CHAPTER 8 - DATA ANALYSIS

We all know that a single experimental measurement in physics does not yield a result that is reproducible to arbitrary accuracy by another single measurement. At the operational level, the equipment, its settings and its operators are not identical in successive experiments, so that successive measurements may produce similar, but not identical results. At the fundamental level, quantum mechanics reminds us that the measurement process itself affects the system being measured. Hence, a sequence of experiments produces a distribution of values for the observable being measured. Further, the values predicted for the measurements also are subject to uncertainties, both through approximations made in analytical calculations and in the limited accuracy of numerical methods.

In this section, we address several questions about the analysis of distributions of numbers, such as are found in data sets. In particular, how do we
•characterize the central value of a distribution?
•assess the reproducibility of a distribution?
•interpret the distribution by means of functional representations?
Each of these tasks is dealt with in Secs. 8.3-8.7. But in order to understand the methodology, we want to have a data set whose properties are known theoretically. In Secs. 8.1-8.2, we use the ideal random walk to provide such a data set. Finally, we analyse data from the random walk in Sec. 8.8.

8.1 Ensemble averages

In Chap. 3, we concentrated on the end-to-end displacement of an ideal chain or random walk. "Ideal" means that the walk is allowed to intersect itself or, for a walk in one dimension, reverse its direction. In a given sample of configurations, there will be a distribution of end-to-end displacements $\boldsymbol{r}_{ee}$, as shown in Fig. 8.1. Each of the sample configurations in Fig. 8.1 has a different value for the magnitude of $\boldsymbol{r}_{ee}$; of course, there are many more possibilities than are present in the figure. From the ensemble of configurations, we can construct a variety of observables that reflect the "center" of the distribution. One such observable is the ensemble average or *mean* (see Sec. 8.3 for others). First, a quick review of a result from Chap. 3.

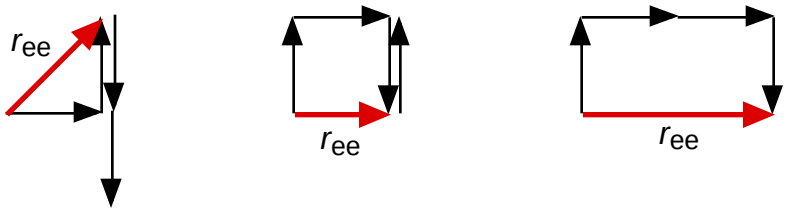

Fig. 8.1 Sample of random walks with four steps on a two-dimensional lattice.

Since $\boldsymbol{r}_{ee}$ is the vector sum of the individual steps $\boldsymbol{a}_i$,

$$\boldsymbol{r}_{ee} = \sum_i \boldsymbol{a}_i, \tag{8.1}$$

then the ensemble average $\langle \boldsymbol{r}_{ee}^2 \rangle$ over all chains with the same number of steps N_{step}, is

$$\langle \boldsymbol{r}_{ee}^2 \rangle = \sum_i \sum_j \langle \boldsymbol{a}_i \cdot \boldsymbol{a}_j \rangle. \tag{8.2}$$

If steps $\boldsymbol{a}_i$ and $\boldsymbol{a}_j$ have random orientations, then the ensemble average of $\boldsymbol{a}_i \cdot \boldsymbol{a}_j$ should vanish for $i \neq j$:

$$\langle \boldsymbol{a}_i \cdot \boldsymbol{a}_j \rangle = 0 \quad \text{for } i \neq j. \tag{8.3}$$

The only non-zero terms on the right-hand side of Eq. (8.2) have $i = j$, and are equal to a^2. Thus, for an ideal random walk

$$\langle \boldsymbol{r}_{ee}^2 \rangle = N_{step} a^2. \tag{8.4}$$

We use N_{step} in Eq. (8.4), rather than N as used previously, to reduce the notational ambiguities in the project of Sec. 8.8.

8.2 Distribution of random walks

For the random walk on a lattice, the distribution of $\boldsymbol{r}_{ee}$ can be denumerated exactly. Consider an ideal walk in 1 dimension, in which the steps all have equal length a. All of the allowed configurations with two steps are shown in Fig. 8.2. The figure shows the order in which the steps are made, and also the end-to-end

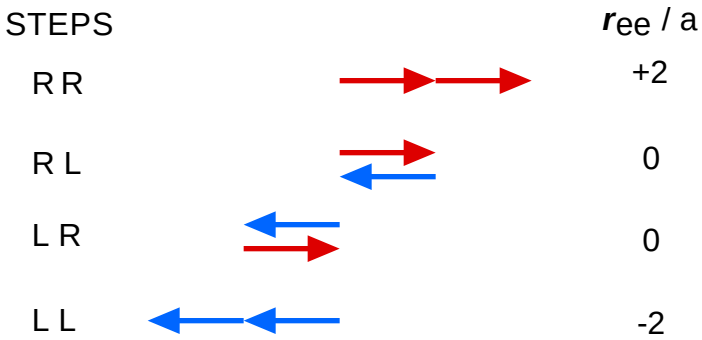

Fig. 8.2 Allowed configurations for a one-dimensional random walk with two steps of fixed length *a*.

displacement once the two-step walk is complete. We see that one walk has $\boldsymbol{r}_{ee}/a = 2$, two walks have $\boldsymbol{r}_{ee}/a = 0$, and one walk has $\boldsymbol{r}_{ee}/a = -2$. For larger number of steps, the distributions look like:

$\boldsymbol{r}_{ee}/a$ =	-4	-3	-2	-1	0	1	2	3	4
$N_{step} = 2$			1		2		1		
$N_{step} = 3$		1		3		3		1	
$N_{step} = 4$	1		4		6		4		1

The number of configurations for a given $\boldsymbol{r}_{ee}/a$ have a familiar form: the are the binomial coefficients. A few minutes' thought about how to count the configurations in an ideal *n*-step walk in one dimension will convince one that the distribution of $\boldsymbol{r}_{ee}/a$ is indeed a binomial distribution $(R + L)^n$. For example, with $N_{step} = 4$

Steps	RRRR	RRRL RRLR RLRR LRRR	RRLL RLRL LRRL LRLR LLRR RLLR	RLLL LRLL LLRL LLLR	LLLL
$(R + L)^3$	RRRR	RRRL RRLR RLRR LRRR	RRLL RLRL LRRL LRLR LLRR RLLR	RLLL LRLL LLRL LLLR	LLLL

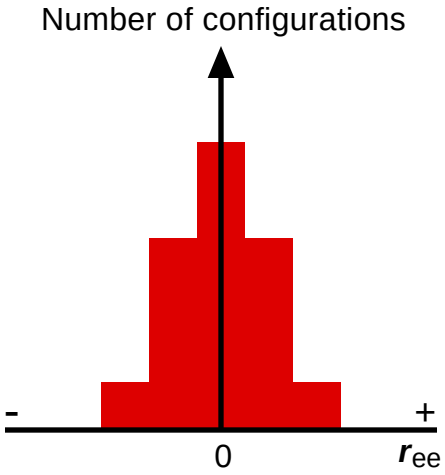

Fig. 8.3 Distribution of $\boldsymbol{r}_{ee}/a$ for a 4-step random walk.

As can be seen from the $N_{step} = 4$ data (schematically from the explicit configurations displayed above, or graphically in Fig. 8.3), the distribution of $\boldsymbol{r}_{ee}/a$ is symmetric around zero, and falls off monotonically as $|\boldsymbol{r}_{ee}/a|$ becomes large. As the number of steps becomes very large, the distribution becomes *Gaussian* or *normal*. In the continuum limit of large N_{step}, one is usually less concerned with the number of discrete states at a specific $\boldsymbol{r}_{ee}/a$, than with the probability of finding the system in a given range of $\boldsymbol{r}_{ee}/a$.

To make the notation a little less cumbersome, we replace r_{ee}/a by the variable x. The probability of finding a walk with displacement between x and $x + dx$ is just

$$P(x)dx = (2\pi\omega^2)^{-1/2} \exp(-x^2/2\omega^2)\, dx \qquad (8.5)$$

where $P(x)$ is the probability density (probability per unit length) and ω is

$$\omega^2 = N_{step}a^2/d. \qquad (8.6)$$

We use ω to represent the width, rather than the conventional σ introduced earlier (Chap. 3), since the project of Sec. 8.8 involves the determination of the standard deviation (σ) and standard error (σ_e) of ω, as found from small samples. The quantity ω is dimension-dependent, and d is the spatial dimension in which the walk is embedded. For one-dimensional walks, obviously $d = 1$.

The distribution in Eq. (8.6) assumes that the chain starts at the origin, and is normalized to unity

$$\int P(x)dx = 1. \qquad (8.7)$$

The distribution $P(x)$ is both centred at, and symmetric about, $x = 0$, so that the expectation of $(x - \langle x \rangle)^2$ is

$$\langle (x - \langle x \rangle)^2 \rangle = \omega^2, \qquad (8.8)$$

as shown earlier in Chap. 3.

8.3 Mean, median and mode

Repeated measurement yields a distribution of values for the quantity being measured. Consider two possible distributions of experimental data shown in Fig. 8.4, which displays the fraction of measurements that fall into each range of an observable x. In part (a) of Fig. 8.4, the observed values of x are spread relatively uniformly over a range of x: the likelihood that a measurement yields a specific value of x does not depend strongly on x. The situation in part (b) is different: some values of x are frequently obtained by the measurements, while other values of x are rarely

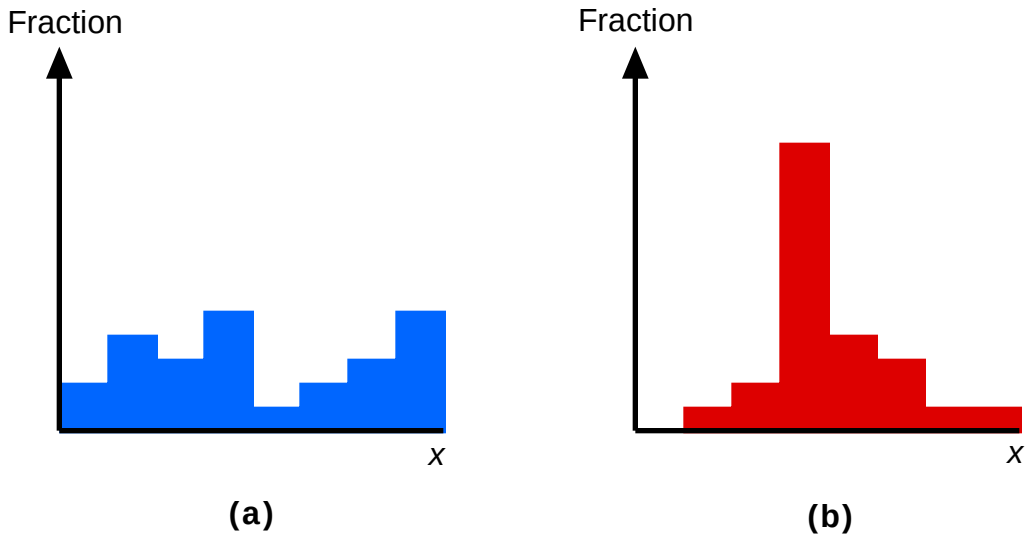

Fig. 8.4 Two different distributions of x, a measured observable. The data are collected into discrete ranges (or bins) of x, and the histograms show the fraction of measurements in which the value of x falls into a particular bin.

obtained. Both distributions are equally important in their own right, but we are often more interested in situations corresponding to part (b): we want to be able to *predict* that an observable will lie in a certain range under specified experimental conditions.

In this section, we deal with several aspects of data analysis:
•What is a good measure of the central value of a distribution?
•How can we determine whether two independently obtained data sets are consistent with each other?
•With what likelihood would a hypothetical description of the measurement produce the measured data?

There are several estimators of the *central value* of a distribution. One of the most common estimators is the mean, which is used extensively throughout these notes. Denoting the mean value of x by $\langle x \rangle$, then

$$\langle x \rangle = N^{-1} \sum_{j=1,N} x_j \qquad (8.9)$$

for a discrete distribution of N elements, labelled by x_i, or

$$\langle x \rangle = \int xP(x)dx \; / \; \int P(x)dx \qquad (8.10)$$

for a continuous distribution $P(x)$. The mean is the *first moment* of the distribution. While this is a useful estimator for distributions such as Fig. 8.4 (b), it is not the only estimator of the central value.

The median x_{med} is defined as the half-way point of the distribution:

$$x_{med} = \begin{cases} x_{(N+1)/2} & \text{if } N \text{ is odd} \\ & \text{(ordered from low to high)} \\ (x_{N/2} + x_{N/2+1}) / 2 & \text{if } N \text{ is even} \end{cases} \qquad (8.11)$$

for a discrete distribution with N elements, or

$$\frac{\int_{-\infty}^{x_{med}} P(x)dx}{\int_{-\infty}^{\infty} P(x)dx} = \frac{\int_{x_{med}}^{\infty} P(x)dx}{\int_{-\infty}^{\infty} P(x)dx} = 1/2 \qquad (8.12)$$

for a continuous distribution. For probability distributions, which are normalized to unity, the denominators in Eqs. (8.10) and (8.12) can be omitted.

The *mode* is the most likely value x in the distribution:

$$P(x_{\text{mode}}) = [\textit{maximum value of } P]. \tag{8.13}$$

These estimators vary in their usefulness for describing the central value. There may be situations, such as the distribution in Fig. 8.4 (a), where none of the estimators is useful in predicting the value of a measurement.

How do we characterize whether a measurement will yield a value of x close to the central value (be it $\langle x \rangle$, x_{med} or x_{mode})? That is, how do we characterize the *width* of the distributions in Fig. 8.4? A quantity that we introduced in other sections of these notes is the dispersion $(\Delta x)^2$, defined by

$$\langle (\Delta x)^2 \rangle = \int (x - \langle x \rangle)^2 P(x) dx \; / \; \int P(x) dx \tag{8.14}$$

for a continuous distribution. It is straightforward to show that Eq. (8.14) is equivalent to

$$\langle (\Delta x)^2 \rangle = \langle x^2 \rangle - \langle x \rangle^2. \tag{8.15}$$

We use the term *dispersion* for $\langle (\Delta x)^2 \rangle$ to avoid conflict with the statisticians' term *variance*, which has a slightly different normalization. For a discrete data set with N elements, the variance is

$$\text{Var}(x_1 \ldots x_N) \equiv \sigma^2 = (N-1)^{-1} \sum_{j=1,N} (x_j - \langle x \rangle)^2, \tag{8.16}$$

where σ is the standard deviation. Just as the mean is the first moment of the distribution, the dispersion and variance are the second moments. At large N, the dispersion and variance converge.

Clearly, if the dispersion or variance of a data set is large compared to $\langle x \rangle^2$, then the likelihood of successive measurements giving similar values of x is small.

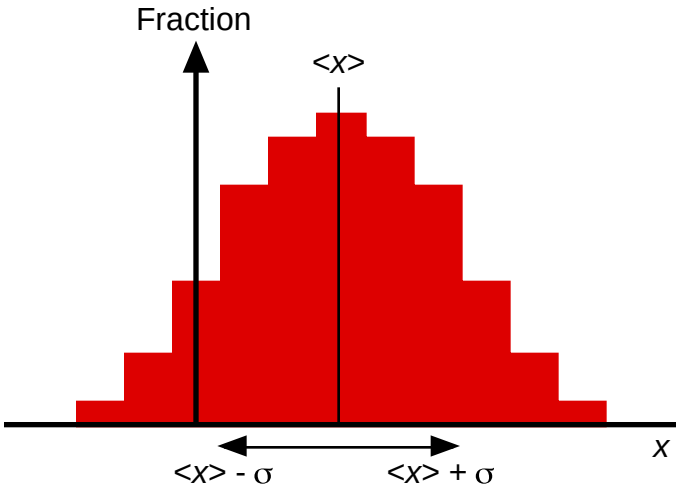

Fig. 8.5 In a Gaussian or normal distribution, 68.3% of the measured values of x lie within $\pm\sigma$ of the mean value $\langle x \rangle$.

8.4 Are data sets consistent?

The mean and dispersion have probabilistic significance. Often, the measured values of x in a large data set obey a normal or Gaussian distribution, illustrated schematically in Fig. 8.5. Integrating the corresponding probability distribution shows that the probability of x to be between $\langle x \rangle - \sigma$ and $\langle x \rangle + \sigma$ is 0.6827. This means that 68.3% of the measured values of x lie within σ of the mean value $\langle x \rangle$.

If the sample size is small, then both the mean and dispersion may vary significantly from one data set to another. But as the sample size becomes large, then $\langle x \rangle$ and σ should converge to their respective asymptotic values. In other words, the width σ of the distribution tends to a specific value as the number of data N in the sample increases; σ does not decrease to zero with increasing N. What does decrease with N is the *uncertainty* in $\langle x \rangle$ and the *uncertainty* in σ. The situation is illustrated in Fig. 8.6, which shows the mean values of three fictitious data sets. Both $\langle x \rangle$ and σ change with the samples, but become ever better known as the size of the total sample increases.

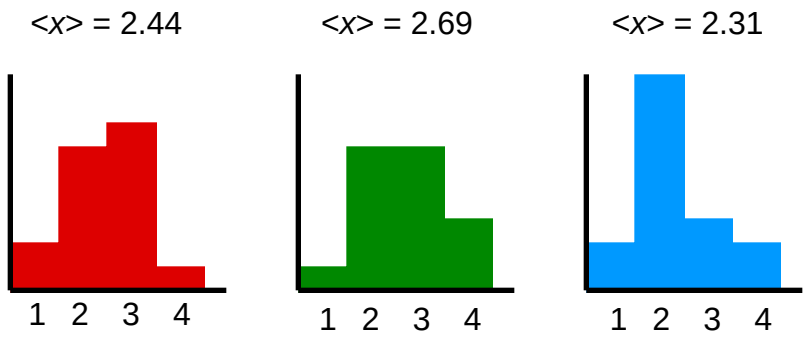

Fig. 8.6 Three independent data sets of the same observable have three values for <*x*> and σ. Each data set has the same number of data points.

It is not uncommon for the standard deviation to be confused with the error in the mean. While <*x*> ± σ may represent the range within which two-thirds of the data points lie, it does not mean that <*x*> is known only to within ±σ. In the data shown in Fig. 8.6, the standard deviation in each of the data sets is a little less than 1.0, but the change in <*x*> from one data set to the next is much less than 1.0. In fact, if all of the data points are combined, for a sample which is triple in size compared to any given individual sample in Fig. 8.6, the resulting mean is 2.48.

In the same way as one constructs the standard deviation of the distribution of *x*, one can construct the standard deviation of the distribution of <*x*>, which is called the *standard error* and which we denote by σ_e. For the data in Fig. 8.6,

$$\sigma_e^2 = [(2.44-2.48)^2 + (2.69-2.48)^2 + (2.31-2.48)^2] / 2$$
$$= 0.037, \qquad (8.17)$$

from which the standard error itself is

$$\sigma_e = 0.19. \qquad (8.18)$$

As one expects from Fig. 8.6, the values of <*x*> span a much narrower range than does the distribution in *x* itself, and the standard error of <*x*> is much smaller than the standard deviation of *x*.

If the distribution is Gaussian, then

$$\sigma_e^2 = \sigma^2 / N, \qquad (8.19)$$

where N is the number of data points in the sample from which $\langle x \rangle$ is obtained. Hence, if $\langle x \rangle$ and σ_e are determined from one data set, then 68.3% of the mean values determined from other data sets would lie between $\langle x \rangle - \sigma_e$ and $\langle x \rangle + \sigma_e$.

Returning to the question in the title of this section: how can we tell if two data sets are consistent? That is, how can we tell if there is a reasonable probability that two data sets with small N could have been chosen independently from the same probability distribution? To answer this question in the detail that it deserves takes us too far into the realm of statistics; as usual, we refer the reader to *Numerical Recipes* for further details. Without too much statistics, however, we can address two aspects of the consistency of distributions.

Mean values

The mean values of two data sets are probably consistent if they fall within one standard error σ_e of each other. Since one actually doesn't *prove* consistency with statistics, a better wording would be that the means values may be *in*consistent if they are separated by more than 2 standard errors. We are told very little about consistency if two mean values $\langle x \rangle$ fall within a *standard deviation* σ of each other: it is the *standard error* σ_e that is important for mean values.

χ^2 statistic

More meaningful tests of the consistency of data sets must involve the comparison of the data on a bin-by-bin basis. Two data sets could easily have a similar mean, and yet have a very different distribution. For example, a random distribution of numbers between zero and one has the same mean as a delta-function centered at $x = 1/2$. A measure of the sameness of distributions is the χ^2 (or chi-square) statistic, defined by

$$\chi^2 = \sum_i \frac{(M_i - E_i)^2}{E_i} \qquad (8.20)$$

The quantities M_i and E_i are numbers: M_i is the number of events *measured* in the i^{th} data bin, and E_i is the number of events *expected* in the i^{th} bin according to some known distribution. Clearly, any term with $M_i = E_i = 0$ should be omitted from Eq. (8.20). Chi-square is small if M_i is close to E_i in each bin. How small must χ^2 be before one could say that the data are consistent?

First, one must normalize χ^2 itself to reflect the fact that χ^2 increases with the number of data bins, for a given data set. To illustrate this, suppose that the total number of data points both measured and expected is the same; that is, $\Sigma M_i = \Sigma E_i$. Then, if all of the data were placed into just one bin, $\chi^2 = 0$! Obviously, this is not a very useful comparison; the fewer the number of data bins, the more information about the distribution has been discarded. However, the larger the number of bins, the more likely it is that M_i will not equal E_i in a given bin, and the larger χ^2 will be. Hence, one should divide χ^2 by a quantity that reflects the way in which the data have been binned and/or fitted. This relevant quantity is called the number of degrees of freedom ν and is given by

$$\nu = [\textit{number of data bins}, N_{bin}] - [\textit{number of free parameters used to predict the } E_i \textit{ distribution}] \quad (8.21)$$

For example, if the total number of expected events is adjusted to equal the total number of observed events (*i.e.*, $\Sigma E_i = \Sigma M_i$), then

$$[\textit{number of free parameters used to predict } E_i] = 1$$

and $\nu = N_{bin} - 1$.

The statistical properties of the chi-square probability function are known [see Press *et al*., (1992), Chaps. 6 and 14]. It is found that the measured distribution is not inconsistent with the expected distribution if $\chi^2 = \nu$. That is,

$$\chi^2 / \nu \sim 1 \quad (8.22)$$

if it is likely that the data agree with expectations. In the case of two data sets X_i and Y_i, the chi-square statistic becomes

$$\chi^2 = \sum_i \frac{(X_i - Y_i)^2}{X_i + Y_i} \quad (8.23)$$

if the number of data points has been adjusted to be the same in both samples (*i.e.,* $\Sigma X_i = \Sigma Y_i$, for which $\nu = N_B - 1$). See Press *et al*. (1992) for $\Sigma X_i \neq \Sigma Y_i$.

8.5 Data fitting

In physics, we often try to fit a data set with a model, perhaps motivated by a microscopic interpretation of the system being measured. At one level, we can ask whether the data are consistent with the model: given a set of predictions E_i for the elements in the data set, what is the probability that the measured values M_i would be found in an experiment? This question is addressed in Sec. 8.4 through the χ^2 statistic. Sometimes the model involves a number of parameters, and then the question becomes: what choice of parameters best represents the data.

As emphasized by Press *et al*. (1992), this question has subtle assumptions that should be clarified. A physicist may perform a measurement with several models in mind, and may wish to find out which model is the "most likely" one. Perhaps there are three models A, B, and C, and it is tempting to ask what is the probability of model A being correct, the probability of model B being correct, *etc*. Unfortunately, the question cannot be approached in this manner. While the model may be motivated by the measurement, in fact statistical deduction assumes that the proposed model is true, and asks: what is the likelihood that the data set could be obtained from the model? In other words, rather than assigning a probability to the model, we obtain a probability for the data set, assuming the model is true. If the probability of the data set being extracted from the model is low, then the model is discarded.

We introduce some symbols to address the issue in more mathematical terms. Suppose that the measurement consists of finding the size of an oilspill as a function of time. We wish to "fit the data" to obtain, say, the rate at which the oil spreads. We measure the area of the spill, which will be called y_i, at a sequence of times t_1, t_2, t_3, ...t_N. Thus, the measurement generates a data set

Time	Area
t_1	y_1
t_2	y_2
t_3	y_3
t_N	y_N

To interpret the data, we construct a model of the spill that gives a function $y(t)$ for the area as a function of time.

Each of the area measurements has an uncertainty, and we assume that the measured values obey a normal distribution around the "true" value, given by $y(t)$. For the time being, we assign the same standard deviation σ to each measurement. Thus, the probability P_i that a given measured value y_i lies in the range Δy around the "true" value is

$$P_i \quad \exp\left[-\frac{1}{2}\left(\frac{y_i - y(t_i)}{\sigma}\right)^2\right]\Delta y \tag{8.24}$$

In Eq. (8.24), σ is the standard deviation of the measured distribution for each value of y; it is not the standard error of the mean value $\langle y \rangle$. The probability that every element of the data set lies in the range Δy around the predicted value $y(t)$ is then

$$P = P_1\, P_2\, P_3\, ...P_N, \tag{8.25}$$

or

$$P \quad \prod_{i=1}^{N}\left\{\exp\left[-\frac{1}{2}\left(\frac{y_i - y(t_i)}{\sigma}\right)^2\right]\Delta y\right\} \tag{8.26}$$

The task, then, is to choose a parameter set that maximizes the likelihood P that the observed data could have been chosen from the model $y(x)$. Maximizing P is equivalent to maximizing its logarithm

$$\ln P = \sum_{i=1}^{N}\left[-\frac{1}{2}\left(\frac{y_i - y(t_i)}{\sigma}\right)^2 + \ln\Delta y\right] + \text{constants} \tag{8.27}$$

or to *minimizing* the negative of its logarithm

$$-\ln P = \sum_{i=1}^{N}\left[\frac{1}{2}\left(\frac{y_i - y(t_i)}{\sigma}\right)^2\right] - N\ln\Delta y \quad - \text{constants} \tag{8.28}$$

Since the last two terms in Eq. (8.28) involve only constants, then

$$[\textit{maximum of } P] \Leftrightarrow [\textit{minimum of } \sum_{i=1}^{N}\frac{[y_i - y(t_i)]^2}{2\sigma^2}\ ] \tag{8.29}$$

The right-hand side of Eq. (8.29) is also refered to as the *least-squares* fit, whose functional form looks similar to the form of χ^2 for discrete data distributions. Hence, we generalize the definition of chi-square to

$$\chi^2 = \sum_i \frac{[y_i - y(t_i)]^2}{\sigma_i^2} \tag{8.30}$$

where the comparisons are now between data that may have physical units. It is simple to obtain Eq. (8.20), which deals with numbers, from (8.30) by recognizing that the standard deviation of a distribution with n elements is just $\sqrt{n}$. Of course, in an experimental measurement, the width of the distribution σ_i may be affected by more that just statistical errors; there may be systematic uncertainties (perhaps associated with the apparatus or its operators) that are included in σ_i.

Fitting the data to a model then has three aspects:

1. The operation of minimizing χ^2.
2. The evaluation of χ^2 per degree of freedom to see if the fit is meaningful.
3. The assignment of uncertainties to the parameters extracted from the fit.

8.6 Linear regression

In the technique of *linear regression*, a linear function $y(x) = a + bx$ is used to represent the "true" data. The corresponding expression for χ^2 becomes

$$\chi^2 = \sum_i \frac{[y_i - a - bx_i]^2}{\sigma_i^2} \tag{8.31}$$

As a function, χ^2 depends on two parameters, a and b. Hence, two conditions must be fulfilled for χ^2 to be at a minimum, namely

$$0 = \frac{\partial \chi^2}{\partial a} = -2 \sum_i \frac{y_i - a - bx_i}{\sigma_i^2} \tag{8.32a}$$

$$0 = \frac{\partial \chi^2}{\partial b} = -2 \sum_i \frac{x_i (y_i - a - bx_i)}{\sigma_i^2} \tag{8.32b}$$

The evaluation of these two equations can be simplified by throwing away the factor of -2 and writing out the sums explicitly:

$$0 = \sum_i y_i/\sigma_i^2 - \sum_i a/\sigma_i^2 - \sum_i bx_i/\sigma_i^2 \qquad (8.33a)$$

$$0 = \sum_i x_i y_i/\sigma_i^2 - \sum_i ax_i/\sigma_i^2 - \sum_i bx_i^2/\sigma_i^2. \qquad (8.33b)$$

Eqs. (8.33) involve 5 independent sums

$$S \equiv \sum_i 1/\sigma_i^2 \qquad S_x \equiv \sum_i x_i/\sigma_i^2 \qquad S_y \equiv \sum_i y_i/\sigma_i^2$$
$$S_{xx} \equiv \sum_i x_i^2/\sigma_i^2 \qquad S_{xy} \equiv \sum_i x_i y_i/\sigma_i^2,$$

which can be used to recast the equations in a more familiar form

$$aS + bS_x = S_y \qquad (8.34a)$$

$$aS_x + bS_{xx} = S_{xy}. \qquad (8.34b)$$

The two unknowns in Eq. (8.34) are a and b, which have the solution

$$a = (S_{xx}S_y - S_xS_{xy}) / \Delta \qquad (8.35a)$$

$$b = (SS_{xy} - S_xS_y) / \Delta \qquad (8.35b)$$

where

$$\Delta = SS_{xx} - S_x^2. \qquad (8.36)$$

The minimum chi-square solutions for a and b are written out explicitly in Eqs. (8.35) and (8.36). But a and b are not exact; they represent a "best fit" of the data, which have uncertainties. The values of a and b also have standard deviations, which we denote by σ_a and σ_b [see Press *et al*. (1992)]:

$$\sigma_a^2 = S_{xx} / \Delta \qquad (8.37a)$$

$$\sigma_b^2 = S / \Delta. \qquad (8.37b)$$

Lastly, once all of the parameters and their uncertainties have been found, we ask: does the fit have any meaning? To answer this, evaluate chi-square per degree of freedom χ^2/ν for the fit, and compare the answer to 1.

8.7 Summary of notation

A rather large number of symbols have been defined in this section, and we summarize them here for the convenience of the reader. Here, *N* is the number of data points in the sample.

Measures of centrality

$\langle x \rangle$ = mean $\quad \langle x \rangle = N^{-1} \sum_{j=1,N} x_j$ (8.9)

x_{med} = median $\quad x_{(N+1)/2}$ or $(x_{N/2} + x_{N/2+1}) / 2$ (ordered list) (8.11)

x_{mode} = mode $\quad P(x_{mode}) =$ [*maximum value of P*]. (8.13)

Width of distributions

$\langle (\Delta x)^2 \rangle$ = dispersion $\quad \langle (\Delta x)^2 \rangle = \langle (x - \langle x \rangle)^2 \rangle$ (8.14)

σ = standard deviation $\quad \sigma^2 = (N-1)^{-1} \sum_{j=1,N} (x_j - \langle x \rangle)^2$ (8.16)

Uncertainties and data fitting

σ_e = standard error $\quad \sigma_e^2 = \sigma^2 / N$ (for normal distribution) (8.19)

χ^2 = chi-square $\quad \chi^2 = \sum_i \dfrac{[y_i - y(t_i)]^2}{\sigma_i^2}$ (8.30)

σ_i = standard deviation of data point *i*

σ_a , σ_b = standard deviation of linear regression parameters *a* and *b* (8.37)

References

Particle Data Group, "Review of Particle Physics", *Phys. Rev. D* 54-I:1-720.

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery *Numerical Recipes in C*, 2nd edition (Cambridge, 1992) .

8.8 Project 8 - Analysis of random numbers

The repeated measurement of an observable generates a data set whose elements are spread over a range of values. In this project, we use a random walk to provide an observable, namely the end-to-end displacement r_{ee}, from which we can generate a data set with known characteristics. We analyse both the distribution of r_{ee} in an individual data set, and the distribution of $R \equiv \langle r_{ee} \rangle$ found from several data sets.

Physical system

The end-to-end displacement of a random walk in one dimension has a Gaussian form at large number of steps given by Eq. (8.5)

$$P(x)dx = (2\pi\omega^2)^{-1/2} \exp(-x^2/2\omega^2)\, dx, \qquad (8.38)$$

where $P(x)$ is a probability density (probability per unit length) and x is introduced as shorthand for r_{ee}. The probability of finding a value of x between x and $x + dx$ is then $P(x)dx$. The width of the distribution (which is also its standard deviation) is ω, where

$$\omega^2 = N_{step} a^2 \qquad (8.39)$$

for one-dimensional walks with a fixed number of steps N_{step} [see Eq. (8.6) for walks in higher dimensions]. In this project, we are interested not only in the width of the distribution, but also in the standard deviation and standard error of the width as extracted from a sample of walks.

Simulation parameters

The mean value of r_{ee} for all chains, large and small, should be zero by symmetry, since the walk can end on either side of the origin. However, the mean

value of r_{ee}^2 does not vanish, and increases with N_{step}. Thus, if we average over sufficiently many walks that the distribution of r_{ee} is fairly smooth, then we expect that the distribution for large N_{step} will be wider than the distribution for small N_{step}. The effect is shown schematically in Fig. 8.7. While the schematic data in Fig. 8.7 are purposefully symmetric about $r_{ee} = 0$, this situation would be highly unusual in a data set. Thus, although $R \equiv \langle r_{ee} \rangle = 0$ in Fig. 8.7, in practice $R \neq 0$. For a given N_{step}, adding more samples to the data set does not change the distribution of r_{ee} appreciably, although it does bring the value of R ever closer to zero.

For a given sample, then, we expect $R \neq 0$. We can repeat the determination of R with another sample, and obtain another value of R, also not equal to zero. By generating one sample after another, a distribution in R can be built up, as shown in Fig. 8.8. A value for $\langle R \rangle$ can be extracted from the distribution in R, and this value comes closer to zero as the number of sample sets increases.

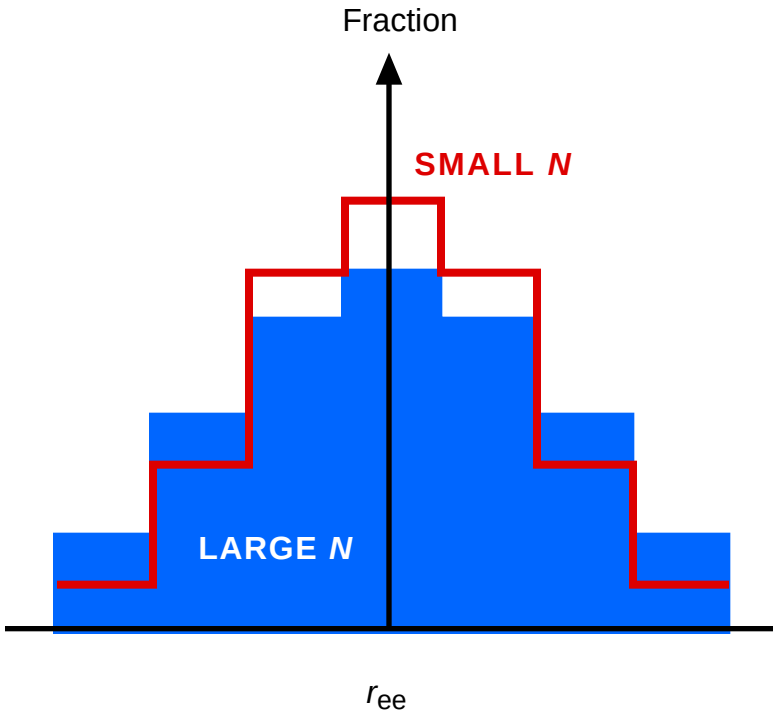

Fig. 8.7 The shape of the distribution in r_{ee} depends upon the number of steps N_{step} in the walk. The solid blue distribution is for large N_{step}, while the red line is the distribution for small N_{step}. Larger walks have a broader distribution: the width grows like $N_{step}^{1/2}$.

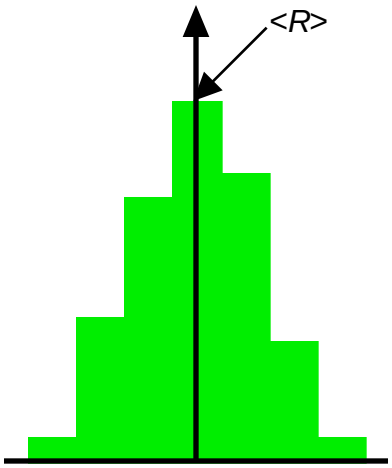

Fig. 8.8 Distribution of $R \equiv \langle r_{ee} \rangle$ obtained from multiple data sets. Each distribution of r_{ee} from Fig. 8.7 produces *one* value of R. As the number of sample sets increases, the mean value of R becomes known ever more accurately.

Code

The code required to generate the random walk data for analysis purposes is *very* short.

1. Perform a random walk in one dimension by throwing a random number to determine whether to walk to the left or right at each step. Start the walk at the origin, and make each step of unit length. The values of r_{ee} that you obtain will either be all odd, or all even, according to the value of N_{step}.

2. Allow the number of steps in the walk, N_{step}, to be a variable. The code executes so fast that the walk size and the sample size can range into the tens of thousands - although running such large samples will not help us understand error analysis.

Analysis

The notation in this project tends to be a little confusing, as one ends up finding dispersions of standard deviations *etc*. To aid the reader, we summarize some of the definitions:

Parameters of the walks:
N_{step} = number of steps in the walk
N_{walk} = number of walks in a sample
N_{sample} = number of samples (each with N_{walk})

Parameters of the distributions:
$R = \langle r_{ee} \rangle$ = the mean value of r_{ee} in a single sample of N_{walk} walks at fixed N_{step}
$Q = \langle r_{ee}^2 \rangle$ = the mean value of r_{ee}^2 in a single sample of N_{walk} walks at fixed N_{step}
σ_{ree}, σ_{ree2} = standard deviation of the distributions of r_{ee} and r_{ee}^2 in a single sample with N_{walk} elements
$\omega = N_{step}^{1/2} a$ = theoretically expected standard deviation (Q) for a distribution of walks at fixed N_{step}
$\langle R \rangle$, $\langle Q \rangle$ = mean values of R and Q obtained from several samples with N_{sample} elements; for large N_{sample}, $\langle R \rangle \rightarrow 0$ and $\langle Q \rangle \rightarrow \omega^2$
σ_R, σ_Q = standard deviation of the distributions of R and Q

DO NOT USE CANNED ROUTINES FOR THE FOLLOWING ANALYSIS; WRITE YOUR OWN CODES!

A single "experiment"

1. This experiment uses:
   - five different values of N_{step} (20, 30, 40, 60, 100)
   - 200 walks for each value of N_{step}: that is, $N_{walk} = 200$
   - only one sample at each N_{step}, comprising 200 walks; $N_{sample} = 1$.

Here, we analyse the properties of a random walk.

2. Choose a value of N_{step} from the set (20, 30, 40, 60, 100), and determine r_{ee} and r_{ee}^2 for one walk.

3. Construct ensemble averages $R \equiv \langle r_{ee} \rangle$ and $Q \equiv \langle r_{ee}^2 \rangle$ as well as their standard deviations σ_{ree} and σ_{ree2} for each N_{step} by generating and averaging over 200 walks (*i.e.*, $N_{walk} = 200$ for each N_{step}). Check that $R \sim 0$ and $Q \sim \omega^2$.

4. Repeat steps 2 and 3 for all five values of N_{step}.

5. Using the data from Steps 2-4, verify that R does not depend on N_{step} and find the N_{step}-dependence of Q. Make a linear fit to $\ln Q$ *vs.* $\ln N_{step}$ to find the power-law behavior (for the weights, use the standard deviation: σ_{ree2} / Q). Determine statistical measures such as chi-square, and the standard deviation of the power-law exponent.

Comparison of many "experiments"

1. This experiment uses:
   - one value of N_{step} (40)
   - 200 walks in each sample; $N_{walk} = 200$
   - fifty samples, each comprising 200 walks; $N_{sample} = 50$.

Here, we compare the results of several different experiments for consistency.

2. For a fixed N_{step} (40) and sample size N_{walk} (200), generate $N_{sample} = 50$ sample values of R and Q.

3. Determine the mean values $\langle R \rangle$ and $\langle Q \rangle$ as well as σ_R and σ_Q, by considering data sets with $N_{sample} = 10$, 20 and 50 samples.
You expect: $\langle R \rangle \rightarrow 0$ and $\langle Q \rangle \rightarrow \omega^2$. What do you expect for σ_R and σ_Q as a function of N_{sample}?

4. Using the data from Step 2, calculate the mean and median of the R and Q distributions for $N_{sample} = 50$. Do these measures agree with one another to within the *standard errors* of R and Q?

5. Divide the data from Step 3 into 2 parts, with 25 samples each. Find the mean $\langle Q \rangle$ and its standard deviation and standard error for both data sets. Do the two values of $\langle Q \rangle$ agree to within their standard error?

Report
Your report should include the following elements:
- a very brief paragraph of the properties of the random walk
- an outline of your code
- the extensive analysis as described above for the single and multiple experiments
- a copy of your code.

Demonstration code

The demonstration code shows the distribution of r_{ee} generated for a random walk of constant step length in one dimension. The code has a set-up page that allows you to select the number of steps in the walk, and the number of walks in the sample. From the shape and magnitude of the distributions, you can see the following:

1. The shape is independent of sample size, for a fixed N_{step}. Choose a specific N_{step} from the menu and run the code at different sample sizes. Obviously the smallest samples produce a somewhat ragged distribution, but one can see that the general shape and magnitude of the distribution do not change.

2. The width increases with the number of steps in the walk. It is best to run this simulation with the largest sample size in the menu, so that the distributions are smooth. By varying the number of steps N_{step} in the walk, one can see that

•the width of the walk increases with N_{step}

•the fraction of walks with $r_{ee} = 0$ decreases with N_{step}.