Part II

If you have read and understood Chapter 5, you are now ready, with surprisingly little difficulty, to study a number of more advanced topics. The chapters of Part II present seven such topics, some of which are applications of the statistical theory already developed, and others of which are further extensions of that theory. All are important, and you will probably study them sooner or later. Because you might not necessarily want to learn them all at once, these topics have been arranged into independent, short chapters that can be studied in any order, as your needs and interests dictate.

Chapter 6

Rejection of Data

This chapter discusses the awkward question of whether to discard a measurement that seems so unreasonable that it looks like a mistake. This topic is controversial; some scientists would argue that discarding a measurement just because it looks unreasonable is *never* justified. Nevertheless, there is a simple test you could at least consider applying if you find yourself confronted with this situation. The test is called *Chauvenet's criterion* and is a nice application of the statistical ideas developed in Chapters 4 and 5.

6.1 The Problem of Rejecting Data

Sometimes, one measurement in a series of measurements appears to disagree strikingly with all the others. When this happens, the experimenter must decide whether the anomalous measurement resulted from some mistake and should be rejected or was a *bona fide* measurement that should be used with all the others. For example, imagine we make six measurements of the period of a pendulum and get the results (all in seconds)

$$3.8,\ 3.5,\ 3.9.\ 3.9,\ 3.4,\ 1.8. \tag{6.1}$$

In this example, the value 1.8 is startlingly different from all the others, and we must decide what to do with it.

We know from Chapter 5 that a legitimate measurement *may* deviate significantly from other measurements of the same quantity. Nevertheless, a legitimate discrepancy as large as that of the last measurement in (6.1) is *very improbable,* so we are inclined to suspect that the time 1.8 s resulted from some undetected mistake or other external cause. Perhaps, for example, we simply misread the last time or our electric timer stopped briefly during the last measurement because of a momentary power failure.

If we have kept very careful records, we may sometimes be able to establish such a definite cause for the anomalous measurement. For example, our records might show that a different stopwatch was used for the last timing in (6.1), and a subsequent check might show that this watch runs slow. In this case, the anomalous measurement should definitely be rejected.

Unfortunately, establishing an external cause for an anomalous result is usually not possible. We must then decide whether or not to reject the anomaly simply by examining the results themselves, and here our knowledge of the Gauss distribution proves useful.

The rejection of data is a controversial question on which experts disagree. It is also an *important* question. In our example, the best estimate for the period of the pendulum is significantly affected if we reject the suspect 1.8 s. The average of all six measurements is 3.4 s, whereas that of the first five is 3.7 s, an appreciable difference.

Furthermore, the decision to reject data is ultimately a subjective one, and the scientist who makes this decision may reasonably be accused by other scientists of "fixing" the data. The situation is made worse by the possibility that the anomalous result may reflect some important effect. Indeed, many important scientific discoveries first appeared as anomalous measurements that looked like mistakes. In throwing out the time 1.8 s in the example (6.1), we just *might* be throwing out the most interesting part of the data.

In fact, faced with data like those in (6.1), our only really honest course is to repeat the measurement many, many times. If the anomaly shows up again, we will presumably be able to trace its cause, either as a mistake or a real physical effect. If it does not recur, then by the time we have made, say, 100 measurements, there will be no significant difference in our final answer whether we include the anomaly or not.

Nevertheless, repeating a measurement 100 times every time a result seems suspect is frequently impractical (especially in a teaching laboratory). We therefore need some criterion for rejecting a suspect result. There are various such criteria, some quite complicated. Chauvenet's criterion provides a simple and instructive application of the Gauss distribution.

6.2 Chauvenet's Criterion

Let us return to the six measurements of the example (6.1):

$$3.8,\ 3.5,\ 3.9,\ 3.9,\ 3.4,\ 1.8.$$

If we assume for the moment that these are six legitimate measurements of a quantity x, we can calculate the mean $\bar{x}$ and standard deviation σ_x,

$$\bar{x} = 3.4 \text{ s} \tag{6.2}$$

and

$$\sigma_x = 0.8 \text{ s}. \tag{6.3}$$

We can now quantify the extent to which the suspect measurement, 1.8, is anomalous. It differs from the mean 3.4 by 1.6, or two standard deviations. If we assume the measurements were governed by a Gauss distribution with center and width given by (6.2) and (6.3), we can calculate the probability of obtaining measurements that differ by at least this much from the mean. According to the probabilities shown

in Appendix A, this probability is

$$\begin{aligned} Prob(\text{outside } 2\sigma) &= 1 - Prob(\text{within } 2\sigma) \\ &= 1 - 0.95 \\ &= 0.05. \end{aligned}$$

In other words, assuming that the values (6.2) and (6.3) for $\bar{x}$ and σ_x are legitimate, we would expect one in every 20 measurements to differ from the mean by at least as much as the suspect 1.8 s does. If we had made 20 or more measurements, we should actually *expect* to get one or two measurements as deviant as the 1.8 s, and we would have no reason to reject it. But we have made only six measurements, so the expected number of measurements as deviant as 1.8 s was actually

$$\begin{aligned} &(\text{expected number as deviant as 1.8 s}) \\ &\quad = (\text{number of measurements}) \times Prob(\text{outside } 2\sigma) \\ &\quad = 6 \times 0.05 = 0.3. \end{aligned}$$

That is, in six measurements we would expect (on average) only one-third of a measurement as deviant as the suspect 1.8 s.

This result provides us with the needed quantitative measure of the "reasonableness" of our suspect measurement. If we choose to regard one-third of a measurement as "ridiculously improbable," then we will conclude that the value 1.8 s was not a legitimate measurement and should be rejected.

The decision of where to set the boundary of "ridiculous improbability" is up to the experimenter. Chauvenet's criterion, as normally given, states that if the expected number of measurements at least as deviant as the suspect measurement is less than one-half, then the suspect measurement should be rejected. Obviously, the choice of one-half is arbitrary, but it is also reasonable and can be defended.

The application of Chauvenet's criterion to a general problem can now be described easily. Suppose you have made N measurements

$$x_1, \ldots, x_N$$

of a single quantity x. From all N measurements, you calculate $\bar{x}$ and σ_x. If one of the measurements (call it x_{sus}) differs from $\bar{x}$ so much that it looks suspicious, then find

$$t_{\text{sus}} = \frac{|x_{\text{sus}} - \bar{x}|}{\sigma_x}, \tag{6.4}$$

the number of standard deviations by which x_{sus} differs from $\bar{x}$. Next, from Appendix A you can find the probability

$$Prob(\text{outside } t_{\text{sus}}\sigma)$$

that a legitimate measurement would differ from $\bar{x}$ by t_{sus} or more standard deviations. Finally, multiplying by N, the total number of measurements, gives

$$\begin{aligned} n &= (\text{expected number as deviant as } x_{\text{sus}}) \\ &= N \times Prob(\text{outside } t_{\text{sus}}\sigma). \end{aligned}$$

If this expected number n is less than one-half, then, according to Chauvenet's criterion, you can reject x_{sus}.

If you do decide to reject x_{sus}, you would naturally recalculate $\bar{x}$ and σ_x using just the remaining data; in particular, your final answer for x would be this new mean, with an uncertainty equal to the new SDOM.

Example: Ten Measurements of a Length

A student makes 10 measurements of one length x and gets the results (all in mm)

$$46,\ 48,\ 44,\ 38,\ 45,\ 47,\ 58,\ 44,\ 45,\ 43.$$

Noticing that the value 58 seems anomalously large, he checks his records but can find no evidence that the result was caused by a mistake. He therefore applies Chauvenet's criterion. What does he conclude?

Accepting provisionally all 10 measurements, he computes

$$\bar{x} = 45.8 \quad \text{and} \quad \sigma_x = 5.1.$$

The difference between the suspect value $x_{sus} = 58$ and the mean $\bar{x} = 45.8$ is 12.2, or 2.4 standard deviations; that is,

$$t_{sus} = \frac{x_{sus} - \bar{x}}{\sigma_x} = \frac{58 - 45.8}{5.1} = 2.4.$$

Referring to the table in Appendix A, he sees that the probability that a measurement will differ from $\bar{x}$ by $2.4\sigma_x$ or more is

$$\begin{aligned} Prob(\text{outside } 2.4\sigma) &= 1 - Prob(\text{within } 2.4\sigma) \\ &= 1 - 0.984 \\ &= 0.016. \end{aligned}$$

In 10 measurements, he would therefore expect to find only 0.16 of one measurement as deviant as his suspect result. Because 0.16 is less than the number 0.5 set by Chauvenet's criterion, he should at least consider rejecting the result.

If he decides to reject the suspect 58, then he must recalculate $\bar{x}$ and σ_x as

$$\bar{x} = 44.4 \quad \text{and} \quad \sigma_x = 2.9.$$

As you would expect, his mean changes a bit, and his standard deviation drops appreciably.

Quick Check 6.1. A student makes 20 measurements of a certain voltage V and computes her mean and standard deviation as $\bar{V} = 51$ and $\sigma_V = 2$ (both in microvolts). The evening before the work is due, she starts to write her report and realizes that one of her measured values was $V_{sus} = 56$. What is the probability of her getting a measurement this deviant from $\bar{V}$? If she decides to use Chauvenet's criterion, should she reject the suspect value?

6.3 Discussion

You should be aware that some scientists believe that data should *never* be rejected without *external* evidence that the measurement in question is incorrect. A reasonable compromise is to use Chauvenet's criterion to identify data that could be *considered* for rejection; having made this identification, you could do all subsequent calculations twice, once including the suspect data and once excluding them, to see how much the questionable values affect your final conclusion.

One reason many scientists are uncomfortable with Chauvenet's criterion is that the choice of one-half as the boundary of rejection (in the condition that $n < \frac{1}{2}$) is arbitrary. Perhaps even more important, unless you have made a very large number of measurements ($N \approx 50$, say), the value of σ_x is extremely uncertain as an estimate for the true standard deviation of the measurements. (See Problem 6.7 for an example.) This means in turn that the number t_{sus} in (6.4) is very uncertain. Because the probability of a measurement outside t standard deviations is very sensitive to t, a large error in t_{sus} causes a very large error in this probability and casts serious doubt on the whole procedure. For both of these reasons, Chauvenet's criterion should be used only as a last resort, when you cannot check your measurements by repeating them.

So far, we have assumed that only one measurement is suspect. What should you do if you have several? Given that the use of Chauvenet's criterion to reject *one* measurement is open to doubt, clearly its use to reject *several* measurements is even more problematic. Nevertheless, if there is absolutely no way you can repeat your measurements (because you have rashly dismantled your equipment before writing your report, for example), you may have to confront this question.

Suppose first that you have two measurements that deviate from the mean by the same large amount. In this case, you could calculate the expected number of measurements this deviant, and if this number is less than one (that is, *two times* one-half), then both measurements could be considered candidates for rejection. If you have two suspect measurements, x_1 and x_2, with x_2 more deviant than x_1, you should first apply Chauvenet's criterion using the value x_1. If the expected number this deviant is less than one, you could reject *both* values. If this expected number is more than one, you certainly should not reject both but instead reapply Chauvenet's criterion using x_2 and, if the expected number this deviant is less than one-half, you could reject just x_2.

Having rejected any measurements that fail Chauvenet's criterion, you would naturally recalculate $\bar{x}$ and σ_x using just the remaining data. The resulting value of σ_x will be smaller than the original one, and with the new σ_x, some more measurements may fail Chauvenet's criterion. However, agreement seems widespread that Chauvenet's criterion should *not* be applied a second time using the recalculated values of $\bar{x}$ and σ_x.

Principal Definitions and Equations of Chapter 6

CHAUVENET'S CRITERION

If you make N measurements $x_1, \ldots, x_N$ of a single quantity x, and if one of the measurements (x_{sus}, say) is suspiciously different from all the others, Chauvenet's criterion gives a simple test for deciding whether to reject this suspect value. First, compute the mean and standard deviation of all N measurements and then find the number of standard deviations by which x_{sus} differs from $\bar{x}$,

$$t_{\text{sus}} = \frac{|x_{\text{sus}} - \bar{x}|}{\sigma_x}.$$

Next, find the probability (assuming the measurements are normally distributed about $\bar{x}$ with width σ_x) of getting a result as deviant as x_{sus}, and, hence, the number of measurements expected to deviate this much,

$$n = N \times \textit{Prob}(\text{outside } t_{\text{sus}}\sigma).$$

If $n < \frac{1}{2}$, then according to Chauvenet's criterion, you can reject the value x_{sus}.

Because there are several objections to Chauvenet's criterion (especially if N is not very large), it should be used only as a last resort, when the measurements of x cannot be checked. The objections to Chauvenet's criterion are even greater if two or more measurements are suspect, but the test *can* be extended to this situation, as described in Section 6.3.

Problems for Chapter 6

For Section 6.2: Chauvenet's Criterion

6.1. ★ An enthusiastic student makes 50 measurements of the heat Q released in a certain reaction. Her average and standard deviation are

$$\overline{Q} = 4.8 \quad \text{and} \quad \sigma_Q = 0.4,$$

both in kilocalories. **(a)** Assuming her measurements are governed by the normal distribution, find the probability that any one measurement would differ from $\overline{Q}$ by 0.8 kcal or more. How many of her 50 measurements should she expect to differ from $\overline{Q}$ by 0.8 kcal or more? If one of her measurements is 4.0 kcal and she decides to use Chauvenet's criterion, would she reject this measurement? **(b)** Would she reject a measurement of 6.0 kcal?

6.2. ★ The Franck-Hertz experiment involves measuring the differences between a series of equally spaced voltages that cause the maximum current through a tube of mercury vapor. A student measures 10 such differences and obtains these results (all in volts):

0.48, 0.45, 0.49, 0.46, 0.44, 0.57, 0.45, 0.47, 0.51, 0.50.

(a) Calculate the mean and standard deviation of these results. (By all means, use the built-in functions on your calculator.)
(b) If he decides to use Chauvenet's criterion, should he reject the reading of 0.57 volts? Explain your reasoning clearly.

6.3. ★ A student makes 14 measurements of the period of a damped oscillator and obtains these results (in tenths of a second):

$$7, 3, 9, 3, 6, 9, 8, 7, 8, 12, 5, 9, 9, 3.$$

Believing that the result 12 is suspiciously high, she decides to apply Chauvenet's criterion. How many results should she expect to find as far from the mean as 12 is? Should she reject the suspect result?

6.4. ★★ A physicist makes 14 measurements of the density of tracks on an emulsion exposed to cosmic rays and obtains the following results (in tracks/cm^2):

$$11, 9, 13, 15, 8, 10, 5, 11, 9, 12, 12, 13, 9, 14.$$

(a) What are his mean and standard deviation? (Use the built-in functions on your calculator.) **(b)** According to Chauvenet's criterion, would he be justified in rejecting the measurement of 5? Explain your reasoning clearly. **(c)** If he does reject this measurement, what does he get for his new mean and standard deviation? (Hint: You can almost certainly recalculate the mean and standard deviation by editing the contents of your calculator's statistical registers rather than re-entering all the data. If you don't know how to do this editing, take a moment to learn.)

6.5. ★★ In the course of a couple of hours, a nuclear engineer makes 12 measurements of the strength of a long-lived radioactive source with the following results, in millicuries:

$$12, 34, 22, 14, 22, 17, 24, 22, 18, 14, 18, 12.$$

(Because the source has a long life, its activity should not change appreciably during the time all the measurements are made.)
(a) What are his mean and standard deviation? (Use the built-in functions on your calculator.) **(b)** According to Chauvenet's criterion, would he be justified in rejecting the value 34 as a mistake? Explain your reasoning clearly. **(c)** If he does reject this measurement, what does he get for his new mean and standard deviation? (Read the hint to Problem 6.4.)

6.6. ★★ Chauvenet's criterion defines a boundary outside which a measurement is regarded as rejectable. If we make 10 measurements and one differs from the mean by more than about two standard deviations (in either direction), that measurement is considered rejectable. For 20 measurements, the corresponding boundary is approximately 2.2 standard deviations. Make a table showing the "boundary of rejectability" for 5, 10, 15, 20, 50, 100, 200, and 1,000 measurements of a quantity that is normally distributed.

6.7. ★★ Based on N measurements of a normally distributed quantity x, the best estimate of the width σ is the standard deviation σ_x of the N measured values. Unfortunately, if N is small, this estimate is fairly uncertain. Specifically, the fractional uncertainty in σ_x as an estimate of σ is given by (5.46) as $1/\sqrt{2(N-1)}$. If

$N = 6$, for example, our estimate for σ is about 30% uncertain. This large uncertainty in σ means you should regard Chauvenet's criterion with considerable skepticism when N is small, as the following example illustrates.

An experimenter measures a certain current six times and obtains the following results (in milliamps):

$$28,\ 25,\ 29,\ 18,\ 29,\ 24.$$

(a) Find the mean $\bar{x}$ and standard deviation σ_x of these measurements. **(b)** If he decided to apply Chauvenet's criterion, would the experimenter reject the value 18? **(c)** Find the uncertainty—call it $\delta\sigma_x$—in σ_x. **(d)** The true value of σ may well be as small as $\sigma_x - \delta\sigma_x$ or as large as $\sigma_x + \delta\sigma_x$. Re-apply Chauvenet's criterion, using each of these two values for σ. Comment on the results.

Chapter 7

Weighted Averages

This chapter addresses the problem of combining two or more separate and independent measurements of a single physical quantity. We will find that the best estimate of that quantity, based on the several measurements, is an appropriate *weighted average* of those measurements.

7.1 The Problem of Combining Separate Measurements

Often, a physical quantity is measured several times, perhaps in several separate laboratories, and the question arises how these measurements can be combined to give a single best estimate. Suppose, for example, that two students, A and B, measure a quantity x carefully and obtain these results:

$$\text{Student } A: \quad x = x_A \pm \sigma_A \tag{7.1}$$

and

$$\text{Student } B: \quad x = x_B \pm \sigma_B. \tag{7.2}$$

Each result will probably itself be the result of several measurements, in which case x_A will be the mean of all A's measurements and σ_A the standard deviation of that mean (and similarly for x_B and σ_B). The question is how best to combine x_A and x_B for a single best estimate of x.

Before examining this question, note that if the discrepancy $|x_A - x_B|$ between the two measurements is much greater than both uncertainties σ_A and σ_B, we should suspect that something has gone wrong in at least one of the measurements. In this situation, we would say that the two measurements are *inconsistent,* and we should examine both measurements carefully to see whether either (or both) was subject to unnoticed systematic errors.

Let us suppose, however, that the two measurements (7.1) and (7.2) are *consistent;* that is, the discrepancy $|x_A - x_B|$ is *not* significantly larger than both σ_A and σ_B. We can then sensibly ask what the best estimate x_{best} is of the true value X, based on the two measurements. Your first impulse might be to use the average $(x_A + x_B)/2$ of the two measurements. Some reflection should suggest, however, that this average is unsuitable if the two uncertainties σ_A and σ_B are unequal. The simple

average $(x_A + x_B)/2$ gives equal importance to both measurements, whereas the more precise reading should somehow be given more weight.

Throughout this chapter, I will assume all systematic errors have been identified and reduced to a negligible level. Thus, all remaining errors are random, and the measurements of x are distributed normally around the true value X.

7.2 The Weighted Average

We can solve our problem easily by using the principle of maximum likelihood, much as we did in Section 5.5. We are assuming that both measurements are governed by the Gauss distribution and denote the unknown true value of x by X. Therefore, the probability of Student A's obtaining his particular value x_A is

$$Prob_X(x_A) \propto \frac{1}{\sigma_A} e^{-(x_A - X)^2/2\sigma_A{}^2}, \tag{7.3}$$

and that of B's getting his observed x_B is

$$Prob_X(x_B) \propto \frac{1}{\sigma_B} e^{-(x_B - X)^2/2\sigma_B{}^2}. \tag{7.4}$$

The subscript X indicates explicitly that these probabilities depend on the unknown actual value.

The probability that A finds the value x_A *and* B the value x_B is just the product of the two probabilities (7.3) and (7.4). In a way that should now be familiar, this product will involve an exponential function whose exponent is the sum of the two exponents in (7.3) and (7.4). We write this as

$$\begin{aligned} Prob_X(x_A, x_B) &= Prob_X(x_A)\, Prob_X(x_B) \\ &\propto \frac{1}{\sigma_A \sigma_B} e^{-\chi^2/2}, \end{aligned} \tag{7.5}$$

where I have introduced the convenient shorthand χ^2 (chi squared) for the exponent

$$\chi^2 = \left(\frac{x_A - X}{\sigma_A}\right)^2 + \left(\frac{x_B - X}{\sigma_B}\right)^2. \tag{7.6}$$

This important quantity is the sum of the squares of the deviations from X of the two measurements, each divided by its corresponding uncertainty.

The principle of maximum likelihood asserts, just as before, that our best estimate for the unknown true value X is that value for which the actual observations x_A, x_B are most likely. That is, the best estimate for X is the value for which the probability (7.5) is maximum or, equivalently, the exponent χ^2 is minimum. (Because maximizing the probability entails minimizing the "sum of squares" χ^2, this method for estimating X is sometimes called the "method of least squares.") Thus, to find the best estimate, we simply differentiate (7.6) with respect to X and set the derivative equal to zero,

$$2\frac{x_A - X}{\sigma_A{}^2} + 2\frac{x_B - X}{\sigma_B{}^2} = 0.$$

The solution of this equation for X is our best estimate and is easily seen to be

$$(\text{best estimate for } X) = \left(\frac{x_A}{\sigma_A{}^2} + \frac{x_B}{\sigma_B{}^2}\right) \Big/ \left(\frac{1}{\sigma_A{}^2} + \frac{1}{\sigma_B{}^2}\right). \tag{7.7}$$

This rather ugly result can be made tidier if we define *weights*

$$w_A = \frac{1}{\sigma_A{}^2} \quad \text{and} \quad w_B = \frac{1}{\sigma_B{}^2}. \tag{7.8}$$

With this notation, we can rewrite (7.7) as the *weighted average* (denoted x_{wav})

$$(\text{best estimate for } X) = x_{\text{wav}} = \frac{w_A x_A + w_B x_B}{w_A + w_B}. \tag{7.9}$$

If the original two measurements are equally uncertain ($\sigma_A = \sigma_B$ and hence $w_A = w_B$), this answer reduces to the simple average $(x_A + x_B)/2$. In general, when $w_A \neq w_B$, the weighted average (7.9) is *not* the same as the ordinary average; it is similar to the formula for the center of gravity of two bodies, where w_A and w_B are the actual weights of the two bodies, and x_A and x_B their positions. In (7.9), the "weights" are the inverse squares of the uncertainties in the original measurements, as in (7.8). If A's measurement is more precise than B's, then $\sigma_A < \sigma_B$ and hence $w_A > w_B$, so the best estimate x_{best} is closer to x_A than to x_B, just as it should be.

Quick Check 7.1. Workers from two laboratories report the lifetime of a certain particle as 10.0 ± 0.5 and 12 ± 1, both in nanoseconds. If they decide to combine the two results, what will be their respective weights as given by (7.8) and their weighted average as given by (7.9)?

Our analysis of two measurements can be generalized to cover any number of measurements. Suppose we have N separate measurements of a quantity x,

$$x_1 \pm \sigma_1, \quad x_2 \pm \sigma_2, \ldots, \quad x_N \pm \sigma_N,$$

with their corresponding uncertainties $\sigma_1, \sigma_2, \ldots, \sigma_N$. Arguing much as before, we find that the best estimate based on these measurements is the weighted average

$$x_{\text{wav}} = \frac{\sum w_i x_i}{\sum w_i}, \tag{7.10}$$

where the sums are over all N measurements, $i = 1, \ldots, N$, and the *weight* w_i of each measurement is the reciprocal square of the corresponding uncertainty,

$$w_i = \frac{1}{\sigma_i^2} \tag{7.11}$$

for $i = 1, 2, \ldots, N$.

Because the weight $w_i = 1/\sigma_i^2$ associated with each measurement involves the *square* of the corresponding uncertainty σ_i, any measurement that is much less precise than the others contributes *very* much less to the final answer (7.10). For example, if one measurement is four times less precise than the rest, its weight is 16 times less than the other weights, and for many purposes this measurement could simply be ignored.

Because the weighted average x_{wav}, is a function of the original measured values $x_1, x_2, \ldots, x_N$, the uncertainty in x_{wav} can be calculated using error propagation. As you can easily check (Problem 7.8), the uncertainty in x_{wav} is

$$\sigma_{\text{wav}} = \frac{1}{\sqrt{\sum w_i}}. \tag{7.12}$$

This rather ugly result is perhaps a little easier to remember if we rewrite (7.11) as

$$\sigma_i = \frac{1}{\sqrt{w_i}}. \tag{7.13}$$

Paraphrasing Equation (7.13), we can say that the uncertainty in each measurement is the reciprocal square root of its weight. Returning to Equation (7.12), we can paraphrase it similarly to say that the uncertainty in the grand answer x_{wav} is the reciprocal square root of *the sum of all the individual weights;* in other words, the total weight of the final answer is the sum of the individual weights w_i.

Quick Check 7.2. What is the uncertainty in your final answer for Quick Check 7.1?

7.3 An Example

Here is an example involving three separate measurements of the same resistance.

Example: Three Measurements of a Resistance

Each of three students measures the same resistance several times, and their three final answers are (all in ohms):

$$\text{Student 1:} \quad R = 11 \pm 1$$
$$\text{Student 2:} \quad R = 12 \pm 1$$
$$\text{Student 3:} \quad R = 10 \pm 3$$

Given these three results, what is the best estimate for the resistance R?

The three uncertainties σ_1, σ_2, σ_3 are 1, 1, and 3. Therefore, the corresponding weights $w_i = 1/\sigma_i^2$ are

$$w_1 = 1, \qquad w_2 = 1, \qquad w_3 = \tfrac{1}{9}.$$

The best estimate for R is the weighted average, which according to (7.10) is

$$R_{\text{wav}} = \frac{\sum w_i R_i}{\sum w_i}$$
$$= \frac{(1 \times 11) + (1 \times 12) + (\frac{1}{9} \times 10)}{1 + 1 + \frac{1}{9}} = 11.42 \text{ ohms.}$$

The uncertainty in this answer is given by (7.12) as

$$\sigma_{\text{wav}} = \frac{1}{\sqrt{\sum w_i}} = \frac{1}{\sqrt{1 + 1 + \frac{1}{9}}} = 0.69.$$

Thus, our final conclusion (properly rounded) is

$$R = 11.4 \pm 0.7 \text{ ohms.}$$

For interest, let us see what answer we would get if we were to ignore completely the third student's measurement, which is three times less accurate and hence nine times less important. Here, a simple calculation gives $R_{\text{best}} = 11.50$ (compared with 11.42) with an uncertainty of 0.71 (compared with 0.69). Obviously, the third measurement does not have a big effect.

Principal Definitions and Equations of Chapter 7

If $x_1, x_2, \ldots, x_N$ are measurements of a single quantity x, with known uncertainties $\sigma_1, \sigma_2, \ldots, \sigma_N$, then the best estimate for the true value of x is the *weighted average*

$$x_{\text{wav}} = \frac{\sum w_i x_i}{\sum w_i}, \qquad \text{[See (7.10)]}$$

where the sums are over all N measurements, $i = 1, \ldots, N$, and the weights w_i are the reciprocal squares of the corresponding uncertainties,

$$w_i = \frac{1}{\sigma_i^2}.$$

The uncertainty in x_{wav} is

$$\sigma_{\text{wav}} = \frac{1}{\sqrt{\sum w_i}}, \qquad \text{[See (7.12)]}$$

where, again, the sum runs over all of the measurements $i = 1, 2, \ldots, N$.

Problems for Chapter 7

For Section 7.2: The Weighted Average

7.1. ★ Find the best estimate and its uncertainty based on the following four measurements of a certain voltage:

$$1.4 \pm 0.5, \quad 1.2 \pm 0.2, \quad 1.0 \pm 0.25, \quad 1.3 \pm 0.2.$$

7.2. ★ Three groups of particle physicists measure the mass of a certain elementary particle with the results (in units of MeV/c^2):

$$1{,}967.0 \pm 1.0, \quad 1{,}969 \pm 1.4, \quad 1{,}972.1 \pm 2.5.$$

Find the weighted average and its uncertainty.

7.3. ★ **(a)** Two measurements of the speed of sound u give the answers 334 ± 1 and 336 ± 2 (both in m/s). Would you consider them consistent? If so, calculate the best estimate for u and its uncertainty. **(b)** Repeat part (a) for the results 334 ± 1 and 336 ± 5. Is the second measurement worth including in this case?

7.4. ★★ Four measurements are made of the wavelength of light emitted by a certain atom. The results, in nanometers, are:

$$503 \pm 10, \quad 491 \pm 8, \quad 525 \pm 20, \quad 570 \pm 40.$$

Find the weighted average and its uncertainty. Is the last measurement worth including?

7.5. ★★ Two students measure a resistance by different methods. Each makes 10 measurements and computes the mean and its standard deviation, and their final results are as follows:

$$\text{Student A:} \quad R = 72 \pm 8 \text{ ohms}$$
$$\text{Student B:} \quad R = 78 \pm 5 \text{ ohms}.$$

(a) Including both measurements, what are the best estimate of R and its uncertainty? **(b)** Approximately how many measurements (using his same technique) would student A need to make to give his result the same weight as B's? (Remember that each student's final uncertainty is the SDOM, which is equal to the $\text{SD}/\sqrt{N}$.)

7.6. ★★ Two physicists measure the rate of decay of a long-lived radioactive source. Physicist A monitors the sample for 4 hours and observes 412 decays; physicist B monitors it for 6 hours and observes 576 decays. **(a)** Find the uncertainties in these two counts using the square-root rule (3.2). (Remember that the square-root rule gives the uncertainty in the actual counted number.) **(b)** What should each physicist report for the decay rate in decays per hour, with its uncertainty? **(c)** What is the proper weighted average of these two rates, with its uncertainty?

7.7. ★★ Suppose that N separate measurements of a quantity x all have the same uncertainty. Show clearly that in this situation the weighted average (7.10) reduces to the ordinary average, or mean, $\bar{x} = \sum x_i/N$, and that the uncertainty (7.12) reduces to the familiar standard deviation of the mean.

7.8. ★★ The weighted average (7.10) of N separate measurements is a simple function of $x_1, x_2, \ldots, x_N$. Therefore, the uncertainty in x_{wav} can be found by error propagation. Prove in this way that the uncertainty in x_{wav} is as claimed in (7.12).

7.9. ★★★ **(a)** If you have access to a spreadsheet program such as Lotus 123 or Excel, create a spreadsheet to calculate the weighted average of three measurements x_i with given uncertainties σ_i. In the first column, give the trial number i, and use columns 2 and 3 to enter the data x_i and the uncertainties σ_i. In columns 4 and 5, put functions to calculate the weights w_i and the products $w_i x_i$; at the bottoms of these columns, you can calculate the sums $\sum w_i$ and $\sum w_i x_i$. Finally, in some convenient position, place functions to calculate x_{wav} and its uncertainty (7.12). Test your spreadsheet using the data of Section 7.3. **(b)** Try to modify your spreadsheet so that it can handle *any number* of measurements up to some maximum (20 say). (The main difficulty is that you will probably need to use some logical functions to make sure that empty cells in column 3 don't get counted as zeros and cause trouble with the function that calculates $w_i = 1/\sigma_i^2$.) Test your new spreadsheet using the data in Section 7.3 and in Problem 7.1.

Chapter 8

Least-Squares Fitting

Our discussion of the statistical analysis of data has so far focused exclusively on the repeated measurement of one single quantity, not because the analysis of many measurements of one quantity is the most interesting problem in statistics, but because this simple problem must be well understood before more general ones can be discussed. Now we are ready to discuss our first, and very important, more general problem.

8.1 Data That Should Fit a Straight Line

One of the most common and interesting types of experiment involves the measurement of several values of two different physical variables to investigate the mathematical relationship between the two variables. For instance, an experimenter might drop a stone from various different heights $h_1, \ldots, h_N$ and measure the corresponding times of fall $t_1, \ldots, t_N$ to see if the heights and times are connected by the expected relation $h = \frac{1}{2}gt^2$.

Probably the most important experiments of this type are those for which the expected relation is *linear*. For instance, if we believe that a body is falling with constant acceleration g, then its velocity v should be a linear function of the time t,

$$v = v_0 + gt.$$

More generally, we will consider any two physical variables x and y that we suspect are connected by a linear relation of the form

$$y = A + Bx, \tag{8.1}$$

where A and B are constants. Unfortunately, many different notations are used for a linear relation; beware of confusing the form (8.1) with the equally popular $y = ax + b$.

If the two variables y and x are linearly related as in (8.1), then a graph of y against x should be a straight line that has slope B and intersects the y axis at $y = A$. If we were to measure N different values $x_1, \ldots, x_N$ and the corresponding values $y_1, \ldots, y_N$ and if our measurements were subject to no uncertainties, then each of the points (x_i, y_i) would lie exactly on the line $y = A + Bx$, as in Figure 8.1(a). In

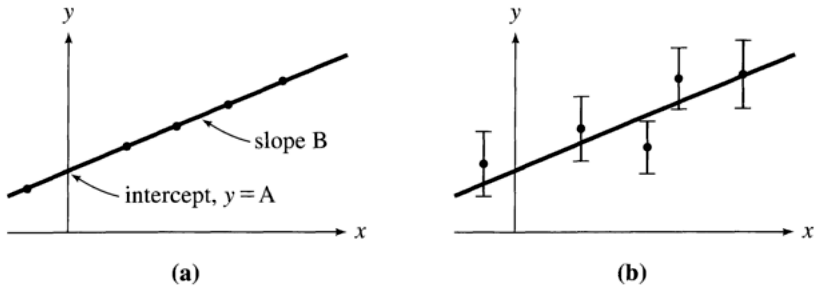


Figure 8.1. **(a)** If the two variables y and x are linearly related as in Equation (8.1), and if there were no experimental uncertainties, then the measured points (x_i, y_i) would all lie exactly on the line $y = A + Bx$. **(b)** In practice, there always are uncertainties, which can be shown by error bars, and the points (x_i, y_i) can be expected only to lie reasonably close to the line. Here, only y is shown as subject to appreciable uncertainties.

practice, there *are* uncertainties, and the most we can expect is that the distance of each point (x_i, y_i) from the line will be reasonable compared with the uncertainties, as in Figure 8.1(b).

When we make a series of measurements of the kind just described, we can ask two questions. First, if we take for granted that y and x *are* linearly related, then the interesting problem is to find the straight line $y = A + Bx$ that best fits the measurements, that is, to find the best estimates for the constants A and B based on the data $(x_1, y_1), \ldots, (x_N, y_N)$. This problem can be approached graphically, as discussed briefly in Section 2.6. It can also be treated analytically, by means of the principle of maximum likelihood. This analytical method of finding the best straight line to fit a series of experimental points is called *linear regression,* or the *least-squares fit for a line,* and is the main subject of this chapter.

The second question that can be asked is whether the measured values $(x_1, y_1), \ldots, (x_N, y_N)$ do really bear out our expectation that y is linear in x. To answer this question, we would first find the line that best fits the data, but we must then devise some measure of *how well* this line fits the data. If we already know the uncertainties in our measurements, we can draw a graph, like that in Figure 8.1(b), that shows the best-fit straight line and the experimental data with their error bars. We can then judge visually whether or not the best-fit line passes sufficiently close to all of the error bars. If we do not know the uncertainties reliably, we must judge how well the points fit a straight line by examining the distribution of the points themselves. We take up this question in Chapter 9.

8.2 Calculation of the Constants *A* and *B*

Let us now return to the question of finding the best straight line $y = A + Bx$ to fit a set of measured points $(x_1, y_1), \ldots, (x_N, y_N)$. To simplify our discussion, we will suppose that, although our measurements of y suffer appreciable uncertainty, the uncertainty in our measurements of x is negligible. This assumption is often reasonable, because the uncertainties in one variable often are much larger than

those in the other, which we can therefore safely ignore. We will further assume that the uncertainties in y all have the same magnitude. (This assumption is also reasonable in many experiments, but if the uncertainties are different, then our analysis can be generalized to weight the measurements appropriately; see Problem 8.9.) More specifically, we assume that the measurement of each y_i is governed by the Gauss distribution, with the same width parameter σ_y for all measurements.

If we knew the constants A and B, then, for any given value x_i (which we are assuming has no uncertainty), we could compute the true value of the corresponding y_i,

$$(\text{true value for } y_i) \;=\; A + Bx_i. \tag{8.2}$$

The measurement of y_i is governed by a normal distribution centered on this true value, with width parameter σ_y. Therefore, the probability of obtaining the observed value y_i is

$$Prob_{A,B}(y_i) \;\propto\; \frac{1}{\sigma_y} e^{-(y_i - A - Bx_i)^2/2\sigma_y^2}, \tag{8.3}$$

where the subscripts A and B indicate that this probability depends on the (unknown) values of A and B. The probability of obtaining our complete set of measurements $y_1, \ldots, y_N$ is the product

$$\begin{aligned} Prob_{A,B}(y_1, \ldots, y_N) \;&=\; Prob_{A,B}(y_1) \cdots Prob_{A,B}(y_N) \\ &\propto\; \frac{1}{\sigma_y^N} e^{-\chi^2/2}, \end{aligned} \tag{8.4}$$

where the exponent is given by

$$\chi^2 \;=\; \sum_{i=1}^{N} \frac{(y_i - A - Bx_i)^2}{\sigma_y^2}. \tag{8.5}$$

In the now-familiar way, we will assume that the best estimates for the unknown constants A and B, based on the given measurements, are those values of A and B for which the probability $Prob_{A,B}(y_1, \ldots, y_N)$ is maximum, or for which the sum of squares χ^2 in (8.5) is a minimum. (This is why the method is known as least-squares fitting.) To find these values, we differentiate χ^2 with respect to A and B and set the derivatives equal to zero:

$$\frac{\partial \chi^2}{\partial A} \;=\; \frac{-2}{\sigma_y^2} \sum_{i=1}^{N} (y_i - A - Bx_i) \;=\; 0 \tag{8.6}$$

and

$$\frac{\partial \chi^2}{\partial B} \;=\; \frac{-2}{\sigma_y^2} \sum_{i=1}^{N} x_i(y_i - A - Bx_i) \;=\; 0. \tag{8.7}$$

These two equations can be rewritten as simultaneous equations for A and B:

$$AN + B\Sigma x_i \;=\; \Sigma y_i \tag{8.8}$$

and

$$A\Sigma x_i + B\Sigma x_i^2 \;=\; \Sigma x_i y_i. \tag{8.9}$$

Here, I have omitted the limits $i = 1$ to N from the summation signs Σ. In the following discussion, I also omit the subscripts i when there is no serious danger of confusion; thus, $\Sigma x_i y_i$ is abbreviated to Σxy and so on.

The two equations (8.8) and (8.9), sometimes called *normal equations,* are easily solved for the least-squares estimates for the constants A and B,

$$A = \frac{\Sigma x^2 \Sigma y - \Sigma x \Sigma xy}{\Delta} \tag{8.10}$$

and

$$B = \frac{N\Sigma xy - \Sigma x \Sigma y}{\Delta}, \tag{8.11}$$

where I have introduced the convenient abbreviation for the denominator,

$$\Delta = N\Sigma x^2 - (\Sigma x)^2. \tag{8.12}$$

The results (8.10) and (8.11) give the best estimates for the constants A and B of the straight line $y = A + Bx$, based on the N measured points $(x_1, y_1), \ldots, (x_N, y_N)$. The resulting line is called the *least-squares fit* to the data, or the *line of regression* of y on x.

Example: Length versus Mass for a Spring Balance

A student makes a scale to measure masses with a spring. She attaches its top end to a rigid support, hangs a pan from its bottom, and places a meter stick behind the arrangement to read the length of the spring. Before she can use the scale, she must calibrate it; that is, she must find the relationship between the mass in the pan and the length of the spring. To do this calibration, she gets five accurate 2-kg masses, which she adds to the pan one by one, recording the corresponding lengths l_i as shown in the first three columns of Table 8.1. Assuming the spring obeys Hooke's law, she anticipates that l should be a linear function of m,

$$l = A + Bm. \tag{8.13}$$

(Here, the constant A is the unloaded length of the spring, and B is g/k, where k is the usual spring constant.) The calibration equation (8.13) will let her find any unknown mass m from the corresponding length l, once she knows the constants A and B. To find these constants, she uses the method of least squares. What are her answers for A and B? Plot her calibration data and the line given by her best fit (8.13). If she puts an unknown mass m in the pan and observes the spring's length to be $l = 53.2$ cm, what is m?

Table 8.1. Masses m_i (in kg) and lengths l_i (in cm) for a spring balance. The "x" and "y" in quotes indicate which variables play the roles of x and y in this example.

Trial number i	"x" Load, m_i	"y" Length, l_i	m_i^2	$m_i l_i$
1	2	42.0	4	84
2	4	48.4	16	194
3	6	51.3	36	308
4	8	56.3	64	450
5	10	58.6	100	586
$N = 5$	$\Sigma m_i = 30$	$\Sigma l_i = 256.6$	$\Sigma m_i^2 = 220$	$\Sigma m_i l_i = 1{,}622$

As often happens in such problems, the two variables are not called x and y, and one must be careful to identify which is which. Comparing (8.13) with the standard form, $y = A + Bx$, we see that the length l plays the role of the dependent variable y, while the mass m plays the role of the independent variable x. The constants A and B are given by (8.10) through (8.12), with the replacements

$$x_i \leftrightarrow m_i \qquad \text{and} \qquad y_i \leftrightarrow l_i.$$

(This correspondence is indicated by the headings "x" and "y" in Table 8.1.) To find A and B, we need to find the sums Σm_i, Σl_i, Σm_i^2, and $\Sigma m_i l_i$; therefore, the last two columns of Table 8.1 show the quantities m_i^2 and $m_i l_i$, and the corresponding sum is shown at the bottom of each column.

Computing the constants A and B is now straightforward. According to (8.12), the denominator Δ is

$$\begin{aligned} \Delta &= N\Sigma m^2 - (\Sigma m)^2 \\ &= 5 \times 220 - 30^2 = 200. \end{aligned}$$

Next, from (8.10) we find the intercept (the unstretched length)

$$\begin{aligned} A &= \frac{\Sigma m^2 \Sigma l - \Sigma m \Sigma ml}{\Delta} \\ &= \frac{220 \times 256.6 - 30 \times 1622}{200} = 39.0 \text{ cm}. \end{aligned}$$

Finally, from (8.11) we find the slope

$$\begin{aligned} B &= \frac{N\Sigma ml - \Sigma m \Sigma l}{\Delta} \\ &= \frac{5 \times 1622 - 30 \times 256.6}{200} = 2.06 \text{ cm/kg}. \end{aligned}$$

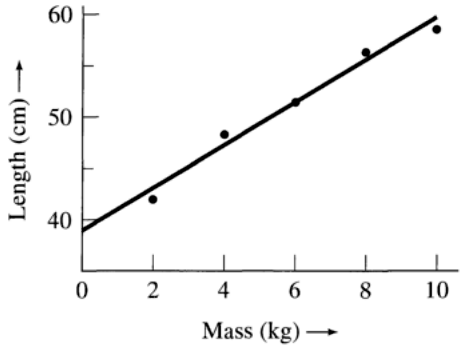


Figure 8.2. A plot of the data from Table 8.1 and the best-fit line (8.13).

A plot of the data and the line (8.13) using these values of A and B is shown in Figure 8.2. If the mass m stretches the spring to 53.2 cm, then according to (8.13) the mass is

$$m = \frac{l - A}{B} = \frac{(53.2 - 39.0)\text{ cm}}{2.06\text{ cm/kg}} = 6.9\text{ kg}.$$

Quick Check 8.1. Find the least-squares best-fit line $y = A + Bx$ through the three points $(x, y) = (-1, 0)$, $(0, 6)$, and $(1, 6)$. Using squared paper, plot the points and your line. [Note that because the three values of x (-1, 0, and 1) are symmetric about zero, $\Sigma x = 0$, which simplifies the calculation of A and B. In some experiments, the values of x can be arranged to be symmetrically spaced in this way, which saves some trouble.]

Now that we know how to find the best estimates for the constants A and B, we naturally ask for the uncertainties in these estimates. Before we can find these uncertainties, however, we must discuss the uncertainty σ_y in the original measurements of $y_1, y_2, \ldots, y_N$.

8.3 Uncertainty in the Measurements of y

In the course of measuring the values $y_1, \ldots, y_N$, we have presumably formed some idea of their uncertainty. Nonetheless, knowing how to calculate the uncertainty by analyzing the data themselves is important. Remember that the numbers $y_1, \ldots, y_N$ are *not* N measurements of the same quantity. (They might, for instance, be the times for a stone to fall from N different heights.) Thus, we certainly do not get an idea of their reliability by examining the spread in their values.

Nevertheless, we can easily estimate the uncertainty σ_y in the numbers $y_1, \ldots, y_N$. The measurement of each y_i is (we are assuming) normally distributed about its true value $A + Bx_i$, with width parameter σ_y. Thus the *deviations* $y_i - A - Bx_i$ are

normally distributed, all with the same central value zero and the same width σ_y. This situation immediately suggests that a good estimate for σ_y would be given by a sum of squares with the familiar form

$$\sigma_y = \sqrt{\frac{1}{N}\Sigma(y_i - A - Bx_i)^2}. \tag{8.14}$$

In fact, this answer can be confirmed by means of the principle of maximum likelihood. As usual, the best estimate for the parameter in question (σ_y here) is that value for which the probability (8.4) of obtaining the observed values $y_1, \ldots, y_N$ is maximum. As you can easily check by differentiating (8.4) with respect to σ_y and setting the derivative equal to zero, this best estimate is precisely the answer (8.14). (See Problem 8.12.)

Unfortunately, as you may have suspected, the estimate (8.14) for σ_y is not quite the end of the story. The numbers A and B in (8.14) are the unknown true values of the constants A and B. In practice, these numbers must be replaced by our *best estimates* for A and B, namely, (8.10) and (8.11), and this replacement slightly reduces the value of (8.14). It can be shown that this reduction is compensated for if we replace the factor N in the denominator by $(N - 2)$. Thus, our final answer for the uncertainty in the measurements $y_1, \ldots, y_N$ is

$$\sigma_y = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N}(y_i - A - Bx_i)^2}, \tag{8.15}$$

with A and B given by (8.10) and (8.11). If we already have an independent estimate of our uncertainty in $y_1, \ldots, y_N$, we would expect this estimate to compare with σ_y as computed from (8.15).

I will not attempt to justify the factor of $(N - 2)$ in (8.15) but can make some comments. First, as long as N is moderately large, the difference between N and $(N - 2)$ is unimportant anyway. Second, that the factor $(N - 2)$ is *reasonable* becomes clear if we consider measuring just two pairs of data (x_1, y_1) and (x_2, y_2). With only two points, we can always find a line that passes *exactly* through both points, and the least-squares fit will give this line. That is, with just two pairs of data, we cannot possibly deduce anything about the reliability of our measurements. Now, since both points lie exactly on the best line, the two terms of the sum in (8.14) and (8.15) are zero. Thus, the formula (8.14) (with $N = 2$ in the denominator) would give the absurd answer $\sigma_y = 0$; whereas (8.15), with $N - 2 = 0$ in the denominator, gives $\sigma_y = 0/0$, indicating correctly that σ_y is undetermined after only two measurements.

The presence of the factor $(N - 2)$ in (8.15) is reminiscent of the $(N - 1)$ that appeared in our estimate of the standard deviation of N measurements of one quantity x, in Equation (5.45). There, we made N measurements $x_1, \ldots, x_N$ of the one quantity x. Before we could calculate σ_x, we had to use our data to find the mean $\bar{x}$. In a certain sense, this computation left only $(N - 1)$ independent measured values, so we say that, having computed $\bar{x}$, we have only $(N - 1)$ *degrees of freedom* left. Here, we made N measurements, but before calculating σ_y we had to compute the *two* quantities A and B. Having done this, we had only $(N - 2)$ degrees of

freedom left. In general, we define the *number of degrees of freedom* at any stage in a statistical calculation as the number of independent measurements *minus* the number of parameters calculated from these measurements. We can show (but will not do so here) that the number of degrees of freedom, *not* the number of measurements, is what should appear in the denominator of formulas such as (8.15) and (5.45). This fact explains why (8.15) contains the factor $(N - 2)$ and (5.45) the factor $(N - 1)$.

8.4 Uncertainty in the Constants *A* and *B*

Having found the uncertainty σ_y in the measured numbers $y_1, \ldots, y_N$, we can easily return to our estimates for the constants A and B and calculate their uncertainties. The point is that the estimates (8.10) and (8.11) for A and B are well-defined functions of the measured numbers $y_1, \ldots, y_N$. Therefore, the uncertainties in A and B are given by simple error propagation in terms of those in $y_1, \ldots, y_N$. I leave it as an exercise for you to check (Problem 8.16) that

$$\sigma_A = \sigma_y \sqrt{\frac{\Sigma x^2}{\Delta}} \tag{8.16}$$

and

$$\sigma_B = \sigma_y \sqrt{\frac{N}{\Delta}} \tag{8.17}$$

where Δ is given by (8.12) as usual.

The results of this and the previous two sections were based on the assumptions that the measurements of y were all equally uncertain and that any uncertainties in x were negligible. Although these assumptions often are justified, we need to discuss briefly what happens when they are not. First, if the uncertainties in y are not all equal, we can use the method of *weighted least squares,* as described in Problem 8.9. Second, if there are uncertainties in x but not in y, we can simply interchange the roles of x and y in our analysis. The remaining case is that in which both x and y have uncertainties—a case that certainly *can* occur. The least-squares fitting of a general curve when both x and y have uncertainties is rather complicated and even controversial. In the important special case of a *straight line* (which is all we have discussed so far), uncertainties in both x and y make surprisingly little difference, as we now discuss.

Suppose, first, that our measurements of x are subject to uncertainty but those of y are not, and we consider a particular measured point (x, y). This point and the true line $y = A + Bx$ are shown in Figure 8.3. The point (x, y) does not lie on the line because of the error—call it Δx—in our measurement of x. Now, we can see

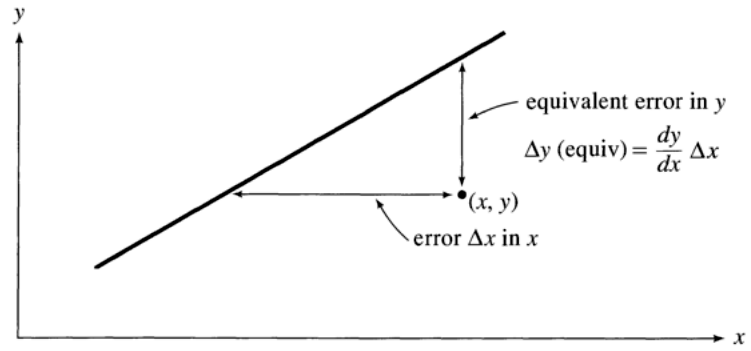


Figure 8.3. A measured point (x, y) and the line $y = A + Bx$ on which the point is supposed to lie. The error Δx in x, with y exact, produces the same effect as an error $\Delta y(\text{equiv}) = (dy/dx)\Delta x$ in y, with x exact. (Here, dy/dx denotes the slope of the expected line.)

easily from the picture that the error Δx in x, with no error in y, produces exactly the same effect as if there had been no error in x but an error in y given by

$$\Delta y(\text{equiv}) = \frac{dy}{dx}\Delta x \tag{8.18}$$

(where "equiv" stands for "equivalent"). The standard deviation σ_x is just the root-mean-square value of Δx that would result from repeating this measurement many times. Thus, according to (8.18), the problem with uncertainties σ_x in x can be replaced with an equivalent problem with uncertainties in y, given by

$$\sigma_y(\text{equiv}) = \frac{dy}{dx}\sigma_x. \tag{8.19}$$

The result (8.19) is true whatever the curve of y vs x, but (8.19) is especially simple if the curve is a straight line, because the slope dy/dx is just the constant B. Therefore, for a straight line

$$\sigma_y(\text{equiv}) = B\sigma_x. \tag{8.20}$$

In particular, if all the uncertainties σ_x are equal, the same is true of the equivalent uncertainties $\sigma_y(\text{equiv})$. Therefore, the problem of fitting a line to points (x_i, y_i) with equal uncertainties in x but no uncertainties in y is equivalent to the problem of equal uncertainties in y but none in x. This equivalence means we can safely use the method already described for either problem. [In practice, the points do not lie *exactly* on the line, and the two "equivalent" problems will not give *exactly* the same line. Nevertheless, the two lines should usually agree within the uncertainties given by (8.16) and (8.17). See Problem 8.17.]

We can now extend this argument to the case that *both* x and y have uncertainties. The uncertainty in x is equivalent to an uncertainty in y as given by (8.20). In addition, y is already subject to its own uncertainty σ_y. These two uncertainties are

independent and must be combined in quadrature. Thus, the original problem, with uncertainties in both x and y, can be replaced with an equivalent problem in which only y has uncertainty, given by

$$\sigma_y(\text{equiv}) = \sqrt{\sigma_y^2 + (B\sigma_x)^2}. \tag{8.21}$$

Provided all the uncertainties σ_x are the same, and likewise all the uncertainties σ_y, the equivalent uncertainties (8.21) are all the same, and we can safely use the formulas (8.10) through (8.17).

If the uncertainties in x (or in y) are not all the same, we can still use (8.21), but the resulting uncertainties will not all be the same, and we will need to use the method of weighted least squares. If the curve to which we are fitting our points is not a straight line, a further complication arises because the slope dy/dx is not a constant and we cannot replace (8.19) with (8.20). Nevertheless, we can still use (8.21) (with dy/dx in place of B) to replace the original problem (with uncertainties in both x and y) by an equivalent problem in which only y has uncertainties as given by (8.21).[1]

8.5 An Example

Here is a simple example of least-squares fitting to a straight line; it involves the constant-volume gas thermometer.

Example: Measurement of Absolute Zero with a Constant-Volume Gas Thermometer

If the volume of a sample of an ideal gas is kept constant, its temperature T is a linear function of its pressure P,

$$T = A + BP. \tag{8.22}$$

Here, the constant A is the temperature at which the pressure P would drop to zero (if the gas did not condense into a liquid first); it is called the *absolute zero of temperature,* and has the accepted value

$$A = -273.15°\text{C}$$

The constant B depends on the nature of the gas, its mass, and its volume.[2] By measuring a series of values for T and P, we can find the best estimates for the constants A and B. In particular, the value of A gives the absolute zero of temperature.

One set of five measurements of P and T obtained by a student was as shown in the first three columns of Table 8.2. The student judged that his measurements of

[1] This procedure is quite complicated in practice. Before we can use (8.21) to find the uncertainty σ_y(equiv), we need to know the slope B (or, more generally, dy/dx), which is not known until we have solved the problem! Nevertheless, we can get a reasonable first approximation for the slope using the method of unweighted least squares, ignoring all of the complications discussed here. This method gives an approximate value for the slope B, which can then be used in (8.21) to give a reasonable approximation for σ_y(equiv).

[2] The difference $T - A$ is called the *absolute temperature.* Thus (8.22) can be rewritten to say that the absolute temperature is proportional to the pressure (at constant volume).

Table 8.2. Pressure (in mm of mercury) and temperature (°C) of a gas at constant volume.

Trial number i	"x" Pressure P_i	"y" Temperature T_i	$A + BP_i$
1	65	−20	−22.2
2	75	17	14.9
3	85	42	52.0
4	95	94	89.1
5	105	127	126.2

P had negligible uncertainty, and those of T were all equally uncertain with an uncertainty of "a few degrees." Assuming his points should fit a straight line of the form (8.22), he calculated his best estimate for the constant A (the absolute zero) and its uncertainty. What should have been his conclusions?

All we have to do here is use formulas (8.10) and (8.16), with x_i replaced by P_i and y_i by T_i, to calculate all the quantities of interest. This requires us to compute the sums ΣP, ΣP^2, ΣT, ΣPT. Many pocket calculators can evaluate all these sums automatically, but even without such a machine, we can easily handle these calculations if the data are properly organized. From Table 8.2, we can calculate

$$\begin{aligned} \Sigma P &= 425, \\ \Sigma P^2 &= 37{,}125, \\ \Sigma T &= 260, \\ \Sigma PT &= 25{,}810, \\ \Delta &= N\Sigma P^2 - (\Sigma P)^2 = 5{,}000. \end{aligned}$$

In this kind of calculation, it is important to keep plenty of significant figures because we have to take differences of these large numbers. Armed with these sums, we can immediately calculate the best estimates for the constants A and B:

$$A = \frac{\Sigma P^2 \Sigma T - \Sigma P \Sigma PT}{\Delta} = -263.35$$

and

$$B = \frac{N \Sigma PT - \Sigma P \Sigma T}{\Delta} = 3.71.$$

This calculation already gives the student's best estimate for absolute zero, $A = -263$°C.

Knowing the constants A and B, we can next calculate the numbers $A + BP_i$, the temperatures "expected" on the basis of our best fit to the relation $T = A + BP$. These numbers are shown in the far right column of the table, and as we would hope, all agree reasonably well with the observed temperatures. We can now take the difference between the figures in the last two columns and calculate

$$\sigma_T = \sqrt{\frac{1}{N-2} \Sigma (T_i - A - BP_i)^2} = 6.7.$$

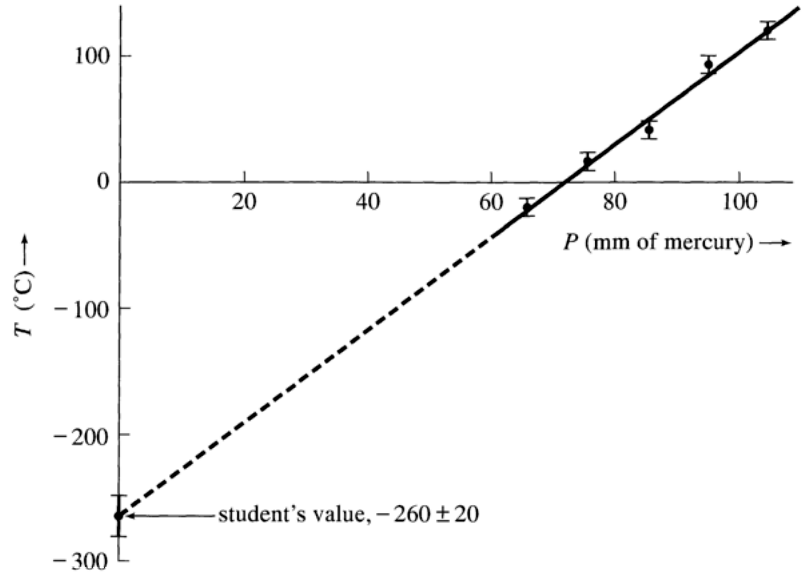


Figure 8.4. Graph of temperature T vs pressure P for a gas at constant volume. The error bars extend one standard deviation, σ_T, on each side of the five experimental points, and the line is the least-squares best fit. The absolute zero of temperature was found by extrapolating the line back to its intersection with the T axis.

This result agrees reasonably with the student's estimate that his temperature measurements were uncertain by "a few degrees."

Finally, we can calculate the uncertainty in A using (8.16):

$$\sigma_A = \sigma_T\sqrt{\Sigma P^2/\Delta} = 18.$$

Thus, our student's final conclusion, suitably rounded, should be

$$\text{absolute zero, } A = -260 \pm 20°\text{C},$$

which agrees satisfactorily with the accepted value, −273°C.

As is often true, these results become much clearer if we graph them, as in Figure 8.4. The five data points, with their uncertainties of ± 7° in T, are shown on the upper right. The best straight line passes through four of the error bars and close to the fifth.

To find a value for absolute zero, the line was extended beyond all the data points to its intersection with the T axis. This process of *extrapolation* (extending a curve beyond the data points that determine it) can introduce large uncertainties, as is clear from the picture. A very small change in the line's slope will cause a large change in its intercept on the distant T axis. Thus, any uncertainty in the data is greatly magnified if we have to extrapolate any distance. This magnification explains why the uncertainty in the value of absolute zero (±18°) is so much larger than that in the original temperature measurements (±7°).

8.6 Least-Squares Fits to Other Curves

So far in this chapter, we have considered the observation of two variables satisfying a linear relation, $y = A + Bx$, and we have discussed the calculation of the constants A and B. This important problem is a special case of a wide class of curve-fitting problems, many of which can be solved in a similar way. In this section, I mention briefly a few more of these problems.

FITTING A POLYNOMIAL

Often, one variable, y, is expected to be expressible as a polynomial in a second variable, x,

$$y = A + Bx + Cx^2 + \cdots + Hx^n. \tag{8.23}$$

For example, the height y of a falling body is expected to be quadratic in the time t,

$$y = y_0 + v_0 t - \tfrac{1}{2} g t^2,$$

where y_0 and v_0 are the initial height and velocity, and g is the acceleration of gravity. Given a set of observations of the two variables, we can find best estimates for the constants $A, B, \ldots, H$ in (8.23) by an argument that exactly parallels that of Section 8.2, as I now outline.

To simplify matters, we suppose that the polynomial (8.23) is actually a quadratic,

$$y = A + Bx + Cx^2. \tag{8.24}$$

(You can easily extend the analysis to the general case if you wish.) We suppose, as before, that we have a series of measurements (x_i, y_i), $i = 1, \ldots, N$, with the y_i all equally uncertain and the x_i all exact. For each x_i, the corresponding true value of y_i is given by (8.24), with A, B, and C as yet unknown. We assume that the measurements of the y_i are governed by normal distributions, each centered on the appropriate true value and all with the same width σ_y. This assumption lets us compute the probability of obtaining our observed values $y_1, \ldots, y_N$ in the familiar form

$$Prob(y_1, \ldots, y_N) \propto e^{-\chi^2/2}, \tag{8.25}$$

where now

$$\chi^2 = \sum_{i=1}^{N} \frac{(y_i - A - Bx_i - Cx_i^2)^2}{\sigma_y^2}. \tag{8.26}$$

[This equation corresponds to Equation (8.5) for the linear case.] The best estimates for A, B, and C are those values for which $Prob(y_1, \ldots, y_N)$ is largest, or χ^2 is smallest. Differentiating χ^2 with respect to A, B, and C and setting these derivatives equal to zero, we obtain the three equations (as you should check; see Problem 8.21):

$$\begin{aligned} AN + B\Sigma x + C\Sigma x^2 &= \Sigma y, \\ A\Sigma x + B\Sigma x^2 + C\Sigma x^3 &= \Sigma xy, \\ A\Sigma x^2 + B\Sigma x^3 + C\Sigma x^4 &= \Sigma x^2 y. \end{aligned} \tag{8.27}$$

For any given set of measurements (x_i, y_i), these simultaneous equations for A, B, and C (known as the *normal equations*) can be solved to find the best estimates for A, B, and C. With A, B, and C calculated in this way, the equation $y = A + Bx + Cx^2$ is called the least-squares polynomial fit, or the polynomial regression, for the given measurements. (For an example, see Problem 8.22.)

The method of polynomial regression generalizes easily to a polynomial of any degree, although the resulting normal equations become cumbersome for polynomials of high degree. In principle, a similar method can be applied to *any* function $y = f(x)$ that depends on various unknown parameters A, B, ... Unfortunately, the resulting normal equations that determine the best estimates for A, B, . . . can be difficult or impossible to solve. However, one large class of problems *can* always be solved, namely, those problems in which the function $y = f(x)$ depends linearly on the parameters A, B, ... These include all polynomials (obviously the polynomial (8.23) is linear in its coefficients A, B, . . .) but they also include many other functions. For example, in some problems y is expected to be a sum of trigonometric functions, such as

$$y = A \sin x + B \cos x. \tag{8.28}$$

For this function, and in fact for any function that is linear in the parameters A, B, . . . , the normal equations that determine the best estimates for A, B, . . . are simultaneous linear equations, which can always be solved (see Problems 8.23 and 8.24).

EXPONENTIAL FUNCTIONS

One of the most important functions in physics is the exponential function

$$y = Ae^{Bx}, \tag{8.29}$$

where A and B are constants. The intensity I of radiation, after passing a distance x through a shield, falls off exponentially:

$$I = I_0 e^{-\mu x},$$

where I_0 is the original intensity and μ characterizes the absorption by the shield. The charge on a short-circuited capacitor drains away exponentially:

$$Q = Q_0 e^{-\lambda t}$$

where Q_0 is the original charge and $\lambda = 1/(RC)$, where R and C are the resistance and capacitance.

If the constants A and B in (8.29) are unknown, we naturally seek estimates of them based on measurements of x and y. Unfortunately, direct application of our previous arguments leads to equations for A and B that cannot be conveniently solved. We can, however, transform the nonlinear relation (8.29) between y and x into a linear relation, to which we can apply our least-squares fit.

To effect the desired "linearization," we simply take the natural logarithm of (8.29) to give

$$\ln y = \ln A + Bx. \tag{8.30}$$

We see that, even though y is not linear in x, $\ln y$ *is*. This conversion of the nonlinear (8.29) into the linear (8.30) is useful in many contexts besides that of least-squares fitting. If we want to check the relation (8.29) graphically, then a direct plot of y against x will produce a curve that is hard to identify visually. On the other hand, a plot of $\ln y$ against x (or of $\log y$ against x) should produce a straight line, which can be identified easily. (Such a plot is especially easy if you use "semilog" graph paper, on which the graduations on one axis are spaced logarithmically. Such paper lets you plot $\log y$ directly without even calculating it.)

The usefulness of the linear equation (8.30) in least-squares fitting is readily apparent. If we believe that y and x should satisfy $y = Ae^{Bx}$, then the variables $z = \ln y$ and x should satisfy (8.30), or

$$z = \ln A + Bx. \tag{8.31}$$

If we have a series of measurements (x_i, y_i), then for each y_i we can calculate $z_i = \ln y_i$. Then the pairs (x_i, z_i) should lie on the line (8.31). This line can be fitted by the method of least squares to give best estimates for the constants $\ln A$ (from which we can find A) and B.

Example: A Population of Bacteria

Many populations (of people, bacteria, radioactive nuclei, etc.) tend to vary exponentially over time. If a population N is decreasing exponentially, we write

$$N = N_0 e^{-t/\tau}, \tag{8.32}$$

where τ is called the population's *mean life* [closely related to the *half-life*, $t_{1/2}$; in fact, $t_{1/2} = (\ln 2)\tau$]. A biologist suspects that a population of bacteria is decreasing exponentially as in (8.32) and measures the population on three successive days; he obtains the results shown in the first two columns of Table 8.3. Given these data, what is his best estimate for the mean life τ?

Table 8.3. Population of bacteria.

Time t_i (days)	Population N_i	$z_i = \ln N_i$
0	153,000	11.94
1	137,000	11.83
2	128,000	11.76

If N varies as in (8.32), then the variable $z = \ln N$ should be linear in t:

$$z = \ln N = \ln N_0 - \frac{t}{\tau}. \tag{8.33}$$

Our biologist therefore calculates the three numbers $z_i = \ln N_i$ $(i = 0, 1, 2)$ shown in the third column of Table 8.3. Using these numbers, he makes a least-squares fit to the straight line (8.33) and finds as best estimates for the coefficients $\ln N_0$ and $(-1/\tau)$,

$$\ln N_0 = 11.93 \quad \text{and} \quad (-1/\tau) = -0.089 \text{ day}^{-1}.$$

The second of these coefficients implies that his best estimate for the mean life is

$$\tau = 11.2 \text{ days.}$$

The method just described is attractively simple (especially with a calculator that performs linear regression automatically) and is frequently used. Nevertheless, the method is not quite logically sound. Our derivation of the least-squares fit to a straight line $y = A + Bx$ was based on the assumption that the measured values $y_1, \ldots, y_N$ were all equally uncertain. Here, we are performing our least-squares fit using the variable $z = \ln y$. Now, if the measured values y_i are all equally uncertain, then the values $z_i = \ln y_i$ are *not*. In fact, from simple error propagation we know that

$$\sigma_z = \left|\frac{dz}{dy}\right| \sigma_y = \frac{\sigma_y}{y}. \tag{8.34}$$

Thus, if σ_y is the same for all measurements, then σ_z varies (with σ_z larger when y is smaller). Evidently, the variable $z = \ln y$ does not satisfy the requirement of equal uncertainties for all measurements, if y itself does.

The remedy for this difficulty is straightforward. The least-squares procedure can be modified to allow for different uncertainties in the measurements, provided the various uncertainties are known. (This method of *weighted least squares* is outlined in Problem 8.9). If we know that the measurements of $y_1, \ldots, y_N$ really are equally uncertain, then Equation (8.34) tells us how the uncertainties in $z_1, \ldots, z_N$ vary, and we can therefore apply the method of weighted least squares to the equation $z = \ln A + Bx$.

In practice, we often cannot be sure that the uncertainties in $y_1, \ldots, y_N$ really are constant; so we can perhaps argue that we could just as well assume the uncertainties in $z_1, \ldots, z_N$ to be constant and use the simple unweighted least squares. Often the variation in the uncertainties is small, and which method is used makes little difference, as in the preceding example. In any event, when the uncertainties are unknown, straightforward application of the ordinary (unweighted) least-squares fit is an unambiguous and simple way to get *reasonable* (if not *best*) estimates for the constants A and B in the equation $y = Ae^{Bx}$, so it is frequently used in this way.

MULTIPLE REGRESSION

Finally, we have so far discussed only observations of *two* variables, x and y, and their relationship. Many real problems, however, have more than two variables to be considered. For example, in studying the pressure P of a gas, we find that it depends on the volume V and temperature T, and we must analyze P as a function of V and T. The simplest example of such a problem is when one variable, z, depends linearly on two others, x and y:

$$z = A + Bx + Cy. \tag{8.35}$$

This problem can be analyzed by a very straightforward generalization of our two-variable method. If we have a series of measurements (x_i, y_i, z_i), $i = 1, \ldots, N$ (with the z_i all equally uncertain, and the x_i and y_i exact), then we can use the principle of maximum likelihood exactly as in Section 8.2 to show that the best estimates for the constants A, B, and C are determined by normal equations of the form

$$\begin{aligned} AN + B\Sigma x + C\Sigma y &= \Sigma z, \\ A\Sigma x + B\Sigma x^2 + C\Sigma xy &= \Sigma xz, \\ A\Sigma y + B\Sigma xy + C\Sigma y^2 &= \Sigma yz. \end{aligned} \tag{8.36}$$

The equations can be solved for A, B, and C to give the best fit for the relation (8.35). This method is called *multiple regression* ("multiple" because there are more than two variables), but we will not discuss it further here.

Principal Definitions and Equations of Chapter 8

Throughout this chapter, we have considered N pairs of measurements $(x_1, y_1), \ldots, (x_N, y_N)$ of two variables x and y. The problem addressed was finding the best values of the parameters of the curve that a graph of y vs x is expected to fit. We assume that only the measurements of y suffered appreciable uncertainties, whereas those for x were negligible. [For the case in which both x and y have significant uncertainties, see the discussion following Equation (8.17).] Various possible curves can be analyzed, and there are two different assumptions about the uncertainties in y. Some of the more important cases are as follows:

A STRAIGHT LINE, $y = A + Bx$; EQUAL WEIGHTS

If y is expected to lie on a straight line $y = A + Bx$, and if the measurements of y all have the same uncertainties, then the best estimates for the constants A and B are:

$$A = \frac{\Sigma x^2 \Sigma y - \Sigma x \Sigma xy}{\Delta}$$

and

$$B = \frac{N\Sigma xy - \Sigma x \Sigma y}{\Delta},$$

where the denominator, Δ, is

$$\Delta = N\Sigma x^2 - (\Sigma x)^2. \qquad \text{[See (8.10) to (8.12)]}$$

Based on the observed points, the best estimate for the uncertainty in the measurements of y is

$$\sigma_y = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N}(y_i - A - Bx_i)^2}. \qquad \text{[See (8.15)]}$$

The uncertainties in A and B are:

$$\sigma_A = \sigma_y \sqrt{\frac{\Sigma x^2}{\Delta}}$$

and

$$\sigma_B = \sigma_y \sqrt{\frac{N}{\Delta}}. \qquad \text{[See (8.16) \& (8.17)]}$$

STRAIGHT LINE THROUGH THE ORIGIN ($y = Bx$); EQUAL WEIGHTS

If y is expected to lie on a straight line through the origin, $y = Bx$, and if the measurements of y all have the same uncertainties, then the best estimate for the constant B is:

$$B = \frac{\Sigma xy}{\Sigma x^2}. \qquad \text{[See Problem 8.5]}$$

Based on the measured points, the best estimate for the uncertainty in the measurements of y is:

$$\sigma_y = \sqrt{\frac{1}{N-1}\Sigma(y_i - Bx_i)^2}$$

and the uncertainty in B is:

$$\sigma_B = \frac{\sigma_y}{\sqrt{\Sigma x^2}}. \qquad \text{[See Problem 8.18]}$$

WEIGHTED FIT FOR A STRAIGHT LINE, $y = A + Bx$

If y is expected to lie on a straight line $y = A + Bx$, and if the measured values y_i have different, known uncertainties σ_i, then we introduce the *weights* $w_i = 1/\sigma_i^2$, and the best estimates for the constants A and B are:

$$A = \frac{\Sigma wx^2 \, \Sigma wy - \Sigma wx \, \Sigma wxy}{\Delta}$$

and

$$B = \frac{\Sigma w \, \Sigma wxy - \Sigma wx \, \Sigma wy}{\Delta},$$

where

$$\Delta = \Sigma w \, \Sigma wx^2 - (\Sigma wx)^2. \qquad \text{[See Problem 8.9]}$$

The uncertainties in A and B are:

$$\sigma_A = \sqrt{\frac{\Sigma wx^2}{\Delta}}$$

and

$$\sigma_B = \sqrt{\frac{\Sigma w}{\Delta}}. \qquad \text{[See Problem 8.19]}$$

OTHER CURVES

If y is expected to be a polynomial in x, that is,

$$y = A + Bx + \ldots + Hx^n,$$

then an exactly analogous method of least-squares fitting can be used, although the equations are quite cumbersome if n is large. (For examples, see Problems 8.21 and 8.22.) Curves of the form

$$y = Af(x) + Bg(x) + \ldots + Hk(x),$$

where $f(x), \ldots, k(x)$ are known functions, can also be handled in the same way. (For examples, see Problems 8.23 and 8.24.)

If y is expected to be given by the exponential function

$$y = Ae^{Bx},$$

then we can "linearize" the problem by using the variable $z = \ln(y)$, which should satisfy the linear relation

$$z = \ln(y) = \ln(A) + Bx. \qquad \text{[See (8.31)]}$$

We can then apply the linear least-squares fit to z as a function of x. Note, however, that if the uncertainties in the measured values of y are all equal, the same is certainly *not* true of the values of z. Then, strictly speaking, the method of weighted least squares should be used. (See Problem 8.26 for an example.)

Problems for Chapter 8

For Section 8.2: Calculation of the Constants *A* and *B*

8.1. ★ Use the method of least squares to find the line $y = A + Bx$ that best fits the three points (1, 6), (3, 5), and (5, 1). Using squared paper, plot the three points and your line. Your calculator probably has a built-in function to calculate A and B; if you don't know how to use it, take a moment to learn and then check your own answers to this problem.

8.2. ★ Use the method of least squares to find the line $y = A + Bx$ that best fits the four points $(-3, 3)$, $(-1, 4)$, (1, 8), and (3, 9). Using squared paper, plot the four points and your line. Your calculator probably has a built-in function to calculate A and B; if you don't know how to use it, take a moment to learn and then check your own answers to this problem.

8.3. ★ The best estimates for the constants A and B are determined by Equations (8.8) and (8.9). The solutions to these equations were given in Equations (8.10) through (8.12). Verify that these are indeed the solutions of (8.8) and (8.9).

8.4. ★ Prove the following useful fact: The least-squares fit for a line through any set of points $(x_1, y_1), \ldots, (x_N, y_N)$ always passes through the "center of gravity" $(\bar{x}, \bar{y})$ of the points, where the bar denotes the average of the N values concerned. [Hint: You know that A and B satisfy Equation (8.8).]

8.5. ★★ **Line Through the Origin.** Suppose two variables x and y are known to satisfy a relation $y = Bx$. That is, $y \propto x$, and a graph of y vs x is a line *through the origin.* (For example, Ohm's law, $V = RI$, tells us that a graph of voltage V vs current I should be a straight line through the origin.) Suppose further that you have N measurements (x_i, y_i) and that the uncertainties in x are negligible and those in y are all equal. Using arguments similar to those of Section 8.2, prove that the best estimate for B is

$$B = \frac{\sum xy}{\sum x^2}.$$

8.6. ★★ For measurements of just two points (x_1, y_1) and (x_2, y_2), the line that best fits the points is obviously the line *through* the points. Prove that the least-squares line for two points does indeed pass through both points. [One way to do this is to use Equations (8.8) and (8.9) to show that either of the points does satisfy the equation $y = A + Bx$.]

8.7. ★★ To find the spring constant of a spring, a student loads it with various masses M and measures the corresponding lengths l. Her results are shown in Table 8.4. Because the force $mg = k(l - l_o)$, where l_o is the unstretched length of the

Table 8.4. Length versus load for a spring; for Problem 8.7.

"x": Load m (grams)	200	300	400	500	600	700	800	900
"y": Length l (cm)	5.1	5.5	5.9	6.8	7.4	7.5	8.6	9.4

spring, these data should fit a straight line, $l = l_o + (g/k)m$. Make a least-squares fit to this line for the given data, and find the best estimates for the unstretched length l_o and the spring constant k. Do the calculations yourself, and then check your answers using the built-in functions on your calculator.

8.8. ★★ A student measures the velocity of a glider on a horizontal air track. She uses a multiflash photograph to find the glider's position s at five equally spaced times as in Table 8.5. **(a)** One way to find v would be to calculate $v = \Delta s/\Delta t$ for each of the four successive two-second intervals and then average them. Show that this procedure gives $v = (s_5 - s_1)/(t_5 - t_1)$, which means that the middle three measurements are completely ignored by this method. Prove this without putting in numbers; then find a numerical answer. **(b)** A better procedure is to make a least-

Table 8.5. Position versus time data; for Problem 8.8.

"x": Time, t (s)	−4	−2	0	2	4
"y": Position, s (cm)	13	25	34	42	56

squares fit to the equation $s = s_0 - vt$ using all five data points. Follow this procedure to find the best estimate for v and compare your result with that from part (a). Do the calculations yourself and check your answers using the built-in functions on your calculator. (The five times have negligible uncertainty; the positions are all equally uncertain.)

8.9. ★★ Weighted Least Squares. Suppose we measure N pairs of values (x_i, y_i) of two variables x and y that are supposed to satisfy a linear relation $y = A + Bx$. Suppose the x_i have negligible uncertainty and the y_i have *different* uncertainties σ_i. (That is, y_1 has uncertainty σ_1, while y_2 has uncertainty σ_2, and so on.) As in Chapter 7, we can define the *weight* of the ith measurement as $w_i = 1/\sigma_i^2$. Review the derivation of the least-squares fit in Section 8.2 and generalize it to cover this situation, where the measurements of the y_i have different weights. Show that the best estimates of A and B are

$$A = \frac{\Sigma wx^2 \Sigma wy - \Sigma wx \Sigma wxy}{\Delta} \tag{8.37}$$

and

$$B = \frac{\Sigma w \Sigma wxy - \Sigma wx \Sigma wy}{\Delta} \tag{8.38}$$

where

$$\Delta = \Sigma w \Sigma wx^2 - (\Sigma wx)^2. \tag{8.39}$$

Obviously, this method of *weighted least squares* can be applied only when the uncertainties σ_i (or at least their relative sizes) are known.

8.10. ★★ Suppose y is known to be linear in x, so that $y = A + Bx$, and we have three measurements of (x, y):

$$(1, 2 \pm 0.5), (2, 3 \pm 0.5), \text{ and } (3, 2 \pm 1).$$

(The uncertainties in x are negligible.) Use the method of weighted least squares, Equations (8.37) to (8.39), to calculate the best estimates for A and B. Compare your results with what you would get if you ignored the variation in the uncertainties, that is, used the unweighted fit of Equations (8.10) to (8.12). Plot the data and both lines, and try to understand the differences.

8.11. ★★ **(a)** If you have access to a spreadsheet program such as Lotus 123 or Excel, create a spreadsheet that will calculate the coefficients A and B for the least-squares fit for up to 10 points $(x_1, y_1), \ldots$. Use the layout of Table 8.1. **(b)** Test your spreadsheet with the data of Problems 8.1 and 8.7.

For Section 8.3: Uncertainty in the Measurements of y

8.12. ★★ Use the principle of maximum likelihood, as outlined in the discussion of Equation (8.14), to show that (8.14) gives the best estimate for the uncertainty σ_y in y in a series of measurements $(x_1, y_1), \ldots, (x_N, y_N)$ that are supposed to lie on a straight line. [Note that in (8.14), A and B are the true values of these two constants. When we replace these true values by our best estimates, as given by (8.10) and (8.11), the expression is decreased, and the N in the denominator must be replaced by $N - 2$ to get the best estimate, as in (8.15).]

8.13. ★★★ If you have a reliable estimate of the uncertainty δy in the measurements of y, then by comparing this estimate with σ_y as given by (8.15), you can assess whether your data confirm the expected linear relation $y = A + Bx$. The quantity σ_y is roughly the average distance by which the points (x_i, y_i) fail to lie on the best-fit line. If σ_y is about the same as the expected uncertainty δy, the data are consistent with the expected linear relation; if σ_y is much larger than δy, there is good reason to doubt the linear relation. The following problem illustrates these ideas.

A student measures the velocity of a glider coasting along a horizontal air track using a multiflash photograph to find its position s at four equally spaced times, as shown in Table 8.6. Assuming the glider moves with constant velocity, he fits these

Table 8.6. Positions and times of a coasting glider; for Problem 8.13.

"x": Time, t (s)	−3	−1	1	3
"y": Position, s (cm)	4.0	7.5	10.3	12.0

data to a line $s = s_0 + vt$. **(a)** Use the method of least squares to find his best estimates for a s_0 and v and for the standard deviation σ_s in the measurements of s. **(b)** Suppose the quality of his photograph was poor and he believed his measurements of s were uncertain by $\delta s \approx 1$ cm. By comparing this value of δs with σ_s, decide if his data are consistent with his assumption that the velocity was constant. Draw a graph of s vs t, with error bars showing his uncertainties $\delta s \approx 1$ cm, to confirm your conclusion. **(c)** Suppose, instead, that he was confident his measurements were good within 0.1 cm. In this case, are his measurements consistent with a constant velocity? After examining your graph of the data and best line, can you suggest an explanation?

For Section 8.4: Uncertainty in the Constants A and B

8.14. ★ Use the method of least squares to find the student's best estimate for the velocity v of the glider in Problem 8.8 and the uncertainty in v. (By all means, use the built-in functions on your calculator to find the best-fit line; unfortunately, most

calculators do not have built-in functions to find the uncertainty in A and B, so you will probably have to do this part of the calculation yourself.)

8.15. ★★ Kundt's tube is a device for measuring the wavelength λ of sound. The experimenter sets up a standing wave inside a glass tube in which he or she has sprinkled a light powder. The vibration of the air causes the powder to move and eventually to collect in small piles at the displacement nodes of the standing wave, as shown in Figure 8.5. Because the distance between the nodes is $\lambda/2$, this lets the

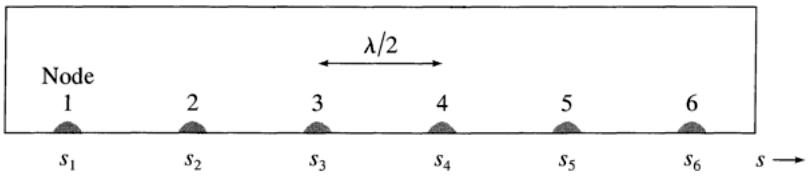


Figure 8.5. Kundt's tube, with small piles of powder at the nodes of a standing wave; for Problem 8.15.

experimenter find λ. A student finds six nodes (numbered $n = 1, \ldots, 6$) as shown in Table 8.7. Because the nodes should be equal distances $\lambda/2$ apart, their positions should satisfy $s_n = A + Bn$, where $B = \lambda/2$. Use the method of least squares to fit the data to this line and find the wavelength and its uncertainty.

Table 8.7. Positions of nodes in Kundt's tube; for Problem 8.15.

"x": Node number n	1	2	3	4	5	6
"y": Position s_n (cm)	5.0	14.4	23.1	32.3	41.0	50.4

8.16. ★★ Use error propagation to verify that the uncertainties in the parameters of a straight line $y = A + Bx$ are given by (8.16) and (8.17). [Hint: Use the general error propagation formula (3.47) and remember that, by assumption, only the y_i are subject to uncertainty. When you differentiate a sum like Σy with respect to y_i, writing the sum out as $y_1 + \ldots + y_N$ may help. Any quantity, such as the denominator Δ, that involves only the x_i has no uncertainty.]

8.17. ★★ The least-squares fit to a set of points $(x_1, y_1), \ldots, (x_N, y_N)$ treats the variables x and y unsymmetrically. Specifically, the best fit for a line $y = A + Bx$ is found on the assumption that the numbers $y_1, \ldots, y_N$ are all equally uncertain, whereas $x_1, \ldots, x_N$ have negligible uncertainty. If the situation were reversed, the roles of x and y would have to be exchanged and x and y fitted to a line $x = A' + B'y$. The resulting two lines, $y = A + Bx$ and $x = A' + B'y$, would be the same if the N points lay *exactly* on a line, but in general the two lines will be slightly different. The following problem illustrates this small difference.

(a) Find the best fit to the line $y = A + Bx$ for the data of Problem 8.1 and the uncertainties in A and B. **(b)** Now reverse the roles of x and y and fit the same data to the line $x = A' + B'y$. If you solve this equation for y in terms of x, you can find values for A and B based on the values of A' and B'. Comment on the difference between the two approaches.

8.18. ★★ Uncertainties for a Line Through the Origin. Consider the situation described in Problem 8.5, in which it is known that $y = Bx$ [that is, the points (x, y) lie on a straight line known to pass through the origin]. The best value for the slope B is given in Problem 8.5. Arguing as in Section 8.3, we can show that the uncertainty σ_y in the measurements of y is given by

$$\sigma_y = \sqrt{\frac{1}{N-1}\Sigma(y_i - Bx_i)^2}.$$

Following the argument sketched in Problem 8.16, prove that the uncertainty in the constant B is given by

$$\sigma_B = \frac{\sigma_y}{\sqrt{\Sigma x^2}}.$$

8.19. ★★★ Uncertainties in Weighted Least-Squares Fits. Consider the method of weighted least squares outlined in Problem 8.9. Use error propagation to prove that the uncertainties in the constants A and B are given by

$$\sigma_A = \sqrt{\frac{\Sigma w x^2}{\Delta}}$$

and

$$\sigma_B = \sqrt{\frac{\Sigma w}{\Delta}}.$$

For Section 8.5: An Example

8.20. ★ A student measures the pressure P of a gas at five different temperatures T, keeping the volume constant. His results are shown in Table 8.8. His data should

Table 8.8. Temperature (in °C) versus Pressure (in mm of mercury); for Problem 8.20.

"x": Pressure P	79	82	85	88	90
"y": Temperature T	8	17	30	37	52

fit a straight line of the form $T = A + BP$, where A is the absolute zero of temperature (whose accepted value is −273°C). Find the best fit to the student's data and hence his best estimate for absolute zero and its uncertainty.

For Section 8.6: Least-Squares Fits to Other Curves

8.21. ★★ Consider the problem of fitting a set of measurements (x_i, y_i), with $i = 1, \ldots, N$, to the polynomial $y = A + Bx + Cx^2$. Use the principle of maximum likelihood to show that the best estimates for A, B, and C based on the data are given by Equations (8.27). Follow the arguments outlined between Equations (8.24) and (8.27).

8.22. ★★ One way to measure the acceleration of a freely falling body is to measure its heights y_i at a succession of equally spaced times t_i (with a multiflash photograph, for example) and to find the best fit to the expected polynomial

$$y = y_0 + v_0 t - \tfrac{1}{2} g t^2. \tag{8.40}$$

Use the equations (8.27) to find the best estimates for the three coefficients in (8.40) and hence the best estimate for g, based on the five measurements in Table. 8.9.

Table 8.9. Height (in cm) versus time (in tenths of a second) for a falling body; for Problem 8.22.

"x": Time t	−2	−1	0	1	2
"y": Height y	131	113	89	51	7

(Note that we can name the times however we like. A more natural choice might seem to be $t = 0, 1, \ldots, 4$. When you solve the problem, however, you will see that defining the times to be symmetrically spaced about $t = 0$ causes approximately half of the sums involved to be zero and greatly simplifies the algebra. This trick can be used whenever the values of the independent variable are equally spaced.

8.23. ★★ Suppose y is expected to have the form $y = Af(x) + Bg(x)$, where A and B are unknown coefficients and f and g are fixed, known functions (such as $f = x$ and $g = x^2$, or $f = \cos x$ and $g = \sin x$). Use the principle of maximum likelihood to show that the best estimates for A and B, based on data (x_i, y_i), $i = 1, \ldots, N$, must satisfy

$$\begin{aligned} A\Sigma[f(x_i)]^2 + B\Sigma f(x_i)g(x_i) &= \Sigma y_i f(x_i), \\ A\Sigma f(x_i)g(x_i) + B\Sigma[g(x_i)]^2 &= \Sigma y_i g(x_i). \end{aligned} \tag{8.41}$$

8.24. ★★ A weight oscillating on a vertical spring should have height given by

$$y = A \cos \omega t + B \sin \omega t.$$

A student measures ω to be 10 rad/s with negligible uncertainty. Using a multiflash photograph, she then finds y for five equally spaced times, as shown in Table 8.10.

Table 8.10. Positions (in cm) and times (in tenths of a second) for an oscillating mass; for Problem 8.24.

"x": Time t	−4	−2	0	2	4
"y": Position y	3	−16	6	9	−8

Use Equations (8.41) to find best estimates for A and B. Plot the data and your best fit. (If you plot the data first, you will have the opportunity to consider how hard it would be to choose a best fit without the least-squares method.) If the student judges that her measured values of y were uncertain by "a couple of centimeters," would you say the data are an acceptable fit to the expected curve?

8.25. ★★ The rate at which a sample of radioactive material emits radiation decreases exponentially as the material is depleted. To study this exponential decay, a student places a counter near a sample and records the number of decays in 15 seconds. He repeats this five times at 10-minute intervals and obtains the results shown in Table 8.11. (Notice that, because it takes nearly 10 minutes to prepare the equipment, his first measurement is made at $t = 10$ min.)

Table 8.11. Number $\nu(t)$ of emissions in a 15-second interval versus total time elapsed t (in minutes); for Problem 8.25.

"x": Elapsed time t	10	20	30	40	50
"y": Number $\nu(t)$	409	304	260	192	170

If the sample does decay exponentially, the number $\nu(t)$ should satisfy

$$\nu(t) = \nu_0 e^{-t/\tau}, \tag{8.42}$$

where τ is the (unknown) mean life of the material in question and ν_0 is another unknown constant. To find the mean life τ, the student takes the natural log of Equation (8.42) to give the linear relation

$$z = \ln(\nu) = \ln(\nu_0) - t/\tau \tag{8.43}$$

and makes a least-squares fit to this line. What is his answer for the mean life τ? How many decays would he have counted in 15 seconds at $t = 0$?

8.26. ★★★ The student of Problem 8.25 decides to investigate the exponential character of radioactive decay further and repeats his experiment, monitoring the decays for a longer total time. He decides to count the decays in 15-second periods as before but makes six measurements at 30-minute intervals, as shown in Table 8.12.

Table 8.12. Number $\nu(t)$ of emissions in a 15-second interval versus total time elapsed t (in minutes); for Problem 8.26.

"x": Time t	10	40	70	100	130	160
"y": Number $\nu(t)$	188	102	60	18	16	5

According to the square-root rule of Section 3.2, the uncertainty in each of the counts ν is $\sqrt{\nu}$, which obviously varies widely during this experiment. Therefore, he decides to make a *weighted* least-squares fit to the straight line (8.43). **(a)** To this end, he must first calculate the logs, $z_i = \ln(\nu_i)$, and their uncertainties. Use error propagation to show that the uncertainty in z_i is $1/\sqrt{\nu_i}$, which means that the weight

of z_i is just ν_i. **(b)** Now make the weighted least-squares fit to the line (8.43). What is his answer for the material's mean life τ and its uncertainty? [The best values for the coefficients A and B are given in Equations (8.37) and (8.38) of Problem 8.9, and their uncertainties are given in Problem 8.19. Make a table showing t, ν, $z = \ln(\nu)$, and the various sums in (8.37) to (8.39).] **(c)** Draw a graph of $z = \ln(\nu)$ vs t, showing the best-fit straight line and the data with error bars. Is your graph consistent with the expectation that the decay rate should decrease exponentially?

Chapter 9

Covariance and Correlation

This chapter introduces the important concept of covariance. Because this concept arises naturally in the propagation of errors, Section 9.1 starts with a quick review of error propagation. This review sets the stage for Section 9.2, which defines covariance and discusses its role in the propagation of errors. Then, Section 9.3 uses the covariance to define the coefficient of linear correlation for a set of measured points $(x_1, y_1), \ldots, (x_N, y_N)$. This coefficient, denoted r, provides a measure of how well the points fit a straight line of the form $y = A + Bx$; its use is described in Sections 9.4 and 9.5.

9.1 Review of Error Propagation

This and the next section provide a final look at the important question of error propagation. We first discussed error propagation in Chapter 3, where we reached several conclusions. We imagined measuring two quantities x and y to calculate some function $q(x, y)$, such as $q = x + y$ or $q = x^2 \sin y$. [In fact, we discussed a function $q(\mathrm{x}, \ldots, z)$ of an arbitrary number of variables $x, \ldots, z$; for simplicity, we will now consider just two variables.] A simple argument suggested that the uncertainty in our answer for q is just

$$\delta q \approx \left|\frac{\partial q}{\partial x}\right| \delta x + \left|\frac{\partial q}{\partial y}\right| \delta y. \tag{9.1}$$

We first derived this approximation for the simple special cases of sums, differences, products, and quotients. For instance, if q is the sum $q = x + y$, then (9.1) reduces to the familiar $\delta q \approx \delta x + \delta y$. The general result (9.1) was derived in Equation (3.43).

We next recognized that (9.1) is often probably an overstatement of δq, because there may be partial cancellation of the errors in x and y. We stated, without proof, that when the errors in x and y are independent and random, a better value for the

uncertainty in the calculated value of $q(x, y)$ is the quadratic sum

$$\delta q = \sqrt{\left(\frac{\partial q}{\partial x}\,\delta x\right)^2 + \left(\frac{\partial q}{\partial y}\,\delta y\right)^2}. \tag{9.2}$$

We also stated, without proof, that whether or not the errors are independent and random, the simpler formula (9.1) always gives an upper bound on δq; that is, the uncertainty δq is never any worse than is given by (9.1).

Chapter 5 gave a proper definition and proof of (9.2). First, we saw that a good measure of the uncertainty δx in a measurement is given by the standard deviation σ_x; in particular, we saw that if the measurements of x are normally distributed, we can be 68% confident that the measured value lies within σ_x of the true value. Second, we saw that if the measurements of x and y are governed by independent normal distributions, with standard deviations σ_x and σ_y, the values of $q(x, y)$ are also normally distributed, with standard deviation

$$\sigma_q = \sqrt{\left(\frac{\partial q}{\partial x}\,\sigma_x\right)^2 + \left(\frac{\partial q}{\partial y}\,\sigma_y\right)^2}. \tag{9.3}$$

This result provides the justification for the claim (9.2).

In Section 9.2, I will derive a precise formula for the uncertainty in q that applies whether or not the errors in x and y are independent and normally distributed. In particular, I will prove that (9.1) always provides an upper bound on the uncertainty in q.

Before I derive these results, let us first review the definition of the standard deviation. The standard deviation σ_x of N measurements $x_1, \ldots, x_N$ was originally defined by the equation

$$\sigma_x^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \bar{x})^2. \tag{9.4}$$

If the measurements of x are normally distributed, then in the limit that N is large, the definition (9.4) is equivalent to defining σ_x as the width parameter that appears in the Gauss function

$$\frac{1}{\sigma_x\sqrt{2\pi}}\, e^{-(x-X)^2/2\sigma_x^2}$$

that governs the measurements of x. Because we will now consider the possibility that the errors in x may not be normally distributed, this second definition is no longer available to us. We can, and will, still define σ_x by (9.4), however. Whether or not the distribution of errors is normal, this definition of σ_x gives a reasonable measure of the random uncertainties in our measurement of x. (As in Chapter 5, I will suppose all systematic errors have been identified and reduced to a negligible level, so that all remaining errors *are* random.)

The usual ambiguity remains as to whether to use the definition (9.4) of σ_x or the "improved" definition with the factor N in the denominator replaced by $(N - 1)$. Fortunately, the discussion that follows applies to either definition, as long as we are consistent in our use of one or the other. For convenience, I will use the definition (9.4), with N in the denominator throughout this chapter.

9.2 Covariance in Error Propagation

Suppose that to find a value for the function $q(x, y)$, we measure the two quantities x and y several times, obtaining N pairs of data, $(x_1, y_1), \ldots, (x_N, y_N)$. From the N measurements $x_1, \ldots, x_N$, we can compute the mean $\bar{x}$ and standard deviation σ_x in the usual way; similarly, from $y_1, \ldots, y_N$, we can compute $\bar{y}$ and σ_y. Next, using the N pairs of measurements, we can compute N values of the quantity of interest

$$q_i = q(x_i, y_i), \qquad (i = 1, \ldots, N).$$

Given $q_1, \ldots, q_N$, we can now calculate their mean $\bar{q}$, which we assume gives our best estimate for q, and their standard deviation σ_q, which is our measure of the random uncertainty in the values q_i.

I will assume, as usual, that all our uncertainties are small and hence that all the numbers $x_1, \ldots, x_N$ are close to $\bar{x}$ and that all the $y_1, \ldots, y_N$ are close to $\bar{y}$. We can then make the approximation

$$\begin{aligned} q_i &= q(x_i, y_i) \\ &\approx q(\bar{x}, \bar{y}) + \frac{\partial q}{\partial x}(x_i - \bar{x}) + \frac{\partial q}{\partial y}(y_i - \bar{y}). \end{aligned} \tag{9.5}$$

In this expression, the partial derivatives $\partial q/\partial x$ and $\partial q/\partial y$ are taken at the point $x = \bar{x}$, $y = \bar{y}$, and are therefore the same for all $i = 1, \ldots, N$. With this approximation, the mean becomes

$$\begin{aligned} \bar{q} &= \frac{1}{N}\sum_{i=1}^{N} q_i \\ &= \frac{1}{N}\sum_{i=1}^{N}\left[q(\bar{x}, \bar{y}) + \frac{\partial q}{\partial x}(x_i - \bar{x}) + \frac{\partial q}{\partial y}(y_i - \bar{y})\right]. \end{aligned}$$

This equation gives $\bar{q}$ as the sum of three terms. The first term is just $q(\bar{x}, \bar{y})$, and the other two are exactly zero. [For example, it follows from the definition of $\bar{x}$ that $\Sigma(x_i - \bar{x}) = 0$.] Thus, we have the remarkably simple result

$$\bar{q} = q(\bar{x}, \bar{y}); \tag{9.6}$$

that is, to find the mean $\bar{q}$ we have only to calculate the function $q(x, y)$ at the point $x = \bar{x}$ and $y = \bar{y}$.

The standard deviation in the N values $q_1, \ldots, q_N$ is given by

$$\sigma_q{}^2 = \frac{1}{N}\Sigma(q_i - \bar{q})^2.$$

Substituting (9.5) and (9.6), we find that

$$\begin{aligned} \sigma_q{}^2 &= \frac{1}{N}\Sigma\left[\frac{\partial q}{\partial x}(x_i - \bar{x}) + \frac{\partial q}{\partial y}(y_i - \bar{y})\right]^2 \\ &= \left(\frac{\partial q}{\partial x}\right)^2 \frac{1}{N}\Sigma(x_i - \bar{x})^2 + \left(\frac{\partial q}{\partial y}\right)^2 \frac{1}{N}\Sigma(y_i - \bar{y})^2 \\ &\quad + 2\frac{\partial q}{\partial x}\frac{\partial q}{\partial y}\frac{1}{N}\Sigma(x_i - \bar{x})(y_i - \bar{y}). \end{aligned} \tag{9.7}$$

The sums in the first two terms are those that appear in the definition of the standard deviations σ_x and σ_y. The final sum is one we have not encountered before. It is called the *covariance*[1] of x and y and is denoted

$$\sigma_{xy} = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}). \tag{9.8}$$

With this definition, Equation (9.7) for the standard deviation σ_q becomes

$$\sigma_q^2 = \left(\frac{\partial q}{\partial x}\right)^2\sigma_x^2 + \left(\frac{\partial q}{\partial y}\right)^2\sigma_y^2 + 2\frac{\partial q}{\partial x}\frac{\partial q}{\partial y}\sigma_{xy}. \tag{9.9}$$

This equation gives the standard deviation σ_q, whether or not the measurements of x and y are independent or normally distributed.

If the measurements of x and y *are* independent, we can easily see that, after many measurements, the covariance σ_{xy} should approach zero: Whatever the value of y_i, the quantity $x_i - \bar{x}$ is just as likely to be negative as it is to be positive. Thus, after many measurements, the positive and negative terms in (9.8) should nearly balance; in the limit of infinitely many measurements, the factor $1/N$ in (9.8) guarantees that σ_{xy} is zero. (After a finite number of measurements, σ_{xy} will not be exactly zero, but it should be *small* if the errors in x and y really are independent and random.) With σ_{xy} zero, Equation (9.9) for σ_q reduces to

$$\sigma_q^2 = \left(\frac{\partial q}{\partial x}\right)^2\sigma_x^2 + \left(\frac{\partial q}{\partial y}\right)^2\sigma_y^2, \tag{9.10}$$

the familiar result for independent and random uncertainties.

If the measurements of x and y are *not* independent, the covariance σ_{xy} need not be zero. For instance, it is easy to imagine a situation in which an overestimate of x will always be accompanied by an overestimate of y, and vice versa. The numbers $(x_i - \bar{x})$ and $(y_i - \bar{y})$ will then always have the same sign (both positive or both negative), and their product will always be positive. Because all terms in the sum (9.8) are positive, σ_{xy} will be positive (and nonzero), even in the limit that we make infinitely many measurements. Conversely, you can imagine situations in which an overestimate of x is always accompanied by an underestimate of y, and vice versa; in this case $(x_i - \bar{x})$ and $(y_i - \bar{y})$ will always have opposite signs, and σ_{xy} will be negative. This case is illustrated in the example below.

When the covariance σ_{xy} is not zero (even in the limit of infinitely many measurements), we say that the errors in x and y are *correlated.* In this case, the uncertainty σ_q in $q(x, y)$ as given by (9.9) is *not* the same as we would get from the formula (9.10) for independent, random errors.

---

[1] The name *covariance* for σ_{xy} (for two variables x, y) parallels the name *variance* for σ_x^2 (for one variable x). To emphasize this parallel, the covariance (9.8) is sometimes denoted σ_{xy}^2, not an especially apt notation, because the covariance can be negative. A convenient feature of the definition (9.8) is that σ_{xy} has the dimensions of xy, just as σ_x has the dimensions of x.

Example: Two Angles with a Negative Covariance

Each of five students measures the same two angles α and β and obtains the results shown in the first three columns of Table 9.1.

Table 9.1. Five measurements of two angles α and β (in degrees).

Student	α	β	$(\alpha - \bar{\alpha})$	$(\beta - \bar{\beta})$	$(\alpha - \bar{\alpha})(\beta - \bar{\beta})$
A	35	50	2	−2	−4
B	31	55	−2	3	−6
C	33	51	0	−1	0
D	32	53	−1	1	−1
E	34	51	1	−1	−1

Find the average and standard deviation for each of the two angles, and then find the covariance $\sigma_{\alpha\beta}$ as defined by (9.8). The students now calculate the sum $q = \alpha + \beta$. Find their best estimate for q as given by (9.6) and the standard deviation σ_q as given by (9.9). Compare the standard deviation with what you would get if you assumed (incorrectly) that the errors in α and β were independent and that σ_q was given by (9.10).

The averages are immediately seen to be $\bar{\alpha} = 33$ and $\bar{\beta} = 52$. With these values, we can find the deviations $(\alpha - \bar{\alpha})$ and $(\beta - \bar{\beta})$, as shown in Table 9.1, and from these deviations we easily find

$$\sigma_\alpha{}^2 = 2.0 \quad \text{and} \quad \sigma_\beta{}^2 = 3.2.$$

[Here I have used the definition (9.4), with the N in the denominator.]

You can see from Table 9.1 that high values of α seem to be correlated with low values of β and vice versa, because $(\alpha - \bar{\alpha})$ and $(\beta - \bar{\beta})$ always have opposite signs. (For an experiment in which this kind of correlation arises, see Problem 9.6.) This correlation means that the products $(\alpha - \bar{\alpha})(\beta - \bar{\beta})$ shown in the last column of the table are all negative (or zero). Thus, the covariance $\sigma_{\alpha\beta}$ as defined by (9.8) is negative,

$$\sigma_{\alpha\beta} = \frac{1}{N}\Sigma(\alpha - \bar{\alpha})(\beta - \bar{\beta}) = \frac{1}{5} \times (-12) = -2.4.$$

The best estimate for the sum $q = \alpha + \beta$ is given by (9.6) as

$$q_{\text{best}} = \bar{q} = \bar{\alpha} + \bar{\beta} = 33 + 52 = 85.$$

To find the standard deviation using (9.9), we need the two partial derivatives, which are easily seen to be $\partial q/\partial\alpha = \partial q/\partial\beta = 1$. Therefore, according to (9.9),

$$\begin{aligned}\sigma_q &= \sqrt{\sigma_\alpha{}^2 + \sigma_\beta{}^2 + 2\sigma_{\alpha\beta}} \\ &= \sqrt{2.0 + 3.2 - 2 \times 2.4} = 0.6.\end{aligned}$$

If we overlooked the correlation between the measurements of α and β and treated them as independent, then according to (9.10) we would get the incorrect answer

$$\begin{aligned}\sigma_q &= \sqrt{{\sigma_\alpha}^2 + {\sigma_\beta}^2} \\ &= \sqrt{2.0 + 3.2} = 2.3.\end{aligned}$$

We see from this example that a correlation of the right sign can cause a dramatic difference in a propagated error. In this case we can see why there is this difference: The errors in each of the angles α and β are a degree or so, suggesting that $q = \alpha + \beta$ would be uncertain by a couple of degrees. But, as we have noted, the positive errors in α are accompanied by negative errors in β, and vice versa. Thus, when we add α and β, the errors tend to cancel, leaving an uncertainty of only a fraction of a degree.

Quick Check 9.1. Each of three students measures the two sides, x and y, of a rectangle and obtains the results shown in Table 9.2. Find the means $\bar{x}$ and $\bar{y}$,

Table 9.2. Three measurements of x and y (in mm); for Quick Check 9.1.

Student	x	y
A	25	33
B	27	34
C	29	38

and then make a table like Table 9.1 to find the covariance σ_{xy}. If the students calculate the sum $q = x + y$, find the standard deviation σ_q using the correct formula (9.9), and compare it with the value you would get if you ignored the covariance and used (9.10). (Notice that in this example, high values of x seem to correlate with high values of y and vice versa. Specifically, student C appears consistently to overestimate and student A to underestimate. Remember also that with just three measurements, the results of any statistical calculation are only a rough guide to the uncertainties concerned.)

Using the formula (9.9), we can derive an upper limit on σ_q that is always valid. It is a simple algebraic exercise (Problem 9.7) to prove that the covariance σ_{xy} satisfies the so-called *Schwarz inequality*

$$|\sigma_{xy}| \leq \sigma_x \sigma_y. \tag{9.11}$$

If we substitute (9.11) into the expression (9.9) for the uncertainty σ_q, we find that

$${\sigma_q}^2 \leq \left(\frac{\partial q}{\partial x}\right)^2 {\sigma_x}^2 + \left(\frac{\partial q}{\partial y}\right)^2 {\sigma_y}^2 + 2\left|\frac{\partial q}{\partial x}\frac{\partial q}{\partial y}\right| \sigma_x \sigma_y$$

$$= \left[\left| \frac{\partial q}{\partial x} \right| \sigma_x + \left| \frac{\partial q}{\partial y} \right| \sigma_y \right]^2;$$

that is,

$$\sigma_q \leqslant \left| \frac{\partial q}{\partial x} \right| \sigma_x + \left| \frac{\partial q}{\partial y} \right| \sigma_y. \tag{9.12}$$

With this result, we have finally established the precise significance of our original, simple expression

$$\delta q \approx \left| \frac{\partial q}{\partial x} \right| \delta x + \left| \frac{\partial q}{\partial y} \right| \delta y \tag{9.13}$$

for the uncertainty δq. If we adopt the standard deviation σ_q as our measure of the uncertainty in q, then (9.12) shows that the old expression (9.13) is really the *upper limit* on the uncertainty. Whether or not the errors in x and y are independent and normally distributed, the uncertainty in q will never exceed the right side of (9.13). If the measurements of x and y are correlated in just such a way that $|\sigma_{xy}| = \sigma_x \sigma_y$, its largest possible value according to (9.11), then the uncertainty in q can actually be as large as given by (9.13), but it can never be any larger.

In an introductory physics laboratory, students usually do not make measurements for which the covariance σ_{xy} can be estimated reliably. Thus, you will probably not have occasion to use the result (9.9) explicitly. If, however, you suspect that two variables x and y may be correlated, you should probably consider using the bound (9.12) instead of the quadratic sum (9.10). Our next topic is an application of covariance that you will almost certainly be able to use.

9.3 Coefficient of Linear Correlation

The notion of covariance σ_{xy} introduced in Section 9.2 enables us to answer the question raised in Chapter 8 of how well a set of measurements $(x_1, y_1), \ldots, (x_N, y_N)$ of two variables supports the hypothesis that x and y are linearly related.

Let us suppose we have measured N pairs of values $(x_1, y_1), \ldots, (x_N, y_N)$ of two variables that we suspect should satisfy a linear relation of the form

$$y = A + Bx.$$

Note that $x_1, \ldots, x_N$ are no longer measurements of one single number, as they were in the past two sections; rather, they are measurements of N different values of some variable (for example, N different heights from which we have dropped a stone). The same applies to $y_1, \ldots, y_N$.

Using the method of least squares, we can find the values of A and B for the line that best fits the points $(x_1, y_1), \ldots, (x_N, y_N)$. If we already have a reliable estimate of the uncertainties in the measurements, we can see whether the measured points do lie reasonably close to the line (compared with the known uncertainties). If they do, the measurements support our suspicion that x and y are linearly related.

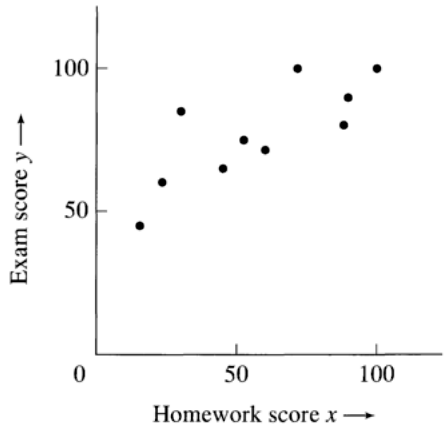


Figure 9.1. A "scatter plot" showing students' scores on exams and homework. Each of the 10 points (x_i, y_i) shows a student's homework score, x_i, and exam score, y_i.

Unfortunately, in many experiments, getting a reliable estimate of the uncertainties in advance is hard, and we must use the data themselves to decide whether the two variables appear to be linearly related. In particular, there is a type of experiment for which knowing the size of uncertainties in advance is *impossible.* This type of experiment, which is more common in the social than the physical sciences, is best explained by an example.

Suppose a professor, anxious to convince his students that doing homework will help them do well in exams, keeps records of their scores on homework and exams and plots the scores on a "scatter plot" as in Figure 9.1. In this figure, homework scores are plotted horizontally and exam scores vertically. Each point (x_i, y_i) shows one student's homework score, x_i, and exam score, y_i. The professor hopes to show that high exam scores tend to be *correlated* with high homework scores, and vice versa (and his scatter plot certainly suggests this is approximately so). This kind of experiment has no uncertainties in the points; each student's two scores are known exactly. The uncertainty lies rather in the extent to which the scores *are correlated;* and this has to be decided from the data.

The two variables x and y (in either a typical physics experiment or one like that just described) may, of course, be related by a more complicated relation than the simple linear one, $y = A + Bx$. For example, plenty of physical laws lead to quadratic relations of the form $y = A + Bx + Cx^2$. Nevertheless, I restrict my discussion here to the simpler problem of deciding whether a given set of points supports the hypothesis of a *linear* relation $y = A + Bx$.

The extent to which a set of points $(x_1, y_1), \ldots, (x_N, y_N)$ supports a linear relation between x and y is measured by the *linear correlation coefficient,* or just *correlation coefficient,*

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y}, \tag{9.14}$$

where the covariance σ_{xy} and standard deviations σ_x and σ_y are defined exactly as before, in Equations (9.8) and (9.4).[2] Substituting these definitions into (9.14), we can rewrite the correlation coefficient as

$$r = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma (x_i - \bar{x})^2 \Sigma (y_i - \bar{y})^2}}. \tag{9.15}$$

As I will show directly, the number r is an indicator of how well the points (x_i, y_i) fit a straight line. It is a number between -1 and 1. If r is close to ± 1, the points lie close to some straight line; if r is close to 0, the points are uncorrelated and have little or no tendency to lie on a straight line.

To prove these assertions, we first observe that the Schwarz inequality (9.11), $|\sigma_{xy}| \leq \sigma_x \sigma_y$, implies immediately that $|r| \leq 1$ or

$$-1 \leq r \leq 1$$

as claimed. Next, let us suppose that the points (x_i, y_i) all lie *exactly* on the line $y = A + Bx$. In this case $y_i = A + Bx_i$ for all i, and hence $\bar{y} = A + B\bar{x}$. Subtracting these two equations, we see that

$$y_i - \bar{y} = B(x_i - \bar{x})$$

for each i. Inserting this result into (9.15), we find that

$$r = \frac{B \Sigma (x_i - \bar{x})^2}{\sqrt{\Sigma (x_i - \bar{x})^2 B^2 \Sigma (x_i - \bar{x})^2}} = \frac{B}{|B|} = \pm 1. \tag{9.16}$$

That is, if the points $(x_1, y_1), \ldots, (x_N, y_N)$ lie perfectly on a line, then $r = \pm 1$, and its sign is determined by the slope of the line ($r = 1$ for B positive, and $r = -1$ for B negative).[3] Even when the variables x and y really are linearly related, we do not expect our experimental points to lie *exactly* on a line. Thus, we do not expect r to be exactly ± 1. On the other hand, we do expect a value of r that is *close to* ± 1, if we believe that x and y are linearly related.

Suppose, on the other hand, there is no relationship between the variables x and y. Whatever the value of y_i, each x_i would then be just as likely to be above $\bar{x}$ as below $\bar{x}$. Thus, the terms in the sum

$$\Sigma (x_i - \bar{x})(y_i - \bar{y})$$

in the numerator of r in (9.15) are just as likely to be positive as negative. Meanwhile, the terms in the denominator of r are all positive. Thus, in the limit that N, the number of measurements, approaches infinity, the correlation coefficient r will

---

[2] Notice, however, that their significance is slightly different. For example, in Section 9.2 $x_1, \ldots, x_N$ were measurements of *one number*, and if these measurements were precise, σ should be small. In the present case $x_1, \ldots, x_N$ are measurements of *different* values of a variable, and even if the measurements are precise, there is no reason to think σ_x will be small. Note also that some authors use the number r^2, called the *coefficient of determination*.

[3] If the line is exactly horizontal, then $B = 0$, and (9.16) gives $r = 0/0$; that is, r is undefined. Fortunately, this special case is not important in practice, because it corresponds to y being a constant, independent of x.

be zero. With a finite number of data points, we do not expect r to be exactly zero, but we do expect it to be *small* (if the two variables really are unrelated).

If two variables x and y are such that, in the limit of infinitely many measurements, their covariance σ_{xy} is zero (and hence $r = 0$), we say that the variables are *uncorrelated.* If, after a finite number of measurements, the correlation coefficient $r = \sigma_{xy}/\sigma_x\sigma_y$ is small, the hypothesis that x and y are uncorrelated is supported.

As an example, consider the exam and homework scores shown in Figure 9.1. These scores are given in Table 9.3. A simple calculation (Problem 9.12) shows that

Table 9.3. Students' scores.

Student i	1	2	3	4	5	6	7	8	9	10
Homework x_i	90	60	45	100	15	23	52	30	71	88
Exam y_i	90	71	65	100	45	60	75	85	100	80

the correlation coefficient for these 10 pairs of scores is $r = 0.8$. The professor concludes that this value is "reasonably close" to 1 and so can announce to next year's class that, because homework and exam scores show good correlation, it is important to do the homework.

If our professor had found a correlation coefficient r close to zero, he would have been in the embarrassing position of having shown that homework scores have no bearing on exam scores. If r had turned out to be close to -1, then he would have made the even more embarrassing discovery that homework and exam scores show a *negative* correlation; that is, that students who do a good job on homework tend to do poorly on the exam.

Quick Check 9.2. Find the correlation coefficient for the data of Quick Check 9.1. Note that these measurements show a positive correlation; that is, high values of x correlate with high values of y, and vice versa.

9.4 Quantitative Significance of r

The example of the homework and exam scores clearly shows that we do not yet have a complete answer to our original question about how well data points support a linear relation between x and y. Our professor found a correlation coefficient $r = 0.8$, and judged this value "reasonably close" to 1. But how can we decide objectively what is "reasonably close" to 1? Would $r = 0.6$ have been reasonably close? Or $r = 0.4$? These questions are answered by the following argument.

Suppose the two variables x and y are in reality *uncorrelated;* that is, in the limit of infinitely many measurements, the correlation coefficient r would be zero.

After a finite number of measurements, r is very unlikely to be exactly zero. One can, in fact, calculate the probability that r will exceed any specific value. We will denote by

$$Prob_N(|r| \geq r_o)$$

the probability that N measurements of two uncorrelated variables x and y will give a coefficient r larger[4] than any particular r_o. For instance, we could calculate the probability

$$Prob_N(|r| \geq 0.8)$$

that, after N measurements of the uncorrelated variables x and y, the correlation coefficient would be at least as large as our professor's 0.8. The calculation of these probabilities is quite complicated and will not be given here. The results for a few representative values of the parameters are shown in Table 9.4, however, and a more complete tabulation is given in Appendix C.

Table 9.4. The probability $Prob_N(|r| \geq r_o)$ that N measurements of two uncorrelated variables x and y would produce a correlation coefficient with $|r| \geq r_o$. Values given are percentage probabilities, and blanks indicate values less than 0.05%.

	r_o										
N	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
3	100	94	87	81	74	67	59	51	41	29	0
6	100	85	70	56	43	31	21	12	6	1	0
10	100	78	58	40	25	14	7	2	0.5		0
20	100	67	40	20	8	2	0.5	0.1			0
50	100	49	16	3	0.4						0

Although we have not shown how the probabilities in Table 9.4 are calculated, we can understand their general behavior and put them to use. The first column shows the number of data points N. (In our example, the professor recorded 10 students' scores, so $N = 10$.) The numbers in each succeeding column show the percentage probability that N measurements of two *uncorrelated* variables would yield a coefficient r at least as big as the number at the top of the column. For example, we see that the probability that 10 uncorrelated data points would give $|r| \geq 0.8$ is only 0.5%, not a large probability. Our professor can therefore say it is *very unlikely* that uncorrelated scores would have produced a coefficient with $|r|$ greater than or equal to the 0.8 that he obtained. In other words, it is very *likely* that the scores on homework and examinations really are correlated.

Several features of Table 9.4 deserve comment. All entries in the first column are 100%, because $|r|$ is always greater than or equal to zero; thus, the probability

[4] Because a correlation is indicated if r is close to $+1$ *or* to -1, we consider the probability of getting the *absolute value* $|r| \geq r_o$.

of finding $|r| \geq 0$ is always 100%. Similarly, the entries in the last column are all zero, because the probability of finding $|r| \geq 1$ is zero.[5] The numbers in the intermediate columns vary with the number of data points N. This variation also is easily understood. If we make just three measurements, the chance of their having a correlation coefficient with $|r| \geq 0.5$, say, is obviously quite good (67%, in fact); but if we make 20 measurements and the two variables really are uncorrelated, the chance of finding $|r| \geq 0.5$ is obviously very small (2%).

Armed with the probabilities in Table 9.4 (or in the more complete table in Appendix C), we now have the most complete possible answer to the question of how well N pairs of values (x_i, y_i) support a linear relation between x and y. From the measured points, we can first calculate the observed correlation coefficient r_o (the subscript o stands for "observed"). Next, using one of these tables, we can find the probability $Prob_N(|r| \geq |r_o|)$ that N uncorrelated points would have given a coefficient at least as large as the observed coefficient r_o. If this probability is "sufficiently small," we conclude that it is very *improbable* that x and y are uncorrelated and hence very *probable* that they really are correlated.

We still have to choose the value of the probability we regard as "sufficiently small." One fairly common choice is to regard an observed correlation r_o as "significant" if the probability of obtaining a coefficient r with $|r| \geq |r_o|$ from uncorrelated variables is less than 5%. A correlation is sometimes called "highly significant" if the corresponding probability is less than 1%. Whatever choice we make, we do *not* get a definite answer that the data are, or are not, correlated; instead, we have a quantitative measure of how improbable it is that they are uncorrelated.

Quick Check 9.3. The professor of Section 9.3 teaches the same course the following year and this time has 20 students. Once again, he records homework and exam scores and this time finds a correlation coefficient $r = 0.6$. Would you describe this correlation as significant? Highly significant?

9.5 Examples

Suppose we measure three pairs of values (x_i, y_i) and find that they have a correlation coefficient of 0.7 (or -0.7). Does this value support the hypothesis that x and y are linearly related?

Referring to Table 9.4, we see that even if the variables x and y were completely uncorrelated, the probability is 51% for getting $|r| \geq 0.7$ when $N = 3$. In other words, it is entirely possible that x and y are uncorrelated, so we have no worthwhile evidence of correlation. In fact, with only three measurements, getting convincing evidence of a correlation would be very difficult. Even an observed coefficient as large as 0.9 is quite insufficient, because the probability is 29% for getting $|r| \geq 0.9$ from three measurements of uncorrelated variables.

---

[5] Although it is *impossible* that $|r| > 1$, it is, in principle, possible that $|r| = 1$. However, r is a continuous variable, and the probability of getting $|r|$ exactly equal to 1 is zero. Thus $Prob_N(|r| \geq 1) = 0$.

If we found a correlation of 0.7 from six measurements, the situation would be a little better but still not good enough. With $N = 6$, the probability of getting $|r| \geq 0.7$ from uncorrelated variables is 12%. This probability is not small enough to rule out the possibility that x and y are uncorrelated.

On the other hand, if we found $r = 0.7$ after 20 measurements, we would have strong evidence for a correlation, because when $N = 20$, the probability of getting $|r| \geq 0.7$ from two uncorrelated variables is only 0.1%. By any standards this is very improbable, and we could confidently argue that a correlation is indicated. In particular, the correlation could be called "highly significant," because the probability concerned is less than 1%.

Principal Definitions and Equations of Chapter 9

COVARIANCE

Given N pairs of measurements $(x_1, y_1), \ldots, (x_N, y_N)$ of two quantities x and y, we define their covariance to be

$$\sigma_{xy} = \frac{1}{N}\Sigma(x_i - \bar{x})(y_i - \bar{y}). \qquad \text{[See (9.8)]}$$

If we now use the measured values to calculate a function $q(x, y)$, the standard deviation of q is given by

$$\sigma_q^2 = \left(\frac{\partial q}{\partial x}\right)^2 \sigma_x^2 + \left(\frac{\partial q}{\partial y}\right)^2 \sigma_y^2 + 2\frac{\partial q}{\partial x}\frac{\partial q}{\partial y}\sigma_{xy}. \qquad \text{[See (9.9)]}$$

If the errors in x and y are independent, then $\sigma_{xy} = 0$, and this equation reduces to the usual formula for error propagation. Whether or not the errors are independent, the Schwarz inequality (9.11) implies the upper bound

$$\sigma_q \leq \left|\frac{\partial q}{\partial x}\right| \sigma_x + \left|\frac{\partial q}{\partial y}\right| \sigma_y. \qquad \text{[See (9.12)]}$$

CORRELATION COEFFICIENT

Given N measurements $(x_1, y_1), \ldots, (x_N, y_N)$ of two variables x and y, we define the correlation coefficient r as

$$r = \frac{\sigma_{xy}}{\sigma_x \sigma_y} = \frac{\Sigma (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma (x_i - \bar{x})^2 \Sigma (y_i - \bar{y})^2}}. \qquad \text{[See (9.15)]}$$

An equivalent form, which is sometimes more convenient, is

$$r = \frac{\Sigma x_i y_i - N\bar{x}\bar{y}}{\sqrt{(\Sigma x_i^2 - N\bar{x}^2)(\Sigma y_i^2 - N\bar{y}^2)}}. \qquad \text{[See Problem 9.10]}$$

Values of r near 1 or -1 indicate strong linear correlation; values near 0 indicate little or no correlation. The probability $Prob_N(|r| > r_o)$ that N measurements of two *uncorrelated* variables would give a value of r larger than any observed value r_o is tabulated in Appendix C. The smaller this probability, the better the evidence that the variables x and y really are correlated. If the probability is less than 5%, we say the correlation is *significant*; if it is less than 1%, we say the correlation is *highly significant.*

Problems for Chapter 9

For Section 9.2: Covariance in Error Propagation

9.1. ★ Calculate the covariance for the following four measurements of two quantities x and y.

x:	20	23	23	22
y:	30	32	35	31

9.2. ★ Each of five students measures the two times (t and T) for a stone to fall from the third and sixth floors of a tall building. Their results are shown in Table 9.5. Calculate the two averages $\bar{t}$ and $\bar{T}$, and find the covariance σ_{tT} using the layout of Table 9.1.

Table 9.5. Five measurements of two times, t and T (in tenths of a second); for Problem 9.2.

Student	t	T
A	14	20
B	12	18
C	13	18
D	15	22
E	16	22

[As you examine the data, note that students B and C get lower-than-average answers for both times, whereas D's and E's answers are both higher than average. Although this difference could be just a chance fluctuation, it suggests B and C may have a systematic tendency to underestimate their times and D and E to overestimate. (For instance, B and C could tend to anticipate the landing, whereas D and E could tend to anticipate the launch.) Under these conditions, we would *expect* to get a correlation of the type observed.]

9.3 ★★ **(a)** For the data of Problem 9.1, calculate the variances σ_x^2 and σ_y^2 and the covariance σ_{xy}. **(b)** If you now decide to calculate the sum $q = x + y$, what will be its standard deviation according to (9.9)? **(c)** What would you have found for the standard deviation if you had ignored the covariance and used Equation (9.10)?

(d) In a simple situation like this, an easier way to find the standard deviation of q is just to calculate four values of q [one for each pair (x, y)] and then find σ_q from these four values. Show that this procedure gives the same answer as you got in part (b).

9.4. ★★ **(a)** For the data of Problem 9.2, calculate the variances σ_t^2 and σ_T^2 and the covariance σ_{tT}. **(b)** If the students decide to calculate the difference $T - t$, what will be its standard deviation according to (9.9)? **(c)** What would they have found for the standard deviation if they had ignored the covariance and used Equation (9.10)? **(d)** In a simple situation like this, an easier way to find the standard deviation of $T - t$ is just to calculate five values of $T - t$ [one for each pair (t, T)] and then find the standard deviation of these five values. Show that this procedure gives the same answer as you got in part (b).

9.5. ★★ Imagine a series of N measurements of two fixed lengths x and y that were made to find the value of some function $q(x, y)$. Suppose each pair is measured with a different tape; that is, the pair (x_1, y_1) is measured with one tape, (x_2, y_2) is measured with a second tape, and so on. **(a)** Assuming the main source of errors is that some of the tapes have shrunk and some stretched (uniformly, in either case), show that the covariance σ_{xy} is bound to be positive. **(b)** Show further, under the same conditions, that $\sigma_{xy} = \sigma_x\sigma_y$; that is, σ_{xy} is as large as permitted by the Schwarz inequality (9.11).

[Hint: Assume that the ith tape has shrunk by a factor λ_i, that is, present length = (design length)/λ_i, so that a length that is really X will be measured as $x_i = \lambda_i X$. The moral of this problem is that there are situations in which the covariance is certainly not negligible.]

9.6. ★★ Here is an example of an experiment in which we would expect a negative correlation between two measured quantities (high values of one correlated with low values of the other). Figure 9.2 represents a photograph taken in a bubble chamber, where charged subatomic particles leave clearly visible tracks. A positive particle called the K^+ has entered the chamber at the bottom of the picture. At point A, it has collided with an invisible neutron (n) and has undergone the reaction

$$K^+ + n \rightarrow K^o + p.$$

The proton's track (p) is clearly visible, going off to the right, but the path of the K^o (shown dotted) is really invisible because the K^o is uncharged. At point B, the K^o decays into two charged pions,

$$K^o \rightarrow \pi^+ + \pi^-,$$

whose tracks are again clearly visible. To investigate the conservation of momentum in the second process, the experimenter needs to measure the angles α and β between the paths of the two pions and the invisible path of the K^o, and this measurement requires drawing in the dotted line that joins A and B. The main source of error in finding α and β is in deciding on the direction of this line, because A and B are often close together (less than a cm), and the tracks that define A and B are rather wide. For the purpose of this problem, let us suppose that this is the *only* source of error.

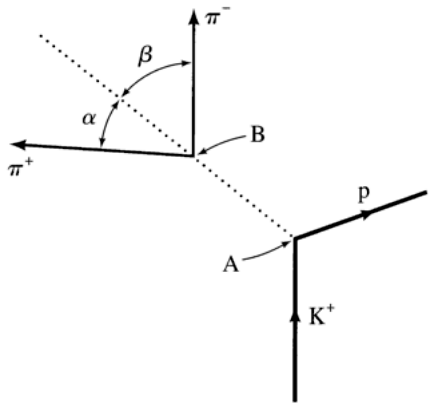


Figure 9.2. Tracks of charged particles in a bubble chamber. The dotted line shows the direction of an invisible K°, which was formed at A and decayed at B; for Problem 9.6.

Suppose several students are given copies of Figure 9.2, and each draws in his best estimate for the line AB and then measures the two angles α and β. The students then combine their results to find the means $\overline{\alpha}$ and $\overline{\beta}$, the standard deviations σ_α and σ_β, and the covariance $\sigma_{\alpha\beta}$. Assuming that the only source of error is in deciding the direction of the line AB, explain why an overestimate of α is inevitably accompanied by an underestimate of β. Prove that $\sigma_\alpha = \sigma_\beta$ and that the covariance $\sigma_{\alpha\beta}$ is negative and equal to the largest value allowed by the Schwartz inequality, $\sigma_{\alpha\beta} = -\sigma_\alpha\sigma_\beta$.

(Hint: Suppose that the ith student draws his line AB to the right of the true direction by an amount Δ_i. Then his value for α will be $\alpha_i = \alpha_{\text{true}} + \Delta_i$. Write the corresponding expression for his value β_i and compute the various quantities of interest in terms of the Δ_i and $\overline{\Delta}$.)

9.7. ★★ Prove that the covariance σ_{xy} defined in (9.8) satisfies the Schwarz inequality (9.11),

$$|\sigma_{xy}| \leqslant \sigma_x\sigma_y. \tag{9.17}$$

[Hint: Let t be an arbitrary number and consider the function

$$A(t) = \frac{1}{N}\Sigma[(x_i - \overline{x}) + t(y_i - \overline{y})]^2 \geqslant 0. \tag{9.18}$$

Because $A(t) \geqslant 0$ whatever the value of t, even its minimum $A_{\min} \geqslant 0$. Find the minimum $A_{\min}$, and set $A_{\min} \geqslant 0$.]

For Section 9.3: Coefficient of Linear Correlation

9.8. ★ Calculate the correlation coefficient r for the following five pairs of measurements:

x	=	1	2	3	4	5
y	=	8	8	5	6	3

Do the calculations yourself, but if your calculator has a built-in function to compute r, make sure you know how it works, and use it to check your value.

9.9. ★ Calculate the correlation coefficient r for the following six pairs of measurements:

$$x = 1 \quad 2 \quad 3 \quad 5 \quad 6 \quad 7$$
$$y = 5 \quad 6 \quad 6 \quad 8 \quad 8 \quad 9$$

Do the calculations yourself, but if your calculator has a built-in function to compute r, make sure you know how it works, and use it to check your value.

9.10. ★★ **(a)** Prove the identity

$$\Sigma(x_i - \bar{x})(y_i - \bar{y}) = \Sigma x_i y_i - N\bar{x}\bar{y}.$$

(b) Hence, prove the correlation coefficient r defined in (9.15) can be written as

$$r = \frac{\Sigma x_i y_i - N\bar{x}\bar{y}}{\sqrt{(\Sigma x_i^2 - N\bar{x}^2)(\Sigma y_i^2 - N\bar{y}^2)}}. \tag{9.19}$$

Many calculators use this result to find r because it avoids the need to store all the data before calculating the means and deviations.

For Section 9.4: Quantitative Significance of r

9.11. ★ In the photoelectric effect, the kinetic energy K of electrons ejected from a metal by light is supposed to be a linear function of the light's frequency f,

$$K = hf - \phi, \tag{9.20}$$

where h and ϕ are constants. To check this linearity, a student measures K for N different values of f and calculates the correlation coefficient r for her results. **(a)** If she makes five measurements ($N = 5$) and finds $r = 0.7$, does she have significant support for the linear relation (9.20)? **(b)** What if $N = 20$ and $r = 0.5$?

9.12. ★ **(a)** Check that the correlation coefficient r for the 10 pairs of test scores in Table 9.3 is approximately 0.8. (By all means, use the built-in function on your calculator, if it has one.) **(b)** Using the table of probabilities in Appendix C, find the probability that 10 *uncorrelated* scores would have given $|r| \geq 0.8$. Is the correlation of the test scores significant? Highly significant?

9.13. ★ A psychologist, investigating the relation between the intelligence of fathers and sons, measures the Intelligence Quotients of 10 fathers and sons and obtains the following results:

Father:	74	83	85	96	98	100	106	107	120	124
Son:	76	103	99	109	111	107	91	101	120	119

Do these data support a correlation between the intelligence of fathers and sons?

9.14. ★ Eight aspiring football players are timed in the 100-meter dash and the 1,500-meter run. Their times (in seconds) are as follows:

Dash:	12	11	13	14	12	15	12	16
Run:	280	290	220	260	270	240	250	230

Calculate the correlation coefficient r. What kind of correlation does your result suggest? Is there, in fact, significant evidence for a correlation?

9.15. ★★ Draw a scatter plot for the six data pairs of Problem 9.9 and the least-squares line that best fits these points. Find their correlation coefficient r. Based on the probabilities listed in Appendix C, would you say these data show a significant linear correlation? Highly significant?

9.16. ★★ **(a)** Draw a scatter plot for the five data pairs of Problem 9.8 and the least-squares line that best fits these points. Find their correlation coefficient r. Based on the probabilities listed in Appendix C, would you say these data show a significant linear correlation? Highly significant? **(b)** Repeat for the following data:

$$x = 1 \quad 2 \quad 3 \quad 4 \quad 5$$
$$y = 4 \quad 6 \quad 3 \quad 0 \quad 2$$

Chapter 10

The Binomial Distribution

The Gauss, or normal, distribution is the only example of a distribution we have studied so far. We will now discuss two other important examples, the binomial distribution (in this chapter) and the Poisson distribution (in Chapter 11).

10.1 Distributions

Chapter 5 introduced the idea of a *distribution,* the function that describes the proportion of times a repeated measurement yields each of its various possible answers. For example, we could make N measurements of the period T of a pendulum and find the distribution of our various measured values of T, or we could measure the heights h of N Americans and find the distribution of the various measured heights h.

I next introduced the notion of the *limiting distribution,* the distribution that would be obtained in the limit that the number of measurements N becomes very large. The limiting distribution can be viewed as telling us the *probability* that one measurement will yield any of the possible values: the probability that one measurement of the period will yield any particular value T; the probability that one American (chosen at random) will have any particular height h. For this reason, the limiting distribution is also sometimes called the *probability distribution.*

Of the many possible limiting distributions, the only one we have discussed is the Gauss, or normal, distribution, which describes the distribution of answers for any measurement subject to many sources of error that are all random and small. As such, the Gauss distribution is the most important of all limiting distributions for the physical scientist and amply deserves the prominence given it here. Nevertheless, several other distributions have great theoretical or practical importance, and this and the next chapter present two examples.

This chapter describes the binomial distribution, which is not of great practical importance to the experimental physicist. Its simplicity, however, makes it an excellent introduction to many properties of distributions, and it is theoretically important, because we can derive the all-important Gauss distribution from it.

10.2 Probabilities in Dice Throwing

The binomial distribution can best be described by an example. Suppose we undertake as our "experiment" to throw three dice and record the number of aces showing. The possible results of the experiment are the answers 0, 1, 2, or 3 aces. If we repeat the experiment an enormous number of times, we will find the limiting distribution, which will tell us the probability that in any one throw (of all three dice) we get ν aces, where $\nu = 0, 1, 2,$ or 3.

This experiment is sufficiently simple that we can easily calculate the probability of the four possible outcomes. We observe first that, assuming the dice are true, the probability of getting an ace when throwing *one* die is $\frac{1}{6}$. Let us now throw all three dice and ask first for the probability of getting three aces ($\nu = 3$). Because each separate die has probability $\frac{1}{6}$ of showing an ace, and because the three dice roll independently, the probability for three aces is

$$Prob(3 \text{ aces in } 3 \text{ throws}) = \left(\frac{1}{6}\right)^3 \approx 0.5\%.$$

Calculating the probability for two aces ($\nu = 2$) is a little harder because we can throw two aces in several ways. The first and second dice could show aces and the third not (A, A, not A), or the first and third could show aces and the second not (A, not A, A), and so on. Here, we argue in two steps. First, we consider the probability of throwing two aces in any definite order, such as (A, A, not A). The probability that the first die will show an ace is $\frac{1}{6}$, and likewise for the second. On the other hand, the probability that the last die will *not* show an ace is $\frac{5}{6}$. Thus, the probability for two aces in this particular order is

$$Prob(A, A, \text{ not } A) = \left(\frac{1}{6}\right)^2 \times \left(\frac{5}{6}\right).$$

The probability for two aces in any other definite order is the same. Finally, there are three different orders in which we could get our two aces: (A, A, not A), or (A, not A, A), or (not A, A, A). Thus, the total probability for getting two aces (in any order) is

$$Prob(2 \text{ aces in } 3 \text{ throws}) = 3 \times \left(\frac{1}{6}\right)^2 \times \left(\frac{5}{6}\right) \approx 6.9\%. \qquad (10.1)$$

Similar calculations give the probabilities for one ace in three throws (34.7%) and for no aces in three throws (57.9%). Our numerical conclusions can be summarized by drawing the probability distribution for the number of aces obtained when throwing three dice, as in Figure 10.1. This distribution is an example of the binomial distribution, the general form of which we now describe.

10.3 Definition of the Binomial Distribution

To describe the general binomial distribution, I need to introduce some terminology. First, imagine making n independent *trials,* such as throwing n dice, tossing n coins, or testing n firecrackers. Each trial can have various outcomes: A die can show any

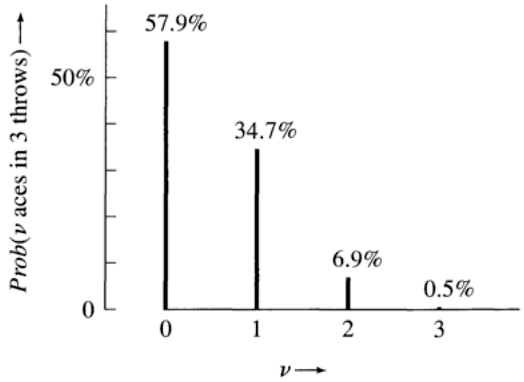


Figure 10.1. Probability of getting ν aces when throwing three dice. This function is the binomial distribution $B_{n,p}(\nu)$, with $n = 3$ and $p = 1/6$.

face from 1 to 6, a coin can show heads or tails, a firecracker can explode or fizzle. We refer to the outcome in which we happen to be interested as a *success.* Thus "success" could be throwing an ace on a die, or a head on a coin, or having a firecracker explode. We denote by p the probability of success in any one trial, and by $q = 1 - p$ that of "failure" (that is, of getting any outcome other than the one of interest). Thus, $p = \frac{1}{6}$ for getting an ace on a die, $p = \frac{1}{2}$ for heads on a coin, and p might be 95% for a given brand of firecracker to explode properly.

Armed with these definitions, we can now ask for the probability of getting ν successes in n trials. A calculation I will sketch in a moment shows that this probability is given by the so-called *binomial distribution:*

$$\begin{aligned} Prob(\nu \text{ successes in } n \text{ trials}) &= B_{n,p}(\nu) \\ &= \frac{n(n-1)\cdots(n-\nu+1)}{1 \times 2 \times \cdots \times \nu} p^{\nu} q^{n-\nu}. \end{aligned} \tag{10.2}$$

Here the letter B stands for "binomial"; the subscripts n and p on $B_{n,p}(\nu)$ indicate that the distribution depends on n, the number of trials made, and p, the probability of success in one trial.

The distribution (10.2) is called the binomial distribution because of its close connection with the well-known binomial expansion. Specifically, the fraction in (10.2) is the *binomial coefficient,* often denoted

$$\binom{n}{\nu} = \frac{n(n-1)\cdots(n-\nu+1)}{1 \times 2 \times \cdots \times \nu} \tag{10.3}$$

$$= \frac{n!}{\nu!(n-\nu)!}, \tag{10.4}$$

where we have introduced the useful *factorial* notation,

$$n! = 1 \times 2 \times \cdots \times n.$$

[By convention, $0! = 1$, and so $\binom{n}{0} = 1$.] The binomial coefficient appears in the binomial expansion

$$\begin{aligned}(p + q)^n &= p^n + np^{n-1}q + \cdots + q^n \\ &= \sum_{\nu=0}^{n} \binom{n}{\nu} p^\nu q^{n-\nu},\end{aligned} \tag{10.5}$$

which holds for any two numbers p and q and any positive integer n (see Problems 10.5 and 10.6).

With the notation (10.3), we can rewrite the binomial distribution in the more compact form

The Binomial Distribution

$$\begin{aligned}Prob(\nu \text{ successes in } n \text{ trials}) &= B_{n,p}(\nu) \\ &= \binom{n}{\nu} p^\nu q^{n-\nu},\end{aligned} \tag{10.6}$$

where, as usual, p denotes the probability of success in one trial and $q = 1 - p$.

The derivation of the result (10.6) is similar to that of the example of the dice in (10.1),

$$Prob(2 \text{ aces in } 3 \text{ throws}) = 3 \times \left(\frac{1}{6}\right)^2 \times \left(\frac{5}{6}\right). \tag{10.7}$$

In fact, if we set $\nu = 2$, $n = 3$, $p = \frac{1}{6}$, and $q = \frac{5}{6}$ in (10.6), we obtain precisely (10.7), as you should check. Furthermore, the significance of each factor in (10.6) is the same as that of the corresponding factor in (10.7). The factor p^ν is the probability of getting all successes in any definite ν trials, and $q^{n-\nu}$ is the probability of failure in the remaining $n - \nu$ trials. The binomial coefficient $\binom{n}{\nu}$ is easily shown to be the number of different orders in which there can be ν successes in n trials. This establishes that the binomial distribution (10.6) is indeed the probability claimed.

Example: Tossing Four Coins

Suppose we toss four coins ($n = 4$) and count the number of heads obtained, ν. What is the probability of obtaining the various possible values $\nu = 0, 1, 2, 3, 4$?

Because the probability of getting a head on one toss is $p = \frac{1}{2}$, the required probability is simply the binomial distribution $B_{n,p}(\nu)$, with $n = 4$ and $p = q = \frac{1}{2}$,

$$Prob(\nu \text{ heads in 4 tosses}) = \binom{4}{\nu}\left(\frac{1}{2}\right)^4.$$

These probabilities are easily evaluated. For example,

$$Prob(0 \text{ heads in 4 tosses}) = 1 \times \left(\frac{1}{2}\right)^4 = 0.0625.$$

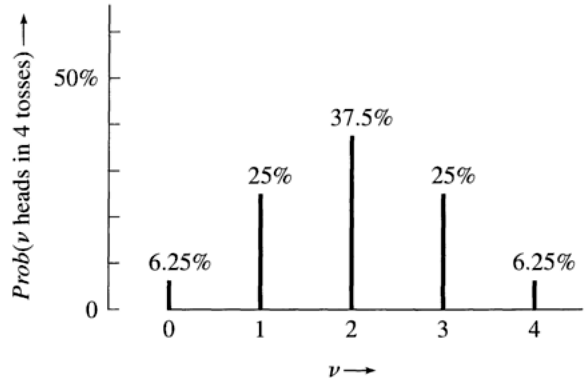


Figure 10.2. The binomial distribution $B_{n,p}(\nu)$ with $n = 4$, $p = \frac{1}{2}$. This gives the probability of getting ν heads when throwing four coins.

All five probabilities are shown in Figure 10.2.

We see that the most probable number of heads is $\nu = 2$, as we would expect. Here, the probabilities are symmetric about this most probable value. That is, the probability for three heads is the same as that for one, and the probability for four heads is the same as that for none. As we will see, this symmetry occurs only when $p = \frac{1}{2}$.

Quick Check 10.1. If you draw one card at random from a full deck of 52 playing cards, the probability of drawing any particular suit is $\frac{1}{4}$. If you draw three times, replacing your card after each draw, what is the probability of drawing three hearts? Of drawing exactly two hearts? Of drawing two or more hearts?

10.4 Properties of the Binomial Distribution

The binomial distribution $B_{n,p}(\nu)$ gives the probability of having ν "successes" in n trials, when p is the probability of success in a single trial. If we repeat our whole experiment, consisting of n trials, many times, then it is natural to ask what our average number of successes $\bar{\nu}$ would be. To find this average, we just sum over all possible values of ν, each multiplied by its probability. That is,

$$\bar{\nu} = \sum_{\nu=0}^{n} \nu B_{n,p}(\nu) \tag{10.8}$$

and is easily evaluated (Problem 10.14) as

$$\bar{\nu} = np. \tag{10.9}$$

That is, if we repeat our series of n trials many times, the average number of successes will be just the probability of success in one trial (p) times n, as you would expect. We can similarly calculate the standard deviation σ_ν in our number of successes (Problem 10.15). The result is

$$\sigma_\nu = \sqrt{np(1-p)}. \tag{10.10}$$

When $p = \frac{1}{2}$ (as in a coin-tossing experiment), the average number of successes is just $n/2$. Furthermore, it is easy to prove for $p = \frac{1}{2}$ that

$$B_{n,1/2}(\nu) = B_{n,1/2}(n - \nu) \tag{10.11}$$

(see Problem 10.13). That is, the binomial distribution with $p = \frac{1}{2}$ is symmetric about the average value $n/2$, as we noticed in Figure 10.2.

In general, when $p \neq \frac{1}{2}$, the binomial distribution $B_{n,p}(\nu)$ is not symmetric. For example, Figure 10.1 is clearly not symmetric; the most probable number of successes is $\nu = 0$, and the probability diminishes steadily for $\nu = 1$, 2, and 3. Also, the average number of successes ($\bar{\nu} = 0.5$) here is not the same as the most probable number of successes ($\nu = 0$).

It is interesting to compare the binomial distribution $B_{n,p}(\nu)$ with the more familiar Gauss distribution $G_{X,\sigma}(x)$. Perhaps the biggest difference is that the experiment described by the binomial distribution has outcomes given by the *discrete*[1] values $\nu = 0, 1, 2, \ldots, n$, whereas those of the Gauss distribution are given by the *continuous* values of the measured quantity x. The Gauss distribution is a symmetric peak centered on the average value $x = X$, which means that the average value X is also the most probable value [that for which $G_{X,\sigma}(x)$ is maximum]. As we have seen, the binomial distribution is symmetric only when $p = \frac{1}{2}$, and in general the average value does not coincide with the most probable value.

GAUSSIAN APPROXIMATION TO THE BINOMIAL DISTRIBUTION

For all their differences, the binomial and Gauss distributions have an important connection. If we consider the binomial distribution $B_{n,p}(\nu)$ for any fixed value of p, then when n is large $B_{n,p}(\nu)$ is closely approximated by the Gauss distribution $G_{X,\sigma}(\nu)$ with the same mean and same standard deviation; that is,

$$B_{n,p}(\nu) \approx G_{X,\sigma}(\nu) \qquad (n \text{ large}) \tag{10.12}$$

with

$$X = np \text{ and } \sigma = \sqrt{np(1-p)}. \tag{10.13}$$

We refer to (10.12) as the Gaussian approximation to the binomial distribution.

---

[1] The word *discrete* (not to be confused with discreet) means "detached from one another" and is the opposite of continuous.

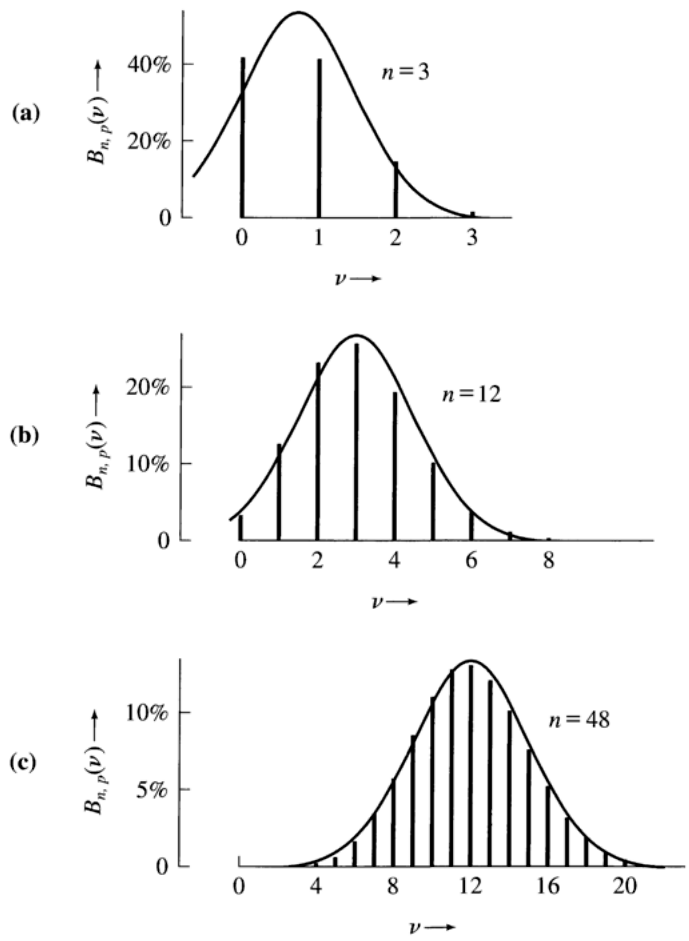


Figure 10.3. The binomial distributions for $p = 1/4$ and $n = 3$, 12, and 48. The continuous curve superimposed on each picture is the Gauss function with the same mean and same standard deviation.

I will not prove the result here,[2] but its truth is clearly illustrated in Figure 10.3, which shows the binomial distribution for $p = \frac{1}{4}$ and for three successively larger values of n ($n = 3$, 12, 48). Superimposed on each binomial distribution is the Gaussian distribution with the same mean and standard deviation. With just three trials ($n = 3$), the binomial distribution is quite different from the corresponding Gaussian. In particular, the binomial distribution is distinctly asymmetric, whereas the Gaussian is, of course, perfectly symmetric about its mean. By the time $n = 12$, the asymmetry of the binomial distribution is much less pronounced, and the two distributions are quite close to one another. When $n = 48$, the difference between the binomial and the corresponding Gauss distribution is so slight that the two are almost indistinguishable on the scale of Figure 10.3(c).

[2]For proofs, see S. L. Meyer, *Data Analysis for Scientists and Engineers* (John Wiley, 1975), p. 226, or H. D. Young, *Statistical Treatment of Experimental Data* (McGraw-Hill, 1962), Appendix C.

That the binomial distribution can be approximated by the Gauss function when n is large is very useful in practice. Calculation of the binomial function with n greater than 20 or so is extremely tedious, whereas calculation of the Gauss function is always simple whatever the values of X and σ. To illustrate this, suppose we wanted to know the probability of getting 23 heads in 36 tosses of a coin. This probability is given by the binomial distribution $B_{36,1/2}(\nu)$ because the probability of a head in one toss is $p = \frac{1}{2}$. Thus

$$Prob(23 \text{ heads in } 36 \text{ tosses}) = B_{36,\,1/2}(23) \tag{10.14}$$

$$= \frac{36!}{23!13!}\left(\frac{1}{2}\right)^{36}, \tag{10.15}$$

which a fairly tedious calculation[3] shows to be

$$Prob(23 \text{ heads}) = 3.36\%.$$

On the other hand, because the mean of the distribution is $np = 18$ and the standard deviation is $\sigma = \sqrt{np(1-p)} = 3$, we can approximate (10.14) by the Gauss function $G_{18,3}(23)$, and a simple calculation gives

$$Prob(23 \text{ heads}) \approx G_{18,3}(23) = 3.32\%.$$

For almost all purposes, this approximation is excellent.

The usefulness of the Gaussian approximation is even more obvious if we want the probability of several outcomes. For example, the probability of getting 23 *or more* heads in 36 tosses is

$$Prob(23 \text{ or more heads}) = Prob(23 \text{ heads}) + Prob(24 \text{ heads}) + \cdots + Prob\ (36 \text{ heads}),$$

a tedious sum to calculate directly. If we approximate the binomial distribution by the Gaussian, however, then the probability is easily found. Because the calculation of Gaussian probabilities treats ν as a continuous variable, the probability for $\nu = 23, 24, \ldots$ is best calculated as $Prob_{\text{Gauss}}(\nu \geqslant 22.5)$, the probability for any $\nu \geqslant 22.5$. Now, $\nu = 22.5$ is 1.5 standard deviations above the mean value, 18. (Remember, $\sigma = 3$, so $4.5 = 1.5\sigma$.) The probability of a result more than 1.5σ above the mean equals the area under the Gauss function shown in Figure 10.4. It is easily calculated with the help of the table in Appendix B, and we find

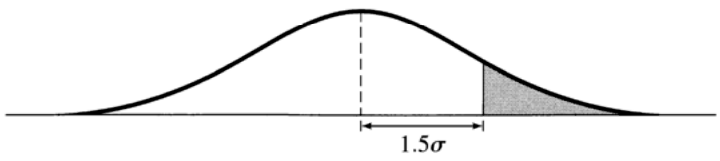


Figure 10.4. The probability of a result more than 1.5σ above the mean is the shaded area under the Gauss curve.

[3] Some hand calculators are preprogrammed to compute $n!$ automatically, and with such a calculator computation of (10.15) is easy. In most such calculators, however, $n!$ overflows when $n \geqslant 70$, so for $n \geqslant 70$ this preprogrammed function is no help.

$$Prob(23 \text{ or more heads}) \approx Prob_{\text{Gauss}}(\nu \geqslant X + 1.5\sigma) = 6.7\%.$$

This value compares well with the exact result (to two significant figures) 6.6%.

10.5 The Gauss Distribution for Random Errors

In Chapter 5, I claimed that a measurement subject to many small random errors will be distributed normally. We are now in a position to prove this claim, using a simple model for the kind of measurement concerned.

Let us suppose we measure a quantity x whose true value is X. We assume our measurements are subject to negligible systematic error but there are n independent sources of random error (effects of parallax, reaction times, and so on). To simplify our discussion, suppose further that all these sources produce random errors of the same fixed size ε. That is, each source of error pushes our result upward or downward by ε, and these two possibilities occur with equal probability, $p = \frac{1}{2}$. For example, if the true value is X and there is just one source of error, our possible answers are $x = X - \varepsilon$ and $x = X + \varepsilon$, both equally likely. If there are two sources of error, a measurement could yield $x = X - 2\varepsilon$ (if both errors happened to be negative), or $x = X$ (if one was negative and one positive), or $x = X + 2\varepsilon$ (if both happened to be positive). These possibilities are shown in Figures 10.5(a) and (b).

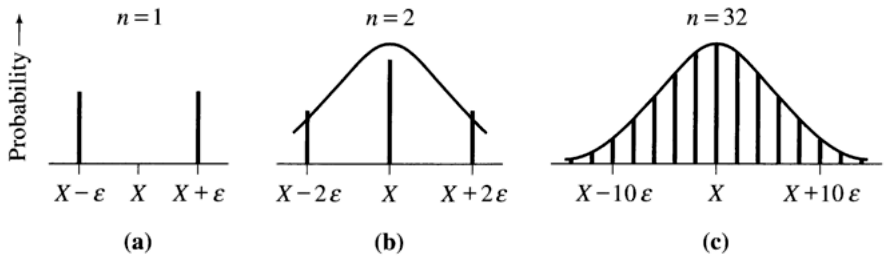


Figure 10.5. Distribution of measurements subject to n random errors of magnitude ε, for $n = 1$, 2, and 32. The continuous curves superimposed on (b) and (c) are Gaussians with the same center and width. (The vertical scales differ in the three graphs.)

In general, if there are n sources of error, our answer could range from $x = X - n\varepsilon$ to $x = X + n\varepsilon$. In a given measurement, if ν sources happen to give positive errors and $(n - \nu)$ negative errors, our answer will be

$$\begin{aligned} x &= X + \nu\varepsilon - (n - \nu)\varepsilon \\ &= X + (2\nu - n)\varepsilon. \end{aligned} \tag{10.16}$$

The probability of this result occurring is just the binomial probability

$$Prob(\nu \text{ positive errors}) = B_{n,1/2}(\nu). \tag{10.17}$$

Thus, the possible results of our measurement are symmetrically distributed around the true value X, and the probabilities are given by the binomial function (10.17).

This distribution is illustrated in Figure 10.5 for $n = 1$, 2, and 32.

I now claim that if the number of sources of error, n, is large and the size of the individual errors, ε, is small, then our measurements are normally distributed. To be more precise, we note that the standard deviation of the binomial distribution is $\sigma_\nu = \sqrt{np(1-p)} = \sqrt{n}/4$. Therefore, according to (10.16), the standard deviation of our measurements of x is $\sigma_x = 2\varepsilon\sigma_\nu = \varepsilon\sqrt{n}$. Accordingly, we let $n \to \infty$ and $\varepsilon \to 0$ in such a way that $\sigma_x = \varepsilon\sqrt{n}$ remains fixed. Two things happen as a result. First, as discussed in the previous section, the binomial distribution approaches the Gauss distribution with center X and width σ_x. This approach is clearly visible in Figures 10.5(b) and (c), on which the appropriate Gauss functions have been superimposed. Second, as $\varepsilon \to 0$, the possible results of our measurement move closer together (as is also clear in Figure 10.5), so that the discrete distribution approaches a continuous distribution that is precisely the expected Gauss distribution.

10.6 Applications; Testing of Hypotheses

Once we know how the results of an experiment should be distributed, we can ask whether the actual results of the experiment *were* distributed as expected. This kind of test of a distribution is an important technique in the physical sciences and perhaps even more so in the biological and social sciences. One important, general test, the chi-squared test, is the subject of Chapter 12. Here, I give two examples of a simpler test that can be applied to certain problems involving the binomial distribution.

TESTING A NEW SKI WAX

Suppose a manufacturer of ski waxes claims to have developed a new wax that greatly reduces the friction between skis and snow. To test this claim, we might take 10 pairs of skis and treat one ski from each pair with the wax. We could then hold races between the treated and untreated members of each pair by letting them slide down a suitable snow-covered incline.

If the treated skis won all 10 races, we would obviously have strong evidence that the wax works. Unfortunately, we seldom get such a clear-cut result, and even when we do, we would like some quantitative measure of the strength of the evidence. Thus, we have to address two questions. First, how can we quantify the evidence that the wax works (or doesn't work)? Second, where would we draw the line? If the treated skis won nine of the races, would this result be conclusive? What if they won eight races? Or seven?

Precisely these same questions arise in a host of similar statistical tests. If we wanted to test the efficacy of a fertilizer, we would organize "races" between treated and untreated plants. To predict which candidate is going to win an election, we would choose a random sample of voters and hold "races" between the candidates among the members of our sample.

To answer our questions, we need to decide more precisely what we should

expect from our tests. In the accepted terminology, we must formulate a *statistical hypothesis.* In the example of the ski wax, the simplest hypothesis is the *null hypothesis,* that the new wax actually makes no difference. Subject to this hypothesis, we can calculate the probability of the various possible results of our test and then judge the significance of our particular result.

Suppose we take as our hypothesis that the ski wax makes no difference. In any one race, the treated and untreated skis would then be equally likely to win; that is, the probability for a treated ski to win is $p = \frac{1}{2}$. The probability that the treated skis will win ν of the 10 races is then the binomial probability:

$$\begin{aligned} \textit{Prob}(\nu \text{ wins in 10 races}) &= B_{10,1/2}(\nu) \\ &= \frac{10!}{\nu!(10-\nu)!}\left(\frac{1}{2}\right)^{10}. \end{aligned} \tag{10.18}$$

According to (10.18), the probability that the treated skis would win all 10 races is

$$\textit{Prob}(10 \text{ wins in 10 races}) = (\tfrac{1}{2})^{10} \approx 0.1\%. \tag{10.19}$$

That is, if our null hypothesis is correct, the treated skis would be *very* unlikely to win all 10 races. Conversely, if the treated skis *did* win all 10 races, the null hypothesis is very unlikely to be correct. In fact, the probability (10.19) is so small, we could say the evidence in favor of the wax is "highly significant," as we will discuss shortly.

Suppose instead that the treated skis had won eight of the 10 races. Here, we would calculate the probability of eight *or more* wins:

$$\begin{aligned} &\textit{Prob}(8 \text{ or more wins in 10 races}) \\ &\quad = \textit{Prob}(8 \text{ wins}) + \textit{Prob}(9 \text{ wins}) + \textit{Prob}(10 \text{ wins}) \approx 5.5\%. \end{aligned} \tag{10.20}$$

For the treated skis to win eight or more races is still quite unlikely but not nearly as unlikely as their winning all 10.

To decide what conclusion to draw from the eight wins, we must recognize that there are really just two alternatives. Either

(a) our null hypothesis is correct (the wax makes no difference), but by chance, an unlikely event has occurred (the treated skis have won eight races)

or

(b) our null hypothesis is false, and the wax does help.

In statistical testing, by tradition we pick some definite probability (5%, for example) to define the boundary below which an event is considered unacceptably improbable. If the probability of the actual outcome (eight or more wins, in our case) is below this boundary, we choose alternative (b), reject the hypothesis, and say the result of our experiment was *significant.*

By common practice, we call a result significant if its probability is less than 5% and call it highly significant if its probability is less than 1%. Because the

probability (10.20) is 5.5%, we see that eight wins for the waxed skis are just *not* enough to give significant evidence that the wax works. On the other hand, we saw that the probability of 10 wins is 0.1%. Because this value is less than 1%, we can say that 10 wins would constitute highly significant evidence that the wax helps.[4]

GENERAL PROCEDURE

The methods of the example just described can be applied to any set of n similar but independent tests (or "races"), each of which has the same two possible outcomes, "success" or "failure." First, a hypothesis is formulated, here simply an assumed value for the probability p of success in any one test. This assumed value of p determines the expected mean number of successes, $\bar{\nu} = np$, in n trials.[5] If the actual number of successes, ν, in our n trials is close to np, there is no evidence against the hypothesis. (If the waxed skis win five out of 10 races, there is no evidence the wax makes any difference.) If ν is appreciably larger than np, we calculate the probability (given our hypothesis) of getting ν or more successes. If this probability is less than our chosen significance level (for example, 5% or 1%), we argue that our observed number is unacceptably improbable (if our hypothesis is correct) and hence that our hypothesis should be rejected. In the same way, if our number of successes ν is appreciably less than np, we can argue similarly, except that we would calculate the probability of getting ν or less successes.[6]

As you should have expected, this procedure does not provide a simple answer that our hypothesis is certainly true or certainly false. But it does give a quantitative measure of the reasonableness of our results in light of the hypothesis, so we can choose an objective, if arbitrary, criterion for rejection of the hypothesis. When experimenters state conclusions based on this kind of reasoning, they must state clearly the criterion used and the calculated probability so that readers can judge the reasonableness of the conclusion for themselves.

AN OPINION POLL

As a second example, consider an election between two candidates, A and B. Suppose candidate A claims that extensive research has established that he is favored by 60% of the electorate, and suppose candidate B asks us to check this claim (in the hope, of course, of showing that the number favoring A is significantly less than 60%).

Here, our statistical hypothesis would be that 60% of voters favor A, so the probability that a randomly selected voter will favor A would be $p = 0.6$. Recognizing that we cannot poll every single voter, we carefully select a random sample of

---

[4] Note the great simplicity of the test just described. We could have measured various additional parameters, such as the time taken by each ski, the maximum speed of each ski, and so on. Instead, we simply recorded which ski won each race. Tests that do not use such additional parameters are called *nonparametric* tests. They have the great advantages of simplicity and wide applicability.

[5] As usual, $\bar{\nu} = np$ is the mean number of successes expected if we were to repeat our whole set of n trials many times.

[6] As we discuss below, in some experiments the relevant probability is the "two-tailed" probability of getting a value of ν that deviates *in either direction* from np by as much as, or more than, the value actually obtained.

600 and ask their preferences. If 60% really favor A, the expected number in our sample who favor A is $np = 600 \times 0.6 = 360$. If in fact 330 state a preference for A, can we claim to have cast significant doubt on the hypothesis that 60% favor A?

To answer this question, we note that (according to the hypothesis) the probability that ν voters will favor A is the binomial probability

$$Prob(\nu \text{ voters favor } A) = B_{n,p}(\nu) \tag{10.21}$$

with $n = 600$ and $p = 0.6$. Because n is so large, it is an excellent approximation to replace the binomial function by the appropriate Gauss function, with center at $np = 360$ and standard deviation $\sigma_\nu = \sqrt{np(1-p)} = 12$.

$$Prob(\nu \text{ voters favor } A) \approx G_{360,12}(\nu). \tag{10.22}$$

The mean number expected to favor A is 360. Thus, the number who actually favored A in our sample (namely 330) is 30 less than expected. Because the standard deviation is 12, our result is 2.5 standard deviations below the supposed mean. The probability of a result this low or lower (according to the table in Appendix B) is 0.6%.[7] Thus, our result is highly significant, and at the 1% level we can confidently reject the hypothesis that A is favored by 60%.

This example illustrates two general features of this kind of test. First, having found that 330 voters favored A (that is, 30 less than expected), we calculated the probability that the number favoring A would be 330 *or less.* At first thought, you might have considered the probability that the number favoring A is precisely $\nu = 330$. This probability is extremely small (0.15%, in fact) and even the most probable result ($\nu = 360$) has a low probability (3.3%). To get a proper measure of how unexpected the result $\nu = 330$ is, we have to include $\nu = 330$ *and* any result that is even further below the mean.

Our result $\nu = 330$ was 30 less than the expected result, 360. The probability of a result 30 or more below the mean is sometimes called a "one-tailed probability," because it is the area under one tail of the distribution curve, as in Figure 10.6(a). In some tests, the relevant probability is the "two-tailed probability" of getting a result that differs from the expected mean by 30 or more *in either direc-*

(a) (b)

Figure 10.6. (a) The "one-tailed" probability for getting a result 30 or more below the mean. (b) The "two-tailed" probability for getting a result that differs from the mean by 30 or more in either direction. (Not to scale.)

[7] Strictly speaking, we should have computed the probability for $\nu \leq 330.5$ because the Gauss distribution treats ν as a continuous variable. This number is 2.46σ below the mean, so the correct probability is actually 0.7%, but this small a difference does not affect our conclusion.

tion, that is, the probability of getting $\nu \leqslant 330$ or $\nu \geqslant 390$, as in Figure 10.6(b). Whether you use the one-tailed or two-tailed probability in a statistical test depends on what you consider the interesting alternative to the original hypothesis. Here, we were concerned to show that candidate *A* was favored by *less* than the claimed 60%, so the one-tailed probability was appropriate. If we were concerned to show that the number favoring *A* was *different* from 60% (in either direction), the two-tailed probability would be appropriate. In practice, the choice of which probability to use is usually fairly clear. In any case, the experimenter always needs to state clearly the probability and significance level chosen and the calculated value of the probability. With this information readers can judge the significance of the results for themselves.

Quick Check 10.2. If I toss a coin 12 times and get 11 heads, do I have significant evidence that the coin is loaded in favor of heads? (Hint: Assuming the coin is true, the probability of getting heads in a single throw is $p = \frac{1}{2}$. Making this assumption, find the probability that I would have obtained 11 or more heads in 12 tosses.

Principal Definitions and Equations of Chapter 10

THE BINOMIAL DISTRIBUTION

We consider an experiment with various possible outcomes and designate the particular outcome (or outcomes) in which we are interested as a "success." If the probability of success in any one trial is p, then the probability of ν successes in n trials is given by the binomial distribution:

$$\begin{aligned} Prob(\nu \text{ successes in } n \text{ trials}) &= B_{n,p}(\nu) \\ &= \binom{n}{\nu} p^{\nu}(1-p)^{n-\nu}, \end{aligned} \qquad \text{[See (10.6)]}$$

where $\binom{n}{\nu}$ denotes the binomial coefficient,

$$\binom{n}{\nu} = \frac{n!}{\nu!\,(n-\nu)!}.$$

If we repeat the whole set of n trials many times, the expected mean number of successes is

$$\bar{\nu} = np \qquad \text{[See (10.9)]}$$

and the standard deviation of ν is

$$\sigma_\nu = \sqrt{np(1-p)}. \qquad \text{[See (10.10)]}$$

THE GAUSSIAN APPROXIMATION TO THE BINOMIAL DISTRIBUTION

When n is large, the binomial distribution $B_{n,p}(\nu)$ is well approximated by the Gauss function with the same mean and standard deviation; that is,

$$B_{n,p}(\nu) \approx G_{X,\sigma}(\nu), \qquad \text{[See (10.12)]}$$

where $X = np$ and $\sigma = \sqrt{np(1-p)}$.

Problems for Chapter 10

For Section 10.2: Probabilities in Dice Throwing

10.1. ★ Consider the experiment of Section 10.2 in which three dice are thrown. Derive the probabilities for throwing no aces and one ace. Verify all four of the probabilities shown in Figure 10.1.

10.2. ★ Compute the probabilities *Prob*(ν aces in two throws) for $\nu = 0$, 1, and 2 in a throw of two dice. Plot them in a histogram.

10.3. ★★ Compute the probabilities *Prob*(ν aces in four throws) for $\nu = 0, 1, \ldots, 4$ in a throw of four dice. Plot them in a histogram. You will need to think carefully about the number of different ways in which you can get two aces in a throw of four dice.

For Section 10.3: Definition of the Binomial Distribution

10.4. ★ **(a)** Compute 5!, 6!, and 25!/23! (Please think for a moment before you do the last one.) **(b)** Explain why we traditionally define 0! = 1. [Hint: According to the usual definition, for $n = 1, 2, \ldots$, the factorial function satisfies $n! = (n+1)!/(n+1)$.] **(c)** Prove that the binomial coefficient defined by Equation (10.3) is equal to

$$\binom{n}{\nu} = \frac{n!}{\nu!(n-\nu)!}.$$

10.5. ★ Compute the binomial coefficients $\binom{3}{\nu}$ for $\nu = 0$, 1, 2, and 3. Hence, write the binomial expansion (10.5) of $(p+q)^3$.

10.6. ★ Compute the binomial coefficients $\binom{4}{\nu}$ for $\nu = 0$, 1, 2, 3, and 4. Hence, write the binomial expansion (10.5) of $(p+q)^4$.

10.7. ★ Compute and plot a histogram of the binomial distribution function $B_{n,p}(\nu)$ for $n = 4$, $p = 1/2$, and all possible ν.

10.8. ★ Compute and plot a histogram of the binomial distribution function $B_{n,p}(\nu)$ for $n = 4$, $p = 1/5$, and all possible ν.

10.9. ★ I draw six times from a deck of 52 playing cards, replacing each card before making the next draw. Find the probabilities, *Prob*(ν hearts), that I would draw exactly ν hearts in six draws, for $\nu = 0, 1, \ldots, 6$.

10.10. ★ A hospital admits four patients suffering from a disease for which the mortality rate is 80%. Find the probabilities of the following outcomes: **(a)** None of the patients survives. **(b)** Exactly one survives. **(c)** Two or more survive.

10.11. ★★ In the game of Yahtzee, a player throws five dice simultaneously. **(a)** Find the probabilities of throwing ν aces, for $\nu = 0, 1, \ldots, 5$. **(b)** What is the probability of throwing three *or more* aces? **(c)** What is the probability of throwing five of a kind?

10.12. ★★ Prove that the binomial coefficient $\binom{n}{\nu}$ is the number of different orders in which ν successes could be obtained in n trials.

For Section 10.4: Properties of the Binomial Distribution

10.13. ★ Prove that for $p = 1/2$, the binomial distribution is symmetric about $\nu = n/2$; that is,

$$B_{n,1/2}(\nu) = B_{n,1/2}(n - \nu).$$

10.14. ★★ Prove that the mean number of successes,

$$\bar{\nu} = \sum_{\nu=0}^{n} \nu B_{n,p}(\nu),$$

for the binomial distribution is just np. [There are many ways to do this, of which one of the best is this: Write the binomial expansion (10.5) for $(p + q)^n$. Because this expansion is true for any p and q, you can differentiate it with respect to p. If you now set $p + q = 1$ and multiply the result by p, you will have the desired result.]

10.15. ★★ **(a)** The standard deviation for any limiting distribution $f(\nu)$ is defined by

$$\sigma_\nu^2 = \overline{(\nu - \bar{\nu})^2}.$$

Prove that this definition is the same as $\overline{\nu^2} - (\bar{\nu})^2$. **(b)** Use this result to prove that, for the binomial distribution,

$$\sigma_\nu^2 = np(1 - p).$$

(Use the same trick as in Problem 10.14, but differentiate twice.)

10.16. ★★ The Gaussian approximation (10.12) to the binomial distribution is excellent for n large and surprisingly good for n small (especially if p is close to $\frac{1}{2}$). To illustrate this claim, calculate $B_{4,1/2}(\nu)$ (for $\nu = 0, 1, \ldots, 4$) both exactly and using the Gaussian approximation. Compare your results.

10.17. ★★ Use the Gaussian approximation (10.12) to find the probability of getting 15 heads if you throw a coin 25 times. Calculate the same probability exactly and compare answers.

10.18. ★★ Use the Gaussian approximation to find the probability of getting 18 or more heads in 25 tosses of a coin. (In using the Gauss probabilities, you should find the probability for $\nu \geqslant 17.5$.) Compare your approximation with the exact answer, which is 2.16%.

For Section 10.6: Applications; Testing of Hypotheses

10.19. ★ In the test of a ski wax described in Section 10.6, suppose the waxed skis had won 9 of the 10 races. Assuming the wax makes no difference, calculate the probability of 9 or more wins. Do 9 wins give significant (5% level) evidence that the wax is effective? Is the evidence highly significant (1% level)?

10.20. ★★ To test a new fertilizer, a gardener selects 14 pairs of similar plants and treats one plant from each pair with the fertilizer. After two months, 12 of the treated plants are healthier than their untreated partners (and the remaining 2 are less healthy). If, in fact, the fertilizer made no difference, what would be the probability that pure chance led to 12 or more successes? Do the 12 successes give significant (5% level) evidence that the fertilizer helps? Is the evidence highly significant (1% level)?

10.21. ★★ Of a certain kind of seed, 25% normally germinates. To test a new germination stimulant, 100 of these seeds are planted and treated with the stimulant. If 32 of them germinate, can you conclude (at the 5% level of significance) that the stimulant helps?

10.22. ★★ In a certain school, 420 of the 600 students pass a standardized mathematics test, for which the national passing rate is 60%. If the students at the school had no special aptitude for the test, how many would you expect to pass, and what is the probability that 420 or more would pass? Can the school claim that its students are significantly well prepared for the test?

Chapter 11

The Poisson Distribution

This chapter presents a third example of a limiting distribution, the Poisson distribution, which describes the results of experiments in which we count events that occur at random but at a definite average rate. Examples of this kind of counting experiment crop up in almost every area of science; for instance, a sociologist might count the number of babies born in a hospital in a three-day period. An important example in physics is the counting of the decays of a radioactive sample; for instance, a nuclear physicist might decide to count the number of alpha particles given off by a sample of radon gas in a ten-second interval.

This kind of counting experiment was discussed in Section 3.2, where I stated but did not prove the "square-root rule": If you count the occurrences of an event of this type in a chosen time interval T and obtain ν counts, then the best estimate for the true average number in time T is, of course, ν, and the uncertainty in this estimate is $\sqrt{\nu}$.

In Sections 11.1 and 11.2, I introduce the Poisson distribution and explore some of its properties. In particular, I prove in Section 11.2 that the standard deviation of the Poisson distribution is the square root of the expected number of events. This result justifies the square-root rule of Section 3.2. Sections 11.3 and 11.4 describe some applications of the Poisson distribution.

11.1 Definition of the Poisson Distribution

As an example of the Poisson distribution, suppose we are given a sample of radioactive material and use a suitable detector to find the number ν of decay particles ejected in a two-minute interval. If the counter is reliable, our value of ν will have no uncertainty. Nevertheless, if we repeat the experiment, we will almost certainly get a different value for ν. This variation in the number ν does not reflect uncertainties in our counting; rather, it reflects the intrinsically random character of the radioactive decay process.

Each radioactive nucleus has a definite probability for decaying in any two-minute interval. If we knew this probability and the number of nuclei in our sample, we could calculate the *expected average number* of decays in two minutes. Nevertheless, each nucleus decays at a random time, and in any given two-minute interval, the number of decays may be different from the expected average number.

Obviously the question we should ask is this: If we repeat our experiment many times (replenishing our sample if it becomes significantly depleted), what distribution should we expect for the number of decays ν observed in two-minute intervals? If you have studied Chapter 10, you will recognize that the required distribution is the binomial distribution. If there are n nuclei and the probability that any one nucleus decays is p, then the probability of ν decays is just the probability of ν "successes" in n "trials," or $B_{n,p}(\nu)$. In the kind of experiment we are now discussing, however, there is an important simplification. The number of "trials" (that is, nuclei) is enormous ($n \sim 10^{20}$, perhaps), and the probability of "success" (decay) for any one nucleus is tiny (often as small as $p \sim 10^{-20}$). Under these conditions (n large and p small), the binomial distribution can be shown to be indistinguishable from a simpler function called the Poisson distribution. Specifically, it can be shown that

$$Prob(\nu \text{ counts in any definite interval}) = P_\mu(\nu), \tag{11.1}$$

where the *Poisson distribution*, $P_\mu(\nu)$, is given by

The Poisson Distribution

$$P_\mu(\nu) = e^{-\mu}\frac{\mu^\nu}{\nu!}. \tag{11.2}$$

In this definition, μ is a positive parameter ($\mu > 0$) that, as I will show directly, is just the expected mean number of counts in the time interval concerned, and $\nu!$ denotes the factorial function (with $0! = 1$).

SIGNIFICANCE OF μ AS THE EXPECTED MEAN COUNT

I will not derive the Poisson distribution (11.2) here but simply assert that it *is* the appropriate distribution for the kind of counting experiment in which we are interested.[1] To establish the significance of the parameter μ in (11.2), we have only to calculate the average number of counts, $\bar{\nu}$, expected if we repeat our counting experiment many times. This average is found by summing over all possible values of ν, each multiplied by its probability:

$$\bar{\nu} = \sum_{\nu=0}^{\infty} \nu P_\mu(\nu) = \sum_{\nu=0}^{\infty} \nu e^{-\mu}\frac{\mu^\nu}{\nu!}. \tag{11.3}$$

The first term of this sum can be dropped (because it is zero), and $\nu/\nu!$ can be replaced by $1/(\nu - 1)!$. If we remove a common factor of $\mu e^{-\mu}$, we get

$$\bar{\nu} = \mu e^{-\mu} \sum_{\nu=1}^{\infty} \frac{\mu^{\nu-1}}{(\nu-1)!}. \tag{11.4}$$

---

[1] For derivations, see, for example, H. D. Young, *Statistical Treatment of Experimental Data* (McGraw-Hill, 1962), Section 8, or S. L. Meyer, *Data Analysis for Scientists and Engineers* (John Wiley, 1975), p. 207.

The infinite sum that remains is

$$1 + \mu + \frac{\mu^2}{2!} + \frac{\mu^3}{3!} + \cdots = e^{\mu}, \tag{11.5}$$

which is just the exponential function e^{μ} (as indicated). Thus, the exponential $e^{-\mu}$ in (11.4) is exactly canceled by the sum, and we are left with the simple conclusion that

$$\boxed{\bar{\nu} = \mu.} \tag{11.6}$$

That is, the parameter μ that characterizes the Poisson distribution $P_{\mu}(\nu)$ is just the *average number of counts expected if we repeat the counting experiment many times.*

Sometimes, we may know in advance the average rate R at which the events we are counting should occur. In this case, the expected average number of events in a time T is just

$$\mu = \text{rate} \times \text{time} = RT.$$

Conversely, if the rate R is unknown, then by counting the number of events in a time T, we can get an estimate for μ and hence for the rate R as $R_{\text{best}} = \mu_{\text{best}}/T$.

Example: Counting Radioactive Decays

Careful measurements have established that a sample of radioactive thorium emits alpha particles at a rate of 1.5 per minute. If I count the number of alpha particles emitted in two minutes, what is the expected average result? What is the probability that I would actually get this number? What is the probability for observing ν particles for $\nu = 0, 1, 2, 3, 4$, and for $\nu \geq 5$?

The expected average count is just the average rate of emissions ($R = 1.5$ per minute) multiplied by the time during which I make my observations ($T = 2$ minutes):

$$(\text{expected average number}) = \mu = 3.$$

This result does not mean, of course, that I should expect to observe exactly three particles in any single trial. On the contrary, the probabilities for observing any number (ν) of particles are given by the Poisson distribution

$$\textit{Prob}(\nu \text{ particles}) = P_3(\nu) = e^{-3}\frac{3^{\nu}}{\nu!}.$$

In particular, the probability that I would observe exactly three particles is

$$\textit{Prob}(3 \text{ particles}) = P_3(3) = e^{-3}\frac{3^3}{3!} = 0.22 = 22\%.$$

Notice that although the expected average result is $\nu = 3$, we should expect to get this number only about once in every five trials.

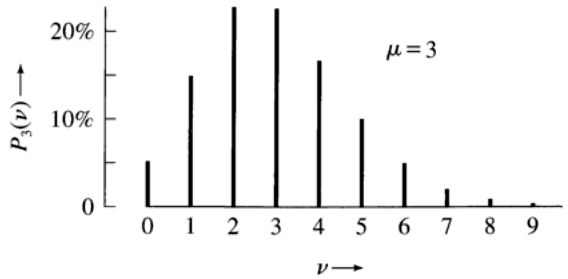


Figure 11.1. The Poisson distribution $P_3(\nu)$ gives the probabilities of observing ν events in a counting experiment for which the expected average count is 3.

The probabilities for any number ν can be calculated in the same way and are (as you might want to check):

Number ν:	0	1	2	3	4
Probability:	5%	15%	22%	22%	17%

These probabilities (up to $\nu = 9$) are plotted in Figure 11.1. The simplest way to find the probability for getting 5 or more counts is to add the probabilities for $\nu = 0, \ldots, 4$ and then subtract the sum from 100% to give[2]

$$\begin{aligned} Prob(\nu \geqslant 5) &= 100\% - (5 + 15 + 22 + 22 + 17)\% \\ &= 19\%. \end{aligned}$$

Quick Check 11.1. On average, each of the 18 hens in my henhouse lays 1 egg per day. If I check the hens once an hour and remove any eggs that have been laid, what is the average number, μ, of eggs that I find on my hourly visits? Use the Poisson distribution $P_\mu(\nu)$ to calculate the probabilities that I would find ν eggs for $\nu = 0$, 1, 2, 3, and $\nu = 4$ or more. What is the most probable number? What is the probability that I would find exactly μ eggs? Verify the probabilities shown in Figure 11.2.

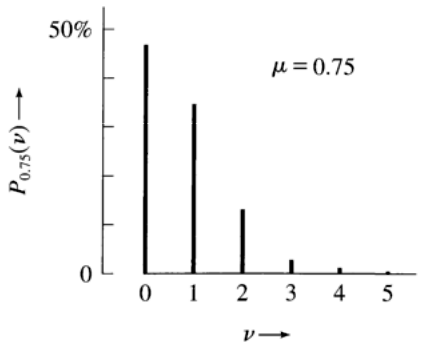


Figure 11.2. The Poisson distribution $P_{0.75}(\nu)$ gives the probabilities of observing ν events in a counting experiment for which the expected average count is 0.75.

---

[2]The correct answer is actually 18.48%, as you can check by keeping a couple of extra decimal places in all the probabilities.

11.2 Properties of the Poisson Distribution

THE STANDARD DEVIATION

The Poisson distribution $P_\mu(\nu)$ gives the probability of getting the result ν in an experiment in which we count events that occur at random but at a definite average rate. We have seen that the parameter μ is precisely the expected average count $\bar{\nu}$. The natural next question is to ask for the standard deviation of the counts ν when we repeat the experiment many times. The standard deviation of any distribution (after a large number of trials) is just the root-mean-square deviation from the mean. That is,

$$\sigma_\nu^2 = \overline{(\nu - \bar{\nu})^2},$$

or, using the result of Problems 10.15(a) or 4.5(a),

$$\sigma_\nu^2 = \overline{\nu^2} - (\bar{\nu})^2. \tag{11.7}$$

For the Poisson distribution, we have already found that $\bar{\nu} = \mu$ and a similar calculation (Problem 11.9) gives $\overline{\nu^2} = \mu^2 + \mu$. Therefore, Equation (11.7) implies that $\sigma_\nu^2 = \mu$ or

$$\boxed{\sigma_\nu = \sqrt{\mu}.} \tag{11.8}$$

That is, the Poisson distribution with mean count μ has standard deviation $\sqrt{\mu}$.

The result (11.8) justifies the square-root rule of Section 3.2. If we carry out a counting experiment once and get the answer ν, we can easily see (using the principle of maximum likelihood, as in Problem 11.11) that the best estimate for the expected mean count is $\mu_{\text{best}} = \nu$. From (11.8), it immediately follows that the best estimate for the standard deviation is just $\sqrt{\nu}$. In other words, if we make one measurement of the number of events in a time interval T and get the answer ν, our answer for the expected mean count in time T is

$$\nu \pm \sqrt{\nu}. \tag{11.9}$$

This answer is precisely the square-root rule quoted without proof in Equation (3.2).

Example: More Radioactive Decays

A student monitors the thorium sample of the previous example for 30 minutes and observes 49 alpha particles. What is her answer for the number of particles emitted in 30 minutes? What is her answer for the rate of emission, R, in particles per minute?

According to (11.9), her answer for the number of particles emitted in 30 minutes is

$$\text{(number emitted in 30 minutes)} = 49 \pm \sqrt{49} = 49 \pm 7.$$

To find the rate in particles per minute, she must divide by 30 minutes. Assuming this 30 minutes has no uncertainty, we find

$$R = \frac{49 \pm 7}{30} = 1.6 \pm 0.2 \text{ particles/min.} \tag{11.10}$$

Notice that the square-root rule gives the uncertainty in the actual counted number ($\sigma_\nu = \sqrt{\nu} = 7$, in this case). A common mistake is to calculate the rate of decay $R = \nu/T$ and then to take the uncertainty in R to be $\sqrt{R}$. A glance at (11.10) should convince you that this procedure is simply not correct. The square-root rule applies only to the actual counted number ν, and the uncertainty in $R = \nu/T$ must be found from that in ν using error propagation as in (11.10).

Quick Check 11.2. The farmer of Quick Check 11.1 observes that in a certain ten-hour period his hens lay 9 eggs. Based on this one observation, what would you quote for the number of eggs expected in ten hours? What would you give for the rate R of egg production, in eggs per hour? (Give the uncertainties in both your answers.)

GAUSSIAN APPROXIMATION TO THE POISSON DISTRIBUTION

In Chapter 10, we compared the Gauss distribution with the binomial distribution. We saw that in most ways, the two distributions are very different; nevertheless, under the right conditions, the Gauss distribution gives an excellent, extremely useful, approximation to the binomial distribution. As we will now see, almost exactly the same can be said about the Gauss and *Poisson* distributions.

The Gauss distribution $G_{X,\sigma}(x)$ gives the probabilities of the various values of a *continuous* variable x; by contrast, the Poisson distribution $P\mu(\nu)$, like the binomial $B_{n,p}(\nu)$, gives the probabilities for a *discrete* variable $\nu = 0, 1, 2, 3, \ldots$. Another important difference is that the Gauss distribution $G_{X,\sigma}(x)$ is specified by *two* parameters, the mean X and the standard deviation σ, whereas the Poisson distribution $P_\mu(\nu)$ is specified by a single parameter, the mean μ, because the width of the Poisson distribution is automatically determined by the mean (namely, $\sigma_\nu = \sqrt{\mu}$). Finally, the Gauss distribution is always bell shaped and symmetric about its mean value, whereas the Poisson distribution has neither of these properties in general. This last point is especially clear in Figure 11.2, which shows the Poisson distribution for $\mu = 0.75$; this curve is certainly not bell shaped, nor is it even approximately symmetric about its mean, 0.75.

Figure 11.1 showed the Poisson distribution for $\mu = 3$. Although this curve is obviously not exactly bell shaped, it is undeniably more nearly so than the curve for $\mu = 0.75$ in Figure 11.2. Figure 11.3 shows the Poisson distribution for $\mu = 9$; this curve is quite nearly bell shaped and close to symmetric about its mean ($\mu = 9$). In fact, it can be proved that as $\mu \to \infty$, the Poisson distribution becomes progres-

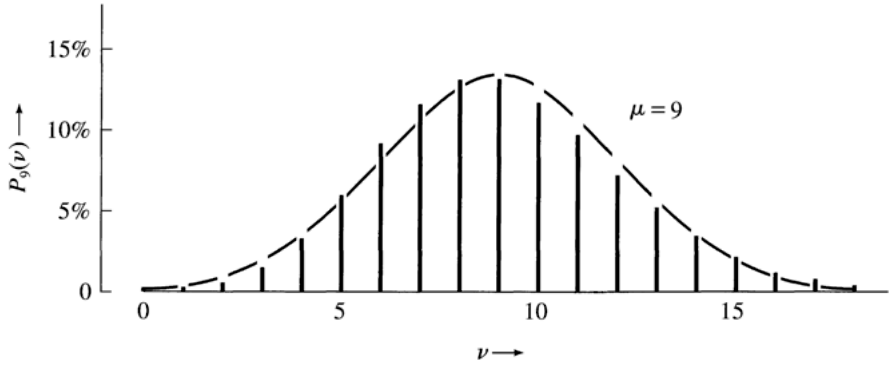


Figure 11.3. The Poisson distribution for $\mu = 9$. The dashed curve is the Gauss distribution with the same mean and standard deviation ($X = 9$ and $\sigma = 3$). As $\mu \to \infty$, the two distributions become indistinguishable; even when $\mu = 9$, they are very close.

sively more bell shaped and approaches the Gauss distribution with the same mean and standard deviation.[3] That is,

$$P_\mu(\nu) \approx G_{X,\sigma}(\nu), \qquad \text{(when } \mu \text{ is large)} \tag{11.11}$$

where

$$X = \mu \qquad \text{and} \qquad \sigma = \sqrt{\mu}.$$

In Figure 11.3, the dashed curve is the Gauss function with $X = 9$ and $\sigma = 3$. You can see clearly how, even when μ is only 9, the Poisson distribution is remarkably close to the appropriate Gauss function; the slight discrepancy reflects the remaining asymmetry in the Poisson function.

The approximation (11.11) is called the *Gaussian approximation to the Poisson distribution.* It is analogous to the corresponding approximation for the binomial distribution (discussed in Section 10.4) and is useful under the same conditions, namely, when the parameters involved are large.

Example: Gaussian Approximation to a Poisson Distribution

To illustrate the Gaussian approximation to the Poisson distribution, consider the Poisson distribution with $\mu = 64$. The probability of 72 counts, for example, is

$$Prob(72 \text{ counts}) = P_{64}(72) = e^{-64}\frac{(64)^{72}}{72!}, \tag{11.12}$$

which a tedious calculation gives as

$$Prob(72 \text{ counts}) = 2.9\%.$$

[3] For proof, see S. L. Meyer, *Data Analysis for Scientists and Engineers* (John Wiley, 1975), p. 227.

According to (11.11), however, the probability (11.12) is well approximated by the Gauss function

$$Prob(72 \text{ counts}) \approx G_{64,8}(72),$$

which is easily evaluated to give

$$Prob(72 \text{ counts}) \approx 3.0\%.$$

If we wanted to calculate directly the probability of 72 *or more* counts in the same experiment, an extremely tedious calculation would give

$$\begin{aligned} Prob(\nu \geq 72) &= P_{64}(72) + P_{64}(73) + \ldots \\ &= 17.3\%. \end{aligned}$$

If we use the approximation (11.11), then we have only to calculate the Gaussian probability for getting $\nu \geq 71.5$ (because the Gauss distribution treats ν as a continuous variable). Because 71.5 is 7.5, or 0.94σ, above the mean, the required probability can be found quickly from the table in Appendix B as

$$\begin{aligned} Prob(\nu \geq 72) \approx Prob_G(\nu \geq 71.5) &= Prob_G(\nu \geq X + 0.94\sigma) \\ &= 17.4\%, \end{aligned}$$

by almost any standard an excellent approximation.

11.3 Applications

As I have emphasized, the Poisson distribution describes the distribution of results in a counting experiment in which events are counted that occur at random but at a definite average rate. In an introductory physics laboratory, the two most common examples are counting the disintegrations of radioactive nuclei and counting the arrival of cosmic ray particles.

Another very important example is an experiment to study an expected limiting distribution, such as the Gauss or binomial distributions, or the Poisson distribution itself. A limiting distribution tells us how many events of a particular type are expected when an experiment is repeated several times. (For example, the Gaussian $G_{X,\sigma}(x)$ tells us how many measurements of x are expected to fall in any interval from $x = a$ to $x = b$.) In practice, the observed number is seldom exactly the expected number. Instead, it fluctuates in accordance with the Poisson distribution. In particular, if the expected number of events of some type is n, the observed number can be expected to differ from n by a number of order $\sqrt{n}$. We will make use of this point in Chapter 12.

In many situations, it is reasonable to expect numbers to be distributed approximately according to the Poisson distribution. The number of eggs laid in an hour on a poultry farm and the number of births in a day at a hospital would both be expected to follow the Poisson distribution at least approximately (though they would probably show some seasonal variations as well). To test this assumption,

you would need to record the number concerned many times over. After plotting the resulting distribution, you could compare it with the Poisson distribution to see how close the fit is. For a more quantitative test, you would use the chi-squared test described in Chapter 12.

Example: Cosmic Ray Counting

As another example of the Poisson distribution, let us consider an experiment with cosmic rays. These "rays" originate as charged particles, such as protons and alpha particles, that enter the Earth's atmosphere from space. Many of these primary particles collide with atoms in the atmosphere and create further, secondary particles, such as mesons and positrons. Some of the particles (both primary and secondary) travel all the way to ground level and can be detected (with a Geiger counter, for example) in the laboratory. In the following problem, I exploit the fact that the number of cosmic rays hitting any given area in a given time should follow the Poisson distribution.

Student A asserts that he has measured the number of cosmic rays hitting a Geiger counter in one minute. He claims to have made the measurement repeatedly and carefully and to have found that, on average, 9 particles hit the counter per minute, with "negligible" uncertainty. To check this claim, Student B counts how many particles arrive in one minute and gets the answer 12. Does this answer cast serious doubt on A's claim that the expected rate is 9? To make a more careful check, Student C counts the number of particles that arrive in ten minutes. From A's claim, she expects to get 90 but actually gets 120. Does this value cast significant doubt on A's claim?

Let us consider B's result first. If A is right, the expected mean count is 9. Because the distribution of counts should be the Poisson distribution, the standard deviation is $\sqrt{9} = 3$. Student B's result of 12 is, therefore, only one standard deviation away from the mean of 9. This amount is certainly not far enough away to contradict A's claim. More specifically, knowing that the probability of any answer ν is supposed to be $P_9(\nu)$, we can calculate the total probability for getting an answer that differs from 9 by 3 or more. This probability turns out to be 40% (see Problem 11.7). Obviously B's result is not at all surprising, and A has no reason to worry.

Student C's result is quite a different matter. If A is right, C should expect to get 90 counts in ten minutes. Since the distribution should be Poisson, the standard deviation should be $\sqrt{90} = 9.5$. Thus, C's result of 120 is more than *three standard deviations* away from A's prediction of 90. With these large numbers, the Poisson distribution is indistinguishable from the Gauss function, and we can immediately find from the table in Appendix A that the probability of a count more than three standard deviations from the mean is 0.3%. That is, if A is right, it is extremely improbable that C would have observed 120 counts. Turning this statement around, we can say something almost certainly has gone wrong. Perhaps A was just not as careful as he claimed. Perhaps the counter was malfunctioning for A or C, introducing systematic errors into one of the results. Or perhaps A made his measurements at a time when the flux of cosmic rays was truly less than normal.

11.4 Subtracting a Background

To conclude this chapter, I discuss a problem that complicates many counting experiments. Often, the events we want to study are accompanied by other "background" events that cannot be distinguished from the events of interest. For example, in studying the disintegrations of a radioactive source, we usually cannot prevent the detector from registering particles from other radioactive materials in the vicinity or from cosmic rays. This means that the number we count includes the events of interest *plus* these background events, and we must somehow subtract out the unwanted background events. In principle at least, the remedy is straightforward: Having found the total counting rate (due to source and background), we must remove the source and find the rate of events due to the background alone; the rate of events from the source is then just the difference of these two measured rates.

In practice, it is surprisingly easy to make a mistake in this procedure, especially in the error analysis. It is usually convenient to measure the total and background counts using different time intervals. Suppose we count a total of ν_{tot} events (source *plus* background) in a time T_{tot}, and then ν_{bgd} background events in a time T_{bgd}. Obviously, we do not simply subtract ν_{bgd} from ν_{tot} because they refer to different time intervals. Instead, we must first calculate the *rates*

$$R_{\text{tot}} = \frac{\nu_{\text{tot}}}{T_{\text{tot}}} \quad \text{and} \quad R_{\text{bgd}} = \frac{\nu_{\text{bgd}}}{T_{\text{bgd}}} \tag{11.13}$$

and then calculate the rate from the source as the difference

$$R_{\text{sce}} = R_{\text{tot}} - R_{\text{bgd}}. \tag{11.14}$$

In estimating the uncertainties in the quantities involved, you must remember that the square-root rule gives the uncertainties in the *counted numbers* ν_{tot} and ν_{bgd}; the uncertainties in the corresponding rates must be found by error propagation, as in the following example.

Example: Radioactive Decays with a Background

A student decides to monitor the activity of a radioactive source by placing it in a liquid scintillation detector. In the course of 10 minutes, the detector registers 2,540 total counts. To allow for the possibility of unwanted background counts, she removes the source and notes that in 3 minutes, the detector registers a further 95 counts. To find the activity of the source, she calculates the two rates of counting, R_{tot} and R_{bgd} (in counts per minute) and their difference $R_{\text{sce}} = R_{\text{tot}} - R_{\text{bgd}}$. What are her answers with their uncertainties? (Assume the two times have negligible uncertainty.)

According to the square-root rule, the two counted numbers with their uncertainties are

$$\nu_{\text{tot}} = 2{,}540 \pm \sqrt{2{,}540} = 2{,}540 \pm 50$$

and

$$\nu_{\text{bgd}} = 95 \pm \sqrt{95} = 95 \pm 10.$$

Dividing these numbers by their corresponding times, we find the rates

$$R_{\text{tot}} = \frac{\nu_{\text{tot}}}{T_{\text{tot}}} = \frac{2{,}540 \pm 50}{10} = 254 \pm 5 \text{ counts/min}$$

and

$$R_{\text{bgd}} = \frac{\nu_{\text{bgd}}}{T_{\text{bgd}}} = \frac{95 \pm 10}{3} = 32 \pm 3 \text{ counts/min.}$$

Finally, the rate due to the source alone is

$$R_{\text{sce}} = R_{\text{tot}} - R_{\text{bgd}} = (254 \pm 5) - (32 \pm 3) = 222 \pm 6 \text{ counts/min.}$$

Notice that, in the last step, the errors are combined in quadrature, because they are certainly independent and random.

Principal Definitions and Equations of Chapter 11

THE POISSON DISTRIBUTION

The Poisson distribution describes experiments in which you count events that occur at random but at a definite average rate. If you count for a chosen time interval T, the probability of observing ν events is given by the Poisson function

$$Prob(\nu \text{ counts in time } T) = P_\mu(\nu) = e^{-\mu}\frac{\mu^\nu}{\nu!}, \qquad \text{[See (11.2)]}$$

where the parameter μ is the expected average number of events in time T; that is,

$$\bar{\nu} = \mu \ (\textit{after many trials}) = RT, \qquad \text{[See (11.6)]}$$

where R is the mean rate at which the events occur.

The standard deviation of the observed number ν is

$$\sigma_\nu = \sqrt{\mu}. \qquad \text{[See (11.8)]}$$

THE GAUSSIAN APPROXIMATION TO THE POISSON DISTRIBUTION

When μ is large, the Poisson distribution $P_\mu(\nu)$ is well approximated by the Gauss function with the same mean and standard deviation; that is,

$$P_\mu(\nu) \approx G_{X,\sigma}(\nu),$$

where $X = \mu$ and $\sigma = \sqrt{\mu}$. [See (11.11)]

SUBTRACTING A BACKGROUND

The events produced by a source subject to an unavoidable background can be counted in a three-step procedure:

(1) Count the total number ν_{tot} (source plus background) in a time T_{tot}, and calculate the total rate $R_{\text{tot}} = \nu_{\text{tot}}/T_{\text{tot}}$.
(2) Remove the source, and measure the number of background events in a time T_{bgd}; then calculate the background rate $R_{\text{bgd}} = \nu_{\text{bgd}}/T_{\text{bgd}}$.
(3) Calculate the rate of the events from the source as the difference $R_{\text{sce}} = R_{\text{tot}} - R_{\text{bgd}}$.

Finally, the uncertainties in the numbers ν_{tot} and ν_{bgd} are given by the square-root rule, and, from these values, the uncertainties in the three rates can be found using error propagation.

Problems for Chapter 11

For Section 11.1: Definition of the Poisson Distribution

11.1. ★ Compute the Poisson distribution $P_\mu(\nu)$ for $\mu = 0.5$ and $\nu = 0, 1, \ldots, 6$. Plot a bar histogram of $P_{0.5}(\nu)$ against ν.

11.2. ★ **(a)** Compute the Poisson distribution $P_\mu(\nu)$ for $\mu = 1$ and $\nu = 0, 1, \ldots, 6$, and plot your results as a bar histogram. **(b)** Repeat part (a) but for $\mu = 2$.

11.3. ★★ A radioactive sample contains 5.0×10^{19} atoms, each of which has a probability $p = 3.0 \times 10^{-20}$ of decaying in any given five-second interval. **(a)** What is the expected average number, μ, of decays from the sample in five seconds? **(b)** Compute the probability $P_\mu(\nu)$ of observing ν decays in any five-second interval, for $\nu = 0, 1, 2, 3$. **(c)** What is the probability of observing 4 or more decays in any five-second interval?

11.4. ★★ In the course of four weeks, a farmer finds that between 10:00 and 10:30 A.M., his hens lay an average of 2.5 eggs. Assuming the number of eggs laid follows a Poisson distribution with $\mu = 2.5$, on approximately how many days do you suppose he found no eggs laid between 10:00 and 10:30 A.M.? On how many days do you suppose there were 2 or less? 3 or more?

11.5. ★★ A certain radioactive sample is expected to undergo three decays per minute. A student observes the number ν of decays in 100 separate one-minute intervals, with the results shown in Table 11.1. **(a)** Make a histogram of these re-

Table 11.1. Occurrences of numbers of decays in one-minute intervals; for Problem 11.5.

No. of decays ν	0	1	2	3	4	5	6	7	8	9
Times observed	5	19	23	21	14	12	3	2	1	0

sults, plotting f_ν (the fraction of times the result ν was found) against ν. **(b)** On the same plot, show the expected distribution $P_3(\nu)$. Do the data seem to fit the expected distribution? (For a quantitative measure of the fit, you could use the chi-squared test, discussed in Chapter 12.)

11.6. ★★★ **(a)** The Poisson distribution, like all distributions, must satisfy a "normalization condition,"

$$\sum_{\nu=0}^{\infty} P_\mu(\nu) = 1. \tag{11.15}$$

This condition asserts that the total probability of observing *all* possible values of ν must be one. Prove it. [Remember the infinite series (11.5) for e^μ.] **(b)** Differentiate (11.15) with respect to μ, and then multiply the result by μ to give an alternative proof that, after infinitely many trials, $\bar{\nu} = \mu$ as in Equation (11.6).

For Section 11.2: Properties of the Poisson Distribution

11.7. ★ **(a)** What is the standard deviation σ_ν (after a large number of trials) of the observed counts ν in a counting experiment in which the expected average count is $\mu = 9$? **(b)** Compute the probabilities $P_9(\nu)$ of obtaining ν counts for $\nu = 7, 8, \ldots, 11$. **(c)** Hence, find the probability of getting a count ν that differs from the expected mean by one or more standard deviations. **(d)** Would a count of 12 cause you to doubt that the expected mean really is 9?

11.8. ★ A nuclear physicist monitors the disintegrations of a radioactive sample with a Geiger counter. She counts the disintegrations in 15 separate five-second intervals and gets the following numbers:

7, 11, 10, 7, 5, 7, 6, 12, 12, 7, 18, 12, 13, 12, 6.

(a) Her best estimate μ_{best} for the true mean count μ is the average of her 15 counts. (For a proof of this claim, see Problem 11.15.) What is her value for μ_{best}? **(b)** The standard deviation σ_ν of her 15 counts should be close to $\sqrt{\mu}$. What is her standard deviation and how does it compare with $\sqrt{\mu_{\text{best}}}$?

11.9. ★★ **(a)** Prove that the average value of ν^2 for the Poisson distribution $P_\mu(\nu)$ is $\overline{\nu^2} = \mu^2 + \mu$. [The easiest way to do this is probably to differentiate the identity (11.15) twice with respect to μ.] **(b)** Hence, prove that the standard deviation of ν is $\sigma_\nu = \sqrt{\mu}$. [Use the identity (11.7).]

11.10. ★★ The average rate of disintegrations from a certain radioactive sample is known to be roughly 20 per minute. If you wanted to measure this rate within 4%, for approximately how long would you plan to count?

11.11. ★★ Consider a counting experiment governed by the Poisson distribution $P_\mu(\nu)$, where the mean count μ is unknown, and suppose that I make a single count and get the value ν. Write down the probability $Prob(\nu)$ for getting this value ν. According to the principle of maximum likelihood, the best estimate for the unknown μ is that value of μ for which $Prob(\nu)$ is largest. Prove that the best estimate μ_{best} is precisely the observed count ν as you would expect. (In this calculation, the unknown μ is the variable and ν is the fixed value I obtained in my one experiment.)

11.12. ★★ The expected mean count in a certain counting experiment is $\mu = 16$. **(a)** Use the Gaussian approximation (11.11) to estimate the probability of getting a count of 10 in any one trial. Compare your answer with the exact value $P_{16}(10)$. **(b)** Use the Gaussian approximation to estimate the probability of getting a count of 10 *or less.* (Remember to compute the Gaussian probability for $\nu \leqslant 10.5$ to allow for the fact that the Gauss distribution treats ν as a continuous variable.) Calculate the exact answer and compare.

Note how, even with μ as small as 16, the Gaussian approximation gives quite good answers and—at least in part (b)—is significantly less trouble to compute than an exact calculation using the Poisson distribution.

11.13. ★★ The expected mean count in a certain counting experiment is $\mu = 400$. Use the Gaussian approximation to find the probability of getting a count anywhere in the range $380 \leqslant \nu \leqslant 420$. Compare your answer with the exact answer, which is 69.47%. Discuss the feasibility of finding this exact probability using your calculator.

11.14. ★★★ In the experiment of Problem 11.5, the expected mean count was known in advance to be $\mu = 3$. Therefore, the mean of the data should be close to 3 and the standard deviation should be close to $\sqrt{3}$. Find the mean and standard deviation of the data and compare them with their expected values.

11.15. ★★★ Consider a counting experiment governed by the Poisson distribution $P_\mu(\nu)$, where the mean count μ is unknown, and suppose that you make N separate trials and get the values

$$\nu_1, \nu_2, \ldots, \nu_N.$$

Write the probability $Prob(\nu_1, \ldots, \nu_N)$ for getting this particular set of values. According to the principle of maximum likelihood, the best estimate for the unknown μ is that value of μ for which $Prob(\nu_1, \ldots, \nu_N)$ is largest. Prove that the best estimate μ_{best} is the average of the observed numbers

$$\mu_{\text{best}} = \frac{1}{N} \sum_{i=1}^{N} \nu_i$$

as you would expect. (In this calculation, the unknown μ is the variable, and $\nu_1, \ldots, \nu_N$ are fixed at the values you obtained in your N trials.)

For Section 11.3: Applications

11.16. ★ Using a Geiger counter, a student records 25 cosmic-ray particles in 15 seconds. **(a)** What is her best estimate for the true mean number of particles in 15 seconds, with its uncertainty? **(b)** What should she report for the *rate* (in particles per minute) with its uncertainty?

11.17. ★ A student counts 400 disintegrations from a radioactive sample in 40 seconds. **(a)** What is his best estimate for the true mean number of disintegrations in 40 seconds, with its uncertainty? **(b)** What should he report for the *rate* of disintegrations (per second) with its uncertainty?

11.18. ★★ Sean and Maria monitor the same radioactive sample. Sean counts for one minute and gets 121 counts. Maria counts for six minutes and gets 576 counts. Each then calculates the *rate* in counts per minute. Do their results agree satisfactorily?

For Section 11.4: Subtracting a Background

11.19. ★★ A heating duct at a nuclear plant is suspected of being contaminated with radioactive dust. To check this suspicion, a health physicist places a detector inside the duct and records 7,075 particles of radiation in 120 minutes. To check for background radiation, he moves his detector some distance from the duct and shields it from any radiation coming from the duct; in this new position, he records 3,015 particles in 60 minutes. What is his final answer for the radiation rate (in particles per minute) due to the contents of the duct alone? Does he have significant evidence for radioactive material in the duct?

11.20. ★★ To measure the activity of a rock thought to be radioactive, a physicist puts the rock beside a detector and counts 225 particles in 10 minutes. To check for background, she removes the rock and then records 90 particles in 6 minutes. She converts both these answers into rates, in particles per hour, and takes their difference to give the activity of the rock alone. What is her final answer, in particles per hour, and what is its uncertainty? Does she have significant evidence that the rock is radioactive?

11.21. ★★★ A physicist needs an accurate measurement of the activity of a radioactive source. To plan the best use of his time, he first makes quick approximate measurements of the total counting rate (source plus background) and the background rate. These approximations give him rough values for the true rates, r_{tot} and r_{bgd}. (Notice that I have used lower-case r for the true rates; this will let you use upper-case R for the values he measures in the main experiment.) To get more accurate values, he plans to count the total number ν_{tot} (source plus background) in a long time T_{tot} and the background number ν_{bgd} in a long time T_{bgd}. From these measurements, he will calculate the rates R_{tot} and R_{bgd} and finally the rate R_{sce} due to the source alone.

The detector he is using is available only for a time T, so that the two times T_{tot} and T_{bgd} are constrained not to exceed

$$T_{\text{tot}} + T_{\text{bgd}} = T. \tag{11.16}$$

Therefore, he must choose the two times T_{tot} and T_{bgd} to minimize the uncertainty in his final answer for R_{sce}.

(a) Show that his final uncertainty will be minimum if he chooses the two times such that

$$\frac{T_{\text{tot}}}{T_{\text{bgd}}} = \sqrt{\frac{r_{\text{tot}}}{r_{\text{bgd}}}}. \tag{11.17}$$

[Hint: Write the expected numbers ν_{tot} and ν_{bgd} in terms of the true rates, r_{tot} and r_{bgd}, and the corresponding times. From these calculate all the uncertainties in-

volved. To minimize the uncertainty in R_{sce}, eliminate T_{bgd} using (11.16), and then differentiate with respect to T_{tot} and set the derivative equal to zero.] Because his preliminary measurements of r_{tot} and r_{bgd} tell him the approximate value of the ratio on the right side, (11.17) tells him how to apportion the available time between T_{tot} and T_{bgd}.

(b) If the available time t is two hours and his preliminary measurements have shown that $r_{tot}/r_{bgd} \approx 9$, how should he choose T_{tot} and T_{bgd}?

Chapter 12

The Chi-Squared Test for a Distribution

By now, you should be reasonably familiar with the notion of limiting distributions. These are the functions that describe the expected distribution of results if an experiment is repeated many times. There are many different limiting distributions, corresponding to the many different kinds of experiments possible. Perhaps the three most important limiting distributions in physical science are the three we have already discussed: the Gauss (or normal) function, the binomial distribution, and the Poisson distribution.

This final chapter focuses on how to decide whether the results of an actual experiment are governed by the expected limiting distribution. Specifically, let us suppose that we perform some experiment for which we believe we know the expected distribution of results. Suppose further that we repeat the experiment several times and record our observations. The question we now address is this: How can we decide whether our observed distribution is consistent with the expected theoretical distribution? We will see that this question can be answered using a simple procedure called the *chi-squared,* or χ^2, *test.* (The Greek letter χ is spelled "chi" and pronounced "kie.")

12.1 Introduction to Chi Squared

Let us begin with a concrete example. Suppose we make 40 measurements $x_1, \ldots, x_{40}$ of the range x of a projectile fired from a certain gun and get the results shown in Table 12.1. Suppose also we have reason to believe these measurements are governed by a Gauss distribution $G_{X,\sigma}(x)$, as is certainly very natural. In this type

Table 12.1. Measured values of x (in cm).

731	772	771	681	722	688	653	757	733	742
739	780	709	676	760	748	672	687	766	645
678	748	689	810	805	778	764	753	709	675
698	770	754	830	725	710	738	638	787	712

of experiment, we usually do not know in advance either the center X or the width σ of the expected distribution. Our first step, therefore, is to use our 40 measurements to compute best estimates for these quantities:

$$(\text{best estimate for } X) = \bar{x} = \frac{\sum x_i}{40} = 730.1 \text{ cm} \tag{12.1}$$

and

$$(\text{best estimate for } \sigma) = \sqrt{\frac{\sum (x_i - \bar{x})^2}{39}} = 46.8 \text{ cm}. \tag{12.2}$$

Now we can ask whether the actual distribution of our results $x_1, \ldots, x_{40}$ is consistent with our hypothesis that our measurements were governed by the Gauss distribution $G_{X,\sigma}(x)$ with X and σ as estimated. To answer this question, we must compute how we would expect our 40 results to be distributed if the hypothesis is true and compare this expected distribution with our actual observed distribution. The first difficulty is that x is a continuous variable, so we cannot speak of the expected number of measurements equal to any one value of x. Rather, we must discuss the expected number in some interval $a < x < b$. That is, we must divide the range of possible values into *bins*. With 40 measurements, we might choose bin boundaries at $X - \sigma$, X, and $X + \sigma$, giving four bins as in Table 12.2.

Table 12.2. A possible choice of bins for the data of Table 12.1. The final column shows the number of observations that fell into each bin.

Bin number k	Values of x in bin		Observations O_k
1	$x < X - \sigma$	(or $x < 683.3$)	8
2	$X - \sigma < x < X$	(or $683.3 < x < 730.1$)	10
3	$X < x < X + \sigma$	(or $730.1 < x < 776.9$)	16
4	$X + \sigma < x$	(or $776.9 < x$)	6

We will discuss later the criteria for choosing bins. In particular, they must be chosen so that all bins contain several measured values x_i. In general, I will denote the number of bins by n; for this example with four bins, $n = 4$.

Having divided the range of possible measured values into bins, we can now formulate our question more precisely. First, we can count the number of measurements that fall into each bin k.[1] We denote this number by O_k (where O stands for "observed number"). For the data of our example, the observed numbers O_1, O_2, O_3, O_4 are shown in the last column of Table 12.2. Next, assuming our measurements are distributed normally (with X and σ as estimated), we can calculate the *expected* number E_k of measurements in each bin k. We must then decide how well the observed numbers O_k compare with the expected numbers E_k.

[1] If a measurement falls exactly on the boundary between two bins, we can assign half a measurement to each bin.

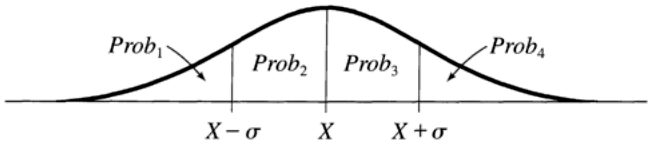


Figure 12.1. The probabilities $Prob_k$ that a measurement falls into each of the bins, $k = 1, 2, 3, 4$, of Table 12.2 are the four areas shown under the Gauss function.

The calculation of the expected numbers E_k is quite straightforward. The *probability* that any one measurement falls in an interval $a < x < b$ is just the area under the Gauss function between $x = a$ and $x = b$. In this example, the probabilities $Prob_1$, $Prob_2$, $Prob_3$, $Prob_4$ for a measurement to fall into each of our four bins are the four areas indicated in Figure 12.1. The two equal areas $Prob_2$ and $Prob_3$ together represent the well-known 68%, so the probability for falling into one of the two central bins is 34%; that is, $Prob_2 = Prob_3 = 0.34$. The outside two areas comprise the remaining 32%; thus $Prob_1 = Prob_4 = 0.16$. To find the expected numbers E_k, we simply multiply these probabilities by the total number of measurements, $N = 40$. Therefore, our expected numbers are as shown in the third column of Table 12.3. That the numbers E_k are not integers serves to remind us that the "expected number" is not what we actually expect in any one experiment; it is rather the expected average number after we repeat our whole series of measurements many times.

Our problem now is to decide how well the expected numbers E_k do represent the corresponding observed numbers O_k (in the last column of Table 12.3). We

Table 12.3. The expected numbers E_k and the observed numbers O_k for the 40 measurements of Table 12.1, with bins chosen as in Table 12.2.

Bin number k	Probability $Prob_k$	Expected number $E_k = NProb_k$	Observed number O_k
1	16%	6.4	8
2	34%	13.6	10
3	34%	13.6	16
4	16%	6.4	6

would obviously not expect *perfect* agreement between E_k and O_k after any finite number of measurements. On the other hand, if our hypothesis that our measurements are normally distributed is correct, we would expect that, in some sense, the deviations

$$O_k - E_k \tag{12.3}$$

would be *small.* Conversely, if the deviations $O_k - E_k$ prove to be *large,* we would suspect our hypothesis is incorrect.

To make precise the statements that the deviation $O_k - E_k$ is "small" or "large," we must decide how large we would *expect* $O_k - E_k$ to be if the measurements really are normally distributed. Fortunately, this decision is easily made. If we imagine repeating our whole series of 40 measurements many times, then the number O_k of measurements in any one bin k can be regarded as the result of a counting experiment of the type described in Chapter 11. Our many different answers for O_k should have an average value of E_k and would be expected to fluctuate around E_k with a standard deviation of order $\sqrt{E_k}$. Thus, the two numbers to be compared are the deviation $O_k - E_k$ and the expected size of its fluctuations $\sqrt{E_k}$.

These considerations lead us to consider the ratio

$$\frac{O_k - E_k}{\sqrt{E_k}}. \tag{12.4}$$

For some bins k, this ratio will be positive, and for some negative; for a few k, it may be appreciably larger than one, but for most it should be of order one, or smaller. To test our hypothesis (that the measurements are normally distributed), it is natural to square the number (12.4) for each k and then sum over all bins $k = 1, \ldots, n$ (here $n = 4$). This procedure defines a number called *chi squared,*

$$\chi^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k}. \tag{12.5}$$

This number χ^2 is clearly a reasonable indicator of the agreement between the observed and expected distributions. If $\chi^2 = 0$, the agreement is perfect; that is, $O_k = E_k$ for all bins k, a situation most unlikely to occur. In general, the individual terms in the sum (12.5) are expected to be of order one, and there are n terms in the sum. Thus, if

$$\chi^2 \leqslant n$$

(χ^2 of order n or less), the observed and expected distributions agree about as well as could be expected. In other words, if $\chi^2 \leqslant n$, we have no reason to doubt that our measurements were distributed as expected. On the other hand, if

$$\chi^2 \gg n$$

(χ^2 significantly greater than the number of bins), the observed and expected numbers differ significantly, and we have good reason to suspect that our measurements were not governed by the expected distribution.

In our example, the numbers observed and expected in the four bins and their differences are shown in Table 12.4, and a simple calculation using them gives

$$\begin{aligned} \chi^2 &= \sum_{k=1}^{4} \frac{(O_k - E_k)^2}{E_k} \\ &= \frac{(1.6)^2}{6.4} + \frac{(-3.6)^2}{13.6} + \frac{(2.4)^2}{13.6} + \frac{(-0.4)^2}{6.4} \\ &= 1.80. \end{aligned} \tag{12.6}$$

Table 12.4. The data of Table 12.1, shown here with the differences $O_k - E_k$.

Bin number k	Observed number O_k	Expected number $E_k = NProb_k$	Difference $O_k - E_k$
1	8	6.4	1.6
2	10	13.6	−3.6
3	16	13.6	2.4
4	6	6.4	−0.4

Because the value of 1.80 for χ^2 is less than the number of terms in the sum (namely, 4), we have no reason to doubt our hypothesis that our measurements were distributed normally.

Quick Check 12.1. Each of the 100 students in a class measures the time for a ball to fall from a third-story window. They calculate their mean $\bar{t}$ and standard deviation σ_t and then group their measurements into four bins, chosen as in the example just discussed. Their results are as follows:

less than $(\bar{t} - \sigma_t)$: 19
between $(\bar{t} - \sigma_t)$ and $\bar{t}$: 30
between $\bar{t}$ and $(\bar{t} + \sigma_t)$: 37
more than $(\bar{t} + \sigma_t)$: 14.

Assuming their measurements are normally distributed, what are the expected numbers of measurements in each of the four bins? What is χ^2, and is there reason to doubt that the measurements *are* distributed normally?

12.2 General Definition of Chi Squared

The discussion so far has focused on one particular example, 40 measurements of a continuous variable x, which denoted the range of a projectile fired from a certain gun. We defined the number χ^2 and saw that it is at least a rough measure of the agreement between our observed distribution of measurements and the Gauss distribution we expected our measurements to follow. We can now define and use χ^2 in the same way for many different experiments.

Let us consider any experiment in which we measure a number x and for which we have reason to expect a certain distribution of results. We imagine repeating the measurement many times (N) and, having divided the range of possible results x into n bins, $k = 1, \ldots, n$, we count the number O_k of observations that actually fall into each bin k. Assuming the measurements really are governed by the expected

distribution, we next calculate the expected number E_k of measurements in the kth bin. Finally, we calculate χ^2 exactly as in (12.5),

$$\chi^2 = \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k}. \tag{12.7}$$

The approximate significance of χ^2 is always the same as in our previous example. That is, if $\chi^2 \leqslant n$, the agreement between our observed and expected distributions is acceptable; if $\chi^2 \gg n$, there is significant disagreement.

The procedure for choosing the bins in terms of which χ^2 is computed depends somewhat on the nature of the particular experiment. Specifically, it depends on whether the measured quantity x is continuous or discrete. I will discuss these two situations in turn.

MEASUREMENTS OF A CONTINUOUS VARIABLE

The example discussed in Section 12.1 involved a continuous variable x, and little more needs to be said. The only limiting distribution we have discussed for a continuous variable is the Gauss distribution, but there are, of course, many different distributions that can occur. For example, in many atomic and nuclear experiments, the expected distribution of the measured variable x (actually an energy) is the Lorentzian distribution

$$f(x) \propto \frac{1}{(x - X)^2 + \gamma^2},$$

where X and γ are certain constants. Another example of a continuous distribution, mentioned in Problem 5.6, is the exponential distribution $\frac{1}{\tau} e^{-t/\tau}$, which gives the probability that a radioactive atom (whose expected mean life is τ) will live for a time t.

Whatever the expected distribution $f(x)$, the total area under the graph of $f(x)$ against x is one, and the probability of a measurement between $x = a$ and $x = b$ is just the area between a and b,

$$Prob(a < x < b) = \int_a^b f(x)\, dx.$$

Thus, if the kth bin runs from $x = a_k$ to $x = a_{k+1}$, the expected number of measurements in the kth bin (after N measurements in all) is

$$\begin{aligned} E_k &= N \times Prob(a_k < x < a_{k+1}) \\ &= N \int_{a_k}^{a_{k+1}} f(x)\, dx. \end{aligned} \tag{12.8}$$

When we discuss the quantitative use of the chi-squared test in Section 12.4, we will see that the expected numbers E_k should not be too small. Although there is no definite lower limit, E_k should probably be approximately five or more,

$$E_k \geqslant 5. \tag{12.9}$$

We must therefore choose bins in such a way that E_k as given by (12.8) satisfies this condition. We will also see that the number of bins must not be too small. For instance, in the example of Section 12.1, where the expected distribution was a Gauss distribution whose center X and width σ were not known in advance, the chi-squared test cannot work (as we will see) with less than four bins; that is, in this example we needed to have

$$n \geqslant 4. \tag{12.10}$$

Combining (12.9) and (12.10), we see that we cannot usefully apply the chi-squared test to this kind of experiment if our total number of observations is less than about 20.

MEASUREMENT OF A DISCRETE VARIABLE

Suppose we measure a discrete variable, such as the now-familiar number of aces when we throw several dice. In practice, the most common discrete variable is an integer (such as the number of aces), and we will denote the discrete variable by v instead of x (which we use for a continuous variable). If we throw five dice, the possible values of v are $v = 0, 1, \ldots, 5$, and we do not actually need to group the possible results into bins. We can simply count how many times we got each of the six possible results. In other words, we can choose six bins, each of which contains just one result.

Nonetheless, it is often desirable to group several different results into one bin. For instance, if we threw our five dice 200 times, then (according to the probabilities found in Problem 10.11) the expected distribution of results is as shown in the first two columns of Table 12.5. We see that here the expected numbers of throws giving four and five aces are 0.6 and 0.03, respectively, both much less than the five or so occurrences required in each bin if we want to use the chi-squared test. This difficulty is easily remedied by grouping the results $v = 3$, 4, and 5 into a single bin. This grouping leaves us with four bins, $k = 1, 2, 3, 4$, which are shown with their corresponding expected numbers E_k, in the last two columns of Table 12.5.

Table 12.5. Expected occurrence of v aces ($v = 0, 1, \ldots, 5$) after throwing five dice 200 times.

Result	Expected occurrences	Bin number k	Expected number E_k
No aces	80.4	1	80.4
One	80.4	2	80.4
Two	32.2	3	32.2
Three	6.4	4	7.0
Four	0.6		
Five	0.03		

Having chosen bins as just described, we could count the observed occurrences O_k in each bin. We could then compute χ^2 and see whether the observed and expected distributions seem to agree. In this experiment, we know that the expected distribution is certainly the binomial distribution $B_{5,1/6}(v)$ *provided* the dice are true

(so that p really is $\frac{1}{6}$). Thus, our test of the distribution is, in this case, a test of whether the dice are true or loaded.

In any experiment involving a discrete variable, the bins can be chosen to contain just one result each, provided the expected number of occurrences for each bin is at least the needed five or so. Otherwise, several different results should be grouped together into a single larger bin that does include enough expected occurrences.

OTHER FORMS OF CHI SQUARED

The notation χ^2 has been used earlier in the book, in Equations (7.6) and (8.5); it could also have been used for the sum of squares in (5.41). In all these cases, χ^2 is a sum of squares with the general form

$$\chi^2 = \sum_1^n \left(\frac{\text{observed value} - \text{expected value}}{\text{standard deviation}}\right)^2. \tag{12.11}$$

In all cases, χ^2 is an indicator of the agreement between the observed and expected values of some variable. If the agreement is good, χ^2 will be of order n; if it is poor, χ^2 will be much greater than n.

Unfortunately, we can use χ^2 to test this agreement only if we know the expected values and the standard deviation, and can therefore calculate (12.11). Perhaps the most common situation in which these values are known accurately enough is the kind of test discussed in this chapter, namely, a test of a distribution, in which E_k is given by the distribution, and the standard deviation is $\sqrt{E_k}$. Nevertheless, the chi-squared test is of very wide application. Consider, for example, the problem discussed in Chapter 8, the measurement of two variables x and y, where y is expected to be some definite function of x,

$$y = f(x)$$

(such as $y = A + Bx$). Suppose we have N measured pairs (x_i, y_i), where the x_i have negligible uncertainty and the y_i have known uncertainties σ_i. Here, the expected value of y_i is $f(x_i)$, and we could test how well y fits the function $f(x)$ by calculating

$$\chi^2 = \sum_1^N \left(\frac{y_i - f(x_i)}{\sigma_i}\right)^2.$$

All our previous remarks about the expected value of χ^2 would apply to this number, and the quantitative tests described in the following sections could be used. This important application will not be pursued here, because only rarely in the introductory physics laboratory would the uncertainties σ_i be known reliably enough (but see Problem 12.14).

12.3 Degrees of Freedom and Reduced Chi Squared

I have argued that we can test agreement between an observed and an expected distribution by computing χ^2 and comparing it with the number of bins used in

collecting the data. A slightly better procedure, however, is to compare χ^2, not with the number of bins n, but instead with the *number of degrees of freedom,* denoted d. The notion of degrees of freedom was mentioned briefly in Section 8.3, and we must now discuss it in more detail.

In general, the number of degrees of freedom d in a statistical calculation is defined as the number of observed data *minus* the number of parameters computed from the data and used in the calculation. For the problems considered in this chapter, the observed data are the numbers of observations O_k in the n bins, $k = 1, \ldots, n$. Thus, the number of observed data is just n, the number of bins. Therefore, in the problems considered here,

$$d = n - c,$$

where n is the number of bins and c is the number of parameters that had to be calculated from the data to compute the expected numbers E_k. The number c is often called the number of *constraints,* as I will explain shortly.

The number of constraints c varies according to the problem under consideration. Consider first the dice-throwing experiment of Section 12.2. If we throw five dice and are testing the hypothesis that the dice are true, the expected distribution of numbers of aces is the binomial distribution $B_{5,1/6}(v)$, where $v = 0, \ldots, 5$ is the number of aces in any one throw. Both parameters in this function—the number of dice, five, and the probability of an ace, $\frac{1}{6}$—are known in advance and do not have to be calculated from the data. When we calculate the expected number of occurrences of any particular v, we must multiply the binomial probability by the total number of throws N (in our example, $N = 200$). This parameter *does* depend on the data. Specifically, N is just the sum of the numbers O_k,

$$N = \sum_{k=1}^{n} O_k. \tag{12.12}$$

Thus, in calculating the expected results of our dice experiment, we have to calculate just one parameter (N) from the data. The number of constraints is, therefore,

$$c = 1,$$

and the number of degrees of freedom is

$$d = n - 1.$$

In Table 12.5, the results of the dice experiment were grouped into four bins (that is, $n = 4$), so that experiment had 3 degrees of freedom.

The equation (12.12) illustrates well the curious terminology of constraints and degrees of freedom. Once the number N has been determined, we can regard (12.12) as an equation that "constrains" the values of $O_1, \ldots, O_n$. More specifically, we can say that, because of the constraint (12.12), only $n - 1$ of the numbers $O_1, \ldots, O_n$ are independent. For instance, the first $n - 1$ numbers $O_1, \ldots, O_{n-1}$ could take any value (within certain ranges), but the last number O_n would be completely determined by Equation (12.12). In this sense, only $n - 1$ of the data are *free* to take on independent values, so we say there are only $n - 1$ independent degrees of freedom.

In the first example in this chapter, the range x of a projectile was measured 40

times ($N = 40$). The results were collected into four bins ($n = 4$) and compared with what we would expect for a Gauss distribution $G_{X,\sigma}(x)$. Here, there were *three* constraints and hence only one degree of freedom,

$$d = n - c = 4 - 3 = 1.$$

The first constraint is the same as (12.12): The total number of observations N is the sum of the observations O_k in all the bins. But here there were two more constraints, because (as is usual in this kind of experiment) we did not know in advance the parameters X and σ of the expected Gauss distribution $G_{X,\sigma}(x)$. Thus, before we could calculate the expected numbers E_k, we had to estimate X and σ using the data. Therefore, there were three constraints in all, so in this example

$$d = n - 3. \tag{12.13}$$

Incidentally, this result explains why we had to use at least four bins in this experiment. We will see that the number of degrees of freedom must always be one or more, so, from (12.13), we clearly had to choose $n \geqslant 4$.

The examples considered here will always have at least one constraint (namely, the constraint $N = \sum O_k$, involving the total number of measurements), and there may be one or two more. Thus, the number of degrees of freedom, d, will range from $n - 1$ to $n - 3$ (in our examples). When n is large, the difference between n and d is fairly unimportant, but when n is small (as it often is, unfortunately), there is obviously a significant difference.

Armed with the notion of degrees of freedom, we can now begin to make our chi-squared test more precise. It can be shown (though I will not do so) that the *expected* value of χ^2 is precisely d, the number of degrees of freedom,

$$(\text{expected average value of } \chi^2) = d. \tag{12.14}$$

This important equation does not mean that we really expect to find $\chi^2 = d$ after any one series of measurements. It means instead that if we could repeat our whole series of measurements infinitely many times and compute χ^2 each time, the average of these values of χ^2 would be d. Nonetheless, even after just *one* set of measurements, a comparison of χ^2 with d is an indicator of the agreement. In particular, if our expected distribution was the *correct* distribution, χ^2 would be very unlikely to be a lot larger than d. Turning this statement around, if we find $\chi^2 \gg d$, we can assert that our expected distribution was most unlikely to be correct.

We have *not* proved the result (12.14), but we can see that some aspects of the result are reasonable. For example, because $d = n - c$, we can rewrite (12.14) as

$$(\text{expected average value of } \chi^2) = n - c. \tag{12.15}$$

That is, for any given n, the expected value of χ^2 will be smaller when c is larger (that is, if we calculate more parameters from the data). This result is just what we should expect. In the example of Section 12.1, we used the data to calculate the center X and width σ of the expected distribution $G_{X,\sigma}(x)$. Naturally, because X and σ were chosen to fit the data, we would expect to find a somewhat better agreement between the observed and expected distributions; that is, these two extra constraints would be expected to reduce the value of χ^2. This reduction is just what (12.15) implies.

The result (12.14) suggests a slightly more convenient way to think about our chi-squared test. We introduce a *reduced chi squared* (or *chi squared per degree of freedom*), which we denote by $\tilde{\chi}^2$ and define as

$$\tilde{\chi}^2 = \chi^2/d. \tag{12.16}$$

Because the expected value of χ^2 is d, we see that the

$$(\text{expected average value of } \tilde{\chi}^2) = 1. \tag{12.17}$$

Thus, whatever the number of degrees of freedom, our test can be stated as follows: If we obtain a value of $\tilde{\chi}^2$ of order one or less, then we have no reason to doubt our expected distribution; if we obtain a value of $\tilde{\chi}^2$ much larger than one, our expected distribution is unlikely to be correct.

Quick Check 12.2. For the experiment of Quick Check 12.1, what is the number of degrees of freedom, and what is the value of the reduced chi squared, $\tilde{\chi}^2$?

12.4 Probabilities for Chi Squared

Our test for agreement between observed data and their expected distribution is still fairly crude. We now need a *quantitative* measure of agreement. In particular, we need some guidance on where to draw the boundary between agreement and disagreement. For example, in the experiment of Section 12.1, we made 40 measurements of a certain range x whose distribution should, we believed, be Gaussian. We collected our data into four bins, and found that $\chi^2 = 1.80$. With three constraints, there was only one degree of freedom ($d = 1$), so the reduced chi squared, $\tilde{\chi}^2 = \chi^2/d$, is also 1.80,

$$\tilde{\chi}^2 = 1.80.$$

The question is now: Is a value of $\tilde{\chi}^2 = 1.80$ sufficiently larger than one to rule out our expected Gauss distribution or not?

To answer this question, we begin by supposing that our measurements *were* governed by the expected distribution (a Gaussian, in this example). With this assumption, we can calculate the *probability* of obtaining a value of $\tilde{\chi}^2$ as large as, or larger than, our value of 1.80. Here, this probability turns out to be

$$Prob(\tilde{\chi}^2 \geq 1.80) \approx 18\%,$$

as we will soon see. That is, if our results were governed by the expected distribution, there would be an 18% probability of obtaining a value of $\tilde{\chi}^2$ greater than or

equal to our actual value 1.80. In other words, in this experiment a value of $\tilde{\chi}^2$ as large as 1.80 is not at all unreasonable, so we would have no reason (based on this evidence) to reject our expected distribution.

Our general procedure should now be reasonably clear. After completing any series of measurements, we calculate the reduced chi squared, which we now denote by $\tilde{\chi}_o^2$ (where the subscript o stands for "observed," because $\tilde{\chi}_o^2$ is the value actually observed). Next, assuming our measurements do follow the expected distribution, we compute the probability

$$Prob(\tilde{\chi}^2 \geq \tilde{\chi}_o^2) \tag{12.18}$$

of finding a value of $\tilde{\chi}^2$ greater than or equal to the observed value $\tilde{\chi}_o^2$. If this probability is high, our value $\tilde{\chi}_o^2$ is perfectly acceptable, and we have no reason to reject our expected distribution. If this probability is unreasonably low, a value of $\tilde{\chi}^2$ as large as our observed $\tilde{\chi}_o^2$ is very unlikely (if our measurements were distributed as expected), and our expected distribution is correspondingly unlikely to be correct.

As always with statistical tests, we have to decide on the boundary between what is reasonably probable and what is not. Two common choices are those already mentioned in connection with correlations. With the boundary at 5%, we would say that our observed value $\tilde{\chi}_o^2$ indicates a significant disagreement if

$$Prob(\tilde{\chi}^2 \geq \tilde{\chi}_o^2) < 5\%,$$

and we would reject our expected distribution at the 5% significance level. If we set the boundary at 1%, then we could say that the disagreement is highly significant if $Prob(\tilde{\chi}^2 \geq \tilde{\chi}_o^2) < 1\%$ and reject the expected distribution at the 1% significance level.

Whatever level you choose as your boundary for rejection, the level chosen should be stated. Perhaps even more important, you should state the probability $Prob(\tilde{\chi}^2 \geq \tilde{\chi}_o^2)$, so that your readers can judge its reasonableness for themselves.

The calculation of the probabilities $Prob(\tilde{\chi}^2 \geq \tilde{\chi}_o^2)$ is too complicated to describe in this book. The results can be tabulated easily, however, as in Table 12.6 or in the more complete table in Appendix D. The probability of getting any particular values of $\tilde{\chi}^2$ depends on the number of degrees of freedom. Thus, we will write the probability of interest as $Prob_d(\tilde{\chi} \geq \tilde{\chi}_o^2)$ to emphasize its dependence on d.

The usual calculation of the probabilities $Prob_d(\tilde{\chi}^2 \geq \tilde{\chi}_o^2)$ treats the observed numbers O_k as continuous variables distributed around their expected values E_k according to a Gauss distribution. In the problems considered here, O_k is a discrete variable distributed according to the Poisson distribution.[2] Provided all numbers involved are reasonably large, the discrete character of the O_k is unimportant, and the Poisson distribution is well approximated by the Gauss function. Under these conditions, the tabulated probabilities $Prob_d(\tilde{\chi}^2 \geq \tilde{\chi}_o^2)$ can be used safely. For this reason, we have said the bins must be chosen so that the expected count E_k in each bin is reasonably large (at least five or so). For the same reason, the number of bins should not be too small.

---

[2] I have argued that finding the number O_k amounts to a counting experiment and hence that O_k should follow a Poisson distribution. If the bin k is too large, then this argument is not strictly correct, because the probability of a measurement in the bin is not much less than one (which is one of the conditions for the Poisson distribution, as mentioned in Section 11.1), so we must have a reasonable number of bins.

Table 12.6. The percentage probability $Prob_d(\tilde{\chi}^2 \geq \tilde{\chi}_o^2)$ of obtaining a value of $\tilde{\chi}^2$ greater than or equal to any particular value $\tilde{\chi}_o^2$, assuming the measurements concerned are governed by the expected distribution. Blanks indicate probabilities less than 0.05%. For a more complete table, see Appendix D.

	$\tilde{\chi}_o^2$												
d	0	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2	3	4	5	6
1	100	62	48	39	32	26	22	19	16	8	5	3	1
2	100	78	61	47	37	29	22	17	14	5	2	0.7	0.2
3	100	86	68	52	39	29	21	15	11	3	0.7	0.2	—
5	100	94	78	59	42	28	19	12	8	1	0.1	—	—
10	100	99	89	68	44	25	13	6	3	0.1	—	—	—
15	100	100	94	73	45	23	10	4	1	—	—	—	—

With these warnings, we now give the calculated probabilities $Prob_d(\tilde{\chi}^2 \geq \tilde{\chi}_o^2)$ for a few representative values of d and $\tilde{\chi}_o^2$ in Table 12.6. The numbers in the left column give six choices of d, the number of degrees of freedom ($d = 1, 2, 3, 5, 10, 15$). Those in the other column heads give possible values of the observed $\tilde{\chi}_o^2$. Each cell in the table shows the percentage probability $Prob_d(\tilde{\chi}^2 \geq \tilde{\chi}_o^2)$ as a function of d and $\tilde{\chi}_o^2$. For example, with 10 degrees of freedom ($d = 10$), we see that the probability of obtaining $\tilde{\chi}^2 \geq 2$ is 3%,

$$Prob_{10}(\tilde{\chi}^2 \geq 2) = 3\%.$$

Thus, if we obtained a reduced chi squared of 2 in an experiment with 10 degrees of freedom, we could conclude that our observations differed significantly from the expected distribution and reject the expected distribution at the 5% significance level (though not at the 1% level).

The probabilities in the second column of Table 12.6 are all 100%, because $\tilde{\chi}^2$ is always certain to be greater than or equal to 0. As $\tilde{\chi}_o^2$ increases, the probability of getting $\tilde{\chi}^2 \geq \tilde{\chi}_o^2$ diminishes, but it does so at a rate that depends on d. Thus, for 2 degrees of freedom ($d = 2$), $Prob_d(\tilde{\chi}^2 \geq 1)$ is 37%, whereas for $d = 15$, $Prob_d(\tilde{\chi}^2 \geq 1)$ is 45%. Note that $Prob_d(\tilde{\chi}^2 \geq 1)$ is always appreciable (at least 32%, in fact), so a value for $\tilde{\chi}_o^2$ of 1 or less is perfectly reasonable and never requires rejection of the expected distribution.

The minimum value of $\tilde{\chi}_o^2$ that does require questioning the expected distribution depends on d. For 1 degree of freedom, we see that $\tilde{\chi}_o^2$ can be as large as 4 before the disagreement becomes significant (5% level). With 2 degrees of freedom, the corresponding boundary is $\tilde{\chi}_o^2 = 3$; for $d = 5$, it is closer to 2 ($\tilde{\chi}_o^2 = 2.2$, in fact), and so on.

Armed with the probabilities in Table 12.6 (and Appendix D), we can now assign a quantitative significance to the value of $\tilde{\chi}_o^2$ obtained in any particular experiment. Section 12.5 gives some examples.

Quick Check 12.3. Each student in a large class times a glider on an air track as it coasts the length of the track. They calculate their mean time and standard deviation and then divide their data into six bins. Assuming their measurements

ought to be normally distributed, they calculate the numbers of measurements expected in each bin and the reduced chi squared, for which they get 4.0. If their measurements really were normally distributed, what would have been the probability of getting a value of $\tilde{\chi}^2$ this large? Is there reason to think the measurements were *not* normally distributed?

12.5 Examples

We have already analyzed rather completely the example of Section 12.1. In this section, we consider three more examples to illustrate the application of the chi-squared test.

Example: Another Example of the Gauss Distribution

The example of Section 12.1 involved a measurement for which the results were expected to be distributed normally. The normal, or Gauss, distribution is so common that we consider briefly another example. Suppose an anthropologist is interested in the heights of the natives on a certain island. He suspects that the heights of the adult males should be normally distributed and measures the heights of a sample of 200 men. Using these measurements, he calculates the mean and standard deviation and uses these numbers as best estimates for the center X and width parameter σ of the expected normal distribution $G_{X,\sigma}(x)$. He now chooses eight bins, as shown in the first two columns of Table 12.7, and groups his observations, with the results shown in the third column.

Table 12.7. Measurements of the heights of 200 adult males.

Bin number k	Heights in bin	Observed number O_k	Expected number E_k
1	less than $X - 1.5\sigma$	14	13.4
2	between $X - 1.5\sigma$ and $X - \sigma$	29	18.3
3	between $X - \sigma$ and $X - 0.5\sigma$	30	30.0
4	between $X - 0.5\sigma$ and X	27	38.3
5	between X and $X + 0.5\sigma$	28	38.3
6	between $X + 0.5\sigma$ and $X + \sigma$	31	30.0
7	between $X + \sigma$ and $X + 1.5\sigma$	28	18.3
8	more than $X + 1.5\sigma$	13	13.4

Our anthropologist now wants to check whether these results are consistent with the expected normal distribution $G_{X,\sigma}(x)$. To this end, he first calculates the probability $Prob_k$ that any one man has height in any particular bin k (assuming a normal distribution). This probability is the integral of $G_{X,\sigma}(x)$ between the bin boundaries and is easily found from the table of integrals in Appendix B. The expected number E_k in each bin is then $Prob_k$ times the total number of men sampled (200). These numbers are shown in the final column of Table 12.7.

To calculate the expected numbers E_k, the anthropologist had to use three parameters calculated from his data (the total number in the sample and his estimates for X and σ). Thus, although there are eight bins, he had three constraints; so the number of degrees of freedom is $d = 8 - 3 = 5$. A simple calculation using the data of Table 12.7 gives for his reduced chi squared

$$\tilde{\chi}^2 = \frac{1}{d}\sum_{i=1}^{8}\frac{(O_k - E_k)^2}{E_k} = 3.5.$$

Because this value is appreciably larger than one, we immediately suspect that the islanders' heights do not follow the normal distribution. More specifically, we see from Table 12.6 that, if the islanders' heights were distributed as expected, then the probability $Prob_5(\tilde{\chi}^2 \geqslant 3.5)$ of obtaining $\tilde{\chi}^2 \geqslant 3.5$ is approximately 0.5%. By any standards, this value is very improbable, and we conclude that the islanders' heights are very unlikely to be normally distributed. In particular, at the 1% (or highly significant) level, we can reject the hypothesis of a normal distribution of heights.

Example: More Dice

In Section 12.2, we discussed an experiment in which five dice were thrown many times and the number of aces in each throw recorded. Suppose we make 200 throws and divide the results into bins as discussed before. Assuming the dice are true, we can calculate the expected numbers E_k as before. These numbers are shown in the third column of Table 12.8.

Table 12.8. Distribution of numbers of aces in 200 throws of 5 dice.

Bin number k	Results in bin	Expected number E_k	Observed number O_k
1	no aces	80.4	60
2	one ace	80.4	88
3	two aces	32.2	39
4	3, 4, or 5 aces	7.0	13

In an actual test, five dice were thrown 200 times and the numbers in the last column of Table 12.8 were observed. To test the agreement between the observed and expected distributions, we simply note that there are three degrees of freedom (four bins minus one constraint) and calculate

$$\tilde{\chi}^2 = \frac{1}{3}\sum_{k=1}^{4}\frac{(O_k - E_k)^2}{E_k} = 4.16.$$

Referring back to Table 12.6, we see that with three degrees of freedom, the probability of obtaining $\tilde{\chi}^2 \geqslant 4.16$ is approximately 0.7%, *if* the dice are true. We conclude that the dice are almost certainly not true. Comparison of the numbers E_k and O_k in Table 12.8 suggests that at least one die is loaded in favor of the ace.

Example: An Example of the Poisson Distribution

As a final example of the use of the chi-squared test, let us consider an experiment in which the expected distribution is the Poisson distribution. Suppose we arrange a Geiger counter to count the arrival of cosmic-ray particles in a certain region. Suppose further that we count the number of particles arriving in 100 separate one-minute intervals, and our results are as shown in the first two columns of Table 12.9.

Table 12.9. Numbers of cosmic-ray particles observed in 100 separate one-minute intervals.

Counts ν in one minute	Occurrences	Bin number k	Observations O_k in bin k	Expected number E_k
None	7	1	7	7.5
One	17	2	17	19.4
Two	29	3	29	25.2
Three	20	4	20	21.7
Four	16	5	16	14.1
Five	8	6	11	12.1
Six	1			
Seven	2			
Eight or more	0			
Total	100			

Inspection of the numbers in column two immediately suggests that we group all counts $\nu \geqslant 5$ into a single bin. This choice of six bins ($k = 1, \ldots, 6$) is shown in the third column and the corresponding numbers O_k in column four.

The hypothesis we want to test is that the number ν is governed by a Poisson distribution $P_\mu(\nu)$. Because the expected mean count μ is unknown, we must first calculate the average of our 100 counts. This value is easily found to be $\bar{\nu} = 2.59$, which gives us our best estimate for μ. Using this value $\mu = 2.59$, we can calculate the probability $P_\mu(\nu)$ of any particular count ν and hence the expected numbers E_k as shown in the final column.

In calculating the numbers E_k, we used two parameters based on the data, the total number of observations (100), and our estimate of μ ($\mu = 2.59$). (Note that because the Poisson distribution is completely determined by μ, we did not have to estimate the standard deviation σ. Indeed, because $\sigma = \sqrt{\mu}$, our estimate for μ automatically gives us an estimate for σ.) There are, therefore, two constraints, which reduces our six bins to four degrees of freedom, $d = 4$.

A simple calculation using the numbers in the last two columns of Table 12.9 now gives for the reduced chi squared

$$\tilde{\chi}^2 = \frac{1}{d} \sum_{k=1}^{6} \frac{(O_k - E_k)^2}{E_k} = 0.35.$$

Because this value is less than one, we can conclude immediately that the agreement between our observations and the expected Poisson distribution is satisfactory. More

specifically, we see from the table in Appendix D that a value of $\tilde{\chi}^2$ as large as 0.35 is very probable; in fact

$$Prob_4(\tilde{\chi}^2 \geq 0.35) \approx 85\%.$$

Thus, our experiment gives us absolutely no reason to doubt the expected Poisson distribution.

The value of $\tilde{\chi}^2 = 0.35$ found in this experiment is actually appreciably less than one, indicating that our observations fit the Poisson distribution very well. This small value does *not,* however, give stronger evidence that our measurements are governed by the expected distribution than would a value $\tilde{\chi}^2 \approx 1$. If the results really are governed by the expected distribution, and if we were to repeat our series of measurements many times, we would expect many different values of $\tilde{\chi}^2$, fluctuating about the average value one. Thus, if the measurements are governed by the expected distribution, a value of $\tilde{\chi}^2 = 0.35$ is just the result of a large chance fluctuation away from the expected mean value. In no way does it give extra weight to our conclusion that our measurements do seem to follow the expected distribution.

If you have followed these three examples, you should have no difficulty applying the chi-squared test to any problems likely to be found in an elementary physics laboratory. Several further examples are included in the problems below. You should certainly test your understanding by trying some of them.

Principal Definitions and Equations of Chapter 12

DEFINITION OF CHI SQUARED

If we make n measurements for which we know, or can calculate, the expected values and the standard deviations, then we define χ^2 as

$$\chi^2 = \sum_1^n \left(\frac{\text{observed value} - \text{expected value}}{\text{standard deviation}}\right)^2. \qquad \text{[See 12.11)]}$$

In the experiments considered in this chapter, the n measurements were the numbers, $O_1, \ldots, O_n$, of times that the value of some quantity x was observed in each of n bins. In this case, the expected number E_k is determined by the assumed distribution of x, and the standard deviation is just $\sqrt{E_k}$; therefore,

$$\chi^2 = \sum_{k=1}^n \frac{(O_k - E_k)^2}{E_k}. \qquad \text{[See (12.7)]}$$

If the assumed distribution of x is correct, then χ^2 should be of order n. If $\chi^2 \gg n$, the assumed distribution is probably incorrect.

DEGREES OF FREEDOM AND REDUCED CHI SQUARED

If we were to repeat the whole experiment many times, the mean value of χ^2 should be equal to d, the number of *degrees of freedom,* defined as

$$d = n - c,$$

where c is the number of *constraints,* the number of parameters that had to be calculated from the data to compute χ^2.

The *reduced* χ^2 is defined as

$$\tilde{\chi}^2 = \chi^2/d. \qquad \text{[See (12.16)]}$$

If the assumed distribution is correct, $\tilde{\chi}^2$ should be of order 1; if $\tilde{\chi}^2 \gg 1$, the data do not fit the assumed distribution satisfactorily.

PROBABILITIES FOR CHI SQUARED

Suppose you obtain the value $\tilde{\chi}_o^2$ for the reduced chi squared in an experiment. If $\tilde{\chi}_o^2$ is appreciably greater than one, you have reason to doubt the distribution on which your expected values E_k were based. From the table in Appendix D, you can find the probability,

$$Prob_d(\tilde{\chi}^2 \geq \tilde{\chi}_o^2),$$

of getting a value $\tilde{\chi}^2$ as large as $\tilde{\chi}_o^2$, assuming the expected distribution is correct. If this probability is small, you have reason to reject the expected distribution; if it is less than 5%, you would reject the assumed distribution at the 5%, or significant, level; if the probability is less than 1%, you would reject the distribution at the 1%, or highly significant, level.

Problems for Chapter 12

For Section 12.1: Introduction to Chi Squared

12.1. ★ Each member of a class of 50 students is given a piece of the same metal (or what is said to be the same metal) and told to find its density ρ. From the 50 results, the mean $\bar{\rho}$ and standard deviation σ_ρ are calculated, and the class decides to test whether the results are normally distributed. To this end, the measurements are grouped into four bins with boundaries at $\bar{\rho} - \sigma_\rho$, $\bar{\rho}$, and $\bar{\rho} + \sigma_\rho$, and the results are shown in Table 12.10.

Table 12.10. Observed densities of 50 pieces of metal arranged in four bins; for Problem 12.1.

Bin number k	Range of bin	Observed number O_k
1	less than $\bar{\rho} - \sigma_\rho$	12
2	between $\bar{\rho} - \sigma_\rho$ and $\bar{\rho}$	13
3	between $\bar{\rho}$ and $\bar{\rho} + \sigma_\rho$	11
4	more than $\bar{\rho} + \sigma_\rho$	14

Assuming the measurements were normally distributed, with center $\bar{\rho}$ and width σ_ρ, calculate the expected number of measurements E_k in each bin. Hence, calculate χ^2. Do the measurements seem to be normally distributed?

12.2. ★★ Problem 4.13 reported 30 measurements of a time t, with mean $\bar{t} = 8.15$ sec and standard deviation $\sigma_t = 0.04$ sec. Group the values of t into four bins with boundaries at $\bar{t} - \sigma_t$, $\bar{t}$, and $\bar{t} + \sigma_t$, and count the observed number O_k in each bin $k = 1, 2, 3, 4$. Assuming the measurements were normally distributed with center at $\bar{t}$ and width σ_t, find the expected number E_k in each bin. Calculate χ^2. Is there any reason to doubt the measurements are normally distributed?

For Section 12.2: General Definition of Chi Squared

12.3. ★ A gambler decides to test a die by throwing it 240 times. Each throw has six possible outcomes, $k = 1, 2, \ldots, 6$, where k is the face showing, and the distribution of his throws is as shown in Table 12.11. If you treat each possible

Table 12.11. Number of occurrences of each face showing on a die thrown 240 times; for Problem 12.3.

Face showing k:	1	2	3	4	5	6
Occurrences O_k:	20	46	35	45	42	52

result k as a separate bin, what is the expected number E_k in each bin, assuming that the die is true? Compute χ^2. Does the die seem likely to be loaded?

12.4. ★★ I throw three dice together a total of 400 times, record the number of sixes in each throw, and obtain the results shown in Table 12.12. Assuming the dice

Table 12.12. Number of occurrences for each result for three dice thrown together 400 times; for Problem 12.4.

Result	Bin number k	Occurrences O_k
No sixes	1	217
One six	2	148
Two or three sixes	3	35

are true, use the binomial distribution to find the expected number E_k for each of the three bins and then calculate χ^2. Do I have reason to suspect the dice are loaded?

12.5. ★★ As a radioactive specimen decays, its activity decreases exponentially as the number of radioactive atoms diminishes. Some radioactive species have mean lives in the millions, and even billions, of years, and for such species the exponential decay of activity is not readily apparent. On the other hand, many species have mean lives of minutes or hours, and for such species the exponential decay is easily observed. The following problem illustrates this latter case.

A student wants to verify the exponential decay law for a material with a known mean life of τ. Starting at time $t = 0$, she counts the decays from a sample in successive intervals of length T; that is, she establishes bins with $0 < t < T$, and $T < t < 2T$, and so on, and she counts the number of decays in each bin. **(a)** According to the exponential decay law, the number of radioactive atoms that remain after time t is $N(t) = N_0 e^{-t/\tau}$, where N_0 is the number of atoms at $t = 0$. Deduce that the number of decays expected in the kth time bin $[(k-1)\,T < t < kT]$ is

$$E_k = N_0(e^{T/\tau} - 1)e^{-kT/\tau}. \tag{12.19}$$

(b) For convenience, the student chooses her time interval T equal to the known value of τ, and she gets the results shown in Table 12.13. What are the expected

Table 12.13. Observed number of decays in successive time intervals of a radioactive sample. Note that all decays that occurred after $t = 4T$ have been included in a single bin; for Problem 12.5.

Bin number k	Time interval	Observed decays O_k
1	$0 < t < T$	528
2	$T < t < 2T$	180
3	$2T < t < 3T$	71
4	$3T < t < 4T$	20
5	$4T < t$	16

numbers E_k, according to (12.19), and what is her value for χ^2? Are her results consistent with the exponential decay law? (Note that the initial number N_0 is easily found from the data, because it must be equal to the total number of decays after $t = 0$.)

For Section 12.3: Degrees of Freedom and Reduced Chi Squared

12.6. ★ **(a)** For the experiment of Problem 12.1, find the number of constraints c and the number of degrees of freedom d. **(b)** Suppose now that the accepted value ρ_{acc} of the density was known and that the students decided to test the hypothesis that the results were governed by a normal distribution centered on ρ_{acc}. For this test, how many constraints would there be, and how many degrees of freedom?

12.7. ★ For each of Problems 12.2 to 12.4, find the number of constraints c and the number of degrees of freedom d.

For Sections 12.4 and 12.5: Probabilities for Chi Squared and Examples

12.8. ★ If we observe the distribution of results in some experiment and know the expected distribution, we can calculate the observed $\tilde{\chi}_o^2$ and find the probability $Prob_d(\tilde{\chi}^2 \geq \tilde{\chi}_o^2)$. If this probability is less than 5%, we can then reject the expected distribution at the 5% level. For example, with two degrees of freedom ($d = 2$),

any value of $\tilde{\chi}_o^2$ greater than 3.0 would justify rejection of the expected distribution at the 5% level; that is, for $d = 2$, $\tilde{\chi}_o^2 = 3.0$ is the critical value above which we can reject the distribution at the 5% level. Use the probabilities in Appendix D to make a table of the corresponding critical values (5% level) of $\tilde{\chi}_o^2$, for $d = 1, 2, 3, 4, 5, 10, 15, 20, 30$.

12.9. ★ As in Problem 12.8, make a table of the critical values of $\tilde{\chi}_o^2$ for rejection of the expected distribution, but this time at the 1% level, for $d = 1, 2, 3, 4, 5, 10, 15, 20, 30$.

12.10. ★★ For the data of Problem 12.1, compute $\tilde{\chi}^2$. If the measurements were normally distributed, what was the probability of getting a value of $\tilde{\chi}^2$ this large or larger? At the 5% significance level, can you reject the hypothesis that the measurements were normally distributed? At the 1% level? (See Appendix D for the needed probabilities.)

12.11. ★★ In Problem 12.3, find the value of $\tilde{\chi}^2$. Can we conclude that the die was loaded at the 5% significance level? At the 1% level? (See Appendix D for the necessary probabilities.)

12.12. ★★ In Problem 12.4, find the value of $\tilde{\chi}^2$. If the dice really are true, what is the probability of getting a value of $\tilde{\chi}^2$ this large or larger? Explain whether the evidence suggests the dice are loaded. (See Appendix D for the necessary probabilities.)

12.13. ★★ Calculate χ^2 for the data of Problem 11.5, assuming the observations should follow the Poisson distribution with mean count $\mu = 3$. (Group all values $\nu \geqslant 6$ into a single bin.) How many degrees of freedom are there? (Don't forget that μ was given in advance and didn't have to be calculated from the data.) What is $\tilde{\chi}^2$? Are the data consistent with the expected Poisson distribution?

12.14. ★★ The chi-squared test can be used to test how well a set of measurements (x_i, y_i) of two variables fits an expected relation $y = f(x)$, provided the uncertainties are known reliably. Suppose y and x are expected to satisfy a linear relation

$$y = f(x) = A + Bx. \tag{12.20}$$

(For instance, y might be the length of a metal rod and x its temperature.) Suppose that A and B are predicted theoretically to be $A = 50$ and $B = 6$, and that five measurements have produced the values shown in Table 12.14. The uncertainty in

Table 12.14. Five measurements of two variables expected to fit the relation $y = A + Bx$; for Problem 12.14.

x (no uncertainty):	1	2	3	4	5
y (all ± 4):	60	56	71	66	86

the measurements of x is negligible; all the measurements of y have the same standard deviation, which is known to be $\sigma = 4$.

(a) Make a table of the observed and expected values of y_i, and calculate χ^2 as

$$\chi^2 = \sum_{i=1}^{5} \left(\frac{y_i - f(x_i)}{\sigma} \right)^2 .$$

(b) Because no parameters were calculated from the data, there are no constraints and hence five degrees of freedom. Calculate $\tilde{\chi}^2$, and use Appendix D to find the probability of obtaining a value of $\tilde{\chi}^2$ this large, assuming y does satisfy (12.20). At the 5% level, would you reject the expected relation (12.20)? (Note that if the constants A and B were not known in advance, you would have to calculate them from the data by the method of least squares. You would then proceed as before, except that there would now be two constraints and only three degrees of freedom.)

12.15. ★★★ Two dice are thrown together 360 times, and the total score is recorded for each throw. The possible totals are 2, 3, . . . , 12, and their numbers of occurrences are as shown in Table 12.15. **(a)** Calculate the probabilities for each

Table 12.15. Observed occurrences of total scores for two dice thrown together 360 times; for Problem 12.15.

Total score:	2	3	4	5	6	7	8	9	10	11	12
Occurrences:	6	14	23	35	57	50	44	49	39	27	16

total and hence the expected number of occurrences (assuming the dice are true). **(b)** Calculate χ^2, d, and $\tilde{\chi}^2 = \chi^2/d$. **(c)** Assuming the dice are true, what is the probability of getting a value of $\tilde{\chi}^2$ this large or larger? **(d)** At the 5% level of significance, can you reject the hypothesis that the dice are true? At the 1% level?

12.16. ★★★ A certain long-lived radioactive sample is alleged to produce an average of 2 decays per minute. To check this claim, a student counts the numbers of decays in 40 separate one-minute intervals and obtains the results shown in Table 12.16. (The mean life is so long that any depletion of the sample is negligible over

Table 12.16. Observed number of decays in one-minute intervals of a radioactive sample; for Problem 12.16.

Number of decays ν:	0	1	2	3	4	5 or more
Times observed:	11	12	11	4	2	0

the course of all these measurements.) **(a)** If the Poisson distribution that governs the decays really does have $\mu = 2$, what numbers E_k should the student expect to have found? (Group all observations with $\nu \geqslant 3$ into a single bin.) Calculate χ^2, d, and $\tilde{\chi}^2 = \chi^2/d$. (Don't forget that μ was not calculated from the data.) At the 5% significance level, would you reject the hypothesis that the sample follows the Poisson distribution with $\mu = 2$? **(b)** The student notices that the actual mean of his results is $\bar{\nu} = 1.35$, and he therefore decides to test whether the data fit a Poisson distribution with $\mu = 1.35$. In this case, what are d and $\tilde{\chi}^2$? Are the data consistent with this new hypothesis?

12.17. ★★★ Chapter 10 described a test for fit to the binomial distribution. We considered n trials, each with two possible outcomes: "success" (with probability p) and "failure" (with probability $1 - p$). We then tested whether the observed number of successes, ν, was compatible with some assumed value of p. Provided the numbers involved are reasonably large, we can also treat this same problem with the chi-squared test, with just two bins—$k = 1$ for successes and $k = 2$ for failures—and one degree of freedom. In the following problem, you will use both methods and compare results. When the numbers are large, you will find the agreement is excellent; when they are small, it is less so but still good enough that chi squared is a useful indicator.

(a) A soup manufacturer believes he can introduce a different dumpling into his chicken dumpling soup without noticeably affecting the flavor. To test this hypothesis, he makes 16 cans labeled "Style X" that contain the old dumpling and 16 cans labeled "Style Y" that contain the new dumpling. He sends one of each type to 16 tasters and asks them which they prefer. If his hypothesis is correct, we should expect eight tasters to prefer X and eight to prefer Y. In the actual test, the number who favored X was $\nu = 11$. Calculate χ^2 and the probability of getting a value this large or larger. Does the test indicate a significant difference between the two kinds of dumpling? Now, calculate the corresponding probability exactly, using the binomial distribution, and compare your results. (Note that the chi-squared test includes deviations away from the expected numbers in both directions. Therefore, for this comparison you should calculate the two-tailed probability for values of ν that deviate from 8 by 3 or more in either direction; that is, $\nu = 11, 12, \ldots, 16$ *and* $\nu = 5, 4, \ldots, 1$.)

(b) Repeat part (a) for the next test, in which the manufacturer makes 400 cans of each style and the number of tasters who prefer X is 225. (In calculating the binomial probabilities, use the Gaussian approximation.)

(c) In part (a), the numbers were small enough that the chi-squared test was fairly crude. (It gave a probability of 13%, compared with the correct value of 21%.) With one degree of freedom, you can improve on the chi-squared test by using an *adjusted* chi-squared, defined as

$$(\text{adjusted } \chi^2) = \sum_{k=1}^{2} \frac{\left(|O_k - E_k| - \frac{1}{2}\right)^2}{E_k}.$$

Calculate the adjusted χ^2 for the data of part (a) and show that the use of this value, instead of the ordinary χ^2, in the table of Appendix D gives a more accurate value of the probability.[3]

[3] We have not justified the use of the adjusted chi squared here, but this example does illustrate its superiority. For more details, see H. L. Alder and E. B. Roessler, *Introduction to Probability and Statistics,* 6th ed. (W. H. Freeman, 1977), p. 263.

Appendixes

Appendix A

Normal Error Integral, I

If the measurement of a continuous variable x is subject to many small errors, all of them random, the expected distribution of results is given by the normal, or Gauss, distribution,

$$G_{X,\sigma}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-X)^2/2\sigma^2},$$

where X is the true value of x, and σ is the standard deviation.

The integral of the normal distribution function, $\int_a^b G_{X,\sigma}(x)dx$, is called the *normal error integral,* and is the probability that a measurement falls between $x = a$ and $x = b$,

$$Prob(a \leqslant x \leqslant b) = \int_a^b G_{X,\sigma}(x)\, dx.$$

Table A shows this integral for $a = X - t\sigma$ and $b = X + t\sigma$. This gives the probability of a measurement within t standard deviations on either side of X,

$$\begin{aligned} Prob(\text{within } t\sigma) &= Prob(X - t\sigma \leqslant x \leqslant X + t\sigma) \\ &= \int_{X-t\sigma}^{X+t\sigma} G_{X,\sigma}(x)\, dx = \frac{1}{\sqrt{2\pi}} \int_{-t}^{t} e^{-z^2/2}\, dz. \end{aligned}$$

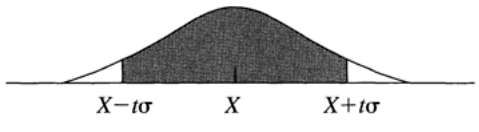
This function is sometimes denoted erf(t), but this notation is also used for a slightly different function.

The probability of a measurement *outside* the same interval can be found by subtraction;

$$Prob(\text{outside } t\sigma) = 100\% - Prob(\text{within } t\sigma).$$

For further discussions, see Section 5.4 and Appendix B.

Table A. The percentage probability, $Prob(\text{within } t\sigma) = \int_{X-t\sigma}^{X+t\sigma} G_{X,\sigma}(x)\,dx$, as a function of t.



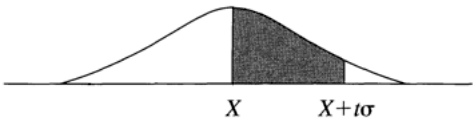
t	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00	0.80	1.60	2.39	3.19	3.99	4.78	5.58	6.38	7.17
0.1	7.97	8.76	9.55	10.34	11.13	11.92	12.71	13.50	14.28	15.07
0.2	15.85	16.63	17.41	18.19	18.97	19.74	20.51	21.28	22.05	22.82
0.3	23.58	24.34	25.10	25.86	26.61	27.37	28.12	28.86	29.61	30.35
0.4	31.08	31.82	32.55	33.28	34.01	34.73	35.45	36.16	36.88	37.59
0.5	38.29	38.99	39.69	40.39	41.08	41.77	42.45	43.13	43.81	44.48
0.6	45.15	45.81	46.47	47.13	47.78	48.43	49.07	49.71	50.35	50.98
0.7	51.61	52.23	52.85	53.46	54.07	54.67	55.27	55.87	56.46	57.05
0.8	57.63	58.21	58.78	59.35	59.91	60.47	61.02	61.57	62.11	62.65
0.9	63.19	63.72	64.24	64.76	65.28	65.79	66.29	66.80	67.29	67.78
1.0	68.27	68.75	69.23	69.70	70.17	70.63	71.09	71.54	71.99	72.43
1.1	72.87	73.30	73.73	74.15	74.57	74.99	75.40	75.80	76.20	76.60
1.2	76.99	77.37	77.75	78.13	78.50	78.87	79.23	79.59	79.95	80.29
1.3	80.64	80.98	81.32	81.65	81.98	82.30	82.62	82.93	83.24	83.55
1.4	83.85	84.15	84.44	84.73	85.01	85.29	85.57	85.84	86.11	86.38
1.5	86.64	86.90	87.15	87.40	87.64	87.89	88.12	88.36	88.59	88.82
1.6	89.04	89.26	89.48	89.69	89.90	90.11	90.31	90.51	90.70	90.90
1.7	91.09	91.27	91.46	91.64	91.81	91.99	92.16	92.33	92.49	92.65
1.8	92.81	92.97	93.12	93.28	93.42	93.57	93.71	93.85	93.99	94.12
1.9	94.26	94.39	94.51	94.64	94.76	94.88	95.00	95.12	95.23	95.34
2.0	95.45	95.56	95.66	95.76	95.86	95.96	96.06	96.15	96.25	96.34
2.1	96.43	96.51	96.60	96.68	96.76	96.84	96.92	97.00	97.07	97.15
2.2	97.22	97.29	97.36	97.43	97.49	97.56	97.62	97.68	97.74	97.80
2.3	97.86	97.91	97.97	98.02	98.07	98.12	98.17	98.22	98.27	98.32
2.4	98.36	98.40	98.45	98.49	98.53	98.57	98.61	98.65	98.69	98.72
2.5	98.76	98.79	98.83	98.86	98.89	98.92	98.95	98.98	99.01	99.04
2.6	99.07	99.09	99.12	99.15	99.17	99.20	99.22	99.24	99.26	99.29
2.7	99.31	99.33	99.35	99.37	99.39	99.40	99.42	99.44	99.46	99.47
2.8	99.49	99.50	99.52	99.53	99.55	99.56	99.58	99.59	99.60	99.61
2.9	99.63	99.64	99.65	99.66	99.67	99.68	99.69	99.70	99.71	99.72
3.0	99.73									
3.5	99.95									
4.0	99.994									
4.5	99.9993									
5.0	99.99994									

Appendix B

Normal Error Integral, II

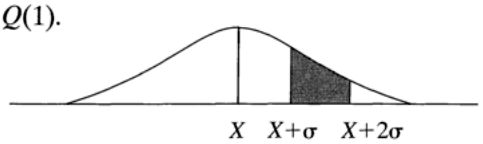
In certain calculations, a convenient form of the normal error integral is

$$Q(t) = \int_X^{X+t\sigma} G_{X,\sigma}(x)\,dx$$
$$= \frac{1}{\sqrt{2\pi}} \int_0^t e^{-z^2/2}\,dz.$$



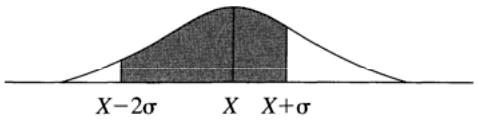
(This integral is, of course, just half the integral tabulated in Appendix A.) The probability $Prob(a \leqslant x \leqslant b)$ of a measurement in any interval $a \leqslant x \leqslant b$ can be found from $Q(t)$ by a single subtraction or addition. For example,

$$Prob(X + \sigma \leqslant x \leqslant X + 2\sigma) = Q(2) - Q(1).$$



Similarly,

$$Prob(X - 2\sigma \leqslant x \leqslant X + \sigma) = Q(2) + Q(1).$$



The probability of a measurement greater than any $X + t\sigma$ is just $0.5 - Q(t)$. For example,

$$Prob(x \geqslant X + \sigma) = 50\% - Q(1).$$

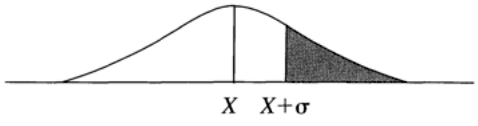
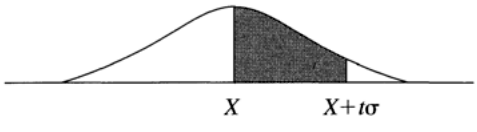


Table B. The percentage probability, $Q(t) = \int_X^{X+t\sigma} G_{X,\sigma}(x)\,dx$, as a function of t.



t	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.00	0.40	0.80	1.20	1.60	1.99	2.39	2.79	3.19	3.59
0.1	3.98	4.38	4.78	5.17	5.57	5.96	6.36	6.75	7.14	7.53
0.2	7.93	8.32	8.71	9.10	9.48	9.87	10.26	10.64	11.03	11.41
0.3	11.79	12.17	12.55	12.93	13.31	13.68	14.06	14.43	14.80	15.17
0.4	15.54	15.91	16.28	16.64	17.00	17.36	17.72	18.08	18.44	18.79
0.5	19.15	19.50	19.85	20.19	20.54	20.88	21.23	21.57	21.90	22.24
0.6	22.57	22.91	23.24	23.57	23.89	24.22	24.54	24.86	25.17	25.49
0.7	25.80	26.11	26.42	26.73	27.04	27.34	27.64	27.94	28.23	28.52
0.8	28.81	29.10	29.39	29.67	29.95	30.23	30.51	30.78	31.06	31.33
0.9	31.59	31.86	32.12	32.38	32.64	32.89	33.15	33.40	33.65	33.89
1.0	34.13	34.38	34.61	34.85	35.08	35.31	35.54	35.77	35.99	36.21
1.1	36.43	36.65	36.86	37.08	37.29	37.49	37.70	37.90	38.10	38.30
1.2	38.49	38.69	38.88	39.07	39.25	39.44	39.62	39.80	39.97	40.15
1.3	40.32	40.49	40.66	40.82	40.99	41.15	41.31	41.47	41.62	41.77
1.4	41.92	42.07	42.22	42.36	42.51	42.65	42.79	42.92	43.06	43.19
1.5	43.32	43.45	43.57	43.70	43.82	43.94	44.06	44.18	44.29	44.41
1.6	44.52	44.63	44.74	44.84	44.95	45.05	45.15	45.25	45.35	45.45
1.7	45.54	45.64	45.73	45.82	45.91	45.99	46.08	46.16	46.25	46.33
1.8	46.41	46.49	46.56	46.64	46.71	46.78	46.86	46.93	46.99	47.06
1.9	47.13	47.19	47.26	47.32	47.38	47.44	47.50	47.56	47.61	47.67
2.0	47.72	47.78	47.83	47.88	47.93	47.98	48.03	48.08	48.12	48.17
2.1	48.21	48.26	48.30	48.34	48.38	48.42	48.46	48.50	48.54	48.57
2.2	48.61	48.64	48.68	48.71	48.75	48.78	48.81	48.84	48.87	48.90
2.3	48.93	48.96	49.98	49.01	49.04	49.06	49.09	49.11	49.13	49.16
2.4	49.18	49.20	49.22	49.25	49.27	49.29	49.31	49.32	49.34	49.36
2.5	49.38	49.40	49.41	49.43	49.45	49.46	49.48	49.49	49.51	49.52
2.6	49.53	49.55	49.56	49.57	49.59	49.60	49.61	49.62	49.63	49.64
2.7	49.65	49.66	49.67	49.68	49.69	49.70	49.71	49.72	49.73	49.74
2.8	49.74	49.75	49.76	49.77	49.77	49.78	49.79	49.79	49.80	49.81
2.9	49.81	49.82	49.82	49.83	49.84	49.84	49.85	49.85	49.86	49.86
3.0	49.87									
3.5	49.98									
4.0	49.997									
4.5	49.9997									
5.0	49.99997									

Appendix C

Probabilities for Correlation Coefficients

The extent to which N points $(x_1, y_1), \ldots, (x_N, y_N)$ fit a straight line is indicated by the linear correlation coefficient

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\Sigma(x_i - \bar{x})^2 \Sigma(y_i - \bar{y})^2}},$$

which always lies in the interval $-1 \leq r \leq 1$. Values of r close to ± 1 indicate a good linear correlation; values close to 0 indicate little or no correlation.

A more quantitative measure of the fit can be found by using Table C. For any given observed value r_o, $Prob_N(|r| \geq |r_o|)$ is the probability that N measurements of two uncorrelated variables would give a coefficient r as large as r_o. Thus, if we obtain a coefficient r_o for which $Prob_N(|r| \geq |r_o|)$ is small, it is correspondingly unlikely that our variables are uncorrelated; that is, a correlation is indicated. In particular, if $Prob_N(|r| \geq |r_o|) \leq 5\%$, the correlation is called *significant*; if it is less than 1%, the correlation is called *highly significant.*

For example, the probability that 20 measurements ($N = 20$) of two uncorrelated variables would yield $|r| \geq 0.5$ is given in the table as 2.5%. Thus, if 20 measurements gave $r = 0.5$, we would have *significant* evidence of a linear correlation between the two variables. For further discussion, see Sections 9.3 to 9.5.

The values in Table C were calculated from the integral

$$Prob_N(|r| \geq |r_o|) = \frac{2\Gamma[(N-1)/2]}{\sqrt{\pi}\Gamma[(N-2)/2]} \int_{|r_o|}^{1} (1 - r^2)^{(N-4)/2}\, dr.$$

See, for example, E. M. Pugh and G. H. Winslow, *The Analysis of Physical Measurements* (Addison-Wesley, 1966), Section 12-8.

Table C. The percentage probability $Prob_N(|r| \geq r_o)$ that N measurements of two uncorrelated variables give a correlation coefficient with $|r| \geq r_o$, as a function of N and r_o. (Blanks indicate probabilities less than 0.05%.)

	r_o										
N	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
3	100	94	87	81	74	67	59	51	41	29	0
4	100	90	80	70	60	50	40	30	20	10	0
5	100	87	75	62	50	39	28	19	10	3.7	0
6	100	85	70	56	43	31	21	12	5.6	1.4	0
7	100	83	67	51	37	25	15	8.0	3.1	0.6	0
8	100	81	63	47	33	21	12	5.3	1.7	0.2	0
9	100	80	61	43	29	17	8.8	3.6	1.0	0.1	0
10	100	78	58	40	25	14	6.7	2.4	0.5		0
11	100	77	56	37	22	12	5.1	1.6	0.3		0
12	100	76	53	34	20	9.8	3.9	1.1	0.2		0
13	100	75	51	32	18	8.2	3.0	0.8	0.1		0
14	100	73	49	30	16	6.9	2.3	0.5	0.1		0
15	100	72	47	28	14	5.8	1.8	0.4			0
16	100	71	46	26	12	4.9	1.4	0.3			0
17	100	70	44	24	11	4.1	1.1	0.2			0
18	100	69	43	23	10	3.5	0.8	0.1			0
19	100	68	41	21	9.0	2.9	0.7	0.1			0
20	100	67	40	20	8.1	2.5	0.5	0.1			0
25	100	63	34	15	4.8	1.1	0.2				0
30	100	60	29	11	2.9	0.5					0
35	100	57	25	8.0	1.7	0.2					0
40	100	54	22	6.0	1.1	0.1					0
45	100	51	19	4.5	0.6						0
	0	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	
50	100	73	49	30	16	8.0	3.4	1.3	0.4	0.1	
60	100	70	45	25	13	5.4	2.0	0.6	0.2		
70	100	68	41	22	9.7	3.7	1.2	0.3	0.1		
80	100	66	38	18	7.5	2.5	0.7	0.1			
90	100	64	35	16	5.9	1.7	0.4	0.1			
100	100	62	32	14	4.6	1.2	0.2				

Appendix D

Probabilities for Chi Squared

If a series of measurements is grouped into bins $k = 1, \ldots, n$, we denote by O_k the number of measurements observed in the bin k. The number *expected* (on the basis of some assumed or expected distribution) in the bin k is denoted by E_k. The extent to which the observations fit the assumed distribution is indicated by the reduced chi squared, $\tilde{\chi}^2$, defined as

$$\tilde{\chi}^2 = \frac{1}{d} \sum_{k=1}^{n} \frac{(O_k - E_k)^2}{E_k},$$

where d is the number of degrees of freedom, $d = n - c$, and c is the number of constraints (see Section 12.3). The expected average value of $\tilde{\chi}^2$ is 1. If $\tilde{\chi}^2 \gg 1$, the observed results do not fit the assumed distribution; if $\tilde{\chi}^2 \leqslant 1$, the agreement is satisfactory.

This test is made quantitative with the probabilities shown in Table D. Let $\tilde{\chi}_o^2$ denote the value of $\tilde{\chi}^2$ actually obtained in an experiment with d degrees of freedom. The number $Prob_d(\tilde{\chi}^2 \geqslant \tilde{\chi}_o^2)$ is the probability of obtaining a value of $\tilde{\chi}^2$ as large as the observed $\tilde{\chi}_o^2$, if the measurements really did follow the assumed distribution. Thus, if $Prob_d(\tilde{\chi}^2 \geqslant \tilde{\chi}_o^2)$ is large, the observed and expected distributions are consistent; if it is small, they probably disagree. In particular, if $Prob_d(\tilde{\chi}^2 \geqslant \tilde{\chi}_o^2)$ is less than 5%, we say the disagreement is *significant* and reject the assumed distribution at the 5% level. If it is less than 1%, the disagreement is called *highly significant,* and we reject the assumed distribution at the 1% level.

For example, suppose we obtain a reduced chi squared of 2.6 (that is, $\tilde{\chi}_o^2 = 2.6$) in an experiment with six degrees of freedom ($d = 6$). According to Table D, the probability of getting $\tilde{\chi}^2 \geqslant 2.6$ is 1.6%, if the measurements were governed by the assumed distribution. Thus, at the 5% level (but not quite at the 1% level), we would reject the assumed distribution. For further discussion, see Chapter 12.

Table D. The percentage probability $Prob_d(\tilde{\chi}^2 \geq \tilde{\chi}_o^2)$ of obtaining a value of $\tilde{\chi}^2 \geq \tilde{\chi}_o^2$ in an experiment with d degrees of freedom, as a function of d and $\tilde{\chi}_o^2$. (Blanks indicate probabilities less than 0.05%.)

	$\tilde{\chi}_o^2$															
d	0	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5	6.0	8.0	10.0	
1	100	48	32	22	16	11	8.3	6.1	4.6	3.4	2.5	1.9	1.4	0.5	0.2	
2	100	61	37	22	14	8.2	5.0	3.0	1.8	1.1	0.7	0.4	0.2			
3	100	68	39	21	11	5.8	2.9	1.5	0.7	0.4	0.2	0.1				
4	100	74	41	20	9.2	4.0	1.7	0.7	0.3	0.1	0.1					
5	100	78	42	19	7.5	2.9	1.0	0.4	0.1							
	0	0.2	0.4	0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0	2.2	2.4	2.6	2.8	3.0
1	100	65	53	44	37	32	27	24	21	18	16	14	12	11	9.4	8.3
2	100	82	67	55	45	37	30	25	20	17	14	11	9.1	7.4	6.1	5.0
3	100	90	75	61	49	39	31	24	19	14	11	8.6	6.6	5.0	3.8	2.9
4	100	94	81	66	52	41	31	23	17	13	9.2	6.6	4.8	3.4	2.4	1.7
5	100	96	85	70	55	42	31	22	16	11	7.5	5.1	3.5	2.3	1.6	1.0
6	100	98	88	73	57	42	30	21	14	9.5	6.2	4.0	2.5	1.6	1.0	0.6
7	100	99	90	76	59	43	30	20	13	8.2	5.1	3.1	1.9	1.1	0.7	0.4
8	100	99	92	78	60	43	29	19	12	7.2	4.2	2.4	1.4	0.8	0.4	0.2
9	100	99	94	80	62	44	29	18	11	6.3	3.5	1.9	1.0	0.5	0.3	0.1
10	100	100	95	82	63	44	29	17	10	5.5	2.9	1.5	0.8	0.4	0.2	0.1
11	100	100	96	83	64	44	28	16	9.1	4.8	2.4	1.2	0.6	0.3	0.1	0.1
12	100	100	96	84	65	45	28	16	8.4	4.2	2.0	0.9	0.4	0.2	0.1	
13	100	100	97	86	66	45	27	15	7.7	3.7	1.7	0.7	0.3	0.1	0.1	
14	100	100	98	87	67	45	27	14	7.1	3.3	1.4	0.6	0.2	0.1		
15	100	100	98	88	68	45	26	14	6.5	2.9	1.2	0.5	0.2	0.1		
16	100	100	98	89	69	45	26	13	6.0	2.5	1.0	0.4	0.1			
17	100	100	99	90	70	45	25	12	5.5	2.2	0.8	0.3	0.1			
18	100	100	99	90	70	46	25	12	5.1	2.0	0.7	0.2	0.1			
19	100	100	99	91	71	46	25	11	4.7	1.7	0.6	0.2	0.1			
20	100	100	99	92	72	46	24	11	4.3	1.5	0.5	0.1				
22	100	100	99	93	73	46	23	10	3.7	1.2	0.4	0.1				
24	100	100	100	94	74	46	23	9.2	3.2	0.9	0.3	0.1				
26	100	100	100	95	75	46	22	8.5	2.7	0.7	0.2					
28	100	100	100	95	76	46	21	7.8	2.3	0.6	0.1					
30	100	100	100	96	77	47	21	7.2	2.0	0.5	0.1					

The values in Table D were calculated from the integral

$$Prob_d(\tilde{\chi}^2 \geq \tilde{\chi}_o^2) = \frac{2}{2^{d/2}\Gamma(d/2)} \int_{\chi_o}^{\infty} x^{d-1} e^{-x^2/2} dx.$$

See, for example, E. M. Pugh and G. H. Winslow, *The Analysis of Physical Measurements* (Addison-Wesley, 1966), Section 12-5.

Appendix E

Two Proofs Concerning Sample Standard Deviations

In Chapter 5, I quoted without proof two important results concerning measurements of a quantity x normally distributed with width σ: (1) The best estimate of σ based on N measurements of x is the sample standard deviation, σ_x, of the N measurements, as defined by (5.45) with the factor of $(N - 1)$ in the denominator. (2) The fractional uncertainity in σ_x as an estimate of σ is $1/\sqrt{2(N-1)}$, as in (5.46). The proofs of these two results are surprisingly awkward and were omitted from Chapter 5. For those who like to see proofs, I give them here.

Consider an experiment that consists of N measurements (all using the same method) of a quantity x, with the results

$$x_1, x_2, \ldots, x_N.$$

The sample standard deviation of these N measurements is defined by

$$\begin{aligned} \sigma_x^2 &= \frac{\Sigma(x_i - \bar{x})^2}{N-1} \\ &= \frac{SS}{N-1}, \end{aligned} \tag{E1}$$

where I have introduced the notation SS for the *Sum of Squares,*

$$\begin{aligned} SS &= \text{Sum of Squares} \\ &= \textstyle\sum_1^N (x_i - \bar{x})^2 \\ &= \textstyle\sum_1^N (x_i)^2 - \frac{1}{N}\left(\sum_1^N x_i\right)^2. \end{aligned} \tag{E2}$$

In writing the last line of (E2), I have used the identity (4.28) from Problem 4.5. Equation (E1) actually defines σ_x *squared*, which is called the *sample variance* of the N measurements. To avoid numerous square root signs, I will work mostly with the variance rather than the standard deviation itself.

Our proofs will be simplified by noting that neither the true width σ nor the sample standard deviation σ_x is changed if we subtract any fixed constant from our measured quantity x. In particular, we can subtract from x its true value X, which gives a quantity normally distributed about a true value of zero. In other words, *we can (and will) assume that the true value of our measured quantity* x is $X = 0$.

To discuss our best estimate for σ and its uncertainty, we must imagine repeating our whole experiment an enormous ("infinite") number of times $\alpha = 1, 2, 3, \ldots$. In this way we generate an immense array of measurements, which we can display as in Table E1.

Table E1. Results of a large number of experiments, $\alpha = 1, 2, 3, \ldots$, each of which consists of N measurements of a quantity x. The ith measurement in the αth experiment is denoted by $x_{\alpha i}$ and is shown in the ith column of the αth row.

Experiment number	1st measurement	2nd measurement	....	ith measurement	....	Nth measurement
1	x_{11}	x_{12}	....	x_{1i}	....	x_{1N}
2	x_{21}	x_{22}	....	x_{2i}	....	x_{2N}
.....	....	....	....	....	....	....
α	$x_{\alpha 1}$	$x_{\alpha 2}$	....	$x_{\alpha i}$	....	$x_{\alpha N}$
.....	....	....	....	....	....	....
.....	....	....	....	....	....	....

For each of the infinitely many experiments, we can calculate the mean and variance. For example, for the αth experiment (or αth "sampling"), we find the sample mean $\bar{x}_\alpha$ by adding all the numbers in the row of the αth experiment and dividing by N. To find the corresponding sample variance, we compute the sum of squares

$$SS_\alpha = \sum_{i=1}^{N} (x_{\alpha i} - \bar{x}_\alpha)^2$$

and divide by $(N - 1)$. Here again, the sum runs over all entries in the appropriate row of data.

We use the usual symbol of a bar, as in $\bar{x}_\alpha$, to indicate an average over all measurements in one sample (that is, in one row of our data). We also need to consider the average of all numbers in a given column of our data (that is, an average over all $\alpha = 1, 2, 3, \ldots$), and we denote this kind of average by two brackets $\langle . . \rangle$. For example, we denote by $\langle x_{\alpha i} \rangle$, or just $\langle x \rangle_i$, the average of all the measurements in the ith column. Because this value is the average of infinitely many measurements of x, it equals the true value X, which we have arranged to be zero. Thus,

$$\langle x \rangle_i = X = 0.$$

Similarly, if we average the *squares* of all the values in any column, we will get the true variance

$$\langle x^2 \rangle_i = \sigma^2. \tag{E3}$$

[Remember that $X = 0$, so that $\langle x^2 \rangle_i$ is the same as $\langle (x - X)^2 \rangle_i$, which is just σ^2.] Armed with these ideas, we are ready for our two proofs.

I Best Estimate for the Width σ

We want to show that, based on N measurements of x, the best estimate for the true width σ is the sample standard deviation of the N measurements. We do this by proving the following proposition: If we calculate the sum of squares SS_α for each sample α and then average the sums over all $\alpha = 1, 2, 3, \ldots$, the result is $(N - 1)$ times the true variance σ^2:

$$\langle SS \rangle = (N - 1)\sigma^2. \tag{E4}$$

Dividing (E4) by $(N - 1)$, we see the true variance is $\sigma^2 = \langle SS \rangle/(N - 1)$. Here, $\langle SS \rangle$ is the average of the sums of squares from infinitely many sets of N measurements. This, in turn, means that the best estimate for σ^2 based on a *single* set of N measurements is just $SS/(N - 1)$, where SS is the sum of squares for that single set of measurements. This is just the variance given in Equation (E1). Thus, if we can prove (E4), we will have established the desired result.

To prove Equation (E4), we start with the sum of squares SS_α for the αth sample. Using (E2), we can write this sum as

$$SS_\alpha = \sum_i (x_{\alpha i})^2 - \frac{1}{N} \sum_i \sum_j x_{\alpha i} x_{\alpha j}. \tag{E5}$$

This equation differs from (E2) only in that I have added the subscripts α and have written out the square in the last term of (E2) as a product of two sums. The double sum in (E5) consists of two parts: First, there are N terms with $i = j$, which can be combined with the single sum in (E5). This leaves $N(N - 1)$ terms with $i \neq j$. Thus, we can rewrite (E5) as

$$SS_\alpha = \left(1 - \frac{1}{N}\right) \sum_i (x_{\alpha i})^2 - \frac{1}{N} \sum_i \sum_{j \neq i} x_{\alpha i} x_{\alpha j}. \tag{E6}$$

The expression (E6) is the sum of squares for a single sample α. To find $\langle SS \rangle$, we have only to average (E6) over all values of α. Because the average of any sum equals the sum of the corresponding averages, (E6) implies that

$$\langle SS \rangle = \frac{N - 1}{N} \sum_i \langle (x_{\alpha i})^2 \rangle - \frac{1}{N} \sum_i \sum_{j \neq i} \langle x_{\alpha i} x_{\alpha j} \rangle. \tag{E7}$$

The N terms in the first sum are all the same, and each is equal to σ^2 [as in (E3)]. Therefore, the first sum is just $N\sigma^2$. With $i \neq j$, the terms of the double sum in (E7) are all zero, $\langle x_{\alpha i} x_{\alpha j} \rangle = 0$. (Remember that x is normally distributed about $X = 0$. Thus, for each possible value of $x_{\alpha i}$, positive and negative values of $x_{\alpha j}$ are equally likely and cancel each other out.) Thus, the whole second sum is zero, and we are left with

$$\langle SS \rangle = (N - 1)\sigma^2,$$

which is precisely the result (E4) we needed to prove.

2 Uncertainty in the Estimate for the Width σ

We have shown that, based on N measurements of x, the best estimate for the true variance σ^2 is the sample variance of the N measurements, as defined in (E1), $SS/(N-1)$. Therefore, the fractional uncertainty in σ^2 is the same as that in the sum of squares SS:

$$\text{(fractional uncertainty in estimate for } \sigma^2) = \text{(fractional uncertainty in } SS). \tag{E8}$$

Because σ is the square root of σ^2, the fractional uncertainty in σ is just half of this:

$$\text{(fractional uncertainty in estimate for } \sigma) = \frac{1}{2}\,\text{(fractional uncertainty in } SS). \tag{E9}$$

To find the uncertainty in SS, we use the result (11.7) that the uncertainty in any quantity q is $\sqrt{\langle q^2\rangle - \langle q\rangle^2}$, where, as usual, the brackets $\langle\,.\,.\rangle$ denote the average of the quantity concerned after an infinite number of measurements. Therefore, the uncertainty in the sum of squares SS is

$$\text{(uncertainty in } SS) = \sqrt{\langle SS^2\rangle - \langle SS\rangle^2}. \tag{E10}$$

We already know [Equation (E4)] that the second average in this square root is $\langle SS\rangle = (N-1)\sigma^2$, so all that remains is to find the first term $\langle SS^2\rangle$.

For any one set of N measurements, the sum of squares SS is given by (E6). Squaring this expression, we find that (I omit all subscripts α to reduce the clutter)

$$\begin{aligned} SS^2 &= \left(\frac{N-1}{N}\right)^2 \sum_i (x_i)^2 \sum_j (x_j)^2 \\ &\quad - 2\frac{N-1}{N^2} \sum_i (x_i)^2 \sum_j \sum_{k\neq j} x_j x_k \\ &\quad + \frac{1}{N^2} \sum_j \sum_{k\neq j} x_j x_k \sum_m \sum_{n\neq m} x_m x_n \\ &= A + B + C. \end{aligned} \tag{E11}$$

To find the average value of (E11), we need to evaluate each of the three averages $\langle A\rangle$, $\langle B\rangle$, and $\langle C\rangle$. The double sum in A contains N terms with $i = j$, each of which averages to $\langle x^4\rangle$, and $N(N-1)$ terms with $i \neq j$, each of which averages to $\langle x^2\rangle^2$. Because x is distributed normally about $X = 0$, $\langle x^2\rangle = \sigma^2$ (as we already knew), and a straightforward integration shows that $\langle x^4\rangle = 3\sigma^4$. Thus, the double sum in A averages to $3N\sigma^4 + N(N-1)\sigma^4 = N(N+1)\sigma^4$, and

$$\langle A\rangle = \frac{(N-1)^2(N+2)}{N}\,\sigma^4.$$

Every term in the sum of B can easily be seen to contain an odd power of x_k (either x_k or $x_k{}^3$). Because x is normally distributed about 0, the average of any odd power

is zero, so

$$\langle B \rangle = 0.$$

The quadruple sum in the term C contains $N(N-1)$ terms in which $j = m$ and $k = n$, each of which averages to $\langle x^2 \rangle^2 = \sigma^4$. There is the same number of terms in which $j = n$ and $k = m$, each of which also averages to σ^4. All the remaining terms contain an odd power of x and average to zero. Thus, the quadruple sum in C averages to $2N(N-1)\sigma^4$, and

$$\langle C \rangle = \frac{2(N-1)}{N}\sigma^4.$$

Adding together the last three equations and inserting the result into (E11), we conclude that

$$\langle SS^2 \rangle = \frac{(N-1)^2(N+2) + 2(N-1)}{N}\sigma^4 = (N^2-1)\sigma^4.$$

Inserting this result into (E10) [and replacing $\langle SS \rangle$ by $(N-1)\sigma^2$], we find that

$$(\text{uncertainty in } SS) = \sqrt{(N^2-1) - (N-1)^2}\,\sigma^2 = \sqrt{2(N-1)}\,\sigma^2.$$

If we divide through by $\langle SS \rangle = (N-1)\sigma^2$, this result implies that the fractional uncertainty in SS is $\sqrt{2/(N-1)}$. Finally, we know [from (E9)] that the fractional uncertainty in our value of σ is half of that in SS. So,

$$(\text{fractional uncertainty in estimate for } \sigma) = \frac{1}{\sqrt{2(N-1)}},$$

which is the required result.