# The European Commission's science and knowledge service

## Joint Research Centre

European Commission

# Step 3: The identification and treatment of outliers

**Giacomo Damioli**

COIN 2018 - 16th JRC Annual Training on Composite Indicators & Scoreboards
05-07/11/2018, Ispra (IT)

European Commission

# 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 (outliers and missing values)**
- Step 2. Selection of indicators
- Step 1. Developing the framework



European Commission

# Outline

Outliers

- Definition and relevance
- Outlier identification
- Outlier treatment techniques

Missing values

- Definition and relevance
- Pre-imputation steps
- Imputation techniques

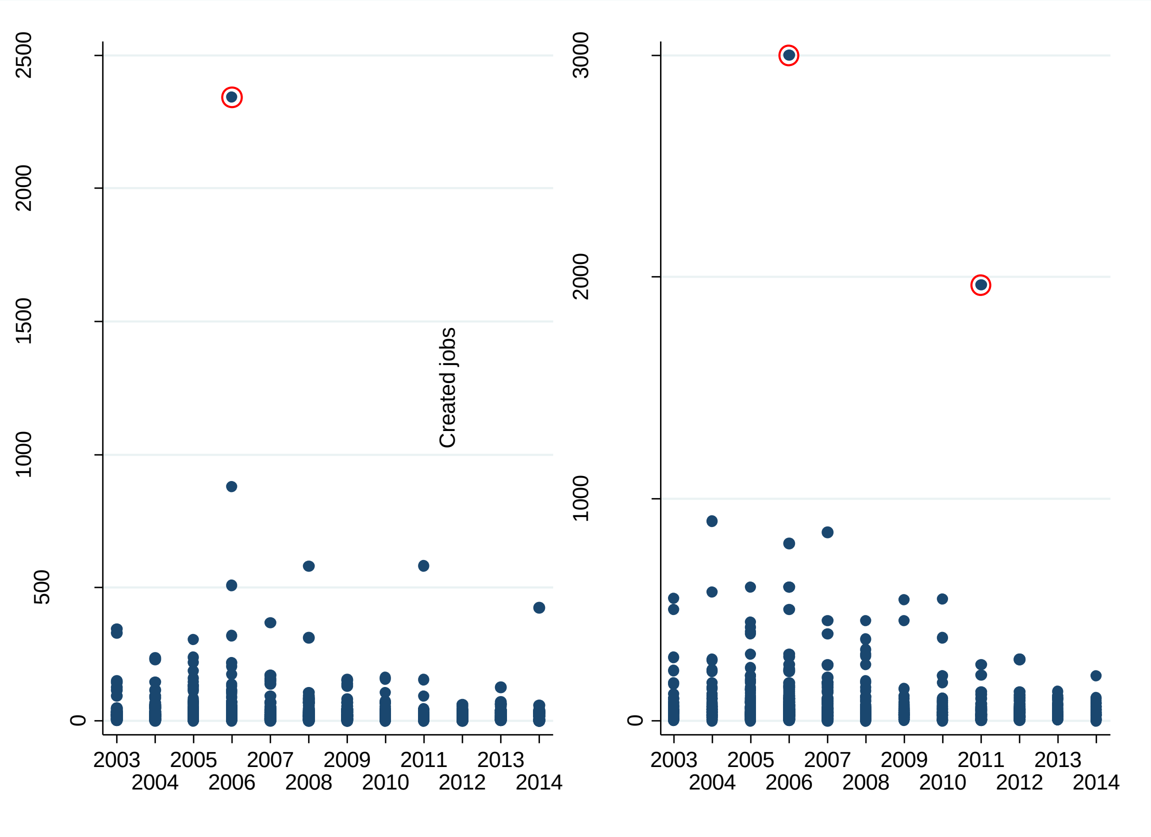
# Outliers – what are they?

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

*"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]






European Commission

# Outliers – why do we care about?

**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.



European Commission

# Outliers – how do we identify them?

Graphical/visual inspection

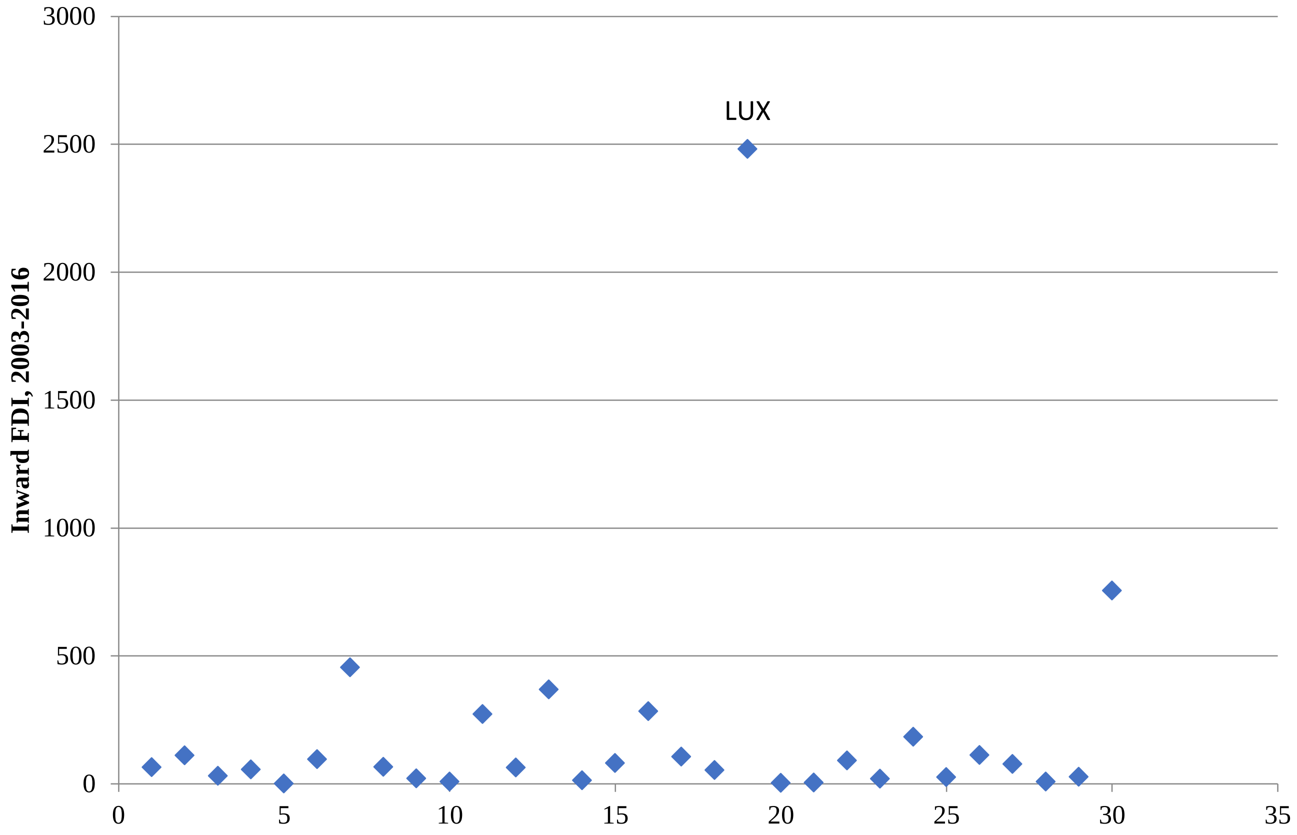
- Simply have a look at the data!

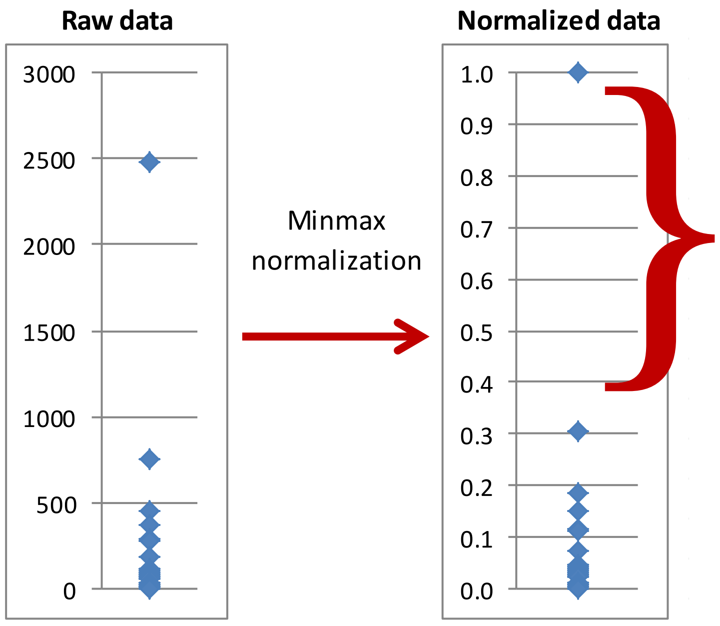
Statistical rules (-of-thumb) 

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

# Outliers – how do we identify them?

✓ simply have a look at the data!




***70% of the scale is "empty"***

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



European Commission

# Outliers – how do we identify them?

✓ 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



European Commission

# Outliers – how do we identify them?

✓ 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

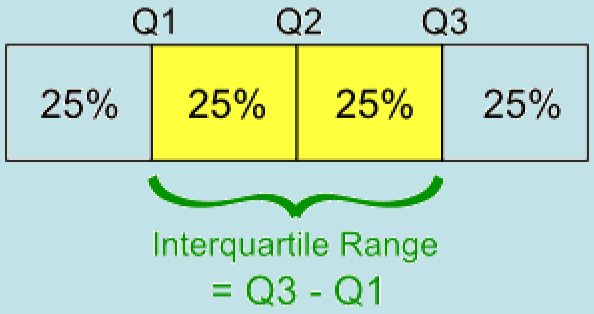
European Innovation Scoreboard 2017 - Methodology report (p. 22)



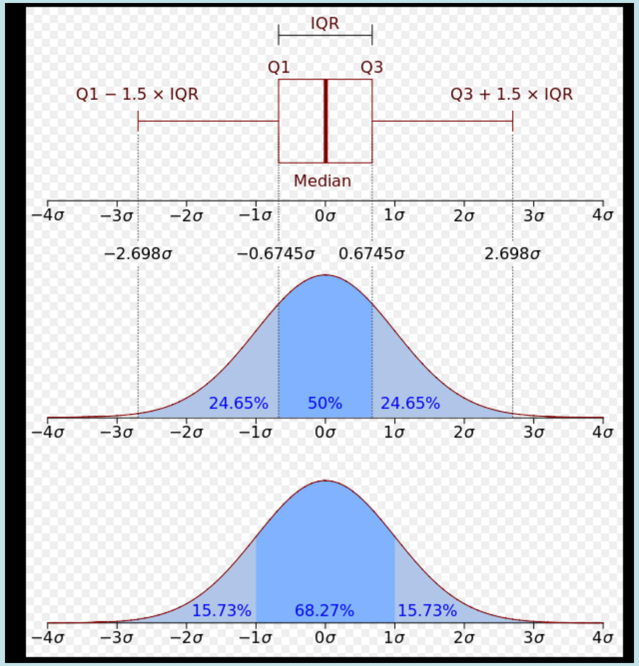
European Commission

# Outliers – how do we identify them?

✓ ± 1.5 * Interquartile range




lower boundary $Q_1 - 1.5(Q_3\text{-}\ Q_1)$




upper boundary $Q_3 + 1.5(Q_3\text{-}\ Q_1)$

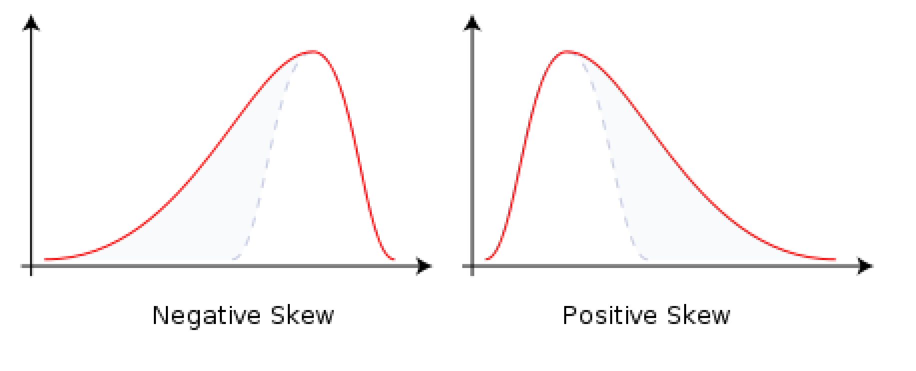
if data are approx. normal, 1.5 corresponds to approx. ± 2.7sd and more than 99% coverage of distribution



European Commission

# Outliers – how do we identify them?

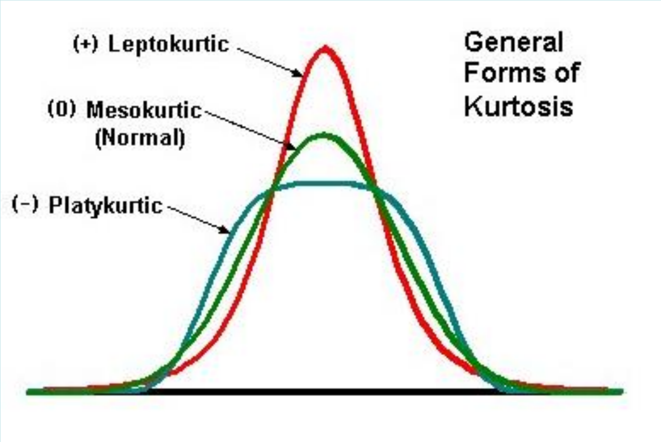
## Skewness and Kurtosis




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 – how do we identify them?

✓ Simultaneous ‘anomalous’ values of Skewness and Kurtosis (JRC preferred option)

- 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 |

# Outliers – how do we identify them?

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 | 3 |

"invasiveness" + -



European Commission

# Outliers – how do we treat them?

To treat or not to treat ….

- Reasons to treat outliers
- Cautions

Methods for the treatment of outliers

- Winsorization
- Trimming
- Box-Cox transformation

# Outliers – should we treat them?

Outlier treatment <u>may</u> 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 … not treating outliers may cause misinterpretations of CIs



European Commission

# Outliers – should we treat them?

## Cautions:

- every transformation alters original data

- carefully ponder the choice of transforming data and do it only if really not avoidable

SPECIAL CASE: normalization based on rankings ➔ no need to treat outliers (outliers are an issue when the distance, not their ordering, is used in CI development)

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

# Outliers – how do we treat them?

## Simplest approaches:

✓Winsorization (JRC preferred treatment in the case of low number of outliers – less than 5): 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 desired ranges (i.e. |skewness| < 2 *or* kurtosis < 3.5).

Winsorization does NOT preserve order relations for the units treated

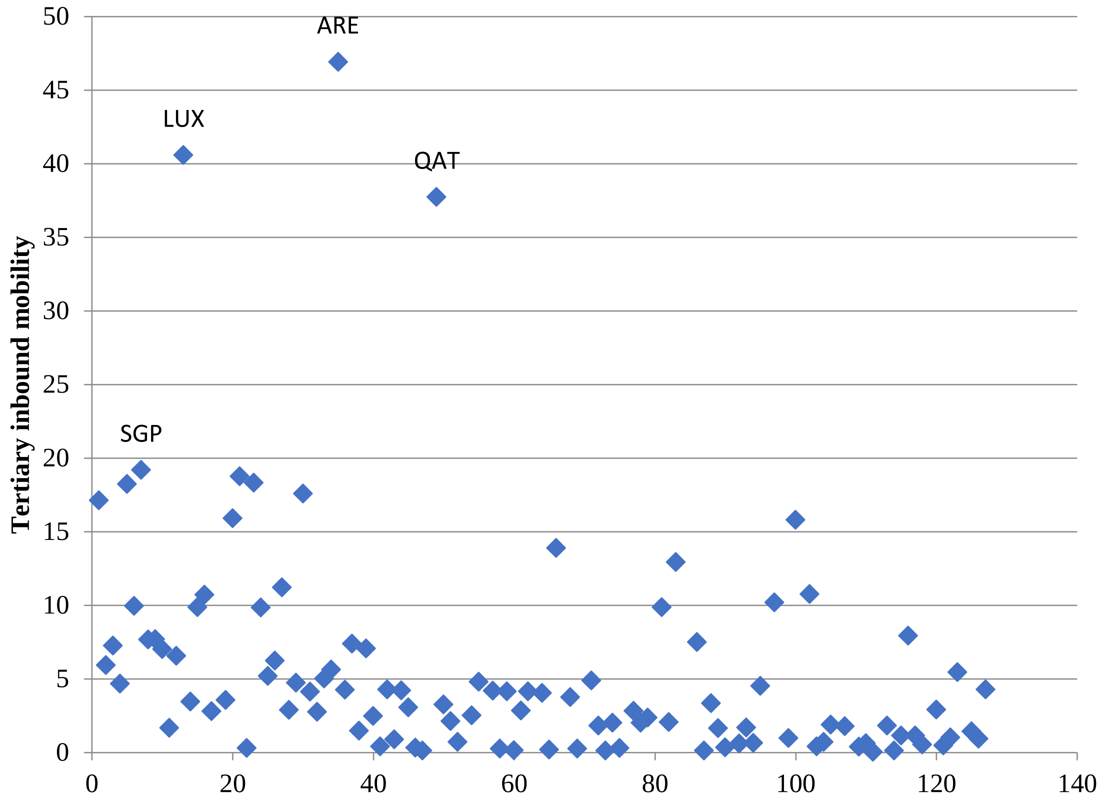
✓Trimming: the most extreme way to treat an outlier is to trim it out from the sample, i.e. to eliminate it



European Commission

# Outliers – how do we treat them?

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

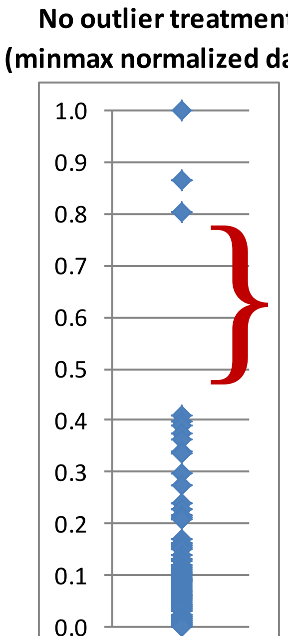





European Commission

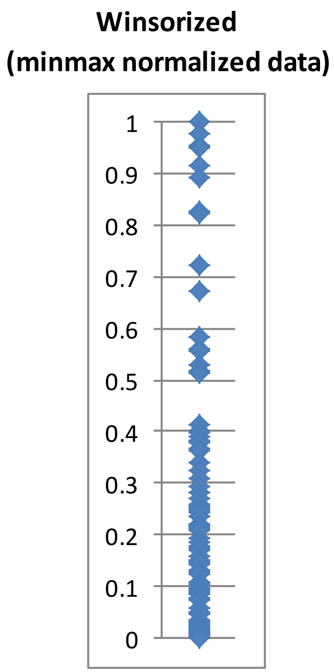
# Outliers – how do we treat them?

## An example - Winsorization




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

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




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

| | Raw data | Winsorized |
|---|---|---|
| Skewness | 3.1 | 1.4 |
| Kurtosis | 11.6 | 1.0 |
| Corr(2.2.3, 2.2.1) | 0.09 | 0.20 |

| 2.2.1 |
|---|
| Tertiary enrolment |



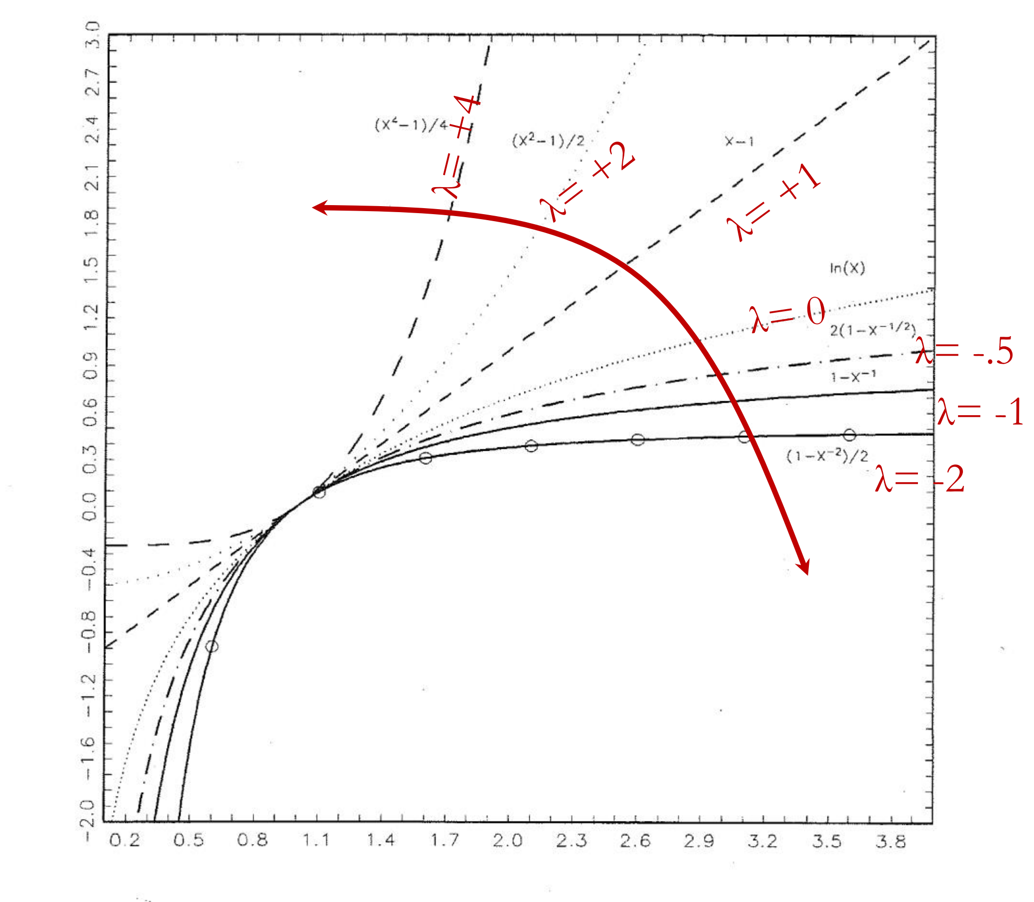
European Commission

# Outliers – how do we treat them?

✓ Box-Cox family of transformations

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

$$x > 0$$



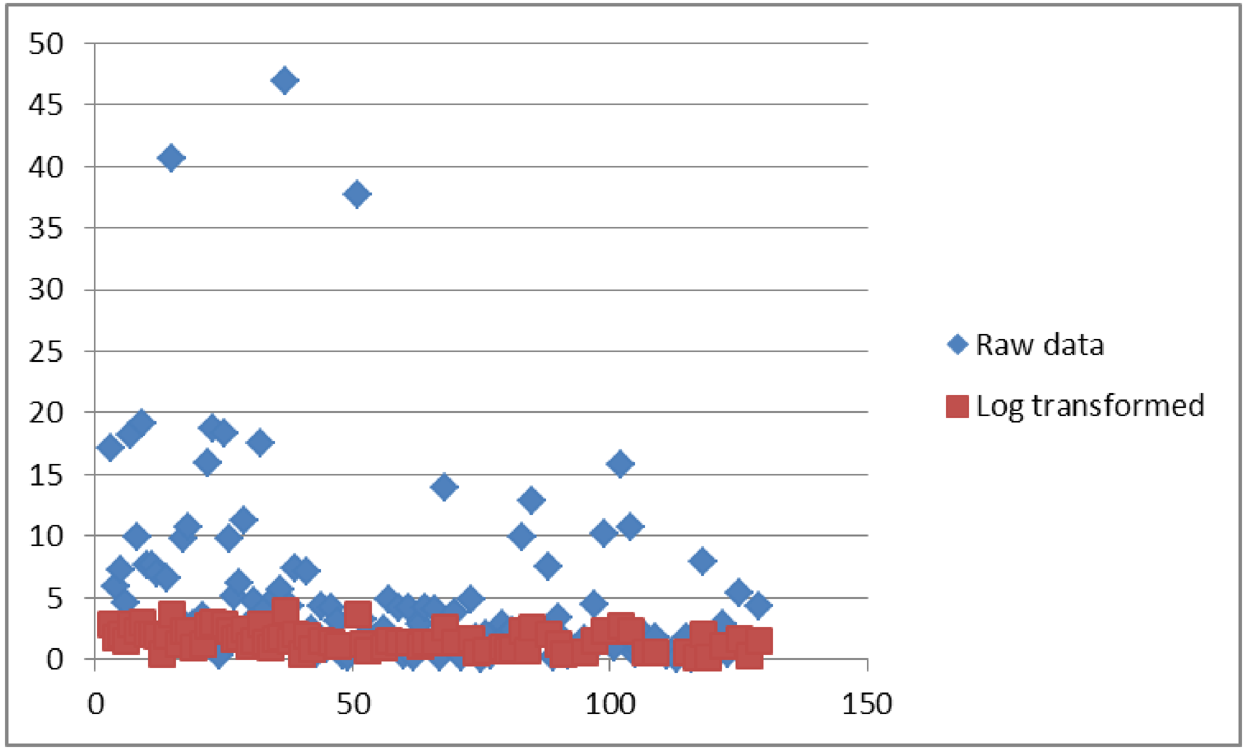

- can ‘compact’ high values if λ<1 (can ‘stretch’ them if λ>1)
- choice of λ should be based on a symmetry measure of the transformed indicator
- often different optimal λ for different indicators
- log transformation (λ=0):
  - case most widely used
  - **JRC preferred method in the case of high number (e.g. 5 or more) of outliers**



European Commission

# Outliers – how do we treat them?

## An example – log transformation




Log-transformation changes all data and "compacts" them

| | Raw data | Winsorized | Log transformed |
|---|---|---|---|
| Skewness | 3.1 | 1.4 | -0.6 |
| Kurtosis | 11.6 | 1.0 | 0.1 |
| Corr(2.2.3, 2.2.1) | 0.09 | 0.20 | 0.28 |

| 2.2.1 |
|---|
| Tertiary enrolment |

# 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 – final remarks and suggested reading

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"




European Commission

# Outline

Outliers

- Definition and relevance
- Outlier identification
- Outlier treatment techniques

Missing values

- Definition and relevance
- Pre-imputation steps
- Imputation techniques

# Missing values – what are they?

"Missing data (or missing values) is defined as the data value that is not stored for a variable in the observation of interest"

[Kang, 2013. The prevention and handling of the missing data]

"Imputation of missing data on a variable is replacing that missing by a value that is drawn from an estimate of the distribution of this variable"

[Dondersa et al., 2006. Review: A gentle introduction to imputation of missing values]

# Missing values – why do we care about?

Ignoring **missing data** may:

- reduce the representativeness of the sample
- reduce the statistical power of the data
- generate biased estimates

These issues may prevent sound CI development

# Missing values – pre-imputation steps

## Before moving to imputation, it is recommended to:

1. Identify **reasons and patterns for missing**, and recode correctly when relevant

-> coding issues (there are often special values, such as 99, -1, 0 …, for missing), questions to be skipped in questionnaires according to previous answers, respondent refusal, item and unit nonresponses, (panel) attrition …

-> what items (i.e. indicators) have more missing? what units (i.e. countries)?

2. Assess the distribution of the missing data, and identify the **type** of ‘missingness’

-> missing completely at random (MCAR), missing at random (MAR), not missing at random (NMAR)

# Missing values – pre-imputation steps

## Are missing values too many?

Rules of thumb: 

- At the indicator level: at least 65% of countries should have valid data
- At the country level: at least 65% of indicators should have valid data

Thresholds have to be considered thoroughly, also reflecting indicator importance and the conceptual framework ... Good correlation between indicators supports more missing values …



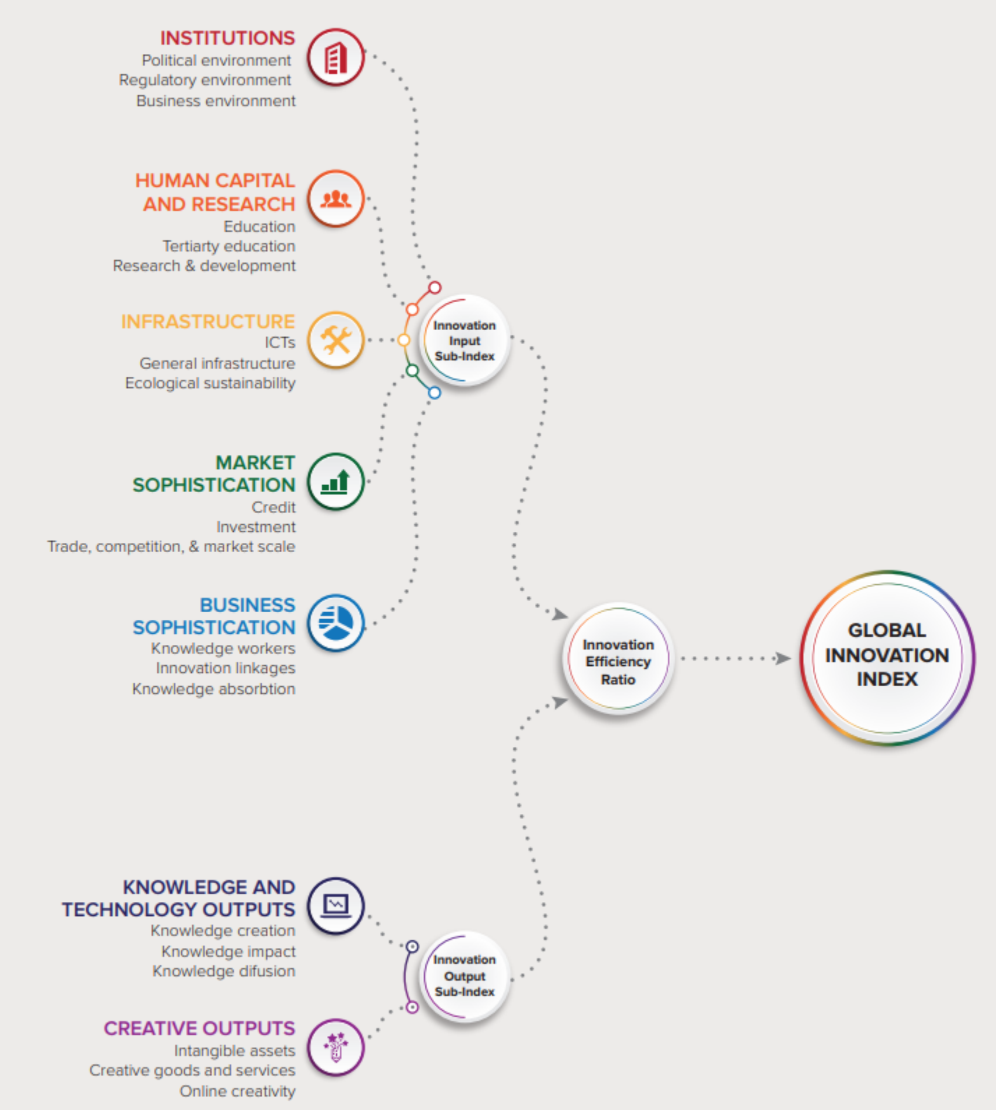
European Commission

# Missing values – pre-imputation steps

## Example. GII 2018

- Indicator-level: at least about 75 countries (out of total 126) with valid cases … exact threshold depends on indicator importance

- Country-level:
  - at least 66% of indicators available in each of the two (Innovation Input and Innovation Output) sub-indexes;
  - scores available for at least two sub-pillars per pillar.



# Missing values – pre-imputation steps


## Types of 'missingness':

1. Missing completely at random (MCAR): 'missingness' not related to any variable


2. Missing at random (MAR): 'missingness' is related to variables having complete information

Example: countries with a democracy more likely to report economic data (GDP, FDI …) than authoritarian countries


3. Not missing at random (NMAR): 'missingness' is related to the variable with missing values

Example: countries with a democracy more likely to report political data (voter turnout, rule of law …)



European Commission

# Missing values – pre-imputation steps

## A toy example of 'missingness' types

| Country | Per-capita GDP | Voter turnout: True unobserved values | MCAR | MAR | NMAR |
|---|---|---|---|---|---|
| 1 | € 60,000 | 70% | N/A | 70% | 70% |
| 2 | € 60,000 | 60% | 60% | 60% | 60% |
| 3 | € 60,000 | 50% | 50% | 50% | N/A |
| 4 | € 40,000 | 70% | 70% | 70% | 70% |
| 5 | € 40,000 | 60% | N/A | 60% | 60% |
| 6 | € 40,000 | 50% | 50% | 50% | N/A |
| 7 | € 20,000 | 70% | 70% | N/A | 70% |
| 8 | € 20,000 | 60% | 60% | N/A | 60% |
| 9 | € 20,000 | 50% | N/A | N/A | N/A |

# Missing values – pre-imputation steps

In the CI development context, ‘missingness’ is typically assumed to be MACR or MAR

- could test for MCAR (t-tests) but not totally accurate

Yet, some good news!!

- Some MAR analysis methods using MNAR data are still pretty good
- Maximum likelihood (ML) and Multiple Imputation (MI) methods are often unbiased with NMAR data even though assume data is MAR

[Schafer and Graham (2002). Missing Data: Our View of the State of the Art]

# Missing values – how should we impute them?

## Imputation methods

- **Deletion Methods** (listwise deletion, pairwise deletion)
- **Single Imputation Methods** (mean/median/mode substitution, hotdeck method, single regression)
- **Model-based Methods** (Maximum Likelihood, Multiple imputation)

# Missing values – how should we impute them?

## Deletion methods

- **Listwise or case deletion**: if a country has a missing value in one or more indicators, then the country is discarded

- **Pairwise deletion**: ignore missing data (no action)

Reasonable only if missing data are very rare and sparse

# Missing values – how should we impute them?

## Listwise deletion

| Country | Pupil-teacher ratio, secondary | Tertiary enrolment | Graduates in science and engineering | Tertiary inbound mobility |
|---|---|---|---|---|
| DNK | 11.3 | 81.5 | 20.4 | 9.9 |
| ~~SGP~~ | ~~14.9~~ | ~~69.8~~ | ~~N/A~~ | ~~19.2~~ |
| FIN | 12.8 | 87.3 | 27.9 | 7.7 |
| ~~DEU~~ | ~~12.1~~ | ~~68.3~~ | ~~N/A~~ | ~~7.7~~ |
| ~~IRL~~ | ~~N/A~~ | ~~77.6~~ | ~~23.8~~ | ~~7.0~~ |
| KOR | 15.6 | 95.3 | 31.9 | 1.7 |
| ~~ISL~~ | ~~N/A~~ | ~~81.3~~ | ~~15.6~~ | ~~6.5~~ |

Pros: the same number countries for every indicator
Cons: reduced sample size and statistical power

## Pairwise deletion

| Country | Pupil-teacher ratio, secondary | Tertiary enrolment | Graduates in science and engineering | Tertiary inbound mobility |
|---|---|---|---|---|
| DNK | 11.3 | 81.5 | 20.4 | 9.9 |
| SGP | 14.9 | 69.8 | ~~N/A~~ | 19.2 |
| FIN | 12.8 | 87.3 | 27.9 | 7.7 |
| DEU | 12.1 | 68.3 | ~~N/A~~ | 7.7 |
| IRL | ~~N/A~~ | 77.6 | 23.8 | 7.0 |
| KOR | 15.6 | 95.3 | 31.9 | 1.7 |
| ISL | ~~N/A~~ | 81.3 | 15.6 | 6.5 |

Pros: simple; retain more data compared to listwise
Cons: it is IMPLICIT imputation; might encourage countries not to report bad performances

# Missing values – how should we impute them?

**Example. Ignoring missing values is *implicit* imputation!!!**

| 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 |

# Missing values – how should we impute them?

## Single Imputation methods

- **Mean** (or median or mode) **substitution**: substitute missing values with the variable mean across countries with valid cases ( or a subgroup of them)
- **Hotdeck method**: substitute missing values with the value(s) of similar countries
- **Single regression**: substitute missing values with regression predicted values



European Commission

# Missing values – how should we impute them?

**Mean substitution**

| 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 |

| 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: simplicity

Cons: distorts distribution, reduces variances ➔ modifies correlations

# Missing values – how should we impute them?

## Hotdeck method

“Missing values of cases with missing data (recipients) are replaced by values extracted from cases (donors) that are similar to the recipient with respect to observed characteristics”

[Beretta and Santaniello, 2016. Nearest neighbor imputation algorithms: a critical evaluation]

**Basic property**: each missing value is replaced with an observed response from a “similar” unit

Various ways to identify “similarity” between units (countries, cities, …) – for instance (Euclidean, Manhattan, …) distances

Pros: use real 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

# Missing values – how should we impute them?

## An example of hot deck imputation - Nearest Neighbor (kNN)

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

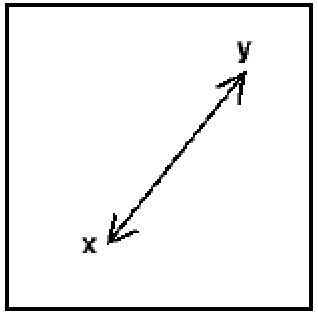
| Country | Expenditure on education | Government expenditure on education per pupil, secondary | School life expectancy |
|---|---|---|---|
| SGP | 2.9 | 16.7 | 12.8 |
| DEU | 4.9 | 23.7 | 17.3 |
| IRL | 5.3 | 26.0 | 19.0 |
| KOR | 4.6 | 23.4 | 16.6 |
| ISL | 7.8 | 18.3 | 19.6 |
| LUX | 4.1 | 19.4 |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

**Step 1. Compute the distance between HKG and other countries**

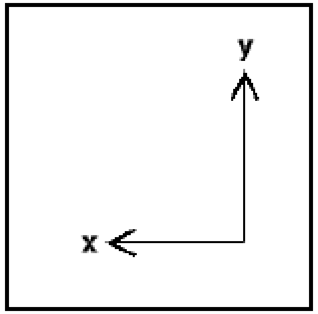
Manhattan distance sometimes preferred over classical Euclidean one if high differences shall not be overweighted
Other distance types do exist (supreme, …)

$$\sqrt{\sum_{i=1}^{N}(x_i - y_i)^2}$$

$$\sum_{i=1}^{N}|x_i - y_i|$$



Euclidean



Manhattan

| 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 |

**Step 2. Impute HKG missing value with the value of the closest country, or the mean value of the k closest countries**

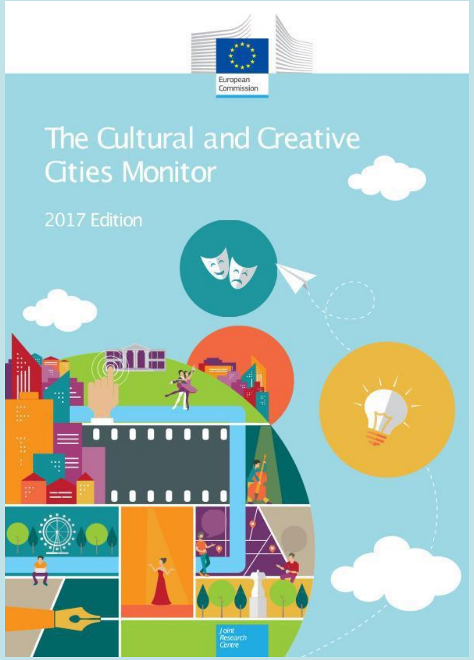
| Number of neighbours | Distance type | Imputed value | |
|---|---|---|---|
| 1NN | Euclidean | 13.9 | |
| 1NN | Manhattan | 13.9 | |
| 2NN | Euclidean | 15.3 | [= (13.9+16.6)/2] |
| 2NN | Manhattan | 13.4 | [= (13.9+12.8)/2] |
| 3NN | Euclidean | 17.4 | [= (13.9+16.6+12.8)/3] |
| 3NN | Manhattan |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

# Missing values – how should we impute them?

## The example of the "Cultural and Creative City Monitor"

- **Mean substitution**: Indicators with missing values imputed with the average of cities having similar population, GDP and employment rates (variables outside the scoreboard, NOT used to capture the culture and creativity of cities)

- **Hot deck**: Remaining missing values imputed with the 3NN method, i.e. using the average of the 3 cities closer (using the Manhattan distance) to the one with the missing value to be imputed in respect to all other variables included in the indicator scoreboard (ie. the 27 variables used to capture the culture and creativity of cities)





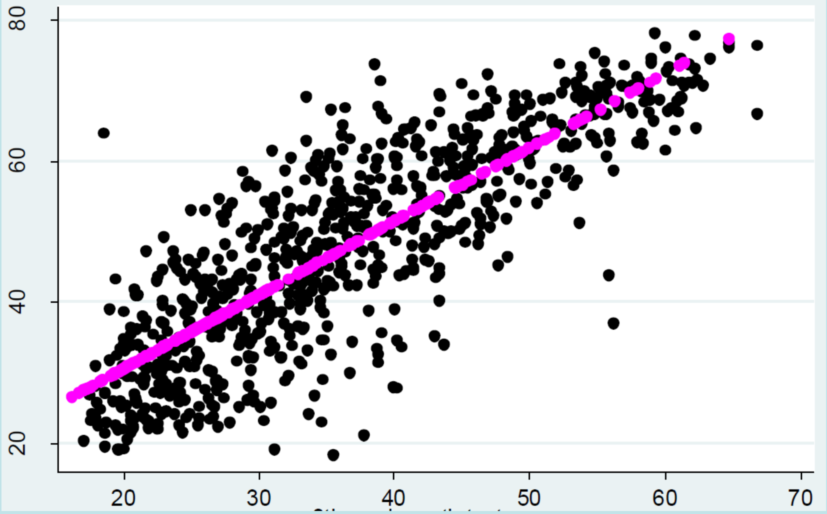

European Commission

# Missing values – how should we impute them?

## Single regression

- use regression model on valid cases of the independent (X) and dependent variables to predict/fill missing values of the dependent variable (Y)

$$Y = X\beta + \varepsilon$$




Pros: simple, uses the most sources of information in comparison to previously discussed methods
Cons: imposes a structure on relationships between variables (e.g linearity …); relies on high correlation between X and Y

# Missing values – how should we impute them?

**Model-based methods: Expectation-Maximization - Maximum Likelihood (EM-ML)**

- The EM algorithm is an iterative procedure to compute Maximum Likelihood estimate in the presence of missing values
- Each iteration of the EM algorithm consists of two processes:
  - ✓ the Expectation-step: the missing data are estimated given the observed data and current estimate of the model parameters
  - ✓ the Maximization-step: the likelihood function is maximized using the estimate of the missing data from the Expectation-step

Pros: works well with good correlation structure; might provide unbiased imputed values also if MNAR
Cons: difficult to communicate, computational-intensive (but increasingly automatised in statistical software)

# Missing values – how should we impute them?

## Model-based methods: Multiple imputation

Multiple imputation methods follow 3 steps:

1) Imputation – Similar to single imputation, missing values are imputed. However, the imputed values are sampled *m times* from their predictive distribution ➔ *m* completed datasets

2) Analysis – Each of the *m* datasets is analyzed ➔ perform CI analysis *m* times

3) Pooling – The *m* results are aggregated into one result by calculating the mean, std. errors and confidence intervals

Pros: might provide unbiased imputed values also if MNAR
Cons: difficult to communicate, computational-intensive (but increasingly automatized in statistical softwares)



European Commission

# Selected Software Packages used in working with missing values

| Software Package | Selected Software Packages used in working with missing values |
|---|---|
| **Freeware** | **link** |
| Amelia | http://gking.harvard.edu/amelia |
| CAT | http://cat.texifter.com/ (for categorical data) |
| EMCOV | https://methodology.psu.edu/publications/books/missing |
| NORM | https://methodology.psu.edu/publications/books/missing |
| MICE | http://www.stefvanbuuren.nl/mi/index.html |
| PAN | http://stat.ethz.ch/~maechler/adv_topics_compstat/MissingData_Imputation.html (Free with R, commercial with S-Plus, for clustered data, including longitudinal data). |
| **Commercial Software** | |
| AMOS | https://www.ibm.com/us-en/marketplace/structural-equation-modeling-sem |
| EQS | http://www.mvsoft.com |
| HLM | http://www.ssicentral.com/hlm/index.html |
| LISREL | http://www.ssicentral.com/index.html |
| Mplus | http://www.statmodel.com |
| SAS | https://www.sas.com/it_it/home.html |
| SOLAS | https://www.statcon.de/shop/en/software/statistics/solas |
| S-Plus | http://www.solutionmetrics.com.au/products/splus/default.html |
| SPSS | http://www-01.ibm.com/software/analytics/spss/products/statistics/modules/ |
| Stata | http://www.stata.com, installing ice or mvis |
| | *Source: Acock, 2005 with author's webpage updates* |

# Missing values – what imputation method?

## Cross-validation

- Many different available methods: which  one to use?

- **Cross-validation**:
  - ✓ impute values of all indicators and countries with valid cases using different methods,
  - ✓ For every method, contrast the imputed/predicted (P) values obtained with the observed (O) values for every valid case (*i*)

    [one way of measuring this difference is the Mean Absolute Percentage Error (MAPE) = $\sum_{i=1}^{N} \frac{|O_i - P_i|}{N}$ ]
  - ✓ choose the method with lowest difference between observed and predicted values

- In JRC experience, cross-validation often indicates to use the EM-ML method

# Missing values – Key lessons

- Pre-imputation steps are important to choose when and how to impute

- How many missing values are there?
  - At the country-level; at the indicator- and pillar-/sub pillar-levels
  - If too many, consider alternative indicators …

- When imputing, avoid as much as possible different imputation methods for different indicators (but be aware that often that's unavoidable!)

- Consider pros and cons of different imputation methods, and assess the sensitivity of rankings to different imputation methods
  - For instance, hotdeck is better when correlation structure of indicators is relatively low, while single regression and multiple-imputation methods rely on strong correlations

# Missing values – suggested reading

## Suggested reading


- Beretta, L. and A. Santaniello (2016). Nearest neighbor imputation algorithms: a critical evaluation. BMC Medical Informatics and Decision Making, 16 (Suppl 3):74.
- Chen, Y. & M.R. Gupta, 2010 EM Demystified: An Expectation-Maximization Tutorial. Department of Electrical Engineering. University of Washington
- Dondersa et al., 2006. Review: A gentle introduction to imputation of missing values. Journal of Clinical Epidemiology. 59: 1087-1091
- Enders, C. K., 2010, Applied Missing Data Analysis. The Guilford Press. Inc: New York, London
- He, Y., 2010, Missing Data Analysis Using Multiple Imputation: Getting to the Heart of the Matter. Circ Cardiovasc Qual Outcomes. 3(1): 98
- Graham, J. W., 2012. Missing data: Analysis and design. New York: Springer.
- Kang, H., 2013, The prevention and handling of the missing data. The Korean Journal of Anesthesiology. 64(5): 402–406.
- Schafer, J. L. & Graham, J.W., 2002, Missing Data: Our View of the State of the Art Psychological Methods. 7(2):147–177

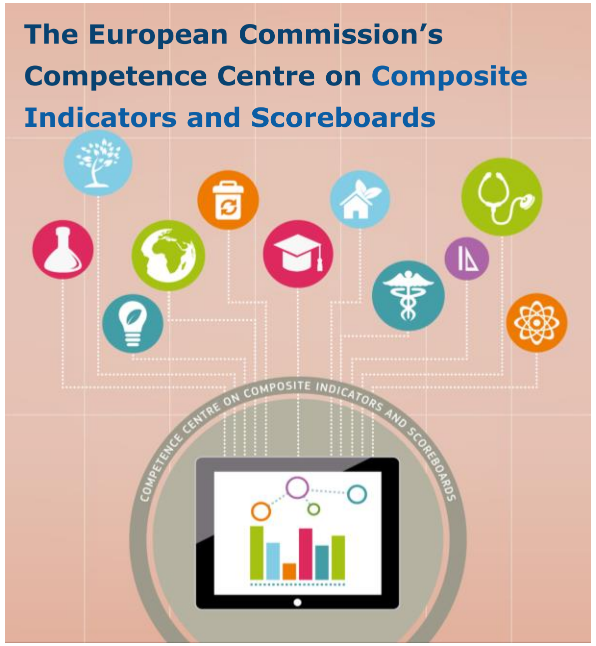
# THANK YOU

## 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