# VALUE\_STRATEGY and VALUE\_STRATEGY\_ARGS Documentation

## Introduction

This document serves as a comprehensive guide for analysts using SDG2. It is crucial to understand the various VALUE\_STRATEGY and VALUE\_STRATEGY\_ARGS options available when filling out the input sheet in Excel for data generation.

Each VALUE\_STRATEGY dictates the type of synthetic data to be generated for a particular column, and VALUE\_STRATEGY\_ARGS provides additional parameters to fine-tune the generated data.

**Note**: All strategies mentioned here are single-column strategies. The VALUE\_STRATEGY\_ARGS must be entered in the format {min = 0, max = 10} where applicable.

## Detailed Strategy Descriptions

### Linear Strategies

#### LinearRandomInteger

* **VALUE\_STRATEGY\_ARGS**: {min = 0, max = 10}
* **Default**: {min = 0, max = 100}
* **Constraints**: Min must be less than max at all times.
* **Description**: Generates a random integer between the specified min and max values.

#### LinearRandomDouble

* **VALUE\_STRATEGY\_ARGS**: {min = 0, max = 10}
* **Default**: {min = 0, max = 1}
* **Constraints**: Min must be less than max at all times.
* **Description**: Generates a random double between the specified min and max values.

### Bool Strategies

#### RandomBoolean

* **VALUE\_STRATEGY\_ARGS**: {p\_true = 0.5}
* **Default**: p\_true = 0.5
* **Description**: Generates a boolean value, true or false. The probability of generating true is specified by p\_true.

### Chrono Strategies

#### LinearRandomLocalDateTime

* **VALUE\_STRATEGY\_ARGS**: {min = "YYYY-MM-ddTHH:mm:ss", max = "YYYY-MM-ddTHH:mm:ss"}
* **Default**: Min is one year before the current date-time, and Max is one day before the current date-time.
* **Description**: Generates a random LocalDateTime value between the specified min and max date-times.

### Finance Strategies

#### Big4CreditCard

* **VALUE\_STRATEGY\_ARGS**: None
* **Description**: Generates a valid credit card number from VISA, MasterCard, Amex, or Discover. The number passes the Luhn Checksum test.

### People Strategies

#### Email

* **VALUE\_STRATEGY\_ARGS**: None
* **Description**: Generates a random email address using the faker library.

#### FullName

* **VALUE\_STRATEGY\_ARGS**: None
* **Description**: Generates a random full name using the faker library.

### Text Strategies

#### LoremIpsumSentence

* **VALUE\_STRATEGY\_ARGS**: {min\_words = 3, max\_words = 10}
* **Default**: {min\_words = 3, max\_words = 10}
* **Description**: Generates lorem ipsum text with a word count between the specified min and max words.

### Statistical Distribution Strategies

#### NormalDistribution

* **VALUE\_STRATEGY\_ARGS**: {mu, sigma}
* **Description**: Generates a random double value based on a normal distribution with mean μμ and standard deviation σσ.

#### LogNormalDistribution

* **VALUE\_STRATEGY\_ARGS**: {mu, sigma}
* **Default**: {mu = 0.0, sigma = 1.0}
* **Description**: Generates a random double value based on a log-normal distribution with mean μμ and standard deviation σσ.

#### ExponentialDistribution

* **VALUE\_STRATEGY\_ARGS**: {lambda}
* **Default**: {lambda = 1.0}
* **Description**: Generates a random double value based on an exponential distribution with rate parameter λλ.

For further clarification or inquiries, please consult the technical support team.

# Statistical Distribution Strategies: Mathematical Background and Output Explanation

## Introduction

The statistical distribution strategies provide a way to generate synthetic data that follows specific statistical distributions. These strategies come with parameters (VALUE\_STRATEGY\_ARGS) that control the shape and properties of the distribution. Understanding the mathematical underpinning of these parameters can offer better control over the synthetic data generated.

Below are the details about each statistical distribution strategy, the mathematics behind the parameters, and what kind of output to expect.

## Normal Distribution (NormalDistribution)

### Mathematical Background

The Normal Distribution is defined by the probability density function (PDF):

f(x∣μ,σ2)=12πσ2exp⁡(−(x−μ)22σ2)f(x∣μ,σ2)=2πσ2

​1​exp(−2σ2(x−μ)2​)

* μμ: Mean of the distribution.
* σσ: Standard deviation of the distribution.

### Parameters (VALUE\_STRATEGY\_ARGS)

* mu: The mean (μμ) of the distribution.
* sigma: The standard deviation (σσ) of the distribution.

### Output Explanation

Generated data will be a series of real numbers that, when aggregated, will approximate the shape of a Normal Distribution centered around μμ with a standard deviation of σσ.

## Log-Normal Distribution (LogNormalDistribution)

### Mathematical Background

The Log-Normal Distribution is defined by the PDF:

f(x∣μ,σ2)=1xσ2πexp⁡(−(ln⁡x−μ)22σ2)f(x∣μ,σ2)=xσ2π

​1​exp(−2σ2(lnx−μ)2​)

* μμ: Mean of the natural logarithm of the distribution.
* σσ: Standard deviation of the natural logarithm of the distribution.

### Parameters (VALUE\_STRATEGY\_ARGS)

* mu: The mean (μμ) of the natural logarithm of the distribution.
* sigma: The standard deviation (σσ) of the natural logarithm of the distribution.

### Output Explanation

Generated data will be a series of real numbers greater than zero. When the natural logarithm of these numbers is taken, the resulting distribution will approximate a Normal Distribution with mean μμ and standard deviation σσ.

## Exponential Distribution (ExponentialDistribution)

### Mathematical Background

The Exponential Distribution is defined by the PDF:

f(x∣λ)=λe−λxf(x∣λ)=λe−λx

* λλ: The rate parameter, which is the reciprocal of the mean (1meanmean1​).

### Parameters (VALUE\_STRATEGY\_ARGS)

* lambda: The rate parameter (λλ) of the distribution.

### Output Explanation

Generated data will be a series of real numbers greater than or equal to zero. The distribution will approximate an Exponential Distribution with a rate parameter λλ.

## Graphical Validation

If you already have a graphical representation of the validation, overlaying it with a plotted line of the ideal distribution serves as a strong verification mechanism. The closer your generated data matches the ideal distribution, the better your synthetic data mimics real-world data distributions.

For further inquiries or clarification, please consult the technical support team.

# Statistical Distribution Strategies: Impact of Parameter Changes

## Introduction

Adjusting the parameters (VALUE\_STRATEGY\_ARGS) for statistical distribution strategies can significantly impact the shape and properties of the generated synthetic data. Below is an exploration into how varying these parameters affects each of the statistical distributions.

## Normal Distribution (NormalDistribution)

### Mathematical Background and Parameters

Refer to the previous section for the formal mathematical background.

### Impact of Changing Parameters

* **Changing mu (Mean)**:
  + **Effect**: Shifts the entire distribution left or right along the x-axis.
  + **Use-case**: Useful when you want to model data that is centered around a different value than zero.
* **Changing sigma (Standard Deviation)**:
  + **Effect**: Alters the "width" or "spread" of the distribution. A higher value will result in a wider distribution.
  + **Use-case**: Useful when the data points are spread out over a larger range.

### Output Explanation

As previously explained, the output will be a series of real numbers approximating a Normal Distribution.

## Log-Normal Distribution (LogNormalDistribution)

### Mathematical Background and Parameters

Refer to the previous section for the formal mathematical background.

### Impact of Changing Parameters

* **Changing mu (Mean of Logarithm)**:
  + **Effect**: Shifts the entire distribution left or right along the x-axis on a logarithmic scale.
  + **Use-case**: Useful for modeling data sets where the median needs to be adjusted.
* **Changing sigma (Standard Deviation of Logarithm)**:
  + **Effect**: Alters the "width" or "spread" of the distribution on a logarithmic scale. A higher value results in a wider distribution.
  + **Use-case**: Useful when the data has a large range and is positively skewed.

### Output Explanation

Generated data will be a series of real numbers greater than zero. Their natural logarithm will approximate a Normal Distribution.