value. Because there are 78 data values, divide each class frequency by 78, and then multiply by 100%. The first class of Table 2-2 has a frequency of 2, so divide 2 by 78 to get 0.0256, and then multiply by 100% to get 2.56%, which we rounded to 2.6%. The sum of the percentages should be 100%, with a small discrepancy allowed for rounding errors, so a sum such as 99% or 101% is acceptable. The sum of the percentages in Table 2-4 is 100.1%.

The sum of the percentages in a relative frequency distribution must be very close to 100%.

Cumulative Frequency Distribution

Another variation of a frequency distribution is a **cumulative frequency distribution** in which the frequency for each class is the sum of the frequencies for that class and all previous classes. Table 2-5 is a cumulative frequency distribution based on Table 2-2. Using the original frequencies of 2, 33, 35, 7, and 1, we add 2 + 33 to get the second cumulative frequency of 35, then we add 2 + 33 + 35 to get the third, and so on. See Table 2-5, and note that in addition to the use of cumulative frequencies, the class limits are replaced by "less than" expressions that describe the new ranges of values.

Table 2-5 Cumulative Frequency Distribution of IQ Scores of Low Lead Group

IQ Score	Cumulative Frequency
Less than 70	2
Less than 90	35
Less than 110	70
Less than 130	77
Less than 150	78

Critical Thinking: Using Frequency Distributions to Understand Data

Earlier, we noted that a frequency distribution can help us understand the *distribution* of a data set, which is the nature or shape of the spread of the data over the range of values (such as bell-shaped). In statistics we are often interested in determining whether the data have a *normal distribution*. (Normal distributions are discussed extensively in Chapter 6.) Data that have an approximately normal distribution are characterized by a frequency distribution with the following features:

Normal Distribution

1. The frequencies start low, then increase to one or two high frequencies, and then decrease to a low frequency.
2. The distribution is approximately symmetric, with frequencies preceding the maximum being roughly a mirror image of those that follow the maximum.

Table 2-6 satisfies these two conditions. The frequencies start low, increase to the maximum of 56, then decrease to a low frequency. Also, the frequencies of 1 and 10 that precede the maximum are a mirror image of the frequencies 10 and 1 that follow the maximum. Real data sets are usually not so perfect as Table 2-6, and judgment must be used to determine whether the distribution comes "close enough" to satisfying those two conditions.

Outlier Ticket Price

The cost of a Metro-North train ride from Grand Central Station in New York City to the Goldens Bridge station is usually $10.

Johns Hopkins student Lydia Alcock purchased a ticket and charged it to her Visa credit card. She was surprised when she later received her Visa bill and saw that the train ride had a price of $23,148,855,308,184,500, or $23 quadrillion. She was also charged another $20 for exceeding her account limit.

The $23 quadrillion train ticket price is an outlier because it is substantially higher than all other train ticket prices. In this case, the amount is clearly an error. If we were analyzing a sample of train ticket prices and the $23 quadrillion price were included as a sample value, we should delete that value because it is an obvious error. Another outlier might be very far away from the vast majority of other sample values, but we should not necessarily exclude it unless we are sure that it is an error. If we think that the outlier might be a valid sample value, it would be wise to analyze the sample data with the outlier included, and then again with the outlier excluded, so that we can tell what effect the outlier has on the results.