The European Commission's science and knowledge service

Joint Research Centre

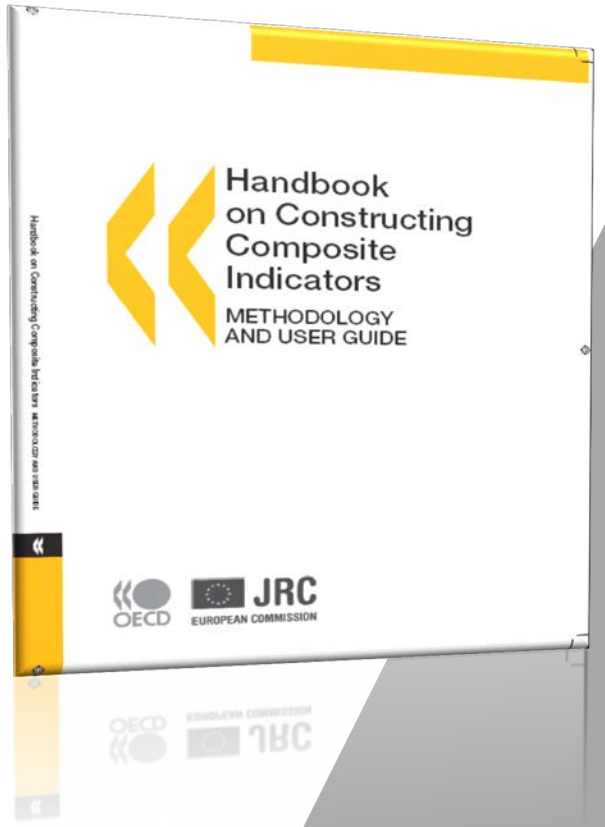
European Commission

Step 3: The identification and treatment of outliers

Giacomo Damioli

COIN 2017 - 15th JRC Annual Training on Composite Indicators & Scoreboards
06-08/11/2017, Ispra (IT)

Decalogue



Step 10. Presentation & dissemination

Step 9. Association with other variables

Step 8. Back to the indicators

Step 7. Robustness & sensitivity

Step 6. Weighting & aggregation

Step 5. Normalization of data

Step 4. Multivariate analysis

Step 3. Data treatment (missing, outliers)

Step 2. Selection of indicators

Step 1. Developing the framework

European Commission

Outline

Introduction of the topic

- Definition and relevance

Outlier identification

- Graphical/visual inspection
- Statistical rules (-of-thumb)

Outlier treatment

- To treat or not to treat: this is the question …
- Winsorization, Trimming, Box-Cox transformation

Definition(s)

"*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, B. and P. Mee (2002) Statistics for social data analysis]

"*An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism*"

[Hawkins, D. (1980) Identification of Outliers]

"*An outlying observation, or "outlier," is one that appears to deviate markedly from other members of the sample in which it occurs*"

[Grubbs, F. E. (1969) Procedures for detecting outlying observations in samples]

Relevance

Outliers:

- often indicate either measurement error or that the population has a heavy-tailed distribution;
- generally spoil basic descriptive statistics such as the MEAN, the STANDARD DEVIATION and CORRELATION COEFFICIENT, thus causing misinterpretations;
- can be either:
  - ❑ univariate, i.e an observation that consists of an extreme value on one variable, or
  - ❑ multivariate , i.e. a combination of unusual values on at least two variables
- **Focus of the course**: mostly concerned with **univariate outliers** in the composite indicator context.

Outlier identification

Graphical/visual inspection

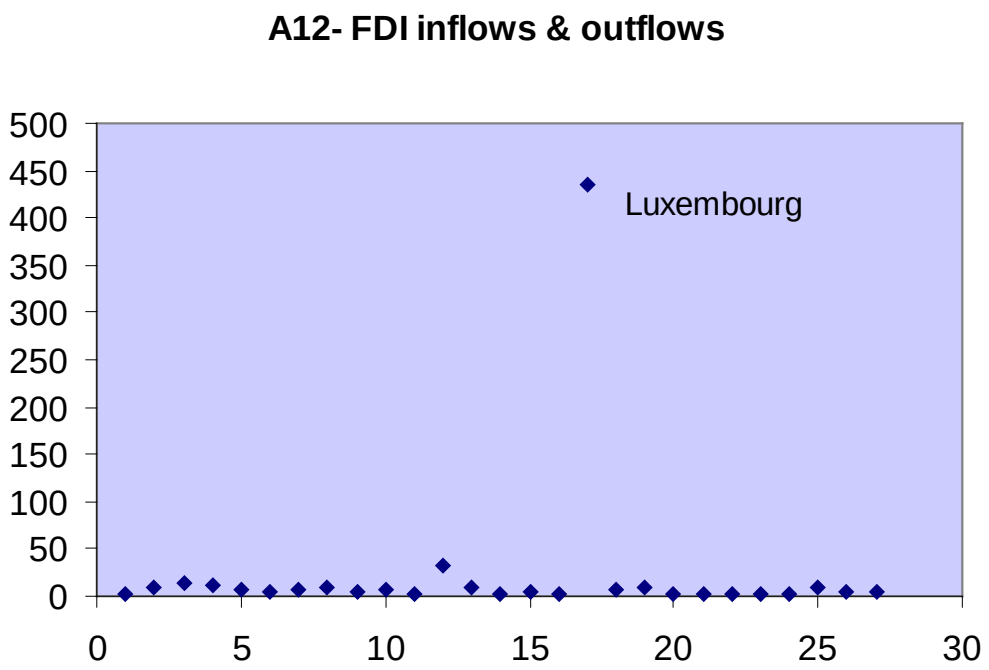
- Simply have a look at the data!

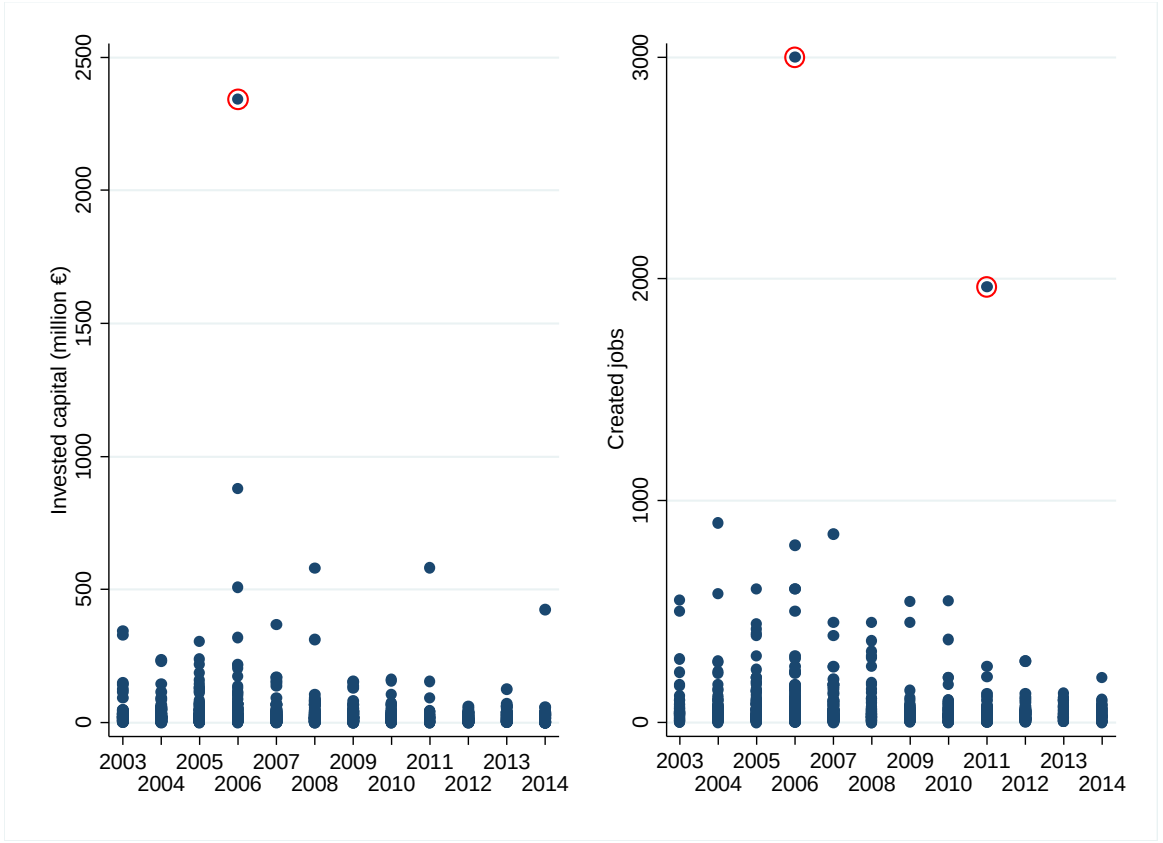
Statistical rules (-of-thumb) 

- z-scores
- ± 1.5 * Interquartile range
- Simultaneous ‘anomalous’ values of Skewness and Kurtosis

Outlier identification

✓ simply have a look at the data!





European Commission

Outlier identification

✓ z-scores

Another way to identify univariate outliers is to convert all values (x_i) of a variable to standard scores (z_i):

$$z_i = \frac{x_i - \mu}{\sigma}$$

Then:

- If the sample size is small (80 or fewer cases), a case is an outlier if

  $|z_i| \geq 2.5$ (or equivalently $|x_i| \geq \mu + 2.5\sigma$)

- If the sample size is larger than 80 cases, a case is an outlier if

  $|z_i| \geq 3$ (or equivalently $|x_i| \geq \mu + 3\sigma$)

} more than 99% coverage of distribution

Outlier identification

✓ z-scores

In practice, this criteria can be applied more or less strictly … for instance the Summary Innovation Index, having the number of cases (i.e. countries) equal to 37, uses a stricter cut-off (i.e. $|z_i| \geq 2$ implying "just" more than 97% coverage of distribution) .

4.2 Methodology for calculating the Summary Innovation Index

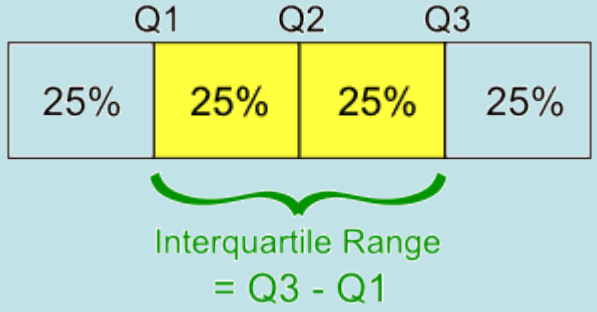
Step 1: Identifying and replacing outliers

Positive outliers are identified as those country scores which are higher than the mean across all countries plus twice the standard deviation. Negative outliers are identified as those country scores which are lower than the mean across all countries minus twice the standard deviation. These outliers are replaced by the respective maximum

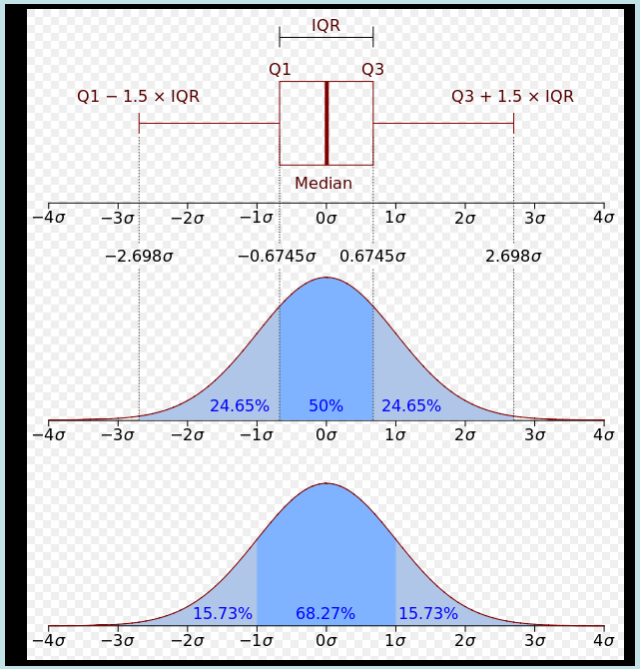
European Innovation Scoreboard 2017 - Methodology report (p. 22)

Outlier identification

✓ ± 1.5 * Interquartile range



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

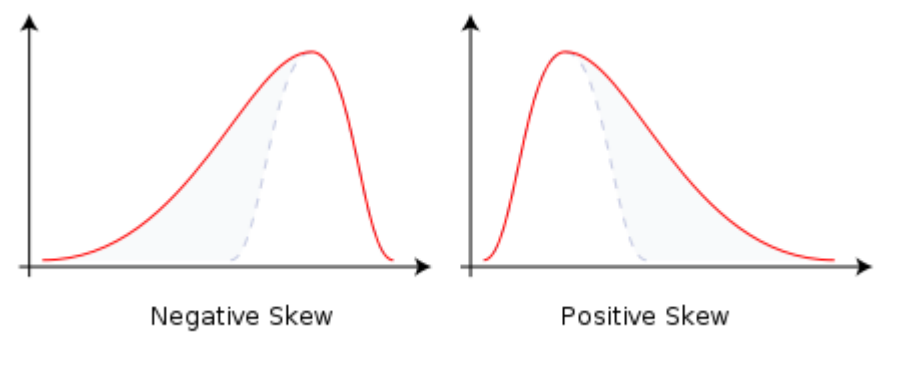


upper boundary $Q_3 + 1.5(Q_3 - Q_1)$

if data are approx. normal, 1.5 corresponds to approx. ± 2.7sd and more than 99% coverage of distribution

Outlier identification

Skewness and Kurtosis

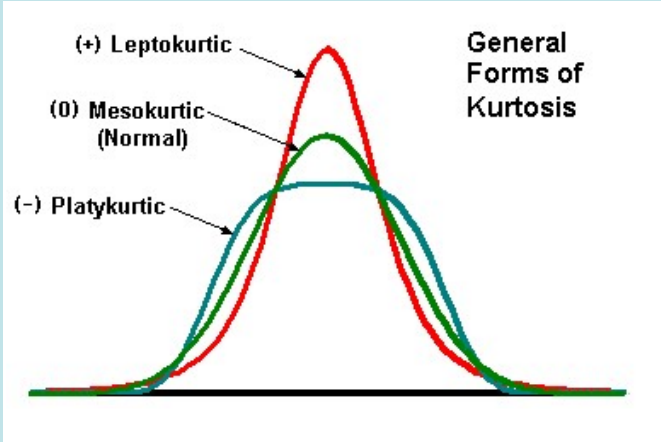


Skewness: measure of the asymmetry of a distribution;

= 0 in the Normal distribution

(+) higher peak around the mean and fatter tails

(-) fatter around the mean and thinner tails



Kurtosis: measure of the thickness of the tails of a distribution;

= 3 in the Normal distribution

Outliers identification

✓ Simultaneous ‘anomalous’ values of Skewness and Kurtosis

- Critical values of skewness and kurtosis (depending on sample size)
- Rule of thumb:  |skewness| > 2 & kurtosis > 3.5

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

Outlier identification

The criterion based on the interquartile range identifies more cases as outliers (is more "invasive") than z-scores, which in its turn identifies more cases as outliers than the criterion based on skewness and kurtosis (is less "invasive")

Global Innovation Index 2017 - A sub-sample (indicators within components 2.1 and 2.2)

	2.1.1	2.1.2	2.1.3	2.1.4	2.1.5	2.2.1	2.2.2	2.2.3
	Expenditure on education	Government expenditure on education per pupil, secondary	School life expectancy	Assessment in reading, mathematics, and science	Pupil-teacher ratio, secondary	Tertiary enrolment	Graduates in science and engineering	Tertiary inbound mobility
Methods for outlier identification	**Number of outliers**							
± 1.5*(Q3-Q1)	4	3	1	0	4	0	3	9
z-scores	0	2	0	0	0	0	2	3
'anomalous' Skewness & Kurtosis	0	0	0	0	0	0	0	1+

Outlier treatment

To treat or not to treat ….

- Reasons to treat outliers
- Cautions

Methods for the treatment of outliers

- Winsorization
- Trimming
- Box-Cox transformation

Outlier treatment

Outlier treatment may be recommended if:

- You are using a model assuming normality (e.g. standard linear regression) … often treatment means discarding outliers in such a context … but this is not the main reason to treat them in the case of CIs
- You are interested in descriptive statistics such as the MEAN, the STANDARD DEVIATION and the CORRELATION COEFFICIENT, which are often spoiled by outliers … neglecting outliers may cause misinterpretations of CIs

Outlier treatment

Cautions:

- every transformation alters original data
- carefully ponder the choice of transforming data and do it only if really not avoidable
- avoid as much as possible 'tailor-made' transformations (different for each indicator)

Outlier treatment

Simplest approaches:

✓Winsorization: modify their values so to make them closer to the other sample values

Typical case: values distorting the indicator distribution are assigned the next highest/lowest value, up to the level where skewness or kurtosis enter within the specified ranges.

Winsorization does NOT preserve order relations for the units treated

✓Trimming: the most extreme way to treat an outlier is to trim it out from the sample, i.e. to eliminate it

Outlier treatment

An example of winsorization: the 2017 Summary Innovation Index

4.2 Methodology for calculating the Summary Innovation Index

Step 1: Identifying and replacing outliers

Positive outliers are identified as those country scores which are higher than the mean across all countries plus twice the standard deviation. Negative outliers are identified as those country scores which are lower than the mean across all countries minus twice the standard deviation. These outliers are replaced by the respective maximum and minimum values observed over all the years and all countries. Table 4 summarises the outliers per indicator and year (negative outliers are shown in italics).

Table 4: Overview of positive and negative outliers

	Positive / *Negative* outlier
Human resources	
1.1.1 New doctorate graduates	SI: 2013-2015; CH: 2008-2015
1.1.2 Percentage population aged 25-34 having completed tertiary education	CY: 2016 *TR: 2009, 2010*
1.1.3 Population aged 25-64 participating in lifelong learning	SE: 2013-2016; CH: 2010-2016
Attractive research systems	
1.2.1 International scientific co-publications per million population	DK: 2016, 2016; IS: 2010, 2012-2016; CH: 2011-2016
1.2.2 Top 10% most cited publications	CH: 2011, 2013
1.2.3 Foreign doctorate students	LU: 2011-2015
Innovation-friendly environment	
1.3.1 Broadband penetration	DK: 2016-2016; FI: 2016;

European Innovation Scoreboard 2017 - Methodology report (p. 22)

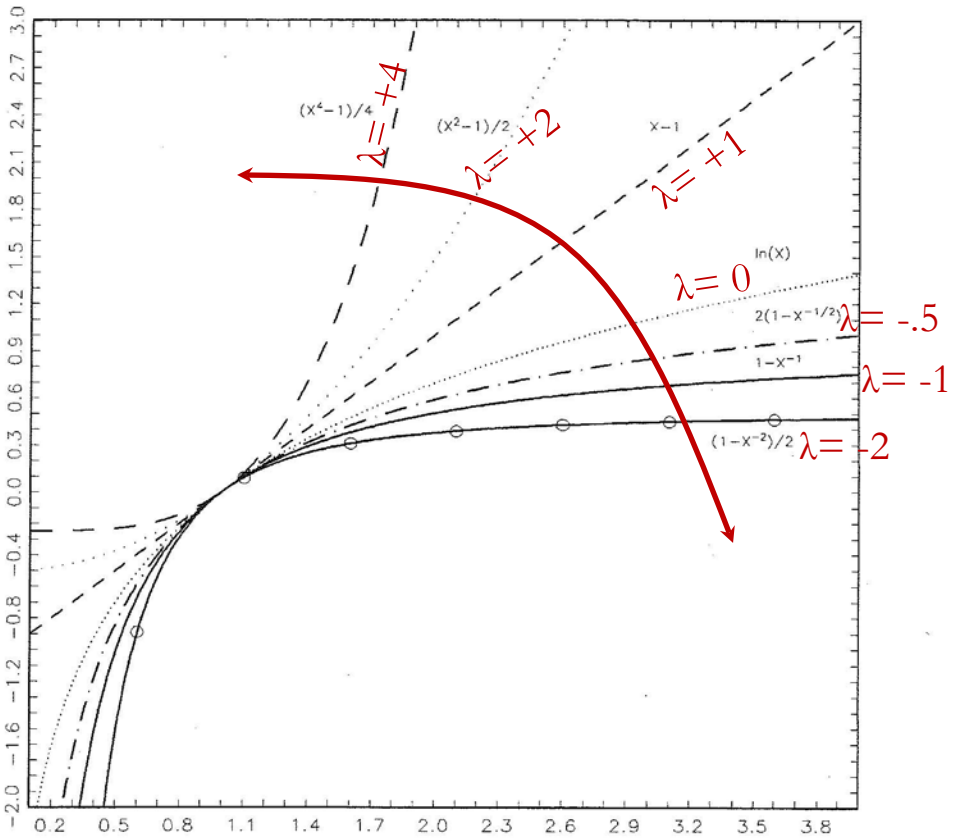
Outlier treatment

✓ Box-Cox family of transformations

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

$$x > 0$$

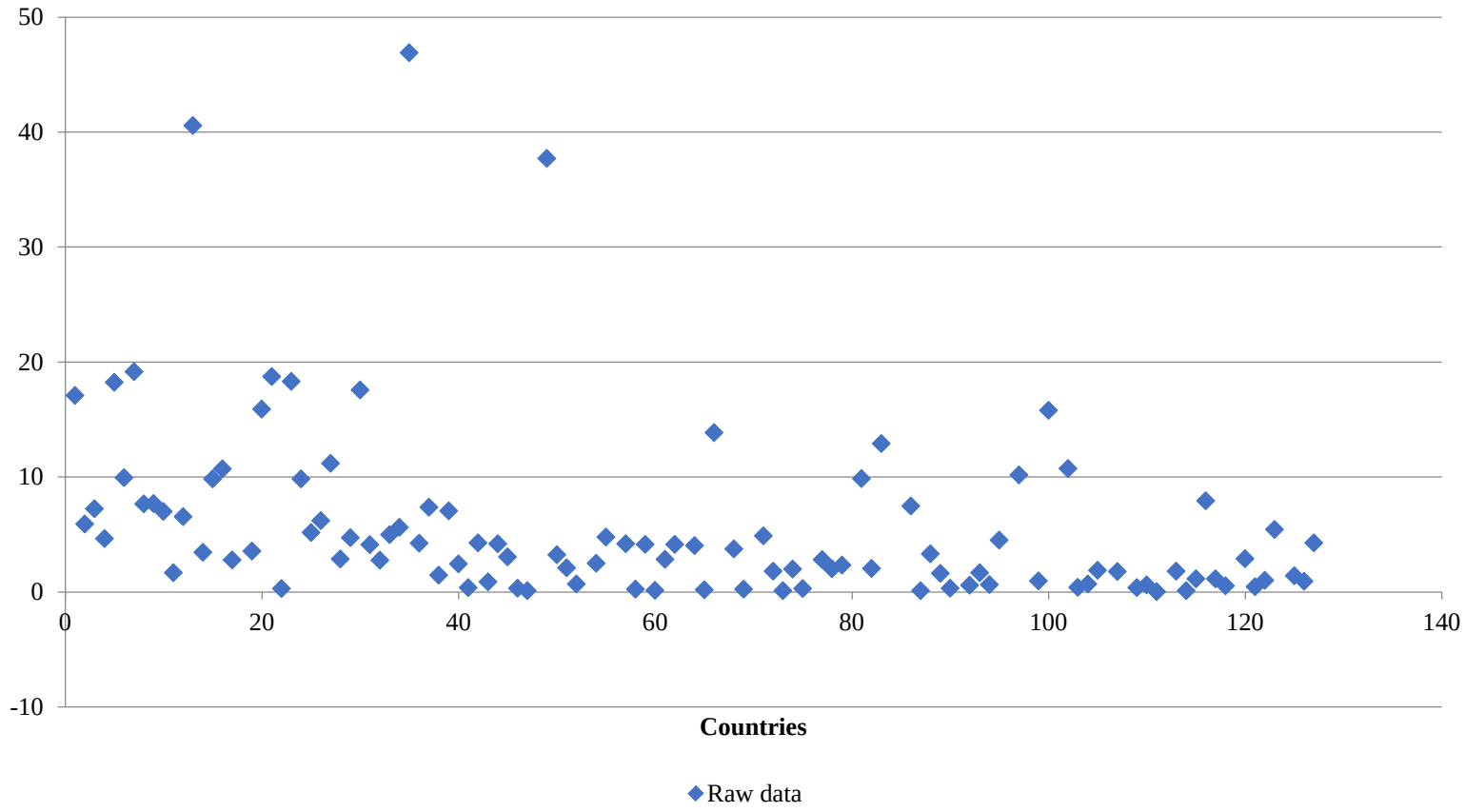
- can ‘compact’ high values if λ<1 (can ‘stretch’ them if λ>1)
- choice of λ should be based on a symmetry measure of the transformed indicator
- often different optimal λ for different indicators
- log transformation case most widely used



European Commission

Outlier treatment

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

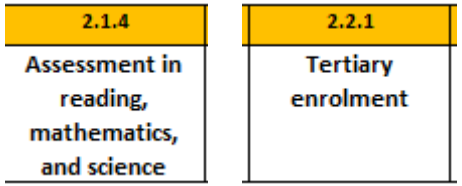


Outlier treatment

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

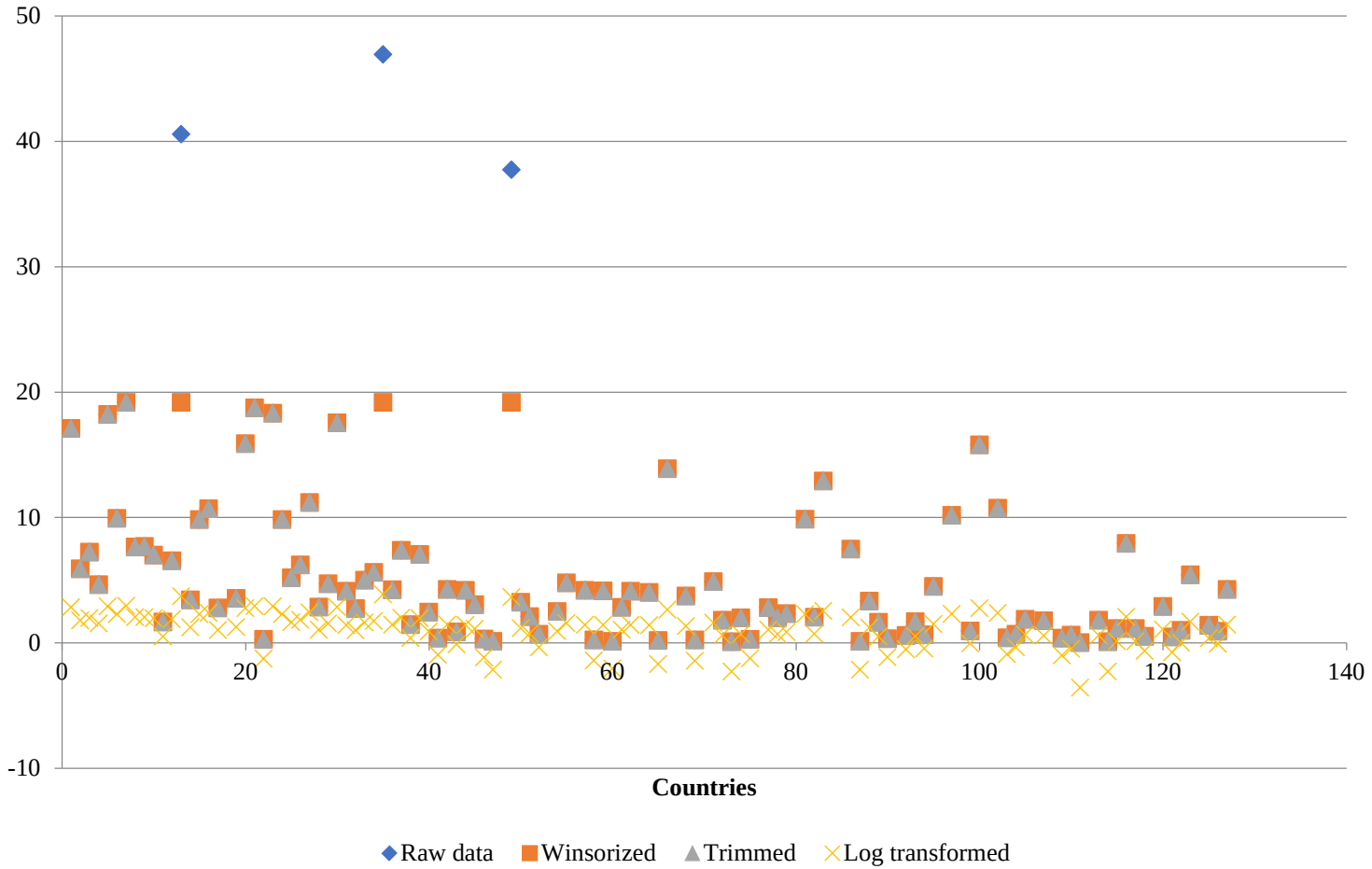
Country	Raw data	Winsorized	Trimmed	Log transformed
CHE	17	17	17	2.8
SWE	5.9	5.9	5.9	1.8
NLD	7.2	7.2	7.2	2.0
USA	4.6	4.6	4.6	1.5
GBR	18.2	18.2	18.2	2.9
DNK	9.9	9.9	9.9	2.3
SGP	19.2	19.2	19.2	3.0
FIN	7.7	7.7	7.7	2.0
DEU	7.7	7.7	7.7	2.0
IRL	7.0	7.0	7.0	1.9
KOR	1.7	1.7	1.7	0.5
ISL	6.5	6.5	6.5	1.9
LUX	40.6	19.2		3.7
JPN	3.4	3.4	3.4	1.2
FRA	0.8	0.8	0.8	2.2

	Raw data	Winsorized	Trimmed	Log transformed
Mean	5.8	5.1	4.7	0.9
Sigma (sd)	7.8	5.4	4.9	1.5
Q1	1.1	1.1	1.0	0.1
Q3	7.1	7.1	6.7	2.0
Skewness	3.1	1.4	1.5	-0.6
Kurtosis	11.6	1.0	1.5	0.1
Corr(2.2.3, 2.1.4)	0.20	0.33	0.39	0.38
Corr(2.2.3, 2.2.1)	0.09	0.20	0.27	0.28



Outlier treatment

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



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

Final remarks

In this class we have considered each variable (indicator) one at a time. Multivariate, simultaneous detection of outliers may also be of interest:

- Forward Search
- Mahalanobis distance

Suggested reading

- Atkinson, A.C., Riani, M. & A. Ceriolin (2004) "Exploring Multivariate Data with the Forward Search" Springer-Verlag – New York.
- Ghosh, D., & A. Vogt (2012) " Outliers: an evaluation of methodologies" *American Statistical Association*. Section on Survey Research Methods – JSM 2012
- 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"

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