AIMMS Language Reference - Available Distributions and Statistical Operators

This file contains only one chapter of the book. For a free download of the complete book in pdf format, please visit www.aimms.com.

AIMMS B.V.
Diakenhuisweg 29-35
2033 AP Haarlem
The Netherlands
Tel.: +31 23 5511512

AIMMS Inc.
11711 SE 8th Street
Suite 303
Bellevue, WA 98005
USA
Tel.: +1 425 458 4024

AIMMS Pte. Ltd.
55 Market Street #10-00
Singapore 048941
Tel.: +65 6521 2827

AIMMS
SOHO Fuxing Plaza No.388
Building D-71, Level 3
Madang Road, Huangpu District
Shanghai 200025
China
Tel.: ++86 21 5309 8733

Email: info@aimms.com
WWW: www.aimms.com

Appendices

Appendix A

Distributions, statistical operators and histogram functions

This chapter

This chapter provides a more elaborate description of the distributions and distribution and sample operators listed in Tables 6.5, 6.6 and 6.7. You can use this information when you want to set up an experiment around your (optimization-based) AIMMS model.

Description of distributions

For each of the available distributions we describe

- its parameters, mean and variance,
- the unit relationship between its parameters and result,
- its shape, and
- its typical use in applications.

Such information may be useful in the selection and use of a distribution to describe the particular statistical behavior of input data of experiments that you want to perform on top of your model. However, a general guideline for choosing the right might be in order and is provided in the next paragraph.

Choosing the right distribution

Whenever your experiment counts a number of occurrences, you should first make a distinction between experiments with replacement (i.e. throwing dice), experiments without replacement (i.e. drawing cards from a deck), or experiments in which independent occurrences take place at random moments (i.e. customers appearing at a desk). Having made this distinction, Table A.1 will help you to select the right distribution for your experiment. In any other case the `Normal` distribution should be considered first. Although this distribution is unbounded, it is declining so rapidly that it can often be used even when the result should be bounded. If the `Normal` distribution does not suffice, the primary selection criterium is existence of bounds: AIMMS provides the user with distributions with no bounds, one (lower) bound and two (upper and lower) bounds. See section A.2 (continuous distributions) for details.

Description of distribution operators

For each of the available distribution and sample operators we provide

- the interpretation of its result, and
- the formula for the computation of the operator.

Such information may be useful when you want to perform an analysis of the results of your experiments.

Option for backward compatibility

All distribution operators that are listed in Section A.3 have been introduced in AIMMS 3.4, although the DistributionCumulative and DistributionInverseCumulative operator were already available under the names CumulativeDistribution and InverseCumulativeDistribution, respectively. Furthermore, in order to obtain a consistent set of distribution functions the prototype for some of them has been slightly adapted. Section A.2 discusses the function prototype of the continuous distribution functions in full detail. Both the old and the new function prototypes are discussed in the AIMMS Function Reference. To make sure that models using distribution functions and developed in an older version of AIMMS are working correctly, you should set the option Distribution_compatibility to 'AIMMS 3.0'.

A.1 Discrete distributions

Discrete distributions

We start our discussion with the discrete distributions available in AIMMS. They are

- the Binomial distribution,
- the HyperGeometric distribution,
- the Poisson distribution,
- the Negative Binomial distribution, and
- the Geometric distribution.

Discrete distributions describing successes

The Binomial, HyperGeometric and Poisson distributions describe the number of times that a particular outcome (referred to as "success") occurs. In the Binomial distribution, the underlying assumption is a fixed number of trials and a constant likelihood of success. In the HyperGeometric distribution, the underlying assumption is "sampling without replacement": A fixed number of trials are taken from a population. Each element of this population denotes a success or failure and cannot occur more than once. In the Poisson distribution the number of trials is not fixed. Instead we assume that successes occur independently of each other and with equal chance for all intervals with the same duration.

	with replacement	**without replacement**	**independent occurrences at random moments**
example	throwing dice	drawing cards	serving customers
# trials until first success / time until first occurrence	`Geometric`	not supported in AIMMS	`Exponential` (continuous)
# trials until n-th success / time until n-th occurrence	`Negative Binomial`	not supported in AIMMS	`Gamma` (continuous)
# successes in fixed # trials / # successes in fixed time	`Binomial`	`Hypergeometric`	`Poisson`

Table A.1: Overview of discrete distributions in AIMMS

Distributions describing trials

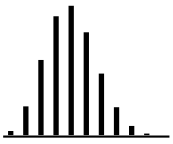
The `Negative Binomial` distribution describes the number of failures before a specified number of successes have occurred. It assumes a constant chance of success for each trial, so it is linked to the `Binomial` distribution. Similarly, the distribution linked to `Poisson` distribution that describes the amount of time until a certain number of successes have occurred is known as the `Gamma` distribution and is discussed in Section A.2. The `Negative Binomial` distribution is a special case of the `Geometric` distribution and describes the number of failures before the first success occurs. Similarly, the `Exponetial` distribution is a special case of the `Gamma` distribution and describes the amount of time until the first occurrence.

Discrete distributions overview

Table A.1 shows the relation between the discrete distributions. The continuous `Exponential` and `Gamma` distribution naturally fit in this table as they represent the distribution of the time it takes before the first/n-th occurrence (given the average time between two consecutive occurrences).

`Binomial` distribution

The `Binomial`(p, n) distribution:

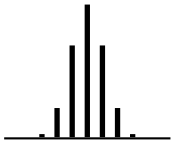


- **Input parameters** : Probability of success p and number of trials n
- **Input check** : integer $n > 0$ and $0 < p < 1$
- **Permitted values** : $\{i \mid i = 0, 1, \ldots, n\}$
- **Formula** : $P(X = i) = \binom{n}{i} p^i (1-p)^{n-i}$
- **Mean** : np
- **Variance** : $np(1-p)$
- **Remarks** : `Binomial`(p, n) = `HyperGeometric`(p, n, ∞)

A typical example for this distribution is the number of defectives in a batch of manufactured products where a fixed percentage was found to be defective in previously produced batches. Another example is the number of persons in a group voting yes instead of no, where the probability of yes has been determined on the basis of a sample.

HyperGeometric distribution

The HyperGeometric(p,n,N) distribution:

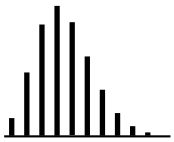


- **Input parameters** : Known initial probability of success p, number of trials n and population size N
- **Input check** : integer $n,N : 0 < n \leq N$, and $p \in \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-1}{N}$
- **Permitted values** : $\{i \mid i = 0,1,\ldots,n\}$
- **Formula** : $P(X=i) = \dfrac{\binom{Np}{i}\binom{N(1-p)}{n-i}}{\binom{N}{n}}$
- **Mean** : np
- **Variance** : $np(1-p)\frac{N-n}{N-1}$

As an example of this distribution, consider a set of 1000 books of which 30 are faulty When considering an order containing 50 books from this set, the HyperGeometric(0.03,50,1000) distribution shows the probability of observing i $(i = 0,1,\ldots,n)$ faulty books in this subset.

Poisson distribution

The Poisson(λ) distribution:

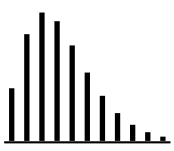


- **Input parameters** : Average number of occurrences λ
- **Input check** : $\lambda > 0$
- **Permitted values** : $\{i \mid i = 0,1,\ldots\}$
- **Formula** : $P(X=i) = \dfrac{\lambda^i}{i!}e^{-\lambda}$
- **Mean** : λ
- **Variance** : λ
- **Remarks** : Poisson$(\lambda) = \lim_{p \downarrow 0}$Binomial$(p, \lambda/p)$

The Poisson distribution should be used when there is an constant chance of a 'success' over time or (as an approximation) when there are many occurrences with a very small individual chance of 'success'. Typical examples are the number of visitors in a day, the number of errors in a document, the number of defects in a large batch, the number of telephone calls in a minute, etc.

Negative Binomial distribution

The NegativeBinomial(p,r) distribution:

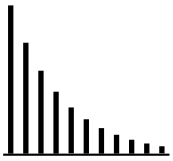


- **Input parameters** : Success probability p and number of successes r
- **Input check** : $0 < p < 1$ and $r = 1,2,\ldots$
- **Permitted values** : $\{i \mid i = 0,1,\ldots\}$
- **Formula** : $P(X=i) = \binom{r+i-1}{i} p^r (1-p)^i$
- **Mean** : $r/p - r$
- **Variance** : $r(1-p)/p^2$

Whenever there is a repetition of the same activity, and you are interested in observing the r-th occurrence of a particular outcome, then the `Negative Binomial` distribution might be applicable. A typical situation is going from door-to-door until you have made r sales, where the probability of making a sale has been determined on the basis of previous experience. Note that the `NegativeBinomial` distribution describes the number of *failures* before the r-th success. The distribution of the number of *trials* i before the r-th success is given by $P_{\texttt{NegativeBinomial(p,r)}}(X = i - r)$.

Geometric distribution

The `Geometric`(p) distribution:



- **Input parameters** : Probability of success p
- **Input check** : $0 < p < 1$
- **Permitted values** : $\{i \mid i = 0, 1, \ldots\}$
- **Formula** : $P(X = i) = (1 - p)^i p$
- **Mean** : $1/p - 1$
- **Variance** : $(1 - p)/p^2$
- **Remarks** : `Geometric`(p) = `NegativeBinomial`$(p, 1)$

The `Geometric` distribution is a special case of the `NegativeBinomial` distribution. So it can be used for the same type of problems (the number of visited doors before the first sale). Another example is an oil company drilling wells until a producing well is found, where the probability of success is based on measurements around the site and comparing them with measurements from other similar sites.

A.2 Continuous distributions

Continuous distributions

In this section we discuss the set of continuous distributions available in AIMMS.

The three distributions with both lower and upper bound are

- the `Uniform` distribution,
- the `Triangular` distribution, and
- the `Beta` distribution.

The five distributions with only a lower bound are

- the `LogNormal` distribution,
- the `Exponential` distribution,
- the `Gamma` distribution,
- the `Weibull` distribution, and
- the `Pareto` distribution.

The three unbounded distributions are

- the `Normal` distribution,

- the Logistic distribution, and
- the Extreme Value distribution.

Parameters of continuous distributions

Every parameter of a continuous distributions can be characterized as either a *shape* parameter β, a *location* parameter l, or a *scale* parameter s. While the presence and meaning of a shape parameter is usually distribution-dependent, location and scale parameters find their origin in the common transformation

$$x \mapsto \frac{x - l}{s}$$

to shift and stretch a given distribution. By choosing $l = 0$ and $s = 1$ the standard form of a distribution is obtained. If a certain distribution has n shape parameters ($n \geq 0$), these shape parameters will be passed as the first n parameters to AIMMS. The shape parameters are then followed by two optional parameters, with default values 0 and 1 respectively. For double-bounded distributions these two optional parameters can be interpreted as a lower and upper bound (the value of the location parameter l for these distributions is equal to the lower bound and the value of the scale parameter s is equal to the difference between the upper and lower bound). For single-bounded distributions the bound value is often used as the location parameter l. In this section, whenever the location parameter can be interpreted as a mean value or whenever the scale parameter can be interpreted as the deviation of a distribution, these more meaningful names are used to refer to the parameters. Note that the LogNormal, Gamma and Exponential distributions are distrubutions that will mostly be used with location parameter equal to 0.

Transformation to standard form

When transforming a distribution to standard form, distribution operators change. Section A.5 (scaling of statistical operators) gives the relationships between distribution operators working on random variables $X(l, s)$ and $X(0, 1)$.

Units of measurement

When a random variable representing some real-life quantity with a given unit of measurement (see also Chapter 32) is distributed according to a particular distribution, some parameters of that distribution are also naturally expressed in terms of this same unit while other parameters are expected to be unitless. In particular, the location and scale parameters of a distribution are measured in the same unit of measurement as the corresponding random variable, while shape parameters (within AIMMS) are implemented as unitless parameters.

Unit notation in this appendix

When you use a distribution function, AIMMS will perform a unit consistency check on its parameters and result, whenever your model contains one or more QUANTITY declarations. In the description of the continuous distributions below, the expected units of the distribution parameters are denoted in square brackets. Throughout the sequel, $[x]$ denotes that the parameter should have the same unit of measurement as the random variable X and [–] denotes that a parameter should be unitless.

A commonly used distribution

In practice, the Normal distribution is used quite frequently. Such widespread use is due to a number of pleasant properties:

- the Normal distribution has no shape parameters and is symmetrical,
- random values are more likely as they are closer to the mean value,
- it can be directly evaluated for any given mean and standard deviation because it is fully specified through the mean and standard deviation parameter,
- it can be used as a good approximation for distributions on a finite interval, because its probability density is declining fast enough (when moving away from the mean),
- the mean and sum of any number of uncorrelated Normal distributions are Normal distributed themselves, and thus have the same shape, and
- the mean and sum of a large number of uncorrelated distributions are always approximately Normal distributed.

Distributions for double bounded variables

For random variables that have a known lower and upper bound, AIMMS provides three continuous distributions on a finite interval: the Uniform, Triangular and Beta distribution. The Uniform (no shape parameters) and Triangular (one shape parameter) distributions should be sufficient for most experiments. For all remaining experiments, the user might consider the highly configurable Beta (two shape parameters) distribution.

Distributions for single bounded variables

When your random variable only has a single bound, you should first check whether the Gamma distribution can be used or whether the Normal distribution is accurate enough. The LogNormal distribution should be considered if the most likely value is near but not at the bound. The Weibull or Gamma distribution ($\beta > 1$), or even the ExtremeValue distribution are alternatives, while the Weibull or Gamma distribution ($\beta \leq 1$) or Pareto distribution should be considered if the bound is the most likely value.

The Gamma distribution

The Gamma (and as a special case thereof the Exponential) distribution is widely used for its special meaning. It answers the question: how long does it take for a success to occur, when you only know the average number of occurrences (like in the Poisson distribution). The Exponential distribution gives the time to the first occurrence, and its generalization, the Gamma(β) distribution gives the time to the β-th occurrence. Note that the sum of a Gamma(β_1, l_1, s) and Gamma(β_2, l_2, s) distribution has a Gamma($\beta_1 + \beta_2, l_1 + l_2, s$) distribution.

The LogNormal distribution

If you assume the logarithm of a variable to be Normal distributed, the variable itself is LogNormal-distributed. As a result, it can be shown that the chance of an outcome in the interval $[x \cdot c_1, x \cdot c_2]$ is equal to the chance of an outcome in the interval $[x/c_2, x/c_1]$ for some x. This might be a reasonable assumption in price developments, for example.

Uniform distribution

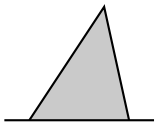
The Uniform(*min*,*max*) distribution:

- **Input parameters** : $min\,[x]$, $max\,[x]$
- **Input check** : $min < max$
- **Permitted values** : $\{x \mid min \le x \le max\}$
- **Standard density** : $f_{(0,1)}(x) = 1$
- **Mean** : $1/2$
- **Variance** : $1/12$

In the Uniform distribution all values of the random variable occur between a fixed minimum and a fixed maximum with equal likelihood. It is quite common to use the Uniform distribution when you have little knowledge about an uncertain parameter in your model except that its value has to lie anywhere within fixed bounds. For instance, after talking to a few appraisers you might conclude that their single appraisals of your property vary anywhere between a fixed pessimistic and a fixed optimistic value.

Triangular distribution

The Triangular(β,*min*,*max*) distribution:

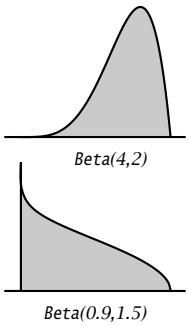


- **Input parameters** : shape β [−],$min\,[x]$, $max\,[x]$
- **Input check** : $min < max$, $0 < \beta < 1$
- **Permitted values** : $\{x \mid min \le x \le max\}$
- **Standard density** : $f_{(\beta,0,1)}(x) = \begin{cases} 2x/\beta & \text{for } 0 \le x \le \beta \\ 2(1-x)/(1-\beta) & \text{for } \beta < x \le 1 \end{cases}$
- **Mean** : $(\beta + 1)/3$
- **Variance** : $(1 - \beta + \beta^2)/18$
- **Remarks** : The shape parameter β indicates the position of the peak in relation to the range, i.e. $\beta = \frac{peak-min}{max-min}$

In the Triangular distribution all values of the random variable occur between a fixed minimum and a fixed maximum, but not with equal likelihood as in the Uniform distribution. Instead, there is a most likely value, and its position is not necessarily in the middle of the interval. It is quite common to use the Triangular distribution when you have little knowledge about an uncertain parameter in your model except that its value has to lie anywhere within fixed bounds and that there is a most likely value. For instance, assume that a few appraisers each quote an optimistic as well as a pessimistic value of your property. Summarizing their input you might conclude that their quotes provide not only a well-defined interval but also an indication of the most likely value of your property.

Beta distribution

The Beta(α,β,*min*,*max*) distribution:



- **Input parameters** : shape α [–], shape β [–], *min* [x], *max* [x]
- **Input check** : $\alpha > 0, \beta > 0, min < max$
- **Permitted values** : $\{x \mid min < x < max\}$
- **Standard density** : $f_{(\alpha,\beta,0,1)}(x) = \dfrac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$
  where $B(\alpha,\beta)$ is the Beta function
- **Mean** : $\alpha/(\alpha+\beta)$
- **Variance** : $\alpha\beta(\alpha+\beta)^{-2}(\alpha+\beta+1)^{-1}$
- **Remarks** : Beta(1,1,*min*,*max*)=Uniform(*min*,*max*)

The Beta distribution is a very flexible distribution whose two shape parameters allow for a good approximation of almost any distribution on a finite interval. The distribution can be made symmetrical, positively skewed, negatively skewed, etc. It has been used to describe empirical data and predict the random behavior of percentages and fractions. Note that for $\alpha < 1$ a singularity occurs at $x = min$ and for $\beta < 1$ at $x = max$.

LogNormal distribution

The LogNormal(β,*min*,*s*) distribution:

- **Input parameters** : shape β [–], lowerbound *min* [x] and scale s [x]
- **Input check** : $\beta > 0$ and $s > 0$
- **Permitted values** : $\{x \mid min < x < \infty\}$
- **Standard density** : $f_{(\beta,0,1)}(x) = \dfrac{1}{\sqrt{2\pi}x\ln(\beta^2+1)} e^{\frac{-(\ln(x^2(\beta^2+1)))}{2\ln(\beta^2+1)}}$
- **Mean** : 1
- **Variance** : β^2

If you assume the logarithm of the variable to be Normal(μ,σ)-distributed, then the variable itself is LogNormal($\sqrt{e^{\sigma^2}-1}, 0, e^{\mu-\sigma^2/2}$)-distributed. This parameterization is used for its simple expressions for mean and variance. A typical example is formed by real estate prices and stock prices. They all cannot drop below zero, but they can grow to be very high. However, most values tend to stay within a particular range. You usually can form some expected value of a real estate price or a stock price, and estimate the standard deviation of the prices on the basis of historical data.

Exponential distribution

The Exponential(*min*,*s*) distribution:

- **Input parameters** : lowerbound *min* [x] and scale s [x]
- **Input check** : $s > 0$
- **Permitted values** : $\{x \mid min \le x < \infty\}$
- **Standard density** : $f_{(0,1)}(x) = \lambda e^{-x}$
- **Mean** : 1
- **Variance** : 1
- **Remarks** : Exponential(*min*,*s*) = Gamma(1,*min*,*s*)
  Exponential(*min*,*s*) = Weibull(1,*min*,*s*)

Assume that you are observing a sequence of independent events with a constant chance of occurring in time, with s being the average time between occurrences. (in accordance with the `Poisson` distribution) The `Exponential`$(0, s)$ distribution gives answer to the question: how long a time do you need to wait until you observe the first occurrence of an event. Typical examples are time between failures of equipment, and time between arrivals of customers at a service desk (bank, hospital, etc.).

Gamma distribution

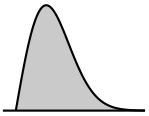
The `Gamma`(β,min,s) distribution:

- **Input parameters** : shape β [–], lowerbound min [x] and scale s [x]
- **Input check** : $s > 0$ and $\beta > 0$
- **Permitted values** : $\{x \mid min < x < \infty\}$
- **Standard density** : $f_{(\beta,0,1)}(x) = x^{\beta-1}e^{-x}/\Gamma(\beta)$ where $\Gamma(\beta)$ is the Gamma function
- **Mean** : β
- **Variance** : β

The `Gamma` distribution gives answer to the question: how long a time do you need to wait until you observe the β-th occurrence of an event (instead of the first occurrence as in the `Exponential` distribution). Note that it is possible to use non-integer values for β and a location parameter. In these cases there is no natural interpretation of the distribution and for $\beta < 1$ a singularity exists at $x = min$, so one should be very careful in using the `Gamma` distribution this way.

Weibull distribution

The `Weibull`(β,min,s) distribution:

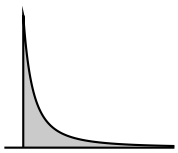


- **Input parameters** : shape β [–], lowerbound min [x] and scale s [x]
- **Input check** : $\beta > 0$ and $s > 0$
- **Permitted values** : $\{x \mid min \leq x < \infty\}$
- **Standard density** : $f_{(\beta,0,1)}(x) = \beta x^{\beta-1}e^{-x^{\beta}}$
- **Mean** : $\Gamma(1 + 1/\beta)$
- **Variance** : $\Gamma(1 + 2/\beta) - \Gamma^2(1 + 1/\beta)$

The `Weibull` distribution is another generalization of the `Exponential` distribution. It has been successfully used to describe failure time in reliability studies, and the breaking strengths of items in quality control testing. By using a value of the shape parameter that is less than 1, the `Weibull` distribution becomes steeply declining and could be of interest to a manufacturer testing failures of items during their initial period of use. Note that in that case there is a singularity at $x = min$.

The Pareto(β,l,s) distribution:

Pareto distribution

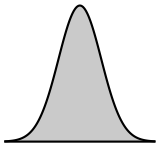


- **Input parameters** : shape β [–], location l [x] and scale s [x]
- **Input check** : $s > 0$ and $\beta > 0$
- **Permitted values** : $\{x \mid l + s < x < \infty\}$
- **Standard density** : $f_{(\beta,0,1)}(x) = \beta / x^{\beta+1}$
- **Mean** : for $\beta > 1$: $\beta/(\beta-1), \infty$ otherwise
- **Variance** : for $\beta > 2$: $\beta(\beta-1)^{-2}(\beta-2)^{-1}, \infty$ otherwise

The Pareto distribution has been used to describe the sizes of such phenomena as human population, companies, incomes, stock fluctuations, etc.

The Normal(μ,σ) distribution:

Normal distribution

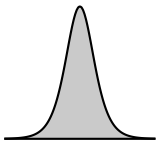


- **Input parameters** : Mean μ [x] and standard deviation σ [x]
- **Input check** : $\sigma > 0$
- **Permitted values** : $\{x \mid -\infty < x < \infty\}$
- **Standard density** : $f_{(0,1)}(x) = e^{-x^2/2}/\sqrt{2\pi}$
- **Mean** : 0
- **Variance** : 1
- **Remarks** : Location μ, scale σ

The Normal distribution is frequently used in practical applications as it describes many phenomena observed in real life. Typical examples are attributes such as length, IQ, etc. Note that while the values in these examples are naturally bounded, a close fit between such data values and normally distributed values is quite common in practice, because the likelihood of extreme values away from the mean is essentially zero in the Normal distribution.

The Logistic(μ,s) distribution:

Logistic distribution



- **Input parameters** : mean μ [x] and scale s [x]
- **Input check** : $s > 0$
- **Permitted values** : $\{x \mid -\infty < x < \infty\}$
- **Standard density** : $f_{(0,1)}(x) = (e^x + e^{-x} + 2)^{-1}$
- **Mean** : 0
- **Variance** : $\pi^2/3$

The Logistic distribution has been used to describe growth of a population over time, chemical reactions, and similar processes. Extreme values are more common than in the somewhat similar Normal distribution

The Extreme Value(l,s) distribution:

Extreme Value distribution

- **Input parameters** : Location l [x] and scale s [x]
- **Input check** : $s > 0$
- **Permitted values** : $\{x \mid -\infty < x < \infty\}$
- **Standard density** : $f_{(0,1)}(x) = e^x e^{-e^x}$

- **Mean** : $\gamma = 0.5772\ldots$ (Euler's constant)
- **Variance** : $\pi^2/6$

Extreme Value distributions have been used to describe the largest values of phenomena observed over time: water levels, rainfall, etc. Other applications include material strength, construction design or any other application in which extreme values are of interest. In literature the Extreme Value distribution that is provided by AIMMS is known as a type 1 Gumbel distribution.

A.3 Distribution operators

Distribution operators

The distribution operators discussed in this section can help you to analyze the results of an experiment. For example, it is expected that the sample mean of a sequence of observations gets closer to the mean of the distribution that was used during the observations as the number of observations increases. To compute statistics over a sample, you can use the sample operators discussed in Section A.4 or you can use the histogram functions that are explained in Section **??** of the Language Reference. The following distribution operators are available in AIMMS:

- the DistributionCumulative(*distr,x*) operator,
- the DistributionInverseCumulative(*distr,α*) operator,
- the DistributionDensity(*distr,x*) operator,
- the DistributionInverseDensity(*distr,α*) operator,
- the DistributionMean(*distr*) operator,
- the DistributionDeviation(*distr*) operator,
- the DistributionVariance(*distr*) operator,
- the DistributionSkewness(*distr*) operator, and
- the DistributionKurtosis(*distr*) operator.

Cumulative distributions ...

DistributionCumulative(*distr,x*) computes the probability that a random variable X drawn from the distribution *distr* is less or equal than x. Its inverse, DistributionInverseCumulative(*distr*, α), computes the smallest x such that the probability that a variable X is greater than or equal to x does not exceed α.

... and their derivatives

The DistributionDensity(*distr,x*) expresses the expected density around x of sample points drawn from a *distr* distribution and is in fact the derivative of DistributionCumulative(*distr,x*). The DistributionInverseDensity(*distr*, α) is the derivative of DistributionInverseCumulative(*distr*, α). Given a random variable X, the DistributionInverseDensity can be used to answer the question of how much a given value x should be increased such that the probability $P(X \leq x)$ is increased with α (for small values of α).

... for discrete distributions

For continuous distributions *distr*, $\alpha \in [0,1]$, and x = DistributionInverseCumulative(*distr*, α) it holds that

$$\begin{aligned} \texttt{DistributionDensity}(\textit{distr}, x) &= \partial\alpha/\partial x \\ \texttt{DistributionInverseDensity}(\textit{distr}, \alpha) &= \partial x/\partial\alpha \end{aligned}$$

Note that the above two relations make it possible to express DistributionInverseDensity in terms of DistributionDensity. Through this relation the DistributionInverseDensity is also defined for discrete distributions.

Distribution statistics

The operators DistributionMean, DistributionDeviation, DistributionVariance, DistributionSkewness and DistributionKurtosis provide the mean, standard deviation, variance, skewness and kurtosis of a given distribution. Note that the values computed using the sample operators converges to the values computed using the corresponding distribution operators as the size of the sample increases (the law of large numbers).

A.4 Sample operators

Sample operators

The statistical sample operators discussed in this section can help you to analyze the results of an experiment. The following operators are available in AIMMS:

- the Mean operator,
- the GeometricMean operator,
- the HarmonicMean operator,
- the RootMeanSquare operator,
- the Median operator,
- the SampleDeviation operator,
- the PopulationDeviation operator,
- the Skewness operator,
- the Kurtosis operator,
- the Correlation operator, and
- the RankCorrelation operator.

Associated units

The results of the Skewness, Kurtosis, Correlation and RankCorrelation operator are unitless. The results of the other sample operators listed above should have the same unit of measurement as the expression on which the statistical computation is performed. Whenever your model contains one or more QUANTITY declarations, AIMMS will perform a unit consistency check on arguments of the statistical operators and their result.

Mean

The following mean computation methods are supported: (arithmetic) mean or average, geometric mean, harmonic mean and root mean square (RMS). The first method is well known and has the property that it is an unbiased estimate of the expectation of a distribution. The geometric mean is defined as the N-th root of the product of N values. The harmonic mean is the reciprocal of the arithmetic mean of the reciprocals. The root mean square is defined as the square root of the arithmetic mean of the squares. It is mostly used for averaging the measurements of a physical process.

- **Operator :** `Mean(`*domain*`,`*expression*`)`
- **Formula :** $\frac{1}{n}\sum_{i=1}^{n} x_i$

- **Operator :** `GeometricMean(`*domain*`,`*expression*`)`
- **Formula :** $\sqrt[n]{\prod_{i=1}^{n} x_i}$

- **Operator :** `HarmonicMean(`*domain*`,`*expression*`)`
- **Formula :** $\frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}}$

- **Operator :** `RootMeanSquare(`*domain*`,`*expression*`)`
- **Formula :** $\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2}$

Median

The median is the middle value of a sorted group of values. In case of an odd number of values the median is equal to the middle value. If the number of values is even, the median is the mean of the two middle values.

- **Operator :** `Median(`*domain*`,`*expression*`)`
- **Formula :** $median = \begin{cases} x_{\frac{N+1}{2}} & \text{if } N \text{ is odd} \\ \frac{1}{2}\left(x_{\frac{N}{2}} + x_{\frac{N+2}{2}}\right) & \text{if } N \text{ is even} \end{cases}$

Standard deviation

The standard deviation is a measure of dispersion about the mean. It is defined as the root mean square of the distance of a set of values from the mean. There are two kinds of standard deviation: the standard deviation of a sample of a population, also known as σ_{n-1} or s, and the standard deviation of a population, which is denoted by σ_n. The relation between these two standard deviations is that the first kind is an unbiased estimate of the second kind. This implies that for large n $\sigma_{n-1} \approx \sigma_n$. The standard deviation of an sample of a population can be computed by means of

- **Operator :** `SampleDeviation(`*domain*`,`*expression*`)`

■ **Formula** : $\sqrt{\frac{1}{n-1}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)}$

whereas the standard deviation of a population can be determined by

■ **Operator** : PopulationDeviation(*domain*,*expression*)

■ **Formula** : $\sqrt{\frac{1}{n}\left(\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2\right)}$

Skewness

The skewness is a measure of the symmetry of a distribution. Two kinds of skewness are distinguished: positive and negative. A positive skewness means that the tail of the distribution curve on the right side of the central maximum is longer than the tail on the left side (skewed "to the right"). A distribution is said to have a negative skewness if the tail on the left side of the central maximum is longer than the tail on the right side (skewed "to the left"). In general one can say that a skewness value greater than 1 of less than -1 indicates a highly skewed distribution. Whenever the value is between 0.5 and 1 or -0.5 and -1, the distribution is considered to be moderately skewed. A value between -0.5 and 0.5 indicates that the distribution is fairly symmetrical.

■ **Operator** : Skewness(*domain*,*expression*)

■ **Formula** : $\frac{\sum_{i=1}^{n}(x_i-\mu)^3}{\sigma_{n-1}^3}$

where μ denotes the mean and σ_{n-1} denotes the standard deviation.

Kurtosis

The kurtosis coefficient is a measure for the peakedness of a distribution. If a distribution is fairly peaked, it will have a high kurtosis coefficient. On the other hand, a low kurtosis coefficient indicates that a distribution has a flat peak. It is common practice to use the kurtosis coefficient of the standard Normal distribution, equal to 3, as a standard of reference. Distributions which have a kurtosis coefficient less than 3 are considered to be platykurtic (meaning flat), whereas distributions with a kurtosis coefficient greater than 3 are leptokurtic (meaning peaked). Be aware that in literature also an alternative definition of kurtosis is used in which 3 is subtracted from the formula used here.

■ **Operator** : Kurtosis(*domain*,*expression*)

■ **Formula** : $\frac{\sum_{i=1}^{n}(x_i-\mu)^4}{\sigma_{n-1}^4}$

where μ denotes the mean and σ_{n-1} denotes the standard deviation.

Correlation coefficient

The correlation coefficient is a measurement for the relationship between two variables. Two variables are positive correlated with each other when the correlation coefficient lies between 0 and 1. If the correlation coefficient lies between -1 and 0, the variables are negative correlated. In case the correlation coefficient is 0, the variables are considered to be unrelated to one another. Positive correlation means that if one variable increases, the other variable increases also. Negative correlation means that if one variable increases, the other variable decreases.

- **Operator :** `Correlation(`*domain*`,`*x_expression*`,` *y_expression*`)`
- **Formula :**

$$\frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left(n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right)\left(n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right)}}$$

Rank correlation

If one wants to determine the relationship between two variables, but their distributions are not equal or the precision of the data is not trusted, one can use the rank correlation coefficient to determine their relationship. In order to compute the rank correlation coefficient the data is ranked by their value using the numbers $1, 2, \ldots, n$. These rank numbers are used to compute the rank correlation coefficient.

- **Operator :** `RankCorrelation(`*domain*`,`*x_expression*`,` *y_expression*`)`
- **Formula :**

$$1 - \frac{6\sum_{i=1}^{n} \left(\mathrm{Rank}(x_i) - \mathrm{Rank}(y_i)\right)^2}{n(n^2-1)}$$

A.5 Scaling of statistical operators

Transforming distributions

Shifting and scaling distribution has an effect on the distribution operators, and on sample operators when the samples are from a specified distribution. Location and scale parameters find their origin in a common transformation

$$x \mapsto \frac{x-l}{s}$$

to shift and stretch a given distribution. By choosing $l = 0$ and $s = 1$ one obtains the standard form of a given distribution, and the relation of operators working on the general and standard form of distributions is as follows:

$$
\begin{aligned}
X(l,s) &= l + sX(0,1) \\
\texttt{DistributionDensity}(x;l,s) &= \tfrac{1}{s}\,\texttt{DistributionDensity}(\tfrac{x-l}{s};0,1) \\
\texttt{DistributionInversDensity}(\alpha;l,s) &= s \cdot \texttt{DistributionInversDensity}(\alpha;0,1) \\
\texttt{DistributionCumulative}(x;l,s) &= \texttt{DistributionCumulative}(\tfrac{x-l}{s};0,1) \\
\texttt{DistributionInverseCumulative}(\alpha;l,s) &= l + s \cdot \texttt{DistributionInverseCumulative}(\alpha;0,1) \\
\texttt{Mean}(X(l,s)) &= l + s \cdot \texttt{Mean}(X(0,1)) \\
\texttt{Median}(X(l,s)) &= l + s \cdot \texttt{Median}(X(0,1)) \\
\texttt{Deviation}(X(l,s)) &= s \cdot \texttt{Deviation}(X(0,1)) \\
\texttt{DistributionVariance}(X(l,s)) &= s^2 \cdot \texttt{DistributionVariance}(X(0,1)) \\
\texttt{Skewness}(X(l,s)) &= \texttt{Skewness}(X(0,1)) \\
\texttt{Kurtosis}(X(l,s)) &= \texttt{Kurtosis}(X(0,1)) \\
\texttt{(Rank)Correlation}(X(l,s),Y) &= \texttt{(Rank)Correlation}(X(0,1),Y)
\end{aligned}
$$

Transformation of the mean

The transformation formula for the Mean holds for both the DistributionMean and the Mean of a sample. However, for the GeometricMean, the HarmonicMean and the RootMeanSquare, only the scale factor can be propagated easily during the transformation. Thus, for a sample taken from a distribution X and a any mean operator M from the GeometricMean, HarmonicMean or RootMeanSquare, it holds that

$$M(X(l,s)) = s \cdot M(X(l,1))$$

but in general

$$M(X(l,s)) \neq l + M(X(0,s))$$

Transformation of the other moments

The transformation formula for the deviation is valid for the DistributionDeviation, the SampleDeviation and PopulationDeviation, while the transformation formulae for the Skewness and Kurtosis hold for both the distribution and sample operators.

A.6 Creating histograms

Histogram

The term histogram typically refers to a picture of a number of observations. The observations are divided over equal-length intervals, and the number of observed values in each interval is counted. Each count is referred to as a frequency, and the corresponding interval is called a frequency interval. The picture of a number of observations is then constructed by drawing, for each frequency interval, the corresponding frequency as a bar. A histogram can thus be viewed as a bar chart of frequencies.

Histogram support

The procedures and functions discussed in this section allow you to create histograms based on a large number of trials in an experiment conducted from within your model. You can set up such an experiment by making use of random data for each trial drawn from one or more of the distributions discussed in the AIMMS Language Reference. The histogram frequencies, created through the functions and procedures discussed in this section, can be displayed graphically using the standard AIMMS bar chart object.

Histogram functions and procedures

AIMMS provides the following procedure and functions for creating and computing histograms.

- `HistogramCreate`(*histogram-id*[,*integer-histogram*][,*sample-buffer-size*])
- `HistogramDelete`(*histogram-id*)
- `HistogramSetDomain`(*histogram-id*,*intervals*[,*left*,*width*]
  [,*left-tail*][,*right-tail*])
- `HistogramAddObservation`(*histogram-id*,*value*)
- `HistogramAddObservations`(*histogram-id*,*values-parameter*)
- `HistogramGetFrequencies`(*histogram-id*,*frequency-parameter*)
- `HistogramGetBounds`(*histogram-id*,*left-bound*,*right-bound*)
- `HistogramGetObservationCount`(*histogram-id*)
- `HistogramGetAverage`(*histogram-id*)
- `HistogramGetDeviation`(*histogram-id*)
- `HistogramGetSkewness`(*histogram-id*)
- `HistogramGetKurtosis`(*histogram-id*)

The *histogram-id* argument assumes an integer value. The arguments *frequency-parameter*, *left-bound* and *right-bound* must be one- dimensional parameters (defined over a set of intervals declared in your model). The optional arguments *integer-histogram* (default 0), *left-tail* (default 1) and *right-tail* (default 1) must be either 0 or 1. The optional argument *sample-buffer-size* must be a positive integer, and defaults to 512.

Creating and deleting histograms

Through the procedures `HistogramCreate` and `HistogramDelete` you can create and delete the internal data structures associated with each individual histogram in your experiment. Upon success, the procedure `HistogramCreate` passes back a unique integer number, the *histogram-id*. This reference is required in the remaining procedures and functions to identify the histogram at hand. The observations corresponding to a histogram can be either continuous or integer-valued. AIMMS assumes continuous observations by default. Through the optional *integer-histogram* argument you can indicate that the observations corresponding to a histogram are integer-valued.

Sample buffer size

For every histogram, AIMMS will allocate a certain amount of memory for storing observations. By default, AIMMS allocates space to store samples of 512 observations at most. Using the optional *sample-buffer-size* argument, you can override the default maximum sample size. As long as the number of observations is still smaller than the sample buffer size, all observations will be stored individually. As soon as the actual number of observations exceeds the sample buffer size, AIMMS will no longer store the individual observations. Instead, all observations are then used to determine the frequencies of frequency intervals. These intervals are determined on the basis of the sample collected so far, unless you have specified interval ranges through the procedure HistogramSetDomain.

Setting the interval domain

You can use the function HistogramSetDomain to define frequency intervals manually. You do so by specifying

- the number of fixed-width *intervals*,
- the lower bound of the *left*-most interval (not including a left-tail interval) together with the (fixed) *width* of intervals to be created (optional),
- whether a *left-tail* interval must be created (optional), and
- whether a *right-tail* interval must be created (optional).

The default for the *left* argument is -INF. *Note that the* left *argument is ignored unless the* width *argument is strictly greater than 0.* Note that the selection of one or both of the tail intervals causes a corresponding increase in the number of frequency intervals to be created.

Use of tail intervals

Whenever an observed value is smaller than the lower bound of the left-most fixed-width interval, AIMMS will update the frequency count of the left-tail interval. If the left-tail interval is not present, then the observed value is lost and the procedure HistogramAddObservation and HistogramAddObservations (to be discussed below) will have a return value of 0. Similarly, AIMMS will update the frequency count of the right-tail interval, when an observation lies beyond the right-most fixed-width interval.

Adjusting the interval domain

Whenever, during the course of an experiment, the number of added observations is still below the sample buffer size, you are allowed to modify the interval ranges. As soon as the number of observations exceeds the sample buffer size, AIMMS will have fixed the settings for the interval ranges, and the function HistogramSetDomain will fail. This function will also fail when previous observations cannot be placed in accordance with the specified interval ranges.

Adding observations

You can use the procedure HistogramAddObservation to add a new observed value (or HistogramAddObservations to add a set of values) to a histogram. Non-integer observations for integer-valued histograms will be rounded to the nearest integer value. The procedure will fail, if the observed value cannot be placed in accordance with the specified interval ranges.

Obtaining frequencies

With the procedure HistogramGetFrequencies, you can request AIMMS to fill a one-dimensional parameter (slice) in your model with the observed frequencies. The cardinality of the index domain of the frequency parameter must be at least as large as the total number of frequency intervals (including the tail interval(s) if created). The first element of the domain set is associated with the left-tail interval, if created, or else the left-most fixed-width interval.

Interval determination

If you have provided the number of intervals through the procedure HistogramSetDomain, AIMMS will create this number of frequency intervals plus at most two tail intervals. Without a custom-specified number of intervals, AIMMS will create 16 fixed-width intervals plus two tail intervals. If you have not provided interval ranges, AIMMS will determine these on the basis of the collected observations. As long as the sample buffer size of the histogram has not yet been reached, you are still allowed to modify the number of intervals prior to any subsequent call to the procedure HistogramGetFrequencies.

Obtaining interval bounds

Through the procedure HistogramGetBounds you can obtain the left and right bound of each frequency interval. The bound parameters must be one-dimensional, and the cardinality of the corresponding domain set must be at least the number of intervals (including possible left- and right-tail intervals). The lower bound of a left-tail interval will be -INF, the upper bound of a right-tail interval will be INF.

Obtaining statistical information

The following functions provided statistical information:

- HistogramGetObservationCount The total number of observations,
- HistogramGetAverage the arithmetic mean,
- HistogramGetDeviation standard deviation,
- HistogramGetSkewness skewness, and
- HistogramGetKurtosis kurtosis coefficient.

Example

In the following example, a number of observable outputs o of a mathematical program are obtained as the result of changes in a single uniformly distributed input parameter InputRate. The interval range of every histogram is set to the interval [0,100] in 10 steps, and it is assumed that the set associated with index i has at least 12 elements.

```
for (o) do
    HistogramCreate( HistogramID(o) );
    HistogramSetDomain( HistogramID(o), intervals: 10, left: 0.0, width: 10.0 );
endfor;

while ( LoopCount <= TrialSize ) do
    InputRate := Uniform(0,1);
    solve MathematicalProgram;
    for (o) do
        HistogramAddObservation( HistogramID(o), ObservableOutput(o) );
    endfor;
```

```
endwhile;



for (o) do
    HistogramGetFrequencies( HistogramID(o), Frequencies(o,i) );
    HistogramGetBounds( HistogramID(o), LeftBound(o,i), RightBound(o,i) );
    HistogramDelete( HistogramID(o) );
endfor;
```