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),

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