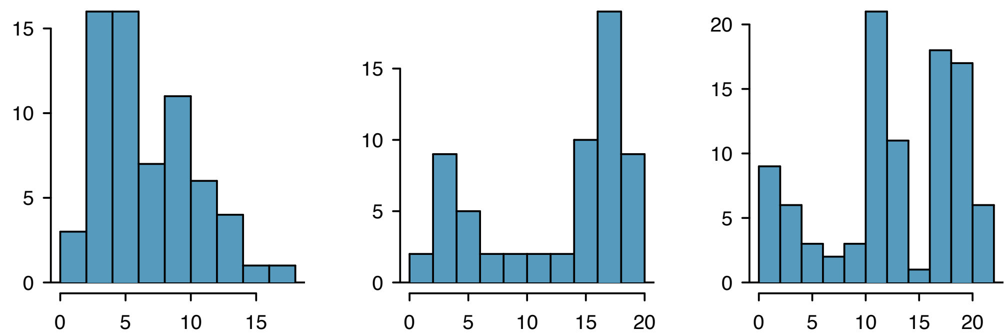

Figure 1.22: Counting only prominent peaks, the distributions are (left to right) unimodal, bimodal, and multimodal.

Exercise 1.27 Height measurements of young students and adult teachers at a K-3 elementary school were taken. How many modes would you anticipate in this height data set?[31]

TIP: Looking for modes
Looking for modes isn't about finding a clear and correct answer about the number of modes in a distribution, which is why *prominent* is not rigorously defined in this book. The important part of this examination is to better understand your data and how it might be structured.

1.6.4 Variance and standard deviation

The mean was introduced as a method to describe the center of a data set, but the variability in the data is also important. Here, we introduce two measures of variability: the variance and the standard deviation. Both of these are very useful in data analysis, even though their formulas are a bit tedious to calculate by hand. The standard deviation is the easier of the two to understand, and it roughly describes how far away the typical observation is from the mean.

We call the distance of an observation from its mean its **deviation**. Below are the deviations for the 1^{st}, 2^{nd}, 3^{rd}, and 50^{th} observations in the `num_char` variable. For computational convenience, the number of characters is listed in the thousands and rounded to the first decimal.

$$\begin{aligned} x_1 - \bar{x} &= 21.7 - 11.6 = 10.1 \\ x_2 - \bar{x} &= 7.0 - 11.6 = -4.6 \\ x_2 - \bar{x} &= 0.6 - 11.6 = -11.0 \\ &\vdots \\ x_{50} - \bar{x} &= 15.8 - 11.6 = 4.2 \end{aligned}$$

[31]There might be two height groups visible in the data set: one of the students and one of the adults. That is, the data are probably bimodal.

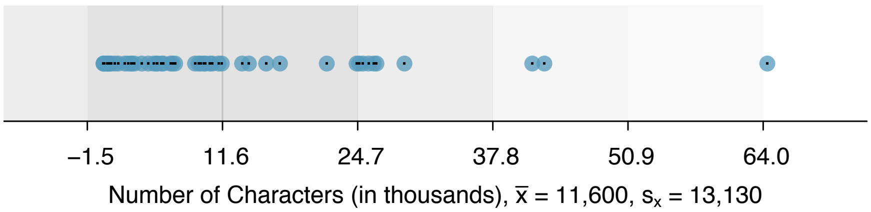

Figure 1.23: In the `num_char` data, 41 of the 50 emails (82%) are within 1 standard deviation of the mean, and 47 of the 50 emails (94%) are within 2 standard deviations. Usually about 70% of the data are within 1 standard deviation of the mean and 95% are within 2 standard deviations, though this rule of thumb is less accurate for skewed data, as shown in this example.

s^2 sample variance

If we square these deviations and then take an average, the result is about equal to the sample **variance**, denoted by s^2:

$$\begin{aligned} s^2 &= \frac{10.1^2 + (-4.6)^2 + (-11.0)^2 + \cdots + 4.2^2}{50 - 1} \\ &= \frac{102.01 + 21.16 + 121.00 + \cdots + 17.64}{49} \\ &= 172.44 \end{aligned}$$

We divide by $n - 1$, rather than dividing by n, when computing the variance; you need not worry about this mathematical nuance for the material in this textbook. Notice that squaring the deviations does two things. First, it makes large values much larger, seen by comparing 10.1^2, $(-4.6)^2$, $(-11.0)^2$, and 4.2^2. Second, it gets rid of any negative signs.

The **standard deviation** is defined as the square root of the variance:

s sample standard deviation

$$s = \sqrt{172.44} = 13.13$$

The standard deviation of the number of characters in an email is about 13.13 thousand. A subscript of $_x$ may be added to the variance and standard deviation, i.e. s_x^2 and s_x, as a reminder that these are the variance and standard deviation of the observations represented by x_1, x_2, ..., x_n. The $_x$ subscript is usually omitted when it is clear which data the variance or standard deviation is referencing.

Variance and standard deviation
The variance is roughly the average squared distance from the mean. The standard deviation is the square root of the variance. The standard deviation is useful when considering how close the data are to the mean.

Formulas and methods used to compute the variance and standard deviation for a population are similar to those used for a sample.[32] However, like the mean, the population values have special symbols: σ^2 for the variance and σ for the standard deviation. The symbol σ is the Greek letter *sigma*.

σ^2 population variance

σ population standard deviation

[32] The only difference is that the population variance has a division by n instead of $n - 1$.

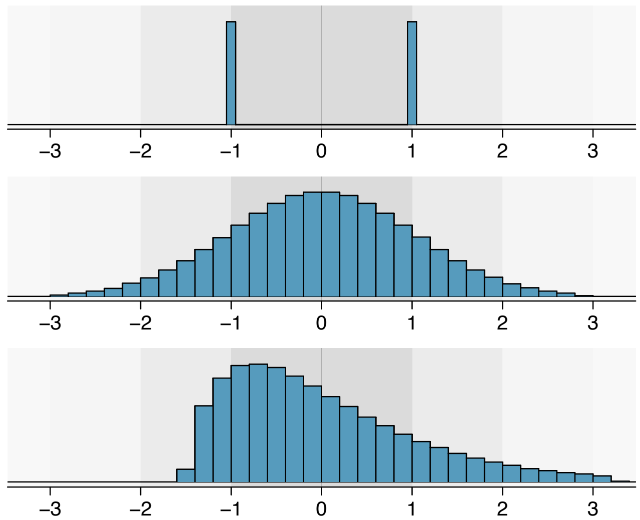

Figure 1.24: Three very different population distributions with the same mean $\mu = 0$ and standard deviation $\sigma = 1$.

TIP: standard deviation describes variability
Focus on the conceptual meaning of the standard deviation as a descriptor of variability rather than the formulas. Usually 70% of the data will be within one standard deviation of the mean and about 95% will be within two standard deviations. However, as seen in Figures 1.23 and 1.24, these percentages are not strict rules.

Exercise 1.28 On page 23, the concept of shape of a distribution was introduced. A good description of the shape of a distribution should include modality and whether the distribution is symmetric or skewed to one side. Using Figure 1.24 as an example, explain why such a description is important.[33]

Example 1.29 Describe the distribution of the `num_char` variable using the histogram in Figure 1.21 on page 24. The description should incorporate the center, variability, and shape of the distribution, and it should also be placed in context: the number of characters in emails. Also note any especially unusual cases.

The distribution of email character counts is unimodal and very strongly skewed to the high end. Many of the counts fall near the mean at 11,600, and most fall within one standard deviation (13,130) of the mean. There is one exceptionally long email with about 65,000 characters.

In practice, the variance and standard deviation are sometimes used as a means to an end, where the "end" is being able to accurately estimate the uncertainty associated with a sample statistic. For example, in Chapter 4 we will use the variance and standard deviation to assess how close the sample mean is to the population mean.

[33]Figure 1.24 shows three distributions that look quite different, but all have the same mean, variance, and standard deviation. Using modality, we can distinguish between the first plot (bimodal) and the last two (unimodal). Using skewness, we can distinguish between the last plot (right skewed) and the first two. While a picture, like a histogram, tells a more complete story, we can use modality and shape (symmetry/skew) to characterize basic information about a distribution.