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,

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

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.

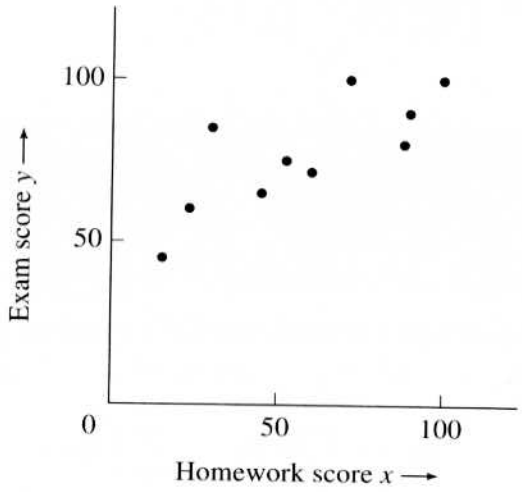

Figure 9.1. A "scatter plot" showing students' scores on exams and homework. Each of the 10 points (x_i, y_i) shows a student's homework score, x_i, and exam score, y_i.

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.

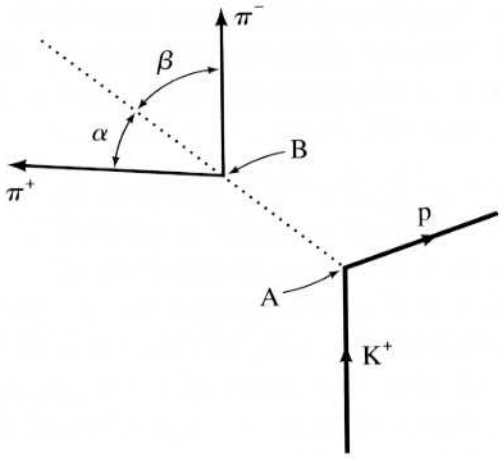

Figure 9.2. Tracks of charged particles in a bubble chamber. The dotted line shows the direction of an invisible K°, which was formed at A and decayed at B; for Problem 9.6.

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),

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

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

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

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:

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