Print

Describing Categorical Data

In previous videos, you learned about methods for describing the shape, centering, and spread of continuous variables. For categorical variables, the descriptive measures are largely based on frequencies.

Consider the Impurity Logistic data. The Outcome variable has two categories, Pass or Fail, based on whether the specification of 7% for the polymer has been met.

Out of 100 batches, 26 of the batches failed to meet the specification, and 74 of the batches passed.

This information is reported in the Frequency Distribution table and is displayed graphically in the bar chart. We can see the information in the bar chart more clearly here.

We've separated the bars, added percents in each category, and added a count axis. The height of the bars corresponds to the frequency of occurrence of each of the categories in the data table.

You can use linked bar charts to explore potential relationships between variables. Here, we've added the histograms for Temp, Catalyst Conc, and Reaction Time, and have selected the bar for Fail in the bar chart for Outcome.

You can see that batches that failed tend to have high values for Temperature and high values for Catalyst Conc.

The Impurity data set also includes information on the production shift and the reactor that was used. Here, you can see that a majority of the failed batches were produced using reactor 3, but approximately the same number of failed batches were produced on each shift.

As this example illustrates, you can use linked bar charts and histograms to explore potential relationships between variables.

When you want to look at many variables at a time, you can use tabular summaries of the data. In this example, you see the breakdown of Fail and Pass across the different reactors. Here, two statistics were used: N and Column %.

N is the number of observations for the particular outcome for the different reactors. For example, in this data set, 1 fail was produced with reactor 1, 4 fails were produced with reactor 2, and 21 fails were produced with reactor 3.

Likewise, the Column % statistic breaks down each outcome by Reactor. You can see that 80.77% of the failed batches were produced using reactor 3.

Here's another way of summarizing the same data. Instead of using Column %, the Row % is shown. This breaks down the reactors for the different outcomes, so you can see that, of the batches produced using reactor 3, 63.64% were fails and 36.36% were passes.

A contingency table is an efficient way to summarize all of this information in one table. In this contingency table, you see the counts (or N), the column percent, and the row percent.

You also see cumulative totals for both variables, the percent for each cell out of the total, and the percent for each level of a variable out of the total.

Let's turn our attention now to graphical displays for categorical data. You can graphically describe the relationship between two categorical variables in a number of ways. For example, side-by-side and stacked bar charts are efficient graphs for showing the distribution of the reactors for the two outcomes.

In both of the graphs shown here, the height of the colored bars is the count of the outcome for the given reactor. You can see that there are more passes than fails, that more of the failed batches were from reactor 3, and that more of the passes were from reactors 1 and 2.

Here's another stacked bar chart, where the roles for Outcome and Reactor are reversed. This display makes it easy to see that very few of the failed batches were produced using reactor 1 and a majority of the failed batches were from reactor 3.

A final display of the data that we'll discuss is the mosaic plot. A mosaic plot is similar to a stacked bar chart. In this mosaic plot, the vertical bars show the percent of fails and passes for each of the three reactors.

The horizontal widths of the bars show the percent of the batches from each of the reactors. There's about the same number of observations from each reactor, so the bars are approximately the same width.

You can add counts or percents to all of these graphs to help interpret what the graph is telling you and to better communicate the story in your data. The percents are the row percents reported in the contingency table you saw earlier.

As you can see, there are many ways of summarizing and graphing categorical data. You learned some core tools in this video and will see other methods for visualizing and exploring categorical data later in this module. You'll also learn some best practices for creating effective visualizations and for sharing your results with others.

In the next videos, you see how to create tabular summaries for categorical data, bar charts, mosaic plots, and contingency tables in JMP.

Close