Journal of Statistical Software

April 2012, Volume 47, Software Review 1. *http://www.jstatsoft.org/*

Reviewer: Barrie Stokes
University of Newcastle

mathStatica 2.5

> **mathStatica** Pty. Ltd., Sydney, Australia. USD 285 (standard), USD 127 (academic), USD 55 (student).
> `http://www.mathStatica.com/`

Introduction

mathStatica is an add-on package for the computer algebra system Mathematica (Wolfram 2003), designed to enable the user to carry out symbolic – and numeric and graphic – calculations in mathematical statistics.

The first version of **mathStatica** was bundled with the Springer hard cover book (Rose and Smith 2002); the current version is 2.5 released in late 2011. Perhaps representing the shift from printed to electronic publishing, the software now comes bundled with the new 2011 electronic edition (Rose and Smith 2011) that is integrated into same.

Key new features of **mathStatica** 2.5 include:

- A new parallel processing engine for greater speed for users with multi-processor machines.
- A new piecewise processing engine providing support for piecewise functions.
- New palettes for nearly 100 predefined continuous and discrete distributions.
- Progress indicators (shows progress for complicated problems).
- Random number generation for essentially any discrete or continuous univariate distribution.
- Enhanced support for non-rectangular domains.
- Extended support for multivariate discrete distributions.

mathStatica combines a modern e-book on mathematical statistics with a software toolkit built on top of Mathematica. The key focus with **mathStatica** is working with arbitrary user-defined univariate and multivariate distributions: it provides automated functions that enable one to find expectations, derive transformations of random variables, find moments, cumulative distribution functions, characteristic functions, order statistics, find Fisher information (with automatic selection of method, depending on regularity conditions), calculate correlations between random variables, marginal distributions, derive the probability density function (PDF) of $\max(X, Y, Z, \ldots)$ and $\min(X, Y, Z, \ldots)$ and so on.

It also embodies extensive algorithmic 'knowledge' about raw and central moments and cumulants and their inter-conversion, the derivation of unbiased estimators, the Johnson and Pearson distributions systems, and the problem of moments of moments (e.g., find the covariance between say the sample mean and the sample variance). **mathStatica** also supports exact kernel density estimation, and there is also a variety of useful statistical plots which allow comparison of data and fitted distributions, for example.

Importantly, **mathStatica** does not provide pre-programmed answers (like one might find in a textbook appendix), rather it provides general algorithms for solving arbitrary problems 'live', rather than hard-coding special known solutions in the way that most software packages usually proceed. This general functionality is possible because **mathStatica** is built on top of a computer algebra system (Mathematica).

There is a very informative **mathStatica** home page at `http://www.mathStatica.com/`, from which the product may be obtained. Prices range from USD 55 for students (single-user) to USD 97 (for academics), or USD 127 for academics (with solutions to the problems in the e-book), to USD 285 for commercial/government users. Pricing includes the software and the Rose and Smith (2011) eBook.

The coverage of **mathStatica** can be indicated by listing the 12 chapter headings in the accompanying e-book 'Mathematical Statistics with Mathematica' which is fully cross-linked, and integrated into Mathematica's on-line help during the installation process. The chapters are *Introduction*, *Continuous Random Variables*, *Discrete Random Variables*, *Distributions of Functions of Random Variables*, *Systems of Distributions*, *Multivariate Distributions*, *Moments of Sampling Distributions*, *Asymptotic Theory*, *Statistical Decision theory*, *Unbiased Parameter Estimation*, *Maximum Likelihood Estimation: Theory*, and *Maximum Likelihood Estimation: Practice*. 'Mathematical Statistics with Mathematica' is a complete textbook, which incorporates all the supporting mathematical theory for the material covered, closely integrated with the Mathematica syntax for carrying out the corresponding symbolic calculations.

An appendix includes sections titled *Is That the Right Answer, Dr Faustus?*, *Working with Packages*, *Working with =, ->, == and :=*, *Working with Lists*, *Working with Subscripts*, *Working with Matrices*, *Working with Vectors*, *Changes to Default Behaviour*, and *Building Your Own **mathStatica** Function*. Following the appendix is a notes chapter which contains all the notes referred to in the main chapters; within a chapter a link in the form of a numerical superscript opens a small window with the referenced information. There is an extensive references section with 167 entries. The final section is the index, which is fully hyperlinked to the rest of the e-book and has a handy A–Z list of navigating links at the top. Each of the over 1300 entries has one or more links which open the corresponding sections of the e-book. For any topic covered, the relevant book section – with full theory and coded examples – is just a click away.

After installation, **mathStatica** automatically appears in Mathematica's list of built-in palettes, making access and startup simple and easy. (Figure 1).

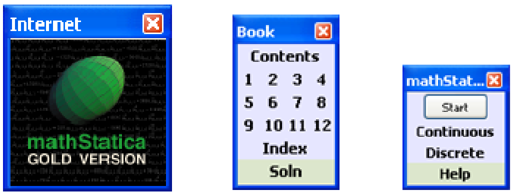


Figure 1: **mathStatica** palettes.

mathStatica in action

Note that although **mathStatica** provides easy palette access to about 100 continuous, discrete, and multivariate distributions, the key point is to enable one to work with arbitrary distributions ...that is, to help solve problems that may not be in text books or journal articles, or that may never have been solved before.

mathStatica adds around 100 new function to Mathematica, but the 'core' functions are `PlotDensity`, `Expect`, `Prob`, and `Transform`. `PlotDensity` does a nice job of plotting continuous (even piecewise) and discrete PDFs and cumulative distribution functions (CDFs), with automatic axis labels and scaling.

Several PDFs with different parameter values can be plotted together, and if the domain changes with these values, `PlotDensity` automatically takes this into account.

`Expect` is the very general and powerful expectation operator, while the syntax `Prob[x,f]` calculates the CDF for any PDF f, continuous (including piecewise), discrete, or multivariate. `Transform[eqn,f]`, in conjunction with `TransformExtremum[eqn,f]`, allows the user to find the PDF of a new random variable Y as a function of X, e.g., `Y=Sqrt[X]`, once the PDF (and domain) of X has been defined. Again, X can be continuous, discrete, or multivariate.

All these functions take full advantage of the assumptions technology in Mathematica which enables constraints on parameters to be taken into account when deriving mathematical results. We start with some examples of the use of the 'core' functions `Expect`, and `Prob`.

Core functions

Suppose the random variable X is trapezoidal with piecewise PDF $f(x)$:

$$\text{In[1]:=}\ \ \mathbf{f} = \begin{cases} \frac{x}{20} & 0 < x < 5 \\ \frac{1}{4} & 5 \le x < 6 \\ \frac{7-x}{4} & 6 \le x < 7 \\ 0 & \text{True} \end{cases}; \qquad \text{domain[f] = \{x, 0, 7\};}$$

The mean of this PDF is calculated via:

```
In[2]:= Expect[x, f]
```

$$\text{Out[2]=}\ \frac{17}{4}$$

Here is the expectation of X^3:

```
In[3]:= Expect[x^3, f]
```

Out[3]= $\frac{4203}{40}$

This is the variance of X^2:

```
In[4]:= Var[x^2, f]
```

Out[4]= $\frac{107\,083}{720}$

With `Prob`, the derivation of the CDF is immediate:

```
In[5]:= cdfF = Prob[x, f]
```

Out[5]=
$$\begin{cases} 1 & x \geq 7 \\ \frac{1}{8}\,(-41-(-14+x)\,x) & 6 < x < 7 \\ \frac{1}{8}\,(-5+2\,x) & 5 < x \leq 6 \\ \frac{x^2}{40} & 0 < x \leq 5 \\ 0 & \text{True} \end{cases}$$

The entropy of the distribution with PDF f is given by:

```
In[6]:= Expect[-Log[f], f]
```

Out[6]= $\frac{3}{8} + \text{Log}[4]$

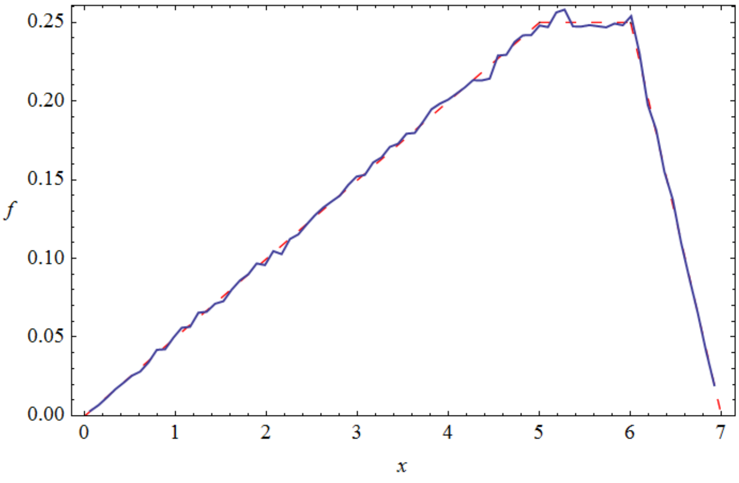
This input generates 200,000 pseudorandom samples from this trapezoidal distribution, taking around one second to execute on a 4-core workstation running Windows 7:

```
In[7]:= data = RandomNumber[200 000, f];
```

mathStatica's `FrequencyPlot` suggests that this data represents a fair sample from the given distribution. A graphical comparison of the density and the sampled data is shown in below:

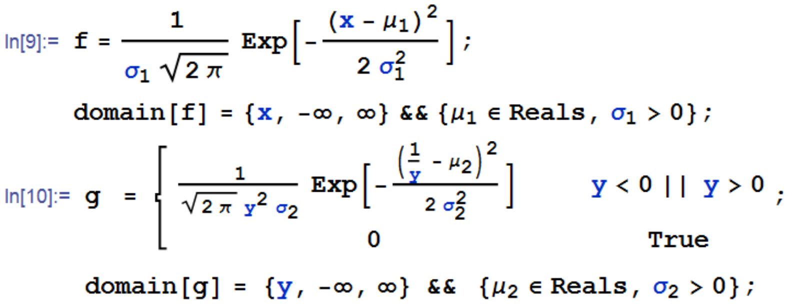
```
In[8]:= FrequencyPlot[data, f]
```

Out[8]=



`TransformProduct`

Let random variable X have PDF $f(x)$, and let random variable Y have PDF $g(y)$:



In[9]:= $$\mathbf{f} = \frac{1}{\sigma_1\sqrt{2\pi}}\ \mathrm{Exp}\left[-\frac{(\mathbf{x}-\mu_1)^2}{2\,\sigma_1^2}\right];$$

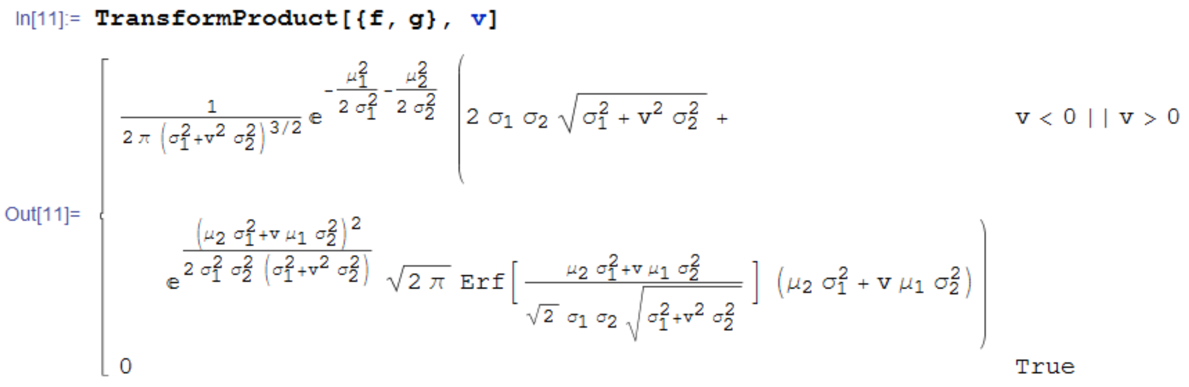
$$\mathrm{domain}[\mathbf{f}] = \{\mathbf{x},\ -\infty,\ \infty\}\ \&\&\ \{\mu_1 \in \mathrm{Reals},\ \sigma_1 > 0\};$$

In[10]:= $$\mathbf{g} = \begin{cases} \frac{1}{\sqrt{2\pi}\,\mathbf{y}^2\,\sigma_2}\ \mathrm{Exp}\left[-\frac{\left(\frac{1}{\mathbf{y}}-\mu_2\right)^2}{2\,\sigma_2^2}\right] & \mathbf{y} < 0\ ||\ \mathbf{y} > 0 \\ 0 & \mathrm{True} \end{cases};$$

$$\mathrm{domain}[\mathbf{g}] = \{\mathbf{y},\ -\infty,\ \infty\}\ \&\&\ \{\mu_2 \in \mathrm{Reals},\ \sigma_2 > 0\};$$

In this example, we show how **mathStatica**'s new `TransformProduct` function solves a celebrated problem (Tooze, Košmelj, and Blejec 2004; Fieller 1932; Geary 1930; Hayya, Armstrong, and Gressis 1975; Hinkley 1969; Kamerud 1978; Marsaglia 1965, 2006; Pham-Gia, Turkkan, and Marchand 2006; Rose and Smith 2002): What is the PDF of the ratio of two general normal distributions, $N(\mu_1,\,\sigma_1)$ and $N(\mu_2,\,\sigma_2)$? To see this, note that density $g(y)$ is the PDF of $Y = 1/Z$ where $Z \sim N(\mu_2,\,\sigma_2)$. Notice also how the domain statement for g allows Y to vary from $-\infty$ to $+\infty$, while the piecewise definition specifies that Y cannot be zero (`||` is the Mathematica syntax for the logical 'Or', i.e., $\vee$).

The PDF of $V = X \cdot Y$ is calculated via: (this takes about 30 seconds)

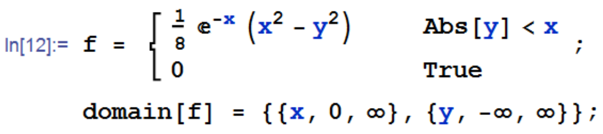


In[11]:= `TransformProduct[{f, g}, v]`

Out[11]= $$\begin{cases} \frac{1}{2\pi\left(\sigma_1^2+v^2\sigma_2^2\right)^{3/2}}\, e^{-\frac{\mu_1^2}{2\sigma_1^2}-\frac{\mu_2^2}{2\sigma_2^2}} \left(2\,\sigma_1\,\sigma_2\sqrt{\sigma_1^2+v^2\,\sigma_2^2} + e^{\frac{\left(\mu_2\,\sigma_1^2+v\,\mu_1\,\sigma_2^2\right)^2}{2\,\sigma_1^2\,\sigma_2^2\left(\sigma_1^2+v^2\,\sigma_2^2\right)}}\sqrt{2\pi}\ \mathrm{Erf}\left[\frac{\mu_2\,\sigma_1^2+v\,\mu_1\,\sigma_2^2}{\sqrt{2}\,\sigma_1\,\sigma_2\sqrt{\sigma_1^2+v^2\,\sigma_2^2}}\right]\left(\mu_2\,\sigma_1^2+v\,\mu_1\,\sigma_2^2\right)\right) & v < 0\ ||\ v > 0 \\ 0 & \mathrm{True} \end{cases}$$

Non-rectangular domains

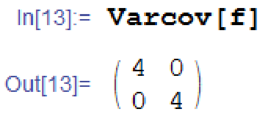
Let the random variables X and Y have joint PDF $f(x,y)$:



In[12]:= $$\mathbf{f} = \begin{cases} \frac{1}{8}\,e^{-\mathbf{x}}\left(\mathbf{x}^2-\mathbf{y}^2\right) & \mathrm{Abs}[\mathbf{y}] < \mathbf{x} \\ 0 & \mathrm{True} \end{cases};$$

$$\mathrm{domain}[\mathbf{f}] = \{\{\mathbf{x},\ 0,\ \infty\},\ \{\mathbf{y},\ -\infty,\ \infty\}\};$$

mathStatica quickly calculates the variance-covariance matrix for this distribution:



In[13]:= `Varcov[f]`

Out[13]= $$\begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}$$

The marginal distribution of Y is found with

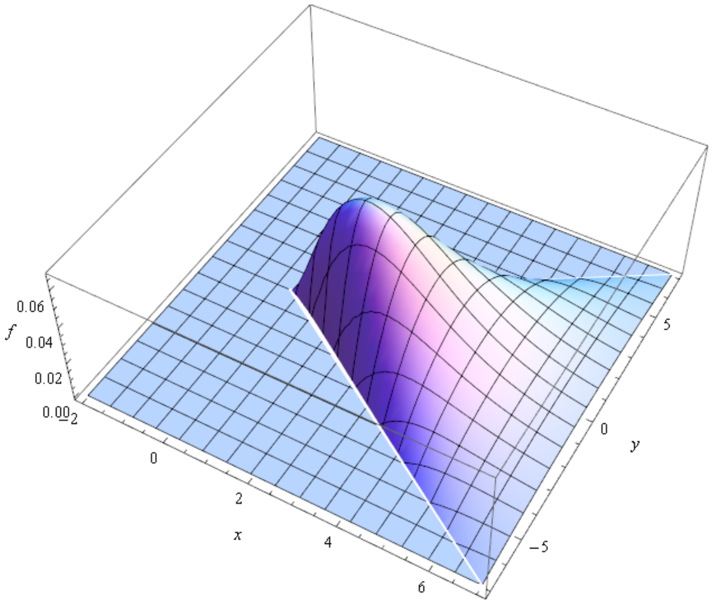


Figure 2: The PDF $\frac{1}{8}e^{-x}(x^2 - y^2)$, $|y| < x$.

```
In[14]:= Marginal[y, f]

         ⎧ - 1/4 ⅇ^y (-1 + y)    y ≤ 0
Out[14]= ⎨
         ⎩ 1/4 ⅇ^-y (1 + y)      True
```

After some deeper thought, **mathStatica** finds the probability that the Euclidean distance from the origin to a random sample is less than 3:

```
In[15]:= Prob[√(x^2 + y^2) < 3, f] // N

Out[15]= 0.301376
```

Figure 2 shows the PDF on its domain of support, $|y| < x$.

Full precision through exact symbolic methods

An important feature of **mathStatica** is that only exact analytic methods are used throughout; there is never any recourse to numerical approximation. This example (see Rose and Smith 2011, Section 5.5) highlights the difference between an exact solution to the kernel density equation, and an approximation via a numerical interpolation.

Suppose we generate some pseudo-random data consisting of 100,000 samples from a Cauchy distribution Cauchy(a, b) with location parameter $a = 0$ and scale parameter $b = 150$, and then 500 samples from the normal distribution $\mu = 20$, and $\sigma = 1/4$:

```
In[16]:= data = Join[RandomReal[CauchyDistribution[0, 150], 100 000],
             RandomReal[NormalDistribution[20, 1/4], 500]];
```

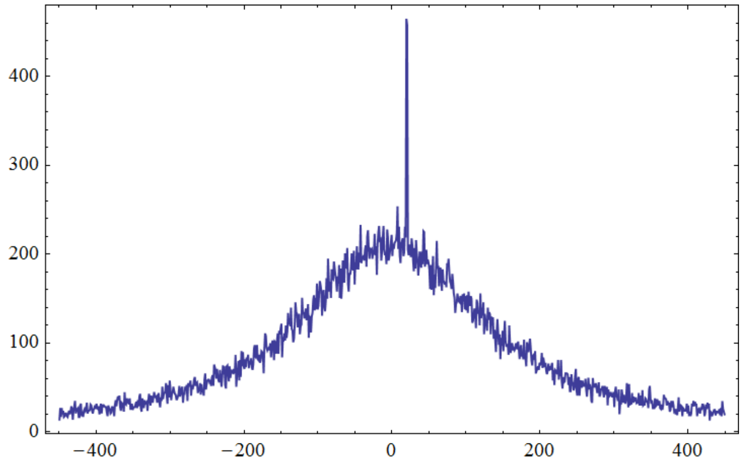


Figure 3: 100,000 Cauchy samples and 500 normal samples.

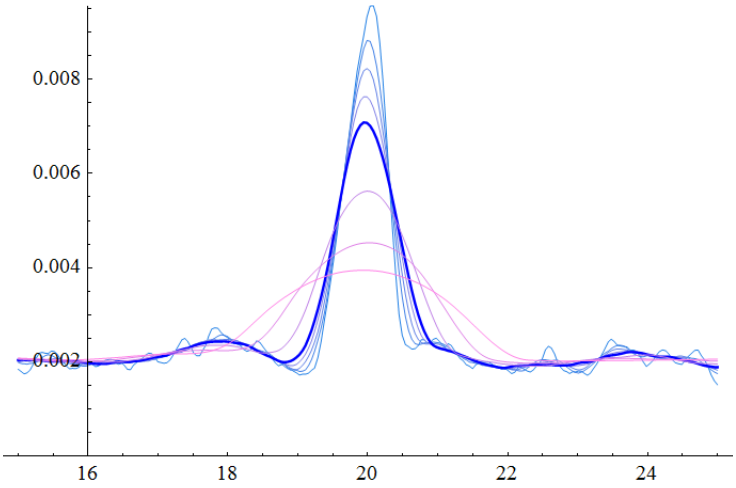


Figure 4: NPKDEPlot smoothed non-parametric kernel density estimate plots (Epanechnikov kernels).

In Figure 3, **mathStatica**'s FrequencyPlot shows the empirical PDF, with the narrow 'spike' at $X = 20$ due to the normal samples. (Plotting over a range of $\pm 3b$ shows 79.5% of the Cauchy sample.)

Rose and Smith (2011), Section 5.5 use this data set to plot a smoothed non-parametric kernel density estimate and investigate the spike. Using **mathStatica**'s NPKDEPlot function with a range of possible bandwidths, we plot over (15, 25) since our interest is in the spike at $X = 20$. For the family of different bandwidths considered, the corresponding family of smoothed non-parametric kernel density estimate plots (using the default Epanechnikov kernel) is generated by:

```
In[18]:= bandwidths = {.2, .35, .45, .55, .65, 1, 1.5, 2};
         NPKDEPlot[data, bandwidths, {15, 25}]
```

In Figure 4 **mathStatica**'s NPKDEPlot shows a range of kernel density estimates for the indi-

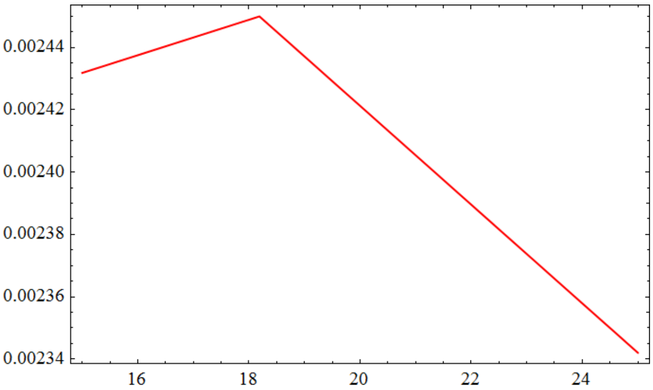


Figure 5: `SmoothKernelDistribution` estimate (Epanechnikov kernel).

cated bandwidths.

```
In[20]:= dist = SmoothKernelDistribution[data, 1, "Epanechnikov"]

Out[20]= DataDistribution[≪SmoothKernel≫, {100 500}]
```

Mathematica 8 has a new `SmoothKernelDistribution` function. Whereas **mathStatica**'s `NPKDEPlot` function is an exact solution to the kernel density equation, `SmoothKernelDistribution` only approximates the solution and, in this instance, it approximates it very poorly indeed, completely failing to pick up the spike at $X = 20$ ...

Figure 5 shows the Wolfram solution to the same plot (bandwidth $= 1$).

Maximum and minimum

mathStatica has new `Maximum` and `Minimum` functions for deriving the PDF of $\max(X, Y, Z, \ldots)$ or $\min(X, Y, Z, \ldots)$. Here are three different distributions defined over three different domains of support. We have $X \sim \text{Triangular}(1/2, 1/2)$ with PDF $f(x)$, $Y \sim \text{Uniform}(1/2, 2/3)$ with PDF $g(y)$, and $Z \sim \text{half} - \text{Halo}$ with PDF $h(z)$:

```
In[22]:= f  =  ⎧ 4/3 (2 - x)        1 < x < 2
               ⎨ 8/3 (x - 1/2)      1/2 < x ≤ 1 ;          domain[f] = {x, 1/2, 2};
               ⎩ 0                  True

         g = 1;                                            domain[g] = {y, 1/2, 3/2};

         h = 1/π 2 √(1 - (z - 2)^2);                       domain[h] = {z, 1, 3};
```

We want to derive the PDF of $W = \max(X, Y, Z)$. The solution PDF, say $\phi(w)$, is obtained via:

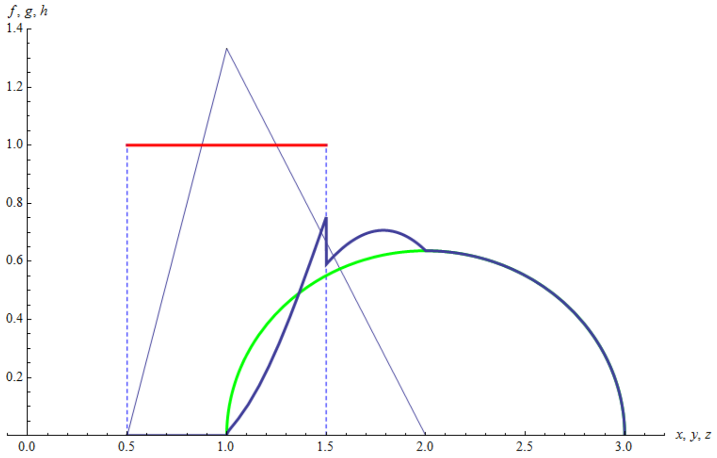
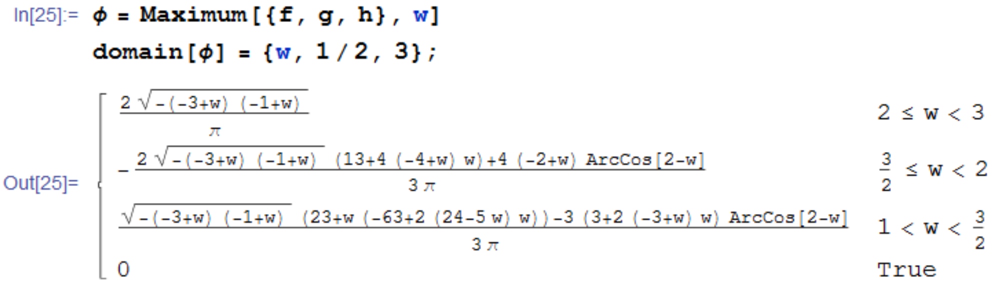


Figure 6: The PDFs f, g, and h with the PDF of $\max(f, g, h)$.



```
In[25]:= ϕ = Maximum[{f, g, h}, w]
         domain[ϕ] = {w, 1 / 2, 3};
```

$$\text{Out[25]=}\begin{cases} \frac{2\sqrt{-(-3+w)\ (-1+w)}}{\pi} & 2 \leq w < 3 \\ -\frac{2\sqrt{-(-3+w)\ (-1+w)}\ \ (13+4\ (-4+w)\ w)+4\ (-2+w)\ \text{ArcCos}[2-w]}{3\ \pi} & \frac{3}{2} \leq w < 2 \\ \frac{\sqrt{-(-3+w)\ (-1+w)}\ \ (23+w\ (-63+2\ (24-5\ w)\ w))-3\ (3+2\ (-3+w)\ w)\ \text{ArcCos}[2-w]}{3\ \pi} & 1 < w < \frac{3}{2} \\ 0 & \text{True} \end{cases}$$

In Figure 6 the PDF of the maximum $\phi(w)$ is plotted together with the underlying PDFs, using `PlotDensity`.

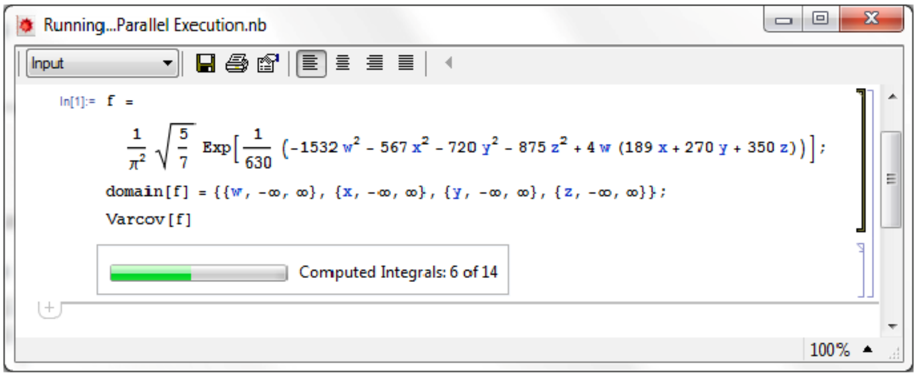
Parallel processing speed-up

mathStatica's new parallel engine will automatically takes advantage of however many cores are available on the host machine. Usually, parallel computing is used to speed-up numerical problems. By contrast, **mathStatica** uses parallel-processing power not only for numerical problems, but for symbolic ones.

Table 1 shows computing times for the variance-covariance matrix of a certain four-variate distribution, with different numbers of processors. The 4×4 variance-covariance matrix requires the calculation of 14 separate symbolic integrals which can be computed in parallel. (This timing test was done on a 2.13 Ghz Intel Xeon PC running Windows 7, with 8 GB of memory).

No. of processors	Absolute times
1	144.66
2	83.75
3	62.39
4	51.43

Table 1: Timing results (in seconds).



In[1]:= f =

$$\frac{1}{\pi^2}\sqrt{\frac{5}{7}}\,\mathrm{Exp}\left[\frac{1}{630}\left(-1532\,w^2-567\,x^2-720\,y^2-875\,z^2+4\,w\,(189\,x+270\,y+350\,z)\right)\right];$$

```
domain[f] = {{w, -∞, ∞}, {x, -∞, ∞}, {y, -∞, ∞}, {z, -∞, ∞}};
Varcov[f]
```

Computed Integrals: 6 of 14

Figure 7: Progress bars.

While this parallel calculation is in progress, incidentally, **mathStatica** thoughtfully displays a progress bar to reassure the user that something is happening! (The screen shot in Figure 7 shows the PDF and domain of the timing test distribution.)

Conclusion

Perhaps the single most important feature of **mathStatica** – which it inherits from Mathematica – is that it is general by design. That is, **mathStatica**'s functions will work on any arbitrary univariate, discrete, or multivariate distribution – defined via its PDF and domain (illustrated several times above) – just as well as on the 100 or so distributions already available with a click on a palette. An inventive programmer can also write new statistical procedures using those provided in **mathStatica** as building blocks.

As a Mathematica user since 1992, and a **mathStatica** user since version 1.0, I find that the "look-and-feel-and-operation" of **mathStatica** is completely consistent with that of Mathematica, which for me is a huge plus. The online help and the e-book are beautifully integrated and cross-linked, and work just like Mathematica's help. The conciseness and expressive power of the Mathematica language, the fine control over graphics, the piecewise function capability, and the gigaNumerics functionality providing great speed and precision for working with real numbers, are all at the user's fingertips.

Most importantly, there is security in being able to carry out totally integrated symbolic computation providing exact results into which numerical values can be substituted if required. Of course, not all problems have closed form solutions, and Mathematica cannot solve all integrals that do have closed form solutions (and none that don't!). If it could, we Bayesian statisticians would have no need for MCMC (Markov chain Monte Carlo) methods to handle very complex models. Even so, **mathStatica** is still a very useful tool to have at one's disposal.

Quibbles? Very few. Sometimes **mathStatica** takes some time to generate an answer, but that's usually because behind the scenes one or more highly non-trivial symbolic integrations are being carried out to find a variance-covariance matrix, or derive a conditional or marginal distribution for an exotic multivariate distribution. In such cases, there is the progress bar to keep you posted. The JohnsonSB entry in the on-line help page is the only one to give a message like "`JohnsonSB[...]` is not currently implemented.", so the promise is there.

Bayesians like myself will note that there is nothing specifically Bayesian in either the functionality or documentation of **mathStatica**, but that said the system is perfectly capable of carrying out many Bayesian analyses, again up to the point where calculating posterior distributions involves integrations that cannot be done analytically – a difficulty which is inherent in the mathematics, and not a particular shortcoming of Mathematica!

I did notice that the plots produced by `FrequencyPlot` and `FrequencyGroupPlot` are automatically `Frame`-d, whereas those generated by `FrequencyPlotDiscrete` are not. Everything else works as advertised, and the whole system has a very impressive integrity.

When combined with the symbolic computational (and numerical and graphical) power of Mathematica itself, **mathStatica** offers the mathematical statistician – and any scientist interested in the "nuts-and-bolts" of basic and advanced statistical ideas – both a first class exposition in the form of the accompanying e-book, and a flexible and highly developed system for carrying out theoretical and numerical statistical manipulations.

References

Fieller EC (1932). "The Distribution of the Index in a Normal Bivariate Population." *Biometrika*, **24**(3/4), 428–440.

Geary RC (1930). "The Frequency Distribution of the Quotient of Two Normal Variates." *Journal of the Royal Statistical Society*, **93**(3), 442–446.

Hayya J, Armstrong D, Gressis N (1975). "A Note on the Ratio of Two Normally Distributed Variables." *Management Science*, **21**(11), 1338–1341.

Hinkley DV (1969). "On the Ratio of Two Correlated Normal Random Variables." *Biometrika*, **56**(3), 635–639.

Kamerud D (1978). "The Random Variable X/Y, X, Y Normal." *The American Mathematical Monthly*, **85**(3), 207–207.

Marsaglia G (1965). "Ratios of Normal Variables and Ratios of Sums of Uniforms Variables." *Journal of the American Statistical Association*, **60**, 193–204.

Marsaglia G (2006). "Ratios of Normal Variables." *Journal of Statistical Software*, **16**(4), 1–10. URL http://www.jstatsoft.org/v16/i04/.

Pham-Gia T, Turkkan N, Marchand E (2006). "Density of the Ratio of Two Normal Random Variables and Applications." *Communications in Statistics – Theory and Methods*, **35**(9), 1569–1591.

Rose C, Smith MD (2002). *Mathematical Statistics with Mathematica*. Springer-Verlag.

Rose C, Smith MD (2011). *Mathematical Statistics with Mathematica*. **mathStatica** Pty. Ltd., Sydney.

Tooze A, Košmelj K, Blejec A (2004). "The Distribution of the Ratio of Jointly Normal Variables." *Metodološki zvezki*, **1**(1), 99–108.

Wolfram S (2003). *The Mathematica Book*. 5th edition. Wolfram Media. URL http://www.wolfram.com/.

Reviewer:

Barrie Stokes
University of Newcastle
Discipline of Clinical Pharmacology and Toxicology
Newcastle, NSW, Australia
E-mail: barrie.stokes@newcastle.edu.au
URL: http://www.newcastle.edu.au/staff/research-profile/Barrie_Stokes/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 47, Software Review 1 *Published:* 2012-04-02
April 2012