SAS/QC® 14.2 User's Guide
The PARETO Procedure

This document is an individual chapter from *SAS/QC® 14.2 User's Guide*.

The correct bibliographic citation for this manual is as follows: SAS Institute Inc. 2016. *SAS/QC® 14.2 User's Guide*. Cary, NC: SAS Institute Inc.

SAS/QC® 14.2 User's Guide

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

November 2016

Chapter 16
The PARETO Procedure

Contents

Overview: PARETO Procedure

The PARETO procedure creates Pareto charts, which display the relative frequencies of quality-related problems in a process or operation. The frequencies are represented by bars that are ordered in decreasing magnitude. Thus, you can use a Pareto chart to decide which subset of problems you should solve first or which problem areas deserve the most attention.

Pareto charts provide a tool for visualizing the Pareto principle,[1] which states that a small subset of problems tend to occur much more frequently than the remaining problems. In Japanese industry, the Pareto chart is one of the "seven basic QC tools" that are heavily used by workers and engineers. Ishikawa (1976) discusses how to construct and interpret a Pareto chart. Examples of Pareto charts are also given by Kume (1985) and Wadsworth, Stephens, and Godfrey (1986).

You can use the PARETO procedure to do the following:

- construct Pareto charts from unsorted raw data (for example, a set of quality problems that have not been classified into categories) or from a set of distinct categories and corresponding frequencies
- construct Pareto charts that are based on the percentage of occurrence of each problem, the frequency (number of occurrences), or a weighted frequency (such as frequency that is weighted by the cost of each problem)
- add a curve that indicates the cumulative percentage across categories
- construct side-by-side Pareto charts or stacked Pareto charts
- construct *comparative Pareto charts*, which enable you to compare the Pareto frequencies across the levels of one or two classification variables. For example, you can compare the frequencies of problems that occur on three different machines for five consecutive days.
- highlight the "vital few" and the "useful many"[2] categories by using different colors for bars that correspond to the n most frequently occurring categories or the m least frequently occurring categories.
- restrict the number of categories that are displayed to the n most frequently occurring categories
- create charts whose bars are oriented vertically or horizontally
- highlight special categories by using different colors for specific bars
- display sample sizes and other statistics on Pareto charts
- label the bars with their frequency values

[1] Both the chart and the principle are named after Vilfredo Pareto (1848–1923), an Italian economist and sociologist. His first work, *Cours d'Économie Politique* (1895–1897), applied what is now termed the *Pareto distribution* to the study of income size.

[2] Juran originally referred to these categories as the "trivial many"; however, because all problems merit attention, the term "useful many" is preferred (Burr 1990).

- create charts as ODS Graphics output or as traditional graphics
- annotate traditional graphics charts
- save traditional graphics output in a graphics catalog for subsequent replay
- save information that is associated with the categories (such as the frequencies) in an output data set
- create variations on traditional Pareto charts, as described by Wilkinson (2006)

A Pareto chart has three axes, whose display depends on whether the Pareto chart is a traditional vertical Pareto or a horizontal bar chart. A horizontal bar chart that is produced by the PARETO procedure is essentially a vertical Pareto chart that is rotated 90 degrees clockwise. Table 16.1 shows how the three axes are displayed on the two types of Pareto charts.

Table 16.1 Pareto Chart Axes

Axis	Displayed on a Vertical Pareto Chart	Displayed on a Horizontal Pareto Chart
Category axis	Horizontally at the bottom of the chart	Vertically at the left side of the chart
Frequency axis	On the left (also called the primary vertical axis)	At the top of the chart (also called the primary horizontal axis)
Cumulative percentage axis	On the right (also called the secondary vertical axis)	At the bottom of the chart (also called the secondary horizontal axis)

Getting Started: PARETO Procedure

Creating a Pareto Chart from Raw Data

NOTE: See *Basic Pareto Chart from Raw Data* in the SAS/QC Sample Library.

In the fabrication of integrated circuits, common causes of failures include improper doping, corrosion, surface contamination, silicon defects, metallization, and oxide defects. The causes of 31 failures were recorded in a SAS data set called Failure1:

```
data Failure1;
   length Cause $ 16;
   label  Cause = 'Cause of Failure';
   input  Cause & $;
   datalines;
Corrosion
Oxide Defect
Contamination
Oxide Defect
Oxide Defect
Miscellaneous
Oxide Defect
```

```
Contamination
Metallization
Oxide Defect
Contamination
Contamination
Oxide Defect
Contamination
Contamination
Contamination
Corrosion
Silicon Defect
Miscellaneous
Contamination
Contamination
Contamination
Miscellaneous
Contamination
Contamination
Doping
Oxide Defect
Oxide Defect
Metallization
Contamination
Contamination
;
```

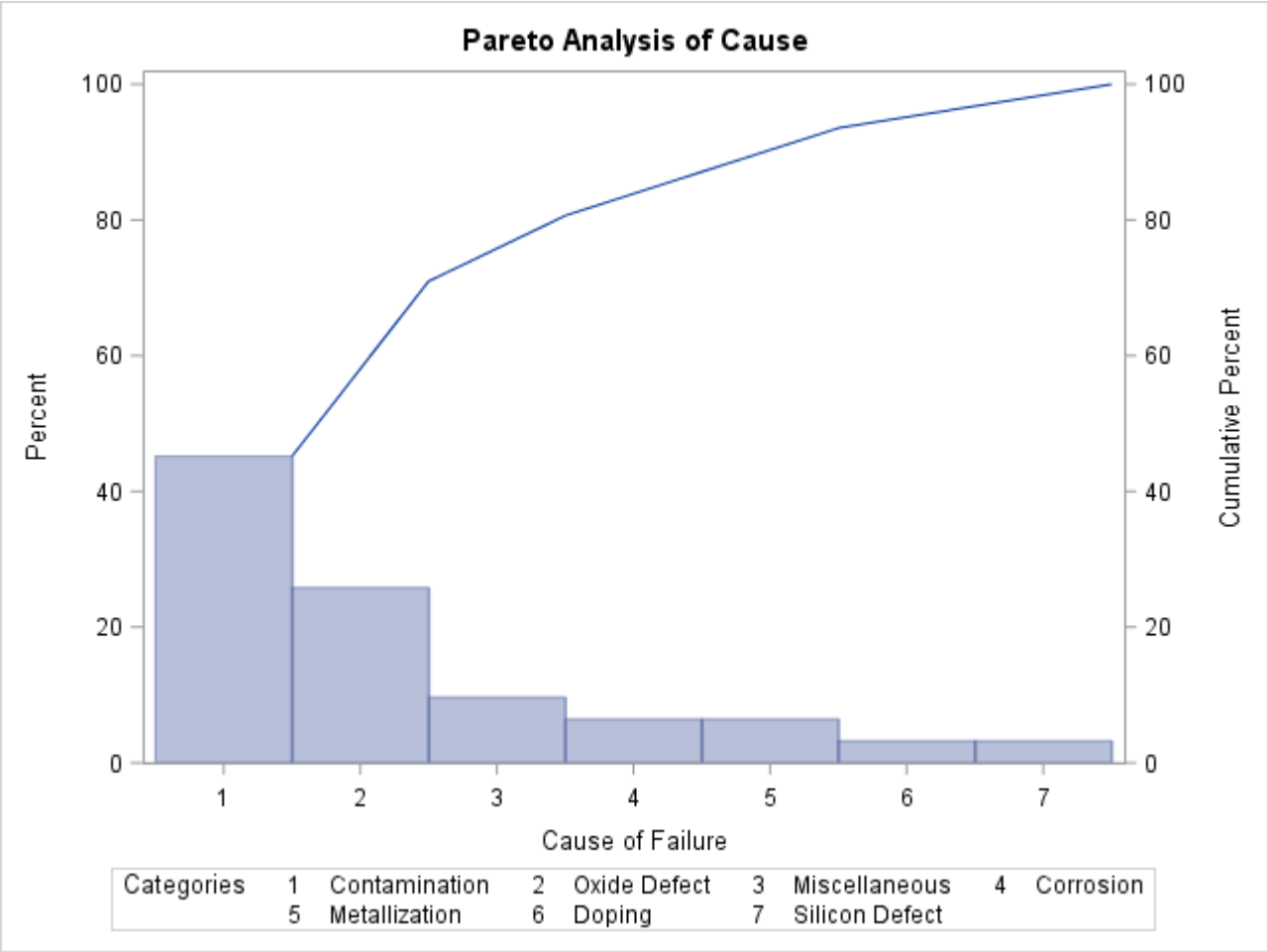
Each of the 31 observations corresponds to a different circuit, and the value of Cause provides the cause for the failure. These are raw data in the sense that more than one observation has the same value of Cause and that the observations are not sorted by Cause. The following statements produce a basic Pareto chart for the failures:

```
ods graphics on;
proc pareto data=Failure1;
   vbar Cause;
run;
```

The PROC PARETO statement (referred to as the PROC statement) invokes the PARETO procedure and identifies the input data set. You specify one or more process variables to be analyzed in the VBAR statement. The ODS GRAPHICS ON statement that is specified before the PROC statement enables ODS Graphics, so the Pareto chart is created using ODS Graphics instead of traditional graphics.

The Pareto chart is shown in Figure 16.1.

Figure 16.1 Pareto Chart for Integrated Circuit Failures in the Data Set Failure1



PROC PARETO has classified the values of Cause into seven distinct categories. The bars represent the percentage of failures in each category, and they are arranged in decreasing order. Thus, the most frequently occurring category is `'Contamination'`, which accounts for 45% of the failures. The Pareto curve indicates the cumulative percentage of failures from left to right; for example, `'Contamination'` and `'Oxide Defect'` together account for 71% of the failures.

If there is insufficient space to label the bars along the category axis, PROC PARETO numbers the bars from left to right and adds a legend to identify the categories, as in Figure 16.1. A category legend is likely to be introduced in the following cases:

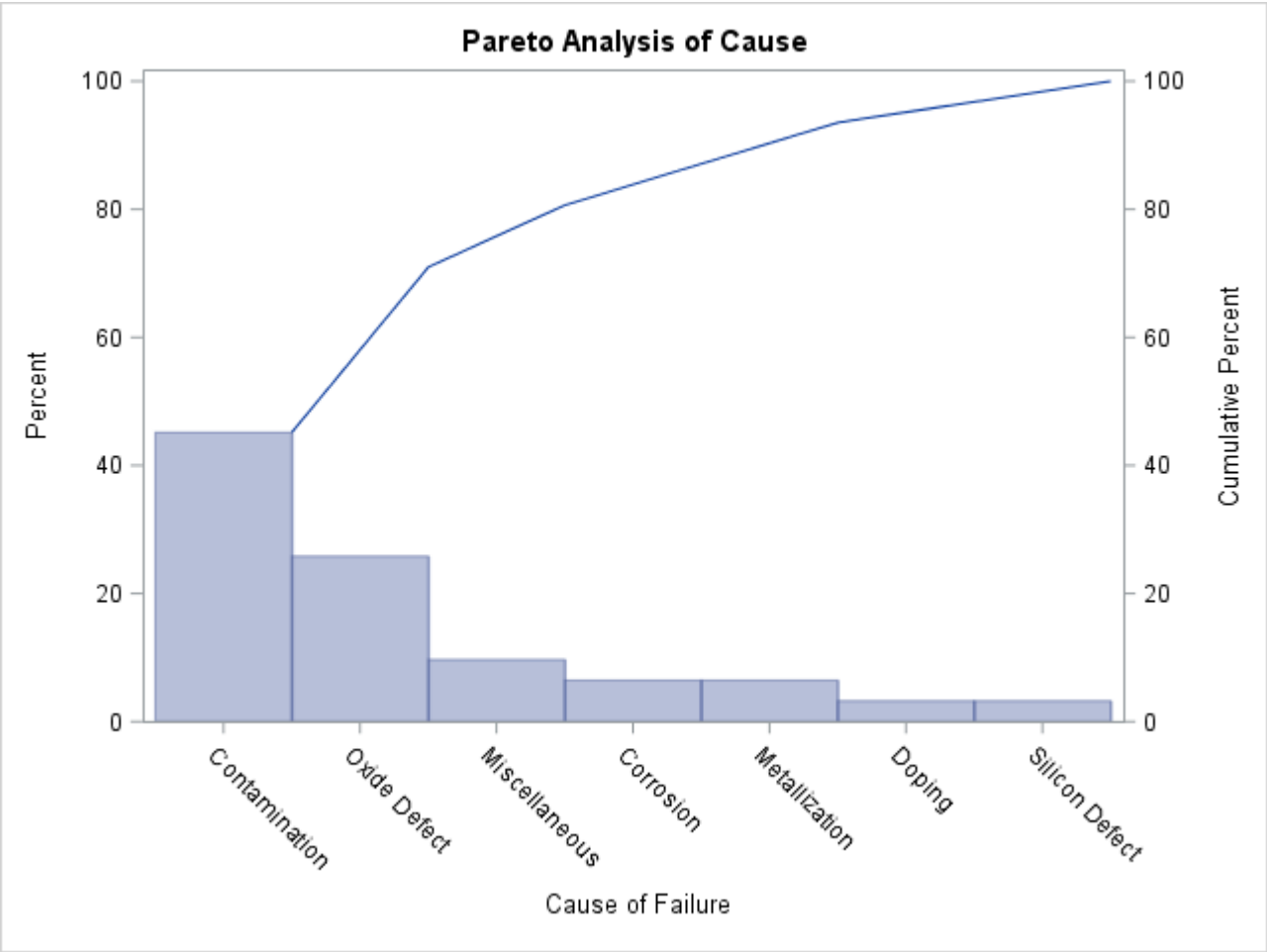
- The number of categories is large.
- The category labels are lengthy. Category labels can be up to 64 characters.
- You specify a large text height. In traditional graphics output, you can specify the text height in the HEIGHT= option in the HBAR or VBAR statement or in the HTEXT= option in a GOPTIONS statement.

The following statements suppress the category legend by specifying the CATLEGEND=OFF option:

```
proc pareto data=Failure1;
   vbar Cause / catlegend=off;
run;
```

A slash (/) is used to separate the process variable Cause from the options that are specified in the VBAR statement. The resulting chart is show in Figure 16.2.

Figure 16.2 Pareto Chart with Category Legend Suppressed



Because the category legend is turned off, PROC PARETO displays the category labels at an angle so that they do not collide.

Creating a Pareto Chart from Frequency Data

NOTE: See *Basic Pareto Chart from Frequency Data* in the SAS/QC Sample Library.

In some situations, a count (frequency) is available for each category, or you can compress a large data set by creating a frequency variable for the categories before applying the PARETO procedure.

For example, you can use the FREQ procedure to obtain the compressed data set Failure2 from the data set Failure1:

```
proc freq data=Failure1;
   tables Cause / noprint out=Failure2;
run;
```

A listing of Failure2 is shown in Figure 16.3.

Figure 16.3 Data Set Failure2, Which Is Created by Using PROC FREQ

Obs	Cause	COUNT	PERCENT
1	Contamination	14	45.1613
2	Corrosion	2	6.4516
3	Doping	1	3.2258
4	Metallization	2	6.4516
5	Miscellaneous	3	9.6774
6	Oxide Defect	8	25.8065
7	Silicon Defect	1	3.2258

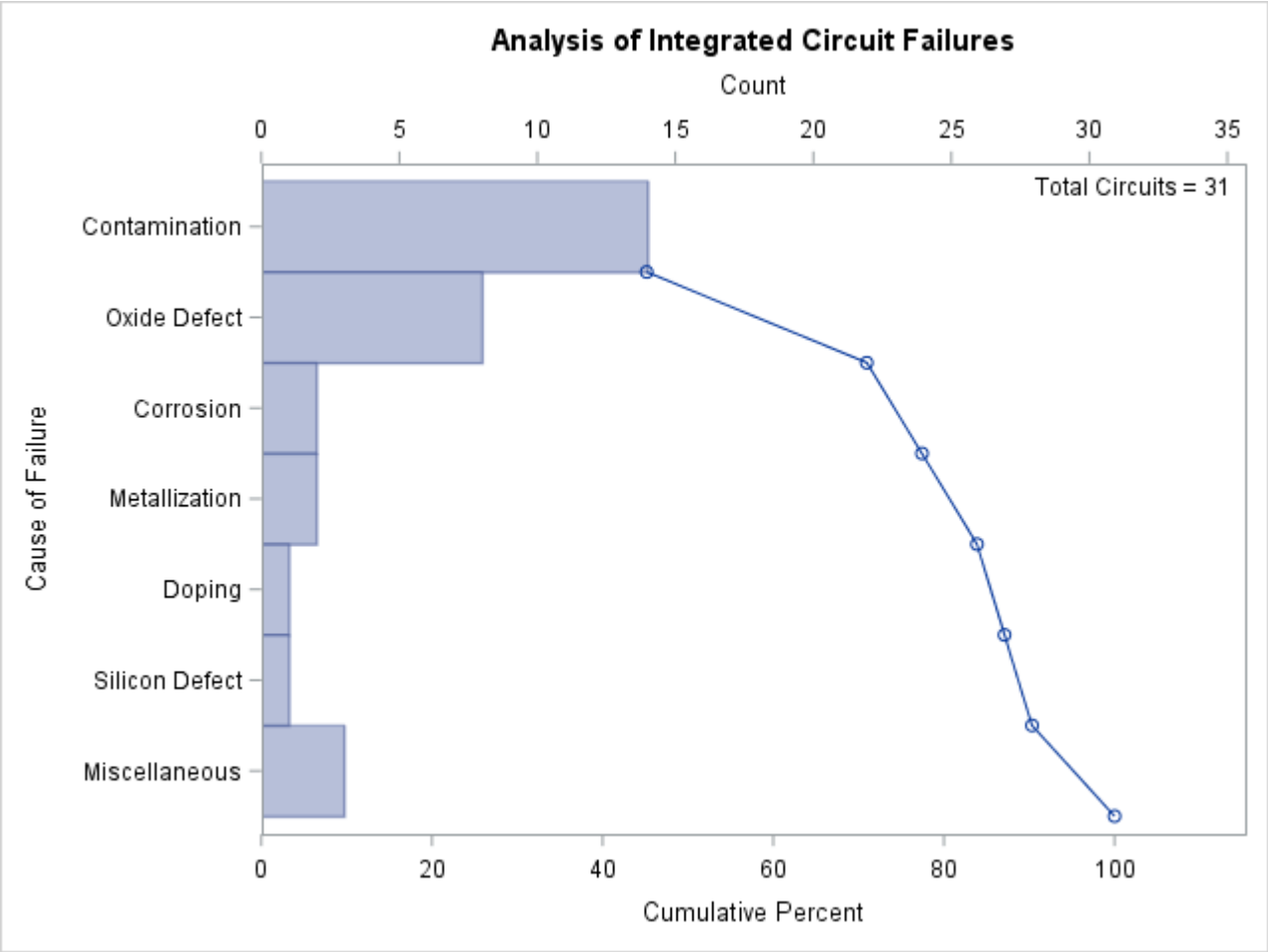
The following statements produce a horizontal Pareto chart for the data in Failure2:

```
title 'Analysis of Integrated Circuit Failures';
proc pareto data=Failure2;
   hbar Cause / freq     = Count
                scale    = count
                last     = 'Miscellaneous'
                nlegend  = 'Total Circuits'
                odstitle = title1
                markers;
run;
```

The frequency variable Count is specified in the FREQ= option. Specifying SCALE=COUNT requests a frequency scale for the frequency axis (at the top of the chart). Specifying LAST=‘Miscellaneous’ causes the `'Miscellaneous'` category to be displayed last regardless of its frequency. The NLEGEND= option adds a sample size legend labeled “Total Circuits.” Specifying ODSTITLE=TITLE replaces the default graph title with the title that is specified in the TITLE statement. The MARKERS option places markers at the points on the cumulative percentage curve.

The chart is displayed in Figure 16.4.

Figure 16.4 Pareto Chart with Frequency Scale



Note that in a horizontal Pareto chart categories are listed in decreasing frequency order from top to bottom on the category axis.

There are two sets of tied categories in this example: `'Corrosion'` and `'Metallization'` each occur twice, and `'Doping'` and `'Silicon Defect'` each occur once. PROC PARETO displays tied categories alphabetically in order of their formatted values. Thus, `'Corrosion'` appears before `'Metallization'`, and `'Doping'` appears before `'Silicon Defect'` in Figure 16.4. This is simply a convention, and no practical significance should be attached to the order in which tied categories are arranged.

Restricting the Number of Pareto Categories

NOTE: See *Pareto Chart with Restricted Number of Categories* in the SAS/QC Sample Library.

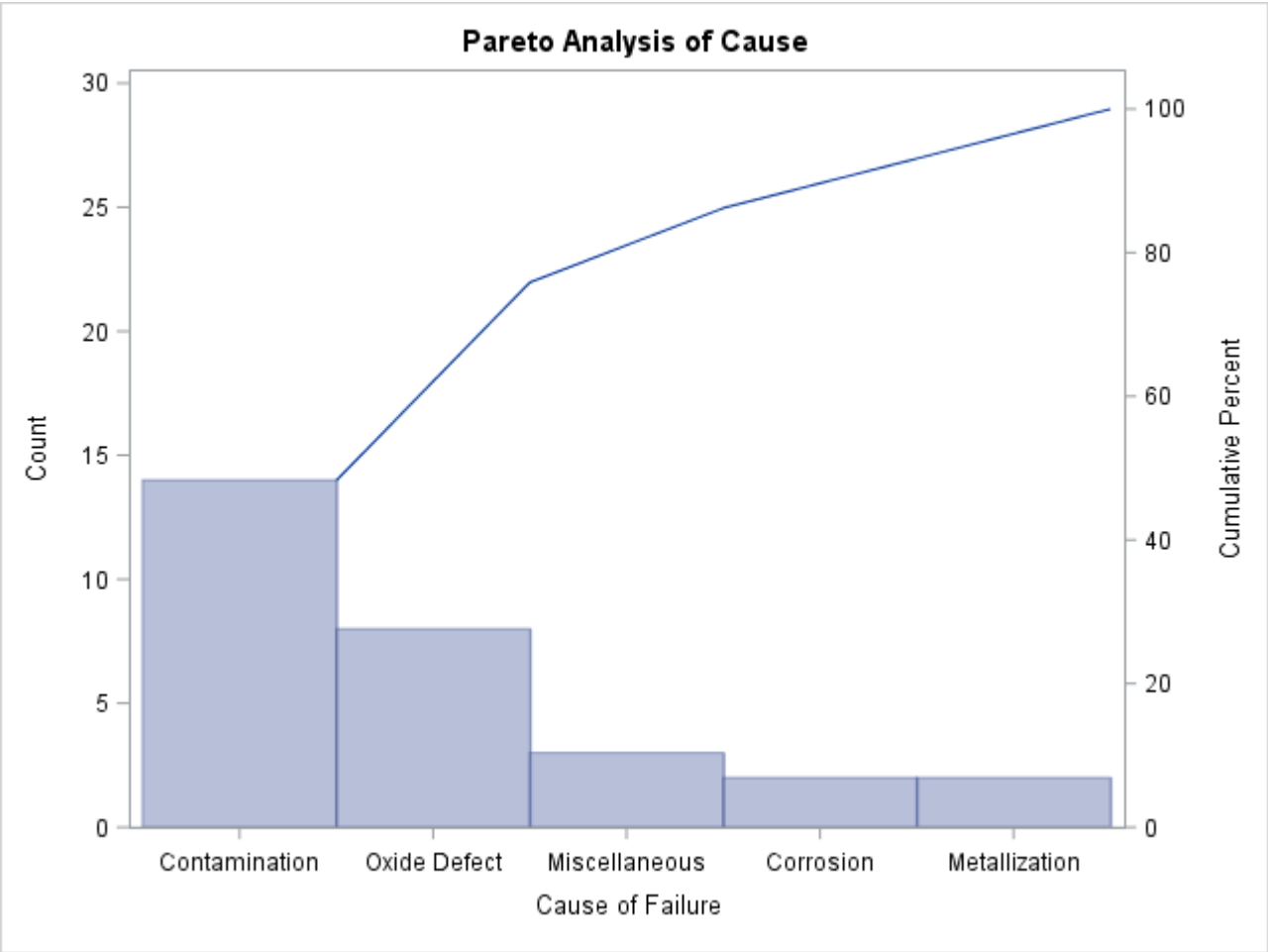
Unlike the previous examples, some applications involve too many categories to display on a chart. The solution presented here is to create a restricted Pareto chart that displays only the most frequently occurring categories.

The following statements create a Pareto chart for the five most frequently occurring levels of Cause in the data set Failure2 (which is listed in Figure 16.3):

```
proc pareto data=Failure2;
   vbar Cause / freq    = Count
                scale   = count
                maxncat = 5;
run;
```

The MAXNCAT= option specifies the number of categories to be displayed. The chart, shown in Figure 16.5, does not display the categories `'Doping'` and `'Silicon Defect'`.

Figure 16.5 Restricted Pareto Chart

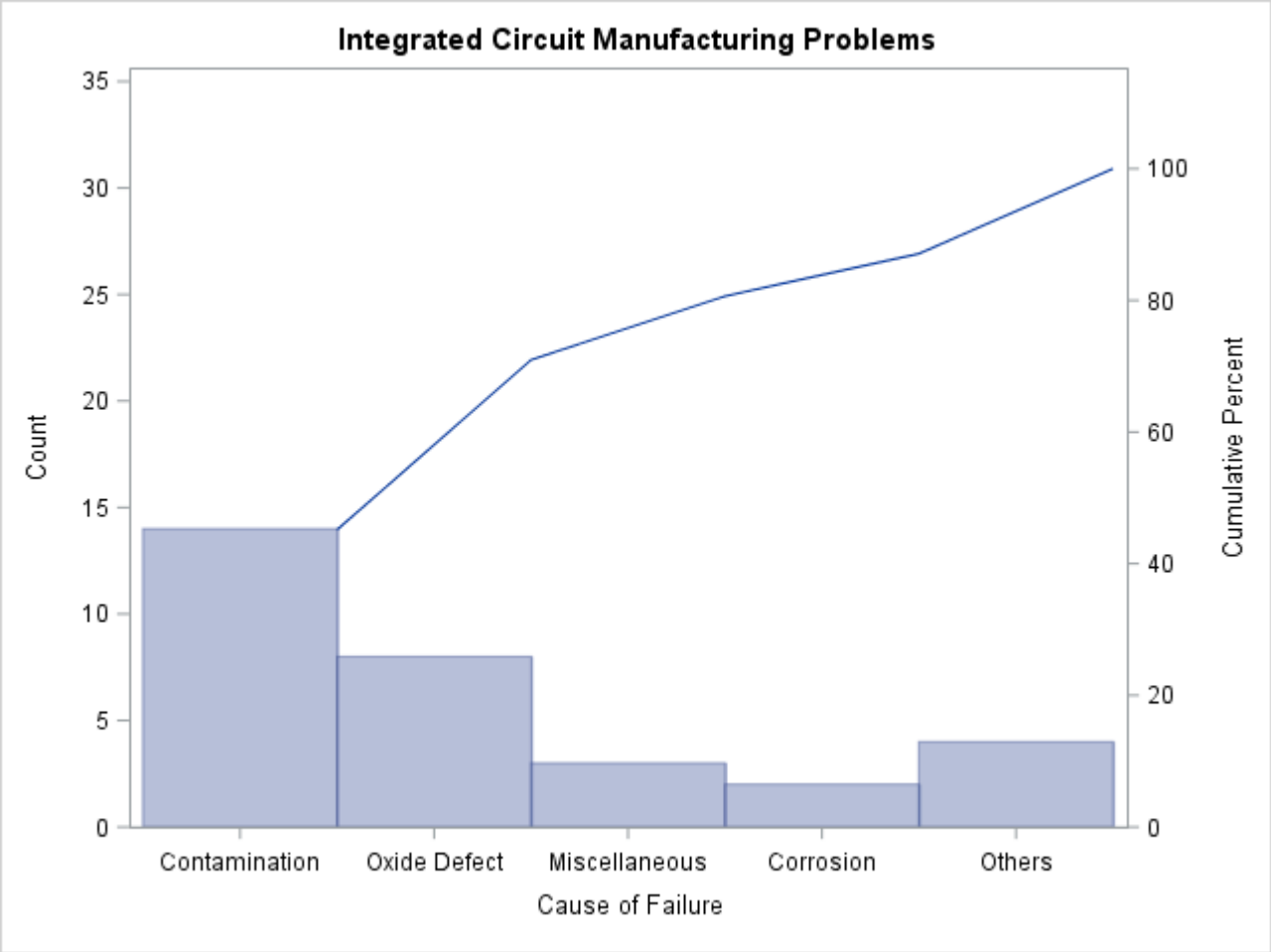


You can also display the most frequently occurring categories and merge the remaining categories into a single *other* category that is represented by a bar. You can specify the name for the new category with the OTHER= option. If, in addition, you specify that name in the LAST= option, the category is positioned at the bottom of the chart. The following statements illustrate both options:

```
title 'Integrated Circuit Manufacturing Problems';
proc pareto data=Failure2;
   vbar Cause / freq     = Count
                scale    = count
                maxncat  = 5
                other    = 'Others'
                last     = 'Others'
                odstitle = title1;
run;
```

The chart is shown in Figure 16.6.

Figure 16.6 Restricted Pareto Chart with *Other* Category



The number of categories displayed is five, which is the number specified in the MAXNCAT= option. The first four categories are the four most frequently occurring problems in Failure2, and the fifth category merges the remaining problems.

Note that `'Corrosion'` and `'Metallization'` both have a frequency of two. When the MAXNCAT= option is applied to categories with tied frequencies, PROC PARETO breaks the tie by using the order of the formatted values. Thus `'Corrosion'` is displayed, whereas `'Metallization'` is merged into the `'Other'` category. The MAXNCAT= and related options are described in the section "Restricted Pareto Charts" on page 1094.

Displaying Summary Statistics on a Pareto Chart

NOTE: See *Displaying Summary Statistics on a Pareto Chart* in the SAS/QC Sample Library.

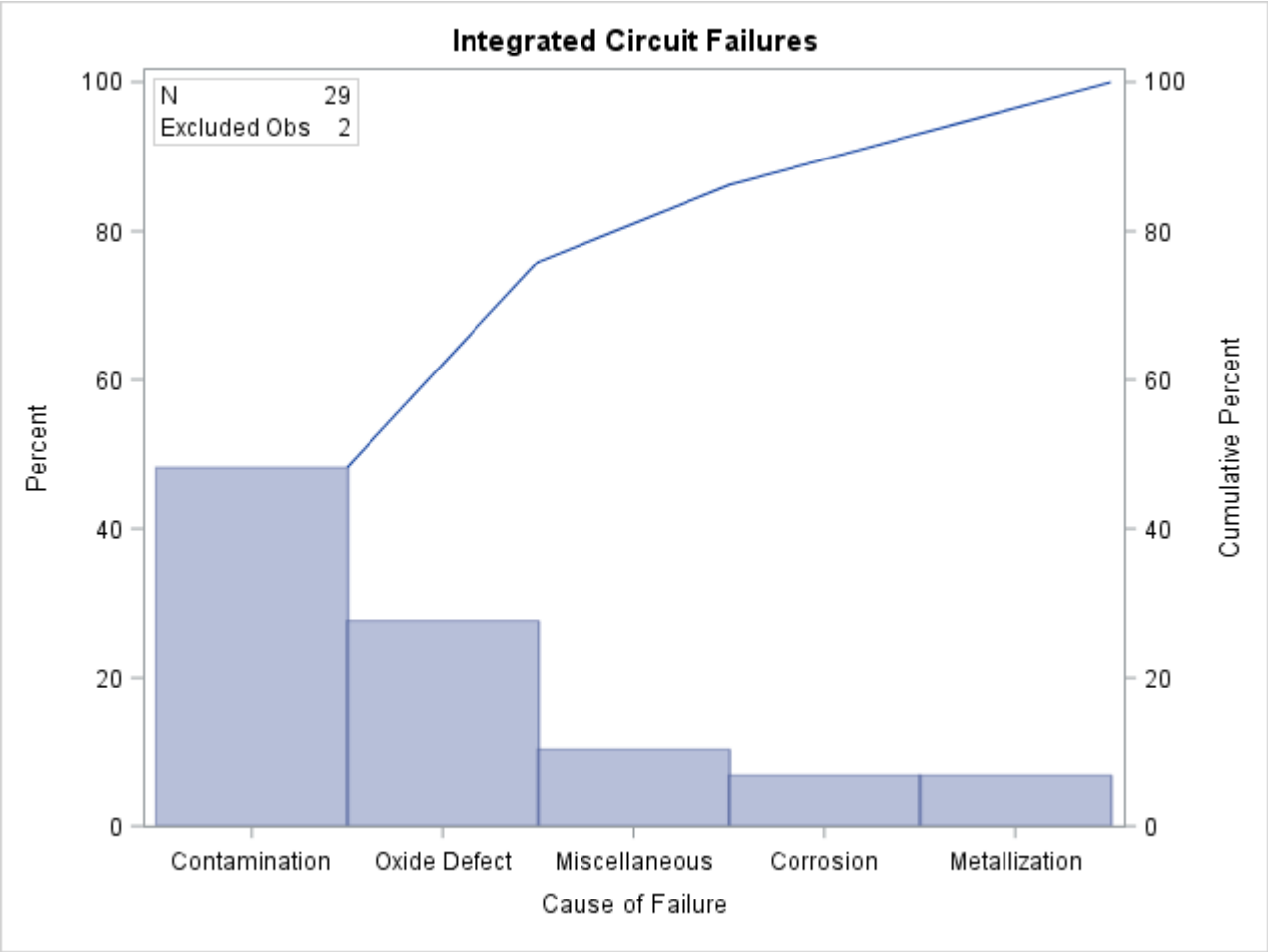
You can use an INSET statement to add a box or table (referred to as an *inset*) of summary statistics on a Pareto chart. The following statements generate a chart from the Failure2 data set and limit the number of categories to five:

```
data Failure2;
   length Cause $ 16 ;
   label  Cause = 'Cause of Failure' ;
   input  Cause $ 1-16 Count;
   datalines;
Contamination    14
Corrosion         2
Doping            1
Metallization     2
Miscellaneous     3
Oxide Defect      8
Silicon Defect    1
;


title 'Integrated Circuit Failures';
proc pareto data=Failure2;
   vbar Cause /
      freq     = Count
      maxncat  = 5
      odstitle = title;
   inset n nexcl;
run;
```

An INSET statement produces an inset on the chart that is created by the preceding HBAR or VBAR chart statement. You specify inset keywords to request summary statistics, and the statistics appear in the order in which you specify the keywords. The keyword N displays the number of categories that are displayed in the chart; the keyword NEXCL displays the number of categories that are excluded. A complete list of keywords available with the INSET statement is provided in the section "INSET Statement Keywords" on page 1062.

The resulting chart is displayed in Figure 16.7.

Figure 16.7 A Pareto Chart with an Inset



Syntax: PARETO Procedure

The following statements are available in the PARETO procedure:

PROC PARETO < *options* > **;**
  BY *variables* **;**
  HBAR (*variable-list*)< / *options* > **;**
  VBAR (*variable-list*)< / *options* > **;**
  INSET *keyword-list* < / *options* > **;**

You must specify the PROC PARETO statement and at least one HBAR or VBAR chart statement. A *chart statement* specifies the process variables that you want to analyze and produces a Pareto chart for each. You can specify any number of chart statements, and all other statements are optional.

The following statements request a vertical Pareto chart for the process variable Reason from the data set Failures. When the process *variable-list* contains only one variable, you do not need to enclose it in parentheses.

```
proc pareto data=Failures;
   vbar Reason;
run;
```

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

PROC PARETO Statement

PROC PARETO < *options* > **;**

The PROC PARETO statement invokes the PARETO procedure. Table 16.2 summarizes the *options* available in the PROC PARETO statement.

Table 16.2 PROC PARETO Statement Options

Option	Description
General Option	
DATA=	Specifies the input SAS data set
Traditional Graphics Options	
ANNOTATE=	Specifies the annotation data set for the frequency axis
ANNOTATE2=	Specifies the annotation data set for the cumulative percentage axis
GOUT=	Specifies the graphics catalog for saving traditional graphics output
Legacy Line Printer Chart Options	
FORMCHAR=	Specifies the formatting characters that are used to construct line printer charts
LINEPRINTER	Creates line printer charts

You can specify the following *options*:

ANNOTATE=*SAS-data-set*
ANNO=*SAS-data-set*

specifies an input data set that contains annotation variables as described in *SAS/GRAPH: Reference*. You can use *SAS-data-set* to customize traditional graphics charts with features such as labels that explain critical categories. The ANNOTATE= data set is associated with the frequency axis. If the annotation is based on data coordinates, you must use the same units as the frequency axis uses. Features provided in this data set are added to every chart that PROC PARETO produces in its current run. This option has no effect when ODS Graphics is enabled.

ANNOTATE2=*SAS-data-set*
ANNO2=*SAS-data-set*

specifies an input data set that contains annotation variables as described in *SAS/GRAPH: Reference*. You can use *SAS-data-set* to customize traditional graphics charts with features such as labels that explain critical categories. The ANNOTATE2= data set is associated with the cumulative percentage axis. If the annotation is based on data coordinates, you must use the same units as the cumulative percentage axis uses. Features provided in this data set are added to every chart that PROC PARETO produces in its current run. This option has no effect when ODS Graphics is enabled.

DATA=*SAS-data-set*
specifies an input data set that contains the process variables and related variables. If you do not specify a DATA= data set, PROC PARETO uses the most recently created data set.

FORMCHAR='*string*'
specifies a list of corner characters and other special characters that enhance the appearance of legacy line printer charts.

If your device supports the ASCII symbol set (1 or 2), use the following list:

```
formchar = 'B3,C4,DA,C2,BF,C3,C5,B4,C0,C1,D9'X
```

The FORMCHAR= option overrides (but does not alter) the FORMCHAR= option that is specified in an OPTIONS statement such as in the following statement:

```
options formchar = 'B3,C4,DA,C2,BF,C3,C5,B4,C0,C1,D9'X;
```

You can place the OPTIONS statement at the top of your SAS program or in an AUTOEXEC.SAS file. The FORMCHAR= has no effect unless you specify LINEPRINTER option.

GOUT=*graphics-catalog*
specifies the graphics catalog in which to save traditional graphics output. This option has no effect when ODS Graphics is enabled.

LINEPRINTER
requests that legacy line printer charts be produced. The HBAR statement does not produce line printer output, so you cannot use an HBAR statement when you specify the LINEPRINTER option.

BY Statement

BY *variables* **;**

You can specify a BY statement with PROC PARETO 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 PARETO 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 *Base SAS Procedures Guide*.

HBAR Statement

HBAR (*variable-list* **)**< / *options* > **;**

The HBAR statement creates a Pareto chart that uses horizontal bars to represent the frequencies of problems in a process or operation. The HBAR statement does not produce line printer charts, so you cannot specify it when you specify the LINEPRINTER option in the PROC PARETO statement.

A horizontal Pareto chart has a vertical category axis. The frequency axis appears at the top of the chart and measures the lengths of the bars on the chart. The cumulative percentage axis is at the bottom of the chart and measures the cumulative percentage curve.

The HBAR statement produces two types of output for Pareto charts:

- It produces ODS Graphics output if ODS Graphics is enabled (for example, by specifying the ODS GRAPHICS ON statement prior to the PROC statement).
- Otherwise, it produces traditional graphics if SAS/GRAPH is licensed.

For more information about producing these different types of graphs. see Chapter 4, "SAS/QC Graphics,"

The *variable-list* specifies the process variables to be analyzed. PROC PARETO creates a chart for each variable, and the values of each variable determine the Pareto categories for that chart. If *variable-list* contains only one process variable, you do not need to enclose it in parentheses.

The variables can be numeric or character, and the maximum length of a character variable is 64. Formatted values determine the categories and are displayed in labels and legends. The maximum format length is 64.

Table 16.3 lists the HBAR statement *options* by function. For complete descriptions, see the section "Dictionary of HBAR and VBAR Statement Options" on page 1071.

Table 16.3 HBAR Statement Options

Option	Description
Data Processing Options	
FREQ=	Specifies the frequency variable
MISSING	Requests that missing values of the process variable be treated as a Pareto category
MISSING1	Requests that missing values of the first CLASS= variable be analyzed as a level
MISSING2	Requests that missing values of the second CLASS= variable be analyzed as a level
OUT=	Creates an output data set that saves the information that is displayed in the Pareto chart
WEIGHT=	Specifies weight variables used to weight frequencies
Options for Restricting the Number of Categories	
LOTHER=	Specifies a label for the OTHER= bar
MAXCMPCT=	Displays only the categories whose cumulative percentage is less than the specified percentage

Table 16.3 (*continued*)

Option	Description
MAXNCAT=	Displays only the categories that have the *n* highest values
MINPCT=	Displays only the categories whose percentages are greater than the specified percentage
OTHER=	Merges all categories that are not displayed
OTHERCVAL=	Specifies an OUT= data set character variable value for the OTHER= category
OTHERNVAL=	Specifies an OUT= data set numeric variable value for the OTHER= category
Options for Displaying Bars	
BARLABEL=	Displays labels for bars
BARS=	Specifies a variable that groups bars for a display by using ODS style colors
CHIGH(*n*)	Specifies color for bars that have the *n* highest values
CLOW(*n*)	Specifies color for bars that have the *n* lowest values
LABOTHER=	Specifies a label for the OTHER= category
LAST=	Specifies the bottommost category
Options for the Cumulative Percentage Curve	
ANCHOR=	Specifies the corner of topmost bar to which the curve is anchored
CMPCTLABEL	Labels curve points with their values
NOCURVE	Suppresses the cumulative percentage curve
NOCUMLABEL	Suppresses the cumulative percentage axis label
NOCUMTICK	Suppresses the cumulative percentage axis tick marks and tick mark labels
Options for Comparative Pareto Charts	
CLASS=	Specifies classification variables
CLASSKEY=	Specifies the key cell
CPROP	Requests proportion-of-frequency bars
INTERTILE=	Specifies the distance in screen percentage units between tiles
MISSING1	Requests that missing values of the first CLASS= variable be analyzed as a level
MISSING2	Requests that missing values of the second CLASS= variable be analyzed as a level
NCOLS=	Specifies the number of columns
NOKEYMOVE	Suppresses the placement of the key cell in the top left corner
NROWS=	Specifies the number of rows
ORDER1=	Specifies the order in which values of the first CLASS= variable are displayed
ORDER2=	Specifies the order in which values of the second CLASS= variable are displayed

Table 16.3 (*continued*)

Option	Description
Options for Controlling Axes	
AXISFACTOR=	Specifies the distance factor between the longest bar and the right frame
FREQAXIS=	Specifies tick mark values for the frequency axis
FREQAXISLABEL=	Labels the frequency axis
CUMAXIS=	Specifies tick mark values for the cumulative percentage axis
CUMAXISLABEL=	Specifies a label for the cumulative percentage axis
FREQOFFSET=	Specifies the frequency axis offset in screen percentage units
GRID	Adds a grid that corresponds to the frequency axis
GRID2	Adds a grid that corresponds to the cumulative percentage axis
NOCHART	Suppresses the Pareto chart
NOFREQLABEL	Suppresses the frequency axis label
NOCUMLABEL	Suppresses the cumulative percentage axis label
NOFREQTICK	Suppresses tick marks and tick mark labels for the frequency axis
NOCUMTICK	Suppresses tick marks and tick mark labels for the cumulative percentage axis
NOCATLABEL	Suppresses the category axis label
SCALE=	Specifies the units in which the frequency axis is scaled
CATOFFSET=	Specifies the category axis offset in screen percentage units
Options for Reference Lines	
CATREF=	Requests reference lines perpendicular to the category axis
CATREFLABELS=	Specifies labels for CATREF= lines
CUMREF=	Requests reference lines perpendicular to the cumulative percentage axis
CUMREFLABELS=	Specifies labels for CUMREF= lines
FREQREF=	Requests reference lines perpendicular to the frequency axis
FREQREFLABELS=	Specifies labels for FREQREF= lines
HREFLABPOS=	Specifies the position of FREQREFLABELS= and CUMREFLABELS= labels
VREFLABPOS=	Specifies the position of CATREFLABELS= labels
Options for Displaying Legends	
BARLEGEND=	Displays a legend for the BARS=, CBARS=, or PBARS= options
BARLEGLABEL=	Displays a label for the BARLEGEND= legend
CATLEGLABEL=	Specifies a label for the Pareto categories legend
CFRAMENLEG	Frames the sample size legend
HLLEGLABEL=	Displays a label for the legend that describes colors and patterns of the highest or lowest bars
NLEGEND=	Requests a sample size legend
NOHLLEG	Suppresses the legend that describes colors and patterns of the highest and lowest bars

Table 16.3 (*continued*)

Option	Description
Options for ODS Graphics Output	
CATLEGEND=	Controls the display of the Pareto categories legend
CHARTTYPE=	Specifies the type of Pareto chart to be produced
MARKERS	Requests markers on the cumulative percentage curve
ODSFOOTNOTE=	Specifies a footnote to be displayed on the chart
ODSFOOTNOTE2=	Specifies a secondary footnote to be displayed on the chart
ODSTITLE=	Specifies a title to be displayed on the chart
ODSTITLE2=	Specifies a secondary title to be displayed on the chart
URL=	Specifies a variable whose values are URLs to be associated with bars
Options for Traditional Graphics	
ANGLE=	Rotates category axis tick mark labels
ANNOKEY	Applies annotation only to the key cell
ANNOTATE=	Specifies an annotation data set that uses frequency axis data units
ANNOTATE2=	Specifies an annotation data set that uses cumulative percentage axis data units
BARLABPOS=	Specifies the position of BARLABEL= labels
BARWIDTH=	Specifies the width (vertical dimension) of the bars in screen percentage units
CAXIS=	Specifies the axis color
CAXIS2=	Specifies the color for the cumulative percentage axis and tick marks
CBARLINE=	Specifies the color for bar outlines
CBARS=	Specifies the color for bars
CCATREF=	Specifies the color for CATREF= lines
CCONNECT=	Specifies the color for the curve
CCUMREF=	Specifies the color for CUMREF= lines
CFRAME=	Specifies the color for the area enclosed by axes and frame
CFRAMESIDE=	Specifies the frame color for row labels
CFRAMETOP=	Specifies the frame color for column labels
CFREQREF=	Specifies the color for FREQREF= lines
CGRID=	Specifies the color for frequency axis grid lines
CGRID2=	Specifies the color for cumulative percentage axis grid lines
CLIPREF	Draws reference lines behind bars
COTHER=	Specifies the color for OTHER= bar
CTEXT=	Specifies the color for text
CTEXTSIDE=	Specifies the color for row labels
CTEXTTOP=	Specifies the color for column labels
CTILES=	Specifies the colors for tile backgrounds
DESCRIPTION=	Specifies a description of the Pareto chart's GRSEG catalog entry
FONT=	Specifies the font for text

Table 16.3 (*continued*)

Option	Description
FRONTREF	Draws reference lines in front of bars
HEIGHT=	Specifies the text height in screen percentage units
HTML=	Specifies a variable whose values create links that are associated with bars in traditional graphics output
INFONT=	Specifies the font for text inside the frame
INHEIGHT=	Specifies the text height in screen percentage units for text inside the frame
INTERBAR=	Specifies the distance between bars in screen percentage units
LCATREF=	Specifies the line type for CATREF= lines
LCUMREF=	Specifies the line type for CUMREF= lines
LFREQREF=	Specifies the line type for FREQREF= lines
LGRID=	Specifies the line type for frequency axis grid lines
LGRID2=	Specifies the line type for cumulative percentage axis grid lines
NAME=	Specifies the name of the Pareto chart's GRSEG catalog entry
NOFRAME	Suppresses the axis frame
PBARS=	Specifies the pattern for the bars
PHIGH(*n*)=	Specifies the pattern for bars that have the *n* highest values
PLOW(*n*)=	Specifies the pattern for bars that have the *n* lowest values
POTHER=	Specifies the pattern for the OTHER= bar
TILELEGEND=	Specifies a legend for the CTILES= colors
TILELEGLABEL=	Specifies label for TILELEGEND= legend
WAXIS=	Specifies the width in pixels for the axes and frame
WBARLINE=	Specifies the width for bar outlines
WGRID=	Specifies the width of frequency axis grid lines
WGRID2=	Specifies the width of cumulative percentage axis grid lines

INSET Statement

INSET *keyword-list* < / *options* > **;**

The INSET statement enables you to enhance a Pareto chart by adding a box or table (called an *inset*) of summary statistics directly to the graph. An inset can display statistics that are calculated by the PARETO procedure or arbitrary values that are provided in a SAS data set.

An INSET statement must follow a chart statement, and it produces an inset on that chart. More than one INSET statement can apply to the same chart statement. When the chart statement produces a comparative chart, an associated INSET statement produces an inset in every cell of the chart. Statistics that are displayed in the inset of a cell are computed from the data that are associated with that cell.

NOTE: When ODS Graphics is enabled, only one INSET statement can be associated with a comparative Pareto chart. Insets are not available with legacy line printer charts, so the INSET statement is not applicable when you specify the LINEPRINTER option in the PROC PARETO statement.

The *keyword-list* can include any of the keywords listed in Table 16.4. Statistics are displayed in the order in which the keywords are specified. Each *keyword-list* entry has the following form:

keyword <=‘*label*’> <(*format*)>

By default, inset statistics are identified with appropriate labels, and numeric values are printed using appropriate formats. However, you can provide customized labels and formats. You provide a customized label by specifying the *keyword* for that statistic followed by an equal sign (=) and the label in quotation marks. Labels can have up to 24 characters. You provide the numeric format in parentheses after the *keyword*. If you specify both a label and a format for a statistic, the label must appear before the format. See Example 16.10.

Note the difference between *keywords* and *options*: *keywords* specify the information to be displayed in an inset, whereas *options* control the appearance of the inset. You can use *options* in the INSET statement to do the following:

- specify the position of the inset
- specify a header for the inset
- specify enhancements for traditional graphics, such as background colors, text colors, text height, text font, and drop shadows

Table 16.5 lists available INSET statement *options*.

The following statements produce a vertical Pareto chart with insets in the upper left (northwest) and upper right (northeast) corners, and a horizontal comparative Pareto chart with insets in each cell.

```
proc pareto data=Failure3;
   vbar Cause / maxncat = 5 other = 'Others';
      inset nothercat / position = nw;
      inset nother / position = ne;
   hbar Cause / class = Stage;
      inset n;
run;
```

INSET Statement Keywords

Table 16.4 lists the *keywords* available in the INSET statement.

Table 16.4 INSET Statement Keywords

Keyword	Description
DATA=*SAS-data-set*	Reads (label, value) pairs from a SAS data set
N	Specifies the sample size
NEXCL	Specifies the number of observations excluded from a restricted Pareto chart
NOTHER	Specifies the number of observations in the OTHER= category
NOTHERCAT	Specifies the number of categories merged to form the OTHER= category

Table 16.4 (*continued*)

Keyword	Description
SUMWGTS	Specifies the sum of weighted frequencies across all categories

The NOTHERCAT and NOTHER statistics are 0 if the OTHER= option is not specified. The NEXCL statistic is 0 if the OTHER= option is specified.

All INSET keywords request a single statistic in an inset, except for the DATA= keyword. The DATA= keyword specifies a SAS data set that contains (label, value) pairs to be displayed in an inset. The data set must contain the variables _LABEL_ (a character variable whose values provide labels for inset entries) and _VALUE_ (which can be character or numeric and provides values displayed in the inset). The label and value from each observation in the DATA= data set occupy one line in the inset. Example 16.11 illustrates the use of the DATA= keyword.

INSET Statement Options

Figure 16.8 illustrates the terms that are used in this section.

Figure 16.8 Inset Terms

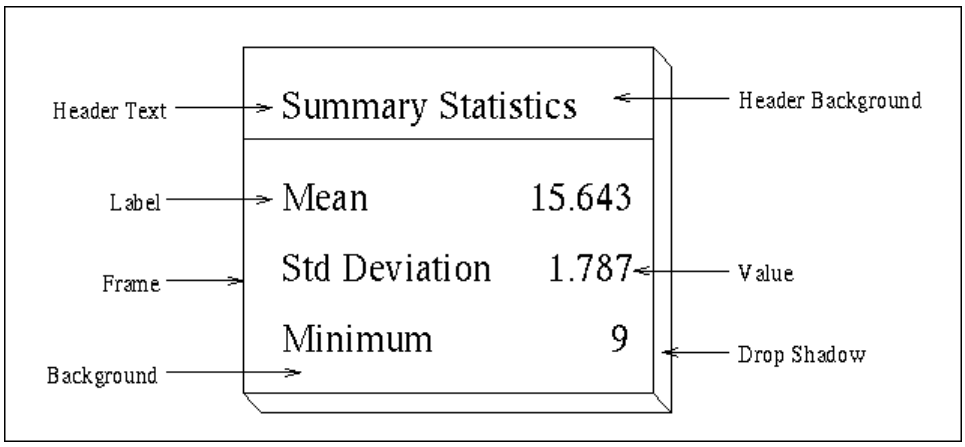


Table 16.5 lists the *options* available in the INSET statement.

Table 16.5 INSET Statement Options

Keyword	Description
General Options	
FORMAT=	Specifies the format for numeric values in the inset
HEADER=	Specifies the header text
NOFRAME	Suppresses the frame around the inset
POSITION=	Specifies the position of the inset

Table 16.5 (*continued*)

Option	Description
Options for ODS Graphics Output	
CFILL	Specifies the color of the inset background
GUTTER=	Specifies the gutter width for an inset in the top or bottom margin
NCOLS=	Specifies the number of columns for an inset in the top or bottom margin
Options for Traditional Graphics	
CFILL=	Specifies the color of the inset background
CFILLH=	Specifies the color of the header background
CFRAME=	Specifies the color of the frame
CHEADER=	Specifies the color of the header text
CSHADOW=	Specifies the color of the drop shadow
CTEXT=	Specifies the color of the inset text
DATA	Specifies that POSITION=(x,y) coordinates are in data units
FONT=	Specifies the text font
HEIGHT=	Specifies the height of the inset text
REFPOINT=	Specifies the reference point of an inset that is positioned by POSITION=(x,y) coordinates

The following entries provide detailed descriptions of *options* in the INSET statement.

General Options

You can specify the following *options* when you use either ODS Graphics or traditional graphics:

FORMAT=*format*
specifies a format for all the values that are displayed in an inset. If you specify a format for a particular statistic, then that format overrides the format you specify in this option.

HEADER= *'string'*
specifies the header text. The *string* cannot exceed 40 characters. If you do not specify this option, no header line appears in the inset.

NOFRAME
suppresses the frame drawn around the inset.

POSITION=*position*
POS=*position*
determines the position of the inset. The *position* can be a compass point keyword (N, NE, E, SE, S, SW, W, or NW), a margin keyword (TM, RM, BM, or LM), or a pair of coordinates *(x,y)*. You can specify coordinates in axis percentage units or axis data units. For more information, see the section "Positioning Insets" on page 1097. By default, POSITION=NW, which positions the inset in the upper left (northwest) corner of the display.

NOTE: You cannot use the POSITION= option to specify coordinates when producing ODS Graphics output.

ODS Graphics Options

You can specify the following *options* when you use ODS Graphics:

CFILL

CFILL=BLANK

specifies the color of the inset background. If you do not specify this option, the inset background is transparent. This means that items that are overlapped by the inset (such as Pareto bars or the cumulative percentage curve) show through the inset. If you specify this option without an argument, the background is opaque and its color is specified by the Color attribute of the GraphBackground style element in the current ODS style. If you specify CFILL=BLANK, the background is opaque and its color is specified by the Color attribute of the GraphWalls style element in the current ODS style.

GUTTER=*value*

specifies the gutter width in screen percentage units for an inset that is located in the top or bottom margin. The gutter is the space between columns of values in an inset.

NCOLS=*n*

specifies the number of columns of (label, value) pairs that are displayed in an inset that is located in the top or bottom margin.

Traditional Graphics Options

You can specify the following *options* when you produce traditional graphics.

CFILL=*color* | **BLANK**

specifies the color of the inset background (including the header background if you do not specify the CFILLH= option). If you specify CFILL=BLANK, the background color is determined by the Color attribute of the GraphWalls style element in the current ODS style. If you do not specify this option, the inset background is transparent. This means that items overlapped by the inset (such as Pareto bars or the cumulative percentage curve) show through the inset.

CFILLH=*color*

specifies the color of the header background. If you do not specify this option, the CFILL= color is used.

CFRAME=*color*

specifies the color of the inset frame. The default color is specified by the ContrastColor attribute of the GraphBorderLines style element in the current ODS style.

CHEADER=*color*

specifies the color of the header text. If you do not specify this option, the CTEXT= color is used.

CSHADOW=*color*

CS=*color*

specifies the color of the drop shadow. See Output 16.11.1 for an example. If you do not specify this option, a drop shadow is not displayed.

CTEXT=*color*

CT=*color*

specifies the color of the text. The default *color* is specified by the Color attribute of the GraphValueText style element in the current ODS style.

DATA

specifies that data coordinates be used in positioning the inset with the POSITION= option. You can specify this option only when you specify POSITION=(*x*,*y*), and you must include it immediately after the coordinates (*x*,*y*). For more information, see the section "Using Coordinates to Position Insets" on page 1099. See Figure 16.11 for an example.

FONT=*font*

specifies the font of the text. The default font is determined by the FontFamily, FontStyle, and FontWeight attributes of the GraphValueText style element in the current ODS style.

HEIGHT=*value*

specifies the height of the text in the inset. The default value is specified by the FontSize attribute of the GraphValueText style element in the current ODS style.

REFPOINT=BR | BL | TR | TL

RP=BR | BL | TR | TL

specifies the reference point for an inset that is positioned by a pair of coordinates (*x*,*y*), which are specified in the POSITION= option. The REFPOINT= option specifies which corner of the inset frame you want positioned at coordinates (*x*,*y*). The keywords BL, BR, TL, and TR represent bottom left, bottom right, top left, and top right, respectively. See Figure 16.12 for an example. By default, REFPOINT=BL.

If you specify the position of the inset as a compass point or margin keyword, this option is ignored. For more information, see "Using Coordinates to Position Insets" on page 1099.

VBAR Statement

VBAR (*variable-list* **)**<*/ options*> **;**

The VBAR statement creates a Pareto chart in which vertical bars represent the frequencies of problems in a process or operation. A vertical Pareto chart has a horizontal category axis. The frequency axis is oriented vertically on the left side of the chart and measures the lengths of the bars on the chart. The cumulative percentage axis is on the right of the chart and measures the cumulative percentage curve.

The VBAR statement produces three types of output for Pareto charts:

- It produces ODS Graphics output if ODS Graphics is enabled (for example, by specifying the ODS GRAPHICS ON statement prior to the PROC statement).
- Otherwise, it produces traditional graphics by default if SAS/GRAPH is licensed.
- It produces legacy line printer charts when you specify the LINEPRINTER option in the PROC statement.

For more information about producing these different types of graphs, see Chapter 4, "SAS/QC Graphics."

The *variable-list* specifies the process variables to be analyzed. A chart is created for each variable, and the values of each variable determine the Pareto categories for that chart. If *variable-list* contains only one process variable, you do not need to enclose it in parentheses.

The variables can be numeric or character, and the maximum length of a character variable is 64. Formatted values are used to determine the categories and are displayed in labels and legends. The maximum format length is 64.

Table 16.6 lists the VBAR statement options by function. For complete descriptions, see the section "Dictionary of HBAR and VBAR Statement Options" on page 1071.

Table 16.6 VBAR Statement Options

Option	Description
Data Processing Options	
FREQ=	Specifies the frequency variable
MISSING	Requests that missing values of the process variable be treated as a Pareto category
MISSING1	Requests that missing values of the first CLASS= variable be analyzed as a level
MISSING2	Requests that missing values of the second CLASS= variable be analyzed as a level
OUT=	Creates an output data set that saves the information that is displayed in the Pareto chart
WEIGHT=	Specifies weight variables that are used to weight frequencies
Options for Restricting the Number of Categories	
LOTHER=	Specifies a label for the OTHER= bar
MAXCMPCT=	Displays only the categories whose cumulative percentage is less than the specified percentage
MAXNCAT=	Displays only the categories that have the *n* highest values
MINPCT=	Displays only the categories that have percentages greater than the specified percentage
OTHER=	Merges all categories that are not displayed
OTHERCVAL=	Specifies an OUT= data set character variable value for the OTHER= category
OTHERNVAL=	Specifies an OUT= data set numeric variable value for the OTHER= category
Options for Displaying Bars	
BARLABEL=	Displays labels for bars
BARS=	Specifies a variable that groups bars for a display by using ODS style colors
CHIGH(*n*)	Specifies the color for bars that have the *n* highest values
CLOW(*n*)	Specifies the color for bars that have the *n* lowest values
LABOTHER=	Specifies a label for the OTHER= category
LAST=	Specifies the bottommost category

Table 16.6 (*continued*)

Option	Description
Options for the Cumulative Percent Curve	
ANCHOR=	Specifies the corner of the topmost bar to which the curve is anchored
CMPCTLABEL	Labels curve points with their values
NOCURVE	Suppresses the cumulative percentage curve
NOCUMLABEL	Suppresses the cumulative percentage axis label
NOCUMTICK	Suppresses cumulative percentage axis tick marks and tick mark labels
Options for Comparative Pareto Charts	
CLASS=	Specifies classification variables
CLASSKEY=	Specifies the key cell
CPROP	Requests proportion-of-frequency bars
INTERTILE=	Specifies the distance in screen percentage units between tiles
MISSING1	Requests that missing values of the first CLASS= variable be analyzed as a level
MISSING2	Requests that missing values of the second CLASS= variable be analyzed as a level
NCOLS=	Specifies the number of columns
NOKEYMOVE	Suppresses the placement of the key cell in the top left corner
NROWS=	Specifies the number of rows
ORDER1=	Specifies the order in which values of the first CLASS= variable are displayed
ORDER2=	Specifies the order in which values of the second CLASS= variable are displayed
Options for Controlling Axes	
AXISFACTOR=	Specifies the distance factor between the longest bar and the top frame
FREQAXIS=	Specifies tick mark values for the frequency axis
FREQAXISLABEL=	Labels the frequency axis
CUMAXIS=	Specifies tick mark values for the cumulative percentage axis
CUMAXISLABEL=	Specifies a label for the cumulative percentage axis
FREQOFFSET=	Specifies the frequency axis offset in screen percentage units
GRID	Adds a grid that corresponds to the frequency axis
GRID2	Adds a grid that corresponds to the cumulative percentage axis
NOCHART	Suppresses the Pareto chart
NOFREQLABEL	Suppresses the frequency axis label
NOCUMLABEL	Suppresses the cumulative percentage axis label
NOFREQTICK	Suppresses tick marks and tick mark labels for the frequency axis
NOCUMTICK	Suppresses tick marks and tick mark labels for the cumulative percentage axis
NOCATLABEL	Suppresses the category axis label
SCALE=	Specifies units in which the frequency axis is scaled

Table 16.6 (*continued*)

Option	Description
CATOFFSET=	Specifies the category axis offset in screen percentage units
Options for Reference Lines	
CATREF=	Requests reference lines perpendicular to the category axis
CATREFLABELS=	Specifies labels for CATREF= lines
CUMREF=	Requests reference lines perpendicular to the cumulative percentage axis
CUMREFLABELS=	Specifies labels for CUMREF= lines
FREQREF=	Requests reference lines perpendicular to the frequency axis
FREQREFLABELS=	Specifies labels for FREQREF= lines
HREFLABPOS=	Specifies the position of FREQREFLABELS= and CUMREFLABELS= labels
VREFLABPOS=	Specifies the position of CATREFLABELS= labels
Options for Displaying Legends	
BARLEGEND=	Displays a legend for the BARS=, CBARS=, or PBARS= options
BARLEGLABEL=	Displays a label for BARLEGEND= legend
CATLEGLABEL=	Specifies a label for the Pareto categories legend
CFRAMENLEG	Frames the sample size legend
HLLEGLABEL=	Displays a label for the legend that describes colors and patterns of the highest and lowest bars
NLEGEND=	Requests a sample size legend
NOHLLEG	Suppresses the legend that describes colors and patterns of the highest and lowest bars
Options for ODS Graphics Output	
CATLEGEND=	Controls the display of the Pareto categories legend
CHARTTYPE=	Specifies the type of Pareto chart produced
MARKERS	Requests markers on the cumulative percentage curve
ODSFOOTNOTE=	Specifies a footnote to be displayed on the chart
ODSFOOTNOTE2=	Specifies a secondary footnote to be displayed on the chart
ODSTITLE=	Specifies a title to be displayed on the chart
ODSTITLE2=	Specifies a secondary title to be displayed on the chart
URL=	Specifies a variable whose values are URLs to be associated with bars
Options for Traditional Graphics	
ANGLE=	Rotates the category axis tick mark labels
ANNOKEY	Applies annotation only to the key cell
ANNOTATE=	Specifies an annotation data set that uses frequency axis data units
ANNOTATE2=	Specifies an annotation data set that uses cumulative percentage axis data units
BARLABPOS=	Specifies the position of the BARLABEL= labels

Table 16.6 (*continued*)

Option	Description
BARWIDTH=	Specifies the width (horizontal dimension) of the bars in screen percentage units
CAXIS=	Specifies the axis color
CAXIS2=	Specifies the color for the cumulative percentage axis and tick marks
CBARLINE=	Specifies the color for bar outlines
CBARS=	Specifies the color for bars
CCATREF=	Specifies the color for CATREF= lines
CCONNECT=	Specifies the color for the curve
CCUMREF=	Specifies the color for CUMREF= lines
CFRAME=	Specifies the color for the area that is enclosed by axes and frame
CFRAMESIDE=	Specifies the frame color for row labels
CFRAMETOP=	Specifies the frame color for column labels
CFREQREF=	Specifies the color for FREQREF= lines
CGRID=	Specifies the color for the frequency axis grid lines
CGRID2=	Specifies the color for the cumulative percentage axis grid lines
CLIPREF	Draws reference lines behind bars
COTHER=	Specifies the color for the OTHER= bar
CTEXT=	Specifies the color for text
CTEXTSIDE=	Specifies the color for row labels
CTEXTTOP=	Specifies the color for column labels
CTILES=	Specifies the colors for tile backgrounds
DESCRIPTION=	Specifies a description of the Pareto chart's GRSEG catalog entry
FONT=	Specifies the text font
FRONTREF	Draws reference lines in front of bars
HEIGHT=	Specifies the text height in screen percentage units
HTML=	Specifies a variable whose values create links that are associated with bars in traditional graphics output
INFONT=	Specifies the font for text inside frame
INHEIGHT=	Specifies the text height in screen percentage units for text inside the frame
INTERBAR=	Specifies the distance between bars in screen percentage units
LCATREF=	Specifies the line type for the CATREF= lines
LCUMREF=	Specifies the line type for the CUMREF= lines
LFREQREF=	Specifies the line type for the FREQREF= lines
LGRID=	Specifies the line type for the frequency axis grid lines
LGRID2=	Specifies the line type for the cumulative percentage axis grid lines
NAME=	Specifies the name of the Pareto chart's GRSEG catalog entry
NOFRAME	Suppresses the axis frame
PBARS=	Specifies a pattern for the bars
PHIGH(*n*)=	Specifies the pattern for the bars that have the *n* highest values

Table 16.6 (*continued*)

Option	Description
PLOW(*n*)=	Specifies the pattern for the bars that have the *n* lowest values
POTHER=	Specifies the pattern for the OTHER= bar
TILELEGEND=	Specifies a legend for the CTILES= colors
TILELEGLABEL=	Specifies the label for the TILELEGEND= legend
TURNVLABEL	Turns and strings vertically the characters in the frequency and cumulative percentage axis labels
WAXIS=	Specifies the width in pixels for the axes and frame
WBARLINE=	Specifies the width for bar outlines
WGRID=	Specifies the width of frequency axis grid lines
WGRID2=	Specifies the width of cumulative percentage axis grid lines
Options for Legacy Line Printer Charts	
CONNECTCHAR=	Specifies the plot character for the cumulative percentage curve segments
HREFCHAR=	Specifies the plot character for category reference lines
VREFCHAR=	Specifies the plot character for frequency and cumulative percentage reference lines
SYMBOLCHAR=	Specifies the plot character for points on the cumulative percentage curve

Dictionary of HBAR and VBAR Statement Options

This section provides detailed descriptions of *options* you can specify after the slash (/) in the HBAR and VBAR statements. For example, to request that the frequency axis of a vertical Pareto chart be scaled by counts, use the SCALE= option as follows:

```
proc pareto data=failure;
   vbar cause / scale = count;
run;
```

This section consists of the following subsections:

- The section "General Options" on page 1072 contains descriptions of general Pareto chart options.
- The section "Options for Traditional Graphics" on page 1086 describes options that apply only when traditional graphics output is produced, as when ODS Graphics is disabled.
- The section "Options for Legacy Line Printer Charts" on page 1093 contains descriptions of options that apply only to legacy line printer charts, which are produced by VBAR statements when you specify the LINEPRINTER option in the PROC PARETO statement.

NOTE: The terminology used in the option descriptions describes vertical Pareto charts. For example, the "tallest" bar is the one that extends farthest along the frequency axis, whether it is oriented vertically or horizontally.

General Options

You can specify the following general options:

ANCHOR=*keyword*

specifies where the Pareto curve is anchored to the first bar on the chart. Table 16.7 describes the position keywords available in the HBAR and VBAR statements.

Table 16.7 ANCHOR= Option Keywords

HBAR Keyword	Anchoring Position
BR	Bottom right corner (default)
LC	Left center
RC	Right center
TL	Top left corner

VBAR Keyword	Anchoring Position
BC	Bottom center
BL	Bottom left corner
TC	Top center
TR	Top right corner (default)

See Output 16.2.1 for an illustration.

AXISFACTOR=*value*

specifies a factor used in scaling the frequency axis. This factor determines (approximately) the ratio of the length of the axis to the length of the tallest bar, and it is used to provide space for the cumulative percentage curve. The *value* must be greater than or equal to 1.

By default, the factor is chosen so that the curve is anchored at the top right corner of the first bar (see also the ANCHOR= option). However, if anchoring to the top of the first bar causes the bars to be flattened excessively, a smaller default factor is used.

This option is not applicable if the cumulative percentage curve is suppressed by the NOCURVE option.

BARLABEL=CMPCT | COUNT | VALUE | (*variable-list***)**

requests that a label be displayed for each bar. You can specify the following values:

CMPCT specifies that the label indicates the cumulative percentage that is associated with that bar. An alternative to BARLABEL=CMPCT is the CMPCTLABEL option, which labels points on the cumulative percentage curve with their values.

COUNT specifies that the label displays the count for the bar, regardless of the SCALE= option setting.

VALUE specifies that the label indicates the height of the bar in the units used by the frequency axis. The units are determined by the SCALE= option setting. See Example 16.8 for an illustration.

(*variable-list*) specifies that the label displays the values of one or more variables from the input data set. If a format is associated with a variable, then the formatted value is displayed. Values can be up to 32 characters long. The variable values must be consistent within observations that correspond to a particular Pareto category. The variables are saved in the OUT= data set. If you specify more than one process variable in the chart statement, you can specify more than one variable in *variable-list*. The BARLABEL= and process variables are matched by their positions in their respective variable lists.

The space in horizontal Pareto charts might be insufficient to display long bar labels. You can specify the AXISFACTOR= option to increase the available space beyond the bars. If you are producing traditional graphics, you can use the BARLABPOS= option to specify how labels are positioned relative to the bars.

BARLEGEND=(*variable-list*)

requests that a legend be added to the chart to explain colors for bars that are specified in the BARS= or CBARS= option, or patterns for bars that are specified in the PBARS= option. The *variable-list* must be enclosed in parentheses even if only one *variable* is specified. See Output 16.4.1 for an illustration.

The values of the variables in *variable-list* provide the explanatory labels used in the legend. If a format is associated with the variable, then the formatted value is displayed. Values can be up to 32 characters long.

This option is not applicable unless you specify one or more of the BARS=, CBARS=, or PBARS= options. In the DATA= data set, the values of the BARLEGEND= variable must be identical in observations for which the value of the BARS=, CBARS=, or PBARS= variable (or the combination of the CBARS= and PBARS= values) is the same. This ensures that the legend derived from the BARLEGEND= variable is consistent.

If you specify more than one process variable in a chart statement and a corresponding list of BARS=, CBARS=, or PBARS= variables, you can specify a list of BARLEGEND= variables. The number of variables in *variable-list* should be less than or equal to the number of process variables. The lists of variables are matched so that the first variable in *variable-list* is applied to the first process variable and the first BARS=, CBARS=, or PBARS= variable; the second variable in *variable-list* is applied to the second process variable and the second BARS=, CBARS=, or PBARS= variable; and so forth. If the process variable list is longer than *variable-list*, the charts for the extra process variables do not display a bar legend.

BARLEGLABEL='*label*'

specifies the *label* to be displayed to the left of the legend that is created by the BARLEGEND= option. See Output 16.4.1 for an illustration.

The BARLEGLABEL= option is applicable only in conjunction with BARS=, CBARS=, or PBARS= variables. The *label* can be up to 16 characters and must be enclosed in quotation marks.

If you do not specify a *label*, the BARLEGEND= variable label is displayed (unless the label is longer than 16 characters, in which case the variable name is displayed). If you do not specify the BARLEGLABEL= option and no label is associated with the BARLEGEND= variable, no legend label is displayed.

BARS=(*variable-list*)

uses different colors to group bars of the Pareto chart for display Bars that correspond to the same value of a variable in *variable-list* are assigned the same color from the ODS style. You cannot specify the BARS= option in conjunction with the CHIGH(*n*) or CLOW(*n*) options.

If you specify more than one process variable, you can specify more than one variable in *variable-list*. The number of variables in *variable-list* should be less than or equal to the number of process variables. The two lists of variables are paired in order of their specification. If a BARS= *variable* is not provided for a process variable, the bars for that chart are filled with the default color from the ODS style.

CATLEGEND=AUTO | OFF | ON

specifies whether a category legend is created for ODS Graphics output. You can specify the following values:

AUTO	creates a category legend only when the labels would be too crowded on the category axis.
OFF	suppresses the category legend.
ON	creates a category legend.

By default, CATLEGEND=AUTO. This option is ignored if ODS Graphics is not enabled.

CATLEGLABEL=‘*label*’

specifies a label for the category legend. A category legend is created when there is insufficient space to label the categories along the category axis or when requested in the CATLEGEND= option. The *label* can be up to 16 characters and must be enclosed in quotation marks. The default label is “Categories:”. See Example 16.3 for an illustration. This option is ignored when no category legend is produced.

CATOFFSET=*value*

specifies the length of the offset at both ends of the category axis (in screen percentage units). You can eliminate the offset by specifying CATOFFSET=0.

CATREF=‘*value-list*’

specifies where reference lines perpendicular to the Pareto category axis are to appear on the chart. Character values can be up to 64 characters and must be enclosed in quotation marks. The values must be values of the process variable regardless of whether the bars are numbered and a category legend is introduced.

CATREFLABELS=‘*label1*’. . .‘*labeln*’

specifies *labels* for the lines that are requested in the CATREF= option. The number of labels must equal the number of lines requested. Labels can be up to 16 characters and must be enclosed in quotation marks.

CFRAMENLEG

CFRAMENLEG=EMPTY

CFRAMENLEG=*color*

displays a frame around the sample size legend that is requested in the NLEGEND option. You can specify this option in the following ways:

(no argument) fills the frame with the background color that is specified by the Color attribute of the GraphBackground style element in the current ODS style.

EMPTY produces a frame that has a transparent background.

color produces a frame whose background is *color* when you are producing traditional graphics.

CHARTTYPE=CUMULATIVE | INTERVALS< (*interval-options*) > | STANDARD

specifies the type of Pareto chart to be produced. This option is supported only for ODS Graphics output. You can specify the following options:

CUMULATIVE creates a cumulative Pareto bar chart.

INTERVALS< (*interval-options*) > creates a Pareto dot plot that includes acceptance intervals, which are computed using simulation. You can specify the following *interval-options* for computing acceptance intervals:

ALPHA=*value*
specifies the significance level for the acceptance intervals. By default, ALPHA=0.05.

NSAMPLES=*n*
specifies the number of random samples used in the simulation. By default, NSAMPLES=2000.

SEED=*n*
specifies the seed value for the random number generator that is used in the simulation. By default, or when you specify $n \leq 0$, a seed value is generated by using the system clock.

STANDARD creates a traditional Pareto chart.

By default, CHARTTYPE=STANDARD.

Wilkinson (2006) describes the advantages of the cumulative Pareto bar chart and the Pareto dot plot that includes acceptance intervals. See Example 16.9 for examples of these alternative Pareto charts.

CHIGH(*n*)

CHIGH(*n*)=*color*

highlights the bars that have the *n* highest frequencies by filling them with a contrasting color from the ODS style. When producing traditional graphics output, you can specify CHIGH(*n*)=*color* to select a specific color. You cannot use the CHIGH(*n*) option in conjunction with a BARS= or CBARS= variable, but you can use it together with the CLOW(*n*) and CBARS=*color* options. See Output 16.3.1 for an illustration.

CLASS=*variable*

CLASS=(*variable1 variable2*)

creates a comparative Pareto chart by using the levels of the *variables*. If you specify two *variables*, then you must enclose in parentheses. See Example 16.1 and Example 16.2.

If you specify a single *variable*, the observations in the input data set are classified by the formatted values (levels) of the *variable*. A Pareto chart is created for the process variable values in each level,

and these component charts (referred to as cells) are arranged in an array. The cells are labeled with the class levels, and uniform horizontal and vertical axes are used to facilitate comparisons.

If you specify two *variables*, the observations in the input data set are cross-classified by the values (levels) of the *variables*. A Pareto chart is created for the process variable values in each cell of the cross-classification, and these charts are arranged in a matrix. The levels of the first *variable* label the rows, and the levels of the second *variable* label the columns. Uniform horizontal and vertical axes are used to facilitate comparisons.

The *variables* can be numeric or character. The maximum length of a character *variable* is 32. If a format is associated with a *variable*, the formatted values determine the levels. Only the first 32 characters of the formatted values are used to determine the levels. You can specify whether missing values are treated as a level by using the MISSING1 and MISSING2 options.

In traditional graphics output, only the level values are displayed in row and column headers. If a label is associated with a *variable*, the label is displayed in a second header that spans the row or column headers.

CLASSKEY='*value*'

CLASSKEY=('*value1*' '*value2*'**)**

specifies the key cell in a comparative Pareto chart, which is created when you specify the CLASS= option. The *key cell* is defined as the cell in which the Pareto bars are arranged in decreasing order. This order then determines the uniform category axis used for all the cells.

If you specify CLASS=*variable*, you can specify CLASSKEY='*value*' to identify the key cell as the level for which the variable is equal to *value*. The *value* can have up to 32 characters, and you must specify a formatted *value*. By default, the levels are sorted as specified by the ORDER1= option, and the key cell is the level that occurs first in this order. The cells are displayed in this order from top to bottom (or left to right, depending on the NCOLS= and NROWS= values), and consequently the key cell is displayed at the top or at the left. The cell you specify in the CLASSKEY= option is displayed at the top or at the left unless you also specify the NOKEYMOVE option.

If you specify CLASS=(*variable1 variable2*), you can specify CLASSKEY=('*value1*' '*value2*') to identify the key cell as the level for which *variable1* is equal to *value1* and *variable2* is equal to *value2*. Here, *value1* and *value2* must be formatted values, and they must be enclosed in quotation marks. By default, the levels of *variable1* are sorted in the order determined by the ORDER1= option, and then within each of these levels, the levels of *variable2* are sorted in the order determined by the ORDER2= option. The default key cell is the combination of levels of *variable1* and *variable2* that occurs first in this order. The cells are displayed in order of *variable1* from top to bottom and in order of *variable2* from left to right. Consequently, the default key cell is displayed in the upper left corner. The cell you specify in the CLASSKEY= option is displayed in the upper left corner unless you also specify the NOKEYMOVE option.

For an example of the use of the CLASSKEY= option, see Output 16.1.3.

CLOW(*n***)**

CLOW(*n***)=***color*

highlights the bars that have the *n* lowest frequencies by filling them with a contrasting color from the ODS style. When producing traditional graphics output, you can specify CLOW(*n*)=*color* to select a specific color. You cannot use the CLOW(*n*)= option in conjunction with a CBARS= variable, but you can use it together with the CBARS=*color* and CHIGH(*n*) options.

CMPCTLABEL

labels points on the cumulative percentage curve with their values. By default, the points are not labeled.

CPROP

CPROP=EMPTY

CPROP=*color*

requests that a proportion-of-frequency bar of the specified color be displayed horizontally across the top of each tile in a comparative Pareto chart. You can specify the following values:

(no argument)	creates bars that are filled with a color from the ODS style.
EMPTY	produces empty bars in traditional graphics output.
color	produces bars that are filled with *color* in traditional graphics output.

The length of the bar relative to the width of the tile indicates the proportion of the total frequency count in the chart that is represented by the tile. You can use the bars to visualize the distribution of frequency count by tile. See Output 16.1.4 for an illustration.

The CPROP= option provides a graphical alternative to the NLEGEND option, which displays the actual count. The CPROP= option is applicable only with comparative Pareto charts.

CUMAXIS=*value-list*

specifies tick mark values for the cumulative percentage axis. The values must be equally spaced and in increasing order, and the first value must be 0. You must scale the values in percentage units, and the last value must be greater than or equal to 100.

CUMAXISLABEL='*label*'

specifies a *label*, up to 40 characters, for the cumulative percentage axis. The default *label* is "Cumulative Percent" or "Cm Pct," depending on the space available.

CUMREF=*value-list*

requests reference lines perpendicular to the cumulative percentage axis at the specified *values*. You must specify the values in cumulative percentage units.

CUMREFLABELS='*label1*'...'*labeln*'

specifies labels for the lines that are requested in the CUMREF= option. The number of labels must equal the number of lines requested. Enclose the labels in quotation marks. Labels can be up to 16 characters.

FREQ=*variable*

specifies a frequency *variable* whose values provide the counts (numbers of occurrences) of the values of the process variable. Specifying a frequency *variable* is equivalent to replicating the observations in the input data set. The *variable* must be a numeric variable that has nonnegative integer values. See "Creating a Pareto Chart from Frequency Data" on page 1049 for an illustration. If you specify more than one process variable in the chart statement, the *variable* values are used with each process variable. If you do not specify this option, each value of the process variable is counted exactly once.

FREQAXIS=*value-list*
specifies tick mark values for the frequency axis. The values must be equally spaced and in increasing order, and the first value must be 0. You must scale the values in the same units as the bars (see the SCALE= option), and the last value must be greater than or equal to the height of the largest bar.

FREQAXISLABEL='*label*'
specifies a label, up to 40 characters, for the frequency axis. If a WEIGHT= variable is specified, its label is the default frequency axis label. Otherwise, the default label depends on the value of the SCALE= option.

FREQOFFSET=*value*
specifies the length in screen percentage units of the offset at the upper end of the frequency axis.

FREQREF=*value-list*
specifies where reference lines perpendicular to the frequency axis are to appear on the chart. You must specify the values in the same units that are used to scale the frequency axis. By default, the frequency axis is scaled in percentage units, but you can specify other units in the SCALE= option. See Output 16.2.3 for an illustration.

FREQREFLABELS='*label1*'...'*labeln*'
specifies labels for the lines that are requested in the FREQREF= option. The number of labels must equal the number of lines requested. Enclose the labels in quotation marks. Labels can be up to 16 characters.

GRID
adds a grid that corresponds to the frequency axis to the Pareto chart. Grid lines are positioned at tick marks on the frequency axis. The lines are useful for comparing the heights of the bars.

GRID2
adds a grid that corresponds to the cumulative percentage axis to the Pareto chart. Grid lines are positioned at tick marks on the cumulative percentage axis. The lines are useful for reading the cumulative percentage curve.

HLLEGLABEL='*label*'
specifies a label for the legend that is automatically created when you use a combination of the CHIGH(*n*), CLOW(*n*), PHIGH(*n*), and PLOW(*n*) options. See Output 16.3.1 for an illustration. The *label* can be up to 16 characters and must be enclosed in quotation marks. The default label is "Bars:".

HREFLABPOS=*n*
specifies the vertical position of labels for reference lines that are associated with horizontal axes, which are specified in the FREQREF= and CUMREF= options in an HBAR statement or the CATREF= option in a VBAR statement. The available positions are described in the following table.

n	**Position**
1	Along top of chart
2	Staggered from top to bottom of chart
3	Along bottom of chart
4	Staggered from bottom to top of chart

By default, HREFLABPOS=1. **NOTE:** HREFLABPOS=2 and HREFLABPOS=4 are not supported for ODS Graphics output.

INTERTILE=*value*

specifies the distance in horizontal screen percentage units between tiles (cells) in a comparative Pareto chart. When ODS Graphics is enabled, the default value is 2%. In traditional graphics, the tiles are contiguous by default. See Output 16.1.3 for an illustration.

LABOTHER= '*other-label*'

is used in conjunction with the BARLABEL=(*variable*) option and specifies a label for the 'other' category that is optionally specified in the OTHER= option.

LAST='*category*'

requests that the bar that corresponds to *category* be displayed last (at the bottom of a horizontal chart or the right end of a vertical chart) regardless of the frequency that is associated with this category. The category must be a formatted value of the process variable and must be enclosed in quotation marks. The *category* can be up to 64 characters. See Figure 16.6 for an illustration.

LOTHER='*label*'

specifies a label for the bar that is defined in the OTHER= option. This label appears in the legend that is specified in the BARLEGEND= option. The *label* must be enclosed in quotation marks and can be up to 32 characters. The default is the value that is specified in the OTHER= option. The LOTHER= option is applicable only when a BARLEGEND= variable is specified.

MARKERS

requests that the points on the cumulative percentage curve be plotted with markers in ODS Graphics output. You can use a SYMBOL statement to plot the points in traditional graphics output.

MAXCMPCT=*percent*

requests that only the Pareto categories that have the highest frequency counts be displayed, where the sum of their corresponding percentages is less than or equal to *percent*. For example, if you specify the following statements, the chart displays only the most frequently occurring categories that account for no more than 90% of the total frequency:

```
proc pareto data=failure;
   vbar cause / maxcmpct = 90;
```

You can use the OTHER= option in conjunction with the MAXCMPCT= option to create and display a new category that combines categories that are not selected by the MAXCMPCT= option. For example, if you specify the following statements, the chart displays the categories that account for no more than 90% of the total frequency, together with a category labeled "Others" that merges the remaining categories:

```
proc pareto data=failure;
   vbar cause / maxcmpct = 90
                other    = 'Others';
```

The MAXCMPCT= option is an alternative to the MINPCT= and MAXNCAT= options.

MAXNCAT=*n*

requests that only the Pareto categories with the *n* highest frequencies be displayed. For example, if you specify the following statements, the chart displays only the categories that have the 20 highest frequencies:

```
proc pareto data=failure;
   vbar cause / maxncat = 20;
```

If the total number of categories is less than 20, all the categories are displayed.

You can use the OTHER= option in conjunction with the MAXNCAT= option to create and display a new category that combines categories that are not selected by the MAXNCAT= option. For example, if you specify the following statements, the chart displays the categories that have the 19 highest frequencies, together with a category labeled "Others" that merges the remaining categories:

```
proc pareto data=failure;
   vbar cause / maxncat = 20
                other= 'Others';
```

See Figure 16.6 for another illustration.

The MAXNCAT= option is an alternative to the MINPCT= and MAXCMPCT= options.

MINPCT=*percent*

requests that only the Pareto categories whose frequency percentages are greater than or equal to *percent* be displayed. For example, if you specify the following statements, the chart displays only categories that have at least 5% of the total frequency:

```
proc pareto data=failure;
   vbar cause / minpct = 5;
```

You can use the OTHER= option in conjunction with the MINPCT= option to create and display a new category that combines categories that are not selected by the MINPCT= option. The merged category that is created by the OTHER= option is displayed even if its total percentage is less than *percent*. For example, if you specify the following statements, the chart displays the categories whose percentages are greater than or equal to 5%, together with a category labeled "Others" that merges the remaining categories:

```
proc pareto data=failure;
   vbar cause / minpct = 5
                other  = 'Others';
```

The MINPCT= option is an alternative to the MAXNCAT= and MAXCMPCT= options.

MISSING

requests that missing values of the process variable be treated as a Pareto category that is represented with a bar on the chart. If the process variable is a character variable, a missing value is defined as a blank internal (unformatted) value. If the process variable is numeric, a missing value is defined as any of the SAS missing values. If you do not specify this option, missing values are excluded from the analysis.

MISSING1

requests that missing values of the first CLASS= variable be treated as a level of the CLASS= variable. If the first CLASS= variable is a character variable, a missing value is defined as a blank internal (unformatted) value. If the first CLASS= variable is numeric, a missing value is defined as any of the SAS missing values. If you do not specify this option, observations in the DATA= data set for which the first CLASS= variable is missing are excluded from the analysis.

MISSING2

requests that missing values of the second CLASS= variable be treated as a level of the CLASS= variable. If the second CLASS= variable is a character variable, a missing value is defined as a blank internal (unformatted) value. If the second CLASS= variable is numeric, a missing value is defined as any of the SAS missing values. If you do not specify this option, observations in the DATA= data set for which the second CLASS= variable is missing are excluded from the analysis.

NCOLS=*n*

NCOL=*n*

specifies the number of columns in a comparative Pareto chart. You can use this option in conjunction with the NROWS= option. See Output 16.2.3 and Output 16.2.4 for an illustration. By default, NCOLS=1 and NROWS=2 if one CLASS= variable is specified, and NCOLS=2 and NROWS=2 if two CLASS= variables are specified.

NLEGEND

NLEGEND=‘*label*’

NLEGEND=(*variable***)**

requests a sample size legend and specifies its form. You can specify the following values:

(no argument) requests a sample size legend and specifies its form as *N*=*n*, where *n* is the total count for the Pareto categories. In a comparative Pareto chart, a legend is displayed in each tile, and *n* is the total count for that particular cell. See Output 16.2.1 for an illustration.

‘*label*’ requests a sample size legend and specifies its form as *label*=*n*, where *n* is the total count for the Pareto categories. The *label* can be up to 32 characters and must be enclosed in quotation marks. For an illustration, see Figure 16.4 or Output 16.1.4.

(*variable***)** requests a sample size legend that is the value of *variable* from the DATA= data set. The formatted length of *variable* cannot exceed 32. If a format is associated with *variable*, then the formatted value is displayed. This option is intended for use with comparative Pareto charts and enables you to display a customized legend inside each tile (this legend does not need to provide a total count). It is assumed that the values of *variable* are identical for all observations in a particular class.

By default, the legend is placed in the upper left corner of the chart. If you specify the NOCURVE option, the legend is placed in the upper right corner of the chart. You can use the CFRAMENLEG= option to frame the sample size legend. No sample size legend is displayed if you do not specify an NLEGEND option.

NOCATLABEL
suppresses the category axis label. This option is useful for avoiding clutter where the meaning of the category axis is apparent from the labels for the Pareto categories. See Output 16.2.2 for an illustration.

NOCHART
suppresses the creation of a Pareto chart. This option is useful when you are simply creating an output data set.

NOCUMLABEL
suppresses the cumulative percentage axis label. This option is useful for avoiding clutter on comparative Pareto charts.

NOCUMTICK
suppresses the cumulative percentage axis label, tick marks, and tick mark labels.

NOCURVE
suppresses the cumulative percentage curve and the cumulative percentage axis. Compare Output 16.2.1 and Output 16.2.2 for an illustration.

NOFREQLABEL
suppresses the frequency axis label.

NOFREQTICK
suppresses the frequency axis label, tick marks, and tick mark labels.

NOHLLEG
suppresses the legend that is generated by the CHIGH(*n*)=, CLOW(*n*)=, PHIGH(*n*)=, and PLOW(*n*)= options.

NOKEYMOVE
suppresses the rearrangement of cells within a comparative Pareto chart that occurs when you use the CLASSKEY= option. By default, the key cell appears in the top left corner of a comparative Pareto chart.

NROWS=*n*
NROW=*n*
specifies the number of rows in a comparative Pareto chart. You can use the NROWS= option in conjunction with the NCOLS= option. See Output 16.2.3 and Output 16.2.4 for an illustration. By default, NROWS=2.

ODSFOOTNOTE=FOOTNOTE | FOOTNOTE1 | '*string*'
adds a footnote to ODS Graphics output. You can specify the following values:

FOOTNOTE (or **FOOTNOTE1**) uses the value of the SAS FOOTNOTE statement as the graph footnote.

'*string*' uses *string* as the footnote. The quoted *string* can contain either of the following escaped characters, which are replaced with the appropriate values from the analysis:

\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 ODS Graphics output. You can specify the following values:

FOOTNOTE2 — uses the value of the SAS FOOTNOTE2 statement as the secondary graph footnote.

'*string*' — uses *string* as the secondary footnote. The quoted *string* can contain any of the following escaped characters, which are replaced with the appropriate values from the analysis:

\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 | LABELFMT | '*string*'
specifies a title for ODS Graphics output. You can specify the following values:

TITLE (or **TITLE1**) — uses the value of the SAS TITLE statement as the graph title.

NONE — suppresses all titles from the graph.

DEFAULT — uses the default ODS Graphics title (a descriptive title that consists of the plot type and the process variable name).

LABELFMT — uses the default ODS Graphics title, but substitutes the process variable label for the process variable name.

'*string*' — uses *string* as the graph title. The quoted *string* can contain the following escaped characters, which are replaced with the appropriate values from the analysis:

\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).

ODSTITLE2=TITLE2 | '*string*'
specifies a secondary title for ODS Graphics output. You can specify the following values:

TITLE2 — uses the value of the SAS TITLE2 statement as the secondary graph title.

'*string*' — uses *string* as the graph title. The quoted *string* can contain the following escaped characters, which are replaced with the appropriate values from the analysis:

\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).

ORDER1=DATA | FORMATTED | FREQ | INTERNAL

specifies the display order for the values of the first CLASS= variable. The levels of the first CLASS= variable are always constructed using the formatted values of the variable, and the formatted values are always used to label the rows (columns) of a comparative Pareto chart. You can specify the following values:

DATA displays the rows (columns) from top to bottom (left to right) in the order in which the values of the first CLASS= variable first appear in the input data set.

FORMATTED displays the rows (columns) from top to bottom (left to right) in increasing order of the formatted values of the first CLASS= variable. For example, suppose you use a numeric CLASS= variable called Day (with values 1, 2, and 3) to create a one-way comparative Pareto chart. Also suppose you use the FORMAT procedure to associate the formatted values 1 = 'Wednesday', 2 = 'Thursday', and 3 = 'Friday' with Day. If you specify ORDER1=FORMATTED, the rows appear in alphabetical order (`'Friday'`, `'Thursday'`, `'Wednesday'`) from top to bottom.

FREQ displays the rows (columns) from top to bottom (left to right) in order of decreasing frequency count. If two or more classes have the same frequency count, the order is determined by the formatted values.

INTERNAL displays the rows (columns) from top to bottom (left to right) in increasing order of the internal (unformatted) values of the first CLASS= variable. If there are two or more distinct internal values that have the same formatted value, the order is determined by the internal value that occurs first in the input data set. In the previous example with variable Day, if you specify ORDER1=INTERNAL, the rows of the comparative chart appear in chronological order (`'Wednesday'`, `'Thursday'`, `'Friday'`) from top to bottom.

By default, ORDER1=INTERNAL.

ORDER2=INTERNAL | FORMATTED | DATA | FREQ

specifies the display order for the values of the second CLASS= variable. The levels of the second CLASS= variable are always constructed using the formatted values of the variable, and the formatted values are always used to label the columns of a two-way comparative Pareto chart.

The PARETO procedure determines the layout of a two-way comparative Pareto chart by first using the ORDER1= option to obtain the order of the rows from top to bottom (recall that ORDER1=INTERNAL by default). Then the ORDER2= option is applied to the observations that correspond to the first row to obtain the order of the columns from left to right. If any columns remain unordered (that is, the categories are unbalanced), the ORDER2= option is applied to the observations in the second row, and so on until all the columns have been ordered.

The values of the ORDER2= option are interpreted as described for the ORDER1= option. By default, ORDER2=INTERNAL.

OTHER='*category*'

specifies a new category that merges all categories that are not selected in the MAXNCAT=, MINPCT=, or MAXCMPCT= options. See the section "Restricting the Number of Pareto Categories" on page 1050 for an illustration.

The *category* should be specified as a formatted value of the process variable. The *category* can be up to 32 characters and must be enclosed in quotation marks. If you specify an OUT= data set, you

should also specify an internal value that corresponds to *category* by specifying the OTHERCVAL= option or the OTHERNVAL= option.

The OTHER= option is not applicable unless you specify the MAXNCAT=, MINPCT=, or MAXCMPCT= option. You can use the COTHER=, LOTHER=, POTHER=, OTHERCVAL=, and OTHERNVAL= options with the OTHER= option.

OTHERCVAL='*value*'
specifies the internal (unformatted) *value* for a character process variable in the OUT= data set that corresponds to the category that is specified in the OTHER= option. The *value* can be up to 64 characters and must be enclosed in quotation marks.

The OTHERCVAL= option is not applicable unless you specify the OTHER= and OUT= options. If you specify the OTHER= option but not the OTHERCVAL= option, the value specified in the OTHER= option is written to the OUT= data set.

OTHERNVAL=*value*
specifies the internal (unformatted) *value* for a numeric process variable in the OUT= data set that corresponds to the category that is specified in the OTHER= option. The OTHERNVAL= option is not applicable unless you specify the OTHER= and OUT= options. If you specify the OTHER= option but not the OTHERNVAL= option, a missing value is written to the OUT= data set.

OUT=*SAS-data-set*
creates an output data set that contains the information that is displayed in the Pareto chart. This data set is useful if you want to create a report to accompany your chart. See Example 16.8 for an illustration.

SCALE=COUNT | FREQUENCY | PERCENT | WEIGHT
specifies the scale for the frequency axis. You can specify the following values:

COUNT or **FREQUENCY** specifies that the scale is counts. See Output 16.1.4 for an illustration. This option is ignored if you specify the WEIGHT= option.

PERCENT specifies that the scale is the percentage of the total frequency or, if you specify the WEIGHT= option, the percentage of the total weight.

WEIGHT scales the vertical axis in the same units as the variable you specify in the WEIGHT= option. This option applies only if you specify the WEIGHT= option.

By default, SCALE=PERCENT. See Output 16.8.1 for an example.

NOTE: Regardless of the value you specify for the SCALE= option, the cumulative percentage axis is scaled in cumulative percentage units.

URL=*variable*
specifies URLs as values of the specified character *variable* (or formatted values of a numeric *variable*). These URLs are associated with bars on the Pareto chart when ODS Graphics output is directed into HTML. The value of *variable* should be the same for each observation that has a particular value of the process variable. The URL= option is not supported for traditional graphics output.

VREFLABPOS=*n*

specifies the vertical positioning of the labels for reference lines that are associated with vertical axes, which are specified in the CATREF= option in an HBAR statement or in the FREQREF= and CUMREF= options in a VBAR statement. If you specify VREFLABPOS=1, the labels are positioned at the left of the chart; if you specify VREFLABPOS=2, the labels are positioned at the right. By default, VREFLABPOS=1.

WEIGHT=*variable-list*

specifies weight variables that are used to construct weighted Pareto charts. Variables in the *variable-list* are paired with the process variables in order of specification. The WEIGHT= variables must be numeric, and their values must be nonnegative (noninteger values are permitted). If a WEIGHT= variable is not provided for a process variable, the weights applied to that process variable are assumed to be 1. See "Weighted Pareto Charts" on page 1094 for computational details.

A WEIGHT= variable is particularly useful for carrying out a Pareto analysis based on cost rather than frequency of occurrence. See Example 16.8 for an illustration.

Options for Traditional Graphics

You can specify the following options only when traditional graphics are produced. The PARETO procedure produces traditional graphics when ODS Graphics is disabled and SAS/GRAPH is licensed.

ANGLE=*value*

specifies an angle in degrees for rotating the labels on the category axis. The *value* is the angle between the baseline of the label and the category axis. See Output 16.1.1 and Output 16.1.2 for an illustration. The *value* must be greater than or equal to –90 and less than 90. The default value is 0.

ANNOKEY

applies the annotation requested in the ANNOTATE= and ANNOTATE2= options only to the key cell in a comparative Pareto chart. By default, annotation is applied to all of the cells.

ANNOTATE=*SAS-data-set*

ANNO=*SAS-data-set*

specifies an input data set that contains annotation variables as described in *SAS/GRAPH: Reference*. You can use the *SAS-data-set* to customize the Pareto charts that are produced by a single HBAR or VBAR statement. (A data set that is specified in the ANNOTATE= option in the PROC PARETO statement customizes charts that are produced by *all* HBAR and VBAR charts.) The *SAS-data-set* is associated with the frequency axis. If the annotation is based on data coordinates, you must use the same units as the frequency axis.

ANNOTATE2=*SAS-data-set*

ANNO2=*SAS-data-set*

specifies an input data set that contains annotation variables as described in *SAS/GRAPH: Reference*. You can use the *SAS-data-set* to customize the Pareto charts that are produced by a single HBAR or VBAR statement. (A data set that is specified in the ANNOTATE2= option in the PROC PARETO statement customizes charts that are produced by *all* HBAR and VBAR charts.) The *SAS-data-set* is associated with the cumulative percentage axis. If the annotation is based on data coordinates, you must use the same units as the cumulative percentage axis.

BARLABPOS=*keyword*

specifies the position for labels that are requested in the BARLABEL= option.

You can specify the following *keywords* in an HBAR statement:

HBAR displays the label right-justified on the bar. If the label is longer than the bar, it is left-justified at the base of the bar.

HFIT right-justifies the label on the bar. If the label is longer than the bar, the label is displayed to the right of the bar.

HLJUST left-justifies the label at the base of the bar.

HRIGHT displays the label to the right of the bar. If there is insufficient space for the label to the right of the bar, the label is right-justified at the right edge of the frame.

HRJUST right-justifies the label at the right edge of the frame.

The default for an HBAR statement is BARLABPOS=HRIGHT.

You can specify the following *keywords* in a VBAR statement:

HCENTER centers the label horizontally above the bar. If the centered label would extend outside the frame, the label is left-justified or right-justified at the edge of the frame.

HLJUST left-justifies the label horizontally above the bar. The label is truncated if necessary.

VBAR displays the label vertically on the bar. If the label is longer than the bar, it extends above the bar.

VFIT displays the label vertically on or above the bar, depending on the available space. If the label is longer than the bar, it is displayed just below the top edge of the frame.

The default for a VBAR statement is to center the labels horizontally above the bars, with a reduction in text height if necessary. Reduction is not applied when the BARLABPOS= option is specified.

BARWIDTH=*value*

specifies the width of the bars in screen percentage units. By default, the bars are made as wide as possible.

CAXIS=*color*

CAXES=*color*

CA=*color*

specifies the color for the axis lines and tick marks. The default color is specified by the ContrastColor attribute of the GraphAxisLines style element in the current ODS style. If the NOGSTYLE option is in effect, *color* is also used for bar outlines and grid lines, unless overridden by the CBARLINE=, CGRID=, or GRID2= option.

CAXIS2=*color*

specifies the color for the tick mark labels and axis label that are associated with the cumulative percentage axis. By default, the color specified in the CTEXT= option (or its default) is used.

CBARLINE=*color*

specifies the color for bar outlines. The default color is specified by the ContrastColor attribute of the GraphOutlines style element in the current ODS style.

CBARS=*color*

CBARS=(*variable-list***)**

specifies how the bars of the Pareto chart are colored. You can specify the following values:

color uses a single color for all the bars. You can use this option in conjunction with the CHIGH(*n*) and CLOW(*n*) options.

variable-list uses a distinct color for each bar (or combination of bars). The colors are specified as values of variables in the *variable-list*. Each variable must be a character variable. You can use the special value `'EMPTY'` to indicate that a bar is not to be colored. Note that *variable-list* must be enclosed in parentheses. You cannot specify a *variable-list* conjunction with the CHIGH(*n*) or CLOW(*n*) option.

If you specify more than one process variable, you can specify more than one CBARS= variable. The number of CBARS= variables should be less than or equal to the number of process variables. The two lists of variables are paired in order of specification.

If no CBARS= color or variable is specified for a process variable, the bars for its chart are displayed in the default color, which is determined by the Color attribute of the GraphData1 style element in the current ODS style.

If you specify one or more CBARS= variables, you can also use the BARLEGEND= option to add a legend to the chart that explains the significance of each color. Furthermore, you can use the PBARS= option to specify patterns in conjunction with the CBARS= option.

CCATREF=*color*

specifies the color for reference lines that are requested in the CATREF= option. The default color is specified by the ContrastColor attribute of the GraphReference style element in the current ODS style.

CCONNECT=*color*

specifies the color for the line segments that connect the points on the cumulative percentage curve. The default color is determined by the ContrastColor attribute of the GraphDataDefault style element in the current ODS style. You can specify the color for the points on the cumulative percentage curve in SYMBOL statement COLOR= option.

CCUMREF=*color*

specifies the color for reference lines that are requested in the CUMREF= option. The default color is specified by the ContrastColor attribute of the GraphReference style element in the current ODS style.

CFRAME=*color*

specifies the color for filling the area that is enclosed by the axes and the frame. The default color is specified by the Color attribute of the GraphWalls style element in the current ODS style. You cannot use the CFRAME= option in conjunction with the NOFRAME option or the CTILES= option.

CFRAMESIDE=*color*

specifies the color for filling the frame area for the row labels, which are displayed along the left side of a comparative Pareto chart. If a label is associated with the classification variable, *color* is also used to fill the frame area for this label. By default, the frame is transparent.

CFRAMETOP=*color*

specifies the color for filling the frame area for the column labels, which are displayed across the top of a comparative Pareto chart. If a label is associated with the classification variable, *color* is also used to fill the frame area for this label. By default, the frame is transparent.

CFREQREF=*color*

specifies the color for reference lines that are requested in the FREQREF= option. The default color is specified by the ContrastColor attribute of the GraphReference style element in the current ODS style.

CGRID=*color*

specifies the color for frequency axis grid lines. If you specify this option, you do not need to specify the GRID option. The default color is specified by the ContrastColor attribute of the GraphGridLines style element in the current ODS style.

CGRID2=*color*

specifies the color for cumulative percentage axis grid lines. If you specify this option, you do not need to specify the GRID2 option. The default color is specified by the ContrastColor attribute of the GraphGridLines style element in the current ODS style.

CLIPREF

draws reference lines that are requested in the CATREF=, CUMREF=, and FREQREF= options behind the bars on the Pareto chart. When the GSTYLE option is in effect, reference lines are drawn in front of the bars by default.

COTHER=*color*

specifies the color for the bar that is defined by the OTHER= option. By default the CFRAME= color is used. The COTHER= option is not applicable unless a BARS= or CBARS= variable is specified.

CTEXT=*color*

CT=*color*

specifies the color for text, such as tick mark labels, axis labels, and legends. The default color is specified by the Color attribute of a style element in the current ODS style. Axis labels use the GraphLabelText style element, and all other text uses the GraphValueText style element.

CTEXTSIDE=*color*

specifies the color for row labels, which are displayed along the left side of a comparative Pareto chart. If you do not specify a *color*, the color specified in the CTEXT= option is used. If neither option is specified, the color is determined by the Color attribute of the GraphValueText style element in the current ODS style.

CTEXTTOP=*color*

specifies the color for column labels, which are displayed across the top of a comparative Pareto chart. If you do not specify a *color*, the color specified in the CTEXT= option is used. If neither option is specified, the color is determined by the Color attribute of the GraphValueText style element in the current ODS style.

CTILES=(*variable***)**

specifies a character variable whose values are the fill colors for the tiles in a comparative Pareto chart. This option generalizes the CFRAME= option, which provides a single color for all of the tiles. The *variable* must be enclosed in parentheses. The values of the *variable* must be identical for all

observations that have the same level of the CLASS= variables. You can use the same color to fill more than one tile. You can use the special value `'EMPTY'` to indicate that a tile is not to be filled.

You cannot use the CTILES= option in conjunction with the NOFRAME or CFRAME= options. You can use the TILELEGEND= option in conjunction with the CTILES= option to add an explanatory legend for the CTILES= colors at the bottom of the chart. See Output 16.5.1 for an illustration.

DESCRIPTION='*string*'

DES='*string*'

specifies a description, up to 256 characters long, for the GRSEG catalog entry for a traditional graphics chart.

FONT=*font*

specifies a font for text that is used in labels and legends. The default font is determined by the FontFamily, FontStyle, and FontWeight attributes of a style element in the current ODS style; axis labels use the GraphLabelText style element and all other text uses the GraphValueText style element.

FRONTREF

draws reference lines that are requested in the CATREF=, FREQREF=, and CUMREF= options in front of the bars on the Pareto chart. When the NOGSTYLE option is in effect, reference lines are drawn behind the bars by default and can be obscured by them.

HEIGHT=*value*

specifies the height in screen percentage units of text for labels and legends. This option takes precedence over the GOPTONS HTEXT= option. The default value is specified by the FontSize attribute of the a style element in the current ODS style; axis labels use the GraphLabelText style element and all other text uses the GraphValueText style element.

HTML=*variable*

specifies a variable whose values create links that are associated with Pareto bars when traditional graphics output is directed into HTML. You can specify a character variable or a formatted numeric variable. The value of the HTML= variable should be the same for each observation that has a particular value of the process variable.

INFONT=*font*

specifies a font for bar labels, cumulative percentage curve labels, and sample size legends. This option takes precedence over the FONT= option and the FTEXT= option in the GOPTIONS statement. The default font is determined by the FontFamily, FontStyle, and FontWeight attributes of the GraphValueText style element in the current ODS style.

INHEIGHT=*value*

specifies the height in screen percentage units of bar labels, cumulative percentage curve labels, and sample size legends. This option takes precedence over the HEIGHT= option and the HTEXT= option in a GOPTIONS statement. The default value is specified by the FontSize attribute of the GraphValueText style element in the current ODS style.

INTERBAR=*value*

specifies the distance in screen percentage units between bars on the chart. By default, the bars are contiguous.

LCATREF=*line-type*
specifies the line type for reference lines that are requested in the CATREF= option. The default line type is specified by the LineStyle attribute of the GraphReference style element in the current ODS style.

LCUMREF=*line-type*
specifies the line type for reference lines that are requested in the CUMREF= option. The default line type is specified by the LineStyle attribute of the GraphReference style element in the current ODS style.

LFREQREF=*line-type*
specifies the line type for lines that are requested in the FREQREF= option. The default line type is specified by the LineStyle attribute of the GraphReference style element in the current ODS style.

LGRID=*line-type*
specifies the line type for frequency axis grid lines. If you specify this option, you do not need to specify the GRID option. The default line type is specified by the LineStyle attribute of the GraphGridLines style element in the current ODS style.

LGRID2=*line-type*
specifies the line type for cumulative percentage axis grid lines. If you specify this option, you do not need to specify the GRID2 option. The default line type is specified by the LineStyle attribute of the GraphGridLines style element in the current ODS style.

NAME='*string*'
specifies the name of the GRSEG catalog entry for a traditional graphics chart, and the name of the graphics output file if one is created. The name can be up to 256 characters long, but the GRSEG name is truncated to eight characters. The default name is "PARETO".

NOFRAME
suppresses the frame that is drawn around the chart by default. You cannot specify the NOFRAME option in conjunction with the CFRAME= or TILES= options.

PBARS=*pattern*
PBARS=(*variable-list***)**
specifies pattern fills for the bars. You can specify the following values:

pattern uses a single pattern for all the bars. You can use this approach in conjunction with the PHIGH(*n*)= and PLOW(*n*)= options.

variable-list uses a distinct pattern for *each* bar (or combination of bars). You provide the patterns as values of variables in the *variable-list*. For example, you might use the solid pattern ('S') to indicate severe problems and the empty pattern ('E') for all other problems. Each variable must be a character variable of length eight, and the *variable-list* must be enclosed in parentheses. You cannot specify a *variable-list* in conjunction with the PHIGH(*n*)= and PLOW(*n*)= options.

If you specify more than one process variable in the chart statement, you can provide more than one variable in the *variable-list*. The number of variables in the *variable-list* should be less than or equal to the number of process variables. The two lists of variables are paired in order of specification. If a variable is not

provided in the *variable-list* for a process variable, the bars for that chart are not filled.

If you specify a *variable-list*, you can also use the BARLEGEND= option to add a legend to the chart that explains the significance of each pattern.

You can use the CBARS= option to specify colors in conjunction with the PBARS= option.

PHIGH(*n*)=*pattern*
specifies the pattern for the bars that have the *n* highest values. You cannot specify this option in conjunction with a PBARS=*variable-list*, but you can specify this option together with the PLOW(*n*)= and PBARS=*pattern* options.

PLOW(*n*)=*pattern*
specifies the pattern for the bars that have the *n* lowest values. You cannot specify this option in conjunction with a PBARS=*variable-list*, but you can use this option together with the PHIGH(*n*)= and PBARS=*pattern* options.

POTHER=*pattern*
specifies the pattern for the bar that is defined by the OTHER= option. This option applies only if you specify a PBARS=*variable-list*.

TILELEGEND=(*variable*)
specifies a *variable* that is used to add a legend for CTILES= colors. The variable can have a formatted length less than or equal to 32. If a format is associated with the variable, then the formatted value is displayed. You must specify the TILELEGEND= option in conjunction with the CTILES= option. If you specify the CTILES= option but do not specify the TILELEGEND= option, a color legend is not displayed.

The values of the CTILES= and TILELEGEND= variables should be consistent for all observations that have the same level of the CLASS= variables. The value of the TILELEGEND= variable is used to identify the corresponding color value of the CTILES= variable in the legend. See Output 16.5.1 for an illustration.

TILELEGLABEL='*label*'
specifies a label for the legend that is created when you specify a TILELEGEND= variable. The *label* can be up to 16 characters and must be enclosed in quotation marks. The default is "Tiles:". See Output 16.5.1 for an illustration.

TURNVLABEL
TURNVLABELS
turns and strings out vertically the characters in the labels for the frequency and cumulative percentage axes. The TURNVLABELS option is valid only in a VBAR statement.

WAXIS=*n*
specifies the line thickness (in pixels) for the axes and frame. This thickness is also used for bar outlines and grid lines, unless overridden by the WBARLINE=, WGRID=, or WGRID2= option. The default line thickness is specified by the LineThickness attribute of the GraphAxisLines style element in the current ODS style.

WBARLINE=*n*
specifies the width for bar outlines. The default outline thickness is specified by the LineThickness attribute of the GraphOutlines style element in the current ODS style.

WGRID=*n*
specifies the width of the frequency axis grid lines. If you specify this option, the GRID option is not required. The default line thickness is specified by the LineThickness attribute of the GraphGridLines style element in the current ODS style.

WGRID2=*n*
specifies the width of the cumulative percentage axis grid lines. If you specify this option, the GRID2 option is not required. The default line thickness is specified by the LineThickness attribute of the GraphGridLines style element in the current ODS style.

Options for Legacy Line Printer Charts

NOTE: The HBAR statement does not produce legacy line printer charts, so the following *options* apply only to the VBAR statement.

CONNECTCHAR='*character*'
CCHAR='*character*'
specifies the plot character for line segments that connect points on the cumulative percentage curve. The default character is a plus sign (+).

HREFCHAR='*character*'
specifies the plot character used to form the lines that are requested in the CATREF= option. The default character is a vertical bar (|).

SYMBOLCHAR='*character*'
specifies the plot character for points on the cumulative percentage curve. The default character is an asterisk (*).

VREFCHAR='*character*'
specifies the character to be used to form the lines that are requested in the FREQREF= and CUMREF= options. The default character is a dash (-).

Details: PARETO Procedure

Terminology

Basic Pareto Charts

A basic Pareto chart (see Figure 16.1) analyzes the unique values of a *process variable*. These values are called *Pareto categories* or *levels*, and they usually represent problems that are encountered during some phase of a manufacturing or service activity.

A basic vertical Pareto chart (as produced by the PARETO procedure's VBAR statement) has one horizontal axis and two vertical axes:

- The *category* axis is displayed horizontally at the bottom of the chart and lists the Pareto categories.

- The *frequency axis* (or *primary vertical axis*) is displayed on the left. The relative frequency of each Pareto category is represented by a vertical bar whose height is measured on the frequency axis. You can use the SCALE= option to scale this axis in percentage, count, or weight units.
- The *cumulative percentage axis* (or *secondary vertical axis*) is displayed on the right. This axis is scaled in cumulative percentage units and is used to read the *cumulative percentage curve*. The height of each point on the curve represents the percentage of the total frequency that is accounted for by the Pareto categories to the left of the point.

A horizontal Pareto chart (as produced by the HBAR statement), is essentially a vertical Pareto chart rotated 90 degrees clockwise. The category axis is displayed vertically on the left. Categories appear in order of decreasing relative frequency from top to bottom. The frequency axis appears at the top of the chart and the cumulative percentage axis appears at the bottom. The relative frequencies of the Pareto categories are represented by horizontal bars. A point on the cumulative percentage curve represents the percentage of the total frequency that is accounted for by the Pareto categories above that point.

NOTE: For the sake of brevity, in this chapter the term *height* refers to the size of a bar as measured along the frequency axis, whether the Pareto chart is oriented vertically or horizontally.

Restricted Pareto Charts

A *restricted Pareto chart* (see Figure 16.6) displays only the n most frequently occurring categories in a data set that contains N categories, where $N > n$. The remaining $N - n$ categories are dropped or are merged into a single "other" category that is created when you specify the OTHER= option. The MAXCMPCT=, MAXNCAT=, and MINPCT= options provide alternative methods for specifying n. See the entries for these options in the section "Dictionary of HBAR and VBAR Statement Options" on page 1071.

Weighted Pareto Charts

A *weighted Pareto chart* (see Example 16.8) displays bars whose heights represent the weighted frequencies of the categories. Typical weights are the cost of repair or the loss incurred by the customer.

The weight W_i for the ith Pareto category is computed as

$$W_i = \sum_{u \in \mathcal{C}_i} w(u) f(u)$$

where $\mathcal{C}_i$ is the set of observations that make up the ith category, $w(u)$ is the value of the weight variable in the uth observation, and $f(u)$ is the value of the frequency variable in the uth observation (taking $f(u) \equiv 1$ if a FREQ= variable is not specified). If SCALE=WEIGHT is specified, the height of the bar for the ith category is W_i. If SCALE=PERCENT is specified, the height of this bar is

$$\frac{100 W_i}{\sum_{j=1}^{N} W_j}$$

where N is the total number of categories.

Comparative Pareto Charts

A *comparative Pareto chart* combines two or more Pareto charts for the same process variable. The component charts are displayed with uniform axes to facilitate comparison. The observations that are represented by a component chart are called a *cell*. The framed areas for the component charts are called *tiles*.

In a *one-way comparative Pareto chart*, each component chart corresponds to a different level of a single classification variable, which is specified in the CLASS= option. The component charts are arranged in a stack or a row, as illustrated in Output 16.1.3, Output 16.1.4, Output 16.2.2, and Output 16.2.3. In a *two-way comparative Pareto chart*, each component chart corresponds to a different combination of levels of two classification variables, which are specified in the CLASS= option. The component charts are arranged in a matrix, as illustrated in Output 16.2.4.

Every comparative Pareto chart has a *key cell*, in which the bars are in decreasing order and whose order is imposed on all the other cells to achieve a uniform category axis. By default, the key cell is the cell in the upper left corner, but you can use the CLASSKEY= option to designate any other cell as the key cell. If you designate another cell as the key cell, the rows and columns of the comparative chart are rearranged so that the key cell appears in the upper left. However, if you require the rows and columns in a particular order, you can specify the NOKEYMOVE option in conjunction with the CLASSKEY= option to suppress the rearrangement.

You can use the NROWS= and NCOLS= options to specify the numbers of rows and columns in a comparative Pareto chart. By default, NROWS=2 and NCOLS=1 for a one-way comparison and NROWS=2 and NCOLS=2 for a two-way comparison. There is no upper limit to the number of rows or columns that you can specify, but in practice the limit is determined by the area of the graphical display. If the numbers of classification variable levels exceed the NROWS= and NCOLS= values, the chart is created on multiple panels or pages.

If the same set of Pareto categories does not occur in each cell of a comparative Pareto chart, the categories are said to be *unbalanced*. In this case, PROC PARETO uses the following convention to construct the uniform category axis. First, the categories that occur in the key cell are arranged on the category axis from left to right (top to bottom for a horizontal chart) and sorted in decreasing order of frequency, with tied levels arranged in order of their formatted values. The categories not in the key cell are assigned frequencies of 0 in the key cell, and they are arranged at the right (bottom) of the category axis, where they are ordered by their formatted values. This arrangement is simply a convention of the PARETO procedure and should not be interpreted to mean that one category is more important than another.

Whether the categories in the input data set are balanced or not, the categories in the OUT= data set are always balanced. PROC PARETO balances this data set by assigning values of 0 to the _COUNT_ and _PCT_ variables as necessary.

Unbalanced categories present a special problem when the MAXNCAT= option is used to restrict the number of categories that are displayed on the chart. For example, suppose that you specify MAXNCAT=12 and there are 15 categories in all, 10 of which occur in the key cell. Because there is no unambiguous method for selecting two of the remaining five categories to complete the restricted list, the PARETO procedure reduces the restricted list to the categories that occur in the key cell and displays only those 10 categories. A warning message is issued in the SAS log.

Labels for Chart Features

Table 16.8 summarizes the methods for labeling the features of Pareto charts.

Table 16.8 Labeling Features of Pareto Charts

Feature	Method for Specifying Label
Titles	TITLE*n* statements, ODSTITLE= option, ODSTITLE2= option
Footnotes	FOOTNOTE*n* statements, ODSFOOTNOTE= option, ODSFOOTNOTE2= option
Category axis	Process variable label
Frequency axis	FREQAXISLABEL= option
Cumulative percentage axis	CUMAXISLABEL= option
Bars	BARLABEL= option
Points on cumulative percentage curve	CMPCTLABEL= option
Rows and columns	CLASS= variable labels
Cells	NLEGEND option or NLEGEND= variable
Category legend	CATLEGLABEL= option
High/low bar legend	HLLEGLABEL= option
Bar color legend	BARLEGLABEL= option
Tile legend	TILELEGLABEL= option
Annotation	ANNOTATE= and ANNOTATE2= data sets

Scaling the Cumulative Percentage Curve

Pareto charts shown in textbooks usually scale the cumulative percentage curve so that it is anchored at the top right corner of the leftmost bar. The upper end of the frequency axis is then extended to accommodate the curve. For an illustration, see Figure 16.1. By default, the PARETO procedure uses the top right corner as the anchor position on a vertical chart and the bottom right corner of the topmost bar as the anchor position on a horizontal chart. You can override the default by specifying the ANCHOR= option.

This method of scaling is not feasible if the number of categories is very large and if the Pareto distribution is uniform. In this case, the bars are excessively compressed relative to the curve. Conversely, this method excessively compresses the curve relative to the bars when you use a count scale for the frequency axis in a comparative Pareto chart and the tallest bar does not occur in the key cell. In either situation, PROC PARETO overrides the textbook scaling method and balances the scales of the bars and the curve.

You can use the AXISFACTOR= option to specify the extent to which the frequency axis should be extended. Alternatively, you can extend the frequency axis by using the FREQAXIS= option to specify the tick mark values for the axis.

Another scaling anomaly is illustrated by the comparative Pareto chart in Output 16.1.4. There, the cumulative percentage curve in the bottom chart is not anchored because a uniform count scale is combined with different sample sizes in the two cells.

Positioning Insets

This section provides details about three different methods of positioning insets using the POSITION= option. You can use the POSITION= option to specify the following:

- compass points
- keywords for margin positions
- coordinates in data units or percentage axis units

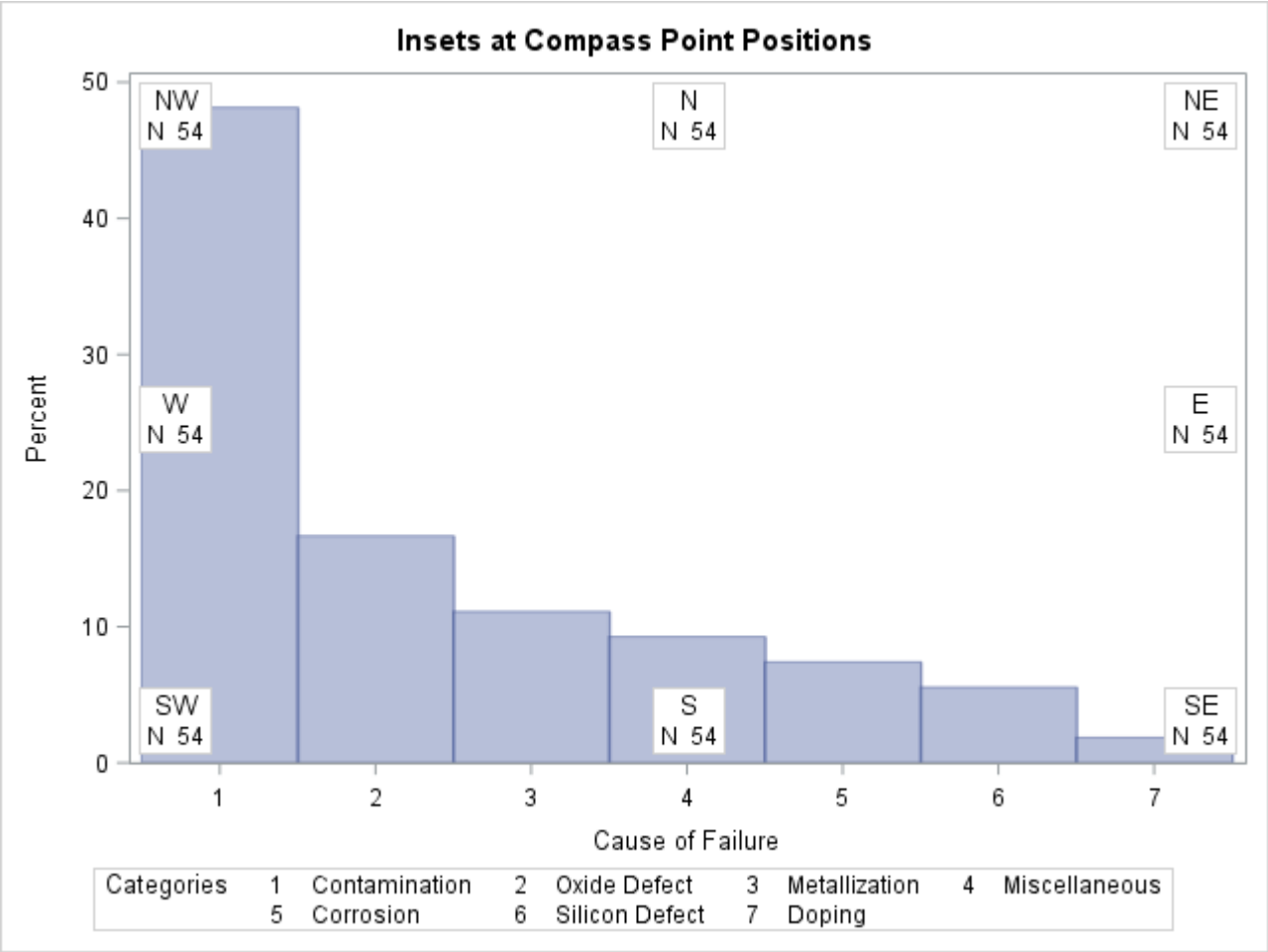
Using Compass Points to Position Insets

NOTE: See *Positioning Insets in Pareto Charts* in the SAS/QC Sample Library.

You can specify the eight compass points N, NE, E, SE, S, SW, W, and NW as keywords for the POSITION= option. The following statements create the display in Figure 16.9, which demonstrates all eight compass positions. The default is NW.

```
proc pareto data=Failure3;
   vbar Cause / freq     = Counts
                odstitle = "Insets at Compass Point Positions"
                nocurve
                ;
   inset n / cfill header='NW' pos=nw;
   inset n / cfill header='N ' pos=n ;
   inset n / cfill header='NE' pos=ne;
   inset n / cfill header='E ' pos=e ;
   inset n / cfill header='SE' pos=se;
   inset n / cfill header='S ' pos=s ;
   inset n / cfill header='SW' pos=sw;
   inset n / cfill header='W ' pos=w ;
run;
```

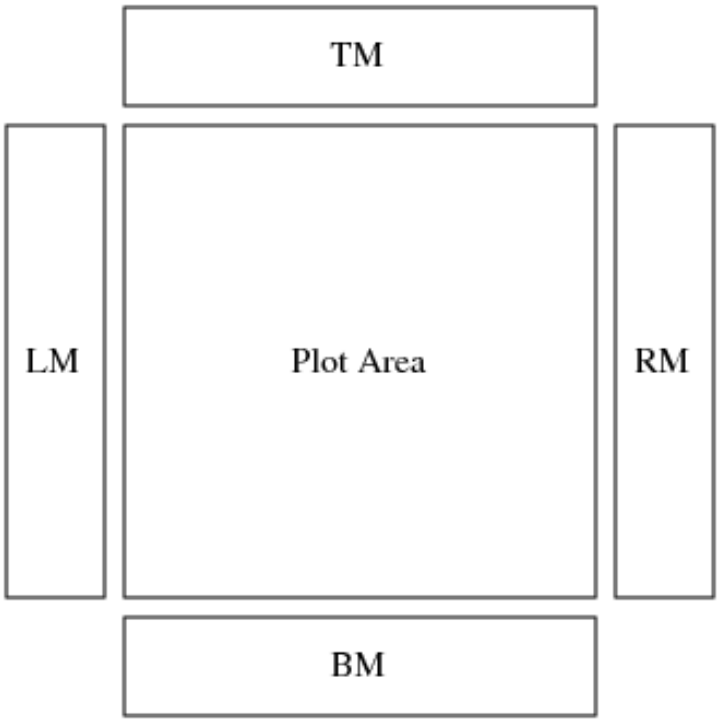
Figure 16.9 Using Compass Points to Position Insets



Positioning Insets in the Margins

You can also use the margin keywords LM, RM, TM, or BM in the INSET statement to position an inset in one of the four margins that surround the plot area, as illustrated in Figure 16.10.

Figure 16.10 Positioning Insets in the Margins



For an example of an inset placed in the right margin, see Output 16.11.1. You might want to place an inset in a margin if it contains a large number of entries (for example the contents of a data set that is specified in the DATA= keyword). If you attempt to display a lengthy inset in the interior of the plot, the inset is likely to collide with the data display.

Insets that are associated with a comparative Pareto chart cannot be positioned in the margins.

Using Coordinates to Position Insets

When you produce traditional graphics, you can also specify the position of the inset with coordinates by specifying POSITION=(x, y). The coordinates can be specified in axis percentage units (the default) or in axis data units.

Data Unit Coordinates

If you specify the DATA option immediately following the coordinates, the inset is positioned using axis data units. Data units along the category axis are based on category numbers. Categories are numbered from left to right (VBAR chart) or top to bottom (HBAR chart), starting with 1.

NOTE: See *Positioning Insets in Pareto Charts* in the SAS/QC Sample Library.

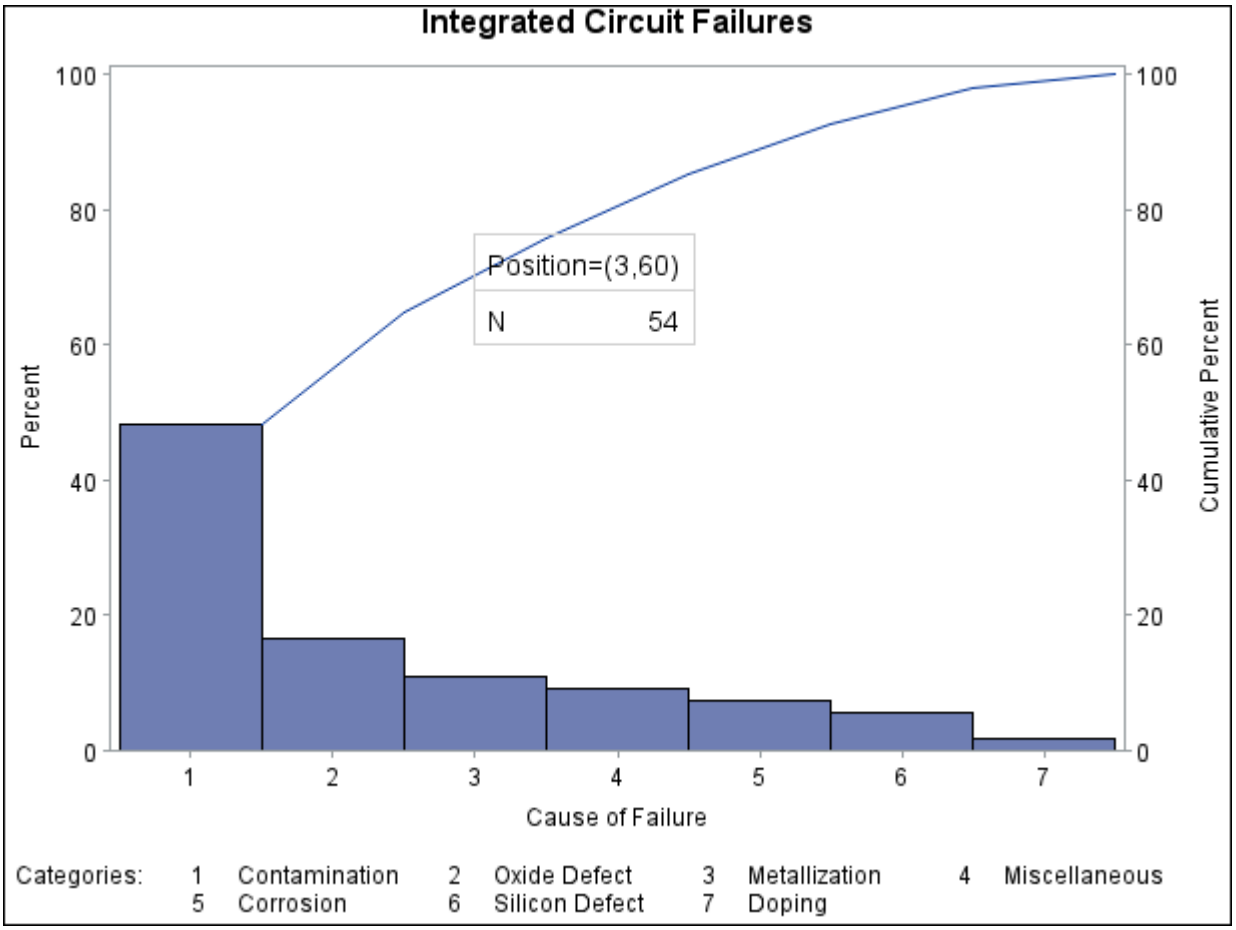
For example, the following statements produce the Pareto chart that is displayed in Figure 16.11:

```
ods graphics off;
title 'Integrated Circuit Failures';
proc pareto data=Failure3;
   vbar Cause / freq = Counts;
   inset n / header   = 'Position=(3,60)'
             position = (3,60) data
             height   = 3;
run;
```

The HEIGHT= option in the INSET statement specifies the text height that is used to display the statistics in the inset.

The bottom left corner of the inset is lined up with the tick mark for the third category on the horizontal axis and at 60 on the vertical axis. By default, the specified coordinates determine the position of the bottom left corner of the inset. You can change this reference point by specifying the REFPOINT= option, as shown in the next section.

Figure 16.11 Inset Positioned Using Data Unit Coordinates



Axis Percentage Unit Coordinates

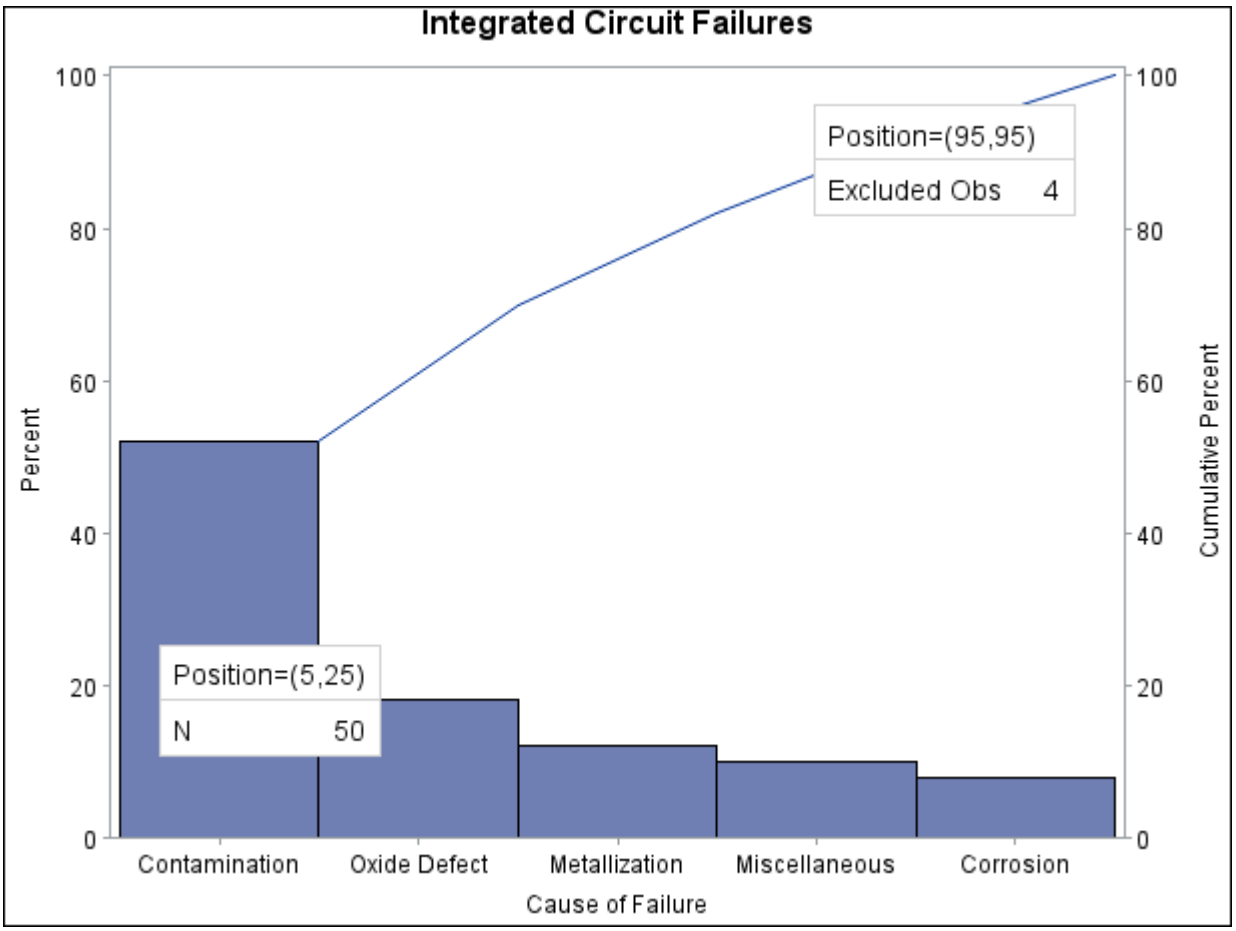
NOTE: See *Positioning Insets in Pareto Charts* in the SAS/QC Sample Library.

If you do not use the DATA option, the inset is positioned using axis percentage units. The coordinates of the bottom left corner of the display are (0, 0), and the coordinates of the upper right corner are (100, 100). For example, the following statements create a Pareto chart that has two insets, both positioned using coordinates in axis percentage units.

```
proc pareto data=Failure3;
   vbar Cause / freq     = Counts
                maxncat  = 5;
   inset n / position = (5,25)
             header   = 'Position=(5,25)'
             height   = 3
             cfill    = blank
             refpoint = tl;
   inset nexcl / position = (95,95)
                 header   = 'Position=(95,95)'
                 height   = 3
                 cfill    = blank
                 refpoint = tr;
run;
```

The chart is shown in Figure 16.12. Notice that the REFPOINT= option is used to determine which corner of the inset is to be placed at the coordinates that are specified in the POSITION= option. The first inset has REFPOINT=TL, so the top left corner of the inset is positioned 5% of the way across the horizontal axis and 25% of the way up the vertical axis. The second inset has REFPOINT=TR, so the top right corner of the inset is positioned 95% of the way across the horizontal axis and 95% of the way up the vertical axis. Coordinates in axis percentage units must be between 0 and 100.

Figure 16.12 Inset Positioned Using Axis Percentage Unit Coordinates



Creating Output Data Sets

The OUT= data set saves the information that is displayed on a Pareto chart. If you specify CLASS= variables, the OUT= data set contains one block of observations for each combination of levels of the CLASS= variables, and each block contains an observation for each Pareto category. The observations are sorted in the order in which the categories are displayed on the chart. The following variables from a DATA= data set are saved in an OUT= data set:

- process variables
- CLASS= variables
- BY variables
- WEIGHT= variables
- the CTILES= variable
- the TILELEGEND= variable

- the NLEGEND= variable
- BARS= or CBARS= variables
- PBARS= variables
- BARLEGEND= variables

In addition, the OUT= data set contains the following variables that are created during the analysis:

- _COUNT_, which saves the frequency count for each Pareto category
- _WCOUNT_, which saves the weighted count for each category. This variable is created only when you specify the WEIGHT= option.
- _PCT_, which saves the percentage of the total count for each category. If you specify the WEIGHT= option, the variable _PCT_ saves the percentage of the total weighted count.
- _CMPCT_, which saves the cumulative percentage for each category

See Output 16.8.2 for an example of an OUT= data set.

If you specify the MAXNCAT=, MAXCMPCT=, or MINPCT= option, the OUT= data set saves only the categories that are displayed on the chart. If you create an OTHER= category that merges the remaining categories, an additional observation is saved with the new category. Because the OTHER= value is defined as a formatted value of the process variable, you should also specify a corresponding internal value, as follows:

- If the process variable is a character variable, specify the internal value in the OTHERCVAL= option. If you do not specify this value, the OTHER= value is saved as the internal value.
- If the process variable is a numeric variable, specify the internal value in the OTHERNVAL= option. If you do not specify this value, an internal missing value is saved.

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 appearance of a graph that ODS Graphics produces is determined by the style that is associated with the ODS destination where the graph is produced. HBAR and VBAR statement options that control the appearance of traditional graphics (listed in the section "Options for Traditional Graphics" on page 1086) are ignored for ODS Graphics output.

When ODS Graphics is in effect, the PARETO procedure assigns a name to graphs it creates. You can use this name to refer to the graph when using ODS. The name is listed in Table 16.9.

Table 16.9 ODS Graphics Produced by the PARETO Procedure

ODS Graph Name	Plot Description
ParetoChart	Pareto chart

See Chapter 4, "SAS/QC Graphics," for more information about ODS Graphics and other methods for producing charts.

Constructing Effective Pareto Charts

The following are recommendations for improving the visual clarity of Pareto charts:

- Decide carefully how the bars should be scaled. The default percentage scale is not always the best choice. For example, a count scale might be more appropriate in a comparative Pareto chart where the total count per cell varies widely from cell to cell and where you want to compare Pareto distributions on an absolute scale rather than a relative scale. You can request a count scale by specifying SCALE=COUNT. In other situations, it might be more appropriate to use a weighted percentage scale or a weighted count scale (specify a WEIGHT= variable and either SCALE=PERCENT or SCALE=WEIGHT).
- Use a weight variable if the counts are dependent on a factor (such as exposure or opportunity) that varies from one category to another. For example, suppose you are creating a Pareto chart for the number of medical claims that are categorized by the job titles of company employees who submit them. The counts can be weighted to adjust for the fact that there are more individuals in some jobs than in others and for the fact that some jobs might be associated with greater health risks than others.
- Use the NOCURVE option to eliminate the cumulative percentage curve in situations where the curve reveals little information about the data. In general, the bars should be more prominent than the curve.
- Maximize the space used for the bars by eliminating unnecessary labels and visual clutter. This is particularly important for comparative Pareto charts. The NOCATLABEL, NOFREQLABEL, and NOCUMLABEL options are useful for this purpose. You can also use the NOFREQTICK and NOCUMTICK options to eliminate tick marks and tick mark labels on the frequency and cumulative percentage axes.
- Make legends more informative by specifying legend labels.
- Avoid filling bars with multiple types of cross-hatched patterns; solid color fills are less distracting. Use color sparingly to emphasize important features (such as the "vital few" categories), and choose bar colors that provide good visual discrimination.
- If you are working with a large data set that involves many categories, limit the number of categories that are displayed to achieve visual clarity.
- If your application involves classification effects, construct more than one Pareto chart for the data by using various combinations of classification variables. (This approach is illustrated in Example 16.2).
- Provide reference lines on comparative Pareto charts to aid visual comparison.

See to Chapter 2 of Cleveland (1985) for a general discussion of the principles of statistical graphics.

Missing Values

By default, observations that have missing values of a process variable are not processed. If you specify the MISSING option, then missing values are treated as a Pareto category.

Likewise, observations that have missing values of the CLASS= variables are not processed by default. Missing values of the first CLASS= variable are treated as a level if the MISSING1 option is specified, and missing values of the second CLASS= variable are treated as a level if the MISSING2 option is specified.

Role of Variable Formats

The categories of a Pareto chart are always determined using formatted values of the process variable, and the format is used to label the categories.

On the chart, the categories are displayed in decreasing order of frequency. If multiple categories have the same count, the tied categories are displayed in order of their formatted values.

When you create a comparative Pareto chart, the formatted levels of the CLASS= variables are used to group the data into cells. There is a cell for each level of the CLASS= variable in a one-way comparative chart, and there is a cell for each combination of levels of the CLASS= variables in a two-way comparative chart.

You can specify the order of the rows and columns that correspond to the classification levels by specifying the ORDER1= and ORDER2= options. The default value of these options is INTERNAL, which means that the order is determined by the internal values of the CLASS= variables. It is possible for a particular formatted value to correspond to more than one internal value. To resolve this ambiguity, the internal value that determines the position of the row or column is the value that occurs first in the input data set.

Other values that you can specify for the ORDER1= and ORDER2= options are FORMATTED, FREQ, and DATA.

Large Data Sets

Although there is no limit to the number of observations that can be read from an input data set, the maximum number of Pareto categories that can be read is 32,767. This limit is a practical issue only if you are creating a restricted Pareto chart from a large data set, because the number of categories that can be displayed is limited by the resolution of your graphical display. The number of categories that can be read is limited by the amount of memory available, because the levels are stored in memory. If you run out of memory, you should first reduce the data by using the FREQ procedure.

Examples: PARETO Procedure

Example 16.1: Creating Before-and-After Pareto Charts

NOTE: See *Before & After Pareto Charts Using a BY Variable* in the SAS/QC Sample Library.

During the manufacture of a metal-oxide semiconductor (MOS) capacitor, causes of failures were recorded before and after a tube in the diffusion furnace was cleaned. This information was saved in a SAS data set named Failure3:

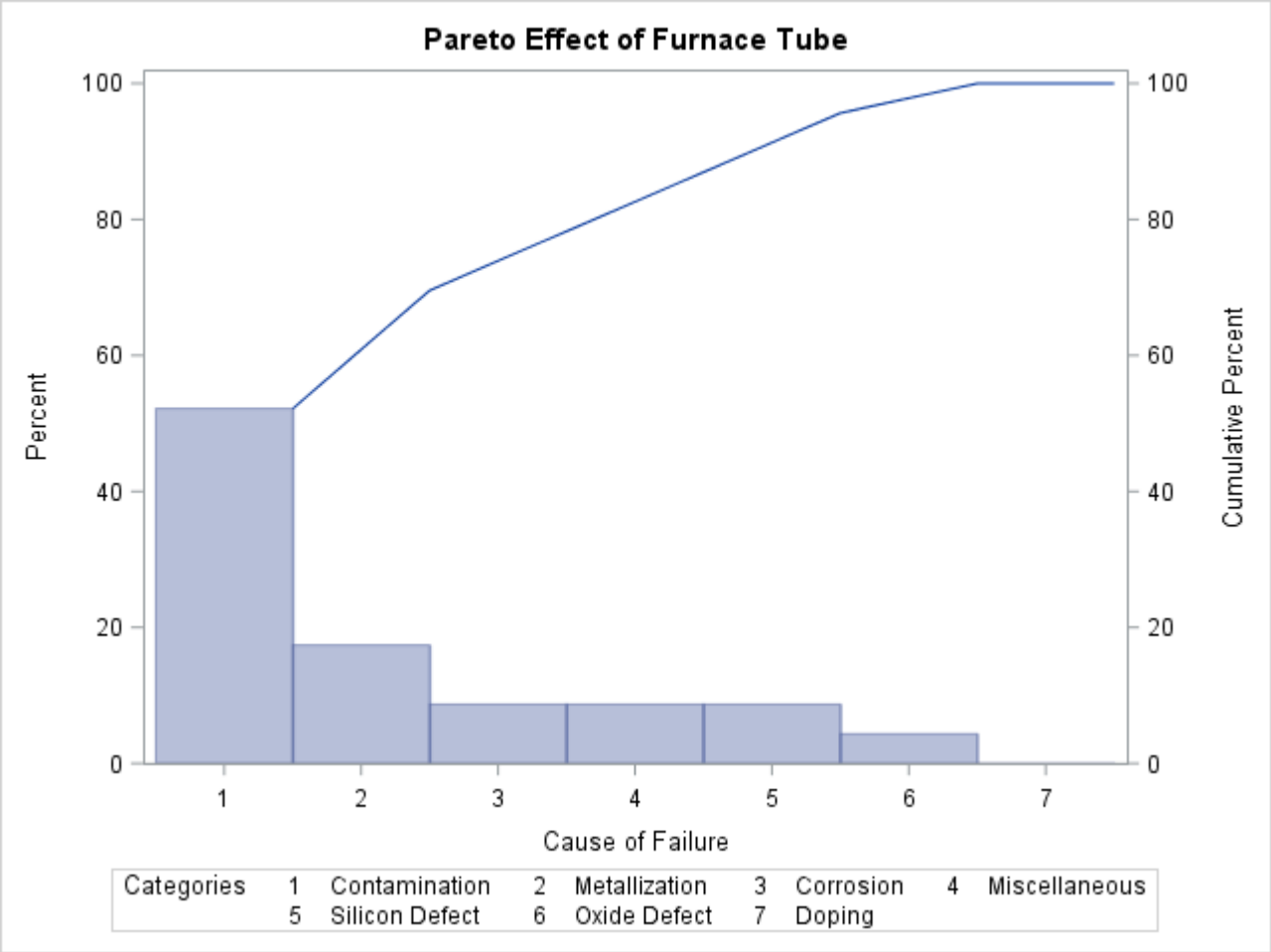
```
data Failure3;
   length Cause $ 16 Stage $ 16;
   label  Cause = 'Cause of Failure';
   input  Stage & $ Cause & $ Counts;
datalines;
Before Cleaning   Contamination    14
Before Cleaning   Corrosion         2
Before Cleaning   Doping            1
Before Cleaning   Metallization     2
Before Cleaning   Miscellaneous     3
Before Cleaning   Oxide Defect      8
Before Cleaning   Silicon Defect    1
After Cleaning    Doping            0
After Cleaning    Corrosion         2
After Cleaning    Metallization     4
After Cleaning    Miscellaneous     2
After Cleaning    Oxide Defect      1
After Cleaning    Contamination    12
After Cleaning    Silicon Defect    2
;
```

To compare distribution of failures before and after cleaning, you can use the BY statement to create two separate Pareto charts, one for the observations in which Stage is equal to `'Before Cleaning'` and one for the observations in which Stage is equal to `'After Cleaning'`:

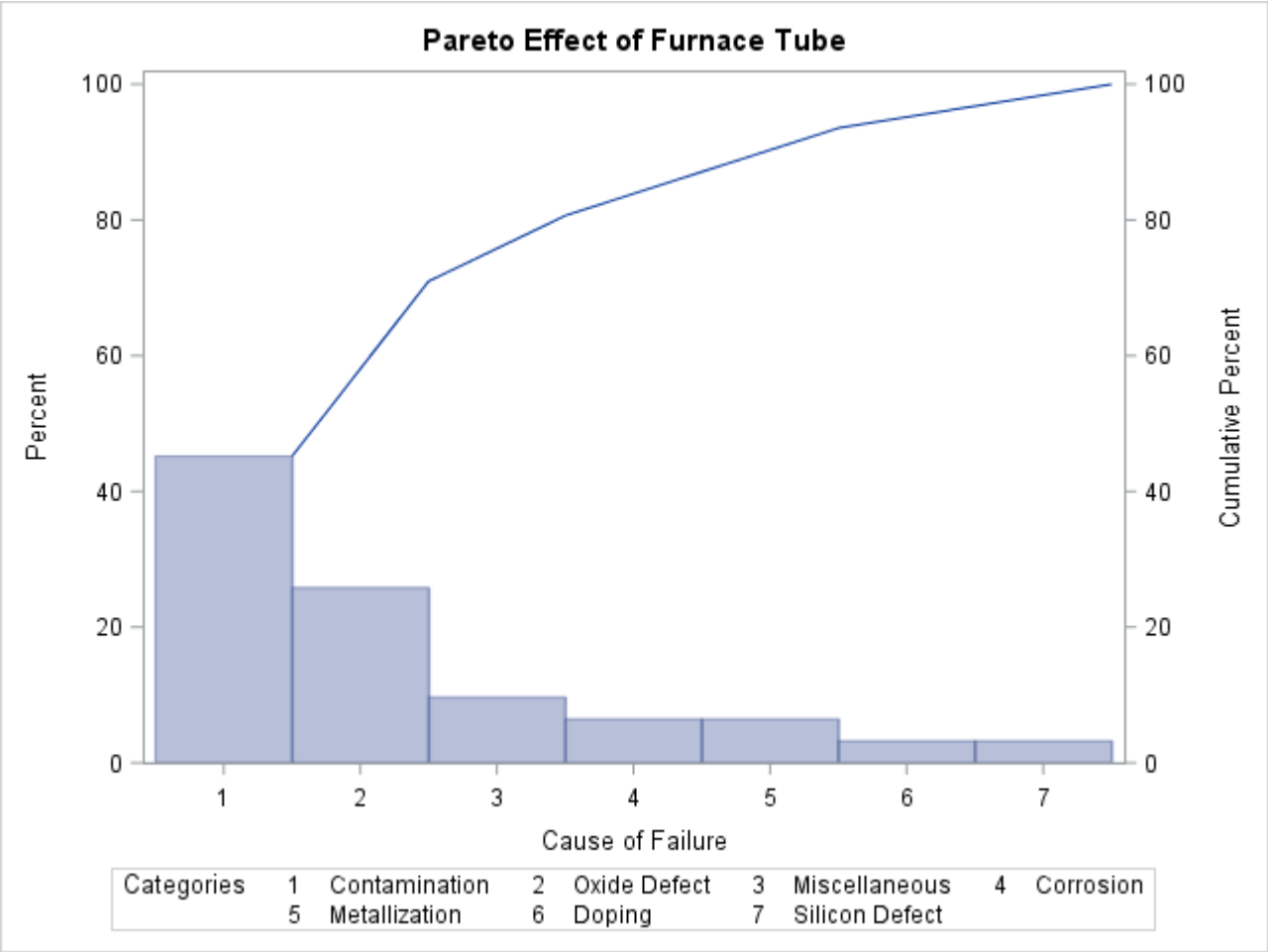
```
proc sort data=Failure3;
   by Stage;
run;

title 'Pareto Effect of Furnace Tube';
proc pareto data=Failure3;
   vbar Cause / freq     = Counts
                odstitle = title;
   by Stage;
run;
```

The SORT procedure sorts the observations in order of the values of Stage. It is not necessary to sort by the values of Cause because this is done by the PARETO procedure. The two charts, displayed in Output 16.1.1 and Output 16.1.2, reveal a reduction in oxide defects after the tube was cleaned. This is a relative reduction, because the frequency axes are scaled in percentage units. Note that the `'After Cleaning'` chart is produced first, based on alphabetical sorting of BY groups.

Output 16.1.1 "After" Analysis Using Stage as a BY Variable



Output 16.1.2 "Before" Analysis Using Stage as a BY Variable

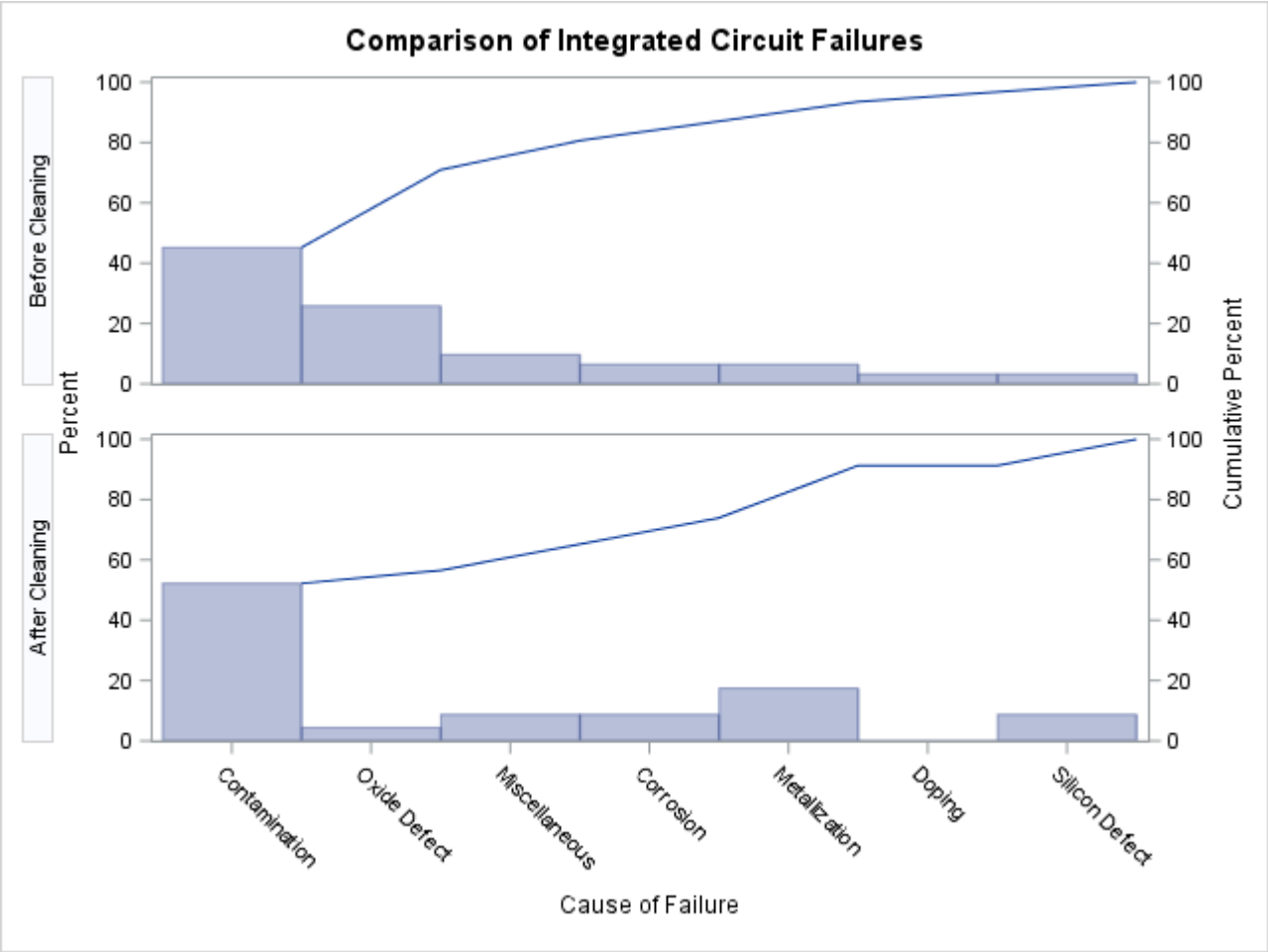


In general, it is difficult to compare Pareto charts that are created by using BY processing because their axes are not necessarily uniform. A better approach is to construct a comparative Pareto chart, as illustrated by the following statements:

```
title 'Comparison of Integrated Circuit Failures';
proc pareto data=Failure3;
   vbar Cause / class      = Stage
                freq       = Counts
                scale      = percent
                intertile  = 5.0
                classkey   = 'Before Cleaning'
                odstitle   = title;
run;
```

The CLASS= option designates Stage as a classification variable, and this directs PROC PARETO to create the one-way comparative Pareto chart shown in Output 16.1.3, which displays a component chart for each level of Stage. The INTERTILE= option separates the cells with an offset of 5 screen percentage units.

Output 16.1.3 Before-and-After Analysis That Uses a Comparative Pareto Chart



In a comparative Pareto chart, there is always one special cell, called the *key cell*, in which the bars are displayed in decreasing order, and whose order determines the uniform category axis that is used for all the cells. The key cell is positioned at the top of the chart. Here, the key cell is the set of observations for which Stage equals `'Before Cleaning'`, as specified by the CLASSKEY= option. By default, the levels are sorted in the order determined by the ORDER1= option, and the key cell is the level that occurs first in this order.

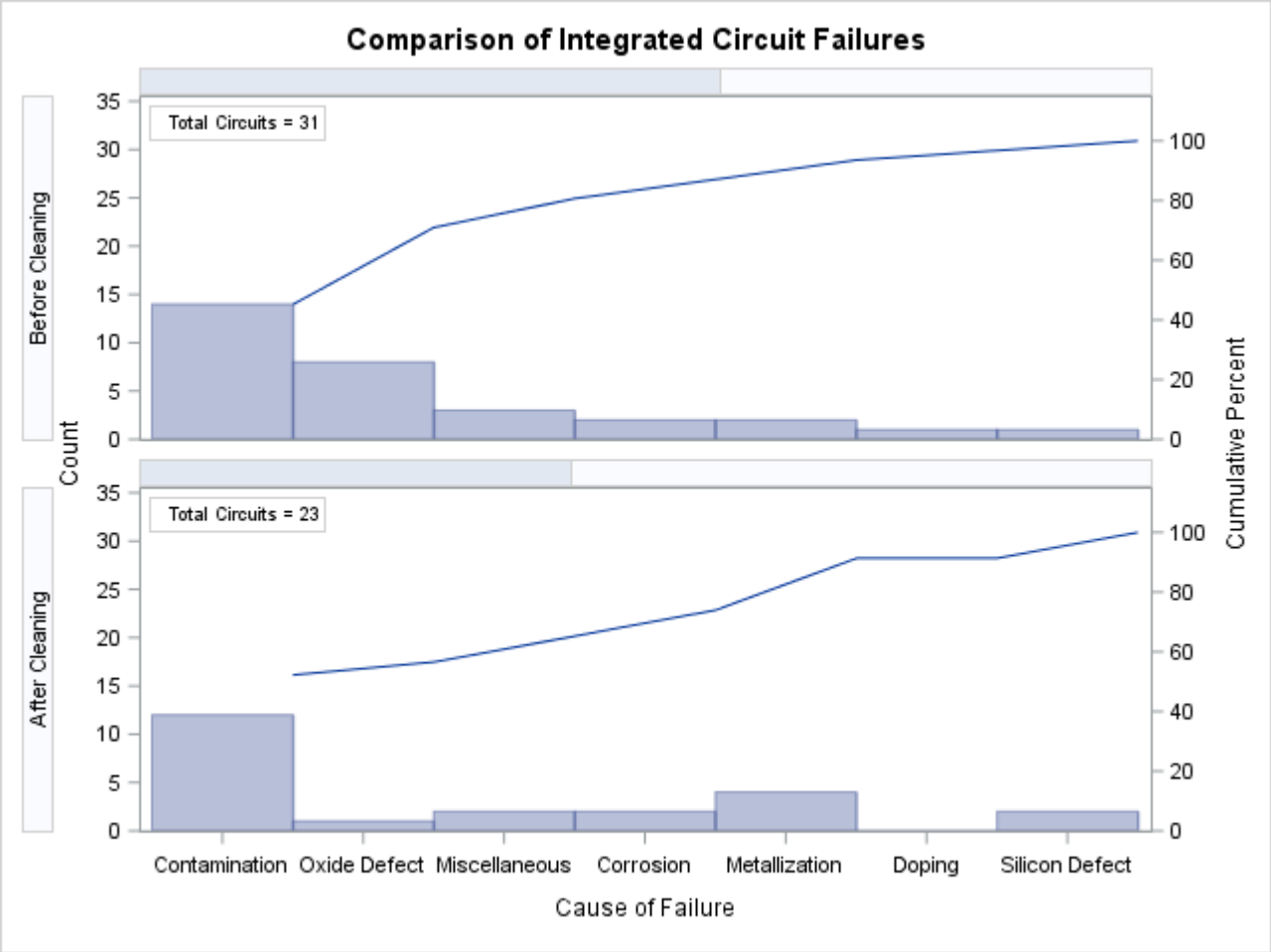
In many applications, it can be more revealing to base comparisons on counts rather than percentages. The following statements construct a chart that has a frequency scale:

```
title 'Comparison of Integrated Circuit Failures';
proc pareto data=Failure3;
   vbar Cause / class     = Stage
                freq      = Counts
                scale     = count
                nlegend   = 'Total Circuits'
                classkey  = 'Before Cleaning'
                odstitle  = title
                cframenleg
                cprop;
run;
```

Specifying SCALE=COUNT scales the frequency axis in count units. The NLEGEND= option adds a sample size legend, and the CFRAMENLEG option frames the legend. The CPROP option adds bars that indicate the proportion of total frequency represented by each cell.

The chart is shown in Output 16.1.4.

Output 16.1.4 Before-and-After Analysis Using Comparative Pareto Chart



Note that the lower cumulative percentage curve in Output 16.1.4 is not anchored to the first bar. This is a consequence of the uniform frequency scale and of the fact that the number of observations in each cell is not the same.

Example 16.2: Creating Two-Way Comparative Pareto Charts

NOTE: See *Basic and Comparative Pareto Charts* in the SAS/QC Sample Library.

During the manufacture of a MOS capacitor, different cleaning processes were used by two manufacturing systems operating in parallel. Process A used a standard cleaning solution, and Process B used a different cleaning mixture that contained less particulate matter. The failure causes that were observed with each process for five consecutive days were recorded and saved in a SAS data set called Failure4:

```
data Failure4;
   length Process $ 9 Cause $ 16;
   label  Cause   = 'Cause of Failure';
   input  Process & $ Day & $ Cause & $ Counts;
   datalines;
Process A   March 1   Contamination    15
Process A   March 1   Corrosion         2
Process A   March 1   Doping            1
Process A   March 1   Metallization     2
Process A   March 1   Miscellaneous     3
Process A   March 1   Oxide Defect      8
Process A   March 1   Silicon Defect    1
Process A   March 2   Contamination    16
Process A   March 2   Corrosion         3
Process A   March 2   Doping            1
Process A   March 2   Metallization     3
Process A   March 2   Miscellaneous     1
Process A   March 2   Oxide Defect      9
Process A   March 2   Silicon Defect    2
Process A   March 3   Contamination    20
Process A   March 3   Corrosion         1
Process A   March 3   Doping            1
Process A   March 3   Metallization     0
Process A   March 3   Miscellaneous     3
Process A   March 3   Oxide Defect      7
Process A   March 3   Silicon Defect    2
Process A   March 4   Contamination    12
Process A   March 4   Corrosion         1
Process A   March 4   Doping            1
Process A   March 4   Metallization     0
Process A   March 4   Miscellaneous     0
Process A   March 4   Oxide Defect     10
Process A   March 4   Silicon Defect    1
Process A   March 5   Contamination    23
Process A   March 5   Corrosion         1
Process A   March 5   Doping            1
Process A   March 5   Metallization     0
Process A   March 5   Miscellaneous     1
Process A   March 5   Oxide Defect      8
Process A   March 5   Silicon Defect    2
Process B   March 1   Contamination     8
Process B   March 1   Corrosion         2
Process B   March 1   Doping            1
Process B   March 1   Metallization     4
Process B   March 1   Miscellaneous     2
Process B   March 1   Oxide Defect     10
Process B   March 1   Silicon Defect    3
Process B   March 2   Contamination     9
Process B   March 2   Corrosion         0
Process B   March 2   Doping            1
Process B   March 2   Metallization     2
Process B   March 2   Miscellaneous     4
Process B   March 2   Oxide Defect      9
```

```
Process B   March 2   Silicon Defect    2
Process B   March 3   Contamination     4
Process B   March 3   Corrosion         1
Process B   March 3   Doping            1
Process B   March 3   Metallization     0
Process B   March 3   Miscellaneous     0
Process B   March 3   Oxide Defect     10
Process B   March 3   Silicon Defect    1
Process B   March 4   Contamination     2
Process B   March 4   Corrosion         2
Process B   March 4   Doping            1
Process B   March 4   Metallization     0
Process B   March 4   Miscellaneous     3
Process B   March 4   Oxide Defect      7
Process B   March 4   Silicon Defect    1
Process B   March 5   Contamination     1
Process B   March 5   Corrosion         3
Process B   March 5   Doping            1
Process B   March 5   Metallization     0
Process B   March 5   Miscellaneous     1
Process B   March 5   Oxide Defect      8
Process B   March 5   Silicon Defect    2
;
```

In addition to the process variable Cause, this data set has two classification variables: Process and Day. The variable Counts is a frequency variable.

This example creates a series of displays that progressively use more of the classification information.

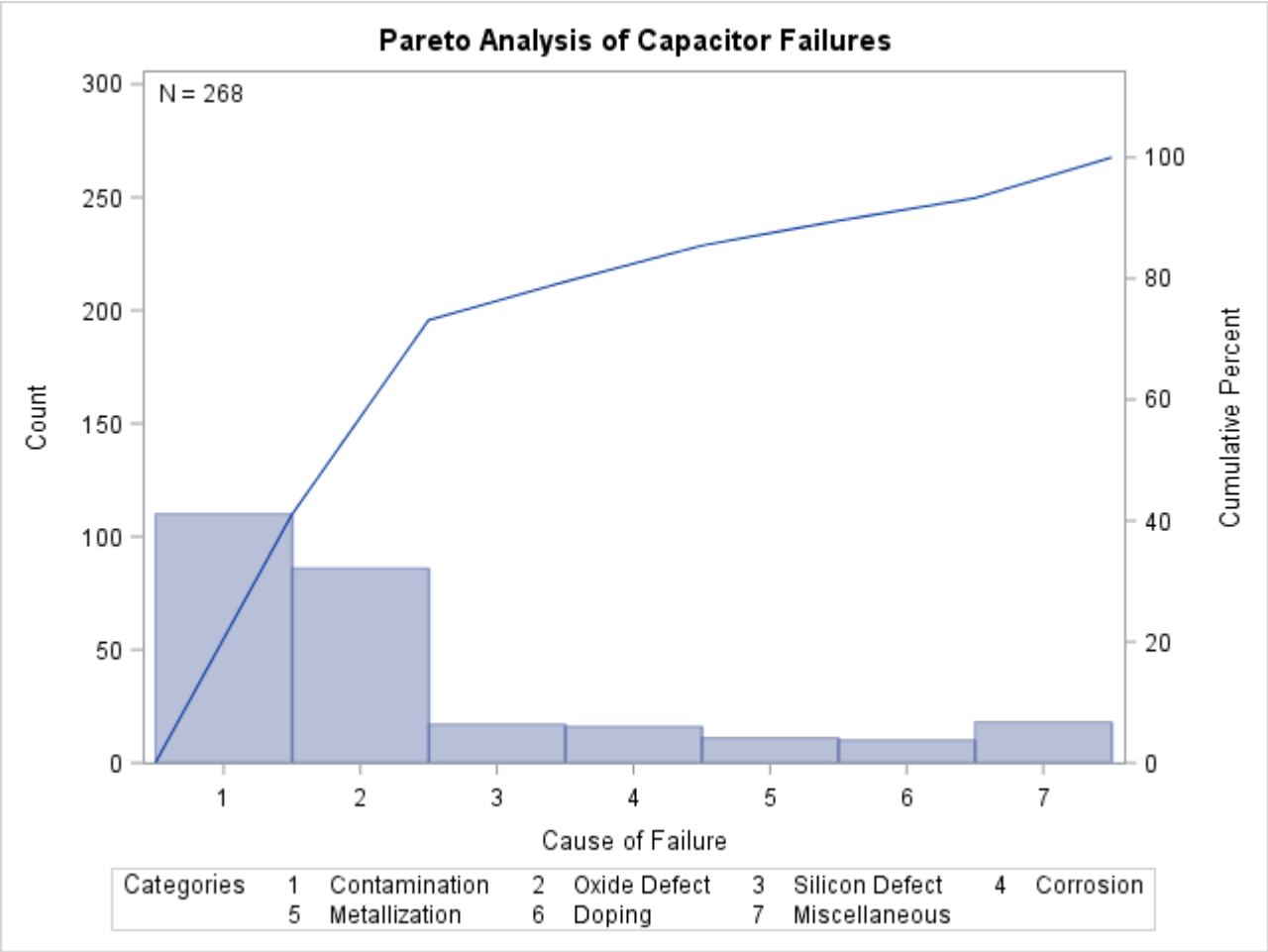
Basic Pareto Chart

The following statements create the first display, which analyzes the process variable without taking into account the classification variables:

```
title 'Pareto Analysis of Capacitor Failures';
proc pareto data=Failure4;
   vbar Cause / freq     = Counts
                last     = 'Miscellaneous'
                scale    = count
                anchor   = bl
                odstitle = title
                nlegend;
run;
```

The chart, shown in Output 16.2.1, indicates that contamination is the most frequently occurring problem.

Output 16.2.1 Pareto Analysis without Classification Variables



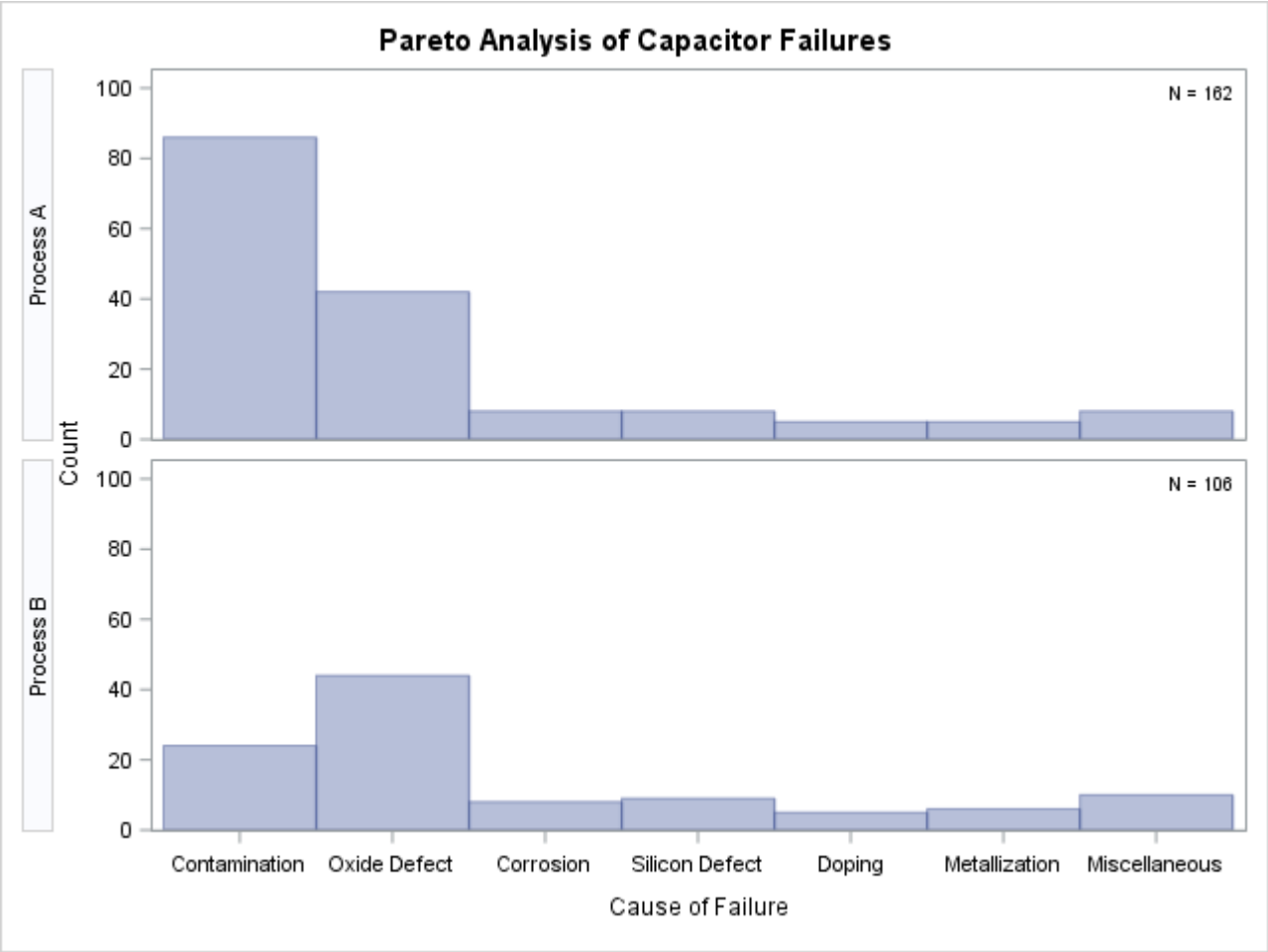
The ANCHOR=BL option anchors the cumulative percentage curve at the bottom left (BL) of the first bar. The NLEGEND option adds a sample size legend.

One-Way Comparative Pareto Chart for Process

The following statements specify Process as a classification variable to create a comparative Pareto chart, which is displayed in Output 16.2.2:

```
proc pareto data=Failure4;
   vbar Cause / class       = Process
                freq        = Counts
                last        = 'Miscellaneous'
                scale       = count
                odstitle    = title
                nocurve
                nlegend;
run;
```

Output 16.2.2 One-Way Comparative Pareto Analysis with CLASS=Process



Each cell corresponds to a level of the CLASS= variable (Process). By default, the cells are arranged from top to bottom in alphabetical order of the formatted values of Process, and the key cell is the top cell. The main difference in the two cells is a decrease in contamination when Process B is used.

The NOCURVE option suppresses the cumulative percentage curve, along with the cumulative percentage axis.

One-Way Comparative Pareto Chart for Day

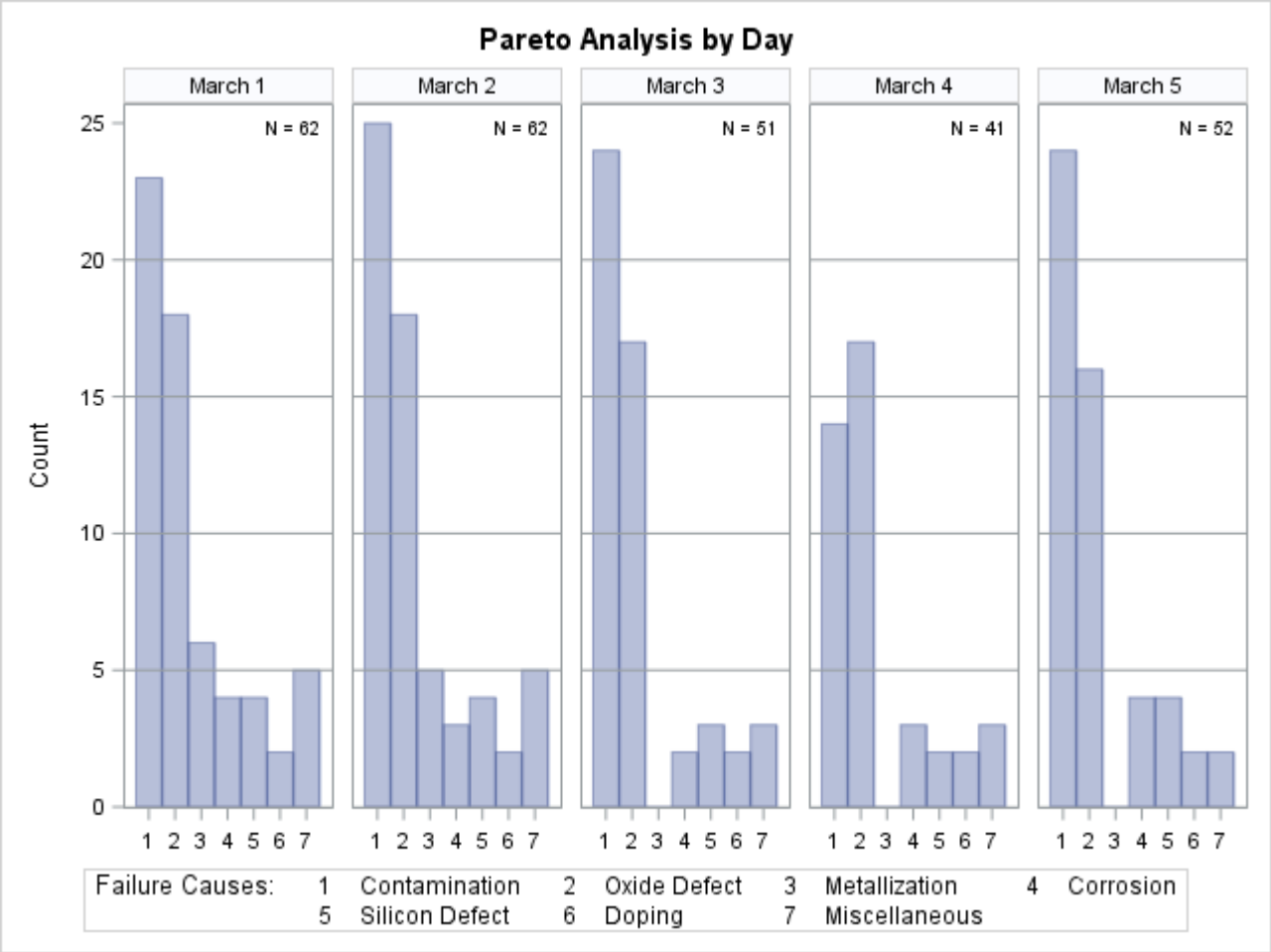
The following statements specify Day as a classification variable:

```
title 'Pareto Analysis by Day';
proc pareto data=Failure4;
   vbar Cause / class       = Day
                freq        = Counts
                last        = 'Miscellaneous'
                scale       = count
                catleglabel = 'Failure Causes:'
                odstitle    = title
                nrows       = 1
```

```
                    ncols       = 5
                    freqref     = 5 10 15 20
                    nocatlabel
                    nocurve
                    nlegend;
  run;
```

The NROWS= and NCOLS= options display the cells in a side-by-side arrangement. The FREQREF= option adds reference lines perpendicular to the frequency axis. The NOCATLABEL option suppresses the category axis labels, and the CATLEGLABEL= option incorporates that information into the category legend label. The chart is displayed in Output 16.2.3.

Output 16.2.3 One-Way Comparative Pareto Analysis with CLASS=Day



By default, the key cell is the leftmost cell. There were no failures due to metallization starting on March 3 (in fact, process controls to reduce this problem were introduced on this day).

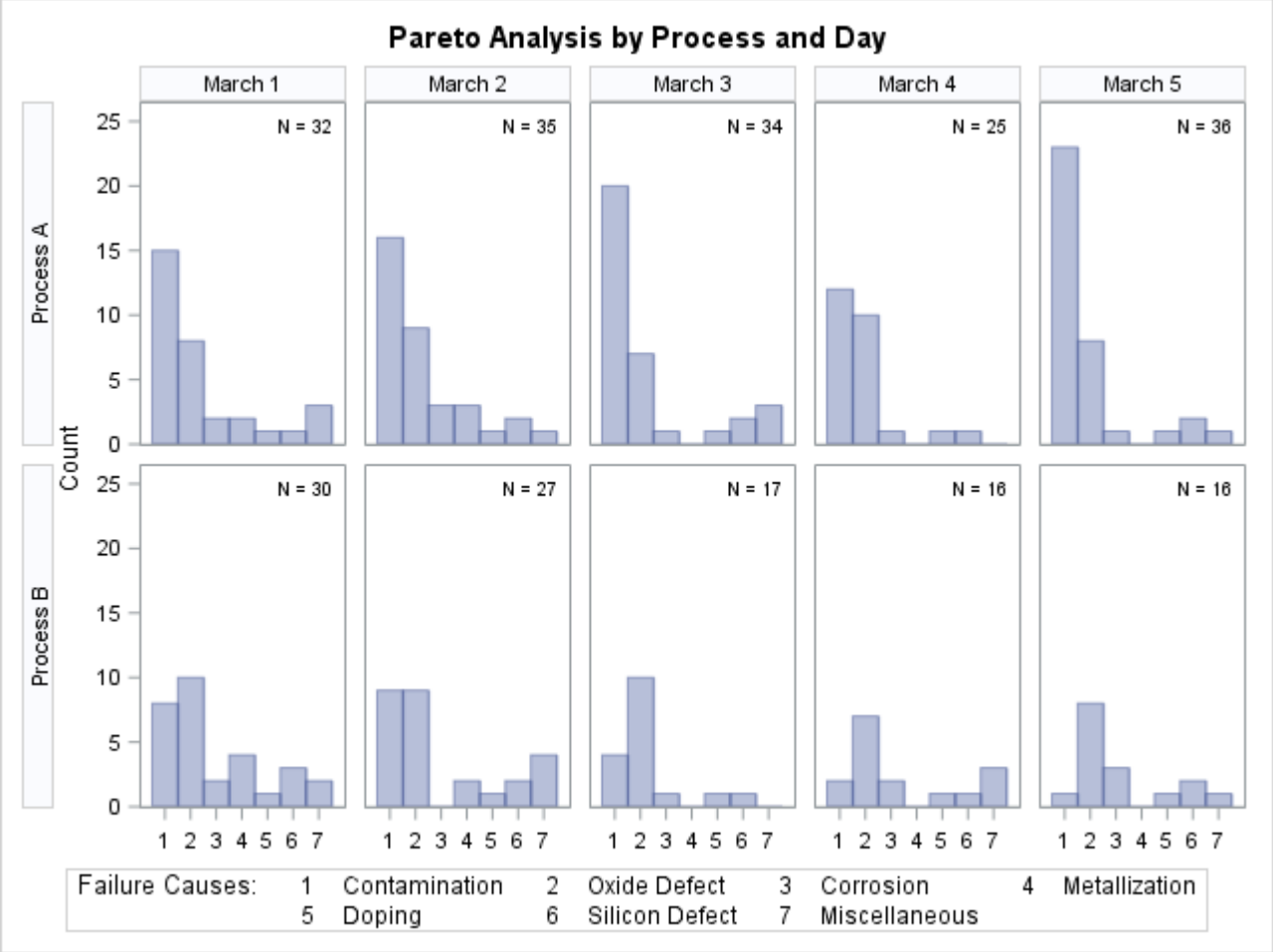
Two-Way Comparative Pareto Chart for Process and Day

The following statements specify both Process and Day as CLASS= variables to create a two-way comparative Pareto chart:

```
title 'Pareto Analysis by Process and Day';
proc pareto data=Failure4;
   vbar Cause / class       = ( Process Day )
                freq        = Counts
                nrows       = 2
                ncols       = 5
                last        = 'Miscellaneous'
                scale       = count
                catleglabel = 'Failure Causes:'
                odstitle    = title
                nocatlabel
                nocurve
                nlegend;
run;
```

The chart is displayed in Output 16.2.4.

Output 16.2.4 Two-Way Comparative Pareto Analysis for Process and Day



The cells are arranged in a matrix whose rows correspond to levels of the first CLASS= variable (Process) and whose columns correspond to levels of the second CLASS= variable (Day). The dimensions of the matrix are specified in the NROWS= and NCOLS= options. The key cell is in the upper left corner.

The chart reveals continuous improvement when Process B is used.

Example 16.3: Highlighting the "Vital Few"

NOTE: See *Highlighting the "Vital Few"* in the SAS/QC Sample Library.

This example is a continuation of Example 16.2.

In some applications you might want to use colors and patterns to highlight the bars that correspond to the most frequently occurring categories, which are referred to as the "vital few."

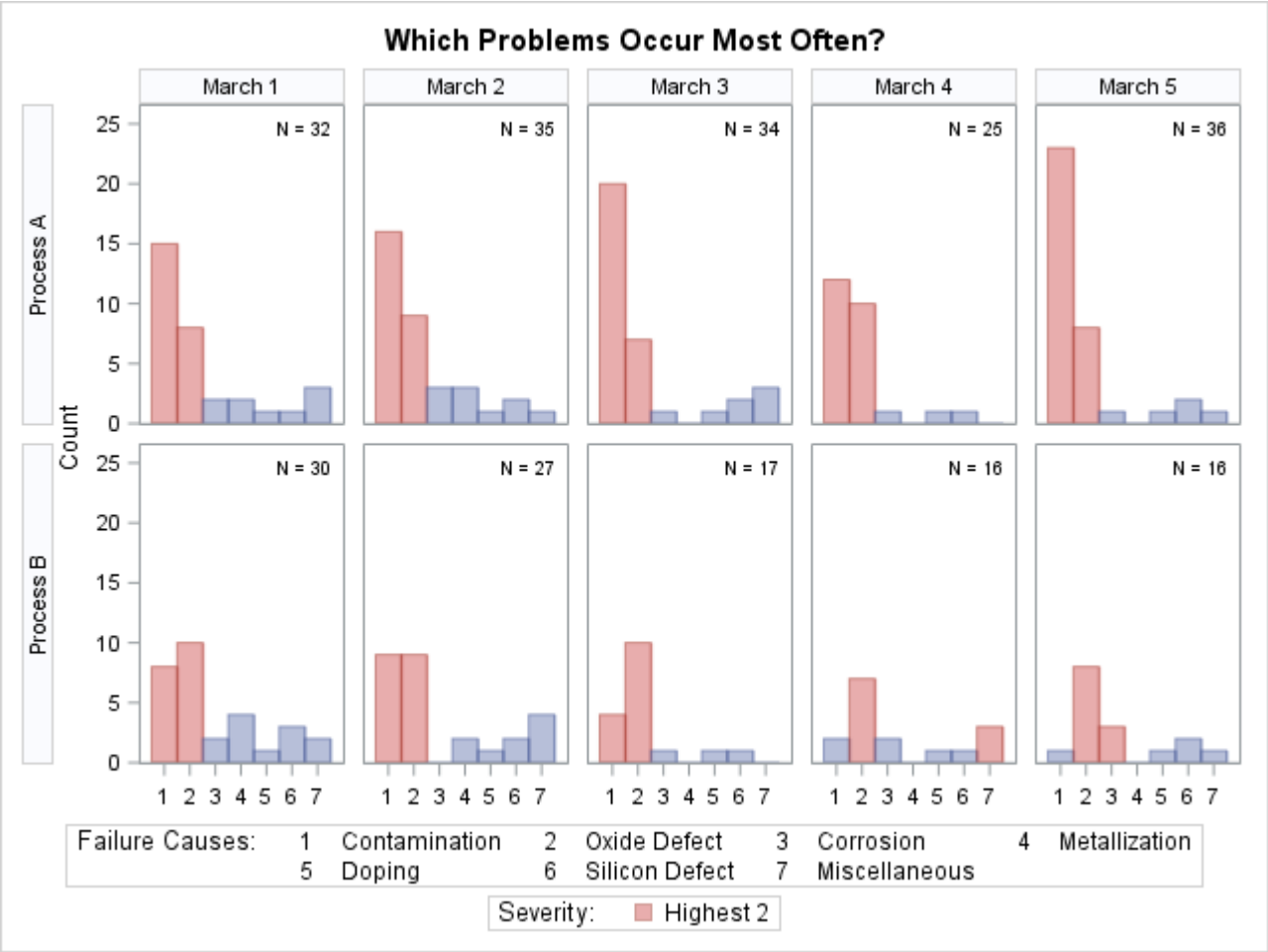
The following statements highlight the two most frequently occurring categories in each cell of the comparative Pareto chart shown in Output 16.2.4:

```
title 'Which Problems Occur Most Often?';
proc pareto data=Failure4;
   vbar Cause / class       = ( Process Day )
                freq        = Counts
                nrows       = 2
                ncols       = 5
                last        = 'Miscellaneous'
                scale       = count
                chigh(2)
                hlleglabel  = 'Severity:'
                catleglabel = 'Failure Causes:'
                odstitle    = title
                nocatlabel
                nocurve
                nlegend;
run;
```

Specifying CHIGH(2) causes the two highest bars in each cell to be filled with a contrasting color from the ODS style. The new chart is displayed in Output 16.3.1. In all but two of the cells, the two vital problems are `'Contamination'` and `'Oxide Defect'`.

You can also highlight the "trivial many" categories (also referred to as the "useful many") with the CLOW(*m*) option. You can use these options in conjunction with the CHIGH(*n*) and BARS= options. For more information, see the entries for these options in the "Dictionary of HBAR and VBAR Statement Options" on page 1071.

Output 16.3.1 Emphasizing the “Vital Few”



Example 16.4: Highlighting Combinations of Categories

NOTE: See *Highlighting Specific Pareto Categories* in the SAS/QC Sample Library.

In some applications, it is useful to classify the categories into groups that are not necessarily related to frequency. This example, which is a continuation of Example 16.2, shows how you use a bar legend to display this classification.

Suppose that contamination and metallization are high-priority problems, oxide defect is a medium-priority problem, and all other categories are low-priority problems. Begin by adding this information to the data set Failure4 as follows:

```
data Failure4;
   length Priority $ 16;
   set Failure4;
   if Cause = 'Contamination' or
      Cause = 'Metallization'
   then
      Priority = 'High';
   else
   if Cause = 'Oxide Defect'
   then
      Priority = 'Medium';
   else
      Priority = 'Low';
run;
```

The variable Priority indicates the priority that is associated with a defect cause.

The following statements specify Priority in both the BARS= and BARLEGEND= options:

```
title 'Which Problems Take Priority?';
proc pareto data=Failure4;
   vbar Cause / class       = ( Process Day )
                freq        = Counts
                nrows       = 2
                ncols       = 5
                last        = 'Miscellaneous'
                scale       = count
                bars        = ( Priority )
                barlegend   = ( Priority )
                barleglabel = 'Priority:'
                catleglabel = 'Failure Causes:'
                odstitle    = title
                nocatlabel
                nocurve
                nlegend;
run;
```

Colors from the ODS style are assigned to the bars based on levels of the BARS= variable. The chart is displayed in Output 16.4.1. The levels of the BARLEGEND= variable are the values that are displayed in the legend labeled “Priority:” at the bottom of the chart.

In general, when you specify BARS= and BARLEGEND= variables, their values must be consistent and unambiguous. Each observation that has a particular value of the process variable should have the same BARS= or BARLEGEND= variable value. For more information, see the entries for the BARS= and BARLEGEND= options in “Dictionary of HBAR and VBAR Statement Options” on page 1071.

Output 16.4.1 Highlighting Selected Subsets of Categories

Example 16.5: Highlighting Combinations of Cells

NOTE: See *Highlighting Tiles in a Comparative Pareto Chart* in the SAS/QC Sample Library.

This example is a continuation of Example 16.4.

In some applications that involve comparative Pareto charts, it is useful to classify the cells into groups. This example shows how you can use traditional graphics to display this type of classification by coloring the cells (also called tiles) and adding a legend.

Suppose you want to enhance Output 16.4.1 by highlighting the two cells for which Process='Process B' and Day='March 4' and 'March 5' to emphasize the improvement displayed in those cells. Begin by adding a tile color variable (Tilecol) and a tile legend variable (Tileleg) to the data set Failure4 as follows:

```
data Failure4;
   length Tilecol $ 8 Tileleg $ 16;
   set Failure4;
   if (Process='Process B') and (Day='March 4' or Day='March 5')
   then do; Tilecol='ywh'; Tileleg = 'Improvement'; end;
   else do; Tilecol='ligr'; Tileleg = 'Status Quo'; end;
run;
```

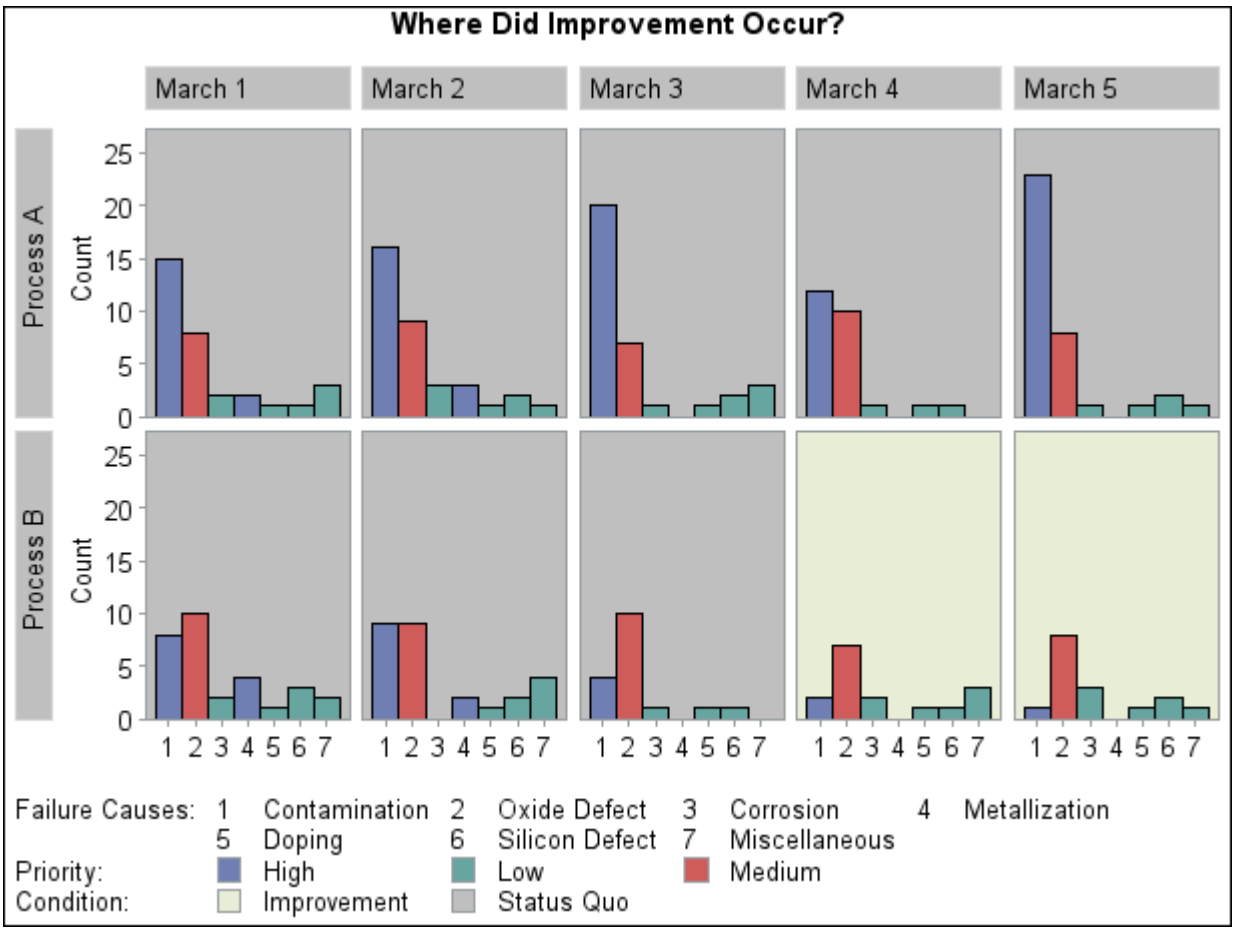
The following statements specify Tilecol as a CTILES= variable and Tileleg as a TILELEGEND= variable. Note that the variable names are enclosed in parentheses.

```
ods graphics off;
title 'Where Did Improvement Occur?';
proc pareto data=Failure4;
   vbar Cause / class        = ( Process Day )
                freq         = Counts
                nrows        = 2
                ncols        = 5
                last         = 'Miscellaneous'
                scale        = count
                catleglabel  = 'Failure Causes:'
     /* options for highlighting bars: */
                bars         = ( Priority )
                barlegend    = ( Priority )
                barleglabel  = 'Priority:'
     /* options for highlighting tiles: */
                ctiles       = ( Tilecol )
                tilelegend   = ( Tileleg )
                tileleglabel = 'Condition:'
                intertile    = 1.0
                cframeside   = ligr
                cframetop    = ligr
                nocatlabel
                nocurve;
run;
```

The ODS GRAPHICS OFF statement before the PROC statement disables ODS Graphics, so the Pareto chart is produced using traditional graphics. The CTILES=, TILELEGEND=, CFRAMESIDE=, and CFRAMETOP= options are valid only for traditional graphics output. See the section "Options for Traditional Graphics" on page 1086 for descriptions of options specific to traditional graphics.

In the chart, shown in Output 16.5.1, the values that are displayed in the legend labeled "Condition:" are the levels of the TILELEGEND= variable.

Output 16.5.1 Highlighting Specific Tiles

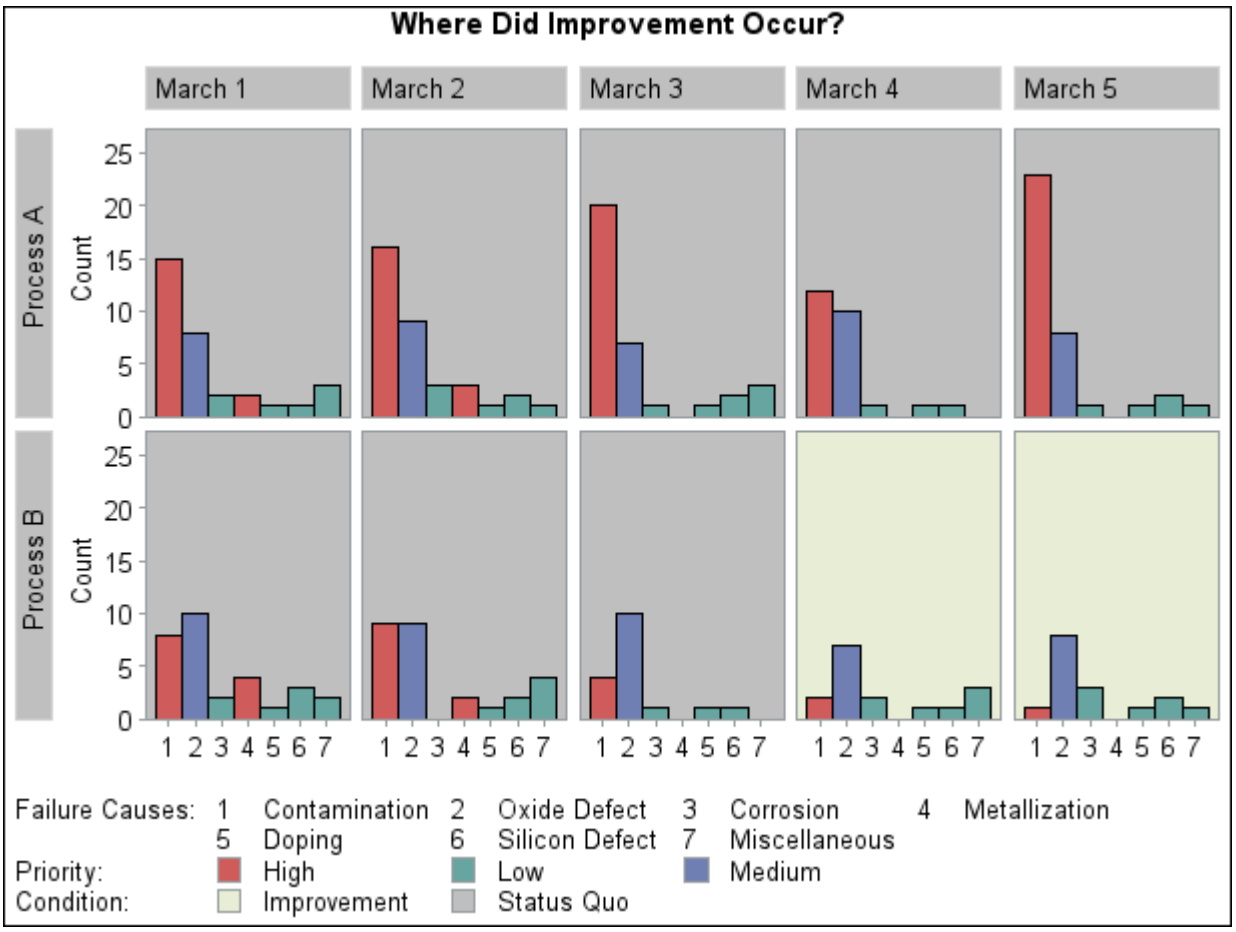


PROC PARETO sequentially assigns colors from a list defined by the ODS style to the levels of the BARS= variable. The first color is associated with the first value of Priority, and so on. When traditional graphics is enabled, you can use the CBARS= option to assign specific colors to Pareto categories. The following statements assign explicit color values to the variable PriorityColor:

```
data Failure4;
   length PriorityColor $ 8;
   set Failure4;
   if Priority = 'High'
   then
      PriorityColor = 'CXD05B5B';
   else
   if Priority = 'Medium'
   then
      PriorityColor = 'CX6F7EB3';
   else
      PriorityColor = 'CX66A5A0';
run;
```

Output 16.5.2 shows the chart that is produced by replacing the BARS= option with CBARS=PriorityColor. The high-priority problems are represented by red bars, and the low-priority problems are represented by green bars.

Output 16.5.2 Assigning Specific Bar Colors



Example 16.6: Ordering Rows and Columns in a Comparative Pareto Chart

NOTE: See *Ordering Rows and Columns in a Comparative Chart* in the SAS/QC Sample Library.

This example illustrates methods for controlling the order of rows and columns in a comparative Pareto chart.

The following statements create a data set named Failure5:

```
proc format;
   value procfmt   1   = 'Process A'
                   2   = 'Process B';
   value dayfmt    1   = 'Monday'
                   2   = 'Tuesday'
                   3   = 'Wednesday'
                   4   = 'Thursday'
                   5   = 'Friday';
run;

data Failure5;
   length Cause $16;
   format Process procfmt. Day dayfmt.;
```

```
   label  Cause   = 'Cause of Failure'
          Process = 'Cleaning Method'
          Day     = 'Day of Manufacture';
   input  Process Day Cause $16. Counts @@;
   datalines;
   1   1   Contamination    15   1   1   Corrosion         2
   1   1   Doping            1   1   1   Metallization     2
   1   1   Miscellaneous     3   1   1   Oxide Defect      8
   1   1   Silicon Defect    1   1   2   Contamination    16
   1   2   Corrosion         3   1   2   Doping            1
   1   2   Metallization     3   1   2   Miscellaneous     1
   1   2   Oxide Defect      9   1   2   Silicon Defect    2
   1   3   Contamination    20   1   3   Corrosion         1
   1   3   Doping            1   1   3   Metallization     0
   1   3   Miscellaneous     3   1   3   Oxide Defect      7
   1   3   Silicon Defect    2   1   4   Contamination    12
   1   4   Corrosion         1   1   4   Doping            1
   1   4   Metallization     0   1   4   Miscellaneous     0
   1   4   Oxide Defect     10   1   4   Silicon Defect    1
   1   5   Contamination    23   1   5   Corrosion         1
   1   5   Doping            1   1   5   Metallization     0
   1   5   Miscellaneous     1   1   5   Oxide Defect      8
   1   5   Silicon Defect    2   2   1   Contamination     8
   2   1   Corrosion         2   2   1   Doping            1
   2   1   Metallization     4   2   1   Miscellaneous     2
   2   1   Oxide Defect     10   2   1   Silicon Defect    3
   2   2   Contamination     9   2   2   Corrosion         0
   2   2   Doping            1   2   2   Metallization     2
   2   2   Miscellaneous     4   2   2   Oxide Defect      9
   2   2   Silicon Defect    2   2   3   Contamination     4
   2   3   Corrosion         1   2   3   Doping            1
   2   3   Metallization     0   2   3   Miscellaneous     0
   2   3   Oxide Defect     10   2   3   Silicon Defect    1
   2   4   Contamination     2   2   4   Corrosion         2
   2   4   Doping            1   2   4   Metallization     0
   2   4   Miscellaneous     3   2   4   Oxide Defect      7
   2   4   Silicon Defect    1   2   5   Contamination     1
   2   5   Corrosion         3   2   5   Doping            1
   2   5   Metallization     0   2   5   Miscellaneous     1
   2   5   Oxide Defect      8   2   5   Silicon Defect    2
;
```

Note that Failure5 is similar to the data set Failure4, which is created in Example 16.2. Here, the classification variables Process and Day are numeric formatted variables, and the formatted values of Day are `'Monday'` through `'Friday'`. In Example 16.2, Process and Day are character variables, and the values of Day are 'March 1' through 'March 5'.

The following statements create a two-way comparative Pareto chart for Cause; in this chart the rows represent levels of Process, and the columns represent levels of Day:

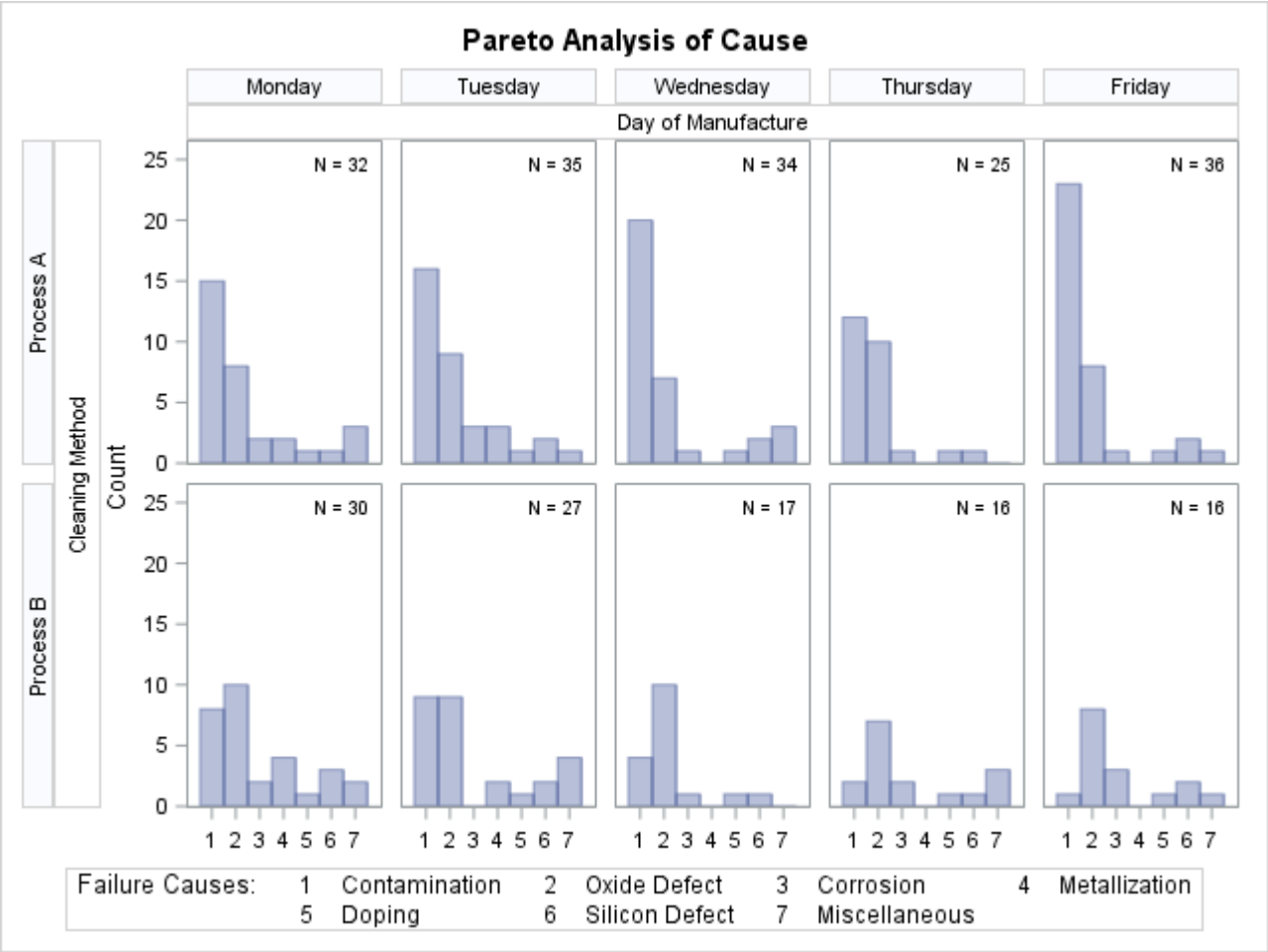
```
title 'Pareto Analysis by Process and Day';
proc pareto data=Failure5;
   vbar Cause / class       = ( Process Day )
                freq        = Counts
                nrows       = 2
                ncols       = 5
                last        = 'Miscellaneous'
                scale       = count
                catleglabel = 'Failure Causes:'
                nocatlabel
                nocurve
                nlegend;
run;
```

The chart is shown in Output 16.6.1. The levels of the classification variables are determined by their formatted values. The default order in which the rows and columns are displayed is determined by the internal values of the classification variables, and consequently the columns appear in the order of the days of the week.

If Day had been defined as a character variable with values `'Monday'` through `'Friday'`, the columns in Output 16.6.1 would have appeared in alphabetical order.

You can override the default order by specifying the ORDER1= or ORDER2= option (or both).

Output 16.6.1 Controlling Row and Column Order



Example 16.7: Merging Columns in a Comparative Pareto Chart

NOTE: See *Merging Columns in a Comparative Pareto Chart* in the SAS/QC Sample Library.

This example is a continuation of Example 16.4 and illustrates a method for merging the columns in a comparative Pareto chart.

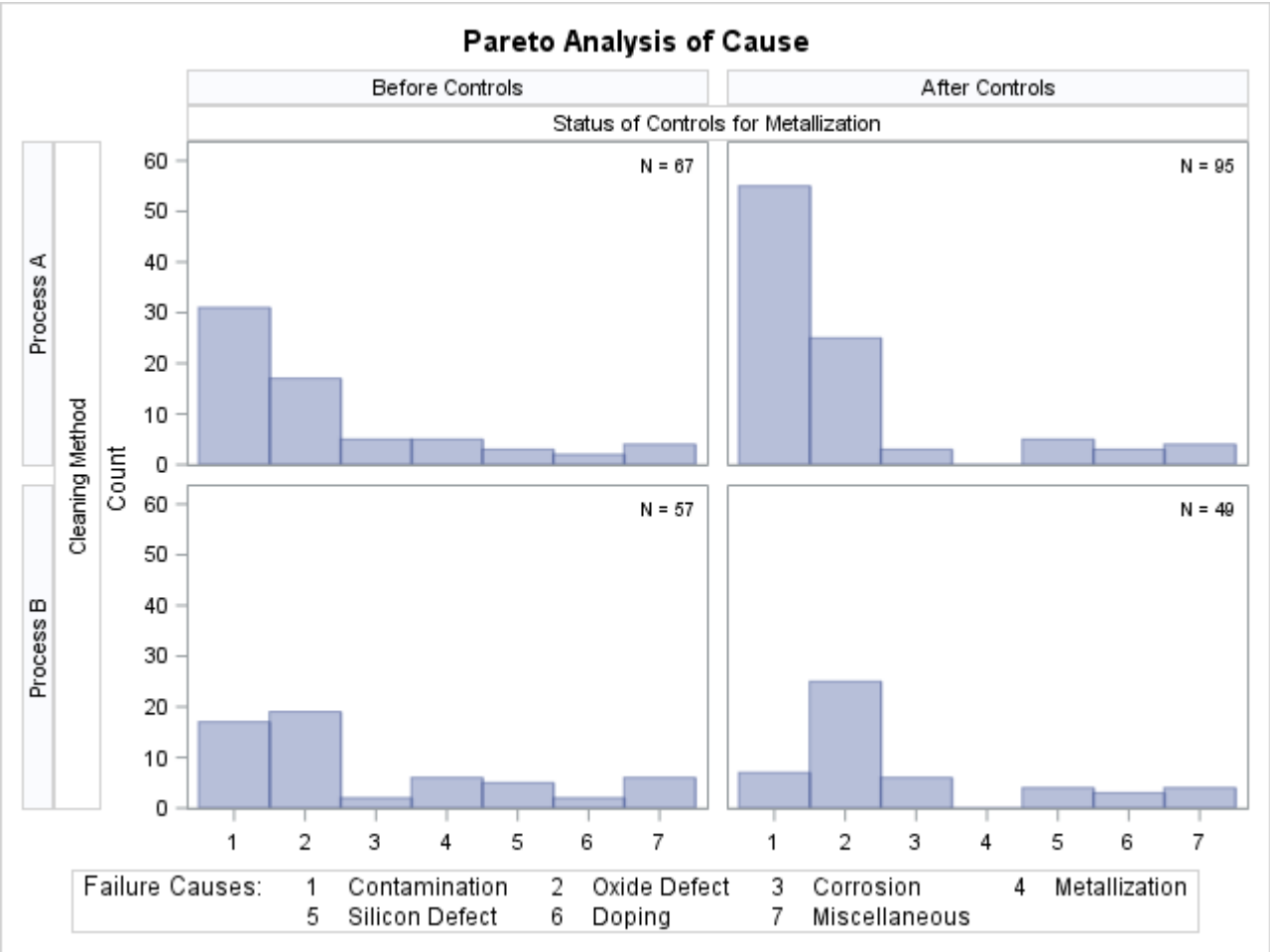
Suppose that controls for metallization were introduced on Wednesday. To show the effect of the controls, the columns for `'Monday'` and `'Tuesday'` are to be merged into a column labeled `'Before Controls'`, and the remaining columns are to be merged into a column labeled `'After Controls'`. The following statements introduce a format named `'cntlfmt'` that merges the levels of Day:

```
proc format;
   value cntlfmt   1-2 = 'Before Controls'
                   3-5 = 'After Controls';
```

The following statements create the chart shown in Output 16.7.1:

```
proc pareto data=Failure5;
   vbar Cause / class       = ( Process Day )
                freq        = Counts
                last        = 'Miscellaneous'
                scale       = count
                catleglabel = 'Failure Causes:'
                nocatlabel
                nocurve
                nlegend;
   format Day cntlfmt.;
   label  Day = 'Status of Controls for Metallization';
run;
```

Output 16.7.1 Merging Classification Levels



The levels of Day are determined by its formatted values, `'Before Controls'` and `'After Controls'`. By default, the order in which the columns are displayed is determined by the internal values. In this example, there are multiple distinct internal values for each level, and PROC PARETO uses the internal value that occurs first in the input data set.

Example 16.8: Creating Weighted Pareto Charts

NOTE: See *Pareto Analysis Based on Cost* in the SAS/QC Sample Library.

In many applications, you can quantify the priority or severity of a problem by using a measure such as the cost of repair or the loss to the customer expressed in man-hours. This example shows how to analyze such data by using a weighted Pareto chart that incorporates the cost.

Suppose that the cost associated with each of the problems in data set Failure5 (see Example 16.6) has been determined and that the costs have been converted to a relative scale. The following statements add the cost information to the data set:

```
data Failure5;
   length Analysis $ 16;
   label Analysis = 'Basis for Analysis';
   set Failure5;
   Analysis = 'Cost';
      if      Cause = 'Contamination'  then Cost = 3.0;
      else if Cause = 'Metallization'  then Cost = 8.5;
      else if Cause = 'Oxide Defect'   then Cost = 9.5;
      else if Cause = 'Corrosion'      then Cost = 2.5;
      else if Cause = 'Doping'         then Cost = 3.6;
      else if Cause = 'Silicon Defect' then Cost = 3.4;
      else                                 Cost = 1.0;
      output;
   Analysis = 'Frequency';
      Cost = 1.0;
      output;
run;
```

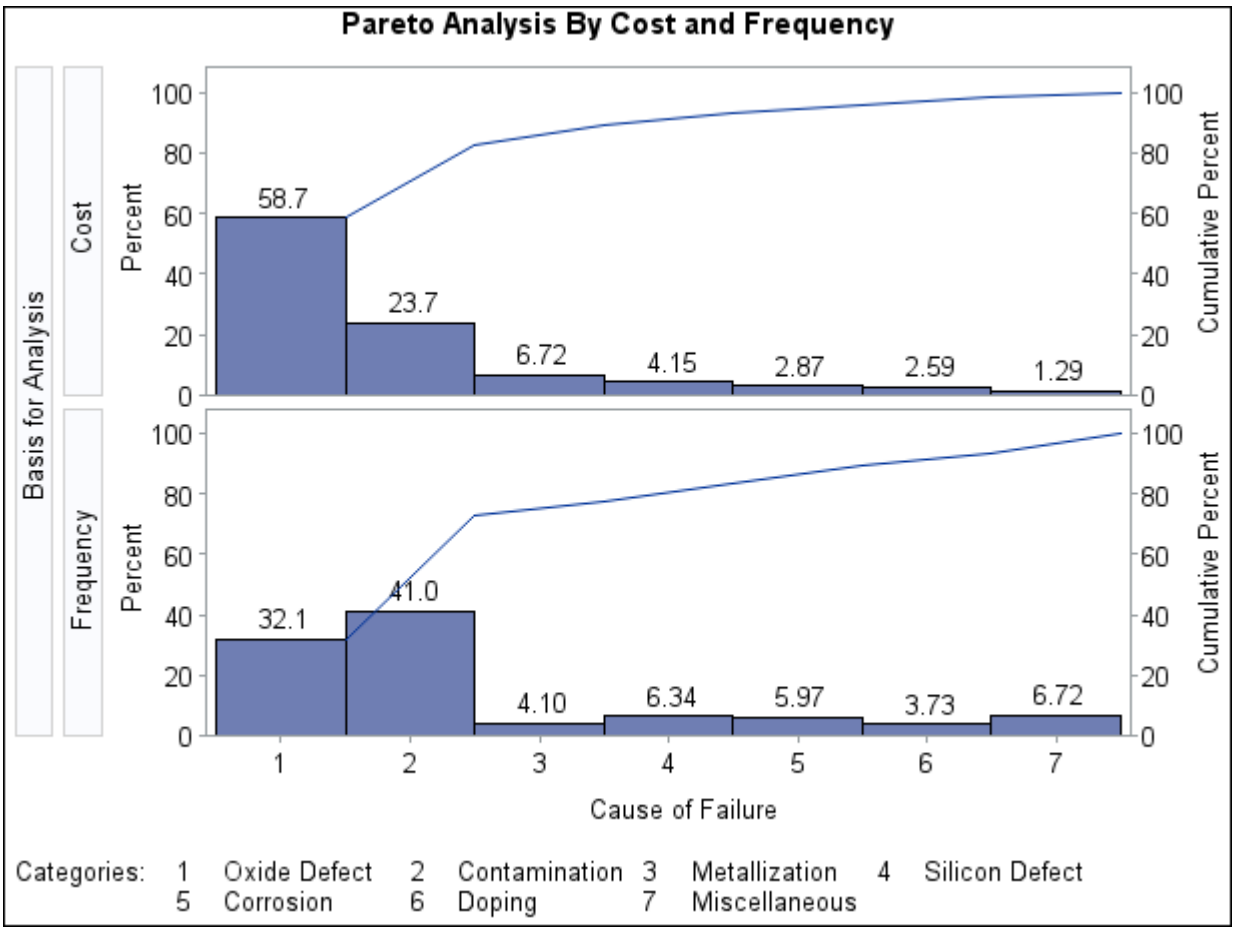
The classification variable Analysis has two levels, `'Cost'` and `'Frequency'`. For Analysis=`'Cost'`, the value of Cost is the relative cost, and for Analysis=`'Frequency'`, the value of Cost is one.

The following statements use Analysis as the classification variable to create a one-way comparative Pareto chart in which the cells are weighted Pareto charts that use Cost as the weight variable:

```
ods graphics off;
goptions vsize=4.25 in htext=2.8 pct htitle=3.2 pct;
title 'Pareto Analysis By Cost and Frequency';
proc pareto data=Failure5;
   vbar Cause / class      = ( Analysis )
                freq       = Counts
                weight     = Cost
                barlabel   = value
                out        = Summary
                intertile  = 1.0;
run;
```

The display is shown in Output 16.8.1.

Output 16.8.1 Taking Cost into Account



Within each cell, the height of a bar is the frequency of the category multiplied by the value of Cost, expressed as a percentage of the total across all categories. Thus, for the cell in which Analysis is equal to `'Frequency'`, the bars simply indicate the frequencies expressed in percentage units. This display shows that the most commonly occurring problem (contamination) is not the most expensive problem (oxide defect). The output data set Summary is listed in Output 16.8.2.

Output 16.8.2 Summary Output Data Set

Pareto Analysis By Cost and Frequency

Obs	Analysis	Cause	Cost	_COUNT_	_WCOUNT_	_PCT_	_CMPCT_
1	Cost	Oxide Defect	9.5	86	817.0	58.6799	58.680
2	Cost	Contamination	3.0	110	330.0	23.7018	82.382
3	Cost	Metallization	8.5	11	93.5	6.7155	89.097
4	Cost	Silicon Defect	3.4	17	57.8	4.1514	93.249
5	Cost	Corrosion	2.5	16	40.0	2.8729	96.122
6	Cost	Doping	3.6	10	36.0	2.5856	98.707
7	Cost	Miscellaneous	1.0	18	18.0	1.2928	100.000
8	Frequency	Oxide Defect	1.0	86	86.0	32.0896	32.090
9	Frequency	Contamination	1.0	110	110.0	41.0448	73.134
10	Frequency	Metallization	1.0	11	11.0	4.1045	77.239
11	Frequency	Silicon Defect	1.0	17	17.0	6.3433	83.582
12	Frequency	Corrosion	1.0	16	16.0	5.9701	89.552
13	Frequency	Doping	1.0	10	10.0	3.7313	93.284
14	Frequency	Miscellaneous	1.0	18	18.0	6.7164	100.000

Example 16.9: Creating Alternative Pareto Charts

NOTE: See *Alternative Pareto Charts* in the SAS/QC Sample Library.

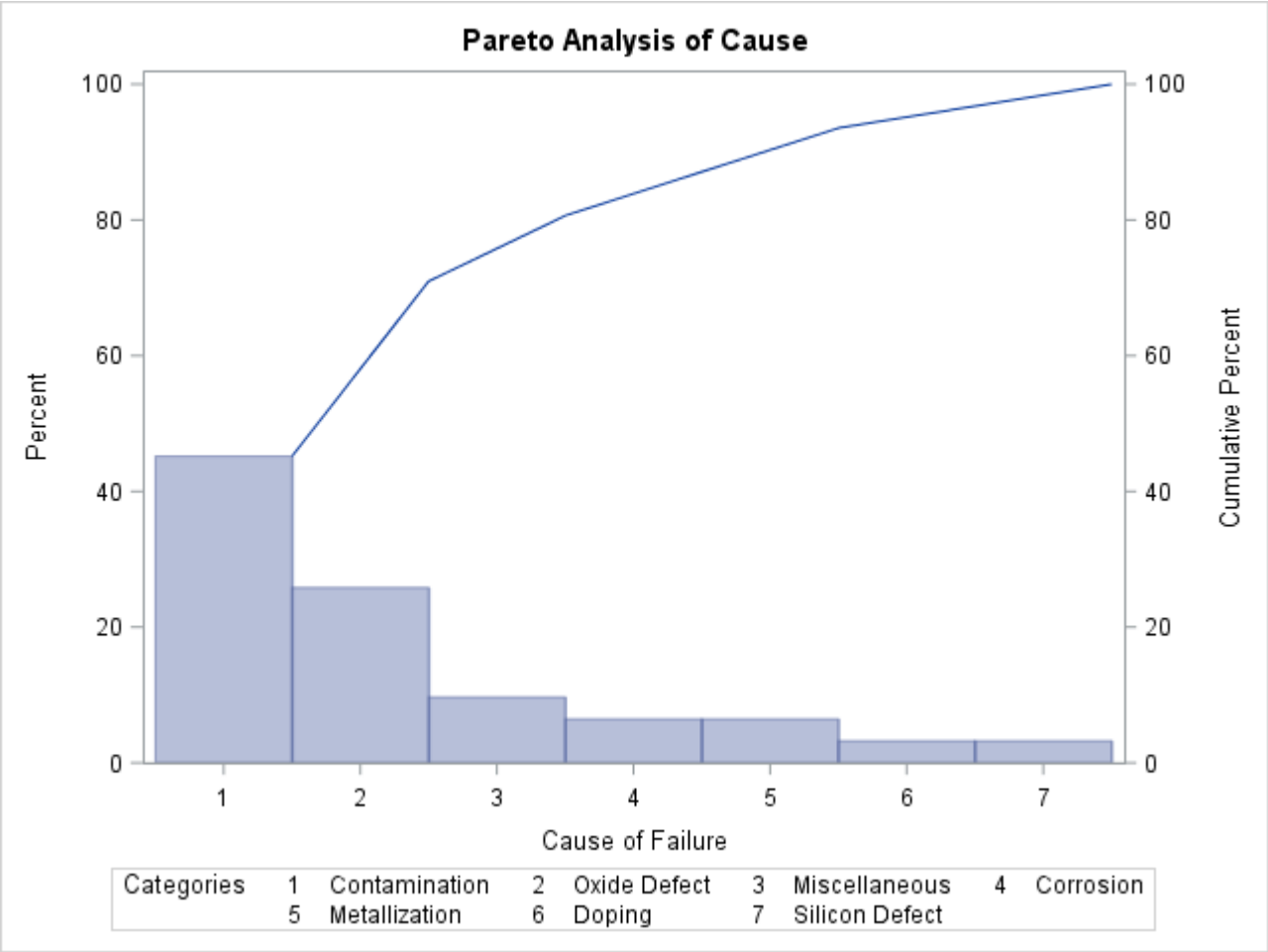
This example uses the Failure1 data set of integrated circuit fabrication failures from the section "Creating a Pareto Chart from Raw Data" on page 1045. The following statements use the CHARTTYPE= option to produce a standard Pareto chart, a cumulative Pareto bar chart, and a Pareto dot plot that includes acceptance intervals for the data:

```
proc pareto data=Failure1;
   vbar Cause;
   vbar Cause / charttype=cumulative;
   vbar Cause / charttype=intervals;
run;
```

NOTE: ODS Graphics must be enabled for you to use the CHARTTYPE= option.

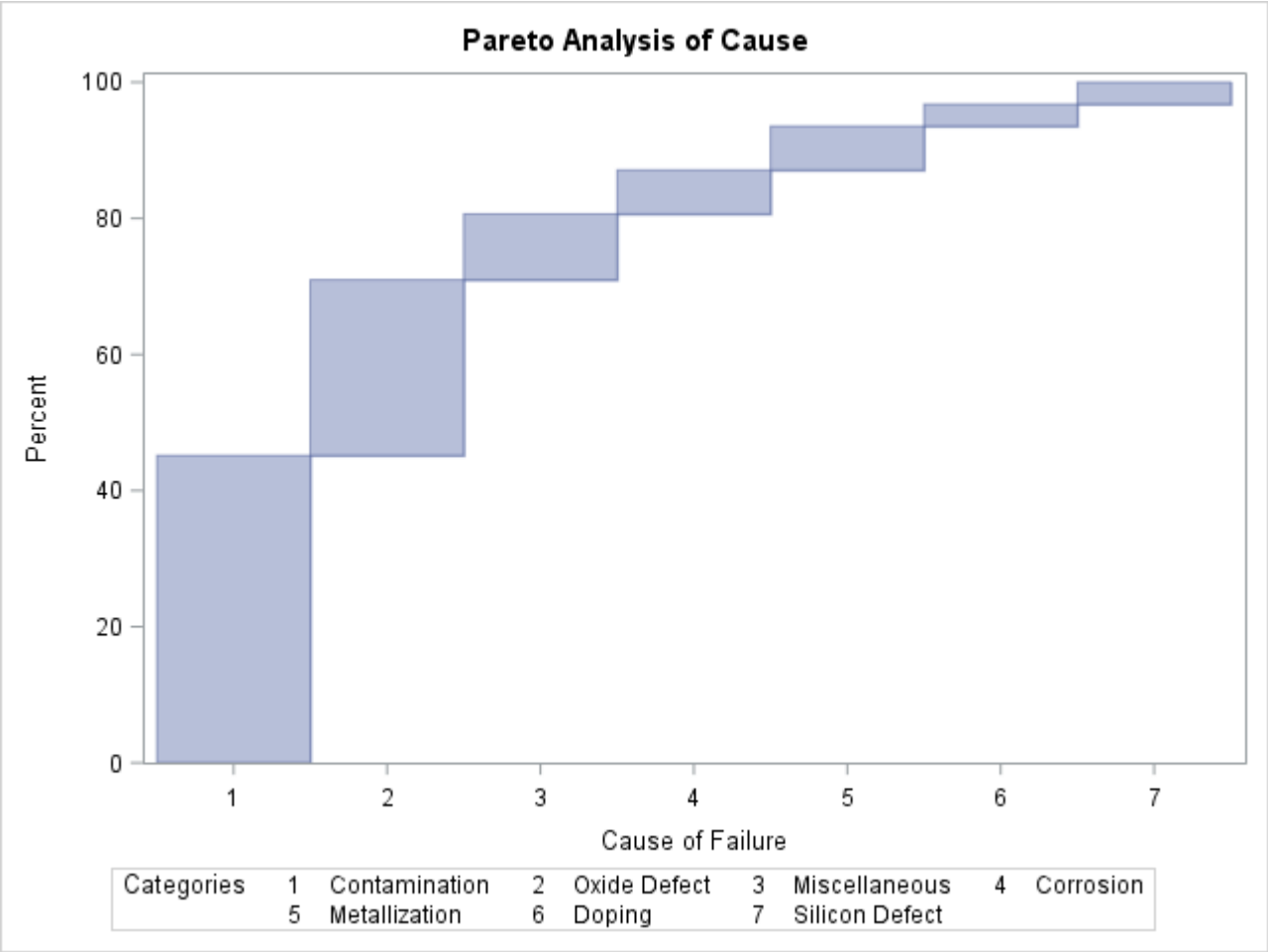
Output 16.9.1 shows the standard Pareto chart that the first VBAR statement produces.

Output 16.9.1 Standard Pareto Chart



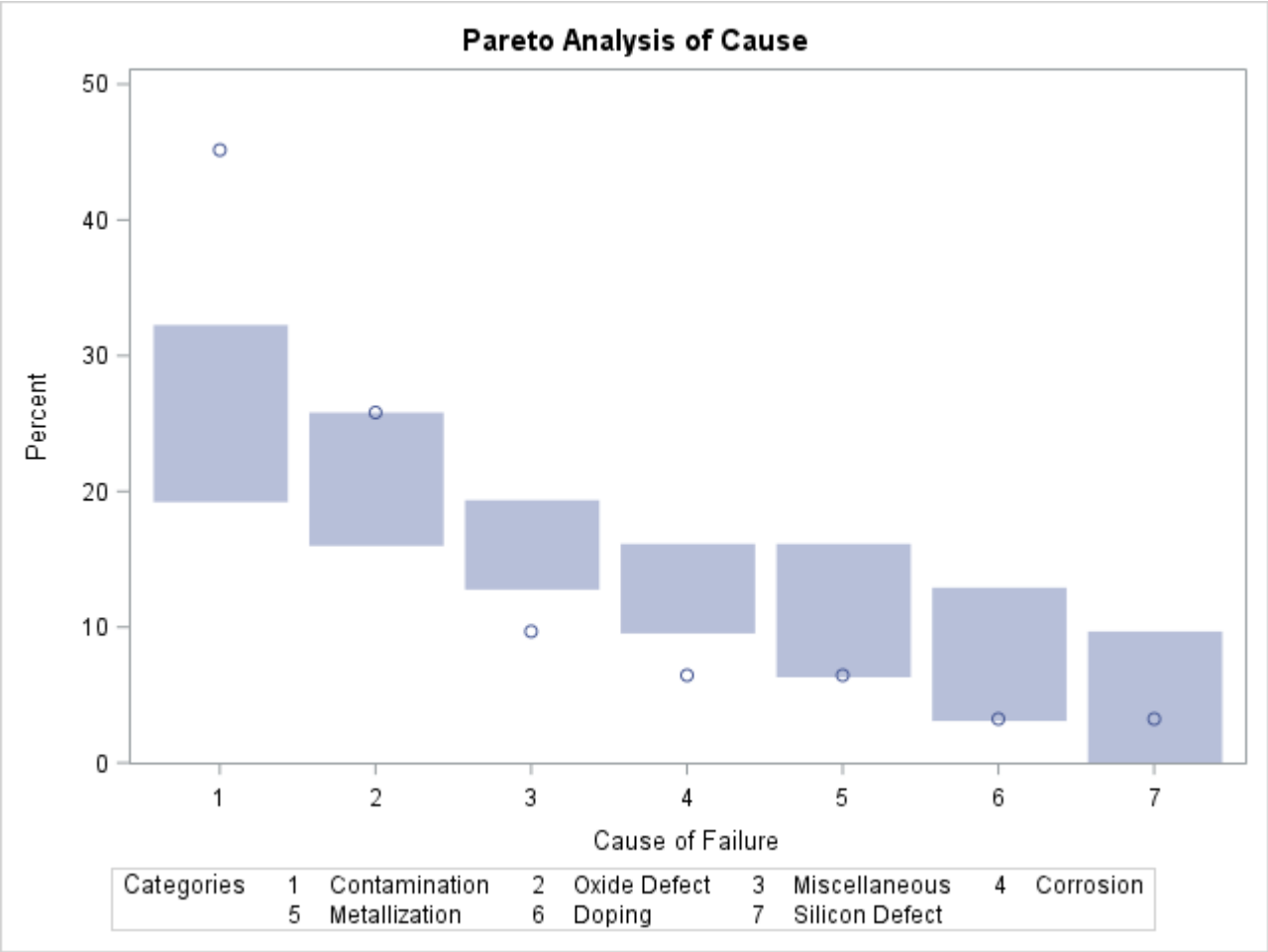
Output 16.9.2 shows the cumulative Pareto bar chart that the second VBAR statement produces.

Output 16.9.2 Cumulative Pareto Bar Chart



Output 16.9.3 shows the Pareto dot plot and acceptance intervals that the third VBAR statement produces.

Output 16.9.3 Pareto Dot Plot and Acceptance Intervals



Output 16.9.3 shows that the most frequently occurring problem, *Contamination*, occurs more frequently than the first-ranked cause from a random sample of seven uniformly distributed causes. This result indicates that addressing contamination problems should be given a high priority.

Example 16.10: Customizing Inset Labels and Formatting Values

NOTE: See *Customizing Inset Labels and Formatting Values* in the SAS/QC Sample Library.

When you add an inset to a Pareto chart, by default each inset statistic is identified by an appropriate label and its value is displayed using an appropriate format. However, you might want to provide your own labels and formats. For example, in Figure 16.7 the default label used for the N statistic is not very descriptive. The following statements produce a comparative Pareto chart whose insets display longer labels for both statistics. A format that uses one decimal place is also specified for each statistic. (These are integer values—the decimals are added only to demonstrate this feature.) Note that a single INSET statement produces an inset in each cell of the comparative Pareto chart.

```
proc pareto data=Failure3;
   vbar Cause /
      class    = Stage
      freq     = Counts
      maxncat  = 5
      classkey = 'Before Cleaning';
   inset n    ='Observations Shown' (4.1)
         nexcl='Observations Excluded' (3.1);
run;
```

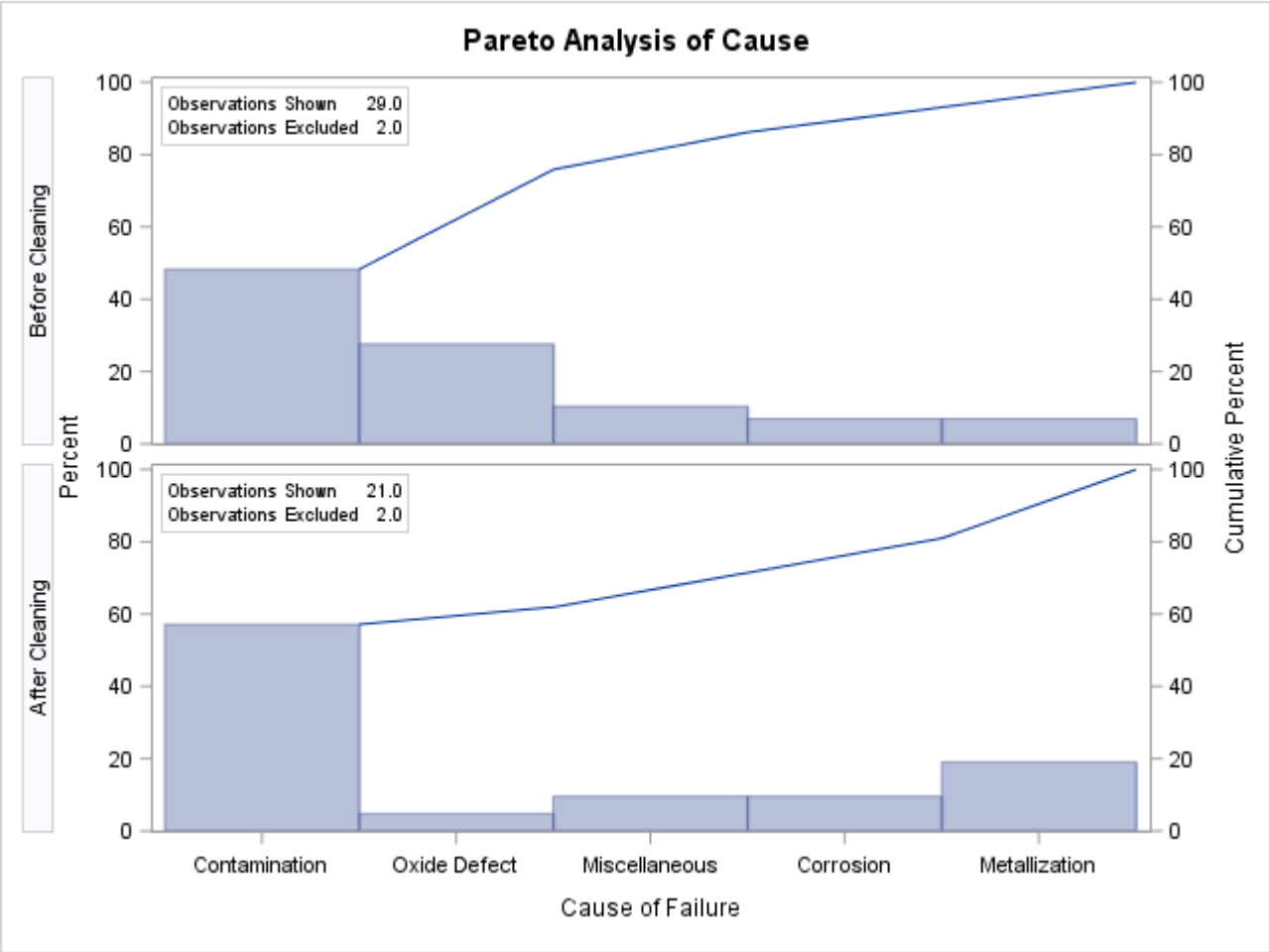
The resulting chart is displayed in Output 16.10.1.

You can provide your own label by specifying the keyword for that statistic followed by an equal sign (=) and the label in quotation marks. The label can have up to 24 characters.

The format 4.1 specified in parentheses after the N keyword displays the statistics by using a field width of four and one decimal place. In general, you can specify any numeric SAS format in parentheses after an inset keyword. You can also use the FORMAT= option to specify a format to be used for all the statistics in the INSET statement. For more information about SAS formats, see *SAS Formats and Informats: Reference*.

NOTE: If you specify both a label and a format for a statistic, the label must appear before the format.

Output 16.10.1 Customizing Labels and Formatting Values in an Inset



Example 16.11: Specifying Inset Headers and Positions

NOTE: See *Specifying Inset Headers and Positions* in the SAS/QC Sample Library.

By default, the first INSET statement that is specified after a chart statement displays an inset in the upper left corner of the chart. You can control the inset position by specifying the POSITION= option. In addition, you can display a header at the top of the inset by specifying the HEADER= option. The following statements create a data set to be used with the INSET DATA= keyword and produce the horizontal Pareto chart shown in Output 16.11.1:

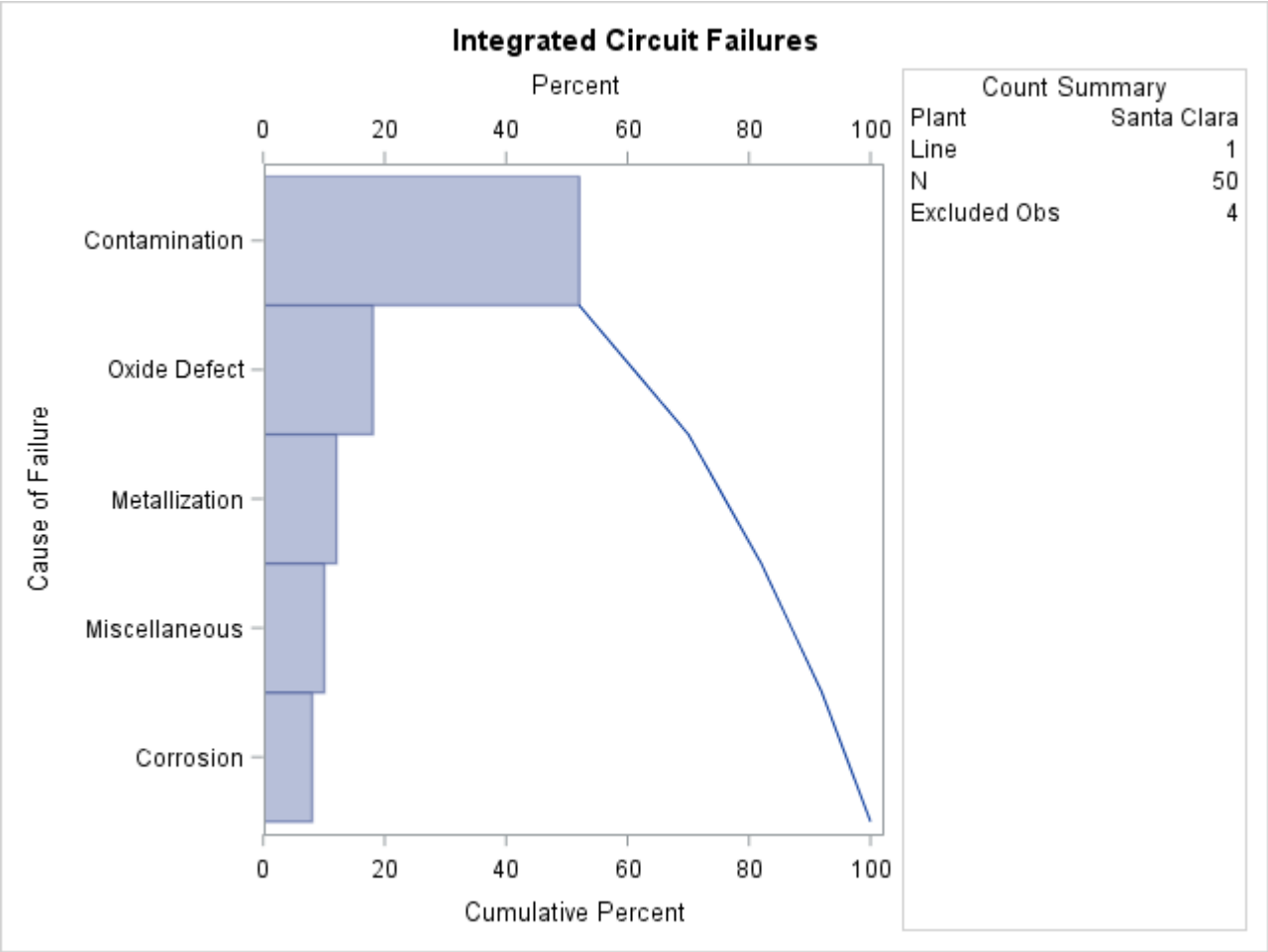
```
data location;
   length _LABEL_ $ 10 _VALUE_ $ 12;
   input _LABEL_ _VALUE_ &;
datalines;
Plant     Santa Clara
Line      1
;

title 'Integrated Circuit Failures';
proc pareto data=Failure3;
   hbar Cause /
      freq     = Counts
      maxncat  = 5
      odstitle = title;
   inset data = location n nexcl /
      position = rm
      header   = 'Count Summary';
run;
```

The header (in this case, “Count Summary”) can be up to 40 characters. The POSITION=RM option is specified to position the inset in the right margin so that it does not interfere with features of the chart. For more information about positioning, see the section “Positioning Insets” on page 1097.

INSET statement options, such as the POSITION= and HEADER= options, are specified after the slash (/). For more information about INSET statement options, see the section “INSET Statement Options” on page 1063.

Note that the contents of the data set location appear before other statistics in the inset. The position of the DATA= keyword in the keyword list determines the position of the data set’s contents in the inset.

Output 16.11.1 Adding a Header and Repositioning the Inset



Example 16.12: Managing a Large Number of Categories

NOTE: See *Managing a Large Number of Categories* in the SAS/QC Sample Library.

The Centers for Disease Control publish a variety of public health statistics. The numbers of deaths in 2010 in the United States that were caused by various types of cancer are recorded in the SAS data set CancerDeaths2010:

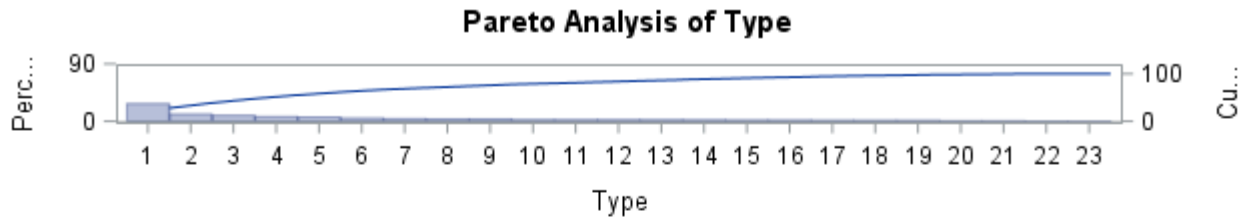
```
data CancerDeaths2010;
   length Type $ 45;
   input Type & @47 Deaths comma7.;
   datalines;
Lip, oral cavity and pharynx                    8,474
Esophagus                                      14,490
Stomach                                        11,390
Colon, rectum and anus                         52,622
Liver and intrahepatic bile ducts              20,305
Pancreas                                       36,888
Larynx                                          3,691
Trachea, bronchus and lung                    158,318
Skin                                            9,154
Breast                                         41,435
Cervix                                          3,939
Uterus                                          8,402
Ovary                                          14,572
Prostate                                       28,561
Kidney and renal pelvis                        13,219
Bladder                                        14,731
Meninges, brain, other central nervous system  14,164
Hodgkin's disease                               1,231
Non-Hodgkin's lymphoma                         20,294
Leukemia                                       22,569
Multiple myeloma and immunoproliferative       11,428
Other lymphoid, hematopoietic and related          68
All other and unspecified                      64,798
;
```

The following statements produce a Pareto chart for the data in CancerDeaths2010:

```
proc pareto data=CancerDeaths2010;
   vbar Type / freq = Deaths;
run;
```

The resulting Pareto chart is shown in Output 16.12.1.

Output 16.12.1 Cancer Deaths Pareto Chart with Default Width



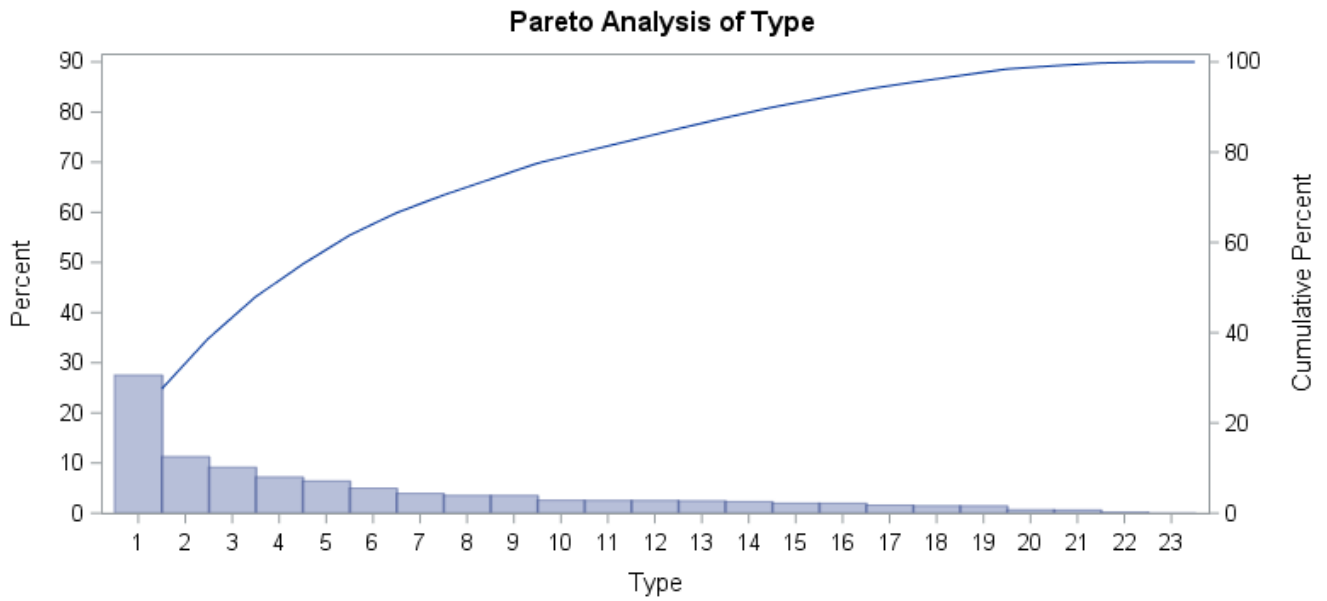
Categories	1	Trachea, bronchus and lung
	2	All other and unspecified
	3	Colon, rectum and anus
	4	Breast
	5	Pancreas
	6	Prostate
	7	Leukemia
	8	Liver and intrahepatic bile ducts
	9	Non-Hodgkin's lymphoma
	10	Bladder
	11	Ovary
	12	Esophagus
	13	Meninges, brain, other central nervous system
	14	Kidney and renal pelvis
	15	Multiple myeloma and immunoproliferative
	16	Stomach
	17	Skin
	18	Lip, oral cavity and pharynx
	19	Uterus
	20	Cervix
	21	Larynx
	22	Hodgkin's disease
	23	Other lymphoid, hematopoietic and related

Note that PROC PARETO has labeled the category axis tick marks with numbers and produced a corresponding category legend. This is done by default when there is not enough room to use category names as tick labels on the category axis. Unfortunately, because some of the category names are long, the legend has room for only one column of entries and therefore occupies an inordinate amount of space. Among the alternatives for addressing this problem are the following:

- replacing the original category names with shorter ones
- increasing the space available for the graph

You can implement the second alternative by specifying the WIDTH= option in the ODS GRAPHICS statement prior to invoking the procedure. (The ODS GRAPHICS statement is documented in the *SAS Output Delivery System: User's Guide*.) Output 16.12.2 shows the Pareto chart that is produced after the graph width is increased.

Output 16.12.2 Cancer Deaths Pareto Chart with Increased Width



Categories				
	1	Trachea, bronchus and lung	2	All other and unspecified
	3	Colon, rectum and anus	4	Breast
	5	Pancreas	6	Prostate
	7	Leukemia	8	Liver and intrahepatic bile ducts
	9	Non-Hodgkin's lymphoma	10	Bladder
	11	Ovary	12	Esophagus
	13	Meninges, brain, other central nervous system	14	Kidney and renal pelvis
	15	Multiple myeloma and immunoproliferative	16	Stomach
	17	Skin	18	Lip, oral cavity and pharynx
	19	Uterus	20	Cervix
	21	Larynx	22	Hodgkin's disease
	23	Other lymphoid, hematopoietic and related		

In a standard Pareto chart, the cumulative percentage curve is anchored at the top of the first category bar. In Output 16.12.2 PROC PARETO has automatically relaxed that rule to avoid excessive compression of the bars. You can use the FREQAXIS= option to specify that the frequency axis extend to 100%, which restores the anchoring of the curve. (For more information about scaling the frequency and cumulative percentage axes, see the section "Scaling the Cumulative Percentage Curve" on page 1096.)

Note also in Output 16.12.2 that the category `'All other and unspecified'` has the second highest frequency. To better indicate the specific types of cancer responsible for the most deaths, you can use the LAST= option to display the `'All other and unspecified'` category last.

The following statements incorporate these changes and add other enhancements to the chart:

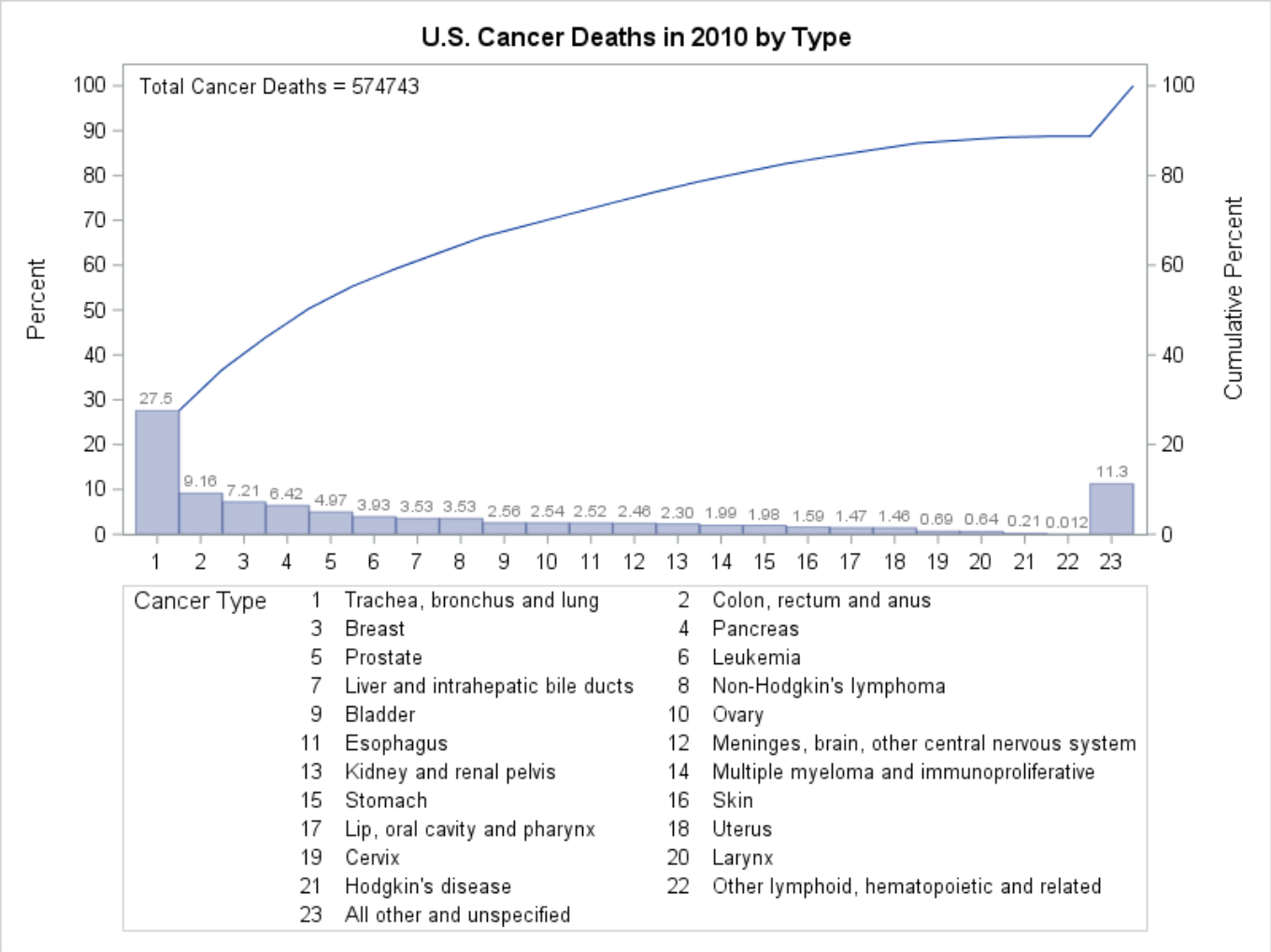
```
ods graphics / width=800px;
title 'U.S. Cancer Deaths in 2010 by Type';
proc pareto data=CancerDeaths2010;
   vbar Type / freq        = Deaths
               barlabel    = value
               last        = 'All other and unspecified'
               nocatlabel
```

```
                    catleglabel = 'Cancer Type'
                    freqaxis    = 0 to 100 by 10
                    nlegend     = 'Total Cancer Deaths'
                    odstitle    = title
                    out         = CSummary;
                    ;
run;
```

The BARLABEL= option labels each bar with its value in frequency axis units, which in this case is the percentage of cancer deaths that were caused by that type of cancer. The NOCATLABEL option saves some space by eliminating the category axis label, and the CATLEGLABEL= option produces a more informative label for the category legend. The NLEGEND= option displays the total sample size with an appropriate label. The ODSTITLE= option replaces the default graph title with the one specified in the TITLE statement. The OUT= option saves a summary of the Pareto chart in the data set CSummary.

The improved Pareto chart is shown in Output 16.12.3, and a listing of CSummary is shown in Output 16.12.4.

Output 16.12.3 Improved Pareto Chart of 2010 Cancer Deaths



Output 16.12.4 CSummary Data Set

U.S. Cancer Deaths in 2010 by Type

Obs	Type	_COUNT_	_PCT_	_CMPCT_
1	Trachea, bronchus and lung	158318	27.5459	27.546
2	Colon, rectum and anus	52622	9.1557	36.702
3	Breast	41435	7.2093	43.911
4	Pancreas	36888	6.4182	50.329
5	Prostate	28561	4.9694	55.298
6	Leukemia	22569	3.9268	59.225
7	Liver and intrahepatic bile ducts	20305	3.5329	62.758
8	Non-Hodgkin's lymphoma	20294	3.5310	66.289
9	Bladder	14731	2.5631	68.852
10	Ovary	14572	2.5354	71.388
11	Esophagus	14490	2.5211	73.909
12	Meninges, brain, other central nervous system	14164	2.4644	76.373
13	Kidney and renal pelvis	13219	2.3000	78.673
14	Multiple myeloma and immunoproliferative	11428	1.9884	80.661
15	Stomach	11390	1.9818	82.643
16	Skin	9154	1.5927	84.236
17	Lip, oral cavity and pharynx	8474	1.4744	85.710
18	Uterus	8402	1.4619	87.172
19	Cervix	3939	0.6853	87.858
20	Larynx	3691	0.6422	88.500
21	Hodgkin's disease	1231	0.2142	88.714
22	Other lymphoid, hematopoietic and related	68	0.0118	88.726
23	All other and unspecified	64798	11.2743	100.000

The Pareto chart in Output 16.12.3 has 23 categories, some of which account for only a small percentage of the total deaths. Often only a relatively few categories that have the highest frequencies are of interest. The PARETO procedure provides options for limiting the number of categories that are displayed on a chart. For an example of restricting the number of categories by using the MAXNCAT= and OTHER= options, see the section "Restricting the Number of Pareto Categories" on page 1050.

The original CancerDeaths2010 data set appears to have been summarized in advance, with the `'All other and unspecified'` category containing the total count for unspecified cancers plus those types that account for fewer deaths than the 22 distinct types that are shown in Output 16.12.3. The `'All other and unspecified'` category has the second highest frequency, accounting for 11.3% of all deaths.

The chart statement options that limit the number of categories to be displayed omit or merge low-frequency categories. In this case, it is more useful to merge the low-frequency categories into the existing `'All other and unspecified'` category. The following DATA step merges each type that accounts for less than 2% of cancer deaths into the `'All other and unspecified'` category:

```
data CSummary;
   set CSummary;
   if _PCT_ < 2.0 then Type='All other and unspecified';
run;
```

The modified CSummary data set is shown in Output 16.12.5.

Output 16.12.5 Modified CSummary Data Set

U.S. Cancer Deaths in 2010 by Type

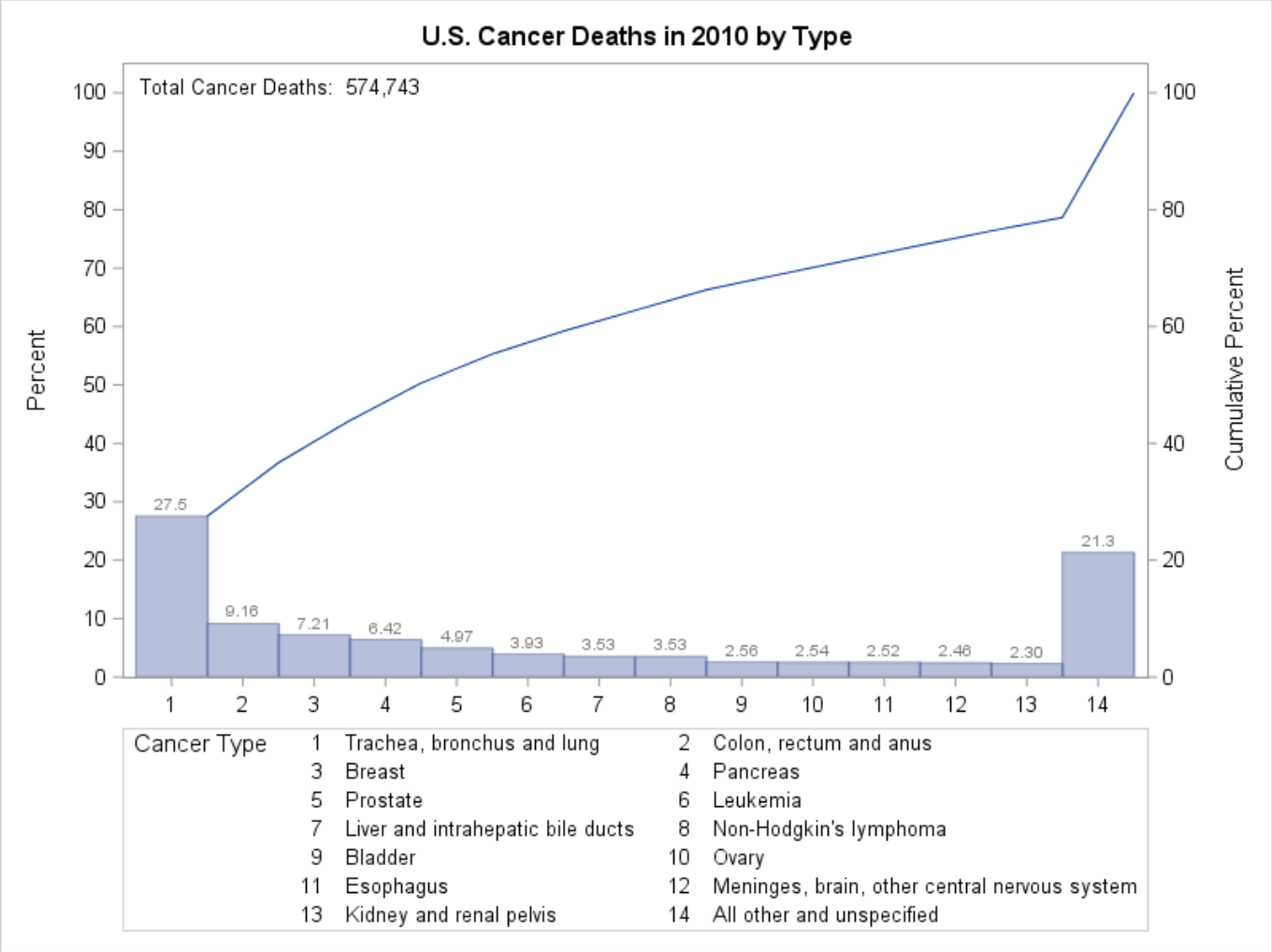
Obs	Type	_COUNT_	_PCT_	_CMPCT_
1	Trachea, bronchus and lung	158318	27.5459	27.546
2	Colon, rectum and anus	52622	9.1557	36.702
3	Breast	41435	7.2093	43.911
4	Pancreas	36888	6.4182	50.329
5	Prostate	28561	4.9694	55.298
6	Leukemia	22569	3.9268	59.225
7	Liver and intrahepatic bile ducts	20305	3.5329	62.758
8	Non-Hodgkin's lymphoma	20294	3.5310	66.289
9	Bladder	14731	2.5631	68.852
10	Ovary	14572	2.5354	71.388
11	Esophagus	14490	2.5211	73.909
12	Meninges, brain, other central nervous system	14164	2.4644	76.373
13	Kidney and renal pelvis	13219	2.3000	78.673
14	All other and unspecified	11428	1.9884	80.661
15	All other and unspecified	11390	1.9818	82.643
16	All other and unspecified	9154	1.5927	84.236
17	All other and unspecified	8474	1.4744	85.710
18	All other and unspecified	8402	1.4619	87.172
19	All other and unspecified	3939	0.6853	87.858
20	All other and unspecified	3691	0.6422	88.500
21	All other and unspecified	1231	0.2142	88.714
22	All other and unspecified	68	0.0118	88.726
23	All other and unspecified	64798	11.2743	100.000

Note that although CSummary contains frequency data, it can contain multiple observations that have the same category value. The following statements create a Pareto chart from the modified CSummary data set:

```
proc pareto data=CSummary;
   vbar Type / freq        = _COUNT_
               last        = 'All other and unspecified'
               barlabel    = value
               nocatlabel
               catleglabel = 'Cancer Type'
               freqaxis    = 0 to 100 by 10
               odstitle    = title;
   inset n='Total Cancer Deaths:'(comma7.) / noframe;
run;
```

Note that the sample size legend in Output 16.12.3 displays the sample size as an unformatted integer. By using an INSET statement instead of the NLEGEND= option, you can specify a format for the sample size. (For a complete description of the INSET statement, see the section "INSET Statement" on page 1061.) The resulting chart is shown in Output 16.12.6.

Output 16.12.6 Cancer Deaths Pareto Chart with Fewer Categories



Output 16.12.6 shows that 21.3% of deaths are assigned to `'All other and unspecified'` and that the bar frequencies sum to 100%.

References

Burr, J. T. (1990). "The Tools of Quality, Part 6: Pareto Charts." *Quality Progress* 23:59–61.

Cleveland, W. S. (1985). *The Elements of Graphing Data*. Monterey, CA: Wadsworth.

Ishikawa, K. (1976). *Guide to Quality Control*. Tokyo: Asian Productivity Organization.

Kume, H. (1985). *Statistical Methods for Quality Improvement*. Tokyo: AOTS Chosakai.

Wadsworth, H. M., Stephens, K. S., and Godfrey, A. B. (1986). *Modern Methods for Quality Control and Improvement*. New York: John Wiley & Sons.

Wilkinson, L. (2006). "Revising the Pareto Chart." *American Statistician* 60:332–334.

Subject Index

Syntax Index