Definition of distributions

by Agner Fog

This document is published at www.agner.org/random, November 2002, as part of a software package.

Introduction

A C++ class library of uniform and non-uniform random number generators is available from www.agner.org/random. The following distributions are included in this library:

- uniform real
- uniform integer
- normal
- bernoulli
- poisson
- binomial, multinomial
- hypergeometric, multivariate hypergeometric
- noncentral hypergeometric, multivariate noncentral hypergeometric
- extended hypergeometric, multivariate extended hypergeometric

Descriptions of most of these distributions can be found in any textbook on statistics. However, the last two distributions are not very well-known and need a more detailed description here.

The methods used for generating variates with these distributions are described in the files `sampmet.pdf` and `nchyp.pdf` (Fog 2002) available from www.agner.org/random.

Uniform distribution, real

`x = Random()`

x ~ uniform(0,1)

x has a uniform distribution over the interval $0 \le x < 1$.

$$f(x) = 1$$

Uniform distribution, integer

```
x = IRandom(min,max)
```

All integer values in the interval $\min \le x \le \max$ are equally probable.

$$f(x) = \frac{1}{\max - \min + 1}$$

Normal distribution

```
x = Normal(m,s)
```

$x \sim \text{normal}(m,s)$

Normal distribution with mean m and standard deviation s. This distribution simulates the sum of many random contributions.

$$f(x) = \frac{1}{\sqrt{2\pi}s} e^{-\frac{(x-m)^2}{2s^2}}$$

Bernoulli distribution

```
x = Bernouilli(p)
```

$x \sim \text{binomial}(1,p)$

Describes a situation with two possible outcomes. The probability of success is p.

$$f(0) = 1 - p, \quad f(1) = p$$

Poisson distribution

```
x = Poisson(L)
```

$x \sim \text{poisson}(\lambda)$

This is the distribution of the number of events in a given time span or a given geographical area when these events are randomly scattered in time or space.

$$f(x) = \frac{\lambda^x}{x!} e^{-\lambda}$$

Binomial distribution

```
x = Binomial(n,p)
```

$x \sim \text{binomial}(n,p)$

This is the distribution of the number of red balls you get when drawing n balls *with replacement* from an urn where p is the fraction of red balls in the urn.

$$\text{f}(x) = \binom{n}{x} p^x (1-p)^{n-x}$$

Multinomial (multivariate binomial) distribution

```
Multinomial(x,p,n,c)
```

$\mathbf{x} \sim \text{mbinomial}(\mathbf{p},n,c)$

This distribution extends the urn experiment *with replacement* to any number of colors, c. p_i is the fraction of balls of color i in the urn. n is the number of balls you take. x_i is the number of balls of color i you get in the sample.

$$\text{f}(\mathbf{x}) = \frac{n!}{\prod_{i=1}^{c} x_i!} \prod_{i=1}^{c} p_i^{x_i}$$

Hypergeometric distribution

```
x = Hypergeometric(n,m,N)
```

$x \sim \text{hyp}(n,m,N)$

This is the distribution of the number of red balls you get when drawing n balls *without replacement* from an urn where m is the number of red balls in the urn and N is the total number of balls in the urn.

$$\text{f}(x) = \frac{\binom{m}{x}\binom{N-m}{n-x}}{\binom{N}{n}}$$

Multivariate hypergeometric distribution

`MultiHypergeo(x,m,n,c)`

$\mathbf{x} \sim \mathrm{mhyp}(\mathbf{m},n,c)$

This distribution extends the urn experiment *without replacement* to any number of colors, c. m_i is the number of balls of color i in the urn. N is the total number of balls in the urn. n is the number of balls you take. x_i is the number of balls of color i you get in the sample.

$$\mathrm{f}(\mathbf{x}) = \frac{\prod_{i=1}^{c}\binom{m_i}{x_i}}{\binom{N}{n}}$$

Noncentral hypergeometric distribution, univariate and multivariate

```
x = NonCentralHypergeometric (n,m,N,w)
MultiNonCentralHypergeo(x,m,w,n,c)
```

$\mathbf{x} \sim \mathrm{mnchyp}(n,\mathbf{m},\boldsymbol{\omega})$

This distribution models an urn experiment without replacement, with bias. Assume that an urn contains N balls of c different colors. m_i is the number of balls of color i. The balls have different weight (or size) which makes the sampling biased. The probability that a particular ball of color i is taken is proportional to its weight w_i. The univariate distribution is the special case $c = 2$. Note that the extended hypergeometric distribution, defined below, is sometimes erroneously called the noncentral hypergeometric distribution.

The univariate noncentral hypergeometric distribution is defined by Wallenius (1963) and the multivariate distribution by Chesson (1976):

$$\mathrm{f}(\mathbf{x}) = \left(\prod_{i=1}^{c}\binom{m_i}{x_i}\right)\int_0^1 \prod_{i=1}^{c}(1-t^{\omega_i/d})^{x_i}\,dt, \text{ where } d = \sum_{i=1}^{c}\omega_i(m_i - x_i)$$

$$\sum x_i = n, \quad \sum m_i = N$$

Description of various properties of this distribution, calculation methods and sampling methods are given in Fog (2002).

Extended hypergeometric distribution, univariate and multivariate

```
x = ExtendedHypergeometric(n,m,N,w)
MultiExtendedHypergeo(x,m,w,n,c)
```

$\mathbf{x} \sim \mathrm{mxhyp}(n,\mathbf{m},\mathbf{w})$

This distribution resembles the noncentral hypergeometric distribution, and unfortunately it is often given the same name (e.g. McCullagh & Nelder 1989). The correct name for this distribution is the extended hypergeometric distribution (Harkness 1965, Johnson & Kotz 1969).

The extended hypergeometric distribution is defined as a conditional distribution. Let x_i be c independent binomial variates with the distributions

$x_i \sim \text{binomial}(m_i, p_i), \quad i = 1..c$

then the distribution of $\mathbf{x}$ on the condition that $\sum x_i = n$ is the (multivariate) extended hypergeometric distribution.

This distribution is not normally associated with the metaphor or taking colored balls from an urn, but here I will apply an urn model, however farfetched, for the sake of analogy with the noncentral hypergeometric distribution. You are taking n balls without replacement from an urn containing balls of c different colors. The balls have different weights w_i which make the sampling biased in favor of the heavier balls. Before taking the balls you assign to each ball a probability p_i of being taken. These probabilities are calculated so that the expected total number of balls taken is n. Now you take or don't take each ball according to the assigned probabilities and count the total number of balls taken. If this number is not equal to n then put all the balls back in the urn and repeat the experiment. You may have to repeat this experiment many times before you have a sample containing exactly n balls.

The relationship between p_i and w_i is given by

$$p_i = \frac{rw_i}{1 + rw_i}$$

where the scale factor r is adjusted so that $\sum m_i p_i = n$.

The distribution function is

$$\mathrm{f}(\mathbf{x}) = \frac{\mathrm{f}_1(\mathbf{x})}{\sum_{\mathbf{y} \in S} \mathrm{f}_1(\mathbf{y})}, \text{ where}$$

$$\mathrm{f}_1(\mathbf{x}) = \prod_{i=1}^{c} \binom{m_i}{x_i} w_i^{x_i}, \text{ and the support}$$

$$S = \left\{ \mathbf{x} \in \mathbb{Z}_{0+}^{c} \mid \sum_{i=1}^{c} x_i = n \right\}$$

Description of various properties of this distribution, calculation methods and sampling methods are given in Fog (2002).

References

Chesson, J (1976): A non-central multivariate hypergeometric distribution arising from biased sampling with application to selective predation. *Journal of Applied Probability*, vol. 13, no. 4, pp. 795-797.

Fog, Agner (2002): *The noncentral and extended hypergeometric probability distributions.* www.agner.org/random/theory/nchyp.pdf.

Harkness, W L (1965): Properties of the Extended Hypergeometric Distribution. *Annals of Mathematical Statistics*, vol. 36, no. 3, pp. 938-945.

Johnson, N L & Kotz, S (1969): *Distributions in Statistics, 1: Discrete Distributions*. Boston: Houghton Mifflin Co.

McCullagh, P & Nelder, J A (1989): *Generalized Linear Models*. 2. ed. London: Chapman & Hall.

Wallenius, K T (1963): *Biased Sampling: The Non-central Hypergeometric Probability Distribution*. Ph.D. thesis, Stanford University.