Chapter 25
Introduction to the PARETO Procedure

Chapter Table of Contents

OVERVIEW ................................................................. 881
GETTING STARTED ..................................................... 883
  Creating a Pareto Chart from Raw Data .......................... 883
  Creating a Pareto Chart Using Frequency Data ............... 885
  Restricting the Number of Pareto Categories .................. 888
Chapter 25
Introduction to the PARETO Procedure

Overview

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

Pareto charts provide a tool for visualizing the Pareto principle,* 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” heavily used by workers and engineers. Ishikawa (1976) discusses how to construct and interpret a Pareto diagram. Examples of Pareto diagrams are also given by Kume (1985) and Wadsworth and others (1986).

You can use the PARETO procedure to

- construct Pareto charts from unsorted raw data (for instance, 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 based on the percentage of occurrence of each problem, the frequency (number of occurrences), or a weighted frequency (such as frequency weighted by the cost of each problem)
- add a curve indicating the cumulative percentage across categories
- construct side-by-side Pareto charts or stacked Pareto charts
- construct comparative Pareto charts that enable you to compare the Pareto frequencies across levels of one or two classification variables. For example, you can compare the frequencies of problems encountered with three different machines for five consecutive days.
- highlight the “vital few” and the “useful many”\(^1\) categories by using different colors for bars corresponding to the \(n\) most frequently occurring categories or the \(m\) least frequently occurring categories.

* 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.

\(^1\) Juran originally referred to these categories as the “trivial many”; however, because all problems merit attention, the term “useful many” is preferable. Refer to Burr (1990).
Part 7. The CAPABILITY Procedure

- highlight special categories by using different colors for specific bars
- create charts using either a high-resolution graphics device or a line printer
- annotate charts created on graphics devices
- save charts created on graphics devices in a graphics catalog for subsequent replay
- display sample sizes on Pareto charts
- display frequencies above the bars
- define characters used for features on plots produced on line printers
- save information associated with the categories (such as the frequencies) in an output data set
- restrict the number of categories displayed to the \( n \) most frequently occurring categories

If you are using the PARETO procedure for the first time, begin by reading the “Getting Started” section that follows.
Getting Started

The examples in this section illustrate basic features of the PARETO procedure.

Creating a Pareto Chart from Raw Data

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.

```sas
data failure1;
  length cause $ 16;
  label cause = 'Cause of Failure';
  input cause;
  cards;
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 there
Part 7. The CAPABILITY Procedure

is more than one observation with the same value of CAUSE, and the observations are not sorted by CAUSE. The following statements produce a basic Pareto chart for the failures:

```
proc pareto data=failure1;
  vbar cause;
run;
```

The PARETO procedure is invoked with the first statement, referred to as the PROC statement. You specify the process variable to be analyzed in the VBAR statement.

The Pareto chart is shown in Figure 25.1.

![Pareto Chart](image)

**Figure 25.1.** Pareto Chart for IC Failures in the Data Set FAILURE1

The procedure has classified the values of CAUSE into seven distinct categories (levels). The bars represent the percent 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 percent of failures from left to right; for example, *Contamination* and *Oxide* together account for 71% of the failures.

If there is sufficient space, the procedure labels the bars along the horizontal axis as in Figure 25.1. Otherwise, as in Figure 25.2, the procedure numbers the bars from left to right and adds a legend identifying the categories.
A category legend is likely to be introduced when

- the number of categories is large
- the category labels are lengthy (as in this example). Category labels can be up to 16 characters.
- a large text height is used. You can specify the height with the HEIGHT= option in the VBAR statement or with the HTEXT= option in a GOPTIONS statement (not shown here).

Creating a Pareto Chart Using Frequency Data

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.

```sas
proc freq data=failure1;
  tables cause / noprint out=failure2;
proc print;
run;
```
A listing of FAILURE2 is shown in Figure 25.3.

<table>
<thead>
<tr>
<th>Obs</th>
<th>cause</th>
<th>COUNT</th>
<th>PERCENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Contamination</td>
<td>14</td>
<td>45.1613</td>
</tr>
<tr>
<td>2</td>
<td>Corrosion</td>
<td>2</td>
<td>6.4516</td>
</tr>
<tr>
<td>3</td>
<td>Doping</td>
<td>1</td>
<td>3.2258</td>
</tr>
<tr>
<td>4</td>
<td>Metallization</td>
<td>2</td>
<td>6.4516</td>
</tr>
<tr>
<td>5</td>
<td>Miscellaneous</td>
<td>3</td>
<td>9.6774</td>
</tr>
<tr>
<td>6</td>
<td>Oxide Defect</td>
<td>8</td>
<td>25.8065</td>
</tr>
<tr>
<td>7</td>
<td>Silicon Defect</td>
<td>1</td>
<td>3.2258</td>
</tr>
</tbody>
</table>

Figure 25.3.  The Data Set FAILURE2 Created Using PROC FREQ

The following statements produce a Pareto chart for the data in FAILURE2:

```sas
title 'Analysis of IC Failures';
symbol value=dot;
proc pareto data=failure2;
  vbar cause / freq = count
                  scale = count
                  interbar = 1.0
                  last = 'Miscellaneous'
                  nlegend = 'Total Circuits'
                  cframenleg = empty;
run;
```

The chart is displayed in Figure 25.4.

Figure 25.4.  Pareto Chart with Frequency Scale
A slash (/) is used to separate the process variable CAUSE from the options specified in the VBAR statement. The frequency variable COUNT is specified with the FREQ= option. Specifying the keyword COUNT with the SCALE= option requests a frequency scale for the vertical axis.

The INTERBAR= option inserts a small space between the bars, and specifying LAST='Miscellaneous' causes the category Miscellaneous to be displayed last regardless of its frequency. The NLEGEND= option adds a sample size legend labeled Total Circuits, and the CFRAMENLEG= option frames the legend. The SYMBOL statement marks points on the curve with dots.

In the preceding statements, adding the keyword LINEPRINTER requests a line printer version of the chart, which is displayed in Figure 25.5.

![Pareto Chart for IC Failures](image)

**Figure 25.5.** Pareto Chart for IC Failures in the Data Set FAILURE2

*In Release 6.12 and previous releases of SAS/QC software, the keyword GRAPHICS was required in the PROC PARETO statement to specify that the chart be created with a graphics device. In Version 7, you can specify the LINEPRINTER option to request line printer plots.*
There are two sets of tied categories in this example; *Corrosion* and *Metallization* each occur twice, and *Doping* and *Silicon Defect* each occur once. The procedure displays tied categories in alphabetical order of their formatted values. Thus, *Corrosion* appears before *Metallization*, and *Doping* appears before *Silicon Defect* in Figure 25.4 and Figure 25.5. 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

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 25.3:

```sas
title 'IC Manufacturing Problems';
symbol value=dot;
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 25.6, does not display the categories *Doping* and *Silicon Defect*. 
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 the name with the LAST= option, the category is positioned at the far right of the chart. The following statements illustrate both options:

```sas
title 'IC Manufacturing Problems';
symbol value=dot;
proc pareto data=failure2;
   vbar cause / freq = count
       scale = count
       maxncat = 5
       other = 'Others'
       last = 'Others';
run;
```

The chart is shown in Figure 25.7.
The number of categories displayed is five, which is the number specified with 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, the procedure 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 “Restricted Pareto Charts” on page 922.

The next chapter describes the syntax of the PARETO procedure and provides examples of additional types of Pareto displays, including comparative Pareto charts.

**Figure 25.7.** Restricted Pareto Chart with Other Category

![Restricted Pareto Chart with Other Category](image-url)