What is a mosaic plot?

A mosaic plot is a graphical display that allows you to examine the relationship among two or more **categorical** variables.

The mosaic plot starts as a square with length one. The square is divided first into horizontal bars whose widths are proportional to the probabilities associated with the first categorical variable. Then each bar is split vertically into bars that are proportional to the conditional probabilities of the second categorical variable. Additional splits can be made if wanted using a third, fourth variable, etc.

Example: The mortality rates aboard the Titanic, which are influenced strongly by age, sex, and passenger class. If you wanted to compare the mortality rates between men and women using a mosaic plot, you would first divide the unit square according to the overall proportion of males and females.

![Mosaic Plot Example](image)

Roughly 35% of the passengers were female, so the first split of the mosaic plot is 35/65.

Next, split each bar vertically according to the proportion who lived and died.

![Mosaic Plot Example](image)

Among females, 67% survived (coded as 1 on this plot) and 33% died (coded as 0). So the female bar shows a 67/33 split. Among males, only 17% survived, so this bar shows a 17/83 split.
Most implementations of the mosaic plot offer as a default a small margin around each cell to make the graph easier to read.

This plot shows you that males were the majority of the deaths and the minority of the survivors. As a general recommendation, variables that represent an exposure or treatment status should usually represent the first split and variables that represent an outcome should represent the second split.

**Practice Analyzing a Mosaic Plot**

Here is a mosaic plot looking at the relationship between passenger class (1\textsuperscript{st}, 2\textsuperscript{nd}, or 3\textsuperscript{rd}) and mortality (1 survived and 0 did not survive).

1. Based on the chart, what categories have relatively similar areas?
2. Based upon the chart, does the class of the passenger affect the likelihood they would survive? How do you tell?
Practice Making a Mosaic Plot

In Michael Friendly’s paper, “A Brief History of Mosaic Plots” (2001), he provides the following description: In statistical graphics, the mosaic display, attributed to Hartigan and Kleiner (1981), is a graphical method to show the values (cell frequencies) in a contingency table cross-classified by one or more “factors”.

A contingency table is simply a table that displays a count (frequency) in each cell that resides at the column and row intersections of two or more categorical variables. Consider a group of individuals for whom data was collected regarding two variables: hair color (black, brown, red, and blond) and eye color (brown, blue, hazel, and green). (This example comes from Friendly’s paper.) Below is a sample two-way (as in two categorical variables) contingency table of this data:

Using this table, we can analyze the question, is there a relationship between hair color and having blue eyes? In order to do this, we need to further break down the data to make a contingency table that isolates our variables.

3. Using the data table above, condense the data to complete the following contingency table.

<table>
<thead>
<tr>
<th>Eye Color</th>
<th>Black</th>
<th>Brown</th>
<th>Red</th>
<th>Blonde</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Not Blue</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>113</td>
</tr>
</tbody>
</table>

4. Using the contingency table, use a piece of graph paper to make a mosaic plot for the data.

5. Based upon the mosaic plot, is there a relationship between hair color and eye color?
Analyzing the Big Picture

The widths of the rectangles represent the proportion of people with each hair color and their heights represent the proportion of people with each eye color within each hair color group. The area of each rectangle is proportional to the frequency of each combined eye color and hair color group. In other words, the areas represent the numbers in the body of the contingency table.

To create a mosaic plot, you begin with a large rectangle and divide it into vertical sections based on the first categorical variable, in this case hair color, and then you add a little space between the sections.

You then divide it into horizontal sections based on the second variable, in this case eye color, and once again add some space between them.

Spaces between the sections are conventional, but not necessary. When these spaces are omitted, the graph is sometimes called a Mondrian diagram.

6. Using the Mosaic or Mondrian diagram above, describe any relationships you can infer.
Going Further …

In this example, we have a mosaic plot and the three-way contingency table on which it is based. Its three categorical variables are Salk polio vaccine (either administered to the patient or not), paralysis (either happened to the patient or not), and age of the patient in years (0-4, 5-9, 10-14, 15-19, 20-39, and 40+). The colors are there to merely reinforce the distinction between paralysis (orange) and no paralysis (green).

![Mosaic plot](image)

We wish to learn if symptom status (paralysis or not) is independent of vaccination status after controlling for age. The goal of the study is to show that vaccination reduces paralysis.

7. Examine the age based data for children aged 5 – 9. Compare the area of children who did not get vaccinated and became paralyzed to the area of the group of children who did get vaccinated and did not get paralyzed. What type of relationship can you determine from area?

8. Examine the data as a whole. Is there a relationship between getting the polio vaccination and paralysis? How can you tell?

References

https://medschool.vanderbilt.edu/cqs/files/cqs/media/DrTsai2_0.pdf
https://www.perceptualedge.com/articles/visual_business_intelligence/are_mosaic_plots_worthwhile.pdf