Chapter 2

Distributions in Physics

Topics

How to handle distributions in physics. Discrete distributions, averages and true means, variance and standard deviation. Continuous distributions, distribution functions and probability densities. Examples.

We need to spend a little time revising the concepts of discrete and continuous distributions and how to describe them mathematically. Many of the parameters involved will be related to physically measurable quantities. More details of these topics are contained in Section 16 of the *Maths Handbook.*

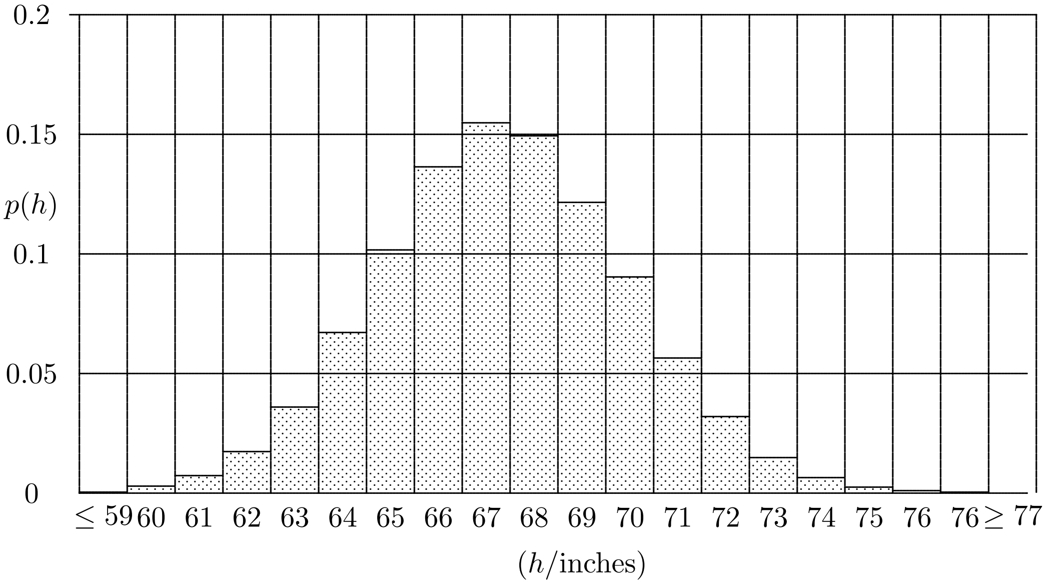
2.1 Discrete Distributions

Let us begin with a splendid old example from my favorite statistics book, *Teach Yourself Statistics.* In 1951, the heights of all the service personnel born in 1933 were measured – typically, they were about 18 years old. There were 58,703 personnel and the distribution of their heights to the nearest inch is listed in Table 2.1 and presented as a histogram in Figure 2.1. The box 60 inches means heights in the interval $60 \leq h < 61$ so that the mean is 60.5.

Table 2.1. Heights of national service personnel born in 1933 in the 1951 intake.

Height (h/inches)	Number	Probability $p(h)$
≤ 59	23	0.0004
60	169	0.0029
61	439	0.0075
62	1030	0.0175
63	2116	0.0360
64	3947	0.0672
65	5965	0.1016
66	8012	0.1365
67	9089	0.1548
68	8763	0.1493
69	7132	0.1215
70	5314	0.0905
71	3320	0.0566
72	1884	0.0321
73	876	0.0149
74	383	0.0065
75	153	0.0026
76	63	0.0011
≥ 77	25	0.0004
Total	58703	1.0000

Table 2.1 The probability distribution of heights of national service personnel in 1951.



We need to be able to characterise such distributions and so let us revise some of these measures.

2.1.1 Averages and True Means

The *average* or *sample average* of a set of N quantities $x_1, x_2, x_3 \ldots x_N$ is defined as

$$\overline{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_N}{N} = \frac{\sum_i x_i}{N}. \qquad (2.1)$$

We can also define a *weighted average* by

$$\overline{x} = \frac{\sum_i w_i x_i}{\sum_i w_i} \qquad (2.2)$$

where the quantities w_i are the *weights*. The values with larger weights have more influence in defining the 'typical' value. The ordinary average (2.1) has the same weight, $w_i = 1$, for every value x_i and so $\sum_i w_i = N$.

Now, consider some physical quantity x, for which we wish to measure the average value as precisely as possible. We could imagine taking a *very large*, or, in the limit, an *infinite* set of measurements. The average of such an infinite set of measurements is known as the *mean* or *true mean* value of x, and is written as μ or μ_x, or $\langle x \rangle$. It is sometimes also written as $\overline{x}$, but we shall use this symbol for the *average of a finite sample*. This may look somewhat pedantic, but it emphasises the point that the sample average of a finite set of measurements provides only an *estimate* of the true mean and *not* the true mean itself. Physicists have to live with such uncertainties and try to improve the precision of their experiments.

2.1.2 Probabilities

Let us define the probability $p(x_i)$ of obtaining a given value of x_i for a discrete distribution. $p(x_i)$ is *normalised* so that the total probability of obtaining any value of x_i is unity. Therefore,

$$p(x_i) = \frac{N(x_i)}{\sum_i N(x_i)}, \tag{2.3}$$

where $N(x_i)$ is the number of occurrences of x_i. The data shown in Table 2.1 are also shown as a set of probabilities $p(h)$ by dividing all the numbers in the second column 2 by 58,703.

Thus, we can equally well write the average as

$$\overline{x} = \sum_i x_i \, p(x_i). \tag{2.4}$$

Taking the extreme values ≤ 59 and ≥ 77 to be 59.5 and 77.5 respectively, the sample average height in Table 2.1 is 67.87 inches $\approx$ 5 feet 8 inches.

2.1.3 Variance and Standard Deviation

For any set of data, an important quantity is the spread of the data about the average value. This quantity has many applications in physics.

The quantity $\varepsilon_i = x_i - \mu$ for any particular measurement x_i is called the *deviation* or, in the case of the analysis of errors, the *error*. Notice that we have written the expression for ε in terms of the true mean μ. For very large samples, the mean deviation is, from the definition of the mean, zero. The mean square deviation σ^2 is, however, finite and is called the *variance*,

$$\text{variance} = \sigma^2 = \langle \varepsilon_i^2 \rangle = \frac{1}{N} \sum_i \varepsilon_i^2 = \frac{1}{N} \sum_i (x_i - \mu)^2, \tag{2.5}$$

that is, the variance is the quantity $(x_i - \mu)^2$ averaged over an infinite number of measurements.

Notice that, in (2.5), we compared the deviation with the *true mean* and normally we will not know the value of μ. If instead we refer the deviations ε_i to the sample average $\varepsilon_i = (x_i - \overline{x})$, the *sample variance* s^2 is defined to be

$$\text{sample variance} = s^2 = \langle \varepsilon_i^2 \rangle = \frac{1}{N-1}\sum_i \varepsilon_i^2$$
$$= \frac{1}{N-1}\sum_i (x_i - \overline{x})^2. \qquad (2.6)$$

The -1 in the $N-1$ takes account of the fact that we have already used one piece of information about the properties of the data-set in determining the sample average $\overline{x}$. Generally speaking, when dealing with large enough samples, the distinction between N and $N-1$ is not important, but it is best to use $N-1$ to be strictly correct if the data are referred to the sample average.

If the data are presented as a histogram of probabilities, as in Table 2.1, we can equally well write the sample variance in terms of the probabilities $p(x_i)$

$$s^2 = \frac{N}{N-1}\sum_i (x_i - \overline{x})^2\, p(x_i). \qquad (2.7)$$

We often need a suitable measure of the typical deviation from the sample mean and we define the *standard deviation*, or the *standard error* in the case of errors, as σ or s, the square root of the variance – this quantity is often referred to as the *root mean square deviation.*

A measure of the spread in heights of the service personnel in Table 2.1 is found from $s^2 = 7.530$ inches^2 and the standard deviation $s = 2.74$ inches. About 70% of the heights lie between $\pm s$.

A Useful simplification to find σ^2
Rather than having to work out $x_i - \mu$, square and add, we can use the following simplifiation.

$$\sigma^2 = \frac{1}{N}\sum_i \varepsilon_i^2 = \frac{1}{N}\sum_i (x_i - \mu)^2,$$
$$= \frac{1}{N}\sum_i x_i^2 - \frac{1}{N}\sum_i 2x_i\mu + \frac{1}{N}\sum_i \mu^2,$$
$$= \frac{1}{N}\sum_i x_i^2 - \frac{2\mu}{N}\sum_i x_i + \mu^2.$$

But, $(\sum_i x_i)/N = \mu$ and so

$$\sigma^2 = \frac{1}{N}\sum_i x_i^2 - \mu^2.$$

2.2 Continuous Distributions

In the case of very large samples, for example, the 6×10^{23} molecules in one mole of gas, we need to think in terms of, say, their velocities as having a *continuous* range of values, rather than each having a discrete value. The probability of finding some particular value *exactly* is now zero, and this is not a helpful idea.

Instead, we use the concept that the probability that v lies *in some narrow range* between, say, v and $v+\mathrm{d}v$. We express this in terms of the *distribution function* $f(v)$, which is the *probability density* per unit range of the velocity v. The probability $\mathrm{d}p$ that v lies in the narrow range of velocity v to $v + \mathrm{d}v$ is defined to be

$$\mathrm{d}p(v) = f(v)\,\mathrm{d}v. \tag{2.8}$$

The probability on the left hand side is a dimensionless number between zero and one. $\mathrm{d}v$ has the dimensions of velocity and so the function $f(v)$ must have dimensions $1/v$, that is, it is the probability per unit range of velocity.

Consider the case of the *one-dimensional velocity distribution* of the molecules in a gas. This means that we consider only, say, the x-component of the particles' velocities. Therefore, the value of v_x can range between $+\infty$ and $-\infty$. The function, which we will derive later, is

$$f(v_x)\,\mathrm{d}v_x = \frac{1}{\alpha(2\pi)^{1/2}}\mathrm{e}^{-v_x^2/2\alpha^2}\mathrm{d}v_x, \tag{2.9}$$

where $\alpha = (kT/m)^{1/2}$, and is shown in Figure 2.5. The key point is that *areas under the $f(v_x)$ curve, and **not** the values of $f(v_x)$ represent the probabilities.* This obviously has to be the case since a probability has to be a dimensionless number and $f(v_x)$ has dimensions [velocity]$^{-1}$. Thus, the little shaded area is $f(v_x)\,\mathrm{d}v_x$ tells us the probability that the velocity of the molecule lies in the velocity range v_x to $v_x + \mathrm{d}v_x$. The total area under the curve represents the total probability that v_x has some value and so must be equal to unity. Therefore,

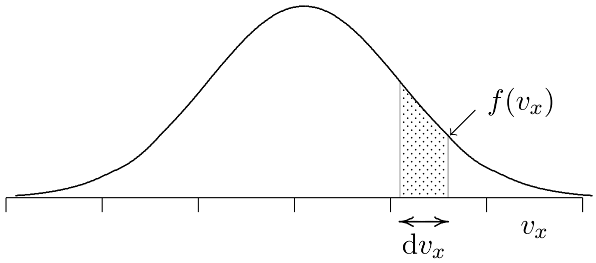
$$\int_{-\infty}^{\infty} f(v_x)\,\mathrm{d}v_x = 1. \tag{2.10}$$

We say that the probability distribution $f(v_x)$ has been *normalised.*

You should check for yourselves that the expression (2.14) has been correctly normalised using the integral relation

$$\int_{-\infty}^{\infty} \mathrm{e}^{-x^2}\,\mathrm{d}x = \sqrt{\pi}. \tag{2.11}$$

Figure 2.5 The probability density $f(v_x) = \dfrac{1}{\alpha(2\pi)^{1/2}}\mathrm{e}^{-v_x^2/2\alpha^2}$



Gaussian or Normal Distribution
The one-dimensional velocity distribution is an example of the *Gaussian* or *Normal* Distribution.

$$f(x)\,\mathrm{d}x = \frac{1}{(2\pi)^{1/2}}\mathrm{e}^{-x^2/2}\,\mathrm{d}x.$$

Continuous probability distributions
Areas under the $f(v_x)$ curve $\int_{x_1}^{x_2} f(v_x)\,\mathrm{d}v_x$, and **not** the values of $f(v_x)$, represent the probabilities.

It takes a little while to get used to the use of continuous functions and it may seem strange that probability density functions have dimensions. The key thing to remember is that the function itself only becomes a probability when multiplied by $\mathrm{d}v_x$. We can make $\mathrm{d}v_x$ as small as we like and so obtain the probability for an infinitesimal range of $\mathrm{d}v_x$ about the value v_x in which we are interested.

It is now straightforward to define the statistical measures we discussed above, but now for continuous probability density distributions. Now we deal with *integrals* rather than sums.

- The *mean value* of v_x is
$$\mu_{v_x} = \int v_x \, \mathrm{d}p = \int_{-\infty}^{\infty} v_x f(v_x) \, \mathrm{d}v_x. \quad (2.12)$$
- The *variance* is
$$\sigma^2 = \int_{-\infty}^{\infty} (v_x - \mu_{v_x})^2 f(v_x) \, \mathrm{d}v_x. \quad (2.13)$$

As an exercise, you should confirm the following results for the velocity distribution shown in Figure 2.5.

- $\mu_{v_x} = 0$;
- $\sigma^2 = \alpha^2$ and $\sigma = \alpha$; $\Longrightarrow$

Hint: integrate $\int_{-\infty}^{\infty} x^2 \mathrm{e}^{-x^2} \, \mathrm{d}x$ by parts to reduce the integral to $\int_{-\infty}^{\infty} \mathrm{e}^{-x^2} \, \mathrm{d}x$, for which the result has already been quoted.