Chapter 4

Statistical Analysis of Random Uncertainties

We have seen that one of the best ways to assess the reliability of a measurement is to repeat it several times and examine the different values obtained. In this chapter and Chapter 5, I describe statistical methods for analyzing measurements in this way.

As noted before, not all types of experimental uncertainty can be assessed by statistical analysis based on repeated measurements. For this reason, uncertainties are classified into two groups: the *random* uncertainties, which *can* be treated statistically, and the *systematic* uncertainties, which *cannot.* This distinction is described in Section 4.1. Most of the remainder of this chapter is devoted to random uncertainties. Section 4.2 introduces, without formal justification, two important definitions related to a series of measured values $x_1, \ldots, x_N$, all of some single quantity x. First, I define the *average* or *mean* $\bar{x}$ of $x_1, \ldots, x_N$. Under suitable conditions, $\bar{x}$ is the best estimate of x based on the measured values $x_1, \ldots, x_N$. I then define the *standard deviation* of $x_1, \ldots, x_N$, which is denoted σ_x and characterizes the average uncertainty in the separate measured values $x_1, \ldots, x_N$. Section 4.3 gives an example of the use of the standard deviation.

Section 4.4 introduces the important notion of the *standard deviation of the mean.* This parameter is denoted $\sigma_{\bar{x}}$ and characterizes the uncertainty in the mean $\bar{x}$ as the best estimate for x. Section 4.5 gives examples of the standard deviation of the mean. Finally, in Section 4.6, I return to the vexing problem of systematic errors.

Nowhere in this chapter do I attempt a complete justification of the methods described. The main aim is to introduce the basic formulas and describe how they are used. In Chapter 5, I give proper justifications, based on the important idea of the normal distribution curve.

The relation of the material of this chapter (statistical analysis) to the material of Chapter 3 (error propagation) deserves mention. From a practical point of view, these two topics can be viewed as separate, though related, branches of error analysis (somewhat as algebra and geometry are separate, though related, branches of mathematics). Both topics need to be mastered, because most experiments require the use of both.

In a few kinds of experiments, the roles of error propagation and of statistical analysis are complementary. That is, the experiment can be analyzed using either

error propagation or statistical methods. Consider an example: Suppose you decide to measure the acceleration of gravity, g, by measuring the period, T, and the length, l, of a simple pendulum. Since $T = 2\pi\sqrt{l/g}$, you can find g as $g = 4\pi^2 l/T^2$. You might decide to repeat this experiment using several different values of l and measuring the corresponding period T for each. In this way, you would arrive at several values for g. To find the uncertainty in these values of g, you could proceed in either of two ways. If you can estimate realistically the uncertainties in your measurements of l and T, you could propagate these uncertainties to find the uncertainties in your values of g. Alternatively, given your several values of g, you could analyze them statistically; in particular, their *standard deviation* will be a good measure of their uncertainty. Unfortunately, you do not truly have a choice of how to find the uncertainty. If the uncertainty can be found in these two ways, you really ought to do so *both* ways to check that they do give, at least approximately, the same answer.

4.1 Random and Systematic Errors

Experimental uncertainties that can be revealed by repeating the measurements are called *random* errors; those that cannot be revealed in this way are called *systematic.* To illustrate this distinction, let us consider some examples. Suppose first that we time a revolution of a steadily rotating turntable. One source of error will be our reaction time in starting and stopping the watch. If our reaction time were always exactly the same, these two delays would cancel one another. In practice, however, our reaction time will vary. We may delay more in starting, and so underestimate the time of a revolution; or we may delay more in stopping, and so overestimate the time. Since either possibility is equally likely, the sign of the effect is *random.* If we repeat the measurement several times, we will sometimes overestimate and sometimes underestimate. Thus, our variable reaction time will show up as a variation of the answers found. By analyzing the spread in results statistically, we can get a very reliable estimate of this kind of error.

On the other hand, if our stopwatch is running consistently slow, then all our times will be underestimates, and no amount of repetition (with the same watch) will reveal this source of error. This kind of error is called *systematic,* because it always pushes our result in the same direction. (If the watch runs slow, we always underestimate; if the watch runs fast, we always overestimate.) Systematic errors cannot be discovered by the kind of statistical analysis contemplated here.

As a second example of random versus systematic errors, suppose we have to measure some well-defined length with a ruler. One source of uncertainty will be the need to interpolate between scale markings; and this uncertainty is probably random. (When interpolating, we are probably just as likely to overestimate as to underestimate.) But there is also the possibility that our ruler has become distorted; and this source of uncertainty would probably be systematic. (If the ruler has stretched, we always underestimate; if it has shrunk, we always overestimate.)

Just as in these two examples, almost all measurements are subject to both random and systematic uncertainties. You should have no difficulty finding more examples. In particular, notice that common sources of random uncertainties are

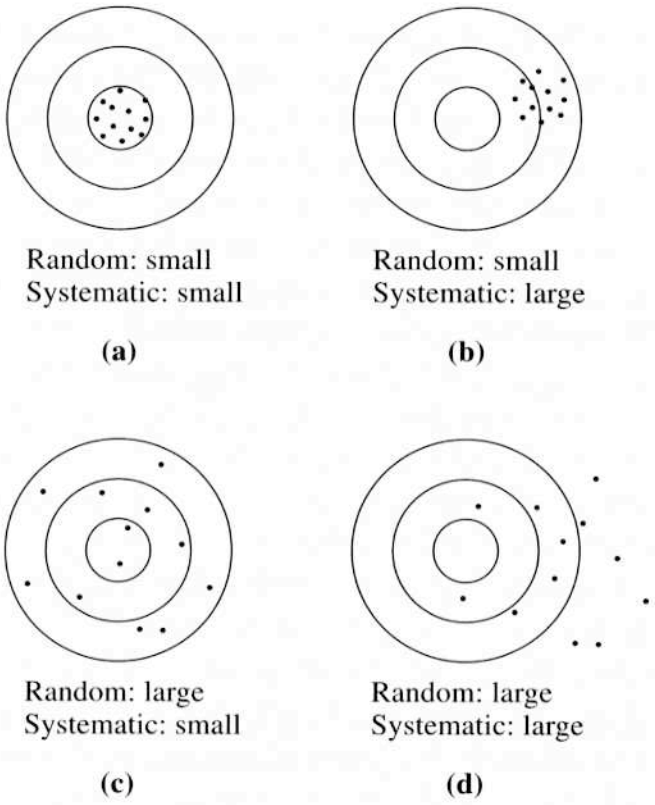

Figure 4.1. Random and systematic errors in target practice. **(a)** Because all shots arrived close to one another, we can tell the random errors are small. Because the distribution of shots is centered on the center of the target, the systematic errors are also small. **(b)** The random errors are still small, but the systematic ones are much larger—the shots are "systematically" off-center toward the right. **(c)** Here, the random errors are large, but the systematic ones are small—the shots are widely scattered but not systematically off-center. **(d)** Here, both random and systematic errors are large.

small errors of judgment by the observer (as when interpolating), small disturbances of the apparatus (such as mechanical vibrations), problems of definition, and several others. Perhaps the most obvious cause of systematic error is the miscalibration of instruments, such as the watch that runs slow, the ruler that has been stretched, or a meter that is improperly zeroed.

To get a better feel for the difference between random and systematic errors, consider the analogy shown in Figure 4.1. Here the "experiment" is a series of shots fired at a target; accurate "measurements" are shots that arrive close to the center. Random errors are caused by anything that makes the shots arrive at randomly different points. For example, the marksman may have an unsteady hand, or fluctuating atmospheric conditions between the marksman and the target may distort the view of the target in a random way. Systematic errors arise if anything makes the shots arrive off-center in one "systematic" direction, for instance, if the gun's sights are misaligned. Note from Figure 4.1 how the results change according to the various combinations of small or large random or systematic errors.

Although Figure 4.1 is an excellent illustration of the effects of random and systematic errors, it is, nonetheless, misleading in one important respect. Because

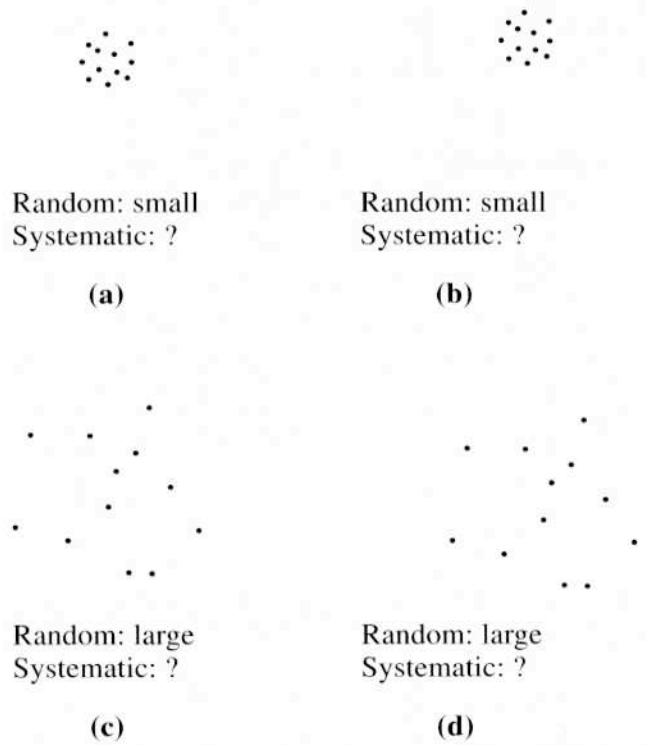

Figure 4.2. The same experiment as in Figure 4.1 redrawn without showing the position of the target. This situation corresponds closely to the one in most real experiments, in which we do not know the true value of the quantity being measured. Here, we can still assess the random errors easily but cannot tell anything about the systematic ones.

each of the four pictures shows the position of the target, we can tell at a glance whether a particular shot was accurate or not. In particular, the difference between the top two pictures is immediately evident. The shots in the left picture cluster around the target's center, whereas those in the right picture cluster around a point well off-center; clearly, therefore, the marksman responsible for the left picture had little systematic error, but the one responsible for the right picture had a lot more. Knowing the position of the target in Figure 4.1 corresponds, in a laboratory measurement, to knowing the true value of the measured quantity, and in the vast majority of real measurements, we do *not* know this true value. (If we knew the true value, we would usually not bother to measure it.)

To improve the analogy of Figure 4.1 with most real experiments, we need to redraw it without the rings that show the position of the target, as in Figure 4.2. In these pictures, identifying the random errors is still easy. (The top two pictures still obviously have smaller random errors than the bottom two.) Determining which marksman had larger systematic errors, however, is *impossible* based on Figure 4.2. This situation is exactly what prevails in most real experiments; by examining the distribution of measured values, we can easily assess the random errors but get no guidance concerning the systematic errors.

The distinction between random and systematic errors is not always clear-cut, and a problem that causes random errors in one experiment may produce systematic errors in another. For example, if you position your head first to one side and then to another to read a typical meter (such as an ordinary clock), the reading on the meter changes. This effect, called *parallax,* means that a meter can be read correctly only if you position yourself directly in front of it. No matter how careful you are, you cannot always position your eye *exactly* in front of the meter; consequently, your measurements will have a small uncertainty due to parallax, and this uncertainty will probably be random. On the other hand, a careless experimenter who places a meter to one side of his seat and forgets to worry about parallax will introduce a systematic error into all his readings. Thus, the same effect, parallax, can produce random uncertainties in one case, and systematic uncertainties in another.

The treatment of random errors is different from that of systematic errors. The statistical methods described in the following sections give a reliable estimate of the random uncertainties, and, as we shall see, provide a well-defined procedure for reducing them. For the reasons just discussed, systematic uncertainties are usually hard to evaluate and even to detect. The experienced scientist has to learn to anticipate the possible sources of systematic error and to make sure that all systematic errors are much less than the required precision. Doing so will involve, for example, checking the meters against accepted standards and correcting them or buying better ones if necessary. Unfortunately, in the first-year physics laboratory, such checks are rarely possible, so the treatment of systematic errors is often awkward. This concept is discussed further in Section 4.6. For now, I will discuss experiments in which all sources of systematic error have been identified and made much smaller than the required precision.

4.2 The Mean and Standard Deviation

Suppose we need to measure some quantity x, and we have identified all sources of systematic error and reduced them to a negligible level. Because all remaining sources of uncertainty are random, we should be able to detect them by repeating the measurement several times. We might, for example, make the measurement five times and find the results

$$71,\ 72,\ 72,\ 73,\ 71 \tag{4.1}$$

(where, for convenience, we have omitted any units).

The first question we address is this: Given the five measured values (4.1), what should we take for our best estimate x_{best} of the quantity x? Reasonably, our best estimate would seem to be the *average* or *mean* $\bar{x}$ of the five values found, and in Chapter 5, I will prove that this choice is normally best. Thus,

$$\begin{aligned} x_{\text{best}} &= \bar{x} \\ &= \frac{71 + 72 + 72 + 73 + 71}{5} \\ &= 71.8. \end{aligned} \tag{4.2}$$

Here, the second line is simply the definition of the mean $\bar{x}$ for the numbers at hand.[1]

More generally, suppose we make N measurements of the quantity x (all using the same equipment and procedures) and find the N values

$$x_1, x_2, \ldots, x_N. \tag{4.3}$$

Once again, the best estimate for x is usually the average of $x_1, \ldots, x_N$. That is,

$$x_{\text{best}} = \bar{x}, \tag{4.4}$$

where

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_N}{N} = \frac{\sum x_i}{N}. \tag{4.5}$$

In the last line, I have introduced the useful sigma notation, according to which

$$\sum_{i=1}^{N} x_i = \sum_{i} x_i = \sum x_i = x_1 + x_2 + \cdots + x_N;$$

the second and third expressions here are common abbreviations, which I will use when there is no danger of confusion.

The concept of the average or mean is almost certainly familiar to most readers. Our next concept, that of the *standard deviation,* is probably less so. The standard deviation of the measurements $x_1, \ldots, x_N$ is an estimate of the *average uncertainty of the measurements* $x_1, \ldots, x_N$ and is determined as follows.

Given that the mean $\bar{x}$ is our best estimate of the quantity x, it is natural to consider the difference $x_i - \bar{x} = d_i$. This difference, often called the *deviation* (or residual) of x_i from $\bar{x}$, tells us *how much the* i^{th} *measurement* x_i *differs from the average* $\bar{x}$. If the deviations $d_i = x_i - \bar{x}$ are all very small, our measurements are all close together and presumably very precise. If some of the deviations are large, our measurements are obviously not so precise.

To be sure you understand the idea of the deviation, let us calculate the deviations for the set of five measurements reported in (4.1). These deviations can be listed as shown in Table 4.1. Notice that the deviations are not (of course) all the same size; d_i is small if the ith measurement x_i happens to be close to $\bar{x}$, but d_i is large if x_i is far from $\bar{x}$. Notice also that some of the d_i are positive and some negative because some of the x_i are bound to be higher than the average $\bar{x}$, and some are bound to be lower.

To estimate the average reliability of the measurements $x_1, \ldots, x_5$, we might naturally try averaging the deviations d_i. Unfortunately, as a glance at Table 4.1 shows, the average of the deviations is zero. In fact, this average will be zero for

[1] In this age of pocket calculators, it is worth pointing out that an average such as (4.2) is easily calculated in your head. Because all the numbers are in the seventies, the same must be true of the average. All that remains is to average the numbers 1, 2, 2, 3, 1 in the units place. These numbers obviously average to $9/5 = 1.8$, and our answer is $\bar{x} = 71.8$.

Table 4.1. Calculation of deviations.

Trial number i	Measured value x_i	Deviation $d_i = x_i - \bar{x}$
1	71	−0.8
2	72	0.2
3	72	0.2
4	73	1.2
5	71	−0.8
	$\Sigma x_i = 359$	$\Sigma d_i = 0.0$
	mean, $\bar{x} = \Sigma x_i/N = 359/5 = 71.8$	

any set of measurements $x_1, \ldots, x_N$ because the definition of the average $\bar{x}$ ensures that $d_i = x_i - \bar{x}$ is sometimes positive and sometimes negative in just such a way that $\bar{d}$ is zero (see Problem 4.4). Obviously, then, the average of the deviations is not a useful way to characterize the reliability of the measurements $x_1, \ldots, x_N$.

The best way to avoid this annoyance is to *square* all the deviations, which will create a set of *positive* numbers, and then average these numbers.[2] If we then take the square root of the result, we obtain a quantity with the same units as x itself. This number is called the *standard deviation* of $x_1, \ldots, x_N$, and is denoted σ_x:

$$\sigma_x = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(d_i)^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2}. \tag{4.6}$$

With this definition, the standard deviation can be described as the *root mean square* (or RMS) deviation of the measurements $x_1, \ldots, x_N$. It proves to be a useful way to characterize the reliability of the measurements. [As we will discuss shortly, the definition (4.6) is sometimes modified by replacing the denominator N by $N - 1$.]

To calculate the standard deviation σ_x as defined by (4.6), we must compute the deviations d_i, square them, average these squares, and then take the square root of the result. For the data of Table 4.1, we start this calculation in Table 4.2.

Table 4.2. Calculation of the standard deviation.

Trial number i	Measured value x_i	Deviation $d_i = x_i - \bar{x}$	Deviation squared d_i^2
1	71	−0.8	0.64
2	72	0.2	0.04
3	72	0.2	0.04
4	73	1.2	1.44
5	71	−0.8	0.64
	$\Sigma x_i = 359$	$\Sigma d_i = 0.0$	$\Sigma d_i^2 = 2.80$
		$\bar{x} = 359/5 = 71.8$	

[2] Another possibility would be to take the absolute values $|d_i|$ and average them, but the average of the d_i^2 proves more useful. The average of the $|d_i|$ is sometimes (misleadingly) called the *average deviation*.

Summing the numbers d_i^2 in the fourth column of Table 4.2 and dividing by 5, we obtain the quantity σ_x^2 (often called the *variance* of the measurements),

$$\sigma_x^2 = \frac{1}{N}\sum d_i^2 = \frac{2.80}{5} = 0.56. \tag{4.7}$$

Taking the square root, we find the standard deviation

$$\sigma_x \approx 0.7. \tag{4.8}$$

Thus the average uncertainty of the five measurements 71, 72, 72, 73, 71 is approximately 0.7.

Unfortunately, the standard deviation has an alternative definition. There are theoretical arguments for replacing the factor N in (4.6) by $(N - 1)$ and defining the standard deviation σ_x of $x_1, \ldots, x_N$ as

$$\sigma_x = \sqrt{\frac{1}{N-1}\sum d_i^2} = \sqrt{\frac{1}{N-1}\sum (x_i - \bar{x})^2}. \tag{4.9}$$

I will not try here to prove that definition (4.9) of σ_x is better than (4.6), except to say that the new "improved" definition is obviously a little larger than the old one (4.6) and that (4.9) corrects a tendency for (4.6) to understate the uncertainty in the measurements $x_1, \ldots, x_N$, especially if the number of measurements N is small. This tendency can be understood by considering the extreme (and absurd) case that $N = 1$ (that is, we make only one measurement). Here, the average $\bar{x}$ is equal to our one reading x_1, and the one deviation is automatically zero. Therefore, the definition (4.6) gives the absurd result $\sigma_x = 0$. On the other hand, the definition (4.9) gives 0/0; that is, with definition (4.9), σ_x is undefined, which correctly reflects our total ignorance of the uncertainty after just one measurement. The definition (4.6) is sometimes called the *population standard deviation* and (4.9) the *sample standard deviation.*

The difference between the two definitions (4.6) and (4.9) is almost always numerically insignificant. You should always repeat a measurement many times (at least five, and preferably many more). Even if you make only five measurements ($N = 5$), the difference between $\sqrt{N} = 2.2$ and $\sqrt{N-1} = 2$ is, for most purposes, insignificant. For example, if we recalculate the standard deviation (4.8) using the improved definition (4.9), we obtain $\sigma_x = 0.8$ instead of $\sigma_x = 0.7$, not a very important difference. Nevertheless, you need to be aware of both definitions. In the physics laboratory, using the more conservative (that is, larger) definition (4.9) is almost always best, but in any case, your laboratory report should state clearly which definition you are using so that your readers can check the calculations for themselves.

Quick Check 4.1. You measure the time for a cart to roll down the same length of track four times and get the following results:

21, 24, 25, 22

(in seconds). Find the average time and the standard deviation as given by the improved definition (4.9).

To understand the notion of the standard deviation, you must be able to calculate it yourself for simple cases such as that in Quick Check 4.1. Most scientific calculators, however, have a built-in function to do the calculation automatically, and you will certainly want to use this function for real experiments that involve numerous measurements. If you are not sure how to use your calculator to obtain standard deviations, take the time to learn, and then use the function to check your answer to Quick Check 4.1. Some calculators give you a choice of the definitions (4.6) or (4.9); some use just (4.9). Make sure you know what yours does.

4.3 The Standard Deviation as the Uncertainty in a Single Measurement

Recall the claim that the standard deviation σ_x characterizes the average uncertainty of the measurements $x_1, \ldots, x_N$ from which it was calculated. In Chapter 5, I will justify this claim by proving the following more precise statement. If you measure the same quantity x many times, always using the same method, and if all your sources of uncertainty are small and random, then your results will be distributed around the true value x_{true} in accordance with the so-called normal, or bell-shaped, curve. In particular, *approximately 68% of your results*[3] *will fall within a distance* σ_x *on either side of* x_{true}; that is, 68% of your measurements will fall in the range $x_{\text{true}} \pm \sigma_x$.

In other words, if you make a *single* measurement (using the same method), the *probability* is 68% that your result will be within σ_x of the correct value. Thus, we can adopt σ_x to mean exactly what we have been calling "uncertainty." If you make one measurement of x, the uncertainty associated with this measurement can be taken to be

$$\delta x = \sigma_x;$$

with this choice, you can be 68% confident that the measurement is within δx of the correct answer.

To illustrate the application of these ideas, suppose we are given a box of similar springs and told to measure their spring constants k. We might measure the spring constants by loading each spring and observing the resulting extension or, perhaps better, by suspending a mass from each spring and timing its oscillations. Whatever method we choose, we need to know k and its uncertainty δk for each spring, but it would be hopelessly time-consuming to repeat our measurements many times for each spring. Instead we reason as follows: If we measure k for the first

[3] As we will see, the exact number is 68.27 . . . %, but stating this kind of number so precisely is obviously absurd. In fact, it is often best to think of this number as "about two thirds."

spring several (say, 5 or 10) times, then the mean of these measurements should give a good estimate of k for the first spring. More important for now, the standard deviation σ_k of these 5 or 10 measurements provides us with an estimate of the uncertainty in our method for measuring k. Provided our springs are all reasonably similar and we use the same method to measure each one, we can reasonably expect the same uncertainty in each measurement.[4] Thus, for each subsequent spring we need to make only one measurement, and we can immediately state that the uncertainty δk is the standard deviation σ_k measured for the first spring, with a 68% confidence that our answer is within σ_k of the correct value.

To illustrate these ideas numerically, we can imagine making 10 measurements on the first spring and obtaining the following measured values of k (in newtons/meter):

$$86, 85, 84, 89, 85, 89, 87, 85, 82, 85. \tag{4.10}$$

From these values, we can immediately calculate $\bar{k} = 85.7$ N/m and, using the definition (4.9),

$$\sigma_k = 2.16 \text{ N/m} \tag{4.11}$$

$$\approx 2 \text{ N/m}. \tag{4.12}$$

The uncertainty in any one measurement of k is therefore approximately 2 N/m. If we now measure the second spring once and obtain the answer $k = 71$ N/m, we can without further ado take $\delta k = \sigma_k = 2$ N/m and state with 68% confidence that k lies in the range

$$(k \text{ for second spring}) = 71 \pm 2 \text{ N/m}. \tag{4.13}$$

4.4 The Standard Deviation of the Mean

If $x_1, \ldots, x_N$ are the results of N measurements of the same quantity x, then, as we have seen, our best estimate for the quantity x is their mean $\bar{x}$. We have also seen that the standard deviation σ_x characterizes the average uncertainty of the separate measurements $x_1, \ldots, x_N$. Our answer $x_{\text{best}} = \bar{x}$, however, represents a judicious combination of all N measurements, and we have every reason to think it will be more reliable than any one of the measurements taken alone. In Chapter 5, I will prove that the uncertainty in the final answer $x_{\text{best}} = \bar{x}$ is given by the standard deviation σ_x *divided by* $\sqrt{N}$. This quantity is called the *standard deviation of the mean,* or SDOM, and is denoted $\sigma_{\bar{x}}$:

$$\sigma_{\bar{x}} = \sigma_x/\sqrt{N}. \tag{4.14}$$

(Other common names are *standard error* and *standard error of the mean.*) Thus, based on the N measured values $x_1, \ldots, x_N$, we can state our final answer for the

[4] If some springs are very different from the first, our uncertainty in measuring them may be different. Thus, if the springs differ a lot, we would need to check our uncertainty by making several measurements for each of two or three different springs.

value of x as

$$(\text{value of } x) = x_{\text{best}} \pm \delta x,$$

where $x_{\text{best}} = \bar{x}$, the mean of $x_1, \ldots, x_N$, and δx is the standard deviation of the mean,

$$\delta x = \sigma_{\bar{x}} = \sigma_x/\sqrt{N}. \tag{4.15}$$

As an example, we can consider the 10 measurements reported in (4.10) of the spring constant k of one spring. As we saw, the mean of these values is $\bar{k} = 85.7$ N/m, and the standard deviation is $\sigma_k = 2.2$ N/m. Therefore, the standard deviation of the mean is

$$\sigma_{\bar{k}} = \sigma_k/\sqrt{10} = 0.7 \text{ N/m}, \tag{4.16}$$

and our final answer, based on these 10 measurements, would be that the spring has

$$k = 85.7 \pm 0.7 \text{ newtons/meter}. \tag{4.17}$$

When you give an answer like this, you must state clearly what the numbers are—namely, the mean and the standard deviation of the mean—so your readers can judge their significance for themselves.

An important feature of the standard deviation of the mean, $\sigma_{\bar{x}} = \sigma_x/\sqrt{N}$, is the factor $\sqrt{N}$ in the denominator. The standard deviation σ_x represents the average uncertainty in the individual measurements $x_1, \ldots, x_N$. Thus, if we were to make some more measurements (using the same technique), the standard deviation σ_x would not change appreciably. On the other hand, the standard deviation of the mean, $\sigma_x/\sqrt{N}$, would slowly decrease as we increase N. This decrease is just what we would expect. If we make more measurements before computing an average, we would naturally expect the final result to be more reliable, and this improved reliability is just what the denominator $\sqrt{N}$ in (4.15) guarantees. This conclusion provides one obvious way to improve the precision of our measurements.

Unfortunately, the factor $\sqrt{N}$ grows rather slowly as we increase N. For example, if we wish to improve our precision by a factor of 10 simply by increasing the number of measurements N, we will have to increase N by a factor of 100—a daunting prospect, to say the least! Furthermore, we are for the moment neglecting systematic errors, and these are *not* reduced by increasing the number of measurements. Thus, in practice, if you want to increase your precision appreciably, you will probably do better to improve your technique than to rely merely on increased numbers of measurements.

Quick Check 4.2. A student makes five measurements of e, the magnitude of the electron's charge, as follows:

$$15, 17, 18, 14, 16,$$

all in units of 10^{-20} coulombs. Find her best estimate for e (as given by the mean) and its uncertainty (as given by the SDOM).

4.5 Examples

In this section, I discuss two examples of simple experiments that make use of the ideas of the past three sections.

Example: Area of a Rectangle

As a first, simple application of the standard deviation of the mean, imagine that we have to measure very accurately the area A of a rectangular plate approximately 2.5 cm $\times$ 5 cm. We first find the best available measuring device, which might be a vernier caliper, and then make several measurements of the length l and breadth b of the plate. To allow for irregularities in the sides, we make our measurements at several different positions, and to allow for small defects in the instrument, we use several different calipers (if available). We might make 10 measurements each of l and b and obtain the results shown in Table 4.3.

Table 4.3. Length and breadth (in mm).

	Measured values	Mean	SD	SDOM
l	24.25, 24.26, 24.22, 24.28, 24.24 24.25, 24.22, 24.26, 24.23, 24.24	$\bar{l} = 24.245$	$\sigma_l = 0.019$	$\sigma_{\bar{l}} = 0.006$
b	50.36, 50.35, 50.41, 50.37, 50.36 50.32, 50.39, 50.38, 50.36, 50.38	$\bar{b} = 50.368$	$\sigma_b = 0.024$	$\sigma_{\bar{b}} = 0.008$

Using the 10 observed values of l, you can quickly calculate the mean $\bar{l}$, the standard deviation σ_l, and the standard deviation of the mean $\sigma_{\bar{l}}$, as shown in the columns labeled mean, SD, and SDOM. In the same way you can calculate $\bar{b}$, σ_b, and $\sigma_{\bar{b}}$. Before doing any further calculations, you should examine these results to see if they seem reasonable. For example, the two standard deviations σ_l and σ_b are supposed to be the average uncertainty in the measurements of l and b. Because l and b were measured in exactly the same way, σ_l and σ_b should not differ significantly from each other or from what we judge to be a reasonable uncertainty for the measurements.

Having convinced yourself that the results so far are reasonable, you can quickly finish the calculations. The best estimate for the length is the mean $\bar{l}$ and the uncertainty is the SDOM $\sigma_{\bar{l}}$; so the final value for l is

$$l = 24.245 \pm 0.006 \text{ mm} \quad (\text{or } 0.025\%);$$

the number in parenthesis is the percentage uncertainty. Similarly, the value for b is

$$b = 50.368 \pm 0.008 \text{ mm} \quad (\text{or } 0.016\%).$$

Finally, the best estimate for the area $A = lb$ is the product of these values, with a fractional uncertainty given by the quadratic sum of those in l and b (assuming the

errors are independent):

$$\begin{aligned} A &= (24.245 \text{ mm} \pm 0.025\%) \times (50.368 \text{ mm} \pm 0.016\%) \\ &= 1221.17 \text{ mm}^2 \pm 0.03\% \\ &= 1221.2 \pm 0.4 \text{ mm}^2. \end{aligned} \tag{4.18}$$

To arrive at the answer (4.18) for A, we calculated the averages $\bar{l}$ and $\bar{b}$, each with an uncertainty equal to the standard deviation of its mean. We then calculated the area A as the product of $\bar{l}$ and $\bar{b}$ and found the uncertainty by propagation of errors. We could have proceeded differently. For instance, we could have multiplied the first measured value of l by the first value of b to give a first answer for A. Continuing in this way we could have calculated 10 answers for A and then have subjected these 10 answers to statistical analysis, calculating $\bar{A}$, σ_A, and finally $\sigma_{\bar{A}}$. If, however, the errors in l and b are independent and random, and if we make enough measurements, this alternative procedure will produce the same result as the first one.[5]

Example: Another Spring

As a second example, consider a case in which a statistical analysis cannot be applied to the direct measurements but can to the final answers. Suppose we wish to measure the spring constant k of a spring by timing the oscillations of a mass m fixed to its end. We know from elementary mechanics that the period for such oscillations is $T = 2\pi\sqrt{m/k}$. Thus, by measuring T and m, we can find k as

$$k = 4\pi^2 m/T^2. \tag{4.19}$$

The simplest way to find k is to take a single, accurately known mass m and make several careful measurements of T. For various reasons, however, timing T for several *different* masses m may be more interesting. (For example, in this way, we could check that $T \propto \sqrt{m}$ as well as measure k.) We might then get a set of readings such as those in the first two lines of Table 4.4.

Table 4.4. Measurement of spring constant k.

Mass m (kg)	0.513	0.581	0.634	0.691	0.752	0.834	0.901	0.950
Period T (s)	1.24	1.33	1.36	1.44	1.50	1.59	1.65	1.69
$k = 4\pi^2 m/T^2$	13.17	12.97	etc.					

It obviously makes no sense to average the various different masses in the top line (or the times in the second line) because they are *not* different measurements of the same quantity. Nor can we learn anything about the uncertainty in our measurements by comparing the different values of m. On the other hand, we can com-

[5]The second procedure has a certain illogic because there is no particular reason to associate the first measurement of l with the first measurement of b. Indeed, we might have measured l eight times and b twelve times; then we couldn't pair off values. Thus, our first procedure is logically preferable.

bine each value of m with its corresponding period T and calculate k, as in the final line of Table 4.4. Our answers for k in the bottom line *are* all measurements of the same quantity and so can be subjected to statistical analysis. In particular, our best estimate for k is the mean, $\bar{k} = 13.16$ N/m, and our uncertainty is the standard deviation of the mean, $\sigma_{\bar{k}} = 0.06$ N/m (see Problem 4.20). Thus, the final answer, based on the data of Table 4.4, is

$$\text{spring constant } k = 13.16 \pm 0.06 \text{ N/m}. \tag{4.20}$$

If we had formed reasonable estimates of the uncertainties in our original measurements of m and T, we could also have estimated the uncertainty in k by using error propagation, starting from these estimates for δm and δT. In this case, it would be a good idea to compare the final uncertainties in k obtained by the two methods.

4.6 Systematic Errors

In the past few sections, I have been taking for granted that all systematic errors were reduced to a negligible level before serious measurements began. Here, I take up again the disagreeable possibility of appreciable systematic errors. In the example just discussed, we may have been measuring m with a balance that read consistently high or low, or our timer may have been running consistently fast or slow. Neither of these systematic errors will show up in the comparison of our various answers for the spring constant k. As a result, the standard deviation of the mean $\sigma_{\bar{k}}$ can be regarded as the *random component* δk_{ran} of the uncertainty δk but is certainly not the total uncertainty δk. Our problem is to decide how to estimate the *systematic component* δk_{sys} and then how to combine δk_{ran} and δk_{sys} to give the complete uncertainty δk.

No simple theory tells us what to do about systematic errors. In fact, the only theory of systematic errors is that they must be identified and reduced until they are much less than the required precision. In a teaching laboratory, however, this goal is often not attainable. Students often cannot check a meter against a better one to correct it, much less buy a new meter to replace an inadequate one. For this reason, some teaching laboratories establish a rule that, in the absence of more specific information, meters should be considered to have some definite systematic uncertainty. For example, the decision might be that all stopwatches have up to 0.5% systematic uncertainty, all balances up to 1%, all voltmeters and ammeters up to 3%, and so on.

Given rules of this kind, there are various possible ways to proceed. None can really be rigorously justified, and we describe just one approach here. (Problems 4.23 to 4.28 contain more examples.) In the last example in Section 4.5, the spring constant $k = 4\pi^2 m/T^2$ was found by measuring a series of values of m and the corresponding values of T. As we have seen, a statistical analysis of the various answers for k gives the random component of δk as

$$\delta k_{\text{ran}} = \sigma_{\bar{k}} = 0.06 \text{ N/m}. \tag{4.21}$$

Suppose now we have been told that the balance used to measure m and the clock used for T have systematic uncertainties up to 1% and 0.5%, respectively. We can

then find the systematic component of δk by propagation of errors; the only question is whether to combine the errors in quadrature or directly. Because the errors in m and T are surely independent and some cancellation is therefore possible, using the quadratic sum is probably reasonable[6]; this choice gives

$$\frac{\delta k_{\text{sys}}}{k} = \sqrt{\left(\frac{\delta m_{\text{sys}}}{m}\right)^2 + \left(2\frac{\delta T_{\text{sys}}}{T}\right)^2} \tag{4.22}$$

$$= \sqrt{(1\%)^2 + (1\%)^2} = 1.4\% \tag{4.23}$$

and hence

$$\begin{aligned} \delta k_{\text{sys}} &= k_{\text{best}} \times (1.4\%) \\ &= (13.16 \text{ N/m}) \times 0.014 = 0.18 \text{ N/m}. \end{aligned} \tag{4.24}$$

Now that we have estimates for both the random and systematic uncertainties in k, we must decide how to state our final conclusion for the spring constant k with its overall uncertainty. Because the method for combining δk_{ran} and δk_{sys} is not completely clear, many scientists leave the two components separate and state a final answer in the form

$$\begin{aligned} (\text{measured value of } k) &= k_{\text{best}} \pm \delta k_{\text{ran}} \pm \delta k_{\text{sys}} \\ &= 13.16 \pm 0.06 \pm 0.18 \text{ N/m} \end{aligned} \tag{4.25}$$

(all of which should probably be rounded to one decimal place). Alternatively, a case can be made that δk_{ran} and δk_{sys} should be combined in quadrature, in which case we could state a single, total uncertainty

$$\begin{aligned} \delta k &= \sqrt{(\delta k_{\text{ran}})^2 + (\delta k_{\text{sys}})^2} \\ &= \sqrt{(0.06)^2 + (0.18)^2} = 0.19 \text{ N/m} \end{aligned} \tag{4.26}$$

and replace the conclusion (4.25) by

$$\begin{aligned} (\text{measured value of } k) &= k_{\text{best}} \pm \delta k \\ &= 13.16 \pm 0.19 \text{ N/m} \end{aligned}$$

or, probably better, 13.2 ± 0.2 N/m.

The expression (4.26) for δk cannot really be rigorously justified. Nor is the significance of the answer clear; for example, we probably cannot claim 68% confidence that the true answer lies in the range $\bar{k} \pm \delta k$. Nonetheless, the expression does at least provide a reasonable estimate of our total uncertainty, given that our apparatus has systematic uncertainties we could not eliminate. In particular, there is one important respect in which the answer (4.26) is realistic and instructive. We saw in Section 4.4 that the standard deviation of the mean $\sigma_{\bar{k}}$ approaches zero as the number of measurements N is increased. This result suggested that, if you have the

[6] Whether we should use the quadratic or ordinary sum really depends on what is meant by the statement that the balance has "up to 1% systematic uncertainty." If it means the error is *certainly* no more than 1% (and likewise for the clock), then direct addition is appropriate, and δk_{sys} is then *certainly* no more than 2%. On the other hand, perhaps an analysis of all balances in the laboratory has shown that they follow a normal distribution, with 68% of them better than 1% reliable (and likewise for the clocks). In this case, we can use addition in quadrature as in (4.22) with the usual significance of 68% confidence.

patience to make an enormous number of measurements, you can reduce the uncertainties indefinitely without having to improve your equipment or technique. We can now see that this suggestion is incorrect. Increasing N can reduce the *random* component $\delta k_{\text{ran}} = \sigma_{\bar{k}}$ indefinitely. But any given apparatus has *some* systematic uncertainty, which is *not* reduced as we increase N. From (4.26) we clearly see that little is gained from further reduction of δk_{ran}, once δk_{ran} is smaller than δk_{sys}. In particular, the total δk can never be made less than δk_{sys}. This fact simply confirms what we already guessed, that in practice a large reduction of the uncertainty requires improvements in techniques or equipment to reduce both the random and systematic errors in each single measurement.

As discussed in Chapter 2, a peculiar feature of the teaching laboratory is that you will probably be asked to measure quantities, such as the acceleration of gravity, for which an accurate, accepted value is already known. In this kind of experiment, the logic of the error analysis is a bit confusing. Probably the most honest course is to ignore the known accepted value until *after* you have done all calculations of your measured value, q_{best}, and its uncertainty. Then, of course, you must ask whether the accepted value lies inside (or at least close to) the range $q_{\text{best}} \pm \delta q$. If it does, you can simply record this agreement in your report. If the accepted value lies well outside the range $q_{\text{best}} \pm \delta q$, however, you have to examine the possible causes of the excessive discrepancy. For example, you might measure g, the acceleration of gravity, and get the results (all in m/s^2),

$$g_{\text{best}} = 9.97, \tag{4.27}$$

with uncertainties

$$\delta g_{\text{ran}} = 0.02 \quad \text{and} \quad \delta g_{\text{sys}} = 0.03,$$

and hence a total uncertainty, as in (4.26), of

$$\delta g = 0.04.$$

Clearly, the accepted value of

$$g = 9.80 \text{ m/s}^2$$

lies far outside the measured range, 9.97 ± 0.04. (More specifically, the discrepancy is 0.17, which is four times the uncertainty.) This result is definitely *not* satisfactory and further analysis is required.

The first thing to check is the possibility that you made a downright mistake in calculating g_{best} or one of the uncertainties δg_{ran} and δg_{sys}. If you can convince yourself that all your calculations were correct, the next possibility is that the accepted value is wrong. In the case of $g = 9.80 \text{ m/s}^2$ this possibility is rather unlikely, but it is entirely possible for plenty of other cases. For example, suppose you were measuring the density of air; because this is strongly dependent on the temperature and pressure, you could easily have looked up the wrong accepted value for this parameter.

Once you have eliminated these suspects, only one possibility is left: You must have overlooked some systematic error so that your value of δg_{sys} is too small. Ideally, you should try to find the culprit, but this search can be hard because of the many possibilities:

(1) Perhaps one of your meters had larger systematic errors than you had allowed for when you calculated δg_{sys}. You can investigate this possibility by determining how large a systematic error in your clock (or voltmeter, or whatever) would be needed to account for the offending discrepancy. If the needed error is not unreasonably large, you have one possible explanation of your difficulty.

(2) Another possible cause of systematic error is that you used an incorrect value for some parameter needed in your calculations. A celebrated example of this was Millikan's famous measurement of the electron's charge, e. Millikan's method depended on the viscosity of air, for which he used a value that was 0.4% too small. This discrepancy caused all of his values of e to be 0.6% too small, an error that was not noticed for nearly 20 years. This kind of mistake sometimes arises in a teaching laboratory when a student uses a value that has too few significant figures. For example, suppose you do an experiment with protons and you expect to have an accuracy better than 1%. If you take the proton's mass to be 1.7×10^{-27} kg (instead of the more exact 1.67×10^{-27} kg), you will have introduced a 2% systematic error, which will almost certainly frustrate your hope for 1% results.

(3) Much harder to analyze is the possibility of a flaw in the design of the experiment. For example, if you had measured g by dropping an object from a great height, air resistance could introduce an appreciable systematic error. [Note, however, that this error would not account for the large value of g in (4.27) because air resistance would cause an acceleration that was too *small.*] Similarly, if you try to measure the half-life of a radioactive material and your sample is contaminated with another material of shorter half-life, you will get an answer that is systematically too short.

Obviously, tracking down the source of systematic errors is difficult and has defied the best efforts of many great scientists. In all probability, your instructors are not going to penalize you too severely if you fail to do so. Nevertheless, they will expect an intelligent discussion of the problem and at least an honest admission that there appear to have been systematic errors that you were unable to identify.

Principal Definitions and Equations of Chapter 4

Suppose that we make N measurements, $x_1, x_2, \ldots, x_N$ of the same quantity x, all using the same method. Provided all uncertainties are random and small, we have the following results:

THE MEAN

The best estimate for x, based on these measurements, is their mean:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i. \qquad \text{[See (4.5)]}$$

THE STANDARD DEVIATION

The average uncertainty of the individual measurements $x_1, x_2, \ldots, x_N$ is given by the standard deviation, or SD:

$$\sigma_x = \sqrt{\frac{1}{N-1}\Sigma(x_i - \bar{x})^2}. \qquad \text{[See (4.9)]}$$

This definition of the SD, often called the *sample* standard deviation, is the most appropriate for our purposes. The *population* standard deviation is obtained by replacing the factor $(N - 1)$ in the denominator by N. You will usually want to calculate standard deviations using the built-in function on your calculator; be sure you know which definition it uses.

The detailed significance of the standard deviation σ_x is that approximately 68% of the measurements of x (using the same method) should lie within a distance σ_x of the true value. (This claim is justified in Section 5.4.) This result is what allows us to identify σ_x as the *uncertainty* in any one measurement of x,

$$\delta x = \sigma_x,$$

and, with this choice, we can be 68% confident that any one measurement will fall within σ_x of the correct answer.

THE STANDARD DEVIATION OF THE MEAN

As long as systematic uncertainties are negligible, the uncertainty in our best estimate for x (namely $\bar{x}$) is the standard deviation of the mean, or SDOM,

$$\sigma_{\bar{x}} = \frac{\sigma_x}{\sqrt{N}}. \qquad \text{[See (4.14)]}$$

If there *are* appreciable systematic errors, then $\sigma_{\bar{x}}$ gives the *random component* of the uncertainty in our best estimate for x:

$$\delta x_{\text{ran}} = \sigma_{\bar{x}}.$$

If you have some way to estimate the systematic component δx_{sys}, a reasonable (but not rigorously justified) expression for the total uncertainty is the quadratic sum of δx_{ran} and δx_{sys}:

$$\delta x_{\text{tot}} = \sqrt{(\delta x_{\text{ran}})^2 + (\delta x_{\text{sys}})^2}. \qquad \text{[See (4.26)]}$$

Problems for Chapter 4

For Section 4.2: The Mean and Standard Deviation

4.1. ★ You measure the time for a ball to drop from a second-floor window three times and get the results (in tenths of a second):

11, 13, 12.