Left (negatively) skewed frequency histograms can be fitted to square Normal or mirrored Gumbel probability functions.

R.J. Oosterbaan, 16-01-2020. On www.waterlog.info public domain

Abstract
Histograms with a left-skewed (negative skewness) frequency distribution can be fitted to square-normal or mirrored Gumbel probability functions. Such histograms reveal a long left tail and a short right tail.

The square-normal distribution is obtained by raising the data to the power two and applying the well know normal distribution to these transformed data. After the procedure of distribution fitting and subsequently restoring the data to their original value one will notice that the contraction of the normal distribution at the right side is stronger than at the let side which result in a left-skewed distribution.

An alternative application of the normal distribution is by raising the data to the power E instead of 2, and finding the optimal value of the exponent E by iteration and selecting that value that produces the highest goodness of fit. For distributions skewed to the left one will find a value of E greater than one, while for distributions skewed to the right the value of E will be less than one. When E equals one, the distribution is symmetrical. For distributions skewed to the right one might also try the log-normal distribution or even the root-normal distribution.

The inverted Gumbel distribution is found by subtracting the standard Gumbel Cumulative Distribution Functon from 1 obtaining the mirrored Gumbel distribution. As the standard Gumbel distribution is skewed to the right, the mirrored distribution will be skewed to the left.

In this article various practical examples will be given using the software program CumFreq for probability distribution fitting.

Contents

1. Introduction

The negatively skewed probability distribution has been described in Reference 1 for the age of death of Australian males [Ref. 1, Fig. 1] and in Reference 2 for the test scores of school children [Ref. 2, Fig. 2].

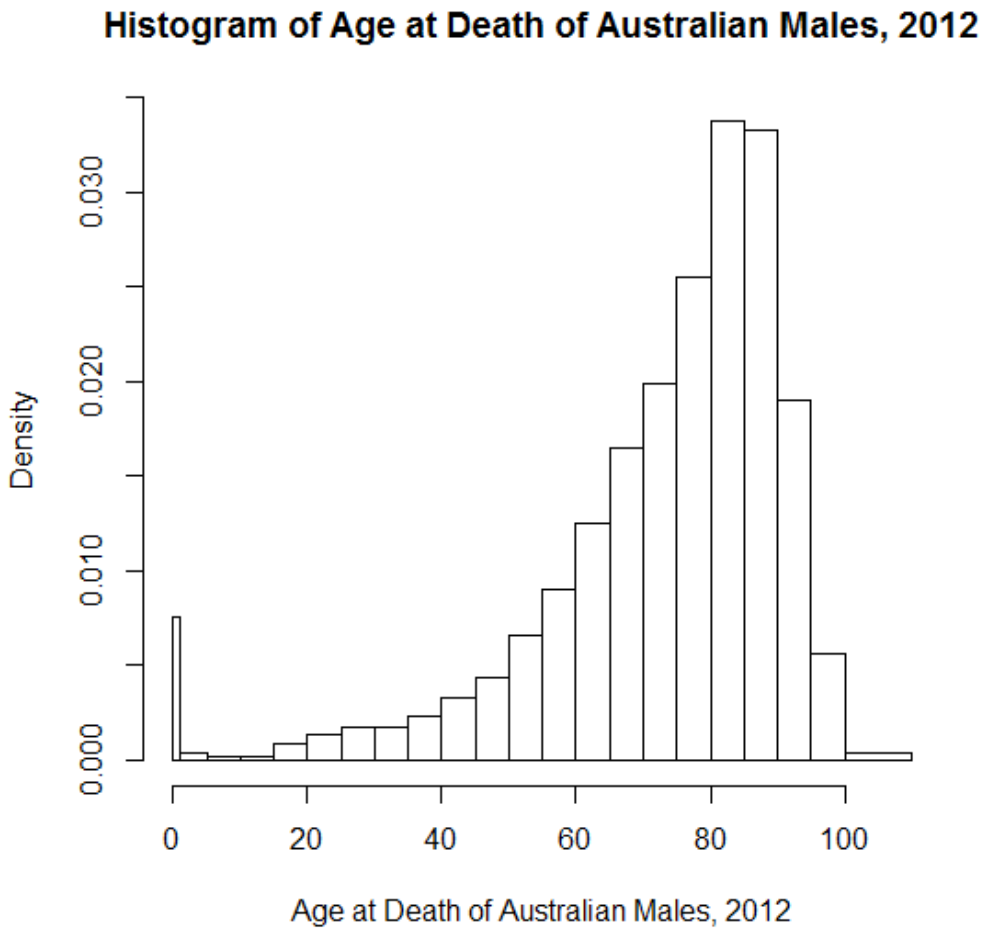

Figure 1. Histogram of age at death showing a distribution with a large left-hand tail and a short right-hand tail [Ref. 1].

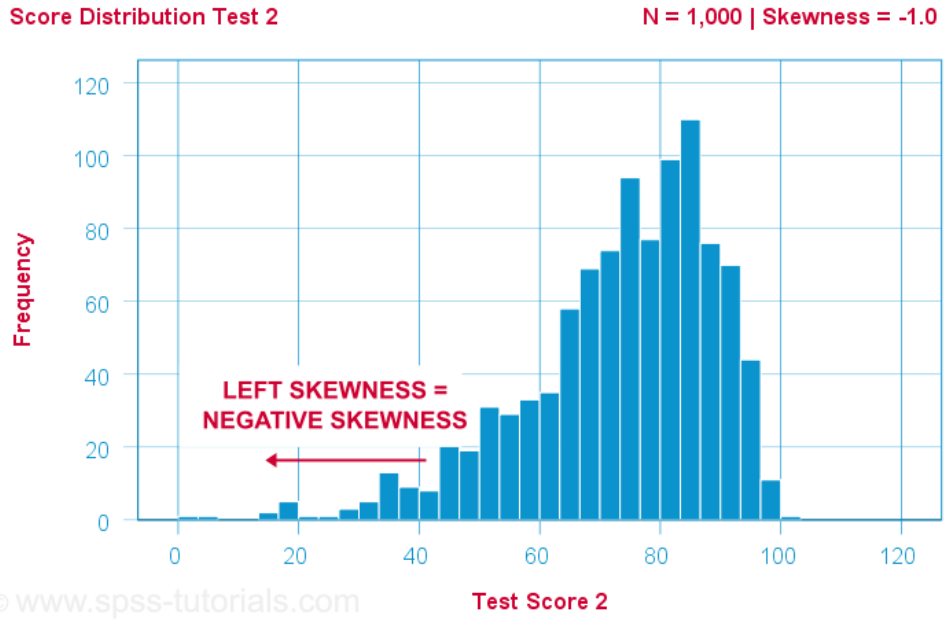

Figure 2. Histogram of test scores showing a distribution with a large left-hand tail and a short right-hand tail [Ref. 2].

2. The square-normal distribution

Figure 3 shows the cumulative distribution function of the test score of children in a school class obtained with the CumFreq model [Ref.3] for distribution fitting expressing the preference for the square-normal probability distribution. It is created using the squared values of the test scores and applying the normal distribution to these squared values. The fit of the data is quite close.

Figure 4 shows the theoretical histogram of the function using a division of the data into 10 categories. It also shows the corresponding probability density function, which is clearly skewed to the left. Further it indicates the observed frequency values in each category.

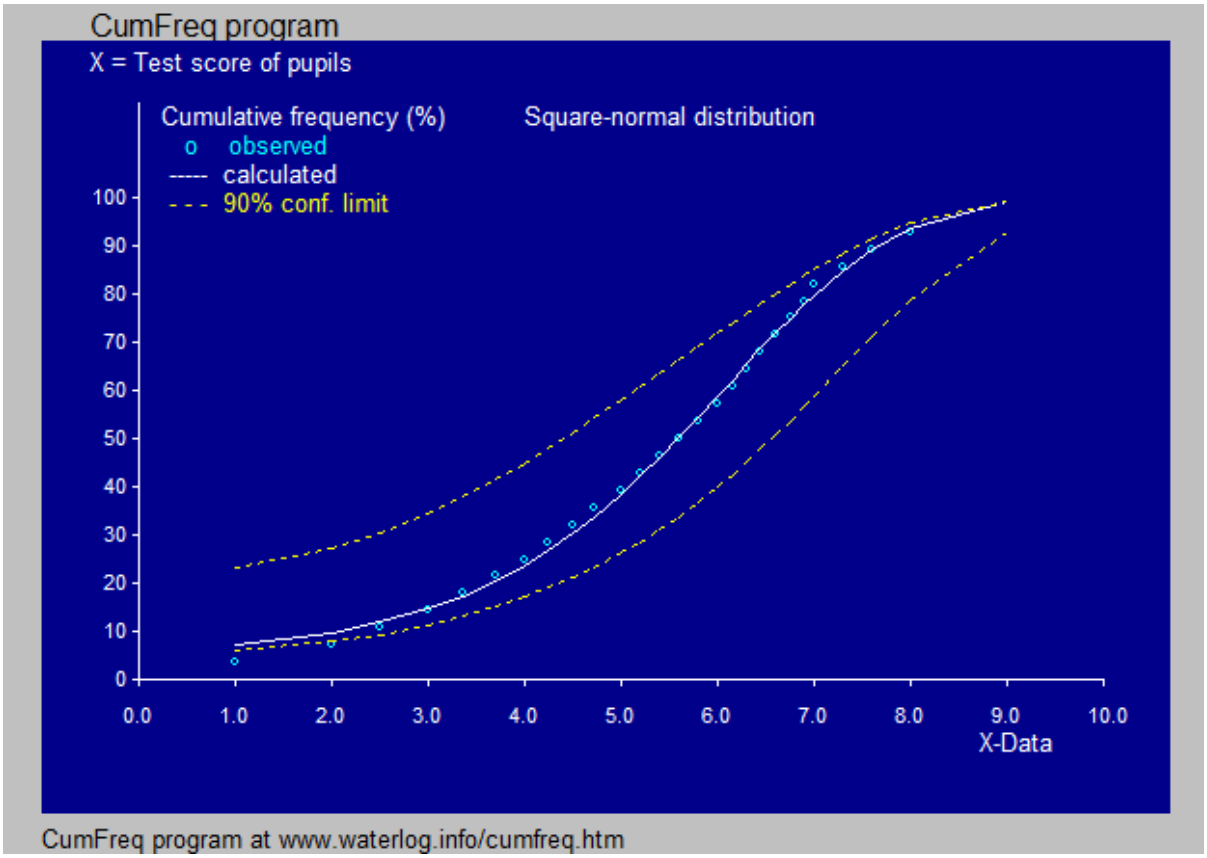

Figure 3. The cumulative distribution function of the test score of children in a school class obtained with the CumFreq model [Ref.3] for distribution fitting expressing the preference for the square-normal probability distribution.

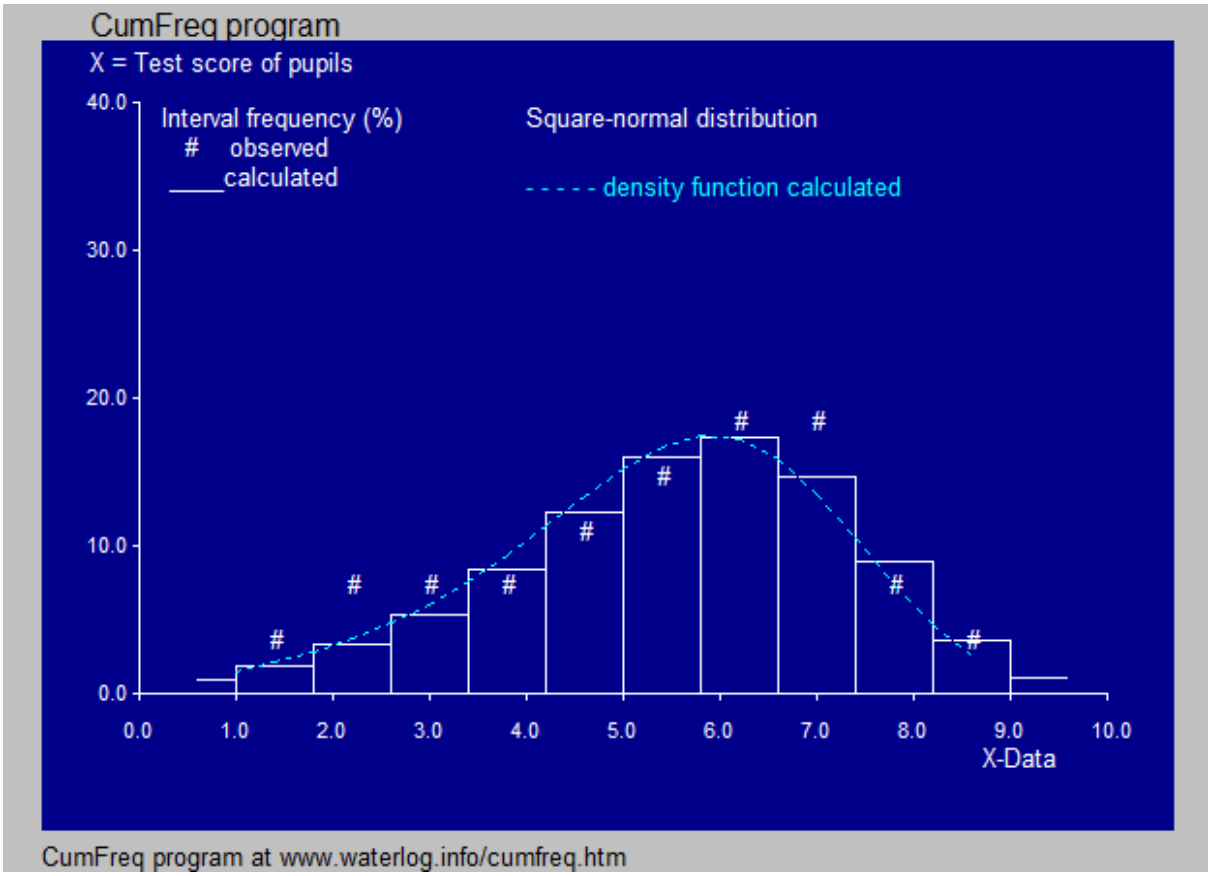

Figure 4. The calculated histogram of the function depicted in figure 3 using a division of the data into 10 categories. It also shows the corresponding probability density function, which is clearly skewed to the left.

2. The mirrored Gumbel distribution

Figure 5 shows the cumulative distribution function of the test score of children in a school class obtained with the CumFreq model [Ref.3] for distribution fitting expressing the preference for the generalized mirrored Gumbel probability distribution. The fit of the data is quite close and the result is practically the same as in figure 3.

Figure 6 shows the theoretical histogram of the function using a division of the data into 10 categories. It also shows the corresponding probability density function, which is clearly skewed to the left. Further it indicates the observed frequency values in each category. Figure 5 and figure 3 are not much different.

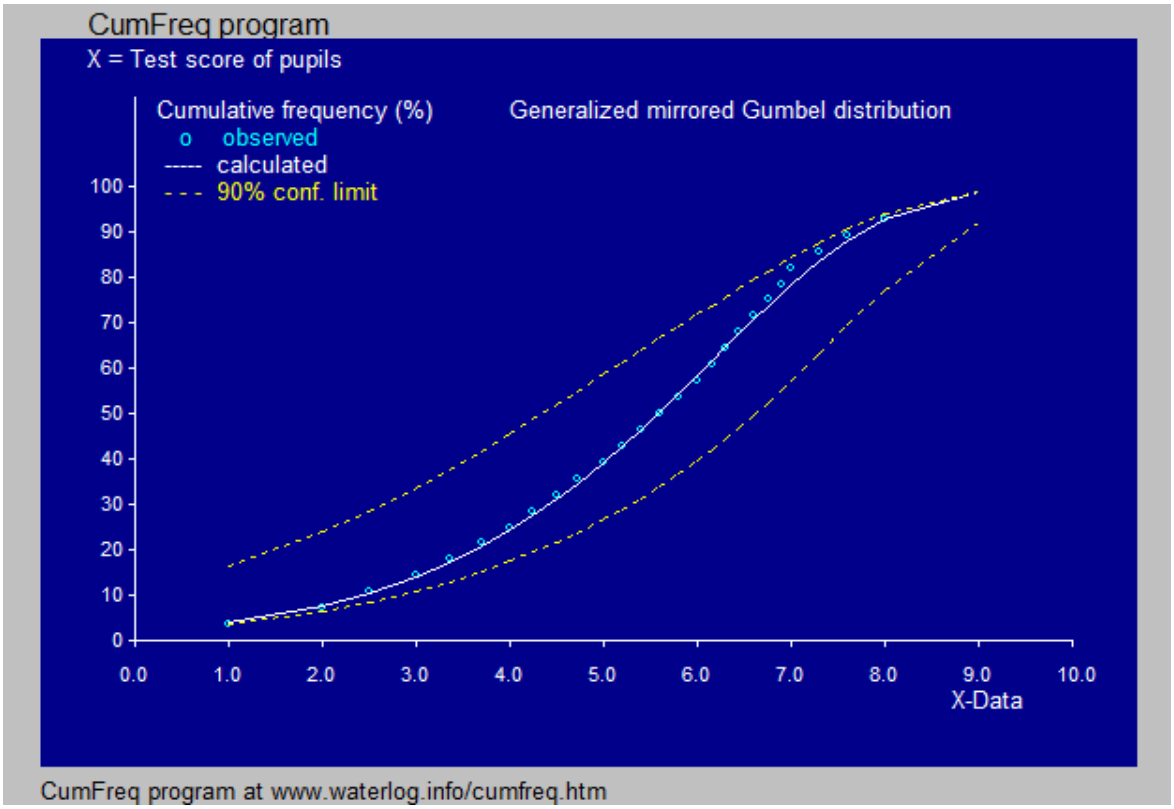

Figure 5. The cumulative distribution function of the test score of children in a school class obtained with the CumFreq model [Ref.3] for distribution fitting expressing the preference for the generalized mirrored Gumbel probability distribution.

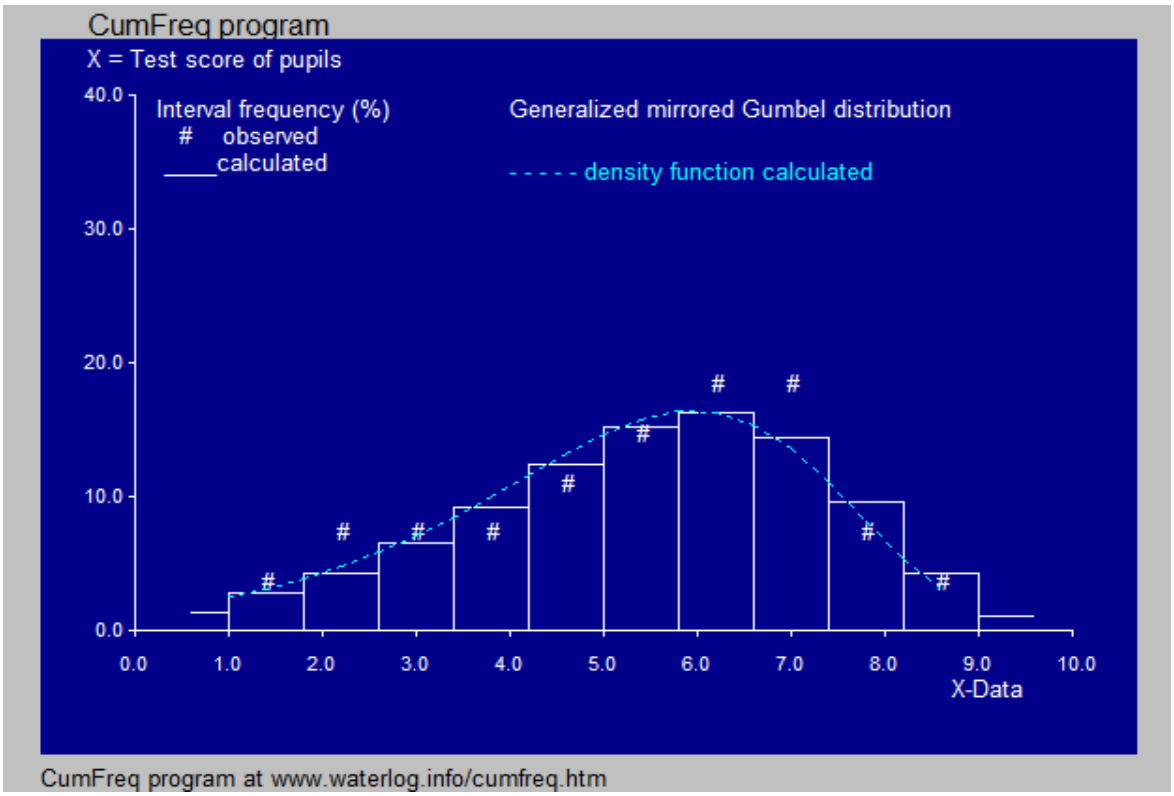

Figure 6. The calculated histogram of the function depicted in figure 5 using a division of the data into 10 categories. It also shows the corresponding probability density function, which is clearly skewed to the left.

The cumulative distribution function of Gumbel is:

$$Fc = 1 - \exp [- \exp \{ - (A*X+B) \}] \quad \{Eq. 1\}$$

Where Fc = cumulative frequency, X = variable under investigation, A and B are parameters. Fc can be estimated from Fc = R / (N+1) where R is the rank number of X arranged in ascending order and N the number of data.

The *mirrored* cumulative distribution function of Gumbel is:

$$Fc = \exp [- \exp \{ - (A*X+B) \}] \quad \{Eq. 2\}$$

The *generalized* mirrored cumulative distribution function of Gumbel is:

$$Fc = \exp [- \exp \{ - (A*Z+B) \}] \quad \{Eq. 3\}$$

Where Z = X^E, E being an exponent to be determined by iteration and optimization, i. e. selecting that value of E that produces the highest goodness of fit

Using the transformation:

$$Ft = - Ln \{ - Ln (1-Fc) \}$$

equation 3 can be written as

$$Ft = A*Z + B$$

so that A and B can be simply found from a linear regression of Ft upon Z.

In the example of figure 4, the values of E, A and B are found to be 0.870, -0.815, and 4.00 respectively.

4. Conclusion

For the determination of probability distributions of data that show a tendency of negative skewness to the left the CumFreq model can be used indicating a preference for the square-normal or mirrored Gumbel distribution.

5 – Addendum (generalized logistic distribution)

Figure 7 shows the cumulative distribution function of the test score of children in a school class obtained with the CumFreq model [Ref.3] for distribution fitting expressing the preference for the versatile generalized logistic probability distribution [Ref.4]. The fit of the data is quite close and the result is practically the same as in figures 3 and 5.

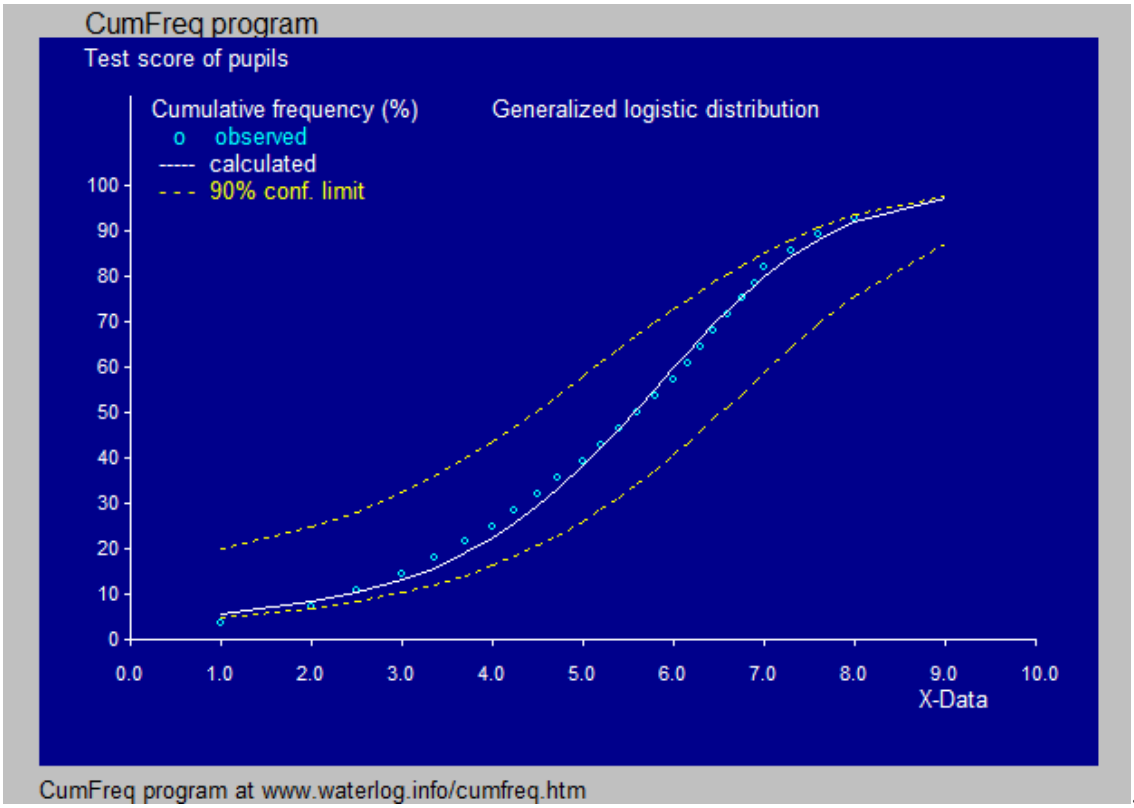

Figure 7. The cumulative distribution function of the test score of children in a school class obtained with the CumFreq model [Ref.3] for distribution fitting expressing the preference for the generalized logistic probability distribution.

6. References

[Ref. 1] Real life examples of distributions with negative skewness. On line: https://stats.stackexchange.com/questions/89179/real-life-examples-of-distributions-with-negative-skewness

[Ref. 2] Skewness – Quick Introduction, Examples & Formulas. On line: https://www.spss-tutorials.com/skewness/

[Ref. 3] CumFreq, free software for probability distribution fitting. On line: https://www.waterlog.info/cumfreq.htm

[Ref.4] https://www.researchgate.net/publication/335022301_FITTING_THE_VERSATILE_LINEARIZED_COMPOSITE_AND_GENERALIZED_LOGISTIC_PROBABILITY_DISTRIBUTION_TO_A_DATA_SET