A New SSD - The gamma-generated Logistic Distribution (`ggLogis`)

David R. Fox
Environmetrics Australia
University of Melbourne

1. INTRODUCTION

Despite the many well-known and documented deficiencies, the species sensitivity distribution (SSD) remains a cornerstone of modern ecotoxicological practice. SSDs are typically used to establish concentrations of toxicants and pollutants in waterbodies that are protective of some nominally high fraction of all potentially affected species.

Being a probability model for toxicity data, the SSD approach is necessarily statistical and its implementation can be mathematically and computationally demanding. Unlike some other branches of science, there are no guiding principles or axiomatic theorems in ecotoxicology that establish the foundations of the SSD paradigm. This means that more assumptions have to be made than would otherwise be the case. For example, it has to be assumed that our toxicity data represent a random sample from the larger population of all toxicity values when we know this to be patently false. Or we must assume that the underlying probability distribution for toxicity values is of a particular theoretical form such as the log-normal or log-logistic distribution when in fact we have no idea what the true functional form of this distribution is. Compounding these problems is the omnipotent and vexing issue of pathologically small sample sizes.

These are serious drawbacks and certainly call into question the legitimacy of SSD modelling. However, ecotoxicologists are generally agreed that, despite the problems, the SSD methodology is the best we currently have and is preferable to the so-called assessment factor (AF) approach that it displaced some 25-30 years ago.

In this note, we introduce the gamma-generated logistic distribution (`ggLogis`) as a new candidate SSD model. The idea of generalising standard statistical distributions is not new. For example, Mudholkar and Srivastava (1993) introduced the exponentiated Weibull distribution to analyse failure rate data while Gupta et al. (1998) proposed a generalization of the standard exponential distribution. A class of generalised distributions known as beta and gamma-generated distributions was introduced by Zografos and Balakrishnan (2009). The first application of this method using the standard logistic distribution appears to be due to Castellares et al. (2015) who provided details of many of its mathematical and statistical properties. As far as we are aware, these methods of generalising existing distributions have not previously been used in the context of SSD modelling.

The standard logistic distribution is often used in ecotoxicology as an SSD. Because it is symmetrical and distributed over the entire real line, the logistic SSD is most often applied to log-transformed toxicity data (meaning the untransformed toxicity data are assumed to follow a log-logistic distribution)

The logistic distribution was also favoured by virtue of the following:

- parameter estimation is relatively straightforward;
- it has a closed-form expression for the cumulative distribution function (*cdf*);
- it admits a variety of shapes.

In the remainder of this note we provide mathematical details of the `ggLogis` distribution including its pdf, *cdf*, moments, and MLE equations. We illustrate the use of the ggLogis distribution as an SSD using the R statistical computing software and compare HCx estimates with those obtained using more conventional SSDs and software tools. Importantly, we show how standard errors and confidence intervals can be obtained without the need to resort to computationally-intensive resampling techniques.

2. MATHEMATIAL AND STATISTICAL PROPERTIES OF THE `ggLogis` DISTRIBUTION

We commence with the definition of a gamma-generated distribution. Let the *cdf* of the root distribution be $F\left(x;\Theta\right)$ and let U be a gamma random variable having shape parameter $a > 0$ and *pdf*

$$f_U(u;a)=\frac{1}{\Gamma(a)}u^{a-1}e^{-u}\quad;\quad \text{u>0}$$

A new random variable *Y* has the gamma-generated root distribution $G\{y;(a;\Theta)\}$ if its *cdf* is given by Equation 1.

$$G_Y(y;a,\Theta)=\frac{1}{\Gamma(a)}\int_0^{-\ln[1-F_X(y;\Theta)]}t^{a-1}e^{-t}\,dt\quad;\quad -\infty<y<\infty \tag{1}$$

To obtain the gamma-generated logistic distribution, we need to replace $F_X(\cdot\,;\Theta)$ in Equation 1 with the *cdf* for the standard logistic distribution (Equation 2).

$$F_X(x,\mu,\sigma)=\frac{1}{1+e^{-\frac{(x-\mu)}{\sigma}}}\quad;\quad -\infty<\{x;\mu\}<\infty;\quad \sigma>0 \tag{2}$$

This gives:

$$G_Y(y;a,\mu,\sigma)=\frac{\Gamma\left[a,-\ln\left(\dfrac{e^{-\frac{(y-\mu)}{\sigma}}}{1+e^{-\frac{(y-\mu)}{\sigma}}}\right)\right]}{\Gamma(a)} \tag{3}$$

where $\Gamma(\xi,z)$ is the incomplete gamma function given by Equation 4.

$$\Gamma(\xi,z)=\int_z^\infty s^{\xi-1}e^{-s}\,ds \tag{4}$$

Differentiating the *pdf* for the gamma-generated logistic distribution with respect to *y* we obtain the *pdf* :

$$g_Y(y;a,\mu,\sigma)=\frac{-\left\{\ln\left[\dfrac{1+e^{-\frac{(y-\mu)}{\sigma}}}{e^{-\frac{(y-\mu)}{\sigma}}}\right]\right\}^a}{4\sigma\Gamma(a)\ln\left[\dfrac{e^{-\frac{(y-\mu)}{\sigma}}}{1+e^{-\frac{(y-\mu)}{\sigma}}}\right]\cosh^2\left(e^{-\frac{(y-\mu)}{2\sigma}}\right)} \tag{5}$$

Furthermore, $g_Y(y;a,\mu,\sigma)$ can be expressed as:

$$g_Y(y;a,\mu,\sigma)=\frac{\left\{-\ln\left[1-F_X(y;\Theta)\right]\right\}^{a-1}f_X(y;\Theta)}{\Gamma(a)} \tag{6}$$

where $g_Y(y;\Theta)$ is the *pdf* corresponding to $G_Y(y;\Theta)$.

The `ggLogis` distribution given by Equation 5 can generate a variety of shapes. Of particular relevance to ecotoxicology and SSD modelling is the ability to generate 'fat' left-tail distributions. (Figure 1).

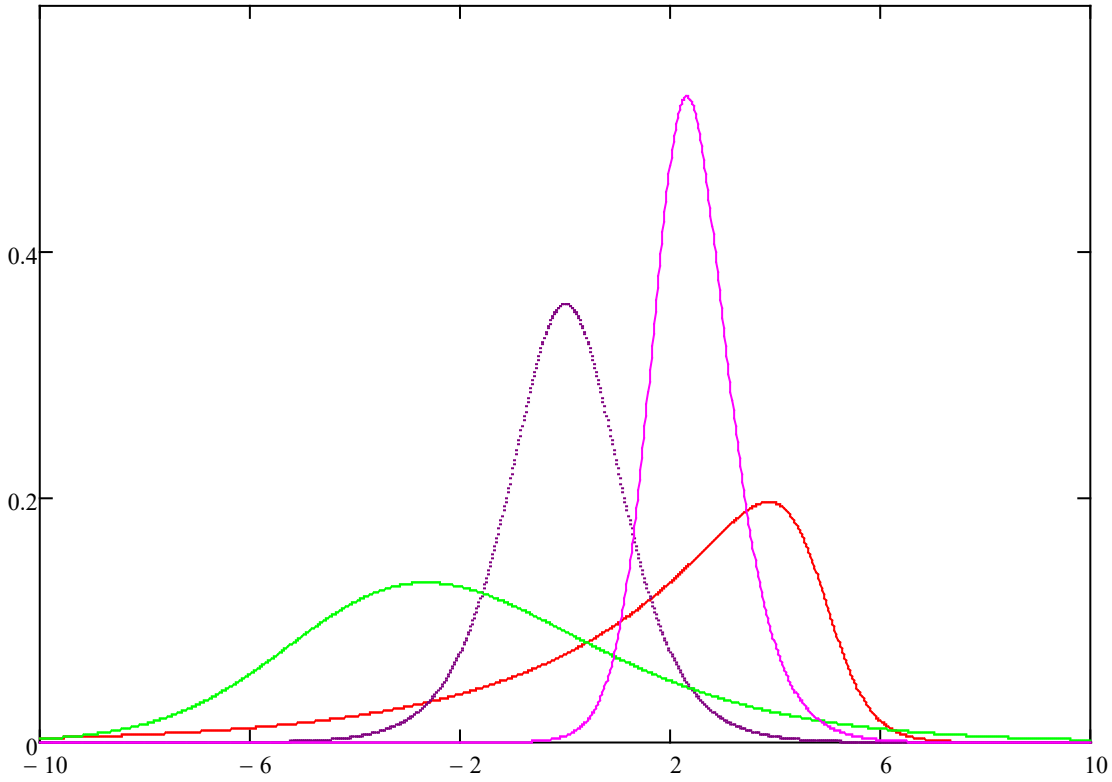

Figure 1. Shapes of the ggLogis distribution for a selection of parameter combinations.

3. MOMENTS OF THE ggLogis DISTRIBUTION

For a random variable Y having *pdf* given by Equation 6, the k^{th} raw moment is:

$$E\left[Y^k\right] = \frac{1}{\Gamma(a)} \int_{-\infty}^{\infty} y^k \left\{-\ln\left[1 - F_Y\left(y;\Theta\right)\right]\right\}^{a-1} f_Y(y;\Theta) \qquad (7)$$

Substituting $Z = -\ln\left[1 - F_Y\left(y;\Theta\right)\right]$ in Equation 7 we obtain:

$$
\begin{aligned}
E\left[Y^k\right] &= \frac{1}{\Gamma(a)}\int_{-\infty}^{\infty}\left[F^{-1}\left(1-e^{-z}\right)\right]^k z^{a-1}\,e^{-z}\,dz \\
&= E_Z\left\{\left[F^{-1}\left(1-e^{-z}\right)\right]^k\right\}
\end{aligned}
\tag{8}
$$

That is, the k^{th} raw moment of Y can be expressed as the expectation of the k^{th} power of $\left[F^{-1}\left(1-e^{-Z}\right)\right]$.

where Z has a gamma(a,1) *pdf*. In our case, $F(\bullet)$ is the logistic *cdf* which, on substitution into Equation 8 yields:

$$
E\left[Y^k\right] = \frac{1}{\Gamma(a)}\int_{0}^{\infty}\left[\ln\left(e^{z}-1\right)\right]^k z^{a-1}\,e^{-z}\,dz \tag{9}
$$

Equation 9 is not particularly useful for method of moments parameter estimation and so we next discuss parameter estimation using maximum likelihood.

4. MAXIMUM LIKELIHOOD ESTIMATION FOR THE ggLogis DISTRIBUTION

The *log-likelihood* function for the `ggLogis` distribution is given by Equations 10(a)-(c).

$$
\sum_{i-1}^{n}\ln\left\{-\ln(w[i)\right\} = n\,\Psi(a) \tag{10a}
$$

$$
\sum_{i=1}^{n}(1-a)(w_i-1)^2 = \sum_{i=1}^{n}(w_i-2w_i^2) \tag{10b}
$$

$$
n\sum_{i=1}^{n}\left\{\frac{\left[-\ln(w_i)\right]^a\ln\left(\dfrac{1-w_i}{w_i}\right)}{\ln(w_i)}\left[(2w_i-1)\ln(w_i)+(a-1)(w_i-1)\right]\right\} = 0 \tag{10c}
$$

where $w_i = \dfrac{1}{1+e^{(x_i-\mu)/\sigma}}$ and $\Psi(\cdot)$ is the digamma function.

There is no explicit solution to the system of non-linear equations 10(a)-(c) and these must be solved numerically. We demonstrate how this is accomplished using the `optim` package in R with the use of an example.

EXAMPLE – PFOS DATA

The data for this example are given in Table 1 and Figure 2.

The estimated HC_5 from a log-logistic distribution fitted to the data is 1.05.

Table 1. PFOS toxicity data.

Species	Concentration
Danio rerio	0.2936
Oryzias latipes	4
Enallagma cyathigerum	7.95
Daphnia magna	8
Xiphorus helleri	40
Chironomus tentans	49.2
Myriophyllum sibiricum	100
Pimephales promelas	300
Moina macrocopa	312.5
Rana pipiens	1242
Myriophyllum spicatam	3300
Selenastrum capricornutum	5300
Daphnia pulicaria	6000
Lemna gibba	6600
Chlorella vulgaris	8200
Scenedesmus obliquus	51000
Navicula pelliculosa	62300
Anabaena flos-aquae	82000

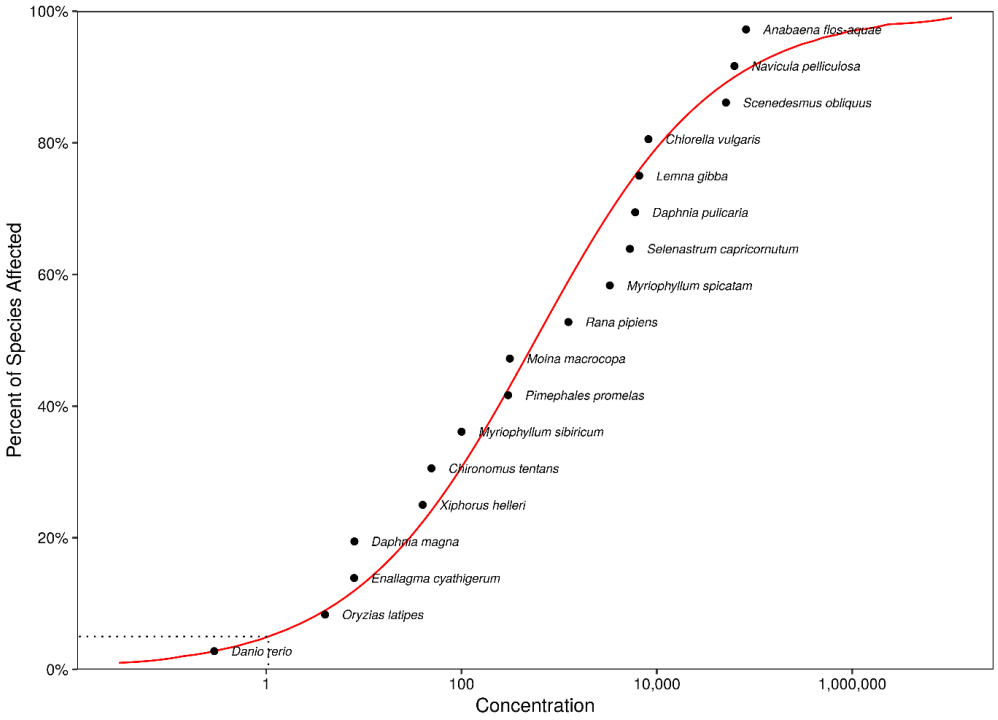

Figure 2. Log-logistic species sensitivity distribution fitted to data of Table 1 using the ssdtools shiny app.

R- code

```
#PFOS data
x<-scan()  #  read in data
0.2936
4
7.95
.
.
62300
82000


y<-log(x)  # fit distribution to log of data
n<-length(x)  # sample size

f<-function(theta){  #  this is the pdf of the ggLogis distribution
  a<-theta[1]
  m<-theta[2]
  s<-theta[3]
  z<-(y-m)/s
  g<-log((1+exp(-z))/exp(-z))^(a-1)/(4*s*gamma(a)*cosh(z/2)^2)
  return(g)
}


ll<-function(theta){  #  this is the log-likelihood function
  a<-theta[1]
  m<-theta[2]
  s<-theta[3]
  loglik<-sum(log(f(theta)))
  return(loglik)
}

theta<-c(0.5,10,1.0)  # initial guess for parameter values
mle<-optim(theta,ll,gr=NULL,method="Nelder-
Mead",control=list(fnscale=-1),hessian = TRUE)
V<-solve(-mle$hessian)  #variance-covariance matrix of estimated
parameters
se<-diag(V)^0.5  #  extract standard errors of parameter estimates
```

Box 1. R-code for fitting ggLogis SSD

Running the code in Box 1 gives the following output:

```
> mle
$par
[1]  0.2861793 10.1569461  1.1213885

$value
[1] -48.70741

$counts
function gradient
     118       NA
```

```
$convergence
[1] 0

$message
NULL
$hessian
           [,1]       [,2]       [,3]
[1,] -240.50822 -14.413248  56.851078
[2,]  -14.41325  -1.925805   1.305431
[3,]   56.85108   1.305431 -18.644832

> V
           [,1]      [,2]       [,3]
[1,]  0.1277794 -0.726718  0.3387383
[2,] -0.7267180  4.678188 -1.8883329
[3,]  0.3387383 -1.888333  0.9542885

> se
[1] 0.3574625 2.1629120 0.9768769
```

A comparison of the fitted ggLogis and regular logistic distributions is shown in Figures

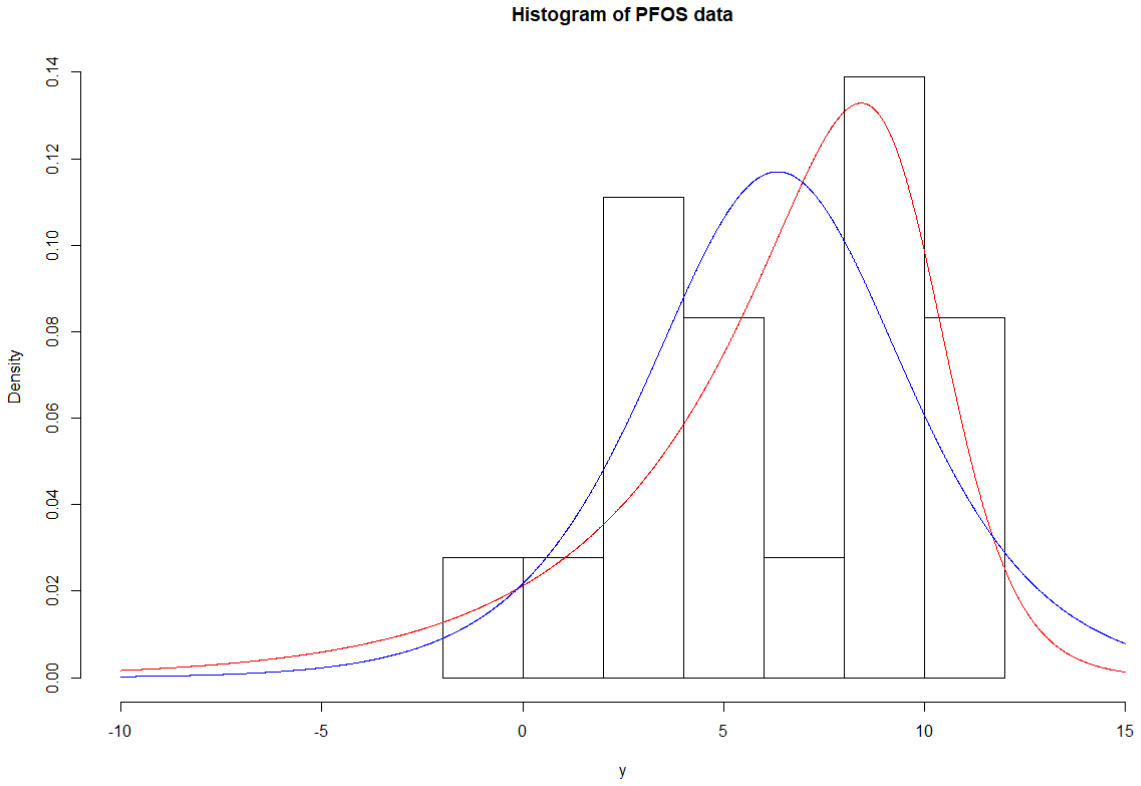

Figure 3. Histogram of log-PFOS data with fitted ggLogis distribution (red curve) and regular logistic distribution (blue curve) overlaid.

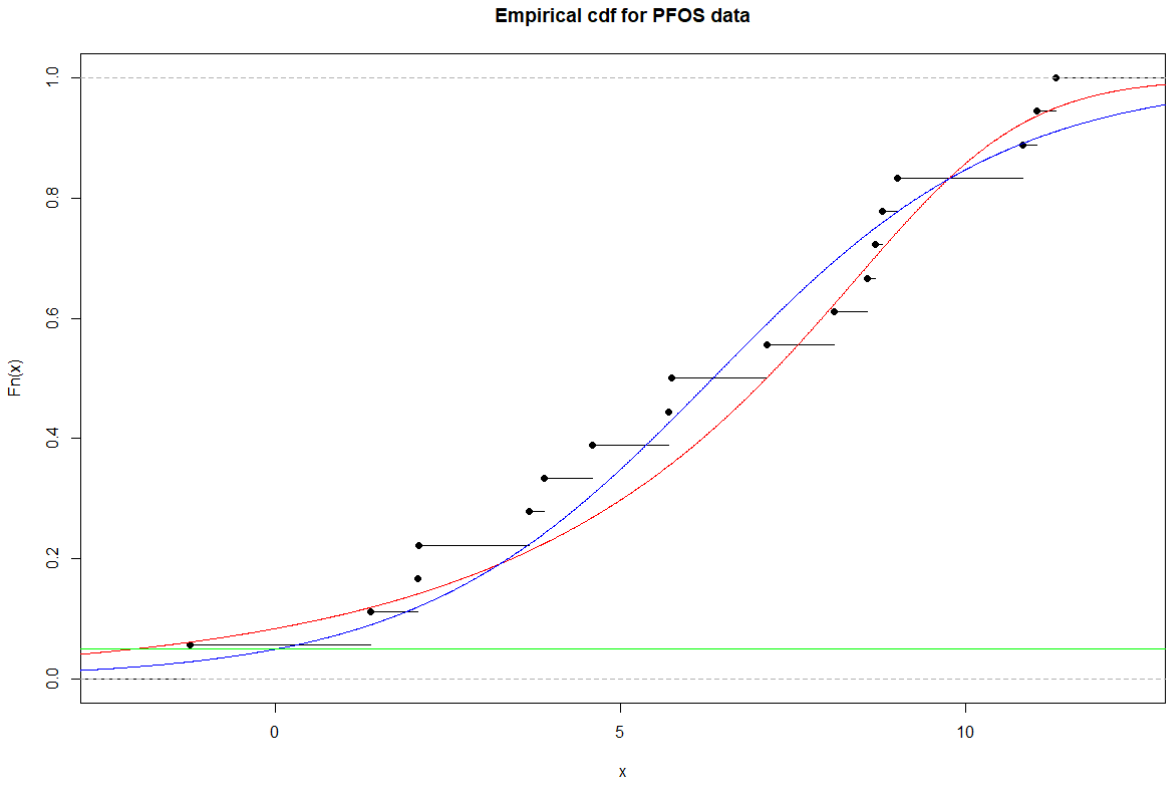

Figure 4. Empirical cdf of log-PFOS data (step-function) with fitted ggLogis distribution (red curve) and regular logistic distribution (blue curve) overlaid. Green horizontal line is at the 0.05 probability level.

5. QUANTILES OF THE ggLogis DISTRIBUTION

The quantiles of any gamma-generated distribution are readily obtained from the definition given by Equation 1. Let ξ_p be the p^{th} quantile of the gamma-generated distribution given by Equation 1 and let $s = -\ln\left[1 - F_X\left(\xi_p\right)\right]$. Thus,

$$G_Y(\xi_p) = \frac{1}{\Gamma(a)} \int_0^{-\ln\left[1-F_X\left(\xi_p\right)\right]} t^{a-1}\, e^{-t}\, dt \;=\; p \qquad (11)$$

or

$$\frac{1}{\Gamma(a)} \int_0^{s} t^{a-1}\, e^{-t}\, dt \;=\; p$$

It is evident from Equation 11 that s is the p^{th} quantile of a gamma distribution having shape parameter a and scale parameter 1, call it γ_p. Hence $\gamma_p = -\ln\left[1 - F_X\left(\xi_p\right)\right]$ which is readily solved for ξ_p (Equation 12).

$$\xi_p = F_X^{-1}\left(1 - e^{-\gamma_p}\right) \tag{12}$$

In R, this is a single-line statement:

```
qlogis(1-exp(-qgamma(p,shape=mle$par[1],scale=1),
     location=mle$par[2],scale=mle$par[3]))
```

where `mle$par` is the vector of parameter estimates obtained from the code in Box 1.

Thus, for the PFOS example above, the HC_5 is:

```
> E<-qlogis(1-exp(-qgamma(0.05,shape=mle$par[1],scale=1)),
+           location=mle$par[2],scale=mle$par[3])
> exp(E)    #  exponentiate to obtain HC5 of untransformed data

[1] 0.1358724
```

Thus, we obtain an estimated HC_5 of 0.1359 from the `ggLogis` distribution compared with 1.05 from the standard logistic distribution.

We next consider confidence intervals for the estimated HC_x .

6. CONFIDENCE INTERVALS FOR AN HCx ESTIMATED FROM THE ggLogis DISTRIBUTION

Most software tools for performing SSD calculations use bootstrapping to obtain confidence intervals for an estimated HC_x (Fox et al. *submitted*). While there is nothing inherently wrong with this, depending on the distribution fitted the method can be time-consuming. When the Hessian matrix is available from the distribution-fitting algorithm (as is the case using the code in Box 1), an approximate confidence interval can be constructed using the delta-method as described below.

Let Σ be the p x p covariance matrix of the parameter estimates $\hat{\Theta}$ for the fitted `ggLogis` distribution where p (=3 here) is the number of model parameters. The covariance matrix of a function $h\left(\hat{\Theta}\right)$ is approximately:

$$Cov\left[h\left(\hat{\Theta}\right)\right] \approx \left[\nabla h\left(\hat{\Theta}\right)\right]^T \Sigma \left[\nabla h\left(\hat{\Theta}\right)\right] \tag{13}$$

where $\nabla(\cdot)$ is the gradient operator. In our case, the function $h\left(\hat{\Theta}\right)$ is given by Equation 12. Since no closed form expression is available for $h\left(\hat{\Theta}\right)$, a numerical approximation will be used. This can be achieved in R using the `grad` function from the `numDeriv` package as shown in Box 2. An approximate $(1-\alpha)100\%$ confidence interval for HC_x is obtained using Equation 14.

$$\widehat{HC}_x \pm t_{n-p,\alpha/2}\, SE\left[\widehat{HC}_x\right] \tag{14}$$

where $t_{n-p,\alpha/2}$ is the quantile from a central T-distribution having *n-p* degrees of freedom (*p* is the number of estimated parameters) and $SE\left[\widehat{HC}_x\right]$ is obtained as the square root of the scalar result in Equation 13.

```
require(numDeriv) # load required package

# create a function to compute the HCx
hcx<-function(p,par){

  E<-qlogis(1-exp(-qgamma(p,shape=par[1],scale=1)),
            location=par[2],scale=par[3])
  return(exp(E))
}

# Evaluate the gradient of the hcx function at the mle for the ggLogis
       g<-grad(hcx,p=0.05,mle$par)  # for the HC5

> g  #  print the gradient vector
[1]  5.6716675  0.1358724 -1.4725093

#  Now compute the approx. SE of the estimated HCx (Equation 13)
se<-sqrt((g)%*%V%*%g)
> se
          [,1]
[1,] 0.4934323

# Compute limits of approx. 95% CI using Equation 14.

> UL<- hcx(0.05,mle$par) + qt(0.975,n-3)          # upper limit

> LL<- max(0,hcx(0.05,mle$par) - qt(0.975,n-3))   # lower limit

>  c(LL,UL)
[1] 0.000000 2.267322
```

Box 2. R code to obtain approximate confidence limits for the HC_x

We see from Box 2 the approximate 95% confidence limits for the HC_5 are {0; 2.267}. Using the `ssdtools` shiny app to fit a standard logistic distribution to the PFOS data resulted in an estimated HC_5 of 1.05 with approximate 95% confidence limits of {0.06; 21.1} based on 5,000 bootstrap samples. We note that the respective confidence intervals overlap and that the `ssdtools` confidence interval is almost 10x wider than the confidence interval from the `ggLogis` distribution. This situation is not improved much if model averaging of the log-logistic, gamma, log-gumbel, and Weibull distributions is used. In this case the point estimate of the HC_5 is 0.453 with approximate 95% confidence limits of {0.03; 18.7}.

REFERENCES

Castellares, F, Santos, MAC, Montenegro, LC, Cordeiro, GM. (2015) A Gamma-Generated Logistic Distribution: Properties and Inference, *American Journal of Mathematical and Management Sciences,* 34:14-39.

Gupta, R. C., Gupta, P. L., and Gupta, R. D., (1998). Modeling failure time data by Lehman alternatives. *Communications in Statistics, Theory and Methods* 27, 887–904.

Mudholkar, G.S., Srivastava, D.K. (1993). Exponentiated Weibull family for analyzing bathtub failure-ratedata. *IEEE Transactions on Reliability*, 42(2): 299–302.

Zografos, K., and Balakrishnan, N. (2009) On families of beta- and generalized gamma-generated distributions and associated inference. *Statist. Methodol.* 6, pp. 344–362.