Unit 3: Histograms

SUMMARY OF VIDEO

Many people are afraid of getting hit by lightning. And while getting hit by lightning is against the odds, it is not against all odds. Hundreds of people are struck by lightning every year in the U.S. What's more, fires started by lightning strikes cause hundreds of millions of dollars of property damage. Meteorologist Raul Lopez and his associates began collecting detailed data on lightning strikes back in the 1980s and soon were overwhelmed by the vast amount of data. In one year, they collected three-quarters of a million flashes in a small area of Colorado. They decided to focus on when lightning strikes occurred. The data on the times of the first lightning strike needed to be organized, summarized, and displayed graphically. One of the statistical tools that Raul Lopez turned to was the graphic display called a histogram. For example, data on the percent of first lightning flashes for each hour of the day is displayed in the histogram in Figure 3.1.

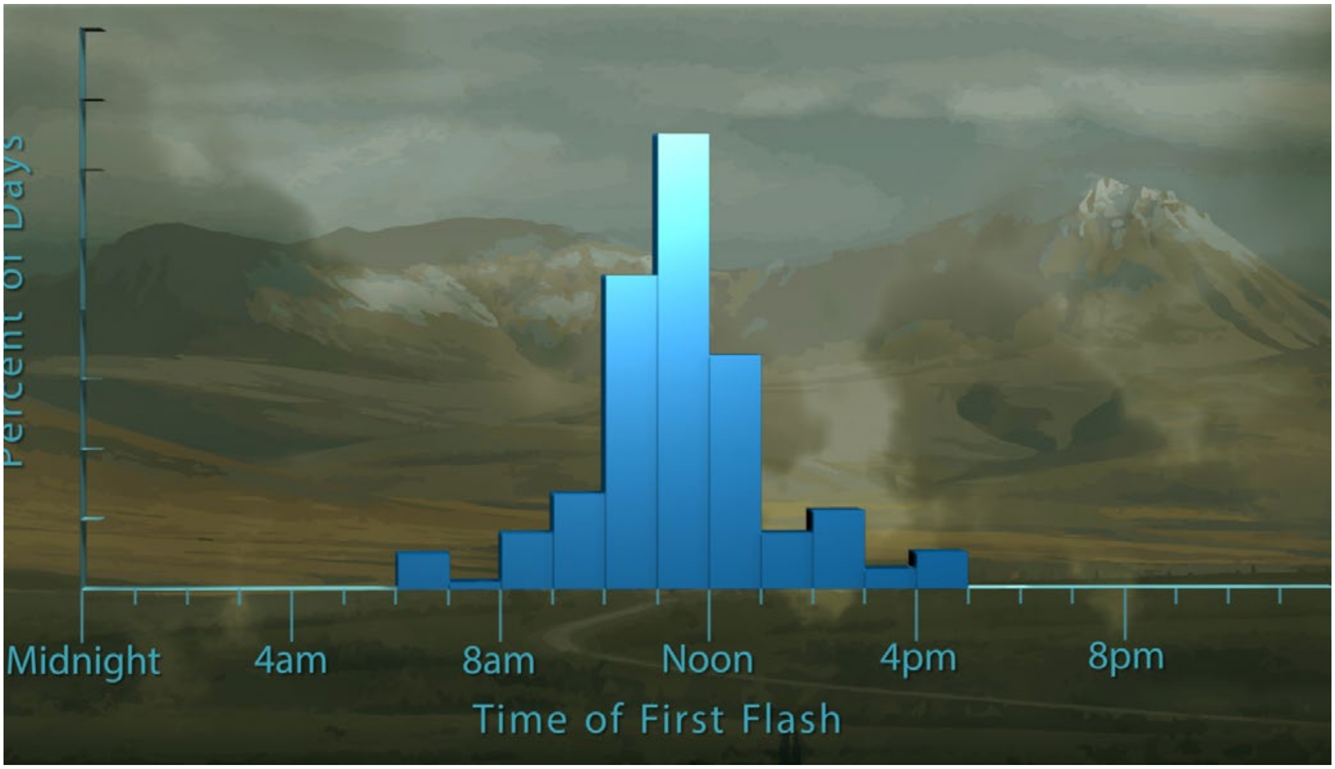

Figure 3.1. Histogram of the time of the first lightning strike.

Before the histogram could be constructed, each day was broken into hours (horizontal axis), the number of first flashes in each hour was counted, and then the counts were converted to percentages (vertical axis). So, in this histogram, each bar represents one hour, and its height is the percentage of days in which the first lightning flash fell in that hour. This histogram has two very striking features. First, it is roughly symmetric about the tallest bar, which represents the percentage of first flashes between 11 a.m. and noon. The second rather surprising feature is how tightly the time of first strike clusters around the center bar, with a range from 10 a.m. to 1 p.m. accounting for most of the days' first strikes. And there are no first strikes at night. This pattern helped explain how lightning storms form in this area. This region is mountainous and winds from the eastern plains carry warm moist air. When the wind hits the mountains it is forced upward where it meets and mixes with colder air higher in the atmosphere forming clouds. And this turns out to be a regular daily occurrence during the Colorado summer.

Lopez and his colleagues next looked at the time of day when the maximum number of lightning flashes occurred. (See Figure 3.2.) They found a similar pattern, with a peak showing that most flashes occur between 4 p.m. and 5 p.m.

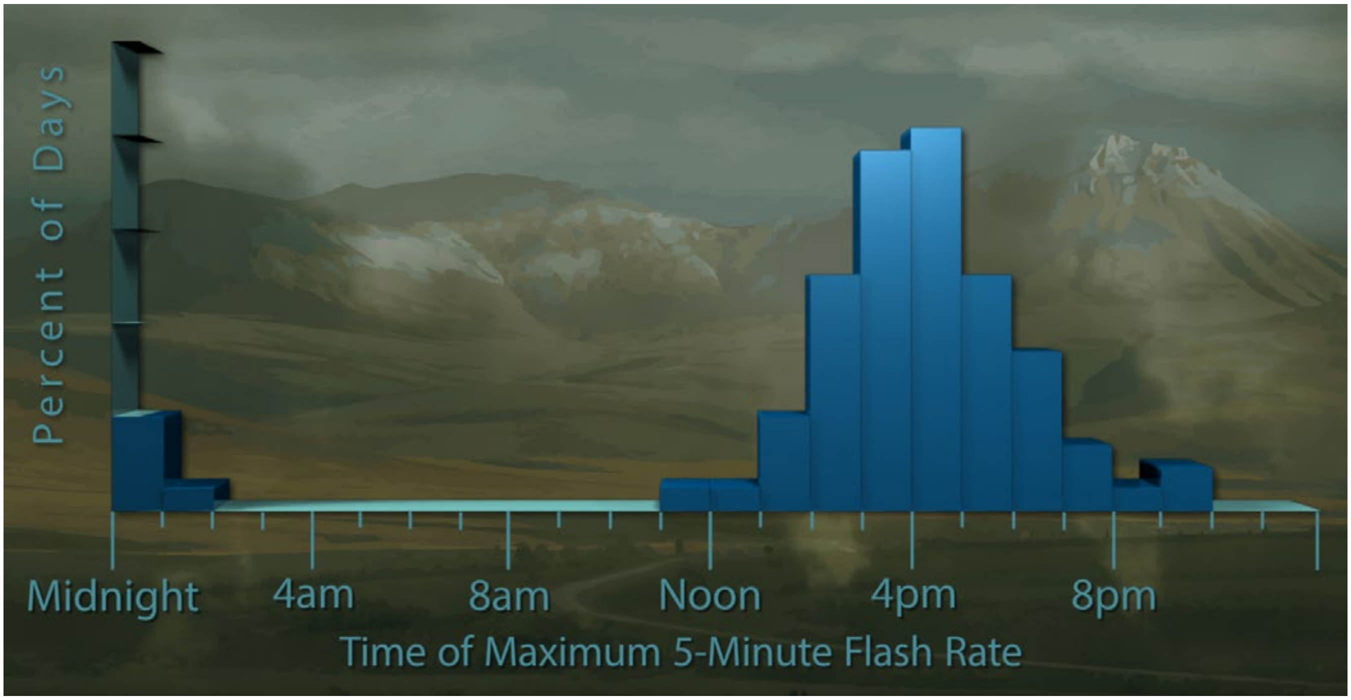

Figure 3.2. Histogram of the time of maximum flash rate.

But there is one big difference from the first flash histogram in Figure 3.1. On a few days the maximum was in the early hours of the morning. Data points like these, which stand out from the overall pattern of the distribution, are called outliers. Outliers are often the most intriguing features of a histogram. Outliers should always be investigated and, if possible, explained.

The explanation that Lopez and his colleagues came up with was that they occur on days when larger weather systems, specifically very strong winds from fast moving weather fronts, overpower the local effect.

Data collection on Colorado lightning has continued since the pioneering work of Raul Lopez and his colleagues. Figure 3.3 shows a histogram produced from more recent data showing the number of people injured or killed by lightning strikes in the last 30 years. It shows the same clustering pattern as Raul Lopez's histograms, but interestingly, the peak time for getting struck by lightning is around 2 p.m., about midway between the peaks of the first strike and maximum activity histograms.

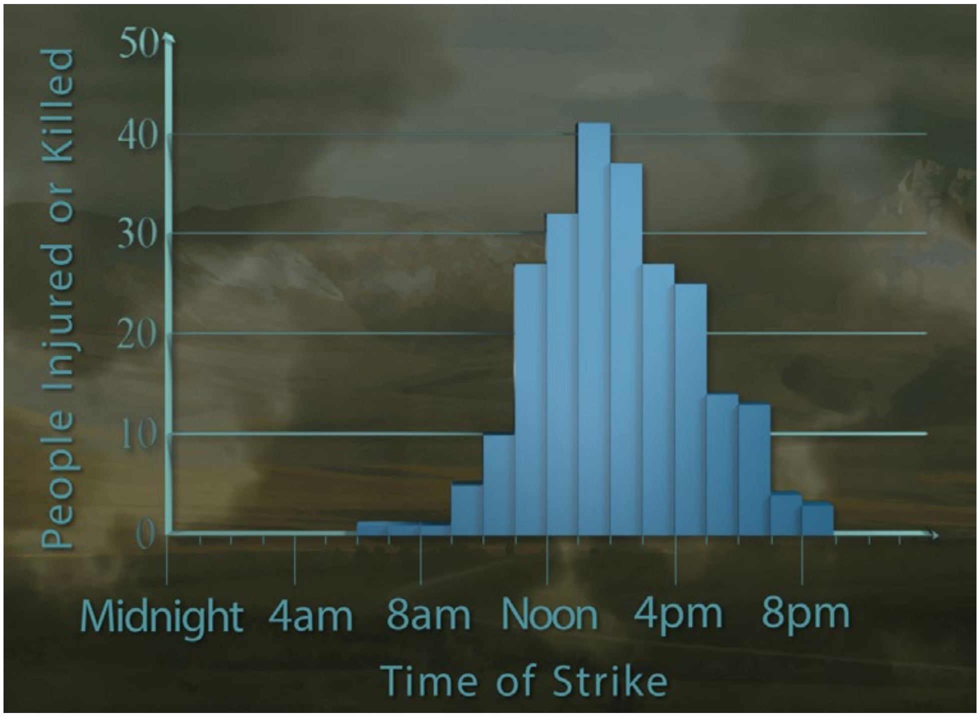

Figure 3.3. Histogram of time when people were struck by lightning.

When constructing histograms it is very important to choose the best class size – that is, the choice of the interval widths for the horizontal axis. Lopez chose one hour for his data, and it works well. But suppose we turn our attention to a different context, the weekday traffic density on a portion of the Massachusetts Turnpike. First, we look at a histogram with class intervals of three hours. (See Figure 3.4.)

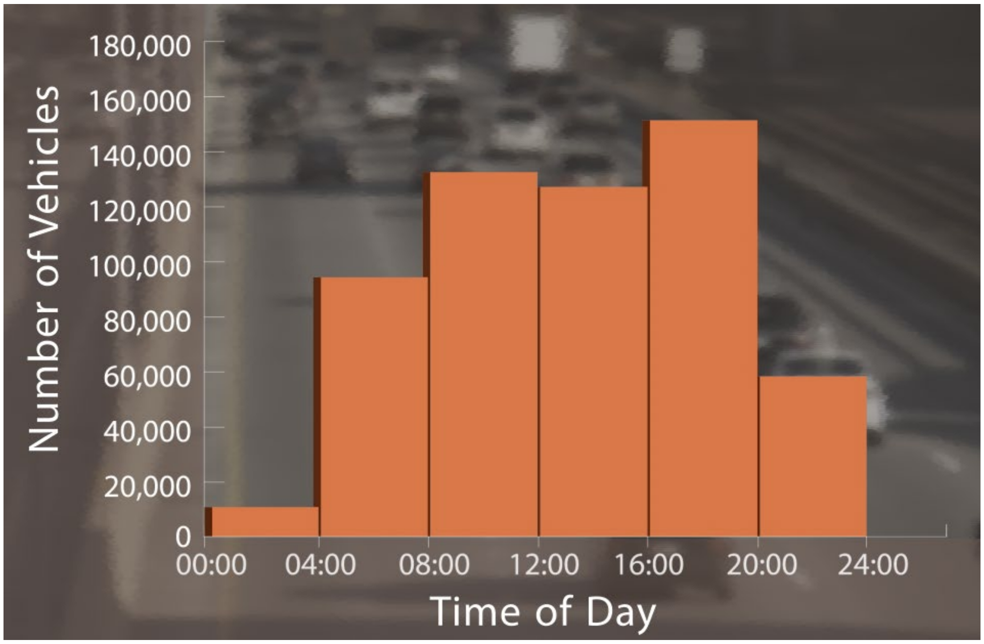

Figure 3.4. Histogram of traffic density in three-hour intervals.

The histogram in Figure 3.4 is not terribly informative. Next, we changed the interval width to one hour, which was better. However, using one-half hour widths as shown in Figure 3.5 is even better. Now, the increased traffic density during morning rush hour and evening rush hour is clearly visible in the pattern of two peaks.

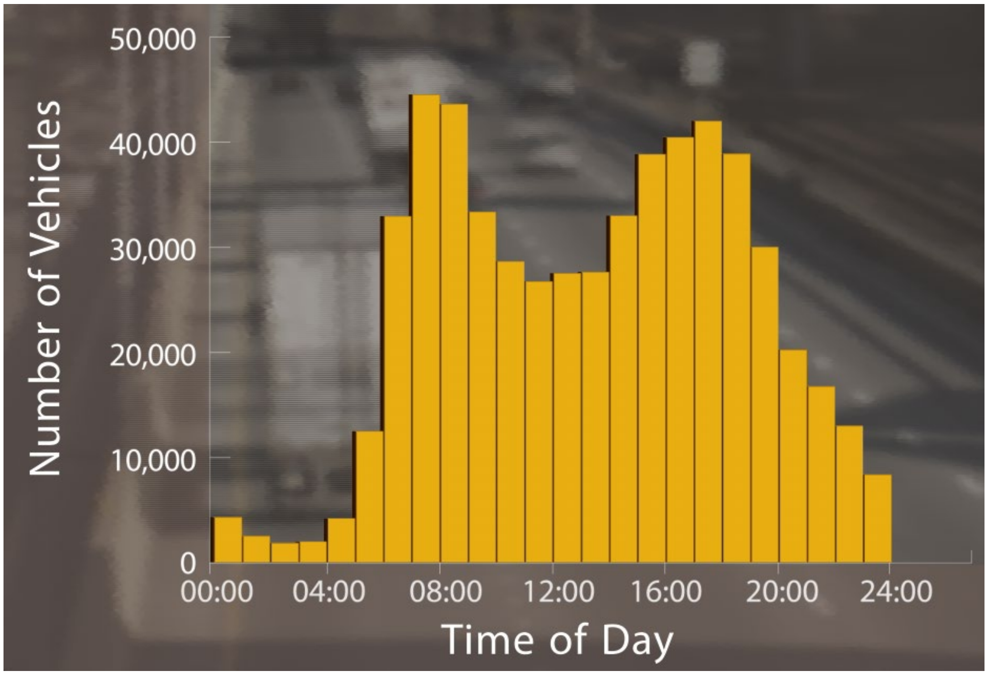

Figure 3.5. Histogram of traffic density in half-hour intervals.

But what if we went even finer-grained and used 5-minute intervals? Take a look at Figure 3.6. Now the peaks begin disappearing again back into the numbers and the histogram becomes less informative.

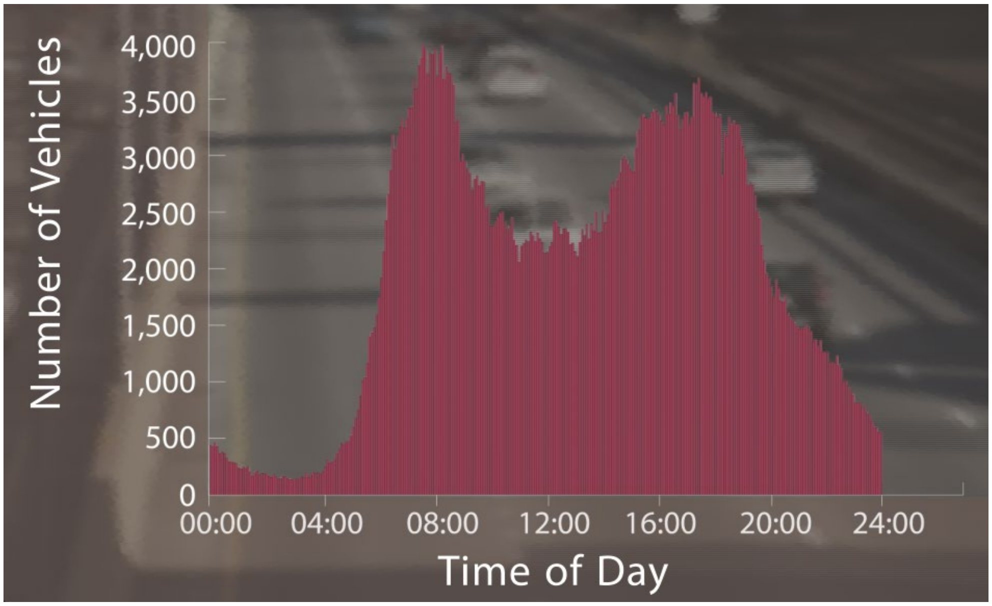

Figure 3.6. Histogram of traffic density in 5-minute intervals.

So, we have seen how histograms can literally show at a glance the essence of a whole lot of numbers. Here is one last example. Figure 3.7 shows a histogram of the weekly wages of workers in the U.S. in the year 1992.

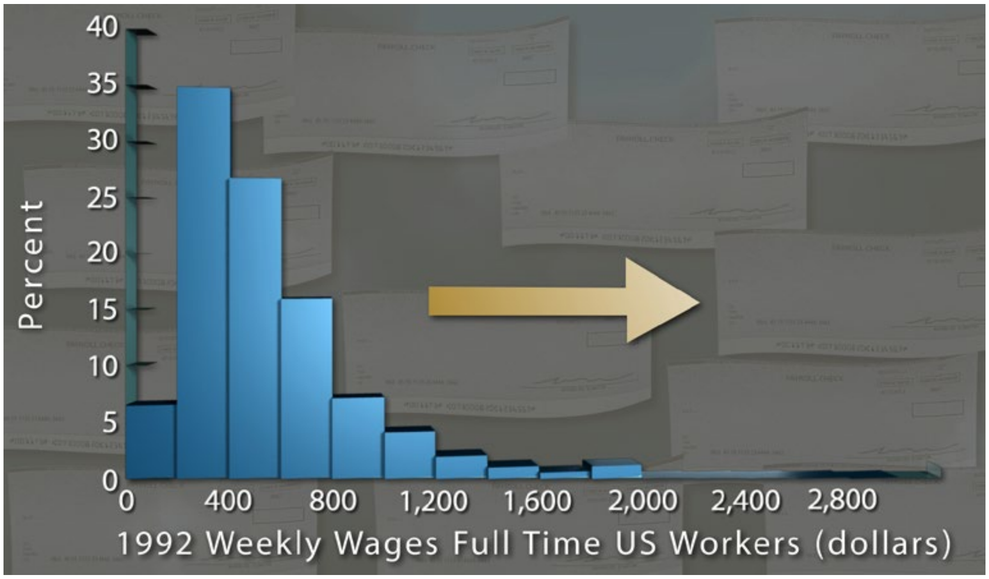

Figure 3.7. Histogram of weekly wages (1992).

Notice how strikingly it is skewed, with most people earning around $450 per week. As you go out to what is called the tail of the distribution (to the right), the salaries get bigger, but the percent of people earning those salaries gets smaller. Statisticians say a distribution like this is skewed to the right, because the right side of the histogram extends much further out than the left side. Now look at the histogram in Figure 3.8 of the same variable, weekly wages, but for the year 2011.

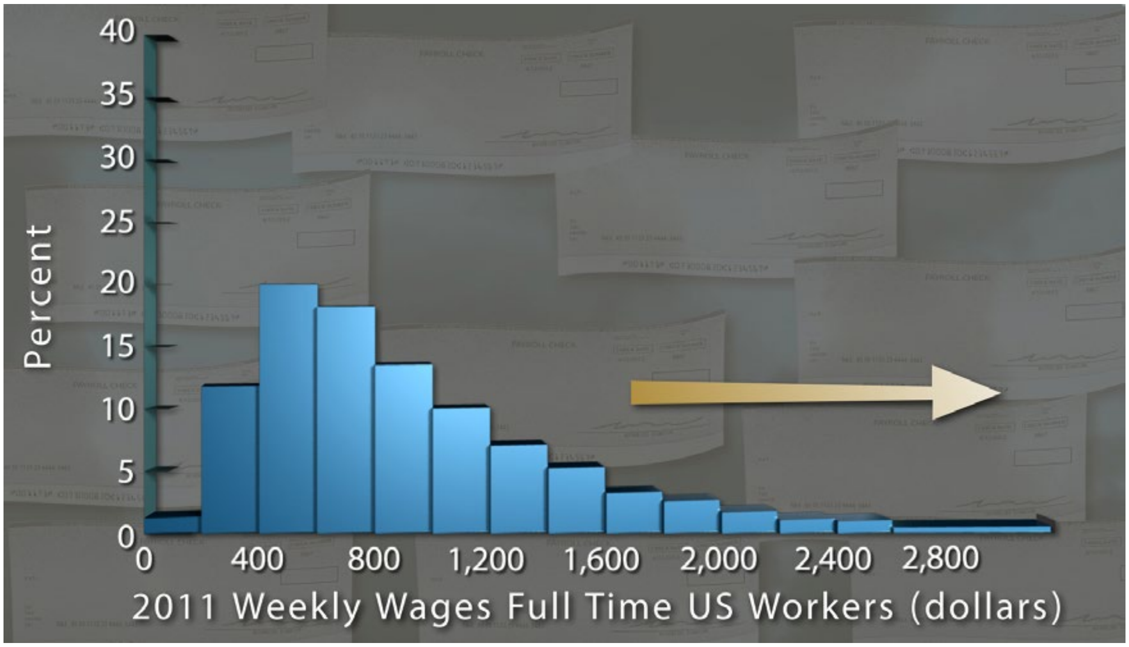

Figure 3.8. Histogram of weekly wages (2011).

Now, the skew has become much more pronounced, and the tail has grown much longer. Suddenly our little discourse on histograms could become highly political!