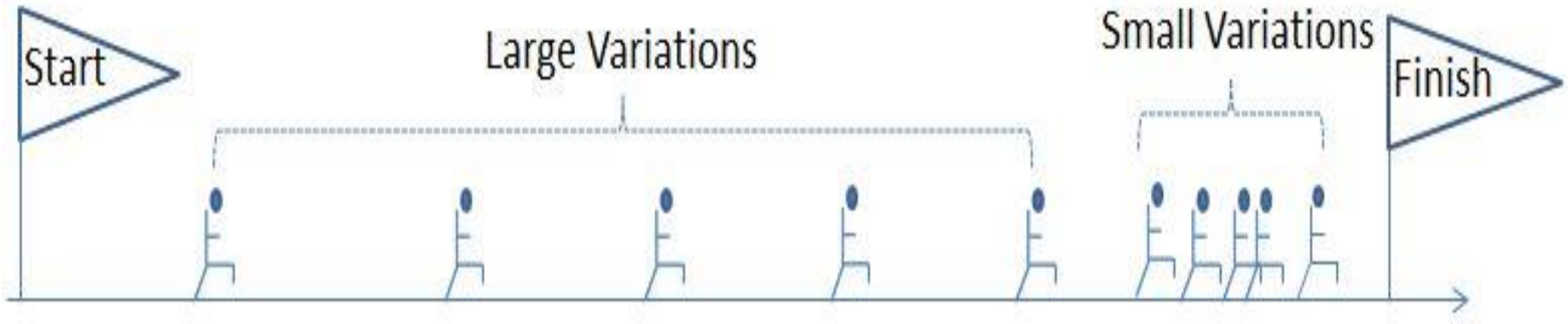
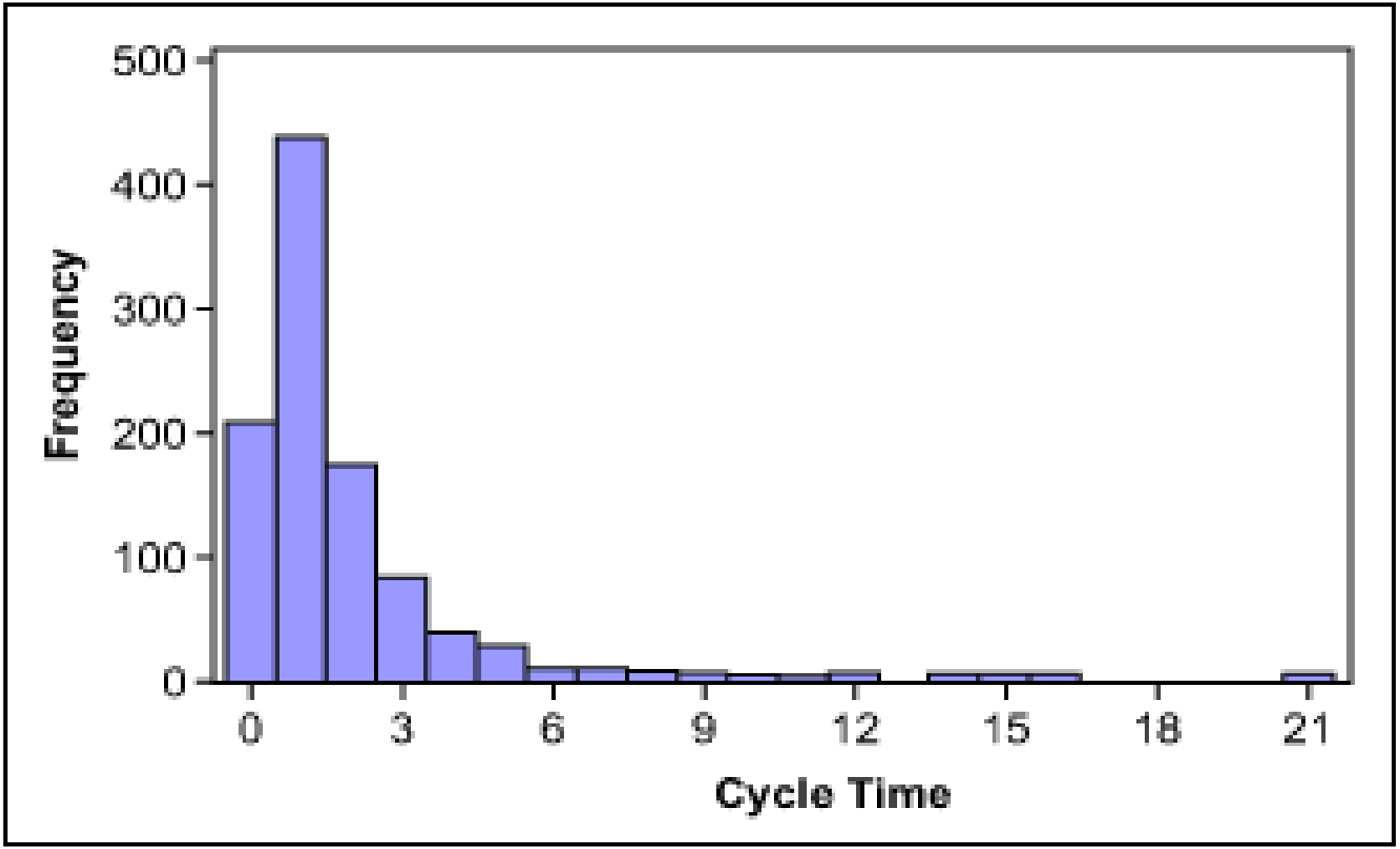
Data Preparation

Handling Skewed Data

How does skewness comes into Picture?



Right Skewed Distribution



Measuring Skewness: Central Moment

The k^{th} central moment (or moment about the mean) of a data is defined as:

$$\mu_k = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^k$$

Skewness

- The skewness is a measure of symmetry.

- The skewness of a data is defined by the following formula, where μ_2 and μ_3 are the second and third central moments:

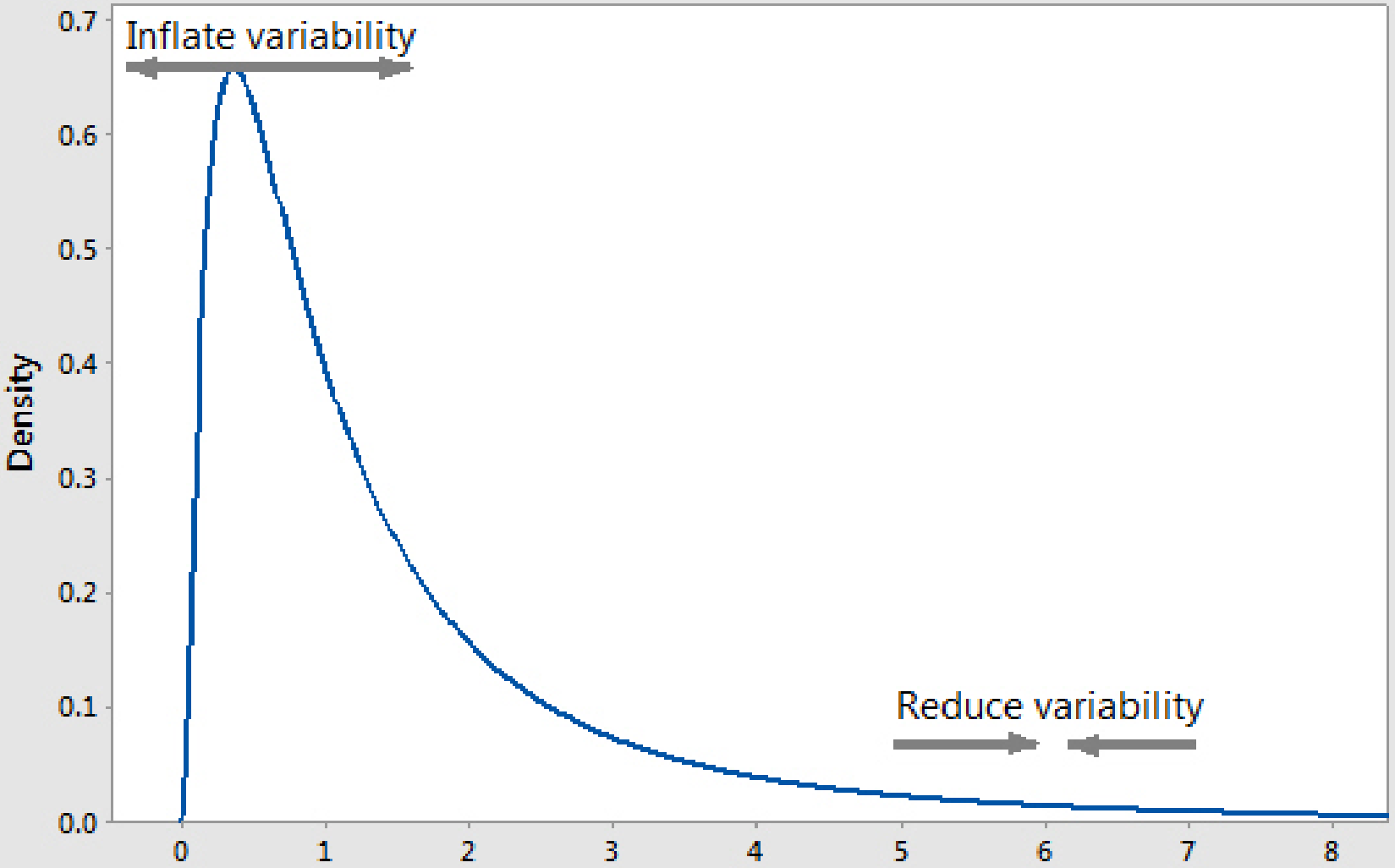
$$\gamma_1 = \mu_3 / \mu_2^{3/2}$$

Interpreting Skewness

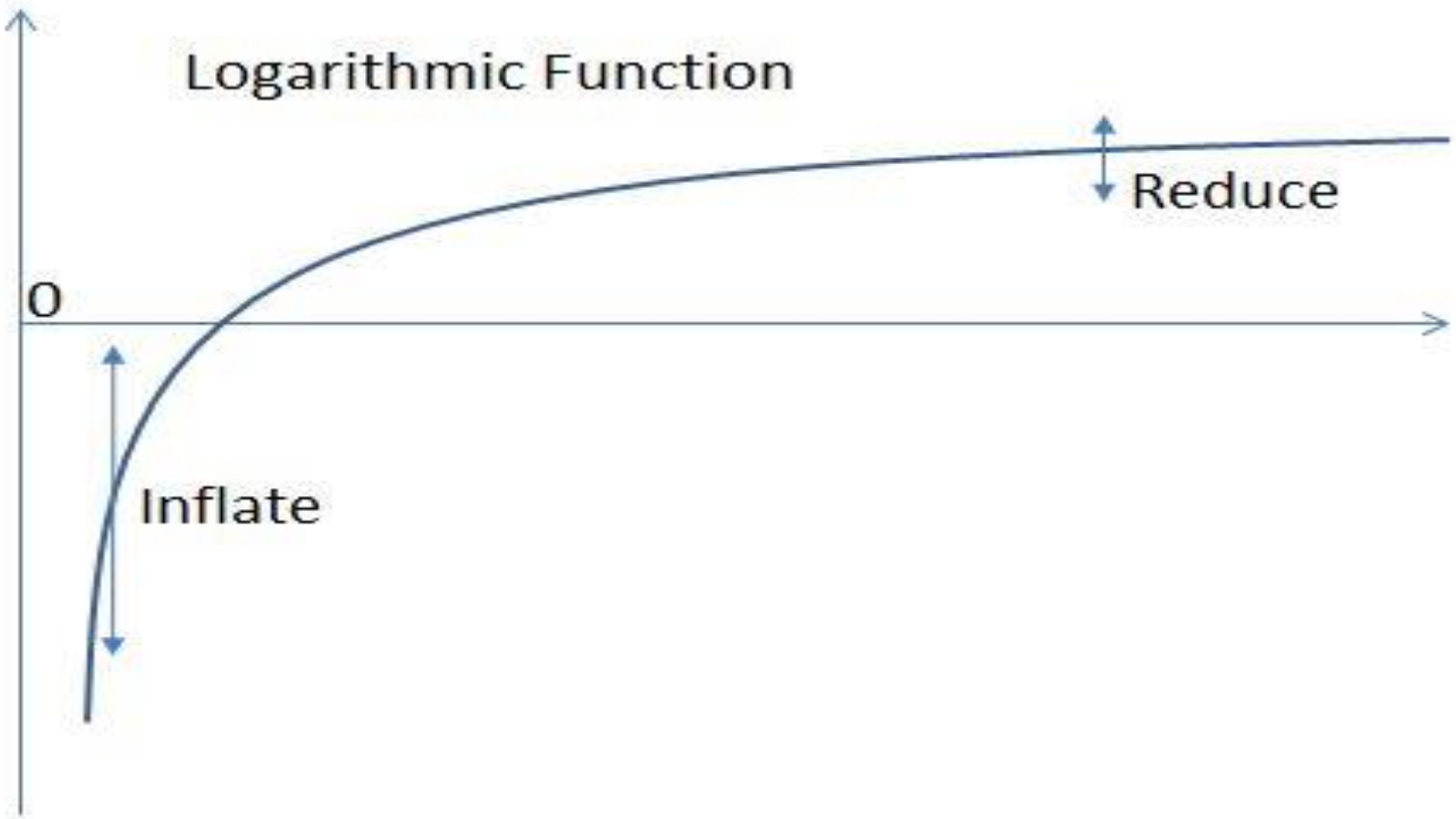
- Negative skewness indicates that the mean of the data values is less than the median, and the data distribution is left-skewed.

- Positive skewness would indicates that the mean of the data values is larger than the median, and the data distribution is right-skewed.

How do you resolve Skewness?



Behavior of Logarithmic Function



Impact of Logarithmic Function

The differences between smaller values will be expanded because the slope of the logarithmic function is steeper when values are small.

The differences between larger values will be reduced because of the very moderate slope of the log distribution for larger values.

If you inflate differences on the left tail and reduce differences on the right side tail, the result will be a symmetrical normal distribution.

Generalization: Box-Cox Transform

- Replacing the data with the log, square root, or inverse may help to remove the skew.

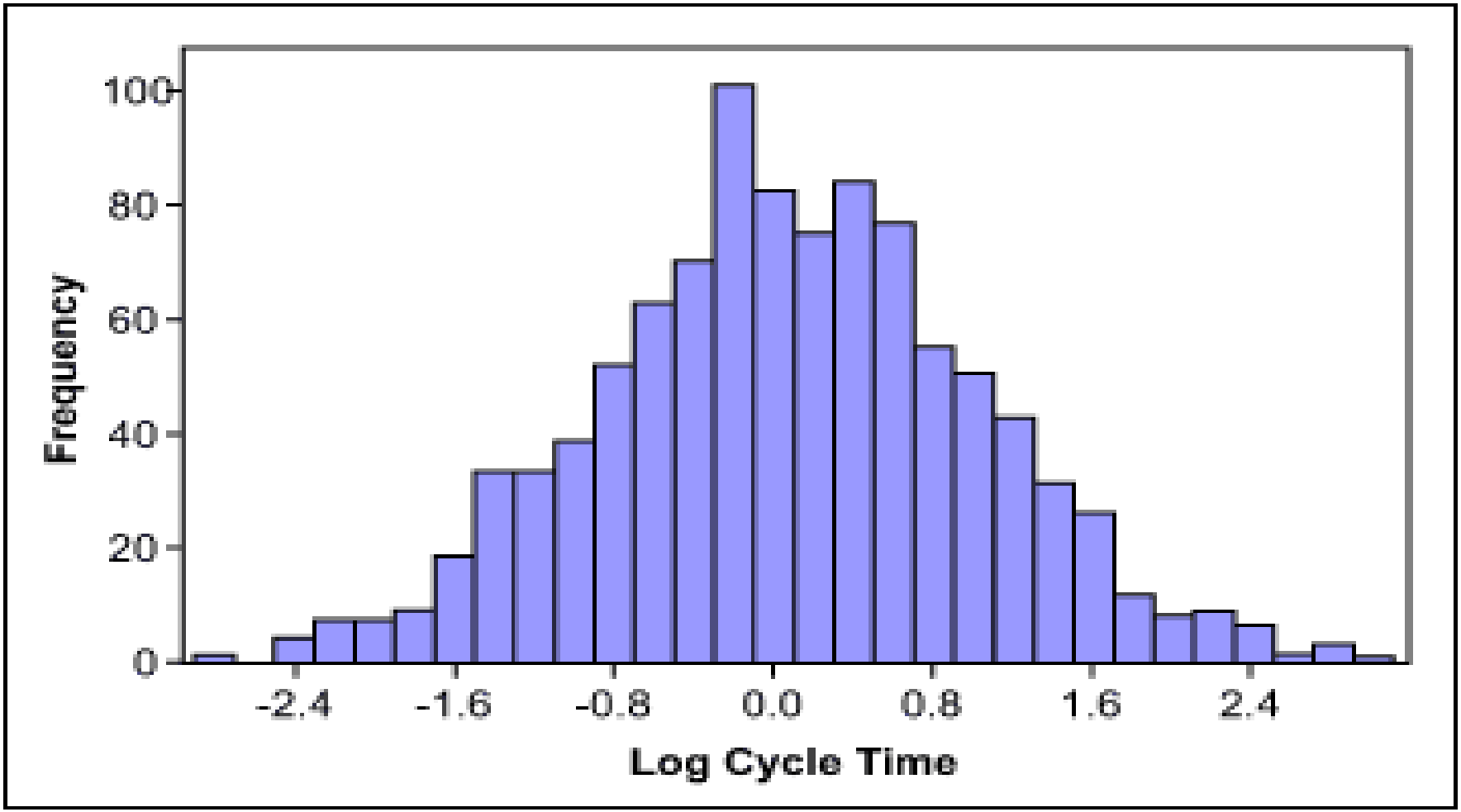
- Box-Cox power transformation provides mathematical generalization to achieve that based on estimated lamda value from given data:

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

Common Box-Cox Transformations

Lambda	X’
-2	$X^{-2} = 1/X^2$
-1	$X^{-1} = 1/X^1$
-0.5	$X^{-0.5} = 1/(\text{Sqrt}(X))$
0	log(X)
0.5	$X^{0.5} = \text{Sqrt}(X)$
1	$X^1 = X$
2	X^2

Box-Cox Transformed Data: Logarithm



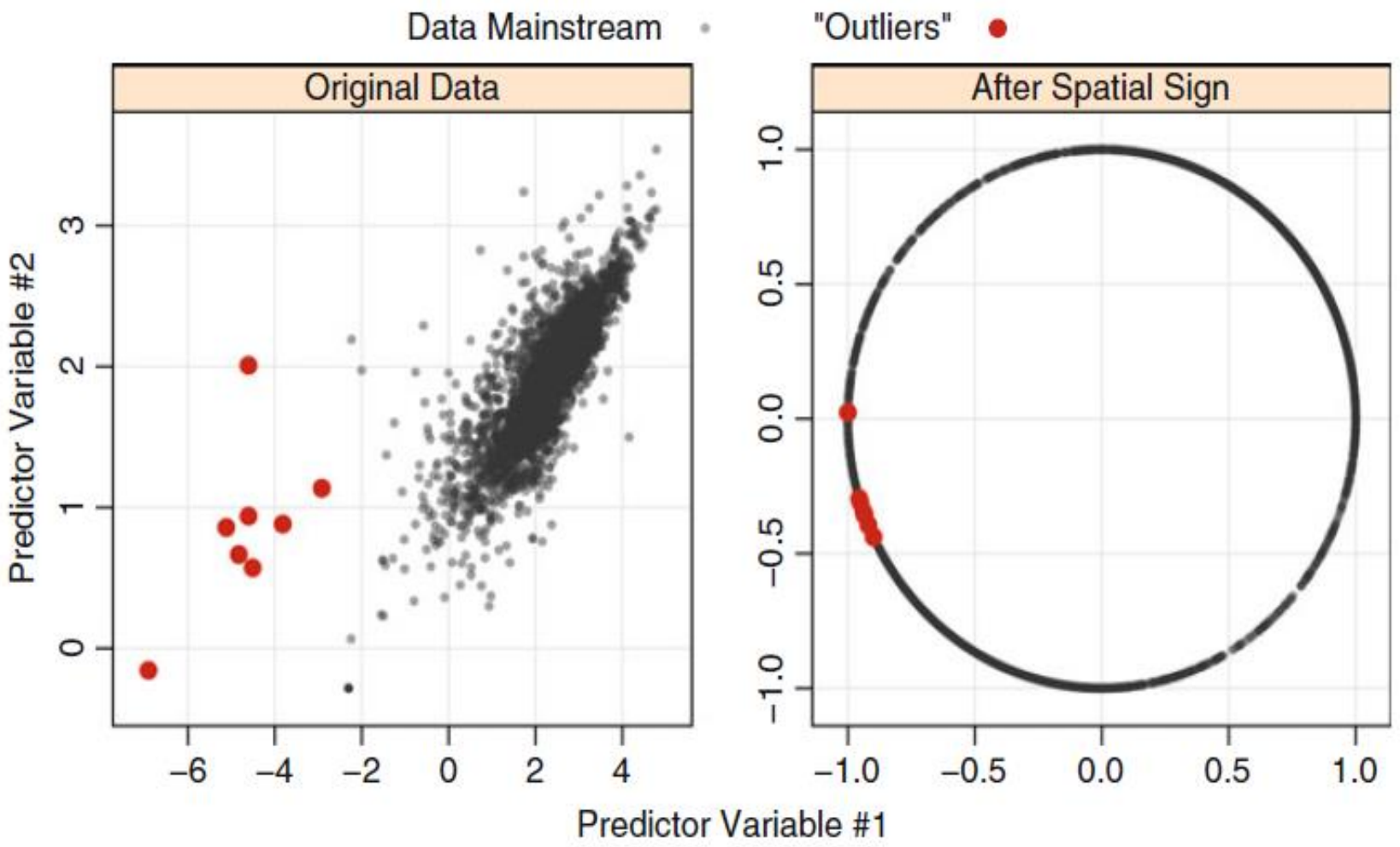
Note: When comparing transformed data, everything under comparison must be transformed in the same way.

Handling Noise/Outliers

Spatial Sign

If a model is considered to be sensitive to outliers, one data transformation that can minimize the problem is the spatial sign.

This procedure projects the predictor values onto a multidimensional sphere. This has the effect of making all the samples the same distance from the center of the sphere.


Data Mainstream
"Outliers"
Original Data
After Spatial Sign
Predictor Variable #2
Predictor Variable #1
0
1
2
3
−6
−4
−2
0
2
4
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0

Handling Missing Values