R Programming/Probability Distributions

This page review the main probability distributions and describe the main R functions to deal with them.

R has lots of probability functions.

- `r` is the generic prefix for random variable generator such as `runif()`, `rnorm()`.
- `d` is the generic prefix for the probability density function such as `dunif()`, `dnorm()`.
- `p` is the generic prefix for the cumulative density function such as `punif()`, `pnorm()`.
- `q` is the generic prefix for the quantile function such as `qunif()`, `qnorm()`.

Contents

Distribution in circular statistics

See also

References

Discrete distributions

Benford Distribution

The Benford Distribution (http://en.wikipedia.org/wiki/Benford_distribution) is the distribution of the first digit of a number. It is due to Benford 1938[1] and Newcomb 1881[2].

```
> library(VGAM)
> dbenf(c(1:9))
[1] 0.30103000 0.17609126 0.12493874 0.09691001 0.07918125 0.06694679 0.05799195 0.05115252 0.04575749
```

Bernoulli

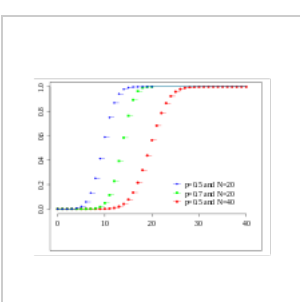
We can draw from a Bernoulli (http://en.wikipedia.org/wiki/Bernoulli_distribution) using `sample()`, `runif()` or `rbinom()` with `size = 1`.

```
> n <- 1000
> x <- sample(c(0,1), n, replace=T)
> x <- sample(c(0,1), n, replace=T, prob=c(0.3,0.7))
> x <- runif(n) > 0.3
> x <- rbinom(n, size=1, prob=0.2)
```

Binomial

We can sample from a binomial distribution (http://en.wikipedia.org/wiki/Binomial_distribution) using the `rbinom()` function with arguments `n` for number of samples to take, `size` defining the number of trials and `prob` defining the probability of success in each trial.

```
> x <- rbinom(n=100,size=10,prob=0.5)
```



Hypergeometric distribution

We can sample `n` times from a hypergeometric distribution (http://en.wikipedia.org/wiki/Hypergeometric_distribution) using the `rhyper()` function.

```
> x <- rhyper(n=1000, 15, 5, 5)
```

Geometric distribution

The geometric distribution (http://en.wikipedia.org/wiki/Geometric_distribution).

```
> N <- 10000
> x <- rgeom(N, .5)
> x <- rgeom(N, .01)
```

Multinomial

The multinomial distribution (http://en.wikipedia.org/wiki/Multinomial_distribution).

```
> sample(1:6, 100, replace=T, prob= rep(1/6,6))
```

Negative binomial distribution

The negative binomial distribution (http://en.wikipedia.org/wiki/Negative_binomial_distribution) is the distribution of the number of failures before k successes in a series of Bernoulli events.

```
> N <- 100000
> x <- rnbinom(N, 10, .25)
```

Poisson distribution

We can draw `n` values from a Poisson distribution (http://en.wikipedia.org/wiki/Poisson_distribution) with a mean set by the argument `lambda`.

```
> x <- rpois(n=100, lambda=3)
```

Zipf's law

The distribution of the frequency of words is known as Zipf's Law (http://en.wikipedia.org/wiki/Zipf%27s_Law). It is also a good description of the distribution of city size[3]. dzipf() and pzipf() (**VGAM**)

```
> library(VGAM)
> dzipf(x=2, N=1000, s=2)
```

Continuous distributions

Beta and Dirichlet distributions

- beta distribution (http://en.wikipedia.org/wiki/Beta_distribution)
- Dirichlet (http://en.wikipedia.org/wiki/Dirichlet_distribution) in **gtools** and **MCMCpack**

```
>library(gtools)
>?rdirichlet
>library(bayesm)
>?rdirichlet
>library(MCMCpack)
>?Dirichlet
```

Cauchy

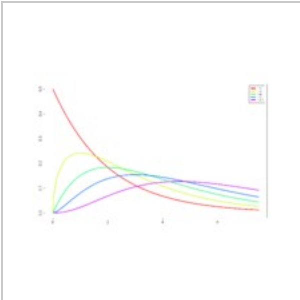
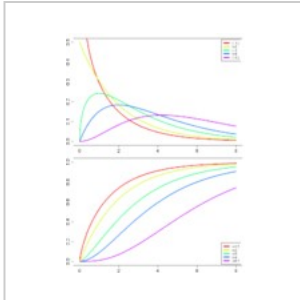
We can sample n values from a Cauchy distribution (http://en.wikipedia.org/wiki/Cauchy_distribution) with a given location parameter x_0 (default is 0) and scale parameter γ (default is 1) using the rcauchy() function.

```
> x <- rcauchy(n=100, location=0, scale=1)
```

Chi Square distribution

Quantile of the Chi square distribution (http://en.wikipedia.org/wiki/Chi-square_distribution) (χ^2 distribution)

```
> qchisq(.95,1)
[1] 3.841459
> qchisq(.95,10)
[1] 18.30704
> qchisq(.95,100)
[1] 124.3421
```

Exponential

We can sample n values from a exponential distribution (http://en.wikipedia.org/wiki/Exponential_distribution) with a given `rate` (default is 1) using the `rexp()` function

```
> x <- rexp(n=100, rate=1)
```

Fisher-Snedecor

We can draw the density of a Fisher distribution (http://en.wikipedia.org/wiki/F-distribution) (F-distribution) :

```
> par(mar=c(3,3,1,1))
> x <- seq(0,5,len=1000)
> plot(range(x),c(0,2),type="n")
> grid()
> lines(x,df(x,df1=1,df2=1),col="black",lwd=3)
> lines(x,df(x,df1=2,df2=1),col="blue",lwd=3)
> lines(x,df(x,df1=5,df2=2),col="green",lwd=3)
> lines(x,df(x,df1=100,df2=1),col="red",lwd=3)
> lines(x,df(x,df1=100,df2=100),col="grey",lwd=3)
> legend(2,1.5,legend=c("n1=1, n2=1","n1=2, n2=1","n1=5, n2=2","n1=100, n2=1","n1=100, n2=100"),col=c("black",
"blue","green","red","grey"),lwd=3,bty="n")
```

Gamma

We can sample n values from a gamma distribution (http://en.wikipedia.org/wiki/Gamma_distribution) with a given `shape` parameter and `scale` parameter θ using the `rgamma()` function. Alternatively a `shape` parameter and `rate` parameter $\beta = 1/\theta$ can be given.

```
> x <- rgamma(n=10, scale=1, shape=0.4)
> x <- rgamma(n=100, scale=1, rate=0.8)
```

Levy

We can sample n values from a Levy distribution (http://en.wikipedia.org/wiki/Levy_distribution) with a given location parameter μ (defined by the argument m, default is 0) and scaling parameter (given by the argument s, default is 1) using the `rlevy()` function.

```
> x <- rlevy(n=100, m=0, s=1)
```

Log-normal distribution

We can sample n values from a log-normal distribution (http://en.wikipedia.org/wiki/Lognormal) with a given `meanlog` (default is 0) and `sdlog` (default is 1) using the `rlnorm()` function

```
> x <- rlnorm(n=100, meanlog=0, sdlog=1)
```

Normal and related distributions

We can sample n values from a normal (http://en.wikipedia.org/wiki/Normal_distribution) or gaussian Distribution with a given `mean` (default is 0) and `sd` (default is 1) using the `rnorm()` function

```
> x <- rnorm(n=100, mean=0, sd=1)
```

Quantile of the normal distribution

```
> qnorm(.95)
[1] 1.644854
> qnorm(.975)
[1] 1.959964
> qnorm(.99)
[1] 2.326348
```

- The **mvtnorm** package includes functions for multivariate normal distributions.
  - `rmvnorm()` generates a multivariate normal distribution.

```
> library(mvtnorm)
> sig <- matrix(c(1, 0.8, 0.8, 1), 2, 2)
> r <- rmvnorm(1000, sigma = sig)
> cor(r)
         [,1]      [,2]
```

```
[1,] 1.0000000 0.8172368
[2,] 0.8172368 1.0000000
```

Pareto Distributions

- Generalized Pareto (http://en.wikipedia.org/wiki/Pareto_distribution) `dgpd()` in **evd**
- `dpareto(), ppareto(), rpareto(), qpareto()` in **actuar**
- The **VGAM** package also has functions for the Pareto distribution.

Student's t distribution

Quantile of the Student t distribution (http://en.wikipedia.org/wiki/Student%27s_t-distribution)

```
> qt(.975,30)
[1] 2.042272
> qt(.975,100)
[1] 1.983972
> qt(.975,1000)
[1] 1.962339
```

The following lines plot the .975th quantile of the t distribution in function of the degrees of freedom :

```
curve(qt(.975,x), from = 2 , to = 100, ylab = "Quantile 0.975 ", xlab = "Degrees of freedom", main = "Student
t distribution")
abline(h=qnorm(.975), col = 2)
```

Uniform distribution

We can sample `n` values from a uniform distribution (http://en.wikipedia.org/wiki/Uniform_distribution_(continuous)) (also known as a rectangular distribution] between two values (defaults are 0 and 1) using the `runif()` function

```
> runif(n=100, min=0, max=1)
```

Weibull

We can sample `n` values from a Weibull distribution (http://en.wikipedia.org/wiki/Weibull_distribution) with a given `shape` and `scale` parameter μ (default is 1) using the `rweibull()` function.

```
> x <- rweibull(n=100, shape=0.5, scale=1)
```

Extreme values and related distribution

- The Gumbel distribution (http://en.wikipedia.org/wiki/Gumbel_distribution)
- The logistic distribution (http://en.wikipedia.org/wiki/Logistic_distribution) : distribution of the difference of two gumbel distributions.

```
plogis, qlogis, dlogis, rlogis
```

- Frechet `dfrechet()` **evd**
- Generalized Extreme Value `dgev()` **evd**
- Gumbel `dgumbel()` **evd**
- `Burr, dburr, pburr, qburr, rburr` in **actuar**

Distribution in circular statistics

- Functions for circular statistics are included in the **CircStats** package.
  - `dvm()` Von Mises (http://en.wikipedia.org/wiki/Von_Mises_distribution) (also known as the nircular normal or Tikhonov distribution) density function
  - `dtri()` triangular density (http://en.wikipedia.org/wiki/Triangular_distribution) function
  - `dmixedvm()` Mixed Von Mises density
  - `dwrpcauchy()` wrapped Cauchy density
  - `dwrpnorm()` wrapped normal density.

See also

- Packages **VGAM**, **SuppDists**, **actuar**, **fBasics**, **bayesm**, **MCMCpack**

References

1. Benford, F. (1938) The Law of Anomalous Numbers. Proceedings of the American Philosophical Society, 78, 551–572.
2. Newcomb, S. (1881) Note on the Frequency of Use of the Different Digits in Natural Numbers. American Journal of Mathematics, 4, 39–40.
3. Gabaix, Xavier (August 1999). "Zipf's Law for Cities: An Explanation". Quarterly Journal of Economics 114 (3): 739–67. doi:10.1162/003355399556133. ISSN 0033-5533. http://pages.stern.nyu.edu/~xgabaix/papers/zipf.pdf.