MATHCHAPTER B

Probability and Statistics

In many of the following chapters, we will deal with probability distributions, average values, and standard deviations. Consequently, we take a few pages here to discuss some basic ideas of probability and show how to calculate average quantities in general.

Consider some experiment, such as the tossing of a coin or the rolling of a die, that has n possible outcomes, each with probability p_j, where $j = 1, 2, \ldots, n$. If the experiment is repeated indefinitely, we intuitively expect that

$$p_j = \lim_{N\to\infty} \frac{N_j}{N} \qquad j = 1, 2, \ldots, n \tag{B.1}$$

where N_j is the number of times that the event j occurs and N is the total number of repetitions of the experiment. Because $0 \leq N_j \leq N$, p_j must satisfy the condition

$$0 \leq p_j \leq 1 \tag{B.2}$$

When $p_j = 1$, we say the event j is a certainty and when $p_j = 0$, we say it is impossible. In addition, because

$$\sum_{j=1}^{n} N_j = N$$

we have the normalization condition,

$$\sum_{j=1}^{n} p_j = 1 \tag{B.3}$$

Equation B.3 means that the probability that some event occurs is a certainty. Suppose now that some number x_j is associated with the outcome j. Then we define the *average*

of x or the *mean* of x to be

$$\langle x \rangle = \sum_{j=1}^{n} x_j p_j = \sum_{j=1}^{n} x_j p(x_j) \tag{B.4}$$

where in the last term we have used the expanded notation $p(x_j)$, meaning the probability of realizing the number x_j. We will denote an average of a quantity by enclosing the quantity in angular brackets.

EXAMPLE B–1

Suppose we are given the following data:

x	$p(x)$
1	0.20
3	0.25
4	0.55

Calculate the average value of x.

SOLUTION: Using Equation B.4, we have

$$\langle x \rangle = (1)(0.20) + (3)(0.25) + (4)(0.55) = 3.15$$

It is helpful to interpret a probability distribution like p_j as a distribution of a unit mass along the x axis in a discrete manner such that p_j is the fraction of mass located at the point x_j (Figure B.1). According to this interpretation, the average value of x is the center of mass of this system.

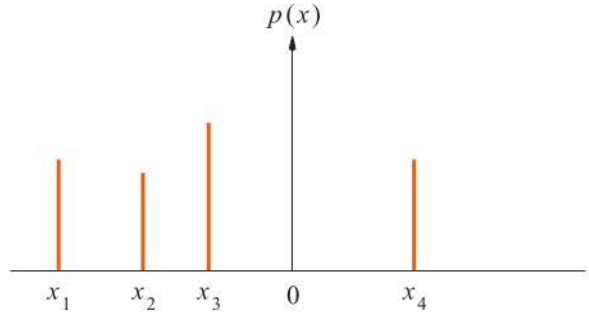

FIGURE B.1
The discrete probability frequency function or probability density, $p(x)$.

Another quantity of importance is

$$\langle x^2 \rangle = \sum_{j=1}^{n} x_j^2 p_j \tag{B.5}$$

The quantity $\langle x^2 \rangle$ is called the *second moment* of the distribution $\{p_j\}$ and is analogous to the moment of inertia.

EXAMPLE B–2

Calculate the second moment of the data given in Example B–1.

SOLUTION: Using Equation B.5, we have

$$\langle x^2 \rangle = (1)^2(0.20) + (3)^2(0.25) + (4)^2(0.55) = 11.25$$

Note from Examples B–1 and B–2 that $\langle x^2 \rangle \neq \langle x \rangle^2$. This nonequality is a general result that we will prove below.

A physically more interesting quantity than $\langle x^2 \rangle$ is the *second central moment*, or the *variance*, defined by

$$\sigma_x^2 = \langle (x - \langle x \rangle)^2 \rangle = \sum_{j=1}^{n} (x_j - \langle x \rangle)^2 p_j \tag{B.6}$$

As the notation suggests, we denote the square root of the quantity in Equation B.6 by σ_x, which is called the *standard deviation*. From the summation in Equation B.6, we can see that σ_x^2 will be large if x_j is likely to differ from $\langle x \rangle$, because in that case $(x_j - \langle x \rangle)$ and so $(x_j - \langle x \rangle)^2$ will be large for the significant values of p_j. On the other hand, σ_x^2 will be small if x_j is not likely to differ from $\langle x \rangle$, or if the x_j cluster around $\langle x \rangle$, because then $(x_j - \langle x \rangle)^2$ will be small for the significant values of p_j. Thus, we see that either the variance or the standard deviation is a measure of the spread of the distribution about its mean.

Equation B.6 shows that σ_x^2 is a sum of positive terms, and so $\sigma_x^2 \geq 0$. Furthermore,

$$\begin{aligned} \sigma_x^2 &= \sum_{j=1}^{n} (x_j - \langle x \rangle)^2 p_j = \sum_{j=1}^{n} (x_j^2 - 2\langle x \rangle x_j + \langle x \rangle^2) p_j \\ &= \sum_{j=1}^{n} x_j^2 p_j - 2 \sum_{j=1}^{n} \langle x \rangle x_j p_j + \sum_{j=1}^{n} \langle x \rangle^2 p_j \end{aligned} \tag{B.7}$$

The first term here is just $\langle x^2 \rangle$ (cf. Equation B.5). To evaluate the second and third terms, we need to realize that $\langle x \rangle$, the average of x_j, is just a number and so can be factored out of the summations, leaving a summation of the form $\sum x_j p_j$ in the second term and

$\sum p_j$ in the third term. The summation $\sum x_j p_j$ is $\langle x \rangle$ by definition and the summation $\sum p_j$ is unity because of normalization (Equation B.3). Putting all this together, we find that

$$\begin{aligned} \sigma_x^2 &= \langle x^2 \rangle - 2\langle x \rangle^2 + \langle x \rangle^2 \\ &= \langle x^2 \rangle - \langle x \rangle^2 \geq 0 \end{aligned} \tag{B.8}$$

Because $\sigma_x^2 \geq 0$, we see that $\langle x^2 \rangle \geq \langle x \rangle^2$. A consideration of Equation B.6 shows that $\sigma_x^2 = 0$ or $\langle x \rangle^2 = \langle x^2 \rangle$ only when $x_j = \langle x \rangle$ with a probability of one, a case that is not really probabilistic because the event j occurs on every trial.

So far we have considered only discrete distributions, but continuous distributions are also important in physical chemistry. It is convenient to use the unit mass analogy. Consider a unit mass to be distributed continuously along the x axis, or along some interval on the x axis. We define the linear mass density $\rho(x)$ by

$$dm = \rho(x)dx$$

where dm is the fraction of the mass lying between x and $x + dx$. By analogy, then, we say that the probability that some quantity x, such as the position of a particle in a box, lies between x and $x + dx$ is

$$\text{Prob}\{x, x + dx\} = p(x)dx \tag{B.9}$$

and that

$$\text{Prob}\{a \leq x \leq b\} = \int_a^b p(x)dx \tag{B.10}$$

In the mass analogy, Prob$\{a \leq x \leq b\}$ is the fraction of mass that lies in the interval $a \leq x \leq b$. The normalization condition is

$$\int_a^b p(x)dx = 1 \tag{B.11}$$

Following Equations B.4 through B.6, we have the definitions

$$\langle x \rangle = \int_a^b x p(x)dx \tag{B.12}$$

$$\langle x^2 \rangle = \int_a^b x^2 p(x)dx \tag{B.13}$$

and

$$\sigma_x^2 = \int_a^b (x - \langle x \rangle)^2 p(x)dx \tag{B.14}$$

EXAMPLE B–3

Perhaps the simplest continuous distribution is the so-called uniform distribution, where

$$p(x) = \begin{cases} \text{constant} = A & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

Show that A must equal $1/(b-a)$. Evaluate $\langle x \rangle$, $\langle x^2 \rangle$, σ_x^2, and σ_x for this distribution.

SOLUTION: Because $p(x)$ must be normalized,

$$\int_a^b p(x)dx = 1 = A\int_a^b dx = A(b-a)$$

Therefore, $A = 1/(b-a)$ and

$$p(x) = \begin{cases} \dfrac{1}{b-a} & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

The mean of x is given by

$$\langle x \rangle = \int_a^b xp(x)dx = \frac{1}{b-a}\int_a^b xdx$$
$$= \frac{b^2-a^2}{2(b-a)} = \frac{b+a}{2}$$

and the second moment of x by

$$\langle x^2 \rangle = \int_a^b x^2 p(x)dx = \frac{1}{b-a}\int_a^b x^2 dx$$
$$= \frac{b^3-a^3}{3(b-a)} = \frac{b^2+ab+a^2}{3}$$

Last, the variance is given by Equation B.6, and so

$$\sigma_x^2 = \langle x^2 \rangle - \langle x \rangle^2 = \frac{(b-a)^2}{12}$$

and the standard deviation is

$$\sigma_x = \frac{(b-a)}{\sqrt{12}}$$

EXAMPLE B–4

The most commonly occurring and most important continuous probability distribution is the *Gaussian distribution*, given by

$$p(x)dx = ce^{-x^2/2a^2}dx \qquad -\infty < x < \infty$$

Find c, $\langle x \rangle$, σ_x^2, and σ_x.

SOLUTION: The constant c is determined by normalization:

$$\int_{-\infty}^{\infty} p(x)dx = 1 = c\int_{-\infty}^{\infty} e^{-x^2/2a^2}dx \tag{B.15}$$

If you look in a table of integrals (e.g., *The CRC Standard Mathematical Tables* or *The CRC Handbook of Chemistry and Physics*; CRC Press: Boca Raton, FL), you won't find the above integral. However, you will find the integral (see also Problem B–6)

$$\int_0^{\infty} e^{-\alpha x^2}dx = \left(\frac{\pi}{4\alpha}\right)^{1/2} \tag{B.16}$$

The reason that you won't find the integral with the limts $(-\infty, \infty)$ is illustrated in Figure B.2a, where $e^{-\alpha x^2}$ is plotted against x. Note that the graph is symmetric about the vertical axis, so that the corresponding areas on the two sides of the axis are equal. Such a function has the mathematical property that $f(x) = f(-x)$ and is called an *even function*. For an even function

$$\int_{-A}^{A} f_{\text{even}}(x)dx = 2\int_0^{A} f_{\text{even}}(x)dx \tag{B.17}$$

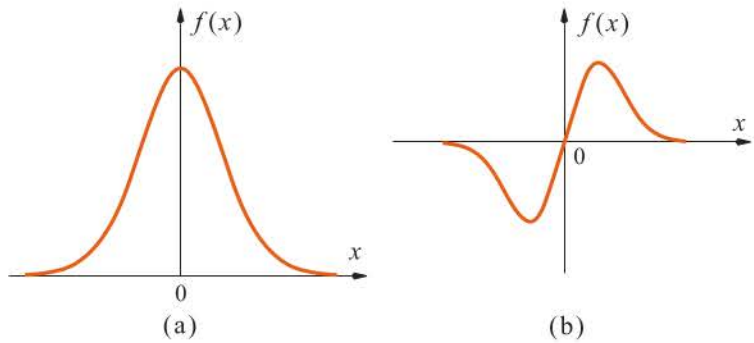

FIGURE B.2

(a) The function $f(x) = e^{-x^2}$ is an even function, $f(x) = f(-x)$. (b) The function $f(x) = xe^{-x^2}$ is an odd function, $f(x) = -f(-x)$.

If we recognize that $p(x) = e^{-x^2/2a^2}$ is an even function and use Equation B.16, then we find that

$$c\int_{-\infty}^{\infty} e^{-x^2/2a^2}dx = 2c\int_{0}^{\infty} e^{-x^2/2a^2}dx$$

$$= 2c\left(\frac{\pi a^2}{2}\right)^{1/2} = 1$$

or $c = 1/(2\pi a^2)^{1/2}$.

The mean of x is given by

$$\langle x\rangle = \int_{-\infty}^{\infty} xp(x)dx = (2\pi a^2)^{-1/2}\int_{-\infty}^{\infty} xe^{-x^2/2a^2}dx \tag{B.18}$$

The integrand in Equation B.18 is plotted in Figure B.2b. Notice that this graph is antisymmetric about the vertical axis and that the area on one side of the vertical axis cancels the corresponding area on the other side. This function has the mathematical property that $f(x) = -f(-x)$ and is called an *odd function*. For an odd function,

$$\int_{-A}^{A} f_{\text{odd}}(x)dx = 0 \tag{B.19}$$

The function $xe^{-x^2/2a^2}$ is an odd function, and so

$$\langle x\rangle = \int_{-\infty}^{\infty} xe^{-x^2/2a^2}dx = 0$$

The second moment of x is given by

$$\langle x^2\rangle = (2\pi a^2)^{-1/2}\int_{-\infty}^{\infty} x^2e^{-x^2/2a^2}dx$$

The integrand in this case is even because $f(x) = x^2e^{-x^2/2a^2} = f(-x)$. Therefore,

$$\langle x^2\rangle = 2(2\pi a^2)^{-1/2}\int_{0}^{\infty} x^2e^{-x^2/2a^2}dx$$

The integral

$$\int_{0}^{\infty} x^2e^{-\alpha x^2}dx = \frac{1}{4\alpha}\left(\frac{\pi}{\alpha}\right)^{1/2} \tag{B.20}$$

can be found in integral tables, and so

$$\langle x^2 \rangle = \frac{2}{(2\pi a^2)^{1/2}} \frac{(2\pi a^2)^{1/2} a^2}{2} = a^2$$

Because $\langle x \rangle = 0$, $\sigma_x^2 = \langle x^2 \rangle$, and so σ_x is given by

$$\sigma_x = a$$

The standard deviation of a normal distribution is the parameter that appears in the exponential. The standard notation for a normalized Gaussian distribution function is

$$p(x)dx = (2\pi\sigma_x^2)^{-1/2} e^{-x^2/2\sigma_x^2} dx \tag{B.21}$$

Figure B.3 shows Equation B.21 for various values of σ_x. Note that the curves become narrower and taller for smaller values of σ_x.

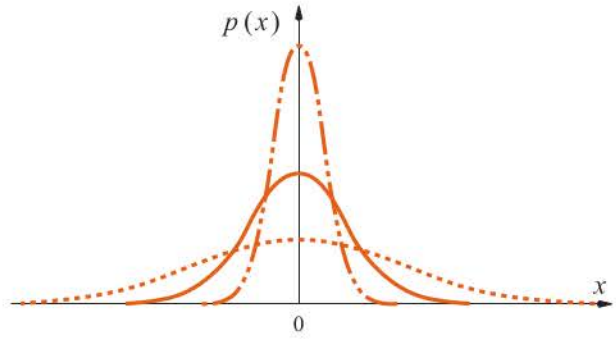

FIGURE B.3
A plot of a Gaussian distribution, $p(x)$, (Equation B.21) for three values of σ_x. The dotted curve corresponds to $\sigma_x = 2$, the solid curve to $\sigma_x = 1$, and the dash–dotted curve to $\sigma_x = 0.5$.

A more general version of a Gaussian distribution is

$$p(x)dx = (2\pi\sigma_x^2)^{-1/2} e^{-(x-\langle x \rangle)^2/2\sigma_x^2} dx \tag{B.22}$$

This expression looks like those in Figure B.3 except that the curves are centered at $x = \langle x \rangle$ rather than $x = 0$. A Gaussian distribution is one of the most important and commonly used probability distributions in all of science.

Problems

B–1. Using the following table

x	$p(x)$
−6	0.05
−2	0.15
0	0.50
1	0.10
3	0.05
4	0.10
5	0.05

calculate $\langle x \rangle$ and $\langle x^2 \rangle$ and show that $\sigma_x^2 > 0$.

B–2. A discrete probability distribution that is commonly used in statistics is the Poisson distribution

$$f_n = \frac{\lambda^n}{n!} e^{-\lambda} \qquad n = 0, 1, 2, \ldots$$

where λ is a positive constant. Prove that f_n is normalized. Evaluate $\langle n \rangle$ and $\langle n^2 \rangle$ and show that $\sigma^2 > 0$. Recall that

$$e^x = \sum_{n=0}^{\infty} \frac{x^n}{n!}$$

B–3. An important continuous distribution is the exponential distribution

$$p(x)dx = ce^{-\lambda x} dx \qquad 0 \leq x < \infty$$

Evaluate c, $\langle x \rangle$, and σ^2, and the probability that $x \geq a$.

B–4. Prove explicitly that

$$\int_{-\infty}^{\infty} e^{-\alpha x^2} dx = 2 \int_{0}^{\infty} e^{-\alpha x^2} dx$$

by breaking the integral from $-\infty$ to ∞ into one from $-\infty$ to 0 and another from 0 to ∞. Let $z = -x$ in the first integral and $z = x$ in the second to prove the above relation.

B–5. By using the procedure in Problem B–4, show explicitly that

$$\int_{-\infty}^{\infty} x e^{-\alpha x^2} dx = 0$$

B–6. Integrals of the type

$$I_n(\alpha) = \int_{-\infty}^{\infty} x^{2n} e^{-\alpha x^2} dx \qquad n = 0, 1, 2, \ldots$$

occur frequently in a number of applications. We can simply either look them up in a table of integrals or continue this problem. First, show that

$$I_n(\alpha) = 2 \int_0^{\infty} x^{2n} e^{-\alpha x^2} dx$$

The case $n = 0$ can be handled by the following trick. Show that the square of $I_0(\alpha)$ can be written in the form

$$I_0^2(\alpha) = 4 \int_0^{\infty} \int_0^{\infty} dx dy e^{-\alpha(x^2+y^2)}$$

Now convert to plane polar coordinates, letting

$$r^2 = x^2 + y^2 \qquad \text{and} \qquad dxdy = rdrd\theta$$

Show that the appropriate limits of integration are $0 \leq r < \infty$ and $0 \leq \theta \leq \pi/2$ and that

$$I_0^2(\alpha) = 4 \int_0^{\pi/2} d\theta \int_0^{\infty} dr r e^{-\alpha r^2}$$

which is elementary and gives

$$I_0^2(\alpha) = 4 \cdot \frac{\pi}{2} \cdot \frac{1}{2\alpha} = \frac{\pi}{\alpha}$$

or that

$$I_0(\alpha) = \left(\frac{\pi}{\alpha}\right)^{1/2}$$

Now prove that the $I_n(\alpha)$ may be obtained by repeated differentiation of $I_0(\alpha)$ with respect to α and, in particular, that

$$\frac{d^n I_0(\alpha)}{d\alpha^n} = (-1)^n I_n(\alpha)$$

Use this result and the fact that $I_0(\alpha) = (\pi/\alpha)^{1/2}$ to generate $I_1(\alpha)$, $I_2(\alpha)$, and so forth.

B–7. Without using a table of integrals, show that all of the odd moments of a Gaussian distribution are zero. Using the results derived in Problem B–6, calculate $\langle x^4 \rangle$ for a Gaussian distribution.

B–8. Consider a particle to be constrained to lie along a one-dimensional segment 0 to a. We will learn in the next chapter that the probability that the particle is found to lie between x

and $x + dx$ is given by

$$p(x)dx = \frac{2}{a} \sin^2 \frac{n\pi x}{a} dx$$

where $n = 1, 2, 3, \ldots$. First show that $p(x)$ is normalized. Now show that the average position of the particle along the line segment is $a/2$. Is this result physically reasonable? The integrals that you need are (*The CRC Handbook of Chemistry and Physics* or *The CRC Standard Mathematical Tables*; CRC Press: Boca Raton, FL)

$$\int \sin^2 \alpha x dx = \frac{x}{2} - \frac{\sin 2\alpha x}{4\alpha}$$

and

$$\int x \sin^2 \alpha x dx = \frac{x^2}{4} - \frac{x \sin 2\alpha x}{4\alpha} - \frac{\cos 2\alpha x}{8\alpha^2}$$

B–9. Show that $\langle x \rangle^2 = a^2/4$ and that the variance associated with the probability distribution given in Problem B–8 is given by $\left(\frac{a}{2\pi n}\right)^2 \left(\frac{\pi^2 n^2}{3} - 2\right)$. The necessary integral is (CRC tables)

$$\int x^2 \sin^2 \alpha x dx = \frac{x^3}{6} - \left(\frac{x^2}{4\alpha} - \frac{1}{8\alpha^3}\right) \sin 2\alpha x - \frac{x \cos 2\alpha x}{4\alpha^2}$$

B–10. Show that

$$\sigma_x = (\langle x^2 \rangle - \langle x \rangle^2)^{1/2}$$

for a particle in a box is less than a, the width of the box, for any value of n. If σ_x is the uncertainty in the position of the particle, could σ_x ever be larger than a?

B–11. All the definite integrals used in Problems B–8 and B–9 can be evaluated from

$$I(\beta) = \int_0^a e^{\beta x} \sin^2 \frac{n\pi x}{a} dx$$

Show that the above integrals are given by $I(0)$, $I'(0)$, and $I''(0)$, respectively, where the primes denote differentiation with respect to β. Using a table of integrals, evaluate $I(\beta)$ and then the above three integrals by differentiation.

B–12. Using the probability distribution given in Problem B–8, calculate the probability that the particle will be found between 0 and $a/2$. The necessary integral is given in Problem B–8.