

**Program Reference Distribution Method**

**for Time Series**

**User Manual**

Tineke de Jonge

August 2017

Erasmus Happiness Economics Research Organization

Erasmus University Rotterdam

The Netherlands

tineke.dejonge@gmail.com

**Contents**

|  |  |
| --- | --- |
| 1. Introduction ……………………………………………………………….……………….... | 1 |
| 2. Requirements for the Program …………………………………………………………......... | 1 |
| 3. Preparation of the Input ……………………………………………………………………... | 1 |
| 3.1 Phase 1: Collection of the time series to be converted ………………………………….. | 1 |
| 3.2 Phase 1: Preparation of a Conversion Scheme ………………………………………….. | 2 |
| 3.3 Phase 2: Preparation of the Input Worksheets …………………………………………... | 4 |
| 4. Conversion of Time Series ………………………………………………………………….. | 5 |
| 4.1 Running the Program Reference Distribution Method ………………………………….. | 5 |
| 4.2 Conversion Results ……………………………………………………………………… | 7 |
| References ……………………………………………………………………………………... | 11 |
| Appendix A The Reference Distribution Method ……………………………………………... | 12 |
| A.1 Application of the Continuum Approach to Derive a Reference Distribution ………… | 12 |
| A.2 Illustration of an Application of the Reference Distribution Method …………………. | 13 |
| Appendix B Visual Basic Code of the Program ……………………………………………….. | 16 |
|  |  |
|  |  |

1. Introduction

The Reference Distribution Method is a method to make the responses to different survey questions on the same topic comparable. The method is based on the idea that, for a given year and a given population, the distribution means after scale transformation for similar questions about a given topic asked in different representative surveys should be approximately the same irrespective of the primary response scales used. In this method the boundaries between the response options are derived from a reference distribution. The method can be applied to combine time series from different surveys on the same topic which span different periods of time into one long time series and to bring the responses to survey questions on the same topic using different response scales to a comparable level De Jonge et al (2016), De Jonge et al (2017, Part IV) and Brulé and Maggino (2017).

In this manual we describe how the program ‘Program Reference Distribution Method for Time Series’ can be used to convert the mean values of time series of measurements to a level on the continuum from 0 to 10 that is comparable to the mean value of a reference distribution. Moreover, we will show how the method can be used to convert the ranks of the response options of a primary discrete scale to a mean interval value on the continuum from 0 to 10, which, for example, for a topic such as ‘satisfaction faction with life’, would result in the mean satisfaction of the ‘fairly satisfied’ or the ‘very satisfied’ on the 0 – 10 continuum. A description of the Reference Distribution Method is given in Appendix A. We advise you to read this description before you continue.

We have used the time series on life satisfaction for the Netherlands to illustrate the application of the program. The time series of these surveys span 1973 – 2015 during which life satisfaction has been measured in several different surveys.

2. Requirements for the Program

The program ‘Program Reference Distribution Method for Time Series’ is built in MS Excel and makes use of a VBA macro to process the assessments. The requirements for the program are

* MS Excel version 2010 or higher[[1]](#footnote-1) enabled for macro’s
* Input files in a pre-described format in Excel

The code of the VBA macro is given in Appendix B.

Note: be aware that the program is not protected. We advise you always to work with a copy of the program.

3. Preparation of the Input

**3.1 Phase 1: Collection of the time series to be converted**

The first phase of the preparation of the input consist of collection the time series of frequency distributions for a given topic that you want to convert. For example in the case of satisfaction with life as whole in The Netherlands, time series from the following surveys and for the following time periods are available:

* World Values Survey (WVS): 10-point numerical scale, 6 waves, 1981 - 2012
* Eurobarometer (EB): 4-point verbal scale, waves in 41 years, 1973 - 2015
* Statistics Netherlands (CBS): 5-point verbal scale, 26 waves in 17 years, 1974 - 2009
* The Netherlands Institute for Social Research (SCP): 5-point verbal scale, 8 waves, 1974 - 2001
* SCP: 10-point numerical scale, 3 waves, 2004 – 2008
* European Social Survey (ESS): 11-point numerical scale, 7 waves, 2002 – 2014

The frequency distribution over the response options for the valid response should be available for each wave of a time series. In addition to this, but optionally, it is required to have available:

* the total response in absolute numbers, including the invalid response
* the valid response in absolute numbers
* the percentage of non-response

If, for a given survey, there are multiple waves per year, than you have to decide whether you 1) want to include all individual waves in the conversion process, 2) want to average the frequency distributions of all waves in a year to have only one frequency distribution per year or 3) to pick out one wave to be included for each year. To give an example, take the case of the standard version of the EB which has, almost without exception, a spring wave and an autumn wave for every year. Here, for the illustrations given in this manual, we made the choice to average the frequency distributions of the EB per year. The time series of mean life satisfaction in The Netherlands before conversion are depicted in Fig. 1.

**Fig. 1 Mean life satisfaction in The Netherlands based on ranks options primary scale**



**3.2 Phase 2: Preparation of a Conversion Scheme**

The second phase of preparation consist of a thorough inspection of the available time series for discontinuities to be solved using the Reference Distribution Method which must result in a conversion scheme. In the CBS survey, for example, a major change took place in 1997 when among other things the survey mode was changed from paper-and-pencil surveying into face-to-face interviews. This change was introduced using a split half measurement with half of the respondents being required to fill in a paper-and-pencil questionnaire and the other half being interviewed. This mode change led to a severe discontinuity in the time series for life satisfaction, and to correct for this, both 1997 results will be used in the conversion process (DeJonge et al 2016, pp. 875-876; DeJonge et al 2017, pp. 90-91).

The verbal scale items used for measuring life satisfaction in the surveys of CBS and SCP did not change over time, however, the surveys themselves underwent some changes (De Jonge et al 2016, pp. 865-866). As a consequence, the time series for both surveys have to be split up for the conversion process in time series each covering a period in which the remained unchanged. Once you know whether it is needed and if so into how many parts each time series has to be split up, you are ready to make the conversion scheme. As an illustration, the conversion scheme for the Dutch time series is given in Fig. 2.

**Fig. 2 Conversion Scheme for Life Satisfaction in The Netherlands**



In the scheme shown in Fig. 2

* in the most left column the year of surveying is given.
* in the second column, it can be seen that the WVS was fielded in six years and that the 2006 wave will be used to estimate a best fit beta distribution that will serve as a primary reference distribution for the other survey items which were fielded in 2006 and which come from the surveys of the EB, CBS, SCP and ESS.
* in column three to six it can be seen in which years the EB, CBS, SCP and ESS surveys were fielded, with a vertical bar as short for at least one wave in every year in the period it covers. The primary reference distribution will be used to derive the boundaries[[2]](#footnote-2) between the response options of each of the scales of these based on the frequency distributions for these items in that year. A best fit beta distribution can be estimated for each other wave of these surveys using the boundaries derived from the primary reference distribution and the frequency distribution for that wave.
* in the next columns the split up of the time series for the CBS and SCP surveys is given. A secondary reference distribution has to be used for the conversion of each of these time series, which has to be chosen from the best fit beta distributions estimated for the time series in column two to six. For example, the best fit beta distribution estimated for the CBS item from the time series in column three and fielded in 1997, will be used as a secondary reference distribution to the derive boundaries between the response options of the response scales of the CBS item, before the mode change, and the SCP item based on frequency distributions of the 1997 waves. These boundaries on their turn, can be used to estimate best fit beta distributions for the waves of the SCP survey fielded in the period from 1997 to 2000 and the waves of the CBS survey fielded in the period from 1994 to 1997.

**3.3 Phase 3: Preparation of the Input Worksheets**

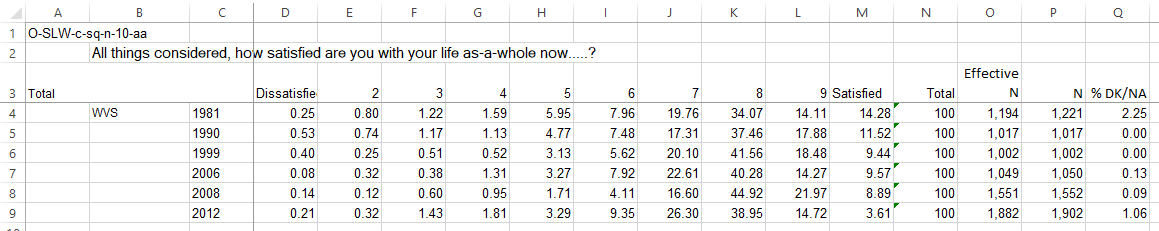
Given the conversion scheme you have made, you can make the worksheets with input for the conversion process. You have to make one worksheet for each column specified in the conversion scheme, which has to meet the requirements given in Table 1.

**Table 1 Requirements input worksheets Reference Distribution Method**

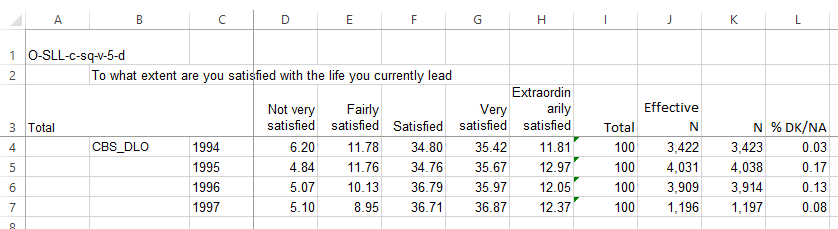
|  |  |  |
| --- | --- | --- |
| **Position** | **Description** | **Obligatory** |
| Cell(1,”A”) | Code of the survey item | N |
| Cell(2,”B”) | Leading question of the survey item | N |
| Cell(3, “A”) to Cells(3,”C”) | Description of the target population of the survey | N |
| Columns(“A:B”), row 4 and below | Explanatory note for each wave in which the item was used | N |
| Column(“C”), row 4 and below | Year of surveying | Y |
| Cell(3,”D”) to Cell(3,3+number of response options) | Labels of response options | Y |
| Cell(3,4 + number of response options) | Keyword “Total” | Y |
| Cell(3,4 + number of response options + 1) | Header for valid response | N |
| Cell(3,4 + number of response options + 2) | Header for total response | N |
| Cell(3,4 + number of response options + 3) | Header for non-response and invalid response | N |
| Cell(4,”D”) to Cell(3+number of waves, 3 + number of response options) | Frequency distributions | Y |
| Column(4 + number of response options), row 4 and below | Sum of frequencies, must be equal to 100 | N |
| Column(4 + number of response options + 1), row 4 and below | Total response in absolute numbers | N |
| Column(4 + number of response options + 2), row 4 and below | Valid response in absolute numbers | N |
| Column(4 + number of response options + 2), row 4 and below | Percentage non-response and invalid response | N |

All input worksheets have to be part of the same workbook. It is advisable to give each worksheet a meaningful name for. Two examples of input worksheets are given in Figs. 3 and 4. In Fig. 3 the input for life satisfaction in The Netherlands measured with the WVS is given. In Fig. 4 you can see the input for life satisfaction measured with the CBS survey in the period from 1994 to 1997.

**Fig. 3 Input worksheet life satisfaction WVS 1981 - 2012**



**Fig. 4 Input worksheet life satisfaction CBS 1994 - 1997**

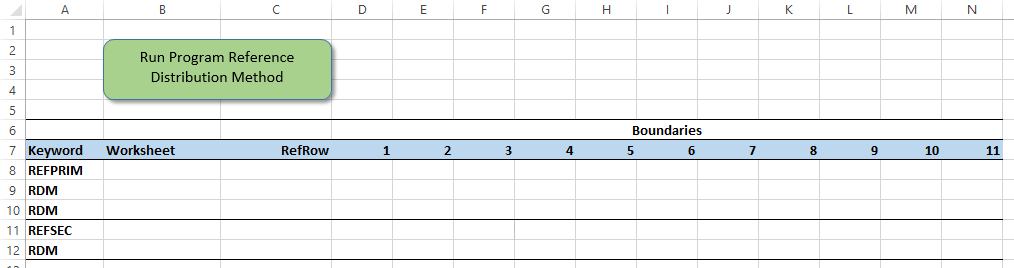


4. Conversion of time series

**4.1 Running the Program Reference Distribution Method**

When you open the “Program Reference Distribution Method” in Excel, you will see the worksheet ‘Input’ as shown in Fig. 5 and a button to start the VBA Macro given in Appendix B.

**Fig. 5 Worksheet ‘Input’ of the “Program Reference Distribution Method”**

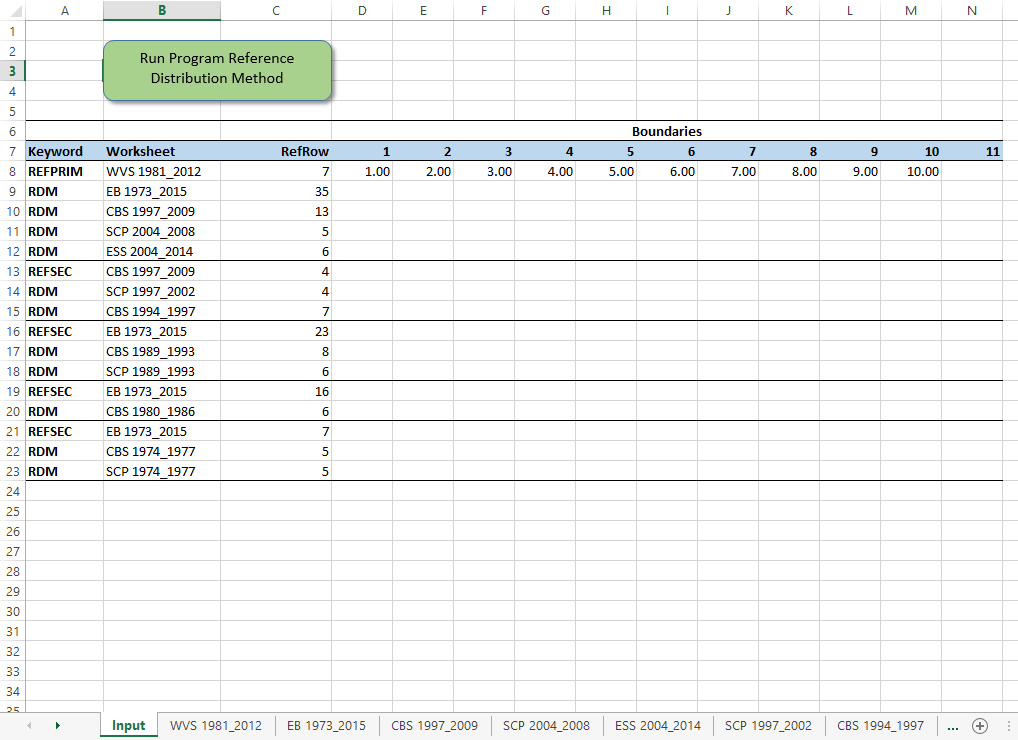


In Fig. 5 you can see the basics of the program for the input. You should add all the input worksheets you have prepared to the program workbook and subsequently edit the contents of the worksheet ‘Input’ on a basis of the conversion scheme and the contents of each input worksheet. The input makes use of three keywords:

* REFPRIM: indication that a primary reference distribution has to be estimated, which must only be used in row 8 which specifies the time series that will serve as a reference for the entire conversion process
* REFSEC: indication that a secondary reference distribution has to be estimated
* RDM: indication that a time series has to be converted using the last estimated reference distribution as a reference

After the keyword the worksheet with the time series the keyword refers to should be given, followed by the number of the row in this worksheet which contains the frequency distribution bridging the step from one time series to another time series. Only in the case of the keyword ‘REFPRIM’ do the boundaries between the response options of the verbal scale that will be used to estimate the primary reference distribution have to be specified. To explicate the editing of the contents of the worksheet ‘Input’ we used the time series on life satisfaction in The Netherlands. In Fig. 6 you can see that the input worksheets for these time series, which were made according to the requirements in Table 1, are added to the program workbook and that the worksheet ‘Input’ has been edited in line with this.

**Fig. 6 Worksheet ‘Input’ for the conversion of time series on life satisfaction in The Netherlands**



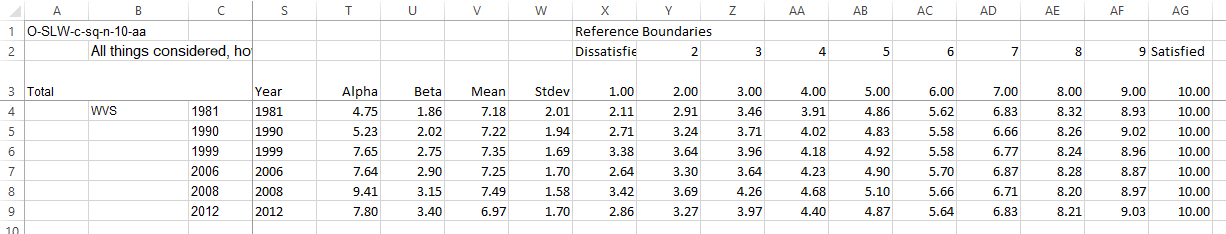
1. In row 8 it is specified that a primary reference distribution has to be estimated based on the frequency distribution given in row 7 of the worksheet ‘WVS 1981\_2012’, under the assumption that the boundaries between the response options of the 10-point numerical scale of the WVS are equidistantly distributed over the continuum from 0 to 10. In Fig. 3 it can be seen that row 7 in the worksheet for the WVS contains the frequency distribution for 2006. The keyword ‘REFPRIM” will also guarantee that a beta distribution will be estimated for the frequency distributions from other waves of the WVS and equidistant boundaries between the response options.
2. In rows 9 to 12 it is specified that the boundaries between the response options of the items taken from the EB, CBS, SCP and ESS have to be derived from the primary reference distribution given the frequency distribution in the specified reference rows. The keyword ‘RDM’ also indicates that with these boundaries beta distributions have to be estimated for all frequency distributions specified in the worksheets mentioned in column “B” of the rows 9 to 12.
3. According to the conversion scheme in Fig. 2, the beta distribution estimated for the CBS frequency distribution in 1997 and the boundaries derived from the primary reference distribution, has to be used as a secondary reference distribution for the time series from SCP covering the period from 1997 to 2000 and from CBS covering the period from 1994 to 1997. This is specified in rows 13 to 15.
4. Likewise secondary reference distributions are used to converted the time series of CBS and SCP covering earlier periods in time as is specified in rows 16 to 22.

With the input worksheets added to the program workbook and the specification of contents of the worksheet ‘Input’, you can start the conversion process by pressing the button ‘Run Program Reference Distribution Method’. Depending on the number of frequency distributions to be converted this may take a few minutes.

**4.2 Conversion Results**

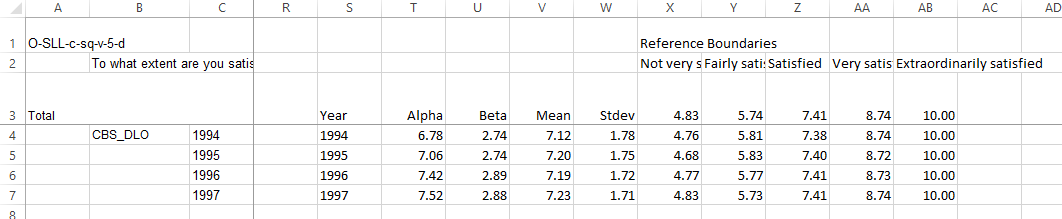
The results of the conversion will be added to each input worksheets. The results contain three blocks of information, which we will discuss below. The first part of the results, consists for each wave of a time series, of the parameters α and β of the best fit beta distribution, followed by the estimated population mean and standard deviation in eqs. (1) and (2) of Appendix A and the points at the 0-10 continuum where the values of the best fit beta distribution are equal to the cumulative frequencies on the primary verbal scales. This is illustrated in Figs. 7 and 8 for the WVS times series for the period 1981 - 2012 and the CBS time series for the period 1994 - 1997.

**Fig. 7 Conversion results for the WVS time series 1981 - 2012, Part 1**



The values of the points where the beta distribution is equal to the cumulative frequencies on the primary verbal scale may differ considerably for the WVS, since the boundaries between the response options of the WVS were assumed to be equidistant, as can be seen in Cell(3,”X”) to Cell(3, “AG”), which may not be realistic. This is not the case for the CBS time series, as can be seen in Fig. 8.

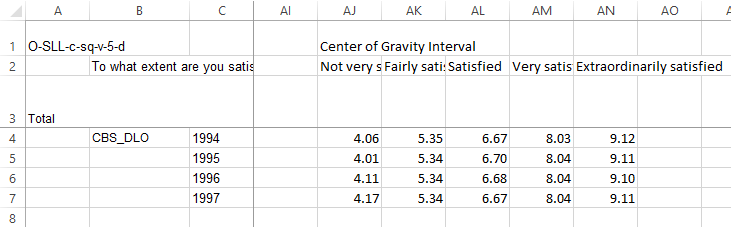
**Fig. 8 Conversion results for the CBS time series 1994 - 1997, Part 1**



In row 3 shown in Fig. 8 the reference boundaries are given that have be derived for the CBS item for the period 1994 – 1997 from the secondary reference distribution based on the results for the 1997 wave of the CBS item in the time series for the period 1997 - 2009. These reference boundaries are equal to 4.83, 5.74, 7.41, 8.74 and 10.00 which do not differ much from the transition points for the response options given in column “X” to “AB” for the individual waves of the CBS item in the period 1994 – 1997.

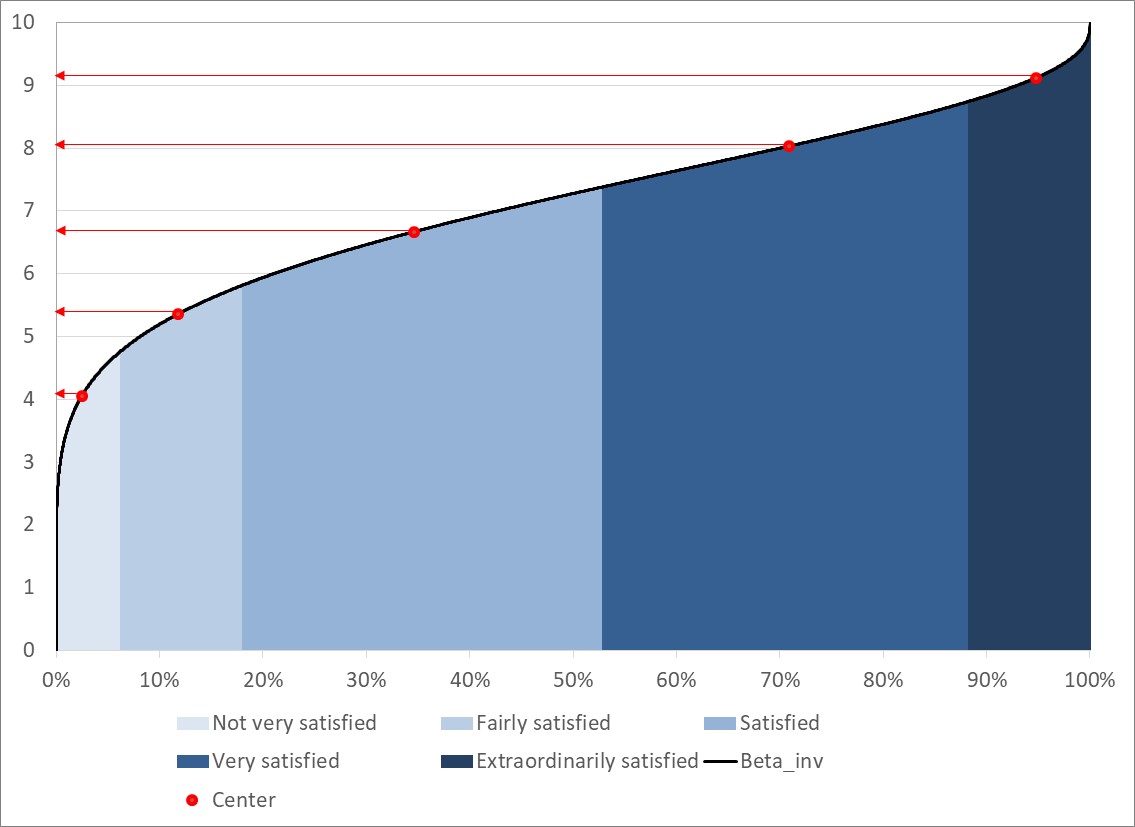
The second part of the conversion results, consists of the centers of gravity estimated for each response option in a given wave. For a given wave, the center of gravity, which can be considered to be the center of the interval on the 0 - 10 continuum for a given response option, is obtained by first calculating the value of the inverse beta distribution for the 10.000 equidistant points in between 0% to 100% specified by the parameters α and β estimated for that wave. Next, the center of gravity for the response options follows from dividing the sum of all values of these inverse beta distributions which are in the interval between the lower and upper boundary for the response option by the number of equidistant points that meet this criterion. This is illustrated in Fig. 9 for the CBS times series for the period 1994 – 1997.

**Fig. 9 Conversion results for the CBS time series 1994 - 1997, Part 2: Centers of gravity response options**



From Fig. 8 it can be seen that the interval between 5.81 and 7.38 represents the response option ‘Satisfied’ in 1994. The center of gravity of this interval is equal to 6.67 as can be seen in Fig. 9. A graphical illustration of the centers of gravity for the CBS item in 1994 is given in Fig. 10. The black line in this figure is the inverse beta distribution with parameters α = 6.78 and β = 2.74. The 0 – 10 continuum is represented by the vertical axis and the colored faces below the inverse beta distribution mark the areas corresponding to the five response options. The dots on the inverse beta distributions mark the centers of gravity, the value of which is indicated by an arrow pointing to the vertical axis.

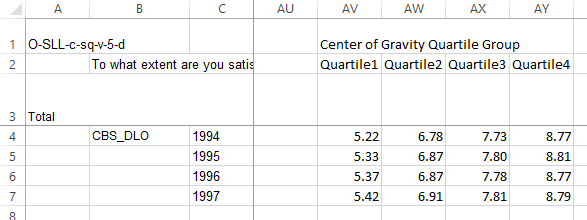
**Fig. 10 Centers of gravity for response options CBS item 1994**



The centers of gravity can be seen as the conversion of the rank of a response option to a numerical value on the continuum from 0 to 10.

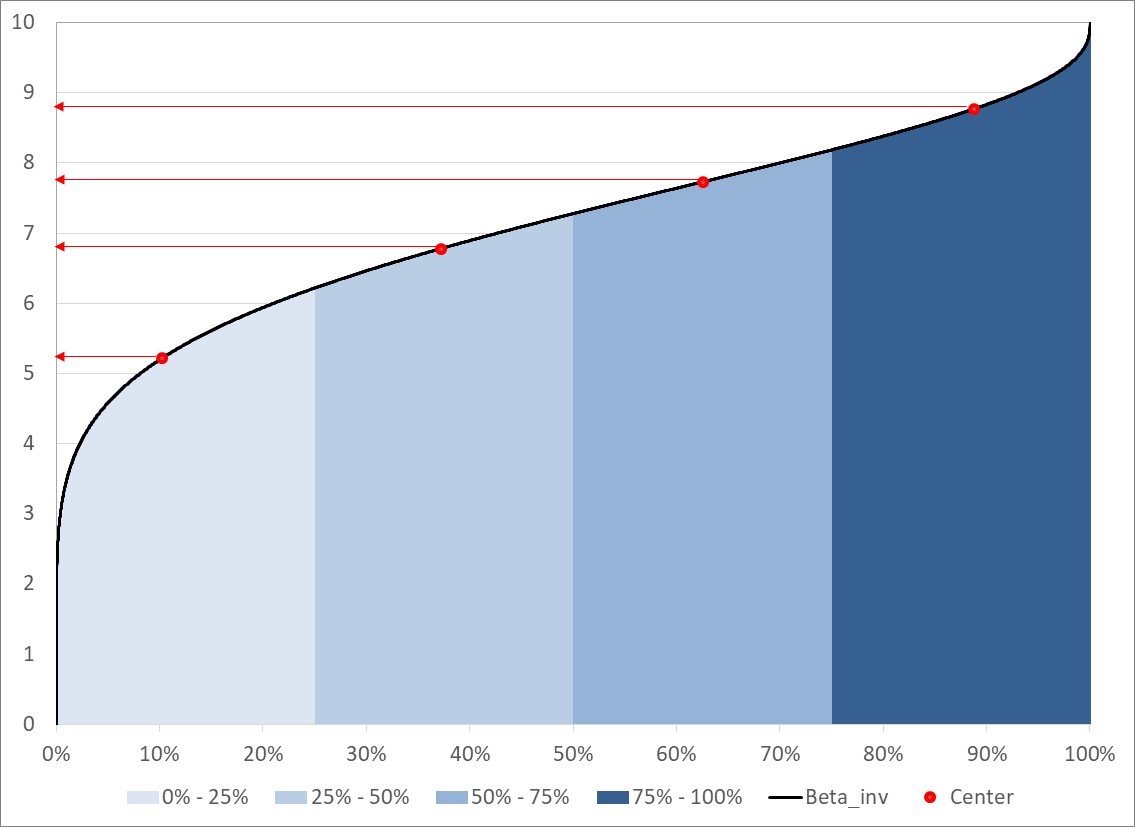
The third part of the conversion results, consists of the centers of gravity estimated for quartile groups of the population. In the case of life satisfaction for example, these centers of gravity are an estimate of the average life satisfaction on the continuum from 0 to 10 for each quartile group. These values are given in Fig. 11 for the CBS time series 1994 – 1997.

**Fig. 11 Conversion results for the CBS time series 1994 - 1997, Part 3: Centers of gravity quartile groups of the population**



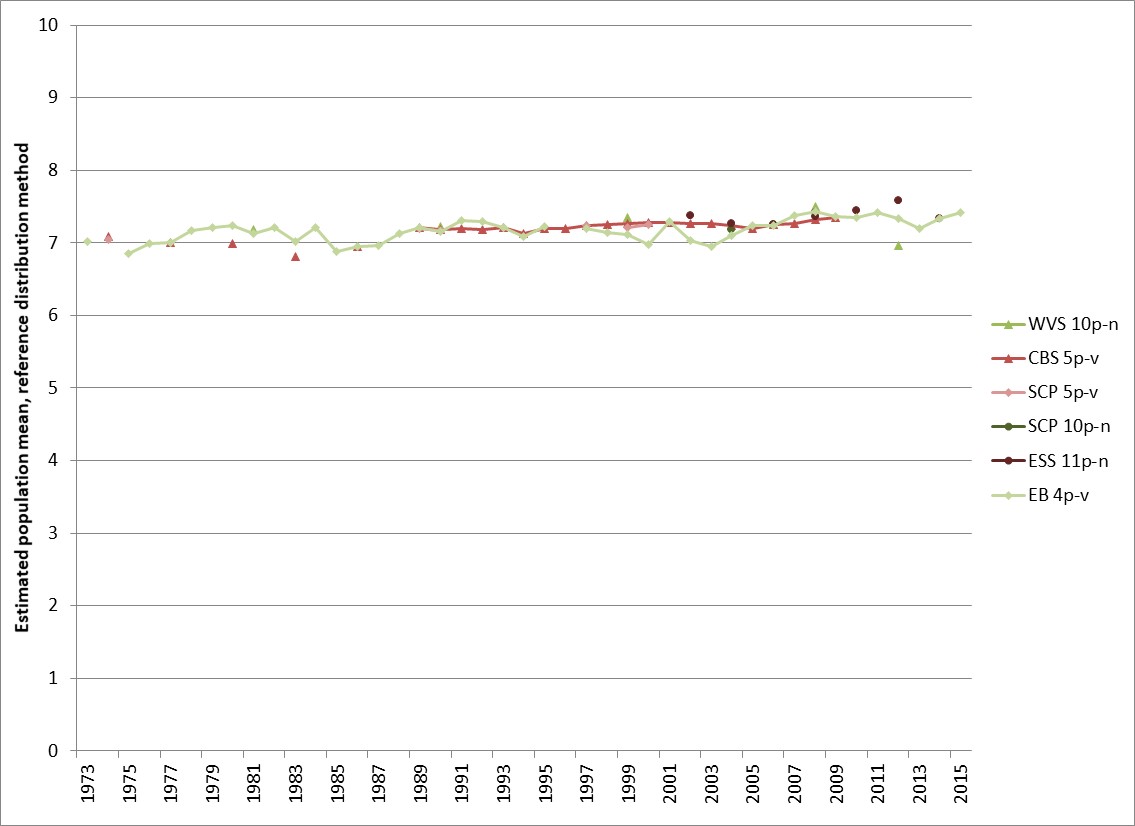
In Fig. 11 it can be seen that estimated value of satisfaction with life is on average 5.22 for the subgroup of the 25% of people in the population who were least satisfied in 1994. The estimate for the average life satisfaction of the 25% of people in the population who were most satisfied is 8.77. A graphical representation for the average life satisfaction per quartile group of the Dutch population in 1994 is given in Fig. 12.

**Fig. 12 Centers of gravity for quartile groups of the population, CBS item 1994**



To finalize, the time series of mean life satisfaction in The Netherlands after conversion are depicted in Fig. 13.

**Fig. 13 Mean life satisfaction in The Netherlands after conversion apply the Reference Distribution Method**



**References**

Brulé, G. & Maggino, F. (Eds.) (2018). *Metrics of Subjective Well-Being: Limits and Improvements*.

Springer. Series: Happiness Studies Book Series (in press). doi: 10.1007/978-3-319-61810-4\_12

De Jonge, T., Veenhoven, R., Kalmijn, W. M., & Arends, L. R. (2016). *Pooling time series based*

*on slightly different questions about the same topic. Forty years of survey research on*

*happiness and life satisfaction in The Netherlands*. Social Indicators Research, 126,

863–891. doi:10.1007/s11205-015-0898-5

De Jonge, T. Veenhoven. R. & Kalmijn. W. (2017). *Diversity in Survey Questions on the Same Topic*. Springer. Social Indicators Research Series. Vol. 68. doi: 10.1007/978-3-319-53261-5

Kalmijn, W.M. (2010). *Quantification of Happiness Inequality*. Dissertation. Erasmus University Rotterdam. The Netherlands. Enschede: Ipskamp Drukkers. Available a

Kalmijn, W.M. Arends. L.R.. & Veenhoven. R. (2011). Happiness Scale Interval Study. Methodological Considerations*. Social Indicators Research* 102(3). 497-515. DOI 10.1007-s11205-010-9688-2.

Kalmijn, W. M. (2013*). From discrete 1 to 10 towards continuous 0 to 10: The continuum*

*approach to estimating the distribution of happiness in a nation*. Social Indicators Research,

110(2), 549–557. doi:10.1007/s11205-011-9943-1.

**Appendix A The Reference Distribution Method**

**A.1 Application of the Continuum Approach to Derive a Reference Distribution**

The Reference Distribution Method is a method which relies heavily on the Continuum Approach (Kalmijn 2010; De Jonge et al 2016; De Jonge et al 2017, Ch.7). The main assumption of the Continuum Approach is that variables are continuously distributed in a population. A beta distribution is the most appropriate to use in the Continuum Approach due to at least three interesting properties it has (Kalmijn, Arends, and Veenhoven, 2011, pp. 509-510):

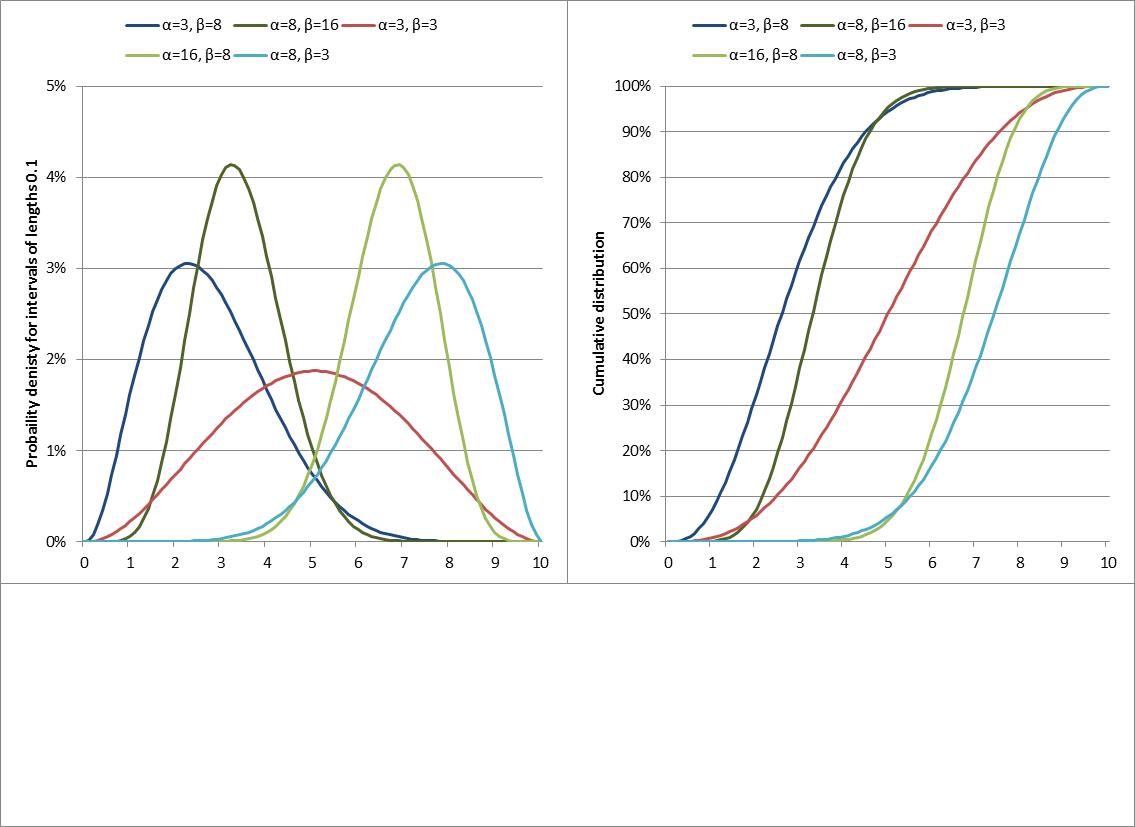
1. it is a continuous distribution, which makes it suitable as a model for the continuous latent happiness variable in the population
2. the random variable has a two-sided bounded domain, which makes it suitable for happiness as it is measured using two-sided bounded primary scales
3. the distribution has two shape parameters, which makes beta distributions cover a wide class of different distribution shapes, including skew distributions, both positive and negative

The family of beta distributions consists of a series of distributions, each member of which is characterized by two shape parameters, *α* and *β*. The mean *μ* of a beta distribution with parameters *α* and *β* on the continuum from 0 to 10 and the standard deviation *σ* are equal to:

(1) (2)

Some examples of the probability density functions and the cumulative distribution functions for different values of *α* and *β* are given in Fig. A.1.

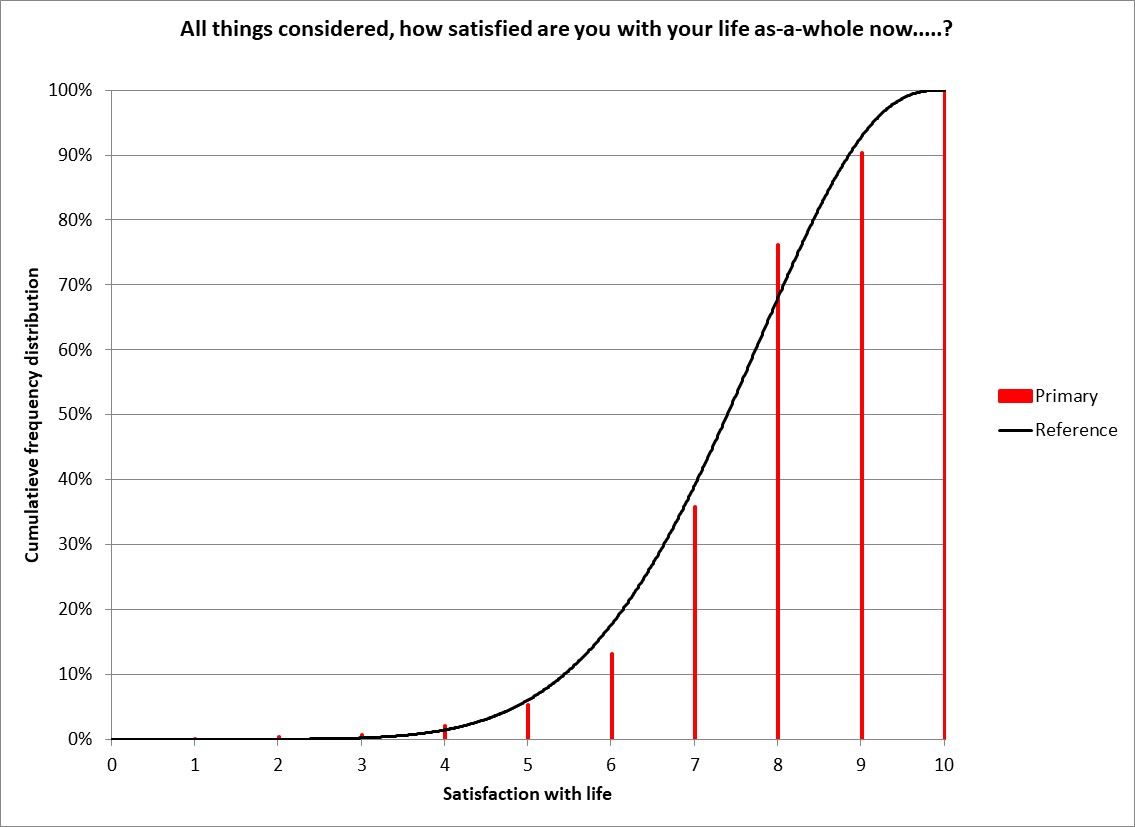
**Fig. A.1 Examples of the probability density β-functions and cumulative distribution β-functions**



A starting point for the Continuum Approach is provided by the cumulative frequencies of a measurement on a discrete primary scale and the values on a continuum from 0 to 10 at which respondents change their judgment from one to the adjacent response option on this primary scale, for example from ‘fairly satisfied’ to ‘satisfied’. In the Continuum Approach the shape parameters *α* and *β* of the beta distribution that best fits the cumulative frequencies and the values on the continuum of the boundaries between the response options of the primary scale are estimated as maximum likelihood estimators. This estimation procedure is described into more detail in Kalmijn (2010, pp. 160-162).

If the Continuum Approach is used to derive a reference distribution based on survey results measured on a discrete scale, this scale should preferably be numerical with ten to eleven response options. In Fig. A.2 we illustrate the use of the Continuum Approach to derive a reference distribution from responses to the question: ‘All things considered, how satisfied are you with your life as-a-whole now.....?’ using an 10-point numerical scale from 0 to 10 with the anchor points of the scale labelled by ‘Dissatisfied’ and ‘Satisfied’. This item was taken from the World Values Survey[[3]](#footnote-3). We fixed 10 equidistant upper boundaries, one for each response option, starting at 1.0 for the response option at the lower end of the scale and ending at 10.0 for the option at the upper end of the scale. The assumption that these boundaries between the response options are equidistantly distributed is a methodological choice (Kalmijn, 2013) which provides a very useful basis for the Reference Distribution Method.

**Fig. A.2 Best fitting beta distribution to the WVS frequency distribution 2006**



The parameters of the beta distribution in Fig. A.2 are *α* = 7.64 and *β* = 2.90, which, according to Eq. 1 corresponds to a mean of 7.24.

The curve in Fig. A.2 is the beta distribution that according to the Continuum Approach fits best to the boundaries and cumulative frequencies distribution of the WVS item in 2006.

**A.2 Illustration of an Application of the Reference Distribution Method**

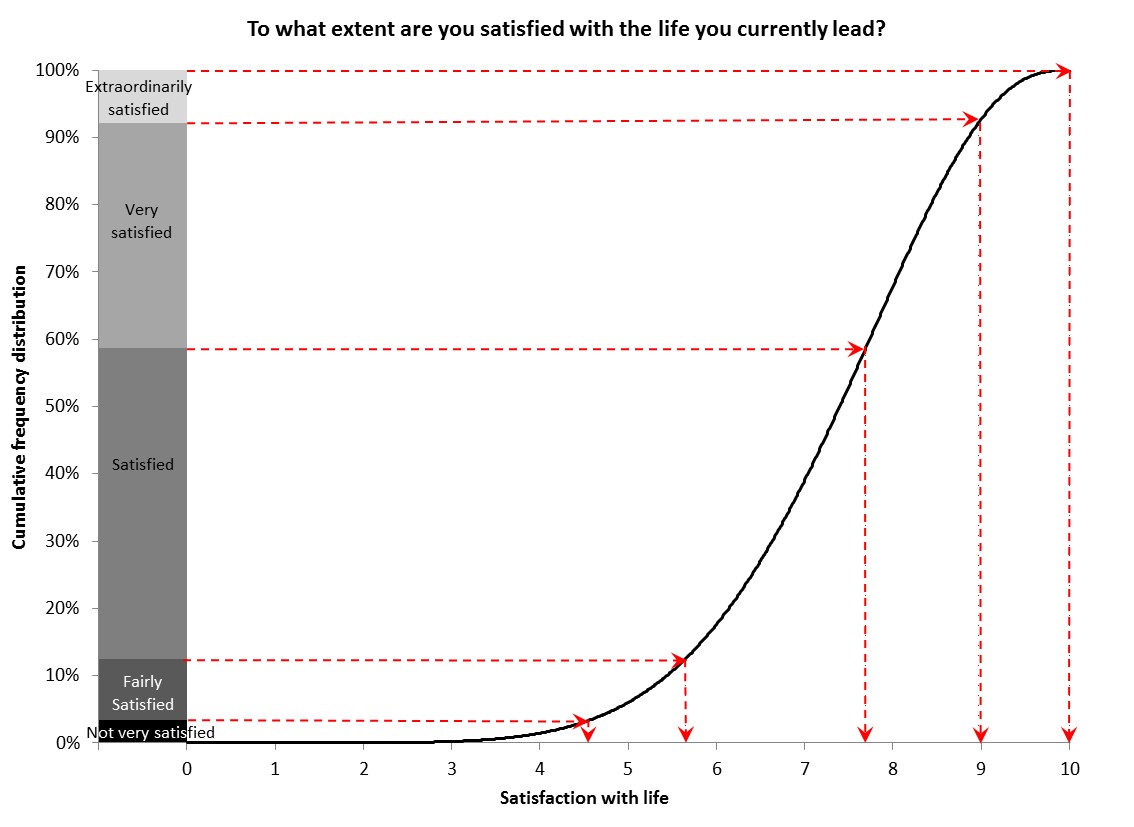
We will use the best fit beta distribution based on the WVS item as a reference for the illustration of the application of the Reference Distribution Method to the life satisfaction item from a survey of Statistics Netherlands. The frequency distribution of the responses to this item in 2006 in The Netherlands is:

* Extraordinarily satisfied 7.8%
* Very satisfied 33.5%
* Satisfied 46.3%
* Fairly satisfied 9.1%
* Not very satisfied 3.4%

Using the Reference Distribution Method, the procedure to determine the reference boundaries between the response options of the CBS item on the continuum from 0 to 10 is as follows, see also Fig. A.3.

* We start with the cumulative frequency distribution of the CBS item for which we want to determine where the boundaries between the response options are positioned on the continuum from 0 to 10.This cumulative frequency distribution is shown as a stacked bar on the left side of Fig. A.3.
* The reference distribution derived from the WVS is depicted to the right side of this stacked bar, plotted against the 0 to 10 continuum which is represented by the horizontal axis.
* A horizontal line is drawn from the cumulative frequency displayed in the stacked bar on the left side of Fig. A.3 for each response option of the response scale, to the point where it touches the reference distribution. At this point the value of the reference distribution is equal to the cumulative distribution on the scale of the CBS item.
* A vertical line is drawn from this latter point to the 0 to 10 continuum on the horizontal axis. The value at which the vertical line touches the horizontal axis is the position of the reference boundary of the corresponding response option.

**Fig. A.3 Illustration of the Reference Distribution Method**



Following this procedure, the reference boundaries for the response options of the CBS item on life satisfaction on the 0 to 10 continuum are, consecutively, 4.6, 5.6, 7.7, 9.0 and 10.0, and, given the reference boundaries at these positions, the reference distribution perfectly fits the cumulative frequency distribution of the CBS item in 2006. The mean 7.2 of the reference distribution derived from the WVS is thus also an estimate of the population mean for 2006 wave of the CBS item.

Reference boundaries found in this way can in their turn be used as input for an application of the Continuum Approach to the cumulative frequencies of the CBS item obtained in other waves. The estimated mean on the 0-10 continuum for each of these waves is equal to the mean of the corresponding best fit beta distribution resulting from the application of the Continuum Approach.

**Appendix B** **Visual Basic Code of the Program**

Public r As Integer 'Number of response options

Public B(0 To 11) As Double 'Boundaries between response options

Public f(1 To 11) As Double 'Frequency distribution

Public fc(1 To 11) As Double 'Cumulative frequency distribution

Public Alpha, Beta 'Parameters beta distribution

Public Ref\_Alpha, Ref\_Beta 'Parameters reference beta distribution

'---------------------------------------------------------------------------------------------

Public Sub MLE()

Dim i, j, k, n As Integer 'Counters

Dim timeseries(1 To 1000, 1 To 11) As Double 'Frequency distributions time series

Dim row\_ref As Integer 'Row from time series with reference information

Dim WS\_Name As String

Application.ScreenUpdating = False

Application.DisplayAlerts = False

'READ INPUT

i = 8

Worksheets("Input").Select

B(0) = 0 'Lower boundary response option ranked 1

While Not IsEmpty(Cells(i, 1))

'MAXIMUM LIKELIHOOD ESTIMATION PARAMETERS BEST FIT BETA DISTRIBUTION TO A PRIMARY SET

‘OF BOUNDARIES AND PRIMARY FREQUENCY DISTRIBUTION TO SERVE AS PRIMAY REFERENCE

‘DISTRIBUTION

If Cells(i, 1).Value = "REFPRIM" Then 'Read frequencies and boundaries and estimate beta distribution

For j = 1 To 11

B(j) = 0

f(j) = 0

Next j

r = 0

j = 1

While Not IsEmpty(Cells(i, j + 3))

B(j) = Cells(i, j + 3).Value

r = r + 1

j = j + 1

Wend

row\_ref = Cells(i, "C").Value

Worksheets(Cells(i, "B").Value).Select

For j = 1 To r

f(j) = Cells(row\_ref, j + 3).Value / 100

Next j

'Start ML\_Estimation beta distribution

Call ML\_Estimation

Ref\_Alpha = Alpha

Ref\_Beta = Beta

Call Time\_Series

End If

'APPLICATION OF THE REFERENCE DISTRIBUTION METHOD FOR A SPECIFIED TIME SERIES AND THE

‘LAST ESTIMATED BEST FIT BETA DISTRIBUTION

If Cells(i, 1).Value = "RDM" Then 'Read frequencies and derive boundaries and

'Estimate best fit beta distributions for time series

row\_ref = Cells(i, "C").Value

WS\_Name = Cells(i, "B").Value

Worksheets(WS\_Name).Select

'Calculation of cumulative distribution

f(1) = Cells(row\_ref, "D").Value / 100

B(1) = Excel.WorksheetFunction.Beta\_Inv(f(1), Ref\_Alpha, Ref\_Beta, 0, 10)

r = 0

While Not Cells(3, r + 4) = "Total"

r = r + 1

If r > 11 Then

MsgBox "Keyword 'Total' not found in worksheet " & WS\_Name, vbOKOnly

Exit Sub

End If

Wend

For j = 2 To r - 1

f(j) = f(j - 1) + Cells(row\_ref, 3 + j).Value / 100

B(j) = Excel.WorksheetFunction.Beta\_Inv(f(j), Ref\_Alpha, Ref\_Beta, 0, 10)

Next j

B(r) = 10

Call Time\_Series

End If

'MAXIMUM LIKELIHOOD ESTIMATION PARAMETERS BEST FIT BETA DISTRIBUTION TO A SECOND SET

‘OF BOUNDARIES AND PRIMARY FREQUENCY DISTRIBUTION TO SERVE AS SECONDARY REFERENCE

‘DISTRIBUTION

If Cells(i, 1).Value = "REFSEC" Then 'Read frequencies and boundaries and estimate beta distribution

row\_ref = Cells(i, "C").Value

Worksheets(Cells(i, "B").Value).Select

Ref\_Alpha = Cells(row\_ref, "T")

Ref\_Beta = Cells(row\_ref, "U")

Worksheets("Input").Select

End If

i = i + 1

Wend

Cells(3, "B").Select

Application.DisplayAlerts = True

Application.ScreenUpdating = True

End Sub

'---------------------------------------------------------------------------------------------

Public Sub ML\_Estimation()

Dim i, j, k, m, n As Integer 'Counters

Dim Mean\_V As Double 'Mean value Weighted Average Method mid-interval values

Dim Var\_V As Double 'Variance Weighted Average Method mid-interval values

Dim ADum(1 To 9) 'Alpha dummies for iteration

Dim BDum(1 To 9) 'Beta dummies for iteration

Dim Alpha\_it(1 To 25) 'Estimators alpha

Dim Beta\_it(1 To 25) 'Estimators beta

Dim logL(0 To 9) 'Log likelihood estimators, LogL(0) is minimum of LogL(1) to LogL(9)

Dim Stepsize As Single 'Step size log likelihood estimation

Dim Centre\_V As Integer 'Centre value

Mean\_V = 0

Var\_V = 0

For j = 1 To r

If j = 1 Then

Mean\_V = Mean\_V + f(j) \* B(j) / 2

Else

Mean\_V = Mean\_V + f(j) \* (B(j - 1) + (B(j) - B(j - 1)) / 2)

End If

Next j

For j = 1 To r

If j = 1 Then

Var\_V = f(j) \* (B(j) / 2 - Mean\_V) ^ 2

Else

Var\_V = Var\_V + f(j) \* ((B(j - 1) + (B(j) - B(j - 1)) / 2) - Mean\_V) ^ 2

End If

Next j

'Rescaling from range 0 - 10 to range 0 - 1

Mean\_V = Mean\_V / 10

Var\_V = Var\_V / 100

'Start value alpha and beta

Alpha\_it(1) = Mean\_V \* (Mean\_V \* (1 - Mean\_V) / Var\_V - 1)

Beta\_it(1) = Alpha\_it(1) \* (1 - Mean\_V) / Mean\_V

For n = 1 To 25

'Set stepsize. Size differs per iteration step

If n = 1 Then Stepsize = 1.28

If n > 1 And Centre\_V = 1 Then Stepsize = Stepsize / 2

For j = 1 To 9

Centre\_V = 0

ADum(1) = Alpha\_it(n)

If ADum(1) - Stepsize > 0.001 Then ADum(2) = ADum(1) - Stepsize Else ADum(2) = 0.001

ADum(3) = ADum(2)

ADum(4) = ADum(2)

ADum(5) = ADum(1)

ADum(6) = ADum(1)

ADum(7) = ADum(1) + Stepsize

ADum(8) = ADum(7)

ADum(9) = ADum(7)

BDum(1) = Beta\_it(n)

If BDum(1) - Stepsize > 0.001 Then BDum(2) = BDum(1) - Stepsize Else BDum(2) = 0.001

BDum(3) = BDum(1)

BDum(4) = BDum(1) + Stepsize

BDum(5) = BDum(2)

BDum(6) = BDum(4)

BDum(7) = BDum(2)

BDum(8) = BDum(1)

BDum(9) = BDum(4)

logL(j) = 0

For m = 1 To r

logL(j) = logL(j) + 1000 \* f(m) \* (-Excel.WorksheetFunction.Log \_

(Excel.WorksheetFunction.BetaDist(B(m), ADum(j), BDum(j), 0, 10) - \_

Excel.WorksheetFunction.BetaDist(B(m - 1), ADum(j), BDum(j), 0, 10)))

Next m

'Determination of minimum value of LogL

If j = 1 Then

logL(0) = logL(1)

End If

If j > 1 And logL(j) < logL(0) Then

logL(0) = logL(j)

End If

Next j

'Selection of alpha and beta next iteration step

If n < 25 Then

For j = 2 To 7

If logL(9 - j) = logL(0) Then

Alpha\_it(n + 1) = ADum(9 - j)

Beta\_it(n + 1) = BDum(9 - j)

Centre\_V = 9 - j

End If

Next j

For j = 8 To 9

If logL(17 - j) = logL(0) Then

Alpha\_it(n + 1) = ADum(17 - j)

Beta\_it(n + 1) = BDum(17 - j)

Centre\_V = 17 - j

End If

Next j

If logL(1) = logL(0) Then

Alpha\_it(n + 1) = ADum(1)

Beta\_it(n + 1) = BDum(1)

Centre\_V = 1

End If

End If

Next n

For j = 2 To 7

If logL(9 - j) = logL(0) Then

Alpha = ADum(9 - j)

Beta = BDum(9 - j)

Centre\_V = 9 - j

End If

Next j

For j = 8 To 9

If logL(17 - j) = logL(0) Then

Alpha = ADum(17 - j)

Beta = BDum(17 - j)

Centre\_V = 17 - j

End If

Next j

If logL(1) = logL(0) Then

Alpha = ADum(1)

Beta = BDum(1)

Centre\_V = 1

End If

End Sub

'---------------------------------------------------------------------------------------------

Public Sub Time\_Series()

Dim i, j, ts As Integer 'Counters

Dim Boundary(1 To 11) As Double 'Boundaries between response options per wave

Dim CGInt(1 To 11), CGQrt(1 To 4) As Double

Dim BetaInvDum As Double

Dim freq As Double

Columns("S:AZ").Clear

Cells(1, "X").Value = "Reference Boundaries"

Cells(1, "AJ").Value = "Center of Gravity Interval"

Cells(1, "AV").Value = "Center of Gravity Quartile Group"

For j = 1 To r

Cells(2, 23 + j).Value = Cells(3, 3 + j).Value

Cells(2, 35 + j).Value = Cells(3, 3 + j).Value

Cells(3, 23 + j).Value = B(j)

Next j

For j = 1 To 4

Cells(2, 47 + j).Value = "Quartile" & j

Next j

Cells(3, "S").Value = "Year"

Cells(3, "T").Value = "Alpha"

Cells(3, "U").Value = "Beta"

Cells(3, "V").Value = "Mean"

Cells(3, "W").Value = "Stdev"

ts = 0 'Number of frequency distributions in time series

While Not IsEmpty(Cells(ts + 4, "D"))

Cells(ts + 4, "S").Value = Cells(ts + 4, "C").Value

ts = ts + 1

Wend

For k = 1 To ts

For j = 1 To r

f(j) = 0

f(j) = Cells(k + 3, 3 + j).Value / 100

'Calculation of cumulative frequencies

If j = 1 Then

fc(j) = f(j)

Else

fc(j) = fc(j - 1) + f(j)

End If

Next j

Call ML\_Estimation

'ESTIMATE PARAMETERS BEST FIT BETA DISTRIBUTION

Cells(k + 3, "T").Value = Alpha

Cells(k + 3, "U").Value = Beta

'ESTIMATE POPULATION MEAN

Cells(k + 3, "V").Value = 10 \* Alpha / (Alpha + Beta)

'ESTIMATE STANDARD DEVIATION

Cells(k + 3, "W").Value = ((Cells(k + 3, "V").Value \* (10 - Cells(k + 3, "V").Value) / \_

(10 \* Alpha / Cells(k + 3, "V").Value) + 1)) ^ 0.5

'ESTIMATE POSITION BOUNDARIES BETWEEN RESPONSE OPTIONS

Boundary(1) = Excel.WorksheetFunction.Beta\_Inv(f(1), Alpha, Beta, 0, 10)

For j = 2 To r - 1

Boundary(j) = Excel.WorksheetFunction.Beta\_Inv(fc(j), Alpha, Beta, 0, 10)

Next j

Boundary(r) = 10

For j = 1 To r

Cells(k + 3, 23 + j) = Boundary(j)

Next j

'ESTIMATE CENTERS OF GRAVITY FOR INTERVALS AND QUARTILE GROUPS WITH STEPSIZE 0.01%

freq = 0

BetaInvDum = 0

For i = 1 To 4

CGQrt(i) = 0

Next i

For i = 1 To r

j = 1

CGInt(i) = 0

While BetaInvDum <= Boundary(i) And freq + 0.0001 <= 1

freq = freq + 0.0001

BetaInvDum = Excel.WorksheetFunction.Beta\_Inv(freq, Alpha, Beta, 0, 10)

CGInt(i) = CGInt(i) + BetaInvDum

j = j + 1

'SUMMATIONS FOR CENTERS OF GRAVITY QUARTILE GROUPS

If freq <= 0.25 Then

CGQrt(1) = CGQrt(1) + BetaInvDum

CGQrt(2) = BetaInvDum

End If

If freq > 0.25 And freq <= 0.5 Then

CGQrt(2) = CGQrt(2) + BetaInvDum

CGQrt(3) = BetaInvDum

End If

If freq > 0.5 And freq <= 0.75 Then

CGQrt(3) = CGQrt(3) + BetaInvDum

CGQrt(4) = BetaInvDum

End If

If freq > 0.75 Then

CGQrt(4) = CGQrt(4) + BetaInvDum

End If

Wend

If i = 1 Then

CGInt(i) = CGInt(i) / j

Else

CGInt(i) = CGInt(i) / (j - 1)

End If

Next i

For j = 1 To r

Cells(k + 3, 35 + j) = CGInt(j)

Next j

For j = 1 To 4

If j = 1 Then

Cells(k + 3, 47 + j) = CGQrt(j) / 2501

Else

Cells(k + 3, 47 + j) = CGQrt(j) / 2500

End If

Next j

Next k

'LAYOUT OF THE OUTPUT

Cells(3, "S").HorizontalAlignment = xlLeft

Range("T3:W3").HorizontalAlignment = xlRight

Range("S4:S1000").Select

Selection.NumberFormat = "0"

Range("S4:S1000").HorizontalAlignment = xlLeft

Range("X3:AZ1000").Select

Selection.NumberFormat = "0.00"

Range("T4:W1000").Select

Selection.NumberFormat = "0.00"

Cells(1, "S").Select

Worksheets("Input").Select

End Sub

1. We have not tested whether the program is applicable in earlier versions of MS Excel. [↑](#footnote-ref-1)
2. A boundary between two response options is the value on the 0 to 10 continuum at which respondents are expected to change their choice from the one to the other option. [↑](#footnote-ref-2)
3. <http://www.europeansocialsurvey.org/> [↑](#footnote-ref-3)