Design of classes implementing continuous distributions

Scope

This document describes the design of classes implementing continuous distributions (and some discrete), i.e. which methods and properties they must provide, relation between properties and methods for a specific distribution and special mathematical functions required to implement those distributions.

Background information

Discrete distributions

A random variable *X* has a *discrete distribution* if there is only a finite (or infinite but countable) set of values, which it can have. In this case the probability of observation of value *x* is defined by the *probability mass function* (PMF), which is:

```
p(x) = \begin{cases}
0 < \mathtt{Pr}[X = x] < 1 ; \forall ; x \in X,\
0 ; \forall ; x \notin X
\end{cases} \newline
\sum_{x \in X} {p(x)} = 1
```

The mean of the distribution is calculated as:

$$ ☑ = \sum_{x \in X} {x \times p(x)} $$

The higher moments are calculated as

$$ \mathtt{Var}(X) = \sigma^2 = \mathtt{E}[(X - \mu)^2] =\sum_{x \in X} {(x - \mu)^2 \times p(x)} \newline ☑ = \mathtt{E}[\left(\frac{X - \mu}{\sigma} \right)^3] =\sum_{x \in X} {\left(\frac{x - \mu}{\sigma} \right)^3 \times p(x)} \newline ☑ = \mathtt{E}[\left(\frac{X - \mu}{\sigma} \right)^4] =\sum_{x \in X} {\left(\frac{x - \mu}{\sigma} \right)^4 \times p(x)} \newline $$

Usually, the *excess kurtosis* ☑ - 3$.

The *cummulative distribution function* (CDF) *F(x)* is the probability of observation of the random variable *X* at the value less than or equal to the given value *x*, i.e.

$$F(x) = \mathtt{Pr}[X \leq x] = \sum_{y \in X,;y \leq x} p(y)$$

CDF has the following properties:

```
F(x) \in (0, 1] ; \forall ; x \in X \newline
F(x) > F(y) ; \forall ; x > y, ; x \in X, ; y \in X \newline
\mathtt{Pr}[a < X \leq b] = F(b) - F(a)
```

The *median*, the *first quartile*, the *third quartile* and the generic k-th of m-quantile are formaly defined as:

```
\mathtt{Median} = x: F(x) = 0.5 \newline
Q1 = Q_4^1 = x : F(x) = 0.25 \newline
Q3 = Q_4^3 = x : F(x) = 0.75 \newline
Q_m^k = x : F(x) = \frac{k}{m} ; \forall ; 0 < k \leq m
```

Since the CDF is not a continuous function, but a *step function*, the calculation of quantiles (including median and quartiles) can be tricky. If the set of the possible value is finite and small they can be calculated by sorting all possible values of the discrete random variable *X*, calculation of the corresponding *F(x)* values sequence and using the linear interpolation between two consecutive values, for which *F(x)* is below and above the required value respectively (see DE001 document). For the large finite sets or inifinite sets of discrete values of a random variable either an *inverse cummulative distribution function* (ICDF) should be used, if known, or a numerical method like bi-section should be applied to CDF.

Continuous distributions

When a random variable *X* does not have a finite set of values, but may have any value within an interal $[a, b]$ the distribution is *continuous* and is described by the *probability denisty function* (PDF) $f(x)$ instead of set of probilities for specific values. For the many common and useful distributions the interval is the entire set of real numbers, i.e. $(-\infin, +\infin)$. Thus, for a continuous distribution there is no probability of any exact value, but the probability of the value being within a specific interval:

$$\mathtt{Pr}[a \leq X \leq b] = \int_a^b f(x)dx \Rightarrow f(x) = \lim_{a \rightarrow x, b \rightarrow x} \frac{\mathtt{Pr}[a \leq X \leq b]}{b - a}$$

The PDF has the following important properties:

```
f(x) > 0 ; \forall ; x \in \mathbb{R} \newline
\int_{- \infin}^{+ \infin} {f(x) dx} = 1
```

Hence, the *arithmetic mean* of the distribution is:

$$ ☑ = \int_{- \infin}^{+ \infin} {x f(x) dx} $$

Similarly, the higher moments are defined as:

$$ \mathtt{Var}(X) = \sigma^2 = \mathtt{E}[(X - \mu)^2] =\int_{- \infin}^{+ \infin} {(x - \mu)^2 f(x) dx} \newline ☑ = \mathtt{E}[\left(\frac{X - \mu}{\sigma} \right)^3] =\int_{- \infin}^{+ \infin} {\left(\frac{x - \mu}{\sigma} \right)^3 f(x) dx} \newline ☑ = \mathtt{E}[\left(\frac{X - \mu}{\sigma} \right)^4] =\int_{- \infin}^{+ \infin} {\left(\frac{x - \mu}{\sigma} \right)^4 f(x) dx} \newline $$

Note: in practical applications *excess kurtosis* ☑ - 3$. Furthermore, for the majority of the common parameteric distributions (like Gaussian, Binomial, Student's t-distibution, etc.) these properties have simple analitical expressions in terms of the parameters of the distribution; and they do not have to be calculated using numerical integration.

The second way to describe the distribution is the *cummulative distribution function* (CDF) $\Phi(x)$, which is the probability of observation of the random variable *X* at the value less than or equal to the given value *x*, i.e.

$$\Phi(x) = \mathtt{Pr}[X \leq x] = \int_{-\infin}^{x} f(y)dy \Rightarrow f(x) = \frac{d}{dx}\Phi(x)$$

which is just an alternative representation of the definition of PDF, since $\mathtt{Pr}[a \leq X \leq b] = \Phi(b) - \Phi(a)$. Naturally, CDF has the following properties:

$$\Phi(x) \in [0, 1] ; \forall ; x \in \mathbb{R} \newline
\Phi(- \infin) = 0 \newline
\Phi(+ \infin) = 1 \newline
\Phi(x) > \Phi(y) ; \forall ; x > y$$

The *inverse cummulative distribution function* (ICDF) $\Phi^{-1}(p)$, also known as *quantile function* (QF) $Q(p)$ is defined on the interval $[0, 1]$ as the value *x* of the random variable *X* for which the cummulative probability $\Phi(x) = \mathtt{Pr}[X \leq x] = p$. Naturally, the this function has the following properties:

$$\Phi^{-1}(p) \in \mathbb{R} ; \forall ; p \in [0, 1] \newline
\Phi^{-1}(0) = - \infin \newline
\Phi^{-1}(1) = + \infin \newline
\Phi^{-1}(p) > \Phi^{-1}(q) ; \forall ; p > q \newline
\Phi(\Phi^{-1}(p)) = p \newline
\Phi^{-1}(\Phi(x)) = x$$

This function allows calcualtion of a generic quantile as:

$$Q_{m}^{k} = Q(\frac{k}{m}) \equiv \Phi^{-1}(\frac{k}{m}) \Rightarrow \newline
\mathtt{Median} = \Phi^{-1}(0.5) \newline
\mathtt{Q1} = Q_{4}^{1} = \Phi^{-1}(0.25) \newline
\mathtt{Q3} = Q_{4}^{3} =\Phi^{-1}(0.75)$$

Gaussian distribution and Z-distribution

The generic *Gaussian distribution*, a.k.a. *normal distribution*, is characterized by its *mean* value μ and *standard deviation* σ. This distribution supports $x \in (-\infin, +\infin)$.

Its PDF ans CDF are:

$$f(x) = \frac{1}{\sigma \sqrt{2 \pi}} \times e^{- \frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2} \newline
\Phi(x) = \frac{1}{2} \left[1 + \mathtt{erf} \left(\frac{x -\mu}{\sigma \sqrt{2}} \right) \right]$$

where the *error function* erf() is defined as:

$$\mathtt{erf}(z > 0) = \frac{2}{\sqrt{\pi}} \int_0^z {e^{- t^2} dt} \newline
\mathtt{erf}(0) = 0 \newline
\mathtt{erf}(z < 0) = - \mathtt{erf}(-z)$$

Note that both the *exponent* and the *error function* are implemented in the module *math* of the Standard Python Library.

The ICDF / QF function is:

$$\Phi^{-1}(p) = \mu + \sigma \sqrt{2} \mathtt{erf}^{-1}(2p - 1)$$

The *inverse error function* is not part of the Standard Python Library, and it must be implemented among other *special functions* (see DE003 document).

The Gaussian distribution has the following statistical properties:

- Mean is μ
- Variance is σ^2
- Skewness is 0
- Excess kurtosis is 0
- Median is μ
- The first quartile Q1 is $\mu - \sigma \sqrt{2} \mathtt{erf}^{-1}(0.5)$
- The third quartile Q3 is $\mu + \sigma \sqrt{2} \mathtt{erf}^{-1}(0.5)$

Note that the value of $\sqrt{2} \mathtt{erf}^{-1}(0.5)$ can be pre-calculated in advance, thus all of these properties can be expressed solely in terms of the *parameters* of the distibution.

The Z-distribuion, a.k.a. *standard normal distribution*, is a special case of μ = 0 and σ = 1.

The Gaussian distribution is important due to two facts:

- Many naturally occuring process show Gaussian distribution or a distribution, which can be closely approximated by a Gaussian one
- *Central Limit Theorem*, with the important practical consequence of which being that the distribution of the *samples means* of independent and identically distributed random variable with finite variance *converge* to a normal distribution

The second point is the core of Z-test, which compares the obtained sample mean with the known population mean as long as the population variance is known as well.

Student's t-distribution

The *Student's t-distribution* is a family of continuous probability distributions that arise when estimating the mean of a normally distributed population in situations where the sample size is small and the population's standard deviation is unknow. It is defined by a single parameter - degrees of freedom $\nu > 0$, which can be any real positive number, although in the majority of the cases it is a natural number (positive integer). This distribution supports $x \in (-\infin, +\infin)$.

Its PDF ans CDF are:

```
f(x) = \frac{\Gamma \left( \frac{\nu + 1}{2} \right)}{\sqrt{\nu \pi} \Gamma \left( \frac{\nu}{2} \right)} \left( 1 + \frac{x^2}{\nu} \right)^{- \frac{\nu + 1}{2}} \equiv \frac{1}{\sqrt{\nu} B \left( \fra
\Phi(x) = \begin{cases}
0.5, ; \mathtt{if} ; x =0 \
1 - \frac{1}{2} I_y \left( \frac{\nu}{2}, \frac{1}{2}\right), ; \mathtt{where} ; y = \frac{\nu}{x^2 + \nu}, ; \mathtt{if} ; x > 0 \
\frac{1}{2} I_y \left( \frac{\nu}{2}, \frac{1}{2}\right), ; \mathtt{where} ; y = \frac{\nu}{x^2 + \nu}, ; \mathtt{if} ; x < 0
\end{cases}
```

Where $\Gamma(z) = \int_0^{+\infin} x^{z-1} e^{-x} dx$ is *gamma function*, which is implemented in the module *math* of the Standard Python Library. $B\left(\frac{1}{2}, \frac{\nu}{2}\right)$ is *beta function* and $I_y\left(\frac{\nu}{2}, \frac{1}{2}\right)$ is *regularized incomplete beta function* (see DE03 for details).

The ICDF / QF has simple analytical formulas only for ν=1, 2, 4:

```
\Phi^{-1}(p) = \begin{cases}
\mathtt{tan}(\pi (p - 1/2)), ; \mathtt{if} ; \nu = 1 \
2 (p - 1/2) \sqrt{\frac{2}{\alpha}}, ; \mathtt{if} ; \nu = 2 \
2 \mathtt{sign}(p - 1/2) \sqrt{q-1}, ; \mathtt{if} ; \nu = 4
\end{cases} \newline
\alpha = 4 p (1 - p) \newline
q = \frac{\mathtt{cos} \left( \frac{1}{3} \mathtt{arccos} (\sqrt{\alpha}) \right)}{\sqrt{\alpha}}
```

In other cases it must be calculated using numerical methods.

Concerning the statistical properties of the distribution

$$ \mathtt{Mean} = \begin{cases} 0, ; \forall ; \nu > 1 \ \mathtt{undefined} ; \forall ; \nu \in (0, 1] \end{cases} \newline ☑ = \begin{cases} \frac{\nu}{\nu - 2}, ; \forall ; \nu > 2 \ \infin ; \forall ; \nu \in (1, 2] \ \mathtt{undefined} ; \forall ; \nu \in (0, 1] \end{cases} \newline ☑ = \begin{cases} 0, ; \forall ; \nu > 3 \ \mathtt{undefined} ; \forall ; \nu \in (0, 3] \end{cases} \newline ☑ = \mathtt{Kurt}[X] - 3 = \begin{cases} \frac{6}{\nu - 4}, ; \forall ; \nu > 4 \ \infin ; \forall ; \nu \in (2, 4] \ \mathtt{undefined} ; \forall ; \nu \in (0, 2] \end{cases} \newline \mathtt{Median} = 0 $$

Chi-squared distribution

The *chi-square distribution* or χ^2-*distribution* with $k \in \mathbb{Z}, ; k > 0$ degrees of freedom is the distribution of a sum of the squares of k independent standard normal random variables. It is a special case of the *gamma distribution*.

This distribution supports $x \in (0, +\infin)$ if *k* < *2*, and $x \in [0, +\infin)$ otherwise.

Its PDF ans CDF are:

```
f(x) = \frac{x^{k/2 - 1} e^{-x/2}}{2^{k/2} \Gamma(k/2)} \newline
\Phi(x) = \frac{\gamma(k/2, x/2)}{\Gamma(k/2)} = P \left( k/2, x/2 \right)
```

where γ() is the *lower incomplete gamma function*, and *P*() is the *regularized lower incomplete gamma function* (see DE03 for details).

The ICDF / QF has no analitical expression and must be calculated using numerical methods from CDF. Concerning the statistical properties of the distribution:

$$ \mathtt{Mean} = k \newline ☑ = 2k \newline ☑ = \sqrt{8/k} \newline ☑ = \mathtt{Kurt}[X] - 3 = \frac{12}{k} \newline \mathtt{Median} \approx k \left(1 - \frac{2}{9k} \right)^3 $$

Note, that the expression for the median value is not precise (assymptotic), as all other quantiles it must be calculated from the CDF using numerical methods.

F-distribution

The *F-distribution* is frequently used as the null distribution of a test statistics, e.g. in F-tests. It is defined by two parameters - degrees of freedom $d_1, d_2 > 0$, which can be any positive real numbers, although in the majority of applications the degrees of freedom are natural numbers (positive integers). Basically, it describes the distribution of the ratio of two indepent random variables with *chi-square distribution*, normalized to their degrees of freedom. However, it is also related to many other distributions, i.e. it describes distribution of the random variable produced as a functional transformation of another random variable or two other indenpenent random variables.

This distribution supports $x \in (0, +\infin)$ if $d_1 = 1$, and $x \in [0, +\infin)$ otherwise.

Its PDF and CDF are:

```
f(x) = \frac{\left( \frac{d_1}{d_2} \right)^{d_1/2} x^{d_1/2 - 1} \left( 1 + \frac{d_1}{d_2} x\right)^{-(d_1 + d_2) / 2}}{\mathtt{B} \left( \frac{d_1}{2}, \frac{d_2}{2} \right)} \newline
\Phi(x) = I_{\frac{d_1 x}{d_1 x + d_2}} \left( \frac{d_1}{2}, \frac{d_2}{2} \right)
```

The ICDF / QF has no analitical expression and must be calculated using numerical methods from CDF. Concerning the statistical properties of the distribution:

$$ \mathtt{Mean} = \begin{cases} \frac{d_2}{d_2 - 2}, ; \forall ; d_2 > 2 \ \mathtt{undefined} ; \mathtt{otherwise} \end{cases} \newline ☑ = \begin{cases} \frac{2 d_2^2 (d_1 + d_2 - 2)}{d_1 (d_2 - 2)^2 (d_2 -4)}, ; \forall ; d_2 > 4 \ \mathtt{undefined} ; \mathtt{otherwise} \end{cases} \newline ☑ = \begin{cases} \frac{(2 d_1 + d_2 -2) \sqrt{8 (d_2 -4)}}{(d_2 - 6) \sqrt{d_1 (d_1 + d_2 -2)}}, ; \forall ; d_2 > 6 \ \mathtt{undefined} ; \mathtt{otherwise} \end{cases} \newline ☑ = \mathtt{Kurt}[X] - 3 = \begin{cases} 12 \frac{d_1 (5 d_2 - 22)(d_1 + d_2 -2) + (d_2 - 4) (d_2 -2)^2}{d_1 (d_2 - 6) (d_2 -8) (d_1 + d_2 - 2)}, ; \forall ; d_2 > 8 \ \mathtt{undefined} ; \mathtt{otherwise} \end{cases} \newline $$

All quantiles, including median must be calculated from CDF using numerical methods

Exponential distribution

The *exponential distribution* is the probability distribution of the time between events in a process, in which events occuur continuously and independently at a constant average rate (Poission point process), which is a particular case of the *gamma distrbution*. It is defined by a single parameter - rate $\lambda > 0$. This distribution supports $x \in [0, +\infin)$.

Its PDF, CDF and ICDF / QF are:

$$f(x) = \lambda e^{- \lambda x} \newline
\Phi(x) = 1 - e^{- \lambda x} \newline
\Phi^{-1}(p) = - \frac{\mathtt{ln}(1 - p)}{\lambda}$$

where ln() is the natural logarthm (base *e*), which is implemented in the Standard Python Library module *math*.

The exponential distribution has the following statistical properties:

- Mean is $1/\lambda$
- Variance is $1/\lambda^2$
- Skewness is 2
- Excess kurtosis is 6
- Median is $\mathtt{ln}(2)/\lambda \approx 0.6931/\lambda$
- The first quartile Q1 is $-\mathtt{ln}(0.75)/\lambda \approx 0.2877/\lambda$
- The third quartile Q3 is $-\mathtt{ln}(0.25)/\lambda \approx 1.3863/\lambda$

Gamma distribution

Gamma distribution is a two-parameter family of continuous probability distributions, defined by two parameters:

- shape $k = \alpha > 0$, and
- scale $\theta > 0$ or rate $\beta = 1/\theta > 0$

This distribution supports $x \in (0, +\infin)$. Its PDF and CDF are:

$$f(x) = \frac{\beta^\alpha x^{\alpha - 1} e^{- \beta x}}{\Gamma(\alpha)} \newline
\Phi(x) = \frac{\gamma(\alpha, \beta x)}{\Gamma(\alpha)} = P(\alpha, \beta x)$$

The ICDF / QF has no analitical expression and must be calculated using numerical methods from CDF. Concerning the statistical properties of the distribution:

- Mean is α/β
- Variance is α/β^2
- Skewness is $2/\sqrt{\alpha}$
- Excess kurtosis is $6/\alpha$

All quantiles, including quartiles and median must be calculated from CDF using numerical methods.

Erlang distribution

The *Erlang distribution* is the distribution of a sum of k independent exponential variables with same mean, e.g. the distribution of the time until kth event of a Poisson process.

This distribution is a particular case of *gamma distribution*, and it supports $x \in [0, +\infin)$. It is defined by two parameters:

- shape $k \in \mathbb{N}$, and
- rate $\lambda > 0$

Its PDF and CDF are:

$$f(x) = \frac{\lambda^k x^{k - 1} e^{- \lambda x}}{(k-1)!} \newline
\Phi(x) = \frac{\gamma(k, \lambda x)}{(k-1)!} = P(k, \lambda x)$$

The ICDF / QF has no analitical expression and must be calculated using numerical methods from CDF. Concerning the statistical properties of the distribution:

- Mean is k/λ
- Variance is k/λ^2
- Skewness is $2/\sqrt{k}$
- Excess kurtosis is $6/k$

All quantiles, including quartiles and median must be calculated from CDF using numerical methods.

Poisson distribution

The *Poisson distribution* is a discrete probability distribution describing the probability of a given number of events occuring in a fixed interval of time or space if these events occur with a known constant mean rate $\lambda > 0$ and indepenently since the previous event.

It supports $k \geq 0, ; k \in \mathbb{Z}$. Its PMF and CDF are defined as:

$$p(k) = \frac{\lambda^k e^{- \lambda}}{k!} \newline
F(k) = \frac{\Gamma(k+1, \lambda)}{k!} = Q(k+1, \lambda)$$

where Γ(x, y) is *upper incomplete gamma function* and *Q*(x, y) is *regularized upper incomplete gamma function*.

The following statistical properties are defined:

- Mean is λ
- Variance is λ
- Skewness is $1/\sqrt{\lambda}$
- Excess kurtosis is $1/\lambda$
- Median is $\approx \lfloor \lambda + 1/3 - 0.02/\lambda \rfloor$

The exact value of the median as well as quartiles and any quantile must be calculated from CDF using numerical methods.

Binomial distribution

The *binomial distribution* is a discrete probability distribution describing the probability of *k* successes in $n \geq 0, ; n \in \mathbb{Z}$ draws with replacement (each draw is independent), where the probability of a success in any draw is $0 \leq p \leq 1$.

It supports $0 \leq k \leq n, ; k \in \mathbb{Z}$. Its PMF and CDF are defined as:

p(k) = C_n^k p^k (1-p)^{n-k}\newline
F(k) = I_{1-p}(n-k, 1 + k)

where

$$C_n^k = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

is a *binomial coefficient*, which is implemented as a function in the Standard Python Library for version >= 3.8.

The following statistical properties are defined:

- Mean is *np*
- Variance is *np(1-p)*
- Skewness is $\frac{1-2p}{\sqrt{np(1-p)}}$
- Excess kurtosis is $\frac{1-6p(1-p)}{np(1-p)}$
- Median is $\lfloor np \rfloor$ or $\lceil np \rceil$

All quantiles, including quartiles must be calculated from CDF using numerical methods.

Note the degenerative cases:

- p = 0, when p(0) = 1 and p(k > 0) = 0
- p = 1, when p(k = n) = 1 and p(k < n) = 0
- n = 0, when p(0) = 1 and k = 0 is the only possible outcome

These cases should not be supported, therefore the moments are always defined.

Geometric distribution

The *geometric distribution* is either of two discrete probability distributions describing:

- the number *X* of Bernoulli trials (indendent draws with replacement and two possible outcomes) needed to get one success - supported on the $k \in \mathbb{N}$
- the number *Y = X -1* of failures before the first success - supported on the $k \geq 0, ; k \in \mathbb{Z}$

In the both cases the degenerative cases of p = 0 and p = 1 should be excluded, thus $p \in (0, 1)$ is the accepted range of the distribution`s parameter.

Its PMF and CDF are defined as:

p(k) = (1-p)^{k-1} p\newline
F(k) = 1 - (1-p)^k

where the first definition of the distribution is used. Therefore, the ICDF / QF is:

$$F^{-1}(z) = \frac{\mathtt{ln}(1-z)}{\mathtt{ln}(1-p)}$$

The following statistical properties are defined:

- Mean is $\frac{1}{p}$
- Variance is $\frac{1-p}{p^2}$
- Skewness is $\frac{2-p}{\sqrt{1-p}}$
- Excess kurtosis is $6 + \frac{p^2}{1-p}$
- Median is $\frac{\mathtt{ln}(0.5)}{\mathtt{ln}(1-p)} = \frac{-\mathtt{ln}(2)}{\mathtt{ln}(1-p)} = \frac{-1}{\mathtt{log}_2(1-p)}$
- Q1 is $\frac{\mathtt{ln}(0.75)}{\mathtt{ln}(1-p)}$
- Q3 is $\frac{\mathtt{ln}(0.25)}{\mathtt{ln}(1-p)} = \frac{-\mathtt{ln}(4)}{\mathtt{ln}(1-p)} = \frac{-2*\mathtt{ln}(2)}{\mathtt{ln}(1-p)} = \frac{-2}{\mathtt{log}_2(1-p)}$

Hypergeometric distribution

The *hypergeometric distribution* is a discrete probability distribution describing the probability of *k* successes in *n* draws without replacement from a finite population of size *N* containing exactly *K* required objects, i.e. a success is the drawing of a required object.

Basically, it is defined by three parameters:

- Population size $N \geq 0, ; N \in \mathbb{Z}$
- Number of 'success' objects $0 \leq K \leq N, ; K \in \mathbb{Z}$
- Number of draws $0 \leq n \leq N, ; n \in \mathbb{Z}$

It supports $\mathtt{max}(0, n + K - N) \leq k \leq \mathtt{min}(n, K), ; k \in \mathbb{Z}$. Its PMF and CDF are defined as:

```
p(k) = \frac{C_K^k C_{N-K}^{n-k}}{C_N^n}\newline
F(k) = 1 - \frac{C_n^{k+1} C_{N-n}^{K-k-1}}{C_N^K} \times {}_3 F_2 \left[ 1, k+1-K, k+1-n; k+2, N+K+2-K-n; 1 \right]
```

where

$$ {}*p F_q[a_1, ..., a_p; b_1, ..., b_q;z] = \sum*{n=0}^{\infin} {\frac{\prod_{k=0}^{n - 1} {(a_1 + k)} \times ... \times \prod_{k=0}^{n - 1} {(a_p + k)}}{\prod_{k=0}^{n - 1} {(b_1 + k)} \times ... \times \prod_{k=0}^{n - 1} {(b_q + k)}} \frac{z^m}{n!}} $$

is the generalized *hypergeometric function*, which is not implemented in the Standard Python Library. Since the hypergeometric distrion is not only discrete but also finite, it is easier to calculate CDF by direct summation of PDF, instead of implementating finite polynomial approximation of infinite series.

The following statistical properties are defined:

- Mean is $n\frac{K}{N}$
- Variance is $n\frac{K}{N}\frac{N-K}{N}\frac{N-n}{N-1}$
- Skewness is $\frac{(N-2K)(N-1)^{1/2}(N-2n)}{\lfloor nK(N-K)(N-n)\rfloor^{1/2}(N-2)}$
- Excess kurtosis is $\frac{(N-1)N^2(N(N+1)-6K(N-K)-6n(N-n))+6nK(N-K)(N-n)(5N-6)}{nK(N-K)(N-n)(N-2)(N-3)}$

All quantiles, including quartiles and median must be calculated from CDF using numerical methods.

In the formulas above there are many possible situations causing *division by zero* error: K = 0, K = N, n = 0, n = N, N = 0, N = 1, N = 2 and N = 3. However, such situations are easily avoidable:

- First of all, for the number of draws n = 0 the only possible number of successes is 0 regardless of the N and K values
- Secondly, for the number of draws n = N there are always K successes (full set is drawn)
- For K = 0 the number of successes is always 0 regardless of the number of draws n
- Finally, for K = N, the number of successes is always n

Thus, these trivial cases can be ignored by modifying the parameters requierements to $1 \leq K; , ; n < N; \Rightarrow ; N \geq 2$, which leaves only 5 possible problematic situations, when the requirement moments can be easily calculated.

Case 1: N = 2, K = n = 1

- ☑ = 0 * \frac{1}{2} + 1 * \frac{1}{2} = \frac{1}{2}$
- $Var(X) = E[(X - \mu)^2] = \left(-\frac{1}{2}\right)^2 * \frac{1}{2} + \left(\frac{1}{2}\right)^2 * \frac{1}{2} = \frac{1}{4}$, hence $\sigma = \frac{1}{2}$
- $Skew = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = (-1)^3 * \frac{1}{2} + (1)^3 * \frac{1}{2} = 0$
- $KurtE\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = (-1)^4 * \frac{1}{2} + (1)^4 * \frac{1}{2} = 1$, hence *ExKurt = Kurt - 3 = - 2*

Case 2: N = 3, K = 1, n = 1

- ☑ = 0 * \frac{2}{3} + 1 * \frac{1}{3} = \frac{1}{3}$
- $Var(X) = E[(X - \mu)^2] = \left(-\frac{1}{3}\right)^2 * \frac{2}{3} + \left(\frac{2}{3}\right)^2 * \frac{1}{3} = \frac{6}{27} = \frac{2}{9}$, hence $\sigma = \frac{\sqrt{2}}{3}$
- $Skew = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = (-\frac{1}{\sqrt{2}})^3 * \frac{2}{3} + (\frac{2}{\sqrt{2}})^3 * \frac{1}{3} = \frac{1}{\sqrt{2}}$
- $KurtE\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = (-\frac{1}{\sqrt{2}})^4 * \frac{2}{3} + (\frac{2}{\sqrt{2}})^4 * \frac{1}{3} = \frac{2+16}{12} = \frac{3}{2} = 1.5$, hence *ExKurt = Kurt - 3 = - 1.5*

Case 3: N = 3, K = 1, n = 2

- ☑ = 0 * \frac{1}{3} + 1 * \frac{2}{3} = \frac{2}{3}$
- $Var(X) = E[(X - \mu)^2] = \left(-\frac{2}{3}\right)^2 * \frac{1}{3} + \left(\frac{1}{3}\right)^2 * \frac{2}{3} = \frac{6}{27} = \frac{2}{9}$, hence $\sigma = \frac{\sqrt{2}}{3}$
- $Skew = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = (-\frac{2}{\sqrt{2}})^3 * \frac{1}{3} + (\frac{1}{\sqrt{2}})^3 * \frac{2}{3} = -\frac{1}{\sqrt{2}}$
- $KurtE\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = (-\frac{2}{\sqrt{2}})^4 * \frac{1}{3} + (\frac{1}{\sqrt{2}})^4 * \frac{2}{3} = \frac{16+2}{12} = \frac{3}{2} = 1.5$, hence *ExKurt = Kurt - 3 = - 1.5*

Case 4: N = 3, K = 2, n = 1

- ☑ = 0 * \frac{1}{3} + 1 * \frac{2}{3} = \frac{2}{3}$
- $Var(X) = E[(X - \mu)^2] = \left(-\frac{2}{3}\right)^2 * \frac{1}{3} + \left(\frac{1}{3}\right)^2 * \frac{2}{3} = \frac{6}{27} = \frac{2}{9}$, hence $\sigma = \frac{\sqrt{2}}{3}$
- $Skew = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = (-\frac{2}{\sqrt{2}})^3 * \frac{1}{3} + (\frac{1}{\sqrt{2}})^3 * \frac{2}{3} = -\frac{1}{\sqrt{2}}$

- $KurtE\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = (-\frac{2}{\sqrt{2}})^4 * \frac{1}{3} + (\frac{1}{\sqrt{2}})^4 * \frac{2}{3} = \frac{16+2}{12} = \frac{3}{2} = 1.5$, hence *ExKurt = Kurt - 3 = - 1.5*

Case 5: N = 3, K = 2, n = 2

- ☑ = 1 * \frac{2}{3} + 2 * \frac{1}{3} = \frac{4}{3}$
- $Var(X) = E[(X-\mu)^2] = \left(-\frac{1}{3}\right)^2 * \frac{2}{3} + \left(\frac{2}{3}\right)^2 * \frac{1}{3} = \frac{6}{27} = \frac{2}{9}$, hence $\sigma = \frac{\sqrt{2}}{3}$
- $Skew = E\left[\left(\frac{X-\mu}{\sigma}\right)^3\right] = (-\frac{1}{\sqrt{2}})^3 * \frac{2}{3} + (\frac{2}{\sqrt{2}})^3 * \frac{1}{3} = \frac{1}{\sqrt{2}}$
- $KurtE\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] = (-\frac{1}{\sqrt{2}})^4 * \frac{2}{3} + (\frac{2}{\sqrt{2}})^4 * \frac{1}{3} = \frac{2+16}{12} = \frac{3}{2} = 1.5$, hence *ExKurt = Kurt - 3 = - 1.5*

To summarize:

- N = 2: Var = 0.25, σ = 0.5, Skew = 0, ExKurt = -2
- N = 2: Var = 2/9, $\sigma = \sqrt{2}/3$, ExKurt = -1.5, and
  - Skew = 1 / $\sqrt{2}$ for K = n = 1 , 2
  - Skew = - 1 / $\sqrt{2}$ for $1 \leq K \neq n \leq 2$

General design patterns

Both types of the random distributions: discrete and continuous - can be implemented as classes with the identical API, compatible with the 1D and 2D statistics classes, with the following conventions:

- The parameters of a distribution are passes as arguments of the initialization method, i.e. during the class instantiation
- The same parameters can be accessed (read-out) and modified (write access) at any time via **getter + setter properties** - with an exception of Z-distribution, which has fixed parameters, coinciding with the statistical properties, therefore - they must be accessible only for reading-out, but not for modification
- An instance of such class provides the following **read-only properties** to access the respective statistical properties of the distribution - with exception for the generic Gaussian distribution, in which case *Mean* and *Sigma* are also parameters of the distribution, and they must be **getter + setter properties**:
  - Mean (arithmetic mean)
  - Var (variance)
  - Sigma (standard deviation)
  - Skew (skewness)
  - Kurt (excess kurtosis)
  - Median (median value of the distribution)
  - Q1 (the first quartile of the distribution)
  - Q3 (the third quartile of the distribution)
  - Min
  - Max
- If these properties have fixed values or can be easily calculated from the parameters of the distribution, such calculations can be performed each time; however, if intensive calculations are requried - the respective values should be cached and re-used
- The positive infinity value (for Max) should be represented as *math.inf* constant, the negative infinity value (for Min) - as *-math.inf*, and the open zero interval (for Min) as *2 * sys.float_info.min*
- All classes should provide the following methods:
  - pdf(Value: int OR float), which must return the value of:
    - PDF for a continuous distribution with the following special cases
      - with accepted range Value > 0 the convention is pdf(Value $\leq$ 0) = 0
      - for the accepted range Value $\geq$ 0 the convention is pdf(Value < 0) = 0
    - PMF for a discrete distribution if the value is integer and within the accepted range, otherwise - strict zero
  - cdf(Value: int or float), which must return CDF, which is
    - a continuous function for the continuous distributions with the return values in
      - the open interval (0, 1) for a distribution w/o minimum accepted value
      - the semi-open interval [0, 1) for a distribution with minimum accepted value, where cdf(Value $\leq$ Min) = 0
    - a step-function, which is 0 for Value < Min, and is constant cdf($\lfloor Value \geq Min \rfloor$) in the semi-open interval $[\lfloor Value \rfloor, \lfloor Value \rfloor + 1)$, with the return values belonging to
      - the semi-open interval [0, 1) for the infinite discrete distribution
      - the closed interval [0, 1] for the finite distribution, where cdf(Value $\geq$ Max) = 1
  - qf(Probability: 0 < float < 1), which must return ICDF / QF; for the discrete distributions the return value is, in principle, a floating point number Value, such that $cdf(\lfloor Value \rfloor) \leq Probability < cdf(\lfloor Value \rfloor + 1)$
  - getQuantile(k: int > 0, m: int > 0), where k < m is a *short-hand* for qf(k/m)
  - getHistogram(min: int OR float, max: int OR float, NBins: int > 1), where max > min; the calculations should be performed in the following manner
    - bin size S is defined as (max - min) / (NBins - 1)
    - for the k-th bin (indexing from 0 to NBins - 1)
      - the left boundary is Left = min + (k - 0.5) * S
      - the right boundary is Right = min + (k + 0.5) * S
      - the central value is Center = min + k * S
      - the binned frequency is F = cdf(Right) - cdf(Left)
    - the return value is a tuple of length NBins with each element being a 2-tuple of (Center, F) pairs
  - random(), which returns:
    - a floating point number R = qf(r), where r is uniformly distributed random value in the range (0,1) in the case of continuous distributions
    - an integer number $R = \lceil qf(r) \rceil$, where r is uniformly distributed random value in the range (0,1) in the case of discrete distributions

Modified bi-section method for calculating ICDF / QF

The function f(x) is monotonically increasing if for a < x < b the function's values are related as f(a) < f(x) < f(b), in which case the solution of the equation f(x) = y can be found numerically using the following iterative procedure:

- Find the *intial guesses* a and b such that f(a) < y < f(b)
- Calculate f((a+b)/2). If it is
    - greater than y, then adjust the upper boundary of the range, b -> (a+b) / 2
    - less than y, then adjust the lower boundary of the range, a -> (a+b) / 2
    - equal to y - the solution is found, which is (a+b) / 2
- Iterately apply the following step, untill the required precision is achieved, i.e. $b - a \leq \delta$ or $|f(\frac{a+b}{2}) - y| \leq \varepsilon$

In the case of *discrete distributions* the boundaries *a* and *b* are **integer numbers**, because CDF is step-wise increasing function is this case, thus the new (lower or upper) boundary is defined as $\lfloor\frac{a+b}{2}\rfloor$, and the iterative narrowing of the range is terminated as soon as a + 1 = b. Furthermore, if at any step $|f(a) - y| \leq \varepsilon$ or $|f(b) - y| \leq \varepsilon$ the corresponding boundary value is returned as the result. Otherwise, as soon as a - b = q situation is reached the return value x is calculated as

$$x = a + \frac{y - f(a)}{f(a+1) - f(a)}$$

Naturally, f(x) is the CDF, where y is the specified cummulative propability *p*.

Furthermore, for the discrete distributions $x_1, x_2, \dots$ (infinite) or $x_1, x_2, \dots, x_N$ (finite) an additional check must be performed, if $cdf(x_1) > p$, in which case $a = x_1 - 1$ and $b = a + 1 = x_1$ are immediately the terminal bounds.

Then, the following procedure should be used for the selection of the *initial guesses*, since one or both boundaries of the values supported by the distribution can be infinite:

- Select an initial point x_0 between *Max* and *Min* boundaries of the values supported by the distibution:
    - If both *Max* and *Min* are finite - select 0.5 * (Max + Min)
    - Else:
        - If *Mean* property is defined for the distribution - select it
        - Otherwise:
            - If *Min* is finite - select $Min + 3 * \sigma$
            - Else if *Max* is finite - select $Max - 3 * \sigma$
            - Otherwise (both bounds are infinite) - select 0 (zero value)
- Note that if the standard deviation is not defined for the distribution, or it is infinite, use $\sigma = 1$ value instead, also if actual standard deviation < 1 for a discrete distribution
- Calculate $cdf(x_0)$ for continuous distribution or $cdf(\lfloor x_0 \rfloor)$ for discrete and compare it with the passed *p* value:
    - $\approx$ p - the solution is found
    - < p - shift to the left (towards *Min*) using the following algorithm:
        - set upper boundary $Right = x_0$
        - select new lower boundary $Left = x_0$ and shift it as:
            - if $Min > -\infin$ set $Left \rightarrow (Min + Left)/2$
            - else:
                - if $Left \leq -\sigma$ set $Left \rightarrow 2 * Left$
                - else set $Left \rightarrow Left - \sigma$
        - compare $cdf(Left)$ with *p*
            - if $cdf(Left) \approx p$ - *Left* is the solution
            - else if $cdf(Left) < p$ - the boundaries *Left* and *Right* are found
            - else set $Right \rightarrow Left$ and keep on shifting *Left* until a solution or the proper frame boundaries are found
    - > p - shift to the right (towards *Max*) using using the following algorithm:
        - set lower boundary $Left = x_0$
        - select new lower boundary $Right = x_0$ and shift it as:
            - if $Max < \infin$ set $Right \rightarrow (Max + Right)/2$
            - else:
                - if $Right \geq \sigma$ set $Right \rightarrow 2 * Left$
                - else set $Right \rightarrow Right + \sigma$
        - compare $cdf(Right)$ with *p*
            - if $cdf(Right) \approx p$ - *Right* is the solution
            - else if $cdf(Right) > p$ - the boundaries *Left* and *Right* are found
            - else set $Left \rightarrow Right$ and keep on shifting *Right* until a solution or the proper frame boundaries are found
- In the case of a discrete distribution the *Left* boundary is *floored* to an integer, and *Right* boundary is *ceiled* to an integer

The described modifications to the bisection method takes extend the method for the cases of:

- step-wise growing functions
- functions being defined not on the entire real numbers set $\mathbb{R}$, i.e. $(-\infin, +\infin)$ but also:
    - intervals with finite lower bound $[a, +\infin)$
    - intervals with finite upper bound $(-\infin, b]$
    - bound (closed) intevals $[a, b]$