Probability in the Engineering and Informational Sciences, **22**, 2008, 261–271. Printed in the U.S.A.
doi: 10.1017/S0269964808000156

A NEW MODEL FOR SYMMETRIC AND SKEWED DATA

SARALEES NADARAJAH
School of Mathematics
University of Manchester
Manchester M60 1QD, UK
E-mail: saralees.nadarajah@manchester.ac.uk

The Normal and Gamma distributions are the most popular models for analyzing symmetric and skewed data, respectively. In this article, a new multimodal distribution is introduced that contains the Normal and Gamma distributions as particular cases and thus could be a better model for both symmetric and skewed data. Various structural properties of this distribution are derived, including its moment-generating function, characteristic function, moments, entropy, asymptotic distribution of the extreme order statistics, method of moment estimates, maximum likelihood estimates, Fisher information matrix, and simulation issues. The superiority of the new distribution is illustrated by means of two real datasets.

1. INTRODUCTION

The statistics literature is filled with hundreds of continuous univariate distributions. Johnson, Kotz, and Balakrishnan [5,6] provide excellent accounts of the known distributions. Undoubtedly, the most popular distributions for symmetric and skewed data are the Normal and Gamma distributions, respectively. It seems, however, that there are no distributions which contain the Normal and Gamma as particular cases (except, of course, for a mixture distribution of the two). Such distributions will be important because they will lead to better models for both symmetric and skewed data.

In this article, we introduce a new multimodal distribution that contains the Normal, Gamma, and the Rayleigh distributions as particular cases (Section 2). We derive various structural properties of this new distribution, including its moment-generating function (Section 3), characteristic function (Section 3), moments

 0269-9648/08 $25.00

(Section 4), entropy (Section 5), asymptotic distribution of the extreme order statistics (Section 6), estimation issues (Section 7), and simulation issues (Section 8). The applicability of the new distribution is illustrated by means of two real datasets (Section 9).

The calculations in this article involve several special functions, including the complementary error function defined by

$$\operatorname{erfc}(x) = 1 - \frac{2}{\sqrt{\pi}} \int_0^x \exp(-t^2)\, dt,$$

the modified Bessel function of the third kind defined by

$$K_\nu(x) = \frac{x^\nu \Gamma(1/2)}{2^\nu \Gamma(\nu + 1/2)} \int_1^\infty \exp(-xt)(t^2 - 1)^{\nu - 1/2}\, dt,$$

the parabolic cylinder function defined by

$$D_p(x) = \frac{\exp(-x^2/4)}{\Gamma(-p)} \int_0^\infty \exp\{-(tx + t^2/2)\} t^{-(p+1)}\, dt,$$

the Kummer function defined by

$$\Psi(a,b;x) = \frac{1}{\Gamma(a)} \int_0^\infty \exp(-xt) t^{a-1} (1+t)^{b-a-1}\, dt,$$

and the confluent hypergeometric function defined by

$${}_1F_1(a;c;x) = \sum_{k=0}^{\infty} \frac{(a)_k}{(c)_k} \frac{x^k}{k!},$$

where $(e)_k = e\ (e+1) \cdots (e+k-1)$ denotes the ascending factorial. The properties of these special functions can be found in Prudnikov, Brychkov, and Marichev [8] and Gradshteyn and Ryzhik [3].

2. PROBABILITY DENSITY FUNCTION

The new distribution is taken to have a probability density function (p.d.f.) that is proportional to the product of Normal and Gamma p.d.f.s; that is,

$$f(x) = Cx^{\alpha-1} \exp(-px^2 - qx) \tag{1}$$

for $0 < x < \infty$, $\alpha > 0$, $p \geq 0$, and $-\infty < q < \infty$, where C denotes the normalizing constant. Note that if $p = 0$, then one must have $q > 0$. Application of Eq. (2.3.15.3) in Prudnikov et al. [8, Vol. 1] shows that the normalizing constant is given by

$$\frac{1}{C} = \Gamma(\alpha)(2p)^{-\alpha/2} \exp\left(\frac{q^2}{8p}\right) D_{-\alpha}\left(\frac{q}{\sqrt{2p}}\right). \tag{2}$$

https://doi.org/10.1017/S0269964808000156 Published online by Cambridge University Press

By special properties of the parabolic cylinder function, (2) can be reduced to

$$\frac{1}{C} = \frac{1}{2}\sqrt{\frac{\pi}{p}}\,\exp\left(\frac{q^2}{4p}\right)\,\mathrm{erfc}\left(\frac{q}{2\sqrt{p}}\right)$$

if $\alpha=1$, to

$$\frac{1}{C} = \frac{1}{2}\sqrt{\frac{q}{p}}\,\exp\left(\frac{q^2}{8p}\right)K_{1/4}\left(\frac{q^2}{8p}\right)$$

if $\alpha=1/2$, and to

$$\frac{1}{C} = \frac{1}{8}\left(\frac{q}{p}\right)^{3/2}\exp\left(\frac{q^2}{8p}\right)\left\{K_{3/4}\left(\frac{q^2}{8p}\right) - K_{1/4}\left(\frac{q^2}{8p}\right)\right\}$$

if $\alpha = 3/2$. The new distribution given by (1) is very flexible and it contains several of the standard distributions as particular cases. The half-Normal distribution is the particular case for $\alpha = 1$; the Gamma distribution is the particular case for $p = 0$; and the Rayleigh distribution is the particular case for $q = 0$. Let us now consider the shape of (1). The first and second derivatives of $\log f$ are

$$\frac{d\log f}{dx} = \frac{\alpha - 1}{x} - 2px - q$$

and

$$\frac{d^2\log f}{dx^2} = -\frac{\alpha - 1}{x^2} - 2p.$$

Setting the first derivative to zero, one obtains the quadratic equation $\alpha - 1 - 2px^2 - qx = 0$ with the solutions $x = \delta_1$ or $x = \delta_2$, where

$$\delta_1 = \frac{-q - \sqrt{q^2 - 8p(1-\alpha)}}{4p}$$

and

$$\delta_2 = \frac{-q + \sqrt{q^2 - 8p(1-\alpha)}}{4p}.$$

Routines calculations show that the following shapes are possible:

- If either $\alpha \leq 1$ and $q \geq 0$ or $\alpha < 1$, $q < 0$, and $2\sqrt{2_p(1-\alpha)} + q \geq 0$, then f is monotonically decreasing.
- If $\alpha = 1$ and $q < 0$, then f attains a maximum at $x = - = q/(2p)$ before decreasing for all $x > -q/(2p)$.
- If $\alpha > 1$, then f attains a maximum at $x = \delta_2$ before decreasing for all $x > \delta_2$.
- If $\alpha < 1$, $q < 0$, and $2\sqrt{2p(1-\alpha)} + q < 0$, then f has a minimum at $x = \delta_1$ and a maximum at $x = \delta_2$ before decreasing for all $x > \delta_2$.

https://doi.org/10.1017/S0269964808000156 Published online by Cambridge University Press

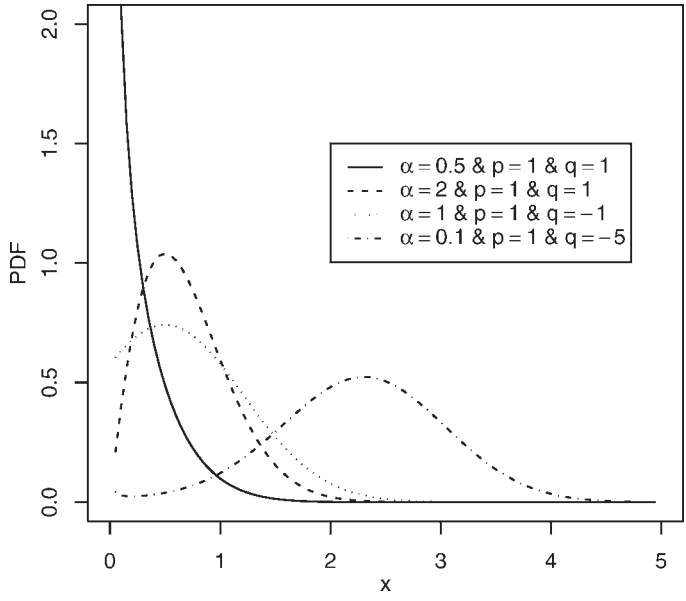

FIGURE 1. Plots of the p.d.f. (1) for selected values of α, p, and q.

Some of the possible shapes are illustrated in Figure 1 for selected values of α, p, and q.

3. CHARACTERISTIC FUNCTION

Here, we derive the moment-generating and the characteristic functions of a random variable X with p.d.f. (1). The moment-generating function (m.g.f.) is defined by $M(t) = \mathrm{E}(\exp(tX))$. It can be calculated easily as

$$M(t) = C\int_0^\infty x^{\alpha-1}\ \exp\{-px^2 - (q-t)x\}\,dx$$

$$= C\Gamma(\alpha)(2p)^{-\alpha/2}\ \exp\left\{\frac{(q-t)^2}{8p}\right\}D_{-\alpha}\left(\frac{q-t}{\sqrt{2p}}\right),$$

where we have applied Eq. (2.3.15.3) in Prudnikov et al. [8, Vol. 1]. Thus, the characteristic function of X defined by $\phi(t) = \mathrm{E}(\exp(itX))$ takes the form

$$\phi(t) = C\Gamma(\alpha)(2p)^{-\alpha/2}\ \exp\left\{\frac{(q-it)^2}{8p}\right\}D_{-\alpha}\left(\frac{q-it}{\sqrt{2p}}\right),$$

where $i = \sqrt{-1}$ is the complex number.

https://doi.org/10.1017/S0269964808000156 Published online by Cambridge University Press

4. MOMENTS

The nth moment of a random variable X with p.d.f. (1) can be calculated easily as

$$\begin{aligned} \mathrm{E}(X^n) &= C\int_0^\infty x^{n+\alpha-1}\ \exp(-px^2-qx)dx \\ &= C\Gamma(n+\alpha)(2p)^{-(n+\alpha)/2}\ \exp\left(\frac{q^2}{8p}\right)D_{-(n+\alpha)}\left(\frac{q}{\sqrt{2p}}\right), \end{aligned} \tag{3}$$

where we have applied Eq. (2.3.15.3) in Prudnikov et al. [8, Vol. 1]. If $\alpha + n \geq 1$ is an integer, then (3) can be reduced to

$$\mathrm{E}(X^n) = C\frac{(-1)^{\alpha+n}\sqrt{\pi}}{2\sqrt{p}}\frac{\partial^{\alpha+n}}{\partial q^{\alpha+n}}\left[\exp\left(\frac{q^2}{4p}\right)\mathrm{erfc}\left(\frac{q}{2\sqrt{p}}\right)\right]$$

by special properties of the parabolic cylinder function.

5. RÉNYI ENTROPY

An entropy of a random variable X is a measure of variation of the uncertainty. The Rényi entropy is defined by

$$\mathcal{J}_R(\gamma) = \frac{1}{1-\gamma}\log\left\{\int f^\gamma(x)\,dx\right\},$$

where $\gamma > 0$ and $\gamma \neq 1$ [9]. It follows easily by the application of Eq. (2.3.15.3) in Prudnikov et al. [8, Vol. 1] that

$$\begin{aligned} \int_0^\infty f^\gamma(x)\,dx &= C^\gamma\int_0^\infty x^{\alpha\gamma-\gamma}\exp(-p\gamma x^2 - q\gamma x)\,dx \\ &= C^\gamma\Gamma(\alpha\gamma-\gamma+1)(2p)^{-(\alpha\gamma-\gamma+1)/2}\ \exp\left(\frac{q^2\gamma}{8p}\right)D_{-(\alpha\gamma-\gamma+1)}\left(\frac{q\sqrt{\gamma}}{\sqrt{2p}}\right). \end{aligned}$$

Thus, the Rényi entropy for (1) is given by

$$\begin{aligned} \mathcal{J}_R(\gamma) = \frac{1}{1-\gamma}\Bigg\{&\gamma\log C + \log\,\Gamma(\alpha\gamma-\gamma+1) \\ &-\frac{\alpha\gamma-\gamma+1}{2}\log(2p) + \frac{q^2\gamma}{8p} + \log\,D_{-(\alpha\gamma-\gamma+1)}\left(\frac{q\sqrt{\gamma}}{\sqrt{2p}}\right)\Bigg\}. \end{aligned}$$

6. ASYMPTOTICS

If $X_1, \ldots, X_n$ is a random sample from (1) and if $\bar{X}=(X_1+\cdots+X_n)/n$ denotes the sample mean, then by the usual central limit theorem, $\sqrt{n}(\bar{X} - E(X))/\sqrt{\mathrm{Var}(X)}$

https://doi.org/10.1017/S0269964808000156 Published online by Cambridge University Press

approaches the standard Normal distribution as $n \to \infty$. Sometimes one would be interested in the asymptotics of the extreme values $M_n = \max(X_1, \ldots, X_n)$ and $m_n = \min(X_1, \ldots, X_n)$. Note from (1) that $f(t) \sim Ct^{\alpha-1} \exp(-pt^2 - qt)$ as $t \to \infty$ and $f(t) \sim Ct^{\alpha-1}$ as $t \to 0$. Thus, it follows by using L'Hospital's rule that

$$\frac{1 - F(t + x/t)}{1 - F(t)} \longrightarrow \exp(-2px)$$

as $t \to \infty$ and

$$\frac{F(xt)}{F(t)} \longrightarrow x^{\alpha}$$

as $t \to 0$. Hence, it follows from Theorem 1.6.2 in Leadbetter, Lindgren, and Rootzén [7] that there must be norming constants $a_n > 0$, b_n, $c_n > 0$, and d_n such that

$$\Pr\{a_n(M_n - b_n) \le x\} \longrightarrow \exp\{-\exp(-2px)\}$$

and

$$\Pr\{c_n(m_n - d_n) \le x\} \longrightarrow 1 - \exp(-x^{\alpha})$$

as $n \to \infty$. The form of the norming constants can also be determined. For instance, using Corollary 1.6.3 in Leadbetter et al. [7], one can see that $a_n = b_n = F^{-1}(1 - 1/n)$, where $F(\cdot)$ is the cumulative distribution function corresponding to (1).

7. ESTIMATION

Here, we consider estimation by the method of moments and the method of maximum likelihood when $X_1, \ldots, X_n$ is a random sample from (1) and we provide expressions for the associated Fisher information matrix. It is clear from (3) that the method of moments estimates are the simultaneous solutions of the equations

$$C\Gamma(1 + \alpha)(2p)^{-(1+\alpha)/2} \exp\left(\frac{q^2}{8p}\right) D_{-(1+\alpha)}\left(\frac{q}{\sqrt{2p}}\right) = \frac{1}{n}\sum_{j=1}^{n} X_j,$$

$$C\Gamma(2 + \alpha)(2p)^{-(2+\alpha)/2} \exp\left(\frac{q^2}{8p}\right) D_{-(2+\alpha)}\left(\frac{q}{\sqrt{2p}}\right) = \frac{1}{n}\sum_{j=1}^{n} X_j^2,$$

and

$$C\Gamma(3 + \alpha)(2p)^{-(3+\alpha)/2} \exp\left(\frac{q^2}{8p}\right) D_{-(3+\alpha)}\left(\frac{q}{\sqrt{2p}}\right) = \frac{1}{n}\sum_{j=1}^{n} X_j^3$$

The log-likelihood is

$$\log L(\alpha,p,q) = n \log C + (\alpha - 1)\sum_{j=1}^{n} \log X_j - p\sum_{j=1}^{n} X_j^2 - q\sum_{j=1}^{n} X_j.$$

https://doi.org/10.1017/S0269964808000156 Published online by Cambridge University Press

The first derivatives with respect to the three parameters are

$$\frac{\partial \log L}{\partial \alpha} = \sum_{j=1}^{n} \log X_j + \frac{n}{C}\frac{\partial C}{\partial \alpha},$$

$$\frac{\partial \log L}{\partial p} = -\sum_{j=1}^{n} X_j^2 + \frac{n}{C}\frac{\partial C}{\partial p},$$

and

$$\frac{\partial \log L}{\partial q} = \sum_{j=1}^{n} X_j + \frac{n}{C}\frac{\partial C}{\partial q}.$$

Thus, the maximum likelihood estimates are the simultaneous solutions of the equations

$$\frac{n}{C}\frac{\partial C}{\partial \alpha} = -\sum_{j=1}^{n} \log X_j, \tag{4}$$

$$\frac{n}{C}\frac{\partial C}{\partial p} = \sum_{j=1}^{n} X_j^2, \tag{5}$$

and

$$\frac{n}{C}\frac{\partial C}{\partial q} = \sum_{j=1}^{n} X_j. \tag{6}$$

The partial derivatives in (4)–(6) can be computed by using the facts

$$D_\nu(z) = 2^{\nu/2} \exp\left(-\frac{z^2}{4}\right)\Psi\left(-\frac{\nu}{2}, \frac{1}{2}; \frac{z^2}{2}\right),$$

$$\Psi(a, c; z) = \frac{\Gamma(1-c)}{\Gamma(1+a-c)}{}_1F_1(a; c; z) + \frac{\Gamma(c-1)}{\Gamma(a)}{}_1F_1(1+a-c; 2-c; z),$$

$$\frac{\partial}{\partial a}{}_1F_1(a; c; z) = \sum_{k=0}^{\infty} \frac{(a)_k \psi(a+k) z^k}{(c)_k k!} - \psi(a){}_1F_1(a; c; z),$$

and

$$\frac{\partial}{\partial z}{}_1F_1(a; c; z) = \frac{a}{c}{}_1F_1(a+1; c+1; z),$$

where $\psi(x) = d \log \Gamma(x)/dx$ is the digamma function. Calculation of the associated Fisher information matrix requires second-order derivatives of $\log L$. All of the

https://doi.org/10.1017/S0269964808000156 Published online by Cambridge University Press

second-order derivatives take the form

$$\frac{\partial^2 \log L}{\partial\theta_i \partial\theta_j} = -\frac{n}{C^2}\frac{\partial C}{\partial\theta_i}\frac{\partial C}{\partial\theta_j} + \frac{n}{C}\frac{\partial^2 C}{\partial\theta_i \partial\theta_j}.$$

Hence, the elements of the Fisher information matrix all take the form

$$E\left(-\frac{\partial^2 \log L}{\partial\theta_i\,\partial\theta_j}\right) = -\frac{n}{C^2}\frac{\partial C}{\partial\theta_i}\frac{\partial C}{\partial\theta_j} - \frac{n}{C}\frac{\partial^2 C}{\partial\theta_i \partial\theta_j}. \tag{7}$$

The second-order partial derivatives in (7) can be computed by using the facts

$$\frac{\partial^2}{\partial a^2}{}_1F_1(a;c;z) = \sum_{k=0}^{\infty} \frac{\left\{\begin{array}{c}\Gamma(a+k)\psi'(a+k) + \Gamma'(a+k)\psi(a+k) \\ -\Gamma(a+k)\psi(a+k)\psi(a)\end{array}\right\} z^k}{\Gamma(a)(c)_k k!}$$

$$- \psi(a)\frac{\partial}{\partial a}{}_1F_1(a;c;z) - \psi'(a){}_1F_1(a;c;z),$$

$$\frac{\partial^2}{\partial a \partial z}{}_1F_1(a;c;z) = \sum_{k=0}^{\infty} \frac{(a)_k \psi(a+k) k z^{k-1}}{(c)_k k!} - \frac{a\psi(a)}{c}{}_1F_1(a+1;c+1;z),$$

and

$$\frac{\partial^2}{\partial z^2}{}_1F_1(a;c;z) = \frac{a(a+1)}{c(c+1)}{}_1F_1(a+2;c+2;z),$$

where $\psi'(\cdot)$ denotes the derivative of the digamma function.

8. SIMULATION

The rejection sampling can be used to simulate from (1) with a gamma p.d.f. chosen as the envelope. The following scheme can be used:

1. Generate a gamma random variable Y that has the p.d.f. $q^{\alpha}y^{\alpha-1}\exp(-qy)/\Gamma(\alpha)$.
2. Generate a uniform [0, 1] random variable U independently of Y.
3. If $U < \exp(-pY^2)$, accept Y as a realization from (1).
4 If $U \geq \exp(-pY^2)$, return to step 1.

Note that there are standard routines for generating gamma random variables.

9. APPLICATION

Here, we illustrate the superiority of the new distribution given by (1) as compared to the Normal and Gamma distributions. We consider two biological datasets from Fry

https://doi.org/10.1017/S0269964808000156 Published online by Cambridge University Press

[2]: One is a symmetric dataset and the other is skewed. We fitted the following three models to each of the datasets:

Model 1: Equation (1) with $\alpha = 1$ corresponding to the half-normal model.

Model 2: Equation (1) with $p = 0$ corresponding to the Gamma model.

Model 3: Equation (1) with no restrictions on the parameters.

The fitting of (1) was performed by the method of maximum likelihood; see Section 7. The quasi-Newton algorithm nlm in the R software package [1,4,10] was used to solve the likelihood equations (4)–(6). The results for the two datasets are described as follows:

Dataset 1. This is a symmetric dataset (see Fig. 2). Model 1 gave the estimates $\hat{p} = 6.778 \times 10^{-7}$ and $\hat{q} = 1.221$ with $-\log L = 365.2$. Model 2 gave the estimates $\hat{\alpha} = 5.786$ and $\hat{q} = 1.835$ with $-\log L = 163.0$. Model 3 gave the estimates $\hat{\alpha} = 0.408$, $\hat{p} = 0.479$, and $\hat{q} = -3.247$ with $-\log L = 148.5$. Thus, it follows by the standard likelihood ratio test [11] that Model 3 should be preferred. This is supported by Figure 2, where the fitted p.d.f.s for the three models and the histogram of the data are shown.

Dataset 2. This is a skewed dataset (see Fig. 3). Model 1 gave the estimates $\hat{p} = 1.338 \times 10^{-6}$ and $\hat{q} = 1.663$ with $-\log L = 733.9$. Model 2 gave the estimates

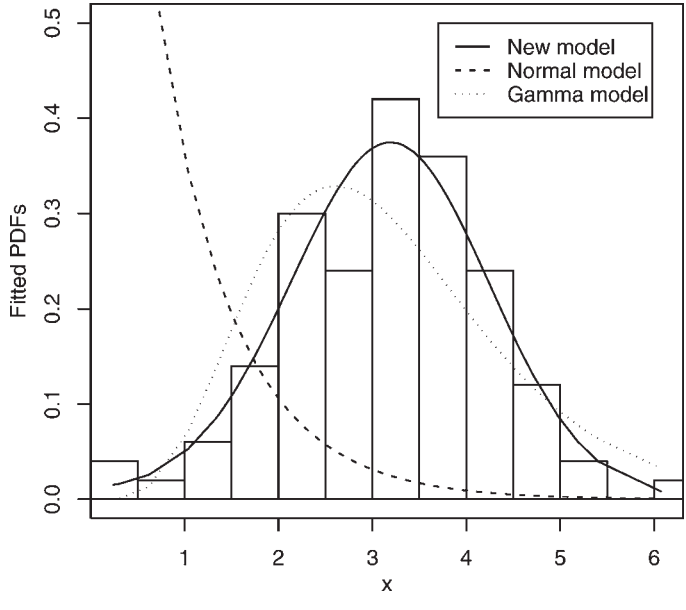

Figure 2. The three fitted p.d.f.s and the histogram for dataset 1.

https://doi.org/10.1017/S0269964808000156 Published online by Cambridge University Press

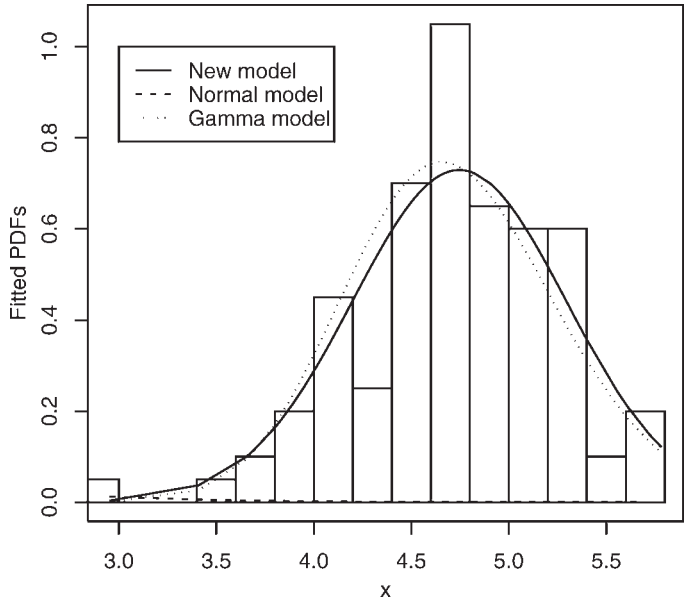

Figure 3. The three fitted p.d.f.s and the histogram for dataset 2.

$\hat{\alpha} = 77.061$ and $\hat{q} = 16.331$ with $-\log L = 79.4$. Model 3 gave the estimates $\hat{\alpha} = 2.468 \times 10^{-6}$, $\hat{p} = 1.693$, and $\hat{q} = -16.283$ with $-\log L = 77.2$. Thus, it follows again by the standard likelihood ratio test [11] that Model 3 should be preferred. This is supported by Figure 3, where the fitted p.d.f.s for the three models and the histogram of the data are shown.

The fittings of the above exercise were repeated to several other biological datasets exhibiting a variety of skewness structures; the conclusions for each dataset were the same.

References

1. Dennis, J.E. & Schnabel, R.B. (1983). *Numerical methods for unconstrained optimization and nonlinear equations*. Englewood Cliffs, NJ: Prentice-Hall.
2. Fry, J.C. (1993). *Biological data analysis: A practical approach*. Oxford: Oxford University Press.
3. Gradshteyn, I.S. & Ryzhik, I.M. (2000). *Table of integrals, series, and products*, 6th ed. San Diego: Academic Press.
4. Ihaka, R. & Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics* 5: 299–314.
5. Johnson, N.L., Kotz, S., & Balakrishnan, N. (1994). *Continuous univariate distributions*, Vol. 1: 2nd ed. New York: Wiley.

https://doi.org/10.1017/S0269964808000156 Published online by Cambridge University Press

6. Johnson, N.L., Kotz, S. & Balakrishnan, N. (1995). *Continuous univariate distributions*, Vol. 2, 2nd ed. New York: Wiley.
7 Leadbetter, M.R., Lindgren, G., & Rootzén, H. (1987). *Extremes and related properties of random sequences and processes*. New York: Springer-Verlag.
8. Prudnikov, A.P., Brychkov, Y.A., & Marichev, O.I. (1986). *Integrals and series*, Vols. 1, 2 and 3. Amsterdam: Gordon & Breach Science.
9. Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. I, pp. 547–561. Berkeley: University of California Press.
10. Schnabel, R.B., Koontz, J.E., & Weiss, B.E. (1985). A modular system of algorithms for unconstrained minimization. *ACM Transactions on Mathematical Software* 11: 419–440.
11. Wald, A. (1923). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, 54: 426–483.