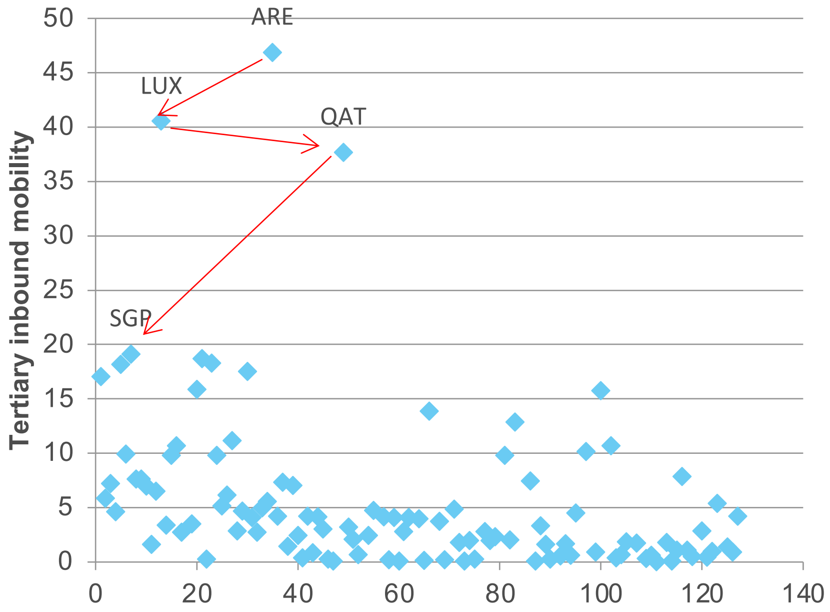
European Commission

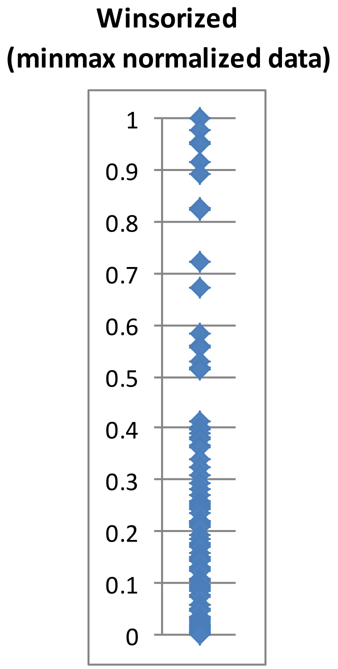
Treatment (I): winsorisation - example

Winsorisation would treat 3 data points (3 outliers)



40% of the scale is "empty"





After winsorization, data-points are more homogeneously spread across the normalised scale

	Raw data	Winsorized
Skewness	3.1	1.4
Kurtosis	11.6	1.0

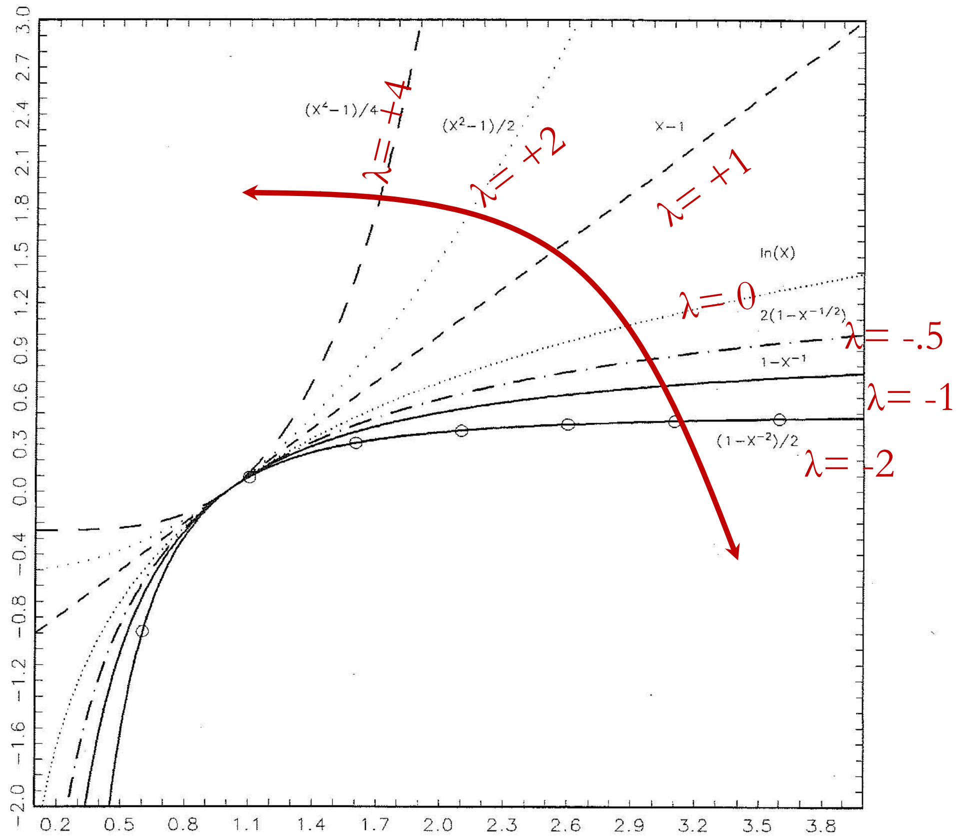
European Commission

Treatment (II): log-transformations (Box-Cox)

Box-Cox transformations:

$$\phi_\lambda(x) = \begin{cases} \dfrac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log x & \text{if } \lambda = 0 \end{cases}$$

$$x > 0$$



Treat and transform **all the values** in the indicator

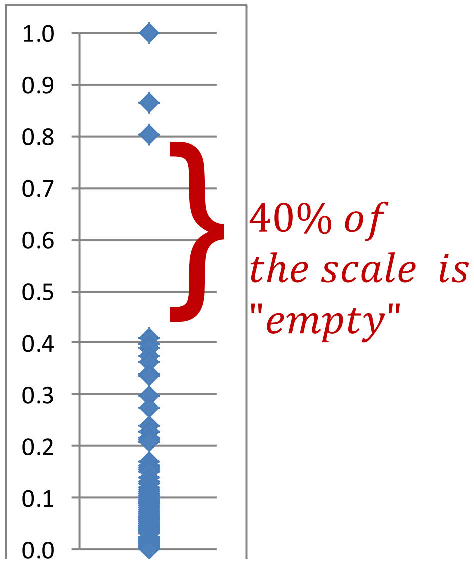
Recommended as an alternative to winsorisation **in case of identifying a high number of outliers (e.g. 5 or more)**

Log-transformation (λ=0): "Long-tail" distributions often resemble normal distributions when log-transformed

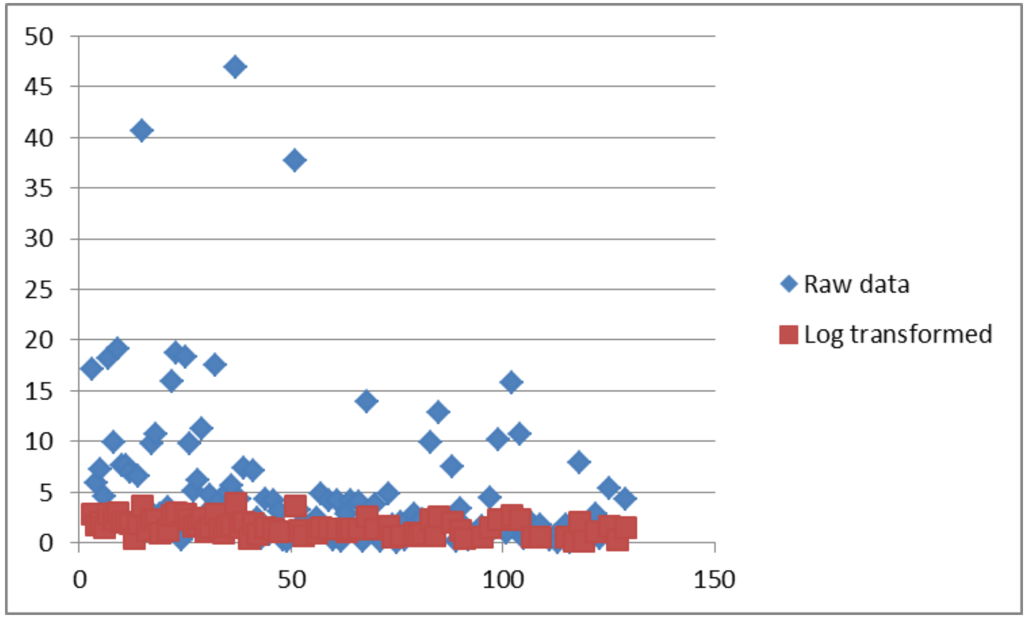
European Commission

Treatment (II): log-transformations - example

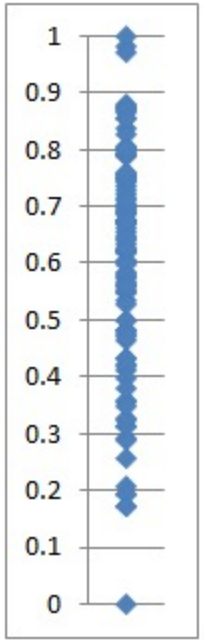
No outlier treatment
(minmax normalized data)



Log-transformation treats ("compacts") all data-points



Winsorized
(minmax normalized data)



After log-transformation data-points spread more homogeneously across the normalised scale

	Raw data	Winsorized	Log transformed
Skewness	3.1	1.4	-0.6
Kurtosis	11.6	1.0	0.1

European Commission

Outliers – takeaways

- ***Look into the data*** and search for potential outliers

- ***Some identification methods are more "invasive" than others***, i.e. tend to identify more cases as potential outliers

(-) -------- Skewness & Kurtosis -------- Z-scores -------- IQR -------- (+)

(less invasive) (more invasive)

- ***Outliers*** often ***spoil/bias/severely affect*** basic ***descriptive statistics*** (mean, variance) ***and correlation coefficients***, thus causing misinterpretations

- Every ***outlier treatment*** method ***alters the original data*** -> Ponder the choice of transforming the data only if necessary (e.g. not needed if normalisation methods rely on rankings/orderings)

- ***Avoid*** as much as possible tailored-made solutions (i.e. ***using different methods to treat different indicators*** across the framework); ***consider assessing the impact of different outlier treatment strategies*** in the uncertainty-sensitivity analysis

Outliers – takeaways (II)

- Bottomline: ***treat as few observations as possible to render the indicator framework ready for <u>normalisation</u>, <u>aggregation</u>, and <u>statistical coherence analysis</u>***

- Our suggestion: identification using critical values of ***skewness & kurtosis*** (more conservative) + treatment using ***winsorisation*** (only outliers are treated) if less than 5 outliers or ***log-transformation*** (all observations are treated) if 5 or more outliers

Missing data - outline

- Definition and identification
- Implications for CIs
- Treatment
- Takeaways

Definition and identification

- Missing data corresponds to a situation in which some of the indicator values for some of the units in our dataset are not reported (deliberately) or not available for analysis

Rows: **units** (cases or observations)

	Var1	Var2	Var3	Var4	Var5	Var6	Var7	Var8	Var9	Var10	Missing per country
C1		x			x						2
C2	x		x					x	x	x	5
C3	x	x				x					3
C4											0
C5											0
C6	x		x		x	x		x		x	6
C7											0
C8											0
C9	x			x		x	x		x	x	6
C10											0
C11						x					1
C12											0
C13											0
C14			x		x						2
C15						x					1
C16			x								1
C17											0
C18		x				x					2
C19											0
C20						x					1
Missing per indicator	4	3	4	1	3	7	1	2	2	3	10

Columns: **indicators** measured for each unit

Definition and identification – underlying mechanisms

MCAR – Missing completely at random

- missingness <u>does not depend on the values in the data matrix</u>, missing or observed
  - observed units are random subsample of original sample - ***values missing randomly***
    - ***e.g.*** survey respondents ***roll a die*** and answer the "earnings" question if "6" shows up *(unrelated to any variable in the data matrix)*

MAR – Missing at random

- missingness <u>depends on observed components</u> and not on the missing components
  - observed units not random sample of original sample - ***values missing systematically***
  - potentially unbalanced data in categories/subpopulations (i.e. contingent emptiness of cells)
    - ***e.g***. ***missing income related to ethnicity and education*** (***fully recorded*** in the data set)

NMAR – Not missing at random

- missingness <u>depends on missing values</u> in the data matrix (either missing values of variable itself or other partially unobserved variables)
  - observed units not random sample of original sample - ***values missing systematically***
    - ***e.g***. ***missing income related to income level***

European Commission

Relevance for CIs

- Missing data (treatment) will have an impact on indicator variances and correlations
- N.B.: “hands-off” approach (not to impute) is equivalent to a **“shadow imputation”** (i.e. unnoticed data treatment == imputing mean-row of normalised indicators in each pillar/dimension when calculating aggregate scores

Country	Pupil-teacher ratio, secondary	Tertiary enrolment	Graduates in science and engineering	Mean
DNK	11.3	81.5	20.4	**37.7**
SGP	14.9	69.8	N/A	**42.4**
FIN	12.8	87.3	27.9	**42.7**
DEU	12.1	68.3	N/A	**40.2**
IRL	N/A	77.6	23.8	**50.7**
KOR	15.6	95.3	31.9	**47.6**
ISL	N/A	81.3	15.6	**48.5**

Country	Pupil-teacher ratio, secondary	Tertiary enrolment	Graduates in science and engineering	Mean
DNK	11.3	81.5	20.4	**37.7**
SGP	14.9	69.8	42.4	**42.4**
FIN	12.8	87.3	27.9	**42.7**
DEU	12.1	68.3	40.2	**40.2**
IRL	50.7	77.6	23.8	**50.7**
KOR	15.6	95.3	31.9	**47.6**
ISL	48.5	81.3	15.6	**48.5**

Note that pillar averages based only on observed values are identical to pillar averages after imputing row mean values

Treatment (I) – mean imputation

Unconditional mean imputation (by column/normalised indicator)

Country	Pupil-teacher ratio, secondary	Tertiary enrolment	Graduates in science and engineering
DNK	11.3	81.5	20.4
SGP	14.9	69.8	N/A
FIN	12.8	87.3	27.9
DEU	12.1	68.3	N/A
IRL	N/A	77.6	23.8
KOR	15.6	95.3	31.9
ISL	N/A	81.3	15.6
Mean	**13.3**	**80.2**	**23.9**

Country	Pupil-teacher ratio, secondary	Tertiary enrolment	Graduates in science and engineering
DNK	11.3	81.5	20.4
SGP	14.9	69.8	**23.9**
FIN	12.8	87.3	27.9
DEU	12.1	68.3	**23.9**
IRL	**13.3**	77.6	23.8
KOR	15.6	95.3	31.9
ISL	**13.3**	81.3	15.6
Mean	**13.3**	**80.2**	**23.9**

Pros: simple, relies on the observed values from the same variable.
Cons: correlations are affected; **variances will be typically underestimated** (as missing values are imputed with 'central values').

Treatment (II): k-nearest neighbours (kNN) algorithm

Replaces missing values for a nonrespondent (***recipient***) with observed values from a respondent (***donor***) "similar" (based on distance metrics) to the recipient with respect to observed characteristics

Step 1. Compute the distance / similarity between recipient and potential donors

Manhattan (absolute) distance preferred option if high differences shall not be overweighed; alternative metrics: *Euclidean* (square), Mahalanobis, etc.

Normalized values

$$d_{ij} = \sum_k \left| {}_k x_i - {}_k x_j \right| \quad \text{Manhattan}$$

Index k goes through all the indicators jointly observed on units i and j

$$d_{ij} = \sqrt{\sum_k \left({}_k x_i - {}_k x_j \right)^2} \quad \text{Euclidean}$$

Country	Expenditure on education	Government expenditure on education per pupil, secondary	School life expectancy
SGP	2.9	16.7	12.8
DEU	4.9	23.7	17.3
IRL	5.3	26.0	19.0
KOR	4.6	23.4	16.6
ISL	7.8	18.3	19.6
LUX	4.1	19.4	13.9
JPN	3.8	25.1	15.4
FRA	5.5	26.8	16.3
HKG	3.3	20.4	N/A

Country	Distance	
	Euclidean	Manhattan
SGP	3.68	4.02
DEU	3.74	5.02
IRL	6.00	7.70
KOR	3.31	4.37
ISL	4.97	6.55
LUX	1.30	1.83
JPN	4.79	5.27
FRA	6.85	8.72
HKG	0	0

closest country (green)
2nd closest country (blue)
3rd closest country (orange)

Step 2. The imputed value for the recipient is the observed on the most similar unit, or the mean value of the *k*-closest units

Number of neighbours	Distance type	Imputed value	
1NN	Euclidean	13.9	
1NN	Manhattan	13.9	
2NN	Euclidean	15.3	[= (13.9+16.6)/2]
2NN	Manhattan	13.4	[= (13.9+12.8)/2]
3NN	Euclidean	17.4	[= (13.9+16.6+12.8)/3]
3NN	Manhattan	17.4	[= (13.9+12.8+16.6)/3]
...	...	...	

Pros: uses actual values (easy to communicate); does not impose a structure on relationships between variables.
Cons: might be computational-intensive; might reduce variance, but typically less than mean substitution.

Treatment (III): expectation-maximisation (EM) algorithm

- Likelihood based approaches: defining a (parametric) model for the observed data and estimating those parameters by Maximum Likelihood (ML)
- ***EM:*** powerful and reliable ***iterative procedure to compute ML estimates from incomplete data sets*** (i.e. missing values filled in with ML estimates based on available data)

- Each iteration of EM until convergence consists of two-steps:
  - ✓ ***M-step***: ML estimation of underlying parameters as if there were no missing data (i.e. maximizing likelihood of the “expected complete-data”)
  - ✓ ***E-step***: calculates conditional expectation of missing data given observed data and current estimated parameters

Pros: appropriate under MAR conditions – often reduced bias even with data NMAR
Cons: highly **dependent on strong correlations (>= 0.6)**; **computational-intensive;** difficult to communicate

European Commission

Treatment (IV): expectation-maximisation & multiple imputation (MI)

- Implies creating ***m-complete data sets*** by imputing *m*-values for each missing cell; the *m*-estimates can be combined (e.g. averaging them)

- Explicitly ***accounting for the uncertainty about the values generated and imputed;*** allow appropriate assessment of imputation uncertainty (i.e. statistical inference on variances, s.e.'s and confidence intervals of point estimates)

e.g. Amelia II software (https://gking.harvard.edu/amelia)

Combines EM algorithm with a bootstrap approach

Special features for time-series-cross-section data

Works from the ***R*** command line or via a graphical user interface

Pros: often reduced bias even with data NMAR; explicitly account for imputation uncertainty

Cons: highly dependent on strong correlations; computational-intensive; difficult to communicate

Missing data – takeaways

- Pre-imputation step: look into the data and try to identify/reflect on the **patterns** of missingness and **coding errors** (e.g. missing data coded as “0”, “-1”, “999”, etc.)

- Imputations often unreliable if ***data set*** contains more than ***1/3 - 40% of missing values***

  @indicator-level: at least ***50% of units*** should have valid data for that indicator – otherwise drop indicator and search for an alternative proxy

  @unit-level: at least ***65-75% of the indicators*** for the unit should have valid data (***apply this threshold at pillar (dimension) level and not only at framework level!***) – otherwise exclude unit from pillar/index score calculations (Step 6: Aggregation)

- ***Consider pros and cons*** of each method and try to ***avoid using different methods*** for different indicators

European Commission

Missing data – takeaways (II)

- ***Imputation algorithms (kNN, EM) should*** not ***be run*** for the whole dataset at once, but ***separately by pillar/dimension*** (i.e. use related variables to improve accuracy/predictive power)

- ***EM*** algorithm performance is dependent on the correlation structure; ***correlations should be strong enough*** (>= 0.6), otherwise you can't make a good prediction using EM!

- ***kNN*** algorithm identifies donors using distances (***not based on correlations***); handy option when correlations are poor

- When using ***kNN***, always ***search for "close" donors*** by keeping the number of selected neighbours low (e.g. k = 2 or 3)

- ***Imputation algorithms are usually applied after normalisation*** for practical reasons: (1) EM: min-max normalisation (e.g. 0-100) helps to easily spot out-of-bound imputed values; (2) kNN: having all data points in a common meaningful scale helps to give indicators the same influence when computing distances and identifying neighbours

European Commission

Missing data – takeaways (III)

- ***Ignoring missing values*** when calculating aggregate scores is nothing else than a subtle form of imputation (i.e. ***"shadow imputation"***); remember that we will be replacing the missing value for a unit with the mean normalised values for the other variables in the pillar!
- When constructing a composite indicator that will be used for benchmarking and monitoring performance across units, ***shadow imputation*** (by row) ***and mean imputation*** (by column) methods would ***provide incentives to not report low performance***
- Consider assessing the ***sensitivity of final rankings to different imputation methods*** (Step 8: Robustness & sensitivity)

Thank you

Marcos.DOMINGUEZ-TORREIRO@ec.europa.eu | jrc-coin@ec.europa.eu

composite-indicators.jrc.ec.europa.eu

European Commission