Using simple transformations of probability distributions to determine their parameters by straightforward linear regression and to be able to fit symmetrical and skewed data sets easily.

R.J. Oosterbaan, December 2021. www.waterlog.info/cumfreq.htm

Abstract
A number of commonly used probability distributions can be applied to data sets in a simple manner by using obvious transformations so that linear functions are obtained from which the distribution parameters can be determined using the standard linear regression method. A condition is that the cumulative frequency (frequency of non-exceeding, or 1 minus the frequency of exceeding) of the data is calculated from the plotting position, which requires ranking of the data in increasing order from the lowest to the highest value.
In this paper, the transformation method is explained by type of distribution and various practical examples are given together with an uncomplicated technique to find an index for goodness of fit employing free probability distribution software.

Content

1. Introduction

1.1 General definitions

Cumulative probability distributions with two parameters may be briefly written as:

$Pc\ (X<Y) = Fc\ (A,\ B,\ Y)$

where:

Pc = probability the X value under consideration to be less than assumed the value Y;
Pc is also called the probability of non-exceeding of X compared to Y
A and B are the distribution's parameters
Fc = cumulative probability distribution function (often abbreviated as CDF) of X with respect to Y based on the values of the parameters A and B

There exist quite a number of probability distributions with two parameters, as will be shown in section 1.3.

The probability of exceeding will be:

Pe (X>Y) = 1 – Fc (A, B, Y)

so that Pc + Pe = 1.

Pe may also be called the mirrored (or inverted) distribution of Pc and vice versa.

1. 2 Distribution skewness

Probability distributions (*figure 1*) may be:
- symmetrical (for example the well known normal distribution)
- skew to the right (positively skewed, there is a prolonged tail to the right)
- skew to the left (negatively skewed, the tail to the left is dominating)

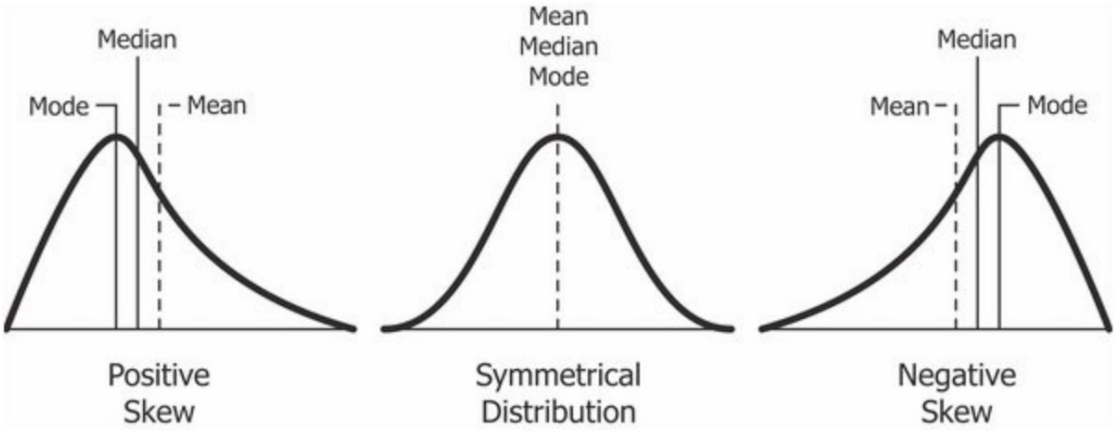

Figure 1. Illustrating skewness an symmetry of probability distributions

In *figure 1* it can be seen that, in case of positive skewness, the extreme values on the right (the higher extremes) are farther away from the mode (the top of the distribution) than the extreme values on the left (the lower extremes). For negative skewness, the reverse is true.

The figure also shows that:
- for symmetrical distributions the mean, the mode and the median values coincide
- for positively skewed distributions we have mean>median>mode
- for positively skewed distributions we have mean<median<mode

When a probability distribution is known to be skewed positively, for example the standard Gumbel distribution, see *section 1.3*, then its mirrored (or inverted) distribution, in this example the mirrored (inverted) Gumbel distribution, is negatively skewed and vice versa.

Hence, the use of the mirrored distribution of frequently used probability distributions enlarges the possibilities of distribution fitting.

1.3 Distributions with 2 parameters that can be found from a linear regression of transformed data (Examples, summary)

For transformations and linearization, see *section 2*.

Definitions:

- a Exp (Z) = e^Z, e being approximately equal to 2.718 . . .
  Found from: $f(x) = e^x \Rightarrow d\{f(x)\} / dx = e^x$
- b pi (greek π): ratio of circumference of a circle to circle diameter ≈ 3.142 . . .
- c ^ stands for: raised to the power by the number following the ^ symbol
- d * stands for multiplication
- e X is data value

The relevant probability distributions are:

1a. Logistic distribution (symmetrical): Fc = 1/[1+Exp(A*X+B)]
1b. Log-logistic distribution (skew to right): Fc = 1/[1+Exp{A*log(X)+B}]

2. Cauchy distribution (symmetrical): Fc = (1/pi)*arctan(A*X+B) + 0.5

3a. Negative exponential distribution (Poisson type, skew to right): Fc = 1 – Exp{– (A*X+B)}
3b. Mirrored (inverted) exponential distribution (skew to left): Fc = Exp{– (A*X+B)}

4a. Gumbel (Fisher-Tippett type I) distribution (skew to right): Fc = Exp[-Exp{– (A*X+B)}]
4b. Mirrored Gumbel distribution (skew to left): Fc = 1 – Exp[-Exp{– (AX+B)}]

5a.. Weibull distribution (skew to right): Fc = 1 - Exp{– (X/D)^A} with D = Exp(–B/A)
5b. Mirrored Weibull distribution (skew to left): Fc = Exp{– (X/D)^A} with D = Exp(–B/A)

1.4 Generalization

The probability distributions can be generalized and made more versatile using Z = X raised to the power E ($Z=X^E$) instead of X and numerically optimizing the E value. This procedure is outside the scope of this paper, but it can be checked in *Reference 1*.

2. Transformation and linearization of distributions with two parameters

The Fc values are found from the plotting positions of the X values in a data set arranged in increasing order (lowest value first, higest value last) as follows:

Fc = R / (N + 1), where R is the rank number of the ordered data, and N is the total number of data present (*Reference 2*). Fc is also called: "plotting position".

The distributions summarized in *section 1.3* can now be transformed an linearized LN = natural logarithm (logarithm on the basis of the number e ≈ 2.718 . . .).

1a. Logistic distribution (symmetrical)
Fc = 1 / [1+Exp(A*X+B)]
Transformation: Ft = Ln(–1+1/Fc)
Hence: Ft = A*X + B (see example in *Section 4.1*)
Parameters A and B are found from a linear regression of Ft on X

1b. Log-logistic distribution (skew to right):
Fc = 1 / [1+Exp{A*Ln(X)+B}]
Transformation: Ft = Ln(–1+1/Fc)
Hence: Ft = A*Ln(X)+ B
Parameters A and B are found from a linear regression of Ft on Ln(X)

2. Cauchy distribution (symmetrical)
Fc = (1/pi)*arctan(A*X+B) + 0.5
Transformation: Ft = tan{pi*(Fc–0.5)}
Hence: Ft = A*X + B
A and B are found from a linear regression of Ft on X

3a. Generalized (negative) exponential distribution
(Poisson-type, skew to right)
Fc = 1 – Exp{– (A*X+B)}
1st Transformation: Xt = Ln(X)
2nd Transformation: Ft = – Ln (1-Fc)
Hence: Ft = A*Xt + B
A and B are found from a linear regression of Ft on Xt

3b. Mirrored exponential distribution generalized
(Skew to left)
Fc = Exp{– (A*X^E+B)}
1st Transformation: Xt = Ln(X^E) = E*Ln(X)
2nd Transformation: Ft = – Ln (Fc)
Hence: Ft = A*Xt + B
A and B are found from a linear regression of Ft on Xt

4a. Gumbel (Fisher-Tippett type I) distribution
(skew to right)
Fc = Exp [– Exp { – (A*X+B) }]
Transformation: Ft = – Ln{ – Ln (Fc) }
Hence: Ft = A*X + B
A and B are found from a linear regression of Ft on X

4b. Mirrored Gumbel distribution (skew to left)
Fc = 1 – Exp [– Exp { – (A*X+B) }]
Transformation: Ft = Ln{–Ln(1–Fc)}
Hence: Ft = A*X + B (see example 4.3 in *Section 4*)
A and B are found from a linear regression of Ft on X

5a. Weibull distribution (skew to right)
Fc = 1 – Exp { – (X / C) ^A }
with C = Exp (– B / A)
1st Transformation: Xt = Ln (X)
2nd Transformation: Ft = Ln { – Ln (1 – Fc) }
Hence: Ft = A*Xt + B (see example 4.2 in *Section 4*)
A and B are found from a linear regression of Ft on Xt

5b. Mirrored Weibull distribution (skew to left)
Fc = Exp {– (X / C) ^A}
with C = Exp (– B / A)
1st Transformation: Xt = Ln (X)
2nd Transformation: Ft = Ln { – Ln (Fc) }
Hence: Ft = A*Xt + B
A and B are found from a linear regression of Ft on Xt

3. Use of free software CumFreq

The free software for probability distribution fitting CumFreq (*Reference 3*) offers the possibility to select a preferred probability distribution (*Figure 2, green rectangle*).

When clicking on the preference button the program presents a list of probability distribution from which a selection can be made (*Figure 3*)

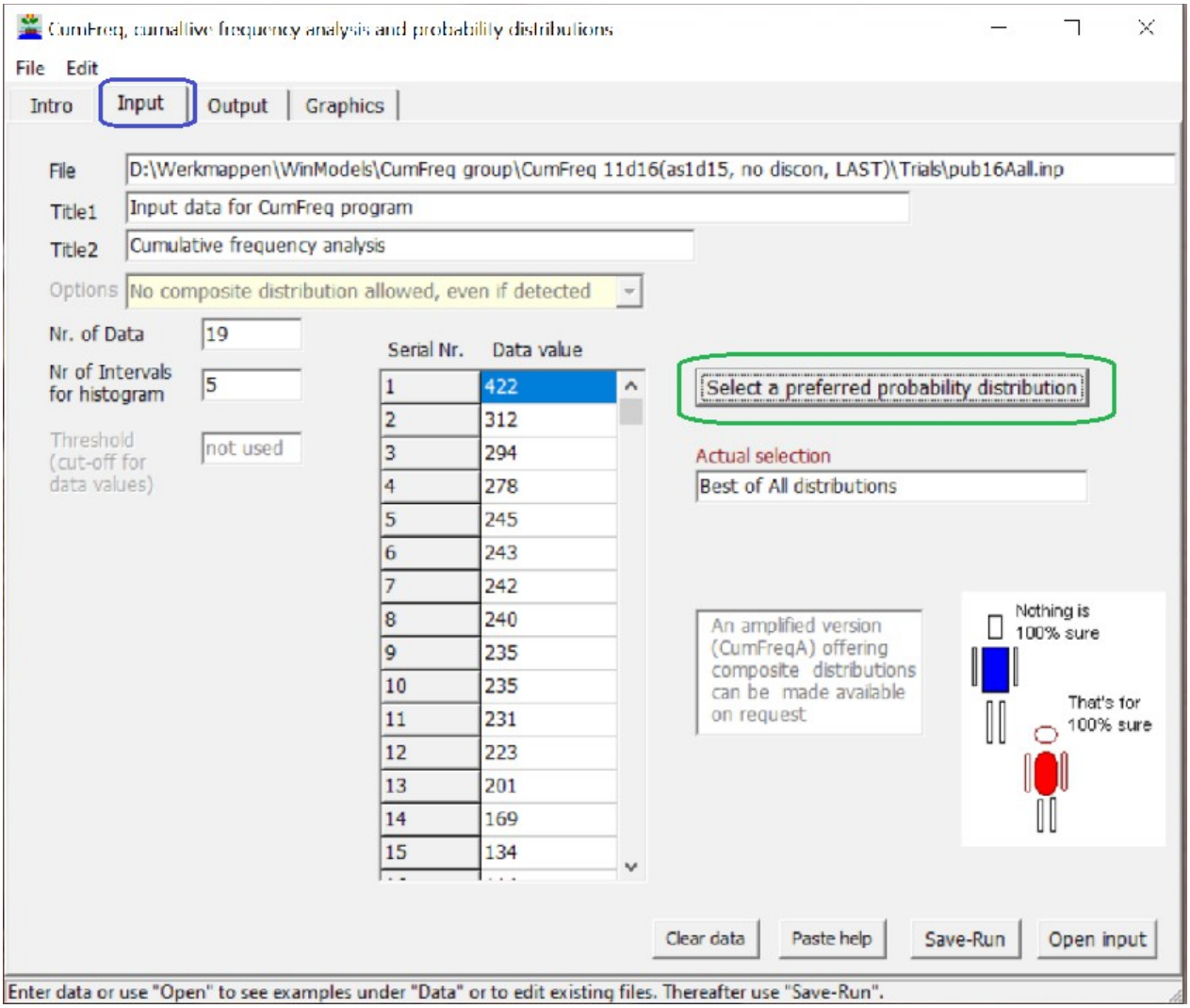

Figure 2, CunFreq input user interface (blue rectangle) giving the option to select a preferred probability distribution (green rectangle). Upon clicking this button, Figure 3 is presented.

Figure 3. Selection options in CumFreq of a probability distribution. The distributions discussed in this paper are indicated with green rectangles.

4. Examples

4.1 Logistic distribution, symmetrical

Figure 4 gives the Cumfreq cumulative probability result (or cumulative distribution function, CDF) for a symmetrical data set.

By linear regression is found that the factor A equals – 0.618 and the term B becomes 4.32, *Figure 5*. (See also the equations under distribution **1.a** in *Section 2*).

Figure 6 depicts the symmetry in the probability density function (PDF).

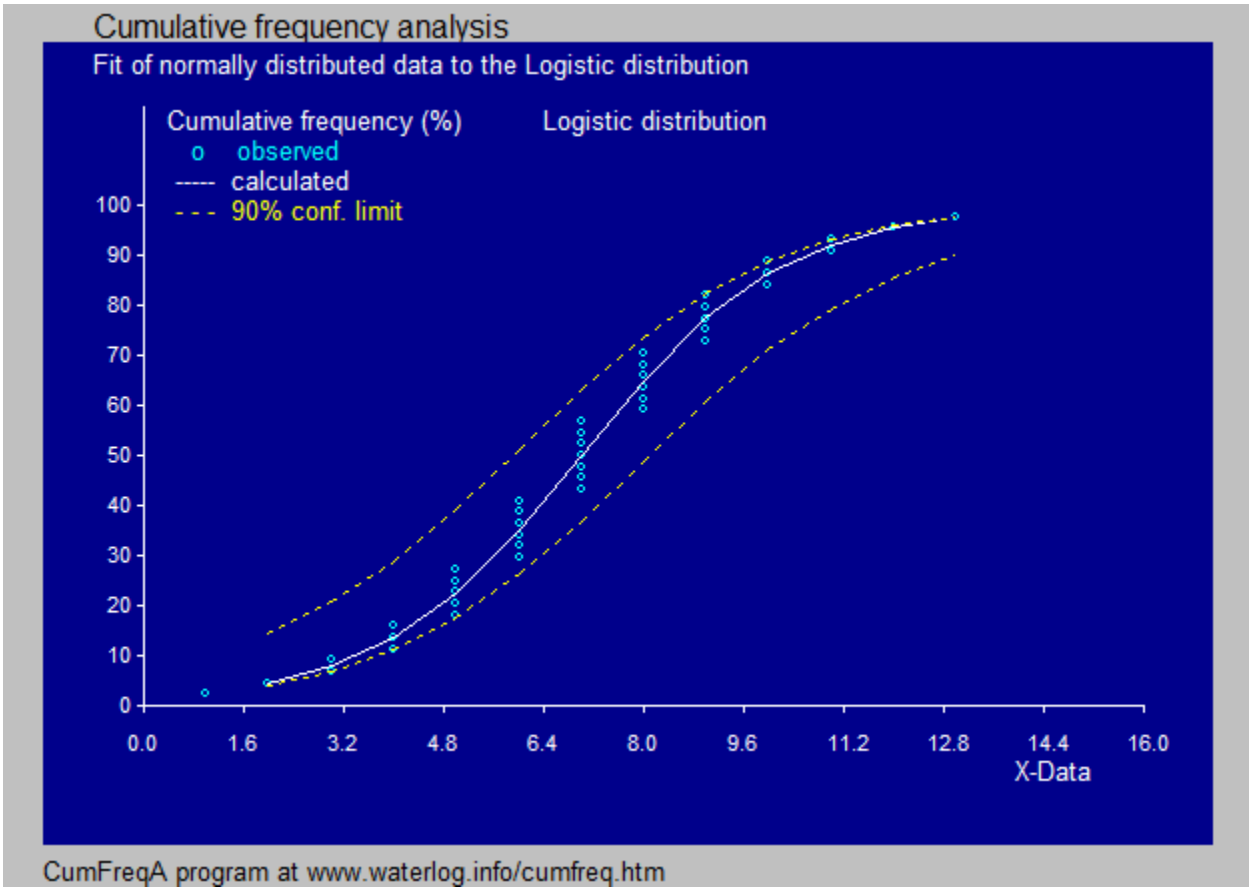

Figure 4. Cumfreq cumulative probability result (or cumulative distribution function, CDF) for a symmetrical data set.

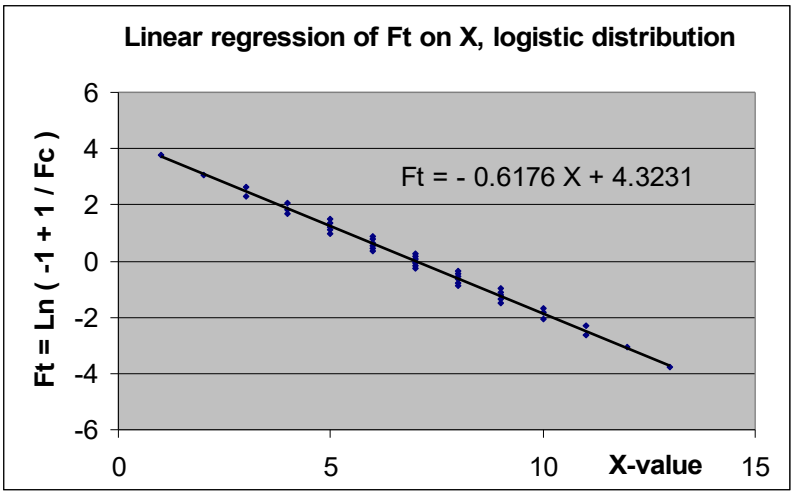

Figure 5. Linear regression of Ft on X, logistic distribution.
According to distribution **1.a** *in* Section 2 we have *Ft = Ln(–1+1/Fc)*

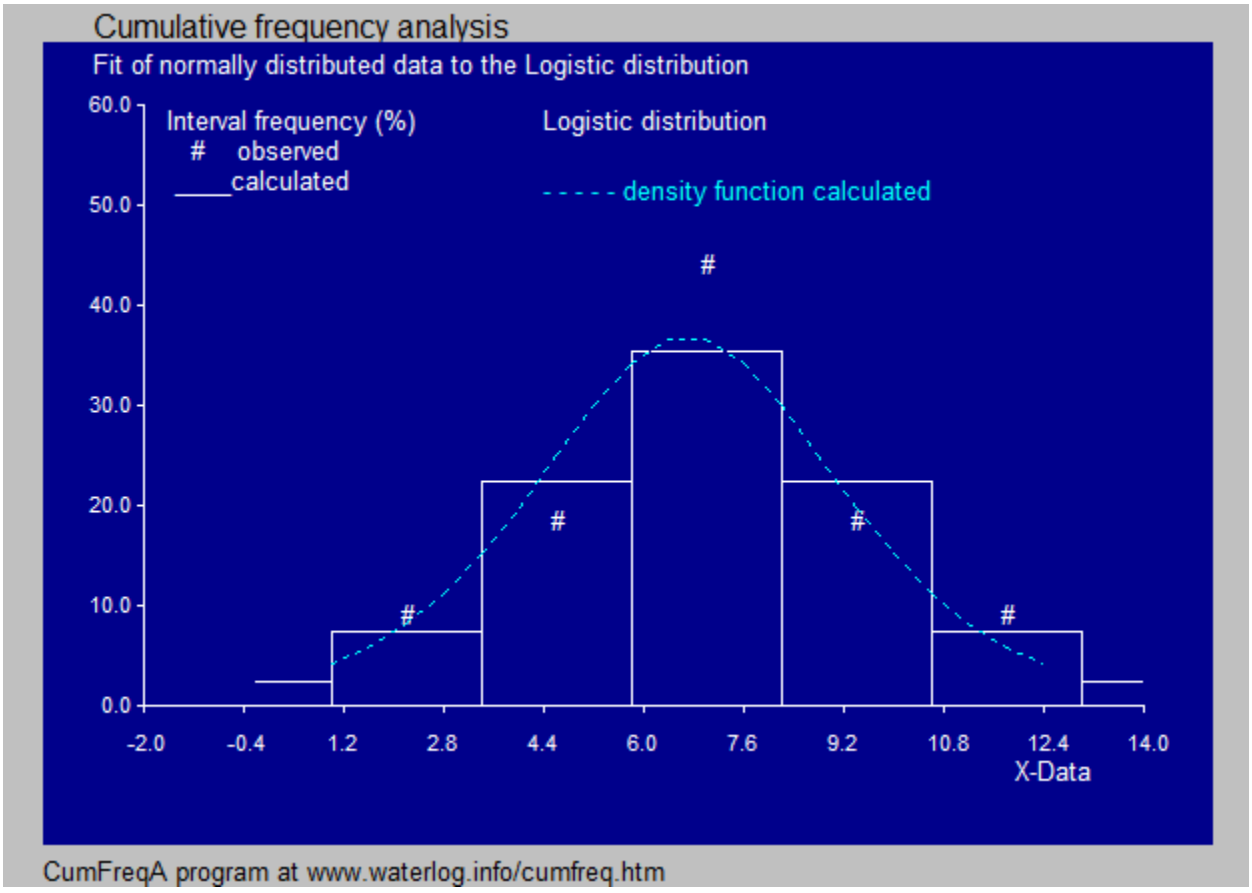

Figure 6. Depicting the symmetry in the probability density function (PDF) using the same data as in Figure 4.

Note. For more information on the logistic distribution see *Reference 4*.

4.2 Weibull distribution, skew to right

Figure 7 gives the Cumfreq cumulative probability result (or cumulative distribution function, CDF) for a positively skewed data set.

By linear regression is found that the factor A equals 1.56 and the term B becomes -6.80, *Figure 8*. (See also the equations under distribution **5.a** in *Section 2*).

Figure 9 depicts the positively skewed probability density function (PDF).

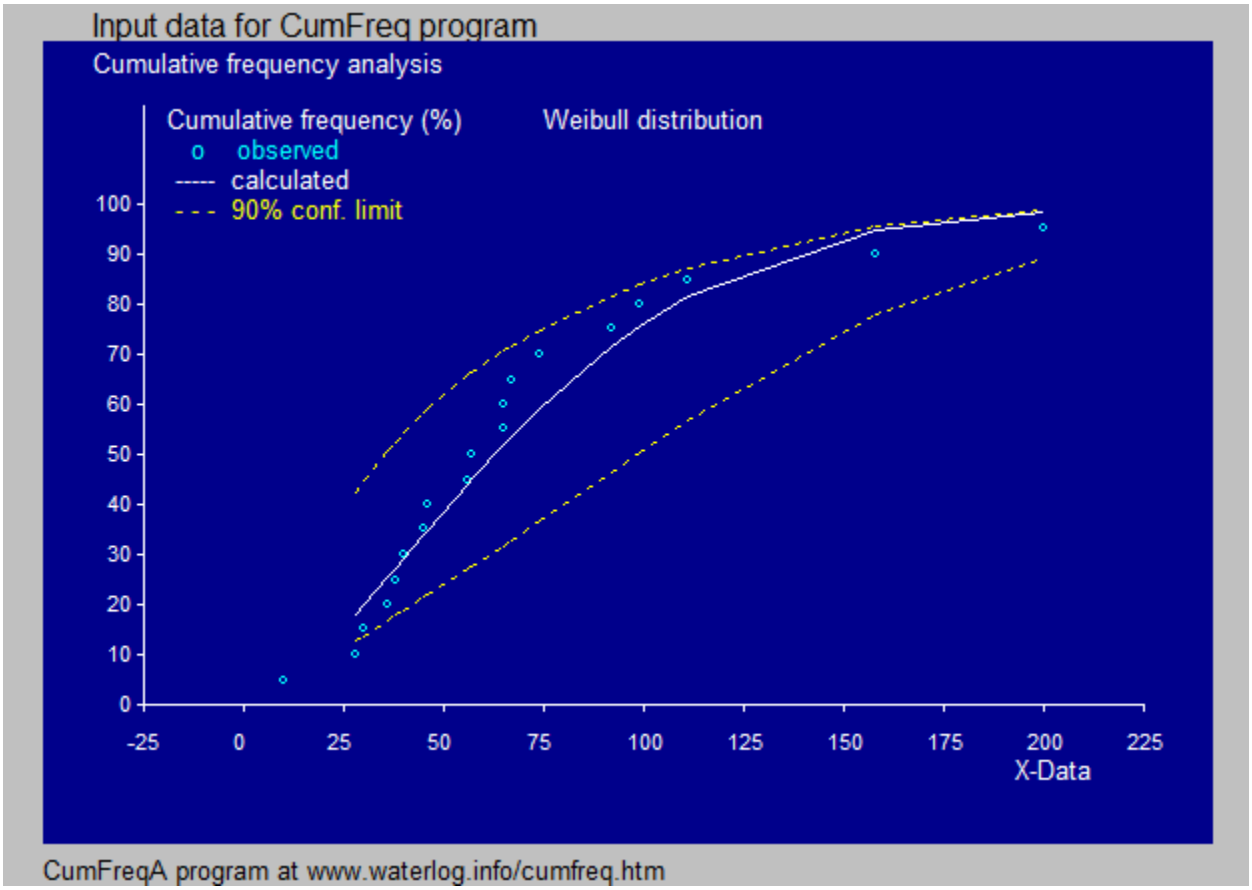

Figure 7. Cumfreq cumulative probability result (or cumulative distribution function, CDF) for a positively skewed data set.

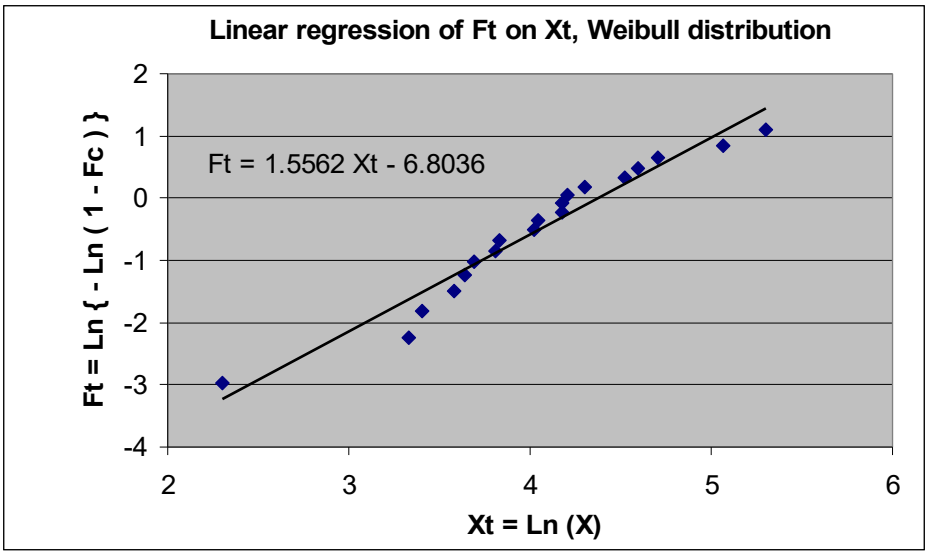

Figure 8. Linear regression of Ft on X, Weibulll distribution.
According to distribution **5.a** *in* Section 2 *we have Xt = ln(X) and Ft = Ln{–Ln(1–Fc)}. It is seen that A = 1.556 and B = – 6.804*
Hence C = Exp (– B/A) = Exp (6.804/1.556) = Exp (4.373) = $e^{4.373} = 79.28$

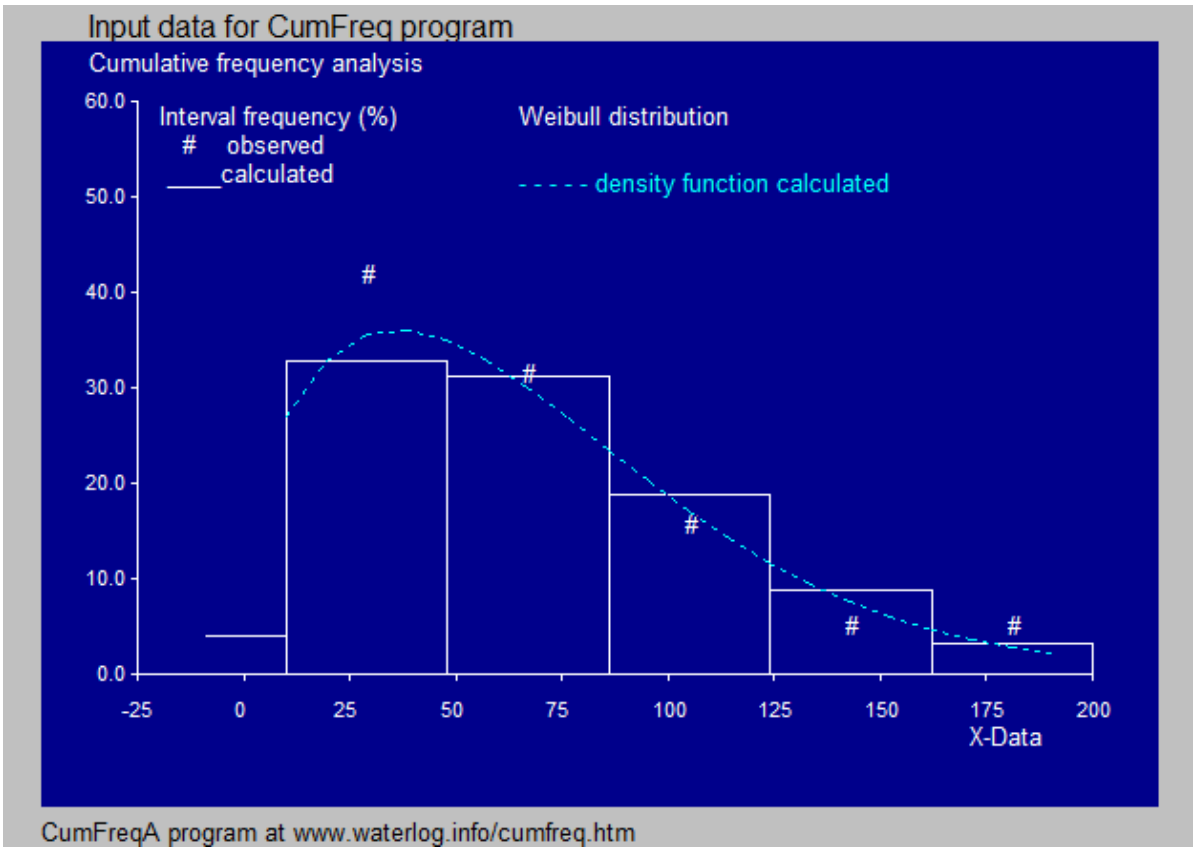

Figure 9. Depicting the positive skewness in the probability density function (PDF) using the same data as in Figure 7.

4.3 Mirrored Gumbel distribution, skew to left

Figure 10 gives the Cumfreq cumulative probability result (or cumulative distribution function, CDF) for a negatively skewed data set.

By linear regression is found that the factor A equals – 0.582 and the term B becomes 3.63, *Figure 11*. (See also the equations under distribution **4.b** in *Section 2*).

Figure 12. depicts the negativey skewed probability density function (PDF).

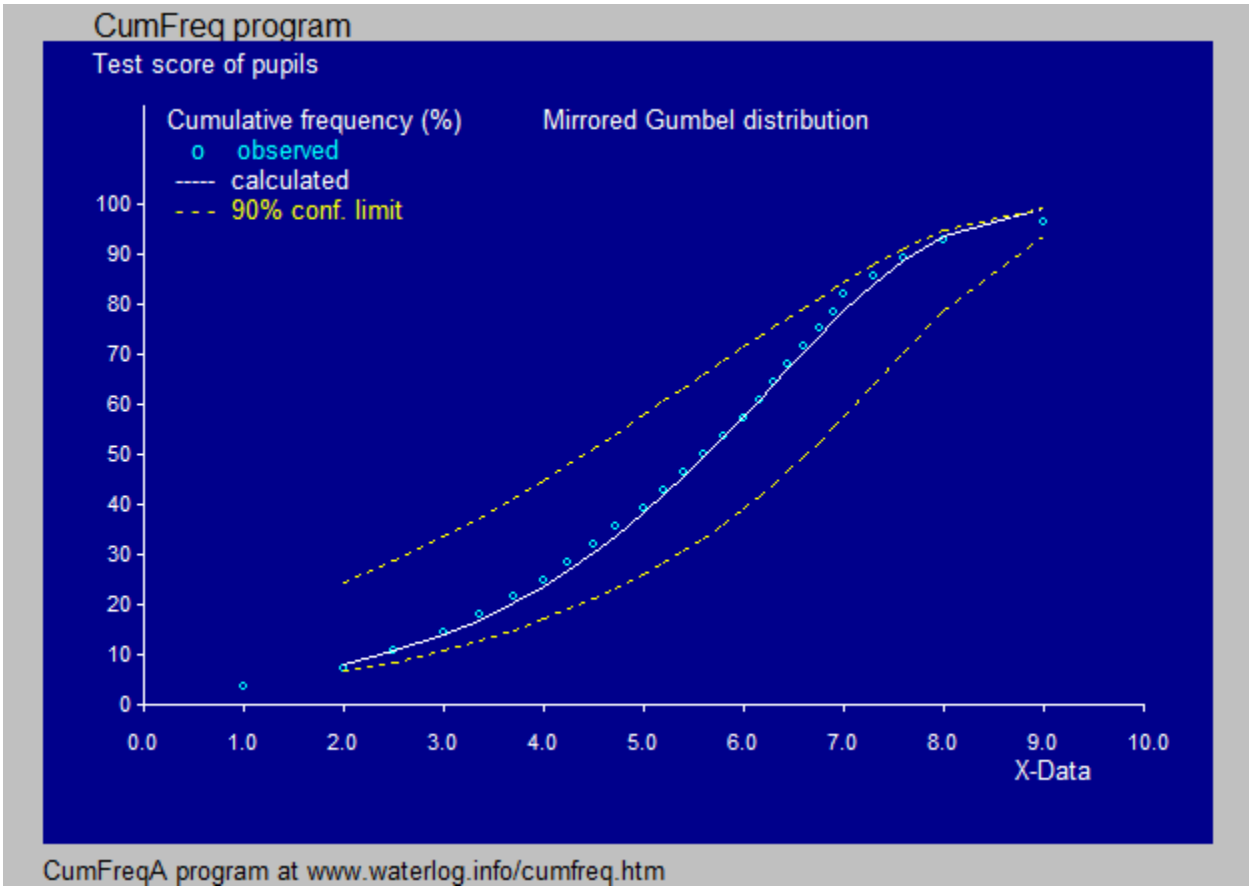

Figure 10. Cumfreq cumulative probability result (or cumulative distribution function, CDF) for a negatively skewed data set.

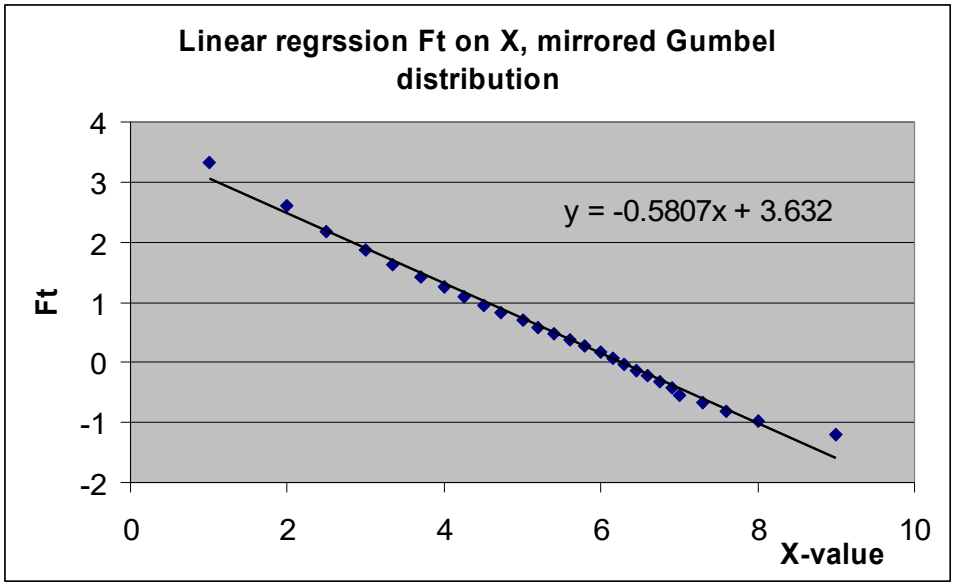

Figure 11. Linear regression of Ft on X, Mirrored Gumbel distribution.
According to distribution **4.b** *in* Section 2 *we have* Ft = –Ln{–Ln(1–Fc)}

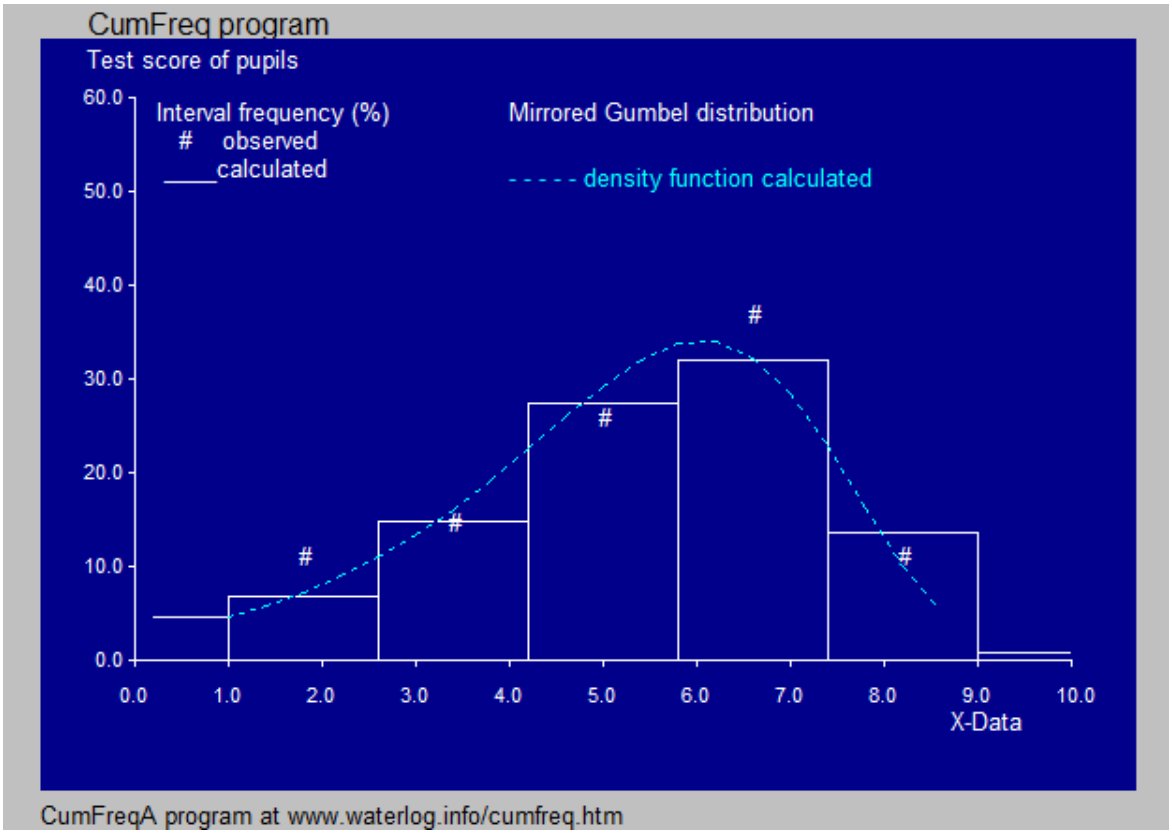

Figure 12. Depicting the negative skewness in the probability density function (PDF) using the same data as in Figure10.

Note. For more information on left skewed probability distributions see *Reference 5*.

5. References

Reference 1.
International Journal of Mathematical and Computational Methods, 4, 1-9.
SOFTWARE FOR GENERALIZED AND COMPOSITE PROBABILITY DISTRIBUTIONS
or:
https://www.waterlog.info/pdf/MathJournal.pdf

Reference 2.
Lasse Makkonen, 2008. Bringing Closure to the Plotting Position Controversy. In:
Communication in Statistics- Theory and Methods 37(3):460-467

Reference 3.
Free software for cumulative frequency analysis (CumFreq) and probability distribution fitting.
On line: https://www.waterlog.info/cumfreq.htm

Reference 4.
FITTING THE VERSATILE LINEARIZED, COMPOSITE, AND GENERALIZED LOGISTIC PROBABILITY DISTRIBUTION TO A DATA SET
or:
https://www.waterlog.info/pdf/logistic.pdf

Reference 5.
Left (negatively) skewed frequency histograms can be fitted to square Normal or mirrored Gumbel probability functions
or:
https://www.waterlog.info/pdf/LeftSkew.pdf