The European Commission's science and knowledge service

Joint Research Centre

European Commission

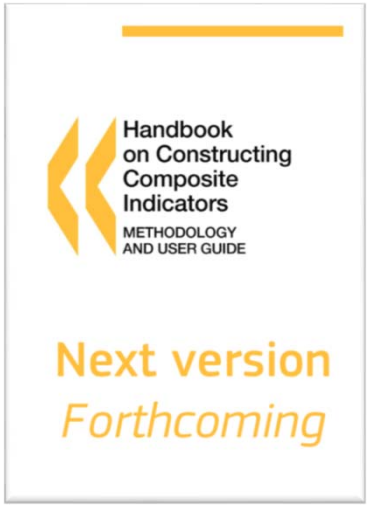
Step 3: The identification and treatment of outliers

Giacomo Damioli

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

European Commission

Ten steps



Step 1 Developing the framework

Step 2 Selection of indicators

Step 3 Data treatment

Step 4 Normalisation

Step 5 Weighting

Step 6 Aggregation

Step 7 Statistical coherence

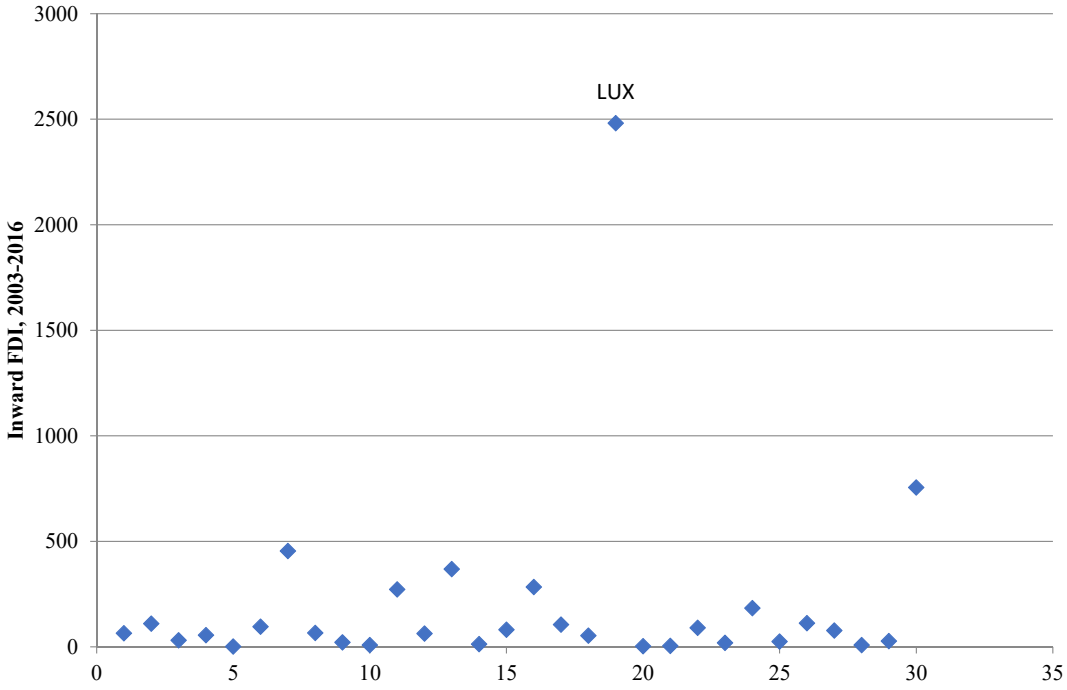
Step 8 Robustness & sensitivity

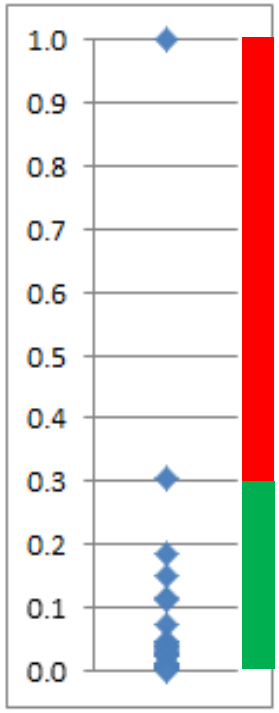
Step 9 Back to the data

Step 10 Visualisation

European Commission

Outliers – identification



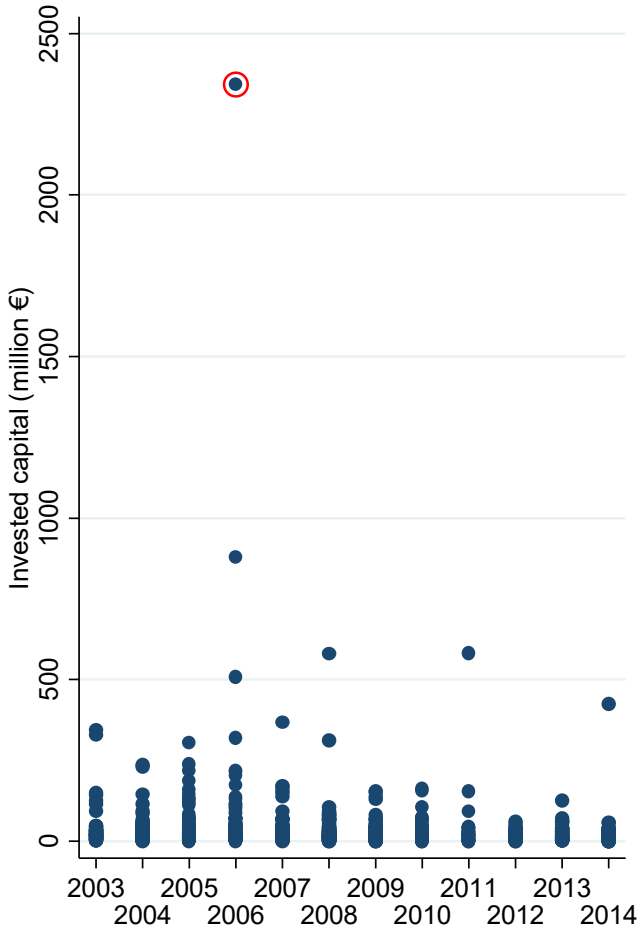


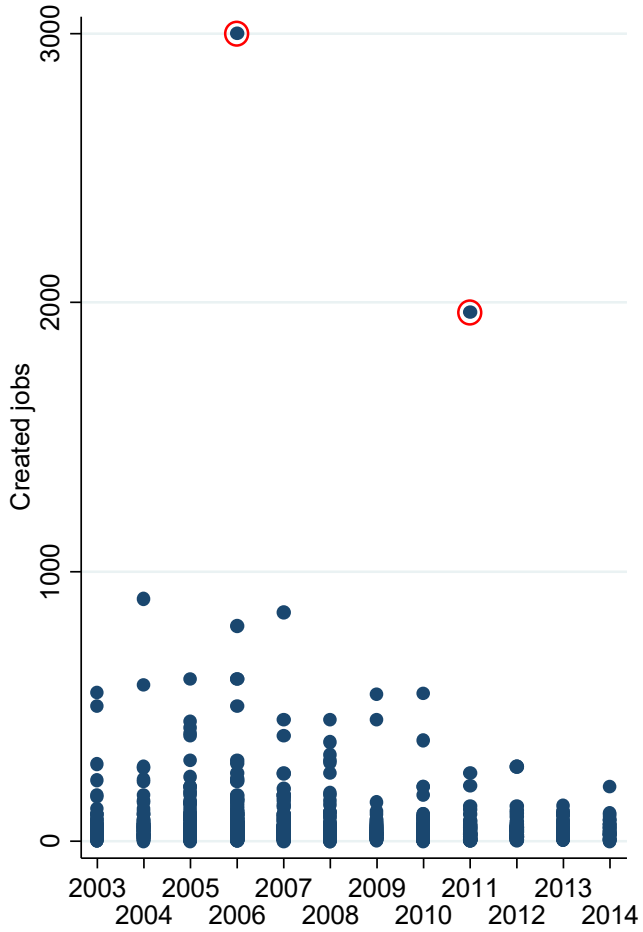
European Commission

Outline

- Definition and relevance
- Outlier identification
- Outlier treatment techniques

Outliers – what are they?





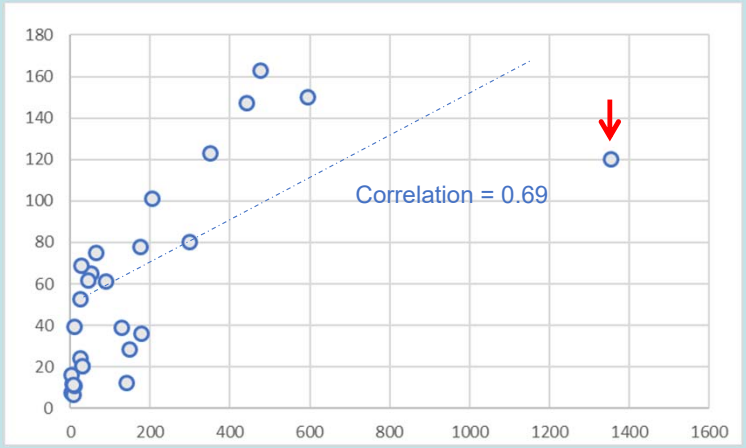
"*An outlier is an observed value that is so extreme (either large or small) that it seems to stand apart from the rest of the distribution*"

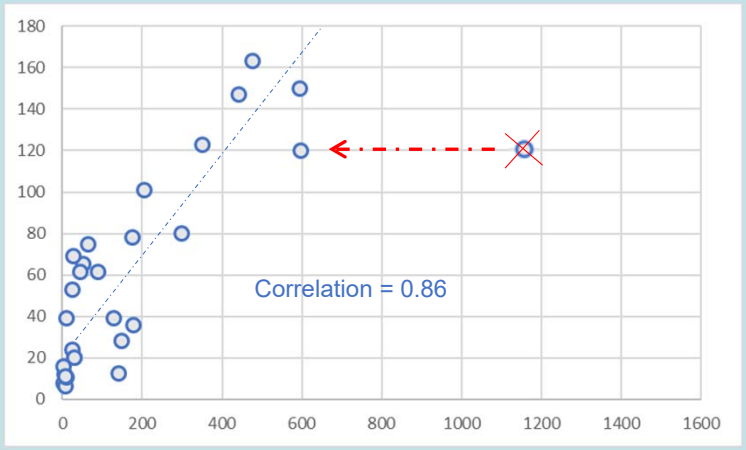
(Knoke & Mee, 2002)

European Commission

Outliers – why do we care about?

Outliers may spoil descriptive statistics (means, standard deviations, correlations …) and cause misinterpretations





Outliers – identification

Graphical/visual inspection

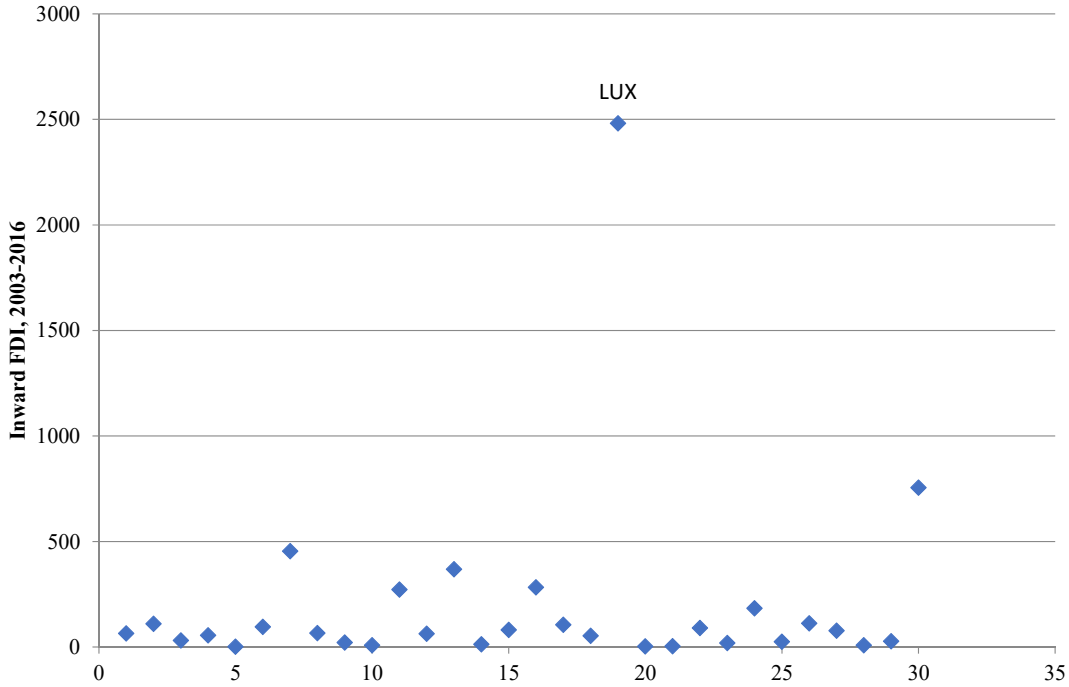
- Simply have a look at the data!

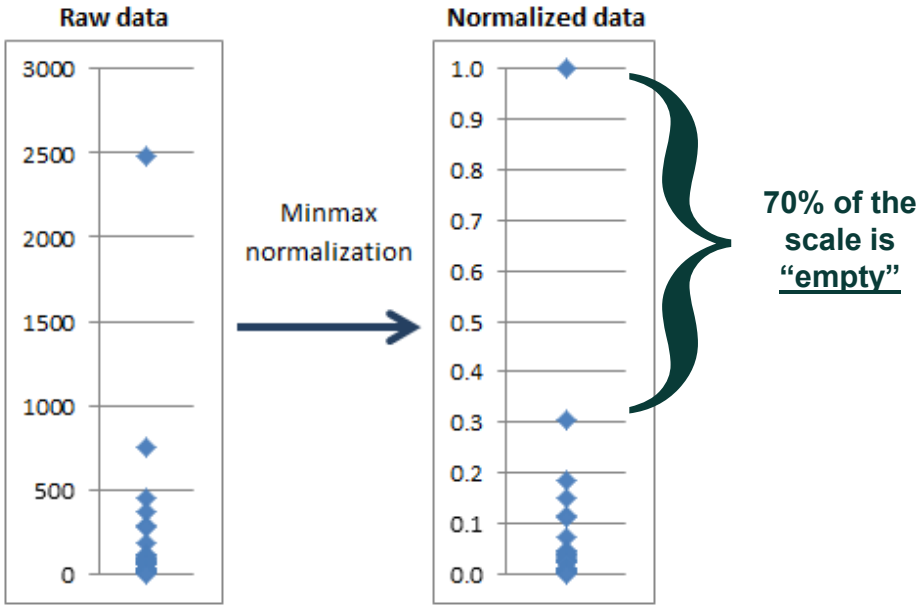
Statistical rules(-of-thumb ) based on

1) z-scores
2) Interquartile range
3) Skewness and Kurtosis

Outliers – identification

- Look at the data!





Ideally less than 20% of the scale should be "empty"

European Commission

Outliers – identification

- z-scores

Standardisation:

- put different variables on the same scale
- change original values into standard scores: $z_i = \frac{x_i - \mu}{\sigma}$

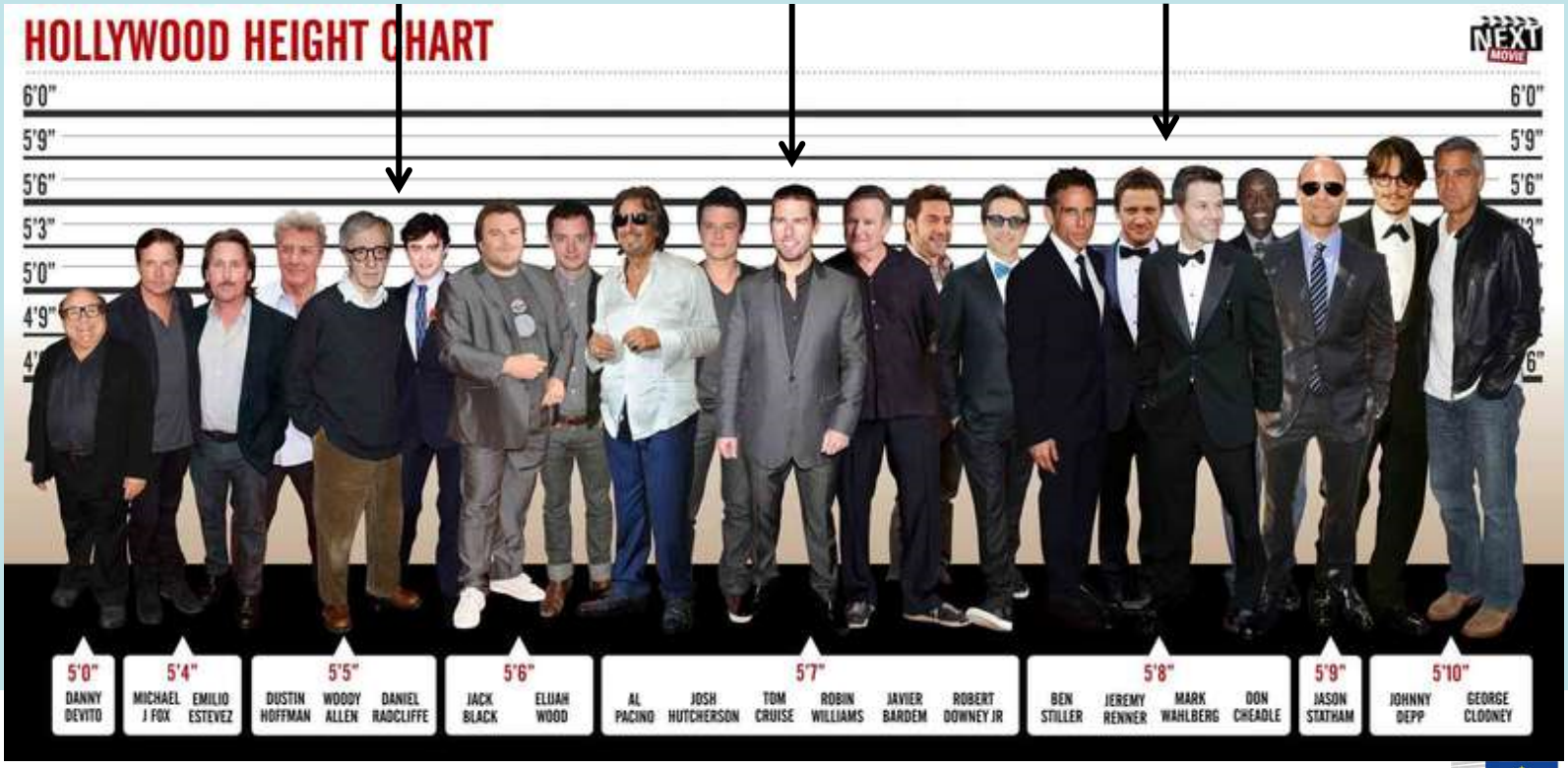
#1. A case x_i is an outlier if:

$$|z_i| \geq 2.5 \text{ for } N \leq 80$$

$$|z_i| \geq 3 \quad \text{for } N > 80$$

Aggarwal (2015)

Outliers – identification



European Commission

Outliers – identification

Quartiles: the 3 special points dividing a distribution into 4 intervals with proportion 0.25

The second quartile (Q_2) is the median

$$\text{Interquartile range (IQR)} = Q_3 - Q_1$$

Outliers – identification

- Interquartile range (IQR)

#2. A case x_i is an outlier if:

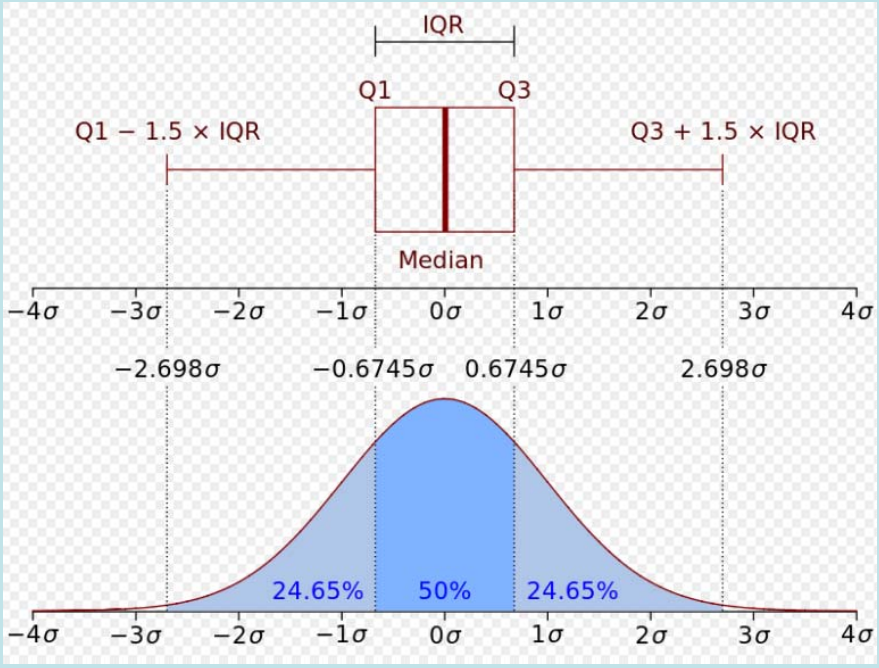
$$x_i \leq Q_1 - 1.5 \text{ x } IQR$$

or

$$x_i \geq Q_3 + 1.5 \text{ x } IQR$$

Aggarwal (2015)

Example with a Normal distribution $N(0,\sigma^2)$

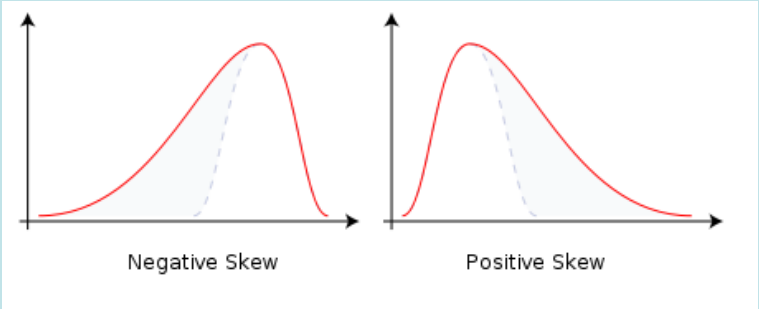


European Commission

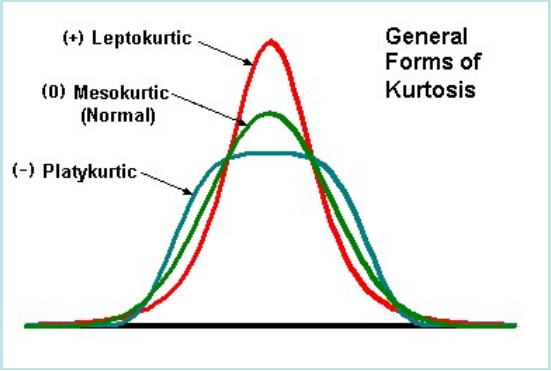
Outliers – identification

- Skewness and Kurtosis

Skewness: measure of the asymmetry of a distribution
(= 0 in the Normal distribution)



Kurtosis: measure of the thickness of the tails of a distribution
(= 3 in the Normal distribution)



(+) higher peak around the mean and fatter tails

(-) fatter around the mean and thinner tails

European Commission

Outliers – identification

- Skewness and Kurtosis

#3. A probability distribution contains one or more outliers if:

|skewness| > 2
and
kurtosis > 3.5

Groeneveld and Meeden (1984)

variable	min	p10	p25	mean	p50	p75	p90	max	sd	cv	skewness	kurtosis	N
Var_1	2,12	2,34	2,61	3,26	2,99	3,66	4,76	5,89	0,92	0,28	**1,17**	**3,63**	133
Var_2	1,91	2,79	3,16	3,90	3,68	4,43	5,40	6,19	0,97	0,25	0,52	2,54	133
Var_3	2,09	2,47	2,65	3,28	3,01	3,62	4,67	6,02	0,90	0,27	**1,28**	**4,07**	133
Var_4	2,20	2,57	3,04	3,62	3,41	4,06	4,94	5,90	0,86	0,24	0,71	2,84	133
Var_5	2,29	2,84	3,20	3,64	3,57	4,05	4,39	5,50	0,61	0,17	0,25	2,80	133
Var_6	2,70	3,10	3,53	4,14	4,16	4,68	5,18	6,01	0,77	0,19	0,17	2,34	133
Var_7	0,00	0,00	0,00	18,55	0,40	3,24	71,09	200,00	44,35	2,39	**2,74**	**9,89**	133
Var_8	1,70	2,46	2,81	3,76	3,54	4,61	5,66	6,21	1,17	0,31	0,53	2,21	133

Yet, even if skewness low …

Outliers – identification

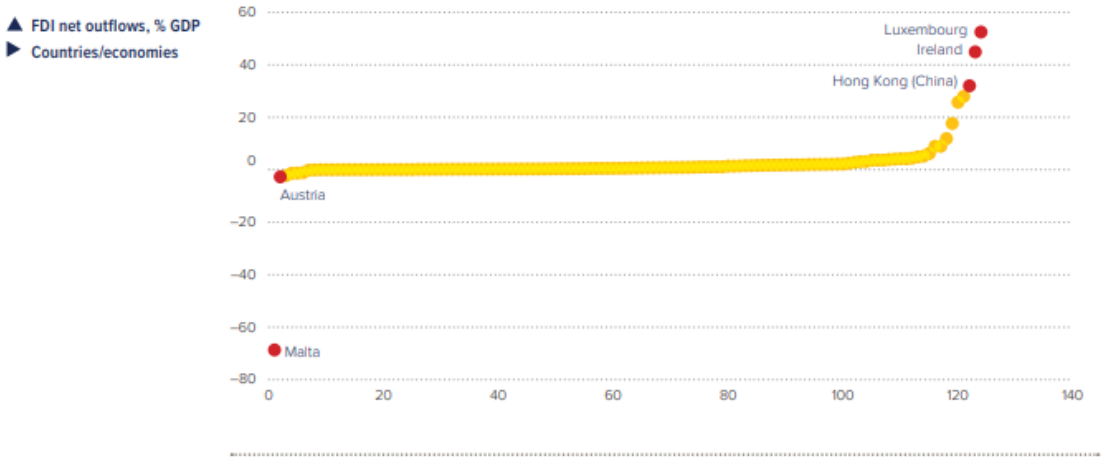
- Skewness and Kurtosis

Note well

- … if skewness low, and
- kurtosis very high (e.g. > 10)
- check for the presence of outliers on both sides of the distribution (as symmetrical outliers reduce skewness)
- Example: Malta in Global Innovation Index 2018

Figure 2.

Malta's outlier performance in FDI net outflows



Source: European Commission, Joint Research Centre, 2018.

Notes: Economies with the highest and lowest FDI outflow scores are highlighted. Skewness = −0.75; kurtosis = 28.16.

European Commission

Outliers – identification

How many outliers does a  identify?

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

(less invasive) (more invasive)

Outliers – JRC suggested strategy

- **JRC suggested outlier identification method**
  - Skewness & Kurtosis

    -> identifies less outliers, change original data as less as possible

Outliers – treatment

To treat or not to treat ….

Methods for outliers' treatment

- Winsorisation
- Trimming
- Box-Cox transformation

Outliers – treatment

Cautions:

- treatment alters original data, check implications on key statistics (mean, std. dev., corr.)
- treat only if really not avoidable

(example: with normalisation based on rankings – see Step 4 – no need to treat outliers)

- avoid as much as possible ‘tailor-made’ transformations (different for each indicator)

Outliers – treatment

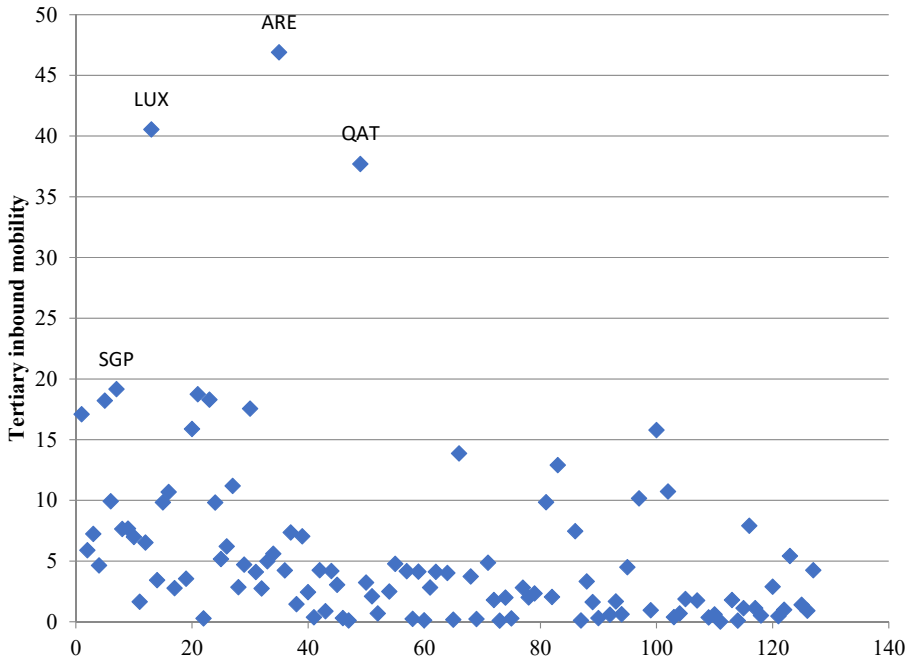
✓Winsorisation: modify outliers' values so to make them closer to other cases' values

✓ Trimming: exclude the outliers and treat them as missing values

✓ Transformations: change the values of all data-points, not only outliers (typical case log-transformation)

Outliers – treatment

An example from the Global Innovation Index 2017 (2.2.3 Tertiary inbound mobility)



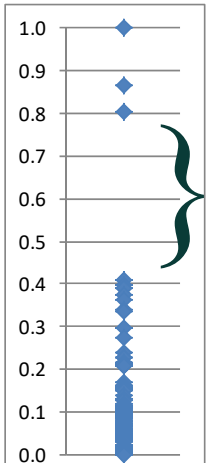
European Commission

Outliers – treatment

Winsorisation: outliers' values are assigned the next highest(/lowest) value (up to when the indicator's distribution satisfies the desired rule-of-thumb for identification)

Trimming: exclude the outliers and treat them as missing values

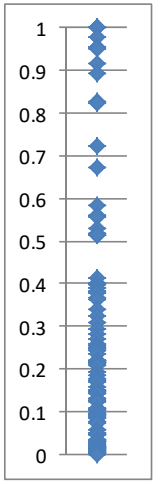
No outlier treatment
(minmax normalized data)



about 40% *of the scale is* "*empty*"

Country	Raw data	Winsorized	Trimmed
CHE	17.1	17.1	17.1
SWE	5.9	5.9	5.9
NLD	7.2	7.2	7.2
USA	4.6	4.6	4.6
GBR	18.2	18.2	18.2
DNK	9.9	9.9	9.9
SGP	19.2	19.2	19.2
FIN	7.7	7.7	7.7
DEU	7.7	7.7	7.7
IRL	7.0	7.0	7.0
KOR	1.7	1.7	1.7
ISL	6.5	6.5	6.5
LUX	40.6	19.2	
JPN	3.4	3.4	3.4
FRA	9.8	9.8	9.8

Winsorized
(minmax normalized data)

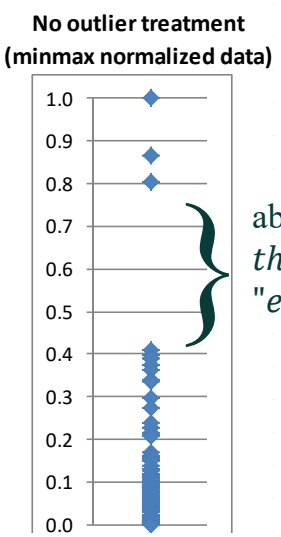


data-points spread homogeneously across the scale

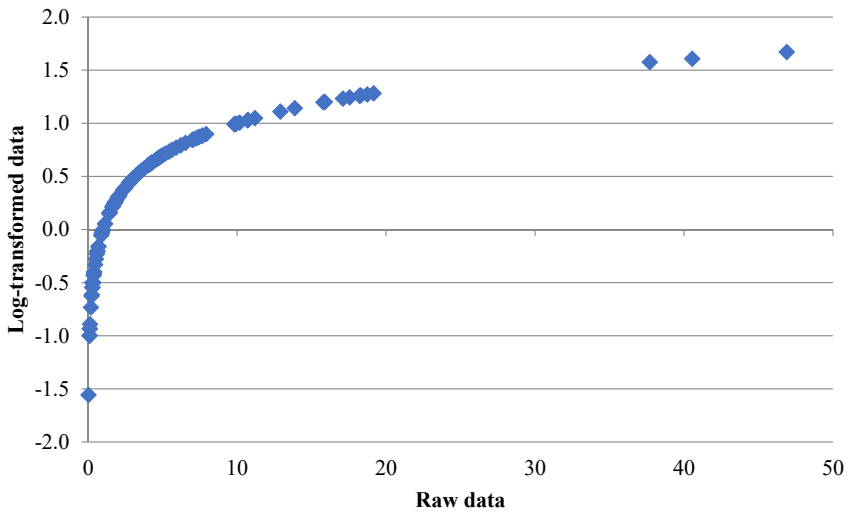
European Commission

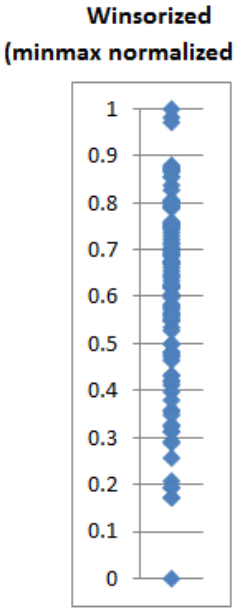
Outliers – treatment

Log-transformation: changes all data and "compacts" them



about 40% *of the scale is "empty"*





data-points spread homogeneously across the scale

European Commission

Outliers – treatment

Check out the consequences of the treatment!

	Raw data	Winsorised	Trimmed	Log transformed
Mean	5.8	5.1	4.7	0.4
Sigma (sd)	7.8	5.4	4.9	0.6
Skewness	3.1	1.4	1.5	-0.6
Kurtosis	11.6	1.0	1.5	0.1
Corr(2.2.3, 2.2.1)	0.09	0.20	0.27	0.28

Tertiary inbound mobility

Tertiary enrolment

Outliers – JRC suggested strategy

- **JRC suggested outlier identification method**
  - Skewness & Kurtosis

    -> identifies less outliers, change original data as less as possible

- **JRC suggested outlier treatment methods**
  - Winsorisation, with less than 5 outliers

    -> change outliers only

  - Log-transformation, with 5 or more outliers

    -> change all observations

Outliers – Key lessons

- Do always identify outliers
- The method based on simultaneous ‘anomalous’ values of Skewness and Kurtosis is the method for outlier identification that identifies the lowest number of outliers (less ‘invasive’)
- Think carefully if and how to treat the identified outliers
- When treating outliers, avoid as much as possible tailored-made treatment of different indicators
- Always assess the consequences of the treatment on the distribution of the treated indicator, as well as on its correlation with other indicators

Outliers –suggested reading

Suggested readings

- Aggarwal, C. C. (2015) Outlier analysis. In *Data mining* (pp. 237-263). Springer, Cham.
- Atkinson, A.C., Riani, M. & A. Ceriolin (2004) Exploring Multivariate Data with the Forward Search. Springer-Verlag.
- Ghosh, D., & A. Vogt (2012) Outliers: an evaluation of methodologies. American Statistical Association. Section on Survey Research Methods – JSM 2012.
- Groeneveld, R. A., & Meeden, G. (1984). Measuring Skewness and Kurtosis. *The Statistician* 33: 391–99.
- Grubbs, F. E. (1969) Procedures for detecting outlying observations in samples" *Technometrics* 11 (1): 1–21.
- Hawkins, D. (1980) Identification of Outliers. Chapman and Hall.
- Knoke, B. & P. Mee (2002) Statistics for social data analysis. Wadsworth Publishing.
- Tukey, John W (1977). Exploratory Data Analysis. Addison-Wesley.

THANK YOU

Any questions?

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