The European Commission's science and knowledge service

Joint Research Centre

Step.4 Normalization

Simone Russo
JRC (Competence Centre on
Composite Indicators and Scoreboards)

Ispra, 6th November 2017

Definition of Normalization

Adjustments of distribution and scale of variables for comparability and robust aggregation.

Standardization

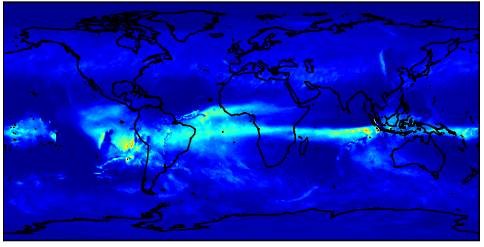
Adjustments of values of difference variables into a common scale (standard-normal $(-\infty, +\infty)$, standard-uniform (0,1))

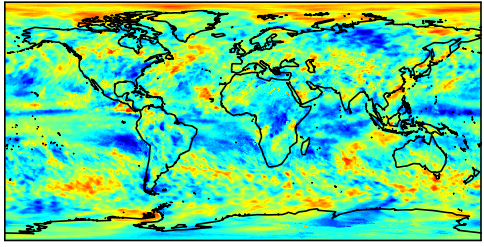
Normalization

Transform variables with different distributions into the same (generally normal) distribution.

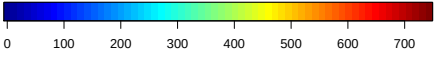
Purpose of normalization

Aug.Sept.Oct. 2016

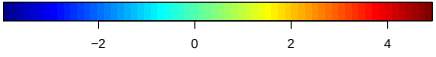




Precipitation (mm)



Norm. Precipitation (std. norm.)



Normalization methods

- Normal Variables: linear transformations (z-score, minmax)
- Non-Normal Variables: empirical or non Linear transformation (quantile normalization, log transformation, box-cox, Johnson method)

Normalization methods

Standardization: adjust values into a common scale

- Z-score support in $[-\infty, +\infty]$
- MinMax support in [0,1]

Advantages: *very simple to calculate*
Disadvantages: *Does not transform distribution of non-Normal Variables (Log-normal, Gumbel, Weibull)*

Normalization methods

Normalization: adjust distributions into a common distribution

- Quantile empirical transformation [0,1], $[-\infty, +\infty]$
- Non Linear Transformation: Box Cox, Johnson distribution family $[-\infty, +\infty]$

Advantages:

- *automatic correction of outliers and adjustment for skewness and kurtosis*
- *transforming variables into standard normal: a condition needed by other statistical methods such as linear regression, PCA, ANOVA.*

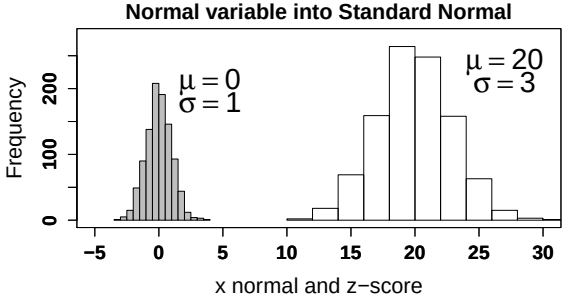
Disadvantages: *Not always simple to be applied*

Normalization Methods

Standardization: z-score

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

The z-score method is used to transform Normal variable with ($\mu \neq 0$ and $\sigma \neq 1$) into a standard normal distribution with $\mu = 0$ and $\sigma = 1$

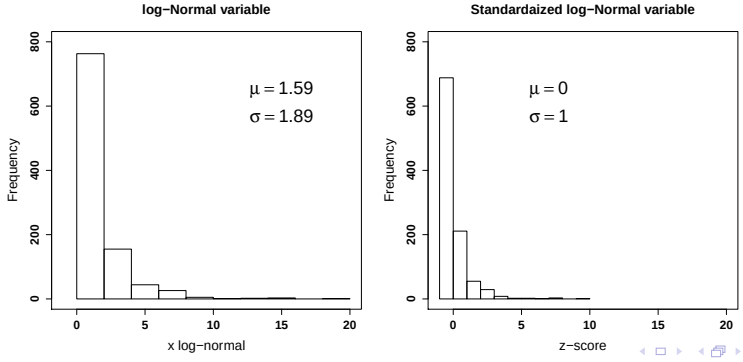


Normalization Methods

Does the z-score method adjust non-normal distribution?

Standardization: z-score

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

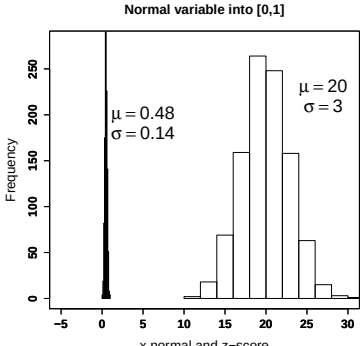


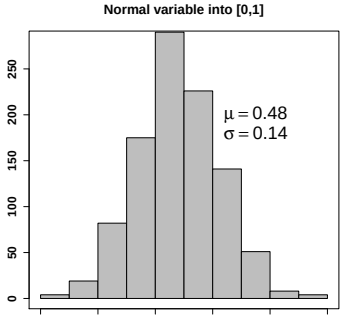
Normalization Methods

Standardization: minmax transformation

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

The minmax method is used to transform variable range into [0,1].





European Commission

Normalization Methods

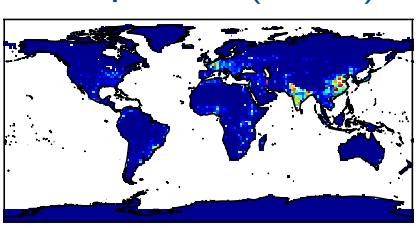
Example: let's consider global population density in million

Standardization: minmax transformation

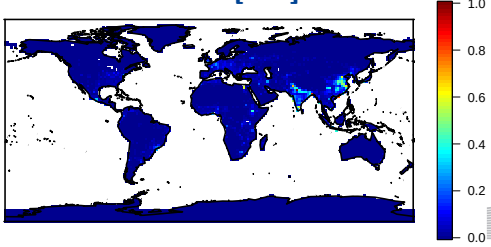
$$z = \frac{x - min(x)}{max(x) - min(x)}$$

The minmax method is used to transform variable range into [0,1].

Pop. dens. (Million)



MinMax [0,1]

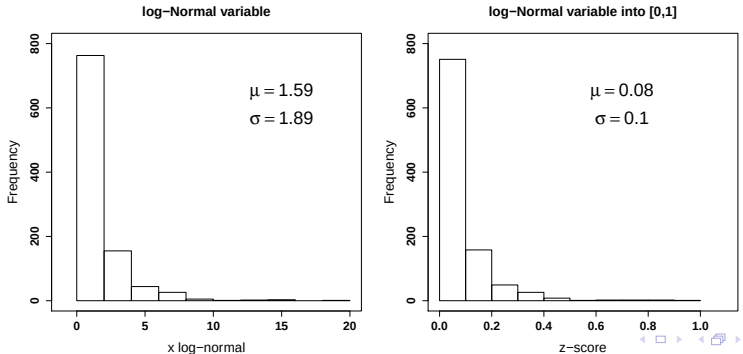


Normalization Methods

Does the minmax method adjust non-normal distribution?

Standardization: minmax transformation

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



Normalization Methods

How can we adjust for distribution?

Quantile Normalization

Definition: *quantile normalization is a technique for making two distributions identical in statistical properties.*

$$u = \frac{rank(x)}{N+1}$$

- where **N** is the sample size;
- rank(x) is the rank associated to each realization.

For example: if $x = (2, -1.5, 3, -6)$ **then**
$rank(x) = (3, 2, 4, 1)$; $u = (0.6, 0.8, 0.4, 0.2)$

Normalization Methods

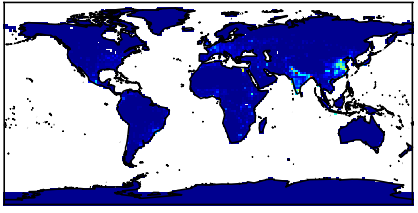
Example: let's consider the global population density in million

Quantile normalization:

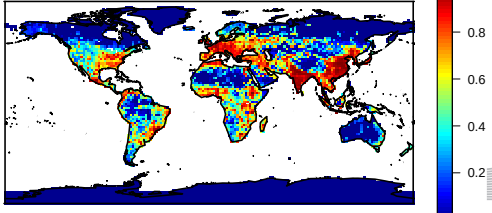
$$u = \frac{rank(x)}{N+1}$$

The Quantile normalization method is used to transform population variable into a standard uniform variable (range into [0,1]).

Pop. dens. (MinMax)



Quantile Norm.

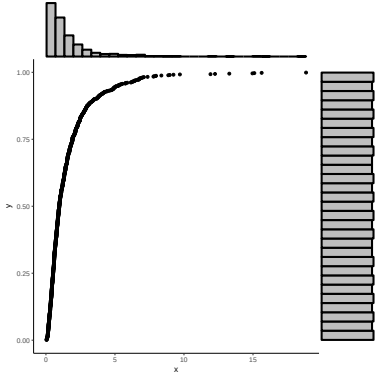


Normalization Methods

How can we adjust for distribution?

Quantile Normalization

From Log-Normal to Uniform: $u = \frac{rank(x)}{N+1}$

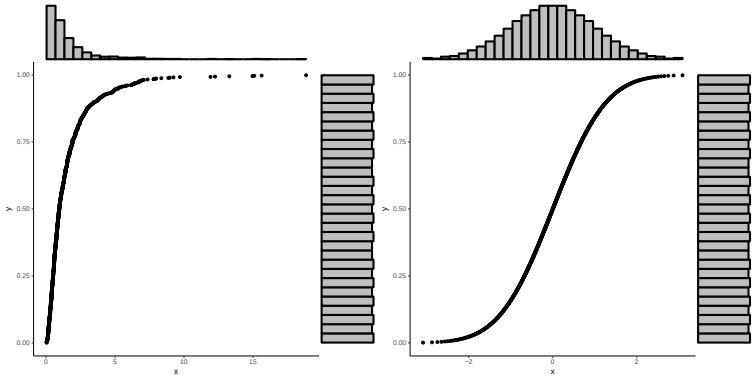


Normalization Methods

How can we adjust for distribution?

Quantile Normalization

From Uniform to Normal: $z = qnorm(u)$



Normalization Methods

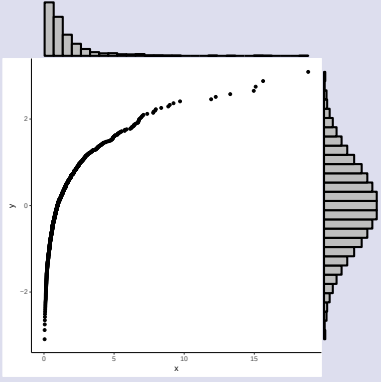
How can we adjust for distribution?

Quantile Normalization

From Log-Normal to Normal: $z = qnorm(\frac{rank(x)}{N+1})$

Advantages:very simple transformation

Disadvantages:Big bias with small sample size ($N < 30$)



Normalization Methods

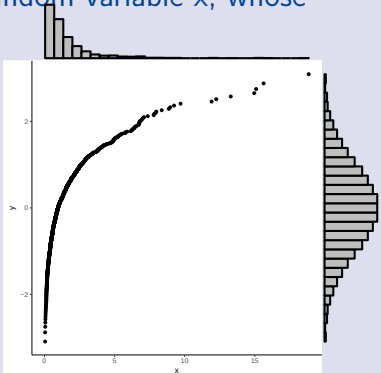
We could re-define Normalization as a mathematical law transforming values of a no-normal distribution into normal.

The Johnson Normalization method (*Johnson 1949*)

Definition: The Johnson normalization consists in finding the best function of transformation to the normal law of a continuous random variable x, whose distribution is not known.

$$Z = \gamma + \delta \times g(\frac{x-\xi}{\lambda})$$

$$Z = 0 + 0.965 \times log(\frac{x-0.0249}{0.944})$$



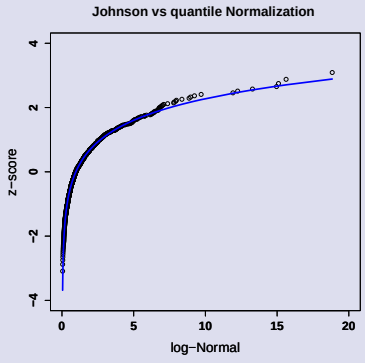
Normalization Methods

Johnson vs Quantile normalization)

The blue curve, in the figure below, is the Johnson function given by the following equation:

$$Z = 0.97 \times log(\frac{x-0.025}{0.94})$$

For more info on the method see:
Johnson 1949; Giovanardi et al 2006

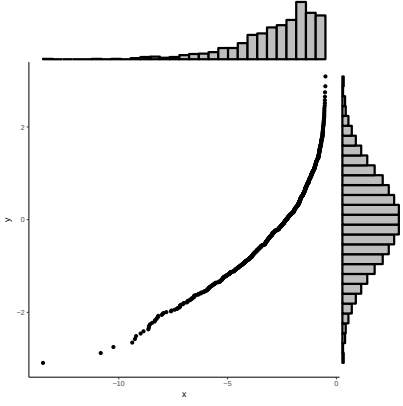


Normalization Methods

The Johnson method works with very complex distribution.

Johnson Normalization with negative skewness

From Weibull to Normal



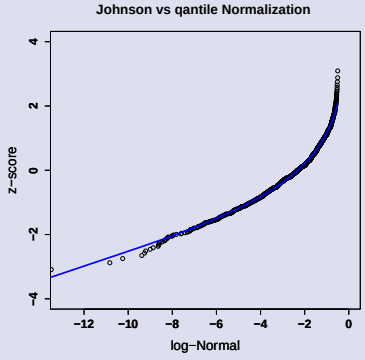
Normalization Methods

Johnson vs Quantile normalization

The blue curve, in the figure below, is the Johnson function given by the following equation:

$$Z = -2.85 + 1.31 \times log\left(\frac{\frac{x+22.85}{22.87}}{1-\frac{x+22.85}{22.87}}\right)$$

For more info on the method see:
Johnson 1949; Giovanardi et al 2006



Normalization vs Standardization

Why do we need to transform variable into distribution with the same properties?

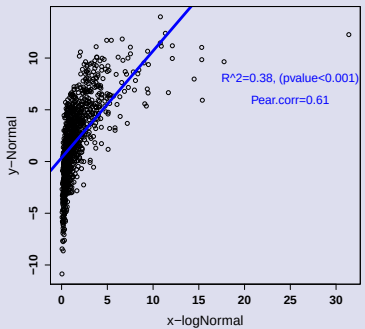
Excercise

Let x be a Log-Normal variable;
Let y be a normal variable dependent on x

In order to study the relationship
between x and y, can we apply a linear regression?

Note that linear regression is
based on Pearson correlation

Statistical methods
based on linear relationships
(e.g. linear regression, ANOVA, PCA)
need normal distributed variables.



Normalization vs Standardization

Why do we need to transform variable into distribution with the same properties?

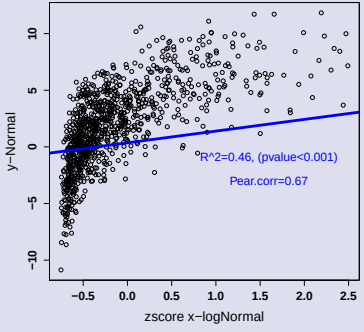
Excercise

Let x be a Log-Normal variable;
Let y be a normal variable dependent on x

In order to study the relationship
between x and y, can we apply a linear regression?

Note that linear regression is
based on Pearson correlation

Statistical methods
based on linear relationships
(e.g. linear regression, ANOVA, PCA)
need normal distributed variables.



Normalization vs Standardization

Why do we need to transform variable into distribution with the same properties?

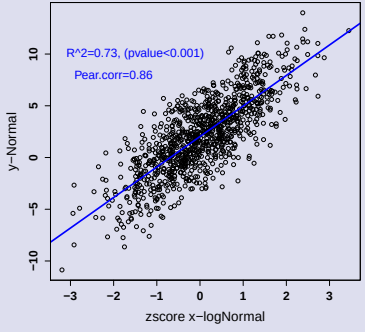
Excercise

Let x be a Log-Normal variable;
Let y be a normal variable dependent on x

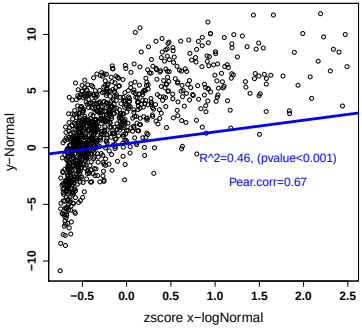
In order to study the relationship
between x and y, can we apply a linear regression?

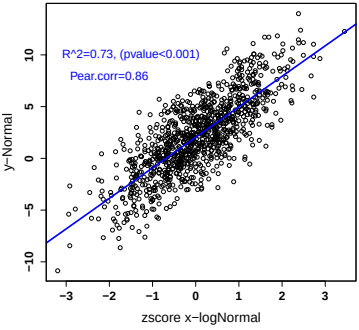
Note that linear regression is
based on Pearson correlation

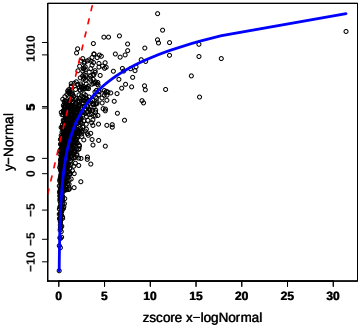
Statistical methods
based on linear relationships
(e.g. linear regression, ANOVA, PCA)
need normal distributed variables.



Normalization vs Standardization







European Commission

Conclusions

- It is important to distinguish between standardization and normalization
- Standardization mainly works with normal variable and adjusts only the scale;
- Normalization is a non-linear process for non-normal variables;
- Normalization adjusts both distribution and scale of a variable;
- When we want to apply multivariate linear methods (linear regression, ANOVA, PCA) to a set of non-normal variables we need to transform both distribution and scale of the variables into standard normal (NORMALIZATION)

THANK YOU

Any questions?

You may contact us at: simone.russo@ec.europa.eu