How to derive a probability distribution from a data set using the simple method of plotting positions and the free CumFreq model

R.J. Oosterbaan
July 2022 www.waterlog.info public domain

Abstract
It is customary to determine the parameters of a probability distribution from the data properties like the mean value, the mode, the median, the standard deviation. Some parameters can not be derived from the data set and need to be found numerically. The methods used for the determination are: the method of moments, the maximum spacing estimation, the method of L-moments and, as a last resort, the maximum likelihood method. Plotting positions are seldom used although they can be very effective and, in addition, they show the observed probabilities next to the simulated ones so that the goodness of fit can be assessed without difficulty. In addition, the method allows a generalization of the probability distribution enhancing its versatility. This paper explains how the method of plotting positions is applicable to a large number of probability distributions so that the best fitting distribution to a data set can be detected. This procedure can be performed using the free CumFreq model or the amplified CumFreqA model, which can give generalizations of probability distribution that increase their versatility. The second model offers the extra possibility to detect composite distributions. Both models also construct confidence belts. From the cumulative probability obtained by the plotting position method, the probability density function can be simply derived and compared with the histogram for which the model gives the choice of the number of intervals. The paper gives a number of illustrative examples.

Contents

1. Introduction
It is customary to determine the parameters of a probability distribution from the data characteristics like the mean value, the mode, the median and the standard deviation. Some parameters can not be derived directly from the data set and need to be found numerically. The methods used for the determination are:

– Method of moments (when it is about mean, standard deviation and/or skewness, *Reference 1*)
– Method of L-moments (*Reference 2*)
– Maximum spacing estimation (*Reference 3*)
– Maximum likelihood (*Reference 4*)

There is a possibility to bypass the above methods using the procedure of plotting positions to estimate the probability directly from the data set. This procedure is described in the next section.

2. The Gumbel method of plotting positions

The Gumbel plotting position (Pp) gives an estimate of the cumulative probability (Cp or probability of non-exceedance) for each of the values in a data set.

Before the Pp can be determined the data set must be arranged in ascending order. Each value Xn in this series with n = 1, 2, 3, ... N (where N is the total number of data) is given the PP value n / (N+1).

Gumbel (1954, *Reference 5*) has shown that Pp is an unbiased estimator of the cumulative probability around the mode of the distribution. In literature there exist other estimates, but Makkonen (2006, *Reference 6*) has proved that the Gumbel Pp is the best of all.

Table 1 shows how in CumFreqA the X-values have been ranked in ascending order and the Pp values are determined. Further, the calculated Cp values have been added by fitting a probability distribution in a way that will be explained later.

Table 1. Observed and calculated cumulative probabilities

X-value Ranked	Cumulative probability (%)	
	Pp	Cp calculated
18.0	7.69	9.76
25.0	15.38	12.30
37.0	23.08	21.25
47.0	30.77	38.49
48.0	38.46	41.13
49.0	46.15	44.00
51.0	53.85	55.32
58.0	61.54	58.93
80.0	69.23	70.62
98.0	76.92	79.37
105.0	84.62	82.37
125.0	92.31	89.39

3. Linearization and generalization of probability distributions

With the method of plotting positions, probability distributions can be linearized. Examples are given for the Gumbel and logistic distribution. Briefly some other linearizations are also presented..

3.1. Linearization of the logistic distribution

The cumulative logistic distribution function can be written as:

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

Using the plotting position Pp, being an estimator of the cumulative probability Cp, instead of Cp, the Pp can be rewritten in linear form as:

$$\ln (1 / Pp) \approx A*X + B \qquad \text{(Equation 1)}$$

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

3.2. Linearization of the Gumbel distribution

The Gumbel distribution can be written as:

$$Cp = \exp[-\exp\{-(A*X+B)\}]$$

where Cp is the cumulative probability distribution.

Taking the natural log (ln) of Cp gives:

$$\ln(Cp) = -\exp\{-(A*X+B)\} \quad \text{or} \quad -\ln(Cp) = \exp\{-(A*X+B)\}$$

Taking the natural log once again yields:

$$\ln\{-\ln(Cp)\} = -(A*X+B)$$

or

$$-\ln\{-\ln(Cp)\} = A*X+B \qquad \text{(Equation 2)}$$

Using the plotting position Pp, being an estimator of the cumulative probability Cp, instead of Cp and setting

$$D = B + \ln\{-\ln(Pp)\}$$

we find:

$$A*X + D \approx 0$$

which is the linearized form of the Gumbel distribution.

The parameters A and D can now be found from a linear regression so that the standard Gumbel distribution is fully defined.

3.3. Linearization of other distributions

The following table gives a brief overview of linearizations for other distributions (Oosterbaan, 2020, *Reference 7*).

Table 2. Linearizations of some cumulative probability distributions

Name of distribution	Cumulative probability Cp	Linearization
Cauchy	Cp = (1/pi)*arctan(A*X+B) + 0.5	Cp* = tan{pi*(Cp-0.5)} Cp* = A*X + B
Exponential (Poisson)	Cp = 1 – Exp{– (A*X+B)}	Cp* = – Ln (1–Cp) Cp* = A*X + B
Fisher-Tippet type III	Cp = Exp[-{(C-X)/Exp(-B/A)}^A]	Xt = Ln(C–X) Cp* = Ln{–Ln(Cp)} Cp* = A*Xt + B
Frechet (F-T type II)	Cp = Exp[-{(X-C)/Exp(-B/A)}^A]	Xt = Ln(X–C) Cp* = Ln{–Ln(Cp)} Cp* = A*Xt + B
Gompertz	Cp= 1 – exp[A*{exp(B*X) – 1}]	Xt = exp(B*X) – 1 Cp* = Ln (1–Cp) Cp* = A*Xt
Kumaraswamy	Cp = 1 – {1 – (X/C) ^B} ^ A	Xt = Ln{(X/C)^B} = B*Ln(X/C) Cp* = Ln(1–Cp) Cp* = A*Xt
Laplace, composite Split in two parts Separated by X=Q	X<Q: Cp = 0.5*Exp{A1*(X–B)} X>Q: Cp = 1 - 0.5*exp{A2*(X–B)}	X<Q: Cp* = ln(2Cp) B = –A1*Q Cp* = A1*X + B X>Q : Cp* = ln(0.5) –ln(1-Cp) Cp* = A2*X
Weibull	Fc = 1 - Exp{– (X/C)^A} with C = Exp(–B/A)	Xt = Ln{Ln(X)} Cp* = Ln{–Ln(1-Cp)} Bt = B/A Cp* = A*Xt + Bt

4. Generalization

The generalization is accomplished by a transformation of the data.

A well known transformation is taking the logarithmic value of the data before applying the normal distribution, obtaining the log-normal distribution.

When the data set is skew to the right, the normal distribution cannot be used because it is symmetrical. However, by employing the logarithmic transformation it may happen that the distribution does become normal.

In this article the transformation is realized by raising the data values to the power (exponent) E. When E < 1 the effect is similar to taking the logarithmic value. However, because the E value may have a large range its versatility is greater than only a single log transformation.

Mathematically generalization can be simply accomplished by replacing in the equations of the cumulative probability the X variable by X^E.

5. Examples of probability distributions

5.1 The logistic distribution

The logistic distribution (*Reference 8*) is symmetrical by nature, but by transforming the data raising them to the power E it can become skew to the right when E < 1 or skew to the left when E > 1. This procedure is called generalization.

The power (exponent) E is to be found by numerical optimization minimizing the sum of the absolute values of the differences between plotting position and simulated cumulative probability.

It may be noted that the cumulative probability at X is the probability that the data value is smaller than X.

Figure 1A gives the cumulative probability distribution function (CPDF or CDF or CPF) obtained by the generalized logistic distribution for data on the height of shrubs.

In *figure 1B* the corresponding histogram and probability density function (PDF) is given. Note that the PDF is the derivative of the CPF.

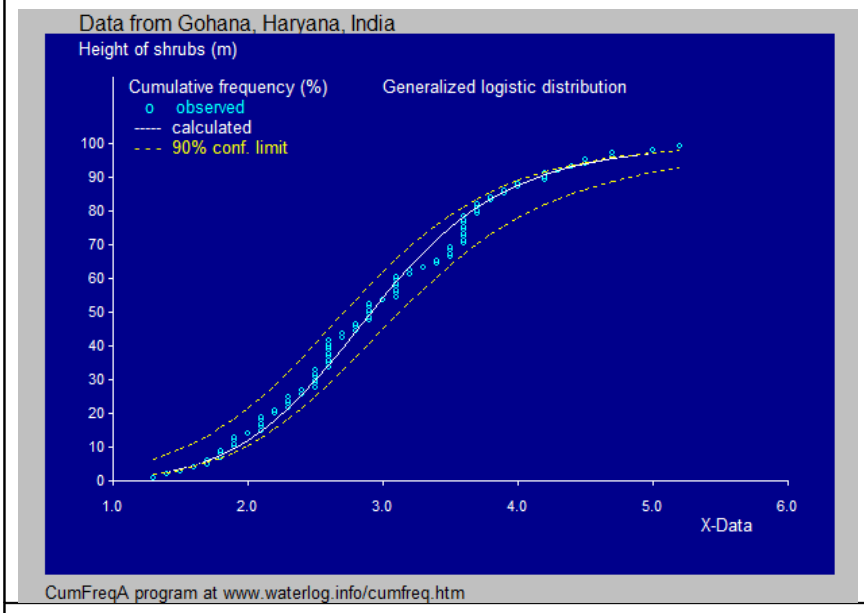


Figure 1A

Generalized logistic CPF of the height of shrubs.

*Cp = 1/{1+ exp (–A*X^E+B)}*

with E = 0.43 (optimized)) while A = –8.47 and B = 13.4 from linear regression (Equation 1)

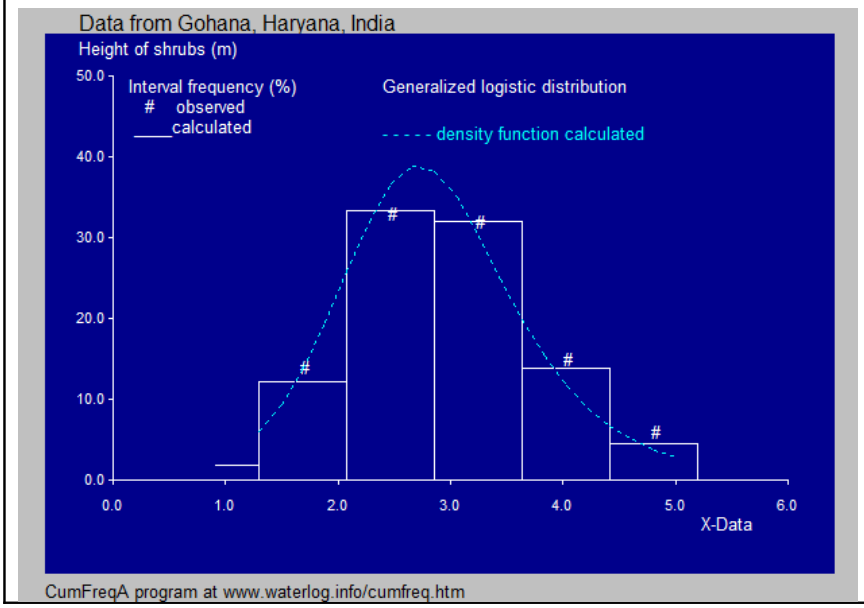


Figure 1B

Histogram and generalized logistic PDF of the height of shrubs.

The PDF is skew to the right, reason why E is less than 1 (see figure 1A*)*

Figure 2A gives the cumulative probability distribution function (CPDF or CDF or CPF) obtained by the generalized logistic distribution for data on the percentage of healthy cows.

In *figure 2B* the corresponding histogram and probability density function (PDF) is given. It is explained that the data are such that a composite probability distribution may be required an that will be further explained in *figure 6C*.

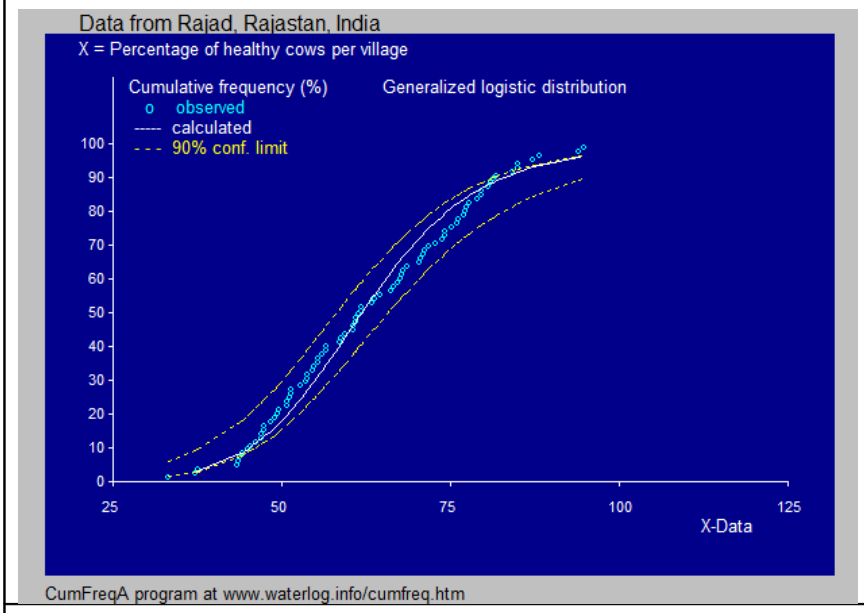


Figure 2A

Generalized logistic CPF of the health of cows..

*Cp = 1/{1+ exp (–A*X^E+B)}*

with E = 2.13 (optimized)) while A = 0.000353 and B = –2.00 from linear regression (Equation 1)

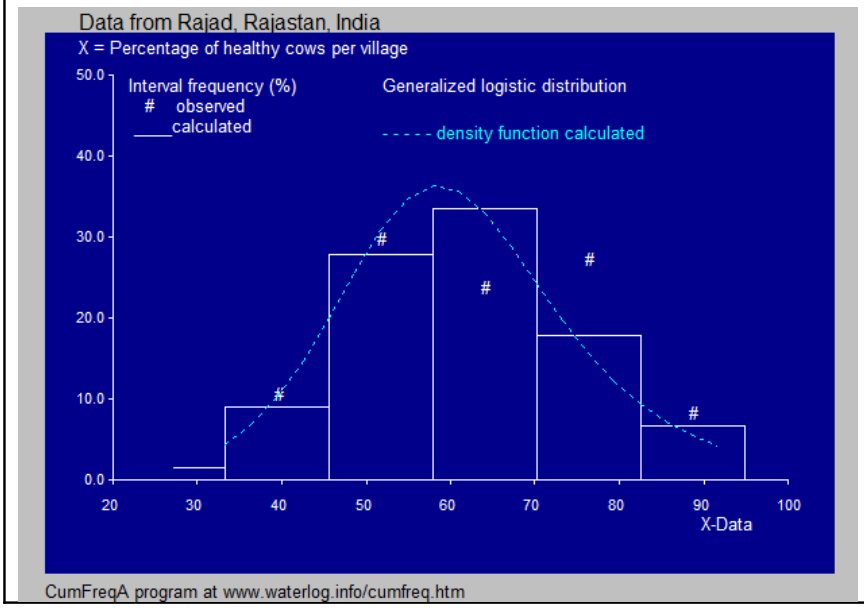


Figure 2B

Histogram and generalized logistic PDF of the health of cows.

The relatively low value of the observed interval frequency (symbol #) between X=60 and X=70 suggests that the distribution is composite as will be explained in figure 6C

Figure 3A gives the cumulative probability distribution function (CPDF or CDF or CPF) obtained by the generalized logistic distribution for data from Mr. Suha that still need to be identified.

In *figure 3B* the corresponding histogram and probability density function (PDF) is given.

I

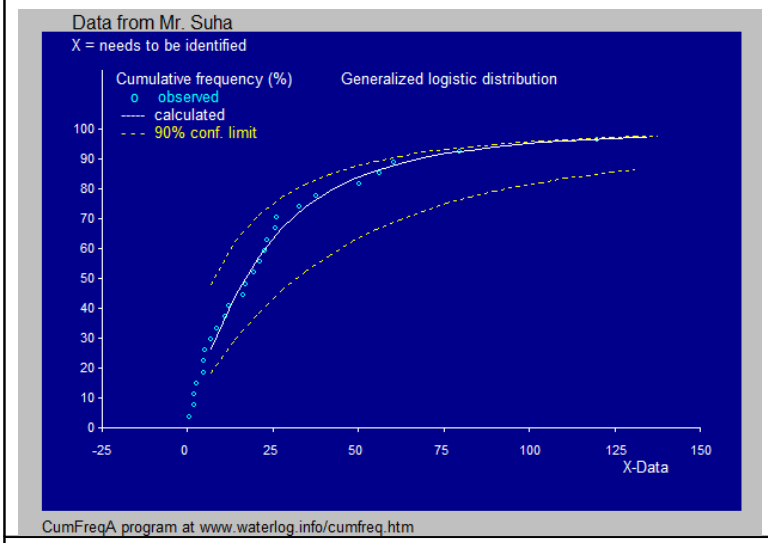


Figure 3A

Generalized logistic CPF of the data that need to be identified.

*Cp = 1/{1+ exp (–A*X^E+B)}*

with E = 0.26 (optimized)) while A = –2.41 and B = 5.01 from linear regression (Equation 1)

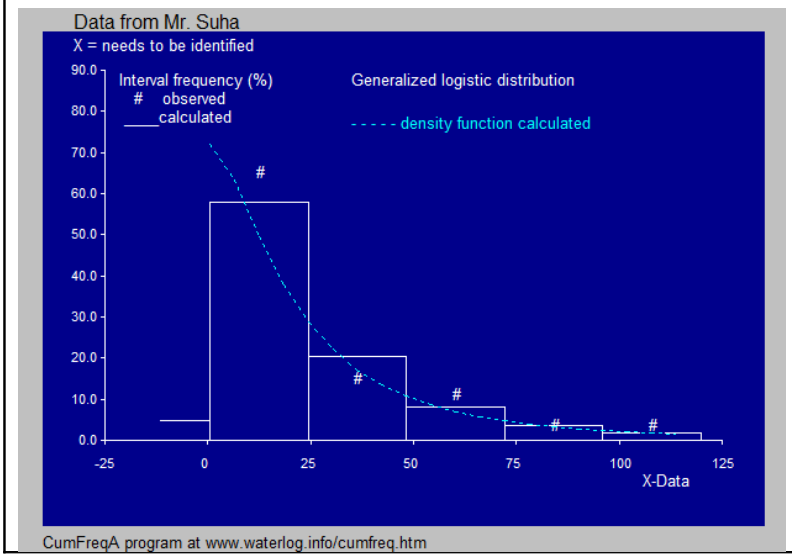


Figure 3B

Histogram and generalized logistic PDF of data that still need to be defined.

The density curve looks exponential and the distribution looks extremely skewed to the right, reason why the exponent E is very low (see figure 3A*).*

Figure 4A gives the cumulative probability distribution function (CPDF or CDF or CPF) obtained by the generalized logistic distribution for the test score of school pupils.

In *figure 4B* the corresponding histogram and probability density function (PDF) is given. The density curve is skew to the left reason why the optimized exponent E (1.65) is greater than 1.

The number of intervals for the test score histogram has been enlarged compared to the previous numbers in order to obtain round numbers from 1 to 10 on the X-axis as the test score runs from 1 to 10.

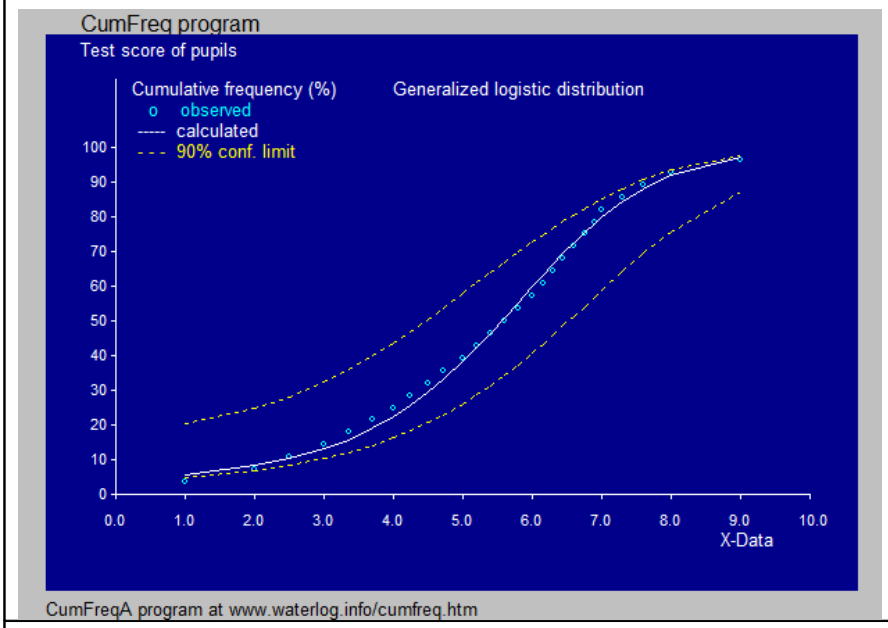


Figure 4A

Generalized logistic CPF of the test score of school children..

*Cp = 1/{1+ exp (–A*X^E+B)}*

with E = 1.65 (optimized) while A = –0.175 and B = 2.96 from linear regression (Equation 1)

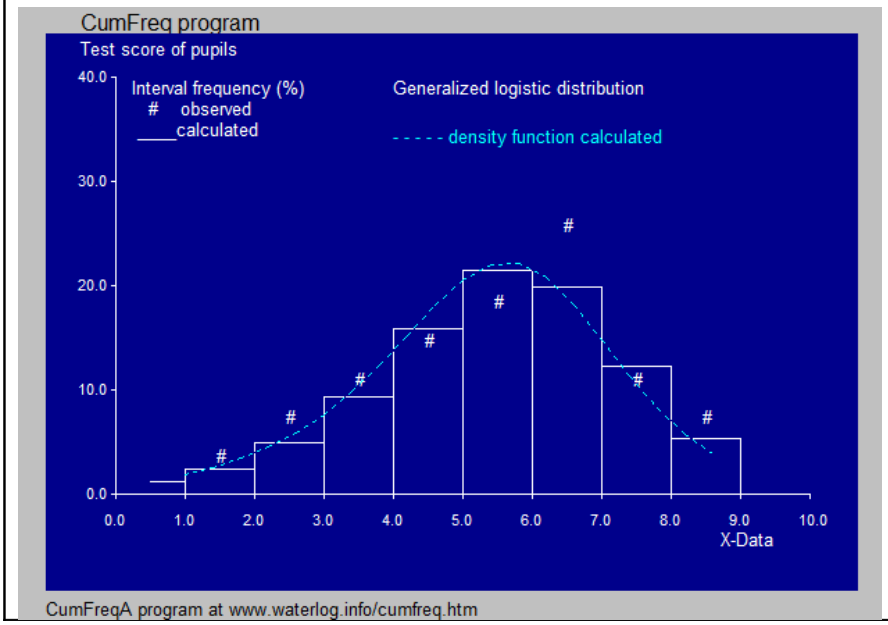


Figure 4B

Histogram and generalized logistic PDF of test scores..

The density curve looks skewed to the left, reason why the exponent E is greater than 1 (see figure 4A)

All previous distributions were generalized logistic. In the next section the generalized (mirrored) Gumbel distribution will be examined using exactly the same data sets.

5.2 The Gumbel and the mirrored Gumbel distribution

Figure 5A gives the cumulative probability distribution function (CPDF or CDF or CPF) obtained by the generalized Gumbel distribution for data on the height of shrubs.

In *figure 5B* the corresponding histogram and probability density function (PDF) is given. Note that the PDF is the derivative of the CPF.

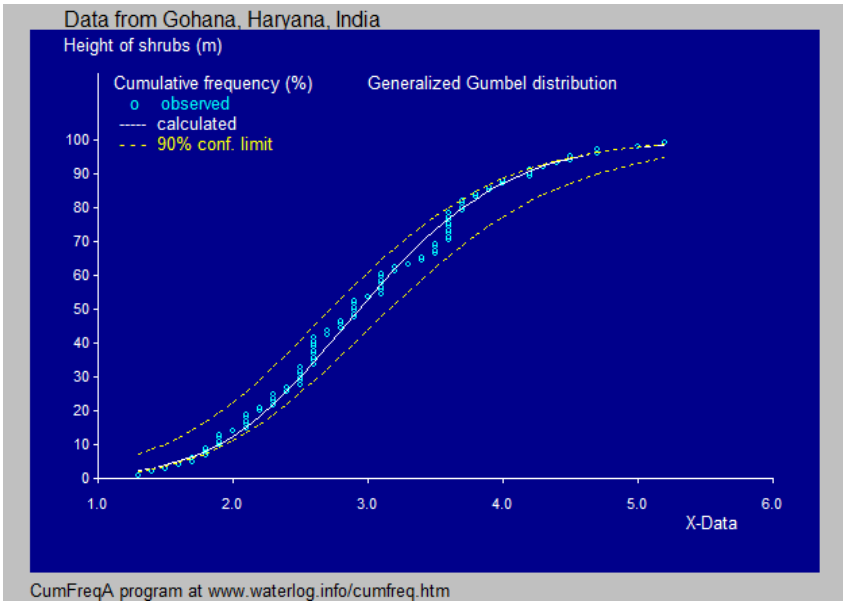


Figure 5A

Generalized Gumbel CPF of the height of shrubs..

*Cp = exp[–exp{– (A*X^E+B)}]*

with E = 1.77 (optimized) while A = 33.0 and B = –1.68 from linear regression (Equation 2))

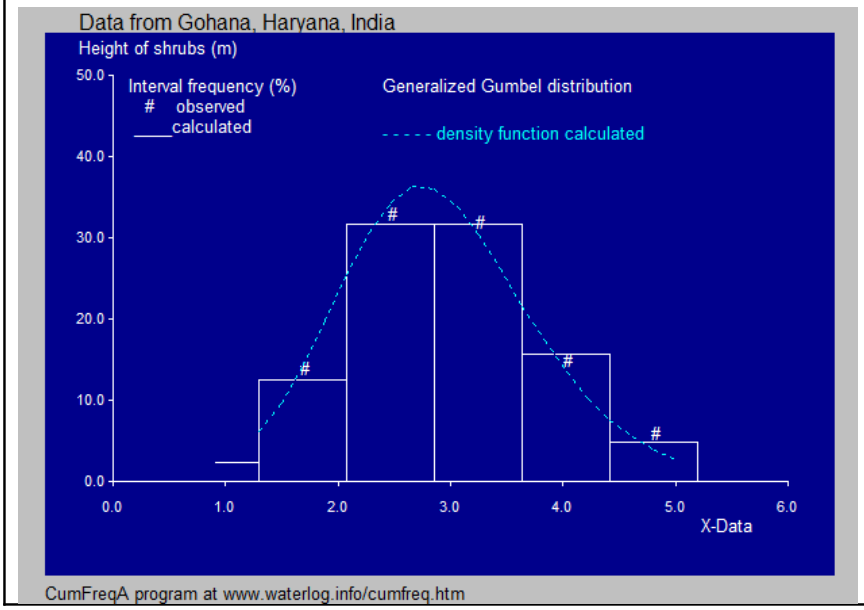


Figure 5B

Histogram and generalized logistic PDF of the height of shrubs.

Figures 5A and 5B for the Gumbel distribution look the same as figures 1A and 1B for the logistic distribution and in both cases the index for goodness of fit (0.99 or 99%) is very high (see the P-P plot in section 6: *Conclusion*). Hence both distributions can be used.

Figure 6A gives the cumulative probability distribution function (CPDF or CDF or CPF) obtained by the generalized Gumbel distribution for data on the health of cows.

In *figure 6B* the corresponding histogram and probability density function (PDF) is given. Note that the PDF is the derivative of the CPF.

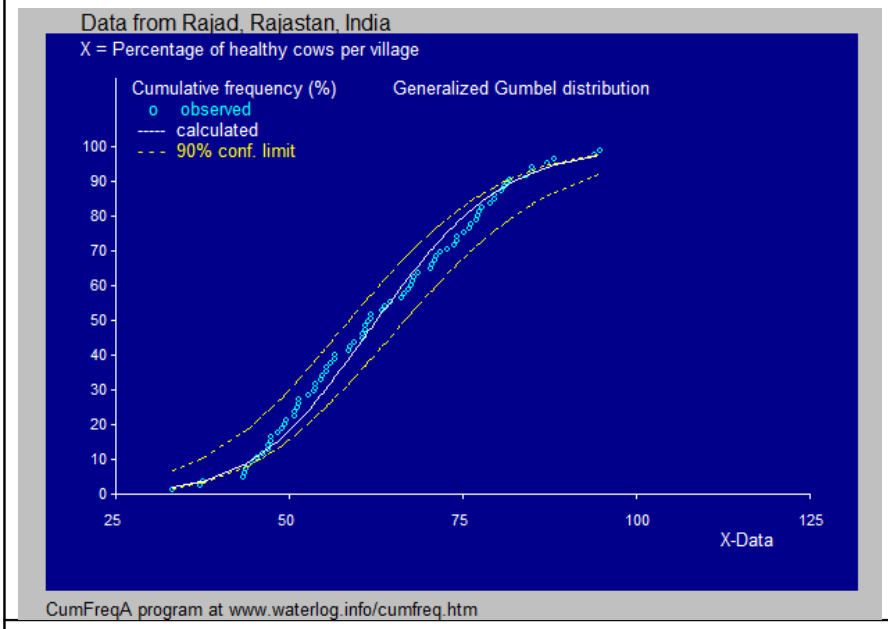


Figure 6A

Generalized Gumbel CPF of the health of cows.

*Cp = exp[–exp{– (A*X^E+B)}]*

with E = 2.13 (optimized) while A = 0.000353 and B = –2.00 from linear regression (Equation 2))

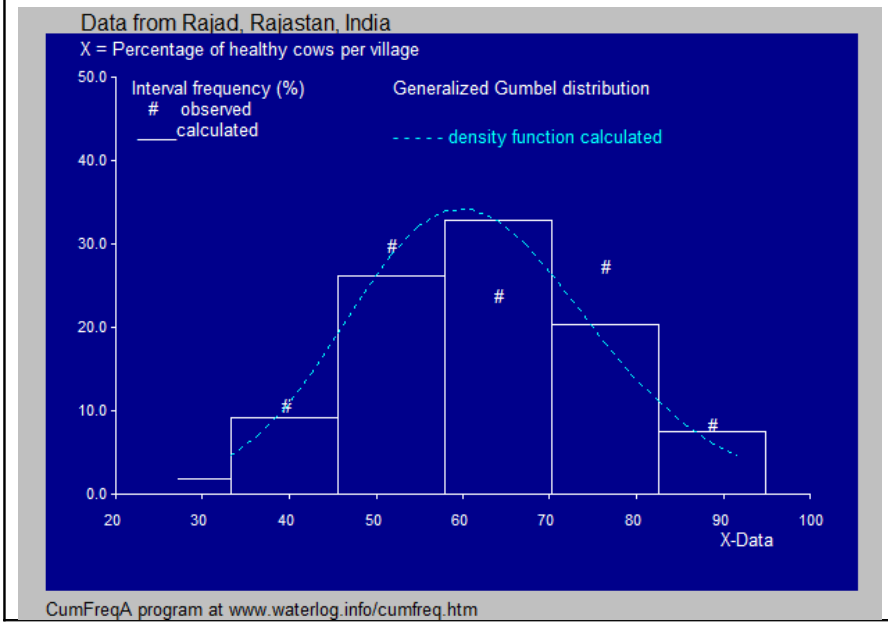


Figure 6B

Histogram and generalized logistic PDF of the health of cows

Figures 6A and 6B for the Gumbel distribution look the same as *figures 1A and 1B* for the logistic distribution and in both cases the index for goodness of fit (0.98 or 98%) is very high (see the P-P plot in *section 6: Conclusion*). Hence both distributions can be used.

Like in *figure 2B*, the relatively low value of the observed interval frequency (symbol #) between X=60 and X=70 suggests that the distribution is composite as will be explained in *figure 6C.*

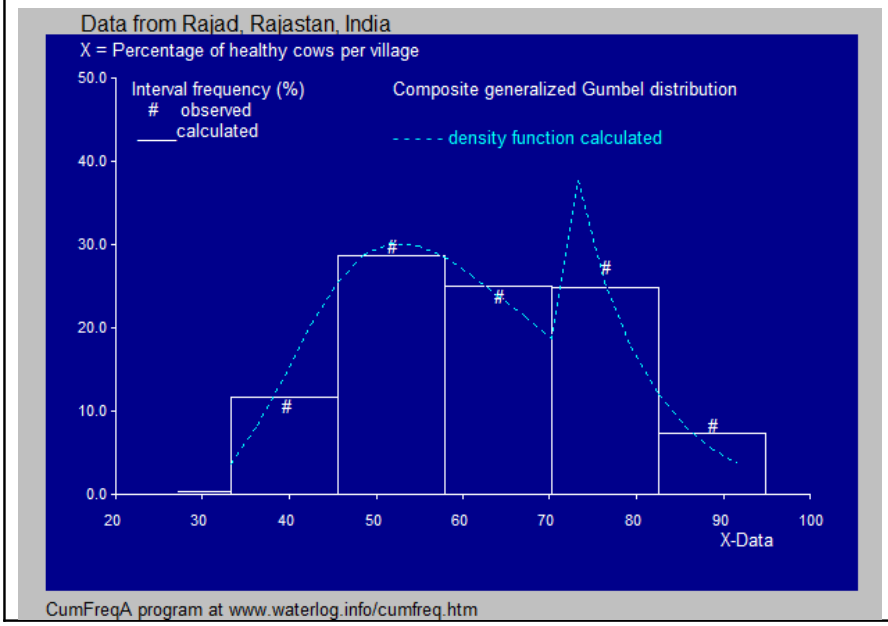


Figure 6C

Probability density function (PDF) of the composite generalized Gumbel distribution.

Explanation hereunder.

The composite distribution used in *figure 6C* is based on a separation of the data set into two parts separated by the point Xs and applying the cumulative distribution separately to the left and to the right of Xs. The generalized Gumbel equations in this case are

X < Xs : Cp = exp[–exp{– (As*X^Es+Bs)}]
with Es = 0.46 (optimized) while As = 1,27 and Bs = –8.2 from linear regression
X > Xs : Freq = exp[–exp{– (Ag*X^Eg+Bg)}]
with Eg = 0.46 (optimized) while Ag = 3.09 and Bg = –2.12 from linear regression
Xs being 7.43

Here the fit of the simulated probability density function (blue dotted curve) to the observed values (symbol #) is better than in *figure 6B*

Figure 7A gives the cumulative probability distribution function (CPDF or CDF or CPF) obtained by the generalized Gumbel distribution for data from Mr Suha that have still to be defined.

In *figure 7B* the corresponding histogram and probability density function (PDF) is given. Note that the PDF is the derivative of the CPF.

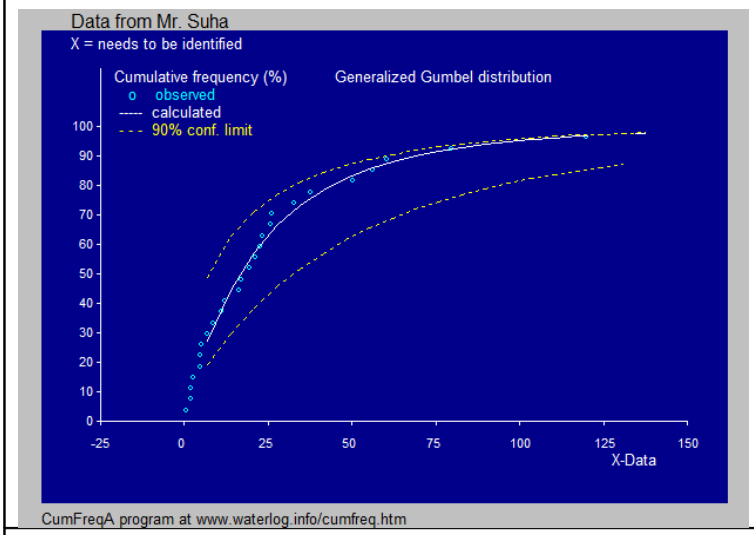


Figure 7A

Generalized Gumbel CPF of the data that need to be defined.

*Cp = exp[–exp{– (A*X^E+B)}]*

with E = 0.51 (optimized) while A = 0.419 and B = –1.38 from linear regression (Equation 2))

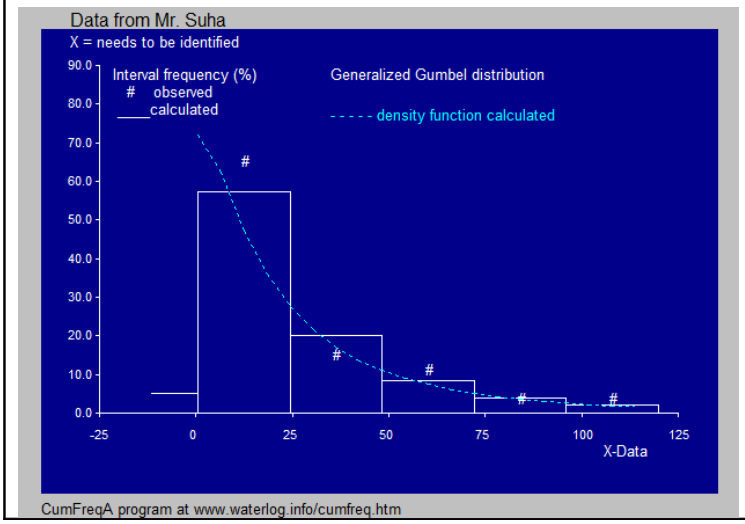


Figure 7B

Histogram and generalized logistic PDF of data that still need to be defined.

The density curve looks exponential and the distribution looks extremely skewed to the right, reason why the exponent E is less than 1 (see figure 7A*).*

Figures 7A and 7B for the Gumbel distribution look the same as *figures 3A and 3B* for the logistic distribution and in both cases the index for goodness of fit (0.99 or 99%) is very high (see the P-P plot in *section 6: Conclusion*). Hence both distributions can be used.

As the density function in figure 4B is skewed to the left, for the case of the test score of pupils *Figure 8A* gives the cumulative probability distribution function (CPDF or CDF or CPF) obtained by the generalized mirrored Gumbel distribution for the test score of pupils.

As the density function in figure 4B for the case of the test score of pupils is skewed to the left, here the *mirrored* Gumbel distribution is used instead of the standard one.

As the standard Gumbel cumulative probability function is Cp = exp[–exp{– (A*X+B)}] (see section 3.2), the *mirrored* Gumbel cumulative probability function is

Cp = **1** – exp[–exp{– (A*X+B)}]
In *figure 8B* the corresponding histogram and probability density function (PDF) is given. Note that the PDF is the derivative of the CPF.

Like in *figure 4B* the number of intervals for the test score histogram has been enlarged compared to the previous numbers in order to obtain round numbers from 1 to 10 on the X-axis as the test score runs from 1 to 10.

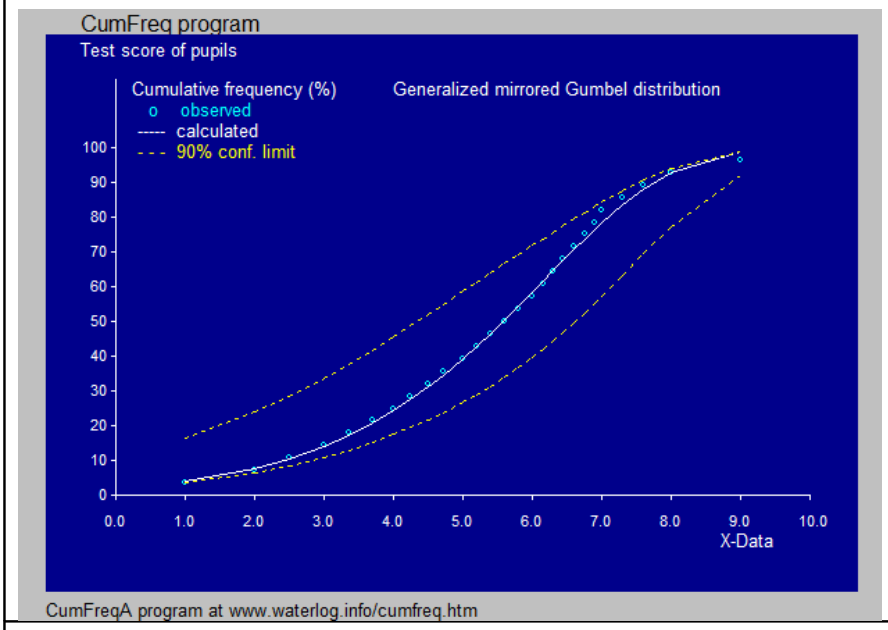


Figure 8A

Generalized mirrored *Gumbel CPF of the test score of pupils.*

*Cp = **1** – exp[–exp{– (A*X^E+B)}]*

with E = 0.870 (optimized) while A = –81.5 and B = 4.00 from linear regression (Equation 2))

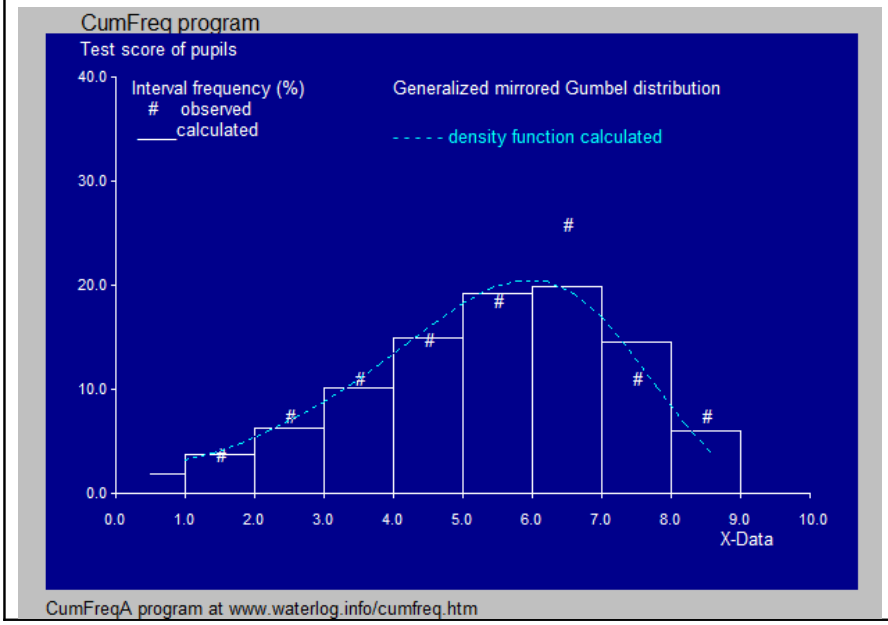


Figure 8B

Histogram and generalized logistic PDF of data that still need to be defined.

Figures 8A and 8B for the *mirrored* Gumbel distribution look the same as *figures 4A and 4B* for the logistic distribution and in both cases the index for goodness of fit (0.99 or 99%) is very high (see the P-P plot in section 6: *Conclusion*). Hence both distributions can be used.

6. Conclusion

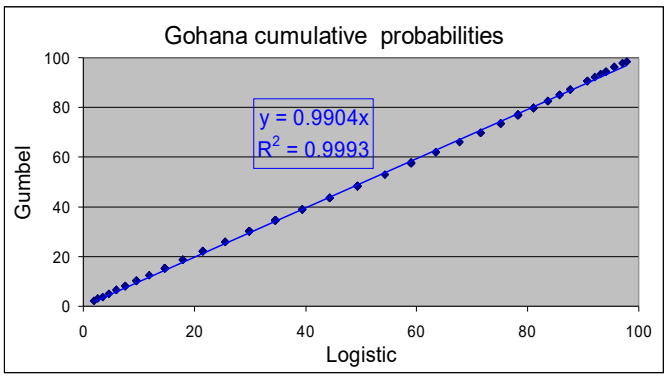
From the previous examples it becomes clear the generalized logistic and the generalized (mirrored) Gumbel probability distribution functions derived with the method of plotting positions have a wide applicability.

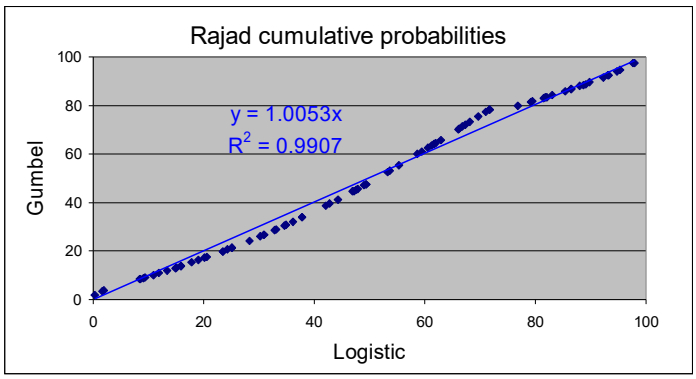
The plotting positions are helpful in simplifying the way in which the parameters of the cumulative distribution functions are determined, yet they are very effective. Often the parameters can be found by linear regression of transformed distribution functions constructed with the help of the method of plotting positions.

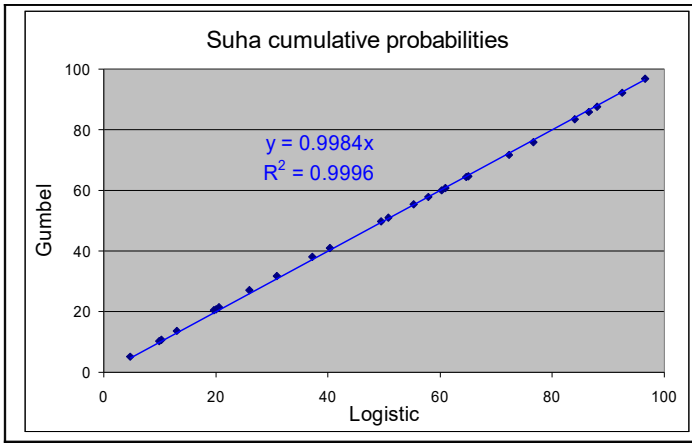
From the cumulative distribution functions thus formulated it is possible to derive the probability density distributions as they are the derivative of the cumulative ones.

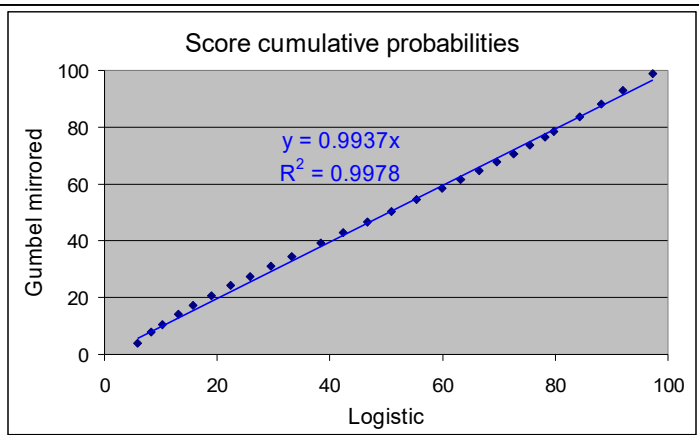
The method of plotting positions also makes it possible to construct the confidence belts shown in the above examples, see the *Appendix*.

The correspondence between the cumulative probabilities of the generalized logistic and of the generalized Gumbel distributions is demonstrated in the next figures of P-P plots which speak for themselves.









7. References

Reference 1.
Kimiko O. Bowman and L. R. Shenton, "Estimator: Method of Moments", pp 2092–2098, *Encyclopedia of statistical sciences*, Wiley (1998).

Reference 2.
Hosking, J.R.M. (1990). *"*L-moments: analysis and estimation of distributions using linear combinations of order statistics*". Journal of the Royal Statistical Society, Series B.* ***52*** *(1): 105–124.*

Reference 3.
Pyke, Ronald (1965). *"*Spacings*". Journal of the Royal Statistical Society, Series B.* ***27*** *(3): 395–449.*

Reference 4.
Kane, Edward J. (1968). *"*Economic Statistics and Econometrics*". New York, NY: Harper & Row. p. 179.*

Reference 5.
Gumbel, E.J. (1954). *"*Statistical theory of extreme values and some practical applications*". Applied Mathematics Series. Vol. 33 (1st ed.). U.S. Department of Commerce, National Bureau of Standards.*

Reference 6.
Lasse Makkonen, 2006. "Plotting Positions in Extreme Value Analysis*". Journal of Applied Meteorology and Climatology Vol. 45*. On line: https://journals.ametsoc.org/doi/10.1175/JAM2349.1

Reference 7.
R.J. Oosterbaan, 2020. "Software for generalized and composite probability distributions" *International Journal of Mathematical and Computational Methods,* **4**, 1-9
On line: https://www.waterlog.info/pdf/MathJournal.pdf

Reference 8
R.J.Oosterbaan 2021. "Fitting the versatile linearized, composite, and generalized logistic probability distribution to a data set". On line: https://www.waterlog.info/pdf/logistic.pdf

Reference 9
R.J.Oosterbaan 2022. "The generalized standard and mirrored Gumbel probability distributions, composite or not, are applicable to many datasets, either symmetrical, skew to the left, or skew to the right". On line:
https://www.waterlog.info/pdf/Composite distribution.pdf

8. Appendix (confidence belts)

In a number of figures with the cumulative distribution depicted, their 90% confidence belts have been drawn. The confidence intervals are found from the (relative) standard deviation (Sd) of the binomial probability distribution [**Ref. A**]:

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

where Pc is the cumulative (non-exceedance) probability ($0<Pc<1$), and N is the number of data.

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

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

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

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

[**Ref. A**] Use of the binomial probability distribution for confidence intervals of cumulative probability distribution functions. On line: https://www.waterlog.info/pdf/binoom.pdf