15

Probability and Statistics

15.1 Probability and Thomas Bayes

The probability $P(A)$ of an outcome in a set A is the sum of the probabilities P_j of all the different (mutually exclusive) outcomes j in A

$$P(A) = \sum_{j \in A} P_j. \tag{15.1}$$

For instance, if one throws two fair dice, then the probability that the sum is 2 is $P(1,1) = 1/36$, while the probability that the sum is 3 is $P(1,2) + P(2,1) = 1/18$.

The set of all possible outcomes is called the **sample space**, and any subset of the sample space is called an **event**.

If A and B are two sets of possible outcomes, then the probability of an outcome in the **union** $A \cup B$ is the sum of the probabilities $P(A)$ and $P(B)$ minus that of their **intersection** $A \cap B$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{15.2}$$

If the outcomes are mutually exclusive, then $P(A \cap B) = 0$, and the probability of the union is the sum $P(A \cup B) = P(A) + P(B)$. The **joint probability** $P(A, B) \equiv P(A \cap B)$ is the probability of an outcome that is in both sets A and B. If the joint probability is the product $P(A \cap B) = P(A)\, P(B)$, then the outcomes in sets A and B are **statistically independent**.

The probability that a result in set B also is in set A is the **conditional probability** $P(A|B)$ or the probability of A given B

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{15.3}$$

Interchanging A and B, we get as the probability of B given A

$$P(B|A) = \frac{P(B \cap A)}{P(A)}. \tag{15.4}$$

Since $A \cap B = B \cap A$, the last two equations (15.3 & 15.4) tell us that

$$P(A \cap B) = P(B \cap A) = P(B|A)\, P(A) = P(A|B)\, P(B) \tag{15.5}$$

in which the last equality is **Bayes's theorem**

$$P(A|B) = \frac{P(B|A)\, P(A)}{P(B)}. \tag{15.6}$$

In this formula, the probability $P(A|B)$ is the **posterior** distribution of the observable A given the B data, while $P(A)$ is the **prior** distribution of A before the B data became available. The probability $P(B|A)$ of the B data given the observable A is the **likelihood** of the data B given A. (Thomas Bayes, 1702–1761).

If a set B of outcomes is contained in a union of n sets A_j that are mutually exclusive,

$$B \subset \bigcup_{j=1}^{n} A_j \quad \text{and} \quad A_i \cap A_k = \emptyset, \tag{15.7}$$

then we must sum over them

$$P(B) = \sum_{j=1}^{n} P(B|A_j)\, P(A_j). \tag{15.8}$$

If, for example, A_j were the probability of selecting an atom with Z_j protons and N_j neutrons, and if $P(B|A_j)$ were the probability that such a nucleus would decay in time t, then the probability that the nucleus of a selected atom would decay in time t would be given by a sum (15.8) over different kinds of atoms. In this case, if we replace A by A_k in the formula (15.5), then we get $P(B \cap A_k) = P(B|A_k)\, P(A_k) = P(A_k|B)\, P(B)$. This last equality and the sum (15.8) give us these forms of Bayes's theorem

$$P(A_k|B) = \frac{P(B|A_k)\, P(A_k)}{\sum_{j=1}^{N} P(B|A_j)\, P(A_j)} \tag{15.9}$$

$$P(B|A_k) = \frac{P(A_k|B)}{P(A_k)} \sum_{j=1}^{N} P(B|A_j)\, P(A_j). \tag{15.10}$$

Example 15.1 (Was the cab blue?) A cab was involved in a hit-and-run accident at night. In the city 85% of the cabs are Green and 15% are Blue.

A witness said the cab was Blue. Tests showed that the witness correctly distinguished Green and Blue cabs only 80% of the time. What is the probability that the guilty cab was Blue? (Kahneman, 2011, p.166)

The probability of a random cab's being Blue is $P(B) = 0.15$, and the probability of a random cab's being Green is $P(G) = 0.85$. The probabilities that the witness would call a Blue or a Green cab Blue are $P(wB|B) = 0.8$ and $P(wB|G) = 0.2$. So the probability $P(wB)$ that the witness said a random cab was Blue is

$$P(wB) = P(wB|B)P(B) + P(wB|G)P(G) = 0.29. \tag{15.11}$$

Now Bayes's theorem (15.6) gives the probability $P(B|wB)$ that a cab the witness said was Blue actually was Blue is

$$P(B|wB) = \frac{P(wB|B)P(B)}{P(wB)} = \frac{0.8\,(0.15)}{0.29} = 0.41 \tag{15.12}$$

which is about half the naive answer of 80%. □

Example 15.2 (Low-base-rate problem) Suppose the incidence of a rare disease in a population is $P(D) = 0.001$. Suppose a test for the disease has a **sensitivity** of 99%, that is, the probability that a carrier of the disease will test positive is $P(+|D) = 0.99$. Suppose the test also is highly **selective** with a false-positive rate of only $P(+|N) = 0.005$. Then the probability that a random person in the population would test positive is by (15.8)

$$P(+) = P(+|D)\,P(D) + P(+|N)\,P(N) = 0.00599. \tag{15.13}$$

So by Bayes's theorem (15.6), the probability that a person who tests positive actually has the disease is only

$$P(D|+) = \frac{P(+|D)\,P(D)}{P(+)} = \frac{0.99 \times 0.001}{0.00599} = 0.165 \tag{15.14}$$

and the probability that a person testing positive actually is healthy is $P(N|+) = 1 - P(D|+) = 0.835$.

Even with an excellent test, screening for rare diseases is problematic. Similarly, screening for rare behaviors, such as disloyalty in the FBI, is dicey with a good test and absurd with a poor one like a polygraph. □

Example 15.3 (Three-door problem) A prize lies behind one of three closed doors. A contestant gets to pick which door to open, but before the chosen door is opened, a door that does not lead to the prize and was not picked by the contestant swings open. Should the contestant switch and choose a different door?

A contestant who picks the wrong door and switches always wins, so $P(W|Sw, WD) = 1$, while one who picks the right door and switches never wins $P(W|Sw, RD) = 0$. Since the probability of picking the wrong door is $P(WD) = 2/3$, the probability of winning if one switches is

$$P(W|Sw) = P(W|Sw, WD)\, P(WD) + P(W|Sw, RD)\, P(RD) = 2/3. \tag{15.15}$$

The probability of picking the right door is $P(RD) = 1/3$, and the probability of winning if one picks the right door and stays put is $P(W|Sp, RD) = 1$. So the probability of winning if one stays put is

$$P(W|Sp) = P(W|Sp, RD)\, P(RD) + P(W|Sp, WD)\, P(WD) = 1/3. \tag{15.16}$$

Thus, one should switch after the door opens. □

If the set A is the interval $(x - dx/2, x + dx/2)$ of the real line, then $P(A) = P(x)\, dx$, and version (15.9) of Bayes's theorem says

$$P(x|B) = \frac{P(B|x)\, P(x)}{\int_{-\infty}^{\infty} P(B|x')\, P(x')\, dx'}. \tag{15.17}$$

Example 15.4 (A tiny poll) We ask 4 likely voters if they will vote for Nancy Pelosi, and 3 say "Yes." If the probability that a random voter will vote for her is y, then the probability that 3 in our sample of 4 will is

$$P(3|y) = 4\, y^3\, (1 - y) \tag{15.18}$$

which is the value $P_b(3, y, 4)$ of the binomial distribution (section 15.3, 15.50) for $n = 3$, $p = y$, and $N = 4$. We don't know the **prior** probability distribution $P(y)$, so we set it equal to unity on the interval $(0, 1)$. Then the continuous form of Bayes's theorem (15.17) and our cheap poll give the probability distribution of the fraction y who will vote for her as

$$\begin{aligned} P(y|3) &= \frac{P(3|y)\, P(y)}{\int_0^1 P(3|y')\, P(y')\, dy'} = \frac{P(3|y)}{\int_0^1 P(3|y')\, dy'} \\ &= \frac{4\, y^3\, (1 - y)}{\int_0^1 4\, y'^3\, (1 - y')\, dy'} = 20\, y^3\, (1 - y). \end{aligned} \tag{15.19}$$

Our best guess then for the probability that she will win the election is

$$\int_{1/2}^{1} P(y|3)\, dy = \int_{1/2}^{1} 20\, y^3\, (1 - y)\, dy = \frac{13}{16} \tag{15.20}$$

which is slightly higher than the naive estimate of 3/4. □

Figure 15.1 Chris Bishop, `https://www.youtube.com/watch?v=8FHBh_OmdsM`

Example 15.5 (Quantum mechanics) But when are two sets A_1 and A_2 of microscopic events mutually exclusive? Suppose a photon can go from a laser through slits 1 and 2 and be detected at point B. Unless we measure which slit the photon goes through, the two passages are not mutually exclusive. So we can't compute the probability $P(B)$ that the photon is detected at point B as the sum (15.8) $P(B) = P(B|A_1)\,P(A_1) + P(B|A_2)\,P(A_2)$ in which $P(A_i)$ is the probability of its going through slit i. We must use the quantum-mechanical formula $P(B) = |\langle B|A_1\rangle + \langle B|A_2\rangle|^2$ in which $\langle B|A_i\rangle$ is the amplitude for the photon to get to B through slit i. □

Example 15.6 (Transitivity) If a, b, c are numbers, then $a > b$ and $b > c$ implies that $a > c$. Such transitivity works for numbers but not always for probabilities. For instance, in a game in which the person throwing the highest number wins, the utility of the red die of Fig. 15.1 is greater than that of the green one; that of the yellow die is greater than that of the red one; that of the purple die is greater than that of the yellow one; and that of the green die is greater than that of the purple one: $1/3 + (2/3)(1/3) = 5/9$. □

15.2 Mean and Variance

In many games, N outcomes x_j can occur with probabilities P_j that sum to unity

$$\sum_{j=1}^{N} P_j = 1. \tag{15.21}$$

The **expected value** $E[x]$ of the outcome x is its **mean** μ or **average** value $\langle x \rangle = \overline{x}$

$$E[x] = \mu = \langle x \rangle = \overline{x} = \sum_{j=1}^{N} x_j \, P_j. \tag{15.22}$$

The **expected value** $E[x]$ also is variously called the **expectation** of x, the **expectation value** of x, the **mean value** of x, and the **average value** of x.

The **ℓth moment** of x is

$$E[x^\ell] = \mu_\ell = \langle x^\ell \rangle = \sum_{j=1}^{N} x_j^\ell P_j \tag{15.23}$$

and its **ℓth central moment** is

$$E[(x-\mu)^\ell] = \nu_\ell = \sum_{j=1}^{N} (x_j - \mu)^\ell P_j \tag{15.24}$$

in which $\mu_0 = \nu_0 = 1$, and $\nu_1 = 0$ (exercise 15.3).

The **variance** $V[x]$ is the second central moment ν_2

$$\begin{aligned} V[x] &\equiv E[(x - \langle x \rangle)^2] = \nu_2 \\ &= \sum_{j=1}^{N} (x_j - \langle x \rangle)^2 \; P_j = \langle x^2 \rangle - \langle x \rangle^2 \end{aligned} \tag{15.25}$$

and the **standard deviation** σ is its square root

$$\sigma = \sqrt{V[x]}. \tag{15.26}$$

If the values of x are distributed continuously according to a **probability distribution** or **density** $P(x)$ normalized to unity

$$\int P(x) \, dx = 1 \tag{15.27}$$

then the average value is

$$E[x] = \mu = \langle x \rangle = \int x\, P(x)\, dx \tag{15.28}$$

and the ℓth moment is

$$E[x^\ell] = \mu_\ell = \langle x^\ell \rangle = \int x^\ell\, P(x)\, dx. \tag{15.29}$$

The ℓth central moment is

$$E[(x-\mu)^\ell] = \nu_\ell = \int (x-\mu)^\ell\, P(x)\, dx. \tag{15.30}$$

The variance of the distribution is the second central moment

$$V[x] = \nu_2 = \int \left(x - \langle x \rangle\right)^2\, P(x)\, dx = \mu_2 - \mu^2 \tag{15.31}$$

and the standard deviation σ is its square root $\sigma = \sqrt{V[x]}$.

Many authors use $f(x)$ for the probability distribution $P(x)$ and $F(x)$ for the cumulative probability $\Pr(-\infty, x)$ of an outcome in the interval $(-\infty, x)$

$$F(x) \equiv \Pr(-\infty, x) = \int_{-\infty}^{x} P(x')\, dx' = \int_{-\infty}^{x} f(x')\, dx' \tag{15.32}$$

a function that is necessarily **monotonic**

$$F'(x) = \Pr'(-\infty, x) = f(x) = P(x) \geq 0. \tag{15.33}$$

Some mathematicians reserve the term probability **distribution** for probabilities like $\Pr(-\infty, x)$ and P_j and call a continuous distribution $P(x)$ a **probability density function** or **PDF**.

Although a probability distribution $P(x)$ is normalized (15.27), it can have **fat tails**, which are important in financial applications (Bouchaud and Potters, 2003). Fat tails can make the variance and even the **mean absolute deviation**

$$E_{\text{abs}} \equiv \int |x - \mu|\, P(x)\, dx \tag{15.34}$$

diverge.

Example 15.7 (Heisenberg's uncertainty principle) In quantum mechanics, the absolute-value squared $|\psi(x)|^2$ of a wave function $\psi(x)$ is the probability distribution $P(x) = |\psi(x)|^2$ of the position x of the particle, and $P(x)\, dx$ is the probability that the particle is found between $x - dx/2$ and $x + dx/2$. The variance $\langle (x - \langle x \rangle)^2 \rangle$ of the position operator x is written as the square $(\Delta x)^2$ of the standard deviation $\sigma = \Delta x$ which is the **uncertainty**

in the position of the particle. Similarly, the square of the uncertainty in the momentum $(\Delta p)^2$ is the variance $\langle (p - \langle p \rangle)^2 \rangle$ of the momentum.

For the wave function (4.74)

$$\psi(x) = \left(\frac{2}{\pi}\right)^{1/4} \frac{1}{\sqrt{a}}\, e^{-(x/a)^2} \tag{15.35}$$

these uncertainties are $\Delta x = a/2$ and $\Delta p = \hbar/a$. They provide a saturated example $\Delta x\, \Delta p = \hbar/2$ of Heisenberg's uncertainty principle

$$\Delta x\, \Delta p \ge \frac{\hbar}{2}. \tag{15.36}$$

□

If x and y are two random variables that occur with a **joint distribution** $P(x, y)$, then the expected value of the linear combination $ax^n y^m + bx^p y^q$ is

$$\begin{aligned} E[ax^n y^m + bx^p y^q] &= \int (ax^n y^m + bx^p y^q)\, P(x, y)\, dxdy \\ &= a \int x^n y^m\, P(x, y)\, dxdy + b \int x^p y^q\, P(x, y)\, dxdy \\ &= a\, E[x^n y^m] + b\, E[x^p y^q]. \end{aligned} \tag{15.37}$$

This result and its analog for discrete probability distributions show that **expected values are linear**.

Example 15.8 (Jensen's inequalities) A **convex** function is one that lies above its tangents:

$$f(x) \ge f(y) + (x - y) f'(y). \tag{15.38}$$

For example, e^x lies above $1 + x$ which is its tangent at $x = 0$. Multiplying both sides of the definition (15.38) by the probability distribution $P(x)$ and integrating over x with $y = \langle x \rangle$, we find that the mean value of a convex function

$$\begin{aligned} \langle f(x) \rangle = \int f(x) P(x) dx &\ge \int \left[f(\langle x \rangle) + (x - \langle x \rangle) f'(\langle x \rangle) \right] P(x) dx \\ &= \int f(\langle x \rangle)\, P(x)\, dx = f(\langle x \rangle) \end{aligned} \tag{15.39}$$

exceeds its value at $\langle x \rangle$. Equivalently, $E[f(x)] \ge f(E[x])$.

For a **concave** function, the inequalities (15.38) and (15.39) reverse, and $f(E[x]) \ge E[f(x)]$. Thus since $\log(x)$ is concave for $x > 0$, we have

$$\log(E[x]) = \log \left(\frac{1}{n} \sum_{i=1}^{n} x_i \right) \ge E[\log(x)] = \sum_{i=1}^{n} \frac{1}{n} \log(x_i). \tag{15.40}$$

Exponentiating both sides, we get the **inequality of arithmetic and geometric means**

$$\frac{1}{n}\sum_{i=1}^{n} x_i \geq \left(\prod_{i=1}^{n} x_i\right)^{1/n} \tag{15.41}$$

(Johan Jensen, 1859–1925). □

The **correlation coefficient** or **covariance** of two variables x and y that occur with a **joint distribution** $P(x, y)$ is

$$C[x, y] \equiv \int P(x, y)(x - \overline{x})(y - \overline{y})\, dxdy = \langle (x - \overline{x})(y - \overline{y})\rangle = \langle x\, y\rangle - \langle x\rangle\langle y\rangle. \tag{15.42}$$

The variables x and y are said to be **independent** if

$$P(x, y) = P(x)\, P(y). \tag{15.43}$$

Independence implies that the covariance vanishes, but $C[x, y] = 0$ does not guarantee that x and y are independent (Roe, 2001, p. 9).

The variance of $x + y$

$$\langle (x + y)^2\rangle - \langle x + y\rangle^2 = \langle x^2\rangle - \langle x\rangle^2 + \langle y^2\rangle - \langle y\rangle^2 + 2\left(\langle x\, y\rangle - \langle x\rangle\langle y\rangle\right) \tag{15.44}$$

is the sum

$$V[x + y] = V[x] + V[y] + 2\, C[x, y]. \tag{15.45}$$

It follows (exercise 15.6) that for any constants a and b the variance of $ax+by$ is

$$V[ax + by] = a^2\, V[x] + b^2\, V[y] + 2\, ab\, C[x, y]. \tag{15.46}$$

More generally (exercise 15.7), the variance of the sum $a_1x_1 + a_2x_2 + \ldots + a_Nx_N$ is

$$V[a_1x_1 + \ldots + a_Nx_N] = \sum_{j=1}^{N} a_j^2\, V[x_j] \;+ \sum_{j,k=1,j<k}^{N} 2a_ja_k\, C[x_j, x_k]. \tag{15.47}$$

If the variables x_j and x_k are independent for $j \neq k$, then their covariances vanish $C[x_j, x_k] = 0$, and the variance of the sum $a_1x_1 + \ldots + a_Nx_N$ is

$$V[a_1x_1 + \ldots + a_Nx_N] = \sum_{j=1}^{N} a_j^2\, V[x_j]. \tag{15.48}$$

15.3 Binomial distribution

If the probability of success is p on each try, then we expect that in N tries the mean number of successes will be

$$\langle n \rangle = N\, p. \tag{15.49}$$

The probability of failure on each try is $q = 1 - p$. So the probability of a particular sequence of successes and failures, such as n successes followed by $N - n$ failures is $p^n\, q^{N-n}$. There are $N!/n!\,(N-n)!$ different sequences of n successes and $N - n$ failures, all with the same probability $p^n\, q^{N-n}$. So the probability of n successes (and $N - n$ failures) in N tries is

$$P_b(n, p, N) = \frac{N!}{n!\,(N-n)!}\, p^n\, q^{N-n} = \binom{N}{n}\, p^n\, (1-p)^{N-n}. \tag{15.50}$$

This **binomial distribution** also is called **Bernoulli's distribution** (Jacob Bernoulli, 1654–1705).

The sum (5.93) of the probabilities $P_b(n, p, N)$ for $n = 0, 1, 2, \ldots, N$ is unity

$$\sum_{n=0}^{N} P_b(n, p, N) = \sum_{n=0}^{N} \binom{N}{n}\, p^n\, (1-p)^{N-n} = (p + 1 - p)^N = 1. \tag{15.51}$$

In Fig. 15.2, the probabilities $P_b(n, p, N)$ for $0 \le n \le 250$ and $p = 0.2$ are plotted for $N = 125, 250, 500$, and 1000 tries.

The mean number of successes

$$\mu = \langle n \rangle_B = \sum_{n=0}^{N} n\, P_b(n, p, N) = \sum_{n=0}^{N} n \binom{N}{n}\, p^n q^{N-n} \tag{15.52}$$

is a partial derivative with respect to p with q held fixed

$$\begin{aligned} \langle n \rangle_B &= p\, \frac{\partial}{\partial p} \sum_{n=0}^{N} \binom{N}{n}\, p^n q^{N-n} \\ &= p\, \frac{\partial}{\partial p}\, (p+q)^N = Np\, (p+q)^{N-1} = Np \end{aligned} \tag{15.53}$$

which verifies the estimate (15.49).

One may show (exercise 15.9) that the variance (15.25) of the binomial distribution is

$$V_B = \langle (n - \langle n \rangle)^2 \rangle = p\, (1-p)\, N. \tag{15.54}$$

Its standard deviation (15.26) is

$$\sigma_B = \sqrt{V_B} = \sqrt{p\,(1-p)\,N}. \tag{15.55}$$

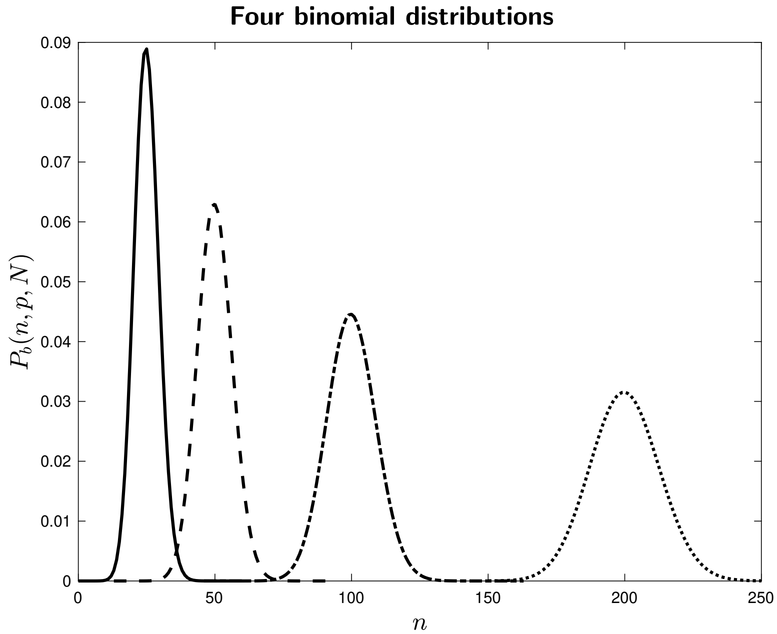

Figure 15.2 The binomial probability distribution $P_b(n, p, N)$ (15.50) is plotted here for $p = 0.2$ and $N = 125$ (solid), 250 (dashes), 500 (dot dash), and 1000 tries (dots). This chapter's codes are in Probability_and_statistics at `github.com/kevinecahill`.

The ratio of the width to the mean

$$\frac{\sigma_B}{\langle n \rangle_B} = \frac{\sqrt{p\,(1-p)\,N}}{Np} = \sqrt{\frac{1-p}{Np}} \tag{15.56}$$

decreases with N as $1/\sqrt{N}$.

Example 15.9 (Avogadro's number) A mole of gas is Avogadro's number $N_A = 6 \times 10^{23}$ of molecules. If the gas is in a cubical box, then the chance that each molecule will be in the left half of the cube is $p = 1/2$. The mean number of molecules there is $\langle n \rangle_b = pN_A = 3 \times 10^{23}$, and the uncertainty in n is $\sigma_b = \sqrt{p\,(1-p)\,N} = \sqrt{3 \times 10^{23}/4} = 3 \times 10^{11}$. So the numbers of gas molecules in the two halves of the box are equal to within $\sigma_b / \langle n \rangle_b = 10^{-12}$ or to 1 part in 10^{12}. □

Example 15.10 (Counting fluorescent molecules) Molecular biologists can insert the DNA that codes for a fluorescent protein next to the DNA that codes for a specific natural protein in the genome of a bacterium. The bacterium then will make its natural protein with the fluorescent protein attached to it, and the labeled protein will produce light of a specific color when suitably illuminated by a laser. The intensity I of the light is proportional to the number n of labeled protein molecules $I = \alpha\, n$, and one can find the constant of proportionality α by measuring the light given off as the bacteria divide. When a bacterium divides, it randomly separates the total number N of fluorescent proteins inside it into its two daughter bacteria, giving one n fluorescent molecules and the other $N - n$. The variance of the difference is

$$\langle (n - (N - n))^2 \rangle = \langle (2n - N)^2 \rangle = 4\,\langle n^2 \rangle - 4\,\langle n \rangle N + N^2. \tag{15.57}$$

The mean number (15.53) is $\langle n \rangle = pN$, and our variance formula (15.54) tells us that

$$\langle n^2 \rangle = \langle n \rangle^2 + p(1-p)N = (pN)^2 + p(1-p)N. \tag{15.58}$$

Since the probability $p = 1/2$, the variance of the difference is

$$\langle (n - (N - n))^2 \rangle = (2p - 1)^2 N^2 + 4p(1-p)N = N. \tag{15.59}$$

Thus the ratio of the variance of the difference of daughters' intensities to the intensity of the parent bacterium reveals the unknown constant of proportionality α (Phillips et al., 2012)

$$\frac{\langle (I_n - I_{N-n})^2 \rangle}{\langle I_N \rangle} = \frac{\alpha^2 \langle (n - (N - n))^2 \rangle}{\alpha \langle N \rangle} = \frac{\alpha^2 N}{\alpha N} = \alpha. \tag{15.60}$$

□

15.4 Coping with big factorials

Because $n!$ increases very rapidly with n, the rule

$$P_b(k+1, p, n) = \frac{p}{1-p}\,\frac{n-k}{k+1}\,P_b(k, p, n) \tag{15.61}$$

is helpful when n is big. But when n exceeds a few hundred, the formula (15.50) for $P_b(k, p, n)$ becomes unmanageable even in quadruple precision. One solution is to work with the logarithm of the expression of interest. The Fortran function log_gamma(x), the C function lgamma(x), the Matlab function gammaln(x), and the Python function loggamma(x) all give $\log(\Gamma(x)) = \log((x-1)!)$ for real x. Using the very tame logarithm of the gamma function, one may compute $P_b(k, p, n)$ even for $n = 10^7$ as

$$\begin{aligned} \binom{n}{k} p^k q^{n-k} = {} & \exp\big[\log\left(\Gamma(n+1)\right) - \log\left(\Gamma(n-k+1)\right) \\ & - \log\left(\Gamma(k+1)\right) + k\log p + (n-k)\log q\big]. \end{aligned} \tag{15.62}$$

Another way to cope with huge factorials is to use Stirling's formula (5.40) $n! \approx \sqrt{2\pi n}\,(n/e)^n$ or Srinivasa Ramanujan's correction (5.41) or Mermin's even more accurate approximations (5.42–5.44).

A third way to cope with the unwieldy factorials in the binomial formula $P_b(k, p, n)$ is to use its limiting forms due to Poisson and to Gauss.

15.5 Poisson's distribution

Poisson approximated the formula (15.50) for the binomial distribution $P_b(n, p, N)$ by taking the two limits $N \to \infty$ and $p = \langle n \rangle / N \to 0$ while keeping n and the product $pN = \langle n \rangle$ constant. Using Stirling's formula $n! \approx \sqrt{2\pi n}\,(n/e)^n$ (6.338) for the two huge factorials $N!$ and $(N-n)!$, we get as $n/N \to 0$ and $\langle n \rangle / N \to 0$ with $\langle n \rangle = pN$ kept fixed

$$\begin{aligned} P_b(n, p, N) &= \binom{N}{n} p^n (1-p)^{N-n} = \frac{N!}{(N-n)!}\,\frac{p^n}{n!}\,(1-p)^{N-n} \\ &\approx \sqrt{\frac{N}{N-n}}\,\left(\frac{N}{e}\right)^N \left(\frac{e}{N-n}\right)^{N-n} \frac{(pN)^n}{n!}\,(1-p)^{N-n} \qquad (15.63) \\ &\approx e^{-n}\,\left(1 - \frac{n}{N}\right)^{-N+n} \frac{\langle n \rangle^n}{n!}\,\left(1 - \frac{\langle n \rangle}{N}\right)^{N-n}. \end{aligned}$$

So using the definition $\exp(-x) = \lim_{N\to\infty} (1 - x/N)^N$ to take the limits

$$\left(1 - \frac{n}{N}\right)^{-N} \left(1 - \frac{n}{N}\right)^{n} \to e^n \quad \text{and} \quad \left(1 - \frac{\langle n \rangle}{N}\right)^{N} \left(1 - \frac{\langle n \rangle}{N}\right)^{-n} \to e^{\langle n \rangle}, \qquad (15.64)$$

we get from the binomial distribution Poisson's estimate

$$P_P(n, \langle n \rangle) = \frac{\langle n \rangle^n}{n!}\, e^{-\langle n \rangle} \qquad (15.65)$$

of the probability of n successes in a very large number N of tries, each with a tiny chance $p = \langle n \rangle / N$ of success. (Siméon-Denis Poisson, 1781–1840. Incidentally, *poisson* means *fish* and sounds like pwahsahn.)

The Poisson distribution is normalized to unity

$$\sum_{n=0}^{\infty} P_P(n, \langle n \rangle) = \sum_{n=0}^{\infty} \frac{\langle n \rangle^n}{n!}\, e^{-\langle n \rangle} = e^{\langle n \rangle}\, e^{-\langle n \rangle} = 1. \qquad (15.66)$$

Its mean μ is the parameter $\langle n \rangle = pN$ of the binomial distribution

$$\mu = \sum_{n=0}^{\infty} n \, P_P(n, \langle n \rangle) = \sum_{n=1}^{\infty} n \, \frac{\langle n \rangle^n}{n!} \, e^{-\langle n \rangle} = \langle n \rangle \sum_{n=1}^{\infty} \frac{\langle n \rangle^{(n-1)}}{(n-1)!} \, e^{-\langle n \rangle}$$
$$= \langle n \rangle \sum_{n=0}^{\infty} \frac{\langle n \rangle^n}{n!} \, e^{-\langle n \rangle} = \langle n \rangle. \tag{15.67}$$

As $N \to \infty$ and $p \to 0$ with $p\,N = \langle n \rangle$ fixed, the variance (15.54) of the binomial distribution tends to the limit

$$V_P = \lim_{\substack{N\to\infty \\ p\to 0}} V_B = \lim_{\substack{N\to\infty \\ p\to 0}} p\,(1-p)\,N = \langle n \rangle. \tag{15.68}$$

Thus the mean and the variance of a Poisson distribution are equal

$$V_P = \langle (n - \langle n \rangle)^2 \rangle = \langle n \rangle = \mu \tag{15.69}$$

as one may show directly (exercise 15.12).

Example 15.11 (Accuracy of Poisson's distribution) If $p = 0.0001$ and $N = 10,000$, then $\langle n \rangle = 1$ and Poisson's approximation to the probability that $n = 2$ is $1/2e$. The exact binomial probability (15.62) and Poisson's estimate are $P_b(2, 0.01, 1000) = 0.18395$ and $P_P(2, 1) = 0.18394$. □

Example 15.12 (Coherent states) The **coherent state** $|\alpha\rangle$ introduced in equation (1.314)

$$|\alpha\rangle = e^{-|\alpha|^2/2} e^{\alpha a^\dagger} |0\rangle = e^{-|\alpha|^2/2} \sum_{n=0}^{\infty} \frac{\alpha^n}{\sqrt{n!}} |n\rangle \tag{15.70}$$

is an eigenstate $a|\alpha\rangle = \alpha|\alpha\rangle$ of the annihilation operator a with eigenvalue α. The probability $P(n)$ of finding n quanta in the state $|\alpha\rangle$ is the square of the absolute value of the inner product $\langle n|\alpha\rangle$

$$P(n) = |\langle n|\alpha\rangle|^2 = \frac{|\alpha|^{2n}}{n!} \, e^{-|\alpha|^2} \tag{15.71}$$

which is a Poisson distribution $P(n) = P_P(n, |\alpha|^2)$ with mean and variance $\mu = \langle n \rangle = V(\alpha) = |\alpha|^2$. □

Example 15.13 (Radiation and cancer) If a cell becomes cancerous only after being hit N times by ionizing radiation, then the probability of cancer $P(\langle n\rangle)_N$ rises with the dose or mean number $\langle n\rangle$ of hits per cell as

$$P(\langle n\rangle)_N = \sum_{n=N}^{\infty} \frac{\langle n\rangle^n}{n!}\, e^{-\langle n\rangle} \tag{15.72}$$

or $P(\langle n\rangle)_N \approx \langle n\rangle^N/N!$ for $\langle n\rangle \ll 1$. As illustrated in Fig. 15.3, although the incidence of cancer $P(\langle n\rangle)_N$ rises linearly (solid) with the dose $\langle n\rangle$ of radiation if a single hit, $N = 1$, can cause a cell to become cancerous, it rises more slowly if the threshold for cancer is $N = 2$ (dot dash), 3 (dashes), or 4 (dots). Most mutations are harmless. The mean number N of harmful mutations that occur before a cell becomes cancerous is about 4, but N varies with the affected organ from 1 to 10 (Martincorena et al., 2017).

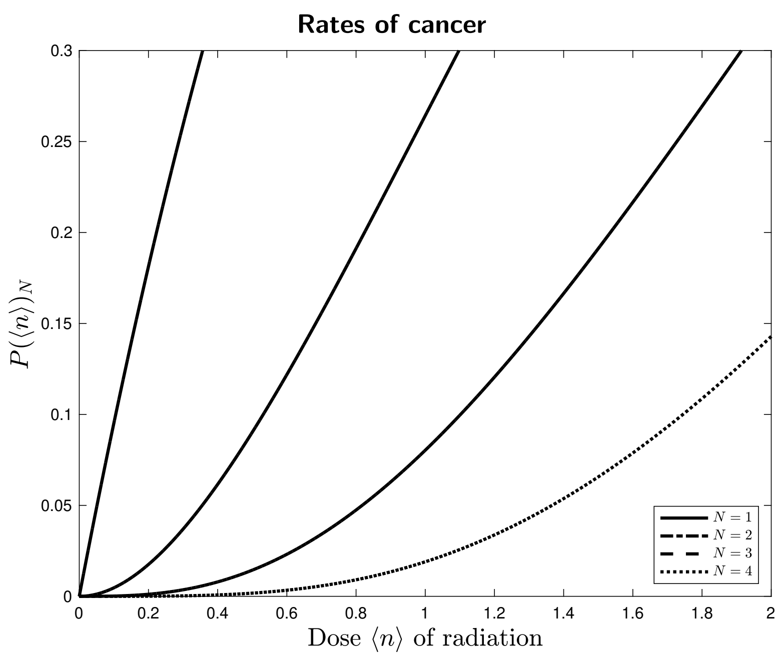

Figure 15.3 The incidence of cancer $P(\langle n\rangle)_N$ rises linearly (solid) with the dose or mean number $\langle n\rangle$ of times a cell is struck by ionizing radiation if a single hit, $N = 1$ (solid), can cause a cell to become cancerous. It rises more slowly if the threshold for cancer is $N = 2$ (dot dash), 3 (dashes), or 4 (dots) hits.

□

15.6 Gauss's distribution

Gauss considered the binomial distribution in the limits $n \to \infty$ and $N \to \infty$ with the probability p fixed. In this limit, all three factorials are huge, and we may apply Stirling's formula to each of them

$$\begin{aligned} P_b(n,p,N) &= \frac{N!}{n!\,(N-n)!}\, p^n\, q^{N-n} \\ &\approx \sqrt{\frac{N}{2\pi n(N-n)}} \left(\frac{N}{e}\right)^N \left(\frac{e}{n}\right)^n \left(\frac{e}{N-n}\right)^{N-n} p^n\, q^{N-n} \\ &= \sqrt{\frac{N}{2\pi n(N-n)}} \left(\frac{pN}{n}\right)^n \left(\frac{qN}{N-n}\right)^{N-n}. \end{aligned} \tag{15.73}$$

This probability $P_b(n,p,N)$ is tiny unless n is near pN which means that $n \approx pN$ and $N-n \approx (1-p)N = qN$ are comparable. So we set $y = n - pN$ and treat y/N as small. Since $n = pN + y$ and $N - n = (1-p)N + pN - n = qN - y$, we can write the square root as

$$\begin{aligned} \sqrt{\frac{N}{2\pi\, n\,(N-n)}} &= \frac{1}{\sqrt{2\pi\, N\,[(pN+y)/N]\,[(qN-y)/N]}} \\ &= \frac{1}{\sqrt{2\pi\, pqN\,(1+y/pN)\,(1-y/qN)}}. \end{aligned} \tag{15.74}$$

Because y remains finite as $N \to \infty$, the limit of the square root is

$$\lim_{N\to\infty} \sqrt{\frac{N}{2\pi\, n\,(N-n)}} = \frac{1}{\sqrt{2\pi\, pqN}}. \tag{15.75}$$

Substituting $pN + y$ for n and $qN - y$ for $N - n$ in (15.73), we find

$$\begin{aligned} P_b(n,p,N) &\approx \frac{1}{\sqrt{2\pi\, pqN}} \left(\frac{pN}{pN+y}\right)^{pN+y} \left(\frac{qN}{qN-y}\right)^{qN-y} \\ &= \frac{1}{\sqrt{2\pi\, pqN}} \left(1+\frac{y}{pN}\right)^{-(pN+y)} \left(1-\frac{y}{qN}\right)^{-(qN-y)} \end{aligned} \tag{15.76}$$

which implies

$$\log\left[P_b(n,p,N)\sqrt{2\pi\, pqN}\right] \approx -(pN+y)\log\left[1+\frac{y}{pN}\right] - (qN-y)\log\left[1-\frac{y}{qN}\right]. \tag{15.77}$$

The first two terms of the power series (5.101) for $\log(1+\epsilon)$ are

$$\log(1+\epsilon) \approx \epsilon - \frac{1}{2}\epsilon^2. \tag{15.78}$$

So applying this expansion to the two logarithms and using the relation $1/p + 1/q = (p+q)/pq = 1/pq$, we get

$$\log\left(P_b(n,p,N)\sqrt{2\pi\, pqN}\right) \approx -(pN+y)\left[\frac{y}{pN} - \frac{1}{2}\left(\frac{y}{pN}\right)^2\right] - (qN-y)\left[-\frac{y}{qN} - \frac{1}{2}\left(\frac{y}{qN}\right)^2\right] \approx -\frac{y^2}{2pqN}. \tag{15.79}$$

Remembering that $y = n - pN$, we get Gauss's approximation to the binomial probability distribution

$$P_{bG}(n,p,N) = \frac{1}{\sqrt{2\pi pqN}} \exp\left(-\frac{(n-pN)^2}{2pqN}\right). \tag{15.80}$$

This probability distribution is normalized

$$\sum_{n=0}^{\infty} \frac{1}{\sqrt{2\pi pqN}} \exp\left(-\frac{(n-pN)^2}{2pqN}\right) = 1 \tag{15.81}$$

almost exactly for $pN > 100$.

Extending the integer n to a continuous variable x, we have

$$P_G(x,p,N) = \frac{1}{\sqrt{2\pi pqN}} \exp\left(-\frac{(x-pN)^2}{2pqN}\right) \tag{15.82}$$

which on the real line $(-\infty, \infty)$ is (exercise 15.13) a normalized probability distribution with mean $\langle x \rangle = \mu = pN$ and variance $\langle (x-\mu)^2 \rangle = \sigma^2 = pqN$. Replacing pN by μ and pqN by σ^2, we get the Standard form of **Gauss's distribution**

$$P_G(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{15.83}$$

This distribution occurs so often in mathematics and in nature that it is often called **the normal distribution**. Its odd central moments all vanish $\nu_{2n+1} = 0$, and its even ones are $\nu_{2n} = (2n-1)!!\,\sigma^{2n}$ (exercise 15.15).

Example 15.14 (Acccuracy of Gauss's distribution) If $p = 0.1$ and $N = 10^4$, then Gauss's approximation to the probability that $n = 10^3$ is $1/(30\sqrt{2\pi})$. The exact binomial probability (15.62) is $P_b(10^3, 0.1, 10^4) = 0.013297$, and Gauss's estimate is $P_G(10^3, 0.1, 10^4) = 0.013298$. □

Example 15.15 (Single-molecule super-resolution microscopy) If the wavelength of visible light were a nanometer, microscopes would yield much sharper images. Each photon from a (single-molecule) fluorophore entering the lens of a microscope would follow ray optics and be focused within

Super-resolution microscopy

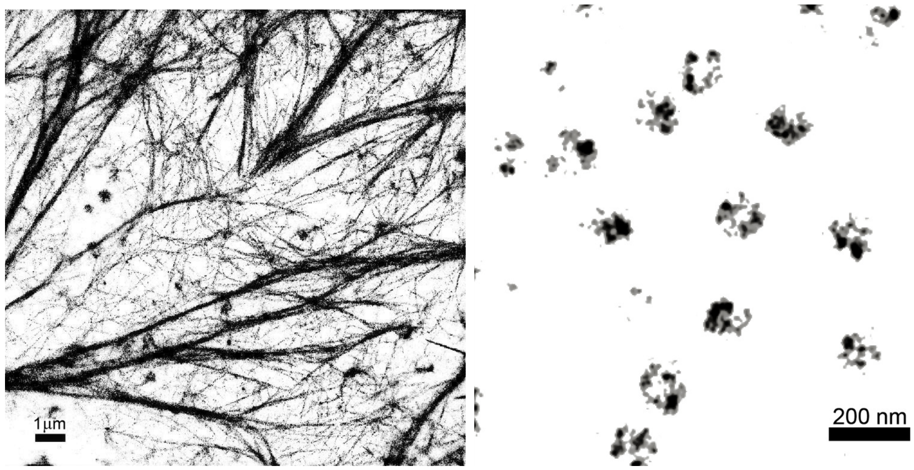

Figure 15.4 Left: dSTORM image of actin filaments in a HeLa cell, courtesy of Hanieh Mazloom Farsibaf and Keith Lidke, University of New Mexico. Right: images of the nuclear-pore-complex protein (Nup98) in the 120 nm-wide nuclear pores of a COS-7 cell, courtesy of Donghan Ma and Fang Huang, Purdue University.

a tiny circle of about a nanometer on a detector. Instead, a photon arrives not at $\boldsymbol{x} = (x_1, x_2)$ but at $\boldsymbol{y_i} = (y_{1i}, y_{2i})$ with gaussian probability

$$P(\boldsymbol{y_i}) = \frac{1}{2\pi\sigma^2}\, e^{-(\boldsymbol{y_i}-\boldsymbol{x})^2/2\sigma^2} \tag{15.84}$$

where $\sigma \approx 150$ nm is about a quarter of a wavelength. What to do?

In the **centroid** method, one collects $N \approx 500$ points $\boldsymbol{y_i}$ and finds the point $\boldsymbol{x}$ that maximizes the joint probability of the N image points

$$P = \prod_{i=1}^{N} P(\boldsymbol{y_i}) = d^N \prod_{i=1}^{N} e^{-(\boldsymbol{y_i}-\boldsymbol{x})^2/(2\sigma^2)} = d^N \, \exp\left[-\sum_{i=1}^{N}(\boldsymbol{y_i}-\boldsymbol{x})^2/(2\sigma^2)\right] \tag{15.85}$$

where $d = 1/2\pi\sigma^2$ by solving for $k = 1$ and 2 the equations

$$\frac{\partial P}{\partial x_k} = 0 = P\,\frac{\partial}{\partial x_k}\left[-\sum_{i=1}^{N}(\boldsymbol{y_i}-\boldsymbol{x})^2/(2\sigma^2)\right] = \frac{P}{\sigma^2}\sum_{i=1}^{N}(y_{ik}-x_k)\,. \tag{15.86}$$

This **maximum-likelihood** estimate of the image point $\boldsymbol{x}$ is the average of the observed points $\boldsymbol{y_i}$

$$\boldsymbol{x} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{y_i}. \tag{15.87}$$

This method is an improvement, but it is biased by auto-fluorescence and out-of-focus fluorophores. Fang Huang and Keith Lidke use **direct stochastic optical reconstruction microscopy** (dSTORM) to locate the image point $\boldsymbol{x}$ of the fluorophore in ways that account for the finite accuracy of their pixilated detector and the randomness of photo-detection (Smith et al., 2010; Huang et al., 2011).

Actin filaments are double helices of the protein actin some 5–9 nm wide. They occur throughout a eukaryotic cell but are concentrated near its surface and determine its shape. Together with tubulin and intermediate filaments, they form a cell's cytoskeleton. The double membrane of a cell's nucleus is studded with 1000 nuclear pore complexes each of which regulates and facilitates the translocation of 1000 molecules per second. Figure 15.4 shows dSTORM images of actin filaments in a HeLa cell (left) and of the nuclear-pore-complex protein (Nup98) in the nuclear pores of a COS-7 cell (right). The finite size of the fluorophore and the motion of the molecules of living cells limit dSTORM's improvement in resolution to a factor of 10 to 20. □

15.7 The error function erf

The probability that a random variable x distributed according to Gauss's distribution (15.83) has a value between $\mu - \delta$ and $\mu + \delta$ is

$$\begin{aligned} P(|x-\mu| < \delta) &= \int_{\mu-\delta}^{\mu+\delta} P_G(x,\mu,\sigma)\,dx = \frac{1}{\sigma\sqrt{2\pi}}\int_{\mu-\delta}^{\mu+\delta} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)dx \\ &= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\delta}^{\delta} \exp\left(-\frac{x^2}{2\sigma^2}\right)dx = \frac{2}{\sqrt{\pi}}\int_0^{\delta/\sigma\sqrt{2}} e^{-t^2}\,dt. \end{aligned} \tag{15.88}$$

The last integral is the error function

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2}\,dt. \tag{15.89}$$

The probability that x lies within δ of the mean μ is

$$P(|x-\mu| < \delta) = \operatorname{erf}\left(\frac{\delta}{\sigma\sqrt{2}}\right). \tag{15.90}$$

In particular, the probabilities that x falls within one, two, or three standard deviations of μ are

$$\begin{aligned} P(|x-\mu| < \sigma) &= \operatorname{erf}(1/\sqrt{2}) = 0.6827 \\ P(|x-\mu| < 2\sigma) &= \operatorname{erf}(2/\sqrt{2}) = 0.9545 \\ P(|x-\mu| < 3\sigma) &= \operatorname{erf}(3/\sqrt{2}) = 0.9973. \end{aligned} \tag{15.91}$$

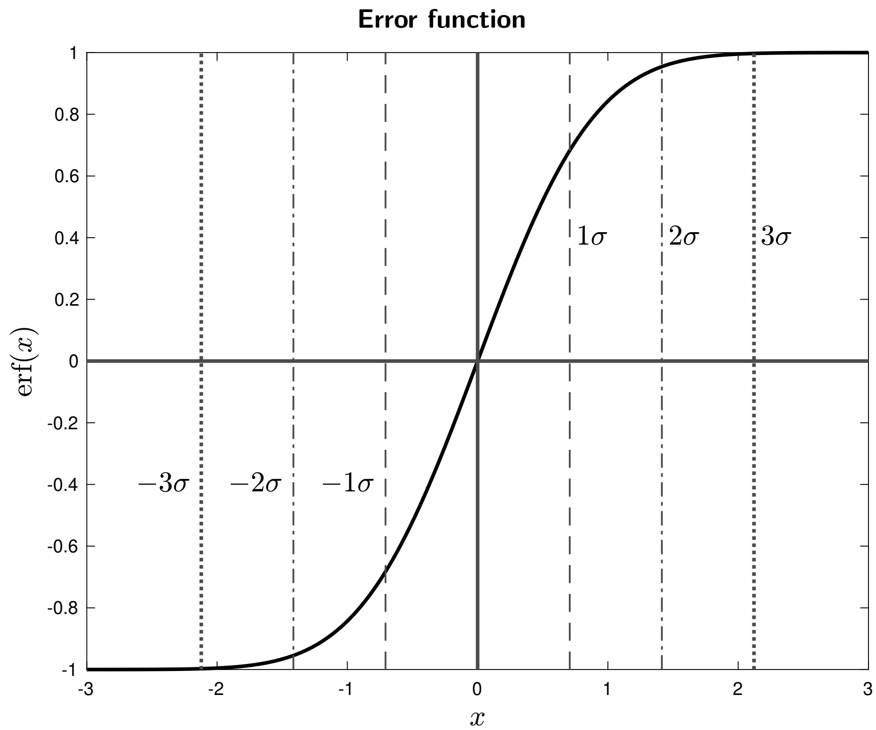

Figure 15.5 The error function $\mathrm{erf}\,(x)$ is plotted for $0 < x < 2.5$. The vertical lines are at $x = \pm\delta/(\sigma\sqrt{2})$ for $\delta = \sigma$, 2σ, and 3σ with $\sigma = 1/\sqrt{2}$.

The error function $\mathrm{erf}\,(x)$ is plotted in Fig. 15.5 in which the vertical lines are at $x = \delta/(\sigma\sqrt{2})$ for $\delta = \sigma$, 2σ, and 3σ.

The probability that x falls between a and b is (exercise 15.16)

$$P(a < x < b) = \frac{1}{2}\left[\mathrm{erf}\left(\frac{b-\mu}{\sigma\sqrt{2}}\right) - \mathrm{erf}\left(\frac{a-\mu}{\sigma\sqrt{2}}\right)\right]. \tag{15.92}$$

In particular, the cumulative probability $P(-\infty, x)$ that the random variable is less than x is for $\mu = 0$ and $\sigma = 1$

$$P(-\infty, x) = \frac{1}{2}\left[\mathrm{erf}\left(\frac{x}{\sqrt{2}}\right) - \mathrm{erf}\left(\frac{-\infty}{\sqrt{2}}\right)\right] = \frac{1}{2}\left[\mathrm{erf}\left(\frac{x}{\sqrt{2}}\right) + 1\right]. \tag{15.93}$$

The complement erfc of the error function is defined as

$$\mathrm{erfc}\,(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2}dt = 1 - \mathrm{erf}\,(x) \tag{15.94}$$

and is numerically useful for large x where round-off errors may occur in subtracting erf(x) from unity. Both erf and erfc are intrinsic functions in Fortran available without any effort on the part of the programmer.

Example 15.16 (Summing Binomial Probabilities) To add up several binomial probabilities when the factorials in $P_b(n, p, N)$ are too big to handle, we first use Gauss's approximation (15.80)

$$P_b(n, p, N) = \frac{N!}{n!\,(N-n)!}\, p^n\, q^{N-n} \approx \frac{1}{\sqrt{2\pi pqN}}\, \exp\left(-\frac{(n-pN)^2}{2pqN}\right). \tag{15.95}$$

Then using (15.92) with $\mu = pN$, we find (exercise 15.14)

$$P_b(n, p, N) \approx \frac{1}{2}\left[\text{erf}\left(\frac{n+\frac{1}{2}-pN}{\sqrt{2pqN}}\right) - \text{erf}\left(\frac{n-\frac{1}{2}-pN}{\sqrt{2pqN}}\right)\right] \tag{15.96}$$

which we can sum over the integer n to get

$$\sum_{n=n_1}^{n_2} P_b(n, p, N) \approx \frac{1}{2}\left[\text{erf}\left(\frac{n_2+\frac{1}{2}-pN}{\sqrt{2pqN}}\right) - \text{erf}\left(\frac{n_1-\frac{1}{2}-pN}{\sqrt{2pqN}}\right)\right] \tag{15.97}$$

which is easy to evaluate. □

Example 15.17 (Polls) Suppose in a poll of 1000 likely voters, 600 have said they would vote for Nancy Pelosi. Repeating the analysis of example 15.4, we see that if the probability that a random voter will vote for her is y, then the probability that 600 in our sample of 1000 will is by (15.95)

$$\begin{aligned} P(600|y) = P_b(600, y) &= \binom{1000}{600} y^{600}\, (1-y)^{400} \\ &\approx \frac{1}{10\sqrt{20\pi y(1-y)}}\, \exp\left(-\frac{20(3-5y)^2}{y(1-y)}\right). \end{aligned} \tag{15.98}$$

So if we conservatively assume that the unknown probability density $P(y)$ that a random voter will vote for her is an unknown constant which cancels, then the probability density that a random voter will vote for her, given that 600 have, is

$$\begin{aligned} P(y|600) &= \frac{P(600|y)P(y)}{\int_0^1 P(600, y')\, P(y')\, dy'} = \frac{P(600|y)}{\int_0^1 P(600, y')\, dy'} \\ &= \frac{[y(1-y)]^{-1/2}\, \exp\left(-\frac{20(3-5y)^2}{y(1-y)}\right)}{\int_0^1 [y'(1-y')]^{-1/2}\, \exp\left(-\frac{20(3-5y')^2}{y'(1-y')}\right)\, dy'}. \end{aligned} \tag{15.99}$$

So we estimate the probability that $y > 0.5$ as the ratio of the integrals

$$P(y > 0.5) \approx \frac{\int_{1/2}^{1} [y(1-y)]^{-1/2} \exp\left(-\frac{20(3-5y)^2}{y(1-y)}\right) dy}{\int_{0}^{1} [y(1-y)]^{-1/2} \exp\left(-\frac{20(3-5y)^2}{y(1-y)}\right) dy}. \tag{15.100}$$

The Mathematica script ratio.nb gives $P(y > 1/2) \approx 0.999999999873$.

The normalized probability distribution (15.100) is negligible except for y near 3/5 (exercise 15.17), where it is approximately Gauss's distribution

$$P(y|600) \approx \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-3/5)^2}{2\sigma^2}\right) \tag{15.101}$$

with mean $\mu = 3/5$ and variance $\sigma^2 = 3/12500 = 2.4 \times 10^{-4}$. The probability that $y > 1/2$ then is by (15.92) approximately

$$\begin{aligned} P(y > 1/2) &\approx \frac{1}{2}\left[\operatorname{erf}\left(\frac{1-\mu}{\sigma\sqrt{2}}\right) - \operatorname{erf}\left(\frac{1/2-\mu}{\sigma\sqrt{2}}\right)\right] \\ &= 0.999999999946. \end{aligned} \tag{15.102}$$

□

15.8 Error analysis

The mean value $\bar{f} = \langle f \rangle$ of a smooth function $f(x)$ of a random variable x is

$$\begin{aligned} \bar{f} &= \int f(x)\, P(x)\, dx \\ &\approx \int \left[f(\mu) + (x-\mu) f'(\mu) + \frac{1}{2}(x-\mu)^2 f''(\mu)\right] P(x)\, dx \\ &= f(\mu) + \frac{1}{2}\sigma^2 f''(\mu) \end{aligned} \tag{15.103}$$

as long as the higher central moments ν_n and the higher derivatives $f^{(n)}(\mu)$ are small. The mean value of f^2 then is

$$\begin{aligned}
\langle f^2 \rangle &= \int f^2(x)\, P(x)\, dx \\
&\approx \int \left[f(\mu) + (x-\mu) f'(\mu) + \frac{1}{2}(x-\mu)^2 f''(\mu) \right]^2 \, P(x)\, dx \\
&\approx \int \left[f^2(\mu) + (x-\mu)^2 f'^{\,2}(\mu) + (x-\mu)^2 f(\mu) f''(\mu) \right] \, P(x)\, dx \\
&= f^2(\mu) + \sigma^2 \, f'^{\,2}(\mu) + \sigma^2 \, f(\mu)\, f''(\mu).
\end{aligned} \tag{15.104}$$

Subtraction of $\bar{f}^2$ gives the variance of the variable $f(x)$

$$\sigma_f^2 = \langle \left(f - \bar{f} \right)^2 \rangle = \langle f^2 \rangle - \bar{f}^{\,2} \approx \sigma^2 \, f'^{\,2}(\mu). \tag{15.105}$$

A similar formula gives the variance of a smooth function $f(x_1, \dots, x_n)$ of several independent variables $x_1, \dots, x_n$ as

$$\sigma_f^2 = \langle \left(f - \bar{f} \right)^2 \rangle = \langle f^2 \rangle - \bar{f}^{\,2} \approx \sum_{i=1}^{n} \sigma_i^2 \, \left(\frac{\partial f(x)}{\partial x_i} \right)^2 \Bigg|_{x=\bar{x}} \tag{15.106}$$

in which $\bar{x}$ is the vector $(\mu_1, \dots, \mu_n)$ of mean values, and $\sigma_i^2 = \langle (x_i - \mu_i)^2 \rangle$ is the variance of x_i.

This formula (15.106) implies that the variance of a sum $f(x, y) = c\,x + d\,y$ is

$$\sigma_{cx+dy}^2 = c^2 \sigma_x^2 + d^2 \sigma_y^2. \tag{15.107}$$

Similarly, the variance formula (15.106) gives as the variance of a product $f(x, y) = x\,y$

$$\sigma_{xy}^2 = \sigma_x^2 \mu_y^2 + \sigma_y^2 \mu_x^2 = \mu_x^2 \mu_y^2 \left(\frac{\sigma_x^2}{\mu_x^2} + \frac{\sigma_y^2}{\mu_y^2} \right) \tag{15.108}$$

and as the variance of a ratio $f(x, y) = x/y$

$$\sigma_{x/y}^2 = \frac{\sigma_x^2}{\mu_y^2} + \sigma_y^2 \frac{\mu_x^2}{\mu_y^4} = \frac{\mu_x^2}{\mu_y^2} \left(\frac{\sigma_x^2}{\mu_x^2} + \frac{\sigma_y^2}{\mu_y^2} \right). \tag{15.109}$$

The variance of a power $f(x) = x^a$ follows from the variance (15.105) of a function of a single variable

$$\sigma_{x^a}^2 = \sigma_x^2 \left(a \mu_x^{a-1} \right)^2. \tag{15.110}$$

In general, the standard deviation σ is the square root of the variance σ^2.

Example 15.18 (Photon density) The 2009 COBE/FIRAS measurement of the temperature of the cosmic microwave background (CMB) radiation is $T_0 = 2.7255 \pm 0.0006$ K. The mass density (5.110) of these photons is

$$\rho_\gamma = \sigma \frac{8\pi^5 \, (k_{\rm B} T_0)^4}{15 h^3 c^5} = 4.6451 \times 10^{-31} \text{ kg m}^{-3}. \tag{15.111}$$

Our formula (15.110) for the variance of a power says that the standard deviation σ_ρ of the photon density is its temperature derivative times the standard deviation σ_T of the temperature

$$\sigma_\rho = \rho_\gamma \, \frac{4\sigma_T}{T_0} = 0.00088 \, \rho_\gamma. \tag{15.112}$$

So the probability that the photon mass density lies within the range

$$\rho_\gamma = \ (4.6451 \pm 0.0041) \times 10^{-31} \text{ kg m}^{-3} \tag{15.113}$$

is 0.68. □

15.9 Maxwell-Boltzmann distribution

It is a small jump from Gauss's distribution (15.83) to the Maxwell-Boltzmann distribution of velocities of molecules in a gas. We start in one dimension and focus on a single molecule that is being knocked forward and backward with equal probabilities by other molecules. If each tiny hit increases or decreases its speed by dv, then after n hits from behind and $N - n$ hits from in front, the speed v_x of a molecule initially at rest would be

$$v_x = n dv - (N - n) dv = (2n - N) dv. \tag{15.114}$$

The probability of this speed is given by Gauss's approximation (15.80) to the binomial distribution $P_b(n, \frac{1}{2}, N)$ as

$$P_{bG}(n, \tfrac{1}{2}, N) = \sqrt{\frac{2}{\pi N}} \exp\left(- \frac{(2n - N)^2}{2N} \right) = \sqrt{\frac{2}{\pi N}} \exp\left(- \frac{v_x^2}{2N dv^2} \right). \tag{15.115}$$

In this formula, the product $N dv^2$ is the variance $\sigma_{v_x}^2$ which is the mean value $\langle v_x^2 \rangle$ because $\langle v_x \rangle = 0$. Kinetic theory says that this variance $\sigma_{v_x}^2 = \langle v_x^2 \rangle$ is $\langle v_x^2 \rangle = kT/m$ in which m is the mass of the molecule, k Boltzmann's constant, and T the temperature. So the probability of the molecule's having velocity v_x is the Maxwell-Boltzmann distribution

$$P_G(v_x) = \frac{1}{\sigma_v\sqrt{2\pi}} \exp\left(-\frac{v_x^2}{2\sigma_v^2}\right) = \sqrt{\frac{m}{2\pi kT}} \exp\left(-\frac{mv_x^2}{2kT}\right) \tag{15.116}$$

when normalized over the line $-\infty < v_x < \infty$.

In three space dimensions, the Maxwell-Boltzmann distribution $P_{MB}(\boldsymbol{v})$ is the product

$$P_{MB}(\boldsymbol{v})d^3v = P_G(v_x)\, P_G(v_y)\, P_G(v_z)d^3v = \left(\frac{m}{2\pi kT}\right)^{3/2} e^{-\frac{1}{2}m\boldsymbol{v}^2/(kT)} 4\pi v^2 dv. \tag{15.117}$$

The mean value of the velocity of a Maxwell-Boltzmann gas vanishes

$$\langle \boldsymbol{v} \rangle = \int \boldsymbol{v}\, P_{MB}(\boldsymbol{v})d^3v = \boldsymbol{0} \tag{15.118}$$

but the mean value of the square of the velocity $\boldsymbol{v}^2 = \boldsymbol{v}\cdot\boldsymbol{v}$ is the sum of the three variances $\sigma_x^2 = \sigma_y^2 = \sigma_z^2 = kT/m$

$$\langle v^2 \rangle = V[\boldsymbol{v}^2] = \int \boldsymbol{v}^2\, P_{MB}(\boldsymbol{v})\, d^3v = 3kT/m \tag{15.119}$$

which is the familiar statement

$$\frac{1}{2}m\langle v^2 \rangle = \frac{3}{2}\, kT \tag{15.120}$$

that each degree of freedom gets $kT/2$ of energy.

15.10 Fermi-Dirac and Bose-Einstein distributions

The commutation and anticommutation relations (11.143)

$$\psi_s(t,\boldsymbol{x})\psi_{s'}(t,\boldsymbol{x}') - (-1)^{2j}\psi_{s'}(t,\boldsymbol{x}')\psi_s(t,\boldsymbol{x}) = 0 \tag{15.121}$$

of Bose fields $(-1)^{2j} = 1$ and of Fermi fields $(-1)^{2j} = -1$ determine the statistics of bosons and fermions.

One can put any number N_n of noninteracting bosons, such as photons or gravitons, into any state $|n\rangle$ of energy E_n. The energy of that state is N_nE_n, and the states $|n, N_n\rangle$ form an independent thermodynamical system for each state $|n\rangle$. The grand canonical ensemble (1.425) gives the probability of the state $|n, N_n\rangle$ as

$$\rho_{n,N_n} = \langle n, N_n|\rho|n, N_n\rangle = \frac{e^{-\beta(E_n-\mu)N_n}}{\sum_{N_n}\langle n, N_n|\rho|n, N_n\rangle}. \tag{15.122}$$

For each state $|n\rangle$, the partition function is a geometric sum

$$Z(\beta, \mu, n) = \sum_{N_n=0}^{\infty} e^{-\beta(E_n-\mu)N_n} = \sum_{N_n=0}^{\infty} \left(e^{-\beta(E_n-\mu)}\right)^{N_n} = \frac{1}{1-e^{-\beta(E_n-\mu)}}. \tag{15.123}$$

So the probability of the state $|n, N_n\rangle$ is

$$\rho_{n,N_n} = \frac{e^{-\beta(E_n-\mu)N_n}}{1-e^{-\beta(E_n-\mu)}}, \tag{15.124}$$

and the mean number of bosons in the state $|n\rangle$ is

$$\langle N_n \rangle = \frac{1}{\beta}\frac{\partial}{\partial\mu}\log Z(\beta, \mu, n) = \frac{1}{e^{\beta(E_n-\mu)}-1}. \tag{15.125}$$

One can put at most one fermion into a given state $|n\rangle$. If like neutrinos the fermions don't interact, then the states $|n, 0\rangle$ and $|n, 1\rangle$ form an independent thermodynamical system for each state $|n\rangle$. So for noninteracting fermions, the partition function is the sum of only two terms

$$Z(\beta, \mu, n) = \sum_{N_n=0}^{1} e^{-\beta(E_n-\mu)N_n} = 1 + e^{-\beta(E_n-\mu)}, \tag{15.126}$$

and the probability of the state $|n, N_n\rangle$ is

$$\rho_{n,N_n} = \frac{e^{-\beta(E_n-\mu)N_n}}{1+e^{-\beta(E_n-\mu)}}. \tag{15.127}$$

So the mean number of fermions in the state $|n\rangle$ is

$$\langle N_n \rangle = \; = \frac{1}{e^{\beta(E_n-\mu)}+1}. \tag{15.128}$$

15.11 Diffusion

We may apply the same reasoning as in the preceding section (15.9) to the diffusion of a gas of particles treated as a random walk with step size dx. In one dimension, after n steps forward and $N-n$ steps backward, a particle starting at $x=0$ is at $x=(2n-N)dx$. Thus as in (15.115), the probability of being at x is given by Gauss's approximation (15.80) to the binomial distribution $P_b(n, \frac{1}{2}, N)$ as

$$P_{bG}(n, \tfrac{1}{2}, N) = \sqrt{\frac{2}{\pi N}}\exp\left(-\frac{(2n-N)^2}{2N}\right) = \sqrt{\frac{2}{\pi N}}\exp\left(-\frac{x^2}{2N dx^2}\right). \tag{15.129}$$

In terms of the diffusion constant

$$D = \frac{N dx^2}{2t} \tag{15.130}$$

this distribution is

$$P_G(x) = \left(\frac{1}{4\pi Dt}\right)^{1/2} \exp\left(-\frac{x^2}{4Dt}\right) \tag{15.131}$$

when normalized to unity on $(-\infty, \infty)$.

In three dimensions, this gaussian distribution is the product

$$P(\boldsymbol{r},t) = P_G(x)\, P_G(y)\, P_G(z) = \left(\frac{1}{4\pi Dt}\right)^{3/2} \exp\left(-\frac{\boldsymbol{r}^2}{4Dt}\right). \tag{15.132}$$

The variance $\sigma^2 = 2Dt$ gives the average of the squared displacement of each of the three coordinates. Thus the mean of the squared displacement $\langle \boldsymbol{r}^2 \rangle$ rises **linearly** with the time as

$$\langle \boldsymbol{r}^2 \rangle = V[\boldsymbol{r}] = 3\sigma^2 = \int \boldsymbol{r}^2\, P(\boldsymbol{r},t)\, d^3r = 6\, D\, t. \tag{15.133}$$

The distribution $P(\boldsymbol{r},t)$ satisfies the **diffusion equation**

$$\dot{P}(\boldsymbol{r},t) = D\, \nabla^2 P(\boldsymbol{r},t) \tag{15.134}$$

in which the dot means time derivative.

15.12 Langevin's theory of brownian motion

Einstein made the first theory of brownian motion in 1905, but Langevin's approach (Langevin, 1908) is simpler. A tiny particle of colloidal size and mass m in a fluid is buffeted by a force $\boldsymbol{F}(t)$ due to the 10^{21} collisions per second it suffers with the molecules of the surrounding fluid. Its equation of motion is

$$m\, \frac{d\boldsymbol{v}(t)}{dt} = \boldsymbol{F}(t). \tag{15.135}$$

Langevin suggested that the force $\boldsymbol{F}(t)$ is the sum of a viscous drag $-\boldsymbol{v}(t)/B$ and a rapidly fluctuating part $\boldsymbol{f}(t)$

$$\boldsymbol{F}(t) = -\, \boldsymbol{v}(t)/B + \boldsymbol{f}(t) \tag{15.136}$$

so that

$$m\, \frac{d\boldsymbol{v}(t)}{dt} = -\, \frac{\boldsymbol{v}(t)}{B} + \boldsymbol{f}(t). \tag{15.137}$$

The parameter $B = \tau/m$ is called the **mobility**. The **ensemble average** (the average over all the particles) of the fluctuating force $\boldsymbol{f}(t)$ is zero

$$\langle \boldsymbol{f}(t) \rangle = \boldsymbol{0}. \tag{15.138}$$

Thus the ensemble average of the velocity satisfies

$$m \frac{d\langle \boldsymbol{v} \rangle}{dt} = -\frac{\langle \boldsymbol{v} \rangle}{B} \equiv -\frac{m\langle \boldsymbol{v} \rangle}{\tau} \tag{15.139}$$

whose solution is

$$\langle \boldsymbol{v(t)} \rangle = \langle \boldsymbol{v(0)} \rangle \, e^{-t/\tau}. \tag{15.140}$$

The instantaneous equation (15.137) divided by the mass m is

$$\frac{d\boldsymbol{v}(t)}{dt} = -\frac{\boldsymbol{v}(t)}{\tau} + \boldsymbol{a}(t) \tag{15.141}$$

in which $\boldsymbol{a}(t) = \boldsymbol{f}(t)/m$ is the acceleration. The ensemble average of the scalar product of the position vector $\boldsymbol{r}$ with this equation is

$$\left\langle \boldsymbol{r} \cdot \frac{d\boldsymbol{v}}{dt} \right\rangle = -\frac{\langle \boldsymbol{r} \cdot \boldsymbol{v} \rangle}{\tau} + \langle \boldsymbol{r} \cdot \boldsymbol{a} \rangle. \tag{15.142}$$

But since the ensemble average $\langle \boldsymbol{r} \cdot \boldsymbol{a} \rangle$ of the scalar product of the position vector $\boldsymbol{r}$ with the random, fluctuating part $\boldsymbol{a}$ of the acceleration vanishes, we have

$$\left\langle \boldsymbol{r} \cdot \frac{d\boldsymbol{v}}{dt} \right\rangle = -\frac{\langle \boldsymbol{r} \cdot \boldsymbol{v} \rangle}{\tau}. \tag{15.143}$$

Now

$$\frac{1}{2} \frac{d\,\boldsymbol{r}^2}{dt} = \frac{1}{2} \frac{d}{dt} \left(\boldsymbol{r} \cdot \boldsymbol{r} \right) = \boldsymbol{r} \cdot \boldsymbol{v} \tag{15.144}$$

and so

$$\frac{1}{2} \frac{d^2 \boldsymbol{r}^2}{dt^2} = \boldsymbol{r} \cdot \frac{d\boldsymbol{v}}{dt} + \boldsymbol{v}^2. \tag{15.145}$$

The ensemble average of this equation is

$$\frac{d^2 \langle \boldsymbol{r}^2 \rangle}{dt^2} = 2 \left\langle \boldsymbol{r} \cdot \frac{d\boldsymbol{v}}{dt} \right\rangle + 2\langle \boldsymbol{v}^2 \rangle \tag{15.146}$$

or in view of (15.143)

$$\frac{d^2 \langle \boldsymbol{r}^2 \rangle}{dt^2} = -2 \frac{\langle \boldsymbol{r} \cdot \boldsymbol{v} \rangle}{\tau} + 2\langle \boldsymbol{v}^2 \rangle. \tag{15.147}$$

We now use (15.144) to replace $\langle \boldsymbol{r} \cdot \boldsymbol{v} \rangle$ with half the first time derivative of $\langle \boldsymbol{r}^2 \rangle$ so that we have

$$\frac{d^2 \langle \boldsymbol{r}^2 \rangle}{dt^2} = -\frac{1}{\tau} \frac{d\langle \boldsymbol{r}^2 \rangle}{dt} + 2\langle \boldsymbol{v}^2 \rangle. \tag{15.148}$$

If the fluid is in equilibrium, then the ensemble average of $\boldsymbol{v}^2$ is given by the Maxwell-Boltzmann value (15.120)

$$\langle \boldsymbol{v}^2 \rangle = \frac{3kT}{m} \tag{15.149}$$

and so the acceleration (15.148) of $\langle \boldsymbol{r}^2 \rangle$ is

$$\frac{d^2 \langle \boldsymbol{r}^2 \rangle}{dt^2} + \frac{1}{\tau} \frac{d\langle \boldsymbol{r}^2 \rangle}{dt} = \frac{6kT}{m} \tag{15.150}$$

which we can integrate.

The general solution (7.12) to a second-order linear inhomogeneous differential equation is the sum of any particular solution to the inhomogeneous equation plus the general solution of the homogeneous equation. The function $\langle \boldsymbol{r}^2(t) \rangle_{pi} = 6kTt\tau/m$ is a particular solution of the inhomogeneous equation. The general solution to the homogeneous equation is $\langle \boldsymbol{r}^2(t) \rangle_{gh} = U + W \exp(-t/\tau)$ where U and W are constants. So $\langle \boldsymbol{r}^2(t) \rangle$ is

$$\langle \boldsymbol{r}^2(t) \rangle = U + W\, e^{-t/\tau} + 6kT\tau t/m \tag{15.151}$$

where U and W make $\langle \boldsymbol{r}^2(t) \rangle$ fit the boundary conditions. If the individual particles start out at the origin $\boldsymbol{r} = \boldsymbol{0}$, then one boundary condition is

$$\langle \boldsymbol{r}^2(0) \rangle = 0 \tag{15.152}$$

which implies that

$$U + W = 0. \tag{15.153}$$

And since the particles start out at $\boldsymbol{r} = \boldsymbol{0}$ with an isotropic distribution of initial velocities, the formula (15.144) for $\dot{r}^2$ implies that at $t = 0$

$$\left. \frac{d\, \langle \boldsymbol{r}^2 \rangle}{dt} \right|_{t=0} = 2\langle \boldsymbol{r}(0) \cdot \boldsymbol{v}(0) \rangle = 0. \tag{15.154}$$

This boundary condition means that our solution (15.151) must satisfy

$$\left. \frac{d\, \langle \boldsymbol{r}^2(t) \rangle}{dt} \right|_{t=0} = -\frac{W}{\tau} + \frac{6kT\tau}{m} = 0. \tag{15.155}$$

Thus $W = -U = 6kT\tau^2/m$, and so our solution (15.151) is

$$\langle \boldsymbol{r}^2(t) \rangle = \frac{6kT\tau^2}{m} \left[\frac{t}{\tau} + e^{-t/\tau} - 1 \right]. \tag{15.156}$$

At times short compared to τ, the first two terms in the power series for the exponential $\exp(-t/\tau)$ cancel the terms $-1 + t/\tau$, leaving

$$\langle \boldsymbol{r}^2(t) \rangle = \frac{6kT\tau^2}{m} \left[\frac{t^2}{2\tau^2} \right] = \frac{3kT}{m} \, t^2 = \langle v^2 \rangle \, t^2. \tag{15.157}$$

But at times long compared to τ, the exponential vanishes, leaving

$$\langle \boldsymbol{r}^2(t) \rangle = \frac{6kT\tau}{m} \, t = 6 \, B \, kT \, t. \tag{15.158}$$

The **diffusion constant** D is defined by

$$\langle \boldsymbol{r}^2(t) \rangle = 6 \, D \, t \tag{15.159}$$

and so we arrive at **Einstein's relation**

$$D = B \, kT \quad \text{or} \quad \zeta \, D = kT \tag{15.160}$$

in which $\zeta = 1/B$ is the **viscous-friction coefficient**. For a fluid or gas of viscosity η, Stokes's formula for ζ is

$$\zeta \equiv \frac{1}{B} = \frac{m}{\tau} = 6\pi\eta r. \tag{15.161}$$

These equations (15.160 and 15.161) express Boltzmann's constant k in terms of three quantities ζ, D, and T that were accessible to measurement in the first decade of the 20th century. They enabled scientists to measure Boltzmann's constant k for the first time. And since Avogadro's number N_A was the known gas constant R divided by k, the number of molecules in a mole was revealed to be $N_A = 6.022 \times 10^{23}$. Chemists could then divide the mass of a mole of any pure substance by 6.022×10^{23} and find the mass of the molecules that composed it. Suddenly the masses of the molecules of chemistry became known, and molecules were recognized as real particles and not tricks for balancing chemical equations.

15.13 Einstein-Nernst relation

If a particle of mass m carries an electric charge q and is exposed to an electric field $\boldsymbol{E}$, then in addition to viscosity $-\boldsymbol{v}/B$ and random buffeting $\boldsymbol{f}$, the constant force $q\boldsymbol{E}$ acts on it

$$m \, \frac{d\boldsymbol{v}}{dt} = -\frac{\boldsymbol{v}}{B} + q\boldsymbol{E} + \boldsymbol{f}. \tag{15.162}$$

The mean value of its velocity satisfies the differential equation

$$\left\langle \frac{d\boldsymbol{v}}{dt} \right\rangle = -\frac{\langle \boldsymbol{v} \rangle}{\tau} + \frac{q\boldsymbol{E}}{m} \tag{15.163}$$

where $\tau = mB$. A particular solution of this inhomogeneous equation is

$$\langle \boldsymbol{v}(t) \rangle_{pi} = \frac{q\tau\boldsymbol{E}}{m} = qB\boldsymbol{E}. \tag{15.164}$$

The general solution of its homogeneous version is $\langle \boldsymbol{v}(t) \rangle_{gh} = \boldsymbol{A} \exp(-t/\tau)$ in which the constant $\boldsymbol{A}$ is chosen to give $\langle \boldsymbol{v}(0) \rangle$ at $t = 0$. So by (7.12), the general solution $\langle \boldsymbol{v}(t) \rangle$ to equation (15.163) is (exercise 15.18) the sum of $\langle \boldsymbol{v}(t) \rangle_{pi}$ and $\langle \boldsymbol{v}(t) \rangle_{gh}$

$$\langle \boldsymbol{v}(t) \rangle = qB\boldsymbol{E} + [\langle \boldsymbol{v}(0) \rangle - qB\boldsymbol{E}] \, e^{-t/\tau}. \tag{15.165}$$

By applying the tricks of the previous section (15.12), one may show (exercise 15.19) that the variance of the position $\boldsymbol{r}$ about its mean $\langle \boldsymbol{r}(t) \rangle$ is

$$\left\langle (\boldsymbol{r} - \langle \boldsymbol{r}(t) \rangle)^2 \right\rangle = \frac{6kT\tau^2}{m} \left(\frac{t}{\tau} - 1 + e^{-t/\tau} \right) \tag{15.166}$$

where for $\langle \boldsymbol{r}(0) \rangle = 0$

$$\langle \boldsymbol{r}(t) \rangle = (q\tau^2 \boldsymbol{E}/m) \left(t/\tau - 1 + e^{-t/\tau} \right) + \tau \left(1 - e^{-t/\tau} \right) \langle \boldsymbol{v}(0) \rangle. \tag{15.167}$$

For times $t \gg \tau$, the variance (15.166) is

$$\left\langle (\boldsymbol{r} - \langle \boldsymbol{r}(t) \rangle)^2 \right\rangle = \frac{6kT\tau t}{m}. \tag{15.168}$$

Since the diffusion constant D is defined by (15.159) as

$$\left\langle (\boldsymbol{r} - \langle \boldsymbol{r}(t) \rangle)^2 \right\rangle = 6 \, D \, t \tag{15.169}$$

we arrive at the Einstein-Nernst relation

$$D = \frac{kT\tau}{m} = kTB = \frac{\mu}{q} kT \tag{15.170}$$

in which the electric mobility is $\mu = qB$.

Example 15.19 (Coronavirus in air) How long does it take for an aerosol particle of radius $r = 0.1\,\mu$m containing SARS-CoV-2 to fall 2 m in the local gravitational field $-g\hat{\boldsymbol{z}}$? In the distance formula (15.167), we replace $q\boldsymbol{E}$ with $-mg\hat{\boldsymbol{z}}$ and get $\langle \boldsymbol{r}(t) \rangle \approx -g\tau t \hat{\boldsymbol{z}} + \tau \langle \boldsymbol{v}(0) \rangle$ for $t \gg \tau$. The parameter τ is given by Stokes's formula (15.161) as $\tau = m/(6\pi\eta r)$ for a gas or a fluid of viscosity η. So the time for the aerosol particle to fall a distance d is $t = 6\pi\eta r d/(gm)$ or $t = 9\eta d/(2g\rho r^2)$ in which ρ is the density of the aerosol.

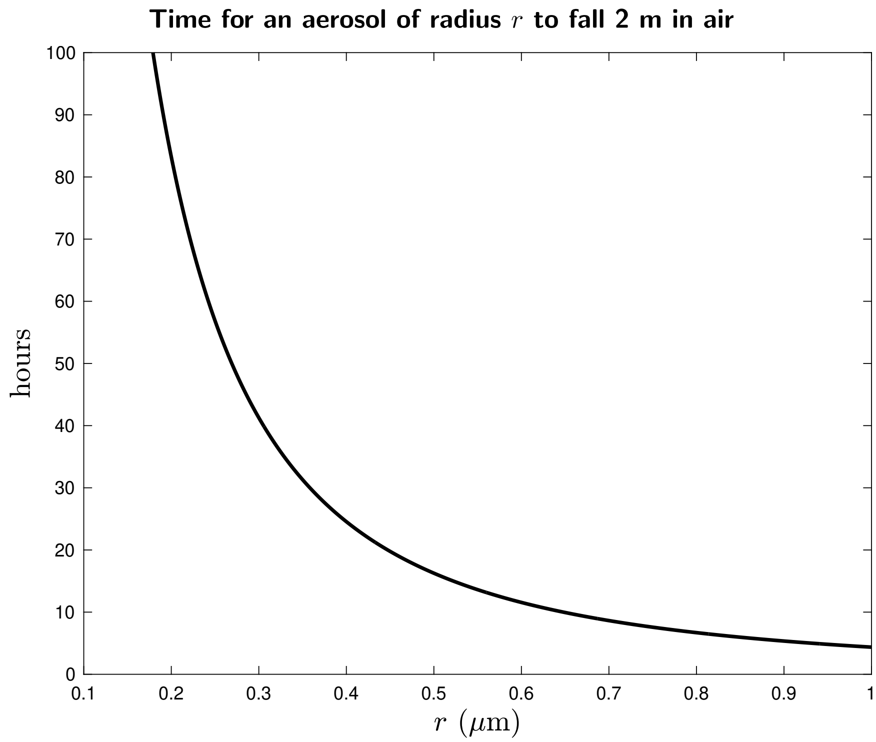

Figure 15.6 The time (15.171) for an aerosol of radius r to fall 2 m is plotted against the radius r in microns. Most airborne infections are spread by aerosols of radius $0.1 < r < 1$ micron.

The Cunningham slip correction C slightly changes the time to fall a distance d to

$$t = \frac{9\eta d}{2g\rho r^2 C} \tag{15.171}$$

in which (Hinds, 1999; Pöhlker et al., 2021)

$$C = 1 + \frac{\lambda}{r}\left(1.17 + 0.525\, e^{-0.78\, r/\lambda}\right) \tag{15.172}$$

and $\lambda = 68$ nm is the mean free path of air.

At $T = 25$ C, the viscosity of air is $\eta = 1.849 \times 10^{-5}$ kg/(m s). If the density of the aerosol particle is that of water, then the time (15.171) for it to fall 2 m varies inversely with the square of its radius and for $0.1 < r < 1\,\mu$m is as plotted in Fig. 15.6. The time to fall 2 m is 1.69×10^6 s or 19.5 days. For $r = 1\,\mu$m, the time to fall 2 m is 4.7 hours. SARS-CoV-2 spreads mainly via aerosol particles of $0.05 < r < 5\,\mu$m with peak transmission carried by those with $0.1 < r < 1$ micron. Thus Covid is airborne (Wang et al., 2021; Pöhlker et al., 2021; Hawks et al., 2021).

The radius σ of the falling ball into which the aerosol diffuses is approximately the square-root of the variance (15.168), and so is only $\sigma = 2.06 \times 10^{-4}$ m for $r = 1\,\mu$m and $\sigma = 6.52 \times 10^{-4}$ m for $r = 0.1\,\mu$m.

The transverse distance $\tau\langle \boldsymbol{v}(0)\rangle$ is negligible because the time τ is less than 3 ms for $r < 5\,\mu$m. So even for coughs with $\langle \boldsymbol{v}(0)\rangle = 10$ m/s, the distance $\tau\langle \boldsymbol{v}(0)\rangle$ is less than 3 cm, although the air flow of a cough can blow aerosols farther. □

15.14 Fluctuation and dissipation

Let's look again at Langevin's equation (15.141)

$$\frac{d\boldsymbol{v}(t)}{dt} + \frac{\boldsymbol{v}(t)}{\tau} = \boldsymbol{a}(t). \tag{15.173}$$

If we multiply both sides by the exponential $\exp(t/\tau)$

$$\left(\frac{d\boldsymbol{v}}{dt} + \frac{\boldsymbol{v}}{\tau}\right) e^{t/\tau} = \frac{d}{dt}\left(\boldsymbol{v}\, e^{t/\tau}\right) = \boldsymbol{a}(t)\, e^{t/\tau} \tag{15.174}$$

and integrate from 0 to t

$$\int_0^t \frac{d}{dt'}\left(\boldsymbol{v}\, e^{t'/\tau}\right) dt' = \boldsymbol{v}(t)\, e^{t/\tau} - \boldsymbol{v}(0) = \int_0^t \boldsymbol{a}(t')\, e^{t'/\tau}\, dt' \tag{15.175}$$

then we get

$$\boldsymbol{v}(t) = e^{-t/\tau}\, \boldsymbol{v}(0) + e^{-t/\tau} \int_0^t \boldsymbol{a}(t')\, e^{t'/\tau}\, dt'. \tag{15.176}$$

Thus the ensemble average of the square of the velocity is

$$\begin{aligned}\langle \boldsymbol{v}^2(t)\rangle = e^{-2t/\tau}\, \langle \boldsymbol{v}^2(0)\rangle + 2e^{-2t/\tau} \int_0^t \langle \boldsymbol{v}(0)\cdot \boldsymbol{a}(t')\rangle\, e^{t'/\tau}\, dt' \quad (15.177)\\ +e^{-2t/\tau} \int_0^t \int_0^t \langle \boldsymbol{a}(u_1)\cdot \boldsymbol{a}(t_2)\rangle\, e^{(u_1+t_2)/\tau}\, du_1 dt_2.\end{aligned}$$

The second term on the RHS is zero, so we have

$$\langle \boldsymbol{v}^2(t)\rangle = e^{-2t/\tau}\, \langle \boldsymbol{v}^2(0)\rangle + e^{-2t/\tau} \int_0^t \int_0^t \langle \boldsymbol{a}(t_1)\cdot \boldsymbol{a}(t_2)\rangle\, e^{(t_1+t_2)/\tau}\, dt_1 dt_2. \tag{15.178}$$

The ensemble average

$$C(t_1, t_2) = \langle \boldsymbol{a}(t_1) \cdot \boldsymbol{a}(t_2) \rangle \tag{15.179}$$

is an example of an **autocorrelation function**.

All autocorrelation functions have some simple properties, which are easy to prove (Pathria, 1972, p. 458):

1. If the system is independent of time, then its autocorrelation function for any given variable $\boldsymbol{A}(t)$ depends only upon the time delay s:

$$C(t, t+s) = \langle \boldsymbol{A}(t) \cdot \boldsymbol{A}(t+s) \rangle \equiv C(s). \tag{15.180}$$

2. The autocorrelation function for $s = 0$ is necessarily nonnegative

$$C(t, t) = \langle \boldsymbol{A}(t) \cdot \boldsymbol{A}(t) \rangle = \langle \boldsymbol{A}(t)^2 \rangle \geq 0. \tag{15.181}$$

   If the system is time independent, then $C(t, t) = C(0) \geq 0$.

3. The absolute value of $C(t_1, t_2)$ is never greater than the average of $C(t_1, t_1)$ and $C(t_2, t_2)$ because

$$\langle |\boldsymbol{A}(t_1) \pm \boldsymbol{A}(t_2)|^2 \rangle = \langle \boldsymbol{A}(t_1)^2 \rangle + \langle \boldsymbol{A}(t_2)^2 \rangle \pm 2 \langle \boldsymbol{A}(t_1) \cdot \boldsymbol{A}(t_2) \rangle \geq 0 \tag{15.182}$$

   which implies that $-2\, C(t_1, t_2) \leq C(t_1, t_1) + C(t_2, t_2) \geq 2\, C(t_1, t_2)$ or

$$2\, |C(t_1, t_2)| \leq C(t_1, t_1) + C(t_2, t_2). \tag{15.183}$$

   For a time-independent system, this inequality is $|C(s)| \leq C(0)$ for every time delay s.

4. If the variables $\boldsymbol{A}(t_1)$ and $\boldsymbol{A}(t_2)$ commute, then their autocorrelation function is symmetric

$$C(t_1, t_2) = \langle \boldsymbol{A}(t_1) \cdot \boldsymbol{A}(t_2) \rangle = \langle \boldsymbol{A}(t_2) \cdot \boldsymbol{A}(t_1) \rangle = C(t_2, t_1). \tag{15.184}$$

   For a time-independent system, this symmetry is $C(s) = C(-s)$.

5. If the variable $\boldsymbol{A}(t)$ is randomly fluctuating with zero mean, then we expect both that its ensemble average vanishes

$$\langle \boldsymbol{A}(t) \rangle = \boldsymbol{0} \tag{15.185}$$

   and that there is some characteristic time scale T beyond which the correlation function falls to zero:

$$\langle \boldsymbol{A}(t_1) \cdot \boldsymbol{A}(t_2) \rangle \to \langle \boldsymbol{A}(t_1) \rangle \cdot \langle \boldsymbol{A}(t_2) \rangle = 0 \tag{15.186}$$

   when $|t_1 - t_2| \gg T$.

In terms of the autocorrelation function $C(t_1, t_2) = \langle \boldsymbol{a}(t_1) \cdot \boldsymbol{a}(t_2) \rangle$ of the acceleration, the variance of the velocity (15.178) is

$$\langle \boldsymbol{v}^2(t) \rangle = e^{-2t/\tau} \langle \boldsymbol{v}^2(0) \rangle + e^{-2t/\tau} \int_0^t \int_0^t C(t_1, t_2)\, e^{(t_1+t_2)/\tau}\, dt_1 dt_2. \quad (15.187)$$

Since $C(t_1, t_2)$ is big only for tiny values of $|t_2 - t_1|$, it makes sense to change variables to

$$s = t_2 - t_1 \quad \text{and} \quad w = \frac{1}{2}(t_1 + t_2). \quad (15.188)$$

The element of area then is by (14.6–14.14)

$$dt_1 \wedge dt_2 = dw \wedge ds \quad (15.189)$$

and the limits of integration are $-2w \le s \le 2w$ for $0 \le w \le t/2$ and $-2(t-w) \le s \le 2(t-w)$ for $t/2 \le w \le t$. So $\langle \boldsymbol{v}^2(t) \rangle$ is

$$\begin{aligned} \langle \boldsymbol{v}^2(t) \rangle = {} & e^{-2t/\tau} \langle \boldsymbol{v}^2(0) \rangle + e^{-2t/\tau} \int_0^{t/2} e^{2w/\tau} dw \int_{-2w}^{2w} C(s)\, ds \\ & + e^{-2t/\tau} \int_{t/2}^{t} e^{2w/\tau} dw \int_{-2(t-w)}^{2(t-w)} C(s)\, ds. \end{aligned} \quad (15.190)$$

Since by (15.186) the autocorrelation function $C(s)$ vanishes outside a narrow window of width $2T$, we may approximate each of the s-integrals by

$$C = \int_{-\infty}^{\infty} C(s)\, ds. \quad (15.191)$$

It follows then that

$$\begin{aligned} \langle \boldsymbol{v}^2(t) \rangle &= e^{-2t/\tau} \langle \boldsymbol{v}^2(0) \rangle + C\, e^{-2t/\tau} \int_0^t e^{2w/\tau} dw \\ &= e^{-2t/\tau} \langle \boldsymbol{v}^2(0) \rangle + C\, e^{-2t/\tau}\, \frac{\tau}{2} \left(e^{2t/\tau} - 1 \right) \\ &= e^{-2t/\tau} \langle \boldsymbol{v}^2(0) \rangle + C\, \frac{\tau}{2} \left(1 - e^{-2t/\tau} \right). \end{aligned} \quad (15.192)$$

As $t \to \infty$, $\langle \boldsymbol{v}^2(t) \rangle$ must approach its equilibrium value of $3kT/m$, and so

$$\lim_{t\to\infty} \langle \boldsymbol{v}^2(t) \rangle = C\, \frac{\tau}{2} = \frac{3kT}{m} \quad (15.193)$$

which implies that

$$C = \frac{6kT}{m\tau} \quad \text{or} \quad \frac{1}{B} = \frac{m^2 C}{6kT}. \quad (15.194)$$

Our final formula for $\langle \boldsymbol{v}^2(t) \rangle$ then is

$$\langle \boldsymbol{v}^2(t) \rangle = e^{-2t/\tau} \langle \boldsymbol{v}^2(0) \rangle + \frac{3kT}{m} \left(1 - e^{-2t/\tau} \right). \tag{15.195}$$

Referring back to the definition (15.161) of the viscous-friction coefficient $f_v = 1/B$, we see that f_v is related to the integral

$$f_v = \frac{1}{B} = \frac{m^2}{6kT} C = \frac{m^2}{6kT} \int_{-\infty}^{\infty} \langle \boldsymbol{a}(0) \cdot \boldsymbol{a}(s) \rangle ds = \frac{1}{6kT} \int_{-\infty}^{\infty} \langle \boldsymbol{f}(0) \cdot \boldsymbol{f}(s) \rangle ds \tag{15.196}$$

of the autocorrelation function of the random acceleration $\boldsymbol{a}(t)$ or equivalently of the random force $\boldsymbol{f}(t)$. This equation relates the dissipation of viscous friction to the random fluctuations. It is an example of a **fluctuation-dissipation theorem**.

If we substitute our formula (15.195) for $\langle \boldsymbol{v}^2(t) \rangle$ into the expression (15.148) for the acceleration of $\langle \boldsymbol{r}^2 \rangle$, then we get

$$\frac{d^2 \langle \boldsymbol{r}^2(t) \rangle}{dt^2} = -\frac{1}{\tau} \frac{d \langle \boldsymbol{r}^2(t) \rangle}{dt} + 2 e^{-2t/\tau} \langle \boldsymbol{v}^2(0) \rangle + \frac{6kT}{m} \left(1 - e^{-2t/\tau} \right). \tag{15.197}$$

The solution with both $\langle \boldsymbol{r}^2(0) \rangle = 0$ and $d\langle \boldsymbol{r}^2(0) \rangle / dt = 0$ is (exercise 15.20)

$$\langle \boldsymbol{r}^2(t) \rangle = \langle \boldsymbol{v}^2(0) \rangle \, \tau^2 \left(1 - e^{-t/\tau} \right)^2 - \frac{3kT}{m} \tau^2 \left(1 - e^{-t/\tau} \right) \left(3 - e^{-t/\tau} \right) + \frac{6kT\tau}{m} t. \tag{15.198}$$

15.15 Fokker-Planck equation

Let $P(\boldsymbol{v}, t)$ be the probability distribution of particles in velocity space at time t, and $\psi(\boldsymbol{v}; \boldsymbol{u})$ be a normalized transition probability that the velocity changes from $\boldsymbol{v}$ to $\boldsymbol{v} + \boldsymbol{u}$ in the time interval $[t, t + \Delta t]$. We take the interval Δt to be much longer than the interval between successive particle collisions but much shorter than the time over which the velocity $\boldsymbol{v}$ changes appreciably. So $|\boldsymbol{u}| \ll |\boldsymbol{v}|$. We also assume that the successive changes in the velocities of the particles is a **Markoff stochastic process**, that is, that the changes are random and that what happens at time t depends only upon the state of the system at time t and not upon the history of the system. We then expect that the velocity distribution at time $t + \Delta t$ is related to that at time t by

$$P(\boldsymbol{v}, t + \Delta t) = \int P(\boldsymbol{v} - \boldsymbol{u}, t) \, \psi(\boldsymbol{v} - \boldsymbol{u}; \boldsymbol{u}) \, d^3\boldsymbol{u}. \tag{15.199}$$

Since $|\boldsymbol{u}| \ll |\boldsymbol{v}|$, we can expand $P(\boldsymbol{v}, t+\Delta t)$, $P(\boldsymbol{v}-\boldsymbol{u}, t)$, and $\psi(\boldsymbol{v}-\boldsymbol{u}; \boldsymbol{u})$ in Taylor series in $\boldsymbol{u}$ like

$$\psi(\boldsymbol{v}-\boldsymbol{u}; \boldsymbol{u}) = \psi(\boldsymbol{v}; \boldsymbol{u}) - \boldsymbol{u} \cdot \nabla_{\boldsymbol{v}} \psi(\boldsymbol{v}; \boldsymbol{u}) + \frac{1}{2} \sum_{i,j} u_i u_j \frac{\partial^2 \psi(\boldsymbol{v}; \boldsymbol{u})}{\partial v_i \partial v_j} \quad (15.200)$$

and get

$$\begin{aligned} P(\boldsymbol{v}, t) + \Delta t \, \frac{\partial P(\boldsymbol{v}, t)}{\partial t} = & \int \left[P(\boldsymbol{v}, t) - \boldsymbol{u} \cdot \nabla_{\boldsymbol{v}} P(\boldsymbol{v}, t) + \frac{1}{2} \sum_{i,j} u_i u_j \frac{\partial^2 P(\boldsymbol{v}, t)}{\partial v_i \partial v_j} \right] \\ & \times \left[\psi(\boldsymbol{v}; \boldsymbol{u}) - \boldsymbol{u} \cdot \nabla_{\boldsymbol{v}} \psi(\boldsymbol{v}; \boldsymbol{u}) + \frac{1}{2} \sum_{i,j} u_i u_j \frac{\partial^2 \psi(\boldsymbol{v}; \boldsymbol{u})}{\partial v_i \partial v_j} \right] d^3\boldsymbol{u}. \end{aligned} \quad (15.201)$$

The normalization of the transition probability ψ and the average changes in velocity are

$$\begin{aligned} 1 &= \int \psi(\boldsymbol{v}; \boldsymbol{u}) \, d^3\boldsymbol{u} \\ \langle u_i \rangle &= \int u_i \, \psi(\boldsymbol{v}; \boldsymbol{u}) \, d^3\boldsymbol{u} \\ \langle u_i u_j \rangle &= \int u_i \, u_j \, \psi(\boldsymbol{v}; \boldsymbol{u}) \, d^3\boldsymbol{u} \end{aligned} \quad (15.202)$$

in which the dependence of the mean values $\langle u_i \rangle$ and $\langle u_i u_j \rangle$ upon the velocity $\boldsymbol{v}$ is implicit. In these terms, the expansion (15.201) is

$$\begin{aligned} \Delta t \, \frac{\partial P(\boldsymbol{v}, t)}{\partial t} = & - \langle \boldsymbol{u} \rangle \cdot \nabla_{\boldsymbol{v}} P(\boldsymbol{v}, t) + \frac{1}{2} \sum_{i,j} \langle u_i u_j \rangle \frac{\partial^2 P(\boldsymbol{v}, t)}{\partial v_i \partial v_j} \\ & - P(\boldsymbol{v}, t) \, \nabla_{\boldsymbol{v}} \cdot \langle \boldsymbol{u} \rangle + P(\boldsymbol{v}, t) \, \frac{1}{2} \sum_{i,j} \frac{\partial^2 \langle u_i u_j \rangle}{\partial v_i \partial v_j} \\ & + \sum_{i,j} \frac{\partial P(\boldsymbol{v}, t)}{\partial v_i} \frac{\partial \langle u_i u_j \rangle}{\partial v_j}. \end{aligned} \quad (15.203)$$

Combining terms, we get the **Fokker-Planck equation** in its most general form (Chandrasekhar, 1943)

$$\Delta t \, \frac{\partial P(\boldsymbol{v}, t)}{\partial t} = - \nabla_{\boldsymbol{v}} \left[P(\boldsymbol{v}, t) \cdot \langle \boldsymbol{u} \rangle \right] + \frac{1}{2} \sum_{i,j} \frac{\partial^2}{\partial v_i \partial v_j} \left[P(\boldsymbol{v}, t) \langle u_i u_j \rangle \right]. \quad (15.204)$$

Example 15.20 (Brownian motion) Langevin's equation (15.137) gives the change $\boldsymbol{u}$ in the velocity $\boldsymbol{v}$ as the viscous drag plus some 10^{21} random tiny accelerations per second

$$\boldsymbol{u} = -\frac{\boldsymbol{v}\,\Delta t}{mB} + \frac{\boldsymbol{f}\Delta t}{m} \tag{15.205}$$

in which B is the mobility of the colloidal particle. The random changes $\boldsymbol{f}\Delta t/m$ in velocity are gaussian, and the transition probability is

$$\psi(\boldsymbol{v};\boldsymbol{u}) = \left(\frac{\beta m^2 B}{4\pi\Delta t}\right)^{3/2} \exp\left(-\frac{\beta m^2 B}{4\Delta t}\left|\boldsymbol{u} + \frac{\boldsymbol{v}\,\Delta t}{mB}\right|^2\right). \tag{15.206}$$

Here $\beta = 1/kT$, and Stokes's formula for the mobility B of a spherical colloidal particle of radius r in a fluid of viscosity η is $1/B = 6\pi r\eta$. The moments (15.202) of the changes $\boldsymbol{u}$ in velocity are in the limit $\Delta t \to 0$

$$\begin{aligned} \langle \boldsymbol{u} \rangle &= -\frac{\boldsymbol{v}\,\Delta t}{mB} \\ \langle u_i\, u_j \rangle &= 2\delta_{ij}\,\frac{kT}{m^2 B}\,\Delta t. \end{aligned} \tag{15.207}$$

So for Brownian motion, the Fokker-Planck equation is

$$\frac{\partial P(\boldsymbol{v},t)}{\partial t} = \frac{1}{mB}\nabla_{\boldsymbol{v}}\left[P(\boldsymbol{v},t)\cdot\boldsymbol{v}\right] + \frac{kT}{m^2 B}\nabla^2_{\boldsymbol{v}}\,P(\boldsymbol{v},t). \tag{15.208}$$

□

15.16 Characteristic and moment-generating functions

The Fourier transform (4.9) of a probability distribution $P(x)$ is its **characteristic function** $\tilde{P}(k)$ sometimes written as $\chi(k)$

$$\tilde{P}(k) \equiv \chi(k) \equiv E[e^{ikx}] = \int e^{ikx}\,P(x)\,dx. \tag{15.209}$$

The probability distribution $P(x)$ is the inverse Fourier transform (4.9)

$$P(x) = \int e^{-ikx}\,\tilde{P}(k)\,\frac{dk}{2\pi} = \int e^{-ikx}\,\chi(k)\,\frac{dk}{2\pi}. \tag{15.210}$$

Example 15.21 (Gauss) The characteristic function of the gaussian

$$P_G(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\,\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{15.211}$$

is by (4.19)

$$\tilde{P}_G(k,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \int \exp\left(ikx - \frac{(x-\mu)^2}{2\sigma^2}\right) dx \tag{15.212}$$
$$= \frac{e^{ik\mu}}{\sigma\sqrt{2\pi}} \int \exp\left(ikx - \frac{x^2}{2\sigma^2}\right) dx = \exp\left(i\mu k - \frac{1}{2}\sigma^2 k^2\right).$$

□

For a discrete probability distribution P_n the characteristic function is

$$\chi(k) \equiv E[e^{ikn}] = \sum_n e^{ikn} P_n. \tag{15.213}$$

The normalization of both continuous and discrete probability distributions implies that their characteristic functions satisfy $\tilde{P}(0) = \chi(0) = 1$.

Example 15.22 (Binomial and Poisson) The characteristic function of the binomial distribution (15.50)

$$P_b(n,p,N) = \binom{N}{n} p^n (1-p)^{N-n} \tag{15.214}$$

is

$$\chi_b(k) = \sum_{n=0}^{N} e^{ikn} \binom{N}{n} p^n (1-p)^{N-n} = \sum_{n=0}^{N} \binom{N}{n} (pe^{ik})^n (1-p)^{N-n}$$
$$= \left(pe^{ik} + 1 - p\right)^N = \left[p\left(e^{ik} - 1\right) + 1\right]^N. \tag{15.215}$$

The Poisson distribution (15.65)

$$P_P(n,\langle n\rangle) = \frac{\langle n\rangle^n}{n!} e^{-\langle n\rangle} \tag{15.216}$$

has the characteristic function

$$\chi_P(k) = \sum_{n=0}^{\infty} e^{ikn} \frac{\langle n\rangle^n}{n!} e^{-\langle n\rangle} = e^{-\langle n\rangle} \sum_{n=0}^{\infty} \frac{(\langle n\rangle e^{ik})^n}{n!} = \exp\left[\langle n\rangle \left(e^{ik} - 1\right)\right]. \tag{15.217}$$

□

The **moment-generating function** is the characteristic function evaluated at an imaginary argument

$$M(k) \equiv E[e^{kx}] = \tilde{P}(-ik) = \chi(-ik). \tag{15.218}$$

For a continuous probability distribution $P(x)$, it is

$$M(k) = E[e^{kx}] = \int e^{kx}\, P(x)\, dx \tag{15.219}$$

and for a discrete probability distribution P_n, it is

$$M(k) = E[e^{kx}] = \sum_n e^{kx_n}\, P_n. \tag{15.220}$$

In both cases, the normalization of the probability distribution implies that $M(0) = 1$.

Derivatives of the moment-generating function and of the characteristic function give the moments μ_n

$$E[x^n] = \mu_n = \left.\frac{d^n M(k)}{dk^n}\right|_{k=0} = (-i)^n \left.\frac{d^n \tilde{P}(k)}{dk^n}\right|_{k=0}. \tag{15.221}$$

Example 15.23 (Three moment-generating functions) The characteristic functions of the binomial distribution (15.215) and those of the distributions of Poisson (15.217) and Gauss (15.211) give us the moment-generating functions

$$M_b(k,p,N) = \left[p\left(e^k - 1\right) + 1\right]^N, \quad M_P(k,\langle n\rangle) = \exp\left[\langle n\rangle\left(e^k - 1\right)\right],$$
$$\text{and } \; M_G(k,\mu,\sigma) = \exp\left(\mu k + \frac{1}{2}\sigma^2 k^2\right). \tag{15.222}$$

Thus by (15.221), the first three moments of these three distributions are

$$\begin{aligned} \mu_{b0} =& 1, \quad \mu_{b1} = Np, \quad \mu_{b2} = N^2 p \\ \mu_{P0} =& 1, \quad \mu_{P1} = \langle n\rangle, \quad \mu_{P2} = \langle n\rangle + \langle n\rangle^2 \\ \mu_{G0} =& 1, \quad \mu_{G1} = \mu, \quad \mu_{G2} = \mu^2 + \sigma^2 \end{aligned} \tag{15.223}$$

(exercise 15.21). □

Since the characteristic and moment-generating functions have derivatives (15.221) proportional to the moments μ_n, their Taylor series are

$$\tilde{P}(k) = E[e^{ikx}] = \sum_{n=0}^{\infty} \frac{(ik)^n}{n!} E[x^n] = \sum_{n=0}^{\infty} \frac{(ik)^n}{n!}\, \mu_n \tag{15.224}$$

and

$$M(k) = E[e^{kx}] = \sum_{n=0}^{\infty} \frac{k^n}{n!} E[x^n] = \sum_{n=0}^{\infty} \frac{k^n}{n!}\, \mu_n. \tag{15.225}$$

The **cumulants** c_n of a probability distribution are the derivatives of the logarithm of its moment-generating function at $k = 0$

$$c_n = \left. \frac{d^n \log M(k)}{dk^n} \right|_{k=0} = (-i)^n \left. \frac{d^n \log \tilde{P}(k)}{dk^n} \right|_{k=0} . \tag{15.226}$$

One may show (exercise 15.23) that the first five cumulants of an arbitrary probability distribution are

$$c_0 = 0, \quad c_1 = \mu, \quad c_2 = \sigma^2, \quad c_3 = \nu_3, \quad \text{and} \quad c_4 = \nu_4 - 3\sigma^4 \tag{15.227}$$

where the ν's are its central moments (15.30). The 3d and 4th **normalized cumulants** are the **skewness** $v = c_3/\sigma^3 = \nu_3/\sigma^3$ and the **kurtosis** $\kappa = c_4/\sigma^4 = \nu_4/\sigma^4 - 3$.

Example 15.24 (Gaussian Cumulants) The logarithm of the moment-generating function (15.222) of Gauss's distribution is $\mu k + \sigma^2 k^2/2$. Thus by (15.226), $P_G(x, \mu, \sigma)$ has no skewness or kurtosis, its cumulants vanish $c_{Gn} = 0$ for $n > 2$, and its fourth central moment is $\nu_4 = 3\sigma^4$. □

15.17 Fat tails

The gaussian probability distribution $P_G(x, \mu, \sigma)$ falls off for $|x - \mu| \gg \sigma$ very fast—as $\exp\big(-(x-\mu)^2/2\sigma^2\big)$. Many other probability distributions fall off more slowly; they have **fat tails**. Rare "black-swan" events—wild fluctuations, market bubbles, and crashes—lurk in their fat tails.

Gosset's distribution, which is known as **Student's t-distribution** with ν degrees of freedom

$$P_S(x, \nu, a) = \frac{1}{\sqrt{\pi}} \frac{\Gamma((1+\nu)/2)}{\Gamma(\nu/2)} \frac{a^\nu}{(a^2 + x^2)^{(1+\nu)/2}} \tag{15.228}$$

has **power-law tails**. Its even moments are

$$\mu_{2n} = (2n-1)!! \, \frac{\Gamma(\nu/2 - n)}{\Gamma(\nu/2)} \left(\frac{a^2}{2} \right)^n \tag{15.229}$$

for $2n < \nu$ and infinite otherwise. For $\nu = 1$, it coincides with the Breit-Wigner or Cauchy distribution

$$P_S(x, 1, a) = \frac{1}{\pi} \frac{a}{a^2 + x^2} \tag{15.230}$$

in which $x = E - E_0$ and $a = \Gamma/2$ is the half-width at half-maximum.

Two representative cumulative probabilities are (Bouchaud and Potters, 2003, pp.15–16)

$$\Pr(x,\infty) = \int_x^\infty P_S(x',3,1)\,dx' = \frac{1}{2} - \frac{1}{\pi}\left[\arctan x + \frac{x}{1+x^2}\right] \tag{15.231}$$

$$\Pr(x,\infty) = \int_x^\infty P_S(x',4,\sqrt{2})\,dx' = \frac{1}{2} - \frac{3}{4}u + \frac{1}{4}u^3 \tag{15.232}$$

where $u = x/\sqrt{2+x^2}$ and a is picked so $\sigma^2 = 1$. William Gosset (1876–1937), who worked for Guinness, wrote as Student because Guinness didn't let its employees publish.

The **log-normal** probability distribution on $(0,\infty)$

$$P_{\ln}(x) = \frac{1}{\sigma x\sqrt{2\pi}} \exp\left[-\frac{\log^2(x/x_0)}{2\sigma^2}\right] \tag{15.233}$$

describes distributions of rates of return (Bouchaud and Potters, 2003, p. 9). Its moments are (exercise 15.26)

$$\mu_n = x_0^n\, e^{n^2\sigma^2/2}. \tag{15.234}$$

The **exponential distribution** on $[0,\infty)$

$$P_e(x) = \alpha e^{-\alpha x} \tag{15.235}$$

has (exercise 15.27) mean $\mu = 1/\alpha$ and variance $\sigma^2 = 1/\alpha^2$. The sum of n independent exponentially and identically distributed random variables $x = x_1 + \cdots + x_n$ is distributed on $[0,\infty)$ as (Feller, 1966, p.10)

$$P_{n,e}(x) = \alpha\frac{(\alpha x)^{n-1}}{(n-1)!}e^{-\alpha x}. \tag{15.236}$$

The sum of the squares $x^2 = x_1^2 + \cdots + x_n^2$ of n independent normally and identically distributed random variables of zero mean and variance σ^2 gives rise to Pearson's **chi-squared distribution** on $(0,\infty)$

$$P_{n,P}(x,\sigma)dx = \frac{\sqrt{2}}{\sigma}\frac{1}{\Gamma(n/2)}\left(\frac{x}{\sigma\sqrt{2}}\right)^{n-1} e^{-x^2/(2\sigma^2)}dx \tag{15.237}$$

which for $x = v$, $n = 3$, and $\sigma^2 = kT/m$ is (exercise 15.28) the Maxwell-Boltzmann distribution (15.117). In terms of $\chi = x/\sigma$, it is

$$P_{n,P}(\chi^2/2)\,d\chi^2 = \frac{1}{\Gamma(n/2)}\left(\frac{\chi^2}{2}\right)^{n/2-1} e^{-\chi^2/2}d\left(\chi^2/2\right). \tag{15.238}$$

It has mean and variance

$$\mu = n \quad \text{and} \quad \sigma^2 = 2n \tag{15.239}$$

and is used in the chi-squared test (Pearson, 1900). The Porter-Thomas distribution $P_{PT}(x) = e^{-x/2}/\sqrt{2\pi x}$ and the exponential distribution $P_e(x)$ (15.235) are special cases of the class (15.237) of chi-squared distributions.

Personal income, the amplitudes of catastrophes, the price changes of financial assets, and many other phenomena occur on both small and large scales. **Lévy** distributions describe such multi-scale phenomena. The characteristic function for a symmetric Lévy distribution is for $\nu \le 2$

$$\tilde{L}_\nu(k, a_\nu) = \exp\left(- a_\nu |k|^\nu \right). \tag{15.240}$$

Its inverse Fourier transform (15.210) is for $\nu = 1$ (exercise 15.29) the **Cauchy** or **Lorentz** distribution

$$L_1(x, a_1) = \frac{a_1}{\pi(x^2 + a_1^2)} \tag{15.241}$$

and for $\nu = 2$ the gaussian

$$L_2(x, a_2) = P_G(x, 0, \sqrt{2a_2}) = \frac{1}{2\sqrt{\pi a_2}} \exp\left(- \frac{x^2}{4a_2} \right) \tag{15.242}$$

but for other values of ν no simple expression for $L_\nu(x, a_\nu)$ is available. For $0 < \nu < 2$ and as $x \to \pm\infty$, it falls off as $|x|^{-(1+\nu)}$, and for $\nu > 2$ it assumes negative values, ceasing to be a probability distribution (Bouchaud and Potters, 2003, pp. 10–13).

15.18 Central limit theorem and Jarl Lindeberg

We have seen in sections (15.9 & 15.11) that unbiased fluctuations tend to distribute the position and velocity of molecules according to Gauss's distribution (15.83). Gaussian distributions occur very frequently. The **central limit theorem** suggests why they occur so often.

Let $x_1, \ldots, x_N$ be N **independent** random variables described by probability distributions $P_1(x_1), \ldots, P_N(x_N)$ with finite means μ_j and finite variances σ_j^2. The P_j's may be all different. The central limit theorem says that as $N \to \infty$ the probability distribution $P^{(N)}(y)$ for the average of the x_j's

$$y = \frac{1}{N} \left(x_1 + x_2 + \ldots + x_N \right) \tag{15.243}$$

tends to a gaussian in y quite independently of what the underlying probability distributions $P_j(x_j)$ happen to be.

Because expected values are linear (15.37), the mean value of the average y is the average of the N means

$$\begin{aligned}\mu_y = E[y] = E[(x_1 + \ldots + x_N)/N] &= \frac{1}{N}\left(E[x_1] + \ldots + E[x_N]\right) \\ &= \frac{1}{N}\left(\mu_1 + \ldots + \mu_N\right).\end{aligned} \tag{15.244}$$

The independence of the random variables $x_1, x_2, \ldots, x_N$ implies (15.43) that their joint probability distribution factorizes

$$P(x_1, \ldots, x_N) = P_1(x_1)P_2(x_2)\ldots P_N(x_N). \tag{15.245}$$

And our rule (15.48) for the variance of a linear combination of *independent* variables says that the variance of the average y is the sum of the variances

$$\sigma_y^2 = V[(x_1 + \ldots + x_N)/N] = \frac{1}{N^2}\left(\sigma_1^2 + \ldots + \sigma_N^2\right). \tag{15.246}$$

The conditional probability (15.3) $P^{(N)}(y|x_1, \ldots, x_N)$ that the average of the x's is y is the delta function (4.36)

$$P^{(N)}(y|x_1, \ldots, x_N) = \delta(y - (x_1 + x_2 + \ldots + x_N)/N). \tag{15.247}$$

Thus by (15.8) the probability distribution $P^{(N)}(y)$ for the average $y = (x_1 + x_2 + \ldots + x_N)/N$ of the x_j's is

$$\begin{aligned}P^{(N)}(y) &= \int P^{(N)}(y|x_1, \ldots, x_N)\, P(x_1, \ldots, x_N)\, d^N x \\ &= \int \delta(y - (x_1 + x_2 + \ldots + x_N)/N)\, P(x_1, \ldots, x_N)\, d^N x\end{aligned} \tag{15.248}$$

where $d^N x = dx_1 \ldots dx_N$. Its characteristic function is then

$$\begin{aligned}\tilde{P}^{(N)}(k) &= \int e^{iky}\, P^{(N)}(y)\, dy \\ &= \int e^{iky}\, \delta(y - (x_1 + x_2 + \ldots + x_N)/N)\, P(x_1, \ldots, x_N)\, d^N x\, dy \\ &= \int \exp\left[\frac{ik}{N}(x_1 + x_2 + \ldots + x_N)\right] P(x_1, \ldots, x_N)\, d^N x \\ &= \int \exp\left[\frac{ik}{N}(x_1 + x_2 + \ldots + x_N)\right] P_1(x_1)P_2(x_2)\ldots P_N(x_N)\, d^N x\end{aligned} \tag{15.249}$$

which is the product

$$\tilde{P}^{(N)}(k) = \tilde{P}_1(k/N)\, \tilde{P}_2(k/N)\, \ldots\, \tilde{P}_N(k/N) \tag{15.250}$$

of the characteristic functions

$$\tilde{P}_j(k/N) = \int e^{ikx_j/N} \, P_j(x_j) \, dx_j \tag{15.251}$$

of the probability distributions $P_1(x_1), \dots, P_N(x_N)$.

The Taylor series (15.224) for each characteristic function is

$$\tilde{P}_j(k/N) = \sum_{n=0}^{\infty} \frac{(ik)^n}{n! \, N^n} \, \mu_{nj} \tag{15.252}$$

and so for big N we can use the approximation

$$\tilde{P}_j(k/N) \approx 1 + \frac{ik}{N} \, \mu_j - \frac{k^2}{2N^2} \, \mu_{2j} \tag{15.253}$$

in which $\mu_{2j} = \sigma_j^2 + \mu_j^2$ by the formula (15.25) for the variance. So we have

$$\tilde{P}_j(k/N) \approx 1 + \frac{ik}{N} \, \mu_j - \frac{k^2}{2N^2} \left(\sigma_j^2 + \mu_j^2\right) \tag{15.254}$$

or for large N

$$\tilde{P}_j(k/N) \approx \exp\left(\frac{ik}{N}\mu_j - \frac{k^2}{2N^2}\sigma_j^2\right). \tag{15.255}$$

Thus as $N \to \infty$, the characteristic function (15.250) for the variable y converges to

$$\begin{aligned} \tilde{P}^{(N)}(k) &= \prod_{j=1}^{N} \tilde{P}_j(k/N) = \prod_{j=1}^{N} \exp\left(\frac{ik}{N}\mu_j - \frac{k^2}{2N^2}\sigma_j^2\right) \\ &= \exp\left[\sum_{j=1}^{N} \left(\frac{ik}{N}\mu_j - \frac{k^2}{2N^2}\sigma_j^2\right)\right] = \exp\left(i\mu_y k - \frac{1}{2}\sigma_y^2 k^2\right) \end{aligned} \tag{15.256}$$

which is the characteristic function (15.212) of a gaussian (15.211) with mean and variance

$$\mu_y = \frac{1}{N} \sum_{j=1}^{N} \mu_j \quad \text{and} \quad \sigma_y^2 = \frac{1}{N^2} \sum_{j=1}^{N} \sigma_j^2. \tag{15.257}$$

The inverse Fourier transform (15.210) now gives the probability distribution $P^{(N)}(y)$ for the average $y = (x_1 + x_2 + \ldots + x_N)/N$ as

$$P^{(N)}(y) = \int_{-\infty}^{\infty} e^{-iky} \, \tilde{P}^{(N)}(k) \, \frac{dk}{2\pi} \tag{15.258}$$

which in view of (15.256) and (15.212) tends as $N \to \infty$ to Gauss's distribution $P_G(y, \mu_y, \sigma_y)$

$$\begin{aligned} \lim_{N\to\infty} P^{(N)}(y) &= \int_{-\infty}^{\infty} e^{-iky} \lim_{N\to\infty} \tilde{P}^{(N)}(k) \frac{dk}{2\pi} \\ &= \int_{-\infty}^{\infty} e^{-iky} \exp\left(i\mu_y k - \frac{1}{2}\sigma_y^2 k^2\right) \frac{dk}{2\pi} \\ &= P_G(y, \mu_y, \sigma_y) = \frac{1}{\sigma_y\sqrt{2\pi}} \exp\left[-\frac{(y-\mu_y)^2}{2\sigma_y^2}\right] \end{aligned} \tag{15.259}$$

with mean μ_y and variance σ_y^2 as given by (15.257). The sense in which the exact distribution $P^{(N)}(y)$ converges to $P_G(y, \mu_y, \sigma_y)$ is that for all a and b the probability $\Pr_N(a < y < b)$ that y lies between a and b as determined by the exact $P^{(N)}(y)$ converges as $N \to \infty$ to the probability that y lies between a and b as determined by the gaussian $P_G(y, \mu_y, \sigma_y)$

$$\lim_{N\to\infty} \Pr_N(a < y < b) = \lim_{N\to\infty} \int_a^b P^{(N)}(y)\, dy = \int_a^b P_G(y, \mu_y, \sigma_y)\, dy. \tag{15.260}$$

This type of convergence is called **convergence in probability** (Feller, 1966, pp. 231, 241–248).

For the special case in which all the means and variances are the same, with $\mu_j = \mu$ and $\sigma_j^2 = \sigma^2$, the definitions in (15.257) imply that $\mu_y = \mu$ and $\sigma_y^2 = \sigma^2/N$. In this case, one may show (exercise 15.31) that in terms of the variable

$$u \equiv \frac{\sqrt{N}(y-\mu)}{\sigma} = \frac{\left(\sum_{n=1}^N x_j\right) - N\mu}{\sqrt{N}\,\sigma} \tag{15.261}$$

$P^{(N)}(y)$ converges to a distribution that is normal

$$\lim_{N\to\infty} P^{(N)}(y)\, dy = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}\, du. \tag{15.262}$$

To get a clearer idea of when the **central limit theorem** holds, let us write the sum of the N variances as

$$S_N \equiv \sum_{j=1}^N \sigma_j^2 = \sum_{j=1}^N \int_{-\infty}^{\infty} (x_j - \mu_j)^2\, P_j(x_j)\, dx_j \tag{15.263}$$

and the part of this sum due to the regions within δ of the means μ_j as

$$S_N(\delta) \equiv \sum_{j=1}^N \int_{\mu_j-\delta}^{\mu_j+\delta} (x_j - \mu_j)^2\, P_j(x_j)\, dx_j. \tag{15.264}$$

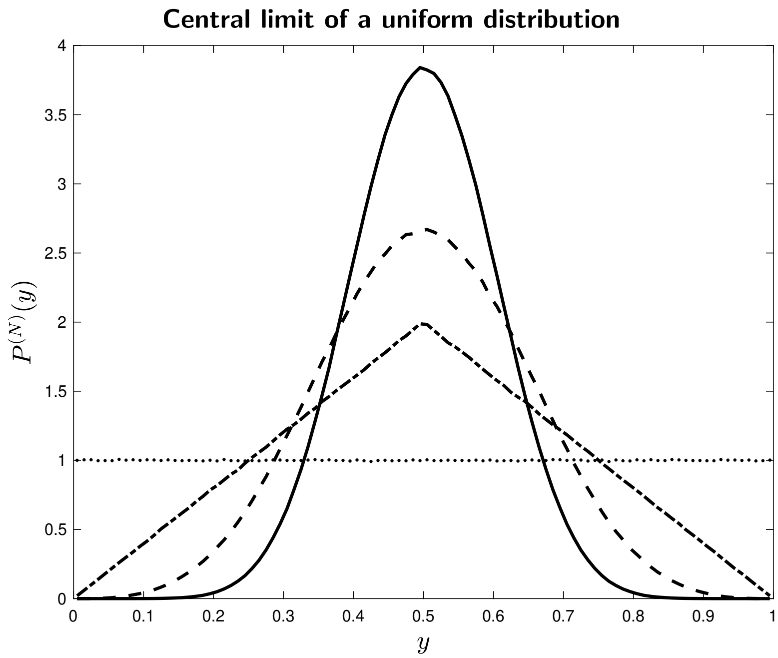

Figure 15.7 The probability distributions $P^{(N)}(y)$ (Eq. 15.248) for the mean $y = (x_1 + \ldots + x_N)/N$ of N random variables drawn from the uniform distribution are plotted for $N = 1$ (dots), 2 (dot dash), 4 (dashes), and 8 (solid). The distributions $P^{(N)}(y)$ rapidly approach gaussians with the same mean $\mu_y = 1/2$ but with shrinking variances $\sigma^2 = 1/(12N)$.

In these terms, Jarl Lindeberg (1876–1932) showed that the exact distribution $P^{(N)}(y)$ converges (in probability) to the gaussian (15.259) as long as the part $S_N(\delta)$ is most of S_N in the sense that for every $\epsilon > 0$

$$\lim_{N\to\infty} \frac{S_N\left(\epsilon\sqrt{S_N}\right)}{S_N} = 1. \tag{15.265}$$

This is **Lindeberg's condition** (Feller 1968, p. 254; Feller 1966, pp. 252–259; Gnedenko 1968, p. 304).

Because we dropped all but the first three terms of the series (15.252) for the characteristic functions $\tilde{P}_j(k/N)$, we may infer that the convergence of the distribution $P^{(N)}(y)$ to a gaussian is quickest near its mean μ_y. If the higher moments μ_{nj} are big, then for finite N the distribution $P^{(N)}(y)$ can have tails that are fatter than those of the limiting gaussian $P_G(y, \mu_y, \sigma_y)$.

Example 15.25 (Illustration of the central-limit theorem) The simplest probability distribution is a random number x uniformly distributed on the interval $(0, 1)$. The probability distribution $P^{(2)}(y)$ of the mean of two such random numbers is the integral

$$P^{(2)}(y) = \int_0^1 dx_1 \int_0^1 dx_2 \; \delta((x_1 + x_2)/2 - y). \tag{15.266}$$

Letting $u_1 = x_1/2$, we find

$$P^{(2)}(y) = 4\int_{\max(0,y-\frac{1}{2})}^{\min(y,\frac{1}{2})} \theta(\tfrac{1}{2} + u_1 - y)\, du_1 = 4y\,\theta(\tfrac{1}{2} - y) + 4(1-y)\,\theta(y - \tfrac{1}{2}) \tag{15.267}$$

which is the dot-dashed triangle in Fig. 15.7. The probability distribution $P^{(4)}(y)$ is the dashed somewhat gaussian curve in the figure, while $P^{(8)}(y)$ is the solid, nearly gaussian curve. □

To work through a more complicated example of the central limit theorem, we first need to learn how to generate random numbers that follow an arbitrary distribution.

15.19 Random-number generators

To generate truly random numbers, one might use decaying nuclei or an electronic device that makes white noise. But people usually settle for **pseudorandom numbers** computed by a mathematical algorithm. Such algorithms are deterministic, so the numbers they generate are not truly random. But for most purposes, they are random enough.

The standard way to generate pseudorandom numbers is to use the random-number generator of a FORTRAN, C, or C++ compiler (`gcc.gnu.org/`). In GFortran, the statement `call random_number(r)` returns an array `r` of random numbers uniformly distributed on the interval $(0, 1)$ and generated by PRNG whose period is $2^{256} - 1$. **Quasirandom numbers** (section 16.3) are somewhat better than pseudorandom ones.

Random-number generators distribute random numbers u uniformly on the interval $(0, 1)$. How do we make them follow an arbitrary distribution $P(r)$? If the distribution is strictly positive $P(r) > 0$ on the relevant interval (a, b), then its integral

$$F(x) = \int_a^x P(r)\, dr \tag{15.268}$$

is a strictly increasing function on (a, b), that is, $a < x < y < b$ implies $F(x) < F(y)$. Moreover, the function $F(x)$ rises from $F(a) = 0$ to $F(b) = 1$

and takes on every value $0 < y < 1$ for exactly one x in the interval (a, b). Thus the inverse function $F^{-1}(y)$

$$x = F^{-1}(y) \quad \text{if and only if} \quad y = F(x) \tag{15.269}$$

is well defined on the interval $(0, 1)$.

Our random-number generator gives us random numbers u that are uniform on $(0, 1)$. We want a random variable r whose probability $\Pr(r < x)$ of being less than any x is $F(x)$. The trick (Knuth, 1981, p. 116) is to generate a uniformly distributed random number u and then replace it with

$$r = F^{-1}(u). \tag{15.270}$$

For then, since $F(x)$ is one-to-one (15.269), the statements $F^{-1}(u) < x$ and $u < F(x)$ are equivalent, and therefore

$$\Pr(r < x) = \Pr(F^{-1}(u) < x) = \Pr(u < F(x)). \tag{15.271}$$

Example 15.26 ($P(r) = 3r^2$) To turn a distribution of random numbers u uniform on $(0, 1)$ into a distribution $P(r) = 3r^2$ of random numbers r, we integrate and find

$$F(x) = \int_0^x P(r)\, dr = \int_0^x 3r^2\, dr = x^3. \tag{15.272}$$

We then set $r = F^{-1}(u) = u^{1/3}$. □

Example 15.27 ($P(r) = 12\,(r - 1/2)^2$) To turn a distribution of random numbers u uniform on $(0, 1)$ into a distribution $P(r) = 12\,(r - 1/2)^2$, we integrate and find

$$F(x) = \int_0^x P(r)\, dr = \int_0^x 12\,(r - 1/2)^2\, dr = 4x^3 - 6x^2 + 3x. \tag{15.273}$$

We set $u = 4r^3 - 6r^2 + 3r$ and solve this cubic equation for r

$$r = \frac{1}{2}\left[1 - (1 - 2u)^{1/3}\right] \tag{15.274}$$

(or we ask Wolfram Alpha for the inverse function to $F(x)$). □

15.20 Illustration of the central limit theorem

To make things simple, we'll take all the probability distributions $P_j(x)$ to be the same and equal to $P_j(x_j) = 3x_j^2$ on the interval $(0, 1)$ and zero elsewhere. Our random-number generator gives us random numbers u that are uniformly distributed on $(0, 1)$, so by the example (15.26) the variable $r = u^{1/3}$ is distributed as $P_j(x) = 3x^2$.

The central limit theorem tells us that the distribution

$$P^{(N)}(y) = \int 3x_1^2\, 3x_2^2\, \dots\, 3x_N^2\, \delta((x_1 + x_2 + \dots + x_N)/N - y)\, d^N x \quad (15.275)$$

of the mean $y = (x_1 + \dots + x_N)/N$ tends as $N \to \infty$ to Gauss's distribution

$$\lim_{N\to\infty} P^{(N)}(y) = \frac{1}{\sigma_y \sqrt{2\pi}} \exp\left(-\frac{(x - \mu_y)^2}{2\sigma_y^2}\right) \quad (15.276)$$

with mean μ_y and variance σ_y^2 given by (15.257). Since the P_j's are all the same, they all have the same mean

$$\mu_y = \mu_j = \int_0^1 3x^3 dx = \frac{3}{4} \quad (15.277)$$

and the same variance

$$\sigma_j^2 = \int_0^1 3x^4 dx - \left(\frac{3}{4}\right)^2 = \frac{3}{5} - \frac{9}{16} = \frac{3}{80}. \quad (15.278)$$

By(15.257), the variance of the mean y is then $\sigma_y^2 = 3/(80N)$. Thus as N increases, the mean y tends to a gaussian with mean $\mu_y = 3/4$ and ever narrower peaks.

For $N = 1$, the probability distribution $P^{(1)}(y)$ is

$$P^{(1)}(y) = \int 3x_1^2\, \delta(x_1 - y)\, dx_1 = 3y^2 \quad (15.279)$$

which is the probability distribution we started with. In Fig. 15.8, this is the quadratic, dotted curve.

For $N = 2$, the probability distribution $P^{(1)}(y)$ is (exercise 15.30)

$$\begin{aligned} P^{(2)}(y) &= \int 3x_1^2\, 3x_2^2\, \delta((x_1 + x_2)/2 - y)\, dx_1\, dx_2 \quad (15.280) \\ &= \theta(\tfrac{1}{2} - y)\, \frac{96}{5}\, y^5 + \theta(y - \tfrac{1}{2}) \left(\frac{36}{5} - \frac{96}{5}\, y^5 + 48y^2 - 36y\right). \end{aligned}$$

You can get the probability distributions $P^{(N)}(y)$ for $N = 2^j$ by running the Fortran or C++ version of the program central_limit_of_3x^2 both of which are in Probability_and_statistics at `github.com/kevinecahill`.

The distributions $P^{(N)}(y)$ for $N = 1$, 2, 4, and 8 are plotted in Fig. 15.8. $P^{(1)}(y) = 3y^2$ is the original distribution. $P^{(2)}(y)$ is trying to be a gaussian, while $P^{(4)}(y)$ and $P^{(8)}(y)$ have almost succeeded. The variance $\sigma_y^2 = 3/80N$ shrinks with N.

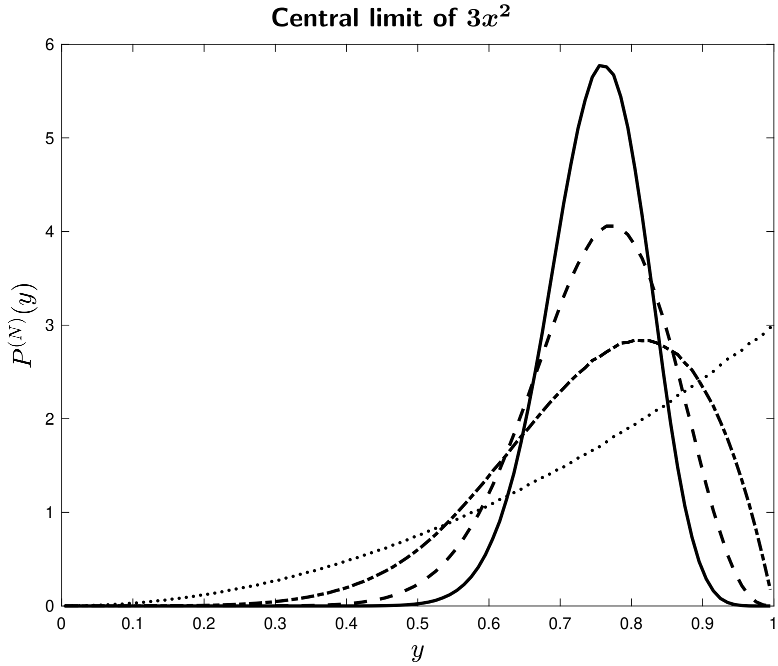

Figure 15.8 The probability distributions $P^{(N)}(y)$ (Eq. 15.275) for the mean $y = (x_1 + \ldots + x_N)/N$ of N random variables drawn from the quadratic distribution $P(x) = 3x^2$ are plotted for $N = 1$ (dots), 2 (dot dash), 4 (dashes), and 8 (solid). The four distributions $P^{(N)}(y)$ rapidly approach gaussians with the same mean $\mu_y = 3/4$ but with shrinking variances $\sigma_y^2 = 3/(80N)$.

The quadratic distribution $P(x) = 12\,(x - 1/2)^2$ of example 15.27 is very different from a gaussian centered at $x = 1/2$. Yet we see in Fig. 15.9 that the probability distributions $P^{(N)}(y)$ (15.248) for the mean $y = (x_1 + \cdots + x_N)/N$ of N random variables drawn from it do converge to such a gaussian.

Although FORTRAN95 is an ideal language for computation, C++ is more versatile and modular, and Java is easier to use.

15.21 Measurements, estimators, and Friedrich Bessel

The exact, physical probability distribution $P(x; \boldsymbol{\theta})$ for a stochastic variable x may depend upon one or more unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$. Experimenters seek to determine the unknown parameters $\boldsymbol{\theta}$, such as the mean μ and the variance σ^2, by collecting data in the form of observed values $\boldsymbol{x} = x_1, \ldots, x_N$ of the stochastic variable x. They assume that the probability distribution for the sequence $\boldsymbol{x} = (x_1, \ldots, x_N)$ is the product of

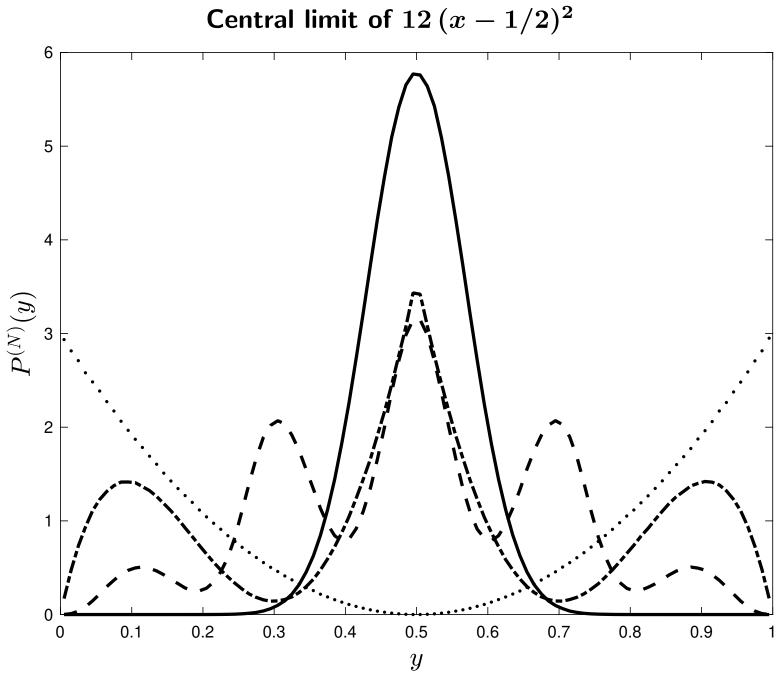

Figure 15.9 The probability distributions $P^{(N)}(y)$ (15.248) for the mean $y = (x_1 + \cdots + x_N)/N$ of N random variables drawn from the quadratic distribution $P(x) = 12\,(x - 1/2)^2$ of example 15.27 are plotted for $N = 1$ (dots), 2 (dot dash), 4 (dashes), and 32 (solid). The distributions have the same mean $\mu_y = 1/2$ and shrinking variances $\sigma_y^2 = 3/(20N)$.

N factors of the physical distribution $P(x;\boldsymbol{\theta})$

$$P(\boldsymbol{x};\boldsymbol{\theta}) = \prod_{j=1}^{N} P(x_j;\boldsymbol{\theta}). \tag{15.281}$$

They approximate the unknown value of a parameter θ_ℓ as the mean value of its **estimator** $u_\ell^{(N)}(\boldsymbol{x})$

$$E[u_\ell^{(N)}] = \int u_\ell^{(N)}(\boldsymbol{x})\, P(\boldsymbol{x};\boldsymbol{\theta})\, d^N x = \theta_\ell + b_\ell^{(N)}(\boldsymbol{\theta}). \tag{15.282}$$

If as $N \to \infty$, the **bias** $b_\ell^{(N)}(\boldsymbol{\theta}) \to 0$, then the estimator $u_\ell^{(N)}(\boldsymbol{x})$ is **consistent**.

Inasmuch as the mean (15.28) is the integral of the physical distribution

$$\mu = \int x\, P(x;\boldsymbol{\theta})\, dx \tag{15.283}$$

a natural estimator for the mean is

$$u_\mu^{(N)}(\boldsymbol{x}) = (x_1 + \ldots + x_N)/N. \tag{15.284}$$

Its expected value is

$$\begin{aligned} E[u_\mu^{(N)}] &= \int u_\mu^{(N)}(\boldsymbol{x})\, P(\boldsymbol{x};\boldsymbol{\theta})\, d^N x = \int \frac{x_1 + \ldots + x_N}{N}\, P(\boldsymbol{x};\boldsymbol{\theta})\, d^N x \qquad (15.285)\\ &= \frac{1}{N}\sum_{k=1}^{N} \int x_k\, P(x_k;\boldsymbol{\theta})\, dx_k \prod_{k\neq j=1}^{N} \int P(x_j;\boldsymbol{\theta})\, dx_j = \frac{1}{N}\sum_{k=1}^{N} \mu = \mu. \end{aligned}$$

Thus the natural estimator $u_\mu^{(N)}(\boldsymbol{x})$ of the mean (15.284) has $b_\ell^{(N)} = 0$, and so it is a consistent and unbiased estimator for the mean.

Since the variance (15.31) of the probability distribution $P(x;\boldsymbol{\theta})$ is the integral

$$\sigma^2 = \int (x-\mu)^2\, P(x;\boldsymbol{\theta})\, dx \tag{15.286}$$

the variance of the estimator u_μ^N is

$$\begin{aligned} V[u_\mu^{(N)}] &= \int \left(u_\mu^{(N)}(\boldsymbol{x}) - \mu\right)^2 P(\boldsymbol{x};\boldsymbol{\theta})\, d^N x = \int \left[\frac{1}{N}\sum_{j=1}^{N}(x_j-\mu)\right]^2 P(\boldsymbol{x};\boldsymbol{\theta})\, d^N x\\ &= \frac{1}{N^2}\sum_{j,k=1}^{N} \int (x_j-\mu)\,(x_k-\mu)\, P(\boldsymbol{x};\boldsymbol{\theta})\, d^N x \qquad (15.287)\\ &= \frac{1}{N^2}\sum_{j,k=1}^{N} \delta_{jk} \int (x_j-\mu)^2\, P(\boldsymbol{x};\boldsymbol{\theta})\, d^N x = \frac{1}{N^2}\sum_{k=1}^{N}\sigma^2 = \frac{\sigma^2}{N} \end{aligned}$$

in which σ^2 is the variance (15.286) of the physical distribution $P(x;\boldsymbol{\theta})$. We'll learn in the next section that no estimator of the mean can have a lower variance than this.

A natural estimator for the variance of the probability distribution $P(x;\boldsymbol{\theta})$ is

$$u_{\sigma^2}^{(N)}(x) = B\sum_{j=1}^{N}\left(x_j - u_\mu^{(N)}(\boldsymbol{x})\right)^2 \tag{15.288}$$

in which $B = B(N)$ is a constant of proportionality. The naive choice $B(N) = 1/N$ leads to a biased estimator. To find the correct value of B, we set the expected value $E[u^{(N)}_{\sigma^2}]$ equal to σ^2

$$E[u^{(N)}_{\sigma^2}] = \int B \sum_{j=1}^{N} \left(x_j - u^{(N)}_{\mu}(\boldsymbol{x}) \right)^2 P(\boldsymbol{x};\boldsymbol{\theta})\, d^N x = \sigma^2 \tag{15.289}$$

and solve for B. Subtracting the mean μ from both x_j and $u^{(N)}_{\mu}(\boldsymbol{x})$, we express σ^2/B as the sum of three terms

$$\frac{\sigma^2}{B} = \sum_{j=1}^{N} \int \left[x_j - \mu - \left(u^{(N)}_{\mu}(\boldsymbol{x}) - \mu \right) \right]^2 P(\boldsymbol{x};\boldsymbol{\theta})\, d^N x = S_{jj} + S_{j\mu} + S_{\mu\mu} \tag{15.290}$$

the first of which is

$$S_{jj} = \sum_{j=1}^{N} \int (x_j - \mu)^2\, P(\boldsymbol{x};\boldsymbol{\theta})\, d^N x = N\sigma^2. \tag{15.291}$$

The cross-term $S_{j\mu}$ is

$$\begin{aligned} S_{j\mu} &= -2 \sum_{j=1}^{N} \int (x_j - \mu) \left(u^{(N)}_{\mu}(\boldsymbol{x}) - \mu \right) P(\boldsymbol{x};\boldsymbol{\theta})\, d^N x \\ &= -\frac{2}{N} \sum_{j=1}^{N} \int (x_j - \mu) \sum_{k=1}^{N} (x_k - \mu)\, P(\boldsymbol{x};\boldsymbol{\theta})\, d^N x = -2\sigma^2. \end{aligned} \tag{15.292}$$

The third term is the variance (15.287) multiplied by N

$$S_{\mu\mu} = \sum_{j=1}^{N} \int \left(u^{(N)}_{\mu}(\boldsymbol{x}) - \mu \right)^2 P(\boldsymbol{x};\boldsymbol{\theta})\, d^N x = N V[u^{N}_{\mu}] = \sigma^2. \tag{15.293}$$

Thus the factor B must satisfy

$$\sigma^2/B = N\sigma^2 - 2\sigma^2 + \sigma^2 = (N-1)\sigma^2 \tag{15.294}$$

which tells us that $B = 1/(N-1)$, which is **Bessel's correction**. Our estimator for the variance $\sigma^2 = E[u^{(N)}_{\sigma^2}]$ of the probability distribution $P(x;\boldsymbol{\theta})$ then is

$$u^{(N)}_{\sigma^2}(\boldsymbol{x}) = \frac{1}{N-1} \sum_{j=1}^{N} \left(x_j - u^{(N)}_{\mu}(\boldsymbol{x}) \right)^2 = \frac{1}{N-1} \sum_{j=1}^{N} \left(x_j - \frac{1}{N} \sum_{k=1}^{N} x_k \right)^2. \tag{15.295}$$

It is consistent and unbiased since $E[u^{(N)}_{\sigma^2}] = \sigma^2$ by construction (15.289). It gives for the variance σ^2 of a single measurement the undefined ratio $0/0$, as it should, whereas the naive choice $B = 1/N$ absurdly gives zero.

On the basis of N measurements $x_1, \ldots, x_N$ we can estimate the mean of the unknown probability distribution $P(x;\boldsymbol{\theta})$ as $\mu_N = (x_1 + \ldots + x_N)/N$. And we can use Bessel's formula (15.295) to estimate the variance σ^2 of the unknown distribution $P(x;\boldsymbol{\theta})$. Our formula (15.287) for the variance $\sigma^2(\mu_N)$ of the mean μ_N then gives

$$\sigma^2(\mu_N) = \frac{\sigma^2}{N} = \frac{1}{N(N-1)} \sum_{j=1}^{N} \left(x_j - \frac{1}{N} \sum_{k=1}^{N} x_k \right)^2 . \tag{15.296}$$

Thus we can use N measurements x_j to estimate the mean μ to within a standard error or standard deviation of

$$\sigma(\mu_N) = \sqrt{\frac{\sigma^2}{N}} = \sqrt{\frac{1}{N(N-1)} \sum_{j=1}^{N} \left(x_j - \frac{1}{N} \sum_{k=1}^{N} x_k \right)^2} . \tag{15.297}$$

Few formulas have seen so much use.

15.22 Information and Ronald Fisher

The elements of the **Fisher information matrix** of a distribution $P(\boldsymbol{x};\boldsymbol{\theta})$ are the averages of products of pairs of the partial logarithmic derivatives $\partial_{\boldsymbol{\theta}} P(\boldsymbol{x};\boldsymbol{\theta})$ integrated over the possible values of the N measurements

$$\begin{aligned} F_{k\ell}(\boldsymbol{\theta}) &\equiv E\left[\frac{\partial \log P(\boldsymbol{x};\boldsymbol{\theta})}{\partial \theta_k} \, \frac{\partial \log P(\boldsymbol{x};\boldsymbol{\theta})}{\partial \theta_\ell} \right] \\ &= \int \frac{\partial \log P(\boldsymbol{x};\boldsymbol{\theta})}{\partial \theta_k} \, \frac{\partial \log P(\boldsymbol{x};\boldsymbol{\theta})}{\partial \theta_\ell} \, P(\boldsymbol{x};\boldsymbol{\theta}) \, d^N x \end{aligned} \tag{15.298}$$

(Ronald Fisher, 1890–1962). Fisher's matrix (exercise 15.32) is not only symmetric $F_{k\ell} = F_{\ell k}$ but also nonnegative (1.43) because for real c_k

$$\sum_{\ell,k=1}^{J} c_k F_{k\ell} c_\ell = \int \left(\sum_{k=1}^{J} c_k \frac{\partial \log P(\boldsymbol{x};\boldsymbol{\theta})}{\partial \theta_k} \right)^2 P(\boldsymbol{x};\boldsymbol{\theta}) \, d^N x \ge 0. \tag{15.299}$$

When the Fisher matrix is positive (1.44), it has an inverse. By differentiating the normalization condition

$$\int P(\boldsymbol{x};\boldsymbol{\theta}) \, d^N x = 1 \tag{15.300}$$

we get

$$0 = \int \frac{\partial P(\boldsymbol{x};\boldsymbol{\theta})}{\partial \theta_k} \, d^N x = \int \frac{\partial \log P(\boldsymbol{x};\boldsymbol{\theta})}{\partial \theta_k} \, P(\boldsymbol{x};\boldsymbol{\theta}) \, d^N x \tag{15.301}$$

which says that the average value of the **score** or the logarithmic derivative of the probability distribution vanishes. Using commas to denote θ derivatives as in

$$(\log P)_{,k} \equiv \frac{\partial \log P}{\partial \theta_k} \quad \text{and} \quad (\log P)_{,k\ell} \equiv \frac{\partial^2 \log P}{\partial \theta_k \partial \theta_\ell} \tag{15.302}$$

and differentiating the identity (15.301), one has (exercise 15.33)

$$0 = \int (\log P)_{,k} \, (\log P)_{,\ell} \, P \, d^N x + \int (\log P)_{,k\ell} \, P \, d^N x \tag{15.303}$$

so that another form of Fisher's information matrix is

$$F_{k\ell}(\boldsymbol{\theta}) = \, - \, E\left[(\log P)_{,k\ell}\right] = \, - \int (\log P)_{,k\ell} \, P \, d^N x. \tag{15.304}$$

Cramér and Rao used Fisher's information matrix to form a lower bound on the covariance (15.42) matrix $C[u_k, u_\ell]$ of any two estimators. Since both θ_ℓ and b_ℓ are constants independent of x, the vanishing (15.301) of the mean of the score implies that the covariance of the ℓth estimator $u_\ell(\boldsymbol{x})$ with the kth score $(\log P(\boldsymbol{x};\boldsymbol{\theta}))_{,k}$ is related to the θ_k-derivative $\langle u_\ell \rangle_{,k}$ of the mean $\langle u_\ell \rangle$

$$\begin{aligned} C[u_\ell, (\log P)_{,k}] &= \int (u_\ell - \theta_\ell - b_\ell) \, (\log P)_{,k} \, P \, d^N x \\ &= \int u_\ell \, (\log P)_{,k} \, P \, d^N x = \int u_\ell \, P_{,k} \, d^N x \\ &= \langle u_\ell \rangle_{,k} = (\theta_\ell + b_\ell)_{,k} = \delta_{\ell k} + b_{\ell,k}. \end{aligned} \tag{15.305}$$

Thus for any *constants* $y_1, y_2, \dots, y_J$ and $w_1, w_2, \dots, w_J$, where J is the number of estimators u_k, we have

$$\int \sum_{\ell,k=1}^{J} y_\ell \, (u_\ell - \theta_\ell - b_\ell) \, \sqrt{P} \, (\log P)_{,k} \sqrt{P} \, w_k \, d^N x = \sum_{\ell,k=1}^{J} y_\ell \, \langle u_\ell \rangle_{,k} \, w_k. \tag{15.306}$$

In matrix notation with $u'_{\ell k} = \langle u_\ell \rangle_{,k}$ and $(\log P)'_k = (\log P)_{,k}$, the square of this equation is

$$\left(\int y \cdot (u - \theta - b) \, \sqrt{P} \, \sqrt{P} \, (\log P)' \cdot w \, d^N x \right)^2 = (y^{\mathsf{T}} \, u' \, w)^2. \tag{15.307}$$

The Schwarz inequality (7.425) says that

$$\int \big(y\cdot(u-\theta-b)\big)^2\,P\,d^Nx \int \big((\log P)'\cdot w\big)^2\,P\,d^Nx \\ \ge \left(\int y\cdot(u-\theta-b)\,\sqrt{P}\,\sqrt{P}\,(\log P)'\cdot w\,d^Nx\right)^2. \tag{15.308}$$

The last two equations (15.307 and 15.308) now give us the key inequality

$$\int \big(y\cdot(u-\theta-b)\big)^2\,P\,d^Nx \int \big((\log P)'\cdot w\big)^2\,P\,d^Nx \ge \big(y^{\mathsf{T}}u'w\big)^2. \tag{15.309}$$

On the left-hand side of this equation, the first term is $y^{\mathsf{T}}Cy$ in which $C_{\ell\ell'} = C[u_\ell, u_{\ell'}]$ is the covariance (15.42) of the estimators $u_\ell(\boldsymbol{x})$ and $u_{\ell'}(\boldsymbol{x})$, and the second term is $w^{\mathsf{T}}Fw$ in which $F_{kk'}$ is Fisher's information matrix (15.298):

$$y^{\mathsf{T}}Cy\ \ w^{\mathsf{T}}Fw \ge \big(y^{\mathsf{T}}u'w\big)^2. \tag{15.310}$$

The Fisher information matrix F is real and symmetric, and its eigenvalues are nonnegative. If all its eigenvalues are positive (as they are unless P is independent of one or more of the θ_k's), then F has an inverse F^{-1}, and we can set $w = F^{-1}u'^{\mathsf{T}}y$. The inequality (15.310) then becomes

$$y^{\mathsf{T}}Cy\ \ y^{\mathsf{T}}u'F^{-1}FF^{-1}u'^{\mathsf{T}}y \ge y^{\mathsf{T}}u'F^{-1}u'^{\mathsf{T}}y\ y^{\mathsf{T}}u'F^{-1}u'^{\mathsf{T}}y. \tag{15.311}$$

Setting $FF^{-1} = I$ and canceling the common factor $y^{\mathsf{T}}u'F^{-1}u'^{\mathsf{T}}y$, we arrive at the **Cramér-Rao inequality**

$$y^{\mathsf{T}}Cy \ge y^{\mathsf{T}}u'F^{-1}u'^{\mathsf{T}}y. \tag{15.312}$$

Recalling the formula (15.305) which expresses u' as $u'_{\ell k} = \langle u_\ell\rangle_{,k} = \delta_{\ell k} + b_{\ell,k}$, we have

$$y_\ell\, C[u_\ell, u_k]\, y_k \ge y_r\,(\delta_{rs} + b_{r,s})\,F^{-1}_{sm}\,(\delta_{mn} + b_{n,m})\,y_n \tag{15.313}$$

or more succinctly

$$C \ge (I + b')F^{-1}(I + b'^{\mathsf{T}}). \tag{15.314}$$

In these inequalities, the y's are arbitrary numbers. Thus setting $y_\ell = \delta_{\ell k}$ and using the symmetry $F_{k\ell} = F_{\ell k}$, we can write the Cramer-Rao inequality (15.313) in terms of the variance $V[u_k] = C[u_k, u_k]$ as

$$V[u_k] = C[u_k, u_k] \ge F^{-1}_{kk} + 2F^{-1}_{k\ell}\,b_{k,\ell} + b_{k,\ell}F^{-1}_{\ell m}b_{k,m}. \tag{15.315}$$

If the estimator u_k is unbiased, this lower bound simplifies to

$$V[u_k] \ge F^{-1}_{kk}. \tag{15.316}$$

Example 15.28 (Cramér-Rao bound for a gaussian) The diagonal elements of Fisher's information matrix for the mean μ and variance σ^2 of Gauss's distribution for N data points $x_1, \ldots, x_N$

$$P_G^{(N)}(\boldsymbol{x}, \mu, \sigma) = \prod_{j=1}^{N} P_G(x_j; \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \exp\left(-\sum_{j=1}^{N} \frac{(x_j-\mu)^2}{2\sigma^2}\right) \tag{15.317}$$

are

$$\begin{aligned} F_{\mu\mu} &= \int \left[\left(\log P_G^{(N)}(\boldsymbol{x}, \mu, \sigma)\right)_{,\mu}\right]^2 P_G^{(N)}(\boldsymbol{x}, \mu, \sigma)\, d^N x \\ &= \sum_{i,j=1}^{N} \int \left(\frac{x_i-\mu}{\sigma^2}\right)\left(\frac{x_j-\mu}{\sigma^2}\right) P_G^{(N)}(\boldsymbol{x}, \mu, \sigma)\, d^N x \\ &= \sum_{i=1}^{N} \int \left(\frac{x_i-\mu}{\sigma^2}\right)^2 P_G^{(N)}(\boldsymbol{x}, \mu, \sigma)\, d^N x = \frac{N}{\sigma^2} \end{aligned} \tag{15.318}$$

and

$$\begin{aligned} F_{\sigma^2\sigma^2} &= \int \left[(\log P_G^{(N)}(\boldsymbol{x}, \mu, \sigma))_{,\sigma^2}\right]^2 P_G^{(N)}(\boldsymbol{x}, \mu, \sigma)\, d^N x \\ &= \sum_{i,j=1}^{N} \int \left[\frac{(x_i-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2}\right]\left[\frac{(x_j-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2}\right] P_G^{(N)}(\boldsymbol{x}, \mu, \sigma)\, d^N x \\ &= \frac{N}{2\sigma^4}. \end{aligned} \tag{15.319}$$

The off-diagonal terms are

$$\begin{aligned} F_{\mu\sigma^2} &= \int (\log P_G^{(N)}(\boldsymbol{x}, \mu, \sigma))_{,\mu} (\log P_G^{(N)}(\boldsymbol{x}, \mu, \sigma))_{,\sigma^2}\; P_G^{(N)}(\boldsymbol{x}, \mu, \sigma)\, d^N x \\ &= \sum_{i,j=1}^{N} \int \left[\frac{x_i-\mu}{\sigma^2}\right]\left[\frac{(x_j-\mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2}\right] P_G^{(N)}(\boldsymbol{x}, \mu, \sigma) d^N x \end{aligned} \tag{15.320}$$

which vanishes as does $F_{\sigma^2\mu} = F_{\mu\sigma^2} = 0$. The inverse of Fisher's matrix then is diagonal with $(F^{-1})_{\mu\mu} = \sigma^2/N$ and $(F^{-1})_{\sigma^2\sigma^2} = 2\sigma^4/N$.

The variance of any unbiased estimator $u_\mu(x)$ of the mean must exceed its Cramér-Rao lower bound (15.316), and so $V[u_\mu] \ge (F^{-1})_{\mu\mu} = \sigma^2/N$. The variance $V[u_\mu^{(N)}]$ of the natural estimator of the mean $u_\mu^{(N)}(\boldsymbol{x}) = (x_1 + \ldots + x_N)/N$ is σ^2/N by (15.287), and so it respects and saturates the lower bound (15.316)

$$V[u_\mu^{(N)}] = E[(u_\mu^{(N)} - \mu)^2] = \sigma^2/N = (F^{-1})_{\mu\mu}. \tag{15.321}$$

One may show (exercise 15.34) that the variance $V[u^{(N)}_{\sigma^2}]$ of Bessel's estimator (15.295) of the variance is (Riley et al., 2006, p. 1248)

$$V[u^{(N)}_{\sigma^2}] = \frac{1}{N}\left(\nu_4 - \frac{N-3}{N-1}\sigma^4\right) \tag{15.322}$$

where ν_4 is the fourth central moment (15.30) of the probability distribution. For the gaussian $P_G(\boldsymbol{x};\mu,\sigma)$ one may show (exercise 15.35) that this moment is $\nu_4 = 3\sigma^4$, and so for it

$$V_G[u^{(N)}_{\sigma^2}] = \frac{2}{N-1}\,\sigma^4. \tag{15.323}$$

Thus the variance of Bessel's estimator of the variance respects but does not saturate its Cramér-Rao lower bound (15.316, 15.319)

$$V_G[u^{(N)}_{\sigma^2}] = \frac{2}{N-1}\,\sigma^4 > \frac{2}{N}\,\sigma^4. \tag{15.324}$$

□

Estimators that saturate their Cramér-Rao lower bounds are **efficient**. The natural estimator $u^{(N)}_{\mu}(\boldsymbol{x})$ of the mean is efficient as well as consistent and unbiased, and Bessel's estimator $u^{(N)}_{\sigma^2}(\boldsymbol{x})$ of the variance is consistent and unbiased but not efficient.

15.23 Maximum likelihood

Suppose we measure some quantity x at various values of another variable t and find the values $x_1, x_2, \ldots, x_N$ at the known points $t_1, t_2, \ldots, t_N$. We might want to fit these measurements to a curve $x = f(t;\boldsymbol{\alpha})$ where $\boldsymbol{\alpha} = \alpha_1, \ldots, \alpha_M$ is a set of $M < N$ parameters. In view of the central limit theorem, we'll assume that the points x_j fall in Gauss's distribution about the values $x_j = f(t_j;\boldsymbol{\alpha})$ with some known variance σ^2. The probability of getting the N values $x_1, \ldots, x_N$ then is

$$P(\boldsymbol{x}) = \prod_{j=1}^{N} P(x_j, t_j, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \exp\left(-\sum_{j=1}^{N}\frac{(x_j - f(t_j;\boldsymbol{\alpha}))^2}{2\sigma^2}\right). \tag{15.325}$$

To find the M parameters $\boldsymbol{\alpha}$, we maximize the likelihood $P(\boldsymbol{x})$ by minimizing the argument of its exponential

$$0 = \frac{\partial}{\partial\alpha_\ell}\sum_{j=1}^{N}(x_j - f(t_j;\boldsymbol{\alpha}))^2 = -2\sum_{j=1}^{N}(x_j - f(t_j;\boldsymbol{\alpha}))\,\frac{\partial f(t_j;\boldsymbol{\alpha})}{\partial\alpha_\ell}. \tag{15.326}$$

If the function $f(t;\boldsymbol{\alpha})$ depends nonlinearly upon the parameters $\boldsymbol{\alpha}$, then we may need to use numerical methods to solve this **least-squares** problem.

But if the function $f(t;\boldsymbol{\alpha})$ depends **linearly** upon the M parameters $\boldsymbol{\alpha}$

$$f(t;\boldsymbol{\alpha}) = \sum_{k=1}^{M} g_k(t)\,\alpha_k \tag{15.327}$$

then the equations (15.326) that determine these parameters $\boldsymbol{\alpha}$ are linear

$$0 = \sum_{j=1}^{N} \left(x_j - \sum_{k=1}^{M} g_k(t_j)\,\alpha_k \right) g_\ell(t_j). \tag{15.328}$$

In matrix notation with G the $N \times M$ rectangular matrix with entries $G_{jk} = g_k(t_j)$, they are

$$G^{\mathsf{T}}\,\boldsymbol{x} = G^{\mathsf{T}}\,G\boldsymbol{\alpha}. \tag{15.329}$$

The basis functions $g_k(t)$ may depend nonlinearly upon the independent variable t. If one chooses them to be sufficiently different that the columns of G are linearly independent, then the rank of G is M, and the nonnegative matrix $G^{\mathsf{T}}\,G$ has an inverse. The matrix G then has a pseudoinverse (1.467)

$$G^{+} = \left(G^{\mathsf{T}}\,G\right)^{-1} G^{\mathsf{T}} \tag{15.330}$$

and it maps the N-vector $\boldsymbol{x}$ into our parameters $\boldsymbol{\alpha}$

$$\boldsymbol{\alpha} = G^{+}\,\boldsymbol{x}. \tag{15.331}$$

The product $G^{+}\,G = I_M$ is the $M \times M$ identity matrix, while

$$G\,G^{+} = P \tag{15.332}$$

is an $N \times N$ projection operator (exercise 15.36) onto the $M \times M$ subspace for which $G^{+}G = I_M$ is the identity operator. Like all projection operators, P satisfies $P^2 = P$.

15.24 Karl Pearson's chi-squared statistic

The argument of the exponential (15.325) in $P(\boldsymbol{x})$ is (the negative of) Karl Pearson's chi-squared statistic (Pearson, 1900)

$$\chi^2 \equiv \sum_{j=1}^{N} \frac{(x_j - f(t_j;\boldsymbol{\alpha}))^2}{2\sigma^2}. \tag{15.333}$$

When the function $f(t;\boldsymbol{\alpha})$ is linear (15.327) in $\boldsymbol{\alpha}$, the N-vector $f(t_j;\boldsymbol{\alpha})$ is $f = G\,\boldsymbol{\alpha}$. Pearson's χ^2 then is

$$\chi^2 = (\boldsymbol{x} - G\,\boldsymbol{\alpha})^2/2\sigma^2. \tag{15.334}$$

Now (15.331) tells us that $\boldsymbol{\alpha} = G^+\,\boldsymbol{x}$, and so in terms of the projection operator $P = G\,G^+$, the vector $\boldsymbol{x} - G\,\boldsymbol{\alpha}$ is

$$\boldsymbol{x} - G\,\boldsymbol{\alpha} = \boldsymbol{x} - G\,G^+\,\boldsymbol{x} = \left(I - G\,G^+\right)\boldsymbol{x} = (I - P)\,\boldsymbol{x}. \tag{15.335}$$

So χ^2 is proportional to the squared length

$$\chi^2 = \tilde{\boldsymbol{x}}^2/2\sigma^2 \tag{15.336}$$

of the vector

$$\tilde{\boldsymbol{x}} = \boldsymbol{x} - G\,\boldsymbol{\alpha} \equiv (I - P)\,\boldsymbol{x}. \tag{15.337}$$

Thus if the matrix G has rank M, and the vector $\boldsymbol{x}$ has N independent components, then the vector $\tilde{\boldsymbol{x}}$ has only $N - M$ independent components.

Example 15.29 (Two Position Measurements) Suppose we measure a position twice with error σ and get x_1 and x_2. If we choose

$$G = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{then its pseudoinverse (1.471) is} \quad G^+ = \frac{1}{2}\begin{pmatrix} 1 & 1 \end{pmatrix}. \tag{15.338}$$

The single parameter

$$\alpha = G^+ x = \frac{1}{2}\begin{pmatrix} 1 & 1 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{1}{2}\,(x_1 + x_2) \tag{15.339}$$

is the average of the two positions, and our formula (15.336) gives χ^2 as

$$\begin{aligned} \chi^2 &= \left\{[x_1 - (x_1 + x_2)/2]^2 + [x_2 - (x_1 + x_2)/2]^2\right\} \Big/ \, 2\sigma^2 \\ &= \left\{[(x_1 - x_2)/2]^2 + [(x_2 - x_1)/2]^2\right\} \Big/ \, 2\sigma^2 \\ &= \left[(x_1 - x_2)/\sqrt{2}\right]^2 /2\sigma^2. \end{aligned} \tag{15.340}$$

Thus instead of having two independent components x_1 and x_2, χ^2 just has one $(x_1 - x_2)/\sqrt{2}$. □

We can see how this happens more generally if we use as basis vectors the $N - M$ orthonormal vectors $|j\rangle$ in the kernel of P (that is, the $|j\rangle$'s annihilated by P)

$$P|j\rangle = 0 \quad 1 \le j \le N - M \tag{15.341}$$

and the M that lie in the range of the projection operator P

$$P|k\rangle = |k\rangle \quad N - M + 1 \le k \le N. \tag{15.342}$$

In terms of these basis vectors, the N-vector $\boldsymbol{x}$ is

$$\boldsymbol{x} = \sum_{j=1}^{N-M} x_j|j\rangle + \sum_{k=N-M+1}^{N} x_k|k\rangle \tag{15.343}$$

and the last M components of the vector $\tilde{\boldsymbol{x}}$ vanish

$$\tilde{\boldsymbol{x}} = (I - P)\,\boldsymbol{x} = \sum_{j=1}^{N-M} x_j|j\rangle. \tag{15.344}$$

Example 15.30 (N position measurements) Suppose the N values of x_j are the measured values of the position $f(t_j;\alpha) = x_j$ of some object. Then $M = 1$, and we choose $G_{j1} = g_1(t_j) = 1$ for $j = 1, \ldots, N$. Now $G^{\mathsf{T}}G = N$ is a 1×1 matrix, the number N, and the parameter α is the mean $\overline{x}$

$$\alpha = G^{+}\,\boldsymbol{x} = \left(G^{\mathsf{T}}\,G\right)^{-1} G^{\mathsf{T}}\,\boldsymbol{x} = \frac{1}{N}\sum_{j=1}^{N} x_j = \overline{x} \tag{15.345}$$

of the N position measurements x_j. So the vector $\tilde{\boldsymbol{x}}$ has components $\tilde{x}_j = x_j - \overline{x}$ and is orthogonal to $G^{\mathsf{T}} = (1, 1, \ldots, 1)$

$$G^{\mathsf{T}}\tilde{\boldsymbol{x}} = \left(\sum_{j=1}^{N} x_j\right) - N\overline{x} = 0. \tag{15.346}$$

The matrix G^{T} has rank 1, and the vector $\tilde{\boldsymbol{x}}$ has $N - 1$ independent components. □

Suppose now that we have determined our M parameters $\boldsymbol{\alpha}$ and have a theoretical fit

$$x = f(t;\boldsymbol{\alpha}) = \sum_{k=1}^{M} g_k(t)\,\alpha_k \tag{15.347}$$

which when we apply it to N measurements x_j gives χ^2 as

$$\chi^2 = (\tilde{\boldsymbol{x}})^2 / 2\sigma^2. \tag{15.348}$$

How good is our fit?

A χ^2 distribution with $N - M$ **degrees of freedom** has by (15.239) mean

$$E[\chi^2] = N - M \tag{15.349}$$

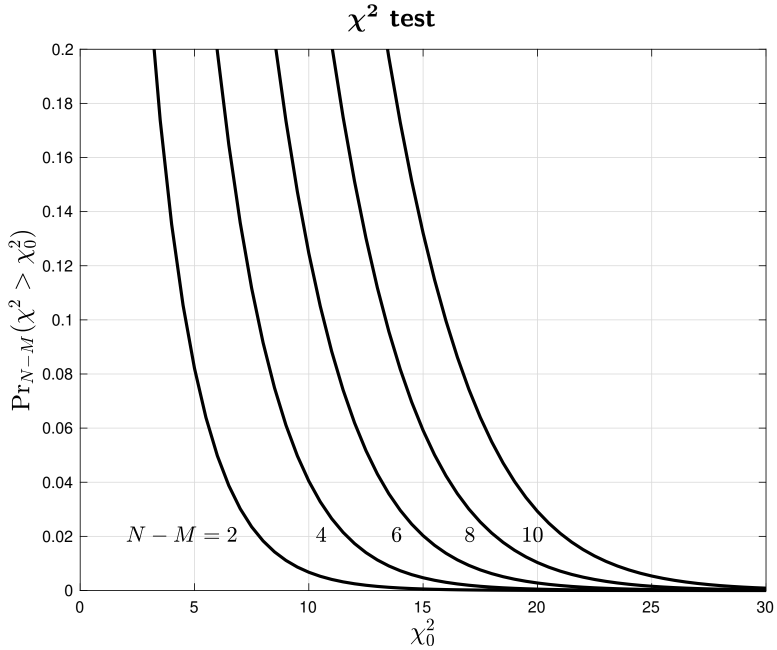

Figure 15.10 The probabilities $\Pr_{N-M}(\chi^2 > \chi_0^2)$ are plotted from left to right for $N - M = 2$, 4, 6, 8, and 10 degrees of freedom as functions of χ_0^2.

and variance

$$V[\chi^2] = 2(N - M). \tag{15.350}$$

So our χ^2 should be about

$$\chi^2 \approx N - M \pm \sqrt{2(N - M)}. \tag{15.351}$$

If it lies within this range, then (15.347) is a good fit to the data. But if it exceeds $N - M + \sqrt{2(N - M)}$, then the fit isn't so good. On the other hand, if χ^2 is less than $N - M - \sqrt{2(N - M)}$, then we may have used too many parameters or overestimated σ. Indeed, by using N parameters with $G\,G^+ = I_N$, we could get $\chi^2 = 0$ every time.

The probability that χ^2 exceeds χ_0^2 is the integral (15.238)

$$\Pr_n(\chi^2 > \chi_0^2) = \int_{\chi_0^2}^{\infty} P_n(\chi^2/2)\, d\chi^2 = \int_{\chi_0^2}^{\infty} \frac{1}{2\Gamma(n/2)} \left(\frac{\chi^2}{2}\right)^{n/2-1} e^{-\chi^2/2} d\chi^2 \tag{15.352}$$

in which $n = N - M$ is the number of data points minus the number of parameters, and $\Gamma(n/2)$ is the gamma function (5.57, 5.67). So an M-parameter fit to N data points has only a chance of ϵ of being good if its

χ^2 is greater than a χ_0^2 for which $\Pr_{N-M}(\chi^2 > \chi_0^2) = \epsilon$. These probabilities $\Pr_{N-M}(\chi^2 > \chi_0^2)$ are plotted in Fig. 15.10 for $N-M = 2$, 4, 6, 8, and 10. In particular, the probability of a value of χ^2 greater than $\chi_0^2 = 20$ respectively is 0.000045, 0.000499, 0.00277, 0.010336, and 0.029253 for $N-M = 2$, 4, 6, 8, and 10.

15.25 Kolmogorov's test

Suppose we want to use a sequence of N measurements x_j to determine the probability distribution that they come from. Our empirical probability distribution is

$$P_e^{(N)}(x) = \frac{1}{N}\sum_{j=1}^{N}\delta(x - x_j). \tag{15.353}$$

Our cumulative probability for events less than x then is

$$\Pr_e^{(N)}(-\infty, x) = \int_{-\infty}^{x} P_e^{(N)}(x')\,dx' = \int_{-\infty}^{x}\frac{1}{N}\sum_{j=1}^{N}\delta(x' - x_j)\,dx'. \tag{15.354}$$

So if we label our events in increasing order $x_1 \le x_2 \le \ldots \le x_N$, then the probability of an event less than x is a staircase

$$\Pr_e^{(N)}(-\infty, x) = \frac{j}{N} \quad \text{for} \quad x_j < x < x_{j+1}. \tag{15.355}$$

Having approximately and experimentally determined our empirical cumulative probability distribution $\Pr_e^{(N)}(-\infty, x)$, we might want to know whether it comes from some hypothetical cumulative probability distribution $\Pr_h(-\infty, x)$. One way to do this is to compute the distance D_N between the two cumulative probability distributions

$$D_N = \sup_{-\infty<x<\infty}\left|\Pr_e^{(N)}(-\infty, x) - \Pr_h(-\infty, x)\right| \tag{15.356}$$

in which **sup** stands for ***supremum*** and means **least upper bound**. Since cumulative probabilities lie between zero and one, it follows (exercise 15.37) that the Kolmogorov distance is bounded by $0 \le D_N \le 1$.

In general, as the number N of data points increases, we expect that our empirical distribution $\Pr_e^{(N)}(-\infty, x)$ should approach the actual or true distribution $\Pr_t(-\infty, x)$ from which the events x_j came. In this case, the Kolmogorov distance D_N should converge to a limiting value D_∞ for the

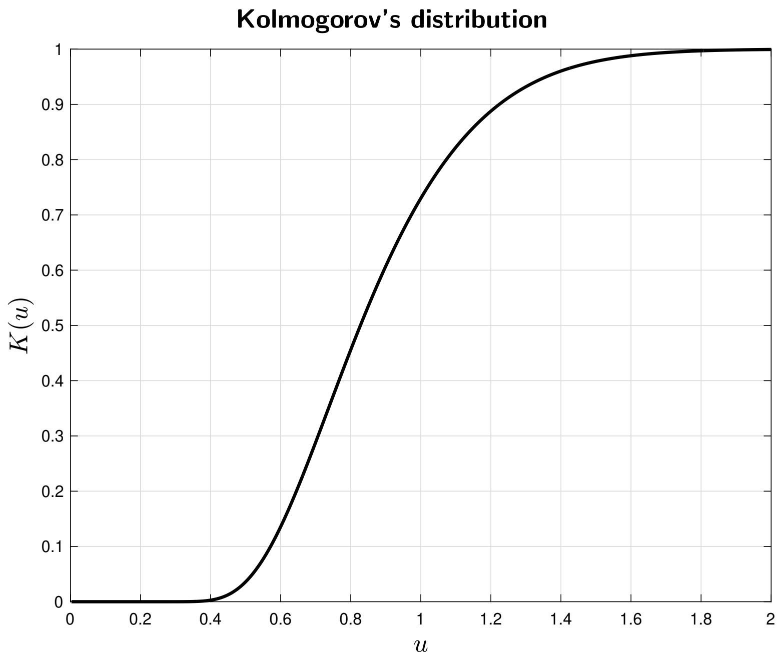

Figure 15.11 Kolmogorov's cumulative probability distribution $K(u)$ defined by (15.358) rises from zero to unity as u runs from zero to about two.

distance between the true distribution $\Pr_t(-\infty, x)$ and the hypothetical distribution $\Pr_h(-\infty, x)$

$$\lim_{N\to\infty} D_N = D_\infty = \sup_{-\infty<x<\infty} |\Pr_t(-\infty, x) - \Pr_h(-\infty, x)| \in [0, 1]. \quad (15.357)$$

If the true distribution $\Pr_t(-\infty, x)$ is the same as the hypothetical distribution $\Pr_h(-\infty, x)$, then we expect that $D_\infty = 0$. This expectation is confirmed by a theorem due to Glivenko (Glivenko, 1933; Cantelli, 1933) according to which the probability that the Kolmogorov distance D_N should go to zero as $N \to \infty$ is unity, $\Pr(D_\infty = 0) = 1$.

The real issue is how fast D_N should decrease with N if our events x_j do come from $\Pr_t(-\infty, x)$. This question was answered by Kolmogorov who showed (Kolmogorov, 1933) that if the events x_j of the empirical distribution $\Pr_e^{(N)}(-\infty, x)$ do come from the hypothetical distribution $\Pr_h(-\infty, x)$, and if $\Pr_h(-\infty, x)$ is continuous, then for large N the probability that $\sqrt{N}\, D_N$ (D_N being the Kolmogorov distance between the empirical $\Pr_e^{(N)}(-\infty, x)$ and hypothetical $\Pr_h(-\infty, x)$ cumulative distributions) is less than u is given

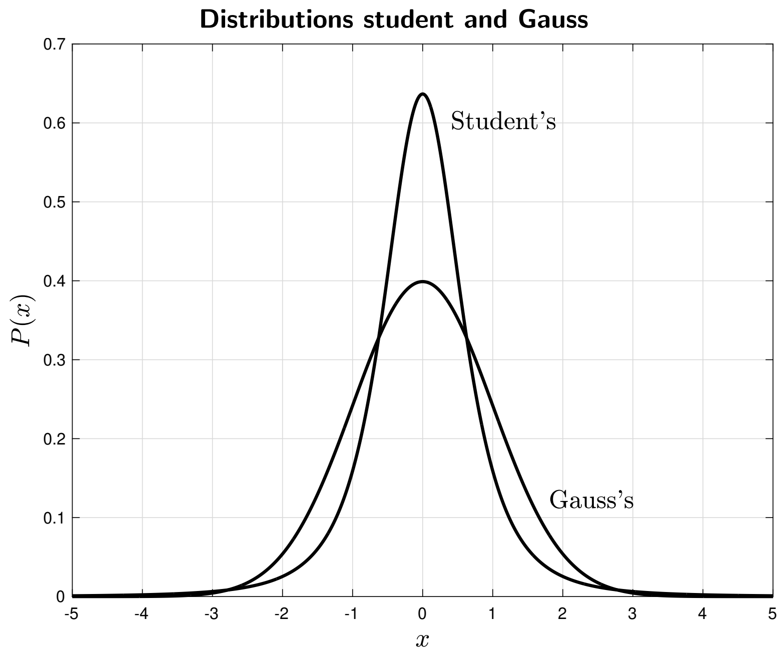

Figure 15.12 The probability distributions of Gauss $P_G(x, 0, 1)$ and Gosset/Student $P_S(x, 3, 1)$ with zero mean and unit variance.

by the **Kolmogorov function** $K(u)$

$$\lim_{N\to\infty} \Pr(\sqrt{N}\, D_N < u) = K(u) \equiv 1 + 2\sum_{k=1}^{\infty}(-1)^k e^{-2k^2u^2}. \qquad (15.358)$$

Amazingly, this upper bound is **universal and independent of the true** $\Pr_t(-\infty, x)$ **and hypothetical** $\Pr_h(-\infty, x)$ **probability distributions**.

But if the events x_j of the empirical distribution $\Pr_e^{(N)}(-\infty, x)$ come from a probability distribution $\Pr_t(-\infty, x)$ that is different from the hypothetical distribution $\Pr_h(-\infty, x)$, then as $N \to \infty$ we should expect that $\Pr_e^{(N)}(-\infty, x) \to \Pr_t(-\infty, x)$, and so that D_N would converge to a positive constant $D_\infty \in (0, 1]$. In this case, we expect that as $N \to \infty$ the quantity $\sqrt{N}\, D_N$ would grow with N as $\sqrt{N}\, D_\infty$.

Example 15.31 (Kolmogorov's Test) How do we use (15.358)? As illustrated in Fig. 15.11, Kolmogorov's distribution $K(u)$ rises from zero to unity on $(0, \infty)$, reaching 0.9993 already at $u = 2$. So if our points x_j come from the hypothetical distribution, then Kolmogorov's theorem (15.358) tells us that as $N \to \infty$, the probability that $\sqrt{N}\, D_N$ is less than 2 is more than

99.9%. But if the experimental points x_j do not come from the hypothetical distribution, then the quantity $\sqrt{N}\,D_N$ should grow as $\sqrt{N}\,D_\infty$ as $N \to \infty$.

To see what this means in practice, I took as the true distribution $P_t(x) = P_G(x, 0, 1)$ which has the cumulative probability distribution (15.93)

$$\Pr_t(-\infty, x) = \frac{1}{2}\left[\operatorname{erf}\left(x/\sqrt{2}\right) + 1\right]. \tag{15.359}$$

I generated $N = 10^m$ experimental points x_j for $m = 1, 2, 3, 4, 5$, and 6 from this true distribution $P_t(x) = P_G(x, 0, 1)$ and computed $u_N = \sqrt{10^m}\,D_{10^m}$ for these points. I found $\sqrt{10^m}\,D_{10^m} = 0.6928, 0.7074, 1.2000, 0.7356, 1.2260$, and 1.0683. All were less than 2, as expected since I had taken the experimental points x_j from the true distribution.

To see what happens when the experimental points do not come from the true distribution $P_t(x) = P_G(x, 0, 1)$, I generated $N = 10^m$ points x_j for $m = 1, 2, 3, 4, 5$, and 6 from Gosset's Student's distribution $P_h(x) = P_S(x, 3, 1)$ defined by (15.228) with $\nu = 3$ and $a = 1$. Both $P_t(x) = P_G(x, 0, 1)$ and $P_h(x) = P_S(x, 3, 1)$ have the same mean $\mu = 0$ and standard deviation $\sigma = 1$, as illustrated in Fig. 15.12. For these points, I computed $u_N = \sqrt{N}\,D_N$ and found $\sqrt{10^m}\,D_{10^m} = 0.7741$, 1.4522, 3.3837, 9.0478, 27.6414, and 87.8147. Only the first two are less than 2, and the last four grow as $\sqrt{N}$, indicating that the x_j had not come from the theoretical distribution. In fact, we can approximate the limiting value of D_N as $D_\infty \approx u_{10^6}/\sqrt{10^6} = 0.0878$. The exact value is (exercise 15.40) $D_\infty = 0.0868552356$.

At the risk of overemphasizing this example, I carried it one step further. I generated $\ell = 1, 2, \ldots 100$ sets of $N = 10^m$ points $x_j^{(\ell)}$ for $m = 2$, 3, and 4 drawn from $P_t(x) = P_G(x, 0, 1)$ and from $P_h(x) = P_S(x, 3, 1)$ and used them to form 100 empirical cumulative probabilities $\Pr_{e,G}^{(\ell,10^m)}(-\infty, x)$ and $\Pr_{e,S}^{(\ell,10^m)}(-\infty, x)$ as defined by (15.353–15.355). Next, I computed the distances $D_{G,G,10^m}^{(\ell)}$ and $D_{S,G,10^m}^{(\ell)}$ of each of these cumulative probabilities from the gaussian distribution $P_G(x, 0, 1)$. I labeled the two sets of 100 quantities $u_{G,G}^{(\ell,m)} = \sqrt{10^m}\,D_{G,G,10^m}^{(\ell)}$ and $u_{S,G}^{(\ell,m)} = \sqrt{10^m}\,D_{S,G,10^m}^{(\ell)}$ in increasing order as $u_{G,G,1}^{(m)} \le u_{G,G,2}^{(m)} \le \ldots \le u_{G,G,100}^{(m)}$ and $u_{S,G,1}^{(m)} \le u_{S,G,2}^{(m)} \le \ldots \le u_{S,G,100}^{(m)}$. I then used (15.353–15.355) to form the cumulative probabilities

$$\Pr_{e,G,G}^{(m)}(-\infty, u) = \frac{j}{N_s} \quad \text{for} \quad u_{G,G,j}^{(m)} < u < u_{G,G,j+1}^{(m)} \tag{15.360}$$

and

$$\Pr_{e,S,G}^{(m)}(-\infty, u) = \frac{j}{N_s} \quad \text{for} \quad u_{S,G,j}^{(m)} < u < u_{S,G,j+1}^{(m)} \tag{15.361}$$

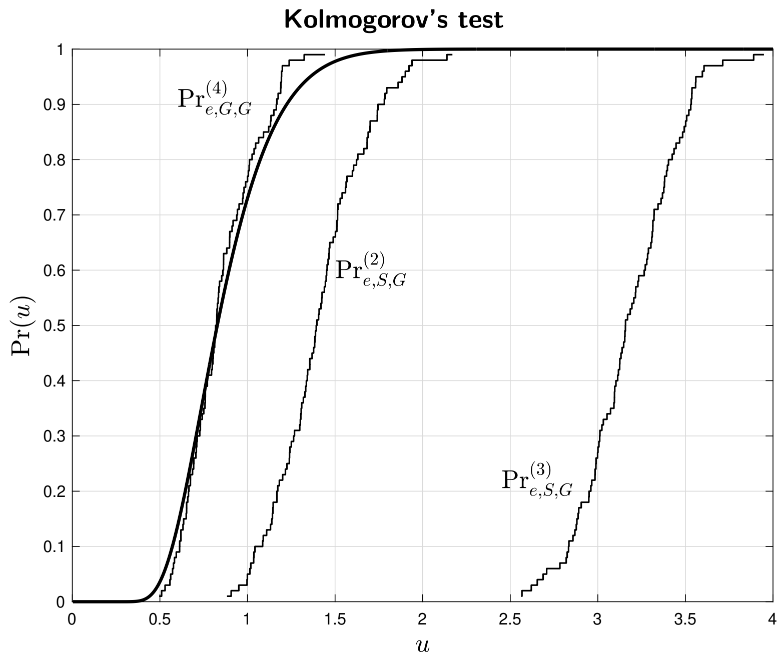

Figure 15.13 Kolmogorov's test is applied to points x_j taken from Gauss's distribution $P_G(x,0,1)$ and from Gosset's Student's distribution $P_S(x,3,1)$ to see whether the x_j came from $P_G(x,0,1)$. The thick smooth curve is Kolmogorov's universal cumulative probability distribution $K(u)$ defined by (15.358). The thin jagged curve that clings to $K(u)$ is the cumulative probability distribution $\Pr^{(4)}_{e,G,G}(-\infty,u)$ made (15.360) from points taken from $P_G(x,0,1)$. The other curves $\Pr^{(m)}_{e,S,G}(-\infty,u)$ for $m=2$ and 3 are made (15.361) from 10^m points taken from $P_S(x,3,1)$.

for $N_s = 100$ sets of 10^m points.

I plotted these cumulative probabilities in Fig. 15.13. The thick smooth curve is Kolmogorov's universal cumulative probability distribution $K(u)$ defined by (15.358). The thin jagged curve that clings to $K(u)$ is the cumulative probability distribution $\Pr^{(4)}_{e,G,G}(-\infty,u)$ made from 100 sets of 10^4 points taken from $P_G(x,0,1)$. As the number of sets increases beyond 100 and the number of points 10^m rises further, the probability distributions $\Pr^{(m)}_{e,G,G}(-\infty,u)$ converge to the universal cumulative probability distribution $K(u)$ and provide a numerical verification of Kolmogorov's theorem. Such curves make poor figures, however, because they hide beneath $K(u)$. The curves labeled $\Pr^{(m)}_{e,S,G}(-\infty,u)$ for $m=2$ and 3 are made from 100 sets of $N = 10^m$ points taken from $P_S(x,3,1)$ and tested as to whether they instead come from $P_G(x,0,1)$. Note that as $N = 10^m$ increases from 100 to

1000, the cumulative probability distribution $\Pr^{(m)}_{e,S,G}(-\infty, u)$ moves farther from Kolmogorov's universal cumulative probability distribution $K(u)$. In fact, the curve $\Pr^{(4)}_{e,S,G}(-\infty, u)$ made from 100 sets of 10^4 points lies beyond $u > 8$, too far to the right to fit in the figure. Kolmogorov's test gets more conclusive as the number of points $N \to \infty$. □

Warning, mathematical hazard: While binned data are ideal for chi-squared fits, they ruin Kolmogorov tests. The reason is that if the data are in bins of width w, then the empirical cumulative probability distribution $\Pr^{(N)}_e(-\infty, x)$ is a staircase function with steps as wide as the bin-width w even in the limit $N \to \infty$. Thus **even if the data come from the theoretical distribution**, the limiting value D_∞ of the Kolmogorov distance will be positive. In fact, one may show (exercise 15.41) that when the data do come from the theoretical probability distribution $P_t(x)$ assumed to be continuous, then the value of D_∞ is

$$D_\infty \approx \sup_{-\infty<x<\infty} \frac{w\, P_t(x)}{2}. \tag{15.362}$$

Thus in this case, the quantity $\sqrt{N}\, D_N$ would diverge as $\sqrt{N}\, D_\infty$ and lead one to believe that the data had not come from $P_t(x)$.

Suppose we have made some changes in our experimental apparatus and our software, and we want to see whether the new data $x'_1, x'_2, \ldots, x'_{N'}$ we took after the changes are consistent with the old data $x_1, x_2, \ldots, x_N$ we took before the changes. Then following equations (15.353–15.355), we can make two empirical cumulative probability distributions—one $\Pr^{(N)}_e(-\infty, x)$ made from the N old points x_j and the other $\Pr^{(N')}_e(-\infty, x)$ made from the N' new points x'_j. Next, we compute the distances

$$\begin{aligned} D^+_{N,N'} &= \sup_{-\infty<x<\infty} \left(\Pr^{(N)}_e(-\infty, x) - \Pr^{(N')}_e(-\infty, x)\right) \\ D_{N,N'} &= \sup_{-\infty<x<\infty} \left|\Pr^{(N)}_e(-\infty, x) - \Pr^{(N')}_e(-\infty, x)\right|. \end{aligned} \tag{15.363}$$

Smirnov (Smirnov 1939; Gnedenko 1968, p. 453) has shown that as $N, N' \to \infty$ the probabilities that

$$u^+_{N,N'} = \sqrt{\frac{NN'}{N+N'}}\, D^+_{N,N'} \quad \text{and} \quad u_{N,N'} = \sqrt{\frac{NN'}{N+N'}}\, D_{N,N'} \tag{15.364}$$

are less than u are

$$\begin{aligned} \lim_{N,N'\to\infty} \Pr(u^{+}_{N,N'} < u) &= 1 - e^{-2u^2} \\ \lim_{N,N'\to\infty} \Pr(u_{N,N'} < u) &= K(u) \end{aligned} \tag{15.365}$$

in which $K(u)$ is Kolmogorov's distribution (15.358).

Further reading

Students can learn about quantum probability and statistics in the book *Quantum Detection and Estimation Theory* (Helstrom, 1976). They can learn more about classical probability and statistics in these books: *Mathematical Methods for Physics and Engineering* (Riley et al., 2006), *An Introduction to Probability Theory and Its Applications I, II* (Feller, 1968, 1966), *Theory of Financial Risk and Derivative Pricing* (Bouchaud and Potters, 2003), and *Probability and Statistics in Experimental Physics* (Roe, 2001).

Exercises

15.1 Find the probabilities that two thrown fair dice give 4, 5, or 6.

15.2 Redo the three-door example for the case in which there are 100 doors, and 98 are opened to reveal empty rooms after one picks a door. Should one switch? What are the odds?

15.3 Show that the zeroth moment μ_0 and the zeroth central moment ν_0 always are unity, and that the first central moment ν_1 always vanishes.

15.4 Compute the variance of the uniform distribution on $(0, 1)$.

15.5 In the formulas (15.25 & 15.31) for the variances of discrete and continuous distributions, show that $E[(x - \langle x \rangle)^2] = \mu_2 - \mu^2$.

15.6 (a) Show that the covariance $\langle (x - \overline{x})(y - \overline{y}) \rangle$ is equal to $\langle x\, y \rangle - \langle x \rangle \langle y \rangle$ as asserted in (15.42). (b) Derive (15.46) for the variance $V[ax + by]$.

15.7 Derive expression (15.47) for the variance of a sum of N variables.

15.8 Find the range of $pq = p(1 - p)$ for $0 \le p \le 1$.

15.9 Show that the variance of the binomial distribution (15.50) is given by (15.54).

15.10 Suppose we ask three likely voters if they will vote for Michael Heinrich, and two say "Yes." What is the probability that he will be reelected? Hint: Imitate example (15.4).

15.11 Redo the polling example (15.18–15.20) for the case of a slightly better poll in which 16 likely voters were asked and 13 said they'd vote for Nancy Pelosi. What's the probability that she'll win the election? (You may use Maple or some other program to do the tedious integral.)

15.12 Without using the fact that the Poisson distribution is a limiting form of the binomial distribution, show from its definition (15.65) and its mean (15.67) that its variance is equal to its mean, as in (15.69).

15.13 Show that Gauss's approximation (15.82) to the binomial distribution is a normalized probability distribution with mean $\langle x \rangle = \mu = pN$ and variance $V[x] = pqN$.

15.14 Derive the approximations (15.96 & 15.97) for binomial probabilities for large N.

15.15 Compute the central moments (15.30) of the gaussian (15.83).

15.16 Derive formula (15.92) for the probability that a gaussian random variable falls within an interval.

15.17 Show that the expression (15.99) for $P(y|600)$ is negligible on the interval $(0, 1)$ except for y near $3/5$.

15.18 Determine the constant A of the homogeneous solution $\langle \boldsymbol{v}(t) \rangle_{gh}$ and derive expression (15.165) for the general solution $\langle \boldsymbol{v}(t) \rangle$ to (15.163).

15.19 Derive equation (15.166) for the variance of the position $\boldsymbol{r}$ about its mean $\langle \boldsymbol{r}(t) \rangle$. You may assume that $\langle \boldsymbol{r}(0) \rangle = \langle \boldsymbol{v}(0) \rangle = 0$ and that $\langle (\boldsymbol{v} - \langle \boldsymbol{v}(t) \rangle)^2 \rangle = 3kT/m$.

15.20 Derive equation (15.198) for the ensemble average $\langle \boldsymbol{r}^2(t) \rangle$ for the case in which $\langle \boldsymbol{r}^2(0) \rangle = 0$ and $d\langle \boldsymbol{r}^2(0) \rangle/dt = 0$.

15.21 Use (15.221) to derive the lower moments (15.223) of the binomial distribution and those of Gauss and Poisson.

15.22 Find the third and fourth moments μ_3 and μ_4 for the distributions of Poisson (15.216) and Gauss (15.211).

15.23 Derive formula (15.227) for the first five cumulants of an arbitrary probability distribution.

15.24 Show that like the characteristic function, the moment-generating function $M(t)$ for an average of several independent random variables factorizes $M(t) = M_1(t/N)\, M_2(t/N) \, \dots \, M_N(t/N)$.

15.25 Derive formula (15.234) for the moments of the log-normal probability distribution (15.233).

15.26 Why doesn't the log-normal probability distribution (15.233) have a sensible power-series about $x = 0$? What are its derivatives there?

15.27 Compute the mean and variance of the exponential distribution (15.235).

15.28 Show that the chi-square distribution $P_{3,G}(v, \sigma)$ with variance $\sigma^2 = kT/m$ is the Maxwell-Boltzmann distribution (15.117).

15.29 Compute the inverse Fourier transform (15.210) of the characteristic function (15.240) of the symmetric Lévy distribution for $\nu = 1$ and 2.

15.30 Show that the integral that defines $P^{(2)}(y)$ gives formula (15.280) with two Heaviside step functions. Hint: keep x_1 and x_2 in the interval $(0, 1)$.

15.31 Derive the normal distribution (15.262) in the variable (15.261) from the central limit theorem (15.259) for the case in which all the means and variances are the same.

15.32 Show that Fisher's matrix (15.298) is symmetric $F_{k\ell} = F_{\ell k}$ and non-negative (1.43), and that when it is positive (1.44), it has an inverse.

15.33 Derive the integral equations (15.301 & 15.303) from the normalization condition $\int P(\boldsymbol{x};\boldsymbol{\theta})\, d^N x = 1$.

15.34 Show that the variance $V[u_{\sigma^2}^{(N)}]$ of Bessel's estimator (15.295) is given by (15.322).

15.35 Compute the fourth central moment (15.30) of Gauss's probability distribution $P_G(x;\mu,\sigma^2)$.

15.36 Show that when the real $N \times M$ matrix G has rank M, the matrices $P = G\,G^+$ and $P_\perp = 1 - P$ are projection operators that are mutually orthogonal $P(I-P) = (I-P)P = 0$.

15.37 Show that Kolmogorov's distance D_N is bounded, $0 \le D_N \le 1$.

15.38 Show that Kolmogorov's distance D_N is the greater of the two Smirnov distances

$$\begin{aligned} D_N^+ &= \sup_{-\infty<x<\infty} \left(\mathrm{Pr}_e^{(N)}(-\infty,x) - \mathrm{Pr}_t(-\infty,x) \right) \\ D_N^- &= \sup_{-\infty<x<\infty} \left(\mathrm{Pr}_t(-\infty,x) - \mathrm{Pr}_e^{(N)}(-\infty,x) \right). \end{aligned} \tag{15.366}$$

15.39 Derive the formulas

$$\begin{aligned} D_N^+ &= \sup_{1\le j\le N} \left(\frac{j}{N} - \mathrm{Pr}_t(-\infty,x_j) \right) \\ D_N^- &= \sup_{1\le j\le N} \left(\mathrm{Pr}_t(-\infty,x_j) - \frac{j-1}{N} \right) \end{aligned} \tag{15.367}$$

for D_N^+ and D_N^-.

15.40 Compute the exact limiting value D_∞ of the Kolmogorov distance between $P_G(x,0,1)$ and $P_S(x,3,1)$. Use the cumulative probabilities (15.359 & 15.231) to find the value of x that maximizes their difference. Using Maple or some other program, you should find $x = 0.6276952185$ and then $D_\infty = 0.0868552356$.

15.41 Show that when the data do come from the theoretical probability distribution (assumed to be continuous) but are in bins of width w, then the limiting value D_∞ of the Kolmogorov distance is given by (15.362).

15.42 Suppose in a poll of 1000 likely voters, 510 have said they would vote for Nancy Pelosi. Redo example 15.17.

15.43 Suppose a hamiltonian has a quadratic term $H = ap^2 + bq^2$. Show that the mean energy at temperature T

$$\langle E \rangle = \int H e^{-\beta H} dp dq \Big/ \int e^{-\beta H} dp dq \tag{15.368}$$

is $kT/2$ if $a > 0$ and $b = 0$, $kT/2$ if $a = 0$ and $b > 0$, and kT if $a > 0$ and $b > 0$.