THE POWER TO KNOW.

# SAS/QC® 14.3 User's Guide

# The RAREEVENTS Procedure

This document is an individual chapter from *SAS/QC® 14.3 User's Guide*.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2017. *SAS/QC® 14.3 User's Guide*. Cary, NC: SAS Institute Inc.

**SAS/QC® 14.3 User's Guide**

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

September 2017

# Chapter 17
# The RAREEVENTS Procedure

**Contents**

## Overview: RAREEVENTS Procedure

The RAREEVENTS procedure produces control charts for rare events. A control chart is a graphical and analytical tool for detecting unusual variation in a process and deciding whether the process is stable and predictable. A rare event is one that occurs infrequently, with a low probability.

In this chapter, a control chart for rare events is referred to as a *rare events chart*. The data that are plotted in a rare events chart represent the times between successive events. Usually these are adverse events that are unwanted outcomes in a process, such as an incorrectly recorded bank deposit, a patient falling in a hospital, or a chemical spill. Rare events charts have gained acceptance in health care quality improvement applications because of their ease of use and suitability to processes that have low defect rates (Benneyan 1999).

An important assumption for a rare events chart is that the events are independent. The occurrence of one event does not affect the probability that another will occur, and the probability of an occurrence is approximately constant over time. Rare events charts should not be used to monitor clusters of events, such as cases of a contagious disease, which violate this assumption. See Woodall (2006) for a thorough discussion of different control charts that are applicable to health care quality improvement.

The data for a rare events chart are often the times between consecutive events, such as the intervals between accidental needle sticks in a hospital. The intervals can be recorded as integer or continuous values. The opportunities for events to occur must be approximately constant over time. For example, the number of times that needles are handled should be about the same each day if you are monitoring the number of days between accidental sticks. Alternatively, the data can be explicit counts of opportunities for occurrence that come between events, such as the number of surgeries performed between occurrences of postsurgical infection. These kinds of data are preferable but often are not available.

A rare events chart has two decision limits: an upper probability limit (UPL) and a lower probability limit (LPL). By default, these are based on a geometric distribution for integer data and an exponential distribution for continuous data. A data value that is greater than the UPL or less than the LPL signals unusual variation in the process. A value that is greater than the UPL indicates that the time between events might be increasing, in which case the events are occurring less frequently. Because the events of interest are usually adverse, this can signal an improvement in the process. Conversely, a value less than the LPL indicates that events are occurring more frequently, which can signal a decline in the process.

You can use the RAREEVENTS procedure to do the following:

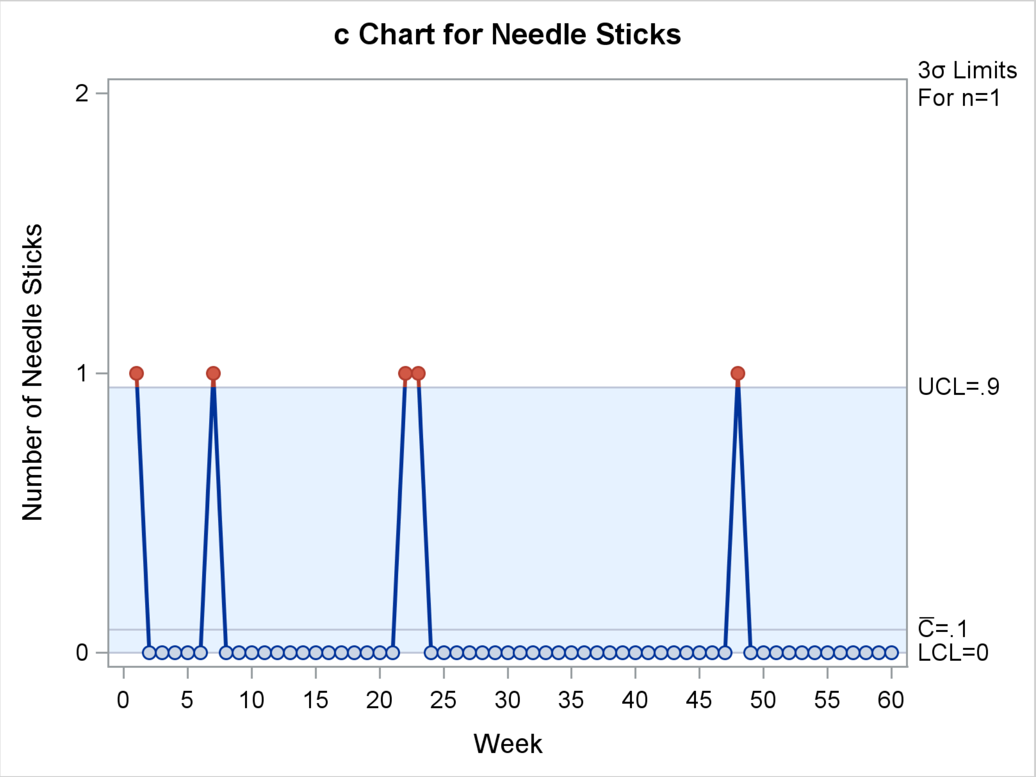
- produce a rare events chart with probability limits that are computed from the data
- create a graph that you can use to compare the distribution of the input data with a reference probability distribution
- specify the probability distribution that is used to compute the probability limits or to compare with the input data
- create a rare events chart that displays distinct sets of probability limits for multiple time phases
- save probability limits in an output data set
- produce a rare events chart that uses preestablished probability limits that are read from a data set
- save process measurements, probability limits, and probability distribution information in an output data set

## Rare Events Charts and *c* Charts

The traditional control chart that is most comparable to a rare events chart is the *c* chart, which is used to monitor counts of unwanted process outcomes. (See the section "CCHART Statement: SHEWHART Procedure" on page 1484 for a detailed description of *c* charts and how to produce them by using the SHEWHART procedure.) However, as explained by Kaminsky et al. (1992), Benneyan (2001a), and others, *c* charts and other traditional control charts do not always perform well when used to monitor rare events.

Figure 17.1 shows a *c* chart of needle sticks per week in a hospital. Weeks are identified on the horizontal axis, and the weekly counts of needle sticks are plotted. The control limits are used to detect unusual variation in the number of needle sticks.

**Figure 17.1** *c* Chart for Needle Sticks per Week

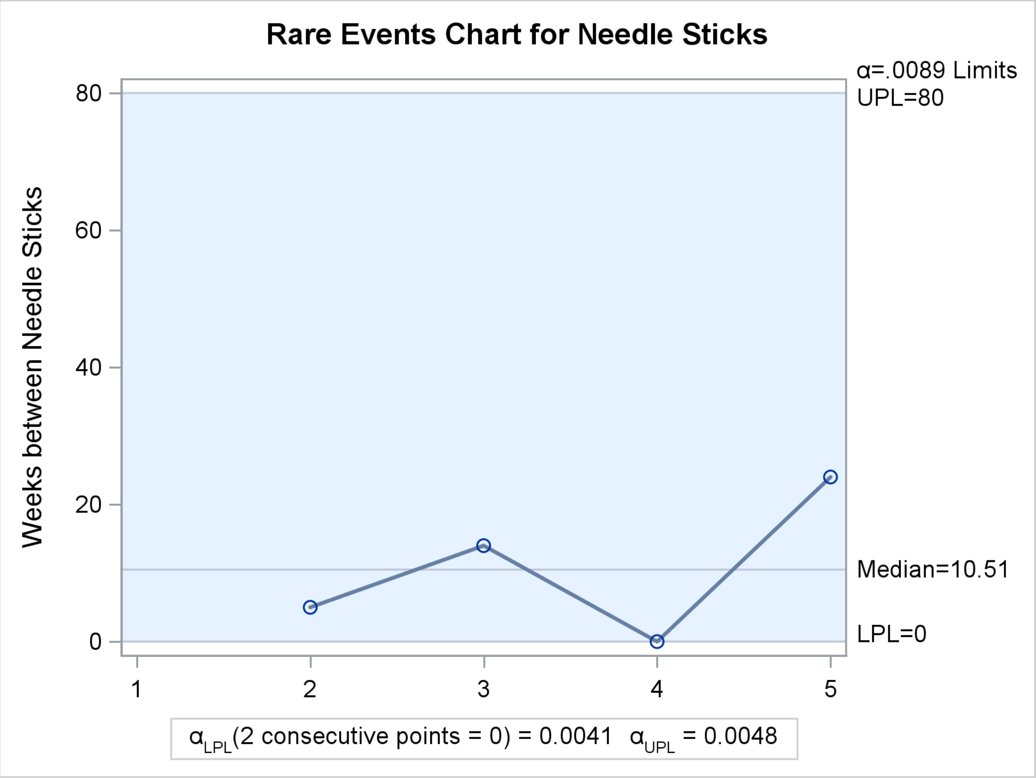


In this case needle sticks are truly rare. Almost all the weekly counts are 0, no week has more than 1, and the mean count is very low. Because the upper control limit (UCL) is less than 1, each individual needle stick signals unusual variation. This *c* chart might be too sensitive to provide useful information.

To address this problem, you could increase the counts (and therefore the UCL) by increasing the length of the time periods over which you accumulate the counts. In this case, grouping the needle stick counts into 15 four-week periods produces a *c* chart with UCL $= 2.07$ and a maximum count of 2, and no unusual variation is signaled. A drawback of this approach is that data are available for analysis only every four weeks, so a change in the process might not be detected quickly. Another possibility would be to modify the way that the control limits are computed by basing them on a discrete distribution other than the Poisson distribution. A better alternative is to use a rare events chart.

Figure 17.2 shows a rare events chart that is used to plot the same needle stick data, which are transformed into weeks between sticks. Individual needle sticks are identified in order of occurrence on the horizontal axis. For each event, the time in weeks since the previous event is plotted.

**Figure 17.2** Rare Events Chart for Weeks between Needle Sticks



The rare events chart does not signal any unusual variation. There would be a signal if two consecutive data values were zero, indicating needle sticks in three consecutive weeks. For more information, see the section "Probability Limits Based on a Geometric Distribution" on page 1183.

When you use a rare events chart, you do not need to wait until the end of a reporting period or collect a large sample of data before plotting a point on the chart. Instead, you can add a point to the chart immediately when an event occurs. Therefore you can construct a useful chart in a more timely manner, which improves your chances of detecting process changes. Because the values that are plotted in a rare events chart are times between events, the simple occurrence of a single event will not signal unusual variation. In summary, a rare events chart is better suited than traditional control charts to detecting changes in the frequency of low-probability events.

## Getting Started: RAREEVENTS Procedure

This example illustrates the basic features of the RAREEVENTS procedure. The data are adapted from Benneyan (1998b). The following statements create a SAS data set named Infections by reading the dates of occurrences of an infectious disease and computing DaysBetween, the numbers of days between successive infections:

```
data Infections;
   input InfectionDate mmddyy10.;
   InfectionNumber = _n_;
   DaysBetween = InfectionDate - lag(InfectionDate);
   format InfectionDate  mmddyy10.;
datalines;
04/17/1995
04/17/1995
04/17/1995
04/19/1995
04/20/1995
05/03/1995
05/05/1995
05/05/1995
05/06/1995
05/07/1995
05/08/1995
05/09/1995
05/09/1995
05/10/1995
05/11/1995
05/27/1995
05/27/1995
05/28/1995
05/29/1995
05/31/1995
06/10/1995
06/11/1995
06/12/1995
06/14/1995
06/16/1995
06/16/1995
06/18/1995
06/21/1995
06/21/1995
;
```

Figure 17.3 shows a partial listing of the Infections data set.

**Figure 17.3** Partial Listing of the Infections Data Set

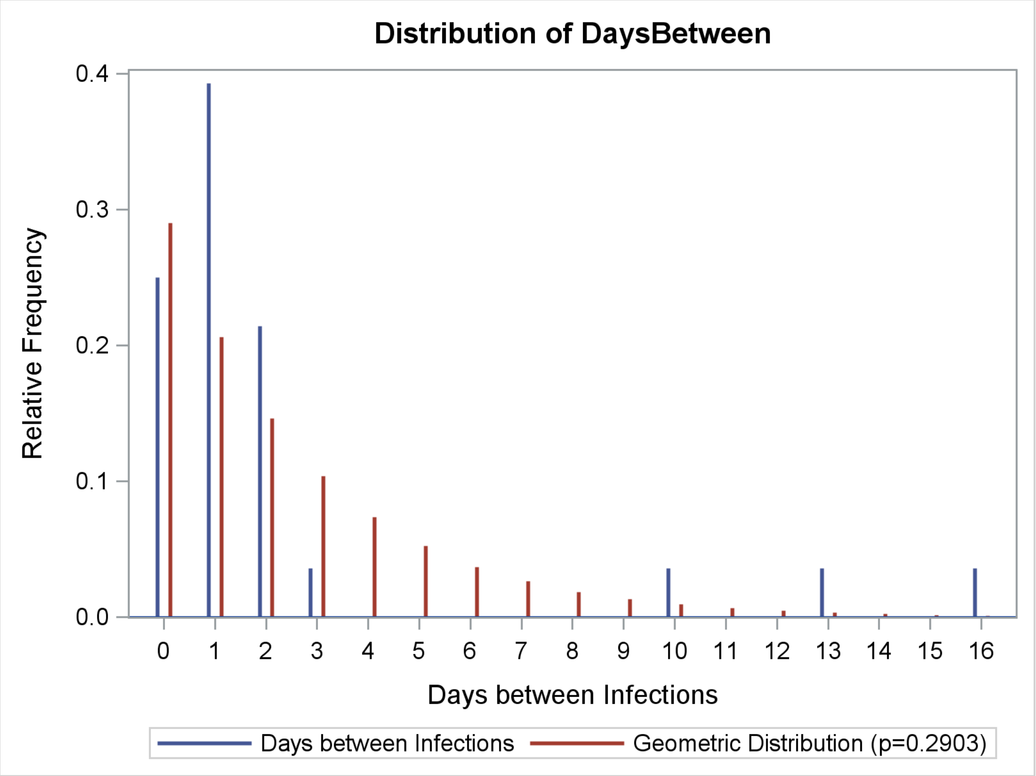
```
InfectionDate  InfectionNumber  DaysBetween
   04/17/1995                1            .
   04/17/1995                2            0
   04/17/1995                3            0
   04/19/1995                4            2
   04/20/1995                5            1
   05/03/1995                6           13
   05/05/1995                7            2
```

The following statements produce a comparison plot and a rare events chart for the variable DaysBetween. Because its values are integers, a geometric distribution is used by default to make the comparison and to compute the probability limits for the rare events chart. The value of parameter *p* for the geometric distribution is estimated from the data. InfectionNumber is an optional index variable whose values are used to label the rare event chart's horizontal axis.

```
ods graphics on;
proc rareevents data=Infections;
   compare DaysBetween;
   chart DaysBetween * InfectionNumber;
   label DaysBetween = 'Days between Infections';
run;
```

The ODS GRAPHICS ON statement enables ODS Graphics, which is necessary for the procedure to produce graphical output. The COMPARE statement produces the needle plot that is shown in Figure 17.4.

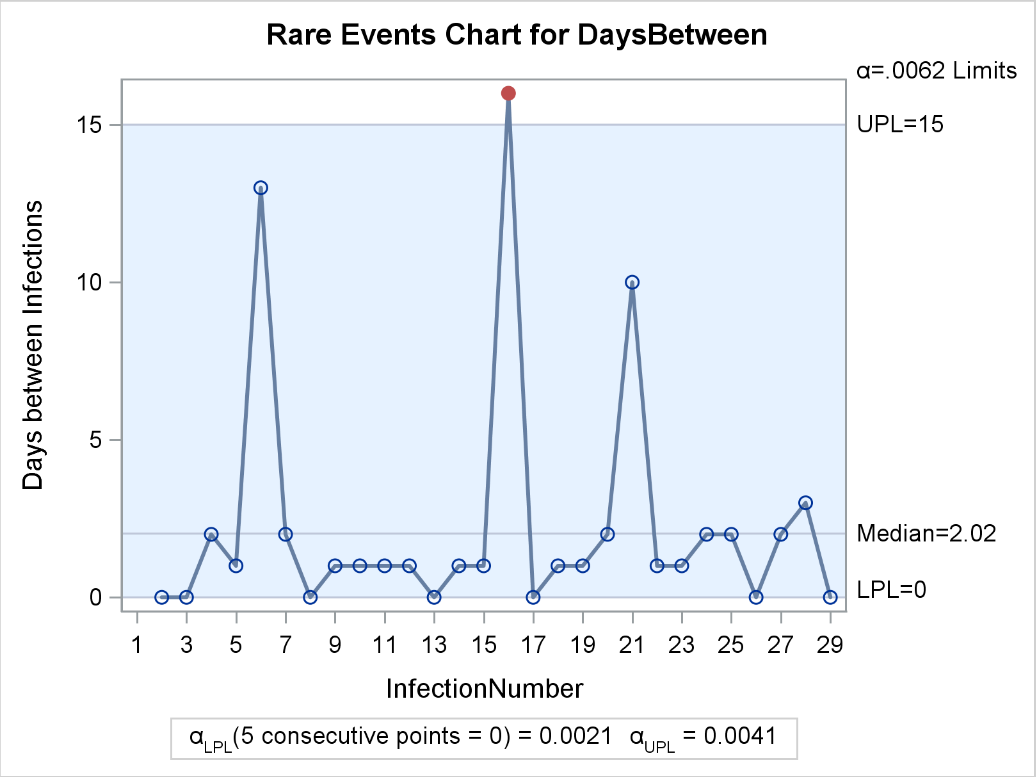
**Figure 17.4** Distribution of Days between Infections



Interpreting a comparison plot of a small data sample can be difficult, but the data have the same general shape as the geometric distribution. The graph does not indicate that the geometric distribution is *not* appropriate for these data.

Figure 17.5 shows the rare events chart of the DaysBetween data that the CHART statement produces.

**Figure 17.5** Rare Events Chart for Urinary Tract Infections



The number of days between infections 15 and 16 exceeds the UPL, signaling unusual variation. Here the unusual variation is welcome, because less frequent infections are desirable.

The median and probability limits for the chart are computed as described in the section "Constructing Rare Events Charts" on page 1182. The chart legend displays the probability, $\alpha_{\mathrm{UPL}}$, that a value from the geometric distribution is greater than the UPL. Note that the LPL in Figure 17.5 is equal to 0, which means that the probability of a DaysBetween value less than the LPL is 0. It is not unusual for the LPL to be equal to the minimum possible data value in a chart of integer data. When this is the case, the procedure checks for sequences of consecutive values equal to the LPL as an indication of unusual variation. The probability, $\alpha_{\mathrm{LPL}}$, of five consecutive 0 values from the geometric distribution is 0.0021, as indicated in the legend. The label outside the upper right corner of the chart shows the overall $\alpha = \alpha_{\mathrm{LPL}} + \alpha_{\mathrm{UPL}}$.

## Syntax: RAREEVENTS Procedure

**PROC RAREEVENTS** < *options* > ;
  **BY** *variables* ;
  **ID** *variables* ;
  **CHART** < / *options* > ;
  **COMPARE** < / *options* > ;

The following sections describe the PROC RAREEVENTS statement and then describe the other statements in alphabetical order.

## PROC RAREEVENTS Statement

**PROC RAREEVENTS** < *options* > **;**

The PROC RAREEVENTS statement invokes the RAREEVENTS procedure and specifies the input data sets. You can specify the following *options*:

**DATA=***SAS-data-set*

specifies an input SAS data set that contains process data, which are measurements of times between events. You cannot specify the TABLE= option together with the DATA= option. For more information about DATA= data sets, see the section "DATA= Data Set" on page 1186.

**LIMITS=***SAS-data-set*

specifies an input SAS data set that contains probability limits for a rare events chart.

**TABLE=***SAS-data-set*

specifies an input SAS data set that contains summary information from a rare events chart. You can produce a TABLE= data set by specifying the OUTTABLE= option in a CHART statement. You can use a TABLE= input data set to display a previously computed rare events chart. You cannot specify the DATA= option together with the TABLE= option. For more information, see the section "TABLE= Data Set" on page 1188.

## BY Statement

**BY** *variables* **;**

You can specify a BY statement with PROC RAREEVENTS to obtain separate analyses of observations in groups that are defined by the BY variables. When a BY statement appears, the procedure expects the input data set to be sorted in order of the BY variables. If you specify more than one BY statement, only the last one specified is used.

If your input data set is not sorted in ascending order, use one of the following alternatives:

- Sort the data by using the SORT procedure with a similar BY statement.
- Specify the NOTSORTED or DESCENDING option in the BY statement for the RAREEVENTS procedure. The NOTSORTED option does not mean that the data are unsorted but rather that the data are arranged in groups (according to values of the BY variables) and that these groups are not necessarily in alphabetical or increasing numeric order.
- Create an index on the BY variables by using the DATASETS procedure (in Base SAS software).

For more information about BY-group processing, see the discussion in *SAS Language Reference: Concepts*. For more information about the DATASETS procedure, see the discussion in the *SAS Visual Data Management and Utility Procedures Guide*.

## ID Statement

**ID** *variables* **;**

The values of the ID *variables* are displayed in tooltips associated with points on a rare events chart when you create HTML output and specify the IMAGEMAP option in the ODS GRAPHICS statement. For more information, see Chapter 21, "Statistical Graphics Using ODS" (*SAS/STAT User's Guide*).

## CHART Statement

**CHART** *process-variable* < $*$ *index-variable* > < / *options* > **;**

The CHART statement produces a rare events chart. The *process-variable* contains measurements of times between events. You can use the optional *index-variable* to label the tick marks on the chart's horizontal axis. When you do not specify an index variable, the tick marks on the horizontal axis are numbered sequentially, starting with 1.

Table 17.1 summarizes the *options* available in the CHART statement.

**Table 17.1** CHART Statement Options

| Option | Description |
|---|---|
| ALPHALPL= | Specifies the probability that is used to compute the lower probability limit |
| ALPHAUPL= | Specifies the probability that is used to compute the upper probability limit |
| DIST= | Specifies the distribution that is used to compute probability limits |
| EXCHART | Displays a chart only if it has points outside the probability limits |
| HAXISLABEL= | Specifies a horizontal axis label for the chart |
| LIMITPHASES= | Specifies the phases for which probability limits are read from the LIMITS= data set |
| NOCHART | Suppresses creation of the rare events chart |
| NOHLABEL | Suppresses the horizontal axis label of the chart |
| NOPHASEREF | Suppresses the vertical reference lines that separate phases |
| NOPHASEREFFILL | Suppresses graph wall fills for phases |
| NOVLABEL | Suppresses the vertical axis label of the chart |
| NPANELPOS= | Specifies the number of horizontal axis plotting positions per panel |
| ODSFOOTNOTE= | Adds a footnote to the chart |
| ODSFOOTNOTE2= | Adds a secondary footnote to the chart |
| ODSTITLE= | Specifies a title for the chart |
| ODSTITLE2= | Specifies a secondary title for the chart |
| OUTLIMITS= | Creates a SAS data set that contains probability limits for the chart |
| OUTTABLE= | Creates a SAS data set that contains a summary of the rare events chart |
| PHASELEGEND | Displays phase labels in a legend across the top of the chart |
| PHASELIMITS | Labels probability limits and center lines with their values within each phase |

**Table 17.1** *(continued)*

| Option | Description |
|---|---|
| READPHASES= | Selects phases from the DATA= or TABLE= data set for processing |
| TOTPANELS= | Specifies the number of panels that are used to display the chart |

You can specify the following options only in the CHART statement. For detailed descriptions of options common to the CHART and COMPARE statements, see the section "Common CHART and COMPARE Statement Options" on page 1180.

**ALPHALPL=$\alpha_{\mathrm{LPL}}$**

specifies the probability ($0 < \alpha_{\mathrm{LPL}} < 1$) that is used to compute the lower probability limit (LPL) for the rare events chart, based on the probability distribution that you specify in the DIST= option. The LPL is computed so that the probability of a measurement from the distribution being less than the LPL is $\alpha_{\mathrm{LPL}}$. By default, $\alpha_{\mathrm{LPL}} = 0.005$.

With a discrete probability distribution, it is not possible in general to compute a LPL for which this probability is exactly $\alpha_{\mathrm{LPL}}$. In that case, the chart includes a legend that shows the $\alpha_{\mathrm{LPL}}$ that corresponds to the computed LPL.

**ALPHAUPL=$\alpha_{\mathrm{UPL}}$**

specifies the probability ($0 < \alpha_{\mathrm{UPL}} < 1$) that is used to compute the upper probability limit (UPL) for the rare events chart, based on the probability distribution that you specify in the DIST= option. The UPL is computed so that the probability of a measurement from the distribution being greater than the UPL is $\alpha_{\mathrm{UPL}}$. By default, $\alpha_{\mathrm{UPL}} = 0.005$.

With a discrete probability distribution, it is not possible in general to compute a UPL for which this probability is exactly $\alpha_{\mathrm{UPL}}$. In that case, the chart includes a legend that shows the $\alpha_{\mathrm{UPL}}$ that corresponds to the computed UPL.

**EXCHART< (LOWER | UPPER) >**

specifies that a rare events chart be displayed only when one or more measurements lie outside the probability limits. If you specify EXCHART(LOWER), then a chart is displayed only when a measurement is less than the lower probability limit. If you specify EXCHART(UPPER), then a chart is displayed only when a measurement is greater than the upper probability limit.

**LIMITPHASES=*value* | ALL**

reads probability limits for one or more phases from the LIMITS= data set.

If you specify LIMITPHASES=*value*, a single set of limits is read from the first observation in the LIMITS= data set (see Table 17.4) for which the following are true:

- The value of _VAR_ matches the process variable name.
- The value of _INDEX_ matches the index variable name, if an index variable is specified in the CHART statement.
- The value of _PHASE_ matches *value*.

If you specify LIMITPHASES=ALL, a set of limits is read for each phase that is specified by the READPHASES= option. The limits for a phase are read from the first observation in the LIMITS= data set for which the following are true:

- The value of _VAR_ matches the process variable name.
- The value of _INDEX_ matches the index variable name, if an index variable is specified in the CHART statement.
- The value of _PHASE_ matches the value of the variable _PHASE_ from the DATA= or TABLE= data set.

If you do not specify the LIMITPHASES= option, then a single set of probability limits is read from the first observation in the LIMITS= data set for which the value of _VAR_ matches the process variable name and the value of _INDEX_ matches the index variable name, if an index variable is specified.

Example 17.2 shows how the LIMITPHASES= and READPHASES= options are used together.

**NOCHART**

suppresses display of the rare events chart. You can use the NOCHART option together with the OUTLIMITS= or OUTTABLE= option to create output data sets without displaying a chart.

**NOPHASEREF**

suppresses phase reference lines. By default, the boundaries between phases are marked by vertical phase reference lines.

**NOPHASEREFFILL**

suppresses graph wall fills for phases. By default, the graph walls for phases are filled with two alternating colors.

**NPANELPOS=***n*

**NPANEL=***n*

specifies the number of horizontal axis plotting positions per panel in the chart. You usually specify this option to display more points in a panel than the default number, which is 50.

You can specify a positive or negative value for *n*. The absolute value of *n* must be at least 5. If *n* is positive, the number of positions is adjusted so that it is approximately equal to *n* and so that all panels display approximately the same number of positions. If *n* is negative, then no balancing is done, and each panel (except possibly the last) displays approximately $|n|$ positions.

**OUTLIMITS=***SAS-data-set*

creates an output SAS data set that contains the probability limits and related information for the rare events chart. For more information about the OUTLIMITS= data set, see the section "OUTLIMITS= Data Set" on page 1189.

**OUTTABLE=***SAS-data-set*

creates an output SAS data set that contains the information plotted in the rare events chart, including the process measurements and the probability limits. For more information about the OUTTABLE= data set, see the section "OUTTABLE= Data Set" on page 1189.

**PHASELEGEND**

produces a legend across the top of the chart that labels each phase with the associated value of the _PHASE_ variable from the input data set.

**PHASELIMITS**

labels the probability limits and center line separately for each phase in the chart.

**READPHASES=***value-list* **| ALL**

selects blocks of consecutive observations to be read from the primary input (DATA= or TABLE=) data set. These blocks are called phases and are defined by the values of the variable _PHASE_, which must be a character variable whose length is no greater than 256.

If you specify READPHASES=*value-list*, only phases whose _PHASE_ values match a value in *value-list* are selected. If you specify READPHASES=ALL, all phases in the input data set are selected.

By default, a separate set of probability limits is computed for each phase. If you specify a LIMITS= data set, you can use the LIMITPHASES= option to read separate sets of limits for different phases.

If you do not specify the READPHASES= option, then the _PHASE_ variable is ignored and a chart without separate phases is produced.

Example 17.2 shows how you can use the READPHASES= option.

**TOTPANELS=***n*

specifies the number of panels that are used to display the chart. By default, the number of panels is determined by the value that you specify in the NPANELPOS= option. If you specify both the TOTPANELS= and NPANELPOS= options, the TOTPANELS= value takes precedence.

## COMPARE Statement

**COMPARE** *process-variable* < */ options* > **;**

The COMPARE statement produces a graph that compares the process data to a reference probability distribution. By default, integer data are displayed in a needle plot. Continuous data are displayed in a histogram. When the reference distribution is an exponential or Weibull distribution, the COMPARE statement also produces a table of goodness-of-fit statistics.

**Table 17.2** COMPARE Statement Options

| Option | Description |
|---|---|
| DIST= | Specifies the reference distribution that is compared to the sample |
| NBINS= | Specifies the number of bins that are used to display the data distribution |
| HAXISLABEL= | Specifies a horizontal axis label for a comparison chart |
| NOHLABEL | Suppresses the horizontal axis label of a comparison chart |
| NOVLABEL | Suppresses the vertical axis label of a comparison chart |
| ODSFOOTNOTE= | Adds a footnote to a comparison chart |
| ODSFOOTNOTE2= | Adds a secondary footnote to a comparison chart |
| ODSTITLE= | Specifies a title for a comparison chart |
| ODSTITLE2= | Specifies a secondary title for a comparison chart |
| PROCESS= | Specifies how integer process data are displayed in a comparison chart |

**Table 17.2** *(continued)*

| Option | Description |
|---|---|
| REFERENCE= | Specifies how an integer reference probability distribution is displayed in a comparison chart |

You can specify the following options only in a COMPARE statement. For detailed descriptions of options common to the CHART and COMPARE statements, see the section "Common CHART and COMPARE Statement Options" on page 1180.

**NBINS=*n***

specifies the number of bins that are used to display the process data in a comparison plot. For integer data, the default number of bins in the comparison plot is

$$\min(\max(n_{\max} - a + 1, 15), 50)$$

where $n_{\max}$ is the maximum data value and $a$ is the minimum possible data value. For continuous data, the default number of histogram bins is based on the data range and the number of observations, using the method of Terrell and Scott (1985).

**PROCESS=BAR | MARKER | NEEDLE**

specifies how integer process data are displayed in a comparison chart. You can specify the following keywords:

| | |
|---|---|
| **BAR** | displays the process data by using bars. |
| **MARKER** | plots the process data by using markers. |
| **NEEDLE** | displays the process data by using needles. |

By default, PROCESS=NEEDLE. The PROCESS= option has no effect on a comparison chart when a continuous reference distribution is in effect.

**REFERENCE=BAR | MARKER | NEEDLE**

specifies how an integer reference data distribution is displayed in a comparison chart. You can specify the following values:

| | |
|---|---|
| **BAR** | displays the reference data distribution by using bars. |
| **MARKER** | displays the reference data distribution by using markers. |
| **NEEDLE** | displays the reference data distribution by using needles. |

By default, REFERENCE=NEEDLE. The REFERENCE= option has no effect on a comparison chart when a continuous reference distribution is in effect.

## Common CHART and COMPARE Statement Options

You can specify the following *options* after a slash (/) in a CHART or COMPARE statement.

**DIST=***distribution*

specifies the probability distribution that is compared to the input data by a COMPARE statement and that is used to compute probability limits for a rare events chart that you create by using a CHART statement. You can specify the following distributions:

**EXPONENTIAL< (***exponential-options***) >**

requests an exponential distribution. You can specify the following *exponential-options*:

**SIGMA=**$\sigma$

specifies the scale parameter for the exponential distribution. By default, $\sigma$ is estimated from the process data.

**THETA=**$\theta$ **| EST**

specifies the threshold parameter for the exponential distribution. By default, $\theta = 0$. The specified value must be greater than or equal to 0. You can specify THETA=EST to compute an estimate of $\theta$ from the process data. If any data value is less than $\theta$, the procedure issues a warning and sets $\theta$ to the minimum data value.

**GEOMETRIC< (***geometric-options***) >**

requests a geometric distribution. You can specify the following *geometric-options*:

**P=***p* **| MLE | MVUE**

specifies the probability of success in a single Bernoulli trial on which the geometric distribution is based. This is the probability that an opportunity for a rare event to occur will actually result in an occurrence. You can specify P=MLE to compute a maximum likelihood estimate (MLE) of *p* or P=MVUE to compute a minimum variance unbiased estimate (MVUE) of *p*. By default, an MVUE is computed if the SHIFT= parameter value is 0 or 1, and an MLE is computed otherwise.

**SHIFT=***a*

specifies the minimum possible value ($a \geq 0$) for the geometric distribution. By default, $a = 0$. If a measurement from the input data represents the time *until* an event occurs (including the event itself) instead of times *between* events, then you should specify $a = 1$. If any data value is less than $a$, the procedure issues a warning and sets $a$ to the minimum data value.

**WEIBULL< (***weibull-options***) >**

requests a Weibull distribution. You can specify the following *weibull-options*:

**C=***c*

specifies the shape parameter for the Weibull distribution. By default, *c* is estimated from the process data.

**SIGMA=**$\sigma$

specifies the scale parameter for the exponential distribution. By default, $\sigma$ is estimated from the process data.

**THETA=**$\theta$ **| EST**

specifies the threshold parameter for the Weibull distribution. By default, $\theta = 0$. The specified value must be greater than or equal to 0. You can specify THETA=EST to compute an estimate of $\theta$ from the process data. If any data value is less than $\theta$, the procedure issues a warning and sets $\theta$ to the minimum data value.

The procedure determines whether the process data have continuous or integer values. By default, an exponential distribution is used for continuous data and a geometric distribution is used for integer data.

**HAXISLABEL='***label***'**

specifies a label for the horizontal axis of the graph.

**NOHLABEL**

suppresses the horizontal axis label in the graph.

**NOVLABEL**

suppresses the vertical axis label in the graph.

**ODSFOOTNOTE=FOOTNOTE | FOOTNOTE1 | '***string***'**

adds a footnote to the graph. If you specify the FOOTNOTE (or FOOTNOTE1) keyword, the value of the SAS FOOTNOTE statement is used as the graph footnote. If you specify a quoted string, that string is used as the footnote. The quoted string can contain the following escape characters, which are replaced by the values indicated:

| | |
|---|---|
| \n | is replaced by the process variable name. |
| \l | is replaced by the process variable label (or name if the process variable has no label). |

**ODSFOOTNOTE2=FOOTNOTE2 | '***string***'**

adds a secondary footnote to the graph. If you specify the FOOTNOTE2 keyword, the value of the SAS FOOTNOTE2 statement is used as the secondary graph footnote. If you specify a quoted string, that string is used as the secondary footnote. The quoted string can contain the following escape characters, which are replaced by the values indicated:

| | |
|---|---|
| \n | is replaced by the process variable name. |
| \l | is replaced by the process variable label (or name if the process variable has no label). |

**ODSTITLE=TITLE | TITLE1 | NONE | DEFAULT | '***string***'**

specifies a title for the graph. You can specify the following values:

| | |
|---|---|
| **TITLE** (or **TITLE1**) | uses the value of the SAS TITLE statement as the graph title. |
| **NONE** | suppresses all graph titles. |

**DEFAULT** uses the default title.

If you specify a quoted string, that string is used as the graph title. The quoted string can contain the following escape characters, which are replaced by the values indicated:

\n is replaced by the process variable name.

\l is replaced by the process variable label (or name if the analysis variable has no label).

**ODSTITLE2=TITLE2 | '*string*'**

specifies a secondary title for the graph. If you specify the TITLE2 keyword, the value of the SAS TITLE2 statement is used as the secondary graph title. If you specify a quoted string, that string is used as the secondary title. The quoted string can contain the following escape characters, which are replaced by the values indicated:

\n is replaced by the process variable name.

\l is replaced by the process variable label (or name if the analysis variable has no label).

# Details: RAREEVENTS Procedure

## Constructing Rare Events Charts

Each point on the rare events chart indicates the value of an individual measurement from the input data set. You compute the lower probability limit (LPL), median, and upper probability limit (UPL) by solving for their values in the following equations, which use the cumulative distribution function (cdf) of the probability distribution that you specify in the DIST= option:

- $\mathrm{cdf}(\mathrm{LPL}) = \alpha_{\mathrm{LPL}}$
- $\mathrm{cdf}(\mathrm{median}) = 0.5$
- $\mathrm{cdf}(\mathrm{UPL}) = 1 - \alpha_{\mathrm{UPL}}$

### Probability Limits Based on an Exponential Distribution

The cumulative distribution function of an exponential distribution with scale parameter $\sigma$ and threshold parameter $\theta$ is

$$\mathrm{cdf}(x) = 1 - \exp\left(-\frac{(x-\theta)}{\sigma}\right)$$

Solving the equations listed previously, the median and probability limits values are as follows:

- $\mathrm{LPL} = \theta - \sigma \ln(1 - \alpha_{\mathrm{LPL}})$
- $\mathrm{median} = \theta + \sigma \ln(2)$
- $\mathrm{UPL} = \theta - \sigma \ln(\alpha_{\mathrm{UPL}})$

## Probability Limits Based on a Geometric Distribution

The cumulative distribution function of a geometric distribution with shift parameter $a$ and probability $p$ is

$$\mathrm{cdf}(x) = 1 - (1 - p)^{x-a+1}$$

Because the geometric distribution is used with integer data, meaningful probability limits must have integer values. Therefore the solutions to the equations listed previously are:

- $\mathrm{LPL} = \left\lfloor \frac{\ln(1-\alpha_{\mathrm{LPL}})}{\ln(1-p)} + a \right\rfloor$
- $\mathrm{median} = \frac{\ln(0.5)}{\ln(1-p)} + a$
- $\mathrm{UPL} = \left\lceil \frac{\ln(\alpha_{\mathrm{UPL}})}{\ln(1-p)} + a - 1 \right\rceil$

The probability of a value from the distribution being greater than the UPL is as close as possible to $\alpha_{\mathrm{UPL}}$ without exceeding it, and the probability of a value from the distribution being less than the LPL is as close as possible to $\alpha_{\mathrm{LPL}}$ without exceeding it. The $\alpha_{\mathrm{UPL}}$ and $\alpha_{\mathrm{LPL}}$ values that correspond to the computed limits are displayed in a legend on the rare events chart.

With integer probability limits, it is not unusual for the computed LPL to be equal to the minimum possible data value, so that no data value can be less than the LPL. In that case, the following value is computed:

$$m = \left\lceil \frac{\ln(\alpha_{\mathrm{LPL}})}{\ln(p)} \right\rceil$$

The probability of a sequence of $m$ consecutive values from the geometric distribution each being equal to the LPL is as close to $\alpha_{\mathrm{LPL}}$ as possible without exceeding it. The RAREEVENTS procedure flags any sequence of $m$ consecutive measurements equal to the LPL as a sign of unusual variation.

## Probability Limits Based on a Weibull Distribution

The cumulative distribution function of a Weibull distribution with scale parameter $\sigma$, shape parameter $c$, and threshold parameter $\theta$ is

$$\mathrm{cdf}(x) = 1 - \exp\left(-\left(\frac{(x-\theta)}{\sigma}\right)^{c}\right)$$

This produces the following probability limits:

- $\mathrm{LPL} = \theta + \sigma \left(-\ln(1 - \alpha_{\mathrm{LPL}})\right)^{1/c}$

- $\text{median} = \theta + \sigma\,(\ln(2))^{1/c}$
- $\text{UPL} = \theta + \sigma\,(-\ln(\alpha_{\text{UPL}}))^{1/c}$

## EDF Goodness-of-Fit Tests

When a continuous reference distribution is in effect, the COMPARE statement provides a series of goodness-of-fit tests based on the empirical distribution function (EDF). For a thorough discussion, see D'Agostino and Stephens (1986).

The empirical distribution function is defined for a set of $n$ independent observations $X_1, \ldots, X_n$ with a common distribution function $F(x)$. Denote the observations ordered from smallest to largest as $X_{(1)}, \ldots, X_{(n)}$. The empirical distribution function, $F_n(x)$, is defined as

$$\begin{array}{lll} F_n(x) = 0, & x < X_{(1)} & \\ F_n(x) = \frac{i}{n}, & X_{(i)} \le x < X_{(i+1)} & i = 1, \ldots, n-1 \\ F_n(x) = 1, & X_{(n)} \le x & \end{array}$$

Note that $F_n(x)$ is a step function that takes a step of height $\frac{1}{n}$ at each observation. This function estimates the distribution function $F(x)$. At any value $x$, $F_n(x)$ is the proportion of observations less than or equal to $x$, while $F(x)$ is the probability of an observation less than or equal to $x$. EDF statistics measure the discrepancy between $F_n(x)$ and $F(x)$.

The computational formulas for the EDF statistics make use of the probability integral transformation $U = F(X)$. If $F(X)$ is the distribution function of $X$, the random variable $U$ is uniformly distributed between 0 and 1.

Given $n$ observations $X_{(1)}, \ldots, X_{(n)}$, the values $U_{(i)} = F(X_{(i)})$ are computed by applying the transformation, as shown in the following sections.

The COMPARE statement provides three EDF tests:

- Kolmogorov-Smirnov
- Anderson-Darling
- Cramér-von Mises

These tests are based on various measures of the discrepancy between the empirical distribution function $F_n(x)$ and the reference parametric cumulative distribution function $F(x)$.

The following sections provide formal definitions of the EDF statistics.

### Kolmogorov-Smirnov Statistic

The Kolmogorov-Smirnov statistic (D) is defined as

$$D = \sup_x |F_n(x) - F(x)|$$

The Kolmogorov-Smirnov statistic belongs to the supremum class of EDF statistics. This class of statistics is based on the largest vertical difference between $F(x)$ and $F_n(x)$.

The Kolmogorov-Smirnov statistic is computed as the maximum of $D^+$ and $D^-$, where $D^+$ is the largest vertical distance between the EDF and the distribution function when the EDF is greater than the distribution function, and $D^-$ is the largest vertical distance when the EDF is less than the distribution function.

$$\begin{aligned} D^+ &= \max_i \left( \tfrac{i}{n} - U_{(i)} \right) \\ D^- &= \max_i \left( U_{(i)} - \tfrac{i-1}{n} \right) \\ D &= \max \left( D^+, D^- \right) \end{aligned}$$

## Anderson-Darling Statistic

The Anderson-Darling statistic and the Cramér-von Mises statistic belong to the quadratic class of EDF statistics. This class of statistics is based on the squared difference $(F_n(x) - F(x))^2$. Quadratic statistics have the following general form:

$$Q = n \int_{-\infty}^{+\infty} (F_n(x) - F(x))^2 \, \psi(x) dF(x)$$

The function $\psi(x)$ weights the squared difference $(F_n(x) - F(x))^2$.

The Anderson-Darling statistic ($A^2$) is defined as

$$A^2 = n \int_{-\infty}^{+\infty} (F_n(x) - F(x))^2 \left[ F(x)\,(1 - F(x)) \right]^{-1} dF(x)$$

Here the weight function is $\psi(x) = [F(x)\,(1 - F(x))]^{-1}$.

The Anderson-Darling statistic is computed as

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} \left[ (2i - 1) \log U_{(i)} + (2n + 1 - 2i) \log \left( \{ 1 - U_{(i)} \right) \right]$$

## Cramér-von Mises Statistic

The Cramér-von Mises statistic ($W^2$) is defined as

$$W^2 = n \int_{-\infty}^{+\infty} (F_n(x) - F(x))^2 \, dF(x)$$

Here the weight function is $\psi(x) = 1$.

The Cramér-von Mises statistic is computed as

$$W^2 = \sum_{i=1}^{n} \left( U_{(i)} - \frac{2i - 1}{2n} \right)^2 + \frac{1}{12n}$$

### Probability Values for EDF Tests

For the probability values ($p$-values) associated with the EDF test statistics, the RAREEVENTS procedure uses internal tables of probability levels similar to those given by D'Agostino and Stephens (1986). If the value is between two probability levels, then linear interpolation is used to estimate the probability value. The probability value depends upon the parameters that are known and the parameters that are estimated for the distribution you are fitting. Table 17.3 summarizes the combinations of estimated parameters for which EDF tests are available.

**Table 17.3** Availability of EDF Tests

| Distribution | Parameters | | | Tests Available |
|---|---|---|---|---|
| | Threshold | Scale | Shape | |
| Exponential | $\theta$ known, | $\sigma$ known | | all |
| | $\theta$ known | $\sigma$ unknown | | all |
| | $\theta$ unknown | $\sigma$ known | | all |
| | $\theta$ unknown | $\sigma$ unknown | | all |
| Weibull | $\theta$ known | $\sigma$ known | $c$ known | all |
| | $\theta$ known | $\sigma$ unknown | $c$ known | $A^2$ and $W^2$ |
| | $\theta$ known | $\sigma$ known | $c$ unknown | $A^2$ and $W^2$ |
| | $\theta$ known | $\sigma$ unknown | $c$ unknown | $A^2$ and $W^2$ |
| | $\theta$ unknown | $\sigma$ known | $c > 2$ known | all |
| | $\theta$ unknown | $\sigma$ unknown | $c > 2$ known | all |
| | $\theta$ unknown | $\sigma$ known | $c > 2$ unknown | all |
| | $\theta$ unknown | $\sigma$ unknown | $c > 2$ unknown | all |

## Input Data Sets

The RAREEVENTS procedure accepts a single primary input data set of either of two types:

- A DATA= data set contains process measurements to be analyzed.
- A TABLE= data set contains a summary of a rare events chart, which consists of the measurements, probability limits, and other information.

These options are mutually exclusive. If you do not specify an option that identifies a primary input data set, PROC RAREEVENTS uses the most recently created SAS data set as a DATA= data set. Valid process measurements are greater than or equal to zero. Missing and negative values are ignored.

You can also specify a LIMITS= data set that contains probability limits for a rare events chart.

### DATA= Data Set

A DATA= data set must include a process variable that contains measurements of the times between rare events. These measurements can be integers (for example, a count of days between events) or continuous values. In addition to the process variable, a DATA= data set can include the following:

- _PHASE_ variable, which is used by the READPHASES= option in the CHART statement
- BY variables
- ID variables
- index variable

The values of the optional index variable are used to label the horizontal axis tick marks on a rare events chart that is produced by a CHART statement. The _PHASE_ and index variables have no application in a COMPARE statement.

## LIMITS= Data Set

A LIMITS= data set contains probability limit information for a rare events chart. Usually, you create a LIMITS= data set by specifying the OUTLIMITS= option in a CHART statement. You can use a LIMITS= data set to specify historical probability limits for a process or custom probability limits that are computed by other means.

Table 17.4 lists the variables that a LIMITS= data set can contain.

**Table 17.4** LIMITS= Data Set Variables

| Variable | Description |
|---|---|
| _ALPHALPL_ | Probability associated with the lower probability limit |
| _ALPHAUPL_ | Probability associated with the upper probability limit |
| _C_ | Shape parameter for a Weibull distribution |
| _DIST_ | Name of the distribution used to compute the probability limits |
| _INDEX_ | Name of the optional index variable |
| _LPL_ | Lower probability limit |
| _MEDIAN_ | Median of the probability distribution |
| _P_ | Probability of success in a single Bernoulli trial on which a geometric distribution is based |
| _PARMEST_ | Specifies whether distribution parameters are estimated or specified |
| _PHASE_ | Phase associated with a set of probability limits |
| _SHIFT_ | Minimum possible value for a geometric distribution |
| _SIGMA_ | Scale parameter for an exponential or Weibull distribution |
| _THETA_ | Threshold parameter for an exponential or Weibull distribution |
| _UPL_ | Upper probability limit |
| _VAR_ | Name of the process variable that contains measurements of times between events |

A LIMITS= data set must contain the variables corresponding to the parameters of the distribution indicated by the value of the _DIST_ variable:

| | |
|---|---|
| EXPONENTIAL | _THETA_, _SIGMA_ |
| GEOMETRIC | _P_ |
| WEIBULL | _THETA_, _SIGMA_, _C_ |

The variable _PARMEST_ contains a code indicating whether the probability distribution parameters are specified or estimated. The _PARMEST_ code is the sum of codes for each parameter. If a parameter is specified, its code is zero. If a parameter is estimated, its code is as show in Table 17.5.

**Table 17.5** LIMITS= Data Set Variables

| Distribution | Parameter | Estimated Code |
|---|---|---|
| Exponential | $\theta$ | 1 |
| | $\sigma$ | 2 |
| Geometric | $p$ | 1 |
| Weibull | $\theta$ | 1 |
| | $\sigma$ | 2 |
| | $c$ | 4 |

For example, the _PARMEST_ value for a Weibull distribution with $\theta$ estimated, $\sigma$ specified, and $c$ estimated is $1 + 0 + 4 = 5$.

## TABLE= Data Set

A TABLE= data set contains a summary of a rare events chart. Usually, you create a TABLE= data set by specifying the OUTTABLE= option in a CHART statement. You can use a TABLE= data set to display a previously created rare events chart or to specify custom probability limits by computing your own _LPL_ and _UPL_ values.

Table 17.6 lists the variables that a TABLE= data set contains.

**Table 17.6** TABLE= Data Set Variables

| Variable | Description |
|---|---|
| _ALPHALPL_ | Probability associated with the lower probability limit |
| _ALPHAUPL_ | Probability associated with the upper probability limit |
| _DIST_ | Name of the distribution used to compute the probability limits |
| _EXLIM_ | Flag that indicates that a probability limit was exceeded |
| *index* | Optional index variable |
| _LPL_ | Lower probability limit |
| _MEDIAN_ | Median of the probability distribution |
| _PHASE_ | Phase to which an observation belongs |
| *process* | Process variable containing measurements of times between events |
| _UPL_ | Upper probability limit |

## Output Data Sets

### OUTLIMITS= Data Set

You can save probability limits and related information in an output data set by specifying the OUTLIMITS= option in a CHART statement. Table 17.7 lists the variables that an OUTLIMITS= data set can contain.

**Table 17.7** OUTLIMITS= Data Set Variables

| Variable | Description |
|---|---|
| _ALPHALPL_ | Probability associated with the lower probability limit |
| _ALPHAUPL_ | Probability associated with the upper probability limit |
| _C_ | Shape parameter for a Weibull distribution |
| _DIST_ | Name of the distribution used to compute the probability limits |
| _INDEX_ | Name of the optional index variable |
| _LPL_ | Lower probability limit |
| _MEDIAN_ | Median of the probability distribution |
| _P_ | Probability of success in a single Bernoulli trial on which a geometric distribution is based |
| _PARMEST_ | Specifies whether distribution parameters are estimated or specified |
| _PHASE_ | Phase associated with a set of probability limits |
| _SHIFT_ | Minimum possible value for a geometric distribution |
| _SIGMA_ | Scale parameter for an exponential or Weibull distribution |
| _THETA_ | Threshold parameter for an exponential or Weibull distribution |
| _UPL_ | Upper probability limit |
| _VAR_ | Name of the process variable that contains measurements of times between events |

When the probability limits are based on an exponential distribution, the OUTLIMITS= data set contains the variables _SIGMA_ and _THETA_. When the probability limits are based on a geometric distribution, the OUTLIMITS= data set contains the variables _P_ and _SHIFT_. When the probability limits are based on a Weibull distribution, the OUTLIMITS= data set contains the variables _C_, _SIGMA_ and _THETA_.

### OUTTABLE= Data Set

You can save process measurements, probability limits, and related information in an output data set by specifying the OUTTABLE= option in a CHART statement. Table 17.8 lists the variables that an OUTTABLE= data set contains.

**Table 17.8** OUTTABLE= Data Set Variables

| Variable | Description |
|---|---|
| _ALPHALPL_ | Probability associated with the lower probability limit |
| _ALPHAUPL_ | Probability associated with the upper probability limit |
| _DIST_ | Name of the distribution used to compute the probability limits |
| _EXLIM_ | Flag that indicates that a probability limit was exceeded |

**Table 17.8** *(continued)*

| Variable | Description |
|---|---|
| *index* | Optional index variable |
| _LPL_ | Lower probability limit |
| _MEDIAN_ | Median of the probability distribution |
| _PHASE_ | Phase to which an observation belongs |
| *process* | Process variable containing measurements of times between events |
| _UPL_ | Upper probability limit |

## ODS Table Names

PROC RAREEVENTS assigns a name to each table that it creates. You can use these names to refer to the tables when you use the Output Delivery System (ODS) to select tables and create output data sets. The ODS table names are listed in Table 17.9.

**Table 17.9** ODS Tables Produced by PROC RAREEVENTS

| ODS Table Name | Description | Statement | Option |
|---|---|---|---|
| GoodnessOfFit | Goodness-of-fit tests for fitted distribution | COMPARE | DIST=EXPONENTIAL (default for continuous data)<br>DIST=WEIBULL |

## ODS Graphics

Before you create ODS Graphics output, ODS Graphics must be enabled (for example, by using the ODS GRAPHICS ON statement). For more information about enabling and disabling ODS Graphics, see the section “Enabling and Disabling ODS Graphics” (Chapter 21, *SAS/STAT User’s Guide*).

The RAREEVENTS procedure assigns a name to each graph that it creates using ODS Graphics. You can use these names to refer to the graphs when you use ODS. The graph names are listed in Table 17.10.

**Table 17.10** ODS Graphics Produced by PROC RAREEVENTS

| ODS Graph Name | Plot Description | Statement or Option |
|---|---|---|
| RareEventsChart | Rare events chart of process data | CHART statement |
| ComparisonPlot | Comparison plot | COMPARE statement |

# Examples: RAREEVENTS Procedure

## Example 17.1: Monitoring Urinary Tract Infections

The data for this example are from Santiago and Smith (2013).

A hospital system tracked the frequency of urinary tract infections (UTIs) acquired by patients while in one of its hospitals. The following statements create a SAS data set with the variable DaysBetween, which contains the number of days between discharges from the hospital of male patients who acquired UTIs while there:

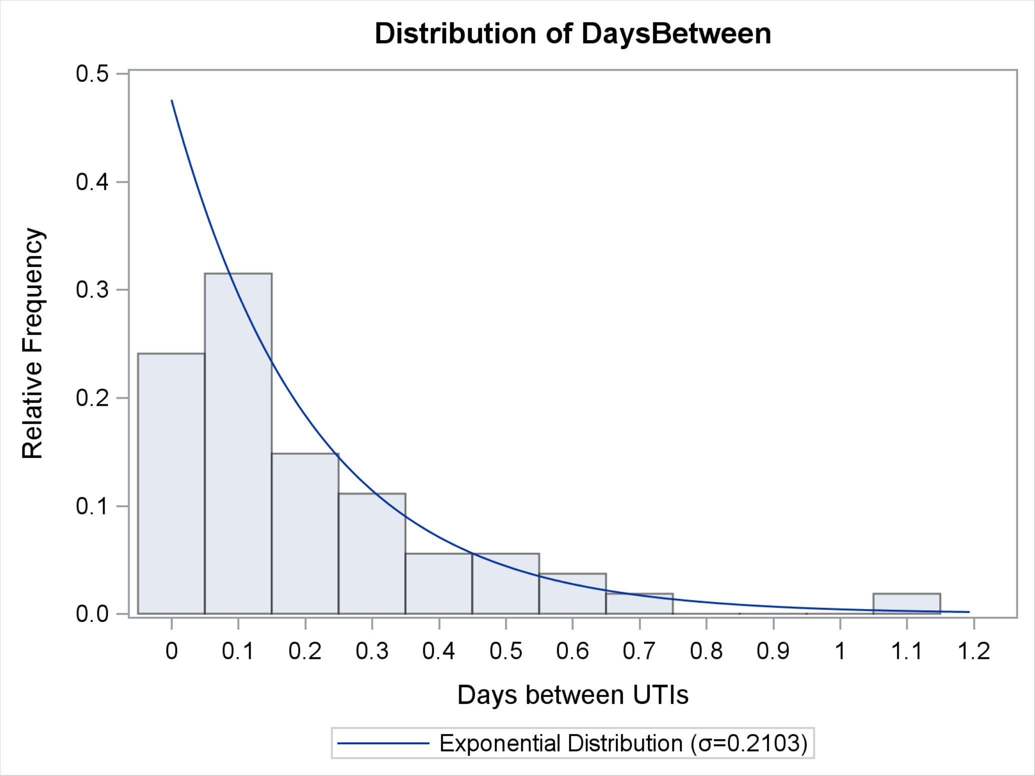
```
data UrinaryTractInfections;
   input DaysBetween @@;
   label DaysBetween = 'Days between UTIs';
datalines;
0.57014 0.07431 0.15278 0.14583 0.13889
0.14931 0.03333 0.08681 0.33681 0.03819
0.24653 0.29514 0.11944 0.05208 0.12500
0.25000 0.40069 0.02500 0.12014 0.11458
0.00347 0.12014 0.04861 0.02778 0.32639
0.64931 0.14931 0.01389 0.03819 0.46806
0.22222 0.29514 0.53472 0.15139 0.52569
0.07986 0.27083 0.04514 0.13542 0.08681
0.40347 0.12639 0.18403 0.70833 0.15625
0.24653 0.04514 0.01736 1.08889 0.05208
0.02778 0.03472 0.23611 0.35972
;
```

The following statements produce a graph that compares the data to a reference distribution whose parameters are estimated from the data. The RAREEVENTS procedure uses an exponential distribution by default because the data are continuous.

```
proc rareevents data=UrinaryTractInfections;
   compare DaysBetween / nbins=12;
run;
```

The NBINS= option specifies that 12 histogram bins be used to display the data. Output 17.1.1 shows the resulting histogram of the data overlaid with the exponential curve.

**Output 17.1.1** Distribution of Intervals between UTIs



Because a continuous distribution is in effect, the COMPARE statement also produces a table of goodness-of-fit statistics, which is shown in Output 17.1.2.

**Output 17.1.2** Goodness-of-Fit Statistics for UTIs

**The RAREEVENTS Procedure**

Goodness-of-Fit Tests for Exponential Distribution

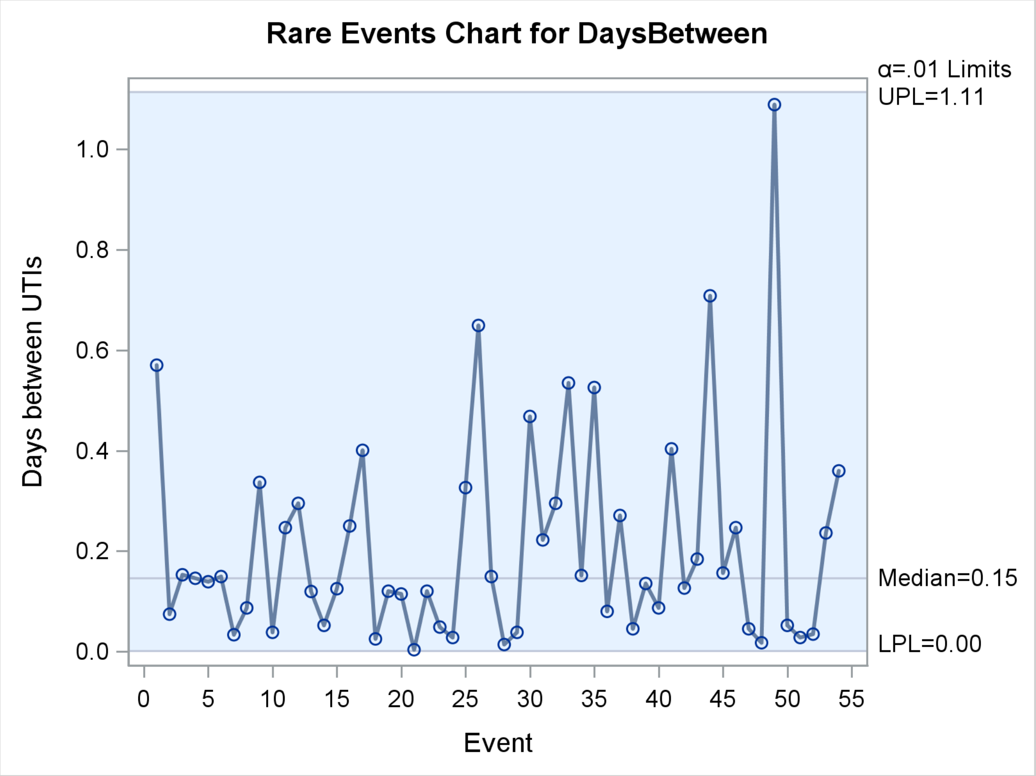
| Test | Statistic | | p Value | |
|---|---|---|---|---|
| Kolmogorov-Smirnov | D | 0.08673920 | Pr > D | >0.500 |
| Cramer-von Mises | W-Sq | 0.04104603 | Pr > W-Sq | >0.500 |
| Anderson-Darling | A-Sq | 0.26919944 | Pr > A-Sq | >0.500 |

The histogram and the goodness-of-fit tests indicate that an exponential distribution is appropriate for the data. The following statements produce a rare events chart for the days between UTIs:

```
proc rareevents data=UrinaryTractInfections;
   chart DaysBetween / totpanels=1;
run;
```

The TOTPANELS= option specifies that all the observations be displayed in a single panel, or page. No index variable is specified, so the DaysBetween values are numbered consecutively, starting with 1. Output 17.1.3 shows the resulting chart.

**Output 17.1.3** Rare Events Chart for Urinary Tract Infections



The rare events chart shows no indication of unusual variation in the incidence of UTIs among male patients.

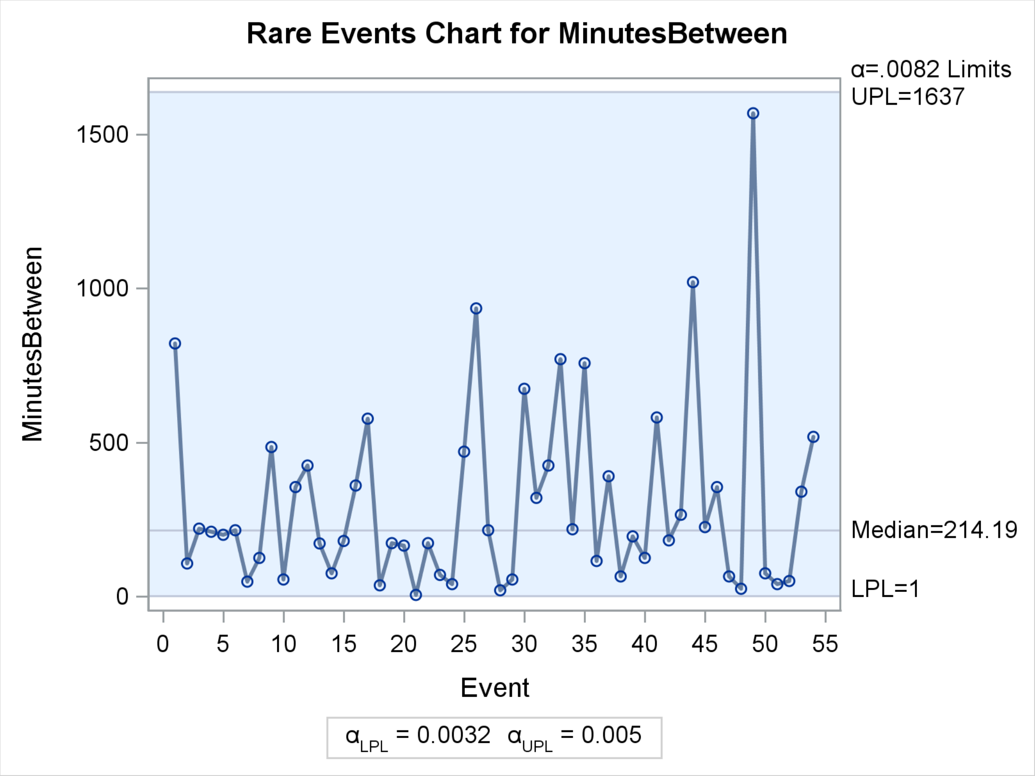
Although Santiago and Smith (2013) provide the data as the (continuous) numbers of days between patient discharges, they could just as well have been recorded as the (integer) number of minutes between discharges. The following statements compute the variable MinutesBetween, which contains counts of the minutes between infections, and produce a rare events chart of the counts. Because the data are integer values, the probability limits are based on a geometric distribution.

```
data UrinaryTractInfections;
   set UrinaryTractInfections;
   MinutesBetween = round( DaysBetween * 1440, 1 );
run;

proc rareevents data=UrinaryTractInfections;
   chart MinutesBetween / totpanels=1;
run;
```

Output 17.1.4 shows the rare events chart for MinutesBetween. The median and probability limits for this chart are very close, but not exactly equal, to the corresponding values measured in days in Output 17.1.3.

**Output 17.1.4** Rare Events Chart for Urinary Tract Infections



## Example 17.2: Airline Crashes

The following statements create a SAS data set that contains data from the National Transportation Safety Board (NTSB) Aviation Accident Database. You can query the database at `http://www.ntsb.gov/_layouts/ntsb.aviation/index.aspx`. These data involve commercial airline crashes that resulted in fatalities and took place in the United States from 1982 through 2016. The DATA step creates a new variable, DaysBetweenCrashes, that records the number of days between successive crashes.

```
data AirCrashes;
   input EventID : $14. EventDate mmddyy10. Location & $32.;
   DaysBetweenCrashes = EventDate - lag(EventDate);
   label DaysBetweenCrashes = 'Days';
datalines;
20020917X01907 01/13/1982 WASHINGTON, DC
20020917X01909 01/23/1982 BOSTON, MA
20020917X03104 07/09/1982 NEW ORLEANS, LA
20020917X04908 11/11/1982 MIAMI, FL
20001214X41967 01/09/1983 BRAINERD, MN
20001214X41968 01/11/1983 DETROIT, MI
20001214X44795 10/11/1983 PINCKNEYVILLE, IL
20001214X45258 12/20/1983 SIOUX FALLS, SD
20001214X39535 05/30/1984 CHALKHILL, PA
20001214X35492 01/09/1985 KANSAS CITY, KS
20001214X35493 01/21/1985 RENO, NV
20001214X36375 05/31/1985 NASHVILLE, TN
```

```
20001214X37434 08/02/1985 DALLAS/FT WORTH, TX
20001214X37757 09/06/1985 MILWAUKEE, WI
20001213X34942 10/04/1986 KELLY AFB, TX
20001213X35148 11/06/1986 TAMPA, FL
20001213X30626 04/13/1987 KANSAS CITY, MO
20001213X31759 08/16/1987 ROMULUS, MI
20001213X32505 11/15/1987 DENVER, CO
20001213X32679 12/07/1987 SAN LUIS OBISPO, CA
20001213X25439 04/28/1988 MAUI, HI
20001213X26528 08/31/1988 DALLAS/FT WORTH, TX
20001213X27734 02/09/1989 SALT LAKE CITY, UT
20001213X27705 02/24/1989 HONOLULU, HI
20001213X27867 03/15/1989 WEST LAFAYETTE, IN
20001213X27869 03/18/1989 SAGINAW, TX
20001213X28786 07/19/1989 SIOUX CITY, IA
20001213X29335 09/20/1989 FLUSHING, NY
20001213X29644 10/07/1989 ORLANDO, FL
20001213X29997 12/27/1989 MIAMI, FL
20001212X22400 01/18/1990 ATLANTA, GA
20001212X22386 01/31/1990 INDIANAPOLIS, IN
20001212X22742 03/13/1990 PHOENIX, AZ
20001212X24506 10/03/1990 CAPE CANAVERAL, FL
20001212X24751 12/03/1990 ROMULUS, MI
20001212X24751 12/03/1990 ROMULUS, MI
20001212X16433 02/01/1991 LOS ANGELES, CA
20001212X16434 02/17/1991 CLEVELAND, OH
20001212X16583 03/03/1991 COLORADO SPGS, CO
20001212X18366 10/12/1991 BRIDGEPORT, CT
20001211X14094 02/15/1992 SWANTON, OH
20001211X14270 03/22/1992 FLUSHING, NY
20001211X14503 04/08/1992 DAYTON, OH
20001211X16222 12/08/1992 FLUSHING, NY
20001211X12079 04/04/1993 CHICAGO, IL
20001206X01727 07/02/1994 CHARLOTTE, NC
20001206X02233 09/08/1994 ALIQUIPPA, PA
20001206X02420 10/31/1994 ROSELAWN, IN
20001206X02586 11/22/1994 BRIDGETON, MO
20001208X05743 05/11/1996 MIAMI, FL
20001208X06203 07/06/1996 PENSACOLA, FL
20001208X06204 07/17/1996 EAST MORICHES, NY
20001208X06132 07/20/1996 RUSSIAN MISSION, AK
20001208X07619 03/27/1997 JAMAICA, NY
20001208X08607 08/07/1997 MIAMI, FL
20001208X09291 12/28/1997 PACIFIC OCEAN
20001212X18961 06/01/1999 LITTLE ROCK, AR
20001212X19260 07/28/1999 LITTLE ROCK, AR
20001212X20339 01/31/2000 Port Hueneme, CA
20001212X20472 02/16/2000 RANCHO CORDOVA, CA
20001212X22314 11/20/2000 MIAMI, FL
20010904X01867 08/05/2001 Washington, DC
20020123X00106 09/11/2001 Shanksville, PA
20020123X00105 09/11/2001 Arlington, VA
20020123X00104 09/11/2001 New York City, NY
20020123X00103 09/11/2001 New York City, NY
```

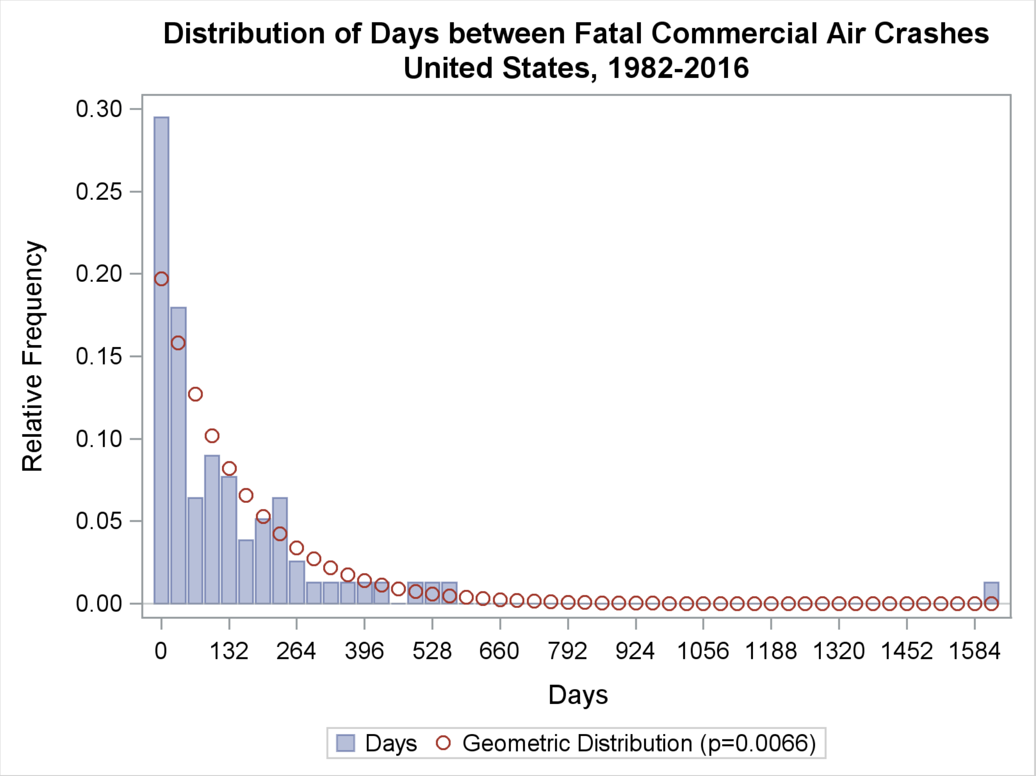
```
   20011130X02321 11/12/2001 Belle Harbor, NY
   20030110X00049 01/08/2003 Charlotte, NC
   20030917X01555 09/12/2003 Norfolk, VA
   20040825X01286 08/13/2004 Florence, KY
   20041020X01659 10/19/2004 Kirksville, MO
   20050609X00744 06/07/2005 Washington, DC
   20051213X01964 12/08/2005 Chicago, IL
   20060106X00018 12/19/2005 Miami, FL
   20060131X00140 01/16/2006 El Paso, TX
   20060828X01244 08/27/2006 Lexington, KY
   20070718X00958 07/10/2007 Tunica, MS
   20090213X13613 02/12/2009 Clarence Center, NY
   20130814X15751 08/14/2013 Birmingham, AL
   ;
```

The following statements produce a comparison plot and a rare events chart for DaysBetweenCrashes:

```
   proc rareevents data=AirCrashes;
      id EventId EventDate Location;
      compare DaysBetweenCrashes /
         process=bar
         reference=marker
         odstitle='Distribution of Days between Fatal Commercial Air Crashes'
         odstitle2='United States, 1982-2016'
         ;
      chart DaysBetweenCrashes /
         odstitle='Days between Fatal Commercial Air Crashes'
         odstitle2='United States, 1982-2016'
         nohlabel
         ;
   run;
```

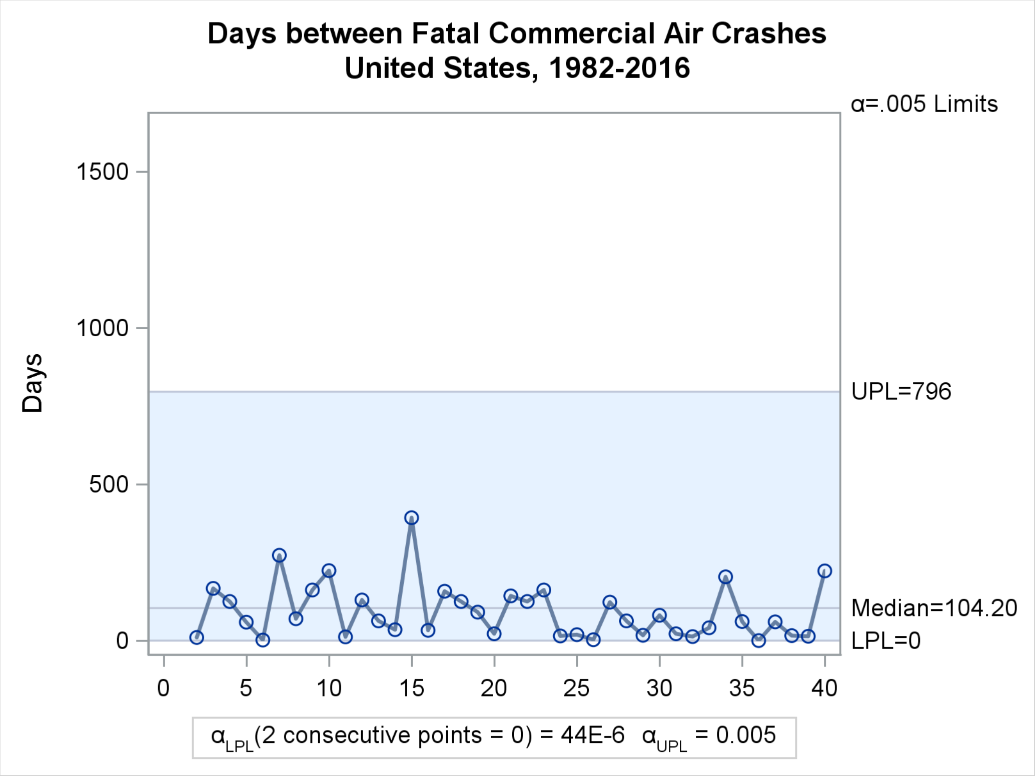
The PROCESS= and REFERENCE= options determine how the process data and reference distribution are displayed in the comparison chart. The ODSTITLE= and ODSTITLE2= options specify titles for the graphs. The NOHLABEL option suppresses the horizontal axis label in the rare events chart. Output 17.2.1 compares the data to a geometric distribution and indicates that the distribution reasonably describes the data.

**Output 17.2.1** Comparison Plot for Days between Crashes

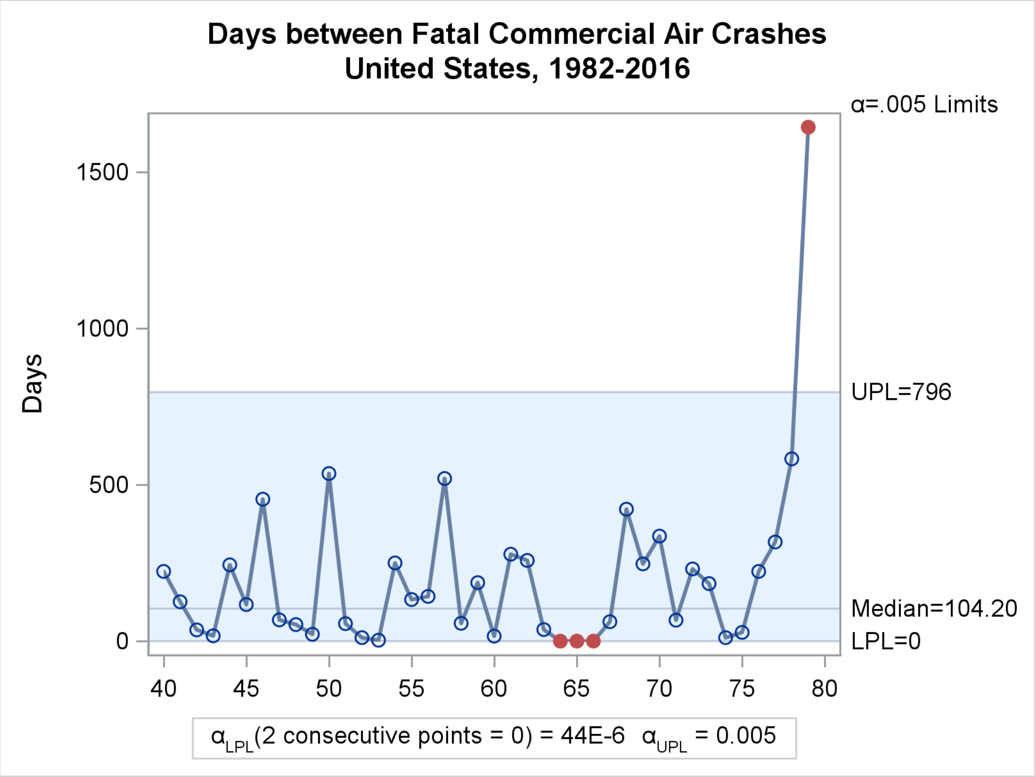


Output 17.2.2 and Output 17.2.3 show the two panels of the rare events chart.

**Output 17.2.2** Rare Events Chart for Air Crashes (Panel 1)



**Output 17.2.3** Rare Events Chart for Air Crashes (Panel 2)



Note that the counts of days between crashes are generally smaller in the first panel of the chart (Output 17.2.2) than in the second panel. Those measurements correspond approximately to the years from 1982 to 1992. There appears to have been a significant change in the process around that time. In Output 17.2.3, the three consecutive measurements of 0 that signal unusual variation correspond to the terrorist attacks on September 11, 2001.

The following statements create a _PHASE_ variable that divides the data into periods before and after December 31, 1992. The observations that correspond to the September 11 crashes are removed from the data, and separate sets of probability limits are computed for the two phases.

```
data AirCrashes2;
   set AirCrashes;
   where EventDate ne '11sep2001'd;
   if EventDate <= '31dec1992'd then
      _PHASE_ = '1982-1992';
   else
      _PHASE_ = '1993-2016';
run;

proc rareevents data=AirCrashes2;
   id EventId EventDate Location;
   chart DaysBetweenCrashes /
      readphases=all
      nochart
      outlimits=AirLimits;
run;
```

The READPHASES=ALL option in the CHART statement specifies that the chart include observations from the input data set for all values of the _PHASE_ variable. The NOCHART option suppresses the creation of the chart, and the OUTLIMITS= option saves the computed probability limits in the data set AirLimits. The AirLimits data set is listed in Output 17.2.4.

**Output 17.2.4** AirLimits Data Set

| _VAR_ | _PHASE_ | _DIST_ | _LPL_ | _MEDIAN_ | _UPL_ | _ALPHALPL_ | _ALPHAUPL_ | _PARMEST_ | _P_ | _SHIFT_ |
|---|---|---|---|---|---|---|---|---|---|---|
| DaysBetweenCrashes | 1982-1992 | GEOMETRIC | 0 | 66.079 | 505 | .000108885 | .004953103 | 1 | 0.010435 | 0 |
| DaysBetweenCrashes | 1993-2016 | GEOMETRIC | 1 | 174.049 | 1330 | .003974563 | .004988181 | 1 | 0.003975 | 0 |

Note the dramatic difference in the _MEDIAN_ values for the two phases.

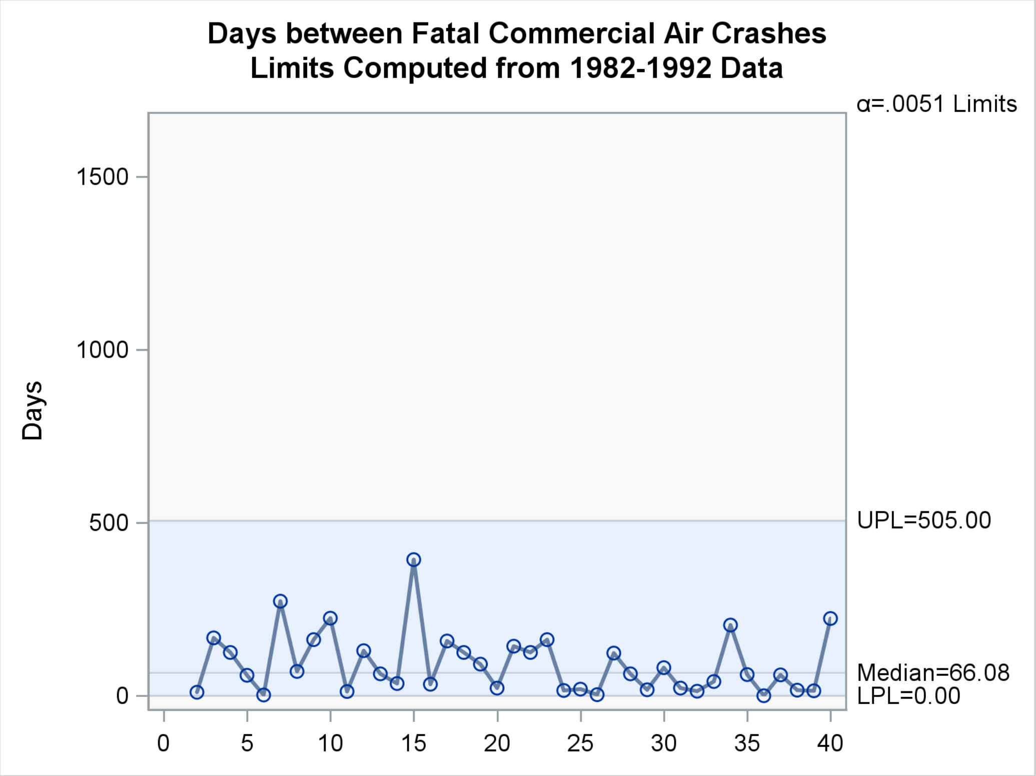
The following statements create a chart of both phases and apply the probability limits that were computed for the first phase:

```
proc rareevents data=AirCrashes2 limits=AirLimits;
   id EventId EventDate Location;
   chart DaysBetweenCrashes /
      readphases=all
      limitphases='1982-1992'
      odstitle='Days between Fatal Commercial Air Crashes'
```

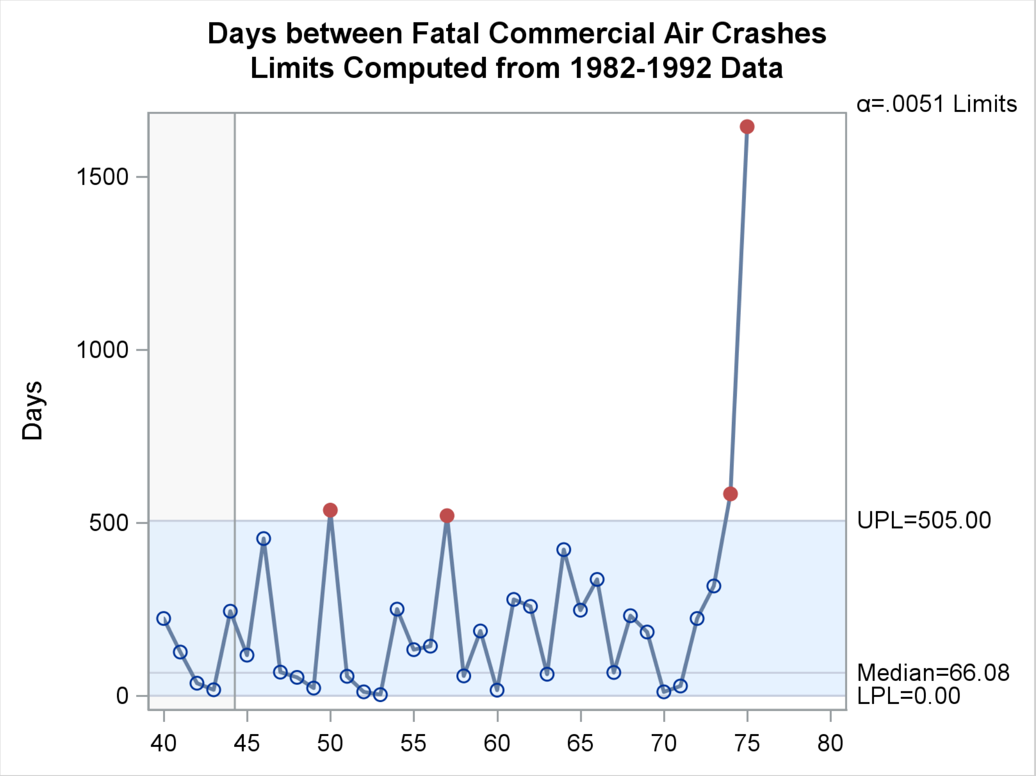
```
        odstitle2='Limits Computed from 1982-1992 Data'
        nohlabel;
run;
```

The LIMITS= option in the PROC statement reads the previously computed probability limits from the AirLimits data set. The LIMITPHASES= option uses the limits for the phase “1982–1992” for the entire chart. The resulting chart is shown in Output 17.2.5 and Output 17.2.6.

**Output 17.2.5** Rare Events Chart for Air Crashes (Panel 1)



**Output 17.2.6** Rare Events Chart for Air Crashes (Panel 2)



This chart emphasizes the process shift that occurred around 1992. In the first phase the process was stable, but 4 of 32 measurements in the second phase exceed the UPL of the first phase. This is strong evidence of a change in the process, with fatal airline crashes becoming less frequent.

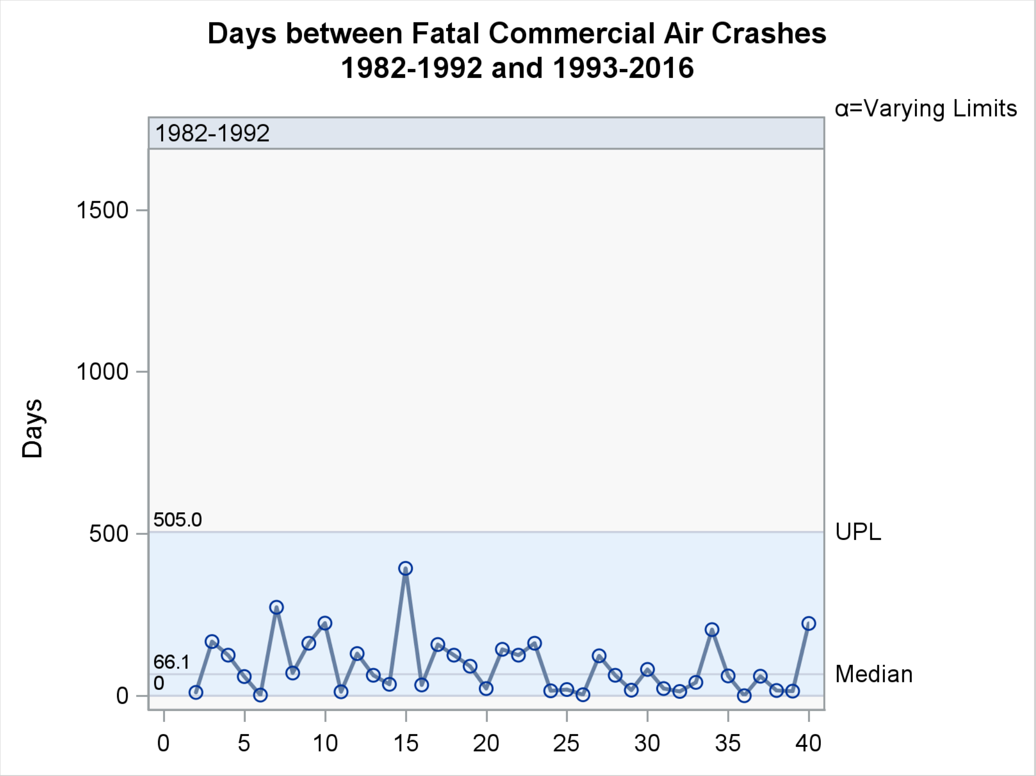
Finally, the following statements produce a rare events chart that uses the probability limits that were computed separately:

```
proc rareevents data=AirCrashes2 limits=AirLimits;
   id EventId EventDate Location;
   chart DaysBetweenCrashes /
      readphases=all
      limitphases=all
      phaselegend
      phaselimits
      odstitle='Days between Fatal Commercial Air Crashes'
      odstitle2='1982-1992 and 1993-2016'
      nohlabel;
run;
```

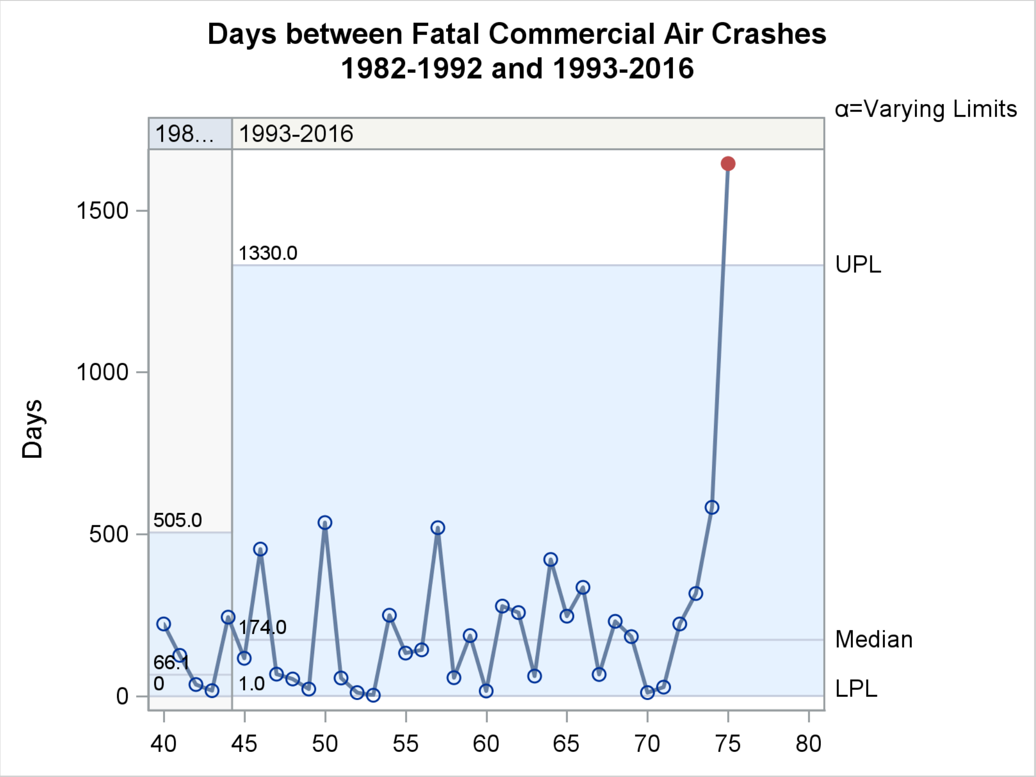
The PHASELEGEND option produces a legend at the top of the chart that labels the phases. The PHASELIMITS option labels the probability limits and center line of each phase.

The resulting chart is shown in Output 17.2.7 and Output 17.2.8.

**Output 17.2.7** Rare Events Chart for Air Crashes (Panel 1)



**Output 17.2.8** Rare Events Chart for Air Crashes (Panel 2)



The time between the two most recent crashes exceeds the UPL for the second phase, which is 1,330 days. The time since the most recent crash, which is not yet reflected on the chart, is also greater than 1,330 days. This indicates that the trend of less frequent fatal commercial air crashes in the United States is not due to random variation but is due to improvements in the process.

## References

Benneyan, J. C. (1998a). "Statistical Quality Control Methods in Infection Control and Hospital Epidemiology, Part I: Introduction and Basic Theory." *Infection Control and Hospital Epidemiology* 19:194–214.

Benneyan, J. C. (1998b). "Statistical Quality Control Methods in Infection Control and Hospital Epidemiology, Part II: Chart Use, Statistical Properties, and Research Issues." *Infection Control and Hospital Epidemiology* 19:265–283.

Benneyan, J. C. (1999). "Geometric-Based *g*-Type Statistical Control Charts for Infrequent Adverse Events: New Quality Control Charts for Hospital Infections." In *Institute of Industrial Engineers Society for Health Systems 1999 Conference Proceedings*, 175–185. Norcross, GA: Institute of Industrial Engineers, Society for Health Systems.

Benneyan, J. C. (2001a). "Number-Between *g*-Type Statistical Control Charts." *Health Care Management Science* 4:305–318.

Benneyan, J. C. (2001b). "Performance of Number-Between *g*-Type Statistical Control Charts for Monitoring Adverse Events." *Health Care Management Science* 4:319–336.

Benneyan, J. C. (2006). "Discussion: Statistical Process Control Methods in Health Care." *Journal of Quality Technology* 38:113–123.

D'Agostino, R. B., and Stephens, M., eds. (1986). *Goodness-of-Fit Techniques*. New York: Marcel Dekker.

Kaminsky, F. C., Benneyan, J. C., Davis, R. D., and Burke, R. J. (1992). "Statistical Control Charts Based on a Geometric Distribution." *Journal of Quality Technology* 24:63–69.

Santiago, E., and Smith, J. (2013). "Control Charts Based on the Exponential Distribution: Adapting Runs Rules for the *t* Chart." *Quality Engineering* 25:85–96.

Terrell, G. R., and Scott, D. W. (1985). "Oversmoothed Nonparametric Density Estimates." *Journal of the American Statistical Association* 80:209–214.

Woodall, W. H. (2006). "The Use of Control Charts in Health-Care and Public-Health Surveillance." *Journal of Quality Technology* 38:89–104.

# Subject Index

# Syntax Index