Technical report on fitting an inverse normal distribution

The cumulative distribution function (cdf) of the inverse normal covers a wide range of possible shapes that make it appealing for modelling cumulative germination curves. Figure 1 shows a range from rapid to symmetric, normal looking germination patterns.

Approximations to the inverse normal distribution function

No closed form exists for the cdf of the inverse normal, except for the relationship given by Eq (2) of the paper.

$$F(t;\tau,\mu,\beta)=\Phi\left(\frac{\sqrt{\beta}\,(t-\tau-\mu)}{\mu\sqrt{t-\tau}}\right)+e^{2\phi}\;\;\Phi\left(-\frac{\sqrt{\beta}\,(t-\tau+\mu)}{\mu\sqrt{t-\tau}}\right)\qquad t\geq\tau. \qquad (2)$$

The contribution from the second half of the approximation to IG(t;τ,μ,λ) given by (2) can sometimes be incorrect simply because of the size of λ/μ. For example, Excel would return an approximation of 0.5 for IG(75;0,75,7500), rather than the correct value of 0.519898, simply because it cannot handle the evaluation of e^{200} Φ(-20;0;1). The accuracy of (2) therefore depends on the coefficient of variation $\varphi=\sqrt{\mu/\lambda}$. The following is a simplification of (2) using a standard approximation to Φ(t) which allows the elimination of the $e^{2\lambda/\mu}$ term:

$$\mathrm{P}(T\leq t)\approx 1-\phi(t_-)\sum_{i=1}^{5}b_i(q_-^i-q_+^i)\,,\qquad \text{if } t-\tau-\mu\geq 0,$$

$$\mathrm{P}(T\leq t)\approx \phi(t_-)\sum_{i=1}^{5}b_i(q_-^i+q_+^i)\,,\qquad \text{if } t-\tau-\mu\leq 0, \qquad (TR1)$$

where ϕ(u) is the standard normal density function evaluated at u,

$$q_-=\frac{1}{1+0.2316419\,t_-}\,,\qquad t_-=\sqrt{\frac{\lambda}{(t-\tau)}}\,\frac{|t-\tau-\mu|}{\mu}\,,$$

$$q_+ = \frac{1}{1 + 0.2316419\, t_+}, \qquad t_+ = \sqrt{\frac{\lambda}{(t-\tau)}}\,\frac{|t-\tau+\mu|}{\mu},$$

and

$b_1 = 0.31938153$
$b_2 = -0.356563782$
$b_3 = 1.781477937$
$b_4 = -1.821255978$
$b_5 = 1.330274429.$

Using (A2.1) we obtain IG(75;0,75,7500) ≈ 0.521485, an over-estimate by 0.001588. Equation (A2.1) is based on an approximation for a standard normal (Abramowitz and Stegun, 1972, 26.2.17) which is accurate to 7.5×10^{-8} and which has certain near-optimal properties. When applied to both parts of (2), however, a different pattern arises. Figure 2 shows the error, defined as (approximate probability - exact probability), for the inverse normal distribution function with $\tau = 0$, $\mu = 75$ and $\lambda = 1000$. For this distribution the coefficient of variation of T is 27.4%. It is clear that the maximum error (0.000405) occurs at $t = \mu$, which is unusual for approximations of this kind.

As we have seen, however, the error can be larger than this. The exact value for $P(T \le \mu)$ can be approximated, using a Taylor expansion, quite accurately by

$$P(T \le \mu) \;=\; 0.5 + \frac{\varphi}{2\sqrt{2\pi}}\left(1 - \frac{\varphi^2}{4} + \frac{3\varphi^4}{16} - \frac{15\varphi^6}{64} + \frac{105\varphi^8}{256}\right) \qquad \text{(TR2)}$$

when $\varphi < 1$. Using this as an accurate approximation, we can show that, at $t = \mu$ (TR1) has a maximum error of 0.001744, and this occurs when $\varphi = 0.065436$. While the approximation is often extremely good, a potential error of 0.001744 could lead to inefficient ML estimates and to larger deviances than are necessarily the case.

We also investigated approximations based on Gaussian quadrature (see Abramowitz and Stegun, 1970, 25.4.30 and Table 25.4) and found that the 16-point method offers an extremely good general

approximation. For example, this method gave IG(75;0,75,7500) ≈ 0.519898; using increased precision, this represents an over-estimate by 4.5×10^{-7}. Numerical integration in packages such as Mathematica use some form of Gaussian quadrature, usually dividing the region of integration into many small intervals recursively until some an appropriate precision goal is met. A single use of Gaussian quadrature would therefore be expected to work well for approximating $P(t_j \le T \le t_{j-1})$ for small time intervals. We also investigated a single application of the16-point method to approximate $IG(t;\tau,\mu,\lambda)$ for $t \le \mu$. For $t>\mu$ we replaced $P(T\le t)$ by $P(T\le\mu) + P(\mu<T\le t)$. The approximation was extremely good: we obtained a maximum error of 4.88×10^{-7} for $\mu = 75$, $\lambda = 7500$ and of 7.86×10^{-9} $\mu = 75$, $\lambda = 1000$. Figure 3 a plot of the error function for the latter example.

The Excel spreadsheet available on our website uses 16-point Gaussian quadrature in a user-defined function to evaluate $P(a < T \le b)$ for the inverse normal distribution. The code is given in Appendix 3.

Appendix 1 and 2 provide additional technical information to the paper.

Time to 50% germination

Clearly an initial solution for the median of an inverse Gaussian can be taken as the mean, $t_{50}^{(0)} = \hat{\tau} + \hat{\mu}$.
Using Newton-Raphson iteration, a first-order approximation is $t_{50}^{(1)} = \hat{\tau} + \hat{\mu} - \dfrac{F(\hat{\tau}+\hat{\mu};\hat{\mu},\hat{\tau},\hat{\lambda}) - 0.5}{f(\hat{\tau}+\hat{\mu};\hat{\mu},\hat{\tau},\hat{\lambda})}$.
Using Eq (2) we obtain

$$t_{50}^{(1)} = \hat{\tau} + \hat{\mu}\left(1 - e^{2\hat{\phi}}\Phi(-2\sqrt{\hat{\phi}};0,1)\sqrt{2\pi/\hat{\phi}}\right).$$

This solution is usually adequate; however further improvement can be obtained by iteration. However, the difference between $t_{50}^{(1)}$ and $t_{50}^{(2)}$ (and successive differences) is usually very small, so we adopt s.e.($t_{50}^{(1)}$) for the approximate standard error of $\hat{t}_{50}$.

$$\text{var}(\hat{t}_{50}^{(1)}) = \mathbf{d}^T\mathbf{V}\mathbf{d} = \sum_{i=1}^{3} d_i^2 V_{ii} + 2\sum_{i<j}^{3} d_i d_j V_{ij}$$

where $\mathbf{d} = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix}$,

$d_1 = 1,$

$$d_2 = \frac{\hat{\mu}}{\hat{\lambda}}\left[1-\sqrt{2\pi}\left(2\sqrt{\frac{\hat{\lambda}}{\hat{\mu}}}-\frac{1}{2}\sqrt{\frac{\hat{\mu}}{\hat{\lambda}}}\right)e^{2\frac{\lambda}{\mu}}\Phi\left(-2\sqrt{\frac{\hat{\lambda}}{\hat{\mu}}}\right)\right],$$

$$d_3 = \sqrt{2\pi}\left[2\sqrt{\frac{\hat{\lambda}}{\hat{\mu}}}-\frac{3}{2}\sqrt{\frac{\hat{\mu}}{\hat{\lambda}}}\right]e^{2\frac{\hat{\lambda}}{\hat{\mu}}}\Phi\left(-2\sqrt{\frac{\hat{\lambda}}{\hat{\mu}}}\right).$$

V is the variance covariance matrix of the parameter estimates in the order $(\hat{\tau},\hat{\lambda},\hat{\mu})$. (The variances and covariances are set to 0 if the corresponding parameter is not estimated.)

Now if $\frac{\hat{\lambda}}{\hat{\mu}}$ is large then the contribution from $e^{2\frac{\hat{\lambda}}{\hat{\mu}}}\Phi\left(-2\sqrt{\frac{\hat{\lambda}}{\hat{\mu}}}\right)$ is calculated wrongly using standard approximations to $\Phi\left(-2\sqrt{\frac{\hat{\lambda}}{\hat{\mu}}}\right)$ (including the one used in GenStat). Thus, we replace this term by:

$$e^{2\frac{\hat{\lambda}}{\hat{\mu}}}\Phi\left(-2\sqrt{\frac{\hat{\lambda}}{\hat{\mu}}}\right) = \frac{b_1 w + b_2 w^2 + b_3 w^2 + b_4 w^4 + b_5 w^5}{\sqrt{2\pi}}$$

where $w = \dfrac{1}{1+0.2316419\times 2\times\sqrt{\frac{\hat{\lambda}}{\hat{\mu}}}}$

and

$b_1 = 0.31938153,$
$b_2 = -0.356563782,$
$b_3 = 1.781477937,$
$b_4 = -1.821255978,$
$b_5 = 1.330274429.$

Example

For the Hunter *et al*. data, the GenStat solutions are

```
                   estimate         s.e.
b                     111.8         51.8
m                     34.08         4.15
lag                   46.39         3.75
gamma                0.9125       0.0263

estimate      ref    correlations
b               1    1.000
m               2    0.841  1.000
lag             3   -0.949 -0.891  1.000
gamma           4   -0.068 -0.024  0.055  1.000
                         1      2      3      4

     Mean  seMean       Sd    seSd      T50   seT50
    80.47   1.888    18.82   2.372    76.03   1.684
```

So t_{50} is calculated in GenStat as:

```
DELETE [REDEFINE=yes] _XDev
CALC _XDev = EDINVNORMAL(0.5;34.08;111.8)
PRINT [IP=*] ':  Inverse Normal Equivalent Deviate = ',_XDev;F=1;SKIP=0
:  Inverse Normal Equivalent Deviate = 29.65
```

We need *lag* + 29.65 = 46.39+ 29.65 = 76.03 hours.

The standard error in GenStat is obtained slightly differently. In Excel it can be approximated as follows.

$$w = \frac{1}{1 + 0.2316419 \times 2 \times \sqrt{\frac{111.8}{34.08}}} = 0.543742,$$

$$e^{2\frac{\lambda}{\mu}} \Phi\left(-2\sqrt{\frac{\lambda}{\mu}}\right) = \frac{b_1 w + b_2 w^2 + b_3 w^2 + b_4 w^4 + b_5 w^5}{\sqrt{2\pi}} = 0.10319,$$

$d_1 = 1$

$$d_2 = \frac{34.08}{111.8}\left\{1 - \sqrt{2\pi}\left[2\sqrt{\frac{111.8}{34.08}} - \frac{1}{2}\sqrt{\frac{34.08}{111.8}}\right]0.10319\right\} = 0.04098,$$

$$d_2 = \sqrt{2\pi}\left(2\sqrt{\frac{111.8}{34.08}} - \frac{3}{2}\sqrt{\frac{34.08}{111.8}}\right)0.10319 = 0.72276.$$

Then

$$\text{var}(t_{50}) = 1^2 \times 3.75^2 + 0.04098^2 \times 51.8^2 + 0.72276 \times 4.15^2 + 2\left\{\begin{array}{l} 1 \times 0.04098 \times (3.75 \times 51.8 \times -0.949) \\ +1 \times 0.72276 \times (3.75 \times 4.15 \times -0.891) \\ +0.04098 \times 0.72276 \times (51.8 \times 4.15 \times 0.841) \end{array}\right\} = 3.121996.$$

Hence s.e.($\hat{t}_{50}$) = 1.767.

Reference

Abramowitz. M. and Stegun, I.A. (editors) (1972**).** Handbook of mathematical functions with formulas, graphs , and mathematical tables. Wahington, D.C. : National Bureau of Standards.

Appendix 1 Ordinary versus weighted least squares

When the sample size $n_.$ is known, the multinomial outcomes have second-order moments as follows. For n_a, the number of seeds germinating in (t_{a-1}, t_a),

$$\begin{aligned} \text{var}(n_a) &= n_. p_a (1 - p_a), \\ \text{cov}(n_a, n_b) &= - n_. p_a p_b, \qquad \text{for } a \neq b. \end{aligned} \quad \text{(A1.1)}$$

Thus, for N_a, the cumulative number of seeds germinating in $(0, t_a)$,

$$\begin{aligned} \text{var}(N_a) &= n_. P_a (1 - P_a), \\ \text{cov}(N_a, N_b) &= n_. P_a (1 - P_b), \qquad \text{for } a < b, \\ \text{cov}(N_a, N_b) &= \text{cov}(N_b, N_a), \qquad \text{for } a > b. \end{aligned} \quad \text{(A1.2)}$$

In (8) E_j is simply the mean $(n_. P_j)$. Ordinary least squares proceeds by forming a vector $(N_a - n_. P_a)$ using the k cumulative counts and minimising $(N_a - n_. P_a)^{\mathrm{T}}(N_a - n_. P_a)$. This assumes (incorrectly) that the covariance matrix of $(N_a - n_. P_a)$ is the identity matrix.

Weighted least squares proceeds as follows.

Form a vector equal to either $(n_a - n_. p_a)$ or $(N_a - n_. P_a)$ as well as the corresponding covariance matrix $\mathbf{V}_{\boldsymbol{n}}$ using (A1.1) or $\mathbf{V}_{\boldsymbol{N}}$ using (A1.2).

Now minimise $(N_a - n_. P_a)^{\mathrm{T}} \, \boldsymbol{V}_N^{-1} (N_a - n_. P_a)$ or $(n_a - n_. p_a)^{\mathrm{T}} \, \boldsymbol{V}_n^{-1} \, (n_a - n_. p_a)$.

It is straightforward to show that both matrix expressions are algebraically equivalent to

$$\mathrm{X}^2 = \sum_{i=1}^{k+1} \frac{(n_i - N p_i)^2}{N p_i}.$$

Appendix 2 Discussion of the analysis of deviance when $\gamma = 1$

If we denote $\hat{F}_j$ as the ML estimate of F_j using the full data set and $\widetilde{F}_j$ as the ML estimate using germinated seeds only, then it can be shown that the difference between (15) and (16) is the sum of two components:

$$Difference_1 = 2\left[\sum_{i=1}^{g} N_{ik} \ln\left(\frac{N_{ik}}{n_{i.}\hat{F}_k}\right) + \sum_{i=1}^{g} n_{i,k+1} \ln\left(\frac{n_{i,k+1}}{n_{i.}(1-\hat{F}_k)}\right)\right], \qquad \text{(A2.1)}$$

$$Difference_2 = 2\sum_{j=1}^{k} n_{.j} \ln\left(\frac{(\widetilde{F}_j - \widetilde{F}_{j-1})/\widetilde{F}_k}{(\hat{F}_j - \hat{F}_{j-1})/\hat{F}_k}\right). \qquad \text{(A2.2)}$$

(A2.1) assesses whether the germinated:ungerminated frequencies are consistent across groups with the probabilities that the assumed common distribution attaches to the two categories, whereas (A2.2) assesses whether the conditional common probability distributions are the same for the two methods of estimation.

Appendix 3 Macro to evaluate P(a ≤ *T* ≤ b) for T distributed as IG(τ,μ,λ)

The macro defines the function **IG(a, b, m, lag, L) = P(a < *T* ≤ b)**

```
Function IG(a As Double, b As Double, m As Double, lag As Double, L As Double)
Dim x(16), w(16)
If b <= lag Then
   IG = 0
   Exit Function
   Else: If a <= lag Then a = lag
End If
x(1) = -0.9894009439
x(2) = -0.944575023073232
x(3) = -0.865631202387831
x(4) = -0.755404408355003
x(5) = -0.617876244402643
x(6) = -0.458016777657227
x(7) = -0.281603550779258
x(8) = -9.50125098376374E-02
w(1) = 0.0271524594
w(2) = 0.0622535239
w(3) = 0.0951585116
w(4) = 0.1246289712
w(5) = 0.1495959888
w(6) = 0.1691565193
w(7) = 0.182603415
w(8) = 0.1894506104
For i = 9 To 16
   w(i) = w(17 - i)
   x(i) = -x(17 - i)
Next i
y = 0
For i = 1 To 16
   t = (b - a) / 2 * x(i) + (b + a) / 2
   f = Sqr(L / (8 * Atn(1) * (t - lag) ^ 3))
   f = f * Exp(-L / (2 * (t - lag) * m ^ 2) * (t - lag - m) ^ 2)
   y = y + w(i) * f
Next i
IG = y * (b - a) / 2
Exit Function
End Function
```

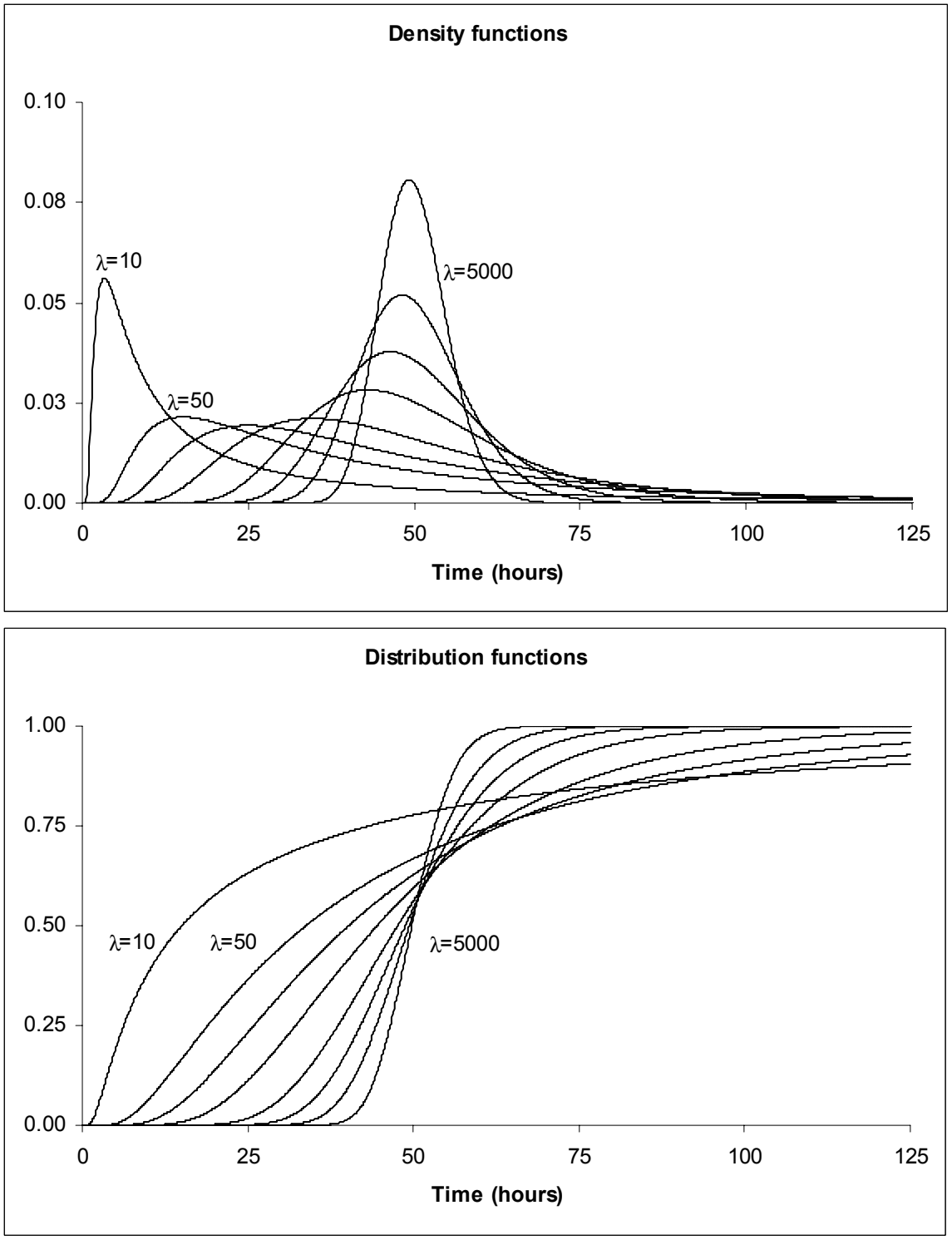

Figure 1 Selected inverse normal distributions; the density functions (top) and distribution functions (bottom) arise using $\tau = 0$, $\mu = 50$ and $\lambda = 10$, 50, 100, 200, 500, 1000, 2000 and 5000 (from left to right)

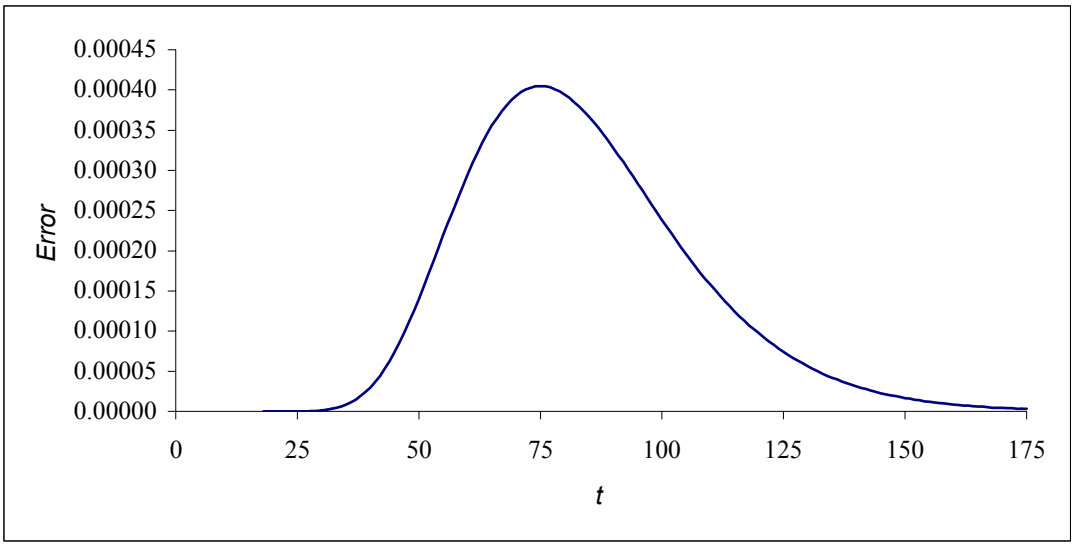

Figure 2 Error made in using approximation (TR1) for the inverse normal distribution function with $\tau = 0$, $\mu = 75$, $\lambda = 1000$.

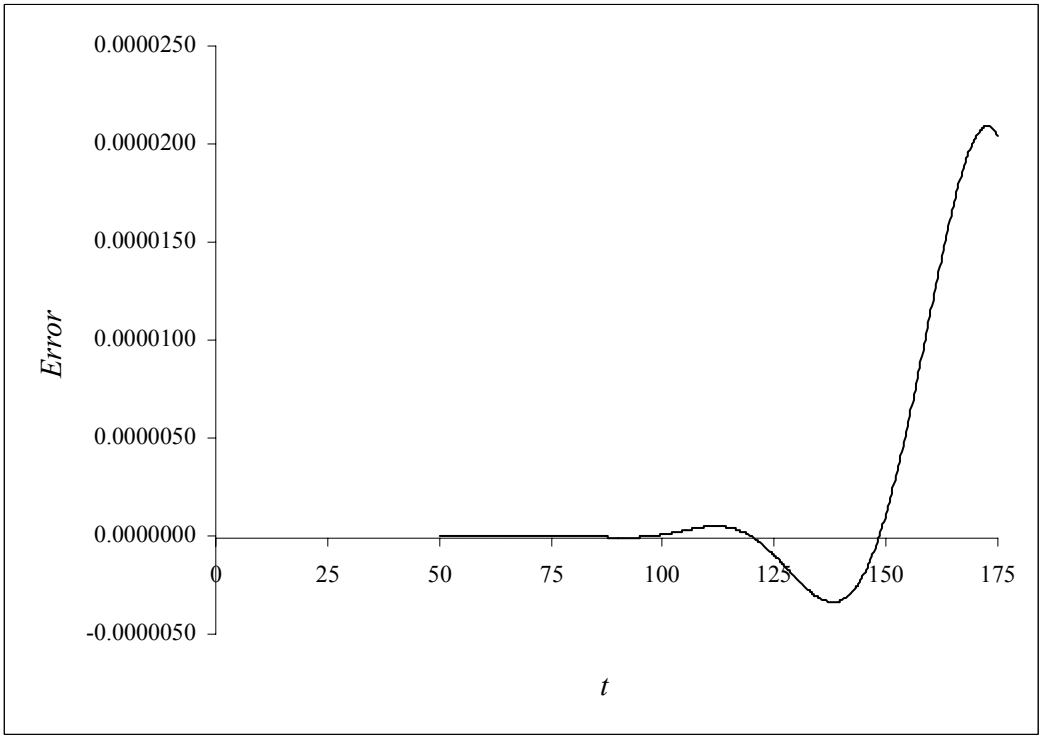

Figure 3 Error made in using 16-point Gaussian quadrature for the inverse Gaussian distribution function with $\tau = 0$, $\mu = 75$ and $\lambda = 1000$.