Graphing Qualitative Variables

Prerequisites

Variables

When Apple Computer introduced the iMac computer in August 1998, the company wanted to learn whether the iMac was expanding Apple's market share. Was the iMac just attracting previous Macintosh owners? Or was it purchased by newcomers to the computer market, and by previous Windows users who were switching over? To find out, 500 iMac customers were interviewed. Each customer was categorized as a previous Macintosh owners, a previous Windows owner, or a new computer purchaser. This section examines graphical methods for displaying the results of the interviews. We'll learn some general lessons about how to graph data that fall into a small number of categories. A later section will consider how to graph numerical data in which each observation is represented by a number in some range. The key point about the qualitative data that occupy us in the present section is that they do not come with a pre-established ordering (the way numbers are ordered). For example, there is no natural sense in which the category of previous Windows users comes before or after the category of previous iMac users. This situation may be contrasted with quantitative data, such as a person's weight. People of one weight are naturally ordered with respect to people of a different weight.

Frequency Tables

All of the graphical methods shown in this section are derived from frequency tables. Table 1 shows a frequency table for the results of the iMac study; it shows the frequencies of the various response categories. It also shows the relative frequencies, which are the proportion of responses in each category. For example, the relative frequency for "none" of 0.17 = 85/500.

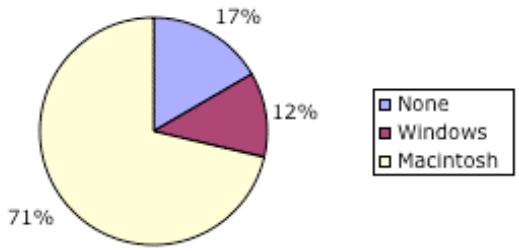
Table 1. Frequency Table for the iMac Data.

Previous Ownership	Frequency	Relative Frequency
None	85	0.17
Windows	60	0.12
Macintosh	355	0.71
Total	500	1.00

Pie Charts

The pie chart in Figure 1 shows the results of the iMac study. In a pie chart, each category is represented by a slice of the pie. The area of the slice is proportional to the percentage of responses in the category. This is simply the relative frequency multiplied by 100. Although most iMac purchasers were Macintosh owners, Apple was encouraged by the 12% of purchasers who were former Windows users, and by the 17% of purchasers who were buying a computer for the first time.

Figure 1. Pie chart of iMac purchases illustrating frequencies of previous computer ownership.



Pie charts are effective for displaying the relative frequencies of a small number of categories. They are not recommended, however, when you have a large number of categories. Pie charts can also be confusing when they are used to compare the outcomes of two different surveys or experiments. In an influential book on the use of graphs, Edward Tufte asserted "The only worse design than a pie chart is several of them."

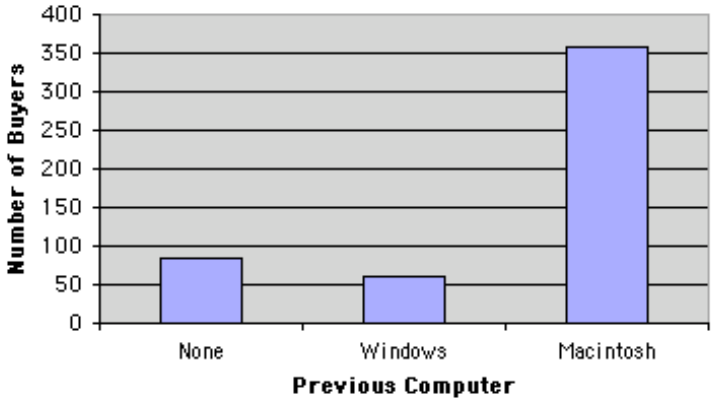
Here is another important point about pie charts. If they are based on a small number of observations, it can be misleading to label the pie slices with percentages. For example, if just 5 people had been interviewed by Apple Computers, and 3 were former Windows users, it would be misleading to display a pie chart with the Windows slice showing 60%. With so few people interviewed, such a large percentage of Windows users might easily have accord since chance can cause large errors with small samples. In this case, it is better to alert the user of the pie chart to the actual numbers involved. The slices should therefore be labeled with the actual frequencies observed (e.g., 3)

instead of with percentages.

Bar charts

Bar charts can also be used to represent frequencies of different categories. A bar chart of the iMac purchases is shown in Figure 2. Frequencies are shown on the Y axis and the type of computer previously owned is shown on the X axis. Typically the Y-axis shows the number of observations rather than the percentage of observations in each category as is typical in pie charts.

Figure 2. Bar chart of iMac purchases as a function of previous computer ownership.



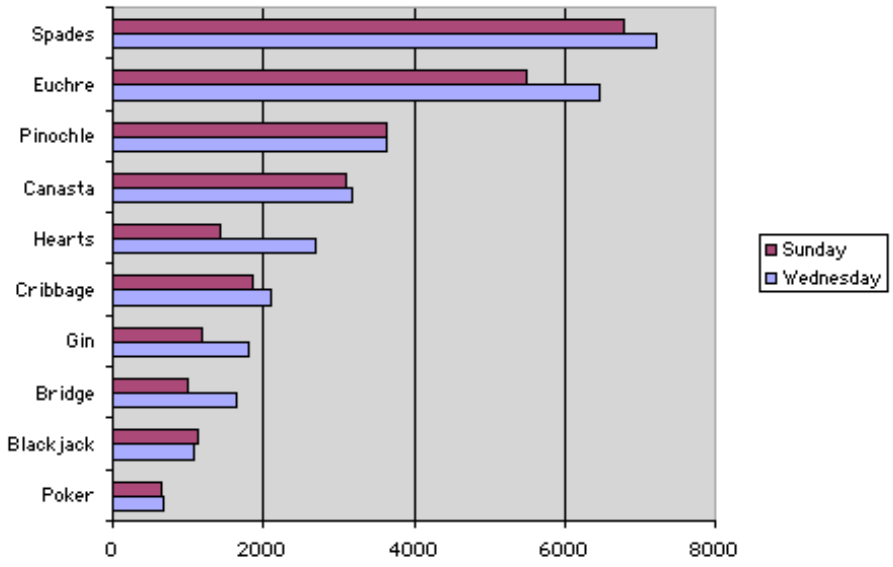
EIA guidelines for statistical graphs.

Comparing Distributions

Often we need to compare the results of different surveys, or of different conditions within the same overall survey. In this case, we are comparing the "distributions" of responses between the surveys or conditions. Bar charts are often excellent for illustrating differences between two distributions. Figure 3 shows the number of people playing card games at the Yahoo web site on a Sunday and on a Wednesday on a day in the Spring of 2001. We see that there were more players overall on Wednesday compared to Sunday. The number of people playing Pinochle was nonetheless the same on these two days. In contrast, there were about twice as many people playing hearts on Wednesday as on Sunday. Facts like these emerge clearly from a well-designed bar chart.

Figure 3. A bar chart of the number of people playing

different card games on Sunday and Wednesday.

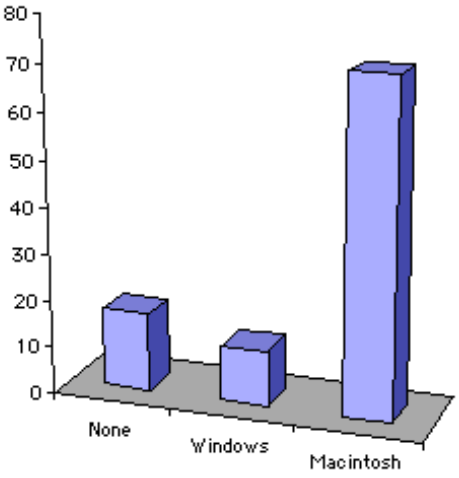


The bars in Figure 3 are oriented horizontally rather than vertically. The horizontal format is useful when you have many categories because there is more room for the category labels. We'll have more to say about bar charts when we consider numerical quantities later in the section Bar Charts.)

Some graphical mistakes to avoid

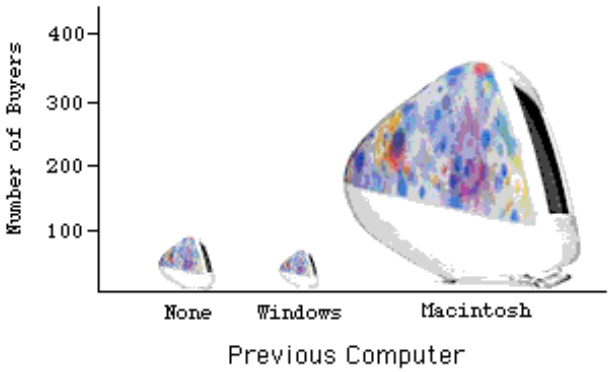
Don't get fancy! People sometimes add features to graphs that don't help to convey their information. For example, 3-dimensional bar charts like the one shown in Figure 4 are usually not as effective as their two-dimensional counterparts.

Figure 4. A three-dimensional version of Figure 2.



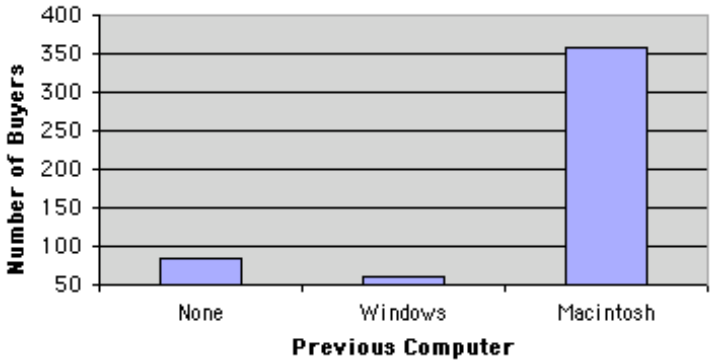
Here is another way that fanciness can lead to trouble. Instead of plain bars, it is tempting to substitute meaningful images. For example, Figure 6 presents the iMac data using pictures of computers. The heights of the pictures accurately represent the number of buyers, yet Figure 6 is misleading because the viewer's attention will be captured by areas. This can exaggerate the size differences between the groups. In terms of percentages, the ratio of previous Macintosh owners to previous Windows owners is about 6 to 1. But the ratio of the two areas in Figure 6 is about 35 to 1. A biased person wishing to hide the fact that many Windows owners purchased iMacs would be tempted to use Figure 6 instead of Figure 2! Edward Tufte coined the term "lie factor" to refer to the ratio of the size of the effect shown in a graph to the size of the effect shown in the data. He suggests that lie factors greater than 1.05 or less than 0.95 produce unacceptable distortion.

Figure 6. A redrawing of Figure 2 with a lie factor greater than 8.



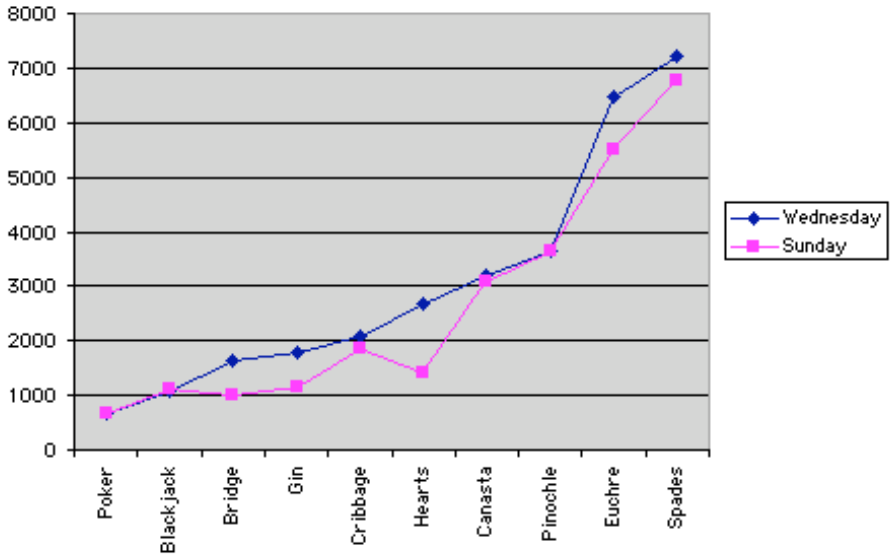
Another distortion in bar charts results from setting the baseline to a value other than zero. The baseline is the bottom of the Y-axis, representing the least number of cases that could have occurred in a category. Normally, this number should be zero. Figure 7 shows the iMac data with a baseline of 50. Once again, the difference in areas suggests a different story than the true differences in percentages. The number of Windows-switchers seems minuscule compared to its true value of 12%.

Figure 7. A redrawing of Figure 2 with a baseline of 50.



Finally, we note that it is a serious mistake to use a line graph when the X-axis contains merely qualitative variables. A line graph is essentially a bar graph with the tops of the bars represented by points joined by lines (the rest of the bar is suppressed). Figure 5 inappropriately shows a line graph of the card game data from Yahoo. The drawback to Figure 5 is that it gives the false impression that the games are naturally ordered in a numerical way.

Figure 5. A line graph of the number of people playing different card games on Sunday and Wednesday.



Summary

Pie charts and bar charts can both be effective methods of portraying qualitative data. Bar charts are better when there are more than just a few categories and for comparing two or more distributions. Be careful to avoid creating misleading graphs.