“ Oh Zeus, free me! ” – Teaching mathematical statistics with *Mathematica* / mathStatica

Colin Rose

Theoretical Research Institute
Bellevue Hill, Sydney, NSW 2023, Australia
Email: colin@tri.org.au

Abstract: **mathStatica** is a general toolset for doing exact (symbolic) mathematical statistics with a computer algebra system. It provides automated statistical operators for taking expectations, finding probabilities, deriving transformations of random variables, finding moments, order statistics, cumulative distribution functions, characteristic functions etc – all for arbitrary user-defined distributions. **mathStatica** v1 accompanies the book: Rose and Smith (2002), *Mathematical Statistics with Mathematica*, Springer-Verlag.

This paper illustrates how **mathStatica** can: free up lecture time by reducing the need to teach laborious and repetitive methods; free students from dreary and monotonous calculations; free both the student and researcher to experiment and play with higher-order concepts in real-time; and for appropriate classes of problems, significantly change the notion of what is difficult, what one can reasonably solve, how one solves it, and perhaps even the notion of what is publishable. We argue that the shift from the traditional numerical approach to a symbolic approach has much broader parallels, namely to an evolving epistemology of statistical knowledge … essentially a shift from a 19th C database conception of knowledge to an algorithmic one.

Introduction

The use of computer software in statistics is, of course, far from new – there already exist hundreds of statistical computer programs. However, existing programs almost always take an essentially numerical / graphical view of the world. By contrast, the **mathStatica** software package has been created with a *symbolic* engine constructed on top of *Mathematica*’s computer algebra system. Living in a *numerical* versus *symbolic* world is not merely an issue of accuracy. Nor is it merely about approximate (numerical) versus exact (symbolic) solutions. More importantly, a symbolic approach to computational statistics significantly changes what one can do, and how one does it. This paper illustrates such concepts over 4 sections, namely: (1) Freedom from hard labour, (2) Expanding the set of problems that one can solve; (3) The evolving epistemology of statistical knowledge; and finally (4) Is that the right answer, Dr Faustus?

1 Freedom from hard labour

A symbolic approach to computational statistics can make solving problems both easier and faster, often dramatically so, in the same way that using a pocket calculator is easier than using a slide rule or log book, for the class of problems for which such tools are relevant. In particular, a symbolic toolkit can free one from laborious mechanical tasks (*e.g.* teaching integration by parts) which are often of little statistical interest in their own right. It seems inevitable that the teaching of techniques that can be automated has the same future as the teaching of long division. This transition from laborious and repetitive mechanics to simple automation means that one can ideally use the time thus saved to either explore higher-order concepts, or trying to solve the remaining subset of problems that do not yield so easily to the pleasures of automation.

A General Toolkit

mathStatica adds over 100 new functions to *Mathematica*. But most of the time, we can get by with just a few of them. Importantly, the tools all operate on general and arbitrary user-specified pdf's, not just standard built-in distributions. To illustrate, let us suppose that random variables X and Y have a Gumbel bivariate Exponential distribution with joint density $f(x, y)$:

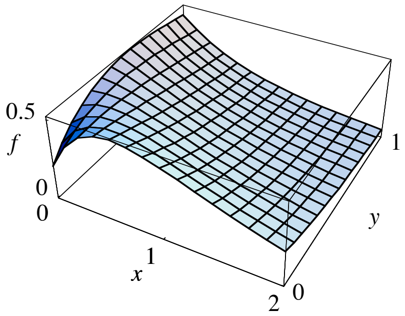
$$f(x, y) \;=\; e^{-2(x+y)}\,(e^{x+y} + \alpha(e^x - 2)(e^y - 2))$$

with domain of support $x > 0$, $y > 0$, where parameter α is such that $-1 < \alpha < 1$. We enter this as:

$$\mathbf{f \;=\; e^{-2\,(x+y)}\;(e^{x+y} + \alpha\;(e^{x} - 2)\;(e^{y} - 2)\;);}$$

```
domain[f] = {{x, 0, ∞}, {y, 0, ∞}}  &&  {-1 < α < 1};
```

Here is a plot of $f(x, y)$ when $\alpha = -0.8$:



Here is the joint distribution function, namely $P(X \le x,\, Y \le y)$:

```
Prob[{x, y}, f]
```

$$e^{-2\,(x+y)}\;(-1 + e^{x})\;(-1 + e^{y})\;(e^{x+y} + \alpha)$$

Here is Cov(X, Y), the covariance between X and Y:

```
Cov[{x, y}, f]
```

$$\frac{\alpha}{4}$$

More generally, here is the variance-covariance matrix:

```
Varcov[f]
```

$$\begin{pmatrix} 1 & \frac{\alpha}{4} \\ \frac{\alpha}{4} & 1 \end{pmatrix}$$

Here is the marginal density of X:

```
Marginal[x, f]
```

$$e^{-x}$$

Here is the conditional density of Y, given $X = x$:

```
Conditional[y, f]
```

$$e^{x-2\,(x+y)}\;(e^{x+y} + (-2 + e^{x})\;(-2 + e^{y})\;\alpha)$$

Here is the bivariate moment generating function $E[e^{t_1 X + t_2 Y}]$:

$$\mathbf{mgf = Expect[}\, e^{t_1 x + t_2 y}\mathbf{,\ f]}$$

— This further assumes that: $\{t_1 < 1,\ t_2 < 1\}$

$$\frac{4 - 2\, t_2 + t_1\, (-2 + (1+\alpha)\, t_2)}{(-2+t_1)\,(-1+t_1)\,(-2+t_2)\,(-1+t_2)}$$

Here is the product moment $E[X^2\, Y^2]$:

$$\mathbf{Expect[}x^2\, y^2\mathbf{,\ f]}$$

$$4 + \frac{9\,\alpha}{4}$$

Multivariate transformations pose no problem either. For instance, let $U = \frac{Y}{1+X}$ and $V = \frac{1}{1+X}$ denote transformations of X and Y, to U and V. Then, using **mathStatica**'s `Transform` function, we can find the joint pdf of random variables U and V, as follows:

$$\mathbf{Transform}\left[\left\{u == \frac{y}{1+x},\ v == \frac{1}{1+x}\right\},\ f\right]$$

$$\frac{e^{\frac{-2-2u+v}{v}}\left(4\, e\, \alpha - 2\, e^{\frac{1}{v}}\, \alpha - 2\, e^{\frac{u+v}{v}}\, \alpha + e^{\frac{1+u}{v}}\, (1+\alpha)\right)}{v^3}$$

Example: **Products of Random Variables (piecewise functions)**

Let random variable $X \sim \text{Pareto}(a, b)$ with pdf $f(x)$:

$$\mathbf{f = a\, b^a\, x^{-(a+1)};} \qquad \mathbf{domain[f] = \{x,\ b,\ \infty\}\ \&\&\ \{a > 0,\ b > 0\};}$$

and let random variable Y have a standard Triangular distribution with pdf $g(y)$ defined in piecewise form:

$$\mathbf{g} = \begin{cases} \frac{2\,y}{c} & 0 \le y \le c \\ \frac{2\,(1-y)}{1-c} & c < y \le 1 \end{cases}; \qquad \mathbf{domain[g] = \{y,\ -\infty,\ \infty\}\ \&\&\ \{0 < c < 1\};}$$

We seek the pdf of the product of the above random variables X and Y, *i.e.* the pdf of $V = X\,Y$. The solution is a piecewise pdf, and it can be simply obtained with **mathStatica** v2 as:

$$\mathbf{TransformProduct[v,\ \{f,\ g\}]}$$

$$\begin{cases} \frac{2\,a\,v}{(2+a)\,b^2\,c} & 0 < v < b\,c \\ \frac{2\,a\left(b^2\,c\,\left(\frac{b\,c}{v}\right)^a - (2+a)\,b\,v + (1+a)\,v^2\right)}{(2+3\,a+a^2)\,b^2\,(-1+c)\,v} & b\,c < v < b \\ \frac{2\,a\,b^a\,(-1+c^{1+a})\,v^{-1-a}}{(2+3\,a+a^2)\,(-1+c)} & v > b \end{cases}$$

2 Expanding the set of problems that one can solve

A symbolic toolset potentially enables one to derive completely new results in real-time. We illustrate with three examples that take just a few seconds to derive here, but which would require an enormous amount of careful work to do manually.

***Example:* Order statistics with non-identical distributions**

Let us suppose we have three completely different distributions defined over three different domains of support. In the following, $f(x)$ is the pdf of an Exponential(λ), $g(x)$ is the pdf of a standard Normal, and $h(x)$ is the pdf of a Uniform(-1, 1) random variable:

$$\mathbf{f} = \frac{1}{\lambda}\, e^{-x/\lambda}; \qquad \text{domain[f]} = \{x,\ 0,\ \infty\}\ \&\&\ \{\lambda > 0\};$$

$$\mathbf{g} = \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}; \qquad \text{domain[g]} = \{x,\ -\infty,\ \infty\};$$

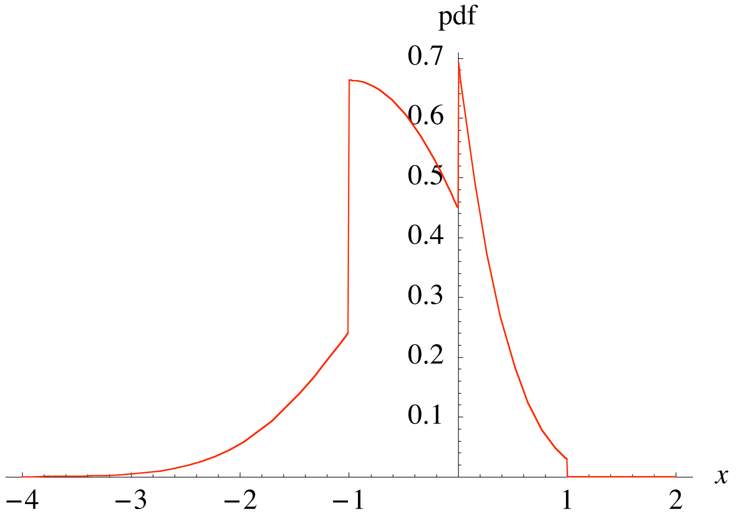
$$\mathbf{h} = \frac{1}{2}; \qquad \text{domain[h]} = \{x,\ -1,\ 1\};$$

mathStatica’s `OrderStat` function (v2) finds the pdf of any order statistic given a random sample, irrespective of whether the values are drawn from identical or non-identical distributions. As a simple illustration, let us find the pdf of min(X, Y, Z), when $X \sim$ Exponential(λ), $Y \sim$ Normal(0, 1) and $Z \sim$ Uniform(-1, 1). The pdf of min(X, Y, Z) is simply the pdf of the first order statistic:

```
OrderStat[1, {f, g, h}]
```

$$\begin{cases} \dfrac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} & x \le -1 \\[2ex] -\dfrac{e^{-\frac{x^2}{2}}\,(-1+x)}{2\sqrt{2\pi}} + \dfrac{1}{4}\,\text{Erfc}\left[\dfrac{x}{\sqrt{2}}\right] & -1 < x \le 0 \\[2ex] \dfrac{e^{-\frac{1}{2}x\left(x+\frac{2}{\lambda}\right)}\left(-\sqrt{\frac{2}{\pi}}\,(-1+x)\,\lambda + e^{\frac{x^2}{2}}\,(1-x+\lambda)\,\text{Erfc}\left[\frac{x}{\sqrt{2}}\right]\right)}{4\lambda} & 0 < x < 1 \end{cases}$$

Here is a plot of the pdf we have just derived (with $\lambda = 1$). One can, of course, easily ‘check’ the solution using Monte Carlo methods.



***Example:* Find the covariance between sample moments**

Let $(X_1, \ldots, X_n)$ denote a random sample of size n drawn from a population random variable X. Find the covariance between the sample mean $\frac{1}{n}\sum_{i=1}^{n} X_i$ and $\left(\frac{1}{n}\sum_{i=1}^{n} X_i^3\right)\left(\sum_{i=1}^{n} X_i^2\right)$.

Solution: The *lingua franca* for such problems is the power sum $s_r = \sum_{i=1}^{n} X_i^r$. Using this notation, we are seeking $\text{Cov}\left(\frac{s_1}{n}, \frac{s_3\, s_2}{n}\right) = \mu_{1,1}\left(\frac{s_1}{n}, \frac{s_3\, s_2}{n}\right)$, i.e. the covariance is just the $\{1, 1\}^{\text{th}}$ product central moment. The solution with **mathStatica** is then simply:

$$\mathtt{CentralMomentToCentral}\left[\{1,\ 1\},\ \left\{\frac{\mathtt{s_1}}{\mathtt{n}},\ \frac{\mathtt{s_3\ s_2}}{\mathtt{n}}\right\}\right]$$

$$\frac{(-1+\mathtt{n})\ \mu_3^2}{\mathtt{n}} + \frac{(-1+\mathtt{n})\ \mu_2\ \mu_4}{\mathtt{n}} + \frac{\mu_6}{\mathtt{n}}$$

***Example:* Find the mixture distribution Poisson(L) $\underset{L}{\bigwedge}$ InverseGaussian(μ, λ)**

Let random variable X have the conditional distribution $X \,\big|\, (P = p) \sim \text{Poisson}(p)$ with pmf $f(x)$:

$$\mathtt{f} = \frac{\mathrm{e}^{-\mathtt{p}}\ \mathtt{p}^{\mathtt{x}}}{\mathtt{x}\,!};$$

$$\mathtt{domain[f]} = \{\mathtt{x},\ 0,\ \infty\}\ \&\&\ \{\mathtt{p} > 0\}\ \&\&\ \{\mathtt{Discrete}\};$$

where parameter P, rather than being fixed, is instead a random variable $P\ \sim$InverseGaussian(μ, λ):

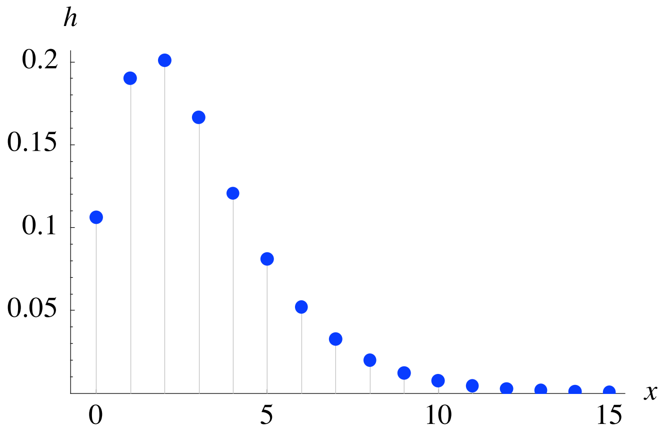
$$\mathtt{g} = \sqrt{\frac{\lambda}{2\,\pi\,\mathtt{p}^3}}\ \mathtt{Exp}\left[-\lambda\ \frac{(\mathtt{p}-\mu)^2}{2\,\mu^2\,\mathtt{p}}\right]; \qquad \mathtt{domain[g]} = \{\mathtt{p},\ 0,\ \infty\}\ \&\&\ \{\mu > 0,\ \lambda > 0\};$$

We seek the parameter-mix distribution $E_P[f(x \mid P = p)]$, and the solution is simply:

```
Expect[f, g]
```

$$\frac{\mathrm{e}^{\lambda/\mu}\ \sqrt{\frac{2}{\pi}}\ \sqrt{\lambda}\ \left(\frac{2}{\lambda} + \frac{1}{\mu^2}\right)^{\frac{1}{4}\,(1-2\,\mathtt{x})}\ \mathtt{BesselK}\left[\frac{1}{2} - \mathtt{x},\ \frac{\sqrt{\lambda\,(\lambda+2\,\mu^2)}}{\mu}\right]}{\mathtt{x}\,!}$$

This is known as Holla's distribution. The diagram below plots the mixture pmf we have just derived, when $\mu = 3$ and $\lambda = 10$. The same approach, of course, applies to deriving any arbitrary desired mixture.



3 The evolving epistemology of statistical knowledge

As in most fields, statistical knowledge has developed in a tree-like manner consisting of main trunks with branches shooting off outwards, slowly filling out knowledge space. Such knowledge trees – whether represented by excellent encyclopaedic texts such as the Johnson, Kotz, Balakrishnan *et al* volumes on statistical distributions, or represented by search engines that index and search published journal papers – all tend to work brilliantly provided that the topic that one is trying to find already exists somewhere on the tree … if so, one can simply pluck the desired fruit from the appropriate branch.

This concept of knowledge has a structure somewhat like a 19thC museum ... that is to say, it works splendidly for classifying your shell collection, but the method fails as soon as one is interested in something that lies outside the tree … something outside what is 'known'. And in the arena of education and research, one is, of course, fundamentally concerned with the quest for new knowledge. Almost by definition, research is the white space that lies *outside* the 'known' tree of knowledge.

The beauty of a symbolic toolkit is that it can escape the 19thC static concept of knowledge – in particular, a symbolic toolkit allows one to construct general algorithms of knowledge, rather than just collecting facts of knowledge. The symbolic approach is a dynamic, live, elegant, flexible and arguably purer concept of knowledge; it is almost the antithesis to the huge collections of tables, volumes, appendices etc that one finds attached to books. This flexibility – this ability to go, as they say in Star Trek, where no-one has gone before – makes the symbolic approach an indispensable part of any toolkit, for anyone in the field of solving new problems.

4 Is that the right answer, Dr Faustus?

As soon as one has a tool that can solve problems that have not been solved before … problems for which no reference exists … the obvious question that arises is: "Is that the right answer?". The joy of computerised problem solving is arguably somewhat Faustian ... the more one uses such tools, the more reliant one becomes on them, and the more dependent one becomes on the accuracy of the software. This is relevant because software is not always infallible! For example, for problems involving integration, symbolic software can sometimes yield solutions that are valid under different assumptions than one intended, or which might simply be plain wrong. Naysayers, Luddites and general followers of the knights who say 'Ni' might immediately respond: "Ni, ni … tis a magic black box – oooer – never trust a black box."

There are at least three replies:

First: books are also black boxes

In case the revelation that symbolic software is not free of error sends some readers running back to trusted reference texts, it should be said that many reference books are essentially extremely primitive black boxes – in effect, manual look-up tables – that equally produce results as if by magic. Even though texts might provide a reference to a source, this cannot remedy that the source itself may sometimes be faulty, or that the result may have been typeset incorrectly, or that the proof may be simply wrong, or that most people do not practically have the time or ability to check a complicated proof should it be provided. All this is just as true for books as it is true for numerical software, as it is indeed for symbolic software. But the symbolic approach has two key advantages, namely …

Second: exact benchmarks

In contrast to numerical black boxes, there exist obvious exact benchmarks for symbolic toolsets. A symbolic toolkit should clearly be able to replicate standard known textbook problems without error. In designing **mathStatica**'s algorithms, the software has been tested against thousands of textbook problems. What is surprising in running such checks is that one discovers that even the most respected reference texts are peppered with errors. As a quick example, consider the k-statistic k_r which is an unbiased estimator of the r^{th} cumulant κ_r; that is, $E[\mathrm{k}_r] = \kappa_r$, for $r = 1, 2, \ldots$. In 1928, Fisher published the product cumulants of the k-statistics, which are now listed in reference bibles such as Stuart and Ord (1994). Here we use **mathStatica** to obtain the solution to the product cumulant $\kappa_{2,2}(\mathrm{k}_3, \mathrm{k}_2)$:

```
CumulantMomentToCumulant[{2, 2}, {KStatistic[3]⟦2⟧, KStatistic[2]⟦2⟧}]
```

$$\frac{288\,n\,\kappa_2^5}{(-2+n)\,(-1+n)^3} + \frac{288\,(-23+10\,n)\,\kappa_2^2\,\kappa_3^2}{(-2+n)\,(-1+n)^3} + \frac{360\,(-7+4\,n)\,\kappa_2^3\,\kappa_4}{(-2+n)\,(-1+n)^3} + \frac{36\,(160-155\,n+38\,n^2)\,\kappa_3^2\,\kappa_4}{(-2+n)\,(-1+n)^3\,n} + \frac{36\,(93-103\,n+29\,n^2)\,\kappa_2\,\kappa_4^2}{(-2+n)\,(-1+n)^3\,n} + \frac{24\,(202-246\,n+71\,n^2)\,\kappa_2\,\kappa_3\,\kappa_5}{(-2+n)\,(-1+n)^3\,n} + \frac{2\,(113-154\,n+59\,n^2)\,\kappa_5^2}{(-1+n)^3\,n^2} + \frac{6\,(-131+67\,n)\,\kappa_2^2\,\kappa_6}{(-2+n)\,(-1+n)^2\,n} + \frac{3\,(117-166\,n+61\,n^2)\,\kappa_4\,\kappa_6}{(-1+n)^3\,n^2} + \frac{6\,(-27+17\,n)\,\kappa_3\,\kappa_7}{(-1+n)^2\,n^2} + \frac{37\,\kappa_2\,\kappa_8}{(-1+n)\,n^2} + \frac{\kappa_{10}}{n^3}$$

By contrast, the solution derived by Fisher in 1928 and which has been continuously published ever since [see Stuart and Ord (1994, eqn 12.70)] turns out to be non-trivially false.

Third: multiple methodologies

The best way to check one's work is to replicate a result via a completely different methodology. Ideally, one should use a different symbolic method. More typically, alternative symbolic methods may not exist or may not be solvable / tractable, in which case a quick numerical or Monte Carlo check is the best practical alternative … Of course, a numerical check does not (and cannot) formally prove that a symbolic solution is correct (other than providing confidence in some statistical sense of 'proof'). But, a numerical check does provide an extremely simple way to prove that a solution is *false* (*i.e.* that something *has* gone horribly wrong). Numerical checks act like a filtering system that eliminates false answers. To illustrate the idea, let us suppose that $X \sim$ Chi-squared(n) with pdf $f(x)$:

$$\mathbf{f} = \frac{\mathbf{x}^{\mathbf{n/2-1}}\,\mathbb{e}^{\mathbf{-x/2}}}{\mathbf{2}^{\mathbf{n/2}}\,\Gamma[\frac{\mathbf{n}}{\mathbf{2}}]}; \qquad \mathbf{domain[f] = \{x, 0, \infty\}\ \&\&\ \{n > 0\};}$$

We wish to find the mean deviation $E\left[\,\left|\,X-\mu\,\right|\,\right]$. Here is the solution derived with **mathStatica**:

```
μ = Expect[x, f];        sol = Expect[Abs[x - μ], f]
```

$$\frac{4\,\mathrm{Gamma}[1+\frac{n}{2},\frac{n}{2}] - 2\,n\,\mathrm{Gamma}[\frac{n}{2},\frac{n}{2}]}{\Gamma[\frac{n}{2}]}$$

By contrast, if we refer to an excellent reference text like Johnson *et al.* (1994, p. 420), the mean deviation is listed as:

$$\mathbf{JKBsol} = \frac{\mathbb{e}^{-\frac{\mathbf{n}}{\mathbf{2}}}\,\mathbf{n}^{\mathbf{n/2}}}{\mathbf{2}^{\frac{\mathbf{n}}{\mathbf{2}}-\mathbf{1}}\,\Gamma[\frac{\mathbf{n}}{\mathbf{2}}]};$$

Let us check if the two solutions are the same, by choosing a value for *n*, say *n* = 6:

```
{sol, JKBsol} /. n → 6.
```

```
{2.6885, 1.34425}
```

Clearly, at least one of the solutions is wrong! Since our original attempt was *symbolic*, we shall now use a *numerical* method to calculate the answer when *n* = 6. Here is the mean deviation as a numerical integral when *n* = 6:

```
NIntegrate[ (Abs[x - μ] f) /. n → 6., {x, 0, ∞}]
```

```
2.6885
```

The numerical answer suggests that **mathStatica**'s symbolic solution is correct, and it proves that the known textbook solution is false. Further experimentation reveals that the solution given in Johnson *et al.* is out by a factor of two. This again highlights how important it is to check all output, from both reference books and computers.

CODA

This paper sets out to provide a tasting of how a symbolic toolkit can benefit the learning and research experience. Symbolic tools make it dramatically easier to solve problems for which they are intended. But, more importantly, symbolic tools change what one can do, and how one does it – moving the research process from a laborious mechanical path to an almost real-time exploration of the unknown. When automated tools can solve new problems that have never been solved before, they give rise to an evolving epistemology of statistical knowledge … a shift from tables, reference books and databases towards a live, flexible, algorithmic concept of knowledge.

REFERENCES

Fisher, R. A. (1928), Moments and product moments of sampling distributions, *Proceedings of the London Mathematical Society*, series 2, volume 30, 199–238 (reprinted in Fisher, R. A. (1950), *Contributions to Mathematical Statistics*, Wiley: New York).

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994), *Continuous Univariate Distributions*, volume 1, 2nd edition, Wiley: New York.

mathStatica (2002-2007), www.mathStatica.com, Sydney.

Rose, C. and Smith, M. D. (2002), *Mathematical Statistics with Mathematica*, Springer-Verlag: New York.

Stuart, A. and Ord, J. K. (1994), *Kendall's Advanced Theory of Statistics*, volume 1, 6th edition, Edward Arnold: London (also Wiley: New York).

RESUMÉ

Colin Rose is director of the Theoretical Research Institute, Sydney. He holds a PhD in economic theory. He is a past visiting scholar at Wolfram Research. His current area of research is exact computational methods in mathematical statistics, in particular, with application to the **mathStatica** project. He is an associate editor of the *Journal of Statistical Software*, a recent guest editor of the *Mathematica Journal*, and an associate editor of *Mathematica in Education and Research.*