Journal of Software for Algebra and Geometry

Probability package for Macaulay2

DOUGLAS A. TORRANCE

vol 14 2024

JSAG **14** (2024), 51–57
https://doi.org/10.2140/jsag.2024.14.51

The Journal of Software for
Algebra and Geometry

`Probability` package for Macaulay2

DOUGLAS A. TORRANCE

ABSTRACT: We introduce the `Probability` package for Macaulay2, which provides an interface for users to compute probabilities and generate random variates from a wide variety of univariate probability distributions.

1. INTRODUCTION. Users of Macaulay2 [9] often have reason to use random elements. Entire packages, e.g., `SpaceCurves` [15], exist to create random algebrogeometric objects. Also, a growing number of packages, e.g., `GraphicalModels` [7], are devoted to the exciting field of algebraic statistics.

Macaulay2 has a built-in method `random` that wraps around two functions in shared libraries that it is linked against: `mpz_urandomm` from GMP [8] for generating random integers and `mpfr_urandomb` from MPFR [6] for generating random real numbers. Both of these use the Mersenne Twister pseudorandom number generator [10] for sampling random variates from uniform discrete and continuous probability distributions, respectively.

There has previously been no support for other probability distributions, and users interested in these distributions would have needed to look elsewhere, for example, to the statistical software R [13]

However, beginning with version 1.20, Macaulay2 has been distributed with the `Probability` package, written by the author, that implements many common probability distributions and provides an interface for users to create additional distributions for use in their work.

This paper is outlined as follows. In Section 2, we review the basics of probability theory. In Section 3, we introduce the `ProbabilityDistribution` class from the package and its methods. Finally, in Section 4, we use the package to perform an example of Pearson's chi-squared test for independence.

2. BASICS OF PROBABILITY THEORY. The basics of probability theory are well-covered in a large variety of undergraduate- and graduate-level textbooks, e.g., [3; 11; 14], but we review the essentials here.

Suppose Ω is a *sample space* containing the possible outcomes of some random experiment, $\mathcal{F}$ is a σ-algebra on Ω (known as the *event space*), and $P : \mathcal{F} \to [0, 1]$ is a σ-additive *probability measure* (or *probability distribution*) satisfying $P(\varnothing) = 0$ and $P(\Omega) = 1$. The triple $(\Omega, \mathcal{F}, P)$ is known as a *probability space*. A *random variable* is measurable function $X : \Omega \to \mathbb{R}$. The probability distribution of this random variable is the pushforward measure X_*P on $\mathbb{R}$, i.e., for any Borel set $A \subset \mathbb{R}$, where

MSC2020: 60-04, 14-04.
Keywords: Macaulay2, probability.
`Probability` version 0.3

distribution	parameters	supp(X)	$f_X(x)$, $x \in \text{supp}(X)$
binomial	n, p	$\{0, \dots, n\}$	$\binom{n}{x} p^x (1-p)^{n-x}$
Poisson	λ	$\{0, 1, \dots\}$	$\frac{\lambda^x}{x!} e^{-\lambda}$
geometric	p	$\{0, 1, \dots\}$	$p(1-p)^x$
negative binomial	r, p	$\{0, 1, \dots\}$	$\frac{\Gamma(x+r)}{\Gamma(r)x!} p^r (1-p)^x$
hypergeometric	m, n, k	$\{0, \dots, m\}$	$\frac{\binom{m}{x}\binom{n}{k-x}}{\binom{m+n}{k}}$

Table 1. Common discrete probability distributions

$$X_* P(A) = P(\{\omega : X(\omega) \in A\}),$$

usually denoted $P(X \in A)$. The smallest closed A for which $P(X \in A) = 1$ is the *support* of X, denoted supp(X). The Radon–Nikodym derivative of $X_* P$ with respect to the measure μ on $\mathbb{R}$, i.e., the function f_X satisfying $P(X \in A) = \int_A f_X \, d\mu$, is known as the *probability density function* of X. The function F_X defined by

$$F_X(x) = P(X \leq x)$$

is the *cumulative distribution function* of X.

There are two main cases. When $X(\Omega)$ is countable and μ is the counting measure, i.e.,

$$\int_A f_X \, d\mu = \sum_{x \in A} f_X(x),$$

then X is said to have a *discrete* distribution. In this case, f_X is often referred to instead as the *probability mass function*. When $X(\Omega)$ is uncountable and μ is the usual Lebesgue measure, then X has a *continuous* distribution. See Tables 1 and 2 for common discrete and continuous probability distributions, respectively.

The *quantile function* is the function $Q_X : [0, 1] \to \mathbb{R} \cup \{-\infty, \infty\}$ defined by

$$Q_X(p) = \inf\{x : F_X(x) \geq p\}.$$

It is useful for generating random variates of X using the *inversion method*. In particular, if u is a random variate of a random variable U with the uniform distribution on the interval [0, 1], then $Q_X(u)$ is a random variate of X [4].

3. THE PROBABILITYDISTRIBUTION CLASS. The `Probability` package defines an abstract class `ProbabilityDistribution` with two subclasses, called `DiscreteProbabilityDistribution` and `ContinuousProbabilityDistribution`. Discrete and continuous random variables are represented by instances of these two classes, respectively.

distribution	parameters	$\operatorname{supp}(X)$	$f_X(x)$, $x \in \operatorname{supp}(X)$
uniform	a, b	$[a, b]$	$\frac{1}{b-a}$
exponential	λ	$[0, \infty)$	$\lambda e^{-\lambda x}$
normal	μ, σ	$\mathbb{R}$	$\frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$
gamma	α, λ	$[0, \infty)$	$\frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}$
chi-squared	ν	$[0, \infty)$	$\frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{\nu/2-1} e^{-x/2}$
Student's t	ν	$\mathbb{R}$	$\frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\,\Gamma(\nu/2)} \left(1 + \frac{x^2}{\nu}\right)^{-(\nu+1)/2}$
F	ν_1, ν_2	$[0, \infty)$	$\frac{\sqrt{(\nu_1 x)^{\nu_1} \nu_2^{\nu_2} / (\nu_1 x + \nu_2)^{\nu_1+\nu_2}}}{x B(\nu_1/2, \nu_2/2)}$
beta	α, β	$[0, 1]$	$\frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha, \beta)}$

Table 2. Common continuous probability distributions

There are four installed methods. If `X` is a `ProbabilityDistribution` object corresponding to a random variable X, then

(a) `density_X` is the probability density function f_X,

(b) `probability_X` is the cumulative distribution function F_X,

(c) `quantile_X` is the quantile function Q_X, and

(d) `random X` is a random variate of X.

To construct a `ProbabilityDistribution` object, use one of the built-in constructor methods, `discreteProbabilityDistribution` or `continuousProbabilityDistribution`.

At minimum, these methods take a function (the probability density function) as input. If the support differs from $\{0, 1, \ldots\}$ (for discrete probability distributions) or $[0, \infty)$ (for continuous probability distributions), then it may be specified using the `Support` option. This option takes a sequence of the form (a, b). In the discrete case, this is interpreted as the set $\{a, \ldots, b\} \subset \mathbb{Z}$, and in the continuous case, the interval $(a, b) \subset \mathbb{R}$. Setting $a = -\infty$ and/or $b = \infty$ is also allowed.

A cumulative distribution function based on the probability density function is installed by default (using `sum` for discrete probability distributions and `integrate` for continuous probability distributions), but one may also be specified using the `DistributionFunction` option.

Similarly, a naïve quantile function (adding probabilities inside a `while` loop until the target probability p is obtained in the discrete case and using the bisection method for numerically approximating a root of $F_X(x) - p$ in the continuous case) is installed by default, but may be specified using the `QuantileFunction` option.

By default, `random` uses the inversion method to generate random variates, but this may be overridden using the `RandomGeneration` option.

Example 3.1 (six-sided die). Consider the following simple example, where X is a discrete random variable corresponding to the result of rolling a fair six-sided die, demonstrating each of the above methods:

```
i1 : needsPackage "Probability";
i2 : X = discreteProbabilityDistribution(x -> 1/6, Support => (1, 6));
i3 : density_X 3
o3 = 1/6
o3 : QQ
i4 : probability_X 3
o4 = 1/2
o4 : QQ
i5 : quantile_X oo
o5 = 3
i6 : random X
o6 = 2
```

Example 3.2 (triangular distribution). We next give an example of a continuous probability distribution, the Bates distribution of the mean of two random variables with the uniform distribution on the unit interval, which has probability density function

$$f_X(x) = \begin{cases} 4x & \text{if } 0 \le x \le \frac{1}{2}, \\ 4(1-x) & \text{if } \frac{1}{2} < x \le 1, \\ 0 & \text{otherwise,} \end{cases}$$

and so its probabilities are areas inside an isosceles triangle:

```
i1 : needsPackage "Probability";
i2 : X = continuousProbabilityDistribution(
         x -> if x < 1/2 then 4*x else 4*(1 - x),
         Support => (0, 1));
i3 : density_X(2/3)
o3 = 4/3
o3 : QQ
i4 : probability_X(2/3)
o4 = .777777777777778
o4 : RR (of precision 53)
i5 : quantile_X oo
o5 = .666666666666668
o5 : RR (of precision 53)
i6 : random X
o6 = .318199004810108
o6 : RR (of precision 53)
```

As is the case when working over $\mathbb{R}$ in any computer system, there can be unexpected results due to floating point rounding errors. This is apparent in the outputs o4 and o5 in the example above, where outputs of $\frac{7}{9}$ and $\frac{2}{3}$ would be expected. Users should take care to be aware of this.

Alternatively, one of the built-in constructor methods for each of the common probability distributions listed in Tables 1 and 2 may be used. For example, `binomialDistribution(n,p)` will return a `DiscreteProbabilityDistribution` object corresponding to a binomially distributed random variable with parameters n and p.

A number of new special functions (e.g, `Beta` and `inverseErf`) were added to Macaulay2 beginning in version 1.20 by the author and Paul Zinn-Justin, wrapping around corresponding functions from the Boost math toolkit [1] and MPFR [6]. Many of these were used to define the density, distribution, and / or quantile functions of these common distributions.

4. CHI-SQUARED TEST FOR INDEPENDENCE. In one of the seminal works in algebraic statistics, Diaconis and Sturmfels [5] considered a data set from [2]. A 12×12 contingency table, represented by the matrix `birthDeath` below, contains birth and death month information for 82 descendants of Queen Victoria. In particular, the rows correspond to birth months (January to December) and the columns correspond to death months. So for example, the 2 in `birthDeath_(3, 2)` means that two of the 82 people represented in the data set were in born in April and died in March.

```
i1 : birthDeath = matrix {
          {1, 0, 0, 0, 1, 2, 0, 0, 1, 0, 1, 0},
          {1, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 2},
          {1, 0, 0, 0, 2, 1, 0, 0 ,0 ,0 ,0, 1},
          {3, 0, 2, 0, 0, 0, 1, 0, 1, 3, 1, 1},
          {2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0},
          {2, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0},
          {2, 0, 2, 1, 0, 0, 0, 0, 1, 1, 1, 2},
          {0, 0, 0, 3, 0, 0, 1, 0, 0, 1, 0, 2},
          {0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 1, 0},
          {1, 1, 0, 2, 0, 0, 1, 0, 0, 1, 1, 0},
          {0, 1, 1, 1, 2, 0, 0, 2, 0, 1, 1, 0},
          {0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0, 0}};

                12       12
o1 : Matrix ZZ   <--- ZZ
```

The question at hand is whether these two variables (birth month and death month) are independent. A standard test is due to Pearson [12]. In particular, given an $m \times n$ contingency table O, let

$$O_{i\bullet} = \sum_{j=1}^{n} O_{ij}, \quad O_{\bullet j} = \sum_{i=1}^{m} O_{ij}, \quad N = \sum_{i=1}^{m}\sum_{j=1}^{n} O_{ij}.$$

Then, defining $E_{ij} = O_{i\bullet}O_{\bullet j}/N$ for all i, j, the test statistic

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

converges in distribution to a chi-squared distribution with $(m-1)(n-1)$ degrees of freedom.

First, we compute the test statistic, noting that $O_{i\bullet}O_{\bullet j}$ and N may be computed nicely by multiplying O by various matrices of ones.

```
i2 : ones = (m, n) -> matrix toList(m : toList(n : 1));
i3 : N = (ones(1, 12) * birthDeath * ones(12, 1))_(0, 0);
i4 : expected = (1/N) * birthDeath * ones(12, 12) * birthDeath;
                12          12
o4 : Matrix QQ   <--- QQ
i5 : chi2 = numeric sum flatten table(12, 12, (i, j) ->
         (birthDeath_(i, j) - expected_(i, j))^2 / expected_(i, j))
o5 = 115.559632730585
o5 : RR (of precision 53)
```

Finally, we use the `Probability` package to compute the p-value of the hypothesis test.

```
i6 : needsPackage "Probability";
i7 : X = chiSquaredDistribution((12 - 1) * (12 - 1))
o7 = chi2(121)
o7 : ContinuousProbabilityDistribution
i8 : probability_X(chi2, LowerTail => false)
o8 = .622505910459144
o8 : RR (of precision 53)
```

In particular, if X is a continuous random variable with the chi-squared distribution with 121 degrees of freedom and χ^2 is the Pearson test statistic computed above, then $P(X > \chi^2) = 0.6225$. So under the null hypothesis that birth and death months are independent, it is quite likely that the sample data may have been obtained, and thus there is not sufficient evidence to reject it in favor of the alternate hypothesis that they are dependent.

ACKNOWLEDGEMENT. The author would like to thank the referees for helpful comments on both this paper and the package.

SUPPLEMENT. The online supplement contains version 0.3 of `Probability`.

REFERENCES.

[1] N. Agrawal, A. Bikineev, M. Borland, P. A. Bristow, M. Guazzone, C. Kormanyos, H. Holin, B. Lalande, J. Maddock, E. Miller, J. Murphy, M. Pulver, J. Råde, G. Sewani, B. Sobotta, N. Thompson, T. van den Berg, D. Walker, , and X. Xhang, "Boost math toolkit", Available at https://www.boost.org/. Version 1.84.0.

[2] D. F. Andrews and A. M. Herzberg, *Data: A collection of problems from many fields for the student and research worker*, Springer, New York, 1985. Zbl

[3] P. Billingsley, *Probability and measure*, 3rd ed., Wiley, New York, 1995. MR Zbl

[4] L. Devroye, *Nonuniform random variate generation*, Springer, New York, 1986. MR Zbl

[5] P. Diaconis and B. Sturmfels, "Algebraic algorithms for sampling from conditional distributions", *Ann. Statist.* **26**:1 (1998), 363–397. MR Zbl

[6] L. Fousse, G. Hanrot, V. Lefèvre, P. Pélissier, and P. Zimmermann, "MPFR: A multiple-precision binary floating-point library with correct rounding", *ACM Trans. Math. Software* **33**:2 (2007), art. id. 13. MR Zbl

[7] L. D. García-Puente, S. Petrović, and S. Sullivant, "Graphical models", *J. Softw. Algebra Geom.* **5** (2013), 1–7. MR Zbl

[8] T. Granlund, "GNU multiple precision arithmetic library", Available at https://gmplib.org/. Version 6.3.0. Zbl

[9] D. Grayson and M. Stillman, "Macaulay 2: A software system for algebraic geometry and commutative algebra", Available at https://macaulay2.com/. Zbl

[10] M. Matsumoto and T. Nishimura, "Mersenne twister: A 623-dimensionally equidistributed uniform pseudo-random number generator", *ACM Trans. Model. Comput. Simul.* **8**:1 (1998), 3–30. Zbl

[11] W. Mendenhall and R. L. Scheaffer, *Mathematical statistics with applications*, Duxbury Press, North Scituate, MA, 1973. MR Zbl

[12] K. Pearson, "X: On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling", *The London*, *Edinburgh*, *and Dublin Philosophical Magazine and Journal of Science* **50**:302 (1900), 157–175. Zbl

[13] R Core Team, "R: A language and environment for statistical computing", 2021, available at https://www.R-project.org/. Zbl

[14] S. Ross, *A first course in probability*, Macmillan, New York, 1976. MR Zbl

[15] M. Zhang, "The SpaceCurves package in Macaulay2", *J. Softw. Algebra Geom.* **8** (2018), 31–48. MR Zbl

RECEIVED: 19 Dec 2022 REVISED: 31 Oct 2023 ACCEPTED: 23 Jan 2024

DOUGLAS A. TORRANCE:

dtorrance@piedmont.edu

Department of Mathematical Sciences, Piedmont University, Demorest, GA, United States