

Finding a fitting bimodal probability distribution for a data set having a histogram with two peaks and a valley in between

R.J. Oosterbaan
November 2022

www.waterlog.info

Abstract

The free CumFreqA program software offers the option to find a composite probability distribution to a data set. It is composed of two different distributions, one for the lower part of the data set and one for the upper part with a separation point in between. When the data set reveals a histogram with two peaks and a valley (depression) in between, the composite distribution can detect a bimodal probability function. In this article demonstrations are given of composite distributions consisting of parts with symmetrical components or components being skew to the left or skew to the right. The results are compared with results found in literature with modernized probability distributions,

Contents

1. Introduction
 - 1.1 The Khaleel case
 - 1.2 The Kilai, Jallal, and Albalawi cases
 - 1.3 The Elbatal case
 - 1.4 The Hassan case
 - 1.5 Notes
2. Bimodal distribution with two symmetrical components
3. Bimodal distribution with two components skew to the left
4. Bimodal distribution with two components skew to the right
5. Mixed bimodal distributions
6. Summary and conclusion
7. References
8. Appendices
 - 8A. Screen-print of the CumFreqA input file showing composite probability distribution options.
 - 8B. Examples of unimodal symmetrical, left skew and right skew distributions.

1. Introduction

Mundher A. Khaleel et al. (2022), have developed new probability distributions with applications to Covid mortality rates in Iraq.

Mutua Kilai et al. (2022), Olayan Albalawi et al (2022) and Muzamil Jallal (2022) have developed new probability distributions with applications to Covid mortality rates in Italy.

Ibrahim Elbatal et al. (2022). have developed a new probability distribution with an analysis of the failure times for a specific product.

Amal S. Hassan et al. (2020). have developed a new probability distribution with an analysis of the failure times of 50 devices.

Their cases appear to be applied to bimodal probability distributions.

In the following sections, first the Khaleel case is analyzed while thereafter the Kilai, Albalawi and Jallal cases are dealt with to be followed by the Elbatal case and finally by the Hassan case.

The cases are compared with composite probability distributions that can be made with the free CumFreqA software and that are often useful in binomial situations.

1.1 The Khaleel case

Mundher A. Khaleel et al. (2022) have developed and used the $[0, 1]$ truncated inverse Weibull Rayleigh probability distribution ($[0,1]$ TIWR) probability distribution with an analysis of the Covid death rates of patients in Iraq. Their result is shown in figure 1. The figure gives the impression that the histogram indicates the presence of a bimodal probability distribution.

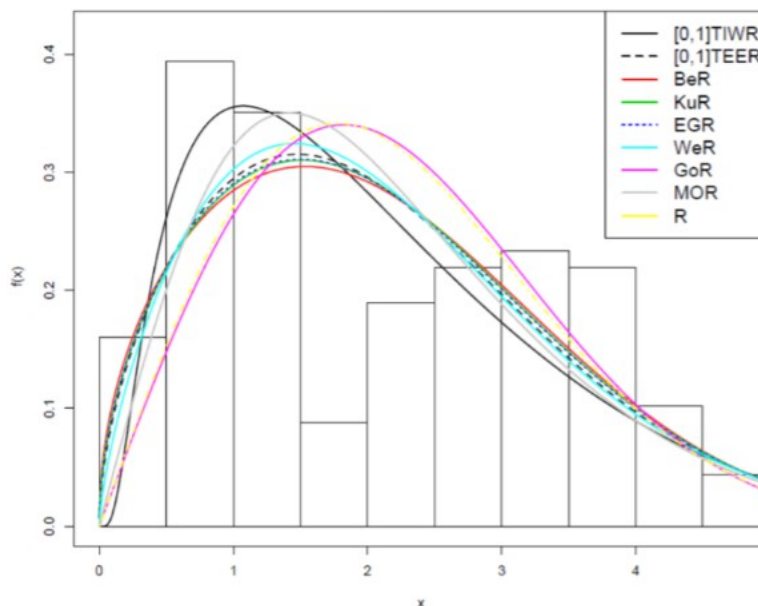


Figure 1. Histogram and probability density functions regarding the Covid death rates in Iraq of various distributions including $[0,1]$ TIWR (Khaleel case). The histogram suggests a bimodal situation, reason why the curves are distant from the observed data in the range of $X = 1.5$ to 2.5 where a valley (depression) in the histogram occurs.

The free CumFreqA software program (Oosterbaan, 2000), which is an amplification of the CumFreq model, using the same Iraqi data and the option to find the best fitting composite distribution (Appendix 8A) , gives as results the composite cumulative probability function (CDF, figure 2) and the histogram based on the observed data together with the theoretically best fitting bimodal probability density curve (PDC, figure 3).

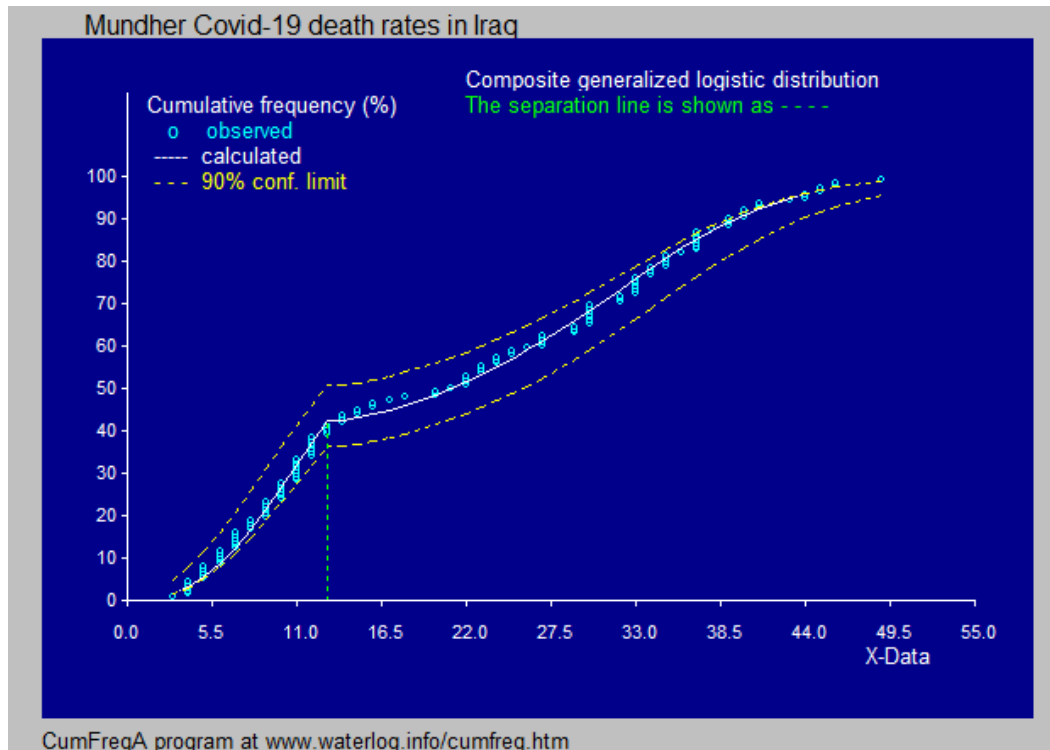


Figure 2. Cumulative probability function (CDF) showing a composite generalized logistic distribution regarding the Covid death rates in Iraq made with CumFreqA. The confidence belt is also shown. The separation point between the two distributions with which the composition is made is at $X = 13$.

The mathematical formulation of the two composing logistic distributions is:

$$X < 13 : \quad \text{CDF} = 1 / \{ 1 + \exp (-260 X^{0.01} + 268) \}$$

$$X > 13 : \quad \text{CDF} = 1 / \{ 1 + \exp (-0.0077 X^{2.48} + 44.2) \}$$

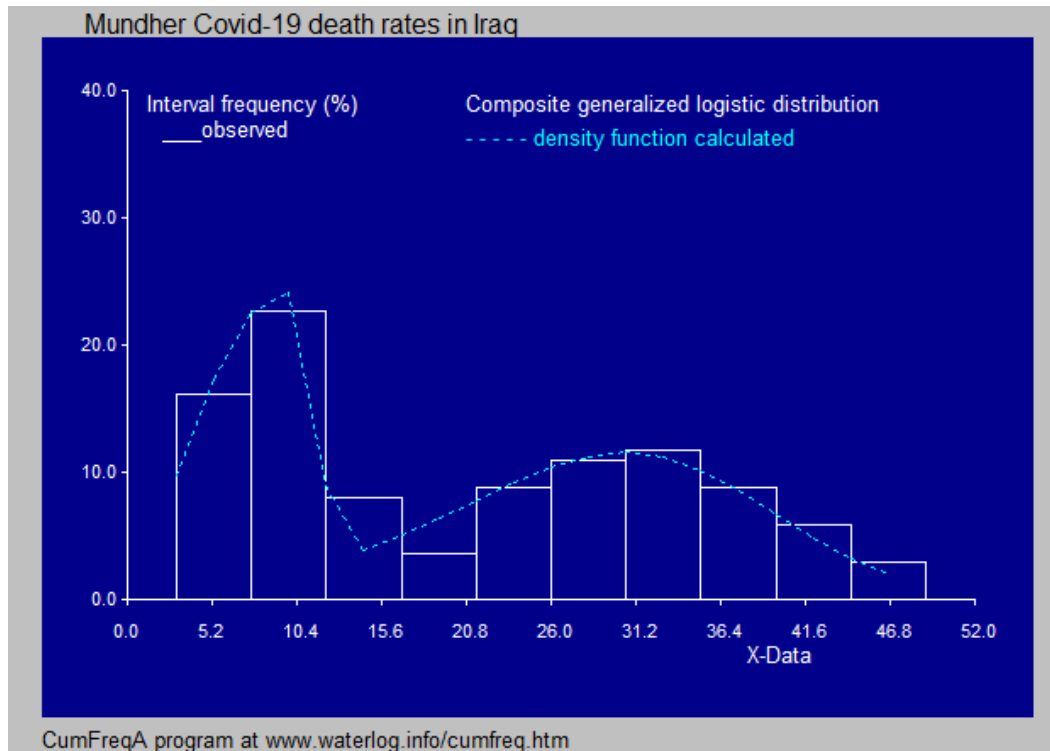


Figure 3. Histogram and probability density curve (PDC) of the composite logistic distribution regarding the Covid death rates in Iraq. The fit of the of the bimodal density function is better than the fit according to the $[0,1]$ TIWR distribution for the Khaleel case in figure 1.

In figure 3 it looks that the lower part of the PDC is skew to the left whereas the upper part is symmetrical.

In general the skewness of the composing parts can be symmetrical, skew to the left or skew to the right. In the sections 2 to 5 different combinations of symmetry and skewness will be demonstrated for bimodal models, beginning with (a) two symmetrical components, then (b) two components skew to the left, (c) two components skew to the right, and (d) various mixtures.

1.2 The Kilai, Jallal, and Albalawi cases

Mutua Kilai et al. (2022) have developed and used the the Exponentiated Generalized Gull Alpha Power Rayleigh (EGGAPR) probability distribution with an analysis of the Covid mortality rates of patients in Italy. Their result is shown in figure 4. The figure, like in the Khaleel case, gives the impression that the histogram indicates the presence of a bimodal probability distribution.

Muzamil Jallal et al. (2022) have developed the Weibull Inverse Power Rayleigh Distribution (WIPRD), with an analysis of the same Covid data. Their result is shown in figure 5. The figure gives the impression that the histogram indicates the presence of a bimodal probability distribution.

Olayan Albalawi et al. (2022) have developed and used the Generalized Logarithmic Transformation Exponential (GLTE) probability distribution with an analysis of the same Covid data. Their result is shown in figure 6. The figure gives the impression that the histogram indicates the presence of a bimodal probability distribution.

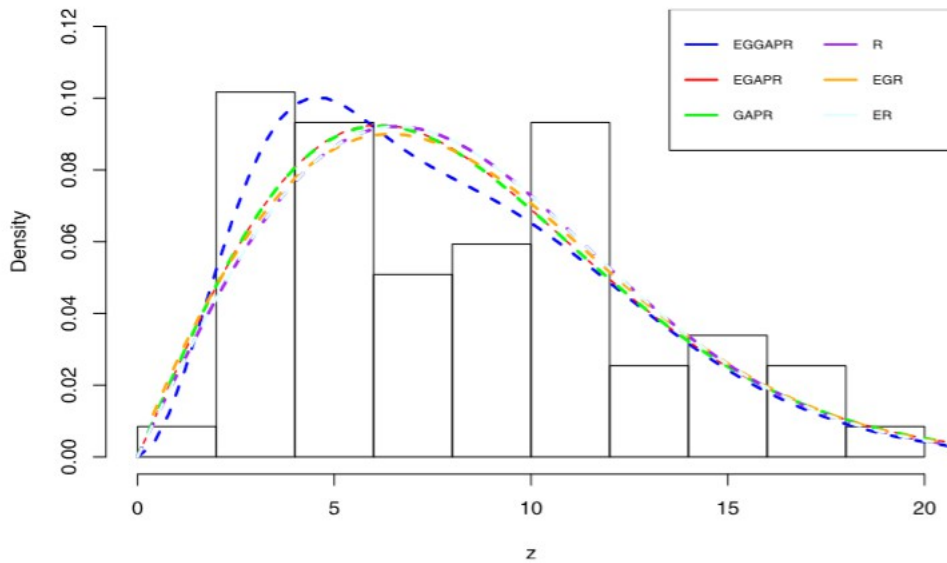


Fig. 12. Fitted densities plot for Italy COVID-19 mortality rates.

Figure 4. Histogram and probability density functions regarding the Covid death rates in Italy of various distributions including EGGAPR (Kilai case). The histogram suggests a bimodal situation, reason why the curves are distant from the observed data in the range of $X = 6$ to 10 where a valley (depression) in the histogram occurs.

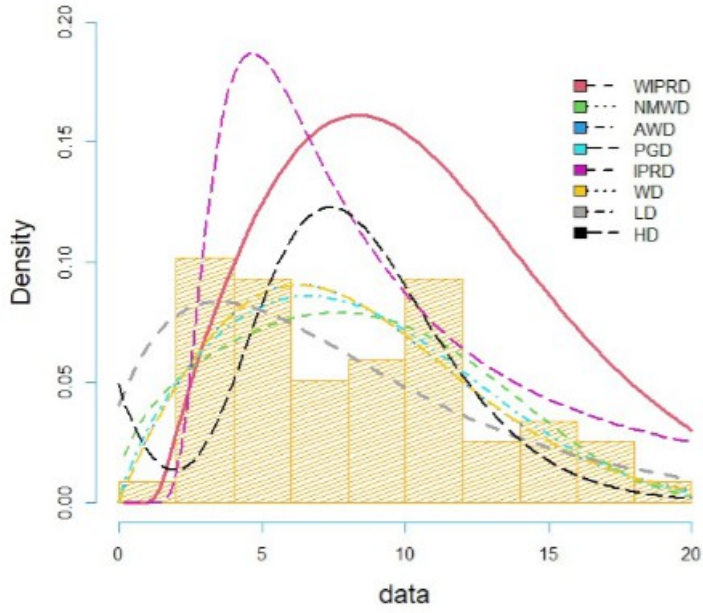


Figure 5. Histogram and probability density functions regarding the Covid death rates in Italy of the WIPRD distribution (Jallal case). The histogram suggests a bimodal situation, reason why the WIPRD curve (red color) is distant from the observed data in the range of $X = 6$ to 10 where a valley (depression) in the histogram occurs.

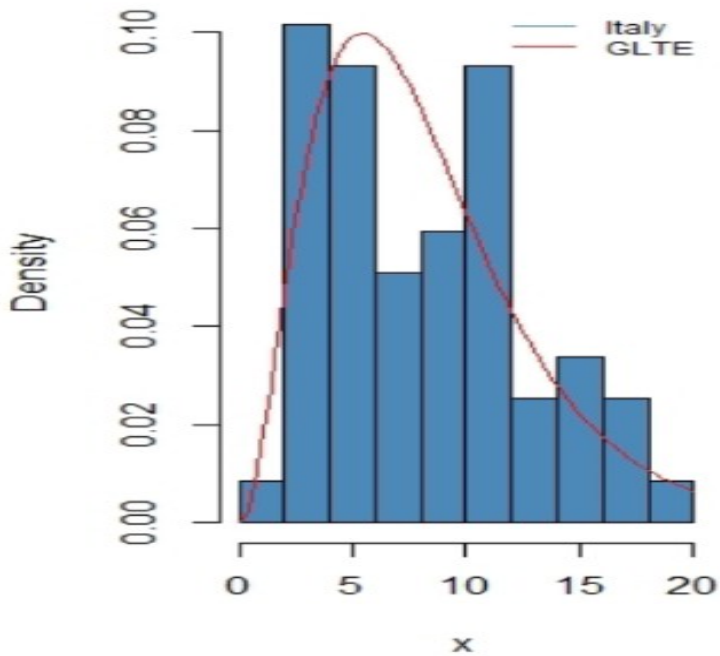


Figure 6. Histogram and probability density functions regarding the Covid death rates in Italy of the GLTE distribution (Albalawi case). The histogram suggests a bimodal situation, reason why the curves are distant from the observed data in the range of $X = 6$ to 10 where a valley (depression) in the histogram occurs.

The free CumFreqA software program (Oosterbaan, 2000), which is an amplification of the CumFreq model, using the same Iraqi data and the option to find the best fitting composite distribution (Appendix 8A) , gives as results the composite cumulative probability function (CDF, figure 7) and the histogram based on the observed data together with the theoretically best fitting bimodal probability density curve (PDC, figure 8).

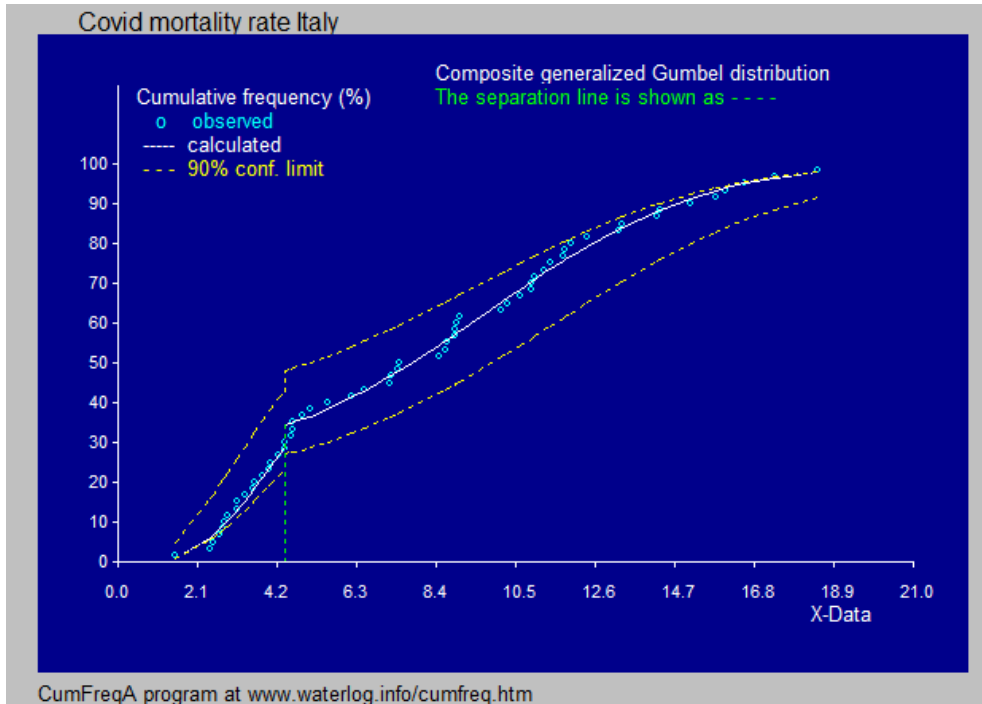


Figure 7. Cumulative probability function (CDF) showing a composite generalized logistic distribution regarding the Covid death rates in Italy made with CumFreqA. The confidence belt is also shown. The separation point between the two distributions with which the composition is made is at $X = 4.4$.

The mathematical formulation of the two composing distributions is:

$$X < 4.4 : \quad \text{CDF} = \text{Exp} \{ - \exp - (2.12 X^{0.40} - 4.06) \}$$

$$X > 4.4 : \quad \text{CDF} = \text{Exp} \{ - \exp - (-0.0067 X^{2.2} - 0.244) \}$$

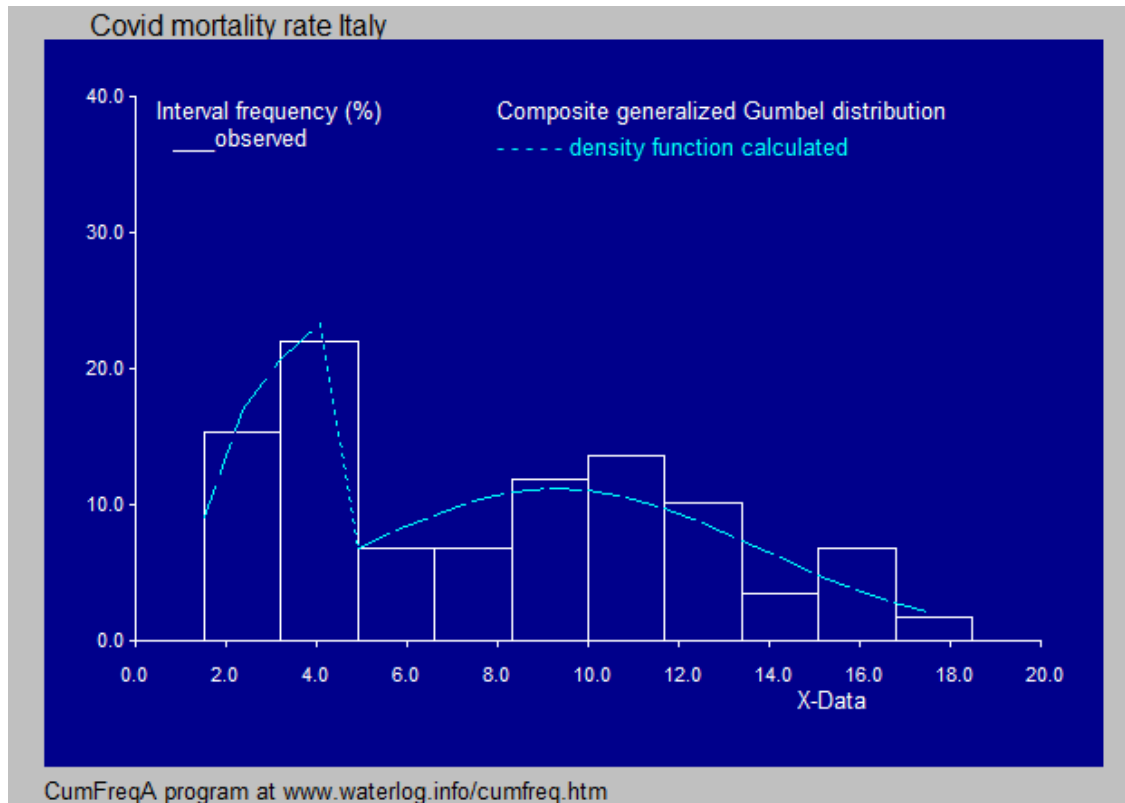


Figure 8. Histogram and probability density curve (PDC) of the composite generalized Gumbel distribution regarding the Covid death rates in Italy. The fit of the of the bimodal density function is better than the fit according to the EGGAPR distribution in figure 4, the WIPRD distribution in figure 5 and the GLTE distribution in figure 6.

In figure 8 it looks that the lower part of the PDC is skew to the left whereas the upper part is skew to the right and not symmetrical as in figure 3.

In general the skewness of the composing parts can be symmetrical, skew to the left or skew to the right. In the next sections different combinations of symmetry and skewness will be demonstrated for bimodal models, beginning with (a) two symmetrical components, then (b) two components skew to the left, (c) two components skew to the right, and (d) various mixtures.

1.3 The Elbatal case

Ibrahim Elbatal et al. (2022). have developed and used the Odd Perks G- Class (OPE) probability distribution with an analysis of the failure times for a specific product. Their result is shown in figure 9. The figure gives the impression that the histogram indicates the presence of a bimodal probability distribution.

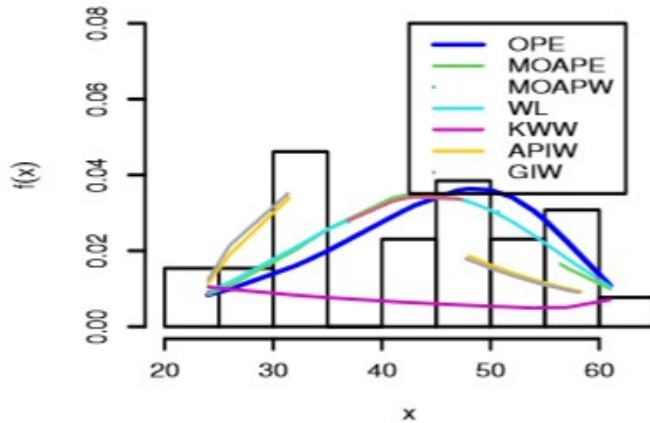


Figure 9. Histogram and probability density functions regarding the failure times for a specific product of various distributions including OPE (blue curve). The histogram suggests a bimodal situation, reason why the curves are distant from the observed data in the range of $X = 30$ to 40 while a valley (depression) in the histogram occurs in the range of $X = 35$ to 40 .

The free CumFreqA software program (Oosterbaan, 2000), which is an amplification of the CumFreq model, using the same Iraqi data and the option to find the best fitting composite distribution (Appendix 8A), gives as results the composite cumulative probability function (CDF, figure 10) and the histogram based on the observed data together with the theoretically best fitting bimodal probability density curve (PDC, figure 11).

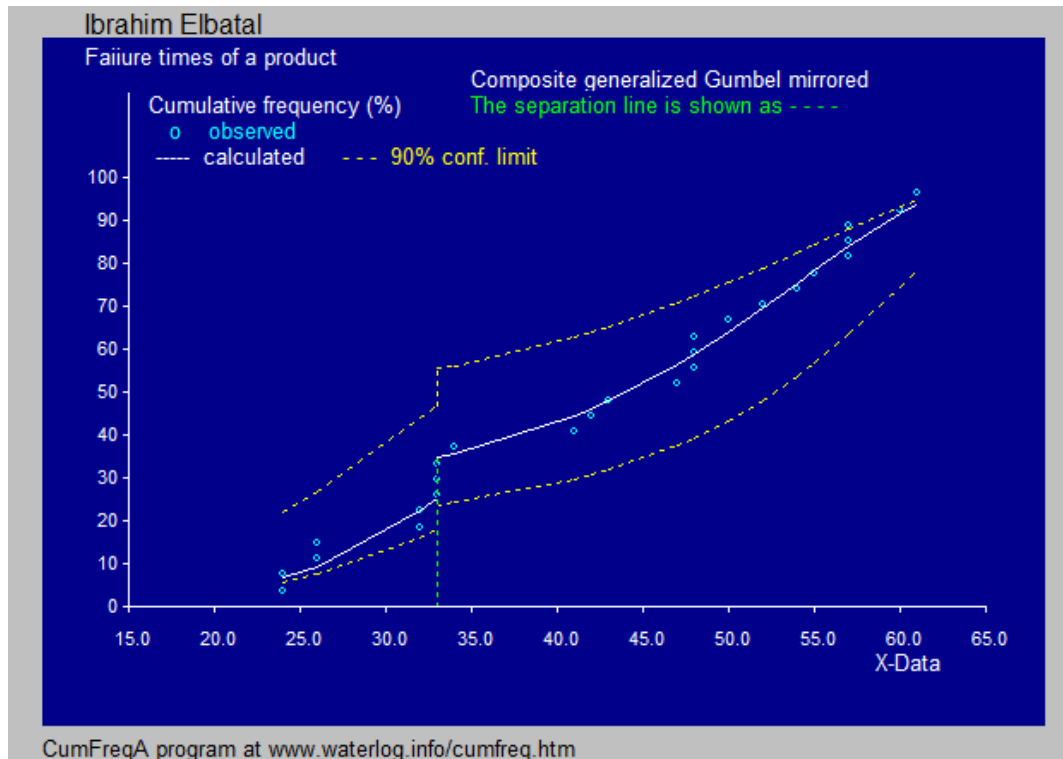


Figure 10. Cumulative probability function (CDF) showing a composite generalized logistic distribution regarding the failure times for a specific product made with CumFreqA. The confidence belt is also shown. The separation point between the two distributions with which the composition is made is at $X = 13$.

The mathematical formulation of the two composing generalized mirrored Gumbel distributions is:

$$X < 33 : \quad \text{CDF} = 1 - \text{Exp} \left\{ - \exp - (-440 X^{0.010} - 457) \right\}$$

$$X > 33 : \quad \text{CDF} = 1 - \text{Exp} \left\{ - \exp - (-0.00098 X^{3.00} + 1.20) \right\}$$

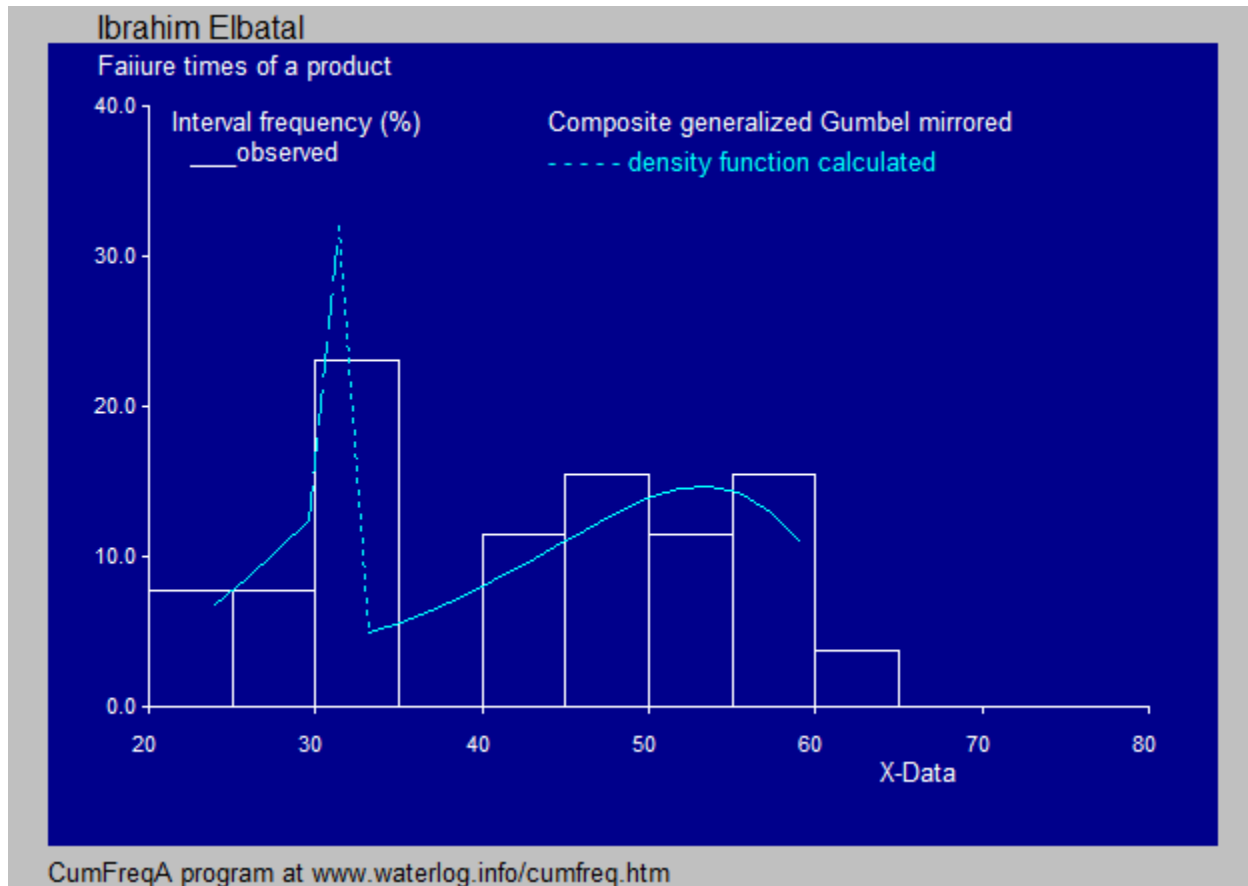


Figure 11. Histogram and probability density curve (PDC) of the composite generalized Gumbel distribution mirrored regarding the failure times for a specific product. The fit of the of the bimodal density function is better than the fit according to the OPE distribution in figure 9.

In figure 11 it looks that both parts of the PDC are skew to the left .

In general the skewness of the composing parts can be symmetrical, skew to the left or skew to the right. In the sections 2 to 5 different combinations of symmetry and skewness will be demonstrated for bimodal models, beginning with (a) two symmetrical components, then (b) two components skew to the left, (c) two components skew to the right, and (d) various mixtures.

1.4 The Hassan case

Amal S. Hassan et al. (2020). have developed and used the Inverted Topp-Leone (ITL) probability distribution with an analysis of the failure times of 50 devices. Their result is shown in figure 12. The figure gives the impression that the histogram indicates the presence of a bimodal probability distribution.

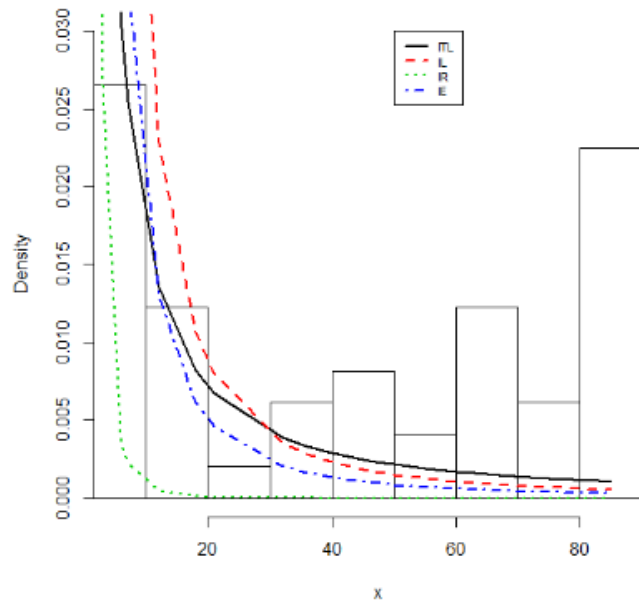


Figure 12. Histogram and probability density functions regarding the failure times of 50 devices of various distributions including ITL (black curve). The histogram suggests a bimodal situation, reason why the curves are distant from the observed data in the range of $X = 60$ to 90 while a valley (depression) in the histogram occurs in the range of $X = 20$ to 50 .

The free CumFreqA software program (Oosterbaan, 2000), which is an amplification of the CumFreq model, using the same failure data and the option to find the best fitting composite distribution (Appendix 8A), gives as a result the histogram based on the observed data together with the theoretically best fitting bimodal probability density curve (PDC, figure 13).

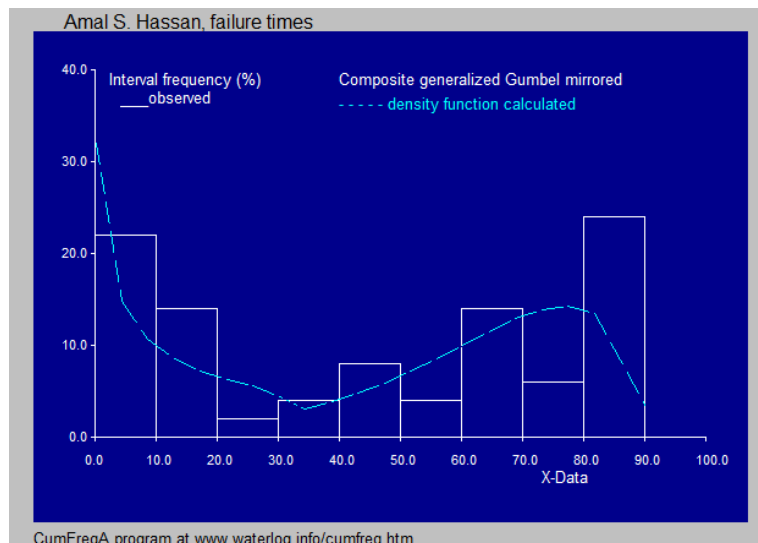


Figure 13. Histogram and probability density curve (PDC) of the composite generalized Gumbel distribution regarding the failure times of 50 devices. The fit of the of the bimodal density function is better than the fit according to the ITL distribution in figure 12.

1.5 Notes

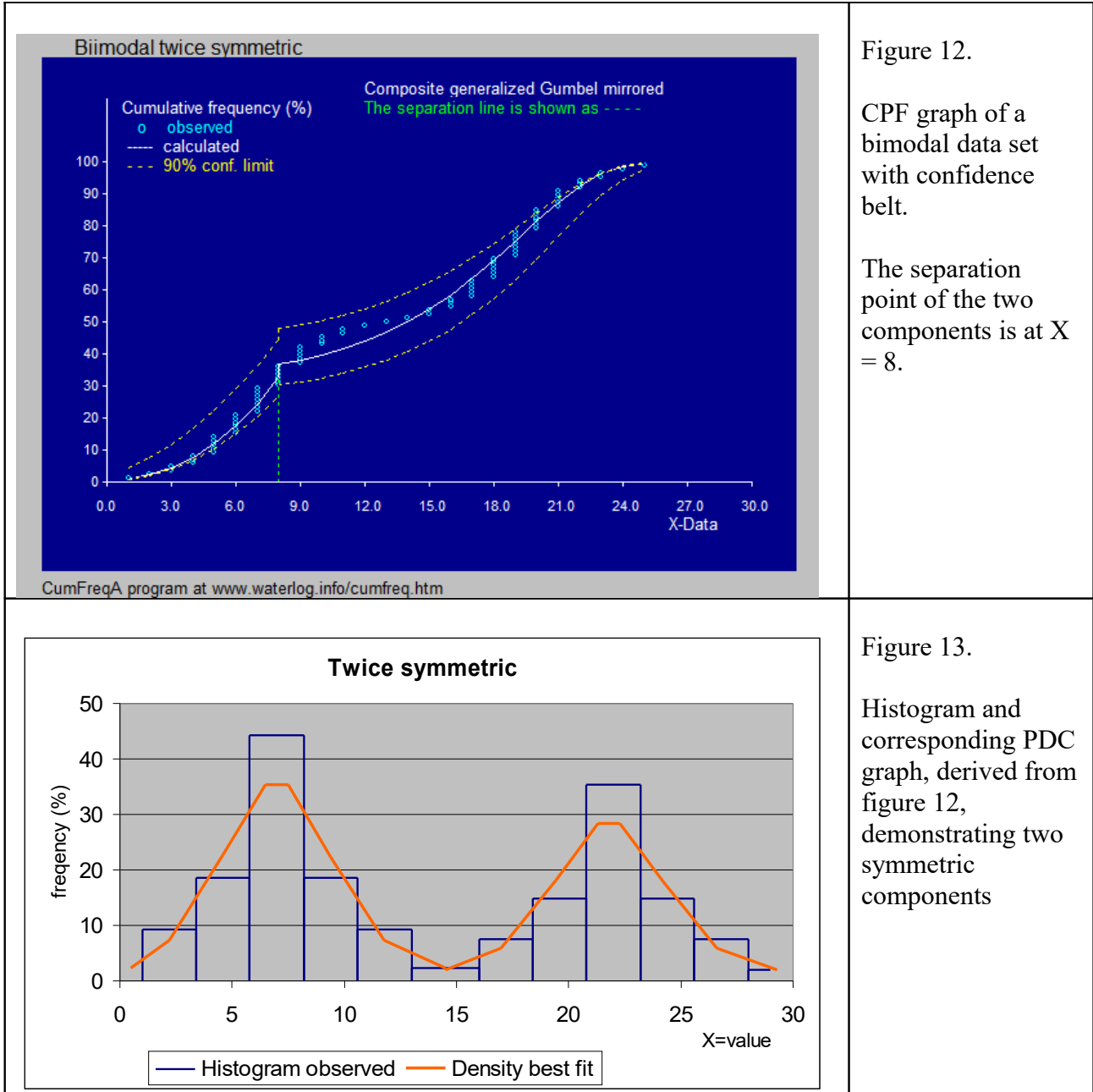
(a) The CDF according to CumFreqA in the Khaleel case (figure 3) is twice generalized logistic, in the Kilai case, the Jallal case and in the Albalawi case (figure 8) it is twice generalized Gumbel while in the Albatal case (figure 11) it is the mirrored version of Gumbel. The CumFreqA program can search for the best fitting bimodal (composite) distribution for each case (see Appendix 8A). Also there are differences in the symmetrical and skewness properties. These are further illustrated in the following sections.

(b) The CumFreq model uses transformations of probability distributions to determine their parameters by straightforward linear regression and to be able to fit symmetrical and skewed data sets easily (Oosterbaan 2021). Further the program uses the simple method of plotting positions to determine cumulative probabilities (Oosterbaan, 2002)

(c) Examples of the histogram and probability density curves (PDC's) of unimodal symmetrical, left skew, and right skew distributions are given in Appendix 8B.

2. Bimodal distribution with two symmetrical components

The CPF and PDC graphs of a bimodal probability distribution with two symmetrical components analyzed with CumFreqA are given in figures 12 and 13 below.



The cumulative probability function CPF is composite generalized Gumbel mirrored:

$$X < 8 : \quad \text{CPF} = 1 - \exp \left[-\exp \left\{ - \left(-2.89 X^{0.40} + 7.57 \right) \right\} \right]$$

$$X > 8 : \quad \text{CPF} = 1 - \exp \left[-\exp \left\{ - \left(-0.0034 X^{2.78} + 88.9 \right) \right\} \right]$$

3. Bimodal distribution with two components skew to the left

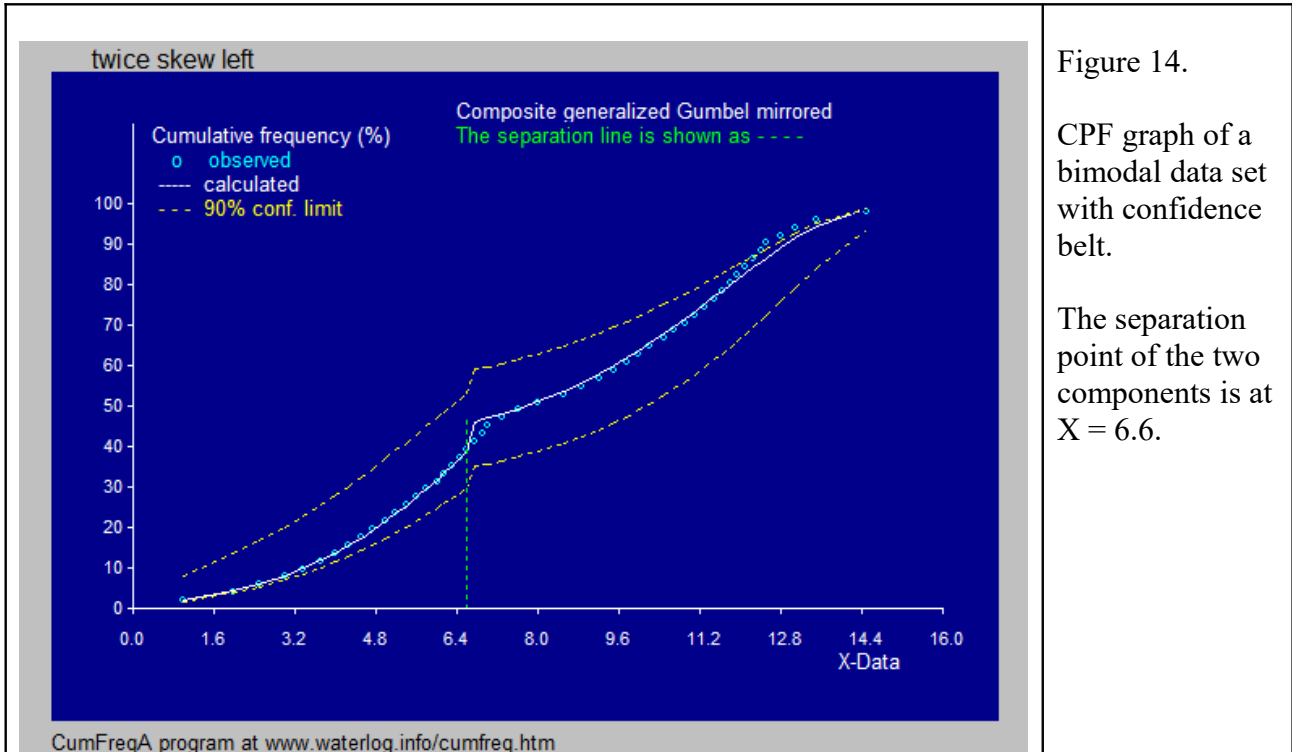


Figure 14.

CPF graph of a bimodal data set with confidence belt.

The separation point of the two components is at $X = 6.6$.



Figure 15.

Histogram and corresponding PDC graph, derived from figure 14, demonstrating two components skewed to the left.

The cumulative probability function CPF is composite generalized Gumbel mirrored:

$$X < 6.6 : \quad \text{CPF} = 1 - \exp \left[-\exp \left\{ - \left(-1.97 X^{0.52} + 5.96 \right) \right\} \right]$$

$$X > 6.6 : \quad \text{CPF} = 1 - \exp \left[-\exp \left\{ - \left(-0.0071 X^{3.00} + 0.70 \right) \right\} \right]$$

4. Bimodal distribution with two components skew to the right

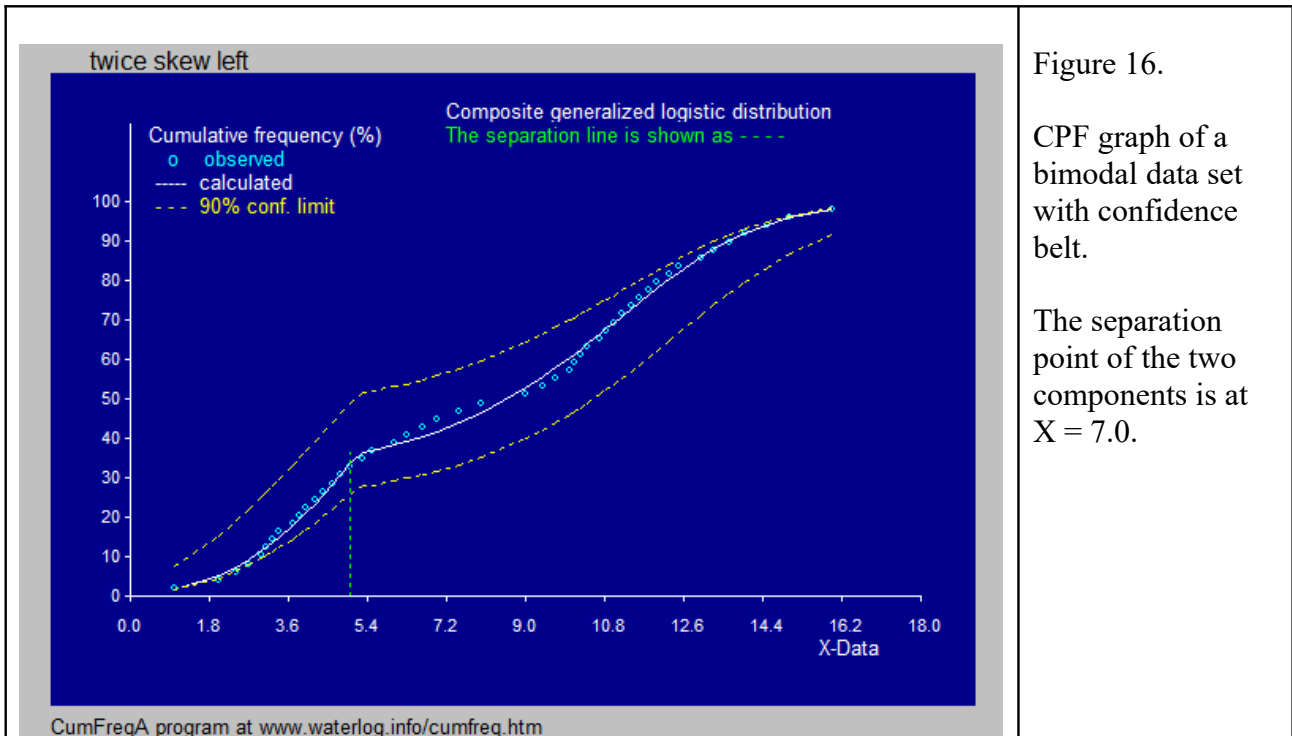


Figure 16.

CPF graph of a bimodal data set with confidence belt.

The separation point of the two components is at $X = 7.0$.

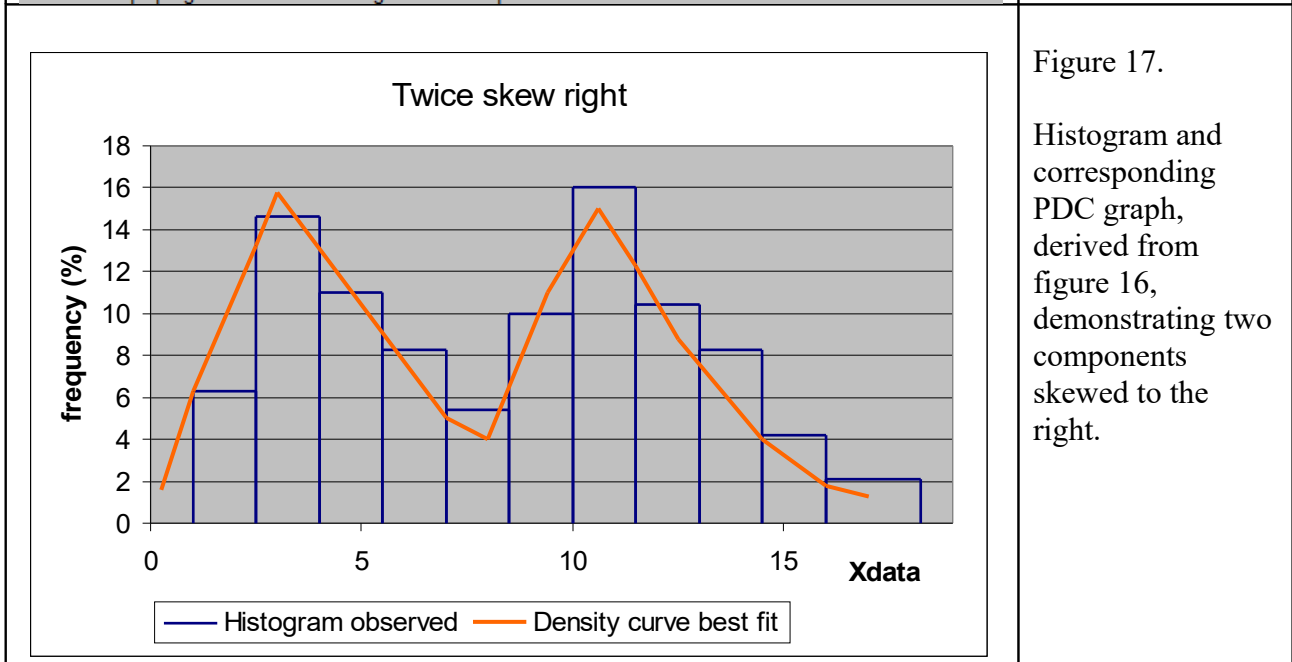


Figure 17.

Histogram and corresponding PDC graph, derived from figure 16, demonstrating two components skewed to the right.

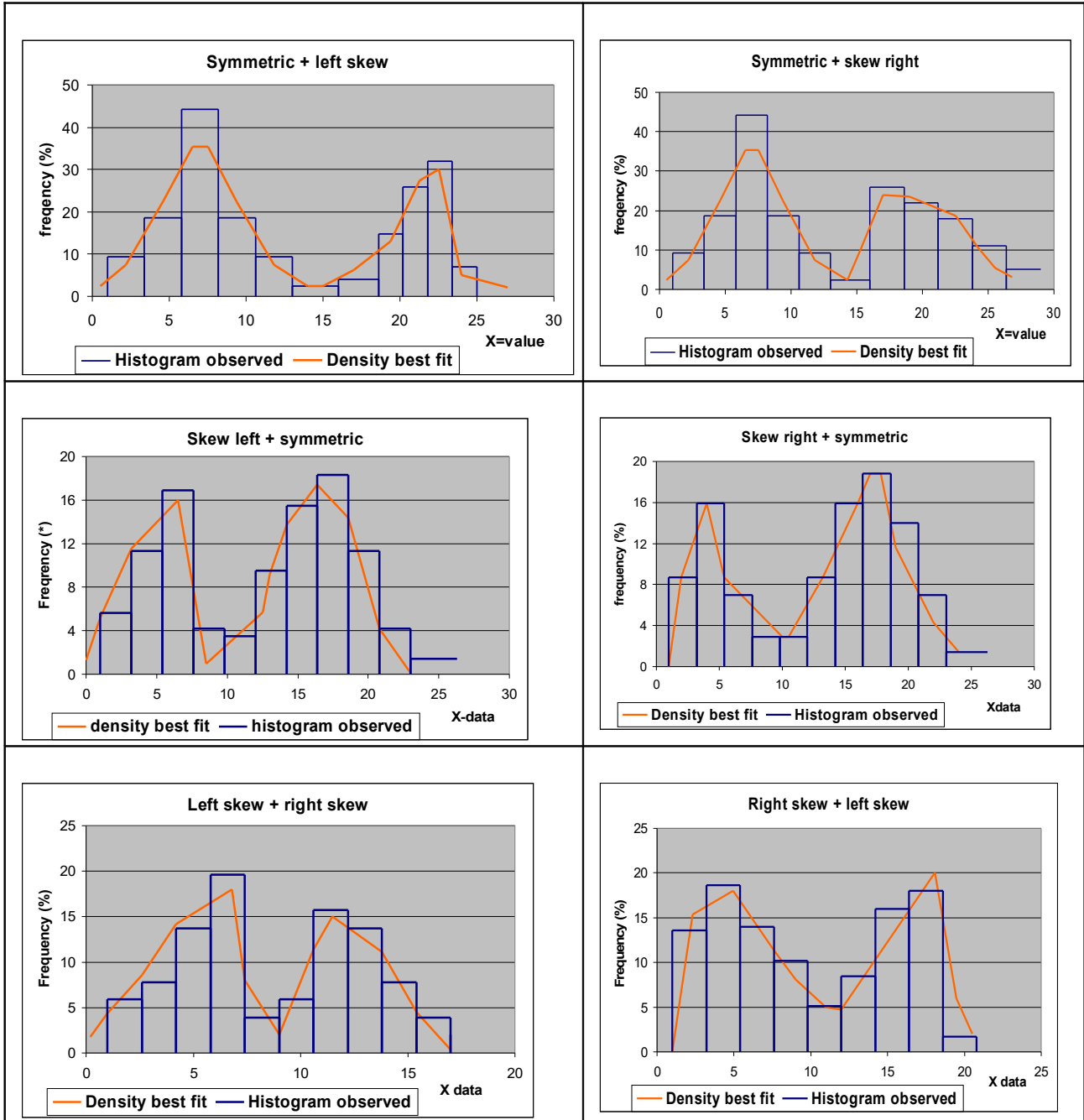
The composite model is twice generalized logistic. The mathematical formulation of the two composing distributions is:

$$X < 7 : \quad \text{CDF} = 1 / \{ 1 + \exp (-260 X^{0.01} + 268) \}$$

$$X > 7 : \quad \text{CDF} = 1 / \{ 1 + \exp (-0.0077 X^{2.48} + 44.2) \}$$

5. Mixed bimodal distributions

The following figures give a brief overview of possible combinations of symmetrical, left skew and right skew bimodal histograms and probability density curves. Only the density functions and the histograms are given because they give more insight in the shape of the distribution.



6. Summary and conclusion

It was seen that the modernized probability distributions like $[0,1]$ TIWR, EGGAPR, WIPRD GLTE, and OPE did not cover the valley (depression) between the two peaks of the bimodal histogram of the data set concerning Covid death rates in Iraq respectively Italy and failure rates of a product. However, the composite generalized logistic distribution respectively the composite generalized Gumbel distribution and the the composite generalized Gumbel mirrored, all three consisting of two different distributions, one for the data below the separation point and the other for the data to the above, were able to create a bimodal probability distribution that followed the peaks and the valley properly

The free CumFreqA software program provides the option to use composite distributions. It is easy to handle by pasting the data copied from an Excel file on to the input menu (see Appendix 8A) and letting the program automatically perform the calculations and produce an output menu with the relevant equations while also the graphics can be inspected.

7. References

Ibrahim Elbatal, Naif Alotaibi, Ehab M. Almetwally, Salem A. Alyami and Mohammed Elgarhy (2022). On Odd Perks-G Class of Distributions: Properties, Regression Model, Discretization, Bayesian and Non-Bayesian Estimation, and Applications. *Symmetry* 2022, 14, 883.

<https://doi.org/10.3390/sym14050883>

Mundher A. Khaleel, Abdulwahab M. Abdulwahaba, Awni M. Gaftana, Moudher Kh. Abdalhammed (2022), A new $[0, 1]$ truncated inverse Weibull-Rayleigh distribution properties with application to COVID-19. *Int. J. Nonlinear Anal. Appl.* 13 (2022) 1, 2933–2946
ISSN: 2008-6822 (electronic). <http://dx.doi.org/10.22075/ijnaa.2022.6026>

Mutua Kilai, Gichuhi A. Waititu, Wanjoya A. Kibira, M.M. Abd El-Raouf, Tahani A. Abushal (2022). A new versatile modification of the Rayleigh distribution for modeling COVID-19 mortality rates. *Results in Physics* 35(1):105260. <http://dx.doi.org/10.1016/j.rinp.2022.105260>

Muzamil Jallal, Aijaz Ahmad, Rajnee Tripathi (2022). Weibull Inverse Power Rayleigh Distribution with Applications Related to Distinct Fields of Science. https://www.researchgate.net/publication/361504378_Weibull_Inverse_Power_Rayleigh_Distribution_with_Applications_Related_to_Distinct_Fields_of_Science

Olayan Albalawi, Naresh Chandra Kabdwal, Qazi J. Azhad, Rashi Hora and Basim S. O. Alsaedi (2022). Estimation of the Generalized Logarithmic Transformation Exponential Distribution under Progressively Type-II Censored Data with Application to the COVID-19 Mortality Rates. *MPDI Mathematics*, 2022, <http://dx.doi.org/10.3390/math10071015>

Amal S. Hassan¹, Mohammed Elgarhy, and Randa Ragab (2020). Statistical Properties and Estimation of Inverted Topp-Leone Distribution. *Journal of Statistics Applications & Probability* 9, No. 2, 319-331 (2020). <http://dx.doi.org/10.18576/jsap/090212>

Oosterbaan, R.J. (2000). CumFreqA, free software program for distribution fitting with composite distribution options. Download from: <https://www.waterlog.info/cumfreq.htm>

Oosterbaan, R.J. (2021). 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. [https://www.waterlog.info/pdf/transform for linear regression.pdf](https://www.waterlog.info/pdf/transform%20for%20linear%20regression.pdf)

Oosterbaan, R.J. (2022). How to derive a probability distribution from a data set using the simple method of plotting positions and the free CumFreq model. [https://www.waterlog.info/pdf/Plotting positions.pdf](https://www.waterlog.info/pdf/Plotting%20positions.pdf)

8. Appendices (8A and 8B)

Appendix 8A. Screen-print of the CumFreqA input file showing composite probability distribution options.

The screenshot shows the CumFreqA software interface. The title bar reads "CumFreqA cumulative frequency analysis with emphasis on composite probability distributions". The menu bar includes "File" and "Edit". The "Input" tab is selected, showing the following fields:

- File: D:\Werkmappen\WinModels\CumFreq group\CumFreqA data used\Various data\normal+skew bimodal\normal bimodal Gu
- Title1: Bimodal twice symmetric
- Title2: (empty)
- Options: Allow a composite distribution, if it can be detected
- Nr. of Data: 85
- Nr of Intervals for histogram: 5
- Threshold (cut-off for data values): not used

A table displays the input data:

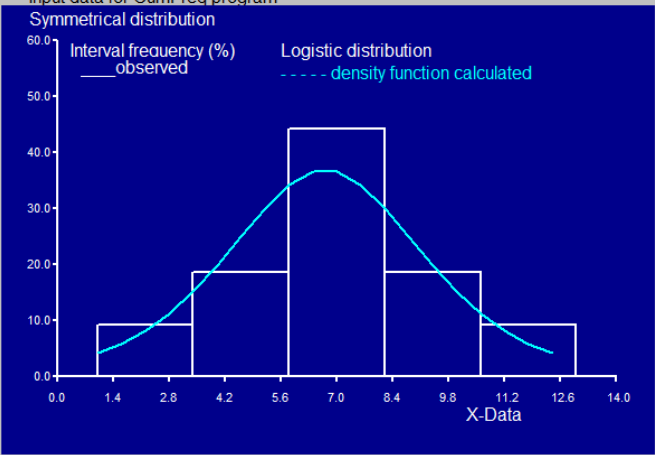
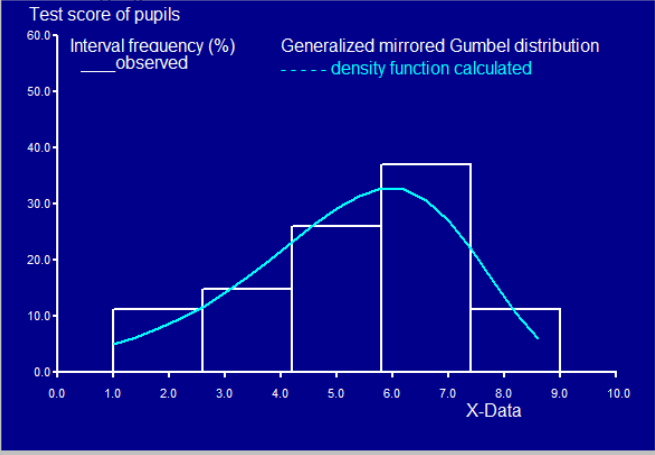
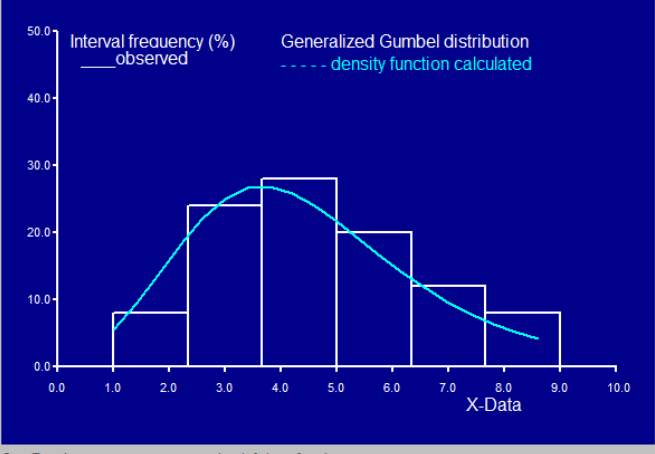
Serial Nr.	Data value
1	1
2	2
3	3
4	3
5	4
6	4
7	4
8	5
9	5
10	5
11	5
12	5
13	6
14	6
15	6

The "Select composite distribution" dropdown menu is open, showing the following options:

- Best fitting of all composite distributions
- Composite logistic distribution
- Composite Gumbel distribution
- Composite Gumbel mirrored
- Logistic + Gumbel
- Logistic + Poisson
- Logistic + Gumbel mirrored
- Gumbel + Logistic
- Gumbel + Poisson
- Gumbel + Gumbel mirrored
- Gumbel mirrored + logistic
- Gumbel mirrored + Gumbel
- Gumbel mirrored + Poisson
- The following use generalized distributions---
- Composite generalized logistic distribution
- Composite generalized Gumbel distribution
- Composite generalized Gumbel mirrored
- Composite generalized Poisson distribution
- Gen. logistic + gen. Gumbel
- Gen logistic + gen. Poisson
- Gen. Gumbel + gen. logistic
- Gen. Gumbel + gen. Poisson
- Gen. Poisson + gen. logistic
- Gen. Poisson + gen. Gumbel
- Generalized Laplace distribution

Buttons at the bottom include "Clear data", "Paste help", "Save-Run", and "Open input". A status bar at the bottom reads: "Enter data or use 'Open' to see examples under 'Data' or to edit existing files. Thereafter use 'Save-Run'."

Appendix 8B. Examples of unimodal symmetrical, left skew and right skew distributions.

<p>Input data for CumFreq program Symmetrical distribution</p>  <p>Interval frequency (%) — observed</p> <p>Logistic distribution ----- density function calculated</p> <p>X-Data</p> <p>CumFreqA program at www.waterlog.info/cumfreq.htm</p>	<p>Histogram of observed data plus best fit of probability density curve (PDC)</p> <p>Unimodal symmetrical</p>
<p>CumFreq program Test score of pupils</p>  <p>Interval frequency (%) — observed</p> <p>Generalized mirrored Gumbel distribution ----- density function calculated</p> <p>X-Data</p> <p>CumFreqA program at www.waterlog.info/cumfreq.htm</p>	<p>Histogram of observed data plus best fit of probability density curve (PDC)</p> <p>Unimodal left skew</p>
<p>Skew right</p>  <p>Interval frequency (%) — observed</p> <p>Generalized Gumbel distribution ----- density function calculated</p> <p>X-Data</p> <p>CumFreqA program at www.waterlog.info/cumfreq.htm</p>	<p>Histogram of observed data plus best fit of probability density curve (PDC)</p> <p>Unimodal right skew</p>