Chapter 1:
Introduction to statistical methods

Zhi-Jie Tan

Wuhan University

2019 spring semester

1, General discussion of the random walk

But there is limitations:

1, Each step has the same length

2, 1 Dimension

Require a more powerful method that can be readily generalized……

1.7 Probability distributions involving several variables

Beyond 1D problem

Considering the case of two variables *u* and *v* , *u* and *v* have M and N possible values

u_i where $i = 1, 2, \ldots, M$

and v_j where $j = 1, 2, \ldots, N$

Let P(u_i, v_j) be the probability that $u=u_i$ and $v=v_j$

The probability that the variables u and v assume any of their possible sets of values must be unity; i.e., one has the normalization requirement

Normalization requires $$\sum_{i=1}^{M} \sum_{j=1}^{N} P(u_i, v_j) = 1 \qquad (1\cdot7\cdot1)$$

where the summation extends over all possible values of u and all possible values of v.

The probability $P_u(u_i)$ that u assumes the value u_i, *irrespective of* the value assumed by the variable v, is the sum of the probabilities of all possible situations consistent with the given value of u_i; i.e.,

$$P_u(u_i) = \sum_{j=1}^{N} P(u_i, v_j) \qquad (1\cdot7\cdot2)$$

where the summation is over all possible values of v_j. Similarly, the probability $P_v(v_j)$ that v assumes the value v_j, irrespective of the value assumed by u, is

$$P_v(v_j) = \sum_{i=1}^{M} P(u_i, v_j) \qquad (1\cdot7\cdot3)$$

A special case:

a variable assumes a certain value does not depend on the value assumed by the other variable. The two variables are statistically "independent".

$$P(u_i, v_j) = P_u(u_i) P_v(v_j)$$

Mean values for the case of two variables

Let us now mention some properties of mean values. If $F(u,v)$ is any function of u and v, then its mean value is defined by

$$\overline{F(u,v)} \equiv \sum_{i=1}^{M} \sum_{j=1}^{N} P(u_i,v_j)F(u_i,v_j) \qquad (1\cdot7\cdot6)$$

Note that if $f(u)$ is a function of u only, it also follows by (1·7·2) that

$$\overline{f(u)} = \sum_i \sum_j P(u_i,v_j)f(u_i) = \sum_i P_u(u_i)f(u_i) \qquad (1\cdot7\cdot7)$$

If F and G are any functions of u and v, then one has the general result

$$\overline{F+G} \equiv \sum_i \sum_j P(u_i,v_j)[F(u_i,v_j) + G(u_i,v_j)]$$
$$= \sum_i \sum_j P(u_i,v_j)F(u_i,v_j) + \sum_i \sum_j P(u_i,v_j)G(u_i,v_j)$$

or

$$\overline{F+G} = \overline{F} + \overline{G} \qquad (1\cdot7\cdot8)$$

i.e., the average of a sum equals simply the sum of the averages.

Special case:

a variable assumes a certain value does not depend on the value assumed by the other variable. The two variables are statistically "independent".

Given any two functions $f(u)$ and $g(v)$, one can also make a general statement about the mean value of their product *if* u and v are statistically independent variables. Indeed, one then finds

$$\begin{aligned}\overline{f(u)g(v)} &= \sum_i \sum_j P(u_i,v_j) f(u_i) g(v_j) \\ &= \sum_i \sum_j P_u(u_i) P_v(v_j) f(u_i) g(v_j) \qquad \text{by (1-7-5)} \\ &= \left[\sum_i P_u(u_i) f(u_i)\right] \left[\sum_j P_v(v_j) g(v_j)\right]\end{aligned}$$

Thus

$$\overline{f(u)g(v)} = \overline{f(u)}\ \overline{g(v)} \qquad (1\text{-}7\text{-}9)$$

i.e., the average of a product equals the product of the averages *if* u and v are *statistically independent*. If u and v are statistically *not* independent, the statement (1-7-9) is in general *not* true.

1.8 Comments on continuous probability distributions

***u* can assume any value in continuous range [a_1--a_2]**
Finding the probability for *u* in [*u*—*u*+*du*].

Dividing [a_1--a_2] into equal intervals with δu

δu ←---------→ $P(u)$
du ←---------→ $\mathcal{P}(u)$

Probability density

To make the connection between the continuous and discrete points of view quite explicit, note that in terms of the original infinitesimal subdivision interval δu,

$$P(u) = \mathcal{P}(u)\,\delta u$$

Similarly, if one considers any interval between u and $u + du$ which is such that du is macroscopically small although $du \gg \delta u$, then this interval contains $du/\delta u$ possible values of u_i for which the probability $P(u_i)$ has essentially the same value—call it simply $P(u)$. Then the probability $P(u)\,du$ of

Mean value for continuous probability distribution: single variable

$\mathcal{P}(u)\,du$:

$$\mathcal{P}(u)\,du \approx P(u_i)\frac{du}{\delta u} = \frac{P(u)}{\delta u}\,du$$

Correlations between continuous … and discrete …

Normalization:

$$\sum_i P(u_i) = 1 \quad \Leftrightarrow \quad \int_{a_1}^{a_2} \mathcal{P}(u)\,du = 1$$

Mean value:

$$\overline{f(u)} = \sum_i P(u_i) f(u_i) \quad \Leftrightarrow \quad \overline{f(u)} = \int_{a_1}^{a_2} \mathcal{P}(u) f(u)\,du$$

Mean value for continuous probability distribution: multiple variables

$\mathcal{P}(u,v)\,du\,dv$:

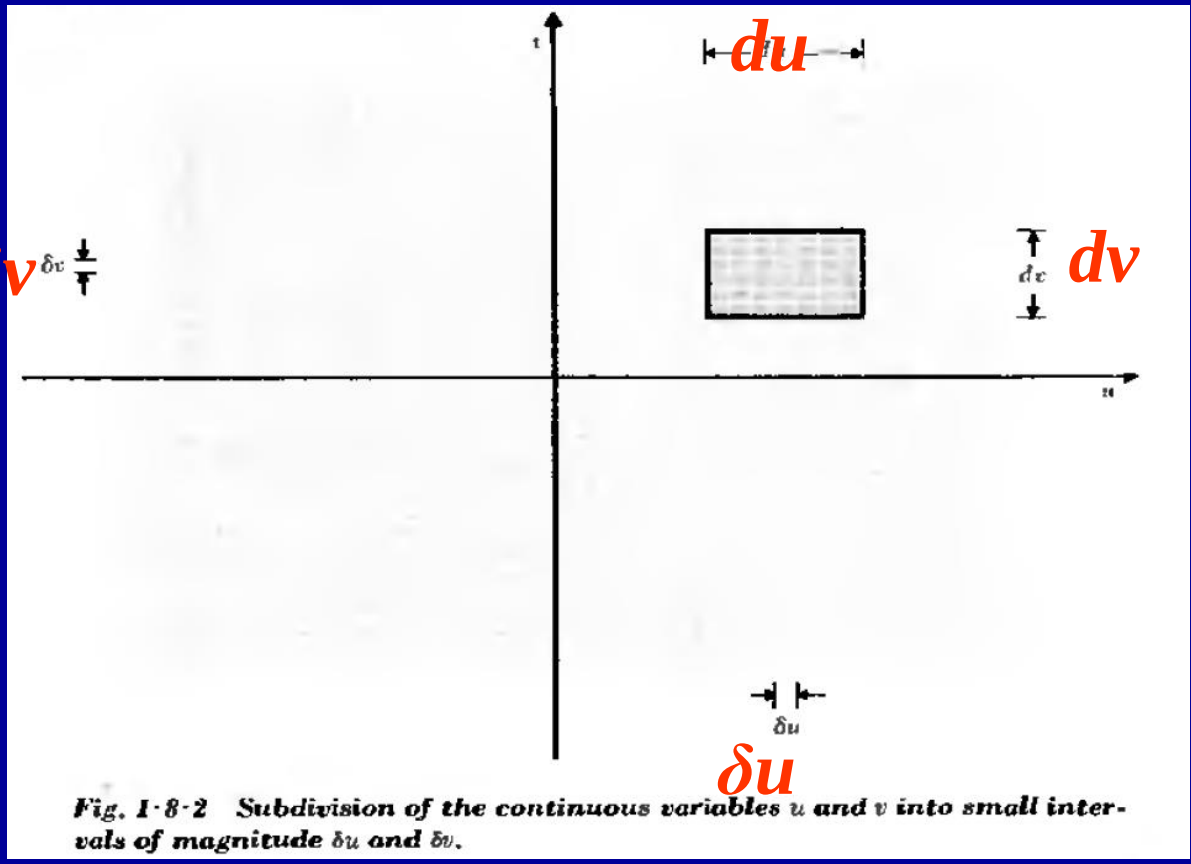


Fig. 1·8·2 Subdivision of the continuous variables u and v into small intervals of magnitude δu and δv.

In analogy to only u:

$$\mathcal{P}(u,v)\,du\,dv = P(u,v)\,\frac{du\,dv}{\delta u\,\delta v}$$

Mean value for continuous probability distribution: multiple variables

Normalization condition:

$$\int_{a_1}^{a_2}\int_{b_1}^{b_2} du\,dv\,\mathcal{P}(u,v) = 1$$

In analogy to case of only u:

$$\overline{F(u,v)} = \int_{a_1}^{a_2}\int_{b_1}^{b_2} du\,dv\,\mathcal{P}(u,v)F(u,v)$$

Generally :

$$\overline{F+G} = \overline{F} + \overline{G}$$

u & v statistically independent

$$\overline{f(u)g(v)} = \overline{f(u)}\;\overline{g(v)}$$

Functions of random variables

A question :

$\varphi(u)$ is some continuous function of u.

If $\mathcal{P}(u)\,du$ is the probability that u lies in the range between u and $u + du$, what is the corresponding probability $W(\varphi)\,d\varphi$ that φ lies in the range between φ and $\varphi + d\varphi$? Clearly, the latter probability is obtained by adding up the probabilities for all those values u which are such that φ lies in the range between φ and $\varphi + d\varphi$; in symbols

$$W(\varphi)\,d\varphi = \int_{d\varphi} \mathcal{P}(u)\,du \qquad (1\cdot8\cdot8)$$

φ in [φ–- φ+dφ]

Here u can be considered a function of φ and the integral extends over all those values of u lying in the range between $u(\varphi)$ and $u(\varphi + d\varphi)$. Thus (1·8·8) becomes simply

$$W(\varphi)\,d\varphi = \int_{\varphi}^{\varphi+d\varphi} \mathcal{P}(u)\left|\frac{du}{d\varphi}\right| d\varphi = \mathcal{P}(u)\left|\frac{du}{d\varphi}\right| d\varphi \qquad (1\cdot8\cdot9)$$

answer :

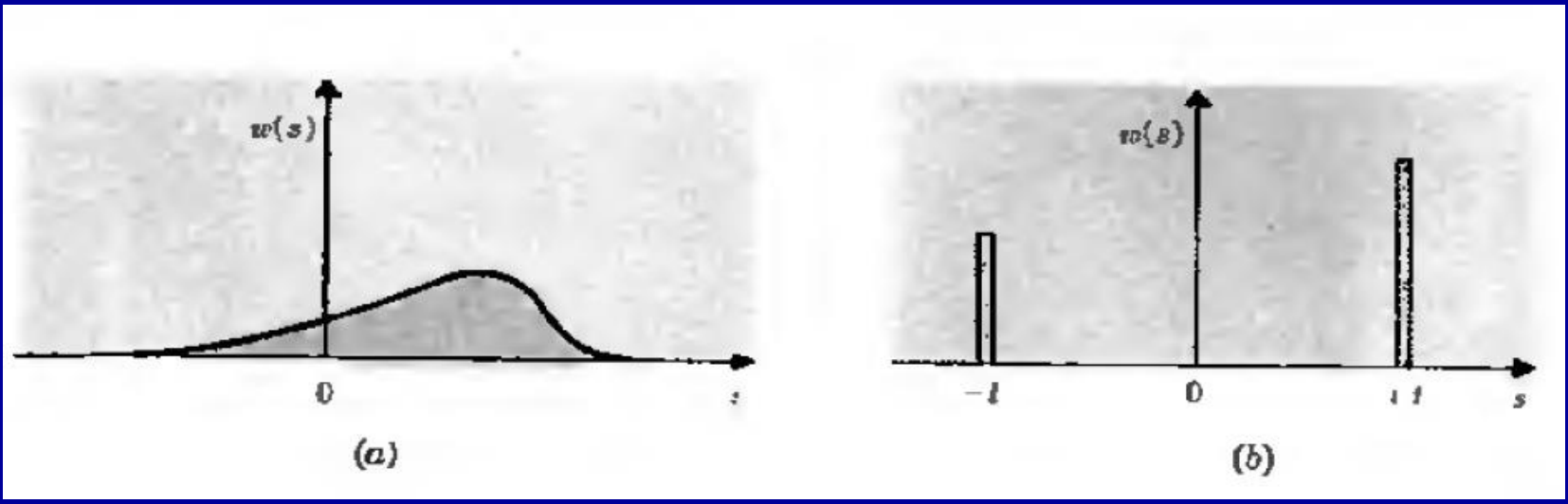
Similar argument for multiple variables!

$$\int_{\phi}^{\phi+d\phi} d\phi = d\phi$$

1.9 General calculation of mean values for random walk

Consider a more general case: step is not fixed

Let $w(s_i)\,ds_i$ be the probability that the ith displacement lies in the range between s_i and $s_i + ds_i$.



Total displacement after N steps $x = ?$

$P\,dx = ?$

The total displacement

For simplicity, w is the same for any i

The total displacement x is equal to

$$x = s_1 + s_2 + \cdots + s_N = \sum_{i=1}^{N} s_i \qquad (1\cdot9\cdot1)$$

Taking mean values of both sides,

$$\bar{x} = \overline{\sum_{i=1}^{N} s_i} = \sum_{i=1}^{N} \bar{s}_i \qquad (1\cdot9\cdot2)$$

where we have used the property (1·7·8). But since $w(s_i)$ is the same for each step, independent of i, each mean value $\bar{s}_i$ is the same. Thus (1·9·2) is simply the sum of N equal terms and becomes

$$\bar{x} = N\bar{s} \qquad (1\cdot9\cdot3)$$

where

$$\bar{s} \equiv \bar{s}_i = \int ds\, w(s)s \qquad (1\cdot9\cdot4)$$

is merely the mean displacement per step.

The dispersion

Next we calculate the dispersion

$$\overline{(\Delta x)^2} \equiv \overline{(x - \bar{x})^2} \tag{1·9·5}$$

By (1·9·1) and (1·9·2) one has

$$x - \bar{x} = \sum_i (s_i - \bar{s})$$

or

$$\Delta x = \sum_{i=1}^{N} \Delta s_i \tag{1·9·6}$$

where

$$\Delta s = s_i - \bar{s} \tag{1·9·7}$$

By squaring (1·9·6) one obtains

$$(\Delta x)^2 = \left(\sum_{i=1}^{N} \Delta s_i\right)\left(\sum_{j=1}^{N} \Delta s_j\right) = \sum_i (\Delta s_i)^2 + \sum_{i} \sum_{j \atop i \neq j} (\Delta s_i)(\Delta s_j) \tag{1·9·8}$$

In the cross terms we make use of the fact that different steps are statistically independent and apply the relation (1·7·9) to write for $i \neq j$

$$\overline{(\Delta s_i)(\Delta s_j)} = \overline{(\Delta s_i)}\,\overline{(\Delta s_j)} = 0 \tag{1·9·10}$$

since

$$\overline{\Delta s_i} = \bar{s}_i - \bar{s} = 0$$

The dispersion

$$(\Delta x)^2 = \left(\sum_{i=1}^{N} \Delta s_i\right)\left(\sum_{j=1}^{N} \Delta s_j\right) = \sum_i (\Delta s_i)^2 + \sum_i \sum_{j \neq i} (\Delta s_i)(\Delta s_j)$$

$$\overline{(\Delta x)^2} = \sum_{i=1}^{N} \overline{(\Delta s_i)^2} \qquad (1\cdot 9\cdot 11)$$

Of course, none of these square terms can be negative. Since the probability distribution $w(s_i)$ is the same for each step, independent of i, it again follows that $\overline{(\Delta s_i)^2}$ must be the same for each step. Thus the sum in (1·9·11) consists merely of N equal terms and becomes simply

$$\overline{(\Delta x)^2} = N\overline{(\Delta s)^2} \qquad (1\cdot 9\cdot 12)$$

where

$$\overline{(\Delta s)^2} \equiv \overline{(\Delta s_i)^2} = \int ds\, w(s)(\Delta s)^2 \qquad (1\cdot 9\cdot 13)$$

is just the dispersion of the displacement per step.

$$\Delta^* x = \left[\overline{\Delta x^2}\right]^{1/2}$$

$$\frac{\Delta^* x}{\bar{x}} = \frac{\Delta^* s}{\bar{s}} \frac{1}{\sqrt{N}}$$

Root mean square from the mean value

1.10 Calculation of the probability distribution

Total displacement

$$x = \sum_{i=1}^{N} s_i$$

To find probability $P(x)dx$ for x in [x, x+dx]

Since the steps are statistically independent, the probability of *a particular* sequence of steps where

the 1st displacement lies in the range between s_1 and $s_1 + ds_1$
the 2nd displacement lies in the range between s_2 and $s_2 + ds_2$
. . .
the Nth displacement lies in the range between s_N and $s_N + ds_N$

is simply given by the product of the respective probabilities, i.e., by

$$w(s_1)\,ds_1 \cdot w(s_2)\,ds_2 \cdot \cdots \cdot w(s_N)\,ds_N$$

To find probability $\mathcal{P}(x)dx$ for x in [x, x+dx]

If we sum this probability over all the possible individual displacements which are consistent with the condition that the total displacement x in (1·10·1) always lies in the range between x and $x + dx$, then we obtain the total probability $\mathcal{P}(x)\,dx$, irrespective of the sequence of steps producing this total displacement. In symbols we can write

$$\mathcal{P}(x)\,dx = \underbrace{\int\cdots\int}_{(dx)} w(s_1)w(s_2)\cdots w(s_N)\,ds_1\,ds_2\cdots ds_N \qquad (1\cdot10\cdot2)$$

where the integration is over all possible values of the variables s_i, subject to the restriction that

$$x < \sum_{i=1}^{N} s_i < x + dx \qquad (1\cdot10\cdot3)$$

In principle, evaluation of the integral (1·10·2) solves completely the problem of finding $\mathcal{P}(x)$.

To find probability $\mathcal{P}(x)dx$ for x in [x, x+dx]

This can readily be done by multiplying the integrand in (1·10·2) by a factor which is equal to unity when the s_i are such that (1·10·3) is satisfied, but which equals zero otherwise. The Dirac δ function $\delta(x - x_0)$, discussed in Appendix A·7, has precisely the selective property that it vanishes whenever $|x - x_0| > \frac{1}{2}|dx|$, while it becomes infinite like $(dx)^{-1}$ in the infinitesimal range where $|x - x_0| < \frac{1}{2}|dx|$; i.e., $\delta(x - x_0)\,dx = 1$ in this latter range. Hence (1·10·2) can equally well be written

$$\mathcal{P}(x)\,dx = \int\!\!\int_{-\infty}^{\infty} \cdots \int w(s_1)w(s_2)\cdots w(s_N)\left[\delta\left(x - \sum_{i=1}^{N} s_i\right)dx\right] ds_1\,ds_2\cdots ds_N \qquad (1\cdot10\cdot4)$$

where there is now *no* further restriction on the domain of integration. At this point we can use the convenient analytical representation of the δ function in terms of the integral (A·7·14); i.e., we can write

$$\delta(x - \Sigma s_i) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dk\, e^{ik[\Sigma s_i - x]} \qquad (1\cdot10\cdot5)$$

To find probability $\mathcal{P}(x)dx$ for x in [x, x+dx]

Substituting this result in (1·10·4) yields:

$$\mathcal{P}(x) = \int\int \cdots \int w(s_1)w(s_2) \cdots w(s_N) \frac{1}{2\pi} \int_{-\infty}^{\infty} dk\, e^{ik(s_1+\cdots+s_N-x)}\, ds_1\, ds_2 \cdots ds_N$$

$$\text{or} \quad \mathcal{P}(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dk\, e^{-ikx} \int_{-\infty}^{\infty} ds_1\, w(s_1)\, e^{iks_1} \cdots \int_{-\infty}^{\infty} ds_N\, w(s_N)\, e^{iks_N} \qquad (1\cdot10\cdot6)$$

where we have interchanged the order of integration and used the multiplicative property of the exponential function. Except for the irrelevant symbol used as variable of integration, each of the last N integrals is identical and equal to

$$Q(k) \equiv \int_{-\infty}^{\infty} ds\, e^{iks} w(s) \qquad (1\cdot10\cdot7)$$

Hence (1·10·6) becomes

$$\mathcal{P}(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dk\, e^{-ikx} Q^N(k) \qquad (1\cdot10\cdot8)$$

Thus the evaluation of two simple (Fourier) integrals solves the problem completely.

1.11 Probability distribution for large N

$$Q(k) \equiv \int_{-\infty}^{\infty} ds\, e^{iks} w(s)$$

$\mathcal{P}(x)$: Solving for large N ??

$$\mathcal{P}(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} dk\, e^{-ikx} Q^N(k)$$

The integrand in (1·10·7) contains the factor e^{iks}, which is an oscillatory function of s and oscillates the more rapidly with increasing magnitude of k. Hence the quantity $Q(k)$ given by the integral in (1·10·7) tends in general to be increasingly small as k becomes large. (See remark below.) If Q is raised to a large power N, it thus follows that $Q^N(k)$ tends to decrease very rapidly with increasing k. To compute $\mathcal{P}(x)$ by Eq. (1·10·8), a knowledge of $Q^N(k)$ for *small* values of k is then sufficient for calculating the integral, since for large values of k the contribution of $Q^N(k)$ to this integral is negligibly small. But for small values of k, it should be possible to approximate $Q^N(k)$ by a suitable expansion in powers of k. Since $Q^N(k)$ is a rapidly varying function of k, it is preferable (as in Sec. 1·5) to seek the more readily convergent power series expansion of its slowly varying logarithm $\ln Q^N(k)$.

$Q(k)$??

The actual calculation is straightforward. We want first to compute $Q(k)$ for small values of k. Expanding e^{iks} in Taylor's series, Eq. (1·10·7) becomes

$$Q(k) \equiv \int_{-\infty}^{\infty} ds\, w(s)\, e^{iks} = \int_{-\infty}^{\infty} ds\, w(s)(1 + iks - \tfrac{1}{2}k^2s^2 + \cdots)$$

or
$$Q(k) = 1 + i\bar{s}k - \tfrac{1}{2}\overline{s^2}k^2 \cdots \qquad (1\cdot11\cdot1)$$

where
$$\overline{s^n} \equiv \int_{-\infty}^{\infty} ds\, w(s)s^n \qquad (1\cdot11\cdot2)$$

is a constant which represents the usual definition of the nth moment of s. Here we assume that $|w(s)| \to 0$ rapidly enough as $|s| \to \infty$ so that these moments are finite. Hence (1·11·1) yields

$$\ln Q^N(k) = N \ln Q(k) = N \ln [1 + \underbrace{i\bar{s}k - \tfrac{1}{2}\overline{s^2}k^2 \cdots}_{y}] \qquad (1\cdot11\cdot3)$$

Using the Taylor's series expansion valid for $y \ll 1$,

$$\ln (1 + y) = y - \tfrac{1}{2}y^2 \cdots$$

$Q(k)$??

Eq. (1 · 11 · 3) becomes, up to terms quadratic in k,

$$\begin{aligned} \ln Q^N &= N[i\bar{s}k - \tfrac{1}{2}\overline{s^2}k^2 - \tfrac{1}{2}(i\bar{s}k)^2 \cdots] \\ &= N[i\bar{s}k - \tfrac{1}{2}(\overline{s^2} - \bar{s}^2)k^2 \cdots] \\ &= N[i\bar{s}k - \tfrac{1}{2}\overline{(\Delta s)^2}k^2 \cdots] \end{aligned}$$

where $$\overline{(\Delta s)^2} \equiv \overline{s^2} - \bar{s}^2 \qquad (1\cdot11\cdot4)$$

Hence we obtain

$$Q^N(k) = e^{iN\bar{s}k - \frac{1}{2}N\overline{(\Delta s)^2}k^2} \qquad (1\cdot11\cdot5)$$

Thus (1 · 10 · 8) becomes

$$\mathcal{P}(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} dk\, e^{i(N\bar{s}-x)k - \frac{1}{2}N\overline{(\Delta s)^2}k^2} \qquad (1\cdot11\cdot6)$$

$$\begin{aligned} \int_{-\infty}^{\infty} du\, e^{-au^2+bu} &= \int_{-\infty}^{\infty} du\, e^{-a[u^2-(b/a)u]} \\ &= \int_{-\infty}^{\infty} du\, e^{-a(u-b/2a)^2+b^2/4a} && \text{by completing the square} \\ &= e^{b^2/4a}\int_{-\infty}^{\infty} dy\, e^{-ay^2} && \text{by putting } y = u - \frac{b}{2a} \\ &= e^{b^2/4a}\sqrt{\frac{\pi}{a}} && \text{by (A·4·2)} \end{aligned}$$

Thus

$$\int_{-\infty}^{\infty} du\, e^{-au^2+bu} = \sqrt{\frac{\pi}{a}}\, e^{b^2/4a} \qquad (1\cdot11\cdot7)$$

$P(x)$??

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}$$

$$\left.\begin{aligned} \mu &\equiv N\bar{s} \\ \sigma^2 &\equiv N\overline{(\Delta s)^2} \end{aligned}\right\}$$

Thus the distribution has the Gaussian form previously encountered in Sec. 1·6. Note, however, the extreme generality of this result. *No matter what* the probability distribution $w(s)$ for each step may be, as long as the steps are *statistically independent* and $w(s)$ falls off rapidly enough as $|s| \to \infty$, the total displacement x will be distributed according to the Gaussian law *if N is sufficiently large.* This very important result is the content of the so-called "central limit theorem," probably the most famous theorem in mathematical probability theory.* The generality of the result also accounts for the fact that so many phenomena in nature (e.g., errors in measurement) obey approximately a Gaussian distribution.

$$\left.\begin{aligned} \bar{x} &= \mu \\ \overline{(\Delta x)^2} &= \sigma^2 \end{aligned}\right\}$$

$$\left.\begin{aligned} \bar{x} &= N\bar{s} \\ \overline{(\Delta x)^2} &= N\overline{(\Delta s)^2} \end{aligned}\right\}$$

Homework: Textbook page 40-43

1.17 1.18

For random polymer

In 3 dimension:

$$P(x,y,z,N)\,dx\,dy\,dz = P(x,N)P(y,N)P(z,N)\,dx\,dy\,dz$$
$$= \left[\frac{\beta}{\pi}\right]^{3/2} e^{-\beta(x^2+y^2+z^2)}\,dx\,dy\,dz.$$

In terms of the vector **r**, you have

$$P(\mathbf{r},N) = P(x,y,z,N) = \left[\frac{\beta}{\pi}\right]^{3/2} e^{-\beta r^2}.$$

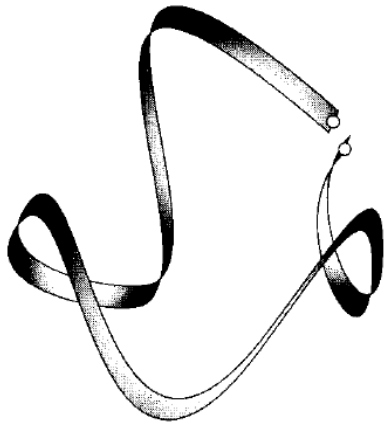


Figure 32.8 For polymer cyclization, the two chain ends must be close together.

Probability of finding a N-mer polymer with end-to-end distance r in 3D

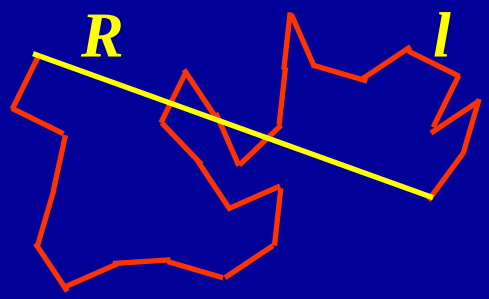
$$4\pi r^2 P(\mathbf{r},N)$$

Polymer cyclization (Jacobson–Stockmayer theory).

$$P_{\text{cyclization}} = \int_0^b P(r,N)\,dr$$
$$= \left[\frac{3}{2\pi N b^2}\right]^{3/2} \int_0^b e^{-3r^2/2Nb^2} 4\pi r^2\,dr. \qquad (32.21)$$

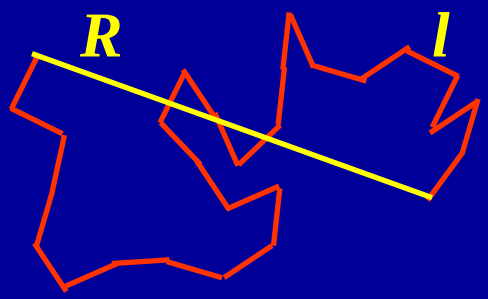
On-class work:

The distance between the starting and end points for *N*-step random walk is *R*. The step length is *l*. Please show that $\langle R^2 \rangle = Nl^2$



An optional problem at home:

If the walk is not purely random, the two adjacent walks are correlated with a energy $k\cos\theta$, how to calculate the distribution of end-to-end distance ?



$P(R)$=???