Distribution Functions

The calculation of the pressure of a gas in the CCR unit *Kinetic Theory* gives us interesting information about the average square speed, and therefore the average energies, of the molecules in a gas, but it does not yield any details about the *distribution* of molecular velocities. Distribution functions are encountered frequently in Chapter 8 and in several later chapters. Here we will discuss distribution functions in general, with some elementary examples from common experience.

Suppose a teacher gave a 25-point quiz to a large number $N$ of students. In order to describe the results of the quiz, the teacher might give the average or median score, but this would not be a complete description. For example, if all $N$ students received 12.5, this is quite a different result than if $N/2$ students received 25 and $N/2$ received 0, though both results have the same average. A more complete description would be to give the number $n_i$ who received the score $s_i$ for all scores $s_i$ between 0 and 25. An alternative would be to divide $n_i$ by the total number of students $N$ to give the fraction $f_i = n_i/N$ receiving the score $s_i$. Both $n_i$ and $f_i$ (which depend on the variable $s$) are called *distribution functions*. The fractional distribution $f_i$ is slightly more convenient to use. The probability that one of the $N$ students selected at random received the score $s_i$ equals the number of students that received that score, $n_i = Nf_i$, divided by the total number $N$; thus this probability equals the distribution function $f_i$.

Note that

$$\sum_i f_i = \sum_i \frac{n_i}{N} = \frac{1}{N} \sum_i n_i$$

and since

$$\sum_i n_i = N$$

we have

$$\sum_i f_i = 1 \quad \text{DF-1}$$

Equation DF-1 is called the *normalization condition* for fractional distribution functions. A possible distribution function for a 25-point quiz is shown in Figure DF-1.

To find the average score, all the scores are added and the result is divided by $N$. Since each score $s_i$ was obtained by $n_i = Nf_i$ students, this procedure is equivalent to

$$\bar{s} = \frac{1}{N} \sum_i s_i n_i = \sum_i s_i f_i \quad \text{DF-2}$$

We will take Equation DF-2 as the definition of the *average* (or *mean*) score $\bar{s}$. Similarly, the average of any function $g(s)$ is defined by
In particular, the mean square score is often useful:

\[ s^2 = \sum_i s_i^2 f_i \]

A useful quantity characterizing a distribution is the standard deviation, \( \sigma \), defined by

\[ \sigma = \left( \frac{\sum_i (s_i - \bar{s})^2 f_i}{n} \right)^{1/2} \]

Note that

\[ \sum_i (s_i - \bar{s})^2 f_i = \sum_i s_i^2 f_i + \bar{s}^2 \sum_i f_i - 2\bar{s} \sum_i s_i f_i = \bar{s}^2 - \bar{s}^2 \]

Therefore

\[ \sigma = (\bar{s}^2 - \bar{s}^2)^{1/2} \]

The standard deviation measures the spread of the values \( s_i \) about the mean. For most distributions there will be few values that differ from \( \bar{s} \) by more than a few multiples of \( \sigma \). In the case of the normal Gaussian distribution, common in the theory of errors, about two-thirds of the values lie within \( \pm \sigma \) of the mean value. A Gaussian distribution is shown in Figure DF-2.

If a student were selected at random from the class and one had to guess that student’s score, the best guess would be the score obtained by the greatest number of students, called the most probable score, \( s_m \). For the distribution in Figure DF-1, \( s_m \) is 16 and the average score, \( \bar{s} \), is 14.17. The root-mean-square score, \( s_{rms} = (\bar{s}^2)^{1/2} \), is 14.9, and the standard deviation \( \sigma \) is 4.6. Note that 66 percent of the scores for this distribution lie within \( \bar{s} \pm \sigma = 14.17 \pm 4.6 \).

Now consider the case of a continuous distribution. Suppose we wanted to know the distribution of heights of a large number of people. For a finite number \( N \), the
number of persons exactly 6 feet tall would be zero. If we assume that height can be determined to any desired accuracy, there is an infinite number of possible heights, and the chance that anybody has a particular exact height is zero. We would therefore divide the heights into intervals $\Delta h$ (for example, $\Delta h$ could be 0.1 ft) and ask what fraction of people have heights that fall in any particular interval. This number depends on the size of the interval. We define the distribution function $f(h)$ as the fraction of the number of people with heights in a particular interval divided by the size of the interval. Thus, for $N$ people, $Nf(h)\Delta h$ is the number of people whose height is in the interval between $h$ and $h + \Delta h$. A possible height-distribution function is plotted in Figure DF-3. The fraction of people with heights in a particular interval is the area of the rectangle $\Delta h f(h)$. The total area represents the sum of all fractions; thus it must equal 1. If $N$ is very large, we can choose $\Delta h$ very small and still have $f(h)$ vary only slightly between intervals. The histogram $f(h)$ versus $h$ approaches a smooth curve as $N \to \infty$ and $\Delta h \to 0$. In many cases of importance, the number of objects $N$ is extremely large and the intervals can be taken as small as measurement allows. The distribution functions $f(h)$ are usually considered to be continuous functions, intervals are written $dh$, and the sums are replaced by integrals. For example, if $f(h)$ is a continuous function, the average height, which we will write as $\langle h \rangle$ for the continuous function $f(h)$, is\(^1\)

$$\langle h \rangle = \int h f(h) \, dh$$  \hfill DF-6

and the normalization condition expressing the fact that the sum of all fractions is 1 is

$$\int f(h) \, dh = 1$$  \hfill DF-7

The CCR unit *Boltzmann Distribution* illustrates a physical application of distribution functions as does the computation of expectation values in Chapters 6 and 7, among others, and in a number of More sections.

\(^1\)The limits on the integration depend on the range of the variable. For this case, $h$ ranges from 0 to $\infty$. We shall often omit explicit indication of the limits when the range of the variable is clear.