

S4 Classes for Distributions—a manual for packages `"distr"`, `"distrEx"`, `"distrSim"`, `"distrTEst"`, version 1.9

Peter Ruckdeschel*

Matthias Kohl†

Thomas Stabla‡

Florian Camphausen§

Mathematisches Institut

Universität Bayreuth

D-95440 Bayreuth

Germany

e-Mail: `peter.ruckdeschel@uni-bayreuth.de`

July 27, 2007

Abstract

`"distr"` is a package for R from version 1.8.1 onwards that is distributed under GPL license 2.0. Its own current version is 1.9. The aim of this package is to provide a conceptual treatment of random variables (r.v.'s) by means of S4-classes. A mother class `Distribution` is introduced with slots for a parameter and for functions `r`, `d`, `p`, and `q` for simulation, respectively for evaluation of density / c.d.f. and quantile function of the corresponding distribution. All distributions of the `"stats"` package are implemented as subclasses of either `AbscontDistribution` or `DiscreteDistribution`, which themselves are again subclasses of `UnivariateDistribution`. By means of these classes, we may automatically generate new objects of these classes for the laws of r.v.'s under standard mathematical univariate transformations and under convolution of independent r.v.'s. From version 1.6 on, `"distr"` has been split up into the smaller packages `"distr"` (only distribution-classes and -methods), `"distrSim"` (standardized treatment of simulations, also under contaminations) and `"distrTEst"` (classes and methods for evaluations of statistical procedures on such simulations).

The latter two of them require package `"setRNG"` by Paul Gilbert to be installed from CRAN.

Additionally, mainly contributed by [4], in `"distrEx"` we extend the functionality of `"distr"`, providing functionals like expectation or variance and distances for distributions. Also, this package contains some first steps to multivariate distributions,

*Universität Bayreuth

†SIRS-Lab GmbH, Jena

‡Hans-Sachs-Gymnasium Nürnberg

§West-LB, Düsseldorf

providing classes for discrete multivariate distributions and for factorized, conditional distributions.

Contents

0	Motivation	4
1	Concept	6
2	Organization in classes	7
2.1	Distribution classes	7
2.1.1	Subclasses	8
2.1.2	Classes for multivariate distributions and for conditional distributions	9
2.1.3	Parameter classes	11
2.2	Simulation classes	11
2.3	Evaluation class	14
2.4	EvaluationList class	15
3	Methods	15
3.1	Affine linear transformations	16
3.2	The group math of unary mathematical operations	16
3.3	Construction of d , p , and q from r	17
3.4	Convolution	17
3.5	Overloaded generic functions	17
3.6	liesInSupport	20
3.7	Simulation (in package distrSim)	20
3.8	Evaluate (in package distrTEst)	26
3.9	Is-Relations	26
3.10	Further methods	27
3.11	Functionals (in package distrEx)	27
3.11.1	Expectation	27
3.11.2	Variance	29
3.11.3	Further functionals	30
3.12	Truncated moments (in package distrEx)	30
3.13	Distances (in package distrEx)	31
3.14	Functions for demos (in package distrEx)	31
3.14.1	CLT for arbitrary summand distribution	31
3.14.2	LLN for arbitrary summand distribution	31
3.14.3	Deconvolution example	32

4	Options	32
4.1	Options for distr	32
4.2	Options for distrEx	33
4.3	Options for distrSim	34
4.4	Options for distrTEst	34
5	Startup Messages	34
6	System/version requirements	35
6.1	System requirements	35
6.2	Required version of R	35
6.3	Dependencies	35
6.4	License	36
7	Details to the implementation	36
8	A general utility	37
9	Odds and Ends	37
9.1	What should be done and what we could do	37
9.2	What should be done but for which we lack the know-how	37
10	Acknowledgement	38
11	Examples	38
11.1	12-fold convolution of uniform (0, 1) variables	38
11.2	Comparison of exact convolution to FFT for normal distributions	40
11.3	Comparison of FFT to RtoDPQ	43
11.4	Comparison of exact and approximate stationary regressor distribution . . .	46
11.5	Truncation and Huberization/winsorization	49
11.6	Distribution of minimum and maximum of two independent random variables	56
11.7	Instructive destructive example	61
11.8	A simulation example	62
11.9	Expectation of a given function under a given distribution	68
11.10	n -fold convolution of absolutely continuous distributions	69

This document appeared in an abridged form in *R-News*, **6**(2) as “S4 Classes for Distributions”, c.f. [8], which in its published form refers to package versions 1.6, resp. 0.4-2. This document takes into account the subsequent revisions and versions.

0 Motivation

R up to now contains powerful techniques for virtually any useful distribution using the suggestive naming convention `[prefix]<name>` as functions where `[prefix]` stands for `r`, `d`, `p`, or `q` and `<name>` is the name of the distribution.

There are limitations of this concept, however: You can only use distributions which are implemented in some library already or for which you yourself have provided an implementation. In many natural settings you want to formulate algorithms once for all distributions, so you should be able to treat the actual distribution `<name>` as sort of a variable.

You may of course paste together prefix and the value of `<name>` as a string and then use `eval(parse(...))`. This is neither very elegant nor flexible, however.

Instead, we would rather like to implement the algorithm by passing an object of some distribution class as argument to the function. Even better though, we would use a generic function and let the S4-dispatching mechanism decide what to do at run-time. In particular, we would like to automatically generate the corresponding functions `r`, `d`, `p`, and `q` for the law of expressions like $X+3Y$ for objects X and Y of class `Distribution`, or, more general, of a transformation of X, Y under a function $f: \mathbb{R}^2 \rightarrow \mathbb{R}$ which is already realized as a function in R.

This is possible with package "distr". As an example, try

```
> library(distr)
> N <- Norm(mean = 2, sd = 1.3)
> P <- Pois(lambda = 1.2)
> Z <- 2*N + 3 + P
> Z
```

Distribution Object of Class: AbscontDistribution

```
> plot(Z, withSweave = TRUE)
> p(Z)(0.4)
```

```
[1] 0.002404514
```

```
> q(Z)(0.3)
```

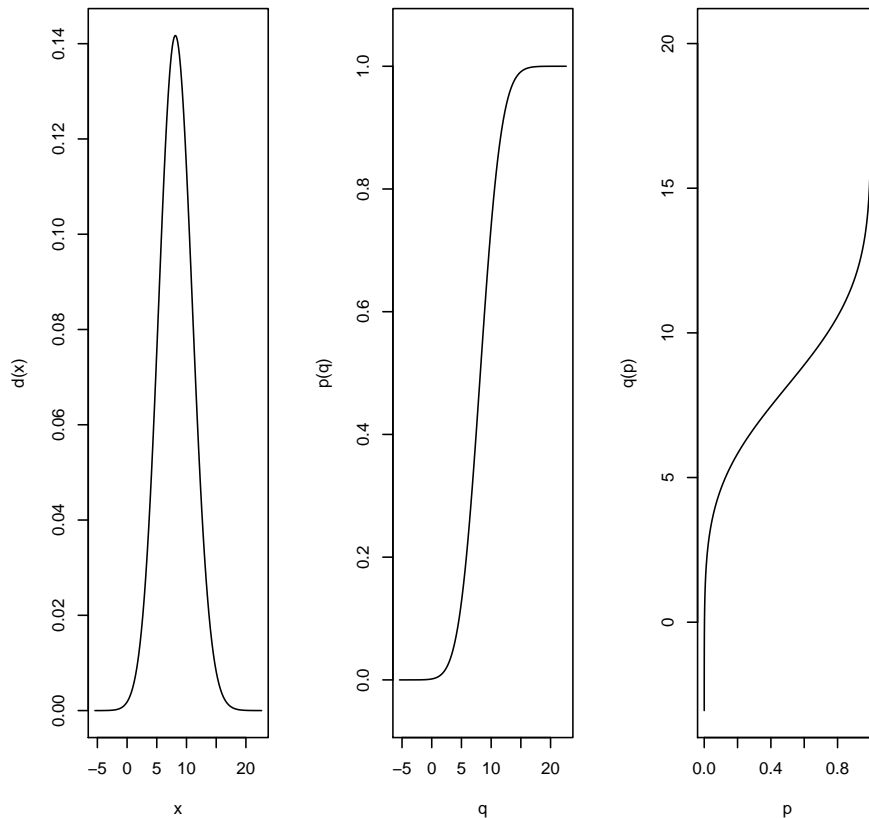
```
[1] 6.709
```

```
> Zs <- r(Z)(50)
> Zs
```

```
[1] 6.241509 4.255878 10.157096 6.552644 10.408005 9.139607 9.492226
[8] 4.719508 9.590499 11.083819 7.460663 9.521220 3.986724 10.073012
[15] 6.074737 8.801221 8.959170 4.500947 6.326799 10.237874 11.903103
```

[22]	11.580105	6.946968	4.408885	9.964763	5.276061	15.118674	10.536000
[29]	4.825118	10.905699	8.768601	7.477989	3.450973	13.651364	5.504031
[36]	10.865972	10.360069	7.676843	9.748552	9.973526	7.946040	8.641474
[43]	9.838526	9.240986	9.799410	7.288146	10.473106	5.972859	13.561151
[50]	8.286181						

Density of AbscontDistributic CDF of AbscontDistributionntile function of AbscontDistri



Comment:

Let N an object of class "Norm" with parameters `mean=2`, `sd=1.3` and let P an object of class "Pois" with parameter `lambda=1.2`. Assigning to Z the expression `2*N+3+P`, a new distribution object is generated —of class "AbscontDistribution" in our case— so that identifying N, P, Z with random variables distributed according to N, P, Z , $\mathcal{L}(Z) = \mathcal{L}(2 * N + 3 + P)$, and writing `p(Z)(0.4)` we get $P(Z \leq 0.4)$, `q(Z)(0.3)` the 30%-quantile of Z , and with `r(Z)(50)` we generate 50 pseudo random numbers distributed according to Z , while the `plot` command generates the above figure.

1 Concept

In developing our packages, we had the following principles in mind: We wanted to be open in our design so that our classes could easily be extended by any volunteer in the R community to provide more complex classes of distributions as multivariate distributions, times series distributions, conditional distributions. As an exercise, the reader is encouraged to implement extrem value distributions from the package "evd"¹. The largest effort will in fact be the documentation...

We also wanted to preserve naming and notation from R-"stats" as far as possible so that any programmer used to S could quickly use our package. Even more so, as the distributions already implemented to R are all well tested and programmed with skills we lack, we use the existing `r`, `d`, `p`, and `q`-functions wherever possible, only wrapping them by small code snippets to our class hierarchy.

Third we wanted to use a suggestive notation for our automatically generated methods `r`, `d`, `p`, and `q`, which we think is now largely achieved. All this should make intensive use of object orientation in order to be able to use inheritance and method overloading. Let us briefly explain why we decided to realize `r`, `d`, `p`, and `q` as part of our class definitions: Doing so, we place ourselves somewhere between pure object orientation where methods would be *slots* —in the language of the S4-concept, confer [2]— and the S4 paradigm where methods “live their own life” apart from the classes, or, to `q`, which should be regarded use [1]’s terminology, we use COOP²-style for `r`, `d`, `p`, and `q` methods, and FOOP³-style for “normal” methods.

The S4-paradigm with methods which are not attached to an object but rather behave differently according to the classes of their arguments is fine if there are particular user-written methods for only some few general distribution classes like `AbscontDistribution`, as in the case for `plot` or `+` (c.f. [5], Section 2.2). During a typical R session with "distr", however, there will be a lot of, mostly automatically generated objects of our distribution classes, each with its own `r`, `d`, `p`, and `q`; this even applies to intermediate expressions like `2*N`, `2*N+3` to eventually produce `Z` in the example in the motivation. Treating `r`, `d`, `p`, and `q` as generic functions, we would need to generate new classes for each expression `2*N`, `2*N+3`, `Z` and, correspondingly, particular S4-methods for `r`, `d`, `p`, and `q` for each of these new classes; apparently, this would produce overly many classes for an effective inheritance structure.

In providing arithmetics for distributions, we have to deviate a little from the paradigm of S as a functional language: For operators like `+`, additional parameters controlling the precision of the results cannot be handily passed as arguments. For this purpose we provide

¹a solution to this “homework” may be found in the sources to "distrEx"

²class-object-orientated programming, as e.g. in C++

³function-object-orientated programming, as in the S4-concept

global options which may be inspected and modified by `distroptions`, `getdistrOption`⁴ in complete analogy to `options`, `getOption`. Finally our concept as to parameters: Contrary to the standard R-functions like `rnorm` we only permit length 1 for parameters like `mean`, because we see the objects as implementations of univariate random variables, for which vector-valued parameters make no sense; rather one could gather several objects with possibly different parameters to a vector/list of distributions. Of course, the original functions `rnorm` etc. remain unchanged and still allow for vector-valued parameters. Kouros Owzar in an off-list mail raised the point, that in case of multiple parameters as in case of the normal or the Γ -distribution, it might be useful to be able to pass these multiple parameters in vectorized form to the generating function. We, too, think that this is a good idea, but even more plan to introduce a further extension package to "`distr`" which will cover statistical models. In this package, this issue will be solved by requiring a map $\theta \mapsto P_\theta$ or, in `S`, a function `function(theta)...` which returns an object of class distribution or subclass, which realizes P_θ . So it will be up to the programmer or user how to realize this map.

2 Organization in classes

Loosely speaking we have three large groups of classes: distribution classes (in "`distr`"), simulation classes (in "`distrSim`") and an evaluation class (in "`distrTEst`"), where the latter two are to be considered only as tools which allow a unified treatment of simulations and evaluation of statistical estimation (perhaps also tests and predictions later) under varying simulation situations. Additionally, package "`distrEx`" provides classes for discrete multivariate distributions and for factorized, conditional distributions, as well as a bundle of functionals and distances (see below).

2.1 Distribution classes

The purpose of the classes derived from the class `Distribution` is to implement the concept of a r.v./distribution as such in R.

All classes derived from `Distribution` have a slot `param` for a parameter, a slot `img` for the range and the constitutive slots `r`, `d`, `p`, and `q`.

From version 1.9 on, up to arguments referring to a parameter of the distribution (like `mean` for the normal distribution), these function slots have the same arguments as those of package "`stats`", i.e.; for a distribution object `X` we may call these functions as

- `r(X)(n)` —except for objects of class `Hyper`, where there is a slot `n` already, so here the argument name to `r` is `nn`.

⁴Upto version 0.4-4, we used a different mechanism to inspect/modify global options of "`distrEx`" (see section 4.2); corresponding functions `distrExoptions`, `getdistrExOption` for package "`distrEx`" are available from version 1.9 on.

- `d(X)(x, log = FALSE)`
- `p(X)(q, lower.tail = TRUE, log.p = FALSE)`
- `q(X)(p, lower.tail = TRUE, log.p = FALSE)`

For the arguments of these function slots see e.g. `rnorm` from package `"stats"`. Note that, as usual, slots `d`, `p`, and `q` are vectorized in their first argument, but are not on the subsequent ones. The idea is to gain higher precision for the upper tails or when multiplying probabilities.

2.1.1 Subclasses

To begin with, we have considered univariate distributions giving the S4-class `UnivariateDistribution`, and as typical subclasses, we have introduced classes for absolutely continuous and discrete distributions — `AbscontDistribution` and `DiscreteDistribution`.

The former, from version 1.9 on, has a slot `gaps` of class `OptionalMatrix`, i.e.; an object which may either be `NULL` or a `matrix`. This slot, if non-`NULL`, contains left and right endpoints of intervals where the density of the object is 0. This slot may be inspected by the accessor `gaps()` and modified by a corresponding replacement method. It may also be filled automatically by `setgaps(object, exactq = 6, ngrid = 50000)`, where upon evaluation of the `d`-slot on a grid of length `ngrid`, all regions in the range⁵ of the distribution where the density is smaller than $10^{-\text{exactq}}$ are set to `gaps`.

For saved objects from earlier versions, we provide the functions `isOldVersion` and `conv2NewVersion` to check whether the object was generated by an older version of this package and to convert such an object to the new format, respectively.

Class `DiscreteDistribution` has a slot `support`, a vector containing the support of the distribution, which is truncated to the lower/upper `TruncQuantile` in case of an infinite support. `TruncQuantile` is a global option of `"distr"` described in section 4. From version 1.9 on, there are methods `p.l` and `q.r` for the left-continuous variant of the cdf, i.e.; $t \mapsto p.l(t) = P(X < t)$, and the right-continuous variant of the quantile function, i.e.;

$$s \mapsto q.r(s) = \sup\{t \mid P(\text{object} \leq t) \leq s\}$$

Also from version 1.9 on, class `DiscreteDistribution` has a subclass `LatticeDistribution` for supports consisting of⁶ an affine linear lattice of form $p + iw$ for $p \in \mathbb{R}$, $w \in \mathbb{R}$,

⁵more precisely: between lower and upper `TruncQuantile`; `TruncQuantile` is a global option of `"distr"` described in section 4

⁶or at least if filled with points carrying no mass have a representation as an affine linear lattice

$w \neq 0$ and $i = 0, 1, \dots, L$, $L \in \mathbb{N} \cup \infty$. This class gains a slot `lattice` of class `Lattice` (see below). The purpose of this class is mainly its use in DFT/FFT methods for convolution. Slot `lattice` may be inspected by the usual accessor function `lattice()`. As by inheritance, all subclasses of `LatticeDistribution` which prior to version 1.9 were direct subclasses of `DiscreteDistribution` gain a slot `lattice`, too, we provide again `isOldVersion` and `conv2NewVersion` methods to check whether the object was generated by an older version of this package and to convert such an object to the new format, respectively. Also note that internally, we suppress lattice points from the support where the probability is 0.

Objects of classes `LatticeDistribution` resp. `DiscreteDistribution` may be generated using the generating functions `LatticeDistribution()` resp. `DiscreteDistribution()`; see also the corresponding help.

As subclasses of these absolutely continuous and discrete classes, we have implemented all parametric families which already exist in the "`stats`" package of R in form of `[prefix]<name>` functions —by just providing wrappers to the original R-functions.

Schematically, the inheritance relations as well as the slots of the corresponding classes may be read off from figure 1. Class `LatticeDistribution` and slot `gaps`, as well as additional classes `AffLinAbscontDistribution`, `AffLinDiscreteDistribution`, `AffLinLatticeDistribution` (c.f. section 3.1) are still lacking in this graphic so far, however.

The most powerful use of our package probably consists in operations to automatically generate new slots `r`, `d`, `p`, and `q` —induced by mathematical transformations. This is discussed in some detail in subsection 3.

2.1.2 Classes for multivariate distributions and for conditional distributions

In "`distrEx`", we provide the following classes for handling multivariate distributions:

Lists of distributions As a first step, we allow distributions to be gathered in lists, giving classes `DistrList` and `UnivarDistrList`, where in case of the latter, all elements must be univariate distributions. For these, the usual indexing operations with `[[.]]` are available. As we will use these lists to construct more general mixture distributions in some subsequent versions, we have moved these routines to package "`distr`" from version 1.9 on.

Multivariate distribution classes Multivariate distributions are much more complicated than univariate ones, which is why but a few exceptional ones have already been implemented to R in packages like "`multnorm`". In particular it is not so clear what a slot `q` should mean and, in higher dimensions slot `p`, and possibly also slot `d` may become

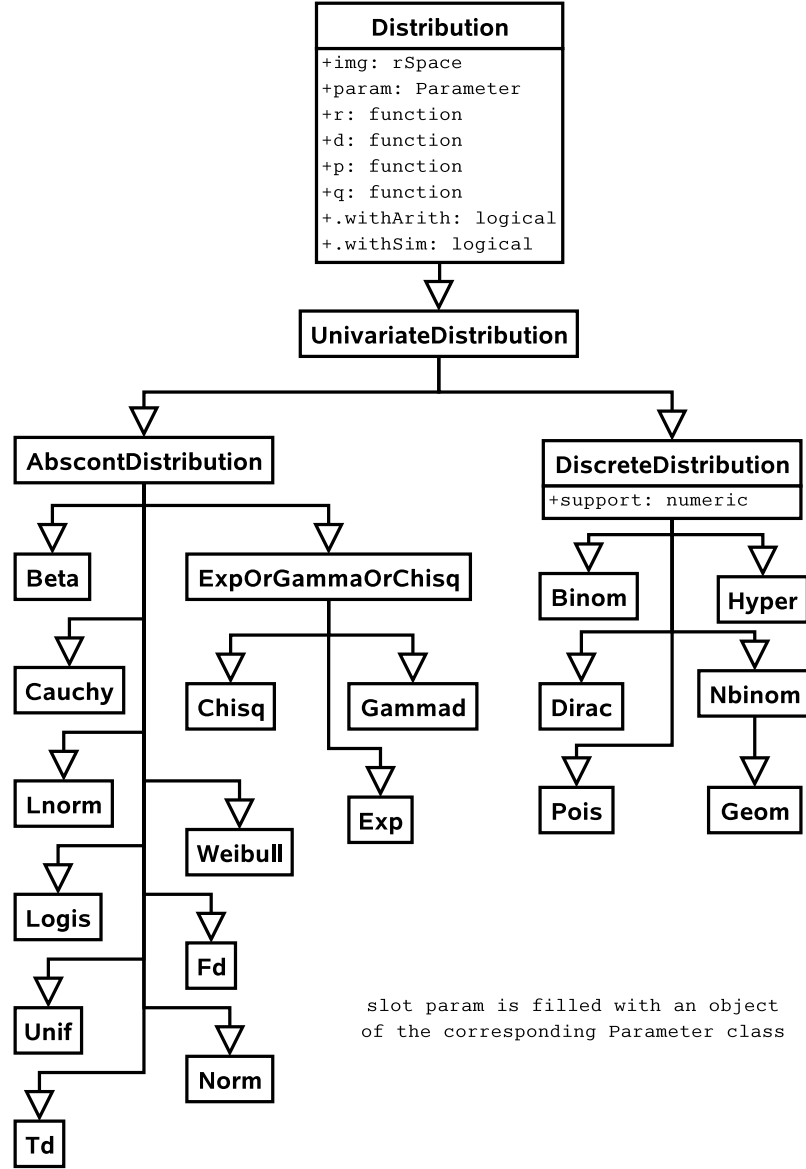


Figure 1: Inheritance relations and slots of the corresponding (sub-)classes for **Distribution** where we do not repeat inherited slots

awkward. So, for multivariate distributions, realized as class **MultivariateDistribution**, we only insist on slot **r**, while the other functional slots may be left void.

The easiest case is the case of a discrete multivariate distribution with finite support which is implemented as class **DiscreteMVDistribution**.

Conditional distribution classes Also arising in multivariate settings only are conditional distributions. In our approach, we realize factorized, conditional distributions where the (factorized) condition is in fact treated as an additional parameter to the distribution. The condition is realized as an object of class **Condition**, which is a slot of corresponding classes **UnivariateCondDistribution**. This latter is the mother class to classes **AbscontCondDistribution** and **DiscreteCondDistribution**. The most important application of these classes so far is regression, where the distribution of the observation given the covariates is just realized as a **UnivariateCondDistribution**.

2.1.3 Parameter classes

As most distributions come with a parameter which often is of own interest, we endow the corresponding slots of a distribution class with an own parameter class, which allows for some checking like “Is the parameter **lambda** of an exponential distribution non-negative?”, “Is the parameter **size** of a binomial a positive integer?”

Consequently, we have a method **liesIn** that may answer such questions by a TRUE/FALSE statement. Schematically, the inheritance relations of class **Parameter** as well as the slots of the corresponding (sub-)classes may be read off in figure 2 where we do not repeat inherited slots. The most important set to be used as parameter domain/sample space (**rSpace**) will be an Euclidean space. So **rSpace** and **EuclideanSpace** are also implemented as classes, the structure of which may be read off in figure 3.

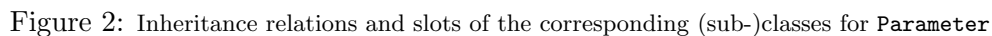
From version 1.9 on, we also have a subclass **Lattice**, which is still lacking in the preceding figure. It has slots **pivot** (of class “numeric”), **width** (of class “numeric” but tested against “==0”) and **Length** (of class “numeric” but tested to be an integer “>0” or **Inf**). All slots may be inspected/modified by the usual accessor/replacement functions.

2.2 Simulation classes

From version 1.6 on, the classes and methods of this subsection are available in package “**distrSim**”.

The aim of simulation classes is to gather all relevant information about a simulation in a correspondingly designed class. To this end we introduce the class **Dataclass** that serves as a common mother class for both “real” and simulated data. As derived classes we then have a simulation class where we also gather all information needed to reconstruct any particular simulation.

From version 1.8 of this package on, we have changed the format how data / simulations are stored: In order to be able to cope with multivariate, regression and (later) time series distributions, we have switched to the common array format **samplesize x obsDim x runs** where **obsDim** is the dimension of the observations. For saved objects from earlier versions, we provide the functions **isOldVersion** and **conv2NewVersion** to check whether the object was generated by an older version of this package and to convert such an object



Finally, coming from robust statistics we also consider situations where the majority of the data stems from an ideal situation/distribution whereas a minority comes from a contam-

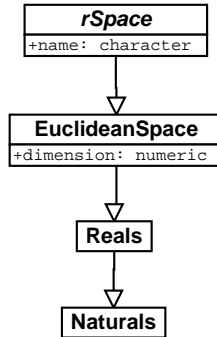


Figure 3: Inheritance relations and slots of the corresponding (sub-)classes for **rSpace**

inating source. To be able to identify ideal and contaminating observations, we also store this information in an indicator variable.

As the actual values of the simulations only play a secondary role, and as the number of simulated variables can become very large, but still easily reproducible, it is not worth storing all simulated observations but rather only the information needed to reproduce the simulation. This can be done by **savedata**.

Schematically, the inheritance relations of class **Dataclass** as well as the slots of the corresponding (sub-)classes may be read off in figure 4 where we do not repeat inherited slots. Also, analogously to package "**distr**", global options for the output by methods **plot** and

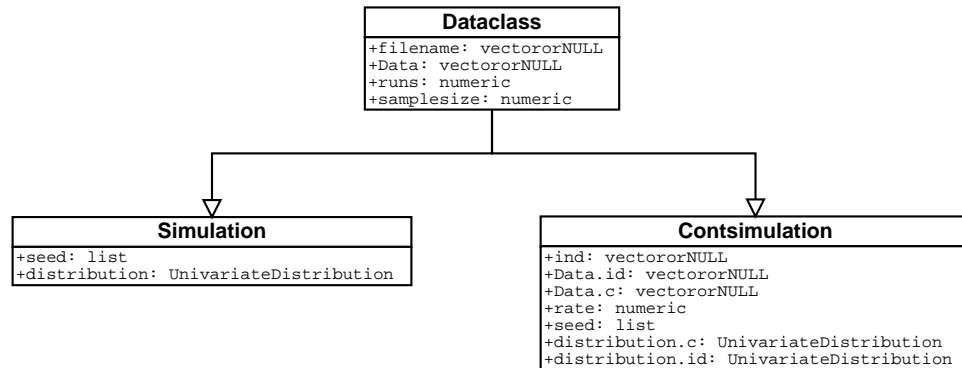


Figure 4: Inheritance relations and slots of the corresponding (sub-)classes for **Dataclass**

`summary` are controlled by `distrSimoptions()` and `getdistrSimoptions()`

2.3 Evaluation class

From version 1.6 on, the class and methods of this subsection are available in package "distrTest".

When investigating properties of a new procedure (e.g. an estimator) by means of simulations, one typically evaluates this procedure on a large set of simulation runs and gets a result for each run. These results are typically not available within seconds, so that it is worth storing them. To organize all relevant information about these results, we introduce a class **Evaluation** the slots of which is filled by method `evaluate` —see subsection 3.8. Schematically, the slots of this class may be read off in figure 5. A corresponding `savedata`

Evaluation
+name: character +filename: character +call.ev: call +result: vectororNULL +estimator: OptionalFunction

Figure 5: Slots of class **Evaluation**

method saves the object of class **Evaluation** in two files in the R-working directory: one using the filename `<filename>` also stores the results; the other one, designed to be “human readable”, comes as a comment file with filename `<filename>.comment` only stores the remaining information. The filename can be specified in the optional argument `fileN` to `savedata`; by default it is concatenated from the `filename` slot of the **Dataclass** object and `<estimatorname>`, which you may either pass as argument `estimatorName` or by default is taken as the R-name of the corresponding R-function specified in slot `estimator`.

From version 1.8 on, slot `result` in class **Evaluation** is of class **DataframeorNULL**, i.e.; may be either a data frame or `NULL`, and slot `call.ev` in class **Evaluation** is of class “**CallorNULL**”, i.e.; may be either a call or `NULL`. Also, we want to gather **Evaluation** objects in a particular data structure **EvaluationList** (see below), so we have to be able to check whether all data sets in the gathered objects coincide. For this purpose, from this version on, class **Evaluation** has an additional slot `Data` of class **Dataclass**. In order not to burden the objects of this class too heavily with uninformative simulated data, in case of a slot `Data` of one of the simulation-type subclasses of **Dataclass**, this `Data` itself has an empty `Data`-slot.

2.4 EvaluationList class

The class and methods of this subsection are available in package "distrTEst". In order to compare different procedures / estimators for the same problem, it is natural to gather several **Evaluation** objects with results of the same range (e.g. a parameter space) generated on the same data, i.e.; on the same **Dataclass** object. To this end, from version 1.8 on, we have introduced class **EvaluationList**. Schematically, the slots of this class may be read off in figure 6. The common **Data** slot of the **Evaluation** objects in an

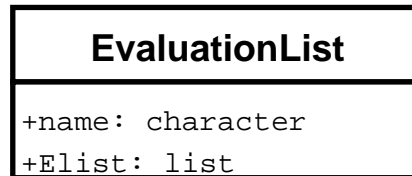


Figure 6: Slots of class **EvaluationList**

EvaluationList object may be accessed by the accessor method **Data**.

3 Methods

We have made available quite general arithmetical operations to our distribution objects, generating new image distributions automatically.

CAVEAT: These arithmetics operate on the corresponding r.v.'s and not on the distributions.

(For the latter, they only would make sense in restricted cases like convex combinations).

Martin Mächler pointed out that this might be confusing. So, this warning is also issued on attaching package "distr", and, by default, again whenever a **Distribution** object, produced by such arithmetics is shown or printed; this also applies to the last line in

```
> A1 <- Norm(); A2 <- Unif()
> A1 + A2
```

Distribution Object of Class: AbscontDistribution

```
Warning message:
arithmetics on distributions are understood as operations on r.v.'s
see 'distrARITH()'; for switching off this warning see '?distroptions' in:
print(object)
```

This behaviour will soon be annoying so you may switch it off setting the global option `WarningArith` to `FALSE` (see section 4).

3.1 Affine linear transformations

We have overloaded the operators `"+"`, `"-"`, `"*"`, `"/"` such that affine linear transformations which involve only single univariate r.v.'s are available; i.e. expressions like $Y = (3 \cdot X + 5) / 4$ are permitted for an object `X` of class `AbscontDistribution` or `DiscreteDistribution` (or some subclass), giving again an object `Y` of class `AbscontDistribution` or `DiscreteDistribution` (in general). Here the corresponding transformations of the `d`, `p`, and `q`-functions are done analytically.

From version 1.9 on, we use subclasses `AffLinAbscontDistribution`, `AffLinDiscreteDistribution`, `AffLinLatticeDistribution` as classes of the return values to enhance accuracy of functionals like `E`, `var`, etc. in package `"distrEx"`. These classes in addition to their counterparts without prefix `"AffLin"` have slots `a`, `b`, and `X0`, to capture the fact that an object of this class is distributed as $a \cdot X0 + b$. Also, we introduce a class union `AffLinDistribution` of classes `AffLinAbscontDistribution` and `AffLinDiscreteDistribution`. Consequently, the result `Y` of `Y <- a1 * X + b1` for an object `X` of (a subclass of) class `AffLinDiscreteDistribution` (if `a != 0`) is of the same class as `X` but with slots `Y@a = a1 * X@a`, `Y@b = b1 + X@b`, `Y@X0 = X@X0`.

3.2 The group math of unary mathematical operations

Also the group `math` of unary mathematical operations is available for distribution classes; so expressions like `exp(sin(3*X+5)/4)` are permitted. The corresponding `r` method consists in simply performing the transformation to the simulated values of `X`. The corresponding (default-) `d`, `p` and `q`-functions are obtained by simulation, using the technique described in the following subsection.

By means of `substitute`, the bodies of the `r`, `d`, `p`, `q`-slots of distributions show the parameter values with which they were generated; in particular, convolutions and applications of the group `math` may be traced in the `r`-slot of a distribution object, compare `r(sin(Norm()) + cos(Unif() * 3 + 2))`.

Initially, it might be irritating that the same "arithmetic" expression evaluated twice in a row gives two different results, compare

```
> A1 <- Norm(); A2 <- Unif()
> d(sin(A1 + A2))(0.1)
```

```
[1] 0.3797933
```

```
> d(sin(A1 + A2))(0.1)
```

```
[1] 0.3786348
```



```
> sin(A1 + A2)
```

Distribution Object of Class: AbscontDistribution

This is due to the fact, that all slots are filled starting from simulations. To explain this, a warning is issued by default, whenever a `Distribution` object, filled by such simulations is shown or printed; this also applies to the last line in the preceding code snippet. This behaviour may again be switched off by setting the global option `WarningSim` to `FALSE` (see section 4).

As they are frequently needed, from version 1.9 on, math operations `abs()`, `exp()`, and—if an R-version $\geq 2.6.0$ is used—also `log()` are implemented in an analytically exact form, i.e.; with exact expressions for slots `d`, `p`, and `q`.

3.3 Construction of `d`, `p`, and `q` from `r`

In order to facilitate automatic generation of new distributions, in particular those arising as image distributions under transformations of correspondingly distributed random variables, we provide ad hoc methods that should be overloaded by more exact ones wherever possible: By means of the function `RtoDPQ` we first generate $10^{\text{RtoDPQ.e}}$ random numbers where `RtoDPQ.e` is a global option of this package and is discussed in section 4. A density estimator is evaluated along this sample, the distribution function is estimated by the empirical c.d.f. and, finally, the quantile function is produced by numerical inversion. Of course the result is rather crude as it relies on the law of large numbers only, but this way all transformations within the group `math` become available. Where laws under transformations can easily be computed exactly—as for affine linear transformations—we replace this procedure by the exactly transformed `d`, `p`, `q`-methods.

3.4 Convolution

A convolution method for two independent r.v.'s is implemented by means of explicit calculations for discrete summands, and by means of DFT/FFT⁷ if one of the summands is absolutely continuous or (from version 1.9 on:) both are lattice distributed with a common lattice as support. This method automatically generates the law of the sum of two independent variables/distributions X and Y of any univariate distributions—or in S4-jargon: the addition operator `"+"` is overloaded for two objects of class `UnivariateDistribution` and corresponding subclasses.

3.5 Overloaded generic functions

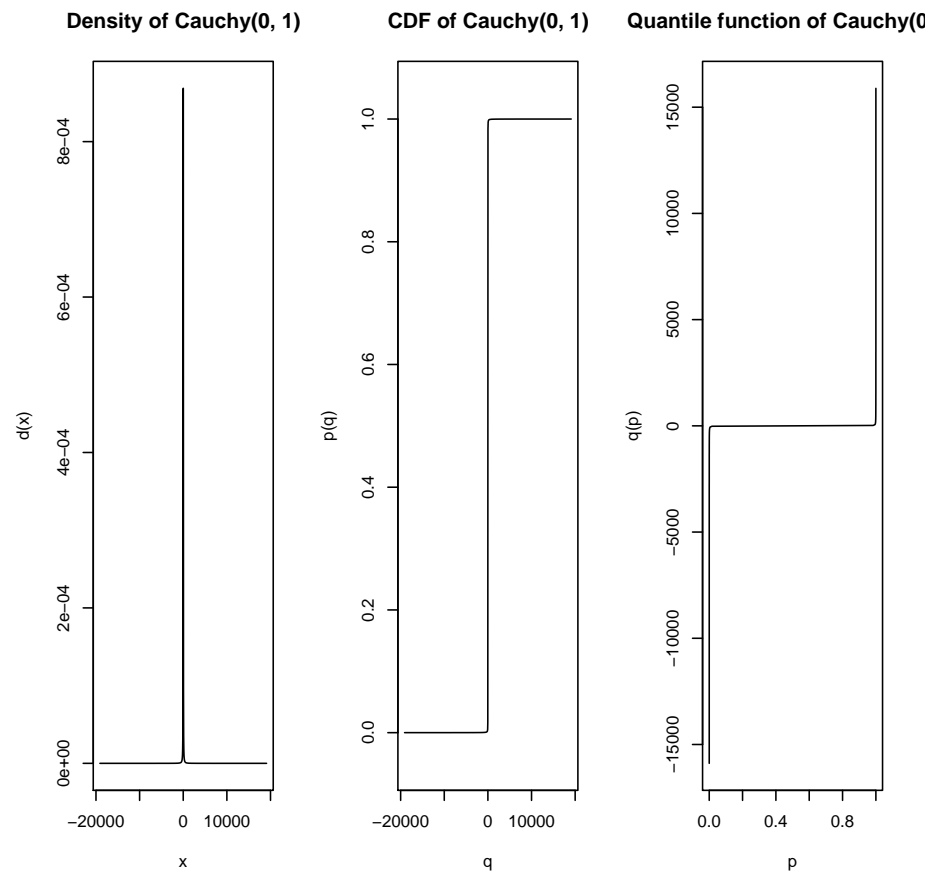
Methods `print`, `plot`, `show` and `summary` have been overloaded for classes `Distribution`, `Dataclass`, `Simulation`, `ContSimulation`, as well as `Evaluation` and `EvaluationList`

⁷Details to be found in [5]

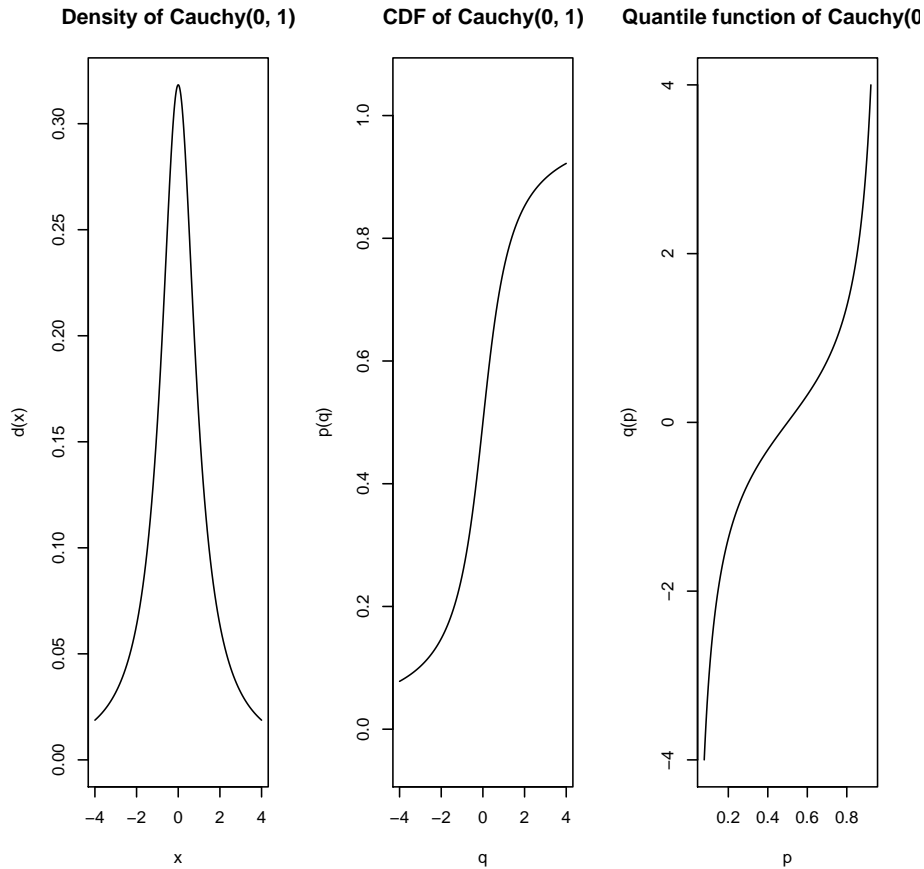
to produce “pretty” output. `print`, `plot`, `show` and `summary` have additional, optional arguments for plotting subsets of the simulations / results: index vectors for the dimensions, the runs, the observations, and the evaluations may be passed using arguments `obs0`, `runs0`, `dims0`, `eval0`, confer `help("<mthd>-methods", package=<pkg>)` where `<mthd>` stands for `plot`, `show`, `print`, or `plot`, and `<pkg>` stands for either `"distrSim"` or `"distrTEst"`.

For an object of class `Distribution`, `plot` displays the density/probability function, the c.d.f. and the quantile function of a distribution. Note that all usual parameters of `plot` remain valid. For instance, you may increase the axis annotations and so on. More important, you may also override the automatically chosen x -region by passing an `xlim` argument:

```
> plot(Cauchy(), withSweave = TRUE)
```



```
> plot(Cauchy(), xlim=c(-4,4), withSweave = TRUE)
```



Moreover you may control optional main, inner titles and subtitles with arguments **main** / **sub** / **inner**. To this end there are preset strings substituted in both expression and character vectors (where in the following **x** denotes the argument with which **plot()** was called)

%A deparsed argument **x**

%C class of argument **x**

%P comma-separated list of parameter values of slot **param** of argument **x**

%Q comma-separated list of parameter values of slot **param** of argument **x** in parenthesis unless this list is empty; then ""

%N comma-separated **<name> = <value>** - list of parameter values of slot **param** of argument **x**

%D time/date at which plot is/was generated

As usual you may control title sizes and colors with `cex.main` / `cex.inner` / `cex.sub` respectively with `col` / `col.main` / `col.inner` / `col.sub`. Additionally it may be helpful to control top and bottom margins with arguments `bmar`, `tmar`. `plot()` can also cope with `log`-arguments. We provide different default symbols for unattained [`pch.u`] / attained [`pch.a`] one-sided limits, which may be overridden by corresponding arguments `pch` / `pch.a` / `pch.u`.

For objects of class `AbscontDistribution`, you may set the number of grid points used by an `ngrid` argument; also the “quantile”-panel takes care of finite left/right endpoints of support and optionally tries to identify constancy region of the `p`-slot.

For objects of class `DiscreteDistributions`, we use `stepfun()` from package “base” as far as possible and (also for panel “q” for `AbscontDistributions`) consequently take over its arguments `do.points`, `verticals`, `col.points` / `col.vert` / `col.hor` and `cex.points`.

As examples consider the following 10 plots:

For objects of class `Dataclass` —or of a corresponding subclass— `plot` plots the sample against the run index and in case of `ContSimulation` the contaminating variables are highlighted by a different color. Additional arguments controlling the plot as in the default `plot` command may be passed, confer `help("plot-methods", package="distrSim")`.

For an object of class `Evaluation`, `plot` yields a boxplot of the results of the evaluation. For an object of class `EvaluationList`, `plot` regroups the list according to the different columns/coordinates of the result of the evaluation; for each such coordinate, a boxplot is generated, containing possibly several procedures, and, if evaluated at a `Contsimulation`, the plots are also grouped into evaluations on ideal and real data. As for the usual `boxplot` function you may pass additional “plot-type” arguments to this particular `plot` method, confer `help("plot-methods", package="distrTEst")`. In particular, the `plot`-arguments `main` and `ylim`, however, may also be transmitted coordinatewise, i.e.; a vector of the same length as the dimension of the result `resDim` (e.g. parameter dimension), respectively a 2 x `resDim` matrix, or they may be transmitted globally, using the usual `S` recycling rules.

3.6 liesInSupport

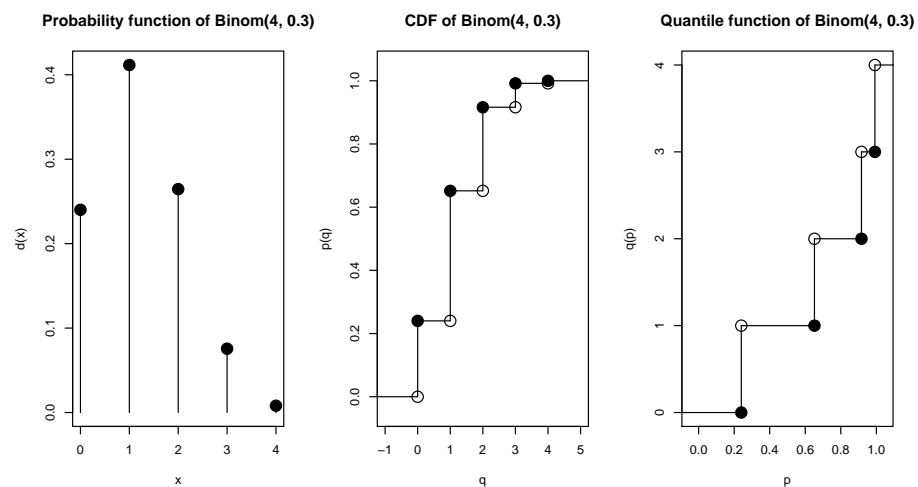
For all discrete distribution classes, we have methods `liesInSupport` to check whether a given vector/ a matrix of points lies in the support of the distribution.

3.7 Simulation (in package "distrSim")

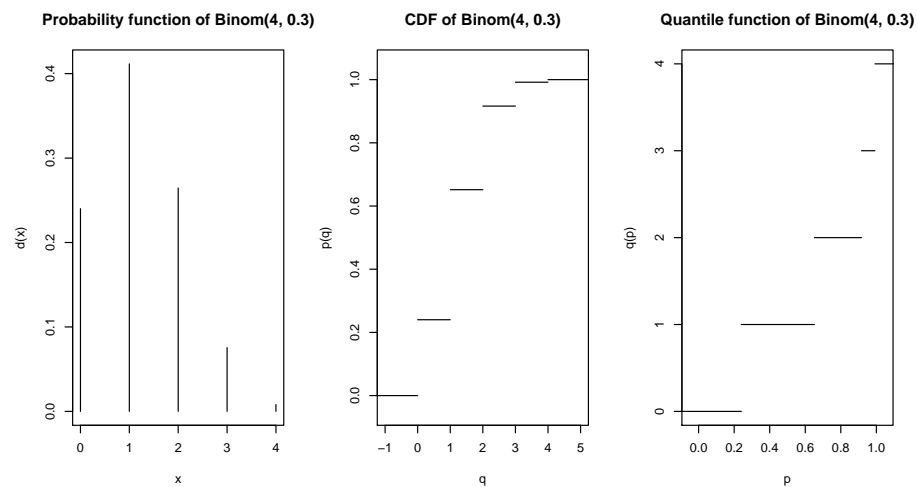
From version 1.6 on, `simulation` is available in package “distrSim”.

For the classes `Simulation` and `ContSimulation`, we normally will not save the current values of the simulation, as they can easily be reproduced knowing the values of the other slots of this class. So when declaring a new object of either of the two classes, the slot `Data` will be empty (`NULL`). To fill it with the simulated values, we have to apply the method `simulate` to the object. This has to be redone whenever another slot of the object is

```
> plot(Binom(size = 4, prob = 0.3), withSweave = TRUE)
```

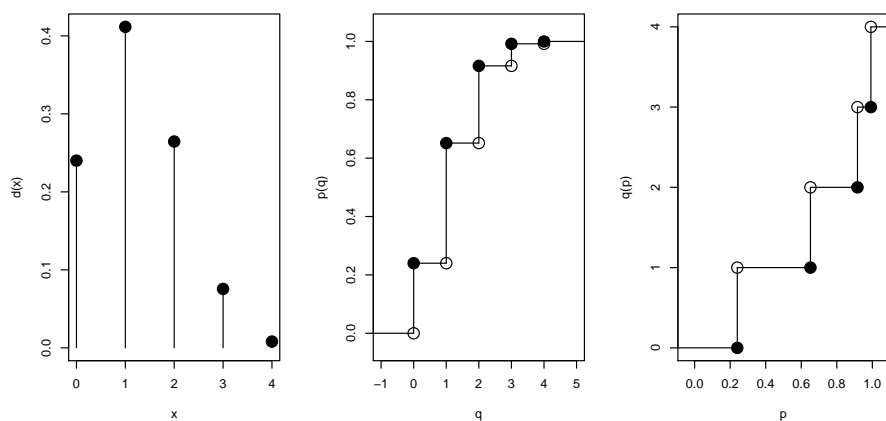


```
> plot(Binom(size = 4, prob = 0.3), do.points = FALSE, verticals = FALSE,
+      withSweave = TRUE)
```

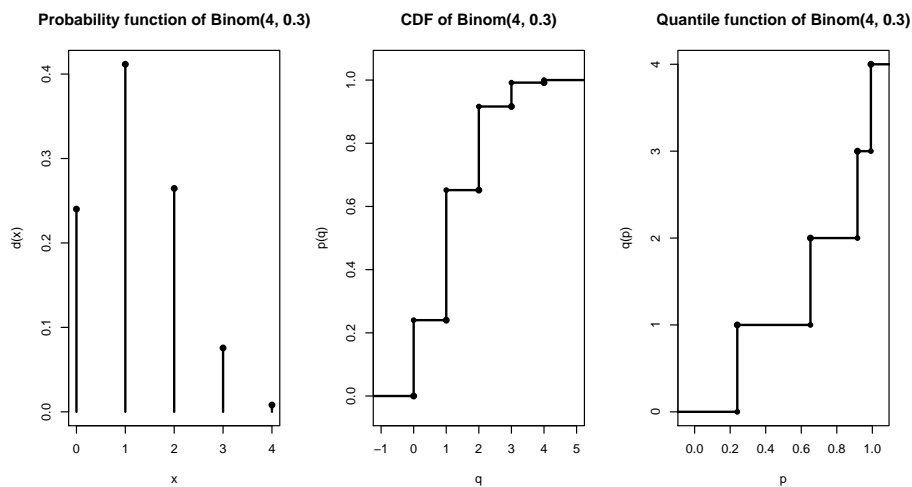


```
> plot(Binom(size = 4, prob = 0.3), main = TRUE, inner = FALSE, cex.main = 1.6,
+       tmar = 6, withSweave = TRUE)
```

Distribution Plot for Binom(size = 4, prob = 0.3)



```
> plot(Binom(size = 4, prob = 0.3), cex.points = 1.2, pch = 20, lwd = 2,
+       withSweave = TRUE)
```

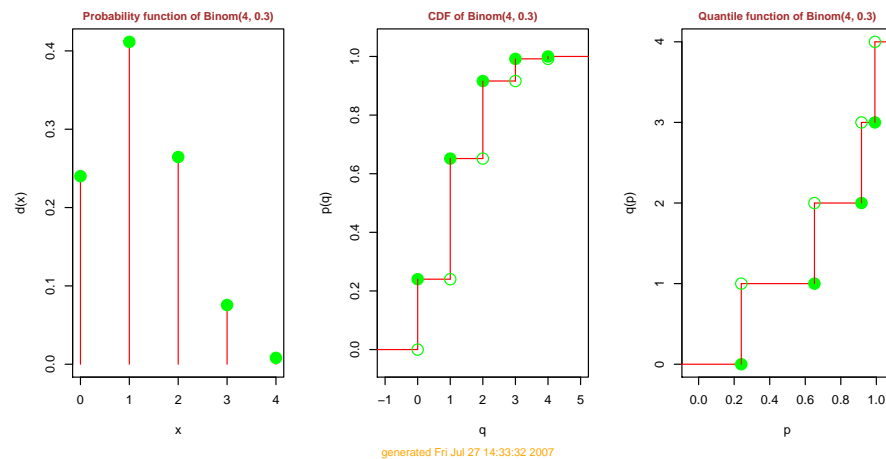


```

> B <- Binom(size = 4, prob = 0.3)
> plot(B, col="red", col.points = "green", main = TRUE, col.main="blue",
+      col.sub = "orange", sub = TRUE, cex.sub = 0.6, col.inner = "brown",
+      withSweave = TRUE)

```

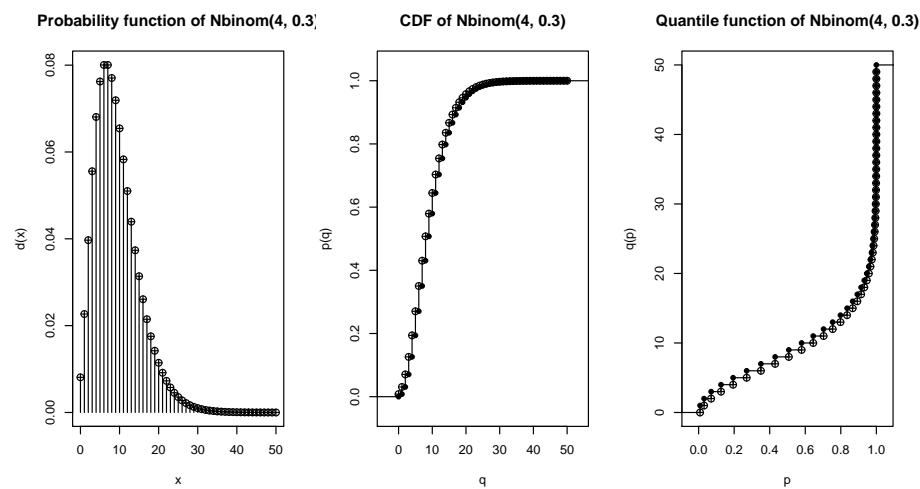
Distribution Plot for B



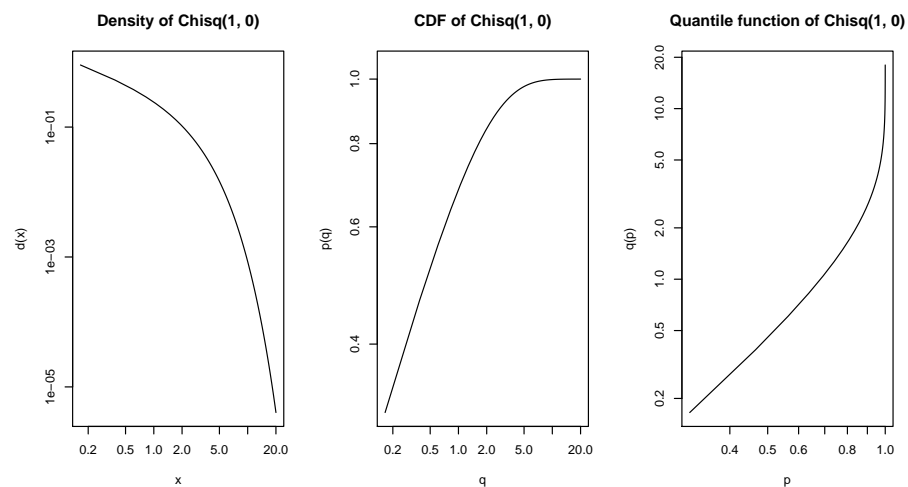
```

> plot(Nbinom(size = 4, prob = 0.3), cex.points = 1.2, pch.u = 20, pch.a = 10,
+      withSweave = TRUE)

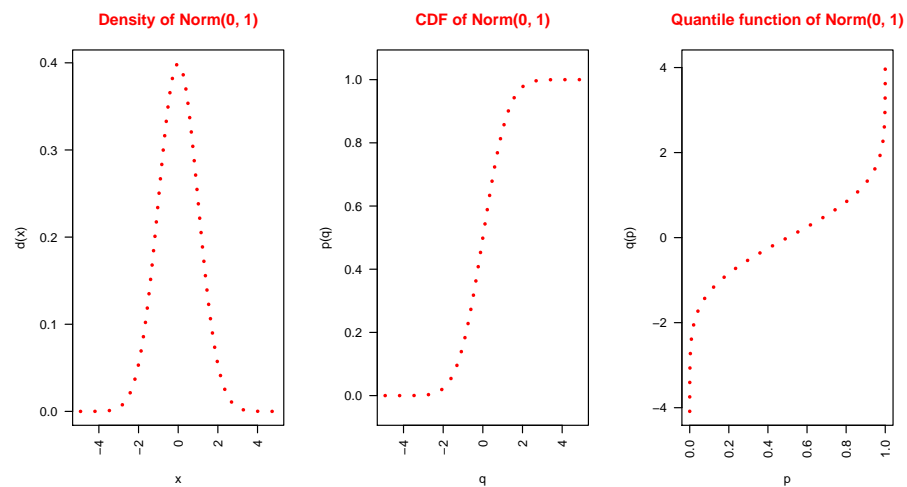
```



```
> plot(Chisq(), log = "xy", ngrid = 100, withSweave = TRUE)
```



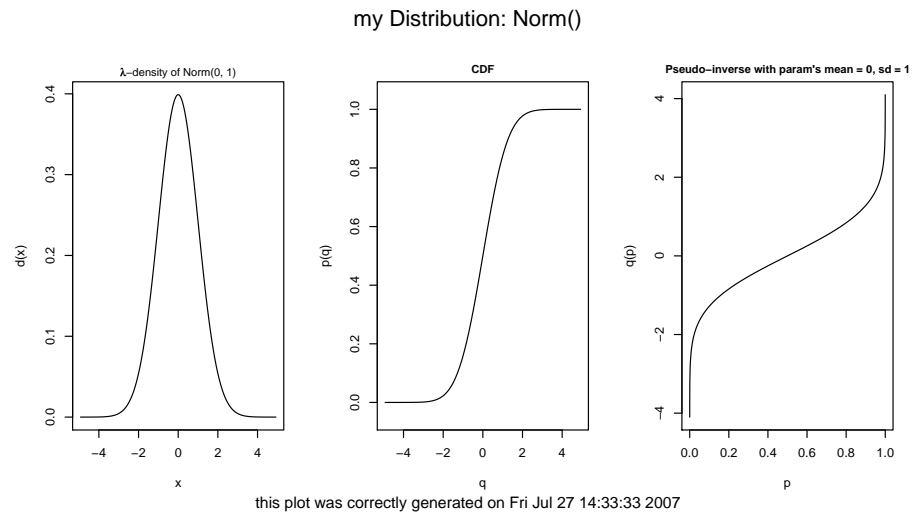
```
> plot(Norm(), lwd=3, col = "red", ngrid = 200, lty = 3, las = 2,
+       withSweave = TRUE)
```




```

> plot(Norm(), main = "my Distribution: \%A",
+       inner = list(expression(paste(lambda, "-density of \%C(\%P)")), "CDF",
+       "Pseudo-inverse with param's \%N"),
+       sub = "this plot was correctly generated on \%D",
+       cex.inner = 0.9, cex.sub = 0.8, withSweave = TRUE)

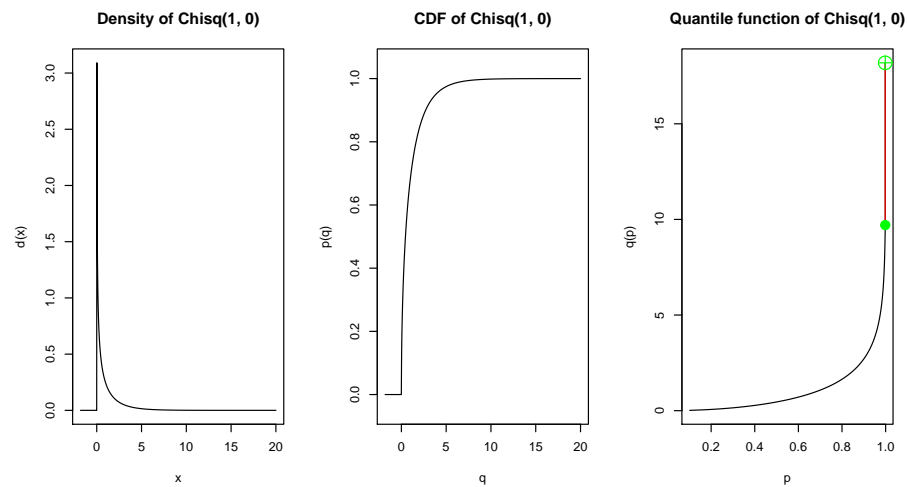
```



```

> Ch <- Chisq(); setgaps(Ch, exactq = 3)
> plot(Ch, cex = 1.2, pch.u = 20, pch.a = 10, col.points = "green",
+       col.vert = "red", withSweave = TRUE)

```



changed. To guarantee reproducibility, we use the slot `seed`.

This slot is controlled and set through **Paul Gilbert's** "`setRNG`" package. By default, `seed` is set to `setRNG()`, which returns the current "state" of the random number generator (RNG). So the user does not need to specify a value for `seed`, and nevertheless may reproduce his samples: He simply uses `simulate` to fill the `Data` slot. If the user wants to, he may also set the `seed` explicitly via the replacement function `seed()`, but has to take care of the correct format himself, confer the documentation of `setRNG`. One easy way to fill the `Data` slot of a simulation `X` with "new" random numbers is

```
> have.distrSim <- suppressWarnings(require("distrSim"))
> if (have.distrSim)
+   {X <- Simulation()
+     seed(X) <- setRNG()
+     simulate(X)
+     print(Data(X)[1:10])
+   } else {
+     cat("\n functionality not (yet) available; ")
+     cat("you have to install package \"distrSim\" first.\n")
+   }

[1] -0.37071086  0.80980890  1.74966290 -0.02922246  2.24155515  2.59616560
[7]  1.51202959 -0.71274533  0.51371994  1.03078964
```

3.8 Evaluate (in package "distrTEst")

From version 1.6 on `evaluate` is available in "`distrTEst`".

In an object of class `Evaluation` we store relevant information about an evaluation of a statistical procedure (estimator/test/predictor) on an object of class `Dataclass`, including the concrete results of this evaluation. An object of class `Evaluation` is generated by an application of method `evaluate` which takes as arguments an object of class `Dataclass` and a procedure of type `function`. As an example, confer Example 11.8. For data of class `Contsimulation`, the result takes a slightly different, combining evaluations on ideal and real data.

3.9 Is-Relations

By means of `setIs`, we have "told" R that a distribution object `obj` of class

- "`Unif`" with $\text{Min} \doteq 0$ and $\text{Max} \doteq 1$ also is a Beta distribution with parameters `shape1 = 1`, `shape2 = 1`
- "`Geom`" also is a negative Binomial distribution with parameters `size = 1`, `prob = prob(obj)`

- "Cauchy" with `location` $\doteq 0$ and `scale` $\doteq 1$ also is a T distribution with parameters `df = 1`, `ncp = 0`
- "Exp" also is a Gamma distribution with parameters `shape = 1`, `scale = 1/rate(obj)` and a Weibull distribution with parameters `shape = 1`, `scale = 1/rate(obj)`
- "Chisq" with non-centrality `ncp` $\doteq 0$ also is a Gamma distribution with parameters `shape = df(obj)/2`, `scale = 2`
- "DiscreteDistribution" (from version 1.9 on) with an equally spaced support also is a "LatticeDistribution"

3.10 Further methods

When iterating/chaining mathematical operations on a univariate distribution, generation process of random variables can become clumsy and slow. To cope with this, we introduce a sort of “Forget-my-past”-method `simplifyr` that replaces the chain of mathematical operations in the `r`-method by drawing with replacement from a large sample ($10^{\text{RtoDPQ.e}}$) of these.

3.11 Functionals (in package "distrEx")

3.11.1 Expectation

The most important contribution of package "distrEx" is a general expectation operator. In basic statistic courses, the expectation E may come as $E[X]$, $E[f(X)]$, $E[X|Y = y]$, or $E[f(X)|Y = y]$. Our operator (or in S4-language “generic function”) `E` covers all of these situations (or *signatures*).

default call The most frequent call will be $E(X)$ where X is an (almost) arbitrary distribution object. More precisely, if X is of a specific distribution class like `Pois`, it is evaluated exactly using analytic terms. Else if it is of class `DiscreteDistribution` we use a sum over the support of X , and if it is of class `AbscontDistribution` we use numerical integration⁸; if we only know that X is of class `UnivariateDistribution` we use Monte-Carlo integration. This also is the default method in for class `MultivariateDistribution`, while for `DiscreteMVDistribution` we again use sums. For an object Y of a subclass of class union `AffLinDistribution`, we determine the expectation as $Y@a * E(Y@X0) + Y@b$ and hence use analytic terms for $X0$ if available.

⁸i.e., we first try (really!): `try(integrate` and if this fails we use Gauß-Legendre integration according to [6], see also `?distrExIntegrate`

with a function as argument we proceed just as without: if X is of class `DiscreteDistribution`, we use a sum over the support of X , and if X is of class `AbscontDistribution` we use numerical integration; else we use Monte-Carlo integration.

in addition: with a condition as argument we simply use the corresponding `d` respective `r` slots with the additional argument `cond`.

exact evaluation is available for X of class `Beta` (for noncentrality 0), `Binom`, `Cauchy`, `Chisq`, `Dirac`, `Exp`, `Fd`, `Gammad`, `Geom`, `Hyper`, `Logis`, `Lnorm`, `Nbinom`, `Norm`, `Pois`, `Td`, `Unif`, `Weibull`.

examples

```
> have.distrEx <- suppressWarnings(require("distrEx"))
> if (have.distrEx)
+   {D4 <- LMCondDistribution(theta = 1)
+     D4 # corresponds to Norm(cond, 1)
+     N <- Norm(mean = 2)
+
+     print(E(D4, cond = 1))
+     print(E(D4, cond = 1, useApply = FALSE))
+     print(E(as(D4, "UnivariateCondDistribution"), cond = 1))
+     print(E(as(D4, "UnivariateCondDistribution"), cond = 1,
+       useApply = FALSE))
+     print(E(D4, function(x){x^2}, cond = 2))
+     print(E(D4, function(x){x^2}, cond = 2, useApply = FALSE))
+     print(E(N, function(x){x^2}))
+     print(E(as(N, "UnivariateDistribution"), function(x){x^2},
+       useApply = FALSE)) # crude Monte-Carlo
+     print(E(D4, function(x, cond){cond*x^2}, cond = 2,
+       withCond = TRUE))
+     print(E(D4, function(x, cond){cond*x^2}, cond = 2,
+       withCond = TRUE, useApply = FALSE))
+     print(E(N, function(x){2*x^2}))
+     print(E(as(N, "UnivariateDistribution"), function(x){2*x^2},
+       useApply = FALSE)) # crude Monte-Carlo
+     Y <- 5 * Binom(4, .25) - 3
+     print(Y); print(E(Y))
+   } else {
+     cat("\n functionality not (yet) available; ")
+   }
```

```

+   cat("you have to install package \"distrEx\" first.\n")
+ }

[1] 0.9999998
[1] 0.9999998
[1] 1.003117
[1] 1.004277
[1] 4.999993
[1] 4.999993
[1] 4.999993
[1] 4.994825
[1] 9.999987
[1] 9.999987
[1] 9.999987
[1] 9.954926
Distribution Object of Class: AffLinLatticeDistribution
[1] 2

```

3.11.2 Variance

The next-common functional is the variance. In order to keep a unified notation we will use the same name as for the empirical variance, i.e., `var`.

masking "stats"-method `var` To cope with the different argument structure of the empirical variance, i.e. `var(x, y = NULL, na.rm = FALSE, use)` and our functional variance, i.e., `var(x, fun = function(t) t, cond, withCond = FALSE, useApply = TRUE, ...)` we have to mask the original "stats"-method:

```

> var <- function(x , ...)
+   {dots <- list(...)}
+   if(hasArg(y)) y <- dots$"y"
+   na.rm <- ifelse(hasArg(na.rm), dots$"na.rm", FALSE)
+   if(!hasArg(use))
+     use <- ifelse (na.rm, "complete.obs", "all.obs")
+   else use <- dots$"use"
+   if(hasArg(y))
+     stats::var(x = x, y = y, na.rm = na.rm, use)
+   else
+     stats::var(x = x, y = NULL, na.rm = na.rm, use)
+   }

```

before registering `var` as generic function. Doing so, if the `x` (or the first) argument of `var` is not of class `UnivariateDistribution`, `var` behaves identically to the "stats" package

default method if `x` is of class `UnivariateDistribution`, `var` just returns the variance of distribution `X` — or of `fun(X)` if a function is passed as argument `fun`, or, if a condition argument `cond` (for $Y = y$), $\text{Var}[X|Y = y]$ respectively $\text{Var}[f(X)|Y = y]$ — just as for `E`. For an object `Y` of a subclass of class union `AffLinDistribution`, we determine the variance as `Y@a^2 * var(Y@X0)` and hence use analytic terms for `X0` if available.

exact evaluation is provided for specific distributions if no function and no condition argument is given: this is available for `X` of class `Beta` (for noncentrality 0), `Binom`, `Cauchy`, `ChisqDirac`, `Exp`, `Fd`, `Gammad`, `Geom`, `Hyper`, `Logis`, `Lnorm`, `Nbinom`, `Norm`, `Pois`, `Unif`, `Td`, `Weibull`.

3.11.3 Further functionals

By the same techniques we provide the following functionals for univariate distributions:

- standard deviation: `sd`
- skewness: `skewness` (code contributed by G. Jay Kerns, gkerns@ysu.edu)
- kurtosis: `kurtosis` (code contributed by G. Jay Kerns, gkerns@ysu.edu)
- median: `median` (not for function/condition arguments)
- median of absolute deviations: `mad` (not for function/condition arguments)
- interquartile range: `IQR` (not for function/condition arguments)

3.12 Truncated moments (in package "distrEx")

For Robust Statistics, the first two truncated moments are very useful. These are realized as generic functions `m1df` and `m2df`: They use the expectation operator for general univariate distributions, but are overloaded for most specific distributions:

- `Binom`
- `Pois`
- `Norm`
- `Exp`
- `Chisq`

3.13 Distances (in package "distrEx")

For several purposes like Goodness-of-fit tests or minimum-distance estimators, distances between distributions are useful. This applies in particular to Robust Statistics. In package "distrEx", we provide the following distances:

- Kolmogoroff distance
- total variation distance
- Hellinger distance
- convex-contamination “distance” (asymmetric!) defined as

$$d(Q, P) := \inf\{r > 0 \mid \exists \text{ probability } H : Q = (1 - r)P + rH\}$$

3.14 Functions for demos (in package "distrEx")

To illustrate the possibilities with packages "distr" and "distrEx" we include two major demos, each with extra code to it

3.14.1 CLT for arbitrary summand distribution

By means of our convolution algorithm as well as with the operators **E** and **sd** an illustration for the CLT is readily written: function `illustrateCLT`, respectively demo `illustCLT`. For plotting, we have particular methods for discrete and absolute continuous distributions. The user may specify a given summand distribution, an upper limit for the consecutive sums to be considered and a pause between the corresponding plots in seconds. From version 1.9 on, we also include a **TclTk**-based version of this demo, where the user may enter the distribution argument (i.e.; the summands’ distribution) into a text line and control the sample size by a slider in some widget: `illustCLT_tcl`

3.14.2 LLN for arbitrary summand distribution

From version 1.9 on, similarly, we provide an illustration for the LLN: function `illustrateLLN`, respectively demo `illustLLN`. The user may specify a vector of sample sizes to be considered, the number of replicates to be drawn and a pause between the corresponding plots in seconds, also, optionally, the limiting expectation (in case of class **Cauchy**: the non-limiting median) is drawn as a line and Chebyshev/CLT-based (pointwise) confidence bands and their respective empirical coverages are displayed.

3.14.3 Deconvolution example

To illustrate conditional distributions and their implementation in "distrEx", we consider the following situation: We consider a signal $X \sim P^X$ which is disturbed by noise $\varepsilon \sim P^\varepsilon$, independent from X ; in fact we observe $Y = X + \varepsilon$ and want to reconstruct X by means of Y . By means of the generating function `PrognCondDistribution` of package "distrEx", for absolutely continuous P^X, P^ε , we may determine the factorized conditional distribution $P^{X|Y=y}$, and based on this either its (posterior) mode oder (posterior) expectation; also see `demo(Prognose, package="distrEx")`.

4 Options

4.1 Options for "distr"

Analogously to the `options` command in R you may specify a number of global “constants” to be used within the package. These include

- **DefaultNrFFTGridPointsExponent**: the binary logarithm of the number of grid-points used in FFT —default 12
- **DefaultNrGridPoints**: number of grid-points used for a continuous variable —default 4096
- **DistrResolution**: the finest step length that is permitted for a grid for a discrete variable —default 1e-06
- **RtoDPQ.e**: For simulational determination of **d**, **p** and **q**, $10^{\text{RtoDPQ.e}}$ random variables are simulated —default 5
- **TruncQuantile**: to work with compact support, random variables are truncated to their lower/upper **TruncQuantile**-quantile —default 1e-05.
From version 1.9 on, for $\varepsilon = \text{TruncQuantile}$, we use calls of form `q(X)(eps, lower.tail = FALSE)` instead of `q(X)(1-eps)` to gain higher precision.
- **warningSim**: controls whether a warning issued at printing/showing a **Distribution** object the slots of which have been filled starting with simulations —default TRUE
- **warningArith**: controls whether a warning issued at printing/showing a **Distribution** object produced by arithmetics operating on distributions —default TRUE

All current options may be inspected by `distroptions()` and modified by `distroptions("<options-name>"=<value>)`.

As options, `distroptions("<options-name>")` returns a list of length 1 with the value of the corresponding option, so here, just as `getOption`, `getdistrOption("<options-name>")` will be preferable, which only returns the value.

4.2 Options for "distrEx"

Up to version 0.4-4 we used the function `distrExOptions(arg = "missing", value = -1)` to manage some global options for "distrEx", i.e.:

`distrExOptions()` returns a list of these options, `distrExOptions(arg=x)` returns option `x`, and `distrExOptions(arg=x,value=y)` sets the value of option `x` to `y`.

From version 1.9 on, we use a mechanism analogue to the `distroptions/getdistrOption` commands: You may specify certain global output options to be used within the package with `distrExoptions/getdistrExOption`. These include

- **MCIterations**: number of Monte-Carlo iterations used for crude Monte-Carlo integration.
- **GLIntegrateTruncQuantile**: If `integrate` fails and there are infinite integration limits, the function `GLIntegrate` is called inside of `distrExIntegrate` with the corresponding quantiles `GLIntegrateTruncQuantile` resp. `1-GLIntegrateTruncQuantile` as finite integration limits.
- **GLIntegrateOrder**: The order used for the Gauß-Legendre integration inside of `distrExIntegrate`.
- **ElowerTruncQuantile**: The lower limit of integration used inside of `E` which corresponds to the `ElowerTruncQuantile`-quantile.
- **EupperTruncQuantile**: The upper limit of integration used inside of `E` which corresponds to the `(1-ElowerTruncQuantile)`-quantile.
- **ErelativeTolerance**: The relative tolerance used inside of `E` when calling `distrExIntegrate`.
- **m1dfLowerTruncQuantile**: The lower limit of integration used inside of `m1df` which corresponds to the `m1dfLowerTruncQuantile`-quantile.
- **m1dfRelativeTolerance**: The relative tolerance used inside of `m1df` when calling `distrExIntegrate`.
- **m2dfLowerTruncQuantile**: The lower limit of integration used inside of `m2df` which corresponds to the `m2dfLowerTruncQuantile`-quantile.
- **m2dfRelativeTolerance**: The relative tolerance used inside of `m2df` when calling `distrExIntegrate`.

4.3 Options for "distrSim"

Just as with to the `distroptions/getdistrOption` commands you may specify certain global output options to be used within the package with `distrSimoptions/getdistrSimOption`. These include

- `MaxNumberOfPlottedObs` the maximal number of observation plotted in a plot of an object of class `Dataclass`; defaults to 4000
- `MaxNumberOfPlottedObsDims`: the maximum number of observations to be plotted in a plot of an object of class `Dataclass` and descendants; defaults to 6.
- `MaxNumberOfPlottedRuns`: the maximum number of runs to be plotted in a plot of an object of class `Dataclass` and descendants (one run/panel); defaults to 6.
- `MaxNumberOfSummarizedObsDims`: the maximum number of observations to be summarized of an object of class `Dataclass` and descendants; defaults to 6.
- `MaxNumberOfSummarizedRuns`: the maximum number of runs to be summarized of an object of class `Dataclass` and descendants; defaults to 6.

4.4 Options for "distrTest"

Just as with to the `distroptions/getdistrOption` commands you may specify certain global output options to be used within the package with `distrTestoptions/getdistrTestOption`. These include

- `MaxNumberOfPlottedEvaluations`: the maximal number of evaluations to be plotted in a plot of an object of class `EvaluationList`; defaults to 6
- `MaxNumberOfPlottedEvaluationDims`: the maximal number of evaluation dimensions to be plotted in a plot of an object of class `Evaluation`; defaults to 6
- `MaxNumberOfSummarizedEvaluations`: the maximal number of evaluations to be summarized of an object of class `EvaluationList`; defaults to 15
- `MaxNumberOfPrintedEvaluations`: the maximal number of evaluations printed of an object of class `EvaluationList`; defaults to 15

5 Startup Messages

For the management of startup messages, from version 1.7, we use package `"startupmsg"`: When loading/attaching packages `"distr"`, `"distrEx"`, `"distrSim"`, or `"distrTest"` for each package a disclaimer is displayed.

You may suppress these start-up banners/messages completely by setting `options("StartupBanner"="off")` somewhere before loading this package by `library` or `require` in your R-code / R-session.

If option `"StartupBanner"` is not defined (default) or setting `options("StartupBanner" = NULL)` or `options("StartupBanner" = "complete")` the complete start-up banner is displayed.

For any other value of option `"StartupBanner"` (i.e., not in `c(NULL, "off", "complete")`) only the version information is displayed.

The same can be achieved by wrapping the `library` or `require` call into either `onlytypeStartupMessages(<code>, atypes="version")` or `suppressStartupMessages(<code>)`.

6 System/version requirements, license, etc.

6.1 System requirements

As our package is completely written in R, there are no dependencies on the underlying OS; of course, there is the usual speed gain possible on recent machines. We have tested our package on a Pentium II with 233 MHz, on Pentium III's with 0.8–2.1 GHz, and on an Athlon with 2.5 GHz giving a reasonable performance.

6.2 Required version of R

Contrary to the hardware required, if you want to use `library` or `require` to use `"distr"` within R code, you need at least R Version 1.8.1, as we make use of name space operations only available from that version on; also, the command `setClassUnion`, which is used in some sources, is only available from that version on.

On the other hand, if the package may be pasted in by `source`, the code works with R from version 1.7.0 on—but without using name-spaces, which is only available from 1.8.0 on. Due to some changes in R from version 1.8.1 to 1.9.0 and from 1.9.1 to 2.0.0, we have to provide different `zip/tar.gz`-Files for these versions.

Versions of `"distr"` from version number 1.5 onwards are only supplied for R Version 2.0.1 `patched` and later. After a reorganization, versions of `"distr"` from version number 1.6 onwards are only supplied for R Version 2.2.0 `patched` and later.

6.3 Dependencies

In package `"distrEx"`, we need [Alec Stephenson's](#) package `"evd"` for the extreme value distributions implemented therein. In package `"distrSim"`, and consequently also in package `"distrTest"` we use [Paul Gilbert's](#) package `"setRNG"` to be installed from [CRAN](#) for the control of the seed of the random number generator in our simulation classes. More

precisely, for our version ≤ 1.6 we need his version $< 2006.2-1$, and for our version ≥ 1.7 we need his version $\geq 2006.2-1$.

From package version 1.7/0.4-3 on, we also need package "startupmsg" by the first of the present authors, which also is available on [CRAN](#).

6.4 License

This software is distributed under the terms of the GNU GENERAL PUBLIC LICENSE Version 2, June 1991, confer

<http://www.gnu.org/copyleft/gpl.html>

7 Details to the implementation

- As the normal distribution is closed under affine transformations, we have overloaded the corresponding methods.
- For the general convolution algorithm for univariate probability distribution functions/densities by means of FFT, which we use in the overloaded "+"-operator, confer [5].
- Exact convolution methods are implemented for the normal, the Poisson, the binomial, the negative binomial, the Gamma (and the **Exp**), and the χ^2 distribution
- Exact formulae for scale transformations are implemented for the Exp-/Gamma-distribution, the Weibull and the log-normal distribution (the latter two from version 1.9 on).
- Exact formulae for affine linear transformations are available for the normal, the logistic and the Cauchy distribution (the latter two from version 1.9 on).
- Instances of any class transparent to the user are initialized by `<classname>([<slotname>=<value>,...])` where except for class **DataClass** in package "distrSim" all classes have default values for all their slots; in **DataClass**, the slot **Data** has to be specified.
- As suggested in [3] all slots are accessed and modified by corresponding accessor- and replacement functions —templates for which were produced by **standardMethods**.

We strongly discourage the use of the @-operator to modify or even access slots r, d, p, and q, confer Example 11.7.

8 A general utility

Following [3], the programmer of **S4**-classes should provide accessor and replacement functions for the inspection/modification of any newly introduced slot. This can be quite a task when you have a lot of classes/slots. As these functions all have the same structure, it would be nice to automatically generate templates for them. Faced with this problem in developing this package, Thomas Stabla has written such a utility, **standardMethods**—which the authors of this package recommend for any developer of **S4**-classes. For more details, see `?standardMethods`.

9 Odds and Ends

9.1 What should be done and what we could do —for version >1.9

- application of analytic FourierTransforms instead of FFT where appropriate —perhaps also to be controlled by a parameter/option
- use the `q`-slot applied to `runif` in `simplifyr` for continuous distributions
- further exact formulae for binary arithmetic operations like `"*"`
- redo the `math`-method for discrete distributions when only slot `r` is given
- generating function for new distribution classes to ease inclusion of new distributions
- goodness of fit tests for distribution-objects
- use of `\S4method` in documentation
- overloading binary operators of group `Math2` for independent distributions
- defining a subgroup of `Math2` of invertible binary operators

9.2 What should be done but for which we lack the know-how

- multivariate distributions
- conditional distributions
- copula

10 Acknowledgement

In order to give our acknowledgements their due place in the manual, we insert them before some rather extensive presentation of examples, because otherwise they would probably get lost or overseen by most of the readers.

We thank Martin Mächler and Josef Leydold for their helpful suggestions in conceiving the package. John Chambers also gave several helpful hints and insights when responding to our requests concerning the **S4**-class concept in **r-devel**/ **r-help**. We got stimulating replies to an RFC on **r-devel** by Duncan Murdoch and Gregory Warnes. We also thank Paul Gilbert for drawing our attention to his package **setRNG** and making it available as stand-alone version. In the last few days before the release on **CRAN**, Kurt Hornik and Uwe Ligges were very kind, helping us to find the clue how to pass all necessary checks by **R CMD check**. We also thank G. Jay Kerns for contributing code for the skewness and kurtosis functionals.

Last not least a big "thank you" to Torsten Hothorn as editor of **R-News**, for his patience with our endless versions until we finally came to a publishable version.

11 Examples

11.1 12-fold convolution of uniform $(0, 1)$ variables

Code also available under

[http://www.uni-bayreuth.de/departments/math/org/
/mathe7/DISTR/NormApprox.R](http://www.uni-bayreuth.de/departments/math/org/mathe7/DISTR/NormApprox.R)

This example shows how easily we may get the distribution of the sum of 12 i.i.d. $\text{ufo}(0, 1)$ -variables minus 6— which was used as a fast generator of $\mathcal{N}(0, 1)$ -variables in times when evaluations of \exp , \log , \sin and \tan were expensive, confer [7], example C, p. 163. The user should not be confused by expressions like $U+U$: this *does not* mean $2U$ but rather convolution of two independent identically distributed random variables.

```
> require(distr)

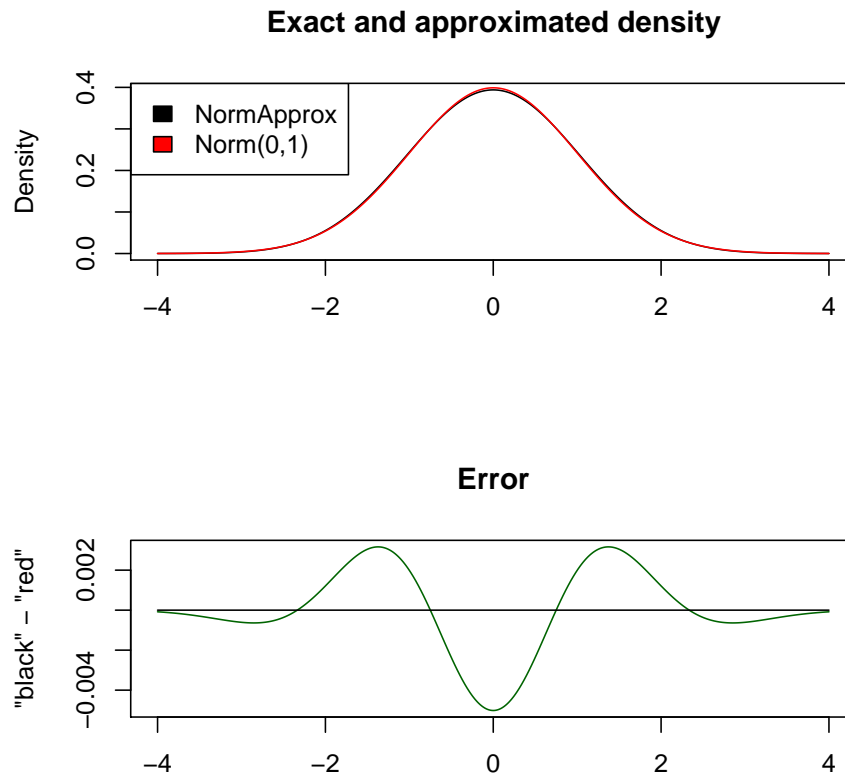
[1] TRUE

> N <- Norm(0,1)
> U <- Unif(0,1)
> U2 <- U + U
> U4 <- U2 + U2
> U8 <- U4 + U4
> U12 <- U4 + U8
> NormApprox <- U12 - 6
> x <- seq(-4,4,0.001)
```

```

> opar <- par()
> par(mfrow = c(2,1))
> plot(x, d(NormApprox)(x),
+      type = "l",
+      xlab = "",
+      ylab = "Density",
+      main = "Exact and approximated density")
> lines(x, d(N)(x),
+      col = "red")
> legend("topleft",
+      legend = c("NormApprox", "Norm(0,1)"),
+      fill = c("black", "red"))
> plot(x, d(NormApprox)(x) - d(N)(x),
+      type = "l",
+      xlab = "",
+      ylab = "\"black\" - \"red\"",
+      col = "darkgreen",
+      main = "Error")
> lines(c(-4,4), c(0,0))
> par(opar)

```



11.2 Comparison of exact convolution to FFT for normal distributions

Code also available under

[http://www.uni-bayreuth.de/departments/math/org/
/mathe7/DISTR/ConvolutionNormalDistr.R](http://www.uni-bayreuth.de/departments/math/org/mathe7/DISTR/ConvolutionNormalDistr.R)

This example illustrates the exactness of the numerical algorithm used to compute the convolution:
We know that $\mathcal{L}(A + B) = \mathcal{N}(5, 13)$ — if the second argument of \mathcal{N} is the variance

```
> require(distr)

[1] TRUE

> ## initialize two normal distributions
> A <- Norm(mean=1, sd=2)
> B <- Norm(mean=4, sd=3)
```



```

> ## convolution via addition of moments
> AB <- A+B
> ## casting of A,B as absolutely continuous distributions
> ## that is, ``forget'' that A,B are normal distributions
> A1 <- as(A, "AbscontDistribution")
> B1 <- as(B, "AbscontDistribution")
> ## for higher precision we change the global variable
> ## "TruncQuantile" from 1e-5 to 1e-8
> oldeps <- getdistrOption("TruncQuantile")
> eps <- 1e-8
> distroptions("TruncQuantile" = eps)
> ## support of A1+B1 for FFT convolution is
> ## [q(A1)(TruncQuantile),
> ## q(B1)(TruncQuantile, lower.tail = FALSE)]
>
> ## convolution via FFT
> AB1 <- A1+B1
> #####
> ## plots of the results
> #####
> par(mfrow=c(1,3))
> low <- q(AB)(1e-15)
> upp <- q(AB)(1e-15, lower.tail = FALSE)
> x <- seq(from = low, to = upp, length = 10000)
> ## densities
> plot(x, d(AB)(x), type = "l", lwd = 5)
> lines(x, d(AB1)(x), col = "orange", lwd = 1)
> title("Densities")
> legend("topleft", legend=c("exact", "FFT"),
+       fill=c("black", "orange"))
> ## cdfs
> plot(x, p(AB)(x), type = "l", lwd = 5)
> lines(x, p(AB1)(x), col = "orange", lwd = 1)
> title("CDFs")
> legend("topleft", legend=c("exact", "FFT"),
+       fill=c("black", "orange"))
> ## quantile functions
> x <- seq(from = eps, to = 1-eps, length = 1000)
> plot(x, q(AB)(x), type = "l", lwd = 5)
> lines(x, q(AB1)(x), col = "orange", lwd = 1)
> title("Quantile functions")

```

```

> legend("topleft", legend=c("exact", "FFT"),
+       fill=c("black", "orange"))
> ## Since the plots of the results show no
> ## recognizable differencies, we also compute
> ## the total variation distance of the densities
> ## and the Kolmogorov distance of the cdfs
>
> ## total variation distance of densities
> total.var <- function(z, N1, N2){
+   0.5*abs(d(N1)(z) - d(N2)(z))
+ }
> dv <- integrate(total.var, lower=-Inf, upper=Inf, rel.tol=1e-8, N1=AB, N2=AB1)
> cat("Total variation distance of densities:\t")

```

Total variation distance of densities:

```

> print(dv) # 4.25e-07

```

4.250016e-07 with absolute error < 1.8e-09

```

> ### meanwhile realized in package "distrEx"
> ### as TotalVarDist(N1,N2)
>
> ## Kolmogorov distance of cdfs
> ## the distance is evaluated on a random grid
> z <- r(Unif(Min=low, Max=upp))(1e5)
> dk <- max(abs(p(AB)(z)-p(AB1)(z)))
> cat("Kolmogorov distance of cdfs:\t", dk, "\n")

```

Kolmogorov distance of cdfs: 2.028446e-07

```

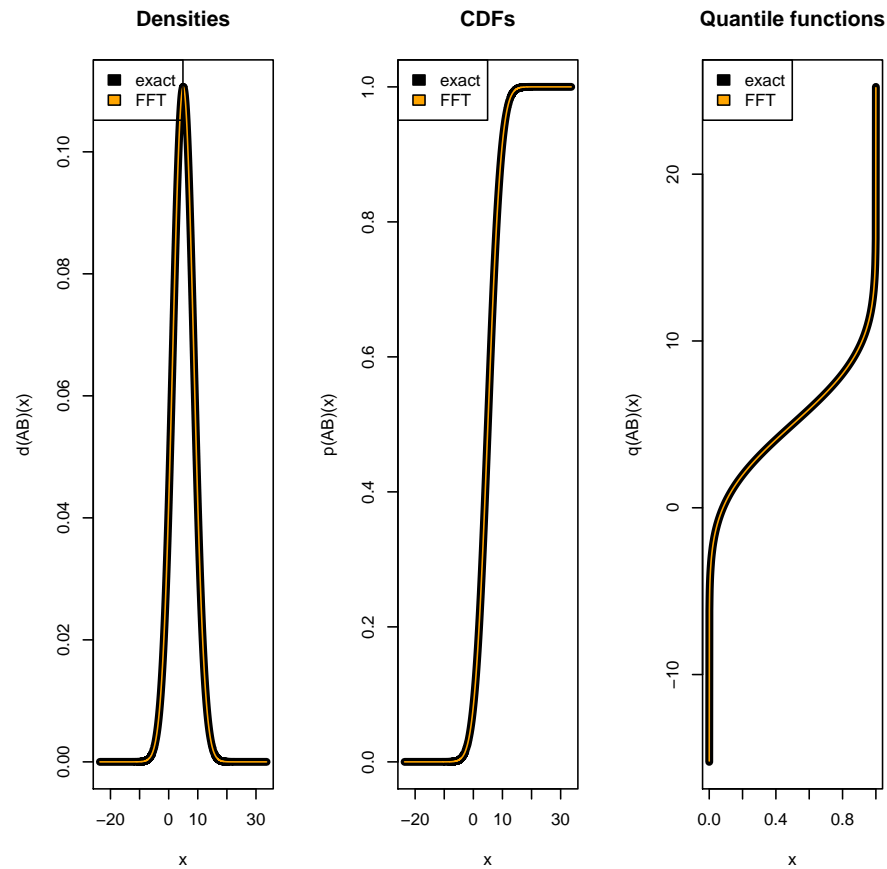
> # 2.03e-07
>

```

```

> ### meanwhile realized in package "distrEx"
> ### as KolmogorovDist(N1,N2)
>
> ## old distroptions
> distroptions("TruncQuantile" = oldeps)

```



11.3 Comparison of FFT to RtoDPQ

Code also available under

[http://www.uni-bayreuth.de/departments/math/org/
/mathe7/DISTR/ComparisonFFTandRtoDPQ.R](http://www.uni-bayreuth.de/departments/math/org/mathe7/DISTR/ComparisonFFTandRtoDPQ.R)

This example illustrates the exactness (or rather not-so-exactness) of the simulational default algorithm used to compute the distribution of transformations of group `math`.

```
> require(distr)

[1] TRUE

> #####
> ## Comparison 1 - FFT and RtoDPQ
> #####
```

```

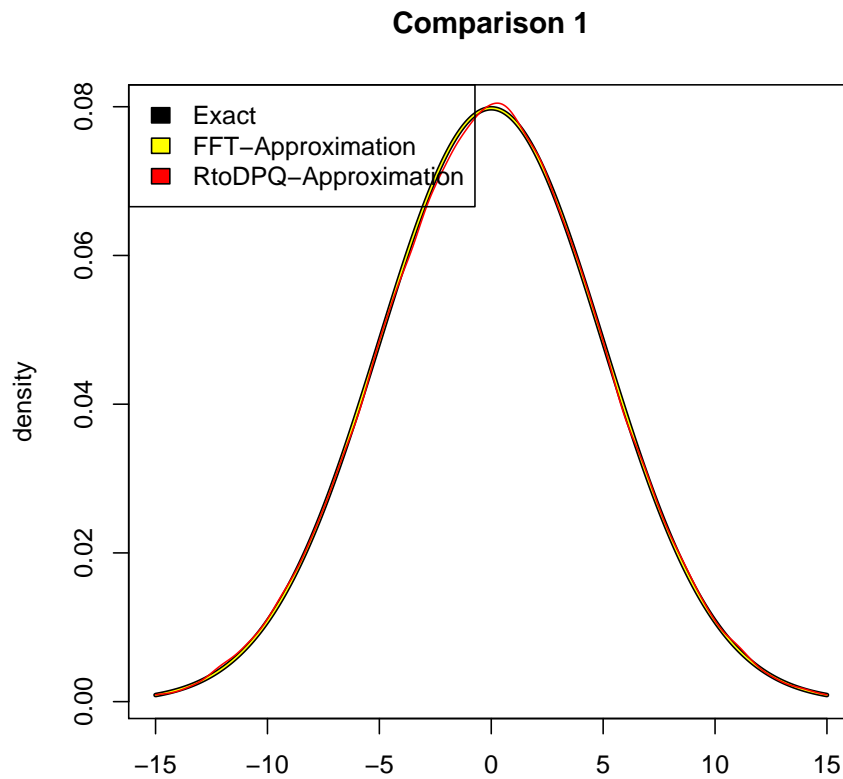
>
> N1 <- Norm(0,3)
> N2 <- Norm(0,4)
> rnew1 <- function(n) r(N1)(n) + r(N2)(n)
> X <- N1 + N2
>     # exact formula -> N(0,5)
> Y <- N1 + as(N2, "AbscontDistribution")
>     # appoximated with FFT
> Z <- new("AbscontDistribution", r = rnew1)
>     # appoximated with RtoDPQ
>
> # density-plot
>
> x <- seq(-15,15,0.01)
> plot(x, d(X)(x),
+      type = "l",
+      lwd = 3,
+      xlab = "",
+      ylab = "density",
+      main = "Comparison 1",
+      col = "black")
> lines(x, d(Y)(x),
+      col = "yellow")
> lines(x, d(Z)(x),
+      col = "red")
> legend("topleft",
+      legend = c("Exact", "FFT-Approximation",
+      "RtoDPQ-Approximation"),
+      fill = c("black", "yellow", "red"))
> #####
> ## Comparison 2 - "Exact" Formula and RtoDPQ
> #####
>
> B <- Binom(size = 6, prob = 0.5) * 10
> N <- Norm()
> rnew2 <- function(n) r(B)(n) + r(N)(n)
> Y <- B + N
>     # "exact" formula
> Z <- new("AbscontDistribution", r = rnew2)
>     # appoximated with RtoDPQ
>

```

```

> # density-plot
>
> x <- seq(-5,65,0.01)
> plot(x, d(Y)(x),
+      type = "l",
+      xlab = "",
+      ylab = "density",
+      main = "Comparison 2",
+      col = "black")
> lines(x, d(Z)(x),
+      col = "red")
> legend("topleft",
+      legend = c("Exact", "RtoDQP-Approximation"),
+      fill = c("black", "red"))

```



11.4 Comparison of exact and approximate stationary regressor distribution

Code also available under

[http://www.uni-bayreuth.de/departments/math/org/
/mathe7/DISTR/StationaryRegressorDistr.R](http://www.uni-bayreuth.de/departments/math/org/mathe7/DISTR/StationaryRegressorDistr.R)

Another illustration for the use of package "distr". In case of a stationary AR(1)-model, for non-normal innovation distribution, the stationary distribution of the observations must be approximated by finite convolutions. That these approximations give fairly good results for approximations down to small orders is exemplified by the Gaussian case where we may compare the approximation to the exact stationary distribution.

```
> require(distr)

[1] TRUE

> ## Approximation of the stationary regressor
> ## distribution of an AR(1) process
> ##       $X_t = \phi X_{t-1} + V_t$ 
> ## where  $V_t$  i.i.d  $N(0,1)$  and  $\phi \in (0,1)$ 
> ## We obtain
> ##       $X_t = \sum_{j=1}^{\infty} \phi^j V_{t-j}$ 
> ## i.e.,  $X_t \sim N(0, 1/(1-\phi^2))$ 
> phi <- 0.5
> ## casting of V as absolutely continuous distributions
> ## that is, ``forget'' that V is a normal distribution
> V <- as(Norm(), "AbscontDistribution")
> ## for higher precision we change the global variable
> ## "TruncQuantile" from 1e-5 to 1e-8
> oldeps <- getdistrOption("TruncQuantile")
> eps <- 1e-8
> distroptions("TruncQuantile" = eps)
> ## Computation of the approximation
> ##       $H = \sum_{j=1}^n \phi^j V_{t-j}$ 
> ## of the stationary regressor distribution
> ## (via convolution using FFT)
> H <- V
> n <- 15
> ## may take some time
> for(i in 1:n){Vi <- phi^i*V; H <- H + Vi }
> ## the stationary regressor distribution (exact)
> X <- Norm(sd=sqrt(1/(1-phi^2)))
```

```

> #####
> ## plots of the results
> #####
> par(mfrow=c(1,3))
> low <- q(X)(1e-15)
> upp <- q(X)(1e-15, lower.tail = FALSE)
> x <- seq(from = low, to = upp, length = 10000)
> ## densities
> plot(x, d(X)(x),type = "l", lwd = 5)
> lines(x , d(H)(x), col = "orange", lwd = 1)
> title("Densities")
> legend("topleft", legend=c("exact", "FFT"),
+       fill=c("black", "orange"))
> ## cdfs
> plot(x, p(X)(x),type = "l", lwd = 5)
> lines(x , p(H)(x), col = "orange", lwd = 1)
> title("CDFs")
> legend("topleft", legend=c("exact", "FFT"),
+       fill=c("black", "orange"))
> ## quantile functions
> x <- seq(from = eps, to = 1-eps, length = 1000)
> plot(x, q(X)(x),type = "l", lwd = 5)
> lines(x , q(H)(x), col = "orange", lwd = 1)
> title("Quantile functions")
> legend( "topleft",
+       legend=c("exact", "FFT"),
+       fill=c("black", "orange"))
> ## Since the plots of the results show no
> ## recognizable differencies, we also compute
> ## the total variation distance of the densities
> ## and the Kolmogorov distance of the cdfs
>
> ## total variation distance of densities
> total.var <- function(z, N1, N2){
+   0.5*abs(d(N1)(z) - d(N2)(z))
+ }
> dv <- integrate(f = total.var, lower = -Inf,
+               upper = Inf, rel.tol = 1e-5,
+               N1=X, N2=H)
> cat("Total variation distance of densities:\t")

```

Total variation distance of densities:

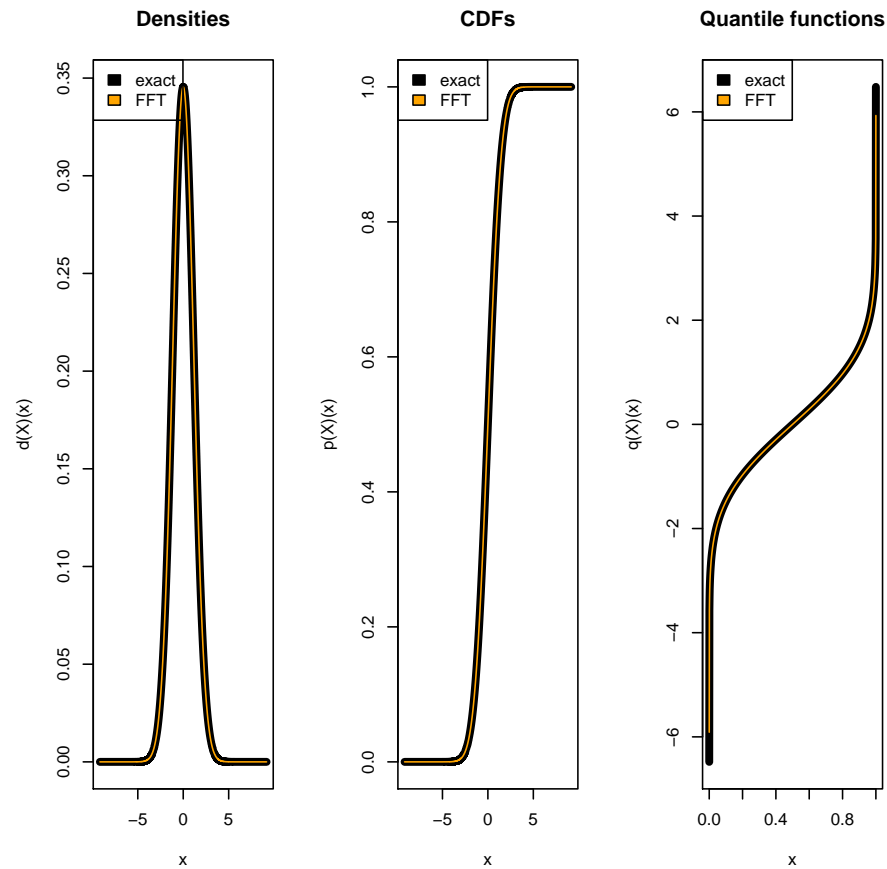
```
> print(dv) # <1.0e-05
```

6.552639e-05 with absolute error < 5.5e-06

```
> ### meanwhile realized in package "distrEx"
> ### as TotalVarDist(N1,N2)
>
>
> ## Kolmogorov distance of cdfs
> ## the distance is evaluated on a random grid
> z <- r(Unif(Min=low, Max=upp))(1e5)
> dk <- max(abs(p(X)(z)-p(H)(z)))
> cat("Kolmogorov distance of cdfs:\t", dk, "\n")
```

Kolmogorov distance of cdfs: 6.538083e-05

```
> # ~5e-06
>
> ### meanwhile realized in package "distrEx"
> ### as KolmogorovDist(N1,N2)
>
>
> ## old distroptions
> distroptions("TruncQuantile" = oldeps)
```

11.5 Truncation and Huberization/winsorization

Code also available under

<http://www.uni-bayreuth.de/departments/math/org/mathe7/DISTR/huberize.R>

<http://www.uni-bayreuth.de/departments/math/org/mathe7/DISTR/truncate.R>

The operations of truncation and Huberization play a crucial role in Robust Statistics, but also arise in many other contexts like censoring etc; they may now be formulated quite generally as shown in this example. With the slots `d`, `p` and `q` of class `UnivariateDistribution` being `OptionalFunction` from version 1.4 on, it would be no problem to return a corresponding distribution object now.

```
> require(distr)
```

```
[1] TRUE
```

```

> if(!isGeneric("Huberize"))
+   setGeneric("Huberize",
+     function(object, lower, upper)
+       standardGeneric("Huberize")
+     )
[1] "Huberize"

> setMethod("Huberize",
+   signature(object = "AbscontDistribution",
+     lower = "numeric", upper = "numeric"),
+   function(object, lower, upper){
+     ## new random number function
+     rnew = function(n){
+       rn = r(object)(n)
+       ifelse(rn < lower, lower,
+         ifelse(rn >= upper, upper, rn))
+     }
+
+     ## new cdf
+     pnew = function(x)
+       ifelse(x < lower, 0,
+         ifelse(x >= upper, 1, p(object)(x)))
+
+     ## new quantile function
+     plower = p(object)(lower)
+     pupper = p(object)(upper)
+     qnew = function(x)
+       ifelse(x < plower,
+         ifelse(x < 0, NA, -Inf),
+         ifelse(x >= pupper,
+           ifelse(x > 1, NA, upper),
+           q(object)(x)))
+
+     new("UnivariateDistribution",
+       r = rnew, p = pnew, q = qnew, d = NULL)
+   })
[1] "Huberize"

> # Example
> # Normal(0,1)-Distribution huberized at -0.5 and 1

```

```

> N = Norm()
> HN = Huberize(N, -0.5, 1)
> # some huberized randomnumbers
> r(HN)(10)

[1] 0.2322964 0.6596313 -0.5000000 1.0000000 1.0000000 0.2558002
[7] 0.4347652 0.9595039 0.2010964 -0.5000000

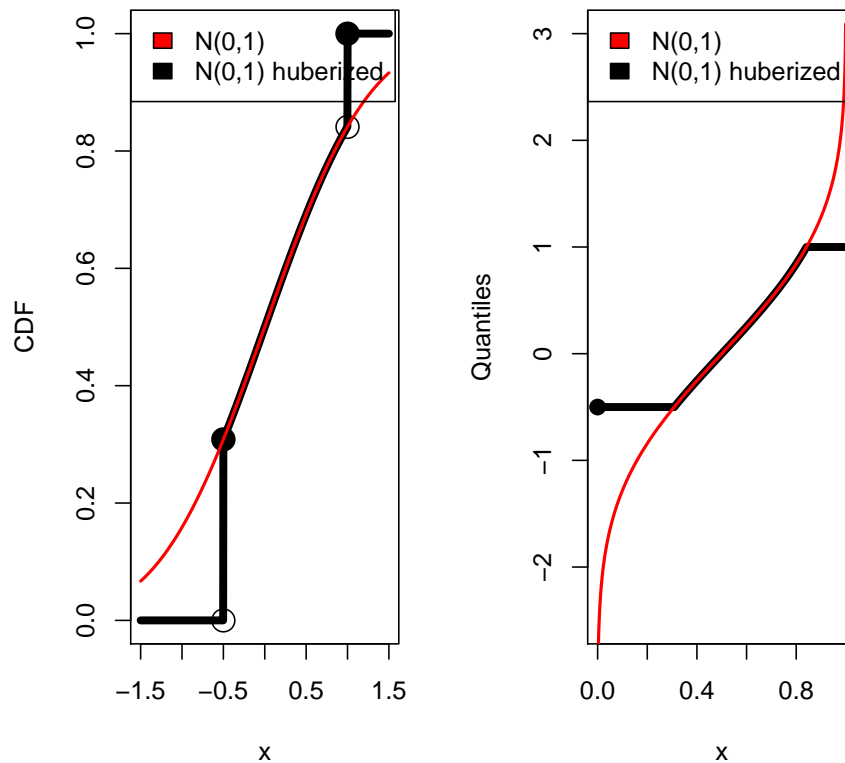
> ## plot is not (yet) available for class
> ## UnivariateDistribution, i.e. distrib's
> ## which are neither a.c. nor discrete
> ## (here HN is a mixture of a.c. and discrete
> ## distributions)
>
> # cdf of huberized Normal-Distribution
> # and of Normal-Distribution
> oldpar = par()
> par(mfrow = c(1,2))
> x = seq(-1.5, 1.5, length = 1000)
> plot(x, p(HN)(x),
+      type = "l",
+      lwd = 5,
+      ylab = "CDF")
> points(c(-0.5,1),p(HN)(c(-0.5,1)),pch=16, cex=2)
> points(c(-0.5,1),c(0,p(HN)(0.999)),pch=21,cex=2)
> lines(x, p(N)(x),
+      lwd = 2,
+      col = "red")
> legend("topleft",
+      legend = c("N(0,1)", "N(0,1) huberized"),
+      fill = c("red", "black"))
> # quantile functions
> x = seq(0, 1, length = 1000)
> plot(x, q(HN)(x),
+      type = "l",
+      lwd = 5,
+      ylab = "Quantiles",
+      ylim = c(-2.5,3))
> lines(c(0,p(HN)(-0.5)), -c(1,1)*0.5, lwd = 5)
> points(0,-0.5,pch=20,cex=2)
> lines(x, q(N)(x),

```

```

+       lwd = 2,
+       col = "red")
> legend("topleft",
+       legend = c("N(0,1)", "N(0,1) huberized"),
+       fill = c("red", "black"))
> par(oldpar)

```



```

> require(distr)

[1] TRUE

> if(!isGeneric("Truncate"))
+   setGeneric("Truncate",
+   function(object, lower, upper)
+     standardGeneric("Truncate")
+   )

```

```
[1] "Truncate"
```

```
> setMethod("Truncate",
+           signature(object = "AbscontDistribution",
+                     lower = "numeric", upper = "numeric"),
+           function(object, lower, upper){
+             ## new random number function
+             rnew = function(n){
+               rn = r(object)(n)
+               while(TRUE){
+                 rn[rn < lower] = NA
+                 rn[rn > upper] = NA
+                 index = is.na(rn)
+                 if(!any(index)) break
+                 rn[index] = r(object)(sum(index))
+               }
+               rn
+             }
+
+             ## new cdf
+             plower = p(object)(lower)
+             pupper = p(object)(upper)
+             pnew = function(x)
+               ifelse(x < lower, 0,
+                     ifelse(x >= upper, 1,
+                           (p(object)(x) - plower)/(pupper - plower)))
+
+             ## new density
+             lostmass = plower + 1 - pupper
+
+             dnew = function(x)
+               ifelse(x < lower, 0,
+                     ifelse(x >= upper, 0,
+                           d(object)(x)/(1-lostmass)))
+
+             # new quantile
+             qfun1 <- function(x){
+               if(x == 0) return(lower)
+               if(x == 1) return(upper)
+               fun <- function(t) pnew(t) - x
+             }
+           }
```

```

+         uniroot(f = fun, interval = c(lower, upper))$root
+     }
+     qfun2 <- function(x)
+         sapply(x, qfun1)
+
+     return(new("AbscontDistribution", r = rnew,
+         d = dnew, p = pnew, q = qfun2))
+ })

[1] "Truncate"

> # Example
> # Normal(0,1)-Distribution truncated at -0.5 and 1
>
> N = Norm()
> Z = Truncate(N, -0.5, 1)
> # the truncated Distribution
>
> plot(Z, lwd=5, withSweave = TRUE)
> # some truncated randomnumbers
>
> r(Z)(10)

[1] -0.3951692  0.9185585  0.6478132  0.1829098  0.8735308  0.8754917
[7]  0.3840936  0.4335924 -0.2598434  0.4525272

> oldpar = par()
> par(mfrow = c(1,2))
> # cdf of truncated Normal-Distribution
> # and of Normal-Distribution
>
> x = seq(-1.5, 1.5, length = 1000)
> plot(x, p(Z)(x),
+     type = "l",
+     lwd = 5,
+     xlab = "",
+     ylab = "CDF")
> lines(x, p(N)(x),
+     lwd = 2,
+     col = "red")
> legend("topleft",
+     legend = c("N(0,1)", "N(0,1) truncated"),

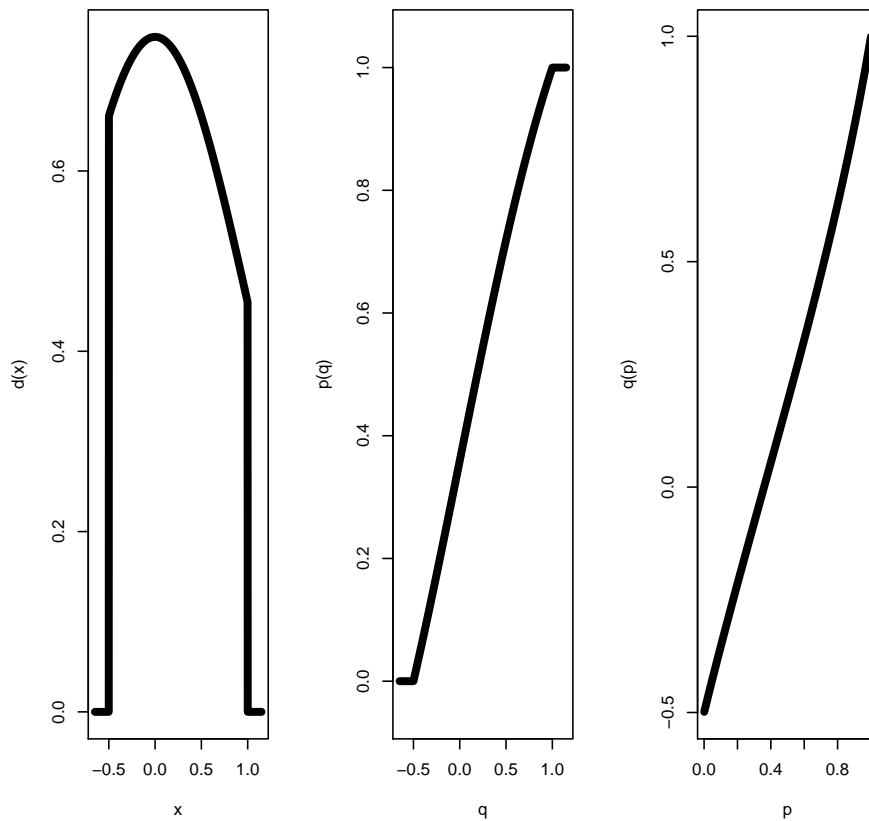
```

```

+         fill = c("red", "black"))
> # density of truncated Normal-Distribution
> # and of Normal-Distribution
>
> x = seq(-1.5, 1.5, length = 1000)
> plot(x, d(Z)(x),
+       type = "l",
+       lwd = 5,
+       xlab = "",
+       ylab = "density")
> lines(x, d(N)(x),
+       lwd = 2,
+       col = "red")
> legend("topleft", legend = c("N(0,1)", "N(0,1) truncated"),
+       fill = c("red", "black"))
> par(oldpar)

```

Density of AbscontDistributio CDF of AbscontDistributiontile function of AbscontDistri



11.6 Distribution of minimum and maximum of two independent random variables

Code also available under

<http://www.uni-bayreuth.de/departments/math/org/mathe7/DISTR/minandmax.R>

As in the preceding example, we illustrate the use of package "distr" by making available widely necessary operations: Minimum and maximum of two independent random variables.

```
> require(distr)

[1] TRUE

> if(!isGeneric("Minimum"))
+   setGeneric("Minimum",
+   function(e1, e2) standardGeneric("Minimum"))
```



```
[1] "Minimum"
```

```
> setMethod("Minimum",
+           signature(e1 = "AbscontDistribution",
+                     e2 = "AbscontDistribution"),
+           function(e1, e2){
+             ## new random number function
+             rnew <- function(n){
+               rn1 <- r(e1)(n)
+               rn2 <- r(e2)(n)
+
+               ifelse(rn1 < rn2, rn1, rn2)
+             }
+
+             ## new cdf
+             pnew <- function(x){
+               p1 <- p(e1)(x)
+               p2 <- p(e2)(x)
+               p1 + p2 - p1 * p2
+             }
+
+             ## new density
+             dnew <- function(x){
+               d1 <- d(e1)(x)
+               d2 <- d(e2)(x)
+               p1 <- p(e1)(x)
+               p2 <- p(e2)(x)
+               d1 + d2 - d1 * p2 - p1 * d2
+             }
+
+             ## new quantile function
+             lower1 <- q(e1)(0)
+             lower2 <- q(e2)(0)
+             upper1 <- q(e1)(1)
+             upper2 <- q(e2)(1)
+             lower <- min(lower1, lower2)
+             upper <- min(upper1, upper2)
+
+             maxquantile = min(q(e1)(1e-6, lower.tail = FALSE),
+                               q(e2)(1e-6, lower.tail = FALSE))
+             minquantile = min(q(e1)(1e-6),
```

```

+               q(e2)(1e-6))
+
+       qfun1 <- function(x){
+         if(x == 0) return(lower)
+         if(x == 1) return(upper)
+         fun <- function(t) pnnew(t) - x
+         uniroot(f = fun,
+               interval = c(maxquantile,
+                           minquantile))$root
+       }
+       qfun2 <- function(x)
+         sapply(x, qfun1)
+
+       return(new("AbscontDistribution", r = rnnew,
+             d = dnnew, p = pnnew, q = qfun2))
+     })

[1] "Minimum"

> if(!isGeneric("Maximum")) setGeneric("Maximum",
+   function(e1, e2) standardGeneric("Maximum"))

[1] "Maximum"

> setMethod("Maximum",
+   signature(e1 = "AbscontDistribution",
+             e2 = "AbscontDistribution"),
+   function(e1, e2){
+     ## new random number function
+     rnnew <- function(n){
+       rn1 <- r(e1)(n)
+       rn2 <- r(e2)(n)
+
+       ifelse(rn1 > rn2, rn1, rn2)
+     }
+
+     ## new cdf
+     pnnew <- function(x){
+       p1 <- p(e1)(x)
+       p2 <- p(e2)(x)
+       p1 * p2
+     }

```

```

+
+     ## new density
+     dnew <- function(x){
+         d1 <- d(e1)(x)
+         d2 <- d(e2)(x)
+         p1 <- p(e1)(x)
+         p2 <- p(e2)(x)
+         d1 * p2 + p1 * d2
+     }
+
+     ## new quantile function
+     lower1 <- q(e1)(0)
+     lower2 <- q(e2)(0)
+     upper1 <- q(e1)(1)
+     upper2 <- q(e2)(1)
+     lower <- max(lower1, lower2)
+     upper <- max(upper1, upper2)
+
+     maxquantile = max(q(e1)(1e-6, lower.tail = FALSE),
+                       q(e2)(1e-6, lower.tail = FALSE))
+
+     minquantile = max(q(e1)(1e-6), q(e2)(1e-6))
+     print(c(minquantile,maxquantile))
+
+     qfun1 <- function(x){
+         if(x == 0) return(lower)
+         if(x == 1) return(upper)
+         fun <- function(t) pnnew(t) - x
+         uniroot(f = fun, interval = c(maxquantile,
+                                       minquantile))$root
+     }
+     qfun2 <- function(x)
+         sapply(x, qfun1)
+
+     return(new("AbscontDistribution", r = rnew,
+         d = dnew, p = pnnew, q = qfun2))
+ })
+
+ [1] "Maximum"
+
+ > # Example
+ >

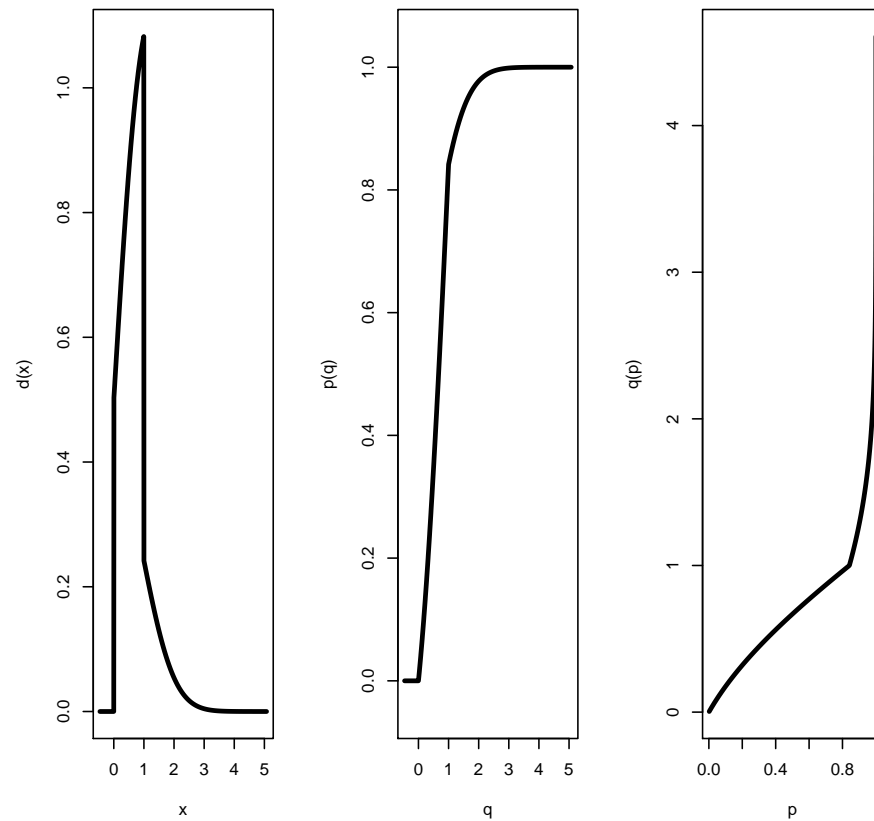
```

```
> N <- Norm(mean = 0, sd = 1)
> U <- Unif(Min = 0, Max = 1)
> Y <- Maximum(N, U)
```

```
[1] 0.000001 4.753424
```

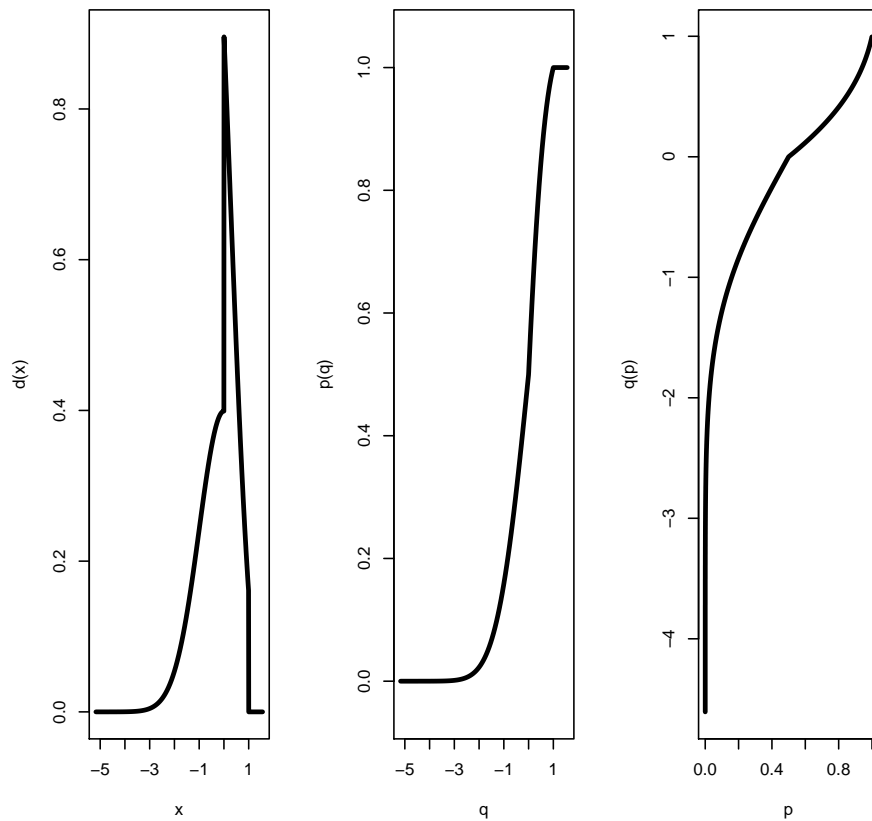
```
> distroptions("TruncQuantile"=1e-6)
> plot(Y, lwd = 3, withSweave = TRUE)
```

Density of AbscontDistributic CDF of AbscontDistributionntile function of AbscontDistri



```
> Z <- Minimum(N,U)
> plot(Z, lwd = 3, withSweave = TRUE)
```

Density of AbscontDistributio CDF of AbscontDistributionntile function of AbscontDistri



11.7 Instructive destructive example

Code also available under

<http://www.uni-bayreuth.de/departments/math/org/mathe7/DISTR/destructive.R>

```
> #####
> ## Demo: Instructive destructive example
> #####
> require(distr)

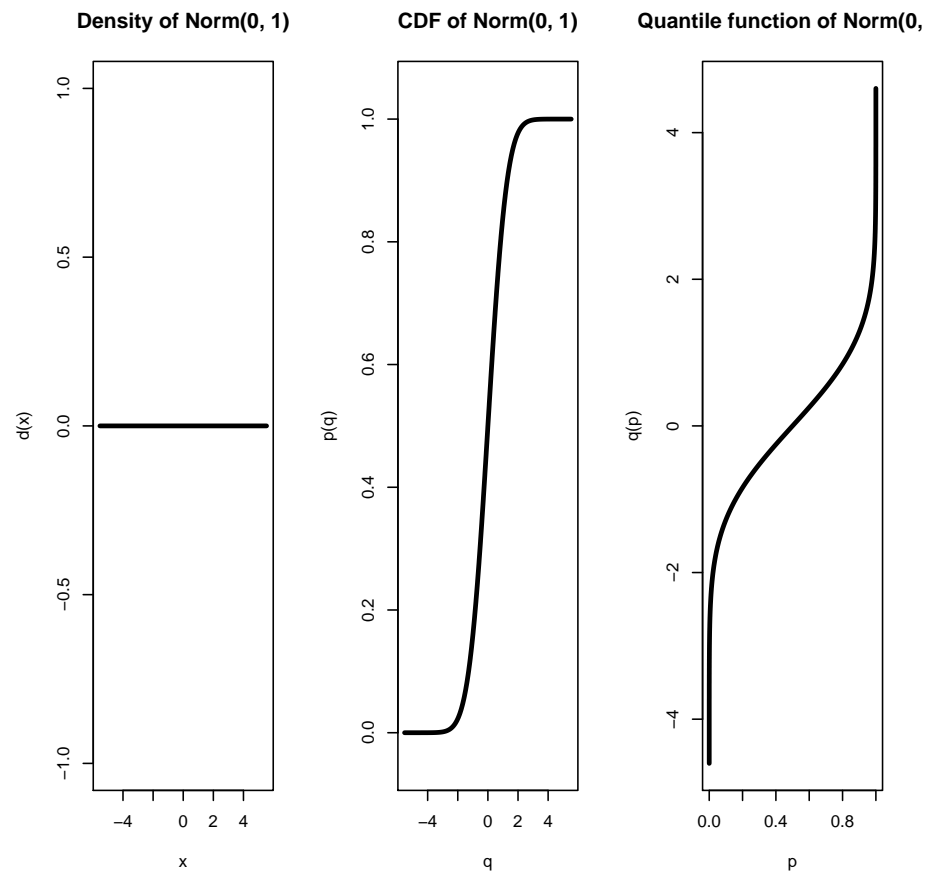
[1] TRUE

> ## package "distr" encourages
> ## consistency but does not
```

```

> ## enforce it---so in general
> ## d o n o t m o d i f y
> ## slots d,p,q,r!
>
> N <- Norm()
> B <- Binom()
> N@d <- B@d
> plot(N, lwd = 3, withSweave = TRUE)

```



11.8 A simulation example

needs packages "distrSim"/"distrTEst"

Code also available under

[http://www.uni-bayreuth.de/departments/math/org/
/mathe7/DISTR/SimulateandEstimate.R](http://www.uni-bayreuth.de/departments/math/org/mathe7/DISTR/SimulateandEstimate.R)

```

> have.distrTEst <- suppressWarnings(require(distrTEst))
>   ### also loads distrSim
> if (have.distrTEst)
+   { sim <- new("Simulation",
+               seed = setRNG(),
+               distribution = Norm(mean = 0, sd = 1),
+               filename="sim_01",
+               runs = 1000,
+               samplesize = 30)
+
+   contsim <- new("Contsimulation",
+                 seed = setRNG(),
+                 distribution.id = Norm(mean = 0, sd = 1),
+                 distribution.c = Norm(mean = 0, sd = 9),
+                 rate = 0.1,
+                 filename="contsim_01",
+                 runs = 1000,
+                 samplesize = 30)
+
+   simulate(sim)
+   simulate(contsim)
+
+   print(sim)
+   summary(contsim)
+   plot(contsim)
+ } else {
+   cat("\n functionality not (yet) available; ")
+   cat("you have to install package \"distrTEst\" first.\n")
+ }

filename of Simulation: sim_01
Seed: Kind: Mersenne-Twister
      Normal Kind: Inversion
      first 6 numbers: 0188430748      0971694936      -0330153779
                      -0302019877      0200422946      -1503041535

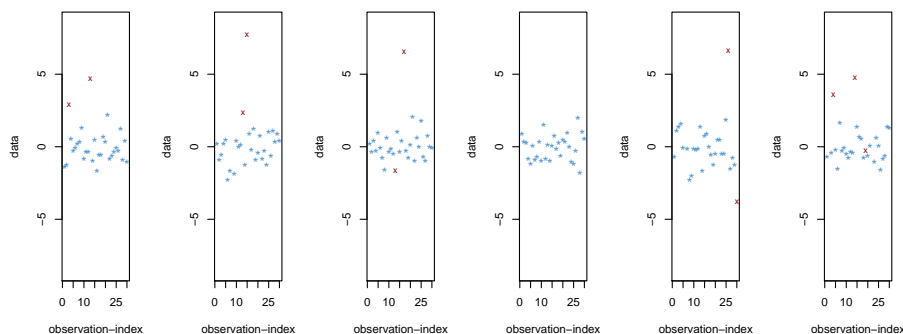
number of runs: 1000
dimension of the observations: 1
size of sample: 30
object was generated by version: 1.9
Distribution:

```

```

Distribution Object of Class: Norm
mean : 0
sd : 1
name of simulation: contsim_01
rate of contamination: 0.100000
real Data:
dimension of the observations: 1
number of runs: 1000
size of sample: 30

```



```

> have.distrTEst <- suppressWarnings(require("distrTEst"))
> if (have.distrTEst)
+   { psim <- function(theta,y,m0){
+     mean(pmin(pmax(-m0, y - theta), m0))
+   }
+   mestimator <- function(x, m = 0.7) {
+     uniroot(f = psim,
+       lower = -20,
+       upper = 20,
+       tol = 1e-10,
+       y = x,
+       m0 = m,
+       maxiter = 20)$root
+   }
+
+   result.id.mean <- evaluate(sim, mean)
+   result.id.mest <- evaluate(sim, mestimator)
+   result.id.median <- evaluate(sim, median)
+

```



```

+
+   result.cont.mean <- evaluate(contsim, mean)
+   result.cont.mest <- evaluate(contsim, mestimator)
+   result.cont.median <- evaluate(contsim, median)
+
+   elist <- EvaluationList(result.cont.mean,
+                           result.cont.mest,
+                           result.cont.median)
+
+   print(elist)
+   summary(elist)
+   plot(elist, cex = 0.7, las = 2)
+ } else {
+   cat("\n functionality not (yet) available; ")
+   cat("you have to install package \"distrTEst\" first.\n")
+ }

```

An EvaluationList Object

name of Evaluation List: a list of "Evaluation" objects

name of Dataobject: object

name of Datafile: contsim_01

An Evaluation Object

estimator: mean

Result: 'data.frame': 1000 obs. of 2 variables:

\$ mean.id: num -0.2571 -0.1581 -0.0351 -0.0396 -0.1863 ...

\$ mean.re: num 0.1007 0.6255 0.2028 -0.0396 0.4015 ...

An Evaluation Object

estimator: mestimator

Result: 'data.frame': 1000 obs. of 2 variables:

\$ mstm.id: num -0.2933 -0.0780 -0.0963 -0.0633 -0.2114 ...

\$ mstm.re: num -0.1524 0.0622 -0.0255 -0.0633 -0.1276 ...

An Evaluation Object

estimator: median

Result: 'data.frame': 1000 obs. of 2 variables:

\$ medn.id: num -0.3372 0.0169 -0.1058 0.0671 -0.1607 ...

\$ medn.re: num -0.2521 0.1521 -0.0627 0.0671 -0.1461 ...

name of Evaluation List: a list of "Evaluation" objects

name of Dataobject: object

name of Datafile: contsim_01

name of Evaluation: object

estimator: mean

Result:

mean.id	mean.re
Min. : -0.602897	Min. : -1.951351
1st Qu.: -0.140631	1st Qu.: -0.314385
Median : -0.004469	Median : -0.004854
Mean : -0.010057	Mean : 0.004532
3rd Qu.: 0.115294	3rd Qu.: 0.335325
Max. : 0.493840	Max. : 1.904512

name of Evaluation: object

estimator: mestimator

Result:

mstm.id	mstm.re
Min. : -0.64719	Min. : -0.653427
1st Qu.: -0.15283	1st Qu.: -0.160285
Median : -0.01076	Median : -0.005312
Mean : -0.01220	Mean : -0.011271
3rd Qu.: 0.13080	3rd Qu.: 0.148967
Max. : 0.55204	Max. : 0.552036

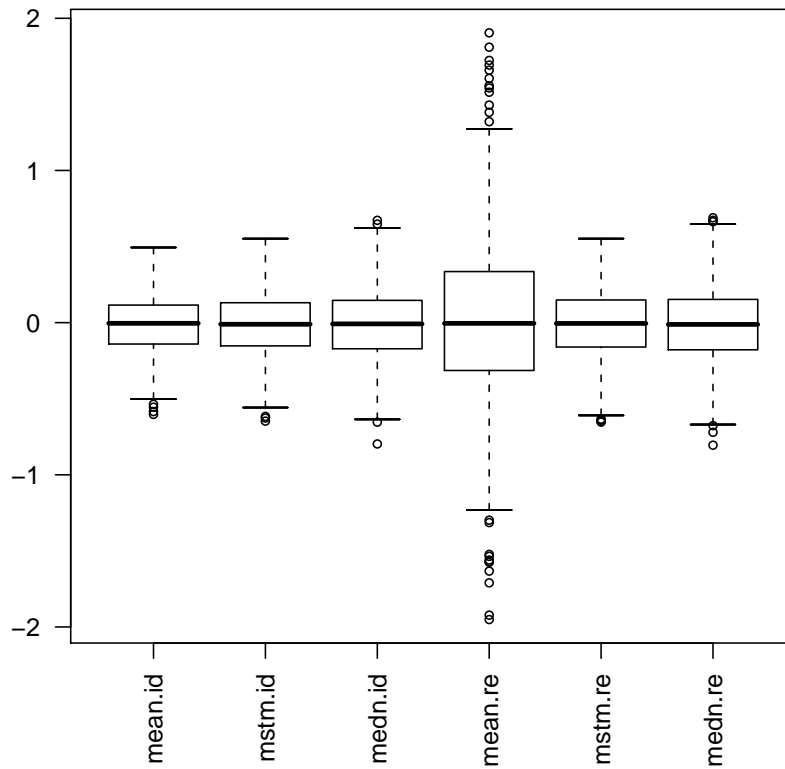
name of Evaluation: object

estimator: median

Result:

medn.id	medn.re
Min. : -0.797128	Min. : -0.80534
1st Qu.: -0.171668	1st Qu.: -0.17828
Median : -0.008259	Median : -0.01176
Mean : -0.015120	Mean : -0.01360
3rd Qu.: 0.146195	3rd Qu.: 0.15249
Max. : 0.672417	Max. : 0.68941

1. coordinate



Output by `plot/show`-method for an object of class `Evaluation`

```
> result.cont.mest
```

An Evaluation Object

name of Dataobject: object

name of Datafile: contsim_01

estimator: mestimator

Result: 'data.frame': 1000 obs. of 2 variables:

```
$ mstm.id: num -0.2933 -0.0780 -0.0963 -0.0633 -0.2114 ...
```

```
$ mstm.re: num -0.1524 0.0622 -0.0255 -0.0633 -0.1276 ...
```

Output by `summary`-method for an object of class `EvaluationList`

```
> summary(elist)
```

name of Evaluation List: a list of "Evaluation" objects

name of Dataobject: object

name of Datafile: contsim_01

name of Evaluation: object

estimator: mean

```
Result:
      mean.id      mean.re
Min.   :-0.602897 Min.   :-1.951351
1st Qu.: -0.140631 1st Qu.: -0.314385
Median :-0.004469 Median :-0.004854
Mean   :-0.010057 Mean    : 0.004532
3rd Qu.: 0.115294 3rd Qu.: 0.335325
Max.    : 0.493840 Max.    : 1.904512
-----
```

```
name of Evaluation: object
estimator: mestimator
```

```
Result:
      mstm.id      mstm.re
Min.   :-0.64719 Min.   :-0.653427
1st Qu.: -0.15283 1st Qu.: -0.160285
Median :-0.01076 Median :-0.005312
Mean   :-0.01220 Mean    :-0.011271
3rd Qu.: 0.13080 3rd Qu.: 0.148967
Max.    : 0.55204 Max.    : 0.552036
-----
```

```
name of Evaluation: object
estimator: median
```

```
Result:
      medn.id      medn.re
Min.   :-0.797128 Min.   :-0.80534
1st Qu.: -0.171668 1st Qu.: -0.17828
Median :-0.008259 Median :-0.01176
Mean   :-0.015120 Mean    :-0.01360
3rd Qu.: 0.146195 3rd Qu.: 0.15249
Max.    : 0.672417 Max.    : 0.68941
```

In this example we present a standard robust simulation study that — in variations — arises in almost every paper on Robust Statistics. We do this with the tools provided by our package...

11.9 Expectation of a given function under a given distribution

Code also available under

[http://www.uni-bayreuth.de/departments/math/org/
/mathe7/DISTR/Expectation.R](http://www.uni-bayreuth.de/departments/math/org/mathe7/DISTR/Expectation.R)

This code is for illustration only; in the mean-time, the expectation- and variance operators implemented in this example have been included to package "distrEx" where their functionality has further been extended. As in examples 11.5 and 11.6, we illustrate the use of package "distr" by implementing a general evaluation of expectation and variance under a given distribution.

```
> have.distrEx <- suppressWarnings(require("distrEx"))
> if (have.distrEx)
+   {
+     # Example
```

```

+     id <- function(x) x
+     sq <- function(x) x^2
+
+     # Expectation and Variance of Binom(6,0.5)
+     B <- Binom(6, 0.5)
+     print(E(B, id))
+     print(E(B, sq) - E(B, id)^2)
+
+     # Expectation and Variance of Norm(1,1)
+     N <- Norm(1, 1)
+     print(E(N, id))
+     print(E(N, sq) - E(N, id)^2)
+   } else {
+     cat("\n functionality not (yet) available; ")
+     cat("you have to install package \"distrEx\" first.\n")
+   }

```

[1] 3
 [1] 1.5
 [1] 0.9999998
 [1] 0.9999944

11.10 n -fold convolution of absolutely continuous distributions

Code also available under

<http://www.uni-bayreuth.de/departments/math/org/mathe7/DISTR/nFoldConvolution.R>

Might be useful for teaching the CLT: a straightforward implementation of the n -fold convolution of an arbitrary implemented absolutely continuous distribution — to show accuracy of our method we compare it to the exact formula valid for n -fold convolution of normal distributions. From version 1.9 this is integrated to package "distr".

```

> #####
> ## Demo: n-fold convolution of absolutely continuous
> ##      probability distributions
> #####
> require(distr)

```

[1] TRUE

```

> if(!isGeneric("convpow"))
+   setGeneric("convpow",

```

```

+     function(D1,...) standardGeneric("convpow"))
> #####
> ## Function for n-fold convolution
> ## -- absolute continuous distribution --
> #####
>
> ##implentation of Algorithm 3.4. of
> # Kohl, M., Ruckdeschel, P., Stabla, T. (2005):
> #   General purpose convolution algorithm for distributions
> #   in S4-Classes by means of FFT.
> # Technical report, Feb. 2005. Also available in
> # http://www.uni-bayreuth.de/departments/math/org/mathe7/
> #   /RUCKDESCHEL/pubs/comp.pdf
>
>
> setMethod("convpow",
+           signature(D1 = "AbscontDistribution"),
+           function(D1, N){
+             if((N < 1)||(!identical(floor(N), N)))
+               stop("N has to be a natural greater than 0")
+
+             m <- getdistrOption("DefaultNrFFTGridPointsExponent")
+
+             ##STEP 1
+
+             lower <- ifelse((q(D1)(0) > - Inf), q(D1)(0),
+                             q(D1)(getdistrOption("TruncQuantile")))
+             upper <- ifelse((q(D1)(1) < Inf), q(D1)(1),
+                             q(D1)(getdistrOption("TruncQuantile"), lower.tail = FALSE))
+
+             ##STEP 2
+
+             M <- 2^m
+             h <- (upper-lower)/M
+             if(h > 0.01)
+               warning(paste("Grid for approxfun too wide, ",
+                             "increase DefaultNrFFTGridPointsExponent", sep=""))
+             x <- seq(from = lower, to = upper, by = h)
+             p1 <- p(D1)(x)
+
+             ##STEP 3

```

```

+
+       p1 <- p1[2:(M + 1)] - p1[1:M]
+
+ ##STEP 4
+
+       ## computation of DFT
+       pn <- c(p1, numeric((N-1)*M))
+       fftpn <- fft(pn)
+
+ ##STEP 5
+
+       ## convolution theorem for DFTs
+       pn <- Re(fft(fftpn^N, inverse = TRUE)) / (N*M)
+       pn <- (abs(pn) >= .Machine$double.eps)*pn
+       i.max <- N*M-(N-2)
+       pn <- c(0,pn[1:i.max])
+       dn <- pn / h
+       pn <- cumsum(pn)
+
+ ##STEP 6(density)
+
+       ## density
+       x <- c(N*lower,seq(from = N*lower+N/2*h,
+           to = N*upper-N/2*h, by=h),N*upper)
+       dnfun1 <- approxfun(x = x, y = dn, yleft = 0, yright = 0)
+
+ ##STEP 7(density)
+
+       standardizer <- sum(dn[2:i.max]) + (dn[1]+dn[i.max+1]) / 2
+       dnfun2 <- function(x) dnfun1(x) / standardizer
+
+ ##STEP 6(cdf)
+
+       ## cdf with continuity correction h/2
+       pnfun1 <- approxfun(x = x+0.5*h, y = pn,
+           yleft = 0, yright = pn[i.max+1])
+
+ ##STEP 7(cdf)
+
+       pnfun2 <- function(x) pnfun1(x) / pn[i.max+1]
+

```

```

+
+     ## quantile with continuity correction h/2
+     yleft <- ifelse(((q(D1)(0) == -Inf)|
+                     (q(D1)(0) == -Inf)),
+                     -Inf, N*lower)
+     yright <- ifelse(((q(D1)(1) == Inf)|
+                       (q(D1)(1) == Inf)),
+                       Inf, N*upper)
+     w0 <- options("warn")
+     options(warn = -1)
+     qnfun1 <- approxfun(x = pnfun2(x+0.5*h),
+                         y = x+0.5*h, yleft = yleft, yright = yright)
+     qnfun2 <- function(x){
+       ind1 <- (x == 0)*(1:length(x))
+       ind2 <- (x == 1)*(1:length(x))
+       y <- qnfun1(x)
+       y <- replace(y, ind1[ind1 != 0], yleft)
+       y <- replace(y, ind2[ind2 != 0], yright)
+       return(y)
+     }
+     options(w0)
+
+     rnew = function(N) apply(matrix(r(e1)(n*N),
+                                     ncol=N), 1, sum)
+
+     return(new("AbscontDistribution", r = rnew,
+             d = dnfun1, p = pnfun2, q = qnfun2))
+ })

[1] "convpow"

> ## initialize a normal distribution
> A <- Norm(mean=0, sd=1)
> ## convolution power
> N <- 10
> ## convolution via FFT
> AN <- convpow(as(A,"AbscontDistribution"), N)
> ## ... for the normal distribution , 'convpow' has an "exact"
> ## method by version 1.9 so the as(.,.) is needed to
> ## see how the algorithm above works
>

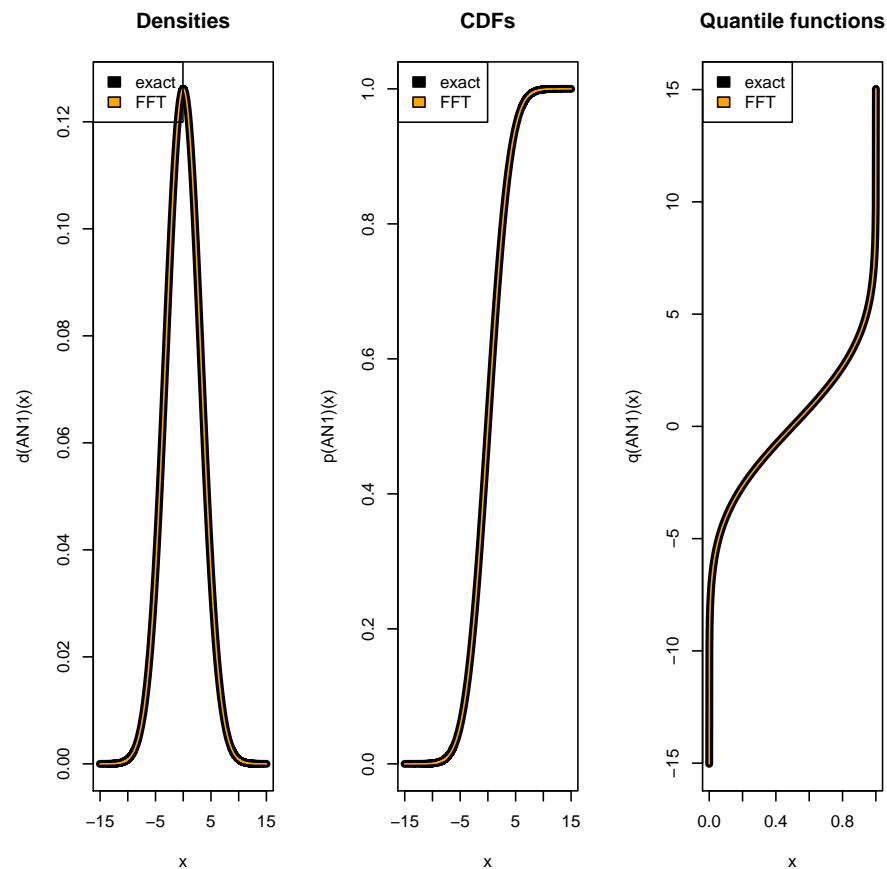
```



```

> ## convolution exact
> AN1 <- Norm(mean=0, sd=sqrt(N))
> ## plots of the results
> eps <- getdistrOption("TruncQuantile")
> par(mfrow=c(1,3))
> low <- q(AN1)(eps)
> upp <- q(AN1)(eps, lower.tail = FALSE)
> x <- seq(from = low, to = upp, length = 10000)
> ## densities
> plot(x, d(AN1)(x), type = "l", lwd = 5)
> lines(x, d(AN)(x), col = "orange", lwd = 1)
> title("Densities")
> legend("topleft", legend=c("exact", "FFT"),
+       fill=c("black", "orange"))
> ## cdfs
> plot(x, p(AN1)(x), type = "l", lwd = 5)
> lines(x, p(AN)(x), col = "orange", lwd = 1)
> title("CDFs")
> legend("topleft", legend=c("exact", "FFT"),
+       fill=c("black", "orange"))
> ## quantile functions
> x <- seq(from = eps, to = 1-eps, length = 1000)
> plot(x, q(AN1)(x), type = "l", lwd = 5)
> lines(x, q(AN)(x), col = "orange", lwd = 1)
> title("Quantile functions")
> legend("topleft",
+       legend = c("exact", "FFT"),
+       fill = c("black", "orange"))

```



References

- [1] Bengtsson H. The R.oo package - object-oriented programming with references using standard R code. In: Hornik K., Leisch F. and Zeileis A. (Eds.) *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*. Vienna, Austria. Published as <http://www.ci.tuwien.ac.at/Conferences/DSC-2003/> 6
- [2] Chambers J.M. *Programming with data. A guide to the S language*. Springer. <http://cm.bell-labs.com/stat/Sbook/index.html> 6
- [3] Gentleman R. *Object Orientated Programming. Slides of a Short Course held in Auckland*. <http://www.stat.auckland.ac.nz/S-Workshop/Gentleman/Methods.pdf> 36, 37
- [4] Kohl M. *Numerical Contributions to the Asymptotic Theory of Robustness*. Dissertation, Universität Bayreuth. See also <http://stamats.de/ThesisMKohl.pdf> 1

- [5] Kohl M., Ruckdeschel P. and Stabla T. General Purpose Convolution Algorithm for Distributions in S4-Classes by means of FFT. unpublished manual [6](#), [17](#), [36](#)
- [6] Press W.H., Teukolsky S.A., Vetterling W.T. and Flannery B.P. *Numerical recipes in C. The art of scientific computing*. Cambridge Univ. Press, 2. Aufl. [27](#)
- [7] Rice J.A. *Mathematical statistics and data analysis*. The Wadsworth & Brooks/Cole Statistics/Probability Series. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, California. [38](#)
- [8] Ruckdeschel P., Kohl M., Stabla T., and Camphausen F. S4 Classes for Distributions. *R-News*, **6**(2): 10–13. http://CRAN.R-project.org/doc/Rnews/Rnews_2006-2.pdf [3](#)