## FITTING THE VERSATILE LINEARIZED, COMPOSITE, AND GENERALIZED LOGISTIC PROBABILITY DISTRIBUTION TO A DATA SET R.J. Oosterbaan, 06-08-2019. On <u>www.waterlog.info</u> public domain.

## Abstract

The logistic probability distribution can be linearized for easy fitting to a data set using a linear regression to determine the parameters.

The logistic probability distribution can also be generalized by raising the data values to the power P that is to be optimized by an iterative method based on the minimization of the squares of the deviations of calculated and observed data values (the least squares or LS method). This enhances its applicability.

When P<1 the logistic distribution becomes skew to the right and when P>1 it becomes skew to the left. When P=1, the distribution is symmetrical and quite like the normal probability distribution.

In addition, the logistic distribution can be made composite, that is: split into two parts with different parameters. Composite distributions can be used favorably when the data set is obtained under different external conditions influencing its probability characteristics.

# Contents

1. The standard logistic cumulative distribution function (CDF) and its linearization

- 2. The logistic cumulative distribution function (CDF) generalized
- 3. The composite generalized logistic distribution
- 4. Fitting the standard, generalized, and composite logistic distribution to a data set, available software
- 5. Construction of confidence belts
- 6. Constructing histograms and the probability density function (PDF)
- 7. Ranking according to goodness of fit
- 8. Conclusion
- 9. References

# 1. The standard cumulative distribution function (CDF) and its linearization

The cumulative logistic distribution function (CDF) can be written as:

$$Fc = 1 / \{1 + e^{(A^*X + B)}\}$$

where

Fc = cumulative logistic distribution function or cumulative frequency

- $e = base of the natural logarithm (Ln), e = 2.71 \dots$
- X = randomly variable data value
- A = scale parameter
- B = place parameter

The CDF can be rewritten in linear form as:

Ln (1 / Fc) = A\*X + B

so that the parameters A and B can be found from a linear regression of Y = Ln (1 / Fc) on X.

### 2. The logistic cumulative distribution function (CDF) generalized

The CDF of the logistic distribution can be generalized replacing X by  $Z = X^{P}$  so that the linearized CDF becomes:

$$Ln (1 / Fc) = A^*Z + B = A^*X^P + B$$

The value of the power P is to be found by numerical and iterative procedures using the least sum of squares of deviation of observed from theoretical (calculated) Fc values (the least squares or LS method).

When P<1 the generalized logistic distribution becomes skew to the right and when P>1 it becomes skew to the left. This makes the generalized logistic distribution versatile and it can be used for extreme values. When P=1, the distribution is symmetrical and quite similar to the normal distribution.



Fig. 1. Probability density function (PDF) skewed to the left (1st picture, P>1), symmetrical (central picture, P=1), skewed to the right (2nd picture, P<1)

### 3. The composite generalized logistic distribution

For the composite distribution, we split it into two parts with a separation point Q:

a) when Z < Q then

Ln (1 / Fc) = 
$$A_1*Z + B_1 = A_1*X^{P_1} + B_1$$

a) when Z > Q then

Ln (1 / Fc) = 
$$A_2 * Z + B_2 = A_2 * X^{P_2} + B_2$$

# 4. Fitting the generalized and composite logistic distribution to a data set, available software

The question that remains is: how to determine Fc?

Fc is normally found from the following equation (also called plotting position):

$$Fc = R/(N+1)$$

where

R= the rank number of the respective X values arranged in an ascending order N= the number of data

Thus the generalized and/or composite logistic CDF can now be fitted to a data set with the free CumFreq program [Ref. 1] respectively CumFreqA program [Ref. 2]. See Figures 2 and 3.





Fig. 2 Example of the generalized logistic distribution.

A = -4.58 B = 10.8 P = 0.21



CumFreq program at www.waterlog.info/cumfreq.htm



 $A_1 = -\ 0.471, \ B_1 = 4.99, \ P_1 = 0.58$ 

 $A_2 = - \ 0.202, \ B_2 = 1.87, \ P_2 = 0.60$ 

According to Fig. 3, the rainfalls smaller than 65 mm/day occur under climatic conditions that differ from those generating rainfalls larger than 65 mm/day.

The data used have been derived from Publication 16, ILRI, Wageningen [Ref. 1]

NOTE 1: The goodness of fit is discussed in section 6.

NOTE 2: The CumFreq [Ref. 2] program offers the user the possibility to select the type of probability distribution (Fig.4), while the amplified CumFreqA program [Ref. 3] also offers the possibility to select the type of composite distribution (Fig. 5)

	Best of All Distributions	CIICK GU
	C Normal (standard)	C Fisher-Tippet III
	C Normal (optimized)	C F-T III mirrored
	C Log-normal (standard)	C Frechet (typical)
	C Log-normal (optimized)	C Frechet mirrored
	C Root-normal	C GEV (Gen. Extreme. Value)
	C Square-normal	C Gompertz (generalized)
	C Generalized normal	C Gumbel (standard)
	C Logistic (standard)	C Gumbel generalized
	C Log-logistic	C Gumbel mirrored
-	C Logistic generalized	C Gumbel mirrored generalized
	C Burr (generalized)	C Kumaraswamy generalized
	C Cauchy (standard)	C Laplace (standard)
	C Cauchy generalized	C Laplace generalized
	C Dagum (generalized)	C Student (1 or 2 d.f.)
	C Exponential (standard)	C Pareto-Lomax
	C Exponential generalized	C Weibull (standard)
	C Exponential general, mirr.	C Weibull mirrored

Fig 4. Possibilities to select a non-composite probability distribution in CumFreq and CumFreqA.

≚ CumFreqA cumulative	e frequency analysis				
File Edit		1			
Intro Input C	Output Graphics	;			
Intro Input C File D:\Werkm Title1 Input data Title2 Cumulative Options Allow a co Nr. of Data 1 Nr of Intervals for histogram 5 Threshold 0 (cut-off for data values)	Dutput Graphics Dutput Graphics Dutput Graphics Duppen\WinModels\ a for CumFreq program frequency analysis Domposite distribution 9	CumFreq gr ram son, if it can Serial Nr. 1 2 3 4 5 5 6 7 8 9 9 10 11	oep\CumFreqA be detected Data value 200 158 111 99 92 74 67 65 65 57 56	11R(as:	11Q, density changed, LAST)\DATA trials\PU16B.INP         Select composite distribution         Composite generalized logistic distribution         ©         Best fitting of all composite distribution         Composite logistic distribution         Composite Gumbel distribution         Composite Gumbel mirrored         Logistic + Gumbel mirrored         Logistic + Gumbel mirrored         Gumbel + Logistic         Gumbel + Logistic         Gumbel mirrored + Boisson        The following use generalized distributions         Composite generalized Gumbel distribution         Composite generalized Gumbel distribution
		12	46		Gen, logistic + gen, Gumbel
		13	45		Gen. Gumbel + gen. logistic
		14	40		Gen. Gumbel + gen. Poisson Gen. Poisson + gen. logistic
		15	38	_	Gen. Poisson + gen. Gumbel Generalized Laplace distribution
F				d	lear data Paste help Save-Run Open input

Fig 5. Possibilities to select a composite probability distribution in CumFreqA. The selection space for the number of intervals for histogram analysis is also visible.

NOTE 3 The Cumfreq programs offers a calculator to determine the probability given an X value and vice versa (Fig. 6)

Calculator	
Enter X-value, then click "Go ->", to obtain cumulative frequency and return period	Fig. 6.
X-value Cum. Freq. Ret. Per.	Calculator incorporated in the CumFreq programs.
90% confidence limits of cum. freq. 90% confidence limits of ret. per. Lower Upper Lower Upper	This facilitates the computation of confidence intervals.
Enter cumulative frequency or return period to obtain X-value	
Cum. Freq. Ret. Per. X-value	
0.9 or 10 Go-> 329.40916	
90% confidence limits of X-Value Lower Upper 248.674530 347.186145	

### 5. Construction of confidence belts

In 2 and 3 the 90% confidence belts of the CDF's have been drawn. The confidence intervals are found from the (relative) standard deviation (Sd) of the binomial probability distribution [Ref. 4]:

$$Sd = sqrt \{Fc(1-Fc)/N\},\$$

where Fc is the cumulative (non-exceedance) frequency ( $0 \le Fc \le 1$ ), and N is the number of data.

There are only two events: Fc, the non-exceedance, or (1-Fc), the exceedance, reason why the binomial distribution is applicable.

The determination of the confidence interval of Fc makes use of Student's t-statistic (t) [Ref 5]. Using 90% confidence limits the t-value is close to 1.7 when N>10.

The binomial distribution is symmetrical when Fc=0.5 (in the center of the distribution), but it becomes more skew when Fc approaches 0 or 1. Therefore Fc can be used as a weight factor in the assignation of Sd to U and L (upper and lower confidence limit respectively):

$$U = Fc + 2*1.7 (1-Fc) Sd$$
  
L = Fc - 2\*1.7 Fc.Sd

### 6. Constructing histograms and the probability density function (PDF)

The PDF is found by differentiating the CDF.

CumFreq can show the histogram and corresponding CDF as demonstrated in the following figures.



Fig.7 Histogram and PDF for the data shown in Fig. 2.

In this case the histogram is made up of 5 intervals, but CumFreq offers the user the option to determine the number of intervals by him/herself.



Fig.8 Histogram and PDF for the data shown in Fig. 3. The separation point is Q = 65

# 7. Ranking according to goodness of fit

CumFreq prepares a list of AVERAGES of absolute values of the differences between observed and calculated cumulative frequency values (in %). This is a measure for goodness of fit. See the examples below, in which it is shown that the generalized logistic distribution ranks in the top 10 with an average of 3.12%

F	CumFreq list of parameters of fit	x
	List of AVERAGES of absolute values of the differences between observed and calcuated cumulative frequency values (in %).	*
	The lower the average, the better the distribution fit. Averages are ranked.	
	Name of this file : PU16BBestNoComp.Lst	
	GEV distribution Average = 2.0892	
	Burr distribution generalized Average = 2.2199	=
	Generalized Laplace distribution Average = 2.2931	
	Log-normal distribution optimized Average = 2.3977	
	Log-normal distribution standard Average = 2.7420	
	Frechet (Fisher-Tippet Type 2) distribution Average = 2.8427	
	Generalized normal distribution Average = 2.9574	
	Laplace distribution Average = 2.9875	
K	Generalized logistic distribution Average = 3.1238	
	Log-logistic distribution Average = 3.2363	
	Generalized Gumbel distribution Average = 3.3002	
	Root-normal distribution optimized Average = 3.5853	
	Dagum distribution generalized	-
	4	

Fig. 9 CumFreq ranks the various distributions according to goodness of fit.

👱 CumFreq list of parameters of fit	x
List of AVERAGES of absolute values of the differences between observed and calcuated cumulative frequency values (in %).	Â
The lower the average, the better the distribution fit. Averages are ranked.	
Name of this file : PU16BAll.Lst	-
Composite generalizd logistic distribution Average = 2.0210	-
Generalized Gumbel + Poisson distribution Average = 2.0417	
Composite generalized Gumbel distribution Average = 2.0473	
Generalized Poisson + Gumbel distribution Average = 2.0584	
GEV distribution Average = 2.0892	
Composite generalized exponential distribution Average = 2.1228	
Composite Gumbel distribution Average = 2.1233	
Burr distribution generalized Average = 2.2199	
Generalized Laplace distribution Average = 2.2931	
Gumbel + Poisson distribution Average = 2.3076	
Composite logistic distribution Average = 2.3665	
Log-normal distribution optimized Average = 2.3977	
Generalized Logistic + Poisson distribution	-
↓	

Fig 10 CumFreqA ranks the various distributions, including the composite ones, according to goodness of fit. The composite generalized logistic distribution ranks highest with an index of 2.02 %, which is better than the top rank in Fig. 9 (the GEV distribution with an index of 2.09 %, ranking on the fifth place here)

# 8. Conclusion

The CumFreq program offers the possibility to select the standard and generalized logistic distribution for data fitting.

More details about CumFreq options are discussed in [Ref. 6].

# 9. References

[Ref. 1]	Frequency and Regression Analysis. Chapter 6 in Publication 16, ILRI, Wageningen. Download from: <u>https://www.waterlog.info/pdf/freqtxt.pdf</u>
[Ref. 2]	CumFreq, free calculator for probability distributions. Download from: <u>https://www.waterlog.info/cumfreq.htm</u>
[Ref. 3]	CumFreqA, amplified Cumfreq calculator with emphasis on composite probability distributions. Download freely from: <u>https://www.waterlog.info/cumfreq.htm</u>
[Ref. 4]	Use of the binomial probability distribution for confidence intervals of Cumulative probability distribution functions. On line: <u>https://www.waterlog.info/pdf/binoom.pdf</u>
[Ref. 5]	Use of Student's t-distribution to determine confidence limits given the average and standard deviation of data in a sample. On line: <u>https://www.waterlog.info/t-tester.htm</u>

[Ref. 6] SOFTWARE FOR GENERALIZED AND COMPOSITE PROBABILITY DISTRIBUTIONS. On line:

https://www.researchgate.net/publication/332466331\_SOFTWARE\_FOR\_GENERALIZED\_ AND\_COMPOSITE\_PROBABILITY\_DISTRIBUTIONS