PharmaSUG 2023 - Paper SI-016

Fitting Logitoid-Normal distributions with MLE estimate by SAS SEVERITY and FCMP procedures

Lili Huang, Bristol Myers Squibb, Berkeley Heights, NJ;
Helen Dong, Bristol Myers Squibb, Berkeley Heights, NJ;
Yuanyuan Liu, Bristol Myers Squibb, Seattle, WA

ABSTRACT

The real-world data in manufacture are usually non-normally distributed, and the behavior of statistical procedure applied depends on the distribution family from which the data are. The knowledge of the distribution family is necessary in exploration of the behavior. The SEVERITY procedure of SAS is capable of fitting distributions with MLE estimate of parameters. One curb of PROC SEVERITY is that the default pool of probability distribution models is quite limited. Distributions such as Johnson family (Johnson Su, Johnson Sl, and Johnson Sb), SHASH, and Logit-Normal are potentially applicable to the manufacture data, unfortunately they are not available for PROC SEVERITY in the latest version SAS/ETS® 14.3. In this paper, four Logitoid-Normal distributions are used as examples in contrast with and Normal distribution to demonstrate that customized distributions could be defined with FCMP procedure, and hence the distribution model parameters can be fitted using PROC SEVERITY.

INTRODUCTION

The chemistry, manufacturing, and controls (CMC) functionality in pharmaceutical industry plays an instrumental part in drug development. The justification of specification of the drug product is expected to elaborate the reliability and robustness of the system and validity of the acceptable criteria of the product. The statistical inference of some aspects of the underlying probabilistic process that generate the data would be of great help towards a deeper knowledge of the behavior of manufacture data and a better understanding and communicating of the process control.

In SAS, the SEVERITY procedure can be used to fit any arbitrary continuous probability distribution if the distribution model is appropriately defined. However, only ten distribution models including Burr, Exponential, Gamma, Pareto, Generalized Pareto, Inverse Gaussian, Lognormal, Tweedie, Scaled Tweedie and Weibull were predefined in the latest SAS version (SAS Institute Inc. 2023). Any other distribution of interest must be defined with the FCMP procedure before it can be invoked by the SEVERITY procedure.

The Normal, Lognormal, Johnson family (Johnson 1949), SHASH (Jones and Pewsey 2009), and Logit-Normal (Atchison and Shen 1980) distributions are gaining preference with manufactured data among numerous continuous distribution models. Normal and Logitoid-Normal distributions, which are simple and comparable in number of parameters, are used as examples in this paper to demonstrate the possibility of customized distribution fitting in SAS using the SEVERITY and FCMP procedures.

STATISTICAL MODELS

PROBABILITY DISTRIBUTION FITTING

The goal of fitting the data to a certain distribution is to identify such a distribution model with parameter values so that they could give the best description of data, which in turn will entail the plausible inference made based on this profile. There are different methods to calculate the parameters in a distribution model, such as method of moments, maximum spacing estimation, and maximum likelihood estimation (Casella and Roger 2002). The maximum likelihood estimation is used in this paper since it is the default method of distribution fitting in the SAS SEVERITY procedure.

The likelihood function of a distribution to a series of sample data is defined as,

$$L(\boldsymbol{\theta}|\mathbf{x}) = \prod_{i=1}^{n} f_X(x_i|\boldsymbol{\theta}) \quad (1)$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_k)$ are the k parameters in the distribution probability density function (PDF) $f_X(x|\boldsymbol{\theta})$ of random variable X, and $\mathbf{x} = (x_1, x_2, \cdots, x_n)$ are the values of n observed sample data.

The distribution parameter $\widehat{\boldsymbol{\theta}} = \left(\hat{\theta}_1, \hat{\theta}_2, \cdots, \hat{\theta}_k\right)$ that maximize the likelihood function over the parameter space, known as maximum likelihood estimator (MLE) by the following expression,

$$\hat{L} = L\left(\widehat{\boldsymbol{\theta}}|\mathbf{x}\right) = \max_{\boldsymbol{\theta}} \prod_{i=1}^{n} f_X(x_i|\boldsymbol{\theta}) \tag{2}$$

A classical approach of decision making is when several distribution models are in consideration, the Akaike information criterion (AIC) (Cavanaugh and Neath 2019) is generally used to aid in selecting the most appropriate ones for the sample data. The AIC is defined as,

$$\mathrm{AIC} = -2\ln\left(\hat{L}\right) + 2k \tag{3}$$

where k is the number of parameters in the distribution model and $\hat{L}$ is the maximum likelihood of the distribution to the sample data. A smaller AIC value indicates a potential better fitting. When sample size n is small, an adjusted formula of AIC, known as AICc (DelSole and Tippett 2021) is utilized instead,

$$\mathrm{AICc} = \mathrm{AIC} + \frac{2k(k+1)}{n-k-1} = -2\ln\left(\hat{L}\right) + \frac{2kn}{n-k-1} \tag{4}$$

AICc converges to AIC for large samples but gives more accurate evaluation of distributions for small samples.

LOGITOID-NORMAL DISTRIBUTIONS

X and Y are continuous random variables. Suppose $\mathrm{Y} = g(\mathrm{X})$, where $g(\cdot)$ is a Logit like function and Y is from a normal distribution, then X is of a Logitoid-Normal distribution.

The simplest Logit-Normal distribution (Atchison and Shen 1980) introduced here is $Y = \ln[X/(100 - X)]$. The probability density function (PDF) of X is,

$$f_X(x|\mu,\sigma) = \begin{cases} 0 & x \le 0 \\ \frac{1}{\sigma\sqrt{2\pi}} \frac{100}{x(100-x)} e^{-\frac{\left[\ln\left(\frac{x}{100-x}\right)-\mu\right]^2}{2\sigma^2}} & 0 < x < 100 \\ 0 & x \ge 100 \end{cases} \tag{5}$$

and the cumulative distribution function (CDF) of X is,

$$F_X(x|\mu,\sigma) = \begin{cases} 0 & x \le 0 \\ \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{\ln\left(\frac{x}{100-x}\right)-\mu}{\sqrt{2}\sigma}\right) & 0 < x < 100 \\ 1 & x \ge 100 \end{cases} \tag{6}$$

where $\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-\frac{t^2}{2}} dt$ is the Gauss error function.

It is commonly known that normal distribution has two distribution parameters μ and σ with $-\infty < \mu < \infty$ and $\sigma > 0$. Likewise, Logit-Normal distribution has the parameters μ and σ as well. The standard logit function $\ln[x/(1-x)]$ with x between 0 and 1 is not used in this paper. Instead, a variant of the logit function is taken: $\ln[x/(100-x)]$ to enlarge the applicable range for data larger than 0 and smaller than 100. To further widen the application range of data, three other Logitoid-Normal distributions are proposed in this work: Upper-bounded Logit-Normal (UBLN), Lower-bounded Logit-Normal (LBLN) and Double-bounded Logit-Normal (DBLN) distributions.

In the Upper-bounded Logit-Normal distribution, an upper bound parameter u is included in the Logit like function, that is, $Y = \ln[X/(u - X)]$. Then the probability density function (PDF) of X is,

$$f_X(x|u,\mu,\sigma) = \begin{cases} 0 & x \le 0 \\ \frac{1}{\sigma\sqrt{2\pi}} \frac{u}{x(u-x)} e^{-\frac{\left[\ln\left(\frac{x}{u-x}\right)-\mu\right]^2}{2\sigma^2}} & 0 < x < u \\ 0 & x \ge u \end{cases} \tag{7}$$

and the cumulative distribution function (CDF) of X is,

$$F_X(x|u,\mu,\sigma) = \begin{cases} 0 & x \le 0 \\ \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{\ln\left(\frac{x}{u-x}\right)-\mu}{\sqrt{2}\sigma}\right) & 0 < x < u \\ 1 & x \ge u \end{cases} \quad (8)$$

The upper-bounded Logit-Normal distribution has three distribution parameters u, μ and σ with $u > 0$, $-\infty < \mu < \infty$ and $\sigma > 0$. It is applicable for data larger than 0 in the distribution fittings.

In the Lower-bounded Logit-Normal distribution, a lower bound parameter l is included in the Logit like function, that is, $Y = \ln[(X-l)/(100-X)]$. Then the probability density function (PDF) of X is,

$$f_X(x|l,\mu,\sigma) = \begin{cases} 0 & x \le l \\ \frac{1}{\sigma\sqrt{2\pi}} \frac{100-l}{(x-l)(100-x)} e^{-\frac{\left[\ln\left(\frac{x-l}{100-x}\right)-\mu\right]^2}{2\sigma^2}} & l < x < 100 \\ 0 & x \ge 100 \end{cases} \quad (9)$$

and the cumulative distribution function (CDF) of X is,

$$F_X(x|l,\mu,\sigma) = \begin{cases} 0 & x \le l \\ \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{\ln\left(\frac{x-l}{100-x}\right)-\mu}{\sqrt{2}\sigma}\right) & l < x < 100 \\ 1 & x \ge 100 \end{cases} \quad (10)$$

The lower-bounded Logit-Normal distribution has three distribution parameters l, μ and σ with $l < 100$, $-\infty < \mu < \infty$ and $\sigma > 0$. It is applicable for data smaller than 100 in distribution fittings.

In the Double-bounded Logit-Normal distribution, upper bound and lower bound parameters u and l are included in the Logit like function, that is, $Y = \ln[(X-l)/(u-X)]$. Then the probability density function (PDF) of X is,

$$f_X(x|u,l,\mu,\sigma) = \begin{cases} 0 & x \le l \\ \frac{1}{\sigma\sqrt{2\pi}} \frac{u-l}{(x-l)(u-x)} e^{-\frac{\left[\ln\left(\frac{x-l}{u-x}\right)-\mu\right]^2}{2\sigma^2}} & l < x < u \\ 0 & x \ge u \end{cases} \quad (11)$$

and the cumulative distribution function (CDF) of X is,

$$F_X(x|u,l,\mu,\sigma) = \begin{cases} 0 & x \le l \\ \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{\ln\left(\frac{x-l}{u-x}\right)-\mu}{\sqrt{2}\sigma}\right) & l < x < u \\ 1 & x \ge u \end{cases} \quad (12)$$

The Double-bounded Logit-Normal distribution has four distribution parameters u, l, μ and σ with $l < u$, $-\infty < \mu < \infty$ and $\sigma > 0$. It is mathematically equivalent to Johonson's SB distribution (Johnson 1949) only with different parameterization and is applicable for all numerical data in the distribution fittings.

SAS PROGRAMMING

CUSTOMIZED LOGITOID-NORMAL DISTRIBUTIONS IN THE FCMP PROCEDURE

Any customized distribution other than those of the default set with PROC SEVERITY must be defined in the FCMP procedure first before they could be used by the SEVERITY procedure. Any distribution model enabled by PROC SEVERITY consists of a set of functions and subroutines that are defined through FCMP procedure. The FCMP procedure is part of Base SAS software. Each function or subroutine must be named as <*distribution-name*>_<*keyword*>, where *distribution-name* is the identifying short name of the distribution and keyword identifies one of the functions or subroutines. Among all functions and subroutines which can be defined in the FCMP procedure, either the CDF or the LOGCDF (Log of CDF) must be defined, and

either the PDF or the LOGPDF (Log of PDF) must be defined, others are optional.

All the complied functions and subroutines of customized distributions, which are defined between statements of proc FCMP and quit, will be stored in the *dists* package of work.loginms library using the following FCMP procedure,

```
proc fcmp outlib=work.logitnms.dists;

quit;
```

Using Upper-bounded Logit-Normal distribution as an example, the PDF function is defined as,

```
function ulogitnm_pdf(x,Upper,Mu,Sigma);
   if ((x<=0) or (x>=Upper)) then v = 0;
   else do;
      y = log(x/(Upper-x));
      c = Upper / (x*(Upper-x));
      v = c * exp(-(y-Mu)**2/(2*Sigma**2))/(Sigma * sqrt(2*constant('PI')));
   end;
   return ( v );
endsub;
```

In this PDF definition, the distribution name is ulogitnm and three distribution parameters are Upper, Mu, and Sigma, which are corresponding to u, μ and σ respectively in equation. The CDF function of Upper-bounded Logit-Normal distribution is defined as,

```
function ulogitnm_cdf(x,Upper,Mu,Sigma);
   if (x<=0) then v = 0;
   else if (x>=Upper) then v = 1;
   else do;
      y = log(x/(Upper-x));
      z = (y - Mu) / Sigma;
      v = 0.5 + 0.5*erf(z/sqrt(2));
   end;
   return ( v );
endsub;
```

The lower boundary subroutine below is necessary for the Upper-bounded Logit-Normal distribution since $u > 0$ and $\sigma > 0$ are required in the model,

```
subroutine ulogitnm_lowerbounds(Upper,Mu,Sigma);
   outargs Upper, Mu, Sigma;
   Upper = 0; Mu = .; Sigma = 0;
endsub;
```

The subroutine of parameter initialization is not required but is useful in most cases when multiple local minima exist in the likelihood function of the distribution. In the Upper-bounded Logit-Normal distribution, u must be larger than the maximum value of the sample data, so the parameter initialization subroutine for the can be defined as,

```
subroutine ulogitnm_parminit(dim,x[*],nx[*],F[*],Ftype,Upper,Mu,Sigma);
   outargs Upper, Mu, Sigma;
   MAXX = x[1]; MINX = x[1];
   do i = 2 to dim;
      if (x[i]>MAXX) then MAXX = x[i];
      if (x[i]<MINX) then MINX = x[i];
   end;
   Upper = MAXX + 0.01*abs(MAXX-MINX);
   NTOT = nx[1];
   SUMY = nx[1] * log(x[1]/(Upper-x[1]));
   SUMY2 = nx[1] * log(x[1]/(Upper-x[1]))**2;
   do i = 2 to dim;
      NTOT = NTOT + nx[i];
      SUMY = SUMY + nx[i] * log(x[i]/(Upper-x[i]));
```

```
        SUMY2 = SUMY2 + nx[i] * log(x[i]/(Upper-x[i]))**2;
    end;
    MEANY = SUMY / NTOT; MEANY2 = SUMY2 / NTOT;
    Mu = MEANY;
    Sigma = sqrt(MEANY2 - MEANY**2);
endsub;
```

All the SAS codes of functions and subroutines for the Normal and four Logitoid-Normal distributions can be found in the supplementary material of this paper.

DISTRIBUTION FITTING USING THE SEVERITY PROCEDURE

The CMPLIB option, which specify the data set that contain complied subroutines and functions defined in the FCMP procedure, must be added as below so that the distribution models can be found in the SEVERITY procedure,

```
options cmplib=(work.logitnms cmplib);
```

In SASHELP.QTR1001, there are 40 data available for variables S0381, S0382 and S0384 and these data will be used as sample data to demonstrate the distribution fitting procedures that are pertinent to this work.

```
data inds(keep=t s0381 s0382 s0384);
   set sashelp.qtr1001;
   where nmiss(s0381,s0382,s0384)<1;
run;
```

In the SEVERITY procedure, the variable whose data are to be fitted should be given in the LOSS statement, and all the distribution names should be put in the DIST statement. The NLOPTIONS statement is used to control the optimization procedure details during the fitting.

```
proc severity data=inds print=all;
   loss S0384;
   dist normal logitnm ulogitnm llogitnm dlogitnm;
   nloptions tech=nmsimp absfconv=1.0e-5 maxiter=10000 maxfunc=500000;
run;
```

One SEVERITY procedure can be applied for fitting multiple distributions for a given set of data of interest. The fitting results for S0384 data using the above procedure are shown in Table 1.

	Statistics			Parameters			
Distribution	-2 Log Likelihood	AIC	AICc	Mu	Sigma	Upper	Lower
normal	171.368	175.368	175.692	4.81200	2.06094		
logitnm	166.904	170.904	171.229	-3.08139	0.46892		
ulogitnm	164.455	170.455	171.122	-0.09388	0.91199	10.11085	
llogitnm	166.269	172.269	172.936	-3.48339	0.68659		1.24605
dlogitnm	143.777*	151.777*	152.920*	-0.65732	2.39709	8.65855	2.10523

Table 1 Distribution fitting results for S0384. The asterisk (*) masks the best model according to each column's criterion.

All the statistics (-2*Log Likelihood, AIC, and AICc) in Table 1 indicate that the Double-bounded Logit-Normal is the best one among the five distributions investigated for S0384. It is obvious that the Double-bounded Logit-Normal distribution will tend to yield a bigger Likelihood (smaller -2*Log Likelihood) value than the Logit-Normal, Upper-bounded Logit-Normal and Lower-Bounded Logit-Normal distributions. The AIC and AICc values do not perform likewise across the set of five distributions, though both AIC and AICc have the number of parameters accounted for to penalize the complexity of distribution functions, the later one has sample sizes weighted in measure of how well the distribution performs in fitting the data.

The SEVERITY procedure is shown as follows for fitting S0382 data with the four Logitoid-Normal and Normal distributions,

```
proc severity data=inds print=all;
   loss S0382;
   dist normal logitnm ulogitnm llogitnm dlogitnm;
   nloptions tech=nmsimp absfconv=1.0e-5 maxiter=10000 maxfunc=500000;
run;
```

The fitting results for S0382 data using the above procedure are shown in Table 2.

	Statistics			Parameters			
Distribution	-2 Log Likelihood	AIC	AICc	Mu	Sigma	Upper	Lower
normal	309.800	313.800	314.124	45.02498	11.62967		
logitnm	309.860	313.860	314.183	-0.20207	0.49939		
ulogitnm	304.026	310.026	310.693	-6.25836	0.24858	22837	
llogitnm	303.440	309.440*	310.107*	-1.01121	0.80140		22.57314
dlogitnm	303.073*	311.073	312.216	-1.38437	0.62217	134.35647	20.22296

Table 2 Distribution fitting results for S0382. The asterisk (*) masks the best model according to each column's criterion.

The Double-bounded Logit-Normal is the best distribution for S0382 according to the -2*Log Likelihood statistics, while the Lower-bounded Logit-Normal is selected as the best one according to the AIC and AICc criterions.

When it comes to fitting S0381 data, the Logit-Normal and Lower-bounded Logit-Normal distributions are not applicable since some of the data are greater than 100. The SEVERITY procedure is displayed as follows for S0381 data by excluding llogitnm and logitnm in DIST statement,

```
proc severity data=inds print=all;
   loss S0381;
   dist normal ulogitnm dlogitnm;
   nloptions tech=nmsimp absfconv=1.0e-5 maxiter=10000 maxfunc=500000;
run;
```

The fitting results for S0381 data using the above procedure are shown in Table 3.

	Statistics			Parameters			
Distribution	-2 Log Likelihood	AIC	AICc	Mu	Sigma	Upper	Lower
normal	348.353	352.353	352.677	66.65101	18.83031		
ulogitnm	342.433	348.433*	349.100*	-5.91231	0.27306	23783	
dlogitnm	340.753*	348.753	349.896	-0.90300	0.97096	137.00112	33.80905

Table 3 Distribution fitting results for S0381. The asterisk (*) masks the best model according to each column's criterion.

The Upper-bounded Logit-Normal is selected as the best one among the three candidate distributions used for investigating S038, per AIC and AICc criterions. It is notifiable that initial parameter values could been assigned arbitrarily as input in the SEVERITY procedure as follows,

```
data inits;
```

```
    length _TYPE_ $8 _MODEL_ $16;
    _TYPE_ = 'EST';
    _MODEL_ = "normal"; Mu = 60; Sigma = 1; output;
    _MODEL_ = "ulogitnm"; Upper = 200000; Mu = -8; Sigma = 0.3; output;
    _MODEL_ = "dlogitnm"; Upper = 150; Lower = 30; Mu = -1; Sigma = 1; output;
run;

proc severity data=inds inest=inits print=all;
   loss S0381;
   dist normal ulogitnm dlogitnm;
   nloptions tech=nmsimp absfconv=1.0e-5 maxiter=10000 maxfunc=500000;
run;
```

	Statistics			Parameters			
Distribution	-2 Log Likelihood	AIC	AICc	Mu	Sigma	Upper	Lower
normal	348.353	352.353	352.677	66.64911	18.83073		
ulogitnm	342.428	348.428*	349.095*	-8.04396	0.27255	199999	
dlogitnm	340.753*	348.753	349.896	-0.90286	0.97076	136.99256	33.80109

Table 4 Distribution fitting results for S0381 with input initial parameter values. The asterisk (*) masks the best model according to each column's criterion.

The improvement in the fitting results for S0381 is barely observed when use input initial parameter values as shown in Table 4, which indicates the initial parameters assigned in the subroutines of parameter initialization defined in the FCMP procedure works well for cases in this work.

CONCLUSION

Three Logitoid-Normal distributions are introduced in this paper and are defined in the SAS FCMP procedure with their respective PDF and CDF functions along with subroutines of parameter boundaries and initializations. Normal and the three Logitoid-Normal distributions are investigated for three distinctive sets of data in the SASHELP library using the SEVERITY procedure. The fitting results demonstrate the instrumental role that number of parameters have played in selection of best distribution models when AIC or AICc is used as criterion.

The distribution models defined in the FCMP procedure in this paper has been working very well with the SEVERITY procedure regardless the initial value input of parameter. They could be used as long as the distributions are applicable for the data to be fitted. The script of model defining part could be used as references for any other distributions intended for fitting by PROC SEVERITY in SAS.

REFERENCES

SAS Institute Inc. 2023. *SAS/ETS® 15.3 User's Guide.* Cary, NC: SAS Institute Inc.

Johnson, N. L. 1949. "Systems of Frequency Curves Generated by Methods of Translation." *Biometrika.* 36(1/2): 149.

Jones, M. C. and Pewsey, A. 2009. "Sinh-arcsinh distributions." *Biometrika.* 96(4): 761.

J Atchison and SM Shen. 1980. "Logistic-normal distributions: Some properties and uses." *Biometrika.* 67(2): 261.

Casella, George and Roger L. Berger. 2002. *Statistical Inference.* 2nd ed. Australia ; Pacific Grove, CA: Thomson Learning.

Joseph E. Cavanaugh and Andrew A. Neath. 2019. "The Akaike information criterion: Background,

derivation, properties, application, interpretation, and refinements." *WIREs Computational Statistics.* 11(3): 1460.

Timothy DelSole and Michael K.Tippett. 2021. "Correcting the corrected AIC." *Statistics & Probability Letters.* 173: 109064.

ACKNOWLEDGMENTS

We would like to thank the employer Bristol Myers Squibb for the generous support over this work and approval to present this paper at PharmaSUG.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Lili Huang
Bristol Myers Squibb
Email: Lili.Huang@bms.com