Distribution Fitting (Univariate Mixture Distributions)

Revised: 12/3/2019

Summary 1
Data Input 3
Analysis Options 4
Tables and Graphs 5
Statistical Model 5
Analysis Summary 6
Plot of Fitted Distribution 7
Estimated Parameters 10
Density Trace 10
Model Comparisons 12
Goodness-of-Fit Test 13
Component Assignments 17
Quantile Plot 19
Tail Areas 20
Critical Values 21
Quantile-Quantile Plot 22
Distribution Functions 1 and 2 23
Save Results 25
References 26

Summary

The **Distribution Fitting (Univariate Mixture Models)** fits a distribution to continuous numeric data that consists of a mixture of 2 or more univariate Gaussian distributions. The components of the mixture may represent different groups in the sample used to fit the overall distribution, or the mixture model may approximate some distribution with a complicated shape.

The procedure fits the distribution, creates graphs, and calculates tail areas and critical values. Tools are also provided for determining how many components are needed to represent a data sample.

The calculations are performed by the “EMCluster” package in R. To run the procedure, R must be installed on your computer together with those packages. For information on downloading and installing R, refer to the document titled “R – Installation and Configuration”.

Sample StatFolio: *univariate mixture.sgp*

Sample Data

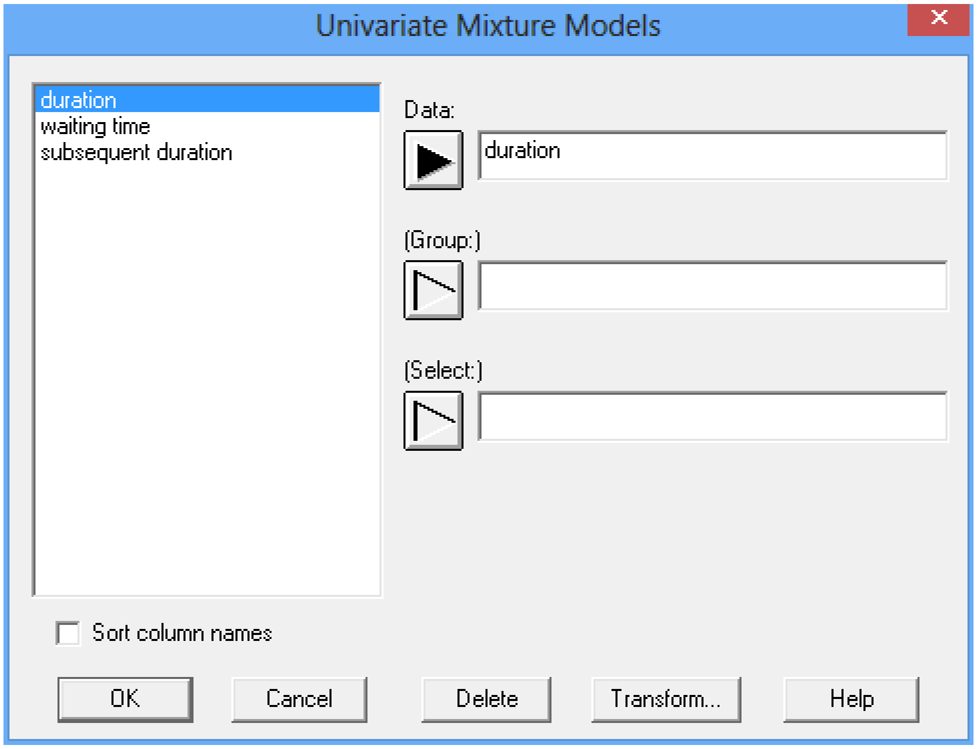
The file *old faithful.sgd* contains measurements of the duration of consecutive eruptions of the Old Faithful Geyser in Yellowstone National Park. The first several rows of that file are shown below:

duration
3.600
1.800
3.333
2.283
4.533
2.883
4.700
3.600
1.950
4.350
…

Duration measures the length of each eruption in minutes..

Data Input

When the procedure is first selected, a data input dialog box is displayed requesting the name of the column containing the data:



Data: duration

(Group:)

(Select:)

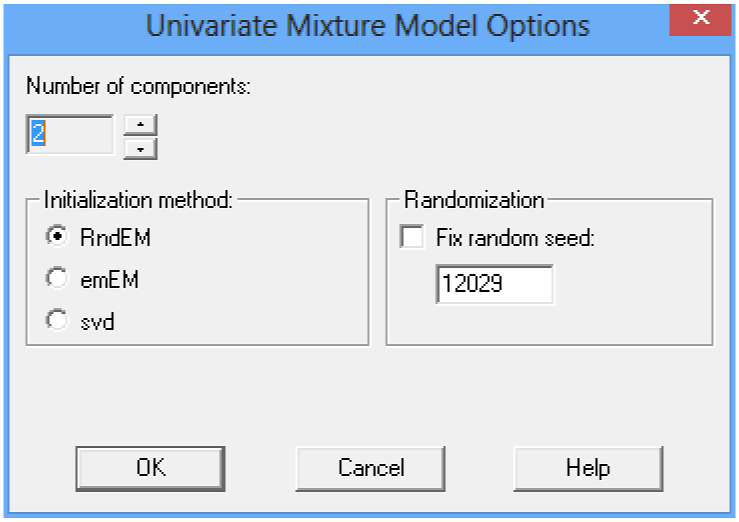
- **Data:** the data to be used to fit the distribution.

- **(Group:)** optional numeric or character column identifying group membership for each observation. This entry has no effect on the fitted model. It is only used to summarize membership percentages in each component of the model.

- **(Select:**) optional subset selection.

Analysis Options

After the data have been specified, the *Analysis Options* dialog box is displayed:



Univariate Mixture Model Options

Number of components: 2

Initialization method:
● RndEM
○ emEM
○ svd

Randomization
□ Fix random seed: 12029

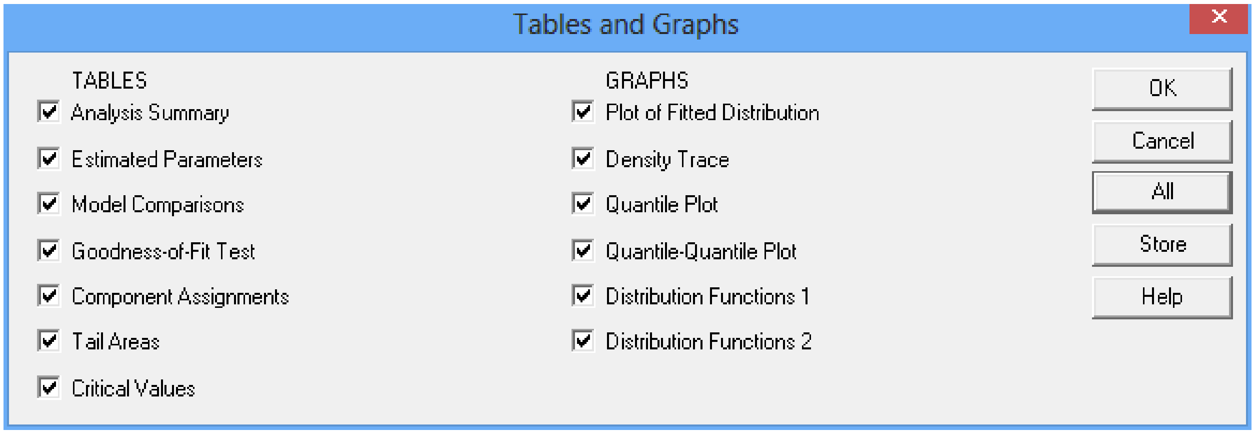
- **Number of components:** *K*, the number of separate univariate Gaussian distributions contained in the model.

- **Initialization method:** method used to initialize the model-fitting algorithm. *RndEm* (Maitra 2009) randomly selects K centers and groups all other data to closest center; then repeats the process and selects best initialization based on log likelihood. *emEM* (Biernacki et al. 2003) consists of both short-EM and long-EM steps. *svd* (Maitra 2001) selects centers from major component space and singular value decomposition of data. For more details, see Wei-Chen Chen and Ranjan Maitra (2015).

- **Randomization:** whether to fix the seed of the random number generator using the value indicated. If the seed is fixed, the same results will be obtained each time the procedure is run (assuming no other options are changed).

Tables and Graphs

The following tables and graphs may be created:



TABLES

☑ Analysis Summary

☑ Estimated Parameters

☑ Model Comparisons

☑ Goodness-of-Fit Test

☑ Component Assignments

☑ Tail Areas

☑ Critical Values

GRAPHS

☑ Plot of Fitted Distribution

☑ Density Trace

☑ Quantile Plot

☑ Quantile-Quantile Plot

☑ Distribution Functions 1

☑ Distribution Functions 2

Statistical Model

The statistical model fit by this procedure is a mixture of *K* univariate Gaussian distributions. It may be written as

$$f(y) = \sum_{i=1}^{K} p_i \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(y-\mu_i)^2}{2\sigma_i^2}} \qquad (1)$$

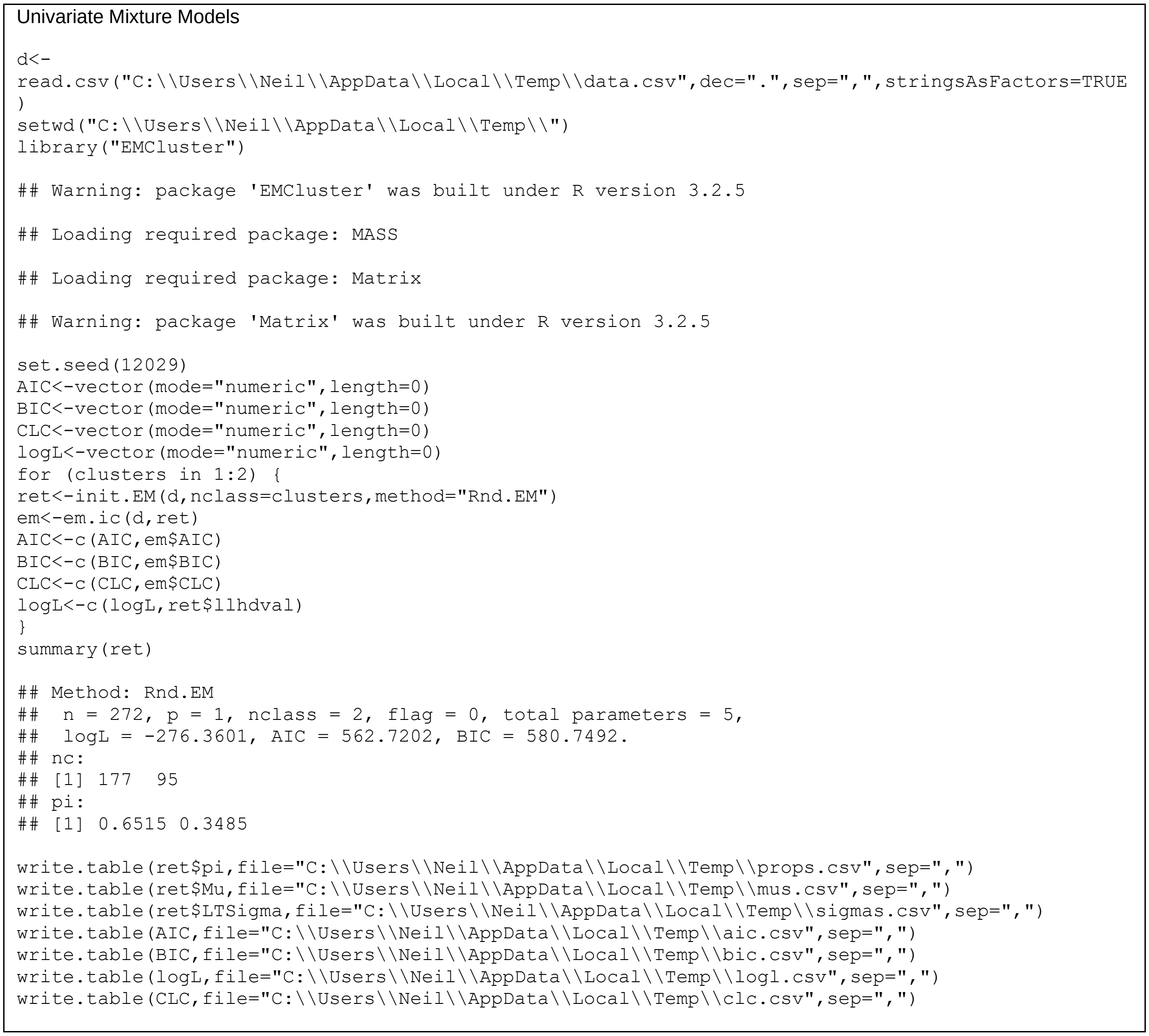
where

$$\sum_{i=1}^{K} p_i = 1 \qquad (2)$$

The model parameters include *K* means μ_i, *K* standard deviations σ_i, and (*K*-1) independent mixing parameters p_i.

Analysis Summary

The *Analysis Summary* displays the R commands that were executed.



Univariate Mixture Models

```
d<-
read.csv("C:\\Users\\Neil\\AppData\\Local\\Temp\\data.csv",dec=".",sep=",",stringsAsFactors=TRUE
)
setwd("C:\\Users\\Neil\\AppData\\Local\\Temp\\")
library("EMCluster")

## Warning: package 'EMCluster' was built under R version 3.2.5

## Loading required package: MASS

## Loading required package: Matrix

## Warning: package 'Matrix' was built under R version 3.2.5

set.seed(12029)
AIC<-vector(mode="numeric",length=0)
BIC<-vector(mode="numeric",length=0)
CLC<-vector(mode="numeric",length=0)
logL<-vector(mode="numeric",length=0)
for (clusters in 1:2) {
ret<-init.EM(d,nclass=clusters,method="Rnd.EM")
em<-em.ic(d,ret)
AIC<-c(AIC,em$AIC)
BIC<-c(BIC,em$BIC)
CLC<-c(CLC,em$CLC)
logL<-c(logL,ret$llhdval)
}
summary(ret)

## Method: Rnd.EM
##  n = 272, p = 1, nclass = 2, flag = 0, total parameters = 5,
##  logL = -276.3601, AIC = 562.7202, BIC = 580.7492.
## nc:
## [1] 177  95
## pi:
## [1] 0.6515 0.3485

write.table(ret$pi,file="C:\\Users\\Neil\\AppData\\Local\\Temp\\props.csv",sep=",")
write.table(ret$Mu,file="C:\\Users\\Neil\\AppData\\Local\\Temp\\mus.csv",sep=",")
write.table(ret$LTSigma,file="C:\\Users\\Neil\\AppData\\Local\\Temp\\sigmas.csv",sep=",")
write.table(AIC,file="C:\\Users\\Neil\\AppData\\Local\\Temp\\aic.csv",sep=",")
write.table(BIC,file="C:\\Users\\Neil\\AppData\\Local\\Temp\\bic.csv",sep=",")
write.table(logL,file="C:\\Users\\Neil\\AppData\\Local\\Temp\\logl.csv",sep=",")
write.table(CLC,file="C:\\Users\\Neil\\AppData\\Local\\Temp\\clc.csv",sep=",")
```

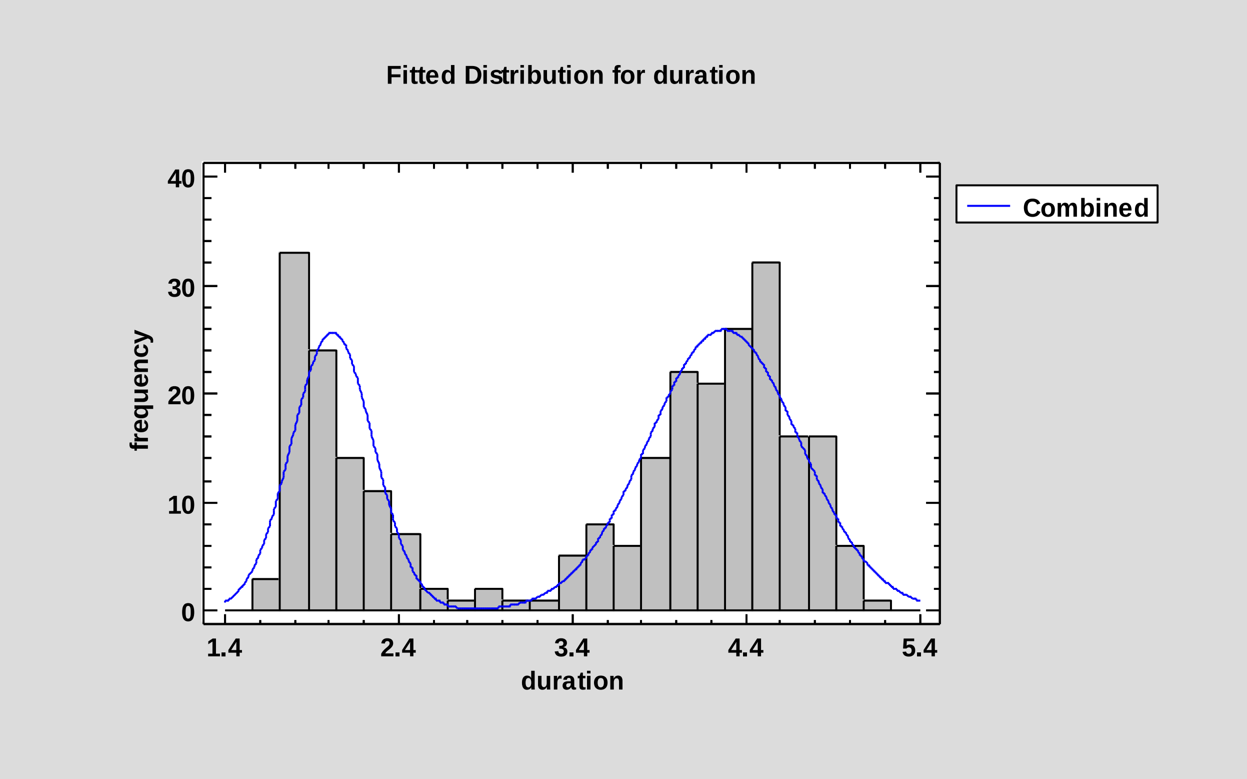
In the lines below *summary(ret)*, it summarizes the results of the model fitting process. Of particular interest are:

1. *n*: the number of observations used to fit the model.
2. *nclass*: the number of components in the fitted model.
3. *total parameters*: the number of estimated parameters.
4. *logL:* the final value of the log likelihood function.
5. *AIC:* the value of the Akaike Information Criterion.
6. *BIC:* the value of the Bayesian Information Criterion.
7. *nc*: the number of observations in each component of the model.

8. *pi*: the estimated proportion of the distribution for each component.

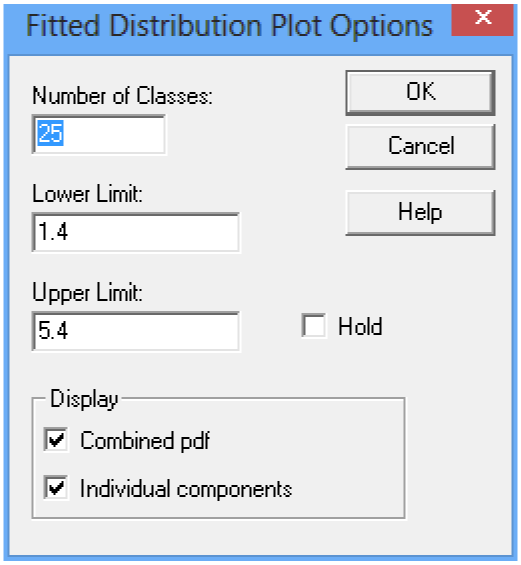
Plot of Fitted Distribution

This plot shows the fitted distribution.



The above plot shows the combined distribution.

Pane Options



Fitted Distribution Plot Options

Number of Classes: 25

Lower Limit: 1.4

Upper Limit: 5.4

□ Hold

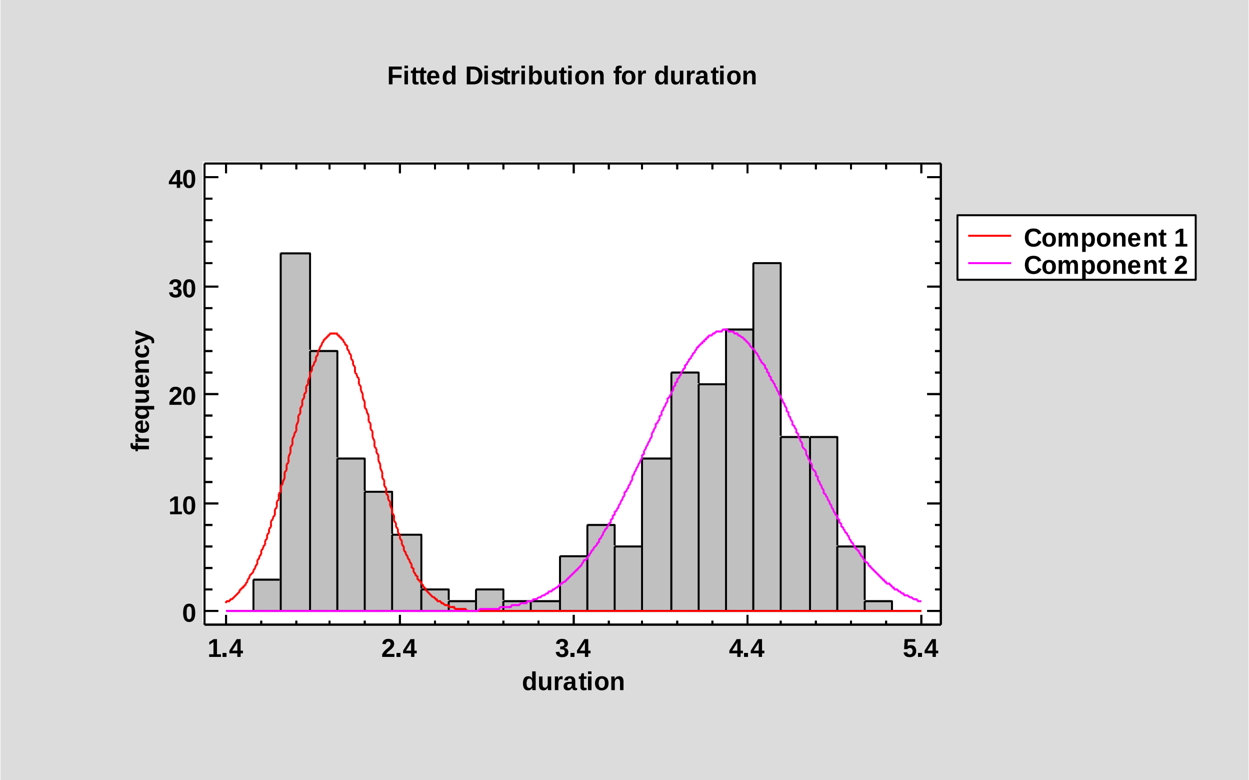
Display

☑ Combined pdf

☑ Individual components

- **Number of classes:** number of classes in the histogram used to display the data.
- **Lower limit:** lower limit of the first class.
- **Upper limit:** upper limit of the last class.
- **Hold:** whether to hold the above values fixed when the data change.
- **Display:** whether to display the combined distribution, the individual components, or both.

The plot below shows the separate components.


Fitted Distribution for duration
frequency
duration
0
10
20
30
40
1.4
2.4
3.4
4.4
5.4
Component 1
Component 2

Estimated Parameters

This table shows the estimated model parameters:

Estimated Parameters

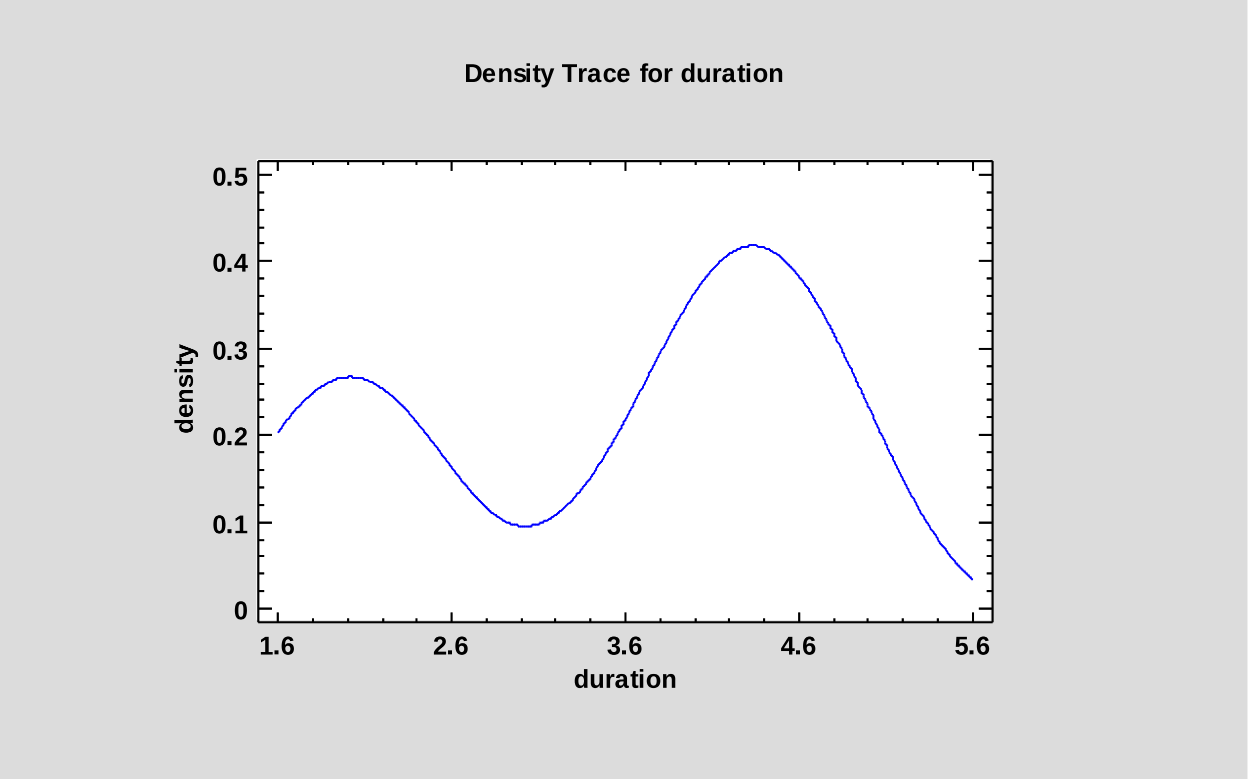
Sample size: 272

Component	*Proportion*	*Mean*	*Sigma*
1	0.348462	2.01874	0.235837
2	0.651538	4.27347	0.436871
Combined		3.48778	1.13927

It includes estimated
s of each component mean, each component standard deviation, and the component proportions p_i. The mean and standard deviation of the combined distribution are also displayed.

Density Trace

The *Density Trace* provides a nonparametric estimate of the probability density function of the population from which the data were sampled. It is created by counting the number of observations that fall within a window of fixed width moved across the range of the data. It may be used as a tool for helping determine how many components may be needed to fit the data.



The estimated density function is given by:

$$f(x) = \frac{1}{hn}\sum_{i=1}^{n} W\left(\frac{x - x_i}{h}\right) \quad (3)$$

where *h* is the width of the window in units of *X* and *W(u)* is a weighting function determined by the selection on the *Pane Options* dialog box. Two forms of weighting function are offered:

Boxcar Function

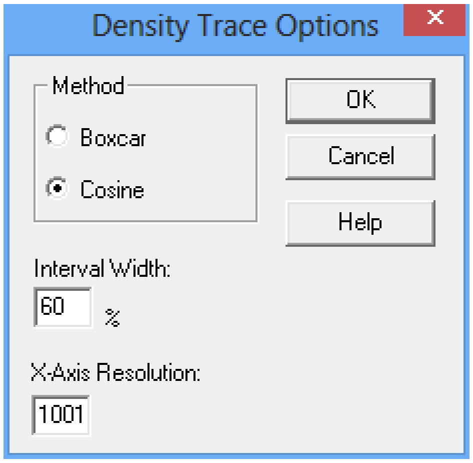
$$W(u) = \begin{cases} 1 & if|u| \le 1/2 \\ 0 & otherwise \end{cases} \quad (4)$$

Cosine Function

$$W(u) = \begin{cases} 1 + \cos(2\pi u) & if|u| < 1/2 \\ 0 & otherwise \end{cases} \quad (5)$$

The latter selection usually gives a smoother result, with the desirable value of *h* depending on the size of the data sample. In the case of the eruption data, it is clear that at least 2 components are needed to model the data.

Pane Options



Density Trace Options

Method
○ Boxcar
● Cosine

Interval Width: 60 %

X-Axis Resolution: 1001

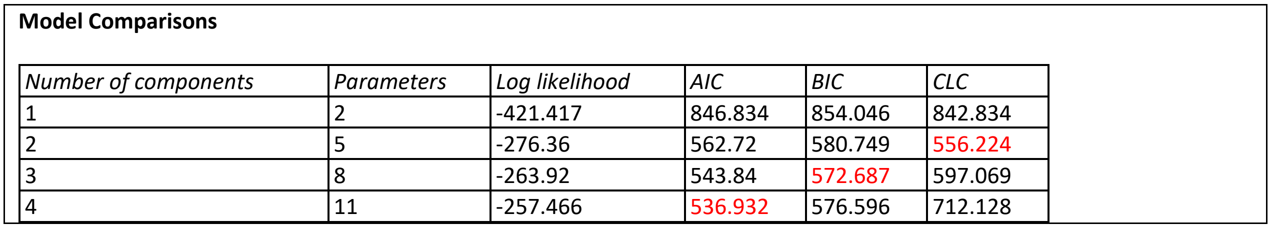
- **Method:** the desired weighting function. The boxcar function weights all values within the window equally. The cosine function gives decreasing weight to observations further from the center of the window. The default selection is determined by the setting on the *EDA* tab of the *Preferences* dialog box accessible from the *Edit* menu.

- **Interval Width:** the width of the window *h* within which observations affect the estimated density, as a percentage of the range covered by the x-axis. *h* = 60% is not unreasonable for a small sample but may not give as much detail as a smaller value in larger samples.

- **X-Axis Resolution**: the number of points at which the density is estimated.

Model Comparisons

This table compares the results of fitting mixtures of Gaussian distributions with components varying between 1 and *K*, the number of components indicated on the *Analysis Options* dialog box.



Model Comparisons

Number of components	*Parameters*	*Log likelihood*	*AIC*	*BIC*	*CLC*
1	2	-421.417	846.834	854.046	842.834
2	5	-276.36	562.72	580.749	556.224
3	8	-263.92	543.84	572.687	597.069
4	11	-257.466	536.932	576.596	712.128

The table includes:

1. **Parameters** - the number of estimated parameters *m* in the model. For a model with K components, *m* = 3*K*-1.

2. **Log likelihood** - the value of the log likelihood function $ln(\hat{L})$. Models with more parameters will always have larger values of the log likelihood function.

3. **AIC** – the value of the Akaike Information Criterion. The AIC is a widely used criterion for model selection and penalizes the likelihood function based on the number of estimated parameters according to

$$AIC = 2m - 2ln(\hat{L}) \tag{6}$$

4. **BIC** – the value of the Bayesian Information Criterion. It is similar to the AIC except that it uses a different equation to penalize the likelihood function:

$$BIC = \ln(n)m - 2ln(\hat{L}) \tag{7}$$

5. **CLC** – the value of the Classification Likelihood Criterion. It is similar to AIC and BIC but penalizes the likelihood function based on an entropy measure (see Biernacki and Govaert 1997).

Models with the smallest values of AIC, BIC and CLC are preferable. In the table above, each criterion selects a model with a different number of components.

Goodness-of-Fit Test

This table shows the result of a chi-square goodness of fit test run to determine whether the fitted distribution adequately models the sample data. The test divides the range of *X* into *q* intervals and compares the observed counts

O_j = number of data values observed in interval *j*

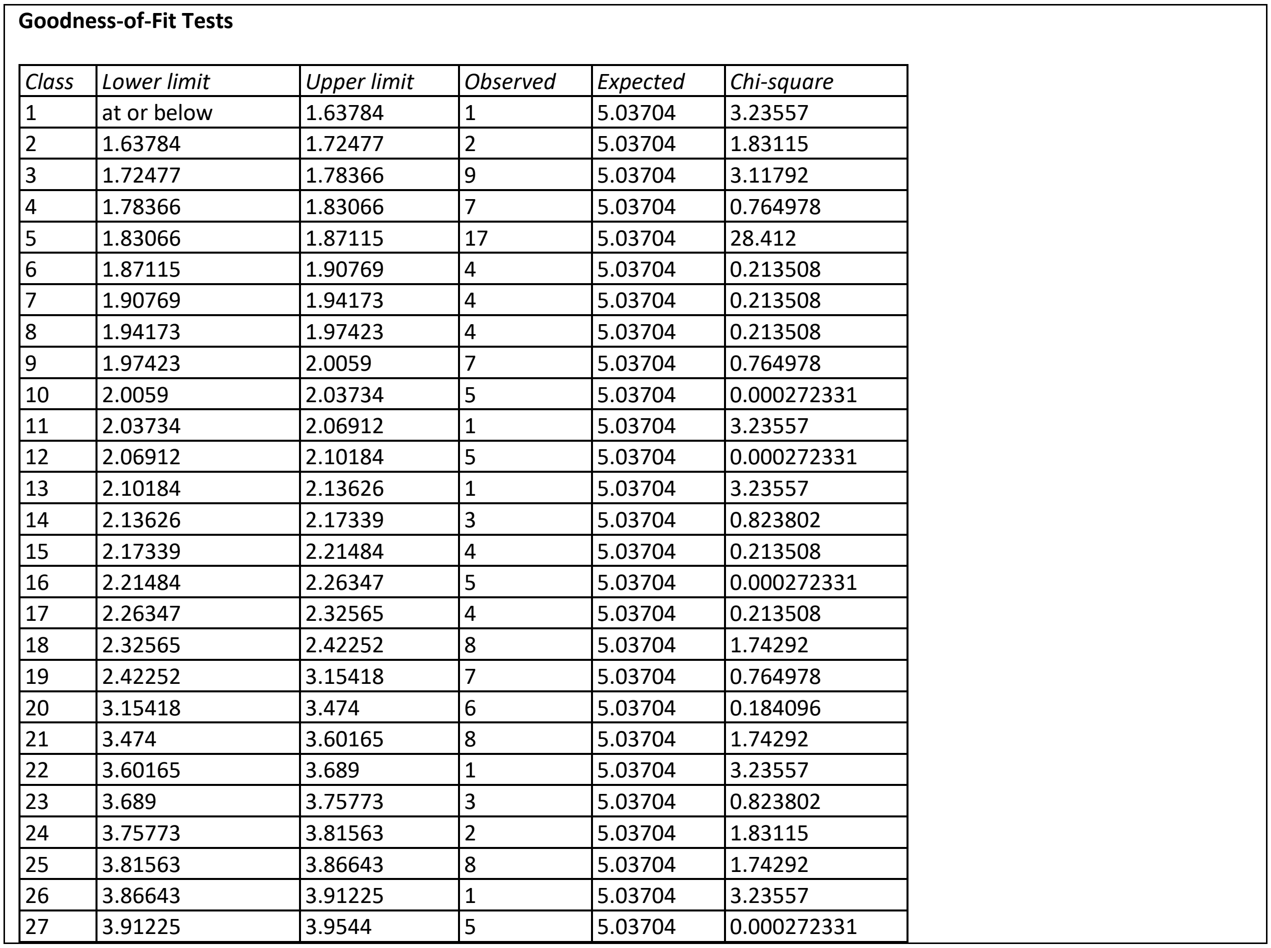
to the number expected given the fitted distribution

E_j = number of data values expected in interval *j* given the fitted distribution.

The test statistic is given by

$$\chi^2 = \sum_{j=1}^{q} \frac{(O_j - E_j)^2}{E_j} \tag{8}$$

which is compared to a chi-squared distribution with *q-m-1* degrees of freedom, where *m* is the number of parameters estimated when fitting the selected distribution.



Goodness-of-Fit Tests

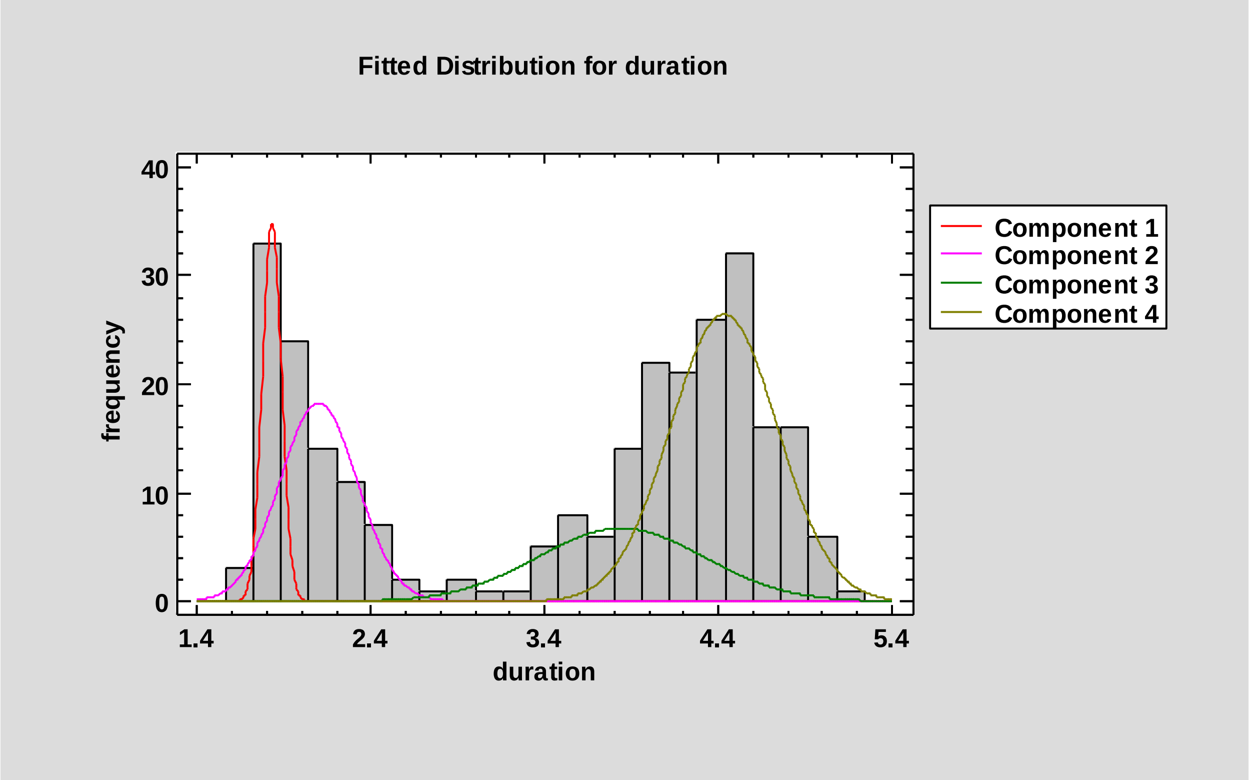
Class	*Lower limit*	*Upper limit*	*Observed*	*Expected*	*Chi-square*
1	at or below	1.63784	1	5.03704	3.23557
2	1.63784	1.72477	2	5.03704	1.83115
3	1.72477	1.78366	9	5.03704	3.11792
4	1.78366	1.83066	7	5.03704	0.764978
5	1.83066	1.87115	17	5.03704	28.412
6	1.87115	1.90769	4	5.03704	0.213508
7	1.90769	1.94173	4	5.03704	0.213508
8	1.94173	1.97423	4	5.03704	0.213508
9	1.97423	2.0059	7	5.03704	0.764978
10	2.0059	2.03734	5	5.03704	0.000272331
11	2.03734	2.06912	1	5.03704	3.23557
12	2.06912	2.10184	5	5.03704	0.000272331
13	2.10184	2.13626	1	5.03704	3.23557
14	2.13626	2.17339	3	5.03704	0.823802
15	2.17339	2.21484	4	5.03704	0.213508
16	2.21484	2.26347	5	5.03704	0.000272331
17	2.26347	2.32565	4	5.03704	0.213508
18	2.32565	2.42252	8	5.03704	1.74292
19	2.42252	3.15418	7	5.03704	0.764978
20	3.15418	3.474	6	5.03704	0.184096
21	3.474	3.60165	8	5.03704	1.74292
22	3.60165	3.689	1	5.03704	3.23557
23	3.689	3.75773	3	5.03704	0.823802
24	3.75773	3.81563	2	5.03704	1.83115
25	3.81563	3.86643	8	5.03704	1.74292
26	3.86643	3.91225	1	5.03704	3.23557
27	3.91225	3.9544	5	5.03704	0.000272331

28	3.9544	3.99377	2	5.03704	1.83115
29	3.99377	4.03099	6	5.03704	0.184096
30	4.03099	4.06652	3	5.03704	0.823802
31	4.06652	4.10074	9	5.03704	3.11792
32	4.10074	4.13392	4	5.03704	0.213508
33	4.13392	4.16632	4	5.03704	0.213508
34	4.16632	4.19814	5	5.03704	0.000272331
35	4.19814	4.22956	1	5.03704	3.23557
36	4.22956	4.26076	7	5.03704	0.764978
37	4.26076	4.29188	4	5.03704	0.213508
38	4.29188	4.32311	3	5.03704	0.823802
39	4.32311	4.35459	9	5.03704	3.11792
40	4.35459	4.3865	5	5.03704	0.000272331
41	4.3865	4.41903	5	5.03704	0.000272331
42	4.41903	4.45238	5	5.03704	0.000272331
43	4.45238	4.48682	3	5.03704	0.823802
44	4.48682	4.52263	9	5.03704	3.11792
45	4.52263	4.5602	6	5.03704	0.184096
46	4.5602	4.60003	11	5.03704	7.0591
47	4.60003	4.64278	4	5.03704	0.213508
48	4.64278	4.68939	3	5.03704	0.823802
49	4.68939	4.7413	8	5.03704	1.74292
50	4.7413	4.80083	9	5.03704	3.11792
51	4.80083	4.87219	5	5.03704	0.000272331
52	4.87219	4.96452	6	5.03704	0.184096
53	4.96452	5.10549	4	5.03704	0.213508
54	above 5.10549		0	5.03704	5.03704

Chi-square statistic = 98.8529 with 48 degrees of freedom   P-Value = 0.0000

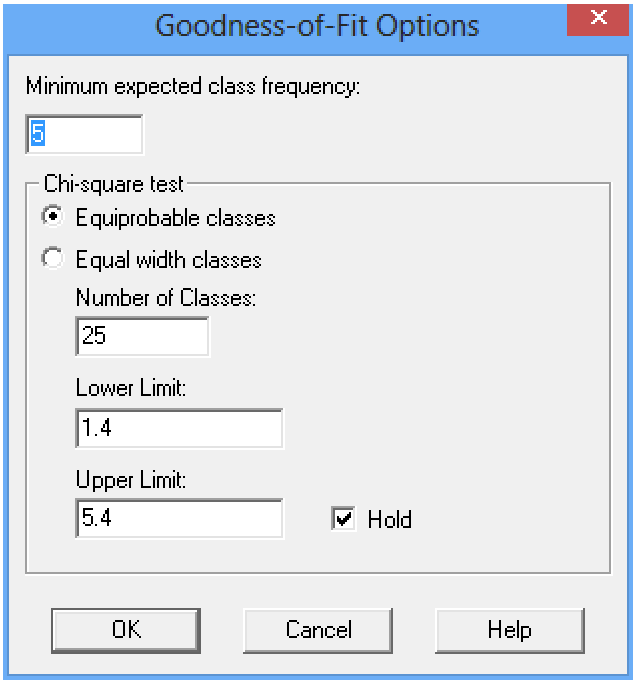
If the P-Value displayed at the bottom of the table is less than 0.05, as in this case, the currently selected model does not adequately fit the observed data at the 5% significance level. Notice also that the rightmost column shows the contribution of each class to the overall chi-square statistic. The class that contributes most to the statistic is the class between 1.83066 minutes and 1.87115 minutes, which contains 17 observations compared to an expected value of just over 5.

To obtain a model for which the chi-square P-Value is greater than 0.05, four components are needed as shown in the plot below:



2 components are now used to represent eruptions with durations less than approximately 2.7 minutes and 2 components for longer eruptions.

Pane Options


Goodness-of-Fit Options

Minimum expected class frequency: 5

Chi-square test
- ● Equiprobable classes
- ○ Equal width classes

Number of Classes: 25

Lower Limit: 1.4

Upper Limit: 5.4 ☑ Hold

When creating classes for the chi-square test, 2 approaches may be used:

1. **Equiprobable classes:** creates classes in which the expected number of observations is the same for all classes. As many classes as possible are created provided that the expected number in each class is greater than or equal to the specified *minimum expected class frequency*. This approach maximizes the power of the chi-square test unless the data are too heavily rounded.

2. **Equal width classes:** creates classes of equal width based on the *number of classes*, *lower limit* and *upper limit*. Any class for which the expected number of observations is less than the specified *minimum expected class frequency* is combined with another class to achieve that minimum. Note: the default number of classes and limits are based on the observed data. To keep the class specification fixed even if the data change, check *Hold*.

Component Assignments

To determine which component each of the observed data values is most likely to belong to, the program compares the height of the weighted component probability density functions at each value. It then assigns an observation to that value for which the weighted pdf is greatest. A table is given showing the component assignments:

Row	*duration*	*Component*
1	3.6	2
2	1.8	1
3	3.333	2
4	2.283	1
5	4.533	2
6	2.883	2
7	4.7	2
8	3.6	2
9	1.95	1
10	4.35	2
11	1.833	1
…	…	…

The output also shows the number of observations associated with each component in the model:

Component Assignments

Group Percentages

Group	*Count*	*Component 1*	*Component 2*
N/A	272	34.93%	65.07%

For the sample data, about 35% of the observations have been assigned to Component #1 and 65% to Component #2.

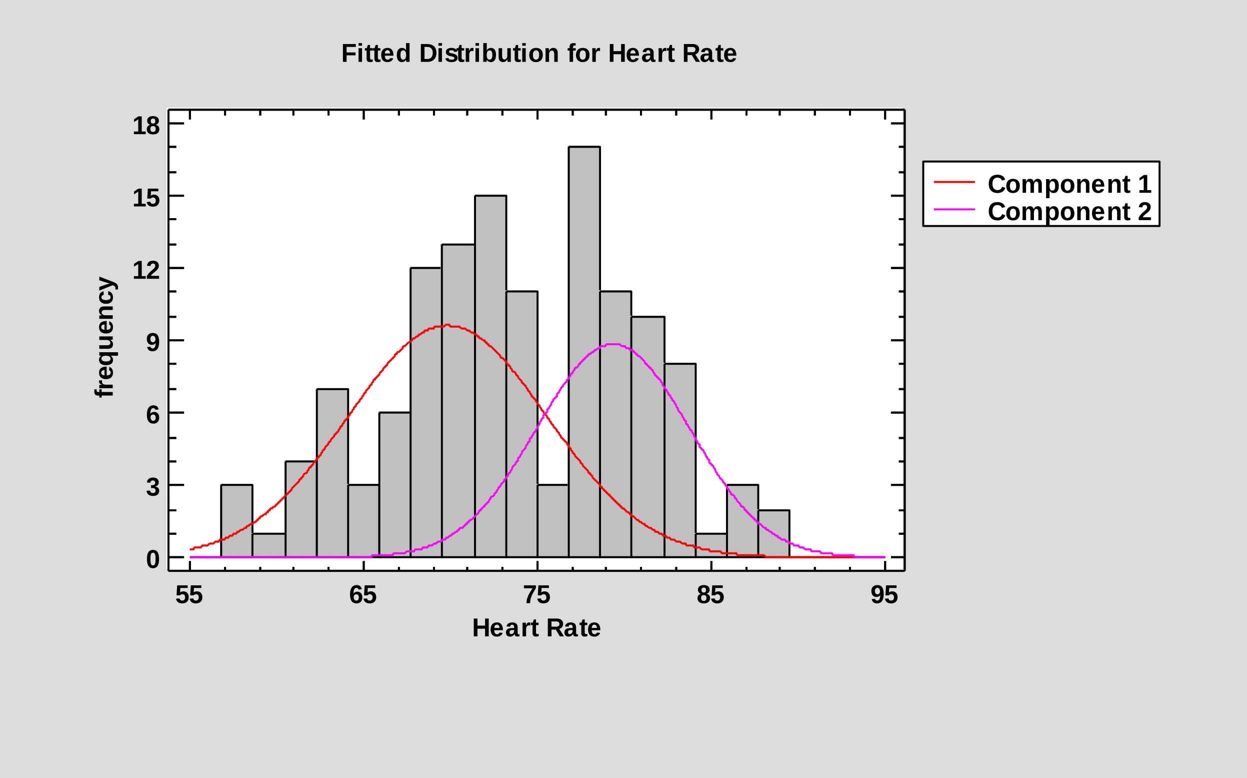
If an entry was made in the *Group* field on the data input dialog box, the assignment percentages will also be displayed by group. For example, fitting a 2-component model to patient heartrates in the *bodytemp.sgd* file shows the following comparison of males and females:

Component Assignments

Group Percentages

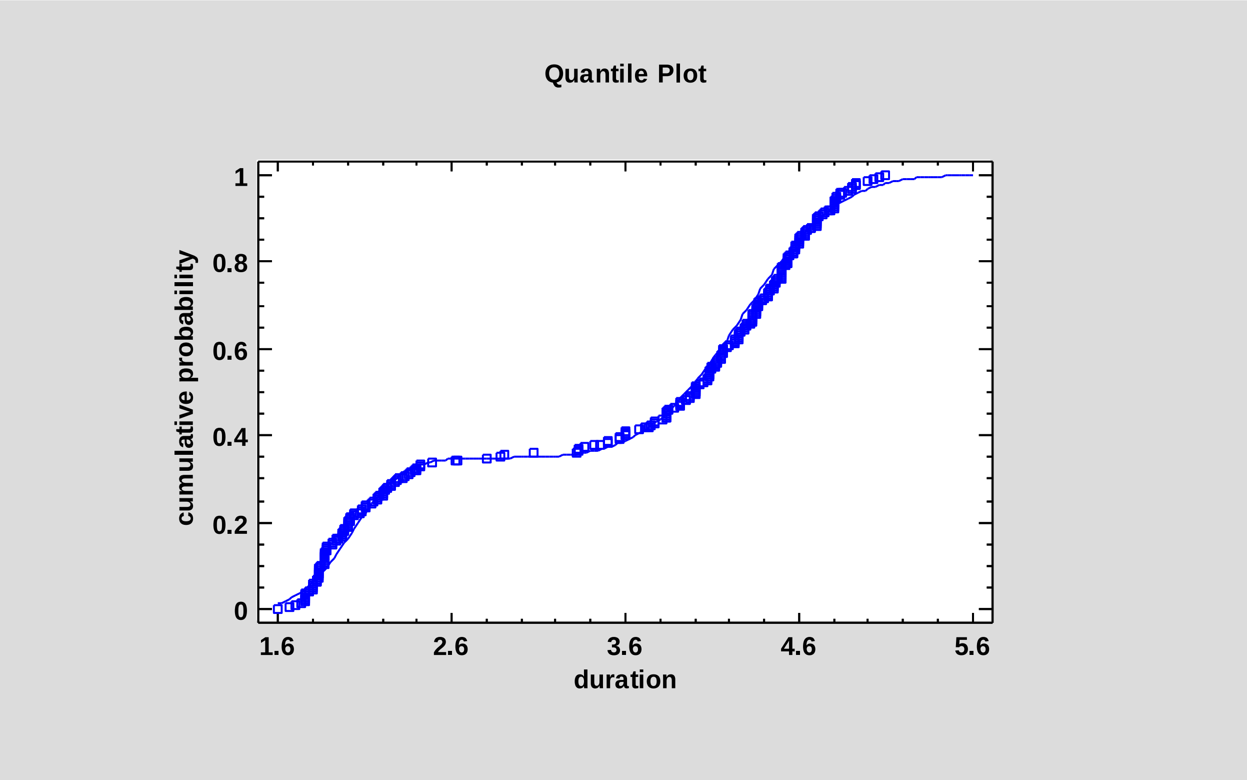
Gender	*Count*	*Component 1*	*Component 2*
Female	65	49.23%	50.77%
Male	65	66.15%	33.85%

Note that females are split almost evenly between the 2 components while 2/3rds of the males are in component #1. A plot of the fitted distribution is shown below:


Fitted Distribution for Heart Rate
frequency
18
15
12
9
6
3
0
55
65
75
85
95
Heart Rate
Component 1
Component 2

Quantile Plot

The *Quantile Plot* shows the fraction of observations at or below X, together with the cumulative distribution function of the fitted distribution



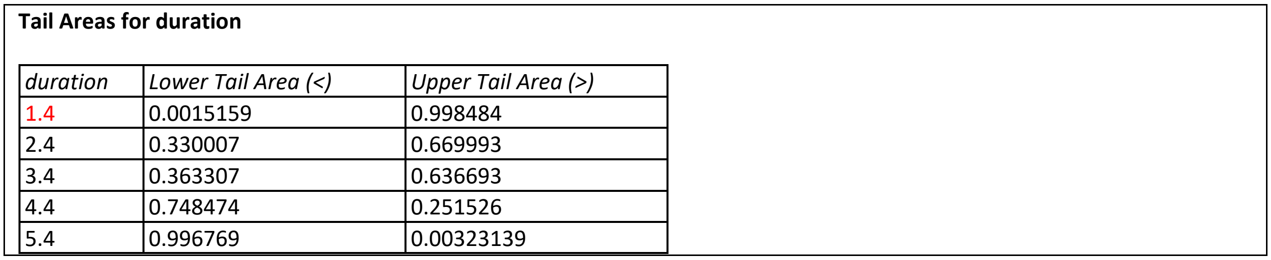
To create the plot, the data are sorted from smallest to largest and plotted at the coordinates

$$\left(x_{(j)}, \frac{j-0.5}{n} \right) \tag{9}$$

Ideally, the points will lie close to the line for the fitted distribution, as is the case in the plot above.

Tail Areas

This pane shows the value of the cumulative distribution at up to 10 values of X.



Tail Areas for duration

duration	*Lower Tail Area (<)*	*Upper Tail Area (>)*
1.4	0.0015159	0.998484
2.4	0.330007	0.669993
3.4	0.363307	0.636693
4.4	0.748474	0.251526
5.4	0.996769	0.00323139

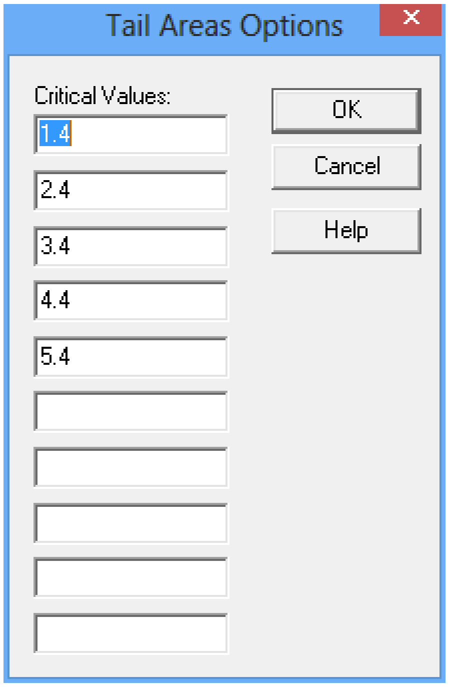
The table displays:

- **Lower Tail Area** – the probability that the random variable is less than or equal to X.

- **Upper Tail Area** – the probability that the random variable is greater than X.

For example, the probability that the duration of an eruption will be less than or equal to *X* = 3.4 is approximately 0.3633.

Pane Options



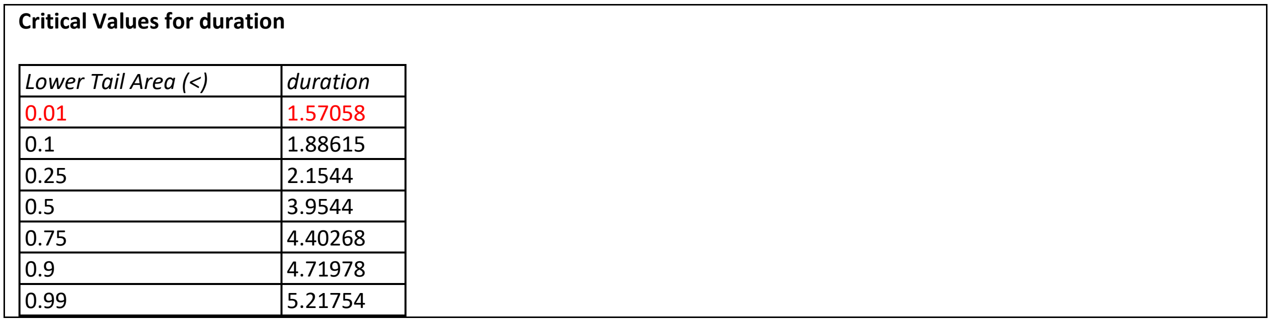
Tail Areas Options

Critical Values:
1.4
2.4
3.4
4.4
5.4

- **Critical Values**: values of X at which the cumulative probability is to be calculated.

Critical Values

This pane calculates the value of the random variable X below which lies a specified probability.

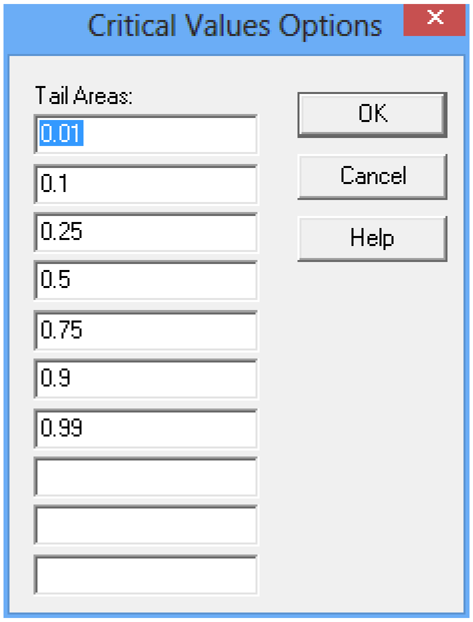


Critical Values for duration

Lower Tail Area (<)	*duration*
0.01	1.57058
0.1	1.88615
0.25	2.1544
0.5	3.9544
0.75	4.40268
0.9	4.71978
0.99	5.21754

The table displays the smallest value of X such that the probability of being less than or equal to X is at least the tail area desired. The table above shows that the c.d.f. of the fitted distribution equals 0.01 at *X* = 1.57058.

Pane Options



Critical Values Options

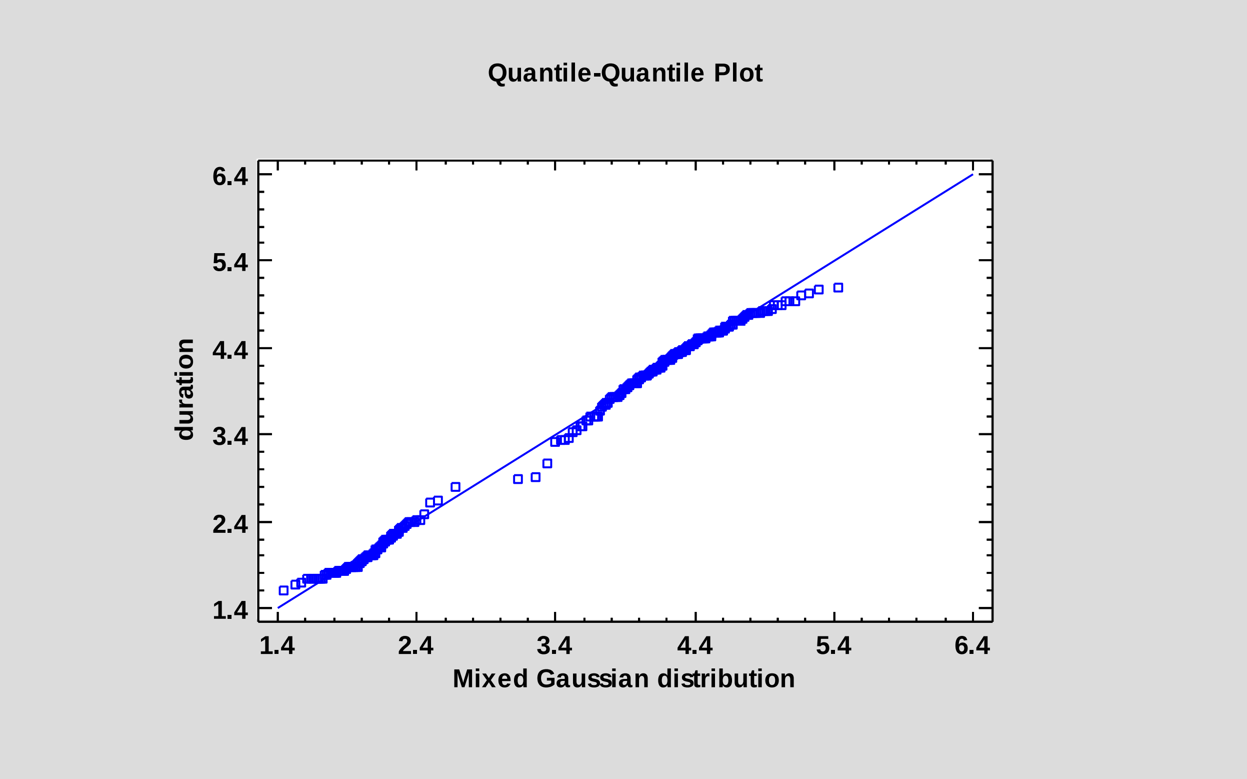
Tail Areas:

0.01
0.1
0.25
0.5
0.75
0.9
0.99

- **Tail Areas**: values of the c.d.f. at which to determine percentiles of the fitted distribution.

Quantile-Quantile Plot

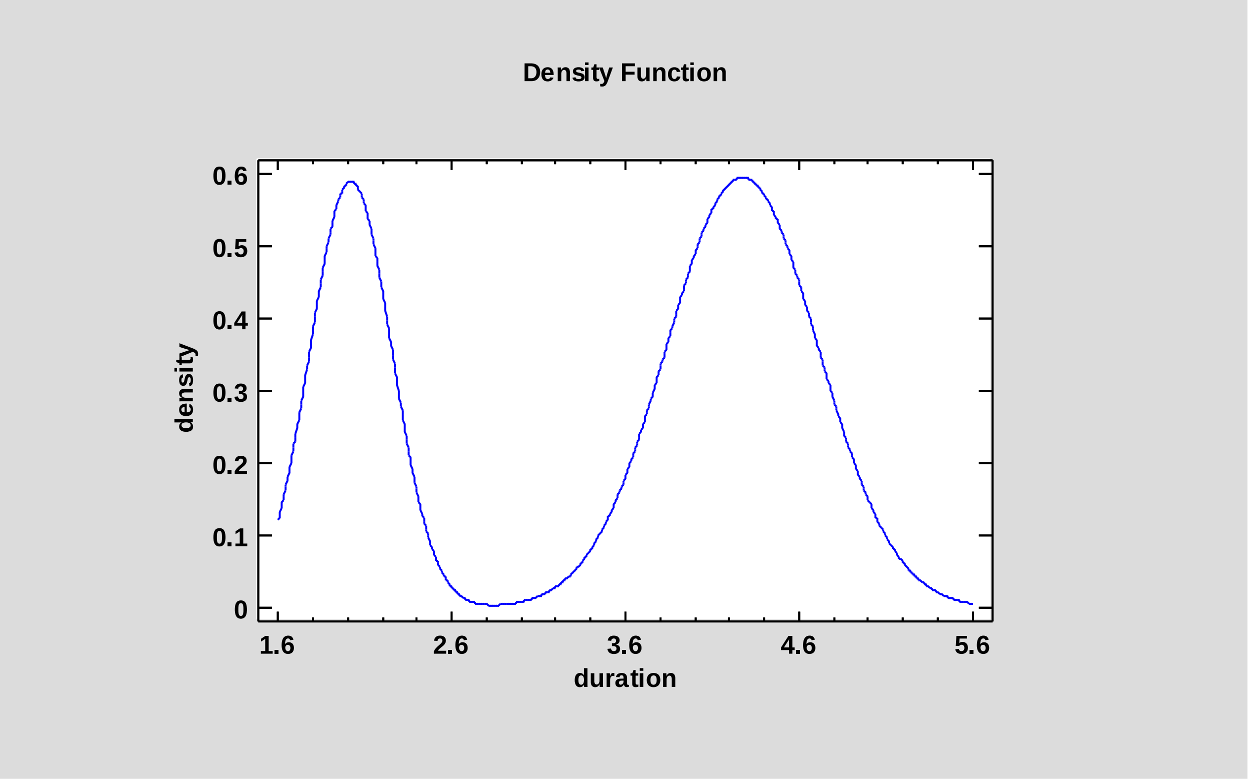
The *Quantile-Quantile Plot* shows the fraction of observations at or below X plotted versus the equivalent percentiles of the fitted distribution.



If the model fits the data well, the points on the plot should lie close to the diagonal line.

Distribution Functions 1 and 2

These two panes plot various functions for the fitted distribution.

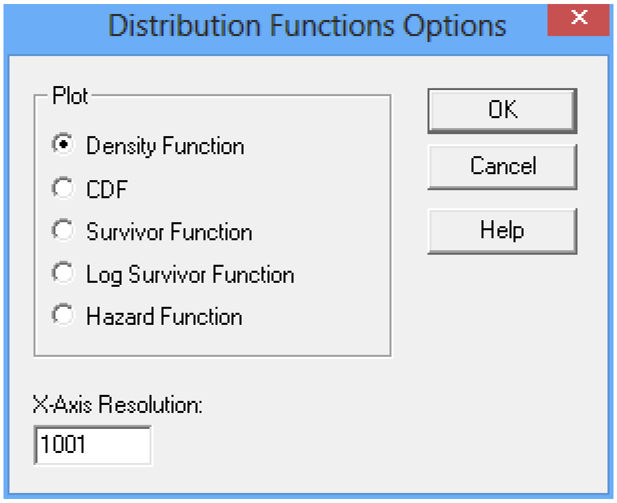


Using *Pane Options*, you may plot any of the following:

1. Probability density or mass function
2. Cumulative distribution function
3. Survivor function
4. Log survivor function
5. Hazard function

For definitions of these functions, see the documentation for *Probability Distributions*.

Pane Options



Distribution Functions Options

Plot

● Density Function

○ CDF

○ Survivor Function

○ Log Survivor Function

○ Hazard Function

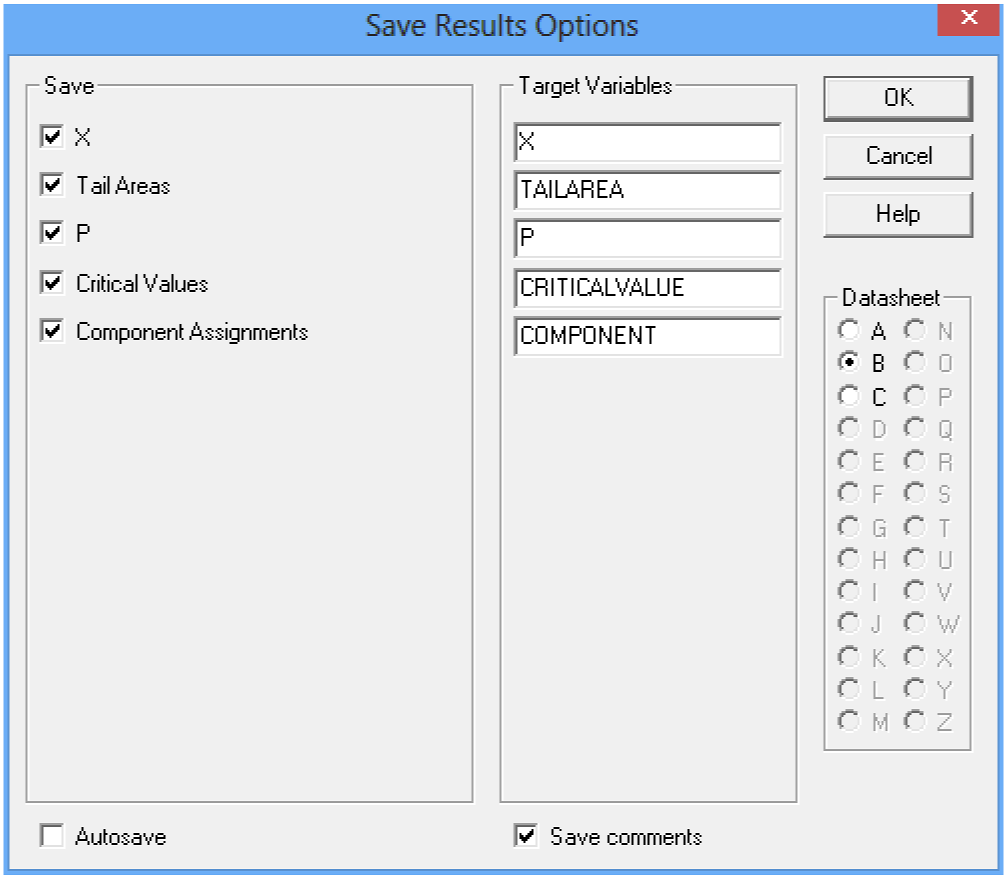
X-Axis Resolution: 1001

- **Plot**: the function to plot.

- **X-Axis Resolution** – the number of X locations at which the function is plotted. Increase this value if the function is not smooth enough.

Save Results

The results of selected calculations may be saved in a Statgraphics datasheet by pressing the *Save Results* button on the analysis toolbar. The following dialog box will be presented:



Save Results Options

Save
- ☑ X
- ☑ Tail Areas
- ☑ P
- ☑ Critical Values
- ☑ Component Assignments

Target Variables
- X
- TAILAREA
- P
- CRITICALVALUE
- COMPONENT

Datasheet

○ A	○ N
● B	○ O
○ C	○ P
○ D	○ Q
○ E	○ R
○ F	○ S
○ G	○ T
○ H	○ U
○ I	○ V
○ J	○ W
○ K	○ X
○ L	○ Y
○ M	○ Z

□ Autosave

☑ Save comments

- **Save:** select the items to be saved.
  - **X and Tail Areas** – from the *Tail Areas* table.
  - **P, Critical Values** – from the *Critical Values* table.
  - **Component Assignments** – most likely component associated with each observation.

- **Target Variables:** enter names for the columns to be created.

- **Datasheet:** the datasheet into which the results will be saved.

- **Autosave:** if checked, the results will be saved automatically each time a saved StatFolio is loaded.

- **Save comments:** if checked, comments for each column will be saved in the second line of the datasheet header.

References

Biernacki, C., Celeux, G, and Govaert, G. (2003). "Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models." Computational Statistics and Data Analysis, **413**, 561-575.

Biernacki C and Govaert G (1997). "Using the classification likelihood to choose the number of clusters." Computing Science and Statistics **29**, 451–457.

Maitra R. (2001). "Clustering massive datasets with applications to software metrics and tomography." Technometrics, **43**(3), 336-346.

Maitra R. (2009). "Initializing Partition-Optimization Algorithms." IEEE/ACM Transactions on Computational Biology and Bioinformatics, **6**, 144-157.

R Package "EMCluster" (2018) - https://cran.r-project.org/web/packages/EMCluster/EMCluster.pdf

Wei-Chen Chen and Ranjan Maitra (2015) – A Quick Guide for the EMCluster Package.