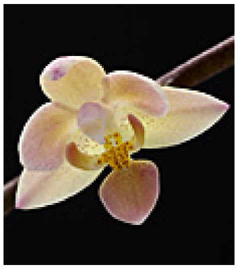
Pareto Charts [04-25]
Finding and Displaying Critical Categories

Introduction

Pareto Charts are a very simple way to graphically show a *priority* breakdown among categories along some dimension/measure of interest. Usually you are interested in some result or outcome among categories such as revenue, size, cost, or safety, for example. By categories, I mean any grouping you care to make, including things like services, products, personnel, zoo animals, or capital equipment. For example, you might have eight operational elements (departments) you want to rank with respect to sales. The eight operations would be the categories and the dimension/measure of interest would be their sales figures. You might also rank these same departments along the dimension of 'health', plotting the number of sick days/per-capita taken per department. You could even go further and plot the *ratios* of sales to sick days/per-capita per department, perhaps investigating sales versus stress? A Pareto plot of each of these situations would give you a graphical picture of each categories' relative importance. More importantly, you will have an external picture of what you're talking about - very helpful.



You can use these 'Pareto' ideas for any set of categories and their associated measures of interest. If you are dealing with groups of people in a city for example, you could categorize them any number of ways, say, country of origin. Then you could look at various measures on these country of origin categories such as: simple counts, median years of schooling, median age, per-capita family income, political affiliations, or any other characteristic you might come up with. For each of these measures you could make a Pareto chart showing the relative rankings of the categories.

If you are dealing with products or services, you might categorize them by revenue, cost, resource constraints, supply chain sequence, or a myriad others. The idea is to not only rank the categories with respect to whatever measure you are interested in, but display those rankings in a compelling manner, to invite the viewer to see what she couldn't see from the raw data or the raw conversation.

Cut to the Chase - Doing Pareto

Let's start with a simple 'Revenue versus Sales Category' data example, and go from there. (Details are illustrated in Example 1 below). What I want to find out is which wealth producing operational element is most important (with respect to revenue), which is second most important, which third, fourth and so on Usually, discovering the importance of a category means that I will direct more attention to it and a Pareto plot is designed to make it easy to discover relative importance.

To continue with this general discussion, assume that associated with each operational element is a revenue figure and its' calculated percentage contribution to the total revenue stream. I start with the largest revenue producer, calculate and plot its percentage contribution. Then I take the second most important sales category, calculate its percentage contribution and then add this percentage to the previous one. This gives me the cumulative percentage contribution of the first two categories. I do this, calculate and add, for each of the subsequent categories, ending up with a chart showing a cumulative percentage adding up to 100%. At each step, I can see from the chart the relative contribution of each operational element to the overall sales total.

Example 1: Revenue versus Operational Element Sales

Suppose I know that for a three month period, the revenue versus sales category figures look like the ones in the table below. I have already sorted them in revenue order, which is what you need to do if revenue is the driving factor. From the table, out of total revenues of $22,000, the Training Contract Category accounts for $6000, which is 27% of the total. Next in revenue importance is CD sales of XML tutorials. The revenue there is $5000 and accounts for $5000/$22000 fraction of revenues or, 22%. When I add the Training Contracts percentage to the CD sales revenue percentage, I get a cumulative total of 27 + 22 = 49%. The third most important revenue source is CDs of Java Graphics at $4000. Individually, this revenue accounts for $4000/$22000 * 100 = 18%, and, when added to the previous cumulative total, I am up to 67%. This means that the first three categories account for almost 70% of sales, and, this might provide some helpful insight on the management of the various operational elements. The graph also provides insight into combinations of revenue contributions as well.

Service/Product Sales Category	$Sales, 3-months	% of Total	Cumulative%
Training Contracts	6000	27	27
CDs of XML Tutorials	5000	22	49
CDs of Java Graphics	4000	18	67
DVDs of Java Tutorial	3000	13	80
CDs of Java Mobile	1000	5	85
Licensing	1000	5	90
Software maintenance	1000	5	95
Speaking Engagements	1000	5	100
Service/Product Sales Category	Total = 22,000		

You can see that a Pareto chart is a combination of a bar chart and a cumulative graph.

Out[10]=

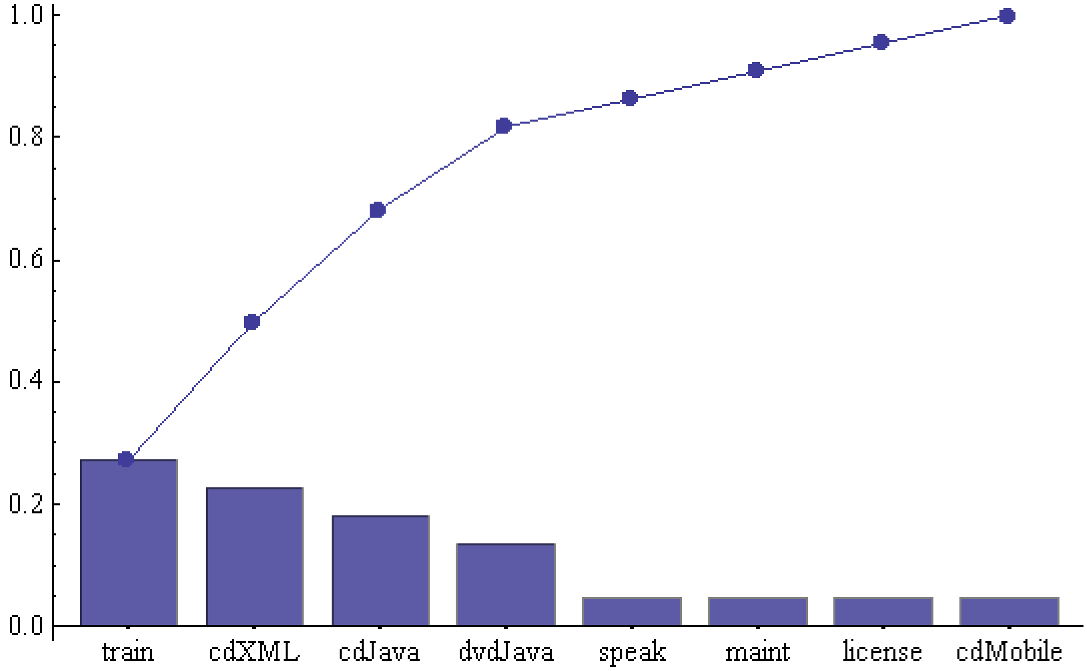


FIGURE 1. Pareto Chart of Revenue versus Sales Category

Pareto's 80-20 Rule (or maybe also the 10-90 Rule)?

It has been observed over a very long time that there seems to be a general rule that some 80% of effects roughly correlate with some 20% of the input or effort. This seems to hold true for large complex systems and less so for small deterministic-like organizations/systems. In general though, there is a lot of evidence for this kind of split taking place in many situations.

Vilfredo Pareto, an Italian economist of the 19th century, first published his analyses of income distributions among elite Florentine Italian families showing that some 80% of the wealth. was controlled by some 20% of the families (times have sure changed, since now its a 90-10 split!). From Pareto's work, a simplified analysis emerged, called "Pareto Analysis" that has been generalized to include all kinds of domains. This analysis gives a rough and ready first cut at the relative importance of a set of categories with respect to some property. Since the Pareto Plot shows all effects in a graded manner, there is actually no restriction to just talk about 80% or 20% levels, but that's a handy reference point.

For example, it seems to be generally the case that for large fuzzy/probabilistic environments we have:

- 20% of manufacturing processes account for some 80% of the defects (also, some 20% of the manufacturing processes often account for some 80% of the revenue, where the processes in both categories may or may not overlap!)
- 20% of the services account for 80% of the revenue
- 20% of the customers account for 80% of the revenue
- 20% of the employees do 80% of the work (or in your shop, maybe it's 10%-90%)!

- 10% of the pilots shoot down 90% of the enemy planes
- 10% of the energy categories provide 90% of the expended energy (e.g. fossil fuels versus all other energy sources)
- 10% of the rail routes provide 90% of the revenue
- 10% of the rail, highway, airline, or waterway routes carry 90% of the traffic for that mode.
- 20% of your employees/colleagues/family/bosses cause 80% of your difficulties!
- As engineers well know, getting to around 80% efficiency is feasible and maybe takes some 20% of the budget, but getting beyond a certain efficiency (with 100% as the target) for a process, takes extraordinary effort (in fact, this last effort is often not considered worthwhile for the ensuing gain)
- To eliminate those last few errors in a very large project (or in a very long document) takes extraordinary effort. That is, glaring errors are easy to fix, that is, the first 90% of the errors are easy to find and fix, but finding those subtle grammar and logic errors becomes progressively more difficult.
- Software bugs are notorious for remaining hidden for 90% of the life of an application and all of a sudden, a different sequence of commands causes a catastrophic error.

You can, no doubt, think of lots of examples from your own experience.

Example 2: Customer QOS: Bank Service Times

The following are 100 random customer waiting times for services at a metropolitan bank, in minutes. These reflect a measure of Quality Of Service (QOS) and are tracked by the bank's management. The sample was assumed to consist of observations of a random selection of customers and so was considered a *random sample*. I got these numbers out of a statistics book so you can consider them hypothetical but instructive. I have already sorted the times from low to high and will do a little Exploratory Data Analysis (EDA) on these values before I *Pareto* them. This is a good chance to use several of these EDA techniques such as: a Stem&Leaf plot, and a Box & Whisker Plot!

Be aware that a Pareto plot is based on categories and a measure of each category. For this bank situation the categories are the specific time intervals and the measure is the number of waiting times falling into each of these intervals. That is, the fact that most customers waited between 4 and 5 minutes while the next most frequent time was between 3 and 4 minutes and so on, may not be what you are after, but that's what Pareto does. In the Pareto approach, the categories *are specific intervals of waiting times* while the dimension of interest is the *count* of times within this interval.

Note: I am illustrating here the *Pareto* approach but, just so you'll know, if what you *really* want is a 'cumulative' distribution of times so that you can tell what percentage of customers wait less than 1 minute, what percentage wait less than 2 minutes, what percentage wait less than, say 8 minutes, and so on, then I have a graph of this situation following the Pareto chart.

```
{0.4, 0.8, 1.1, 1.3, 1.4, 1.6, 1.8, 1.8, 2., 2.2, 2.3, 2.4,
 2.5, 2.7, 2.8, 2.9, 2.9, 3.1, 3.2, 3.4, 3.5, 3.6, 3.7,
 3.7, 3.8, 3.8, 3.9, 3.9, 4., 4., 4.1, 4.2, 4.3, 4.3, 4.3,
 4.4, 4.4, 4.5, 4.5, 4.5, 4.6, 4.7, 4.7, 4.8, 4.9, 5.,
 5.1, 5.1, 5.2, 5.2, 5.3, 5.4, 5.4, 5.5, 5.6, 5.6, 5.7,
 5.8, 5.8, 5.8, 6.1, 6.1, 6.2, 6.3, 6.3, 6.3, 6.4, 6.5,
 6.5, 6.6, 6.7, 6.7, 6.8, 7., 7.2, 7.2, 7.3, 7.4, 7.4,
 7.5, 7.7, 7.8, 7.9, 8., 8.1, 8.3, 8.4, 8.6, 8.6, 8.7,
 9.1, 9.2, 9.3, 9.5, 9.8, 9.9, 10.2, 10.7, 10.9, 11.6}
```

FIGURE 2. 100 Customer Waiting Times in Bank Lines (in minutes)

A Stem&Leaf Plot of the Bank Waiting Times - Showing the Distribution of Wait Times

The graph below is another of John Tukey’s inventions, called a stem and leaf plot. The diagram below is interpreted as follows: the first line indicates that there were two wait times of 0.4 and 0.8 minutes. The second line show 6 waiting times with values of 1.1, 1.3, 1.4, 1.6, 1.8, 1.8 minutes. The third line has 9 leaves: 2.0, 2.2, 2.3, . . . The last line shows the longest wait time of 11.6 minutes. The stem and leaf plot uses a ‘stem’, such as ‘1’ and appends digits to it. So, the second line has a stem of ‘1’ to which is appended ‘leaves of 1,3, 4, 6, 8, and interpreted as noted above. The diagram shows the waiting times distribution.

As a side note, the example of wait times with its stem and leaf plot reveals what looks like an approximation to a Normal Curve. Another phase of analysis might explore this possibility?

```
Stem | Leaves             Counts
---------------------------------
   0 | 48                    2
   1 | 134688                6
   2 | 023457899             9
   3 | 12456778899          11
   4 | 00123334455567789    17
   5 | 011223445667888      15
   6 | 1123334556778        13
   7 | 0223445789           10
   8 | 0134667               7
   9 | 123589                6
  10 | 279                   3
  11 | 6                     1
Stem units: 1
```

FIGURE 3. Stem & Leaf Plot of Bank Customer Waiting Times

A Box & Whisker Plot of the Bank Waiting Times

The same 100 values are now condensed into another type of display called a Box and Whisker

Plot. The mid line of the box is the median (this is the value that divides the data set in half, also called Q2, the 50% quartile), the upper and lower crossbars are at the 25% (Q1) and 75%(Q3) quartiles. The bottom whisker extends down from the lower quartile (3.8) to the smallest point that is not an outlier. In this case that will turn out to be the value 0.4. The upper whisker extends from the upper quartile (7.2) to the largest value that is not an outlier, which in this case is 11.6.

(See the tutorial on this site called *Five Number Summaries and Box & Whisker Plots*). A five number summary is used to produce the box and whisker plot, if done manually.The diagram below was done using *Mathematica* 6.0.

Median = 5.25 (half the wait times are less than this, and half are more)

lower quartile = 3.8 (also called the lower hinge - 1/4 of the data values are less or equal than this)

upper quartile = 7.2 (also called the upper hinge - 1/4 of the data values are equal to or greater than this)

adjacent points = {0.4 and 11.6} (these are the smallest and largest values of the data set that are not outliers)

For this data set of 100 values, the values of the waiting times lie within reasonable bounds, so, in other words, this set had no detectable outliers. Even the smallest and largest values in the set, 0.4 and 11.6, were not incompatible with the bulk of the data set (which is all we have to go on at this stage of an investigation).

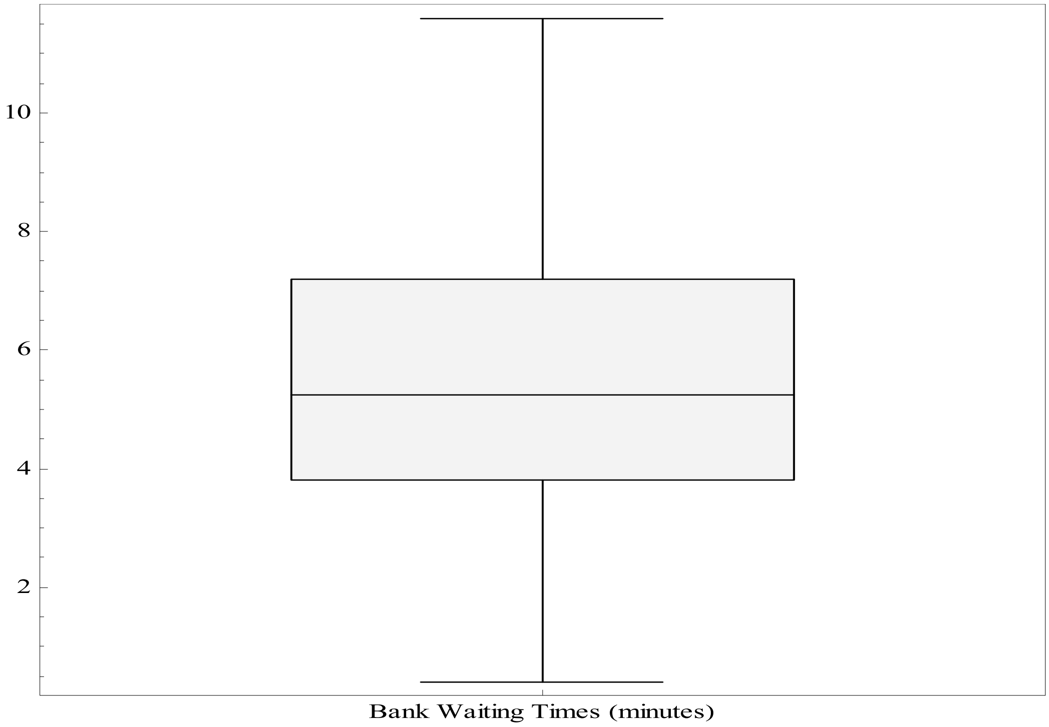


FIGURE 4. Box & Whisker Plot of Bank Waiting Times (minutes)

Pareto Plot of Bank Waiting Times (but be careful what you wish for)!

When I said that the Pareto Plot would show how many values were in each 'category' under consideration you need to be sure these categories are what you are interested in. The plot below takes its' 'categories' as each unit time interval and counts the number of values *for that time interval.*

So, the plot below simply tallies up the number of values lying between 4 and 4.9, for example, and plots that number, 17, as a percentage of the total. That is the first 'bar' of the chart below. This plot simply shows you the frequency of 1 minute time intervals. You could also see this from the stem and leaf plot more clearly perhaps. So this Pareto plot says that the wait times between 4 minutes and 4.9 minutes are most frequent, followed by times between 5 minutes and 5.9, followed by 6- 6.9, and then 3-3.9, and so on. This is interesting, but may not be what you are after. Maybe you want the cumulative percentage of customers who wait, say less than 8 minutes for service? That is shown in the chart after this one.

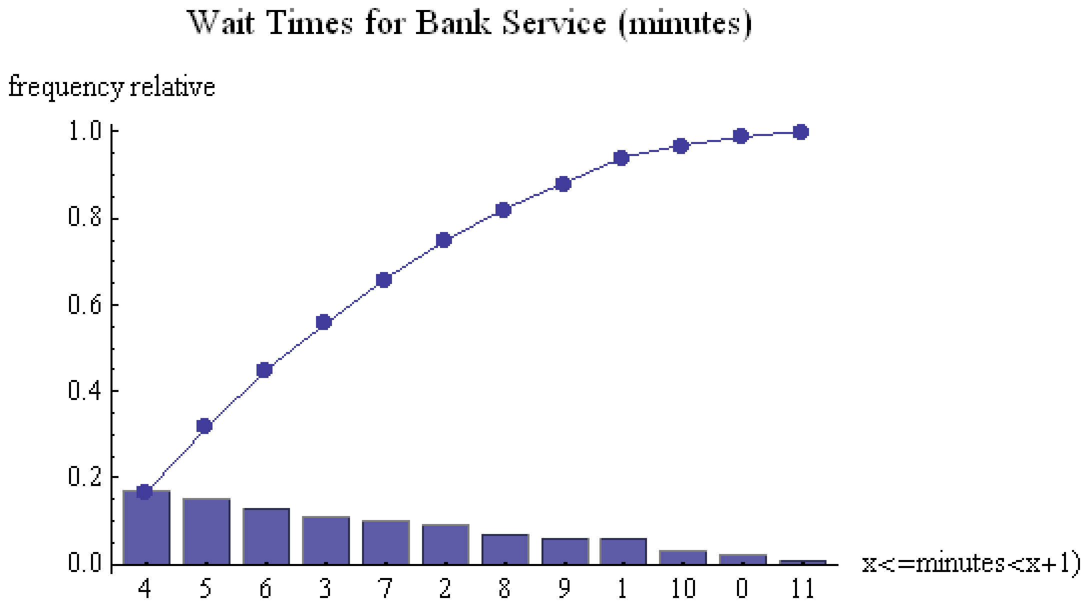


FIGURE 5. Pareto Chart of Bank Customer Waiting Times (individual time frequencies)

Alternatively: What Percentage of Customers Wait Less Than 8 Minutes?

If I want to know the percentage of customer waiting less than, say, 8 minutes, that is, starting from 0 *up to* 8 minutes, I need another chart. You can easily construct the table below from the Stem & Leaf chart, Figure 3 on page 5.

Wait Time Intervals	Interval Count	Percent of Total	Cumulative%
0 ≤ t <1	2	2	2
1≤ t <2	6	6	10
2 ≤ t < 3	7	7	17
3 ≤ t < 4	11	11	28
4 ≤ t < 5	17	17	45
5 ≤ t < 6	15	15	60
6 ≤ t < 7	13	13	73
7 ≤ t < 8	10	10	83
Wait Time Intervals	Interval Count	Percent of Total	Cumulative%

Wait Time Intervals	Interval Count	Percent of Total	Cumulative%
$8 \leq t < 9$	7	7	90
$9 \leq t < 10$	6	6	96
$10 \leq t < 11$	3	3	99
$11 \leq t < 12$	1	1	100
Wait Time Intervals	Interval Count	Percent of Total	Cumulative%

Interval Counts (also called 'bin' counts)

Here are the waiting time intervals I am considering, as shown in the table above: $0 \leq t < 1$, $1 \leq t <2$, $2 \leq t < 3$, . . ., $11\leq t <12$

For example, for wait times starting at 0 and less that 1 minute, I have 2 such times, 0.4, and 0.8. For the number of wait times starting at 1 minute and extending to 1.9, I find 6 such values, and, for the number of wait times between 4 minutes and 4.9, inclusive, I see 17.

Here are those counts for each interval: {2, 6, 9, 11, 17, 15, 13, 10, 7, 6, 3, 1}, also shown in the table above.

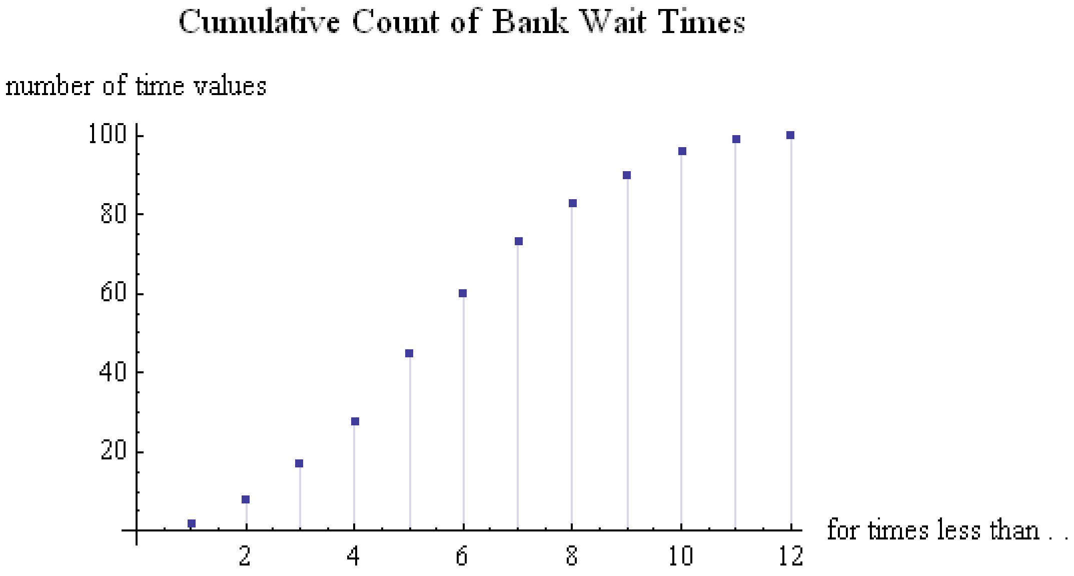
Cumulative Interval Counts

Now I accumulate these bin counts so I can find the cumulative number of wait times, say, less than 1 minute. That matches the number '2' below. If I want the number of wait times less that 2 minutes I see 8 = 2+ 6. For the cumulative number of wait times less than 3 minutes I see 17= 2 + 8 + 7, and so on. This is also shown in the table above.

{2, 8, 17, 28, 45, 60, 73, 83, 90, 96, 99, 100}

Plotting the Cumulative Counts of Wait Times

If you want to know what percentage of customers experience wait times less than, say, 6 minutes, check out the chart below and locate 6 on the horizontal axis, go up to the dot and then see where that dot lines up on the vertical axis. Looks like 60% to me (actually I cheated and looked at the cumulative table, but I could have used the drawing!). For wait times less that 8 minutes, I read off about 85%. That is 85% of my customers experience a waiting time of less than 8 minutes. (The actual value is 83% but sometimes a graph is easier to present).



Example 2: A Generic Pareto Plot

The example below shows how you can *Pareto* most anything. What happens is that the software package (or you!) totals up effects from each category, orders the categories by number of effects, computes a percentage of the total, and plots those cumulative percentages. To show how general this approach is, here is a set of letters of the alphabet. The categories could simply be *similar* letters, and the dimension/effect of interest is their count. That is, the effect/measure is: how *many* a's, b's, and so on are there?

letters = {d, d, d, d, e, e, e, b b, c, c, a, a, f}.

So, there are 14 letters of which 4 are 'd's, 3 'e', 2 'b', 2 'c', 2 'a', and 1 'f'. So, each of the counts are scaled by '14' to get a relative frequency and then plotted. The category of 'd' accounts for 4/ 14 * 100% = 29%. Next comes the category of 'e' of which there are 3, for a percentage of 3/14 * 100 = 21%. Then comes the 'b' category with 2 entries for a percentage of 2/14 * 100. = 14%. If I total the first three categories, I would have a cumulative percentage of 29 + 21 + 14 = 65%. This tells me that the first three categories account for some 64% of the total letter count.

```
In[2]:= ParetoPlot[{a, b, c, d, d, d, e, d, e, e, f, a, b, c}]
```

Out[2]=

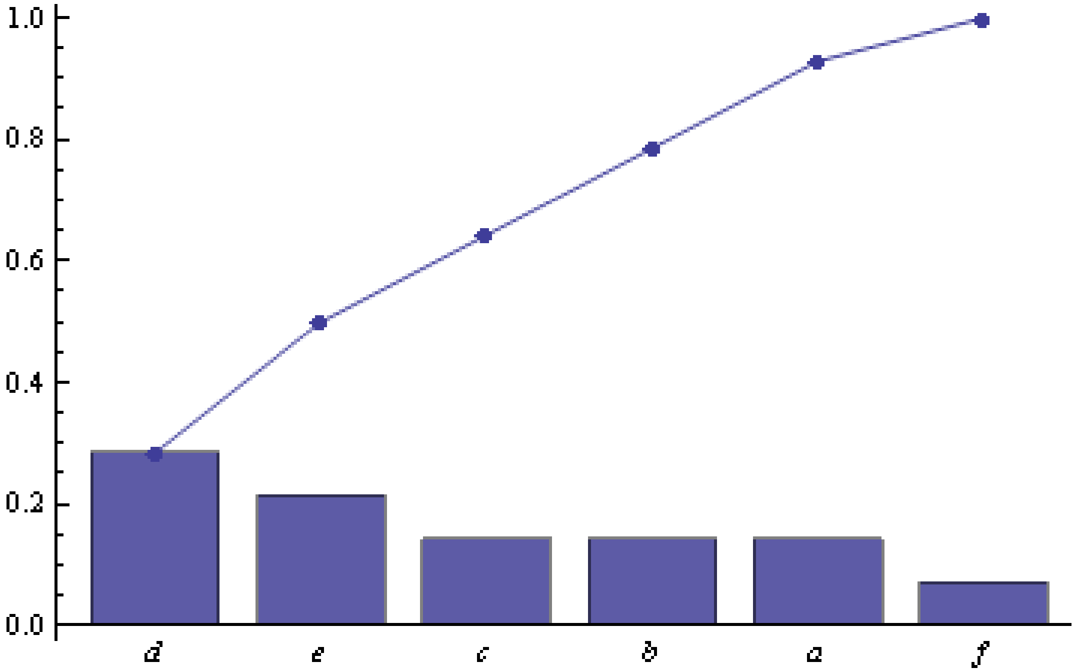


FIGURE 6. I Can Plot Anything I Can Categorize (along some dimension)!

Summary

If you want to show yourself and others the priorities of effects of some process, and you can categorize those effects, a Pareto chart might help. The Pareto chart shows the number of effects of the most important category first, then the next most important, and so on. Since the cumulative effects are also plotted, it is easy to see which categories contribute the most to the effect of interest.

References

Rucker, R.(2007) *Introduction to EDA*, available at web site: milagrosoft.com

Rucker, R.(2007) *Five Number Summaries Box Plots*, available at web site: milagrosoft.com

Tukey, John (1977) *Exploratory Data Analysis*, Addison Wesley