Print

Pareto Plots

In these next videos, we turn our attention to graphing categorical data.

You've learned that bar charts are an efficient tool for visualizing the frequencies of categorical data.

Take, for example, the Chemical Manufacturing scenario. The categorical variable, Performance, has two values, Accept and Reject.

Rejected batches did not have a yield > 80%. Approximately 15.6% of the batches were rejected.

Bar charts can also be used for exploratory data analysis.

When you look at the distribution of Performance across the other categorical variables, you can see that Performance seems to be related to Vessel Size, Amine Supplier, and Base Particle Size.

You can also use bar charts to identify the most commonly occurring problems or issues. For example, let's say that you're studying engine defects. Here, the most common defect is End play, followed by Ignition.

To make it easier to identify the largest defect categories, you can sort the data in descending order of the frequency of occurrence. Now, you can see that the top three issues are End play, Ignition, and Gas line.

Sorting is particularly useful when you have many levels. For example, here there are 21 engine defect categories.

Can you easily identify the top three or four issues using this bar chart?

By sorting the data in descending order of the frequency of occurrence, you can now easily identify the top issues.

You can also see that there are many defects that don't occur very often.

Instead of using a bar chart to graph these data, you can use a Pareto plot.

A Pareto plot is a sorted bar chart that also displays a curve for the cumulative frequency or cumulative percent. This curve helps you identify the top few issues that account for the majority of the problems.

The Pareto plot, or Pareto chart, is based on the Pareto principle, also known as the 80/20 rule. This rule states that approximately 80% of the consequences result from 20% of the causes.

This Pareto plot is a little hard to interpret because there are so many defect categories that don't occur very often. The defects with the highest counts don't stand out.

In this graph, we've combined the smallest 12 categories. They are now grouped together in an "Others" bar. It is now easier to see the top defects. Note that we could have combined a different number of categories. This was an arbitrary decision based on these data and the story we are trying to tell with the data.

Here, we've labeled the cumulative percent curve.

You can see that the top five issues account for more than 80% of the total engine defects, and that the top four issues account for over 75%. If you are on a team that is addressing engine defects, this gives you a good place to start.

In this example, we've ignored the fact that some defects might be costlier than others, or might be weighted as more important by the business or the customer. This information can also be included in the analysis to guide your decision making.

Let's say you are studying scrapped parts.

You have data, collected over a three-month period, about scrapped parts. During this time, 811 batches had one or more scrapped parts.

The data table includes the following information for these 811 batches: the number of pieces scrapped per batch, the product line, the product family, and the total value of the scrapped parts per batch. The data are in the file Scrapped Parts.jmp.

This sorted bar chart shows that 54% of the batches with scrapped parts were from product lines A2 and A3.

The third highest is B2, at 9.9%.

When you add the number of parts scrapped per batch, product lines A2 and A3 account for 62% of the parts scrapped, and B2 now accounts for 14%. Here, the same ordering is used, so you can see how the values change.

How does this picture change when you use the total value of the scrapped parts?

In terms of value, product lines A2 and A3 are now much lower in comparison to B2, which accounts for 31% of the total value of the scrapped parts.

These data can be explored from different angles, and you revisit this scenario in a practice.

This example illustrates that the biggest problem isn't necessarily the problem that should be addressed first. You need to consider the cost of the problem and the importance to the business and the customer.

Also, you should keep in mind that some problems might be much easier to tackle than others. As with all analyses, your knowledge of the product, the process, and the business environment, along with what the data say, should guide your decision making.

In this video, we revisited bar charts for graphing categorical data. You learned that sorted bar charts can make it easier to identify the most frequently occurring categories. You also learned about Pareto plots. Pareto plots are a special type of bar chart for identifying the top issues or opportunities.

In future videos, you revisit mosaic plots, for graphing two categorical variables at a time, and you learn about additional tools for graphing categorical data with many levels (packed bar charts and tree maps).

Close