13

Probability and Statistics

13.1 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

$$P(A) = \sum_{j \in A} P_j. \tag{13.1}$$

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 $P(1,2) + P(2,1) = 1/18$.

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** $A \cap B$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{13.2}$$

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 $P(A, B) = P(A)\, P(B)$, 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** $P(A|B)$, the probability of A given B

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{13.3}$$

Also $P(B|A) = P(A \cap B)/P(A)$. The substitution $B \to B \cap C$ in (13.3)

gives $P(A|B,C) = P(A \cap B \cap C)/P(B \cap C)$. If we multiply (13.3) by $P(B)$, we get

$$P(A,B) = P(A \cap B) = P(B|A)\,P(A) = P(A|B)\,P(B). \tag{13.4}$$

Combination of (13.3 & 13.4) gives **Bayes's theorem** (Riley et al., 2006, p. 1132)

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \tag{13.5}$$

(Thomas Bayes, 1702–1761).

If the set B of outcomes or events is contained in the union of N mutually exclusive sets A_j of outcomes, then we must sum over them

$$P(B) = \sum_{j=1}^{N} P(B|A_j)\,P(A_j). \tag{13.6}$$

The probabilities $P(A_j)$ are called ***a priori*** probabilities. In this case, Bayes's theorem is (Roe, 2001, p. 119)

$$P(A_k|B) = \frac{P(B|A_k)\,P(A_k)}{\sum_{j=1}^{N} P(B|A_j)\,P(A_j)}. \tag{13.7}$$

If there are several B's, then a third form of Bayes's theorem is

$$P(A_k|B_\ell) = \frac{P(B_\ell|A_k)\,P(A_k)}{\sum_{j=1}^{N} P(B_\ell|A_j)\,P(A_j)}. \tag{13.8}$$

Example 13.1 (The 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 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 (13.6)

$$P(+) = P(+|D)\,P(D) + P(+|N)\,P(N) = 0.005993. \tag{13.9}$$

And by Bayes's theorem (13.5), the probability that a person who tests positive actually has the disease is only

$$P(D|+) = \frac{P(+|D)\,P(D)}{P(+)} = \frac{0.99 \times 0.001}{0.005993} = 0.165 \tag{13.10}$$

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 drug use in the CIA or disloyalty in the army, is difficult with a good test and absurd with a poor one like a polygraph. □

Example 13.2 (The 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?

We note that 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 does $P(W|Sw, RD) = 0$. Since the probability of picking the wrong door is $P(WD) = 2/3$, the probability of winning if one switches is

$$P(W|Sw) = P(W|Sw, WD)\, P(WD) + P(W|Sw, RD)\, P(RD) = 2/3. \tag{13.11}$$

The probability 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

$$P(W|Sp) = P(W|Sp, RD)\, P(RD) + P(W|Sp, WD)\, P(WD) = 1/3. \tag{13.12}$$

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 the second version (13.7) of Bayes's theorem says

$$P(x|B) = \frac{P(B|x)\, P(x)}{\int_{-\infty}^{\infty} P(B|x')\, P(x')\, dx'}. \tag{13.13}$$

Example 13.3 (A Tiny Poll) We ask 4 people 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

$$P(3|y) = 4\, y^3\, (1 - y). \tag{13.14}$$

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 (13.13) and our cheap poll give the probability distribution of the fraction

y who will vote for her as

$$P(y|3) = \frac{P(3|y)\,P(y)}{\int_0^1 P(3|y')\,P(y')\,dy'} = \frac{P(3|y)}{\int_0^1 P(3|y')\,dy'}$$
$$= \frac{4\,y^3\,(1-y)}{\int_0^1 4\,y'^3\,(1-y')\,dy'} = 20\,y^3\,(1-y). \tag{13.15}$$

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

$$\int_{1/2}^1 P(y|3)\,dy = \int_{1/2}^1 20\,y^3\,(1-y)\,dy = \frac{13}{16} \tag{13.16}$$

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

13.2 Mean and Variance

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

$$\sum_{j=1}^N P_j = 1. \tag{13.17}$$

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

$$E[x] = \mu = \langle x\rangle = \overline{x} = \sum_{j=1}^N x_j\,P_j. \tag{13.18}$$

The **expected value** $E[x]$ also is called the **expectation** of x or **expectation value** of x.

The **ℓth moment** is

$$E[x^\ell] = \mu_\ell = \langle x^\ell\rangle = \sum_{j=1}^N x_j^\ell P_j \tag{13.19}$$

and the **ℓth central moment** is

$$E[(x-\mu)^\ell] = \nu_\ell = \sum_{j=1}^N (x_j-\mu)^\ell P_j \tag{13.20}$$

where always $\mu_0 = \nu_0 = 1$ and $\nu_1 = 0$ (exercise 13.2).

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

$$V[x] \equiv E[(x - \langle x \rangle)^2] = \nu_2 = \sum_{j=1}^{N} (x_j - \langle x \rangle)^2 \; P_j \tag{13.21}$$

which one may write as (exercise 13.4)

$$V[x] = \langle x^2 \rangle - \langle x \rangle^2 \tag{13.22}$$

and the standard deviation σ is its square-root

$$\sigma = \sqrt{V[x]}. \tag{13.23}$$

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

$$\int P(x)\, dx = 1 \tag{13.24}$$

then the mean value is

$$E[x] = \mu = \langle x \rangle = \int x\, P(x)\, dx \tag{13.25}$$

and the ℓth moment is

$$E[x^\ell] = \mu_\ell = \langle x^\ell \rangle = \int x^\ell\, P(x)\, dx. \tag{13.26}$$

The ℓth central moment is

$$E[(x - \mu)^\ell] = \nu_\ell = \int (x - \mu)^\ell\, P(x)\, dx. \tag{13.27}$$

The variance of the distribution is the second central moment

$$V[x] = \nu_2 = \int (x - \langle x \rangle)^2 \; P(x)\, dx = \mu_2 - \mu^2 \tag{13.28}$$

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 $(-\infty, x)$

$$F(x) \equiv \Pr(-\infty, x) = \int_{-\infty}^{x} P(x')\, dx' = \int_{-\infty}^{x} f(x')\, dx' \tag{13.29}$$

a function that is necessarily **monotonic**

$$F'(x) = \Pr'(-\infty, x) = f(x) = P(x) \geq 0. \tag{13.30}$$

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. But usage of the Maxwell-Boltzmann distribution is too widespread in physics for me to observe this distinction.

Although a probability distribution $P(x)$ is normalized (13.24), 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**

$$E_{\text{abs}} \equiv \int |x - \mu| \, P(x) \, dx \tag{13.31}$$

diverge.

Example 13.4 (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 (3.70)

$$\psi(x) = \left(\frac{2}{\pi}\right)^{1/4} \frac{1}{\sqrt{a}} \, e^{-(x/a)^2}. \tag{13.32}$$

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

$$\Delta x \, \Delta p \geq \frac{\hbar}{2}. \tag{13.33}$$

□

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

$$\begin{aligned} E[ax^n y^m + bx^p y^q] &= \int (ax^n y^m + bx^p y^q) \, P(x, y) \, dxdy \\ &= a \int x^n y^m \, P(x, y) \, dxdy + b \int x^p y^q \, P(x, y) \, dxdy \\ &= a \, E[x^n y^m] + b \, E[x^p y^q]. \end{aligned} \tag{13.34}$$

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

The **correlation coefficient** or **covariance** of two variables x and y that

occur with a **joint distribution** $P(x, y)$ is

$$C[x, y] \equiv \int P(x, y)(x - \overline{x})(y - \overline{y})\, dxdy = \langle (x - \overline{x})(y - \overline{y}) \rangle = \langle x\, y \rangle - \langle x \rangle \langle y \rangle. \tag{13.35}$$

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

$$P(x, y) = P(x)\, P(y). \tag{13.36}$$

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 $x + y$

$$\langle (x + y)^2 \rangle - \langle x + y \rangle^2 = \langle x^2 \rangle - \langle x \rangle^2 + \langle y^2 \rangle - \langle y \rangle^2 + 2\, (\langle x\, y \rangle - \langle x \rangle \langle y \rangle) \tag{13.37}$$

is the sum

$$V[x + y] = V[x] + V[y] + 2\, C[x, y]. \tag{13.38}$$

It follows (exercise 13.6) that for any constants a and b the variance of $ax+by$ is

$$V[ax + by] = a^2\, V[x] + b^2\, V[y] + 2\, ab\, C[x, y]. \tag{13.39}$$

More generally (exercise 13.7), the variance of the sum $a_1 x_1 + a_2 x_2 + \cdots + a_N x_N$ is

$$V[a_1 x_1 + \cdots + a_N x_N] = \sum_{j=1}^{N} a_j^2\, V[x_j] \; + \sum_{j,k=1, j<k}^{N} 2 a_j a_k\, C[x_j, x_k]. \tag{13.40}$$

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 $a_1 x_1 + \cdots + a_N x_N$ is

$$V[a_1 x_1 + \cdots + a_N x_N] = \sum_{j=1}^{N} a_j^2\, V[x_j]. \tag{13.41}$$

13.3 The 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

$$\langle n \rangle = N\, p. \tag{13.42}$$

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

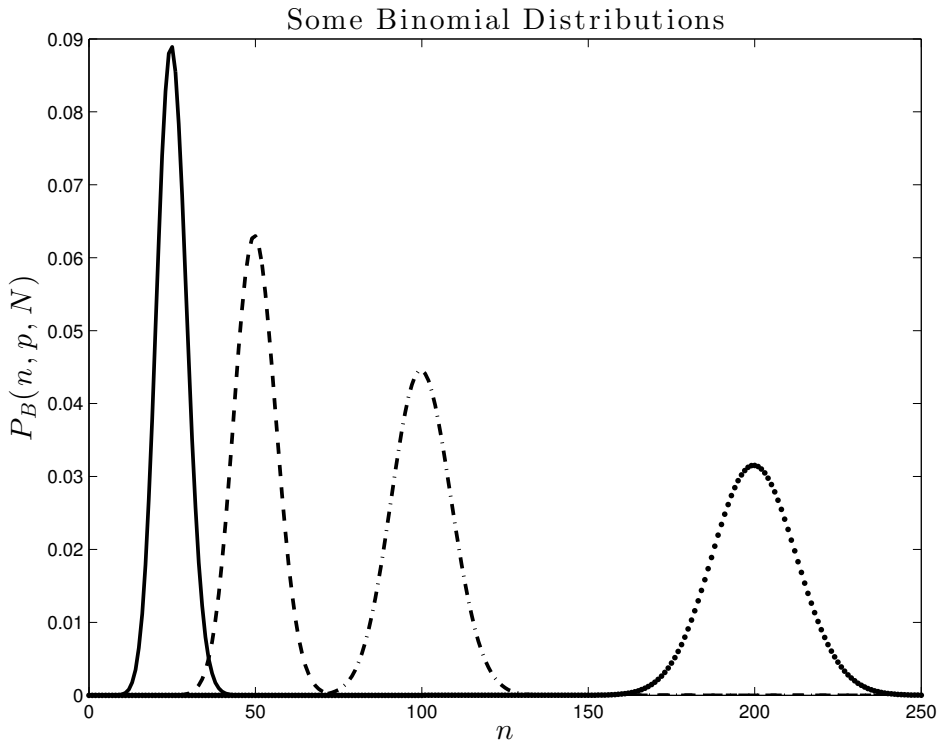

Figure 13.1 If the probability of success on any try is p, then the probability $P_B(n, p, N)$ of n successes in N tries is given by equation (13.43). For $p = 0.2$, this binomial probability distribution $P_B(n, p, N)$ is plotted against n for $N = 125$ (solid), 250 (dashes), 500 (dot dash), and 1000 tries (dots).

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

$$P_B(n, p, N) = \frac{N!}{n!\,(N-n)!}\, p^n\, q^{N-n} = \binom{N}{n} p^n\, (1-p)^{N-n}. \tag{13.43}$$

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

The sum of the probabilities $P_B(n, p, N)$ for all possible values of n is unity

$$\sum_{n=0}^{N} P_B(n, p, N) = (p + 1 - p)^N = 1. \tag{13.44}$$

In fig. 13.1, 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

$$\mu = \langle n \rangle_B = \sum_{n=0}^{N} n\, P_B(n,p,N) = \sum_{n=0}^{N} n \binom{N}{n} p^n q^{N-n} \tag{13.45}$$

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

$$\begin{aligned} \langle n \rangle_B &= p \frac{\partial}{\partial p} \sum_{n=0}^{N} \binom{N}{n} p^n q^{N-n} \\ &= p \frac{\partial}{\partial p}\, (p+q)^N = Np\,(p+q)^{N-1} = Np \end{aligned} \tag{13.46}$$

which verifies the estimate (13.42).

One may show (exercise 13.9) that the variance (13.21) of the binomial distribution is

$$V_B = \langle (n - \langle n \rangle)^2 \rangle = p\,(1-p)\,N. \tag{13.47}$$

Its standard deviation (13.23) is

$$\sigma_B = \sqrt{V_B} = \sqrt{p\,(1-p)\,N}. \tag{13.48}$$

The ratio of the width to the mean

$$\frac{\sigma_B}{\langle n \rangle_B} = \frac{\sqrt{p\,(1-p)\,N}}{Np} = \sqrt{\frac{1-p}{Np}} \tag{13.49}$$

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

Example 13.5 (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}. □

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

$$P_B(n+1,p,N) = \frac{p}{1-p}\,\frac{N-n}{n+1}\,P_B(n,p,N) \tag{13.50}$$

is helpful when N is big. But when N exceeds a few hundred, the formula (13.43) for $P_B(n,p,N)$ becomes unmanageable even in quadruple precision.

One way of computing $P_B(n,p,N)$ for large N is to use Srinivasa Ramanujan's correction (4.39) to Stirling's formula $N! \approx \sqrt{2\pi N}(N/e)^N$

$$N! \approx \sqrt{2\pi N}\left(\frac{N}{e}\right)^N \left(1 + \frac{1}{2N} + \frac{1}{8N^2}\right)^{1/6}. \tag{13.51}$$

When N and $N-n$, but not n, are big, one may use (13.51) for $N!$ and $(N-n)!$ in the formula (13.43) for $P_B(n,p,N)$ and so may show (exercise 13.11) that

$$P_B(n,p,N) \approx \frac{(pN)^n}{n!}\, q^{N-n}\, R_2(n,N) \tag{13.52}$$

in which

$$R_2(n,N) = \left(1 - \frac{n}{N}\right)^{n-1/2} \left(1 + \frac{1}{2N} + \frac{1}{8N^2}\right)^{1/6} \times \left[1 + \frac{1}{2(N-n)} + \frac{1}{8(N-n)^2}\right]^{-1/6} \tag{13.53}$$

tends to unity as $N \to \infty$ for any fixed n.

When all three factorials in $P_B(n,p,N)$ are huge, one may use Ramanujan's approximation (13.51) to show (exercise 13.12) that

$$P_B(n,p,N) \approx \sqrt{\frac{N}{2\pi n(N-n)}} \left(\frac{pN}{n}\right)^n \left(\frac{qN}{N-n}\right)^{N-n} R_3(n,N) \tag{13.54}$$

where

$$R_3(n,N) = \left(1 + \frac{1}{2n} + \frac{1}{8n^2}\right)^{-1/6} \left(1 + \frac{1}{2N} + \frac{1}{8N^2}\right)^{1/6} \times \left[1 + \frac{1}{2(N-n)} + \frac{1}{8(N-n)^2}\right]^{-1/6} \tag{13.55}$$

tends to unity as $N \to \infty$, $N - n \to \infty$, and $n \to \infty$.

Another way of coping with the unwieldy factorials in the binomial formula $P_B(n,p,N)$ is to use limiting forms of (13.43) due to Poisson and to Gauss.

13.4 The Poisson Distribution

Poisson took the two limits $N \to \infty$ and $p = \langle n \rangle / N \to 0$. So we let N and $N - n$, but not n, tend to infinity, and use (13.52) for the binomial distribution (13.43). Since $R_2(n,N) \to 1$ as $N \to \infty$, we get

$$P_B(n,p,N) \approx \frac{(pN)^n}{n!}\, q^{N-n} = \frac{\langle n \rangle^n}{n!}\, q^{N-n}. \tag{13.56}$$

Now $q = 1 - p = 1 - \langle n\rangle/N$, and so for any fixed n we have

$$\lim_{N\to\infty} q^{N-n} = \lim_{N\to\infty} \left(1 - \frac{\langle n\rangle}{N}\right)^N \left(1 - \frac{\langle n\rangle}{N}\right)^{-n} = e^{-\langle n\rangle}. \tag{13.57}$$

Thus as $N \to \infty$ with pN fixed at $\langle n\rangle$, the binomial distribution becomes the Poisson distribution

$$P_P(n, \langle n\rangle) = \frac{\langle n\rangle^n}{n!}\, e^{-\langle n\rangle}. \tag{13.58}$$

(Siméon-Denis Poisson, 1781–1840. Incidentally, *poisson* means *fish* and sounds like pwahsahn.)

The Poisson distribution is normalized to unity

$$\sum_{n=0}^{\infty} P_P(n, \langle n\rangle) = \sum_{n=0}^{\infty} \frac{\langle n\rangle^n}{n!}\, e^{-\langle n\rangle} = e^{\langle n\rangle}\, e^{-\langle n\rangle} = 1. \tag{13.59}$$

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

$$\begin{aligned}\mu = \sum_{n=0}^{\infty} n\, P_P(n, \langle n\rangle) &= \sum_{n=1}^{\infty} n\, \frac{\langle n\rangle^n}{n!}\, e^{-\langle n\rangle} = \langle n\rangle \sum_{n=1}^{\infty} \frac{\langle n\rangle^{(n-1)}}{(n-1)!}\, e^{-\langle n\rangle} \\ &= \langle n\rangle \sum_{n=0}^{\infty} \frac{\langle n\rangle^n}{n!}\, e^{-\langle n\rangle} = \langle n\rangle. \end{aligned} \tag{13.60}$$

As $N \to \infty$ and $p \to 0$ with $p\,N = \langle n\rangle$ fixed, the variance (13.47) of the binomial distribution tends to the limit

$$V_P = \lim_{\substack{N\to\infty \\ p\to 0}} V_B = \lim_{\substack{N\to\infty \\ p\to 0}} p\,(1-p)\,N = \langle n\rangle. \tag{13.61}$$

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

$$V_P = \langle (n - \langle n\rangle)^2\rangle = \langle n\rangle = \mu \tag{13.62}$$

as one may show directly (exercise 13.13).

Example 13.6 (Coherent States) The **coherent state** $|\alpha\rangle$ introduced in equation (2.138)

$$|\alpha\rangle = e^{-|\alpha|^2/2} e^{\alpha a^\dagger} |0\rangle = e^{-|\alpha|^2/2} \sum_{n=0}^{\infty} \frac{\alpha^n}{\sqrt{n!}} |n\rangle \tag{13.63}$$

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 $\langle n|\alpha\rangle$

$$P(n) = |\langle n|\alpha\rangle|^2 = \frac{|\alpha|^{2n}}{n!}\, e^{-|\alpha|^2} \tag{13.64}$$

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

13.5 The Gaussian Distribution

Gauss considered the binomial distribution in the limit $N \to \infty$ with the probability p fixed. In this limit, the binomial probability

$$P_B(n, p, N) = \frac{N!}{n!\,(N-n)!}\, p^n\, q^{N-n} \tag{13.65}$$

is very tiny unless n is near pN which means that $n \approx pN$ and $N - n \approx (1-p)N = qN$ are comparable. So the limit $N \to \infty$ effectively is one in which n and $N - n$ also tend to infinity. The approximation (13.54)

$$P_B(n, p, N) \approx \sqrt{\frac{N}{2\pi n(N-n)}}\, \left(\frac{pN}{n}\right)^n \left(\frac{qN}{N-n}\right)^{N-n} R_3(n, N) \tag{13.66}$$

applies in which $R_3(n, N) \to 1$ as N, $N-n$, and n all increase without limit.

Because the probability $P_B(n, p, N)$ is negligible unless $n \approx pN$, 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 may write the square-root as

$$\begin{aligned}\sqrt{\frac{N}{2\pi\, n\,(N-n)}} &= \frac{1}{\sqrt{2\pi\, N\, [(pN+y)/N]\, [(qN-y)/N]}} \\ &= \frac{1}{\sqrt{2\pi\, pqN\,(1+y/pN)\,(1-y/qN)}}.\end{aligned} \tag{13.67}$$

Since y remains finite as $N \to \infty$, we get in this limit

$$\lim_{N\to\infty} \sqrt{\frac{N}{2\pi\, n\,(N-n)}} = \frac{1}{\sqrt{2\pi\, pqN}}. \tag{13.68}$$

Substituting $pN + y$ for n and $qN - y$ for $N - n$ in (13.66), we find

$$\begin{aligned} P_B(n,p,N) &\approx \frac{1}{\sqrt{2\pi\, pqN}} \left(\frac{pN}{pN+y}\right)^{pN+y} \left(\frac{qN}{qN-y}\right)^{qN-y} \\ &= \frac{1}{\sqrt{2\pi\, pqN}} \left(1+\frac{y}{pN}\right)^{-(pN+y)} \left(1-\frac{y}{qN}\right)^{-(qN-y)} \end{aligned} \tag{13.69}$$

which implies

$$\ln\left[P_B(n,p,N)\sqrt{2\pi\, pqN}\right] \approx -(pN+y)\ln\left[1+\frac{y}{pN}\right] - (qN-y)\ln\left[1-\frac{y}{qN}\right]. \tag{13.70}$$

The first two terms of the power series (4.88) for $\ln(1+\epsilon)$ are

$$\ln(1+\epsilon) \approx \epsilon - \frac{1}{2}\epsilon^2. \tag{13.71}$$

So using this expansion for $\ln(1+y/pN)$ and also for $\ln(1-y/qN)$, we get

$$\begin{aligned} \ln\left(P_B(n,p,N)\sqrt{2\pi\, pqN}\right) &\approx -(pN+y)\left[\frac{y}{pN} - \frac{1}{2}\left(\frac{y}{pN}\right)^2\right] \\ &\quad - (qN-y)\left[-\frac{y}{qN} - \frac{1}{2}\left(\frac{y}{qN}\right)^2\right] \approx -\frac{y^2}{2pqN}. \end{aligned} \tag{13.72}$$

Gauss's approximation to the binomial probability distribution thus is

$$P_{BG}(n,p,N) = \frac{1}{\sqrt{2\pi pqN}} \exp\left(-\frac{(n-pN)^2}{2pqN}\right) \tag{13.73}$$

in which we've replaced y by $n - pN$ and $1 - p$ by q.

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

$$P_G(x,p,N) = \frac{1}{\sqrt{2\pi pqN}} \exp\left(-\frac{(x-pN)^2}{2pqN}\right) \tag{13.74}$$

which is (exercise 13.14) 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**

$$P_G(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{13.75}$$

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 13.16).

Actin Fibers in HELA Cells

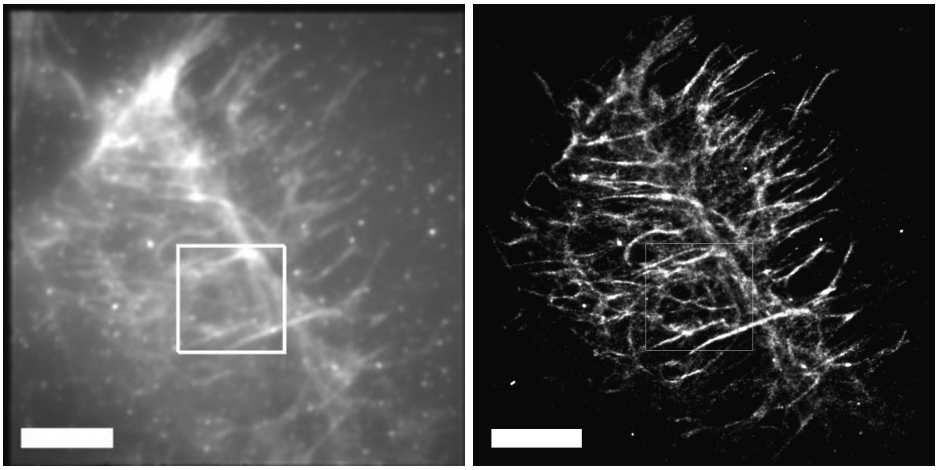

Figure 13.2 Conventional (left, fuzzy) and dSTORM (right, sharp) images of actin fibers in HELA cells. The actin is labeled with Alexa Fluor 647 Phalloidin. The white rectangles are 5 microns in length. Images courtesy of Fang Huang and Keith Lidke.

Example 13.7 (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 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

$$P(\boldsymbol{y_i}) = \frac{1}{2\pi\sigma^2}\, e^{-(\boldsymbol{y_i}-\boldsymbol{x})^2/2\sigma^2} \tag{13.76}$$

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

$$P = \prod_{i=1}^{N} P(\boldsymbol{y_i}) = d^N \prod_{i=1}^{N} e^{-(\boldsymbol{y_i}-\boldsymbol{x})^2/(2\sigma^2)} = d^N \exp\left[-\sum_{i=1}^{N} (\boldsymbol{y_i}-\boldsymbol{x})^2/(2\sigma^2)\right] \tag{13.77}$$

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

$$\frac{\partial P}{\partial x_k} = 0 = P\,\frac{\partial}{\partial x_k}\left[-\sum_{i=1}^{N} (\boldsymbol{y_i}-\boldsymbol{x})^2/(2\sigma^2)\right] = \frac{P}{\sigma^2}\sum_{i=1}^{N} (y_{ik} - x_k)\,. \tag{13.78}$$

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

the observed points $\boldsymbol{y_i}$

$$\boldsymbol{x} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{y_i}. \tag{13.79}$$

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.

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. Figure 13.2 shows conventional (left, fuzzy) and dSTORM (right, sharp) images of actin filaments. 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. □

13.6 The Error Function ERF

The probability that a random variable x distributed according to Gauss's distribution (13.75) has a value between $\mu - \delta$ and $\mu + \delta$ is

$$\begin{aligned} P(|x-\mu|<\delta) &= \int_{\mu-\delta}^{\mu+\delta} P_G(x,\mu,\sigma)\,dx = \frac{1}{\sigma\sqrt{2\pi}}\int_{\mu-\delta}^{\mu+\delta} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx \\ &= \frac{1}{\sigma\sqrt{2\pi}}\int_{-\delta}^{\delta} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx = \frac{2}{\sqrt{\pi}}\int_0^{\delta/\sigma\sqrt{2}} e^{-t^2}\,dt. \end{aligned} \tag{13.80}$$

The last integral is the error function

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x e^{-t^2} dt \tag{13.81}$$

so in terms of it the probability that x lies within δ of the mean μ is

$$P(|x-\mu|<\delta) = \operatorname{erf}\left(\frac{\delta}{\sigma\sqrt{2}}\right). \tag{13.82}$$

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

$$\begin{aligned} P(|x-\mu|<\sigma) &= \operatorname{erf}(1/\sqrt{2}) = 0.6827 \\ P(|x-\mu|<2\sigma) &= \operatorname{erf}(2/\sqrt{2}) = 0.9545 \\ P(|x-\mu|<3\sigma) &= \operatorname{erf}(3/\sqrt{2}) = 0.9973. \end{aligned} \tag{13.83}$$

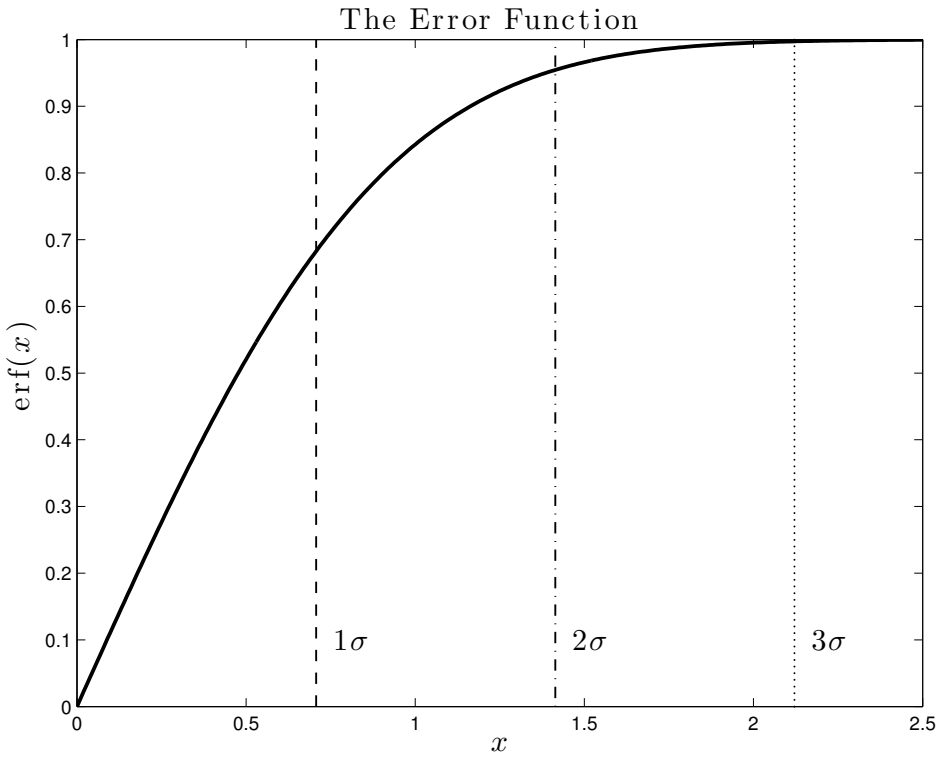

Figure 13.3 The error function $\mathrm{erf}\,(x)$ is plotted for $0 < x < 2.5$. The vertical lines are at $x = \delta/(\sigma\sqrt{2})$ for $\delta = \sigma$, 2σ, and 3σ with $\sigma = 1/\sqrt{2}$.

The error function $\mathrm{erf}\,(x)$ is plotted in Fig. 13.3 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 13.17)

$$P(a < x < b) = \frac{1}{2}\left[\mathrm{erf}\left(\frac{b-\mu}{\sigma\sqrt{2}}\right) - \mathrm{erf}\left(\frac{a-\mu}{\sigma\sqrt{2}}\right)\right]. \tag{13.84}$$

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

$$P(-\infty, x) = \frac{1}{2}\left[\mathrm{erf}\left(\frac{x}{\sqrt{2}}\right) - \mathrm{erf}\left(\frac{-\infty}{\sqrt{2}}\right)\right] = \frac{1}{2}\left[\mathrm{erf}\left(\frac{x}{\sqrt{2}}\right) + 1\right]. \tag{13.85}$$

The complement erfc of the error function is defined as

$$\mathrm{erfc}\,(x) = \frac{2}{\sqrt{\pi}}\int_x^\infty e^{-t^2} dt = 1 - \mathrm{erf}\,(x) \tag{13.86}$$

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 13.8 (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 (13.73)

$$P_B(n,p,N) = \frac{N!}{n!\,(N-n)!}\,p^n\,q^{N-n} \approx \frac{1}{\sqrt{2\pi pqN}}\,\exp\left(-\frac{(n-pN)^2}{2pqN}\right). \tag{13.87}$$

Then using (13.84) with $\mu = pN$, we find (exercise 13.15)

$$P_B(n,p,N) \approx \frac{1}{2}\left[\text{erf}\left(\frac{n+\frac{1}{2}-pN}{\sqrt{2pqN}}\right) - \text{erf}\left(\frac{n-\frac{1}{2}-pN}{\sqrt{2pqN}}\right)\right] \tag{13.88}$$

which we can sum over the integer n to get

$$\sum_{n=n_1}^{n_2} P_B(n,p,N) \approx \frac{1}{2}\left[\text{erf}\left(\frac{n_2+\frac{1}{2}-pN}{\sqrt{2pqN}}\right) - \text{erf}\left(\frac{n_1-\frac{1}{2}-pN}{\sqrt{2pqN}}\right)\right] \tag{13.89}$$

which is easy to evaluate. □

Example 13.9 (Polls) Suppose in a poll of 1000 likely voters, 600 have said they would vote for Nancy Pelosi. Repeating the analysis of example 13.3, 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 (13.87)

$$\begin{aligned} P(600|y) = P_B(600,y) &= \binom{1000}{600} y^{600}\,(1-y)^{400} \\ &\approx \frac{1}{10\sqrt{20\pi y(1-y)}}\,\exp\left(-\frac{20(3-5y)^2}{y(1-y)}\right) \end{aligned} \tag{13.90}$$

and so the probability density that a fraction y of the voters will vote for her is

$$\begin{aligned} P(y|600) &= \frac{P(600|y)}{\int_0^1 P(600,y')\,dy'} \\ &= \frac{[y(1-y)]^{-1/2}\,\exp\left(-\frac{20(3-5y)^2}{y(1-y)}\right)}{\int_0^1 [y'(1-y')]^{-1/2}\,\exp\left(-\frac{20(3-5y')^2}{y'(1-y')}\right)\,dy'}. \end{aligned} \tag{13.91}$$

This normalized probability distribution is negligible except for y near 3/5

(exercise 13.18), where it is approximately Gauss's distribution

$$P(y|600) \approx \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y-3/5)^2}{2\sigma^2}\right) \tag{13.92}$$

with mean $\mu = 3/5$ and variance

$$\sigma^2 = \frac{3}{12500} = 2.4 \times 10^{-4}. \tag{13.93}$$

The probability that $y > 1/2$ then is by (13.84)

$$\begin{aligned} P(\tfrac{1}{2} < y < 1) &= \frac{1}{2}\left[\operatorname{erf}\left(\frac{1-\mu}{\sigma\sqrt{2}}\right) - \operatorname{erf}\left(\frac{1/2-\mu}{\sigma\sqrt{2}}\right)\right] \\ &= \frac{1}{2}\left[\operatorname{erf}\left(\frac{20}{\sqrt{1.2}}\right) - \operatorname{erf}\left(\frac{-5}{\sqrt{1.2}}\right)\right] \approx 1. \end{aligned} \tag{13.94}$$

The probability that $y < 1/2$ is 5.4×10^{-11}. □

13.7 The Maxwell-Boltzmann distribution

It is a small jump from Gauss's distribution (13.75) 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 hit fore and aft with equal probabilities by other molecules. If each hit increases or decreases its speed by dv, then after n aft hits and $N - n$ fore hits, the speed v_x of a molecule initially at rest would be

$$v_x = ndv - (N-n)dv = (2n-N)dv. \tag{13.95}$$

The probability of this speed is given by Gauss's approximation (13.73) to the binomial distribution $P_B(n, \frac{1}{2}, N)$ as

$$P_{BG}(n, \tfrac{1}{2}, N) = \sqrt{\frac{2}{\pi N}} \exp\left(-\frac{(2n-N)^2}{2N}\right) = \sqrt{\frac{2}{\pi N}} \exp\left(-\frac{v_x^2}{2Ndv^2}\right). \tag{13.96}$$

This argument applies to any physical variable subject to unbiased random fluctuations. It is why Gauss's distribution describes statistical errors and why it occurs so often in Nature as to be called the normal distribution.

We now write the argument of the exponential in terms of the temperature T and Boltzmann's constant k by setting $N = kT/(m\,dv^2)$ so that

$$-\frac{\frac{1}{2}v_x^2}{Ndv^2} = -\frac{\frac{1}{2}mv_x^2}{mNdv^2} = -\frac{\frac{1}{2}mv_x^2}{kT}. \tag{13.97}$$

Then with $dv_x = 2dv$, we have

$$P_G(v_x)dv_x = \sqrt{\frac{2m}{\pi kT}}\, dv\, \exp\left(-\frac{\frac{1}{2}mv_x^2}{kT}\right) = \sqrt{\frac{m}{2\pi kT}}\, dv_x\, \exp\left(-\frac{\frac{1}{2}mv_x^2}{kT}\right). \tag{13.98}$$

Gauss's distribution is normalized to unity because it is the limit of the binomial distribution (13.44)

$$\int_{-\infty}^{\infty} \sqrt{\frac{m}{2\pi kT}}\, \exp\left(-\frac{\frac{1}{2}mv_x^2}{kT}\right)\, dv_x = 1 \tag{13.99}$$

as you may verify by explicit integration.

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

$$P_{MB}(\boldsymbol{v})d^3v = P_G(v_x)\, P_G(v_y)\, P_G(v_z)d^3v = \left(\frac{m}{2\pi kT}\right)^{3/2}\, e^{-\frac{1}{2}m\boldsymbol{v}^2/(kT)} 4\pi v^2 dv. \tag{13.100}$$

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

$$\langle \boldsymbol{v} \rangle = \int \boldsymbol{v}\, P_{MB}(\boldsymbol{v})d^3v = \boldsymbol{0} \tag{13.101}$$

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 $\sigma_x^2 = \sigma_y^2 = \sigma_z^2 = kT/m$

$$\langle v^2 \rangle = V[\boldsymbol{v}^2] = \int \boldsymbol{v}^2\, P_{MB}(\boldsymbol{v})\, d^3v = 3kT/m \tag{13.102}$$

which is the familiar statement

$$\frac{1}{2}m\langle v^2 \rangle = \frac{3}{2}\, kT \tag{13.103}$$

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

13.8 Diffusion

We may apply the same reasoning as in the preceding section (13.7) 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 (13.96), the probability of being at x is given by Gauss's approximation (13.73) to the binomial

distribution $P_B(n, \frac{1}{2}, N)$ as

$$P_{BG}(n, \tfrac{1}{2}, N) = \sqrt{\frac{2}{\pi N}}\, \exp\left(-\frac{(2n-N)^2}{2N}\right) = \sqrt{\frac{2}{\pi N}}\, \exp\left(-\frac{x^2}{2N dx^2}\right). \tag{13.104}$$

In terms of the diffusion constant

$$D = \frac{N dx^2}{2t} \tag{13.105}$$

this distribution is

$$P_G(x) = \left(\frac{1}{4\pi D t}\right)^{1/2} \exp\left(-\frac{x^2}{4Dt}\right) \tag{13.106}$$

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

In three dimensions, this gaussian distribution is the product

$$P(\boldsymbol{r}, t) = P_G(x)\, P_G(y)\, P_G(z) = \left(\frac{1}{4\pi D t}\right)^{3/2} \exp\left(-\frac{\boldsymbol{r}^2}{4Dt}\right). \tag{13.107}$$

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

$$\langle \boldsymbol{r}^2 \rangle = V[\boldsymbol{r}] = 3\sigma^2 = \int \boldsymbol{r}^2\, P(\boldsymbol{r}, t)\, d^3r = 6\, D\, t. \tag{13.108}$$

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

$$\dot{P}(\boldsymbol{r}, t) = D\, \nabla^2 P(\boldsymbol{r}, t) \tag{13.109}$$

in which the dot means time derivative.

13.9 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

$$m\, \frac{d\boldsymbol{v}(t)}{dt} = \boldsymbol{F}(t). \tag{13.110}$$

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

$$\boldsymbol{F}(t) = -\boldsymbol{v}(t)/B + \boldsymbol{f}(t) \tag{13.111}$$

so that

$$m\,\frac{d\boldsymbol{v}(t)}{dt} = -\frac{\boldsymbol{v}(t)}{B} + \boldsymbol{f}(t). \tag{13.112}$$

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

$$\langle \boldsymbol{f}(t) \rangle = \mathbf{0}. \tag{13.113}$$

Thus the ensemble average of the velocity satisfies

$$m\,\frac{d\langle \boldsymbol{v} \rangle}{dt} = -\frac{\langle \boldsymbol{v} \rangle}{B} \tag{13.114}$$

whose solution with $\tau = mB$ is

$$\langle \boldsymbol{v(t)} \rangle = \langle \boldsymbol{v(0)} \rangle\, e^{-t/\tau}. \tag{13.115}$$

The instantaneous equation (13.112) divided by the mass m is

$$\frac{d\boldsymbol{v}(t)}{dt} = -\frac{\boldsymbol{v}(t)}{\tau} + \boldsymbol{a}(t) \tag{13.116}$$

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

$$\left\langle \boldsymbol{r} \cdot \frac{d\boldsymbol{v}}{dt} \right\rangle = -\frac{\langle \boldsymbol{r} \cdot \boldsymbol{v} \rangle}{\tau} + \langle \boldsymbol{r} \cdot \boldsymbol{a} \rangle. \tag{13.117}$$

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

$$\left\langle \boldsymbol{r} \cdot \frac{d\boldsymbol{v}}{dt} \right\rangle = -\frac{\langle \boldsymbol{r} \cdot \boldsymbol{v} \rangle}{\tau}. \tag{13.118}$$

Now

$$\frac{1}{2}\frac{d\,\boldsymbol{r}^2}{dt} = \frac{1}{2}\frac{d}{dt}\,(\boldsymbol{r} \cdot \boldsymbol{r}) = \boldsymbol{r} \cdot \boldsymbol{v} \tag{13.119}$$

and so

$$\frac{1}{2}\frac{d^2\boldsymbol{r}^2}{dt^2} = \boldsymbol{r} \cdot \frac{d\boldsymbol{v}}{dt} + \boldsymbol{v}^2. \tag{13.120}$$

The ensemble average of this equation gives us

$$\frac{d^2\langle \boldsymbol{r}^2\rangle}{dt^2} = 2\left\langle \boldsymbol{r}\cdot\frac{d\boldsymbol{v}}{dt}\right\rangle + 2\langle \boldsymbol{v}^2\rangle \tag{13.121}$$

or in view of (13.118)

$$\frac{d^2\langle \boldsymbol{r}^2\rangle}{dt^2} = -2\frac{\langle \boldsymbol{r}\cdot\boldsymbol{v}\rangle}{\tau} + 2\langle \boldsymbol{v}^2\rangle. \tag{13.122}$$

We now use (13.119) 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

$$\frac{d^2\langle \boldsymbol{r}^2\rangle}{dt^2} = -\frac{1}{\tau}\frac{d\langle \boldsymbol{r}^2\rangle}{dt} + 2\langle \boldsymbol{v}^2\rangle. \tag{13.123}$$

If the fluid is in equilibrium, then the ensemble average of $\boldsymbol{v}^2$ is given by the Maxwell-Boltzmann value (13.103)

$$\langle \boldsymbol{v}^2\rangle = \frac{3kT}{m} \tag{13.124}$$

and so the acceleration (13.123) is

$$\frac{d^2\langle \boldsymbol{r}^2\rangle}{dt^2} + \frac{1}{\tau}\frac{d\langle \boldsymbol{r}^2\rangle}{dt} = \frac{6kT}{m}. \tag{13.125}$$

which we can integrate.

The general solution (6.13) 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

$$\langle \boldsymbol{r}^2(t)\rangle = U + W\,e^{-t/\tau} + 6kT\tau t/m \tag{13.126}$$

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

$$\langle \boldsymbol{r}^2(0)\rangle = 0 \tag{13.127}$$

which implies that

$$U + W = 0. \tag{13.128}$$

And since the particles start out at $\boldsymbol{r} = \boldsymbol{0}$ with an isotropic distribution of initial velocities, the formula (13.119) for $\dot{r}^2$ implies that at $t = 0$

$$\left.\frac{d\,\langle \boldsymbol{r}^2\rangle}{dt}\right|_{t=0} = 2\langle \boldsymbol{r}(0)\cdot\boldsymbol{v}(0)\rangle = 0. \tag{13.129}$$

This boundary condition means that our solution (13.126) must satisfy

$$\left.\frac{d\,\langle \boldsymbol{r}^2(t)\rangle}{dt}\right|_{t=0} = -\frac{W}{\tau} + \frac{6kT\tau}{m} = 0. \tag{13.130}$$

Thus $W = -U = 6kT\tau^2/m$, and so our solution (13.126) is

$$\langle \boldsymbol{r}^2(t)\rangle = \frac{6kT\tau^2}{m}\,\left[\frac{t}{\tau} + e^{-t/\tau} - 1\right]. \tag{13.131}$$

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

$$\langle \boldsymbol{r}^2(t)\rangle = \frac{6kT\tau^2}{m}\,\left[\frac{t^2}{2\tau^2}\right] = \frac{3kT}{m}\,t^2 = \langle v^2\rangle\, t^2. \tag{13.132}$$

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

$$\langle \boldsymbol{r}^2(t)\rangle = \frac{6kT\tau}{m}\,t = 6\,B\,kT\,t. \tag{13.133}$$

The **diffusion constant** D is defined by

$$\langle \boldsymbol{r}^2(t)\rangle = 6\,D\,t \tag{13.134}$$

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

$$D = B\,kT \tag{13.135}$$

which often is written in terms of the **viscous-friction coefficient** ζ

$$\zeta \equiv \frac{1}{B} = \frac{m}{\tau} \tag{13.136}$$

as

$$\zeta\,D = kT. \tag{13.137}$$

This equation expresses 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. It 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.

13.10 The 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

$$m\frac{d\boldsymbol{v}}{dt} = -\frac{\boldsymbol{v}}{B} + q\boldsymbol{E} + \boldsymbol{f}. \tag{13.138}$$

The mean value of its velocity will then satisfy the differential equation

$$\left\langle \frac{d\boldsymbol{v}}{dt} \right\rangle = -\frac{\langle \boldsymbol{v} \rangle}{\tau} + \frac{q\boldsymbol{E}}{m} \tag{13.139}$$

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

$$\langle \boldsymbol{v}(t) \rangle_{pi} = \frac{q\tau \boldsymbol{E}}{m} = qB\boldsymbol{E}. \tag{13.140}$$

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 (6.13), the general solution $\langle \boldsymbol{v}(t) \rangle$ to equation (13.139) is (exercise 13.19) the sum of $\langle \boldsymbol{v}(t) \rangle_{pi}$ and $\langle \boldsymbol{v}(t) \rangle_{gh}$

$$\langle \boldsymbol{v}(t) \rangle = qB\boldsymbol{E} + [\langle \boldsymbol{v}(0) \rangle - qB\boldsymbol{E}]\, e^{-t/\tau}. \tag{13.141}$$

By applying the tricks of the previous section (13.9), one may show (exercise 13.20) that the variance of the position $\boldsymbol{r}$ about its mean $\langle \boldsymbol{r}(t) \rangle$ is

$$\left\langle (\boldsymbol{r} - \langle \boldsymbol{r}(t) \rangle)^2 \right\rangle = \frac{6kT\tau^2}{m} \left(\frac{t}{\tau} - 1 + e^{-t/\tau} \right) \tag{13.142}$$

where $\langle \boldsymbol{r}(t) \rangle = (q\tau^2 \boldsymbol{E}/m) \left(t/\tau - 1 + e^{-t/\tau} \right)$ if $\langle \boldsymbol{r}(0) \rangle = \langle \boldsymbol{v}(0) \rangle = 0$. So for times $t \gg \tau$, this variance is

$$\left\langle (\boldsymbol{r} - \langle \boldsymbol{r}(t) \rangle)^2 \right\rangle = \frac{6kT\tau t}{m}. \tag{13.143}$$

Since the diffusion constant D is defined by (13.134) as

$$\left\langle (\boldsymbol{r} - \langle \boldsymbol{r}(t) \rangle)^2 \right\rangle = 6\, D\, t \tag{13.144}$$

we arrive at the Einstein-Nernst relation

$$D = BkT = \frac{qB}{q} kT = \frac{\mu}{q} kT \tag{13.145}$$

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

13.11 Fluctuation and Dissipation

Let's look again at Langevin's equation (13.116) but with u as the independent variable

$$\frac{d\boldsymbol{v}(u)}{du} + \frac{\boldsymbol{v}(u)}{\tau} = \boldsymbol{a}(u). \tag{13.146}$$

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

$$\left(\frac{d\boldsymbol{v}}{du} + \frac{\boldsymbol{v}}{\tau}\right) e^{u/\tau} = \frac{d}{du}\left(\boldsymbol{v}\, e^{u/\tau}\right) = \boldsymbol{a}(u)\, e^{u/\tau} \tag{13.147}$$

and integrate from 0 to t

$$\int_0^t \frac{d}{du}\left(\boldsymbol{v}\, e^{u/\tau}\right) du = \boldsymbol{v}(t)\, e^{t/\tau} - \boldsymbol{v}(0) = \int_0^t \boldsymbol{a}(u)\, e^{u/\tau}\, du \tag{13.148}$$

then we get

$$\boldsymbol{v}(t) = e^{-t/\tau}\, \boldsymbol{v}(0) + e^{-t/\tau} \int_0^t \boldsymbol{a}(u)\, e^{u/\tau}\, du. \tag{13.149}$$

Thus the ensemble average of the square of the velocity is

$$\langle \boldsymbol{v}^2(t)\rangle = e^{-2t/\tau}\, \langle \boldsymbol{v}^2(0)\rangle + 2e^{-2t/\tau} \int_0^t \langle \boldsymbol{v}(0)\cdot \boldsymbol{a}(u)\rangle\, e^{u/\tau}\, du \tag{13.150}$$
$$+e^{-2t/\tau} \int_0^t \int_0^t \langle \boldsymbol{a}(u_1)\cdot \boldsymbol{a}(u_2)\rangle\, e^{(u_1+u_2)/\tau}\, du_1 du_2.$$

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

$$\langle \boldsymbol{v}^2(t)\rangle = e^{-2t/\tau}\, \langle \boldsymbol{v}^2(0)\rangle + e^{-2t/\tau} \int_0^t \int_0^t \langle \boldsymbol{a}(u_1)\cdot \boldsymbol{a}(u_2)\rangle\, e^{(u_1+u_2)/\tau}\, du_1 du_2. \tag{13.151}$$

The ensemble average

$$C(u_1, u_2) = \langle \boldsymbol{a}(u_1)\cdot \boldsymbol{a}(u_2)\rangle \tag{13.152}$$

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:

$$C(t, t+s) = \langle \boldsymbol{A}(t)\cdot \boldsymbol{A}(t+s)\rangle \equiv C(s). \tag{13.153}$$

2. The autocorrelation function for $s = 0$ is necessarily non-negative

$$C(t, t) = \langle \boldsymbol{A}(t)\cdot \boldsymbol{A}(t)\rangle = \langle \boldsymbol{A}(t)^2\rangle \geq 0. \tag{13.154}$$

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

$$\langle |\boldsymbol{A}(t_1) \pm \boldsymbol{A}(t_2)|^2 \rangle = \langle \boldsymbol{A}(t_1)^2 \rangle + \langle \boldsymbol{A}(t_2)^2 \rangle \pm 2\langle \boldsymbol{A}(t_1) \cdot \boldsymbol{A}(t_2) \rangle \geq 0 \quad (13.155)$$

which implies that

$$-C(t_1, t_2) \leq \frac{1}{2}\left(C(t_1, t_1) + C(t_2, t_2)\right) \geq C(t_1, t_2) \quad (13.156)$$

or

$$|C(t_1, t_2)| \leq \frac{1}{2}\left(C(t_1, t_1) + C(t_2, t_2)\right). \quad (13.157)$$

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

$$C(t_1, t_2) = \langle \boldsymbol{A}(t_1) \cdot \boldsymbol{A}(t_2) \rangle = \langle \boldsymbol{A}(t_2) \cdot \boldsymbol{A}(t_1) \rangle = C(t_2, t_1). \quad (13.158)$$

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

$$\langle \boldsymbol{A}(t) \rangle = \boldsymbol{0} \quad (13.159)$$

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

$$\langle \boldsymbol{A}(t_1) \cdot \boldsymbol{A}(t_2) \rangle \to \langle \boldsymbol{A}(t_1) \rangle \cdot \langle \boldsymbol{A}(t_2) \rangle = 0 \quad (13.160)$$

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

In terms of the autocorrelation function $C(u_1, u_2) = \langle \boldsymbol{a}(u_1) \cdot \boldsymbol{a}(u_2) \rangle$ of the acceleration, the variance of the velocity (13.151) is

$$\langle \boldsymbol{v}^2(t) \rangle = e^{-2t/\tau} \langle \boldsymbol{v}^2(0) \rangle + e^{-2t/\tau} \int_0^t \int_0^t C(u_1, u_2)\, e^{(u_1+u_2)/\tau}\, du_1 du_2. \quad (13.161)$$

Since $C(u_1, u_2)$ is big only for tiny values of $|u_2 - u_1|$, it makes sense to change variables to

$$s = u_2 - u_1 \quad \text{and} \quad w = \frac{1}{2}(u_1 + u_2). \quad (13.162)$$

The element of area then is by (12.6–12.14)

$$du_1 \wedge du_2 = dw \wedge ds \quad (13.163)$$

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

$$\langle \boldsymbol{v}^2(t) \rangle = e^{-2t/\tau} \langle \boldsymbol{v}^2(0) \rangle + e^{-2t/\tau} \int_0^{t/2} e^{2w/\tau} dw \int_{-2w}^{2w} C(s)\, ds + e^{-2t/\tau} \int_{t/2}^{t} e^{2w/\tau} dw \int_{-2(t-w)}^{2(t-w)} C(s)\, ds. \tag{13.164}$$

Since by (13.160) the autocorrelation function $C(s)$ vanishes outside a narrow window of width $2T$, we may approximate each of the s-integrals by

$$C = \int_{-\infty}^{\infty} C(s)\, ds. \tag{13.165}$$

It follows then that

$$\begin{aligned} \langle \boldsymbol{v}^2(t) \rangle &= e^{-2t/\tau} \langle \boldsymbol{v}^2(0) \rangle + C\, e^{-2t/\tau} \int_0^t e^{2w/\tau} dw \\ &= e^{-2t/\tau} \langle \boldsymbol{v}^2(0) \rangle + C\, e^{-2t/\tau} \frac{\tau}{2} \left(e^{2t/\tau} - 1 \right) \\ &= e^{-2t/\tau} \langle \boldsymbol{v}^2(0) \rangle + C \frac{\tau}{2} \left(1 - e^{-2t/\tau} \right). \end{aligned} \tag{13.166}$$

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

$$\lim_{t \to \infty} \langle \boldsymbol{v}^2(t) \rangle = C \frac{\tau}{2} = \frac{3kT}{m} \tag{13.167}$$

which implies that

$$C = \frac{6kT}{m\tau} \quad \text{or} \quad \frac{1}{B} = \frac{m^2 C}{6kT}. \tag{13.168}$$

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

$$\langle \boldsymbol{v}^2(t) \rangle = e^{-2t/\tau} \langle \boldsymbol{v}^2(0) \rangle + \frac{3kT}{m} \left(1 - e^{-2t/\tau} \right). \tag{13.169}$$

Referring back to the definition (13.136) of the viscous-friction coefficient $\zeta = 1/B$, we see that ζ is related to the integral

$$\zeta = \frac{1}{B} = \frac{m^2}{6kT} C = \frac{m^2}{6kT} \int_{-\infty}^{\infty} \langle \boldsymbol{a}(0) \cdot \boldsymbol{a}(s) \rangle ds = \frac{1}{6kT} \int_{-\infty}^{\infty} \langle \boldsymbol{f}(0) \cdot \boldsymbol{f}(s) \rangle ds \tag{13.170}$$

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 (13.169) for $\langle \boldsymbol{v}^2(t)\rangle$ into the expression (13.123) for the acceleration of $\langle \boldsymbol{r}^2\rangle$, then we get

$$\frac{d^2\langle \boldsymbol{r}^2(t)\rangle}{dt^2} = -\frac{1}{\tau}\frac{d\langle \boldsymbol{r}^2(t)\rangle}{dt} + 2e^{-2t/\tau}\,\langle \boldsymbol{v}^2(0)\rangle + \frac{6kT}{m}\left(1 - e^{-2t/\tau}\right). \quad (13.171)$$

The solution with both $\langle \boldsymbol{r}^2(0)\rangle = 0$ and $d\langle \boldsymbol{r}^2(0)\rangle/dt = 0$ is (exercise 13.21)

$$\langle \boldsymbol{r}^2(t)\rangle = \langle \boldsymbol{v}^2(0)\rangle\,\tau^2\left(1 - e^{-t/\tau}\right)^2 - \frac{3kT}{m}\,\tau^2\left(1 - e^{-t/\tau}\right)\left(3 - e^{-t/\tau}\right) + \frac{6kT\tau}{m}\,t. \quad (13.172)$$

13.12 Characteristic and Moment-Generating Functions

The Fourier transform (3.9) of a probability distribution $P(x)$ is its **characteristic function** $\tilde{P}(k)$ sometimes written as $\chi(k)$

$$\tilde{P}(k) \equiv \chi(k) \equiv E[e^{ikx}] = \int e^{ikx}\,P(x)\,dx. \quad (13.173)$$

The probability distribution $P(x)$ is the inverse Fourier transform (3.9)

$$P(x) = \int e^{-ikx}\,\tilde{P}(k)\,\frac{dk}{2\pi}. \quad (13.174)$$

Example 13.10 (Gauss) The characteristic function of the gaussian

$$P_G(x,\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}}\,\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \quad (13.175)$$

is by (3.18)

$$\begin{aligned}\tilde{P}_G(k,\mu,\sigma) &= \frac{1}{\sigma\sqrt{2\pi}}\int \exp\left(ikx - \frac{(x-\mu)^2}{2\sigma^2}\right)\,dx \quad (13.176)\\ &= \frac{e^{ik\mu}}{\sigma\sqrt{2\pi}}\int \exp\left(ikx - \frac{x^2}{2\sigma^2}\right)\,dx = \exp\left(i\mu k - \frac{1}{2}\sigma^2k^2\right).\end{aligned}$$

□

For a discrete probability distribution P_n the characteristic function is

$$\chi(k) \equiv E[e^{ikx}] = \sum_n e^{ikx_n}\,P_n. \quad (13.177)$$

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

Example 13.11 (Poisson) The Poisson distribution (13.58)

$$P_P(n, \langle n \rangle) = \frac{\langle n \rangle^n}{n!} e^{-\langle n \rangle} \tag{13.178}$$

has the characteristic function

$$\chi(k) = \sum_{n=0}^{\infty} e^{ikn} \frac{\langle n \rangle^n}{n!} e^{-\langle n \rangle} = e^{-\langle n \rangle} \sum_{n=0}^{\infty} \frac{(\langle n \rangle e^{ik})^n}{n!} = \exp\left[\langle n \rangle \left(e^{ik} - 1\right)\right]. \tag{13.179}$$

□

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

$$M(k) \equiv E[e^{kx}] = \tilde{P}(-ik) = \chi(-ik). \tag{13.180}$$

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

$$M(k) = E[e^{kx}] = \int e^{kx} P(x)\, dx \tag{13.181}$$

and for a discrete probability distribution P_n, it is

$$M(k) = E[e^{kx}] = \sum_n e^{kx_n} P_n. \tag{13.182}$$

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

$$E[x^n] = \mu_n = \left.\frac{d^n M(k)}{dk^n}\right|_{k=0} = (-i)^n \left.\frac{d^n \tilde{P}(k)}{dk^n}\right|_{k=0}. \tag{13.183}$$

Example 13.12 (Gauss and Poisson) The moment-generating functions for the distributions of Gauss (13.175) and Poisson (13.178) are

$$M_G(k, \mu, \sigma) = \exp\left(\mu k + \frac{1}{2}\sigma^2 k^2\right) \quad \text{and} \quad M_P(k, \langle n \rangle) = \exp\left[\langle n \rangle \left(e^k - 1\right)\right]. \tag{13.184}$$

They give as the first three moments of these distributions

$$\mu_{G0} = 1, \quad \mu_{G1} = \mu, \quad \mu_{G2} = \mu^2 + \sigma^2 \tag{13.185}$$

$$\mu_{P0} = 1, \quad \mu_{P1} = \langle n \rangle, \quad \mu_{P2} = \langle n \rangle + \langle n \rangle^2 \tag{13.186}$$

(exercise 13.22). □

Since the characteristic and moment-generating functions have derivatives (13.183) proportional to the moments μ_n, their Taylor series are

$$\tilde{P}(k) = E[e^{ikx}] = \sum_{n=0}^{\infty} \frac{(ik)^n}{n!} E[x^n] = \sum_{n=0}^{\infty} \frac{(ik)^n}{n!} \mu_n \tag{13.187}$$

and

$$M(k) = E[e^{kx}] = \sum_{n=0}^{\infty} \frac{k^n}{n!} E[x^n] = \sum_{n=0}^{\infty} \frac{k^n}{n!} \mu_n. \tag{13.188}$$

The **cumulants** c_n of a probability distribution are the derivatives of the logarithm of its moment-generating function

$$c_n = \left. \frac{d^n \ln M(k)}{dk^n} \right|_{k=0} = (-i)^n \left. \frac{d^n \ln \tilde{P}(k)}{dk^n} \right|_{k=0}. \tag{13.189}$$

One may show (exercise 13.24) that the first five cumulants of an arbitrary probability distribution are

$$c_0 = 0, \quad c_1 = \mu, \quad c_2 = \sigma^2, \quad c_3 = \nu_3, \quad \text{and} \quad c_4 = \nu_4 - 3\sigma^4 \tag{13.190}$$

where the ν's are its central moments (13.27). The 3d and 4th **normalized cumulants** are the **skewness** $\zeta = c_3/\sigma^3 = \nu_3/\sigma^3$ and the **kurtosis** $\kappa = c_4/\sigma^4 = \nu_4/\sigma^4 - 3$.

Example 13.13 (Gaussian Cumulants) The logarithm of the moment-generating function (13.184) of Gauss's distribution is $\mu k + \sigma^2 k^2/2$. Thus by (13.189), $P_G(x, \mu, \sigma)$ has no skewness or kurtosis, its cumulants vanish $c_{Gn} = 0$ for $n > 2$, and its fourth central moment is $\nu_4 = 3\sigma^4$. □

13.13 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

$$P_S(x, \nu, a) = \frac{1}{\sqrt{\pi}} \frac{\Gamma((1+\nu)/2)}{\Gamma(\nu/2)} \frac{a^\nu}{(a^2 + x^2)^{(1+\nu)/2}} \tag{13.191}$$

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

$$\mu_{2n} = (2n-1)!! \, \frac{\Gamma(\nu/2 - n)}{\Gamma(\nu/2)} \left(\frac{a^2}{2}\right)^n \tag{13.192}$$

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

$$P_S(x, 1, a) = \frac{1}{\pi} \frac{a}{a^2 + x^2} \tag{13.193}$$

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, p.15–16)

$$\Pr(x, \infty) = \int_x^\infty P_S(x', 3, 1) \, dx' = \frac{1}{2} - \frac{1}{\pi} \left[\arctan x + \frac{x}{1 + x^2}\right] \tag{13.194}$$

$$\Pr(x, \infty) = \int_x^\infty P_S(x', 4, \sqrt{2}) \, dx' = \frac{1}{2} - \frac{3}{4}u + \frac{1}{4}u^3 \tag{13.195}$$

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 $(0, \infty)$

$$P_{\ln}(x) = \frac{1}{\sigma x \sqrt{2\pi}} \exp\left[-\frac{\ln^2(x/x_0)}{2\sigma^2}\right] \tag{13.196}$$

describes distributions of rates of return (Bouchaud and Potters, 2003, p. 9). Its moments are (exercise 13.27)

$$\mu_n = x_0^n \, e^{n^2\sigma^2/2}. \tag{13.197}$$

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

$$P_e(x) = \alpha e^{-\alpha x} \tag{13.198}$$

has (exercise 13.28) 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)

$$P_{n,e}(x) = \alpha \frac{(\alpha x)^{n-1}}{(n-1)!} e^{-\alpha x}. \tag{13.199}$$

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 $(0, \infty)$

$$P_{n,P}(x,\sigma)dx = \frac{\sqrt{2}}{\sigma}\frac{1}{\Gamma(n/2)}\left(\frac{x}{\sigma\sqrt{2}}\right)^{n-1} e^{-x^2/(2\sigma^2)}dx \tag{13.200}$$

which for $x = v$, $n = 3$, and $\sigma^2 = kT/m$ is (exercise 13.29) the Maxwell-Boltzmann distribution (13.100). In terms of $\chi = x/\sigma$, it is

$$P_{n,P}(\chi^2/2)\,d\chi^2 = \frac{1}{\Gamma(n/2)}\left(\frac{\chi^2}{2}\right)^{n/2-1} e^{-\chi^2/2} d\left(\chi^2/2\right). \tag{13.201}$$

It has mean and variance

$$\mu = n \quad \text{and} \quad \sigma^2 = 2n \tag{13.202}$$

and is used in the chi-squared test (Pearson, 1900).

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 $\nu \leq 2$

$$\tilde{L}_\nu(k, a_\nu) = \exp\left(-a_\nu |k|^\nu\right). \tag{13.203}$$

Its inverse Fourier transform (13.174) is for $\nu = 1$ (exercise 13.30) the **Cauchy** or **Lorentz** distribution

$$L_1(x, a_1) = \frac{a_1}{\pi(x^2 + a_1^2)} \tag{13.204}$$

and for $\nu = 2$ the gaussian

$$L_2(x, a_2) = P_G(x, 0, \sqrt{2a_2}) = \frac{1}{2\sqrt{\pi a_2}} \exp\left(-\frac{x^2}{4a_2}\right) \tag{13.205}$$

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).

13.14 The Central Limit Theorem and Jarl Lindeberg

We have seen in sections (13.7 & 13.8) that unbiased fluctuations tend to distribute the position and velocity of molecules according to Gauss's distribution (13.75). 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

$$y = \frac{1}{N}\,(x_1 + x_2 + \cdots + x_N) \tag{13.206}$$

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 (13.34), the mean value of the average y is the average of the N means

$$\begin{aligned}\mu_y = E[y] = E[(x_1 + \cdots + x_N)\,/N] &= \frac{1}{N}\,(E[x_1] + \cdots + E[x_N]) \\ &= \frac{1}{N}\,(\mu_1 + \cdots + \mu_N)\,. \end{aligned} \tag{13.207}$$

Similarly, our rule (13.41) for the variance of a linear combination of *independent* variables tells us that the variance of the average y is

$$\sigma_y^2 = V[(x_1 + \cdots + x_N)\,/N] = \frac{1}{N^2}\left(\sigma_1^2 + \cdots + \sigma_N^2\right). \tag{13.208}$$

The independence of the random variables $x_1, x_2, \ldots, x_N$ implies (13.36) that their joint probability distribution factorizes

$$P(x_1, \ldots, x_N) = P_1(x_1)P_2(x_2)\cdots P_N(x_N). \tag{13.209}$$

We can use a delta function (3.36) to write the probability distribution $P^{(N)}(y)$ for the average $y = (x_1 + x_2 + \cdots + x_N)/N$ of the x_j's as

$$P^{(N)}(y) = \int P(x_1, \ldots, x_N)\,\delta((x_1 + x_2 + \cdots + x_N)/N - y)\,d^N x \tag{13.210}$$

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

$$\begin{aligned}\tilde{P}^{(N)}(k) &= \int e^{iky}\,P^{(N)}(y)\,dy \\ &= \int e^{iky} \int P(x_1, \ldots, x_N)\,\delta((x_1 + x_2 + \cdots + x_N)/N - y)\,d^N x\,dy \\ &= \int \exp\left[\frac{ik}{N}(x_1 + x_2 + \cdots + x_N)\right] P(x_1, \ldots, x_N)\,d^N x \\ &= \int \exp\left[\frac{ik}{N}(x_1 + x_2 + \cdots + x_N)\right] P_1(x_1)P_2(x_2)\cdots P_N(x_N)\,d^N x\end{aligned} \tag{13.211}$$

is then the product

$$\tilde{P}^{(N)}(k) = \tilde{P}_1(k/N)\,\tilde{P}_2(k/N)\cdots\tilde{P}_N(k/N) \tag{13.212}$$

of the characteristic functions

$$\tilde{P}_j(k/N) = \int e^{ikx_j/N}\, P_j(x_j)\, dx_j \tag{13.213}$$

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

The Taylor series (13.187) for each characteristic function is

$$\tilde{P}_j(k/N) = \sum_{n=0}^{\infty} \frac{(ik)^n}{n!\, N^n}\, \mu_{nj} \tag{13.214}$$

and so for big N we can use the approximation

$$\tilde{P}_j(k/N) \approx 1 + \frac{ik}{N}\, \mu_j - \frac{k^2}{2N^2}\, \mu_{2j} \tag{13.215}$$

in which $\mu_{2j} = \sigma_j^2 + \mu_j^2$ by the formula (13.22) for the variance. So we have

$$\tilde{P}_j(k/N) \approx 1 + \frac{ik}{N}\, \mu_j - \frac{k^2}{2N^2}\, (\sigma_j^2 + \mu_j^2) \tag{13.216}$$

or for large N

$$\tilde{P}_j(k/N) \approx \exp\left(\frac{ik}{N}\mu_j - \frac{k^2}{2N^2}\sigma_j^2\right). \tag{13.217}$$

Thus as $N \to \infty$, the characteristic function (13.212) for the variable y converges to

$$\begin{aligned} \tilde{P}^{(N)}(k) = \prod_{j=1}^{N} \tilde{P}_j(k/N) \;&= \prod_{j=1}^{N} \exp\left(\frac{ik}{N}\mu_j - \frac{k^2}{2N^2}\sigma_j^2\right) \\ &= \exp\left[\sum_{j=1}^{N}\left(\frac{ik}{N}\mu_j - \frac{k^2}{2N^2}\sigma_j^2\right)\right] = \exp\left(i\mu_y k - \frac{1}{2}\sigma_y^2 k^2\right) \end{aligned} \tag{13.218}$$

which is the characteristic function (13.176) of a gaussian (13.175) with mean and variance

$$\mu_y = \frac{1}{N}\sum_{j=1}^{N} \mu_j \quad \text{and} \quad \sigma_y^2 = \frac{1}{N^2}\sum_{j=1}^{N} \sigma_j^2. \tag{13.219}$$

The inverse Fourier transform (13.174) now gives the probability distribution $P^{(N)}(y)$ for the average $y = (x_1 + x_2 + \cdots + x_N)/N$ as

$$P^{(N)}(y) = \int_{-\infty}^{\infty} e^{-iky}\, \tilde{P}^{(N)}(k)\, \frac{dk}{2\pi} \tag{13.220}$$

which in view of (13.218) and (13.176) tends as $N \to \infty$ to Gauss's distribution $P_G(y, \mu_y, \sigma_y)$

$$\begin{aligned}\lim_{N\to\infty} P^{(N)}(y) &= \int_{-\infty}^{\infty} e^{-iky} \lim_{N\to\infty} \tilde{P}^{(N)}(k)\, \frac{dk}{2\pi} \\ &= \int_{-\infty}^{\infty} e^{-iky} \exp\left(i\mu_y k - \frac{1}{2}\sigma_y^2 k^2\right) \frac{dk}{2\pi} \\ &= P_G(y, \mu_y, \sigma_y) = \frac{1}{\sigma_y\sqrt{2\pi}} \exp\left[-\frac{(y-\mu_y)^2}{2\sigma_y^2}\right]\end{aligned} \tag{13.221}$$

with mean μ_y and variance σ_y^2 as given by (13.219). The sense in which $P^{(N)}(y)$ 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 $P^{(N)}(y)$ converges as $N \to \infty$ to the probability that y lies between a and b as determined by the gaussian $P_G(y, \mu_y, \sigma_y)$

$$\lim_{N\to\infty} \Pr_N(a < y < b) = \lim_{N\to\infty} \int_a^b P^{(N)}(y)\, dy = \int_a^b P_G(y, \mu_y, \sigma_y)\, dy. \tag{13.222}$$

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 (13.219) imply that $\mu_y = \mu$ and $\sigma_y^2 = \sigma^2/N$. In this case, one may show (exercise 13.32) that in terms of the variable

$$u \equiv \frac{\sqrt{N}(y-\mu)}{\sigma} = \frac{\left(\sum_{n=1}^{N} x_j\right) - N\mu}{\sqrt{N}\,\sigma} \tag{13.223}$$

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

$$\lim_{N\to\infty} P^{(N)}(y)\, dy = \frac{1}{\sqrt{2\pi}}\, e^{-u^2/2}\, du. \tag{13.224}$$

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

$$S_N \equiv \sum_{j=1}^{N} \sigma_j^2 = \sum_{j=1}^{N} \int_{-\infty}^{\infty} (x_j - \mu_j)^2\, P_j(x_j)\, dx_j \tag{13.225}$$

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

$$S_N(\delta) \equiv \sum_{j=1}^{N} \int_{\mu_j-\delta}^{\mu_j+\delta} (x_j - \mu_j)^2\, P_j(x_j)\, dx_j. \tag{13.226}$$

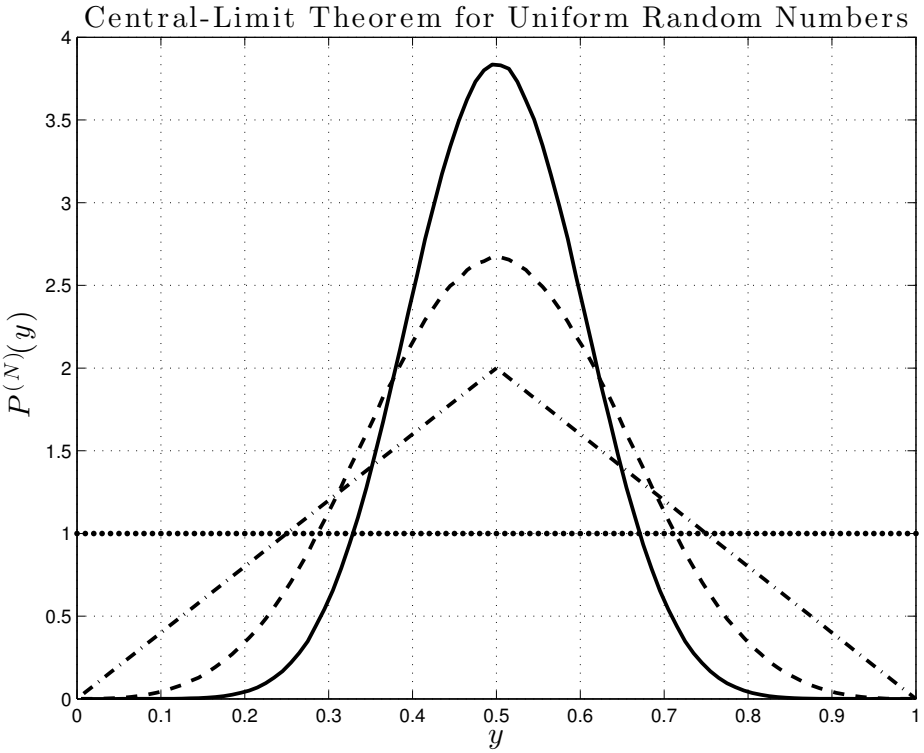

Figure 13.4 The probability distributions $P^{(N)}(y)$ (Eq. 13.210) for the mean $y = (x_1 + \cdots + x_N)/N$ of N random variables drawn from the uniform distribution are plotted for $N = 1$ (dots), 2 (dot dash), 4 (dashes), and 8 (solid). The distributions $P^{(N)}(y)$ rapidly approach gaussians with the same mean $\mu_y = 1/2$ but with shrinking variances $\sigma^2 = 1/12N$.

In terms of these definitions, Jarl Lindeberg (1876–1932) showed that $P^{(N)}(y)$ converges (in probability) to the gaussian (13.221) as long as the part $S_N(\delta)$ is most of S_N in the sense that for every $\epsilon > 0$

$$\lim_{N\to\infty} \frac{S_N\left(\epsilon\sqrt{S_N}\right)}{S_N} = 1. \tag{13.227}$$

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 (13.214) 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 13.14 (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

$$P^{(2)}(y) = \int_0^1 dx_1 \int_0^1 dx_2 \; \delta((x_1 + x_2)/2 - y). \tag{13.228}$$

Letting $u_1 = x_1/2$, we find

$$P^{(2)}(y) = 4 \int_{\max(0, y-\frac{1}{2})}^{\min(y, \frac{1}{2})} \theta(\tfrac{1}{2} + u_1 - y)\, du_1 = 4y\, \theta(\tfrac{1}{2} - y) + 4(1 - y)\, \theta(y - \tfrac{1}{2}) \tag{13.229}$$

which is the dot-dashed triangle in Fig. 13.4. 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.

13.15 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 **pseudo-random 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 easiest way to generate pseudo-random numbers is to use a random-number algorithm that is part of one's favorite FORTRAN, C, or C^{++} compiler. To run it, one first gives it a random starting point called a **seed**, which is a number or a vector. For instance, to start the GNU or Intel FORTRAN90 compiler, one includes in the code the line

```
call random_seed()
```

before using the line

```
call random_number(x)
```

to generate a random number x uniformly distributed on the interval $0 < x < 1$, or an array of such random numbers.

Some applications require random numbers of very high quality. For such applications, one might use Lüscher's RANLUX (Lüscher, 1994; James, 1994).

Most random-number generators are periodic with very long periods. The **Mersenne Twister** (Saito and Matsumoto, 2007) has the exceptionally long period $2^{19937} - 1 \gtrsim 4.3 \times 10^{6001}$. Matlab uses it.

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

$$F(x) = \int_a^x P(x')\, dx' \tag{13.230}$$

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 $F^{-1}(y)$

$$x = F^{-1}(y) \quad \text{if and only if} \quad y = F(x) \tag{13.231}$$

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 x is $F(x)$. The trick (Knuth, 1981, p. 116) is to set

$$r = F^{-1}(u) \tag{13.232}$$

so that $\Pr(r < x) = \Pr(F^{-1}(u) < x)$. For by (13.231) $F^{-1}(u) < x$ if and only if $u < F(x)$. So $\Pr(r < x) = \Pr(F^{-1}(u) < x) = \Pr(u < F(x)) = F(x)$. The trick works.

Example 13.15 ($P(r) = 3r^2$) 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

$$F(x) = \int_0^x P(x')\, dx' = \int_0^x 3x'^2\, dx' = x^3. \tag{13.233}$$

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

13.16 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 (13.15) the variable $r = u^{1/3}$ is distributed as $P_j(x) = 3x^2$.

The central limit theorem tells us that the distribution

$$P^{(N)}(y) = \int 3x_1^2\, 3x_2^2 \ldots 3x_N^2\, \delta((x_1 + x_2 + \cdots + x_N)/N - y)\, d^N x \quad (13.234)$$

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

$$\lim_{N\to\infty} P^{(N)}(y) = \frac{1}{\sigma_y\sqrt{2\pi}} \exp\left(-\frac{(x-\mu_y)^2}{2\sigma_y^2}\right) \quad (13.235)$$

with mean μ_y and variance σ_y^2 given by (13.219). Since the P_j's are all the same, they all have the same mean

$$\mu_y = \mu_j = \int_0^1 3x^3 dx = \frac{3}{4} \quad (13.236)$$

and the same variance

$$\sigma_j^2 = \int_0^1 3x^4 dx - \left(\frac{3}{4}\right)^2 = \frac{3}{5} - \frac{9}{16} = \frac{3}{80}. \quad (13.237)$$

By(13.219), 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

$$P^{(1)}(y) = \int 3x_1^2\, \delta(x_1 - y)\, dx_1 = 3y^2 \quad (13.238)$$

which is the probability distribution we started with. In Fig. 13.5, this is the quadratic, dotted curve.

For $N = 2$, the probability distribution $P^{(1)}(y)$ is (exercise 13.31)

$$\begin{aligned} P^{(2)}(y) &= \int 3x_1^2\, 3x_2^2\, \delta((x_1 + x_2)/2 - y)\, dx_1\, dx_2 \quad (13.239)\\ &= \theta(\tfrac{1}{2} - y)\, \frac{96}{5}\, y^5 + \theta(y - \tfrac{1}{2}) \left(\frac{36}{5} - \frac{96}{5}\, y^5 + 48y^2 - 36y\right). \end{aligned}$$

You can get the probability distributions $P^{(N)}(y)$ for $N = 2^j$ by running the FORTAN95 program

```
program clt
  implicit none ! avoids typos
  character(len=1)::ch_i1
```

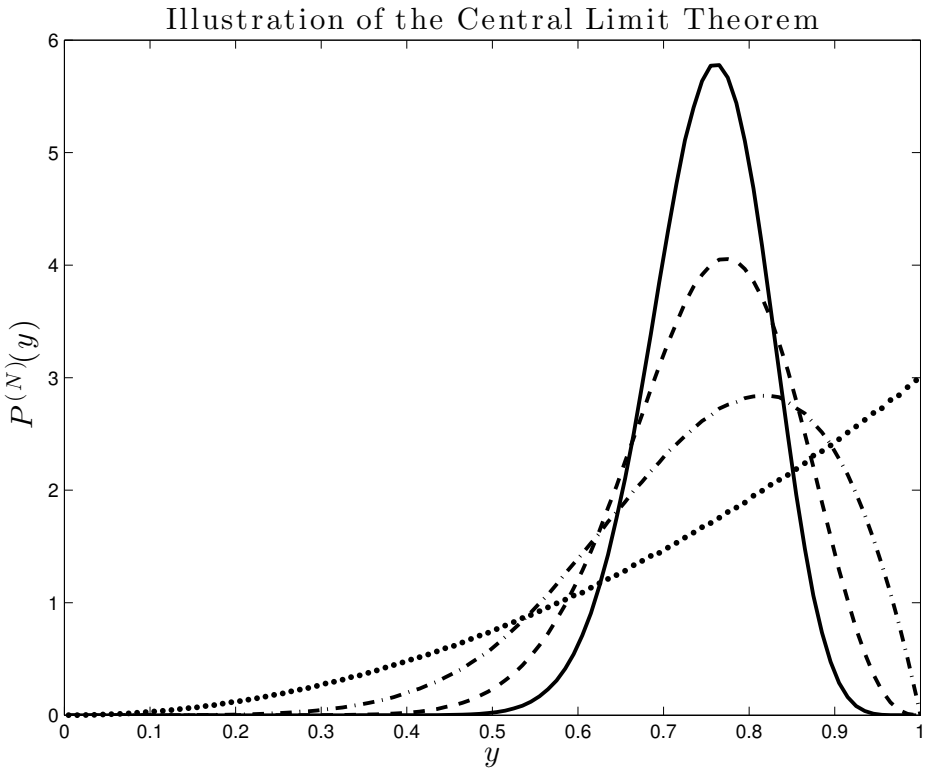

Figure 13.5 The probability distributions $P^{(N)}(y)$ (Eq. 13.234) for the mean $y = (x_1 + \cdots + x_N)/N$ of N random variables drawn from the quadratic distribution $P(x) = 3x^2$ are plotted for $N = 1$ (dots), 2 (dot dash), 4 (dashes), and 8 (solid). The four distributions $P^{(N)}(y)$ rapidly approach gaussians with the same mean $\mu_y = 3/4$ but with shrinking variances $\sigma_y^2 = 3/80N$.

```
integer,parameter::dp = kind(1.d0) !define double precision
integer::j,k,n,m
integer,dimension(100)::plot = 0
real(dp)::y
real(dp),dimension(100)::rplot
real(dp),allocatable,dimension(:)::r,u
real(dp),parameter::onethird = 1.d0/3.d0
write(6,*)'What is j?'; read(5,*) j
allocate(u(2**j),r(2**j))
call init_random_seed() ! set new seed, see below
do k = 1, 10000000  ! Make the N = 2**j plot
```

```
     call random_number(u)
     r = u**onethird
     y = sum(r)/2**j
     n = 100*y + 1
     plot(n) = plot(n) + 1
  end do
  rplot = 100*real(plot)/sum(plot)
  write(ch_i1,"(i1)") j ! turns integer j into character ch_i1
  open(7,file='plot'//ch_i1) ! opens and names files
  do m = 1, 100
     write(7,*) 0.01d0*(m-0.5), rplot(m)
  end do
end program clt
subroutine init_random_seed()
  implicit none
  integer i, n, clock
  integer, dimension(:), allocatable :: seed
  call random_seed(size = n) ! find size of seed
  allocate(seed(n))
  call system_clock(count=clock) ! get time of processor clock
  seed = clock + 37 * (/ (i-1, i=1, n) /) ! make seed
  call random_seed(put=seed) ! set seed
  deallocate(seed)
end subroutine init_random_seed
```

The distributions $P^{(N)}(y)$ for $N = 1$, 2, 4, and 8 are plotted in Fig. 13.5. $P^{(1)}(y) = 3y^2$ 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.

Although FORTRAN95 is an ideal language for computation, C++ is more versatile, more modular, and more suited to large projects involving many programmers. An equivalent C++ code written by Sean Cahill is:

```
#include <stdlib.h>
#include <time.h>
#include <math.h>
#include <string>
#include <iostream>
#include <fstream>
#include <sstream>
#include <iomanip>
```

```
#include <valarray>

using namespace std;

// Fills the array val with random numbers between 0 and 1
void rand01(valarray<double>& val)
{
  // Records the size
  unsigned int size = val.size();

  // Loops through the size
  unsigned int i=0;
  for (i=0; i<size; i++)
  {
    // Generates a random number between 0 and 1
    val[i] = static_cast<double>(rand()) / RAND_MAX;
  }
}

void clt ()
{
  // Declares local constants
  const int PLOT_SIZE      = 100;
  const int LOOP_CALC_ITR  = 10000000;
  const double ONE_THIRD   = 1.0 / 3.0;

  // Inits local variables
  double y=0;
  int i=0, j=0, n=0;

  // Gets the value of J
  cout << "What is J? ";
  cin >> j;

  // Bases the vec size on J
  const int VEC_SIZE = static_cast<int>(pow(2.0,j));

  // Inits vectors
  valarray<double> plot(PLOT_SIZE);
  valarray<double> rplot(PLOT_SIZE);
```

```
  valarray<double> r(VEC_SIZE);

  // Seeds random number generator
  srand ( time(NULL) );

  // Performs the calculations
  for (i=0; i<LOOP_CALC_ITR; i++)
  {
    rand01(r);
    r = pow(r, ONE_THIRD);
    y = r.sum() / VEC_SIZE;
    n = static_cast<int>(100 * y);
    plot[n]++;
  }

  // Normalizes RPLOT
  rplot = plot * (100.0 / plot.sum());

  // Opens a data file
  ostringstream fileName;
  fileName << "plot_" << j << ".txt";
  ofstream fileHandle;
  fileHandle.open (fileName.str().c_str());

  // Sets precision
  fileHandle.setf(ios::fixed,ios::floatfield);
  fileHandle.precision(7);

  // Writes the data to a file
  for (i=1; i<=PLOT_SIZE; i++)
   fileHandle << 0.01*(i-0.5) << "    " << rplot[i-1] << endl;

  // Closes the data file
  fileHandle.close();
}
```

13.17 Measurements, Estimators, and Friedrich Bessel

A probability distribution $P(x;\boldsymbol{\theta})$ for a stochastic variable x may depend upon one or more unknown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$ such as the mean μ and the variance σ^2.

Experimenters seek to determine the unknown parameters $\boldsymbol{\theta}$ by collecting data in the form of 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 N factors of the physical distribution $P(x;\boldsymbol{\theta})$

$$P(\boldsymbol{x};\boldsymbol{\theta}) = \prod_{j=1}^{N} P(x_j;\boldsymbol{\theta}). \tag{13.240}$$

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

$$E[u_\ell^{(N)}] = \int u_\ell^{(N)}(\boldsymbol{x})\, P(\boldsymbol{x};\boldsymbol{\theta})\, d^N x = \theta_\ell + b_\ell^{(N)} \tag{13.241}$$

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

Inasmuch as the mean (13.25) is the integral of the physical distribution

$$\mu = \int x\, P(x;\boldsymbol{\theta})\, dx \tag{13.242}$$

a natural estimator for the mean is

$$u_\mu^{(N)}(\boldsymbol{x}) = (x_1 + \cdots + x_N)/N. \tag{13.243}$$

Its expected value is

$$\begin{aligned} E[u_\mu^{(N)}] &= \int u_\mu^{(N)}(\boldsymbol{x})\, P(\boldsymbol{x};\boldsymbol{\theta})\, d^N x = \int \frac{x_1 + \cdots + x_N}{N}\, P(\boldsymbol{x};\boldsymbol{\theta})\, d^N x \qquad (13.244)\\ &= \frac{1}{N}\sum_{k=1}^{N} \int x_k\, P(x_k;\boldsymbol{\theta})\, dx_k \prod_{k\neq j=1}^{N} \int P(x_j;\boldsymbol{\theta})\, dx_j = \frac{1}{N}\sum_{k=1}^{N} \mu = \mu. \end{aligned}$$

Thus the natural estimator $u_\mu^{(N)}(\boldsymbol{x})$ of the mean (13.243) has $b_\ell^{(N)} = 0$, and so it is a consistent and unbiased estimator for the mean.

Since the variance (13.28) of the probability distribution $P(x;\boldsymbol{\theta})$ is the integral

$$\sigma^2 = \int (x-\mu)^2\, P(x;\boldsymbol{\theta})\, dx \tag{13.245}$$

the variance of the estimator u_μ^N is

$$V[u_\mu^N] = \int \left(u_\mu^{(N)}(\boldsymbol{x}) - \mu\right)^2 P(\boldsymbol{x};\boldsymbol{\theta})\, d^N x = \int \left[\frac{1}{N}\sum_{j=1}^{N}(x_j - \mu)\right]^2 P(\boldsymbol{x};\boldsymbol{\theta})\, d^N x$$

$$= \frac{1}{N^2}\sum_{j,k=1}^{N}\int (x_j - \mu)\,(x_k - \mu)\, P(\boldsymbol{x};\boldsymbol{\theta})\, d^N x \tag{13.246}$$

$$= \frac{1}{N^2}\sum_{j,k=1}^{N}\delta_{jk}\int (x_j - \mu)^2\, P(\boldsymbol{x};\boldsymbol{\theta})\, d^N x = \frac{1}{N^2}\sum_{j,k=1}^{N}\delta_{jk}\,\sigma^2 = \frac{\sigma^2}{N}$$

in which σ^2 is the variance (13.245) 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

$$u_{\sigma^2}^{(N)}(x) = B\sum_{j=1}^{N}\left(x_j - u_\mu^{(N)}(\boldsymbol{x})\right)^2 \tag{13.247}$$

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 $E[u_{\sigma^2}^{(N)}]$ equal to σ^2

$$E[u_{\sigma^2}^{(N)}] = \int B\sum_{j=1}^{N}\left(x_j - u_\mu^{(N)}(\boldsymbol{x})\right)^2 P(\boldsymbol{x};\boldsymbol{\theta})\, d^N x = \sigma^2 \tag{13.248}$$

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

$$\frac{\sigma^2}{B} = \sum_{j=1}^{N}\int \left[x_j - \mu - \left(u_\mu^{(N)}(\boldsymbol{x}) - \mu\right)\right]^2 P(\boldsymbol{x};\boldsymbol{\theta})\, d^N x = S_{jj} + S_{j\mu} + S_{\mu\mu} \tag{13.249}$$

the first of which is

$$S_{jj} = \sum_{j=1}^{N}\int (x_j - \mu)^2\, P(\boldsymbol{x};\boldsymbol{\theta})\, d^N x = N\sigma^2. \tag{13.250}$$

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

$$S_{j\mu} = -2\sum_{j=1}^{N}\int (x_j-\mu)\left(u_\mu^{(N)}(\boldsymbol{x})-\mu\right)P(\boldsymbol{x};\boldsymbol{\theta})\,d^Nx \tag{13.251}$$

$$= -\frac{2}{N}\sum_{j=1}^{N}\int (x_j-\mu)\sum_{k=1}^{N}(x_k-\mu)\,P(\boldsymbol{x};\boldsymbol{\theta})\,d^Nx = -2\sigma^2.$$

The third term is the related to the variance (13.246)

$$S_{\mu\mu} = \sum_{j=1}^{N}\int \left(u_\mu^{(N)}(\boldsymbol{x})-\mu\right)^2 P(\boldsymbol{x};\boldsymbol{\theta})\,d^Nx = NV[u_\mu^N] = \sigma^2. \tag{13.252}$$

Thus the factor B must satisfy

$$\sigma^2/B = N\sigma^2 - 2\sigma^2 + \sigma^2 = (N-1)\sigma^2 \tag{13.253}$$

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

$$u_{\sigma^2}^{(N)}(\boldsymbol{x}) = \frac{1}{N-1}\sum_{j=1}^{N}\left(x_j - u_\mu^{(N)}(\boldsymbol{x})\right)^2 = \frac{1}{N-1}\sum_{j=1}^{N}\left(x_j - \frac{1}{N}\sum_{k=1}^{N}x_k\right)^2. \tag{13.254}$$

It is consistent and unbiased since $E[u_{\sigma^2}^{(N)}] = \sigma^2$ by construction (13.248). 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 $\mu_N = (x_1+\cdots+x_N)/N$. And we can use Bessel's formula (13.254) to estimate the variance σ_N^2 of the unknown distribution $P(x;\boldsymbol{\theta})$. Our formula (13.246) for the variance $\sigma^2(\mu_N)$ of the mean μ_N then gives

$$\sigma^2(\mu_N) = \frac{\sigma_N^2}{N} = \frac{1}{N(N-1)}\sum_{j=1}^{N}\left(x_j - \frac{1}{N}\sum_{k=1}^{N}x_k\right)^2. \tag{13.255}$$

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

$$\sigma(\mu_N) = \sqrt{\frac{\sigma_N^2}{N}} = \sqrt{\frac{1}{N(N-1)}\sum_{j=1}^{N}\left(x_j - \frac{1}{N}\sum_{k=1}^{N}x_k\right)^2}. \tag{13.256}$$

Few formulas have seen so much use.

13.18 Information and Ronald Fisher

The **Fisher information matrix** of a distribution $P(\boldsymbol{x};\boldsymbol{\theta})$ is the mean of products of its partial logarithmic derivatives

$$\begin{aligned} F_{k\ell}(\boldsymbol{\theta}) &\equiv E\left[\frac{\partial \ln P(\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta_k}\,\frac{\partial \ln P(\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta_\ell}\right] \\ &= \int \frac{\partial \ln P(\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta_k}\,\frac{\partial \ln P(\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta_\ell}\,P(\boldsymbol{x};\boldsymbol{\theta})\,d^N x \end{aligned} \tag{13.257}$$

(Ronald Fisher, 1890–1962). Fisher's matrix (exercise 13.33) is symmetric $F_{k\ell} = F_{\ell k}$ and nonnegative (1.38), and when it is positive (1.39), it has an inverse. By differentiating the normalization condition

$$\int P(\boldsymbol{x};\boldsymbol{\theta})\,d^N x = 1 \tag{13.258}$$

we have

$$0 = \int \frac{\partial P(\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta_k}\,d^N x = \int \frac{\partial \ln P(\boldsymbol{x};\boldsymbol{\theta})}{\partial\theta_k}\,P(\boldsymbol{x};\boldsymbol{\theta})\,d^N x \tag{13.259}$$

which says that the mean value of the logarithmic derivative of the probability distribution, a quantity called the **score**, vanishes. Using the notation $P_{,k} \equiv \partial P/\partial\theta_k$ and $(\ln P)_{,k} \equiv \partial \ln P/\partial\theta_k$ and differentiating again, one has (exercise 13.34)

$$\int (\ln P)_{,k}\,(\ln P)_{,\ell}\,P\,d^N x = -\int (\ln P)_{,k,\ell}\,P\,d^N x \tag{13.260}$$

so that another form of Fisher's information matrix is

$$F_{k\ell}(\boldsymbol{\theta}) = -E\left[(\ln P)_{,k,\ell}\right] = -\int (\ln P)_{,k,\ell}\,P\,d^N x. \tag{13.261}$$

Cramér and Rao used Fisher's information matrix to form a lower bound on the covariance (13.35) matrix $C[u_k, u_\ell]$ of any two estimators. To see how this works, we use the vanishing (13.259) of the mean of the score to write the covariance of the kth score $V_k \equiv (\ln P(\boldsymbol{x};\boldsymbol{\theta}))_{,k}$ with the ℓth estimator $u_\ell(\boldsymbol{x})$ as a derivative $\langle u_\ell\rangle_{,k}$ of the mean $\langle u_\ell\rangle$

$$\begin{aligned} C[V_k, u_\ell] &= \int (\ln P)_{,k}\,(u_\ell - b_\ell - \theta_\ell)\,P\,d^N x = \int (\ln P)_{,k}\,u_\ell\,P\,d^N x \\ &= \int P_{,k}(\boldsymbol{x};\boldsymbol{\theta})\,u_\ell(\boldsymbol{x})\,d^N x = \langle u_\ell\rangle_{,k}. \end{aligned} \tag{13.262}$$

Thus for any two sets of constants y_k and w_ℓ, we have with $P = \sqrt{P}\,\sqrt{P}$

$$\sum_{\ell,k=1}^{m} y_k\,\partial_k\langle u_\ell\rangle\,w_\ell = \int \sum_{\ell,k=1}^{m} y_k\,(\ln P)_{,k}\sqrt{P}\;(u_\ell - b_\ell - \theta_\ell)\,w_\ell\,\sqrt{P}\,d^N x. \quad (13.263)$$

We can suppress some indices by grouping the y_j's, the w_j's, and so forth into the vectors $Y^\mathsf{T} = (y_1, \ldots, y_m)$, $W^\mathsf{T} = (w_1, \ldots, w_m)$, $U^\mathsf{T} = (u_1, \ldots, u_m)$, $B^\mathsf{T} = (b_1, \ldots, b_m)$, and $\Theta^\mathsf{T} = (\theta_1, \ldots, \theta_m)$, and by grouping the $\partial_k\langle u_\ell\rangle$'s into a matrix $(\nabla\overline{U})_{kl}$ which by (13.241) is

$$(\nabla\overline{U})_{kl} \equiv \partial_k\langle u_\ell\rangle = \partial_k\,(\theta_\ell + b_\ell) = \delta_{kl} + \partial_k b_\ell. \quad (13.264)$$

In this compact notation, our relation (13.263) is

$$Y^\mathsf{T}\,\nabla\overline{U}\,W = \int Y^\mathsf{T}(\nabla \ln P)\,\sqrt{P}\,(U^\mathsf{T} - B^\mathsf{T} - \Theta^\mathsf{T})W\,\sqrt{P}\,d^N x. \quad (13.265)$$

Squaring, we apply Schwarz's inequality (6.379)

$$\begin{aligned}\left[Y^\mathsf{T}\,\nabla\overline{U}\,W\right]^2 &= \left[\int Y^\mathsf{T}(\nabla \ln P)\,\sqrt{P}\,(U^\mathsf{T} - B^\mathsf{T} - \Theta^\mathsf{T})W\,\sqrt{P}\,d^N x\right]^2 \quad (13.266)\\ &\le \int \left[Y^\mathsf{T}(\nabla \ln P)\,\sqrt{P}\right]^2 d^N x \int \left[(U^\mathsf{T} - B^\mathsf{T} - \Theta^\mathsf{T})W\,\sqrt{P}\right]^2 d^N x\\ &= \int \left[Y^\mathsf{T}\nabla \ln P\right]^2\,P\,d^N x \int \left[(U^\mathsf{T} - B^\mathsf{T} - \Theta^\mathsf{T})W\right]^2\,P\,d^N x.\end{aligned}$$

In the last line, we recognize the first integral as $Y^\mathsf{T}FY$, where F is Fisher's matrix (13.257), and the second as $W^\mathsf{T}CW$ in which C is the covariance of the estimators

$$C_{k\ell} \equiv C[U,U]_{k\ell} = C[u_k - b_k - \theta_k, u_\ell - b_\ell - \theta_\ell]. \quad (13.267)$$

So (13.266) says

$$\left(Y^\mathsf{T}\nabla\overline{U}W\right)^2 \le Y^\mathsf{T}FY\;\;W^\mathsf{T}CW. \quad (13.268)$$

Thus as long as the symmetric non-negative matrix F is positive and so has an inverse, we can set the arbitrary constant vector $Y = F^{-1}\nabla\overline{U}\,W$ and get

$$(W^\mathsf{T}\nabla\overline{U}^\mathsf{T}F^{-1}\nabla\overline{U}W)^2 \le W^\mathsf{T}\nabla\overline{U}^\mathsf{T}F^{-1}\nabla\overline{U}W\;\;W^\mathsf{T}CW. \quad (13.269)$$

Canceling a common factor, we obtain the **Cramér-Rao inequality**

$$W^\mathsf{T}C\,W \ge W^\mathsf{T}\nabla\overline{U}^\mathsf{T}F^{-1}\,\nabla\overline{U}\,W \quad (13.270)$$

often written as

$$C \ge \nabla\overline{U}^\mathsf{T}\,F^{-1}\,\nabla\overline{U}. \quad (13.271)$$

By (13.264), the matrix $\nabla\overline{U}$ is the identity matrix I plus the gradient of the bias B

$$\nabla\overline{U} = I + \nabla B. \tag{13.272}$$

Thus another form of the Cramér-Rao inequality is

$$C \geq (I + \nabla B)^{\mathsf{T}} F^{-1} (I + \nabla B) \tag{13.273}$$

or in terms of the arbitrary vector W

$$W^{\mathsf{T}} C W \geq W^{\mathsf{T}} (I + \nabla B)^{\mathsf{T}} F^{-1} (I + \nabla B) W. \tag{13.274}$$

Letting the arbitrary vector W be $W_j = \delta_{jk}$, one arrives at (exercise 13.35) the **Cramér-Rao lower bound on the variance** $V[u_k] = C[u_k, u_k]$

$$V[u_k] \geq (F^{-1})_{kk} + \sum_{\ell=1}^{m} 2(F^{-1})_{k\ell}\partial_l b_k + \sum_{\ell,j=1}^{m} (F^{-1})_{\ell j}\partial_l b_k \partial_j b_k. \tag{13.275}$$

If the estimator $u_k(\boldsymbol{x})$ is unbiased, then this lower bound simplifies to

$$V[u_k] \geq (F^{-1})_{kk}. \tag{13.276}$$

Example 13.16 (Cramér-Rao Bound for a Gaussian) The elements of Fisher's information matrix for the mean μ and variance σ^2 of Gauss's distribution for N data points $x_1, \ldots, x_N$

$$P_G^{(N)}(\boldsymbol{x}, \mu, \sigma) = \prod_{j=1}^{N} P_G(x_j; \mu, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \exp\left(-\sum_{j=1}^{N} \frac{(x_j - \mu)^2}{2\sigma^2}\right) \tag{13.277}$$

are

$$\begin{aligned}
F_{\mu\mu} &= \int \left[\left(\ln P_G^{(N)}(\boldsymbol{x}, \mu, \sigma)\right)_{,\mu}\right]^2 P_G^{(N)}(\boldsymbol{x}, \mu, \sigma)\, d^N x \\
&= \sum_{i,j=1}^{N} \int \left(\frac{x_i - \mu}{\sigma^2}\right)\left(\frac{x_j - \mu}{\sigma^2}\right) P_G^{(N)}(\boldsymbol{x}, \mu, \sigma)\, d^N x \\
&= \sum_{i=1}^{N} \int \left(\frac{x_i - \mu}{\sigma^2}\right)^2 P_G^{(N)}(\boldsymbol{x}, \mu, \sigma)\, d^N x = \frac{N}{\sigma^2} \\
F_{\mu\sigma^2} &= \int (\ln P_G^{(N)}(\boldsymbol{x}, \mu, \sigma))_{,\mu} (\ln P_G^{(N)}(\boldsymbol{x}, \mu, \sigma))_{,\sigma^2} P_G^{(N)}(\boldsymbol{x}, \mu, \sigma)\, d^N x \\
&= \sum_{i,j=1}^{N} \int \left[\frac{x_i - \mu}{\sigma^2}\right]\left[\frac{(x_j - \mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2}\right] P_G^{(N)}(\boldsymbol{x}, \mu, \sigma)\, d^N x = 0
\end{aligned} \tag{13.278}$$

$F_{\sigma^2\mu} = F_{\mu\sigma^2} = 0$, and

$$\begin{aligned} F_{\sigma^2\sigma^2} &= \int \left[(\ln P_G^{(N)}(\boldsymbol{x}, \mu, \sigma))_{,\sigma^2} \right]^2 P_G^{(N)}(\boldsymbol{x}, \mu, \sigma)\, d^N x \\ &= \sum_{i,j=1}^{N} \int \left[\frac{(x_i - \mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2} \right] \left[\frac{(x_j - \mu)^2}{2\sigma^4} - \frac{1}{2\sigma^2} \right] P_G^{(N)}(\boldsymbol{x}, \mu, \sigma)\, d^N x \\ &= \frac{N}{2\sigma^4}. \end{aligned} \tag{13.279}$$

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 (13.276), and so $V[u_\mu] \geq (F^{-1})_{\mu\mu} = \sigma^2/N$. The variance $V[u_\mu^{(N)}]$ of the natural estimator of the mean $u_\mu^{(N)}(\boldsymbol{x}) = (x_1 + \cdots + x_N)/N$ is σ^2/N by (13.246), and so it respects and saturates the lower bound (13.276)

$$V[u_\mu^{(N)}] = E[(u_\mu^{(N)} - \mu)^2] = \sigma^2/N = (F^{-1})_{\mu\mu}. \tag{13.280}$$

One may show (exercise 13.36) that the variance $V[u_{\sigma^2}^{(N)}]$ of Bessel's estimator (13.254) of the variance is (Riley et al., 2006, p. 1248)

$$V[u_{\sigma^2}^{(N)}] = \frac{1}{N} \left(\nu_4 - \frac{N-3}{N-1} \sigma^4 \right) \tag{13.281}$$

where ν_4 is the fourth central moment (13.26) of the probability distribution. For the gaussian $P_G(\boldsymbol{x}; \mu, \sigma)$ one may show (exercise 13.37) that this moment is $\nu_4 = 3\sigma^4$, and so for it

$$V_G[u_{\sigma^2}^{(N)}] = \frac{2}{N-1}\, \sigma^4. \tag{13.282}$$

Thus the variance of Bessel's estimator of the variance respects but does not saturate its Cramér-Rao lower bound (13.276, 13.279)

$$V_G[u_{\sigma^2}^{(N)}] = \frac{2}{N-1}\, \sigma^4 > \frac{2}{N}\, \sigma^4. \tag{13.283}$$

□

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

13.19 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

$$P(\boldsymbol{x}) = \prod_{j=1}^{N} P(x_j, t_j, \sigma) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^N \exp\left(-\sum_{j=1}^{N} \frac{(x_j - f(t_j; \boldsymbol{\alpha}))^2}{2\sigma^2}\right). \tag{13.284}$$

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

$$0 = \frac{\partial}{\partial \alpha_\ell} \sum_{j=1}^{N} (x_j - f(t_j; \boldsymbol{\alpha}))^2 = -2 \sum_{j=1}^{N} (x_j - f(t_j; \boldsymbol{\alpha})) \, \frac{\partial f(t_j; \boldsymbol{\alpha})}{\partial \alpha_\ell}. \tag{13.285}$$

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 $\boldsymbol{\alpha}$

$$f(t; \boldsymbol{\alpha}) = \sum_{k=1}^{M} g_k(t) \, \alpha_k \tag{13.286}$$

then the equations (13.285) that determine these parameters $\boldsymbol{\alpha}$ are linear

$$0 = \sum_{j=1}^{N} \left(x_j - \sum_{k=1}^{M} g_k(t_j) \, \alpha_k\right) g_\ell(t_j). \tag{13.287}$$

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

$$G^{\mathsf{T}} \boldsymbol{x} = G^{\mathsf{T}} G \boldsymbol{\alpha}. \tag{13.288}$$

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 (1.399)

$$G^{+} = \left(G^{\mathsf{T}} G\right)^{-1} G^{\mathsf{T}} \tag{13.289}$$

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

$$\boldsymbol{\alpha} = G^+ \boldsymbol{x}. \tag{13.290}$$

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

$$G G^+ = P \tag{13.291}$$

is an $N \times N$ projection operator (exercise 13.38) onto the $M \times M$ subspace for which $G^+G = I_M$ is the identity operator. Like all projection operators, P satisfies $P^2 = P$.

13.20 Karl Pearson's Chi-Squared Statistic

The argument of the exponential (13.284) in $P(\boldsymbol{x})$ is (the negative of) Karl Pearson's chi-squared statistic (Pearson, 1900)

$$\chi^2 \equiv \sum_{j=1}^{N} \frac{(x_j - f(t_j; \boldsymbol{\alpha}))^2}{2\sigma^2}. \tag{13.292}$$

When the function $f(t; \boldsymbol{\alpha})$ is linear (13.286) in $\boldsymbol{\alpha}$, the N-vector $f(t_j; \boldsymbol{\alpha})$ is $f = G\,\boldsymbol{\alpha}$. Pearson's χ^2 then is

$$\chi^2 = (\boldsymbol{x} - G\,\boldsymbol{\alpha})^2 / 2\sigma^2. \tag{13.293}$$

Now (13.290) 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

$$\boldsymbol{x} - G\,\boldsymbol{\alpha} = \boldsymbol{x} - G\,G^+\,\boldsymbol{x} = \left(I - G\,G^+\right)\boldsymbol{x} = (I - P)\,\boldsymbol{x}. \tag{13.294}$$

So χ^2 is proportional to the squared length

$$\chi^2 = \tilde{\boldsymbol{x}}^2 / 2\sigma^2 \tag{13.295}$$

of the vector

$$\tilde{\boldsymbol{x}} \equiv (I - P)\,\boldsymbol{x}. \tag{13.296}$$

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 13.17 (Two Position Measurements) Suppose we measure a position twice with error σ, get x_1 and x_2, and choose $G^{\mathsf{T}} = (1, 1)$. Then

the single parameter α is their average $\alpha = (x_1 + x_2)/2$, and χ^2 is

$$\begin{aligned}\chi^2 &= \Big\{[x_1 - (x_1+x_2)/2]^2 + [x_2 - (x_1+x_2)/2]^2\Big\}\Big/ 2\sigma^2 \\ &= \Big\{[(x_1 - x_2)/2]^2 + [(x_2 - x_1)/2]^2\Big\}\Big/ 2\sigma^2 \\ &= \Big[(x_1 - x_2)/\sqrt{2}\Big]^2 /2\sigma^2. \end{aligned} \tag{13.297}$$

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)

$$P|j\rangle = 0 \quad 1 \le j \le N - M \tag{13.298}$$

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

$$P|k\rangle = |k\rangle \quad N - M + 1 \le k \le N. \tag{13.299}$$

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

$$\boldsymbol{x} = \sum_{j=1}^{N-M} x_j|j\rangle + \sum_{k=N-M+1}^{N} x_k|k\rangle \tag{13.300}$$

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

$$\tilde{\boldsymbol{x}} = (I - P)\,\boldsymbol{x} = \sum_{j=1}^{N-M} x_j|j\rangle. \tag{13.301}$$

Example 13.18 (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 $\overline{x}$

$$\alpha = G^{+}\,\boldsymbol{x} = \left(G^{\mathsf{T}}\,G\right)^{-1} G^{\mathsf{T}}\,\boldsymbol{x} = \frac{1}{N}\sum_{j=1}^{N} x_j = \overline{x} \tag{13.302}$$

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 $G^{\mathsf{T}} = (1, 1, \ldots, 1)$

$$G^{\mathsf{T}}\tilde{\boldsymbol{x}} = \left(\sum_{j=1}^{N} x_j\right) - N\overline{x} = 0. \tag{13.303}$$

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

$$x = f(t;\boldsymbol{\alpha}) = \sum_{k=1}^{M} g_k(t)\,\alpha_k \tag{13.304}$$

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

$$\chi^2 = (\tilde{\boldsymbol{x}})^2 / 2\sigma^2. \tag{13.305}$$

How good is our fit?

A χ^2 distribution with $N-M$ **degrees of freedom** has by (13.202) mean

$$E[\chi^2] = N - M \tag{13.306}$$

and variance

$$V[\chi^2] = 2(N-M). \tag{13.307}$$

So our χ^2 should be about

$$\chi^2 \approx N - M \pm \sqrt{2(N-M)}. \tag{13.308}$$

If it lies within this range, then (13.304) 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 (13.201)

$$\Pr_n(\chi^2 > \chi_0^2) = \int_{\chi_0^2}^{\infty} P_n(\chi^2/2)\,d\chi^2 = \int_{\chi_0^2}^{\infty} \frac{1}{2\Gamma(n/2)} \left(\frac{\chi^2}{2}\right)^{n/2-1} e^{-\chi^2/2} d\chi^2 \tag{13.309}$$

in which $n = N - M$ is the number of data points minus the number of parameters, and $\Gamma(n/2)$ is the gamma function (5.102, 4.62). 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. 13.6 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.

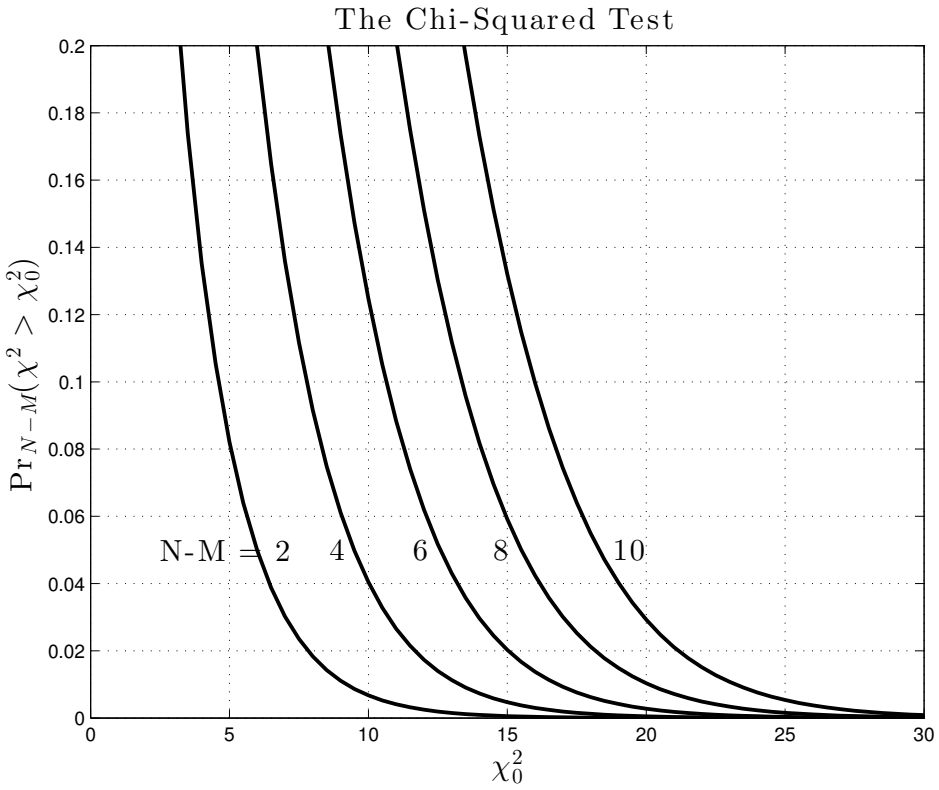

Figure 13.6 The probabilities $\Pr_{N-M}(\chi^2 > \chi_0^2)$ are plotted from left to right for $N - M = 2$, 4, 6, 8, and 10 degrees of freedom as functions of χ_0^2.

13.21 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

$$P_e^{(N)}(x) = \frac{1}{N}\sum_{j=1}^{N}\delta(x - x_j). \tag{13.310}$$

Our cumulative probability for events less than x then is

$$\Pr_e^{(N)}(-\infty, x) = \int_{-\infty}^{x} P_e^{(N)}(x')\,dx' = \int_{-\infty}^{x}\frac{1}{N}\sum_{j=1}^{N}\delta(x' - x_j)\,dx'. \tag{13.311}$$

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

$$\Pr_e^{(N)}(-\infty, x) = \frac{j}{N} \quad \text{for} \quad x_j < x < x_{j+1}. \tag{13.312}$$

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, theoretical cumulative probability distribution $\Pr_t(-\infty, x)$. One way to do this is to compute the distance D_N between the two cumulative probability distributions

$$D_N = \sup_{-\infty<x<\infty} \left| \Pr_e^{(N)}(-\infty, x) - \Pr_t(-\infty, x) \right| \tag{13.313}$$

in which **sup** stands for ***supremum*** and means **least upper bound**. Since cumulative probabilities lie between zero and one, it follows (exercise 13.39) that the Kolmogorov distance is bounded by

$$0 \leq D_N \leq 1. \tag{13.314}$$

The simpler Smirnov distances

$$\begin{aligned} D_N^+ &= \sup_{-\infty<x<\infty} \left(\Pr_e^{(N)}(-\infty, x) - \Pr_t(-\infty, x) \right) \\ D_N^- &= \sup_{-\infty<x<\infty} \left(\Pr_t(-\infty, x) - \Pr_e^{(N)}(-\infty, x) \right) \end{aligned} \tag{13.315}$$

provide (exercise 13.40) an expression for D_N as the greater of the two

$$D_N = \max(D_N^+, D_N^-). \tag{13.316}$$

Using our explicit expression (13.312) for the empirical cumulative probability $\Pr_e^{(N)}(-\infty, x)$ and the monotonicity (13.30) of cumulative probabilities such as $\Pr_t(-\infty, x)$, one may show (exercise 13.41) that the Smirnov distances are given by

$$\begin{aligned} D_N^+ &= \sup_{1\leq j\leq N} \left(\frac{j}{N} - \Pr_t(-\infty, x_j) \right) \\ D_N^- &= \sup_{1\leq j\leq N} \left(\Pr_t(-\infty, x_j) - \frac{j-1}{N} \right). \end{aligned} \tag{13.317}$$

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 empirical distribution $\Pr_e(-\infty, x)$ from which the events x_j came. In this case, the

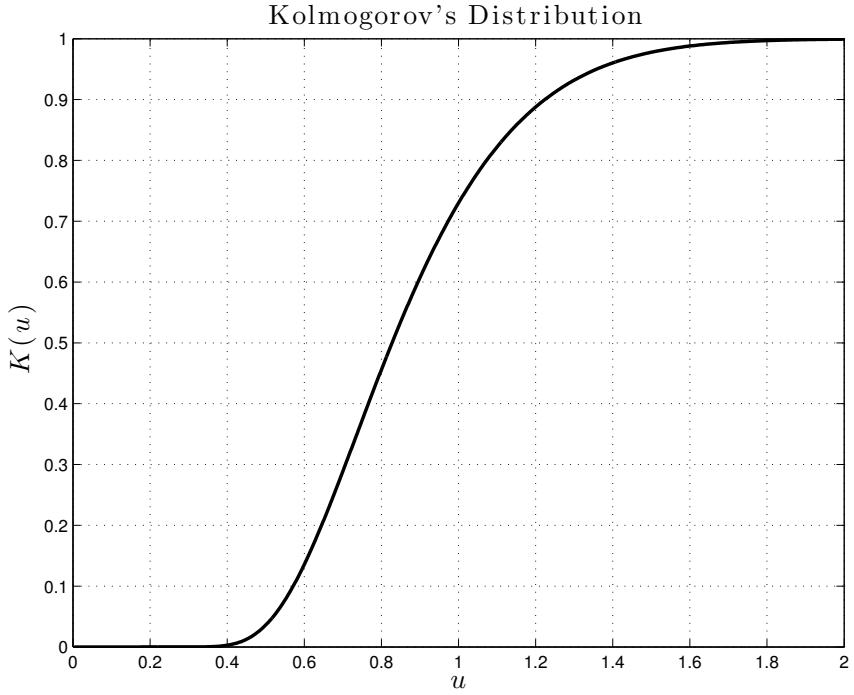

Figure 13.7 Kolmogorov's cumulative probability distribution $K(u)$ defined by (13.320) rises from zero to unity as u runs from zero to about two.

Kolmogorov distance D_N should converge to a limiting value D_∞

$$\lim_{N\to\infty} D_N = D_\infty = \sup_{-\infty<x<\infty} |\Pr_e(-\infty, x) - \Pr_t(-\infty, x)| \in [0, 1]. \quad (13.318)$$

If the empirical distribution $\Pr_e(-\infty, x)$ is the same as the theoretical distribution $\Pr_t(-\infty, x)$, 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

$$\Pr(D_\infty = 0) = 1. \quad (13.319)$$

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 theoretical distribution $\Pr_t(-\infty, x)$ is continuous, then for large N (and for $u > 0$) the probability that $\sqrt{N}\, D_N$

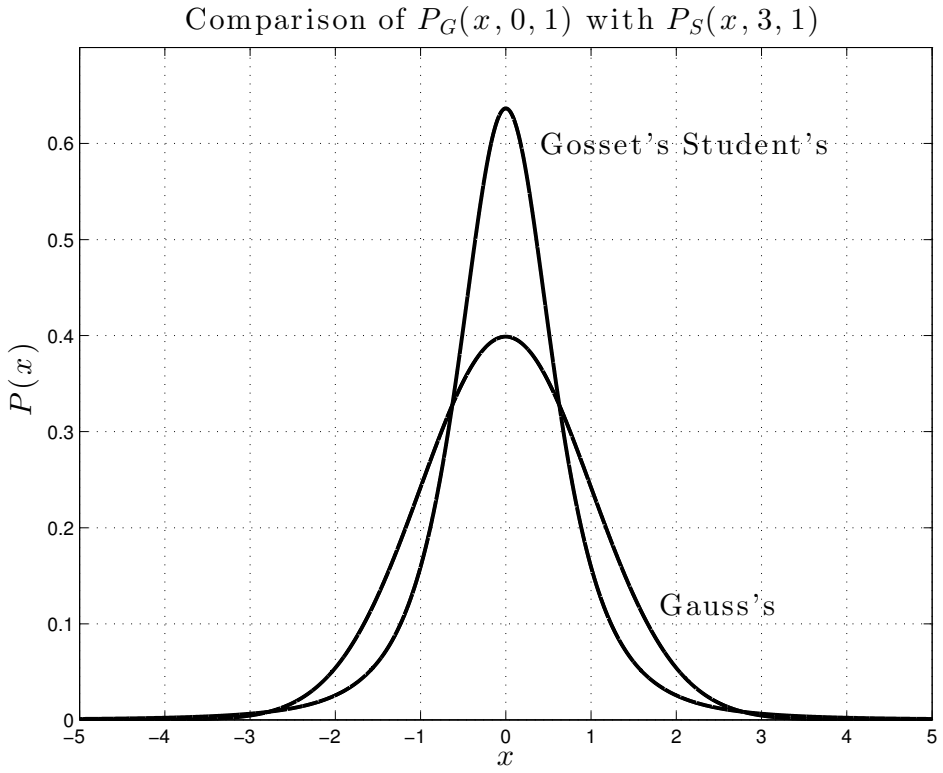

Figure 13.8 The probability distributions of Gauss $P_G(x,0,1)$ and Gosset/Student $P_S(x,3,1)$ with zero mean and unit variance.

is less than u is given by the **Kolmogorov function** $K(u)$

$$\lim_{N\to\infty} \Pr(\sqrt{N}\, D_N < u) = K(u) \equiv 1 + 2\sum_{k=1}^{\infty} (-1)^k e^{-2k^2u^2} \tag{13.320}$$

which is **universal and independent of the particular probability distributions** $\Pr_e(-\infty, x)$ and $\Pr_t(-\infty, x)$.

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

Example 13.19 (Kolmogorov's Test) How do we use (13.320)? As illustrated in Fig. 13.7, 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 theoretical distribution, then Kolmogorov's theorem (13.320) 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 theoretical 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 theoretical distribution $P_t(x) = P_G(x, 0, 1)$ which has the cumulative probability distribution (13.85)

$$\Pr_t(-\infty, x) = \frac{1}{2}\left[\text{erf}\left(x/\sqrt{2}\right) + 1\right]. \tag{13.321}$$

I generated $N = 10^m$ points x_j for $m = 1$, 2, 3, 4, 5, and 6 from the theoretical 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 theoretical distribution.

To see what happens when the experimental points do not come from the theoretical 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 $P_S(x, 3, 1)$ defined by (13.191) with $\nu = 3$ and $a = 1$. Both $P_t(x) = P_G(x, 0, 1)$ and $P_S(x, 3, 1)$ have the same mean $\mu = 0$ and standard deviation $\sigma = 1$, as illustrated in Fig. 13.8. 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 13.42) $D_\infty = 0.0868552356$.

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 $P_G(x, 0, 1)$ and from $P_S(x, 3, 1)$ 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 (13.310–13.312). 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 $u_{G,G,1}^{(m)} \le u_{G,G,2}^{(m)} \le \cdots \le u_{G,G,100}^{(m)}$ and $u_{S,G,1}^{(m)} \le u_{S,G,2}^{(m)} \le \cdots \le u_{S,G,100}^{(m)}$. I then used (13.310–13.312) to form the cumulative probabilities

$$\Pr_{e,G,G}^{(m)}(-\infty, u) = \frac{j}{N_s} \quad \text{for} \quad u_{G,G,j}^{(m)} < u < u_{G,G,j+1}^{(m)} \tag{13.322}$$

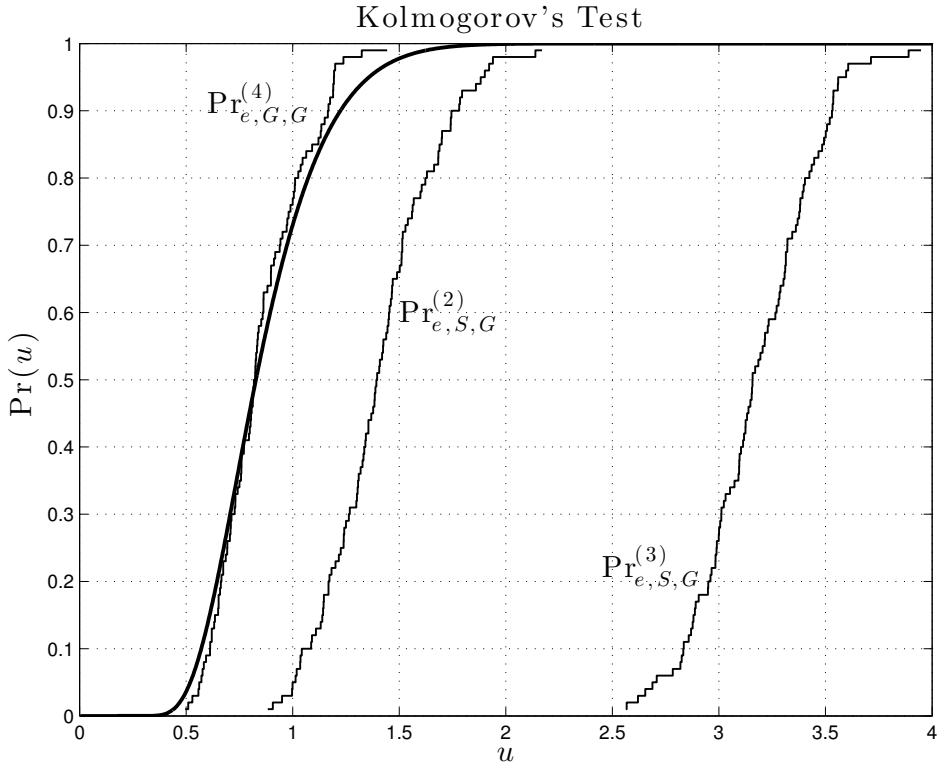

Figure 13.9 Kolmogorov's test is applied to points x_j taken from Gauss's distribution $P_G(x, 0, 1)$ and from Gosset's Student's distribution $P_S(x, 3, 1)$ to see whether the x_j came from $P_G(x, 0, 1)$. The thick smooth curve is Kolmogorov's universal cumulative probability distribution $K(u)$ defined by (13.320). The thin jagged curve that clings to $K(u)$ is the cumulative probability distribution $\Pr^{(4)}_{e,G,G}(-\infty, u)$ made (13.322) 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 (13.323) from 10^m points taken from $P_S(x, 3, 1)$.

and

$$\Pr^{(m)}_{e,S,G}(-\infty, u) = \frac{j}{N_s} \quad \text{for} \quad u^{(m)}_{S,G,j} < u < u^{(m)}_{S,G,j+1} \tag{13.323}$$

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

I plotted these cumulative probabilities in Fig. 13.9. The thick smooth curve is Kolmogorov's universal cumulative probability distribution $K(u)$ defined by (13.320). 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 13.43) that when the data do come from the theoretical probability distribution $P_t(x)$ assumed to be continuous, then the value of D_∞ is

$$D_\infty \approx \sup_{-\infty<x<\infty} \frac{w\, P_t(x)}{2}. \tag{13.324}$$

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 (13.310–13.312), 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

$$\begin{aligned} D^+_{N,N'} &= \sup_{-\infty<x<\infty} \left(\Pr^{(N)}_e(-\infty, x) - \Pr^{(N')}_e(-\infty, x) \right) \\ D_{N,N'} &= \sup_{-\infty<x<\infty} \left| \Pr^{(N)}_e(-\infty, x) - \Pr^{(N')}_e(-\infty, x) \right| \end{aligned} \tag{13.325}$$

which are analogous to (13.313–13.316). Smirnov (Smirnov 1939; Gnedenko

1968, p. 453) has shown that as $N, N' \to \infty$ the probabilities that

$$u^+_{N,N'} = \sqrt{\frac{NN'}{N+N'}}\, D^+_{N,N'} \quad \text{and} \quad u_{N,N'} = \sqrt{\frac{NN'}{N+N'}}\, D_{N,N'} \tag{13.326}$$

are less than u are

$$\begin{aligned} \lim_{N,N'\to\infty} \Pr(u^+_{N,N'} < u) &= 1 - e^{-2u^2} \\ \lim_{N,N'\to\infty} \Pr(u_{N,N'} < u) &= K(u) \end{aligned} \tag{13.327}$$

in which $K(u)$ is Kolmogorov's distribution (13.320).

Further Reading

Students can learn more about probability and statistics in *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

13.1 Find the probabilities that the sum on two thrown fair dice is 4, 5, or 6.

13.2 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.

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

13.4 In the formulas (13.21 & 13.28) for the variances of discrete and continuous distributions, show that $E[(x - \langle x\rangle)^2] = \mu_2 - \mu^2$.

13.5 A **convex** function is one that lies above its tangents: if $f(x)$ is convex, then $f(x) \geq f(y)+(x-y)f'(y)$. For instance, $e^x \geq 1+x$. Show that for any convex function $f(x)$ that $f(x) \geq f(\langle x\rangle) + (x - \langle x\rangle)f'(\langle x\rangle)$ and so that $\langle f(x)\rangle \geq f(\langle x\rangle)$ or $E[f(x)] \geq f(E[x])$ (Johan Jensen 1859–1925).

13.6 (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 (13.35). (b) Derive (13.39) for the variance $V[ax + by]$.

13.7 Derive expression (13.40) for the variance of a sum of N variables.

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

13.9 Show that the variance of the binomial distribution (13.43) is given by (13.47).

13.10 Redo the polling example (13.14–13.16) for the case of a slightly better poll in which 16 people 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.)

13.11 For the case in which N and $N - n$ are big, derive (13.52 & 13.53) from (13.43 & 13.51).

13.12 For the case in which N, $N-n$, and n are big, derive (13.54 & 13.55) from (13.43 & 13.51).

13.13 Without using the fact that the Poisson distribution is a limiting form of the binomial distribution, show from its definition (13.58) and its mean (13.60) that its variance is equal to its mean, as in (13.62).

13.14 Show that Gauss's approximation (13.74) to the binomial distribution is a normalized probability distribution with mean $\langle x \rangle = \mu = pN$ and variance $V[x] = pqN$.

13.15 Derive the approximations (13.88 & 13.89) for binomial probabilities for large N.

13.16 Compute the central moments (13.27) of the gaussian (13.75).

13.17 Derive formula (13.84) for the probability that a gaussian random variable falls within an interval.

13.18 Show that the expression (13.91) for $P(y|600)$ is negligible on the interval $(0, 1)$ except for y near $3/5$.

13.19 Determine the constant A of the homogeneous solution $\langle \boldsymbol{v}(t) \rangle_{gh}$ and derive expression (13.141) for the general solution $\langle \boldsymbol{v}(t) \rangle$ to (13.139).

13.20 Derive equation (13.142) 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 $\langle (\boldsymbol{v} - \langle \boldsymbol{v}(t) \rangle)^2 \rangle = 3kT/m$.

13.21 Derive equation (13.172) for the ensemble average $\langle \boldsymbol{r}^2(t) \rangle$ for the case in which $\langle \boldsymbol{r}^2(0) \rangle = 0$ and $d\langle \boldsymbol{r}^2(0) \rangle / dt = 0$.

13.22 Use (13.183) to derive the lower moments (13.185 & 13.186) of the distributions of Gauss and Poisson.

13.23 Find the third and fourth moments μ_3 and μ_4 for the distributions of Poisson (13.178) and Gauss (13.175).

13.24 Derive formula (13.190) for the first five cumulants of an arbitrary probability distribution.

13.25 Show that like the characteristic function, the moment-generating function $M(t)$ for an average of several independent random variables factorizes $M(t) = M_1(t/N)\, M_2(t/N) \cdots M_N(t/N)$.

13.26 Derive formula (13.197) for the moments of the log-normal probability distribution (13.196).

13.27 Why doesn't the log-normal probability distribution (13.196) have a sensible power-series about $x = 0$? What are its derivatives there?

13.28 Compute the mean and variance of the exponential distribution (13.198).

13.29 Show that the chi-square distribution $P_{3,G}(v, \sigma)$ with variance $\sigma^2 = kT/m$ is the Maxwell-Boltzmann distribution (13.100).

13.30 Compute the inverse Fourier transform (13.174) of the characteristic function (13.203) of the symmetric Lévy distribution for $\nu = 1$ and 2.

13.31 Show that the integral that defines $P^{(2)}(y)$ gives formula (13.239) with two Heaviside functions. Hint: keep x_1 and x_2 in the interval $(0, 1)$.

13.32 Derive the normal distribution (13.224) in the variable (13.223) from the central limit theorem (13.221) for the case in which all the means and variances are the same.

13.33 Show that Fisher's matrix (13.257) is symmetric $F_{k\ell} = F_{\ell k}$ and non-negative (1.38), and that when it is positive (1.39), it has an inverse.

13.34 Derive the integral equations (13.259 & 13.260) from the normalization condition $\int P(\boldsymbol{x}; \boldsymbol{\theta})\, d^N x = 1$.

13.35 Derive the Cramér-Rao lower bound (13.275) on the variance $V[t_k]$ from the inequality (13.270).

13.36 Show that the variance $V[u^{(N)}_{\sigma^2}]$ of Bessel's estimator (13.254) is given by (13.281).

13.37 Compute the fourth central moment (13.27) of Gauss's probability distribution $P_G(x; \mu, \sigma^2)$.

13.38 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 $P(I - P) = (I - P)P = 0$.

13.39 Show that Kolmogorov's distance D_N is bounded as in (13.314).

13.40 Show that Kolmogorov's distance D_N is the greater of D_N^+ and D_N^-.

13.41 Derive the formulas (13.317) for D_N^+ and D_N^-.

13.42 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 (13.321 & 13.194) to find the value of x that maximizes their difference. Using Maple or some other program, you should find $x = 0.6276952185$ and then $D_\infty = 0.0868552356$.

13.43 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 (13.324).

13.44 Suppose in a poll of 1000 likely voters, 510 have said they would vote for Nancy Pelosi. Redo example 13.9.