Population Pyramid Statlet

statgraphics 18®
centurion

Revised: 10/9/2017

Summary

The *Population Pyramid Statlet* is designed to compare the distribution of population counts (or similar values) between 2 groups. It may be used to display that distribution at a single point in time, or it may show changes over time in a dynamic manner. In the latter case, various options are offered for smoothing the data and for dealing with missing values.

Sample StatFolio: *pyramid.sgp*

Sample Data

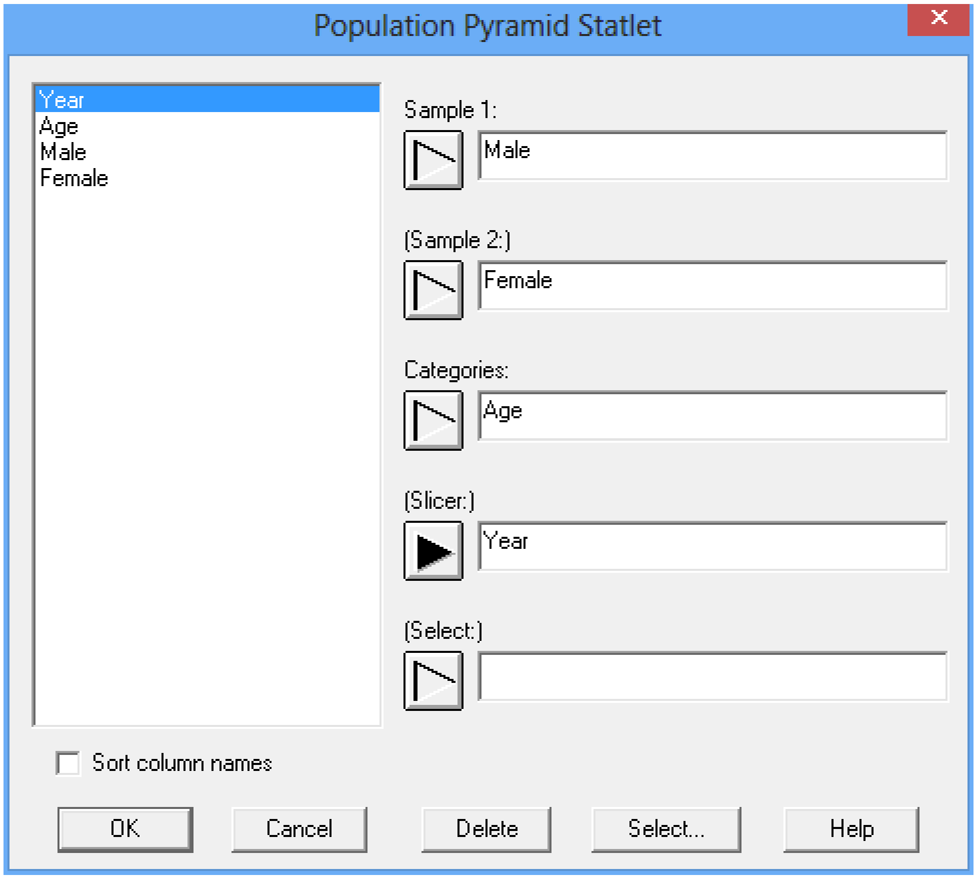
The file *us population.sgd* contains population data for the United States over *n* = 63 years (1950-2012). It was obtained from the United Nations (un.org).

The first several rows of that file are shown below:

Year	Age	Male	Female
1950	0-4	8236164	7990587
1950	5-9	6714555	6485130
1950	10-14	5660399	5458869
1950	15-19	5311342	5305256
1950	20-24	5606293	5875535
1950	25-29	5972078	6270182
1950	30-34	5624723	5892284
1950	35-39	5517544	5728842
1950	40-44	5070269	5133704
1950	45-49	4526366	4544099
1950	50-54	4128648	4143540
1950	55-59	3630046	3605074
1950	60-64	3037838	3021637
1950	65-69	2424561	2578375
1950	70-74	1628829	1783120
1950	75-79	1506756	1770995
1950	80-84	500345	624225
1950	85+	236828	340073
1951	0-4	8775000	8446000

Data Input

The data input dialog box requests the names of the columns containing the data values to be plotted:



Sample 1: Male

(Sample 2:) Female

Categories: Age

(Slicer:) Year

(Select:)

□ Sort column names

- **Sample 1:** name of a numeric column containing the population values for *n* time periods and *p* categories. This column is required.

- **Sample 2:** name of a second numeric column containing similar population values for a second sample. This column is normally supplied but is not required.

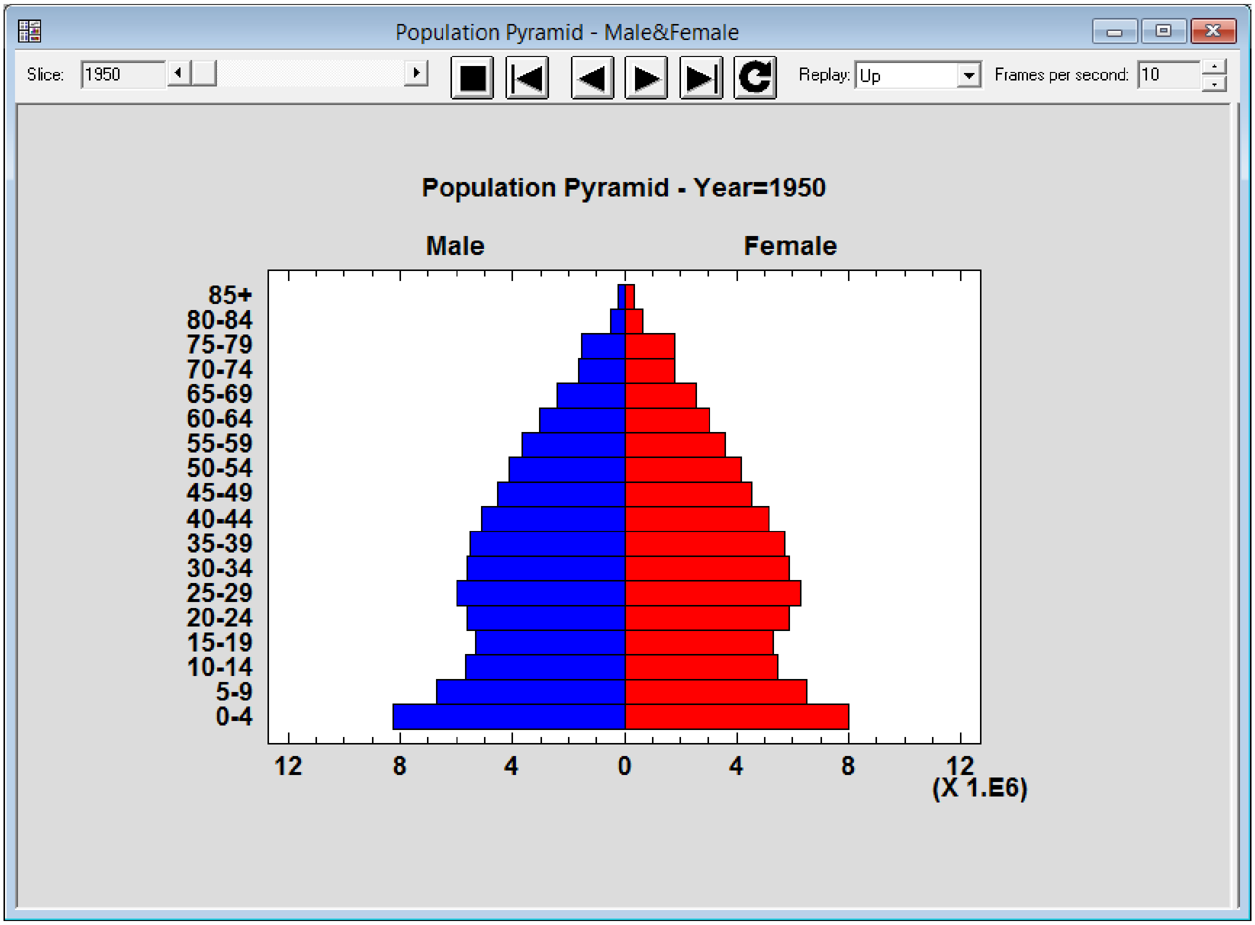
- **Categories:** name of the numeric or non-numeric column used to define population categories. There should be *p* unique values of this variable. The categories will be plotted in the order that they first appear.

- **Slicer:** name of the numeric column used to define subsets of the data. This variable, often a measure of time, is changed dynamically to illustrate changes in the data. There should be *n* unique values of this variable. If there is only 1 time period in the file, this column may be omitted.

- **Select:** optional subset selection.

Statlet

The output of this procedure is displayed in a dynamic Statlet window. When first created, the window displays data for the first time period (or first value of the *Slicer*) as shown below:



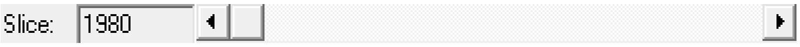
Population Pyramid - Male&Female

Slice: 1950

Replay: Up

Frames per second: 10

The length of the bars represent population.



Slice: 1980

Slice scrollbar: used to change the time period at which the data are displayed.

▶ **Forward button**: used to start a timer which plots the data for each time period in increasing order.

◀ **Backward button**: used to start a timer which plots the data for each time period in decreasing order.

▶| **Fast foward button**: advances to the last time period.

⏮ **Rewind button**: rewinds to the first time period.

⟲ **Replay button**: causes the sequence of time periods to be replayed over and over.

■ **Stop button**: stops the timer or replay.

Replay: Up

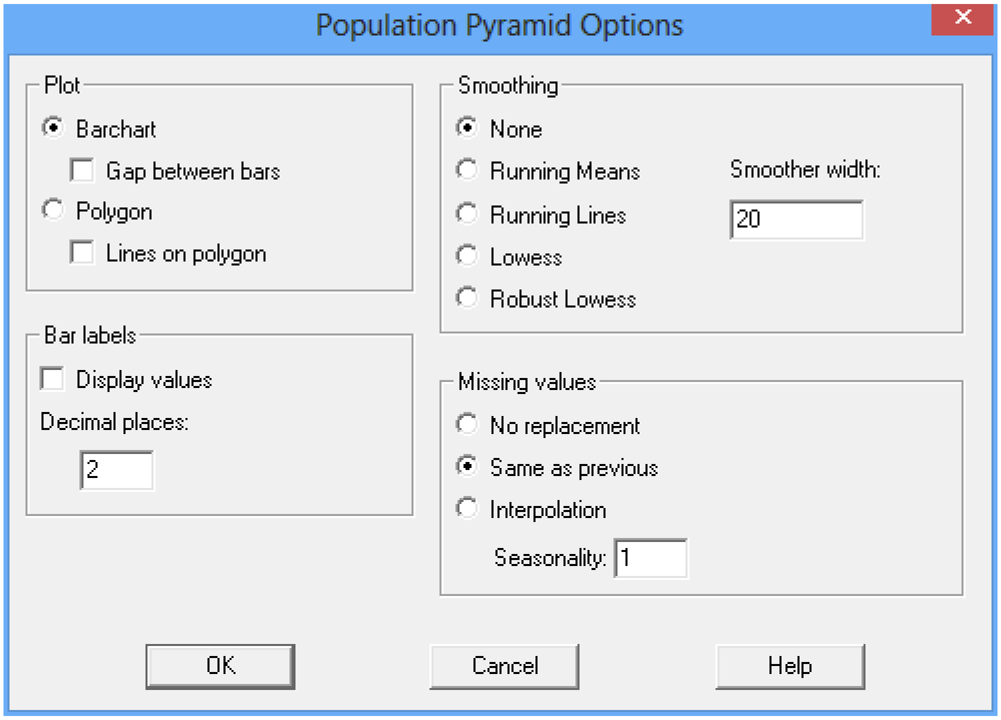
Replay pulldown list: specifies the direction for the time sequence when the replay button is pushed.

Frames per second: 1

Frames per second spinner: specifies the rate at which the time period is changed.

Analysis Options

The *Analysis Options* dialog box allows for various special effects to be applied:



Population Pyramid Options

Plot

● Barchart

☐ Gap between bars

○ Polygon

☐ Lines on polygon

Bar labels

☐ Display values

Decimal places: 2

Smoothing

● None

○ Running Means

○ Running Lines

○ Lowess

○ Robust Lowess

Smoother width: 20

Missing values

○ No replacement

● Same as previous

○ Interpolation

Seasonality: 1

The following options are available:

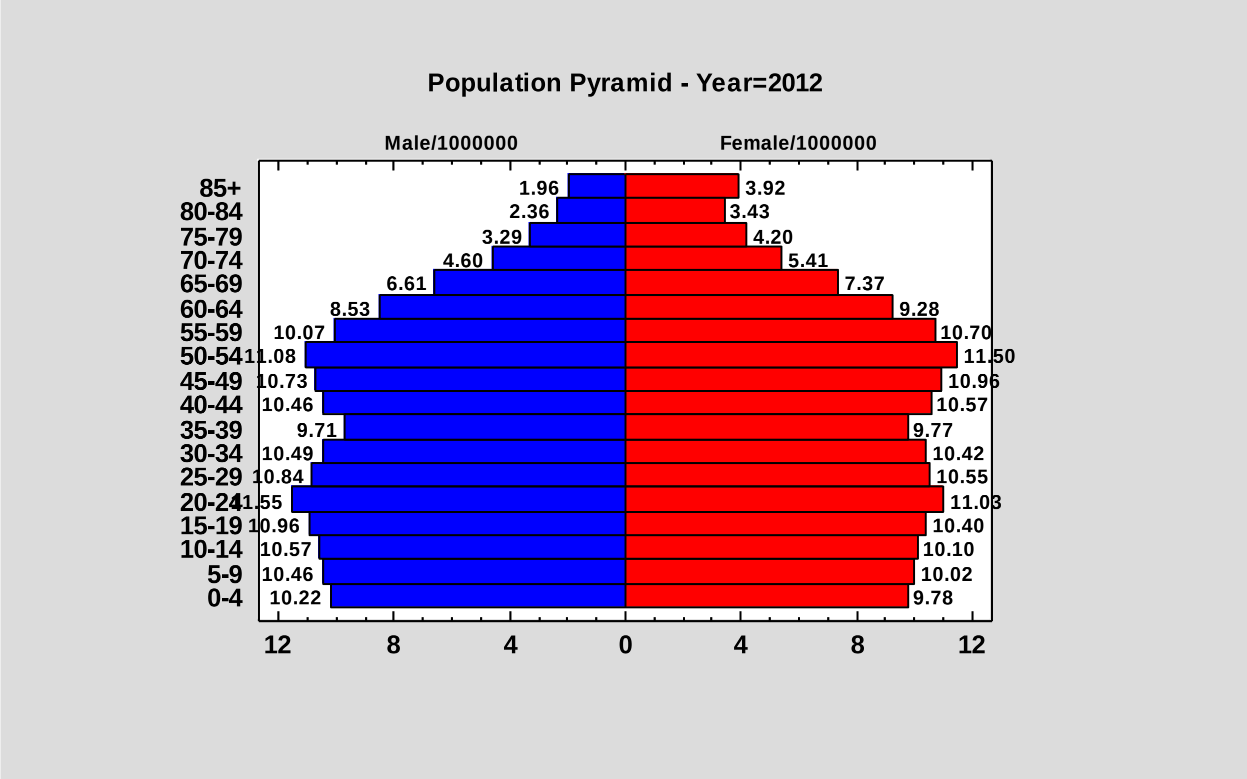
- **Plot**: specifies the type of plot to be created and selected options for that plot..

- **Bar labels**: indicates whether the data values should be plotted alongside each bar.

- **Smoothing**: smooths each time series using one of four methods. These are the same methods used to smooth X-Y scatterplots as described in the PDF document titled *Graphics Options*. If the data contain a large amount of sampling error, smoothing the time series will cause the points to move more smoothly as time is changed.

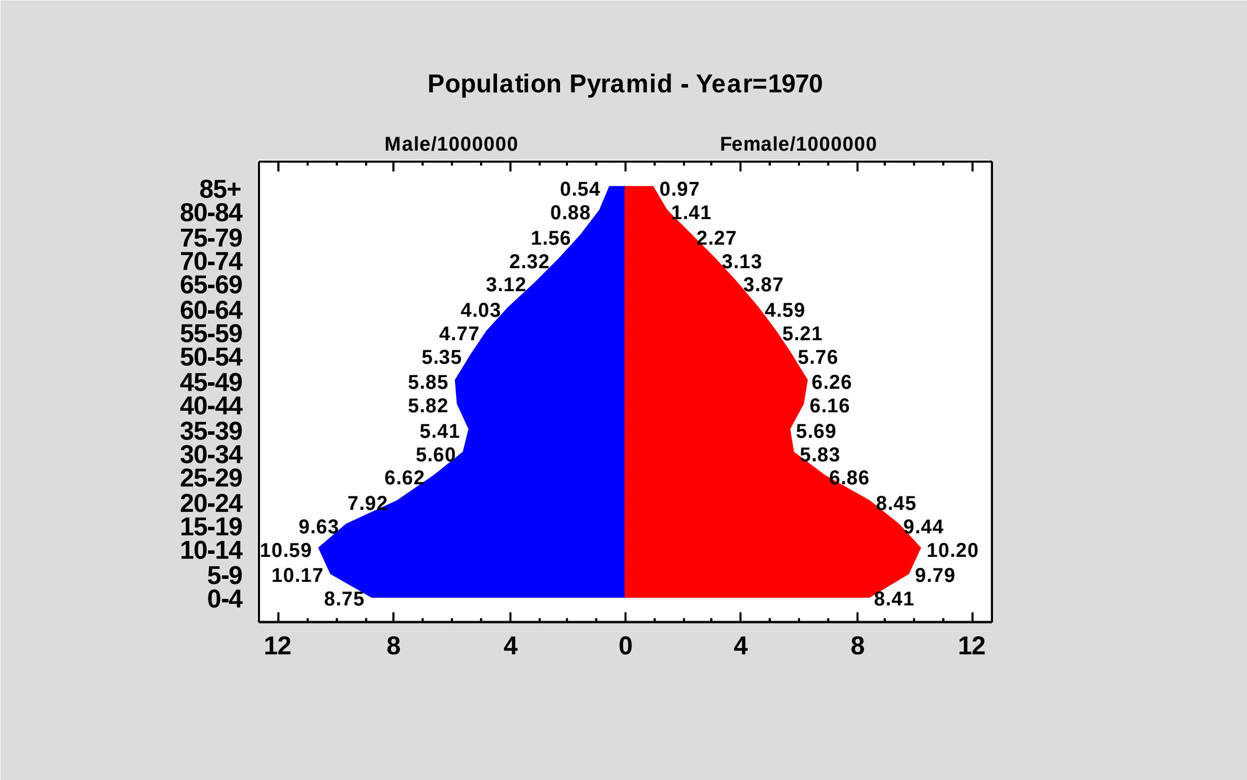
- **Missing values**: specifies how missing values should be treated. By default, missing values are not plotted, so that bars may appear and disappear as time changes. Selecting *Same as previous* will cause missing values to be replaced with the closest previous value which is not missing, which will cause bars to pause in one place but not disappear. Interpolation fills in missing values using an interpolation of 4 adjacent values, as described in the *Calculations* section of this document. If the data are seasonal, indicate the length of seasonality *s* to be used in the interpolation (for seasonal monthly data, *s* = 12). For nonseasonal data, *s* = 1.

The plot below shows the output for the final time period, replacing missing values with the previous values and adding labels.



Graphics Options has been used to reduce the font size of the labels.

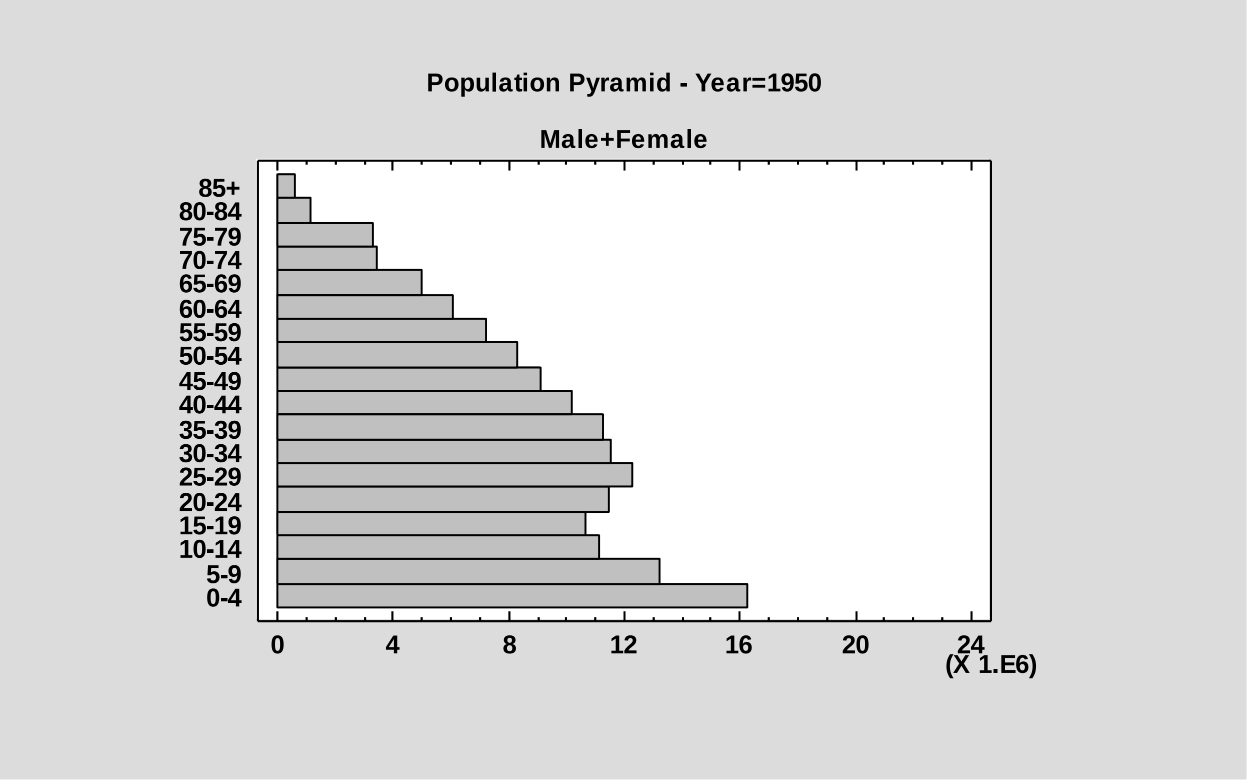
If *Polygon* is chosen for the plot type, the bars will be replaced by lines connecting the data values.


Population Pyramid - Year=1970
Male/1000000
Female/1000000
85+
80-84
75-79
70-74
65-69
60-64
55-59
50-54
45-49
40-44
35-39
30-34
25-29
20-24
15-19
10-14
5-9
0-4
0.54
0.88
1.56
2.32
3.12
4.03
4.77
5.35
5.85
5.82
5.41
5.60
6.62
7.92
9.63
10.59
10.17
8.75
0.97
1.41
2.27
3.13
3.87
4.59
5.21
5.76
6.26
6.16
5.69
5.83
6.86
8.45
9.44
10.20
9.79
8.41
12
8
4
0
4
8
12

]

Plotting One Sample Only

Although population pyramids are usually created to compare the distribution of 2 samples, the Statlet only requires one sample. It then creates a one-sided image as illustrated below:



Calculations

The *interpolation* method may be used to replace a limited number of missing values in each time series, provided there are not too many missing values close together. Before the data is analyzed, missing values are replaced by interpolated values, determined using the following rule:

1. If y_t, the observation at time *t*, is missing, find the two observations in the same season that precede time *t* (y_{t-s} and y_{t-2s}) and the two observations in the same season that come after time *t* (y_{t+s} and y_{t+2s}).

2. If none of the four observations are missing, then the replacement value for y_t is:

$$y_t = \frac{-3y_{t-2s} + 12y_{t-s} + 12y_{t+s} - 3y_{t+2s}}{18} \qquad (1)$$

3. If y_{t+2s} is missing but the other three are not, then the replacement value for y_t is:

$$y_t = \frac{-y_{t-2s} + 3y_{t-s} + y_{t+s}}{3} \qquad (2)$$

4. If y_{t+s} is missing but the other three are not, then the replacement value for y_t is:

$$y_t = \frac{-3y_{t-2s} + 8y_{t-s} + y_{t+s}}{6} \qquad (3)$$

5. If y_{t-s} is missing but the other three are not, then the replacement value for y_t is:

$$y_t = \frac{y_{t-2s} + 8y_{t+s} - 3y_{t+2s}}{6} \qquad (4)$$

6. If y_{t-2s} is missing but the other three are not, then the replacement value for y_t is:

$$y_t = \frac{y_{t-s} + 3y_{t+s} - y_{t+s}}{3} \qquad (5)$$

7. If y_{t+s} and y_{t+2s} are missing but the other two are not, then the replacement value for y_t is:

$$y_t = -y_{t-2s} + 2y_{t-s} \qquad (6)$$

8. If $y_{t\text{-}s}$ and y_{t+2s} are missing but the other two are not, then the replacement value for y_t is:

$$y_t = \frac{y_{t-2s} + 2y_{t+s}}{3} \tag{7}$$

9. If $y_{t\text{-}s}$ and y_{t+s} are missing but the other two are not, then the replacement value for y_t is:

$$y_t = \frac{y_{t-2s} + y_{t+2s}}{2} \tag{8}$$

10. If $y_{t\text{-}2s}$ and y_{t+2s} are missing but the other two are not, then the replacement value for y_t is:

$$y_t = \frac{y_{t-s} + y_{t+s}}{2} \tag{9}$$

11. If $y_{t\text{-}2s}$ and y_{t+s} are missing but the other two are not, then the replacement value for y_t is:

$$y_t = \frac{2y_{t-s} + y_{t+2s}}{3} \tag{10}$$

12. If $y_{t\text{-}2s}$ and $y_{t\text{-}s}$ are missing but the other two are not, then the replacement value for y_t is:

$$y_t = 2y_{t+s} - y_{t+2s} \tag{11}$$

If more than 2 of the four observations are missing, the missing value will not be replaced.

The interpolated values are designed to perfectly reproduce a quadratic trend (if only one observation is missing) or a linear trend (if two observations are missing), provided no noise is present.